Journal of Medical Internet Research

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

Contents

Reviews

Innovations in Deaf Health Care Communication: Systematic Review of Sign Language Recognition Systems (e70417)
Milena Marcolino, Lucca Oliveira, Lucas Valle, Luiza Rosa, Gabriela Sanches, Natalia Santos, Michelle Costa, Elidea Bernardino, Raniere Cordeiro, Raquel Prates, Zilma Reis, Mario Campos. 28

Machine Learning in Predicting the Risk of Esophagogastric Variceal Bleeding Among Patients With Liver Cirrhosis: Systematic Review and Meta-Analysis (e78203)
Yuan Lian, Xinping Qiu, Geli Wang, Nannan Liu, Jing Zhao, Shanshan Song, Shiqi Wang, Mingjun Sun. 47

The Role of Digital Biomarkers in Physiological Signal-Based Depression Assessment: Systematic Review and Meta-Analysis (e76432)
Hyeongsuk Lee, Seung-Gul Kang, SeonHeui Lee. 70

Accuracy of Radiomics-Based Machine Learning for Predicting Risk of Recurrence in Non–Small Cell Lung Cancer: Systematic Review and Meta-Analysis (e84223)
Junpei Wu, Ye Zhang, Jiaye Wang, Chengshui Chen, Shiyu Hu, Wenyu Chen. 95

Multimodal AI for Alzheimer Disease Diagnosis: Systematic Review of Datasets, Models, and Modalities (e85414)
Ziwen Yu, Anthony Mulholland, Tianyan Huang, Qiang Liu. 109

Effectiveness of Mobile Health for Improving Medication Adherence in Patients With Cancer: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e85949)
Wei Liu, Li Wang, Qian Zhang, Jing Yang. 145

Artificial Intelligence Tools for Automating Evidence Synthesis: Scoping Review (e81597)
Sashika Harasgama, Helen Pearce, Cameron Appel, Liam Loftus, Helena Painter, Isla Kuhn, Justine Karpusheff, Aji Ceesay, John Ford. . . .
1 7 1

Smartphone-Based Digital Phenotyping Across Health Conditions: Scoping Review (e84146)
Arlen Dumas, Joanne Hokayem, Georgia Goodman, Krishna Venkatasubramanian, Peter Chai. 184

Effects of Virtual Reality–Based Interventions for Promoting Physical Activity in Patients With Heart Failure: Systematic Review (e86567)
Jeong-Ah Ahn, Jung Lee, Kyoung-A Kim. 214

XSL•FO
RenderX

The Current Landscape of Remote Digital Symptom Monitoring for Patients With Lung Cancer: Scoping Review (e83666)
Hao Huang, Naomi Takemura, Hammoda Abu-Odah, Shuhan Li, Ting Mao, Zengjie Ye, Janelle Yorke. . . . 229

Association Between Short-Form Video Use and Mental Health: Systematic Review and Meta-Analysis (e82503)
Di Tang, Xin Zhang, Pengpeng Gou, Jie Feng, Rui Hu, Kim-wai Sum. . . . 245

Telehealth Approaches for Pediatric Otitis Media and Clinical Outcomes: Scoping Review (e85416)
Masao Noda, Ryohei Akiyoshi, Makoto Hosoya, Chikako Shinkawa, Ryota Koshu, Hidekane Yoshimura, Yukifumi Monden, Hiroaki Fushiki, Yasuhiro Tanaka, Makoto Ito. . . . 265

The Performance of Artificial Intelligence in Classifying Molecular Markers in Adult-Type Gliomas Using Histopathological Images: Systematic Review (e78377)
Obada Almaabreh, Rukaya Al-Dafi, Aliya Tabassum, Ahmad Othman, Alaa Abd-alrazaq. . . . 289

eHealth Literacy and Type 2 Diabetes Prevention Among At-Risk Populations: Mechanistic Systematic Review Using Theory-Driven Thematic Analysis (e77788)
Jingyi Li, Arina Azlan, Nurzihan Hassim, Yuan Wang, Ruina Guo. . . . 309

Artificial Intelligence Applications in Medical Devices for Personalized Health Care Solutions: Systematic Review (e72410)
Hemanth Ponnambalath Mohanadas, A Manikandan, Ahmad Ismail, Nick Tucker, Saravana Jaganathan. . . . 344

Effectiveness of Interventions for Internet, Smartphone, and Gaming Addictions: Umbrella Review and Meta–Meta-Analysis (e81705)
Minggang Zhang, Tong Lan, Tao Song, Xiaochun Wang. . . . 368

Effectiveness of Virtual Reality–Based Early Rehabilitation Strategies on Pain, Sleep, Anxiety, Balance, Cognition, and Limb Motor Function in Adult Intensive Care Unit Patients: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e81865)
Fei Wu, Yunting Wu, Yana Xing, Weixin Cai, Ran Zhang. . . . 382

Artificial Intelligence for Medicines Information: Scoping Review of Clinical Applications and Digital Health Inequalities (e77747)
Shahd Al-Arkee, Josephine Falade, Vibhu Paudyal. . . . 403

Machine Learning in Left Ventricular Hypertrophy Detection: Systematic Review and Meta-Analysis (e76637)
Yilin Li, Ke Zhao, Jing Wu. . . . 417

Accuracy of Medical Image–Based Deep Learning for Detecting Microvascular Invasion in Hepatocellular Carcinoma: Systematic Review and Meta-Analysis (e82000)
Wei Feng, Bo Qu, Shuo Han. . . . 435

The Predictive Value of Machine Learning for Postoperative Delirium in Cardiac Surgery: Systematic Review and Meta-Analysis (e72304)
Yi Guo, Hong Xu, Ankui Wang, Mingming Zhang, Shuai Zhang, Peng Xie. . . . 464

Digital Inequalities in the Use of eHealth Services in European Public Health Care Systems: Systematic Review of Observational Studies (e81841)
Garazi Monasterio, Marcos Fernández-López, Erika Valero, Unai Martin, Amaia Ayala-García. . . . 514

XSL•FO
RenderX

Data-Driven Guideline Adherence in Data Representation and Compliance Measurement: Scoping Review (e79937)
Minh Hoang, Candice Donnelly, Christina Igasto, Amith Shetty, Malcolm Pradhan, Tim Shaw. . . . 537

Effects of Digital Health Interventions to Promote Safer Sex Behaviors Among Youth: Systematic Review and Bayesian Network Meta-Analysis (e87071)
Yiran Zhu, Wenwen Peng, Die Hu, Edmond Choi, Maritta Välimäki, Ci Zhang, Xianhong Li. . . . 552

Characterization of Models for Identifying Physical and Cognitive Frailty in Older Adults With Diabetes: Systematic Review and Meta-Analysis (e84617)
Xia Wang, Shujie Meng, Xiang Xiao, Liu Lu, Hongyan Chen, Yong Li, Rong Zhang, Qiwu Jiang, Shan Liu, Ru Gao. . . . 585

Predictive Performance of Artificial Intelligence Algorithms for Gestational Diabetes Mellitus in Pregnant Women: Systematic Review and Meta-Analysis (e79729)
Yingni Liang, Anran Dai, Meiyan Luo, Zhuolian Zheng, Jiayu Shen, Yinhua Su, Zhongyu Li. . . . 606

Barriers to Digital Health Adoption in Older Adults: Scoping Review Informed by Innovation Resistance Theory (e75591)
Yosefa Birati, Roy Tzemah-Shahar. . . . 629

Machine Learning Techniques Used for the Identification of Sociodemographic Factors Associated With Cancer: Systematic Literature Review (e79187)
Liz González-Infante, Gaston Marquez, Solange Parra-Soto, Mónica Cardona-Valencia, Carla Taramasco. . . . 653

Patient Benefits in the Context of Sepsis-Related AI-Based Clinical Decision Support Systems: Scoping Review (e76772)
Pascal Raszke, Godwin Giebel, Jürgen Wasem, Michael Adamzik, Hartmuth Nowak, Lars Palmowski, Philipp Heinz, Nina Timmesfeld, Marianne Tokic, Frank Brunkhorst, Nikola Blase. . . . 684

Diagnostic Performance of Deep Learning and Radiomics in Extracranial Carotid Plaque Detection: Systematic Review and Meta-Analysis (e77092)
Lingjie Ju, Yongsheng Guo, Haiyong Guo, Ruijuan Liu, Yiyang Wang, Siyu Wang, Na Ma, Junhong Ren. . . . 699

Evidence for Digital Health Tools Designed to Support the Triage of Musculoskeletal Conditions in Primary, Urgent, and Emergency Care Settings: Scoping Review (e81578)
Linda Truong, James Wrightson, Raphaël Vincent, Eunice Lui, Jamon Couch, Ellen Wang, Cobie Starcevich, Dean Giustini, Alex Haagaard, Elena Lopatina, Niels van Berkel, Michael Rathleff, Clare Ardern. . . . 725

The Diagnostic Value of Image-Based Machine Learning for Osteoporosis: Systematic Review and Meta-Analysis (e75965)
Rui Zhao, Haolin Yang, Yangbo Li, Xiaoyun Li, Zhijie Yang, Yanping Lin, Jiachun Huang, Lei Wan, Hongxing Huang. . . . 746

The Application of Mobile Health in Self-Management Among Patients Undergoing Dialysis: Scoping Review (e76880)
Qin Xu, Yulin Xu, Xiaoqin Liu, Xiaolin Ma. . . . 769

Accuracy of Deep Learning in Diagnosing Chronic Obstructive Pulmonary Disease: Systematic Review and Meta-Analysis (e83459)
Hui Yang, Yijiu Wu, Tong Wu, Jingyan Ji, Sitao Lei, Weibin Xu. . . . 792

Assessment of the Diagnostic Performance and Clinical Impact of AI in Hepatic Steatosis: Systematic Review and Meta-Analysis (e78310)
Jiamei Song, Dan Liu, Jitong Li, Haoru Cong, Ruixue Deng, Yihan Lu, Jiayi Sun, Jingzhou Zhang. . . . 810

AI for Detecting and Predicting Postpartum Depression: Scoping Review (e77376)
Mais Alkhateeb, Ajisha Nayeem, Arfan Ahmed, Mohammed Alsahli, Javaid Sheikh, Alaa Abd-Alrazaq. . . . 837

XSL•FO
RenderX

Face-to-Face Versus Digital, Telephone-Delivered, and Self-Help Cognitive Behavioral Therapy for Irritable Bowel Syndrome: Systematic Review and Bayesian Indirect Treatment Comparison Meta-Analysis (e75833)
Qing-Feng Tao, Can Hua, Xiao Zhuo, Jian-Jiao Mou, Chao-Rong Xie, Yu-Xin Zhang, Bei Lv, Xin-Ying Niu, Min Chen, Hui Zheng. 864

Digital Interventions Targeting Healthy and Sustainable Eating Behavior: Systematic Review and Meta-Analysis (e80821)
Käbi Vanwinkelen, Bram Spruyt, Tim Smits. 880

Diet-Related Health Recommender Systems for Patients With Chronic Health Conditions: Scoping Review (e77726)
Xiaolan Dong, Bei Yun, Anni Pakarinen, Zhuting Zheng, Hao Niu, Tian Jin, Changrong Yuan, Jingting Wang. 905

AI-Supported Digital Microscopy Diagnostics in Primary Health Care Laboratories: Scoping Review (e78500)
Joar von Bahr, Antti Suutala, Vinod Diwan, Andreas Mårtensson, Johan Lundin, Nina Linder. 925

Effectiveness of Machine Learning in Detecting Vessels Encapsulating Tumor Clusters in Hepatocellular Carcinoma: Systematic Review and Meta-Analysis (e82839)
Huili Shui, Wenyu Wu, Zhenming Xie, Bing Yang, Jia Deng, Dongxin Tang. 945

Machine Learning Prediction Models for Preeclampsia: Systematic Review and Meta-Analysis (e78714)
Lu Liu, Qixuan Zhu, Yichi Zong, Xueyuan Chen, Wei Zhang, Jun Wang. 957

Communication Strategies to Promote Patient Engagement in Telemedicine: Systematic Review (e85456)
Yangna Hu, Cindy Ngai, Rui Jiang. 983

Products, Performance, and Technological Development of Ambulatory Oxygen Therapy Devices: Scoping Review (e81077)
Shohei Kawachi, Mariana Hoffman, Lorena Romero, Magnus Ekström, Jerry Krishnan, Anne Holland. 1005

The Development and Use of AI Chatbots for Health Behavior Change: Scoping Review (e79677)
Lingyi Fu, Ryan Burns, Yuhuan Xie, Jincheng Shen, Shandian Zhe, Paul Estabrooks, Yang Bai. 1023

Efficacy of Brain-Computer Interface Therapy for Upper Limb Rehabilitation in Chronic Stroke: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e79132)
HongJie Chen, GuoJun Yun. 1043

Behavioral Determinants and Effectiveness of Digital Behavior Change Interventions for the Prevention of Sexually Transmitted Infections and HIV: Overview of Systematic Reviews (e74201)
Giuliano Duarte-Anselmi, Susana Sanduvete-Chaves, Salvador Chacón-Moscoso, Daniel López-Arenas. 1066

Effect of Digital Health Interventions on College Students' Lifestyle Behaviors: Systematic Review (e82192)
Qingyuan Zhou, Jiajun Jiang, Zhihua Yin, Ruishi Fan. 1089

Key Components and Barriers in Web-Based Suicide Prevention Gatekeeper Training: Systematic Narrative Review (e81572)
Olivier Ferlatte, Emmanuelle Gareau, Keven Lee, Kinda Wassef, John Oliffe, Hannah Kia, Brock Dumville. 1114

Interventions That Use Highly Visual Social Media Platforms to Tackle Unhealthy Body Image in Adolescents and Young Adults: Systematic Review of Randomized Controlled Trials and Quasi-Experimental Studies (e80141)
Anthony Copez-Lonzoy, Janina Bazalar-Palacios, Nikol Mayo-Puchoc, Gustavo Calderon de la Cruz, Liliana Cruz-Ausejo, Evelyn Goicochea-Ríos, Juan Bazo-Alvarez. 1134

XSL•FO
RenderX

Text-Based Depression Estimation Using Machine Learning With Standard Labels: Systematic Review and Meta-Analysis (e82686)
Shengming Zhang, Chaohai Zhang, Jiaxin Zhang. . . . 1149

AI and Wearables for Early Detection of Cognitive Impairment and Dementia: Systematic Review (e86262)
Ander Cejudo, Markel Arrojo, Cristina Martín, Aitor Almeida. . . . 1169

Design Practices for Data Dashboards in Health Care: Scoping Review (e77361)
Heike Vornhagen, Stephen Barrett, Ciara Carroll, Lydia Iladiva, Gregory Martin, Declan McKeown, Jennifer Martin. . . . 1195

The Effects of Digital Health Interventions for Family Members in Intensive Care Units: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e83294)
Chuchu Zhang, Weijing Sui, Weilin Jiang, Yibing Shen, Zhenzhen Huang, Ran Yan, Jia Yi, Jiayu Zhang, Chaoyi Zhang, Yiyu Zhuang, Xiaoyan Gong. . . . 1205

AI-Enabled Wearables for Motor Function Assessment and Rehabilitation in Parkinson Disease: Scoping Review (e85596)
Shengting Li, Siqi Chen, Xiaosong Yu, Huixiang Shang, Tingting Tu, Mingtao Quan. . . . 1232

Effectiveness of Telerehabilitation Interventions for Self-Management of Tinnitus: Update of a Systematic Review (e83529)
Sara Demoen, Elise Van Kerchove, Annick Timmermans, Vincent Van Rompaey, Sarah Michiels, Annick Gilles. . . . 1256

Digital Health Interventions to Promote Physical Activity Among Adolescents: Systematic Review (e82395)
Rui Fan, Jia Jiang, Qing Zhou, Xin Zhang, Zhou Wu, Liu Ji. . . . 1293

Digital Technologies That Support Meaningful Connections in Care Homes: Scoping Review (e88181)
Deborah Muldrew, Rosemary Bradley, Kelly Conway. . . . 1320

Association Between Digital Biomarkers of Health and Anxiety: Systematic Review and Meta-Analysis (e73812)
Yolanda Lau, Natalia Chemas, Heema Ajeet Gokani, Rachel Morrell, Harisd Phannarus, Claudia Cooper, Zuzana Walker, Harriet Demnitz-King, Natalie Marchant. . . . 1344

Effects of Digital Health Interventions on Functional and Psychological Outcomes in Older Patients With Hip Fractures: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e79563)
Wei Fan, Qi Zhang, Qunfeng Lu. . . . 1372

Use of Health and Welfare Technology in Palliative Care: State-of-the-Art Review (e79637)
Viktoria Zander, Maja Holm, Monir Mazaheri, Christine Gustafsson, Sara Landerdahl Stridsberg, Ragnhild Hedman. . . . 1396

The Effects of Digital Health Interventions on Motor Symptoms, Nonmotor Symptoms, and Quality of Life in Patients With Parkinson Disease: Systematic Review and Meta-Analysis of Randomized Controlled Trials (e79935)
Ruwen Liu, Sirui Zhang, Yi Xiao, Yangfan Cheng, Huifang Shang. . . . 1414

Effects of Wearable Devices on Parkinson Disease: Systematic Review and Meta-Analysis of Randomized Controlled Trials Within the International Classification of Functioning, Disability, and Health Framework (e85914)
Jiarong Wu, Wanli Zang, Mingqing Fang, Ningkun Xiao, Xianzuo Zhang, Su Wang, Qiuxia Zhang. . . . 1442

Determinants and Health Outcomes of Digital Health Literacy in Patients With Cardiovascular Disease: Systematic Review and Meta-Analysis (e89102)
Eun-Jung Kim, Da-Young Kim, Ha-Jin Kim, Youn-Jung Son. . . . 1466

XSL•FO
RenderX

Robot-Assisted Therapy for Upper Limb Rehabilitation After Stroke: Umbrella Review (e79363)
Sijia Liu, Xianghong Zhang, Lijun Zhou, Jiashan Zhang, Jialin Liu, Chengqi He. . . . 1488

Technology-Based Interventions for Prevention of Type 2 Diabetes Following Gestational Diabetes: Systematic Review and Meta-Analysis (e78841)
Claire Eades, Anh Nguyen-Hoang, Louise Hoyle, Dawn Cameron, Josie Evans. . . . 1510

Predictive Value of Machine Learning for Poststroke Mortality Risk: Systematic Review and Meta-Analysis (e83821)
Yujie Chen, Zhujing Ou, Yiting Deng, Anling Luo, Xuzi Li, Yujia Yang, Qin Yang, Xintong Wu, Muke Zhou. . . . 1534

Machine Learning for Predicting Stroke Risk Stratification Using Multiomics Data: Systematic Review (e85654)
Hae Yoo, Hyerim Shin, Eun-Jung Kim, Youn-Jung Son. . . . 3778

Immersive Technologies for Cognitive Rehabilitation in Dementia and Mild Cognitive Impairment: Systematic Review (e84349)
Jaromir Konecny, Giuseppe Lanza, Serafino Buono, Raffaele Ferri, Antonina Luca, Radek Martinek, Sabrina Musso, Aurora Palmigiano, Barbora Prauzkova, Angelica Quercia, Francesco Rundo, Ramachandran Subramanian, Alessandro Serretti, Michal Prauzek. . . . 4323

Artificial Intelligence for Predicting Treatment Response in Patients With Anxiety Disorders After Cognitive Behavioral Therapy: Systematic Review and Meta-Analysis (e86079)
Jiawen Liu, Junhui Wang, Zhaobin Wu, Mohamad Adam Assim. . . . 4527

Original Papers

Securing Federated Learning With Blockchain in the Medical Field: Systematic Literature Review (e79052)
Xudong Wang, Yi Xie, Xiaoliang Chen, Jiaming Yang, Ruiyuan Li, Weihang Gao, Zineng Yan, Hong Zhou, Zhewei Ye. . . . 479

The Phases of Living Evidence Synthesis Using AI: Living Evidence Synthesis (Version 1) (e76130)
Xuping Song, Zhenjie Lian, Rui Wang, Ruixin Li, Zhenzhen Yang, Xufei Luo, Lei Feng, Zhiming Ma, Zhen Pu, Qi Wang, Long Ge, Caihong Li, Yaolong Chen, Kehu Yang, John Lavis. . . . 668

Artificial Intelligence, Connected Care, and Enabling Digital Health Technologies in Rare Diseases With a Focus on Lysosomal Storage Disorders: Scoping Review (e73612)
Alberta Spreafico, Luca Neri, Kim Angel, Cinzia Bellettato, Maurizio Scarpa. . . . 1562

Association Between Telemedicine Adoption and Physician Job Satisfaction: Cross-Sectional Study (e82285)
Xiaohui Zhai, Zhongliang Zhou, Zhichao Wang, Guanping Liu, Peter Coyte. . . . 1593

Social Media Intervention Based on the Information-Motivation-Behavioral Skills Model Promotes HIV Testing and Reduces High-Risk Behaviors Among Men Who Have Sex With Men in Resource-Limited Settings in China: Randomized Controlled Trial (e84279)
Huanzhuo Mai, Jianyuan Liu, Liping Cheng, Hao Liang, Ping Cui, Qin Yao, Qin Cao, Heng Zhang, Wenping Liao, Shiran He, Xing Yang, Ping Cen, Jiegang Huang. . . . 1840

Determinants of the Uptake and Frequency of Use of a Web Portal Digital Health Intervention in Patients With Type 2 Diabetes and/or Coronary Heart Disease: Secondary Analysis of a Randomized Controlled Trial (e80895)
Maximilian Scholl, Claas Lendt, Sebastian Appelbaum, Bianca Biallas, Katja Brenk-Franz, Chloé Chermette, Friederike Frank, Angeli Gawlik, Lisa Giesen, Martina Heßbrügge, Lucas Küppers, Larisa Pilic, Marcus Redaèlli, Lara Schneider, Frank Vitinius, Stefan Wilm, Uwe Konerding. . . .
1 8 6 0

XSL•FO
RenderX

Internet-Based Problem Management Plus Intervention for Antenatal Depression: Randomized Controlled Trial (e81998)
Shuanghui Zheng, Shiyu Wang, Xiaoyu Shui, Jing Zhao, Ying Lin, Qian Huang, Siqi Li, Yang Chen, Zhijie Zou, Jin Zheng, Xiaoli Chen. 1880

Digital Intervention (MiVacunaLA 2.0) to Promote COVID-19 Vaccine Acceptance Among Hispanic Children: Community-Based Randomized Controlled Trial (e78103)
Luisa Blanco, Alexandra Klomhaus, Rebecca Dudovitz, Denise Marquez, Ray Lopez-Chang, Evelyn Aleman, Bas Weerman, Michael Moldoff, Keith Norris, Yelba Castellon-Lopez. 1898

Designing a Carbohydrate Counting App for Young Adults With Type 1 Diabetes: Usability Testing Interview Study (e86024)
Asmaa Housni, Aidan Shulkin, Alexandra Katz, Giuliana Giannini, Amélie Roy-Fleming, Meranda Nakhla, Courtney South, Anne-Sophie Brazeau. 1914

Effectiveness of Postdischarge Telephone Calls in Reducing Hospital Utilization: Quasi-Randomized Controlled Trial (e80529)
Sandie Du, Abigail Anada, Maria Montenegro, Maurie Maitland, Sean Chuan, Sonia Jose, Harveer Sihota, Delia Cooper, Joanna Rivera, Dianne Burditt, Minetaro Naruki-van Velzen, Megan MacPherson. 1928

Effects of Self-Compassion and Mindfulness Interventions on Mental Health and Work-Related Outcomes Among Japanese Workers: Randomized Controlled Trial (e79991)
Takumu Kurosawa, Koichiro Adachi, Ryu Takizawa. 1942

Effects of Internet-Based Dementia Risk Reduction Education on Risk and Protective Factor Knowledge, Intentions, and Health Behaviors: Randomized Controlled Trial (e79405)
Anthony Levinson, Stephanie Ayers, Sandra Clark, Rebekah Woodburn, Amy Schneeberg, Dima Hadid, Nick Kates, Gillian Strudwick, Roland Grad, Alexandra Papaioannou, Maureen Dobbins, Henry Siu, Dante Duarte, Karen Saperson, Sharon Marr, Doug Oliver, Sarah Neil-Sztramko.
1 9 6 2

Enhancing LGBTQ+ Inclusivity in an AI-Powered Sexual Health Chatbot: User-Centered Design Approach Through a Nonprofit and Academic Partnership (e78621)
William Liem, Elizabeth Casline, Julianna Lorenzo, Jacob Gordon, Andrés Avila, Attia Taylor, Nicole Levitz, Michael O'Keefe, Kathryn Macapagal. 1982

The Impact of a Health Coaching App on the Subjective Well-Being of Individuals With Multimorbidity: Mixed Methods Study (e78738)
Isabelle Symes, Alexandra Burton, Daniela Mercado, Feifei Bu. 2003

Problems and Barriers Regarding the Admission, Financing, and Service Provision of Digital Health Apps: Qualitative Stakeholder Survey (e73332)
Felix Plescher, Godwin Giebel, Carina Abels, Silke Neusser, Bonnie Kampka, Kirstin Börchers, Jürgen Wasem, Nikola Blase. 2021

Effects of Using a Smartphone App Combined With Behavior Change Techniques on the Level of Physical Activity Among Adults and Older Adults: Sequential Multiple Assignment Randomized Trial (e73388)
Maria Simoes, Neli Proença, Vinícius Lauria, Matheus do Nascimento, Ricardo Padovani, Victor Dourado. 2035

Effect of Lung Cancer Screening, Smoking Cessation, and Cessation Smartphone App to Health-Related Quality of Life Among Heavy Smokers: Randomized Controlled Trial (e81687)
Antti Kurtti, Sanna Iivanainen, Riitta Kaarteenaho, Heidi Andersen, Antti Jekunen, Tuula Vasankari, Jussi Koivunen. 2055

Patient and Care Team Perspectives of Barriers to and Facilitators for the Implementation of a Digital Health Program for Depression in Primary Care: Qualitative Study (e72003)
Andrea Nederveld, Elise Robertson, Angela Lanigan, Elisabeth Callen, Tarin Clay, Ben Fehnert, Lambros Chrones, Michael Martin, Margaret McCue, Christina Hester, Melissa Filippi. 2073

XSL•FO
RenderX

Internet Health Care Service Use Behavioral Pattern Among Older Adults and the Role of the Technology Acceptance and Social Ecological Theory Model: Cross-Sectional Survey (e78037)
Rui Li, Xinyu Xu, Qingsong Li, Haobiao Liu, Ting Zhou, Abebe Amhare, Peiyu Liu, Jing Tang, Wei Wang, Fuju Zheng, Jing Han. 2085

Therapeutic Effects of a WeChat Mini-Program on Metabolic Dysfunction–Associated Fatty Liver Disease: Randomized Controlled Trial (e76204)
Chao Sun, Guangyu Chen, Cuicui Shi, Haixia Cao, Ruixu Yang, Jing Zeng, Xiaoyan Duan, Xin Sun, Jian-Gao Fan. 2101

Cognitive-Behavioral Therapy−Based Massed Brief Psychoeducational Group via Videoconference for Social Anxiety: Randomized Controlled Trial (e79825)
Lele Feng, Wei Liu, Liechuan Cui, Deborah Dobson, Xinfeng Tang. 2122

Smartphone-Based Digital Eczema Education Program for Atopic Dermatitis in Children Aged 0 to 6 Years: Multicenter, Randomized, Parallel Controlled Clinical Study (e79559)
Huan Yang, Hong Shu, Liu-hui Wang, Ping Li, Yun-ling Li, Qin-feng Li, Xiu-ping Han, Jing Tian, Jing Chang, Hua Qian, Jing-ping Chen, Xin-qiang Ding, Pan-qian Wu, Li-min Dou, Zhen Luo, Wei Li, Yang-yang Lin, Lin Li, Shu-zhen Yue, Yang Gu, Li Yang, Xiao-hong Sun, Xiao-yan Luo, Lin Ma, Hua Wang. 2138

Bayesian-Based Pharmacokinetic Framework Integrated with Therapeutic Drug Monitoring for Assessing Adherence to Antiseizure Medications: A Clinical Trial Simulation Study (e77917)
Xiao-Qin Liu, Zi-Ran Li, Wei-Wei Lin, Juan Wang, Fu-Qing Gu, Jun-Jie Ding, Zheng Jiao. 2151

A Real-Life Digital Intervention for Personalized Nutrition in Adults With Overweight or Obesity: Remote Randomized Controlled Trial (e73367)
Jelle de Jong, Femke Hoevenaars, Lotte Peters, Charlotte Berendsen, Wilrike Pasman, Martien Caspers, Remon Dulos, Suzan Wopereis. . .
2 1 6 5

A Web-Based Cancer Prevention Intervention for Rural Emerging Adults: Mixed Methods Development and Pilot-Testing Study (e80803)
Echo Warner, Alishia Kinsey, Barbara Walkosz, Julia Berteletti, Kayla Nuss, Annelise Small, W Woodall, Deanna Kepka, Douglas Taren, Meghan Skiba, Dolores Guest, Cindy Blair, Judith Gordon, David Wetter, Evelinn Borrayo, Kimberly Henry, Andrew Sussman, David Buller. 2182

Effects of Artificial Intelligence Recognition–Based Telerehabilitation on Exercise Capacity in Patients With Hypertension: Randomized Controlled Trial (e81400)
Qiuru Yao, Baizhi Qiu, Longlong He, Qin Wang, Jihua Zou, Donghui Liang, Shuyang Wen, Yingchao Liu, Gege Li, Jinjing Hu, Huan Ma, Guozhi Huang, Qing Zeng. 2201

Digital Engagement Significantly Enhances Weight Loss Outcomes in Adults With Obesity Treated With Tirzepatide: Retrospective Cohort Study of a Digital Weight Loss Service (e83718)
Hans Johnson, Ashley Clift, Daniel Reisel, David Huang. 2219

The Relationship Between Physician Self-Disclosure and Patient Acquisition in Digital Health Markets: Cross-Sectional Study (e84963)
Quanchen Liu, Pengqing Yin, Jing Fan. 2235

Website Use and Associations With Behavior Change and Weight Loss in Cancer Survivors and Their Partners: Secondary Analysis of a Randomized Controlled Trial (e86908)
Harleen Kaur, Dori Pekmezi, Tracy E Crane, David Farrell, Laura Q Rogers, Wendy Demark-Wahnefried. 2251

Traditional Rehabilitation Experiences, Unmet Needs, and Perspectives on Virtual Reality–Based Rehabilitation Among Patients With Stroke in China: Qualitative Thematic Analysis and Semistructured Interview Study (e84532)
Xite Zheng, Lu Xing, Haitao Lu, Shimeng Hao, Fen Liu. 2268

XSL•FO
RenderX

Effectiveness of the Mobile-Based Diabetes Little Helper Video Intervention on Medication Adherence Among Older Adults Living With Type 2 Diabetes Mellitus, Henan, China: Randomized Controlled Trial (e78731)
Meng Wang, Khuan Lee, Hui Thew, Siti Sowtali, Qiuhuan Jiang, Zhengyan Zhao, Yu Liu, Poh Lim. 2284

Added Technology for Weight Management in Cardiac Rehabilitation: Biopsychosocial Factors and Rehabilitees' Talk in a Mixed Method Study (e78347)
Heli Lahtio, Teemu Paajanen, Hilkka Korpi, Ari Heinonen, Tuulikki Sjögren. 2302

Effectiveness of the Components of a Digital Multiple Health Behavior Intervention Among University Students (Buddy): Factorial Randomized Trial (e88884)
Katarina Åsberg, Oskar Lundgren, Hanna Henriksson, Pontus Henriksson, Ann Eldh, Preben Bendtsen, Marie Löf, Marcus Bendtsen. 2319

Telepathology and Mobile Health System for Province-Wide Pathology Consultation in Henan, China: Retrospective Evaluation Study (e75172)
Jinming Shi, Ming Ye, Dongxu Sun, Xianying He, Yaoen Lu, Linlin Wang, Haotian Chen, Wenchao Wang, Jie Zhao, Fangfang Cui. 2342

Creation of EmpowerMe Website to Promote Self-Efficacy in Survivors of Stroke: Co-Design Study (e76756)
Elizabeth Lynch, Adrian O'Malley, Zoe Adey-Wakeling, Billie Bonevski, Dominique Cadilhac, Saran Chamberlain, Leonid Churilov, Robyn Clark, Richard Cullen, Erin Godecke, Gillian Harvey, Fiona Jones, Natasha Lannin, Stacy Larcombe, Jarrad Law, Annette McGrath, Lisa Murphy, Katie Nesbitt, Karly Zacharia, Niranjan Bidargaddi, Coralie English. 2359

Insights and Recommendations From Moderators and Community Members for Keeping Online Peer Support Safe: Thematic Analysis (e81943)
Hannah Jones, Grace Lavelle, Elly Aylwin-Foster, Ciara Regan, Alan Simpson, Ewan Carr, Matthew Hotopf, Vanessa Lawrence. 2374

Effect of a Digital-Driven Physician-Pharmacist Collaborative Model for Diabetes in Primary Health Care: Cluster Randomized Trial (e77470)
Jie Xiao, Qing Wang, Shenglan Tan, Lei Chen, Daxiong Xiang, Haiyan Yuan, Xia Li, Shuting Huang, Bingjie Tang, Yan Guo, Haiying Huang, Danhui Zhao, Yue Li, Li Wang, Qun Li, Juan Liu, Ping Xu. 2394

Coproducing an Online Platform for People With Long-Term Physical Health Conditions: Development and Usability Study (e79666)
Hannah Jones, Grace Lavelle, Elly Aylwin-Foster, Ewan Carr, Alan Simpson, Matthew Hotopf, Vanessa Lawrence. 2410

Effectiveness of the Components of a Digital Multiple Health Behavior Change Intervention Among Individuals Seeking Help Online (Coach): Factorial Randomized Trial (e88881)
Joel Crawford, Katarina Åsberg, Jenny Blomqvist, Oskar Lundgren, Hanna Henriksson, Pontus Henriksson, Preben Bendtsen, Marie Löf, Marcus Bendtsen. 2428

Preferences for Personalized Text Message Appointment Reminders Among Outpatients in a Universal Health System: Cross-Sectional Study (e81324)
Shih-Huai Hsiao, Yu-Chen Chuang, Shu-Ling Ma, Hui-Tzu Lin, I-Chen Lee, Pei-Shih Chen. 2447

A Gamified Mobile Health Intervention to Promote Physical Activity, Executive Function, and Mental Health in College Students: Randomized Controlled Trial (e82769)
Ming Yang, Yingjun Guo, Zhen Li, Lei Jiang, Yanan Gao, Sen Yang, Zhixiong Zhou. 2461

Digital Peer Support to Increase Walking Among Older Adults: Cluster Randomized Trial (e75708)
Atsushi Nakagomi, Noriyuki Abe, Takayuki Ueno, Gemmei Izuka, Yohei Kawasaki, Katsunori Kondo. 2482

Enhancing Adherence to Home-Based Expiratory Muscle Strength Training in Parkinson Disease: Randomized Controlled Trial of an mHealth Intervention (e78022)
Martin Srp, Martina Hoskovcova, Rebeka Lagnerova, Katerina Dvorakova, Radim Kliment, Jan Muzik, Radim Krupicka, Ota Gal, Evzen Ruzicka.
2 4 9 5

XSL•FO
RenderX

Comparing the Associations of Internet Addiction and Internet Gaming Disorder With Psychopathological Symptoms: Cross-Sectional Study of Three Independent Adolescent Samples (e82414)
Ying-ying Li, A-qian Hu, Ling-li Yi, Zi-xin Mao, Qiu-yue Lü, Juan Wang, Wei Wei, Yue-qi Huang, Shu Huang, Wen-jing Dai, Meng-xuan Qiao, Jia-jun Xu, Qiang Wang, Xiao-jing Li, Fu-gang Luo, Wei Deng, Yu-zheng Hu, Tao Li, Wan-jun Guo. 2511

Digital Phenotyping for Adolescent Mental Health: Feasibility Study Using Machine Learning to Predict Mental Health Risk From Active and Passive Smartphone Data (e72501)
Balasundaram Kadirvelu, Teresa Bellido Bel, Aglaia Freccero, Martina Di Simplico, Dasha Nicholls, A Faisal. 2526

The Effectiveness of the Headspace App for Improving Sleep: Randomized Controlled Trial (e56287)
Zoltan Torok, Larisa Gavrilova, Amish Patel, Matthew Zawadzki. 2553

Associations Between Smartphone-Based Finger Tapping and Cognitive Performance in Older Adults: Observational Study (e82463)
Huitong Ding, Taylor Orwig, Jian Rong, Heaven Tatere, Chunyu Liu, Eric Schramm, Chathurangi Pathiravasan, Joanne Murabito, Honghuang Lin. 2567

Impact of Push Notifications on Physical Activity and Sodium Intake Among Patients with Hypertension: Microrandomized Trial of a Just-in-Time Adaptive Intervention (e78218)
Jessica Golbus, Michael Dorsch, Yuxuan Chen, Tanima Basu, Evan Luff, Predrag Klasnja, Mark Newman, Lesli Skolarus, Walter Dempsey, Brahmajee Nallamothu. 2579

Effects of an eHealth Cardiac Exercise Rehabilitation Platform for Patients After Percutaneous Coronary Intervention Based on the Persuasive Systems Design Model: Randomized Controlled Trial (e71450)
Yang Liu, Xiting Huang, Ziying Dai, Zhili Jiang, Wenxiao Wu, Jing Wang, Zhiqian Wang, Luyao Yu, Hanyu Li, Lihua Huang. 2590

Feasibility, Acceptability, and Perspectives Regarding the Use of Activity Tracking Wearable Devices Among Home Health Aides: Mixed Methods Study (e77510)
Ian Solano-Kamaiko, Michael Dicpinigaitis, Melissa Tan, Irene Yang, Kexin Cheng, Ronica Peramsetty, Michelle Shum, Yanira Escamilla, Jennifer Bayly, Meghan Turchioe, Ariel Avgar, Aditya Vashistha, Nicola Dell, Madeline Sterling. 2612

Characteristics Influencing Support for the National Health Service COVID-19 App in England and Wales: Findings From a Longitudinal Survey (e76863)
Josephine Exley, Paul Boadu, Kasim Allel, Bob Erens, Nicholas Mays, Mustafa Al-Haboubi. 2623

Patterns and Characteristics of Mobile App Use to Promote Wellness and Manage Illness: Cross-Sectional Study (e71363)
Hayriye Gulec, David Smahel, Yi Huang. 2639

Effects of an 8-Week App-Based Mindfulness Intervention on Mental Health in Working Women: Randomized Controlled Trial (e62814)
Riko Uwagawa, Koichiro Adachi, Mariko Shimoda, Ryu Takizawa. 2661

Changing Habits With the Happy Hands App: Qualitative Focus Group Study of a Hand Osteoarthritis Self-Management Intervention (e82773)
Kristine Fjeldstad, Anne Tveter, Eivor Rasmussen, Lena Olden, Sissel Nyheim, Thalita Blanck, Rikke Killingmo, Ingvild Kjeken. 2676

Facilitators and Barriers to Implementing Mobile Mental Health Interventions: Qualitative Study of the Consolidated Framework for Implementation Research in Pediatric Oncology Providers (e87533)
Shannon Hong, Michaela Patton, Krysta Barton, Tonya Palermo, Karen Mulholland, Eric Chow, Nancy Lau. 2688

XSL•FO
RenderX

Mobile Health App Attitudes and Adoption Among Oncology Providers: Cross-Sectional National Survey (e85583)
Nancy Lau, Kavin Srinakarin, Shannon Hong, Homer Aalfs, Michael Roth, Karly Ingram, Amy Berkman, Sherilynn Chan, Joanna Patten, Wade Iwata, Mallory Taylor, Jesse Fann, Eric Chow, Tonya Palermo. 2702

Patient Sharing of Digital Health Data in the Veterans Health Administration: Cross-Sectional Analysis (e80517)
Mark Zocchi, Jessica Lipschitz, Felicia Bixler, Stephanie Robinson, Bella Etingen, Timothy Hogan, Ndindam Ndiwane, Saige Calkins, Terry Newton, Nilesh Shah, Ben Kragen, Alexander Young, Stephanie Shimada. 2720

Leveraging Social Media to Achieve Population-Level Reach of Lung Cancer Screening-Eligible Individuals: A RE-AIM Framework Perspective (e80281)
Lisa Carter-Bawa, Jamie Ostroff, Susan Rawl, Erin Hirsch, Smita Banerjee, Andrew Ciupek, Robert Comer, Minal Kale, Katherine Leopold, Patrick Monahan, James Slaven Jr, Francis Valenzona, Renda Wiener, Ana Vielma. 2732

Changes in Workplace Productivity and Estimated Cost Savings During Internet-Based Cognitive Behavioral Therapy in the Irish National Health Service: Naturalistic, Repeated-Measures, Retrospective Survey Study (e80689)
Chi Lee, Siobhán Harty, Angel Enrique, Alba Jiménez-Díaz, Garrett Hisler, Daniel Duffy, Derek Richards. 2745

Optimization of University Counseling Consent Forms With Large Language Models: Multidimensional Comparative Evaluation (e86502)
Jianchen Luo, Jing Ma, Danni Zhan, Yuhong Zhou, Jiayu Li, Lan Zhang, Wentao Wang. 2761

Well-Being and Cognitive Factors Influencing Health Care Workers' Adherence to Internet-Based Stress Management: Mixed Methods Analysis of a Nonrandomized Controlled Study (e73306)
Miqi Li, Quan Wang, Danli Zheng, Danni Feng, Xiaorong Lang. 2778

Goal Setting and Anchoring Effects on Meditation Using a Digital Platform: Large-Scale Digital Field Study (e85801)
Michael Bowen, Michael Beam. 2792

Bidirectionality Between Perceived Immediate and Long-Term Benefits and Losses and Internet Gaming Disorder Among Chinese Adolescent Gamers: Prospective Longitudinal Study (e74030)
Siman Li, Jianxin Zhang, Ji-bin Li, Joseph Lau, Yanqiu Yu. 2803

Interactions of Technology and Obsessive-Compulsive Disorder Symptomatology in Adults: Qualitative Interview Study (e85033)
Lucas Occhino-Moede, Kaitlyn Sulivan-Pascual, Kendall Phelan, Harrison Wang, Daniel Mokhtar, Elisa Liu, Erica Schug, Megan Mirkis, Thomas Baek, Tamerlane Visher, Ujjwal Pasupulety, Adam Frank. 2817

Assessing Usage and Usability of a Narrative-Based Psychoeducational Digital Intervention to Improve Medication Adherence Among Individuals With Schizophrenia in a Stable Phase: Mixed Methods Study (e59175)
Dian Zhu, Fangyuan Chang, Hongyi Yang, Yiwen Wei, Zhao Liu. 2830

Effectiveness of and Mechanisms of Change in a Self-Help Web- and App-Based Resilience Intervention on Perceived Stress in the General Working Population: Randomized Controlled Trial (e78335)
Sandy Hannibal, Dörte Behrendt, Michèle Wessa, Sarah Schäfer, Nina Dalkner, Dirk Lehr. 2845

Predictors of Professional Responses in Nonprofit Mental Health Forums: Interpretable Machine Learning Analysis (e74359)
Shuang Geng, Yanghui Li, Jie Wang, Peixuan Chen, Xusheng Wu, Zhiqun Zhang. 2869

Effect of AI-Based Natural Language Feedback on Engagement and Clinical Outcomes in Fully Self-Guided Internet-Based Cognitive Behavioral Therapy for Depression: 3-Arm Randomized Controlled Trial (e76902)
Mirai So, Yoichi Sekizawa, Sora Hashimoto, Masami Kashimura, Hajime Yamakage, Norio Watanabe. 2889

XSL•FO
RenderX

Leisure Screen Time, Internet Gaming Disorder, and Mental Health Among Chinese Adolescents: Large-Scale Cross-Sectional Study (e80737)
Qin Deng, Linna Sha, Jiaojiao Hou, Xunying Zhao, Rong Xiang, Jiangbo Zhu, Yang Qu, Jinyu Zhou, Ting Yu, Xin Song, Sirui Zheng, Tao Han, Bin Yang, Mengyu Fan, Xia Jiang. 2909

Longitudinal Between- and Within-Person Associations Among Screen Time, Bedtime, and Daytime Sleepiness Among Adolescents: Three-Wave Prospective Panel Study (e78972)
Michał Tkaczyk, Albert Ksinan, David Smahel. 2923

Tailored Internet-Delivered Mindfulness-Based Interventions for Patients With Hepatocellular Carcinoma After Transarterial Chemoembolization: Qualitative Study (e78337)
Zengxia Liu, Min Li, Yong Jia, Li Chen. 2939

Associations Among Cyberbullying Victimization, Inhibitory Control, Neural Activation of Error Processing, and Mental Health Problems in Adolescents: Neuroimaging, Retrospective Longitudinal Cohort Study Using the Adolescent Brain Cognitive Development Data (e75126)
Xuanyu Zhang, Chengyan Xie, Yu Chen, Boyu Qiu. 2953

Adherence, Acceptability, and Sexual Health Outcomes of the Odeya App–Based Intervention for Sexual Distress in Women With Endometriosis: Randomized Controlled Mixed Methods Trial (e86042)
Selina Kronthaler, Eden Kosman, Tatjana Tissen-Diabaté, Elena Mühle, Luzia Weber- Schallauer, Therese Pross, Maria Karsten, Klaus Beier, Laura Hatzler. 2970

Mental Health Professionals' Perceptions of Benefits and Disadvantages of Telehealth: International Mixed Methods Study (e75905)
Madeline Montoya, Cary Kogan, Brigitte Khoury, José García-Pacheco, Rebeca Robles, Tahilia Rebello, Maya Kulygina, Geoffrey Reed. . . . 2 9 9 5

Misinformation in Social Media Narratives on Highly Pathogenic Avian Influenza: Systematic Content Analysis of Facebook and Instagram Posts (e73275)
Ahmed Al-Rawi, Betty Ackah, Abdelrahman Fakida, Kelley Lee. 3015

Patient Perceptions of Ozempic (Semaglutide) for Weight Loss: Mixed Methods Analysis of Online Medication Reviews (e78391)
Abanoub Armanious, Rachel-Mae Hunter, Kristi Griffiths, Hannah Bowrey, Robyn Brown, Morgan James. 3030

Public Emotional and Thematic Responses to Major Emergencies on Social Media, 2024-2025: Cross-Sectional Convergent Mixed Methods Study (e84648)
Xingrong Guo, Yiqian Fan, Yiming Guo. 3045

Cocreating Principles for Digital Health Equity: Cross-Sectional, Qualitative Study for Participatory Human-Centered Design in Catalonia (e84129)
Jordi Piera-Jiménez, Núria Vilarasau Creus, Ada Maymó Costa, Xabier Michelena, Andrea Climent Fageda, Alèxia Farré, László Herczeg, Lekshmy Parameswaran, Gerard Carot-Sans, Luis Valle. 3073

Telemedicine Adoption for Managing Chronic and Rare Diseases in Indonesia During and Beyond the COVID-19 Era: Qualitative Study (e83462)
Christine Pratama, Rima Nurlianti, Cornelius Herstatt, Moritz Goeldner. 3096

Telehealth Intervention to Reduce Sedentary Behavior in Older Adults With Type 2 Diabetes: Development and Feasibility Study (e80827)
Xiaoyan Zhang, Dan Yang, Sihan Chen, Lei Wang, Xin Li, Ziyan Wang, Meiqi Meng, Xuejing Li, Hongzhan Jiang, Yufang Hao, Jiayin Luo. . . 3 1 1 2

XSL•FO
RenderX

Association Between Physician Communication Features and Patient Outcomes in Telemedicine: Retrospective Cross-Sectional Observational Study (e86977)
Yipei Wang, Shu Wang, Ke Zhang, Zhijie Liu, Qingbian Ma, Hong Ji, Zheng Hou, Tracy Liu, Xuedong Xu, Xinxia Wu, Changxiao Jin. 3127

Digital Primary Health in Rwanda: Qualitative Study of User Experiences and Implementation Lessons From Babyl's Telemedicine Platform (e84832)
Sabine Musange Furere, Gashaija Absolomon, Nathalie Umutoni, Stella Matutina Umuhoza, Felix K Rubuga, Yvonne Nsaba-Uwera, Patrick Singa Muhoza, Regis Hitimana, Jeanine Condo, James Humuza. 3145

Telehealth Use and Modality Choice Among US Adults: Shorrocks-Shapley Decomposition of a 2022 Cross-Sectional National Survey (e81879)
Corneliu Bolbocean, Corey Hayes, Cari Bogulski. 3162

Attitudes Toward Video Consultations From the Perspective of Physicians and Psychotherapists in German Outpatient Care After the COVID-19 Pandemic: Survey Study (e73757)
Lara Kleinschmidt, Juergen Wasem, Nikola Blase, Beatrice Nauendorf, Juliane Malsch, Matthias Brittner, Paul Brandenburg, André Aeustergerling, Theresa Hüer. 3177

Communication Challenges and Mitigation Strategies in Primary Care Virtual Consultations: Qualitative Study (e79399)
Ahmed Alboksmaty, Tetiana Lunova, Ara Darzi, Ana-Luisa Neves. 3193

Gender Concordance and Patient Outcomes in Indian Telemedicine: Retrospective Cross-Sectional Quantitative Study of 286,000 Consultations (e78311)
Nafisa Vaz, Vishalkumar Jani. 3207

Virtual Delivery of Early Psychosis Care: Retrospective Cohort Study of Factors Associated With Initial Engagement (e81313)
Trinity Vey, Nicole Kozloff, George Foussias, Wei Wang, Albert Wong, Aristotle Voineskos, Nicole Davis-Faroque, Alexia Polillo. 3220

Rising to the Challenge of Early Screening in Primary Health Care Through the Web Italian Network for Autism Spectrum Disorder (Win4ASD) in the Pediatric Population: Retrospective Observational Study (e74302)
Noemi Buo, Eleonora Rosi, Silvia Busti Ceccarelli, Mariarosa Ferrario, Valerio Maiorca, Erika Morandi, Nicole Viganó, Ivan Limosani, Laura Falcone, Massimo Molteni, Paola Colombo. 3231

The Effect of Veteran Race and Socioeconomic Status on Enrollment in Remote Patient Monitoring for Hypertension: Retrospective Observational Cross-Sectional Study (e78423)
Hannah Friedman, Hillary Mull, Jenice Guzman-Clark, Daniel Sturgeon, Marva Foster. 3243

Impact of Tele-Intensive Care Units on the Clinical Outcomes of Critically Ill Patients With COVID-19: Retrospective Cohort Study (e64996)
Thais Midega, Fabio Hohmann, Renato Chaves, Guacyra Almeida, Vivian Leoneza, Jennifer Cabral, Bianca Vitalino, Emanuelle Camboim, Nelma Machado, Ricardo Melo, Jorge Feliciano, Breno Cardoso, Leonardo Ferraz, Thiago Corrêa, Maura Santos, Renata Albaladejo Morbeck, Adriano Pereira. 3258

Supporting Access to Care Through Peripheral Devices and Patient-Generated Health Data: Qualitative Study (e88405)
Ashley Griffin, Caroline Gray, Victoria Ngo, Camila Chaudhary, Stephanie Shimada, Amanda Midboe, Donna Zulman. 3269

Telehealth Delivery of the Homeostasis–Enrichment–Plasticity Approach for Premature Infants With Developmental Risks: Exploratory Feasibility Study (e86883)
Aymen Balikci, Gamze Sirma, Izgi-Miray Demirbag, Teresa May-Benson, Hatice Sozen, Ayse Aracikul Balikci, Gul Ilbay. 3283

Willingness to Share Internet Use Data for Research on Early Disease Detection: Cross-Sectional Survey (e85637)
Christina Derksen, Mel Ramasawmy, Sanjula Arora, Wynsee Lau, Suzanne Scott. 3301

XSL•FO
RenderX

From Searching to Coping, How Chinese Patients With Breast Cancer Navigate Web-Based Health Information: Semistructured Interview Study (e80363)
Jiahui Liu, Xingfeng Li, Xuying Li, Jiejun Chen, Dan Zhang, Cuiling Qiu. 3322

Sociodemographic Drivers of Recruitment and Attrition in Digital Neurological Research: Longitudinal Cohort Study (e83432)
Peyman Nejat, Ashley Bachman, Vicki Stubbs, Joseph Duffy, John Stricker, Vitaly Herasevich, David Jones, Rene Utianski, Hugo Botha. . . . 3 3 3 8

Using Epidemiological Test Diagnostics to Select Fraud Detection Methods: Secondary Analysis of Quantitative Cross-Sectional Survey Data (e85161)
Rachel Willard-Grace, Tali Klima, Mansi Dedhia, Emily Lo, Annie Nisnevich, Allison Gray, Holly Henry. 3352

The Feasibility of Smartwatch Micro–Ecological Momentary Assessment for Tracking Eating Patterns of Malaysian Children and Adolescents in the South-East Asian Community Observatory Child Health Update 2020: Cross-Sectional Study (e73435)
Richard Lane, Louise Millard, Ruth Salway, Chris Stone, Andy Skinner, Sophia Brady, Jeevitha Mariapun, Sutha Rajakumar, Amutha Ramadas, Hussein Rizal, Laura Johnson, Tin Su, Miranda Armstrong. 3363

Integrating a Large Language Model to Streamline Nursing Handover Documentation Across Multiple Hospitals in Taiwan: Development and Implementation Study (e81604)
Ray-Jade Chen, Mai-Szu Wu, Lung-Wen Tsai, Shy-Shin Chang, Shu-Tai Shen Hsiao, Yu-Sheng Lo. 3377

Digital Health Technology Use Among Rehabilitation Professionals in China: Multi-Province Cross-Sectional Survey (e90623)
Shuning Duan, Zi-Ru Wang, Xier Chen, Guoxin Ni. 3399

Associations Between Short-Video Platform Use and Health Across Health Distribution and Usage Behaviors in China: Cross-Sectional Questionnaire Study (e86526)
Yangyang Pan, Kangkang Zhang, Yilin Wei, Yangzhen Huang, Chengxu Long, Chenxin Yang, Shangfeng Tang. 3417

User Profiles and Engagement in a Hypertension Self-Management App: Cross-Sectional Survey (e83075)
Felix Muehlensiepen, Susann May, Frances Seifert, Eileen Wengemuth, Olen Johannsen, Martin Middeke, Martin Heinze, Pascal Petit, Nicolas Vuillerme, Sebastian Spethmann, Dunja Bruch. 3433

Understanding User Intent in Code-Mixed Sexual and Reproductive Health Queries in Urban India: Hierarchical Classification Approach Using Large Language Models (e86545)
Sumon Dey, Manvi S, Aradhana Thapa, Meet Shah, Zeel Mehta, Shraddha Kapile, Tanvi Divate, Suhani Jalota, Azra Ismail. 3447

Longitudinal Effects of a Smartphone Game (Tumaini) for HIV Prevention Among Kenyan Adolescents: 45-Month Trajectories of Condom Use–Related Proximal Outcomes From a Randomized Controlled Trial (e83982)
Kate Winskell, Gaëlle Sabben, Haowen Qin, Calvin Mbeda, Sophie Goldenberg, Ken Ondeng'e, Richard Ndivo, Judith Arego, Xinwei He, Robert Bednarczyk, Kelli Komro, Robert Lyles, Victor Mudhune. 3469

The Effect of Text Messaging on the Postoperative Pain Experience in Pediatric Patients Undergoing Thoracic Surgery: Randomized Controlled Trial (e81806)
Yun Shi, Qianqiu Wang, Aihua Liu, Li Jiang. 3486

Designing Electronic Problem-Solving Training for Individuals With Traumatic Brain Injury: Mixed Methods, Community-Based, Participatory Research Case Study (e83995)
Matthew Schmidt, Yueqi Weng, Shannon Juengst, Alexandra Holland. 3498

XSL•FO
RenderX

Toward Patient-Centric Digital Monitoring of Obstructive Sleep Apnea: Mixed Methods Study (e82460)
James Timmis, Kerstin Schorr, Rana Yüksel, Tim van den Broek, Sebastiaan Overeem, Dagmar Smid, Willem van den Brink, Nina Haring. . . 3 5 2 3

Effectiveness of Educational Videos in Encouraging Preferences for Guideline-Based Cancer Screening in Japan: Three-Arm Pseudorandomized Controlled Trial (e82322)
Kosuke Sakai, Yuko Furuya, Shoko Nakazawa, Kota Fukai, Kei Sano, Masayuki Tatemichi. 3539

Seeking But Not Discussing Online Health Information With Physicians: Cross-Sectional Survey Study of eHealth Literacy–Empowerment Profiles and Patient-Centered Communication (e78836)
Qianfeng Lu, Wen Jiao, Angela Chang, Peter Schulz. 3553

Impact of GPT-4–Generated Discharge Letters on Patients' Medical Comprehension: Prospective Crossover Study (e81243)
Friederike Holderried, Alessandra Sonanini, Christian Stegemann–Philipps, Anne Herrmann–Werner, Philipp Spitzer, Martina Guthoff, Nils Heyne, Konstantin Sering, Martin Holderried, Felix Eisinger. 3571

Use, Utility, and User Experience of Cloud-Based Medical Imaging in Pulmonary Nodule Care in China: Mixed Methods Study (e86745)
Ziguo Chen, Junhan Wu, Yunying Chen, Weitao Zhuang, Jiaxuan Huang, Weifeng Zhong, Zijie Li, Shuhuan Xie, Chaofan Liu, Guojie Lu, Guibin Qiao. 3588

The WONE Index as a Multidimensional Assessment of Stress Resilience: A Development and Validation Study (e81714)
Lydia Roos, Destiny Gilliland, Kelsey Julian, Reeva Misra. 3601

Program Theory and Core Outcome Set Development for a Technology-Assisted Counseling Intervention in Dementia: Multimethods Study (e81669)
Dorothee Bauernschmidt, Anja Bieber, Ronja Hubrich, Janina Wittmann, Gabriele Meyer. 3631

Validation of the Simplified Chinese Clear Communication Index Using Diabetes Education Materials: Instrument Adaptation and Validation Study (e83935)
Lu Liu, Yi Chen, Xiaoqi Li, Hanyu Du, Rong Gui, Huiqing Jin, Fengfei Chen, Yujun Lin, Qiuhuan Huang, Yingge Tong. 3657

A Digital Tool for Assessing the Distinct Effects of Depression, Anxiety, and Attention-Deficit/Hyperactivity Disorder (ADHD) on Children's Emotional Cognitive Bias: Cross-Sectional Study (e86286)
Sang-Eon Park, Jisu Chung, Sang Lee. 3673

Factors Influencing Continuance Intention for Online Consultations Among Survivors of Cancer: Grounded Theory Study (e84644)
Yutang Yao, Musi Zhang, Shanshan Peng, Zhuzhong Cheng, Yun Duan. 3689

The Structure of Psychopathology on Reddit: Network Analysis of Mental Health Communities in Relation to the ICD Diagnostic System (e80958)
Bojan Evkoski, Srebrenka Letina, Petra Kralj Novak. 3702

Experiences of Alert Fatigue and Its Contributing Factors in Hospitals: Qualitative Study (e78676)
Nicki Newton, Adeola Bamgboje-Ayodele, Rowena Forsyth, Amina Tariq, Jing Huang, Ramyasri Yannam, Daniel Lalor, Andrew Sobey, Melissa Baysari. 3719

Improving the Understandability of Clinical Guidelines: Development and Evaluation of a GPT-4–Based Pipeline (e81915)
Matthew Jones, Melissa Torgbi, Harish Tayyar Madabushi. 3733

XSL•FO
RenderX

Interorganizational Mechanisms for Developing and Implementing Clinical Decision Support Systems in Primary Care: Exploratory, Qualitative Case Study (e83084)
Jesse Santema, Jeroen van Wijngaarden, Eric Hiddink, Fleur Deken, Maaike Kleinsmann, Hendrikus van Os. 3744

Assessing Health Care Professionals' Perceptions of a New System in Clinical Workflows: Systems Engineering Initiative for Patient Safety–Based Consensual Qualitative Research (e86166)
Ye-Eun Park, Minsu Ock, Jae-Ho Lee, Dae-Hyun Ko, Hak-Jae Lee, Taezoon Park, Junsang Yoo, Yura Lee. 3762

Advancing Nursing Data Integration Through a Nursing Minimum Dataset for the Conceptual and Technical Development of a "Fall Prevention" Data Module: Development Study (e82417)
Sarah Milkov, Antonia Schmidt, Anja Burmann, Niklas Tschorn, Marcel Klötgen, Wolfgang Deiters, Christian Potthoff, Kirsten Neveling, Yvonne Weber, Maren Keuchel, Daniela Holle. 3797

Challenges of Standard Pediatric Epilepsy Monitoring and the Potential Benefits of Contactless Sensor Technologies: Exploratory Qualitative Study (e83009)
Natalie Öhl, Dilbar Strack, Stefan Kusnik, Isabell Link, Marlene Hofmann, Tobias Steigleder, Alina Weisser, Christoph Ostgathe, Regina Trollmann, Maria Heckel. 3810

Development and User-Centered Evaluation of Smart Systems for Loneliness Monitoring in Older Adults: Mixed Methods Study (e81027)
Yi Zhou, Jessica Rees, Faith Matcham, Ashay Patel, Michela Antonelli, Anthea Tinker, Sebastien Ourselin, Wei Liu. 3824

Guide on Selection of Optimal Motivational Themes for Use in a Clinical Trial Recruiting Black US Adults: Survey Study (e75857)
Tamunotonye Harry, Jingyi Cao, Zaib Hussain, Ruth-Alma Turkson-Ocran, Stephen Juraschek, Timothy Lahey, Timothy Plante, Yuanyuan Feng. 3846

Ethical Handling of Occupational Health and Safety Data in the Fire Service: Empirical Interview and Focus Group Study of Firefighter and Fire Service Leadership Privacy Preferences (e84465)
Rachel Topazian, Aleksandra Wec, Joseph Ali, Shannon Frattaroli, Cassandra Crifasi. 3867

The Ethics of Leveraging Routinely Collected Patient Data for AI Development: Mixed Methods Study (e79863)
Menno Maris, Joanna Klopotowska, Ronald Cornet, Mariëtte Van den Hoven, Joris Lieverse, Daniel Fernández-Llaneza, Marieke Bak, LEAPfROG Consortium. 3879

Ethical Knowledge, Challenges, and Institutional Strategies Among Medical AI Developers and Researchers: Focus Group Study (e79613)
Sophia Fantus, Jinxu Li, Tianci Wang, Lu Tang. 3900

Educating Students About Digital Health Research Ethics: Curricula Review and Expert Interview Study (e82861)
Yier Zhu, Brian McInnis, Tanya Punater, Brittany York, Camille Nebeker. 3912

Factors Influencing Universal Coverage of AI-Assisted Cervical Cancer Screening: Qualitative Study Based on the Macro Model of Health System (e75372)
Lu Ji, Xinke Zhou, Lan Yao. 3927

Strategy for Hepatitis B and C Virus Testing Campaigns Through Web Services and Digital Advertising in Japan: Nationwide Cross-Sectional Study With Correspondence Analysis (e89585)
Kosuke Sakai, Shoko Nakazawa, Yuko Furuya, Kota Fukai, Kei Sano, Masaaki Korenaga, Masayuki Tatemichi. 3941

XSL•FO
RenderX

Opposition to Youth e-Cigarette Prevention Campaigns on Twitter and TikTok: Cross-Platform Observational Mixed Methods Analysis (e83791)
Chandler Carter, Simon Page, Mateusz Borowiecki, Ganna Kostygina, Sherry Emery, Miao Feng. . . . 3955

Developing and Validating a Machine Learning Algorithm to Predict the Risk of Incident Opioid Use Disorder Among OneFlorida+ Patients: Prognostic Modeling Study (e79482)
Jabed Faysal, Weihsuan Lo-Ciganic, Walid Gellad, Yonghui Wu, Christopher Harle, Khoa Nguyen, James Huang, Gerald Cochran, Debbie Wilson, Stephanie Staras, Siegfried Schmidt, Eric Rosenberg, Danielle Nelson, Shunhua Yan, Gary Reisfield, William Greene, Courtney Kuza, Md Hasan. . . . 3968

Breast Cancer Screening Knowledge and Sentiments in Singaporean Women: Mixed Methods Study Using Topic Modeling, Sentiment Analysis, and Structured Questionnaire Data (e78439)
Peh Ho, Zi Lim, Jenny Liu, Nur Mohamed Riza, Ying Chew, Yi Lim, Hui Tan, Su-Ann Goh, Han Oh, Chi Chin, Sing Kwek, Zhi Zhang, Desmond Ong, Swee Quek, Sujith Wijerathne, Philip Iau, Mikael Hartman, Jingmei Li. . . . 3983

Suicidal Thoughts and Behaviors Among Chinese Adolescents in Relation to Negative Life Events, Internet Addiction, and Sexual Abuse: Cross-Sectional Study (e85371)
Juanfang Zhu, Qinqin Jiang, Feng Zhang, Hao Liu, Shaojie Yu, Jinhai Sun, Zhe Zhao, Lijuan Liu, Lei Yuan. . . . 4003

Exploring Reddit Discourse and Information Needs Surrounding Extreme Heat: Topic, Sentiment, and Engagement Analysis (e82426)
Melissa MacKay, Soroush Zamani Alavijeh, Sydney Gosselin, Fattane Zarrinkalam, Jennifer McWhirter. . . . 4019

Detection of Antithrombotic-Related Bleeding in Older Inpatients: Multicenter Retrospective Study Using Structured and Unstructured Electronic Health Record Data (e77809)
Claire Coumau, Frederic Gaspar, Mehdi Zayene, Elliott Bertrand, Lorenzo Alberio, Christian Lovis, Patrick Beeler, Fabio Rinaldi, Monika Lutters, Marie-Annick Le Pogam, Chantal Csajka, SwissMADE Collaborators. . . . 4034

Behavioral Dynamics of AI Trust and Health Care Delays Among Adults: Integrated Cross-Sectional Survey and Agent-Based Modeling Study (e82170)
Xueyao Cai, Weidong Li, Wenjun Shi, Yuchen Cai, Jianda Zhou. . . . 4053

Screen Time and Chronic Pain Health: Mendelian Randomization Study (e78233)
Jiahui Jiang, Chunyan Pu, Jiarui Cai, Chuan Yu, Zhenmi Liu, Chenghan Xiao. . . . 4073

Genre-Specific Gaming Addiction and Flourishing in Adolescents: Cross-Sectional Survey Study (e89319)
Yan Luo, Xiaofei Chen, Yaoyu Liu, Wanjia Hong, Miner Fan, David Thomas, Sean Mullen. . . . 4090

Implementation of Reference Framework for Preventive Care for Older Adults: Cross-sectional Qualitative Study (e78814)
Claire Zhong, Mingtao Chen, Chung Lo, Man YIM, Xiaoshu Zhang, William Wong, Junjie Huang, Martin Wong. . . . 4110

Institutionalizing Digital Parenting Programs in Low Resource Settings in China: Comparative Case Study of Health Care and Education Sectors Using the RE-AIM Framework (e79848)
Xinyu Shi, Ruochen Ruan, Yi Qie, Jamie Lachman, Na Zhong, Zuyi Fang. . . . 4131

A Complex Digital Health Intervention to Support People With HIV: Organizational Readiness Survey Study and Preimplementation Planning for a Hybrid Effectiveness-Implementation Study (e76327)
Jacqueline Hodges, Wendy Cohn, Amanda Castel, Tabor Flickinger, Ava Waldman, Michelle Hilgart, Olivia Kirby, Sylvia Caldwell, Karen Ingersoll. . . . 4152

Health Communication Campaign Performance During the HEALing Communities Study: Cross-Sectional Examination of Digital Advertising Methods (e75458)
Nicky Lewis, Jennifer Reynolds, Diane Krause, Philip Reeves, Jamie Luster, Michael Stein, Sharon Walsh, Amy Farmer, Michelle Lofwall, Monica Roberts, Hilary Surratt, Brooke Crockett, Kara Stephens, Kelli Bursey, Kristin Mattson, Michael Slater. . . . 4171

XSL•FO
RenderX

Engagement With Meditation Apps: Cross-Sectional Survey of Use and Associations (e71960)
Julia Adams, Jonathan Davies, Prai Wattanatakulchat, Julieta Galante, Felicity Miller, Simon D'Alfonso, Nicholas Van Dam. 4187

Automated Inference of Social Anxiety From Behavior in Social Virtual Reality: Cross-Sectional Observational Study (e79147)
Gayoung Son, Marius Rubo. 4210

Temporal Annotation of German Clinical Language in Real and Synthetic Clinical Documents: Corpus Development and Baseline Tagger Validation Study (e71458)
Luise Modersohn, Udo Hahn. 4228

Association of Electronic Health Literacy With Self-Care and Health Outcomes Among Patients With Type 2 Diabetes Mellitus: Cross-Sectional Study (e77856)
Phoenix Mo, Alice Kong, Luyao Xie, Virginia Chan, Joseph Lau. 4254

The Relationship Between Electronic Health Literacy and Health-Related Quality of Life Among Chinese Older Adults: Cross-Sectional Study (e84700)
Yongqiang Wang, Baozhen Dai, Jiazhen Yao. 4267

Validation of the Updated Digital Health Literacy Instrument and Development of a Short Form: Online Survey Study of the General Population (e86879)
Rosalie van der Vaart, Inge Tuitert, Job van 't Veer, Constance Drossaert. 4288

Digital Engagement and Cognitive Function Among Older Adults in China: Cross-Sectional Questionnaire Study and Moderated Mediation Model Analysis (e83955)
Yongqi Du, Qing Niu, Gangrui Tan, Jianqian Chao, Shengxuan Jin, Leixia Wang. 4302

Investigating the Effect of Hospital Infection Control Informatization on Optimizing Microbiological Specimen Submission Before Antibiotic Therapy: Failure Mode and Effects Analysis (e78118)
Jianxiong Wu, Weijiang Zhan, Jun Hua, Qinling Ge, Yuexian Zhu, Huixian Yu, Min Zhao, Xiaoyan Zhan, Bingwei Zhu, Tongqi Xiang, Longxi Lu, Tieying Dai. 4362

Quality of Cancer-Related Clinical Coding in Primary Care in North Central London: Mixed Methods Quality Improvement Project (e73205)
Afsana Bhuiya, Graham Roberts, Katie Tucker, Stefanie Bonfield, Georgia Black. 4375

Multimodal Large Language Models for Cystoscopic Image Interpretation and Bladder Lesion Classification: Comparative Study (e87193)
Yung-Chi Shih, Cheng-Yang Wu, Shi-Wei Huang, Chung-You Tsai. 4396

Remote Patient Monitoring in Louisiana Medicare Beneficiaries With Diabetes or Hypertension: Retrospective Cohort Study of 2016-2020 Claims Data (e80782)
Debra Winberg, Charles Stoecker, Jian Li, Lizheng Shi. 4417

Extended Grammar of Systematized Nomenclature of Medicine – Clinical Terms for Semantic Representation of Clinical Data: Methodological Study (e80314)
Christophe Gaudet-Blavignac, Julien Ehrsam, Monika Baumann, Adel Bensahla, Mirjam Mattei, Yuanyuan Zheng, Christian Lovis. 4431

Establishment and Optimization of a Patient-Reported Outcome–Based Electronic-Diary for Symptoms Evaluation in Patients With Gastroesophageal Reflux Disorder: Prospective Cohort Study (e83680)
Yun-Chun Chen, Yen-Po Wang, Jui-Hsuan Hung, Da-Wei Wang, Shang-Liang Wu, Li-Fen Chen, Yueh-Hsin Ping, Mei-Lien Pan, Ching-Liang Lu. 4446

XSL•FO
RenderX

Quality of Life Trajectories With Integration Into Electronic Health Records for High-Resolution Patient Outcomes: Algorithm Development and Validation Study (e79834)
Martin Dugas, Robin Fleige, Max Blumenstock, Stephan Feder, Tobias Dittrich, Niels Siegel, Celine Bergmann, Luis Wettach, Sandra Sauer, Pavlina Lenga, Sandro Krieg, Susanne Dugas-Breit, Lucy Kessler, Kosima Kosmalla, Fleur Fritz-Kebede, Martin Loos, Hans-Christoph Friederich, Thomas Pausch, Matthias Ganzinger. 4459

Communicative Behaviors in an Internet-Based Intervention for Individuals With Autism: Mixed Methods Analysis (e76527)
Britta Westerberg, Karin Jacobson, Maria Unenge Hallerbäck, Susanne Bejerot, Fredrik Holländare. 4470

Initial Insights Into an Institutional Secure Large Language Model for Magnetic Resonance Imaging Examination Requests: Retrospective Study (e82579)
James Hallinan, Naomi Leow, Yi Low, Aric Lee, Wilson Ong, Matthew Chan, Ganakirthana Devi, Stephanie He, Daniel Loh, Desmond Lim, Xi Low, Mei Lim, Clement Yong, Weizhong Sng, Ee Teo, Jiong Tan, Naresh Kumar, Andrew Makmur, Yonghan Ting. 4484

Artificial Intelligence in Health Professions Education: Qualitative Study of Student Experiences (e82432)
Marina Guirguis, Salomon Fotsing, Joanne Fevry, Christine Landry, Diane Bouchard-Lamothe, Jennifer Lacroix, Alireza Jalali. 4495

Improving Retrieval Augmented Generation for Health Care by Fine-Tuning Clinical Embedding Models: Development and Evaluation Study (e82997)
Kamyar Arzideh, Henning Schäfer, Ahmad Idrissi-Yaghir, Cynthia Schmidt, Bahadir Eryilmaz, Mikel Bahn, Amin Turki, Olivia Pollok, Eva Hartmann, Philipp Winnekens, Katarzyna Borys, Johannes Haubold, Felix Nensa, René Hosch. 4509

Anticipating Moral and Economic Considerations, Opportunities, and Potential Frictions for AI in Medical Imaging: Multistakeholder Cocreation Study (e83407)
Martin Schilder, Alexandra Keyser, Susan van Hees, Alessandro Sbrizzi, Wouter Boon. 4548

Forecasting Waitlist Trajectories for Patients With Metabolic Dysfunction–Associated Steatohepatitis Cirrhosis: A Neural Network Competing Risk Analysis (e68247)
Gopika Punchhi, Yingji Sun, Eunice Tan, Naomi Hlaing, Chang Liu, Sumeet Asrani, Sirisha Rambhatla, Mamatha Bhat. 4567

End-to-End Platform for Electrocardiogram Analysis and Model Fine-Tuning: Development and Validation Study (e81116)
Lucas Bickmann, Lucas Plagwitz, Antonius Büscher, Lars Eckardt, Julian Varghese. 4581

Predicting the Intention to Use Generative Artificial Intelligence for Health Information: Comparative Survey Study (e75648)
Jörg Matthes, Anne Reinhardt, Selma Hodzic, Jaroslava Ka ková, Alice Binder, Ljubisa Bojic, Helle Maindal, Corina Paraschiv, Knud Ryom. .
4 5 9 9

Key Information Influencing Patient Decision-Making About AI in Health Care: Survey Experiment Study (e75615)
Xuan Zhu, Austin Stroud, Sarah Minteer, Dong Yoo, Jennifer Ridgeway, Maryam Mooghali, Jennifer Miller, Barbara Barry. 4613

Intervention in Health Misinformation Using Large Language Models for Automated Detection, Thematic Analysis, and Inoculation: Case Study on COVID-19 (e75500)
Samira Malek, Christopher Griffin, Robert Fraleigh, Robert Lennon, Vishal Monga, Lijiang Shen. 4639

Developing an AI-Assisted Tool That Identifies Patients With Multimorbidity and Complex Polypharmacy to Improve the Process of Medication Reviews: Qualitative Interview and Focus Group Study (e74304)
Aseel Abuzour, Samantha Wilson, Alan Woodall, Frances Mair, Asra Aslam, Andrew Clegg, Eduard Shantsila, Mark Gabbay, Michael Abaho, Danushka Bollegala, Harriet Cant, Alan Griffiths, Layik Hama, Gary Leeming, Emma Lo, Simon Maskell, Maurice O'Connell, Olusegun Popoola, Sam Relton, Roy Ruddle, Pieta Schofield, Matthew Sperrin, Tjeerd Van Staa, Iain Buchan, Lauren Walker. 4661

Evaluating and Validating Large Language Models for Health Education on Developmental Dysplasia of the Hip: 2-Phase Study With Expert Ratings and a Pilot Randomized Controlled Trial (e73326)
Hui Ouyang, Gan Lin, Yiyuan Li, Zhixin Yao, Yating Li, Han Yan, Fang Qin, Jinghui Yao, Yun Chen. 4676

XSL•FO
RenderX

Data Poisoning Vulnerabilities Across Health Care Artificial Intelligence Architectures: Analytical Security Framework and Defense Strategies (e87969)
Farhad Abtahi, Fernando Seoane, Ivan Pau, Mario Vega-Barbas. . . . 4698

Implementing an Artificial Intelligence Decision Support System in Radiology: Prospective Qualitative Evaluation Study Using the Nonadoption Abandonment Scale-Up, Spread, and Sustainability (NASSS) Framework (e80342)
Sundresan Naicker, Paul Schmidt, Bruce Shar, Amina Tariq, Ashleigh Earnshaw, Steven McPhail. . . . 4718

Integrated Prediction System for Individualized Ovarian Stimulation and Ovarian Hyperstimulation Syndrome Prevention: Algorithm Development and Validation (e78245)
Jingjing Chen, Jianjuan Zhao, Huiyu Qiu, Yanhui Liu, Yunqi Zhang, Qicheng Sun, Yan Yi, Hongying Tang, Jing Zhao, Bin Xu, Qiong Zhang, Ge Yang, Hui Li, Junjie Liu, Zhongzhou Yang, Shaolin Liang, Yanping Li, Jing Fu. . . . 4735

Patterns of AI Use in Clinical Work by Hospitalists: Survey Study (e85973)
Prabhava Bagla, Jasmah Hanna, Bhargav Marthambadi, Stacey Watkins. . . . 4750

Psychosis Risk and Generative Artificial Intelligence Use Frequency, Motivations, and Delusion-Like Experiences: Cross-Sectional Survey Study (e85038)
Benjamin Buck, Anne Maheux. . . . 4763

Scalable and Robust Artificial Intelligence for Spine Alignment Assessment: Multicenter Study Enabled by Real-Time Data Transformation (e78396)
Guilin Chen, Nan Meng, Yipeng Zhuang, Zhe Chen, Zhen Bian, Zhaoyang Gong, Jiawei Shi, Tao Huang, Xihe Kuang, Pengyu Lu, Cong Nie, Qifeng Yu, Zefu Chen, Hui Jiang, Zhongmin Zhang, Chaojun Zheng, Yu Liang, Nan Wu, Jason Cheung, Jianguo Zhang, Teng Zhang. . . . 4774

Measuring the Impact of AI on Report-Drafting Efficiency in Chest Computed Tomography Interpretation: Retrospective Analysis (e77967)
Weiqi Liu, You Wu, Wei Yu, Mark Bittle, Zhuozhao Zheng, Hadi Kharrazi. . . . 4788

Enhancing Detection of Message Intents in a Mobile Health Smoking-Cessation Intervention Using Large Language Model Fine-Tuning, Data Downsampling, and Error Correction: Algorithm Development and Validation (e83437)
Shagoto Rahman, Cornelia Pechmann, Ian Harris. . . . 4801

Evaluation of an Artificial Intelligence Conversational Chatbot to Enhance HIV Preexposure Prophylaxis Uptake: Development and Usability Internal Testing (e79671)
Jun Tao, Ellie Pavlick, Amaris Grondin, Josue Bustamante, Harrison Martin, Hannah Parent, Natalie Fenn, Alexi Almonte, Amanda Maguire-Wilkerson, Mofan Gu, Jack Rusley, Bryce Perler, Tyler Wray, Amy Nunn, Philip Chan. . . . 4829

Developing a Quality Evaluation Index System for Health Conversational Artificial Intelligence: Mixed Methods Study (e83188)
Weizhen Liao, Meng Li, Chengyu Ma, Youli Han, Dan Wang, Haopeng Liu, Yi Wang, Zijie Feng, Huichao Wang, Yiru Guan. . . . 4845

Developing a Service Quality Index System for AI Health Care Chatbots: Mixed Methods Study (e83051)
Yu Gu, Xinyi Wang. . . . 4866

Comparison of Emotional Content in Text Responses From Physicians and AI Chatbots to Patient Health Queries: Cross-Sectional Study (e85516)
Daniel Burns, Channing Bice, Paul Johnson, Nicholas Chia, Timothy Robinson. . . . 4878

Young Adults' Interactions With Food and Nutrition Content on Social Media and Implications for Intervention Design: Semistructured Interview Study (e89344)
Hao Tang, Amy Ahern, Marie Spreckley, Andrea Smith. . . . 4891

XSL•FO
RenderX

Assistive Robotics for Healthy Aging: A Foundational Phenomenological Co-Design Exercise (e77179)
Stephen Potter, Mark Hawley, Angela Higgins, Farshid Amirabdollahian, Mauro Dragone, Alessandro Di Nuovo, Praminda Caleb-Solly. 4906

Adoption of Internet of Things in Health Care: Weighted and Meta-Analytical Review of Theoretical Frameworks and Predictors (e64091)
Inês Veiga, Tiago Oliveira, Mijail Naranjo-Zolotov, Ricardo Martins, Stylianos Karatzas. 4924

Patient Perceptions of Blockchain-Based Health Information Exchange: User-Centered Design Study (e78849)
Richard Guse, Shanshan Hu, Scott Thiebes, Christina Erler, Catherine Caridia, Wilhelm Stork, Martin Gersch, Ali Sunyaev. 4955

Resistance to Blockchain Adoption in Health Care Organizations: Evidence From a Cross-Sectional Study (e77933)
Wilfried Bazomanza Nzabandora, Sehl Mellouli, Rohit Nishant. 4974

Digital Discourse, Secondary Victimization, and Psychological Harm: Mixed-Methods Analysis of System Justification in the #MeToo Movement (e75533)
Harsh Parekh, Shriya Thakkar, Paras Bhatt, Patricia Akello. 4993

Analysis of Multilevel Factors Mobilizing the Spectrum of Interorganizational Knowledge Sharing for Facilitating Digital Transformation at Scale: Qualitative Study (e83345)
Hajar Mozaffar, Robin Williams, Kathrin Cresswell. 5011

Quality Challenges in Municipal Telecare Call Center Services: Qualitative Evaluation Using the Anchored, Realistic, Cocreated, Human, Integrated, and Evaluated (ARCHIE) Framework (e76054)
Linda Grøndal-Eeles, Janne Dugstad, Hilde Eide, Etty Nilsen. 5031

Exploring the Dynamics of Actors, Structural Factors, and Bricolage in the Implementation and Sustainability of eHealth Solutions: Qualitative Multiple-Case Study (e79999)
Susanne Eriksen, Christine Øye, Anne Dahler. 5045

“I Want to Spend My Time Living”—Experiences With a Digital Outpatient Service With a Mobile App for Tailored Care Among Adults With Long-Term Health Service Needs: Qualitative Study Using Thematic Analysis (e79155)
Heidi Holmen, Erik Fosse. 5060

Feasibility and Acceptability of AI-Powered Tools for Early Autism Screening in Egypt: Semistructured Focus Group Study (e82564)
Pratheepan Yogarajah, Ammal Metwally, Priyanka Chaurasia, Ghada Elshaarawy, Shereen El Khateeb, Engy Ashaat, Amal Elsaeid, Nahed Elghareeb, Amira ElRifay. 5072

What Patients With Asthma Share When No One Listens: Multimethod Observational Study of Patient Narratives on Reddit (e77027)
Elena Curto-Sánchez, Gabriela Salazar-Palacios, Ana Martín-Varillas, Estela Prieto-Maíllo, Jacinto Ramos-González, Ignacio Dávila-González, Domingo Palacios-Ceña, Juan Cuenca-Zaldivar. 5089

Moderators’ Role in Fostering Health Practitioner Engagement Within an Online Active Learning Program (The Community Fracture Capture Learning Hub): Qualitative Relational Content Thematic Analysis (e83764)
Ahmed Fathalla, Shanton Chang, Ralph Audehm, Cherie Chiang, Alexandra Gorelik, Christopher Yates, Steve Snow, Rahul Barmanray, Sarah Price, Lucy Collins, John Wark. 5106

Quantifying Innovation in Stroke: Large Language Model Bibliometric Analysis (e70754)
Adam Marcus, Georgina Lockwood-Taylor, Daniel Rueckert, Paul Bentley. 5123

XSL•FO
RenderX

Electrocardiogram-Based Mental Stress Detection Amid Everyday Activities Using Machine Learning: Model Development and Validation Study (e80450)
Buelent Uendes, Alex Antonides, Sjors van de Ven, Denise van der Mee, Eco de Geus, Mark Hoogendoorn. 5137

Examining the Association Between Internet Use and Perceived Stress in Adults: Longitudinal Observational Study Combining Web Tracking Data With Questionnaires (e78775)
Mohammad Belal, Nguyen Luong, Talayeh Aledavood, Juhi Kulshrestha. 5163

The Impact of Evaluation Strategy on Sepsis Prediction Model Performance Metrics in Intensive Care Data: Retrospective Cohort Study (e72083)
Dang-Khoa Do, Patrick Rockenschaub, Sebastian Boie, Oliver Kumpf, Hans-Dieter Volk, Felix Balzer, Falk von Dincklage, Gregor Lichtner. . .
5 1 8 9

Physical Activity Recommendations Tailored by a Predictive Model for Adults With High Blood Pressure: Observational Study (e78492)
Yuhui Yang, Manqing Chen, Weiwei Hu, Yifan Fu, Xingyan Li, Zhenli Liao, Hongman Feng, Yaling Zhao, Leilei Pei, Baibing Mi, Fangyao Chen.
5 2 0 7

Comparative Analysis of Prescriptions and Pharmacy Services in Internet-Based Psychiatric Hospital During and After the COVID-19 Pandemic: Retrospective Cross-Sectional Observational Study (e74059)
Guowei Deng, Hui Xia, De-wei Shang, Yuguan Wen, Jinqing Hu, Yaqian Tan. 5250

Time Savings Through an AI Speech Assistant for Nursing Documentation: Pre-Post Time-Motion Study in German Long-Term Care (e86078)
Katja Schwabe, Drin Ferizaj, Susann Neumann, Sandra Strube-Lahmann, Nils Lahmann. 5271

Computerized Self-Reported Medical History Taking to Support Early Rule Out of Major Adverse Cardiac Events in Patients With Acute Chest Pain: Post Hoc Analysis of the CLEOS-CPDS Prospective Cohort Study (e76087)
Helge Brandberg, Carl Sundberg, Jonas Spaak, Sabine Koch, Thomas Kahan. 5284

Acceptability, Feasibility, and Perceived Effectiveness of Video-Based Patient Records for Supporting Care Delivery to Older Adults With Frailty: Nonrandomized Mixed Methods Pilot Study (e77318)
Phoebe Averill, Rachael Lear, Ricky Odedra, Susannah Long, Alex Taylor, Pi-Jung Charville, Jessica Fernandes, Uzoamaka Ekeogu, Jessica Leombruno, Sophia Ellis, Erik Mayer. 5329

Psychotherapists' Trust, Distrust, and Generative AI Practices in Psychotherapy: Qualitative Study (e88932)
Jinyan Kuang, Amber Pope, Yixuan Zhang. 5349

The Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB) for the Evaluation of Medical Large Language Model Question-Answer Studies: Development and Pilot Validation (e87057)
Carrie Ye, Joseph Mitchell, Daniel Baumgart, Zechen Ma, Angela Fung, Daniela Orellana, Juel Chowdhury, Abdullah Abass, Steven Katz, Jacob Jaremko, Pierre Boulanger, Claire Barber, Gillian Lemermeyer, Hosna Jabbari, Lili Mou, Maryam Mirzaei, Mary Githumbi, Puneeta Tandon, Randy Goebel, Rhys Clark, Whitney Hung, Marjan Abbasi, Farhad Maleki, Scott Klarenbach, Mohamed Abdalla. 5365

Disclaimers and Referral Patterns for Medical Advice Across Urgency Levels: Large Language Model Evaluation Study (e84668)
Florian Reis, Louis Agha-Mir-Salim, Richard Hickstein, Moritz Reis, Sophie Piper, Felix Balzer, Sebastian Boie. 5379

Effectiveness of AI-Assisted Patient Health Education Using Voice Cloning and ChatGPT: Prospective Randomized Controlled Trial (e81387)
Yan Sun, Shangqing Xu, Hongying Jin, Xiaoyan Han, Kangqi Jin, Yimei Zhang, Xiaoli Ma, Huaping Wei, Minjie Ma. 5392

XSL•FO
RenderX

Reliability of Large Language Model Generated Clinical Reasoning in Assisted Reproductive Technology: Blinded Comparative Evaluation Study (e85206)
Dou Liu, Ying Long, Sophia Zuoqiu, Di Liu, Kang Li, Yiting Lin, Hanyi Liu, Rong Yin, Tian Tang. 5406

Framing the Convergence of One Health and Digital Health in the Global South With a Gender-Sensitive Foresight Perspective: Delphi Study Using Latent Semantic Analysis (e78702)
Jude Kong, Nicola Bragazzi. 5508

Channel Allocation and Equity in Preventive Campaigns for Older Adults: Agent-Based Modeling Study (e88429)
Jihye Lee, Juyoung Park, Yuna Kim, Duk-Jo Kong. 5522

Commentary

Translating Telehealth Communication Research Into Patient-Centered, Implementable Practice (e93690)
Rachel Pittmann, Paula Koppel, David Barrett. 1588

Viewpoints

Integrating Generative AI Into Patient-Centered Clinical Decision Support: Viewpoint on Research and Practice Considerations (e81628)
Prashila Dullabh, Courtney Zott, Nicole Gauthreaux, Caroline Peterson, Abigail Aronoff, Kistein Monkhouse, Dean Sittig. 1608

Deep Research Agents: Major Breakthrough or Incremental Progress for Medical AI? (e88195)
Matthew Wong, Ariel Ong, David Merle, Pearse Keane. 1619

From Knowledge Graphs to Digital Twins: Perspectives on Modeling Patient Outcomes for Health Care Quality Assessment (e81946)
Anna-Katharina Nitschke, Juan Diaz Ochoa, Simone Neumaier, Markus Knott. 1631

Extrinsic Trust as a Contractual Framework for Accountable AI in Health Care: Viewpoint (e83903)
Anthony Kelly. 1653

Digital Twins for Just-in-Time Adaptive Interventions (JITAIs): Framework for Optimizing and Continually Improving JITAIs (e72830)
Asim Gazi, Daiqi Gao, Susobhan Ghosh, Ziping Xu, Anna Trella, Predrag Klasnja, Susan Murphy. 1665

AI Agents and Epidemic Intelligence on Respiratory Infectious Diseases: Toward a Conceptual Framework Integrating Decision Support (e86936)
Liuyang Yang, Liyu Shan, Xiaolin Cao, Jinzhao Cui, Michael Tong, Yan Niu, Ting Zhang. 1686

Scalable Agile Framework for Execution in AI for Medical AI Ethics Policy Design in Small- and Medium-Sized Enterprises (e80028)
Ion Nemteanu, Adir Mancebo Jr, Leslie Joe, Ryan Lopez, Patricia Lopez, Warren Pettine. 1694

From Agents to Governance: Essential AI Skills for Clinicians in the Large Language Model Era (e86550)
Weiping Cao, Qing Zhang, Jialin Liu, Siru Liu. 1717

XSL•FO
RenderX

Collaborative and Cooperative Hospital “In-House” Medical Device Development and Implementation in the AI Age: The European Responsible AI Development (EURAID) Framework Compatible With European Values (e80754)
Anett Schönfelder, Maria Eberlein-Gonska, Manfred Hülsken-Giesler, Florian Jovy-Klein, Jakob Kather, Elisabeth Kohoutek, Thomas Lennefer, Elisabeth Liebert, Myriam Lipprandt, Rebecca Mathias, Hannah Muti, Julius Obergassel, Thomas Reibel, Ulrike Rösler, Moritz Schneider, Larissa Schlicht, Hannes Schlieter, Malte Schmieding, Nils Schweingruber, Martin Sedlmayr, Reinhard Strametz, Barbara Susec, Magdalena Wekenborg, Eva Weicken, Katharina Weitz, Anke Diehl, Stephen Gilbert. 1726

AI Triage in Primary Care: Building Safer and More Equitable Real-World Evidence (e88396)
Aymn Alamoudi, Evangelos Kontopantelis, Salwa Zghebi, Benjamin Brown. 1749

Assessing the Evolution and Influence of Medical Open Databases on Biomedical Research and Health Care Innovation: A 25-Year Perspective With a Focus on Privacy and Privacy-Enhancing Technologies (e58954)
Albert Yang, Mei-Lien Pan, Henry Lu, Chung-Yueh Lien, Da-Wei Wang, Chih-Hsiung Chen, Der-Cherng Tarng, Dau-Ming Niu, Shih-Hwa Chiou, Chun-Ying Wu, Ying Sun, Shih-Ann Chen, Shuu-Jiun Wang, Wayne Sheu, Chi-Hung Lin. 5421

Africa’s Digital Health Revolution: The Digital Fit-Viability Model to Move From Innovation to Scaled Implementation (e63495)
Afra Jiwa, Antony Ngatia, Karim Benali, Niclas Boehmer, Sangu Delle, Patrick Emedom-Nnamdi, Chris Fofie, Christine O’Brien, Tobi Olatunji, Kate Obayabgona, Milind Tambe, Richard Fletcher, Adeline Boatin, Bethany Hedt-Gauthier. 5436

Tutorials

Using an Adult-Designed Wearable for Pediatric Monitoring: Practical Tutorial and Application in School-Aged Children With Obesity (e76166)
Alessandra Angelucci, Andrea Aliverti, Matteo Vandoni, Daniela Lucini, Cristiana Larizza, Vittoria Carnevale Pellino, Gianvincenzo Zuccotti, Valeria Calcaterra. 1761

SMARTCLOTH Prototype for Dietary Management in Patients With Diabetes Mellitus: Tutorial on Human-Centered Design Methodology for Health Care Hardware Development (e75744)
Jose Palomares, Rafael Molina-Luque, Fernando León-García, Irene Casares-Rodríguez, María García-Rodríguez, María Villena Esponera, Guillermo Molina-Recio. 1781

Developing a Trauma-Informed Social Media Campaign to Disseminate Endometriosis-Specific Qualitative Art-Based Research Findings: Tutorial (e83491)
Kerry Marshall, Hargun Dhillon, A Howard, Heather Noga, Grace Yang, William Zhu, Jessica Sutherland, Sarah Lett, Anna Leonova, Paul Yong, Natasha Orr. 1807

Team-Based Analysis of Large-Scale Qualitative Data: Tutorial Using a Nationwide SMS Text Messaging Poll of Youth (e72526)
Melissa DeJonckheere, Samantha Chuisano, Marika Waselewski, Kendrin Sonneville, Tammy Chang. 1829

Corrigenda and Addendas

Correction: Characterization of Models for Identifying Physical and Cognitive Frailty in Older Adults With Diabetes: Systematic Review and Meta-Analysis (e92717)
Xia Wang, Shujie Meng, Xiang Xiao, Liu Lu, Hongyan Chen, Yong Li, Rong Zhang, Qiwu Jiang, Shan Liu, Ru Gao. 5224

XSL•FO
RenderX

Correction: User Profiles and Engagement in a Hypertension Self-Management App: Cross-Sectional Survey (e93515)
Felix Muehlensiepen, Susann May, Frances Seifert, Eileen Wengemuth, Olen Johannsen, Martin Middeke, Martin Heinze, Pascal Petit, Nicolas Vuillerme, Sebastian Spethmann, Dunja Bruch. . . . 5226

Correction: Leveraging Social Media to Achieve Population-Level Reach of Lung Cancer Screening-Eligible Individuals: A RE-AIM Framework Perspective (e94664)
Lisa Carter-Bawa, Jamie Ostroff, Susan Rawl, Erin Hirsch, Smita Banerjee, Andrew Ciupek, Robert Comer, Minal Kale, Katherine Leopold, Patrick Monahan, James Slaven Jr, Francis Valenzona, Renda Wiener, Ana Vielma. . . . 5228

Correction: Integrating Text and Image Analysis: Exploring GPT-4V's Capabilities in Advanced Radiological Applications Across Subspecialties (e91415)
Felix Busch, Tianyu Han, Marcus Makowski, Daniel Truhn, Keno Bressem, Lisa Adams. . . . 5230

Correction: Evidence for Digital Health Tools Designed to Support the Triage of Musculoskeletal Conditions in Primary, Urgent, and Emergency Care Settings: Scoping Review (e92722)
Linda Truong, James Wrightson, Raphaël Vincent, Eunice Lui, Jamon Couch, Ellen Wang, Cobie Starcevich, Dean Giustini, Alex Haagaard, Elena Lopatina, Niels van Berkel, Michael Rathleff, Clare Ardern. . . . 5232

Correction: Acceptability of Health Information Technology by Health Care Professionals: Where We Are Now and How We Can Fill the Gap (e89383)
Corinne Isnard-Bagnis, Stéphane Mouchabac, Riadh Lebib, Hervé Bismut, Pierre Geoffroy. . . . 5234

Correction: Combining Artificial Intelligence and Human Support in Mental Health: Digital Intervention With Comparable Effectiveness to Human-Delivered Care (e88640)
Clare Palmer, Emily Marshall, Edward Millgate, Graham Warren, Michael Ewbank, Elisa Cooper, Samantha Lawes, Alastair Smith, Chris Hutchins-Joss, Jessica Young, Malika Bouazzaoui, Morad Margoum, Sandra Healey, Louise Marshall, Shaun Mehew, Ronan Cummins, Valentin Tablan, Ana Catarino, Andrew Welchman, Andrew Blackwell. . . . 5236

Correction: Effectiveness of a Web-Based Medication Education Course on Pregnant Women's Medication Information Literacy and Decision Self-Efficacy: Randomized Controlled Trial (e91835)
Suya Li, Hui-Jun Chen, Jie Zhou, Yi-Bei Zhouchen, Rong Wang, Jinyi Guo, Sharon Redding, Yan-Qiong Ouyang. . . . 5238

Correction: Culturally Adapted Guided Internet-Based Cognitive Behavioral Therapy for Hong Kong People With Depressive Symptoms: Randomized Controlled Trial (e88495)
Jia-Yan Pan, Jonas Rafi. . . . 5240

Correction: Ingroup Favoritism Surrounding COVID-19 Vaccinations in the Hispanic Communities: Experimental Study (e90652)
Juwon Hwang, Asya Cooley, Skye Cooley, Robert Hinck. . . . 5244

Correction: A Supervised Explainable Machine Learning Model for Perioperative Neurocognitive Disorder in Liver-Transplantation Patients and External Validation on the Medical Information Mart for Intensive Care IV Database: Retrospective Study (e88952)
Zhendong Ding, Linan Zhang, Yihan Zhang, Jing Yang, Yuheng Luo, Mian Ge, Weifeng Yao, Ziqing Hei, Chaojin Chen. . . . 5246

Correction: Infoveillance of COVID-19 Infections in Dentistry Using Platform X: Descriptive Study (e86469)
Alghalia Al-Mansoori, Ola Al Hayk, Sharifa Qassmi, Sarah Aziz, Fatima Haouari, Tawanda Chivese, Faleh Tamimi, Alaa Daud. . . . 5248

Research Letters

Quality of Conventional versus Artificial Intelligence Oral Surgery Consent Forms: Comparative Analysis (e59851)
Jan Gaessler, Bernhard Remschmidt, Ann-Kathrin Jopp, Behrouz Arefnia, Adrian Franke, Marcus Rieder. . . . 5305

XSL•FO
RenderX

National Institutes of Health–Funded Artificial Intelligence and Machine Learning Research, 2019 2023: Cross-Sectional Study (e84861)
Joshua Le, Joseph Morrison, Atul Malhotra, Shamim Nemati, Gabriel Wardi, James Ford. 5309

Transformer-Based Topic Modeling: Characterizing Cannabis Product Adverse Experiences Self-Reported as Requiring Medical Attention on Reddit (e82661)
Tim Mackey, Matthew Nali, Meng Larsen, Zhuoran Li, Cassandra Taylor, Beverly Wolpert, Catharine Trice. 5314

Sex Differences in Medical Specialist Physicians' Electronic Health Record In-Basket Workloads and the Implications for Compensation and Equity: Retrospective Cross-Sectional Study (e79172)
Ashwini Nadkarni, Laura Glick, Molly Becker, Michael Healey, David Stein, Mahyar Heydarpour, Lisa Rotenstein. 5319

Trends in Venture Capital Investment in AI-Driven Biopharmaceutical Startups (e84968)
Abhishek Bazaz, Yunan Ji, Mariana Socal, So-Yeon Kang. 5324

News and Perspectivess

Wearable Air Samplers Reveal How Wildfire Shapes the Exposome (e93193)
Virginia Gewin. 5444

Influencing the Influencers: How Health Experts Are Partnering With Content Creators to Fight Misinformation Online (e93450)
Wendy Glauser. 5447

Is Going Analog Good for Children and Teens' Mental Health and Well-Being? (e94018)
Sara Novak. 5450

Physician, Restore Thyself? The Digital Gap in Physician Well-Being Support (e93338)
Jenny Cato. 5453

Using AI to Assess Potential Zoonotic Threats (e93261)
Cliff Dominy. 5456

As Social Media Scales Back Fact-Checking, Can Technologies Fill the Gap? (e95730)
Wendy Glauser. 5459

Further Promise and Potential for Precision Medicine in Oncology (e95657)
Shalini Narang. 5462

Emerging Risks of AI-to-AI Interactions in Health Care: Lessons From Moltbook (e96199)
Tejas Athni. 5465

Our AI-Powered Discoveries Are Trapped in a Predigital System (e96018)
Boon-How Chew. 5468

Data Governance Lessons From an Unvalidated Dataset (e94518)
Cliff Dominy. 5471

The Right to Understand in Health Care AI (e95090)
Anshu Ankolekar. 5474

XSL•FO
RenderX

When Old Diseases Return: Cholera, Crisis, and Digital Surveillance in Fragile Settings (e94818)
Isabelle Basbouss-Serhal. 5478

How Digital Diaries Are Changing Communication and Recovery in the Intensive Care Unit and Beyond (e92661)
Jenna Congdon. 5481

Digital Narrative Medicine: Translating Stories Into Medical Education and Better Patient Care (e93037)
Candice Sage. 5484

How Digital Health Technology Is Extending Independent Living (e93767)
Jenna Congdon. 5488

What Health Care Organizations Have Learned From Telecommunication Outages (e91456)
Catharine Solomon. 5491

A Frontline Worker's Take on Hybrid Care Implementation in the Hospital Setting (e90879)
Jenna Congdon. 5494

WHOOP, There It Is: Lessons From WHOOP's FDA Warning Letter (e90882)
Blythe Karow. 5497

UnitedXR Europe 2025: Aligning Health Care Extended Reality (e90727)
Jose Costa. 5501

When Lived Experience Designs the Intervention (e91371)
Trevor van Mierlo. 5505

XSL•FO
RenderX

Innovations in Deaf Health Care Communication: Systematic Review of Sign Language Recognition Systems

Milena Soriano Marcolino[1,2,3], MD, PhD; Lucca Fagundes Ramos de Oliveira[2,4], MD; Lucas Rocha Valle[1,2]; Luiza Marinho Motta Santa Rosa[2,5], MD; Gabriela Teodora de Souza Sanches[1,2]; Natalia Sales Santos[6], MSc; Michelle Ralil da Costa[3], PhD; Elidea Lucia Almeida Bernardino[7], PhD; Raniere Alislan Almeida Cordeiro[7], PhD; Raquel Oliveira Prates[6], PhD; Zilma Silveira Nogueira Reis[1], MD, PhD; Mario Fernando Montenegro Campos[6], PhD

[1]Medical School, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[2]Telehealth Center, University Hospital, Universidade Federal de Minas Gerais, Avenida Professor Alfredo Balena, 190, Santa Efigênia, Belo Horizonte, Brazil

[3]Institute for Health Technology Assessment, Porto Alegre, Brazil

[4]Medical School, Universidade Federal de Ouro Preto, Ouro Preto, Brazil

[5]Faculdade de Ciências Médicas de Minas Gerais, Belo Horizonte, Brazil

[6]Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[7]Faculty of Arts and Sciences, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Corresponding Author:
Milena Soriano Marcolino, MD, PhD
Medical School, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Abstract

Background: Deaf individuals often face communication challenges when interacting with those who can hear. Within health care settings, these challenges may pose risks to their safety, potentially resulting in misdiagnoses, treatment errors, and decreased quality of care.

Objective: This study aims to systematically review the evidence on communication systems reported in the literature that use human-computer interaction techniques to support communication between deaf individuals who use sign language and hearing health professionals in health care settings. The review focuses on systems that are either currently in use or proposed for use in health care and that have been tested using human participants or videos of human users.

Methods: A comprehensive search was performed via MEDLINE, Web of Science, ACM, IEEE Xplore, Scopus, and Google Scholar in March 2025. The inclusion criteria comprised studies developing a sign language recognition system within a health care context and testing with human users. Eligible studies underwent screening by 2 independent investigators (LRV and LMMSR or LFRdO and GTdSS), with any disagreements resolved by a senior researcher (MSM).

Results: The search retrieved 21,778 publications, and screening of reference lists identified 2 additional studies, resulting in a total of 23 studies meeting the eligibility criteria. Most systems (15/23, 65.2%) were image-based, while 34.8% (8/23) relied on sensors (glove-based or depth-sensing). Applications varied across health care settings, including general hospital care (10/23, 43.5%), emergencies (8/23, 34.8%), and primary care (4/23, 17.4%). All systems were in the development and testing stage, with no data on security and psychological impacts. Accuracy ranged from 25% to 100% for image-based and 72% to 99.7% for sensor-based systems. Bidirectionality and facial expression recognition, crucial for effective communication, were largely overlooked.

Conclusions: Image-based systems were more common than sensor-based ones, though both showed wide variability in accuracy in recognizing and interpreting signs. Most systems failed to address critical aspects such as bidirectional communication and the recognition of facial expressions, essential for effective communication. None fully addresses the requirements for integration into health care settings. These findings highlight the need for further research on implementation, usability, and impact on the quality of care for deaf patients.

International Registered Report Identifier (IRRID): RR2-10.2196/55427

(J Med Internet Res 2026;28:e70417) doi:10.2196/70417

KEYWORDS

computer neural network; artificial intelligence; biomedical technology; sign language; hearing loss; deafness; communication barriers; gestures; PRISMA; Preferred Reporting Items for Systematic Reviews and Meta-Analyses

XSL•FO
RenderX

Introduction

Sign languages are a form of communication characterized by the successful coordination of gestures, body, head, and hand motions, with facial expressions. They are naturally developed and highly structured systems, governed by a set of linguistic rules, different from spoken languages with no standardized written forms [1]. They enable comprehensive and rich linguistic systems for formulating words and phrases, making them the primary mode of communication for many hearing-impaired people [2]. According to the World Health Organization, over 430 million people worldwide experience disabling hearing loss, a number expected to exceed 700 million by 2050 [3]. Despite the World Federation of the Deaf estimating that there are over 200 sign languages and over 70 million deaf individuals who use them, reliable data on sign language users remains scarce [4].

Even though sign language is the primary mode of communication for millions of deaf individuals, they still encounter significant barriers in daily interactions, particularly in critical areas such as health care, education, and public services, where most professionals do not understand or use sign language [5-7]. Specifically in health care, this communication gap can lead to misdiagnoses, inadequate treatment, and reduced quality of care [8].

In response to this challenge, and recognizing the significance of sign language in fostering social integration for the deaf community, researchers have developed sign language recognition (SLR) systems. SLR refers to the identification and interpretation of sign language gestures and movements [9]. These systems aim to facilitate real-time translation between sign and spoken languages, to build a communication bridge between deaf individuals who communicate through sign language and ordinary people, and to enhance accessibility and inclusion across multiple domains, including health care [10]. In this pursuit, machine learning and other artificial intelligence (AI) techniques have become important emerging tools to help overcome communication barriers, enhancing not only accessibility but also valuing the cultural identity of the deaf community.

Regarding automatic SLR systems, it is crucial to highlight the use of advanced algorithms and other AI techniques to interpret the complex gestures and movements inherent to sign language communication. With extensive use of advanced computer vision algorithms and machine learning techniques, these systems are becoming increasingly able to automatically recognize and translate signs into text of the target language, thereby facilitating effective communication between deaf individuals and those who do not understand sign language [11-13].

The SLR systems comprise different types of methodologies and technologies to recognize and translate sign language. In this paper, we organize them into 2 main categories: image-based and sensor-based. Even though strictly speaking, cameras are sensors, we used sensor-based to identify the approaches that use other sensor modalities. Image-based approaches use computer vision techniques coupled with deep learning models to analyze video streams from cameras that capture hand gestures, body movements, and facial expressions to accurately detect and interpret signs [11,13,14]. Sensor-based systems use wearable devices equipped with sensors, such as accelerometers and gyroscopes, to capture hand and body movements and translate signs into written or spoken language [11,12,15]. These systems can be divided into 2 subgroups: glove-based systems and depth-sensor systems. Glove-based systems use specialized gloves embedded with sensors to capture fine-grained hand movements, enabling real-time translation into spoken or written language. In contrast, depth-sensor systems rely on depth-sensing cameras, such as Kinect or light detection and ranging, to obtain 3D information about hand and body movements. By analyzing the 3D pose of different body parts, these systems allow more precise recognition and interpretation of sign language gestures [16].

Each of the aforementioned types of systems has strengths and limitations concerning accuracy in detecting and interpreting signs, portability, cost, and accessibility. Additionally, user-friendliness is an important consideration in the development and implementation of SLR systems. It is critical for ensuring accessibility and inclusion for deaf individuals. As these systems are intended to be used by people with varying levels of technical expertise, their design must prioritize simplicity, intuitiveness, and ease of use [17]. This is particularly true in health care settings, where effective communication with medical staff is essential. Furthermore, given the sensitive nature of health care interactions, it is crucial to assess how these systems address ethical concerns, such as patient data privacy, and practical challenges related to accessibility, including cost and implementation feasibility in diverse health care settings.

Therefore, this study aims to systematically review the evidence on the translation systems developed for deaf people who communicate through sign language with hearing health professionals in a health care context, which are already in use or proposed for use and have been tested with human users or videos of human users. The main research question was: What technologies have been developed and tested in real-world settings to translate sign and oral languages, facilitating communication between deaf patients who primarily use sign language and health care workers? The specific questions are as follows:

- In which context of health care have these technologies been used?
- Which languages (sign and oral) can these technologies translate?
- Which technologies are required for it to be used on-site? How were they developed?
- How were they deployed and tested?
- How has the communication between health care workers and deaf people been improved by using these technologies?
- How was the efficacy of these technologies evaluated?
- Is the system or technology “bidirectionally” interactive?
- How do these systems address ethical concerns in health care settings, such as patient privacy and data security?

XSL•FO
RenderX

Methods

Overview

The research protocol was registered in the Open Science Framework and previously published in detail [18,19]. It followed guidance from the Cochrane Guidelines and the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [20,21] (Checklist 1). A multidisciplinary team comprising researchers from health and computing domains, along with linguistic specialists in sign language, collaboratively conducted the systematic review. Two members of the team are linguistic specialists and sign language researchers, one of them is deaf.

Search Strategy

Independent researchers performed a literature search using Web of Science, MEDLINE, IEEE Xplore, ACM, Scopus, and Google Scholar. A preliminary search strategy, developed by 5 authors (MSM, LFRO, LRV, ROP, and ZSNR), incorporated MeSH and defined text words relevant to the topic (Multimedia Appendix 1). The last search was conducted on March 27, 2025.

All studies, regardless of publication date or language, from inception onward were considered. In the case of unpublished studies, the authors were contacted at least 3 times via email or other social networks used by researchers to request additional information. Reference lists of eligible studies were examined to identify additional eligible studies.

Study Selection

Prospective, retrospective, or descriptive studies that address the development of communication systems specifically designed for deaf individuals in health care encounters and that involve testing with human users or videos of human users were included. Human users could be people of any age who are deaf and whose primary communication modality is sign language.

Studies that do not address the specified research questions, do not mention testing with human users or videos of human users, or do not mention use in health care encounter contexts were excluded. Short communications, conference abstracts, and correspondences were not excluded.

Independent researchers blindly screened the studies. Titles and abstracts of identified studies were individually reviewed to assess eligibility. Full-text versions of papers that were not excluded at this initial stage were read for a thorough examination. Subsequently, potentially pertinent studies were independently evaluated to ascertain if they aligned with the inclusion criteria. Any disagreements were resolved by a senior researcher (MSM). Whenever necessary, corresponding authors were contacted to obtain data not included in the publication using email and ResearchGate.

Search data for the identified studies and information for each stage of study selection were registered in detail, following the guidelines of the PRISMA methodology [20,21].

Data Extraction

A data extraction table was custom-designed for this study and independently piloted by 2 researchers, as well as data extraction. The extraction was checked by 2 other reviewers, one researcher with a computer science background and a senior researcher with a health science background. Conflicts were resolved by consensus or by consulting a senior researcher (MSM). Details on the variables extracted were previously published [18,19].

Furthermore, the researchers responsible for each study included in the systematic review were contacted via email (with up to 3 attempts) and ResearchGate (at least 1 attempt) to obtain updates on the current status of their systems and to request any additional information regarding their application in real-world contexts.

The definitions used regarding the type of SLR system, the corpus formation, and the health context are available in Multimedia Appendix 2. To extract data from papers that met the eligibility criteria, the authors of the study developed a codebook with clear definitions for all variables to ensure consistent data collection (Multimedia Appendix 3).

Data Analysis

A qualitative synthesis was performed to analyze the data, and a narrative synthesis of the evidence was conducted to provide an overview of the results. The results are summarized according to system types: image-based and sensor-based.

Results

Search Results and Study Selection

The search retrieved 21,778 publications, 2021 of which were duplicates. Most studies were excluded after title and abstract analysis (n=19,605). In total, 6 reports were unavailable, of which 4 did not respond despite repetitive contact attempts with corresponding authors, and 2 did not make any contact details available (Multimedia Appendix 4), thus leaving 146 selected for full-text screening. Of those, 22 studies were selected for inclusion after applying the eligibility criteria. In 2 cases, publications by the same research groups were identified. The first case involved a deep learning–based system for recognizing emergency gestures in Indian Sign Language to support communication with hearing-impaired individuals [22,23]. The journal paper published in IEEE Access in 2022 represents an extended and more comprehensive version of the earlier conference paper. It includes additional models (such as 3D convolutional neural network [CNN] and You Only Look Once [YOLO] v5), a more detailed methodology, a larger dataset, and more robust performance metrics, including mean average precision, precision, and recall. Given its methodological completeness and broader evaluation, only the IEEE Access paper was included in this review, while the conference version was excluded to avoid duplication of data and analysis [22].

The second case involved 2 publications from a Brazilian research group describing a system for recognizing Brazilian Sign Language (Libras) in the health context [15,24]. The 2024 journal paper incorporates a more advanced architecture (multiple-stream versus 2-stream), enhanced performance results, and a more detailed methodology, while using the same dataset as the earlier work. Thus, only the 2024 version was included [15].

XSL•FO
RenderX

Additionally, 2 studies were obtained from reference list screening, totaling 23 studies ultimately included in the review (Figure 1) [12-15,22,25-42].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart [21].

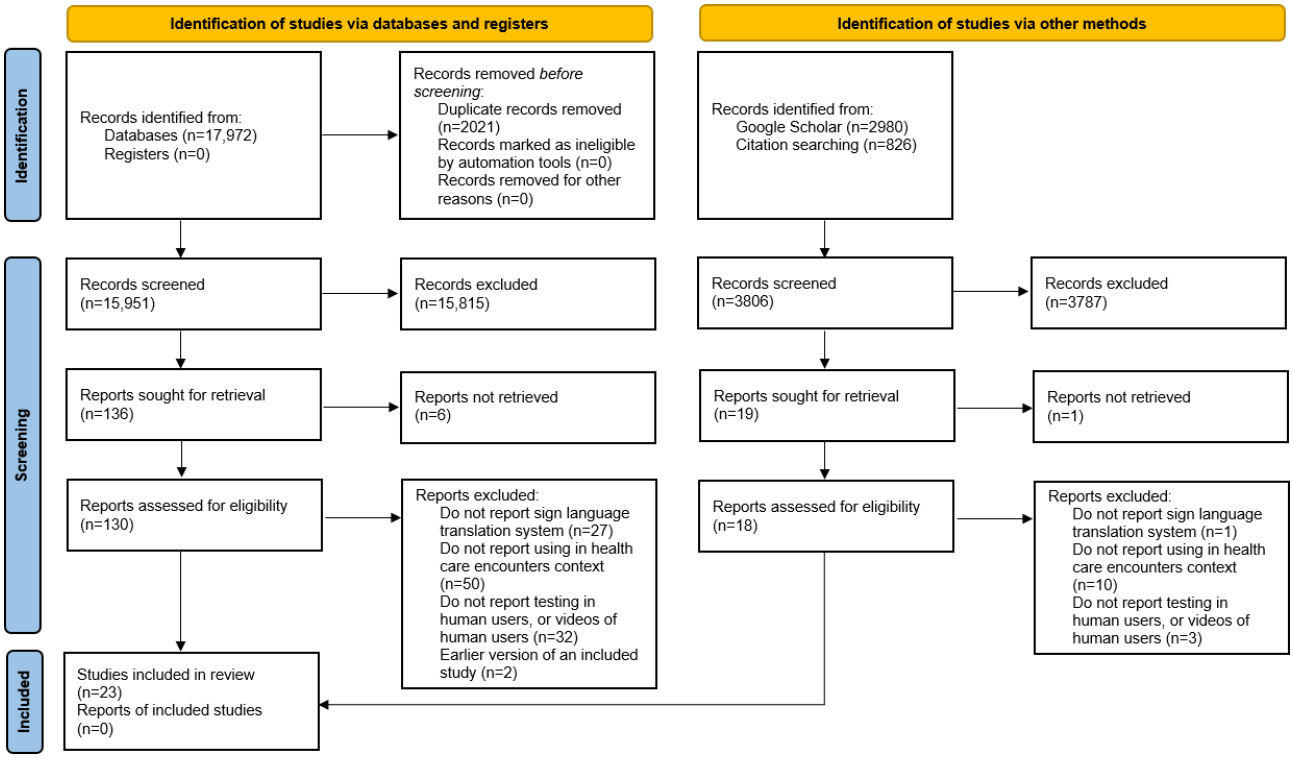

Characteristics of Included Studies

The main characteristics of the included studies are summarized in Tables 1 and 2 and Multimedia Appendix 5 [12-15,22,25-42]. They were published in 18 journals and conferences from 2015 to 2024. All of them were published in English. In 15 studies, the system was classified as image-based [13-15,22,25-28,30,31,36-39,42], and in 7, it was sensor-based. Of these, 4 were depth-sensing [29,33,35,41], and 3 were glove-based [32,34,40].

XSL•FO
RenderX

Table . Main architectural, functional, and technical characteristics of sign language recognition systems.

	Languages involved (oral and sign)	Technology needed		Infrastructure needed
		Hardware or software	Development technology[a]	
Image-based				
Xia et al (2022) [13]	Chinese and Chinese Sign Language	Kendryte K210, 8-megapixel camera, microphone arrays, operating system FreeRTOS, LabelImg, Autodesk Inventor	TensorFlow, Keras, Darknet	SLA[b] 3D printing technology and heart-speaker device
Pikoulis et al (2022) [14]	Greek and Greek Sign Language	No information	Sentence-BERT[c], Google Mediapipe	Computer and camera
da Silva et al (2024) [15]	Brazilian Portuguese and Brazilian Sign Language (Libras)	Intel Core i3 and all experiments were performed on Ubuntu 18.04 LTS[d]	LSTM[e] network, OpenPose, Keras, cuDNN[f] 8.5	Computer and smartphone HD[g] camera
Das et al (2023) [25]	English and Indian Sign Language	No information	CNN[h], BiLSTM[i]	Computer and camera
Ko et al (2019) [26]	Korean and Korean Sign Language	No information	OpenPose, PyTorch	Computer and camera
Barrientos-Villalta et al (2022) [27]	Peruvian Spanish and Peruvian Sign Language	Google Cloud, Storage, Google Cloud Run	Google Mediapipe, LSTM neural network	Mobile devices and internet
Ramírez Sánchez et al (2021) [28]	Spanish and Mexican Sign Language	Webcam	Google MediaPipe, OpenCV, CNN, HMM[j], Viterbi algorithm	Webcam and computer
Gandhi et al (2021) [30]	English and Indian Sign Language	Mobile phones with their camera resolutions	CNN using Image Stacking (VGG-16[k] and Resnet50) CNN + LSTM and LSTM with OpenPose	Mobile phone and camera
Uchil et al (2019) [31]	English and Indian Sign Language	Intel Core i5 CPU[l] running macOS Mojave and Core i3 CPU running Windows 7, smartphone	OpenPose library, OpenCV	Mobile phone camera
Areeb et al (2022) [22]	English and Indian Sign Language	No information	3D CNN, pretrained VGG-16, LSTM (RNN-LSTM)[m] scheme, YOLO[n] v5	Computer and camera
Adithya and Rajesh (2020) [36]	English and Indian Sign Language	Digital camera	LSTM network	Computer and camera
Ihsan et al (2024) [37]	English and American Sign Language	Intel Core i5, Visual Studio Code	Pretrained CNN (MobileNetV2) model, BiLSTM model, Mediapipe, TensorFlow, Keras	Computer, camera, and smartphones
Das et al (2024) [38]	English and Indian Sign Language	No information	CNNs, LSTM, long-term recurrent convolutional network model	Computer and mobile phone camera
Faisal et al (2023) [39]	Arabic and Saudi Sign Language	Robot Operating System, IClone Pro, 3DXchange, Unity	Self-developed concise 3D graph convolutional network, time delay neural network model, DTW[o] algorithm, FastSpeech2, Kaldi toolkit, MediaPipe	Portable electronic devices or computers equipped with a camera
Bellil et al (2024) [42]	Arabic and Algerian Sign Language	iPhone XS MAX	1D-CNN, MediaPipe	Computer and smartphone
Sensor-based (depth-sensing)				

XSL•FO
RenderX

	Languages involved (oral and sign)	Technology needed		Infrastructure needed
		Hardware or software	Development technology[a]	
Hisham and Hamouda (2019) [29]	Egyptian Arabic and Egyptian Arabic Sign Language	Kinect SDK, Windows 7, Visual Basic.NET, C# and C++	Bayesian network, AdaBoosting, DTW, and HMM	Computer and Microsoft Kinect sensor
Sarhan et al (2015) [33]	Arabic and Arabic Sign Language	No information	HMM	Computer and Microsoft Kinect sensor
Süzgün et al (2015) [35]	English and Turkish Sign Language	Personal computer and a touch display	DTW algorithm	Computer and Microsoft Kinect sensor
Dewasurendra et al (2020) [41]	Sinhala and Sri Lankan Sign Language	Webcam, mobile device with camera, microphone, GPS	EfficientNet-Lite0, Pillow (Python image processing library), MaryTTS Framework, CMU[p] Sphinx 4 toolkit (Sphinx4+ SphinxTrain) and natural language processing, TensorFlow	Kinect and Leap Motion Controller
Sensor-based (glove-based)				
Deji Dere et al (2022) [32]	English and American Sign Language	Arduino Nano 33 BLE, Edge Impulse software	1D-CNN model	Inertial Measurement Unit
Guo et al (2023) [34]	English and American Sign Language	INMO AIR with Android 10 Go	Android-MediaPipe, VOSK[q] library	Mobile phone and augmented reality glasses[r]
Luqman and Mahmoud (2020) [40]	Arabic and Arabic Sign Language	No information	KenLM (statistical language model), MADAMIRA[s] (morphological analyzer and disambiguation tool)	Computer and camera
Hybrid				
Sosa-Jiménez et al (2022) [12]	Mexican Spanish and Mexican Sign Language	Intel Core i7 running Windows 7, Intel Core i5 running Windows 8.1, Microsoft Visual Studio 2015	Markov models (probabilistic) and neural networks	Computer and Microsoft Kinect sensor

[a]Artificial intelligence and imaging processing.

[b]SLA: stereolithography.

[c]BERT: bidirectional encoder representations from transformers.

[d]LTS: long-term support.

[e]LSTM: long short-term memory.

[f]cuDNN: CUDA deep neural network library.

[g]HD: high definition.

[h]CNN: convolutional neural network.

[i]BiLSTM: bidirectional long short-term memory.

[j]HMM: hidden Markov model.

[k]VGG-16: visual geometry group 16 layers.

[l]CPU: central processing unit.

[m]RNN-LSTM: recurrent neural network with a long short-term memory.

[n]YOLO: You Only Look Once.

[o]DTW: dynamic time warping.

[p]CMU: speech recognition toolkit.

[q]VOSK: offline speech recognition toolkit [43].

[r]Mobile phones and augmented reality glasses: portable electronic devices that enable interactive digital experiences.

[s]MADAMIRA: a system for morphological analysis and disambiguation of Arabic [44].

XSL•FO
RenderX

Table . Main characteristics of the recognition approaches of sign language recognition systems.

	Corpus formation			Captures facial expressions and body movement	Health context
	Type	Isolated words	Sentences		
Image-based					
Xia et al (2022) [13]	Isolated words	19	0	N/A[a]	General hospital care
Pikoulis et al (2022) [14]	Isolated words and sentences	6319 words with 1374 of them being unique (excluding repetitions)	1029 simple sentences with 945 of them being unique (excluding repetitions).	N/A	Psychiatric interviews
da Silva et al (2024) [15]	Isolated words	50	0	Yes	General hospital care
Das et al (2023) [25]	Isolated words	8	0	N/A	Emergency situations
Ko et al (2019) [26]	Isolated words and sentences	419	105	Yes	Emergency situations
Barrientos-Villalta et al (2022) [27]	Isolated words	17	0	Yes	General hospital care
Ramírez Sánchez et al (2021) [28]	Isolated words and sentences	49	20	Yes	Primary care consultations
Gandhi et al (2021) [30]	Isolated words	20	0	Yes	General hospital care
Uchil et al (2020) [31]	Isolated words	20	0	Yes	General hospital care
Areeb et al (2022) [22]	Isolated words	8	0	N/A	Emergency situations
Adithya and Rajesh (2020) [36]	Isolated words	8	0	No	Emergency situations
Ihsan et al (2024) [37]	Isolated words	30	0	N/A	General hospital care
Das et al (2024) [38]	Isolated words	6	0	N/A	Primary care consultations
Faisal et al (2023) [39]	Isolated words	293	0	N/A	Primary care consultations
Bellil et al (2024) [42]	Isolated words	10	0	No	General hospital care
Sensor-based (depth-sensing)					
Hisham and Hamouda (2019) [29]	Isolated words	42	0	No	General hospital care
Sarhan et al (2015) [33]	Isolated words	16	0	No	Emergency consultations
Süzgün et al (2015) [35]	Isolated words	33	0	N/A	General hospital care
Dewasurendra et al (2020) [41]	Isolated words	0	N/A	No	Emergency situations
Sensor-based (glove-based)					
Deji Dere et al (2022) [32]	Isolated words	5	0	No	Emergency situations
Guo et al (2023) [34]	Isolated words	More than 550	0	N/A	Emergency situations
Luqman and Mahmoud (2020) [40]	Isolated words and sentences	3327 sign words and 30,296 singular words with their plurals	600	Yes	General hospital care

XSL•FO
RenderX

	Corpus formation			Captures facial expressions and body movement	Health context
	Type	Isolated words	Sentences		
Hybrid					
Sosa-Jiménez et al (2022) [12]	Isolated words	43	0	Yes	Primary care consultations

[a]N/A: not applicable.

The system developed by Sosa-Jiménez et al [12] is the only one classified as hybrid, as it leverages Kinect's red, green, blue, and depth sensors to recognize signs through both visual input and 3D skeletal tracking. Because it is the only hybrid system, it is presented separately in Table 1 and Multimedia Appendix 5; however, in the following sections, it will be discussed alongside the sensor-based systems (depth-sensing).

Aspects Related to System Development

Languages Involved, Direction of Communication, and Capturing of Facial Expressions

Image-Based

The systems studied supported a range of spoken-sign language pairs, including Arabic (Algerian and Saudi), English, Chinese, Brazilian, Portuguese, Greek, Indian, Korean, Peruvian, and Mexican Spanish, and their respective sign languages (Table 2). Only 3 papers successfully implemented a bidirectional communication between spoken and sign languages [15,35,39], while another study attempted bidirectionality but did not achieve it [13].

Only 2 studies explicitly addressed structural and cultural differences between spoken and sign languages in their translation processes [15,28]. Additionally, 7 systems incorporated the recognition of facial expressions, facial key points, and other nonmanual features (such as audio phonemes and body posture) to enhance translation accuracy and communication effectiveness [15,26-28,30,31,39].

Sensor-Based

These systems also supported multiple spoken-sign language pairs, including combinations involving Arabic (and its Egyptian variant), English, Turkish, Sinhala, and Mexican Spanish, along with their respective national sign languages (Table 1).

In total, 3 studies reported successful implementation of bidirectional communication between spoken and sign languages [12,34,41]. One study took linguistic differences into account by applying rule-based morphological and syntactic processing to accommodate features specific to Arabic Sign Language—such as subject-initial word order, lack of inflection, and the use of spatial and visual cues—thus enabling transformation into grammatically correct Arabic sentences [40].

Two sensor-based systems also linked corpus construction to the use of facial expression and body movement capture for more precise recognition [12,32]. As anticipated, no glove-based system used such an approach.

Corpus Used to Generate the Language Database for System Development

Image-Based

Overall, most systems relied on corpora composed of isolated words or terms, except for 3 studies, which included full sentences [14,26,28]. Two studies defined their corpus based on real-world contexts [15,26], while others relied on datasets designed by the authors [13,30], consulted health professionals [14], or extracted content from dictionaries [31]. Four systems focused on essential or emergency vocabulary, often including the alphabet, numbers, and isolated signs relevant to health care, obtained through internet searches or expert input [22,25,36].

Sensor-Based

One glove-based system developed its corpus through consultations with health professionals and internet searches [32], while another was based on real emergency call transcripts [34]. A third glove-based system included an extensive set of 600 health-related sentences covering various types (nominal, verbal, and questions) with over 3000 signs [40]. A fourth one focused on basic signs only (fingerspelling the alphabet, numbers, and isolated sign words) [32].

Among depth-sensing systems, 3 built their corpora using isolated words or terms [29,33], while others developed their corpora in consultation with health care professionals [12,35] or constructed its corpus using full sentences commonly used during emergency calls [41].

Health Care Context

Image-Based

Among the 15 studies in this group, 46.7% (7/15) were based on the general hospital environment [13,15,27,30,31,37,42], 36.7% (4/15) focused on emergency services or situations [22,25,26,36], 20% (3/15) were based on primary care consultations [28,38,39], and 6.7% (1/15) on psychiatric interviews [14].

Sensor-Based

In total, 2 of the glove-based systems used emergency services or situations as a basis [32,34], and the remaining one used the general hospital environment [40], while 2 depth-sensing systems were based on care provided in a general hospital environment [29,35], 1 used the context of primary care consultations [12], and 2 used emergency services [33,41].

XSL•FO
RenderX

Approaches and Techniques for System Development

Image-Based

In the image-based systems reviewed, different datasets have been created, involving capturing images in many lighting conditions and backgrounds, recording of signs made by users fluent in sign languages, extracting keyframes from videos [15,22,25-28,33-38], and image stacking [30]. These datasets have been divided into training, validation, and test sets for developing and evaluating the proposed systems [13,14].

Data processing included methods such as hierarchical classification, feature extraction through neural networks, human key point estimation, and pose, hand, and skeleton coordinates of people in videos [14,15,25,27]. Different models were used, such as CNNs, long short-term memory, hidden Markov model (HMM), OpenPose, 3D graph convolutional network, long-term recurrent convolutional network, 1D-CNN, and bidirectional long short-term memory, both individually and in combination, for the classification and prediction of sign language signals [15,22,26,27,30,37-39,42].

Additionally, only 1 paper presents an avatar module that transforms a hearing person's text into sign language [39]. The system comprises 3 modules: a signal recognition module, a speech recognition and synthesis module, and an avatar module. Each module performs specific tasks to ensure the integrated functioning of the system.

Sensor-Based

In sensor-based systems, development methods involved capturing sign language signs with a Kinect, Leap Motion Controller, and cameras on wearable devices such as smart glasses [32-35,41]. Then, the captured data were processed and analyzed to extract relevant characteristics, such as position, movement, and shape of the hand; wrist trajectory; and other specific features of the sign [32,34].

Different techniques have been applied in the development of these systems, including machine learning models such as random forest, naïve Bayes classifier, Ada-Boosting, dynamic time warping, and HMMs [12,29,35]. Furthermore, preprocessing, segmentation, and feature extraction techniques have also been used to improve the effectiveness and accuracy of the systems in detecting and interpreting signs [32-34].

One of the systems also integrated machine translation processes by applying morphological and syntactic analysis to restructure sentence structure and ensure grammatical agreement [40].

Technologies Involved

The technologies involved in system development are shown in Multimedia Appendix 6.

The Infrastructure and Training Required for Use

Image-Based

Desktop and notebook computers, cameras, sensors, and supporting devices and equipment were needed to implement 3 proposed systems [25,27,36,42], whereas a mobile phone was required for 4 of them [27,30,31,42]. Beyond internet connectivity, 1 proposal [29] required downloading apps. In the image-based group, 2 systems explicitly mentioned using a phone's camera [30,31], and 3 mentioned using a webcam or external camera [13,36,37]. None of the studies addressed the details of the training required for the translation system. In total, 3 papers failed to address the structure and training required for use [14,22,26].

Sensor-Based

Three systems, all depth sensing [29,33,35], addressed the need for a computer connected to a power source. Only 1 depth-sensing system pointed out the need for a mobile phone with an integrated camera [41]. One glove-based [34] proposal required downloading an app, and this was the only one that described the use of a phone's camera, additionally requiring augmented reality glasses with a Bluetooth interface. Another one, from the depth-sensing type [33], mentioned the existence of an external camera device. Microsoft Kinect is a standard device for sensor-based systems that uses depth sensing [45]. None of the studies addressed the details of the training required for the translation system. One paper did not discuss the necessary structure and training for implementation [40].

Multidisciplinary Partnerships

Image-Based

Various disciplines were represented among the research teams, who reported back to university departments of informatics [14,30], applied sciences [27,42], engineering [13,14,30,37,38], technology [26,30,31,37], computer science [22,38,39,42], and biomedical physics [37]. Only 4 studies mentioned the participation of linguistic specialists [13,15,26,32] during corpus development or database recording to improve sign language representation, and only 1 mentioned having involved physicians [14] in selecting the words to create the corpus. Along with sign language interpreters, hearing-impaired native sign language speakers [11,14,26] were engaged to record the videos.

Sensor-Based

The sensor-based systems encompassed a great variety of backgrounds, including university departments of statistics [12], AI [12,32], informatics [12,41], applied sciences [12], engineering [29,33,35], technology [33,41], and computer science [32,34,40]. To contribute to the enhancement of sign language representation, some systems involved linguistic specialists [35] in the development of a corpus or database, and also deaf native sign language speakers [32,34] and sign language interpreters worked on recording the videos. Physicians [12,32,35] and deaf individuals [35] were also involved in selecting the words for the corpus.

Sample Size (Sample Size of Individuals to Develop the System, Users, and Content)

Image-Based

The sample size (number of individuals) used to develop the systems ranged from 2 [31] to 26 [22,25,36] among systems that used the same people for development and testing, and 1 study did not inform the sample size used in the development phase [38]. For systems that used different individuals for development and testing, the number used in the development

XSL•FO
RenderX

phase ranged from 14 [26] to 33 [39], and the number of individuals used in the test phase ranged from 2 [39] to 6 [26]. Four studies [27,28,30,42] did not mention whether they used the same individuals for both development and testing. The dataset's contents ranged from 21 scripts [14] to 145,035 videos (Table 2) [39].

Sensor-Based

The sample size (number of individuals) used to develop the systems ranged from 2 [32,40] to 6 [34] among systems that used the same people for development and testing. For systems that used different people for development and testing, the number of people to develop ranged from 2 [40] to 12 [12] (2 studies [33,46] did not inform the sample size in development), and the number of users to test ranged from 2 [39] to 10 [29].

The content of the datasets ranged from 10 videos (5 for the training phase and 5 for the testing phase) [32] to 1260 samples (840 samples for the training set and 420 samples for the testing set) [29]. In total, 2 studies did not inform the size of the content of their datasets (Table 2) [40,41].

Cost

Image-Based

None of the included studies provided quantitative data on the costs associated with the development of their SLR systems, either in terms of technological infrastructure or human resources.

Sensor-Based

As with image-based studies, no quantitative analysis has been reported regarding the costs for developing the SLR systems. However, one of the studies mentioned that the development cost of the device "SmartCall" was much less when compared to those of similar works (which used sensors embedded in smartwatches), suggesting that researchers in low- and middle-income countries could readily prototype their own devices [32].

Testing

Image-Based

The tests have been conducted in simulated environments involving different training networks, varying noise levels, and different scenarios [13,14]. Data splitting strategies into training and testing sets [15,22,25,36-39] and comparing different methods [18] have been used to evaluate the accuracy and effectiveness of the developed systems in detecting and interpreting signs.

Accuracy measures in detecting and interpreting signs ranged from 25% [27] to 100% [42]. Precision ranged from 91.5% [25] to 100% [42] (5 systems) [22,24,25,37,42]. Recall ranged from 90.1% [25] to 100% [42] (4 systems) [24,25,37,42]. F_1-score ranged from 90.7% [25] to 100% [42] (4 systems) [24,25,37,42]. No study presented specificity and sensitivity (Multimedia Appendix 5).

A common difficulty was the failure of some systems to predict similar hand movements, given the proximity of the key points [22,30,31]. For that reason, double-handed signs were more accurately classified than single-handed ones [22].

Sensor-Based

The tests were conducted using datasets, in which different users performed sign language signs repeatedly in different scenarios [12,32,35]. Testing also involved splitting data into training, validation, and test sets and evaluating the performance of the systems [29,32,33]. Additionally, usability and user acceptance tests have been conducted to assess the effectiveness and practicality of the systems in real-life situations [34,35].

Accuracy measures in detecting and interpreting signs ranged from 80.5% [33] to 99.8% [41] (8 systems) [12,29,32-35,40,41], being 80.5% [33] to 99.8% [41] for systems with depth-sensing sensors and 72% [32] to 92% [40] for systems with hand glove sensors. Only 1 study [12] presented precision (">90%"). Recall ranged from 90.1% [25] to 98.8% [24] (3 systems) [24,25,37]. F_1-score ranged from 85% [32] to 88.6% [12] (2 systems) [12,32]. Only 1 study [12] presented specificity and sensitivity (99.8% and 87.9%, respectively; Multimedia Appendix 5).

Evaluation or Use in Real Context

Image-Based

Among image-based studies, authors of 11 of the studies [13-15,22,25-28,31,36,38,39] did not respond to the emails sent by the research team, so it was impossible to update the current status of their systems. Authors of 2 of the studies informed that their systems have not been implemented in real scenarios or commercial settings [30,42]. One author mentioned that his team is in the process of developing a smartphone app for interpreting medical signs and that they will proceed to real-world implementation once it is ready [37].

Sensor-Based

Among sensor-based studies, the authors of 3 of the studies [29,33,41] did not respond to the emails sent by the research team, so it was not possible to update the current status of their systems. Authors of 4 other studies informed that their systems have never been deployed or commercialized [12,34,35,40]. Authors of the other study announced that their system is currently undergoing major design improvements and patent applications and that they hope to make it commercial very soon [32]. Authors of one of the studies reported that they open-sourced their dataset and their source code implementation to facilitate adoption of their system and accelerate technology transfer among research groups [32].

Emergency Responses Capability

Regarding emergency circumstances, 13 studies [12-15,22,25,26,28,30,32,33,35,36] analyzed the capabilities of the system to respond in urgent situations according to various parameters.

Image-Based

Several image-based systems assessed key performance indicators relevant to urgent health care contexts, including recognition speed for signs and phrases [12,13], the system's ability to reliably alert health care providers [12,14,34], and the clarity of communication output [12,26]. Some studies also

XSL•FO
RenderX

addressed design considerations critical for emergency use, such as error tolerance (the system's resilience to technical failures) [25], and the inclusion of medical vocabulary specific to emergency scenarios [30]. One system also incorporated structured protocols aimed at guiding interactions with deaf patients during emergencies [15]. However, a common limitation was the small size of the vocabulary recognized, which restricted the ability to handle diverse emergency dialogues [22,36].

Sensor-Based

Sensor-based systems also demonstrated potential for real-time emergency support. For instance, one system was specifically designed to improve access to emergency services through depth-sensing technology [41], while another emphasized robustness through error-tolerant functionality [12]. Similar to image-based tools, sensor-based systems faced limitations related to the restricted number of medical terms recognized, which could hinder effective communication during critical situations [32,34].

Technological Characteristics

Image-Based

None of the studies reported on critical aspects such as system reliability (graceful degradation or recovery after crashes), data security, video storage, or deletion policies after translation, or objective metrics comparing communication outcomes with and without the system. Translation time was superficially addressed in 1 paper [22]. Authors compared 3 different models (a 3D CNN; a pretrained visual geometry group 16 layers and a recurrent neural network with a long short-term memory scheme; and a model based on YOLO v5, an advanced object detection algorithm). When comparing the proposed techniques, the YOLO-based model was faster, with an image processing rate of 40 - 90 frames per second. Therefore, it could be used for emergency sign recognition with only a few milliseconds of delay.

Sensor-Based

Only 3 studies addressed translation time in sensor-based systems [29,34,41]. One study compared different machine learning classifiers and reported the following processing times: 8.09 seconds for the naïve Bayes classifier, 15.3 seconds for random forest, 17.63 seconds for dynamic time warping, and 20.8 seconds for HMM [29]. Another system reported an average latency of 0.55 seconds for translating 550 signs, with mobile translation taking 122 milliseconds and phoneme streaming 206 milliseconds to render the animation [34]. The third study reported that the system required 3 to 4 seconds from the end of the gesture to generate the output. In test scenarios, it consistently outperformed Google Cloud, with response times of 2.5, 2.6, and 2.2 seconds, compared to 4.1, 4.2, and 3.7 seconds for Google Cloud [41].

As with image-based systems, none of the sensor-based systems discussed reliability, data privacy, postprocessing video storage, or communication effectiveness metrics.

User Experience

Image-Based

User experience was not explicitly evaluated in any of the image-based studies. A few of them commented on some factors related to the usability of the developed systems, both about their facilities and difficulties. These feedbacks are listed in Multimedia Appendix 5.

Sensor-Based

The evaluation of users' experience was subjectively reported in 4 sensor-based studies [34,35,40,41]. The first one, classified as glove-based, conducted a user study on the signers' experience. The study involved 12 participants, and Quality of Experience was rated on a 5-point scale, ranging from 0=worst evaluation)to 5=best evaluation, to assess the usefulness of the system. The obtained results were 4.2 for accessibility, 4.3 for usability, and 4.6 for overall experience [34].

Additionally, a second study, classified as depth-sensing, developed a questionnaire score using the same scale for 5 statements. In total, 5 participants used the system, and afterward, responded to the questionnaire. The questionnaire results obtained for each statement question were 4.85 for "I was able to express my situation using HospiSign," 4.85 for "HospiSign is easy to use," 5.00 for "Most people could learn to use HospiSign quickly," 4.45 for "Options were sufficiently clear," and 5.00 for "HospiSign would help the hearing impaired" [35].

Another study, from the glove-based category, provided a manual evaluation performed by 3 native Arabic speakers. Each of the 600 sentences was classified by them as understandable, somehow understandable, and not understandable according to grammatical and semantic metrics. The respective results were 80%, 12% and 8%, demonstrating an acceptable translation provision in 92% of the sample [40].

Additionally, a third study, classified as depth-sensing, conducted a user satisfaction survey with 10 users, and the results were classified as high, medium, and low satisfaction. The results reported were user registration: high, text-to-call: high, call-to-sign: medium, sign-to-call: medium, and prerecorded message: high [41].

Finally, a fourth study, classified as depth-sensing as well, developed a questionnaire score using the same scale for 5 statements. In total, 5 participants used the system, and afterward, responded to the questionnaire, and the results obtained for each statement were 4.85 for "I was able to express my situation using HospiSign," 4.85 for "HospiSign is easy to use," 5.00 for "Most people could learn to use HospiSign quickly," 4.45 for "Options were sufficiently clear," and 5.00 for "HospiSign would help the hearing impaired" [35].

There were also some comments on the usability of the developed systems, concerning the user experience. These feedbacks are listed in Multimedia Appendix 5.

Ethical Issues Addressed

Ethical concerns related to the use of communication systems for deaf and hard-of-hearing individuals were largely

XSL•FO
RenderX

underexplored across the included studies. Importantly, none of the included studies discussed broader ethical implications of the systems themselves, such as data privacy and security, the handling and potential storage of video or biometric information, or the autonomy and psychological impact of using these technologies. Despite their role in sensitive health care contexts, no system reported design strategies for ensuring user dignity, minimizing potential emotional distress, or addressing users' preferences regarding data management. Additionally, none of the studies evaluated whether addressing ethical aspects influenced system usability or user trust, including transparency, control over personal data, and informed decision-making.

Discussion

Principal Findings

The development of SLR systems represents a promising step toward improving communication between deaf individuals and health care providers. Despite the vast literature in SLR systems, many computer science studies focus on creating algorithms for recognizing (and less often translating) signed content [1]. These teams often lack deaf individuals with firsthand experience of the challenges the technology aims to address, and they may not fully understand the linguistic complexities of the language. Additionally, the algorithms are typically trained on datasets that do not reflect real-world scenarios, making these 1D approaches to sign language processing of limited practical value [1]. In our analysis, the comprehensive search resulted in over 21,000 papers, but only 23 studies met our inclusion criteria. All analyzed systems are in the development and testing stage, with no real application yet, and this review highlights significant challenges that hinder the implementation of these systems in real-world settings. These challenges include variability in methodologies, limitations in dataset quality, and the underrepresentation of key communication elements, such as bidirectionality and facial expressions. Furthermore, the absence of cost analysis and the lack of multilingual integration reflect gaps in the current body of research. To address these issues, it is essential to evaluate both the technical aspects and the practical implications of SLR systems, aiming for inclusive and effective solutions.

Image-based systems accounted for the majority of the studies. This predominance highlights the versatility of image-based systems, which capture facial expressions, head movements, lip reading, and hand gestures using simple devices like mobile phone cameras [47]. In contrast, it is possible to infer that sensor-based systems, which rely on advanced technologies such as depth-sensing cameras and sensor-embedded gloves, incur higher costs [48]. However, a critical gap identified across studies was the lack of clear cost-related information. Only 1 study briefly mentioned that its device, "SmartCall," was more affordable compared to similar solutions, suggesting its feasibility for low- and middle-income countries [32].

Given the lack of sufficient cost-related data in the studies reviewed, it would be challenging to provide accurate inferences regarding the cost-effectiveness of these technologies. Furthermore, as costs vary significantly depending on the country and context, it would be difficult to generalize these findings across all regions. This is a notable limitation, and we believe that future studies should address this gap by providing more detailed and context-specific cost analyses.

Corpus Definition and Number of Samples

Corpus definition and dataset size were highly heterogeneous across the reviewed studies, reflecting diverse development contexts and objectives. Most systems relied on isolated words, limiting applicability in real-world health care communication, while some incorporated phrases or real clinical contexts, occasionally guided by health professionals or dictionaries [14,26,28,40]. A few image-based systems also integrated facial expressions and body movements, enhancing linguistic representation [15,26-28,30,31,39]. Among sensor-based systems, corpus construction lacked multimodal features, particularly in glove-based systems, which are limited in capturing facial and body expressions.

Sample sizes ranged widely from as few as 10 to over 145,000 videos, raising concerns about the adequacy of data for training reliable models. This lack of standardization in corpus design and dataset size limits the comparability, scalability, and generalizability of SLR systems in clinical settings. Establishing minimum standards for corpus development could improve system robustness and support broader implementation in health care.

Key Aspects of Effective Communication

Bidirectionality and facial expression recognition are 2 critical components for effective communication, but were not included in the majority of publications. Most systems were unidirectional, translating only from sign language to written text, impairing its applicability in real-world contexts. While such systems may enable health care practitioners to understand what a deaf patient is communicating, they fail to facilitate communication in the opposite direction, thereby undermining true interactive dialogue.

Facial expression recognition is a vital element for capturing linguistic prosody, and although face landmarks were considered in 8 papers [15,26-28,30,31,39,40], only 1 of them explicitly addressed it as an indicator of grammatical tense [28]. This omission hinders the natural and nuanced communication required in health care contexts.

Notably, only 3 studies explicitly acknowledged the cultural and structural distinctions between sign languages and oral languages [15,28,40]. This omission is significant, as sign languages are not mere gestural representations of spoken language, but complete linguistic systems with their own grammar, syntax, and cultural context. Failure to consider these differences can lead to inaccurate or overly literal translations, potentially compromising communication quality and user trust in assistive systems. The limited attention to these linguistic specificities suggests that many of the reviewed systems may have been developed from a technical standpoint, with insufficient involvement of deaf communities or sign language experts.

Given the numerous challenges involved, it is nearly impossible to develop SLR tools that achieve 100% accuracy with a large

XSL•FO
RenderX

vocabulary [49]. For example, the same sign can have large changes in shape when it is in different locations in the sentence [50]. In the studies analyzed, only 1 demonstrated perfect accuracy (100%), but this was based on an image-based system, which was tested only with isolated words [12]. Studies that involved a relevant number of sentences and words tended to perform worse, with accuracy sometimes too low to enable effective communication [14,26].

Multilingual Challenges

Sign languages are not universal, and they are not mutually intelligible, and they can present significant differences between countries, and sometimes even within the same country [51]. While some studies proposed extending their systems to multiple languages [12,28,41], none successfully implemented multilingual translation. The inherent diversity of sign languages—each with unique grammar, vocabulary, and body configurations—poses substantial challenges to creating universal systems [27,52,53]. Furthermore, there was a predominance of English corpora [25,30-32,34,36,37]. This is probably driven by high research investments, as well as the global reach and accessibility of English [54], which not only facilitates better comprehension within the medical academic community but also enables broader dissemination and open sharing of the corpus [55]. This indicates a gap in accessibility for non-English–speaking regions, emphasizing the importance of expanding datasets to include diverse languages and cultures. To address this gap, it is crucial to foster collaboration between researchers from different countries, sharing sign language corpora and visual data, so that a truly multilingual system can be developed.

Multidisciplinary Partnerships

As Bragg et al [1] described, an interdisciplinary approach is crucial to sign language processing. Representatives of the deaf community and health care practitioners must be involved in the team to better understand the needs of the community and the specific purposes for which the technology is intended. Linguistics plays a key role in identifying the structures of sign languages that the algorithms need to process. Natural language processing and machine translation offer valuable techniques for modeling, analyzing, and translating. Computer vision is necessary for recognizing signed content, while computer graphics are essential for generating it. Finally, human-computer interaction and design are vital for developing comprehensive systems that meet the needs of the community and integrate seamlessly into people's daily lives [1].

The inclusion of multidisciplinary teams was a notable feature in several studies, combining expertise from linguistics, inclusive design, AI, and health care practitioners. However, the participation of deaf community representatives was limited, potentially undermining cultural and linguistic relevance.

Ethical and Data Security Concerns

As these systems are intended to support real-time communication in clinical settings, their design must go beyond technical optimization to consider ethical issues such as privacy, autonomy, and emotional safety, dimensions largely overlooked in the reviewed studies. The reviewed literature does not adequately address the psychological impact of using these systems. Issues such as the psychological effects on users, especially those who are deaf or hard of hearing, were not discussed in any of the studies. These concerns are particularly important, as assistive technology should promote autonomy and well-being without causing additional stress or emotional burdens for users [56].

Data security is another critical concern that received little attention. Many SLR systems collect sensitive information, such as facial images, gestures, and body movements, which may contain identifiable or private data. However, few studies addressed how these data are stored, protected, or deleted. In health care contexts—where confidentiality is paramount—this lack of attention to privacy and data protection represents a significant oversight.

These omissions highlight a critical gap: as assistive communication technologies are implemented in real-world health care settings, particularly among structurally marginalized populations like the deaf community, ethical design and responsible implementation are essential. Future studies should incorporate ethical considerations from the early stages of system development and report these aspects transparently.

User Experience

Overall, few studies provided reports on the users' experience with the developed systems, which were structurally evaluated and described in only 2 of them [22,36], and only with a small number of participants (12 in Areeb et al [22] and 5 in Adithya and Rajesh [36]). The other studies focused on usability metrics of the developed systems [12,14,15,25,29]. The issues raised corroborate difficulties described in the literature. All recognition systems face the challenge of hand tracking and struggle with self-occlusion between fingers, with additional difficulties in single-hand signals [57]. Such an aspect was mentioned especially by the authors of image-based systems, since in their approaches they did not attempt 3D perception to mitigate occlusion [14,15]. Furthermore, another significant obstacle is the interpersonal variation between signers, which includes not only their physical characteristics but also the way they perform gestures. Minor differences in signaling are important constraints for AI and recognition tools, justifying the reports regardless of the system's type [12,26].

Although the studies that presented user evaluation [34,35] reported that the systems performed well, showing positive impacts on accessibility and effective communication, the evidence is limited, as only 2 studies were conducted, and both involved a small number of users. User experience goes beyond the system's accuracy and usability, and according to International Organization for Standardization [58], it refers to a "person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service," including "all the users' emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviours and accomplishments that occur before, during and after use." Thus, generating a positive user experience for the deaf users is essential for the adoption and use of the proposed systems. While studies typically focus on accuracy and usability metrics, it is equally important to evaluate other factors, such as learning

XSL•FO
RenderX

cost and response speed, which can significantly affect user adoption, especially in health care contexts, where ease of use and time efficiency are critical.

Saeed et al [59], in a review on system-based sensory gloves for SLR, reported that user comfort was a challenge, as using a sensory glove required the user to wear a bulky glove containing sensors, cables, and a circuit board, which limited the user's hand mobility. However, we believe this issue is gradually being reduced, as sensors continue to evolve, becoming smaller and more efficient. Furthermore, the ability to connect these sensors via Bluetooth has further minimized the bulkiness, offering more comfort and flexibility for the user.

The design of interfaces for automatic sign language translation systems, specifically their integration within web-based and social media applications, necessitates careful consideration of numerous factors. Research conducted by Debevc et al [46] and Kožuh et al [60] addresses methodologies for enhancing the interaction experience for deaf and hard-of-hearing users within digital communication platforms [46,60]. The authors examine methods to improve accessibility [46,60].

A recurring concern highlighted in these studies is the potential for picture-in-picture windows to disrupt the user experience for nondeaf individuals. Therefore, the implementation of picture-in-picture displays should be judiciously considered and minimized. The domain of automatic sign language translation is characterized by rapid advancements and multifaceted complexities. The techniques used to optimize the user experience for these systems, while of significant relevance, warrant a dedicated review to comprehensively assess their efficacy and implications.

Limitations and Strengths of the Review

This review is primarily qualitative, as the diversity in methodologies, technologies, and parameter calculations precluded quantitative synthesis. Additionally, there are no standardized tools to assess the risk of bias in these types of studies included. Furthermore, this review focused on systems that recognize sign language. Although these systems do not fully represent the portion of the deaf community that prefers lip reading, writing, or other methods of communication, they do include individuals who use sign language. Finally, 4 publications could not be included, as they were not retrieved despite at least 3 contact attempts regarding 2 studies and due to the unavailability of authors' contact details of the other 2 studies.

The strengths of this systematic review include prospective registration and publication of a protocol [18,19]. Furthermore, it adheres to the Cochrane Guidelines and the PRISMA statement [20,21], and it has a broad search strategy encompassing multiple languages and publication types. A multidisciplinary team, including a deaf linguist, ensured diverse perspectives and representation.

Additionally, this study is innovative, as it focuses specifically on SLR for the health context, an area in which social inclusion and understanding differences are extremely important. Although other reviews were previously published, the majority of them cannot be considered true systematic reviews [49,59,61,62]. They lack a prepublication of a protocol, a robust methodology, a broad search strategy, paired screening and data extraction, and comprehensiveness. They usually focus on technical aspects of a broader and more general analysis of sign language translation systems and use restricted publication periods and language (English only) in the search. Overall, they did not include a search in MEDLINE, which is a crucial database for health-related research, and presented methodological shortcomings, such as not mentioning the use of paired reviewers for study selection and data extraction. Furthermore, they have not reported extracting data on user experience, cost, and data security. Unlike these more general reviews, our study is unique in focusing on the health care context. Therefore, our study fills these gaps, providing a more robust, detailed, and comprehensive contribution specifically to the health field.

Final Considerations

Image-based systems dominated the field, demonstrating greater accessibility but facing challenges with lighting conditions and environmental variability. While sensor-based systems showed higher precision, their reliance on specialized equipment limits scalability. Critical aspects, such as bidirectional communication and facial expression recognition, remain underexplored, hindering the practical implementation of these systems in health care settings.

Based on the evidence reviewed, effective SRL systems for real-world health care scenarios are scarce. Future research should focus on developing adaptive systems capable of recognizing diverse sign languages and addressing the varying communication needs of the deaf community. It should prioritize inclusive and participatory design processes that respect the linguistic integrity of sign languages and address sociocultural nuances critical to effective communication. Expanding datasets, incorporating user feedback, and fostering multidisciplinary collaborations will be pivotal for creating inclusive and scalable solutions.

Additionally, future research should incorporate a comprehensive assessment of the proposed systems, addressing both technical and user-centered aspects. This assessment should consider device compatibility (especially for smartphone-based systems), required learning time, translation accuracy across different signers and accents [63], user trust in automated translation, its effect on self-expression, and privacy implications. These evaluations should not only be conducted once the system is fully implemented but also throughout the development process (ie, formative evaluation) [64].

Another significant gap identified in the literature is the lack of a standardized evaluation framework for SLR systems. The inclusion of consistent and well-defined indicators, such as user satisfaction and economic viability, would provide a more comprehensive understanding of the systems' effectiveness and their potential for widespread adoption. User satisfaction is a crucial success factor, and including economic indicators would help assess whether the systems can be feasibly adopted in different settings, especially in low- and middle-income countries. The implementation of the framework would not only enhance the comparability of findings but also provide valuable

XSL•FO
RenderX

insights for policymakers and stakeholders in the health care industry.

Furthermore, we suggest that the development of future SLR solutions involves experts in ethics, law, and data security to ensure compliance with legal and ethical standards and to protect users during the implementation of these technologies. Studies should provide more detailed information about how ethical and data security issues are managed and the potential impacts on user trust and adoption.

One notable strength identified in this review was the initiative of a research group to open-source both their dataset and system code [32]. This practice promotes transparency, reproducibility, and collaboration—key elements for the advancement of SLR systems. By allowing other researchers to access and build upon existing work, open data and open-source tools can accelerate the refinement of models and the development of more accurate and user-friendly solutions. In the long term, such collaborative practices may contribute to the creation of robust, scalable systems that are truly usable in health care settings, bridging communication gaps and improving care for deaf individuals. Encouraging the adoption of open science principles within the SLR research community could significantly enhance innovation and standardization efforts in this emerging field.

Conclusions

This review highlights the ongoing development of communication systems designed to assist deaf individuals who use sign language to improve their interaction with health care providers. There was a predominance of image-based approaches over sensor-based ones, though both have demonstrated substantial variability in accuracy in recognizing and interpreting signs. However, critical aspects such as bidirectional communication and the recognition of facial expressions, which are essential for effective communication, were notably absent in most studies. None of the systems has reported to have addressed all aspects critical to integration into health care settings. These findings underscore the need for further research, especially regarding the practical implementation of these systems, their usability, and their overall effect on the quality of care for deaf patients.

Acknowledgments

This manuscript was prepared by the authors without the assistance of artificial intelligence tools.

Funding

This study was funded in part by the Institute for Health Assessment and Translation for Highly Relevant Chronic and Neglected Diseases (IATS CARE)/National Institute of Science and Technology for Health Technology Assessment (Instituto de Avaliação de Tecnologias em Saúde) or National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico; grant 465518/2014-1), Financiadora de Estudos e Projetos (grant 01.21.0097.00), and SignumWeb Comunicação Inclusiva Ltda, as part of the project “Captar-Libras: Sistema de Comunicação por vídeos para surdos aplicado ao pré-atendimento médico.” MSM, MFMC, and LRV were supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (grants 311742/2025-4, 316841/2021-8, and 134878/2023-0, respectively). The funding sources had no role in the design of this study and will not have any role during its execution, analyses, interpretation of the data, or decision.

Data Availability

All data generated or analyzed during this study are included in this published paper and its Multimedia Appendices 1,2,3,4,5,6.

Authors' Contributions

MSM, LFRdO, LRV, ROP, and ZSNR developed a draft of the literature search, which LFRdO and LRV executed. MSM, ROP, ZSNR, and MFMC guided the construction of the protocol, and MSM is a guarantor of the review. LFRdO, LRV, LMMSR, GTdSS, NSS, and MRdC contributed to the work’s extraction, analysis, or interpretation of data. MSM, LFRdO, LRV, LMMSR, MRdC, and GTdSS drafted the manuscript; all the other authors performed critical revisions. ELAB and RAAC are professors of linguistics and experts in sign language. They reviewed the entire manuscript to ensure the language was consistent with the literature on disability and deafness. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File, 2988 KB - jmir_v28i1e70417_app1.docx]

Multimedia Appendix 2
The definitions used regarding the type of sign language recognition system, the corpus formation, and the health context.

XSL•FO
RenderX

[DOCX File, 17 KB - jmir_v28i1e70417_app2.docx]

Multimedia Appendix 3
Codebook.
[DOCX File, 2489 KB - jmir_v28i1e70417_app3.docx]

Multimedia Appendix 4
Studies not retrieved.
[DOCX File, 14 KB - jmir_v28i1e70417_app4.docx]

Multimedia Appendix 5
Development and testing of different sign language recognition systems.
[DOCX File, 30 KB - jmir_v28i1e70417_app5.docx]

Multimedia Appendix 6
Technologies necessary for sign language recognition system implementation.
[DOCX File, 16 KB - jmir_v28i1e70417_app6.docx]

Checklist 1
PRISMA checklist.
[DOCX File, 32 KB - jmir_v28i1e70417_app7.docx]

References

1. Bragg D, Koller O, Bellard M, et al. Sign language recognition, generation, and translation: an interdisciplinary perspective. Presented at: Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility; Oct 28-30, 2019. [doi: 10.1145/3308561.3353774]
2. Kendon A. Gesture Visible Action as Utterance: Cambridge University Press; 2004.
3. Deafness and hearing loss. World Health Organization. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss [accessed 2024-11-14]
4. World Federation of the Deaf. URL: https://wfdeaf.org [accessed 2025-12-05]
5. Ubido J, Huntington J, Warburton D. Inequalities in access to healthcare faced by women who are deaf. Health Soc Care Community 2002 Jul;10(4):247-253. [doi: 10.1046/j.1365-2524.2002.00365.x] [Medline: 12193168]
6. Kuenburg A, Fellinger P, Fellinger J. Health care access among deaf people. J Deaf Stud Deaf Educ 2016 Jan;21(1):1-10. [doi: 10.1093/deafed/env042] [Medline: 26405210]
7. Mansutti I, Achil I, Rosa Gastaldo C, Tomé Pires C, Palese A. Individuals with hearing impairment/deafness during the COVID-19 pandemic: a rapid review on communication challenges and strategies. J Clin Nurs 2023 Aug;32(15-16):4454-4472. [doi: 10.1111/jocn.16572] [Medline: 36320127]
8. Rannefeld J, O'Sullivan JL, Kuhlmey A, Zoellick JC. Deaf and hard-of-hearing patients are unsatisfied with and avoid German health care: results from an online survey in German Sign Language. BMC Public Health 2023 Oct 18;23(1):2026. [doi: 10.1186/s12889-023-16924-w] [Medline: 37848898]
9. Rastgoo R, Kiani K, Escalera S. Sign language recognition: a deep survey. Expert Syst Appl 2021 Feb;164:113794. [doi: 10.1016/j.eswa.2020.113794]
10. Papastratis I, Chatzikonstantinou C, Konstantinidis D, Dimitropoulos K, Daras P. Artificial intelligence technologies for sign language. Sensors (Basel) 2021 Aug 30;21(17):5843. [doi: 10.3390/s21175843] [Medline: 34502733]
11. Parton BS. Sign language recognition and translation: a multidisciplined approach from the field of artificial intelligence. J Deaf Stud Deaf Educ 2006;11(1):94-101. [doi: 10.1093/deafed/enj003] [Medline: 16192405]
12. Sosa-Jiménez CO, Ríos-Figueroa HV, Solís-González-Cosío AL. A prototype for Mexican Sign Language recognition and synthesis in support of a primary care physician. IEEE Access 2022;10:127620-127635. [doi: 10.1109/ACCESS.2022.3226696]
13. Xia K, Lu W, Fan H, Zhao Q. A sign language recognition system applied to deaf-mute medical consultation. Sensors (Basel) 2022 Nov 24;22(23):9107. [doi: 10.3390/s22239107] [Medline: 36501809]
14. Pikoulis EV, Bifis A, Trigka M, Constantinopoulos C, Kosmopoulos D. Context-aware automatic sign language video transcription in psychiatric interviews. Sensors (Basel) 2022 Mar 30;22(7):2656. [doi: 10.3390/s22072656] [Medline: 35408270]
15. da Silva DRB, de Araújo TMU, do Rêgo TG, Brandão MAC, Gonçalves LMG. A multiple stream architecture for the recognition of signs in Brazilian sign language in the context of health. Multimed Tools Appl 2024;83(7):19767-19785. [doi: 10.1007/s11042-023-16332-7]

XSL•FO
RenderX

16. S Kumar S, Wangyal T, Saboo V, Srinath R. Time series neural networks for real time sign language translation. Presented at: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA); Dec 17-20, 2018. [doi: 10.1109/ICMLA.2018.00043]
17. Alaghband M, Maghroor HR, Garibay I. A survey on sign language literature. Mach Learn Appl 2023 Dec;14:100504. [doi: 10.1016/j.mlwa.2023.100504]
18. Sign language recognition system for deaf patients: a systematic review. Open Science Framework. 2023. URL: https://osf.io/fpemr/overview [accessed 2026-01-28]
19. Marcolino MS, Oliveira LFR, Valle LR, et al. Sign language recognition system for deaf patients: protocol for a systematic review. JMIR Res Protoc 2025 Jan 23;14(1):e55427. [doi: 10.2196/55427] [Medline: 39847417]
20. Higgins JPT. Cochrane Handbook for Systematic Reviews of Interventions Version 65 (Updated August 2024): Cochrane; 2024.
21. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
22. Areeb QM, Nadeem M, Alroobaea R, Anwer F. Helping hearing-impaired in emergency situations: a deep learning-based approach. IEEE Access 2022;10:8502-8517. [doi: 10.1109/ACCESS.2022.3142918]
23. Areeb QM, Nadeem M. Deep learning based hand gesture recognition for emergency situation: a study on Indian Sign Language. Presented at: 2021 International Conference on Data Analytics for Business and Industry (ICDABI); Oct 25-26, 2021. [doi: 10.1109/ICDABI53623.2021.9655842]
24. da Silva DRB, Araujo TMU, Rêgo TD, Brandão MAC. A two-stream model based on 3D convolutional neural networks for the recognition of brazilian sign language in the health context. Presented at: WebMedia '20: Proceedings of the Brazilian Symposium on Multimedia and the Web; Nov 30 to Dec 4, 2020. [doi: 10.1145/3428658.3430980]
25. Das S, Biswas SK, Purkayastha B. Indian Sign Language recognition system for emergency words by using shape and deep features. Presented at: 2023 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON); Feb 10-11, 2023. [doi: 10.1109/IEMECON56962.2023.10092312]
26. Ko SK, Kim CJ, Jung H, Cho C. Neural sign language translation based on human keypoint estimation. Appl Sci (Basel) 2019;9(13):2683. [doi: 10.3390/app9132683]
27. Barrientos-Villalta GF, Quiroz P, Ugarte W. Peruvian Sign Language recognition using recurrent neural networks. Presented at: International Conference on Advanced Research in Technologies, Information, Innovation and Sustainability; Sep 12-15, 2022. [doi: 10.1007/978-3-031-20319-0_34]
28. Ramírez Sánchez JE, Rodríguez AA, Mendoza MG. Real-time Mexican Sign Language interpretation using CNN and HMM. Presented at: Mexican International Conference on Artificial Intelligence; Oct 25-30, 2021. [doi: 10.1007/978-3-030-89817-5_4]
29. Hisham B, Hamouda A. Supervised learning classifiers for Arabic gestures recognition using Kinect V2. SN Appl Sci 2019 Jul;1(7):1-21. [doi: 10.1007/s42452-019-0771-2]
30. Gandhi J, Gandhi P, Gosar A, Chaudhari S. Video recognition techniques for Indian Sign Language in healthcare domain. Presented at: 2021 2nd International Conference for Emerging Technology (INCET); May 21-23, 2021. [doi: 10.1109/INCET51464.2021.9456116]
31. Uchil AP, Jha S, Sudha BG. Vision-based deep learning approach for dynamic Indian Sign Language recognition in healthcare. Presented at: Computational Vision and Bio-Inspired Computing: ICCVBIC 2019; Sep 25-26, 2019. [doi: 10.1007/978-3-030-37218-7_43]
32. Deji Dere M, Dere RO, Adesina A, Yauri AR. SmartCall: a real-time, sign language medical emergency communicator. Presented at: 2022 5th Information Technology for Education and Development (ITED); Nov 1-3, 2022. [doi: 10.1109/ITED56637.2022.10051420]
33. Sarhan NA, El-Sonbaty Y, Youssef SM. HMM-based Arabic Sign Language recognition using Kinect. Presented at: 2015 Tenth International Conference on Digital Information Management (ICDIM); Dec 23-24, 2015. [doi: 10.1109/ICDIM.2015.7381873]
34. Guo Y, Zhao J, Ding B, et al. Sign-to-911: emergency call service for sign language users with assistive AR glasses. Presented at: ACM MobiCom '23: Proceedings of the 29th Annual International Conference on Mobile Computing and Networking; Oct 2-5, 2023. [doi: 10.1145/3570361.3613260]
35. Süzgün M, Özdemir H, Camgöz N, et al. HospiSign: an interactive sign language platform for hearing impaired. J Naval Sci Eng 2015;11(3):75-92 [FREE Full text]
36. Adithya V, Rajesh R. Hand gestures for emergency situations: a video dataset based on words from Indian Sign Language. Data Brief 2020 Aug;31:106016. [doi: 10.1016/j.dib.2020.106016] [Medline: 32715044]
37. Ihsan MA, Eram AF, Nahar L, Kadir MA. MediSign: an attention-based CNN-BiLSTM approach of classifying word level signs for patient-doctor interaction in hearing impaired community. IEEE Access 2024;12:33803-33815. [doi: 10.1109/ACCESS.2024.3370684]
38. Das HV, Mohan K, Paul L, Kumaresan S, Nair CS. Transforming consulting atmosphere with Indian Sign Language translation. Multimed Tools Appl 2024;83(5):13543-13555. [doi: 10.1007/s11042-023-15214-2]

XSL•FO
RenderX

39. Faisal M, Alsulaiman M, Mekhtiche M, et al. Enabling two-way communication of deaf using Saudi Sign Language. IEEE Access 2023;11:135423-135434. [doi: 10.1109/ACCESS.2023.3337514]
40. Luqman H, Mahmoud SA. A machine translation system from Arabic Sign Language to Arabic. Univ Access Inf Soc 2020 Nov;19(4):891-904. [doi: 10.1007/s10209-019-00695-6]
41. Dewasurendra D, Kumar A, Perera I, Jayasena D, Thelijjagoda S. Emergency communication application for speech and hearing-impaired citizens. Presented at: 2020 From Innovation to Impact (FITI); Dec 15-15, 2020. [doi: 10.1109/FITI52050.2020.9424899]
42. Bellil L, Ghiri MG, Boulesnane A. Empowering deaf community in healthcare communication: 1D-CNN-based Algerian Sign Language recognition system. Presented at: 2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS); Apr 24-25, 2024. [doi: 10.1109/PAIS62114.2024.10541233]
43. Vosk: speech recognition toolkit. Alphacephei. URL: https://alphacephei.com/vosk/ [accessed 2026-03-24]
44. Arabic language disambiguation for natural language processing applications. Columbia University. URL: https://inventions.techventures.columbia.edu/technologies/arabic-language-disambiguation--cu14012 [accessed 2026-03-24]
45. Shotton J, Fitzgibbon A, Cook M, et al. Real-time human pose recognition in parts from single depth images. Presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jun 20-25, 2011. [doi: 10.1109/CVPR.2011.5995316]
46. Debevc M, Kosec P, Holzinger A. Improving multimodal web accessibility for deaf people: sign language interpreter module. Multimed Tools Appl 2011 Aug;54(1):181-199. [doi: 10.1007/s11042-010-0529-8]
47. Madhiarasan M, Roy PP. A comprehensive review of sign language recognition: different types, modalities, and datasets. arXiv. Preprint posted online on Apr 7, 2022. [doi: 10.48550/arXiv.2204.03328]
48. Haji Mohd MN, Mohd Asaari MS, Lay Ping O, Rosdi BA. Vision-based hand detection and tracking using fusion of kernelized correlation filter and single-shot detection. Appl Sci (Basel) 2023;13(13):7433. [doi: 10.3390/app13137433]
49. Al-Qurishi M, Khalid T, Souissi R. Deep learning for sign language recognition: current techniques, benchmarks, and open issues. IEEE Access 2021;9:126917-126951. [doi: 10.1109/ACCESS.2021.3110912]
50. Yang R, Sarkar S, Loeding B. Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming. IEEE Trans Pattern Anal Mach Intell 2009;32(3):462-477. [doi: 10.1109/TPAMI.2009.26]
51. Amin MS, Rizvi STH, Hossain MM. A comparative review on applications of different sensors for sign language recognition. J Imaging 2022 Apr 2;8(4):98. [doi: 10.3390/jimaging8040098] [Medline: 35448225]
52. Figg B. Deafness and other communication disorders. J Consum Health Internet 2023 Apr 3;27(2):199-209. [doi: 10.1080/15398285.2023.2204717]
53. Bihari Dash D. Significance of globalisation and English language. Int J Stud Engl Lang Lit 2022;10(5):10-16. [doi: 10.20431/2347-3134.1005002]
54. Nye B, Jessy Li J, Patel R, et al. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. Presented at: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Jul 15-20, 2018. [doi: 10.18653/v1/P18-1019]
55. Zheng L, Liang B, Jiang A. Recent advances of deep learning for sign language recognition. Presented at: 2017 International Conference on Digital Image Computing; Nov 29 to Dec 1, 2017. [doi: 10.1109/DICTA.2017.8227483]
56. Jiménez-Arberas E, Díez E. Psychosocial impact of assistive devices and other technologies on deaf and hard of hearing people. Int J Environ Res Public Health 2021 Jul 7;18(14):7259. [doi: 10.3390/ijerph18147259] [Medline: 34299710]
57. Adeyanju IA, Bello OO, Adegboye MA. Machine learning methods for sign language recognition: a critical review and analysis. Intell Syst Appl 2021 Nov;12:200056. [doi: 10.1016/j.iswa.2021.200056]
58. Ergonomics of human-system interaction: Part 210: Human-centered design for interactive systems. International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-1:v1:en [accessed 2025-04-13]
59. Saeed ZR, Zainol ZB, Zaidan BB, Alamoodi AH. A systematic review on systems-based sensory gloves for sign language pattern recognition: an update from 2017 to 2022. IEEE Access 2022;10:123358-123377. [doi: 10.1109/ACCESS.2022.3219430]
60. Kožuh I, Hintermair M, Holzinger A, Volčič Z, Debevc M. Enhancing universal access: deaf and hard of hearing people on social networking sites. Univ Access Inf Soc 2015 Nov;14(4):537-545. [doi: 10.1007/s10209-014-0354-3]
61. Boggaram A, Boggaram A, Sharma A, Srinivasa Ramanujan A, Bharathi R. Sign language translation systems. Int J Softw Sci Comput Intell 2022 Jan;14(1):1-33. [doi: 10.4018/IJSSCI.311448]
62. Waghmare PP. Techniques for generating sign language a comprehensive review. J Inst Eng India Ser B 2024 Dec;105(6):1789-1803. [doi: 10.1007/s40031-024-01118-8]
63. Rowley K, Cormier K. Accent or not? Language attitudes towards regional variation in British Sign Language. Appl Linguist Rev 2023 Jul 26;14(4):919-943. [doi: 10.1515/applirev-2020-0144]
64. Preece J, Rogers Y, Sharp H. Interaction Design: Beyond Human-Computer Interaction, 6th edition: John Wiley & Sons; 2023.

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
CNN: convolutional neural network
HMM: hidden Markov model
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SLR: sign language recognition
YOLO: You Only Look Once

Edited by Y Li; submitted 22.Dec.2024; peer-reviewed by A Holzinger, S Singh, Z Yang; revised version received 03.Aug.2025; accepted 05.Aug.2025; published 09.Apr.2026.

Please cite as:
Marcolino MS, Oliveira LFRD, Valle LR, Rosa LMMS, Sanches GTDS, Santos NS, Costa MRD, Bernardino ELA, Cordeiro RAA, Prates RO, Reis ZSN, Campos MFM
Innovations in Deaf Health Care Communication: Systematic Review of Sign Language Recognition Systems
J Med Internet Res 2026;28:e70417
URL: https://www.jmir.org/2026/1/e70417
doi:10.2196/70417

© Milena Soriano Marcolino, Lucca Fagundes Ramos de Oliveira, Lucas Rocha Valle, Luiza Marinho Motta Santa Rosa, Gabriela Teodora de Souza Sanches, Natalia Sales Santos, Michelle Ralil da Costa, Elidea Lucia Almeida Bernardino, Raniere Alislan Almeida Cordeiro, Raquel Oliveira Prates, Zilma Silveira Nogueira Reis, Mario Fernando Montenegro Campos. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Machine Learning in Predicting the Risk of Esophagogastric Variceal Bleeding Among Patients With Liver Cirrhosis: Systematic Review and Meta-Analysis

Yuan Lian[1*], MS; Xinping Qiu[1], MBBS; Geli Wang[1], MBBS; Nannan Liu[1], MBBS; Jing Zhao[1], MBBS; Shanshan Song[1], MBBS; Shiqi Wang[1], MS; Mingjun Sun[1,2*], MD

[1]Department of Gastroenterology, Fuxin Mining General Hospital of Liaoning Health Industry Group, No.43,Health Care Street,Haizhou District, Fuxin, China

[2]Department of Gastroenterology, The First Hospital of China Medical University, Shenyang, China

[*]these authors contributed equally

Corresponding Author:
Yuan Lian, MS
Department of Gastroenterology, Fuxin Mining General Hospital of Liaoning Health Industry Group, No.43,Health Care Street,Haizhou District, Fuxin, China

Abstract

Background: Liver cirrhosis (LC) can lead to several complications. Esophageal variceal bleeding (EVB) and esophagogastric variceal bleeding (EGVB) are particularly severe, leading to a high risk of mortality. Early identification of esophageal varices and esophagogastric varices is essential. Several studies have constructed prediction models for EVB and EGVB in patients with LC. However, robust systematic evidence to prove their performance is lacking.

Objective: We included original studies that developed prediction models for esophageal or gastric variceal bleeding in patients with LC under different modeling variables. This study aimed to review the predictive performance of various models for EVB or EGVB in patients with LC, providing insights into the development or updating of simplified scoring tools in the future.

Methods: PubMed, Web of Science, Embase, and the Cochrane Library were searched up to August 21, 2024, to collect original full-text studies on machine learning (ML) in the prediction of EVB and EGVB in patients with LC. The models were evaluated using the Prediction Model Risk of Bias Assessment Tool (PROBAST). Subgroup analyses were carried out based on the modeling variables.

Results: In total, 21 studies were included, with 7011 patients with LC, among whom 1412 (20.14%) developed EVB and 733 (10.45%) developed EGVB. The meta-analysis results suggested that the pooled c-index, sensitivity, and specificity of the prediction model for predicting EVB in the validation set were 0.85 (95% CI 0.77 - 0.92), 0.93 (95% CI 0.87 - 0.96), and 0.66 (95% CI 0.46 - 0.82), respectively. The pooled c-index, sensitivity, and specificity of the prediction model for predicting EGVB in the validation set were 0.89 (95% CI 0.85 - 0.94), 0.77 (95% CI 0.66 - 0.85), and 0.81 (95% CI 0.67 - 0.90), respectively. The subgroup analysis based on modeling variables revealed that, for predicting EVB, the c-index in the validation set was 0.84 (95% CI 0.80 - 0.88) for models based on clinical features, 0.82 (95% CI 0.69 - 0.96) for radiomics-based models, 0.78 (95% CI 0.67 - 0.89) for models based on radiomics and clinical features, and 0.97 (95% CI 0.95 - 1.00) for models based on endoscopic features. Subgroup analyses based on modeling variables revealed that, for predicting EGVB, the c-index in the validation set was 0.91 (95% CI 0.86 - 0.96) for models based on clinical features and 0.85 (95% CI 0.75 - 0.96) for models based on radiomics and clinical features.

Conclusions: ML methods are feasible for predicting EVB and EGVB in patients with LC. Nevertheless, the number of included original studies is limited. In the future, more studies with larger sample sizes are needed to promote the application of ML in the early assessment of EVB and EGVB in patients with LC in clinical practice.

Trial Registration: PROSPERO CRD42024585100; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024585100

(J Med Internet Res 2026;28:e78203) doi:10.2196/78203

KEYWORDS

liver cirrhosis; hepatic; prediction model; esophageal variceal bleeding; esophagogastric variceal bleeding; radiomics; machine learning

XSL•FO
RenderX

Introduction

Liver cirrhosis (LC) often occurs after long-term inflammation, resulting in fibrotic tissue and regenerative nodules and, subsequently, portal hypertension [1]. Approximately 1 million deaths worldwide are attributable to LC every year [2]. LC is the 11th leading cause of death worldwide and the third leading cause of death among individuals aged 45 to 64 years. Together with liver cancer, it represents 3.5% of global deaths [2]. LC contributed to 2.4% of deaths worldwide in 2019 [3]. LC is a notable cause of both morbidity and mortality in patients with chronic liver diseases, leading to hepatocellular carcinoma (HCC) and liver decompensation. LC-induced portal hypertension is associated with complications such as ascites, hepatic encephalopathy, acute kidney injury, hepatorenal syndrome, and portal hypertension–related bleeding, such as esophagogastric variceal bleeding (EGVB) [1]. Thus, LC becomes a major health care burden worldwide.

The incidence of EGVB is high among the primary complications of LC. It is the second most prevalent complication in patients with LC, after ascites. EGVB, a frequent cause of bleeding, is regarded as a medical emergency. This complication, even when well managed, is associated with a mortality rate of approximately 20% within 6 weeks. If an infection occurs, the risk of mortality increases further [1]. EGVB is a critical condition and often recurs. In severe cases, it may be life threatening [4]. Currently, the treatment methods for EGVB are diverse. For stable patients after volume resuscitation, cautious transfusion is advised, with a transfusion threshold of 7 g/dL. For hemostatic treatment, vasoactive agents such as octreotide, somatostatin, or terlipressin, together with endoscopic interventions such as variceal ligation, should be used. A transjugular intrahepatic portosystemic shunt is used for resuscitation when uncontrolled bleeding occurs. It is advisable to administer prophylactic antibiotics, such as ceftriaxone, for 7 days or until the patient is discharged. For patients at high risk of recurrent bleeding, specifically those with a Child-Turcotte-Pugh B score between 8 and 9 points and active bleeding or a Child-Turcotte-Pugh C score ranging from 10 to 13 points, pre-emptive transjugular intrahepatic portosystemic shunt within 72 hours of admission could prolong their survival. For patients who are not at high risk, β-blockers combined with variceal ligation are recommended [1]. These conditions not only intensify patients' pain but also shorten their lifespan and increase their economic burden. Thus, early identification of the risk of esophagogastric varices (EGV) in patients with LC is crucial for preventing EGVB, lowering its threat to health and life and the economic burden on families.

Endoscopy, which is interventional, is the primary tool for diagnosing esophageal variceal bleeding (EVB) and EGVB. As the gold standard for detecting EVB and EGVB, endoscopy is also recommended for diagnosing and screening for EGV and assessing the risk of bleeding [5]. Nonetheless, endoscopic examination is an invasive procedure, and a portion of patients may refuse this examination due to low tolerance. Moreover, repeated procedures might diminish long-term adherence. Furthermore, this examination may also induce EGVB in low-risk patients. Thus, recent studies have focused on the development of noninvasive or minimally invasive tools to identify the risks of EGV at an early stage. However, effective tools for the early prediction of EGV remain lacking. Recently, as machine learning (ML) has been increasingly used in medical diagnostics, some studies have investigated the feasibility of using ML techniques to evaluate the risk of EGVB in patients with LC. Nevertheless, their detection performance varies notably with the modeling approach or the variables used. Currently, there is insufficient evidence to demonstrate the feasibility of using ML to identify EVB and EGVB in patients with LC. Thus, this study included original studies that developed prediction models for esophageal or gastric variceal bleeding in patients with LC under different modeling variables. The aim of this study was to review the performance of ML in the early identification of EVB and EGVB in patients with LC, providing evidence for the advancement and updating of early risk assessment tools for EGV.

Methods

Study Registration

This study was performed based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Checklist 1) and was prospectively registered with PROSPERO (ID CRD42024585100).

Eligibility Criteria

Before screening the original studies, we established comprehensive eligibility criteria (Table 1).

XSL•FO
RenderX

Table . Eligibility criteria for screening the original studies for this systematic review.

Items	Inclusion criteria	Exclusion criteria
Population	Patients with LC[a] as the study participants.	Studies that did not strictly distinguish between patients with liver cirrhosis and those with hepatocellular carcinoma.
Intervention	Studies that developed prediction models for the risk of EVB[b] or EGVB[c].	Studies that solely explored risk factors without creating a complete prediction model, studies that only centered on validated mature scales, studies on the predictive accuracy of a single factor for outcomes, and studies that did not strictly distinguish EVB or EGVB from other upper gastrointestinal bleeding.
Comparison	None	None
Outcome	Studies reporting any of the following outcome measures to evaluate the prediction model: c-index, sensitivity, specificity, accuracy, recall, precision, confusion matrix, F_1-score, and calibration curve.	None
Study design	Studies adopting case-control, cohort, and cross-sectional designs; and original studies reported in English.	Conference abstracts that were published without peer review and for studies that repeatedly trained and validated models based on the same dataset, only the most recently published study was included.

[a]LC: liver cirrhosis.

[b]EVB: esophageal variceal bleeding

[c]EGVB: esophagogastric variceal bleeding.

Data Sources and Search Strategy

PubMed, Web of Science, Embase, and the Cochrane Library were searched up to August 21, 2024. Both subject terms and free terms were used to design search terms. No restrictions regarding region or period were imposed. Multimedia Appendix 1 provides the detailed search strategy.

Study Selection and Data Extraction

Articles retrieved from the databases were imported into EndNote (Clarivate Analytics). First, duplicates were removed. By reading the titles and abstracts, potentially eligible original studies were chosen. The full texts of these studies were then downloaded and read to determine eligible articles. Subsequently, a standardized spreadsheet was developed to extract data (eg, study title, first author, publication year, and study type) from the included articles. Two investigators independently screened the studies, extracted the data, and then cross-checked the results. Any disagreements were resolved through consultation with a third investigator.

Risk of Bias Assessment

The risk of bias (ROB) in the included studies was appraised using the Prediction Model Risk of Bias Assessment Tool (PROBAST). This tool encompasses 4 domains: participant, predictor variable, outcome, and statistical analysis. Questions in each domain could be answered with 3 responses: yes or probably yes, no or probably no, and no information. A domain was deemed to have high ROB if any question was answered with no or probably no. A domain was deemed to have low ROB if all questions were answered with yes or probably yes. The overall ROB was low if all domains were considered low ROB. The overall ROB was high if at least one domain was considered to have a high ROB. Two investigators independently appraised the ROB using the PROBAST, and their results were cross-checked. Disagreements, if any, were resolved by a third investigator.

Synthesis Methods

For evaluating the overall accuracy, a meta-analysis of the c-indices of ML models was performed. In certain original studies, the 95% CI and SEs for the c-index were missing. We referred to the study by Debray et al [6] to estimate the SEs. Heterogeneity across studies was assessed using the I^2 statistic. A random-effects model was used if I^2 was greater than 50%, whereas a fixed effects model was adopted if I^2 was less than 50%. Additionally, the sensitivity and specificity, derived from the diagnostic quadrangle table, were pooled through bivariate mixed effects models. However, such a table was not available in many of the original studies. In this case, the number of cases with sensitivity, specificity, and precision was combined to compute the estimates. Sensitivity and specificity were extracted based on the optimal Youden index. This meta-analysis was performed in Stata (version 15.0; StataCorp LLC).

Results

Study Selection

In total, 1215 articles were retrieved from the databases. Following the removal of 144 (11.85%) duplicates, 1071 (88.15%) articles were left. In total, 1041(85.68%) articles were excluded due to unrelated topics, including meta-analyses, reviews, guidelines, animal experiments, case reports, letters,

XSL•FO
RenderX

and registration protocols. The full texts of the remaining 30 (2.47%) articles were then downloaded and read. We further excluded 4 (0.33%) conference abstracts that had not undergone peer review, 1 (0.08%) study that analyzed only risk factors without developing a complete ML model, 2 (0.16%) studies without an accuracy assessment of model outcomes, and 2 (0.16%) studies published based on the same dataset of another 2 studies. Finally, 21 (1.73%) articles were included [7-27] (Figure 1).

Figure 1. .PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the literature screening process.

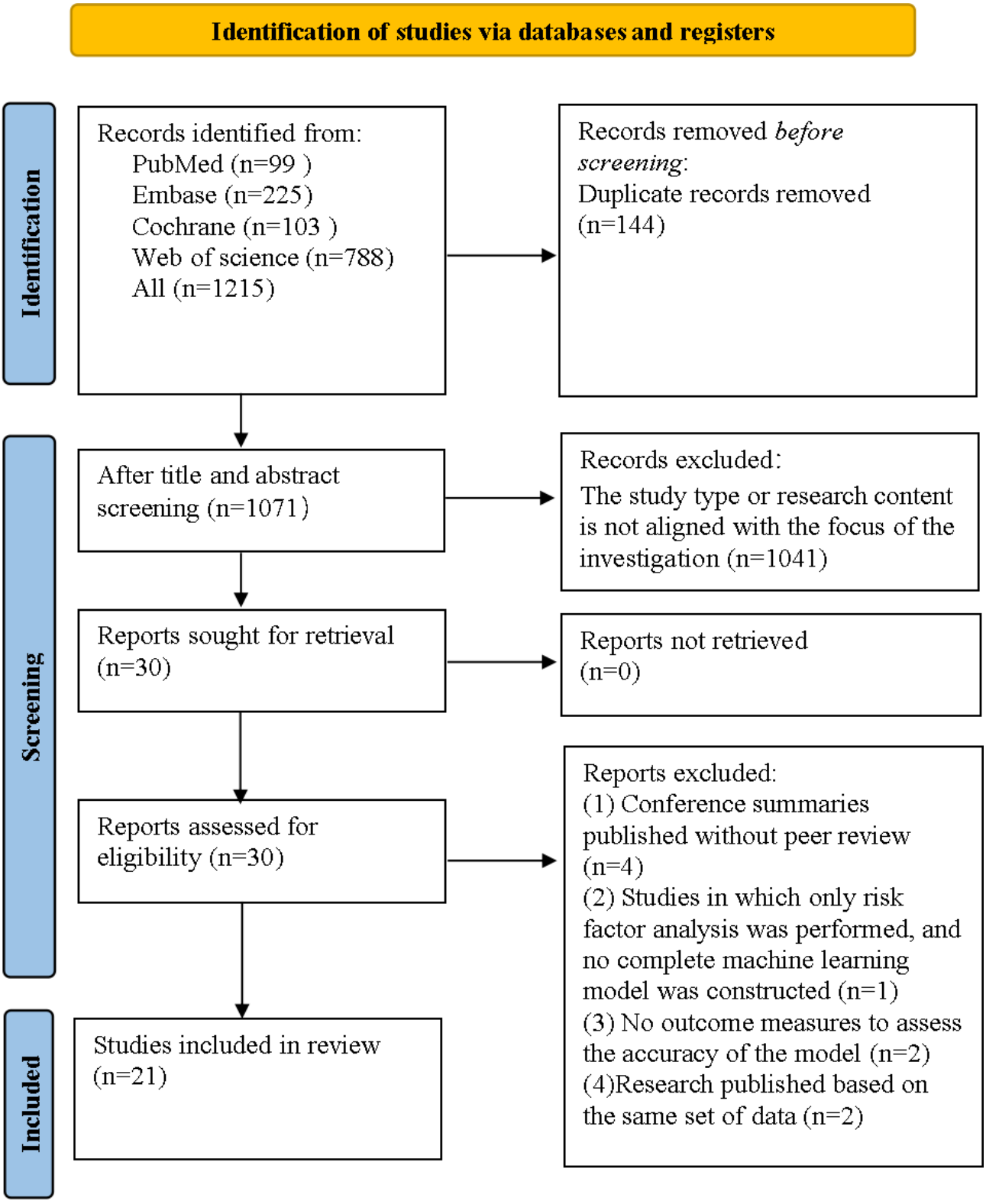

XSL•FO
RenderX

Study Characteristics

The included studies were published between 2014 and 2024, involving a total of 7011 patients with LC. Among these studies, 12 studies included 1412 cases of EVB [8,10-15,19-22,25], and 9 studies included 733 cases of EGVB [7-9,11,12,16,17,26,27]. These studies were conducted in 4 countries: 18 studies in China [7-20,23,25-27], 1 study in Egypt [21], 1 study in Iran [22], and 1 study in South Korea [24]. Among the included studies, 14 were case-control studies [7,12-16,18-20,22,24-27], 6 were cohort studies [8-11,21,23], and 1 was a cross-sectional study [15]. There were 16 single-center studies [7-9,12-17,19-23,25,27] and 5 multicenter studies [10,11,18,24,26]. Nine articles [7-9,11,12,16,17,26,27] centered on EGVB, and 12 articles focused on EVB [10,13-15,18-25]. Fourteen studies clearly outlined the method used to generate the validation set. Only 3 studies used external validation [10,18,19]. Two studies used time-based split [8,25], while the remaining studies conducted internal validation with random sampling. The models involved were mainly logistic regression models, while only a few studies constructed other types of models. Fifteen models were constructed based on clinical features [7-9,11,15-25]. Three models were based on radiomics [13,26,27]. Four models were based on radiomics and clinical features [12,14,26,27], and 1 model was based on endoscopic features in 1 study [10]. The detailed results are presented in Table 2.

XSL•FO
RenderX

Table . Basic characteristics of the included studies.

First author	Year of publication	Country of authors	Study type	Patient source	Predicted events	Event cases (n=2145), n (30.59%)	Cases (n=7011), n	Cases in training set (n=5554), n	Generation method of validation set	Cases in the validation set (n=1457), n	Model type	Modeling variables
Liu et al [7]	2022	China	Case-control study	Single center	EGVB[a]	77	330	220	Random sampling	110	LR[b]	Clinical features (emotional stimuli, improper diet, overwork, lower temperature, and increased abdominal pressure)
Zhang et al [8]	2022	China	Prospective cohort study	Single center	EGVB	52	172	128	Time-based split	44	Cox[c]	Clinical features (vPPG[d], PLT[e], albumin, and INR)[f]
Tan et al [9]	2023	China	Prospective study (cohort)	Single center	EGVB	74	230	160	Random sampling	70	LR	Clinical features (right liver lobe volume, diameters of maximum esophagogastric varices, and portal vein system)
Wang et al [10]	2023	China	Retrospective cohort study	Multicenter	EVB[g]	121	502	275	Random sampling and external validation	227	DL[h], XGBoost[i], GLM[j], GBM[k], and RF[l]	Endoscopic features (EV[m] bleeding image)

XSL•FO
RenderX

First author	Year of publication	Country of authors	Study type	Patient source	Predicted events	Event cases (n=2145), n (30.59%)	Cases (n=7011), n	Cases in training set (n=5554), n	Generation method of validation set	Cases in the validation set (n=1457), n	Model type	Modeling variables
Hou et al [11]	2023	China	Prospective cohort study	Multicenter	EGVB	131	1100	999	Random sampling	101	Cox	Clinical features (gender, alcohol and smoking history, decompensation, ascites, location and size of varicose veins, ALT[n], GGT[o], HCT[p], and NLR[q] levels, and RBC[r] count)
Luo et al [12]	2023	China	Retrospective case-control study	Single center	EGVB	88	211	149	Random sampling	62	LR	Radiomics and clinical features (albumin, fibrinogen, portal vein thrombosis, aspartate aminotransferase, and spleen thickness)
Meng et al [13]	2021	China	Retrospective case-control study	Single center	EVB	39	173	121	Random sampling	52	Cox	Radiomics
Liu et al [14]	2022	China	Retrospective case-control study	Single center	EVB	69	317	222	Random sampling	95	LR	Radiomics and clinical features (hemoglobin)

XSL•FO
RenderX

First author	Year of publication	Country of authors	Study type	Patient source	Predicted events	Event cases (n=2145), n (30.59%)	Cases (n=7011), n	Cases in training set (n=5554), n	Generation method of validation set	Cases in the validation set (n=1457), n	Model type	Modeling variables
Xu et al [15]	2024	China	Case-control study	Single center	EVB	114	241	160	Random sampling	81	LR	Clinical features (spleen length and platelet distribution width PDW[s])
Lin et al [17]	2024	China	Retrospective cross-sectional study	Single center	EGVB	70	210	210	None[t]		LR	Clinical features (LSM[u] and GBS[v] scoring)
Zhang et al [18]	2023	China	Retrospective analysis	Multicenter	EVB	77	714	527	External validation	187	Cox	Clinical features (the size of varices, red wale marks, ascites, spleen thickness, γ-glutamyltransferase, and hematocrit)
Yang et al [19]	2023	China	Retrospective case-control study	Single center	EVB	98	200	140	External validation	60	LR	Clinical features (SSM[w] and LSM)
Fu et al [20]	2023	China	Case-control study	Single center	EVB	40	78	78	None		LR	Clinical features (age, gender, platelet count, Child-Pugh classification, and normalized iodine density)

XSL•FO
RenderX

First author	Year of publication	Country of authors	Study type	Patient source	Predicted events	Event cases (n=2145), n (30.59%)	Cases (n=7011), n	Cases in training set (n=5554), n	Generation method of validation set	Cases in the validation set (n=1457), n	Model type	Modeling variables
Samy et al [21]	2022	Egypt	Prospective cohort study	Single center	EVB	42	85	85	None		LR	Clinical features (main coronary vein size, paraesophageal collateral, short gastric collateral, gastric varicosities, child class C, and size of EVs >4 mm)
Salahshour et al [22]	2020	Iran	Retrospective case-control study	Single center	EVB	50	124	124	None		LR	Clinical features (presence of SGC[x], SGC size, presence of EV, and MELD[y] score)
Li et al [16]	2022	China	Retrospective case-control study	Single center	EGVB	85	142	142	None		LR	Clinical features (spleen volume expansion rate, serum sodium level, hemoglobin level, and prothrombin time)
Xu et al [23]	2014	China	Prospective cohort study	Single center	EVB	51	416	416	None		LR	Clinical features (ΔMELD, LGVV[z], and LGVBFD)[aa]

XSL•FO
RenderX

First author	Year of publication	Country of authors	Study type	Patient source	Predicted events	Event cases (n=2145), n (30.59%)	Cases (n=7011), n	Cases in training set (n=5554), n	Generation method of validation set	Cases in the validation set (n=1457), n	Model type	Modeling variables
Heo et al [24]	2019	Korea	Retrospective case-control study	Multicenter	EVB	61	262	262	None		LR	Clinical features (ARFI[ab] velocity, spleen diameter, platelet count, ASPS[ac], Child-Pugh class, EV status, portal hypertensive gastropathy, etiology of cirrhosis, age)
Liu et al [25]	2022	China	Retrospective case-control study	Single center	EVB	650	1099	823	Time-based split	276	LR	Clinical features (PLT, Hb[ad], ALB[ae], GLB[af], FBG[ag], Cl[ah], and CTPD[ai])
Yang et al [26]	2019	China	Retrospective case-control study	Multicenter	EGVB	91	295	236	Random sampling	59	RF	Radiomics and clinical features (sex, portal vein thrombosis, and portal hypertension)

XSL•FO
RenderX

First author	Year of publication	Country of authors	Study type	Patient source	Predicted events	Event cases (n=2145), n (30.59%)	Cases (n=7011), n	Cases in training set (n=5554), n	Generation method of validation set	Cases in the validation set (n=1457), n	Model type	Modeling variables
Zhang et al [27]	2023	China	Retrospective study	Single center	EGVB	65	110	77	Random sampling	33	LR	Radiomics and clinical features (PLT)

[a]EGVB: esophagogastric variceal bleeding.
[b]LR: logistic regression.
[c]Cox: Cox regression.
[d]vPPG: virtual portal pressure gradient.
[e]PLT: platelet.
[f]INR: international normalized ratio.
[g]EVB: esophageal variceal bleeding.
[h]DL: deep learning.
[i]XGBoost: Extreme Gradient Boosting.
[j]GLM: general linear regression.
[k]GBM: gradient boost machine.
[l]RF: random forest.
[m]EV: esophageal variceal.
[n]ALT: alanine aminotransferase.
[o]GGT: γ-glutamyl transferase.
[p]HCT: hematocrit.
[q]NLR: neutrophil-lymphocyte ratio.
[r]RBC: red blood cell.
[s]PDW: platelet distribution width.
[t]In the generation method of the validation set, “None” indicates that there is no independent validation set.
[u]LSM: liver stiffness measurement.
[v]GBS: Glasgow-Blatchford score.
[w]SSM: spleen stiffness measurement.
[x]SGC: short gastric.
[y]MELD: Model for End-stage Liver Disease.
[z]LGVV: left gastric vein blood flow velocity.
[aa]LGVBFD: left gastric vein blood flow direction.
[ab]ARFI: acoustic radiation force impulse.
[ac]ASPS: ARFI–spleen diameter-to-platelet ratio scores.
[ad]Hb: hemoglobin.
[ae]ALB: albumin.
[af]GLB: globulin.
[ag]FBG: fasting blood glucose.
[ah]Cl: serum chloride.
[ai]CTPD: computed tomography portal vein diameter.

ROB Assessment

The ROB in the included models was assessed. Fourteen models were from case-control studies, introducing a high ROB in participant selection. Fifteen models were built based on clinical features, resulting in a higher ROB in the predictive factors. Three models were based on radiomics, and thus, it was challenging to assess ROB. In terms of results, the included studies clearly outlined diagnostic criteria for EVB and EGVB, and no evidence of high ROB was noted in these diagnostic standards.

Of the models included, the number of cases was small in 9 training sets, resulting in a higher ROB. For 7 models, the number of cases in the validation set could not be estimated, leading to an unclear ROB. Regarding the handling of missing values, missing values were deleted in 6 models, leading to a high ROB. Only one study applied univariate analysis to select

XSL•FO
RenderX

variables for the models, and 7 studies did not provide details on their variable selection methods. Among the included models, only one model was cross-validated, and thus, a high ROB was identified in the study. The full assessment results are available in Figure 2.

Figure 2. Risk of bias assessment results of the included studies by the Prediction Model Risk of Bias Assessment Tool (PROBAST).

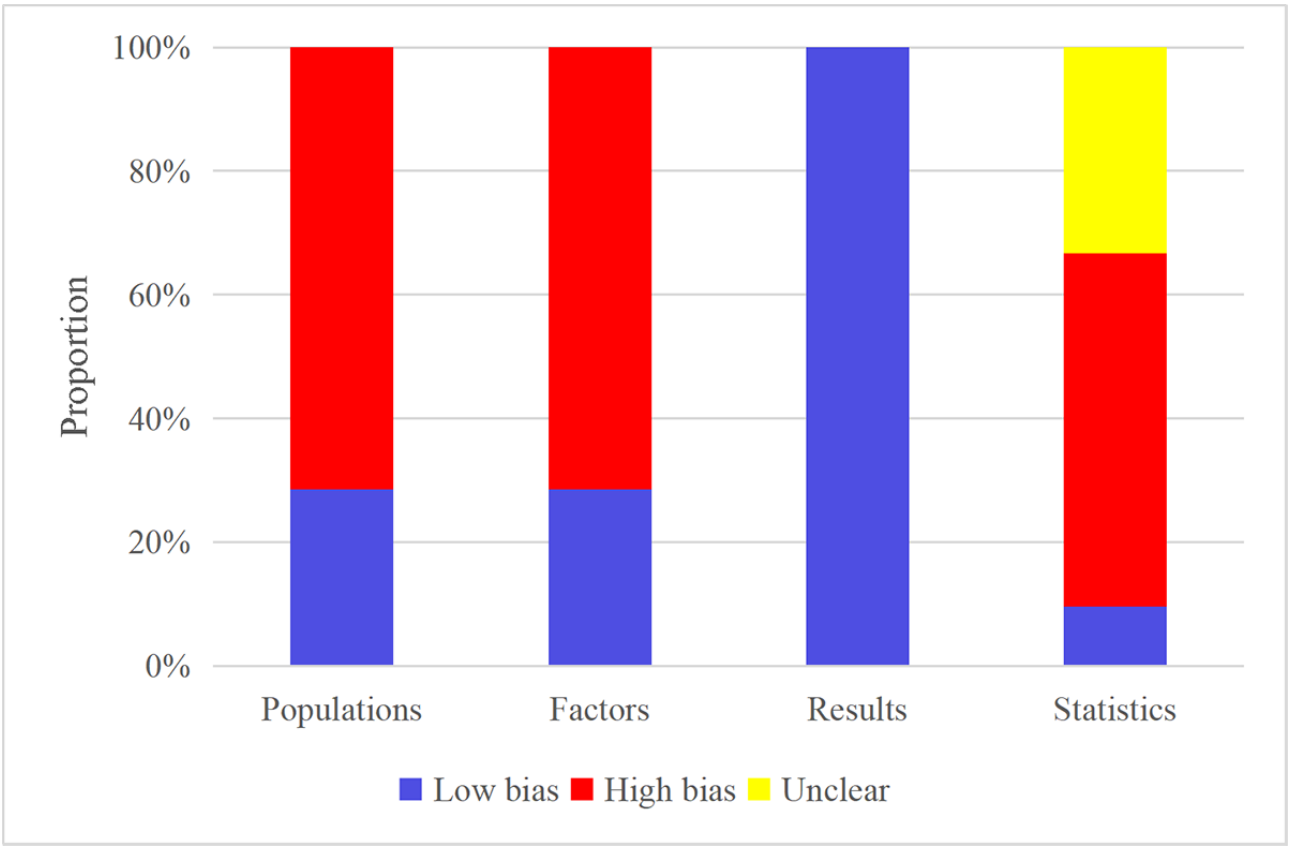

Meta-Analysis

Esophageal Variceal Bleeding

Pooled Results

Twelve models were created for predicting EVB. The meta-analysis was performed using a random-effects model. The pooled c-index, sensitivity, and specificity in the training set were 0.88 (95% CI 0.83 - 0.92), 0.93 (95% CI 0.83 - 0.97), and 0.75 (95% CI 0.65 - 0.84), respectively (Figures 3 and 4). The pooled c-index, sensitivity, and specificity in the validation set were 0.85 (95% CI 0.77 - 0.92), 0.93 (95% CI 0.87 - 0.96), and 0.66 (95% CI 0.46 - 0.82), respectively (Figures 5 and 6).

XSL•FO
RenderX

Figure 3. Forest plot for meta-analysis of c-index of machine learning for predicting esophageal variceal bleeding and esophagogastric variceal bleeding in patients with liver cirrhosis in the training set [7-9,11,12,13-16,17-19,20,21,22-24,26,27].

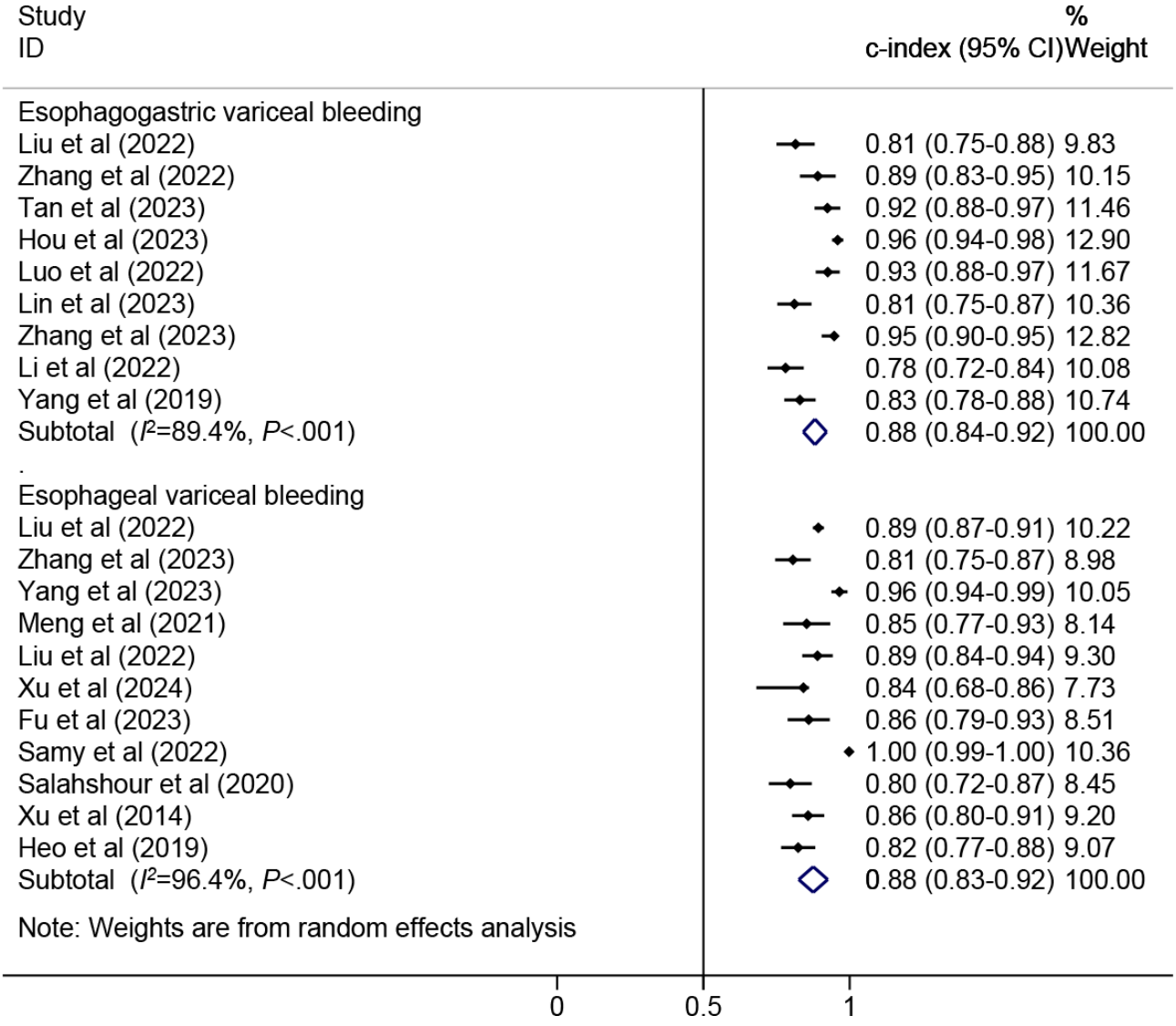

XSL•FO
RenderX

Figure 4. Forest plot for meta-analysis of sensitivity and specificity of machine learning for predicting esophageal variceal bleeding in patients with liver cirrhosis in the training set [14,15,18-25].

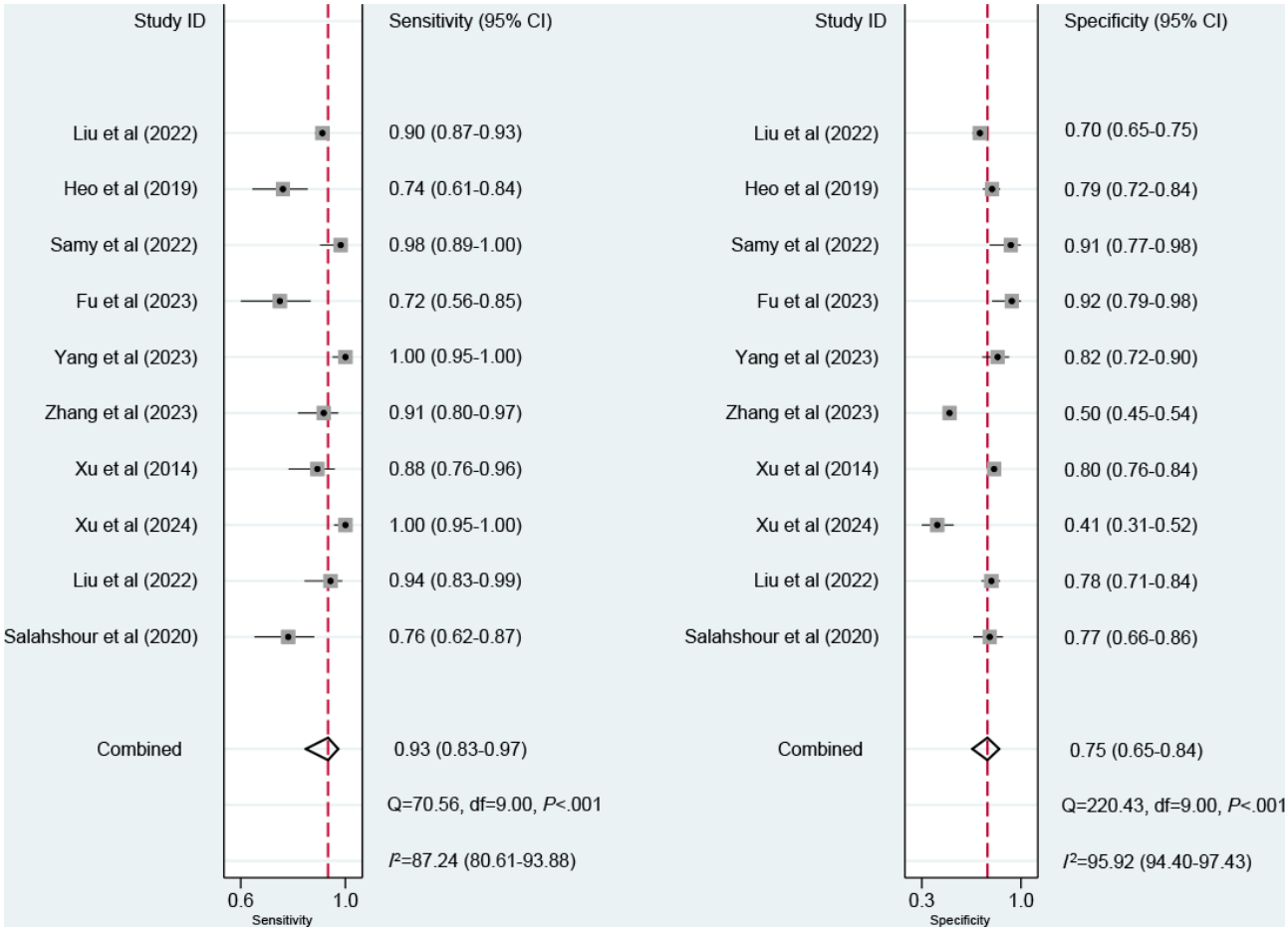

XSL•FO
RenderX

Figure 5. Forest plot for meta-analysis of c-index of machine learning for predicting esophageal variceal bleeding and esophagogastric variceal bleeding in patients with liver cirrhosis in the validation set [7-15,18,19,25-27].

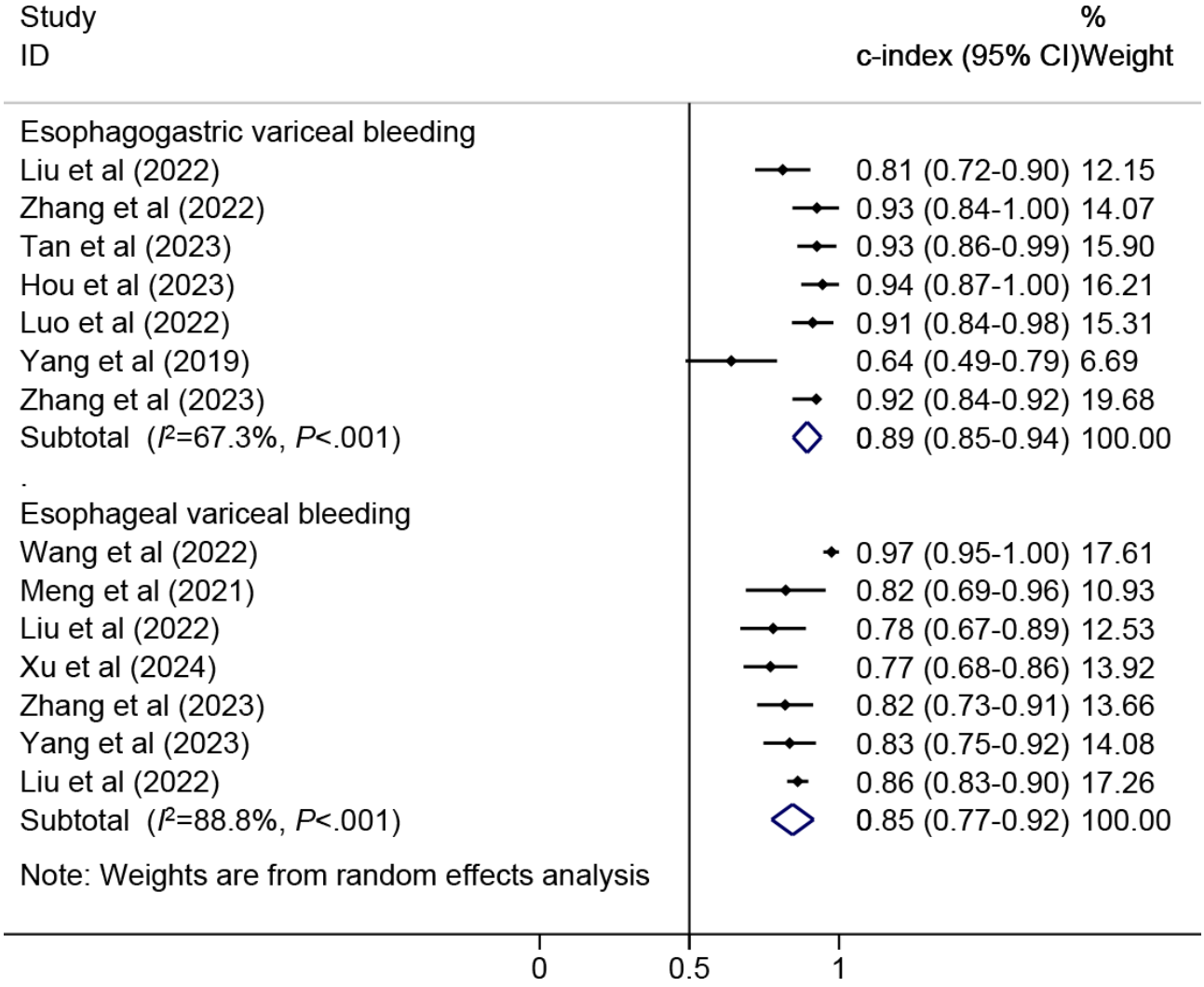

XSL•FO
RenderX

Figure 6. Forest plot for meta-analysis of sensitivity and specificity of machine learning for predicting esophageal variceal bleeding in patients with liver cirrhosis in the validation set [10,14,15,18,19].

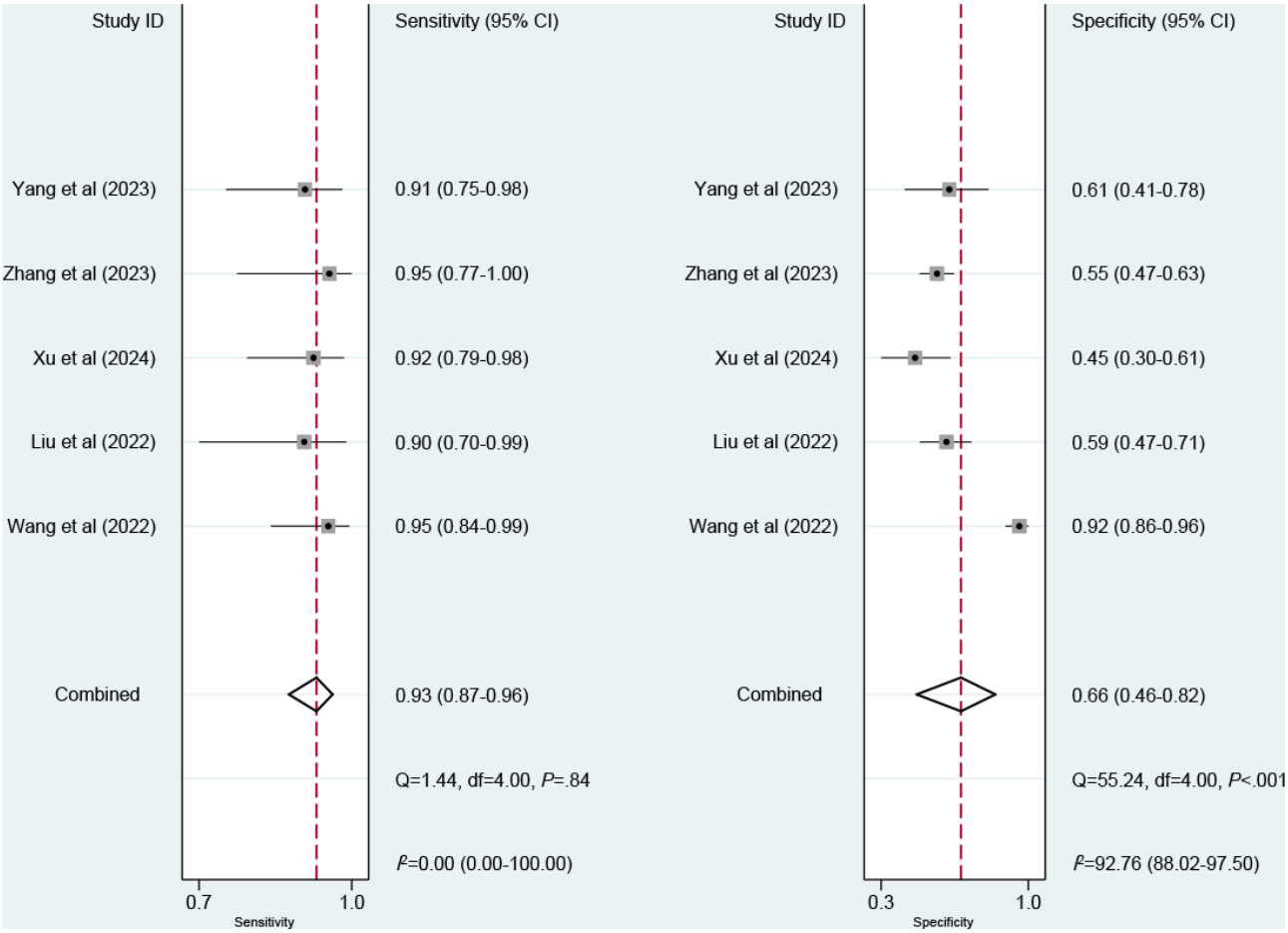

Subgroup Analysis

A subgroup analysis was performed based on modeling variables and the generation method of the validation set. In the subgroup analysis by model variables, models were grouped based on clinical features, radiomics-based models, models based on radiomics and clinical features, and models based on endoscopic features. In the training set, the c-index was 0.88 (95% CI 0.82 - 0.93) for models based on clinical features, 0.85 (95% CI 0.77 - 0.93) for radiomics-based models, and 0.89 (95% CI 0.84 - 0.94) for models based on radiomics and clinical features (Multimedia Appendix 2). In the validation set, the c-index was 0.84 (95% CI 0.80 - 0.88) for models based on clinical features, 0.82 (95% CI 0.69 - 0.96) for radiomics-based models, 0.78 (95% CI 0.67 - 0.89) for models based on radiomics and clinical features, and 0.97 (95% CI 0.95 - 1.00) for models based on endoscopic features (Multimedia Appendix 3). In the subgroup analysis by the generation method of the validation set, models were grouped as random sampling, external validation, and time-based split. The c-index was 0.78 (95% CI 0.72 - 0.85) for models based on random sampling, 0.88 (95% CI 0.77 - 1.00) for external validation, and 0.86 (95% CI 0.83 - 0.90) for time-based split (Multimedia Appendix 4).

Esophagogastric Variceal Bleeding

Pooled Results

Nine models were created for predicting EGVB. The meta-analysis was conducted using the random-effects model. The pooled results revealed that the c-index, sensitivity, and specificity in the training set were 0.88 (95% CI 0.84 - 0.92), 0.90 (95% CI 0.78 - 0.96), and 0.79 (95% CI 0.58 - 0.91), respectively (Figures 3 and 7). The pooled c-index, sensitivity, and specificity in the validation set were 0.89 (95% CI 0.85 - 0.94), 0.77 (95% CI 0.66 - 0.85), and 0.81 (95% CI 0.67 - 0.90), respectively (Figures 5 and 8).

XSL•FO
RenderX

Figure 7. Forest plot for meta-analysis of sensitivity and specificity of machine learning for predicting esophagogastric variceal bleeding in patients with liver cirrhosis in the training set [7,9,12,16,17,27].

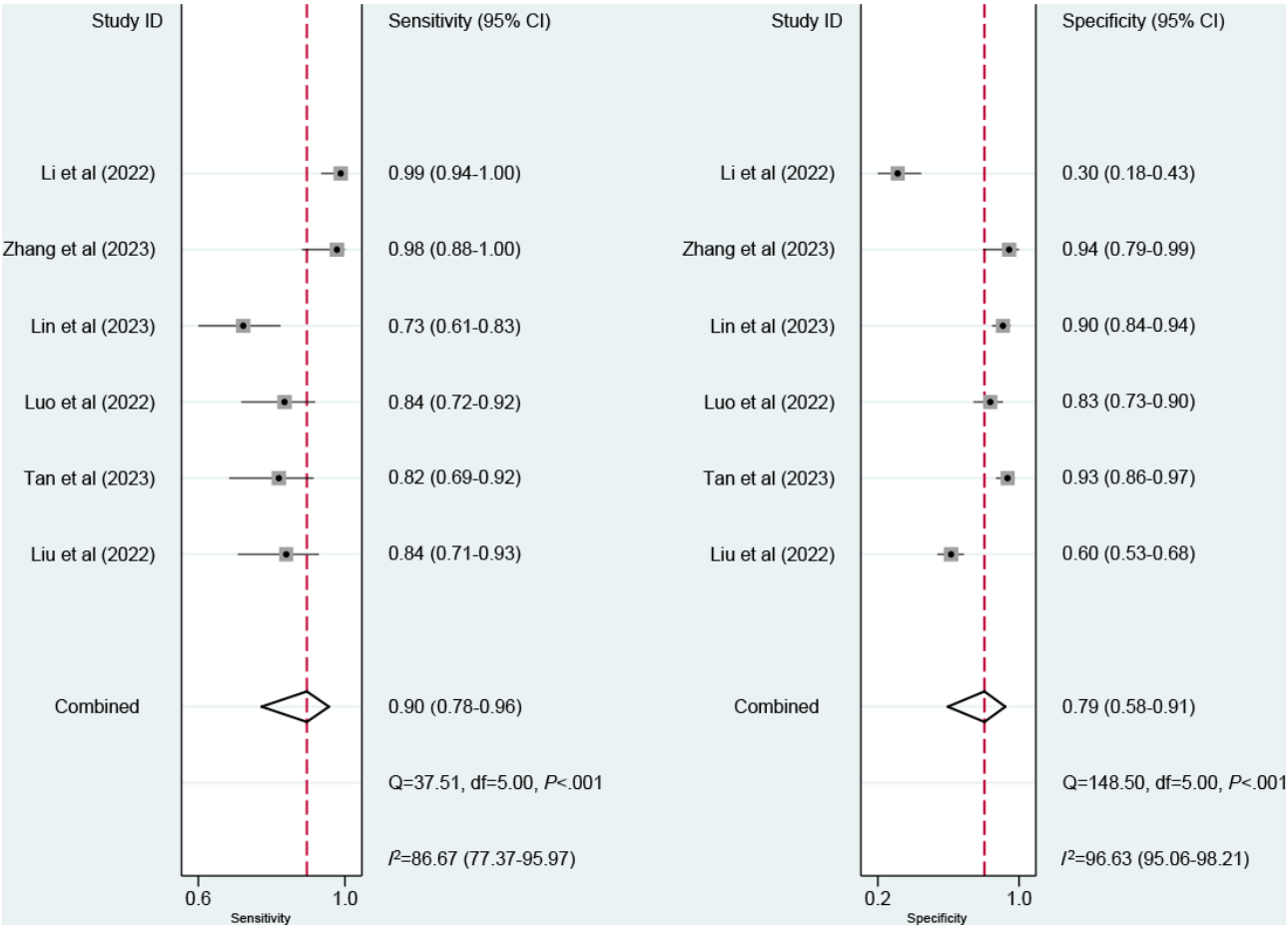

XSL•FO
RenderX

Figure 8. Forest plot for meta-analysis of sensitivity and specificity of machine learning for predicting esophagogastric variceal bleeding in patients with liver cirrhosis in the validation set [7,9,12,26,27].

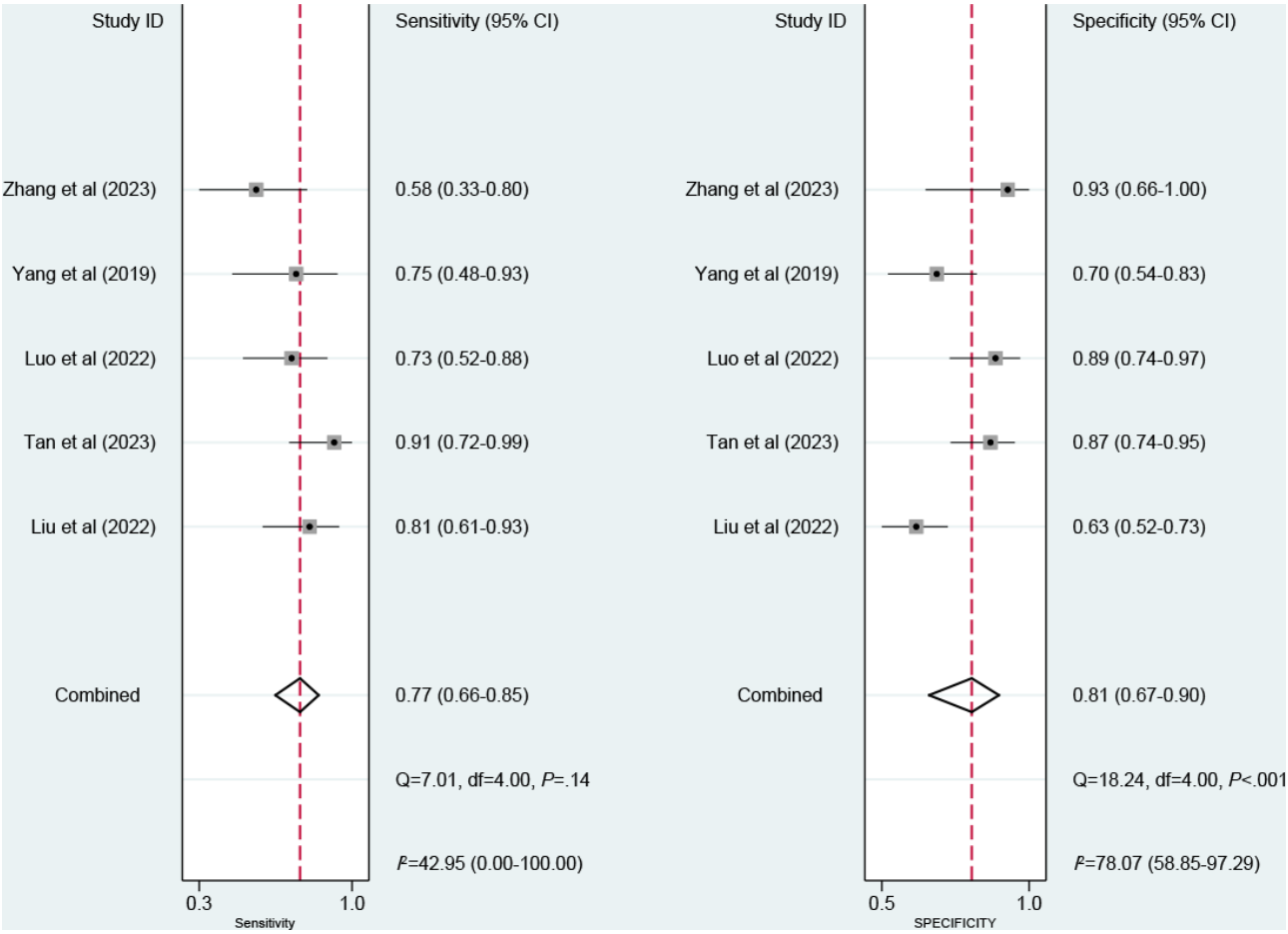

Subgroup Analysis

A subgroup analysis was performed based on modeling variables and the generation method of the validation set. In the subgroup analysis by model variables, models were grouped as models based on clinical features and models based on radiomics and clinical features. In the training set, the c-index was 0.87 (95% CI 0.80 - 0.93) for models based on clinical features and 0.90 (95% CI 0.84 - 0.97) for models based on radiomics and clinical features (Multimedia Appendix 5). In the validation set, the c-index was 0.91 (95% CI 0.86 - 0.96) for models based on clinical features, 0.85 (95% CI 0.75 - 0.96) for models based on radiomics and clinical features (Multimedia Appendix 6). In the subgroup analysis by the generation method of the validation set, models were grouped as random sampling and time-based split. The c-index was 0.89 (95% CI 0.83 - 0.94) for models based on random sampling and 0.93 (95% CI 0.85 - 1.00) for time-based split (Multimedia Appendix 7).

Discussion

Principal Findings

The risk of EVB and EGVB in patients with LC is high. Our study demonstrates that ML may be a feasible method for the early prediction of the risk of EVB and EGVB. In the validation set, the c-index, sensitivity, and specificity for ML models in the prediction of EVB were 0.85 (95% CI 0.77 - 0.92), 0.93 (95% CI 0.87 - 0.96), and 0.66 (95% CI 0.46 - 0.82), respectively. Moreover, in the validation set, the c-index, sensitivity, and specificity for ML models in the prediction of EGVB were 0.89 (95% CI 0.85 - 0.94), 0.77 (95% CI 0.66 - 0.85), and 0.81 (95% CI 0.67 - 0.90), respectively. Notably, 2 studies by Yang and Xu reported a sensitivity of 1.00 (100%) in the training set, but their specificity did not reach 1.00. Therefore, no perfect separation was observed. Furthermore, we also found that in the validation set, both studies reported sensitivities greater than 0.9 in the validation set, indicating no excessive overfitting. Therefore, this result indicates that the c-index and sensitivity of ML models are relatively high.

Comparison With Previous Studies

In patients with LC, EVB and EGVB represent a noticeable challenge that must be addressed in clinical research. Previous studies have established prediction models for EVB and EGVB. For instance, Malik et al [28] explored the use of ML for EVB or grading in patients with LC. Their results indicate that ML is highly promising for predicting esophageal varices in patients with LC and may reduce the necessity for minor surgeries. Nevertheless, their study points out that there is considerable heterogeneity and methodological limitations in their included studies, resulting in a high or unclear ROB. In the future, large-scale prospective trials are needed, and the evaluation standards for ML should be standardized to improve the practicality of these models in clinical settings.

XSL•FO
RenderX

Radiomics has recently attracted increasing attention in research on LC. In the study by Xue et al [29], transfer learning (TL) radiomics using multimodal ultrasound imaging was used for classifying the stage of liver fibrosis (LF). Their results revealed that in both the gray scale modality (GM) and elastogram modality (EM), TL exhibits superior diagnostic performance to non-TL approaches, with significantly higher areas under the receiver operating characteristic curve (AUCs; all $P<.01$). Single-modal GM and EM both outperform liver stiffness measurement (LSM) and serum indices for staging LF (all $P<.001$). The prediction performance of multimodal GM+EM is higher than GM+LSM, GM and EM alone, LSM, and biomarkers (all $P<.05$). The AUC of multimodal GM+EM for staging LF is 0.950 for S4, 0.932 for S3 or more, and 0.930 for S2 or more. Yan et al [30] used multiphase magnetic resonance imaging–based radiomics to forecast the histological grading of HCC. According to their results, 33.8% of the cases were high-grade HCCs. The levels of α-fetoprotein (odds ratio 1.89) and tumor size (>5 cm; odds ratio 2.33) are strongly associated with HCC grade. The performance of the combined model is high in predicting the grade of HCC in the test dataset (AUC=0.801), with favorable calibration and clinical utility. Park et al [31] reviewed the applications of radiomics and deep learning (DL) in liver diseases. According to their results, radiomics and DL are promising techniques for imaging-based assessment of liver diseases. Recent studies have reported the potential uses of radiomics and DL in staging LF, detecting portal hypertension, describing focal liver lesions, and predicting malignant liver tumors. Consequently, radiomics has been receiving growing attention in the clinical diagnosis and treatment of LC.

In our study, the included variables primarily consisted of interpretable clinical features and radiomic features. The accuracy of radiomics-based models for predicting EVB and EGVB in patients with LC is relatively favorable. In the included studies, 5 models were constructed based on radiomics features [12-14,26,27], among which 1 model was based on radiomics features [13] and 4 models were based on radiomics and clinical features [12,14,26,27]. The 5 studies describe the detection of bleeding in LC based on computed tomography (CT) or magnetic resonance imaging scans. Luo et al [12] developed a model based on 5 CT features from the liver and 3 CT features from the spleen. Its AUCs were 0.817 and 0.741 in the training and validation cohorts, respectively. Moreover, the clinical-radiomics model also exhibits notably predictive performance, with an AUC of 0.925 in the training cohort and 0.912 in the validation cohort. In the study by Meng et al [13], the Rad-score Liver-Spleen, based on 10 features, demonstrated strong discriminative performance. Its c-index is 0.853 (95% CI 0.776 - 0.904) in the training cohort and 0.822 (95% CI 0.749 - 0.875) in the validation cohort. According to Liu et al [14], the AUCs of their model are 0.89 (95% CI 0.84 - 0.94) and 0.78 (95% CI 0.68 - 0.87) in the training and validation datasets, respectively. These studies demonstrate that radiomics-based ML models have high predictive accuracy.

Although only a few studies on radiomic features were included in this study, their findings reveal that radiomics is essential in the diagnosis and treatment of LC. Given the small number of studies, their effect sizes were not pooled. However, the c-index of radiomics-based models was examined in the training and validation sets for both EVB and EGVB. For EVB, the c-index values were 0.85 (95% CI 0.77 - 0.93) and 0.82 (95% CI 0.69 - 0.96) for radiomics-based models in the training and validation sets, respectively. For EGVB, the c-index values were 0.90 (95% CI 0.84 - 0.97) and 0.85 (95% CI 0.75 - 0.96) for models based on radiomics and clinical features in the training and validation sets, respectively. The data indicate that radiomics appears to have a favorable performance for early prediction of EVB and EGVB. Therefore, future studies should focus on applying radiomics for the early differentiation of EVB and EGVB. However, due to the limited number of studies on radiomics, the results should be interpreted with caution. The detection performance of models constructed with traditional clinical features appears to be similar to that of models based on radiomics features. However, due to the limited number of studies on radiomics, more radiomics methods need to be explored to assess the risk of EVB and EGVB in patients with LC. Therefore, more studies need to be included in future research.

Furthermore, there was only one study based on endoscopic features. Endoscopy is broadly used in the management of EGV in patients with LC. The European Society of Gastrointestinal Endoscopy Guideline [32] offers a comprehensive overview of the role of endoscopy in the diagnosis and treatment of variceal bleeding. However, given the limited number of studies based on endoscopic images, it is impossible to assess their performance. Overall, this study demonstrates that ML has high predictive accuracy. Nonetheless, a high ROB is observed. Moreover, it is difficult to assess the ROB of models based on endoscopic features. Therefore, we believe that ML is a feasible approach to predicting EVB and EGVB in patients with LC, but it comes with a high ROB.

We conducted subgroup analyses by modeling variables, including those based on clinical characteristics and those based on imaging, to further elucidate the sources of heterogeneity. However, subgroup analyses revealed substantial heterogeneity within the subgroups. Due to the limited number of included articles, we were unable to conduct subgroup analysis by additional variables. Furthermore, subgroup analysis remains a significant challenge for meta-analyses on ML models. Given the limited number of included articles, the interpretation of differences in results across different types of ML is limited. Heterogeneity between different model types is a concern in meta-analyses of ML models. Even within the same model, differences in training rules can contribute to heterogeneity. Furthermore, the number of cases remains a potential influencing factor. Overall, this study aimed to quantitatively summarize ML models for the early prediction of esophageal or gastric variceal bleeding in LC, better demonstrating the application value of ML. However, high heterogeneity was observed. Although we attempted to address the heterogeneity from multiple perspectives, it was difficult to further quantify the impact of these differences on the results due to population differences, varying methods of assessing bleeding, and a limited number of studies. Therefore, this may impose certain

XSL•FO
RenderX

limitations on the interpretation of the results. Future research should further consider the impact of these factors on the results.

This study reveals a clear association between high AUC values and high ROB. Although several studies reported extremely high discrimination accuracy (AUC >0.90), PROBAST assessment showed that these studies generally had a high or unclear ROB. In the field of ML, near-perfect discrimination performance (AUC≈1.0) in small-sample, single-center studies is often a warning sign of methodological flaws rather than reflecting true performance. Specifically, among models with AUC ≥0.95, most were derived from single-center studies with a sample size of fewer than 200. These small sample studies with high AUC values mainly had the following problems in the PROBAST assessment: (1) the high ROB in participant selection was attributable to a retrospective design and case-control design; (2) the high ROB in analysis may be explained by the lack of appropriate validation strategy and treatment measures for overfitting; and (3) high ROB in predictors was due to insufficient transparency in variable selection and treatment. This suggests that limitations in methodology may lead to overestimation of model performance. Especially, extremely high AUC values observed in small-sample studies may reflect overfitting of the model to the training data rather than true clinical utility. This phenomenon may be due to multiple factors, such as occasional outperformance due to insufficient sample size, overfitting due to inadequate validation, and definitions of predictors and outcomes. The high ROB for the analysis domain in the PROBAST assessment (particularly the lack of measures to address overfitting) supports this explanation. Models with a c-index or AUC above 0.95 may be due to a small sample size and single-center design, with a potential risk of overfitting. Notably, high discrimination was also observed in larger, more rigorously designed multicenter trials [10].

These findings have important implications for clinical translation. First, for exceptionally high-performing models, the methodological rigor should be rigorously assessed. Second, future studies should include a sufficient sample size, perform cross-validation or external validation, and report calibration performance rather than focusing solely on discrimination. Finally, multicenter collaborations and prospective designs are crucial for generating generalizable predictive models. The discrimination performance of radiomics-based models in this study was high (AUC 0.91 - 0.93). However, given the limited number of related studies (n=5) with small sample sizes and insufficient validation, these results should be interpreted with caution.

Strengths and Limitations

This study is the first to explore ML in predicting the risk of EVB and EGVB in patients with LC, offering a strong reference for the future development or updating of intelligent tools. Nonetheless, certain limitations must be considered. First, owing to the small number of radiomics studies, this study only summarized the c-index of radiomics-based models and did not discuss their sensitivities and specificities. Hence, the results should be interpreted with caution. Second, only some of the studies explicitly described their validation sets, while others did not perform any model validation. Third, the models were primarily developed using internal validation methods, with no independent external validation, which, to some extent, restricts the interpretation of the results. Fourth, while current ML techniques demonstrate excellent sensitivity in validation datasets, the false-positive rate remains high. Future research should not only improve sensitivity but also optimize specificity. Fifth, given significant methodological and clinical heterogeneity in the current evidence, our results merely summarize the current state of application of ML. It has the potential to serve as an early predictive tool for esophageal or gastric variceal bleeding in LC. Future research should further improve the methodology and model performance to establish definitive clinical benchmarks. Sixth, although we conducted subgroup analysis based on the generation method of the validation set, there are few studies on external validation. The results of external validation need to be interpreted carefully. Seventh, there was excessive heterogeneity in the meta-analysis of sensitivity and specificity. Hence, the pooled estimates should be considered descriptive averages rather than true predictive performance. The results need to be interpreted with caution. Seventh, 18 of the 21 included studies were from China, which may limit the generalizability of the results to other regions. Therefore, the generalizability of the results still needs to be validated by more data. Finally, although our results show high AUC and c-index values, a large number of the included studies have a high ROB (particularly in the analysis domain, eg, small sample size, a lack of external validation, and potential overfitting). Hence, our results should be interpreted with caution.

Conclusions

The performance of ML-based models for predicting EVB and EGVB in patients with LC is relatively favorable, particularly those based on radiomics. However, as the number of the included original studies is limited, it is impossible to explore the impact of validation set generation methods, model types, and other factors on outcomes. Therefore, future studies should include multicenter cohorts and integrate multi-omics databases to develop ML prediction tools with improved predictive performance and broader applicability, thereby developing specialized prevention strategies to reduce the risk of EVB and EGVB in patients with LC.

Funding

The authors declare that they did not receive any funding from any source.

XSL•FO
RenderX

Authors' Contributions

YL contributed to conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing the original draft, reviewing and editing the article, visualization, supervision, and project administration. MS contributed to conceptualization, methodology, visualization, supervision, project administration, and funding acquisition. XQ contributed to conceptualization, methodology, validation, project administration, and funding acquisition; GW contributed to validation, formal analysis, and investigation. NL contributed to formal analysis, investigation, resources, and data curation. JZ contributed to resources and data curation. SS contributed to software, validation, formal analysis, investigation, writing the original draft, and reviewing and editing the article. SW contributed to software, validation, formal analysis, investigation, writing the original draft, and reviewing and editing the article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Literature search strategy.
[PNG File, 218 KB - jmir_v28i1e78203_app1.png]

Multimedia Appendix 2
Forest plot of meta-analysis of c-index of machine learning models based on different modeling variables for predicting esophageal variceal bleeding in patients with liver cirrhosis in the training set.
[PNG File, 78 KB - jmir_v28i1e78203_app2.png]

Multimedia Appendix 3
Forest plot of meta-analysis of c-index of machine learning models based on different modeling variables for predicting esophageal variceal bleeding in patients with liver cirrhosis in the validation set.
[PNG File, 80 KB - jmir_v28i1e78203_app3.png]

Multimedia Appendix 4
Forest plot of meta-analysis of c-index of machine learning models based on different methods of generating the validation set for predicting esophageal variceal bleeding patients with liver cirrhosis in the validation set.
[PNG File, 68 KB - jmir_v28i1e78203_app4.png]

Multimedia Appendix 5
Forest plot of meta-analysis of c-index of machine learning models based on different modeling variables for predicting esophagogastric variceal bleeding patients with liver cirrhosis in the training set.
[PNG File, 72 KB - jmir_v28i1e78203_app5.png]

Multimedia Appendix 6
Forest plot of meta-analysis of c-index of machine learning models based on different modeling variables for predicting esophagogastric variceal bleeding patients with liver cirrhosis in the validation set.
[PNG File, 63 KB - jmir_v28i1e78203_app6.png]

Multimedia Appendix 7
Forest plot of meta-analysis of c-index of machine learning models based on different methods of generating the validation set for predicting esophagogastric variceal bleeding patients with liver cirrhosis in the validation set.
[PNG File, 55 KB - jmir_v28i1e78203_app7.png]

Checklist 1
PRISMA checklist.
[PDF File, 613 KB - jmir_v28i1e78203_app8.pdf]

References

1. Ginès P, Krag A, Abraldes JG, Solà E, Fabrellas N, Kamath PS. Liver cirrhosis. Lancet 2021 Oct 9;398(10308):1359-1376. [doi: 10.1016/S0140-6736(21)01374-X] [Medline: 34543610]
2. Asrani SK, Devarbhavi H, Eaton J, Kamath PS. Burden of liver diseases in the world. J Hepatol 2019 Jan;70(1):151-171. [doi: 10.1016/j.jhep.2018.09.014] [Medline: 30266282]

XSL•FO
RenderX

3. Huang DQ, Terrault NA, Tacke F, et al. Global epidemiology of cirrhosis - aetiology, trends and predictions. Nat Rev Gastroenterol Hepatol 2023 Jun;20(6):388-398. [doi: 10.1038/s41575-023-00759-2] [Medline: 36977794]
4. Cherian JV, Deepak N, Ponnusamy RP, Somasundaram A, Jayanthi V. Non-invasive predictors of esophageal varices. Saudi J Gastroenterol 2011;17(1):64-68. [doi: 10.4103/1319-3767.74470] [Medline: 21196656]
5. Garcia-Tsao G, Abraldes JG, Berzigotti A, Bosch J. Portal hypertensive bleeding in cirrhosis: risk stratification, diagnosis, and management: 2016 practice guidance by the American Association for the study of liver diseases. Hepatology 2017 Jan;65(1):310-335. [doi: 10.1002/hep.28906] [Medline: 27786365]
6. Debray TP, Damen JA, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Stat Methods Med Res 2019 Sep;28(9):2768-2786. [doi: 10.1177/0962280218785504] [Medline: 30032705]
7. Liu R, Sun Y, Xu K, Shi H, Sheng S, Kong D. A histogram model to predict the risk of bleeding from oesophageal and gastric variceal rupture in cirrhosis. J Coll Physicians Surg Pak 2022 May;32(5):586-590. [doi: 10.29271/jcpsp.2022.05.586] [Medline: 35546692]
8. Zhang S, Song W, Yang B, et al. A non-invasive model based on the virtual portal pressure gradient to predict the first variceal hemorrhage in cirrhotic patients. Hepatol Int 2022 Aug;16(4):926-935. [doi: 10.1007/s12072-022-10344-5] [Medline: 35585289]
9. Tan BG, Tang Z, Ou J, et al. An improved model based on quantitative features of right liver lobe, maximum varices, and portal vein system measured on magnetic resonance imaging to predict oesophagogastric variceal haemorrhage secondary to hepatitis B-related cirrhosis. Quant Imaging Med Surg 2023 Dec 1;13(12):7741-7752. [doi: 10.21037/qims-23-353] [Medline: 38106265]
10. Wang Y, Hong Y, Wang Y, et al. Automated multimodal machine learning for esophageal variceal bleeding prediction based on endoscopy and structured data. J Digit Imaging 2023 Feb;36(1):326-338. [doi: 10.1007/s10278-022-00724-6]
11. Hou Y, Yu H, Zhang Q, et al. Machine learning-based model for predicting the esophagogastric variceal bleeding risk in liver cirrhosis patients. Diagn Pathol 2023 Feb 23;18(1):29. [doi: 10.1186/s13000-023-01293-0] [Medline: 36823660]
12. Luo R, Gao J, Gan W, Xie WB. Clinical-radiomics nomogram for predicting esophagogastric variceal bleeding risk noninvasively in patients with cirrhosis. World J Gastroenterol 2023 Feb 14;29(6):1076-1089. [doi: 10.3748/wjg.v29.i6.1076] [Medline: 36844133]
13. Meng D, Wei Y, Feng X, et al. CT-based radiomics score can accurately predict esophageal variceal rebleeding in cirrhotic patients. Front Med (Lausanne) 2021;8:745931. [doi: 10.3389/fmed.2021.745931] [Medline: 34805214]
14. Liu H, Sun J, Liu G, Liu X, Zhou Q, Zhou J. Establishment of a non-invasive prediction model for the risk of oesophageal variceal bleeding using radiomics based on CT. Clin Radiol 2022 May;77(5):368-376. [doi: 10.1016/j.crad.2022.01.046] [Medline: 35241274]
15. Xu J, Tan L, Jiang N, et al. Assessment of nomogram model for the prediction of esophageal variceal hemorrhage in hepatitis B-induced hepatic cirrhosis. Eur J Gastroenterol Hepatol 2024 Jun 1;36(6):758-765. [doi: 10.1097/MEG.0000000000002750] [Medline: 38683192]
16. Li J, Li J, Ji Q, et al. Nomogram based on spleen volume expansion rate predicts esophagogastric varices bleeding risk in patients with hepatitis B liver cirrhosis. Front Surg 2022;9:1019952. [doi: 10.3389/fsurg.2022.1019952] [Medline: 36468077]
17. Lin LY, Zeng DW, Liu YR, Zhu YY, Huang LL. Diagnostic value of liver stiffness measurement combined with risk scores for esophagogastric variceal bleeding in patients with hepatitis B cirrhosis. Eur J Radiol 2024 Apr;173:111385. [doi: 10.1016/j.ejrad.2024.111385] [Medline: 38377895]
18. Zhang Q, Niu S, Yang L, et al. A novel prognostic model for predicting the risk of first variceal hemorrhage in patients with HBV-related cirrhosis. Front Cell Infect Microbiol 2023;13:1062172. [doi: 10.3389/fcimb.2023.1062172] [Medline: 36733609]
19. Yang LB, Gao X, Li H, et al. Non-invasive model for predicting high-risk esophageal varices based on liver and spleen stiffness. World J Gastroenterol 2023 Jul 7;29(25):4072-4084. [doi: 10.3748/wjg.v29.i25.4072]
20. Fu S, Chen D, Zhang Z, Shen R. Predictive value of spectral computed tomography parameters in esophageal variceal rupture and bleeding in cirrhosis. Turk J Gastroenterol 2023 Apr;34(4):339-345. [doi: 10.5152/tjg.2023.21908] [Medline: 37089047]
21. Samy M, Gamal D, Othman MHM, Ahmed SA. Assessment of variceal bleeding in cirrhotic patients: accuracy of multi-detector computed tomography. Egypt J Radiol Nucl Med 2022 Dec;53(1):62. [doi: 10.1186/s43055-022-00738-8]
22. Salahshour F, Mehrabinejad MM, Rashidi Shahpasandi MH, et al. Esophageal variceal hemorrhage: the role of MDCT characteristics in predicting the presence of varices and bleeding risk. Abdom Radiol (NY) 2020 Aug;45(8):2305-2314. [doi: 10.1007/s00261-020-02585-5] [Medline: 32447415]
23. Xu XD, Dai JJ, Qian JQ, Pin X, Wang WJ. New index to predict esophageal variceal bleeding in cirrhotic patients. World J Gastroenterol 2014 Jun 14;20(22):6989-6994. [doi: 10.3748/wjg.v20.i22.6989] [Medline: 24944493]
24. Heo JY, Kim BK, Park JY, et al. Multicenter retrospective risk assessment of esophageal variceal bleeding in patients with cirrhosis: an acoustic radiation force impulse elastography-based prediction model. Gut Liver 2019 Mar 15;13(2):206-214. [doi: 10.5009/gnl18292] [Medline: 30602219]

XSL•FO
RenderX

25. Liu CH, Liu S, Zhao YB, et al. Development and validation of a nomogram for esophagogastric variceal bleeding in liver cirrhosis: a cohort study in 1099 cases. J Dig Dis 2022 Oct;23(10):597-609. [doi: 10.1111/1751-2980.13145] [Medline: 36400743]
26. Yang JQ, Zeng R, Cao JM, et al. Predicting gastro-oesophageal variceal bleeding in hepatitis B-related cirrhosis by CT radiomics signature. Clin Radiol 2019 Dec;74(12):976. [doi: 10.1016/j.crad.2019.08.028] [Medline: 31604574]
27. Zhang Y, Zhao Q, Wen J. Splenic CT radiomics nomogram predicting the risk of upper gastrointestinal hemorrhage in cirrhosis. Journal of Radiation Research and Applied Sciences 2023 Mar;16(1):100486. [doi: 10.1016/j.jrras.2022.100486]
28. Malik S, Tenorio BG, Moond V, Dahiya DS, Vora R, Dbouk N. Systematic review of machine learning models in predicting the risk of bleed/grade of esophageal varices in patients with liver cirrhosis: a comprehensive methodological analysis. J Gastroenterol Hepatol 2024 Oct;39(10):2043-2059. [doi: 10.1111/jgh.16645] [Medline: 38886175]
29. Xue LY, Jiang ZY, Fu TT, et al. Transfer learning radiomics based on multimodal ultrasound imaging for staging liver fibrosis. Eur Radiol 2020 May;30(5):2973-2983. [doi: 10.1007/s00330-019-06595-w] [Medline: 31965257]
30. Yan Y, Si Z, Chun C, et al. Multiphase MRI-based radiomics for predicting histological grade of hepatocellular carcinoma. J Magn Reson Imaging 2024 Nov;60(5):2117-2127. [doi: 10.1002/jmri.29289] [Medline: 38344854]
31. Park HJ, Park B, Lee SS. Radiomics and deep learning: hepatic applications. Korean J Radiol 2020 Apr;21(4):387-401. [doi: 10.3348/kjr.2019.0752] [Medline: 32193887]
32. Gralnek IM, Camus Duboc M, Garcia-Pagan JC, et al. Endoscopic diagnosis and management of esophagogastric variceal hemorrhage: European Society of Gastrointestinal Endoscopy (ESGE) Guideline. Endoscopy 2022 Nov;54(11):1094-1120. [doi: 10.1055/a-1939-4887] [Medline: 36174643]

Abbreviations

AUC: area under the receiver operating characteristic curve
CT: computed tomography
DL: deep learning
EGVB: esophagogastric variceal bleeding
EM: elastogram modality
EVB: esophageal variceal bleeding
GM: gray scale modality
HCC: hepatocellular carcinoma
LC: liver cirrhosis
LF: liver fibrosis
LSM: liver stiffness measurement
ML: machine learning
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction Model Risk of Bias Assessment Tool
ROB: risk of bias
TL: transfer learning

Edited by A Coristine; submitted 28.May.2025; peer-reviewed by H Lin, L Jiang, R Luo; revised version received 14.Feb.2026; accepted 17.Feb.2026; published 08.Apr.2026.

Please cite as:
Lian Y, Qiu X, Wang G, Liu N, Zhao J, Song S, Wang S, Sun M
Machine Learning in Predicting the Risk of Esophagogastric Variceal Bleeding Among Patients With Liver Cirrhosis: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e78203
URL: https://www.jmir.org/2026/1/e78203
doi:10.2196/78203

© Yuan Lian, Xinping Qiu, Geli Wang, Nannan Liu, Jing Zhao, Shanshan Song, Shiqi Wang, Mingjun Sun. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Role of Digital Biomarkers in Physiological Signal-Based Depression Assessment: Systematic Review and Meta-Analysis

Hyeongsuk Lee[1], RN, PhD; Seung-Gul Kang[2], MD, PhD; SeonHeui Lee[1], RN, PhD

[1]Research Institute of AI and Nursing Science, College of Nursing, Gachon University, 191 Hambangmoe-ro, Yeonsu-gu, Incheon, Republic of Korea
[2]Department of Psychiatry, Gil Medical Center, Gachon University College of Medicine, Incheon, Republic of Korea

Corresponding Author:
SeonHeui Lee, RN, PhD
Research Institute of AI and Nursing Science, College of Nursing, Gachon University, 191 Hambangmoe-ro, Yeonsu-gu, Incheon, Republic of Korea

Abstract

Background: Digital biomarkers are increasingly being used to support depression assessment by providing objective, continuous, and real-time physiological and behavioral data. However, most existing studies have focused on individual biomarkers, such as sleep or cardiac parameters, while integrative evaluations that capture the multidimensional nature of depression remain limited.

Objective: This systematic review evaluated digital biomarkers for depression and synthesized evidence on differences between individuals with depression and controls.

Methods: Eligible studies included observational or interventional studies examining digital biomarkers for depression with validated outcome measures. We searched major international and Korean databases, including MEDLINE, PsycINFO, CINAHL, IEEE Xplore, Web of Science, Cochrane Library, KISS, RISS, KMbase, and KoreaMed, from inception to December 28, 2025. Risk of bias was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 tool and the Scottish Intercollegiate Guidelines Network checklist. Meta-analyses were conducted using random-effects models with the Hartung-Knapp-Sidik-Jonkman method, and other outcomes were narratively summarized.

Results: The search yielded 39,617 records, of which 132 studies involving 57,852 participants met the inclusion criteria. These studies encompassed various digital biomarkers, including sleep, physical activity, cardiac measures, smartphone-derived data, speech, GPS data, and circadian rhythms. A meta-analysis of 22 studies (6947 participants) revealed that individuals with depression had significantly longer sleep onset latency (5 studies; n=292; +4.75 min, 95% CI 2.46-7.04; *P*=.005; 95% prediction interval [PI] 0.01-10.27) and time in bed (3 studies; n=236; +31.81 min, 95% CI 18.22-45.39; *P*=.01; 95% PI 2.28-55.16). Physical activity counts were also significantly lower (5 studies; n=462; standardized mean difference −0.71, 95% CI −1.33 to −0.09; *P*=.03; 95% PI −2.18 to 0.71). Although individuals with depression showed a lower sleep efficiency, higher mean heart rate, and lower SD of normal-to-normal intervals, these differences were not statistically significant. Other digital markers yielded inconsistent results. Overall, these findings indicate that no single digital biomarker sufficiently captures depression-related changes. Instead, the results support the superiority of personalized, multimodal approaches. However, the generalizability of these findings is limited by the lack of standardized data collection protocols and high clinical heterogeneity across studies, as reflected in wide PIs.

Conclusions: Certain digital biomarkers, particularly sleep onset latency and physical activity counts, showed consistent average differences between the depression and control groups. However, wide PIs indicate substantial variability across settings, suggesting that no single marker is sufficient for reliable detection. This study advances the field by providing a comprehensive meta-analysis of multidimensional digital biomarkers, establishing a quantitative foundation for objective depression screening and monitoring. These findings support the use of personalized, multimodal digital phenotyping approaches and highlight the need for standardized, clinically interpretable frameworks for real-world depression monitoring.

Trial Registration: PROSPERO CRD42024518136; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024518136

(J Med Internet Res 2026;28:e76432) doi:10.2196/76432

KEYWORDS

digital biomarkers; depression; wearable electronic devices; sleep quality; ambulatory monitoring

XSL•FO
RenderX

Introduction

Depression is a significant mental health issue that affects over 300 million people worldwide and is one of the leading causes of disability-adjusted life years [1]. However, the current diagnostic process for depression largely relies on self-reported questionnaires and subjective clinical judgment, raising concerns regarding its accuracy and consistency [2]. These "snapshot" evaluations often fail to capture the dynamic, fluctuating nature of depressive symptoms in real-world settings, leading to delayed interventions and suboptimal treatment outcomes. To bridge this gap, digital biomarkers have emerged as a transformative objective approach, enabling the moment-by-moment quantification of individual-level human phenotypes in situ using data from personal digital devices [3,4].

Digital biomarkers, derived from smartphones, wearables, and ambient sensors, provide continuous, noninvasive, and high-frequency longitudinal data [3,5]. These markers encompass a wide range of clinical dimensions, including sleep patterns, physical activity levels, heart rate variability (HRV), vocal characteristics, and social interaction data. Recent advancements in sensor technology have significantly enhanced the precision of these metrics, offering unprecedented insights into the physiological and behavioral underpinnings of mood disorders [6,7]. These digital biomarkers possess unique properties that indicate their potential to complement or even replace traditional subjective methods of diagnosing depression. However, as the field matures, a critical challenge has surfaced regarding the inconsistency across various research findings [8,9].

While numerous studies have identified potential biomarkers, significant variability in hardware, data collection duration, and analytical pipelines has led to fragmented findings, raising concerns regarding their reproducibility and generalizability [10-12]. Although attempts have been made to conduct systematic reviews focused on digital biomarkers related to depression, these efforts have often been hindered by issues, such as the limited number of related studies or data heterogeneity, making meta-analyses infeasible [4]. Consequently, existing review reports are confined to specific biomarkers (eg, HRV or sleep data) or merely categorize and summarize findings without deeper integration [13,14]. Given that depression is a systemic disorder characterized by complex interactions between biological rhythms and behavioral shifts, isolated metrics are insufficient to capture its full multidimensionality. The lack of comprehensive systematic reviews poses a critical barrier to understanding the clinical utility and practical implementation of digital biomarkers.

There is, therefore, an urgent clinical and scientific need for a comprehensive, multimodal meta-analysis. Such an investigation is essential to distinguish "robust indicators" from "context-specific noise" and to establish the pooled effect sizes necessary for developing reliable diagnostic algorithms. By synthesizing evidence across diverse domains, including sleep, physical activity, cardiac measures, smartphone usage, speech, GPS, and circadian parameters, this systematic review aims to provide a comprehensive evaluation of the digital biomarker landscape. Specifically, this systematic review performs meta-analyses to quantify group differences between individuals with depression and controls without depression, thereby strengthening the evidence base for personalized, preemptive depression management in the digital health era.

Methods

Study Design and Registration

The protocol was prospectively registered in the PROSPERO database (CRD42024518136) and prepared following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) protocol checklist (Checklist 1).

Ethical Considerations

This systematic review was exempted by the Institutional Review Board (IRB) of Gachon University (IRB number: 1044396 - 202403 HR-052-01).

Eligibility Criteria

The inclusion criteria were as follows: (1) inclusion of participants with depression, (2) use of digital biomarkers to assess depression severity, and (3) reporting diagnostic concordance with validated assessment tools. The exclusion criteria were as follows: (1) non-English or non-Korean articles, (2) duplication, (3) inaccessible full text, and (4) reviews or qualitative studies.

Information Sources and Search Strategy

This systematic review involved a search of major academic databases, including the Cochrane Library (Wiley), MEDLINE (Ovid), PsycINFO (Ovid), CINAHL (EBSCOhost), IEEE Xplore (IEEE), Web of Science (Clarivate), and Korean academic databases, such as KISS, RISS, KMbase, and KoreaMed. The final search of all sources was conducted on December 28, 2025. Trial registries, gray literature, and author contacts were omitted as the primary search provided sufficient data. Additionally, we manually screened reference lists and removed duplicates using EndNote 20.

The search strategy combined Medical Subject Headings (MeSH) and free-text keywords related to depression and digital biomarkers. These terms were adapted for each database to maximize search sensitivity. The key terms included "depressi*," "MDD," "wearable," "application," "smartwatch," "biomarker*," "sleep*," "speech," "behavioral parameter*," "electroencephalogram," and "electrocardiogram." The search process followed the PRISMA Search Strategy (PRISMA-S) extension [15]. The full search strategy, including specific search strings, limits applied, and the number of records retrieved per database, is provided in Multimedia Appendix 1.

Selection Process

Two independent reviewers screened the titles and abstracts after removing duplicates. Potentially relevant studies and manually identified records underwent full-text assessment. Any disagreements were resolved through consensus with a third reviewer.

XSL•FO
RenderX

Data Collection Process and Data Extraction

Data were extracted by 2 independent reviewers using standardized forms. Extracted characteristics included author, year, country, study design, number of participants, age, sex, biomarker measurement device, measurement period, depression indicators, and analytical methods.

Study Risk of Bias Assessment

Quality was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool to evaluate diagnostic accuracy and the Scottish Intercollegiate Guidelines Network (SIGN) tool for case-control studies. All eligible studies were included irrespective of their quality scores.

Effect Measures and Synthesis Methods

A meta-analysis was performed on quantitatively synthesizable digital biomarkers, and a systematic review was conducted on other biomarkers. All statistical analyses were performed in R software (version 4.5.0; R Foundation for Statistical Computing) using the "meta" and "pimeta" packages. Pooled effects, expressed as mean differences (MDs) or standardized mean differences (SMDs), were estimated using a random-effects model with the Hartung-Knapp-Sidik-Jonkman (HKSJ) method [16,17]. Restricted maximum likelihood was used for τ^2 estimation. Results have been presented with 95% CIs. To account for the distribution of true effects across different settings, 95% prediction intervals (PIs) were additionally calculated using the parametric bootstrap approach proposed by Nagashima et al [18], which is robust for small study numbers.

Parameters that could not be meta-analyzed due to reporting inconsistencies or insufficient data were narratively synthesized. These included specific sleep measures (eg, sleep fragmentation, rapid eye movement [REM] sleep, REM latency, and slow-wave sleep [SWS]); physical activity (eg, light physical activity [LPA] and energy expenditure); cardiac HRV indices (eg, the root mean square of successive differences [RMSSD], low frequency [LF], high frequency [HF], and LF/HF ratio); and parameters related to speech, GPS, and circadian rhythms.

Reporting Bias Assessment and Certainty Assessment

Reporting bias was considered qualitatively, as a formal statistical assessment (eg, funnel plot) was infeasible due to fewer than 10 studies per outcome. This was supported by comprehensive database searches and manual screening of reference lists. The certainty of evidence was assessed using the GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) approach, considering risk of bias, inconsistency, imprecision, and indirectness [19].

Results

Study Selection

The initial search yielded 39,617 studies. After removing duplicates and excluding records automatically identified as ineligible through journal-type filtering, 21,915 studies remained. Following the screening of titles and abstracts, 21,649 studies were excluded. Before conducting a full-text review, a manual search through reference lists and citation tracking of relevant systematic reviews identified 17 additional studies. Consequently, full-text reviews were conducted on 283 articles, of which 132 studies involving 57,852 participants met the inclusion criteria (Figure 1).

XSL•FO
RenderX

Figure 1. Flow diagram of the study selection process for examining the role of digital biomarkers in depression assessment.

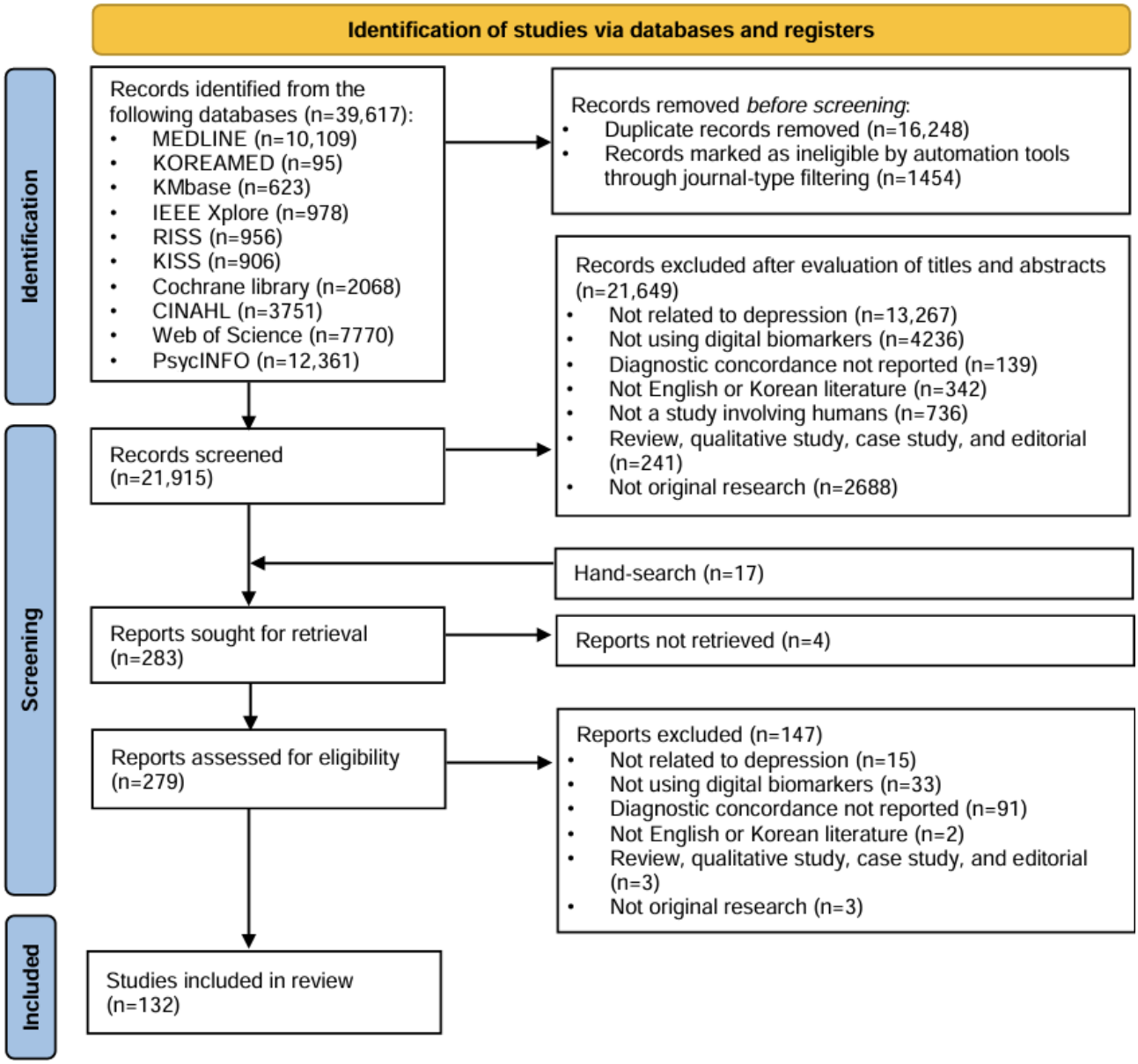

Study Characteristics

The characteristics of the included studies are summarized in Tables S1 and S2 in Multimedia Appendix 2. The included studies were categorized based on the types of digital biomarkers investigated. A total of 87 studies used single parameters, including sleep parameters (n=23) [20-42], cardiac parameters (n=19) [43-61], physical activity parameters (n=16) [62-77], smartphone-based parameters (n=9) [78-86], speech parameters (n=10) [87-96], circadian rhythm parameters (n=7) [97-103], electroencephalogram parameters (n=2) [104,105], and video-based parameters (n=1) [106]. Furthermore, 45 studies used multiple parameters to assess digital biomarkers [4,107-150]. Of the 132 studies, 63 (47.7%) used wearable devices to continuously measure biomarkers in daily life, typically for more than a week. Study designs encompassed large-scale cohorts and group comparisons (eg, depression vs control). Analytical approaches varied, including regression analyses for associations, evaluations of diagnostic accuracy, and assessments of group differences (Multimedia Appendix 3).

Validated depression assessment tools were used, with the most frequent being the Patient Health Questionnaire (PHQ), followed by the Hamilton Depression Rating Scale (HDRS), Beck Depression Inventory (BDI), Geriatric Depression Scale (GDS), Korean version of the GDS (SGDS-K), and Center for Epidemiologic Studies Depression Scale (CES-D) (Multimedia Appendix 4). Biomarker categories and their detailed features are illustrated in Multimedia Appendix 5 and Table S3 in Multimedia Appendix 2.

Risk of Bias in Studies

Based on the SIGN assessment, all 73 case-control studies demonstrated acceptable internal validity with clearly defined groups. However, participation rates for each group showed substantial variability, ranging from 20.1% to 100%, and many studies (70/73, 96%) did not report comparisons between participants and nonparticipants. Additionally, potential

XSL•FO
RenderX

confounding factors were insufficiently addressed in several studies (Table S4 in Multimedia Appendix 2).

Among diagnostic accuracy studies assessed using QUADAS-2, most studies (58/59, 98%) showed a low risk of bias in the index test, reference standard, and flow and timing domains. In contrast, great concern regarding applicability was identified because study populations often did not align with intended clinical targets, limiting the generalizability of the findings to real-world settings (Table 1).

XSL•FO
RenderX

Table . Study quality assessment using QUADAS-2[a].

First author	Year	Risk of bias[b]				Applicability concerns[b]		
		Patient selection	Index test	Reference standard	Flow and timing	Patient selection	Index test	Reference standard
Peng [21]	2023	1	1	1	1	1	1	1
Hasanzadeh [28]	2020	1	1	1	1	1	1	1
Ding [30]	2019	1	1	1	1	1	1	1
Coutts [50]	2020	2	1	1	1	2	1	1
Roh [55]	2014	1	1	1	1	2	1	1
Zhang [57]	2012	2	1	1	1	2	1	1
Espino-Salinas [63]	2022	1	1	1	1	2	1	1
Hsueh [70]	2021	1	1	1	1	1	1	1
Jakobsen [71]	2020	1	1	1	1	1	1	1
Zhao [72]	2019	1	1	1	1	2	1	1
Ku [74]	2018	1	1	1	1	1	1	1
Fadul [78]	2023	1	1	1	1	2	1	1
Auerbach [79]	2022	1	1	1	1	1	1	1
Otte Andersen [80]	2022	1	1	1	1	2	1	1
Opoku Asare [81]	2021	2	2	1	1	1	1	1
Chikersal [82]	2021	2	1	1	1	2	1	1
Zhang [83]	2021	1	1	1	1	1	1	1
Pedrelli [84]	2020	2	1	1	1	2	1	1
Mastoras [85]	2019	1	1	1	1	2	1	1
Saeb [86]	2015	2	1	1	1	2	1	1
Wiseman [87]	2025	1	1	1	1	1	1	1
Kim [90]	2023	1	1	1	1	1	1	1
Zhao [92]	2022	1	1	1	1	1	1	1
Ye [93]	2021	1	1	1	1	1	1	1
Klangpornkun [94]	2021	1	2	1	1	1	2	1
Demiroglu [95]	2020	1	1	1	1	1	1	1
Yamamoto [96]	2020	1	1	1	1	1	1	1
Choi [99]	2021	1	1	1	1	1	1	1
Anik [104]	2024	2	1	1	1	2	1	1
Tian [105]	2025	1	1	1	1	1	1	1
Islam [106]	2024	1	1	1	1	1	1	1
Makhmutova [107]	2022	1	1	1	1	1	1	1
Price [108]	2024	1	1	1	1	1	1	1
Griffiths [110]	2022	1	1	1	1	2	1	1
Rykov [113]	2021	2	1	1	1	2	1	1

XSL•FO
RenderX

First author	Year	Risk of bias[b]				Applicability concerns[b]		
		Patient selection	Index test	Reference standard	Flow and timing	Patient selection	Index test	Reference standard
Ahmed [119]	2022	2	1	1	1	2	1	1
Choi [120]	2022	2	1	1	1	2	1	1
Mahendran [121]	2019	2	1	1	1	2	1	1
Xu [122]	2019	2	1	1	1	2	1	1
Lu [123]	2018	2	1	1	1	2	1	1
Farhan [124]	2016	1	1	1	1	2	1	1
Karimi [125]	2025	1	1	1	1	1	1	1
Jacobson [126]	2020	2	1	1	1	2	1	1
Wang [127]	2018	1	1	1	1	2	1	1
Zhou [129]	2022	1	1	1	1	1	1	1
Williamson [130]	2019	1	1	1	1	2	1	1
Asare [132]	2022	1	1	1	1	2	1	1
Narziev [133]	2020	1	1	1	1	2	1	1
Kim [134]	2019	1	1	1	1	2	1	1
Di Matteo [137]	2021	1	1	1	1	1	1	1
Sverdlov [138]	2021	1	1	1	1	1	1	1
Minaeva [140]	2020	1	1	1	1	1	1	1
Dai [141]	2022	1	1	1	1	1	1	1
Tazawa [142]	2020	1	1	1	1	1	1	1
Thati [143]	2023	2	1	1	1	2	1	1
Jiang [144]	2024	1	1	1	1	1	1	1
Bai [148]	2021	2	1	1	1	2	1	1
Borelli [149]	2025	1	1	1	1	1	1	1
Chen [150]	2024	1	1	1	1	1	1	1

[a]QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2.

[b]1: low risk; 2: high risk.

Results of Individual Studies and Syntheses

Meta-analysis results are presented in Figure 2. A detailed overview of the parameters that could not be quantitatively synthesized and their reported associations with depression is provided in Multimedia Appendix 6.

XSL•FO
RenderX

Figure 2. Forest plots of sleep-related parameters (A: total sleep time [20,33,109,115-117,147,150]; B: sleep efficiency [20,33,39,115-117,147,150]; C: wake after sleep onset [20,33,39,115,116]; D: sleep onset latency [20,39,40,115,116]; E: time in bed [20,115,132]), physical activity–related parameters (F: step counts [69,132,147]; G: physical activity counts [20,76,111,134,150]; H: sedentary time [64,111,150]; I: moderate-to-vigorous activity [64,69,109,147]), and cardiac parameters (J: mean heart rate [43,52,54,58,61]; K: SD of normal-to-normal intervals [58,61,114,147]). Pooled estimates have been calculated using a random-effects model with the Hartung-Knapp-Sidik-Jonkman adjustment. The between-study variance (^{2}) has been estimated via the restricted maximum-likelihood method. 95% prediction intervals (PIs) have been derived using the Nagashima parametric bootstrap method.

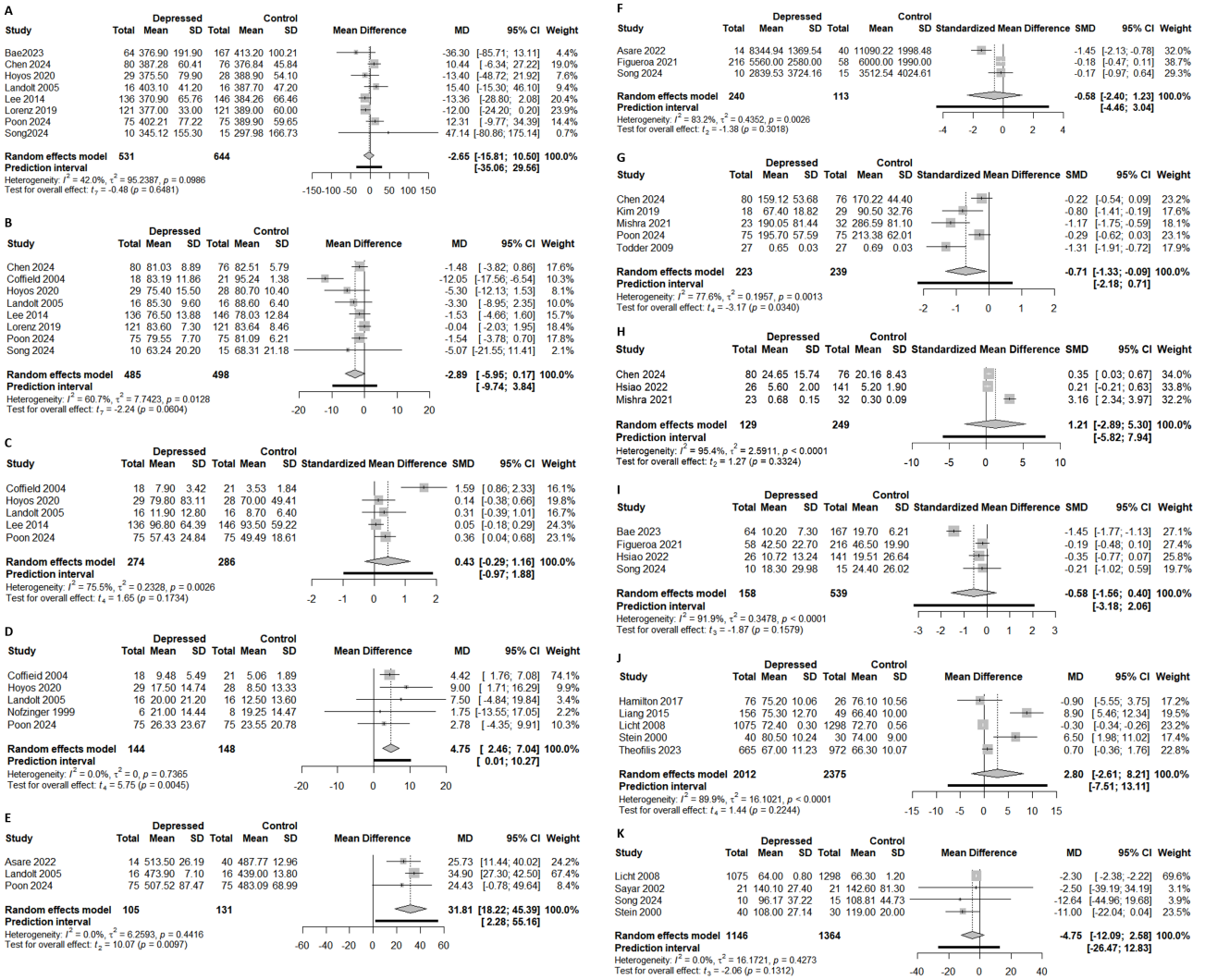

Sleep-Related Parameters and Depression

Total Sleep Time

Total sleep time (TST) showed no significant difference between the depression and control groups (−2.65 min, 95% CI −15.81 to 10.50; *P*=.65; 95% PI −35.06 to 29.56; Figure 2A). Some studies suggested an association between TST and depressive symptoms [109,112,131,135], with TST identified as the most predictive variable in 21% of patients in a predictive model [131]. Treatment studies also found a negative correlation between reduced depressive symptoms and longer TST [112]. Conversely, other studies [4,20,22,36,41,115,132,136,147,150], including those focusing on older adults and medication-treated patients [26,116], reported no significant differences.

Sleep Efficiency

Sleep efficiency (SE) was lower in the depression group but not statistically significant (−2.89%, 95% CI −5.95 to 0.17; *P*=.06; 95% PI −9.74 to 3.84; Figure 2B). Several studies reported an improvement in SE with symptom reduction [35,39] and negative correlations across diverse populations [24,26,38,148]. However, some studies found no significant differences [20,115-117,147,150] or considered SE to be a nonsignificant predictor after adjustment [36,37]. Overall, SE tended to be lower in patients with depression, although the significance varied according to the analytical method.

Wake After Sleep Onset

Wake after sleep onset (WASO) tended to be longer in the depression group, though the pooled effect was not statistically significant (SMD 0.43, 95% CI −0.29 to 1.16; *P*=.17; 95% PI −0.97 to 1.88; Figure 2C). Previous evidence has been mixed: several studies reported prolonged WASO in the depression group [20,39,132], with 1 showing posttreatment improvement [34], while others found no consistent differences [33,115,116].

Sleep Fragmentation

Sleep fragmentation and frequent awakenings were identified as predictors of depression [24,39,110].

XSL•FO
RenderX

Sleep Onset Latency

Sleep onset latency (SOL) was significantly longer in the depression group (4.75 min, 95% CI 2.46-7.04, *P*=.005; 95% PI 0.01-10.27; Figure 2D). Most studies supported this trend [35,36,116], although some reported no significant differences in the adjusted models or after symptom resolution [34,39]. Overall, a prolonged SOL was consistently associated with depressive symptoms.

Time in Bed

Time in bed (TIB) was significantly longer in the depression group (31.81 min, 95% CI 18.22-45.39; *P*=.01; 95% PI 2.28-55.16; Figure 2E). Most studies corroborated this finding, with linear mixed models consistently associating longer TIB with depressive symptoms [26,115,132,145]. Although the 95% PI remained entirely above zero, its wide range suggests that the extent of TIB prolongation may vary substantially across individuals or settings.

REM Sleep and REM Latency

Most studies on REM sleep relied on small hospital-based samples, limiting generalizability. One study found that REM sleep was significantly reduced in treated patients, particularly during the first third of the night [41]. While some research identified REM sleep as a potential predictor of depression [110], others found no group differences [115,116] or explanatory value [26,31]. Findings on REM latency were also inconsistent; 1 study showed that antidepressant-treated patients had increased REM latency [41], whereas 2 studies found no difference between the depression and control groups [33,115]. In older adults, REM latency was reported to be comparable or longer [116]. Longitudinal analyses suggested that SOL may serve as a more reliable marker than REM latency [26].

Non–Rapid Eye Movement Sleep and SWS

Evidence on non–rapid eye movement (NREM) and SWS remains limited. While 1 study found a significant association between NREM sleep and depression [26], 2 studies reported no such relationship [41,116]. Similarly, no studies to date have identified a significant association between SWS and depression [33,41,115,116].

Sleep Onset, Midpoint, and Offset

Findings regarding sleep timing (onset, midpoint, and offset) varied. Several studies, particularly those using actigraphy, reported delayed timing in individuals with depression [23,112,113,116,150]. For instance, a study of hospitalized patients showed a weak negative correlation between sleep onset time and depression (r=−0.381), indicating a delay before discharge [112]. Other research in working adults and middle-aged women found later sleep midpoints and offsets associated with higher depression scores [23,113,116]. However, these results remain inconsistent across measurement methods and study populations [26,116,127,136,148].

Physical Activity Parameters and Depression

Step Counts

The meta-analysis showed no significant difference in daily step counts between the depression and control groups (SMD −0.58, 95% CI −2.40 to 1.23; *P*=.30; 95% PI −4.46 to 3.04; Figure 2F), although the point estimate suggested fewer steps for the depression group. Longitudinal studies indicated that higher step counts were associated with reduced depression severity, with counts tending to increase during recovery [65,142]. In older adults, more daily steps were associated with a lower risk of future depressive symptoms [63]. However, some studies reported a limited predictive value of step counts [145,147].

Physical Activity Counts

Physical activity counts were lower in the depression group (SMD −0.71, 95% CI −1.33 to −0.09; *P*=.03; 95% PI −2.18 to 0.71; Figure 2G). This finding is consistent with actigraphy-based studies reporting reduced activity levels in individuals with depression [4,38,63,71,149,150]. Strong negative correlations with symptom severity were reported, particularly in hospitalized patients [112]. Longitudinal analyses showed that activity counts decreased during depressive episodes and increased with symptom improvement, supporting their relevance as state-dependent markers [4]. Notably, daytime activity levels were sensitive to clinical improvement, whereas nighttime activity measures did not reflect symptom changes [76]. Despite these promising results, the wide PI encompassing zero underscores the need for standardized approaches to account for potential inconsistencies in future observations.

Sedentary Time

Sedentary time showed no significant difference between the depression and control groups (SMD 1.21, 95% CI −2.89 to 5.30; *P*=0.33; 95% PI −5.82 to 7.94; Figure 2H). Findings regarding sedentary time were inconsistent. While some studies linked increased sedentary time to depression [66], others found no association after adjusting for moderate-to-vigorous physical activity (MVPA) or overall activity levels [74,135,149]. Longitudinal and intervention-based analyses further suggested that sedentary time alone is a relatively weak and nonspecific marker of depressive symptoms compared with mobility-related indicators such as activity counts or homestay [136]. This limitation may be particularly evident in older populations, where prolonged sedentary time may occur regardless of the depressive status [64,111].

MVPA Findings

The meta-analysis showed no significant difference in MVPA between the depression and control groups (SMD −0.58, 95% CI −1.56 to 0.40; *P*=.16; 95% PI −3.18 to 2.06; Figure 2I). However, individual studies suggested a possible link, with higher MVPA associated with reduced depression severity [68], particularly in older adults [109]. Notably, 1 study observed significant reductions in MVPA on weekends among individuals with depression [69].

LPA Findings

The results of LPA were mixed. Some studies reported that a greater time spent in LPA was associated with fewer depressive symptoms [68,74], whereas others found no significant association [110,111,123,147,149]. Overall, LPA may be more effective when interpreted alongside MVPA or step data.

XSL•FO
RenderX

Energy Expenditure

Energy expenditure, measured in kilocalories or metabolic equivalents of tasks (METs), has shown mixed findings. One study reported significantly lower energy expenditure in patients with depression than in controls [132], and a longitudinal study found increases in energy expenditure with symptom improvement [142]. Conversely, another study observed no group differences [117]. Further studies are required to assess its value as a biomarker.

Cardiac Parameters and Depression

Time Domain

Mean Heart Rate

The meta-analysis showed a higher mean heart rate (mHR) during depression, although it was not statistically significant (2.80 beats per min, 95% CI −2.61 to 8.21; *P*=.22; 95% PI −7.51 to 13.11; Figure 2J). Notably, elevated nighttime mHR was linked to depression severity, whereas daytime mHR showed no consistent association [44,113].

SD of Normal-to-Normal Intervals

The SD of normal-to-normal intervals (SDNN) was lower in the depression group, but the meta-analysis result was not significant (−4.75 ms, 95% CI −12.09 to 2.58; *P*=.13; 95% PI −26.47 to 12.83; Figure 2K). While some studies reported significance [54,61,125], others did not [49,56]. Recent wearable electrocardiogram (ECG) research suggests that the SDNN tends to be reduced in individuals at risk of depression, reflecting autonomic dysregulation; however, its standalone discriminative power may be limited compared with other short-term HRV indices [125].

RMSSD Findings

The RMSSD tended to be lower in the depression group but was not consistently significant across studies [49,56,61,114]. Although some research on first-episode patients reported a significant association [54], others found no correlation between depression scores and the RMSSD [45,47,57,145]. Recent digital phenotyping studies using wearable ECG or multimodal sensing reported reduced RMSSD-related features in high-risk individuals, though their utility was often context-dependent or enhanced when combined with other features [125,147,149].

Proportion of Normal-to-Normal Intervals

Most studies showed no significant group differences in the proportion of normal-to-normal intervals (pNN) [43,54,112], although this has been highlighted in diagnostic models [53,55]. Wearable ECG-based screening studies similarly reported that pNN-related features contributed modestly to multivariate models but showed limited standalone discriminative ability for depression [125].

Mean RR Interval

The findings for the mean RR interval (RRI) were mixed. Two studies reported shorter RRIs in patients with severe depression [45,54], while 2 other studies found no relationship [56,59]. The mean RRI tended to be shorter in patients with depression, suggesting an autonomic imbalance; however, the small number of available studies (2/4, 50%) limits the generalizability of the findings [45,54], highlighting the need for further large-scale research.

SD of Heart Rate

Studies reported mixed findings regarding the SD of heart rate (HR). While some studies found that a reduced SD of HR was associated with greater depression severity [44], others showed an opposite association or a nonsignificant association [113,131].

Frequency Domain

HF Power

HF power is generally lower in individuals with depression [43,45,49,53,56,59,61] and may predict future depressive symptoms [49,51,147]. However, some studies reported nonsignificant associations in multivariate models [54,57,61]. A recent wearable ECG study suggested that while HF-related parasympathetic indices contribute to depression classification models, they may lose independent significance when integrated with other HRV features [125].

LF Power

LF power has shown inconsistent results. While several studies reported lower LF in depression [45,49,51,59,61], others found higher values or no significant associations [54,56,57]. Wearable ECG-based studies indicated that LF-related features contribute variably to depression classification models, reflecting substantial heterogeneity and limited standalone interpretability [125].

LF/HF Ratio

Some studies reported a higher LF/HF ratio in the depression group [45,47,53,54], whereas others found no significant associations after adjustment [56,57,61]. Digital phenotyping evidence suggests that while the LF/HF ratio may reflect altered autonomic balance, its discriminative performance remains inconsistent across populations and analytic models [125,147]. Overall, the LF/HF ratio may show an increasing trend with depression but requires further validation as a reliable biomarker.

Very Low Frequency and Ultra-Low Frequency

Very low frequency (VLF) power was generally lower in individuals with depression [51,57,59,61] and was identified as a key feature in diagnostic models [57], although some studies reported nonsignificant associations after adjustment [49]. Ultra-low frequency (ULF) power also tended to decrease with greater depression severity [51,61]; however, 1 study found no significant difference between the depression and control groups [59]. This limited evidence warrants further investigation.

Total Power

Total power findings were inconsistent across studies. While 1 study reported significant effects on depression severity [54], other studies found no such differences [49,56,61]. However, its utility as a standalone biomarker is not well-supported.

Smartphone Parameters and Depression

Phone Usage Frequency

Findings regarding usage frequency were mixed. One study reported higher use in individuals with depression, particularly

XSL•FO
RenderX

students [86], while other studies found no such association [80,145]. Multimodal studies suggested that raw usage volume may not differ between groups; instead, temporal patterns, screen time, and communication regularity appear more relevant [4,81,138,149]. Furthermore, physical activity and sleep features often outperformed usage frequency in predictive models [84,133,137].

Phone Usage Duration

Results regarding phone usage duration were conflicting. While some studies observed longer usage in individuals with depression [86,122,127], others reported a shorter duration [132] or no differences [145]. Its predictive value appeared more pronounced in younger populations [122] but generally remained inferior to physiological indicators, such as HRV or physical activity [148,149]. These demographic variations and the relative inconsistency across findings limit its potential as a reliable biomarker.

Phone Calls

The frequency of phone calls was lower in some individuals with depression [138], though this effect was modest and highly dependent on age and social context [4,149]. Younger users preferred text-based communication over traditional calls [82,84,133]. Consequently, call frequency was insufficient as a biomarker but may complement indicators, such as activity or sleep [4,131,145,149].

Light Exposure

Low light exposure was shown to be linked to relapse [135] or depression in older adults [134], yet other studies reported no effects [133,137,142]. While potentially relevant for specific subgroups, this parameter requires further validation.

Number of Bluetooth-Connected Devices

Used as a proxy for social contact, this parameter was shown to correlate negatively with depression [146] and was integrated into predictive models [82,122]. While useful for assessing social activity, broader application requires more robust evidence.

Typing Patterns

Typing behavior analyzed via machine learning showed high accuracy in predicting depression [78,86], though further validation is needed to establish its reliability as a biomarker.

Speech Parameters and Depression

Speech parameters were categorized into speech flow and voice acoustic parameters.

Speech Flow Parameters

Speech Rate

Slower speech rates are consistently associated with greater depression severity [88,96], with recent evidence confirming them as robust indicators of major depressive disorder (MDD) that correlate with objective executive dysfunction [87,144]. Models incorporating speech rate and duration outperformed acoustic-only models [95], though further validation across diverse populations is needed.

Speech Duration

Shorter speaking times were linked to higher depression severity [95,127,137,139,146]. Recent automated assessments and smartphone-derived data identified reduced active speaking time as a primary predictor of depression [137,144,150], establishing speech duration as a potential digital biomarker.

Pause Time

While some associations between depression and pause duration were nonsignificant [129], the majority of studies reinforced that increased pausing effectively captures psychomotor retardation and contributes to high-accuracy multimodal detection [87,88,96,144]. These findings underscore the potential of pause-related features as a reliable objective marker.

Voice Acoustic Parameters

Mel-Frequency Cepstral Coefficients

Mel-frequency cepstral coefficients were significantly associated with the severity of depression in most studies [91-93,129]. Recent multimodal analyses further confirmed their effectiveness when integrated with facial and cardiovascular patterns [144], although a study found text-based features more predictive than mel-frequency cepstral coefficients [95].

Fundamental Frequency

Findings on fundamental frequency were mixed. Some studies reported significant differences or enhanced effectiveness through multimodal integration [92,93,129,144], whereas others found no significant associations [89,91], necessitating further clarification of its role.

Jitter

Jitter was significantly higher in some individuals with depression [89], but other studies did not identify it as a significant variable [95,129]. Recent evidence suggests that it provides discriminative value when integrated into multimodal frameworks [144], warranting further investigation.

Shimmer

Shimmer showed significant associations with depression in some studies [89,129], while others found no such link [95]. Its predictive power was notably enhanced within multimodal frameworks, contributing to more robust detection than when used alone [144].

GPS Parameters and Depression

Total Distance

Most studies found a negative correlation between depression and total distance traveled [4,79,123,132,139,141,146]. Although not always statistically significant [86,124,145], reduced mobility was a promising digital biomarker, particularly when integrated into multimodal detection models [149].

Location Variance

Lower location variance was consistently associated with higher depression severity [4,82,86,124,132,145]. While 1 study noted inconsistencies depending on smartphone types [123], location variance remains a key feature in high-accuracy multimodal models for detecting depressive symptoms [149].

XSL•FO
RenderX

Time Spent at Home

A significant positive correlation exists between depression severity and time spent at home [4,86,123,124,127,132,149]. Notably, early changes in "homestay" were identified as a critical predictor of symptom improvement [136], demonstrating strong potential for longitudinal depression monitoring.

Location Entropy and Normalized Location Entropy

Reduced location and normalized entropy were generally associated with higher depression scores [4,86,123,124,132,149], reflecting less diverse movements. However, some studies found no significant associations [127,145].

Number of Locations Visited

Fewer visited locations correlated negatively with depression in most studies [4,123,127,132,137,149], though 1 study found no significant correlation [86].

Time Spent Moving

While associations between time spent moving and depression were inconsistent [86,123,124,145], recent longitudinal data identified early changes in moving time as a key predictor of symptom improvement [136]. These features are considered essential in multimodal frameworks for depression detection [149].

Average Moving Speed

Average moving speed was identified as a key feature in a depression prediction model [4,126], although its correlation often varied by device type [123]. While its standalone predictive power may be limited, it remains a crucial component in high-accuracy multimodal frameworks [149].

Circadian Rhythm Parameters and Depression

Interdaily Stability

Most studies found no significant group differences [97,99,100,117,118,150], whereas 2 studies reported lower interdaily stability and greater depression severity [103,113]. While interdaily stability may reflect irregular daily routines, current evidence is limited.

Intradaily Variability

Generally, intradaily variability was found to be unrelated to depression [97,99,100,117,118,150], despite 1 study linking higher intradaily variability to greater severity [103]. Its utility as a standalone biomarker remains restricted.

Midline Estimating Statistic of Rhythm

Midline estimating statistic of rhythm was lower in individuals with greater depression severity [102,148] and was identified as an important predictor in other models [99,101]. Consistently, recent evidence indicates that patients with MDD exhibit significantly lower midline estimating statistic of rhythm than controls [150], potentially reflecting reduced energy levels, though some studies found no associations [97,140].

Amplitude, Acrophase, and Relative Amplitude

Lower amplitude was observed in the depression group in some studies [101,102], while others found no significant differences [97,99,140,148,150]. Regarding acrophase, most studies found no association [97,101,102,113,148], but a recent study noted a significantly later acrophase in MDD, suggesting a delayed circadian phase [150]. Relative amplitude was generally lower in individuals with depression, suggesting flatter activity cycles, although these findings often did not reach statistical significance [97,99,100,117].

Pseudo F-Statistic

The pseudo F-statistic (F-pseudo), which measures the circadian rhythm strength, was lower in individuals with depressive symptoms [101,102], suggesting weaker or irregular rhythms. However, other studies reported no significant associations [99,113].

Most Active 10-Hour Period and Least Active 5-Hour Period

The most active 10-hour period and the least active 5-hour period showed no significant differences between the depression and control groups across multiple studies [97,99,117]. These parameters currently provide no evidence of reliability as biomarkers for depression.

Reporting Bias and Certainty of Evidence

The qualitative assessment of reporting bias suggested a low likelihood of missing relevant studies, supported by comprehensive multidatabase searches and manual reference screening. According to the GRADE approach, the certainty of evidence ranged from low to very low across the key digital biomarkers (Table 2). The certainty was low for SOL, TIB, and physical activity counts, whereas it was very low for TST, SE, WASO, and mHR. The overall certainty was mainly downgraded due to inconsistency across studies and imprecision associated with wide CIs or PIs.

XSL•FO
RenderX

Table . GRADE (Grading of Recommendations, Assessment, Development and Evaluation) summary of findings for key digital biomarkers in individuals with depression.

Certainty assessment	Studies, n	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Individuals with depression, n	Controls (no depression), n	Effect, relative (95% CI)	Effect, absolute (95% CI)	Certainty	Importance
Total sleep time	8	Nonrandomized studies	Not serious	Serious[a]	Not serious	Serious[b]	None	531	644	—[c]	MD[d] 2.65 min fewer (15.81 fewer to 10.5 more)	Very low[a,b]	Critical
Sleep efficiency	8	Nonrandomized studies	Not serious	Serious[e]	Not serious	Serious[b]	None	485	498	—	MD 2.89% lower (5.95 lower to 0.17 higher)	Very low[b,e]	Critical
Wake after sleep onset	5	Nonrandomized studies	Not serious	Serious[e]	Not serious	Serious[b]	None	274	286	—	SMD[f] 0.43 SD more (0.29 fewer to 1.16 more)	Very low[b,e]	Critical
Sleep onset latency	5	Nonrandomized studies	Not serious	Not serious	Not serious	Not serious	None	144	148	—	MD 4.75 min more (2.46 more to 7.04 more)	Low	Critical
Time in bed	3	Nonrandomized studies	Not serious	Not serious	Not serious	Not serious	None	105	131	—	MD 31.81 min more (18.22 more to 45.39 more)	Low	Critical
Physical activity counts	5	Nonrandomized studies	Not serious	Not serious	Not serious	Not serious	None	223	239	—	SMD 0.71 SD lower (1.33 lower to 0.09 lower)	Low	Critical
Mean heart rate	5	Nonrandomized studies	Not serious	Serious[e]	Not serious	Serious[b]	None	2012	2375	—	MD 2.8 bpm higher (2.61 lower to 8.21 higher)	Very low[b,e]	Critical

[a]Downgraded for serious inconsistency due to conflicting findings across studies and a wide prediction interval.

[b]Downgraded for serious imprecision due to a wide CI crossing the null effect.

[c]Not applicable.

[d]MD: mean difference.

XSL•FO
RenderX

[e]Downgraded for serious inconsistency due to moderate heterogeneity and variable results across studies.

[f]SMD: standardized mean difference.

Discussion

Principal Findings

This systematic review synthesized digital biomarkers for depression across diverse domains—sleep, physical activity, cardiac parameters, smartphone usage, speech, GPS data, and circadian rhythms—to identify more consistent indicators across multiple digital signals. Our meta-analysis identified prolonged SOL, increased TIB, and reduced activity counts as the most consistent behavioral features associated with depression. These findings support the hypothesis that digital phenotyping can capture objective manifestations of depression, particularly sleep initiation difficulties and reduced energy expenditure, which are often difficult to quantify through traditional self-reports.

A key contribution of this systematic review is the application of quantitative meta-analyses in a field where such synthesis was previously considered methodologically challenging [10,11]. Unlike prior meta-analyses that focused on single domains or summarized findings narratively because of methodological heterogeneity [13,14], this review systematically evaluates digital biomarkers across diverse domains. Consistent with earlier reports, the results indicate that depression is not characterized by a single physiological signature but rather by a constellation of behavioral and biological changes [6,149].

Furthermore, unlike recent scoping reviews on general mental illness [151], our analysis is specifically tailored to MDD. Importantly, the interpretation of these findings must consider the substantial heterogeneity observed across studies. The CIs reflect the average effect across the included studies, whereas the PIs indicate the range of effects that may occur in future settings. For several outcomes, wide PIs crossing the null value suggest that effect sizes may vary considerably depending on the study population, device type, measurement protocol, and analytical approach. These findings indicate that group-level average effects may not generalize consistently across contexts, supporting the use of personalized, longitudinal multibiomarker models for effective monitoring and intervention.

Key Digital Biomarkers

Several biomarkers emerged as important indicators of depression across multiple domains. In the sleep domain, individuals with depression showed significantly longer SOL and increased TIB compared to controls without depression. These findings suggest that the sleep-wake cycle in MDD is characterized more by structural fragmentation than by simple reductions in sleep duration. Specifically, increased TIB likely reflects hallmark symptoms such as psychomotor retardation and lethargy. While polysomnography remains the gold standard [152], these results demonstrate that wearable devices can capture such patterns in ecologically valid, naturalistic settings.

The absence of significant differences in TST, SE, and WASO, which are parameters frequently associated with depression in previous reviews [4], highlights the clinical heterogeneity of MDD [153,154]. Depression may manifest as either insomnia or hypersomnia, depending on the subtype, potentially neutralizing average effects in pooled analyses [155]. The wide PIs further indicate substantial heterogeneity, likely driven by differences in study populations, medication use, and device-specific scoring. These findings underscore that absolute sleep quantity alone is an insufficient marker, and the relationship between sleep and depression is shaped by complex biopsychosocial factors [153,154]. Consequently, temporal and qualitative features, such as SOL and TIB, may serve as more clinically informative indicators than total sleep volume [9,156].

With respect to sleep architecture, evidence regarding REM-related parameters remains inconclusive. While some studies suggest that shortened REM latency may precede depression [157], results for REM duration and frequency are inconsistent [26,31,41,115,116]. Age-related changes [33] and reliance on single-night laboratory measurements further limit generalizability. These findings highlight the need for longitudinal, real-world assessments to clarify the relationship between sleep architecture and depression.

Physical activity counts emerged as a sensitive state-dependent marker, showing significant reductions in depression groups. Compared with simple step counts, activity counts better capture movement intensity and frequency [65,142,158], reflecting the energy deficits associated with MDD. Although daily step counts were not significantly different in pooled analyses, longitudinal evidence suggests that they may be sensitive to individual recovery trajectories [65,142]. Several studies also reported more pronounced reductions during daytime hours and on weekends, emphasizing the importance of temporal patterns rather than simple daily averages [158,159]. In contrast, MVPA, sedentary time, and LPA showed inconsistent results. This lack of robust significance likely reflects high heterogeneity and the limited specificity of these markers when used in isolation.

GPS-derived parameters, including total distance traveled, location entropy, time spent at home, and number of locations visited, capture behavioral and environmental changes associated with depression. Mobility patterns, such as reduced diversity in visited locations (entropy) and increased homestay durations, have emerged as promising indicators of depressive symptoms [160]. Recent longitudinal evidence suggests that early changes in homestay and mobility patterns may predict symptom improvement [136], highlighting the value of GPS-derived features in multimodal monitoring frameworks [149].

Cardiac parameters showed less consistent results. Individuals with depression tended to exhibit higher nocturnal mHR and lower HRV, although these differences were not statistically significant in pooled analyses. These trends may reflect physiological hyperarousal and autonomic dysregulation [161,162], but their effects appear to vary across populations and study conditions. Accordingly, these markers may be more informative when interpreted within multimodal models that account for age, medication use, and comorbid conditions. Despite this overall inconsistency, short-term HRV indices and

XSL•FO
RenderX

frequency-domain features (eg, HF and VLF) remain meaningful components in diagnostic and multimodal frameworks, particularly as sensitive indicators of autonomic imbalance in first-episode depression or high-risk individuals [54,125,149].

Speech parameters, including slower speech rates and longer pauses, consistently distinguished individuals with MDD from controls in several studies [87,144]. These acoustic features likely reflect psychomotor retardation.

Implications for Personalized Digital Phenotyping

Digital biomarkers offer several advantages over traditional self-reported measures by providing objective, continuous data that reduce recall bias and capture early physiological and behavioral changes [163,164]. Passive monitoring enables large-scale, low-burden data collection, supporting early detection and personalized, data-driven interventions [5,8,165].

However, the substantial variability observed across individuals highlights the need to interpret these signals within a personalized framework, as traditional group-based models may be inadequate in many contexts [166]. The predictive power of biomarkers varies with age. TIB, daily step counts, and MVPA are more predictive in older adults [11,109,167], whereas smartphone usage patterns show stronger associations in younger individuals [11,122]. High-resolution temporal data, such as nighttime HRV or weekend reductions in MVPA, may provide additional insights into symptom dynamics, underscoring the importance of temporal patterns in digital measurements [69]. Monitoring sleep-related biomarkers in relation to changes in individual symptoms may further improve the prediction and tracking of depressive episodes [6]. Together, these findings suggest that digital monitoring strategies should focus on detecting deviations from an individual's baseline rather than relying on universal thresholds.

Methodological Limitations of the Included Studies and Evidence Certainty

Despite identifying several robust biomarkers, parameters, such as LPA, sedentary time, and specific cardiac measures, showed inconsistent results. This inconsistency, together with the wide PIs observed across studies, likely reflects not only diverse device technologies and measurement protocols, but also the symptomatic variability across depression subtypes [155,168]. While wearable devices offer objective physiological data, they do not directly capture emotional states [169]. Furthermore, the predominance of cross-sectional designs limits our ability to determine the temporal or causal relationships between these digital signals and depressive symptoms [170].

Risk-of-bias assessments also indicated several methodological concerns. Although participation rates were somewhat low in certain studies [24,34,59,62,128], they were generally acceptable across the majority of the included research [4,25,27,31,32,35,38,42,45,49,52-54,56,60,67,68,76,88,91,101,111,112,114,115,117,131,145]. However, comparisons between participants and nonparticipants were often not reported, raising the possibility of selection bias. For example, individuals with more severe depressive symptoms, lower motivation, or limited familiarity with wearable technologies may have been less likely to participate or remain in the studies. In addition, adjustment for confounding variables was limited in several studies. For diagnostic accuracy studies, most domains showed low risk of bias, but concerns regarding applicability were common because study populations did not always reflect real-world clinical settings.

Consistent with these methodological limitations, the GRADE assessment indicated low to very low certainty of evidence for the key digital biomarkers. This was primarily due to inconsistency across studies and imprecision associated with wide CIs and PIs. These findings suggest that the pooled estimates should be interpreted cautiously and that further well-designed, standardized studies are needed to strengthen the evidence base.

Limitations of This Review

This systematic review has several limitations. First, substantial heterogeneity across studies limited the precision and generalizability of the pooled estimates, reflecting differences in devices, study populations, monitoring periods, and analytic methods. Second, the overall certainty of evidence was low to very low according to the GRADE framework, which reduces confidence in the pooled estimates. Third, a formal statistical assessment of reporting bias was not feasible because fewer than 10 studies were available for each meta-analysis. Finally, many studies included nonclinical or convenience samples, which may limit the generalizability of the findings to real-world clinical populations.

Future Directions for Digital Biomarker Research

Future research should move beyond the identification of individual markers toward the development of integrated, clinically actionable digital phenotyping systems. The consistent associations observed for indicators, such as SOL, TIB, and activity counts, suggest that certain behavioral signals may serve as foundational components of continuous mental health monitoring. The next phase of research should therefore focus on embedding these markers within longitudinal, multimodal frameworks that support personalized clinical decision-making and improve the precision of depression monitoring [149,150,155], with the potential to inform more proactive intervention strategies in real-world settings.

Achieving this transition will require greater methodological standardization across devices, measurement protocols, and analytic pipelines. Reducing reliance on proprietary, nontransparent algorithms and promoting device-agnostic, reproducible approaches will be essential for ensuring clinical validity and interoperability across health care systems. Integration of digital biomarker data into electronic health records may further enable real-time, context-aware decision support while minimizing additional cognitive burden on clinicians.

More broadly, digital biomarkers may support earlier detection and more adaptive treatment strategies by providing high-frequency, objective data on symptom trajectories [136,164]. At the population level, scalable signals such as physical activity and mobility patterns may also facilitate screening and risk stratification in settings with limited access to mental health professionals. Future work should therefore prioritize longitudinal, multimodal designs; representative

XSL•FO
RenderX

clinical populations; standardized measurement protocols; and transparent reporting of analytic methods.

Conclusion

This systematic review provides a comprehensive synthesis of multimodal digital biomarkers for depression. Unlike previous reviews that focused on single signals or technical feasibility, this systematic review advances the field by establishing a standardized framework for objective clinical decision-making through a rigorous meta-analysis. The findings indicate that while certain markers, particularly SOL and physical activity counts, show consistent average differences, their effects vary substantially across settings, as reflected by wide PIs. These results suggest that depression cannot be reliably characterized by a single digital biomarker. Instead, a multimodal, personalized approach that integrates physiological, behavioral, and contextual signals is likely to be more effective for real-world applications. More broadly, this systematic review demonstrates that quantitative synthesis in digital phenotyping is feasible despite substantial heterogeneity and that meaningful signals can be identified when methodological rigor and transparent reporting are applied. Establishing standardized, clinically interpretable digital biomarker frameworks will be essential for advancing objective, continuous, and personalized assessments of depression in routine care.

Funding

This study was supported by a grant (24202MFDS201) from the Ministry of Food and Drug Safety in 2024. The funder was not involved in the study design, data collection, analysis, interpretation, or writing of the manuscript.

Disclaimer

Generative artificial intelligence tools were not used in the preparation, writing, or editing of this manuscript.

Data Availability

No original individual participant data were collected for this systematic review. All the data used in this systematic review and meta-analysis were extracted from previously published studies. Therefore, individual-level datasets cannot be deposited in a public repository. The list of the included studies and extracted variables is available from the corresponding author upon reasonable request.

Authors' Contributions

SHL contributed to the study conception and design. HL and SHL assisted with the statistical analysis and provided administrative support. HL, SHL, and SGK wrote the first draft of the manuscript. All authors contributed to the writing and preparation of the final manuscript. All authors had full access to all the data in the study and had the final responsibility for the decision to submit the manuscript for publication. HL and SHL accessed and verified the data reported in this manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search terms and strategy.
[DOCX File, 63 KB - jmir_v28i1e76432_app1.docx]

Multimedia Appendix 2
Detailed study characteristics and quality assessment of the included studies.
[DOCX File, 165 KB - jmir_v28i1e76432_app2.docx]

Multimedia Appendix 3
Participants and analytical methods in the studies.
[DOCX File, 23 KB - jmir_v28i1e76432_app3.docx]

Multimedia Appendix 4
Depression measurement tools included in the studies.
[DOCX File, 34 KB - jmir_v28i1e76432_app4.docx]

Multimedia Appendix 5
Digital biomarkers included in the studies.
[DOCX File, 79 KB - jmir_v28i1e76432_app5.docx]

XSL•FO
RenderX

Multimedia Appendix 6
Digital biomarkers not included in the meta-analysis.
[DOCX File, 63 KB - jmir_v28i1e76432_app6.docx]

Checklist 1
PRISMA checklist.
[DOCX File, 278 KB - jmir_v28i1e76432_app7.docx]

References

1. Li S, Xu Y, Zheng L, et al. Sex difference in global burden of major depressive disorder: findings from the Global Burden of Disease Study 2019. Front Psychiatry 2022;13:70890. [doi: 10.3389/fpsyt.2022.789305]
2. Pu J, Yu Y, Liu Y, et al. ProMENDA: an updated resource for proteomic and metabolomic characterization in depression. Transl Psychiatry 2024 May 30;14(1):229. [doi: 10.1038/s41398-024-02948-2] [Medline: 38816410]
3. Macias Alonso AK, Hirt J, Woelfle T, Janiaud P, Hemkens LG. Definitions of digital biomarkers: a systematic mapping of the biomedical literature. BMJ Health Care Inform 2024 Apr 8;31(1):345. [doi: 10.1136/bmjhci-2023-100914] [Medline: 38589213]
4. Aledavood T, Luong N, Baryshnikov I, et al. Multimodal digital phenotyping study in patients with major depressive episodes and healthy controls (mobile monitoring of mood): observational longitudinal study. JMIR Ment Health 2025 Feb 21;12:e63622. [doi: 10.2196/63622] [Medline: 39984168]
5. Ikäheimonen A, Luong N, Baryshnikov I, et al. Predicting and monitoring symptoms in patients diagnosed with depression using smartphone data: observational study. J Med Internet Res 2024 Dec 3;26:e56874. [doi: 10.2196/56874] [Medline: 39626241]
6. Jacobson NC, Weingarden H, Wilhelm S. Digital biomarkers of mood disorders and symptom change. NPJ Digit Med 2019;2(1):3. [doi: 10.1038/s41746-019-0078-0] [Medline: 31304353]
7. Jung HW, Kim DY, Lee I, et al. Key features of digital phenotyping for monitoring mental disorders: systematic review. J Med Internet Res 2025 Nov 5;27:e77331. [doi: 10.2196/77331] [Medline: 41191793]
8. Zhuparris A, Maleki G, van Londen L, et al. A smartphone- and wearable-based biomarker for the estimation of unipolar depression severity. Sci Rep 2023 Nov 1;13(1):18844. [doi: 10.1038/s41598-023-46075-2] [Medline: 37914808]
9. Zierer C, Behrendt C, Lepach-Engelhardt AC. Digital biomarkers in depression: a systematic review and call for standardization and harmonization of feature engineering. J Affect Disord 2024 Jul 1;356:438-449. [doi: 10.1016/j.jad.2024.03.163] [Medline: 38583596]
10. De Angel V, Lewis S, White K, et al. Digital health tools for the passive monitoring of depression: a systematic review of methods. NPJ Digit Med 2022 Jan 11;5(1):3. [doi: 10.1038/s41746-021-00548-8] [Medline: 35017634]
11. Zarate D, Stavropoulos V, Ball M, de Sena Collier G, Jacobson NC. Exploring the digital footprint of depression: a PRISMA systematic literature review of the empirical evidence. BMC Psychiatry 2022 Jun 22;22(1):421. [doi: 10.1186/s12888-022-04013-y] [Medline: 35733121]
12. Leaning IE, Ikani N, Savage HS, et al. From smartphone data to clinically relevant predictions: a systematic review of digital phenotyping methods in depression. Neurosci Biobehav Rev 2024 Mar;158:105541. [doi: 10.1016/j.neubiorev.2024.105541] [Medline: 38215802]
13. Ho FYY, Poon CY, Wong VWH, et al. Actigraphic monitoring of sleep and circadian rest-activity rhythm in individuals with major depressive disorder or depressive symptoms: a meta-analysis. J Affect Disord 2024 Sep 15;361:224-244. [doi: 10.1016/j.jad.2024.05.155] [Medline: 38851435]
14. Bizzozero-Peroni B, Díaz-Goñi V, Jiménez-López E, et al. Daily step count and depression in adults: a systematic review and meta-analysis. JAMA Netw Open 2024 Dec 2;7(12):e2451208. [doi: 10.1001/jamanetworkopen.2024.51208] [Medline: 39680407]
15. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
16. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14(1):25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
17. Higgins JPT, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.5 (updated August 2024): Cochrane; 2024. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook [accessed 2026-03-11]
18. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019 Jun;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]

XSL•FO
RenderX

19. Schünemann H, Brożek J, Guyatt G, Oxman A. GRADE Handbook - Handbook for Grading the Quality of Evidence and the Strength of Recommendations Using the GRADE Approach: GRADE Working Group; 2013. URL: https://gdt.gradepro.org/app/handbook/handbook.html [accessed 2026-03-11]
20. Poon CY, Cheng YC, Wong VWH, et al. Directional associations among real-time activity, sleep, mood, and daytime symptoms in major depressive disorder using actigraphy and ecological momentary assessment. Behav Res Ther 2024 Feb;173:104464. [doi: 10.1016/j.brat.2023.104464] [Medline: 38159415]
21. Peng D, Liu W, Luo Y, Mao Z, Zheng WL, Lu BL. Deep depression detection with resting-state and cognitive-task EEG. Annu Int Conf IEEE Eng Med Biol Soc 2023 Jul;2023:1-4. [doi: 10.1109/EMBC40787.2023.10340667] [Medline: 38083722]
22. Comas M, Solis Flores A, Lovato N, et al. The relationship between anxiety, subjective and objective sleep, chronotype and circadian rhythms with depressive symptoms in insomnia disorder. Brain Sci 2023 Apr 4;13(4):613. [doi: 10.3390/brainsci13040613] [Medline: 37190578]
23. Bowman MA, Kline CE, Buysse DJ, et al. Longitudinal association between depressive symptoms and multidimensional sleep health: the SWAN Sleep Study. Ann Behav Med 2021 Jun 28;55(7):641-652. [doi: 10.1093/abm/kaaa107] [Medline: 33410460]
24. Smith P, Kandakatla A, Frankel CW, et al. Sleep quality, depressive symptoms, and transplant outcomes: follow-up analyses from the ADAPT prospective pilot study. Gen Hosp Psychiatry 2021;72:53-58. [doi: 10.1016/j.genhosppsych.2021.06.011] [Medline: 34298477]
25. Mohammadi Y, Moradi MH. Prediction of depression severity scores based on functional connectivity and complexity of the EEG signal. Clin EEG Neurosci 2021 Jan;52(1):52-60. [doi: 10.1177/1550059420965431] [Medline: 33040603]
26. Zhang Y, Folarin AA, Sun S, et al. Predicting depressive symptom severity through individuals' nearby Bluetooth device count data collected by mobile phones: preliminary longitudinal study. JMIR Mhealth Uhealth 2021 Jul 30;9(7):e29840. [doi: 10.2196/29840] [Medline: 34328441]
27. Chen F, Zhao L, Li B, Yang L. Depression evaluation based on prefrontal EEG signals in resting state using fuzzy measure entropy. Physiol Meas 2020 Oct 6;41(9):095007. [doi: 10.1088/1361-6579/abb144] [Medline: 33021227]
28. Hasanzadeh F, Mohebbi M, Rostami R. Graph theory analysis of directed functional brain networks in major depressive disorder based on EEG signal. J Neural Eng 2020 Mar 27;17(2):026010. [doi: 10.1088/1741-2552/ab7613] [Medline: 32053813]
29. Rajpurkar P, Yang J, Dass N, et al. Evaluation of a machine learning model based on pretreatment symptoms and electroencephalographic features to predict outcomes of antidepressant treatment in adults with depression: a prespecified secondary analysis of a randomized clinical trial. JAMA Netw Open 2020 Jun 1;3(6):e206653. [doi: 10.1001/jamanetworkopen.2020.6653] [Medline: 32568399]
30. Ding X, Yue X, Zheng R, Bi C, Li D, Yao G. Classifying major depression patients and healthy controls using EEG, eye tracking and galvanic skin response data. J Affect Disord 2019 May 15;251:156-161. [doi: 10.1016/j.jad.2019.03.058] [Medline: 30925266]
31. Gould CE, Karna R, Jordan J, et al. Subjective but not objective sleep is associated with subsyndromal anxiety and depression in community-dwelling older adults. Am J Geriatr Psychiatry 2018 Jul;26(7):806-811. [doi: 10.1016/j.jagp.2018.03.010] [Medline: 29709510]
32. Haghighi M, Ludyga S, Rahimi B, et al. In patients suffering from major depressive disorders, quantitative EEG showed favorable changes in left and right prefrontal cortex. Psychiatry Res 2017 May;251:137-141. [doi: 10.1016/j.psychres.2017.02.012] [Medline: 28199912]
33. Lee HJ, Lee JS, Kim T, Yoon IY. Relationship between sleep disturbances and cognitive impairments in older adults with depression. Sleep Medicine and Psychophysiology 2014 Jun 30;21(1):5-13. [doi: 10.14401/KASMED.2014.21.1.5]
34. Maglione JE, Ancoli-Israel S, Peters KW, et al. Subjective and objective sleep disturbance and longitudinal risk of depression in a cohort of older women. Sleep 2014 Jul 1;37(7):1179-1187. [doi: 10.5665/sleep.3834] [Medline: 25061246]
35. Pillai V, Steenburg LA, Ciesla JA, Roth T, Drake CL. A seven day actigraphy-based study of rumination and sleep disturbance among young adults with depressive symptoms. J Psychosom Res 2014 Jul;77(1):70-75. [doi: 10.1016/j.jpsychores.2014.05.004] [Medline: 24913345]
36. Maglione JE, Ancoli-Israel S, Peters KW, et al. Depressive symptoms and subjective and objective sleep in community-dwelling older women. J Am Geriatr Soc 2012 Apr;60(4):635-643. [doi: 10.1111/j.1532-5415.2012.03908.x] [Medline: 22428562]
37. Bei B, Milgrom J, Ericksen J, Trinder J. Subjective perception of sleep, but not its objective quality, is associated with immediate postpartum mood disturbances in healthy women. Sleep 2010 Apr;33(4):531-538. [doi: 10.1093/sleep/33.4.531] [Medline: 20394323]
38. Goyal D, Gay C, Lee K. Fragmented maternal sleep is more strongly correlated with depressive symptoms than infant temperament at three months postpartum. Arch Womens Ment Health 2009 Aug;12(4):229-237. [doi: 10.1007/s00737-009-0070-9] [Medline: 19396527]
39. Coffield TG, Tryon WW. Construct validation of actigraphic sleep measures in hospitalized depressed patients. Behav Sleep Med 2004;2(1):24-40. [doi: 10.1207/s15402010bsm0201_3] [Medline: 15600222]

XSL•FO
RenderX

40. Nofzinger EA, Nichols TE, Meltzer CC, et al. Changes in forebrain function from waking to REM sleep in depression: preliminary analyses [of 18F]FDG PET studies. Psychiatry Research: Neuroimaging 1999 Aug;91(2):59-78. [doi: 10.1016/S0925-4927(99)00025-6]
41. Luthringer R, Toussaint M, Schaltenbrand N, et al. A double-blind, placebo-controlled evaluation of the effects of orally administered venlafaxine on sleep in inpatients with major depression. Psychopharmacol Bull 1996;32(4):637-646. [Medline: 8993085]
42. Roemer RA, Shagass C, Dubin W, Jaffe R, Siegal L. Quantitative EEG in elderly depressives. Brain Topogr 1992;4(4):285-290. [doi: 10.1007/BF01135566] [Medline: 1510872]
43. Theofilis P, Oikonomou E, Lazaros G, et al. The association of depression with QT duration: a comparison between individuals younger or older than 65 years. Psychosom Med 2023 Jan 1;85(1):18-25. [doi: 10.1097/PSY.0000000000001152] [Medline: 36516316]
44. Siddi S, Bailon R, Giné-Vázquez I, et al. The usability of daytime and night-time heart rate dynamics as digital biomarkers of depression severity. Psychol Med 2023 Jun;53(8):3249-3260. [doi: 10.1017/S0033291723001034] [Medline: 37184076]
45. Subramaniam V, R N. Relationship between heart rate variability and major depressive disorder in young adults. Swiss Arch Neurol Psychiatr Psychother 2023;174:126-129. [doi: 10.4414/sanp.2023.03284]
46. Lee SY, Waring ME, Park CL, Blake EC. Do depressive symptoms predict blood pressure control in US veterans? J Gen Intern Med 2022 Jan;37(1):57-63. [doi: 10.1007/s11606-021-06709-5] [Medline: 33772439]
47. Chung AH, Gevirtz RN, Gharbo RS, Thiam MA, Ginsberg JPJ. Pilot study on reducing symptoms of anxiety with a heart rate variability biofeedback wearable and remote stress management coach. Appl Psychophysiol Biofeedback 2021 Dec;46(4):347-358. [doi: 10.1007/s10484-021-09519-x] [Medline: 34308526]
48. da Estrela C, McGrath J, Booij L, Gouin JP. Heart rate variability, sleep quality, and depression in the context of chronic stress. Ann Behav Med 2021 Mar 16;55(2):155-164. [doi: 10.1093/abm/kaaa039] [Medline: 32525208]
49. Lee MS, Seo YE, Mok YE, Lee SH. Heart rate variability after treatment for depression in North Korean defectors. Appl Psychophysiol Biofeedback 2021 Mar;46(1):11-18. [doi: 10.1007/s10484-020-09491-y] [Medline: 33074464]
50. Coutts LV, Plans D, Brown AW, Collomosse J. Deep learning with wearable based heart rate variability for prediction of mental and general health. J Biomed Inform 2020 Dec;112(112):103610. [doi: 10.1016/j.jbi.2020.103610] [Medline: 33137470]
51. Huang M, Shah A, Su S, et al. Association of depressive symptoms and heart rate variability in Vietnam war-era twins: a longitudinal twin difference study. JAMA Psychiatry 2018 Jul 1;75(7):705-712. [doi: 10.1001/jamapsychiatry.2018.0747] [Medline: 29799951]
52. Hamilton JL, Alloy LB. Physiological markers of interpersonal stress generation in depression. Clin Psychol Sci 2017;5(6):911-929. [doi: 10.1177/2167702617720211] [Medline: 29123970]
53. Sun G, Shinba T, Kirimoto T, Matsui T. An objective screening method for major depressive disorder using logistic regression analysis of heart rate variability data obtained in a mental task paradigm. Front Psychiatry 2016;7:180. [doi: 10.3389/fpsyt.2016.00180] [Medline: 27867364]
54. Liang CS, Lee JF, Chen CC, Chang YC. Reactive heart rate variability in male patients with first-episode major depressive disorder. Prog Neuropsychopharmacol Biol Psychiatry 2015 Jan 2;56:52-57. [doi: 10.1016/j.pnpbp.2014.08.004] [Medline: 25149628]
55. Roh T, Hong S, Yoo HJ. Wearable depression monitoring system with heart-rate variability. Annu Int Conf IEEE Eng Med Biol Soc 2014;2014:562-565. [doi: 10.1109/EMBC.2014.6943653] [Medline: 25570021]
56. Lee CK, Yoo SK. An analysis of the relationship between self-reported anxiety, depressiveness and parameters of heart rate variability based on photoplethysmography [Article in Korean]. Sci Emot Sens 2012;3:345-354 [FREE Full text]
57. Zhang ZX, Tian XW, Lim JS. Neuro-fuzzy network-based depression diagnosis algorithm using optimal features of HRV. The Journal of the Korea Contents Association 2012 Feb 28;12(2):1-9. [doi: 10.5392/JKCA.2012.12.02.001]
58. Licht CMM, de Geus EJC, Zitman FG, Hoogendijk WJG, van Dyck R, Penninx B. Association between major depressive disorder and heart rate variability in the Netherlands Study of Depression and Anxiety (NESDA). Arch Gen Psychiatry 2008 Dec;65(12):1358-1367. [doi: 10.1001/archpsyc.65.12.1358] [Medline: 19047522]
59. Kojima M, Hayano J, Fukuta H, et al. Loss of fractal heart rate dynamics in depressive hemodialysis patients. Psychosom Med 2008 Feb;70(2):177-185. [doi: 10.1097/PSY.0b013e31816477a1] [Medline: 18256338]
60. Stroup-Benham CA, Markides KS, Black SA, Goodwin JS. Relationship between low blood pressure and depressive symptomatology in older people. J Am Geriatr Soc 2000 Mar;48(3):250-255. [doi: 10.1111/j.1532-5415.2000.tb02642.x] [Medline: 10733049]
61. Stein PK, Carney RM, Freedland KE, et al. Severe depression is associated with markedly reduced heart rate variability in patients with stable coronary heart disease. J Psychosom Res 2000;48(4-5):493-500. [doi: 10.1016/s0022-3999(99)00085-9] [Medline: 10880671]
62. Wang X, Pathiravasan CH, Zhang Y, et al. Association of depressive symptom trajectory with physical activity collected by mHealth devices in the Electronic Framingham Heart Study: cohort study. JMIR Ment Health 2023 Jul 14;10:e44529. [doi: 10.2196/44529] [Medline: 37450333]

XSL•FO
RenderX

63. Espino-Salinas CH, Galván-Tejada CE, Luna-García H, et al. Two-dimensional convolutional neural network for depression episodes detection in real time using motor activity time series of Depresjon dataset. Bioengineering (Basel) 2022 Sep 9;9(9):458. [doi: 10.3390/bioengineering9090458] [Medline: 36135004]
64. Hsiao C, Hsueh MC, Liao Y. Associations between objectively measured sedentary behavior patterns and depressive symptoms in older adults: a cross sectional study. Ment Health Phys Act 2022 Oct;23:100471. [doi: 10.1016/j.mhpa.2022.100471]
65. Ramsey CM, Lynch KG, Gehrman PR, et al. Daily steps and depressive symptoms: a longitudinal evaluation of patients with major depressive disorder in the precision medicine in mental health care study. J Affect Disord 2022 Mar 1;300:334-340. [doi: 10.1016/j.jad.2021.12.116] [Medline: 34979178]
66. Hussenoeder FS, Conrad I, Pabst A, et al. Physical activity and mental health: the connection between step count and depression, anxiety and quality of sleep. Psychol Health Med 2023;28(9):2419-2429. [doi: 10.1080/13548506.2022.2159453] [Medline: 36529963]
67. Michalak J, Niemeyer H, Tschacher W, Baumann N, Chi Zhang X, Adolph D. Subjective and objective measures of activity in depressed and non-depressed individuals in everyday life. J Exp Psychopathol 2022 Apr;13(2):20438087221092582. [doi: 10.1177/20438087221092582]
68. Li YM, Konstabel K, Mõttus R, Lemola S. Temporal associations between objectively measured physical activity and depressive symptoms: an experience sampling study. Front Psychiatry 2022;13:920580. [doi: 10.3389/fpsyt.2022.920580] [Medline: 35923450]
69. Figueroa CA, Vittinghoff E, Aguilera A, Fukuoka Y. Differences in objectively measured daily physical activity patterns related to depressive symptoms in community dwelling women - mPED trial. Prev Med Rep 2021 Jun;22:101325. [doi: 10.1016/j.pmedr.2021.101325] [Medline: 33659156]
70. Hsueh MC, Stubbs B, Lai YJ, Sun CK, Chen LJ, Ku PW. A dose response relationship between accelerometer assessed daily steps and depressive symptoms in older adults: a two-year cohort study. Age Ageing 2021 Feb 26;50(2):519-526. [doi: 10.1093/ageing/afaa162] [Medline: 32980870]
71. Jakobsen P, Garcia-Ceja E, Riegler M, et al. Applying machine learning in motor activity time series of depressed bipolar and unipolar patients compared to healthy controls. PLoS ONE 2020;15(8):e0231995. [doi: 10.1371/journal.pone.0231995] [Medline: 32833958]
72. Zhao N, Zhang Z, Wang Y, et al. See your mental state from your walk: recognizing anxiety and depression through Kinect-recorded gait data. PLoS ONE 2019;14(5):e0216591. [doi: 10.1371/journal.pone.0216591] [Medline: 31116785]
73. Averill IR, Crowe M, Frampton CM, et al. Clinical response to treatment in inpatients with depression correlates with changes in activity levels and psychomotor speed. Aust N Z J Psychiatry 2018 Jul;52(7):652-659. [doi: 10.1177/0004867417753549] [Medline: 29417833]
74. Ku PW, Steptoe A, Liao Y, Sun WJ, Chen LJ. Prospective relationship between objectively measured light physical activity and depressive symptoms in later life. Int J Geriatr Psychiatry 2018 Jan;33(1):58-65. [doi: 10.1002/gps.4672] [Medline: 28181713]
75. Holmquist S, Mattsson S, Schele I, Nordström P, Nordström A. Low physical activity as a key differentiating factor in the potential high-risk profile for depressive symptoms in older adults. Depress Anxiety 2017 Sep;34(9):817-825. [doi: 10.1002/da.22638] [Medline: 28489319]
76. Todder D, Caliskan S, Baune BT. Longitudinal changes of day-time and night-time gross motor activity in clinical responders and non-responders of major depression. World J Biol Psychiatry 2009;10(4):276-284. [doi: 10.3109/15622970701403081] [Medline: 19921969]
77. Lemke MR, Wendorff T, Mieth B, Buhl K, Linnemann M. Spatiotemporal gait patterns during over ground locomotion in major depression compared with healthy controls. J Psychiatr Res 2000;34(4-5):277-283. [doi: 10.1016/s0022-3956(00)00017-0] [Medline: 11104839]
78. Fadul R, Alfalahi H, Shehhi AA, Hadjileontiadis L. Depressive disorder remote detection through touchscreen typing behaviour. Annu Int Conf IEEE Eng Med Biol Soc 2023 Jul;2023:1-4. [doi: 10.1109/EMBC40787.2023.10340393] [Medline: 38082634]
79. Auerbach RP, Srinivasan A, Kirshenbaum JS, Mann JJ, Shankman SA. Geolocation features differentiate healthy from remitted depressed adults. J Psychopathol Clin Sci 2022 May;131(4):341-349. [doi: 10.1037/abn0000742] [Medline: 35230855]
80. Otte Andersen T, Skovlund Dissing A, Rosenbek Severinsen E, et al. Predicting stress and depressive symptoms using high-resolution smartphone data and sleep behavior in Danish adults. Sleep 2022 Jun 13;45(6):zsac067. [doi: 10.1093/sleep/zsac067] [Medline: 35298650]
81. Opoku Asare K, Terhorst Y, Vega J, Peltonen E, Lagerspetz E, Ferreira D. Predicting depression from smartphone behavioral markers using machine learning methods, hyperparameter optimization, and feature importance analysis: exploratory study. JMIR Mhealth Uhealth 2021 Jul 12;9(7):e26540. [doi: 10.2196/26540] [Medline: 34255713]
82. Chikersal P, Doryab A, Tumminia M, et al. Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: a machine learning approach with robust feature selection. ACM Trans Comput Hum Interact 2021;28(1):1-41. [doi: 10.1145/3422821]

XSL•FO
RenderX

83. Zhang Y, Folarin AA, Sun S, et al. Relationship between major depression symptom severity and sleep collected using a wristband wearable device: multicenter longitudinal observational study. JMIR Mhealth Uhealth 2021 Apr 12;9(4):e24604. [doi: 10.2196/24604] [Medline: 33843591]
84. Pedrelli P, Fedor S, Ghandeharioun A, et al. Monitoring changes in depression severity using wearable and mobile sensors. Front Psychiatry 2020;11:584711. [doi: 10.3389/fpsyt.2020.584711] [Medline: 33391050]
85. Mastoras RE, Iakovakis D, Hadjidimitriou S, et al. Touchscreen typing pattern analysis for remote detection of the depressive tendency. Sci Rep 2019 Sep 16;9(1):13414. [doi: 10.1038/s41598-019-50002-9] [Medline: 31527640]
86. Saeb S, Zhang M, Karr CJ, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res 2015 Jul 15;17(7):e175. [doi: 10.2196/jmir.4273] [Medline: 26180009]
87. Wiseman M, Yep R, Wood Alexander M, et al. Objective speech measures capture depressive symptoms and associated cognitive difficulties. Transl Psychiatry 2025 Nov 17;15(1):525. [doi: 10.1038/s41398-025-03728-2] [Medline: 41257835]
88. Wadle LM, Ebner-Priemer UW, Foo JC, et al. Speech features as predictors of momentary depression severity in patients with depressive disorder undergoing sleep deprivation therapy: ambulatory assessment pilot study. JMIR Ment Health 2024 Jan 18;11:e49222. [doi: 10.2196/49222] [Medline: 38236637]
89. Silva WJ, Lopes L, Galdino MKC, Almeida AA. Voice acoustic parameters as predictors of depression. J Voice 2024 Jan;38(1):77-85. [doi: 10.1016/j.jvoice.2021.06.018] [Medline: 34353686]
90. Kim AY, Jang EH, Lee SH, Choi KY, Park JG, Shin HC. Automatic depression detection using smartphone-based text-dependent speech signals: deep convolutional neural network approach. J Med Internet Res 2023 Jan 25;25:e34474. [doi: 10.2196/34474] [Medline: 36696160]
91. Wang Y, Liang L, Zhang Z, et al. Fast and accurate assessment of depression based on voice acoustic features: a cross-sectional and longitudinal study. Front Psychiatry 2023;14:1195276. [doi: 10.3389/fpsyt.2023.1195276]
92. Zhao Q, Fan HZ, Li YL, et al. Vocal acoustic features as potential biomarkers for identifying/diagnosing depression: a cross-sectional study. Front Psychiatry 2022;13:815678. [doi: 10.3389/fpsyt.2022.815678] [Medline: 35573349]
93. Ye J, Yu Y, Wang Q, et al. Multi-modal depression detection based on emotional audio and evaluation text. J Affect Disord 2021 Dec 1;295:904-913. [doi: 10.1016/j.jad.2021.08.090] [Medline: 34706461]
94. Klangpornkun N, Ruangritchai M, Munthuli A, et al. Classification of depression and other psychiatric conditions using speech features extracted from a Thai psychiatric and verbal screening test. Annu Int Conf IEEE Eng Med Biol Soc 2021 Nov;2021:651-656. [doi: 10.1109/EMBC46164.2021.9629571] [Medline: 34891377]
95. Demiroglu C, Beşirli A, Ozkanca Y, Çelik S. Depression-level assessment from multi-lingual conversational speech data using acoustic and text features. J Audio Speech Music Proc 2020 Dec;2020(1):17. [doi: 10.1186/s13636-020-00182-4]
96. Yamamoto M, Takamiya A, Sawada K, et al. Using speech recognition technology to investigate the association between timing-related speech features and depression severity. PLoS ONE 2020;15(9):e0238726. [doi: 10.1371/journal.pone.0238726] [Medline: 32915846]
97. Ali FZ, Parsey RV, Lin S, Schwartz J, DeLorenzo C. Circadian rhythm biomarker from wearable device data is related to concurrent antidepressant treatment response. NPJ Digit Med 2023 Apr 29;6(1):81. [doi: 10.1038/s41746-023-00827-6] [Medline: 37120493]
98. Braund TA, Zin MT, Boonstra TW, et al. Smartphone sensor data for identifying and monitoring symptoms of mood disorders: a longitudinal observational study. JMIR Ment Health 2022 May 4;9(5):e35549. [doi: 10.2196/35549] [Medline: 35507385]
99. Choi JG, Ko I, Han S. Depression level classification using machine learning classifiers based on actigraphy data. IEEE Access 2021;9:116622-116646. [doi: 10.1109/ACCESS.2021.3105393]
100. George SV, Kunkels YK, Booij S, Wichers M. Uncovering complexity details in actigraphy patterns to differentiate the depressed from the non-depressed. Sci Rep 2021 Jun 29;11(1):13447. [doi: 10.1038/s41598-021-92890-w] [Medline: 34188115]
101. Smagula SF, Ancoli-Israel S, Blackwell T, et al. Circadian rest-activity rhythms predict future increases in depressive symptoms among community-dwelling older men. Am J Geriatr Psychiatry 2015 May;23(5):495-505. [doi: 10.1016/j.jagp.2014.06.007] [Medline: 25066948]
102. Maglione JE, Ancoli-Israel S, Peters KW, et al. Depressive symptoms and circadian activity rhythm disturbances in community-dwelling older women. Am J Geriatr Psychiatry 2014 Apr;22(4):349-361. [doi: 10.1016/j.jagp.2012.09.003] [Medline: 23567424]
103. Luik AI, Zuurbier LA, Hofman A, Van Someren EJW, Tiemeier H. Stability and fragmentation of the activity rhythm across the sleep-wake cycle: the importance of age, lifestyle, and mental health. Chronobiol Int 2013 Dec;30(10):1223-1230. [doi: 10.3109/07420528.2013.813528] [Medline: 23971909]
104. Anik IA, Kamal AHM, Kabir MA, Uddin S, Moni MA. A robust deep-learning model to detect major depressive disorder utilizing EEG signals. IEEE Trans Artif Intell 2024;5(10):4938-4947. [doi: 10.1109/TAI.2024.3394792]
105. Tian F, Zhang H, Tan Y, et al. An on-board executable multi-feature transfer-enhanced fusion model for three-lead EEG sensor-assisted depression diagnosis. IEEE J Biomed Health Inform 2025 Jan;29(1):152-165. [doi: 10.1109/JBHI.2024.3487012] [Medline: 39466874]

XSL•FO
RenderX

106. Islam R, Bae SW. FacePsy: an open-source affective mobile sensing system - analyzing facial behavior and head gesture for depression detection in naturalistic settings. Proc ACM Hum-Comput Interact 2024 Sep 24;8(MHCI):1-32. [doi: 10.1145/3676505]
107. Makhmutova M, Kainkaryam R, Ferreira M, Min J, Jaggi M, Clay I. Predicting changes in depression severity using the PSYCHE-D (Prediction of Severity Change-Depression) model involving person-generated health data: longitudinal case-control observational study. JMIR Mhealth Uhealth 2022 Mar 25;10(3):e34148. [doi: 10.2196/34148] [Medline: 35333186]
108. Price GD, Heinz MV, Collins AC, Jacobson NC. Detecting major depressive disorder presence using passively-collected wearable movement data in a nationally-representative sample. Psychiatry Res 2024 Feb;332:115693. [doi: 10.1016/j.psychres.2023.115693] [Medline: 38194801]
109. Bae S, Jang M, Kim GM, et al. Nonlinear associations between physical function, physical activity, sleep, and depressive symptoms in older adults. J Clin Med 2023 Sep 16;12(18):6009. [doi: 10.3390/jcm12186009] [Medline: 37762948]
110. Griffiths C, da Silva KM, Leathlean C, Jiang H, Ang CS, Searle R. Investigation of physical activity, sleep, and mental health recovery in treatment resistant depression (TRD) patients receiving repetitive transcranial magnetic stimulation (rTMS) treatment. J Affect Disord Rep 2022 Apr;8:100337. [doi: 10.1016/j.jadr.2022.100337] [Medline: 35619990]
111. Mishra R, Park C, York MK, et al. Decrease in mobility during the COVID-19 pandemic and its association with increase in depression among older adults: a longitudinal remote mobility monitoring using a wearable sensor. Sensors (Basel) 2021 Apr 29;21(9):3090. [doi: 10.3390/s21093090] [Medline: 33946664]
112. Raoux N, Benoit O, Dantchev N, et al. Circadian pattern of motor activity in major depressed patients undergoing antidepressant therapy: relationship between actigraphic measures and clinical course. Psychiatry Res 1994 Apr;52(1):85-98. [doi: 10.1016/0165-1781(94)90122-8] [Medline: 8047624]
113. Rykov Y, Thach TQ, Bojic I, Christopoulos G, Car J. Digital biomarkers for depression screening with wearable devices: cross-sectional study with machine learning modeling. JMIR Mhealth Uhealth 2021 Oct 25;9(10):e24872. [doi: 10.2196/24872] [Medline: 34694233]
114. Sayar K, Güleç H, Gökçe M, Ismail AK. Heart rate variability in depressed patients. Psychiatry Clin Psychopharmacol 2002;12(3):130-133 [FREE Full text]
115. Landolt HP, Gillin JC. Similar sleep EEG topography in middle-aged depressed patients and healthy controls. Sleep 2005 Feb;28(2):239-247. [doi: 10.1093/sleep/28.2.239] [Medline: 16171249]
116. Hoyos CM, Gordon C, Terpening Z, et al. Circadian rhythm and sleep alterations in older people with lifetime depression: a case-control study. BMC Psychiatry 2020 Apr 29;20(1):192. [doi: 10.1186/s12888-020-02606-z] [Medline: 32349697]
117. Lorenz N, Spada J, Sander C, Riedel-Heller SG, Hegerl U. Circadian skin temperature rhythms, circadian activity rhythms and sleep in individuals with self-reported depressive symptoms. J Psychiatr Res 2019 Oct;117:38-44. [doi: 10.1016/j.jpsychires.2019.06.022] [Medline: 31279242]
118. Vanderlind WM, Beevers CG, Sherman SM, et al. Sleep and sadness: exploring the relation among sleep, cognitive control, and depressive symptoms in young adults. Sleep Med 2014 Jan;15(1):144-149. [doi: 10.1016/j.sleep.2013.10.006] [Medline: 24332565]
119. Ahmed A, Ramesh J, Ganguly S, Aburukba R, Sagahyroon A, Aloul F. Investigating the feasibility of assessing depression severity and valence-arousal with wearable sensors using discrete wavelet transforms and machine learning. Information 2022;13(9):406. [doi: 10.3390/info13090406]
120. Choi J, Lee S, Kim S, Kim D, Kim H. Depressed mood prediction of elderly people with a wearable band. Sensors (Basel) 2022;22(11):4174. [doi: 10.3390/s22114174]
121. Mahendran N, Vincent DR, Srinivasan K, et al. Sensor-assisted weighted average ensemble model for detecting major depressive disorder. Sensors (Basel) 2019 Nov 6;19(22):4822. [doi: 10.3390/s19224822] [Medline: 31698678]
122. Xu X, Chikersal P, Doryab A, et al. Leveraging routine behavior and contextually-filtered features for depression detection among college students. Proc ACM Interact Mob Wearable Ubiquitous Technol 2019 Sep 9;3(3):1-33. [doi: 10.1145/3351274]
123. Lu J, Shang C, Yue C, et al. Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018 Mar 26;2(1):1-21. [doi: 10.1145/3191753]
124. Farhan AA, Yue C, Morillo R, et al. Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. Presented at: 2016 IEEE Wireless Health (WH); Oct 25-27, 2026. [doi: 10.1109/WH.2016.7764553]
125. Karimi S, Nateghi M, Cestero GI, et al. Prescreening depression using wearable electrocardiogram and photoplethysmogram data from a psycholinguistic experiment. Physiol Meas 2025 Aug 2;46(8):085004. [doi: 10.1088/1361-6579/adf6fe] [Medline: 40752509]
126. Jacobson NC, Chung YJ. Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. Sensors (Basel) 2020 Jun 24;20(12):3572. [doi: 10.3390/s20123572] [Medline: 32599801]
127. Wang R, Wang W, Dasilva A, et al. Tracking depression dynamics in college students using mobile phone and wearable sensing. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018 Mar;2(1):1-26. [doi: 10.1145/3191775] [Medline: 39449996]

XSL•FO
RenderX

128. Di Matteo D, Fotinos K, Lokuge S, et al. The relationship between smartphone-recorded environmental audio and symptomatology of anxiety and depression: exploratory study. JMIR Form Res 2020 Aug 13;4(8):e18751. [doi: 10.2196/18751] [Medline: 32788153]
129. Zhou Y, Yao X, Han W, Wang Y, Li Z, Li Y. Distinguishing apathy and depression in older adults with mild cognitive impairment using text, audio, and video based on multiclass classification and shapely additive explanations. Int J Geriat Psychiatry 2022 Nov;37(11):2345-2356. [doi: 10.1002/gps.5827]
130. Williamson JR, Young D, Nierenberg AA, Niemi J, Helfer BS, Quatieri TF. Tracking depression severity from audio and video based on speech articulatory coordination. Comput Speech Lang 2019 May;55:40-56. [doi: 10.1016/j.csl.2018.08.004] [Medline: 32773961]
131. Shah RV, Grennan G, Zafar-Khan M, et al. Personalized machine learning of depressed mood using wearables. Transl Psychiatry 2021 Jun 9;11(1):338. [doi: 10.1038/s41398-021-01445-0] [Medline: 34103481]
132. Opoku Asare K, Moshe I, Terhorst Y, et al. Mood ratings and digital biomarkers from smartphone and wearable data differentiates and predicts depression status: a longitudinal data analysis. Pervasive Mob Comput 2022 Jul;83:101621. [doi: 10.1016/j.pmcj.2022.101621]
133. Narziev N, Goh H, Toshnazarov K, Lee SA, Chung KM, Noh Y. STDD: short-term depression detection with passive sensing. Sensors (Basel) 2020 Mar 4;20(5):1396. [doi: 10.3390/s20051396] [Medline: 32143358]
134. Kim H, Lee S, Lee S, Hong S, Kang H, Kim N. Depression prediction by using ecological momentary assessment, Actiwatch data, and machine learning: observational study on older adults living alone. JMIR Mhealth Uhealth 2019 Oct 16;7(10):e14149. [doi: 10.2196/14149] [Medline: 31621642]
135. Kumagai N, Tajika A, Hasegawa A, et al. Predicting recurrence of depression using lifelog data: an explanatory feasibility study with a panel VAR approach. BMC Psychiatry 2019 Dec 11;19(1):391. [doi: 10.1186/s12888-019-2382-2] [Medline: 31829206]
136. Funkhouser CJ, Weiner LS, Crowley RN, et al. Early changes in passively sensed homestay predict depression symptom improvement during digital behavioral activation. Behav Res Ther 2025 Oct;193:104815. [doi: 10.1016/j.brat.2025.104815] [Medline: 40614686]
137. Di Matteo D, Fotinos K, Lokuge S, et al. Automated screening for social anxiety, generalized anxiety, and depression from objective smartphone-collected data: cross-sectional study. J Med Internet Res 2021 Aug 13;23(8):e28918. [doi: 10.2196/28918] [Medline: 34397386]
138. Sverdlov O, Curcic J, Hannesdottir K, et al. A study of novel exploratory tools, digital technologies, and central nervous system biomarkers to characterize unipolar depression. Front Psychiatry 2021;12:640741. [doi: 10.3389/fpsyt.2021.640741] [Medline: 34025472]
139. Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. Psychiatr Rehabil J 2015 Sep;38(3):218-226. [doi: 10.1037/prj0000130] [Medline: 25844912]
140. Minaeva O, Riese H, Lamers F, Antypa N, Wichers M, Booij SH. Screening for depression in daily life: development and external validation of a prediction model based on actigraphy and experience sampling method. J Med Internet Res 2020 Dec 1;22(12):e22634. [doi: 10.2196/22634] [Medline: 33258783]
141. Dai R, Kannampallil T, Zhang J, Lv N, Ma J, Lu C. Multi-task learning for randomized controlled trials: a case study on predicting depression with wearable data. Proc ACM Interact Mob Wearable Ubiquitous Technol 2022;6(2):1-23. [doi: 10.1145/3534591]
142. Tazawa Y, Liang KC, Yoshimura M, et al. Evaluating depression with multimodal wristband-type wearable device: screening and assessing patient severity utilizing machine-learning. Heliyon 2020 Feb;6(2):e03274. [doi: 10.1016/j.heliyon.2020.e03274] [Medline: 32055728]
143. Thati RP, Dhadwal AS, Kumar P, Sainaba P. Multimodal depression detection: using fusion strategies with smart phone usage and audio-visual behavior. Int J Artif Intell Tools 2023 Mar;32(2):2340008. [doi: 10.1142/S0218213023400080]
144. Jiang Z, Seyedi S, Griner E, et al. Multimodal mental health digital biomarker analysis from remote interviews using facial, vocal, linguistic, and cardiovascular patterns. IEEE J Biomed Health Inform 2024 Mar;28(3):1680-1691. [doi: 10.1109/JBHI.2024.3352075] [Medline: 38198249]
145. Moshe I, Terhorst Y, Opoku Asare K, et al. Predicting symptoms of depression and anxiety using smartphone and wearable data. Front Psychiatry 2021;12:625247. [doi: 10.3389/fpsyt.2021.625247] [Medline: 33584388]
146. Wang R, Chen F, Chen Z, et al. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. Presented at: UbiComp '14: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing; Sep 13-17, 2014. [doi: 10.1145/2632048.2632054]
147. Song S, Seo Y, Hwang S, Kim HY, Kim J. Digital phenotyping of geriatric depression using a community-based digital mental health monitoring platform for socially vulnerable older adults and their community caregivers: 6-week living lab single-arm pilot study. JMIR Mhealth Uhealth 2024 Jun 17;12:e55842. [doi: 10.2196/55842] [Medline: 38885033]
148. Bai R, Xiao L, Guo Y, et al. Tracking and monitoring mood stability of patients with major depressive disorder by machine learning models using passive digital data: prospective naturalistic multicenter study. JMIR Mhealth Uhealth 2021 Mar 8;9(3):e24365. [doi: 10.2196/24365] [Medline: 33683207]

XSL•FO
RenderX

149. Borelli JL, Wang Y, Li FH, et al. Detection of depressive symptoms in college students using multimodal passive sensing data and light gradient boosting machine: longitudinal pilot study. JMIR Form Res 2025 Jun 3;9:e67964. [doi: 10.2196/67964] [Medline: 40460426]
150. Chen J, Chan NY, Li CT, et al. Multimodal digital assessment of depression with actigraphy and app in Hong Kong Chinese. Transl Psychiatry 2024 Mar 18;14(1):150. [doi: 10.1038/s41398-024-02873-4] [Medline: 38499546]
151. Shen S, Qi W, Zeng J, et al. Passive sensing for mental health monitoring using machine learning with wearables and smartphones: scoping review. J Med Internet Res 2025 Aug 14;27:e77066. [doi: 10.2196/77066] [Medline: 40811794]
152. Lee YJ, Lee JY, Cho JH, Kang YJ, Choi JH. Performance of consumer wrist-worn sleep tracking devices compared to polysomnography: a meta-analysis. J Clin Sleep Med 2025 Mar 1;21(3):573-582. [doi: 10.5664/jcsm.11460] [Medline: 39484805]
153. Poon K, Ho MSH, Tai APL, Leung MK, Chan MCM, Lau WKW. Mediating roles of sleep quality and resilience in the relationships between chronotypes and mental health symptoms. Sci Rep 2024 Mar 11;14(1):5874. [doi: 10.1038/s41598-024-56688-w] [Medline: 38467740]
154. Röver C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. BMC Med Res Methodol 2015 Nov 14;15:99. [doi: 10.1186/s12874-015-0091-1] [Medline: 26573817]
155. Su Y, Li M, Caron J, Li D, Meng X. Differential effects of lifetime stressors on major depressive disorder severity: a longitudinal community-based cohort study. Eur Psychiatry 2024 Oct 4;67(1):e66. [doi: 10.1192/j.eurpsy.2024.1783] [Medline: 39363747]
156. Price GD, Heinz MV, Song SH, Nemesure MD, Jacobson NC. Using digital phenotyping to capture depression symptom variability: detecting naturalistic variability in depression symptoms across one year using passively collected wearable movement and sleep data. Transl Psychiatry 2023 Dec 9;13(1):381. [doi: 10.1038/s41398-023-02669-y] [Medline: 38071317]
157. Yasugaki S, Okamura H, Kaneko A, Hayashi Y. Bidirectional relationship between sleep and depression. Neurosci Res 2025 Feb;211:57-64. [doi: 10.1016/j.neures.2023.04.006] [Medline: 37116584]
158. Meng Y, Ma N, Shi Y, et al. The association of physical activity and sedentary behavior with depression in US adults: NHANES 2007-2018. Front Public Health 2024;12:1404407. [doi: 10.3389/fpubh.2024.1404407] [Medline: 38975361]
159. Hayashi M, Takeshima M, Hosoya T, Kume Y. 24-hour rest-activity rhythm in middle-aged and older persons with depression. Int J Environ Res Public Health 2023 Mar 27;20(7):5275. [doi: 10.3390/ijerph20075275] [Medline: 37047891]
160. Edler JS, Winter M, Steinmetz H, Cohrdes C, Baumeister H, Pryss R. Predicting depressive symptoms using GPS-based regional data in Germany with the CORONA HEALTH app during the COVID-19 pandemic: cross-sectional study. Interact J Med Res 2024 Dec 3;13:e53248. [doi: 10.2196/53248] [Medline: 39625745]
161. Alshanskaia EI, Zhozhikashvili NA, Polikanova IS, Martynova OV. Heart rate response to cognitive load as a marker of depression and increased anxiety. Front Psychiatry 2024;15:1355846. [doi: 10.3389/fpsyt.2024.1355846] [Medline: 39056018]
162. Xu L, Zhai X, Shi D, Zhang Y. Depression and coronary heart disease: mechanisms, interventions, and treatments. Front Psychiatry 2024;15:1328048. [doi: 10.3389/fpsyt.2024.1328048] [Medline: 38404466]
163. Monaco F, Vignapiano A, Piacente M, et al. Innova4Health: an integrated approach for prevention of recurrence and personalized treatment of major depressive disorder. Front Artif Intell 2024;7:1366055. [doi: 10.3389/frai.2024.1366055] [Medline: 38774832]
164. Choo M, Park D, Cho M, Bae S, Kim J, Han DH. Exploring a multimodal approach for utilizing digital biomarkers for childhood mental health screening. Front Psychiatry 2024;15:1348319. [doi: 10.3389/fpsyt.2024.1348319] [Medline: 38666089]
165. Mandryk RL, Birk MV, Vedress S, et al. Remote assessment of depression using digital biomarkers from cognitive tasks. Front Psychol 2021;12:767507. [doi: 10.3389/fpsyg.2021.767507] [Medline: 34975656]
166. Hurwitz E, Butzin-Dozier Z, Master H, et al. Harnessing consumer wearable digital biomarkers for individualized recognition of postpartum depression using the All of Us research program data set: cross-sectional study. JMIR Mhealth Uhealth 2024 May 2;12:e54622. [doi: 10.2196/54622] [Medline: 38696234]
167. Chen J, Lai TF, Lin LJ, Park JH, Liao Y. Is overall and timing-specific physical activity associated with depression in older adults? Front Public Health 2023;11:567123. [doi: 10.3389/fpubh.2023.1241170]
168. Hernández-Vicente A, Hernando D, Marín-Puyalto J, et al. Validity of the Polar H7 heart rate sensor for heart rate variability analysis during exercise in different age, body composition and fitness level groups. Sensors (Basel) 2021 Jan 29;21(3):902. [doi: 10.3390/s21030902] [Medline: 33572800]
169. Sato W, Murata K, Uraoka Y, Shibata K, Yoshikawa S, Furuta M. Emotional valence sensing using a wearable facial EMG device. Sci Rep 2021 Mar 11;11(1):5757. [doi: 10.1038/s41598-021-85163-z] [Medline: 33707605]
170. Caldiroli A, Capuzzi E, Affaticati LM, et al. Candidate biological markers for social anxiety disorder: a systematic review. Int J Mol Sci 2023 Jan 3;24(1):835. [doi: 10.3390/ijms24010835] [Medline: 36614278]

Abbreviations

BDI: Beck Depression Inventory

XSL•FO
RenderX

CES-D: Center for Epidemiologic Studies Depression Scale
ECG: electrocardiogram
GDS: Geriatric Depression Scale
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
HDRS: Hamilton Depression Rating Scale
HF: high frequency
HKSJ: Hartung-Knapp-Sidik-Jonkman
HR: heart rate
HRV: heart rate variability
IDAS: Inventory of Depression and Anxiety Symptoms
IDS-SR: Inventory of Depressive Symptomatology
IRB: Institutional Review Board
LF: low frequency
LPA: light physical activity
MD: mean difference
MDD: major depressive disorder
MET: metabolic equivalent of task
mHR: mean heart rate
MVPA: moderate-to-vigorous physical activity
NREM: non–rapid eye movement
PHQ: Patient Health Questionnaire
pNN: proportion of normal-to-normal intervals
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2
REM: rapid eye movement
RMSSD: root mean square of successive differences
RRI: RR interval
SDNN: SD of normal-to-normal intervals
SE: sleep efficiency
SGDS-K: shortened Korean version of the Geriatric Depression Scale
SIGN: Scottish Intercollegiate Guidelines Network
SMD: standardized mean difference
SOL: sleep onset latency
SWS: slow-wave sleep
TIB: time in bed
TST: total sleep time
ULF: ultra-low frequency
VLF: very low frequency
WASO: wake after sleep onset

Edited by S Brini; submitted 23.Apr.2025; peer-reviewed by S Oh, V Manea; revised version received 11.Feb.2026; accepted 13.Feb.2026; published 02.Apr.2026.

Please cite as:
Lee H, Kang SG, Lee S
The Role of Digital Biomarkers in Physiological Signal-Based Depression Assessment: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e76432
URL: https://www.jmir.org/2026/1/e76432
doi:10.2196/76432

© Hyeongsuk Lee, Seung-Gul Kang, SeonHeui Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Accuracy of Radiomics-Based Machine Learning for Predicting Risk of Recurrence in Non–Small Cell Lung Cancer: Systematic Review and Meta-Analysis

Junpei Wu[1,2], MSc; Ye Zhang[3], MSc; Jiaye Wang[4], MSc; Chengshui Chen[5], PhD; Shiyu Hu[2*], MSc; Wenyu Chen[2,5*], PhD

[1]Department of Pulmonary and Critical Care Medicine, Yiwu Central Hospital, Yiwu, China

[2]Department of Respiratory, The Affiliated Hospital of Jiaxing University, 1882 Zhonghuan South Road, Nanhu District, Jiaxing, China

[3]Health Management Center, The Affiliated Hospital of Jiaxing University, Jiaxing, China

[4]Department of Internal Medicine, Zhejiang Chinese Medical University, Hangzhou, China

[5]Department of Respiratory, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China

[*]these authors contributed equally

Corresponding Author:
Wenyu Chen, PhD
Department of Respiratory, The Affiliated Hospital of Jiaxing University, 1882 Zhonghuan South Road, Nanhu District, Jiaxing, China

Abstract

Background: During the diagnosis and treatment of non–small cell lung cancer (NSCLC), detecting the risk of its recurrence in an early phase is still challenging. Recent studies have investigated the radiomics-based machine learning (ML) models for detecting the risk of recurrence in NSCLC. However, there is still insufficient systematic evidence to prove its efficiency.

Objective: This study is designed to systematically evaluate the effectiveness of radiomics-based ML in predicting the risk of recurrence in NSCLC, aiming to provide evidence-based support for the subsequent development of scoring tools to forecast recurrence risk.

Methods: For acquiring research on radiomics-based models for forecasting the risk of recurrence in NSCLC, Cochrane Library, Web of Science, PubMed, and Embase were systematically retrieved, up to October 24, 2025. Studies on analyzing the recurrence of NSCLC using radiomics-based ML were included, while those in which only texture analysis was conducted or radiomics-based ML was not constructed were excluded. The Radiomics Quality Score (RQS) was used to appraise the eligible studies. Subgroup analyses were conducted according to the variables of the model, the background of treatment, the stage of lung cancer, and the pathological type.

Results: Ultimately, 30 eligible studies in total were included, covering 7964 patients with NSCLC. According to the meta-analysis, the c-index of radiomics-based ML models for forecasting the risk of recurrence in NSCLC was 0.850 (95% CI 0.834 - 0.866, 95% prediction interval [PI] 0.623 - 1.004) in the training set. Specifically, the pooled c-index was 0.876 (95% CI 0.853 - 0.900) among the patients receiving the stereotactic body radiation therapy and 0.825 (95% CI 0.804 - 0.848) among those who received surgeries combined with other adjuvant treatment regimens. The c-index of the radiomics-based ML models combined with clinical features for forecasting the risk of recurrence in NSCLC was 0.833 (95% CI 0.822 - 0.854, 95% PI 0.717 - 0.945) in the training set. In contrast, the c-index of radiomics-based ML models for forecasting the risk of recurrence in NSCLC was 0.878 (95% CI 0.854 - 0.902, 95% PI 0.681 - 1.000) in the validation set. The c-index of radiomics-based ML models combined with clinical features for forecasting the risk of recurrence in NSCLC was 0.854 (95% CI 0.830 - 0.878, 95% PI 0.655 - 0.992) in the validation set. The average RQS across the included studies was 27.4%, revealing methodological limitations and an absence of standardization.

Conclusions: This study is the first to confirm that radiomics-based ML models effectively predict the risk of recurrence in NSCLC. This study provides evidence-based support for the subsequent development or updating of radiomics-based ML models. However, the current methodological application of radiomics remains concerning. Therefore, in the future, research should standardize the workflow for implementing radiomics-based ML and incorporate multicenter imaging data to enhance its generalizability.

Trial Registration: PROSPERO CRD42025631191; https://www.crd.york.ac.uk/PROSPERO/view/CRD42025631191

(J Med Internet Res 2026;28:e84223) doi:10.2196/84223

XSL•FO
RenderX

KEYWORDS

non–small cell lung cancer; NSCLC; machine learning; ML; radiomics; recurrence; prognostic prediction

Introduction

As documented in the Cancer Statistics Report 2024, lung cancer serves as a major contributor to global cancer-related mortality and ranks second in the incidence of new cancer cases among male individuals [1]. With an 85% incidence among individuals with lung cancer, non–small cell lung cancer (NSCLC) emerges as the most frequently observed pathological subtype, and its incidence is gradually increasing [2]. Hence, enhancing the prevention, screening, and management of NSCLC is clinically significant.

The current treatment methods for NSCLC are constantly evolving. Early-stage NSCLC surgeries and radiotherapy, which are traditional treatment approaches, are still widely used and have helped numerous patients achieve promising therapeutic outcomes. In recent years, emerging therapies such as molecular targeted therapy, antitumor angiogenesis therapy, and immunotherapy [3-5] for NSCLC have significantly progressed. Those innovative treatment methods offer patients more options and significantly improve their survival rates and quality of life. Although those treatment methods have led to significant survival benefits for patients with NSCLC, there are still some serious adverse prognoses, posing significant challenges in the formulation of treatment plans in clinical practice [6,7]. In particular, recurrence is a serious adverse outcome in NSCLC treatment. According to the epidemiological statistics on lung cancer, the 5-year cumulative recurrence rate of NSCLC (stage I) after surgery is 25% to 30%, and it is as high as 60% to 70% for NSCLC (stage IIIA) [8,9]. Therefore, early identification of the risk of recurrence in NSCLC is clinically significant for formulating individualized therapeutic schedules as well as rehabilitation plans.

Unfortunately, up to now, there has been a lack of tools for effectively predicting the recurrence of NSCLC in clinical practice. In recent years, with artificial intelligence evolving, radiomics has received widespread attention in clinical practice. Radiomics refers to the process of converting a group of high-data-content medical images into digital ones and combining machine learning (ML) to aid in the diagnosis of diseases, prediction of progression, and prognosis of diseases [10]. Radiomics was originally proposed by Lambin et al [11] from the Netherlands in 2012. In comparison with traditional models, radiomics-based models are capable of extracting imaging features through high-throughput methods, which can more comprehensively reflect the spatial heterogeneity of tumors and are equipped with higher information dimensions and discriminative capabilities. Therefore, it has shown significant advantages in multiple studies. The study by Chen et al [12] has revealed a high c-index of radiomics-based ML models for forecasting the risk of recurrence in NSCLC (the stage of early epidermal growth factor receptor mutation). In the study by Yan et al [13], a radiomics-based ML is effective for predicting lymph node metastasis in NSCLC.

Beyond conventional ML algorithms based on radiomics, recent research has also explored more advanced deep learning techniques and AI algorithms capable of perceiving uncertainty for medical prediction. These models use dimensionality and techniques of size reduction for feature selection, maintaining robustness even in datasets with incomplete or noisy information. For instance, deep ML algorithms integrating feature elimination and dimensionality reduction have been successfully applied to the diagnosis of thyroid cancer with randomly missing data. Additionally, big data-driven supervised ML models based on similar mechanisms have demonstrated high diagnostic performance in nonalcoholic steatohepatitis [14,15]. These advances highlight the potential of incorporating high-dimensional feature learning and uncertainty modeling into frameworks of medical prediction. However, their application in forecasting the recurrence of NSCLC remains limited, and no systematic review has been conducted to synthesize current evidence.

Against this backdrop, some researchers have sought to establish radiomics-based ML models to forecast the risk of recurrence in NSCLC. However, current studies are restricted by limited sample size, heterogeneous design, and data sourced from a single center [16]. These studies are absent from systematic quantitative integration and heterogeneity analysis, thereby failing to provide high-quality and evidence-based support. This poses a challenge to the development of intelligent models for predicting the risk of recurrence in NSCLC. Therefore, this study is designed to systematically evaluate the predictive performance of radiomics-based ML models for the recurrence risk of NSCLC and their current application in this field. This study provides evidence-based support for the future standardized application of radiomics and the development or update of prediction and assessment tools based on radiomics-based ML.

Methods

Registration

This study has been prospectively registered on the International Prospective Register of Systematic Reviews (PROSPERO) and was implemented strictly following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines to guarantee transparency and repeatability (Checklist 1) [17].

Eligibility Criteria

Inclusion Criteria

The inclusion criteria were as follows: (1) individuals pathologically or clinically diagnosed with NSCLC; (2) studies on the recurrence in NSCLC using radiomics-based ML; (3) studies reporting outcome indicators of the accuracy of ML; (4) cohort studies, case-control studies, or cross-sectional studies; and (5) studies reported in English.

XSL•FO
RenderX

Exclusion Criteria

The exclusion criteria were as follows: (1) unpublished conference abstracts were included; (2) in studies, an ML model only based on clinical features was constructed for predicting the risk of recurrence in NSCLC, and a radiomics-based ML model was not established; (3) studies failed to strictly distinguish NSCLC from other types of lung cancers; and (4) in studies, outcome indicators were absent from appraising the predictive precision of an ML model, including c-index, sensitivity, specificity, accuracy, recovery rate, precision rate, confusion matrix, and F_1-score.

Data Sources and Search Strategy

To ensure transparency, the literature search was conducted in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search Extension) guidelines (Checklist 2) [18]. PubMed, Cochrane Library, Embase, and Web of Science were systematically searched for studies on predicting NSCLC recurrence based on radiomics from the inception of the databases to October 24, 2025, with no restrictions on region or year of publication. No specialized study registries were searched, nor were relevant experts contacted to obtain unpublished data. The reference lists of all included studies were manually checked (backward searching) to avoid missing potentially relevant studies. The search strategy was a combination of subject terms with relevant free terms. The subject terms comprised lung neoplasms (MeSH), machine learning (MeSH), and recurrence (MeSH). The search strategy for each database and the retrieval process are detailed in Table S2 in Multimedia Appendix 1 and Checklist 2. Prepublished search filters or search strategies for other literature reviews were employed, irrespective of language or study design.

Literature Screening and Data Collection

The retrieved studies were imported into EndNote (Clarivate). After the elimination of duplicates based on title, author, and publication year, the studies were systematically reviewed for their titles and abstracts. Ultimately, the eligible studies were determined through a full-text review. In the included studies, recurrence was defined as recurrence of tumors within or adjacent to the treatment zone, mediastinal recurrence, or distant metastasis following curative-intent treatment such as surgery, chemoradiotherapy, or stereotactic body radiotherapy. Recurrence was generally identified according to the criteria of response evaluation criteria in solid tumors or confirmed via histopathology. To ensure accuracy, in several studies, double-blind evaluation was performed independently by 2 radiologists or oncologists, with disagreements resolved through negotiation. Only studies reporting a clear definition of recurrence and an explicit criterion for follow-up were included in this meta-analysis. To address the high dimensionality of radiomics data, procedures of multistep feature selection and dimensionality reduction were used across the included studies. Commonly applied methods included reproducibility filtering via intraclass correlation coefficient, redundancy removal based on correlation, univariate statistical screening, and feature selection based on ML, such as least absolute shrinkage and selection operator, recursive feature elimination, or principal component analysis. These approaches reduced feature space, enhanced the steadiness of models, thereby ensuring that only the most robust and clinically relevant features were utilized in the construction of prediction models.

Before extracting the data, a standardized spreadsheet for data extraction was created, including title, DOI, first author, country of first author, publication year, study design, follow-up duration, source of radiomics, number of participants involved in image segmentation, image segmentation software, case number of NSCLC in the training set, case number of NSCLC in the validation set, method of generating the validation set, presence or absence of overfitting, total number of participants in the validation set, method of variable selection, method of model construction, presence or absence of a radiomics score, presence or absence of other features combined for model construction (such as clinical features and genetic features), presence or absence of assessment of clinical applicability, presence or absence of evaluation of model calibration, presence or absence of overfitting assessment of the model, c-index, sensitivity, specificity, and diagnostic contingency table.

Literature screening and data extraction were implemented independently by two investigators (J Wu and SH), followed by a cross-verification to ensure consistency. All discrepancies were resolved by a third researcher (YZ) to reach a consensus.

Evaluation for Study Quality

Two investigators used the Radiomics Quality Score (RQS) to assess the methodological quality and analyze risk of bias among the included studies, followed by a cross-verification. Any discrepancy was resolved by a third researcher to make a final decision. The RQS covered 16 items and described the entire process of radiomics, with a maximum score of 36 points.

Synthesis Methods

Meta-analysis was performed for the c-index, an index for appraising the overall precision of ML models. In some original studies, when a 95% CI and a SE of the c-index were not reported, the SEs were estimated based on the study by Debray et al [19]. Among the ML models, discrepancies were identified in the selected variables, with parameters varying. Therefore, in terms of meta-analysis of the c-index, this study privileged the random effects model (REM). Furthermore, the results were calibrated using the Hartung-Knapp-Sidik-Jonkman approach, and the prediction intervals were presented. Subsequently, the bivariate mixed effects model was additionally used to conduct a meta-analysis for sensitivity and specificity. The meta-analysis was carried out via Stata 15.0 (StataCorp LLC).

Results

Literature Selection

A total of 5423 reports were retrieved, among which 1007 duplicate studies were removed, with 4416 studies remaining. A total of 4383 studies were screened out according to the titles and abstracts, due to irrelevance to the study topic or an unmatched study design. The full texts of the remaining 33 studies that were preliminarily in line with the inclusion criteria were obtained. Through the review of the full texts, 2 unpublished conference abstracts and 1 study without outcome

XSL•FO
RenderX

indicators were excluded. Ultimately, 30 studies were included. The process of literature selection is depicted in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart for literature selection.

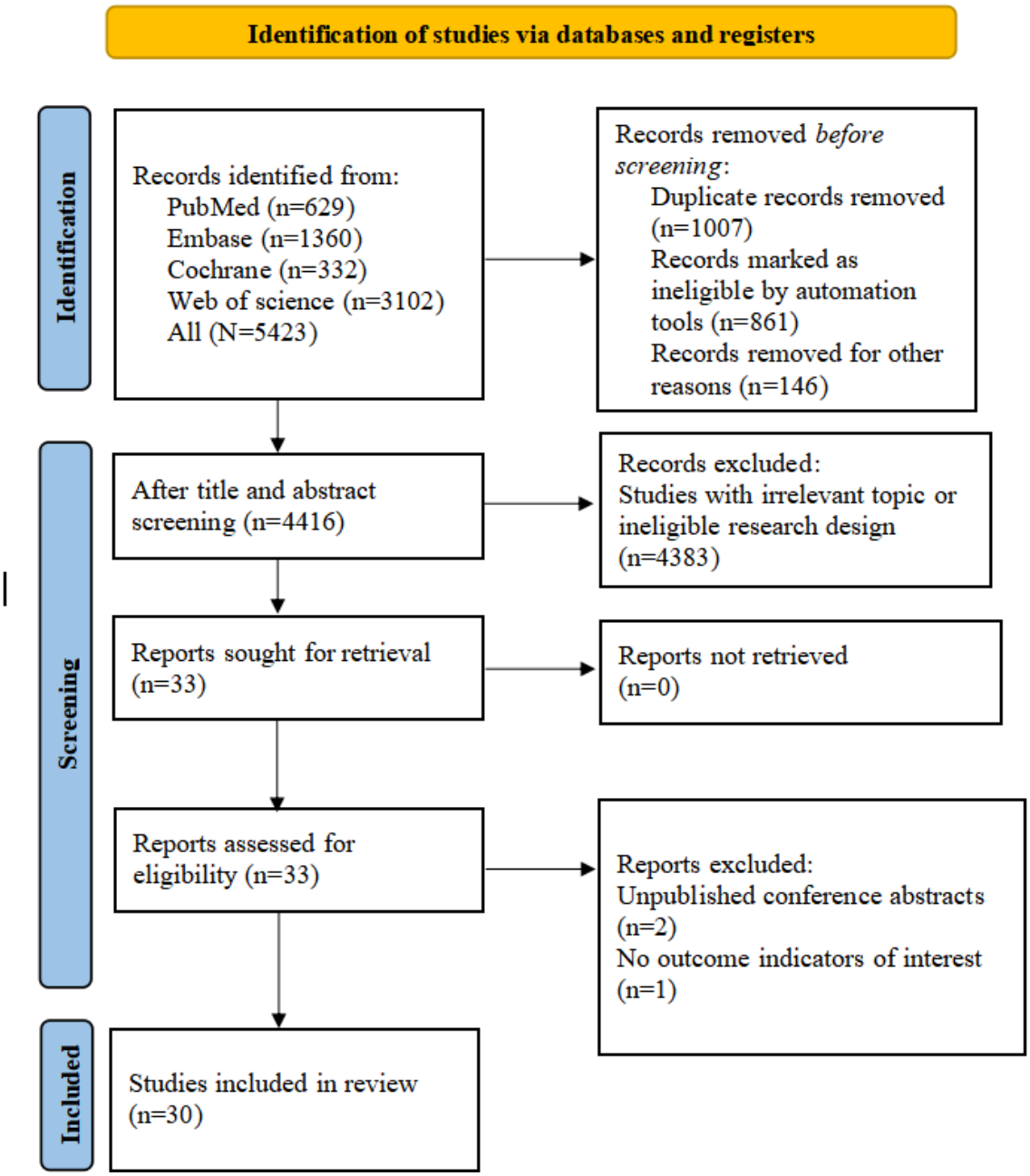

Features of the Included Studies

Those 30 selected studies were all published between 2019 and 2025, in 9 countries. The study types were all cohort studies. The study data were from multiple centers in 11 studies, from a single center in 17 studies, and from a registration database in 2 studies. A total of 8 studies clearly delineated the pathological types of lung adenocarcinoma and/or lung squamous cell carcinoma, while 22 did not delineate the pathological type of the tumor, merely describing it as NSCLC. A total of 19 studies included early-stage lung cancer (based on tumor node metastasis [TNM] staging in stages I-II), 9 included locally advanced or advanced lung cancer (based on TNM staging in stages III-IV), and 2 did not specify the tumor stage. Among the included studies, 16 discussed tumor recurrence after the stereotactic body radiation therapy (SBRT), and 22 discussed tumor recurrence after surgery and/or radiotherapy and chemotherapy. The follow-up time was mainly concentrated between 2 and 70 months. The image sources of the 26 studies mainly included computed tomography (CT) and positron emission tomography–computed tomography (PET-CT). Among them, 24 studies used CT, 2 studies used PET-CT, and 4 studies used both CT and PET-CT. The generation method of the validation set mainly involved random sampling and cross-validation. Among them, 6 studies included external validation (Table S2 in Multimedia Appendix 1).

Evaluation of the Quality of the Selected Studies

None of the included studies described the use of multiple segmentation approaches, repeated measurements at different

XSL•FO
RenderX

time points, or pretests with varying image parameters. None of them conducted prospective registration, provided a description of the degree of consistency with the "gold standard" method, or conducted cost-effectiveness analysis. Therefore, none of the included studies received any points in these categories. In 7 studies, only 1 doctor was involved in the image segmentation, and thus, they did not receive any points. Only the study by Aonpong et al [20] discussed the correlation between radiomics and tumor biology. A total of 16 studies did not report the radiomics score and therefore received a score of 0 in this category. All the included literature reported the discriminant statistics and their statistical significance, but only 15 studies applied the multiple sampling method. Accordingly, in this domain, 15 studies were awarded 2 points, while the remaining studies received 1 point. A total of 5 studies reported calibration statistics and were awarded 2 points, whereas the remaining studies received no points. Among the 30 included studies, all verified the radiomics features, with 1 study using 3 external datasets for verification, another study using 2 external datasets, and 7 studies using 1 external dataset for verification. A total of 7 studies performed decision curve analysis to evaluate the potential clinical utility of the models, each receiving 2 points. A total of 24 studies disclosed images and segmentation, earning 2 points each. One study disclosed data codes, receiving 3 points. A total of 5 studies did not disclose codes, data, or segmentation methods and therefore received no points. The final average score was 27.5% (Table S3 in Multimedia Appendix 1).

Meta-Analysis: Radiomics-Based ML

In the training set, the c-index was pooled by REM (τ^2=0.0116; τ=0.1035; I^2=88.3%), and the c-index of the radiomics-based ML for predicting the risk of recurrence in NSCLC was 0.850 (95% CI 0.834 - 0.866, 95% prediction interval [PI] 0.623 - 1.000; Figure 2).

Figure 2. Forest plot for meta-analysis of the c-index of radiomics-based machine learning (ML) for predicting risk of recurrence in non–small cell lung cancer (NSCLC) in the training set. The forest plot was analyzed using a random effects model and corrected by the Knapp-Hartung method [20-35]. DBP: diastolic blood pressure; OR: odds ratio; SBP: systolic blood pressure.

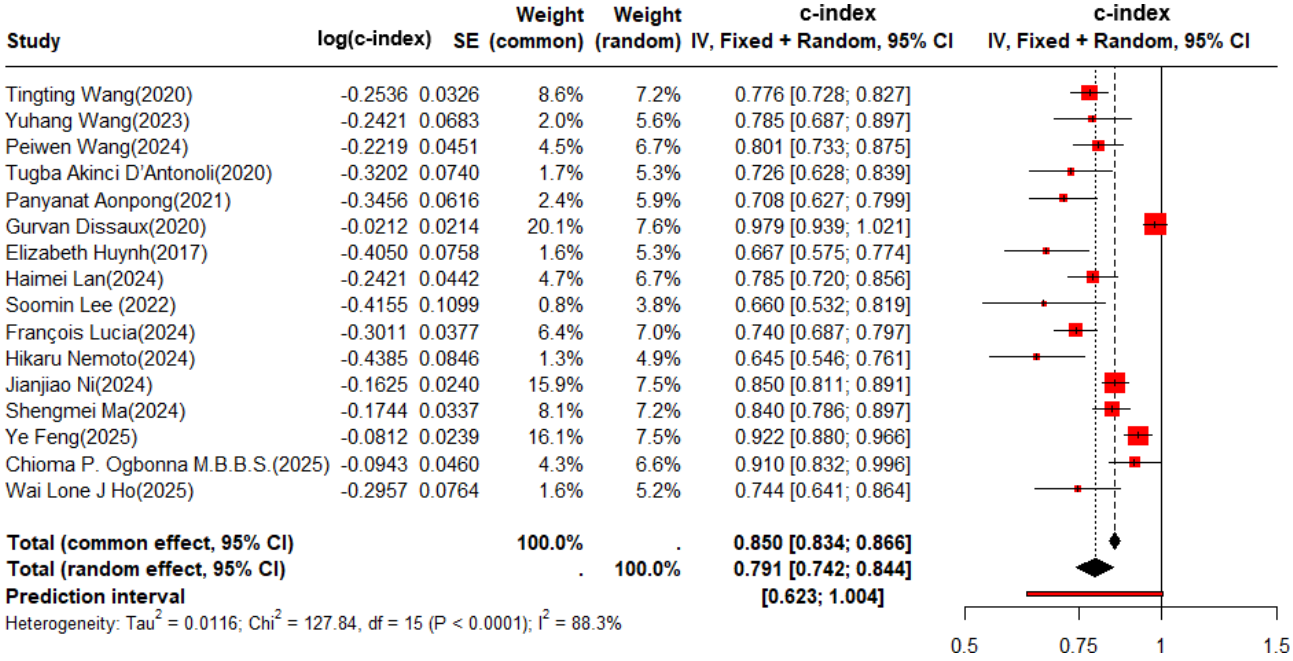

The pooled c-index was 0.876 (95% CI 0.853 - 0.900) among the patients receiving SBRT, and 0.825 (95% CI 0.804 - 0.848) among those who received surgeries combined with other adjuvant treatment regimens (Figure S1 in Multimedia Appendix 1). The pooled c-index was 0.815 (95% CI 0.776 - 0.856) among the patients with adenocarcinoma and squamous cell carcinoma as the pathological type, and 0.856 (95% CI 0.839 - 0.874) among other patients with unspecified specific pathological types (Figure S2 in Multimedia Appendix 1). The pooled c-index was 0.851 (95% CI 0.832 - 0.870) among the patients with early-stage NSCLC, and 0.848 (95% CI 0.818 - 0.879) among the patients at a locally advanced stage and an advanced stage (Figure S3 in Multimedia Appendix 1).

In the validation set, the c-index was pooled by REM (τ^2=0.0072; τ=0.0750; I^2=69.3%), and the c-index of the radiomics-based ML for predicting the risk of recurrence in NSCLC was 0.878 (95% CI: 0.854 - 0.902, 95% PI 0.681 - 1.000; Figure 3).

XSL•FO
RenderX

Figure 3. Forest plot for meta-analysis of the c-index of radiomics-based machine learning for predicting risk of recurrence in non–small cell lung cancer in the validation set. The forest plot was analyzed using a random effects model and corrected by the Knapp-Hartung method [21-30,36]. IV: independent variable.

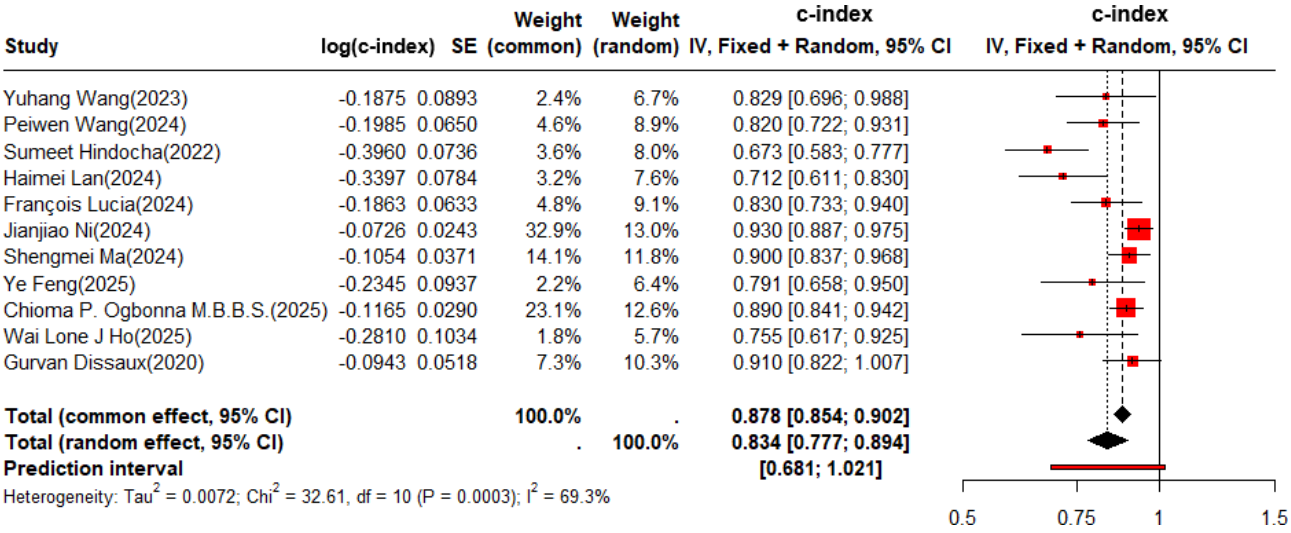

The pooled c-index was 0.907 (95% CI 0.877 - 0.937) among the patients receiving SBRT, and 0.819 (95% CI 0.780 - 0.859) among those who received surgeries combined with other adjuvant treatment regimens (Figure S4 in Multimedia Appendix 1). The pooled c-index was 0.858 (95% CI 0.807 - 0.913) among the patients with adenocarcinoma-squamous cell carcinoma as the pathological type, and 0.883 (95% CI 0.856 - 0.910) among other patients with unspecified specific pathological types (Figure S5 in Multimedia Appendix 1). Among the patients with early-stage NSCLC, the pooled c-index was 0.896 (95% CI 0.870 - 0.922), and among the patients at a locally advanced stage and an advanced stage, the pooled c-index was 0.759 (95% CI 0.702 - 0.821; Figure S6 in Multimedia Appendix 1). In the subgroup with fewer than 100 cases, the pooled c-index was 0.851 (95% CI 0.815 - 0.888), while in the subgroup with more than 100 cases, it was 0.897 (95% CI 0.865 - 0.929; Figure S7 in Multimedia Appendix 1).

ML Based on Radiomics Combined With Clinical Features

In the training set, the c-index was pooled by REM (τ^2=0.0042; τ=0.0628; I^2=74.9%), and the c-index of the ML based on radiomics combined with clinical features for predicting the risk of recurrence in NSCLC was 0.833 (95% CI 0.822 - 0.854, 95% PI 0.717 - 0.945; Figure 4).

XSL•FO
RenderX

Figure 4. Forest plot for meta-analysis of the c-index of radiomics-based machine learning combined with clinical features for predicting risk of recurrence in non–small cell lung cancer in the training set. The forest plot was analyzed using a random effects model and corrected by the Knapp-Hartung method [20-22,24-27,30-32,37-48]. IV: independent variable.

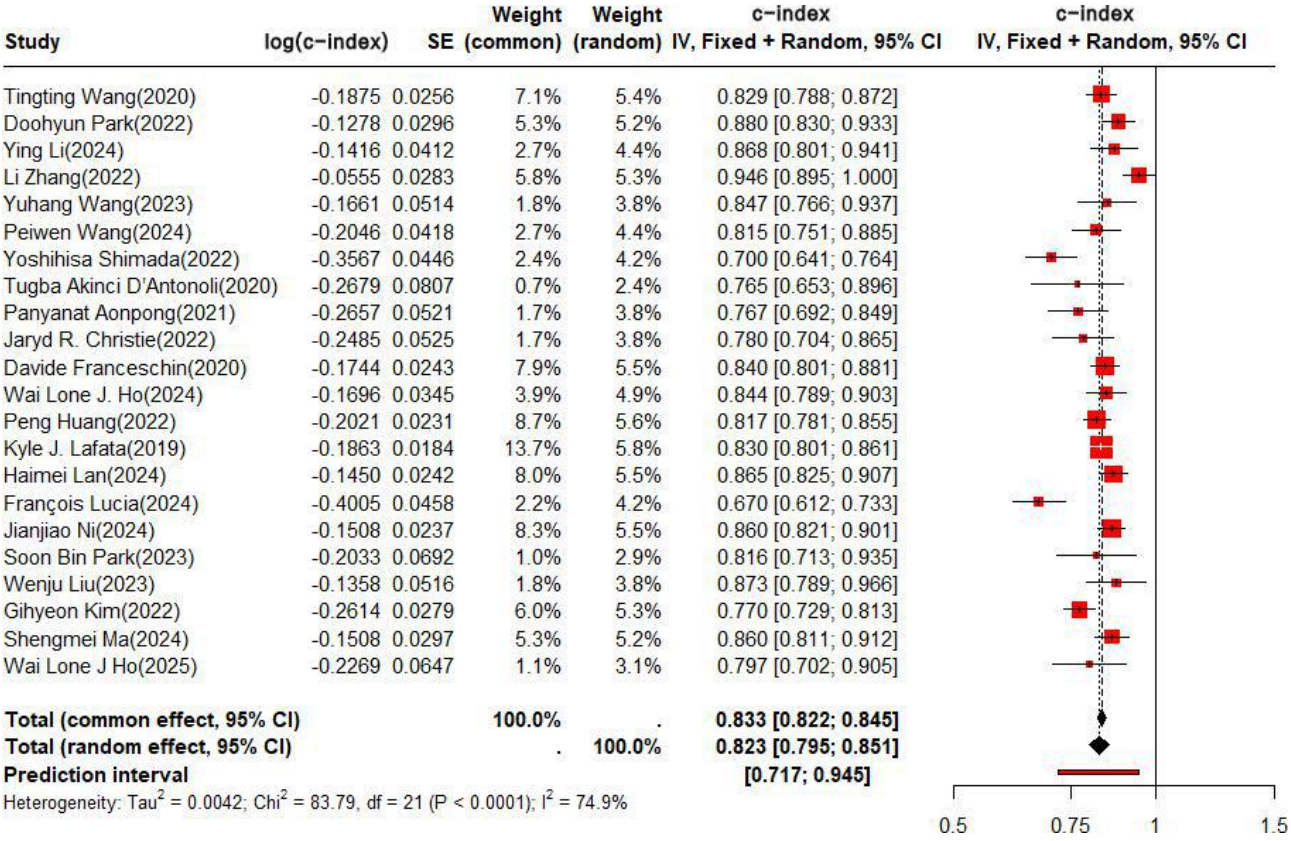

The pooled c-index was 0.828 (95% CI 0.808 - 0.848) among the patients receiving SBRT, and 0.836 (95% CI 0.823 - 0.850) among the patients who received surgeries combined with other adjuvant treatment regimens (Figure S8 in Multimedia Appendix 1). The pooled c-index was 0.882 (95% CI 0.858 - 0.907) among the patients with adenocarcinoma-squamous cell carcinoma, and 0.823 (95% CI 0.810 - 0.837) among other patients with unspecified specific pathological types (Figure S9 in Multimedia Appendix 1). Among the NSCLC patients, the pooled c-index was 0.842 (95% CI 0.829 - 0.855), and among the patients at a locally advanced stage and an advanced stage, the pooled c-index was 0.802 (95% CI 0.779 - 0.826; Figure S10 in Multimedia Appendix 1).

In the validation set, the c-index was pooled by REM (τ^2=0.0079; τ=0.1069; I^2=65.1%), and the c-index of the ML based on radiomics combined with clinical features for predicting the risk of recurrence in NSCLC was 0.854 (95% CI 0.830 - 0.878, 95% PI 0.655 - 0.992; Figure 5).

XSL•FO
RenderX

Figure 5. Forest plot for meta-analysis of the c-index of radiomics-based machine learning combined with clinical features for predicting risk of recurrence in non–small cell lung cancer in the validation set. The forest plot was analyzed using a random effects model and corrected by the Knapp-Hartung method [21,22,24-27,30,36,38,39,41,44,46]. IV: independent variable.

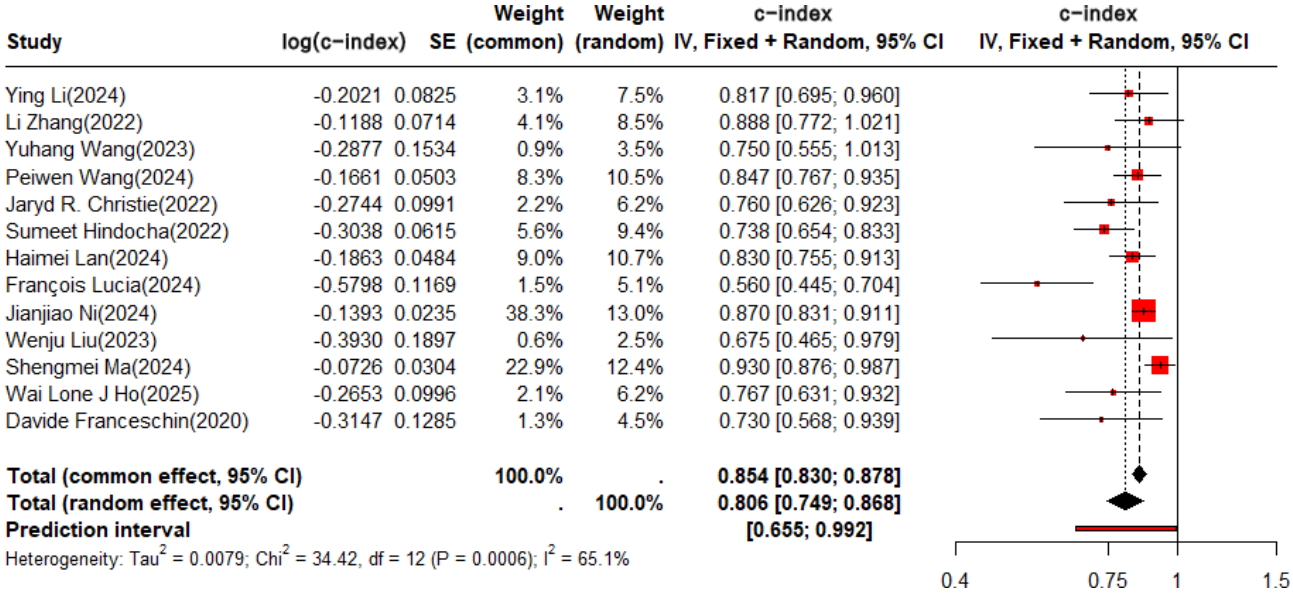

The pooled c-index was 0.851 (95% CI 0.814 - 0.890) among the patients receiving SBRT, and 0.855 (95% CI 0.8824 - 0.888) among the patients who received surgeries combined with other adjuvant treatment regimens (Figure S11 in Multimedia Appendix 1). The pooled c-index was 0.889 (95% CI 0.850 - 0.930) among the patients with adenocarcinoma-squamous cell carcinoma, and 0.831 (95% CI 0.801 - 0.862) among other patients with unspecified specific pathological types (Figure S12 in Multimedia Appendix 1). Among the patients with early-stage NSCLC, the pooled c-index was 0.870 (95% CI 0.842 - 0.897), and among the patients at a locally advanced stage and an advanced stage, the pooled c-index was 0.788 (95% CI 0.738 - 0.842; Figure S13 in Multimedia Appendix 1). The pooled c-index in the subgroup with fewer than 100 cases was 0.855 (95% CI 0.823 - 0.888), while in the subgroup with more than 100 cases, it was 0.852 (95% CI 0.816 - 0.889; Figure S14 in Multimedia Appendix 1).

Discussion

Principal Findings

That meta-analysis of 30 cohort studies demonstrated that the radiomics-based ML achieved pooled c-indexes of 0.850 in the training set and 0.878 in the validation set for predicting recurrence in NSCLC. These findings indicate that radiomics, as a noninvasive and reproducible imaging analysis technique, holds clinical potential for predicting the risk of recurrence in NSCLC. Most studies, for the purpose of improving the prediction accuracy and strengthening the interpretability of the radiomics-based ML, tend to combine clinical features with radiomics features. However, this study revealed that although these attempts indeed enhanced the c-index of the ML model, they failed to show remarkable advantages. It is plausible that radiomics features may inherently capture certain clinically relevant information associated with recurrence, while the clinical features added in the above studies vary greatly and fail to fully reflect the biological characteristics of patients. Therefore, the additional clinical features of patients lead to a limited improvement in the predictive efficiency of the model, which, to some extent, reflects that the radiomics-based model may be more stable.

To thoroughly ascertain the predictive efficiency of the radiomics-based model, subgroup analyses were conducted from multiple dimensions such as stage, treatment modality, and pathological classification. Regarding disease stage, the c-index of the model in patients with NSCLC at an early stage was 0.96, which was higher than that in patients with locally advanced NSCLC (c-index=0.759). The tumors in early-stage NSCLC are usually well-defined in terms of boundaries and relatively intact in structure, and the imaging features of the tumors are more easily recognized and quantified by the model, thereby improving the predictive stability and discriminative capacity of the models. In contrast, the tumors in patients with locally advanced NSCLC often invade adjacent areas such as the mediastinum, pericardium, and chest wall, or present lymph node metastases such as at the pulmonary hilum and mediastinum, and some patients may also suffer from complications such as obstructive pneumonia. These factors contribute to more complex imaging features and increased heterogeneity, which may interfere with the extraction and learning of the model for key features and thereby affect the prediction accuracy [49,50]. Additionally, patients with locally advanced NSCLC often undergo treatment interventions, including neoadjuvant chemotherapy and radiotherapy, which may further change the expression patterns of imaging features [51].

The subgroup analyses, particularly those with wide CIs, should be interpreted with caution. This study included 7 studies using SBRT and 18 studies utilizing surgical treatment. Although the c-index of the 2 groups was similar, the CI of the SBRT group was significantly wider, suggesting that the stability of the prediction results in this group was poorer. These findings are

XSL•FO
RenderX

exploratory and may reflect variability due to small sample sizes or heterogeneity among included studies. Therefore, the subgroup results should be considered hypothesis-generating rather than conclusive. Future research with larger datasets and consistent analytical frameworks is warranted to confirm these preliminary trends. Moreover, SBRT serves as a relatively new and highly technology-dependent treatment modality, and there are significant differences in implementation plans, dose settings, and patient selection criteria among different studies, which may also lead to increased uncertainty in the performance evaluation of the model [52]. These results imply that the stability of current radiomics-based models in patients undergoing SBRT or other nonsurgical treatments still requires further verification. In the future, more rigorous prospective investigations with larger sample sizes are needed to assess the applicability and reliability of the radiomics-based model in patients with NSCLC receiving SBRT treatment, thereby facilitating its integration into personalized treatment plans.

In terms of pathological classification, a total of 22 studies in this study only mentioned NSCLC without further subdividing pathological subtypes (squamous cell carcinoma and adenocarcinoma). The findings revealed that the radiomics-based prediction model exhibited superior predictive accuracy in studies classifying adenocarcinoma and squamous cell carcinoma than in those without specific subtype classification. Adenocarcinoma and squamous cell carcinoma are the 2 most prevalent pathological subtypes; furthermore, due to their histological heterogeneity, they may be more identifiable in imaging features, thereby enabling the model to extract more discriminative imaging features and improving the prediction accuracy [53,54]. In contrast, in studies that did not explicitly include information on pathological subtypes, the predictive performance of their models may be limited due to high sample heterogeneity and inconsistent imaging features. In training and validating subsequent radiomics-based models, explicit histological classifications should be included as much as possible to strengthen the stability and the practical applicability of models. Moreover, given the distinct biological characteristics, therapeutic responses, and prognoses across subtypes, combining histopathological information for multimodal fusion modeling may enhance the efficacy of the radiomics-based prediction models. In summary, while the subgroup analysis is supported by the meta-analysis and the sensitivity analysis, the findings should be considered preliminary and hypothesis-generating. In the future, additional prospective studies with a large sample size and based on multicenter data are warranted to validate these findings.

Radiomics has achieved considerable progress in predicting the risk of recurrence in NSCLC. However, due to the inherent complexity of radiomics and the limited reproducibility of the numerous processes involved, it still faces many challenges in clinical applications. Particularly, in image segmentation, feature extraction, and dataset analysis, the application of radiomics is still limited to academic research [55-57]. To begin with, a primary technical challenge in current radiomics research is the absence of standardized imaging protocols [58]. Different medical institutions use scanning equipment and image acquisition parameters that vary greatly. Even for the same patient, the imaging quality and image features obtained at different times or with different devices may vary, thereby affecting the stability of feature extraction and the generalization ability of models [55,59-62]. Standardizing the processing of image data across devices and centers may be the key to improving the clinical applicability of models. Currently, some studies attempt to reduce the influence of image differences through image normalization, resampling, and image domain adaptation [59], but more confirmatory research is needed. In the future, pilot studies with small samples may be conducted prior to formal model construction to evaluate the actual impact of image protocol differences on model construction. Given the subjectivity and consistency issue of the image segmentation process, the stability of models is affected, and manual segmentation is influenced by operator expertise, leading to variability [58,63,64]. Therefore, more studies attempt to apply deep learning methods for automatic or semiautomatic segmentation and to promote the application of automatic segmentation algorithms in multidisease and multimodal data to improve efficiency and reproducibility [65-67].

Moreover, radiomics usually involves high-dimensional small-sample data, so it is frequently associated with overfitting, whereby the radiomics-based model performs well on training data but underperforms on new data. This issue may be alleviated through techniques such as feature selection and data augmentation. In the previous study by Ye et al [68], the robustness of the model was improved through data augmentation and feature selection, and kernelization methods were used, and principal component analysis was performed to lower the feature dimension to effectively reduce the risk of overfitting. In summary, as an emerging tool for imaging analysis, radiomics holds potential for clinical practice. Nonetheless, before radiomics is applied in clinical practice, the above-mentioned problems should be addressed.

It should be noted that the overall methodological quality of the included studies, as reflected by the relatively low mean RQS (27.4%), indicates substantial limitations in current radiomics-based ML research. Many of the included studies were absent from external validation, prospective design, or standardized feature extraction procedures, which may introduce bias and restrict the generalizability of the pooled diagnostic accuracy. Therefore, the results of this meta-analysis should be interpreted with caution. Future studies are encouraged to adhere to standardized radiomics workflows and reporting guidelines, and to perform multicenter validations to enhance methodological rigor and clinical reliability. Furthermore, the overall RQS of the included studies was relatively low, primarily manifested in the following aspects: biological relevance, cost-benefit analysis, comparison with the gold standard, and prospective design. Biological relevance involves the link of imaging features with genetic or molecular features; however, only 1 included study reported the potential correlation of gene levels and imaging features. None of the original studies included were prospectively registered, and some were absent from external validation or solely relied on internal validation based on single-center data. These deficiencies collectively led to the low RQS.

XSL•FO
RenderX

Strengths and Limitations

Although this meta-analysis demonstrates the potential of radiomics-based ML for forecasting the risk of recurrence in NSCLC, certain limitations still need to be resolved to strengthen the reliability and clinical applicability of radiomics-based models for the prediction of the risk of recurrence in NSCLC. First, this study primarily utilizes retrospective data, potentially introducing selection bias and other issues, thereby limiting the generalizability of the research results. Second, although subgroup analyses are conducted, it is still difficult to completely eliminate heterogeneity. Finally, some studies are limited to the sample sizes, lack external validation, and lack long-term follow-up analysis of actual clinical prognosis, which restricts the assessment of the practical applicability of radiomics-based models.

Conclusion

This systematic review, for the first time, confirms that radiomics-based ML models effectively predict the risk of recurrence in NSCLC, and clinically relevant subgroups (eg, model type, treatment background, stage, and pathology) were explored. Differing from individual original studies or extensive reviews focusing on radiomics with mixed heterogeneous end points, this study concentrates on recurrence, a clinically operable outcome indicator, thereby clarifying areas where model performance is most pronounced and methodological gaps are identified. In clinical practice, robust radiomics-based ML tools may be leveraged to support individualized monitoring intensity, aid in formulating treatment regimens, and optimize risk stratification in clinical trials. Nonetheless, the current application of radiomics methodologies remains insufficient. Therefore, in the future, studies should focus on standardizing the implementation procedure of radiomics-based ML and integrating multicenter imaging data to enhance generalizability.

Funding

This work was supported by the Key Construction Disciplines of Provincial and Municipal Co construction of Zhejiang (2023-SSGJ-002); the National Oncology Clinical Key Speciality (2023-GJZK-001); the Jiaxing Key Laboratory of Precision Treatment for Lung Cancer, the Early Diagnosis and Comprehensive Treatment of Lung Cancer Innovation Team Building Project, and the Scientific Technology Plan Program for Healthcare in Zhejiang Province (2023KY329, 2023KY1196, and 24 2023RC098); the Science and Technology Project of Jiaxing (2025CGZ041); and Zhejiang Province Postdoctoral Research Project (ZJ2024156).

Data Availability

The datasets used and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: J Wu
Formal analysis and investigation: WC, SH, J Wang, CC
Funding acquisition: YZ
Investigation: WC, SH, J Wu, CC
Methodology: J Wu
Resources: J Wang
Supervision: WC
Writing – original draft: J Wu
Writing – review and editing: SH
All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.
SH (hushiyu0317@163.com) and WC (00135116@zjxu.edu.cn) are co-corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental tables and figures: the detailed literature search strategies (Table S1), characteristics of the included studies (Table S2), Radiomics Quality Score assessments (Table S3), and forest plots depicting subgroup analyses for the meta-analysis of machine learning models predicting recurrence in non–small cell lung cancer (Figures S1-S14).
[7Z File, 556 KB - jmir_v28i1e84223_app1.7z]

Checklist 1
PRISMA checklist.
[DOCX File, 272 KB - jmir_v28i1e84223_app2.docx]

XSL•FO
RenderX

Checklist 2
PRISMA-S checklist.
[DOCX File, 22 KB - jmir_v28i1e84223_app3.docx]

References

1. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin 2023 Jan;73(1):17-48. [doi: 10.3322/caac.21763] [Medline: 36633525]
2. Leiter A, Veluswamy RR, Wisnivesky JP. The global burden of lung cancer: current status and future trends. Nat Rev Clin Oncol 2023 Sep;20(9):624-639. [doi: 10.1038/s41571-023-00798-3] [Medline: 37479810]
3. Liu L, Soler J, Reckamp KL, Sankar K. Emerging targets in non–small cell lung cancer. Int J Mol Sci 2024 Sep 18;25(18):10046. [doi: 10.3390/ijms251810046] [Medline: 39337530]
4. Mithoowani H, Febbraro M. Non-small-cell lung cancer in 2022: a review for general practitioners in oncology. Curr Oncol 2022 Mar 9;29(3):1828-1839. [doi: 10.3390/curroncol29030150] [Medline: 35323350]
5. Araghi M, Mannani R, Heidarnejad Maleki A, et al. Recent advances in non–small cell lung cancer targeted therapy; an update review. Cancer Cell Int 2023 Aug 11;23(1):162. [doi: 10.1186/s12935-023-02990-y] [Medline: 37568193]
6. Lin X, Xie M, Yao J, et al. Immune-related adverse events in non-small cell lung cancer: occurrence, mechanisms and therapeutic strategies. Clin Transl Med 2024 Mar;14(3):e1613. [doi: 10.1002/ctm2.1613] [Medline: 38451000]
7. Jairam V, Lee V, Park HS, et al. Treatment-related complications of systemic therapy and radiotherapy. JAMA Oncol 2019 Jul 1;5(7):1028-1035. [doi: 10.1001/jamaoncol.2019.0086] [Medline: 30946433]
8. Monsalve AF, Hoag JR, Resio BJ, et al. Variable impact of prior cancer history on the survival of lung cancer patients. Lung Cancer 2019 Jan;127:130-137. [doi: 10.1016/j.lungcan.2018.11.040] [Medline: 30642541]
9. Riely GJ, Wood DE, Ettinger DS, et al. Non–small cell lung cancer, version 4.2024, NCCN clinical practice guidelines in oncology. J Natl Compr Canc Netw 2024 May;22(4):249-274. [doi: 10.6004/jnccn.2204.0023] [Medline: 38754467]
10. Mohamed SK, Walsh B, Timilsina M, et al. On predicting recurrence in early stage non–small cell lung cancer. AMIA Annu Symp Proc 2021;2021:853-862. [Medline: 35308971]
11. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012 Mar;48(4):441-446. [doi: 10.1016/j.ejca.2011.11.036] [Medline: 22257792]
12. Chen J, Chen A, Yang S, Liu J, Xie C, Jiang H. Accuracy of machine learning in preoperative identification of genetic mutation status in lung cancer: a systematic review and meta-analysis. Radiother Oncol 2024 Jul;196:110325. [doi: 10.1016/j.radonc.2024.110325] [Medline: 38734145]
13. Yan HJ, Zhao JS, Zuo HD, et al. Dual-region computed tomography radiomics-based machine learning predicts subcarinal lymph node metastasis in patients with non–small cell lung cancer. Ann Surg Oncol 2024 Aug;31(8):5011-5020. [doi: 10.1245/s10434-024-15197-w] [Medline: 38520581]
14. Tutsoy O, Sumbul HE. A novel deep machine learning algorithm with dimensionality and size reduction approaches for feature elimination: thyroid cancer diagnoses with randomly missing data. Brief Bioinform 2024 May 23;25(4):bbae344. [doi: 10.1093/bib/bbae344] [Medline: 39007597]
15. Xia F, Wei W, Wang J, Duan Y, Wang K, Zhang C. Machine learning model for non-alcoholic steatohepatitis diagnosis based on ultrasound radiomics. BMC Med Imaging 2024 Aug 20;24(1):221. [doi: 10.1186/s12880-024-01398-y] [Medline: 39164667]
16. Libling WA, Korn R, Weiss GJ. Review of the use of radiomics to assess the risk of recurrence in early-stage non–small cell lung cancer. Transl Lung Cancer Res 2023 Jul 31;12(7):1575-1589. [doi: 10.21037/tlcr-23-5] [Medline: 37577298]
17. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
18. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
19. Debray TP, Damen JA, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Stat Methods Med Res 2019 Sep;28(9):2768-2786. [doi: 10.1177/0962280218785504] [Medline: 30032705]
20. Aonpong P, Iwamoto Y, Han XH, Lin L, Chen YW. Improved genotype-guided deep radiomics signatures for recurrence prediction of non–small cell lung cancer. Annu Int Conf IEEE Eng Med Biol Soc 2021 Nov;2021:3561-3564. [doi: 10.1109/EMBC46164.2021.9630703] [Medline: 34892008]
21. Wang P, Luo Z, Luo C, Wang T. Application of a comprehensive model based on CT radiomics and clinical features for postoperative recurrence risk prediction in non-small cell lung cancer. Acad Radiol 2024 Jun;31(6):2579-2590. [doi: 10.1016/j.acra.2023.11.028]
22. Wang Y, Ding Y, Liu X, et al. Preoperative CT-based radiomics combined with tumour spread through air spaces can accurately predict early recurrence of stage I lung adenocarcinoma: a multicentre retrospective cohort study. Cancer Imaging 2023 Sep 7;23(1):83. [doi: 10.1186/s40644-023-00605-3] [Medline: 37679806]

XSL•FO
RenderX

23. Dissaux G, Visvikis D, Da-Ano R, et al. Pretreatment 18F-FDG PET/CT radiomics predict local recurrence in patients treated with stereotactic body radiotherapy for early-stage non-small cell lung cancer: a multicentric study. J Nucl Med 2020 Jun;61(6):814-820. [doi: 10.2967/jnumed.119.228106] [Medline: 31732678]
24. Lan H, Wei C, Xu F, et al. 2.5D peritumoural radiomics predicts postoperative recurrence in stage I lung adenocarcinoma. Front Oncol 2024;14:1382815. [doi: 10.3389/fonc.2024.1382815] [Medline: 39267836]
25. Lucia F, Louis T, Cousin F, et al. Multicentric development and evaluation of [18F]FDG PET/CT and CT radiomic models to predict regional and/or distant recurrence in early-stage non-small cell lung cancer treated by stereotactic body radiation therapy. Eur J Nucl Med Mol Imaging 2024 Mar;51(4):1097-1108. [doi: 10.1007/s00259-023-06510-y] [Medline: 37987783]
26. Ni J, Chen H, Yu L, et al. Predicting regional recurrence and prognosis in stereotactic body radiation therapy-treated clinical stage I non-small cell lung cancer using a radiomics model constructed with surgical data. Int J Radiat Oncol Biol Phys 2024 Nov 15;120(4):1096-1106. [doi: 10.1016/j.ijrobp.2024.06.010] [Medline: 38936632]
27. Ma S, Li J, Chen Y, et al. Machine learning based on clinical information and integrated CT radiomics to predict local recurrence of stage IA lung adenocarcinoma after microwave ablation. J Vasc Interv Radiol 2024 Dec;35(12):1823-1832. [doi: 10.1016/j.jvir.2024.08.018] [Medline: 39208929]
28. Feng Y, Xu Y, Wang J, et al. Prediction of early recurrence after bronchial arterial chemoembolization in non-small cell lung cancer patients using dual-energy CT: an interpretable model based on SHAP methodology. Acad Radiol 2025 Oct;32(10):6320-6329. [doi: 10.1016/j.acra.2025.07.039] [Medline: 40784844]
29. Ogbonna CP, Breen WG, Le Noach P, et al. Radiomics-based prediction of local recurrence after stereotactic body radiation therapy for early-stage non-small cell lung cancer. Ann Am Thorac Soc 2025 Aug;22(8):1236-1243. [doi: 10.1513/AnnalsATS.202410-1047OC] [Medline: 40208641]
30. Ho WLJ, Fetisov N, Hall LO, Goldgof D, Schabath MB. Utilizing clinicopathological and radiomic features for risk stratification of lung cancer recurrence. Acad Radiol 2025 Sep;32(9):5557-5566. [doi: 10.1016/j.acra.2025.04.062] [Medline: 40379589]
31. Wang T, Deng J, She Y, et al. Radiomics signature predicts the recurrence-free survival in stage I non-small cell lung cancer. Ann Thorac Surg 2020 Jun;109(6):1741-1749. [doi: 10.1016/j.athoracsur.2020.01.010] [Medline: 32087134]
32. Akinci D'Antonoli T, Farchione A, Lenkowicz J, et al. CT radiomics signature of tumor and peritumoral lung parenchyma to predict nonsmall cell lung cancer postsurgical recurrence risk. Acad Radiol 2020 Apr;27(4):497-507. [doi: 10.1016/j.acra.2019.05.019] [Medline: 31285150]
33. Nemoto H, Saito M, Satoh Y, et al. Evaluation of the performance of both machine learning models using PET and CT radiomics for predicting recurrence following lung stereotactic body radiation therapy: a single-institutional study. J Appl Clin Med Phys 2024 Jul;25(7):e14322. [doi: 10.1002/acm2.14322] [Medline: 38436611]
34. Lee S, Jung J, Hong H, Kim BS. Prediction of two-year recurrence-free survival in operable NSCLC patients using radiomic features from intra- and size-variant peri-tumoral regions on chest CT images. Diagnostics (Basel) 2022 May 25;12(6):1313. [doi: 10.3390/diagnostics12061313] [Medline: 35741123]
35. Huynh E, Coroller TP, Narayan V, et al. Associations of radiomic data extracted from static and respiratory-gated CT scans with disease recurrence in lung cancer patients treated with SBRT. PLoS ONE 2017;12(1):e0169172. [doi: 10.1371/journal.pone.0169172] [Medline: 28046060]
36. Hindocha S, Charlton TG, Linton-Reid K, et al. Gross tumour volume radiomics for prognostication of recurrence & death following radical radiotherapy for NSCLC. NPJ Precis Oncol 2022 Oct 27;6(1):77. [doi: 10.1038/s41698-022-00322-3] [Medline: 36302938]
37. Park D, Oh D, Lee M, et al. Importance of CT image normalization in radiomics analysis: prediction of 3-year recurrence-free survival in non-small cell lung cancer. Eur Radiol 2022 Dec;32(12):8716-8725. [doi: 10.1007/s00330-022-08869-2] [Medline: 35639142]
38. Li Y, Zhao J, Li R, et al. Predicting prognosis in patients with stage IA lung adenocarcinoma with a micropapillary component using a nomogram based on computed tomography radiomics and clinicopathologic factors: a retrospective analysis. Transl Lung Cancer Res 2024 Oct 31;13(10):2585-2602. [doi: 10.21037/tlcr-24-544] [Medline: 39507031]
39. Zhang L, Lv L, Li L, et al. Radiomics signature to predict prognosis in early-stage lung adenocarcinoma (≤3 cm) patients with no lymph node metastasis. Diagnostics (Basel) 2022 Aug 6;12(8):1907. [doi: 10.3390/diagnostics12081907] [Medline: 36010257]
40. Shimada Y, Kudo Y, Maehara S, et al. Radiomics with artificial intelligence for the prediction of early recurrence in patients with clinical stage IA lung cancer. Ann Surg Oncol 2022 Dec;29(13):8185-8193. [doi: 10.1245/s10434-022-12516-x] [Medline: 36070112]
41. Christie JR, Daher O, Abdelrazek M, et al. Predicting recurrence risks in lung cancer patients using multimodal radiomics and random survival forests. J Med Imaging (Bellingham) 2022 Nov;9(6):066001. [doi: 10.1117/1.JMI.9.6.066001] [Medline: 36388142]
42. Lafata KJ, Hong JC, Geng R, et al. Association of pre-treatment radiomic features with lung cancer recurrence following stereotactic body radiation therapy. Phys Med Biol 2019 Jan 8;64(2):025007. [doi: 10.1088/1361-6560/aaf5a5] [Medline: 30524018]

XSL•FO
RenderX

43. Park SB, Kim KU, Park YW, Hwang JH, Lim CH. Application of 18 F-fluorodeoxyglucose PET/CT radiomic features and machine learning to predict early recurrence of non-small cell lung cancer after curative-intent therapy. Nucl Med Commun 2023 Feb 1;44(2):161-168. [doi: 10.1097/MNM.0000000000001646] [Medline: 36458424]
44. Liu W, Qiao X, Ge H, et al. Recurrence patterns are significantly associated with the 18F-FDG PET/CT radiomic features of patients with locally advanced non-small cell lung cancer treated with chemoradiotherapy. Oncol Lett 2023 Jul;26(1):317. [doi: 10.3892/ol.2023.13903] [Medline: 37332327]
45. Kim G, Moon S, Choi JH. Deep learning with multimodal integration for predicting recurrence in patients with non-small cell lung cancer. Sensors (Basel) 2022 Aug 31;22(17):6594. [doi: 10.3390/s22176594] [Medline: 36081053]
46. Franceschini D, Cozzi L, De Rose F, et al. A radiomic approach to predicting nodal relapse and disease-specific survival in patients treated with stereotactic body radiation therapy for early-stage non-small cell lung cancer. Strahlenther Onkol 2020 Oct;196(10):922-931. [doi: 10.1007/s00066-019-01542-6] [Medline: 31722061]
47. Huang P, Illei PB, Franklin W, et al. Lung cancer recurrence risk prediction through integrated deep learning evaluation. Cancers (Basel) 2022 Aug 27;14(17):4150. [doi: 10.3390/cancers14174150] [Medline: 36077686]
48. Ho WLJ, Fetisov N, Hall LO, Goldgof D, Schabath MB. Evaluating clinical and radiomic features for predicting lung cancer recurrence pre- and post-tumor resection. Proc SPIE Int Soc Opt Eng 2024 Feb;12926:1292623. [doi: 10.1117/12.3006091] [Medline: 38993353]
49. Lou F, Sima CS, Rusch VW, Jones DR, Huang J. Differences in patterns of recurrence in early-stage versus locally advanced non–small cell lung cancer. Ann Thorac Surg 2014 Nov;98(5):1755-1760. [doi: 10.1016/j.athoracsur.2014.05.070] [Medline: 25110337]
50. Zhang N, Liang R, Gensheimer MF, et al. Early response evaluation using primary tumor and nodal imaging features to predict progression-free survival of locally advanced non–small cell lung cancer. Theranostics 2020;10(25):11707-11718. [doi: 10.7150/thno.50565] [Medline: 33052242]
51. Huber RM, Kauffmann-Guerrero D, Hoffmann H, Flentje M. New developments in locally advanced nonsmall cell lung cancer. Eur Respir Rev 2021 Jun 30;30(160):200227. [doi: 10.1183/16000617.0227-2020] [Medline: 33952600]
52. Schneider BJ, Daly ME, Kennedy EB, et al. Stereotactic body radiotherapy for early-stage non–small-cell lung cancer: American Society of Clinical Oncology endorsement of the American Society for Radiation Oncology evidence-based guideline. J Clin Oncol 2018 Mar 1;36(7):710-719. [doi: 10.1200/JCO.2017.74.9671] [Medline: 29106810]
53. Liu Y, Chang Y, Zha X, et al. A combination of radiomic features, imaging characteristics, and serum tumor biomarkers to predict the possibility of the high-grade subtypes of lung adenocarcinoma. Acad Radiol 2022 Dec;29(12):1792-1801. [doi: 10.1016/j.acra.2022.02.024] [Medline: 35351366]
54. Yang M, Shi L, Huang T, et al. Value of contrast-enhanced magnetic resonance imaging-T2WI-based radiomic features in distinguishing lung adenocarcinoma from lung squamous cell carcinoma with solid components >8 mm. J Thorac Dis 2023 Feb 28;15(2):635-648. [doi: 10.21037/jtd-23-142] [Medline: 36910079]
55. Recht MP, Dewey M, Dreyer K, et al. Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations. Eur Radiol 2020 Jun;30(6):3576-3584. [doi: 10.1007/s00330-020-06672-5] [Medline: 32064565]
56. Ponsiglione A, Stanzione A, Spadarella G, et al. Ovarian imaging radiomics quality score assessment: an EuSoMII radiomics auditing group initiative. Eur Radiol 2023 Mar;33(3):2239-2247. [doi: 10.1007/s00330-022-09180-w] [Medline: 36303093]
57. Park JE, Kim D, Kim HS, et al. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to Radiomics Quality Score and TRIPOD statement. Eur Radiol 2020 Jan;30(1):523-536. [doi: 10.1007/s00330-019-06360-z] [Medline: 31350588]
58. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. Int J Radiat Oncol Biol Phys 2018 Nov 15;102(4):1143-1158. [doi: 10.1016/j.ijrobp.2018.05.053] [Medline: 30170872]
59. Zwanenburg A, Vallières M, Abdalah MA, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 2020 May;295(2):328-338. [doi: 10.1148/radiol.2020191145] [Medline: 32154773]
60. Berenguer R, Pastor-Juan MDR, Canales-Vázquez J, et al. Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters. Radiology 2018 Aug;288(2):407-415. [doi: 10.1148/radiol.2018172361] [Medline: 29688159]
61. Welch ML, McIntosh C, Haibe-Kains B, et al. Vulnerabilities of radiomic signature development: the need for safeguards. Radiother Oncol 2019 Jan;130:2-9. [doi: 10.1016/j.radonc.2018.10.027] [Medline: 30416044]
62. Meyer M, Ronald J, Vernuccio F, et al. Reproducibility of CT radiomic features within the same patient: influence of radiation dose and CT reconstruction settings. Radiology 2019 Dec;293(3):583-591. [doi: 10.1148/radiol.2019190928] [Medline: 31573400]
63. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology 2016 Feb;278(2):563-577. [doi: 10.1148/radiol.2015151169] [Medline: 26579733]
64. van Velden FHP, Kramer GM, Frings V, et al. Repeatability of radiomic features in non–small-cell lung cancer [(18)F]FDG-PET/CT studies: impact of reconstruction and delineation. Mol Imaging Biol 2016 Oct;18(5):788-795. [doi: 10.1007/s11307-016-0940-2] [Medline: 26920355]

XSL•FO
RenderX

65. Parmar C, Rios Velazquez E, Leijenaar R, et al. Robust radiomics feature quantification using semiautomatic volumetric segmentation. PLoS ONE 2014;9(7):e102107. [doi: 10.1371/journal.pone.0102107] [Medline: 25025374]
66. Orlhac F, Soussan M, Maisonobe JA, Garcia CA, Vanderlinden B, Buvat I. Tumor texture analysis in 18F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. J Nucl Med 2014 Mar;55(3):414-422. [doi: 10.2967/jnumed.113.129858] [Medline: 24549286]
67. Nestle U, Kremp S, Schaefer-Schuler A, et al. Comparison of different methods for delineation of 18F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non–small cell lung cancer. J Nucl Med 2005 Aug;46(8):1342-1348. [Medline: 16085592]
68. Ye J, Luo J, Xu S, Wu W. One-slice CT image based kernelized radiomics model for the prediction of low/mid-grade and high-grade HNSCC. Comput Med Imaging Graph 2020 Mar;80:101675. [doi: 10.1016/j.compmedimag.2019.101675] [Medline: 31945637]

Abbreviations

CT: computed tomography
ML: machine learning
NSCLC: non–small cell lung cancer
PET-CT: positron emission tomography–computed tomography
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search Extension
REM: random effects model
RQS: Radiomics Quality Score
SBRT: stereotactic body radiation therapy
TNM: tumor node metastasis

Edited by S Brini; submitted 16.Sep.2025; peer-reviewed by M Liao, O Tutsoy; revised version received 13.Jan.2026; accepted 15.Jan.2026; published 02.Apr.2026.

Please cite as:
Wu J, Zhang Y, Wang J, Chen C, Hu S, Chen W
Accuracy of Radiomics-Based Machine Learning for Predicting Risk of Recurrence in Non–Small Cell Lung Cancer: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e84223
URL: https://www.jmir.org/2026/1/e84223
doi:10.2196/84223

© Junpei Wu, Ye Zhang, Jiaye Wang, Chengshui Chen, Shiyu Hu, Wenyu Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Multimodal AI for Alzheimer Disease Diagnosis: Systematic Review of Datasets, Models, and Modalities

Ziwen Yu[1*], MSc; Anthony Mulholland[1*], PhD; Tianyan Huang[2]; Qiang Liu[1], PhD

[1]School of Engineering Mathematics and Technology, University of Bristol, Tankard's Close, Ada Lovelace Building, Bristol, United Kingdom

[2]Medical Physics and Biomedical Engineering, University College London, London, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Qiang Liu, PhD
School of Engineering Mathematics and Technology, University of Bristol, Tankard's Close, Ada Lovelace Building, Bristol, United Kingdom

Abstract

Background: Early detection of Alzheimer disease (AD) is essential for timely intervention; yet, diagnostic performance varies widely across modalities and datasets. Recent multimodal artificial intelligence (AI) models have made significant progress, but the evidence base remains fragmented due to heterogeneous datasets, modeling frameworks, and reporting quality.

Objective: This systematic review aimed to analyze studies on multimodal AI models for AD diagnosis, prognosis, and risk prediction over 5 years. We evaluated dataset characteristics, modality combinations, modeling strategies, performance metrics, and methodological limitations. We further discuss real-world implications and translational pathways.

Methods: Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines, we systematically searched PubMed, IEEE Xplore, Scopus, ACM Digital Library, Cochrane, and arXiv, with the final datasets last searched on November 15, 2025. Studies applying multimodal machine learning or deep learning to AD, mild cognitive impairment, and dementia outcomes were included, whereas studies using a single modality or lacking sufficient methodological detail were excluded. QUADAS-2 (Revised Quality Assessment of Diagnostic Accuracy Studies tool) assessed risk of bias. Extracted performance results were synthesized across 4 major multimodal dataset families.

Results: A total of 66 studies met the inclusion criteria. Across datasets, multimodal models consistently outperformed single-modal baselines. Alzheimer's Disease Neuroimaging Initiative–based diagnosis achieved an average accuracy of 92.5% (SD 3.8%), while mild cognitive impairment–conversion models achieved an average area under the curve (AUC) of 0.922 (SD 0.045), and several fusion architectures reported AUCs above 0.95. In contrast, UK Biobank risk-prediction studies reported an average AUC of 0.84 (SD 0.056), and this reflects performance in large, population-based datasets. DementiaBank speech-language studies achieved an average AUC of 0.813 (SD 0.042), and cross-lingual AD detection achieved an accuracy of 77% (SD 6.5%). Self-collected multimodal datasets demonstrated average accuracies around 96% (SD 2.4%), but their generalizability is limited due to small sample sizes and single-center designs.

Conclusions: This systematic review demonstrates that multimodal AI models consistently outperform single-modal models for AD diagnosis, prognosis, and risk prediction by integrating complementary biological, clinical, and behavioral information. Unlike prior reviews, this review provides a unified synthesis across heterogeneous clinical, imaging, genetic, and linguistic datasets, enabling cross-domain comparison of modeling strategies and performance. However, the generalizability of reported performance was limited due to substantial heterogeneity in dataset composition, outcome definitions, and validation, and prevalent risks of bias. By evaluating these factors, this review clarifies where current evidence is robust and where caution is warranted. The findings highlight the need for standardized multimodal benchmarks, transparent evaluation protocols, and clinically grounded model design to enable reliable real-world deployment. Overall, this work advances the field by framing multimodal AI not only as a performance-driven tool but also as a translational framework for equitable, interpretable, and scalable AD diagnosis.

Trial Registration: PROSPERO CRD420251241895; https://www.crd.york.ac.uk/PROSPE-RO/view/CRD420251241895

(J Med Internet Res 2026;28:e85414) doi:10.2196/85414

KEYWORDS

alzheimer's disease, neurodegenerative disease; multimodal dataset; machine learning; deep learning; multimodal fusion; computer-aided diagnosis; early diagnosis

XSL•FO
RenderX

Introduction

Alzheimer disease (AD) is the most prevalent neurodegenerative disorder and the leading cause of dementia worldwide [1]. With an aging global population, AD has become one of the most costly and deadly diseases of the 21st century, imposing profound emotional, financial, and caregiving burdens on patients, families, and health systems. By 2050, the number of people with AD is projected to rise from 55 million in 2020 to approximately 139 million [1].

The progression of AD includes the preclinical stage, mild cognitive impairment (MCI), and symptomatic stages, with varying degrees of symptom severity. The preclinical stage is a key window for intervention, during which neuropathological changes have commenced, but clinical symptoms remain largely undetectable [2]. Despite advances in awareness and screening, up to 75% of dementia cases remain undiagnosed worldwide, particularly in low- and middle-income countries [3]. This persistent diagnostic gap highlights the need for low-cost, scalable, and accurate early detection tools to enable timely intervention and slow disease progression [4].

Artificial intelligence (AI) has emerged as a promising approach for improving the early detection and management of AD. By systematically integrating and analyzing multimodal data, AI-based diagnostic frameworks offer powerful tools to enhance early detection accuracy and facilitate timely intervention.

Recent work has used transformer-based models to integrate imaging, genetic, and linguistic data. Multimodal transformers combining magnetic resonance imaging (MRI) or positron emission tomography (PET) with clinical features and cognitive assessments have reported improved diagnostic accuracy and interpretability [5-7]. In parallel, GPT-style architectures, BERT (Bidirectional Encoder Representations From Transformers) variants, and domain-adapted language models improve extraction of linguistic and semantic markers linked to early cognitive decline [8,9]. Self-supervised speech models also perform strongly for detecting MCI and early AD from spontaneous speech [10]. Together, these advances reflect a shift toward unified, more interpretable, and clinically translatable multimodal systems that capture both biological and behavioral aspects of AD.

Traditional machine learning (ML) [11], ensemble methods [12], deep learning [13], and reinforcement learning (RL) [14] can perform well on unimodal data, but clinical diagnosis integrates structural and behavioral information [15]. Unimodal AI can therefore diverge from clinical workflows and miss complementary signals (eg, MRI for structural change plus speech features for cognitive decline [16]), increasing the risk of modality-specific overfitting and poorer real-world performance. Accordingly, recent work has shifted toward multimodal integration for AD diagnosis, yet many studies emphasize incremental accuracy gains while underaddressing generalizability, interpretability, and cost-effectiveness needed for adoption. The literature also remains fragmented: recent reviews often cover multimodal clinical phenotyping datasets [17] and multimodal linguistic cognitive-impairment datasets [18] separately, obscuring cross-modal insights such as how imaging and speech biomarkers might jointly improve early detection.

Recent multimodal methods have substantially improved AD detection. However, a comprehensive systematic review that integrates evidence across both clinical and linguistic modalities, fusion strategies, and critically evaluates methodological quality, dataset diversity, and reporting transparency is still lacking. To address these gaps, this review investigates how multimodal models are applied to AD diagnosis, prognosis, and risk prediction and compares performance across different modality combinations and dataset families published between 2019 and 2025. We also examine modeling and fusion strategies alongside validation practices and assess methodological quality and risk of bias using QUADAS-2 (Revised Quality Assessment of Diagnostic Accuracy Studies tool). Furthermore, key multimodal combinations within public datasets are analyzed in relation to their diagnostic performance, and datasets are categorized to evaluate their suitability for AD research and clinical translation. Overall, this review provides a comprehensive synthesis of multimodal AI in AD diagnosis, bridges previously disconnected research streams, and offers practical guidance for future model development and clinical adoption.

Methods

Study Design

This review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines [19], with the search procedures reported following PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) [20] and developed using the principles outlined in the Cochrane Handbook [21]. These methods were applied to systematically identify and evaluate studies on computer-aided AD diagnosis, with a particular focus on those using multimodal clinical phenotyping datasets and multimodal linguistic-based cognitive impairment datasets.

Source of the Study and Search Criteria

We developed and internally reviewed independent search strategies (no external peer review). We manually searched multiple databases to identify AI-driven multimodal approaches for AD diagnosis, rather than using an integrated multidatabase platform. As this review targets methodological advances mainly reported in peer-reviewed computational literature, we did not search trial registries (ClinicalTrials.gov, World Health Organization's International Clinical Trials Registry Platform). We also avoided validated or published filters, instead iteratively refining customized controlled-vocabulary and free-text terms for AD or dementia, multimodal data, and AI through pilot screening to maximize sensitivity.

Searches were performed in PubMed (447 records; January 1, 2019, to November 13, 2025), Scopus (1086 records; all years through November 13, 2025, filtered to PUBYEAR > 2018), IEEE Xplore (2229 records; January 1, 2020 to November 13, 2025), ACM Digital Library (2067 records; 1 January 2020, to 15 November 2025), Cochrane Library (1061 records; all available years through November 15, 2025), and arXiv (1081

XSL•FO
RenderX

records; all available years through November 15, 2025). We included the verbatim search strings for all databases, and because arXiv does not support bulk export, an arXiv search Python (Python Software Foundation) script is provided in Multimedia Appendix 1 [11,14,22-44].

Eligibility Criteria

The inclusion criteria were if studies were considered eligible if they met all the following conditions: (1) focused on AD, MCI, or related dementias as the primary clinical outcome; (2) applied AI or ML methods for computer-aided diagnosis, classification, or prediction; (3) used multimodal data, defined as any combination of at least two distinct modalities (eg, neuroimaging, clinical phenotyping, genetics, or linguistic features); (4) reported quantitative evaluation metrics; and (5) written in English.

The exclusion criteria were if studies met any of the following conditions: (1) single-modal approaches using only a single imaging modality, cognitive test, or biomarker, without any multimodal integration; (2) works without reported performance metrics or with insufficient methodological detail; (3) works not addressing diagnosis, classification, or prediction (eg, treatment response, drug trials, and lifestyle interventions); (4) duplicate publications or overlapping datasets without providing additional methodological contribution; and (5) non-English publications.

Selection Process

The study selection process followed the PRISMA 2020 guidelines, and the protocol was registered. The final search update was conducted in November 2025. All records retrieved from the databases were first imported into Zotero, where duplicates were automatically detected and removed.

The initial search identified 7435 records. After removing 3047 duplicates, 4388 records remained for title and abstract screening. A total of 4021 records were obviously irrelevant at the title and abstract level, 252 studies were excluded for the following main reasons:

- Focused on outcomes unrelated to AD diagnosis, classification, or prediction (eg, drug trials, treatment response, lifestyle interventions; n=140).
- Used unimodal data without multimodal integration (n=46).
- No sufficient methodological details (n=47).

Finally, 66 studies were included in the systematic synthesis, and all were successfully retrieved (reports not retrieved=0).

Overview of AI-Assisted AD Diagnosis

The workflow of AI-assisted AD diagnosis involves 3 stages, as illustrated in Figure 1. The initial stage involves comprehensive data acquisition, where information is collected from multiple modalities, including neuroimaging, biomarkers, genetics, and speech or behavioral signals. The second stage involves feature extraction and model development, followed by an interpretable and explainable analysis to ensure that AI models can effectively support clinical decision-making.

XSL•FO
RenderX

Figure 1. Overview of the AI pipeline for AD diagnosis. Multimodal inputs undergo preprocessing and feature extraction before model training for classification or regression tasks. Model interpretability supports explanation and performance evaluation. AD: Alzheimer disease; AI: artificial intelligence; LIME: Local Interpretable Model-Agnostic Explanations; SHAP: Shapley Additive Explanations; XAI: explainable artificial intelligence.

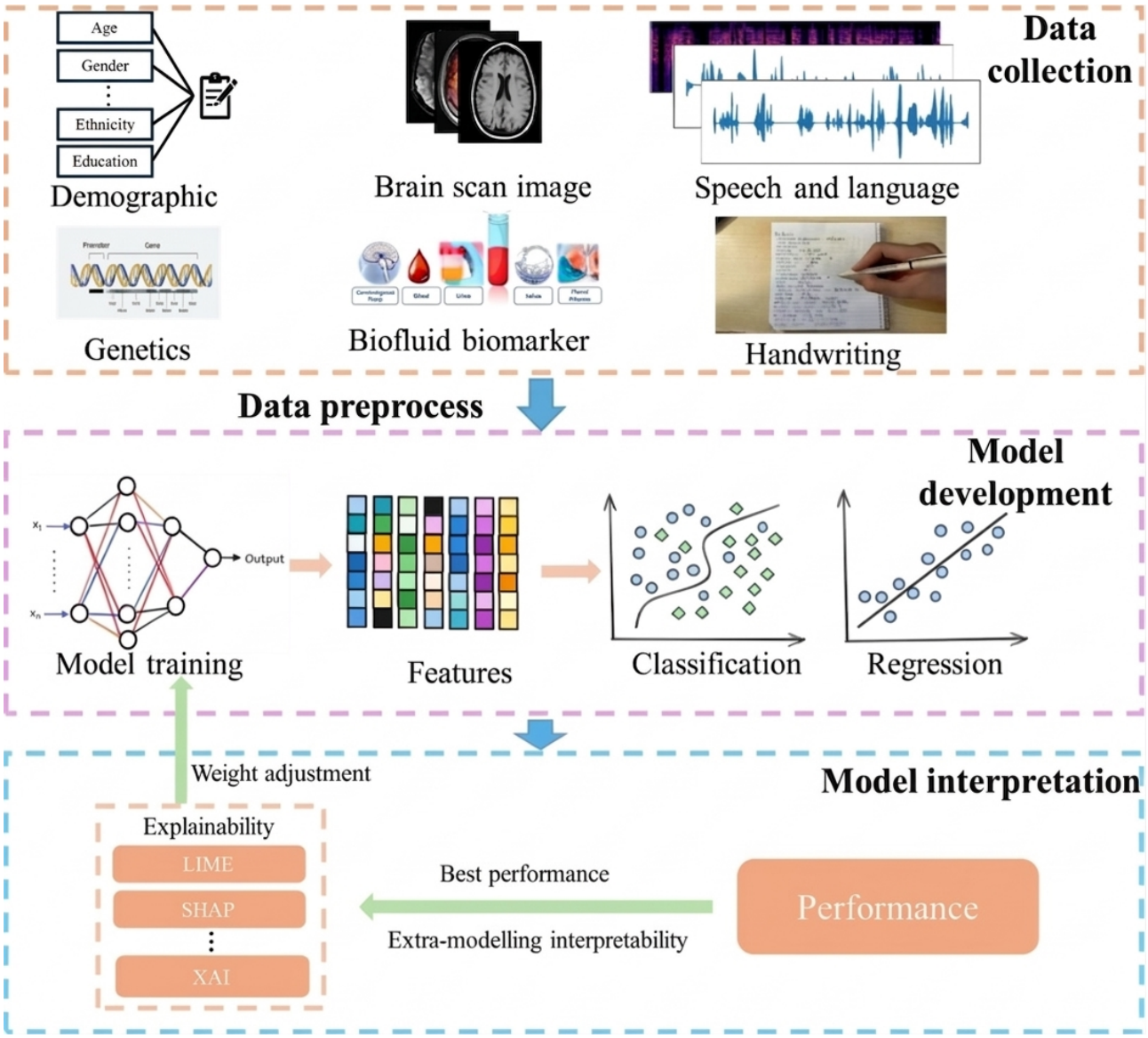

Performance Evaluation Metrics

To ensure the clinical applicability and scientific rigor of computer-aided diagnosis models for AD, it is essential to systematically evaluate their performance using a variety of quantitative metrics. We have summarized all performance evaluation metrics in Multimedia Appendix 2.

Risk of Bias and Quality Assessment

We assessed methodological quality with QUADAS-2 [45], evaluating risk of bias in 4 domains. Patient selection raised the main concern: 61% (40/66) of outcomes were high risk due to poor reporting or nonrepresentative sampling. For the index test, 76% (50/66) were unclear risk because procedures and decision thresholds were insufficiently described, and 20% (13/66) were high risk. The reference standard showed 76% (50/66) unclear risk from limited methodological detail, with no high-risk ratings. Flow and timing had the greatest uncertainty: 85% (56/66) were unclear owing to missing information on testing intervals and participant flow. Figure 2 summarizes domain-level risk distributions; Multimedia Appendix 3 reports study-level assessments.

XSL•FO
RenderX

Figure 2. Summary of the QUADAS-2 plot across the 66 included studies in the domain. QUADAS-2: Revised Quality Assessment of Diagnostic Accuracy Studies tool.

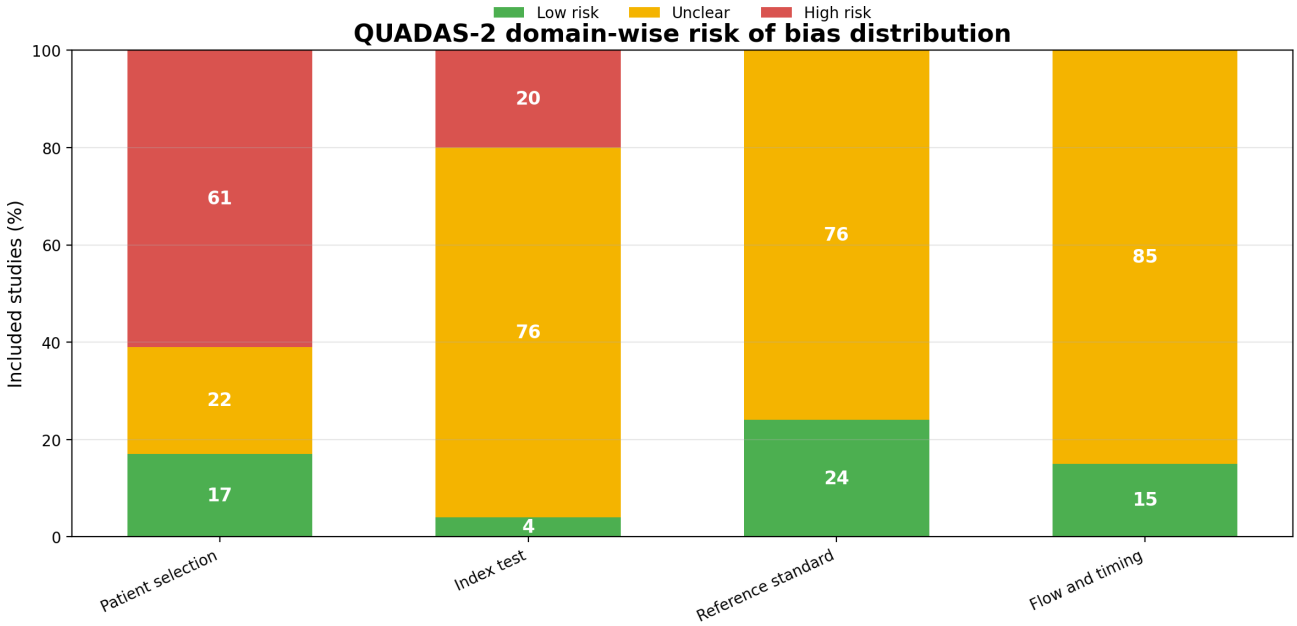

Given frequent unclear and high risk in key domains, we interpreted diagnostic performance cautiously, especially without external validation or a clearly defined reference standard. Future benchmarking should emphasize transparent reporting, prespecified thresholds, and multicenter evaluation to reduce bias and improve reproducibility.

Results

Overview

Following study selection (the complete selection process is summarized in the PRISMA 2020 flow diagram, Figure 3), we first summarize the overall profile of the included literature to contextualize the subsequent synthesis. Figure 4 provides a temporal overview (2019 - 2025) of modeling approaches across included studies, illustrating how methodological focus has shifted over time and informing interpretation of the evidence base.

XSL•FO
RenderX

Figure 3. Flow diagram of PRISMA. AD: Alzheimer disease; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

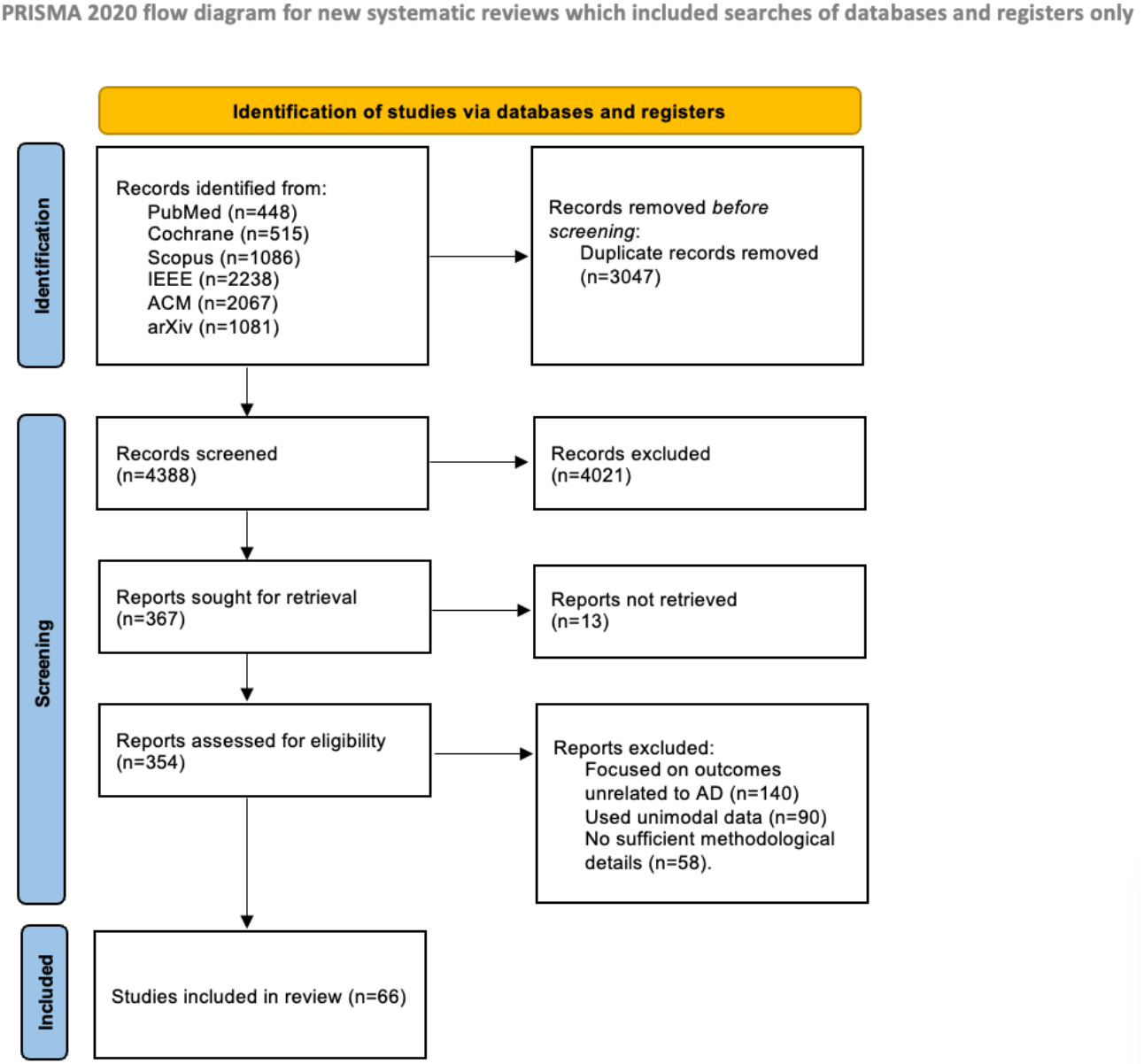

XSL•FO
RenderX

Figure 4. Temporal trends of machine-learning methods used for AD diagnosis (2019 - 2025). AD: Alzheimer disease.

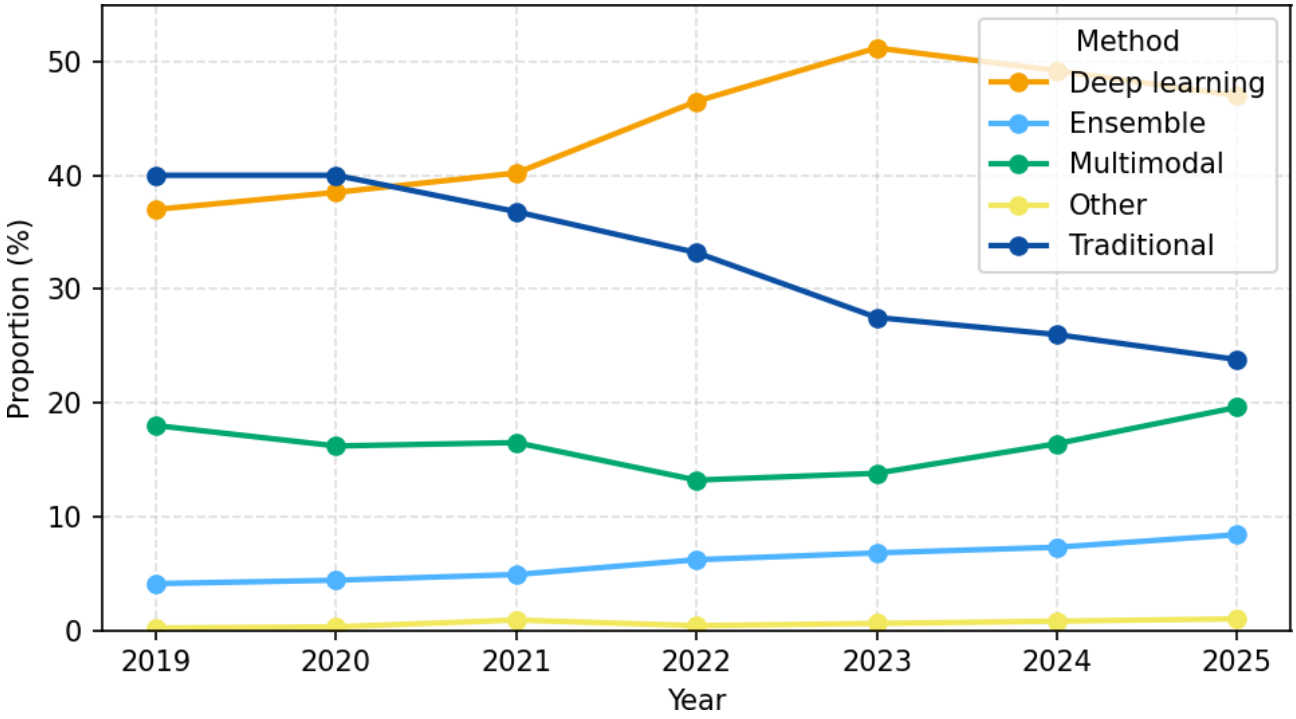

Single Modality

Overview

If readers are already familiar with traditional ML and deep learning approaches, they may wish to proceed directly to the next section, which focuses on multimodal data integration for AD diagnosis.

A concise overview of these baseline methods is provided in Multimedia Appendix 4, which also includes a summary of RL. As most RL studies address sequential decision-making tasks rather than direct diagnostic modeling, their methodological details are presented in Multimedia Appendix 4, to maintain focus on multimodal diagnostic frameworks in the main text.

Deep Learning

Compared with traditional ML, deep learning enables hierarchical feature extraction, capturing complex patterns in high-dimensional data. It is therefore widely used to process and integrate AD-related multimodal inputs, including neuroimaging, clinical scores, genetics, and speech. Key approaches and findings are summarized below.

Recurrent neural networks are effective for modeling sequential data such as longitudinal clinical records and speech signals, but they are susceptible to vanishing gradients in long sequences [46-49]. Long short-term memory networks address this limitation through gated memory mechanisms, enabling more stable training and improved capture of long-term dependencies. Consequently, long short-term memory models have been widely applied in AD research for analyzing temporal and sequential modalities [50-53].

The transformer model, which leverages attention mechanisms, dynamically assigns different weights to input features based on their relative importance. Each layer of the transformer consists of multiple attention heads, allowing the model to capture diverse feature representations by attending to various aspects. Transformer models use attention mechanisms to weight input features and capture diverse representations through multihead attention, enabling efficient and scalable training [54]. Owing to these advantages, they have been widely adopted in AD diagnosis and multimodal learning, where their encoder-decoder architecture facilitates effective integration of heterogeneous data sources [13,55-57].

Ensemble Learning

Ensemble learning improves generalization and robustness by combining multiple base models, including bagging and boosting methods such as AdaBoost (Adaptive Boosting), XGBoost (Extreme Gradient Boosting), and LightGBM (Light Gradient-Boosting Machine), and has been widely applied in AD detection and progression prediction [12,58-60]. However, ensemble models may introduce redundant features, offer limited gains on small datasets, and incur higher computational costs, which can restrict real-time or resource-constrained deployment.

Summarization for Single Modality

Traditional single-modality ML approaches can achieve high performance in AD-related tasks; however, they are constrained by several inherent limitations:

First, regarding information completeness, structural MRI alone has limited sensitivity to functional and molecular changes and cannot fully capture AD-related cognitive and behavioral alterations. Combining MRI with PET, neuropsychological tests, speech, electroencephalography (EEG), and genetic or biomarker data provides a more complete, multidimensional view of disease progression and patient heterogeneity [61].

Second, regarding model robustness, in multimodal data, residual noise in one modality may persist despite denoising, but other modalities can provide complementary signals that

XSL•FO
RenderX

improve robustness. Leveraging multisensory-style integration, multimodal models better reflect biological cognition and can yield more reliable decisions [62].

Third, in cross-modal learning, transformer architectures use cross-modal attention to learn associations between modalities. Some studies apply them in weakly supervised or cross-modal guided settings, using one modality to constrain or guide representation learning in another [63].

Fourth, in real-world decision-making, multimodal learning better matches real-world diagnosis, which integrates multiple information sources. Using diverse modalities aligns models with clinical workflows and improves translational potential in practice [64].

Therefore, this review analyzes the methodological strengths and limitations of multimodal models for computer-assisted AD diagnosis, focusing on how dataset grouping and classification choices affect evaluation. By classifying datasets, we enable model comparisons under a unified setup, allowing for more direct assessment of generalizability and cross-dataset stability.

Multimodal Data

Multimodal Dataset Overview

High-quality data plays a pivotal role in training AI models for computer-aided diagnosis and detection. Robust datasets not only enhance the generalization ability of models but also help mitigate the risk of overfitting. Commonly used datasets for AI-assisted diagnosis of AD can be broadly categorized into 2 types. The first type of multimodal clinical phenotyping dataset, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI), UK Biobank, and the Open Access Series of Imaging Studies (OASIS), focuses on neuroimaging modalities, including MRI, functional MRI, genetic data, and electronic health records. The second type—multimodal cognitive-linguistic behavioral dataset centers on sequential data modalities, including audio, video, and transcribed language data, such as the Pitt Corpus and ADReSS (Alzheimer's Dementia Recognition Through Spontaneous Speech). Each dataset type provides unique features that contribute to the comprehensive modeling of AD progression and diagnosis. The commonly used datasets, along with their population demographics and associated modalities, are summarized in Table 1.

XSL•FO
RenderX

Table . Commonly used dataset.

Dataset	Population demographics		Modalities	Link
	Male	Age (years)		
UK Biobank	23,000 (46)[a]	56.5 (8.1)[b]	MRI[c], fMRI[d], genetic, lifestyle scores, activity monitor, and EHR[e]	[65]
ADNI[f]				
ADNI-1	469 (57.3)[a]	75 (6.9)[b]	MRI, PET[g], genetic, and EHR	[66]
ADNI-GO[h] and 2	473 (53)[a]	72.5 (7.2)[b]	MRI, PET, genetic, and EHR	[66]
ADNI-3	471 (49)[a]	74.9 (8.1)[b]	MRI, PET, genetic, and EHR	[66]
OASIS[i]				
OASIS-1	177 (42.5)[a]	57 (39)[b]	MRI, PET, CT, genetic, and EHR	[67]
OASIS-2	60 (40)[a]	78 (18)[b]	MRI, PET, CT, genetic, and EHR	[67]
OASIS-3	622 (45)[a]	69 (26.56)[b]	MRI, PET, CT, genetic, and EHR	[67]
OASIS-4	663 subjects	57.5 (36.5)[b]	MRI, PET, CT, genetic, and EHR	[67]
NACC[j]	23,625 (44.62)[a]	73.3 (10.5)[b]	MRI, PET, genetics, and EHR	[68]
FHS[k]	718 (42)[a]	80.76 (8.2)[b]	MRI, genetic, and EHR	[69]
AIBL[l]	289 (43.72)[a]	73.5 (7.03)[b]	MRI, PET, genetic, and EHR	[70]
Dementia bank				
Pitt corpus	HC[m]: 104, AD[n]: 208/552	—[o]	Audio and text	[71]
ADReSS[p]	HC: 78, AD: 78/156	—	Audio and text	[71]
ADReSSo[q]	HC: 115, AD: 122/237	—	Audio and text	[71]
ADReSS-M[r]	HC: 143, AD: 148/291	—	Audio and text	[71]
TAUKADIAL[s]	106/507	—	Audio and text	[71]
Multimodal dementia corpus	HC: 10, AD: 12/816	—	Audio, typed, and hand-written	[72]
ADReFV[t]	AD: 25	66.68 (2.08)[b]	Video	[73]
GENCODE[u]	Human: 78,686	—	Genetics	[74]

[a]n (%).

[b]Mean (SD).

[c]MRI: magnetic resonance imaging.

[d]fMRI: functional magnetic resonance imaging.

[e]EHR: electronic health record.

[f]ADNI: Alzheimer's Disease Neuroimaging Initiative.

[g]PET: positron emission tomography.

[h]ADNI-GO: Alzheimer's Disease Neuroimaging Initiative – Grand Opportunity.

[i]OASIS: Open Access Series of Imaging Studies.

[j]NACC: National Alzheimer's Coordinating Centre.

XSL•FO
RenderX

[k]FHS: Framingham Heart Study.

[l]AIBL: Australian Imaging, Biomarkers and Lifestyle Study.

[m]HC: health control.

[n]AD: Alzheimer disease.

[o]Not available.

[p]ADReSS: Alzheimer's Dementia Recognition Through Spontaneous Speech.

[q]ADReSSo: Alzheimer's Dementia Recognition Through Spontaneous Speech 2021 Challenge.

[r]ADReSS-M: Multilingual Alzheimer's Dementia Recognition Through Spontaneous Speech Challenge.

[s]TAUKADIAL: Speech-Based Cognitive Assessment in Chinese and English.

[t]ADReFV: Alzheimer's Disease Recognition from Face & Voice.

[u]GENCODE: gene code.

To better understand current research trends, this paper reviewed studies from the past 5 years on multimodal AI models for AD diagnosis. The literature was categorized by dataset type. (1) Multimodal clinical phenotyping datasets: ADNI dominates this category, used in about 80% of studies, while others, such as UK Biobank, OASIS, National Alzheimer's Coordinating Centre (NACC), Framingham Heart Study, and Australian Imaging, Biomarkers and Lifestyle Study, account for the remaining 20%, mainly for supplementary analysis or external validation. (2) Multimodal cognitive-linguistic behavioral datasets: The ADReSS series is most widely used, representing around 70% of studies. The remaining 30% use datasets often for complementary analysis or benchmarking.

Multimodal Clinical Phenotyping Datasets

Multimodal clinical phenotyping datasets integrate MRI, PET, or diffusion tensor imaging, biomarkers from blood, cerebrospinal fluid, or genomics, and standardized cognitive assessments. This review summarizes representative resources, highlighting their modalities, distinguishing features, and contributions to diagnostic modeling (Table 2).

XSL•FO
RenderX

Table . Multimodal clinical phenotyping datasets related paper. The exceptionally high performances reported in some of these studies can be attributed to specific methodological factors: (1) for two studies [75] and [76], the absence of external validation likely inflated the results; (2) for another study [77], the use of a small but highly controlled dataset, extensive sample expansion, multimodal feature fusion, and pronounced disease-related electrophysiological signatures contributed to the elevated accuracy; and (3) for yet another study [78], the integration of rich gait features with optimized machine learning techniques in a controlled experimental setting facilitated unusually high performance.

Study	Datasets	Model type	Type of task	Modalities	Outcomes	Validation	Results	Limitation
Xue et al [79], 2024	NACC[a], ADNI[b], AIBL[c], FHS[d], PPMI[e], and OASIS[f]	Transformer-based multimodal model	Differential diagnosis of 10 etiologies	MRI[g] (T1, T2, FLAIR[h]), clinical, neuropsychological tests, and PET[i]	Differential diagnosis probabilities and AD[j]/MCI[k]/NC[l] classification	Internal: NACC held out; external: ADNI and FHS	Etiology classification AUROC[m] 0.96; strong alignment with PET biomarkers and neuropathology	Imbalanced etiologies, training label subjectivity, and limited racial diversity
Shi et al [80], 2018	ADNI	MM-SDPN[n]	Classification (AD vs NC, MCI vs NC)	T1 MRI and FDG-PET[o]	Classification accuracy	Comparative vs single-modality and state-of-the-art models	Outperformed single-modality DPN[p]/SDPN[q] and concatenated models	ADNI-only dataset (limits generalizability); ROI[r]-based features rather than voxel-wise
Allwright et al [81], 2023	UK Biobank	XGBoost[s]	Risk prediction (Incident AD)	Demographics, lifestyle, genetics, and medical history	Prediction of incident AD (2-10 years) and risk factor ranking	Internal: nested 3-fold cross-validation; external: held-out validation set	AUROC 0.77, APOE-ε4[t] identified as the strongest risk factor, and liver enzymes or frailty as predictors	*ICD-10*[u] underascertainment, healthy volunteer bias, and observational design
Gu et al [82], 2025	UK Biobank	LightGBM[v]	Risk prediction (incident dementia in ASCVD[w] patients)	Clinical, biological assays, cognitive tests, and physical measures	All-cause incident dementia, AD, and VD[x] incidence	Temporal: train (2006-2009) and test (2010 cohort)	5-year dementia AUC[y] 0.903, AD AUC 0.775, and accuracy 0.851	Sample mostly European descent, static baseline features, and potential overfitting
You et al [83], 2022	UK Biobank	LightGBM	Risk prediction (5/10-year horizon)	Demographics, lifestyle, blood biomarkers, and genetics	Incident all-cause dementia and AD prediction	Internal: 5-fold cross-validation	AUC 0.848 (all-cause), 0.862 (AD), and outperformed CAIDE[z] and DRS[aa]	Limited external validation, population predominantly White, and feature selection fully data-driven
Calvo et al [84], 2024	UK Biobank	Multivariable logistic regression	Risk association analysis	Questionnaire, ICD[ab] records, and genotypes	Odds of AD related to menopause type	Single cohort: multivariable adjustment	Early bilateral oophorectomy associated with 4-fold AD odds (OR[ac] 4.12) and HT[ad] use protective	Low case numbers in subgroups; self-reported HT use and healthy volunteer bias
Yi et al [85], 2025	UK Biobank, ADNI, PPMI, and IXI[ae]	3D-ViT[af]	BAG[ag] estimation and GWAS[ah]	T1-weighted MRI, genetics (SNP[ai] and xQTL[aj])	BAG and drug target prioritization	External: ADNI, PPMI, and IXI	MAE[ak] ≈ 2.6 and identified 7 high-confidence drug targets (eg, MAPT[al] and TNFSF12[am])	European-ancestry bias; lack of biological "ground truth" for brain age

XSL•FO
RenderX

Study	Datasets	Model type	Type of task	Modalities	Outcomes	Validation	Results	Limitation
Yousefzadeh et al [86], 2024	UK Biobank (retina cohort)	VGG-16[an] classifier + LAVA[ao] (XAI[ap])	Binary classification and explainability	Retinal fundus images	AD vs NC classification and neuron-level explanations	Internal: nested 5-fold cross-validation	Accuracy 71.4% and identified 7 latent clusters linking vascular and cognitive decline	Small AD sample size (n=100), cross-sectional design, and UK Biobank volunteer bias
Gong et al [87], 2023	UK Biobank	SuperBigFLICA (semisupervised Bayesian fusion)	Phenotype discovery	Multimodal MRI (47 modalities)	Latent components predictive of non-imaging phenotypes	Internal: train, validation, or test split	Up to 46% improvement over expert IDPs[aq] and interpretable multimodal modes	Linear modeling constraints and UK Biobank population bias
Lian et al [88], 2022	ADNI-1, ADNI-2, and AIBL	Attention-guided HybNet (3D FCN[ar] + hybrid network)	Diagnosis and prognosis	Structural T1 MRI	AD vs NC classification; pMCI[as] vs sMCI[at] prediction	External: trained ADNI-1, and validated ADNI-2 and AIBL	ADNI-2 AD vs NC accuracy 0.919 (AUC 0.965) and outperformed ROI/VBM[au] methods	Heavy preprocessing reliance, limited demographic diversity, and potential overfitting
Lian et al [89], 2022	ADNI-1 and ADNI-2	MWAN[av]	Joint regression of clinical scores	Structural T1 MRI and clinical scores	MMSE[aw], CDRSB[ax], and ADAS-Cog[ay] prediction	Cross-validation: across ADNI-1 and ADNI-2	Lower RMSE[az] and higher correlation coefficients than single-task baselines	Restricted to the ADNI cohorts and potential overfitting to the modest sample size
Li et al [90], 2019	ADNI-1, ADNI-GO/2[ba], and AIBL	3D CNN[bb] + Cox proportional hazards	Time-to-event prognosis	Hippocampal MRI patches and clinical variables	Progression from MCI to AD and risk stratification	External: trained ADNI-1, and validated ADNI-GO/2 and AIBL	C-index 0.864 (combined model) and significant risk-based stratification of MCI	Focus on the hippocampus only and potential cohort and scanner bias (1.5T vs 3T)
Qiu et al [15], 2022	NACC, ADNI, and ADCP[bc]	Multimodal deep learning (3D CNN + FCN)	Multiclass classification	Structural MRI, demographics, and neuropsychology	Diagnosis (NC, MCI, AD, and nADD[bd]) and saliency maps	External: trained NACC, and validated ADNI and independent cohorts	Performance comparable to neurologists and saliency aligned with pathology	Retrospective design and heterogeneity in protocols across cohorts
Oh et al [91], 2023	ADNI	LEAR[be] framework (CNN + RL[bf] + XAI)	Diagnosis and interpretation	Structural T1 MRI	AD vs non-AD classification and counterfactual maps	Internal: cross-validation on ADNI	Improved accuracy and generalization, and localized plausible atrophy patterns	Single-cohort (ADNI) and XAI evaluation, partly qualitative
Lian et al [92], 2020	ADNI-1 and ADNI-2	Hierarchical FCN	Diagnosis and atrophy localization	Structural T1 MRI	AD vs NC, MCI vs NC, and atrophy pattern mapping	External: trained ADNI-1 and tested ADNI-2	Improved accuracy vs conventional features and interpretable atrophy maps	ADNI-only; strong reliance on preprocessing and registration
Avsec et al [93], 2021	Genomic reference datasets	Enformer (transformer)	Genomic prediction	DNA sequence	Gene expression and chromatin state prediction	Internal: held-out chromosomes	Improved capture of long-range regulatory effects vs previous models	Limited to available cell types and assays

XSL•FO
RenderX

Study	Datasets	Model type	Type of task	Modalities	Outcomes	Validation	Results	Limitation
Yang et al [94], 2021	ADNI	Deep learning and super learner	Prognosis	MRI, cognitive, and biomarkers	Diagnostic classification and prognostic risk signature	Internal: cross-validation within ADNI	Derived signature distinguished diagnostic groups and progression risk	Limited external validation and restricted to the ADNI research cohort
Lee et al [95], 2024	ADNI and UK or Singapore Clinics	PPM[bg]	Prognosis (MCI to AD)	MRI (gray matter) and cognitive tests	Individualized prognostic index	External: independent real-world memory clinics	Accuracy ≈81.7%, AUC ≈0.84; and index predicted conversion better than atrophy alone	Heterogeneity in real-world clinical data and potential site effects
Zhu et al [96], 2021	ADNI and AIBL	DA-MIDL[bh]	Diagnosis	Structural MRI patches	AD vs NC and MCI vs NC	External: trained ADNI and tested AIBL	Higher accuracy and generalizability than baselines, and attention maps aligned with pathology	Reliance on structural MRI and potential dataset-specific overfitting
Zhang et al [97], 2024	ADNI	GCN[bi], SHAP[bj], and automatic fusion	Diagnosis	Cognitive, MRI, PET, and risk factors	AD vs non-AD diagnosis, and multimodal feature selection	Internal: two ADNI multimodal cohorts	Accuracies of 95.9% and 91.9%, and efficient selection of clinically important features	Complex model deployment and reliance on ADNI data
Velazquez and Lee [75], 2022	ADNI EMCI[bk]	Ensemble (random forest and CNN)	Prediction of conversion	DTI[bl] (ADC[bm] maps) and EHR[bn]	EMCI to AD conversion prediction	Internal: held-out test set	98.81% accuracy and feature importance explainability provided	Small converter sample size and potential overfitting
Zhang et al [76], 2024	ADNI	Multimodal learning machine (ELM[bo] ensemble)	Diagnosis	MRI features and neuropsychological tests	NC, MCI, and AD classification	Internal: cross-validation on ADNI	>98% accuracy and F_1-score, and no observed bias between MCI and AD	A single research cohort and very high accuracy require external verification
Bi et al [98], 2020	ADNI	Cluster evolutionary random forest	Diagnosis	Resting-state fMRI[bp] and SNP	AD vs control classification and biomarker identification	Comparative vs competing methods	Identified significant brain region-gene pairs and effective classification	Small multimodal sample size and complex hyperparameters
Bi et al [99], 2022	ADNI	Weighted evolutionary random forest	Pathogen detection	Resting-state fMRI and SNP	MCI identification and pathogenic factor extraction	Comparative vs state-of-the-art methods	Superior MCI identification performance, and highlighted key ROIs and SNPs	High-dimensional fusion features, small N, and overfitting risk
Hashmi and Barukab [100], 2023	OASIS	Deep RL and neural network	Staging	Structural MRI	4-class dementia staging	Internal: augmented vs baseline	RL augmentation improved accuracy by ≈6% and recall by ≈13%	Single open dataset; focus on MRI only
Wang et al [101], 2024	ADNI-1/2/3	Multimodal DL[bq] with an interaction layer	Prognosis (MCI to AD)	MRI, clinical, and genetics (SNP)	4-year conversion prediction	External: generalized to ADNI-3	AUC 0.962 (cross-validation), 0.939 (test); interaction effects improved accuracy	ADNI-only and potential overfitting despite cross-validation

XSL•FO
RenderX

Study	Datasets	Model type	Type of task	Modalities	Outcomes	Validation	Results	Limitation
Hatami et al [102], 2024	ADNI	DNN[br] and RL (data augmentation)	Classification	Structural MRI	AD vs NC classification	Comparative vs baseline augmentation approaches	Precision ≈0.95; RL-guided augmentation outperformed baselines	Single research cohort and no external clinical validation
Tabarestani et al [103], 2020	ADNI	Distributed multitask regression	Longitudinal progression	MRI, PET, CSF[bs], EEG[bt], and clinical	Prediction of longitudinal cognitive scores	Comparative vs unimodal or multimodal methods	Reduced errors, particularly in sparse or incomplete longitudinal data	Model complexity and potential sensitivity to hyperparameters
Burkhart et al [104], 2024	ADNI and Singapore Memory Clinic	Unsupervised multimodal trajectory modeling	Prognosis	Cognitive, amyloid PET, and MRI	Cognitive health clustering and progression prediction	External: real-world memory clinic data	Better stratification than standard clinical assessments and robust to missing data	Unsupervised complexity and reliance on ADNI for training
El-Sappagh et al [105], 2021	ADNI	Random forest and SHAP (multilayer)	Diagnosis and progression	11 modalities (MRI, PET, CSF, and clinical)	Multiclass diagnosis and MCI progression detection	Internal: cross-validation	Diagnosis accuracy 93.95%, progression accuracy 87.08%, and interpretable	High complexity and challenges for routine care deployment
Lee et al [106], 2024	ADNI and 4 Korean hospitals	GBM[bu]	Conversion prediction	MRI (T1, T2-FLAIR[bv]), amyloid PET, and clinical	MCI to AD conversion (4-year)	Internal: nested cross-validation with modality combinations	T1 and amyloid PET is the best combination, and T2-FLAIR did not improve prediction	Small multicenter sample and site and scanner heterogeneity
Yuan et al [107], 2021	ADNI	Multimodal cotraining (random forest)	MCI subtype classification	Structural MRI and SNP	sMCI vs pMCI classification	External: ADNI-2 independent test set	Accuracy 85.5% and cotraining outperformed single modality	Dependence on feature selection and ADNI-only
Cirincione et al [108], 2024	TADPOLE[bw] (ADNI)	Ensemble integration	Prediction	MRI, PET, clinical, and cognitive	Future dementia prediction in MCI	Internal: held-out test set	AUC 0.81, and outperformed XGBoost and deep learning baselines	Single research dataset and the complexity of multimodal integration
Cassani and Falk [109], 2020	Clinical EEG	Feature engineering and ML	Diagnosis and severity	Resting-state EEG	AD vs normal, and mild vs moderate AD classification	Internal: cross-validation	Modulation spectral features outperformed traditional EEG features	Small sample size, resting state only, and single center
Cilia et al [110], 2021	Custom (Naples)	Deep transfer learning (CNN)	Diagnosis	Online handwriting (dynamic)	Early AD detection	Internal: cross-validation	Dynamic features (color-encoded) are superior to shape-only images	Single-center dataset and task-specific protocol
Kmetzsch et al [111], 2022	PREV-DEMALS[bx]	Supervised variational autoencoder	Disease progression modeling	MRI and microRNA	Disease progression score (FTD[by]/ALS[bz])	Validation: synthetic data and cohort evaluation	Outperformed competing models in capturing progression trajectory	Small sample (rare disease) and cross-sectional data used for progression

XSL•FO
RenderX

Study	Datasets	Model type	Type of task	Modalities	Outcomes	Validation	Results	Limitation
Mengoudi et al [112], 2020	UCL[ca] and Insight 46	Self-supervised deep neural network	Diagnosis	Eye-tracking (gaze or pupil)	Dementia vs control classification	Comparative vs handcrafted features	Self-supervised features are more sensitive than handcrafted metrics	Modest sample size, mixed dementia subtypes, and specialized hardware
Tsai et al [113], 2024	Taiwan NHI[cb]	MAND[cc]	Incidence prediction	EHR (ICD codes) and demographics	Dementia incidence risk	Internal: held-out test set	AUC 0.901 and outperformed traditional CTR[cd] models	Coding errors in administrative data are specific to the Taiwan NHI
Park et al [22], 2024	Korean memory clinics	SVM[ce]	Diagnosis (MCI vs HC[cf])	VR[cg] biomarkers, MRI, and neuropsychological tests	MCI vs healthy control classification	Internal: train or test split	VR, MRI AUC 0.89, and VR biomarkers comparable to MRI alone	Small sample (n=54) and VR hardware requirement
Wu et al [114], 2022	Clinical EEG	WiGMM[ch]	Severity detection	Resting-state EEG	Unsupervised dementia degree detection	Internal: latent structure analysis	Captured latent dementia degrees matching clinical status	Unsupervised labeling requires careful interpretation
Zhang et al [115], 2025	Chinese memory clinics	FCRN[ci] and MLP[cj] (patch-based)	Diagnosis	MRI, PET, clinical, and genotype	AD vs normal and MCI vs normal classification	Internal: cross-validation	Accuracy ≈96% (AD), ≈92% (MCI), and interpretable probability maps	Single-country clinical cohorts and limited ethnic diversity
Fabietti et al [77], 2023	Mouse models	Ensemble machine learning	Early detection (animal)	LFP[ck]	AD vs control mouse classification	Internal: channel masking robustness tests	Accuracy 99.4% and robust to artifacts	Preclinical animal model results and small sample size
Seifallahi et al [78], 2022	Single center	SVM	Diagnosis	Kinect V2 (gait or TUG[cl])	AD vs healthy control classification	Internal: leave-one-out cross-validation	Accuracy 98.68% using 12 skeletal features	Small sample, case-control design may overestimate performance
Fan et al [116], 2024	CVD[cm] patients (Wuhan)	ViT[cn] (MRI) and XGBoost (clinical)	VCI[co] diagnosis	MRI (T1, T2-FLAIR) and clinical	Vascular cognitive impairment diagnosis	External: independent CVD dataset	The hybrid model has an AUC of 0.965 and is comparable to expert neurologists	CVD-specific cohort and complex ViT and XGBoost pipeline
Beebe-Wang et al [117], 2021	Aging cohort (US)	Nonlinear ML and SHAP	Imminent prediction (3 years)	Clinical, neuropsychological	Incident dementia within 3 years	Internal: cross-validation	Sparse model (4 tests) comparable to full battery	Prediction limited to a 3-year horizon and a single health system

XSL•FO
RenderX

Study	Datasets	Model type	Type of task	Modalities	Outcomes	Validation	Results	Limitation
Battineni et al [118], 2021	Public MRI dataset	Gradient boosting	Classification	MRI features and demographics	AD vs non-AD classification	Internal: cross-validation	Accuracy 97.58% (gradient boosting performed best)	Small public dataset and lack of external validation

[a]NACC: National Alzheimer's Coordinating Centre.
[b]ADNI: Alzheimer's Disease Neuroimaging Initiative.
[c]AIBL: Australian Imaging, Biomarkers, and Lifestyle Study.
[d]FHS: Framingham Heart Study.
[e]PPMI: Parkinson Progression Markers Initiative.
[f]OASIS: Open Access Series of Imaging Studies.
[g]MRI: magnetic resonance imaging.
[h]FLAIR: fluid-attenuated inversion recovery.
[i]PET: positron emission tomography.
[j]AD: Alzheimer disease.
[k]MCI: mild cognitive impairment.
[l]NC: normal control.
[m]AUROC: area under the receiver operating characteristic curve.
[n]MM-SPDN: multimodal stacked deep polynomial network.
[o]FDG-PET: fluorodeoxyglucose-positron emission tomography.
[p]DPN: deep polynomial network.
[q]SPDN: stacked deep polynomial network.
[r]ROI: region of interest.
[s]XGBoost: Extreme Gradient Boosting.
[t]APOE-ε4: apolipoprotein E epsilon 4 allele.
[u]*ICD-10*: *International Statistical Classification of Diseases, Tenth Revision.*
[v]LightGBM: Light Gradient-Boosting Machine.
[w]ASCVD: atherosclerotic cardiovascular disease.
[x]VD: vascular dementia.
[y]AUC: area under the curve.
[z]CAIDE: Cardiovascular Risk Factors, Aging, and Incidence of Dementia.
[aa]DRS: Dementia Risk Score.
[ab]*ICD*: *International Classification of Diseases.*
[ac]OR: odds ratio.
[ad]HT: hormone therapy.
[ae]IXI: Information Extraction From Images.
[af]3D-ViT: 3D vision transformer.
[ag]BAG: brain age gap.
[ah]GWAS: genome-wide association study.
[ai]SNP: single-nucleotide polymorphism.
[aj]xQTL: molecular quantitative trait locus.
[ak]MAE: mean absolute error.
[al]MAPT: microtubule-associated protein tau.
[am]TNFSF12: Tumor Necrosis Factor (Ligand) Superfamily, Member 12.
[an]VGG-16: Visual Geometry Group 16-Layer Network.
[ao]LAVA: Granular Neuron-Level Explainer.
[ap]XAI: explainable artificial intelligence.
[aq]IDP: imaging-derived phenotype.
[ar]FCN: fully convolutional network.
[as]pMCI: progressive mild cognitive impairment.
[at]sMCI: stable mild cognitive impairment.
[au]VBM: voxel-based morphometry.
[av]MWAN: multi-task weakly-supervised attention.

XSL•FO
RenderX

[aw]MMSE: Mini-Mental State Examination.
[ax]CDRSB: Clinical Dementia Rating–Sum of Boxes.
[ay]ADAS-Cog: Alzheimer Disease Assessment Scale–Cognitive Subscale.
[az]RMSE: root mean square error.
[ba]ADNI-GO/2: Alzheimer's Disease Neuroimaging Initiative – Grand Opportunity / Phase 2.
[bb]CNN: convolutional neural network.
[bc]ADPC: Alzheimer Disease Prediction Challenge.
[bd]nADD: non-Alzheimer disease dementia.
[be]LEAR: learn-explain-reinforce.
[bf]RL: reinforcement learning.
[bg]PPM: Predictive Prognostic Model.
[bh]DA-MIDL: dual attention multi-instance deep learning.
[bi]GCN: graph convolutional network.
[bj]SHAP: Shapley Additive Explanations.
[bk]EMCI: early mild cognitive impairment.
[bl]DTI: diffusion tensor imaging.
[bm]ADC: apparent diffusion coefficient.
[bn]EHR: electronic health record.
[bo]ELM: extreme learning machine.
[bp]fMRI: functional magnetic resonance imaging.
[bq]DL: deep learning.
[br]DNN: deep neural network.
[bs]CSF: cerebrospinal fluid.
[bt]EEG: electroencephalography.
[bu]GBM: Gradient Boosting Machine.
[bv]T2-FLAIR: T2-weighted fluid-attenuated inversion recovery.
[bw]TADPOLE: The Alzheimer Disease Prediction of Longitudinal Evolution.
[bx]PREV-DEMALS: Predict to Prevent Frontotemporal Lobar Degeneration and Amyotrophic Lateral Sclerosis.
[by]FTD: frontotemporal dementia.
[bz]ALS: amyotrophic lateral sclerosis.
[ca]UCL: University College London.
[cb]NHI: National Health Insurance.
[cc]MAND: Multimodal Attention Network.
[cd]CTR: clinical trial registration.
[ce]SVM: support vector machine.
[cf]HC: healthy control.
[cg]VR: virtual reality.
[ch]WiGMM: Warped Infinite Gaussian Mixture.
[ci]FCRN: fully convolutional residual network.
[cj]MLP: multilayer perceptron.
[ck]LFP: local field potentials.
[cl]TUG: Timed Up and Go.
[cm]CVD: cardiovascular disease.
[cn]ViT: vision transformer.
[co]VCI: vascular cognitive impairment.

UK Biobank Dataset

UK Biobank enables population-level association studies and early-risk modeling. It has been widely used in AD diagnosis research, including the following notable studies:

Recent UK Biobank–based studies have applied diverse multimodal ML and deep learning approaches for AD risk prediction and diagnosis, integrating neuroimaging, genetic, clinical, and lifestyle data. These models generally achieved moderate to high performance (area under the curve [AUC] ≈0.77 - 0.90) and demonstrated improved diagnostic utility compared with conventional assessment methods [79,81-83]. Several studies further emphasized the importance of genetic and hormonal factors in risk stratification [84,85]. In addition, explainable and semisupervised frameworks have enhanced model interpretability and scalability for population-level analysis, facilitating clinically relevant phenotyping and disease monitoring [86,87].

XSL•FO
RenderX

This section describes multimodal model implementation in the UK Biobank. As shown in the analysis and Table 2, UK Biobank data support AD diagnosis and risk prediction, but limitations remain: class imbalance, which may bias training, and the need for external validation to confirm generalizability beyond the UK Biobank cohort.

ADNI Dataset

ADNI provides a rich and diverse collection of demographic information, multimodal data, and clinical assessments. Owing to its comprehensive scope and longitudinal design, it has become one of the most widely adopted benchmark datasets for computer-aided diagnosis of AD. The following studies exemplify its use:

Recent ADNI-based studies have developed a wide range of multimodal and deep learning frameworks integrating neuroimaging, genetic, cognitive, and clinical data for AD diagnosis and MCI-to-AD progression prediction. Attention-based, multitask, ensemble, and time-to-event models have enabled accurate localization of disease-related regions, improved prognostic modeling, and enhanced interpretability through explainable artificial intelligence techniques such as SHAP (Shapley Additive Explanations) and counterfactual analysis [15,75,76,88-92,97]. Several approaches further incorporated RL, semisupervised learning, and data augmentation to improve robustness and generalizability in heterogeneous and imbalanced datasets [98-103,107]. These models typically achieved high diagnostic and prognostic performance (AUC up to ≈0.96), with some demonstrating strong external validation and clinical relevance [93-96,101,104-106,108]. Nevertheless, existing reviews and benchmarking studies have highlighted persistent limitations, including dataset bias, inconsistent evaluation protocols, and limited cross-center validation, underscoring the need for standardized and reproducible multimodal frameworks [119].

While ADNI provides a comprehensive and standardized multimodal resource for AD research and supports robust model performance, several limitations remain. These include class imbalance, underrepresentation of racially diverse populations, and limited external validation, which may bias model training and restrict generalizability across clinical settings.

Self-Collected Datasets

While public datasets such as ADNI provide standardized benchmarks, self-collected datasets enable more flexible acquisition of targeted modalities. Representative studies include the following.

Studies based on self-collected datasets have explored diverse multimodal fusion strategies. EEG- and local field potentials–based models, as well as hybrid MRI–PET–biomarker frameworks, demonstrated high diagnostic and staging accuracy and supported interpretable risk mapping [77,109,111,114,115]. In parallel, behavioral and digital biomarkers derived from handwriting, eye tracking, virtual reality, and motion capture have enabled noninvasive and low-cost screening with strong classification performance [22,78,110,112]. Large-scale real-world health records and hybrid deep learning models further facilitated population-level risk prediction and vascular cognitive impairment assessment, achieving robust AUC values above 0.90 [113,120]. Overall, self-collected datasets have expanded the scope of multimodal AD research by enabling flexible modality integration and novel biomarker discovery, while remaining constrained by limited sample sizes and heterogeneous acquisition protocols.

Self-collected datasets offer distinct advantages, including targeted modality acquisition, novel biomarker discovery, such as microRNA, local field potentials, and handwriting, and enhanced real-world clinical utility. However, self-collected datasets typically endure limited sample sizes, which increases susceptibility to overfitting and compromises generalizability across diverse populations.

Multimodal Linguistic-Based Cognitive Impairment Datasets

Beyond multimodal clinical phenotyping datasets, multimodal linguistic-based cognitive impairment datasets represent an equally vital research resource. These datasets offer a noninvasive and cost-effective methodology for detecting cognitive decline, particularly valuable for identifying early-stage or subtle impairments where traditional neuroimaging or biomarker data may yield inconclusive results. Capturing spontaneous or semistructured speech and language patterns pushes the development of AI in speech data. Recent work is shown in Table 3.

XSL•FO
RenderX

Table . Multimodal linguistic-based cognitive impairment datasets related papers.

Study	Datasets	Model type	Type of task	Modalities	Outcomes	Validation	Results	Limitation
Ilias et al [121], 2023	ADReSS[a] and ADReSSo[b]	Multimodal transformer (BERT[c] and DeiT[d]) with optimal transport	Dementia detection (AD[e] vs non-AD)	Audio (spectrograms) and text (transcripts)	Classification metrics and calibration	Internal: ADReSS or ADReSSo	Accuracy ≈91.25%, F_1-score ≈91.06%; improved calibration vs baselines	Small, curated datasets, English-only, and potential overfitting
Poor et al [122], 2024	I-CONECT[f]	Multimodal cross-transformer with coattention	MCI[g] prediction (MCI vs NC[h])	Audio, text, and vision (facial video)	AUC[i] scores	Internal: cross-validation	Trimodal AUC 85.3%, and outperformed unimodal (60.9%) and bimodal (76.3%) models	Single cohort (I-CONECT), cross-sectional, and complex architecture
Lin and Washington [123], 2024	DementiaBank (Pitt)	Wav2vec (audio) and Word2Vec (text)	Dementia classification	Audio, text, and timestamps	Accuracy and AUROC[j]	Internal: cross-validation	Text augmentation improved accuracy to ≈80% (AUROC 90%), and timestamps added minimal value	Single corpus: timestamps lacked resolution, and a modest sample size
Ortiz-Perez et al [124], 2023	DementiaBank (Pitt)	Multimodal ensemble (CNN[k] and transformer)	Prediction of dementia signs	Audio and text	Classification accuracy	Internal: held-out test sets	Text-only transformer best (accuracy 90.36%) and audio contributed less than text	Single English dataset, broad diagnosis category, and task constrained to picture description
Ilias and Askounis [125], 2022	ADReSS (DementiaBank)	Transformer (BERT) and Siamese Network	AD identification and severity estimation	Text (transcripts)	Accuracy and interpretability (LIME[l])	Internal: cross-validation	Single-task accuracy 87.50%, multitask accuracy 86.25%, and distinct linguistic patterns identified	Small dataset, text only, MMSE[m] treated as categorical, and no acoustic information
Wen et al [126], 2023	DementiaBank (Pitt)	Transformer and causal counterfactual XAI[n]	AD detection	Text (part-of-speech tag features)	Accuracy; F_1-score; feature importance	Internal: cross-validation	Accuracy 92.2%, F_1-score 0.955, identified 12 key part-of-speech features linked to AD	Text only (part-of-speech), reliance on tagging accuracy, and no acoustic or imaging data
Chen et al [127], 2023	DementiaBank (Pitt)	SpeechFormer + + (hierarchical transformer)	Paralinguistic AD detection	Audio (acoustic features)	Accuracy; F_1-score	Internal: held-out test sets	Outperformed standard transformers and CNN/RNN[o] baselines and SOTA[p] performance	Single corpus, complex computation, audio only, and no cross-lingual evaluation

XSL•FO
RenderX

Study	Datasets	Model type	Type of task	Modalities	Outcomes	Validation	Results	Limitation
Zheng et al [128], 2022	DementiaBank (Pitt)	N-gram, AWD-LSTM[q], or neural models	Dementia detection	Text (context words, stop words, and part-of-speech)	Classification accuracy	Internal: held-out test data	Combined model (vocabulary and grammar) accuracy 81.54%, and grammar contributes comparably to context	Specific to task or language, and moderate performance vs multimodal approaches
Nambiar et al [129], 2022	DementiaBank (Pitt)	Deep Classifiers (BERT/ALBERT[r] + BiLSTM[s])	Early dementia detection	Text (transcripts)	Accuracy; F_1-score	Internal: train and test splits	BERT + BiLSTM accuracy 0.812; ALBERT + BiLSTM F_1-score 0.81; contextual embeddings superior	Text only; reliance on manual transcripts; single dataset
Priyadarshinee et al [130], 2023	ADReSSo-2021	ML[t] classifiers (SVM[u], RF[v], and NN[w])	AD detection	Audio and text (transcripts)	Classification accuracy	Internal: held-out test set	Text features (accuracy 88.7%) outperformed audio, and file-level features were superior to frame-level	Benchmarking context, single task, and single language
Liu et al [131], 2023	ADReSS, ADReSSo, and the local Chinese dataset	Ensemble ML (VAD[x] pause and acoustic)	AD detection	Audio (acoustic and VAD pause features)	Accuracy	Internal: cross-validation; cross-lingual (Chinese)	Ensemble improved accuracy by ≈8% on ADReSS, and accuracy 80% on the local Chinese dataset	Small local dataset (n=10), handcrafted features, and ensemble complexity
Shah et al [23], 2023	ADReSS-M	Logistic regression and SVR	Cross-lingual AD detection; MMSE regression	Audio (duration, pause, and intelligibility) and metadata	Accuracy and RMSE[y]	External: Greek test set	English cross-validation accuracy 74.7%, Greek Test accuracy 69.57%, and MMSE RMSE 4.77 (Greek)	Small Greek sample, modest accuracy, and simple ML models vs deep learning
Mahajan and Baths [132], 2021	ADReSS	Bimodal framework (CNN-LSTM[z] and Speech-GRU[aa])	AD detection	Audio and text	Classification accuracy	Internal: cross-validation	Bimodal enriched model improved performance by ≈6.25% over acoustic baselines	Small dataset, potential overfitting, and single task (picture description)
Mei et al [133], 2023	ADReSS-M	Bilingual wav2vec 2.0 + XGBoost[ab]	Cross-lingual AD detection and MMSE prediction	Audio (acoustic, silence, and low-frequency bands)	Accuracy and RMSE	External: Greek test set	Accuracy 73.9% (Greek), MMSE RMSE 4.610, and low-frequency speech aided transfer	Very small Greek sample, speech-only, and challenge context
Meerza et al [134], 2022	ADReSS	FL[ac] (LSTM[ak] and feed-forward)	Privacy-preserving AD diagnosis	Audio (Mel-frequency and pause features)	Accuracy and fairness metrics	Internal: simulated FL clients	FL accuracy close to the centralized baseline, and q-FedAvg improved fairness	Simulated clients, single dataset, and relies on feature extraction

XSL•FO
RenderX

Study	Datasets	Model type	Type of task	Modalities	Outcomes	Validation	Results	Limitation
Chen et al [135], 2023	ADReSS-M	SVM or NN on pretrained features	Cross-lingual AD detection	Audio (paralinguistic and XLSR-53[ae]), and text (ASR[af]	Accuracy and RMSE	External: Greek test set	Accuracy 69.6% (Greek), RMSE 4.788, and paralinguistic features transferable	Performance below monolingual systems and reliance on ASR quality
Ilias et al [121], 2023	ADReSS	Multimodal transformer (ViT[ag], BERT, and GMU[ah])	AD detection	Audio (spectrograms) and text	Accuracy and F_1-score	Internal: cross-validation	High eighties or low nineties accuracy, ViT is best for acoustic, and fusion surpassed SOTA	Small dataset, binary classification focus, and external generalization untested
Tamm et al [136], 2023	ADReSS-M[ai]	Sequence models (transfer learning)	Cross-lingual AD detection and MMSE	Audio features and demographics	Accuracy and RMSE	External: Greek test set	Accuracy 82.6% (Greek), RMSE 4.345, and ranked second in the challenge	Small Greek sample, acoustic only, and transfer limited to English-Greek
Woszczyk et al [137], 2022	ADReSS	Transformers vs traditional ML	AD detection	Audio and text	Classification accuracy	Internal: held-out test data	Data augmentation improved performance and was comparable to SOTA	Augmentations tuned for ADReSS and a single speech task

XSL•FO
RenderX

Study	Datasets	Model type	Type of task	Modalities	Outcomes	Validation	Results	Limitation
Jin et al [138], 2023	ADReSS-M	CONSEN[aj] ensemble (acoustic and disfluency)	Multilingual AD detection and MMSE	Audio (acoustic embeddings and disfluency)	Accuracy and RMSE	External: Greek test set	First place in the challenge, accuracy 86.69% (Greek), and RMSE 3.727	Challenge dataset, ensemble complexity, and reliance on diarization quality

[a]ADReSS: Alzheimer Dementia Recognition Through Spontaneous Speech.
[b]ADReSSo: Alzheimer's Dementia Recognition Through Spontaneous Speech only.
[c]BERT: Bidirectional Encoder Representations From Transformers.
[d]DeiT: Data-Efficient Image Transformers.
[e]AD: Alzheimer disease.
[f]I-CONECT: Identifying Cognition in the Elderly Through Conversational Engagement.
[g]MCI: mild cognitive impairment.
[h]NC: normal control.
[i]AUC: area under the curve.
[j]AUROC: area under the receiver operating characteristic curve.
[k]CNN: convolutional neural network.
[l]LIME: Local Interpretable Model-Agnostic Explanations.
[m]MMSE: Mini-Mental State Examination.
[n]XAI: explainable artificial intelligence.
[o]RNN: recurrent neural network.
[p]SOTA: state of the art.
[q]AWD-LSTM: Average stochastic gradient descent weight-dropped long short-term memory
[r]ALBERT: A Lite Bidirectional Encoder Representations From Transformers.
[s]BiLSTM: bidirectional long short-term memory.
[t]ML: machine learning.
[u]SVM: support vector machine.
[v]RF: random forest.
[w]NN: neural network.
[x]VAD: voice activity detection.
[y]RMSE: root mean square error.
[z]CNN-LSTM: convolutional neural network long short-term memory.
[aa]Speech-GRU: Speech Gated Recurrent Unit.
[ab]XGBoost: Extreme Gradient Boosting.
[ac]FL: federated learning.
[ad]LSTM: long short-term memory.
[ae]XLSR-53: cross-lingual speech representation-version 53
[af]ASR: automatic speech recognition.
[ag]ViT: vision transformer.
[ah]GMU: gated multimodal unit.
[ai]ADReSS-M: Alzheimer Dementia Recognition through Spontaneous Speech – Multimodal.
[aj]CONSEN: complementary and simultaneous ensemble.

Recent studies have shown that multimodal fusion of speech and text using transformer-based architectures substantially improves AD detection performance, with F_1-scores above 0.90 on ADReSS and ADReSSo (Alzheimer's Dementia Recognition Through Spontaneous Speech 2021 Challenge) datasets [121,132,139]. Linguistic feature engineering and interpretable language models further enhanced classification accuracy, achieving up to 92.2% accuracy and F_1-scores of 0.955 using compact part-of-speech features [124-126,128,130]. Cross-lingual approaches based on language-agnostic and transfer learning methods enabled moderate generalization, with accuracies ranging from 69% to 73.9% in English-Greek transfer settings [23,127,133,136]. To support real-world deployment, lightweight and hierarchical models achieved around 80% accuracy with reduced computational cost [131,135]. In addition, data augmentation and ensemble strategies improved robustness in low-resource scenarios, yielding F_1-score gains of 5% - 7% and competitive challenge performance (accuracy 86.69%) [123,137,138].

XSL•FO
RenderX

Summarization Based on All Multimodal Datasets and Quantitative Analysis

Table 2 and Table 3 summarize the recent state-of-the-art models across the 2 major types of multimodal datasets, extracted according to the Cochrane Handbook. Full QUADAS-2 forms are available in Multimedia Appendix 5. Based on these results, the following quantitative synthesis compares performance trends across all multimodal datasets. Across the 4 major dataset categories, modality choices and model performance show clear dataset-dependent patterns as shown in Table 4. UK Biobank studies mainly combine MRI, clinical variables, and genetic features, with 2 diagnosis studies reporting an average accuracy of 71.4% (SD 5.2%) and 4 risk-prediction studies reaching an average AUC of 0.84 (SD 0.056). ADNI studies use the most comprehensive modality integrations, with 3 diagnosis studies averaging 92.5% (SD 3.8%) accuracy, 3 MCI-conversion studies achieving a mean AUC of 0.922 (SD 0.045), and risk-prediction studies reaching an average AUC of 0.81 (SD 0.06); these tasks collectively achieve the strongest results, with fusion models frequently reporting AUC values above 0.95. DementiaBank studies differ fundamentally by focusing on speech- and language-based modalities; 9 diagnosis studies report an average AUC of 0.813 (SD 0.042), and 5 cross-lingual AD-detection studies show a mean accuracy of 77% (SD 6.5%), where transformer architectures consistently outperform classical approaches, with models such as BERT + DeiT (Data-Efficient Image Transformers), BERT + ViT (vision transformer), and RoBERTa + (Robustly Optimized Bidirectional Encoder Representations From Transformers Approach) DNN (deep neural network) showing F_1-scores exceeding 0.90. Self-collected datasets are typically smaller and more heterogeneous; 3 diagnosis studies report an average accuracy of 96% (SD 2.4%), and lightweight models such as EEGNet or ViT-based hybrids demonstrate strong predictive capacity when applied to EEG or structural MRI.

XSL•FO
RenderX

Table . Summary of representative modality combinations and top-performing models in multimodal AI[a]-aided AD[b] diagnosis.

Dataset and task	Counts	Average performance	Best performance modalities	Related article
UK Biobank				
Diagnosis	2	Accuracy=71.4%	Retinal fundus images	[79,81-87,140]
Risk prediction	4	AUC[c]=84%	Clinical, biological assays, cognitive tests, and physical measures	[79,81-87,140]
Other	3	N/A[d]	Multimodal MRI[e] (T1, T2, MRI, etc)	[79,81-87,140]
ADNI[f]				
Diagnosis	3	Accuracy=92.5%	Structural MRI features and neuropsychological tests	[15,75,76,79,80,89-92,94-108,119,141]
MCI[g] conversion	3	AUC=92.2%	Structural MRI, clinical variables, and genetics (SNP[h])	[15,75,76,79,80,89-92,94-108,119,141]
MMSE[i] regression	2	No integration	Whole-brain T1-weighted MRI and clinical scores	[15,75,76,79,80,89-92,94-108,119,141]
Risk prediction	7	AUC=81%	MRI, PET[j], clinical, and cognitive	[15,75,76,79,80,89-92,94-108,119,141]
Other	13	N/A	N/A	[15,75,76,79,80,89-92,94-108,119,141]
Dementia bank				
Diagnosis	9	AUC=81.3%	Text transcripts → part-of-speech feature vectors	Table 3
Cross-lingual AD detection	5	Accuracy=77%	Multimodal acoustic fusion	Table 3
Other	6	N/A	N/A	Table 3
Self-collected datasets				
Diagnosis	3	Accuracy=96%	MRI, PET, clinical, and genotype	[22,77,78,106,109-115,117,120,142]
Other	6	No integration	Different task	[22,77,78,106,109-115,117,120,142]

[a]AI: artificial intelligence.

[b]AD: Alzheimer disease.

[c]AUC: area under the curve.

[d]N/A: not available.

[e]MRI: magnetic resonance imaging.

[f]ADNI: Alzheimer Disease Neuroimaging Initiative.

[g]MCI: mild cognitive impairment.

[h]SNP: single-nucleotide polymorphism.

[i]MMSE: Mini-Mental State Examination.

[j]PET: positron emission tomography.

To interpret these results and limit metric inflation, note that purely internal cross-validation tends to overestimate performance: AUC is typically ≈5‐15 points higher than with external validation. Small or tightly controlled datasets also report accuracies ≈10%‐20% above those in large, heterogeneous cohorts. Severe class imbalance can further raise accuracy while lowering F_1-score or sensitivity; without correction, imbalance may inflate results by ≈5%‐12%. Cross-sectional models often score higher in single-timepoint evaluations, whereas longitudinal designs usually yield lower but more stable estimates, which are more informative for follow-up and clinical use.

These findings should be interpreted in light of substantial heterogeneity and risk of bias. Variation in sample composition, task definitions, and evaluation procedures across datasets limits direct comparison of performance metrics. QUADAS-2 also indicated frequent unclear and high-risk in-patient selection, reference standards, and flow or timing, especially in studies using only internal validation or selected samples. Reported metrics, therefore, likely represent upper-bound estimates rather than expected real-world performance, and apparent gains often

XSL•FO
RenderX

reflect dataset-specific effects rather than generalizable model superiority.

Overall, the evidence shows that modality effectiveness varies substantially across datasets, transformer models deliver the highest gains in speech-language tasks, and large clinical phenotyping datasets such as UK Biobank and ADNI still rely mainly on traditional machine-learning or custom fusion frameworks rather than modern cross-modal transformers. This gap highlights an opportunity to develop transformer-based multimodal integration approaches tailored to large, heterogeneous clinical datasets.

Multimodal Fusion Taxonomy

A structured multimodal fusion taxonomy clarifies the performance of different integration strategies across datasets (Tables 2 and 3). A total of 4 main paradigms are commonly used: early, intermediate, late, and attention- or graph-based fusion.

Early fusion concatenates low-level features and performs well for aligned modalities such as MRI + PET, often achieving AUC>0.95 in ADNI studies, but is sensitive to missing data and feature-scale heterogeneity. Intermediate fusion combines latent representations from modality-specific encoders and is effective for heterogeneous inputs such as MRI + speech or EEG + clinical data, as demonstrated by high performance in ADReSS-based models, although it may be unstable in small datasets. Late fusion aggregates model outputs and is robust to missing modalities, performing well in large datasets such as the UK Biobank, but underuses fine-grained cross-modal interactions.

Across paradigms, limited modality availability and high acquisition costs remain key challenges, underscoring the need for adaptive and clinically feasible fusion strategies.

Discussion

Principal Findings

This review synthesized multimodal AI studies for AD across diverse dataset families, including clinical phenotyping and cognitive-linguistic datasets. Multimodal fusion generally outperformed unimodal baselines, but the gain is dataset-dependent and should be interpreted cautiously. Strong performance in curated cohorts and constrained speech benchmarks may not generalize to population-based or multicenter settings. QUADAS-2 also indicated frequent risk of bias and unclear reporting across domains, likely inflating metrics and limiting comparability. Accordingly, headline accuracy and AUC should be treated as upper-bound estimates unless supported by external validation and transparent reporting.

Challenges and Future Directions

In recent years, multimodal models have demonstrated remarkable potential in computer-aided diagnosis and risk prediction for AD. While these methods have achieved significant successes, several challenges remain that warrant careful examination. In this section, this systematic review summarizes the common limitations identified in existing studies and proposes directions for future research to advance the field.

Clinical and Translational Implications

Multimodal AI could support AD diagnosis through several clinical pathways. In memory clinics, models combining MRI, cognitive scores, and blood biomarkers could triage referrals, prioritizing patients for specialist review or PET. In general practice, speech-based and routine clinical-feature models could be embedded in consultations to flag early cognitive change. In radiology, MRI-clinical fusion could act as a second reader, reducing interobserver variability and supporting less experienced clinicians. Where imaging or specialist access is limited, speech, digital questionnaires, and basic clinical data could enable telemedicine-based screening and follow-up. At the population level, these models could support risk stratification and targeted monitoring. To enable real-world deployment, research should prioritize external multicenter validation, integration with electronic health records, and evaluation of regulatory feasibility, cost-effectiveness, and clinical impact.

Ethical and Regulatory Implications

Deploying multimodal AI for AD diagnosis requires ethical and regulatory safeguards. As datasets often combine imaging, clinical records, genomics, and speech, they fall under strict privacy regimes (eg, General Data Protection Regulation in the European Union; HIPAA [Health Insurance Portability and Accountability Act] in the United States), requiring explicit consent, data minimization, and secure handling, with added complexity for sensitive modalities such as speech and genomic data. Clinical deployment is also shaped by medical-AI governance frameworks (eg, the European Union AI Act, Food and Drug Administration Software as a Medical Device guidance, and UK Medicines and Healthcare products Regulatory Agency Good Machine Learning Practice), which emphasize transparency, risk management, and postdeployment monitoring. Fairness is essential because demographic imbalance can yield uneven performance across age, ethnicity, and language groups. Interpretability (eg, imaging attention maps and linguistic saliency) supports clinical accountability and aligns with explainability expectations. Future work should incorporate privacy-preserving methods, bias audits, and regulatory-aligned validation pipelines to enable responsible clinical integration.

Data Privacy and Data-Sharing Constraints

Access to multimodal AD data remains severely restricted by privacy regulations and ethical constraints, which limit data sharing and external validation. This restricts the sharing and usage of comprehensive datasets needed for robust external validation and generalizability.

Federated learning (FL) provides a technically viable privacy-preserving solution; however, differences in data formats and institutional infrastructures still impede its large-scale deployment. For instance, Meerza et al [134] pioneered FL for AD speech diagnosis using mel-frequency cepstral coefficients and pause features, maintaining model performance while ensuring privacy through

XSL•FO
RenderX

q-FedAvg/q-FedSGD optimization. Nambiar [129] validated an ALBERT (A Lite Bidirectional Encoder Representations From Transformers) + BiLSTM (bidirectional long short-term memory) hybrid model on the ADReSS dataset, achieving strong performance without compromising data confidentiality. In parallel, multi-institutional collaborations leveraging publicly available datasets such as ADNI, UK Biobank, and OASIS have enabled richer external validation while adhering to rigorous privacy standards [15,79,88,95,100,139,140].

Despite encouraging results, FL still lacks harmonized protocols and interoperable platforms. This limits cross-center reproducibility and weakens clinical credibility. International collaboration also remains constrained by regulatory differences. Future work should prioritize unified federated frameworks with standardized protocols and privacy-preserving methods to enable secure global data collaboration [143,144].

As most datasets lack fully matched modalities per participant, multimodal fusion often relies on representation- or population-level integration rather than early fusion. Early fusion requires paired samples and is therefore infeasible across datasets. By contrast, late fusion and embedding-level integration can train unimodal models separately and combine them via meta-learners, cross-modal transformers, or probabilistic ensembles. Domain adaptation, transfer learning, and harmonization can also combine heterogeneous cohorts at the population level to improve generalizability. A standardized benchmark could further support this by defining shared preprocessing, label taxonomies, and evaluation metrics, enabling meaningful comparison or representation-stage fusion even without subject-level pairing.

Data Imbalance

Severe class imbalance remains a major obstacle, biasing training toward the majority class and inflating accuracy while masking low sensitivity to early disease. In addition, datasets such as the UK Biobank are dominated by White European ancestry, limiting generalizability across racially and ethnically diverse populations. Addressing this requires both technical mitigation and proactive recruitment of underrepresented groups so models better reflect population heterogeneity.

Researchers have applied data-level interventions such as generative adversarial network–based augmentation, diffusion models, and resampling [123,137,138,145,146]; algorithm-level solutions, including cost-sensitive, loss-focused, ensemble, and class-weighted training schemes [147-152]; and evaluation-focused remedies [153] have been developed to mitigate biases.

Current methods frequently introduce new challenges, such as overfitting or inadequate performance in minority classes. Moreover, efforts to increase diversity remain inadequate. Future directions should focus on novel adaptive resampling methods, generative methods for synthetic minority data creation, and dedicated efforts to include and characterize underrepresented populations to ensure equitable and robust clinical applicability across diverse populations.

Lack of Standardized and Longitudinal Data

Differences in acquisition protocols and diagnostic criteria across datasets limit comparability of imaging, cognitive, and biomarker outcomes. Longitudinal evidence is also constrained: even in relatively standardized resources such as ADNI, limited long-term follow-up hampers modeling the temporal dynamics of disease progression.

Future work should standardize key acquisition elements and diagnostic criteria across longitudinal studies and strengthen coordination across institutions. Building on this, a multimodal benchmark spanning imaging, clinical, biomarker, behavioral, and linguistic modalities would enable cross-dataset validation, improve comparability, and support reproducible evaluation of new models. These steps would strengthen temporal modeling and provide more reliable evidence for clinical translation.

Dataset-Specific Limitations

Data imbalance is prevalent across many AD datasets, but the nature of this issue varies substantially between cohorts. This review, therefore, outlines the dataset-specific limitations of commonly used AD cohorts and corpora.

ADNI participants are generally healthier, with fewer comorbidities and a restricted age range (55 - 90 y), limiting representativeness. Protocol differences across centers and evolving diagnostic standards introduce heterogeneity, while frequent reliance on subsets hampers comparability [154].

Of UK Biobank, dementia outcomes are derived mainly from health records, leading to potential misclassification and delayed ascertainment. Participants show strong volunteer bias, and PET or cerebrospinal fluid biomarkers are limited to a small subset, constraining multimodal analyses [155].

OASIS provides open neuroimaging data but with relatively small AD/MCI sample sizes and inconsistent modality coverage. Limited longitudinal depth and cross-scanner variability further reduce reproducibility [156].

Of NACC, data are aggregated from multiple centers with heterogeneous recruitment and diagnostic protocols, making harmonization challenging. The cohort is clinic-based rather than population-representative, and missing biomarker modalities are common [157].

Although high-quality, Australian Imaging, Biomarkers and Lifestyle Study is smaller than ADNI and NACC and is often used only for validation. Regional recruitment and protocol differences reduce ethnic diversity and cross-cohort comparability [158].

Of Pitt Corpus, this is the most widely used speech dataset, but remains small and imbalanced. Tasks are constrained, limiting ecological validity, and cross-linguistic generalizability is poor [159].

Of the ADReSS series, the ADReSS benchmarks provide standardized speech corpora but are modest in size and restricted to English. Narrow task design and small training partitions raise concerns of overfitting and limited external validity [18].

XSL•FO
RenderX

Of self-collected cohorts, locally collected datasets often involve small, single-site samples with heterogeneous acquisition protocols. Missing modalities, limited follow-up, and selection bias further restrict their generalizability [153].

Dataset challenges are compounded by unrepresentative cohorts, incomplete modalities, and poor cross-center consistency, limiting model robustness and cross-dataset generalization in AD diagnosis. Future work should improve data coordination and standardization, enable more practical sharing mechanisms, and adopt cross-cohort validation where feasible. Strengthening data quality and access is essential for translating multimodal AI methods into clinical use.

Model Interpretability and Explainability

A major limitation of multimodal ML models in clinical AD diagnosis is limited interpretability and transparency. Many high-performing models provide insufficient insight into their decision processes, which can hinder clinical adoption and reduce confidence among end users.

Efforts that have been made toward model interpretability include designing inherently transparent models. For example, some studies demonstrate emerging explainability strategies, including hybrid neuro-symbolic models [160] that generate interpretable reports and post hoc methods such as SHAP, LIME (Local Interpretable Model-Agnostic Explanations), gradient-based saliency, and graph-masking techniques [161,162], which collectively enhance transparency in multimodal AD diagnosis.

Current interpretability methods often fail to produce explanations that clinicians can use reliably. Future work should prioritize clinically grounded explainability, including interactive visualizations and concise workflow-aligned natural-language summaries. Hybrid designs that combine deep learning with structured reasoning can further improve transparency by making decision logic explicit. For deployment, models should also report prediction uncertainty and demonstrate compatibility with clinical systems and regulatory requirements.

Beyond technical advances, incorporating patient and public involvement can improve multimodal AI development for AD. Patients and caregivers can help shape evaluation and result communication, not just act as end users, aligning explanations with patient priorities and addressing transparency and fairness. Engaging patient and public involvement earlier in model design may therefore support more interpretable and clinically usable diagnostic tools.

Heterogeneous Multiview Learning Problem

Integrating data across studies is challenging because single datasets rarely cover all modalities, forcing combinations such as ADNI with UK Biobank. However, differences in cohorts, imaging protocols, and cognitive assessment frameworks create substantial heterogeneity that limits direct pooling and comparability.

This heterogeneity hinders building unified models that generalize across nonoverlapping cohorts, so single-dataset models often fail out of domain. Platform-agnostic methods that tolerate missing or inconsistent modalities are therefore needed. Proposed solutions include shared latent-space learning [163], multibranch networks [164], and mixture-of-experts architectures [165] to support partial fusion and cross-dataset adaptation, but most still assume strong cross-domain alignment or require substantial retraining under domain shift.

Despite recent progress, multimodal methods often assume strict cross-domain alignment and require extensive retraining under domain shift or missing modalities. Future work should develop robust, platform-agnostic frameworks that adapt to changing modality availability and distribution shifts with minimal performance loss and advance representation learning to derive stable joint embeddings from heterogeneous data.

Uncertainty Quantification and Clinical Applicability

Although multimodal AD models have advanced, most studies still omit uncertainty quantification (eg, confidence or prediction intervals). Models typically provide deterministic outputs without communicating reliability, despite clinicians relying on uncertainty to guide management and treatment decisions. Future work should embed uncertainty metrics into diagnostic models to better align with clinical needs and improve interpretability, reliability, and real-world adoption.

Risks of Data Leakage in Multimodal AI Modeling

Another limitation is data leakage, which can inflate performance. Common forms include subject-level leakage (samples from the same participant in both training and test sets), patch-level overlap in MRI slice and patch models, and transcript or utterance-level leakage in speech datasets when multiple segments come from 1 individual. Many studies did not report whether participant-independent splits were enforced. Clearer reporting of partitioning and rigorous participant-level cross-validation are therefore essential to ensure real-world generalizability.

Conclusions

This review synthesizes evidence on multimodal AI approaches for AD across clinical, neuroimaging, genetic, and linguistic data, systematically comparing modeling strategies, validation practices, and performance trends across heterogeneous datasets. In contrast to prior modality-specific reviews, the findings show that multimodal models generally outperform unimodal approaches, although performance varies substantially with dataset characteristics, modality availability, and cross-source alignment. High accuracies are often reported in curated or internally validated cohorts, whereas population-based and externally validated studies yield more modest but clinically realistic results, reflecting substantial heterogeneity and risk of bias.

Despite these limitations, the evidence demonstrates that multimodal AI captures complementary biological and behavioral signals relevant to AD, offering clear advantages for diagnosis and risk prediction. Transformer-based architectures and speech- or behavior-derived modalities show promise for scalable and noninvasive early detection. However, meaningful clinical translation will require harmonized benchmarking, transparent reporting, and rigorous external validation. Overall,

XSL•FO
RenderX

this review advances the field by contextualizing performance gains within their methodological constraints and by outlining practical directions for developing robust, interpretable, and generalizable multimodal AI systems. These insights support the responsible integration of AI into real-world dementia screening, risk prediction, and early intervention strategies.

Acknowledgments

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT (Generative Artificial Intelligence for Digital Twins) taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing. The GenAI tool used was ChatGPT-5.2. Responsibility for the final manuscript lies entirely with the authors. Declaration submitted by: JMIR Publications. We used ChatGPT-5.2 (OpenAI) to conduct a grammatical review of the abstract and conclusion sections.

Funding

This work received no specific financial or nonfinancial support. No funder or sponsor had any role in the design of the review; data collection, analysis, or interpretation; writing of this paper; or the decision to submit for publication.

Data Availability

This systematic review did not generate any new datasets. All data extracted and analyzed in this systematic review were obtained from publicly available publications included in the review. No additional unpublished or proprietary data were used.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full database search strategies for PubMed, Scopus, IEEE Xplore, and ACM Digital Library, including complete Boolean queries, search fields, filters, and publication date limits used for study identification.
[DOCX File, 999 KB - jmir_v28i1e85414_app1.docx]

Multimedia Appendix 2
Performance evaluation for AD diagnosis. AD: Alzheimer disease.
[DOCX File, 22 KB - jmir_v28i1e85414_app2.docx]

Multimedia Appendix 3
Complete QUADAS-2 risk-of-bias assessments for all included studies, summarizing judgments across patient selection, index test, reference standard, and flow or timing, with detailed study-level ratings. QUADAS-2: Revised Quality Assessment of Diagnostic Accuracy Studies Tool.
[DOCX File, 38 KB - jmir_v28i1e85414_app3.docx]

Multimedia Appendix 4
Overview of traditional machine-learning models applied in Alzheimer disease research, including SVM, decision trees, HMMs, KNN, logistic regression, GMMs, and foundational CNN or RL descriptions, with methodological principles and limitations. CNN: convolutional neural network; GMM: gaussian mixture models; HMM: hidden markov model ; KNN: k-nearest neighbors; RL: reinforcement learning; SVM: support vector machine.
[DOCX File, 42 KB - jmir_v28i1e85414_app4.docx]

Multimedia Appendix 5
Cochrane Handbook 5.3.3–aligned data-extraction tables summarizing study design, datasets, participants, modalities, preprocessing, model architectures, validation schemes, outcomes, and limitations for all included studies.
[DOCX File, 185 KB - jmir_v28i1e85414_app5.docx]

Checklist 1
Completed PRISMA 2020, PRISMA-S checklist, and PRISMA expanded checklist specifying reporting locations for all required items, including eligibility criteria, search methods, extraction procedures, bias assessments, and synthesis reporting. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for literature searches.
[PDF File, 11026 KB - jmir_v28i1e85414_app6.pdf]

XSL•FO
RenderX

References

1. Scheltens P, De Strooper B, Kivipelto M, et al. Alzheimer's disease. Lancet 2021 Apr 24;397(10284):1577-1590. [doi: 10.1016/S0140-6736(20)32205-4] [Medline: 33667416]
2. 2024 Alzheimer's disease facts and figures. Alzheimers Dement 2024 May;20(5):3708-3821. [doi: 10.1002/alz.13809]
3. World alzheimer report 2024. : Alzheimer's Disease International; 2024 URL: https://www.alzint.org/resource/world-alzheimer-report-2024/ [accessed 2026-02-10]
4. Kaštelan S, Gverović Antunica A, Puzović V, et al. Non-invasive retinal biomarkers for early diagnosis of Alzheimer's disease. Biomedicines 2025 Jan 24;13(2):283. [doi: 10.3390/biomedicines13020283] [Medline: 40002697]
5. Castellano G, Esposito A, Lella E, Montanaro G, Vessio G. Automated detection of Alzheimer's disease: a multi-modal approach with 3D MRI and amyloid PET. Sci Rep 2024 Mar 3;14(1):5210. [doi: 10.1038/s41598-024-56001-9] [Medline: 38433282]
6. Bi Y, Abrol A, Fu Z, Calhoun VD. A multimodal vision transformer for interpretable fusion of functional and structural neuroimaging data. Hum Brain Mapp 2024 Dec 1;45(17):e26783. [doi: 10.1002/hbm.26783] [Medline: 39600159]
7. Yu Q, Ma Q, Da L, et al. A transformer-based unified multimodal framework for Alzheimer's disease assessment. Comput Biol Med 2024 Sep;180:108979. [doi: 10.1016/j.compbiomed.2024.108979] [Medline: 39098237]
8. Leng Y, He Y, Amini S, et al. A GPT-4o-powered framework for identifying cognitive impairment stages in electronic health records. npj Digit Med 2025 Jul 3;8(1):401. [doi: 10.1038/s41746-025-01834-5]
9. Balabin H, Tamm B, Spruyt L, et al. Natural language processing-based classification of early Alzheimer's disease from connected speech. Alzheimer's Dement 2025 Feb;21(2):e14530. [doi: 10.1002/alz.14530] [Medline: 39868827]
10. Yang X, Hong K, Zhang D, Wang K. Early diagnosis of Alzheimer's disease based on multi-attention mechanism. In: Fati SM, editor. PLOS ONE 2024;19(9):e0310966. [doi: 10.1371/journal.pone.0310966]
11. Wijeratne PA, Alexander DC. Learning transition times in event sequences: the event-based hidden markov model of disease progression. Inf Process Med Imaging 2021 Jun;12729(14):583-595. [doi: 10.1007/978-3-030-78191-0_45]
12. Huh YJ, Park JH, Kim YJ, Kim KG. Ensemble learning-based Alzheimer's disease classification using electroencephalogram signals and clock drawing test images. Sensors (Basel) 2025 May 2;25(9):2881. [doi: 10.3390/s25092881] [Medline: 40363322]
13. Karasu E, Baytaş İ. Conversion-aware forecasting of Alzheimer's disease via featurewise attention. Pattern Anal Applic 2025 Jun;28(2):64. [doi: 10.1007/s10044-025-01447-4]
14. Xiao X, Li Y, Wu Q, et al. Development and validation of a novel predictive model for dementia risk in middle-aged and elderly depression individuals: a large and longitudinal machine learning cohort study. Alz Res Therapy 2025 May 13;17(1):103. [doi: 10.1186/s13195-025-01750-6]
15. Qiu S, Miller MI, Joshi PS, et al. Multimodal deep learning for Alzheimer's disease dementia assessment. Nat Commun 2022 Jun 20;13(1):3404. [doi: 10.1038/s41467-022-31037-5] [Medline: 35725739]
16. Chakravarthi BA, Shivakanth G. Integrating multimodal AI techniques and MRI preprocessing for enhanced diagnosis of Alzheimer's disease: clinical applications and research horizons. IEEE Access 2025;13:63519-63531. [doi: 10.1109/ACCESS.2025.3557533]
17. Elazab A, Wang C, Abdelaziz M, et al. Alzheimer's disease diagnosis from single and multimodal data using machine and deep learning models: achievements and future directions. Expert Syst Appl 2024 Dec;255:124780. [doi: 10.1016/j.eswa.2024.124780]
18. Ding K, Chetty M, Noori Hoshyar A, Bhattacharya T, Klein B. Speech based detection of Alzheimer's disease: a survey of AI techniques, datasets and challenges. Artif Intell Rev 2024 Oct 12;57(12):325. [doi: 10.1007/s10462-024-10961-6]
19. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
20. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
21. Li T, Higgins JP, Deeks JJ. Collecting data. In: Cochrane Handbook for Systematic Reviews of Interventions 2019:109-141. [doi: 10.1002/9781119536604]
22. Park B, Kim Y, Park J, et al. Integrating biomarkers from virtual reality and magnetic resonance imaging for the early detection of mild cognitive impairment using a multimodal learning approach: validation study. J Med Internet Res 2024 Apr 17;26:e54538. [doi: 10.2196/54538] [Medline: 38631021]
23. Shah Z, Qi SA, Wang F, et al. Exploring language-agnostic speech representations using domain knowledge for detecting Alzheimer's dementia. ICASSP 2023 - 2023 IEEE Int Conf Acoustics, Speech Signal Process (ICASSP) 2023:1-2. [doi: 10.1109/ICASSP49357.2023.10095593]
24. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/BF00994018]
25. Sharma A, Kaur S, Memon N, Jainul Fathima A, Ray S, Bhatt MW. Alzheimer's patients detection using support vector machine (SVM) with quantitative analysis. Neurosci Inf 2021 Nov;1(3):100012. [doi: 10.1016/j.neuri.2021.100012]

XSL•FO
RenderX

26. Gao X, Liu H, Shi F, Shen D, Liu M. Brain status transferring generative adversarial network for decoding individualized atrophy in Alzheimer's disease. IEEE J Biomed Health Inform 2023 Oct;27(10):4961-4970. [doi: 10.1109/JBHI.2023.3304388] [Medline: 37607152]
27. Lazli L. Improved Alzheimer disease diagnosis with a machine learning approach and neuroimaging: case study development. JMIRx Med 2025 Apr 21;6:e60866. [doi: 10.2196/60866] [Medline: 40257754]
28. Hossain F, Halder RK, Uddin MN. An integrated machine learning based adaptive error minimization framework for Alzheimer's stage identification. Intell-Based Med 2025;11:100243. [doi: 10.1016/j.ibmed.2025.100243]
29. Fulkar B, Dhale T, Pacharaney U, Deshmukh S. Early detection of chronic diseases using machine and deep learning algorithms. 2025 4th Int Conf Sentiment Anal Deep Learn (ICSADL) 2025:1656-1661. [doi: 10.1109/ICSADL65848.2025.10933005]
30. Sathiya A, Basha CH, S V, Sharmila P JJ, S P, Indhumathi R. Enhancing Alzheimer's disease detection using optimized attribute selection and random forest classifier for improved accuracy. 2025 Int Conf Visual Anal Data Visualization (ICVADV) 2025:1174-1179. [doi: 10.1109/ICVADV63329.2025.10961844]
31. Saleh AW, Gupta G, Khan SB, Alkhaldi NA, Verma A. An Alzheimer's disease classification model using transfer learning Densenet with embedded healthcare decision support system. Decis Anal J 2023 Dec;9:100348. [doi: 10.1016/j.dajour.2023.100348]
32. Baucum M, Khojandi A, Papamarkou T. Hidden markov models as recurrent neural networks: an application to Alzheimer's disease. 2021 IEEE 21st Int Conf Bioinf Bioeng (BIBE) 2021:1-6. [doi: 10.1109/BIBE52308.2021.9635256]
33. Cai Z, Zeng D, Marder KS, Honig LS, Wang Y. Dynamic classification of latent disease progression with auxiliary surrogate labels. arXiv. Preprint posted online on Dec 11, 2024. [doi: 10.3390/math11102335]
34. Chen Y, Pham TD. Development of a brain MRI-based hidden Markov model for dementia recognition. Biomed Eng Online 2013;12 Suppl 1(Suppl 1):S2. [doi: 10.1186/1475-925X-12-S1-S2] [Medline: 24564961]
35. Vats NA, Yadavalli A, Gurugubelli K, Vuppala AK. Acoustic features, BERT model and their complementary nature for Alzheimer's dementia detection. IC3 '21 2021 Aug 5:267-272. [doi: 10.1145/3474124.3474162]
36. Xiao R, Cui X, Qiao H, et al. Early diagnosis model of Alzheimer's disease based on sparse logistic regression with the generalized elastic net. Biomed Signal Process Control 2021 Apr;66:102362. [doi: 10.1016/j.bspc.2020.102362]
37. Ablimit A, Botelho C, Abad A, Schultz T, Trancoso I. Exploring dementia detection from speech: cross corpus analysis. ICASSP 2022 - 2022 IEEE Int Conf Acoust, Speech Signal Proc (ICASSP) 2022:6472-6476. [doi: 10.1109/ICASSP43922.2022.9747167]
38. Lahmiri S. Integrating convolutional neural networks, kNN, and Bayesian optimization for efficient diagnosis of Alzheimer's disease in magnetic resonance images. Biomed Signal Process Control 2023 Feb;80:104375. [doi: 10.1016/j.bspc.2022.104375]
39. Suwalka D, Pandita D, Godse S, Patil RR, Salam Khan A, Kumar A. AI applications and simulation-based learning integrating future of nursing education. 2024 Int Conf Intell Innovative Pract Eng Manage (IIPEM) 2024:1-6. [doi: 10.1109/IIPEM62726.2024.10925639]
40. Chaudhari A, Saratkar S, Thute T. AI-enhanced imaging techniques for understanding Alzheimer's progression. 2025 Int Conf Mach Learn Auton Syst (ICMLAS) 2025:1174-1179. [doi: 10.1109/ICMLAS64557.2025.10969042]
41. Ango R, C KKR, Fatima S, Nag A. Brain connectivity analysis in Alzheimer's disease using graph convolutional network. 2024 4th Int Conf Soft Comput Secur Appl (ICSCSA) 2024:133-139. [doi: 10.1109/ICSCSA64454.2024.00028]
42. Chattopadhyay T, Joshy NA, Ozarkar SS, et al. Deep learning algorithms for Alzheimer's disease detection based on diffusion MRI: tests in Indian and North American cohorts. Alzheimer's Dementia 2024 Dec;20(S2):e089294. [doi: 10.1002/alz.089294] [Medline: 39737627]
43. Ma D, Zhang H, Wang L. Editorial: deep learning methods and applications in brain imaging for the diagnosis of neurological and psychiatric disorders. Front Neurosci 2024;18:1497417. [doi: 10.3389/fnins.2024.1497417] [Medline: 39411146]
44. Williams C, Anik FI, Hasan MM, et al. Advancing brain-computer interface closed-loop systems for neurorehabilitation: A systematic review of AI and machine learning innovations in biomedical engineering (preprint). JMIR Biomed Eng 2025 Nov 5;10:e72218. [doi: 10.2196/72218] [Medline: 41191851]
45. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
46. de Swart WK, Loog M, Krijthe JH. A comparative study of methods for dynamic survival analysis. Front Neurol 2025;16:1504535. [doi: 10.3389/fneur.2025.1504535] [Medline: 40040908]
47. Kang MK, Hong KS, Yang D, Kim HK. Multi-scale neural networks classification of mild cognitive impairment using functional near-infrared spectroscopy. Biocybern Biomed Eng 2025 Jan;45(1):11-22. [doi: 10.1016/j.bbe.2024.12.001]
48. Abir SI, et al. EEG functional connectivity and deep learning for automated diagnosis of Alzheimer's disease and schizophrenia. JCSTS 2025 Jan 26;7(1):82-99. [doi: 10.32996/jcsts.2025.7.1.7]
49. Sathish R, Muthukumar R, Dhivya K, Karthikkumar S. Deep learning and IoT-enabled framework for accurate classification and monitoring of alzheimer's disease based on eeg signal analysis. 2025 Fifth Int Conf Adv Electr, Comput, Commun Sustainable Technol (ICAECT) 2025:1-8. [doi: 10.1109/ICAECT63952.2025.10958882]

XSL•FO
RenderX

50. Dubey AK, Kapoor R, Saraswat M. Optimized machine learning for medical data analysis and disease prediction. 2024 Int Conf Artif Intell Emerging Tech (Global AI Summit 2024:1282-1286. [doi: 10.1109/GlobalAISummit62156.2024.10948001]
51. K P, Chitla VB, Aftab A, Kamath S. LSTM-based assistance for people with Alzheimer's disease. 2025 Int Conf Intell Innovative Tech Comput, Electr Electron (IITCEE) 2025:1-5. [doi: 10.1109/IITCEE64140.2025.10915385]
52. Pan J, Fan Z, Smith GE, Guo Y, Bian J, Xu J. Federated learning with multi-cohort real-world data for predicting the progression from mild cognitive impairment to Alzheimer's disease. Alzheimer's Dement 2025 Apr;21(4):e70128. [doi: 10.1002/alz.70128] [Medline: 40219846]
53. Zuo Y, Zhang B, Dong Y, et al. Glypred: lysine glycation site prediction via CCU–LightGBM–BiLSTM framework with multi-head attention mechanism. J Chem Inf Model 2024 Aug 26;64(16):6699-6711. [doi: 10.1021/acs.jcim.4c01034]
54. Zhu M, Xu Z, Zhang Q, Liu Y, Gu D, Xu SD. GCSTormer: gated swin transformer with channel weights for image denoising. Expert Syst Appl 2025 Jul;284:127924. [doi: 10.1016/j.eswa.2025.127924]
55. Han X, Xue R, Feng J, et al. Hypergraph foundation model for brain disease diagnosis. IEEE Trans Neural Netw Learning Syst 2025;36(10):17702-17716. [doi: 10.1109/TNNLS.2025.3554755]
56. Lu SY, Zhang YD, Yao YD. A regularized transformer with adaptive token fusion for Alzheimer's disease diagnosis in brain magnetic resonance images. Eng Appl Artif Intell 2025 Sep;155:111058. [doi: 10.1016/j.engappai.2025.111058]
57. Li X, Zhu W, Qiu P, Dumitrascu OM, Youssef A, Wang Y. A BERT-style self-supervised learning CNN for disease identification from retinal images. arXiv. Preprint posted online on Apr 25, 2025. [doi: 10.48550/arXiv.2504.18049]
58. Mahapatra C. Exploring advanced applications of artificial intelligence in neuropharmacology: a comprehensive overview. Biol Life Sci. Preprint posted online on May 8, 2025. [doi: 10.20944/preprints202505.0369.v1]
59. Ren H, Zheng Y, Li C, et al. Using machine learning to predict cognitive decline in older adults from the Chinese longitudinal healthy longevity survey: model development and validation study. JMIR Aging 2025 Apr 30;8:e67437. [doi: 10.2196/67437] [Medline: 40305830]
60. Shah YAR, Qureshi SM, Qureshi HA, Shah SUR, Ahmad A, Shiwlani A. Advances in artificial intelligence and machine learning for neurodegenerative disease: a literature review. WJRR 2024 Sep 5;19(3):4-18. [doi: 10.31871/WJRR.19.3.8]
61. Fatima G, Ashiquzzaman A, Kim SS, Kim YR, Kwon HS, Chung E. Vascular and glymphatic dysfunction as drivers of cognitive impairment in Alzheimer's disease: insights from computational approaches. Neurobiol Dis 2025 May;208:106877. [doi: 10.1016/j.nbd.2025.106877] [Medline: 40107629]
62. Yang X, Dang X, Cai J, Li J, Wang X, Heng P. Temporal‐multimodal consistency alignment for Alzheimer's cognitive assessment prediction. Med Phys Mex Symp Med Phys 2025 Jun;52(6):5064-5080. [doi: 10.1002/mp.17767]
63. Sadeghian R, Haider F, Fraser K, Tasaki S, Muniz-Terrera G. Editorial: methods in artificial intelligence for dementia 2024. Front Dement 2024;3:1444825. [doi: 10.3389/frdem.2024.1444825] [Medline: 39081595]
64. Kale M, Wankhede N, Pawar R, et al. AI-driven innovations in Alzheimer's disease: integrating early diagnosis, personalized treatment, and prognostic modelling. Ageing Res Rev 2024 Nov;101:102497. [doi: 10.1016/j.arr.2024.102497]
65. UKbiobank. URL: https://ukbiobank.ac.uk [accessed 2026-02-07]
66. ADNI. URL: https://adni.loni.usc.edu [accessed 2026-02-07]
67. Open access series of imaging studies (OASIS). Washington University in St Louis. URL: https://sites.wustl.edu/oasisbrains/ [accessed 2026-02-07]
68. NACC. URL: https://naccdata.org/ [accessed 2026-02-07]
69. Framingham Heart Study. URL: https://www.framinghamheartstudy.org/ [accessed 2026-02-07]
70. aibl. URL: https://aibl.csiro.au [accessed 2026-02-07]
71. TalkBank. URL: https://dementia.talkbank.org/ [accessed 2026-02-07]
72. Gkoumas D, Wang B, Tsakalidis A, et al. A longitudinal multi-modal dataset for dementia monitoring and diagnosis. Lang Resour Eval 2024;58(3):883-902. [doi: 10.1007/s10579-023-09718-4] [Medline: 39323983]
73. Xu T, Wang X, Lun X, Pan H, Wang Z. ADReFV: face video dataset based on human‐computer interaction for Alzheimer's disease recognition. Comput Animation Virtual 2023 Jan;34(1):e2127. [doi: 10.1002/cav.2127]
74. GENCODE. URL: https://www.gencodegenes.org/ [accessed 2026-02-07]
75. Velazquez M, Lee Y. Multimodal ensemble model for Alzheimer's disease conversion prediction from early mild cognitive impairment subjects. Comput Biol Med 2022 Dec;151(Pt A):106201. [doi: 10.1016/j.compbiomed.2022.106201] [Medline: 36370583]
76. Zhang M, Cui Q, Lü Y, Yu W, Li W. A multimodal learning machine framework for Alzheimer's disease diagnosis based on neuropsychological and neuroimaging data. Comput Ind Eng 2024 Nov;197:110625. [doi: 10.1016/j.cie.2024.110625]
77. Fabietti M, Mahmud M, Lotfi A, et al. Early detection of Alzheimer's disease from cortical and hippocampal local field potentials using an ensembled machine learning model. IEEE Trans Neural Syst Rehabil Eng 2023;31:2839-2848. [doi: 10.1109/TNSRE.2023.3288835]
78. Seifallahi M, Mehraban AH, Galvin JE, Ghoraani B. Alzheimer's disease detection using comprehensive analysis of Timed Up and Go Test via Kinect V.2 camera and machine learning. IEEE Trans Neural Syst Rehabil Eng 2022;30:1589-1600. [doi: 10.1109/TNSRE.2022.3181252] [Medline: 35675251]
79. Xue C, Kowshik SS, Lteif D, et al. AI-based differential diagnosis of dementia etiologies on multimodal data. Nat Med 2024 Oct;30(10):2977-2989. [doi: 10.1038/s41591-024-03118-z] [Medline: 38965435]

XSL•FO
RenderX

80. Shi J, Zheng X, Li Y, Zhang Q, Ying S. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. IEEE J Biomed Health Inform 2018 Jan;22(1):173-183. [doi: 10.1109/JBHI.2017.2655720]
81. Allwright M, Mundell HD, McCorkindale AN, et al. Ranking the risk factors for Alzheimer's disease; findings from the UK Biobank study. Aging Brain 2023;3:100081. [doi: 10.1016/j.nbas.2023.100081] [Medline: 37384134]
82. Gu Z, Liu S, Ma H, et al. Estimation of machine learning-based models to predict dementia risk in patients with atherosclerotic cardiovascular diseases: UK Biobank study. JMIR Aging 2025 Feb 26;8:e64148. [doi: 10.2196/64148] [Medline: 40009844]
83. You J, Zhang YR, Wang HF, et al. Development of a novel dementia risk prediction model in the general population: a large, longitudinal, population-based machine-learning study. eClinicalMedicine 2022 Nov;53:101665. [doi: 10.1016/j.eclinm.2022.101665]
84. Calvo N, McFall GP, Ramana S, et al. Associated risk and resilience factors of Alzheimer's disease in women with early bilateral oophorectomy: data from the UK Biobank. J Alzheimers Dis 2024 Nov;102(1):119-128. [doi: 10.3233/JAD-240646] [Medline: 39497303]
85. Yi F, Yuan J, Somekh J, et al. Genetically supported targets and drug repurposing for brain aging: a systematic study in the UK Biobank. Sci Adv 2025 Mar 14;11(11):eadr3757. [doi: 10.1126/sciadv.adr3757] [Medline: 40073132]
86. Yousefzadeh N, Tran C, Ramirez-Zamora A, Chen J, Fang R, Thai MT. Neuron-level explainable AI for Alzheimer's disease assessment from fundus images. Sci Rep 2024 Apr 2;14(1):7710. [doi: 10.1038/s41598-024-58121-8] [Medline: 38565579]
87. Gong W, Bai S, Zheng YQ, Smith SM, Beckmann CF. Supervised phenotype discovery from multimodal brain imaging. IEEE Trans Med Imaging 2023 Mar;42(3):834-849. [doi: 10.1109/TMI.2022.3218720] [Medline: 36318559]
88. Lian C, Liu M, Pan Y, Shen D. Attention-guided hybrid network for dementia diagnosis with structural MR images. IEEE Trans Cybern 2022 Apr;52(4):1992-2003. [doi: 10.1109/TCYB.2020.3005859] [Medline: 32721906]
89. Lian C, Liu M, Wang L, Shen D. Multi-task weakly-supervised attention network for dementia status estimation with structural MRI. IEEE Trans Neural Netw Learning Syst 2022 Aug;33(8):4056-4068. [doi: 10.1109/TNNLS.2021.3055772]
90. Li H, Habes M, Wolk DA, Fan Y. Alzheimer's disease neuroimaging initiative and the Australian Imaging Biomarkers and Lifestyle Study of Aging. A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. Alzheimer's Dementia 2019 Aug;15(8):1059-1070. [doi: 10.1016/j.jalz.2019.02.007] [Medline: 31201098]
91. Oh K, Yoon JS, Suk HI. Learn-explain-reinforce: counterfactual reasoning and its guidance to reinforce an Alzheimer's disease diagnosis model. IEEE Trans Pattern Anal Mach Intell 2023 Apr;45(4):4843-4857. [doi: 10.1109/TPAMI.2022.3197845] [Medline: 35947563]
92. Lian C, Liu M, Zhang J, Shen D. Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. IEEE Trans Pattern Anal Mach Intell 2020 Apr;42(4):880-893. [doi: 10.1109/TPAMI.2018.2889096] [Medline: 30582529]
93. Avsec Ž, Agarwal V, Visentin D, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods 2021 Oct;18(10):1196-1203. [doi: 10.1038/s41592-021-01252-x]
94. Yang L, Wang X, Guo Q, et al. For the Alzheimer's disease neuroimaging initiative. deep learning based multimodal progression modeling for Alzheimer's disease. Stat Biopharm Res 2021 Jul 3;13:337-343. [doi: 10.1080/19466315.2021.1884129]
95. Lee LY, Vaghari D, Burkhart MC, et al. Robust and interpretable AI-guided marker for early dementia prediction in real-world clinical settings. eClinicalMedicine 2024 Aug;74:102725. [doi: 10.1016/j.eclinm.2024.102725]
96. Zhu W, Sun L, Huang J, Han L, Zhang D. Dual attention multi-instance deep learning for Alzheimer's disease diagnosis with structural MRI. IEEE Trans Med Imaging 2021 Sep;40(9):2354-2366. [doi: 10.1109/TMI.2021.3077079]
97. Zhang M, Cui Q, Lü Y, Li W. A feature-aware multimodal framework with auto-fusion for Alzheimer's disease diagnosis. Comput Biol Med 2024 Aug;178:108740. [doi: 10.1016/j.compbiomed.2024.108740] [Medline: 38901184]
98. Bi XA, Hu X, Wu H, Wang Y. Multimodal data analysis of Alzheimer's disease based on clustering evolutionary random forest. IEEE J Biomed Health Inform 2020 Oct;24(10):2973-2983. [doi: 10.1109/JBHI.2020.2973324] [Medline: 32071013]
99. Bi XA, Xing Z, Zhou W, Li L, Xu L. Pathogeny detection for mild cognitive impairment via weighted evolutionary random forest with brain imaging and genetic data. IEEE J Biomed Health Inform 2022 Jul;26(7):3068-3079. [doi: 10.1109/JBHI.2022.3151084] [Medline: 35157601]
100. Hashmi A, Barukab O. Dementia classification using deep reinforcement learning for early diagnosis. Appl Sci (Basel) 2023 Jan 22;13(3):1464. [doi: 10.3390/app13031464]
101. Wang Y, Gao R, Wei T, et al. Predicting long-term progression of Alzheimer's disease using a multimodal deep learning model incorporating interaction effects. J Transl Med 2024 Mar 11;22(1):265. [doi: 10.1186/s12967-024-05025-w]
102. Hatami M, Yaghmaee F, Ebrahimpour R. Investigating the potential of reinforcement learning and deep learning in improving Alzheimer's disease classification. Neurocomputing 2024 Sep;597:128119. [doi: 10.1016/j.neucom.2024.128119]
103. Tabarestani S, Aghili M, Eslami M, et al. A distributed multitask multimodal approach for the prediction of Alzheimer's disease in a longitudinal study. Neuroimage 2020 Feb 1;206:116317. [doi: 10.1016/j.neuroimage.2019.116317] [Medline: 31678502]

XSL•FO
RenderX

104. Burkhart MC, Lee LY, Vaghari D, et al. Unsupervised multimodal modeling of cognitive and brain health trajectories for early dementia prediction. Sci Rep 2024 May 10;14(1):10755. [doi: 10.1038/s41598-024-60914-w] [Medline: 38729989]
105. El-Sappagh S, Alonso JM, Islam SMR, Sultan AM, Kwak KS. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. Sci Rep 2021 Jan 29;11(1):2660. [doi: 10.1038/s41598-021-82098-3] [Medline: 33514817]
106. Lee MW, Kim HW, Choe YS, et al. A multimodal machine learning model for predicting dementia conversion in Alzheimer's disease. Sci Rep 2024 May 29;14(1):12276. [doi: 10.1038/s41598-024-60134-2]
107. Yuan S, Li H, Wu J, Sun X. Classification of mild cognitive impairment with multimodal data using both labeled and unlabeled samples. IEEE/ACM Trans Comput Biol and Bioinf 2021 Nov 1;18(6):2281-2290. [doi: 10.1109/TCBB.2021.3053061]
108. Cirincione A, Lynch K, Bennett J, et al. Prediction of future dementia among patients with mild cognitive impairment (MCI) by integrating multimodal clinical data. Heliyon 2024 Sep 15;10(17):e36728. [doi: 10.1016/j.heliyon.2024.e36728] [Medline: 39281465]
109. Cassani R, Falk TH. Alzheimer's disease diagnosis and severity level detection based on electroencephalography modulation spectral "patch" features. IEEE J Biomed Health Inform 2020 Jul;24(7):1982-1993. [doi: 10.1109/JBHI.2019.2953475] [Medline: 31725401]
110. Cilia ND, D'Alessandro T, De Stefano C, Fontanella F, Molinara M. From online handwriting to synthetic images for Alzheimer's disease detection using a deep transfer learning approach. IEEE J Biomed Health Inform 2021 Dec;25(12):4243-4254. [doi: 10.1109/JBHI.2021.3101982]
111. Kmetzsch V, Becker E, Saracino D, et al. Disease progression score estimation from multimodal imaging and MicroRNA data using supervised variational autoencoders. IEEE J Biomed Health Inform 2022 Dec;26(12):6024-6035. [doi: 10.1109/JBHI.2022.3208517]
112. Mengoudi K, Ravi D, Yong KXX, et al. Augmenting dementia cognitive assessment with instruction-less eye-tracking tests. IEEE J Biomed Health Inform 2020 Nov;24(11):3066-3075. [doi: 10.1109/JBHI.2020.3004686] [Medline: 32749977]
113. Tsai H, Yang TW, Ou KH, Su TH, Lin C, Chou CF. Multimodal attention network for dementia prediction. IEEE J Biomed Health Inform 2024 Nov;28(11):6918-6930. [doi: 10.1109/JBHI.2024.3438885]
114. Wu EQ, Peng XY, Chen SD, Zhao XY, Tang ZR. Detecting Alzheimer's dementia degree. IEEE Trans Cogn Dev Syst 2022 Mar;14(1):116-125. [doi: 10.1109/TCDS.2020.3015131]
115. Zhang H, Ni M, Yang Y, et al. Patch-based interpretable deep learning framework for Alzheimer's disease diagnosis using multimodal data. Biomed Signal Process Control 2025 Feb;100:107085. [doi: 10.1016/j.bspc.2024.107085]
116. Fan CC, Yang H, Zhang C, et al. Graph reasoning module for Alzheimer's disease diagnosis: a plug-and-play method. IEEE Trans Neural Syst Rehabil Eng 2023;31:4773-4780. [doi: 10.1109/TNSRE.2023.3337533]
117. Beebe-Wang N, Okeson A, Althoff T, Lee SI. Efficient and explainable risk assessments for imminent dementia in an aging cohort study. IEEE J Biomed Health Inform 2021 Jul;25(7):2409-2420. [doi: 10.1109/JBHI.2021.3059563]
118. Battineni G, Hossain MA, Chintalapudi N, et al. Improved Alzheimer's disease detection by MRI using multimodal machine learning algorithms. Diagnostics (Basel) 2021 Nov 13;11(11):2103. [doi: 10.3390/diagnostics11112103]
119. Nguyen H, Chu NN. An introduction to deep learning research for Alzheimer's disease. IEEE Consumer Electron Mag 2021 May 1;10(3):72-75. [doi: 10.1109/MCE.2020.3048254]
120. Fan F, Song H, Jiang J, et al. Development and validation of a multimodal deep learning framework for vascular cognitive impairment diagnosis. iScience 2024 Oct;27(10):110945. [doi: 10.1016/j.isci.2024.110945]
121. Ilias L, Askounis D, Psarras J. Detecting dementia from speech and transcripts using transformers. Comput Speech Lang 2023 Apr;79:101485. [doi: 10.1016/j.csl.2023.101485]
122. Poor FF, Dodge HH, Mahoor MH. A multimodal cross-transformer-based model to predict mild cognitive impairment using speech, language and vision. Comput Biol Med 2024 Nov;182:109199. [doi: 10.1016/j.compbiomed.2024.109199] [Medline: 39332117]
123. Lin K, Washington PY. Multimodal deep learning for dementia classification using text and audio. Sci Rep 2024 Jun 16;14(1):13887. [doi: 10.1038/s41598-024-64438-1]
124. Ortiz-Perez D, Ruiz-Ponce P, Tomás D, Garcia-Rodriguez J, Vizcaya-Moreno MF, Leo M. A deep learning-based multimodal architecture to predict signs of dementia. Neurocomputing 2023 Sep;548:126413. [doi: 10.1016/j.neucom.2023.126413]
125. Ilias L, Askounis D. Explainable identification of dementia from transcripts using transformer networks. IEEE J Biomed Health Inform 2022 Aug;26(8):4153-4164. [doi: 10.1109/JBHI.2022.3172479] [Medline: 35511841]
126. Wen B, Wang N, Subbalakshmi K, Chandramouli R. Revealing the roles of part-of-speech taggers in Alzheimer disease detection: scientific discovery using one-intervention causal explanation. JMIR Form Res 2023 May 2;7:e36590. [doi: 10.2196/36590] [Medline: 37129944]
127. Chen X, Pu Y, Li J, Zhang WQ. Cross-lingual Alzheimer's disease detection based on paralinguistic and pre-trained features. ICASSP 2023 - 2023 IEEE Int Conf Acoustics, Speech Signal Proc (ICASSP) 2023:1-2. [doi: 10.1109/ICASSP49357.2023.10095522]
128. Zheng C, Bouazizi M, Ohtsuki T. An evaluation on information composition in dementia detection based on speech. IEEE Access 2022;10:92294-92306. [doi: 10.1109/ACCESS.2022.3203068]

XSL•FO
RenderX

129. Nambiar AS, Likhita K, Pujya K, Gupta D, Vekkot S, Lalitha S. Comparative study of deep classifiers for early dementia detection using speech transcripts. 2022 IEEE 19th India Counc Int Conf (INDICON) 2022:1-6. [doi: 10.1109/INDICON56171.2022.10039705]
130. Priyadarshinee P, Clarke CJ, Melechovsky J, Lin CMY, B. T. B, Chen JM. Alzheimer's dementia speech (audio vs. text): multi-modal machine learning at high vs. low resolution. Appl Sci (Basel) 2023;13(7):4244. [doi: 10.3390/app13074244]
131. Liu J, Fu F, Li L, et al. Efficient pause extraction and encode strategy for Alzheimer's disease detection using only acoustic features from spontaneous speech. Brain Sci 2023 Mar 11;13(3):477. [doi: 10.3390/brainsci13030477] [Medline: 36979287]
132. Mahajan P, Baths V. Acoustic and language based deep learning approaches for Alzheimer's dementia detection from spontaneous speech. Front Aging Neurosci 2021;13:623607. [doi: 10.3389/fnagi.2021.623607] [Medline: 33613269]
133. Mei K, Ding X, Liu Y, et al. The USTC system for ADReSS-M challenge. ICASSP 2023 - 2023 IEEE Int Conf Acoustics, Speech Signal Proc (ICASSP) 2023:1-2. [doi: 10.1109/ICASSP49357.2023.10094714]
134. Ali Meerza SI, Li Z, Liu L, Zhang J, Liu J. Fair and privacy-preserving Alzheimer's disease diagnosis based on spontaneous speech analysis via federated learning. 2022 44th Ann Int Conf IEEE Eng Med Biol Soc (EMBC) 2022:1362-1365. [doi: 10.1109/EMBC48229.2022.9871204]
135. Chen W, Xing X, Xu X, Pang J, Du L. SpeechFormer++: a hierarchical efficient framework for paralinguistic speech processing. IEEE/ACM Trans Audio Speech Lang Process 2023;31:775-788. [doi: 10.1109/TASLP.2023.3235194]
136. Tamm B, Vandenberghe R, Van Hamme H. Cross-lingual transfer learning for alzheimer's detection from spontaneous speech. ICASSP 2023 - 2023 IEEE Int Conf Acoust, Speech Signal Process (ICASSP) 2023:1-2. [doi: 10.1109/ICASSP49357.2023.10096770]
137. Woszczyk D, Hedlikova A, Akman A, Demetriou S, Schuller B. Data augmentation for dementia detection in spoken language. Proc Interspeech 2022 2022:2858-2862. [doi: 10.21437/Interspeech.2022-10210]
138. Jin L, Oh Y, Kim H, et al. CONSEN: complementary and simultaneous ensemble for Alzheimer's disease detection and MMSE score prediction. ICASSP 2023 - 2023 IEEE Int Conf Acoustics, Speech Signal Proc (ICASSP) 2023:1-2. [doi: 10.1109/ICASSP49357.2023.10096253]
139. Ilias L, Askounis D. Context-aware attention layers coupled with optimal transport domain adaptation and multimodal fusion methods for recognizing dementia from spontaneous speech. Knowl Based Syst 2023 Oct;277:110834. [doi: 10.1016/j.knosys.2023.110834]
140. Azevedo T, Bethlehem RAI, Whiteside DJ, et al. Identifying healthy individuals with Alzheimer's disease neuroimaging phenotypes in the UK Biobank. Commun Med 2023 Jul 20;3(1):100. [doi: 10.1038/s43856-023-00313-w]
141. Liang S, Chen T, Ma J, Ren S, Lu X, Du W. Identification of mild cognitive impairment using multimodal 3D imaging data and graph convolutional networks. Phys Med Biol 2024 Dec 7;69(23):235002. [doi: 10.1088/1361-6560/ad8c94]
142. Jahan S, Abu Taher K, Kaiser MS, et al. Explainable AI-based Alzheimer's prediction and management using multimodal data. PLOS ONE 2023;18(11):e0294253. [doi: 10.1371/journal.pone.0294253] [Medline: 37972072]
143. Jahan S, Saif Adib M, Huda SM, et al. Federated explainable AI-based Alzheimer's disease prediction with multimodal data. IEEE Access 2025;13:43435-43454. [doi: 10.1109/ACCESS.2025.3547343]
144. Myrzashova R, Alsamhi SH, Shvetsov AV, Hawbani A, Guizani M, Wei X. BCFTL: blockchain-enabled multimodal federated transfer learning for decentralized Alzheimer's diagnosis. IEEE Internet Things J 2025;12(15):29656-29669. [doi: 10.1109/JIOT.2025.3569652]
145. Chen K, Weng Y, Huang Y, et al. A multi‐view learning approach with diffusion model to synthesize FDG PET from MRI T1WI for diagnosis of Alzheimer's disease. Alzheimers Dement 2025 Feb;21(2):e14421. [doi: 10.1002/alz.14421]
146. Lin W, Lin W, Chen G, et al. Bidirectional mapping of brain MRI and PET with 3D Reversible GAN for the diagnosis of Alzheimer's disease. Front Neurosci 2021;15:646013. [doi: 10.3389/fnins.2021.646013] [Medline: 33935634]
147. Gupta B, Jegannathan GK, Alam MS, et al. Multimodal lightweight neural network for Alzheimer's disease diagnosis integrating neuroimaging and cognitive scores. Neurosci Inf 2025 Sep;5(3):100218. [doi: 10.1016/j.neuri.2025.100218]
148. Chen Z, Wang Z, Zhao M, et al. A new classification network for diagnosing Alzheimer's disease in class-imbalance MRI datasets. Front Neurosci 2022 Aug 25;16:807085. [doi: 10.3389/fnins.2022.807085]
149. Sarma M, Chatterjee D. Multistage diagnosis of Alzheimer's disease from clinical data using 'deep ensemble learning'. JAIAI 2024;01(1):122-138. [doi: 10.54364/JAIAI.2024.1109]
150. Mujahid M, Rehman A, Alam T, Alamri FS, Fati SM, Saba T. An efficient ensemble approach for Alzheimer's disease detection using an adaptive synthetic technique and deep learning. Diagnostics (Basel) 2023 Jul 26;13(15):2489. [doi: 10.3390/diagnostics13152489] [Medline: 37568852]
151. Dubey Y, Bhongade A, Palsodkar P, Fulzele P. Efficient explainable models for Alzheimer's disease classification with feature selection and data balancing approach using ensemble learning. Diagnostics (Basel) 2024 Dec 10;14(24):2770. [doi: 10.3390/diagnostics14242770] [Medline: 39767131]
152. Mandawkar U, Diwan T. Hybrid cuttle Fish-Grey wolf optimization tuned weighted ensemble classifier for Alzheimer's disease classification. Biomed Signal Process Control 2024 Jun;92:106101. [doi: 10.1016/j.bspc.2024.106101]
153. Jasodanand VH, Kowshik SS, Puducheri S, et al. AI-driven fusion of multimodal data for Alzheimer's disease biomarker assessment. Nat Commun 2025 Aug 11;16(1):7407. [doi: 10.1038/s41467-025-62590-4] [Medline: 40789853]

XSL•FO
RenderX

154. Weiner MW, Kanoria S, Miller MJ, et al. Overview of Alzheimer's disease neuroimaging Initiative and future clinical trials. Alzheimer's Dement 2025 Jan;21(1):e14321. [doi: 10.1002/alz.14321] [Medline: 39711072]
155. Wilkinson T, Schnier C, Bush K, et al. Identifying dementia outcomes in UK Biobank: a validation study of primary care, hospital admissions and mortality data. Eur J Epidemiol 2019 Jun;34(6):557-565. [doi: 10.1007/s10654-019-00499-1]
156. Thulasimani V, Shanmugavadivel K, Cho J, Easwaramoorthy SV. A review of datasets, optimization strategies, and learning algorithms for analyzing Alzheimer's dementia detection. Neuropsychiatr Dis Treat 2024;20:2203-2225. [doi: 10.2147/NDT.S496307] [Medline: 39588176]
157. Chan KCG, Xia F, Kukull WA. NACC data: who is represented over time and across centers, and implications for generalizability. Alzheimer's Dement 2025 Sep;21(9):e70657. [doi: 10.1002/alz.70657] [Medline: 40968249]
158. Fowler C, Rainey-Smith SR, Bird S, et al. Fifteen years of the Australian Imaging, Biomarkers and Lifestyle (AIBL) study: progress and observations from 2,359 older adults spanning the spectrum from cognitive normality to Alzheimer's disease. J Alzheimer's Dis Rep 2021 Mar 11;5(1):443-468. [doi: 10.3233/ADR-210005]
159. Yang Q, Li X, Ding X, Xu F, Ling Z. Deep learning-based speech analysis for Alzheimer's disease detection: a literature review. Alz Res Therapy 2022 Dec 14;14(1):186. [doi: 10.1186/s13195-022-01131-3]
160. He Y, Wang Z, Zhang Y, et al. NeuroSymAD: a neuro-symbolic framework for interpretable Alzheimer's disease diagnosis. . Preprint posted online on Mar 1, 2025. [doi: 10.48550/arXiv.2503.00510]
161. Sadeghi A, Hajati F, Argha A, Lovell NH, Yang M. Interpretable graph-based models on multimodal biomedical data integration: a technical review and benchmarking. arXiv. Preprint posted online on May 3, 2025. [doi: 10.48550/arXiv.2505.01696]
162. Mahamud E, Assaduzzaman M, Islam J, Fahad N, Hossen MJ, Ramanathan TT. Enhancing Alzheimer's disease detection: an explainable machine learning approach with ensemble techniques. Intell-Based Med 2025;11:100240. [doi: 10.1016/j.ibmed.2025.100240]
163. Zhou T, Liu M, Thung KH, Shen D. Latent representation learning for Alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data. IEEE Trans Med Imaging 2019 Oct;38(10):2411-2422. [doi: 10.1109/TMI.2019.2913158]
164. Sharma R, Sibille L, Fahmi R. Multi‐branch convolutional neural network for Alzheimer's disease versus normal control classification using PET images. Alzheimer's Dementia 2023 Jun;19(S3):e061092. [doi: 10.1002/alz.061092]
165. Zhang J, Yu X, Chen T, et al. BrainNet-moe: brain-inspired mixture-of-experts learning for neurological disease identification. . Preprint posted online on Mar 5, 2025. [doi: 10.48550/ARXIV.2503.07640]

Abbreviations

AD: Alzheimer disease
AdaBoost: Adaptive Boosting
ADNI: Alzheimer's Disease Neuroimaging Initiative
ADReSS: Alzheimer's Dementia Recognition Through Spontaneous Speech
ADReSSo: Alzheimer's Dementia Recognition Through Spontaneous Speech 2021 Challenge
AI: artificial intelligence
ALBERT: A Lite Bidirectional Encoder Representations From Transformers
AUC: area under the curve
BERT: Bidirectional Encoder Representations From Transformers
BiLSTM: bidirectional long short-term memory
DeiT: Data-Efficient Image Transformers
DNN: DNN
EEG: electroencephalography
FL: federated learning
HIPAA : Health Insurance Portability and Accountability Act
LightGBM: Light Gradient-Boosting Machine
LIME: Local Interpretable Model-Agnostic Explanations
MCI: mild cognitive impairment
ML: machine learning
MRI: magnetic resonance imaging
NACC: National Alzheimer's Coordinating Centre
OASIS: Open Access Series of Imaging Studies
PET: positron emission tomography
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
QUADAS-2: Revised Quality Assessment of Diagnostic Accuracy Studies tool
RL: reinforcement learning

XSL•FO
RenderX

RoBERTa+: Robustly Optimized Bidirectional Encoder Representations From Transformers Approach
SHAP: Shapley Additive Explanations
ViT: vision transformer
XGBoost: Extreme Gradient Boosting

Edited by S Brini; submitted 07.Oct.2025; peer-reviewed by F Elkourdi, MM Sikder, O Oyetunji, S Jiang; revised version received 09.Jan.2026; accepted 09.Jan.2026; published 25.Mar.2026.

Please cite as:
Yu Z, Mulholland A, Huang T, Liu Q
Multimodal AI for Alzheimer Disease Diagnosis: Systematic Review of Datasets, Models, and Modalities
J Med Internet Res 2026;28:e85414
URL: https://www.jmir.org/2026/1/e85414
doi:10.2196/85414

© Ziwen Yu, Anthony Mulholland, Tianyan Huang, Qiang Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of Mobile Health for Improving Medication Adherence in Patients With Cancer: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Wei Chao Liu[1], MN; Li Ping Wang[1,2], PhD; Qian Yu Zhang[1], MN; Jing Han Yang[1], MN

[1]School of Public Health and Nursing, Hangzhou Normal University, No. 2318 Yuhangtang Road, Hangzhou Normal University Cangqian Campus, Hangzhou, China

[2]Engineering Research Center of Mobile Health Management System, Ministry of Education, Hangzhou, China

Corresponding Author:
Li Ping Wang, PhD
School of Public Health and Nursing, Hangzhou Normal University, No. 2318 Yuhangtang Road, Hangzhou Normal University Cangqian Campus, Hangzhou, China

Abstract

Background: Medication adherence among patients with cancer is generally low. Mobile health (mHealth) has gradually been applied to improve this situation, but systematic evidence of its effectiveness remains lacking.

Objective: We aimed to evaluate the effect of mHealth on improving medication adherence among patients with cancer.

Methods: This systematic review included randomized controlled trials (RCTs) evaluating the impact of mHealth on medication adherence among patients with cancer. Systematic searches were conducted in PubMed, Web of Science, CINAHL, Cochrane Library, Embase, Sinomed, CNKI, Cqvip, and ClinicalTrials.gov from inception to December 31, 2025. Two researchers independently performed literature screening, data extraction, and risk of bias assessment. Effects were pooled using a random-effects model (Hartung-Knapp-Sidik-Jonkman), and standardized mean differences (SMDs) and odds ratios (ORs) with 95% CIs have been reported. Evidence quality was assessed using the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) framework.

Results: A total of 17 RCTs (1309 participants) from 8 countries published between 2016 and 2025 were included. mHealth interventions included mobile apps, websites, and text messaging services. The meta-analysis revealed that compared with controls, mHealth interventions significantly improved medication adherence rates (OR 3.47, 95% CI 1.92-6.26; *P*=.002), medication adherence scores (SMD 1.01, 95% CI 0.51-1.52; *P*=.001), self-efficacy (SMD 0.90, 95% CI 0.29-1.51; *P*=.01), and service satisfaction while reducing symptom burden (SMD −0.38, 95% CI −0.61 to −0.14; *P*=.008). However, mHealth had no significant effect on health literacy (SMD 0.51, 95% CI –1.50 to 2.52; *P*=.29). Subgroup analysis revealed that interventions lasting <3 months outperformed those lasting ≥3 months in improving adherence scores (SMD 1.37, 95% CI 0.78-1.96 vs SMD 0.49, 95% CI −0.39 to 1.37; χ^2_1=5.98; *P*=.01). Regarding intervention format, text messaging services demonstrated superior efficacy compared with mobile apps and websites (SMD 1.53, 95% CI −5.49 to 8.55 vs SMD 1.01, 95% CI 0.42-1.61 and SMD 0.11, 95% CI −0.34 to 0.56, respectively; χ^2_2=10.28; *P*=.006). Across cancer types, mHealth most significantly improved adherence scores in patients with breast cancer (SMD 1.29, 95% CI −5.25 to 7.83), outperforming the findings in patients with leukemia and other cancer types (SMD 0.28, 95% CI −0.87 to 1.42 and SMD 1.09, 95% CI 0.10-2.08, respectively; χ^2_2=8.86; *P*=.01).

Conclusions: Our findings confirm that mHealth plays a positive role in improving medication adherence, enhancing patient self-efficacy, increasing patient satisfaction with services, and alleviating symptom burden. However, these findings should be interpreted with caution owing to substantial heterogeneity, a moderate risk of bias, and a low certainty of evidence. Future research should enhance methodological quality by conducting multicenter, large-sample, high-quality RCTs and should explore the long-term effects and cost-effectiveness of mHealth across diverse health care settings and patient populations to clarify its role and value within comprehensive cancer care management systems.

Trial Registration: PROSPERO CRD420251162181; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251162181

(J Med Internet Res 2026;28:e85949) doi:10.2196/85949

KEYWORDS

mobile health; medication adherence; cancer; randomized controlled trial; systematic review; meta-analysis

XSL•FO
RenderX

Introduction

With the continuous rise in global cancer incidence, cancer has become one of the major public health threats to human health. In recent years, the rapid advancement of targeted therapy, immunotherapy, and endocrine therapy has led to the increasingly widespread clinical application of oral anticancer drugs [1]. Compared with intravenous treatment, oral therapy offers advantages, such as convenient administration and lower cost, making it a first-line treatment option in comprehensive cancer management [2]. Statistics indicate that approximately 60% of patients with cancer require 12 - 24 weeks of oral endocrine therapy, targeted therapy, immunotherapy, or chemotherapy [3].

Medication adherence refers to whether patients take their drugs as prescribed by health care professionals and whether they consistently continue the prescribed medications [4]. For oral therapies, treatment efficacy is largely dependent on patient adherence [5]. Unlike intravenous therapy, oral treatment relies primarily on patients and caregivers administering medications independently at home, lacking real-time supervision by health care providers. Compounded by factors, such as prolonged treatment duration, significant drug toxicity and side effects, heavy financial burden, and increased psychological stress, patients with cancer are prone to behaviors such as missed doses, unauthorized dose adjustment, and premature discontinuation [6,7]. Studies indicate that medication adherence among patients with cancer is below 50%, reflecting an overall low level [8].

Poor medication adherence not only reduces the effectiveness of oral therapy but may also cause a series of adverse effects for patients. First, regarding symptom control, patients who fail to take medication regularly or who arbitrarily adjust dosages are prone to fluctuations in blood drug concentrations. This makes it difficult to achieve sustained suppression of tumor growth, thereby diminishing treatment efficacy and exacerbating the burden of symptoms, such as pain, fatigue, and nausea, severely impacting the quality of life of patients [9]. Second, regarding treatment safety, poor medication adherence may increase the risk of adverse events. Failure to follow prescribed regimens not only disrupts established treatment schedules but may also result in insufficient accumulation of effective drug doses, complicating the pattern of adverse reaction occurrence. Clinicians struggle to accurately distinguish drug-induced adverse reactions from natural disease progression, impairing clinical judgment and treatment decisions. This delays optimal intervention timing and increases the likelihood of patients requiring emergency care or unplanned hospitalizations due to complications [10]. Finally, regarding long-term disease outcomes, the greatest concern with poor medication adherence lies in inducing tumor resistance and elevating the risk of recurrence and progression. When tumor cells are repeatedly exposed to drug concentrations insufficient to inhibit their growth, the development of tumor resistance becomes inevitable. This ultimately leads to treatment failure, increasing the risk of tumor recurrence and disease progression, thereby threatening the long-term survival of patients [11]. Consequently, effectively improving medication adherence among patients with cancer has become one of the most pressing challenges in contemporary oncology treatment management.

With the continuous advancement of mobile communication technology, mobile health (mHealth) has emerged as a novel health care service model that is being increasingly applied in chronic disease and cancer management. Utilizing wireless devices, such as smartphones, health monitoring wristbands, and smart pill bottles, mHealth provides services, including medication reminders, symptom monitoring, health education, and remote follow-ups, offering the advantages of cost-effectiveness, convenience, and unrestricted access regardless of time or location [12]. Recent studies indicate that mHealth interventions positively influence medication adherence among patients with cancer [13].

However, current research findings on the impact of mHealth interventions on medication adherence among patients with cancer remain inconsistent. Studies vary greatly in intervention formats and implementation durations, and their effectiveness across different cancer types remains unclear. Although a 2022 systematic review indicated that mHealth interventions can effectively improve medication adherence among chronic disease populations, including patients with cancer [14], systematic evaluations specifically targeting patients with cancer are currently lacking. Furthermore, this review did not perform subgroup analyses based on patient characteristics, intervention formats, or duration, leaving the impact of these variables on adherence improvement unclear. There is a critical need to systematically synthesize existing evidence to more accurately assess the practical efficacy of mHealth interventions in enhancing medication adherence among patients with cancer.

We aim to conduct a systematic review and meta-analysis to systematically integrate evidence on the impact of mHealth interventions on medication adherence among patients with cancer. Through subgroup analyses, we further explore the effects of intervention format, duration, and cancer type on intervention outcomes, thereby providing evidence-based guidance for clinical practice and mHealth intervention design.

Methods

Registration

This systematic review adheres to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist (Checklist 1) and has been registered in PROSPERO (CRD42024522261) [15].

Search Strategy

The initial literature search strategy for this systematic review was designed according to the preregistered protocol. However, the original protocol included only Chinese and English studies and used insufficient search terms, resulting in limitations of the search strategy. Therefore, we optimized the search strategy in the final draft to enhance comprehensiveness and reduce language bias. These modifications have been explicitly reported as post-hoc adjustments in accordance with research standards.

This systematic review's literature search has been reported according to the PRISMA-S statement [16]. A combination of

XSL•FO
RenderX

subject headings and free-text terms was used for systematic retrieval, covering the following databases and registries: PubMed (NCBI), Embase (Embase.com), Web of Science (Clarivate), Cochrane Library (Wiley), CINAHL (EBSCO), China National Knowledge Infrastructure (CNKI), CqVIP (VIP Information), Sinomed (China Biology Medicine Disc; CBMdisc Platform), and ClinicalTrials.gov (NLM). The literature search timeframe was set from database inception to June 28, 2025. To ensure the timeliness of the search results, the retrieval strategy was optimized and rerun on December 31, 2025, to incorporate newly published literature.

The search strategy for this systematic review was developed by refining and optimizing strategies from prior studies [17], and it underwent peer review by experts in library and information retrieval. Search parameters were set to include only randomized controlled trials (RCTs) and peer-reviewed articles, with no restrictions on publication year or language, and no predefined filters were applied. To maximize retrieval volume, snowball citation searching was performed. This involved screening the reference lists of the included studies (backward citation searching) as well as identifying and reviewing studies that cited the included articles (forward citation searching). It should be noted that the literature search for this review was conducted exclusively using the aforementioned method and did not include searches of printed conference proceedings or government websites. Detailed search strategies are provided in Multimedia Appendix 1.

Inclusion and Exclusion Criteria

Inclusion Criteria

The inclusion criteria were predefined using the PICOS (participants, interventions, control/comparison measures, outcomes, study type) framework [15]. The criteria were as follows: (1) participants: patients with a pathologically confirmed cancer diagnosis who were currently taking medication; (2) interventions: mHealth interventions designed to improve medication adherence among patients with cancer, including apps, websites, and text messaging services; (3) control/comparison measures: placebo controls, routine or standard interventions, and other non-mHealth interventions; (4) outcomes: primary outcomes, including medication adherence rate and mean medication adherence score, and secondary outcomes, including health literacy, self-efficacy, symptom burden, and satisfaction; and (5) study type: studies with a RCT design.

Exclusion Criteria

We excluded the following studies: (1) non–peer-reviewed research, such as research proposals, conference abstracts, and reviews, typically lacking rigorous methodological validation; (2) those for which full-text access was unavailable; and (3) those lacking accessible or convertible medication adherence data.

Definition and Classification of mHealth

The mHealth interventions included in this review varied in their implementation formats. To facilitate subsequent integration and comparison, we predefined the following classification criteria based on the primary delivery platform and user interaction model of the intervention content [18]:

1. Mobile apps: Interventions are primarily implemented through dedicated software installed on smartphones or tablets. Such apps can be accessed via mainstream app stores (eg, App Store and Google Play) or through links provided by the research team. They typically feature interactive interfaces (eg, menus, buttons, and data input fields) and support offline functionality for at least some features.
2. Websites: Interventions are primarily delivered through mobile device browsers or web applications (without requiring local software installation). Key characteristics include: access via URL, interfaces optimized for mobile screens (responsive design), and typical requirement of a persistent internet connection.
3. Text messaging services: Interventions are primarily delivered via short messaging or multimedia messaging systems, push text, or multimedia content. These services rely on cellular network messaging systems using asynchronous communication. Owing to character length limitations, the content is typically concise.

Study Selection

After importing the retrieved studies into EndNote 20.6 (Clarivate Analytics) and removing duplicates using the software, 2 researchers (JHY and QYZ) were invited to cross-review the titles and abstracts to complete the initial screening and full-text reading of the remaining literature. When disagreements arose between the 2 researchers regarding inclusion decisions, a third researcher (WCL) was consulted to participate in discussions. Where necessary, the third researcher made the final determination on whether to include a study. For studies lacking medication adherence data or where full-text access was unavailable, we first attempted to contact the corresponding author to obtain the data. If the author did not respond or the data could not be obtained, the study was excluded from the meta-analysis.

Data Extraction

For included studies, data were extracted using Microsoft Excel software, encompassing the following: (1) general information (eg, first author's name, nationality, publication year, theoretical framework, study design, and background); (2) characteristics of patients with cancer (eg, sample size, mean age/age range, and disease type); (3) intervention characteristics (eg, intervention format, intervention content, intervention duration, and intervention frequency); and (4) primary outcomes and measurement tools.

Definition and Classification of Outcome Indicators

Primary Outcome Indicators

Medication adherence was defined as the degree of alignment between a patient's actual medication-taking behavior and the prescribed treatment regimen.

To clarify how this metric was measured across studies, we documented the following three operational dimensions during data extraction:

XSL•FO
RenderX

1. Measurement tools: Categorized into 1 of 5 types: electronic monitoring devices (eg, smart pillboxes), pharmacy dispensing/dispensing records, standardized self-assessment scales (eg, Morisky scale), researcher-designed questionnaires or interviews, and pill counting methods.
2. Data format and definition: Extracted each study's specific criteria for defining "adherence" (eg, "number of missed doses in the past 7 days" or "proportion of days adhering to the prescription ≥80%") and recorded the reporting format of results (continuous variable or dichotomous variable).
3. Processing procedures: For continuous data (eg, mean adherence score), directly extract mean, SD, and sample size; for dichotomous data (eg, adherence rate), extract the number of adherence events and the total sample size.

Based on this classification, we conducted separate meta-analyses for studies reporting continuous and dichotomous variables. This approach avoids introducing additional assumptions and potential bias from forced data format conversion. When pooling continuous data, since different adherence scoring tools (with varying scales) were used across studies, we selected the standardized mean difference (SMD) as the pooled effect measure. This allowed all studies to be included in a unified statistical comparison model. It is important to emphasize that while SMD primarily addresses statistical comparability, it does not eliminate potential conceptual differences underlying varying operational definitions. Such differences may influence the clinical interpretation of results, necessitating caution in analysis. For dichotomous variables, we selected the odds ratio (OR) as the effect measure owing to its greater prevalence and interpretability in epidemiological research.

Secondary Outcome Indicators

The secondary outcome indicators were as follows:

1. Symptom burden: The overall experience of patients with cancer regarding symptoms arising from the disease and treatment in terms of frequency, severity, and functional impact [19].
2. Self-efficacy: The strength of confidence patients with cancer have in their ability to successfully perform and adhere to specific disease management and treatment-related behaviors [20].
3. Service satisfaction: The level of satisfaction with health care services among patients with cancer [21].
4. Health literacy: The ability of patients with cancer to access, understand, evaluate, and apply health information to make appropriate health decisions for maintaining and promoting their own health [22].

Risk of Bias, Level of Evidence, and Methodological Quality Assessment

Two researchers (QYZ and JHY) assessed the risk of bias in RCTs using the 7 domains of The Cochrane Collaboration's tool for risk of bias assessment (version 2.0) [23]. The Cochrane risk of bias assessment tool encompasses 5 key domains: randomization process, intervention deviation, missing data, outcome assessment, and selective reporting. Risks within each domain are categorized as low risk, high risk, or unclear risk. Studies meeting low risk in all domains are classified as grade A, those meeting criteria in some domains are classified as grade B, and those failing to meet criteria in any domain are classified as grade C. This systematic review included only grade A or B literature, and evidence quality was assessed using GRADE Pro software (Evidence Prime Inc). This tool categorizes evidence quality into 4 levels: high, moderate, low, and very low. Factors leading to downgrading encompass 5 aspects: risk of bias, imprecision, inconsistency, indirectness, and publication bias. High-quality evidence indicates very high certainty that the true effect is close to the estimated effect; moderate-quality evidence reflects moderate confidence in the effect estimate, with the true effect potentially close to but possibly substantially different from the estimate; low-quality evidence indicates limited certainty about the effect estimate, with the true effect likely substantially different from the estimate; and very low–quality evidence indicates almost no confidence in the effect estimate, with the true effect highly likely to be substantially different from the estimate [24]. When 2 researchers (QYZ and JHY) disagreed on the risk of bias or evidence grade of a study, a third researcher (WCL) was consulted. If necessary, the third researcher made the final decision on whether to include the study.

Data Analysis

A narrative synthesis approach was used to summarize study characteristics, including research design, publication year, country, participant characteristics, and intervention features. For outcome measures unsuitable for meta-analysis, a narrative synthesis approach was also applied for analysis.

Although the preregistration protocol specified the use of Stata (Version 17.0; StataCorp) and RevMan (Version 5.4; Cochrane) for statistical analysis, the final statistical analysis was conducted in RStudio (Posit) using a conservative random-effects model based on the assumption of between-study heterogeneity. Effect sizes were pooled using the Hartung-Knapp-Sidik-Jonkman (HKSJ) method [25]. To enhance methodological rigor and clinical interpretability, multiple analytical measures were introduced during the optimization phase: the Egger test was used to assess small-sample effects, 95% prediction intervals (PIs) were incorporated, τ^2 was used to quantify heterogeneity, and the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) approach was used to assess evidence quality [26].

For continuous variables, the SMD was used to calculate the pooled effect size, interpreted according to Cohen *d* criteria: 0.2, 0.5, and 0.8 represent small, moderate, and large effects, respectively. For dichotomous variables, the OR was used. Each pooled effect size is accompanied by a 95% CI. Additionally, we report PIs to quantify the variability of actual effects across different populations or settings [27].

The Cochrane Collaboration recommends the HKSJ method for random-effects meta-analysis [28]. When facing insufficient statistical power due to a small number of included studies, random-effects meta-analysis based on the HKSJ method outperforms the standard DerSimonian-Laird method [27]. By

XSL•FO
RenderX

using a t-distribution and adjusting SEs using the q statistic, the HKSJ method provides more robust effect size CI estimates. Particularly with limited studies, HKSJ-estimated CIs are more accurate as they better reflect the inherent uncertainty and imprecision of effect estimates [29]. Model selection should be based on conceptual assumptions about the true effect distribution rather than descriptive statistics like I^2 [26]. Given the presence of differences in participant characteristics, interventions, and measurement tools across the included studies, a conservative random-effects model was used for all analyses. This systematic review used restricted maximum likelihood estimation to account for between-study variance, thereby more accurately reflecting the impact of true heterogeneity on effect estimates. Heterogeneity was quantified using Q-tests, I^2, and τ^2 measures [27]. All statistical analyses were conducted in the R Studio environment using R software (version 4.5.2), utilizing packages including metafor, forestplot, and ggplot2.

Subgroup analyses of postintervention data were based on priority intervention criteria: (1) intervention duration, (2) cancer type, and (3) intervention type. Sensitivity analyses were conducted to assess result stability. Funnel plots were generated, and Egger regression intercept tests were applied to examine small-sample effects. Note that these methods aim to detect small-sample effects rather than directly assess publication bias [30,31]. Small-sample effects may stem from publication bias, selective reporting, genuine heterogeneity between studies, and methodological differences. According to methodological standards, funnel plots require at least 10 studies for reliable interpretation; otherwise, the reliability of interpretations is low [30]. Egger regression intercept analysis was applied to all outcome measures in this systematic review.

Deviations From the Registered Protocol

This systematic review strictly followed the protocol registered in advance on the PROSPERO platform (CRD42024522261). During methodological refinement and peer review, we identified the following limitations in the original protocol and made two subsequent revisions accordingly:

1. Optimization of the search strategy: The original protocol initially proposed a search strategy combining subject headings with free-text terms. Following peer and expert review, we adopted recommendations to “expand search terms to reduce the risk of omission.” This involved supplementing search terms, such as adding “online*” and “drug therapy computer assisted,” to enhance search comprehensiveness.
2. Change in the statistical analysis tool: Although the original protocol specified Stata (version 17.0) and RevMan (version 5.4) for data analysis, during implementation, we ultimately selected RStudio to perform all statistical analyses. This decision was made to leverage RStudio’s flexibility in data visualization and statistical model implementation.

Results

Literature Search Results

This literature search identified a total of 848 studies, including 846 from databases and clinical trial registries and 2 from reference lists. After deduplication, 703 studies were ultimately retained. Following preliminary screening of titles and abstracts against the inclusion criteria, 30 studies underwent full-text assessment. After full-text evaluation, 13 studies were excluded: 6 due to incomplete data, 5 due to unavailable full texts, and 2 due to differing outcome measures. Ultimately, 17 studies [19-22,32-44] were included (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

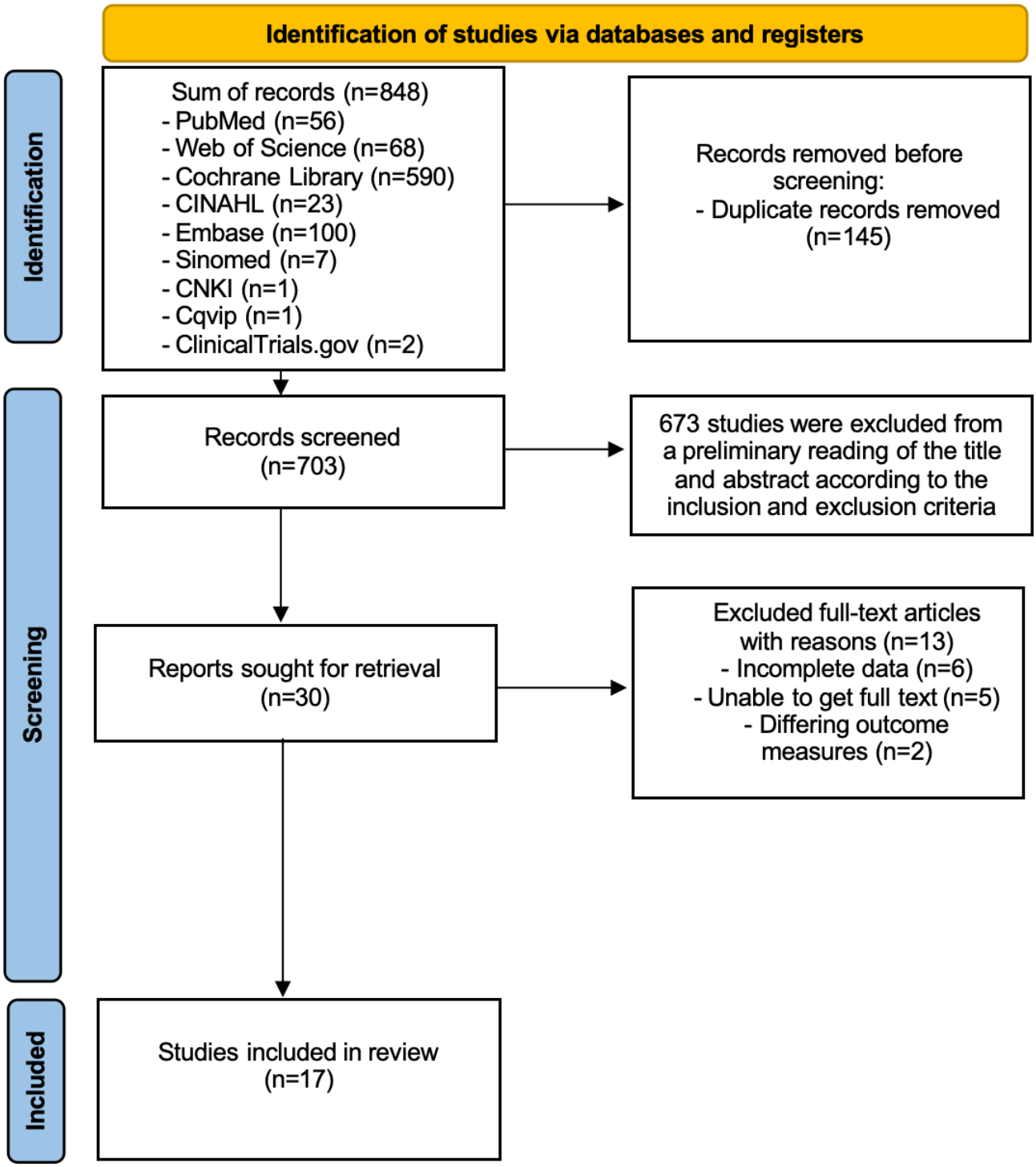

Characteristics of the Included Studies

This systematic review included 17 RCTs. Six studies were published in Simplified Chinese [20,21,36,37,42,44], while the remaining 11 were English-language publications [19,22,32-35,38-41,43], spanning publication dates from 2016 to 2025. Studies originated from 8 countries: the United States, China, Turkey, Australia, the Netherlands, South Korea, Iran, and Finland, and all were conducted in hospital and home settings.

The studies included a total of 1309 participants, with sample sizes ranging from 28 to 166 cases. Cancer types included breast cancer (3 studies) [33,39,43]; leukemia (4 studies) [22,35,40,41]; unspecified cancer types (7 studies) [19,32,34,37,38,42,44]; and thyroid cancer, pheochromocytoma, and prostate cancer (1 study each) [20,21,36].

Regarding intervention formats, the primary methods used were mobile apps (10 studies) [19,33,34,36,37,40-44], text messaging services (5 studies) [20,32,35,38,39], and websites (2 studies) [21,22]. Additionally, 2 studies utilized smart pill bottles as supplementary tools, integrated with either mobile apps or text messaging services [39,43]. Regarding intervention content, mobile apps and websites primarily enhanced medication adherence through online Q&A sessions, multiplayer online

XSL•FO
RenderX

games, and science-popularization video pushes, while text messaging services focused on delivering personalized medication reminders.

Regarding intervention providers, 7 studies explicitly reported that physicians and/or nurses delivered interventions via mHealth platforms [20,21,37,39,40,42,44], while the remaining 10 did not specify the type of health care professional [19,22,32-36,38,41,43]. The intervention duration ranged from 3 weeks to 9 months, with only 3 studies reporting frequency (daily to 3 times weekly) [20,33,40]. Regarding theoretical frameworks, only 1 study designed interventions based on social cognitive theory [32].

All studies assigned control groups to receive usual care. Twelve studies described usual care components, including health education, medication guidance, and regular follow-up [19-21,32-40], while the remaining 5 did not provide specific details [22,41-44]. Detailed study characteristics are summarized in Table 1.

XSL•FO
RenderX

Table . Characteristics of the included studies.

Author (year), country	Study design	Theory	Study setting	Cancer type	Sample, n	Age (years)	Details of the experimental group (format, content, duration, and frequency)	Control group	Main outcomes	Outcome indicator
Spoelstra et al [32] (2016), United States	RCT[a]	Social cognitive theory	Home	Breast cancer and other cancers	69	Experimental: mean 60.1 (SD 10.1); Control: mean 59.9 (SD 11.2)	Format: text messaging service; Content: text message reminders to take patient cancer pills; Duration: 3 weeks; Frequency: NA[b]	Usual care	The difference in the score between the experimental and control groups was 0.29 for adherence, and the difference was not statistically significant.	Symptom burden: MDASI-C[c]; Medication adherence: pill count
Kim et al [33] (2018), Korea	RCT	—[d]	Hospital	Breast cancer	72	Experimental: 49.8; Control: 52.1	Format: app (mobile game); Content: education for preventing the side effects of anticancer drugs; Duration: 3 weeks; Frequency: 3 times a week, 30 minutes each time	Usual care	The mobile game group showed better drug adherence (MARS[e]; mean 7.6, SD 0.7 vs mean 6.5, SD 0.5; *P*<.001).	Medication adherence: MARS; Symptom burden: BREF[f]
Verweij et al [22] (2023), Netherlands	Cluster RCT	—	Hospital and home	Leukemia	86	Range: 18 - 65	Format: website; Content: provides accurate and easy-to-understand information about the disease, medication, side effects, and the effect on daily life, and optimizes medication compliance; Duration: 6 months; Frequency: NA	Usual care	No statistical difference was found in the mean total score of patient medication compliance between groups (*P*=.922).	Health literacy: eHEALS[g]; Medication adherence: MARS; Symptom burden: EORTC QLQ-CML24[h]

XSL•FO
RenderX

Author (year), country	Study design	Theory	Study setting	Cancer type	Sample, n	Age (years)	Details of the experimental group (format, content, duration, and frequency)	Control group	Main outcomes	Outcome indicator
Greer et al [19] (2020), United States	RCT	—	Hospital and home	Hematologic cancer, lung cancer, breast cancer, glioma, sarcoma, gastric cancer, genital cancers, and melanoma	166	Range: 21-88; Experimental: mean 52.85 (SD 13.74); Control: mean 53.76 (SD 12.08)	Format: app; Content: a personalized medication dosing schedule, an adherence and symptom reporting module, educational resources for symptom management and other cancer-related topics, and Fitbit integration for tracking physical activity; Duration: 12 weeks; Frequency: NA	Usual care	Patient-reported adherence (MMAS-4[i]), symptom burden, quality of life, satisfaction with treatment, and health care utilization did not differ significantly between groups.	Medication adherence: MMAS-4; Symptom burden: MDASI-C; Service satisfaction: FACIT-TS-PS[j]
Wang et al [21] (2020), China	RCT	—	Home	Abdominal aortic aneurysm	82	Range: 36-71; Experimental: mean 51.83 (SD 4.56); Control: mean 52.53 (SD 4.23)	Format: website/app; Content: health monitoring, entering, and querying; Duration: 6 months; Frequency: NA	Usual care	The medication compliance of the observation group was significantly higher than that of the control group ($P<.05$).	Medication adherence: self-made medication adherence questionnaire; Service satisfaction: self-made follow-up satisfaction survey questionnaire

XSL•FO
RenderX

Author (year), country	Study design	Theory	Study setting	Cancer type	Sample, n	Age (years)	Details of the experimental group (format, content, duration, and frequency)	Control group	Main outcomes	Outcome indicator
Karaaslan-Eşer et al [34] (2021), Turkey	Cluster RCT	—	Hospital and home	Gastrointestinal cancers, lung cancer, breast cancer, and renal cancer	84	Experimental: mean 60.33 (SD 9.31); Control: mean 62.14 (SD 9.97)	Format: app; Content: providing information about oral anticancer agents and symptom-specific recommendations; Duration: 6 months; Frequency: NA	Usual care	The mean score of the intervention group increased over the first-, third-, and sixth-month measurements ($P<.05$).	Medication adherence: OCAS[k]; Symptom burden: MSAS[l]
Kekäl et al [35] (2016), Finland	RCT	—	Hospital	Leukemia	68	Range: 25 - 83	Format: text messaging service; Content: patients were offered daily text messages to their mobile phones based on the dosing regimen of their treatment to help them remember the time to take their medication; Duration: 9 months; Frequency: NA	Usual care	The groups were similar in adherence as assessed by MMAS class ($P>.05$).	Medication adherence: MMAS
Wang et al [36] (2022), China	RCT	—	Hospital and home	Thyroid cancer	98	Experimental: mean 48.2 (SD 8.0); Control: mean 46.6 (SD 9.4)	Format: app; Content: send medication reminders; researchers monitor patients' follow-up status in the background and provide timely reminders; Duration: 60 days; Frequency: NA	Usual care	The medication compliance of the study group was significantly better than that of the control group ($P<.01$).	Medication adherence: MMAS-8; Health literacy: Medication Belief Questionnaire

XSL•FO
RenderX

Author (year), country	Study design	Theory	Study setting	Cancer type	Sample, n	Age (years)	Details of the experimental group (format, content, duration, and frequency)	Control group	Main outcomes	Outcome indicator
Che et al [37] (2022), China	RCT	—	Hospital and home	Pheochromocytoma	76	Range: 28-65; Experimental: mean 42.65 (SD 3.79); Control: mean 43.26 (SD 3.92)	Format: app; Content: includes a disease knowledge module, medical interaction module, cutting-edge information module, and appointment follow-up module; Duration: 4 weeks; Frequency: NA	Usual care	The medication adherence score was higher in the intervention group than in the control group (P<.05).	Medication adherence: MMAS-8
He et al [20] (2023), China	RCT	—	Hospital and home	Prostate cancer	80	Range: 61-87; Experimental: mean 74.95 (SD 6.78); Control: mean 72.05 (SD 6.56)	Format: text messaging service; Content: at 7:00 every morning, a greeting message and medication reminder are sent to the WeChat group; Duration: 6 months; Frequency: once a day	Usual care	At 1, 3, and 6 months, medication adherence was higher in the education group than in the control group, with statistically significant differences (P<.05).	Medication adherence: MMAS-8
Başoğlu et al [38] (2024), Turkey	RCT	—	Home	Colorectal cancer, breast cancer, and others	48	Experimental: mean 72.00 (SD 7.86); Control: mean 70.52 (SD 5.32)	Format: text messaging service; Content: send patients 1 educational video and 4 text messages per week via WhatsApp; Duration: 4 weeks; Frequency: NA	Usual care	There was no significant difference between the mean scores of the intervention and control groups regarding the pretest application of the OCAS (P>.05).	Medication adherence: OCAS

XSL•FO
RenderX

Author (year), country	Study design	Theory	Study setting	Cancer type	Sample, n	Age (years)	Details of the experimental group (format, content, duration, and frequency)	Control group	Main outcomes	Outcome indicator
Graetz et al [39] (2024), United States	Cluster RCT	—	Hospital and home	Breast cancer	28	Range: 33-82; Experimental: mean 58.1 (SD 10.1); Control: mean 60.2 (SD 12.3)	Format: smart pill bottle + text messaging service; Content: text message reminders for missed or incorrect doses; smart pill bottle provides real-time medication adherence monitoring; Duration: 90 days; Frequency: NA	Usual care	Among participants who received the intervention, the average adherence rate was 79.1% for usual care participants and 89.0% for PRO[m] participants (P=.05).	Medication adherence: Nomi smart pill bottle system; Symptom burden: self-made symptom burden questionnaire
Dang et al [40] (2025), Australia	RCT	—	Home	Leukemia	28	Range: 18 - 65	Format: app; Content: medication-taking schedule, individualized daily medication reminders, and weekly side effect surveys; Duration: 12 weeks; Frequency: once a day	Usual care	There was no difference in the proportion of patients who had optimal adherence (medication rate of adherence ≥90%) between the intervention and control groups (100% for both groups).	Medication adherence: pharmacy refills; Symptom burden: FACT-G[n]; Self-efficacy: PRO

XSL•FO
RenderX

Author (year), country	Study design	Theory	Study setting	Cancer type	Sample, n	Age (years)	Details of the experimental group (format, content, duration, and frequency)	Control group	Main outcomes	Outcome indicator
Hosseinpour et al [41] (2025), Iran	RCT	—	Hospital	Leukemia	70	Range: 10 - 21	Format: app; Content: offers features such as updates (eg, adjusting the volume of alerts, selecting alert types, setting alarms based on dates, and adding games to encourage adolescent engagement); Duration: 8 weeks; Frequency: NA	Usual care	The adherence score in the intervention group differed significantly from that in the control group (P<.05).	Medication adherence: MARS
Lin et al [42] (2018), China	RCT	—	Home	Colorectal cancer, lung cancer, liver cancer, and others	97	Range: 29 - 74	Format: app; Content: the app provides medication instructions, including information on usage methods, dosage, and side effects; Duration: 3 months; Frequency: NA	Usual care	Compared with the control group, the observation group demonstrated significantly higher compliance (P<.05).	Medication adherence: self-made medication adherence questionnaire; Self-efficacy: CPSS[o]

XSL•FO
RenderX

Author (year), country	Study design	Theory	Study setting	Cancer type	Sample, n	Age (years)	Details of the experimental group (format, content, duration, and frequency)	Control group	Main outcomes	Outcome indicator
Park et al [43] (2022), Korea	RCT	—	Home	Breast cancer	57	Experimental: mean 52.07 (SD 9.34); Control: mean 54.74 (SD 7.87)	Format: app + smart pill bottle; Content: participants were reminded to take their medication as per the beeping and blinking function of the smart pill bottle cap, which indicated the scheduled time; they also received smartphone notifications that the time of taking the medication was recorded automatically; Duration: 4 weeks; Frequency: NA	Usual care	Mean medication adherence rates were higher in the experimental group than in the control group (P=.004).	Medication adherence: Pillsy mobile app; Self-efficacy: medication self-efficacy scale

XSL•FO
RenderX

Author (year), country	Study design	Theory	Study setting	Cancer type	Sample, n	Age (years)	Details of the experimental group (format, content, duration, and frequency)	Control group	Main outcomes	Outcome indicator
Zhao et al [44] (2023), China	RCT	—	Home	NA	100	Experimental: mean 60.67 (SD 6.59); Control: mean 60.82 (SD 6.76)	Format: app; Content: remind patients to take their medication on time via mobile apps or text messages, and provide guidance on dosage, frequency, and other aspects to ensure patients take their medication accurately and on schedule; Duration: 3 months; Frequency: NA	Usual care	Following the intervention, the Morisky medication adherence scores of patients in the experimental group were higher than the scores of patients in the control group, with a statistically significant difference (P<.05).	Self-efficacy: SCS[p]; Medication adherence: MMAS-8

[a]RCT: randomized controlled trial.

[b]NA: not applicable.

[c]MDASI-C: M.D. Anderson Symptom Inventory.

[d]Not reported.

[e]MARS: Medication Adherence Rating Scale.

[f]BREF: The World Health Organization Quality of Life-BREF Scale.

[g]eHEALS: eHealth Literacy Scale.

[h]EORTC QLQ-CML24: European Organization for Research and Treatment of Cancer Quality of Life - Chronic Myeloid Leukemia 24-item.

[i]MMAS: Morisky Medication Adherence Scale.

[j]FACIT-TS-PS: Functional Assessment of Chronic Illness Treatment-Treatment Satisfaction-Patient Satisfaction.

[k]OCAS: Oral Chemotherapy Adherence Scale.

[l]MSAS: Memorial Symptom Assessment Scale.

[m]PRO: patient-reported outcomes.

[n]FACT-G: Functional Assessment of Cancer Therapy-General.

[o]CPSS: Chronic Pain Self-Efficacy Scale.

[p]SCS: Self-Control Scale.

Risk of Bias in Studies

Ten studies reported randomization methods, while 5 described allocation concealment methods. Since improving medication adherence among patients with cancer using mHealth technology constitutes a behavioral intervention, blinding participants and researchers proved challenging. Consequently, none of the studies used blinding, resulting in a high risk of performance bias. Furthermore, as outcome measures were primarily assessed using subjective scales, the risk of measurement bias remains unclear. Only 2 studies used intention-to-treat analysis to address missing outcome data. No studies demonstrated selective reporting bias or other types of bias, resulting in an overall quality rating of grade B (Figure 2).

XSL•FO
RenderX

Figure 2. Risk of bias. (A) Risk of bias graph; (B) Risk of bias summary [19-22,32-44].

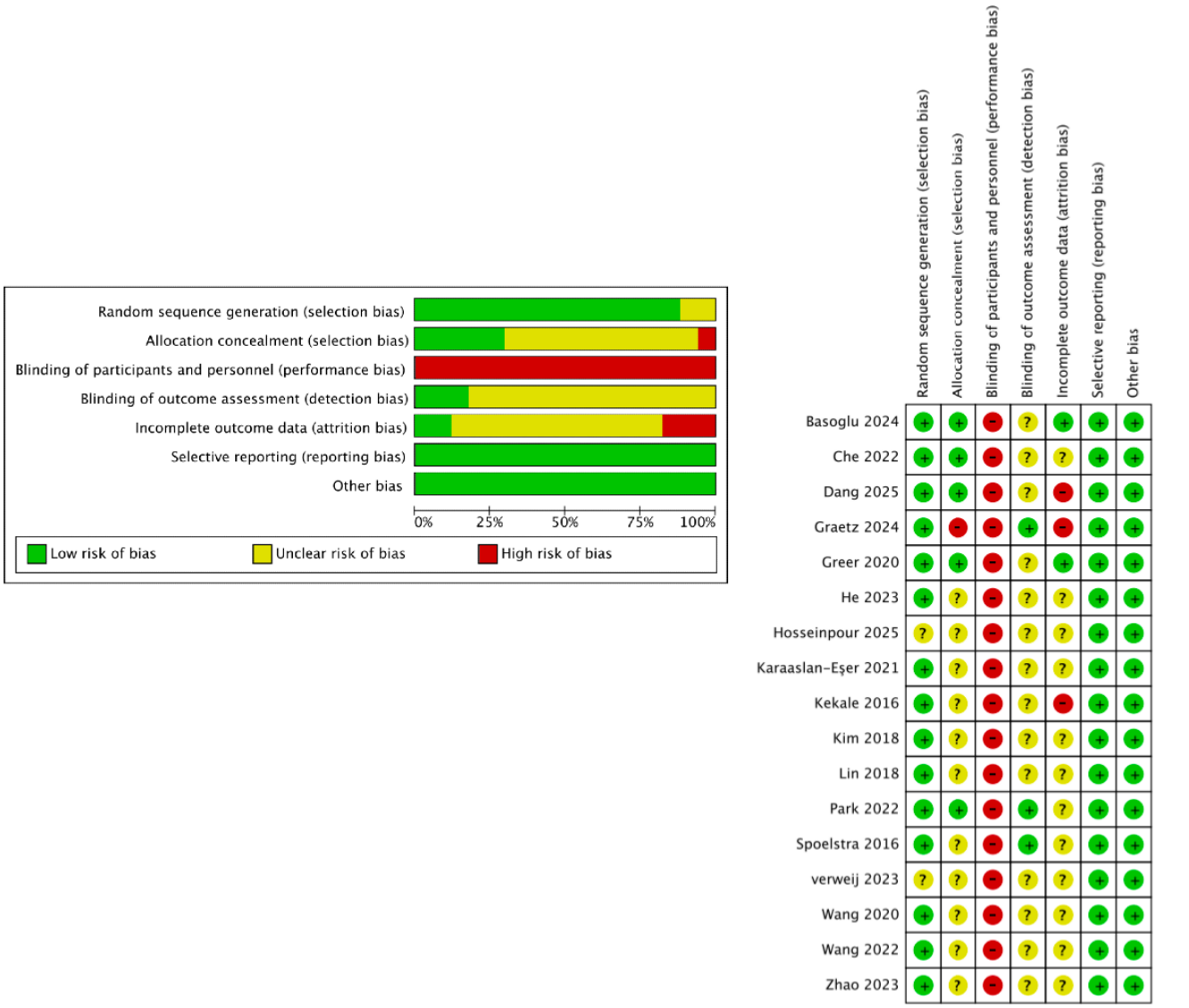

Evidence Quality Assessment

This systematic review used the GRADE evidence rating tool to assess the quality of evidence for each outcome measure. Overall, evidence quality ranged from moderate to very low. For the impact of mHealth interventions on medication adherence rates, the evidence quality was moderate, downgraded primarily due to risk of bias and imprecise data. For the effect of mHealth on the mean medication adherence score, the evidence quality was very low, downgraded due to high risk of bias, inconsistent results, and imprecise data. For the effect of mHealth on symptom burden, the evidence quality was low, downgraded due to risk of bias and imprecise data.

Regarding the effect of mHealth on self-efficacy, the evidence quality was also low, downgraded due to risk of bias and imprecision. The GRADE evidence summaries for each outcome are presented in Table 2, with detailed downgrade justifications available in Multimedia Appendix 2.

XSL•FO
RenderX

Table . GRADE (Grading of Recommendations Assessment, Development, and Evaluation) evidence.

Outcomes	Studies (design), n	Risk of bias	Inconsistency	Indirectness	Imprecision	Publication bias	Patients[a]		Effect		Certainty
							Mobile health	Usual care	Relative, OR[b] (95% CI)	Absolute, value (95% CI)	
Medication adherence rate	7 (RCT[c])	Serious	Not serious	Not serious	Serious	Undetected	269/313 (85.9)	190/295 (64.4)	3.47 (1.92-6.26)	219 more per 1000 (from 132 more to 275 more)	Moderate
Mean medication adherence scores	10 (RCT)	Serious	Serious	Not serious	Serious	Undetected	361	340	—[d]	SMD[e] 1.01 (0.51 to 1.52)	Very low
Symptom burden	7 (RCT)	Serious	Not serious	Not serious	Serious	Undetected	267	234	—	SMD −0.38 (−0.61 to −0.14)	Low
Self-efficacy	6 (RCT)	Serious	Not serious	Not serious	Serious	Undetected	214	191	—	SMD 0.98 (0.25 to 1.72)	Low

[a]Data are presented as n/N (%) or n.

[b]OR: odds ratio.

[c]RCT: randomized controlled trial.

[d]Not applicable.

[e]SMD: standardized mean difference.

Meta-Analysis Results

Effect of mHealth on the Medication Adherence Rate

Seven studies evaluated the impact of mHealth on the medication adherence rate [19,21,32,35,36,39,42]. Despite low heterogeneity (I^2=18.3%; τ^2=0.0658), we used a random-effects model to provide a more conservative and generalizable effect estimate. This decision was based on differences in adherence measures, intervention content, and measurement tools across the included studies. Pooled results demonstrated that mHealth significantly improved the medication adherence rate compared to controls (OR 3.47, 95% CI 1.92-6.26, 95% PI 1.46-8.21; I^2=18.3%; τ^2=0.0658; *P*=.002; Figure 3). The 95% CI confirmed the statistical significance of the average effect, with low heterogeneity among studies (I^2=18.3%; τ^2=0.0658). Furthermore, the 95% PI was entirely above 1, indicating that the benefit of mHealth on the medication adherence rate is likely reproducible in future studies. Note that the small number of included studies may render this 95% CI unreliable. The quality of evidence regarding mHealth's impact on the medication adherence rate was moderate (Table 2 and Multimedia Appendix 2). Despite downgrading due to risk of bias and imprecision, mHealth still demonstrated a significant effect on the medication adherence rate.

Figure 3. Forest plot of the effect of mobile health on the medication adherence rate [19,21,32,35,36,39,42]. OR: odds ratio.

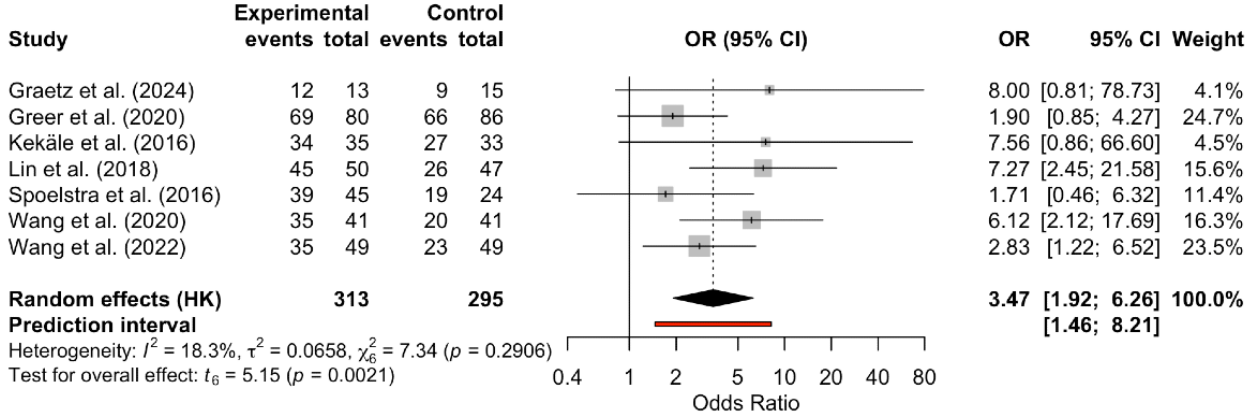

XSL•FO
RenderX

Effect of mHealth on Mean Medication Adherence Scores

Ten studies evaluated the impact of mHealth on mean medication adherence scores [20,22,33,34,37,38,40,41,43,44]. Given the high heterogeneity in the analyses (I^2=83.2%; τ^2=0.3975), we used a random-effects model for pooling to provide a more conservative and generally applicable effect estimate. The pooled analysis revealed that mHealth interventions significantly improved patients' mean medication adherence scores (SMD 1.01, 95% CI 0.51-1.52; I^2=83.2%; τ^2=0.3975; *P*=.001; Figure 4). However, the PI (95% PI −0.50 to 2.52) for this effect estimate was wide and included zero, suggesting that the true effect may vary considerably around zero in future studies. This implies that mHealth may not yield significant effects for certain populations or in different contexts.

Figure 4. Forest plot of the effect of mobile health on mean medication adherence scores [20,22,33,34,37,38,40,41,43,44]. SMD: standardized mean difference.

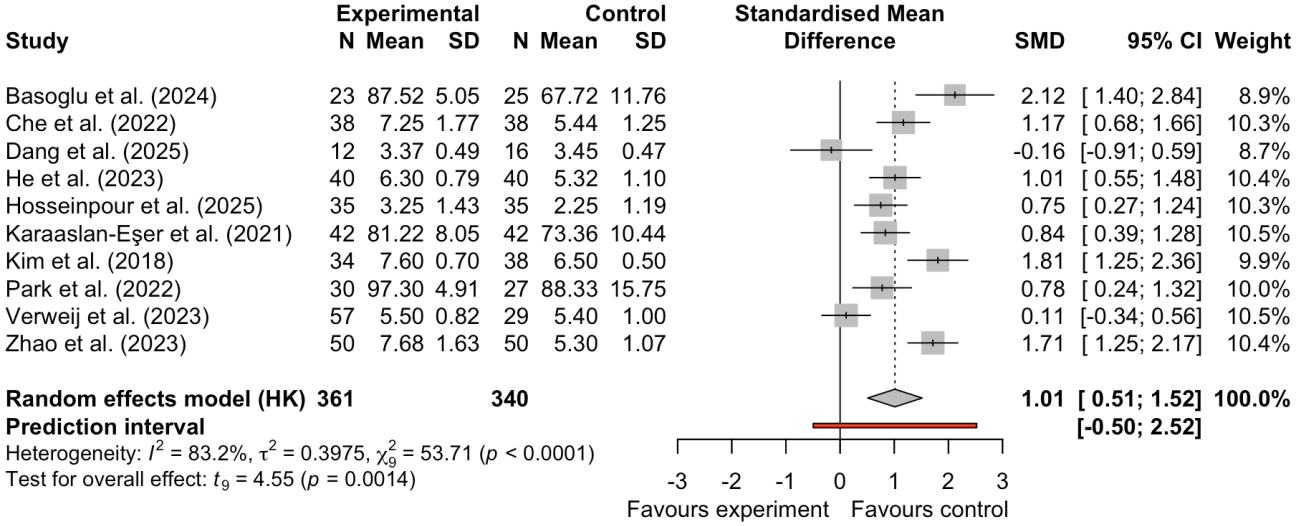

This result should be interpreted with caution. First, the extremely high heterogeneity (I^2=83.2%; τ^2=0.3975) indicates substantial differences in effect sizes across studies, potentially attributable to variations in intervention protocols, study populations, or measurement tools. Second, although the pooled effect was statistically significant, the quality of evidence for this outcome was rated as very low according to the GRADE criteria (Table 2 and Multimedia Appendix 2), primarily due to multiple downgrades for risk of bias, inconsistency, and imprecision. Therefore, while this analysis suggests that mHealth positively impacts average medication adherence scores, the certainty of this conclusion is constrained by the quality and heterogeneity of the existing evidence.

Sensitivity analyses by excluding individual studies showed that excluding the study by Dang et al [40] did not substantially alter the intervention effect on average medication adherence scores (SMD 1.12, 95% CI 0.64-1.60, 95% PI −0.24 to 2.49; I^2=81.9%; τ^2=0.3086; *P*<.001; see Figure S1 in Multimedia Appendix 3), further confirming the robustness of the original meta-analysis results.

To explore the effects of intervention duration, intervention type, and participant characteristics on mean medication adherence scores, we conducted exploratory subgroup analyses. Regarding intervention duration, the subgroup with duration <3 months (SMD 1.37, 95% CI 0.78-1.96; *P*=.001) showed a significantly greater effect than the subgroup with duration ≥3 months (SMD 0.49, 95% CI −0.39 to 1.37; *P*=.006), with a statistically significant between-group difference (χ^2_1=5.98; *P*=.01). Regarding intervention type, significant differences in effect size were observed among text messaging services (SMD 1.53, 95% CI −5.49 to 8.55; *P*=.01), apps (SMD 1.01, 95% CI 0.42-1.61; *P*<.001), and websites (SMD 0.11, 95% CI −0.34 to 0.56), and there were significant differences in effectiveness (χ^2_2=10.28, *P*=.006), with text messaging services yielding the highest point estimate but exhibiting an extremely wide CI. Subgroup differences by cancer type were also statistically significant (χ^2_2=8.86; *P*=.01). The breast cancer subgroup yielded the highest point estimate (SMD 1.29, 95% CI −5.25 to 7.83; *P*=.009), followed by other cancer types (SMD 1.09, 95% CI 0.10-2.08; *P*=.65) and leukemia (SMD 0.28, 95% CI −0.87 to 1.42; *P*=.07). It is important to emphasize that all analyses were post hoc exploratory comparisons, were not adjusted for randomization, and may be influenced by confounding factors (eg, shorter interventions potentially involving more frequent clinician-patient interactions). The observed intergroup differences represent only suggestive correlations and cannot be interpreted as causal relationships. Furthermore, limited numbers of studies and small sample sizes within some subgroups may have resulted in insufficient statistical power and unstable results. Therefore, these findings should be regarded as observational evidence for generating hypotheses and require further validation in future studies.

Effect of mHealth on Symptom Burden

Seven studies evaluated the impact of mHealth technologies on symptom burden in patients with cancer [19,22,32-34,39,40]. Despite low heterogeneity (I^2=4.8%; τ^2=0.0089), we used a random-effects model to obtain a more conservative and generalizable effect estimate. This decision was based on the variability in intervention content and measurement tools across the included studies. Compared with controls, mHealth

XSL•FO
RenderX

interventions yielded modest effects on symptom burden improvement (SMD −0.38, 95% CI −0.61 to −0.14, 95% PI −0.71 to −0.04; I^2=4.8%; τ^2=0.0089; P=.008; Figure 5). The 95% CI confirmed the statistical significance of the pooled effect, with low study heterogeneity (I^2=4.8%). Furthermore, the 95% PI did not cross zero, indicating that the effect of mHealth on symptom burden improvement is likely reproducible in future studies. Note that the small number of included studies may render this 95% PI unreliable. The quality of evidence regarding mHealth's impact on symptom burden was rated as low (Table 2 and Multimedia Appendix 2). This downgrade resulted from risk of bias and imprecision, indicating a small effect of mHealth on symptom burden.

Figure 5. Forest plot of the effect of mobile health on symptom burden [19,22,32-34,39,40]. SMD: standardized mean difference.

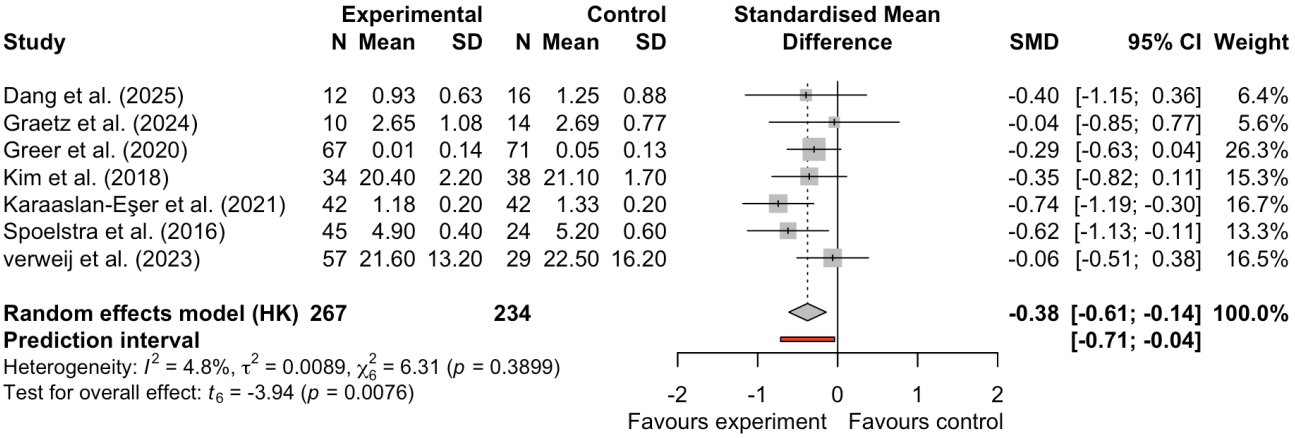

Effect of mHealth on Self-Efficacy

Six studies evaluated the impact of mHealth on self-efficacy among patients with cancer [32,38,40,42-44]. Due to high heterogeneity in this meta-analysis (I^2=75.9%; τ^2=0.2443), we used a random-effects model to obtain a more conservative and generalizable effect estimate. Pooled results demonstrated that mHealth significantly enhanced patient self-efficacy compared with controls (SMD 0.90, 95% CI 0.29-1.51, 95% PI −0.50 to 2.30; I^2=75.9%; τ^2=0.2443; P=.01; Figure 6). The 95% CI confirmed the statistical significance of the average effect, but the 95% PI spanned zero, indicating that mHealth may not improve self-efficacy in future studies or different populations. Note that the small number of included studies may render this 95% PI unreliable. The evidence quality for self-efficacy was low (Table 2 and Multimedia Appendix 2). The evidence rating was downgraded due to risk of bias and imprecision. mHealth still significantly influenced self-efficacy. Sensitivity analysis by excluding individual studies showed that removing the study by Dang et al [40] did not substantially alter the effect of mHealth interventions on mean self-efficacy scores (SMD 0.98, 95% CI 0.25-1.72, 95% PI −0.63 to 2.59; I^2=81.9%; τ^2=0.268; P=.02; Figure S5 in Multimedia Appendix 3), further confirming the robust stability of the original meta-analysis results (Figure 6).

Figure 6. Forest plot of the effect of mobile health on self-efficacy [32,38,40,42-44]. SMD: standardized mean difference.

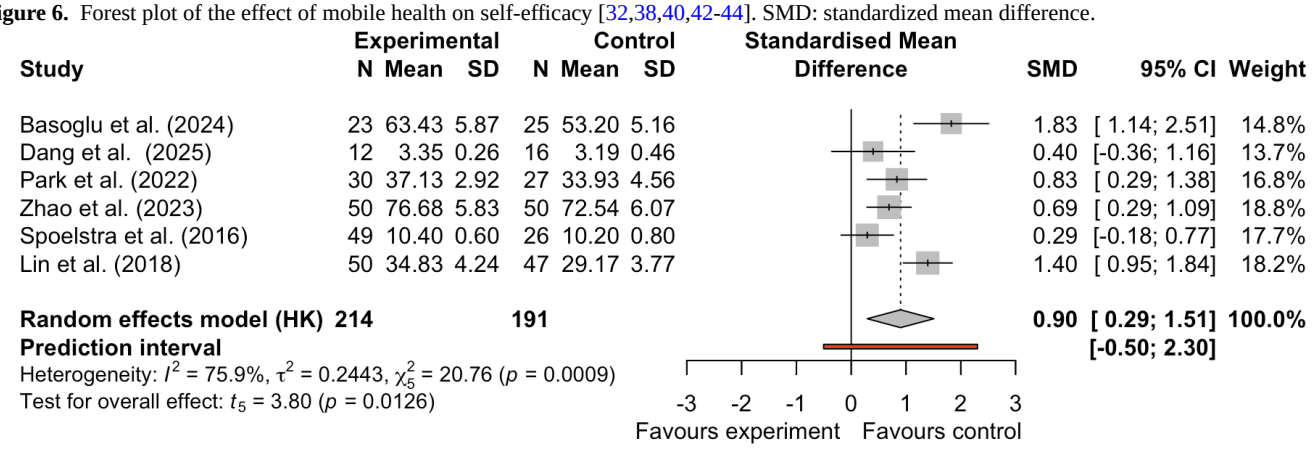

Effect of mHealth on Service Satisfaction

Two studies evaluated the impact of mHealth on service satisfaction, but neither provided the mean and SD data required for a meta-analysis of service satisfaction. After multiple attempts to contact the corresponding authors of these studies, the requested data could not be obtained. Therefore, we conducted a descriptive analysis of the available data [19,21]. Wang et al [21] reported that during the 2-month follow-up period after the intervention, patients were surveyed using a satisfaction questionnaire developed by the hospital's mHealth platform. The results showed significantly higher satisfaction scores in the intervention group compared with the control group. In contrast, Greer et al [19] used the validated Functional Assessment of Chronic Illness Treatment-Therapy Satisfaction-Patient Satisfaction (FACIT-TS-PS) scale to assess

XSL•FO
RenderX

patient satisfaction, finding no statistically significant difference between intervention and control group scores.

Effect of mHealth on Health Literacy

Two studies assessed health literacy levels [22,36]. Despite low heterogeneity (I^2=5.5%; τ^2=0.0028), given the differing measurement tools used across studies (Verweij et al [22] assessed patient health literacy using the validated eHealth Literacy Scale; Wang et al [36] used the validated Medication Beliefs Questionnaire), we used a random-effects model to obtain a more conservative and generalizable effect estimate. The meta-analysis revealed no statistically significant effect of interventions on health literacy (SMD 0.51, 95% CI −1.50 to 2.52; I^2=5.5%; τ^2=0.0028; *P*=.29; Figure 7). However, this result must be interpreted with caution for the following primary reasons. First, the inclusion of only 2 studies severely limited the statistical power of the analysis and the reliability of its conclusions. Second, the PI (95% PI −1.61 to 2.63) and heterogeneity parameter (τ^2=0.0028) calculated from this limited number of studies were highly unstable, rendering them incapable of providing meaningful estimates for future study effects or true heterogeneity.

Figure 7. Forest plot of the effect of mobile health on health literacy [22,36]. SMD: standardized mean difference.

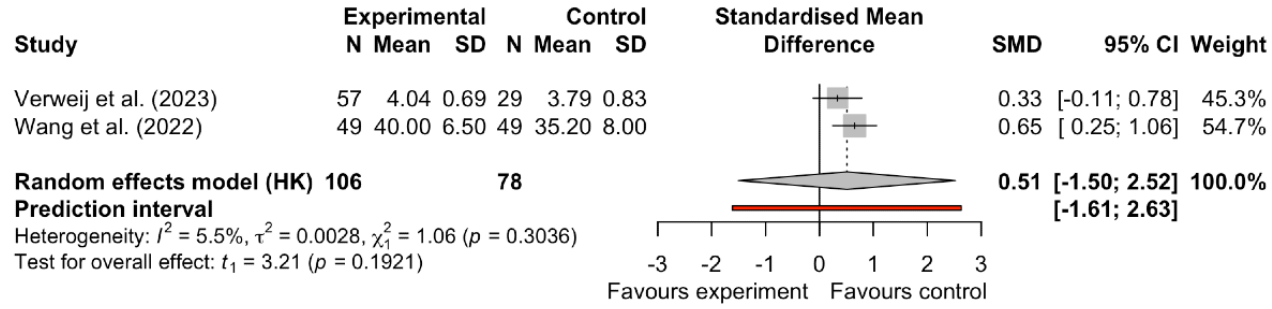

Small Study Effects Assessment

To assess the small-sample effect, we followed the methodological guidance by Sterne et al [30], combining visual inspection of funnel plots with the Egger regression intercept test to analyze 4 outcome measures: mean medication adherence scores, medication adherence, symptom burden, and self-efficacy.

It should be noted that, except for the mean medication adherence score, fewer than 10 studies were included for each of the other outcome measures, which is below the ideal threshold for reliable assessment proposed by Lau et al [45]. Nevertheless, funnel plots were constructed for each measure (Figures S6-S9 in Multimedia Appendix 3) for preliminary exploration. Visual inspection revealed no obvious asymmetry in the funnel plots; however, this visual assessment itself carries considerable uncertainty due to the limited number of studies.

Subsequent quantitative analysis using the Egger test also failed to indicate statistically significant small-sample effects (medication adherence average score: z=1.64; *P*=.75; symptom burden: z=0.18; *P*=.91; self-efficacy: z=1.29; *P*=.78; and medication adherence rate: z=1.53; *P*=.26).

However, these findings must be interpreted with caution. The primary limitation lies in the insufficient number of included studies. First, a small sample size significantly reduces the statistical power of the Egger test, increasing the risk of type II error (ie, failing to detect a genuinely existing small-sample effect) [30]. Second, neither the symmetry of the funnel plot nor the nonsignificance of the Egger test provides sufficient evidence to rule out publication bias [45]. Furthermore, funnel plot asymmetry itself may stem from clinical heterogeneity or differences in methodological quality rather than specifically indicating publication bias. Therefore, based on the current limited data, the findings of this systematic review cannot rule out the possibility of potential publication bias or other forms of selective reporting bias.

Discussion

Summary of Research Findings

This systematic review aimed to systematically evaluate the impact of mHealth interventions on medication adherence among patients with cancer. The results indicate that mHealth interventions can improve medication adherence rates and mean adherence scores among patients with cancer. They also demonstrate positive effects in reducing symptom burden, enhancing self-efficacy, and increasing service satisfaction. However, mHealth showed no significant impact on health literacy. Subgroup analysis results indicate that the effectiveness of mHealth interventions is influenced by intervention timing, intervention type, and cancer type. Furthermore, variations in the intervention protocols and assessment tools used across the included studies may have impacted the reliability of the findings.

mHealth Effectively Improves Medication Adherence

Regarding the medication adherence rate, the 95% PI following the mHealth intervention ranged from 1.46 to 8.21, indicating that the benefits of mHealth for the medication adherence rate are reproducible in future studies. In terms of the mean medication adherence score, the mHealth intervention demonstrated a significant effect, consistent with previous research findings [46]. From the perspective of specific effects and real-world effect variability, the 95% PI for the mean medication adherence score ranged from −0.50 to 2.52, encompassing zero. This indicates that while this study demonstrates a positive impact of mHealth interventions on the mean medication adherence score, the effect may not be significant in certain populations or different contexts.

XSL•FO
RenderX

Furthermore, the included studies utilized specific channels, such as mobile apps, websites, and text messaging services, to deliver knowledge and skills related to oral medication through diverse formats, including real-time online Q&A sessions, multiplayer online games, educational videos, and push notifications. These interventions also provided timely medication reminders. Both the content and methods effectively enhanced medication adherence among patients with cancer by addressing intrinsic knowledge and skill acquisition as well as extrinsic behavioral supervision. Based on this, it can be inferred that improved medication adherence contributes to optimizing the overall effectiveness of cancer treatment, such as reducing symptom burden and enhancing patient self-efficacy. Data analysis results also support this inference, showing that mHealth interventions not only significantly reduced patients' symptom burden but also improved their self-efficacy and service satisfaction. However, the impact of mHealth interventions on health literacy was not statistically significant (P=.29), contradicting previous research findings [47]. This discrepancy may stem from the inclusion of only 2 studies in the health literacy outcome analysis, severely limiting statistical power and reducing the reliability of conclusions. Finally, this systematic review used subgroup analyses to examine the effects of intervention format, timing, and cancer type on mHealth intervention outcomes.

Text Messaging Services Have the Best Intervention Outcomes

Mobile apps and websites offer health education information in diverse formats, including videos, audio, and text. Moreover, the information exchange between these platforms and patients with cancer—that is, between information providers and recipients—is often bidirectional and interactive. In contrast, text messaging services, which are somewhat outdated, provide information in a relatively monotonous format dominated by text, and the information transmission through this channel is predominantly one-way [48]. Theoretically, mobile apps and websites should yield superior intervention outcomes compared with text messages. However, our subgroup analysis revealed that among the 3 mHealth intervention formats, text messaging services demonstrated greater effectiveness than mobile apps and websites. The findings of this systematic review align with the research by Boima et al [49], which demonstrated that text messaging interventions were more effective than mobile apps and web-based interventions in reducing medication nonadherence events among patients with cardiovascular diseases.

The underlying reason for this discrepancy may be related to the age characteristics of patients with cancer. As shown in Table 1, the average age of patients included in this systematic review was relatively high, with 6 studies having an average age exceeding 55 years [20,32,34,38,39,44] and 2 studies having an average age exceeding 70 years [20,38]. Although text messaging services represent a simple, one-way communication method, their content is often concise, clear, and straightforward, making it easier to understand and implement. This characteristic reduces cognitive load during information reception and processing, potentially aligning better with the information format preferences and cognitive functional characteristics of older patients with cancer. Furthermore, compared with mobile apps and websites, text messaging services impose lower demands on mobile devices and network connectivity, enhancing their practical feasibility and applicability. These factors may explain why text messaging interventions outperformed mobile app and website interventions among patients with cancer, who are predominantly older adults.

This finding also raises the following questions: Have existing mHealth interventions delivered via mobile apps and websites sufficiently considered the cognitive characteristics and information preferences of older adults during their design and development? Could this oversight inadvertently contribute to and exacerbate the so-called digital divide?

The Optimal Intervention Duration for mHealth is Less Than 3 Months

This systematic review conducted subgroup analyses based on different intervention durations using 3 months as the cutoff point. The selection of this cutoff point was primarily based on 2 considerations. First, from the perspective of clinical treatment patterns, the duration of oral therapy for patients with cancer typically ranges from 12 to 24 weeks. Three months corresponds to the midpoint of the treatment course, marking a critical transition from the early to the later stages of therapy [50]. Second, prior studies indicate that medication adherence often declines around 3 months after initiation due to factors like drug side effects and forgetfulness [8]. Subgroup analysis revealed that mHealth interventions lasting <3 months significantly improved the mean medication adherence scores of patients with cancer compared with those lasting ≥3 months. The findings of this systematic review align with the study by Al-Arkee et al [51], which demonstrated that mHealth interventions lasting <3 months were more effective than those lasting ≥3 months in improving the average medication adherence scores of patients with cardiovascular diseases.

From a behavioral science perspective, the superior effectiveness of mHealth interventions lasting <3 months, compared with those lasting ≥3 months, may stem from the behavioral habituation cycle. Research by Lally et al [52] indicates that the median time required to establish a new behavior is 66 days. This suggests that the critical phase for behavioral acquisition occurs mainly within the first 2 to 3 months. During this window, mHealth interventions can effectively promote the formation and consolidation of medication-taking behaviors. They do so by providing timely reminders, feedback, and external frameworks. Once the target behavior is automated, continued external prompts may offer diminishing marginal benefits [53]. After habit formation, long-term interventions (≥3 months) typically offer few additional benefits. In some cases, they may even reduce effectiveness by reinforcing external dependence or causing intervention fatigue. In summary, mHealth interventions of <3 months match the natural process of habit formation. By maximizing gains during the most sensitive adaptation period, they significantly improve medication adherence.

XSL•FO
RenderX

mHealth Interventions Show the Most Effective Outcomes in Improving Medication Adherence Among Patients With Breast Cancer

This systematic review included subgroup analyses for different cancer types. Notably, the results indicate that mHealth interventions demonstrate greater efficacy in improving medication adherence among patients with breast cancer compared to patients with leukemia and those with other cancer types. However, no studies have specifically examined the impact of cancer type on the effectiveness of mHealth interventions in enhancing medication adherence. Therefore, future research should conduct in-depth analyses of the differential effects of such interventions across various cancer types.

Limitations

This systematic review has several limitations. First, it included only 17 studies, and the small sample size may lead to imprecise 95% CIs and biased estimates, thereby weakening the reliability of the results. Additionally, the limited number of studies resulted in insufficient statistical power, potentially overlooking potential small-sample effects. Although the existing evidence is insufficient to prove the existence of small-sample effects, we cannot completely rule out the influence of publication bias or selective reporting bias.

Second, among the 17 included studies, only the study by Spoelstra et al [32] used a theoretical model (social cognitive theory) to guide the design of the mHealth platform. Interventions targeting medication adherence are fundamentally interventions targeting medication-taking behavior. From a behavioral science perspective, it is necessary to leverage well-established health behavior theoretical models to explore and analyze the underlying logic and influencing factors of medication adherence among patients with cancer. This approach aims to uncover the core drivers behind medication-taking behavior, thereby enabling the development of more precise and effective medication adherence intervention strategies.

Furthermore, all 17 included studies used uniform interventions for all patients. No tailored interventions were designed or implemented based on individual patient characteristics. Future design of mHealth strategies for cancer medication adherence should incorporate methods, such as latent profile analysis, user profiling, and in-depth interviews, to thoroughly analyze patient traits and needs. Integrating patient profiles into platform design will enable personalized, precision-based scientific interventions [54,55].

Conclusions

Building upon previous systematic reviews, this systematic review systematically reviewed and comprehensively analyzed the impact of mHealth interventions on medication adherence among patients with cancer. Overall, existing evidence suggests that mHealth interventions hold potential advantages in improving medication adherence, self-efficacy, and service satisfaction, as well as reducing symptom burden. However, considerable heterogeneity exists across certain outcome measures, and the overall evidence quality remains moderate to low. Additionally, the included studies exhibited a moderate risk of bias, necessitating cautious interpretation of these findings. It is currently inconclusive to assert superiority over conventional care.

Despite these limitations, the significance of this systematic review lies not only in confirming the efficacy of mHealth interventions for improving medication adherence among patients with cancer but also in revealing their strategic value within the context of digital health care transformation. As chronic disease management shifts from hospital-centered to patient-centered models, mHealth care offers a viable pathway for achieving continuous care, personalized support, and low-cost, scalable interventions. Future research should enhance methodological rigor by conducting multicenter, large-sample, high-quality RCTs. These studies should explore the long-term efficacy and cost-effectiveness of mHealth care across diverse health care settings and patient populations to clarify its role and value within comprehensive cancer management systems.

Acknowledgments

We thank all contributors and reviewers for their insights.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this systematic review.

Data Availability

All data supporting the findings of this systematic review are included in the main text and its supplementary information. Further details or additional datasets can be obtained from the corresponding author upon reasonable request.

Authors' Contributions

WCL participated in conceptualization, data curation, formal analysis, investigation, methodology, writing – original draft, and writing – review and editing. LPW participated in conceptualization, supervision, and writing – review and editing. QYZ participated in formal analysis, investigation, methodology, visualization, and writing – original draft. JHY participated in methodology and writing – original draft.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File, 14 KB - jmir_v28i1e85949_app1.docx]

Multimedia Appendix 2
GRADE (Grading of Recommendations Assessment, Development, and Evaluation) evidence quality assessment.
[DOCX File, 15 KB - jmir_v28i1e85949_app2.docx]

Multimedia Appendix 3
Subgroup analysis, sensitivity analysis plots, and funnel plots for all results.
[DOCX File, 1597 KB - jmir_v28i1e85949_app3.docx]

Checklist 1
PRISMA checklist.
[PDF File, 507 KB - jmir_v28i1e85949_app4.pdf]

References

1. Lohi SM, Chindhalore CA, Dakhale G, Gajbhiye S. Cost variation and affordability of oral targeted cancer therapy in India: comparison of branded and janaushadhi products. Cureus 2025 Dec;17(12):e100287. [doi: 10.7759/cureus.100287] [Medline: 41613654]
2. Yan L, Xiaomeng W, Xiaoya Z, Xiaoyun H, Chunlan X. Summary of best evidence for prevention and management of oral chemotherapy drugs exposure in cancer patients at home [Article in Chinese]. Journal of Nursing Science 2025(4):92-96 [FREE Full text] [doi: 10.3870/j.issn.1001-4152.2025.04.092]
3. Yunyi C, Zifang J, Zefan H. Research progress on medication literacy of cancer patients receiving oral antineoplastic drugs [Article in Chinese]. Journal of Nursing Science 2024;39(14):126-129 [FREE Full text] [doi: 10.3870/j.issn.1001-4152.2024.14.126]
4. Cramer JA, Roy A, Burrell A, et al. Medication compliance and persistence: terminology and definitions. Value Health 2008;11(1):44-47. [doi: 10.1111/j.1524-4733.2007.00213.x] [Medline: 18237359]
5. Uyehara N, Nguyen V, Yu S, et al. The impact of patient characteristics and social drivers of health factors on oral oncolytic adherence. J Manag Care Spec Pharm 2025 Mar;31(3):306-320. [doi: 10.18553/jmcp.2025.31.3.306] [Medline: 40021464]
6. Jafari M, Shahverdian A, Sadigh G, Van Etten RA. Impact of patient personality on adherence to oral anticancer medications: an opportunity? JMIR Cancer 2024 Oct 30;10:e57199. [doi: 10.2196/57199] [Medline: 39475848]
7. van den Bemt P, Blijham MY, Ten Broek L, et al. Patient reported medication-related problems, adherence and waste of oral anticancer medication over time. Eur J Hosp Pharm 2025 Aug 21;32(5):421-426. [doi: 10.1136/ejhpharm-2024-004205] [Medline: 39019578]
8. Tongyu G, Yan S, Fang Z. Single-center evaluation of adherence to immunosuppressive drugs and assessment of influencing factors in 152 renal transplant recipients [Article in Chinese]. China Medicine and Pharmacy 2025;15(12):121-124 [FREE Full text] [doi: 10.20116/j.issn2095-0616.2025.12.25]
9. DiMatteo MR. Variations in patients' adherence to medical recommendations: a quantitative review of 50 years of research. Med Care 2004 Mar;42(3):200-209. [doi: 10.1097/01.mlr.0000114908.90348.f9] [Medline: 15076819]
10. Ashraf H, Warren E, Latner JP, et al. Effect of medication adherence and early follow-up on unplanned readmission post-myocardial infarction: quality improvement in a safety net hospital. Crit Pathw Cardiol 2021 Jun 1;20(2):71-74. [doi: 10.1097/HPC.0000000000000231] [Medline: 32657972]
11. Frezzetti D, Caridi V, Marra L, et al. The impact of inadequate exposure to epidermal growth factor receptor-tyrosine kinase inhibitors on the development of resistance in non-small-cell lung cancer cells. Int J Mol Sci 2024 Apr 29;25(9):4844. [doi: 10.3390/ijms25094844] [Medline: 38732063]
12. Xu Q, Xu Y, Liu X, Ma X. The application of mobile health in self-management among patients undergoing dialysis: scoping review. J Med Internet Res 2026 Jan 2;28:e76880. [doi: 10.2196/76880] [Medline: 41481914]
13. Fahrni ML, Saman KM, Alkhoshaiban AS, Naimat F, Ramzan F, Isa KAM. Patient-reported outcome measures to detect intentional, mixed, or unintentional non-adherence to medication: a systematic review. BMJ Open 2022 Sep 19;12(9):e057868. [doi: 10.1136/bmjopen-2021-057868] [Medline: 36123061]
14. Zhou TT, Wang R, Gu SJ, et al. Effectiveness of mobile medical apps in ensuring medication safety among patients with chronic diseases: systematic review and meta-analysis. JMIR Mhealth Uhealth 2022 Nov 22;10(11):e39819. [doi: 10.2196/39819] [Medline: 36413386]

XSL•FO
RenderX

15. Linder LA, Wu YP, Macpherson CF, et al. Oral medication adherence among adolescents and young adults with cancer before and following use of a smartphone-based medication reminder app. J Adolesc Young Adult Oncol 2019 Apr;8(2):122-130. [doi: 10.1089/jayao.2018.0072] [Medline: 30307778]
16. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
17. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
18. Xu X, Wang F, Liao S, Liu J, Xiao L. Tele-cognitive behavioral therapy for the treatment of diabetes-related distress in idividuals with diabetes mellitus: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2025 Dec 24;27:e80476. [doi: 10.2196/80476] [Medline: 41442672]
19. Greer JA, Jacobs JM, Pensak N, et al. Randomized trial of a smartphone mobile app to improve symptoms and adherence to oral therapy for cancer. J Natl Compr Canc Netw 2020 Feb;18(2):133-141. [doi: 10.6004/jnccn.2019.7354] [Medline: 32023526]
20. He Y, Wang Q, Yan S, Wang L, Fu R. Application of mobile health education in improving medication compliance of elderly patients with endocrine therapy for prostate cancer [Article in Chinese]. The Medical Forum 2023;27(24):12-15. [doi: 10.19435/j.1672-1721.2023.24.004]
21. Wang B, Wang H, Sun H, et al. The effects of application of hospital information platform of interactively medical treatment based on “Internet +” on medication compliance and quality of life during follow-up of patients with abdominal aortic aneurysm surgery [Article in Chinese]. China Medical Equipment 2020;17(9):35-39. [doi: 10.3969/J.ISSN.1672-8270.2020.09.008]
22. Verweij L, Ector G, Smit Y, et al. Effectiveness of digital care platform CMyLife for patients with chronic myeloid leukemia: results of a patient-preference trial. BMC Health Serv Res 2023 Mar 8;23(1):228. [doi: 10.1186/s12913-023-09153-9] [Medline: 36890512]
23. Higgins JPT, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration’s tool for assessing risk of bias in randomised trials. BMJ 2011 Oct 18;343(oct18 2):d5928. [doi: 10.1136/bmj.d5928] [Medline: 22008217]
24. Joanna Briggs Institute Levels of Evidence and Grades of Recommendation Working Party. Supporting document for the Joanna Briggs Institute levels of evidence and grades of recommendation. The Joanna Briggs Institute. 2014. URL: https://jbi.global/sites/default/files/2019-05/JBI%20Levels%20of%20Evidence%20Supporting%20Documents-v2.pdf [accessed 2026-03-23]
25. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
26. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
27. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019 Jun;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
28. Higgins JPT, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions, 2nd edition: John Wiley & Sons; 2019. [doi: 10.1002/9781119536604]
29. Fleiss JL. The statistical basis of meta-analysis. Stat Methods Med Res 1993;2(2):121-145. [doi: 10.1177/096228029300200202] [Medline: 8261254]
30. Sterne JAC, Sutton AJ, Ioannidis JPA, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011 Jul 22;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
31. Shi L, Lin L. The trim-and-fill method for publication bias: practical guidelines and recommendations based on a large database of meta-analyses. Medicine (Abingdon) 2019;98(23):e15987. [doi: 10.1097/MD.0000000000015987]
32. Spoelstra SL, Given CW, Sikorskii A, et al. Proof of concept of a mobile health short message service text message intervention that promotes adherence to oral anticancer agent medications: a randomized controlled trial. Telemed J E Health 2016 Jun;22(6):497-506. [doi: 10.1089/tmj.2015.0126] [Medline: 26716365]
33. Kim HJ, Kim SM, Shin H, Jang JS, Kim YI, Han DH. A mobile game for patients with breast cancer for chemotherapy self-management and quality-of-life improvement: randomized controlled trial. J Med Internet Res 2018 Oct 29;20(10):e273. [doi: 10.2196/jmir.9559] [Medline: 30578205]
34. Karaaslan-Eşer A, Ayaz-Alkaya S. The effect of a mobile application on treatment adherence and symptom management in patients using oral anticancer agents: a randomized controlled trial. Eur J Oncol Nurs 2021 Jun;52:101969. [doi: 10.1016/j.ejon.2021.101969] [Medline: 33991868]
35. Kekäle M, Söderlund T, Koskenvesa P, Talvensaari K, Airaksinen M. Impact of tailored patient education on adherence of patients with chronic myeloid leukaemia to tyrosine kinase inhibitors: a randomized multicentre intervention study. J Adv Nurs 2016 Sep;72(9):2196-2206. [doi: 10.1111/jan.12978] [Medline: 27113362]
36. Wang M, Su W, Zhang J, et al. Effects of mobile health intervention on endocrine therapy compliance andthyroid stimulating hormone compliance rate in postoperative patients withthyroid stimulating hormone compliance rate in postoperative

XSL•FO
RenderX

patients withthyroid cancerthyroid cancer [Article in Chinese]. Oncology Progress 2022;20(9):927-930. [doi: 10.11877/j.issn.1672-1535.2022.20.09.15]
37. Che H. The effect of a mobile health-based chronic disease management system on medication adherence, negative emotions, and blood pressure improvement in pheochromocytoma patients [Article in Chinese]. Chinese General Practice Nursing 2022;20(8):1114-1115. [doi: 10.12104/j.issn.1674-4748.2022.08.030]
38. Başoğlu S, Polat Ü. The effect of education and monitoring via tele-nursing to elderly cancer patients using oral anticancer agents on self-efficacy and medication adherence: a randomized controlled trial. Semin Oncol Nurs 2024 Oct;40(5):151692. [doi: 10.1016/j.soncn.2024.151692] [Medline: 39030135]
39. Graetz I, Hernandez S, Arshad S, et al. Leveraging mobile health to improve capecitabine adherence among women with breast cancer: a pilot randomized controlled trial. JCO Oncol Pract 2024 Oct;20(10):1376-1383. [doi: 10.1200/OP.24.00031] [Medline: 38917401]
40. Dang TH, Wickramasinghe N, Jayaraman PP, et al. Digital solution to support medication adherence and self-management in patients with cancer (SAMSON): pilot randomized controlled trial. JMIR Form Res 2025 Feb 19;9:e65302. [doi: 10.2196/65302] [Medline: 39969972]
41. Hosseinpour M, Shirinabadi Farahani A, Varzeshnejad M, Nasiri M. Check the impact of mobile health on medication adherence in adolescents with leukemia. Asian Pac J Cancer Prev 2025 Mar 1;26(3):751-755. [doi: 10.31557/APJCP.2025.26.3.751] [Medline: 40156390]
42. Lin H, Zheng Z, Chen Y, et al. Practical experience in implementing health education for cancer pain patients using the “317 Care” app in the context of “Internet Plus” [Article in Chinese]. Chinese Rural Health Service Administration 2018;38(6):771-773. [doi: 10.3969/j.issn.1005-5916.2018.06.024]
43. Park HR, Kang HS, Kim SH, Singh-Carlson S. Effect of a smart pill bottle reminder intervention on medication adherence, self-efficacy, and depression in breast cancer survivors. Cancer Nurs 2022;45(6):E874-E882. [doi: 10.1097/NCC.0000000000001030] [Medline: 34661562]
44. Zhao D, Gong Y, Ye Q, et al. The application effectiveness of a mobile healthcare-based chronic disease management system in cancer patient treatment [Article in Chinese]. Journal of Frontiers of Medicine 2023;13(21):74-76. [doi: 10.3969/j.issn.2095-1752.2023.21.023]
45. Lau J, Ioannidis JPA, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. BMJ 2006 Sep 16;333(7568):597-600. [doi: 10.1136/bmj.333.7568.597] [Medline: 16974018]
46. Psihogios AM, El-Khatib A, Matos K, et al. Human-centered design of a personalized digital health intervention to improve oral chemotherapy adherence in adolescents and young adults with hematologic cancers. Pediatr Blood Cancer 2025 Aug;72(8):e31756. [doi: 10.1002/pbc.31756] [Medline: 40346736]
47. Zisopoulou S, Kazantzaki G, Adamakidou T, et al. Ehealth literacy, beliefs about medication and adherence in older adults with chronic diseases: a cross-sectional study. Adv Exp Med Biol 2026;1489:187-196. [doi: 10.1007/978-3-032-03394-9_19] [Medline: 41252006]
48. Wiljer D, Catton P. Multimedia formats for patient education and health communication: does user preference matter? J Med Internet Res 2003;5(3):e19. [doi: 10.2196/jmir.5.3.e19] [Medline: 14517110]
49. Boima V, Doku A, Agyekum F, Tuglo LS, Agyemang C. Effectiveness of digital health interventions on blood pressure control, lifestyle behaviours and adherence to medication in patients with hypertension in low-income and middle-income countries: a systematic review and meta-analysis of randomised controlled trials. EClinicalMedicine 2024 Mar;69:102432. [doi: 10.1016/j.eclinm.2024.102432] [Medline: 38333367]
50. Chiavarini M, Giacchetta I, Rosignoli P, Fabiani R. E-health and m-health in obesity management: a systematic review and meta-analysis of RCTs. Nutrients 2025 Jul 1;17(13):2200. [doi: 10.3390/nu17132200] [Medline: 40647304]
51. Al-Arkee S, Mason J, Lane DA, et al. Mobile apps to improve medication adherence in cardiovascular disease: systematic review and meta-analysis. J Med Internet Res 2021 May 25;23(5):e24190. [doi: 10.2196/24190] [Medline: 34032583]
52. Lally P, van Jaarsveld CHM, Potts HWW, Wardle J. How are habits formed: modelling habit formation in the real world. Euro J Social Psych 2010 Oct;40(6):998-1009. [doi: 10.1002/ejsp.674]
53. Pongchaiyakul C, Driessen S. Evaluating the effect of mobile applications “My A:Care” and “Smart Coach” on adherence to lipid-lowering treatment in patients with dyslipidemia: a prospective, randomized, open-label clinical study. Front Digit Health 2025;7:1502990. [doi: 10.3389/fdgth.2025.1502990] [Medline: 40692655]
54. Amico KR, Mugavero M, Krousel-Wood MA, Bosworth HB, Merlin JS. Advantages to using social-behavioral models of medication adherence in research and practice. J Gen Intern Med 2018 Feb;33(2):207-215. [doi: 10.1007/s11606-017-4197-5] [Medline: 29204968]
55. Yuan Q, Yongfeng T, Dinghua Y, et al. Research progress on the application of “patient profile” based on big data in the field of nursing [Article in Chinese]. West China Medical Journal 2025;40(8):1351-1355. [doi: 10.7507/1002-0179.202410248]

Abbreviations:

GRADE: Grading of Recommendations Assessment, Development, and Evaluation

XSL•FO
RenderX

HKSJ: Hartung-Knapp-Sidik-Jonkman
mHealth: mobile health
OR: odds ratio
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SMD: standardized mean difference

Edited by S Brini; submitted 16.Oct.2025; peer-reviewed by Q Wang, SA Kristina; revised version received 26.Feb.2026; accepted 03.Mar.2026; published 30.Mar.2026.

Please cite as:
Liu WC, Wang LP, Zhang QY, Yang JH
Effectiveness of Mobile Health for Improving Medication Adherence in Patients With Cancer: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e85949
URL: https://www.jmir.org/2026/1/e85949
doi:10.2196/85949

© Wei Chao Liu, Li Ping Wang, Qian Yu Zhang, Jing Han Yang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Artificial Intelligence Tools for Automating Evidence Synthesis: Scoping Review

Sashika Harasgama[1], BSc, MScPH, MD; Helen Pearce[1], BA; Cameron Appel[1], BAc, MSc; Liam Loftus[1], BM BCh; Helena Painter[1], DTM&H, BSc, MBChB; Isla Kuhn[2], MA, MSc; Justine Karpusheff[3], PGCE, MSc, PhD; Aji Ceesay[3], BA, MSc; John Ford[1], DTM&H, MBChB, MSc, PhD

[1]Wolfson Institute of Population Health, Queen Mary University of London, Whitechapel Campus, London, United Kingdom

[2]Medical Library, University of Cambridge, Cambridge, United Kingdom

[3]The Health Foundation, London, United Kingdom

Corresponding Author:
Sashika Harasgama, BSc, MScPH, MD
Wolfson Institute of Population Health, Queen Mary University of London, Whitechapel Campus, London, United Kingdom

Abstract

Background: Rapidly and accurately synthesizing large volumes of evidence is a time- and resource-intensive process. Once published, reviews often risk becoming outdated, limiting their usefulness for decision makers. Recent advancements in artificial intelligence (AI) have enabled researchers to automate stages of the evidence synthesis process, from literature searching and screening to data extraction and analysis. As previous reviews on this topic have been published, a significant number of tools have been further developed and evaluated. Furthermore, as generative AI increasingly automates evidence synthesis, understanding how it is studied and applied is crucial, given both its benefits and risks.

Objective: This review aimed to map the current landscape of evaluated AI tools used to automate evidence synthesis.

Methods: Following the Joanna Briggs Institute methodology for scoping reviews, we searched Ovid MEDLINE, Ovid Embase, Scopus, and Web of Science in February 2025 and conducted a gray literature search in April 2025. We included articles published in any language from January 2021 onward. Two reviewers independently screened citations using Rayyan, and data were extracted based on study design and key AI-related technical features.

Results: We identified 7841 unique citations through database searches and 19 records through gray literature searching. A total of 222 articles were included in the review. We identified 65 AI tools and 25 open-source models or machine learning (ML) algorithms that automate parts of or the whole evidence synthesis pathway. A total of 54.1% (n=120) of the studies were published in 2024, reflecting a trend toward researching general-purpose large language models (LLMs) for evidence synthesis automation. The most popular tool studied was generative pretrained transformer models, including its conversational interface ChatGPT (n=70, 31.5%). Moreover, 31.1% (n=69) studied tools automated by traditional ML algorithms. No studies compared traditional ML tools to LLM-based tools. In addition, 61.7% (n=137) and 26.1% (n=58) studied AI-assisted automation of title and abstract screening and data extraction, respectively, the 2 most intensive stages and, therefore, amenable to automation. Technical performance outcomes were the most frequently reported, with only 4.1% (n=9) of studies reporting time- or workload-specific outcomes. Few studies pragmatically evaluated AI tools in real-world evidence synthesis settings.

Conclusions: This review comprehensively captures the broad, evolving suite of AI automation tools available to support evidence synthesis, leveraged by increasingly complex AI approaches that range from traditional ML to LLMs. The notable shift toward studying general-purpose generative AI tools reflects how these technologies are actively transforming evidence synthesis practice. The lack of studies in our review comparing different AI approaches for specific automation stages or evaluating their effectiveness pragmatically represents a significant research gap. Optimal tool selection will likely depend on the review topic and methodology and researcher priorities. While they offer potential for reducing workload, ongoing evaluation to mitigate AI bias and to ensure the integrity of reviews is essential for safeguarding evidence-based decision-making.

(J Med Internet Res 2026;28:e81597) doi:10.2196/81597

KEYWORDS

artificial intelligence; machine learning; large language models; automation; evidence synthesis; systematic reviews as a topic; ChatGPT

XSL•FO
RenderX

Introduction

High-quality evidence synthesis is essential for guiding policy and practice. However, producing such evidence reviews is not only time- and resource-intensive but also challenging to keep up to date due to the volume of studies being published each year. The recent COVID-19 pandemic highlighted the challenges of having accurate, contemporaneous, and rapidly synthesized data, and the opportunities that automation presents in mitigating these [1,2]. Previous studies have estimated that an average systematic review can take approximately 67 weeks to complete [3], often too long for decision makers. The evolving language and text capabilities of artificial intelligence (AI) have increased the scope of automation within the evidence review process, with studies showing that automated tools can help complete systematic reviews in days to weeks [4,5] and significantly reduce workload [6,7].

The advent of evidence synthesis automation was driven by machine learning (ML), the technology that underpins most AI evidence synthesis tools today. ML is a subset of AI that focuses on creating algorithmic systems that can learn from data, recognize patterns, and make decisions, often with model characteristics manually selected by humans [8]. An AI discipline called natural language processing enables computers to understand, interpret, and work with human language (such as written text) and, when used in conjunction with ML, can automate certain language-related tasks [8]. Deep learning is a subset of ML that uses neural networks (computational models that consist of layers of interconnected processing units) to automatically learn more complex patterns from large datasets [8].

Time- and workload-intensive steps within the evidence synthesis pathway, particularly title and abstract screening and data extraction, have been made more efficient through automation tools using machine and deep learning. However, the introduction of generative AI (ie, AI that can autonomously produce text, speech, or other outputs) in the form of commercially available large language models (LLMs), such as ChatGPT (OpenAI) or Claude (Anthropic), has created an opportunity for more creative and complex evidence synthesis automation. Studies are currently trialing LLM-based methods to automate individual stages, such as data extraction, to entire clinical evidence synthesis pipelines [9,10].

LLMs are leveraged by deep learning models called transformers, which use ‘attention’ mechanisms to determine how important each word is in relation to others. They are trained on massive amounts of written data and can capture nuance and context in language more effectively than traditional ML models. Another popular automation approach uses transformers called Bidirectional Encoder Representations from Transformers (BERT), which reads inputs bidirectionally and is pretrained on large sets of text data and then fine-tuned to excel at domain-specific tasks. These newer AI tools may outperform current automation tools in some tasks; however, researchers are still exploring their potential for evidence synthesis, and their performance is yet to be thoroughly validated over time [11,12]. In fact, the use of advanced and generative AI approaches is not without contention and legitimate risks, with the research community fearing loss of academic integrity in the process [11,13,14].

Reviews published to date on this topic have primarily focused on ML tools and approaches used for automating systematic reviews in particular [12,15-17]. For example, Khalil et al [15] performed a scoping review of automation tools for systematic reviews up to mid-2021 and found 10 validated tools that all used ML. Jimenez et al [16] undertook a mapping review of ML tools to assist with systematic reviews and identified 63 tools at the time of publication in December 2022. Khalil et al [18] also conducted a review in 2024 on automation tools for scoping reviews. While these reviews were comprehensive, the rate at which AI technology evolves often means they themselves risk being outdated. A more recent review published by Lieberum et al [12] focused solely on the use of LLMs for performing systematic reviews and found that in half the included studies, LLMs had promising applications, particularly in screening. However, the authors noted that despite the optimism, LLMs are still not quite ready for direct integration into research practice.

The rapid development and likely increasing adoption of generative AI tools to automate evidence synthesis and inform policy and practice warrant a more contemporaneous scoping of the literature. It presents significant risks to the quality and methodological rigor of evidence synthesis, despite its opportunities. Therefore, it is essential to understand the breadth and scope of AI use in this advancing field.

In this scoping review, we aimed to systematically map AI tools available for all types of evidence synthesis, across all its stages, and to describe the current landscape by exploring the underlying automation approaches used by the tools and identifying trends in automated evidence synthesis research.

Methods

Ethical Considerations

Ethical approval was not required for this scoping review, as it involved analysis of previously published studies and did not include primary data collection.

Review Design

We used the Joanna Briggs Institute methodological guidance for scoping reviews [19]. Using the “PCC” mnemonic (population, concept, and context), our review aimed to focus on current AI tools (*P*) used to automate (*C*) the evidence synthesis pathway (*C*). The reporting of this paper was guided by the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist [20], available in Checklist 1. Our review was not registered a priori in PROSPERO.

We aimed to identify studies across steps of the evidence synthesis pathway: (1) searching for the evidence, (2) screening the evidence, (3) extracting the data, (4) assessing the quality of the evidence, (5) analyzing the data, and (6) writing the review. We also aimed to identify which AI methods or techniques were most leveraged for tool development and to

XSL•FO
RenderX

determine which parts of the evidence synthesis pathway were the most automated. Our secondary objectives included discussing outcomes typically used to assess performance.

Search Strategy

The search strategy was adapted from 2 prior reviews' search strategies [15,16] and iteratively refined through an initial search of Ovid MEDLINE, based on analysis of titles, abstracts, and keywords returned as well as cross-referencing with key articles. The final search was executed across four databases: Ovid MEDLINE, Ovid Embase, Scopus, and Web of Science. We searched for studies published from January 1, 2021, to February 14, 2025, with no language restrictions. Details of our search strategy are available in Part S1 in Multimedia Appendix 1. Our search strategy combined three groups of search terms: (1) evidence review and systematic review terms; (2) AI methods, including ML, deep learning, and other techniques; and (3) terms reflecting the act of automation, such as "support," "assist," and "perform." We also conducted a gray literature search in April 2025 after data extraction of our included studies using Google's search engine and citation snowballing methods with the tool Litmaps [21] to identify further relevant studies from our included studies.

Eligibility Criteria

Full details of the eligibility criteria are displayed in Textbox 1. All primary and secondary studies that evaluated the use of AI tools for evidence synthesis in health and care research were included. We did not exclude studies if they did not pertain to health and care research specifically; however, we excluded studies if the AI tool was deemed not transferable to that setting. Studies that did not provide sufficient evaluation of the tool, through technical or other performance-related metrics, were also excluded.

Textbox 1. Eligibility criteria using the population, intervention, comparison, and outcome (PICO) framework.

Inclusion criteria

- Population and setting
  - Primary or secondary studies that evaluate the use of artificial intelligence (AI) tools for evidence synthesis in health and care research
  - Tool developed in any country
- Intervention
  - AI tools defined as having two characteristics: adaptivity and autonomy (based on the Department for Science, Innovation and Technology guidance)
- Comparison
  - If comparison, to manual research methods
  - If comparison, to other automated tools
- Outcome
  - Any technical or performance-related outcome assessing the AI tool
  - Would be dependent on evidence synthesis task but could include outcomes such as
    - Sensitivity, specificity, precision, and area under the curve
    - Number of relevant studies identified in screening
    - Quality of data extraction
    - Accuracy of meta-analysis
    - Quality and accuracy of writing of evidence synthesis report

Exclusion criteria

- Population or setting
  - Studies do not provide any evaluation of AI tools (only descriptions)
- Intervention
  - Automation tools that are not AI or where it is unclear if it is AI

We defined AI according to the regulatory definition provided by the UK Government Department of Science, Innovation and Technology's policy paper [22]. It defines AI as having two distinct characteristics of adaptivity and autonomy, meaning that AI can continually learn and infer patterns not envisioned by human programmers and can make decisions without the intent or oversight of a human. We excluded studies that discussed automated tools but did not display these

XSL•FO
RenderX

characteristics. We also defined evidence synthesis according to the Cochrane definition, which "involves combining information from multiple studies investigating the same topic to comprehensively understand their findings" [23]. No studies were excluded based on comparators or outcomes.

Citations identified through our search were imported and deduplicated using EndNote [24], with Rayyan [25] used to identify further duplicates. Using Rayyan for the entire screening process, records were initially screened via title and abstract by 1 reviewer. Full-text articles were then uploaded and screened for eligibility by another reviewer, with queries or discrepancies resolved by discussion with a third researcher.

Data Extraction and Synthesis

Data from included studies were extracted into a Microsoft Excel template. Extracted information included baseline study characteristics, such as authors, year of publication, study type, tool name, and outcomes measured. We also extracted specific information regarding the AI method used for tool development, web links to the tools or source code if available, as well as paywall features of the tool if relevant. Consistent with Joanna Briggs Institute scoping review methodology, we did not perform a critical appraisal of the included studies. Findings were thematically synthesized and presented visually, alongside being narratively described with descriptive statistics.

Results

Overview

A total of 11,226 studies were retrieved from database searches, with 3385 citations removed as duplicates (Figure 1). A total of 7841 citations were screened by their title and abstract, and 7499 were excluded after being deemed irrelevant. The remaining 342 citations underwent full-text screening, and 208 articles were included based on our eligibility criteria. An additional 19 records were identified through snowballing methods and a gray literature search, with 14 records being included. In total, 222 articles were included in the final review.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram illustrating the study selection process, including the number of records identified, screened, assessed for eligibility, and included in the final review. AI: artificial intelligence.

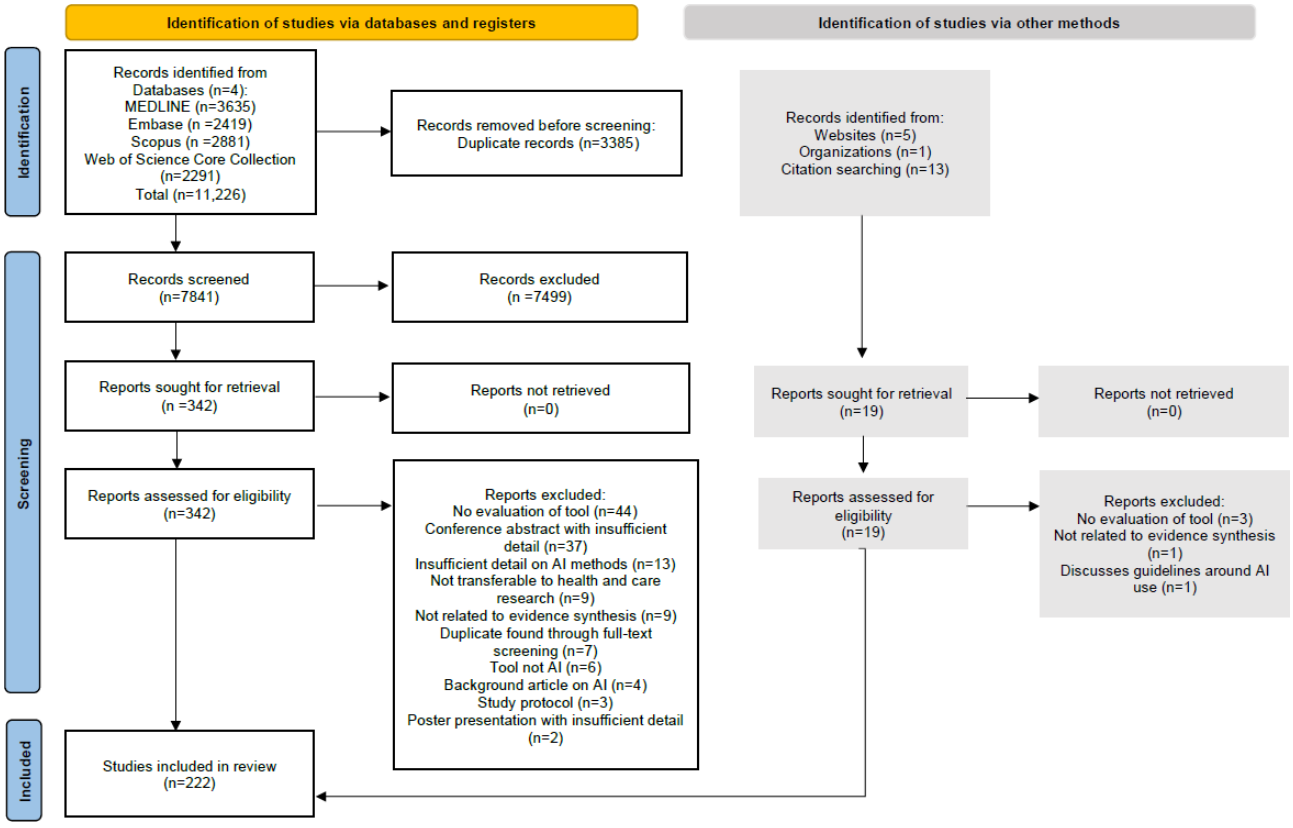

Study Characteristics

We included 206 (92.8%) primary studies and 16 (7.2%) secondary studies, with 2 (0.9%) stating they were explicitly AI assisted in their methods [26,27]. Three (1.4%) studies included were written in a language other than English [26,28,29]. A total of 11.7% (26/222) of studies were conference abstracts or posters. We identified 65 (29.3%) distinct AI tools and 25 (11.3%) open-source models or algorithms to automate the evidence synthesis process (Table S3 and S4 in Multimedia Appendix 1). Forty-two (18.9%) of the included studies had no specific tool name or data availability for their model.

Included studies could be divided into three broad categories: (1) methodological studies whereby researchers develop a novel algorithm or model and evaluate its effectiveness for automation; (2) evaluation, diagnostic accuracy, or feasibility studies of existing tools; and (3) comparative studies comparing the performance of AI tools to human researchers or similar tools executing the same task. Ten (4.5%) studies compared the performance of popular general-purpose LLMs, notably

XSL•FO
RenderX

ChatGPT, Claude, and Gemini, across various tasks, including screening and writing [30-40]. No studies compared traditional ML tools to LLM-based tools. There was only 1 (0.5%) study that evaluated an AI tool using a randomized trial study design [41].

A total of 54.1% (120/222) of included studies were published in 2024, followed by 16.7% (37/222) in 2023 and 14% (31/222) in 2022. Traditional ML, as an automation method, has remained relatively stable across the years, consistently supporting screening and extraction tasks (Figure 2). The use of BERTs has gradually increased since 2021, likely reflecting a growing interest in fine-tuned transformer models for domain-specific tasks. General-purpose LLMs saw a steep increase in 2024, likely coinciding with the widespread availability and maturity of models such as GPT-4 and Claude.

Figure 2. Year-on-year change in the study of three artificial intelligence methods for evidence synthesis automation. BERT: Bidirectional Encoder Representations from Transformers; LLM: large language model; ML: machine learning.

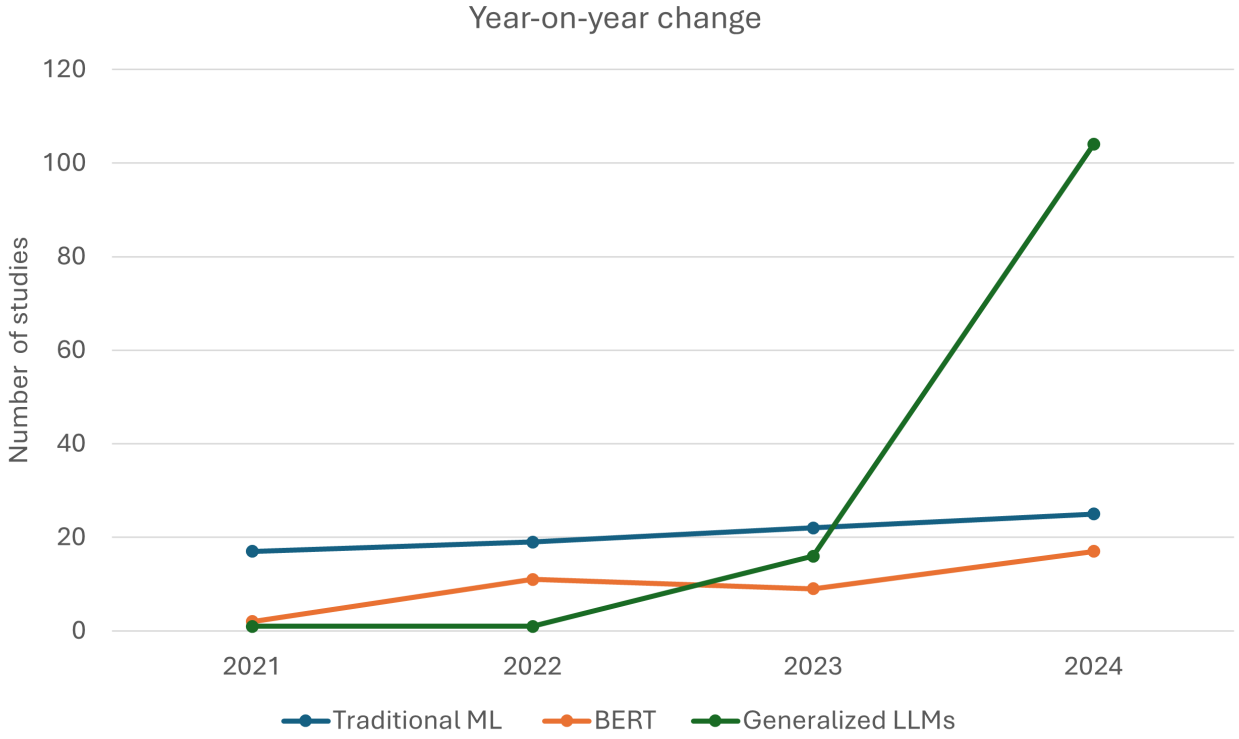

Evidence Synthesis Pathway

Many studies explored multiple areas of the evidence synthesis pathway, with most studies focusing on title and abstract screening (137/222, 61.7%), followed by data extraction (58/222, 26.1%), then literature searching (42/222, 18.9%; Figure 3). The proportion of studies per evidence synthesis stage reflects how amenable it is to automation. We categorized all 65 tools by evidence synthesis stage in Table S5 in Multimedia Appendix 1.

Figure 3. Percentage of studies examining the automation of distinct evidence synthesis stages using artificial intelligence.

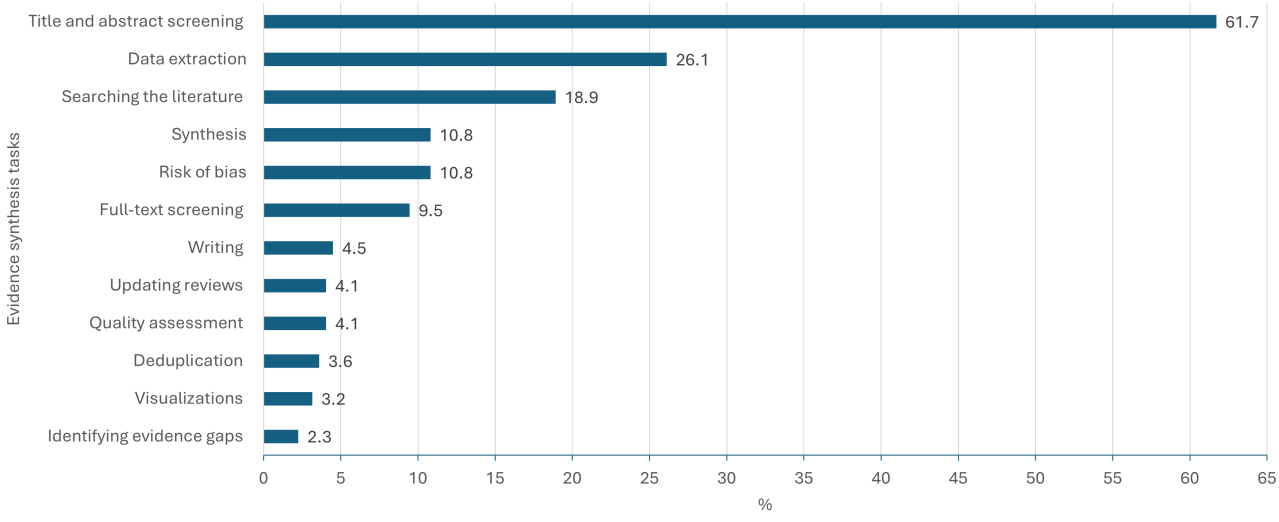

XSL•FO
RenderX

In total, 52.7% (117/222) of included studies researched tools leveraged by transformers. Some examples include general-purpose LLMs such as OpenAI's generative pretrained transformer (GPT) models or Mistral's 8×22B, as well as BERT models such as BioBERT or PubMedBERT (pretrained on predominantly biomedical text) that were fine-tuned for specific classification or screening tasks. In addition, 31.1% (69/222) evaluated tools underpinned by traditional ML algorithms alone, such as RCT Tagger, Research Screener, and Abstrackr. Some tools used an ensemble of ML and transformer-based methods, such as ASReview.

Evolution of AI Methods

A range of AI approaches were used to automate the evidence synthesis pathway, and these have evolved in computational power or complexity. The schematic in Figure 4 demonstrates two distinct categories of learning: traditional ML using classifier or clustering algorithms and deep learning leveraged by classical neural networks or transformers. Transformers are further divided into their subcategories of BERTs and general-purpose LLMs (ie, not designed specifically for evidence synthesis). Language and text processing abilities of the AI tools increase with complexity and computational power.

Figure 4. Schematic overview of artificial intelligence methods for evidence synthesis, ranging from traditional machine learning to neural networks and transformer-based large language models. AI: artificial intelligence.

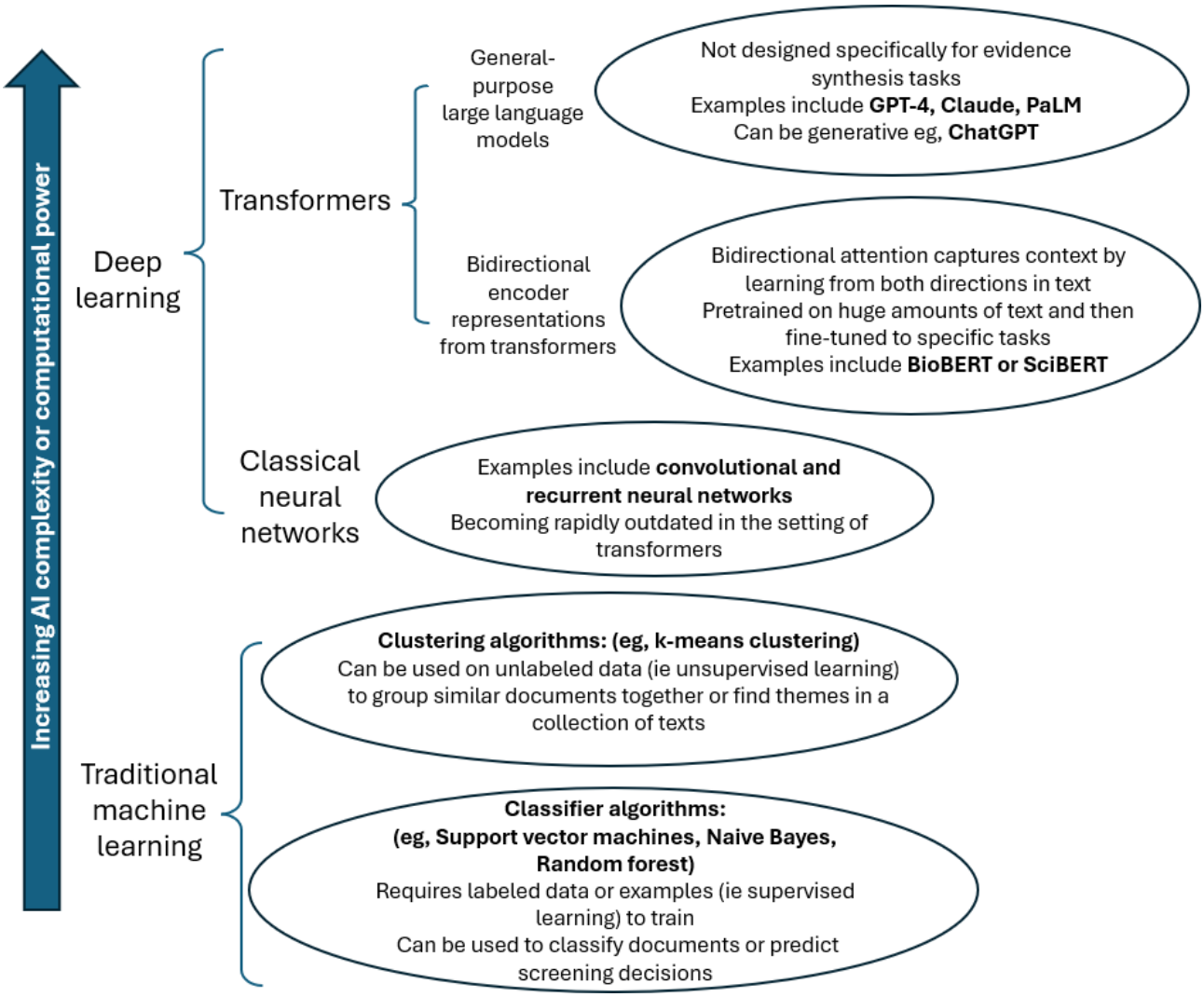

Both traditional ML and general-purpose LLMs were equally used to automate screening, with a similar distribution noted in data extraction (Figure 5). Traditional ML approaches being frequently leveraged for these stages likely highlight their ongoing suitability for evidence synthesis. Meanwhile, general-purpose LLMs can be applied across the entire evidence synthesis pathway. Generative LLMs are the only AI tool at present capable of producing text and therefore solely represent the automation of writing in evidence synthesis (Figure 5).

XSL•FO
RenderX

Figure 5. Visualization of included studies depicting the approximate number of studies by evidence synthesis stage and automation method. AI: artificial intelligence.

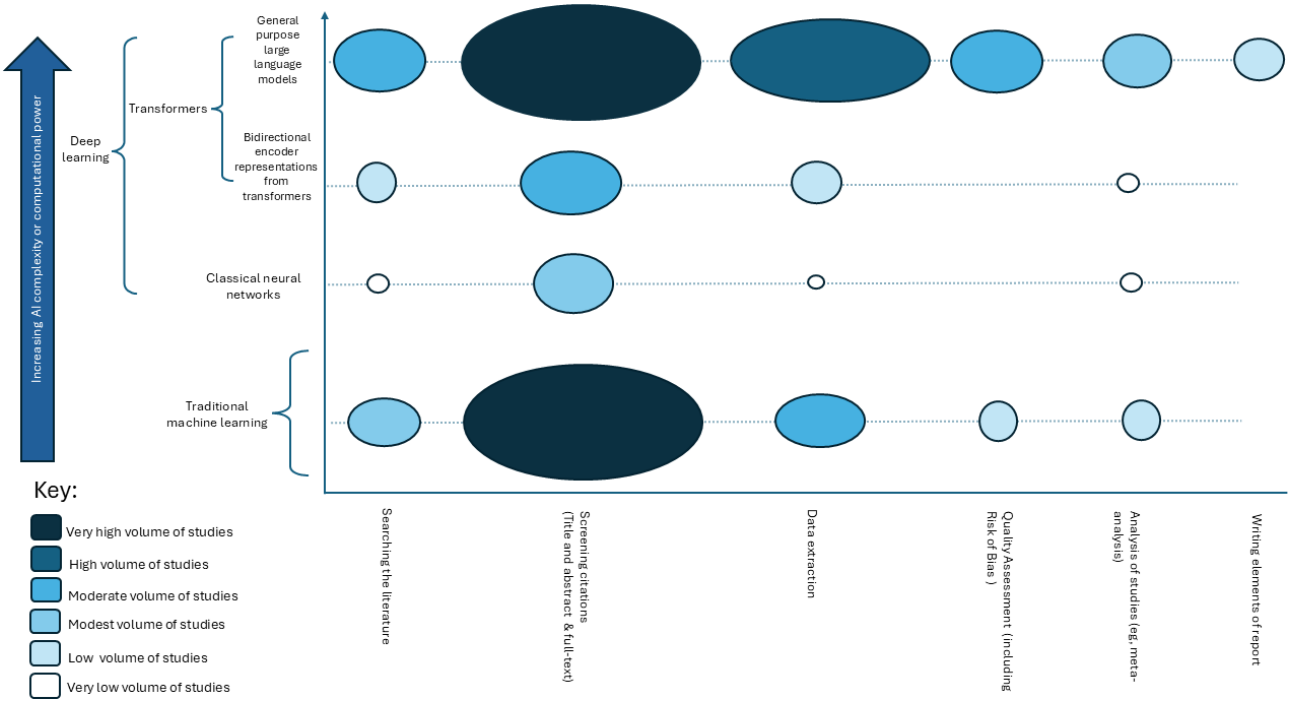

Distribution of Commonly Studied Tools

Of 222 studies, 70 (31.5%) used GPT models or ChatGPT. Other most cited tools included Anthropic's Claude (16/222, 7.2%), Rayyan (13/222, 5.9%), ASReview (12/222, 5.4%), Abstrackr (9/222, 4.1%), Google's Gemini (8/222, 3.6%), Covidence (5/222, 2.3%), DistillerSR (6/222, 2.7%), Colandr (3/222, 1.4%), and EPPI Reviewer (2/222, 0.9%). An alphabetized list of tools and their categorization across evidence synthesis tasks is available in Tables S3 and S5 in Multimedia Appendix 1. A list of all open-source models and algorithms is also available in Table S4 in Multimedia Appendix 1.

Outcomes

The most common outcomes measured were those designed to assess the technical performance of AI tools and models. Sensitivity and recall, which are the same measure, were used in 22.5% (50/222) and 23% (51/222) of studies, respectively. Other frequently reported metrics included specificity, accuracy, precision, area under the curve or area under the receiver operating characteristic curve, and F_1-score.

Using screening as an example, the area under the curve or the area under the receiver operating characteristic curve measures overall classification performance, indicating how well a model can distinguish between studies for inclusion versus exclusion. The F_1-score measures the balance between precision (how many papers selected by the model were actually relevant) and recall (how many relevant papers were successfully identified).

We also found specific outcomes designed to assess the researchers' workload or time saved. A popular outcome measure included Work Saved Over Sampling (WSS), which quantifies the percentage of manual screening work that can be avoided by using an AI tool while still maintaining the ability to identify all or most relevant studies. For example, a WSS of 30% means researchers can skip reviewing 30% of the citations while still identifying all relevant studies. Of 222 studies, 9 (4.1%) studies used WSS as a metric to assess the effectiveness of their automated screening tool [42-50].

Agreement between human researchers and AI tools, measured using metrics such as Cohen κ, was also used to assess the validity of AI decisions, although few included studies (12/222, 5.4%) used this approach.

Discussion

Principal Findings

This scoping review builds on previous reviews by comprehensively mapping the latest automation tools available across the evidence synthesis pathway and makes an important contribution to understanding how these tools are currently being researched and integrated into evidence synthesis workflows. We identified 65 distinct AI tools, some targeting specific aspects of the evidence synthesis pathway and others covering multiple stages, along with 25 open-source ML algorithms or BERT models designed or fine-tuned for evidence synthesis automation tasks. While published studies have traditionally focused on ML approaches for title and abstract screening and data extraction—the stages most readily automated—recent years have seen a notable shift toward transformer-based approaches for these same stages. Particularly in the past 2 years, there has been a rapid increase in published articles exploring the use of general-purpose LLMs, such as GPT, for evidence synthesis, despite these models not being developed specifically for it. This is notable given the risks generative AI poses due to its training methods and bias. We found that far fewer studies focused on optimizing the search for relevant studies or the analysis of the data or literature.

XSL•FO
RenderX

What the Results Mean

While transformer-based approaches are expanding the types of complex tasks that can be automated, task-specific ML methods continue to be developed and evaluated for well-defined problems. For example, priority screening using ML-based approaches has been validated thoroughly in the literature [43,51]. On the basis of the results from our review, a wide range of AI tools are being continually deployed and studied at different points across the evidence synthesis pathway, likely reflecting that no one tool or approach is superior to the other and that there is still room for improvement within evidence synthesis automation. While several studies attempted to compare tools with similar underlying AI methods, the lack of studies comparing distinct AI approaches, particularly traditional ML- to LLM-based tools, means we have no clear understanding of their comparative effectiveness.

It is likely that the choice of AI tools will be shaped by the expertise of the evidence synthesized, as well as the purpose and aim of the review itself. Extensively validated traditional ML tools or BERT models fine-tuned to specific research domains may be preferred by researchers or organizations who have a well-defined evidence synthesis task for automation, due to their reliability. For example, researchers conducting systematic reviews of drug or intervention effectiveness may find several well-evaluated tools at their disposal, as most available ML and BERT-based tools excel at interpreting and extracting data from randomized trials with clear eligibility criteria.

General-purpose LLMs, on the other hand, offer flexible language understanding and generative capabilities that can assist researchers in rapidly processing large volumes of text, identifying relevant information, and synthesizing preliminary insights across diverse topics. While these tools may allow for more nuanced interactions with complex evidence, they are likely less acceptable to traditional academic standards.

Most tools identified in this review were designed to optimize systematic reviews, predominantly in the biomedical domain. This is likely because of the very strict methodology that is applied to such reviews, which allows for replicability and, therefore, optimal comparison to the gold standard, manually performed reviews. Limited research explored AI automation in complex, multidisciplinary fields such as public health or social sciences, and few tools addressed nonsystematic review methods, such as narrative, realist, or integrative reviews. This disparity likely reflects a difference in data structure; more standardized data found in quantitative reviews is easier to process and is more machine readable than qualitative data, which tends to be heterogeneous and context dependent.

BERT models continue to be used as a method of automation for screening and data extraction-related tasks, with popular models such as SciBERT and BioBERT being fine-tuned to perform classification tasks for biomedical research questions notably. There is early emerging evidence, however, that general-purpose LLMs are equal if not superior to BERTs at most evidence synthesis tasks [37,52,53]. This potentially indicates to researchers that general-purpose LLMs, even if not designed specifically for evidence synthesis, may be a suitable option for automation if unable to invest the time and resources required to fine-tune BERT models.

Most studies included in our review used technical outcomes to assess tool performance, rather than directly comparing outcomes between AI tools and human researchers performing the same task. While sensitivity and specificity measures are important, there are implications for feasibility if studies do not assess AI tools pragmatically. Adoption of AI tools is likely going to be higher if there is a clear indication that the tool will save a researcher's time and improve workload for similar outputs.

Comparison With Previous Literature

Similar to the findings in Khalil and Jimenez's reviews [15,16], we found that the task most amenable to automation within evidence synthesis was title and abstract screening, with 62% of included studies exploring both ML and deep learning tools to automate this stage. We also found that most ML-based tools identified previously in both reviews remain active and usable today, with several having significant updates incorporating more advanced AI methods such as EPPI Reviewer, Covidence, and Rayyan. Our review was also uniquely able to identify several LLM-based tools that are built specific for evidence synthesis, such as Elicit, Scite, and Consensus. We also found in our review that general-purpose LLMs were predominantly applied to screening and searching of the literature, with GPT being the most frequently cited general-purpose LLM, consistent with Lieberum's review [12].

Unlike previous reviews, however, which studied ML- and LLM-based tools in isolation, our review demonstrates comparatively the strong shift in the evidence base to studying general-purpose LLMs for evidence synthesis tasks in the last 5 years, particularly for screening, extraction, and, uniquely, writing-related tasks. However, the evidence base for AI-generated or enhanced scientific writing is still evolving and remains a widely debated topic in the research community due to ethical and logistical issues attached to the practice [54-57]. Most journals at present ask for disclosure of AI assistance in the writing process.

Our review also uniquely sought to identify comparative studies between different AI approaches, particularly between traditional ML- and LLM-based tools, as previous reviews focused on distinct approaches in isolation. However, we were unable to identify any such studies within our search time frame, highlighting an important gap in the research. Recent domain-specific evidence has shown that LLM-based tools may outperform ML tools in evidence synthesis tasks such as screening, although these findings are limited due to a lack of generalizability [58].

Strengths and Limitations

There are several strengths to this scoping review. First, it offers a broad and comprehensive overview of the field, including over 200 studies. We clearly mapped AI tool use across the entire evidence synthesis pathway, providing complete lists of tools available for specific or multiple tasks. It also uses previous reviews' frameworks to expand on the topic of automation in evidence synthesis, adding consistently to this evidence base.

XSL•FO
RenderX

The search strategy was robust, covering multiple academic databases as well as gray literature sources, increasing the likelihood of capturing both peer-reviewed and nontraditional publications.

However, there are several limitations. Given the rapidly evolving nature of AI in evidence synthesis, it is possible that several recent publications were not captured. There is also temporal bias: general-purpose LLMs have only gained widespread adoption in the past 2 years, meaning fewer high-quality peer-reviewed studies are available compared to evaluations of older AI approaches. In addition, this review included studies that explicitly evaluated AI methods or tools to automate evidence synthesis, thereby excluding studies that may have researched AI without clear or proper disclosure of it. This potentially limited our evidence base and underestimated the actual AI usage in evidence synthesis. Additionally, a lack of granularity in reporting the underlying AI methods or technical approaches restricted the meaningful categorization and comparison of tools. Poor reporting of tool development methods also limits generalizability to health and care, particularly when models were trained within specific research domains. Publication bias may also exist, as tools developed in academic settings are more likely to appear in scholarly journals, whereas commercial tools may be underrepresented in the literature. Finally, an inherent limitation of scoping review methodology is the lack of critical appraisal of the evidence base, which limits our understanding of the quality of studies included.

Recommendations for Policy and Practice

Until we can put in place mitigations that are effective in reducing error and bias in AI tools, there will be a need for a human researcher in the loop [59]. Therefore, the tools highlighted in this paper should be used as adjuncts to humans, rather than replacements, to maintain high-quality and rigorous reviews.

Funders and researchers should continue to support efforts to build an evidence base in this field that can provide a feedback loop for developers to improve upon the tools' effectiveness, feasibility, and acceptability. In particular, assessing these tools in real-world settings and comparing traditional ML methods to newer LLM-based approaches is pertinent to understanding automation opportunities and challenges. Researchers and funders should foster knowledge sharing on the effective use of emerging AI tools, such as through the International Collaboration of the Automation of Systematic Reviews [60]. This can be through training resources, collaborative learning opportunities, and sharing best practices. This will help build collective understanding and capacity across the research community.

Almost all tools found had paid upgrades or monthly subscriptions. The lack of comprehensive tools that are free to use is concerning from both an accessibility and equity angle and is likely to influence how such tools are integrated into research practice.

Quality standards are needed to ensure ethical tool development, and researchers and stakeholders should adhere to guidelines that stipulate methods to properly and appropriately integrate automation tools into evidence synthesis, such as the recommendations and guidance for Responsible AI in Evidence Synthesis [61].

Decision makers who use AI-assisted evidence syntheses should be aware of the issues around error and bias and should exercise caution when using reviews generated by tools that claim to perform the whole process.

LLMs also have significant electricity demands due to their scale, data requirements, and computational demands [62,63]. Estimates suggest that generative AI processing, from data center operation and cooling, could consume over 8% of the United States' electricity and 5% of Europe's by 2030 [64]. Furthermore, manufacturing generative processing units poses significant environmental impacts, particularly for communities near production like in Taichung, Taiwan [64]. While some developers are exploring LLMs using renewable energy, such developments remain in their infancy [65].

In the setting of many organizations and institutions requiring consideration of sustainability in their practice, it is important that we review the environmental implications of LLMs if implemented regularly into evidence synthesis workflows. Therefore, we recommend that sustainability research be undertaken around the use of such AI tools, which is not at all considered in the current evidence base.

Conclusions

Having accurate, high-quality, and contemporaneous reviews of evidence is essential to making the most informed decisions about health and care. Many of the processes involved in evidence synthesis—particularly identifying relevant studies, screening titles and abstracts, and extracting structured data—involve repetitive, rule-based tasks that are well-suited to automation using AI. With the ongoing development of ML tools and the introduction of LLMs, researchers now have a large suite of AI-based automation tools available to them, as highlighted in our review. While some tools are yet to be comprehensively evaluated, validated, and compared, given the risks generative AI poses to evidence synthesis integrity, the evidence base highlights many potential avenues to improve manual workload. As we continue to integrate such tools into research workflows, it is vital that we continue to monitor outcomes, maintain transparency, and always have a researcher-in-the-loop to ensure that evidence reviews continue to be of a high standard.

Acknowledgments

The Health Foundation was not involved in the design and methodology of this review. They provided final comments to the paper that were incorporated into the manuscript. The authors would like to acknowledge Starry Chow (The University of Hong

XSL•FO
RenderX

Kong—Li Ka Shing Faculty of Medicine) for their input in translation during the full-text screening process. Generative artificial intelligence was not used to inform this review's design or methodology or generate any text or figures. It was only used to refine writing for grammar and clarity.

Funding

"An overview and evaluation of the use of automated tools in rapid evidence synthesis" project is supported by the Health Foundation, an independent charitable organization working to build a healthier United Kingdom (reference FR-0006738).

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

Authors' Contributions

Conceptualization: SH (equal), JF (lead)
Data curation: SH (equal), H Pearce (equal), CA (supporting), IK (supporting)
Formal analysis: SH (lead), JF (supporting)
Funding acquisition: JF
Investigation: SH (lead), H Pearce (equal), IK (supporting)
Methodology: SH (lead), JF (equal)
Project administration: SH (lead), JF (equal)
Supervision: JF
Visualization: SH (lead), H Pearce (supporting), JF (supporting)
Writing – original draft: SH (lead)
Writing – review & editing: H Pearce (supporting), H Painter (supporting), CA (supporting), LL (supporting), IK (supporting), JK (supporting), AC (supporting), JF (equal)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials including detailed database search strategies, PRISMA-S checklist, and categorized tables of AI tools and open-source models identified in the review.
[DOCX File, 49 KB - jmir_v28i1e81597_app1.docx]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 28 KB - jmir_v28i1e81597_app2.docx]

References

1. Khalil H, Tamara L, Rada G, Akl EA. Challenges of evidence synthesis during the 2020 COVID pandemic: a scoping review. J Clin Epidemiol 2022 Feb;142:10-18. [doi: 10.1016/j.jclinepi.2021.10.017] [Medline: 34718121]
2. Tercero-Hidalgo JR, Khan KS, Bueno-Cavanillas A, et al. Artificial intelligence in COVID-19 evidence syntheses was underutilized, but impactful: a methodological study. J Clin Epidemiol 2022 Aug;148:124-134. [doi: 10.1016/j.jclinepi.2022.04.027] [Medline: 35513213]
3. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 2017 Feb 27;7(2):e012545. [doi: 10.1136/bmjopen-2016-012545] [Medline: 28242767]
4. Clark J, Glasziou P, Del Mar C, Bannach-Brown A, Stehlik P, Scott AM. A full systematic review was completed in 2 weeks using automation tools: a case study. J Clin Epidemiol 2020 May;121:81-90. [doi: 10.1016/j.jclinepi.2020.01.008] [Medline: 32004673]
5. Clark J, McFarlane C, Cleo G, Ishikawa Ramos C, Marshall S. The impact of systematic review automation tools on methodological quality and time taken to complete systematic review tasks: case study. JMIR Med Educ 2021 May 31;7(2):e24418. [doi: 10.2196/24418] [Medline: 34057072]
6. Abogunrin S, Muir JM, Zerbini C, Sarri G. How much can we save by applying artificial intelligence in evidence synthesis? Results from a pragmatic review to quantify workload efficiencies and cost savings. Front Pharmacol 2025;16:1454245. [doi: 10.3389/fphar.2025.1454245] [Medline: 39959426]

XSL•FO
RenderX

7. Rogers K, Miller A, Girgis A, Clark EC, Neil-Sztramko SE, Dobbins M. Leveraging AI to optimize maintenance of health evidence and offer a one-stop shop for quality-appraised evidence syntheses on the effectiveness of public health interventions: quality improvement project. J Med Internet Res 2025 Jul 29;27(1):e69700. [doi: 10.2196/69700] [Medline: 40729661]
8. Ofori-Boateng R, Aceves-Martins M, Wiratunga N, Moreno-Garcia CF. Towards the automation of systematic reviews using natural language processing, machine learning, and deep learning: a comprehensive review. Artif Intell Rev 2024;57(8):200. [doi: 10.1007/s10462-024-10844-w]
9. Liu J, Lai H, Zhao W, et al. AI-driven evidence synthesis: data extraction of randomized controlled trials with large language models. Int J Surg 2025;111(3):2722-2726. [doi: 10.1097/JS9.0000000000002215]
10. Wang Z, Cao L, Danek B, Jin Q, Lu Z, Sun J. Accelerating clinical evidence synthesis with large language models. NPJ Digit Med 2025 Aug 8;8(1):509. [doi: 10.1038/s41746-025-01840-7] [Medline: 40775042]
11. Clark J, Barton B, Albarqouni L, et al. Generative artificial intelligence use in evidence synthesis: a systematic review. Res synth methods 2025 Jul;16(4):601-619. [doi: 10.1017/rsm.2025.16]
12. Lieberum JL, Toews M, Metzendorf MI, et al. Large language models for conducting systematic reviews: on the rise, but not yet ready for use-a scoping review. J Clin Epidemiol 2025 May;181:111746. [doi: 10.1016/j.jclinepi.2025.111746] [Medline: 40021099]
13. Siemens W, von Elm E, Binder H, et al. Opportunities, challenges and risks of using artificial intelligence for evidence synthesis. BMJ Evid Based Med 2025 Dec 1;30(6):381-384. [doi: 10.1136/bmjebm-2024-113320] [Medline: 39788693]
14. Ong AY, Merle DA, Wagner SK, Keane PA. Exploring the dilemma of AI use in medical research and knowledge synthesis: a perspective on deep research tools. J Med Internet Res 2025 Jul 15;27(1):e75666. [doi: 10.2196/75666] [Medline: 40663724]
15. Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. J Clin Epidemiol 2022 Apr;144:22-42. [doi: 10.1016/j.jclinepi.2021.12.005] [Medline: 34896236]
16. Cierco Jimenez R, Lee T, Rosillo N, et al. Machine learning computational tools to assist the performance of systematic reviews: a mapping review. BMC Med Res Methodol 2022 Dec 16;22(1):322. [doi: 10.1186/s12874-022-01805-4] [Medline: 36522637]
17. Roth S, Wermer-Colan A. Machine learning methods for systematic reviews:: a rapid scoping review. Dela J Public Health 2023 Nov;9(4):40-47. [doi: 10.32481/djph.2023.11.008] [Medline: 38173960]
18. Khalil H, Pollock D, McInerney P, et al. Automation tools to support undertaking scoping reviews. Res Synth Methods 2024 Nov;15(6):839-850. [doi: 10.1002/jrsm.1731] [Medline: 38885942]
19. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020 Oct;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
20. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
21. Litmaps. URL: https://www.litmaps.com/ [accessed 2025-05-12]
22. A pro-innovation approach to AI regulation. GOV.UK. URL: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper [accessed 2024-12-20]
23. Evidence synthesis - what is it and why do we need it? Cochrane. 2019 Sep 13. URL: https://www.cochrane.org/news/evidence-synthesis-what-it-and-why-do-we-need-it [accessed 2024-12-20]
24. EndNote. URL: https://endnote.com/ [accessed 2025-05-12]
25. Rayyan OM. AI and the evolution of journalistic practices. J Inf Stud Technol 2025 Jul 31;2. [doi: 10.5339/jist.2025.15]
26. Cardoso Sampaio R, Chagas V, Sinimbu Sanchez C. An artificial intelligence (AI)-assisted scoping review of emerging uses of AI in qualitative research and its ethical considerations. Rev Pesq Qual 2024;12(30):01-28. [doi: 10.33361/RPQ.2024.v.12.n.30.729]
27. Teperikidis E, Boulmpou A, Potoupni V, Kundu S, Singh B, Papadopoulos C. Does the long-term administration of proton pump inhibitors increase the risk of adverse cardiovascular outcomes? A ChatGPT powered umbrella review. Acta Cardiol 2023 Nov;78(9):980-988. [doi: 10.1080/00015385.2023.2231299] [Medline: 37431972]
28. Guo Y, Zhang X, Sun W, Deng H. Application of automated literature screening tools in systematic reviews. Med J Peking Union Med Coll Hosp 2024;15(4):921-926. [doi: 10.12290/xhyxzz.2023-0257]
29. Esposito C, Dell'Omo G, Di Ianni D, Di Procolo P. Human vs. ChatGPT. Is it possible obtain comparable results in the analysis of a scientific systematic review? Recenti Prog Med 2024 Sep;115(9):420-425. [doi: 10.1701/4334.43184] [Medline: 39269357]
30. AlSagri HS, Farhat F, Sohail SS, Saudagar AKJ. ChatGPT or Gemini: who makes the better scientific writing assistant? J Acad Ethics 2025 Sep;23(3):1121-1135. [doi: 10.1007/s10805-024-09549-0]
31. De Cassai A, Dost B, Karapinar YE, et al. Evaluating the utility of large language models in generating search strings for systematic reviews in anesthesiology: a comparative analysis of top-ranked journals. Reg Anesth Pain Med 2025 Jan 19;Published online January 2025:4. [doi: 10.1136/rapm-2024-106231] [Medline: 39828514]
32. Delgado-Chaves FM, Jennings MJ, Atalaia A, et al. Transforming literature screening: the emerging role of large language models in systematic reviews. Proc Natl Acad Sci U S A 2025 Jan 14;122(2):e2411962122. [doi: 10.1073/pnas.2411962122] [Medline: 39761403]

XSL•FO
RenderX

33. Konet A, Thomas I, Gartlehner G, et al. Performance of two large language models for data extraction in evidence synthesis. Res Synth Methods 2024 Sep;15(5):818-824. [doi: 10.1002/jrsm.1732] [Medline: 38895747]
34. Li M, Sun J, Tan X. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. Syst Rev 2024 Aug 21;13(1):219. [doi: 10.1186/s13643-024-02609-x] [Medline: 39169386]
35. Omar M, Nassar S, Hijazi K, Glicksberg BS, Nadkarni GN, Klang E. Generating credible referenced medical research: a comparative study of openAI's GPT-4 and Google's gemini. Comput Biol Med 2025 Feb;185:109545. [doi: 10.1016/j.compbiomed.2024.109545] [Medline: 39667055]
36. Rathi H, Malik A, Behera DC, Kamboj G. Msr28 use of large language model (LLM) for full-text screening in systematic literature reviews: a comparative analysis. Value Health 2024 Jun;27(6):S264. [doi: 10.1016/j.jval.2024.03.1461]
37. Sanghera R, Thirunavukarasu AJ, Khoury ME, et al. High-performance automated abstract screening with large language model ensembles. arXiv. Preprint posted online on Nov 22, 2024. [doi: 10.48550/arXiv.2411.02451]
38. Šuster S, Baldwin T, Verspoor K. Zero- and few-shot prompting of generative large language models provides weak assessment of risk of bias in clinical trials. Res Synth Methods 2024 Nov;15(6):988-1000. [doi: 10.1002/jrsm.1749] [Medline: 39176994]
39. Thelwall M. Is Google Gemini better than ChatGPT at evaluating research quality? J. Data Inf. Sci. 2025 Apr 1;10(2):1-5. [doi: 10.2478/jdis-2025-0014]
40. Trad F, Charafeddine J, Chkahtoura M, Rahme M, Fuleihan GEH, Chehab A. Streamlining systematic reviews in medical research: a novel application of large language models. J Bone Miner Res 2024;39(Supplement 1). [doi: 10.1093/jbmr/zjae038] [Medline: 38477812]
41. Arno A, Thomas J, Wallace B, Marshall IJ, McKenzie JE, Elliott JH. Accuracy and efficiency of machine learning-assisted risk-of-bias assessments in "real-world" systematic reviews: a noninferiority randomized controlled trial. Ann Intern Med 2022 Jul;175(7):1001-1009. [doi: 10.7326/M22-0092] [Medline: 35635850]
42. Bravo A, Bennetts L, Atanasov P. Accelerating the early identification of relevant studies in title and abstract screening. Presented at: 2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC); Nov 12-14, 2021. [doi: 10.1109/ISCSIC54682.2021.00034]
43. Ferdinands G, Schram R, de Bruin J, et al. Performance of active learning models for screening prioritization in systematic reviews: a simulation study into the Average Time to Discover relevant records. Syst Rev 2023 Jun 20;12(1):100. [doi: 10.1186/s13643-023-02257-7] [Medline: 37340494]
44. Yao X, Kumar MV, Su E, Flores Miranda A, Saha A, Sussman J. Evaluating the efficacy of artificial intelligence tools for the automation of systematic reviews in cancer research: a systematic review. Cancer Epidemiol 2024 Feb;88:102511. [doi: 10.1016/j.canep.2023.102511] [Medline: 38071872]
45. Chai KEK, Lines RLJ, Gucciardi DF, Ng L. Research Screener: a machine learning tool to semi-automate abstract screening for systematic reviews. Syst Rev 2021 Apr 1;10(1):93. [doi: 10.1186/s13643-021-01635-3] [Medline: 33795003]
46. van Dinter R, Catal C, Tekinerdogan B. A multi-channel convolutional neural network approach to automate the citation screening process. Appl Soft Comput 2021 Nov;112:107765. [doi: 10.1016/j.asoc.2021.107765]
47. Kusa W, Hanbury A, Knoth P. Automation of citation screening for systematic literature reviews using neural networks: a replicability study. Presented at: Advances in Information Retrieval - 44th European Conference on (IR) Research, (ECIR) 2022; Apr 10-14, 2022. [doi: 10.1007/978-3-030-99736-6_39]
48. Muthu S. The efficiency of machine learning-assisted platform for article screening in systematic reviews in orthopaedics. Int Orthop 2023 Feb;47(2):551-556. [doi: 10.1007/s00264-022-05672-y] [Medline: 36562816]
49. Akinseloyin O, Jiang XR, Palade V. A question-answering framework for automated abstract screening using large language models. J Am Med Inform Assoc 2024 Sep 1;31(9):1939-1952. [doi: 10.1093/jamia/ocae166] [Medline: 39042516]
50. Feng Y, Liang S, Zhang Y, et al. Automated medical literature screening using artificial intelligence: a systematic review and meta-analysis. J Am Med Inform Assoc 2022 Jul 12;29(8):1425-1432. [doi: 10.1093/jamia/ocac066] [Medline: 35641139]
51. Tsou AY, Treadwell JR, Erinoff E, Schoelles K. Machine learning for screening prioritization in systematic reviews: comparative performance of Abstrackr and EPPI-Reviewer. Syst Rev 2020 Apr 2;9(1):73. [doi: 10.1186/s13643-020-01324-7] [Medline: 32241297]
52. Jahan I, Laskar MTR, Peng C, Huang J. Evaluation of ChatGPT on biomedical tasks: a zero-shot comparison with fine-tuned generative transformers. arXiv. Preprint posted online on Aug 24, 2023. [doi: 10.48550/arXiv.2302.10198]
53. Zhong Q, Ding L, Liu J, Du B, Tao D. Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT. arXiv. Preprint posted online on Mar 2, 2023. [doi: 10.48550/arXiv.2302.10198]
54. Draxler F, Werner A, Lehmann F, et al. The AI ghostwriter effect: when users do not perceive ownership of AI-generated text but self-declare as authors. ACM Trans Comput-Hum Interact 2024 Apr 30;31(2):1-40. [doi: 10.1145/3637875]
55. Hosseini M, Gordijn B, Kaebnick GE, Holmes K. Disclosing generative AI use for writing assistance should be voluntary. Res Ethics 2025 Jun 21;21(4):728-735. [doi: 10.1177/17470161251345499] [Medline: 40950847]
56. Cohen JF, Moher D. Generative artificial intelligence and academic writing: friend or foe? J Clin Epidemiol 2025 Mar;179:111646. [doi: 10.1016/j.jclinepi.2024.111646] [Medline: 39706536]

XSL•FO
RenderX

57. Formosa P, Bankins S, Matulionyte R, Ghasemi O. Can ChatGPT be an author? Generative AI creative writing assistance and perceptions of authorship, creatorship, responsibility, and disclosure. AI & Soc 2025 Jun;40(5):3405-3417. [doi: 10.1007/s00146-024-02081-0]
58. Dai ZY, Wang FQ, Shen C, et al. Accuracy of large language models for literature screening in thoracic surgery: diagnostic study. J Med Internet Res 2025 Mar 11;27:e67488. [doi: 10.2196/67488] [Medline: 40068152]
59. Li Y, Datta S, Rastegar-Mojarad M, et al. Enhancing systematic literature reviews with generative artificial intelligence: development, applications, and performance evaluation. J Am Med Inform Assoc 2025 Apr 1;32(4):616-625. [doi: 10.1093/jamia/ocaf030] [Medline: 40036547]
60. International Collaboration for the Automation of Systematic Reviews. URL: https://icasr.github.io/ [accessed 2025-05-12]
61. Thomas J, Flemyng E, Noel-Storr A. Responsible AI in Evidence Synthesis (RAISE): guidance and recommendations. Open Science Framework. 2024 Aug 29. URL: https://osf.io/fwaud/ [accessed 2025-04-24]
62. Rillig MC, Ågerstrand M, Bi M, Gould KA, Sauerland U. Risks and benefits of large language models for the environment. Environ Sci Technol 2023 Mar 7;57(9):3464-3466. [doi: 10.1021/acs.est.3c01106] [Medline: 36821477]
63. Ji Z, Jiang M. A systematic review of electricity demand for large language models: evaluations, challenges, and solutions. Renew Sustain Energy Rev 2026 Jan;225:116159. [doi: 10.1016/j.rser.2025.116159]
64. Hosseini M, Gao P, Vivas-Valencia C. A social-environmental impact perspective of generative artificial intelligence. Environ Sci Ecotechnol 2025 Jan;23:100520. [doi: 10.1016/j.ese.2024.100520] [Medline: 39811486]
65. Singh A, Patel NP, Ehtesham A, Kumar S, Khoei TT. A survey of sustainability in large language models: applications, economics, and challenges. arXiv. Preprint posted online on Mar 29, 2023. [doi: 10.48550/arXiv.2412.04782]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
GPT: generative pretrained transformer
LLM: large language model
ML: machine learning
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
WSS: Work Saved Over Sampling

Edited by A Mavragani, S Brini; submitted 01.Aug.2025; peer-reviewed by E Ting, H Khalil; revised version received 27.Oct.2025; accepted 03.Nov.2025; published 30.Mar.2026.

Please cite as:
Harasgama S, Pearce H, Appel C, Loftus L, Painter H, Kuhn I, Karpusheff J, Ceesay A, Ford J
Artificial Intelligence Tools for Automating Evidence Synthesis: Scoping Review
J Med Internet Res 2026;28:e81597
URL: https://www.jmir.org/2026/1/e81597
doi:10.2196/81597

© Sashika Harasgama, Helen Pearce, Cameron Appel, Liam Loftus, Helena Painter, Isla Kuhn, Justine Karpusheff, Aji Ceesay, John Ford. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Smartphone-Based Digital Phenotyping Across Health Conditions: Scoping Review

Arlen Dumas[1], BS; Joanne Hokayem[2], BS; Georgia Goodman[3], BS; Krishna Venkatasubramanian[1], PhD; Peter Chai[2], MD, MMS

[1]Department of Computer Science and Statistics, University of Rhode Island, 9 Greenhouse Road, Kingston, RI, United States

[2]The Fenway Institute, Boston, MA, United States

[3]UCLA (University of California, Los Angeles) Health, Los Angeles, CA, United States

Corresponding Author:
Krishna Venkatasubramanian, PhD
Department of Computer Science and Statistics, University of Rhode Island, 9 Greenhouse Road, Kingston, RI, United States

Abstract

Background: Smartphone-based digital phenotyping uses built-in sensors and usage patterns to passively capture behavioral and environmental data relevant to health and has been applied extensively in mental health and chronic disease contexts.

Objective: This review synthesizes studies that use smartphone-based digital phenotyping, defined as approaches that rely exclusively on onboard smartphone sensors to characterize specific health conditions. To our knowledge, this work provides the most comprehensive cross-condition synthesis of smartphone-based digital phenotyping to date, spanning mental health, physical health, and substance use disorders (SUDs), and highlighting common practices, gaps, and opportunities for future research.

Methods: We conducted a scoping review of English-language, peer-reviewed papers published between 2012 and 2025 in Google Scholar, IEEE Xplore, ACM Digital Library, and PubMed using terms such as "mobile sensing" and "digital phenotyping." Eligible papers used onboard smartphone sensors to assess health and went beyond self-report. Studies that did not rely on smartphone auxiliary sensing modalities or digital phenotyping were excluded.

Results: We performed a descriptive synthesis of study characteristics, sensors, and health domains. Of 111 papers identified, 65 met inclusion criteria. Most studies were observational and relied on passive sensing. Sample sizes ranged from fewer than 10 to over 18,000 participants, with a median of 52 (IQR=26 - 126). Mental health conditions were most frequently examined, including depression (n=16), bipolar disorder (n=11), stress or anxiety (n=10), and schizophrenia (n=8). Less commonly studied conditions included SUDs (n=7), Parkinson disease (n=4), and sleep apnea (n=2). Sensor streams varied widely and included diverse passive smartphone data sources capturing mobility, communication, device usage, environmental context, and user interaction patterns. Ground-truth measurements most commonly relied on validated clinical scales (eg, Patient Health Questionnaire-9, Young Mania Rating Scale [YMRS], and Pittsburgh Sleep Quality Index; n=41), followed by ecological momentary assessments (n=18), clinician-confirmed diagnoses (n=9), and physiological measures such as polysomnography (n=3). Across studies, recurring methodological limitations included incomplete or inconsistent sensor descriptions, limited reporting of data quality (eg, sampling rates and missingness), and heterogeneous validation practices. These issues limit comparability and reproducibility and underscore the need for clearer reporting standards and greater data availability.

Conclusions: This scoping review provides the first comprehensive synthesis of smartphone-only digital phenotyping studies spanning mental health, physical health, and SUDs. Unlike prior reviews, this work maps behavioral associations derived exclusively from smartphone sensors across a broad range of health domains. The primary contribution of this review lies in its consolidation of behavioral associations observed across studies, enabling researchers to correlate new findings to the existing evidence base and identify opportunities for replication, extension, or clinical translation. Collectively, these findings highlight both the promise of smartphone-based digital phenotyping in real-world settings and the need for improved standardization to support translation into clinical and public health applications.

(J Med Internet Res 2026;28:e84146) doi:10.2196/84146

KEYWORDS

digital phenotyping; smartphones; digital health; scoping review; health behaviors; mobile phone

XSL•FO
RenderX

Introduction

Since their introduction in 1994, smartphones have penetrated almost every aspect of human life worldwide [1]. These devices are ubiquitous in many cultures and have transformed access to the internet and communication with others. Like many other sectors, health care has been directly impacted by the use of smartphones [2,3]. Patients can now conduct virtual visits with their health care providers [4], have access to a wide range of sources of medical information, and can access health records on smartphones [3]. Recent survey data in the United States suggests that 62% of individuals would like to access their medical records through mobile-facing apps, and in 2022, 48% of patients accessed medical records via the internet at least once [5]. Additionally, increased use of wearable and other health-related tracking devices has positioned smartphones as a centralized hub for health data and behavior tracking [6,7]. This data suggests that individuals access and use smartphones to assess and modulate individual health, positioning these devices as important platforms to interact with individuals [6,8].

One potential strategy to leverage smartphones to understand health behavior is through digital phenotyping, a strategy that leverages data from mobile computing devices that most people now have with them (eg, smartphones, smartwatches, rings, and other wearable devices) to capture a person's behavior, and through them, understand any health conditions present [9]. More specifically, digital phenotyping is the collection and analysis of data generated through human interaction with mobile computing devices. It encompasses (1) passive input collected from sensors on a smartphone, such as continuous GPS monitoring, accelerometry data; and (2) active input, such as ecological momentary assessments (EMAs) completed by the user of the smartphone. By integrating such passive and active inputs, digital phenotyping aims to explore behaviors that correlate with specific health conditions [10].

Smartphone-based digital phenotyping is a form of digital phenotyping in which smartphones are the only devices used for collecting sensor data from individuals. Smartphones offer a unique advantage due to their widespread use, enabling researchers to collect a rich, high-volume dataset that can be used to assess behaviors (also known as phenotypes) in the context of a specific health condition in an unobtrusive way [11]. Data that has been captured via smartphones can be analyzed through a variety of methods to provide insight into a variety of mental and physical health conditions [12]. Compared to traditional approaches of monitoring and diagnosing complex health conditions (eg, schizophrenia and bipolar disorder) that require considerable human interventions, smartphone-based digital phenotyping offers an opportunity to passively track these conditions over time [13]. The other potential advantage of smartphone-based digital phenotyping is the opportunity to democratize diagnoses of various health conditions simply by assessing smartphone use parameters in the context of the user's known actions [14]. Despite its promise, multiple challenges exist to operationalize smartphone-based digital phenotyping, account for variability in data, and generalize findings across patient populations [9].

This scoping review aims to comprehensively describe the current state of smartphone-based digital phenotyping work. Our scope is more expansive than existing scoping reviews as it covers any and all health conditions that have been explored as part of smartphone-based phenotyping. By exploring the breadth of health conditions studied by smartphone-based digital phenotyping, we can (1) show the versatility of the smartphone-based digital phenotyping to address the needs of a whole variety of health conditions, and (2) obtain a broader understanding of how smartphone-based digital phenotyping can be used across different clinical contexts. Further, in this work, in keeping with the spirit of digital phenotyping, we describe the relationship between smartphone data and behaviors that are exhibited in the presence of specific medical conditions for the works reviewed. We avoid going into algorithmic or machine learning (ML) methods in this scoping review, as that can sometimes promote the discovery without understanding the mindset when looking at prior work. As most of the studies reviewed are observational, the relationships reported between smartphone-derived behaviors and health conditions should be interpreted as associational rather than causal.

This scoping review aims to synthesize the existing literature on smartphone-only digital phenotyping. Specifically, we focused on studies that exclusively use smartphone sensors to characterize health conditions. Our objectives are to consolidate evidence across health domains, describe how behavioral signals are captured and assessed using different smartphone sensor streams, and to identify recurring methodological limitations and gaps in existing literature. By taking a cross-condition perspective, this review clarifies where evidence is concentrated, where it remains sparse, and how future research can better support replication, comparability, and translation into clinical and public health settings.

Methods

We conducted a scoping review to understand the current state of smartphone-based digital phenotyping. No review protocol was registered for this study. Methods adhered to PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [15]. The search strategy was reported in accordance with PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) guidelines [16]. We systematically queried Google Scholar, IEEE Xplore, ACM Digital Library, and PubMed for peer-reviewed papers published between 2012 and October 2025. The revised search strategy yielded approximately 3700 records across the 4 databases before screening. Search terms included "mobile sensing," "digital phenotyping," "smartphone sensing," "personal sensing," and "smartphone assessment." Terms were developed through discussion among this study's team, consisting of experts in digital phenotyping, digital health technologies, behavioral medicine, and computer science. A sample PubMed search string is as follows: ("digital phenotyping" OR "smartphone sensing" OR "mobile sensing") AND (mental OR behavioral OR health OR monitoring). Full search strategies for each database are provided in Multimedia Appendix 1. In addition to database searching, we used backward reference screening

XSL•FO
RenderX

of included papers to identify additional studies; we did not search study registries or gray literature. The search strategy followed PRISMA-S guidelines and was informed by prior reviews in digital phenotyping and mobile sensing. We did not use any standardized search filters.

All records retrieved from databases were exported or saved in PDF format from their respective databases and managed manually. Before formal screening, records were manually reviewed to remove duplicate entries and studies that were clearly outside the scope of the review. A substantial proportion of retrieved records were excluded at this stage due to the use of wearable or auxiliary sensors, non–health-related applications, review papers, study protocols, or a primary focus on algorithmic development rather than health characterization. Following this initial filtering step, 111 records remained and proceeded to formal title and abstract screening. This approach is consistent with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidance and is reflected in the PRISMA flow diagram.

Screening followed a 2-stage process consistent with PRISMA-ScR [15]. Two authors (AD and JH) independently screened a total of 111 retrieved PDF records for relevance to smartphone-based sensing and health-related outcomes. A total of 43 studies were identified from PubMed, 21 from IEEE Xplore, 26 from ACM digital library, and 21 from Google Scholar (N=111). After this initial title and abstract review, the same reviewers examined the full text of the remaining studies to confirm that they met the inclusion criteria, which required the use of smartphone-based sensing and a defined health-related outcome. Screening was performed manually, and any disagreements between reviewers were resolved through discussion until consensus was reached. No formal interrater reliability statistic (eg, Cohen k) was calculated, as the screening process was exploratory and descriptive in nature. Reasons for exclusion at each stage, including the use of auxiliary sensors, EMA-only studies, or the absence of health-related outcomes, are summarized in the Results section.

We included only English-language peer-reviewed papers that (1) primarily used smartphone-based sensor data as the input for digital phenotyping studies, (2) described the use of smartphone-based digital phenotyping with relation to a predescribed health condition. We excluded (1) non-English language literature, (2) conference abstracts, white papers, or other policy papers, (3) papers that described digital phenotyping through the use of smartphones as a platform to exclusively perform EMAs or other surveys without using on-board smartphone sensor data.

To find studies for this scoping review, we first determined the eligibility criteria for our studies. For a work to be considered within our scope for this scoping review, it must meet all three of the inclusion criteria listed here: (1) the studies must use digital phenotyping approaches that exclusively use smartphone-based sensor data as input; (2) the studies must not include approaches that exclusively use smartphones as a method for manual data entry, such as EMAs, surveys, etc; and (3) the studies must use digital phenotyping explicitly in the characterization of one or more well-defined health conditions. This includes studies that looked at monitoring the general well-being of an individual.

We did not include studies in this scoping review that solely focused on algorithmic advances or enhanced methods to capture data without analyzing it in the context of specific health conditions.

Using the keywords listed above, we identified 111 studies, of which 46 did not meet the inclusion criteria. This resulted in 65 studies that form the basis of this review. Table 1 shows the frequency of the health conditions represented. The most commonly studied condition was depression (n=18), followed by bipolar disorder (n=10) and stress (n=9). The least represented conditions were substance use disorder (cannabis; n=1) and attention-deficit/hyperactivity disorder (ADHD, n=1).

Before we proceed further, a note about terminology: in the rest of the paper we use the terms study and studies interchangeably. As each of the 65 studies we looked at in this review reported on 1 study, the 2 terms are synonymous in the context of this study.

XSL•FO
RenderX

Table . This table depicts the counts of health conditions considered by the studies in this scoping review. For studies that cover multiple health conditions, each condition is counted (ie, studies that consider both anxiety and depression are present in the counts for both conditions).

Condition	Count
Depression	18
Bipolar disorder	10
Stress	9
Schizophrenia	8
Anxiety	7
SUD[a] - alcohol	5
Parkinson disease	4
Sleep apnea	3
SUD - opioids	2
Well-being	2
ADHD[b]	1
SUD - cannabis	1

[a]SUD: substance use disorder.

[b]ADHD: attention-deficit/hyperactivity disorder.

Results

Overview

Before we delve into the details of the findings of our scoping review, we quickly present an overview of prior scoping reviews in the digital phenotyping space and argue why our work is different. We then delve into the findings of our scoping review. Figure 1 depicts a PRISMA flowchart of the review process.

XSL•FO
RenderX

Figure 1. PRISMA flow diagram illustrating the screening and selection process for studies included in this scoping review. The initial search across 4 databases (Google Scholar, IEEE Xplore, ACM Digital Library, and PubMed) identified 111 records published between January 1, 2012, and October 31, 2025. After the removal of nonrelevant records during title and abstract screening, 65 full-text papers met the inclusion criteria for smartphone-based digital phenotyping using on-device sensor data and a defined health-related outcome. Studies were excluded for reasons including use of auxiliary sensors or wearables as primary data sources, EMA-only designs without sensor data, absence of a health-related condition or outcome, or non–peer-reviewed publication type. The final set of 65 studies formed the basis of the analysis presented in this review. EMA: ecological momentary assessment; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

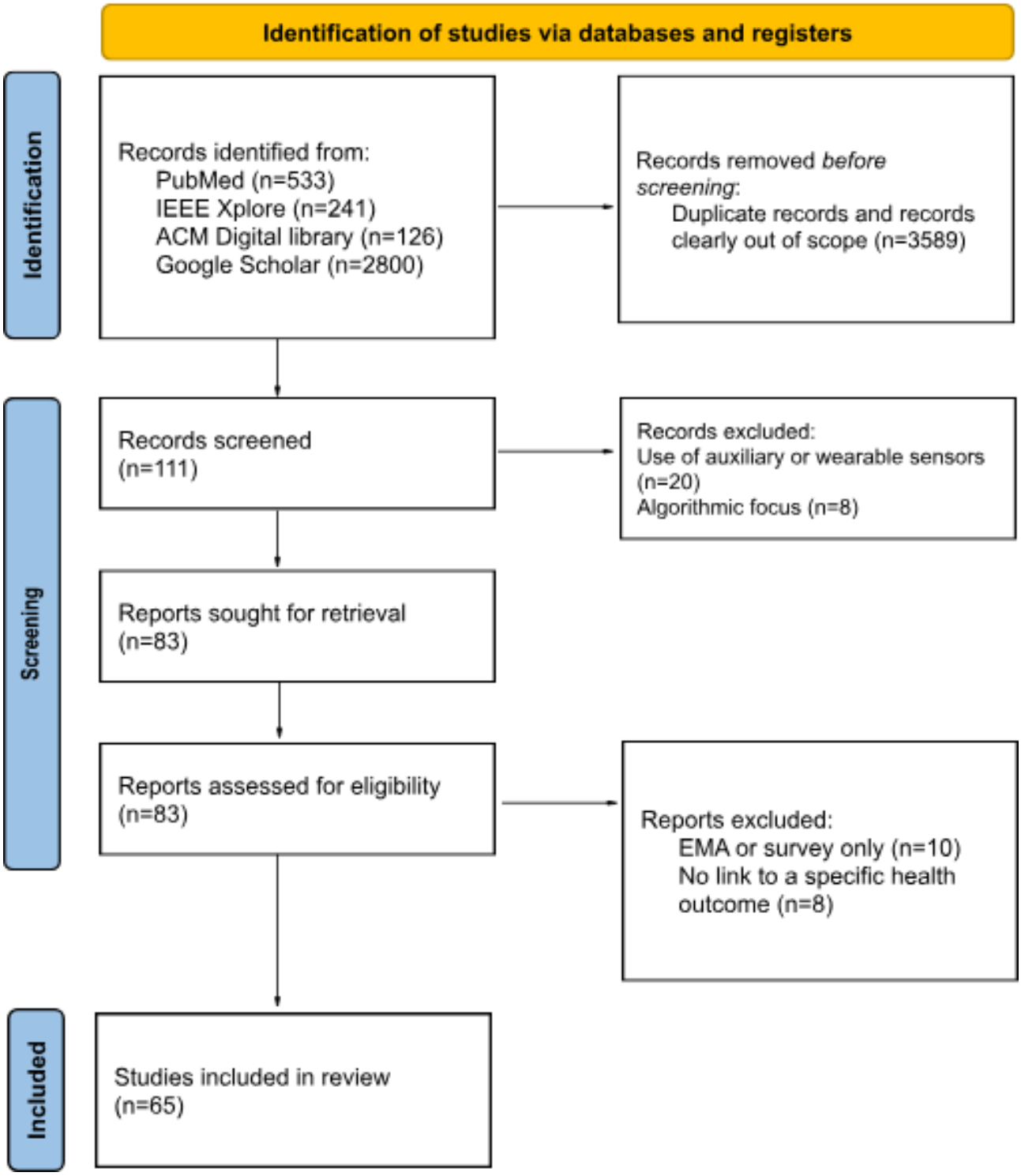

Prior Scoping Reviews

There are existing reviews that have explored the use of digital phenotyping. Our search yielded 21 prior reviews. We classify these reviews based on the types of sensing technology used: smartphones, smartwatches, and auxiliary sensors. We define auxiliary sensors to be any sensing device outside of smartphones and smartwatches, such as smart textiles or sensors worn on the body. A brief overview of the reviews reported on in this section can be found in Table 2, which we now elaborate on. These prior reviews were not part of the review process but

XSL•FO
RenderX

are only presented to show the distinction between the existing reviews and our scoping review.

Table . Summary of existing digital phenotyping reviews (n=19). This table summarizes prior reviews and reviews examining digital phenotyping across different sensing modalities and health domains. Each citation lists the number of studies covered, primary sensing technologies (eg, smartphones, smartwatches, or auxiliary sensors), publication year, and main health focus.

Citation	Studies (n)	Sensing modalities used	Year	Area of interest
[17]	119	Smartphones	2018	Health and well-being
[18]	18	Smartphones	2019	Bipolar disorder
[19]	224	Smartwatches	2020	Presurgical risk assessment, postsurgical outcome prediction
[20]	62	Smartphones, smartwatches, auxiliary sensors	2021	Bipolar disorder
[21]	92	Smartphones, smartwatches, auxiliary sensors	2021	Depression, bipolar disorder, anxiety, stress
[22]	51	Smartphones, smartwatches	2022	Depression
[23]	46	Smartphones, smartwatches, auxiliary sensors	2022	Machine learning practices
[24]	31	Smartphones, smartwatches	2022	Available datasets and smartphone apps
[25]	81	Smartphones, auxiliary sensors	2023	Parkinson disease, Alzheimer disease
[26]	11	Smartphones, smartwatches, auxiliary sensors	2023	Anxiety
[14]	40	Smartphones	2023	Schizophrenia, depression, general well-being
[27]	29	Smartphones, smartwatches	2023	Mood disorders, schizophrenia-spectrum
[28]	47	Smartphones	2023	Social behavior sensing
[29]	59	Smartphones, smartwatches	2023	Depression, anxiety, stress, schizophrenia, mood, general well-being
[30]	35	Smartphones	2024	Youth depression and anxiety
[31]	24	Smartphones	2024	Machine learning practices
[32]	40	Smartphones	2024	Depression, stress, anxiety
[33]	112	Smartphones	2025	Digital phenotyping pipeline
[34]	16	Smartphones, smartwatches, auxiliary sensors	2025	Stress monitoring
This scoping review	65	Smartphones	2025	Depression, bipolar disorder, schizophrenia, anxiety, stress, general well-being, substance use, sleep apnea, Parkinson disease

Earlier reviews predominantly examined mental health outcomes such as depression, bipolar disorder, and stress, whereas more recent work expanded to neurological, physiological, and multimodal applications. The final row (gray) highlights the present scoping review, which uniquely synthesizes 65 smartphone-only studies spanning 9 health domains. About a third of the reviews (n=7) focus exclusively on using smartphones for digital phenotyping. Three of these reviews focus on describing phenotyping in the context of specific health conditions: depression [31], bipolar disorder [18], and a combination of depression, stress, and anxiety [32]. The remaining reviews include a scoping review of 35 studies on depression and anxiety in youth [30], a review of 47 studies using smartphones for social behavior sensing [28], a large review of 112 studies characterizing digital phenotyping pipelines [33], and a systematic review of 119 studies on

XSL•FO
RenderX

smartphone-based health sensing [17]. Additionally, Bufano et al [27] report on 29 clinical studies using smartphones for psychiatric relapse monitoring.

Another third of the reviews (n=6) cover a combination of smartphone and smartwatch-based digital phenotyping. Similar to the previous category, 2 of the studies focus on phenotyping work addressing mental health conditions: depression [22] and depression, anxiety, stress, schizophrenia, mood, and general well-being [29]. One review covers a purely smartwatch-based digital phenotyping to assess presurgical risk and postsurgical care [19]. The last review in this category reports on publicly available applications and datasets for digital phenotyping that use both smartphones and smartwatches [24].

The remaining reviews (n=8) report on digital phenotyping using a combination of smartphones, smartwatches, and auxiliary sensors. Like the prior 2 sections, the reviews in this category focused on mental health conditions: Alzheimer disease [25]; depression, bipolar disorder, anxiety, and stress [21]; bipolar disorder [20]; and anxiety [26]. The studies by Dlima et al [23] and Lee et al [14] focus on the use of ML to interpret phenotyping data, while Bufano et al [27] examine combined sensing modalities in psychiatric relapse studies. Most recently, Kallio et al [34] provide a review of approximately 16 studies focused on continuous stress monitoring in cognitively demanding work environments using smartphones, smartwatches, and ambient sensors. Their review emphasizes real-world deployment, unobtrusiveness, and user acceptance, expanding the digital phenotyping literature into long-term occupational health monitoring.

Most of these reviews (n=15) cover research that uses more sensing modalities than those available on smartphones. Those that do cover purely smartphone-based digital phenotyping (n=7) are scoped in terms of specific mental health conditions, youth populations, or derived behavioral features.

Our review is different from these prior reviews for three reasons: (1) it focuses exclusively on smartphone-based digital phenotyping; (2) its scope is much more expansive, in that it covers any and all health conditions (and not just mental health conditions); and (3) it focuses on describing the relationships between behaviors captured through phenotyping in the context of specific health conditions and not on algorithms used to extract the behaviors.

Findings of This Scoping Review

Our aim is to understand the current state of smartphone-based digital phenotyping search which led to the identification of 111 records. After title or abstract screening, 65 studies met the inclusion criteria and proceeded to full-text review; all 65 were included in the synthesis. Across the 65 included studies, we conducted a descriptive synthesis to summarize the sample sizes, study durations, and participant populations. The median sample size across all studies was 52 participants (IQR=26 - 126), with a median study duration of 8 weeks (IQR 3 wk to 6 mo). Studies commonly used multiple different measurements for ground truth. Clinical rating scales (eg, Patient Health Questionnaire-9, Hamilton Depression Rating Scale, YMRS, Generalized Anxiety Disorder-7, and Positive and Negative Affect Schedule) were used in 63% (41/65) studies, EMAs in 28% (18/65) studies, clinician-confirmed diagnoses in 14% (9/65) studies, and physiological validation measures (eg, polysomnography or gait laboratory testing) in 5% (3/65) studies. Each of these studies was additionally described by the study setting. Controlled laboratory studies are conducted under precise conditions. Naturalistic “daily life” studies capture behavior in real-world environments. Hybrid study designs intentionally combine both settings within the same study [35].

One way of examining smartphone-based digital phenotyping research is by considering the conditions that studies target. These works can be grouped into 6 broad categories: mood disorders, stress-related disorders, schizophrenia, disorders from the use of psychoactive substances, nervous system disorders, and general well-being management.

In all of these efforts, the smartphone sensors were used to collect a variety of contextual information about the individual, such as call or text logs, phone or app usage, location, accelerometer, and microphone data. These were used to extract phenotypes, such as social functioning, physical activity, movement patterns, and sleep patterns, which were associated with the specific health condition being considered. These phenotypes are then analyzed using some form of ML algorithm to answer questions about specific health conditions. To build such models, knowing the ground truth of the health condition of the participants is essential. This ground truth was usually assessed using EMAs or an appropriate scale based on the condition being tracked. In our description below, we do not focus on the intricacies of the ML algorithms used, as they are largely similar across most of the studies. Rather, as much as possible, we focus on the explanations that these models offer about the relationship between sensing, phenotypes, and the health condition being considered.

For each health condition we discuss in this section, we categorize the work done in one or more of the following focus categories: monitoring, diagnosis, and intervention. Monitoring involves collecting and analyzing smartphone-based digital phenotypes to track the health condition being considered. Diagnosis uses smartphone-based digital phenotypes to identify the presence of the health condition or risks thereof. Intervention involves implementing strategies or treatments, often as a response to monitoring a health condition. Not every health condition is worked on in all 3 categories.

Mood Disorders

Mood is defined as a pervasive and sustained feeling that is experienced internally [36]. It impacts nearly all aspects of a person’s behavior externally. Mood disorders are described by marked disruptions in emotions. For example, experiencing severe lows called depression or highs called hypomania or mania [36]. Though there are a variety of mood disorders, smartphone-based digital phenotyping work in the context of mood disorders has targeted only 2 major conditions in this regard: depression and bipolar disorder, which we now describe.

Depression

Depression (otherwise known as depressive disorder and major depressive disorder) involves low moods and a loss of pleasure

XSL•FO
RenderX

or interest in activities over a prolonged period of time [37]. Over the years, studies have shown that passive sensor data collected from smartphones can be associated with depression [38-40]. Much of the digital phenotype effort to explore and characterize depression has focused on 4 types of phenotyping efforts: tracking social interactions, tracking location diversity, vocal characteristics, and tracking sleep quality. For these approaches, the ground truth was collected via EMA or self-reported using pertinent scales, such as the Personal Health Questionnaire [41], the Center for Epidemiological Studies-Depression Scale [42], the Children's Depression Rating Scale-Revised [43], and the Clinical Global Impressions-Severity Scale [44]. Phenotyping work done in the depression space was primarily focused on the monitoring of depression, with some work in the diagnosis focus category. Table 3 provides a summary of the studies reviewed in this section.

XSL•FO
RenderX

Table . Summary of smartphone-based digital phenotyping studies focused on depression (n=16). Studies examined depressive symptoms and mood-related behaviors through passive smartphone sensing. Common phenotypes included sociability, mobility, activity, and sleep patterns. Sensors most frequently used were GPS, accelerometer, screen state, and communication logs, with PHQ-9[a] as the most common ground-truth measure.

Citation	Condition	Phenotypes considered	Sensors used	Ground truth	Participants (n)	Duration	Focus	Study setting
[38]	Depression	Sociability, movement patterns	App use, gyroscope	PHQ-9	558 adults	10.7 days average	Monitoring	Naturalistic setting (app recruitment)
[39]	Depression	Movement patterns	GPS, app logs, call logs, text logs, IMU[b]	PHQ-9	38 adults	2 years	Monitoring	Naturalistic setting (Spain)
[40]	Depression	Movement	GPS, weather, light	DASS[c], PANAS-X[d]	31 adults	2 weeks	Monitoring	Naturalistic setting (United States)
[45]	Depression	Sociability	App use, call logs, GPS, battery, screen state	CES-D[e]	120 adults	1 month	Monitoring	Naturalistic setting (China)
[46]	Depression	Activity, sociability	GPS, screen state	PHQ-9	40 adults	2 weeks	Monitoring	Naturalistic setting (United States)
[47]	Depression	Sleep, activity, sociability	Call logs, accelerometer, light, GPS, screen state	PHQ-9	47 adults with diabetes	20 weeks	Monitoring	Naturalistic setting (India)
[48]	Depression	Sociability, activity, movement	Call or text logs, GPS, accelerometer, light, screen state	PHQ-9, HAM-D[f], HAM-A[g]	13 adolescents	8 weeks	Monitoring	Naturalistic setting (United States)
[49]	Depression	Sociability, mobility	GPS, gyroscope, call or text logs	CDRS-R[h]	24 adolescents	8 weeks	Monitoring	Naturalistic setting (South Korea)
[50]	Depression	Activity, movement	GPS	PHQ-9	48 college students	10 weeks	Monitoring	Naturalistic setting (United States)
[51]	Depression	Movement patterns	GPS	PHQ-8[i]	28 adults	2 months	Monitoring	Naturalistic setting (app store recruitment)
[52]	Depression	Activity, sociability	Accelerometer, GPS, WiFi, call or text logs	PHQ-9	126 adults	9 months	Monitoring	Naturalistic setting (app store recruitment)
[53]	Depression	Movement patterns	Accelerometer, GPS	PHQ-9	33 adults	11 weeks	Monitoring	Naturalistic setting (Bangladesh)
[54]	Depression	Activity, sleep disruption	GPS, accelerometer, logs, screen state	Self-reported mood	100 college students	28 days	Monitoring	Naturalistic setting (United States)
[55]	Depression	Sociability	Microphone, text, GPS	PHQ-9	302 college students	6 months	Monitoring, diagnosis	Hybrid setting (United States)
[56]	Depression	Psychological stress	Microphone	BDI[j]	1814 adults	2 years	Monitoring, diagnosis	Naturalistic setting (Japan)

XSL•FO
RenderX

Citation	Condition	Phenotypes considered	Sensors used	Ground truth	Participants (n)	Duration	Focus	Study setting
[57]	Depression	Sociability	Microphone, call or text logs, GPS	PHQ-9, GAD-7[k]	70 college students	28 weeks	Diagnosis	Hybrid setting (United States)

[a]PHQ-9: Patient Health Questionnaire-9.

[b]IMU: Inertial measurement unit.

[c]DASS: Depression Anxiety Stress Scales.

[d]PANAS-X: Positive and Negative Affect Schedule—Expanded Version.

[e]CES-D: Center for Epidemiologic Studies Depression Scale.

[f]HAM-D: Hamilton Depression Rating Scale.

[g]HAM-A: Hamilton Anxiety Rating Scale.

[h]CDRS-R: Children's Depression Rating Scale-Revised.

[i]PHQ-8: Patient Health Questionnaire-8.

[j]BDI: Beck Depression Inventory.

[k]GAD-7: Generalized Anxiety Disorder-7.

Concerning monitoring, one of the most common ways of characterizing depression and depressive episodes is by looking at changes in social interaction. The social interaction of a person can be approximated by measuring the number of incoming calls a person receives as well as the amount of time people spend on social media apps [45]. It was found that a lower volume of incoming calls and a higher frequency of social media app use were indicative of depression, indicating that individuals who were depressed were not communicating with others and were instead turning to their phones as a potential form of self-soothing [45]. Similar findings on the presence of inverse correlation between number and length of phone calls, text messages, and overall phone use patterns and depression were found in 4 different studies [46-48,55]. However, in 1 study, the authors looked at an adolescent population in South Korea (as opposed to adults in the aforementioned work) and found that those with depression received more calls compared to the control participants [49]. Though the study did find that a person spending more time on smartphones may be associated with depression, as mentioned above.

Phone call and text logs as a proxy for social connection were not the only way to infer and characterize depression. Another common approach was to look at location features (based on the GPS and inertial measurement unit). It has repeatedly been found that movement patterns may be a similarly strong proxy to mental state [50,51]. A case in point is a study that looked at individuals with spinal cord injuries. For such individuals, there was a direct correlation between a decrease in the number of locations visited (presumably due to their condition) and depression [58]. This was found to be true not just for adults but also for younger populations as well [48,59]. A similar pattern was also reported in other studies that showed that the variance, number of location clusters, and location entropy all decreased in relation to increased depression scores [50,51]. When the location data was analyzed by making assumptions about the location context, such as duration of time away from home, the information could be used to provide interventions for individuals with depression [52]. Beyond just diversity of location, in 1 study, the researchers found that physical activity in general was associated with depression, with lower physical activity (in terms of distance variance from any starting point, greater resting time, greater time using the phone in a lying position, fewer steps, less exercise, and greater staying at home) being associated with increased depression [53].

A few studies have also looked at using sleeping patterns to characterize depression. Measuring sleep quality from smartphones, unlike wearable devices (eg, smartwatches), is not straightforward. Consequently, researchers often make broad assumptions about the daily routines of a typical person and check to see if smartphone usage patterns can detect a break in this routine. For instance, in one of these studies, a higher number of instances of "screen on" states recorded from 12 AM to 6 AM was interpreted as indicating a lower sleep quality or disturbed sleep schedule in individuals [47]. This pattern of being up at night was found to be associated with the presence of depression. In a similar study on college students, it was found that higher variance in sleep patterns was associated with depression [54].

Concerning diagnosis, in recent years, an interesting and different approach to detecting the presence of depression in individuals has been explored using vocal characteristics. In 2 different studies, the authors looked at changes in vocalization and tone based on features such as vocal fry, vocal jitter, and monotony measured via unscripted and scripted audio recordings to determine the presence of depression [55,57]. A similar study also analyzed voice features during phone calls to determine the presence of depression [56]. In one of the aforementioned studies, the researchers tried to extend their vocal characteristics-based depression models to the most extreme cases where individuals are experiencing suicidal ideation [57]. They found that such models were not sufficient and required additional context to predict suicidal risk [57].

Bipolar Disorder

Bipolar disorder is a mental health condition that causes individuals to experience extreme mood swings, typically between emotional highs (known as mania or hypomania) and lows (depression) [60]. There are several types of bipolar and

XSL•FO
RenderX

related disorders, not all of which require for both manic and depressive episodes to be present, but all of which require clear changes in mood, energy, and activity levels [60]. Though very similar to the phenotypes used in characterizing depression, the way they are used for characterizing bipolar disorder is somewhat different. As discussed in the prior section, we know that it is possible to identify depression through a variety of methods, but it is important to note that depressive episodes in bipolar disorder may not be as intense or last as long [60]. In addition, mixed states (having depressive symptoms and manic symptoms at the same time) and manic episodes are just as important to recognize in individuals with bipolar disorder, meaning that while some techniques may be useful between both disorders, they are not the same disorder and cannot be treated as such. With respect to smartphone-based digital phenotyping, considerable work has also focused on exploring diseases such as bipolar disorder. These efforts have primarily looked at 4 types of phenotyping efforts: tracking phone usage, tracking location diversity and routine, and tracking sleep quality. Further, the ground truth for bipolar disorder was collected via EMA or self-reported using a combination of scales such as Hamilton Depression Rating Scale [61] and YMRS [62]. Phenotyping work in bipolar disorder was primarily focused on the monitoring category, with some work for both diagnosis and intervention. Table 4 provides a summary of the research reviewed in this section.

XSL•FO
RenderX

Table . Summary of smartphone-based digital phenotyping studies focused on bipolar disorder (n=11). Studies used passive smartphone sensors to assess behavioral and physiological markers associated with bipolar disorder. Commonly examined features included sociability, activity levels, and sleep-related changes. Ground-truth measures primarily relied on clinical scales such as HAM-D[a], YMRS[b], and QIDS-SR[c].

Citation	Condition	Phenotypes considered	Sensors used	Ground truth	Participants (n)	Duration	Focus	Study setting
[63]	BD[d]	Physical activity, sociability	GPS, cell tower pings, accelerometer, call logs, text logs	Daily self-reported mood, HAM-D, YMRS	13 adults with BD	12 months	Monitoring	Naturalistic setting (Germany)
[64]	BD	Sociability	Accelerometer, GPS, call data	HAM-D, YMRS	10 adults with BD	12 weeks	Monitoring	Naturalistic setting (Austria)
[65]	BD	Physical activity, sociability	Screen state, text logs, call logs, cell tower pings	HAM-D, YMRS	29 adults with BD	12 weeks	Monitoring	Naturalistic setting (Denmark)
[66]	BD	Sociability	Call logs, text logs	Self-reported mood, YMRS, HAM-D	26 adults with BD, 12 controls	28 days	Monitoring	Naturalistic setting (United States)
[67]	BD	Movement, activity, sociability	Accelerometer, GPS, WiFi, Bluetooth, call logs	Mood self-report, HAM-D, MSS[e], ADS[f]	10 inpatients with BD	8 weeks	Monitoring	Naturalistic setting (Austria)
[68]	BD	Activity, movement patterns	GPS, WiFi, cell tower pings	QIDS-SR	29 adults with BD, 20 controls	3 months	Monitoring	Naturalistic setting (United Kingdom)
[69]	BD	Movement, sleep disruption	GPS	ASRM[g], DASS-21[h]	159 adults with BD	1 year	Monitoring	Hybrid setting (Taiwan)
[70]	BD	Activity, sociability, movement	Accelerometer, microphone, battery, screen state, logs, calls	SRM-5[i]	7 adults with BD	4 weeks	Monitoring	Naturalistic setting (Germany)
[71]	BD	Sociability	Call or text logs, microphone, screen state, accelerometer, cell tower pings	Mood self- report	28 adults with BD	12 weeks	Monitoring, diagnosis	Naturalistic setting (United States)
[72]	BD	Sleep, activity	Accelerometer, light sensor	PSQI[j]	22 adults with BD, 23 controls	3 weeks	Intervention	Hybrid setting (Austria)
[73]	BD	Sociability, stress, sleep	Accelerometer, WiFi, logs, screen state, Bluetooth, microphone	Mood self-report, HAM-D, YMRS	18 adults with BD	5 months	Monitoring, diagnosis, intervention	Naturalistic setting (app store recruitment)

[a]HAM-D: Hamilton Depression Rating Scale.

[b]YMRS: Young Mania Rating Scale.

[c]QIDS-SR: Quick Inventory of Depressive Symptomatology—self-report.

XSL•FO
RenderX

[d]BD: bipolar disorder.

[e]MSS: Mania Self-Rating Scale.

[f]ADS: Common Depression Scale.

[g]ASRM: Altman Self-Rating Mania Scale.

[h]DASS-21: Depression Anxiety Stress Scales-21.

[i]SRM-5: Social Rhythm Metric Scale-5.

[j]PSQI: Pittsburgh Sleep Quality Index.

Concerning monitoring, like tracking depression, several efforts in recent years have focused on tracking a bipolar individual's mood over time. This information can be used to determine if a person is having a depressive or manic episode. In this regard, one of the common phenotypes is smartphone use pattern [64,65,71,73]. In 1 study, the researchers used features of phone calls themselves, such as total length, average call length, number of unique calls, total number of calls, etc, along with sounds emanating during a phone call to determine the current mood or change in moods of bipolar individuals [64]. The study indicated the presence of some kind of association between bipolar states and these phone usage features, but those exact associations were not clear. In 2 other similar studies, it was found that severe depressive symptoms of bipolar disorder were associated with longer "screen on" state per day, more received calls per day, fewer outgoing calls per day, fewer answered incoming calls per day, and less movement [65,66]. On the other hand, manic symptoms showed the exact opposite pattern as described above.

Another approach to characterizing bipolar disorder has been to look at movement or physical activity using smartphone-based GPS data [67-69]. In another study, researchers found that physical activity changes as mood states transition from one to another [67]. When transitioning from a manic state to a baseline normal state, physical activity decreases, but it increases when transitioning from a depressed state to baseline [67]. In a study focused on quantifying depressive states in those with bipolar disorder, the researchers derived a variety of features from GPS data from an individual's smartphone and were able to use them to predict acute depressive states with high accuracy [68].

Concerning diagnosis, in the context of smartphone-based bipolar disorder diagnosis, researchers looked at vocal patterns along with self-reported surveys on a variety of mood and sleep patterns, along with smartphone usage metadata from individuals with bipolar disorder to detect manic and depressive states [71]. Overall, it was found that while it was possible to classify bipolar states on voice features alone, combining different passively sensed and self-report data with voice features creates a more robust model in terms of detecting subtler manifestations of bipolar states (eg, hypomania as opposed to mania) [71]. Another potential indicator of a bipolar person's mood is in patterns of app use on the smartphone by a bipolar individual [73]. In particular, the way that bipolar individuals interact with their phones has a strong correlation to what moods they are experiencing at the time. For instance, in 1 study, the researchers found that increased social interaction is associated with a decrease in self-reported stress level, irritability, and mixed mood. This indicated that increased social interaction is a good marker for normal social functioning outside of mood swings. Further, patterns in the average number of apps being used can be used to detect the presence of manic or depressive episodes, with a higher number of apps used being positively associated with self-reported stress, and a decrease in the average number of apps used was suggested to indicate a depressive episode [73].

Concerning intervention applications, finally, sleep disturbances have been shown to be early warning signs of manic or depressive episodes in individuals with bipolar disorder [72]. As stated above, smartphone use (screen on or off, and accelerometer readings) can provide insight into lower sleep quality and satisfaction with sleep, alluding to potential sleep disturbances [73]. In a different study, the researchers showed that a smartphone-based app could be a valid tool for measuring sleep onset and wake-up time, to measure sleep disturbances. This information can then be used for the management of the depressive symptoms of bipolar disorder [72].

Stress-Related Disorders

Stress is the perception of an individual that their environmental taxes them or exceeds their adaptive capacity [74]. Stress-related disorders are maladaptive responses to short- or long-term exposures to physical or emotional stressors. Smartphone-based phenotyping work for stress-related disorders can be broadly categorized into 2 groups: anxiety disorders and stress quantification. We describe these below. Table 5 provides a summary of the reviewed work in this section.

XSL•FO
RenderX

Table . Summary of smartphone-based digital phenotyping studies focused on anxiety, stress, and schizophrenia (n=18). Studies span sociability, movement or activity, sleep, and psychological stress phenotypes using signals such as accelerometer, GPS, microphone, app logs, and communication metadata. Ground-truth measures include standardized scales, self-reports, and clinical or laboratory validations.

Citation	Condition	Phenotypes considered	Sensors used	Ground truth	Participants (n)	Duration	Focus	Study setting
[75]	Anxiety	Sociability, movement patterns	Accelerometer, call logs, text logs	SIAS[a], DASS-21[b], PANAS[c]	59 college students	2 weeks	Monitoring	Naturalistic setting (United States)
[76]	Anxiety	Sociability, physical activity	App log, power level, screen, GPS, call or text logs	SIAS, UCLA-LS[d]	127 adults	30 days	Monitoring	Naturalistic setting (China)
[77]	Anxiety	Sociability, physical activity	Light, accelerometer, gyroscope, app logs	STAI[e]	20 healthy adults	30 days	Monitoring	Hybrid setting (Japan)
[78]	Stress	Movement, activity, sociability, psychological stress	Accelerometer, GPS, microphone, WiFi, call or text logs, app logs	Mood self-report	30 employees	8 weeks	Monitoring	Naturalistic setting (Italy)
[79]	Stress	Sleep, activity, sociability	Microphone, GPS, WiFi, accelerometer, light sensor	PSS[f], PHQ-9[g], UCLA-LS	47 young adults	10 weeks	Monitoring	Naturalistic setting (United States)
[80]	Stress	Physical activity, sociability	Accelerometer	OLBI[h]	30 employees	8 weeks	Monitoring	Naturalistic setting (Italy)
[81]	Stress	Sociability	Call or text logs, Bluetooth	Stress self-report	117 graduate students	7 months	Monitoring	Naturalistic setting (United States)
[82]	Stress	Activity, sociability, sleep, movement	Accelerometer, microphone, light, GPS, Bluetooth	PHQ-9, PSS, UCLA-LS	48 students	10 weeks	Monitoring	Naturalistic setting (United States)
[83]	Stress	Activity, sociability, stress, movement	GPS, WiFi, microphone, accelerometer	GPA[i], stress or mood self-report	48 students	10 weeks	Monitoring	Naturalistic setting (United States)
[84]	Stress	Psychological stress	Keyboard pressure	Mood self-report, event context	11 adults	1 laboratory session	Diagnosis	Controlled laboratory setting (United States)
[85]	Schizophrenia	Sociability	Call logs, text logs, microphone	BPRS[j], hospital relapse records	61 adults with schizophrenia	1 year	Monitoring	Naturalistic setting (United States)
[86]	Schizophrenia	Activity, sociability	Call logs, text logs, microphone	Social function diary	5 adults with schizophrenia	5 days	Monitoring	Naturalistic setting (location not stated)

XSL•FO
RenderX

Citation	Condition	Phenotypes considered	Sensors used	Ground truth	Participants (n)	Duration	Focus	Study setting
[87]	Schizophrenia	Movement patterns	GPS	Self-reported activity or context	85 adults with schizophrenia, 58 controls	7 days	Monitoring	Naturalistic setting (United States)
[88]	Schizophrenia	Sociability	GPS, accelerometer, screen, call or text logs	PHQ-9, GAD-7[k], self-reported symptoms	45 adults with schizophrenia, 43 controls	3 months	Diagnosis	Naturalistic setting (United States)
[89]	Schizophrenia	Movement, sociability	Accelerometer, GPS, call or text logs, screen, battery	Symptom report	17 adults with schizophrenia	3 months	Diagnosis	Naturalistic setting (United States)
[90]	Schizophrenia	Movement, activity, sociability, sleep	Accelerometer, GPS, app logs, light, call or text logs	Mood self-report	21 adults with schizophrenia	8 months	Diagnosis	Naturalistic setting (United States)
[91]	Schizophrenia	Activity level	Accelerometer	Self-reported activity	50 adults with schizophrenia, 70 controls	6 days	Diagnosis	Naturalistic setting (United States)
[92]	Schizophrenia	Sleep quality	GPS, screen, call or text logs, accelerometer	PSQI[l], laboratory sleep assessment	17 schizophrenia	3 months	Diagnosis	Naturalistic setting (United States)

[a]SIAS: Social Interaction Anxiety Scale.

[b]DASS-21: Depression Anxiety Stress Scales-21.

[c]PANAS: Positive and Negative Affect Schedule.

[d]UCLA-LS: University of California, Los Angeles Loneliness Scale.

[e]STAI: Spielberger State-Trait Anxiety Inventory.

[f]PSS: Perceived Stress Scale.

[g]PHQ-9: Patient Health Questionnaire-9.

[h]OLBI: Oldenburg Burnout Inventory.

[i]GPA: Grade Point Average.

[j]BPRS: Brief Psychiatric Rating Scale.

[k]GAD-7: Generalized Anxiety Disorder-7.

[l]PSQI: Pittsburgh Sleep Quality Index.

Anxiety Disorder

Anxiety disorders usually involve a persistent feeling of anxiety or dread, which can interfere with daily life. People living with such disorders experience frequent anxiety for extended periods of time [93]. For these efforts, the ground truth was collected via EMA or self-reported using pertinent scales, such as the General Anxiety Disorder Scale (Generalized Anxiety Disorder-7) [94]. There were relatively fewer studies in this space, all of which focused on monitoring.

Concerning monitoring, in 1 study, researchers showed that social function derived based on basic statistics collected about phone calls made, received, missed, diversity of callers, etc, can be used to detect the presence of social anxiety disorder in an individual [75]. Other studies found that lack of social connection in terms of lower call use and increased social media application use [76], and use of phones in the dark [77] were associated with higher levels of anxiety disorders. Further, days of the week may also be important in recognizing anxiety, as self-reported anxiety appears to increase toward the end of the weekend and decrease toward the end of the week, possibly aligning with perceived stress or uncertainty about workloads during the week [77]. Overall, the results for characterizing anxiety disorders were very similar to characterizing depression, such that discerning between one and the other may be challenging. For instance, researchers showed a weak and positive relationship between the predicted social anxiety disorder symptom and depression when capturing social functioning of an individual based on phone call usage statistics [75].

XSL•FO
RenderX

Stress Quantification

Psychological stress is another condition that has been targeted as part of smartphone-based digital phenotyping research. Though experiencing psychological stress is not an uncommon condition for most individuals, prolonged or significant stress has been linked to a variety of physical and mental illnesses [78]. Prolonged stress may create a space for burgeoning mental illnesses such as depression and anxiety to flourish [37,78,79]. Characterizing stress typically relies on the fusion of a variety of behavioral information about an individual, such as their movement patterns, physical activity, social interaction, and sleep patterns, much like works described in the section above. Phenotyping work related to stress was in the monitoring focus category.

Concerning monitoring, in looking at monitoring stress, researchers looked at automatic stress detection in completely unconstrained environments. The study showed that it was possible to extract time and frequency domain features from accelerometer data to characterize routine and associate these features with self-reported stress levels [80]. In a similar study, the researchers explored more complex fusion models where the relationship between self-reported daily stress and sensor-derived activity changes, speech duration, changes in location, and sleep patterns was associated with changes in participants' stress levels over the course of the study [78]. In a similar work, the authors found that speech duration, movement information (measured using GPS and WiFi), and sleep duration were associated with self-reported daily stress levels, while their activity information (measured using an accelerometer) was associated with changes in their subjective loneliness [79]. Another approach to measuring stress proposed the fusion of information from call and text logs (this included observing broader patterns from calls or text, such as percentage of calls done during the night, percentage of initiated calls during the night, text response rate, text response latency, and the percentage of initiated texts) combined with proximity information (collected from Bluetooth signals from neighboring devices) along with weather information (temperatures, precipitation, humidity, visibility, etc) to accurately predict the presence of stress [81].

All these aforementioned works focused on work-related stress. Another group of individuals whose stress has been studied in some detail is college students. In 1 study, the researchers developed an app called StudentLife for continuously sensing the day-to-day impact of college workload on stress, sleep, activity, mood, sociability, mental well-being, and academic performance among a single class of 48 students across a 10-week term [82]. Analysis of the data collected shows that students at the start of the term had high positive affect and conversation levels, low stress, and healthy sleep and daily activity patterns. However, as the term progressed and the workload increased, stress appreciably increased while positive affect, sleep, conversation, and activity dropped off [82]. A follow-up study looked at a more general approach to stress quantification in college students, using a geo-fencing approach on the GPS data from the smartphones carried by students [83]. The various spaces students frequented were designated as "study spaces" and "party locations"; and in each of these locations, the student's behavior is recorded based on smartphone activity and audio collected from the phone's microphone. The results indicated that students who reported decreasing stress levels throughout the term (self-reported) obtained higher GPAs [83].

Finally, researchers used keyboard pressure as a proxy for detecting the presence of stress. Their idea was that the increase in muscle tension in response to an individual experiencing stress would lead to increased keyboard pressure when typing. In this study, the researchers asked individuals to recall a stressful and relaxing event in the recent past and write about it using an iPhone with a modified QWERTY keyboard. They then logged the timing, position, and pressure values associated with each keyboard interaction. Subsequently, the differences in typing pressure across relaxed and stressful expressive writing tasks were associated with differences in self-reported stress [84].

Delusional Disorders

A delusion is a false belief based on an inaccurate interpretation of an external reality despite evidence to the contrary [95]. In the context of delusional disorders, smartphone-based phenotyping work has exclusively focused on schizophrenia. Schizophrenia is a significant and often life-altering mental disorder characterized by episodes, which include persistent delusions and/or hallucinations, disorganized thinking and behaviors, and instances of agitation or negative symptoms. The term negative symptoms describes a lessening or absence of normal behaviors and functions related to motivation and interest, or verbal or emotional expression, which can manifest as blunted affect, reduction in speech, goal-directed activities, motivation, and experience of pleasure [96]. We summarize our findings in this space below. Table 5 provides a summary of the reviewed research in this section.

Schizophrenia

Typically, smartphone-based digital phenotyping schizophrenia research has focused on identifying predictors of negative symptoms. This is done by tracking data such as call patterns, location, physical activity, and sleep quality. The ground truth in most of these studies was collected via EMA. Phenotyping work with respect to schizophrenia space was in both monitoring and diagnosis focus categories.

Concerning monitoring, relatively few works have looked into applying phenotyping to people with schizophrenia. In 1 study, the researchers used patterns in incoming and outgoing calls and their duration, incoming and outgoing text messages, and captured the frequency and duration of nearby voices captured through a smartphone microphone to determine exactly how much social activity an individual with schizophrenia was engaging in. Overall, the researchers reported that reductions in the number and duration of outgoing calls, as well as the number of text messages, were associated with relapses [85].

In examining negative symptoms of schizophrenia, researchers explored the feasibility of clustering GPS location as a proxy for social functioning and determining where individuals with schizophrenia are spending their time. Subsequently, they added contextual labels to the location clusters to attain a general

XSL•FO
RenderX

understanding of social functioning for a specific individual [86]. In a similar study, researchers looked at GPS-derived indicators of location to determine if they could be used as a proxy for negative symptoms. They found that GPS-derived estimations of average distance traveled over time, distance from home, and the percentage of GPS samples at home were found to be highly associated with negative symptoms. Essentially, they found that less mobility was associated with greater negative symptom severity—possibly due to diminished motivation [87].

Concerning diagnosis, methods outside of GPS-based locations have also been explored to determine the level of social functioning among people with schizophrenia. In this case, routines and disturbances within those routines are seen as indicators of a relapse and may be used in the future as a marker for potential intervention [89]. Researchers have looked at smartphone sensor data (eg, call or text logs, phone screen status, accelerometer, and GPS information) as a proxy for social functioning to determine the relationship between measuring social rhythm (using phenotyping) and schizophrenia. They found that increased stability in social rhythm and routine was associated with improved symptom scores in schizophrenia. This also implies that using phenotyping data related to sociability, mobility, and physical activity, as we have described above, in detecting depression, anxiety, and bipolar disorder can all be associated with identifying negative symptoms of schizophrenia [88].

In another study involving adults with schizophrenia, the researchers looked at mobility before a relapse and found that visual hallucinations were associated with a change in sleeping patterns and the number of voice and nonvoice sounds captured, while auditory hallucinations decreased with new places visited and an increase in call duration [90]. Further, compared to individuals who did not require clinical intervention, there was a significantly higher rate of anomalies in the 2 weeks directly preceding the relapse. In a similar study, the researchers looked at movement information using purely accelerometry data from the smartphone to identify negative symptoms of schizophrenia. The findings of the study indicate that smartphone accelerometers, which are markers of movement and vigor (measured via accelerometry), hold promise in detecting negative symptoms of schizophrenia [91]. In an additional study, the researchers explored the use of accelerometer data to estimate sleep duration and sleep quality in individuals with schizophrenia. The results showed that these attributes were reliable to indicate the presence of sleep disturbances, which can be early warning signs for psychosis and relapse [92].

Disorders From the Use of Psychoactive Substances

A substance use disorder (SUD) is a mental disorder in which a person experiences an inability to control their use of psychoactive substances such as alcohol, cannabis, opioids, etc [97]. Understanding these SUDs is an emerging aspect of smartphone-based digital phenotyping. We group the effort in this space into 2 broad categories: alcohol use, and cannabis and opioid use. As with the previous efforts at phenotyping, these approaches use sensor data available from smartphones and associate behaviors such as social functioning, movement patterns, and physical activity to psychoactive substance use. Table 6 provides a summary of the research reviewed in this section. Phenotyping work in SUDs was in both monitoring and diagnosis focus categories for both alcohol use and cannabis and opioid use.

XSL•FO
RenderX

Table . Summary of smartphone-based digital phenotyping studies focused on substance use disorders (n=7). Studies examined behavioral and physiological markers related to alcohol, cannabis, and opioid use. Phenotypes primarily involved movement, sociability, and respiratory features derived from accelerometer, GPS, microphone, and app log data. Ground-truth measures included self-reported use, craving reports, and laboratory validation of BAC[a].

Citation	Condition	Phenotypes considered	Sensors used	Ground truth	Participants (n)	Duration	Focus	Study setting
[98]	SUD[b]–alcohol use	Movement, sociability	Accelerometer, battery, call or text logs, light, WiFi, Bluetooth, GPS, gyroscope, screen state	Self-reported alcohol use	30 adults	28 days	Monitoring	Naturalistic setting (location unclear)
[99]	SUD–alcohol use	Movement, sociability	Accelerometer, GPS, screen state, app logs, call or text logs, keystroke data	Mood and craving self-report	24 adults	30 days	Monitoring	Naturalistic setting (United States)
[100]	SUD–alcohol use	Physical activity	Accelerometer	Alcohol use self-report	7 adults	2 weeks	Diagnosis	Naturalistic setting (United States)
[101]	SUD–alcohol use	Physical activity	Accelerometer, gyroscope	Urine-based BAC	65 adults	1 in-laboratory session (5 h)	Diagnosis	Naturalistic setting (United States)
[102]	SUD–cannabis use	Activity, movement, sociability	GPS, accelerometer, call logs	Acute cannabis intoxication self-report	57 adults	30 days	Monitoring	Naturalistic setting (United States)
[103]	SUD–opioid use	Respiratory changes	Microphone	Simulated overdose events	209 injection events	3 - 5 min postevent	Diagnosis	Hybrid setting (United States)
[104]	SUD–opioid use	Movement patterns	GPS	Drug craving self-report	189 adults with OUD[c]	16 weeks	Diagnosis	Naturalistic setting (United States)

[a]BAC: blood alcohol concentration.

[b]SUD: substance use disorder.

[c]OUD: opioid use disorder.

Alcohol Use

One of the main areas of interest in smartphone-based phenotyping around the use of psychoactive substances is in studying alcohol use. These efforts take 2 basic forms: determining the level of intoxication and predicting cravings. Phenotyping work related to alcohol use falls under the monitoring and diagnosis categories.

Concerning monitoring, using phenotyping to determine how intoxicated an individual is has been the most common form of research in this space. One notable study in this domain looked at a wide range of mobile phone sensor data (eg, accelerometry, keyboard usage, screen state, call or text logs, battery status, etc) from young adults with a history of hazardous drinking over a period of a month. It found that time of day, day of week, screen duration, and frequency of keyboard presses had significant associations with heavy drinking, indicating behavioral changes during these episodes [98].

In a slightly different study, the researchers looked at alcohol craving in individuals with alcohol-associated liver disease via phone call status, movement patterns, phone screen state, and keystroke information. They found that the level of craving was directly proportional to loneliness, stress, and anxiety as measured primarily through a decrease in the entropy of their movement patterns [99].

Concerning diagnosis, a couple of other efforts looked at blood alcohol levels based on gait information inferred from smartphone sensors [100,101]. The researchers conducted these studies in an uncontrolled environment where it was not clear where the phone was located on the individual's body. These works used accelerometers and gyroscopes, respectively. Overall, they found some positive correlation between gait

XSL•FO
RenderX

measured through smartphone sensors and the level of intoxication of the individual.

Cannabis and Opioid Use

Not much work has been done when it comes to characterizing cannabis and opioid use using smartphone-based phenotyping. The extant work in this space can be grouped into the same 2 categories as the alcohol case: determining the level of intoxication and predicting cravings. Similar to alcohol use, phenotyping work related to cannabis and opioid use fell under the monitoring and diagnosis categories.

Concerning monitoring, for cannabis, researchers have looked at identifying slowed psychomotor function as a result of consuming cannabis. They measured psychomotor function in an individual based on their movement pattern and physical activity measured using GPS, accelerometer, and call logs. They found a smaller travel boundary and lower activity changes at times when the individual felt "high." Additionally, stronger body movements indicated by changes in the accelerometer may be indicative of impaired motor control or unsteady movements typically associated with higher levels of intoxication [102].

Concerning diagnosis, in a study related to the context of opioid use, the researchers explored the feasibility of using smartphone data to monitor respiratory changes indicative of an impending opioid overdose. They showed that they could convert a phone microphone into a short-range active sonar using frequency shift to measure chest motion and respiration. This work was done in a controlled operating room environment of the operating room, where overdose conditions were simulated; this method showed promising accuracy in identifying respiratory depression [103].

Finally, the increased interest in using smartphones to aid in the understanding and treatment of SUDs has created an opportunity to assess substance craving as reported in 1 study. By looking at the past 5 hours of solely GPS-based movement data from individuals with opioid-use disorder and correlating it with environmental exposure (measured as a function of the affluence of the locations they visited), the researchers could predict craving behaviors up to 90 minutes in the future for both heroin and cocaine. They were even able to predict stress in the individuals (indicated through EMA responses) using this information [104].

Sleep Disorders

Smartphone-based digital phenotyping research has also specifically looked at sleep disorders. However, it is somewhat limited. The main focus in phenotyping for sleep disorders has been on sleep apnea. Sleep apnea is a common condition in which a person's breathing stops and starts again throughout the night [105]. This is a potentially serious disorder that not only impacts sleep quality, but it may also lead to increased cardiovascular risk and lead to premature death if not treated correctly [105]. Smartphone-based approaches primarily monitor breathing and snoring patterns using the smartphone microphone. The ground truth for these studies was established by recruiting individuals with previously diagnosed sleep apnea or by conducting polysomnography as part of the study. Table 7 provides a summary of sleep disorder research reviewed in this section.

XSL•FO
RenderX

Table . Summary of smartphone-based digital phenotyping studies focused on sleep disorders, neurological disorders, ADHD[a], and general well-being (n=10). Studies in these domains used smartphone sensors to assess sleep quality, neurological motor symptoms, attention patterns, and general well-being. Common sensor modalities included microphones, accelerometers, and gyroscopes, with validation against polysomnogram, clinical scales, or self-reported measures shown to be strongly associated with polysomnography results for central apnea, hypopnea, and obstructive apnea [103].

Citation	Condition	Phenotypes considered	Sensors used	Ground truth	Participants (n)	Duration	Focus	Study setting
[106]	Sleep apnea	Sleep disruption, sleep patterns	Microphone	Polysomnogram taken before the study	55 samples	One 8-hour sleep period	Diagnosis	Hybrid setting (location not stated)
[107]	Sleep apnea	Sleep disruption, sleep patterns	Microphone	Polysomnogram score	55 sleep samples	One 8-hour sleep period	Diagnosis	Controlled laboratory setting (location not stated)
[108]	Parkinson disease	Tremors, FOG[b]	Accelerometer	TUG[c] test	20 adults with PD[d]	One in-laboratory assessment	Diagnosis	Controlled laboratory setting (location not stated)
[109]	Parkinson disease	FOG	Accelerometer, gyroscope	Heel-mounted foot switch, pressure sensor mat	12 adults with PD, 12 healthy controls	One in-laboratory assessment	Diagnosis	Controlled laboratory setting (Singapore)
[110]	Parkinson disease	Tremors	Accelerometer, gyroscope	UPDRS[e]	23 adults with PD	One in-laboratory assessment	Diagnosis	Controlled laboratory setting (Greece)
[111]	Parkinson disease	Tremors	Accelerometer	Diagnosis of PD or ET[f], reevaluation after 1 year	17 adults with PD, 16 adults with ET, 12 healthy controls	One in-laboratory assessment	Diagnosis	Controlled laboratory setting (location not stated)
[112]	ADHD	Loss of concentration	Text logs	Self-reported symptoms	2 adults with ADHD, 2 healthy controls	10 weeks	Monitoring	Naturalistic setting (United States)
[113]	General well-being	Movement patterns, physical activity level, sociability	Microphone, accelerometer, GPS, text logs, call logs	Self-reported mood	18,000 adults	3 years	Monitoring	Naturalistic setting (app store recruitment)
[114]	General well-being	Physical activity level, sleep disruption, sociability	Accelerometer, microphone, screen state, battery level	Self-reported physical and mental well-being	27 adults	19 days	Intervention	Naturalistic setting (app store recruitment)

[a]ADHD: attention-deficit/hyperactivity disorder.

[b]FOG: freezing of gait.

[c]TUG: Timed Up and Go.

[d]PD: Parkinson disease.

[e]UPDRS: Unified Parkinson Disease Rating Scale.

[f]ET: essential tremor.

XSL•FO
RenderX

Sleep Apnea

Traditionally, measuring the presence of sleep apnea requires an overnight stay at a sleep laboratory in order for a formal diagnosis to take place [115]. This stringent requirement may discourage individuals from actively pursuing treatment if they are unsure of the outcome of a sleep study or may perceive it to be too much of a hassle. As a result, smartphone-based digital phenotyping aims to remain as unobtrusive as possible. Phenotyping work for sleep apnea was primarily in the diagnosis focus category.

Concerning diagnosis, in 1 study, the researchers proposed a method that used accelerometer data and compared movements to a threshold to determine the extent of sleep movement. This was combined with snoring-related sound data from the smartphone microphone to determine the quality of sleep, which was divided into 3 levels of apnea: no apnea, mild apnea, and severe apnea [106]. In a similar study, researchers monitored the breathing patterns of a person using a smartphone app placed at the bedside. Their approach includes collecting information on breathing, the sound of snoring, and breathing pauses to determine the presence of sleep apnea and hypopnea, by parsing out white noise and various sounds that occur during the night [107].

The team that developed the means to determine the feasibility of using a smartphone microphone to monitor respiratory changes indicative of an impending opioid overdose also developed a similar approach that turned the microphone into a short-range active sonar to detect chest and abdomen movements. The changes in chest and abdomen movement detected from the reflected sound waves then allowed for apnea detection, which was then shown to be strongly associated with polysomnography results for central apnea, hypopnea, and obstructive apnea [103].

Neurological Disorders

A neurological disorder is any disorder of the nervous system. Examples of symptoms include paralysis, muscle weakness, poor coordination, loss of sensation, seizures, confusion, pain, and altered levels of consciousness [116]. From a smartphone-based digital phenotyping standpoint, work on neurological disorders has primarily focused on characterizing Parkinson disease. Table 7 provides a summary of the research reviewed in this section.

Parkinson disease

Parkinson disease is a movement disorder of the nervous system that progressively worsens over time. As neurons in the brain weaken and die, individuals with the disorder begin to endure tremors, impaired balance, stiffness in the limbs and trunk of the body, and difficulty with movement [117]. Like sleep apnea, phenotyping work for Parkinson disease was also primarily in the diagnosis focus category.

Concerning diagnosis, an aspect of Parkinson disease that has been explored in the smartphone-based phenotyping context is the detection of real-time freezing of gait (stopping suddenly while walking) [108]. In a similar study, 20 individuals with Parkinson disease were asked to perform a video-recorded Timed Up and Go test with and without dual-tasks while carrying a smartphone. The video and accelerometer recordings from the smartphone were synchronized in order to assess the reliability of the freezing of gait detection system in comparison to clinician ratings, with reasonable results [117]. In a similar study, the researchers looked at the accuracy of smartphone-based gait analysis using the built-in triaxial gyroscope and accelerometer and validated it against 2 heel contact-based measurements. These findings highlighted that smartphone-based gait analysis could serve as an alternative to conventional gait analysis methods (eg, foot-switch systems or sensor-embedded walkways), particularly when those methods are inconvenient (eg, cost-prohibitive) [109].

Smartphones have been evaluated as a way to measure hand tremor in Parkinson disease patients [110]. Tremors are a common symptom of Parkinson disease. In this study, accelerometers and gyroscopes were used to detect and measure hand tremors, which were then associated with clinical scores in the hand tremor components of the Unified Parkinson Disease Rating Scale. The results suggest a relatively strong correlation between the patients' Unified Parkinson Disease Rating Scale hand tremor scores and the quantitative measurement derived from smartphones [110]. In another similar study, the goal was to develop a diagnostic test to differentiate between Parkinson disease and essential tremor based on time-frequency differences using smartphone-based accelerometer data without the use of neuroimaging techniques [111]. In this case, the smartphone was placed over the back of the outstretched hand of a person with Parkinson disease while at rest. The results demonstrated that not only was the phone-based approach effective, but it was also able to diagnose the presence of Parkinson disease and essential tremor in several participants, each of whom had not been officially diagnosed [111].

Other Efforts

A small amount of work has also been done in characterizing ADHD and assessing the general mental health and well-being of individuals [112-114]. Table 7 provides a summary of the research covered in this section. Phenotyping work for ADHD was primarily focused on monitoring, whereas general well-being had work in both monitoring and diagnosis focus categories.

Concerning monitoring, with respect to ADHD, researchers collected text data and associated it with weekly self-report ADHD symptom scores. The ground truth for the ADHD symptoms was self-reported by the participants. Overall, the results of the study indicated that specific ADHD inattention, sluggish cognitive tempo, and hyperactivity symptoms could be predicted fairly accurately using the smartphone data collected from college students in a pilot study [112].

Further, smartphone-based phenotyping also focused on general well-being. These studies aimed to associate self-reported mental well-being with routine captured through a smartphone. This was a particularly large and longitudinal study. Overall, the study found that positive or negative affect for an individual could be predicted using passive sensing data (such as GPS, phone or call logs, accelerometer, and microphone data). These

XSL•FO
RenderX

results suggest that such approaches could reduce the need for self-report [113].

Concerning intervention, in the 1 intervention study, researchers deployed an app called BeWell [114]. BeWell monitored user behavior, such as sleep, physical activity, and social interaction, based on accelerometer, screen on or off, charging, and microphone data to give individuals a well-being score. The well-being scores range between 0 and 100 and are calculated for each of the 3 dimensions (sleep patterns, physical activity, and social interaction). A score of 100 indicates the person is matching or exceeding recommended guidelines. The idea is to use these scores to eventually provide feedback to individuals to maintain their well-being over time [114].

Sensor Modalities Used Across Different Health Domains

Finally, the reviewed studies showed wide variation in sensor use across health domains. While some investigations relied on a single data stream (eg, the work [69] that exclusively used GPS to track bipolar disorder or the research [91] that assessed the activity level of adults with schizophrenia using only the smartphone accelerometer), others combined multiple sources such as GPS, accelerometry, and microphone data to capture a richer behavioral context. As shown in Figure 2, GPS and accelerometer emerged as the 2 most common data streams. Accelerometers appeared as input 58% (38/65) of studies, making them the most frequently used modality overall. GPS followed close behind, appearing in 53% (35/65) studies. Communication logs were also common, with call or text logs appearing in roughly 44% (28/65) of studies. These streams formed the core of most smartphone-based digital phenotyping work due to their ability to intuitively capture mobility, activity, and daily routines. Other smartphone sensing modalities were used far less often. Microphone-based sensing appeared in 27% (18/65) of studies and was fairly evenly spread between health conditions, indicating the potential for widespread use across conditions. Screen state data also appeared in 27% (18/65) of studies, typically as a proxy for phone engagement and sleep-wake cycles. Gyroscope data were used relatively infrequently, appearing in just 10% (7/65) of studies. Environmental and proximity sensing (ambient light, WiFi, cell tower pings, app usage, and Bluetooth) appeared sparingly. At most, these streams appeared in 15% (10/65) of studies in the case of ambient light and just 6% (4/65) of studies in the case of cell tower pings. Battery percentage was used surprisingly infrequently, with only 7% (5/65) of studies using this stream. We speculate this may be due to internal differences in how Android (Google LLC) and iOS (Apple Inc) devices may collect data, but this is not made clear by the authors, who elect to use (or not use) the battery state as a data stream. Keystroke dynamics appeared the least out of all streams being used in just 3% (2/65) of studies. This is likely due to keystroke dynamics being used for specific, tailored tasks as opposed to sensing broad behavioral changes.

Figure 2. Heatmap illustrating the frequency with which different smartphone sensor modalities were used across 6 major health domains in the reviewed literature. Brighter and warmer-colored cells indicate more frequent usage of a given sensor type within that domain.

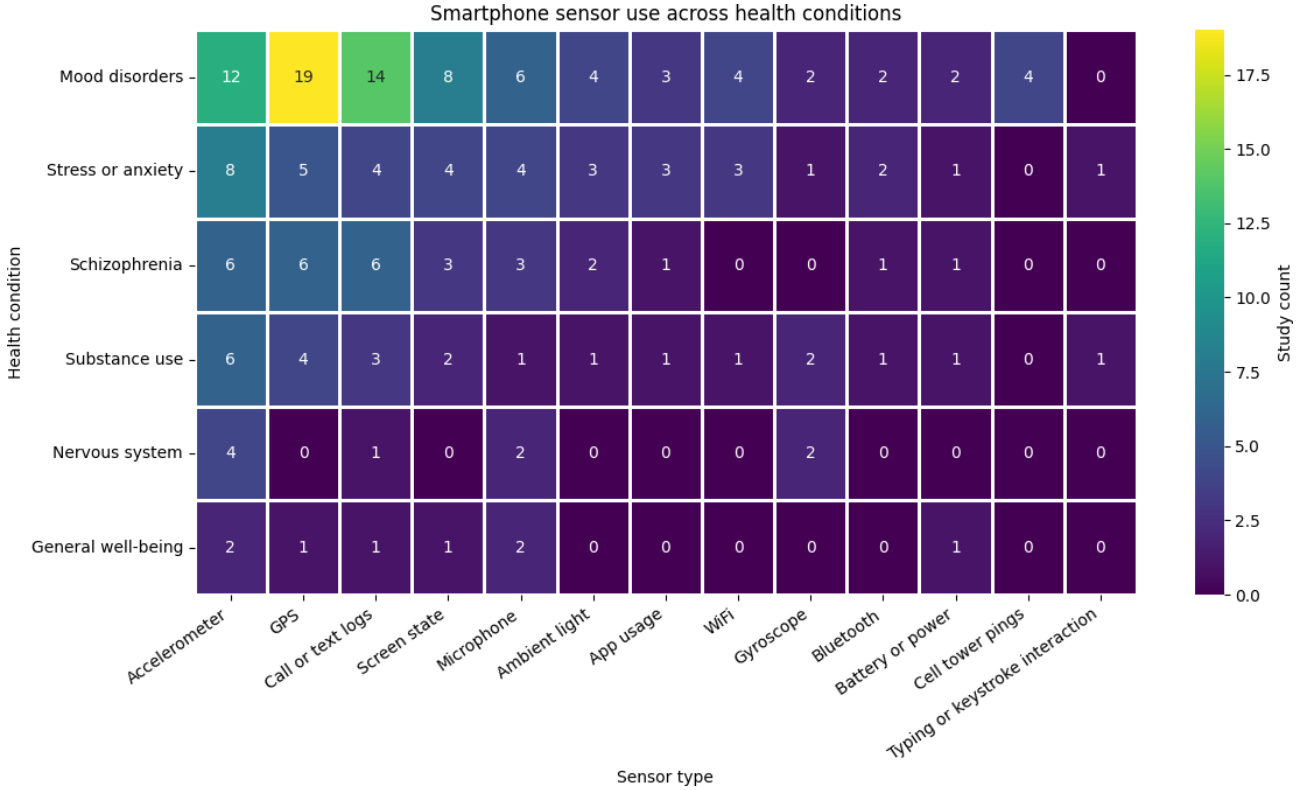

XSL•FO
RenderX

Discussion

Principal Findings

This scoping review synthesized research that used smartphone-based digital phenotyping to understand health-related conditions. Most studies in our review (n=45) used smartphone-based digital phenotyping to help with mental health conditions. Within the wider health scope, depression and bipolar disorder are of particular interest to researchers due to the ability to effectively track behavioral and mood fluctuations over time through the use of passive smartphone sensors. In contrast, relatively few studies explored other domains such as substance use (n=7), neurological disorders (n=4), or sleep disorders (n=7).

When classified by focus, about 63.5% (41/65) of the studies were conducted with the intent to monitor, which we have defined as tracking symptoms, behaviors, and environmental contexts over time. A smaller subset of these studies (n=15, 23.1%) looked at phenotypes for the purpose of diagnosis, which involves using smartphone-derived features to identify the presence or severity of health conditions. Finally, a few studies totaling 3% (2/65) implemented or evaluated intervention-based approaches in which smartphone-derived behaviors were used to inform treatment strategies or real-time interventions.

Collectively, these findings demonstrate that while smartphone-based phenotyping has some potential across monitoring, diagnostic, and intervention-focused studies for a variety of health conditions, there are several aspects of this research domain that could be improved, which we discuss below.

Methodological Inconsistencies and Gaps

We found that 15.3% (10/65) of included research studies lacked a clear description of which sensor streams were being used in their phenotyping work and why. This lack of detail about the streams takes several forms: (1) instances where the sensor streams are mentioned as being in scope but are not mentioned in the collection process described [38,39,52,68,112]; (2) instances where a variety of sensor streams are measured from participants but only a subset are used in the phenotyping work [80,92]; and (3) instances where the type of sensor stream collected to capture a phenotype is not even specified [70]. These limitations hinder continued iteration on the described work and suggest the need for guidelines for how to report data surrounding smartphone-based digital phenotyping work to enhance reproducibility, learn from the successes and failures of groups conducting this work, and ultimately advance these tools toward broader impacts. Each of these instances prevents us from fully understanding the scope of the work being done and hinders future research from properly expanding upon the findings of these works. We recommend that, at a minimum, research groups should describe the types of smartphones used in the study, their operating system, the specific sensor streams, and the strategies in which the data was obtained. This methodology could benefit others who seek to leverage smartphone-based digital phenotyping to detect and understand disease trajectories.

These reporting issues extend beyond sensor specifications. Despite all studies using smartphones as the sole sensing device, few provided sufficient technical or procedural detail to allow for replication. Only 9.2% (6/65) of the studies quantitatively reported missingness at the stream level [38,39,52,59,68,92]. Similarly, only 6.1% (4/65) of the studies explicitly reported sampling rates or duty cycles [52,82,83,85], with all other studies leaving the frequency of data capture and transmission unclear. The most common detail reported was phone operating system with 33.8% (22/65) of the studies stating the phone operating system used in data collection, indicating whether the study used Android, iOS, or mixed devices [38,39,48,51-53,59,64,68,70,71,78,80,82,83,85-87,89,90,112,113]. Finally, only 10.7% (7/65) of the studies discussed participant compliance or attrition, offering information about adherence, dropout, or data completeness at the participant level [38,39,52,68,71,82,85].

ML models are commonly used in smartphone-based digital phenotyping work to determine the presence of specific health conditions when monitoring or diagnosing them. Of the 65 studies we assessed as part of this work, about 90.7% (n=59) used ML models, while about 9.3% (n=6) did not [54,72,84,106,109,111]. For our discussion on the use of ML models, we will be noting statistics as being out of 59 so as not to unfairly include studies that do not use ML. About 69.4% (41/59) of the studies that used ML models in their smartphone-based phenotyping work delved deeper into how these models were learning and what derived features may mean in the context of a given health outcome. However, we found that a substantial number, about 30.5% (18/59) of the studies, did not engage with ML models beyond the perfunctory [38,45,53,56,66,69,77,80,81,86,100,101,103,107,108,110,112,118]. These studies reported how accurate these models are with respect to detecting the presence of a health condition in a dataset. While this kind of analysis is important to understanding if models are working, they often eschew any other kind of analysis on why the model is making its decision. We found that these emphases on how accurately an ML model performed did not provide any insight into the validity of the models themselves or an understanding of a person's behavior being used as phenotypes for characterizing a specific health condition. We believe these points point to a tendency to discover without understanding in a substantial portion of the studies that used ML approaches, which is a cause for numerous errors and bad science [119].

We believe that when ML models are discussed, an effort should be made in exploring the various features used to determine the presence of specific health conditions in addition to reporting on the global accuracy of detection. Reporting on the features allows us, at the very least, to sanity check if the model is making decisions based on known behavioral patterns that individuals with specific health conditions exhibit. However, more importantly, such an analysis has the potential to provide insights into newer ways of identifying these health conditions. This would advance the science of detecting specific health conditions and thus truly leverage the pattern-matching capabilities of ML models.

XSL•FO
RenderX

Finally, we found that despite the amount of work in smartphone-based digital phenotyping, there is a lack of publicly available datasets in this space. The primary issue with not having any data available is that researchers are required to spend a significant amount of time recruiting participants, collecting data, and cleaning data before they perform any kind of analysis. If data were to be made available, innovation and further research would be significantly easier. Only 10.7% (7/65) of studies discuss data availability. Of the studies that provide access to their data, 2 use the Dartmouth StudentLife dataset, which is openly available [82,83]. About 4.6% (3/65) of the studies indicate that the data are available upon request to the authors [49,72,87]. One study indicates that the information is individually identifiable and it must be requested through the institutional ethics board [88]. Only 1.5% (1/65) of the studies provide a working link to the GPS data collected during the study [50]. This suggests that researchers should consider strategies to provide data for other teams to advance smartphone-based digital phenotyping as an innovative strategy to address multiple health conditions. In this process, researchers should consult with their respective institutional ethics boards and ethicists to understand the implications of data sharing, especially for groups whose phenotyping data can be perceived as being more sensitive (eg, specific geolocation and text message data). In some cases, it may be necessary to restrict data access for certain streams of phenotyping data to protect the confidentiality of participants. At the same time, considerations surrounding consent during enrollment for future data sharing to advance the field should be encouraged among research teams. We encourage authors to disclose how to access the data. In the case that data is not readily accessible and must be requested through specific channels, we feel that authors should state this. In case the data is not available under any circumstances, we feel that authors should disclose this as well.

In this scoping review, we provide a comprehensive synthesis of smartphone-based digital phenotyping research across a broader range of health conditions than has been examined in prior reviews. By focusing exclusively on data collected from smartphone sensors, this work differs from existing reviews that either incorporate auxiliary sensing modalities or limit their scope to a small number of predominantly mental health conditions. From a real-world perspective, these results underscore the potential of smartphones as scalable, unobtrusive tools for monitoring, diagnosis, and intervention across mental health, physical health, and SUDs while also highlighting the need for improved reporting practices and validation to support responsible deployment.

Limitations

This scoping review has a few limitations. First, we excluded non-English and non–peer-reviewed literature, which may have resulted in missing relevant work conducted in other regions or early-stage research not yet published in academic journals. Second, we only searched Google Scholar, IEEE Xplore, ACM Digital Library, and PubMed databases for identifying the studies for this scoping review. Again, searching other databases could have revealed more studies in this space, which we did not include in this review. Third, as a scoping review, we aimed to map available literature to our 3 given focus categories rather than assessing the quality or strength of each study included. As an additional aspect to this point, studies included in this scoping review vary widely in multiple aspects, including the length of the study, the number of participants, and the study setting, which makes direct comparison between studies challenging. Finally, most studies in this review were observational and conducted in relatively small samples, often within specific demographic groups (eg, students). Due to this, the behavioral associations identified should be interpreted as correlational rather than causal, and their applicability to broader or more diverse populations remains an open question.

Despite these limitations, this scoping review provides a comprehensive overview of smartphone-based digital phenotyping across health domains and highlights important gaps that future work must address, particularly around reporting standards, validation practices, and data sharing.

Conclusion

In this scoping review, we examined the work done in the domain of smartphone-based digital phenotyping in determining the presence of various health conditions. Our scope was much more expansive than prior scoping reviews in this space and includes any and all health conditions that have been explored using smartphone-based phenotyping work. Further, and most importantly, in this work, we focus on describing the relationships between behaviors exhibited in the context of specific health conditions and not on algorithms used to extract these behaviors. Overall, based on a scoping review of 65 studies, we found that smartphone-based digital phenotyping is a powerful and informative tool that can give us insights into a variety of health conditions. Though most of the work has focused on applying smartphone-based digital phenotyping to mental health issues, such as depression, bipolar disorder, and stress, some effort has also gone into phenotyping sleep and neurological conditions.

Smartphone-based digital phenotyping offers substantial promise for understanding health-related behaviors in everyday life. Despite this promise, continued progress depends on addressing key issues surrounding data quality, reproducibility, and reporting standards. Further, as smartphone-based phenotyping scales to larger and more heterogeneous populations, privacy and consent challenges will become increasingly important. Ensuring ethical collection, sharing, and use of sensitive behavioral data will be essential to maintaining participant trust and supporting the responsible growth of this research domain. Smartphones are no longer simple communication tools, but they are emerging as primary instruments for continuous health monitoring.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank the reviewers for providing excellent comments that were very helpful in improving the quality of this scoping review. Responsibility for the final manuscript lies entirely with the authors. Generative artificial intelligence tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by AD. No artificial intelligence tool was used for this paper.

Funding

This work was supported by the National Institutes of Health grant DP2DA056107. The funders had no role in the decision to publish or in the preparation of this paper.

Authors' Contributions

Conceptualization: AD, KV, PC
Data curation: AD, JH
Formal analysis: AD, JH
Funding acquisition: PC
Investigation: AD, KV
Methodology: AD, KV, PC
Project administration: GG, PC
Resources: PC
Supervision: KV, PC
Writing – original draft: AD, KV, PC
Writing – review and editing: AD, KV

Conflicts of Interest

None declared.

Multimedia Appendix 1
Presents the full electronic search strategies used to identify relevant records.
[DOCX File, 1672 KB - jmir_v28i1e84146_app1.docx]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 74 KB - jmir_v28i1e84146_app2.docx]

References

1. Mobile technology and home broadband 2021. Pew Research Center. 2021 Jun 3. URL: https://www.pewresearch.org/internet/2021/06/03/mobile-technology-and-home-broadband-2021 [accessed 2026-02-27]
2. Ventola CL. Mobile devices and apps for health care professionals: uses and benefits. P T 2014 May;39(5):356-364. [Medline: 24883008]
3. Strawley C, Richwine C. Individuals' access and use of patient portals and smartphone health apps, 2022. In: ASTP Health IT Data Brief: Office of the Assistant Secretary for Technology Policy; 2023, Vol. 2012. 69. [Medline: 39413226]
4. Zulman DM, Verghese A. Virtual care, telemedicine visits, and real connection in the era of COVID-19: unforeseen opportunity in the face of adversity. JAMA 2021 Feb 2;325(5):437-438. [doi: 10.1001/jama.2020.27304] [Medline: 33528520]
5. Trends in individuals' access, viewing, and use of online medical records and other technology for health needs: 2022. Assistant Secretary for Technology Policy. 2019 May. URL: https://www.healthit.gov/data/data-briefs/trends-individuals-access-viewing-and-use-online-medical-records-and-other [accessed 2026-02-27]
6. Wall C, Hetherington V, Godfrey A. Beyond the clinic: the rise of wearables and smartphones in decentralising healthcare. NPJ Digit Med 2023 Nov 25;6(1):219. [doi: 10.1038/s41746-023-00971-z] [Medline: 38007554]
7. Dorsey ER, Chan YFY, McConnell MV, et al. The use of smartphones for health research. Acad Med 2017 Feb;92(2):157-160. [doi: 10.1097/ACM.0000000000001205] [Medline: 27119325]
8. Ernsting C, Dombrowski SU, Oedekoven M, et al. Using smartphones and health apps to change and manage health behaviors: a population-based survey. J Med Internet Res 2017 Apr 5;19(4):e101. [doi: 10.2196/jmir.6838] [Medline: 28381394]
9. Onnela JP. Opportunities and challenges in the collection and analysis of digital phenotyping data. Neuropsychopharmacology 2021 Jan;46(1):45-54. [doi: 10.1038/s41386-020-0771-3] [Medline: 32679583]

XSL•FO
RenderX

10. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu Rev Clin Psychol 2017 May 8;13:23-47. [doi: 10.1146/annurev-clinpsy-032816-044949] [Medline: 28375728]
11. Harari GM, Lane ND, Wang R, Crosier BS, Campbell AT, Gosling SD. Using smartphones to collect behavioral data in psychological science: opportunities, practical considerations, and challenges. Perspect Psychol Sci 2016 Nov;11(6):838-854. [doi: 10.1177/1745691616650285] [Medline: 27899727]
12. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA 2017 Oct 3;318(13):1215-1216. [doi: 10.1001/jama.2017.11295] [Medline: 28973224]
13. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology 2016 Jun;41(7):1691-1696. [doi: 10.1038/npp.2016.7] [Medline: 26818126]
14. Lee K, Lee TC, Yefimova M, et al. Using digital phenotyping to understand health-related outcomes: a scoping review. Int J Med Inform 2023 Jun;174:105061. [doi: 10.1016/j.ijmedinf.2023.105061] [Medline: 37030145]
15. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
16. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
17. Cornet VP, Holden RJ. Systematic review of smartphone-based passive sensing for health and wellbeing. J Biomed Inform 2018 Jan;77:120-132. [doi: 10.1016/j.jbi.2017.12.008] [Medline: 29248628]
18. Fraccaro P, Beukenhorst A, Sperrin M, et al. Digital biomarkers from geolocation data in bipolar disorder and schizophrenia: a systematic review. J Am Med Inform Assoc 2019 Nov 1;26(11):1412-1420. [doi: 10.1093/jamia/ocz043] [Medline: 31260049]
19. Jayakumar P, Lin E, Galea V, et al. Digital phenotyping and patient-generated health data for outcome measurement in surgical care: a scoping review. JPM 2020;10(4):282. [doi: 10.3390/jpm10040282]
20. Saccaro LF, Amatori G, Cappelli A, Mazziotti R, Dell'Osso L, Rutigliano G. Portable technologies for digital phenotyping of bipolar disorder: a systematic review. J Affect Disord 2021 Dec 1;295:323-338. [doi: 10.1016/j.jad.2021.08.052] [Medline: 34488086]
21. Hilty DM, Armstrong CM, Luxton DD, Gentry MT, Krupinski EA. A scoping review of sensors, wearables, and remote monitoring for behavioral health: uses, outcomes, clinical competencies, and research directions. J Technol Behav Sci 2021;6(2):278-313. [doi: 10.1007/s41347-021-00199-2]
22. De Angel V, Lewis S, White K, et al. Digital health tools for the passive monitoring of depression: a systematic review of methods. NPJ Digit Med 2022;5(1):3. [doi: 10.1038/s41746-021-00548-8]
23. Dlima SD, Shevade S, Menezes SR, Ganju A. Digital phenotyping in health using machine learning approaches: scoping review. JMIR Bioinform Biotech 2022;3(1):e39618. [doi: 10.2196/39618]
24. Mendes JPM, Moura IR, Van de Ven P, et al. Sensing apps and public data sets for digital phenotyping of mental health: systematic review. J Med Internet Res 2022 Feb 17;24(2):e28735. [doi: 10.2196/28735] [Medline: 35175202]
25. Alfalahi H, Dias SB, Khandoker AH, Chaudhuri KR, Hadjileontiadis LJ. A scoping review of neurodegenerative manifestations in explainable digital phenotyping. NPJ Parkinsons Dis 2023 Mar 30;9(1):49. [doi: 10.1038/s41531-023-00494-0] [Medline: 36997573]
26. Dobson R, Li LL, Garner K, Tane T, McCool J, Whittaker R. The use of sensors to detect anxiety for in-the-moment intervention: scoping review. JMIR Ment Health 2023 Feb 2;10(1):e42611. [doi: 10.2196/42611] [Medline: 36729590]
27. Bufano P, Laurino M, Said S, Tognetti A, Menicucci D. Digital phenotyping for monitoring mental disorders: systematic review. J Med Internet Res 2023;25:e46778. [doi: 10.2196/46778]
28. Zhang H, Ibrahim A, Parsia B, Poliakoff E, Harper S. Passive social sensing with smartphones: a systematic review. Computing 2023 Jan;105(1):29-51. [doi: 10.1007/s00607-022-01112-2]
29. Moura I, Teles A, Viana D, Marques J, Coutinho L, Silva F. Digital phenotyping of mental health using multimodal sensing of multiple situations of interest: a systematic literature review. J Biomed Inform 2023 Feb;138:104278. [doi: 10.1016/j.jbi.2022.104278] [Medline: 36586498]
30. Beames J, Han J, Shvetcov A, et al. Use of smartphone sensor data in detecting and predicting depression and anxiety in young people (12-25 years): a scoping review. SSRN J 2024;11:e47206. [doi: 10.2139/ssrn.4798261]
31. Leaning IE, Ikani N, Savage HS, et al. From smartphone data to clinically relevant predictions: a systematic review of digital phenotyping methods in depression. Neurosci Biobehav Rev 2024 Mar;158:105541. [doi: 10.1016/j.neubiorev.2024.105541] [Medline: 38215802]
32. Choi A, Ooi A, Lottridge D. Digital phenotyping for stress, anxiety, and mild depression: systematic literature review. JMIR mHealth uHealth 2024;12(1):e40689. [doi: 10.2196/40689]
33. Linardon J, Chen K, Gajjar S, et al. Smartphone digital phenotyping in mental health disorders: a review of raw sensors utilized, machine learning processing pipelines, and derived behavioral features. Psychiatry Res 2025 Jun;348:116483. [doi: 10.1016/j.psychres.2025.116483] [Medline: 40187059]
34. Kallio J, Vildjiounaite E, Tervonen J, Bordallo López M. A survey on sensor-based techniques for continuous stress monitoring in knowledge work environments. ACM Trans Comput Healthcare 2025 Jul 31;6(3):1-31. [doi: 10.1145/3720551]

XSL•FO
RenderX

35. Booth BM, Mundnich K, Feng T, et al. Multimodal human and environmental sensing for longitudinal behavioral studies in naturalistic settings: framework for sensor selection, deployment, and management. J Med Internet Res 2019 Aug 20;21(8):e12832. [doi: 10.2196/12832] [Medline: 31432781]
36. Sekhon S, Gupta V. Mood Disorder: StatPearls Publishing; 2025. URL: https://www.ncbi.nlm.nih.gov/books/NBK558911 [accessed 2026-02-27]
37. Depressive disorder (depression). World Health Organization. 2025 Aug 29. URL: https://www.who.int/news-room/fact-sheets/detail/depression [accessed 2026-02-27]
38. Choudhary S, Thomas N, Ellenberger J, Srinivasan G, Cohen R. A machine learning approach for detecting digital behavioral patterns of depression using nonintrusive smartphone data (complementary path to Patient Health Questionnaire-9 assessment): prospective observational study. JMIR Form Res 2022 May 16;6(5):e37736. [doi: 10.2196/37736] [Medline: 35420993]
39. Berrouiguet S, Ramírez D, Barrigón ML, et al. Combining continuous smartphone native sensors data capture and unsupervised data mining techniques for behavioral changes detection: a case series of the evidence-based behavior (eB2) study. JMIR mHealth uHealth 2018 Dec 10;6(12):e197. [doi: 10.2196/mhealth.9472] [Medline: 30530465]
40. Jacobson NC, Chung YJ. Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. Sensors (Basel) 2020 Jun 24;20(12):3572. [doi: 10.3390/s20123572] [Medline: 32599801]
41. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
42. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol Meas 1977;1(3):385-401. [doi: 10.1177/014662167700100306]
43. Poznanski EO, Grossman JA, Buchsbaum Y, Banegas M, Freeman L, Gibbons R. Children's Depression Rating Scale–Revised. APA PsycTests 1984. [doi: 10.1037/t55280-000]
44. Busner J, Targum SD. The clinical global impressions scale: applying a research tool in clinical practice. Psychiatry (Edgmont) 2007 Jul;4(7):28-37. [Medline: 20526405]
45. Wang Y, Ren X, Liu X, Zhu T. Examining the correlation between depression and social behavior on smartphones through usage metadata: empirical study. JMIR mHealth uHealth 2021 Jan 6;9(1):e19046. [doi: 10.2196/19046] [Medline: 33404512]
46. Saeb S, Zhang M, Karr CJ, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res 2015 Jul 15;17(7):e175. [doi: 10.2196/jmir.4273] [Medline: 26180009]
47. Sarda A, Munuswamy S, Sarda S, Subramanian V. Using passive smartphone sensing for improved risk stratification of patients with depression and diabetes: cross-sectional observational study. JMIR mHealth uHealth 2019 Jan 29;7(1):e11041. [doi: 10.2196/11041] [Medline: 30694197]
48. Cao J, Truong AL, Banu S, Shah AA, Sabharwal A, Moukaddam N. Tracking and predicting depressive symptoms of adolescents using smartphone-based self-reports, parental evaluations, and passive phone sensor data: development and usability study. JMIR Ment Health 2020;7(1):e14045. [doi: 10.2196/14045]
49. Kim JS, Wang B, Kim M, et al. Prediction of diagnosis and treatment response in adolescents with depression by using a smartphone app and deep learning approaches: usability study. JMIR Form Res 2023;7(1):e45991. [doi: 10.2196/45991]
50. Saeb S, Lattie EG, Schueller SM, Kording KP, Mohr DC. The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ 2016;4:e2537. [doi: 10.7717/peerj.2537] [Medline: 28344895]
51. Canzian L, Musolesi M. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Association for Computing Machinery; 2015:1293-1304. [doi: 10.1145/2750858.2805845]
52. Wahle F, Kowatsch T, Fleisch E, Rufer M, Weidt S. Mobile sensing and support for people with depression: a pilot trial in the wild. JMIR mHealth uHealth 2016 Sep 21;4(3):e111. [doi: 10.2196/mhealth.5960] [Medline: 27655245]
53. Masud MT, Mamun MA, Thapa K, Lee DH, Griffiths MD, Yang SH. Unobtrusive monitoring of behavior and movement patterns to detect clinical depression severity level via smartphone. J Biomed Inform 2020 Mar;103:103371. [doi: 10.1016/j.jbi.2019.103371] [Medline: 31935462]
54. Melcher J, Lavoie J, Hays R, et al. Digital phenotyping of student mental health during COVID-19: an observational study of 100 college students. J Am Coll Health 2023 Apr;71(3):736-748. [doi: 10.1080/07448481.2021.1905650] [Medline: 33769927]
55. Tlachac M, Flores R, Reisch M, Kayastha R, Taurich N, Melican V, et al. StudentSADD: rapid mobile depression and suicidal ideation screening of college students during the coronavirus pandemic. Proc ACM Interact Mob Wearable Ubiquitous Technol 2022;6(2):1-32. [doi: 10.1145/3534604]
56. Higuchi M, Nakamura M, Shinohara S, et al. Effectiveness of a voice-based mental health evaluation system for mobile devices: prospective study. JMIR Form Res 2020 Jul 20;4(7):e16455. [doi: 10.2196/16455] [Medline: 32554367]
57. Tlachac M, Flores R, Toto E, Rundensteiner E. Early mental health uncovering with short scripted and unscripted voice recordings. In: Deep Learning Applications, Volume 4: Springer; 2022:79-110. [doi: 10.1007/978-981-19-6153-3_4]
58. Mercier HW, Hamner JW, Torous J, Onnela JP, Taylor JA. Digital phenotyping to quantify psychosocial well-being trajectories after spinal cord injury. Am J Phys Med Rehabil 2020;99(12):1138-1144. [doi: 10.1097/PHM.0000000000001506]

XSL•FO
RenderX

59. MacLeod L, Suruliraj B, Gall D, et al. A mobile sensing app to monitor youth mental health: observational pilot study. JMIR mHealth uHealth 2021;9(10):e20638. [doi: 10.2196/20638]
60. Bipolar disorder. National Institute of Mental Health. URL: https://www.nimh.nih.gov/health/topics/bipolar-disorder [accessed 2026-02-27]
61. Hamilton M. Development of a rating scale for primary depressive illness. Br J Soc Clin Psychol 1967 Dec;6(4):278-296. [doi: 10.1111/j.2044-8260.1967.tb00530.x] [Medline: 6080235]
62. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. Br J Psychiatry 1978 Nov;133(5):429-435. [doi: 10.1192/bjp.133.5.429] [Medline: 728692]
63. Beiwinkel T, Kindermann S, Maier A, et al. Using smartphones to monitor bipolar disorder symptoms: a pilot study. JMIR Ment Health 2016;3(1):e2. [doi: 10.2196/mental.4560]
64. Grunerbl A, Muaremi A, Osmani V, et al. Smartphone-based recognition of states and state changes in bipolar disorder patients. IEEE J Biomed Health Inform 2014;19(1):140-148. [doi: 10.1109/JBHI.2014.2343154]
65. Faurholt-Jepsen M, Vinberg M, Frost M, et al. Behavioral activities collected through smartphones and the association with illness activity in bipolar disorder. Int J Methods Psychiatr Res 2016 Dec;25(4):309-323. [doi: 10.1002/mpr.1502]
66. Ryan KA, Babu P, Easter R, et al. A smartphone app to monitor mood symptoms in bipolar disorder: development and usability study. JMIR Ment Health 2020 Sep 22;7(9):e19476. [doi: 10.2196/19476] [Medline: 32960185]
67. Grünerbl A, Oleksy P, Bahle G, Haring C, Weppner J, Lukowicz P. Towards smart phone based monitoring of bipolar disorder. In: Proceedings of the Second ACM Workshop on Mobile Systems, Applications, and Services for HealthCare: Association for Computing Machinery; 2012:1-6. [doi: 10.1145/2396276.2396280]
68. Palmius N, Tsanas A, Saunders KEA, et al. Detecting bipolar depression from geographic location data. IEEE Trans Biomed Eng 2016;64(8):1761-1771. [doi: 10.1109/TBME.2016.2611862]
69. Tseng YC, Lin ECL, Wu CH, Huang HL, Chen PS. Associations among smartphone app-based measurements of mood, sleep and activity in bipolar disorder. Psychiatry Res 2022 Apr;310:114425. [doi: 10.1016/j.psychres.2022.114425] [Medline: 35152069]
70. Abdullah S, Matthews M, Frank E, Doherty G, Gay G, Choudhury T. Automatic detection of social rhythms in bipolar disorder. J Am Med Inform Assoc 2016 May;23(3):538-543. [doi: 10.1093/jamia/ocv200] [Medline: 26977102]
71. Faurholt-Jepsen M, Busk J, Frost M, et al. Voice analysis as an objective state marker in bipolar disorder. Transl Psychiatry 2016 Jul 19;6(7):e856. [doi: 10.1038/tp.2016.123] [Medline: 27434490]
72. Fellendorf FT, Hamm C, Dalkner N, et al. Monitoring sleep changes via a smartphone app in bipolar disorder: practical issues and validation of a potential diagnostic tool. Front Psychiatry 2021;12:641241. [doi: 10.3389/fpsyt.2021.641241] [Medline: 33841209]
73. Alvarez-Lozano J, Osmani V, Mayora O, Frost M, Bardram J, Faurholt-Jepsen M, et al. Tell me your apps and I will tell you your mood: correlation of apps usage with bipolar disorder state. In: Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments: Association for Computing Machinery; 2014:1-7. [doi: 10.1145/2674396.2674408]
74. Cohen S, Janicki-Deverts D, Miller GE. Psychological stress and disease. JAMA 2007 Oct 10;298(14):1685-1687. [doi: 10.1001/jama.298.14.1685] [Medline: 17925521]
75. Jacobson NC, Summers B, Wilhelm S. Digital biomarkers of social anxiety severity: digital phenotyping using passive smartphone sensors. J Med Internet Res 2020;22(5):e16875. [doi: 10.2196/16875]
76. Gao Y, Li A, Zhu T, Liu X, Liu X. How smartphone usage correlates with social anxiety and loneliness. PeerJ 2016;4:e2197. [doi: 10.7717/peerj.2197]
77. Fukazawa Y, Ito T, Okimura T, Yamashita Y, Maeda T, Ota J. Predicting anxiety state using smartphone-based passive sensing. J Biomed Inform 2019 May;93:103151. [doi: 10.1016/j.jbi.2019.103151]
78. Maxhuni A, Hernandez-Leal P, Morales EF, Sucar LE, Osmani V, Mayora O. Unobtrusive stress assessment using smartphones. IEEE Trans on Mobile Comput 2020;20(6):2313-2325. [doi: 10.1109/TMC.2020.2974834]
79. Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. Psychiatr Rehabil J 2015;38(3):218-226. [doi: 10.1037/prj0000130]
80. Garcia-Ceja E, Osmani V, Mayora O. Automatic stress detection in working environments from smartphones' accelerometer data: a first step. IEEE J Biomed Health Inform 2016 Jul;20(4):1053-1060. [doi: 10.1109/JBHI.2015.2446195] [Medline: 26087509]
81. Bogomolov A, Lepri B, Ferron M, Pianesi F, Pentland A. Daily stress recognition from mobile phone data, weather conditions and individual traits. In: MM '14: Association for Computing Machinery; 2014:477-486. [doi: 10.1145/2647868.2654933]
82. Wang R, Chen F, Chen Z, Li T, Harari G, Tignor S, et al. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Association for Computing Machinery; 2014:3-14. [doi: 10.1145/2632048.2632054]
83. Wang R, Harari G, Hao P, Zhou X, Campbell AT. SmartGPA: how smartphones can assess and predict academic performance of college students. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Association for Computing Machinery; 2015:295-306. [doi: 10.1145/2750858.2804251]

XSL•FO
RenderX

84. Exposito M, Hernandez J, Picard RW. Affective keys: towards unobtrusive stress sensing of smartphone users. In: Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct: Association for Computing Machinery; 2018:139-145. [doi: 10.1145/3236112.3236132]
85. Buck B, Scherer E, Brian R, et al. Relationships between smartphone social behavior and relapse in schizophrenia: a preliminary report. Schizophr Res 2019 Jun;208:167-172. [doi: 10.1016/j.schres.2019.03.014] [Medline: 30940400]
86. Difrancesco S, Fraccaro P, van der Veer SN, et al. Out-of-home activity recognition from GPS data in schizophrenic patients. In: 2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS): IEEE; 2016:324-328. [doi: 10.1109/CBMS.2016.54]
87. Depp CA, Bashem J, Moore RC, et al. GPS mobility as a digital biomarker of negative symptoms in schizophrenia: a case control study. NPJ Digit Med 2019;2(1):108. [doi: 10.1038/s41746-019-0182-1] [Medline: 31728415]
88. Henson P, Barnett I, Keshavan M, Torous J. Towards clinically actionable digital phenotyping targets in schizophrenia. NPJ Schizophr 2020 May 5;6(1):13. [doi: 10.1038/s41537-020-0100-1] [Medline: 32372059]
89. Barnett I, Torous J, Staples P, Sandoval L, Keshavan M, Onnela JP. Relapse prediction in schizophrenia through digital phenotyping: a pilot study. Neuropsychopharmacology 2018 Jul;43(8):1660-1666. [doi: 10.1038/s41386-018-0030-z] [Medline: 29511333]
90. Wang R, Aung MS, Abdullah S, Brian R, Campbell AT, Choudhury T, et al. CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Association for Computing Machinery; 2016:886-897. [doi: 10.1145/2971648.2971740]
91. Strauss GP, Raugh IM, Zhang L, et al. Validation of accelerometry as a digital phenotyping measure of negative symptoms in schizophrenia. Schizophrenia 2022;8(1):37. [doi: 10.1038/s41537-022-00241-z]
92. Staples P, Torous J, Barnett I, et al. A comparison of passive and active estimates of sleep in a cohort with schizophrenia. NPJ Schizophr 2017 Oct 16;3(1):37. [doi: 10.1038/s41537-017-0038-0] [Medline: 29038553]
93. Anxiety disorder. National Institute of Mental Health. URL: https://www.nimh.nih.gov/health/topics/anxiety-disorders [accessed 2026-02-27]
94. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
95. Joseph SM. Delusional Disorder: StatPearls Publishing; 2025. URL: https://www.ncbi.nlm.nih.gov/books/NBK539855 [accessed 2026-02-27]
96. Correll CU, Schooler NR. Negative symptoms in schizophrenia: a review and clinical guide for recognition, assessment, and treatment. Neuropsychiatr Dis Treat 2020 Feb 21;16:519-534. [doi: 10.2147/NDT.S225643] [Medline: 32110026]
97. Drug addiction (substance use disorder). Mayo Clinic. URL: https://www.mayoclinic.org/diseases-conditions/drug-addiction/symptoms-causes/syc-20365112 [accessed 2026-03-16]
98. Bae S, Ferreira D, Suffoletto B, et al. Detecting drinking episodes in young adults using smartphone-based sensors. Proc ACM Interact Mob Wearable Ubiquitous Technol 2017 Jun;1(2):1-36. [doi: 10.1145/3090051] [Medline: 35146236]
99. Wu T, Sherman G, Giorgi S, et al. Smartphone sensor data estimate alcohol craving in a cohort of patients with alcohol-associated liver disease and alcohol use disorder. Hepatol Commun 2023 Dec 1;7(12):e0329. [doi: 10.1097/HC9.0000000000000329] [Medline: 38055637]
100. Arnold Z, Larose D, Agu E. Smartphone inference of alcohol consumption levels from gait. In: 2015 International Conference on Healthcare Informatics (ICHI): IEEE; 2015:417-426. [doi: 10.1109/ICHI.2015.59]
101. Li R, Balakrishnan GP, Nie J, et al. Estimation of blood alcohol concentration from smartphone gait data using neural networks. IEEE Access 2021;9:61237-61255. [doi: 10.1109/ACCESS.2021.3054515]
102. Bae SW, Chung T, Islam R, et al. Mobile phone sensor-based detection of subjective cannabis intoxication in young adults: a feasibility study in real-world settings. Drug Alcohol Depend 2021 Nov;228:108972. [doi: 10.1016/j.drugalcdep.2021.108972]
103. Nandakumar R, Gollakota S, Sunshine JE. Opioid overdose detection using smartphones. Sci Transl Med 2019 Jan 9;11(474):eaau8914. [doi: 10.1126/scitranslmed.aau8914] [Medline: 30626717]
104. Epstein DH, Tyburski M, Kowalczyk WJ, et al. Prediction of stress and drug craving ninety minutes in the future with passively collected GPS data. NPJ Digit Med 2020;3(1):26. [doi: 10.1038/s41746-020-0234-6]
105. Sleep apnea. Mayo Clinic. URL: https://www.mayoclinic.org/diseases-conditions/sleep-apnea/symptoms-causes/syc-20377631 [accessed 2026-02-28]
106. Alqassim S, Ganesh M, Khoja S, Zaidi M, Aloul F, Sagahyroon A. Sleep apnea monitoring using mobile phones. In: 2012 IEEE 14th International Conference on E-Health Networking, Applications and Services (Healthcom): IEEE; 2012:443-446. [doi: 10.1109/HealthCom.2012.6379457]
107. Tiron R, Lyon G, Kilroy H, et al. Screening for obstructive sleep apnea with novel hybrid acoustic smartphone app technology. J Thorac Dis 2020 Aug;12(8):4476-4495. [doi: 10.21037/jtd-20-804] [Medline: 32944361]
108. Capecci M, Pepa L, Verdini F, Ceravolo MG. A smartphone-based architecture to detect and quantify freezing of gait in Parkinson's disease. Gait Posture 2016 Oct;50:28-33. [doi: 10.1016/j.gaitpost.2016.08.018] [Medline: 27567449]

XSL•FO
RenderX

109. Ellis RJ, Ng YS, Zhu S, et al. A validated smartphone-based assessment of gait and gait variability in Parkinson's disease. PLoS ONE 2015;10(10):e0141694. [doi: 10.1371/journal.pone.0141694]
110. Kostikis N, Hristu-Varsakelis D, Arnaoutoglou M, Kotsavasiloglou C. Smartphone-based evaluation of parkinsonian hand tremor: quantitative measurements vs clinical assessment scores. Annu Int Conf IEEE Eng Med Biol Soc 2014;2014:906-909. [doi: 10.1109/EMBC.2014.6943738] [Medline: 25570106]
111. Barrantes S, Egea AJS, Rojas HAG, et al. Differential diagnosis between Parkinson's disease and essential tremor using the smartphone's accelerometer. PLoS ONE 2017;12(8):e0183843. [doi: 10.1371/journal.pone.0183843] [Medline: 28841694]
112. Ware S, Knouse LE, Draz I, Enikeeva A. Predicting ADHD symptoms using smartphone sensing data. In: Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers: Association for Computing Machinery; 2022:503-505. [doi: 10.1145/3544793.3563430]
113. Servia-Rodríguez S, Rachuri KK, Mascolo C, Rentfrow PJ, Lathia N, Sandstrom GM. Mobile sensing at the service of mental well-being: a large-scale longitudinal study. In: Proceedings of the 26th International Conference on World Wide Web: International World Wide Web Conferences Steering Committee; 2017:103-112. [doi: 10.1145/3038912.3052618]
114. Lane ND, Lin M, Mohammod M, et al. BeWell: sensing sleep, physical activities and social interactions to promote wellbeing. Mobile Netw Appl 2014 Jun;19(3):345-359. [doi: 10.1007/s11036-013-0484-5]
115. What are sleep disorders? American Psychiatric Association. URL: https://www.psychiatry.org/patients-families/sleep-disorders/what-are-sleep-disorders [accessed 2026-02-28]
116. Thakur KT, Albanese E, Giannakopoulos P, Jette N, Linde M, Prince MJ, et al. Neurological disorders. In: Mental, Neurological, and Substance Use Disorders: Disease Control Priorities, Third Edition (Volume 4): The International Bank for Reconstruction and Development / The World Bank; 2016:87-107. [doi: 10.1596/978-1-4648-0426-7_ch5]
117. Parkinson's disease. National Institute of Neurological Disorders and Stroke. URL: https://www.ninds.nih.gov/health-information/disorders/parkinsons-disease [accessed 2026-02-28]
118. Charron E, White A, Carlston K, et al. Prospective acceptability of digital phenotyping among pregnant and parenting people with opioid use disorder: a multisite qualitative study. Front Psychiatry 2023;14:1137071. [doi: 10.3389/fpsyt.2023.1137071] [Medline: 37139320]
119. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y) 2023 Sep 8;4(9):100804. [doi: 10.1016/j.patter.2023.100804] [Medline: 37720327]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
EMA: ecological momentary assessment
GAD-7: Generalized Anxiety Disorder-7
ML: machine learning
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
SUD: substance use disorder
YMRS: Young Mania Rating Scale

Edited by S Brini; submitted 15.Sep.2025; peer-reviewed by D Lottridge, V Vojdanovski; accepted 28.Jan.2026; published 24.Mar.2026.

Please cite as:
Dumas A, Hokayem J, Goodman G, Venkatasubramanian K, Chai P
Smartphone-Based Digital Phenotyping Across Health Conditions: Scoping Review
J Med Internet Res 2026;28:e84146
URL: https://www.jmir.org/2026/1/e84146
doi:10.2196/84146

© Arlen Dumas, Joanne Hokayem, Georgia Goodman, Krishna Venkatasubramanian, Peter Chai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effects of Virtual Reality–Based Interventions for Promoting Physical Activity in Patients With Heart Failure: Systematic Review

Jeong-Ah Ahn[1], PhD; Jung Eun Lee[2], PhD; Kyoung-A Kim[3], PhD

[1]College of Nursing and Research Institute of Nursing Science, Ajou University, Suwon, Republic of Korea

[2]College of Nursing, University of Rhode Island, Kingston, RI, United States

[3]Suwon Women's University, 72, Onjeong-ro, Gwonseon-gu, Suwon, Republic of Korea

Corresponding Author:
Kyoung-A Kim, PhD
Suwon Women's University, 72, Onjeong-ro, Gwonseon-gu, Suwon, Republic of Korea

Abstract

Background: Heart failure (HF) is a progressive chronic condition associated with reduced physical and functional capacity, psychological burden, cognitive decline, and diminished quality of life (QOL). Although exercise-based cardiac rehabilitation is beneficial, participation remains low due to accessibility, physical constraints, and motivational barriers. Virtual reality (VR)–based interventions, including immersive platforms and exergaming, may enhance accessibility and engagement and promote physical activity through interactive experiences. However, evidence regarding their effectiveness in patients with HF remains fragmented.

Objective: This systematic review synthesized current evidence on the effects of VR-based interventions on physical activity, psychosocial outcomes, and self-management behaviors in patients with HF.

Methods: We systematically searched PubMed, CINAHL, Embase, and Scopus for studies published within the past 10 years. Randomized controlled trials (RCTs) and non-RCT interventional studies involving adults with HF who participated in VR-based interventions were eligible. Outcomes included physical activity or exercise capacity, psychological well-being, self-management, and QOL. The reviewers screened articles, extracted data, and assessed risk of bias using version 2 of the Cochrane risk-of-bias tool for randomized trials for RCTs and the Risk of Bias in Nonrandomized Studies of Interventions tool for non-RCTs. The review adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 and PRISMA-S (PRISMA extension for reporting literature searches) guidelines.

Results: A total of 10 studies met the inclusion criteria, comprising 7 (70%) RCTs and 3 (30%) non-RCTs. Studies were conducted across multiple countries and predominantly included older adults (mean age ≥65 years). Most interventions were home based, with exergaming as the most frequent modality, followed by immersive VR cycling and digital coaching programs, delivered over 4 to 12 weeks. Across studies, VR-based interventions were associated with improvements in exercise capacity (n=6, 60% of the studies), physical activity (n=5, 50%), and QOL (n=4, 40%). Three of the studies (30%) reported reductions in depressive symptoms, whereas effects on anxiety and self-efficacy were inconsistent. Adherence and usability were high across studies, and no intervention-related adverse events were reported. However, the risk of bias was rated as "some concerns" or "high" in several domains, and heterogeneity in intervention design and outcome measurement, along with small samples, limited pooled synthesis and overall certainty of evidence.

Conclusions: VR-based interventions show promise as accessible and engaging approaches to promote physical activity and support rehabilitation in patients with HF, particularly in home-based settings. Across the included studies, VR interventions were generally associated with improvements in exercise capacity, physical activity, QOL, and depressive symptoms, with high adherence and no reported safety concerns. However, interpretation is limited by heterogeneity in intervention design, small sample sizes, and methodological constraints. Future research should prioritize larger, rigorously designed trials to support sustained clinical impact.

(J Med Internet Res 2026;28:e86567) doi:10.2196/86567

KEYWORDS

virtual reality; heart failure; physical activity; systematic review; rehabilitation

Introduction

Heart failure (HF) is a chronic, progressive condition that affects millions of people worldwide, imposing a substantial burden on individuals, families, and health care systems. As the final pathway of various cardiovascular diseases, the prevalence of HF continues to rise due to advances in cardiac care that have improved survival rates, along with the global aging population [1,2]. HF is associated not only with physiological decline but

XSL•FO
RenderX

also with multidimensional challenges that impact patients' psychological, cognitive, and social well-being [1]. Common symptoms such as fatigue, dyspnea, and exercise intolerance limit physical activity and may lead to deconditioning, reduced mobility, and loss of independence [3]. In addition, depression, anxiety, cognitive impairment, and diminished quality of life (QOL) are highly prevalent and further complicate disease management and self-care [4,5].

Maintaining adequate physical activity is an essential component of HF management and rehabilitation. Evidence consistently indicates that regular exercise improves exercise tolerance, muscle strength, endothelial function, and overall cardiovascular health [6]. Moreover, physical activity has positive effects on psychological outcomes such as mood, motivation, and perceived QOL. However, many patients with HF remain inactive due to fear of symptom exacerbation, lack of motivation, comorbidities, and limited access to supervised rehabilitation programs [7]. Although traditional cardiac rehabilitation (CR) is effective, participation rates remain low because of transportation barriers, cost and time burdens, and reduced self-efficacy [8,9]. Therefore, innovative and patient-centered approaches are needed to enhance participation and promote sustainable physical activity among individuals with HF.

In this context, virtual reality (VR) has emerged as a promising technological tool for health promotion and disease management. VR refers to computer-generated, interactive, 3D environments that simulate real or imagined scenarios, allowing users to engage in immersive experiences [10]. These environments can elicit strong sensory and emotional engagement, making activities more enjoyable, motivating, and meaningful. The immersive nature of VR creates a sense of presence—an illusion of "being there"—which enhances focus, reduces distraction, and increases adherence to therapeutic exercises [11]. In health care, VR has been applied in various fields, including pain management, mental health therapy, motor rehabilitation, and chronic disease management [12-15]. Advances in hardware and software have made VR systems increasingly accessible, portable, and cost-effective, expanding their potential for clinical and home-based rehabilitation.

VR-based interventions have demonstrated positive outcomes across diverse patient populations. In neurological and musculoskeletal rehabilitation, VR programs have been shown to improve balance, motor coordination, and motivation [16-18]. In patients with chronic conditions such as stroke, Parkinson disease, and chronic obstructive pulmonary disease, VR-based exercise interventions have contributed to enhanced physical performance and QOL [19,20]. Moreover, VR provides a safe and controlled environment for graded physical activity tailored to patients' physical capacities and psychological needs. The real-time feedback and gamified nature of VR exercises can foster self-efficacy and promote behavior change [21]. These features are particularly valuable for patients with HF, who often face both physical limitations and psychological barriers to exercise participation. Therefore, VR-based rehabilitation may represent a paradigm shift from conventional, clinic-based rehabilitation programs toward more interactive and personalized care.

Despite growing evidence supporting the benefits of VR-based interventions, their application in HF populations remains relatively limited. Therefore, a comprehensive review of current evidence is required to understand both the potential and the limitations of VR in this population. Given the multifaceted nature of HF and the capacity of VR-based programs to improve physical activity, a systematic review on this topic has significant clinical implications. Evaluating the effects of VR interventions on physical activity and related outcomes among patients with HF can help clarify whether VR serves as an effective adjunct or alternative to traditional rehabilitation. Furthermore, identifying the types of VR interventions, their key components, and outcome measures can inform best practices for future program design.

Therefore, this systematic review aimed to synthesize current evidence on the effects of VR-based interventions for promoting physical activity, psychosocial outcomes, and self-management behaviors in patients with HF. By synthesizing and critically appraising existing evidence, this study aimed to establish a scientific foundation for incorporating VR technologies into chronic disease self-management strategies for patients with HF.

Methods

Search Strategy

This systematic review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [22] and PRISMA-S (PRISMA extension for reporting literature searches) guidelines [23]. A comprehensive search was conducted in 4 electronic databases: PubMed, CINAHL, Embase, and Scopus. Each database was searched individually on its own platform (PubMed via the National Center for Biotechnology Information, CINAHL via EBSCOhost, Embase via Elsevier, and Scopus via Elsevier); no simultaneous multidatabase searching on a single platform was conducted. The search covered the period from January 1, 2016, to December 31, 2025, and the final search was completed on January 2, 2026. All retrieved articles were compiled, and duplicates were removed. The bibliographic management software EndNote (version 21; Clarivate Analytics) was used to manage and organize references.

The search terms were developed based on the population, intervention, comparison, and outcomes framework: "heart failure" (population); "virtual reality," "immersive virtual reality," "augmented reality," "mixed reality," "exergaming," or "exergame" (intervention); "standard care" or "traditional exercise programs" (comparison); and "physical activity," "exercise capacity," "psychological and cognitive outcomes," "self-management behaviors," and "quality of life" (outcomes), including relevant synonyms and related terms. Search terms were combined using Boolean operators ("AND" and "OR") and adapted to each database's indexing system. Both MeSH (Medical Subject Headings) and free-text keywords were used to ensure a comprehensive search. The search strategy was further expanded to include a wider range of controlled vocabulary terms and text words related to HF (eg, "cardiac failure" and "cardiac insufficiency"), virtual and extended reality

XSL•FO
RenderX

technologies (eg, "virtual reality exposure therapy," "mixed reality," "smart glasses," and "head-mounted display"), exergaming and active video gaming, and physical activity–related outcomes (eg, "exercise" and "fitness"). The search strategy applied limits to publication year (past 10 years) and language (English only). The publication period was limited to the last 10 years to capture current evidence. The full reproducible search strategies for each database are provided in Multimedia Appendix 1 in accordance with PRISMA-S recommendations. The search strategy was peer reviewed by an independent librarian to ensure comprehensiveness and accuracy. Two reviewers independently conducted the database searches and verified the search outputs.

The inclusion criteria were (1) studies involving patients with HF, (2) VR-based interventional studies, (3) randomized controlled trials (RCTs) or non-RCTs, and (4) articles published in peer-reviewed journals. Eligible studies were required to include at least one outcome related to physical activity (eg, exercise capacity, activity level, or performance), whereas additional outcomes such as psychological status, self-management behaviors, or QOL were also considered. Studies that did not aim to promote or measure physical activity in patients with HF were excluded. Exclusion criteria included reviews, protocols, conference abstracts, noninterventional studies, and studies that did not report physical activity or related outcomes, as well as those lacking VR components.

Two independent reviewers conducted the screening and selection process. Any discrepancies regarding study inclusion or exclusion were resolved through discussion and consensus. When consensus could not be reached, a third reviewer was consulted. Data were extracted for all outcomes related to physical activity, exercise capacity, psychosocial status, self-management behaviors, and QOL across all reported measures and time points in each study. Physical activity–related outcomes were considered primary outcomes for synthesis and interpretation, while psychosocial and self-management outcomes were treated as secondary outcomes.

In addition to outcome data, information on study characteristics (country and study design), participant characteristics (sample size, age, and sex), and intervention characteristics (type, duration, frequency, and setting) was extracted using a predefined data extraction form. When information was missing or unclear, data were reported as not available, and no assumptions or imputations were made; study authors were not contacted for clarification.

Risk-of-Bias Assessment

The methodological quality and risk of bias of the included studies were evaluated using standardized tools appropriate to each study design. For RCTs, version 2 of the Cochrane risk-of-bias tool for randomized trials was used [24]. This tool assesses five domains of potential bias: (1) the randomization process, (2) deviations from intended interventions, (3) missing outcome data, (4) measurement of outcomes, and (5) selection of the reported results. For nonrandomized studies, the Risk of Bias in Nonrandomized Studies of Interventions tool was used [25]. The Risk of Bias in Nonrandomized Studies of Interventions evaluates seven domains: (1) confounding, (2) selection of participants, (3) classification of interventions, (4) deviations from intended interventions, (5) missing data, (6) measurement of outcomes, and (7) selection of the reported results.

Two reviewers independently assessed each study using these tools. Any disagreements were resolved through discussion or, when necessary, consultation with a third reviewer. The overall risk of bias was categorized into 3 levels: low risk, some concerns or moderate, or high or serious risk.

Synthesis Methods

Studies included in the narrative synthesis met the inclusion criteria and reported outcomes related to physical activity or exercise capacity in patients with HF. Studies were grouped by intervention type, study setting, and outcome domain and synthesized descriptively using reported data without conversion or imputation; meta-analysis was not conducted due to heterogeneity across studies, and potential sources of heterogeneity were explored descriptively.

Results

General Characteristics and Research Methodology of the Included Studies

A PRISMA flow diagram (Figure 1) was used to illustrate the study selection process. A total of 832 records were identified through PubMed, CINAHL, Embase, and Scopus. Of these 832 records, after removing 274 (32.9%) duplicates, 558 (67.1%) studies remained for title and abstract screening. Of these 558 articles, 540 (96.8%) articles were excluded for not meeting the inclusion criteria. The full texts of 18 studies were then reviewed, and 8 (44.4%) were excluded due to ineligible populations, interventions, or outcome reporting—specifically, studies aiming to improve knowledge and self-care, to manage pain, to compare hemodynamic parameters, to confirm system usability and satisfaction, or to assess symptom experience—resulting in a final selection of 10 (55.6%) articles. These comprised 7 RCTs and 3 nonrandomized single-arm studies. Table 1 and Multimedia Appendix 2 [26-35] summarize the general characteristics and research methodologies of the included studies. The studies were conducted in various countries, including Sweden, the United States, Germany, Israel, the Netherlands, Brazil, Italy, Romania, and Spain. Sample sizes ranged from 10 to 605 participants, and most studies (6/10, 60%) included older adults, with a mean age of ≥65 years. In total, 80% (n=8) of the interventions were implemented in home-based settings, whereas 20% (n=2) were conducted in hospital environments.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the study selection process. VR: virtual reality.

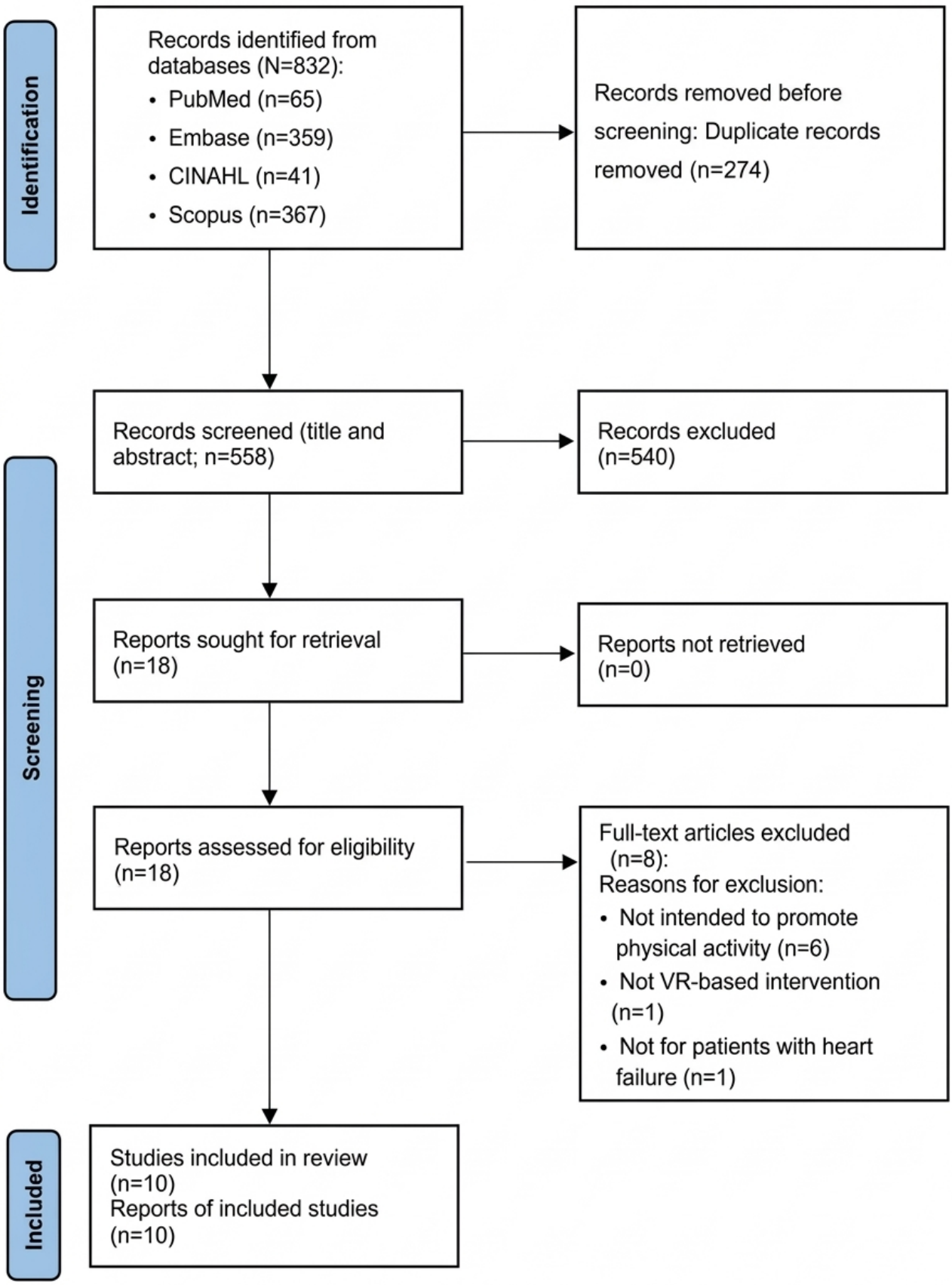

XSL•FO
RenderX

Table . Characteristics of the included studies (N=10).

Variable and category	Studies, n (%)
Disease of participants	
Heart failure	10 (100)
Country[a]	
Sweden	5 (50)
United States	4 (40)
Germany	3 (30)
Israel	3 (30)
The Netherlands	3 (30)
Brazil	2 (20)
Italy	2 (20)
Others (Romania and Spain)	2 (20)
Mean age of participants (y)	
<65	3 (30)
≥65	6 (60)
Not mentioned	1 (10)
Sample size	
<50	4 (40)
50-100	3 (30)
>100	3 (30)
Use of theoretical framework	
Yes	1 (10)
No	9 (90)
Study design	
Randomized controlled trial	7 (70)
1-group (single-arm) study	3 (30)
Setting	
Home	8 (80)
Hospital	2 (20)
Type of intervention	
Exergame	6 (60)
Immersive VR[b]-assisted cycle ergometer	2 (20)
Sensor-controlled digital game	1 (10)
Virtual application for cardiac rehabilitation	1 (10)
Intervention period	
During hospitalization	2 (20)
4 wk	2 (20)
12 wk	6 (60)
Follow-up frequency	
Once	7 (70)
3 times	3 (30)

[a]Some studies were conducted across multiple countries simultaneously. Therefore, the total number of countries listed exceeds the total number of included studies (N=10) as each multinational study was counted once per country.

XSL•FO
RenderX

[b]VR: virtual reality.

Only 10% (1/10) of the studies [29] applied a theoretical framework—the Fogg behavior model—to guide intervention design, whereas the remaining studies were empirically developed. Intervention durations varied, with most (6/10, 60%) lasting 12 weeks, 20% (2/10) lasting 4 weeks, and 20% (2/10) being conducted during hospitalization. The frequency of follow-up assessments ranged from a single posttest to up to 3 follow-up evaluations.

Regarding the types of VR-based interventions, exergaming was the most common approach, including Nintendo Wii–based programs (5/10, 50%) and a mobile exergame (Heart Farming; 1/10, 10%), followed by immersive VR-assisted cycle ergometer training (2/10, 20%), sensor-controlled digital games (1/10, 10%), and a virtual assistant–based CR program (1/10, 10%). A total of 40% (4/10) of the included studies [30-33] originated from the same RCT, the HF-Wii study. These publications represented secondary or subgroup analyses conducted in multinational cohorts (Sweden, Italy, the Netherlands, Israel, Germany, and the United States) that explored distinct outcome domains such as exercise capacity, psychological well-being, and cognitive predictors. To prevent double counting of data, these HF-Wii–derived studies were narratively synthesized and summarized separately rather than aggregated in quantitative comparisons.

Risk of Bias

Of the 7 RCTs, 6 studies were rated as having some concerns, whereas 1 study showed a high risk of bias. Of the 3 non-RCTs, 2 studies showed a moderate risk of bias, whereas 1 feasibility study was judged to have a serious risk of bias. The summarized assessments are presented in Figure 2 [29-35] and in Figure 3 [26-28].

Figure 2. Summary of the risk-of-bias assessment using version 2 of the Cochrane risk-of-bias tool for randomized controlled trials.

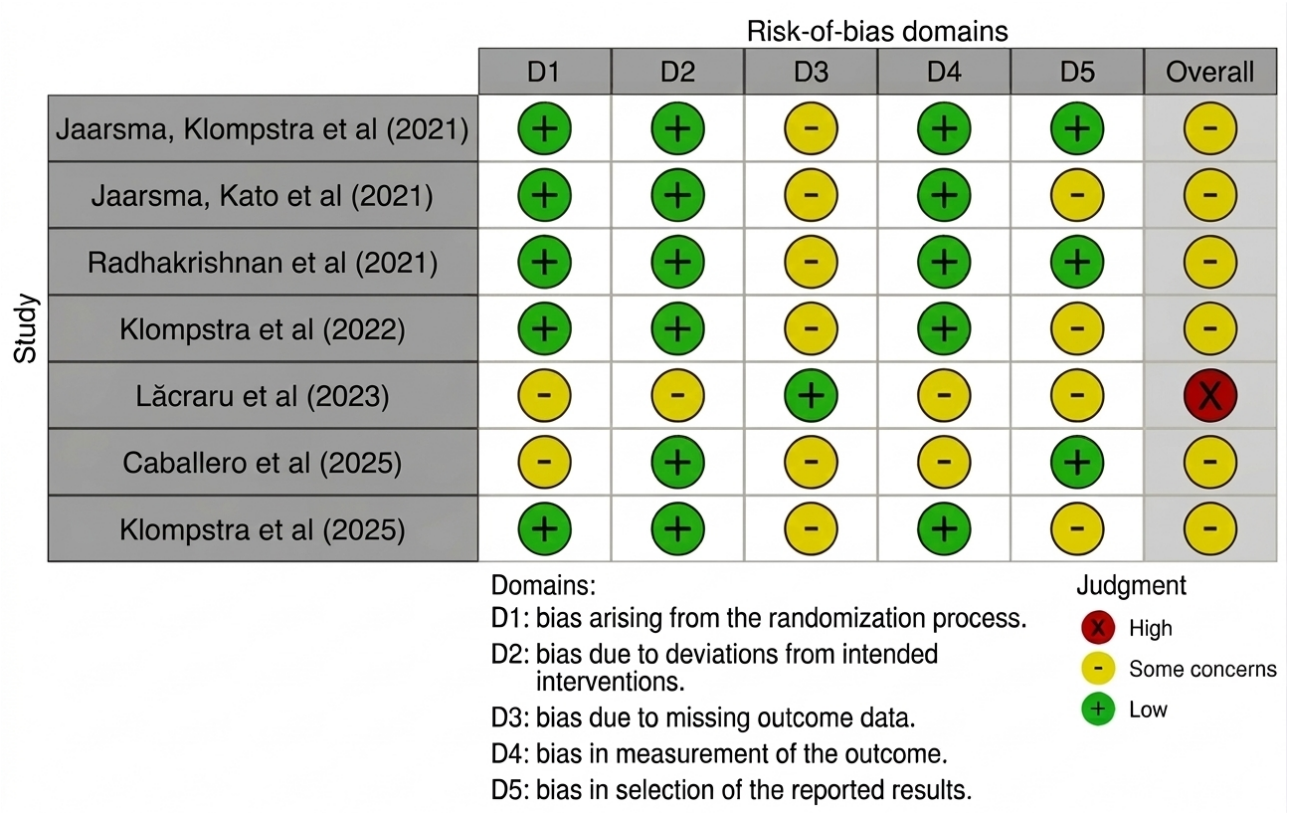

XSL•FO
RenderX

Figure 3. Summary of the risk-of-bias assessment using the Risk of Bias in Nonrandomized Studies of Interventions tool.

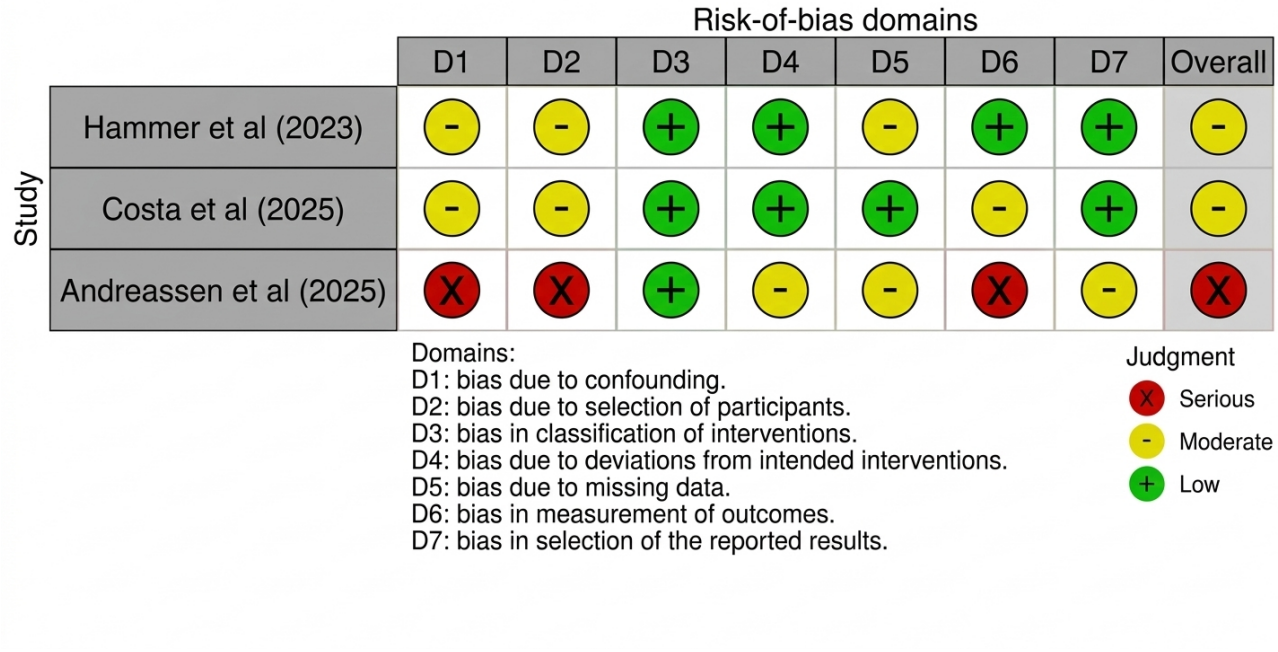

Effects of VR-Based Interventions in Patients With HF

Table 2 and Multimedia Appendix 2 [26-35] show the detailed measurements, tools, and main results related to the effects of VR-based interventions in patients with HF.

XSL•FO
RenderX

Table . Outcome measurements of the included studies (N=10).

Variable[a] and category	Studies, n (%)
Primary outcomes (physical activity related)	
Exercise capacity (6MWT[b] and VO_2 max[c])	6 (60)
Exercise motivation	3 (30)
Exercise self-efficacy	3 (30)
Self-reported physical activity	2 (20)
Physical activity behaviors	2 (20)
Muscle strength	1 (10)
Daily walking distance	1 (10)
Mobilization experience	1 (10)
Sedentary time	1 (10)
Secondary outcomes	
Anxiety and depression	5 (50)
Quality of life	4 (40)
Well-being	2 (20)
HF[d] self-management behaviors	2 (20)
HF self-management knowledge	1 (10)
HF self-efficacy	1 (10)
Cognitive function	1 (10)
HF hospitalization	1 (10)
Feasibility (acceptability, adherence, and retention)	3 (30)
Usability (user experience, system usability, and technology acceptance)	3 (30)
Safety	1 (10)

[a]In some cases, multiple reports or publications were derived from the same study. These entries share the same study population or dataset.

[b]6MWT: 6-minute walk test.

[c]VO_2 max: maximal oxygen consumption.

[d]HF: heart failure.

Primary and Secondary Outcome Measurements

The most frequently assessed primary outcomes were exercise capacity—defined as physiological performance indicators such as the 6-minute walk test (6MWT) or maximal oxygen consumption (VO_2 max)—reported in 60% (6/10) of the studies. In contrast, physical activity, treated as a distinct behavioral domain, was measured through self-reported activity levels, behavioral activity logs, muscle strength tests, mobilization experience, or sedentary time. Other primary outcomes included exercise motivation and self-efficacy (3/10, 30% of the studies each).

Secondary outcomes included anxiety and depression (5/10, 50%), QOL (4/10, 40%), well-being (2/10, 20%), and HF self-management behaviors (2/10, 20%), as well as intervention-related variables such as feasibility, usability, and safety.

Effects on Physical Outcomes

Overall, VR-based interventions demonstrated positive trends in improving physical activity and exercise performance among patients with HF. In total, 50% (5/10) of the studies, which used home-based exergaming Nintendo Wii programs, reported improvements in 6MWT distances, with 2 of these studies showing statistically significant gains over 3 to 6 months ($P<.05$). For example, Klompstra et al [32] reported significant increases in exercise capacity and reductions in fatigue and dyspnea after 3 to 6 months, although these effects diminished by 12 months. Similarly, Hammer et al [26] found that VR exergaming improved 6MWT performance and QOL in patients with HF supported by left ventricular assist devices ($P=.02$). In addition, the mobile exergame (Heart Farming) demonstrated that daily walking targets were manageable and achievable, as confirmed by in-app tracking of walking distance, supporting the feasibility of integrating regular walking activity into daily routines [27].

The vCare virtual assistant app trial by Lăcraru et al [34] demonstrated significant improvements in VO_2 max ($P=.002$)

XSL•FO
RenderX

and reductions in low-density lipoprotein cholesterol (P<.05) compared with standard care.

Among hospitalized patients, Caballero et al [35] and Costa et al [28] found that immersive VR-assisted mobilization and exercise were safe, feasible, and associated with high levels of exercise enjoyment and positive usability ratings.

Effects on Psychological and Cognitive Outcomes

Across the included studies, VR-based programs showed mixed but generally positive psychological effects. In total, 50% (5/10) of the studies assessed anxiety and depression, with 3 of these studies reporting significant postintervention reductions (P<.05). Lăcraru et al [34] found significant improvements in depression scores (P=.03), whereas anxiety levels remained unchanged. In contrast, Klompstra et al [32] observed no significant changes in anxiety or depression despite improvements in physical outcomes.

A total of 20% (2/10) of the studies evaluated overall well-being, reporting small to moderate improvements immediately following the intervention. Cognitive function was assessed in one secondary analysis [31], which identified lower baseline cognitive scores as a predictor of nonimprovement in exercise performance (odds ratio 0.87, 95% CI 0.80-0.94), suggesting that cognitive status may influence the effectiveness of VR-based interventions.

Effects on Self-Management and QOL Outcomes

In total, 20% (2/10) of the studies focused on self-management behaviors. The sensor-controlled digital game intervention (Heart Health Mountain) by Radhakrishnan et al [29] significantly improved HF self-management knowledge (P<.05), QOL (P<.01), and adherence behaviors such as daily weight monitoring and physical activity (r=0.9; P<.001). Self-efficacy improved over 24 weeks, although between-group differences were not statistically significant. Additionally, reduced hospitalization rates were observed in both intervention and control groups over the 6-month follow-up period. The vCare digital coaching program [34] did not directly measure self-management behaviors but functioned as a self-management support tool. The intervention improved VO_2 max (P=.002), QOL (P=.007), and depressive symptoms, whereas participants reported high usability and motivation for continued engagement in home-based rehabilitation.

QOL was assessed in 40% (4/10) of the studies [26,29,32,34], all of which reported improvements following VR-based interventions.

The Heart Health Mountain program [29] demonstrated significant increases in QOL at 6, 12, and 24 weeks (P<.01), along with enhanced self-efficacy and adherence to daily self-care activities.

Similarly, the vCare virtual assistant app [34] improved QOL (P=.007) and reduced depressive symptoms (P=.03), suggesting that personalized, feedback-driven home rehabilitation can yield psychosocial benefits in addition to physiological gains. In a study of left ventricular assist device–supported patients with HF, Hammer et al [26] found significant improvements in both exercise capacity and QOL after a 4-week home-based exergaming program.

Finally, Klompstra et al [32] observed notable improvements in QOL and reductions in fatigue and dyspnea during the 3- to 6-month follow-up period, although these effects diminished by 12 months.

Feasibility, Usability, and Safety of VR-Based Interventions

Feasibility was reported in 30% (3/10) of the studies [26,27,29], showing high adherence, acceptability, and retention rates for home-based exergaming programs. Usability outcomes, assessed in the vCare [34] and VR-assisted cycling studies [28], indicated good to excellent user acceptance, with mean System Usability Scale scores exceeding 68 points. Similarly, the mobile exergame (Heart Farming) was perceived as easy to use, adaptable to individual needs, and engaging, with no major acceptability concerns reported [27]. Among hospitalized patients, immersive VR-assisted mobilization and exercise were also reported to be safe, feasible, and well tolerated [28,35].

No adverse events, cybersickness, or device-related safety concerns were reported in any of the included studies, indicating that VR-based interventions were well tolerated and safe for patients with HF.

Discussion

Principal Findings

This systematic review synthesized current evidence on the effects of VR-based interventions on physical activity, psychological outcomes, and self-management behaviors in patients with HF. Overall, the findings suggest that VR-based interventions can effectively enhance physical activity, exercise performance, and selected psychosocial outcomes in this population. Improvements in exercise capacity and QOL were the most consistently reported outcomes, while effects on psychological and behavioral domains were more variable across studies. The results also confirm the feasibility, usability, and safety of VR-based interventions in both hospital and home settings, demonstrating their potential as complementary or alternative strategies to traditional CR programs for patients with HF.

Consistent with prior meta-analyses on VR-based rehabilitation in chronic disease populations [20,36,37], the included studies demonstrated that VR can improve exercise capacity and physical activity levels in patients with HF. Exercise capacity, measured primarily through the 6MWT and VO_2 max, showed significant improvements in several studies (4/10, 40%), indicating the physiological benefits of engaging in VR-mediated physical activity. Home-based exergaming programs, such as the HF-Wii intervention, encouraged patients to exercise safely and independently while maintaining adherence levels comparable to those of center-based CR [32]. These findings align with those of previous research reporting that gamified exercise enhances motivation, perceived enjoyment, and long-term engagement with physical activity among individuals with chronic cardiovascular conditions [6,21].

XSL•FO
RenderX

The ability of VR to provide immersive and interactive exercise experiences may help overcome common barriers to physical activity in HF populations, including fear of symptom exacerbation, low motivation, and limited access to supervised programs [7,8]. By integrating features such as real-time feedback, adjustable difficulty levels, and visual immersion, VR facilitates safe, individualized exercise experiences that enhance self-efficacy and engagement. Several trials (3/10, 30%) in this review reported high adherence rates (>85%) and positive user feedback regarding enjoyment, indicating that VR-based training is both acceptable and feasible for older adults. These results suggest that VR can serve as a practical extension of center-based CR carried out in the home, especially for individuals who face logistical or psychosocial barriers that hinder participation in traditional center-based programs.

Notably, the interventions in this review ranged from nonimmersive exergaming (eg, Nintendo Wii) to more immersive VR-assisted cycling systems. Although both formats generally demonstrated beneficial effects, the current evidence does not clearly indicate that higher immersion consistently produces superior outcomes, partly because intervention content, exercise intensity, and patient characteristics varied considerably across studies. Therefore, the level of immersion should be viewed as one of several interacting design features rather than the sole driver of effectiveness. Therefore, future research should explicitly compare different levels of immersion to identify which VR design characteristics—such as sensory stimulation, interaction fidelity, or gamification—are most strongly associated with adherence and clinical outcomes.

Nevertheless, the magnitude of improvement in exercise capacity varied across studies, and the sustainability of these effects beyond short-term follow-up remains uncertain. Some trials demonstrated attenuation of benefits after 6 to 12 months [32], implying that continued engagement or periodic reinforcement may be necessary to maintain gains. Future studies should explore strategies such as adaptive difficulty adjustment, remote coaching, and integration with wearable monitoring technologies to support long-term adherence and sustained physical activity.

In addition to physical outcomes, several studies (4/10, 40%) revealed potential psychological and cognitive benefits of VR-based interventions. Improvements in depressive symptoms and emotional well-being were observed in 30% (3/10) of the studies [26,29,34], consistent with previous findings showing that VR-based exercise can enhance mood, reduce stress, and foster social connectedness [11,17,38]. These effects may be mediated by mechanisms such as endorphin release from physical activity, distraction from illness-related distress, and enhanced self-efficacy through interactive goal attainment [39].

Moreover, VR environments provide an engaging, multisensory experience that may alleviate anxiety and fear associated with exercise, particularly among patients with HF who often perceive exertion as a potential trigger for worsening symptoms [4,5]. By enabling safe, controlled participation in virtual exercise, VR interventions may help reduce psychological resistance to activity. Therefore, the findings of this review support the potential of VR in promoting not only physical rehabilitation but also emotional adjustment and self-confidence in managing chronic illness.

Cognitive function, though less frequently assessed, appears to influence the efficacy of VR interventions. Jaarsma et al [30] reported that patients with lower baseline cognitive performance were less likely to achieve improvements in physical outcomes, suggesting that cognitive impairment may hinder engagement with VR technologies. This is likely because VR requires users to attend to visual-spatial cues, remember instructions, and coordinate motor actions; therefore, attention, working memory, and executive functioning play an important role in task performance and learning using VR [40]. When these abilities are reduced, patients may struggle to follow VR tasks or maintain engagement, which may attenuate treatment effects. Prior evidence indicates that cognitive decline, common in HF due to cerebral hypoperfusion, negatively affects self-care and rehabilitation adherence [41]. Accordingly, future research should consider cognitive screening to guide intervention tailoring and the potential value of incorporating cognitive training features into VR programs.

Although relatively few studies (2/10, 20%) examined self-management outcomes, there is available evidence suggesting that VR-based interventions can positively influence health-related behaviors and disease knowledge. The Heart Health Mountain program [29] demonstrated significant improvements in HF knowledge, QOL, and adherence to daily self-care tasks such as weight monitoring and physical activity logging. These results highlight VR's educational potential—its interactive and gamified format can promote active learning, repetition, and reinforcement of self-management principles [15]. Integrating personalized behavioral feedback into VR platforms may further enhance self-regulation and motivation, supporting long-term lifestyle modification.

Nevertheless, behavior changes were inconsistent across studies. Some trials 2/10 (20%) reported only transient improvements that diminished over time, possibly due to limited intervention duration or insufficient strategies to maintain engagement [32]. Future research should investigate hybrid models that combine VR-based education with telemonitoring, remote coaching, or community support to promote sustained behavior change and long-term adherence. In addition, only 10% (1/10) of the studies used a clear theoretical framework to guide intervention design. Given that many VR programs aimed not only to increase physical activity but also to influence psychosocial or self-management behaviors, theory-driven approaches will be essential in future research to clarify mechanisms of change and enhance the sustainability of intervention effects.

One of the most encouraging findings of this review is the strong evidence for the feasibility, usability, and safety of VR-based interventions for patients with HF. No adverse cardiovascular events, cybersickness, or equipment-related injuries were reported across the 10 included studies. Adherence and retention rates were high, particularly in home-based interventions, indicating good acceptability among older adults. Usability assessments consistently yielded satisfactory results, suggesting that the VR systems used were intuitive and well tolerated even

XSL•FO
RenderX

among participants with limited experience using digital technologies [18].

These findings are consistent with those of previous systematic reviews of VR-based rehabilitation in other chronic disease populations, where safety, enjoyment, and user engagement were identified as key facilitators of program success [16,20]. The use of commercially available, low-cost gaming systems (eg, Wii) in several studies (5/10, 50%) further supports the scalability and cost-effectiveness of VR interventions for HF management. As portable VR systems become more widely available, their integration into community and home settings will be increasingly feasible, offering opportunities to expand access to rehabilitation for underserved and mobility-limited populations.

Although 40% (4/10) of the included publications were derived from the same HF-Wii RCT, the findings across these reports were not identical. This variation can largely be explained by differences in follow-up duration and outcome selection. The main HF-Wii trial evaluated overall changes in exercise capacity at 3 months and reported no significant improvement, whereas the subsequent substudies targeted specific subgroups; additional follow-up periods (eg, 6 and 12 months); or secondary outcomes such as QOL, fatigue, and physical activity patterns. Consequently, some beneficial effects—particularly those observed at 3 to 6 months—appeared only in certain subanalyses and did not consistently persist across all publications. These differences highlight that the HF-Wii findings should not be interpreted as repeated independent evidence but rather as multiple analytic perspectives derived from a single parent dataset.

The findings of this review have several important implications. Clinically, VR-based interventions may serve as valuable adjuncts to traditional CR programs by offering flexible, engaging means to enhance physical activity and self-management, particularly for patients unable to attend in-person sessions. Health professionals can leverage VR platforms for remote monitoring, individualized goal setting, and motivational reinforcement. Integration of VR systems with wearable sensors and telehealth platforms could enable real-time feedback and progress tracking, fostering a more personalized and continuous rehabilitation experience [6].

From a research perspective, future RCTs should use larger sample sizes, extended follow-up periods, and standardized outcome metrics to allow for cross-study comparison and meta-analyses. Multidomain interventions combining physical, cognitive, and psychosocial training within VR environments may yield synergistic effects on overall health and well-being. Additionally, cost-effectiveness analyses and qualitative studies on patient experiences will be essential to inform clinical implementation and policy decisions. As VR technology continues to evolve, ensuring accessibility, ease of use, and cultural adaptability will be key to maximizing its impact on HF care.

Limitations

Despite this systematic review presenting promising results, several methodological limitations should be acknowledged. First, the number of high-quality RCTs was limited, and most studies (4/10, 40%) involved small sample sizes, reducing statistical power and generalizability. Second, there was substantial heterogeneity in intervention design, duration, intensity, and outcome measures, which precluded meta-analysis and limited the ability to draw direct comparisons across studies. The diversity of VR modalities—ranging from nonimmersive exergames to fully immersive environments—also complicates interpretation of specific active components responsible for the observed benefits. Third, follow-up periods were typically short (4 - 12 weeks), making it difficult to determine whether the observed improvements translate into sustained lifestyle changes or long-term clinical outcomes such as reduced hospitalization or mortality. Fourth, only 10% (1/10) of the studies used a theoretical framework to guide intervention development and behavioral engagement. Future studies should integrate behavior change theories and standardized outcome measures to strengthen conceptual and methodological rigor. In addition, 40% (4/10) of the included studies originated from the same parent trial (HF-Wii), resulting in overlapping participant samples and repeated use of the same trial dataset. To minimize potential bias from this overlap, these studies were treated as secondary analyses of a single RCT and were narratively summarized rather than combined quantitatively. Nonetheless, because these publications draw on the same underlying sample, their collective contribution should be interpreted with caution as they do not represent independent evidence. Finally, the search period was restricted to the last 10 years to reflect contemporary VR technology and current CR practice. While this approach increases clinical relevance to modern practice, it may also have excluded earlier pioneering VR studies.

Conclusions

This systematic review advances the research field by providing the first focused synthesis of recent VR-based interventions specifically designed to promote physical activity in patients with HF while concurrently evaluating psychosocial and self-management outcomes. This study critically assessed research methodologies, intervention characteristics, and clinical effectiveness within the HF-specific context. By clarifying both the demonstrated short-term benefits and the methodological limitations of existing studies, this review contributes a conceptual and empirical road map for the future development of standardized VR-supported rehabilitation in HF. Importantly, the findings highlight the real-world potential of VR as a scalable, safe, and engaging home-based strategy capable of addressing persistent barriers to traditional CR. As digital health infrastructure continues to expand, evidence-based VR interventions could become an integral component of comprehensive HF self-management and rehabilitation programs. Nevertheless, because only 10% (1/10) of the included studies explicitly incorporated a formal behavior change theory, future large-scale, rigorously designed RCTs should integrate theoretical frameworks to guide intervention development and evaluation. Theory-driven designs will be essential for enhancing methodological rigor and ensuring the long-term sustainability of physical, psychosocial, and self-management outcomes in VR-based HF rehabilitation.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank Yujin Kwon, a medical research librarian at Ajou University, for her guidance on systematic literature review methods in this study. They are also grateful to the editor and reviewers who helped improve this manuscript.

Funding

This work was supported by a National Research Foundation of Korea grant funded by the Korean government (Ministry of Science and ICT; RS-2025-24532981). This funding source had no role in the study design, data collection, data analysis, data interpretation, or writing of the manuscript.

Data Availability

All data generated or analyzed during this systematic review are included in this published article and its supplementary files.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies.
[PDF File, 118 KB - jmir_v28i1e86567_app1.pdf]

Multimedia Appendix 2
Detailed data extraction of each study.
[PDF File, 199 KB - jmir_v28i1e86567_app2.pdf]

Checklist 1
PRISMA checklist.
[PDF File, 150 KB - jmir_v28i1e86567_app3.pdf]

Checklist 2
PRISMA-S checklist.
[PDF File, 225 KB - jmir_v28i1e86567_app4.pdf]

Checklist 3
SWiM reporting items.
[PDF File, 63 KB - jmir_v28i1e86567_app5.pdf]

References

1. Authors/Task Force Members, McDonagh TA, Metra M, et al. 2023 focused update of the 2021 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure: developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur J Heart Fail 2024 Jan;26(1):5-17. [doi: 10.1002/ejhf.3024] [Medline: 38169072]
2. Savarese G, Lund LH. Global public health burden of heart failure. Card Fail Rev 2017 Apr;3(1):7-11. [doi: 10.15420/cfr.2016:25:2] [Medline: 28785469]
3. Zheng X, Zheng Y, Ma J, et al. Effect of exercise-based cardiac rehabilitation on anxiety and depression in patients with myocardial infarction: a systematic review and meta-analysis. Heart Lung 2019 Jan;48(1):1-7. [doi: 10.1016/j.hrtlng.2018.09.011] [Medline: 30366575]
4. Al-Sutari MM, Abdalrahim MS. Symptom burden and quality of life among patients with heart failure. SAGE Open Nurs 2024 Mar 28;10:23779608241242023. [doi: 10.1177/23779608241242023] [Medline: 38559894]
5. Freedland KE, Carney RM, Rich MW, Steinmeyer BC, Skala JA, Dávila-Román VG. Depression and multiple rehospitalizations in patients with heart failure. Clin Cardiol 2016 May;39(5):257-262. [doi: 10.1002/clc.22520] [Medline: 26840627]
6. Molloy C, Long L, Mordi IR, et al. Exercise-based cardiac rehabilitation for adults with heart failure. Cochrane Database Syst Rev 2024 Mar 7;3(3):CD003331. [doi: 10.1002/14651858.CD003331.pub6] [Medline: 38451843]
7. Dalal HM, Doherty P, Taylor RS. Cardiac rehabilitation. BMJ 2015 Sep 29;351:h5000. [doi: 10.1136/bmj.h5000] [Medline: 26419744]

XSL•FO
RenderX

8. Ades PA, Keteyian SJ, Wright JS, et al. Increasing cardiac rehabilitation participation from 20% to 70%: a road map from the Million Hearts cardiac rehabilitation collaborative. Mayo Clin Proc 2017 Feb;92(2):234-242. [doi: 10.1016/j.mayocp.2016.10.014] [Medline: 27855953]
9. Anderson L, Oldridge N, Thompson DR, et al. Exercise-based cardiac rehabilitation for coronary heart disease: Cochrane systematic review and meta-analysis. J Am Coll Cardiol 2016 Jan 5;67(1):1-12. [doi: 10.1016/j.jacc.2015.10.044] [Medline: 26764059]
10. Laver KE, Lange B, George S, et al. Virtual reality for stroke rehabilitation. Cochrane Database Syst Rev 2025 Jun 20;6(6):CD008349. [doi: 10.1002/14651858.CD008349.pub5] [Medline: 40537150]
11. Maples-Keller JL, Bunnell BE, Kim SJ, Rothbaum BO. The use of virtual reality technology in the treatment of anxiety and other psychiatric disorders. Harv Rev Psychiatry 2017;25(3):103-113. [doi: 10.1097/HRP.0000000000000138] [Medline: 28475502]
12. Goudman L, Jansen J, Billot M, et al. Virtual reality applications in chronic pain management: systematic review and meta-analysis. JMIR Serious Games 2022 May 10;10(2):e34402. [doi: 10.2196/34402] [Medline: 35536641]
13. Hajesmaeel Gohari S, Gozali E, Niakan Kalhori SR. Virtual reality applications for chronic conditions management: a review. Med J Islam Repub Iran 2019;33:67. [doi: 10.34171/mjiri.33.67] [Medline: 31456991]
14. Tabassum A, Ghaznavi I, Abd-Alrazaq A, Qadir J. Exploring the application of AI and extended reality technologies in metaverse-driven mental health solutions: scoping review. J Med Internet Res 2025 Aug 19;27:e72400. [doi: 10.2196/72400] [Medline: 40829151]
15. Triberti S, Repetto C, Riva G. Psychological factors influencing the effectiveness of virtual reality-based analgesia: a systematic review. Cyberpsychol Behav Soc Netw 2014 Jun;17(6):335-345. [doi: 10.1089/cyber.2014.0054] [Medline: 24892195]
16. Cano Porras D, Siemonsma P, Inzelberg R, Zeilig G, Plotnik M. Advantages of virtual reality in the rehabilitation of balance and gait: systematic review. Neurology 2018 May 29;90(22):1017-1025. [doi: 10.1212/WNL.0000000000005603] [Medline: 29720544]
17. Kim KA, Ahn JA. Effectiveness of immersive virtual reality simulation programs using head-mounted displays in promoting physical activity in older adults: a systematic review. Clin Simul Nurs 2024 Sep;94:101593. [doi: 10.1016/j.ecns.2024.101593]
18. Rodríguez-Mansilla J, Bedmar-Vargas C, Garrido-Ardila EM, et al. Effects of virtual reality in the rehabilitation of Parkinson's disease: a systematic review. J Clin Med 2023 Jul 26;12(15):4896. [doi: 10.3390/jcm12154896] [Medline: 37568298]
19. Matamala-Gomez M, Donegan T, Bottiroli S, Sandrini G, Sanchez-Vives MV, Tassorelli C. Immersive virtual reality and virtual embodiment for pain relief. Front Hum Neurosci 2019 Aug 21;13:279. [doi: 10.3389/fnhum.2019.00279] [Medline: 31551731]
20. Rutkowski S, Rutkowska A, Kiper P, et al. Virtual reality rehabilitation in patients with chronic obstructive pulmonary disease: a randomized controlled trial. Int J Chron Obstruct Pulmon Dis 2020 Jan 13;15:117-124. [doi: 10.2147/COPD.S223592] [Medline: 32021150]
21. García-Bravo S, Cuesta-Gómez A, Campuzano-Ruiz R, et al. Virtual reality and video games in cardiac rehabilitation programs. A systematic review. Disabil Rehabil 2021 Feb;43(4):448-457. [doi: 10.1080/09638288.2019.1631892] [Medline: 31258015]
22. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. J Clin Epidemiol 2021 Jun;134:178-189. [doi: 10.1016/j.jclinepi.2021.03.001] [Medline: 33789819]
23. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
24. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
25. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct 12;355:i4919. [doi: 10.1136/bmj.i4919] [Medline: 27733354]
26. Hammer Y, Shaul AA, Ben-Avraham B, et al. Exergaming in patients with a left ventricular assist device: a feasibility study. ESC Heart Fail 2023 Feb;10(1):738-741. [doi: 10.1002/ehf2.14203] [Medline: 36251532]
27. Andreassen M, Santaularia N, Berglund A, et al. Feasibility of a mobile exergame for patients with heart failure. Games Health J 2025 Oct;14(5):380-388. [doi: 10.1089/g4h.2024.0093] [Medline: 40201979]
28. Costa AS, Barbosa CB, Guizilini S, et al. Virtual reality and physical activity in patients with heart failure: technology validation and user satisfaction – pilot study. Int J Cardiovasc Sci 2025 Feb 19;38:e20240067. [doi: 10.36660/ijcs.20240067]
29. Radhakrishnan K, Julien C, Baranowski T, et al. Feasibility of a sensor-controlled digital game for heart failure self-management: randomized controlled trial. JMIR Serious Games 2021 Nov 8;9(4):e29044. [doi: 10.2196/29044] [Medline: 34747701]
30. Jaarsma T, Klompstra L, Ben Gal T, et al. Effects of exergaming on exercise capacity in patients with heart failure: results of an international multicentre randomized controlled trial. Eur J Heart Fail 2021 Jan;23(1):114-124. [doi: 10.1002/ejhf.1754] [Medline: 32167657]

XSL•FO
RenderX

31. Jaarsma T, Perkiö Kato N, Ben Gal T, et al. Factors associated with lack of improvement in submaximal exercise capacity of patients with heart failure. ESC Heart Fail 2021 Dec;8(6):4539-4548. [doi: 10.1002/ehf2.13584] [Medline: 34480412]
32. Klompstra L, Hägglund E, Jaarsma T, Kato NP, Strömberg A. Effects of exergaming and yoga on exercise capacity and physical and mental health in heart failure patients: a randomized sub-study. Eur J Cardiovasc Nurs 2025 Apr 11;24(3):389-398. [doi: 10.1093/eurjcn/zvae155] [Medline: 39743240]
33. Klompstra L, Jaarsma T, Piepoli MF, et al. Objectively measured physical activity in patients with heart failure: a sub-analysis from the HF-Wii study. Eur J Cardiovasc Nurs 2022 Jun 30;21(5):499-508. [doi: 10.1093/eurjcn/zvab133] [Medline: 34993536]
34. Lăcraru AE, Busnatu S, Pană MA, et al. Assessing the efficacy of a virtual assistant in the remote cardiac rehabilitation of heart failure and ischemic heart disease patients: case-control study of Romanian adult patients. Int J Environ Res Public Health 2023 Feb 22;20(5):3937. [doi: 10.3390/ijerph20053937] [Medline: 36900948]
35. Caballero LG, Fraga IB, Tremea CEM, et al. Experience of heart failure patients in mobilization with virtual reality: mixed methods study. Rev Esc Enferm USP 2025 Sep 15;59:e20250114. [doi: 10.1590/1980-220X-REEUSP-2025-0114en] [Medline: 40955916]
36. Cugusi L, Prosperini L, Mura G. Exergaming for quality of life in persons living with chronic diseases: a systematic review and meta-analysis. PM R 2021 Jul;13(7):756-780. [doi: 10.1002/pmrj.12444] [Medline: 32592238]
37. Kwon SH, Park JK, Koh YH. A systematic review and meta-analysis on the effect of virtual reality-based rehabilitation for people with Parkinson's disease. J Neuroeng Rehabil 2023 Jul 20;20(1):94. [doi: 10.1186/s12984-023-01219-3] [Medline: 37475014]
38. García-Betances RI, Jiménez-Mixco V, Arredondo MT, Cabrera-Umpiérrez MF. Using virtual reality for cognitive training of the elderly. Am J Alzheimers Dis Other Demen 2015 Feb;30(1):49-54. [doi: 10.1177/1533317514545866] [Medline: 25107931]
39. Martín-Rodríguez A, Gostian-Ropotin LA, Beltrán-Velasco AI, et al. Sporting mind: the interplay of physical activity and psychological health. Sports (Basel) 2024 Jan 22;12(1):37. [doi: 10.3390/sports12010037] [Medline: 38275986]
40. Lachowicz M, Serweta-Pawlik A, Żurek A, et al. Application of short and long-term virtual reality training for visuospatial memory development in amateur e-athletes. Virtual Real 2025;29:136. [doi: 10.1007/s10055-025-01209-x]
41. Goh FQ, Kong WKF, Wong RCC, et al. Cognitive impairment in heart failure-a review. Biology (Basel) 2022 Jan 23;11(2):179. [doi: 10.3390/biology11020179] [Medline: 35205045]

Abbreviations

6MWT: 6-minute walk test
CR: cardiac rehabilitation
HF: heart failure
MeSH: Medical Subject Headings
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for reporting literature searches
QOL: quality of life
RCT: randomized controlled trial
SWiM: Synthesis Without Meta-Analysis
VO_2 max: maximal oxygen consumption
VR: virtual reality

Edited by S Brini; submitted 27.Oct.2025; peer-reviewed by AJ Umeano, C Aladeokin, Y Zhang; revised version received 22.Feb.2026; accepted 24.Feb.2026; published 24.Mar.2026.

Please cite as:
Ahn JA, Lee JE, Kim KA
Effects of Virtual Reality–Based Interventions for Promoting Physical Activity in Patients With Heart Failure: Systematic Review
J Med Internet Res 2026;28:e86567
URL: https://www.jmir.org/2026/1/e86567
doi:10.2196/86567

© Jeong-Ah Ahn, Jung Eun Lee, Kyoung-A Kim. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any

XSL•FO
RenderX

medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Current Landscape of Remote Digital Symptom Monitoring for Patients With Lung Cancer: Scoping Review

Hao Huang[1,2], MSN; Naomi Takemura[1,2], PhD; Hammoda Abu-Odah[1,2], PhD; Shuhan Li[1,2], MSN; Ting Mao[1,2], MSN; Zengjie Ye[3], PhD; Janelle Yorke[1,2,4], PhD

[1]School of Nursing, The Hong Kong Polytechnic University, 11 Yuk Choi Road, Hung Hom, Kowloon, Hong Kong, China (Hong Kong)

[2]JC STEM Lab of Digital Oncology Care Enhancement (DOCE), The Hong Kong Jockey Club Charities Trust, Hong Kong, China (Hong Kong)

[3]School of Nursing, Guangzhou Medical University, Guangzhou, China

[4]Division of Nursing, Midwifery and Social Work, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, United Kingdom

Corresponding Author:
Janelle Yorke, PhD
School of Nursing, The Hong Kong Polytechnic University, 11 Yuk Choi Road, Hung Hom, Kowloon, Hong Kong, China (Hong Kong)

Abstract

Background: Remote digital symptom monitoring systems (rSMS) have been increasingly used in recent years to monitor symptoms, health-related quality of life, and other patient-reported outcomes in lung cancer. Previous studies have demonstrated variability in study design, types of rSMS, and outcomes used to assess benefits for patients and health care systems. However, there remains a lack of synthesized evidence pertaining to the similarities and differences among rSMS, including their theoretical underpinnings, key functional components, and reported benefits and limitations.

Objective: This review aims to identify and synthesize existing research to map the current landscape of rSMS in lung cancer, including the theoretical foundations for its development and implementation, as well as its types, applications, and outcomes.

Methods: This scoping review followed the Joanna Briggs Institute scoping review framework and adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines. A comprehensive literature search was conducted from database inception to October 16, 2025, across 7 English-language databases and 3 Chinese-language databases (CNKI, WanFang, and SinoMed). Eligible studies were peer-reviewed original research articles examining rSMS among adults with lung cancer. Data were independently screened and extracted by 2 reviewers, with discrepancies resolved by a third reviewer. Quantitative data were extracted using a standardized form and synthesized descriptively. Content analysis was performed to analyze the qualitative data.

Results: A total of 41 studies involving 11,765 patients and 85 health care providers were included. Twelve studies focused exclusively on advanced-stage lung cancer. Participants were generally middle-aged to older adults (mean ages 51 - 74 y), with male participants typically comprising 30% to 50% across studies. Most studies were conducted in the United States (n=19). We identified 32 patient-reported outcome measures that were used either as core rSMS components or as study outcomes. Four common functional modules were observed across rSMS: data collection, data analysis, response systems, and patient education. Qualitative evidence was limited; the most frequently reported benefit was the promotion of patient-centered care. Health care providers raised concerns about uncertain effectiveness and increased workload.

Conclusions: This scoping review highlights the promising role of rSMS in lung cancer care and provides a structured map of current evidence. It adds to prior literature in 3 ways. First, it summarizes how and how often theoretical frameworks are reported and applied in rSMS development and implementation. Second, it synthesizes and categorizes four common functional modules across systems. Third, it differentiates measures embedded as rSMS components from those used as evaluation outcomes. These contributions clarify current practices and methodological gaps and underscore the importance of theory-informed design, functional clarity, and stakeholder engagement in the development of patient-centered, clinically meaningful, and sustainable rSMS platforms.

Trial Registration: OSF Registries 9637t; https://osf.io/9637t/overview

(J Med Internet Res 2026;28:e83666) doi:10.2196/83666

KEYWORDS

digital health; electronic patient-reported outcome measure; ePROM; lung cancer; symptom experience; scoping review; symptom monitoring

XSL•FO
RenderX

Introduction

Lung cancer remains the leading cause of morbidity and mortality in men and women worldwide [1,2]. Patients with lung cancer experience a range of debilitating symptoms, including dyspnea, hoarseness, persistent cough, and fatigue, all of which substantially diminish their health-related quality of life (HRQL) throughout the disease trajectory [3,4]. Consequently, identifying and monitoring these symptoms is critical to establishing a foundation for effective interventions and optimizing patient outcomes.

Patient-reported outcome measures (PROMs) have been adopted extensively to monitor and evaluate symptoms [3,5]. These are standardized questionnaires completed by patients to capture self-assessments of symptoms, HRQL, mental well-being, and functional status [6]. In recent years, PROMs have been largely recognized as highly valid and reliable instruments for assessing health care outcomes from the patient's perspective [7-9]. Moreover, several guidelines have been established to facilitate the standardized use and enhance the effectiveness of electronic PROMs (ePROMs) in clinical practice [10-13].

Advancements in digital health technologies have facilitated the evolution of ePROMs, defined as tools for the remote, digital collection of patient-reported outcomes outside clinical settings [14,15]. The application of ePROMs overcomes traditional scalability limitations, including reliance on in-person data collection, manual workflows that delay reporting, and fragmented communication between patients and clinicians.

A variety of terms have been used to describe electronic symptom monitoring interventions, including, but not limited to, electronic symptom self-reporting systems [16], remote symptom monitoring systems [17], electronic patient-reported outcome–based symptom management [18], and patient symptom self-reporting systems [19]. Despite differences in terminology, these interventions share the common feature of enabling remote monitoring of patient symptoms through electronic platforms. To promote uniformity and standardization, this review adopts the term "remote digital symptom monitoring system (rSMS)" to refer to systems designed for digitally monitoring patient-reported outcomes.

The implementation of rSMS into routine cancer care for patients with lung cancer has gained attention in recent years, driven by its potential to enhance person-centered care [18,20]. However, evidence regarding its clinical impact remains inconsistent, reflecting heterogeneity in the evidence base, where benefits have been observed in certain domains but not consistently across all outcomes. While the efficacy of rSMS in alleviating symptom-related outcomes is widely reported, 1 randomized controlled trial (RCT) found no significant reductions in symptom burden among patients with lung cancer [21]. Furthermore, limitations in improving outcomes such as HRQL [22] and overall survival have also been documented among patients with lung cancer [17].

The integration of rSMS into lung cancer care has surged in recent years, yet the field remains marked by heterogeneity in system design, PROMs used, and clinical application. Studies vary widely in scope and sophistication: some prioritize basic functionalities, such as symptom data aggregation via mobile platforms to streamline communication between patients and health care providers [22,23], while other studies use advanced technologies to deliver real-time feedback to health care providers or patients, facilitating effective symptom management [17,24,25]. The most advanced systems integrate automated clinical alerts triggered by predefined symptom thresholds, potentially improving HRQL through timely interventions [26-28]. However, this diversity has resulted in fragmented implementation, with no unifying framework to categorize systems or evaluate their alignment with lung cancer–specific needs. It is important to note that theoretical frameworks play a critical role in enhancing the methodological rigor and reproducibility of rSMS intervention designs and implementations. Theory-driven interventions aimed at promoting behavior change are known to enhance intervention benefits [29-31].

To address these gaps, this scoping review focuses on the development and implementation of theory-driven rSMS interventions in lung cancer care. It aims to identify relevant studies and map how rSMS interventions have been designed, applied, and evaluated in this population, guided by the following questions: (1) "What theoretical frameworks have been applied to studies of rSMS in lung cancer?"; (2) "What are the most commonly used PROMs for rSMS in lung cancer?"; (3) "What are the key technology platforms, functionalities, and characteristics of rSMS?"; (4) "What are the experiences and perceptions of patients and health care providers regarding the implementation of rSMS?"; and (5) "What are the benefits and limitations of various rSMS interventions on clinical outcomes in patients with lung cancer?"

Methods

Review Design

This scoping review followed the Joanna Briggs Institute scoping review framework [32] and adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines (Checklist 1) [33]. This protocol was registered on the Open Science Framework [34].

Study Sources and Strategies

The literature search reporting adhered to the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) guidelines (Checklist 2) [35]. Comprehensive literature searches were conducted from inception to April 3, 2025, in the following English and Chinese electronic databases: (1) PubMed, (2) CINAHL (via EBSCOhost), (3) Cochrane Library, (4) Embase, (5) Scopus, (6) Web of Science, (7) PsycINFO, (8) CNKI, (9) WanFang, and (10) SinoMed. The search strategy was developed collaboratively by the research team and peer-reviewed by an experienced librarian at the institution. Given the broad scope and objectives of this scoping review, the key search terms selected were "lung cancer," "symptom monitoring," "electronic patient-reported outcome," and "remote monitoring." Of note, no limits were applied for date, time period, search filters, or

XSL•FO
RenderX

study design to maximize comprehensiveness. Both MeSH terms and free-text searches were used to ensure a thorough search. Additionally, the reference lists of included studies were reviewed to ensure no relevant evidence was overlooked [33]. If any included articles were found to contain potentially relevant information not provided in the original publication, the research team planned to obtain it by contacting the corresponding authors. The detailed search strategy used in each database is presented in Multimedia Appendix 1. The literature search was rerun on October 16, 2025, to update results.

Eligibility Criteria

This study used the PCC (population, concept, and context) strategy to establish eligibility criteria in line with the Joanna Briggs Institute recommendations [32].

Population

Studies involving adults aged 18 years or older, diagnosed with any type of lung cancer (primary or secondary) across all disease stages (stages I-IV), and receiving any treatment modality, regardless of therapeutic intent (curative or palliative), were included. Studies involving mixed cancer groups were eligible if lung cancer–specific data—particularly regarding symptom severity—were extractable.

Concept

Studies focusing on the use of rSMS in lung cancer, where health-related outcomes are reported directly by patients through digital platforms.

Context

Studies examining the use of rSMS in patients with lung cancer, remotely from home or community-based settings.

Types of Studies

This review included peer-reviewed primary studies (quantitative or qualitative) published in English or Chinese. Conference abstracts, reviews, comments, letters, and case reports were excluded, as these types of studies typically provide insufficient methodological detail or lack the ability to extract outcomes relevant to this review. Studies involving populations with mixed cancers, where lung cancer–specific data could not be isolated, were also excluded.

Study Screening and Selection

All search results were exported to EndNote 2025 (Clarivate Analytics), where duplicates were removed by 2 reviewers independently (HH and TM). Titles and abstracts were thoroughly screened, followed by a full-text review to determine final inclusion or exclusion by 2 reviewers independently (HH and TM). Disagreements were discussed and resolved through consultation with 2 independent members of the team (NT and HA-O). The selection process is illustrated in a PRISMA-ScR flowchart [33].

Data Extraction

Data relevant to the research questions were systematically extracted using a predesigned table developed by the research team. Specifically, the following details were extracted from each included study: (1) study characteristics (author, publication year, country, setting, sample size, and study design); (2) theoretical frameworks or models applied; (3) technology platforms for rSMS delivery and their key features (eg, accessibility, integration, and user interface); (4) clinical workflows for rSMS implementation (eg, with real-time clinician alerts and intervention, and automated responses to patients); (5) type of PROMs used and their symptom or health outcome domains; and (6) reported benefits and limitations of rSMS from patients, health care providers, and system perspectives.

For qualitative studies, thematic data on patients' and health care providers' experiences (eg, barriers, facilitators, and satisfaction) were also extracted.

Data Synthesis

Quantitative data, including study characteristics, the use of PROMs, and reported benefits and barriers, were summarized in tabular and narrative formats, with a narrative interpretation provided. Qualitative data, such as user experiences, were analyzed using content analysis and synthesized narratively [36].

Results

Study Selection

A total of 495 articles were retrieved through the initial search, comprising 491 from English-language databases and 4 from Chinese-language databases. After screening titles and abstracts, followed by a thorough full-text review, 80 English-language articles were deemed eligible, while all Chinese-language articles were excluded due to irrelevant content or study designs that did not meet the inclusion criteria. Additionally, 3 studies were identified through citation searching. Ultimately, 41 studies were included in the final analysis. The search process is depicted in a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart (Figure 1; Checklist 3).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. ePRO: electronic patient-reported outcome.

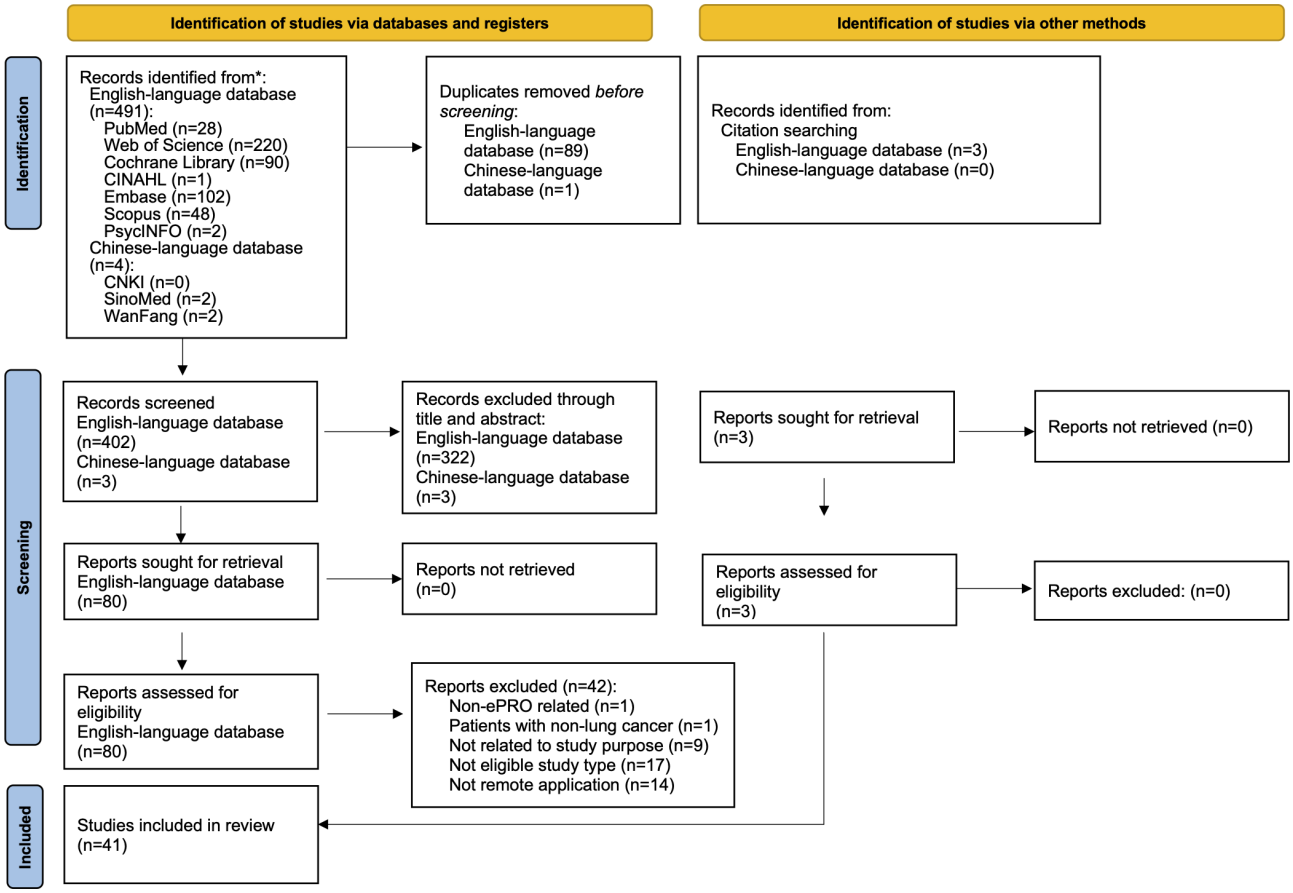

Study Characteristics

The 41 studies were published between 2007 and 2025. The majority of studies were conducted in the United States (n=19, 46%), followed by China (n=7, 17%), Denmark (n=3, 7%), the United Kingdom (n=3, 7%), and 1 (2%) study each from Australia, Germany, Spain, South Korea, Finland, France, Singapore, Italy, and Switzerland. The characteristics of the included studies are presented in Multimedia Appendix 2.

The majority of the included studies (n=19, 46%) used an RCT design, with 17 (41%) representing original RCTs [17,18,20-23,26,37-46] and 2 (5%) constituting secondary analyses derived from previously published RCTs [47,48]. The remaining publications encompassed prospective studies (n=8, 20%) [19,27,49-54], cohort studies (n=4, 10%) [55-57], mixed methods studies (n=4, 10%) [24,58-60], nonrandomized feasibility studies (n=4, 10%) [61-64], and qualitative studies (n=2, 5%) [65,66].

Among all the included studies, 5 (12%) enrolled health care providers as participants [49,50,58,60,63], whereas all others exclusively enrolled patients with lung cancer. Twelve (29%) studies focused exclusively on advanced-stage lung cancer [17,21,28,37,38,49,52,53,58,61,67], 7 (17%) included patients across all lung cancer stages [44,50,51,57,63,65,68], 3 (7%) targeted stages I to III [18,39,46], 1 (2%) concentrated on stages I to II [20], 1 (2%) focused on stage III [64], and 1 (2%) addressed stages II to IV [42]. Fifteen (37%) studies did not specify lung cancer stages [19,22-24,26,27,40,41,43,45,54-56,59,62].

Symptom monitoring across the included studies predominantly targeted respiratory complaints (cough, dyspnea, hemoptysis, voice changes, and chest pain), constitutional symptoms (fatigue, appetite loss, weight change, sleep disturbance, and cognitive difficulties), and gastrointestinal toxicities (nausea, vomiting, constipation, diarrhea, dysphagia, and sore mouth), as well as pain [17,18,20,21,39,42,46,51,55,61,62,64]. Psychological distress (anxiety, depression, and sadness) was less frequently assessed [21,42,45,46,55,64,67]. In addition, several research teams incorporated treatment-related adverse effects, such as side-effect bother, facial swelling, tumor growth sensation, and fever [17,42,45,61].

Application of Theoretical Framework

Three studies noted the application of established theoretical frameworks to guide the rSMS development or intervention implementation (Table 1).

XSL•FO
RenderX

Table . Application of theoretical frameworks or guidelines in the development and implementation of a remote digital symptom monitoring system (rSMS) for lung cancer care.

Author (Year)	Framework or guideline	Purpose	Application
Maguire et al (2015) [59]	Holistic framework to improve the uptake and impact of eHealth technologies [69]	Development of the rSMS	Prospective
Girgis et al (2022) [24]	RE-AIM[a] framework [70]	Guide the implementation	Retrospective
Mooney et al (2024) [26]	Chronic care model [71]	Development of the rSMS	Prospective

[a]RE-AIM: reach, effectiveness, adoption, implementation, and maintenance.

Development of the rSMS

The application of theoretical foundations varied across studies. Maguire et al [59] applied the holistic framework to improve the uptake and impact of eHealth technologies and to guide the development of their rSMS [69]. This framework includes 6 key foundational components, namely stakeholder engagement, iterative evaluation, intertwined development and implementation, transformation of health care organizations, persuasive design techniques, and advanced methods for impact assessment. The chronic care model [71], a patient-centered framework that emphasizes proactive, planned, and systematic interventions, was prospectively applied to design the rSMS intervention and evaluation strategies by Mooney et al [26].

Implementation of rSMS

Girgis et al [24] retrospectively applied the RE-AIM (reach, effectiveness, adoption, implementation, and maintenance) implementation science framework to guide and summarize successful intervention implementation, while concurrently identifying implementation barriers within this framework [70].

PROMs Used Within the rSMS and/or as Study Outcome Measures

A total of 32 PROMs applied across the included studies were categorized into two groups: (1) PROMs integrated into rSMS for monitoring purposes and (2) outcome measures used to evaluate the effectiveness of the rSMS intervention (Multimedia Appendix 3). Specifically, 19 PROMs were integrated within rSMS as components to monitor 3 domains—symptom assessment, HRQL, and psychological well-being. Thirteen PROMs were used as outcome measures to assess the effectiveness of the rSMS (Figure 2). Among all the 32 PROMs, three were used as both outcome measures and as components within the rSMS, including the Edmonton Symptom Assessment Scale, MD Anderson Symptom Inventory for Lung Cancer, and EuroQoL 5D-5L. Consistent with this review's focus on rSMS development and implementation within lung cancer care, detailed analysis is restricted to PROMs integrated within the rSMS, while those functioning solely as outcome measures are comprehensively cataloged in Multimedia Appendix 4.

XSL•FO
RenderX

Figure 2. Frequency of patient-reported outcome measures applied as components in remote digital symptom monitoring systems (rSMS) and as outcome measures in studies on lung cancer symptom monitoring. CTCAE: Common Terminology Criteria for Adverse Events; EORTC-QLQ-C30: European Organisation for Research and Treatment of Cancer Core Quality of Life Questionnaire; ESAS: Edmonton Symptom Assessment Scale; ESRA-C: Electronic Self Report Assessment for Cancer; FACIT-TS-S: Functional Assessment of Chronic Illness Therapy-Treatment Satisfaction-Patient Satisfaction; FACT-G: Functional Assessment of Cancer Therapy–General; FACT-L: Functional Assessment of Cancer Therapy–Lung; FLSI: Functional Assessment of Cancer Therapy–Lung Symptom Index; HADS: Hospital Anxiety and Depression Scale; MDASI-LC: MD Anderson Symptom Inventory for Lung Cancer; MSAS-SF: Memorial Symptom Assessment Scale–Short Form; PRO-CTCAE: Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events; PRO-CTCAE-k: Korean version of the Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events; PROMIS: Patient-Reported Outcomes Measurement Information System; PRSM: Patient-Reported Symptom Monitoring; QLQ-LC13: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire Lung Cancer Module; RSC: Rotterdam Symptom Checklist–Activity Subscale; SDS-15: Symptom Distress Scale–15; SIQOL: Single Item Quality of Life scale; STAI-Y: State-Trait Anxiety Inventory–Form Y; SUPPH-29: Strategies Used by Patients to Promote Health–29.

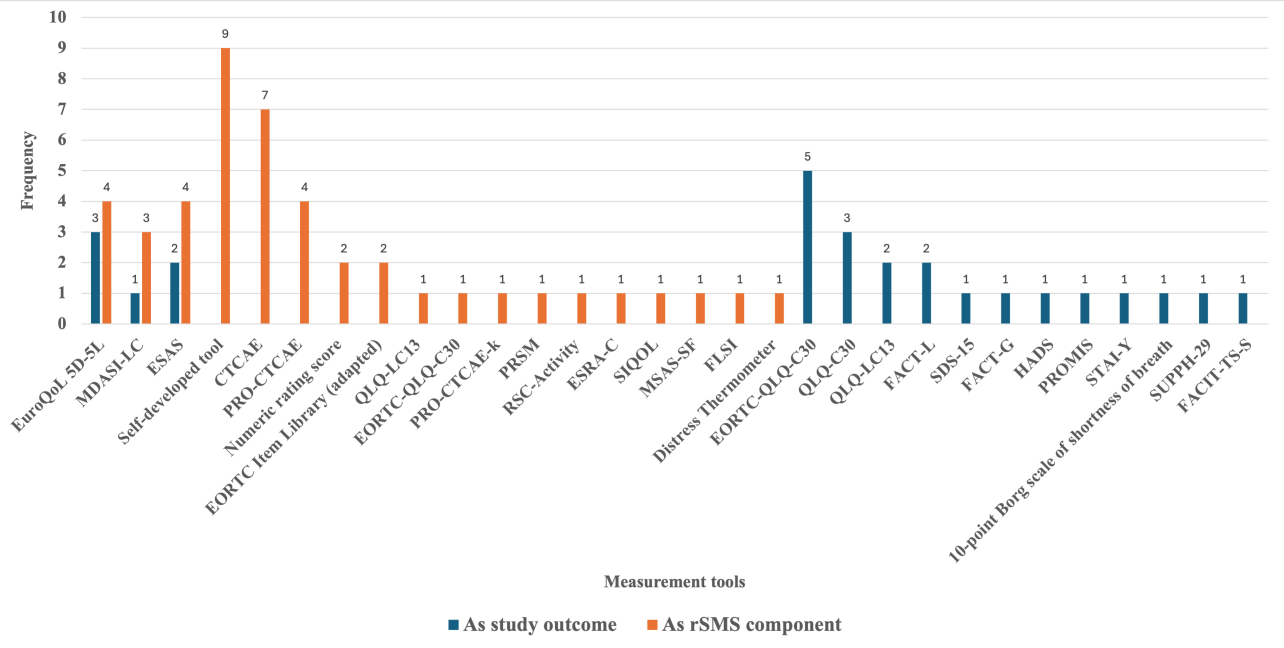

Applied Technology Platforms and the Functional Modules: Technology Platforms

The included studies used a variety of platforms for rSMS, including web-based systems, mobile phones, and wearable devices. Most studies (n=21, 51%) used web-based systems for rSMS implementation [9,17,19,22,24,27,28,43,44,51,52,54,56-58,61,62,65-67]. Mobile phones were the second most used digital platform (n=14, 34%), facilitating real-time responses and notifications between health care providers and patients [20,21,23,26,37-41,45,46,55,59,60]. Six (15%) additional studies implemented rSMS through personal electronic devices—encompassing personal computers, smartphones, or other specialized devices [18,49,50,53,63,64].

Functional Modules of the rSMS

The functional modules of rSMS were categorized into 4 key stages: data collection, data analysis, care response, and patient support (Figure 3).

XSL•FO
RenderX

Figure 3. Functional modules of the remote digital symptom monitoring system. PRO: patient-reported outcome.

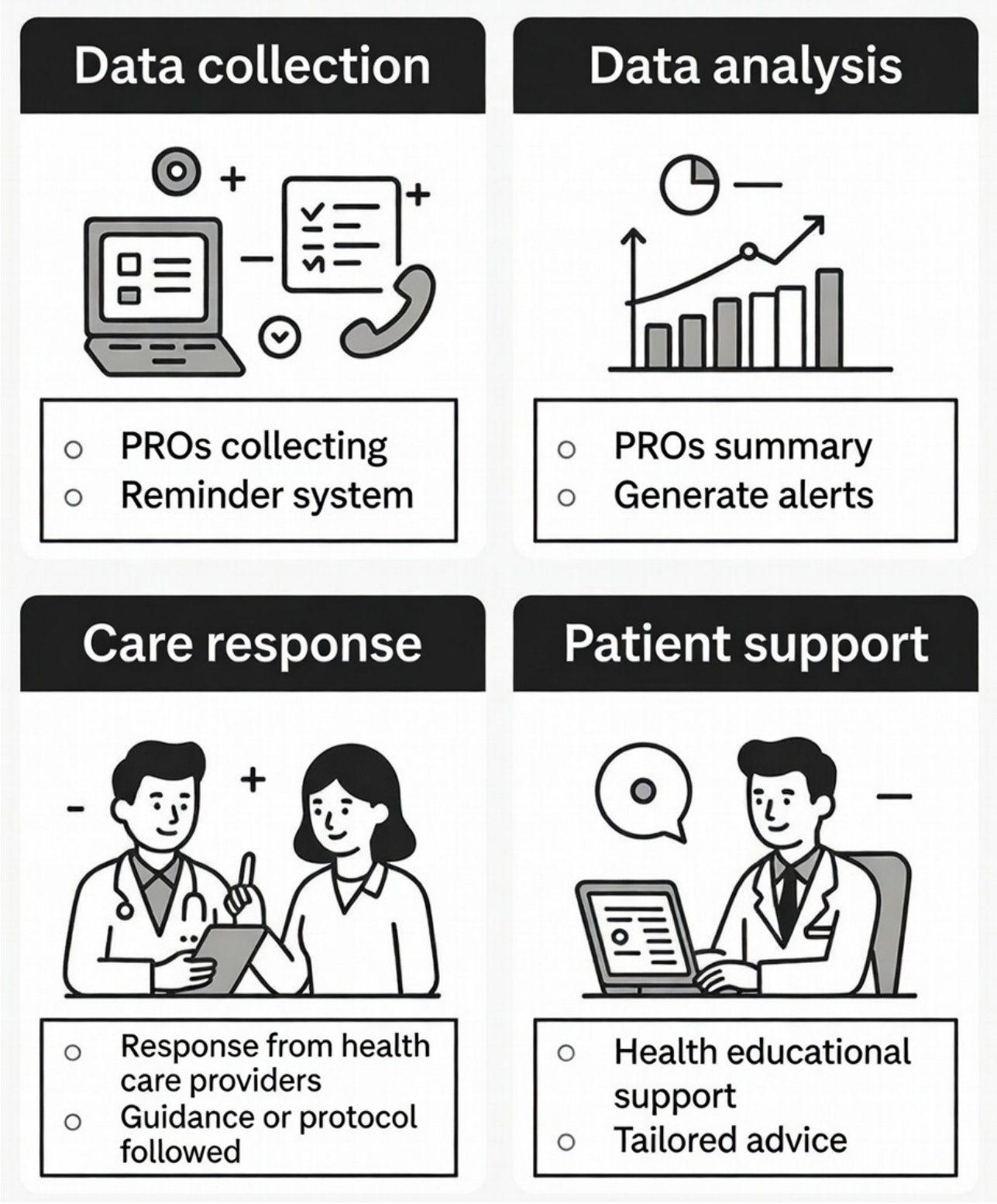

Data Collection

This stage involved symptom assessment tools to collect data, including symptom burden, symptom severity, HRQL, and other patient-reported outcomes. Additionally, 9 (22%) studies incorporated reminder systems to prompt timely patient assessments via emails, phone calls, or mobile app notifications [24,43,51,53,55,61,63,65,66].

Data Analysis

Following data collection, the data analysis stage was reported in all 41 studies. Of these, 10 (24%) studies reported systems generating summaries [22,23,38,40,42-44,60,62,72], which were provided to patients and/or health care providers, while others presented raw data or synthesized data manually. Additionally, 29 (71%) studies included notification or alert systems to inform the health care team when scores reached

XSL•FO
RenderX

predefined thresholds [17-20,24,27,28,39-42,45,46,49,50,52-55,57-61,63-67]. Typically, this module automatically conducted data analysis and generated alerts when predefined clinical criteria were met, which were then transmitted to the clinical teams—a mechanism that has been effectively implemented in current practice [17,42,49].

Care Response

The clinical response stage required mandatory responses from health care providers within a predefined time window, as reported in 11 (27%) studies [18,26,28,37,39,44-46,55,57,63]. A representative implementation by Dai et al's [18] study stipulated that when the rSMS identified a patient-reported symptom score that exceeded 4 points (indicating moderate-to-severe severity), surgeons who were involved in the patient's care were obligated to respond within 24 hours in accordance with the study protocol.

Patient Support

There were 16 (39%) studies that provided additional health education support based on the varying levels of symptom burden reported by patients [20,23,24,26,27,37,40,41,44,46,49,54,58,59,63,64]. For instance, the MyChristie-MyHealth program categorizes patients using a color-coded system based on symptom severity and provides tailored support advice [50]. Asymptomatic ("green") patients received a text message indicating that no action was required, mild symptoms ("blue") triggered an automated link to self-care advice, moderate symptoms ("orange") prompted patients to seek support from their oncology team within 1 week, and patients with severe symptoms ("red") received a message recommending urgent medical advice within 24 hours.

Experiences and Perceptions of Patients and Health Care Professionals

User experiences of both patients and health care providers were evaluated through semistructured interviews across 5 (12%) studies [45,59,60,63,66]. Questionnaire-based surveys were also used to measure satisfaction levels among patients and health care providers [62,63]. Additionally, qualitative interviews were conducted to examine the feasibility of interventions and to identify facilitating and hindering factors from the patients' perspective [65]. The findings from the content analysis are detailed in Multimedia Appendix 5 and visually synthesized in Figure 4, which highlights the key benefits and concerns expressed by patients and health care professionals regarding rSMS use.

Figure 4. Perceived benefits and concerns of the remote symptom monitoring system (rSMS) application among patients and health care professionals.

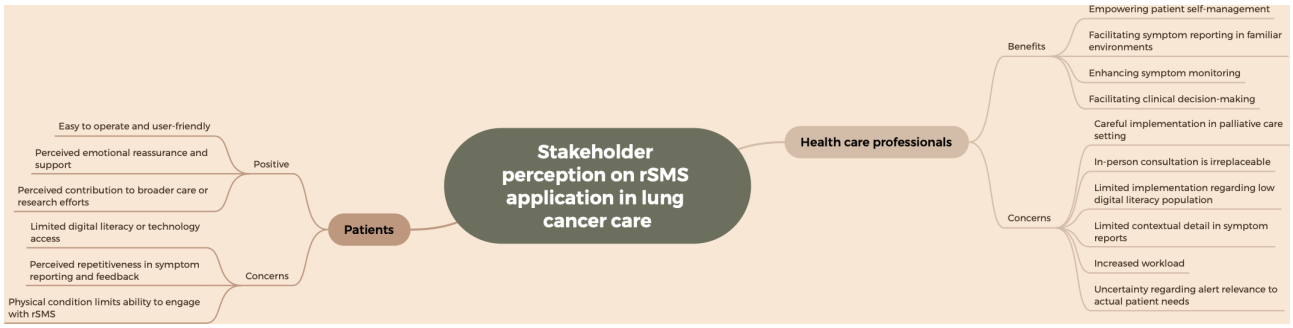

Perceptions of rSMS Applications Among Health Care Professionals

Health care providers expressed diverse benefits and barriers of rSMS. First, some health care providers in palliative care settings perceived that proactive monitoring could empower certain patients by granting them greater control over their conditions and enhancing their disease management abilities, as the system provided positive feedback that motivated patients and guided their next steps [60]. However, the same study included no patient outcome data, and most health care providers expressed doubts about implementing such interventions in palliative care settings. This aligns with the concerns voiced by clinicians in other research [63]. Additionally, the remote reporting feature allowed patients to report symptoms in familiar surroundings, potentially fostering a more relaxed reporting experience compared to hospital settings [66]. Second, for health care providers, patients' previsit self-reports enabled the care teams to prepare more effectively and facilitated early identification of symptom changes [59]. Regular monitoring of symptom fluctuations also supported disease progression tracking, aiding timely adjustments to care strategies and clinical decision-making [63,66].

Concerns about rSMS implementation were also reported. Some physicians emphasized that rSMS should supplement rather than replace clinical judgment [60,66]. Similarly, other clinicians highlighted that traditional in-person consultations and hands-on experience should not be supplanted by remote reporting, as reliance solely on patient self-reports without clinical observation poses significant challenges [60,66]. Additional concerns included the selective applicability of rSMS, particularly regarding older patients' unfamiliarity with or resistance to electronic systems, as well as its limited benefits for palliative care groups [60]. Regarding rSMS itself, some health care providers identified limitations in its ability to support rigorous clinical decisions, noting that symptom reports often lack contextual details such as frequency or type, necessitating further patient interaction [66]. Additionally, some health care providers reported that rSMS added to their existing workload, making timely responses challenging [59,66]. Furthermore, health care providers expressed concerns about alerts, questioning whether all alerts are necessary and effectively reflect patients' actual symptom burdens, while also worrying that excessive alerts might increase workloads [63].

XSL•FO
RenderX

Perceptions of rSMS Applications Among Patients

Patients reported that rSMS was generally user-friendly and easy to operate, enabling them to conveniently report their symptoms and functional status [59,65]. Timely response systems and tailored health advice were considered highly helpful [45,59,65]. Some patients felt that rSMS provided a sense of safety, encouragement, and support, making them feel attended to by health care providers outside clinical settings [45,59,65]. Additionally, patients reported feeling that they were not only receiving help during the intervention, but also contributing to helping others [45].

However, some patients expressed concerns about rSMS. System-related barriers included hardware issues (eg, unstable access in rural areas), a preference for paper-based surveys, and unfamiliarity with electronic systems [9,65]. Some perceived symptom assessments and advice as redundant with existing knowledge, reducing motivation to complete surveys [59,65]. Others questioned rSMS efficacy, describing it as "self-limited" [59,65]. High symptom burden was also cited as an engagement barrier [65].

Clinical Insights

The implementation of rSMS among patients with lung cancer has demonstrated substantial benefits, including symptom alleviation, improvements in HRQL, enhanced communication between patients and health care professionals, strengthened patient-centered care experiences, and broader health system advantages. Nevertheless, some studies have reported nonsignificant effects on clinical outcomes, underscoring the variability of evidence (Figure 5).

Figure 5. Key clinical domains impacted by remote symptom monitoring systems (rSMS).

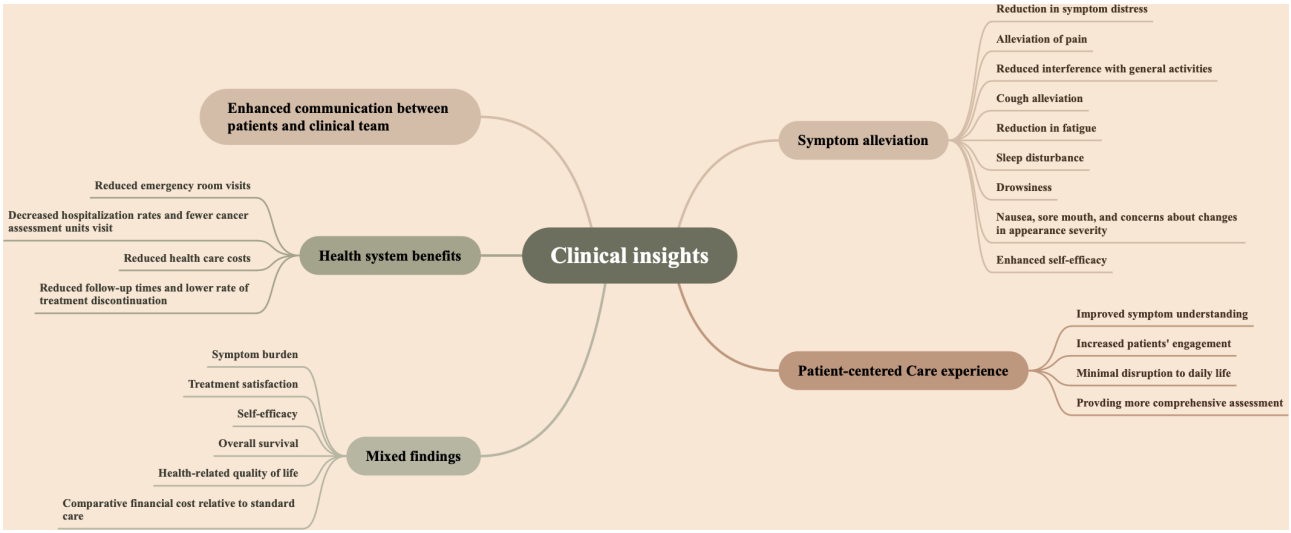

Symptom-Related Outcome

The rSMS intervention demonstrated significant reductions in overall symptoms among patients with lung cancer [18,20,22,26,39,41,46,52]. Symptom distress emerged as the most frequently documented alleviation and was evident across 4 (10%) studies [22,41,46,59]. This was followed by a significant decrease in patient-reported pain, noted in 2 RCTs [20,41]. Additionally, rSMS was reported to notably reduce interference with general activities, work, and walking, as observed in 2 RCTs [18,20]. Significant clinical improvements were reported in cough [20], fatigue [20,41], sleep disturbance [20,41], and drowsiness [59]. One (2%) study extended these findings to additional symptom categories—including nausea, sore mouth, and concerns about changes in appearance severity [41]. Beyond physical symptom management, 1 (2%) study observed enhanced patient self-efficacy, attributing these gains to rSMS-facilitated psychological distress reduction and improved adaptive functioning [59]. Conversely, another research team reported no significant improvements in symptom burden, treatment satisfaction, or self-efficacy [21]. Similarly, no significant rSMS-related benefits in overall survival were found in a separate study [17].

Health-Related Quality of Life

Five (12%) studies observed improvements in HRQL among patients using a symptom monitoring system [17,40,42,43,46]. However, 4 (10%) studies revealed no statistically significant improvements in HRQL among rSMS intervention groups [21,22,37,63].

Communication With Health Care Providers

Five (12%) studies reported that the rSMS application enhanced communication or improved discussions between patients and health care providers [22,49,50,52,57]. All these studies used questionnaire-based investigations to assess patients' perceptions. Additionally, a pooled analysis suggested that improved communication may constitute one of the potential reasons for the success of the intervention program [67]. For health care providers, rSMS primarily supported clinical workflows and optimized patient communication. This includes reducing time spent on follow-up tracking [39], enhancing work efficiency [39,40,58], and increasing confidence [24] in symptom monitoring and management.

Patient-Centered Care Experience

Several studies have demonstrated that rSMS contributes to modern patient-centered care through various approaches. Three

XSL•FO
RenderX

(7%) studies reported improved symptom understanding using a patient-reported experience measure in rSMS for lung cancer care [44,49,50]. Additionally, this intervention enhanced patient-care team interactions, increasing patients' engagement in their own care [50]. Furthermore, 2 RCTs reported that rSMS functioned effectively without disrupting daily lives, yielding high patient satisfaction in lung cancer care [19,39]. Similarly, rSMS enhanced patients' perceived benefit, making them feel their cancer care significantly improved [19,53]. Implementation also provided significant clinical benefits for specific lung cancer groups, such as patients receiving immune checkpoint inhibitors, who were reported to have a more comprehensive assessment, including toxicity grading and long-term surveillance of symptoms [40].

Health System Benefits

Reduced emergency room visits were reported in 2 (5%) studies [40,43]. Similarly, decreased hospitalization rates and fewer visits to cancer assessment units were reported in 2 (5%) other studies [56,73]. Patt et al further identified that rSMS was associated with a lower total cost of care [56]. One RCT study reported that enhanced treatment experiences and automated platforms contributed to reduced follow-up times and a lower rate of treatment discontinuation [40]. Additionally, financial costs were compared between rSMS implementation and the standard of care, with no statistically significant differences observed [45]. In the PROStep feasibility study [38], which integrated both self-reported surveys and passive activity monitoring for patients with lung and gastrointestinal cancers—whereas most other platforms in our review relied on active reporting only—the intervention group reported a high perceived understanding of symptoms and functional status by their oncology team; this did not differ from the control group [38].

Discussion

Principal Findings

This scoping review provides an overview of how rSMS has been applied in lung cancer care, with particular attention to the integration of PROMs and functional modules. The synthesis of existing studies indicates that rSMS interventions show encouraging potential to improve symptom surveillance and support timely clinical responses; however, the current evidence remains fragmented and methodologically limited. Notably, few studies have adopted theory-driven approaches in the design, implementation, or evaluation of rSMS, leaving important conceptual and practical gaps. In addition, variability in patient and health care provider engagement highlights ongoing challenges related to usability and comprehension. These findings set the stage for a more detailed discussion of the opportunities and challenges in advancing rSMS, as well as directions for future research.

Only 3 (7%) studies incorporated theory-driven approaches in designing the intervention program or evaluating the outcomes [24,26,59]. Existing evidence from implementation science underscores that theory-driven approaches enhance methodological transparency, clarify intervention mechanisms, and promote sustainability [74,75]. The current limited application of theoretical frameworks may contribute to inconsistencies in intervention implementation and hinder the comparability and validity of outcome measurements across studies. Future studies should integrate theoretical guidance at both the design and evaluation stages to maximize the effectiveness and sustainability of interventions.

To ensure the reliability of patient-reported data and the validity of study outcomes, rSMS should incorporate PROMs that have undergone rigorous psychometric validation. However, as illustrated in Figure 2, the majority of PROMs used in rSMS studies were self-developed by research teams. Of the 9 self-developed tools identified, only 4 (10%) studies explicitly reported having conducted formal validation procedures [26,42,55]. The limited reporting of validation efforts may affect the interpretability and generalizability of findings derived from these tools, underscoring the need for greater emphasis on validated PROMs in future rSMS designs.

This review further delineates four key functional modules characterizing current rSMS applications in lung cancer care, including data collection, data analysis, the response system, and patient support. While most platforms offer real-time alert capabilities, more than half report limited clinical action in response to these alerts. This lack of action is primarily attributed to barriers such as competing clinical demands and unclear delineation of responsibilities [66]. This highlights a critical need for integration strategies that link real-time symptom alerts to actionable care pathways and effective feedback mechanisms, ensuring digital monitoring enables timely intervention rather than functioning as a passive data repository.

The qualitative findings identified in this review remain limited, with reported outcomes demonstrating variability [45,59,60,63,65,66]. While health care providers value rSMS for symptom monitoring, they noted limitations, including reduced effectiveness among patients with limited digital literacy—particularly older patients with lung cancer—and perceptions of data subjectivity, possibly due to misunderstandings about the system [60,66]. Patients acknowledge rSMS benefits but express notable concerns regarding the rSMS system's trustworthiness [65]. A deeper understanding of these perceptions and experiences is needed to inform strategies for patient engagement and health care provider integration. Importantly, current evidence predominantly originates from Western high-income settings with advanced medical infrastructure, highlighting the need for research in diverse resource environments to comprehensively understand rSMS implementation.

Clinical outcomes associated with rSMS interventions were found to be heterogeneous. Several trials reported meaningful benefits, including reductions in symptom distress and improvements in pain, fatigue, or sleep disturbance [18,20,22,26,39,41,46,52]. In contrast, other studies observed limited or no impact on broader end points such as global quality of life, overall survival, or treatment satisfaction [21,22,37,63]. These discrepancies may be explained by heterogeneity in rSMS design and implementation, including differences in clinical context, monitoring frequency, patient reporting requirements,

XSL•FO
RenderX

and the extent of clinician engagement. Context-related factors may also contribute to this variability, with effectiveness potentially differing across specific care settings. For example, prior work has highlighted that the use of rSMS in palliative care requires particular caution to avoid imposing additional burdens on patients receiving end-of-life care [60].

Future Directions

In the era of precision medicine, rSMS integration can be optimized through person-centered interventions tailored to the diverse disease trajectories of patients with lung cancer. Such a tailored approach could enhance symptom assessment and management efficacy. For example, identifying variations in symptom burden can enable the integration of personalized education and technical support into the intervention program. The color-coded system identified in this review exemplifies a tailored tool for assessing these variations, thereby optimizing the application of rSMS [50].

Improving patients' understanding of rSMS implementation could facilitate its successful adoption. While the current exploratory evidence is limited, studies indicate that patients frequently hesitate to seek medical assistance despite receiving alerts [49,50]. The observed reluctance to seek medical help may reflect a limited comprehension of the intervention's objectives and functionality. As patients' understanding of rSMS can influence their adherence and engagement, future studies should prioritize preintervention education on rSMS features and clinical aims to reinforce engagement.

Limitations

This review has several limitations. First, the included studies were predominantly published in English. Although Chinese literature underwent initial screening, no studies met the inclusion criteria. Second, consistent with the guidelines of scoping review methodology, no formal quality appraisal of the included studies was performed, and therefore, the methodological quality of the evidence could not be ascertained. Third, due to the considerable heterogeneity in study designs, interventions, and outcome measures, it was challenging to draw definitive conclusions regarding clinical effectiveness.

Conclusions

This review advances the field by offering a comprehensive synthesis of the landscape of rSMS in lung cancer care. Unlike previous reviews that focused primarily on clinical outcomes or technical feasibility, this review adds to prior work by summarizing the theoretical foundations guiding rSMS, integrating functional modules, and distinguishing PROMs used as embedded system components from those reported as outcome measures. Together, these contributions clarify the current landscape and highlight methodological gaps. This is innovative because it provides a theoretical lens for comparing rSMS, not only their outcomes. To ensure long-term sustainability and clinical relevance, future research should adopt theory-driven approaches, refine functional integration, and incorporate multistakeholder perspectives to develop rSMS platforms that are patient-centered, sustainable, and adaptable across diverse clinical contexts while minimizing clinician burden.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT (Generative AI Delegation Taxonomy; 2025), the following tasks were delegated to GAI tools under full human supervision: visualization. The GAI tool used was Grok. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

The research work described in this paper was conducted in the Jockey Club Science, Technology, Engineering, and Mathematics Lab of Digital Oncology Care Enhancement funded by the Hong Kong Jockey Club Charities Trust. The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

All data generated or analyzed during this study are included in this published article (and the multimedia appendices).

Authors' Contributions

HH drafted the manuscript. NT and HA-O provided suggestions and revisions regarding the logic, presentation, and data extraction. SL and TM assisted with literature screening and data extraction. JY contributed to the conceptualization, revision, and proofreading of the draft.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of search strategy used across all databases.
[DOCX File, 26 KB - jmir_v28i1e83666_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Characteristics of included studies.
[DOCX File, 136 KB - jmir_v28i1e83666_app2.docx]

Multimedia Appendix 3
Patient-reported outcome measures serve as both components and outcome measures within remote digital symptom monitoring system.
[DOCX File, 93 KB - jmir_v28i1e83666_app3.docx]

Multimedia Appendix 4
Patient-reported outcome measures serve as outcome indicators.
[PNG File, 298 KB - jmir_v28i1e83666_app4.png]

Multimedia Appendix 5
Content analysis of qualitative studies.
[DOCX File, 43 KB - jmir_v28i1e83666_app5.docx]

Checklist 1
PRISMA-ScR checklist
[PDF File, 86 KB - jmir_v28i1e83666_app6.pdf]

Checklist 2
PRISMA-S checklist
[PDF File, 85 KB - jmir_v28i1e83666_app7.pdf]

Checklist 3
PRISMA checklist
[PDF File, 312 KB - jmir_v28i1e83666_app8.pdf]

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA A Cancer J Clinicians 2024 May;74(3):229-263. [doi: 10.3322/caac.21834]
2. Au PCM, Lee AWM, Lee VHF, Wong ICK, Hui RYM, Cheung CL. The trends in lung cancer prevalence, incidence, and survival in Hong Kong over the past two decades (2002–2021): a population-based study. Lancet Reg Health West Pac 2024 Apr;45:101030. [doi: 10.1016/j.lanwpc.2024.101030]
3. Mao S, Liu L, Miao C, et al. Electronic symptom monitoring for home-based palliative care: a systematic review. Palliat Med 2024 Sep;38(8):790-805. [doi: 10.1177/02692163241257578] [Medline: 38835179]
4. Lizán L, Pérez-Carbonell L, Comellas M. Additional value of patient-reported symptom monitoring in cancer care: a systematic review of the literature. Cancers (Basel) 2021 Sep 15;13(18):4615. [doi: 10.3390/cancers13184615] [Medline: 34572842]
5. van den Hurk CJG, Mols F, Eicher M, et al. A narrative review on the collection and use of electronic patient-reported outcomes in cancer survivorship care with emphasis on symptom monitoring. Curr Oncol 2022 Jun 17;29(6):4370-4385. [doi: 10.3390/curroncol29060349] [Medline: 35735458]
6. U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. Health Qual Life Outcomes 2006 Dec;4(1):79. [doi: 10.1186/1477-7525-4-79]
7. Antunes B, Harding R, Higginson IJ, EUROIMPACT. Implementing patient-reported outcome measures in palliative care clinical practice: a systematic review of facilitators and barriers. Palliat Med 2014 Feb;28(2):158-175. [doi: 10.1177/0269216313491619] [Medline: 23801463]
8. Churruca K, Pomare C, Ellis LA, et al. Patient-reported outcome measures (PROMs): a review of generic and condition-specific measures and a discussion of trends and issues. Health Expect 2021 Aug;24(4):1015-1024. [doi: 10.1111/hex.13254] [Medline: 33949755]
9. Brunelli C, Alfieri S, Zito E, et al. Patient voices: multimethod study on the feasibility of implementing electronic patient-reported outcome measures in a comprehensive cancer center. JMIR Cancer 2025 Jan 22;11:e56625. [doi: 10.2196/56625] [Medline: 39842002]

XSL•FO
RenderX

10. Ivatury SJ, Hazard-Jenkins HW, Brooks GA, McCleary NJ, Wong SL, Schrag D. Translation of patient-reported outcomes in oncology clinical trials to everyday practice. Ann Surg Oncol 2020 Jan;27(1):65-72. [doi: 10.1245/s10434-019-07749-2] [Medline: 31452053]
11. Lavallee DC, LeRouge C, Austin EJ, et al. Guidelines for integrating electronic patient-reported outcomes into health systems—final report. : Agency for Healthcare Research and Quality; 2021 URL: https://digital.ahrq.gov/sites/default/files/docs/citation/r01hs023785-lavallee-final-report-2021.pdf [accessed 2026-03-05]
12. Principles for selecting, developing, modifying, and adapting patient-reported outcome instruments for use in medical device evaluation: guidance for industry and food and drug administration staff, and other stakeholders. : US Food and Drug Administration; 2022 URL: https://www.fda.gov/media/141565/download [accessed 2026-03-05]
13. Kaneyasu T, Hoshino E, Naito M, et al. How to select and understand guidelines for patient-reported outcomes: a scoping review of existing guidance. BMC Health Serv Res 2024 Mar 6;24:334. [doi: 10.1186/s12913-024-10707-8]
14. García Abejas A, Serra Trullás A, Sobral MA, Canelas D, Leite Costa F, Salvador Verges À. Improving the understanding and managing of the quality of life of patients with lung cancer with electronic patient-reported outcome measures: scoping review. J Med Internet Res 2023 May 3;25(1):e46259. [doi: 10.2196/46259] [Medline: 37021695]
15. Salmani H, Nasiri S, Alemrajabi M, Ahmadi M. Advancing patient-centered cancer care: a systematic review of electronic patient-reported outcome measures. Front Rehabil Sci 2024;5:1427712. [doi: 10.3389/fresc.2024.1427712] [Medline: 39386354]
16. Cho Y, Zhang H, Harris MR, Gong Y, Smith EL, Jiang Y. Acceptance and use of home-based electronic symptom self-reporting systems in patients with cancer: systematic review. J Med Internet Res 2021 Mar 12;23(3):e24638. [doi: 10.2196/24638] [Medline: 33709929]
17. Friis RB, Pappot H, Hjollund NH, et al. Remote symptom monitoring of patients with advanced lung cancer (the ProWide study): a randomized controlled trial. JCO Oncol Pract 2025 Jun;21(6):813-823. [doi: 10.1200/OP-24-00562] [Medline: 39657078]
18. Dai W, Wang Y, Liao J, et al. Electronic patient-reported outcome-based symptom management versus usual care after lung cancer surgery: long-term results of a multicenter, randomized, controlled trial. J Clin Oncol 2024;42:2126-2131. [doi: 10.1200/JCO.23.01854]
19. Ma TM, Yang T, Philipson R, Kishan AU, Lee P, Raldow AC. Web-based symptom monitoring with patient-reported outcomes during definitive radiation therapy with chemotherapy (SYMPATHY): a prospective single-center phase 1 study. Adv Radiat Oncol 2023;8(3):101073. [doi: 10.1016/j.adro.2022.101073] [Medline: 37273926]
20. Lv C, Lu F, Zhou X, et al. Efficacy of a smartphone application assisting home-based rehabilitation and symptom management for patients with lung cancer undergoing video-assisted thoracoscopic lobectomy: a prospective, single-blinded, randomised control trial (POPPER study). Int J Surg 2025 Jan 1;111(1):597-608. [doi: 10.1097/JS9.0000000000001845] [Medline: 38905505]
21. Yount SE, Rothrock N, Bass M, et al. A randomized trial of weekly symptom telemonitoring in advanced lung cancer. J Pain Symptom Manage 2014 Jun;47(6):973-989. [doi: 10.1016/j.jpainsymman.2013.07.013] [Medline: 24210705]
22. Strasser F, Blum D, von Moos R, et al. The effect of real-time electronic monitoring of patient-reported symptoms and clinical syndromes in outpatient workflow of medical oncologists: E-MOSAIC, a multicenter cluster-randomized phase III study (SAKK 95/06). Ann Oncol 2016 Feb;27(2):324-332. [doi: 10.1093/annonc/mdv576] [Medline: 26646758]
23. Lee M, Kang D, Kang E, et al. Efficacy of the PRO-CTCAE mobile application for improving patient participation in symptom management during cancer treatment: a randomized controlled trial. Support Care Cancer 2023 May 6;31(6):321. [doi: 10.1007/s00520-023-07779-3] [Medline: 37148373]
24. Girgis A, Bamgboje-Ayodele A, Rincones O, et al. Stepping into the real world: a mixed-methods evaluation of the implementation of electronic patient reported outcomes in routine lung cancer care. J Patient Rep Outcomes 2022 Jun 20;6(1):70. [doi: 10.1186/s41687-022-00475-6] [Medline: 35723827]
25. Pietanza MC, Basch EM, Lash A, et al. Harnessing technology to improve clinical trials: study of real-time informatics to collect data, toxicities, image response assessments, and patient-reported outcomes in a phase II clinical trial. J Clin Oncol 2013 Jun 1;31(16):2004-2009. [doi: 10.1200/JCO.2012.45.8117] [Medline: 23630218]
26. Mooney K, Gullatte M, Iacob E, et al. Essential components of an electronic patient-reported symptom monitoring and management system: a randomized clinical trial. JAMA Netw Open 2024 Sep 3;7(9):e2433153. [doi: 10.1001/jamanetworkopen.2024.33153] [Medline: 39269704]
27. Wujcik D, Dudley WN, Dudley M, Gupta V, Brant J. Electronic patient symptom management program to support patients receiving cancer treatment at home during the COVID-19 pandemic. Value Health 2022 Jun;25(6):931-936. [doi: 10.1016/j.jval.2022.01.023] [Medline: 35339378]
28. Chih MY, DuBenske LL, Hawkins RP, et al. Communicating advanced cancer patients' symptoms via the Internet: a pooled analysis of two randomized trials examining caregiver preparedness, physical burden, and negative mood. Palliat Med 2013 Jun;27(6):533-543. [doi: 10.1177/0269216312457213] [Medline: 22988042]
29. Rothman AJ. Capitalizing on opportunities to refine health behavior theories. Health Educ Behav 2009 Oct;36(5 Suppl):150S-155S. [doi: 10.1177/1090198109340514] [Medline: 19830884]

XSL•FO
RenderX

30. Michie S, Johnston M, Francis J, Hardeman W, Eccles M. From theory to intervention: mapping theoretically derived behavioural determinants to behaviour change techniques. Appl Psychol 2008 Oct;57(4):660-680. [doi: 10.1111/j.1464-0597.2008.00341.x]
31. Davis R, Campbell R, Hildon Z, Hobbs L, Michie S. Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. Health Psychol Rev 2015;9(3):323-344. [doi: 10.1080/17437199.2014.941722] [Medline: 25104107]
32. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020 Oct;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
33. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
34. Current digital symptom monitoring implementation among lung cancer patients: a scoping review. Open Science Framework. URL: https://osf.io/9637t/overview [accessed 2026-03-13]
35. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
36. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
37. Yang GM, Ke Y, Ng XH, Neo PSH, Cheung YB. Proactive symptom monitoring to initiate timely palliative care for patients with advanced cancer: a randomized controlled trial. Support Care Cancer 2025 Mar 5;33(3):249. [doi: 10.1007/s00520-025-09311-1] [Medline: 40038129]
38. Parikh RB, Schriver E, Ferrell WJ, et al. Remote patient-reported outcomes and activity monitoring to improve patient-clinician communication regarding symptoms and functional status: a randomized controlled trial. JCO Oncol Pract 2023 Dec;19(12):1143-1151. [doi: 10.1200/OP.23.00048] [Medline: 37816198]
39. Dai W, Feng W, Zhang Y, et al. Patient-reported outcome-based symptom management versus usual care after lung cancer surgery: a multicenter randomized controlled trial. J Clin Oncol 2022 Mar 20;40(9):988-996. [doi: 10.1200/JCO.21.01344]
40. Zhang L, Zhang X, Shen L, Zhu D, Ma S, Cong L. Efficiency of electronic health record assessment of patient-reported outcomes after cancer immunotherapy: a randomized clinical trial. JAMA Netw Open 2022 Mar 1;5(3):e224427. [doi: 10.1001/jamanetworkopen.2022.4427] [Medline: 35357459]
41. Mooney KH, Beck SL, Wong B, et al. Automated home monitoring and management of patient-reported symptoms during chemotherapy: results of the symptom care at home RCT. Cancer Med 2017 Mar;6(3):537-546. [doi: 10.1002/cam4.1002] [Medline: 28135050]
42. Denis F, Lethrosne C, Pourel N, et al. Randomized trial comparing a web-mediated follow-up with routine surveillance in lung cancer patients. J Natl Cancer Inst 2017 Sep 1;109(9). [doi: 10.1093/jnci/djx029] [Medline: 28423407]
43. Basch E, Deal AM, Kris MG, et al. Symptom monitoring with patient-reported outcomes during routine cancer treatment: a randomized controlled trial. J Clin Oncol 2016 Feb 20;34(6):557-565. [doi: 10.1200/JCO.2015.63.0830]
44. Berry DL, Hong F, Halpenny B, et al. The electronic self report assessment and intervention for cancer: promoting patient verbal reporting of symptom and quality of life issues in a randomized controlled trial. BMC Cancer 2014 Jul 12;14(1):513. [doi: 10.1186/1471-2407-14-513] [Medline: 25014995]
45. Pongiglione B, Cucciniello M, Petracca F, et al. A mobile supportive care app for patients with metastatic lung cancer: the Lung Cancer App (LuCApp) randomized controlled trial. Support Care Cancer 2025 Jun 30;33(7):641. [doi: 10.1007/s00520-025-09682-5] [Medline: 40586961]
46. Jing P, Liang Y, Tan Z, et al. Physiological and psychological symptom management based on electronic patient-reported outcomes: the TD-WELLBEING randomized clinical trial. Br J Cancer 2025 Oct;133(7):937-944. [doi: 10.1038/s41416-025-03110-5] [Medline: 40775446]
47. Dubenske LL, Chih MY, Dinauer S, Gustafson DH, Cleary JF. Development and implementation of a clinician reporting system for advanced stage cancer: initial lessons learned. J Am Med Inform Assoc 2008;15(5):679-686. [doi: 10.1197/jamia.M2532] [Medline: 18579837]
48. DuBenske LL, Gustafson DH, Namkoong K, et al. CHESS improves cancer caregivers' burden and mood: results of an eHealth RCT. Health Psychol 2014 Oct;33(10):1261-1272. [doi: 10.1037/a0034216] [Medline: 24245838]
49. Arriola E, Jaal J, Edvardsen A, et al. Feasibility and user experience of digital patient monitoring for real-world patients with lung or breast cancer. Oncologist 2024 Apr 4;29(4):e561-e569. [doi: 10.1093/oncolo/oyad289] [Medline: 38007400]
50. Payne A, Horne A, Bayman N, et al. Patient and clinician-reported experiences of using electronic patient reported outcome measures (ePROMs) as part of routine cancer care. J Patient Rep Outcomes 2023 May 4;7(1):42. [doi: 10.1186/s41687-023-00544-4] [Medline: 37140730]
51. Cheng X, Yang Y, Shentu Y, et al. Remote monitoring of patient recovery following lung cancer surgery: a messenger application approach. J Thorac Dis 2021 Feb;13(2):1162-1171. [doi: 10.21037/jtd-21-27] [Medline: 33717589]
52. Zylla DM, Gilmore GE, Steele GL, et al. Collection of electronic patient-reported symptoms in patients with advanced cancer using Epic MyChart surveys. Support Care Cancer 2020 Jul;28(7):3153-3163. [doi: 10.1007/s00520-019-05109-0] [Medline: 31701269]

XSL•FO
RenderX

53. Iivanainen S, Alanko T, Vihinen P, et al. Follow-up of cancer patients receiving anti-PD-(L) 1 therapy using an electronic patient-reported outcomes tool (KISS): prospective feasibility cohort study. JMIR Form Res 2020 Oct 28;4(10):e17898. [doi: 10.2196/17898] [Medline: 33112242]
54. Sewell M, Boerner T, Harrington C, et al. Remote symptom monitoring in thoracic surgery patients after discharge. Ann Surg 2025 Jun 1;281(6):1063-1069. [doi: 10.1097/SLA.0000000000006619] [Medline: 39774469]
55. Yu H, Lei C, Wei X, et al. Electronic symptom monitoring after lung cancer surgery: establishing a core set of patient-reported outcomes for surgical oncology care in a longitudinal cohort study. Int J Surg 2024 Oct 1;110(10):6591-6600. [doi: 10.1097/JS9.0000000000001855] [Medline: 38896873]
56. Patt DA, Patel AM, Bhardwaj A, et al. Impact of remote symptom monitoring with electronic patient-reported outcomes on hospitalization, survival, and cost in community oncology practice: the Texas two-step study. JCO Clin Cancer Inform 2023 Sep;7(7):e2300182. [doi: 10.1200/CCI.23.00182] [Medline: 37897263]
57. Mody GN, Stover AM, Wang M, et al. Electronic patient-reported outcomes monitoring during lung cancer chemotherapy: a nested cohort within the PRO-TECT pragmatic trial (AFT-39). Lung Cancer 2021 Dec;162:1-8. [doi: 10.1016/j.lungcan.2021.09.020] [Medline: 34634754]
58. Schmalz O, Jacob C, Ammann J, et al. Digital monitoring and management of patients with advanced or metastatic non-small cell lung cancer treated with cancer immunotherapy and its impact on quality of clinical care: interview and survey study among health care professionals and patients. J Med Internet Res 2020 Dec 21;22(12):e18655. [doi: 10.2196/18655] [Medline: 33346738]
59. Maguire R, Ream E, Richardson A, et al. Development of a novel remote patient monitoring system: the advanced symptom management system for radiotherapy to improve the symptom experience of patients with lung cancer receiving radiotherapy. Cancer Nurs 2015;38(2):E37-E47. [doi: 10.1097/NCC.0000000000000150] [Medline: 24836956]
60. Cox A, Illsley M, Knibb W, et al. The acceptability of e-technology to monitor and assess patient symptoms following palliative radiotherapy for lung cancer. Palliat Med 2011 Oct;25(7):675-681. [doi: 10.1177/0269216311399489] [Medline: 21474620]
61. Friis RB, Hjollund NH, Mejdahl CT, Pappot H, Skuladottir H. Electronic symptom monitoring in patients with metastatic lung cancer: a feasibility study. BMJ Open 2020 Jun 17;10(6):e035673. [doi: 10.1136/bmjopen-2019-035673] [Medline: 32554725]
62. Basch E, Iasonos A, Barz A, et al. Long-term toxicity monitoring via electronic patient-reported outcomes in patients receiving chemotherapy. J Clin Oncol 2007 Dec 1;25(34):5374-5380. [doi: 10.1200/JCO.2007.11.2243]
63. Stover AM, Deal AM, Medley CJ, et al. Feasibility, acceptability, and utility of remote patient-reported outcomes monitoring in patients with lung cancer: a Moovcare© study. Clin Lung Cancer 2025 Dec;26(8):e660-e669.e1. [doi: 10.1016/j.cllc.2025.07.016] [Medline: 40850860]
64. Nuamek T, Kwateng PAN, Payne A, et al. Integrating electronic patient-reported outcome measures (ePROMs) into personalised follow-up for patients after radiotherapy: a feasibility study. Tech Innov Patient Support Radiat Oncol 2025 Sep;35:100333. [doi: 10.1016/j.tipsro.2025.100333]
65. Boisson-Walsh A, Cox C, O'Leary M, et al. A qualitative study of electronic patient-reported outcome symptom monitoring after thoracic surgery. J Surg Res 2024 Nov;303:744-755. [doi: 10.1016/j.jss.2024.09.051] [Medline: 39461326]
66. Schougaard LMV, Friis RB, Grytnes R, et al. Exploring the nurses' perspective on using remote electronic symptom monitoring in clinical decision-making among patients with metastatic lung cancer. Semin Oncol Nurs 2023 Oct 26;39(6):151517. [doi: 10.1016/j.soncn.2023.151517] [Medline: 39492326]
67. Gustafson DH, DuBenske LL, Atwood AK, et al. Reducing symptom distress in patients with advanced cancer using an e-alert system for caregivers: pooled analysis of two randomized clinical trials. J Med Internet Res 2017 Nov 14;19(11):e354. [doi: 10.2196/jmir.7466] [Medline: 29138131]
68. Roberts NA, Pelecanos A, Alexander K, Wyld D, Janda M. Implementation of patient-reported outcomes in a medical oncology setting (the iPROMOS Study): type II hybrid implementation study. J Med Internet Res 2024 Aug 27;26:e55841. [doi: 10.2196/55841] [Medline: 39190468]
69. van Gemert-Pijnen JEWC, Nijland N, van Limburg M, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011 Dec 5;13(4):e111. [doi: 10.2196/jmir.1672] [Medline: 22155738]
70. Glasgow RE, Harden SM, Gaglio B, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019;7:64. [doi: 10.3389/fpubh.2019.00064] [Medline: 30984733]
71. Yeoh EK, Wong MCS, Wong ELY, et al. Benefits and limitations of implementing chronic care model (CCM) in primary care programs: a systematic review. Int J Cardiol 2018 May 1;258:279-288. [doi: 10.1016/j.ijcard.2017.11.057] [Medline: 29544944]
72. Nipp RD, Horick NK, Qian CL, et al. Effect of a symptom monitoring intervention for patients hospitalized with advanced cancer: a randomized clinical trial. JAMA Oncol 2022 Apr 1;8(4):571-578. [doi: 10.1001/jamaoncol.2021.7643] [Medline: 35142814]
73. Girgis A, Durcinoska I, Arnold A, et al. Web-based patient-reported outcome measures for personalized treatment and care (PROMPT-Care): multicenter pragmatic nonrandomized trial. J Med Internet Res 2020 Oct 29;22(10):e19685. [doi: 10.2196/19685] [Medline: 33118954]

XSL•FO
RenderX

74. Ridde V, Pérez D, Robert E. Using implementation science theories and frameworks in global health. BMJ Glob Health 2020;5(4):e002269. [doi: 10.1136/bmjgh-2019-002269] [Medline: 32377405]
75. Proctor E, Luke D, Calhoun A, et al. Sustainability of evidence-based healthcare: research agenda, methodological advances, and infrastructure support. Implementation Sci 2015 Dec;10(1):88. [doi: 10.1186/s13012-015-0274-5]

Abbreviations

ePROM: electronic patient-reported outcome measure
HRQL: health-related quality of life
PCC: population, concept, and context
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Literature Search
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analysis Extension for Scoping Reviews
PROM: patient-reported outcome measure
RCT: randomized controlled trial
RE-AIM: reach, effectiveness, adoption, implementation, and maintenance
rSMS: remote digital symptom monitoring system

Edited by S Brini; submitted 06.Sep.2025; peer-reviewed by IM Rasmussen, SS Kollala; accepted 05.Jan.2026; published 24.Mar.2026.

Please cite as:
Huang H, Takemura N, Abu-Odah H, Li S, Mao T, Ye Z, Yorke J
The Current Landscape of Remote Digital Symptom Monitoring for Patients With Lung Cancer: Scoping Review
J Med Internet Res 2026;28:e83666
URL: https://www.jmir.org/2026/1/e83666
doi:10.2196/83666

© Hao Huang, Naomi Takemura, Hammoad Abu-Odah, Shuhan Li, Ting Mao, Zengjie Ye, Janelle Yorke. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Association Between Short-Form Video Use and Mental Health: Systematic Review and Meta-Analysis

Di Tang[1,2], PhD; Xin Zhang[3], PhD; Pengpeng Gou[4,5], MSc; Jie Feng[6], Prof Dr; Rui Hu[7], PhD; Kim-wai Raymond Sum[1], Prof Dr

[1]Department of Sports Science and Physical Education, Faculty of Education, Chinese University of Hong Kong, G09, Kwok Sports Building, Hong Kong, China (Hong Kong)

[2]The Nethersole School of Nursing, Faculty of Medicine, Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

[3]Faculty of Education, Central China Normal University, Wuhan, Hubei, China

[4]School of Sports Training, Wuhan Sports University, Wuhan, Hubei, China

[5]Center of Strength and Conditioning, Wuhan Sports University, Wuhan, Hubei, China

[6]Faculty of Sports Science, Ningbo University, Ningbo, Zhejiang, China

[7]School of Sports Science and Physical Education, Southwestern University of Finance and Economics, Chengdu, Sichuan, China

Corresponding Author:
Kim-wai Raymond Sum, Prof Dr
Department of Sports Science and Physical Education, Faculty of Education, Chinese University of Hong Kong, G09, Kwok Sports Building, Hong Kong, China (Hong Kong)

Abstract

Background: Short-form videos (SFVs) have emerged as a dominant trend in digital content sharing over the past decade, gaining rapid global popularity. An increasing number of studies have explored the association between SFV use and mental health, yet current empirical evidence remains inconsistent.

Objective: This study aimed to provide a comprehensive synthesis examining the relationship between SFV usage and mental health outcomes, distinguishing between problematic and routine usage behaviors.

Methods: Web of Science, PubMed, Scopus, Embase, PsycInfo, SportDiscus, and ProQuest were searched for relevant literature up to January 3, 2026. Statistical metrics reflecting the relationship between SFV use and mental health outcome indicators were extracted for meta-analysis, such as Pearson correlation coefficients, Spearman's rank correlation coefficients, and β coefficients. For intervention and longitudinal studies, we extracted the baseline correlation coefficients. The Joanna Briggs Institute Critical Appraisal Checklist was used to assess the risk of bias.

Results: A total of 58 studies, with a cumulative sample of 96,676 participants, were included in the analysis. Results showed no significant correlations between SFV use and positive psychological states, but significant positive correlations with negative psychological states, including depression (r=0.24, 95% CI 0.15-0.33), anxiety (r=0.26, 95% CI 0.17-0.35), stress (r=0.41, 95% CI 0.38-0.56), negative affect (r=0.22, 95% CI 0.10.33), loneliness (r=0.33, 95% CI 0.25-0.41), and boredom (r=0.42, 95% CI 0.29-0.53). Subgroup analyses revealed substantial differences between usage patterns: problematic use demonstrated significant negative associations with subjective well-being (r=−0.25, 95% CI −0.35 to −0.14) and positive correlations with adverse mental health outcomes, while routine usage showed no significant associations with negative affect and exhibited a negative but nonsignificant correlation with anxiety (r=−0.02, 95% CI −0.18 to 0.14). Additionally, while some studies reported significant associations between routine SFV use (time spent or frequency) and mental health outcomes, these findings were less consistent and showed smaller effect sizes compared to the more robust correlations found with problematic use measures. This suggests that the relationship between time spent on SFV and mental health may be less stable than that of problematic use patterns, highlighting the importance of considering qualitative aspects of usage rather than merely quantitative metrics.

Conclusions: This meta-analysis suggests that SFV use is associated with adverse mental health outcomes. Future research should use objective measurement instruments to capture contemporary SFV usage patterns and differentiate between distinct types of engagement. Additionally, studies with greater geographical diversity and longitudinal or experimental designs are needed to establish causality and examine temporal changes in these associations.

(J Med Internet Res 2026;28:e82503) doi:10.2196/82503

KEYWORDS

short-form video; social media; mobile phone; addiction; mental health; meta-analysis

XSL•FO
RenderX

Introduction

In the past decade, short-form videos (SFVs) have emerged as a dominant trend in digital content sharing, rapidly gaining global popularity. Platforms such as TikTok (known as Douyin in China) have transformed the way users create and consume content by enabling the production and sharing of videos lasting only a few seconds to a few minutes [1]. By 2022, TikTok had surpassed 1.2 billion monthly active users, with the average user spending 168 minutes per day on the platform [2]. In China, approximately 95% of internet users engage in SFV consumption [3]. Historically, video-sharing platforms like YouTube primarily hosted long-form content [4]. However, in response to the growing popularity of SFVs, major platforms such as YouTube, Instagram, and Snapchat have integrated similar features to meet user demand [5]. SFVs have gained attention due to their ease of use and convenience, alongside their engaging nature, becoming a defining feature of modern social media. These platforms not only facilitate passive content consumption but also actively encourage user participation and creativity, reshaping self-expression and cultural exchange in the digital age [6,7].

Compared to other forms of social media, SFV apps have quickly captured significant market share by addressing users' psychological needs, such as fostering social connections, providing emotional gratification, and fulfilling the desire for recognition [6]. One key factor driving this success is the availability of intuitive video editing tools, which empower users to easily become content creators, thereby enhancing engagement and participation [7]. For instance, Instagram reported that by 2020, more than 500 million accounts used its "Stories" feature daily, underscoring the growing appeal of SFV consumption [8]. The preference for SFVs can also be attributed to the fast-paced nature of modern life, where users prioritize convenience and brevity. Platforms offering SFV content are particularly appealing because they are user-friendly and rely on algorithm-driven personalization to deliver tailored content based on individual preferences [5]. This personalized approach not only improves user satisfaction but also boosts traffic, accelerating the widespread adoption of these platforms.

The diverse range of content available in SFVs further contributes to its widespread popularity. These videos span a variety of topics, including beauty, cooking, entertainment, education, health, and technology [5]. While traditional social media platforms often focus on lifestyle-oriented content, platforms like TikTok have differentiated themselves by emphasizing creative, unconventional, and often humorous videos. The majority of these videos are not professionally produced, which enhances their authenticity and relatability. TikTok, for example, primarily targets younger audiences with content such as music, dance, and viral comedic trends. In contrast, Douyin, its Chinese counterpart, caters to a broader demographic by featuring content that resonates with older users and focuses more heavily on everyday life [8]. This dual strategy has enabled TikTok to expand its user base and appeal to diverse cultural and demographic groups, cementing its role as a global leader in SFV content.

Despite the significant economic value brought by SFVs and their role in enriching modern entertainment, their rapid growth and potential for overuse have raised concerns among researchers about underlying health risks [6]. The convenience and entertainment value of SFVs make it easy for the public, especially younger users, to overuse them and struggle to control their viewing habits, resulting in increased time spent on these platforms. This trend exacerbates worries regarding negative health outcomes [6,9]. Current research indicates that excessive use of SFVs exhibits similarities to internet addiction or smartphone addiction, potentially leading to addictive behaviors and health issues, such as impairing social activities, interpersonal relationships, emotional stability, and psychological well-being [1,2,10,11]. Studies have shown that the algorithmic recommendation systems of SFV apps contribute to an information cocoon that increases users' feelings of loneliness, which may subsequently lead to higher levels of depression [12]. Additionally, problematic use of SFVs has been associated with suicidal ideation and self-injurious behaviors [13]. Moreover, numerous studies highlight a positive correlation between SFV usage and negative emotions such as anxiety and stress, while revealing a negative correlation with well-being, self-esteem, and sleep quality [14-17]. However, the literature is far from conclusive. Some studies have reported no significant correlation between SFV usage and specific mental health indicators, such as depression and anxiety [17-19]. For example, Banjanin et al [20] found no significant correlation between depression and time spent on social media. Chung et al [21] and Zhai et al [22] found a positive relationship between SFVs and life satisfaction, and Chao et al [23] discovered that different usage patterns may have opposed effects on mental health. This body of research underscores the complexities surrounding the relationships in this domain [17].

In the current investigations examining the relationship between SFV usage and psychological health indicators, the selected metrics for assessing SFV usage can be broadly categorized into 2 types. The first type pertains to an individual's general engagement with SFVs, encompassing several dimensions such as frequency, duration, and other indices reflecting the extent of usage [9,24-26]. While usage frequency quantifies the time an individual allocates to SFVs, it does not capture the psychological quality of that engagement. The second focuses on problematic usage, centering on addiction severity, dependency, and compulsive behavior [27-29]. Unlike simple usage intensity, problematic usage is characterized by psychological dependence, emotional distress, and functional impairment. Consequently, high-frequency use may simply reflect strong interest or leisure engagement without necessarily implying maladaptive outcomes. Therefore, treating these 2 types of metrics as interchangeable is methodologically problematic. Specifically, the association between usage behavior and psychological health may vary significantly depending on whether usage is operationalized as addiction or problematic dependency versus simply as daily time spent or frequency. For instance, research has reported that active usage of TikTok is significantly and negatively correlated with anxiety [30], whereas problematic usage of SFVs was reported to have a significant positive correlation with anxiety in another study [31]. This distinction is likely a primary driver of the

XSL•FO
RenderX

inconsistencies observed in the existing literature. Furthermore, variations in research quality and design (such as sample size, outcome selection, and measurement methodologies) may further obscure the true relationship between SFV usage and mental health.

Despite the growing body of literature on the relationship between SFV usage and health, there remains a lack of systematic synthesis of existing findings. Given the inconsistencies in prior findings, this study aims to clarify the relationship between SFV use and mental health through a systematic review and meta-analysis. Additionally, by distinguishing between different usage patterns, this research seeks to investigate potential heterogeneity in the impact of SFV usage on mental health outcomes.

Methods

The systematic review adhered to the 2020 guidelines established by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [32] (Checklist 1). In addition, the review protocol was registered in advance in the International Prospective Register of Systematic Reviews database (Registration ID: CRD420251015620).

Search Strategy

Web of Science, PubMed, Scopus, Embase, PsycInfo, SportDiscus, and ProQuest were searched to identify relevant literature from the inception of each database to January 3, 2026. The search terms combined variations of and specific health-related outcomes, such as depression, well-being, and anxiety. A detailed description of the search strategy can be found in the Multimedia Appendix 1. All references retrieved from the searches were organized using EndNote 20 (Clarivate) software.

Eligibility Criteria

The selection criteria imposed no restrictions on participants' sex or age. Regarding study design, we primarily included cross-sectional studies. However, longitudinal studies or interventional trials were also deemed eligible if they performed a cross-sectional analysis of baseline data—specifically, analyzing the relationship between SFV use and mental health outcomes before the commencement of any intervention or follow-up period—and reported the relevant correlation coefficients.

We operationally defined SFVs as video content lasting from several seconds to a few minutes [33,34], without imposing restrictions on specific platforms or content categories. However, for multifunctional video platforms not exclusively dedicated to SFVs (eg, Instagram), inclusion was restricted to studies that specifically isolated usage behaviors related to the SFV feature (eg, intensity of use or addiction to Instagram Reels), rather than general platform usage. Metrics for SFV use were categorized into regular usage indicators, including frequency, daily and weekly duration, as well as problematic use, which encompassed measures of video addiction, dependency levels, and indicators of excessive use. Regarding mental health, studies were required to assess mental health outcomes quantitatively. These constructs could include subjective measures (eg, self-reported or perceived health) or objective measures (eg, standardized tests). To ensure systematic categorization, identified health constructs were classified into mental health domains based on the World Health Organization's International Classification of Diseases [35].

Crucially, this study focuses exclusively on users' autonomous SFV usage behaviors. Therefore, specific interventional experiments involving forced exposure to predetermined content or artificial viewing tasks were excluded. Instead, included studies operationalized autonomous usage primarily through self-reported measures of usage frequency, duration, or validated intensity scales. Furthermore, conference abstracts lacking full texts or complete data, review studies, qualitative research, editorials, and letters were excluded from consideration. Non-English or non-Chinese literature was also excluded.

Study Selection

After removing duplicates, 2 reviewers (DT and XZ) independently screened the titles and abstracts. Records that did not align with the research objectives, such as those failing to report mental health outcomes, not focusing on SFV users, or having ineligible publication types or study designs, were excluded. Subsequently, the full texts of the remaining articles were retrieved and assessed independently by the 2 reviewers to determine final inclusion based on the criteria. During the literature screening process, any disagreements between the 2 reviewers were resolved through discussion to reach a consensus. In cases where consensus could not be achieved, a third reviewer (RH) was available to make a final determination. To ensure the consistency of the screening process, we calculated the interrater reliability using Cohen kappa coefficient, which indicated strong agreement between the reviewers (κ=0.83).

Risk of Bias

Two reviewers (DT and XZ) independently evaluated the quality of the included studies using the Joanna Briggs Institute (JBI) critical appraisal checklist for cross-sectional studies. Disagreements were resolved by discussion or, if necessary, consultation with a third reviewer (RH). The assessment covered three key domains: (1) sample and setting (inclusion criteria and subject description); (2) measurement validity (exposure, condition, and outcome measurement); and (3) confounding and analysis (identification or management of confounders and statistical appropriateness). Although some included studies used a longitudinal or interventional design, our analysis specifically concentrated on baseline data collected before any intervention. Therefore, we restricted our quality assessment to the cross-sectional component of these studies to evaluate the validity and reliability of the baseline data sources, confirming the appropriateness of the JBI cross-sectional checklist for this evaluation.

Data Extraction and Analysis

The extracted data included study design, sample size, participants' age, country, SFV platforms, and measured mental health outcomes. Additionally, statistical metrics reflecting the relationship between SFV use and mental health outcome indicators were extracted, such as Pearson correlation

XSL•FO
RenderX

coefficients, Spearman's rank correlation coefficients, and β coefficients. For intervention and longitudinal studies, we extracted the correlation coefficients at baseline.

Data analysis was carried out using Review Manager version 5.3 (The Cochrane Collaboration) and R version 4.4.2 (R Core Team). To achieve an approximate normal distribution for data synthesis, correlation coefficients (r) were converted to Fisher Z-scores. In instances where correlation coefficients were not reported directly, we used formulas recommended by the Cochrane Handbook for Systematic Reviews to transform other measures of effect size into correlation coefficients [36]. Quantitative synthesis was performed only when the relationship between a specific mental health outcome and SFVs was reported by more than 3 studies. Otherwise, a narrative discussion was conducted for the outcomes.

Statistical heterogeneity was quantified using the I^2 statistic and the Q test. Random-effects models were used when the heterogeneity of the studies was high (I^2>50%), while fixed-effects models were applied when the heterogeneity was low (I^2≤50%).

Based on the existing research, we conducted further subgroup analyses of SFV use by categorizing usage indicators into two types: problematic use, which included measurements of SFV addiction, dependence, and overuse; and routine use, which included daily use duration or frequency. We further explored the relationship between these 2 types of usage indicators and mental health to determine whether there were differences in how these indicators were associated with mental health outcomes.

Potential publication bias was assessed through visual inspection of funnel plots, where asymmetry was evaluated. In accordance with the recommendations of the Cochrane Handbook, for outcomes including 10 or more studies, we further conducted quantitative assessments using Egger's linear regression test and Begg's rank correlation test. A P value of less than .05 was considered to indicate statistically significant publication bias.

Sensitivity analyses were performed to evaluate the robustness of the pooled results. First, we conducted a "leave-one-out" analysis to determine if any single study disproportionately influenced the overall effect size. Second, we assessed the impact of study quality on the outcomes by excluding studies classified as low quality based on the critical appraisal checklist.

Results

Search Results

A total of 4293 articles were retrieved from the databases. Figure 1 illustrates the identification process. Ultimately, 58 articles were included in the review, of which 53 articles were incorporated into the meta-analysis. Figure 1 presents the identification process using the PRISMA flow diagram.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow chart of the identification process. SFV: short-form video.

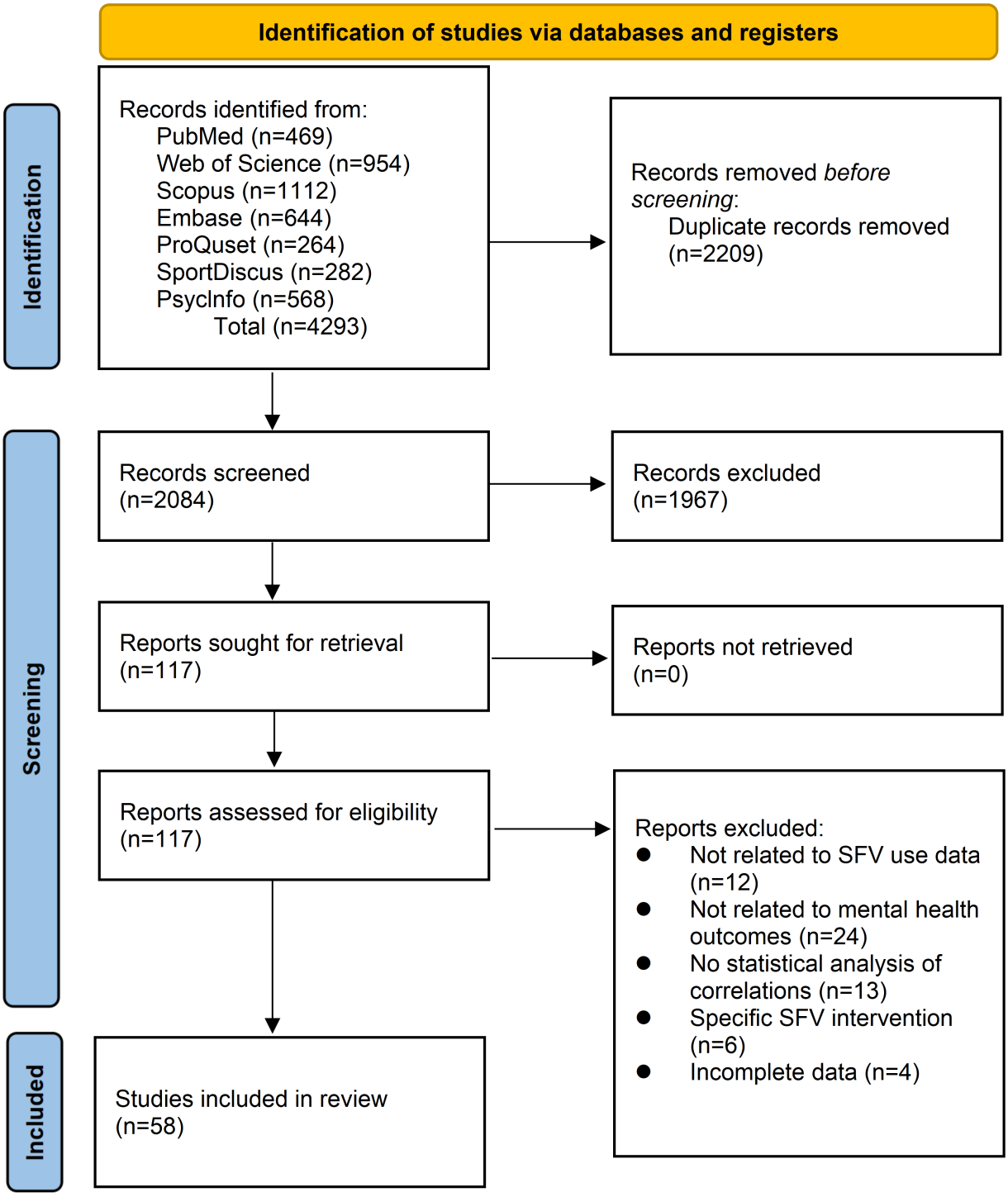

Characteristics of Studies

Among the 58 included studies, the majority were conducted in China (n=55), while a smaller number originated from other countries, including the United Arab Emirates (n=1), Pakistan (n=1), and the United States (n=1). The total accumulated sample size across all studies was 96,676, with individual study sample sizes ranging from 54 to 16,765. The mean age of participants ranged from 11.69 (SD 1.11) to 70.65 (SD 6.40) years, with some studies targeting specific populations, such as middle school students [37] and college students [29,38], while others included broader age ranges, such as 18 - 95 years [26]. Only 16 articles reported the specific SFV platforms used. The primary platforms identified were TikTok, YouTube (YouTube Shorts), and Instagram (Instagram Reels). All studies were published between 2021 and 2025.

In terms of study design, 52 studies used a cross-sectional approach, 4 used a prospective design, 1 was an intervention study, and 1 adopted a mixed methods framework. SFV use

XSL•FO
RenderX

was analyzed in various forms across the studies. SFV addiction was the most frequently examined behavior, reported in 52 studies, followed by daily use in 15 studies, problematic use in 5 studies, and overuse or dependence in a smaller subset [38-40]. Primary mental health outcomes assessed in these studies included depression (n=19), anxiety (n=13), loneliness (n=7), and well-being (n=12). Additional reported outcomes included stress, boredom, affect, suicidal ideation, and emotional dysregulation. Detailed study characteristics are shown in Multimedia Appendix 2.

Risk of Bias

Quality assessment of the 58 included studies yielded an overall mean quality score of 79.1% (SD 14.3%). Among the included studies, 27 (46.6%) achieved high quality scores of greater than or equal to 87.5%. The quality assessment revealed significant variability across methodological domains. The highest compliance rates were observed for item 2 (detailed description of study subjects and setting; 96.6%), item 7 (valid and reliable outcome measurement; 96.6%), and item 8 (appropriate statistical analysis; 94.8%). Conversely, substantial methodological limitations were identified in item 1 (clear definition of inclusion criteria; 60.3%). Detailed quality assessment results were provided in Multimedia Appendix 3.

Meta-Analysis

A total of 53 studies were included in the meta-analysis, which examined psychological outcomes categorized into negative psychological states (including depression, stress, anxiety, loneliness, boredom, and negative affect) and positive psychological states (including subjective well-being and positive affect). Studies that were not included in the meta-analysis were synthesized narratively.

SFV Use and Positive Psychological States

As shown in Figure 2, no significant correlations were observed between SFV use and either subjective well-being or positive affect. However, the 2 subgroups, problematic and routine use, exhibited marginally significant heterogeneity in their associations with subjective well-being (P=.08; I^2=68.3%). After excluding the study by Wen [40], problematic SFV use demonstrated a significant negative correlation with subjective well-being (r=−0.25, 95% CI −0.35 to −0.14).

XSL•FO
RenderX

Figure 2. Association between short-form video use and (A) well-being and (B) positive affect [21-23,26,27,34,40-48].

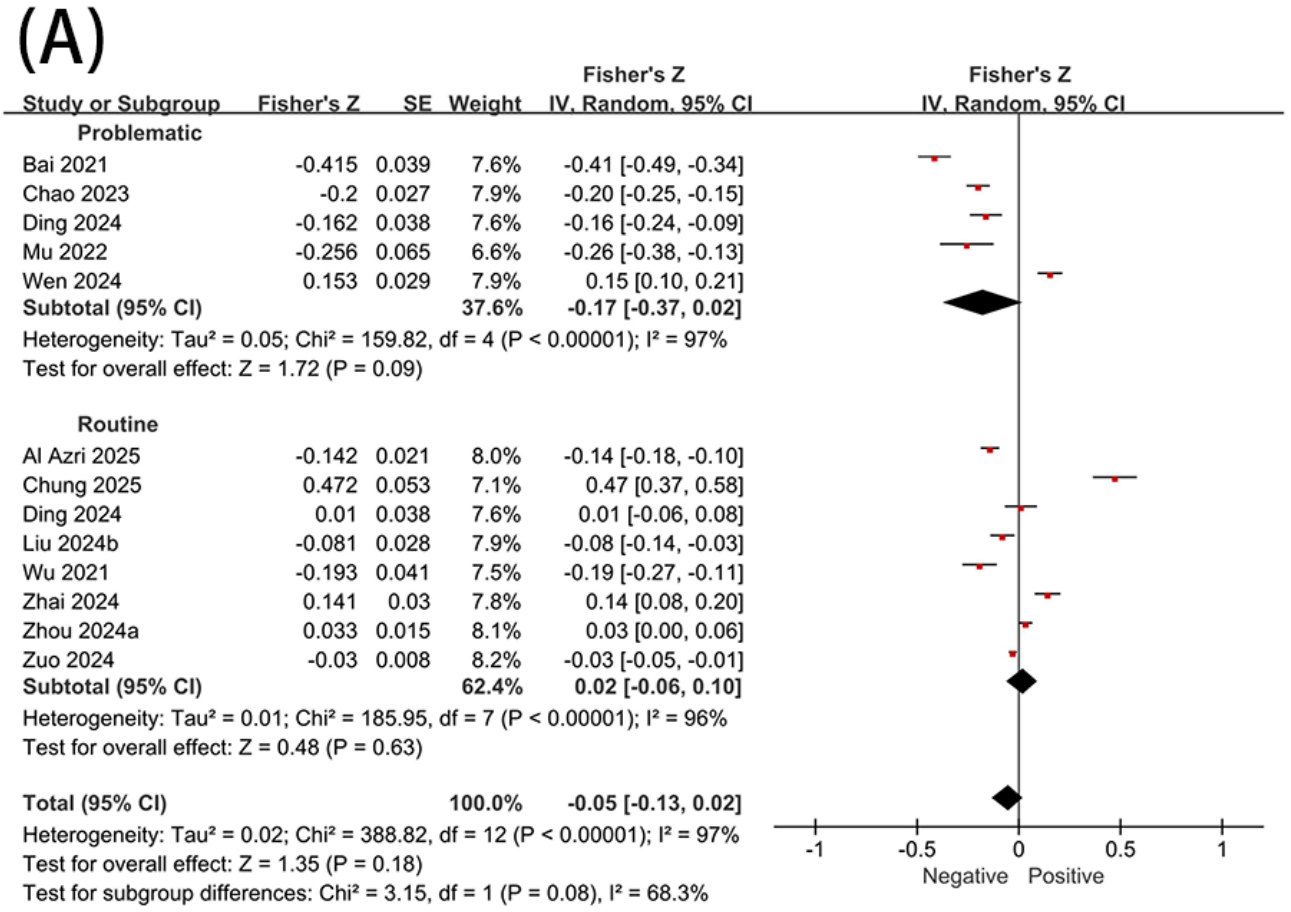

SFV Use and Negative Psychological States

As shown in Figure 3A, SFV use demonstrated a small but significant positive correlation with depression (r=0.24, 95% CI 0.15-0.33). Subgroup analysis revealed significant heterogeneity between the problematic and routine group (P<.001; I^2=96.2%). Problematic use demonstrated a significant moderate positive correlation with depression (r=0.34, 95% CI 0.28-0.39), while routine use showed a weak but significant positive association with depression (r=0.09, 95% CI 0.01-0.17). However, in the leave-one-out sensitivity analysis, the association between general use and depression became

XSL•FO
RenderX

statistically nonsignificant when [25,25], Zhang 2024 [49], Xia 2024 [50], or [44] were excluded individually. This suggests that the current evidence regarding the positive correlation between daily use and depression is not robust, and these findings should be interpreted with caution.

SFV use was significantly positively correlated with anxiety (r=0.26, 95% CI 0.17-0.35). Significant heterogeneity was found between the subgroups of problematic and routine use (P<.001; I^2=95.6%). Problematic use showed a moderate to large positive correlation with anxiety (r=0.38, 95% CI 0.33-0.43), while routine use demonstrated a nonsignificant negative association with anxiety (Figure 3B).

XSL•FO
RenderX

Figure 3. Association between short-form video use and depression (A) and anxiety (B) [13,22,23,25,28,29,31,38,39,43-45,47,49-62].

XSL•FO
RenderX

SFV use was found to be significantly associated with stress (r=0.41, 95% CI 0.38-0.56). Significant heterogeneity was also found between the subgroups of problematic and routine use (P=.006; I^2=86.6%). Specifically, problematic use demonstrated a moderate to large positive correlation with stress (r=0.44, 95% CI 0.36-0.51), whereas routine use showed a small to moderate positive association with stress (r=0.23, 95% CI 0.1-0.35). However, this result should be interpreted with caution as the routine use subgroup comprised only 1 study (Figure 4A). Regarding negative affect, SFV use was also found to be significantly positively associated (r=0.22, 95% CI 0.1-0.33). Problematic use exhibited a moderate and significant positive correlation with negative affect (r=0.3, 95% CI 0.150.43), while no significant association was observed between routine and negative affect (Figure 4B).

XSL•FO
RenderX

Figure 4. Association between short-form video use and (A) stress and (B) negative affect [3,14,15,22,23,47,48,57,58,63-66].

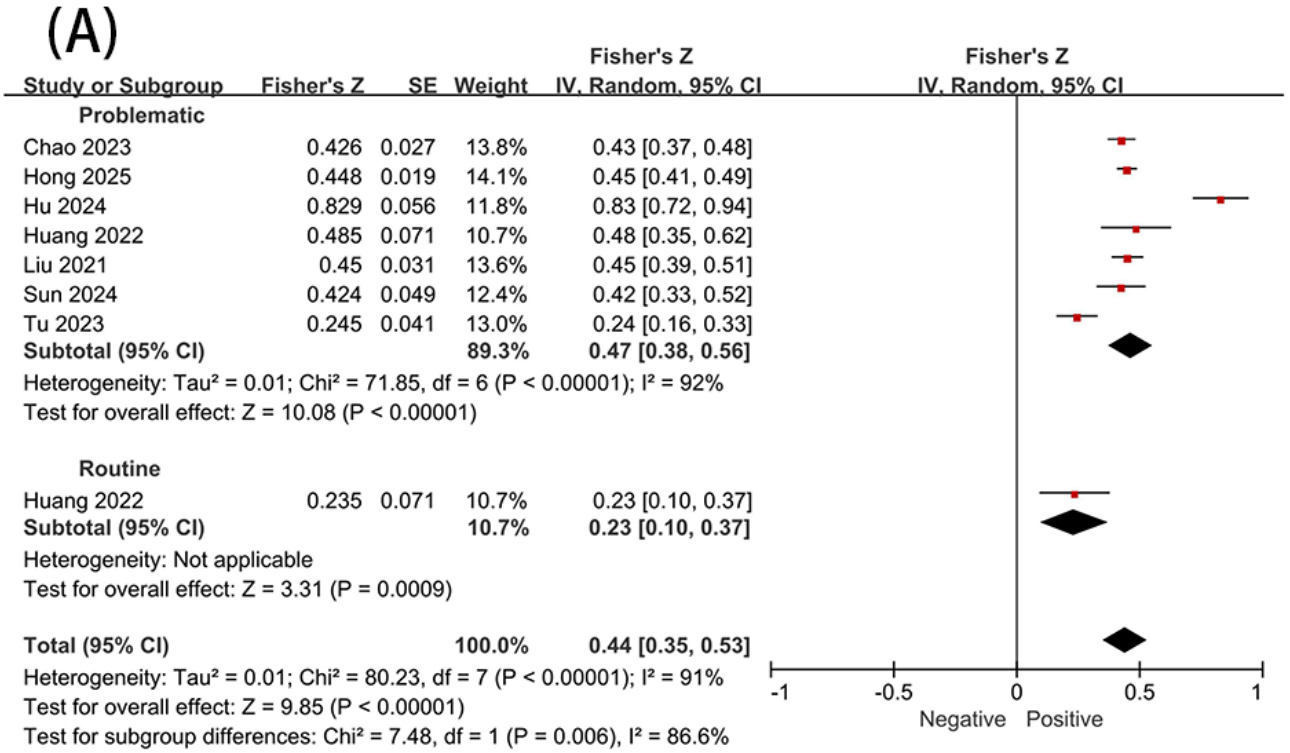

For SFV use and its associations with loneliness and boredom, only studies examining the relationships between problematic use and these 2 variables were sufficient for analysis. Therefore, no further subgroup analyses were conducted. Problematic SFV use demonstrated significant positive correlations with both loneliness (r=0.33, 95% CI 0.25-0.41) and boredom (r=0.42, 95% CI 0.29-0.53) (Figure 5).

Additionally, in studies not included in the meta-analysis, we found that problematic SFV use was positively associated with suicidal attempts (r=0.07; P<.05) [13]. SFV addiction showed positive associations with emotional suppression (r=0.29; P<.001) [37], alexithymia (r=0.339; P<.001) [67], emptiness (r=0.66, P<.001) [68], and neuroticism (r=0.3; P<.01) [69], (r=0.27; P<.01) [70], and (r=0.39; P<.001) [28]. And SFV

XSL•FO
RenderX

addiction was also found to be significantly associated with adolescent emotional dysregulation [71].

Figure 5. Association between short-form video use and (A) loneliness and (B) boredom [23,27,40,51,59,69,72-78].

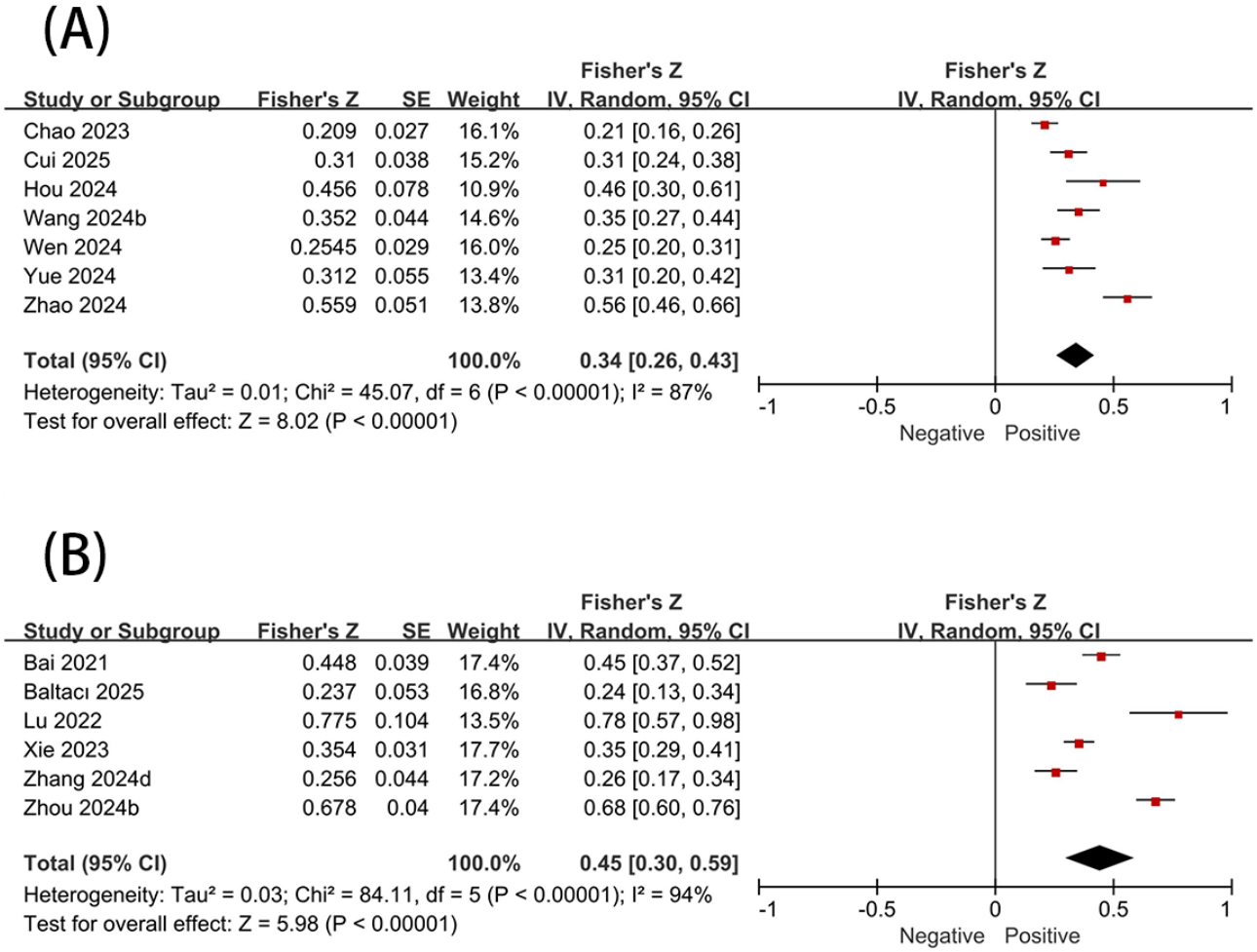

Publication Bias and Sensitivity Analysis

Funnel plots for all analyses showed no obvious asymmetry (Multimedia Appendix 4). Furthermore, for outcomes including 10 or more studies, Egger and Begg tests consistently yielded *P* values greater than .05, indicating no significant evidence of publication bias. Additionally, only 1 study was classified as low quality (JBI score <50%) [79]. Although this study was identified as low quality, it was primarily excluded from the final meta-analysis because insufficient studies reported the specific mental health indicators it analyzed to allow for pooling. Instead, it was synthesized narratively. Consequently, a sensitivity analysis comparing results before and after excluding low-quality studies was not feasible. Leave-one-out sensitivity analyses showed that the statistical significance of our main results remained consistent, confirming their overall robustness. However, in the subgroups examining the relationship between depression and routine use, and well-being and problematic use, excluding specific studies altered the significance of the results, as detailed in the Results section. Consequently, the findings concerning these specific associations appear unstable and warrant cautious interpretation.

Discussion

Principal Findings

This systematic review with meta-analysis provides a comprehensive synthesis of the relationship between SFV use and mental health. It is the first to distinguish between problematic and routine usage behaviors, revealing their divergent associations with different mental health outcomes. Analyses of current evidence indicated that SFV use demonstrated no significant associations with indicators of positive mental health, including subjective well-being and positive affect. Conversely, significant associations emerged between SFV use and various negative mental health states, including depression, anxiety, stress, negative affect, loneliness, and boredom.

Notably, our subgroup analyses revealed substantial differences between problematic and routine usage. Specifically, problematic SFV use demonstrated a significant negative association with subjective well-being and positive correlations with adverse mental health outcomes. In contrast, routine usage behaviors showed no significant associations with negative affect, and surprisingly, exhibited a significant negative association with anxiety, suggesting that nonproblematic routine engagement with SFV platforms may even serve an

XSL•FO
RenderX

anxiety-relieving function for some users [22,30]. In addition, current evidence on the relationship between routine use and depression or stress remains inconclusive. The pooled correlations between routine SFV use and depression were sensitive to the removal of individual studies, with statistical significance changing in several leave-one-out analyses, suggesting that these findings lack robustness. These findings also align with established literature indicating that heightened usage duration and elevated dependency levels are associated with increased vulnerability to adverse mental health outcomes and diminished positive psychological functioning compared to moderate or low-frequency usage patterns [23,44]. The marked distinction between problematic and routine usage in relation to mental health outcomes underscores the importance of considering qualitative differences in how individuals engage with SFV platforms, rather than treating SFV use as a monolithic construct. Overall, these findings support a more nuanced conceptualization of SFV use and its relationship with mental health, emphasizing that both the nature and extent of engagement are critical determinants of psychological outcomes.

Overall, in analyzing the relationship between problematic SFV use and mental health, the results of this meta-analysis broadly align with previous findings on problematic internet, smartphone, or social media use, both demonstrating correlations with decreased mental health and various mental health issues [80-83]. Prior research has also identified significant positive correlations between screen time and internalizing and externalizing behavior problems [84]. This suggests that despite differences in content and usage patterns between SFVs and traditional social media or video platforms, problematic SFV use may share similar mechanisms of impact on mental health with problematic internet, smartphone, and social media use.

However, it is noteworthy that our study found the strength of the association between problematic SFV use and negative mental health outcomes to be higher than previously established correlations for problematic phone use, internet use, social media use, and screen time [80-82,85,86]. In this meta-analysis, problematic SFV use shows moderate correlations with depression, anxiety, stress, loneliness, and boredom. This numerical difference might support that SFVs constitute a distinct media format, emphasizing the necessity for specialized research focused on SFVs rather than simply categorizing them under general “screen time” or “social media use” when considering their relationship with mental health.

First, the algorithmic recommendation mechanisms of SFV platforms may more intensely reinforce user immersion and platform stickiness [87-89]. Unlike traditional media, SFV platforms continuously deliver highly personalized content aligned with users’ interests through sophisticated algorithms, creating reinforcement loops that potentially lead to information cocoons and cognitive biases [90], which may further exacerbate negative mental health outcomes [12,91].

Furthermore, the short duration and rapid switching nature of SFVs may significantly impact users’ attention regulation and cognitive processing abilities. Long-term exposure potentially leads to attention fragmentation and increased cognitive burden, affecting daily functioning [78,92]. The highly condensed nature of SFV content intensifies information input, worsening cognitive load and creating information overload that may harm mental health [40].

Additionally, the rise of SFV platforms has transformed user roles, with increasing numbers of users shifting from passive information recipients to content creators compared to traditional social media or mobile phone usage [7,93]. This transition requires users to spend more time creating content on the platform, fundamentally changing usage patterns. Particularly on platforms like TikTok, users can not only create content but also monetize their videos, establishing content creation as a career option [94,95]. This behavioral pattern transformation potentially introduces psychological impacts distinctly different from those associated with traditional video consumption behaviors. For long-term content creators on SFV platforms, metrics such as view counts and video revenue may produce additional psychological effects that differ significantly from the psychological experiences of regular users who primarily engage in viewing activities [95-98]. However, current research has not effectively differentiated between these behavioral patterns, instead using simple measures of usage frequency and dependence to quantify SFV engagement without a nuanced distinction of qualitatively different usage patterns. This limitation may partly explain the substantial heterogeneity observed in our meta-analysis. Previous research on social media and internet usage has established that different usage patterns may exhibit entirely distinct relationships with mental health outcomes, such as differences between active versus passive engagement styles [22,30,86]. Future research needs to explore not only variations in usage intensity but also examine different usage patterns and their differential impacts on mental health.

Furthermore, the discussion regarding causality warrants further elaboration in the future. Although existing evidence reveals significant associations between SFV use and mental health, the cross-sectional nature of the included studies limits our ability to determine the direction of causality. This relationship is likely complex and multidimensional. First, psychological states may serve as antecedents to SFV use. For instance, individuals may engage with SFVs as a coping mechanism or for emotional regulation in response to boredom, loneliness, or stress [99,100]. Second, SFV use may precede psychological states. For example, problematic use may lead to sleep displacement or increased social comparison, thereby precipitating stress or anxiety [44,101]. Finally, a reciprocal relationship is plausible, analogous to the “internet paradox” [102] or the “reinforcing spirals model” [103]. In this scenario, adverse psychological states may precipitate problematic use, which subsequently aggravates the initial mental health concerns. Future longitudinal inquiries are indispensable for elucidating these temporal dynamics.

The effect sizes for correlations between both routine usage and problematic usage with mental health outcomes exhibited substantial heterogeneity in our findings. This variability likely stems from the wide age ranges, demographic characteristics, usage patterns, and intensity levels represented across existing studies. This heterogeneity suggests that the impact of SFV use on mental health may be influenced by multiple moderating factors, consistent with previous meta-analytic findings for other

XSL•FO
RenderX

digital media. Future research should prioritize more specific investigations to analyze particular effect sizes for specific usage patterns and demographic groups. The differences in measurement instruments highlight the importance of developing and implementing standardized, validated assessment tools for SFV usage patterns and addiction. Specifically, researchers should consider establishing measures that distinguish between content creation and content consumption to more accurately reflect the usage dynamics of modern SFV platforms.

Limitations and Future Directions

Several limitations of the current meta-analysis warrant consideration. First, the included studies were predominantly conducted in China, with limited data from other countries and regions. Given that research has demonstrated the importance of considering cultural and educational backgrounds when investigating problematic smartphone use, caution should be exercised when generalizing these findings to populations in different countries. Previous research on social media has revealed significant variations in associations across different countries and regions [81,85], highlighting the need for future cross-cultural investigations of SFV use.

Second, the limited number of available studies precluded detailed subgroup analyses, assessments of publication bias, and examination of potential moderators such as age, gender, and educational level. Understanding these factors would provide valuable insights into how SFV use and mental health relationships vary across demographic profiles. Future studies should therefore prioritize comprehensive reporting of both participant characteristics (age, gender, and education) and SFV usage patterns (platforms and content types). Particularly important is the distinction between content creators and passive viewers, alongside demographic information. Such detailed data are essential for developing a more nuanced understanding of these relationships.

Third, most of the included studies relied on self-report measures with considerable heterogeneity in assessment instruments, and many studies lacked validated methods or objective standard measurements for SFV use, such as actual usage time and frequency data obtained directly from platform analytics. To enhance cross-study comparability, there is an urgent need to develop and validate standardized measurement tools specifically designed for SFV use or addiction. Future research should differentiate between various SFV usage patterns and explore the potential differential associations between qualitatively distinct forms of engagement and mental health outcomes. Additionally, researchers should consider using more objective approaches to assess users' mental health status, including a diversified indicator system comprising physiological measures, biofeedback techniques, and behavioral data.

Finally, most of the included studies employed cross-sectional designs, which preclude causal inferences and directional conclusions regarding the relationships observed. Future research should prioritize longitudinal investigations to further elucidate causal relationships and examine whether these associations change over time.

Conclusion

This meta-analysis reveals that problematic SFV use is significantly associated with adverse mental health outcomes and reduced psychological well-being. Our findings highlight the distinct relationships between problematic versus routine usage patterns and mental health. Future research should develop standardized assessment tools that better capture contemporary SFV usage patterns and differentiate between varied usage behaviors. Longitudinal and experimental studies are needed to clarify causal relationships and examine whether these associations change over time. Additionally, as current evidence is predominantly based on Chinese samples, there is a critical need for data from diverse geographic regions to understand how these relationships may vary across different cultural contexts and user populations worldwide.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for this study.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

DT and XZ conceptualized and designed the study. DT, XZ, RH, and PG performed the literature search, study selection, data extraction, and quality assessment. DT and JF conducted the statistical analysis and interpreted the results. DT drafted the initial manuscript with XZ's substantial input. XZ, JL, RH, PG, JF, and KRS critically revised the manuscript for important intellectual content and performed content revision. All authors reviewed and approved the final version of the manuscript before submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.

XSL•FO
RenderX

[DOCX File, 11 KB - jmir_v28i1e82503_app1.docx]

Multimedia Appendix 2
Characteristics of the included studies.
[DOCX File, 99 KB - jmir_v28i1e82503_app2.docx]

Multimedia Appendix 3
Quality assessment of included studies.
[DOCX File, 29 KB - jmir_v28i1e82503_app3.docx]

Multimedia Appendix 4
Publication bias.
[DOCX File, 494 KB - jmir_v28i1e82503_app4.docx]

Checklist 1
PRISMA checklist.
[DOCX File, 270 KB - jmir_v28i1e82503_app5.docx]

References

1. Zhang X, Wu Y, Liu S. Exploring short-form video application addiction: socio-technical and attachment perspectives. Telemat Inform 2019 Sep;42:101243. [doi: 10.1016/j.tele.2019.101243]
2. Jiang A, Li S, Wang H, et al. Assessing short-video dependence for e-mental health: development and validation study of the short-video dependence scale. J Med Internet Res 2025 Mar 4;27:e66341. [doi: 10.2196/66341] [Medline: 40053762]
3. Sun R, Zhang MX, Yeh C, Ung COL, Wu AMS. The metacognitive-motivational links between stress and short-form video addiction. Technol Soc 2024 Jun;77:102548. [doi: 10.1016/j.techsoc.2024.102548]
4. Davidson J, Liebald B, Liu J, et al. The Youtube video recommendation system. Presented at: Proceedings of the 4th ACM Conference on Recommender Systems (RecSys 2010); Sep 26-30, 2010; Barcelona, Spain p. 293-296. [doi: 10.1145/1864708.1864770]
5. Wright C. Are beauty bloggers more influential than traditional industry experts? J Promot Commun 2017 Dec 31;5(3) [FREE Full text]
6. Zhang N, Hazarika B, Chen K, Shi Y. A cross-national study on the excessive use of short-video applications among college students. Comput Human Behav 2023 Aug;145:107752. [doi: 10.1016/j.chb.2023.107752]
7. Kuss DJ, Griffiths MD. Social networking sites and addiction: ten lessons learned. Int J Environ Res Public Health 2017 Mar 17;14(3):311. [doi: 10.3390/ijerph14030311] [Medline: 28304359]
8. Wang Y. Humor and camera view on mobile short-form video apps influence user experience and technology-adoption intent, an example of TikTok (DouYin). Comput Human Behav 2020 Sep;110:106373. [doi: 10.1016/j.chb.2020.106373]
9. Shimoga SV, Erlyana E, Rebello V. Associations of social media use with physical activity and sleep adequacy among adolescents: cross-sectional survey. J Med Internet Res 2019 Jun 18;21(6):e14290. [doi: 10.2196/14290] [Medline: 31215512]
10. Andreassen CS, Pallesen S. Social network site addiction - an overview. Curr Pharm Des 2014;20(25):4053-4061. [doi: 10.2174/13816128113199990616] [Medline: 24001298]
11. Qu D, Liu B, Jia L, et al. The longitudinal relationships between short video addiction and depressive symptoms: a cross-lagged panel network analysis. Comput Human Behav 2024 Mar;152:108059. [doi: 10.1016/j.chb.2023.108059]
12. He Y, Liu D, Guo R, Guo S. Information cocoons on short video platforms and its influence on depression among the elderly: a moderated mediation model. Psychol Res Behav Manag 2023;16:2469-2480. [doi: 10.2147/PRBM.S415832] [Medline: 37426388]
13. Yu Z, Zhu X, Li Y. The association between problematic short video use and suicidal ideation and self-injurious behaviors: the mediating roles of sleep disturbance and depression. BMC Public Health 2024 Jun 25;24(1):1689. [doi: 10.1186/s12889-024-19191-5] [Medline: 38915039]
14. Huang Q, Hu M, Chen H. Exploring stress and problematic use of short-form video applications among middle-aged Chinese adults: the mediating roles of duration of use and flow experience. Int J Environ Res Public Health 2021 Dec 23;19(1):132. [doi: 10.3390/ijerph19010132] [Medline: 35010389]
15. Liu Y, Ni X, Niu G. Perceived stress and short-form video application addiction: a moderated mediation model. Front Psychol 2021;12:747656. [doi: 10.3389/fpsyg.2021.747656] [Medline: 35002843]
16. Ye JH, Wu YT, Wu YF, Chen MY, Ye JN. Effects of short video addiction on the motivation and well-being of Chinese vocational college students. Front Public Health 2022;10:847672. [doi: 10.3389/fpubh.2022.847672] [Medline: 35619803]
17. Kelly Y, Zilanawala A, Booker C, Sacker A. Social media use and adolescent mental health: findings from the UK Millennium Cohort Study. EClinicalMedicine 2018 Dec;6:59-68. [doi: 10.1016/j.eclinm.2018.12.005] [Medline: 31193561]

XSL•FO
RenderX

18. Liu M, Kamper-DeMarco KE, Zhang J, Xiao J, Dong D, Xue P. Time spent on social media and risk of depression in adolescents: a dose-response meta-analysis. Int J Environ Res Public Health 2022 Apr 24;19(9):5164. [doi: 10.3390/ijerph19095164] [Medline: 35564559]
19. Zhang Y, Li S, Yu G. The relationship between social media use and fear of missing out: a meta-analysis [Article in Chinese]. Acta Psychol Sinica 2021;53(3):273-290. [doi: 10.3724/SP.J.1041.2021.01228]
20. Banjanin N, Banjanin N, Dimitrijevic I, Pantic I. Relationship between internet use and depression: focus on physiological mood oscillations, social networking and online addictive behavior. Comput Human Behav 2015 Feb;43:308-312. [doi: 10.1016/j.chb.2014.11.013]
21. Chung D, Meng Y, Wang J. The role of short-form video apps in mitigating occupational burnout and enhancing life satisfaction among healthcare workers: a serial multiple mediation model. Healthcare (Basel) 2025 Feb 7;13(4):355. [doi: 10.3390/healthcare13040355] [Medline: 39997230]
22. Zhai G, Su J, Chen Z, et al. The relationships between short video usage and subjective well-being: mediation models and network analysis. Behav Sci (Basel) 2024 Nov 12;14(11):1082. [doi: 10.3390/bs14111082] [Medline: 39594382]
23. Chao M, Lei J, He R, Jiang Y, Yang H. TikTok use and psychosocial factors among adolescents: comparisons of non-users, moderate users, and addictive users. Psychiatry Res 2023 Jul;325:115247. [doi: 10.1016/j.psychres.2023.115247] [Medline: 37167877]
24. Sharifian N, Zahodne LB. Daily associations between social media use and memory failures: the mediating role of negative affect. J Gen Psychol 2021;148(1):67-83. [doi: 10.1080/00221309.2020.1743228] [Medline: 32281502]
25. Zhu C, Jiang Y, Lei H, Wang H, Zhang C. The relationship between short-form video use and depression among Chinese adolescents: examining the mediating roles of need gratification and short-form video addiction. Heliyon 2024;10(9):e30346. [doi: 10.1016/j.heliyon.2024.e30346] [Medline: 38707427]
26. Zuo X, Wang R, Hong Z. Does using short video apps impacts life satisfaction: a perspective from psycho-social mechanism. Curr Psychol 2024 Jun;43(24):21242-21256. [doi: 10.1007/s12144-024-05896-8]
27. Bai J, Mo K, Peng Y, et al. The relationship between the use of mobile social media and subjective well-being: the mediating effect of boredom proneness. Front Psychol 2021;11:568492. [doi: 10.3389/fpsyg.2020.568492] [Medline: 33584406]
28. Yang C, Du J, Li X, et al. Association between parental phubbing and short-form video addiction: a moderated mediation analysis among Chinese adolescents. J Affect Disord 2025 Jan 15;369:523-530. [doi: 10.1016/j.jad.2024.10.023] [Medline: 39389113]
29. Zhang D, Yang Y, Guan M. A cross-lagged analysis of the relationship between short video overuse behavior and depression among college students. Front Psychol 2024;15:1345076. [doi: 10.3389/fpsyg.2024.1345076] [Medline: 39086426]
30. Liu Q, Wen J, Wang N, Wang M. Longitudinal associations between TikTok use, self-concept clarity, and anxiety among Chinese emerging adults: exploring differential impacts of active and passive TikTok use. Emerg Adulthood 2025 Apr;13(2):265-277. [doi: 10.1177/21676968241305653]
31. Jiang Y, Feng Y, Qi J, He R, Chao M. The relationship between bullying victimization and academic performance among adolescents: the chain mediating roles of social anxiety and short video addiction. Psychol Sch 2024;61(10):3859-3872. [doi: 10.1002/pits.23267]
32. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
33. Yang Y, Liu RD, Ding Y, Lin J, Ding Z, Yang X. Time distortion for short-form video users. Comput Human Behav 2024 Feb;151:108009. [doi: 10.1016/j.chb.2023.108009]
34. Wu Y, Wang X, Hong S, Hong M, Pei M, Su Y. The relationship between social short-form videos and youth's well-being: it depends on usage types and content categories. Psychol Pop Media 2021;10(4):467-477. [doi: 10.1037/ppm0000292]
35. International Statistical Classification of Diseases and Related Health Problems: Alphabetical Index: World Health Organization; 2004. URL: https://books.google.co.in/books?id=yUH0zAEACAAJ [accessed 2026-02-19]
36. Higgins JPT, Thomas J, editors. Cochrane Handbook for Systematic Reviews of Interventions: Wiley-Blackwell; 2019. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook [accessed 2026-02-19]
37. Yang J, Ti Y, Ye Y. Offline and online social support and short-form video addiction among Chinese adolescents: the mediating role of emotion suppression and relatedness needs. Cyberpsychol Behav Soc Netw 2022 May;25(5):316-322. [doi: 10.1089/cyber.2021.0323] [Medline: 35486859]
38. Zhang Q, Wang A, Zhang J. Relationship between social short-form video usage and mental health-related behaviors among college students [Article in Chinese]. Chin J Sch Health 2023;44(4):586-589. [doi: 10.16835/j.cnki.1000-9817.2023.04.025]
39. Zhang D, Wu J, Li J, Yang Y. Influence of short video overuse behavior on depression of college students: the chain mediating effect of self-concept clarity and meaning in life [Article in Chinese]. Chin J Behav Med Brain Sci 2024;33(4):342-346. [doi: 10.3760/cma.j.cn371468-20230921-00125]
40. Wen X, Zhou Y, Li Y, Li X, Qu P. Perceived overload on short video platforms and its influence on mental health among the elderly: a moderated mediation model. Psychol Res Behav Manag 2024;17:2347-2362. [doi: 10.2147/PRBM.S459426] [Medline: 38882234]

XSL•FO
RenderX

41. Ding J, Hu Z, Zuo Y, Xv Y. The relationships between short video addiction, subjective well-being, social support, personality, and core self-evaluation: a latent profile analysis. BMC Public Health 2024 Dec 18;24(1):3459. [doi: 10.1186/s12889-024-20994-9] [Medline: 39695521]
42. Mu H, Jiang Q, Xu J, Chen S. Drivers and consequences of short-form video (SFV) addiction amongst adolescents in China: stress-coping theory perspective. Int J Environ Res Public Health 2022 Oct 29;19(21):14173. [doi: 10.3390/ijerph192114173] [Medline: 36361050]
43. Al Azri M, Al Harrasi A, Al Ghailani A, et al. The moderating roles of social comparison in the relationship between Instagram reels use and mental health outcomes: a cross-sectional study in Oman. Sci Rep 2025 Nov 11;15(1):39457. [doi: 10.1038/s41598-025-22938-8] [Medline: 41219267]
44. Liu M, Zhuang A, Norvilitis JM, Xiao T. Usage patterns of short videos and social media among adolescents and psychological health: a latent profile analysis. Comput Human Behav 2024 Feb;151:108007. [doi: 10.1016/j.chb.2023.108007]
45. Zhou D, Li W. Don't say it's too late in life: how short video viewing affects the mental health of the elderly [Article in Chinese]. Doc In Knowl 2024;41(36–45):64. [doi: 10.13366/j.dik.2024.05.036]
46. Tian X, Bi X, Chen H. How short-form video features influence addiction behavior? Empirical research from the opponent process theory perspective. Inf Technol People 2023 Jan 13;36(1):387-408. [doi: 10.1108/ITP-04-2020-0186]
47. Wu Y, Bai Y, Liu X, Xu W, Liu Y. Gender differences in the relationship between short-form video addiction and adolescent depression: the mediating role of attentional bias. Cyberpsychol Behav Soc Netw 2025 Mar;28(3):169-177. [doi: 10.1089/cyber.2024.0442] [Medline: 39772883]
48. Drivas M, Reed OS, Berndt-Goke M. #WhatIEatInADay: the effects of viewing food diary TikTok videos on young adults' body image and intent to diet. Body Image 2024 Jun;49:101712. [doi: 10.1016/j.bodyim.2024.101712] [Medline: 38636388]
49. Zhang R, Su Y, Lin Z, Hu X. The impact of short video usage on the mental health of elderly people. BMC Psychol 2024 Oct 31;12(1):612. [doi: 10.1186/s40359-024-02125-6] [Medline: 39482710]
50. Xia LL, Li JD, Duan F, et al. Effects of online game and short video behavior on academic delay of gratification - mediating effects of anxiety, depression and retrospective memory. Psychol Res Behav Manag 2023;16:4353-4365. [doi: 10.2147/PRBM.S432196] [Medline: 37908679]
51. Baltacı Ö, Açar I. Does leisure boredom predict short video addiction in adolescents? Psychiatr Q 2025 Jun 12. [doi: 10.1007/s11126-025-10172-4] [Medline: 40504300]
52. Dong P, Zhang X, Yin W, et al. Relationship Between Short Video Addiction Tendency and Depression Among Rural Older Adults: Cross-Sectional Study. J Med Internet Res 2025 Jun 30;27:e75938. [doi: 10.2196/75938] [Medline: 40587851]
53. Li L, Li X, Li Y, Liu XP, Huang L. Types of short video addiction among college freshmen: effects on career adaptability, insomnia, and depressive symptoms. Acta Psychol (Amst) 2024 Aug;248:104380. [doi: 10.1016/j.actpsy.2024.104380] [Medline: 38955033]
54. Peng B, Zhu X, Han T, Zhao S, Zheng J, Jia F. Stepping away from the scroll: a chain mediation model from physical exercise and adolescent short video addiction. BMC Psychol 2025 Dec 22;13(1):1366. [doi: 10.1186/s40359-025-03710-z] [Medline: 41430732]
55. Xu X, Li H, Bai R, Liu Q. Do boys and girls display different levels of depression in response to mobile phone addiction? examining the longitudinal effects of four types of mobile phone addiction. Psychol Res Behav Manag 2024;17:4315-4329. [doi: 10.2147/PRBM.S487298] [Medline: 39711982]
56. Li C, Wang Y. Short-form video applications usage and functionally dependent adults' depressive symptoms: a cross-sectional study based on a national survey. Risk Manag Healthc Policy 2024;17:3099-3111. [doi: 10.2147/RMHP.S491498] [Medline: 39676824]
57. Liu L, Wang Y. The influence of watching short videos on the mental health of the middle-aged and elderly in China [Article in Chinese]. Mod Prev Med 2024;51:111-117. [doi: 10.20043/j.cnki.MPM.202305090]
58. Cheng X, Su X, Yang B, Zarifis A, Mou J. Understanding users' negative emotions and continuous usage intention in short video platforms. Electron Commer Res Appl 2023 Mar;58:101244. [doi: 10.1016/j.elerap.2023.101244]
59. Hou A, Hou X. Enjoyment or loneliness: the influence of cognitive absorption on addiction to short-form video apps [Article in Chinese]. Bull Educ Psychol 2024;55:537-556. [doi: 10.6251/BEP.202403_55(3).0005]
60. Li G, Geng Y, Wu T. Effects of short-form video app addiction on academic anxiety and academic engagement: the mediating role of mindfulness. Front Psychol 2024;15:1428813. [doi: 10.3389/fpsyg.2024.1428813] [Medline: 39444830]
61. Mao M, Liao F. Undergraduates short form video addiction and learning burnout association involving anxiety symptoms and coping styles moderation. Sci Rep 2025 Jul 7;15(1):24191. [doi: 10.1038/s41598-025-09656-x] [Medline: 40624123]
62. Wang P, Ouyang M, Yin L, et al. Gender differences in the relationship between parent-adolescent attachment anxiety and adolescent short-form video addiction: the mediating role of fear of missing out. Curr Psychol 2024 Aug;43(29):24216-24228. [doi: 10.1007/s12144-024-06060-y]
63. Hong X, Du W, Quan X, Yu P, Chen L. Network analysis of the interactions of short video features and personal characteristics for short video addiction in Chinese young adults. Acta Psychol (Amst) 2025 Oct;260:105482. [doi: 10.1016/j.actpsy.2025.105482] [Medline: 40885173]
64. Hu H, Huang M. How stress influences short video addiction in China: an extended compensatory internet use model. Front Psychol 2024;15:1470111. [doi: 10.3389/fpsyg.2024.1470111] [Medline: 39583000]

XSL•FO
RenderX

65. Tu W, Nie Y, Liu Q. Does the effect of stress on smartphone addiction vary depending on the gender and type of addiction? Behav Sci (Basel) 2023 Sep 30;13(10):810. [doi: 10.3390/bs13100810] [Medline: 37887460]
66. Deng J, Wang M, Mu W, et al. The relationship between addictive use of short-video platforms and marital satisfaction in older Chinese couples: an asymmetrical dyadic process. Behav Sci (Basel) 2024 Apr 25;14(5):364. [doi: 10.3390/bs14050364] [Medline: 38785855]
67. Li F, Wang G, Zhu D, et al. Parental neglect and short-form video application addiction in Chinese adolescents: the mediating role of alexithymia and the moderating role of refusal self-efficacy. Child Abuse Negl 2023 Sep;143:106345. [doi: 10.1016/j.chiabu.2023.106345] [Medline: 37451180]
68. Mu W, Kong L, He A. Understanding emptiness in a Chinese Sample: cross-cultural validation, latent profile analysis, and association with short-form video addiction. Int J Ment Health Addiction 2025 Oct;23(5):3895-3916. [doi: 10.1007/s11469-024-01327-8]
69. Cui S, Jiang J, Mu L. Impact mechanism of adult offspring's phubbing behavior on elderly short video addiction: a moderated mediation model. Applied Research Quality Life 2024 Apr;20(2):439-458. [doi: 10.1007/s11482-024-10405-x]
70. Asad K, Ali F, Awais M. Personality traits, narcissism and TikTok addiction: a parallel mediation approach. nt J Media Inf Lit 2022;7(2):293-304. [doi: 10.13187/ijmil.2022.2.293]
71. Wang J, Wang M, Lei L. Longitudinal links among paternal and maternal harsh parenting, adolescent emotional dysregulation and short-form video addiction. Child Abuse Negl 2023 Jul;141:106236. [doi: 10.1016/j.chiabu.2023.106236] [Medline: 37192589]
72. Wang P, Yue Y, Ouyang M, Yin L, Yin Y, Li B. Do narcissistic people exhibit more authentic self-disclosure to Generative AI? The roles of short-form video addiction, loneliness, and usage intention. Soc Sci Comput Rev 2025 Dec;43(6):1159-1180. [doi: 10.1177/08944393241308511]
73. Yue H, Yang G, Bao H, Bao X, Zhang X. Linking negative cognitive bias to short‐form video addiction: the mediating roles of social support and loneliness. Psychol Sch 2024 Oct;61(10):4026-4040. [doi: 10.1002/pits.23260]
74. Zhao Z, Kou Y. Effects of loneliness on short video addiction among college students: the chain mediating role of social support and physical activity. Front Public Health 2024;12:1484117. [doi: 10.3389/fpubh.2024.1484117] [Medline: 39600403]
75. Lu L, Liu M, Ge B, Bai Z, Liu Z. Adolescent addiction to short video applications in the mobile internet era. Front Psychol 2022;13:893599. [doi: 10.3389/fpsyg.2022.893599] [Medline: 35619797]
76. Xie J, Xu X, Zhang Y, et al. The effect of short-form video addiction on undergraduates' academic procrastination: a moderated mediation model. Front Psychol 2023;14:1298361. [doi: 10.3389/fpsyg.2023.1298361] [Medline: 38162977]
77. Zhang Y, Bu R, Li X. Social exclusion and short video addiction: the mediating role of boredom and self-control. Psychol Res Behav Manag 2024;17:2195-2203. [doi: 10.2147/PRBM.S463240] [Medline: 38832345]
78. Zhou L, Lv X, Zhou Y, Li J, Yu Z, Gao X. A network analysis perspective on the relationship between boredom, attention control, and problematic short video use among a sample of Chinese young adults. Int J Ment Health Addiction 2025 Dec;23(6):5022-5042. [doi: 10.1007/s11469-024-01392-z]
79. Abulibdeh ES, Alneyadi S, Skaik H, El-Saleh MS, Libdeh FIA, Naser K. Short videoaddiction and its relationship with students' academic achievement and well-being: a pilot study. Presented at: 2024 Global Digital Health Knowledge Exchange & Empowerment Conference (gDigiHealth.KEE); Sep 24-26, 2024. [doi: 10.1109/gDigiHealth.KEE62309.2024.10761759]
80. Cai Z, Mao P, Wang Z, Wang D, He J, Fan X. Associations between problematic internet use and mental health outcomes of students: a meta-analytic review. Adolesc Res Rev 2023;8(1):45-62. [doi: 10.1007/s40894-022-00201-9] [Medline: 36744121]
81. Huang C. A meta-analysis of the problematic social media use and mental health. Int J Soc Psychiatry 2022 Feb;68(1):12-33. [doi: 10.1177/0020764020978434] [Medline: 33295241]
82. Shannon H, Bush K, Villeneuve PJ, Hellemans KG, Guimond S. Problematic social media use in adolescents and young adults: systematic review and meta-analysis. JMIR Ment Health 2022 Apr 14;9(4):e33450. [doi: 10.2196/33450] [Medline: 35436240]
83. Yigiter MS, Demir S, Dogan N. The relationship between problematic social media use and depression: a meta-analysis study. Curr Psychol 2024 Mar;43(9):7936-7951. [doi: 10.1007/s12144-023-04972-9]
84. Eirich R, McArthur BA, Anhorn C, McGuinness C, Christakis DA, Madigan S. Association of screen time with internalizing and externalizing behavior problems in children 12 years or younger: a systematic review and meta-analysis. JAMA Psychiatry 2022 May 1;79(5):393-405. [doi: 10.1001/jamapsychiatry.2022.0155] [Medline: 35293954]
85. Ahmed O, Walsh EI, Dawel A, Alateeq K, Espinoza Oyarce DA, Cherbuin N. Social media use, mental health and sleep: a systematic review with meta-analyses. J Affect Disord 2024 Dec 15;367:701-712. [doi: 10.1016/j.jad.2024.08.193] [Medline: 39242043]
86. Godard R, Holtzman S. Are active and passive social media use related to mental health, wellbeing, and social support outcomes? A meta-analysis of 141 studies. J Comput Mediat Commun 2023 Nov 8;29(1):zmad055. [doi: 10.1093/jcmc/zmad055]
87. Liao M. Analysis of the causes, psychological mechanisms, and coping strategies of short video addiction in China. Front Psychol 2024;15:1391204. [doi: 10.3389/fpsyg.2024.1391204] [Medline: 39165759]

XSL•FO
RenderX

88. Qin Y, Omar B, Musetti A. The addiction behavior of short-form video app TikTok: the information quality and system quality perspective. Front Psychol 2022;13:932805. [doi: 10.3389/fpsyg.2022.932805] [Medline: 36148123]
89. Jia Y, Zheng K, Liu X, Guo X. How short video platforms retain customers: focusing on the roles of user stickiness and flow experience. Asia Pac J Mark Logist 2025 Oct 16;37(8):2308-2326. [doi: 10.1108/APJML-05-2024-0645]
90. Geschke D, Lorenz J, Holtz P. The triple-filter bubble: using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. Br J Soc Psychol 2019 Jan;58(1):129-149. [doi: 10.1111/bjso.12286] [Medline: 30311947]
91. Sunstein CR. Infotopia How Many Minds Produce Knowledge: Oxford University Press; 2006. URL: https://academic.oup.com/book/50393 [accessed 2026-02-19] [doi: 10.1093/oso/9780195189285.001.0001]
92. Firth J, Torous J, Stubbs B, et al. The "online brain": how the Internet may be changing our cognition. World Psychiatry 2019 Jun;18(2):119-129. [doi: 10.1002/wps.20617] [Medline: 31059635]
93. Omar B, Dequan W. Watch, share or create: the influence of personality traits and user motivation on TikTok mobile video usage. Int J Interact Mob Technol 2020;14(4):121. [doi: 10.3991/ijim.v14i04.12429]
94. Abidin C. Mapping internet celebrity on TikTok: exploring attention economies and visibility labours. Cult Sci J 2020 Jan;12(1):77-103. [doi: 10.5334/csci.140]
95. Kaye DBV, Chen X, Zeng J. The co-evolution of two Chinese mobile short video apps: parallel platformization of Douyin and TikTok. Mob Media Commun 2021 May;9(2):229-253. [doi: 10.1177/2050157920952120]
96. Arriagada A, Ibáñez F. "You need at least one picture daily, if not, you're dead:" content creators and platform evolution in the social media ecology. Social Media + Society 2020 Jul;6(3). [doi: 10.1177/2056305120944624]
97. Duffy BE, Meisner C. Platform governance at the margins: social media creators' experiences with algorithmic (in)visibility. Media Cult Soc 2023 Mar;45(2):285-304. [doi: 10.1177/01634437221111923]
98. Duffy BE, Wissinger E. Mythologies of creative work in the social media age: fun, free, and "just being me". Int J Commun 2017 Oct 27;11:20-4671 [FREE Full text]
99. Menon D. Factors influencing Instagram Reels usage behaviours: an examination of motives, contextual age and narcissism. Telemat Inform Rep 2022 Mar;5:100007. [doi: 10.1016/j.teler.2022.100007]
100. Deng X, Yu Z. An extended hedonic motivation adoption model of TikTok in higher education. Educ Inf Technol 2023 Oct;28(10):13595-13617. [doi: 10.1007/s10639-023-11749-x]
101. Tingrong Y, Gen Z. Magic mirror, who is the fairest one of all? testing the mediating effect between short-video social media exposure and appearance anxiety. BMC Psychol 2025 Mar 21;13(1):283. [doi: 10.1186/s40359-025-02601-7] [Medline: 40119477]
102. Kraut R, Patterson M, Lundmark V, Kiesler S, Mukophadhyay T, Scherlis W. Internet paradox: A social technology that reduces social involvement and psychological well-being? Am Psychol 1998;53(9):1017-1031. [doi: 10.1037/0003-066X.53.9.1017] [Medline: 9841579]
103. Slater MD. Reinforcing spirals: the mutual influence of media selectivity and media effects and their impact on individual behavior and social identity. Commun Theory 2007 Aug;17(3):281-303. [doi: 10.1111/j.1468-2885.2007.00296.x]

Abbreviations

JBI: Joanna Briggs Institute
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SFV: short-form video

Edited by A Schwartz, M Balcarras; submitted 16.Aug.2025; peer-reviewed by JY Fam, R Zhang, S Huiqi; accepted 23.Jan.2026; published 18.Mar.2026.

Please cite as:
Tang D, Zhang X, Gou P, Feng J, Hu R, Sum KWR
Association Between Short-Form Video Use and Mental Health: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e82503
URL: https://www.jmir.org/2026/1/e82503
doi:10.2196/82503

© Di Tang, Xin Zhang, Pengpeng Gou, Jie Feng, Rui Hu, Kim-wai Raymond Sum. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN

XSL•FO
RenderX

1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Telehealth Approaches for Pediatric Otitis Media and Clinical Outcomes: Scoping Review

Masao Noda[1*], MD, PhD, MBA; Ryohei Akiyoshi[2*], MD, PhD; Makoto Hosoya[3], MD, PhD; Chikako Shinkawa[4], MD, PhD; Ryota Koshu[1], MD; Hidekane Yoshimura[5], MD, PhD; Yukifumi Monden[6], MD, PhD; Hiroaki Fushiki[7], MD, PhD; Yasuhiro Tanaka[2], MD, PhD; Makoto Ito[1], MD, PhD

[1]Department of Otolaryngology and Head and Neck Surgery, Jichi Medical University, 3311-1, Yakushiji, Shimotsuke, Japan

[2]Department of Otolaryngology and Head and Neck Surgery, Dokkyo Medical University Saitama Medical Center, Saitama, Japan

[3]Department of Otolaryngology and Head and Neck Surgery, Keio University, Tokyo, Japan

[4]Department of Otolaryngology and Head and Neck Surgery, Yamagata University, Yamagata, Japan

[5]Department of Otolaryngology – Head and Neck Surgery, Shinshu University School of Medicine, Matsumoto, Japan

[6]Department of Pediatrics, Jichi Medical University Hospital, Shimotsuke, Japan

[7]Department of Otolaryngology, Mejiro University Ear Institute Clinic, Saitama, Japan

[*]these authors contributed equally

Corresponding Author:
Masao Noda, MD, PhD, MBA
Department of Otolaryngology and Head and Neck Surgery, Jichi Medical University, 3311-1, Yakushiji, Shimotsuke, Japan

Abstract

Background: Otitis media (OM) is a common pediatric infection worldwide. Conventionally, accurate diagnosis depends on in-person pneumatic otoscopy, which is not always accessible, contributing to delayed care and inappropriate prescribing, especially in underserved settings. Rapid advances in telemedicine and digital tools have accelerated the development of remote approaches for assessing pediatric ear diseases, while diagnostic quality maintenance, care models, and real-world outcomes have not been comprehensively mapped in children.

Objective: This study aimed to map existing telehealth technologies and operational models used for pediatric OM and report their diagnostic and implementation outcomes to guide practice and further research.

Methods: PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guided this scoping review (protocol not registered). MEDLINE (via PubMed), Scopus, Web of Science, and the Cochrane Central Register of Controlled Trials were searched for original English-language studies (published between January 1, 2010, and February 9, 2026) reporting the clinical implementation of telemedicine or digital health in children (<18 years) with suspected or confirmed OM. We excluded reviews, editorials, and protocols; conference abstracts; adult-only studies; and purely technical evaluations. Two reviewers independently extracted and charted the study characteristics: telemedicine model, technology, users, comparators, outcomes, and limitations. Findings were synthesized using practice-oriented mapping that aligned technologies with care models and implementation conditions.

Results: Fifty-two studies across 18 countries and various settings met the inclusion criteria. Sample sizes ranged from 6 to 3950, with heterogeneous reporting units (children, ears, episodes, or screening assessments). Asynchronous store-and-forward tele-otoscopy was the most common approach. The rest used synchronous or hybrid models. Image capture by trained personnel and review by experienced clinicians yielded substantial diagnostic consistency with in-person microscopy (κ 0.68 - 0.89, sensitivity 72% - 94%, specificity 93% - 98%, where available). However, the diagnostic yield was highly dependent on the training level: structured instruction improved video capture by parents and nonspecialists, whereas brief or written-only guidance resulted in low rates of diagnostically useful videos. Telemedicine approaches improved access, supported perioperative follow-up, and, in some contexts, reduced reexaminations and promoted more judicious antibiotic use. Televisits without otoscopy were associated with lower confirmation rates of middle ear effusion during tympanostomy tube placement. Evidence was heterogeneous, with predominantly small single-site studies, variable reference standards and operator training, and rapidly evolving device ecosystems.

Conclusions: This review provides a practice-oriented map of telehealth approaches for pediatric OM. Tele-otoscopy and adjunct digital tools are feasible, achieving diagnostic accuracy comparable to in-person assessments while enhancing access and service efficiency. However, important evidence gaps remain, including the need for large multisite trials, evaluation of long-term child outcomes, economic evaluations, and robust external validation of artificial intelligence–based diagnostic tools. Standardization of image capture protocols and integration into hybrid care models should be prioritized for scaling up.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e85416) doi:10.2196/85416

KEYWORDS

pediatrics; acute otitis media; otitis media with effusion; telemedicine; telehealth; smartphone otoscopy; digital otoscopy; artificial intelligence; tympanostomy

Introduction

Telemedicine has expanded rapidly across pediatrics over the past decade, further accelerated by the COVID-19 pandemic. Telehealth performance is disease specific, determined by whether robust remote proxies for the examination exist and are embedded in workflows [1,2]. In practice, the feasibility and safety of remote pediatric care depend on whether key physical findings can be reliably captured and shared. Equity concerns related to devices, connectivity, and digital literacy also underscore the need to track and address the digital divide [3]. For ear disease, this principally entails dependable tympanic membrane visualization and measures of middle ear status [4,5].

Within this context, otitis media (OM) is among the most common pediatric infections worldwide and a leading cause of health care visits, antibiotic prescriptions, and tympanostomy tube placement [6,7]. In 2021, the estimated global incident cases of OM in children were 297 million, highlighting the substantial population burden [6]. OM also contributes substantially to pediatric antibiotic exposure; a previous study reports that in a large outpatient cohort, antibiotics were prescribed in 44.8% of OM encounters [8]. Recent evidence from large outpatient datasets suggests that antibiotic prescribing patterns can differ between telemedicine and in-person visits for acute respiratory infections, raising concerns that diagnostic uncertainty—particularly when objective otoscopic findings are unavailable—may influence management [9]. Acute OM (AOM) and OM with effusion (OME) impose substantial morbidity, reduced quality of life, and increased health care costs across diverse settings [7,10], with a global review reporting annual costs as high as US $5 billion in the United States [10]. Accurate diagnosis hinges on direct visualization of the tympanic membrane and assessment of middle ear status—ideally by pneumatic otoscopy—which requires specific training and equipment [7,11]. Limited access to clinicians trained in pediatric otoscopy, particularly in rural and underserved regions, can contribute to delayed assessment and potentially inappropriate treatment [7].

Concurrently, digital health technologies relevant to ear disease have proliferated [12-14], including smartphone-connected and USB video-otoscopes, non-smartphone digital otoscopes, remote monitoring platforms, and early artificial intelligence–assisted decision support [15,16]. Beyond clinician-facilitated workflows, newer models increasingly rely on caregiver- or patient-captured otoscopic media for asynchronous review; however, interpretability can vary markedly depending on who captures the recordings, indicating that real-world deployment requires systems that standardize capture quality through clear thresholds, training, and feedback loops [17]. When paired with appropriate workflows, these tools enable asynchronous review by experts or real-time supervision and can be complemented by adjunct tests such as tympanometry or acoustic methods [18-23]. Emerging studies suggest that telemedicine can approximate in-person diagnostic performance for pediatric OM when image capture is standardized and operators are trained while also reducing barriers to care and supporting antibiotic stewardship [20,21,24]. However, important uncertainties remain regarding feasibility, usability, and sustainable integration into clinical practice.

Given the rapidly expanding yet fragmented evidence base, we conducted a scoping review to (1) map the technologies and care models used in telehealth for pediatric OM, (2) summarize diagnostic and implementation outcomes, and (3) identify evidence gaps to guide future research and clinical implementation. This review further aims to identify the key barriers and enablers of feasibility and to provide actionable guidance for implementing pediatric OM telehealth—across technologies, care models, and settings.

Methods

Overall Study Design

A scoping review was performed according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist [25]. The review process progressed through the stages outlined in the framework of Arksey and O'Malley [26], which included formulating the research question; identifying relevant studies; selecting studies; extracting and organizing the data; and synthesizing, summarizing, and presenting the findings. The protocol for this scoping review was neither registered nor published. Core decisions were defined a priori (research questions, eligibility criteria, databases, and data items). Minor refinements to the screening guidance, including operational definitions of clinical implementation and tele-otoscopy, were made iteratively during pilot screening to improve consistency and were applied uniformly thereafter. In addition, reporting of the literature search was guided by the PRISMA-S statement (an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews) to enhance transparency and reproducibility of the search methods [27].

Information Sources and Search Strategy

MEDLINE (via PubMed), Scopus, Web of Science Core Collection, and the Cochrane Central Register of Controlled Trials (CENTRAL) were searched for studies published between January 1, 2010, to February 9, 2026 (publication cutoff date), using a predefined strategy developed by 3 authors (MN, RA, and MH). Searches were last run on February 9, 2026, as the final rerun during the revision process; no publication date restrictions were applied beyond the search date, and no automated alerts were used. Each database was searched separately on its native platform rather than via a single multidatabase interface.

XSL•FO
RenderX

The strategy combined controlled vocabulary and free-text terms for three concepts: (1) otitis media, (2) pediatrics, and (3) telehealth or digital health. The start date (January 2010) was selected because smartphone-enabled imaging and telehealth implementations relevant to pediatric otoscopy began to emerge and diffuse in clinical practice around that period. Full strategies for each database are provided exactly as run in Multimedia Appendix 1. The strategy was developed de novo and was not adapted from prior reviews, and no published search filters were used.

In addition, forward citation searching was performed in Web of Science (using "Times Cited" and "Cited Reference Search") for key included articles, and reference lists of included studies were manually screened. Study registries were not searched, and study authors, experts, or manufacturers were not contacted for additional or missing information. The search strategy was not formally peer-reviewed using a dedicated tool but was iteratively reviewed within the author team. Duplicates were removed by automated matching of title, author, and publication year, followed by manual verification.

Eligibility Criteria

We defined the eligibility criteria using the population, concept, and context (PCC) framework developed by the Joanna Briggs Institute (JBI):

- Population: children and adolescents aged <18 years with suspected or confirmed OM (AOM/OME/recurrent acute OM [RAOM]), including mixed-age studies if pediatric data were separable
- Concept: telehealth or digital health approaches for middle ear evaluation and/or OM management, including tele-otoscopy (image/video capture with remote review), tele-otoscopy plus adjunct testing (tympanometry, acoustic methods), and other clinically implemented digital tools relevant to OM assessment or follow-up
- Context: any clinical or community setting (primary care, ED, specialty clinics, schools, outreach programs, home)
- Types of sources: original research reporting clinical implementation outcomes (diagnostic, clinical, implementation, or economic outcomes)

We excluded conference abstracts, protocols, reviews or editorials, adult-only studies, and purely technical evaluations without clinical implementation.

Data Items and Operational Definitions

We used "tele-otoscopy" to describe capture of otoscopic media (images/videos) for remote interpretation (store and forward or real time). We distinguished this from "video visits without otoscopy," in which no ear imaging was obtained. "Adjunct digital diagnostics" referred to nonimaging tools to assess middle ear status (eg, tympanometry and smartphone acoustics).

Study Inclusion Criteria and Exclusion Criteria

Studies were considered eligible if they were published in English between January 1, 2010, and February 9, 2026, and focused on telemedicine, telehealth, or other digital health interventions for pediatric OM. Eligible designs included qualitative, quantitative, or mixed-method approaches, provided that they reported outcomes of implementation in clinical practice. Studies that enrolled children and adolescents aged <18 years with suspected or confirmed OM—in which patients were assessed, monitored, or treated using telemedicine-supported interventions—were included.

Conference abstracts, reviews, editorials, and study protocols were excluded. Studies were also excluded if they focused exclusively on adult populations or conditions other than OM or those that did not involve the clinical application of telemedicine. Purely technical studies—including device prototype development and retrospective artificial intelligence model performance evaluations without prospective or clinical validation—and conference-only publications were excluded. Studies centered solely on audiology or hearing assessment without otoscopic or middle ear evaluation were also excluded.

Selection of Sources of Evidence

Titles and abstracts were independently screened by 2 reviewers (MN and RA). Full-text articles were retrieved for studies deemed potentially eligible and assessed against the predefined inclusion and exclusion criteria. Studies were excluded at the full-text stage for reasons including lack of clinical implementation, exclusive focus on retrospective algorithm performance, conference-only publications, absence of pediatric populations, or absence of OM-specific outcomes. Any disagreements were resolved through discussion and consensus. As part of the revision process, the rerun search results were screened using the same procedures and criteria as the original search.

Data Charting Process and Items

Data extraction was independently performed by 2 reviewers (MN and RA) using a standardized charting form. Extracted items included bibliographic details (author and year) and study design; country and setting (clinic, community, school, or home; level of care); population characteristics (age range, sample size, and OM phenotype when specified); care model (asynchronous, synchronous, or hybrid) and workflow description; operator roles (who captured and who interpreted) and training protocols (duration and content and any quality thresholds); technology and device or system characteristics (type, smartphone requirement, platform, and manufacturer when available); comparator and reference standard (otomicroscopy, pneumatic otoscopy, or surgical findings); outcomes (diagnostic performance, clinical outcomes, implementation outcomes, economic outcomes, and equity-relevant outcomes when reported); and study limitations, potential biases, and funding or conflict of interest statements when reported. Any discrepancies between reviewers were resolved through discussion and consensus. Some included studies reported diagnostic accuracy outcomes; thus, we conducted a targeted assessment for the subset of diagnostic accuracy studies using the revised Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool. Two reviewers independently assessed the risk of bias and applicability, with disagreements being resolved by consensus; detailed judgments are provided in Multimedia Appendix 2.

XSL•FO
RenderX

Synthesis of Results

We summarized the extracted data descriptively (counts/percentages) and synthesized the findings narratively. We created evidence maps (technology category × care model; operator capture vs interpretation) and an outcome-domain frequency figure to visualize evidence distribution and gaps. To apply an implementation science lens, we mapped the reported implementation outcomes and determinants across included studies to the reach, effectiveness, adoption, implementation, and maintenance (RE-AIM) framework. A summary of the RE-AIM framework analysis of each study included in this review is provided in Multimedia Appendix 3.

Results

Selection of Sources of Evidence

The database search retrieved 8356 articles from MEDLINE via PubMed (n=1687), Scopus (n=4810), Web of Science Core Collection (n=1775), and CENTRAL (n=84). In addition, 1911 records were identified through forward citation searching in Web of Science (and backward reference checking). Following the removal of duplicates, 5537 articles remained for the screening of titles and abstracts; of these, 5402 irrelevant articles were excluded, and 135 articles remained. The full texts of 135 articles were subsequently evaluated for eligibility; of these, 83 articles were excluded for the following reasons: nonpediatric population (n=9), absence of telemedicine interventions (n=17), review or commentary articles (n=13), unrelated to OM (n=13), not implemented in clinical practice (n=23), not in English (n=1), and unavailability of full text (n=7). Ultimately, 52 studies were included in this scoping review (Figure 1, PRISMA flow diagram).

Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram.

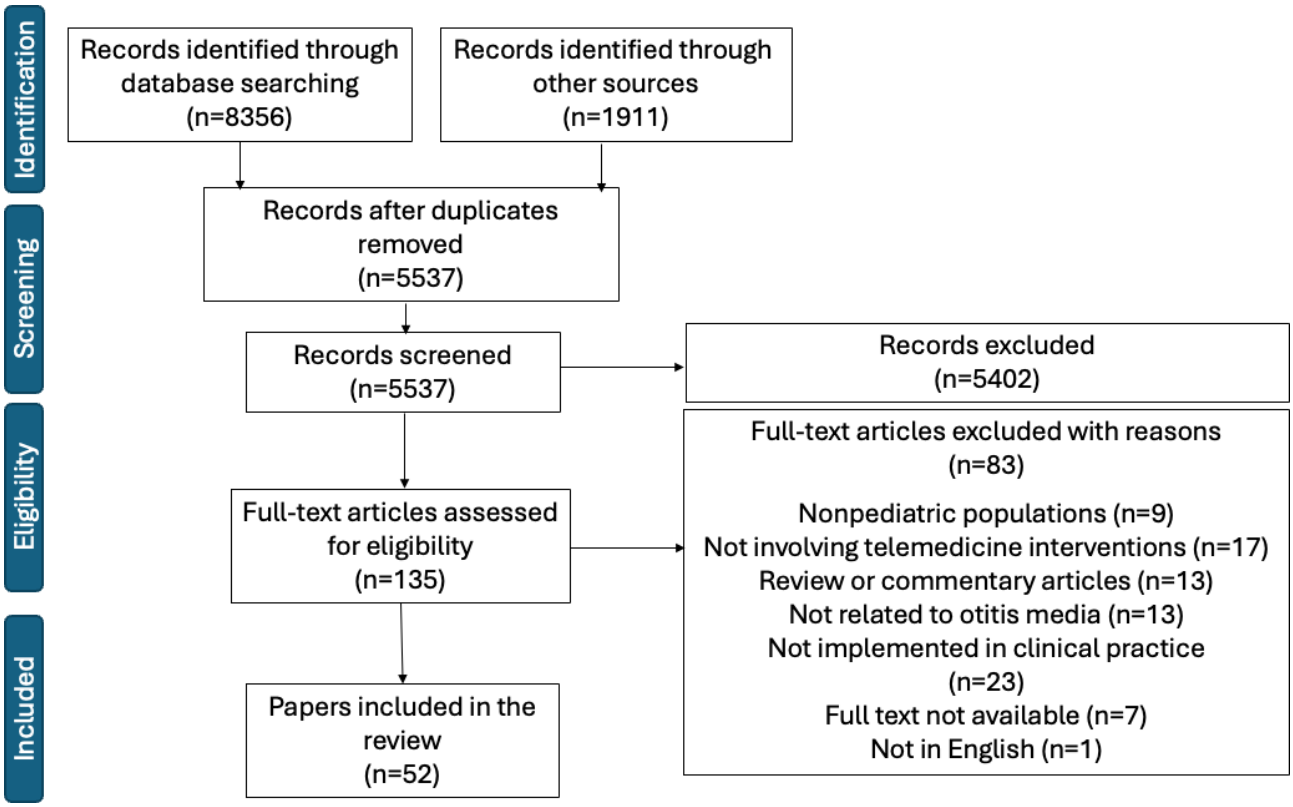

Characteristics of Sources of Evidence

The 52 included studies were published between 2012 and 2026, representing a geographically diverse evidence base encompassing high-, low-, and middle-income settings (summarized in Multimedia Appendix 2) [12-16,22-24,28-71]. QUADAS-2 judgements for diagnostic accuracy studies are summarized in Multimedia Appendix 4. The largest number of included studies (n=15) were conducted in the United States [39,41,42,44,49,52-55,58-60,66,67,71], followed by Australia (n=10) [14,22,23,28,31,32,61,64,65,70], the United Kingdom (n=5) [45,50,51,62,68], India (n=4) [35,40,46,57], and China (n=2) [16,56], with additional studies from other countries, including developing countries (Table 1).

XSL•FO
RenderX

Table . Geographic distribution of included pediatric otitis media telemedicine studies by country (N=52).

Country	Included studies, n (%)[a]	Author (year)
United States	15 (28.8%)	Rappaport et al (2016) [66], Richards et al (2015) [71], Shah et al (2018) [39], Chan et al (2019) [41], Chan et al (2019) [42], Hakimi et al (2019) [44], Jayawardena et al (2020) [49], Kolb et al (2021) [52], Kleinman et al (2021) [53], Don et al (2021) [54], Ni et al (2021) [67], Manayan et al (2022) [55], Schafer et al (2022) [58], Smola et al (2022) [59], Chan et al (2022) [60]
Australia	10 (19.2%)	Smith et al (2012) [28], Phillips et al (2014) [70], Smith et al (2015) [31], Jacups et al (2017) [32], Habib et al (2023) [61], Alenezi et al (2024) [14], Quick et al (2024) [22], Habib et al (2024) [23], Bowers et al (2025) [64], Altamimi et al (2025) [65]
United Kingdom	5 (9.6%)	Fordington and Holland Brown (2020) [45], Cottrell et al (2020) [50], Schuster-Bruce et al (2021) [51], Patel et al (2024) [62], Stancel-Lewis et al (2025) [68]
India	4 (7.7%)	Ramkumar et al (2018) [35], Bhavana et al (2018) [40], Gupta et al (2020) [46], Deshmukh et al (2022) [57]
China	2 (3.8%)	Meng et al (2022) [56], Jin and Fan (2024) [16]
France	2 (3.8%)	Venail et al (2018) [38], Dubois et al (2024) [15]
Nepal	2 (3.8%)	Mandavia et al (2018) [36], Gyawali et al (2024) [63]
South Africa	2 (3.8%)	Biagio et al (2014) [29], Lundberg et al (2017) [34]
Netherlands	1 (1.9%)	Prins-van Ginkel et al (2017) [33]
Brazil	1 (1.9%)	Wagner et al (2023) [13]
Canada	1 (1.9%)	Mousseau et al (2018) [37]
Finland	1 (1.9%)	Erkkola-Anttinen et al (2019) [12]
Greenland	1 (1.9%)	Demant et al (2019) [43]
Hungary	1 (1.9%)	Pannonhalmi et al (2025) [24]
Taiwan	1 (1.9%)	Wu et al (2014) [30]
Turkey	1 (1.9%)	Durgut et al (2020) [48]
Republic of Korea	1 (1.9%)	Cha et al (2020) [47]
Kenya	1 (1.9%)	Yancey et al (2019) [69]

[a]Percentages may not sum to 100% due to rounding.

Designs encompassed diagnostic accuracy comparisons, pragmatic service evaluations, prospective and retrospective cohorts, case series, and education and simulation trials; most studies either enrolled children exclusively or included defined pediatric subsets. Across studies, sample sizes ranged from small pilot cohorts to large-scale community and school-based screening programs.

Results of Individual Sources of Evidence

Telemedicine Models, Users, and Settings

Across the included studies, asynchronous store-and-forward approach was the most commonly used approach, accounting for 35 studies (67.3%; Table 2).

XSL•FO
RenderX

Table . Telemedicine care model classification across included studies on pediatric otitis media (N=52).

Telemedicine model	Included studies, n (%)[a]	Author (year)
Asynchronous[b]	35 (67.3%)	Smith et al (2012) [28], Biagio et al (2014) [29], Phillips et al (2014) [70], Smith et al (2015) [31], Rappaport et al (2016) [66], Prins-van Ginkel et al (2017) [33], Lundberg et al (2017) [34], Mandavia et al (2018) [36], Shah et al (2018) [39], Bhavana et al (2018) [40], Chan et al (2019) [42], Demant et al (2019) [43], Erkkola-Anttinen et al (2019) [12], Yancey et al (2019) [69], Fordington and Holland Brown (2020) [45], Gupta et al (2020) [46], Cha et al (2020) [47], Durgut et al (2020) [48], Cottrell et al (2020) [50], Don et al (2021) [54], Ni et al (2021) [67], Manayan et al (2022) [55], Meng et al (2022) [56], Deshmukh et al (2022) [57], Smola et al (2022) [59], Chan et al (2022) [60], Habib et al (2023) [61], Alenezi et al (2024) [14], Patel et al (2024) [62], Dubois et al (2024) [15], Jin and Fan (2024) [16], Habib et al (2024) [23], Gyawali et al (2024) [63], Altamimi et al (2025) [65], Pannonhalmi et al (2025) [24]
Hybrid[c]	6 (11.5%)	Ramkumar et al (2018) [35], Jayawardena et al (2020) [49], Wagner et al (2023) [13],Quick et al (2024) [22], Bowers et al (2025) [64], Stancel-Lewis et al (2025) [68]
Synchronous[d]	11 (21.2%)	Wu et al (2014) [30], Richards et al (2015) [71], Jacups et al (2017) [32], Mousseau, et al (2018) [37], Venail et al (2018) [38], Chan et al (2019) [41], Hakimi et al (2019) [44], Schuster-Bruce et al (2021) [51], Kolb et al (2021) [52], Kleinman et al (2021) [53], Schafer et al (2022) [58]

[a]Percentages may not total 100% because of rounding.

[b]Asynchronous: capture of images or videos or data with later remote interpretation.

[c]Hybrid: combination of stored media and real-time guidance or interconsultation within the care process.

[d]Synchronous: real-time assessment or interaction (typically video visit or live otoscopy visualization).

In this approach, images or videos were captured by nonphysicians or caregivers and interpreted remotely by otolaryngologists. This includes laptop-tethered video-otoscopy in primary care [29], smartphone otoscopy with consultant review in rural clinics [36], parent-submitted videos in specialty care [39], nurse or technician capture with specialist interpretation in community programs [61], and a multidisciplinary pathway using scheduled remote reviews [14].

Synchronous models accounted for 11 of 52 studies (21.2%) and included real-time postoperative reviews and acute assessments [32,37,52,53,58], whereas blended stored media with live guidance or interconsultation accounted for 6 of 52 studies (11.5%) [13,22,35,49,64,68]. Telemedicine-supported otologic care was implemented across a variety of settings, including primary care clinics, emergency departments, specialty outpatient services, schools, community screening programs, and home environments.

Technologies and Devices

A diverse range of hardware and software solutions were reported across the included studies (Table 3).

XSL•FO
RenderX

Table . Summary of technologies and key outcomes across included pediatric otitis media telemedicine studies (N=52).

References	Implementation/clinical outcome	Diagnostic performance	Typical setting	Representative device/system	Technology category
[12,36,37,41,44,49-51,53,54,71]	Repeat confirmatory examinations decreased from 97.9% to 27.2% when digital otoscopy was used in training clinics [53]. Image acquisition required a median of ~6 minutes per case in rural outpatient workflows [36]. For home tympanostomy tube surveillance, clinicians completed remote reviews in 1 - 3 minutes, families reported mean savings of ~US $127 per episode, and satisfaction was high [54].	Diagnostic concordance with in-person otoscopy was 95%, with κ=0.89, sensitivity 0.94, and specificity 0.96 [36]. In a randomized comparison, trainee diagnostic accuracy improved by 11.2% using digital otoscopy, and interrater agreement rose to Fleiss κ[a]=0.69 versus Fleiss κ=0.40 with a conventional otoscope [53]. Parent-captured videos yielded 40% diagnostic-quality submissions, allowed 87% detection or exclusion of AOM[b] during acute episodes, and achieved κ=0.69 between raters [12].	Emergency, outpatient, school, home; clinicians, trainees, parents	CellScope Oto, Cupris TYM, TYMPA system, endoscope-i	Smartphone-connected digital otoscopes or otoendoscopes
[29,32,34]	Short, focused training improved diagnostic agreement and enabled reliable store-and-forward workflows in primary care [29,34].	Agreement with otomicroscopy was Cohen κ=0.68 - 0.75, with 72.4% - 79.3% sensitivity and 93.2% - 98% specificity [29]. In facilitator-collected videootoscopy reviewed by a general practitioner, repeat video review improved weighted κ from 0.69 to 0.82 (OMgrade), and specificity reached 0.98 [34].	Primary care; CHWs[c] or GPs[d]	Dino-Lite Pro Earscope	USB video-otoscopes
[42]	Users constructed the funnel in a mean (SD) time of 2.8 (0.9) minutes and rated usability as 8.9/10; performance was consistent across phone models following calibration.	Discrimination of middle ear effusion achieved an AUC[e] 0.898, with 84.6% sensitivity and 81.9% specificity, outperforming commercial acoustic reflectometry (AUC 0.776). Caregivers' classifications closely matched that of clinicians in 24/25 ears [42].	Home; caregivers	Paper funnel + smartphone microphone	Smartphone-based acoustic system

XSL•FO
RenderX

References	Implementation/clinical outcome	Diagnostic performance	Typical setting	Representative device/system	Technology category
[60]	Materials cost was <US $30 while maintaining functional parity with a commercial device; the clinical cohort excluded infants younger than 9 months, defining the initial applicability window.	Classification by 5 audiologists, 86% (2%) agreement; all Type B correctly identified. Bland-Altman in 45 intact ears: peak admittance bias=–0.02 (0.14) mL, peak pressure bias=–1 (17) daPa, and ear canal volume bias=–0.09 (0.25) mL [60].	Outpatient; audiologists	Open-source prototype vs GSI TympStar	Low-cost smartphone tympanometer
[33,45,70]	Adoption was high, with 87% of families downloading the app and 74% using it between visits, and clinicians reported improved information exchange during care [45].	A clinic-based app hearing test correlated with pure-tone audiometry (r=–0.656) [45]. In a cohort comparison, the app arm recorded a higher proportion of symptom days (44%) than the paper diary arm (32.5%) and achieved near-complete questionnaire return [33].	Home; parents/patients	Diary tracking, Hear Glue Ear, Mimi	mHealth applications
[15,16,47]	Systems supported self-examination and automated triage, yet generalizability across sites, devices, and pediatric populations remains limited pending broader external validation.	Internal validation of a CNN[f] model reported sensitivity and specificity >98% with AUROC[g] 1.00 across classes; external testing showed reduced sensitivity for AOM (58.3%) and OME (58.3%) with high specificity (100% and 98.1%) [15]. A low-cost USB device with DenseNet[h] achieved 95.7% overall accuracy; for OME[i], sensitivity 0.77 and specificity 0.99; for cholesteatoma, sensitivity 0.79 [16]. A rule-based model reached Top-1 accuracy of 69.5% versus 83% by human telemedicine reviewers [47].	Mixed otology; clinicians and self-examination users	Karl Storz Smart Scope + i-Nside (Inception-v2), DenseNet USB endoscope, rule-based Android	Artificial intelligence–assisted digital otoscopy

XSL•FO
RenderX

References	Implementation/clinical outcome	Diagnostic performance	Typical setting	Representative device/system	Technology category
[13,22,64]	Image capture success reached 96.3% during remote sessions [22]. School programs achieved >85% coverage and reduced costs relative to standard pathways and were rated highly by families, supporting feasibility for integrated tele-audiology [13,64].	In school-entry children, remote testing matched face-to-face within ±10 dB for air conduction at 1 kHz (98%) and 4 kHz (97.8%), bone conduction matched in 100%/95.7%, otoscopy agreement was κ=0.60, and tympanometry agreement was κ=0.90 [64]. Live video-otoscopy showed substantial agreement (κ=0.68 for videos; κ=0.83 for still images) and identified more landmarks (2.79 vs 1.68) [22]. Using a multisensor kit in pediatrics, ear-specific sensitivity was 64.4% (tympanic membrane) and 66.1% (ear canal), with specificity consistently above 70% [13].	School and pediatric tele-audiology	TytoPro; Otometrics A450 + Aurical Otocam 300 + Interacoustics MT10	Multisensor or hybrid diagnostic kits
[14,65]	Median time to diagnosis and care plan fell to 28 days compared with 450 days in standard care (P<.001) [14]. A hospital cost analysis estimated ~US $67 per initial visit (vs US $155) with a break-even point at 223 appointments and about 3-fold greater specialist throughput [65].	Performance was driven by the integrated, multitest pathway rather than a single diagnostic metric [14,65].	Multidisciplinary teleclinic	"Ear Portal": HearScope + HearScreen + Titan (WBT[j]) + DPOAE[k]	Integrated multidisciplinary pathway
[28,31,43,46,57,69]	In a 45-camp program, 3000 individuals were screened; 54% were referred, 13% presented to hospital, and roughly half of presenters required surgery; procedures were delivered at about 40% of usual cost [46]. Longitudinal regional programs achieved ~85% coverage and reduced referrals and waits [28,31].	In CHW-led screening, overall sensitivity was 96.9% with condition-specific concordance of CSOM[l] 96.2%, wax 98%, ASOM[m] 93.9%, foreign body 100%, and OME 100% [57]. When operators received only written instructions, diagnostically useful video yield was 18.1% with substantial interrater agreement (Fleiss κ=0.67) and higher usefulness in older children [43].	Community/schools; CHWs, teachers	ENTraview + ClickMedix; Mebird T5 video-otoscopy + EMRs[n] support follow-up	Community and school screening programs

XSL•FO
RenderX

References	Implementation/clinical outcome	Diagnostic performance	Typical setting	Representative device/system	Technology category
[48,55]	High false-positive rates led to a 66% referral rate in a cohort with 27% hearing loss prevalence; performance was limited by ambient noise and lack of standardized calibration [55].	The Hearing Test app showed κ=0.059, indicating poor agreement with conventional audiometry [48]. The hearScreen app achieved 85% sensitivity and 41% specificity, with ambient noise at 500 Hz significantly increasing false positives (*P*=.01) [55].	Clinics/non-soundproof rooms	Hearing Test, hearScreen	App-based hearing screening

[a]κ: Cohen kappa.

[b]AOM: acute otitis media.

[c]CHW: community health worker.

[d]GP: general practitioner.

[e]AUC: area under the receiver operating characteristic curve.

[f]CNN: convolutional neural network.

[g]AUROC: area under the receiver operating characteristic curve.

[h]DenseNet: Dense Convolutional Network.

[i]OME: otitis media with effusion.

[j]WBT: wideband tympanometry.

[k]DPOAE: distortion product otoacoustic emissions.

[l]CSOM: chronic suppurative otitis media.

[m]ASOM: acute suppurative otitis media.

[n]EMR: electronic medical record.

The most frequently used solution was smartphone-connected digital otoscopes and otoendoscopes, such as CellScope Oto, Cupris TYM, TYMPA system (TympaHealth), and endoscope-i, which were applied in emergency, outpatient, school, and home settings [11,36,37,41,44,49-51,53,54,71]. Earlier clinical evaluation of smartphone-enabled otoscopy for OM diagnosis and management were reported, supporting feasibility in routine care [66]. USB video-otoscopes, including the Dino-Lite Pro Earscope, were used to support store-and-forward apps in clinics [29,34]. Non-smartphone digital otoscopes, such as Welch Allyn Digital MacroView Otoscope, TeleHealth Flexican Otoscope, and Welch Allyn USB Otoscope, were used for real-time follow-up and teleconsultations [22,32].

Apart from imaging, several digital tools were used. A smartphone-only acoustic system, using a paper funnel and the phone's microphone, enabled the detection of middle ear effusion (MEE) [42]. Mobile health (mHealth) apps facilitated diary-based episode detection and home management [33,45]. By contrast, although acceptability was high, a culturally tailored multimedia or text messaging intervention for Aboriginal children with tympanic membrane perforation did not improve clinic attendance or short-term ear health outcomes [70]. A low-cost, open-source smartphone tympanometer was evaluated against the GSI TympStar systems (Grason-Stadler) [60]. Decision support innovations included the Karl Storz Smart Scope with the i-Nside app (Inception-v2) and an inexpensive (~US $10) USB endoscope integrated with a Dense Convolutional Network (DenseNet) classifier for self-examination [15,16]. A consumer Android-tethered otoendoscope, coupled with a rule-based model, demonstrated performance comparable to that of clinicians in diverse otology settings [47]. Community programs frequently used Medtronic's ENTraview integrated with ClickMedix, and 1 program used the Mebird T5 SO (Black Bee Intelligent Manufacturing [Shenzhen] Technology Co., Ltd.) for asynchronous case discussion [35,46,56,57]. A multisensor kit (TytoPro) enabled remote pediatric interconsultation [13]. The school-based hybrid audiology program integrated Madsen A450 audiometry (Otometrics Natus Medical), Natus Aurical Otocam 300 video-otoscopy, and Interacoustics MT10 tympanometry with videoconferencing [64]. For training, Otosim2 (OtoSim Inc.) via AudioProConnect and a low-cost KZYEE USB otoscope improved teaching and landmark recognition [38,44]. Teslong pen endoscopes and the Mimi Hearing Test app (Mimi Hearing Technologies) were piloted for supervised home capture and hearing checks [49,63]. A structured "Ear Portal" pathway combined hearX HearScope video-otoscopy, HearScreen audiometry, Interacoustics Titan wideband tympanometry, and distortion product otoacoustic emissions into a scheduled multidisciplinary review [14,65].

Technology Category × Care Model

Figure 2 maps technology categories against care models. Most studies used asynchronous store-and-forward workflows, primarily tele-otoscopy imaging alone or tele-otoscopy combined with adjunct testing (Figure 2). Hybrid models were less common and tended to involve multisensor kits or tele-otoscopy plus adjunct testing. Notable gaps included limited evidence for integrated pathway models beyond a small number

XSL•FO
RenderX

of programs and sparse evaluation of outcomes beyond diagnostic performance, including long-term child outcomes and cost-effectiveness.

Figure 2. Evidence gap map of technology category by care model in pediatric otitis media telemedicine. mHealth: mobile health.

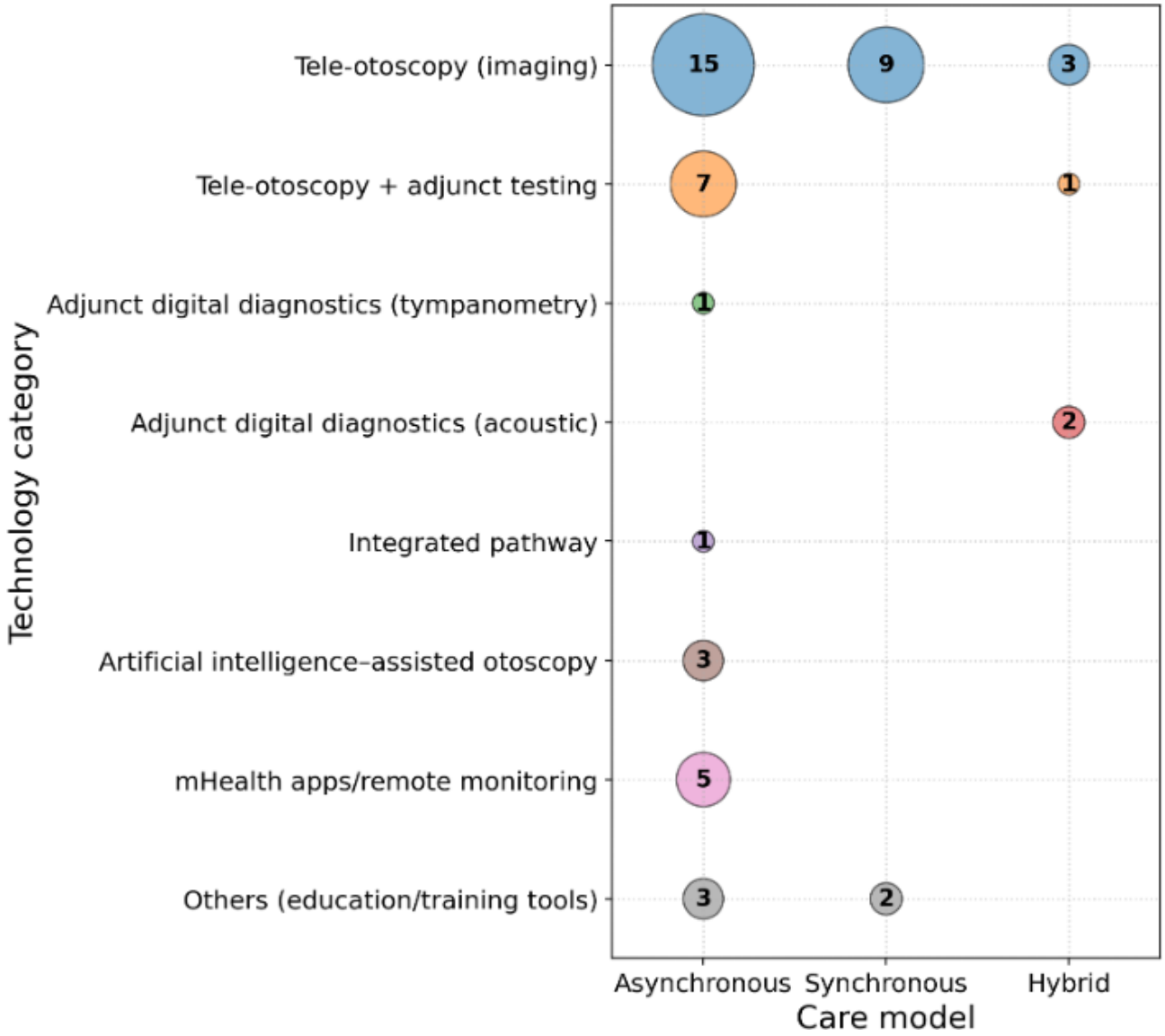

Operators and Workflow Roles

Figure 3 summarizes who captured otoscopic media (clinicians, nurses/community facilitators, caregivers/parents) and who interpreted it. Studies consistently indicated that diagnostic yield depended strongly on operator training and workflow support, which includes checklists, feedback loops, and cerumen management. Evidence gaps included limited direct comparisons of training intensity and limited reporting of minimum image quality thresholds.

XSL•FO
RenderX

Figure 3. Operator workflow map showing who captures otoscopic media and who interprets it in pediatric otitis media tele-otoscopy. CHW: community health worker, ENT: ear, nose, and throat.

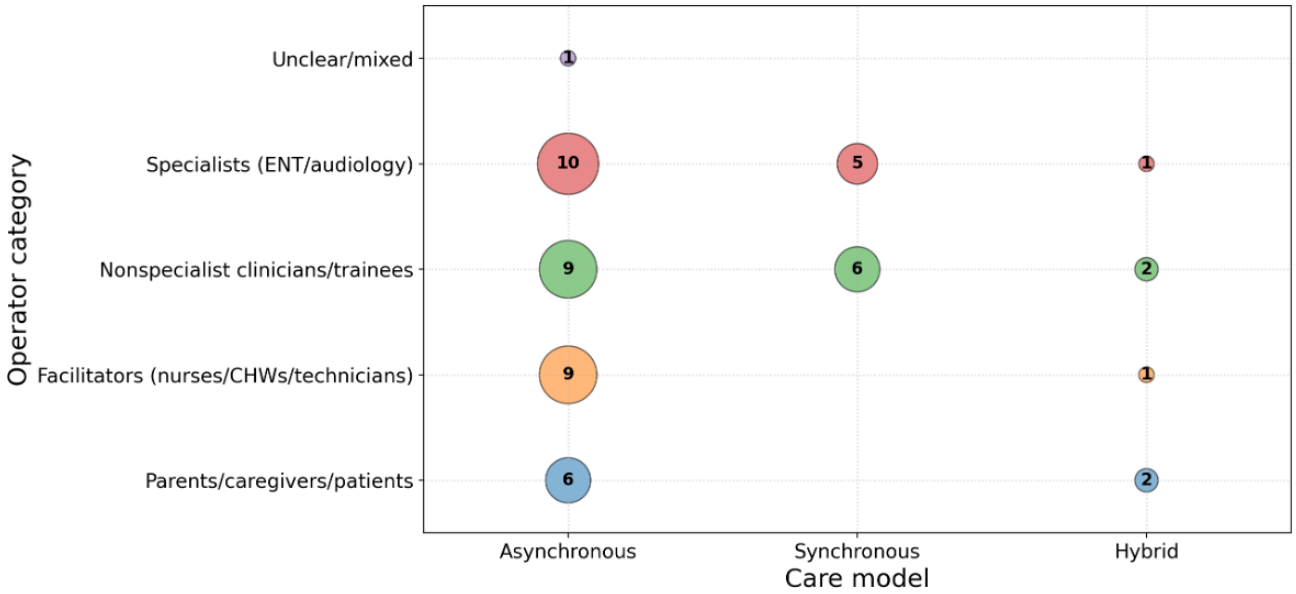

Diagnostic Performance

Across diagnostic accuracy studies, the reported κ, sensitivity, and specificity varied by key sources of clinical and methodological heterogeneity, including pediatric age strata, OM phenotype definitions (AOM vs OME vs mixed), reference standards, operator training and experience, and device and workflow-specific image quality constraints (Table 4). To improve interpretability, we summarized diagnostic performance by capture-interpretation model and reference standard.

XSL•FO
RenderX

Table . Key sources of clinical and methodological heterogeneity across diagnostic accuracy studies in pediatric otitis media tele-otoscopy.

Study and category	Key heterogeneity						Performance (κ[a]/Se[b]/Sp[c]/AUC[d])
	Age	Spectrum	Operator/training	Reference standard	Device	Quality	
(A) Trained operator/expert interpretation							
Biagio et al (2014) [29]	2-16 y	Normal/abnormal OM	Trained capture → clinician read (2 sessions)	Otomicroscopy	Asynchronous video-otoscopy	Acceptable/excellent ratings (NR criteria)	κ 0.68 - 0.75; Se 72.4% - 79.3%; Sp 93.2% - 98%
Lundberg et al (2017) [34]	Pediatrics (NR)[e]	Normal/AOM[f]/AOM[g]/CSOM[h] (abnormal pooled)	Facilitator capture → GP[i] read (repeat)	Otologist otomicroscopy	Video-otoscopy vs otoscopy	NPD[j] due to wax/degree of cooperation/low quality	wκ[k] 0.76→0.82 (video) vs 0.69 (otoscopy); Se 0.77 - 0.81; Sp 0.94 - 0.98
Mandavia et al (2018) [36]	Pediatrics (NR)	Screening (NR)	Trainees capture → remote consultant read	On-site specialist otoscopy/referral	Smartphone otoscopy (Cupris TYM)	NR	Concordance 95% (κ=0.89); Se 0.94; Sp 0.96; referral concordance 100%
(B) Parent/nonspecialist capture (training dependent)							
Erkkola-Anttinen et al (2019) [12]	6-35 mo	AOM-focused	Parent capture (hands-on teaching) → remote raters	Clinician assessment (NR)	Device: smartphone video-otoscopy	Sufficient quality 62% - 64% during intervention; diagnostic quality 40%	Detect/exclude AOM κ=0.69 (87% possible)
Shah et al (2018) [39]	3 mo to 17 y	Mixed pathology	Parent or physician capture → physician read (brief tutorial)	Pneumatic otoscopy (± tympanometry subset)	iPhone/CellScope	Wax common; wax-alert κ≈0.01	Parent κ=0.42; physician κ=0.74 (vs pneumatic otoscopy)
Demant et al (2019) [43]	6-72 mo	Usefulness outcome	Nonspecialist HW capture (written instructions only) → ENT[l] raters	NR	Smartphone otoscopy (Cupris TYM)	Useful videos 6.6% (<24 mo), 22.7% (24-48 mo), 30.7% (>48 mo)	Modified Fleiss κ 0.73 (<24 mo), 0.48 (24-48 mo), 0.60 (>48 mo)
(C) Different outcomes/surrogate standards/different modalities							
Kleinman et al (2021) [53]	Pediatrics (NR)	Education/clinical trial	Trainees	Education/supervisor (NR gold standard)	Digital vs traditional otoscope	NR	Accuracy +11.2%; Fleiss κ 0.40→0.69; repeat confirmatory examinations 97.9%→27.2%
Rappaport et al (2016) [66]	6 mo to 18 y	AOM via ED[m] diagnosis (ICD-9[n] surrogate)	ED capture → blinded physician reviewers	ED diagnosis	Smartphone vs digital otoscope	NR	Overall κ=0.375; smartphone κ=0.368; digital κ=0.381
Chan et al (2019) [42]	18 mo to 17 y (median 5 y)	Chronic OME/recurrent AOM vs controls	Clinician testing;parental use tested after brief demo	Surgical fluid (myringotomy) and/or ENT pneumatic otoscopy	Smartphone acoustic method (paper funnel; ML)[o]	Phone model calibration	AUC 0.898; Se 84.6%; Sp 81.9%; parents matched clinician 24/25 ears

XSL•FO
RenderX

Study and category	Key heterogeneity						Performance (κ[a]/Se[b]/Sp[c]/AUC[d])
	Age	Spectrum	Operator/training	Reference standard	Device	Quality	
Durgut et al (2020) [48]	5-15 y	OME (50 children; 88 ears with OME + 12 controls)	In-clinic testing	Booth pure-tone audiometry	Android Hearing Test app	Environmental noise limitation	Very poor agreement: κ=0.059 (100 ears); no significant correlation 500 - 4000 Hz

[a]κ: Cohen kappa.
[b]Se: sensitivity.
[c]Sp: specificity.
[d]AUC: area under the receiver operating characteristic curve.
[e]NR: not reported.
[f]OME: otitis media with effusion.
[g]AOM: acute otitis media.
[h]CSOM: chronic suppurative otitis media.
[i]GP: general practitioner.
[j]NPD: not possible to diagnose.
[k]wκ: weighted kappa.
[l]ENT: ear, nose, and throat.
[m]ED: emergency department.
[n]ICD-9: International Classification of Diseases, 9th Revision.
[o]ML: machine learning.

Trained Operator Capture with Expert Interpretation Against Clinical Reference Standards

When images were captured by trained operators and interpreted by experienced clinicians, store-and-forward tele-otoscopy showed substantial agreement with in-person otomicroscopy. In a primary care cohort, asynchronous video-otoscopy yielded Cohen κ values of 0.68 - 0.75 across raters and 2 review sessions, 72.4% - 79.3% and specificities of 93.2% - 98% [29]. In another clinic, repeated readings improved weighted κ from 0.69 to 0.82, with specificity up to 0.98 [34]. Smartphone otoscopy used by trainees, with remote consultant review, achieved 95% ear-level concordance with on-site specialist otoscopy (κ=0.89; sensitivity 0.94; specificity 0.96) and 100% concordance with referral decisions [36]. For training, pediatric emergency department residents demonstrated similar diagnostic accuracy when using a smartphone device compared with a traditional otoscope (0.74 vs 0.69; nonsignificant). However, novice users tended to favor the conventional tools, as they felt confident with them [37]. A separate academic comparison showed that digital otoscopy improved trainee accuracy by 11.2%, increased interrater agreement (from Fleiss κof 0.40 to 0.69), and reduced repeat confirmatory examinations from 97.9% to 27.2% [53].

Parent and Nonspecialist Capture (Training- and Cooperation-Dependent Performance)

Parent and nonspecialist performance was highly dependent on training and cooperation. After structured teaching, parents produced videos of sufficient quality in 62% - 64% of submissions during the intervention periods, and the evaluators showed substantial agreement for detecting or excluding AOM (κ=0.69) [12]. In contrast, a brief tutorial yielded lower agreement with pneumatic otoscopy (parent κ=0.42 vs physician κ=0.74), while automated cerumen alerts showed negligible agreement (κ=0.01), highlighting the advantage of structured teaching [39]. Without hands-on training, local health care workers produced diagnostically useful videos in only 18% of cases, with insertion errors and cerumen as major barriers; usefulness increased proportionally with child age [43]. Educational approaches also included a smartphone-adaptable video-otoscopy quiz developed to support otoscopic image interpretation skills, reflecting ongoing efforts to standardize competency development for tele-otoscopy workflows [67].

Alternative Modalities and/or Surrogate Outcomes (Interpret with Caution)

Beyond optical otoscopy workflows, a smartphone-based acoustic method to detect MEE achieved an area under the receiver operating characteristic curve (AUC) of 0.898, with approximately 85% sensitivity and 82% specificity under surgical and/or specialist reference standards, and caregiver performance approximated clinician performance in a subset [42]. A smartphone tympanometer also demonstrated approximately 86% classification agreement with commercial devices, with only minor bias in peak admittance and pressure, supporting the feasibility of adjunct middle ear assessments beyond optical imaging [60]. In contrast, a pediatric hearing test app demonstrated very poor agreement with booth audiometry (κ≈0.06) [48,55]. These findings underscore that performance estimates are not directly comparable across studies when index test modality, outcome definition, and reference standards differ. Artificial intelligence–supported classifiers have shown high performance on curated datasets but lower

XSL•FO
RenderX

sensitivity for OM and limited external validation across devices and sites [15,16,47].

Clinical and Implementation Outcomes

Comparison With Traditional Approaches

Tele-enabled programs have broadened access and accelerated clinical decision-making. A school or community program completed 2111 assessments in 1053 children, demonstrating 85% coverage. In another evaluation, it reduced waiting times and tertiary referrals as the program expanded [28,31]. A general practitioner-to-ear, nose, and throat (ENT) specialist smartphone endoscopy pathway was able to manage 83% of the 53 referrals remotely, with a median specialist response time of 48 (IQR 22-148) minutes [50]. In a pediatric emergency department, a smartphone otoscope showed substantial within-physician agreement versus a traditional otoscope and changed the reported tympanic membrane view in 12% - 16% of examinations, including clinically relevant diagnostic changes to or from AOM in 6% - 7% of examinations [71]. Community screening of 3000 individuals generated 54% of referrals, but only 13% were ultimately presented for hospital evaluation; nearly half of those evaluated required surgery, underscoring the barriers of pathway follow-through [46]. A telehealth-supported surgical model achieved 80% clinical resolution, 88% hearing improvement, and reduced travel costs for postoperative review [32]. Home tympanostomy tube surveillance demonstrated high interrater and intrarater agreement, rapid clinician review (1 - 3 min), and high family satisfaction with reported time and cost savings [54]. The Ear Portal pathway substantially shortened time to diagnosis and care plan (median 28 [IQR 19.8] days), and an economic analysis showed that it lowered per-visit costs, reached a break-even point after ~223 appointments, and increased specialist throughput [14,65]. A community service integrating smartphone otoscopy with tablet audiometry discharged most patients at their first visit at a lower cost than other services using conventional methods, with specialist review rarely leading to changed decisions [62]. However, televisits, which were performed predominantly without ear imaging, produced similar rates of tube recommendations but substantially lower documentation of effusion at surgery compared with encounters that incorporated ear imaging, indicating a potential risk of suboptimal decision-making when otoscopy is omitted [52]. During the COVID-19 pandemic, telehealth encounters were overall less likely to end with tube recommendations, underscoring the limitations imposed by restricted physical examination [58]. A multisensor interconsultation device demonstrated high agreement for several systems but lower sensitivity for ear canal and tympanic membrane findings than conventional methods, reinforcing the necessity of dedicated otoscopy using generalist kits [13]. In pediatric telehealth, live video-otoscopy provided superior image quality and landmark visualization compared to still images, supporting synchronous visualization when feasible [22]. A school-entry hybrid model demonstrated close agreement with face-to-face audiometry and tympanometry, with acceptable agreement for otoscopy [64].

Outcome Domains Reported

Table 5 summarizes the distribution of outcome domains reported across included studies (N=52). Diagnostic performance was the most frequently reported domain (38/52, 73.1%), followed by implementation and usability outcomes (29/52, 55.8%). Clinical and management impact (22/52, 42.3%) and audiology and tympanometry outcomes (19/52, 36.5%) were reported less often, while image or capture quality was reported in 15 of 52 studies (28.8%). Training and education and economic outcomes were each reported in 8 of 52 studies (15.4%), and artificial intelligence or automation was reported in 4 studies (7.7%; Table 5). Reports of longer-term patient outcomes and adverse events were uncommon, indicating priorities for future research.

Table . Outcome domains reported across included studies (N=52).

Outcome domain	Studies reporting domain, n (%)[a]
Diagnostic performance	38 (73.1%)
Implementation/usability	29 (55.8%)
Clinical/management impact	22 (42.3%)
Audiology/tympanometry outcomes	19 (36.5%)
Image/capture quality	15 (28.8%)
Training/education	8 (15.4%)
Economic outcomes	8 (15.4%)
Artificial intelligence/automation	4 (7.7%)

[a]Studies may contribute to multiple domains; therefore, percentages do not sum to 100%.

Applications and Implementation Feasibility

Tele-enabled tools were applied in 4 management scenarios for pediatric ear diseases. For AOM diagnosis, asynchronous video-otoscopy demonstrated substantial agreement with otomicroscopy when images were captured by trained operators, as reported in South Africa using a laptop-tethered Dino-Lite Pro Earscope and specialist review [29] and in a primary care study in which repeated readings of stored video sequences closely tracked the microscopic reference [34]. At the bedside, smartphone-based digital otoscopy (CellScope Oto) demonstrated accuracy comparable to, or modestly better than,

XSL•FO
RenderX

that of trainees using a conventional otoscope in a pediatric ED [37]. With structured instructions, parents successfully captured diagnostically useful otoscopy videos at home, supporting remote triage for suspected AOM [12]. For posttympanostomy tube surveillance and postoperative follow-up, home digital otoscopy with CellScope Oto enabled asynchronous specialist review and was rated highly acceptable by families. In remote Australian communities, a telehealth-supported surgical pathway cut travel and costs for postoperative care [32,54]. For screening and triage in underserved settings, school- and community-based programs—ranging from mobile ear-screening services in Queensland to large-scale campaigns using the ENTraview smartphone otoscope integrated with the ClickMedix platform—efficiently detected pathology, shortened time to specialist input, and improved surgical case prioritization, although barriers such as transport and out-of-pocket costs limited follow-through [28,46]. Additionally, a recent parent-led home ear health check pilot targeting children with complex needs highlighted feasibility and potential access benefits for populations facing barriers to clinic-based otoscopy [68].

Although feasibility was consistently high across settings and devices, the diagnostic yield depended on operator training, child cooperation, and cerumen management. Parents performed better after structured teaching, but accuracy declined with only brief instructions or when wax removal was not addressed [39]. Hands-on practice with a live tele-otoscopy simulation using Otosim2 with the AudioProConnect platform preserved accuracy and shortened task time, whereas written-only instructions under field conditions produced a low proportion of diagnostically useful videos [43]. Clinicians generally rated smartphone otoscopy imagery as adequate and highlighted supervisory advantages of shared visualization, including fewer repeat confirmatory examinations during training encounters [41,53]. Conversely, environmental noise and calibration issues materially influenced app-based hearing tools, which demonstrated poor agreement with booth-based audiometry in children with OME, and a smartphone screening app demonstrated high sensitivity but low specificity when ambient noise was not well controlled [48,55].

Risk of Bias and Study Limitations

Limitations include small, single-site samples; convenience sampling; mixed adult and pediatric cohorts in some studies; lack of blinded reference standards; heterogeneous operator training; device or discontinuation issues; and industrial involvement in some implementations. Parent and nonspecialist performance was sensitive to training and cerumen management, while omitting otoscopy in televisits risked misclassification of MEE [52].

Key Practical Takeaways

Across the included studies, the following points were notably identified as key practical takeaways:

1. Diagnostic performance depended strongly on who captured the otoscopic media and on the training and quality protocol rather than on the device alone.

2. Predefined image quality thresholds, multiview capture, and repeat capture when needed materially affected interpretability and downstream clinical decisions.

3. Telehealth encounters conducted without otoscopic imaging risk underdetecting middle ear pathology, including effusion, and may therefore alter management decisions when objective findings are required.

4. Hub-based or hybrid workflows—such as facilitated capture with remote review—appear to support access and feasibility while maintaining diagnostic reliability in many settings.

5. Evidence remains limited regarding scale-up outcomes (including sustainability/maintenance), formal economic evaluations (particularly in lower-middle-income countries), and external validation of artificial intelligence–enabled tools across devices and sites.

Discussion

Principal Findings

This scoping review mapped a rapidly evolving evidence base for telehealth in pediatric OM. Three consistent signals were observed. First, when images were captured using standardized techniques by trained operators and interpreted by experienced clinicians, store-and-forward tele-otoscopy demonstrated diagnostic performance similar to in-person examination. In primary care and community clinics, agreement with otomicroscopy was frequently in the substantial range, with high specificity for excluding the disease [29,34,36]. Second, operator training proved pivotal: performance declined markedly with brief, unsupervised capture by parents or lay workers; agreement with pneumatic otoscopy was low; and fewer than 20% of videos were diagnostically useful [43]. However, structured instruction or simulation improved the performance of workers without hands-on training [38,39,43]. Example of practical structured training can remain minimal, such as a 5- to 10-minute tutorial, a 1-page capture checklist, and brief hands-on coaching including cerumen management. Embedding a simple feedback loop on image quality further improves yield without adding substantial workload. Third, adjunct digital tools are promising: a smartphone-only acoustic method that infers tympanic membrane mobility detected MEE with the AUC approaching 0.90, whereas digital otoscopy platforms improved trainee-supervisor agreement and reduced duplicate examinations [42,53]. The contribution of this review is an implementation-facing synthesis that links “which tool,” “which care model,” and “under what conditions” to observed diagnostic performance, thereby informing scalable hybrid workflows.

Interpretation and Implications

Diagnostic Performance and Care Models

When image acquisition was standardized and performed by trained facilitators, remote specialists and generalists demonstrated diagnostic agreement with in-person reference standards that are clinically feasible for triage and management in many settings [29,34,36,50]. Programs embedding store-and-forward video-otoscopy, tympanometry, and audiometry in school- and community-based screening have

XSL•FO
RenderX

demonstrated high coverage, efficient specialist sorting, reduced wait times, and identified surgical candidates [28,31,64]. In primary care, asynchronous tele-referral pathways allowed most cases to be managed without face-to-face review while offering timely specialist input and educational feedback to referrers [50].

Simultaneously, telehealth visits performed without otoscopic visualization carry a risk of misclassification. In a pandemic-era cohort, children evaluated for RAOM via video-only visits were less likely to have effusion confirmed at the time of tympanostomy tube placement than office-based comparators (39% vs 79%), highlighting that history alone is insufficient for surgical decision-making [52]. These findings support hybrid models—telehistory plus rapid access to digital otoscopy (clinic kiosk, drive-through nurse station, and community facilitators) or validated home solutions—especially when management depends on objective middle ear findings [52,54,68].

Operators, Training, and Human Factors

Evidence has consistently demonstrated that the person capturing the image is as critical as the device used. Across included studies, capture quality and diagnostic yield varied substantially by operator type, training intensity, child cooperation, and cerumen burden [12,39,43]. Hands-on coaching and structured capture approaches were associated with higher proportions of interpretable recordings in parents and nonspecialists, whereas written-only or brief instruction yielded lower rates of diagnostically useful media in field settings [12,39,43]. In simulation settings, trained facilitators using Otosim2 with the AudioProConnect platform maintained expert accuracy without time penalties [38]. In clinical supervision settings, digital otoscopy improved trainee accuracy by approximately 11 percentage points and increased trainee-supervisor agreement from kappa 0.40 to 0.69, while reducing confirmatory reexaminations from ~98% to ~27% [53]. A randomized comparison in pediatric emergency department demonstrated similar accuracy for smartphones versus conventional otoscopy (0.74 vs 0.69) but revealed lower confidence among inexperienced users, emphasizing the value of familiarization [37]. Smartphone otoscopy may meaningfully affect clinical impressions. In a pediatric emergency department study, CellScope Oto altered the reported view in 12% to 16% of exams, with clinically relevant diagnostic changes in approximately 6%; notably, it was rated by clinicians as easy to use and valuable for teaching [71].

Given the established effectiveness of video-otoscopy, careful consideration of the methods and timing of use may enable promising integration [72]. Standardized training packages and minimum quality thresholds should be considered prerequisites for scaling, particularly for parents, community health workers, and very young children, in whom cooperation and ear canal size may complicate capture [12,39,43].

Adjunct Technologies and Artificial Intelligence–Based Decision Support

Smartphone-only acoustic reflectometry with embedded machine learning demonstrated high accuracy for effusion detection and performed comparably across platforms and even in parents' hands [42]. This suggests a low-barrier screening pathway that does not depend on optical access, which is particularly valuable in toddlers with narrow ear canals or when video quality is suboptimal [42,43]. Early studies of automated image classification demonstrated throughput advantages, but clinical readiness remains limited due to scarce external validation across devices, sites, and pediatric populations and incomplete reporting of training datasets and labeling methods, which constrains generalizability and risk-of-bias assessment [15,16,47]. Although numerous models were developed using artificial intelligence to classify middle ear diseases on the basis of tympanic membrane findings or short-wave infrared imaging, there were few studies addressing its actual clinical implementation [15,16,73-76]. It is therefore necessary to consider not only the accuracy of these models but also the optimal point and method of integration within the clinical workflow [15,47]. Most studies also provide limited information on deployment pathways, including regulatory considerations, governance for performance monitoring, and safeguards when image quality is insufficient, and direct comparisons with human experts under identical capture conditions remain uncommon [15,47]. Accordingly, priority next steps include multisite, device-agnostic external validation using real-world pediatric data and prospective evaluations of workflow-level impact and safety rather than reliance on curated test-set metrics alone. In parallel, self-management apps for OME demonstrated high acceptability and reasonable correlation with clinic audiometry, supporting family engagement and between-visit monitoring rather than diagnosis [45].

Safety, Stewardship, and Workflow

Across randomized and comparative studies, smartphone otoscopy did not increase AOM diagnosis or antibiotic prescription in emergency settings compared to conventional otoscopy, and in some trainee contexts, it was associated with fewer prescriptions, likely due to better shared visualization and supervision [41,53]. Programs that integrate remote reviews into perioperative pathways have demonstrated high postoperative success rates and measurable cost savings for families in remote regions [32,54]. Collectively, these findings support protocolized tele-otoscopy within antibiotic stewardship approaches and access pathways that do not rely exclusively on family-owned smartphones, such as facilitated capture in schools, community programs, and clinics [28,29,34,61,64]. Additionally, remote hearing screening initiatives, when paired with follow-up during rehabilitation, may integrate synergistically with OM care, ultimately contributing to comprehensive pediatric ear health services [64,77-79]. In this review, reporting on governance and medicolegal arrangements was uncommon, representing a gap in implementation evidence. Implementation should include explicit consent, minimum image quality thresholds with default to in-person examination when unmet, and red-flag escalation [52,58,72]. Future studies should report on adverse events and medicolegal frameworks.

Equity and Access

Tele-otoscopy is particularly valuable in indigenous and rural settings with few specialists, where school-based or outreach models led by local health workers can achieve over 80%

XSL•FO
RenderX

screening coverage and facilitate timely surgical intervention [28]. Jacups et al [80] reported that the telehealth model was the most cost-effective method for providing ENT surgery to children living in remote regions, primarily by reducing patient travel. However, equity implications depend strongly on implementation design, including whether capture is facilitated and whether connectivity and device requirements are minimized [28,43,50]. Not all low-burden mHealth interventions translate into short-term clinical improvements; for example, multimedia or text messaging alone did not increase clinic attendance or improve ear outcomes over 6 weeks in a high-risk remote population, despite good acceptability [70]. Digital divide factors—device availability, connectivity and bandwidth requirements, caregiver digital and health literacy, and accessibility beyond geography (including language support and cultural adaptation)—were infrequently reported or measured systematically across included studies [43,50]. Age-related constraints are also salient in pediatrics, with lower capture quality reported in children under 24 months, which can shift burden onto caregivers when programs rely primarily on home capture [43]. Care model choice has direct equity consequences. In lower-middle-income countries, implementation models pairing screening with video-otoscopy and electronic medical record capture can broaden the scope beyond hearing thresholds alone by identifying otologic pathology, even among children with normal hearing, while enabling structured follow-up workflows [69]. Hub-based or facilitated capture workflows (community-hosted capture hubs or school programs) can reduce reliance on family-owned devices and high-bandwidth connections while enabling standardized training and explicit image quality thresholds [28,29,34]. In contrast, home-based models can improve convenience but may disproportionately disadvantage families facing barriers in time, literacy, language, connectivity, or device access unless hardware and support are provisioned [43,68].

Therefore, equity-focused deployment should prioritize community-hosted capture hubs, offline-capable apps, and hardware provision [28,43,50]. In pediatrics, careful consideration of caregivers' role is essential, as their involvement can influence capture quality and follow-through [12,65]. Importantly, few included studies conducted formal health equity analyses or reported outcomes stratified by socioeconomic status, geography, or ethnicity, limiting conclusions on differential benefits across populations and underscoring the need for equity endpoints and subgroup reporting in future evaluations [43,50].

Limitations

This scoping review emphasized breadth rather than quantitative synthesis and therefore did not include a meta-analysis; however, we conducted a targeted QUADAS-2 appraisal for diagnostic accuracy studies. Limitations included the enrollment of mixed-age cohorts, lack of blinded reference standards, and the rapid evolution of device ecosystems, which may reduce the generalizability of device-specific findings. Additionally, language and publication restrictions may have led to the exclusion of relevant gray literature. Although we searched multiple bibliographic databases and used both controlled vocabulary and free-text terms, relevant studies may have been missed due to terminology variation across disciplines and incomplete indexing of emerging digital health tools. Future studies should prioritize large, multicenter evaluations that incorporate standardized blinded reference standards to strengthen diagnostic accuracy across diverse pediatric populations and care settings. External validation efforts must address varied device ecosystems and real-world image quality to ensure generalizability. Comparative research is warranted to establish the relative performance and cost-effectiveness of emerging digital otoscopy, acoustic, and decision support tools. This scoping review protocol was not prospectively registered. Core methodological decisions were defined a priori; minor refinements to screening guidance were made during pilot screening to improve consistency. Although scoping reviews do not always require registration, the absence of prospective registration may reduce transparency and should be considered when interpreting the findings. In addition, interpretation of diagnostic accuracy is constrained by substantial clinical and methodological heterogeneity across studies, which limits cross-setting comparability of κ, sensitivity, and specificity. Equity-relevant reporting was uncommon, restricting conclusions on differential benefits across populations. Finally, while implementation outcomes were frequently described, most studies did not evaluate sustainability or scale-up using an explicit implementation framework, and evidence for long-term maintenance remains limited.

Future Research Agenda

Based on mapped gaps across technologies, care models, and operators, priority research directions include (1) multisite pragmatic evaluations with standardized reporting of pediatric age strata, OM phenotype, reference standards, operator training content and competency, and explicit image quality thresholds; (2) comparative evaluation of hub-based facilitated capture versus home-based capture models, including equity-relevant endpoints and subgroup analyses; (3) workflow-focused outcomes beyond accuracy; (4) sustainability and scale-up studies using an implementation framework, including maintenance, governance, and medicolegal arrangements; and (5) for artificial intelligence–enabled tools, external validation across devices or sites with real-world image quality, transparent dataset description, and prospective evaluation of clinical impact within the care pathway rather than standalone discrimination metrics.

Conclusions

Telehealth for pediatric OM now spans mature technologies and care models. Store-and-forward tele-otoscopy is feasible and, with standardized image capture and brief structured training, can approach in-person diagnostic performance. Implementation quality requires predefined image quality thresholds and a default to in-person care when those thresholds are not met. It also requires integration into hybrid workflows and basic governance covering consent, documentation, and data protection. Programs should ensure equitable access, especially in resource-limited settings. Video-only encounters without otoscopic imaging are inadequate when decisions require objective middle ear findings. Key evidence gaps include multicenter child-level outcomes, economic

XSL•FO
RenderX

evaluations—particularly in low- and middle-income countries—and external validation of artificial intelligence–based diagnostic tools across devices and sites. These priorities should guide the next phase of research and scale-up.

Acknowledgments

This work was undertaken as part of the activities of the Domestic Scientific Committee of the Japanese Society of Pediatric Otorhinolaryngology. We would like to thank Editage (www.editage.com) for English language editing. The authors are involved in a research project supported by the Japan Science and Technology Agency (JST) under the COI-NEXT (Collaborative Creation of New Social Systems) Program (Grant Number JPMJPF2403). However, this specific study did not receive direct financial support from that grant.

Disclosure of delegation to generative artificial intelligence (GenAI): The authors declare the use of GenAI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing and translation. The GenAI tool used was ChatGPT-5.2. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by Corresponding Author. In accordance with the GAIDeT taxonomy, GenAI tools were used solely for reformatting and limited language editing tasks under full human supervision.

Funding

This study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

All data underlying this scoping review were derived from published studies cited in the Reference list. The study-level data extraction table and the QUADAS-2 assessments for diagnostic accuracy studies are provided in the supplementary materials (Multimedia Appendix 2, Multimedia Appendix 3, and Multimedia Appendix 4. Additional materials (the data charting form/codebook and analytic files used to generate summary tables/figures) are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: MN, RA
Data curation: MN, RA
Formal analysis: MN, RA
Investigation: RA, MN, MH
Methodology: MN, RA, MH
Project administration: MN, RA
Supervision: HY, HF, YT, MI
Validation: RA, MN, MH
Visualization: MN, RA
Writing – original draft: MN, RA
Writing – review & editing: MN, RA, MH, CS, RK, HY, YM, HF, YT, MI

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full search strategies for this scoping review.
[DOCX File, 18 KB - jmir_v28i1e85416_app1.docx]

Multimedia Appendix 2
Study-level characteristics and data charting for included studies (N=52). This appendix presents the data extraction (data charting) table for all included studies in this scoping review. Variables include study design and setting, country, participant characteristics (pediatric population), otitis media phenotype, telehealth care model (asynchronous/synchronous/hybrid), device or technology category, operator or capture and interpretation workflow, training and image quality considerations, outcome domains reported, and funding or conflict of interest statements where available.
[XLSX File, 50 KB - jmir_v28i1e85416_app2.xlsx]

XSL•FO
RenderX

Multimedia Appendix 3
RE-AIM mapping of implementation-related outcomes and reporting gaps. This appendix summarizes implementation-relevant constructs using the Reach, Effectiveness, Adoption, Implementation, and Maintenance (RE-AIM) framework across included studies. The table highlights commonly reported elements and underreported domains to inform future research and reporting in telehealth-supported pediatric otitis media care.
[XLSX File, 10 KB - jmir_v28i1e85416_app3.xlsx]

Multimedia Appendix 4
Revised Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2). This appendix reports the QUADAS-2 risk of bias and applicability assessments for the subset of included studies that evaluated diagnostic accuracy. Judgments are provided across QUADAS-2 domains (patient selection, index test, reference standard, flow and timing) and applicability domains (patient selection, index test, reference standard), with signaling questions and domain-level ratings.
[XLSX File, 12 KB - jmir_v28i1e85416_app4.xlsx]

Checklist 1
PRISMA-ScR checklist.
[PDF File, 134 KB - jmir_v28i1e85416_app5.pdf]

References

1. Bajwa NM, Perron NJ, Braillard O, et al. Has telemedicine come to fruition? Parents' and pediatricians' perceptions and preferences regarding telemedicine. Pediatr Res 2024 Oct;96(5):1332-1339. [doi: 10.1038/s41390-024-03172-w] [Medline: 38555380]
2. Howland K, Edvardsson K, Lees H, Hooker L. Telehealth use in the well-child health setting. A systematic review of acceptability and effectiveness for families and practitioners. Int J Nurs Stud Adv 2025 Jun;8:100277. [doi: 10.1016/j.ijnsa.2024.100277] [Medline: 39734584]
3. Luke MJ, Craig S, Pak-Gorstein S, et al. Narrowing the digital divide: framework for creating telehealth equity dashboards. Interact J Med Res 2024 Sep 4;13:e57435. [doi: 10.2196/57435] [Medline: 39231423]
4. Marom T, Pitaro J, Shah UK, et al. Otitis media practice during the COVID-19 pandemic. Front Cell Infect Microbiol 2021;11:749911. [doi: 10.3389/fcimb.2021.749911] [Medline: 35071032]
5. Metcalfe C, Muzaffar J, Orr L, Coulson C. A systematic review of remote otological assessment using video-otoscopy over the past 10 years: reliability and applications. Eur Arch Otorhinolaryngol 2021 Dec;278(12):4733-4741. [doi: 10.1007/s00405-020-06596-2] [Medline: 33486567]
6. Wang H, Zeng X, Miao X, et al. Global, regional, and national epidemiology of otitis media in children from 1990 to 2021. Front Pediatr 2025 Jul 2;13:1513629. [doi: 10.3389/fped.2025.1513629]
7. Schilder AGM, Chonmaitree T, Cripps AW, et al. Otitis media. Nat Rev Dis Primers 2016 Sep 8;2(1):16063. [doi: 10.1038/nrdp.2016.63] [Medline: 27604644]
8. Csonka P, Palmu S, Heikkilä P, Huhtala H, Korppi M. Outpatient antibiotic prescribing for 357,390 children with otitis media. Pediatr Infect Dis J 2022 Dec 1;41(12):947-952. [doi: 10.1097/INF.0000000000003693] [Medline: 36102742]
9. Wittman SR, Hoberman A, Mehrotra A, Sabik LM, Yabes JG, Ray KN. Antibiotic receipt for pediatric telemedicine visits with primary care vs direct-to-consumer vendors. JAMA Netw Open 2024 Mar 4;7(3):e242359. [doi: 10.1001/jamanetworkopen.2024.2359] [Medline: 38483387]
10. Bhatia R, Chauhan A, Rana M, Kaur K, Pradhan P, Singh M. Economic burden of otitis media globally and an overview of the current scenario to alleviate the disease burden: a systematic review. Int Arch Otorhinolaryngol 2024 Jul;28(3):e552-e558. [doi: 10.1055/s-0043-1767802] [Medline: 38974618]
11. El Feghaly RE, Nedved A, Katz SE, Frost HM. New insights into the treatment of acute otitis media. Expert Rev Anti Infect Ther 2023 May;21(5):523-534. [doi: 10.1080/14787210.2023.2206565] [Medline: 37097281]
12. Erkkola-Anttinen N, Irjala H, Laine MK, Tähtinen PA, Löyttyniemi E, Ruohola A. Smartphone otoscopy performed by parents. Telemed J E Health 2019 Jun;25(6):477-484. [doi: 10.1089/tmj.2018.0062] [Medline: 30040525]
13. Wagner R, Lima TC, Silva MD, et al. Assessment of pediatric telemedicine using remote physical examinations with a mobile medical device: a nonrandomized controlled trial. JAMA Netw Open 2023 Feb 1;6(2):e2252570. [doi: 10.1001/jamanetworkopen.2022.52570] [Medline: 36729459]
14. Alenezi EM, Veselinović T, Tao KF, et al. Ear Portal: an urban-based ear, nose, and throat, and audiology referral telehealth portal to improve access to specialist ear health services for children. J Telemed Telecare 2024 Dec;30(10):1581-1589. [doi: 10.1177/1357633X231158839] [Medline: 36916156]
15. Dubois C, Eigen D, Simon F, et al. Development and validation of a smartphone-based deep-learning-enabled system to detect middle-ear conditions in otoscopic images. NPJ Digit Med 2024 Jun 20;7(1):162. [doi: 10.1038/s41746-024-01159-9] [Medline: 38902477]

XSL•FO
RenderX

16. Jin X, Fan M. EarMonitor: non-clinical assessment of ear health conditions using a low-cost endoscope camera on smartphones. Proc ACM Hum-Comput Interact 2024 Sep 24;8(MHCI):1-20. [doi: 10.1145/3676499]
17. Appelberg E, Viitasalo S, Sinkkonen ST. Patient self-recorded digital otoscopy in otology diagnostics. Laryngoscope Investig Otolaryngol 2025 Dec;10(6):e70318. [doi: 10.1002/lio2.70318] [Medline: 41347050]
18. Haynes SC, Marcin JP. Pediatric telemedicine: lessons learned during the coronavirus disease 2019 pandemic and opportunities for growth. Adv Pediatr 2022 Aug;69(1):1-11. [doi: 10.1016/j.yapd.2022.04.002] [Medline: 35985702]
19. Curfman A, McSwain SD, Chuo J, et al. Pediatric telehealth in the COVID-19 pandemic era and beyond. Pediatrics 2021 Sep;148(3):e2020047795. [doi: 10.1542/peds.2020-047795] [Medline: 34215677]
20. Moberly AC, Zhang M, Yu L, et al. Digital otoscopy versus microscopy: how correct and confident are ear experts in their diagnoses? J Telemed Telecare 2018 Aug;24(7):453-459. [doi: 10.1177/1357633X17708531] [Medline: 28480781]
21. Damery L, Lescanne E, Reffet K, Aussedat C, Bakhos D. Interest of video-otoscopy for the general practitioner. Eur Ann Otorhinolaryngol Head Neck Dis 2019 Feb;136(1):13-17. [doi: 10.1016/j.anorl.2018.10.016] [Medline: 30392875]
22. Quick ME, Brennan-Jones CG, Kuthubutheen J. Remote paediatric ear examination comparing video-otoscopy and still otoscopy clinician rated outcomes. Int J Pediatr Otorhinolaryngol 2024 Feb;177:111871. [doi: 10.1016/j.ijporl.2024.111871] [Medline: 38266378]
23. Habib AR, Crossland G, Sacks R, Singh N, Patel H. Tele-otology for Aboriginal and Torres Strait Islander people living in rural and remote areas. Laryngoscope 2024 Dec;134(12):5096-5102. [doi: 10.1002/lary.31624] [Medline: 38982868]
24. Pannonhalmi Á, Posta B, Perényi Á, et al. Clinical validation of a video-otoscopy-based medical device for the remote diagnosis of ear complaints. Sensors (Basel) 2025 Jan 27;25(3):758. [doi: 10.3390/s25030758] [Medline: 39943397]
25. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
26. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
27. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
28. Smith AC, Armfield NR, Wu WI, Brown CA, Perry C. A mobile telemedicine-enabled ear screening service for Indigenous children in Queensland: activity and outcomes in the first three years. J Telemed Telecare 2012 Dec;18(8):485-489. [doi: 10.1258/jtt.2012.gth114] [Medline: 23209267]
29. Biagio L, Swanepoel DW, Laurent C, Lundberg T. Video-otoscopy recordings for diagnosis of childhood ear disease using telehealth at primary health care level. J Telemed Telecare 2014 Sep;20(6):300-306. [doi: 10.1177/1357633X14541038] [Medline: 24958356]
30. Wu CJ, Wu SY, Chen PC, Lin YS. An innovative smartphone-based otorhinoendoscope and its application in mobile health and teleotolaryngology. J Med Internet Res 2014 Mar 3;16(3):e71. [doi: 10.2196/jmir.2959] [Medline: 24590187]
31. Smith AC, Brown C, Bradford N, Caffery LJ, Perry C, Armfield NR. Monitoring ear health through a telemedicine-supported health screening service in Queensland. J Telemed Telecare 2015 Dec;21(8):427-430. [doi: 10.1177/1357633X15605407] [Medline: 26377121]
32. Jacups SP, Newman D, Dean D, Richards A, McConnon KM. An innovative approach to improve ear, nose and throat surgical access for remote living Cape York Indigenous children. Int J Pediatr Otorhinolaryngol 2017 Sep;100:225-231. [doi: 10.1016/j.ijporl.2017.07.011] [Medline: 28802377]
33. Prins-van Ginkel AC, de Hoog ML, Uiterwaal C, Smit HA, Bruijning-Verhagen PC. Detecting acute otitis media symptom episodes using a mobile app: cohort study. JMIR Mhealth Uhealth 2017 Nov 28;5(11):e181. [doi: 10.2196/mhealth.7505] [Medline: 29183869]
34. Lundberg T, Biagio de Jager L, Swanepoel DW, Laurent C. Diagnostic accuracy of a general practitioner with video-otoscopy collected by a health care facilitator compared to traditional otoscopy. Int J Pediatr Otorhinolaryngol 2017 Aug;99:49-53. [doi: 10.1016/j.ijporl.2017.04.045] [Medline: 28688565]
35. Ramkumar V, Rajendran A, Nagarajan R, Balasubramaniyan S, Suresh DK. Identification and management of middle ear disorders in a rural cleft care program: a telemedicine approach. Am J Audiol 2018 Nov 19;27(3S):455-461. [doi: 10.1044/2018_AJA-IMIA3-18-0015] [Medline: 30452749]
36. Mandavia R, Lapa T, Smith M, Bhutta MF. A cross-sectional evaluation of the validity of a smartphone otoscopy device in screening for ear disease in Nepal. Clin Otolaryngol 2018 Feb;43(1):31-38. [doi: 10.1111/coa.12898] [Medline: 28485038]
37. Mousseau S, Lapointe A, Gravel J. Diagnosing acute otitis media using a smartphone otoscope; a randomized controlled trial. Am J Emerg Med 2018 Oct;36(10):1796-1801. [doi: 10.1016/j.ajem.2018.01.093] [Medline: 29544905]
38. Venail F, Akkari M, Merklen F, et al. Evaluation of otoscopy simulation as a training tool for real-time remote otoscopy. Int J Audiol 2018 Mar;57(3):194-200. [doi: 10.1080/14992027.2017.1416190] [Medline: 29256826]
39. Shah MU, Sohal M, Valdez TA, Grindle CR. iPhone otoscopes: currently available, but reliable for tele-otoscopy in the hands of parents? Int J Pediatr Otorhinolaryngol 2018 Mar;106:59-63. [doi: 10.1016/j.ijporl.2018.01.003] [Medline: 29447893]

XSL•FO
RenderX

40. Bhavana K, Ahmad M, Sharma P. Smartphone otoscopy sans attachment: a paradigm shift in diagnosing ear pathologies. OTO Open 2018;2(3):2473974X18786496. [doi: 10.1177/2473974X18786496] [Medline: 31565683]
41. Chan KN, Silverstein A, Bryan LN, McCracken CE, Little WK, Shane AL. Comparison of a smartphone otoscope and conventional otoscope in the diagnosis and management of acute otitis media. Clin Pediatr (Phila) 2019 Mar;58(3):302-306. [doi: 10.1177/0009922818812480] [Medline: 30463434]
42. Chan J, Raju S, Nandakumar R, Bly R, Gollakota S. Detecting middle ear fluid using smartphones. Sci Transl Med 2019 May 15;11(492):eaav1102. [doi: 10.1126/scitranslmed.aav1102] [Medline: 31092691]
43. Demant MN, Jensen RG, Bhutta MF, Laier GH, Lous J, Homøe P. Smartphone otoscopy by non-specialist health workers in rural Greenland: a cross-sectional study. Int J Pediatr Otorhinolaryngol 2019 Nov;126:109628. [doi: 10.1016/j.ijporl.2019.109628] [Medline: 31445480]
44. Hakimi AA, Lalehzarian AS, Lalehzarian SP, Azhdam AM, Nedjat-Haiem S, Boodaie BD. Utility of a smartphone-enabled otoscope in the instruction of otoscopy and middle ear anatomy. Eur Arch Otorhinolaryngol 2019 Oct;276(10):2953-2956. [doi: 10.1007/s00405-019-05559-6] [Medline: 31317322]
45. Fordington S, Brown TH. An evaluation of the Hear Glue Ear mobile application for children aged 2-8 years old with otitis media with effusion. Digit Health 2020;6:2055207620966163. [doi: 10.1177/2055207620966163] [Medline: 33194218]
46. Gupta N, Chawla N, Gupta D, Dhawan N, Janaki VR. Community triage of otology patients using a store-and-forward telemedicine device: a feasibility study. Ear Nose Throat J 2020 Mar 22:145561320908953. [doi: 10.1177/0145561320908953] [Medline: 32200652]
47. Cha D, Shin SH, Kim J, et al. Feasibility of asynchronous and automated telemedicine in otolaryngology: prospective cross-sectional study. JMIR Med Inform 2020 Oct 19;8(10):e23680. [doi: 10.2196/23680] [Medline: 33027033]
48. Durgut O, Ekim B, Dikici O, Solmaz F, Ağırgöl B, Özbakan A. Evaluation of hearing thresholds by using a mobile application in children with otitis media with effusion. Audiol Neurootol 2020;25(3):120-124. [doi: 10.1159/000505309] [Medline: 31962328]
49. Jayawardena ADL, Mankarious LA, Keamy DG Jr, Cohen MS. Pediatric, family-centered, “at-home” otologic physical examination in the COVID-19 era. Otolaryngol Head Neck Surg 2020 Nov;163(5):1061-1063. [doi: 10.1177/0194599820934776] [Medline: 32513060]
50. Cottrell E, George A, Coulson C, Chambers R. Telescopic otology referrals: evaluation of feasibility and acceptability. Laryngoscope Investig Otolaryngol 2020 Apr;5(2):221-227. [doi: 10.1002/lio2.367] [Medline: 32337353]
51. Schuster-Bruce JR, Ali A, Van M, Rogel-Salazar J, Ofo E, Shamil E. A randomised trial to assess the educational benefit of a smartphone otoscope in undergraduate medical training. Eur Arch Otorhinolaryngol 2021 Jun;278(6):1799-1804. [doi: 10.1007/s00405-020-06373-1] [Medline: 32960350]
52. Kolb CM, Born K, Banker K, Barth P, Aaronson NL. Comparing telehealth with office-based visits for common pediatric otolaryngology complaints. Int J Pediatr Otorhinolaryngol 2021 Jun;145:110712. [doi: 10.1016/j.ijporl.2021.110712] [Medline: 33887549]
53. Kleinman K, Psoter KJ, Nyhan A, Solomon BS, Kim JM, Canares T. Evaluation of digital otoscopy in pediatric patients: a prospective randomized controlled clinical trial. Am J Emerg Med 2021 Aug;46:150-155. [doi: 10.1016/j.ajem.2021.04.030] [Medline: 33945977]
54. Don DM, Koempel JA, Fisher LM, Wee CP, Osterbauer B. Prospective evaluation of a smartphone otoscope for home tympanostomy tube surveillance: a pilot study. Ann Otol Rhinol Laryngol 2021 Feb;130(2):125-132. [doi: 10.1177/0003489420938820] [Medline: 32627572]
55. Manayan RC, Ladd-Luthringshauser OH, Packer A, et al. Ambient noise limits efficacy of smartphone-based screening for hearing loss in children at risk. Am J Otolaryngol 2022;43(1):103223. [doi: 10.1016/j.amjoto.2021.103223] [Medline: 34560596]
56. Meng X, Dai Z, Wang Y, et al. Application of smartphone otoscope in telemedicine in rural medical consortium in eastern China in the COVID-19 era. Front Public Health 2022;10:879410. [doi: 10.3389/fpubh.2022.879410] [Medline: 35664106]
57. Deshmukh BG, Bhisegaonkar DA, Phatak P, et al. Accuracy of the novel ENTraview mobile device to increase access to screening and provisional diagnosis of ear disease in underserved communities across India. Epidem Int 2022 Mar 27;07(1):21-24. [doi: 10.24321/2455.7048.202207]
58. Schafer A, Mousset M, Kelly N, Althubaiti A, Bourgeois T, Elmaraghy CA. Assessing the use of telehealth for the surgical management of recurrent otitis media. Int J Pediatr Otorhinolaryngol 2022 Feb;153:111036. [doi: 10.1016/j.ijporl.2021.111036] [Medline: 34998205]
59. Smola C, Shah N, Fowler A, Gubbels A, Monroe K. Assessment of pediatric residents’ comfort level with CellScope Oto to examine pediatric ear exams. Clin Pediatr (Phila) 2022 Jul;61(7):485-489. [doi: 10.1177/00099228221086374] [Medline: 35446181]
60. Chan J, Najafi A, Baker M, et al. Performing tympanometry using smartphones. Commun Med (Lond) 2022;2(1):57. [doi: 10.1038/s43856-022-00120-9] [Medline: 35721828]
61. Habib AR, Perry C, Crossland G, et al. Inter-rater agreement between 13 otolaryngologists to diagnose otitis media in Aboriginal and Torres Strait Islander children using a telehealth approach. Int J Pediatr Otorhinolaryngol 2023 May;168:111494. [doi: 10.1016/j.ijporl.2023.111494] [Medline: 37003013]

XSL•FO
RenderX

62. Patel A, Forde C, Doal S, et al. A new community-to-hospital specialist otology pathway using telemedicine: a pilot study of 50 patients. Clin Otolaryngol 2024 Sep;49(5):652-659. [doi: 10.1111/coa.14192] [Medline: 38899484]
63. Gyawali BR, Obaidullah, Ghimire S, et al. Reliability of a tailored tele-medicine model for otorhinolaryngology-head and neck surgery (ORL-HNS) clinical practice in remote areas of Nepal: a pilot study from tertiary referral center. Ear Nose Throat J 2024 Oct 15:1455613241291168. [doi: 10.1177/01455613241291168] [Medline: 39405038]
64. Bowers P, Graydon K, Tomlin D, Rance G. The reliability of remote diagnostic hearing assessment in school-entry-aged children. Int J Audiol 2025 Mar;64(3):261-270. [doi: 10.1080/14992027.2024.2333392] [Medline: 38551077]
65. Altamimi AA, Brennan-Jones CG, Robinson M, et al. A hospital-based asynchronous ENT telehealth service for children with otitis media: cost-minimisation and improved access. J Telemed Telecare 2025 Jul;31(6):818-825. [doi: 10.1177/1357633X231223994] [Medline: 38295365]
66. Rappaport KM, McCracken CC, Beniflah J, et al. Assessment of a smartphone otoscope device for the diagnosis and management of otitis media. Clin Pediatr (Phila) 2016 Aug;55(9):800-810. [doi: 10.1177/0009922815593909] [Medline: 26156976]
67. Ni G, Curtis S, Kaplon A, et al. Development of video otoscopy quiz using a smartphone adaptable otoscope. J Otol 2021 Apr;16(2):80-84. [doi: 10.1016/j.joto.2020.08.001] [Medline: 33777119]
68. Stancel-Lewis J, Joseph K, Laister S, et al. Parent-led ear health checks at home for children with complex needs pilot-project: a new approach to improving access to ear health. Int J Audiol 2025 Aug;64(8):803-812. [doi: 10.1080/14992027.2025.2456021] [Medline: 39962803]
69. Yancey KL, Cheromei LJ, Muhando J, Reppart J, Netterville JL, Jayawardena ADL. Pediatric hearing screening in low-resource settings: incorporation of video-otoscopy and an electronic medical record. Int J Pediatr Otorhinolaryngol 2019 Nov;126:109633. [doi: 10.1016/j.ijporl.2019.109633] [Medline: 31421355]
70. Phillips JH, Wigger C, Beissbarth J, McCallum GB, Leach A, Morris PS. Can mobile phone multimedia messages and text messages improve clinic attendance for Aboriginal children with chronic otitis media? A randomised controlled trial. J Paediatr Child Health 2014 May;50(5):362-367. [doi: 10.1111/jpc.12496] [Medline: 24612007]
71. Richards JR, Gaylor KA, Pilgrim AJ. Comparison of traditional otoscope to iPhone otoscope in the pediatric ED. Am J Emerg Med 2015 Aug;33(8):1089-1092. [doi: 10.1016/j.ajem.2015.04.063] [Medline: 25979304]
72. Cortés Fuentes IA, Zelada B Ú, Zitko M P. Smartphone-based video-otoscopy performed by non-ENT physicians as a complementary tool to conventional otoscopy for tympanic membrane diagnosis. Speech Lang Hear 2024 Jan 2;27(1):14-19. [doi: 10.1080/2050571X.2023.2209766]
73. Kashani RG, Młyńczak MC, Zarabanda D, et al. Shortwave infrared otoscopy for diagnosis of middle ear effusions: a machine-learning-based approach. Sci Rep 2021 Jun 15;11(1):12509. [doi: 10.1038/s41598-021-91736-9] [Medline: 34131163]
74. Tran TT, Fang TY, Pham VT, Lin C, Wang PC, Lo MT. Development of an automatic diagnostic algorithm for pediatric otitis media. Otol Neurotol 2018 Sep;39(8):1060-1065. [doi: 10.1097/MAO.0000000000001897] [Medline: 30015751]
75. Khan MA, Kwon S, Choo J, et al. Automatic detection of tympanic membrane and middle ear infection from oto-endoscopic images via convolutional neural networks. Neural Netw 2020 Jun;126:384-394. [doi: 10.1016/j.neunet.2020.03.023] [Medline: 32311656]
76. Noda M, Yoshimura H, Okubo T, et al. Feasibility of multimodal artificial intelligence using GPT-4 vision for the classification of middle ear disease: qualitative study and validation. JMIR AI 2024 May 31;3:e58342. [doi: 10.2196/58342] [Medline: 38875669]
77. Dharmar M, Simon A, Sadorra C, et al. Reducing loss to follow-up with tele-audiology diagnostic evaluations. Telemed J E Health 2016 Feb;22(2):159-164. [doi: 10.1089/tmj.2015.0001] [Medline: 26544032]
78. Landolfi E, Continisio GI, Del Vecchio V, et al. NeonaTal Assisted TelerehAbilitation (T.A.T.A. web app) for hearing-impaired children: a family-centered care model for early intervention in congenital hearing loss. Audiol Res 2022 Mar 28;12(2):182-190. [doi: 10.3390/audiolres12020021] [Medline: 35447741]
79. Ramkumar V, Joshi B D, Prabhakar A, Hall JW, Vaidyanath R. Development and beta validation of an mHealth-based hearing screener (SRESHT) for young children in resource-limited countries: pilot validation study. JMIR Form Res 2025 Jan 13;9:e53460. [doi: 10.2196/53460] [Medline: 39805111]
80. Jacups SP, Kinchin I, McConnon KM. Ear, nose, and throat surgical access for remote living Indigenous children: what is the least costly model? J Eval Clin Pract 2018 Dec;24(6):1330-1338. [doi: 10.1111/jep.13044] [Medline: 30311714]

Abbreviations

AOM: acute otitis media
AUC: area under the receiver operating characteristic curve
CSOM: chronic suppurative otitis media
DenseNet: Dense Convolutional Network
ENT: ear, nose, and throat
JBI: Joanna Briggs Institute

XSL•FO
RenderX

MEE: middle ear effusion
mHealth: mobile health
OM: otitis media
OME: otitis media with effusion
PCC: population, concept, and context
PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
QUADAS-2: revised Quality Assessment of Diagnostic Accuracy Studies tool
RAOM: recurrent acute otitis media
RE-AIM: reach, effectiveness, adoption, implementation, and maintenance
ROC: receiver operating characteristic

Edited by S Brini; submitted 07.Oct.2025; peer-reviewed by D Singh, M Mansoor; revised version received 14.Feb.2026; accepted 23.Feb.2026; published 18.Mar.2026.

Please cite as:
Noda M, Akiyoshi R, Hosoya M, Shinkawa C, Koshu R, Yoshimura H, Monden Y, Fushiki H, Tanaka Y, Ito M
Telehealth Approaches for Pediatric Otitis Media and Clinical Outcomes: Scoping Review
J Med Internet Res 2026;28:e85416
URL: https://www.jmir.org/2026/1/e85416
doi:10.2196/85416

© Masao Noda, Ryohei Akiyoshi, Makoto Hosoya, Chikako Shinkawa, Ryota Koshu, Hidekane Yoshimura, Yukifumi Monden, Hiroaki Fushiki, Yasuhiro Tanaka, Makoto Ito. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Performance of Artificial Intelligence in Classifying Molecular Markers in Adult-Type Gliomas Using Histopathological Images: Systematic Review

Obada Almaabreh[1], MD; Rukaya Al-Dafi[1], MD; Aliya Tabassum[2], PhD; Ahmad Othman[1], MD; Alaa Abd-alrazaq[3,4], PhD

[1]Faculty of Medicine, Yarmouk University, Irbid, Jordan

[2]Department of Computer Science and Engineering, College of Engineering, Qatar University, Doha, Qatar

[3]AI Center for Precision Health, Weill Cornell Medical College in Qatar, 103 Ezdan Oasis, Alwakrah, Doha, Qatar

[4]Department of Biomedical Sciences, College of Health Sciences, Qatar University, Doha, Qatar

Corresponding Author:
Alaa Abd-alrazaq, PhD
AI Center for Precision Health, Weill Cornell Medical College in Qatar, 103 Ezdan Oasis, Alwakrah, Doha, Qatar

Abstract

Background: Adult-type gliomas are among the most prevalent and lethal primary central nervous system tumors, where prompt and accurate diagnosis is essential for maximizing survival prospects. Molecular classification, particularly the detection of isocitrate dehydrogenase (IDH) mutations and 1p/19q codeletions, has become crucial for accurate diagnosis and prognosis. Artificial intelligence (AI) has emerged as a promising adjunct in enhancing diagnostic accuracy using histopathological images. Existing reviews mostly focused on radiology rather than histopathology, and no comprehensive systematic review has specifically evaluated AI performance exclusively from histopathological images for detecting these two molecular markers.

Objective: This study aims to systematically evaluate the performance of AI models in detecting and classifying IDH mutation status and 1p/19q gene codeletion in adult-type gliomas using histopathological images.

Methods: A systematic review was conducted in accordance with PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Extension for Diagnostic Test Accuracy) guidelines. Seven databases (MEDLINE, PsycINFO, Embase, IEEE Xplore, ACM Digital Library, Scopus, and Google Scholar) were searched for studies published between 2015 and 2025. Eligible studies used AI models on histopathological images for molecular classification of adult-type gliomas and reported performance metrics. Study selection, data extraction, and risk of bias assessment using a modified QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) tool were conducted independently by two reviewers. Extracted data were synthesized narratively.

Results: A total of 2453 reports were identified, with 22 studies meeting the inclusion criteria. The pooled average accuracy, sensitivity, specificity, and area under the curve (AUC) across studies were 85.46%, 84.55%, 86.03%, and 86.53%, respectively. Hybrid models demonstrated the highest diagnostic performance (accuracy 92.80% and sensitivity 89.62%). In general, AI models that used multimodal data outperformed those that used unimodal data in terms of sensitivity (90.15% vs 84.31%) and AUC (88.93% vs 86.29%). Furthermore, models had a better overall performance in identifying IDH mutations than 1p/19q codeletions, with higher accuracy (86.13% vs 81.63%), specificity (86.61% vs 78.11%), and AUC (86.74% vs 85.15%). Unexpectedly, AI models designed for binary classification exhibited lower performance than those for multiclass classification in terms of both accuracy (91.98% vs 84.02%) and sensitivity (93.41% vs 80.18%). However, these differences should be interpreted as descriptive trends rather than statistically validated superiority, as formal between-group comparisons were not feasible.

Conclusions: AI models show strong potential as complementary tools for the molecular classification of adult-type gliomas using histopathology images, particularly for IDH mutation detection. However, these findings are constrained by the limited number of studies, the focus on adult-type gliomas, lack of meta-analysis, and restriction to English-language publications. While AI offers valuable diagnostic support, it must be integrated with expert clinical judgment. Future research should prioritize larger, more diverse datasets and multimodal AI frameworks and extend to other brain tumor types for broader applicability.

Trial Registration: PROSPERO CRD420250653668; https://www.crd.york.ac.uk/PROSPERO/view/CRD420250653668

(J Med Internet Res 2026;28:e78377) doi:10.2196/78377

XSL•FO
RenderX

KEYWORDS

glioma; brain tumors; molecular markers; artificial intelligence; histopathology; isocitrate dehydrogenase; IDH mutation; 1p/19q codeletion; systematic review

Introduction

Background

Gliomas are the most frequent primary tumors affecting the central nervous system (CNS) in adults, with a variable prognosis depending on the specific subtype and histological grade [1]. The global prevalence of CNS neoplasms in 5 years is approximately 771,110 cases. Primary brain tumors account for 1.7% of all cancers, with a global incidence of 3.9 cases per 100,000 people each year. Among these, gliomas are the most common malignant type, accounting for 75% of all malignant CNS tumors, with an incidence of 6 cases per 100,000 people annually [2]. Gliomas are among the tumors that are difficult to treat, with a 5-year overall survival of less than 35% [3]. Patients with brain tumors present with variable clinical symptoms according to the part of the brain affected. However, they usually share general symptoms that are nonspecific to anatomic location (eg, seizures, headaches).

Gliomas are neuroectodermal tumors arising from glial cells or their precursors and encompass astrocytomas, oligodendrogliomas, and ependymomas [3]. Mutations in isocitrate dehydrogenase 1 and 2 (IDH1 and IDH2) are considered early events in gliomagenesis and occur more frequently in lower-grade gliomas [3]. Diffuse gliomas harboring IDH1/2 mutations are associated with significantly improved prognosis compared with IDH–wild-type diffuse gliomas [3].

Codeletion of chromosomal arms 1p and 19q results from an unbalanced centromeric translocation, t(1;19)(q10;p10) [3]. In combination with IDH mutation, 1p/19q codeletion is a defining molecular feature required for the diagnosis of oligodendroglioma, IDH mutant, and 1p/19q codeleted. This molecular alteration is associated with a favorable prognosis among diffuse gliomas and predicts enhanced responsiveness to alkylating chemotherapy [3].

Conventional classification of gliomas relied on histologic features as a gold standard. Initially, pathologists utilized microscopic analysis of histochemically stained (eg, hematoxylin and eosin) sections for the diagnosis of gliomas [4]. However, ancillary immunohistochemistry has been increasingly used over the past few decades to enhance diagnostic accuracy [4]. With either method, pathologists followed a diagnostic algorithm to rule out nonneoplastic lesions (eg, infarcts) or other types of malignancies (eg, metastatic) based on histologic features. Clinical information (eg, age) and radiologic findings (eg, location) are also valuable clues in the diagnosis [4]. Nonetheless, the histologic-based classification system has a few limitations, particularly interobserver variability, in addition to insufficient or nonrepresentative tissue sampling [5]. Although histologic-based classification can be accurate for prototypic tumors, classification of tumors encountered in clinical practice is often more challenging due to a degree of mixed features in the same sample [5].

In order to address these limitations, molecular-based diagnosis has emerged as a possible solution. Recent advancements in genetics have helped to identify specific gene alterations involved in the development of different types of gliomas. Detection of these gene alterations is the base of molecular-based diagnosis. This method offers improved accuracy and made it possible to distinguish between primary and secondary gliomas, which was not possible with the previous histologic-based method [6]. Therefore, the World Health Organization (WHO) integrated molecular features into its most recent classification of CNS tumors, published in 2021, based on the presence or absence of IDH1 and IDH2 mutations, in addition to chromosome 1p/19q codeletion [7]. It categorized adult-type gliomas into three categories: astrocytoma with IDH mutation, oligodendroglioma with IDH mutation and 1p/19q codeletion, and glioblastoma with IDH wild-type [7]. Furthermore, a combined approach using molecular and histologic features is used for the grading of IDH-mutant astrocytomas as CNS WHO grade 2, 3, or 4. The new WHO classification also led to a more accurate prognosis of gliomas, as the absence of IDH mutation is often a sign of more aggressive tumors [7].

Neuroimaging provides noninvasive insight into the molecular landscape of diffuse gliomas. Advanced magnetic resonance imaging (MRI) features can suggest IDH mutation status, as IDH-mutant tumors typically exhibit lower perfusion and less aggressive radiological behavior than IDH–wild-type gliomas [5-7]. Magnetic resonance spectroscopy allows in vivo detection of the IDH-specific oncometabolite D-2-hydroxyglutarate, enabling preoperative identification of IDH-mutant tumors [7]. In addition, oligodendrogliomas defined by combined IDH mutation and 1p/19q codeletion display characteristic imaging features, including cortical localization, calcifications, and relatively increased perfusion, reflecting distinct tumor biology [4,7]. These radiogenomic associations support the integration of imaging into molecular stratification of adult-type diffuse gliomas [4-6].

As gliomas pose diagnostic challenges due to their complex histological and molecular features, artificial intelligence (AI) has emerged as a promising tool to enhance diagnostic accuracy and efficiency. AI models were able to offer rapid and accurate results with accuracy exceeding 90%, which can further improve the access and speed of molecular diagnosis [8]. In general, AI implementation in health care can lead to improved data synthesis (eg, patient data, medical literature) [9]. Additionally, AI can process multimodal data (two or more different data sources such as histopathological images, radiological data, clinical variables, genomic markers, and other relevant data types) to provide more accurate diagnosis and predictions. Moreover, it offers means for augmenting human performance, as human limitations can prevent processing large quantities of information and making the optimal judgment at all times. Thus, AI can improve care consistency, increase precision, accelerate discoveries, and minimize disparities [9].

XSL•FO
RenderX

Research Gaps and Aim

The use of AI in the diagnosis of brain tumors and gliomas has been widely studied, with many reviews summarizing the results of numerous studies. However, existing reviews exhibit several limitations. For instance, many reviews focused on AI implementation in other diagnostic methods of gliomas, such as radiology-based [10,11], surgery-based [12], and DNA-based diagnoses [13]. As for reviews focusing on AI implementation in the molecular diagnosis of gliomas, they fall into one of the following categories: not systematic [14], outdated [15], restricted by databases or search terms [11,16], lacking subtyping of gliomas [17], not focused on genetic markers (IDH mutations and 1p/19q codeletion) [11], or limited to a specific subfield of AI (eg, deep learning or machine learning) [16]. To address the limitations of the previous reviews, this study aims to systematically assess the performance of AI in detecting and classifying IDH mutation status and 1p/19q gene codeletion in adult-type gliomas using histopathological images.

Methods

This review was conducted in accordance with the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Extension for Diagnostic Test Accuracy) [18] (see Checklist 1) and registered with the international Prospective Register of Systematic Reviews (PROSPERO) under the number CRD420250653668.

Search Strategy

The literature search was performed on November 15, 2024, using the following digital databases: MEDLINE, PsycINFO, Embase, IEEE Xplore, ACM Digital Library, Scopus, and Google Scholar. A biweekly automated search was set for 4 months, concluding on March 15, 2025. Given the large number of results from Google Scholar, which are ranked by relevance, we reviewed only the first 100 entries (10 pages). Our query utilized a combination of proper keywords and Boolean operators, including but not limited to ((“Artificial Intelligence” OR “Machine Learning” OR “Deep Learning”) AND (“Histopatholog*” OR “Whole-Slide Imag*”) AND (“Glioma*” OR “Astrocytoma*” OR “Oligodendroglioma*” OR “Glioblastoma*”)). The search was restricted to studies published between 2015 and 2025. The search query for each database is shown in Multimedia Appendix 1. To identify additional studies, we screened the reference lists of included studies (ie, backward reference list checking) and checked studies that cited the included studies (ie, forward reference list checking).

Study Selection

The selection process of relevant studies consisted of three steps. The first step involved removing duplicates using EndNote. In the second step, the studies were assigned to two authors, who screened the papers independently by title and abstract. The final step involved each author reading the consequent papers in full text independently. Discussions were held to resolve any differences or ambiguities in steps two and three.

Study Eligibility Criteria

Studies fulfilling the following criteria were included: (1) original articles, theses, dissertations, and conference papers written in English, (2) studies published in 2015 onward, (3) studies focused on patients with adult-type glioma with no restrictions on age, (4) studies that used histopathological images gathered from tissue biopsies, tumor excision, or digital histopathological images, (5) studies that assessed the performance of AI algorithms in the detection and classification of the molecular markers, specifically the IDH mutation status and the 1p/19q gene codeletion status using histopathology data, and (6) studies that reported results related to the performance of the developed AI models. No restriction was applied regarding study settings, study design, and country of publication. Studies were excluded if they (1) used data modalities other than histopathology images (eg, radiological images, stimulated Raman histology), (2) did not employ AI algorithms for molecular marker detection or classification, (3) were not human studies, or (4) were categorized as editorials, letters to editors, posters, conference abstracts, reviews, commentaries, short communications, and brief reports.

Data Extraction

Two authors independently extracted data from the eligible studies using predefined tables in Microsoft Excel sheet. Any discrepancies in the extracted data were resolved through discussion. The extracted data included four main categories: (1) study characteristics, including surname of first author, publication type, publication year, and country of publication; (2) participants and dataset characteristics, including number of participants, mean age, female percentage, dataset size, data source, magnification, slide image resolution, level of analysis, number of classes, and ground truth assessor; (3) AI model characteristics, covering problem-solving approaches, AI classifiers, type of classification, aim of classifier, data input to AI algorithms, features extraction methods, type of validation, and performance metrics; (4) results, including accuracy, sensitivity, specificity, and area under the curve (AUC). The data extraction form is shown in Multimedia Appendix 2.

Risk of Bias Assessment

To evaluate the quality and reliability of the included studies, two reviewers independently conducted a risk of bias assessment using an adapted version of the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) tool [19]. The original QUADAS-2 tool was modified to better suit the context of our review, which focuses on AI applications. Specifically, we incorporated relevant elements from the PROBAST (Prediction Model Risk of Bias Assessment Tool) to ensure that the assessment adequately addressed the methodological nuances of AI-based diagnostic models [20]. Each domain included four tailored signaling questions designed to determine the potential for bias and to evaluate applicability to the review’s research question. The risk of bias was rated separately for each domain, while concerns regarding applicability were assessed for the first three domains. For example, the “Participants” domain examined issues such as sample representativeness, exclusions, and subgroup balance. The “Index Test” domain evaluated the transparency and consistency of the AI model’s design and

XSL•FO
RenderX

feature use. The “Reference Standard” focused on the validity and consistency of outcome definitions and assessor qualifications, while the “Analysis” domain addressed data inclusion, preprocessing, set division, and model performance evaluation. In addition to risk of bias, we also assessed concerns regarding applicability in the first three domains, evaluating whether the participants, AI model, and outcome definitions were aligned with the review’s objectives. To refine and validate our modified tool, we conducted a pilot assessment using five studies, which allowed us to fine-tune the criteria before full application. Subsequently, two reviewers independently assessed all included studies using the finalized tool (Multimedia Appendix 3). Any discrepancies in assessments were resolved through discussion and consensus.

Data Synthesis

A narrative approach was used to synthesize the data extracted from the included studies. Specifically, we used texts and tables to summarize and describe the characteristics of the included studies and results. We could not perform a meta-analysis because fewer than two studies reported the confusion matrices or other necessary details (eg, number of cases and controls in the test set) required to calculate the numerators and denominators for the same performance measures (accuracy, sensitivity, specificity, and AUC), AI algorithms, and measured outcomes (detection of IDH and 1p/19q). Instead of meta-analysis, we calculated the traditional mean for each metric (ie, accuracy, sensitivity, specificity, and AUC).

Results

Search Results

As shown in Figure 1, the literature searches identified a total of 2453 reports from databases. After removing duplicate records (n=583), we screened the titles and abstracts of 1870 reports. This led to the exclusion of 1753 reports. Of the remaining 117 reports, we could not retrieve the full text for 5 reports. Subsequently, we checked the eligibility of the remaining 112 reports. This led to the exclusion of 91 reports for the following reasons: not related to glioma (n=7), not related to molecular classification (n=53), not pertaining to histopathology (n=11), inappropriate publication type (n=19), and not in the English language (n=1). We included an additional study by checking the reference list of the included studies. Ultimately, only 22 reports met the inclusion criteria and were included in the review [21-42].

XSL•FO
RenderX

Figure 1. Flowchart of the study selection process.

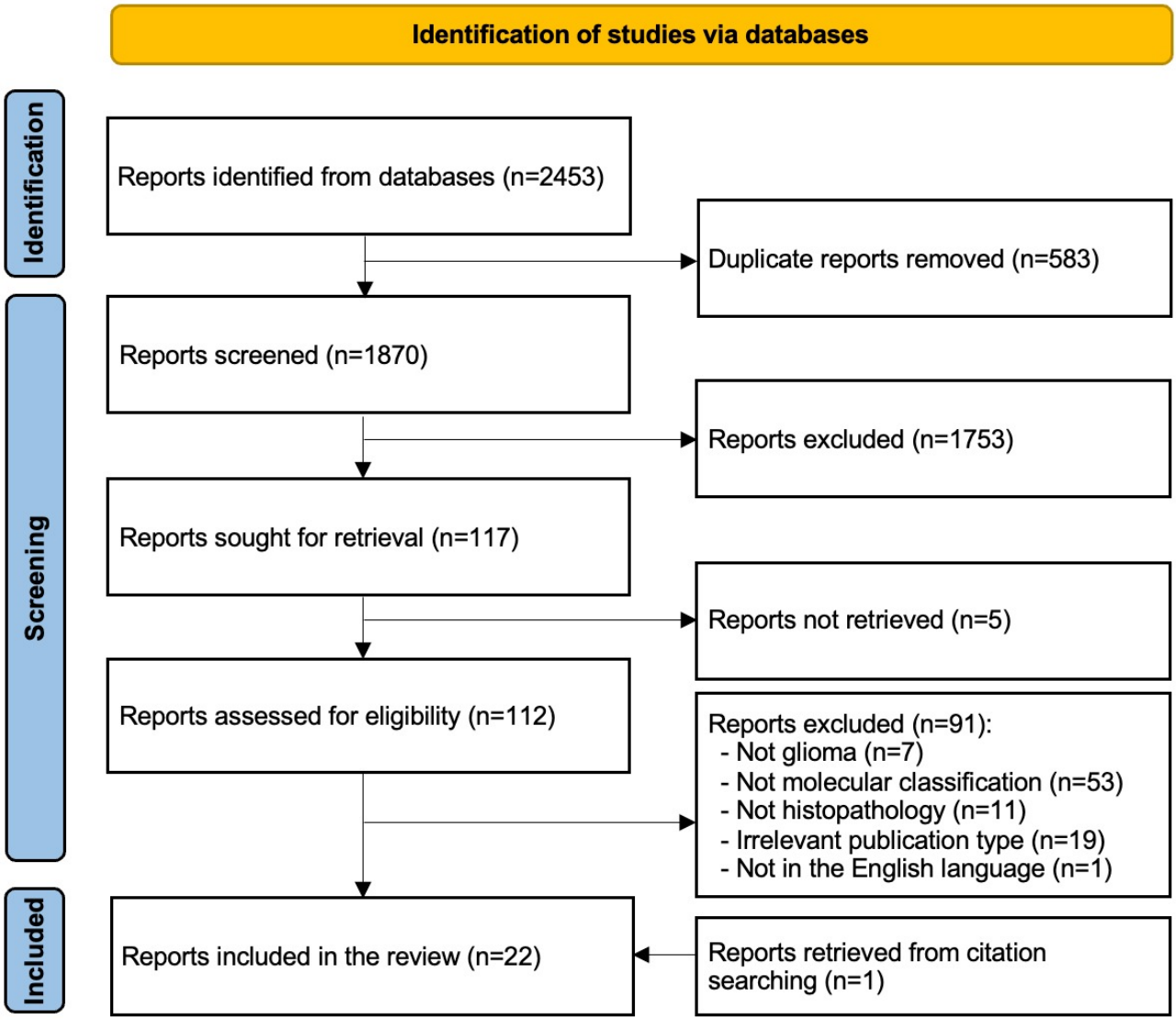

Characteristics of the Included Studies

The included studies were published between 2015 and 2024 (Table 1). The majority of studies were published in 2023 (9/22, 40.91%), followed by 2024 (5/22, 22.73%). Only 3 of the 22 studies (13.64%) are conference papers, whereas 16 (72.73%) are journal articles and 3 (13.64%) are preprints. Studies were carried out in different countries, with the United States and China being the most frequent contributors (6/22, 27.27%) and (5/22, 22.73%), respectively. The number of participants in the included studies ranges from 29 to 2845, with an average of 765.1 (SD 815.6). The mean of female participation rate is 42.45% with a range between 34.5% and 49.3%. The dataset size (ie, number of whole-slide image or patches) varied widely (47 - 27,000), with an average of 2464 (SD 5851). It is worth mentioning that the unit of analysis used to compute performance metrics was predominantly whole-slide image level (21/22, 95.5%), while it was at the patch level in 1 (4.5%) study. The participants' age ranged from 42.2 to 54.5 with an average of 48.03 (SD 4.2; Table 1). The characteristics of each included study are described in Multimedia Appendix 4.

XSL•FO
RenderX

Table . Characteristics of included studies.

Features	Number of studies	References
Year of publication, n (%)		
2024	5 (22.73)	[24,28,35,37,42]
2023	9 (40.91)	[21,22,25,27,30-33,40]
2022	2 (9.09)	[26,34]
2021	3 (13.64)	[29,38,41]
2020	2 (9.09)	[23,36]
2019	1 (4.55)	[39]
Publication type, n (%)		
Journal article	16 (72.73)	[23,24,26,27,29-31,33-38,40-42]
Conference paper	3 (13.64)	[21,28,32]
Preprint	3 (13.64)	[22,25,39]
Country of publication, n (%)		
United States	6 (27.27)	[28,29,32,34,38,39]
China	5 (22.73)	[23,25,35,40,42]
Germany	2 (9.09)	[21,27]
Australia	2 (9.09)	[36,41]
India	1 (4.55)	[22]
Luxembourg	1 (4.55)	[24]
Canada	1 (4.55)	[26]
Switzerland	1 (4.55)	[30]
South Korea	1 (4.55)	[31]
Saudi Arabia	1 (4.55)	[33]
Japan	1 (4.55)	[37]
Number of participants[a], mean (SD); range (%)	765.1 (815.6); 29 - 2845 (95.45)	[21,22,24-42]
Female percentage[b] (%), mean (SD); range (%)	42.45 (0.03); 34.5 - 49.3 (59.09)	[24,25,29,31,33-37,39-42]
Dataset size[c], mean (SD); range (%)	2464 (5851); 47 - 27,000 (95.45)	[22-42]
Mean age[d] (year), mean (SD); range (%)	48.03 (4.2); 42.2 - 54.5 (59.09)	[24,25,29,31,33-37,39-42]

[a]One study did not report the mean age [23].

[b]Studies did not report the female percentage [21-23,26-28,30,32,38].

[c]One study did not report the dataset size [21].

[d]Studies did not report the mean age [21-23,26-28,30,32,38].

Features of Histopathological Images

As presented in Table 2, the histopathological images were captured at various magnification levels, including 2.5×, 5×, 10×, 20×, 40×, and 100×. Among these, 20× magnification was the most commonly used (12/22, 54.55%). While 14 (63.64%) studies focused on architectural-level analysis, 13 (59.09%) studies focused on cellular-level analysis. Slide resolutions varied across studies, with 256×256 pixels being the most common (8/22, 36.36%). Datasets were sourced from either open-source databases (16/22, 72.73%) or closed-source datasets (13/22, 59.09%) (ie, data were collected by the study authors or obtained from previous studies). The open datasets used include The Cancer Genome Atlas (15/22, 68.18%) and the Digital Brain Tumor Atlas (1/22, 4.55%), with some studies combining open and closed sources. The histopathological images were labeled into 2 classes in about two-thirds of the studies (14/22, 63.64%). The features of histopathological images in each study are described in Multimedia Appendix 5.

XSL•FO
RenderX

Table . Features of histopathological images[a].

Features	Number of studies, n (%)	References
Magnification		
2.5×	1 (4.55)	[34]
5×	2 (9.09)	[32,34]
10×	4 (18.18)	[24,29,34,36]
20×[b]	*12* (54.55)	[22,24-26,28,31,32,34,35,40-42]
40×	5 (22.73)	[24,30,31,35,41]
100×	1 (4.55)	[24]
Not reported	7 (31.82)	[21,23,27,33,37-39]
Level of analysis		
Architecture level	*14* (63.64)	[22,24-26,28,29,31,32,34-36,40-42]
Cell level	13 (59.09)	[22,24-26,28,30-32,34,35,40-42]
Not reported	7 (31.82)	[21,23,27,33,37-39]
Slide image resolution		
224×224	4 (18.18)	[29,31,37,42]
256×256	*8* (36.36)	[22,23,25,28,32,34,36,37]
512×512	3 (13.64)	[21,23,38]
1024×1024	5 (22.73)	[23,33,35,36,40]
Other	7 (31.82)	[21,23,24,26,30,41,42]
Not reported	2 (9.09)	[27,39]
Data sources		
Open source	*16* (72.73)	[21-23,25,27-29,31-34,36-39,41]
Closed source	13 (59.09)	[24-27,30,31,33-36,40-42]
Open dataset name		
TCGA[c]	*15* (68.18)	[21,23,25,27-29,31-34,36-39,41]
DBTA[d]	1 (4.55)	[22]
Not applicable	6 (27.27)	[24,26,30,35,40,42]
Number of classes		
Two	*14* (63.64)	[22,23,28-37,41,42]
Three	2 (9.09)	[21,26]
Four	4 (18.18)	[25,27,38,39]
Five	2 (9.09)	[24,33]
Six	1 (4.55)	[40]

[a]Studies used both data sources, open and closed [25,31,33,34,36,41].

[b]Values in italics indicate highest score.

[c]TCGA: The Cancer Genome Atlas.

[d]DBTA: Digital Brain Tumor Atlas.

Features of AI

AI was primarily used for binary classification in 86.36% (19/22) of studies, where models differentiate between two outcomes (eg, IDH-mutant vs wild-type), whereas only 22.73% (5/22) employed multiclass classification, which predicts among 3 or more glioma subtypes (Table 3). The majority of studies (20/22, 90.91%) used AI for IDH subtyping (ie, distinguishing between IDH mutant and IDH wild-type), whereas only 6 (27.27%) studies applied AI for subtyping IDH mutant tumors in relation to 1p/19q gene codeletion. Convolutional neural networks (CNNs) were the most commonly used classifier and feature extraction method, appearing in 54.55% (12/22) and 90.91% (20/22) of studies. For specific classifiers, CNNs are

XSL•FO
RenderX

the most utilized, appearing in 54.55% (12/22) of studies, followed by multiple instance learning (MIL) (which handles weakly labeled slide-level data by learning from sets of image patches without requiring pixel-level annotations) at 50% (11/22), transformers at 18.18% (4/22), and hybrid models (ie, single end-to-end architectures that integrate convolutional components with attention-based mechanisms [eg, CNN-transformer models] within a unified computational graph and are trained jointly) also at 18.18% (4/22). For specific feature extraction models, ResNet50 leads with 46.67% (10/22) usage, followed by densely connected convolutional network 121 (DenseNet121) and InceptionV3, both at 19.05% (4/22). Among the included studies, most models were developed using histopathological images alone (17/22, 77.3%), while a smaller number (5/22, 22.7%) incorporated additional modalities such as MRI scans (2/22, 9.1%), demographic information (3/22, 13.6%), and clinical variables (2/22, 9.1%). The most widely used method for model validation was k-fold cross-validation, used in 50% (11/22) of studies, followed by the train-test split approach in 45.45% (10/22) of studies. AI model performance was assessed using multiple metrics, with accuracy being the most frequently reported (19/22, 86.36%), followed by AUC and sensitivity, each used in 68.18% (15/22) of studies. A detailed summary of the AI models used in each study is provided in Multimedia Appendix 6.

XSL•FO
RenderX

Table . Features of artificial intelligence (AI).

Feature	Number of studies, n (%)	References
Type of classification		
Binary[a]	*19* (86.36)	[22,23,25,27-42]
Multiclass	5 (22.73)	[21,24,26,33,40]
Aim of AI algorithms[b]		
Subtyping IDH[c]	*20* (90.91)	[21-29,32-42]
Subtyping 1p/19q codeletion	6 (27.27)	[25,27,30,31,38,39]
Classifier		
Convolutional neural networks	*12* (54.55)	[24,26,29,30,33-36,38-40,42]
Multiple instance learning	11 (50)	[21-23,25,27,28,31-33,37,42]
Transformers	4 (18.18)	[30,33,35,42]
Hybrid models	2 (9.09)	[35,42]
Ensemble model	2 (9.09)	[34,37]
Logistic regression	2 (9.09)	[31,36]
Random forest	1 (4.55)	[41]
Multilayer perceptron	1 (4.55)	[21]
Specific classifier		
AttMIL[d]	*7* (31.82)	[22,23,25,27,28,33,37]
ResNet50	5 (22.73)	[24,35,36,40,42]
Traditional MIL[e]	4 (18.18)	[21,31,33,42]
InceptionV3	4 (18.18)	[24,33,35,36]
ViT[f]	4 (18.18)	[24,33,35,42]
DenseNet121[g]	3 (13.64)	[24,34,35]
VGG19[h]	3 (13.64)	[24,26,36]
Features extraction methods		
Convolutional neural networks	*20* (90.91)	[21-26,28-40,42]
Transformers	4 (18.18)	[33,35,37,42]
Hybrid models	4 (18.18)	[27,35,37,42]
MIL	1 (4.55)	[42]
Random forest	1 (4.55)	[41]
Specific features extraction model		
ResNet50	*10* (45.45)	[22,24,25,28,31,35-37,40,42]
DenseNet121	5 (22.73)	[21,22,24,34,35]
InceptionV3	4 (18.18)	[24,33,36,37]
ResNet18	3 (13.64)	[24,29,32]
VGG19	2 (9.09)	[24,26,36]
ViT	2 (9.09)	[33,42]
Not reported	2 (9.09)	[23,30]
Data modality		
Unimodel data	17 (77.3)	[22-35,39,40,42]
Multimodel data	5 (22.7)	[21,36-38,41]

XSL•FO
RenderX

Feature	Number of studies, n (%)	References
Types of validation		
k-fold cross-validation	*11* (50)	[21,23,25,27,28,31,33,37-39,41]
Training-test split	10 (45.45)	[22,24,26,29,30,32,34-36,40]
External validation	7 (31.82)	[25,27,29,31,34,40,42]
Performance measures		
Accuracy	*19* (86.36)	[21-25,27,28,31-42]
AUC[i]	15 (68.18)	[21-23,26,28,29,31,33-36,39-42]
Sensitivity	15 (68.18)	[23-25,27,28,30-32,34-37,39,40,42]
F_1-score	9 (40.91)	[21,24,25,31,32,35,37,40,42]
Precision	8 (36.36)	[23,24,27,30-32,35,37]
Specificity	8 (36.36)	[25,27,28,34,36,39,40,42]
Others	8 (36.36)	[21,22,25,27-29,37,39]

[a]Values in italic indicate highest score.

[b]Studies reported results for both IDH mutation status and 1p/19q codeletion status [25,27,38,39].

[c]IDH: isocitrate dehydrogenase.

[d]AttMIL: attention-based multiple instance learning.

[e]MIL: multiple instance learning.

[f]ViT: vision transformer.

[g]DenseNet121: densely connected convolutional network 121.

[h]VGG: visual geometry group.

[i]AUC: area under the curve.

Results of the Risk of Bias and Applicability

In the selection of participants domain, nearly a quarter of the included studies (5/22, 23%) provided sufficient information to confirm the use of an appropriate consecutive or random sampling of eligible participants. Nearly all studies (21/22, 95%) reported a sufficient sample size. Additionally, 8 (36%) studies avoided inappropriate exclusions, while over half (12/22, 55%) demonstrated balanced subgroup representation. Consequently, a small proportion of the studies (2/22, 9%) were assessed as having a low risk of bias, while a larger portion (8/22, 36%) were rated as having an unclear risk of bias in the "selection of participants" domain, as shown in Figure 2.

In terms of matching participants to the predefined requirements of the review question, a low level of concern was identified in 95% (21/22) of the included studies, as shown in Figure 3.

Figure 2. Results of the assessment of risk of bias in the included studies.

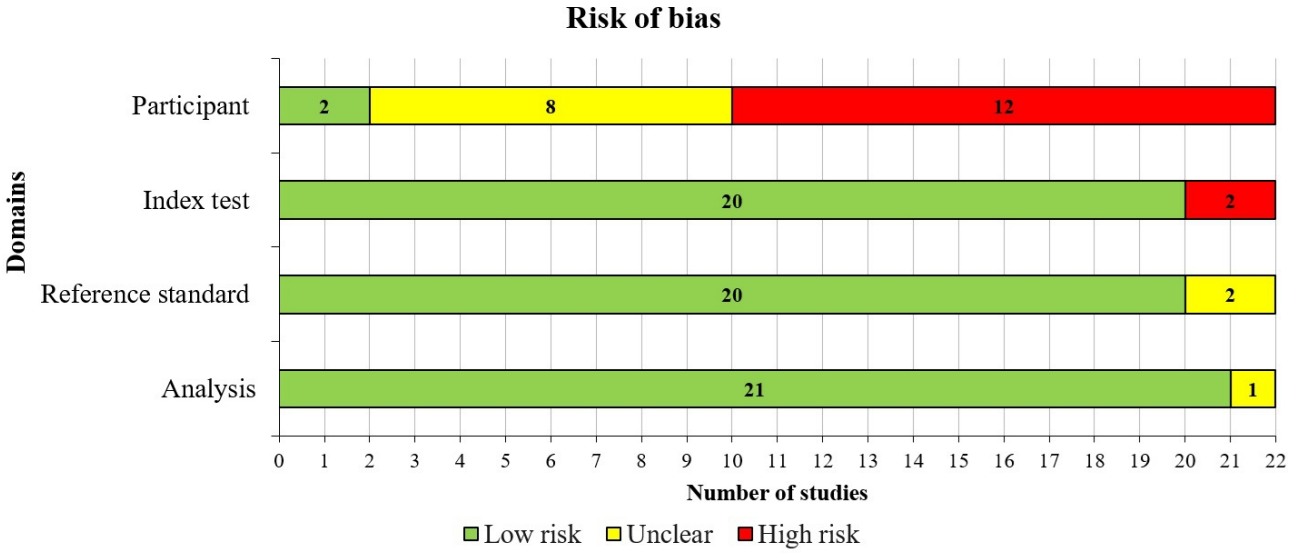

XSL•FO
RenderX

Figure 3. Results of the assessment of applicability concerns in the included studies.

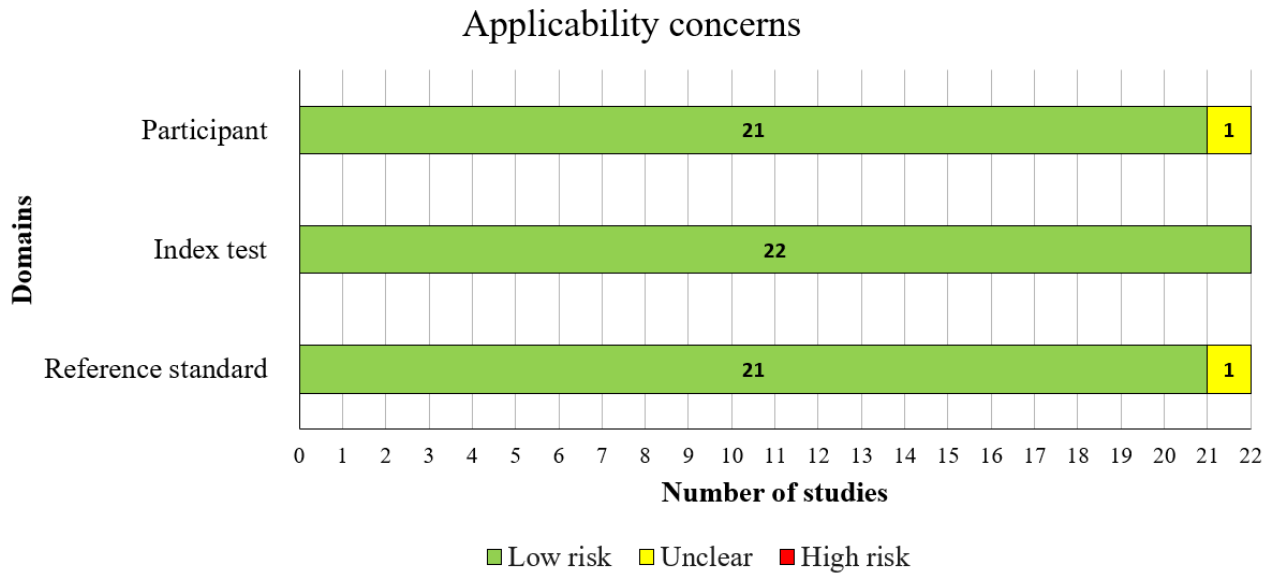

A substantial majority of the included studies comprehensively detailed their AI models, with 21 of the 22 (95%) studies providing thorough descriptions. An overwhelming majority, all 22 studies (100%) clearly reported the features (predictors) used. Moreover, all 22 studies (100%) ensured that these features were sourced without prior knowledge of the outcome data. Consistency in feature assessment across participants was observed in 21 of the 22 (95%) studies. Consequently, the potential for bias in the "index test" domain was assessed as low in 20 of the 22 (91%) studies**,** as shown in Figure 2. In addition, all 22 (100%) studies were found to have minimal concerns regarding the alignment between the model's predictors and the review question's criteria, as illustrated in Figure 3.

In most of the included studies (20/22, 91%), the outcome of interest, histopathology image analysis, was conducted by an expert (eg, pathologist or neuropathologist), ensuring the necessary expertise to accurately classify the outcomes. Nearly all studies (22/22, 100%) applied consistent and uniform outcome definitions across all participants, with outcome classification (ie, image labeling) blinded to the AI model predictions. An overwhelming majority (22/22, 100%) determined the outcome without prior knowledge of the predictor information. Furthermore, in all 22 (100%) studies, the diagnostic process was carried out over an appropriate duration, using the same diagnostic criteria for all participants. As a result, the risk of bias in the "reference standard" domain was considered low in the majority of studies (20/22, 91%), as shown in Figure 2. Additionally, a similar proportion of studies (21/22, 95%) exhibited minimal concerns regarding discrepancies in outcome definition, timing, and determination methods, as shown in Figure 3.

Finally, a significant majority of the studies (22/22, 100%) ensured the inclusion of all enrolled participants in the data analysis. A substantial number of these studies (21/22, 95%) reported no missing values, or if missing values were present, they were handled appropriately (eg, through multiple imputation).

Similarly, a high proportion of studies (22/22, 100%) adopted suitable measures to evaluate the performance of their models. The confusion matrix was presented, or more than one evaluation measure was used, and the selected measures were deemed appropriate.

Nearly half of the studies (21/22, 95%) demonstrated an appropriate split among training, validation, and test sets. The chosen distribution aligned with best practices in the field (eg, 70% - 80% for training, 10% - 15% for validation, and 10% - 20% for testing).

Consequently, the conduct and interpretation of the analysis did not introduce bias in most studies, with 21/22 (95%) having a low risk of bias in the "analysis" domain, as shown in Figure 2. A detailed breakdown of the "risk of bias" and "applicability concerns" for each domain in every study is available in Multimedia Appendix 7.

Results of the Included Studies

As shown in Table 4, the accuracy of AI models in detecting or subtyping molecular markers in adult-type gliomas ranged from 58% to 100%, with a mean accuracy of 85.46%. Sensitivity ranged from 62% to 100%, averaging 84.55%, while specificity ranged from 46% to 100%, with a mean of 86.03%. The AUC values ranged from 46% to 99%, with an average of 86.53%. The hybrid model demonstrated the highest accuracy (92.80%) and sensitivity (89.62%), while the CNN model achieved the highest specificity (89.30%). Logistic regression outperformed other models in terms of AUC (93.05%). Among specific CNN architectures, DenseNet121 achieved the highest accuracy (90.82%), sensitivity (89.28%), and AUC (91.05%). Moreover, AI models that used multimodal data outperformed those that used unimodal data in terms of sensitivity (90.15% vs 84.31%) and AUC (88.93% vs 86.29%). In outcome prediction, AI models showed better performance in detecting IDH mutations compared to 1p/19q codeletions, with higher accuracy (86.13% vs 81.63%), specificity (86.61% vs 78.11%), and AUC (86.74% vs 85.15%). AI models also performed better in IDH subtyping for multiclass classification compared to binary classification

XSL•FO
RenderX

in terms of accuracy (91.98% vs 84.02%) and sensitivity (93.41% vs 80.18%). Conversely, AI models for binary classification yielded a lower performance than multiclass classification in terms of accuracy (91.98% vs 84.02%) and sensitivity (93.41% vs 80.18%). However, these differences should be interpreted as descriptive trends rather than statistically validated superiority, as formal between-group comparisons were not feasible.

XSL•FO
RenderX

Table . Summary of overall performance of all models based on number of classes, type of data, and artificial intelligence (AI) algorithms in the target variable[a].

Groups	Accuracy			Sensitivity			Specificity			AUC[b]		
	Number of studies	Mean % (SD)	Range	Number of studies	Mean % (SD)	Range	Number of studies	Mean % (SD)	Range	Number of studies	Mean % (SD)	Range
Algorithms												
CNNs[c]	8	85.60 (0.09)	0.72 - 1.00	8	84.09 (0.11)	0.64 - 1.00	7	*89.30*[d] (0.09)	0.75 - 1.00	8	86.54 (0.06)	0.76 - 0.99
MIL[e]	10	80.83 (0.09)	0.58 - 0.91	8	82.54 (0.13)	0.62 - 0.99	3	68.69 (0.23)	0.46 - 0.94	10	81.73 (0.16)	0.46 - 0.97
Transformer	4	88.27 (0.03)	0.83 - 0.92	3	81.62 (0.10)	0.69 - 0.92	—	—	—	3	92.37 (0.04)	0.89 - 0.96
Hybrid model	3	*92.80* (0.05)	0.85 - 0.97	3	*89.62* (0.08)	0.81 - 0.97	—	—	—	2	92.20 (0.04)	0.89 - 0.95
Ensemble model	2	86.35 (0.12)	0.78 - 0.95	2	89.60 (0.05)	0.86 - 0.93	—	—	—	2	92.00 (0.09)	0.85 - 0.99
Logistic regression	2	87.15 (0.01)	0.86 - 0.88	2	86.00 (0.12)	0.78 - 0.94	—	—	—	2	*93.05* (0.00)	0.93 - 0.931
CNN models												
ResNet (ResNet18)	7	87.04 (0.09)	0.72 - 1.00	6	87.51 (0.11)	0.69 - 1.00	5	*91.48* (0.07)	0.84 - 1.00	6	90.19 (0.06)	0.81 - 0.99
VGG[f] (VGG16,19)	4	89.67 (0.09)	0.79 - 1.00	4	87.03 (0.13)	0.69 - 1.00	2	90.65 (0.13)	0.81 - 1.00	2	81.55 (0.01)	0.81 - 0.82
DenseNet121[g]	3	*90.82* (0.10)	0.79 - 0.97	3	*89.28* (0.10)	0.78 - 0.97	—	—	—	2	*91.05* (0.08)	0.86 - 0.97
InceptionV3	3	80.15 (0.05)	0.77 - 0.86	3	75.82 (0.11)	0.64 - 0.86	—	—	—	3	83.58 (0.03)	0.80 - 0.87
Type of data												
Unimodel data	20	*85.67* (0.09)	0.58 - 1.00	16	84.31 (0.11)	0.62 - 1.00	9	*86.27* (0.15)	0.46 - 1.00	16	86.29 (0.10)	0.46 - 0.99
Multimodel data	5	83.64 (0.05)	0.78 - 0.90	2	*90.15* (0.06)	0.86 - 0.94	—	—	—	4	*88.93* (0.03)	0.85 - 0.93
Outcome prediction												
1p/19q	6	81.63 (0.12)	0.58 - 1.00	6	*87.64* (0.09)	0.75 - 1.00	3	78.11 (0.28)	0.46 - 1.00	5	85.15 (0.19)	0.46 - 0.99
IDH[h]	19	*86.13* (0.08)	0.72 - 1.00	14	83.96 (0.11)	0.62 - 1.00	8	*86.61* (0.12)	0.52 - 1.00	16	*86.74* (0.08)	0.52 - 0.99
Class prediction												
Binary IDH	16	84.02 (0.07)	0.72 - 1.00	12	80.18 (0.11)	0.62 - 1.00	8	*86.61* (0.12)	0.52 - 1.00	15	*86.87* (0.08)	0.52 - 0.99
Multiclass IDH	3	*91.98* (0.07)	0.75 - 1.00	2	*93.41* (0.05)	0.85 - 1.00	—	—	—	—	—	—
Overall	*20*	*85.46* (0.08)	*0.58 - 1.00*	*16*	*84.55* (0.01)	*0.62 - 1.00*	*9*	*86.03* (0.15)	*0.46 - 1.00*	*17*	*86.53* (0.10)	*0.46 - 0.99*

[a]Specificity not reported in all studies.

[b]AUC: area under the curve.

[c]CNNs: convolutional neural networks.

XSL•FO
RenderX

[d]Values in italics indicate highest score.

[e]MIL: multiple instance learning.

[f]VGG: visual geometry group.

[g]DenseNet121: densely connected convolutional network 121.

[h]IDH: isocitrate dehydrogenase.

Discussion

Principal Findings

This review aimed to systematically evaluate the performance of AI models in detecting and classifying IDH mutation status and 1p/19q codeletion in adult-type gliomas using histopathological images. Our analysis indicates that these models demonstrate promising diagnostic capabilities, achieving pooled mean values of 85.46% for accuracy, 84.55% for sensitivity, 86.03% for specificity, and 86.53% for the area under the receiver operating characteristic curve. As noticed, the performance across all metrics was comparable, suggesting that AI models offer balanced diagnostic capabilities. Despite these promising results, they remain below the commonly cited thresholds for clinical adoption, which often exceed 90% for high-stakes diagnostic tasks [42]. As such, the current evidence is not sufficient to support stand-alone clinical decision-making. Therefore, the results of this meta-analysis should be interpreted as early indicators of feasibility rather than clinically actionable performance, underscoring the substantial methodological and validation work still required before these models can be safely integrated into clinical workflows.

These results are generally consistent with prior reviews. Specifically, a systematic review conducted by Farahani et al [43] reported a similar performance in IDH prediction using MRI, with pooled sensitivity of 84% and specificity of 87% and an AUC of 0.89. Further, Chen et al [44] found pooled sensitivity of 83%, specificity of 84%, and AUC of 0.90 for predicting IDH mutations using machine learning–based radiomics. Taken together, these comparisons suggest that our findings are in line with the broader body of research, reinforcing the reliability and clinical promise of AI-based histopathological approaches for glioma molecular classification.

Individual studies showed considerable variability across performance metrics; for instance, specificity values ranged from 46% to 100%. This variability likely stems from population differences, sample sizes, model architectures, and the inherent challenge of predicting rare molecular alterations (eg, telomerase reverse transcriptase mutations) from histopathology images, compounded by class imbalance, domain shift (ie, performance degradation when models are applied to data from different sources), and the absence of distinct histological correlates. While accuracy and sensitivity remain higher for well-characterized targets (eg, IDH, 1p/19q), specificity suffers when the model encounters rare or ambiguous cases in external validation.

Hybrid models achieved the highest mean accuracy (92.80%) and sensitivity (89.62%), while logistic regression models achieved the top mean AUC just slightly higher than hybrid models (93.05% vs 92.20%). These results suggest that hybrid models deliver the strongest overall performance. Hybrid models are defined as AI frameworks that integrate multiple deep learning architectures to leverage the complementary strengths of each component. In the context of glioma diagnosis, most recent reviews have highlighted the effectiveness of hybrid models combining CNNs with transformer-based architectures [45-47]. CNNs excel at extracting local spatial features from histopathological images, such as cellular morphology and texture patterns, while transformers are adept at capturing long-range dependencies and global contextual information through their self-attention mechanisms. By combining these capabilities, hybrid CNN-transformer models can achieve a more comprehensive understanding of complex tissue structures and molecular characteristics within gliomas, leading to improved predictive performance. Our findings align with a recent review by Redlich et al [45], which analyzed 83 studies on AI-based methods for whole-slide histopathology image analysis of gliomas. The review highlighted that hybrid models combining CNNs with other architectures, including transformers, consistently outperform single-architecture models in tasks such as subtyping, grading, and molecular marker prediction [45]

Among pure CNN backbones, DenseNet121 demonstrated the best performance, achieving an accuracy of 90.82%, sensitivity of 89.28%, and an AUC of 91.05%. DenseNet121's superior performance may be due to its unique architectural features, notably dense connectivity, which enhance training dynamics, feature representation, and parameter efficiency.

In terms of data modalities employed by AI models, our analysis demonstrated that multimodal approaches exhibited superior performance in sensitivity (90.15% vs 84.31%) and AUC (88.93% vs 86.29%) compared to unimodal models. This suggests that multimodal models benefit from richer and more diverse feature representations, thereby enhancing their diagnostic capability. However, unimodal models achieved slightly higher accuracy (85.67% vs 83.64%), which may be attributed to the fact that none of the five multimodal studies included in the accuracy analysis employed hybrid modeling. Instead, they utilized simple model ensembles or comparisons involving architectures such as multilayer perceptron, logistic regression, random forests, and CNNs. Our findings are consistent with existing literature. For instance, a systematic review by Alleman et al [48] concluded that multimodal deep learning approaches enhance prognostic accuracy in glioma patients compared to unimodal models, particularly in predicting overall survival. Similarly, d'Este et al [49] demonstrated that combining multimodal imaging with AI techniques improves visualization of glioma infiltration, surpassing the capabilities of single-modality approaches. Furthermore, the GlioMT model, introduced by Byeon et al [50], which integrates imaging and clinical data, exhibited superior performance in predicting adult-type diffuse gliomas compared to conventional CNNs and visual transformers.

XSL•FO
RenderX

Regarding outcome prediction and AI model performance in detecting molecular markers, our analysis revealed that AI models demonstrated stronger performance in identifying IDH mutations compared to 1p/19q codeletions. Specifically, IDH detection yielded higher accuracy (86.13% vs 81.63%), specificity (86.61% vs 78.11%), and AUC (86.74% vs 85.15%). In contrast, AI models achieved slightly better sensitivity for 1p/19q codeletion detection (87.64% vs 83.96%). The discrepancy in performance between IDH mutation and 1p/19q codeletion detection may be attributed to several factors. IDH mutations often result in more distinct histopathological features, making them more amenable to detection by AI models. In contrast, 1p/19q codeletions may not produce as pronounced morphological changes, posing a greater challenge for AI-based detection. The findings of published studies align with our results. For example, the review by Farahani et al [43] showed higher performance in identifying IDH mutations compared to 1p/19q codeletions using MRI in terms of sensitivity (84% vs 76%) and specificity (87% vs 85%).

In our systematic review, CNNs and MIL algorithms were the most predominantly used in AI models, likely due to their suitability for histopathological image analysis. CNNs are designed to capture spatial hierarchies in image data through convolutional layers, making them effective at identifying morphological patterns such as cellular structures and tissue architecture that are critical in tumor classification. Histopathological slides of gliomas are typically of high resolution and contain complex visual features, which CNNs can learn to recognize with high accuracy. Meanwhile, MIL frameworks address the challenge of weakly labeled data, which is common in medical imaging, where only slide-level labels (eg, tumor type or mutation status) are available, rather than precise annotations at the pixel or region level. MIL allows models to learn from sets of image patches (instances) while associating predictions with the whole slide (bag), making it particularly useful in computational pathology where exhaustive labeling is impractical.

Practical and Research Implications

While our results are encouraging, they remain not optimal and arguably lower than that acceptable in clinical practice [42]. Thus, our findings should be interpreted with caution due to several limitations. First, 15 of the 22 studies included in the review utilized the same dataset, raising concerns regarding the generalizability of the results. Second, many of the studies were based on small sample sizes, limiting the robustness and reliability of the conclusions. Third, several key findings were derived from fewer than 5 studies, reducing the strength of the evidence. Lastly, the aggregation of results using simple arithmetic means rather than formal meta-analytic techniques may have limited the ability to draw definitive conclusions about the overall efficacy of AI models.

Given these limitations, AI should not be used as a stand-alone diagnostic tool for the detection or subtyping of adult-type gliomas but rather as a complementary approach alongside conventional methods. Future research should prioritize the development of AI models using larger and more diverse datasets to enhance performance and generalizability. Ensuring adequate sample sizes and improving the transparency of reporting, such as including confusion matrices, will also facilitate more rigorous meta-analyses. Notably, the majority of studies reviewed developed AI models exclusively from histopathological images. Developing multimodal AI models that incorporate additional data types (eg, radiological and clinical information) may further enhance diagnostic accuracy. Thus, continued efforts are needed to advance the integration of multimodal data in AI model development for the detection and subtyping of adult-type gliomas.

In this review, the majority of studies used AI algorithms that were either MIL or CNNs, indicating a research gap in exploring the performance and effectiveness of other types of AI algorithms (eg, transformers, hybrid models, ensemble models [ie, approaches that combine predictions from multiple independently trained models using averaging, voting, or weighted fusion strategies], or logistic regression). Additionally, we observed a lack of research assessing the performance of AI models that use multimodal data, which may have affected the accuracy of the results. Moreover, the number of studies evaluating AI models' performance in detecting IDH mutations was noticeably greater than those that used 1p/19q as a predicted outcome. These studies predominantly employed binary IDH classification (ie, IDH mutation vs IDH wild-type), whereas studies using IDH multiclass classification (eg, astrocytoma, oligodendroglioma) were markedly fewer. Future research should address these gaps to produce more accurate and representative findings.

Our findings that AI hybrid models performed better in detecting adult-type gliomas than other models underscore the need for more advanced approaches to explore their effectiveness and their potential to achieve diagnostic outcomes comparable to those of histopathologists. However, their real-world applicability requires consideration of computational and operational demands. These models often require greater processing power (eg, GPU-enabled servers or high-performance cloud computing) and longer inference times compared to conventional CNNs. Such requirements may pose challenges for pathology laboratories lacking advanced digital infrastructure. Nonetheless, many modern clinical centers are increasingly equipped with GPU-accelerated digital pathology ecosystems, and cloud-based deployment can significantly reduce the need for on-site hardware. From a cost-benefit perspective, the additional computational resources required by hybrid models may be justified given their superior accuracy and robustness, particularly for molecular prediction tasks where diagnostic errors carry serious clinical implications. However, successful implementation in routine practice will depend on streamlined model optimization, efficient inference pipelines, and prospective validation of runtime performance to ensure that diagnostic gains meaningfully outweigh operational costs.

As our findings indicated superior performance of multimodal AI models in sensitivity and AUC, we suggest a complementary approach: multimodal AI models may serve as effective initial diagnostic tools due to their comprehensive data integration, while unimodal models can be employed subsequently for more accurate and specific outcome determination. Moreover, our results reflect a stronger performance of AI models in the

XSL•FO
RenderX

detection of IDH mutations compared to 1p/19q codeletion in all metrics except sensitivity. Therefore, we suggest that while IDH mutation detection should remain a primary focus for model refinement and deployment due to its balanced high performance, models targeting 1p/19q codeletions may serve as effective initial diagnostic tools in adult-type gliomas due to their higher sensitivity, which is critical in early disease detection.

Given that the current review focused on IDH mutations and 1p/19q codeletions, we encourage researchers to undertake further reviews covering other glioma molecular markers, including MGMT methylation, ATRX, and telomerase reverse transcriptase mutations. Moreover, further research should extend to other types of brain tumors, such as meningiomas, pituitary adenomas, and schwannomas. These tumors differ significantly in terms of biological behavior, treatment options, and clinical outcomes, warranting a comprehensive evaluation of AI's potential in diagnosing and subtyping each tumor type.

Limitations

This systematic review is subject to several limitations. First, the generalizability of the findings is limited, as approximately 68.18% of the included studies relied on the same dataset, and the number of included studies is relatively small (n=22). This raises concerns about the external validity of the results and their applicability to the broader population of adult-type glioma patients. To address the potential duplication of patients or slides across studies (ie, those relying on The Cancer Genome Atlas), we reviewed the data sources reported by each study to identify where overlap was likely. Because individual-level data were not available, we were unable to verify or remove duplicate cases. Instead, we treated studies drawing from the same dataset as nonindependent samples and interpreted pooled estimates with caution. As a result, our summary metrics should be viewed as descriptive indicators rather than independent, unbiased effect estimates. Second, the review focused exclusively on adult-type gliomas, thereby restricting the ability to draw conclusions about the performance of AI in detecting other adult brain tumors or brain tumors in the general population. Third, the analysis was confined to studies utilizing histopathological images, limiting the assessment of AI performance with alternative imaging modalities, such as MRI or computed tomography scans. Fourth, meta-analysis was not feasible due to incomplete reporting of key performance metrics across studies. Instead, mean values were calculated to provide a descriptive summary, though this approach lacks the statistical rigor and precision of formal meta-analysis. Finally, the inclusion of only English-language studies may have introduced language bias and led to the exclusion of relevant research published in other languages, potentially affecting the comprehensiveness of the review.

Conclusion

This systematic review demonstrates that AI models applied to histopathological images show strong potential for the molecular classification of adult-type gliomas, particularly in detecting IDH mutations and 1p/19q codeletions. Our findings indicate that while overall diagnostic performance is promising—especially for IDH mutations—variability in model performance and methodological limitations across studies temper the generalizability of results. Hybrid models and multimodal approaches emerged as particularly effective strategies, offering enhanced sensitivity and discriminative capability. However, the limited use of diverse datasets, the predominance of unimodal designs, and the lack of standardized reporting remain critical barriers to clinical translation. Given these constraints, AI models should currently serve as complementary tools rather than stand-alone diagnostic systems, functioning primarily to assist pathologists in pattern recognition and decision support while final diagnoses remain clinician-led. Future research should prioritize methodological standardization, broader data integration, and validation across larger, heterogeneous populations to support the clinical deployment of AI in glioma diagnostics.

Acknowledgments

The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to generative artificial intelligence (GAI) tools under full human supervision: Proofreading and editing. The GAI tool used was ChatGPT-4. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. All intellectual content, study design, data collection, analysis, and final interpretations are the sole responsibility of the authors.

Funding

The publication of this article was funded by the Weill Cornell Medicine–Qatar Health Sciences Library.

Data Availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Authors' Contributions

Concept for the paper and experimental design: AA, RAD, and OA
Data collection and extraction: AA, RAD, and OA
Data analysis: AA and OA

XSL•FO
RenderX

Risk of bias assessment: AT
Writing original draft preparation: AO, RAD, and OA
Writing-review and editing: AA, AO, RAD, and OA
Supervision: AA
All authors agreed to submission of the manuscript.

Conflicts of Interest

AA is an associate editor of *JMIR Nursing* at the time of publication. All other authors declare no conflict of interest.

Multimedia Appendix 1
Search strategy.
[DOCX File, 30 KB - jmir_v28i1e78377_app1.docx]

Multimedia Appendix 2
Data extraction form.
[DOCX File, 23 KB - jmir_v28i1e78377_app2.docx]

Multimedia Appendix 3
Modified version of QUADAS-2.
[DOCX File, 28 KB - jmir_v28i1e78377_app3.docx]

Multimedia Appendix 4
Characteristics of each included study.
[DOCX File, 43 KB - jmir_v28i1e78377_app4.docx]

Multimedia Appendix 5
Features of histopathological images in each study.
[DOCX File, 39 KB - jmir_v28i1e78377_app5.docx]

Multimedia Appendix 6
Features of artificial intelligence in each study.
[DOCX File, 44 KB - jmir_v28i1e78377_app6.docx]

Multimedia Appendix 7
Reviewers' judgments about each domain in "risk of bias" and "applicability concerns" for each included study.
[DOCX File, 218 KB - jmir_v28i1e78377_app7.docx]

Checklist 1
PRISMA-DTA checklist.
[DOC File, 67 KB - jmir_v28i1e78377_app8.doc]

References

1. Ostrom QT, Cioffi G, Waite K, Kruchko C, Barnholtz-Sloan JS. CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2014–2018. Neuro Oncol 2021 Oct 5;23(12 Suppl 2):iii1-iii105. [doi: 10.1093/neuonc/noab200] [Medline: 34608945]
2. Reynoso-Noverón N, Mohar-Betancourt A, Ortiz-Rafael J. Epidemiology of brain tumors. In: Principles of Neuro-Oncology: Springer; 2020:15-25. [doi: 10.1007/978-3-030-54879-7_2]
3. Lapointe S, Perry A, Butowski NA. Primary brain tumours in adults. The Lancet 2018 Aug;392(10145):432-446. [doi: 10.1016/S0140-6736(18)30990-5]
4. Perry A, Wesseling P. Histologic classification of gliomas. Handb Clin Neurol 2016;134:71-95. [doi: 10.1016/B978-0-12-802997-8.00005-0] [Medline: 26948349]
5. Chen R, Smith-Cohn M, Cohen AL, Colman H. Glioma subclassifications and their clinical significance. Neurotherapeutics 2017 Apr;14(2):284-297. [doi: 10.1007/s13311-017-0519-x] [Medline: 28281173]
6. Kalidindi N, Or R, Babak S, Mason W. Molecular classification of diffuse gliomas. Can J Neurol Sci 2020 Jul;47(4):464-473. [doi: 10.1017/cjn.2020.10] [Medline: 31918786]

XSL•FO
RenderX

7. Antonelli M, Poliani PL. Adult type diffuse gliomas in the new 2021 WHO Classification. Pathologica 2022 Dec;114(6):397-409. [doi: 10.32074/1591-951X-823] [Medline: 36534419]
8. Hollon T, Jiang C, Chowdury A, et al. Artificial-intelligence-based molecular classification of diffuse gliomas using rapid, label-free optical imaging. Nat Med 2023 Apr;29(4):828-832. [doi: 10.1038/s41591-023-02252-4] [Medline: 36959422]
9. Koski E, Murphy J. AI in healthcare. Stud Health Technol Inform 2021 Dec 15;284:295-299. [doi: 10.3233/SHTI210726] [Medline: 34920529]
10. van Kempen EJ, Post M, Mannil M, et al. Accuracy of machine learning algorithms for the classification of molecular features of gliomas on MRI: a systematic literature review and meta-analysis. Cancers (Basel) 2021 May 26;13(11):2606. [doi: 10.3390/cancers13112606] [Medline: 34073309]
11. Farahani S, Hejazi M, Moradizeyveh S, Di Ieva A, Fatemizadeh E, Liu S. Diagnostic accuracy of deep learning models in predicting glioma molecular markers: a systematic review and meta-analysis. Diagnostics (Basel) 2025 Mar 21;15(7):797. [doi: 10.3390/diagnostics15070797] [Medline: 40218147]
12. Puustinen S, Vrzáková H, Hyttinen J, et al. Hyperspectral imaging in brain tumor surgery-evidence of machine learning-based performance. World Neurosurg 2023 Jul;175:e614-e635. [doi: 10.1016/j.wneu.2023.03.149] [Medline: 37030483]
13. Singh J, Sahu S, Mohan T, et al. Current status of DNA methylation profiling in neuro-oncology as a diagnostic support tool: a review. Neuro oncol Pract 2023 Dec;10(6):518-526. [doi: 10.1093/nop/npad040] [Medline: 38009119]
14. Liu Y, Wu M. Deep learning in precision medicine and focus on glioma. Bioeng Transl Med 2023 Sep;8(5):e10553. [doi: 10.1002/btm2.10553]
15. Zhao J, Huang Y, Song Y, et al. Diagnostic accuracy and potential covariates for machine learning to identify IDH mutations in glioma patients: evidence from a meta-analysis. Eur Radiol 2020 Aug;30(8):4664-4674. [doi: 10.1007/s00330-020-06717-9] [Medline: 32193643]
16. Lv Q, Liu Y, Sun Y, Wu M. Insight into deep learning for glioma IDH medical image analysis: a systematic review. Medicine (Baltimore) 2024 Feb 16;103(7):e37150. [doi: 10.1097/MD.0000000000037150] [Medline: 38363910]
17. Guha A, Halder S, Shinde SH, et al. How does deep learning/machine learning perform in comparison to radiologists in distinguishing glioblastomas (or grade IV astrocytomas) from primary CNS lymphomas? A meta-analysis and systematic review. Clin Radiol 2024 Jun;79(6):460-472. [doi: 10.1016/j.crad.2024.03.007] [Medline: 38614870]
18. McInnes MDF, Moher D, Thombs BD, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA 2018 Jan 23;319(4):388-396. [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
19. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
20. Wolff RF, Moons KGM, Riley RD, et al. Probast: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019 Jan 1;170(1):51-58. [doi: 10.7326/M18-1376] [Medline: 30596875]
21. Albuquerque T, Fang ML, Wiestler B, et al. Multimodal context-aware detection of glioma biomarkers using MRI and WSI. 2023 Presented at: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops; Oct 8-12, 2023; Vancouver, BC, Canada p. 157-167. [doi: 10.1007/978-3-031-47425-5_15]
22. Chitnis SR, Liu S, Dash T, et al. Domain-specific pre-training improves confidence in whole slide image classification. 2023 Presented at: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Jul 24-27, 2023. [doi: 10.1109/EMBC40787.2023.10340659]
23. Cui D, Liu Y, Liu G, Liu L. A multiple-instance learning-based convolutional neural network model to detect the IDH1 mutation in the histopathology images of glioma tissues. J Comput Biol 2020 Aug;27(8):1264-1272. [doi: 10.1089/cmb.2019.0410] [Medline: 31905004]
24. Despotovic V, Kim SY, Hau AC, et al. Glioma subtype classification from histopathological images using in-domain and out-of-domain transfer learning: an experimental study. Heliyon 2024 Mar 15;10(5):e27515. [doi: 10.1016/j.heliyon.2024.e27515] [Medline: 38562501]
25. Fang Z, Liu Y, Wang Y, et al. Deep learning predicts biomarker status and discovers related histomorphology characteristics for low-grade glioma. arXiv. Preprint posted online on Oct 11, 2023. [doi: 10.48550/arXiv.2310.07464]
26. Faust K, Lee MK, Dent A, et al. Integrating morphologic and molecular histopathological features through whole slide image registration and deep learning. Neurooncol Adv 2022;4(1):vdac001. [doi: 10.1093/noajnl/vdac001] [Medline: 35156037]
27. Hewitt KJ, Löffler CML, Muti HS, et al. Direct image to subtype prediction for brain tumors using deep learning. Neurooncol Adv 2023;5(1):vdad139. [doi: 10.1093/noajnl/vdad139] [Medline: 38106649]
28. Innani S, Baheti B, Nasrallah MP, Bakas S. Weakly supervised IDH-status glioma classification from H&E-stained whole slide images. 2024 Presented at: 2024 IEEE International Symposium on Biomedical Imaging (ISBI); May 27-30, 2024. [doi: 10.1109/ISBI56570.2024.10635869]
29. Jiang S, Zanazzi GJ, Hassanpour S. Predicting prognosis and IDH mutation status for patients with lower-grade gliomas using whole slide images. Sci Rep 2021 Aug 19;11(1):16849. [doi: 10.1038/s41598-021-95948-x] [Medline: 34413349]

XSL•FO
RenderX

30. Jungo P, Hewer E. Code-free machine learning for classification of central nervous system histopathology images. J Neuropathol Exp Neurol 2023 Feb 21;82(3):221-230. [doi: 10.1093/jnen/nlac131] [Medline: 36734664]
31. Kim GJ, Lee T, Ahn S, Uh Y, Kim SH. Efficient diagnosis of IDH-mutant gliomas: 1p/19qNET assesses 1p/19q codeletion status using weakly-supervised learning. NPJ Precis Onc 2023;7(1):94. [doi: 10.1038/s41698-023-00450-4]
32. Krebs O, Agarwal S, Tiwari P. Self-supervised deep learning to predict molecular markers from routine histopathology slides for high-grade glioma tumors. Presented at: Proceedings Volume 12471, Medical Imaging 2023: Digital and Computational Pathology; Feb 19-23, 2023. [doi: 10.1117/12.2653929]
33. Li Z, Cong Y, Chen X, et al. Vision transformer-based weakly supervised histopathological image analysis of primary brain tumors. iScience 2023 Jan;26(1):105872. [doi: 10.1016/j.isci.2022.105872]
34. Liechty B, Xu Z, Zhang Z, et al. Machine learning can aid in prediction of IDH mutation from H&E-stained histology slides in infiltrating gliomas. Sci Rep 2022 Dec 31;12(1):22623. [doi: 10.1038/s41598-022-26170-6] [Medline: 36587030]
35. Liu X, Hu W, Diao S, Abera DE, Racoceanu D, Qin W. Multi-scale feature fusion for prediction of IDH1 mutations in glioma histopathological images. Comput Methods Programs Biomed 2024 May;248:108116. [doi: 10.1016/j.cmpb.2024.108116] [Medline: 38518408]
36. Liu S, Shah Z, Sav A, et al. Isocitrate dehydrogenase (IDH) status prediction in histopathology images of gliomas using deep learning. Sci Rep 2020;10(1):7733. [doi: 10.1038/s41598-020-64588-y]
37. Nakagaki R, Debsarkar SS, Kawanaka H, Aronow BJ, Prasath VBS. Deep learning-based IDH1 gene mutation prediction using histopathological imaging and clinical data. Comput Biol Med 2024 Sep;179:108902. [doi: 10.1016/j.compbiomed.2024.108902] [Medline: 39038392]
38. Pei L, Jones KA, Shboul ZA, Chen JY, Iftekharuddin KM. Deep neural network analysis of pathology images with integrated molecular data for enhanced glioma classification and grading. Front Oncol 2021;11:668694. [doi: 10.3389/fonc.2021.668694] [Medline: 34277415]
39. Rathore S, Iftikhar M, Nasrallah M, Gurcan M, Rajpoot N, Mourelatos Z. Tmod-35. Prediction of overall survival, and molecular markers in gliomas via analysis of digital pathology images using deep learning. Neuro Oncol 2019 Nov 11;21(Supplement_6):vi270. [doi: 10.1093/neuonc/noz175.1134]
40. Wang W, Zhao Y, Teng L, et al. Neuropathologist-level integrated classification of adult-type diffuse gliomas using deep learning from whole-slide pathological images. Nat Commun 2023;14(1):6359. [doi: 10.1038/s41467-023-41195-9]
41. Wang D, Liu C, Wang X, et al. Automated machine-learning framework integrating histopathological and radiological information for predicting IDH1 mutation status in glioma. Front Bioinform 2021;1:718697. [doi: 10.3389/fbinf.2021.718697] [Medline: 36303770]
42. Zhao Y, Wang W, Ji Y, et al. Computational pathology for prediction of isocitrate dehydrogenase gene mutation from whole slide images in adult patients with diffuse glioma. Am J Pathol 2024 May;194(5):747-758. [doi: 10.1016/j.ajpath.2024.01.009] [Medline: 38325551]
43. Farahani S, Hejazi M, Tabassum M, Di Ieva A, Mahdavifar N, Liu S. Diagnostic performance of deep learning for predicting glioma isocitrate dehydrogenase and 1p/19q co-deletion in MRI: a systematic review and meta-analysis. Eur Radiol 2025 Aug 16. [doi: 10.1007/s00330-025-11898-2] [Medline: 40817944]
44. Chen X, Lei J, Wang S, Zhang J, Gou L. Diagnostic accuracy of a machine learning-based radiomics approach of MR in predicting IDH mutations in glioma patients: a systematic review and meta-analysis. Front Oncol 2024;14:1409760. [doi: 10.3389/fonc.2024.1409760]
45. Redlich JP, Feuerhake F, Weis J, et al. Applications of artificial intelligence in the analysis of histopathology images of gliomas: a review. Npj Imaging 2024 Jul 1;2(1):16. [doi: 10.1038/s44303-024-00020-8] [Medline: 40603567]
46. Nobel SMN, Swapno S, Islam MB, et al. A novel mixed convolution transformer model for the fast and accurate diagnosis of glioma subtypes. Adv Intell Syst 2025 May;7(5):2400566. [doi: 10.1002/aisy.202400566]
47. Henry EU, Emebo O, Omonhinmin CA. Vision transformers in medical imaging: a review. arXiv. Preprint posted online on Nov 18, 2022. [doi: 10.48550/arXiv.2211.10043]
48. Alleman K, Knecht E, Huang J, Zhang L, Lam S, DeCuypere M. Multimodal deep learning-based prognostication in glioma patients: a systematic review. Cancers (Basel) 2023 Jan 16;15(2):545. [doi: 10.3390/cancers15020545] [Medline: 36672494]
49. d'Este SH, Nielsen MB, Hansen AE. Visualizing glioma infiltration by the combination of multimodality imaging and artificial intelligence, a systematic review of the literature. Diagnostics (Basel) 2021 Mar 25;11(4):592. [doi: 10.3390/diagnostics11040592] [Medline: 33806195]
50. Byeon Y, Park YW, Lee S, et al. Interpretable multimodal transformer for prediction of molecular subtypes and grades in adult-type diffuse gliomas. NPJ Digit Med 2025 Mar 5;8(1):140. [doi: 10.1038/s41746-025-01530-4] [Medline: 40044878]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
CNN: convolutional neural network
CNS: central nervous system

XSL•FO
RenderX

DBTA: Digital Brain Tumor Atlas
DenseNet121: densely connected convolutional network 121
IDH: isocitrate dehydrogenase
MIL: multiple instance learning
MRI: magnetic resonance imaging
PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Extension for Diagnostic Test Accuracy
PROBAST: Prediction Model Risk of Bias Assessment Tool
PROSPERO: International Prospective Register of Systematic Reviews
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2
WHO: World Health Organization

Edited by A Schwartz; submitted 01.Jun.2025; peer-reviewed by BS Ibrahim, K Dhou, Y Lu; accepted 06.Jan.2026; published 13.Mar.2026.

Please cite as:
Almaabreh O, Al-Dafi R, Tabassum A, Othman A, Abd-alrazaq A
The Performance of Artificial Intelligence in Classifying Molecular Markers in Adult-Type Gliomas Using Histopathological Images: Systematic Review
J Med Internet Res 2026;28:e78377
URL: https://www.jmir.org/2026/1/e78377
doi:10.2196/78377

© Obada Almaabreh, Rukaya Al-Dafi, Aliya Tabassum, Ahmad Othman, Alaa Abd-alrazaq. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

eHealth Literacy and Type 2 Diabetes Prevention Among At-Risk Populations: Mechanistic Systematic Review Using Theory-Driven Thematic Analysis

Jingyi Li[1,2], MA; Arina Anis Azlan[1,3], PhD; Nurzihan Hassim[1], PhD; Yuan Wang[1,4], MSocSci; Ruina Guo[1], MJC

[1]Centre for Research in Media and Communication, Faculty of Social Sciences and Humanities, National University of Malaysia, UKM, Selangor, Malaysia

[2]Shanxi Technology and Business University, Taiyuan, Shanxi, China

[3]Komunikasi Kesihatan (Healthcomm) - UKM Research Group, National University of Malaysia, Selangor, Malaysia

[4]Yuncheng University, Yuncheng, Shanxi, China

Corresponding Author:
Arina Anis Azlan, PhD
Centre for Research in Media and Communication, Faculty of Social Sciences and Humanities, National University of Malaysia, UKM, Selangor, Malaysia

Abstract

Background: Type 2 diabetes (T2D) is emerging as a growing global public health crisis. Early and effective interventions can reduce T2D incidence among at-risk populations. Compared with traditional approaches, digital health technologies offer promising opportunities for prevention, with eHealth literacy (eHL) emerging as a critical determinant of digital prevention outcomes.

Objective: This systematic review aims to synthesize and explain the pathways and mechanisms through which eHL supports T2D prevention among at-risk populations.

Methods: We searched Scopus, Web of Science, and PubMed databases for English-language original research published between January 1, 2000, and August 14, 2025. Studies included were prevention research involving eHL engagement among populations at risk for T2D. Nonoriginal literature, such as editorials and abstracts, as well as research protocols, was excluded. The findings were synthesized using a thematic analysis approach, integrating the Theoretical Domains Framework with the eHL model. Two reviewers independently screened literature and extracted data, and discrepancies were resolved by a third reviewer. The Mixed Methods Appraisal Tool was used to assess risk of bias.

Results: This review included 28 studies (n=13,100), mostly quantitative and published within the past decade, targeting people with prediabetes, prior gestational diabetes, and overweight/metabolic risk. Study quality was moderate to high (Mixed Methods Appraisal Tool 60% - 100%) with no high risk of bias. eHL supported prevention mainly through knowledge (28/28), behavioral regulation (16/28), social influences (15/28), environmental resources (12/28), and goals (11/28), while emotions, memory, attention, decision process, and beliefs about competence were rarely addressed. Health literacy (27/28), information literacy (20/28), and communicative eHL (20/28) were most common; critical eHL and media literacy were not addressed. Studies reported positive outcomes: high engagement, weight loss (≥5%), improved glycemic markers, and enhanced lifestyle behaviors.

Conclusions: This is the first systematic exploration of eHL mechanism pathways in T2D prevention via theoretical mapping. We found interventions yield positive effects despite highly uneven mechanism application: extant research relies excessively on knowledge and behavioral pathways while underemphasizing emotional support, autonomy, and critical evaluation—factors linked to long-term adherence. We provide a mechanism-based framework and identify critical gaps, including the absence of focus on critical eHL and media literacy. This review is limited by substantial variation across studies that did not allow for meta-analysis and by the limited evidence base on eHL. Future interventions should explore and test emotional and autonomy support, information discernment training, and accessibility optimization in T2D prevention. These comprehensive, equity-focused intervention approaches will help ensure that eHL becomes a truly effective public health tool that benefits everyone, especially at-risk and vulnerable populations.

Trial Registration: PROSPERO CRD42025630395; https://www.crd.york.ac.uk/PROSPERO/view/CRD42025630395

(J Med Internet Res 2026;28:e77788) doi:10.2196/77788

KEYWORDS

eHealth literacy; digital health literacy; type 2 diabetes prevention; type 2 diabetes mellitus; self-management; chronic disease self-care

XSL•FO
RenderX

Introduction

Background

Type 2 diabetes (T2D) is becoming a growing public health crisis worldwide [1], accounting for 90% to 95% of all diabetes cases [2]. The global number of adults with diabetes is expected to surge from 589 million in 2024 to 853 million in 2050 [2], placing a huge burden on health care systems worldwide.

The development of T2D is driven by a variety of factors, among which unhealthy dietary habits and a sedentary lifestyle are the main triggers [3,4]. In addition, at-risk factors include overweight, obesity, insulin resistance, metabolic syndrome, dyslipidemia, and hypertension [5-7], genetics [8,9], nonalcoholic fatty liver disease, history of gestational diabetes [10], chronic stress [11], depression [12], sleep deprivation [13], etc.

Long-term exposure to or persistent presence of the aforementioned risk factors may increase the risk of developing dysglycemia and progressing to prediabetes. Prediabetes is an independent predictor of progression to T2D, and without timely intervention, individuals with prediabetes face a significantly increased risk of developing diabetes. Therefore, populations exhibiting one or more of these characteristics constitute the “at-risk population” for T2D [14].

Focusing prevention efforts on this at-risk population holds significant public health implications. Approximately 5% - 10% of individuals with prediabetes progress to diabetes annually, with a lifetime risk reaching up to 70% [15]. Notably, prediabetes and related risk states are not irreversible. Research indicates that individuals can regain normal glucose regulation through intensive lifestyle interventions [16], offering a critical window for preventive action. Furthermore, targeting prevention efforts toward at-risk populations demonstrates favorable cost-effectiveness in economic evaluations [17]. Despite the significant benefits of preventive interventions, the landmark Diabetes Prevention Program (DPP) study proved that intensive lifestyle intervention can reduce the risk of T2D by 58% in at-risk populations [18]. However, traditional face-to-face DPP programs face severe challenges, such as insufficient geographical and resource distribution limiting accessibility [19], high promotion cost [20], and low completion rate [21], which make it difficult to achieve comprehensive coverage. These limitations highlight the urgent need for innovative prevention models.

Digital health technologies offer promising solutions to overcome traditional barriers to prevention [22]. Interventions based on digital platforms have shown potential to improve access and reduce implementation costs [23]. However, with the rapid growth of web-based health information, there are significant differences in information quality and availability [24]. Some information may even mislead users and cause anxiety, which in turn has a negative impact on health decision-making [25].

In this digital context, eHealth literacy (eHL) has become a key factor in determining the effectiveness of digital prevention. Norman and Skinner [26] defined eHL as "the ability to seek, discover, understand, and evaluate health information from electronic resources, and to apply the knowledge gained to address or solve health problems” and divided it into 6 dimensions. In 2018, Paige et al [27] proposed a conceptual model in which eHL is described as an interpersonal skill and divided into 4 dimensions. Recent conceptual updates emphasize “the ability to use digital technologies to achieve health goals in effective, safe and beneficial ways” [28], reflecting the evolving nature of eHL in digital health settings.

Studies have confirmed that eHL significantly affects diabetes prevention behaviors [29]. High eHL can promote a healthy diet, regular exercise, and good sleep [30], and improve the use of the medical system [31]. Lack of eHL may hinder disease prevention and health management [32].

Despite the increasing importance of eHL, there is still a lack of mechanistic understanding and theory-driven synthesis regarding how eHL influences the prevention of T2D in at-risk populations. On the one hand, existing systematic reviews have generally studied the relationship between eHL and health behavior [33,34]; on the other hand, they have only examined the impact of digital health intervention or telemedicine on prevention, lacking attention to literacy. Or focus on disease management in patients with established disease [35-40].

This study aims to address this research gap by conducting a mechanistic systematic review using theory-driven thematic analysis. The main objective was to study the effect path of eHL on the prevention of T2D in people at risk of T2D and the participation ways of different dimensions. The findings are expected to provide evidence-based guidance to health educators, intervention designers, and policy makers to develop T2D prevention strategies tailored to at-risk populations.

Research Questions

This mechanistic systematic review aimed to explore the pathways of eHL in the prevention of T2D among at-risk populations to fill the existing literature gap. Specifically, this study sought to answer the following research question (RQ):

- RQ1. What are the mechanistic actions of eHealth literacy in type 2 diabetes prevention among at-risk populations?
- RQ2. How do the dimensions of eHealth literacy facilitate type 2 diabetes prevention among at-risk populations?

Methods

Overview

The study protocol for this systematic review was registered with PROSPERO (registration number CRD42025630395 on January 16, 2025). This study was written and reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guideline [41] (Checklist 1) and the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search) [41] extension for reporting literature searches (Checklist 2).

Data Resources and Search Strategy

Scopus, Web of Science (WOS), and PubMed were selected as the databases for this review. The literature search was

XSL•FO
RenderX

conducted from January 1, 2000, to August 14, 2025. The year 2000 was chosen as the starting year because the concept of eHealth was formally proposed in 2001 [42]. Although the term eHL was proposed by Norman and Skinner in 2006 [26], earlier studies may have used related terms such as digital health literacy, health information literacy, etc, so a broad time frame was used to ensure the comprehensibility of the search.

In addition to database searching, we did not search trial or study registries; conduct any targeted searching or browsing of online or print resources; perform reference list screening or citation tracking; obtain additional studies or data by contacting authors, experts, or manufacturers; or use any other information sources or search methods. The search strategy was developed de novo, without adapting or reusing strategies from existing reviews, and no search update was performed (Multimedia Appendix 1). The strategy was peer-reviewed by an information scientist with extensive expertise in systematic review methods and database search strategy design.

Search terms were generated around the research topic. Two researchers (WY and RG) independently screened and identified relevant studies using retrieval strings based on the Population, Interest/Intervention, and Context model combined with Boolean logic. Disagreements were resolved by discussion and, if necessary, consensus was reached by consulting other investigators (AAA and ZZH). All keywords were searched in the database. A comprehensive list of search queries is presented in Table 1. Individual search strings for each database are shown in Checklist 2.

Table . Keywords search strings.

Concept	Keywords/search terms
eHealth	“eHealth” OR “e-health” OR “mHealth” OR “m-health” OR “mobile health” OR “digital health” OR “internet health” OR “telehealth” OR “telemedic*” OR “computer health” OR “computer-based health” OR “web health” OR “web-based health” OR “web based health” OR “online health” OR “online-based health” OR “online based health” OR “health” OR “media” OR “computer” OR “virtual” OR “intelligent” OR “tech-nolog*” OR “Web 2.0” OR “web based” OR “online based” OR “phone*” OR “app*”
Literacy	“literac*” OR “comprehension” OR “skill*” OR “ability” OR “knowledge” OR “seek*“OR “find*” OR “search*” OR “understand*” OR “apprais*” OR “evaluat*” OR “assess*” OR “access*” OR “apply” OR “use*” OR “communication” OR “interaction” OR “efficacy” OR “engagement”
Type 2 diabetes prevention	“diabetes prevent*” OR “diabetes prevention” OR “type 2 diabetes pre-vent*” OR “type II diabetes prevent*” OR “T2D prevent*” OR “T2DM prevent*” OR “prediabetes prevent*” OR “non-insulin-dependent diabetes mellitus prevent*” OR “adult-onset diabetes prevent*” OR “NIDDM prevent*”
At-risk population	“at-risk” OR “high-risk” OR “risk population” OR “prediabetes” OR “overweight” OR “fat” OR “obesity” OR “sedentary behavi*” OR “sedentary lifestyle*” OR “physical inactiv*” OR “lack of exercise” OR “genetics” OR “family history” OR “gestational diabetes” OR “hypertens*” OR “dyslipid*” OR “insulin resistance” OR “acanthosis nigricans” OR “non-alcoholic fatty liver disease” OR “NAFLD” OR “metabolic syn-drome” OR “chronic stress” OR “depression” OR “poor sleep” OR “western diet” OR “unhealthy diet” OR “suboptimal diet”

Selection Criteria

Papers included in the review had to meet prespecified inclusion criteria. First, selected articles should be original research articles written in English and published between January 1, 2000, and August 14, 2025. Other literature types, such as reviews, editorials, conference papers, letters, briefings, reviews, book chapters, monographs, or research protocols, were excluded from this literature review. Second, people at risk for T2D. Patients included those with prediabetes, family history of diabetes, overweight or obesity, history of gestational diabetes, sedentary or lack of exercise, hypertension, dyslipidemia, insulin resistance, acanthosis nigricans, nonalcoholic fatty liver disease, metabolic syndrome, chronic stress, depression, inadequate sleep, or unhealthy diet. In addition, studies were excluded if they were (1) not specifically targeting T2D, but referring to diabetes or type 1 diabetes; (2) conducted via nonelectronic means such as text messages and phone calls; (3) only describing barriers or benefits of online prevention without explicitly mentioning prevention measures related to eHL; and (4) irrelevant to T2D prevention. Detailed inclusion and exclusion criteria are provided in Table 2.

XSL•FO
RenderX

Table . Inclusion and exclusion criteria.

Components	Inclusion	Exclusion
Population	Individuals at risk for type 2 diabetes, including those with prediabetes: a family history of diabetes; overweight or obesity; a history of gestational diabetes; a sedentary lifestyle or physical inactivity; hypertension; dyslipidemia; insulin resistance; acanthosis nigricans; nonalcoholic fatty liver disease; metabolic syndrome; chronic stress; depression; insufficient sleep; or an unhealthy diet.	• Refers to diabetes without specifying people with type 2 diabetes, people with type 1 diabetes, and people without risk of type 2 diabetes.
Interest	Studies explicitly involving eHealth literacy in type 2 diabetes prevention among at-risk populations, either as an overall construct or through specific dimensions, using digital or online approaches.	• Not specifically targeting type 2 diabetes, but referring to diabetes or type 1 diabetes. • Nonelectronic means such as text messages and phone calls.
Context	Studies must report the impact of eHealth literacy on the prevention of type 2 diabetes among at-risk populations.	• Only describe barriers or benefits of online prevention, without explicitly mentioning prevention related to eHealth literacy. • Not associated with type 2 diabetes prevention.
Timeline	January 1, 2000, to August 14, 2025	• Before 2000.
Article type	Primary research journal articles on quantitative research, qualitative research, and mixed methods research	• Reviews, editorials, conference papers, letters, notes, commentaries, book chapters, books, or study protocols.
Language	English	• Non-English.

Systematic Review Process

All papers retrieved from WOS, Scopus, and PubMed databases were imported into Rayyan Citation Manager [43], and duplicates were removed. Two authors (WY and RG) independently screened studies according to prespecified inclusion and exclusion criteria. Studies that may meet the criteria will be screened for full text for a final decision on inclusion in our review. Any inconsistencies were resolved through discussion and consensus among the reviewers. If consensus was not reached, a third reviewer (AAA) made the final decision.

Risk of Bias Assessment and Reporting Bias Assessment

The risk of bias of all included studies was independently assessed by 2 reviewers (WY and RG) and cross-validated by each other. Any disagreements were resolved by discussion with a third evaluator (AAA). The Mixed Methods Appraisal Tool (MMAT, version 2018) was used to assess the quality of the eligible studies. This tool is primarily used to assess the quality of mixed methods studies and also includes the assessment of qualitative and quantitative methods. MMAT included 2 screening questions for each study and then 5 questions for each design, depending on the category of study design. Each question was rated as “no,” “yes,” and “unable to judge.” The risk of bias was determined for each study based on the percentage of questions answered “yes,” with scores ranging from 0% to 100% [44]. Prior to conducting the assessment, reviewers completed the official MMAT online training tutorial developed by Pace et al [45] to ensure consistent and informed use of the tool.

Theoretical Underpinnings

This systematic review uses the Theoretical Domains Framework (TDF) and the eHL model as its theoretical foundation for constructing a coding framework, organizing and interpreting evidence, and identifying behavioral mechanisms and literacy dimensions that are adequately covered versus relatively underrepresented in current intervention designs during the thematic synthesis process.

The TDF encompasses 14 domains, including knowledge, skills, social or professional roles and identities, beliefs about competence, optimism, reinforcement, intentions, goals, memory, attention and decision processes, environmental background and resources, social influences, and emotions and behavioral regulation [46]. This framework provides a systematic and actionable theoretical classification to elucidate the potential mechanisms through which eHL may influence diabetes prevention-related behaviors, further pointing to intervention targets for optimization.

To comprehensively describe the dimensional coverage and mechanistic actions of eHL in the included studies, this research uses 2 complementary eHL models to balance breadth and depth. At the breadth level, the Norman and Skinner framework conceptualizes eHL as 6 foundational literacies (traditional literacy, health literacy, information literacy, scientific literacy, computer literacy, and media literacy) [26], facilitating cross-comparisons of eHL dimensional coverage across studies.

XSL•FO
RenderX

At the depth level, the interactive framework of Paige et al [27] refines eHL into functional eHL, communicative eHL, critical eHL, and translational eHL, emphasizing interaction, comprehension, and application transfer within specific digital health contexts. The integration of these models enables this review to provide a more systematic interpretation and comparative analysis of eHL's multidimensional mechanisms within the context of T2D prevention.

Data Extraction and Analysis

To gain a comprehensive understanding of the essential characteristics and trends of the included studies, we developed a set of standardized data extraction tables for summarizing key study characteristics. The table collected detailed information from the included literature, including author, year of publication, country/region, age of participants, study methods, and main intervention outcomes.

In addition, we use thematic analysis as the main method to identify patterns or themes in qualitative data [47]. It is widely regarded as one of the most appropriate methods for integrating mixed study designs [48]. Thematic analysis is a descriptive strategy that enables flexible reduction of data and can be used in combination with other data analysis techniques [49].

Therefore, we conducted a thematic analysis to identify common themes that emerged in the data. These themes were derived from the objectives of this review.

First, addressing RQ 1, two reviewers (WY and RG) classified relevant findings by mapping intervention components and reported implementation strategies to the 14 domains of the TDF [46], to summarize which behavior change mechanisms were reflected across studies.

Second, addressing RQ 2, we examined how different dimensions of eHL contribute to the prevention process by mapping eHL-related content from included studies onto 2 complementary eHL models [26,27].

In addition, we used the same thematic analysis method to generalize the intervention outcomes reported by the studies into weight change, diabetes-related measures, behavior change, self-efficacy, and engagement and acceptability.

For studies independently reviewed by 2 researchers (WY and RG), if consensus cannot be reached, the final decision will be made by a third reviewer (AAA).

Results

Study Screen

Overview

In accordance with PRISMA 2020 reporting guidelines, we conducted a comprehensive search on WOS, Scopus, and PubMed to find 5409 articles published in English. During the initial title and abstract screening process, we identified and removed 2593 duplicate records, resulting in 2816 articles progressing to the next stage.

After further review of titles and abstracts, 2702 articles were excluded due to lack of relevance to eHL or T2D prevention in at-risk populations, leaving 114 articles for full-text screening. Of these, 86 articles were subsequently excluded. Ultimately, 28 studies were included. The PRISMA screening flow chart is shown in Figure 1 for details.

XSL•FO
RenderX

Figure 1. Flowchart of systematic review search results. eHL: eHealth literacy.

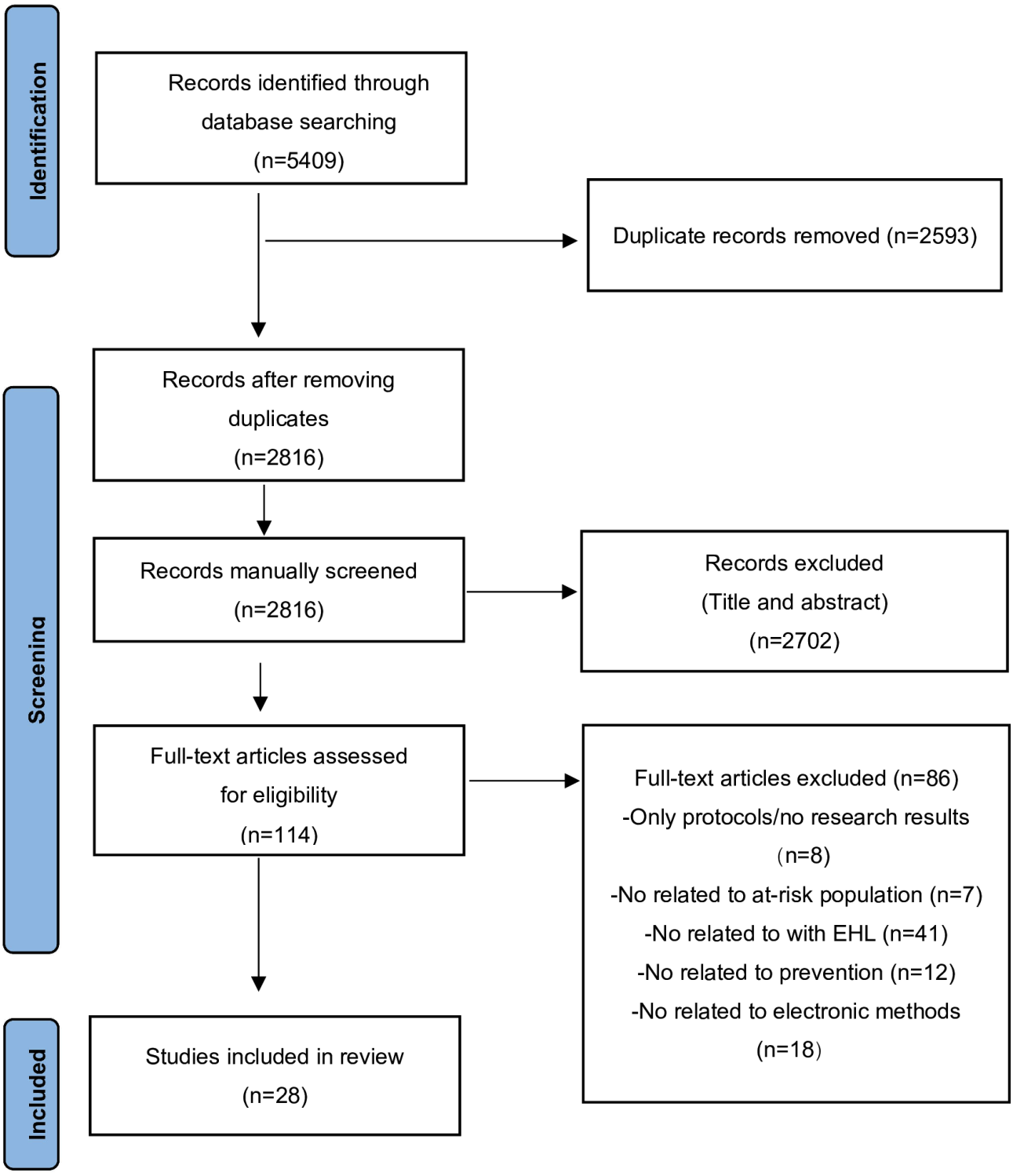

Description of Studies

This review includes 28 studies that met the inclusion criteria, with their characteristics detailed in Table 3.

The majority of studies were conducted within the last decade (27/28), with a particular focus on 2019, 2022, and 2023 [50-62]. One such study was conducted in 2010 [63].

A total of 86% (24/28) of the studies were from the United States. In addition, there was one study from Canada (n=1), one study from Malaysia (n=1), one study from Taiwan, China (n=1), and one study from Singapore (n=1).

In terms of population at-risk, 23 studies focused on people with prediabetes [50-72], 3 studies focused on women with a history of gestational diabetes [59,63,73], and 10 studies examined individuals who were overweight, obese, had excessive BMI, or had metabolic syndrome [53,56,58,59,67-69,71,74,75]. In addition, one study included individuals who had previously given birth to a baby weighing

XSL•FO
RenderX

more than 9 lb (about 3.7 kg) and had hypertension or dyslipidemia [63].

Seven studies [51,53,54,58,60,64,76] specifically addressed special populations, including female veterans [64], low-income individuals enrolled in Medicaid [51], self-identified as Asian [53], adolescents [54], older at-risk individuals [58,76], and Chinese Americans [60].

A total of 18 studies used web-based DPP interventions [51,53,55,57-60,62-64,66,67,69,70,72,74,76,77], and one study specifically recruited a population not previously enrolled in a DPP program [52].

A total of 28 intervention studies were ultimately included, encompassing 13,100 participants. Intervention durations ranged from 12 weeks to 3 years, with participant ages spanning adolescents to older adults (approximately 14.6 to 69.2 y).

XSL•FO
RenderX

Table . Characteristics and key findings of included studies on type 2 diabetes prevention among at-risk populations.

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Vadheim et al (2010) [63]	United States	Quantitative nonrandomized	29	A history of prediabetes, gestational diabetes mellitus (GDM), or delivery of a baby weighing more than 9 pounds, hypertension, or dyslipidemia	51.4	81.4% (22/27) women	Remote videoconference	16 weeks	A total of 88% (14/16) of participants completed the 16-week program; More than 40% (12/29) of participants achieved the 7% weight loss goal.
Moin et al (2015) [64]	United States	Qualitative study	15	Female veterans with prediabetes	56.8	100% (15/15) women	Online virtual platform, pedometer, tracker, wireless scale	16 weeks	Perception: the plan was well aligned with perceived health needs; The program is convenient; Integration of projects into daily life; A sense of responsibility; Hate records; Prefer traditional offline; Difficult to understand; Mean weight loss was 5.24%.
Sepah et al (2015) [74]	United States	Quantitative nonrandomized	220	BMI>24 kg/m^2 over 18 years of age	43.6	17.3% (38/220) men	Online web platform, digital tracking tools (wireless scale and pedometer tracking)	2 years	Users had significant reductions in body weight and HbA_{1c}[a] that were sustained 2 years later.

XSL•FO
RenderX

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Block et al (2015) [65]	United States	RCT[b]	339	Prediabetes	55	68.7% (233/339) men	Alive-PD app, personalized website, and interactive email delivery	6 months	Reductions in HbA_{1c}, fasting glucose, and body weight were significantly better in the intervention group than in the control group. Health behaviors and lifestyle improved significantly, and program participation was high.
Sepah et al (2017) [66]	United States	Quantitative nonrandomized	220	Prediabetes	__[c]	82.7% (182/220) women	An Internet-connected desktop computer or mobile device	3 years	There was a significant decrease in body weight and glycated hemoglobin at week 16, year 1, and year 2. At the third year, there was a significant reduction in body weight and HbA_{1c}.
Stein and Brooks (2017) [75]	United States	Quantitative nonrandomized	70	Overweight and obese adults	47	74.5% (35/47) women	Artificial Intelligence Health Coaching app (HCAI app)	15 weeks	A total of 75.7% (53/70) of the users lost weight in this plan, and the percentage of healthy diets increased by 31%. The response rate of the user trust survey is 100%.

XSL•FO
RenderX

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Michaelides et al (2018) [67]	United States	Quantitative nonrandomized	59	Overweight or obese adults with prediabetes	51	81% (48/59) women	The Noom application	65 weeks	Weight and BMI were significantly reduced, with 53% (25/47) of the completion group and 66% (21/32) of the maintenance group achieving ≥5% weight loss.
Everett et al (2018) [68]	United States	Quantitative nonrandomized	55	Adults with prediabetes and BMI between 24 and 40 kg/m^2	55	60% (33/55) women	Noom application, Sweetch application, and Digital Weight Scales (DBWS)	3 months	The retention rate and acceptance of the study were relatively high. Participants lost 1.6 kilograms in weight, their BMI decreased by 0.6 $kilograms/m^2$, their waist circumference reduced by 1.4 centimeters, and their physical activity significantly increased. Fasting blood glucose showed no significant change compared with the baseline.

XSL•FO
RenderX

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Castro Sweet et al (2018) [69]	United States	Quantitative nonrandomized	501	Have prediabetes or metabolic syndrome	68.8	64% women[d]	Connected devices, wireless scales	12 months	The participants lost an average of 6.5% in weight over 16 weeks. At the 6th month, the weight decreased by 8.0%, and at the 12th month, it decreased by 7.5%. The participants' blood sugar control was improved and their total cholesterol was reduced. Participants' self-reported improvements in well-being, depression, and self-care.
Srivastava et al (2019) [50]	United States	Quantitative nonrandomized	10	Prediabetes	56.1	70% (7/10) women	Cloud-based type 2 diabetes prevention module	6 months	The success rate of the module is approximately 60%. The average weight loss of users is approximately 9.0%, and their weekly physical activity time is 192 minutes.
Kim et al (2019) [51]	United States	Quantitative nonrandomized	227	Low-income people with prediabetes enrolled in Medicaid	48.2	81.3% women[d]	Online digital platforms	12 months	The intervention group had an average weight loss of 4.4% and an average BMI change of –1.6 over 12 months.

XSL•FO
RenderX

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Griauzde et al (2019) [52]	United States	Mixed methods	69	People with prediabetes who did not participate in DPP[e]	51.7	64% (44/69) women	Apps, Fitbit devices	12 weeks	8 out of 13 respondents liked some features of the app, and 13 out of 13 respondents reported changes in health behaviors.
Alwashmi et al (2019) [53]	United States	Quantitative nonrandomized	273	Prediabetes, BMI>25 kg/m^2, or self-identified as Asian	54	70.3% (192/273) women	Interactive mobile computing (eg, smartphone applications), wearable tracking devices (eg, activity trackers), remote health monitoring hardware (eg, electronic scales)	4 months	The average weight loss was 6.5%, the average BMI decreased by 1.9 kg/m^2, the frequency of exercise increased by 1.7 days per week, and the monthly absenteeism rate decreased by nearly half a day.
Toro-Ramos et al (2020) [70]	United States	RCT	202	Prediabetes	56.6	71.3% (144/202) women	The Noom application	12 months	The changes in body weight and BMI in the intervention group were significantly reduced, but the effect on HbA_{1c} was not significant in the intention-to-treat analysis.

XSL•FO
RenderX

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Ferrara et al (2020) [77]	United States	RCT	158	People with a risk of more than 9 points were screened by the tool	52.2	72.8% (115/158) women	Canary health online platform	12 months	The online group achieved significant weight loss within 6 months, with an average weight change of –1.97 kilograms. 12 months –1.27 kilograms; At 6 months and 12 months, the average total calorie intake, dietary fat intake, and saturated fat intake of the online group all decreased significantly. However, it did not lead to significant changes in physical activity.
Ekezie et al (2021) [73]	United States	Qualitative study	23	Women with previous Gestational Diabetes Melilitus	—	100% (23/23) women	Baby steps app	11 months	Changes in lifestyle; Increasing physical activity and steps; Promoting competition and communication; Increase motivation; Bring in peer support.
Islam et al (2022) [54]	United States	Qualitative study	14	Adolescents with prediabetes	14.6	36% (5/14) women	Zoom	12 weeks	Participants indicated that the online adolescent health education program was beneficial and applicable to their lives, and gave them a deeper understanding of healthy eating and active lifestyle.

XSL•FO
RenderX

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Katula et al (2022) [55]	United States	RCT	599	Prediabetes	55.4	61.4% (368/599) women	Internet-connected device (laptop, tablet, or smartphone)	12 months	HbA_{1c}, body weight, and cholesterol/high-HDL[f] ratio decreased significantly. Reductions in HbA_{1c} and body weight were greater in the older participants aged ≥65 years.
Lim et al (2022) [56]	Singapore	RCT	148	Adults with prediabetes and BMI ≥23 kg/m^2	53.1	39.9% (59/148) women	nBuddy Diabetes app	6 months	The intervention group was 2.1 times more likely to achieve euglycemia (HbA_{1c}<5.7%) than the control group. Within-group improvements in systolic and diastolic blood pressure, total cholesterol, and low-density lipoprotein (LDL) cholesterol were significant in the intervention group. The intervention group had significantly greater reductions in daily intake of total calories, carbohydrates, total fat, saturated fat, and sugar than the control group.

XSL•FO
RenderX

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Batten et al (2022) [57]	Canada	Quantitative nonrandomized	1095	Prediabetes	53.6	67.7% (741/1095) women	Smartphone applications, wireless scales, wearable tracking devices	12 months	After at least 9 months of participation, the average weight loss was 11.4 pounds, with a rate of 5.5%. Mean weekly physical activity increased to 132.9 minutes.
Fitzpatrick et al (2022) [58]	United States	Quantitative nonrandomized	3904	65-75 years with prediabetes and obesity	69.2	55.2% (2155/3904) women	Electronic patient portal, wireless scale, and pedometer	24 months	The average weight loss was 8.6 pounds, and participation was high.
Collins et al (2023) [59]	United States	Quantitative nonrandomized	2390	Adults with BMI >25 and a history of gestational diabetes or prediabetes	54.9	89.2% (2132/2389) women	Video conferencing, secure online portals	12 months	HbA_{1c} monitoring decreased significantly, and those who completed the program lost an average weight of 5.66%. Participants showed statistically significant changes in their self-reported confidence in their ability to perform all 18 health promotion behaviors.

XSL•FO
RenderX

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Yeh et al (2023) [60]	United States	Mixed methods	13	Chinese Americans with prediabetes	66	85% (11/13) women	Facebook groups, Qualtrics links	12 months	The completion rate was high. The completion rate of the maintenance phase decreased slightly; the weekly modules were generally considered to be clear and easy to understand. All of the participants (100%) liked the online DPP, and the overall satisfaction of the participants was high. Body weight decreased significantly at 2, 3, 4, 8, and 10 months.

XSL•FO
RenderX

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Zahedani et al (2023) [61]	United States	Quantitative nonrandomized	2217	Prediabetes	49	51% women[d]	January AI app, wearing CGM (Freestyle Libre, Abbott), heart rate monitor (Apple Watch or Fitbit)	About 3 months	The GMI[g] was significantly reduced, and the number of prediabetic hyperglycemia events was the largest reduction. Prediabetic individuals lost an average of 2.5 pounds, roughly doubled their physical activity, and increased their heart rate by 6.3 minutes/day; Calorie, carbohydrate, sugar, and saturated fat intakes all decreased, whereas protein, total fat, and dietary fiber intakes increased.
Chung et al (2023) [62]	Taiwan, China	RCT	121	Prediabetes	58.1	52.9% (64/121) women	TCM mHealth app	About 4 months	The TCM mobile health app intervention was superior to the control group in blood glucose control (HbA_{1c} reduction), body constitution (Yang deficiency, phlegm stasis improvement), body energy improvement, health-related quality of life (PCS, MCS), and BMI reduction.

XSL•FO
RenderX

Study	Country	Study design	Sample size	Characteristics of at-risk populations	Average age (years)	Gender	Intervention delivery modality	Time frame	Key findings
Gerber et al (2024) [76]	United States	RCT	20	Older people with excessive BMI	65.4	45% (9/20) women	Zoom, Fitbit Inspire 2 /Fitbit Aria Air, Fitbit App	6 months	Meeting participation was high, with 74% calorie intake compliance; Body weight and BMI changed significantly. The intervention group achieved an 8.4% reduction in body weight at 6 months while BMI decreased by 2.3 units at 3 months.
Ng et al (2025) [71]	Malaysia	RCT	82	Overweight/obese prediabetic patients	49.8	53.7% (44/82) women	Blood glucose meters and wearable fitness trackers, PRIME app, WhatsApp	6 months	The dietary quality score, vegetable score, and sodium score were significantly improved in the intervention group. The intervention group showed beneficial changes in refined grains, vegetable consumption, and sodium intake.
Audet et al (2025) [72]	United States	Qualitative study	27	Prediabetes	56	89% (24/27) women	HP-DPP program (Zoom+REDCap app)	12 months	The participants' knowledge, motivation, and skills were improved.

[a]HbA_{1c}: glycosylated hemoglobin A_{1c}.

[b]RCT: randomized controlled trial.

[c]Not available.

[d]N values are not available.

[e]DPP: Diabetes Prevention Program.

[f]HDL: density lipoprotein cholesterol.

[g]GMI: glucose management index.

Most studies were quantitative (n=22), including randomized controlled trials (n=8) [55,56,62,65,70,71,76,77] and nonrandomized quantitative studies (n=14). These include retrospective studies [50,53,57], interventional studies [51,61], observational studies [58,59,67-69,75], experimental studies [66,74], and pilot studies [63].

In addition to this, 2 studies used a mixed methods design [52,60] and another 4 studies were qualitative [54,64,72,73].

XSL•FO
RenderX

Among the 28 included studies, the delivery modalities used to deliver T2D prevention interventions were categorized into 4 groups. The first 10 involved apps for diabetes prevention [53,56,61,62,65,67,70,72,73,75] and the second 13 involved the use of web-based platforms or websites [50,51,53,55,58-60,64-66,69,74,77]. In addition, 11 trials used wearable devices or digital tracking tools such as a wireless scale [52,53,57,58,61,64,68,69,72,74,76]. Six studies used video conferencing or social software such as Zoom and Facebook [54,59,60,63,72,76].

Bias Reporting of the Eligible Studies

The MMAT was used to assess the risk of bias across the 28 included studies, with none identified as having a high risk of bias. The overall methodological quality was rated as moderate to high, with MMAT scores ranging from 60% to 100% across the studies (Table 4).

XSL•FO
RenderX

Table . Quality assessment.

Study	Research design	Q1	Q2	Q3	Q4	Q5	Quality	Inclusion in the review
Vadheim et al (2010) [63]	QN (NR)[a]	1	1	1	0	1	80%	✓
Moin et al (2015) [64]	QA[b]	1	1	1	1	1	100%	✓
Sepah et al (2015) [74]	QN (NR)	1	1	0	0	1	60%	✓
Block et al (2015) [65]	QN (RC)[c]	1	1	1	1	1	100%	✓
Sepah et al (2017) [66]	QN (NR)	1	1	1	0	1	80%	✓
Stein and Brooks (2017) [75]	QN (NR)	1	1	1	0	1	80%	✓
Michaelides et al (2018) [67]	QN (NR)	1	1	1	0	1	80%	✓
Everett et al (2018) [68]	QN (NR)	1	1	1	0	1	80%	✓
Castro Sweet et al (2018) [69]	QN (NR)	1	1	1	0	1	80%	✓
Srivastava et al (2019) [50]	QN (NR)	1	1	0	0	1	60%	✓
Kim et al (2019) [51]	QN (NR)	1	1	1	1	1	100%	✓
Griauzde et al (2019) [52]	MX	1	1	1	0	1	80%	✓
Alwashmi et al (2019) [53]	QN (NR)	1	1	1	0	1	80%	✓
Toro-Ramos et al (2020) [70]	QN (RC)	1	1	1	1	1	100%	✓
Ferrara et al (2020) [77]	QN (RC)	1	1	0	1	1	80%	✓
Ekezie et al (2021) [73]	QA	1	1	1	1	1	100%	✓
Islam et al (2022) [54]	QA	1	1	1	1	1	100%	✓
Katula et al (2022) [55]	QN (RC)	0	1	1	1	1	80%	✓
Lim et al (2022) [56]	QN (RC)	1	1	1	1	1	100%	✓
Batten et al (2022) [57]	QN (NR)	1	1	1	0	1	80%	✓
Fitzpatrick et al (2022) [58]	QN (NR)	1	1	1	1	1	100%	✓
Collins et al (2023) [59]	QN (NR)	1	1	0	1	1	80%	✓
Yeh et al (2023) [60]	MX[d]	1	1	1	0	1	80%	✓
Zahedani et al (2023) [61]	QN (NR)	1	1	1	0	1	80%	✓

XSL•FO
RenderX

Study	Research design	Q1	Q2	Q3	Q4	Q5	Quality	Inclusion in the review
Chung et al (2023) [62]	QN (RC)	1	1	1	1	1	100%	✓
Gerber et al (2024) [76]	QN (RC)	1	1	1	1	1	100%	✓
Ng et al (2025) [71]	QN (RC)	0	1	1	0	1	60%	✓
Audet et al (2025) [72]	QA	1	1	1	1	1	100%	✓

[a]QN (NR): Quantitative nonrandomized.

[b]QA: Qualitative assessment.

[c]QN (RC): Quantitative randomized controlled trials.

[d]MX: Mixed Method.

The Mechanism of Action of eHL

Knowledge

Figure 2 illustrates the pathways and mechanisms through which eHL supports T2D prevention among at-risk populations.

Figure 2. Mechanistic pathways of eHL in type 2 diabetes prevention among at-risk populations [26,27]. eHL: eHealth literacy.

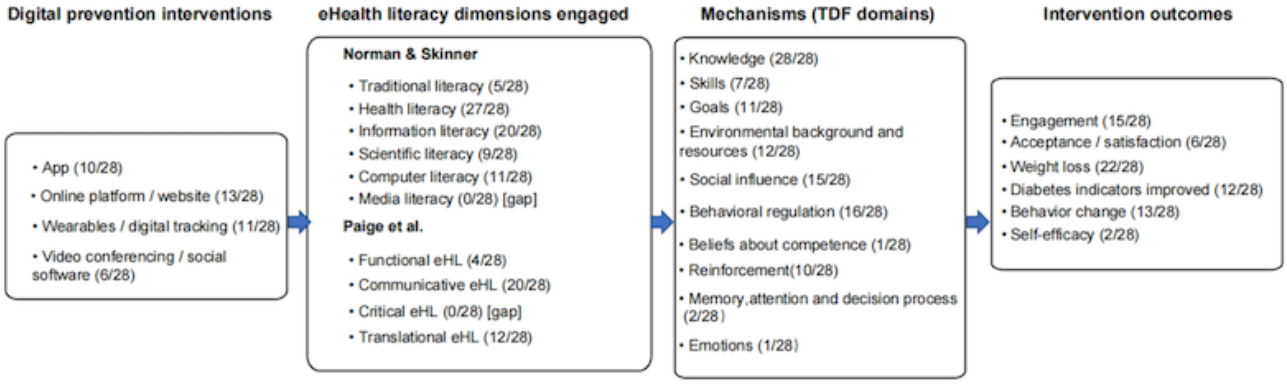

Overview

All 28 included studies (100%) addressed the mechanistic actions of health literacy in the prevention of T2D. The content covered disease cognition, nutrition and diet, physical activity, and standardized DPP courses. These interventions impact diabetes prevention through 3 pathways as follows.

Information Acquisition and Understanding—Disease Knowledge (n=3)

Such as prediabetes and strategies to prevent progression of T2D [52], understanding prediabetes [62], and the pathophysiology and complications of prediabetes [71], the acquisition of disease knowledge builds the knowledge base for prevention.

Behavior Guidance—Diet and Exercise Knowledge (n=8)

Nutrition and diet knowledge was the most common content, Content covering topics such as dietary sources of saturated fats and sugars, portion control [65], weight management, diabetes prevention, healthy meal planning, the relationship between carbohydrates and glycemic response [56], balanced meals following the US Department of Agriculture's My Plate program [57,58], diet [60], reducing foods that raise blood glucose, restricting calories, and increasing fiber intake [61], the Dietary Approaches to Stop Hypertension diet [62], practical nutrition advice for people with prediabetes, including energy balance, portion control, healthy eating based on local foods, glycemic index, fiber intake, variety of fat and protein foods, alcohol intake, etc [71].

Physical activity knowledge (n=6) included teaching classes related to physical activity, such as aerobic exercise, strength training, yoga, dance, pilates, water aerobics, etc [63], physical exercise guidelines [65], educational videos on physical activity [56], benefits and recommendations for increasing activity [61,62] of physical activity [71]. The guidance of diet and exercise knowledge promotes prevention from information to practice.

Digital Intervention Program—Systematizing Knowledge (n=18)

A total of 18 studies [51,53,55,57-60,62-64,66,67,69,70,72,74,76,77] were based on the DPP or its digital adaptations, including 2 from the Prevent project [64,74] and 5 from the Omada project [55,57,58,66,69]. And 11 derived from DPP and its derivative programs

XSL•FO
RenderX

[51,53,59,60,62,63,67,70,72,76,77], whose curricula deliver knowledge through integrated multimodal digital channels.

Skills

Overview

Seven studies [52,54,56,62,67,70,77] reported that eHL promotes prevention through skill building.

Digital Tool Operation Skills (n=6)

eHL is reflected in the ability of participants to learn and master digital intervention tools. Six studies provided training in tool use [52,56,62,67,70,77]. These included learning how to operate the Noom App and interacting with a web-based coach [67,70], receiving instructions for using the App and Fitbit devices [52], learning how to use the online platform through face-to-face group coaching, goal setting, and behavioral techniques for self-monitoring diet, physical activity, and weight using the online platform [77], and downloading and familiarizing yourself with mobile health applications [56,62].

Health Behavior–Related Skills (n=1)

One study [54] reported participants learning practical health management skills through a digital platform, including reading nutrition labels, making plate plans, and controlling portion sizes.

Goals

Overview

A total of 11 studies [54,56-58,62,65,68-71,73,75] provide the function of goal setting in diabetes prevention. There are 3 main patterns of goal setting in research.

Algorithm-Driven Personalized Target Recommendation (n=3)

Three studies used algorithms to generate personalized goals for participants [62,65,68]. In a high-quality study, the Alive-PD system recommended weekly that participants choose one eating habit goal and one physical activity goal from among the suggested goals. In addition to long-term goal setting, individually relevant small-step goals were recommended weekly according to the participant. For example, participants who consumed more than the recommended level of added sugars were provided with weekly targets based on their reported sources of added sugars. The target-recommendation algorithm adapts to the participant's previous performance. They were also prompted on a weekly basis to report on the completion of goals in the previous week. The system tracks the number and type of goals that respondents successfully achieve [65]. In another study, Sweetch used advanced algorithms to provide personalized, contextualized, immediate, and timely advice to guide users toward recommended activity, weight loss, and dietary goals in a way that aligns with the user's actual lifestyle. In another high-quality study, the Sweetch technique was able to learn which types of messages led to better compliance for specific users in specific contexts. Breaking down the weekly 150-minute physical training goal into small pieces throughout the day makes the goal easier to achieve. These snippets will continue to be personalized for each user. Personalized goals are constantly adjusted according to the actual behavior of the user to best fit the actual life capabilities of the user [68]. Another study set up a "milestone" function so that participants could review their goals and thus make adjustments to meet their monthly goals. For example, if a participant sets a fasting glucose target of 60 to 99 mg/dL, a pop-up window displays achievement when the target is reached. Participants can also view weekly or month-long bar and line charts to compare actual versus expected values at different time points and actual versus ideal behavior in order to adjust desired goals as needed [62].

Goal Setting Assisted by a Web-Based Coach (n=4)

Four studies [57,58,70,71] involved health coaches assisting participants to help users set specific, measurable, achievable, and realistic goals. Alternatively, in one study, the intervention group was instructed to set specific, measurable, realistic, achievable, and timely goals by selecting small but meaningful changes to fit into their daily lives, which were constantly reviewed and revised during subsequent follow-up [71].

Autonomous Goal Tracking and Self-Regulation (n=5)

In addition, goal setting can be done through personal records and individual settings [54,56,69,73,75]. In one study, users were guided through goal-setting modules on weight loss and food choice [75]. In another study, individuals were able to set step goals within an app [56].

Environmental Background and Resources

Overview

Twelve studies created empowering prevention resources for participants by providing access to digital tracking tools and wearable devices [52,53,55,56,61,65,66,69,71,73,74,76]. These devices fall into 3 main categories as follows.

Weight Monitoring Equipment (n=7)

Seven studies equipped participants with wireless or Internet-connected scales to automatically track weight changes and transfer data to a database or personal dashboard [52,55,66,69,71,74,76].

Motion Monitoring Equipment (n=11)

Exercise monitoring equipment was the most widely used form of resources, with a total of 11 items. Including a pedometer, wrist wear activity tracker, and other intelligent wearable devices [52,53,55,56,65,69,71,73,74,76]. One study [55] allowed participants to use their own wearable devices (eg, Fitbit and Apple Watch).

Other Health Monitoring Tools (n=2)

In addition to weight and exercise tracking, 2 studies combined more integrated digital health tools such as continuous glucose monitors, which track blood glucose fluctuations in real time, and heart rate monitors [61]. Alternatively, in the study by Ng et al [71], a glucose meter was provided for self-blood glucose monitoring, and data on the glucose meter and tracker were gradually tracked and synchronized into the application.

XSL•FO
RenderX

Social Influence

A total of 15 studies identified social influence as a key mechanistic pathway through which eHL facilitated diabetes prevention behaviors [51,53,55,57-59,62,64,66,67,69,71,72,74,76]. The forms can be divided into 2 types. One group used closed digital communities, with 12 studies interacting through web-based groups, private social forums, private chat groups, third-party messaging tools, where participants posted updates, responded to comments, shared lifestyle change experiences, and received feedback from wellness coaches or other members to get online peer support [51,53,55,57,58,62,64,66,67,69,71,74]. For example, in one study, the intervention group was invited to join a peer support chat group hosted by WhatsApp. They could ask questions related to prediabetes and exchange experiences on lifestyle modification, including best practices, challenges faced, and ways to overcome, throughout the study period [71]. In addition to digital communities, 3 studies used structured online discussion activities to enhance interactivity and interaction during continuous live lectures [59,72,76]. For example, in one study, group members explored interactively around each member's individual case and provided integration opportunities to advise each other based on feedback or their own experience [76].

Behavioral Regulation

Of the included studies, 16 studies included behavioral regulatory mechanisms as an important component of promoting lifestyle change [50,52,53,56-60,62,63,66,67,70-72,75].

Thirteen studies asked participants to regularly self-record health-related behaviors such as body weight, dietary intake, physical activity, and sleep [50,52,56,59,60,62,63,66,67,70-72,75]. Among them, one high-quality study used the Artificial Intelligence Health Coaching (HCAI) application to prompt users to enter their weight weekly and record the intake of meals and snacks. The system automatically classifies foods as "healthy," "unhealthy," or "neutral" to help users form a clearer perception of diet [75]. Another high-quality study introduced the "SPACE" model, which encouraged participants to record five types of daily health-related habits—Sleep, Presence, Activity, Creativity, and Eating. To enhance overall self-awareness [52].

Three other studies used photo-capable "food diaries" in addition to self-recording measures, in which participants were asked to upload photos of each meal, snack, or drink to an App. Health coaches reviewed the records weekly and provided personalized feedback, thereby reinforcing behaviors [53,57,58].

Beliefs About Competence

Only one study addressed the mechanistic pathway of beliefs about capabilities in diabetes prevention [53]. Based on the concept of social cognitive theory, the Transform program aims to help participants improve self-efficacy, enhance participation, and influence behavior.

Reinforcement

Overview

Among the included studies, 10 reported using reinforcement strategies to enhance participants' self-efficacy and support the maintenance of health behaviors. These strategies included structured encouragement, coach-delivered encouragement, and incentive-based approaches such as points-based rewards and gamification [50,59,62,64-68,76,77].

Systematic Encouragement (n=4)

Four studies used programmed systematic incentive strategies to continuously encourage and strengthen participants' participation through algorithmically driven feedback [50,64,65,76]. For example, users can see progress and send encouragement notifications. Users are provided with daily and weekly summaries, with weekly summaries graphically providing the amount consumed, total calories consumed, calories burned, and projected weight loss goals for the week for each food group [50].

Coach Encouragement (n=4)

Four studies emphasized reinforcement strategies at the interpersonal level, which continuously strengthened participants' competency beliefs through encouragement of structured programs by health coaches [66], daily information encouragement [67], progress encouragement [77], and social support [59]. In a high-quality study, participants received daily messages from their coaches, such as "No pains, no gains. Focus on the application as much as possible. Record as much as you can, as this gives your coach a sense of your health habits and provides background information for all discussions. The longer you use the app, the more likely you are to stick to your health goals" [67].

Points and Gamification Encouragement (n=2)

In one study, participants earned points for setting and achieving goals, which were redeemed for small cash rewards to encourage them to continue reporting health behaviors and achieving goals [68]. Another study added a digital gold reward system to the group chat system, where participants could collect digital currency by completing questionnaires, quizzes, or meeting goals and redeem it for actual prizes. Researchers have further embedded gamification elements to improve participants' participation in prevention programs [62].

Memory, Attention, and Decision Processes

Two studies referred to studies related to memory, attention, and decision processes [52,65]. One study helps participants to maintain focus and self-regulation by providing behavioral strategies of mindfulness and decompression [65]. In addition, it predicts the energy and willpower of the individual for a future day. Help individuals understand and control the factors that affect their health behaviors and strengthen their decision-making awareness [52].

Emotions

Only one study incorporated emotion-related mechanisms in prevention by using artificial intelligence (AI)–enabled guidance to provide compassionate, feedback-informed support. Providing

XSL•FO
RenderX

strategic help and emotional support, AI incorporates interactive elements such as reflection, legitimation, respect, support, and partnership to make users feel valued [50].

The Dimensions of eHL and Intervention Outcomes

The classical framework [26] divides eHL into 6 dimensions. Later, Paige et al [27] proposed a new classification after 2018, which emphasized the interactive use, communication, and behavioral transformation of digital health information in real situations.

Traditional Literacy

Five studies addressed traditional literacy [51,54,63,65,69]. Four of these directly intervened in participants' reading comprehension through a forced reading task. For example, in the study by Block et al [65], participants were required to read 2-weekly health notes written by certified diabetes educators and behavioral specialists. In addition, one study conducted an adaptive intervention at the language and cultural level according to the traditional literacy level of participants [51]. The study adapted the DPP curriculum for low-income people, including rewriting content at the fourth and fifth grade reading level, cultural adjustments, and Spanish translation, and adding bilingual and biculture-based health coaches in English and Spanish to help them better understand prevention programs.

Health Literacy

A total of 27 studies addressed health literacy [50-60,62-77], and interventions were broadly divided into 2 aspects.

First, most studies used health education courses as the intervention core (n=16) [50,51,53,55,57-59,63,64,66,67,69,74-77], covering topics such as healthy diet [54,62,63,65,71], physical activity [54,62,63,65], knowledge of prediabetes [71], etc. Three studies integrated the official Centers for Disease Control and Prevention (CDC)-DPP curriculum.

Second, 7 studies incorporated personalized content into health literacy interventions [6,8,12,19,20,26,27]. In particular, 2 of these studies introduced AI and algorithm-driven personalized health content. The research of Stein and Brooks [75] uses AI to simulate health coaches to provide counseling support, provide strategic help and emotional support, and AI integrates interactive elements such as reflection, legitimization, respect, support, and partnership to make users feel valued. And provide personalized content. The research of Everett et al [68] uses the advanced Sweetch algorithm to provide users with personalized, contextualized, immediate, and timely recommendations to guide users to achieve recommended activity, weight loss, and diet goals in a way that conforms to users' actual living habits. Personalized goals are constantly adjusted according to the actual behavior of the user to best fit the actual life capabilities of the user. Notifications constitute "teachable moments" that can bring about sustained behavioral change.

Information Literacy

A total of 20 studies addressed information literacy [50,52-54,56-58,60,62-67,69-72,75,76]. Among them, information recording is the most common type, and participants in 11 studies actively record personal health data such as weight, diet, step count, physical activity, sleep, etc, through apps, dashboards, or online platforms [52,56,60,63,66,67,69,70,72,75,76]. Of these, 4 studies specifically used "health diaries" to enhance information literacy through daily journaling [50,53,62,71]. For example, the study by Srivastava et al [50] entered participants' food eaten, exercise completed, educational sessions completed, and body weight into a diary, which resulted in automated daily and weekly performance summaries and performance-based tailored daily and weekly recommendations.

Second, information was understood and evaluated through quizzes. In 11 studies, participants were required to complete a course quiz or answer questions [50,53,54,57,58,60,62,64,65,69,71]. For example, in the study by Fitzpatrick et al [58], participants were required to complete questionnaires, quizzes, and open-ended questions. Course completion was defined as completion of the quiz attached to the end of the course content.

Science Literacy

Among the included studies, 9 studies involved the cultivation of scientific literacy [50,52,56,61-63,71,75,76]. Of these, 7 studies focused on dietary knowledge based on scientific principles [50,56,61,63,71,75,76]. For example, in the study of Stein and Brooks [75], the AI-based text interactive HCAI system will classify the food entered by the user into 3 categories of "healthy," "unhealthy," and "neutral" to facilitate the user to establish a food recognition mechanism. In the study by Ng et al [71], the application module covered multiple aspects of healthy eating and practical nutrition advice for people with prediabetes, including content on energy balance, portion control, healthy eating based on Malaysian indigenous foods, glycemic index, dietary fiber intake, fat and protein food types, and alcohol intake.

In addition, 2 studies addressed scientific mechanisms of disease [62,71]. For example, in the study of Chung et al [62], the TCM mHealth app included knowledge about constitution classification, meridian energy, and descriptive illustrations and video teaching of Qigong types such as Baduanjin and abdominal breathing. The Ng et al [71] study provided knowledge on the pathophysiology and complications of prediabetes, health benefits, and specific recommendations of physical activity, health risks of smoking, stress management strategies, and motivation training for lifestyle changes in the app. These science literacy training strategies help participants to understand the causal relationship between health behaviors and disease prevention from a scientific perspective, thereby improving health decision-making ability.

Media Literacy

Of the 28 included studies, none explicitly addressed the cultivation of media literacy, which is the ability to critically evaluate digital health information sources, identify misleading content, or understand algorithmic mechanisms, among other things.

XSL•FO
RenderX

Computer Literacy

According to the analysis of the included studies, 11 studies involved the cultivation of computer literacy [52,53,56-58,62,63,67,70,71,77]. These interventions mainly help participants master the skills of using digital health tools through platform operation training and multimedia content creation and uploading.

First, platform operation training is the basis of computer literacy training [52,56,62,63,67,70,71,77]. For example, in the study by Lim et al [56], participants in the intervention group were introduced to the nBuddy Diabetes mobile app. They were asked to download the nBuddy Diabetes app and receive training, and in the study by Michaelides et al [67] and Toro-Ramos et al [70], participants were asked to learn how to use the Noom app and interact online with a health coach.

Second, 3 studies involved multimedia content creation and uploading. For example, participants were asked to track eating behavior by taking and uploading eating photos [53,57,58].

Functional eHealth Literacy

Of the included studies, 4 studies addressed functional eHL [51,54,63,65].

First, the ability to read health information is a core competency of functional literacy [54,63,65]. For example, in the study by Vadheim et al [63], each participant was asked to read books that estimated the fat and calorie content of specific foods and to correctly record fat, body weight, calories, and minutes of activity in an electronic diary. The study by Islam et al [54] designed practice activities in workshops to help participants master important nutrition-related skills, such as reading food nutrition labels.

Second, culturally adaptive content adjustment reflects the attention paid to people with different levels of functional literacy [51]. In the study by Kim et al [51], the DPP curriculum was adapted for low-income populations, including rewriting content to fourth- and fifth-grade reading levels, cultural adjustment and Spanish translation, and adding bilingual and bicultural health coaches in English and Spanish. This design lowers the reading threshold and ensures that participants of different educational backgrounds and language abilities can effectively understand and use digital health information.

Communicative eHealth Literacy

According to the analysis of the included studies, 20 studies related to communicative eHL [51,53-59,64-67,69-76].

A total of 11 studies demonstrated the cultivation of participants' collaboration ability [51,54,57-59,66,69,71,73,74,76]. These studies promote participant prevention through discussion, providing social support, sharing experiences, etc. For example, in the study of Islam et al [54], the study set up a "brainstorming" session, where peer leaders shared questions/prompts related to healthy eating, physical activity, and healthy lifestyle barriers (stress, difficult emotions, school/work schedule, family stress, etc), and participants were invited to share ideas. In the study by Collins et al [59], program participants were allowed to interact during live lectures to share successes/challenges and provide support to each other.

Translational eHealth Literacy

Of the included studies, 12 involved translational eHL [50,52-54,56,61,62,65,68,70,71,76]. The common feature of these studies is the immediate translation of digital health information and self-monitoring data into concrete, executable behavioral goals through regular individualized feedback and the dynamic revision of these goals in subsequent stages. For example, in the study of Block et al [65], the system generates personalized dietary and exercise goals for participants every week and automatically adjusts small step goals for the next week based on their actual achievement in the previous week. Participants are required to report their goal completion every week, and the system continuously updates the follow-up plan accordingly. In contrast, in the study by Alwashmi et al [53], this regulatory process was mainly accomplished by human coaches. The coaches reviewed participants' follow-up weekly, provided specific recommendations for diet and meal plan changes, and helped participants set more feasible goals for the coming week, thereby reinforcing behavior change on the basis of progress.

Critical eHealth Literacy

Among the 28 studies included, none addressed critical eHL, which is the ability to evaluate the credibility, relevance, and risks of health information shared and received online.

Intervention Outcomes

Engagement

Overall, studies reported high levels of participant engagement [50,52,55,58-60,63-65,68,69,74-77]. Of note, 2 studies achieved 100% participation [52,76]. In one of these studies, app+ completion rates reached 100% in the intervention group. Participants used app+ for an average of 37 days over 12 weeks [52]. In the other study, meeting attendance was 100%, participants had a mean age of 65.4 years, wore a fitness tracker daily, were weighed more than 3 times per week, and recorded food intake more than 3 times per week. During the 6-month reporting period, participants adhered to the target calorie range 74% of the time. The initial daily step goal was 5000, and 92% of participants maintained this goal through 6 months [76].

In terms of course completion, participation was better [50,58,59,63,69,77]. Among them, in one study, 88% of participants completed 16 weeks of courses [63]. In another study, 95% completed at least 4 weekly education courses, and 92% completed 9 or more courses [69]. Across different educational stages, nearly 65% of high school or equivalent people completed the first stage of courses. For participants who graduated from college, the completion rate of phase 1 reached 75.5% [59].

In terms of online platform use, participants similarly demonstrated strong engagement. The study showed that all participants completed at least 4 core modules, and participants logged in to the online intervention 76 times on average, weighed 89 times, sent 46.5 group messages, and sent 20.5 messages to the health coach [64]. Another study reported that

XSL•FO
RenderX

70.8% of intervention group participants interacted with the program at least once a week. After 6 months, 70.6% were still actively using the Alive-PD system [65]. In another study, participants logged on an average of 22.5 times weekly, weighed themselves 7.3 times weekly, performed 11 diet/activity tracking sessions, and had 1 private message with a health coach. In addition, 96% (n=455) of participants were enrolled in the digital DPP program for the full 12 months [58].

Acceptance or Satisfaction

Six of the included studies [52,54,60,64,68,75] explored acceptance or satisfaction. Six studies received positive reviews from participants, and in 2 of these studies, participants clearly expressed a preference for prevention programs [52,60]. In the study by Moin et al [64], women veterans with prediabetes after the prevention program rated the program as a good fit with perceived health needs, “it’s perfect,” “it can keep its own pace,” “flexible, it can be free from time constraints,” the program was integrated into daily life and made people feel responsible. In the study of Everett et al [68], Sweetch mobile platform had a high acceptance among participants. Most participants agreed or strongly agreed that they would like to use the Sweetch app often (74%), thought the app was easy to use (83%), thought the features were well integrated (72%), thought most people could learn to use the app quickly (89%), and felt confident in using the app (77%). The value and practicability of the project were highly recognized by the participants. In the study by Islam et al [54], participants indicated that the online adolescent health education program was beneficial and applicable to their lives, giving them a deeper understanding of healthy eating and an active lifestyle. When asked to describe items with adjectives, the most commonly used words were “interesting” and “useful.” Participants appreciated the Let’s Move activities and discussed how the activities had helped them understand healthy eating and share what they had learned with friends and family. Most participants found the game both fun and rewarding, enjoyed teamwork, enjoyed competition, felt that the game helped participants become more comfortable with each other, and appreciated that the game challenged them to remember important concepts they had learned.

In addition, 3 studies [60,64,72] also pointed out the challenges of eHL participation in prevention. In the study by Moin et al [64], participants expressed an aversion to records, a lack of openness online, and a lack of computer literacy, leading to comprehension difficulties.

In the study of Yeh et al [60], computer literacy also encountered challenges. Some participants experienced difficulties in the first 2 weeks, including difficulty recording food intake and low Facebook group participation, with only 2 people interacting. Participants had barriers to software use. In the study by Audet et al [72], older participants still complained about the use of video technology in group meetings and expressed difficulties in developing computer literacy.

In addition, the study of Yeh et al [60] also pointed out the challenges brought by cultural differences. Chinese-American participants expressed challenges in recording weight in prevention programs. Because in Chinese culture, meals are usually shared in a family format, and there are no individual portion sizes.

Body Weight

Of the 28 included studies, 22 reported on weight management results, showing a significant effect of eHL in promoting weight management in people at risk of diabetes [50,51,53,55-61,63-70,74-77].

In short-term intervention studies (≤6 mo) [50,53,56,61,63-65,68,75-77], weight loss has been substantial. Of these, 6 studies explicitly reported that participants lost 5% or more of their body weight [50,53,63-65,76]. For example, in the study by Srivastava et al [50], all 6 users who completed the study achieved the CDC target of 5% weight loss at 6 months, with an average weight loss of approximately 9.0%. In another high-quality study [76], weight was reduced by 5.3% at 3 months and 8.4% at 6 months in the intervention group, with 90% of participants achieving and maintaining the target weight loss of 5% or more and a reduction in BMI from baseline 31.3 to 29.06.

Weight loss was maintained in medium- and long-term intervention (≥9 mo) studies. Five studies [55,57,59,67,70] reported weight changes in participants from 9 months to 65 weeks later. For example, Katula et al [55] showed that patients in the d-DPP group lost an average of 5.49% of their initial weight at 12 months and were 61% more likely to have clinically significant weight loss (≥5%) than those in the control group, and that greater participation was associated with greater weight loss, especially among participants 65 years of age or older.

Six studies reported trends in weight change at different time points [51,58,60,66,69,74]. For example, Castro Sweet et al [69] showed that participants lost an average of 6.5% of body weight at 16 weeks, increased to 8.0% at 6 months, and remained at 7.5% at 12 months, with an average weight loss of 13 - 14 pounds. Long-term follow-up by Sepah et al [74] showed a mean weight loss of 4.9% after 1 year and 4.3% after 2 years among patients who completed the program, demonstrating sustained weight loss maintenance. Of note, in the study by Fitzpatrick et al [58], the average weight loss was 8.6 pounds (4.0%) within 12 months, which was clinically significant, but at 24 months, the percentage of weight loss was 5.7 pounds, which was no longer statistically significant, and long-term weight maintenance was still a challenge.

Diabetes Indicators

Of the 28 included studies, 12 reported improvements in measures of glycemic control [51,55,56,59,61,62,65,66,68-70,74]. Ten of these studies showed positive results [51,55,56,59,61,62,65,66,69,74], and 2 studies [68,70] reported no significant change from baseline in fasting plasma glucose and glycosylated hemoglobin (HbA_{1c}).

Seven studies reported reductions in HbA_{1c} [51,55,59,62,66,69,74]. Among them, 3 studies [62,66,69] found sustained and sustained reductions in HbA_{1c} over time after eHL engagement prevention studies. For example, in the study by Sepah et al [74], the mean reduction in glycated hemoglobin levels was 0.38% after one year and 0.43% after 2 years. In a

XSL•FO
RenderX

long-term study by Sepah et al [66], significant reductions in glycated hemoglobin levels were observed at week 16 and at years 1, 2, and 3.

In addition, 5 studies reported improvements in fasting plasma glucose [56,61,62,65,69]. In a high-quality study, participants' fasting blood glucose decreased by 0.41 at 6 months, and their Framingham 8-year risk for diabetes decreased from 16% to 11%. Most encouragingly, 40.5% of the intervention group achieved normal fasting glucose levels at 6 months.

In terms of other diabetes-related measures, one study showed significant within-group improvements in systolic and diastolic blood pressure, total cholesterol, and low-density lipoprotein cholesterol after 6 months in the intervention group [56], and another study combined a TCM mobile health app intervention. It was shown that participants performed better than the control group in terms of body constitution (improvement of Yang deficiency and sputum stasis), body energy improvement, and health-related quality of life (PCS and MCS) after 4 months. Prevention that demonstrates eHL engagement appears to be more likely to improve physical energy and health-related quality of life [62].

Behavioral Change

Results of a total of 13 studies showed that eHL participation in prevention research promoted the improvement of health behaviors of participants [50,52-54,56,61,65,68,72,73,75,77]. Of these, 8 studies involved improvements in dietary behavior among participants [52,54,56,61,65,72,75,77]. For example, in the study by Lim et al [56], the daily intake of total energy, carbohydrates, total fat, saturated fat, and sugar decreased significantly after 6 months in the intervention group. Another pilot study [54] found that participants learned about healthy eating during the program and were able to share what they learned with family and friends. As one participant commented, "The program helped me to adjust my eating habits, to know what to eat, how much to eat, and how much to plan."

Second, 9 studies showed changes in physical activity among participants [50,52,53,57,61,65,68,72,73]. For example, in the study by Ekezie et al [73], participants significantly increased their physical activity levels and average daily steps. A wrist-worn activity monitor linked to a digital app tracked daily steps in real time, and the "physical activity challenge" feature of the program was considered a key enabler of goal attainment, helping participants stay focused and motivated to continue achieving their goals over a longer period of time, they said. Another study further validated the improvement in motor behavior. Zahedani et al [61] found that participants with prediabetes had a significant increase in their daily physical activity after 3 months of intervention, with the average duration of physical activity approximately doubling. As an objective measure, the mean daily heart rate rose by 6.3 minutes/day.

Self-Efficacy

In addition, 2 studies also reported significant improvements in self-efficacy among participants [59,65]. In the study by Collins et al [59], participants showed statistically significant improvements in the ability to perform 18 self-reported health promotion behaviors.

Discussion

Principal Findings

To the best of our knowledge, this is the first systematic literature review of T2D prevention in at-risk populations, which aims to deeply explore the mechanism of eHL in T2D prevention in at-risk populations and the participation and effects of multidimensional prevention. Our core finding is that eHL, although extremely uneven in its application, still achieves significant preventive effects. This reveals systemic problems in current prevention research and future directions for improvement.

In this review, we mapped intervention components and reported implementation strategies from the included studies to the TDF domains and quantified how frequently each domain was reflected across the 28 studies. The use of these pathways is extremely uneven. Almost all studies relied on the "knowledge" approach (28/28, 100%), while behavioral regulation (16/28, 57.1%) and social influence (15/28, 53.6%) were also widely used. However, few studies involved emotions (1/28, 3.6%), beliefs about competence (1/28, 3.6%), memory, attention, and decision processes (2/28, 7.1%). This implies that current research constructs a "surface" rather than a "deep" prevention model, with much focus on knowledge that is easily transmitted and behaviors that are easily monitored, but the potential to stimulate intrinsic motivation for change has been largely neglected.

In all the studies included in this review, health literacy was central in terms of understanding. The original intention of such prevention is consistent with the CDC in the United States, which is to translate scientific knowledge into action to improve public health [78]. In addition, 18 studies showed a trend toward more standardized prevention content based on DPP content. Similarly, behavioral regulation and social incentives are relatively easy to achieve through regular recording with various types of incentives. Perhaps the reason for the widespread adoption of easy-to-implement mechanisms lies in their operability. The convenience of such preventive measures may obscure a fundamental question: Is the easy-to-implement path the most long-lasting mechanism?

Our findings suggest that, in terms of implementation, the included studies focused primarily on the goals of algorithmic systems and digital coaches assisting design, and that this design difference has theoretical and practical implications for self-determination theory. Self-determination theory emphasizes autonomy as a key intrinsic motivation for promoting behavior change [79]. Overreliance on algorithms or digital coaches may undermine participant autonomy. Second, the emotional and psychological pathways are also neglected. This is consistent with previous findings that psychological factors have been underreported in digital health care interventions [80]. For example, negative emotions such as the frustration caused by weight loss and blood glucose rebound were obviously ignored in the research, and negative emotions are often one of the reasons for interrupting self-control behavior [81]. Such neglect tends to lead to long-term prevention failure.

XSL•FO
RenderX

Second, eHL is by no means a trivial set of skills in a highly digital environment. Rather, it combines knowledge and skills from a variety of fields [82]. Our review found that the framework of Norman et al [26] and Paige et al [27] is highly fragmented and selectively applied in practice.

Specifically, health literacy and information literacy were the 2 most frequently involved dimensions. On the contrary, although the attention to scientific literacy increased year by year, the attention to traditional literacy was completely absent after 2022. Media literacy is more worrisome, and none of the prevention studies addressed the cultivation of media literacy.

There are many explanations for the uneven distribution of dimension involvement in prevention. First, the concept of eHL has been discussed and applied in the field of public health and health communication since it was proposed by Norman et al [26]. In contrast, media literacy stems more from the field of media studies and has not yet been fully integrated. This has led to a critical gap: the Internet has become an important channel for the dissemination of health information for the prevention of T2D, but the requirements of media literacy, such as how to identify the correct medical information, have not been systematically incorporated into preventive medical interventions.

However, the interest in scientific literacy has been increasing in recent years, especially after the COVID-19 pandemic, which is a positive sign to some extent. However, it may simply be because of the external pressure of the global nature of the pandemic.

In addition, the lack of traditional literacy, perhaps due to the popularity of digitalization, researchers have neglected that online functions still require basic reading and writing skills. It has been found that a disadvantage in written and oral language skills can be a barrier to accessing web-based health information [83]. This means that current prevention studies may be inadvertently excluding participants who have low literacy and difficulty reading and typing. People with lower education levels are generally at higher risk of chronic diseases [84]. Thus, the lack of traditional literacy is not only a study design issue, but also a health equity issue. To ignore traditional literacy is to exclude the most vulnerable from digital health help, exacerbating divisions.

In 2018, Paige et al [27] classified the new model as a further improvement of the old model, from “instrumental skills” to “interactive application ability.” However, this study found further “selective neglect” in the study: communicative eHL and translational eHL were more involved in prevention, but functional eHL was less involved, and critical eHL was completely missing. Once again, this demonstrates the “selective” application mechanism of eHL in practical prevention research.

Encouragingly, research has begun to directly acknowledge and respond to inequalities in functional literacy. The diabetes prevention curriculum was adapted for low-income people, rewritten to fourth to fifth grade reading difficulty, culturally adjusted, and translated into Spanish, and added bilingual and bicultural health coaches. The breakthrough of this approach is that it takes the initiative to consider and lower the threshold of health information, so that participants with different cultural backgrounds and language abilities have the opportunity to participate in digital health, which is particularly critical in the field of diabetes prevention. Because the lack of functional eHL can further lead to the emergence of a digital divide, there are significant differences in the need for technical training among different populations. Among the included studies, studies involving, for example, older adults [58,76], their training in the use of digital tools need more targeted attention.

Second, communicative eHL had the highest frequency, also responding to the absence of the “emotional information factor” in the path effect. For those at risk, the experience of being “understood and supported” is itself a source of motivation to continue to participate. This motivation is further strengthened by the translational eHealth hormone. Participants were trained and guided repeatedly, leading to more long-term, routine behavior change. An emerging trend observed in this study is that AI is gradually becoming an important tool for eHL interventions. In one of the included studies, an AI health guidance application (HCAI application) provided a human-like guidance experience and personalized content [75]. AI is playing an increasingly important role in enhancing diabetes prevention and management [85]. The introduction of AI, to some extent, strengthens “translational eHealth literacy,” and AI shows great potential in long-term prevention.

The absence of critical eHL, however, is consistent with media literacy. This is also one of the biggest concerns of this review. The current digital environment is extremely complex, and it is very important for individuals to be able to include and judge the credibility and reliability of information sources. The included studies were mainly conducted within the framework of guidance from medical professionals, and participants were not required to assess whether such programs were disinformative. However, once this support is withdrawn, and the participants’ lack of autonomy is fostered, it is extremely easy to lead to difficulties in self-management. Therefore, we suggest that future prevention research could embed training modules on “how to avoid false judgments” while maintaining “learning to make good decisions,” for example, scoring the credibility of health information, exercises on checking health information, etc, mitigating the double risks of information overload and misinformation.

More urgently, the review 5 years ago already noted that while health literacy was the most assessed domain, none of the studies evaluated all 6 domains of the eHL model [86]. The results of this review still show that none of the 28 studies designed and evaluated eHL as a comprehensive, unified concept. Although the path application of eHL was not balanced with the dimensions of participation, our review suggests that eHL is capable of achieving positive outcomes in terms of T2D prevention outcomes for at-risk populations. The cultivation of eHL can improve the participation, satisfaction, weight management, diabetes-related physiological indicators, health behavior change, and self-efficacy of the prevention of T2D in the risk population. This finding is consistent with existing results on the impact of eHL on health behaviors [87-92].

XSL•FO
RenderX

Therefore, the importance and effectiveness of eHL in the prevention of T2D in the at-risk population are clear. But our findings need to highlight how eHL can be a more sustainable long-term driver that needs to be addressed by health communication. In the future of T2D prevention, especially with the rapid development of AI, there is an urgent need to incorporate "autonomous" literacy training without professional guidance into prevention design, rather than continue to assume that participants will automatically have these abilities. Only in this way can eHL help public health mitigate the imbalances brought about by the information age, especially in at-risk and vulnerable populations, and reduce the incidence of T2D in the long term.

This review is novel in providing a theory-driven, mechanistic synthesis of how eHL supports T2D prevention in at-risk populations. We mapped and quantified intervention components and implementation strategies against the TDF and multidimensional eHL frameworks, revealing a marked imbalance in mechanisms: studies largely rely on knowledge and easily operationalized behavioral pathways, while underrepresenting mechanisms critical for long-term maintenance (eg, emotions, support toward autonomy/self-efficacy, and cognitive decision processes). Our contribution is a practical mechanism-by-dimension framework and an evidence-gap agenda, highlighting common omissions such as critical eHL and media literacy, and limited attention to functional literacy and access barriers as determinants of equity. In real-world implementation, future digital/AI-enabled prevention should move beyond information delivery to embed affective and autonomy-supportive design, misinformation-resilience (critical appraisal) training, and accessibility adaptations for vulnerable groups; AI should be integrated cautiously to avoid undermining autonomy and to enhance long-term sustainability and health equity.

Limitations

This review has several limitations. First, although the study aimed to explore the impact of eHL on the prevention of T2D, we did not conduct a meta-analysis due to the substantial heterogeneity observed among the included studies in terms of research design, intervention approaches, and outcome measures.

Second, as "eHealth literacy" remains an emerging concept in the research landscape, the volume of available literature is still limited. Consequently, the relatively small sample of studies included in this review may restrict the comprehensiveness of our understanding of the field.

Third, the thematic analysis used in this review was qualitative in nature. Most of the included studies did not directly investigate the impact of eHL on T2D prevention but instead provided indirect evidence or descriptions of relevant interventions. As a result, the mechanisms through which different dimensions of eHL contribute to the prevention process were derived primarily through interpretive synthesis rather than quantitative validation. This type of analysis may be subject to researcher subjectivity and carries a risk of reporting bias. To mitigate this risk and enhance the transparency of the findings, all authors were involved in reviewing each stage of the data analysis process.

Accordingly, several strategies are proposed for future research. First, to address the high degree of heterogeneity across studies, future research should consider conducting more randomized controlled trials with clearly defined control groups and standardized intervention protocols. Such studies would help generate high-quality empirical data suitable for quantitative synthesis and lay the groundwork for future meta-analyses.

Second, future studies should conceptualize eHL as an integrated and unified construct rather than examining its dimensions in isolation. Researchers are encouraged to adopt standardized measurement tools, such as the eHL Scale, for systematic assessment. Moreover, rather than focusing solely on the indirect role of eHL in digital interventions, future studies should investigate its direct effects as a core variable. Doing so will help clarify its independent contributions and support the standardization and theoretical development of eHL in the context of disease prevention.

Third, to reduce potential bias and enhance the credibility and validity of findings, future research may consider incorporating external peer review or expert consensus methods in the analytical process. For example, using the Delphi method to collect structured input from multidisciplinary experts can enhance the robustness and transparency of the results. These strategies help ensure that the conclusions are not solely reliant on the subjective judgments of the primary research team.

Conclusions

This mechanistic systematic review integrated the TDF and eHL models through theory-driven thematic analysis. It focuses on the pathways through which eHL influences prevention among individuals at risk for T2D. Findings indicate that eHL supports sustained engagement and self-management through mechanisms such as knowledge and skill acquisition, goal setting, behavioral regulation and self-monitoring, resource usage, and social influence, and is associated with outcomes including weight control, improved glycemic indicators, self-efficacy, and lifestyle change. Most studies reported high engagement rates and satisfaction levels, suggesting good acceptability and feasibility of intervention programs.

However, the exploration of mechanisms for T2D prevention across current studies remains uneven. Emotional support, beliefs about competence, and autonomy were underaddressed, while critical eHL and media literacy dimensions were largely absent. The default assumption of traditional and functional eHL's importance may inadvertently exclude resource-limited, at-risk populations, exacerbating health inequities and the digital divide.

Future digital DPPs must better serve diverse populations. Our findings reveal that most programs focus heavily on information delivery but neglect critical factors that help people sustain behavior change: managing emotional setbacks, supporting personal autonomy, and distinguishing reliable health information from misinformation. We recommend that programs incorporate emotional support to address frustration and disappointment, empower users to make independent choices rather than relying solely on algorithm recommendations, and teach simple skills for evaluating web-based health information

XSL•FO
RenderX

credibility. Additionally, programs should include media literacy training, use engagement mechanisms such as incentives and encouragement, and adapt designs to accommodate different ages, education levels, and digital access capabilities. Only through such comprehensive, equity-focused approaches can eHL serve as an effective public health tool for broader populations, particularly vulnerable and at-risk groups.

Funding

This study did not receive external funding.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

JL, AAA, and ZZH contributed to the conceptualization and methodology of the study. JL was responsible for data curation, investigation, formal analysis, provision of resources, and project administration. WY and RG conducted the screening and selection of studies, performed data extraction, and carried out the critical appraisal. Supervision was provided by AAA and ZZH. JL prepared the original draft of the manuscript, and JL, AAA, and ZZH contributed to the review and editing of the manuscript. All authors have read and approved the final version of the manuscript. No generative artificial intelligence tools were used in the writing of this manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Searching strategies.
[DOCX File, 22 KB - jmir_v28i1e77788_app1.docx]

Checklist 1
PRISMA 2020 checklist.
[PDF File, 169 KB - jmir_v28i1e77788_app2.pdf]

Checklist 2
PRISMA-S checklist.
[DOCX File, 15 KB - jmir_v28i1e77788_app3.docx]

References

1. Lu X, Xie Q, Pan X, et al. Type 2 diabetes mellitus in adults: pathogenesis, prevention and therapy. Sig Transduct Target Ther 2024 Oct 2;9(1):262. [doi: 10.1038/s41392-024-01951-9]
2. IDF global clinical practice recommendations for managing type 2 diabetes 2025. Diabetes Res Clin Pract 2025 Jun;224:112238. [doi: 10.1016/j.diabres.2025.112238]
3. Udler MS, McCarthy MI, Florez JC, Mahajan A. Genetic risk scores for diabetes diagnosis and precision medicine. Endocr Rev 2019 Dec 1;40(6):1500-1520. [doi: 10.1210/er.2019-00088] [Medline: 31322649]
4. Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nat Rev Endocrinol 2018 Feb;14(2):88-98. [doi: 10.1038/nrendo.2017.151] [Medline: 29219149]
5. DeFronzo RA, Ferrannini E. Insulin resistance: a multifaceted syndrome responsible for NIDDM, obesity, hypertension, dyslipidemia, and atherosclerotic cardiovascular disease. Diabetes Care 1991 Mar 1;14(3):173-194. [doi: 10.2337/diacare.14.3.173]
6. Reaven GM. Banting lecture 1988. role of insulin resistance in human disease. Diabetes 1988 Dec;37(12):1595-1607. [doi: 10.2337/diab.37.12.1595] [Medline: 3056758]
7. Rossetti G, Ermer JA, Stentenbach M, et al. A common genetic variant of a mitochondrial RNA processing enzyme predisposes to insulin resistance. Sci Adv 2021 Sep 24;7(39):eabi7514. [doi: 10.1126/sciadv.abi7514] [Medline: 34559558]
8. Cowie CC, Harris MI, Silverman RE, Johnson EW, Rust KF. Effect of multiple risk factors on differences between blacks and whites in the prevalence of non-insulin-dependent diabetes mellitus in the United States. Am J Epidemiol 1993 Apr 1;137(7):719-732. [doi: 10.1093/oxfordjournals.aje.a116732] [Medline: 8484363]
9. Ferrannini E, Gastaldelli A, Matsuda M, et al. Influence of ethnicity and familial diabetes on glucose tolerance and insulin action: a physiological analysis. J Clin Endocrinol Metab 2003 Jul;88(7):3251-3257. [doi: 10.1210/jc.2002-021864] [Medline: 12843172]

XSL•FO
RenderX

10. Tam WH, Ma RW, Yang X, et al. Cardiometabolic risk in Chinese women with prior gestational diabetes: a 15-year follow-up study. Gynecol Obstet Invest 2012;73(2):168-176. [doi: 10.1159/000329339] [Medline: 22179684]
11. Lisco G, Giagulli VA, De Pergola G, et al. Chronic stress as a risk factor for type 2 diabetes: endocrine, metabolic, and immune implications. Endocr Metab Immune Disord Drug Targets 2024 Mar;24(3):321-332. [doi: 10.2174/1871530323666230803095118]
12. Wicke FS, Otten D, Schulz A, et al. Current and past depression as risk factors for incident type 2 diabetes mellitus and pre-diabetes in men and women: evidence from a longitudinal community cohort. Diabetol Metab Syndr 2024;16(1). [doi: 10.1186/s13098-024-01273-4]
13. Darraj A. The link between sleeping and type 2 diabetes: a systematic review. Cureus 2023 Nov;15(11):e48228. [doi: 10.7759/cureus.48228] [Medline: 38050514]
14. Hu S, Ji W, Zhang Y, Zhu W, Sun H, Sun Y. Risk factors for progression to type 2 diabetes in prediabetes: a systematic review and meta-analysis. BMC Public Health 2025 Mar 31;25(1):1220. [doi: 10.1186/s12889-025-21404-4]
15. Tabák AG, Herder C, Rathmann W, Brunner EJ, Kivimäki M. Prediabetes: a high-risk state for diabetes development. The Lancet 2012 Jun;379(9833):2279-2290. [doi: 10.1016/S0140-6736(12)60283-9]
16. Nah EH, Chu J, Kim S, Cho S, Kwon E. Efficacy of lifestyle interventions in the reversion to normoglycemia in Korean prediabetics: one-year results from a randomised controlled trial. Prim Care Diabetes 2019 Jun;13(3):212-220. [doi: 10.1016/j.pcd.2018.11.017] [Medline: 30583933]
17. Herman WH, Hoerger TJ, Brandle M, et al. The cost-effectiveness of lifestyle modification or metformin in preventing type 2 diabetes in adults with impaired glucose tolerance. Ann Intern Med 2005 Mar 1;142(5):323-332. [doi: 10.7326/0003-4819-142-5-200503010-00007] [Medline: 15738451]
18. Diabetes Prevention Program (DPP) Research Group. The diabetes prevention program (DPP): description of lifestyle intervention. Diabetes Care 2002 Dec;25(12):2165-2171. [doi: 10.2337/diacare.25.12.2165] [Medline: 12453955]
19. Auster-Gussman LA, Lockwood KG, Graham SA, Stein N, Branch OH. Reach of a fully digital diabetes prevention program in health professional shortage areas. Popul Health Manag 2022 Aug;25(4):441-448. [doi: 10.1089/pop.2021.0283] [Medline: 35200043]
20. Parsons AS, Raman V, Starr B, Zezza M, Rehm CD. Medicare underpayment for diabetes prevention program: implications for DPP suppliers. Am J Manag Care 2018 Oct;24(10):475-478. [Medline: 30325189]
21. Ely EK, Gruss SM, Luman ET, et al. A national effort to prevent type 2 diabetes: participant-level evaluation of CDC's national diabetes prevention program. Diabetes Care 2017 Oct 1;40(10):1331-1341. [doi: 10.2337/dc16-2099]
22. Lee EC, Grigorescu V, Enogieru I. Updated National Survey Trends in Telehealth Utilization and Modality (2021-2022): Issue Brief: Washington (DC): Office of the Assistant Secretary for Planning and Evaluation (ASPE); 2023. [Medline: 38913813]
23. Iyengar V, Wolf A, Brown A, Close K. Challenges in diabetes care: can digital health help address them? Clin Diabetes 2016 Jul;34(3):133-141. [doi: 10.2337/diaclin.34.3.133] [Medline: 27621530]
24. Haseldine C, O'Donoghue G, Kearney PM, et al. Characteristics of participants in the first fully online national diabetes prevention programme: a quantitative survey. HRB Open Res 2023;6:61. [doi: 10.12688/hrbopenres.13807.4] [Medline: 39931144]
25. Peng RX. How online searches fuel health anxiety: investigating the link between health-related searches, health anxiety, and future intention. Comput Human Behav 2022 Nov;136:107384. [doi: 10.1016/j.chb.2022.107384]
26. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9]
27. Paige SR, Stellefson M, Krieger JL, Anderson-Lewis C, Cheong J, Stopka C. Proposing a transactional model of eHealth literacy: concept analysis. J Med Internet Res 2018 Oct 2;20(10):e10175. [doi: 10.2196/10175]
28. Milanti A, Norman C, Chan DNS, So WKW, Skinner H. eHealth literacy 3.0: updating the Norman and Skinner 2006 model. J Med Internet Res 2025 Mar 11;27(1):e70112. [doi: 10.2196/70112]
29. Nagori A, Keshvani N, Patel L, Dhruve R, Sumarsono A. Electronic health Literacy gaps among adults with diabetes in the United States: role of socioeconomic and demographic factors. Prev Med Rep 2024 Nov;47:102895. [doi: 10.1016/j.pmedr.2024.102895] [Medline: 39823076]
30. Hsu W, Chiang C, Yang S. The effect of individual factors on health behaviors among college students: the mediating effects of eHealth literacy. J Med Internet Res 2014 Dec 12;16(12):e287. [doi: 10.2196/jmir.3542] [Medline: 25499086]
31. Busse TS, Nitsche J, Kernebeck S, et al. Approaches to improvement of digital health literacy (eHL) in the context of person-centered care. Int J Environ Res Public Health 2022 Jul 7;19(14):8309. [doi: 10.3390/ijerph19148309]
32. Watkins I, Xie B. eHealth literacy interventions for older adults: a systematic review of the literature. J Med Internet Res 2014 Nov 10;16(11):e225. [doi: 10.2196/jmir.3318] [Medline: 25386719]
33. Kim K, Shin S, Kim S, Lee E. The relation between eHealth literacy and health-related behaviors: systematic review and meta-analysis. J Med Internet Res 2023 Jan 30;25:e40778. [doi: 10.2196/40778]
34. Jacobs RJ, Lou JQ, Ownby RL, Caballero J. A systematic review of eHealth interventions to improve health literacy. Health Informatics J 2016 Jun;22(2):81-98. [doi: 10.1177/1460458214534092] [Medline: 24916567]

XSL•FO
RenderX

35. Verweel L, Newman A, Michaelchuk W, Packham T, Goldstein R, Brooks D. The effect of digital interventions on related health literacy and skills for individuals living with chronic diseases: a systematic review and meta-analysis. Int J Med Inform 2023 Sep;177:105114. [doi: 10.1016/j.ijmedinf.2023.105114]
36. McDaniel CC, Kavookjian J, Whitley HP. Telehealth delivery of motivational interviewing for diabetes management: a systematic review of randomized controlled trials. Patient Educ Couns 2022 Apr;105(4):805-820. [doi: 10.1016/j.pec.2021.07.036] [Medline: 34366228]
37. Bingham JM, Black M, Anderson EJ, et al. Impact of telehealth interventions on medication adherence for patients with type 2 diabetes, hypertension, and/or dyslipidemia: a systematic review. Ann Pharmacother 2021 May;55(5):637-649. [doi: 10.1177/1060028020950726]
38. Bassi G, Mancinelli E, Dell'Arciprete G, Rizzi S, Gabrielli S, Salcuni S. Efficacy of eHealth interventions for adults with diabetes: a systematic review and meta-analysis. Int J Environ Res Public Health 2021 Aug 26;18(17):8982. [doi: 10.3390/ijerph18178982] [Medline: 34501572]
39. Duong T, Olsen Q, Menon A, et al. Digital health interventions to prevent type 2 diabetes mellitus: systematic review. J Med Internet Res 2025 Apr 25;27:e67507. [doi: 10.2196/67507] [Medline: 40280555]
40. Suhlrie L, Ayyagari R, Mba C, et al. The effectiveness of telemedicine in the prevention of type 2 diabetes mellitus: a systematic review and meta-analysis of interventions. Diabetes Metab Syndr Clin Res Rev 2025 May;19(5):103252. [doi: 10.1016/j.dsx.2025.103252]
41. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
42. Eysenbach G. What is e-health? J Med Internet Res 2001;3(2):E20. [doi: 10.2196/jmir.3.2.e20] [Medline: 11720962]
43. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 5;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
44. Hong QN, Fàbregues S, Bartlett G, et al. The mixed methods appraisal tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018 Dec 18;34(4):285-291. [doi: 10.3233/EFI-180221]
45. Pace R, Pluye P, Bartlett G, et al. Testing the reliability and efficiency of the pilot mixed methods appraisal tool (MMAT) for systematic mixed studies review. Int J Nurs Stud 2012 Jan;49(1):47-53. [doi: 10.1016/j.ijnurstu.2011.07.002] [Medline: 21835406]
46. Atkins L, Francis J, Islam R, et al. A guide to using the theoretical domains framework of behaviour change to investigate implementation problems. Implement Sci 2017 Jun 21;12(1):77 [FREE Full text] [doi: 10.1186/s13012-017-0605-9] [Medline: 28637486]
47. Naeem M, Ozuem W, Howell K, Ranfagni S. A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. Int J Qual Methods 2023 Oct;22. [doi: 10.1177/16094069231205789]
48. Genapathy S, Mohamad E, Damanhuri H, et al. Communication matters: a systematic literature review on malaria prevention behaviour intervention. J Komun. Malays J Commun 2024;40(4):269-286. [doi: 10.17576/JKMJC-2024-4004-15]
49. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. Nurs Health Sci 2013 Sep;15(3):398-405. [doi: 10.1111/nhs.12048] [Medline: 23480423]
50. Srivastava P, Verma A, Geronimo C, Button TM. Behavior stages of a physician- and coach-supported cloud-based diabetes prevention program for people with prediabetes. SAGE Open Med 2019;7:2050312119841986. [doi: 10.1177/2050312119841986] [Medline: 31105938]
51. Kim SE, Castro Sweet CM, Cho E, Tsai J, Cousineau MR. Evaluation of a digital diabetes prevention program adapted for low-income patients, 2016-2018. Prev Chronic Dis 2019 Nov 27;16:E155. [doi: 10.5888/pcd16.190156] [Medline: 31775010]
52. Griauzde D, Kullgren JT, Liestenfeltz B, et al. A mobile phone-based program to promote healthy behaviors among adults with prediabetes who declined participation in free diabetes prevention programs: mixed-methods pilot randomized controlled trial. JMIR Mhealth Uhealth 2019 Jan 9;7(1):e11267. [doi: 10.2196/11267] [Medline: 30626566]
53. Alwashmi MF, Mugford G, Abu-Ashour W, Nuccio M. A digital diabetes prevention program (transform) for adults with prediabetes: secondary analysis. JMIR Diabetes 2019 Jul 26;4(3):e13904. [doi: 10.2196/13904] [Medline: 31350833]
54. Islam S, Elaiho C, Arniella G, Rivera S, Vangeepuram N. A pilot study to examine the feasibility and acceptability of a virtual adaptation of an in-person adolescent diabetes prevention program. IJERPH 2022 Sep 27;19(19):12286. [doi: 10.3390/ijerph191912286]
55. Katula JA, Dressler EV, Kittel CA, et al. Effects of a digital diabetes prevention program: an RCT. Am J Prev Med 2022 Apr;62(4):567-577. [doi: 10.1016/j.amepre.2021.10.023] [Medline: 35151522]
56. Lim SL, Ong KW, Johal J, et al. A smartphone app-based lifestyle change program for prediabetes (D'LITE study) in a multiethnic asian population: a randomized controlled trial. Front Nutr 2022;8:780567. [doi: 10.3389/fnut.2021.780567]
57. Batten R, Alwashmi MF, Mugford G, Nuccio M, Besner A, Gao Z. A 12-month follow-up of the effects of a digital diabetes prevention program (VP transform for prediabetes) on weight and physical activity among adults with prediabetes: secondary analysis. JMIR Diabetes 2022 Jan 14;7(1):e23243. [doi: 10.2196/23243]
58. Fitzpatrick SL, Mayhew M, Rawlings AM, et al. Evaluating the Implementation of a digital diabetes prevention program in an integrated health care delivery system among older adults: results of a natural experiment. Clin Diabetes 2022 Jul 1;40(3):345-353. [doi: 10.2337/cd21-0114]

XSL•FO
RenderX

59. Collins C, McCallum E, Nordby K, et al. Enrollment, engagement, and effectiveness of a large-scale diabetes prevention program using synchronous distance technology. Diabetology 2023 Dec 11;4(4):574-585. [doi: 10.3390/diabetology4040051]
60. Yeh MC, Lau W, Keady CA, et al. Evaluation of feasibility and acceptability of a web-based diabetes prevention program (DPP) for diabetes risk reduction in Chinese Americans in New York City. Front Public Health 2023;11:1199746. [doi: 10.3389/fpubh.2023.1199746]
61. Zahedani AD, McLaughlin T, Veluvali A, et al. Digital health application integrating wearable data and behavioral patterns improves metabolic health. NPJ Digit Med 2023 Nov 25;6(1):216. [doi: 10.1038/s41746-023-00956-y] [Medline: 38001287]
62. Chung HW, Tai CJ, Chang P, Su WL, Chien LY. The effectiveness of a traditional Chinese medicine–based mobile health app for individuals with prediabetes: randomized controlled trial. JMIR Mhealth Uhealth 2023 Jun 20;11:e41099. [doi: 10.2196/41099]
63. Vadheim LM, McPherson C, Kassner DR, et al. Adapted diabetes prevention program lifestyle intervention can be effectively delivered through telehealth. Diabetes Educ 2010;36(4):651-656. [doi: 10.1177/0145721710372811] [Medline: 20534873]
64. Moin T, Ertl K, Schneider J, et al. Women veterans' experience with a web-based diabetes prevention program: a qualitative study to inform future practice. J Med Internet Res 2015 May 25;17(5):e127. [doi: 10.2196/jmir.4332] [Medline: 26006697]
65. Block G, Azar KM, Romanelli RJ, et al. Diabetes prevention and weight loss with a fully automated behavioral intervention by email, web, and mobile phone: a randomized controlled trial among persons with prediabetes. J Med Internet Res 2015 Oct 23;17(10):e240. [doi: 10.2196/jmir.4897] [Medline: 26499966]
66. Sepah SC, Jiang L, Ellis RJ, McDermott K, Peters AL. Engagement and outcomes in a digital diabetes prevention program: 3-year update. BMJ Open Diabetes Res Care 2017;5(1):e000422. [doi: 10.1136/bmjdrc-2017-000422] [Medline: 28948027]
67. Michaelides A, Major J, Pienkosz Jr E, Wood M, Kim Y, Toro-Ramos T. Usefulness of a novel mobile diabetes prevention program delivery platform with human coaching: 65-week observational follow-up. JMIR Mhealth Uhealth 2018 May 3;6(5):e93. [doi: 10.2196/mhealth.9161]
68. Everett E, Kane B, Yoo A, Dobs A, Mathioudakis N. A novel approach for fully automated, personalized health coaching for adults with prediabetes: pilot clinical trial. J Med Internet Res 2018;20(2):e72. [doi: 10.2196/jmir.9723]
69. Castro Sweet CM, Chiguluri V, Gumpina R, et al. Outcomes of a digital health program with human coaching for diabetes risk reduction in a medicare population. J Aging Health 2018 Jun;30(5):692-710. [doi: 10.1177/0898264316688791] [Medline: 28553807]
70. Toro-Ramos T, Michaelides A, Anton M, et al. Mobile delivery of the diabetes prevention program in people with prediabetes: randomized controlled trial. JMIR Mhealth Uhealth 2020 Jul 8;8(7):e17842. [doi: 10.2196/17842] [Medline: 32459631]
71. Ng CM, Teoh KW, Chong CW, Ng YL, Bell JS, Lee SWH. A digital health-supported prediabetes intervention, management and evaluation (PRIME) program improves healthy eating index score: a cluster randomized controlled trial. Nutr Res 2025 Jun;138:151-161. [doi: 10.1016/j.nutres.2025.04.005]
72. Audet CM, Roller LJ, Aliyu MH, et al. Evaluation of a workplace diabetes prevention program delivered via distance learning: a qualitative study. Health Educ Behav 2025 Apr;52(2):156-165. [doi: 10.1177/10901981241285433]
73. Ekezie W, Dallosso H, Saravanan P, Khunti K, Hadjiconstantinou M. Experiences of using a digital type 2 diabetes prevention application designed to support women with previous gestational diabetes. BMC Health Serv Res 2021 Aug 5;21(1):772. [doi: 10.1186/s12913-021-06791-9] [Medline: 34348719]
74. Sepah SC, Jiang L, Peters AL. Long-term outcomes of a web-based diabetes prevention program: 2-year results of a single-arm longitudinal study. J Med Internet Res 2015 Apr 10;17(4):e92. [doi: 10.2196/jmir.4052] [Medline: 25863515]
75. Stein N, Brooks K. A fully automated conversational artificial intelligence for weight loss: longitudinal observational study among overweight and obese adults. JMIR Diabetes 2017 Nov 1;2(2):e28. [doi: 10.2196/diabetes.8590] [Medline: 30291087]
76. Gerber S, Silver RE, Das SK, et al. Development and feasibility of an eHealth diabetes prevention program adapted for older adults—results from a randomized control pilot study. Nutrients 2024 Mar 23;16(7):930. [doi: 10.3390/nu16070930]
77. Ferrara A, McDonald JC, Brown SD, et al. Comparative effectiveness of 2 diabetes prevention lifestyle programs in the workplace: the city and county of San Francisco diabetes prevention trial. Prev Chronic Dis 2020 May 28;17:E38. [doi: 10.5888/pcd17.190396] [Medline: 32463786]
78. Burd C, Gruss S, Albright A, Zina A, Schumacher P, Alley D. Translating knowledge into action to prevent type 2 diabetes: medicare expansion of the national diabetes prevention program lifestyle intervention. Milbank Q 2020 Mar;98(1):172-196. [doi: 10.1111/1468-0009.12443] [Medline: 31994260]
79. Ryan RM, Deci EL. Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness: The Guilford Press; 2017. [doi: 10.1521/978.14625/28806]
80. Patel SB, Iqbal FM, Lam K, Acharya A, Ashrafian H, Darzi A. Characterizing behaviors that influence the implementation of digital-based interventions in health care: systematic review. J Med Internet Res 2025 Jun 12;27:e56711. [doi: 10.2196/56711] [Medline: 40505130]
81. Chester DS, Lynam DR, Milich R, Powell DK, Andersen AH, DeWall CN. How do negative emotions impair self-control? a neural model of negative urgency. Neuroimage 2016 May;132:43-50. [doi: 10.1016/j.neuroimage.2016.02.024]
82. Levin-Zamir D, Bertschi I. Media health literacy, eHealth literacy, and the role of the social environment in context. Int J Environ Res Public Health 2018 Aug 3;15(8):1643. [doi: 10.3390/ijerph15081643]

XSL•FO
RenderX

83. Zanchetta MS, Poureslami IM. Health literacy within the reality of immigrants' culture and language. Can J Public Health 2006;97 Suppl 2:S26-S30. [Medline: 16805158]
84. Liao YS, Tsai WC, Chiu LT, Kung PT. Educational attainment affects the diagnostic time in type 2 diabetes mellitus and the mortality risk of those enrolled in the diabetes pay-for-performance program. Health Policy 2023 Dec;138:104917. [doi: 10.1016/j.healthpol.2023.104917] [Medline: 37776765]
85. Khalifa M, Albadawy M. Artificial intelligence for diabetes: enhancing prevention, diagnosis, and effective management. Comput Methods Programs Biomed Update 2024;5:100141. [doi: 10.1016/j.cmpbup.2024.100141]
86. El Benny M, Kabakian-Khasholian T, El-Jardali F, Bardus M. Application of the eHealth literacy model in digital health interventions: scoping review. J Med Internet Res 2021 Jun 3;23(6):e23473. [doi: 10.2196/23473] [Medline: 34081023]
87. Lee EH, Lee YW, Kang EH, Kang HJ. Relationship between electronic health literacy and self-management in people with type 2 diabetes using a structural equation modeling approach. J Nurs Res 2024 Jan 1;32(1):e315. [doi: 10.1097/jnr.0000000000000588]
88. Lee S. Association between Korean adults' electronic health literacy and active participation in health decision-making. Comput Inform Nurs 2025 Jan 1;43(1):e01175. [doi: 10.1097/CIN.0000000000001175]
89. Cheun J. How eHealth Literacy Impacts Patient-Provider Relationships: A Study on Trust, Self-Care, and Patient Satisfaction [Doctoral]: University of North Texas; 2017. [doi: 10.12794/metadc1011860]
90. Bello CB, Balogun MO, Ogundipe L, Olubiyi SK, Bamigboye TO, Esan DT. Influence of eHealth literacy and health promotion behavior on body mass index of workers in the public sector. SAGE Open Nurs 2024;10:23779608241274253. [doi: 10.1177/23779608241274253] [Medline: 39165911]
91. Sadeghi S, Borzu ZA, Shirzadi S, Jahanfar S, Abadi YG, Taziki M. Investigating e-health literacy and its relationship with self-care behaviors among Iranian middle-aged type 2 diabetes: a cross-sectional study. BMC Public Health 2025 Dec;25(1):1376. [doi: 10.1186/s12889-025-22648-w]
92. Shudayfat T, Hani SB, Shdaifat E, Al-Mugheed K, Alsenany SA, Farghaly Abdelaliem SM. Electronic health literacy and its association with lifestyle behavior among undergraduate students: a cross-sectional survey. Digit Health 2023;9:20552076231185429. [doi: 10.1177/20552076231185429] [Medline: 37434731]

Abbreviations

AI: artificial intelligence
CDC: Centers for Disease Control and Prevention
DPP: Diabetes Prevention Program
eHL: eHealth literacy
HbA_{1c}: glycosylated hemoglobin
HCAI: Artificial Intelligence Health Coaching
MMAT: Mixed Methods Appraisal Tool
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search
RQ: research question
T2D: type 2 diabetes
TDF: Theoretical Domains Framework
WOS: Web of Science

Edited by N Cahill, S Brini; submitted 20.May.2025; peer-reviewed by M Tanhapour, S Ranjbaran, V Mohan; revised version received 11.Jan.2026; accepted 13.Jan.2026; published 10.Mar.2026.

Please cite as:
Li J, Azlan AA, Hassim N, Wang Y, Guo R
eHealth Literacy and Type 2 Diabetes Prevention Among At-Risk Populations: Mechanistic Systematic Review Using Theory-Driven Thematic Analysis
J Med Internet Res 2026;28:e77788
URL: https://www.jmir.org/2026/1/e77788
doi:10.2196/77788

© Jingyi Li, Arina Anis Azlan, Nurzihan Hassim, Yuan Wang, Ruina Guo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is

XSL•FO
RenderX

properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Artificial Intelligence Applications in Medical Devices for Personalized Health Care Solutions: Systematic Review

Hemanth Ponnambalath Mohanadas[1], MS; A Manikandan[2], PhD; Ahmad Fauzi Ismail[3], PhD; Nick Tucker[4], PhD; Saravana Kumar Jaganathan[5,6,7], PhD

[1]Device Development, Abbott Diabetes Care, Alameda, CA, United States

[2]Centre for Material Chemistry, Department of Chemistry, Karpagam Academy of Higher Education, Coimbatore, India

[3]School of Chemical and Energy Engineering, Advanced Membrane Technology Research Centre (AMTEC), University of Technology Malaysia, Skudai, Malaysia

[4]School of Engineering and Physical Sciences, College of Science, University of Lincoln, Lincoln, United Kingdom

[5]Institute of Research and Development, Duy Tan University, 03 Quang Trung, Da Nang, Vietnam

[6]School of Engineering & Technology, Duy Tan University, Da Nang, Vietnam

[7]Biomedical Engineering Research Group, School of Engineering, University of Leicester, Leicester, United Kingdom

Corresponding Author:
Saravana Kumar Jaganathan, PhD
Institute of Research and Development, Duy Tan University, 03 Quang Trung, Da Nang, Vietnam

Abstract

Background: The integration of artificial intelligence (AI) in medical devices is transforming health care by enabling enhanced personalization and precision medicine. AI-driven medical devices can tailor treatments based on individual patient profiles, including genetic data, medical history, and physiological parameters. This advancement holds the potential to refine therapeutic interventions, improve patient outcomes, and streamline health care delivery. However, challenges such as data quality, algorithmic bias, patient privacy, and regulatory complexities hinder the full realization of AI-driven personalization. By 2030, the global AI in health care market is projected to exceed US $187.95 billion, growing at a compound annual growth rate of 37% from US $15.1 billion in 2022.

Objective: This review aims to explore the scope and impact of AI-driven personalization in medical devices. It seeks to analyze key technological innovations that have enabled AI integration, identify the critical challenges impeding progress, and evaluate strategies to address these challenges. Additionally, it highlights future research directions and innovation opportunities in this evolving field.

Methods: A systematic review was conducted, drawing from scholarly literature, industry analyses, and regulatory advisories. Relevant studies and case examples were analyzed to assess the current applications of AI in medical devices, the barriers to its implementation, and best practices for overcoming these barriers. Ethical, technical, and regulatory considerations were also examined. The review included studies published between 2016 and 2023, covering over 100 peer-reviewed articles and reports.

Results: The review highlights significant advancements in AI-driven medical devices, including applications in diagnostics, treatment personalization, wearable health monitoring, and smart prosthetics. AI-based diagnostic tools have achieved up to 98.88% accuracy in multiclass disease classification from X-ray images and 95% accuracy in insulin injection site recognition. It identifies key challenges such as data security risks, algorithmic biases, regulatory constraints, and integration issues with existing health care infrastructures. Currently, more than 70% of clinical decisions rely on diagnostic tests, yet AI-driven automation could reduce diagnostic delays by up to 50%. Several strategies, including improved data validation techniques, regulatory frameworks for AI approval, and ethical guidelines, were found to be effective in mitigating these challenges. Case studies demonstrate how AI has enhanced medical device functionality and patient outcomes.

Conclusions: AI-driven personalization in medical devices holds immense potential to revolutionize health care, offering more precise, adaptive, and patient-centered solutions. However, successful implementation requires addressing technical, ethical, and regulatory challenges. Emerging technologies such as quantum computing could improve AI-driven medical diagnoses by 10 - 20 times in processing efficiency, while blockchain-based patient data management could reduce security breaches by more than 30%. This review serves as a valuable resource for researchers, health care professionals, policymakers, and industry leaders, fostering informed discussions and guiding future advancements in AI-enabled personalized medicine.

(J Med Internet Res 2026;28:e72410) doi:10.2196/72410

XSL•FO
RenderX

KEYWORDS

artificial intelligence; personalized medicine; medical devices; digital health; predictive analytics

Introduction

Background

The intersection of artificial intelligence (AI) with medical device technology marks a cornerstone in the evolution of health care, introducing a new epoch where personalization and precision medicine are not just aspirational goals but achievable realities [1]. AI refers to computational systems that perform tasks typically requiring human intelligence, such as learning and decision-making. Personalization in this context is defined as the adaptation of medical devices and interventions to individual patient characteristics, enabled by AI algorithms. AI-driven personalization thus describes the integration of AI into medical devices to dynamically tailor diagnostics, treatments, and monitoring to each patient. This transformative synergy is at the forefront of a health care revolution, offering personalized health care solutions that were once deemed futuristic. Despite these advancements, existing research lacks a comprehensive analysis of how AI-driven personalization is transforming medical devices, particularly in terms of regulatory challenges, clinical adoption, and long-term impact on patient outcomes. While studies have explored AI applications in diagnostics and predictive analytics, a holistic evaluation of AI's role across the entire medical device spectrum, spanning real-time monitoring, adaptive therapeutics, and regulatory compliance, remains underexplored.

Based on the Precedence Research report, the global AI in the health care market was valued at US $15.1 billion in 2022 and is projected to exceed US $187.95 billion by 2030, growing at a compound annual growth rate of 37% during the forecast period. This substantial growth highlights the increasing adoption of AI technologies in health care, driven by their potential to enhance patient care and operational efficiency. AI-driven personalization in medical devices epitomizes a leap toward a future where medical care is uniquely customized to everyone's genetic, environmental, and lifestyle specifics, revolutionizing patient care and treatment outcomes [2]. However, despite rapid AI integration, real-world deployment remains constrained by technical limitations, algorithmic biases, and evolving regulatory frameworks. This underscores the need for a systematic evaluation of AI-driven medical devices, addressing both technological breakthroughs and the challenges hindering large-scale implementation.

For instance, additive manufacturing (3D printing) is being harnessed to create bespoke implants and prosthetics tailored to the unique anatomical features of patients, enhancing the effectiveness and comfort of these medical solutions [3]. While promising, the clinical validation of these AI-driven medical solutions remains a major hurdle, with only a fraction of AI-based medical devices receiving regulatory approval due to concerns over accuracy, interpretability, and patient safety.

AI, at its core, involves the development of algorithms and systems that can perform tasks requiring human intelligence, such as learning, reasoning, problem-solving, and decision-making [4]. In the medical device sector, AI technologies are used to process vast amounts of health data to uncover patterns and insights that inform clinical decisions. This includes analyzing genomic data, electronic health records, biometric readings, and real-time physiological data [5]. Advanced AI algorithms and machine learning (ML) techniques are crucial in creating medical devices that can adapt, predict, and respond to the specific needs of individual patients, offering unprecedented levels of accuracy and efficiency [6]. Medical devices integrated with AI range from diagnostic tools capable of early disease detection to therapeutic devices that adjust treatment protocols in real time [7]. These devices not only improve the precision of diagnostics and treatment but also facilitate continuous monitoring and proactive health care management. Applications of AI in medical devices are diverse, including AI-powered imaging tools for cancer detection [8], smart prosthetics that adjust to physiological changes [9], and wearable technologies for real-time health tracking [10]. The regulation of AI in medical devices is a complex but essential aspect to ensure patient safety and efficacy of the devices [11]. Regulatory bodies such as the US Food and Drug Administration (FDA) and the European Medicines Agency are continuously evolving their frameworks to keep pace with rapid advancements in AI technology. These regulations cover various aspects, such as algorithm transparency, data security, patient privacy, and postmarket surveillance [12]. Despite these efforts, regulatory inconsistencies across regions, coupled with the lack of standardized validation protocols for AI-based medical devices, remain a critical research gap. Understanding these challenges is vital for the ethical and effective deployment of AI-driven medical devices.

The scope of AI-driven personalization in medical devices is broad, encompassing a diverse array of applications that promise to enhance disease diagnosis, treatment efficacy, and patient monitoring. For instance, AI-powered imaging tools are revolutionizing diagnostics by providing deeper insights into medical images, enabling early detection of conditions such as cancer with higher precision than ever before [13]. In the realm of treatment, we are witnessing the emergence of smart prosthetics and implants that adapt to patients' physiological changes and usage patterns, thereby improving functionality and comfort [14]. Furthermore, wearable technologies equipped with AI algorithms are transforming patient monitoring by offering continuous, real-time health tracking and analysis, facilitating early intervention and personalized health management strategies [15].

Despite the great promise of AI in personalizing medical devices, there are many challenges to overcome, including technical, ethical, and regulatory issues. Technical challenges include ensuring the accuracy, reliability, and robustness of AI algorithms, as well as securing high-quality, representative data to train these models effectively [16]. Ethical considerations revolve around issues such as algorithmic bias, which can lead to unequal treatment outcomes, and the imperative to maintain patient privacy and data security in an era of ubiquitous data

XSL•FO
RenderX

collection [17]. Furthermore, navigating the regulatory landscape presents its own set of hurdles, as policymakers and regulatory bodies strive to keep pace with the rapid advancements in AI technology while ensuring patient safety and the efficacy of AI-integrated medical devices [18].

This study aims to bridge these gaps by systematically reviewing the role of AI in medical devices, with a focus on its applications, regulatory landscape, and clinical implementation. The objectives of this review are threefold: (1) to explore and categorize AI-driven advancements in medical devices, including their technological underpinnings and real-world use cases; (2) to identify and critically assess key challenges, technical, ethical, and regulatory, affecting AI adoption in medical devices; and (3) to propose strategic recommendations for overcoming these barriers, facilitating a smoother transition of AI-powered devices from research to clinical practice. Our hypothesis is that while AI-driven personalization has the potential to revolutionize medical device functionality, its widespread adoption remains limited due to unresolved challenges in data validation, regulatory approval, and integration into existing health care infrastructures. By systematically analyzing these aspects, this review aims to provide actionable insights for researchers, policymakers, and industry leaders striving to harness AI for personalized medicine.

Historical Development of AI in Medical Devices

The integration of AI into medical devices has roots that trace back to the mid-20th century, with the advent of computers and early AI research [19]. Initially, these endeavors focused on automating simple tasks and analyzing medical data. A pivotal moment came with the development of expert systems in the 1970s and 1980s, such as MYCIN, a computer-based consultation system designed to assist physicians in diagnosing bacterial infections and recommending antibiotics [20]. These systems laid the groundwork for the incorporation of AI in health care by demonstrating the potential of machine-based decision-making support.

The subsequent decades saw incremental advancements, with AI applications expanding into diagnostic imaging, patient monitoring, and robotic surgery. However, the real explosion of AI in medical devices coincided with the advent of more sophisticated ML algorithms and the exponential growth in computational power and data availability in the 21st century [21]. This era ushered in devices capable of complex analyses and decisions, from predicting patient outcomes to personalizing treatment plans based on individual patient data.

Basic Principles of AI and ML Relevant to Health Care

AI encompasses a broad range of technologies that enable machines to sense, comprehend, act, and learn. ML, a subset of AI, involves the development of algorithms that can learn from and make predictions or decisions based on data. This is particularly relevant in health care, where ML models are trained on vast datasets of medical records, images, and genetic information to identify patterns and anomalies [6].

Among the diverse techniques under the ML umbrella, 2 approaches are notably significant for personalizing medical devices. Supervised learning, with its basis in training on labeled datasets, proves indispensable in diagnostic realms such as imaging, where algorithms learn from vast databases of prediagnosed images to identify diseases. This training allows machines to recognize complex patterns and anomalies in patient data, making precise predictions or decisions that support clinical diagnoses. On the other hand, unsupervised learning ventures into the analysis of unlabeled data, uncovering hidden patterns without predefined outcomes. This approach is particularly valuable in genomic studies and identifying patient subgroups for targeted treatment approaches, where the absence of labeled data does not hinder the algorithm's ability to generate insights. Together, these methodologies underscore a transformative shift toward more personalized, data-informed health care solutions, leveraging the intrinsic patterns within vast datasets to cater to individual patient needs [22].

Further advancement within ML, deep learning, delves into complex neural networks with multiple layers, offering profound insights into the intricate datasets typical in health care, such as detailed medical imaging [23]. For instance, deep learning algorithms are used in radiology to enhance the accuracy of detecting abnormalities in X-rays and computed tomography (CT) scans. These algorithms can identify subtle signs of diseases like lung cancer at stages earlier than human radiologists might detect [24]. Another example is the use of deep learning in pathology, where neural networks analyze tissue samples to detect cancerous cells with high precision, significantly reducing the risk of misdiagnosis [25]. In cardiology, deep learning applications include analyzing electrocardiograms to predict the likelihood of heart conditions such as atrial fibrillation or heart failure [26]. These predictive models can alert health care providers to intervene early, potentially preventing serious cardiac events. This progression underscores the pivotal role of AI and ML in not only advancing medical technology but also in shaping the future of personalized health care, illustrating a journey from data to diagnosis that is increasingly nuanced and tailored to individual patient needs.

The Evolution of Medical Devices From Generic to Personalized Solutions

The journey of medical devices from one-size-fits-all to personalized solutions reflects broader shifts toward precision medicine and patient-centered care [27]. Initially, medical devices were designed for the average patient, with limited ability to adjust to individual needs [28]. However, the recognition that health care efficacy can significantly improve by considering individual differences in genetics, environment, and lifestyle catalyzed the shift toward personalization [29].

Personalized medical devices leverage AI and ML to adapt their functionality to individual patients. For example, wearable devices now monitor health metrics in real-time, adjusting treatment regimens accordingly [30]. Similarly, AI-enabled implants can adjust their behavior based on changes in the patient's physiological data, optimizing therapeutic outcomes [31].

This evolution is not just technological but also philosophical, representing a shift in health care paradigms toward more holistic, tailored care. It underscores the recognition of the complex interplay between various factors affecting health and

XSL•FO
RenderX

the need for treatments that accommodate this complexity [32]. Personalized medical devices, powered by AI, are at the forefront of this shift, offering the promise of health care that is not only more effective but also more attuned to the unique needs and circumstances of each patient.

In summary, the integration of AI into medical devices is a culmination of decades of technological advancement and a reflection of a broader shift toward personalized health care. By leveraging the power of AI and ML, medical devices are evolving to meet the unique needs of individual patients, promising a new era of precision medicine that is more effective, efficient, and patient-centered.

Methods

Overview

To comprehensively review the current state of AI-driven personalization in medical devices, a systematic literature search was conducted across multiple databases, including PubMed, IEEE Xplore, Scopus, Web of Science, Embase, and Google Scholar, ensuring broad coverage of relevant studies. To minimize potential gaps, additional sources such as preprint repositories and conference proceedings were also considered. The search strategy incorporated a diverse range of keywords and synonyms, including "artificial intelligence in medical devices," "personalized healthcare," "AI-driven diagnostics," "machine learning in medical technology," "smart prosthetics," "ethical considerations in AI-driven healthcare," and "regulatory frameworks for AI-integrated medical devices." Boolean operators (AND, OR) were systematically applied, and medical subject heading terms were used in PubMed to enhance precision, with truncation and wildcard operators used where necessary. The search was limited to peer-reviewed studies published between January 2016 and December 2023 to ensure relevance to contemporary developments. While studies were restricted to English-language publications, a cross-verification process was implemented to identify key non-English papers referenced in systematic reviews and meta-analyses, allowing for their inclusion through reliable academic translation tools. This structured search strategy was designed to maximize coverage, ensure methodological rigor, and minimize potential biases, thereby providing a comprehensive and objective review of AI-driven personalization in medical devices.

Eligibility Criteria

Studies were included if they focused on AI-driven personalization in medical devices, were published in peer-reviewed journals, and provided empirical data, whether quantitative or qualitative. Additionally, studies needed to be available in full text and in the English language. Exclusion criteria encompassed non-English articles, opinion pieces, editorials, letters, studies without accessible full text, and articles unrelated to medical devices or AI. This approach ensured the selection of high-quality and relevant studies for the review.

Study Selection and Agreement Assessment

The study selection process was conducted by the authors, who were divided into 2 independent groups. Each group systematically screened the titles and abstracts of all identified articles based on predefined inclusion and exclusion criteria. Full-text articles were retrieved for studies deemed potentially relevant. After the initial screening, both groups compared their selections, and any discrepancies were resolved through discussion. In cases where disagreements persisted, a third reviewer (SKJ) was consulted to reach a final decision, ensuring a transparent and unbiased selection process. The overall agreement rate before resolution was 90%, reflecting a high level of consistency in study selection. The detailed study selection process is depicted in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram, which illustrates the number of studies identified, screened, excluded, and included in the review. Following study selection, articles were categorized into three key themes: (1) case studies illustrating state-of-the-art AI-integrated medical devices, (2) applications of AI-driven personalization in various medical devices, and (3) challenges associated with AI-driven medical personalization. This thematic categorization enabled a structured synthesis of findings, ensuring clarity in the presentation of results.

Data Extraction

Data extraction was performed using a standardized form developed in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Scoping Reviews) to ensure consistency and methodological rigor. The extracted data included study characteristics (authors, year of publication, and country of origin), AI techniques used (ML models, algorithms, and tools), application in medical devices (type of device and clinical use case), outcomes measured (accuracy, effectiveness, and patient outcomes), and key findings (main results and conclusions). Two authors (HPM and MA) independently extracted data following this predefined protocol to enhance reliability. Any discrepancies were resolved through discussion, and if necessary, by consulting additional authors to achieve consensus. The extracted data were then synthesized according to the predefined thematic categories to align with the structure of the Results and Discussion sections. This structured approach ensured a transparent, reproducible, and comprehensive data collection process.

Reference Screening

In addition to the database searches, the reference lists of all included articles were manually screened to identify further relevant studies. This snowballing technique ensured comprehensive coverage of the literature.

Results

Overview

The results of this review paper are categorized into three major sections: (1) case studies of state-of-the-art AI-integrated medical devices, (2) applications of AI-driven personalization in medical devices, and (3) challenges in AI-driven personalization (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram illustrating the systematic review process for artificial intelligence–driven personalization in medical devices. AI: artificial intelligence.

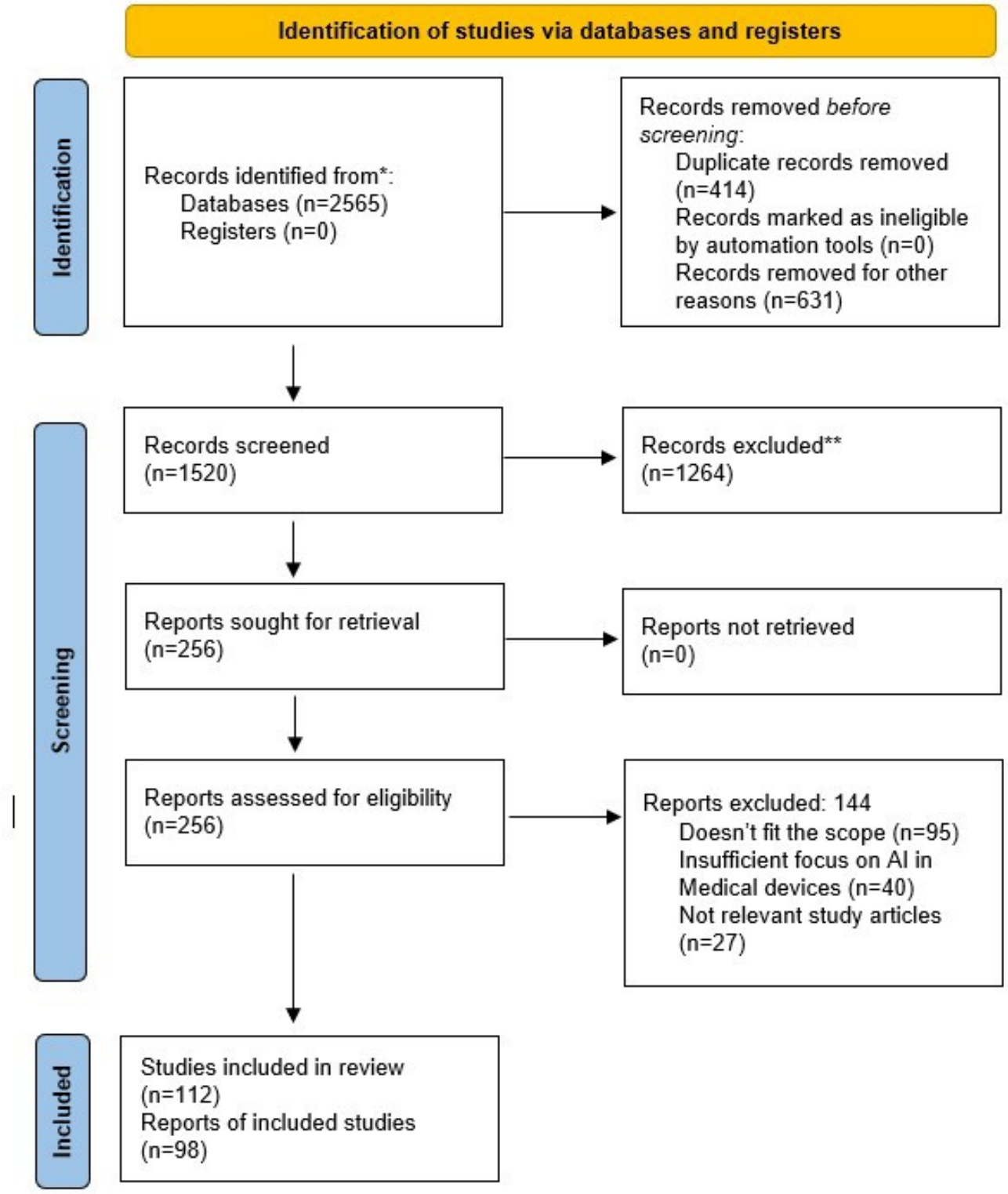

Case Studies of State-of-the-Art AI-Integrated Medical Devices

Smart Insulin Pumps

Smart insulin pumps represent a transformative advancement in diabetes management, addressing many challenges associated with traditional methods that often require frequent manual adjustments [33]. These sophisticated devices use advanced ML algorithms to process real-time data from continuous glucose monitors (CGMs). By analyzing trends and fluctuations in glucose levels, the pumps autonomously adjust insulin delivery, significantly enhancing glycemic control [34]. Adding to this innovation, recent research has successfully applied AI in smart insulin pens, achieving more than 95% accuracy in both detecting proper injection sites and identifying lipodystrophies using a sensor with dual LEDs, thereby

XSL•FO
RenderX

promoting better injection practices and reducing tissue complications [35]. Furthermore, another pivotal development includes a smart AI-based self-care device that predicts future blood glucose fluctuations to proactively manage type-1 diabetes, significantly reducing the risk of hypoglycemia and hyperglycemia by intelligently controlling insulin dosage during automatic injections [36]. The first FDA-approved smart insulin pen in 2014, which records and wirelessly transmits insulin dosing data, exemplifies these technological advances, providing clinical data to assist individualized diabetes management and overcoming barriers such as poor adherence and clinical inertia [37]. This dynamic approach not only alleviates the daily burden on patients but also improves their quality of life by minimizing the risk of glucose variability-related complications and enhancing adherence to injection therapy. Studies have shown that such automated systems can lead to better long-term health outcomes, demonstrating a critical evolution in diabetes care technology [38].

AI-Powered Prosthetics

AI-powered prosthetic limbs are revolutionizing the field of prosthetics, providing amputees with unprecedented levels of functionality and adaptability. These sophisticated devices integrate sensors and advanced AI algorithms that interpret the user's muscle signals and movements, allowing the limb to respond in a natural and intuitive manner. As these AI systems interact with the user, they continuously learn and adapt, fine-tuning their responses to better align with individual needs. This technology not only captures and interprets complex muscle signals but also adapts to the user's unique physiological changes over time, enhancing the effectiveness of rehabilitative practices, as demonstrated in recent studies on surface electromyographic-based systems for hand rehabilitation using AI-powered gesture recognition [39]. Additionally, innovations in lower limb assistive technologies, such as adaptable central pattern generators and personalized motion planning, are enhancing capabilities in gait analysis and stability, crucial for preventing falls and ensuring safe mobility in various environments [40]. Furthermore, AI applications extend beyond mobility aids to the assessment of medical conditions, where tools have significantly improved the evaluation of lower limb radiographs before and after knee arthroplasty, speeding up assessments while maintaining high accuracy [41]. This broad spectrum of AI applications illustrates a transformative impact on patient care, making daily activities and medical diagnostics more efficient and tailored to individual health needs.

Diagnostic and Imaging Tools

AI-integrated diagnostic tools have made significant strides, particularly in the field of medical imaging, where deep learning algorithms excel at detecting subtle patterns indicative of diseases like cancer, often outperforming human radiologists in accuracy. These sophisticated tools adapt their analyses based on individual patient data, greatly enhancing diagnostic precision and enabling timely interventions. For instance, deep learning's integration within imaging modalities has notably advanced pancreatic cancer diagnostics, leading to more precise detection and improved clinical decision-making [42]. Additionally, the application of AI in optical imaging techniques, such as optical coherence tomography and photoacoustic imaging, facilitates high-contrast, low-cost visualization of tumor tissues, thereby enhancing both the detection and theranostics of cancer, which is paving the way for precision oncology [43]. Moreover, the development of a novel deep learning framework for multiclass diagnosis of lung diseases from chest X-ray images shows significant improvements in accuracy and efficiency, exemplifying how AI can transform pulmonary diagnostics and thereby ensure better patient outcomes through enhanced feature extraction capabilities [44]. These advancements underscore the transformative impact of AI in medical diagnostics, revolutionizing early disease detection and treatment strategies, and thereby playing a crucial role in enhancing personalized and effective health care solutions.

AI-Powered Humanoids

AI-powered humanoids are revolutionizing various sectors by merging advanced robotics with deep learning technologies to perform complex tasks. These humanoids, equipped with AI, are capable of high-level functions such as speech recognition, comprehensive comprehension, and sophisticated interactions, making them invaluable in industries like health care, education, and space exploration. For instance, AI-based 3D-printed humanoids represent a significant leap forward, offering customizable interactions and decision-making capabilities that mimic human actions [45]. Similarly, AI-powered smart glasses enhance human-robot interaction by allowing dynamic adaptation to varied walking terrains, demonstrating the potential of embedded devices in everyday environments [46]. Furthermore, discussions about responsibly regulating AI to prevent ethical dilemmas and enhance user safety emphasize the need for knowledge about AI's operational boundaries [47]. In customer service, AI robots are researched for their emotional and functional impact on customer engagement, showcasing their potential to deliver meaningful and personalized service experiences [48]. Moreover, in health care, AI-powered avatars are being evaluated for their acceptance by patients, particularly in mental health services where they can act as assistants to physicians in resource-limited settings [49]. Emotional intelligence in robots, allowing them to recognize and respond to human emotions, is another frontier being explored to make interactions with humanoids more natural and effective [50]. Collectively, these advancements are not just enhancing the capabilities of AI humanoids but are also setting the stage for future innovations that could redefine human-machine collaboration.

Other AI-Powered Applications

AI-powered applications are expanding their reach and transforming industries with innovative and practical solutions. In dentistry, the integration of AI and augmented reality is revolutionizing both practice and education, offering new possibilities for enhanced clinical precision and interactive learning environments [51]. Similarly, the application of AI in the fabrication of 3D-printed materials is propelling the creation of patient-specific wearable devices and smart biomedical implants, streamlining manufacturing processes and advancing personalized medicine [52]. In the field of orthopedics, AI algorithms are being used to standardize the measurement of

XSL•FO
RenderX

implant alignments in radiography, improving the reliability and efficiency of postoperative assessments [53]. The introduction of AI-powered chatbots in medical consultations offers immediate, personalized medical advice, drastically enhancing patient engagement and health care delivery [54]. Moreover, AI is pivotal in field operations, where it aids in the digitalization and enhancement of inspection rounds, significantly optimizing process monitoring and anomaly detection in industrial settings [55]. Additionally, the development of AI-powered neural implants is opening new frontiers in the treatment of conditions like Alzheimer disease, enhancing communication capabilities and patient care through advanced brain-machine interfaces [56]. Each of these applications mentioned in Table 1 not only highlights the diverse capabilities of AI across various sectors but also underscores the significant impact of intelligent technology in improving operational efficiency, accuracy, and personalized solutions.

XSL•FO
RenderX

Table . Case studies: artificial intelligence applications in personalized health care devices.

Case studies	Device type	Specific AI[a] technology	Applications	Outcomes
Exploiting AI to make insulin pens smart: injection site recognition and lipodystrophy detection [35]	Smart insulin pen	Machine learning (one-class SVM[b])	Diabetes management	Achieved more than 95% accuracy in injection site recognition and lipodystrophy detection, promoting better adherence to injection therapy and improving patient quality of life
Smart AI-based self-care device for patients with diabetes [36]	Insulin pump	AI-based predictive algorithms	Managing type-1 diabetes	Intelligently controls insulin dosage during automatic injection, predicts future blood glucose levels, monitors and reports critical conditions, sends summarized reports to health care centers, maintains a timeline of data and insulin dosages
Empowering hand rehabilitation with AI-powered gesture recognition: a study of an sEMG[c]-based system [39]	Wearable rehabilitation glove	Deep learning (1D-CNN[d], InceptionTime [an advanced deep learning architecture designed for time series classification])	Hand rehabilitation	Enhanced hand gesture recognition with 90.89% accuracy, aiding in the rehabilitation of stroke patients through mirror therapy and task-oriented therapy. Improved motor function and usability in both clinical and home settings.
The role and efficiency of an AI-powered software in the evaluation of lower limb radiographs before and after total knee arthroplasty [41]	Radiograph analysis tool	AI-powered software (LAMA, U-net-based convolutional neural network)	Evaluation of lower limb radiographs to assess alignment and angles before and after total knee arthroplasty	Good to excellent agreement with expert assessments (ICC[e]=0.78 - 1.00), significant time efficiency (twice as fast as experts), but struggled with certain measurements (eg, joint line convergence angle, Mikulicz line, and patients with high BMI)
AI-driven diagnosis of pancreatic cancer [42]	Diagnostic imaging	Deep learning and ML[f]	Early detection and accurate diagnosis of pancreatic cancer	Improved accuracy in diagnosis, early detection, better clinical decision-making, and reduced mortality risk
An advanced deep learning framework for multiclass diagnosis from chest x-ray images [44]	Diagnostic tool	Deep learning (CNN[g])	Multiclass diagnosis of lung diseases (fibrosis, opacity, tuberculosis, normal, viral pneumonia, and COVID-19 pneumonia)	Achieved superior performance with an accuracy of 98.88%, precision of 0.9870, recall of 0.9904, F_1-score of 0.9887, and AUC of 0.9939
AI enables reliable and standardized measurements of implant alignment in long leg radiographs with total knee arthroplasties [53]	Diagnostic tool	AI algorithm for image analysis	Evaluation of long leg radiographs (LLR) post-total knee arthroplasties (TKA)	The AI algorithm showed excellent reliability (ICC>0.97) and reproducibility (96 % reproducible, 92.1 % reliable) in measuring hip–knee–ankle (HKA), femoral component (FCA), and tibial component (TCA) angles.
Deep neural networks improve radiologists' performance in breast cancer screening [57]	Diagnostic tool	Deep CNN	Breast cancer screening exam classification	Achieved an AUC[h] of 0.895; model as accurate as experienced radiologists; hybrid model (AI+ radiologist) more accurate than either alone

XSL•FO
RenderX

Case studies	Device type	Specific AI[a] technology	Applications	Outcomes
Deep learning for the classification of small (≤2 cm) pulmonary nodules on CT[i] imaging: a preliminary study [58]	Diagnostic tool (CT imaging)	Deep learning (CT-lungNET, a deep learning-based malignancy prediction model)	Classification of small pulmonary nodules on CT imaging	Improved AUROC[j] (0.85 vs 0.82 for AlexNET), faster processing time, and enhanced diagnostic performance for nonradiologists
AI for cardiac diseases diagnosis and prediction using ECG[k] images on embedded systems [59]	Diagnostic tool	Deep learning algorithms (MobileNetV2 and VGG16)	Analysis and classification of ECG signals to predict cardiovascular diseases	High validation accuracy of 0.95 for MobileNetV2 and VGG16; after implementation on Raspberry Pi, accuracy slightly decreased to 0.94 and 0.90, respectively. Capable of real-time monitoring using smart mobile tools, such as mobile phones, smartwatches, and connected T-shirts.
Personalized diabetes monitoring platform leveraging IoMT[l] and AI for noninvasive estimation [60]	Wearable device	Deep neural network (GlucoNet)	Noninvasive real-time blood glucose estimation	The AI-powered platform demonstrated an MAPE[m] of 17.8 % with 100 % of predictions falling in the clinically acceptable zones A and B of the Clarke-error grid. Clinically validated on over 350 patients, capable of handling large datasets, and does not require frequent calibration.
A predictive machine learning tool for asthma exacerbations: results from a 12-week, open-label study using an electronic multidose dry powder inhaler with integrated sensors [61]	Electronic Multi-Dose Dry Powder Inhaler (eMDPI)	Machine learning (gradient-boosting trees)	Predicting impending asthma exacerbations	The AI model predicted asthma exacerbations with an ROC[n] AUC of 0.83. The model was based on inhaler use data and inhalation parameters, providing a proactive approach to asthma management by identifying exacerbations up to 5 days in advance.

[a]AI: artificial intelligence.

[b]SVM: support vector machine.

[c]sEMG: surface electromyographic.

[d]1D-CNN: One-Dimensional Convolutional Neural Network.

[e]ICC: intraclass correlation coefficient.

[f]ML: machine learning.

[g]CNN: convolutional neural network.

[h]AUC: area under the curve.

[i]CT: computed tomography.

[j]AUROC: area under the receiver operating characteristic.

[k]ECG: electrocardiogram.

[l]IoMT: Internet of Medical Things.

[m]MAPE: mean absolute percentage error.

[n]ROC: receiver operating characteristic.

Applications of AI-Driven Personalization in Medical Devices

Overview

The integration of AI into medical devices has opened a new frontier in health care, enabling personalized medicine that promises more accurate diagnoses, tailored treatments, and continuous patient monitoring. This section delves into the specific applications of AI-driven personalization across various categories of medical devices, highlighting the transformative impact on patient care and health management.

XSL•FO
RenderX

Diagnostic Devices: Improvements in Early Detection and Accuracy

About 70% of clinical decisions rely on diagnostic tests, which play a key role in evidence-based patient care. However, the current method of sending samples to centralized labs often causes delays of several days in getting results. This delay can limit the usefulness of diagnostics and increase patient stress while waiting for important test results [62]. AI-driven diagnostic devices mark a transformative advance in this realm of medical diagnostics, significantly enhancing early detection and improving accuracy. These devices leverage sophisticated algorithms to sift through and analyze complex medical data, including imaging scans [63], blood tests [64], and genetic information [65]. This capability allows them to detect patterns and anomalies that might be overlooked by human evaluators.

In the field of medical imaging, AI algorithms have demonstrated exceptional proficiency. They interpret data from X-rays, magnetic resonance imaging, and CT scans more quickly and accurately than traditional methods. These algorithms excel at detecting subtle abnormalities that are often indiscernible to the human eye, thus facilitating earlier and more accurate diagnoses of conditions such as cancer, cardiovascular diseases, and neurological disorders [66]. The table below (Table 2) summarizes the distribution of FDA-approved diagnostic devices by clinical specialty, underscoring the dominant role of radiology in this field. According to the FDA article released on December 20, 2024, a total of 1016 diagnostic devices have been approved, with an overwhelming 76% (777/1016) categorized under radiology. This statistic highlights the transformative advancements in AI-powered radiology tools, which have played a pivotal role in improving early detection, especially in imaging modalities like X-rays, CT scans, and magnetic resonance imaging. The remaining categories include devices for genetic testing, blood diagnostics, and other fields, reflecting the breadth of AI integration across diverse diagnostic platforms.

Table . Distribution of US Food and Drug Administration–approved artificial intelligence or machine learning–enabled medical devices across medical specialties [67].

Categories	Number of devices (N=1016), n (%)
Radiology	777 (76)
Cardiovascular	104 (10)
Neurology	42 (4)
Other categories	27 (3)
Anesthesiology	17 (2)
Hematology	17 (2)
Clinical chemistry	9 (1)
Gastroenterology – urology	14 (1)
Ophthalmic	9 (1)

AI-powered diagnostic tools are reshaping medical imaging by surpassing human radiologists in detecting subtle disease markers, particularly in cancer diagnostics. Deep learning models, tailored to patient-specific data, have improved the early detection of diseases like pancreatic cancer and enhanced clinical decision-making [42]. AI also advances optical imaging techniques, such as optical coherence tomography and photoacoustic imaging, by offering affordable, high-contrast tumor visualization, which supports precision oncology [43]. Additionally, AI frameworks for multiclass lung disease diagnosis via chest X-rays have boosted both accuracy and efficiency, transforming pulmonary care [44].

For example, a study comparing the diagnostic performance of 14 radiologists in breast cancer screening with a deep learning model illustrates the benefits of deep convolutional neural networks (CNNs). This study demonstrates that CNNs can significantly enhance radiologists' performance, achieving an area under the curve of 0.895. Area under the curve is a statistical measure used to gauge the performance of binary classification models, representing the likelihood that the model correctly distinguishes between positive and negative cases. The study also highlights the successful integration of CNNs with radiologists' evaluations, leading to improved diagnostic outcomes [57]. Similarly, in lung cancer, AI tools applied to low-dose CT scans can identify small pulmonary nodules that may be early indicators of cancer. The precision of AI in these applications not only improves the outcomes for patients through earlier treatment but also reduces the number of false positives and unnecessary biopsies, thereby optimizing health care resources [58]. In ophthalmology, AI-driven tools are used to diagnose diseases like diabetic retinopathy and glaucoma by analyzing retinal images with high accuracy. Various AI systems, including IDx-DR (Digital Diagnostics Inc.), a tool used to diagnose diabetic retinopathy, and EyeArt (Eyenuk Inc.), a system that autonomously analyzes patients' retinal images, which have been approved for clinical use, have demonstrated high sensitivity and specificity in identifying diabetic retinopathy through color fundus photography [68]. These models are trained on large datasets of retinal images to identify specific patterns, such as microaneurysms [69], hemorrhages [70], and exudates [71], which are indicative of diabetic retinopathy. The AI algorithms can assess the severity of retinopathy by evaluating changes in the retina's vascular structure, offering a noninvasive method to monitor the progression of diabetes. Similarly, AI applications in

XSL•FO
RenderX

pulmonology interpret chest X-rays and CT scans to identify signs of lung cancer at early stages [72]. During the COVID-19 pandemic, the integration of AI-driven diagnostic tools significantly enhanced the detection and management of the virus using imaging and blood tests, as highlighted by several studies. For instance, AI algorithms applied to chest X-rays and CT scans swiftly improved the diagnostic accuracy for COVID-19, enabling better prediction of disease severity and aiding in patient management [73]. In parallel, simultaneously, innovative AI applications analyzing changes in blood parameters demonstrated the potential of routine blood tests in COVID-19 diagnosis, achieving high classification performance and offering a rapid, cost-effective alternative to standard reverse transcription polymerase chain reaction [74,75]. These advancements underline the transformative impact of AI in health care, providing crucial support in diagnosing and managing COVID-19 effectively.

AI algorithms are making significant inroads in the field of genetic testing as well. They analyze genetic data to predict the risk of genetic disorders and conditions swiftly and with high precision. A notable example is the AI-based clinical decision support tool, Fabric GEM (Fabric Genomics, Inc.), which enhances the diagnosis of rare genetic diseases by effectively prioritizing genetic variants and integrating deep phenotyping through clinical natural language processing. This tool substantially improves the diagnostic accuracy and efficiency for rare genetic disorders, as demonstrated in a retrospective cohort study [76]. Similarly, another study reviews AI and ML approaches using gene expression and variant data to advance personalized medicine, highlighting the potential of these technologies to enhance disease diagnosis, risk assessment, and treatment outcomes across diverse populations and a range of diseases, including cancer, cardiovascular, and neurodegenerative diseases [77]. Furthermore, a separate research effort using supervised ML showcases how models such as support vector machines and K-nearest neighbors can be applied to multifactorial genetic inheritance disorders to predict conditions such as cancer, dementia, and diabetes with high accuracy. Support vector machine is a classification technique that finds the hyperplane that maximizes the margin between different classes in the data, while K-nearest neighbors classifies data based on the majority class among the k-nearest samples, providing a straightforward but powerful approach for predictive modeling [78]. Despite these advancements, the deployment of AI in diagnostic devices comes with challenges such as ensuring data diversity to prevent bias, enhancing the interpretability of AI decisions for clinical use, and navigating complex regulatory and ethical considerations [79,80].

Therapeutic Devices: Dynamic Treatment Adaptations and Real-Time Monitoring

AI-enabled therapeutic devices are revolutionizing patient care by adapting treatments in real-time, based on continuous monitoring and dynamic feedback. This capability leads to more effective and personalized treatment plans, enhancing overall patient outcomes [30]. By leveraging real-time data analysis, these devices can respond immediately to changes in a patient's condition, ensuring that treatments are both timely and precisely tailored to individual needs.

One key application of AI in therapeutic devices is in the management of diabetes through AI-driven real-time insulin pumps, which use continuous monitoring and feedback to deliver personalized and effective treatment plans. This approach is underscored by recent research that introduces a closed-loop electronic system using a proportional-integral-derivative controller and WiFi-controlled voltage to automatically adjust insulin levels based on glucose sensor readings, enhancing diabetes management [81]. These types of dynamic management help maintain optimal blood sugar levels, significantly reducing the risk of complications associated with diabetes. Research developed for an AI expert system based on CGM data that not only automates insulin delivery but also facilitates forward-thinking and collaborative health management between sensors and patients [82].

Another significant advancement is in cardiac care, where AI-powered devices such as pacemakers and defibrillators adjust their functioning based on the patient's activity level and physiological signals [83]. This ensures that the heart operates as efficiently as possible, providing personalized care that adapts to the daily life and needs of the patient, thereby enhancing treatment efficacy and patient safety. Complementing these innovations, a recent study introduces a neural network capable of identifying the manufacturer and model group of cardiac rhythm devices from chest radiographs with exceptional accuracy. This tool, which achieves 99.6% accuracy for manufacturer identification and 96.4% for model groups, outperforms traditional methods used by cardiologists, offering a quicker and more precise approach to managing patients with implanted cardiac devices [84]. Furthermore, another pivotal study uses CNNs to predict the risk of pacemaker-induced cardiomyopathy using clinical data available prior to pacemaker implantation. Achieving an accuracy rate of 75.8%, this model highlights the potential for AI to proactively manage patients' health, potentially preventing the onset of cardiomyopathy post-implantation [85].

Expanding on this, AI technologies are being increasingly integrated into various cardiac monitoring tools, leveraging advanced algorithms and wireless connectivity to deliver expert-level diagnostics and continuous heart health monitoring. For example, continuous electrocardiogram recordings help redefine arrhythmia phenotypes and predict the success of antiarrhythmic therapies, while wearable activity trackers measure vital physiological indices that can detect irregularities such as atrial fibrillation. Additionally, smartphone apps contribute to reducing response times for sudden cardiac events, showcasing the integration of novel monitoring devices with ML for scalable cardiovascular management. These advancements in cardiovascular monitoring technologies not only promise to improve personalized care but also pose new challenges and opportunities for the future of cardiac health [86].

These AI advancements in cardiac care not only enhance diagnostic and predictive capabilities but also significantly improve the personalized management and safety of patients requiring cardiac rhythm management. Together, these applications of AI in therapeutic devices demonstrate a shift toward more adaptive, responsive medical treatments that

XSL•FO
RenderX

promise to improve patient outcomes by integrating sophisticated technology with real-time health data [87]. Supporting this, a study develops predictive models based on mobile clinical data for patients with heart failure, significantly reducing false alerts and improving the management of patient decompensation through telemonitoring systems [88]. These advances support the ongoing development and refinement of therapeutic devices that not only anticipate health needs but also actively engage with them, paving the way for highly personalized and proactive treatment strategies.

Wearable Technology: Continuous Health Tracking and Disease Management

AI-enabled wearable technologies are revolutionizing health management by providing continuous tracking and dynamic feedback, allowing for unprecedented personalization in medical care. Equipped with advanced AI algorithms, these devices analyze data from everyday activities to offer personalized health insights and recommendations, enhancing the effectiveness and timeliness of interventions [89]. This capability helps users manage their health more proactively and enables early detection of potential health issues. Recent advancements in wearable health technology have significantly improved the performance of wearable sensors, enabling real-time health monitoring and early disease detection. Additionally, for chronic conditions like hypertension and heart disease, wearable technologies equipped with AI can predict exacerbations, enabling preemptive medical interventions that tailor treatments to the nuances of an individual's physiological data [90]. This advanced approach not only optimizes patient care but also integrates seamlessly with the broader goals of precision medicine, underscoring the profound impact of AI in enhancing daily health management.

Wearable devices like fitness trackers and smartwatches have become essential tools in everyday health management, capturing a wide array of data such as heart rate [91], sleep patterns [92], and physical activity levels [93]. These devices leverage sophisticated AI algorithms to analyze this wealth of information, providing personalized insights and actionable recommendations. This functionality not only supports users in maintaining or improving their lifestyle but also plays a crucial role in the early detection of potential health issues [10]. Notably, changes in heart rate variability and sleep quality, as measured by these wearables, may indicate stress or cardiovascular issues. A study found that heart rate variability, tracked through wearable devices, effectively predicts the effectiveness of therapeutic interventions for anxiety and depression, underscoring its value in preemptive health measures [94]. This proactive approach to health monitoring highlights the significant role that wearable technology plays in preventive health care, empowering users to manage their well-being with precision and foresight.

Figure 2 illustrates various innovative CGM biosensors designed for CGM using different platforms such as interstitial fluid, tears, saliva, and sweat. These wearable devices represent the cutting edge of personalized health monitoring technologies, integrating advanced sensor systems and AI to provide real-time health data, crucial for managing chronic conditions like diabetes [95].

XSL•FO
RenderX

Figure 2. Representative examples of wearable continuous glucose monitoring biosensors [95]. (**A**) An electrochemical nonenzymatic continuous glucose monitoring system. Reproduced from Yoon et al [96] (**B**) A fluorescence-based continuous glucose monitoring biosensor. Reproduced from Sawayama and Takeuchi [97]. (**C**) Commercialized Senseonics continuous glucose monitoring sensor. (**D**) A contact lens biosensor for real-time glucose monitoring. Reproduced from Moreddu et al [98] (**E**) A wearable mouthguard continuous glucose monitoring biosensor. Reproduced from Arakawa et al [99] (**F**) A smart wristband for monitoring glucose in sweat. Reproduced from Gao et al [100]. Full figure reproduced from Jin et al. [95], licensed under CC BY 4.0. CGM: continuous glucose monitoring; GF-PEG: glucose-responsive fluorescent polyethylene glycol; ISF: interstitial fluid.

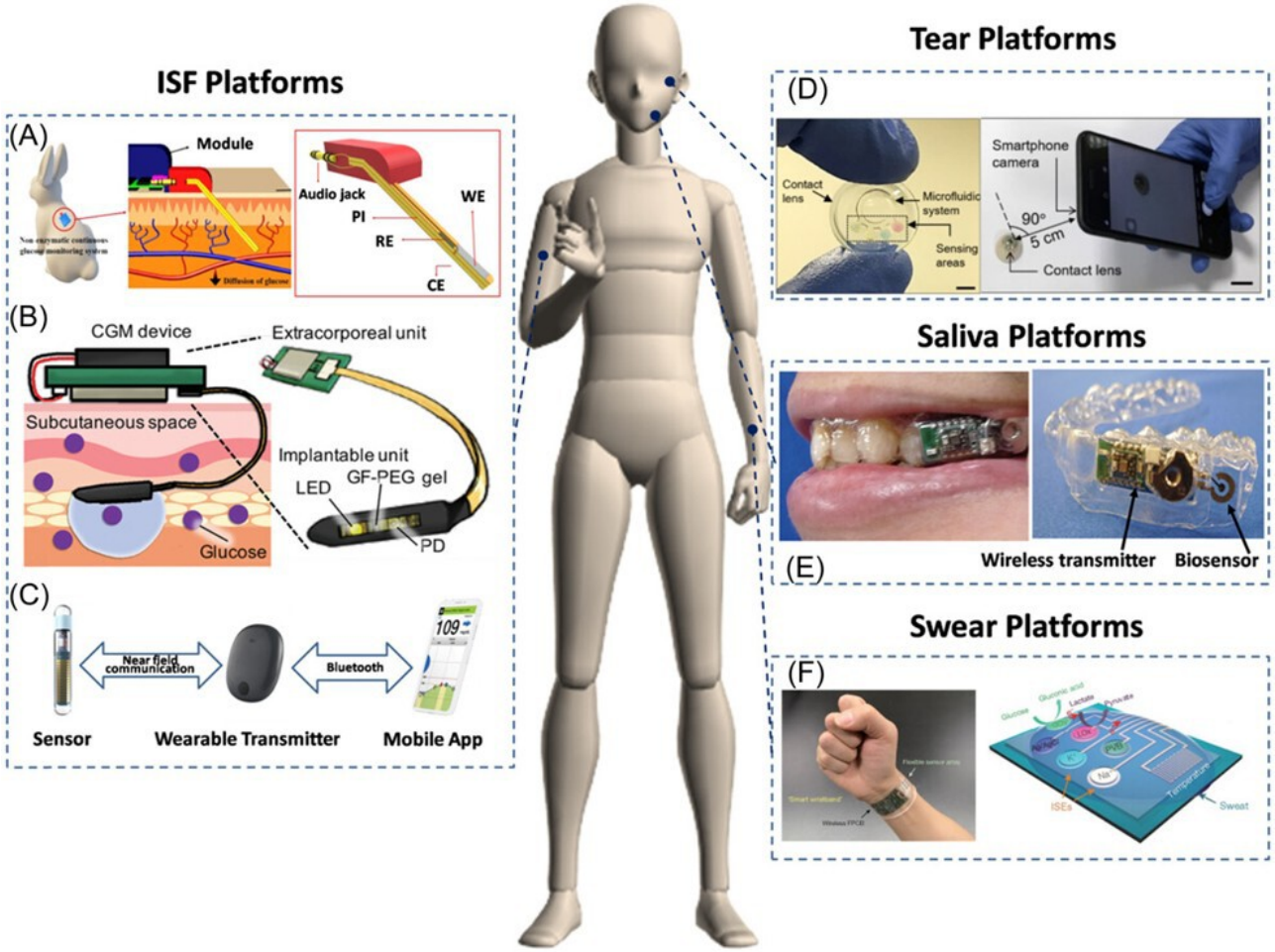

Personalized eHealth programs have demonstrated significant lifestyle improvements among patients with cardiac issues through tailored interventions based on data captured by wearable devices. For instance, the "Do Cardiac Health Advanced New Generation Ecosystem (Do CHANGE 2)" program effectively demonstrates the impact of personalized eHealth interventions, which use wearable technology to promote lifestyle changes and improve the quality of life among patients with cardiac issues [101]. This intervention, documented to significantly enhance lifestyle behaviors through continuous health behavior assessments and personalized feedback, underscores the potential of wearable devices in facilitating substantial health improvements and personalized care. Furthermore, the AI4FoodDB project has innovatively used wearable devices to gather and analyze dietary and lifestyle data, offering groundbreaking personalization in nutritional advice, which significantly impacts lifestyle improvements [102]. Similarly, personalized activity eCoaching, enhanced by AI and semantic ontology, exemplifies how technological advancements can foster more accurate health monitoring and lifestyle recommendations, effectively encouraging healthier behavior patterns [103]. Additionally, real-time models to predict stress-induced blood pressure spikes using wearable devices underline the potential of AI-driven technologies to proactively manage and mitigate health risks [104]. These examples collectively underscore the profound influence of AI-enhanced wearable technology in shaping future health care paradigms, focusing on personalized care and proactive health management strategies.

In the realm of chronic disease management, wearable technologies equipped with AI are particularly transformative. A systematic review of studies involving patients with various chronic conditions demonstrated that wearable devices could facilitate active self-management and real-time health monitoring, thereby playing a crucial role in managing chronic diseases. This data can then be leveraged through AI algorithms to provide actionable insights, which may aid in the early detection and management of potential complications associated with chronic diseases [105]. For instance, wearables have been shown to impact health positively by aiding in the management of diseases like diabetes, where they can monitor and provide feedback on blood sugar levels, or in cardiovascular diseases, where they assist in monitoring heart function [106]. Devices designed for monitoring hypertension and heart disease continuously track vital signs like blood pressure and heart

XSL•FO
RenderX

rhythm, transmitting this data in real time to health care providers [107]. AI plays a critical role in this process by analyzing trends and variations in the data to predict potential exacerbations. This proactive approach allows health care providers to offer preemptive medical interventions, potentially reducing hospital visits and improving patient outcomes [108].

This evolution in wearable technology emphasizes a shift from reactive to proactive and personalized medical care. The continuous monitoring facilitated by these devices, combined with AI-driven analysis, ensures that each patient receives care that is tailored not just to a condition but to their unique physiological makeup and lifestyle. This not only enhances patient engagement by providing them with real-time, actionable health insights but also significantly improves the efficacy of interventions, aligning perfectly with the goals of precision medicine [109]. As these technologies advance, they promise to further refine our approach to health management, making preventative care more accessible and effective than ever before.

Prosthetics and Implants: Customization and Adaptive Functionalities

The advent of AI-driven prosthetics and implants marks a significant leap forward in the field of medical devices, enhancing the integration of technology and human physiology. These advanced devices harness AI to provide functionalities that closely emulate natural human movements, offering a more intuitive and seamless experience for users [110]. This technological evolution not only improves the quality of life for individuals but also pushes the boundaries of what medical devices can achieve in terms of personalization and performance.

In the realm of smart prosthetics, AI plays a critical role in enabling these devices to learn from and adapt to the user's specific movement patterns. This capability allows for fluid and natural motions, such as a prosthetic arm that adjusts its grip based on the object it is interacting with. This is achieved through advanced sensors and actuators that continually collect and analyze data, allowing the prosthetic to make real-time adjustments. For example, shared control systems in assistive robots predict user intent and react accordingly, enhancing the functionality of the prosthetic while maintaining a balance between user control and automated responses [111]. Similarly, demonstration units for upper-limb prostheses provide valuable insights during the calibration and learning phases, further personalizing the user experience [112]. A recent study developed an automated pipeline using ML methods to design custom total knee replacement implants from CT scans. This pipeline demonstrated high accuracy and repeatability, significantly reducing the time and cost associated with traditional manual processes, highlighting the potential for broader applications in custom medical implants (Figure 3, which illustrates the workflow of the CT-3D surface model prediction process for custom total knee replacement design) [113].

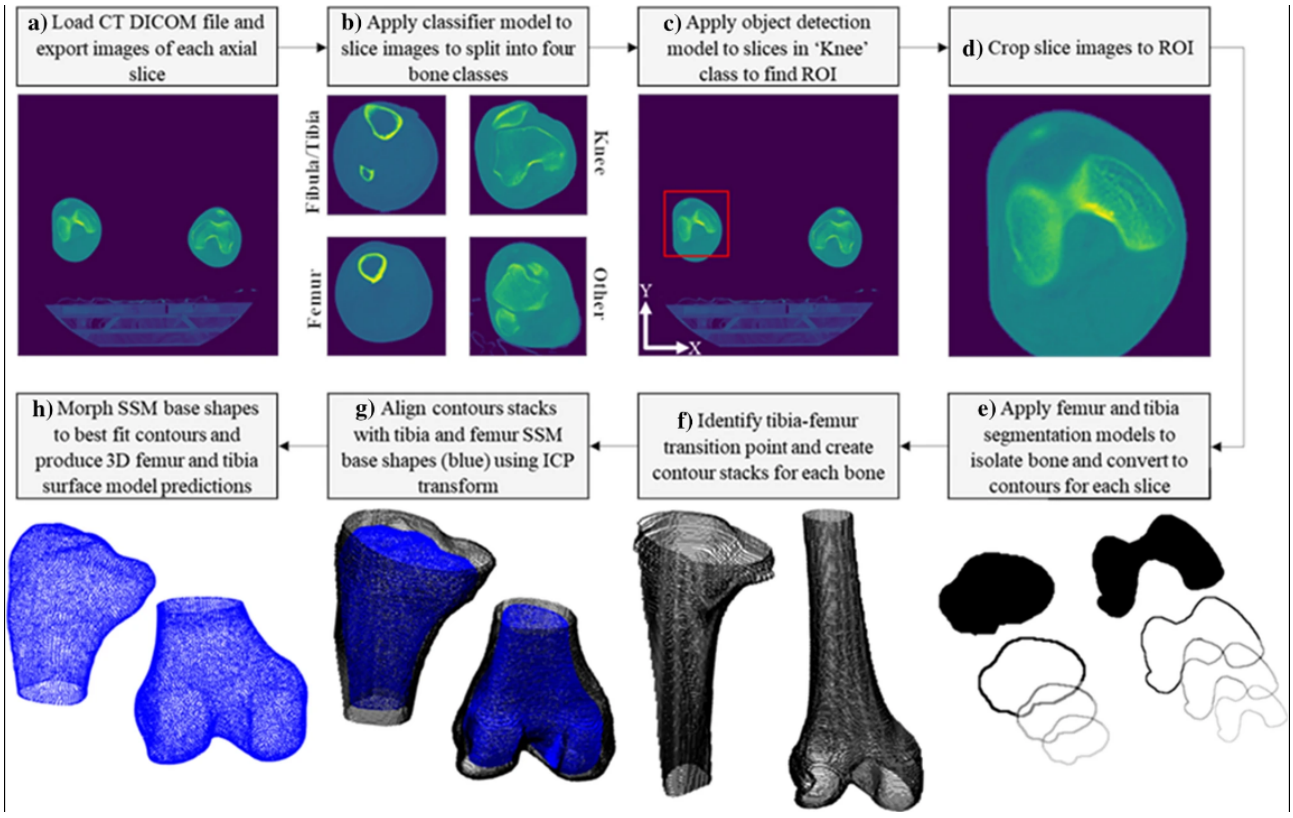

Figure 3. Workflow of computed tomography-3D surface model prediction process for custom total knee replacement design: (**A**) Load digital imaging and communications in medicine, (**B**) classify slices, (**C**) identify region of interest, (**D**) crop slices to region of interest, (**E**) segment femur and tibia bones, (**F**) create contour stacks, (**G**) align contour stacks with statistical shape model base shapes, (**H**) morph statistical shape models to fit contour points and create 3D model predictions [113]. Full figure reproduced from Burge et al. [95,113], licensed under CC BY 4.0. CT: computed tomography; DICOM: digital imaging and communications in medicine; ICP: iterative closest point; ROI: region of interest; SSM: statistical shape model.

XSL•FO
RenderX

The scope of AI's impact extends beyond prosthetics to include adaptive implants, which are revolutionizing areas such as orthopedics. Adaptive implants also benefit from AI integration, particularly in orthopedics, where implants can adjust over time to changes in the body or the mechanical loads they must endure. This adaptability can significantly enhance the implant's longevity and the comfort it offers, making it a vital feature in modern medical practice [114]. For instance, recent advancements in knee arthroplasty involve using ML techniques to automate the customization of stiffness-matched knee implants, ensuring they closely mimic the mechanical properties of the patient's own bones, thereby improving post-surgical outcomes and promoting natural bone remodeling [115]. Furthermore, new designs in transfemoral prosthetic sockets incorporate a combination of rigid frames and silicone structures that not only offer better biomechanical coupling with the residual limb but also integrate smart technologies for health monitoring and feedback [116]. These AI-enhanced implants can respond to the body's changes over time, adjusting to shifts in load and movement. In spinal and joint replacement surgeries, AI assisted in the design of implants that align with the patient's unique biomechanical structure, potentially enhancing the longevity and comfort of these devices [117]. This holistic approach to prosthetics and implants, driven by AI, represents a significant step forward in personalized medical devices, offering enhanced functionality and improved patient comfort through adaptive capabilities.

The potential of AI in prosthetics and implants extends into holistic health management, where these technologies facilitate continuous monitoring and condition management, contributing to preventive health care. Innovations such as electromyography and gyroscope integration in sports training assist systems exemplify how AI-driven devices can enhance physical activities and training effectiveness [118]. Each advancement not only exemplifies the integration of AI in improving device functionality but also highlights the shift toward more personalized, responsive medical care, aligning with the broader goals of enhancing patient outcomes and optimizing therapeutic efficacy.

Challenges in AI-Driven Personalization

Overview

The integration of AI into medical devices heralds significant advancements in personalized health care. However, this integration is not without its challenges, which span technical, ethical, social, and regulatory domains. Addressing these challenges is crucial for the successful implementation and acceptance of AI-driven personalization in medical devices.

Technical Challenges

In the realm of AI-driven personalization in health care, technical challenges significantly impact the development and implementation of these innovative technologies. One of the foundational hurdles is the issue of data quality and availability. Effective AI models rely on high-quality, comprehensive datasets, yet health care often struggles with accessing such data [119]. Patient information is frequently dispersed across various institutions and platforms, creating considerable gaps in data necessary for AI training. Additionally, privacy concerns and stringent regulatory frameworks can restrict access to vital data, compounding the difficulty in ensuring data quality, specifically in terms of accuracy, completeness, and relevance, due to varied data collection methodologies and the potential for human error [120,121]. For instance, the integration of wearables in clinical settings necessitates additional validation to confirm their accuracy, reliability, and effectiveness in enhancing outcomes for chronic conditions. Issues such as data privacy, standardization of device functionalities, and integration into existing health care infrastructures also need to be addressed to leverage their potential in chronic disease management [106].

Another significant technical challenge is the development of reliable and valid AI algorithms that perform consistently well across diverse patient populations and clinical environments [122]. Algorithmic reliability is crucial for clinician acceptance and patient safety. Validation of these algorithms demands extensive testing to ensure they deliver accurate predictions and recommendations under various circumstances [123]. However, the opaque nature of some AI models complicates this validation process, leading to concerns about their interpretability and the feasibility of auditing their decision-making processes [124].

Moreover, the successful deployment of AI-driven devices hinges on their seamless integration with existing health care systems and workflows. This integration presents substantial technical challenges, as health care infrastructures are often intricate and composed of outdated software and hardware that may not readily accommodate new AI functionalities [32]. Achieving interoperability between AI systems and electronic health records is essential for real-time data exchange and analysis, yet often faces obstacles due to compatibility issues [125]. These technical challenges must be navigated carefully to realize the full potential of AI in personalizing medical treatment and enhancing patient care.

Ethical and Social Challenges

The integration of AI into medical devices, while promising, brings with it a myriad of ethical and social challenges that require diligent attention and strategic management. Privacy and data security concerns are at the forefront of these challenges [126]. The extensive collection, storage, and analysis of sensitive patient data inherent in AI applications necessitate rigorous cybersecurity measures to protect against unauthorized access and breaches [127]. Ensuring the confidentiality, integrity, and availability of this data is not only a technical requirement but also a moral imperative, demanding strict adherence to evolving privacy regulations and standards [120].

Another significant ethical concern is addressing algorithmic bias and fairness. AI systems, inherently dependent on the data with which they are trained, can perpetuate existing biases if the data reflect historical inequities in health care [128]. Such biases could result in AI-driven medical devices performing differently across diverse patient groups, potentially exacerbating health disparities rather than alleviating them [129]. To counteract this, there needs to be a proactive effort in the selection and scrutiny of training datasets and the design of AI models, ensuring they are both inclusive and equitable.

XSL•FO
RenderX

Furthermore, the deployment of AI in health care raises complex issues around patient consent and autonomy. It is essential that patients are thoroughly informed about how their data will be used, the involvement of AI in their treatment, and the implications of decisions made by AI systems [130]. This is particularly challenging given the complexity of AI technologies and the potential for misunderstandings among patients lacking technical expertise. Moreover, care must be taken to ensure that AI-enhanced medical devices support, rather than undermine, patient autonomy, allowing individuals to remain central in making informed decisions about their health care [131].

These ethical and social challenges underscore the need for a balanced approach to the deployment of AI in medical devices, where innovation is matched by an unwavering commitment to ethical standards, patient rights, and social justice. As AI technology continues to evolve, so must the strategies for addressing the ethical implications it brings to the health care landscape.

Regulatory and Policy Challenges

Navigating the complex regulatory landscape for AI-enabled medical devices presents a significant challenge as authorities worldwide strive to balance innovation with patient safety. Regulatory bodies are actively developing frameworks that ensure the safety and efficacy of these technologies [12]. This entails rigorous processes for AI development, validation, and ongoing post-market monitoring [132]. The task is particularly daunting due to the rapid technological advances in AI, which require regulators to continually adapt and update guidelines to keep pace with innovation [133].

In addition to regulatory hurdles, the standardization and certification of AI-driven medical devices are critical yet complex issues. Establishing universal protocols for development, testing, and certification is essential to ensure the reliability and interoperability of these devices [134]. However, achieving consensus on these standards involves intricate negotiations among diverse stakeholders from the health care, technology, and regulatory sectors. These standards must not only be robust enough to ensure safety and efficacy but also flexible enough to adapt to future technological innovations [135].

Moreover, the international dimension of AI-driven personalization in medical devices adds another layer of complexity. Global collaboration is necessary to tackle issues like data sharing, privacy protection, and the alignment of regulatory standards [136]. Harmonizing these standards across different jurisdictions can facilitate the widespread adoption of AI technologies in health care [137]. Yet, this requires a concerted effort to reconcile varying regulatory philosophies, health care systems, and cultural attitudes toward privacy and autonomy.

Addressing these regulatory and policy challenges necessitates a multidisciplinary approach that involves experts from health care, technology, ethics, and law. By fostering open collaboration and dialogue among these stakeholders, the sector can effectively navigate the intricacies of AI-driven personalization in medical devices. This collaborative approach is vital for unleashing the transformative potential of AI in health care, enhancing global health outcomes while ensuring compliance with rigorous safety standards.

Discussion

Principal Findings

The comprehensive evaluation of AI-integrated medical devices highlights a remarkable evolution in health care technology, spanning multiple domains including diagnostics, therapeutics, prosthetics, and robotics. The results demonstrate how AI-powered solutions are reshaping health care delivery, enabling real-time monitoring, data-driven interventions, and enhanced patient outcomes. In diabetes management, smart insulin pumps exemplify the shift toward precision medicine. These devices leverage real-time data from CGMs and ML algorithms to autonomously modulate insulin delivery, addressing long-standing challenges such as glycemic variability and poor treatment adherence. Complementing this, AI-enabled insulin pens that achieve over 95% accuracy in detecting injection sites and lipodystrophies illustrate how intelligent devices can mitigate complications and support better patient compliance [35]. Further advancements, such as predictive self-care devices, highlight how AI facilitates proactive diabetes management, reducing the risk of hypoglycemia and hyperglycemia [138].

AI-powered prosthetics demonstrate a significant leap in rehabilitation and mobility aids. These systems, particularly those incorporating surface electromyographic-based systems and deep learning frameworks, adaptively interpret complex muscle signals, enhancing patient-robot interaction and optimizing motor function restoration. Innovations in gait analysis and lower limb assistive technologies further contribute to fall prevention and stability in dynamic environments [39]. Beyond mobility, AI-powered tools also enhance clinical workflow, with AI-driven radiograph analysis improving the speed and accuracy of pre- and post-operative orthopedic evaluations.

The diagnostic sector has witnessed transformative developments, particularly with deep learning models outperforming traditional diagnostic approaches. In medical imaging, AI-enabled systems detect cancerous lesions and other pathologies with higher sensitivity and specificity compared to manual interpretation. AI-enhanced frameworks have been pivotal in diagnosing pancreatic and lung cancers, improving early detection rates and clinical decision-making. The successful application of deep learning in chest X-ray analysis, achieving accuracy rates nearing 99%, reflects the capacity of AI to revolutionize pulmonary diagnostics. Moreover, the incorporation of AI into ophthalmology and cardiology diagnostics has led to non-invasive, precise evaluations of diabetic retinopathy and cardiovascular conditions.

Comparison With Prior Work

In comparison to prior literature, our findings are consistent with earlier studies that highlight the critical role of AI in personalizing medical interventions. For example, the diagnostic accuracy of AI models surpassing traditional radiological

XSL•FO
RenderX

assessments aligns with recent meta-analyses showing AI's superior sensitivity and specificity in medical imaging [43]. Additionally, the effectiveness of AI-powered wearables and biosensors in cardiovascular and metabolic monitoring complements existing literature on personalized chronic disease management [101].

The reviewed case studies build upon previous research by providing concrete evidence of AI's integration into rehabilitation devices, surgical planning tools, and patient communication platforms, such as AI-powered chatbots. Furthermore, AI-driven technologies are making significant contributions to medical manufacturing, education, and patient communication. AI-enhanced 3D printing supports the creation of personalized implants and wearable devices, while AI-powered chatbots revolutionize patient consultation by delivering timely, accurate medical advice.

Additionally, our review underscores how AI customization in prosthetics and implants contributes to reducing production times and improving biomechanical outcomes, which reflects a growing trend observed in related engineering and biomedical studies. The reviewed studies further elaborate on AI's role in adaptive implants and intelligent prosthetic sockets equipped with biosensors, offering continuous feedback, improving comfort, and enhancing functional outcomes for users [114].

Strengths and Limitations

One of the strengths of this review is the comprehensive coverage of AI applications across a broad spectrum of medical devices, ranging from diagnostic imaging to rehabilitation tools and patient engagement platforms. The systematic methodology ensured the inclusion of peer-reviewed studies from diverse health care domains, enhancing the robustness and generalizability of our findings. Additionally, by using a thematic categorization, we provided clarity in synthesizing complex data. The use of a standardized data extraction protocol based on PRISMA-ScR guidelines ensured methodological rigor and minimized bias. Further, the incorporation of snowballing techniques allowed us to capture relevant studies that may have been missed during the initial database searches, contributing to the depth and breadth of our review.

The review has a number of limitations. The restriction to English-language publications may have excluded valuable studies in other languages, potentially introducing language bias. To address this, we conducted cross-verification using key non-English papers referenced in systematic reviews and meta-analyses. Additionally, the study did not fully explore how specific patient demographics, such as age, gender, socioeconomic status, and comorbidities, may influence responses to AI-driven medical devices, which could affect generalizability and clinical relevance. Another limitation lies in the rapidly evolving nature of AI technology, which may render some findings subject to change as new developments emerge. Additionally, while our review emphasizes AI's potential in health care, real-world implementation challenges such as regulatory compliance, data privacy concerns, and the need for interdisciplinary collaboration remain critical issues for future research.

Future Directions

Future research should continue to explore the integration of emerging technologies such as quantum computing, blockchain, and AI-driven personalization into health care. These technologies offer opportunities to enhance the precision, security, and scalability of medical devices and health care delivery [139,140].

Quantum computing holds significant promise for advancing the speed and complexity of data processing, which could improve the performance of AI algorithms used in diagnostics, disease monitoring, and predictive analytics [141]. This computational capacity may enable the development of adaptive medical devices that respond dynamically to patients' evolving health needs [142].

Blockchain technology has the potential to improve data security and integrity by enabling decentralized, tamper-proof health information systems [143]. This could support the development of more trustworthy AI models, particularly in sensitive domains such as patient data management and cross-institutional interoperability [131].

In addition, future research should investigate the role of AI-driven personalized medical devices in addressing global health disparities. Portable AI-enabled diagnostic tools and wearable technologies may help expand health care access in underserved and remote regions [144]. Such innovations could also inform targeted public health strategies, contributing to a more equitable global health landscape. As these technologies become more integrated into health care ecosystems, continued interdisciplinary collaboration will be essential to navigate ethical, regulatory, and technical challenges.

Conclusions

The integration of AI into medical devices is fundamentally reshaping modern health care by enabling the shift toward personalized and precision medicine. AI-powered innovations, from smart insulin pumps to diagnostic imaging tools and wearable biosensors, are transforming how health care providers diagnose, treat, and monitor patients. These technologies harness advanced ML algorithms and real-time data to tailor medical interventions to individual needs, enhancing treatment efficacy and improving patient outcomes. The dynamic capabilities of AI in automating tasks, recognizing patterns in complex datasets, and adapting to physiological feedback exemplify its critical role in evolving health care delivery toward a more patient-centric model.

The technological landscape continues to expand with advancements in deep learning, natural language processing, and edge computing, which have accelerated the development of AI-enhanced therapeutic and assistive devices. Wearable technologies now enable continuous health monitoring, empowering patients to actively participate in managing chronic conditions such as diabetes and cardiovascular diseases. Meanwhile, AI-driven prosthetics and adaptive implants offer improved functionality and comfort by personalizing device behavior to the user's biomechanical and physiological characteristics. This fusion of AI with medical devices not only augments clinical workflows but also enhances the quality of

XSL•FO
RenderX

life for patients through intuitive, real-time, and highly responsive health care solutions.

Despite these strides, realizing the full potential of AI in personalizing medical devices requires addressing key challenges related to data privacy, algorithmic bias, system interoperability, and regulatory compliance. Ensuring AI's ethical deployment, maintaining data integrity, and aligning with global safety standards are essential to prevent disparities in health care delivery. Future innovations, including the integration of quantum computing to optimize AI algorithms and blockchain for secure data management, hold promises for overcoming these barriers. As the health care ecosystem embraces interdisciplinary collaboration across technology, ethics, and policy, AI-driven medical devices will increasingly define the future of global health—advancing precision medicine to be more accessible, equitable, and effective for all populations.

Conflicts of Interest

None declared.

Checklist 1
PRISMA checklist.
[DOCX File, 32 KB - jmir_v28i1e72410_app1.docx]

References

1. Johnson KB, Wei WQ, Weeraratne D, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci 2021 Jan;14(1):86-93. [doi: 10.1111/cts.12884] [Medline: 32961010]
2. Li YH, Li YL, Wei MY, Li GY. Innovation and challenges of artificial intelligence technology in personalized healthcare. Sci Rep 2024;14(1):18994-19728. [doi: 10.1038/s41598-024-70073-7]
3. Mohanadas HP, Nair V, Doctor AA, et al. A systematic analysis of additive manufacturing techniques in the bioengineering of in vitro cardiovascular models. Ann Biomed Eng 2023 Nov;51(11):2365-2383. [doi: 10.1007/s10439-023-03322-x] [Medline: 37466879]
4. Korteling JEH, van de Boer-Visschedijk GC, Blankendaal RAM, Boonekamp RC, Eikelboom AR. Human- versus artificial intelligence. Front Artif Intell 2021;4:622364. [doi: 10.3389/frai.2021.622364] [Medline: 33981990]
5. Khatiwada P, Yang B, Lin JC, Blobel B. Patient-generated health data (PGHD): understanding, requirements, challenges, and existing techniques for data security and privacy. J Pers Med 2024 Mar 3;14(3):282. [doi: 10.3390/jpm14030282] [Medline: 38541024]
6. Ahmed Z, Mohamed K, Zeeshan S, Dong X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database (Oxford) 2020 Jan 1;2020:baaa010. [doi: 10.1093/database/baaa010] [Medline: 32185396]
7. Manickam P, Mariappan SA, Murugesan SM, et al. Artificial intelligence (AI) and internet of medical things (IoMT) assisted biomedical systems for intelligent healthcare. Biosensors (Basel) 2022 Jul 25;12(8):562. [doi: 10.3390/bios12080562] [Medline: 35892459]
8. Iqbal MJ, Javed Z, Sadia H, et al. Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future. Cancer Cell Int 2021 May 21;21(1):270. [doi: 10.1186/s12935-021-01981-1] [Medline: 34020642]
9. Malcangi M. AI-based methods and technologies to develop wearable devices for prosthetics and predictions of degenerative diseases. Methods Mol Biol 2021;2190:337-354. [doi: 10.1007/978-1-0716-0826-5_17] [Medline: 32804375]
10. Nahavandi D, Alizadehsani R, Khosravi A, Acharya UR. Application of artificial intelligence in wearable devices: opportunities and challenges. Comput Methods Programs Biomed 2022 Jan;213:106541. [doi: 10.1016/j.cmpb.2021.106541] [Medline: 34837860]
11. Pesapane F, Volonté C, Codari M, Sardanelli F. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. Insights Imaging 2018 Oct;9(5):745-753. [doi: 10.1007/s13244-018-0645-y] [Medline: 30112675]
12. Gilbert S, Anderson S, Daumer M, Li P, Melvin T, Williams R. Learning from experience and finding the right balance in the governance of artificial intelligence and digital health technologies. J Med Internet Res 2023 Apr 14;25(1):e43682. [doi: 10.2196/43682] [Medline: 37058329]
13. Deepa R, Arunkumar S, Jayaraj V, Sivasamy A. Healthcare's new Frontier: AI-driven early cancer detection for improved well-being. AIP Adv 2023 Nov 1;13(11):11. [doi: 10.1063/5.0177640]
14. Kulkarni PG, Paudel N, Magar S, et al. Overcoming challenges and innovations in orthopedic prosthesis design: an interdisciplinary perspective. Biomed Mater Devices 2023 May 12;2(1):1-12. [doi: 10.1007/s44174-023-00087-8] [Medline: 37363137]
15. Junaid SB, Imam AA, Abdulkarim M, et al. Recent advances in artificial intelligence and wearable sensors in healthcare delivery. Appl Sci (Basel) 2022;12(20):10271. [doi: 10.3390/app122010271]

XSL•FO
RenderX

16. Albahri AS, Duhaim AM, Fadhel MA, et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion. Information Fusion 2023 Aug;96:156-191. [doi: 10.1016/j.inffus.2023.03.008]
17. Nasir S, Khan RA, Bai S. Ethical framework for harnessing the power of AI in healthcare and beyond. IEEE Access 2024;12:31014-31035. [doi: 10.1109/ACCESS.2024.3369912]
18. Kasula BY. Framework development for artificial intelligence integration in healthcare: optimizing patient care and operational efficiency. Transactions on Latest Trends in IoT 2023;6(6):77-83 [FREE Full text]
19. Nakamura T, Sasano T. Artificial intelligence and cardiology: current status and perspective. J Cardiol 2022 Mar;79(3):326-333. [doi: 10.1016/j.jjcc.2021.11.017] [Medline: 34895982]
20. Schurink CAM, Lucas PJF, Hoepelman IM, Bonten MJM. Computer-assisted decision support for the diagnosis and treatment of infectious diseases in intensive care units. Lancet Infect Dis 2005 May;5(5):305-312. [doi: 10.1016/S1473-3099(05)70115-8] [Medline: 15854886]
21. George AH, George AS. From pulse to prescription: exploring the rise of AI in medicine and its implications. Partners Universal International Innovation Journal 2023;1(6):38-54 [FREE Full text]
22. Guni A, Normahani P, Davies A, Jaffer U. Harnessing machine learning to personalize web-based health care content. J Med Internet Res 2021 Oct 19;23(10):e25497. [doi: 10.2196/25497] [Medline: 34665146]
23. Li M, Jiang Y, Zhang Y, Zhu H. Medical image analysis using deep learning algorithms. Front Public Health 2023;11:1273253. [doi: 10.3389/fpubh.2023.1273253] [Medline: 38026291]
24. Al-qaness MAA, Zhu J, AL-Alimi D, et al. Chest X-ray images for lung disease detection using deep learning techniques: a comprehensive survey. Arch Computat Methods Eng 2024 Aug;31(6):3267-3301. [doi: 10.1007/s11831-024-10081-y]
25. Chen C, Zheng S, Guo L, et al. Identification of misdiagnosis by deep neural networks on a histopathologic review of breast cancer lymph node metastases. Sci Rep 2022;12(1):13482. [doi: 10.1038/s41598-022-17606-0]
26. Khurshid S, Friedman S, Reeder C, et al. ECG-based deep learning and clinical risk factors to predict atrial fibrillation. Circulation 2022 Jan 11;145(2):122-133. [doi: 10.1161/CIRCULATIONAHA.121.057480] [Medline: 34743566]
27. Abdulrahman Mohammed Al M, Zakiuddin A, Abdullah M. A. Paradigm shift in healthcare through technology and patient-centeredness. Int Arch Public Health Community Med 2018;2(1):1710015. [doi: 10.23937/iaphcm-2017/1710015]
28. Dhawan AP. Collaborative paradigm of preventive, personalized, and precision medicine with point-of-care technologies. IEEE J Transl Eng Health Med 2016;4:2800908. [doi: 10.1109/JTEHM.2016.2635126] [Medline: 28560119]
29. Gameiro GR, Sinkunas V, Liguori GR, Auler-Júnior JOC. Precision medicine: changing the way we think about healthcare. Clinics (Sao Paulo) 2018 Dec 3;73:e723. [doi: 10.6061/clinics/2017/e723] [Medline: 30517307]
30. Haick H, Tang N. Artificial intelligence in medical sensors for clinical decisions. ACS Nano 2021 Mar 23;15(3):3557-3567. [doi: 10.1021/acsnano.1c00085] [Medline: 33620208]
31. Engineering innovations in healthcare: a review of developments in the USA. Eng Sci Technol J 2023;4(6):381-400. [doi: 10.51594/estj.v4i6.638]
32. Esmaeilzadeh P. Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: a perspective for healthcare organizations. Artif Intell Med 2024 May;151:102861. [doi: 10.1016/j.artmed.2024.102861] [Medline: 38555850]
33. Klonoff DC, Kerr D. Smart pens will improve insulin therapy. J Diabetes Sci Technol 2018 May;12(3):551-553. [doi: 10.1177/1932296818759845] [Medline: 29411641]
34. Vettoretti M, Cappon G, Facchinetti A, Sparacino G. Advanced diabetes management using artificial intelligence and continuous glucose monitoring sensors. Sensors (Basel) 2020 Jul 10;20(14):3870. [doi: 10.3390/s20143870] [Medline: 32664432]
35. Torre E, Francini L, Cordelli E, et al. Exploiting AI to make insulin pens smart: injection site recognition and lipodystrophy detection. 2022 Presented at: 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS); Jul 21-23, 2022; Shenzen, China p. 211-216. [doi: 10.1109/CBMS55023.2022.00044]
36. Almulla SK, Zaatar O, Ahmed HS, ul Misbah Khanum N, Jarndal A. Smart artificial-intelligence based self-care-device for diabetic patients. 2022 Presented at: 2022 Advances in science and engineering technology international conferences (ASET); Feb 21-24, 2022; Dubai, United Arab Emirates p. 1-5. [doi: 10.1109/ASET53988.2022.9734874]
37. Kesavadev J, Saboo B, Krishna MB, Krishnan G. Evolution of insulin delivery devices: from syringes, pens, and pumps to DIY artificial pancreas. Diabetes Ther 2020 Jun;11(6):1251-1269. [doi: 10.1007/s13300-020-00831-z] [Medline: 32410184]
38. Bozzetto L, Corrado A, Scidà G. Dietary treatment of type 1 diabetes: beyond carbohydrate counting to fight cardiovascular risk. Nutr Metab Cardiovasc Dis 2023 Feb;33(2):299-306. [doi: 10.1016/j.numecd.2022.12.012] [Medline: 36642609]
39. Guo K, Orban M, Lu J, Al-Quraishi MS, Yang H, Elsamanty M. Empowering hand rehabilitation with AI-powered gesture recognition: a study of an sEMG-based system. Bioengineering (Basel) 2023 May 6;10(5):557. [doi: 10.3390/bioengineering10050557] [Medline: 37237627]
40. Mehr JK, Akbari M, Faridi P, Xing H, Mushahwar VK, Tavakoli M. Artificial ‐ intelligence ‐ powered lower limb assistive devices: future of home care technologies. Adv Intell Syst 2023 Jun;5(6):2200361. [doi: 10.1002/aisy.202200361]

XSL•FO
RenderX

41. Pagano S, Müller K, Götz J, et al. The role and efficiency of an AI-powered software in the evaluation of lower limb radiographs before and after total knee arthroplasty. J Clin Med 2023 Aug 24;12(17):5498. [doi: 10.3390/jcm12175498] [Medline: 37685563]
42. Hameed BS, Krishnan UM. Artificial intelligence-driven diagnosis of pancreatic cancer. Cancers (Basel) 2022 Oct 31;14(21):5382. [doi: 10.3390/cancers14215382] [Medline: 36358800]
43. Xu M, Chen Z, Zheng J, Zhao Q, Yuan Z. Artificial intelligence-aided optical imaging for cancer theranostics. Semin Cancer Biol 2023 Sep;94:62-80. [doi: 10.1016/j.semcancer.2023.06.003] [Medline: 37302519]
44. Sanida MV, Sanida T, Sideris A, Dasygenis M. An advanced deep learning framework for multi-class diagnosis from chest X-ray images. J 2024;7(1):48-71. [doi: 10.3390/j7010003]
45. Papadopoulos GT, Antona M, Stephanidis C. Towards open and expandable cognitive AI architectures for large-scale multi-agent human-robot collaborative learning. IEEE Access 2021;9:73890-73909. [doi: 10.1109/ACCESS.2021.3080517]
46. Rossos D, Mihailidis A, Laschowski B. AI-powered smart glasses for sensing and recognition of human-robot walking environments. bioRxiv Bioengineering. Preprint posted online on Oct 29, 2023. [doi: 10.1101/2023.10.24.563804]
47. Robbins S. AI and the path to envelopment: knowledge as a first step towards the responsible regulation and use of AI-powered machines. AI & Soc 2020 Jun;35(2):391-400. [doi: 10.1007/s00146-019-00891-1]
48. Hlee S, Park J, Park H, Koo C, Chang Y. Understanding customer's meaningful engagement with AI-powered service robots. ITP 2023 Apr 11;36(3):1020-1047. [doi: 10.1108/ITP-10-2020-0740]
49. Moriuchi E. Leveraging the science to understand factors influencing the use of AI-powered avatar in healthcare services. J Technol Behav Sci 2022;7(4):588-602. [doi: 10.1007/s41347-022-00277-z]
50. Erol BA, Majumdar A, Benavidez P, Rad P, Choo KKR, Jamshidi M. Toward artificial emotional intelligence for cooperative social human–machine interaction. IEEE Trans Comput Soc Syst 2019;7(1):234-246. [doi: 10.1109/TCSS.2019.2922593]
51. Mladenovic R. AI-powered and "augmented" dentistry: applications, implications and limitations. In: Augmented Reality and Artificial Intelligence: The Fusion of Advanced Technologies: Springer Nature Switzerland; 2023:211-226 URL: https://doi.org/10.1007/978-3-031-27166-3_12 [accessed 2026-03-02]
52. Zhu Z, Ng DWH, Park HS, McAlpine MC. 3D-printed multifunctional materials enabled by artificial-intelligence-assisted fabrication technologies. Nat Rev Mater 2021;6(1):27-47. [doi: 10.1038/s41578-020-00235-2]
53. Schwarz GM, Simon S, Mitterer JA, et al. Artificial intelligence enables reliable and standardized measurements of implant alignment in long leg radiographs with total knee arthroplasties. Knee Surg Sports Traumatol Arthrosc 2022 Aug;30(8):2538-2547. [doi: 10.1007/s00167-022-07037-9] [Medline: 35819465]
54. Itelman E, Golovchiner G, Barsheshet A, Kornowski R, Erez A. Balancing innovation and professionalism: the emerging role of AI-powered chatbots in medical consultation. Heart Rhythm 2024 Oct;21(10):2037-2039. [doi: 10.1016/j.hrthm.2024.04.010] [Medline: 38588991]
55. Samanoodi R, Alrasheed A, Felemban A, et al. AI-powered mobility application for field inspection rounds. 2023 Oct 2 Presented at: ADIPEC; Oct 2-5, 2023. [doi: 10.2118/216996-MS]
56. Sudharson NA, Joseph M, Kurian N, Varghese KG, Wadhwa S, Thomas HA. AI-powered neural implants. Br Dent J 2023 Mar;234(6):359-360. [doi: 10.1038/s41415-023-5698-8] [Medline: 36964340]
57. Wu N, Phang J, Park J, et al. Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans Med Imaging 2020 Apr;39(4):1184-1194. [doi: 10.1109/TMI.2019.2945514] [Medline: 31603772]
58. Chae KJ, Jin GY, Ko SB, et al. Deep learning for the classification of small (≤2 cm) pulmonary nodules on CT imaging: a preliminary study. Acad Radiol 2020 Apr;27(4):e55-e63. [doi: 10.1016/j.acra.2019.05.018] [Medline: 31780395]
59. Mhamdi L, Dammak O, Cottin F, Dhaou IB. Artificial intelligence for cardiac diseases diagnosis and prediction using ECG images on embedded systems. Biomedicines 2022 Aug 19;10(8):8. [doi: 10.3390/biomedicines10082013] [Medline: 36009560]
60. Padmavilochanan D, Pathinarupothi RK, Menon KAU, et al. Personalized diabetes monitoring platform leveraging IoMT and AI for non-invasive estimation. Smart Health (2014) 2023 Dec;30:100428. [doi: 10.1016/j.smhl.2023.100428]
61. Lugogo NL, DePietro M, Reich M, et al. A predictive machine learning tool for asthma exacerbations: results from a 12-week, open-label study using an electronic multi-dose dry powder inhaler with integrated sensors. J Asthma Allergy 2022;15:1623-1637. [doi: 10.2147/JAA.S377631] [Medline: 36387836]
62. McRae MP, Rajsri KS, Alcorn TM, McDevitt JT. Smart diagnostics: combining artificial intelligence and in vitro diagnostics. Sensors (Basel) 2022 Aug 24;22(17):6355. [doi: 10.3390/s22176355] [Medline: 36080827]
63. Luchini C, Pea A, Scarpa A. Artificial intelligence in oncology: current applications and future perspectives. Br J Cancer 2022 Jan;126(1):4-9. [doi: 10.1038/s41416-021-01633-1] [Medline: 34837074]
64. Rahmani F. The application of machine learning in the detection of Covid-19 in blood tests. Int J New Chem 2021;8(5):103-126 [FREE Full text]
65. Krittanawong C, Johnson KW, Choi E, et al. Artificial intelligence and cardiovascular genetics. Life (Basel) 2022 Feb 14;12(2):279. [doi: 10.3390/life12020279] [Medline: 35207566]
66. Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med 2021 Apr 7;4(1):65. [doi: 10.1038/s41746-021-00438-z] [Medline: 33828217]

XSL•FO
RenderX

67. Artificial intelligence-enabled medical devices. FDA. 2024 Dec 20. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices [accessed 2026-02-05]
68. Parmar UPS, Surico PL, Singh RB, et al. Artificial intelligence (AI) for early diagnosis of retinal diseases. Medicina (Kaunas) 2024 Mar 23;60(4):527. [doi: 10.3390/medicina60040527] [Medline: 38674173]
69. Chudzik P, Majumdar S, Calivá F, Al-Diri B, Hunter A. Microaneurysm detection using fully convolutional neural networks. Comput Methods Programs Biomed 2018 May;158:185-192. [doi: 10.1016/j.cmpb.2018.02.016] [Medline: 29544784]
70. Wang H, Yuan G, Zhao X, et al. Hard exudate detection based on deep model learned information and multi-feature joint representation for diabetic retinopathy screening. Comput Methods Programs Biomed 2020 Jul;191:105398. [doi: 10.1016/j.cmpb.2020.105398] [Medline: 32092614]
71. Murugan R, Roy P. MicroNet: microaneurysm detection in retinal fundus images using convolutional neural network. Soft Comput 2022 Feb;26(3):1057-1066. [doi: 10.1007/s00500-022-06752-2]
72. Chiu HY, Chao HS, Chen YM. Application of artificial intelligence in lung cancer. Cancers (Basel) 2022 Mar 8;14(6):1370. [doi: 10.3390/cancers14061370] [Medline: 35326521]
73. Fusco R, Grassi R, Granata V, et al. Artificial intelligence and COVID-19 using chest CT scan and chest X-ray images: machine learning and deep learning approaches for diagnosis and treatment. J Pers Med 2021 Sep 30;11(10):993. [doi: 10.3390/jpm11100993] [Medline: 34683133]
74. de Freitas Barbosa VA, Gomes JC, de Santana MA, et al. Heg.IA: an intelligent system to support diagnosis of Covid-19 based on blood tests. Res Biomed Eng 2022 Mar;38(1):99-116. [doi: 10.1007/s42600-020-00112-5]
75. Kukar M, Gunčar G, Vovko T, et al. COVID-19 diagnosis by routine blood tests using machine learning. Sci Rep 2021 May 24;11(1):10738. [doi: 10.1038/s41598-021-90265-9] [Medline: 34031483]
76. De La Vega FM, Chowdhury S, Moore B, et al. Artificial intelligence enables comprehensive genome interpretation and nomination of candidate diagnoses for rare genetic diseases. Genome Med 2021 Oct 14;13(1):153. [doi: 10.1186/s13073-021-00965-0] [Medline: 34645491]
77. Vadapalli S, Abdelhalim H, Zeeshan S, Ahmed Z. Artificial intelligence and machine learning approaches using gene expression and variant data for personalized medicine. Brief Bioinform 2022 Sep 20;23(5):bbac191. [doi: 10.1093/bib/bbac191] [Medline: 35595537]
78. Ghazal TM, Al Hamadi H, Umar Nasir M, et al. Supervised machine learning empowered multifactorial genetic inheritance disorder prediction. Comput Intell Neurosci 2022;2022(1):1051388. [doi: 10.1155/2022/1051388] [Medline: 35685134]
79. Goodman KE, Morgan DJ, Hoffmann DE. Clinical algorithms, antidiscrimination laws, and medical device regulation. JAMA 2023 Jan 24;329(4):285-286. [doi: 10.1001/jama.2022.23870] [Medline: 36602795]
80. Saw SN, Ng KH. Current challenges of implementing artificial intelligence in medical imaging. Phys Med 2022 Aug;100:12-17. [doi: 10.1016/j.ejmp.2022.06.003] [Medline: 35714523]
81. Medanki S, Dommati N, Bodapati HH, et al. Artificial intelligence powered glucose monitoring and controlling system: pumping module. World J Exp Med 2024 Mar 20;14(1):87916. [doi: 10.5493/wjem.v14.i1.87916] [Medline: 38590308]
82. Marian CV. Artificial intelligence expert system based on continuous glucose monitoring (CGM) data for auto-adaptive adjustment therapy protocol – how to make sensors and patients to think forward and work together? Presented at: 2021 International Conference on e-Health and Bioengineering (EHB); Nov 18-19, 2021. [doi: 10.1109/EHB52898.2021.9657707]
83. Choudhury AD, Banerjee R, Kimbahune S, Pal A. New Frontiers of Cardiovascular Screening Using Unobtrusive Sensors, AI, and IoT: Academic Press; 2022. [doi: 10.1016/C2019-0-05510-0]
84. Howard JP, Fisher L, Shun-Shin MJ, et al. Cardiac rhythm device identification using neural networks. JACC Clin Electrophysiol 2019 May;5(5):576-586. [doi: 10.1016/j.jacep.2019.02.003] [Medline: 31122379]
85. Oida M, Mizutani T, Hasumi E, et al. Prediction of pacemaker-induced cardiomyopathy using a convolutional neural network based on clinical findings prior to pacemaker implantation. Sci Rep 2024 Mar 22;14(1):6916. [doi: 10.1038/s41598-024-57418-y] [Medline: 38519537]
86. Krittanawong C, Rogers AJ, Johnson KW, et al. Integration of novel monitoring devices with machine learning technology for scalable cardiovascular management. Nat Rev Cardiol 2021 Feb;18(2):75-91. [doi: 10.1038/s41569-020-00445-9] [Medline: 33037325]
87. Venkatesan D, Elangovan A, Winster H, et al. Diagnostic and therapeutic approach of artificial intelligence in neuro-oncological diseases. Biosensors and Bioelectronics: X 2022 Sep;11:100188. [doi: 10.1016/j.biosx.2022.100188]
88. Larburu N, Artetxe A, Escolar V, Lozano A, Kerexeta J. Artificial intelligence to prevent mobile heart failure patients decompensation in real time: monitoring-based predictive model. Mob Inf Syst 2018 Nov 5;2018(1):1-11. [doi: 10.1155/2018/1546210]
89. Shajari S, Kuruvinashetti K, Komeili A, Sundararaj U. The emergence of AI-based wearable sensors for digital health technology: a review. Sensors (Basel) 2023 Nov 29;23(23):9498. [doi: 10.3390/s23239498] [Medline: 38067871]
90. Xie Y, Lu L, Gao F, et al. Integration of artificial intelligence, blockchain, and wearable technology for chronic disease management: a new paradigm in smart healthcare. CURR MED SCI 2021 Dec;41(6):1123-1133. [doi: 10.1007/s11596-021-2485-0] [Medline: 34950987]

XSL•FO
RenderX

91. Karimi F, Amoozgar Z, Reiazi R, Hosseinzadeh M, Rawassizadeh R. Longitudinal analysis of heart rate and physical activity collected from smartwatches. CCF Trans Pervasive Comp Interact 2024 Mar;6(1):18-35. [doi: 10.1007/s42486-024-00147-y]
92. Arora A, Chakraborty P, Bhatia MPS. Intervention of wearables and smartphones in real time monitoring of sleep and behavioral health: an assessment using adaptive neuro-fuzzy technique. Arab J Sci Eng 2022 Feb;47(2):1999-2024. [doi: 10.1007/s13369-021-06078-5]
93. Esakia A, McCrickard S, Harden S, Horning M, Ramalingam N. Using smartwatches to facilitate a group dynamics-based statewide physical activity intervention. Int J Hum Comput Stud 2020 Oct;142:102501. [doi: 10.1016/j.ijhcs.2020.102501]
94. Lin Z, Zheng J, Wang Y, et al. Prediction of the efficacy of group cognitive behavioral therapy using heart rate variability based smart wearable devices: a randomized controlled study. BMC Psychiatry 2024 Mar 6;24(1):187. [doi: 10.1186/s12888-024-05638-x] [Medline: 38448895]
95. Jin X, Cai A, Xu T, Zhang X. Artificial intelligence biosensors for continuous glucose monitoring. Interdisciplinary Materials 2023 Mar;2(2):290-307 [FREE Full text] [doi: 10.1002/idm2.12069]
96. Yoon H, Xuan X, Jeong S, Park JY. Wearable, robust, non-enzymatic continuous glucose monitoring system and its in vivo investigation. Biosens Bioelectron 2018 Oct 15;117:267-275. [doi: 10.1016/j.bios.2018.06.008] [Medline: 29909198]
97. Sawayama J, Takeuchi S. Long‐term continuous glucose monitoring using a fluorescence‐based biocompatible hydrogel glucose sensor. Adv Healthcare Materials 2021 Feb;10(3):2001286. [doi: 10.1002/adhm.202001286]
98. Moreddu R, Wolffsohn JS, Vigolo D, Yetisen AK. Laser-inscribed contact lens sensors for the detection of analytes in the tear fluid. Sensors and Actuators B: Chemical 2020 Aug;317:128183. [doi: 10.1016/j.snb.2020.128183]
99. Arakawa T, Tomoto K, Nitta H, et al. A wearable cellulose acetate-coated mouthguard biosensor for in vivo salivary glucose measurement. Anal Chem 2020 Sep 15;92(18):12201-12207. [doi: 10.1021/acs.analchem.0c01201] [Medline: 32927955]
100. Gao W, Emaminejad S, Nyein HYY, et al. Fully integrated wearable sensor arrays for multiplexed in situ perspiration analysis. Nature New Biol 2016 Jan 28;529(7587):509-514. [doi: 10.1038/nature16521] [Medline: 26819044]
101. Broers ER, Widdershoven J, Denollet J, et al. Personalized eHealth program for life-style CHANGE: results from the "Do Cardiac Health Advanced New Generated Ecosystem (Do CHANGE 2)" randomized controlled trial. Psychosom Med 2020 May;82(4):409-419. [doi: 10.1097/PSY.0000000000000802] [Medline: 32176191]
102. Romero-Tapiador S, Lacruz-Pleguezuelos B, Tolosana R, et al. AI4FoodDB: a database for personalized e-Health nutrition and lifestyle through wearable devices and artificial intelligence. Database (Oxford) 2023 Jul 18;2023:baad049. [doi: 10.1093/database/baad049] [Medline: 37465917]
103. Chatterjee A, Pahari N, Prinz A, Riegler M. AI and semantic ontology for personalized activity eCoaching in healthy lifestyle recommendations: a meta-heuristic approach. BMC Med Inform Decis Mak 2023 Dec 1;23(1):278. [doi: 10.1186/s12911-023-02364-4] [Medline: 38041041]
104. Kargarandehkordi A, Slade C, Washington P. Personalized AI-driven real-time models to predict stress-induced blood pressure spikes using wearable devices: proposal for a prospective cohort study. JMIR Res Protoc 2024 Mar 25;13(1):e55615. [doi: 10.2196/55615] [Medline: 38526539]
105. Mattison G, Canfell O, Forrester D, et al. The influence of wearables on health care outcomes in chronic disease: systematic review. J Med Internet Res 2022 Jul 1;24(7):e36690. [doi: 10.2196/36690] [Medline: 35776492]
106. Ahmed A, Aziz S, Qidwai U, Abd-Alrazaq A, Sheikh J. Performance of artificial intelligence models in estimating blood glucose level among diabetic patients using non-invasive wearable device data. Comput Methods Programs Biomed Update 2023;3:100094. [doi: 10.1016/j.cmpbup.2023.100094]
107. Kario K. Management of hypertension in the digital era: small wearable monitoring devices for remote blood pressure monitoring. Hypertension 2020 Sep;76(3):640-650. [doi: 10.1161/HYPERTENSIONAHA.120.14742] [Medline: 32755418]
108. Wilmink G, Dupey K, Alkire S, et al. Artificial intelligence-powered digital health platform and wearable devices improve outcomes for older adults in assisted living communities: pilot intervention study. JMIR Aging 2020 Sep 10;3(2):e19554. [doi: 10.2196/19554] [Medline: 32723711]
109. Hughes A, Shandhi MMH, Master H, Dunn J, Brittain E. Wearable devices in cardiovascular medicine. Circ Res 2023 Mar 3;132(5):652-670. [doi: 10.1161/CIRCRESAHA.122.322389] [Medline: 36862812]
110. Wang C, He T, Zhou H, Zhang Z, Lee C. Artificial intelligence enhanced sensors - enabling technologies to next-generation healthcare and biomedical platform. Bioelectron Med 2023 Aug 2;9(1):17. [doi: 10.1186/s42234-023-00118-1] [Medline: 37528436]
111. Menon A, Olascoaga LIG, Balanaga V, et al. Shared control of assistive robots through user-intent prediction and hyperdimensional recall of reactive behavior. 2023 Presented at: 2023 IEEE International Conference on Robotics and Automation (ICRA); May 29-2, 2023; London, United Kingdom p. 12638-12644. [doi: 10.1109/ICRA48891.2023.10161509]
112. Potenza DC, Grazioso A, Gaetani F, Mantriota G, Reina G. A smart demonstration unit for upper-limb myoelectric prostheses. Smart Health (2014) 2024 Jun;32:100481. [doi: 10.1016/j.smhl.2024.100481]
113. Burge TA, Jeffers JRT, Myant CW. Applying machine learning methods to enable automatic customisation of knee replacement implants from CT data. Sci Rep 2023 Feb 27;13(1):3317. [doi: 10.1038/s41598-023-30483-5] [Medline: 36849812]

XSL•FO
RenderX

114. Pandelani T, Nemavhola FJ, Subramaniyan AK. 3D-printed smart implants in orthopedic surgery. In: Digital Design and Manufacturing of Medical Devices and Systems: Springer; 2024:187-211. [doi: 10.1007/978-981-99-7100-8_10]
115. Burge TA, Munford MJ, Kechagias S, Jeffers JRT, Myant CW. Automating the customization of stiffness-matched knee implants using machine learning techniques. Int J Adv Manuf Technol 2023 Jun;126(7-8):3725-3737. [doi: 10.1007/s00170-023-11357-6]
116. Paterno L, Filosa M, Anselmino E, et al. Soft transfemoral prosthetic socket with EMG sensing and augmenting feedback: a case study. Research Square. Preprint posted online on May 19, 2023. [doi: 10.21203/rs.3.rs-2889998/v1]
117. Ramakrishna VAS, Chamoli U, Rajan G, Mukhopadhyay SC, Prusty BG, Diwan AD. Smart orthopaedic implants: a targeted approach for continuous postoperative evaluation in the spine. J Biomech 2020 May 7;104:109690. [doi: 10.1016/j.jbiomech.2020.109690] [Medline: 32139096]
118. Lin KC, Cheng IL, Huang YC, et al. The effects of the badminton teaching-assisted system using electromyography and gyroscope on learners' badminton skills. IEEE Trans Learning Technol 2023;16(5):780-789. [doi: 10.1109/TLT.2023.3292215]
119. Aminizadeh S, Heidari A, Dehghan M, et al. Opportunities and challenges of artificial intelligence and distributed systems to improve the quality of healthcare service. Artif Intell Med 2024 Mar;149:102779. [doi: 10.1016/j.artmed.2024.102779] [Medline: 38462281]
120. Thapa C, Camtepe S. Precision health data: requirements, challenges and existing techniques for data security and privacy. Comput Biol Med 2021 Feb;129:104130. [doi: 10.1016/j.compbiomed.2020.104130] [Medline: 33271399]
121. Ross P, Spates K. Considering the safety and quality of artificial intelligence in health care. Jt Comm J Qual Patient Saf 2020 Oct;46(10):596-599. [doi: 10.1016/j.jcjq.2020.08.002] [Medline: 32878718]
122. Bernstam EV, Shireman PK, Meric-Bernstam F, et al. Artificial intelligence in clinical and translational science: successes, challenges and opportunities. Clin Transl Sci 2022 Feb;15(2):309-321. [doi: 10.1111/cts.13175] [Medline: 34706145]
123. Choudhury A. Toward an ecologically valid conceptual framework for the use of artificial intelligence in clinical settings: need for systems thinking, accountability, decision-making, trust, and patient safety considerations in safeguarding the technology and clinicians. JMIR Hum Factors 2022 Jun 21;9(2):e35421. [doi: 10.2196/35421] [Medline: 35727615]
124. Antoniadi AM, Du Y, Guendouz Y, et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. Appl Sci (Basel) 2021;11(11):5088. [doi: 10.3390/app11115088]
125. Manias G, Op Den Akker H, Azqueta A, et al. IHELP: personalised health monitoring and decision support based on artificial intelligence and holistic health records. 2021 Presented at: 2021 IEEE Symposium on Computers and Communications (ISCC); Sep 5-8, 2021; Athens, Greece p. 1-8. [doi: 10.1109/ISCC53001.2021.9631475]
126. Larson DB, Magnus DC, Lungren MP, Shah NH, Langlotz CP. Ethics of using and sharing clinical imaging data for artificial intelligence: a proposed framework. Radiology 2020 Jun;295(3):675-682. [doi: 10.1148/radiol.2020192536] [Medline: 32208097]
127. Almalawi A, Khan AI, Alsolami F, Abushark YB, Alfakeeh AS. Managing security of healthcare data for a modern healthcare system. Sensors (Basel) 2023 Mar 30;23(7):3612. [doi: 10.3390/s23073612] [Medline: 37050672]
128. McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. Lancet Digit Health 2020 May;2(5):e221-e223. [doi: 10.1016/S2589-7500(20)30065-0] [Medline: 33328054]
129. Chen RJ, Wang JJ, Williamson DFK, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng 2023 Jun;7(6):719-742. [doi: 10.1038/s41551-023-01056-8] [Medline: 37380750]
130. Lorenzini G, Shaw DM, Arbelaez Ossa L, Elger BS. Machine learning applications in healthcare and the role of informed consent: ethical and practical considerations. Clin Ethics 2023 Dec;18(4):451-456. [doi: 10.1177/14777509221094476]
131. Maccaro A, Stokes K, Statham L, et al. Clearing the fog: a scoping literature review on the ethical issues surrounding artificial intelligence-based medical devices. J Pers Med 2024 Apr 23;14(5):443. [doi: 10.3390/jpm14050443] [Medline: 38793025]
132. Higgins DC, Johner C. Validation of artificial intelligence containing products across the regulated healthcare industries. Ther Innov Regul Sci 2023 Jul;57(4):797-809. [doi: 10.1007/s43441-023-00530-4] [Medline: 37202591]
133. Fraser AG, Biasin E, Bijnens B, et al. Artificial intelligence in medical device software and high-risk medical devices - a review of definitions, expert recommendations and regulatory initiatives. Expert Rev Med Devices 2023 Jun;20(6):467-491. [doi: 10.1080/17434440.2023.2184685] [Medline: 37157833]
134. Lewis D, Hogan L, Filip D, Wall PJ. Global challenges in the standardization of ethics for trustworthy AI. JICTS 2020;8(2):123-150. [doi: 10.13052/jicts2245-800X.823]
135. Winter JS, Davidson E. Harmonizing regulatory regimes for the governance of patient-generated health data. Telecomm Policy 2022 Jun;46(5):102285. [doi: 10.1016/j.telpol.2021.102285]
136. Federico CA, Trotsyuk AA. Biomedical data science, artificial intelligence, and ethics: navigating challenges in the face of explosive growth. Annu Rev Biomed Data Sci 2024 Aug;7(1):1-14. [doi: 10.1146/annurev-biodatasci-102623-104553] [Medline: 38598860]
137. Palaniappan K, Lin EYT, Vogel S. Global regulatory frameworks for the use of artificial intelligence (AI) in the healthcare services sector. Healthcare (Basel) 2024 Feb 28;12(5):562. [doi: 10.3390/healthcare12050562] [Medline: 38470673]

XSL•FO
RenderX

138. Mackenzie SC, Sainsbury CAR, Wake DJ. Diabetes and artificial intelligence beyond the closed loop: a review of the landscape, promise and challenges. Diabetologia 2024 Feb;67(2):223-235. [doi: 10.1007/s00125-023-06038-8] [Medline: 37979006]
139. Farouk A, Alahmadi A, Ghose S, Mashatan A. Blockchain platform for industrial healthcare: vision and future opportunities. Comput Commun 2020 Mar;154:223-235. [doi: 10.1016/j.comcom.2020.02.058]
140. Mahajan HB. Emergence of healthcare 4.0 and blockchain into secure cloud-based electronic health records systems: solutions, challenges, and future roadmap. Wireless Pers Commun 2022 Oct;126(3):2425-2446. [doi: 10.1007/s11277-022-09535-y]
141. Aggarwal L, Sachdeva S, Goswami P. Quantum healthcare computing using precision based granular approach. Appl Soft Comput 2023 Sep;144:110458. [doi: 10.1016/j.asoc.2023.110458]
142. Davids J, Lidströmer N, Ashrafian H. Artificial intelligence in medicine using quantum computing in the future of healthcare. In: Artificial Intelligence in Medicine: Springer International Publishing; 2022:423-446. [doi: 10.1007/978-3-030-64573-1_338]
143. Tagde P, Tagde S, Bhattacharya T, et al. Blockchain and artificial intelligence technology in e-health. Environ Sci Pollut Res Int 2021 Oct;28(38):52810-52831. [doi: 10.1007/s11356-021-16223-0] [Medline: 34476701]
144. Mbunge E, Muchemwa B, Jiyane S, Batani J. Sensors and healthcare 5.0: transformative shift in virtual care through emerging digital health technologies. Global Health Journal 2021 Dec;5(4):169-177. [doi: 10.1016/j.glohj.2021.11.008]

Abbreviations

AI: artificial intelligence
CGM: continuous glucose monitor
CNN: convolutional neural network
CT: computed tomography
FDA: US Food and Drug Administration
ML: machine learning
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR : Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Scoping Reviews

Edited by J Sarvestan; submitted 09.Feb.2025; peer-reviewed by G Wilson, M Al-Raeei, MM Khan; revised version received 17.Apr.2025; accepted 18.Apr.2025; published 04.Mar.2026.

Please cite as:
Ponnambalath Mohanadas H, Manikandan A, Ismail AF, Tucker N, Jaganathan SK
Artificial Intelligence Applications in Medical Devices for Personalized Health Care Solutions: Systematic Review
J Med Internet Res 2026;28:e72410
URL: https://www.jmir.org/2026/1/e72410
doi:10.2196/72410

© Hemanth Ponnambalath Mohanadas, A Manikandan, Ahmad Fauzi Ismail, Nick Tucker, Saravana Kumar Jaganathan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of Interventions for Internet, Smartphone, and Gaming Addictions: Umbrella Review and Meta–Meta-Analysis

Minggang Zhang[1*], PhD; Tong Lan[2*], MSc; Tao Song[1], PhD; Xiaochun Wang[1], PhD

[1]School of Psychology, Shanghai University of Sport, 399 ChangHai Road, Shanghai, 200438, P.R. China, Shanghai, China

[2]Psychological Counseling Center, Nanchang Institute of Technology, Nanchang, China

[*]these authors contributed equally

Corresponding Author:
Xiaochun Wang, PhD
School of Psychology, Shanghai University of Sport, 399 ChangHai Road, Shanghai, 200438, P.R. China, Shanghai, China

Abstract

Background: Digital addiction, including internet, smartphone, and gaming addiction, has emerged as a significant global health concern. Although a wide range of interventions has been evaluated, the fragmented and siloed nature of existing meta-analyses limits a clear understanding of the comparative effectiveness of different interventions across addiction subtypes.

Objective: This umbrella review and meta–meta-analysis aimed to estimate the overall effectiveness of interventions for digital addiction and examine differential effects according to addiction subtype, intervention modality, study design, and control condition.

Methods: A systematic search of 5 electronic databases (PubMed, Web of Science, Scopus, APA PsycInfo, and the Cochrane Library) was conducted from inception to June 24, 2025. Eligible studies were systematic reviews with meta-analyses evaluating interventions for internet, smartphone, or gaming addiction. Random-effects models were applied to synthesize standardized mean differences (SMDs). Methodological quality and certainty of evidence were assessed using A Measurement Tool to Assess Systematic Reviews 2 and the Grading of Recommendations Assessment, Development, and Evaluation framework.

Results: A total of 29 meta-analyses, comprising 52 effect sizes and 66,530 participants, were included (I^2=95.13%). Overall, interventions demonstrated a large and statistically significant effect in reducing digital addiction symptoms (SMD=–1.44, 95% CI –1.67 to –1.21; *P*=.003). Subgroup analyses indicated that the largest effects were observed for internet addiction (SMD=–1.70, 95% CI –1.99 to –1.42), followed by gaming addiction (SMD=–0.82, 95% CI –1.09 to –0.56) and smartphone addiction (SMD=–0.80, 95% CI –1.39 to –0.21). Exercise-based interventions, particularly those integrated with psychological approaches, showed large effect sizes (SMD=–3.14, 95% CI –4.30 to –1.97); however, this finding was based on a very limited number of effect sizes and should be interpreted cautiously. In addition, randomized controlled trials yielded larger effects than mixed study designs, and no-intervention controls were associated with larger effect sizes than mixed control conditions. The certainty of evidence was generally low.

Conclusions: Interventions for digital addiction are effective, although their magnitude of benefit varies by addiction subtype and intervention modality. These findings support the use of tailored and multimodal intervention strategies while highlighting the need for more rigorous, high-quality, and balanced evidence across different forms of digital addiction.

Trial Registration: PROSPERO CRD420251084061; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251084061

(J Med Internet Res 2026;28:e81705) doi:10.2196/81705

KEYWORDS

internet addiction; smartphone addiction; gaming disorder; umbrella review; meta–meta-analysis; intervention effectiveness; PRISMA

Introduction

In the digital era, behavioral addictions associated with the excessive use of internet-based technologies have emerged as major public health concerns. A recent global meta-analysis synthesizing 504 prevalence studies involving more than 2 million individuals across 64 countries reported pooled prevalence rates of 14.2% for internet addiction, 26.99% for smartphone addiction, and 6.04% for gaming addiction, with pronounced increases observed during the COVID-19 pandemic [1]. Consistent with these findings, pandemic-restricted analyses have shown prevalence rates rising to 30.7% for smartphone addiction and 10.6% for internet addiction [2,3], underscoring a growing public health burden that spans age groups, sexes, and cultural contexts. The clinical relevance of this burden is increasingly reflected in international diagnostic frameworks

XSL•FO
RenderX

[4]. In 2019, the World Health Organization incorporated gaming disorder into the *International Classification of Diseases, 11th Revision* (*ICD-11*), formally assigning it to the "Disorders due to addictive behaviors" section [5]. Although internet and smartphone addiction have not yet been formally recognized within the *ICD-11*, accumulating evidence links all 3 conditions to emotional dysregulation, sleep disturbances, impaired academic and occupational functioning, and reduced quality of life [6-9].

Over the past decade, a broad range of interventions has been evaluated to address digital addictions, including cognitive behavioral therapy (CBT), group counseling, exercise training, mindfulness-based approaches, and multicomponent programs. A number of meta-analyses have synthesized evidence on the effectiveness of these interventions in reducing symptoms of internet, gaming, and smartphone addiction. For example, a meta-analysis of 59 randomized trials reported large overall effects for interventions targeting internet addiction and identified combined CBT and exercise interventions as particularly effective [10]. Similarly, exercise-based interventions have demonstrated substantial reductions in smartphone dependence among children and adolescents, with optimal effects observed for moderate-intensity aerobic exercise performed approximately 3 times per week over a 12-week period [11]. In the context of gaming disorder, a synthesis of 38 studies involving 9524 participants reported a moderate but highly heterogeneous treatment effect [12].

Despite these encouraging findings, the evidence base for interventions targeting digital addictions remains fragmented. Most meta-analyses to date focus exclusively on a single type of addiction, such as internet use, smartphone dependence, or gaming disorder, while adopting diverse outcome measures, inconsistent risk-of-bias frameworks, and varying inclusion criteria [12-14]. This fragmentation limits the comparability of intervention effects across different behavioral addictions and hinders the development of unified, evidence-based clinical guidelines. Consequently, a higher-level synthesis is needed to systematically integrate and compare findings across multiple meta-analyses.

Umbrella reviews, which synthesize evidence from multiple systematic reviews and meta-analyses, provide a robust methodological framework for addressing these challenges [15]. Although recent umbrella reviews have made initial progress, important gaps remain. For instance, one umbrella review pooled 5 meta-analyses addressing digital addiction, encompassing internet, smartphone, and gaming addiction and general screen use. While cognitive behavioral and exercise-based interventions emerged as promising, the study did not differentiate effects by addiction subtype or conduct quantitative meta–meta-analysis comparisons, limiting conclusions regarding relative effectiveness [16]. Another umbrella review focused on internet and smartphone addiction and confirmed potential benefits of psychotherapeutic and exercise-based interventions; however, it similarly lacked quantitative comparisons across addiction categories or intervention modalities [17].

To address these limitations, this study conducted an umbrella review and meta–meta-analysis of interventions targeting internet, smartphone, and gaming addictions. Specifically, we aimed to (1) quantify the overall effectiveness of major intervention modalities on addiction subtypes and (2) compare effect sizes across internet, smartphone, and gaming addictions. By providing a comparative, high-level synthesis of the current intervention landscape, this study sought to inform clinicians, researchers, and policymakers working to mitigate the psychosocial and health consequences of digital-age behavioral addictions.

Methods

Protocol and Registration

This umbrella review was preregistered in PROSPERO (CRD420251084061). This review was conducted and reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement [18].

Search Strategy

Five electronic databases (PubMed, Web of Science, Scopus, APA PsycInfo, and the Cochrane Library) were systematically searched from inception to June 24, 2025. The search strategy included a combination of 4 key blocks of terms related to internet addiction, intervention, and meta-analyses. Table S1 in Multimedia Appendix 1 provides our electronic search strategy.

Eligibility Criteria

We included systematic reviews with a meta-analysis examining the effects of interventions on internet, smartphone, or gaming addiction. For addiction subtype classification, studies were categorized as targeting internet addiction, smartphone addiction, or gaming addiction based on the primary addiction construct explicitly defined by the authors, regardless of the device used (eg, gaming on a smartphone). No reclassification of primary studies was performed to avoid introducing additional subjectivity.

Specific inclusion criteria were also defined using the population, intervention, comparator, outcome, and study design framework as follows:

- Population—reviews were eligible if they included studies that enrolled individuals with addictive or excessive or problematic internet or smartphone use or gaming.
- Intervention—reviews were eligible if they included studies that investigated any type of psychological, behavioral, or physical activity intervention, among other types, that was implemented for individuals as an addiction intervention. Studies evaluating pharmacological interventions were not eligible for inclusion.
- Comparator—reviews were eligible if they included controlled studies in which the addiction intervention was compared with any type of control group.
- Outcomes—reviews were eligible if they included studies that reported the results of at least 1 meta-analysis of an intervention for internet, smartphone, or gaming addiction at a postintervention time point vs controls. The outcome measures needed to use valid and reliable rating scales suitable for internet, smartphone, or gaming addiction.

XSL•FO
RenderX

When more than one measure was described by a study, the primary outcome was used. If the primary outcome was not specified, the data from the measure reported first were extracted. Meta-analyses without any effect size were excluded.

- Study design—reviews were eligible if they assessed quantitative randomized controlled trials (RCTs) only and if they assessed a mixed evidence base (RCTs with other designs, such as non-RCTs and quasi-experimental studies). Reviews were included if they were published in English in peer-reviewed journals. Conferences, theses, and commentaries were not included.

Data Extraction

Study characteristics and outcome measures were extracted by MZ and TL. Data entry was independently checked by TS after the data were entered into a Microsoft Excel spreadsheet. The following data from two domains were extracted: (1) study characteristics (author, year of publication, country, mean age of participants, number of included studies, and sample size [percentage of participants by gender]) and (2) interventions (addiction types, intervention types, control group conditions, and study design). All the extracted data were combined with those from the original review.

Evaluations of Methodological Quality and Quality of Evidence

The methodological quality of eligible systematic reviews with meta-analyses was evaluated using a critical appraisal tool: A Measurement Tool to Assess Systematic Reviews 2 (AMSTAR 2) [19]. AMSTAR 2 contains 16 checklist items rated as "yes" (1 point), "partial yes" (0.5 points), and "no" (0 points) and is used to identify the risk of bias for systematic reviews that involve randomized and/or nonrandomized intervention studies. Among the AMSTAR 2 items, 7 are considered "critical" and 9 are considered "noncritical." Reviews were scored as follows: "critically low" (>1 critical weakness and 3 noncritical weaknesses), "low" (>1 critical weakness and <3 noncritical weaknesses), "moderate" (1 critical weakness and <3 noncritical weaknesses), or "high" (no critical weaknesses and <3 noncritical weaknesses).

The certainty of evidence was assessed using the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach [20]. Five domains were considered: risk of bias, inconsistency, indirectness, imprecision, and publication bias. Each outcome of internet-related addiction was assigned a certainty rating of "high," "moderate," "low," or "very low" based on these criteria. Two authors independently performed the AMSTAR 2 and GRADE evaluations, with the third author solving any discrepancies through consensus.

Statistical Analysis

Calculation of Effect Sizes

A meta-analysis was performed in this meta-review through Stata/MP (version 18.0; StataCorp) to calculate individual study and pooled effect sizes. To synthesize the findings on the effectiveness of interventions for internet, smartphone, and gaming addictions, meta-analyses were performed by combining the effect sizes and 95% CIs reported in each review via a random-effects model. Standardized effect sizes (standardized mean differences; SMDs) with 95% CIs were used as the effect measure for the meta-analyses. For consistency across studies, all effect sizes were converted such that negative values reflected beneficial intervention effects. All statistical tests were 2 tailed, and statistical significance was set at a P value of less than .05.

Testing Homogeneity and Potential Publication Bias

The heterogeneity of effect sizes was tested using the Cochran Q test and I^2 statistic, which indicates heterogeneity in percentages [21]. Significant heterogeneity was deemed present if the P value of the Q test was less than .10, and 0% indicated no heterogeneity, whereas 25%, 50%, and 75% indicated low, moderate, and high heterogeneity, respectively. To consider potential publication bias, a visual inspection of a funnel plot was conducted, and we also applied the nonparametric trim-and-fill method using a random-effects model. The method was implemented using a linear estimator with imputation on the right side of the funnel plot, reflecting a scenario in which studies with smaller or nonsignificant (less negative) effect sizes may be underrepresented.

Subgroup Analyses

For subgroup analyses, we extracted subgroup-specific effect sizes and corresponding data as reported in the included systematic reviews and pooled these results across reviews. As in the original review, the subgroup analyses were planned according to addiction type, intervention type, control group type, and study design. We used a random-effects analysis to estimate the effect size and compare subgroups.

Results

Search Results and Study Characteristics

A total of 1719 studies were identified through the database searches, and after the titles and abstracts were initially checked, 49 (2.9%) of the studies were included for full-text screening. We ultimately included 29 review and meta-analysis studies that met the full inclusion criteria. The PRISMA flowchart and checklist are presented in Figure 1 and Checklist 1, respectively, and the list of reasons for exclusion after full-text review is presented in Table S2 in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

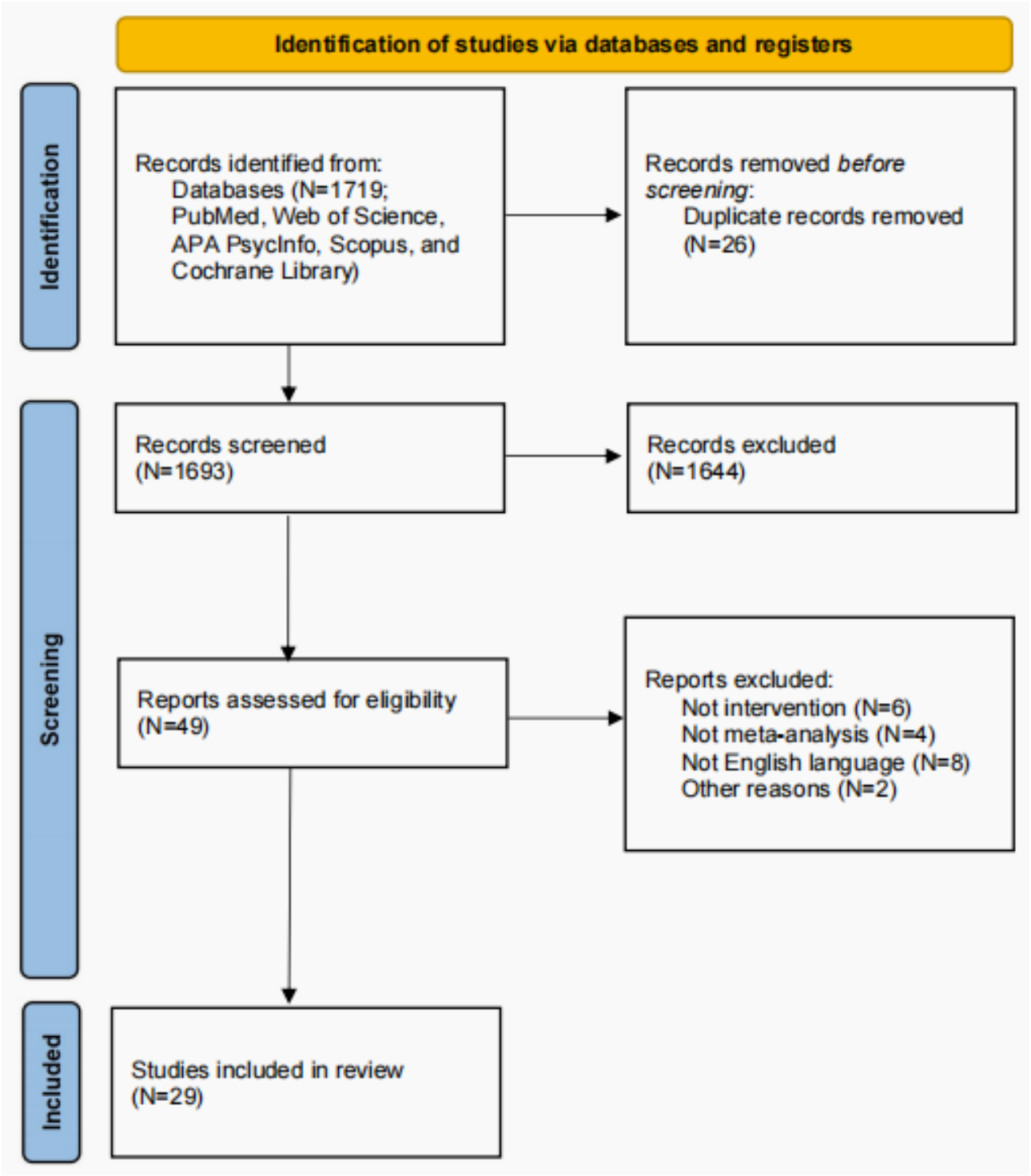

The eligible studies were published from 2017 to 2025 and were conducted in China (15/29, 51.7%), South Korea (6/29, 20.7%), Austria (2/29, 6.9%), Australia (1/29, 3.4%), Lithuania (1/29, 3.4%), Spain (1/29, 3.4%), the United States (1/29, 3.4%), Norway (1/29, 3.4%), and Turkey (1/29, 3.4%). Of these 29 studies, 18 (62.1%) assessed the intervention effect on internet addiction by calculating 38 effect sizes, 8 (27.6%) assessed the intervention effect on internet gaming addiction by calculating 10 effect sizes, and 4 (13.8%) assessed the intervention effect on smartphone addiction by calculating 4 effect sizes. A total of 31 effect sizes were calculated to assess psychological interventions for overall digital addiction, 14 effect sizes were calculated to assess exercise-based interventions, and 7 effect sizes were calculated to assess other types of interventions. In total, 41.4% (12/29) of the studies conducted a meta-analysis of RCTs, 58.6% (17/29) of the studies assessed mixed study designs, 24.1% (7/29) of the studies specifically assessed no-intervention controls, and 75.9% (22/29) of the studies assessed mixed controls. A summary of the participant demographics is provided in Table S3 in Multimedia Appendix 1.

XSL•FO
RenderX

Regarding the methodological quality and quality of evidence of all included studies, the percentages of studies with low and critically low methodological quality were 13.8% (4/29) and 86.2% (25/29), respectively, according to the AMSTAR 2 score (Table S4 in Multimedia Appendix 1). The percentages of reviews with moderate, low, and very low quality of evidence were 3.4% (1/29), 20.7% (6/29), and 75.9% (22/29), respectively, according to the GRADE score (Table S5 in Multimedia Appendix 1).

The visual and statistical results indicated significant asymmetry in the funnel plot (β=–3.00; *P*<.001), suggesting potential publication bias or small-study effects (Figure 2). The trim-and-fill procedure identified 14 potentially missing studies, resulting in a total of 66 studies (n=52, 78.8% observed; n=14, 21.2% imputed; Figure S1 in Multimedia Appendix 1). The observed pooled effect size was SMD=–1.439 (95% CI –1.67 to –1.21). After imputing the potentially missing studies, the adjusted pooled effect size was SMD=–1.09 (95% CI –1.38 to –0.81). Although the adjusted effect size was smaller in magnitude, it remained statistically significant, indicating a robust intervention effect despite potential publication bias.

Figure 2. Funnel plot of all included effect sizes. REML: random effect model; SMD: standardized mean difference.

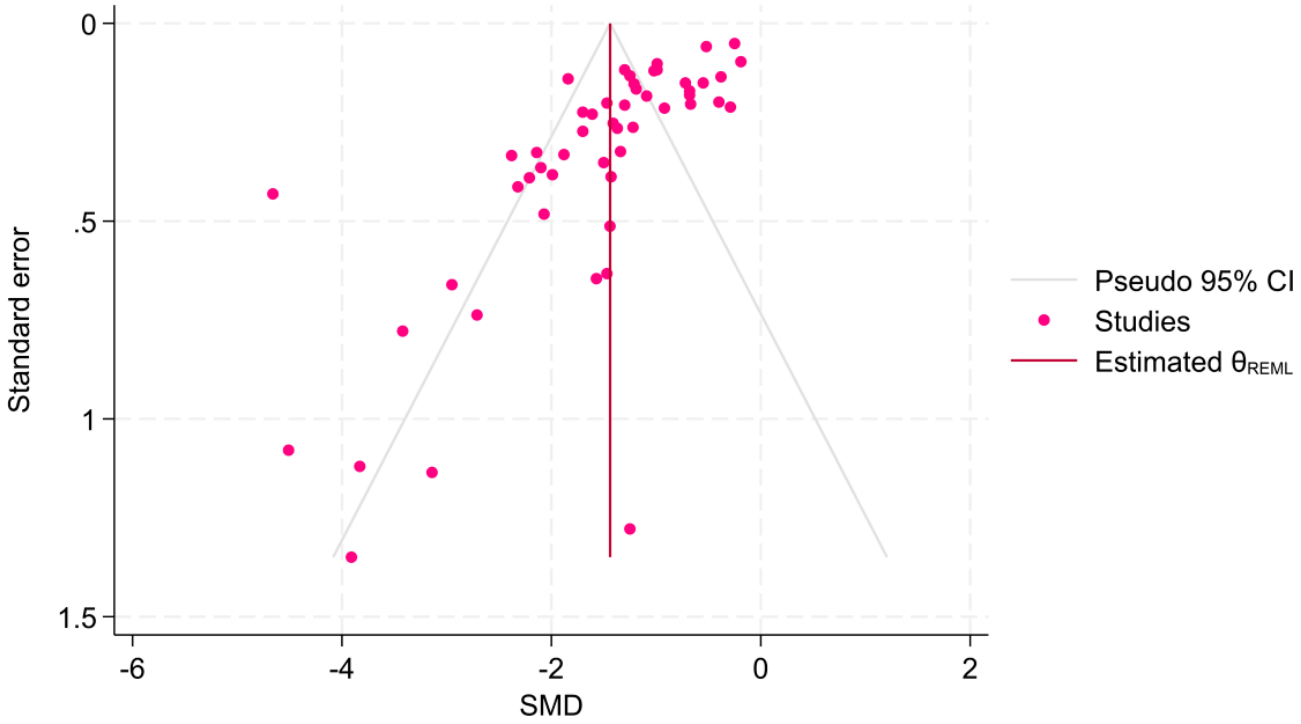

Meta–Meta-Analysis Results

The overall analysis revealed that interventions significantly reduced digital addiction symptoms, with an SMD=–1.44 (95% CI –1.67 to –1.21; *P*=.003; I^2=95.13%; Figure 3 [10,12,13,22-47]). Table 1 summarizes the results of the overall and subgroup meta-analyses. For different addiction types, the subgroup analysis revealed a significant effect (Q_b=21.43; *df*=2; *P*=.003), and the interventions had the largest reducing effect on internet addiction (SMD=–1.70, 95% CI –1.99 to –1.42), followed by internet gaming addiction (SMD=–0.82, 95% CI –1.09 to –0.56) and smartphone addiction (SMD=–0.80, 95% CI –1.39 to –0.21).

XSL•FO
RenderX

Figure 3. Forest plot for overall addiction and subaddictions [10,12,13,22-47]. SMD: standardized mean difference.

XSL•FO
RenderX

Table . Results of overall and subgroup meta-analyses.

	Number of effects	SMD[a] (95% CI)	I^2 (%)	Test of subgroup differences	
				Q_b (*df*)	*P* value
Overall	52	–1.44 (–1.67 to –1.21)	95.13	—[b]	—
Subgroup analysis of all data					
Addiction				21.43 (2)	.003
Internet addiction	38	–1.7 (–1.99 to –1.42)	93.96		
Internet gaming addiction	10	–0.82 (–1.09 to –0.56)	88.43		
Smartphone addiction	4	–0.8 (–1.39 to –0.21)	86.9		
Intervention				58.68 (5)	.002
Psychological	31	–1.16 (–1.38 to –0.95)	92.53		
Exercise	14	–2.01 (–2.52 to –1.50)	93.9		
Exercise+psychological	2	–3.14 (–4.30 to –1.97)	—		
Biofeedback	2	–2.7 (–5.91 to 0.51)	88.6		
Behavioral	2	–0.46 (–0.65 to –0.26)	—		
Electrotherapy	1	–1.25 (–3.75 to 1.25)	—		
Study design				11.89 (1)	.008
RCT[c]	29	–1.83 (–2.20 to –1.46)	91.7		
Mixed	23	–1.06 (–1.29 to –0.82)	94.05		
Control				17.97 (1)	.007
No intervention	16	–2.45 (–3.04 to –1.86)	85.65		
Mixed	36	–1.12 (–1.30 to –0.93)	92.12		
Subgroup analysis for internet addiction					
Intervention				13.33 (4)	.01
Psychological	20	–1.34 (–1.63 to –1.06)	91.6		
Exercise	13	–2.08 (–2.63 to –1.53)	92.87		
Exercise+psychological	2	–3.14 (–4.30 to –1.97)	—		
Biofeedback	2	–2.7 (–5.91 to 0.51)	88.6		
Electrotherapy	1	–1.25 (–3.75 to 1.25)	—		
Control				12.44 (1)	.006
No intervention	16	–2.45 (–3.04 to –1.86)	85.65		
Mixed	22	–1.31 (–1.54 to –1.08)	90.41		
Study design				9.22 (1)	.005
RCT	24	–2.05 (–2.45 to –1.65)	88.14		
Mixed	14	–1.24 (–1.57 to –0.92)	94.13		
Subgroup analysis for gaming addiction					
Intervention				6.43 (1)	.01
Psychological	8	–0.93 (–1.24 to –0.62)	88.37		
Behavioral	2	–0.46 (–0.65 to –0.26)	—		
Study design				0.36 (1)	.55
RCT	3	–1.02 (–1.75 to –0.28)	83.03		

XSL•FO
RenderX

	Number of effects	SMD[a] (95% CI)	I^2 (%)	Test of subgroup differences	
				Q_b (*df*)	*P* value
Mixed	7	–0.77 (–1.06 to –0.49)	88.64		
Subgroup analysis for smartphone addiction					
Intervention				23.60 (1)	.005
Psychological	3	–0.41 (–0.68 to –0.13)	—		
Behavioral	1	–1.3 (–1.53 to –1.07)	—		
Study design				0.00 (1)	.99
RCT	2	–0.81 (–1.80 to 0.18)	94.26		
Mixed	2	–0.82 (–1.93 to 0.28)	66.68		

[a]SMD: standardized mean difference.

[b]Not applicable.

[c]RCT: randomized controlled trial.

For the intervention types, the overall subgroup analysis revealed significant differences in efficacy among the included interventions ($Q_b(5)=58.68$; $P=.002$). Although only 2 effect sizes were calculated to assess the efficacy of a psychological intervention plus exercise, this intervention had the greatest effect on overall addiction (SMD=–3.14, 95% CI –4.30 to –1.97), followed by biofeedback (SMD=–2.70, 95% CI –5.91 to 0.51) and exercise (SMD=–2.01, 95% CI –2.52 to –1.50) interventions. Specifically, the interventions also had significant effects on reducing addictive symptoms for internet addiction ($Q_b(4)=13.33$; $P=.01$), internet gaming addiction ($Q_b(1)=6.43$; $P=.01$), and smartphone addiction ($Q_b(1)=23.60$; $P=.005$). For internet addiction, the intervention with the largest reducing effect was psychological plus exercise (SMD=–3.14, 95% CI –4.30 to –1.97), followed by biofeedback (SMD=–2.70, 95% CI –5.91 to 0.51) and exercise (SMD=–2.08, 95% CI –2.63 to –1.53) interventions. For internet gaming addiction, only 2 types of interventions were assessed. Compared with those of the behavioral intervention, the larger effect sizes of the psychological intervention indicated that it had greater effects on reducing addictive symptoms (SMD=–0.93, 95% CI –1.24 to –0.62). Finally, for smartphone addiction, only 3.4% (1/29) of the studies assessed a behavioral intervention and reported greater effects than those of a psychological intervention (SMD=–0.41, 95% CI –0.68 to –0.31).

With respect to study design, the subgroup analysis revealed that RCTs found a significantly greater effect in reducing overall addictive symptoms (SMD=–1.83, 95% CI –2.20 to –1.46) and internet addiction symptoms (SMD=–2.05, 95% CI –2.25 to –1.65) but not smartphone addiction ($Q_b(1)=0.00$; $P=.99$) or internet gaming addiction ($Q_b(1)=0.36$; $P=.55$) symptoms.

For the control conditions, the subgroup analysis revealed that a no-intervention control had a significant and greater effect on reducing overall and internet addiction with the same included effect sizes (SMD=–2.45, 95% CI –3.04 to –1.86) than a mixed control. No subgroup analyses were conducted for internet gaming addiction or smartphone addiction, as only 1 control was included.

The results of the subgroup analyses for different addiction types, intervention types, study designs, and control group types are presented as forest plots in Figures S2 to S11 in Multimedia Appendix 1.

Discussion

Principal Findings

In this umbrella review and meta–meta-analysis, evidence from 29 systematic reviews and meta-analyses was synthesized to evaluate the effectiveness of interventions targeting internet, smartphone, and gaming addictions. Overall, interventions demonstrated large and statistically significant effects in reducing digital addiction symptoms. Among the 3 addiction subtypes, interventions targeting internet addiction yielded the largest pooled effect, whereas those targeting gaming and smartphone addiction produced more modest but still clinically meaningful reductions. These findings support the overall efficacy of interventions for digital behavioral addictions and suggest that responsiveness to intervention may differ across addiction subtypes. Notably, to our knowledge, this study represents the first quantitative umbrella synthesis that directly compares effect sizes across different categories of digital addiction, thereby providing a more granular and comparative perspective on intervention effectiveness.

Our findings indicate clear differences in intervention effects across addiction subtypes. Interventions for internet addiction showed the largest pooled effect (SMD=–1.70), followed by interventions for gaming addiction (SMD=–0.82) and smartphone addiction (SMD=–0.80). These results are broadly consistent with those of prior meta-analyses reporting moderate effects for gaming disorder and smartphone addiction [11,12], while extending previous work by demonstrating the comparatively greater responsiveness of internet addiction to intervention. One potential explanation is that internet addiction, often conceptualized as a broader and more heterogeneous construct, encompasses diverse behaviors such as social media use, web browsing, and streaming, which may be more amenable to general behavioral regulation strategies [48,49]. In contrast, gaming and smartphone addiction may involve more entrenched

XSL•FO
RenderX

reward-based and habit-forming mechanisms frequently embedded within immersive or socially reinforced contexts, thereby posing greater challenges for behavior change [50,51]. These distinctions underscore the importance of tailoring intervention strategies to the specific behavioral and cognitive characteristics of each addiction subtype.

With respect to intervention modalities, our results highlight a notable and often underemphasized finding. Exercise-based interventions, both as stand-alone approaches and in combination with psychological therapies, were associated with particularly large effect sizes. The combined psychological and exercise intervention demonstrated the largest pooled effect (SMD=–3.14), followed by exercise-only interventions (SMD=–2.01) and biofeedback-based approaches (SMD=–2.70). This pattern diverges from much of the existing literature, which has predominantly emphasized cognitive behavioral and psychotherapeutic interventions [17]. Nevertheless, our findings are consistent with emerging evidence supporting the benefits of physical activity for reducing internet and smartphone addiction symptoms [14,22]. Potential mechanisms underlying these effects include improvements in executive functioning, stress regulation, emotional control, and sleep quality [52,53]. At the same time, the strong performance of psychological-only interventions reinforces their established role in treatment and suggests that psychological and exercise-based modalities may exert complementary effects.

Subgroup analyses further revealed substantial heterogeneity in intervention effects not only across addiction subtypes but also within each subtype, depending on intervention modality, study design, and control condition. For internet addiction, combined psychological and exercise-based interventions yielded the largest effects. In contrast, behavioral interventions appeared more effective than psychological interventions for smartphone addiction, aligning with prior evidence regarding the role of physical activity in enhancing self-regulation and reducing screen-based behaviors [14,22]. For gaming addiction, relatively larger effects were observed for psychological interventions compared with behavioral approaches. These patterns support the increasingly accepted view that different forms of digital addiction may be underpinned by distinct neurocognitive and behavioral profiles and, therefore, require intervention strategies tailored to the specific structure and reinforcement mechanisms of each disorder [51].

Beyond intervention modality, methodological characteristics, including study design and control condition, emerged as important moderators of treatment effects. Across all 3 addiction subtypes, RCTs yielded larger effect sizes than non-RCT or mixed designs, with this difference being most pronounced for internet addiction. This finding reinforces the importance of experimental rigor in evaluating intervention efficacy [15,54]. Regarding control conditions, studies using no-intervention or waitlist controls consistently reported larger effect sizes than those using active or mixed control groups, particularly for internet and smartphone addiction. While this pattern may partly reflect expectancy effects or the absence of placebo balancing, it also raises concerns about potential effect size inflation in less rigorously controlled designs [55,56]. Similar trends have been documented in psychotherapy research, where studies using no-treatment controls often report larger effects than those using active or placebo comparators [57]. These observations highlight the necessity of incorporating well-designed active or attention-matched control conditions in future trials to enhance interpretability and comparability.

Although the meta–meta-analytic findings provide strong support for the effectiveness of interventions, caution is warranted when interpreting the apparent superiority of certain intervention types. In this synthesis, interventions for internet addiction accounted for a disproportionately large share of the available effect sizes, whereas smartphone addiction and gaming addiction were represented by substantially fewer meta-analyses and primary studies. This imbalance likely reflects historical research priorities and differences in diagnostic recognition, particularly given the relatively recent inclusion of gaming disorder in the *ICD-11* [5,58]. However, such disparities limit the robustness of cross-subtype comparisons. Specifically, the smaller evidence base for smartphone and gaming addiction may reduce statistical power, increase sensitivity to individual study characteristics, and obscure true intervention effects. Consequently, the comparatively lower pooled effects observed for these subtypes should not be interpreted as definitive evidence of inferior treatment efficacy. Addressing this limitation will require greater investment in high-quality, subtype-specific intervention research.

Several contextual factors should also be considered when interpreting these findings. First, most included meta-analyses were conducted in middle- and high-income countries, which may limit the generalizability of the results to settings affected by the digital divide, where access to digital technologies and intervention resources differs substantially. Second, reverse causation cannot be fully excluded, as individuals with more severe addiction symptoms may be more likely to participate in intervention studies, potentially inflating observed effects, particularly in non-RCT designs. Moreover, intervention effectiveness may vary by gender, age, and regional context; however, most included reviews did not report stratified analyses for these variables. Future studies should address these gaps to improve the external validity of digital addiction interventions.

Conceptual overlap among digital addiction constructs should also be considered. In practice, behaviors such as gaming may occur on multiple platforms, including smartphones and computers, which may blur the boundaries among internet, smartphone, and gaming addiction. This overlap could partially attenuate observed differences in intervention effectiveness across addiction subtypes and should be taken into account when interpreting comparative results. Future research should also establish a more detailed distinction among these potential impacts. Finally, the substantial heterogeneity observed in the overall analysis (I^2=95.13%) likely reflects multiple sources of variability beyond intervention type. These include differences in diagnostic criteria and measurement instruments, variation in symptom severity and age composition of samples, and cultural and regional differences in digital use patterns. Additionally, the inclusion of both RCTs and mixed study designs may have further contributed to heterogeneity. Such

XSL•FO
RenderX

variability underscores the need for cautious interpretation of pooled estimates.

This study has several key strengths. First, it is the first umbrella review and meta–meta-analysis to systematically compare the effectiveness of interventions across 3 distinct digital addiction types, addressing a major gap in prior literature. Second, it evaluated multiple intervention modalities, offering a broader understanding of the treatment landscape beyond the dominant CBT-centric models. Third, by applying rigorous tools such as AMSTAR 2 and GRADE, it appraised not only effect sizes but also the credibility of the evidence. Notably, the use of these tools revealed that all of the included meta-analyses were rated as low or critically low in methodological quality, and most effect estimates were graded as low or very low in certainty. This outcome reflects the stringent criteria of both appraisal systems. AMSTAR 2 places strong emphasis on protocol registration, risk-of-bias assessment in primary studies, and reporting transparency—criteria often underreported in existing meta-analyses [19]. Similarly, GRADE tends to downgrade evidence derived from observational studies or indirect comparisons and penalizes for inconsistency, imprecision, and publication bias, which were common across the included reviews. While these findings do not negate the observed intervention effects, they underscore the need to improve the methodological transparency and consistency of future meta-analyses in this domain [59].

Nevertheless, several limitations must be acknowledged. All of the included meta-analyses were rated as low or critically low in methodological quality, and the overall certainty of evidence was predominantly classified as very low. Substantial heterogeneity and indications of publication bias further limit the interpretability and generalizability of the findings. In addition, studies focusing on gaming and smartphone addiction were underrepresented relative to those addressing internet addiction, and several intervention approaches were supported by only a small number of studies. These limitations highlight the need for more rigorous, targeted, and subtype-specific intervention research. Finally, the larger effect sizes observed in studies using no-intervention controls are consistent with prior psychotherapy research and likely reflect expectancy and placebo-related effects. Future intervention trials should prioritize the use of active or attention-matched control conditions to reduce bias and improve the interpretability of treatment effects.

Despite these limitations, the findings of this study have important implications for clinical and applied practice. Given the comparatively strong effects observed for integrated interventions, particularly those combining psychological therapy with exercise, clinicians and educators may consider adopting multimodal treatment strategies, especially in school-based or youth-oriented settings. Moreover, the observed benefits of stand-alone exercise and biofeedback interventions underscore the potential value of nonpharmacological, resource-efficient approaches that may be scalable in low-resource or community contexts. Tailoring interventions to specific addiction subtypes, developmental stages, and cultural backgrounds while maintaining rigorous study designs and standardized outcome measures may further enhance the effectiveness and ecological validity of future digital addiction interventions.

Conclusions

This umbrella review demonstrates that interventions targeting internet, smartphone, and gaming addictions are generally effective. Integrated psychological-exercise interventions appear particularly promising; however, the current evidence remains preliminary due to the limited number of available studies. High-quality evidence is especially scarce for smartphone and gaming addictions. Future research should prioritize methodological rigor, expand intervention trials for underrepresented addiction subtypes, and develop tailored strategies based on distinct addiction profiles. Collectively, these findings provide an evidence-based foundation for designing more precise, scalable, and context-sensitive interventions in both clinical and educational settings.

Acknowledgments

Generative artificial intelligence was not used to draft any portion of this manuscript.

Funding

This work was supported by the Ministry of Education of the People’s Republic of China (grant 21JZDW007), the Shanghai Municipal Education Commission “Primary and Secondary School Kindergarten Student Care Project Research” special project (Runxin Project), and the China Scholarship Council (grant 202408310317).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

MZ and TL were responsible for the conception, design, and hypothesis. MZ, TL, and TS conducted the systematic search, data extraction, and assessment of risk of bias. MZ and TL completed all statistical analyses. MZ, TL, and XW contributed to the drafting of the manuscript and its revisions, including the final version. XW supervised the work. All authors read and approved a draft of the manuscript and the final version.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy; Excluded studies; characteristics of the included studies; A Measurement Tool to Assess Systematic Reviews 2; and Grading of Recommendations Assessment, Development, and Evaluation tool.
[DOC File, 519 KB - jmir_v28i1e81705_app1.doc]

Checklist 1
PRISMA checklist.
[DOCX File, 23 KB - jmir_v28i1e81705_app2.docx]

References

1. Meng SQ, Cheng JL, Li YY, et al. Global prevalence of digital addiction in general population: a systematic review and meta-analysis. Clin Psychol Rev 2022 Mar;92:102128. [doi: 10.1016/j.cpr.2022.102128] [Medline: 35150965]
2. Liu X, Gui Z, Chen ZM, et al. Global prevalence of internet addiction among university students: a systematic review and meta-analysis. Curr Opin Psychiatry 2025 May 1;38(3):182-199. [doi: 10.1097/YCO.0000000000000994] [Medline: 40009750]
3. Alimoradi Z, Lotfi A, Lin CY, Griffiths MD, Pakpour AH. Estimation of behavioral addiction prevalence during COVID-19 pandemic: a systematic review and meta-analysis. Curr Addict Rep 2022;9(4):486-517. [doi: 10.1007/s40429-022-00435-6] [Medline: 36118286]
4. Olson JA, Sandra DA, Colucci É, et al. Smartphone addiction is increasing across the world: a meta-analysis of 24 countries. Comput Human Behav 2022 Apr;129:107138. [doi: 10.1016/j.chb.2021.107138]
5. Brand M, Antons S, Bőthe B, et al. Current advances in behavioral addictions: from fundamental research to clinical practice. Am J Psychiatry 2025 Feb 1;182(2):155-163. [doi: 10.1176/appi.ajp.20240092] [Medline: 39659159]
6. Amin MH, Alhadi HA, Mohamed AE, et al. Impact of smartphone addiction on health status, mental well-being, and sleep quality among medical students in Sudan. BMC Psychiatry 2024 Dec 31;24(1):965. [doi: 10.1186/s12888-024-06377-9] [Medline: 39741229]
7. Ye XL, Zhang W, Zhao FF. Depression and internet addiction among adolescents: a meta-analysis. Psychiatry Res 2023 Aug;326:115311. [doi: 10.1016/j.psychres.2023.115311] [Medline: 37348449]
8. Alimoradi Z, Lin CY, Broström A, et al. Internet addiction and sleep problems: a systematic review and meta-analysis. Sleep Med Rev 2019 Oct;47:51-61. [doi: 10.1016/j.smrv.2019.06.004] [Medline: 31336284]
9. Sun Y, Wang Z, Liu T. Association of internet addiction with psychiatric symptom levels and sleep disorders: a systematic review and meta-analysis. Front Psychol 2025 Apr 17;16:1573058. [doi: 10.3389/fpsyg.2025.1573058] [Medline: 40313890]
10. Zhang X, Zhang J, Zhang K, et al. Effects of different interventions on internet addiction: a meta-analysis of random controlled trials. J Affect Disord 2022 Sep 15;313:56-71. [doi: 10.1016/j.jad.2022.06.013] [Medline: 35760188]
11. Chen L, Ren Y, Zhu P, Yang Y, Zhu F. Meta-analysis on intervention effects of exercise on Chinese children and adolescents with mobile phone dependence. Front Psychol 2025 Apr 11;16:1487944. [doi: 10.3389/fpsyg.2025.1487944] [Medline: 40290548]
12. Danielsen PA, Mentzoni RA, Låg T. Treatment effects of therapeutic interventions for gaming disorder: a systematic review and meta-analysis. Addict Behav 2024 Feb;149:107887. [doi: 10.1016/j.addbeh.2023.107887] [Medline: 37826910]
13. Zhu Y, Chen H, Li J, Mei X, Wang W. Effects of different interventions on internet addiction: a systematic review and network meta-analysis. BMC Psychiatry 2023 Dec 8;23(1):921. [doi: 10.1186/s12888-023-05400-9] [Medline: 38066462]
14. Li Z, Xia X, Sun Q, Li Y. Exercise intervention to reduce mobile phone addiction in adolescents: a systematic review and meta-analysis of randomized controlled trials. Front Psychol 2023 Dec 13;14:1294116. [doi: 10.3389/fpsyg.2023.1294116] [Medline: 38192396]
15. Gates M, Gates A, Pieper D, et al. Reporting guideline for overviews of reviews of healthcare interventions: development of the PRIOR statement. BMJ 2022 Aug 9;378:e070849. [doi: 10.1136/bmj-2022-070849] [Medline: 35944924]
16. Lu P, Qiu J, Huang S, et al. Interventions for digital addiction: umbrella review of meta-analyses. J Med Internet Res 2025 Feb 11;27:e59656. [doi: 10.2196/59656] [Medline: 39933164]
17. Yang H, Guo H, Zhu Z, et al. Intervention of internet addiction and smartphone addiction: an umbrella review of systematic reviews and meta-analyses. Curr Addict Rep 2023;11(9):1-24. [doi: 10.1007/s40429-023-00536-w]
18. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
19. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017 Sep 21;358:j4008. [doi: 10.1136/bmj.j4008] [Medline: 28935701]

XSL•FO
RenderX

20. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011 Apr;64(4):383-394. [doi: 10.1016/j.jclinepi.2010.04.026] [Medline: 21195583]
21. Cumpston M, Li T, Page MJ, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane Handbook for Systematic Reviews of Interventions. Cochrane Database Syst Rev 2019 Oct 3;10(10):ED000142. [doi: 10.1002/14651858.ED000142] [Medline: 31643080]
22. Zhou Z, Wan Y, Li C, et al. Effectiveness of sports intervention: a meta-analysis of the effects of different interventions on adolescent internet addiction. J Affect Disord 2024 Nov 15;365:644-658. [doi: 10.1016/j.jad.2024.08.064] [Medline: 39147163]
23. Augner C, Vlasak T, Aichhorn W, Barth A. Tackling the “digital pandemic”: The effectiveness of psychological intervention strategies in problematic Internet and smartphone use-A meta-analysis. Aust N Z J Psychiatry 2022 Mar;56(3):219-229. [doi: 10.1177/00048674211042793] [Medline: 34477009]
24. Chun JS, Shim HS, Kim S. A meta-analysis of treatment interventions for internet addiction among Korean adolescents. Cyberpsychol Behav Soc Netw 2017 Apr;20(4):225-231. [doi: 10.1089/cyber.2016.0188]
25. Gao W, Zhou H, Zou X, Li M, Lin X. The impact of exercise and exercise-based combined therapies on adolescent internet addiction: a systematic review and network meta-analysis. Curr Psychol 2025 Apr;44(8):7095-7110. [doi: 10.1007/s12144-025-07719-w]
26. Goslar M, Leibetseder M, Muench HM, Hofmann SG, Laireiter AR. Treatments for internet addiction, sex addiction and compulsive buying: a meta-analysis. J Behav Addict 2020 Apr 1;9(1):14-43. [doi: 10.1556/2006.2020.00005] [Medline: 32359229]
27. Jiang YS, Liu TH, Qin D, Wang ZP, He XY, Chen YN. Effects of non-pharmacological interventions on youth with internet addiction: a systematic review and meta-analysis of randomized controlled trials. Front Psychiatry 2023;14:1327200. [doi: 10.3389/fpsyt.2023.1327200] [Medline: 38274427]
28. Tian JJ, He XY, Guo Z. Optimal non-pharmacological interventions for reducing problematic internet use in youth: a systematic review and Bayesian network meta-analysis. Behav Sci (Basel) 2025 Jan 20;15(1):98. [doi: 10.3390/bs15010098] [Medline: 39851902]
29. Kim J, Lee S, Lee D, et al. Psychological treatments for excessive gaming: a systematic review and meta-analysis. Sci Rep 2022;12(1). [doi: 10.1038/s41598-022-24523-9]
30. Kim S, Noh D. The current status of psychological intervention research for internet addiction and internet gaming disorder. Issues Ment Health Nurs 2019 Apr;40(4):335-341. [doi: 10.1080/01612840.2018.1534910] [Medline: 30742546]
31. Kurnaz MF, Koçtürk N. Are CBT-based interventions effective for pathologic technology use? A meta-analysis of experimental studies. Curr Psychol 2024 Mar;43(12):11127-11138. [doi: 10.1007/s12144-023-05206-8]
32. Liu J, Nie J, Wang Y. Effects of group counseling programs, cognitive behavioral therapy, and sports intervention on internet addiction in East Asia: a systematic review and meta-analysis. IJERPH 2017;14(12):1470. [doi: 10.3390/ijerph14121470]
33. Liu S, Xiao T, Yang L, Loprinzi PD. Exercise as an alternative approach for treating smartphone addiction: a systematic review and meta-analysis of random controlled trials. IJERPH 2019;16(20):3912. [doi: 10.3390/ijerph16203912]
34. Lo CKM, Chan KL, Yu L, Chui WWH, Ip P. Long-term effects of psychosocial interventions on internet-related disorders: a meta-analysis. Comput Human Behav 2023 Jan;138:107465. [doi: 10.1016/j.chb.2022.107465]
35. Malinauskas R, Malinauskiene V. A meta-analysis of psychological interventions for Internet/smartphone addiction among adolescents. J Behav Addict 2019 Dec 1;8(4):613-624. [doi: 10.1556/2006.8.2019.72] [Medline: 31891316]
36. Ock CM, Lee HS, Chae J, Kim H. Effectiveness of non-pharmacological interventions on gaming disorder: a systematic review and meta-analysis. Psychiatry Investig 2025 May;22(5):490-503. [doi: 10.30773/pi.2024.0358] [Medline: 40404379]
37. Reangsing C, Wongchan W, Trakooltorwong P, Thaibandit J, Oerther S. Effects of cognitive behavioral therapy (CBT) on addictive symptoms in individuals with internet gaming disorders: a systematic review and meta-analysis. Psychiatry Res 2025 Jun;348:116425. [doi: 10.1016/j.psychres.2025.116425] [Medline: 40054044]
38. Kwon MS, Han SJ, Yeun YR. Effectiveness of smartphone addiction intervention for university students: a systematic review and meta analysis. J Korea Converg Soc 2020;11(5):399-412. [doi: 10.15207/JKCS.2020.11.5.399]
39. Stevens MWR, King DL, Dorstyn D, Delfabbro PH. Cognitive–behavioral therapy for internet gaming disorder: a systematic review and meta‐analysis. Clin Psychology and Psychoth 2019 Mar;26(2):191-203. [doi: 10.1002/cpp.2341]
40. Vicente-Escudero JL. Paradox or efficacy. Meta-analysis of internet-based psychological interventions to reduce internet addiction, anxiety and depression in internet-addicted people. Curr Psychol 2025 Jun;44(12):12807-12819. [doi: 10.1007/s12144-025-08096-0]
41. Wang H, Li X, Lok GKI, et al. Family-based therapy for internet addiction among adolescents and young adults: a meta-analysis. J Behav Addict 2024 Jun 26;13(2):295-312. [doi: 10.1556/2006.2024.00015] [Medline: 38635339]
42. Yan Y, Qin X, Liu L, Zhang W, Li B. Effects of exercise interventions on Internet addiction among college students: a systematic review and meta-analysis of randomized controlled trials. Addict Behav 2025 Jan;160:108159. [doi: 10.1016/j.addbeh.2024.108159] [Medline: 39303655]

XSL•FO
RenderX

43. Yeun YR, Han SJ. Effects of psychosocial interventions for school-aged children's internet addiction, self-control and self-esteem: meta-analysis. Healthc Inform Res 2016 Jul;22(3):217-230. [doi: 10.4258/hir.2016.22.3.217] [Medline: 27525163]
44. Zhang K, Lu X, Zhang X, et al. Effects of psychological or exercise interventions on problematic mobile phone use: a systematic review and meta-analysis. Curr Addict Rep 2023;10(2):230-253. [doi: 10.1007/s40429-023-00471-w]
45. Zhang M, Meng SQ, Hasan AJ, et al. Network meta-analysis of the effectiveness of different interventions for internet addiction in college students. J Affect Disord 2024 Oct;363:26-38. [doi: 10.1016/j.jad.2024.07.032]
46. Zhang Y, Li G, Liu C, Chen H, Guo J, Shi Z. Mixed comparison of interventions for different exercise types on students with internet addiction: a network meta-analysis. Front Psychol 2023;14. [doi: 10.3389/fpsyg.2023.1111195]
47. Wang X, Zhang Y, Lin J, et al. Treatments of internet gaming disorder and comorbid mental disorders: a systematic review and meta-analysis. Comput Human Behav 2023 Dec;149:107947. [doi: 10.1016/j.chb.2023.107947]
48. Montag C, Wegmann E, Sariyska R, Demetrovics Z, Brand M. How to overcome taxonomical problems in the study of internet use disorders and what to do with "smartphone addiction"? J Behav Addict 2021 Jan 15;9(4):908-914. [doi: 10.1556/2006.8.2019.59] [Medline: 31668089]
49. Dong G, Potenza MN. A cognitive-behavioral model of internet gaming disorder: theoretical underpinnings and clinical implications. J Psychiatr Res 2014 Nov;58:7-11. [doi: 10.1016/j.jpsychires.2014.07.005] [Medline: 25062755]
50. Stockings E, Shakeshaft A, Farrell M. Community approaches for reducing alcohol-related harms: an overview of intervention strategies, efficacy, and considerations for future research. Curr Addict Rep 2018 Jun;5:274-286. [doi: 10.1007/s40429-018-0210-2]
51. Brand M, Wegmann E, Stark R, et al. The Interaction of Person-Affect-Cognition-Execution (I-PACE) model for addictive behaviors: update, generalization to addictive behaviors beyond internet-use disorders, and specification of the process character of addictive behaviors. Neurosci Biobehav Rev 2019 Sep;104:1-10. [doi: 10.1016/j.neubiorev.2019.06.032] [Medline: 31247240]
52. Patterson MS, Spadine MN, Graves Boswell T, et al. Exercise in the treatment of addiction: a systematic literature review. Health Educ Behav 2022 Apr 29:10901981221090155. [doi: 10.1177/10901981221090155] [Medline: 35484950]
53. Chen Y, Lu J, Wang L, Gao X. Effective interventions for gaming disorder: a systematic review of randomized control trials. Front Psychiatry 2023 Feb 6;14:1098922. [doi: 10.3389/fpsyt.2023.1098922] [Medline: 36815197]
54. Ioannidis JP. Integration of evidence from multiple meta-analyses: a primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. CMAJ 2009 Oct 13;181(8):488-493. [doi: 10.1503/cmaj.081086] [Medline: 19654195]
55. Michopoulos I, Furukawa TA, Noma H, et al. Different control conditions can produce different effect estimates in psychotherapy trials for depression. J Clin Epidemiol 2021 Apr;132:59-70. [doi: 10.1016/j.jclinepi.2020.12.012] [Medline: 33338564]
56. Laws KR, Pellegrini L, Reid JE, Drummond LM, Fineberg NA. The inflating impact of waiting-list controls on effect size estimate. Front Psychiatry 2022 Jun 22;13:877089. [doi: 10.3389/fpsyt.2022.877089] [Medline: 35815038]
57. Mohr DC, Ho J, Hart TL, et al. Control condition design and implementation features in controlled trials: a meta-analysis of trials evaluating psychotherapy for depression. Transl Behav Med 2014 Dec;4(4):407-423. [doi: 10.1007/s13142-014-0262-3] [Medline: 25584090]
58. Reed GM, First MB, Billieux J, et al. Emerging experience with selected new categories in the ICD-11: complex PTSD, prolonged grief disorder, gaming disorder, and compulsive sexual behaviour disorder. World Psychiatry 2022 Jun;21(2):189-213. [doi: 10.1002/wps.20960] [Medline: 35524599]
59. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926. [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]

Abbreviations

AMSTAR 2: A Measurement Tool to Assess Systematic Reviews 2
CBT: cognitive behavioral therapy
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
***ICD-11*:** *International Classification of Diseases, 11th Revision*
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SMD: standardized mean difference

XSL•FO
RenderX

Edited by A Schwartz, M Balcarras; submitted 01.Aug.2025; peer-reviewed by H Rajesh, H Thongdaeng, Z Zhang; revised version received 18.Jan.2026; accepted 04.Feb.2026; published 05.Mar.2026.

Please cite as:
Zhang M, Lan T, Song T, Wang X
Effectiveness of Interventions for Internet, Smartphone, and Gaming Addictions: Umbrella Review and Meta–Meta-Analysis
J Med Internet Res 2026;28:e81705
URL: https://www.jmir.org/2026/1/e81705
doi:10.2196/81705

© Minggang Zhang, Tong Lan, Tao Song, Xiaochun Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of Virtual Reality–Based Early Rehabilitation Strategies on Pain, Sleep, Anxiety, Balance, Cognition, and Limb Motor Function in Adult Intensive Care Unit Patients: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Fei Wu[1,2], MSN; Yunting Wu[1,2], MSN; Yana Xing[1,3], MSN; Weixin Cai[1*], MSN; Ran Zhang[1,3*], MSN

[1]Nursing Department, Beijing Tiantan Hospital, Capital Medical University, No. 119, South 4th Ring West Road, Fengtai District, Beijing, China, Beijing, Beijing, China

[2]School of Nursing, Capital Medical University, Beijing, Beijing, China

[3]Department of Critical Care Medicine, Beijing Tiantan Hospital, Capital Medical University, Beijing, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Ran Zhang, MSN
Nursing Department, Beijing Tiantan Hospital, Capital Medical University, No. 119, South 4th Ring West Road, Fengtai District, Beijing, China, Beijing, Beijing, China

Abstract

Background: Early rehabilitation is vital for functional recovery in critically ill patients. Virtual reality–based early rehabilitation intervention (VR-ERI) is an emerging strategy, but evidence on its feasibility, safety, and efficacy remains inconsistent and unsynthesized.

Objective: We synthesized evidence from randomized controlled trials (RCTs) on the feasibility and safety of VR-ERI in adult critically ill patients and evaluated its effects on functional outcomes during intensive care unit (ICU) stay and at short-term follow-up (≤3 months post ICU).

Methods: Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines, we searched 10 databases (eg, PubMed, Web of Science, Cochrane Library, and Embase) from inception to October 5, 2025, for Chinese and English publications. We included RCTs comparing VR-ERI with control measures initiated early (within 72 h of ICU admission) in adults. FW and YW independently screened studies and extracted data; YW and YX independently conducted the revised Cochrane Risk of Bias tool assessment. Data were synthesized narratively or via meta-analysis in R Studio using a random-effects model (Hartung-Knapp-Sidik-Jonkman adjustment). Effects were expressed as standardized mean differences (SMD; with Hedges g correction) with 95% CIs and 95% prediction intervals (95% PIs). Subgroup and sensitivity analyses were performed to explore heterogeneity and assess robustness. The Grading of Recommendations Assessment, Development, and Evaluation framework was used to assess the quality of evidence.

Results: Sixteen RCTs (published 2020 - 2025) involving 1356 patients were included. Bias assessment found 1 study at low risk, 5 with some concerns, and 10 at high risk. Meta-analysis suggested potential trends for VR-ERI in improving ICU anxiety (SMD −0.86, 95% CI −1.85 to 0.13, 95% PI −3.75 to 2.03; very low certainty) and subjective sleep quality (SMD 3.36, 95% CI 0.77-5.94; very low certainty), with a more pronounced effect in the Richards-Campbell Sleep Questionnaire–assessed subgroup (SMD 5.12, 95% CI 0.54-9.71). At follow-up, VR-ERI showed trends toward improved balance (Berg Balance Scale: SMD 0.97, 95% CI 0.74-1.20, 95% PI 0.37-1.58; moderate certainty), limb motor function (Fugl-Meyer: SMD 1.40, 95% CI −0.23 to 3.02; low certainty), and cognitive function (SMD 0.78, 95% CI 0.16-1.39; low certainty). No significant differences were found for objective sleep measures or ICU pain (low to very low certainty). No serious adverse events were reported; only a few studies mentioned mild reactions, such as dizziness, nausea, and fatigue.

Conclusion: This review indicates VR-ERI's potential to improve anxiety, subjective sleep, balance, cognition, and motor function in early ICU rehabilitation, while its effects on pain and objective sleep remain unclear and safety protocols need refinement. Given the high risk of bias, substantial heterogeneity, and imprecision, the overall certainty of evidence is low. Thus, VR-ERI may serve as a nonpharmacological adjunct, but its clinical translation requires consideration of cost and patient suitability, and more rigorous research is needed.

Trial Registration: PROSPERO CRD420251114439; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251114439

(J Med Internet Res 2026;28:e81865) doi:10.2196/81865

XSL•FO
RenderX

KEYWORDS

virtual reality; intensive care unit; early rehabilitation; randomized controlled trial; systematic review; meta-analysis

Introduction

Background

Patients in the intensive care unit (ICU) often experience a series of physical, cognitive, and psychological impairments after discharge due to factors such as physiological dysfunction, surgical trauma, pain, and reduced social interaction [1]. In 2010, the Society of Critical Care Medicine first introduced the term "post-intensive care syndrome (PICS)" to define this phenomenon [2,3]. The manifestations commonly include weakness, pain, anxiety, sleep disturbances, and cognitive dysfunction, which can persist for months or even years. Studies indicate that 47.9% of patients still exhibit cognitive impairment 3 - 6 months after transfer from the ICU [4]; 68% (355/523) report persistent pain after discharge, with pain intensity remaining at a moderate level even 1 year later [5]. Additionally, 55% of patients experience sleep disturbances within the first month after discharge [6]. Furthermore, 12.3% (98,530/799,645) of critically ill patients are diagnosed with depression or anxiety after discharge, and these patients demonstrate higher long-term mortality rates [7]. PICS severely compromises patients' quality of life, reduces functional independence, and imposes a significant care burden on families and society [8].

To reduce long-term impairments after critical illness, the Society of Critical Care Medicine recommends that health care professionals assess patients and coordinate care strategies for injury management beginning at ICU admission [9]. As clinical treatment advances, early functional recovery has become a key focus for preventing PICS [10]. Early rehabilitation is generally initiated within 72 hours of ICU admission, following achievement of hemodynamic and respiratory stability (typically within 24 - 48 h) [11,12]. It commonly includes early physical activity [13,14], neurocognitive stimulation [15,16], psychological intervention [17], and multimodal programs [18]. Existing meta-analyses confirm that routine early rehabilitation started within 72 hours may improve physical and cognitive function, help prevent PICS without increasing adverse events [19], and reduce the incidence of ICU-acquired weakness [20]. However, no differences have been found in cognitive-related delirium-free days or Hospital Anxiety and Depression Scale scores. Another meta-analysis on nurse-led early rehabilitation showed it significantly shortens ICU length of stay, although effects on functional outcomes, such as mobility, muscle strength, and mortality, were not significant [21]. These inconsistencies may relate to factors including the time-intensive nature of current measures, insufficient standardization in implementation, and low patient participation and adherence.

Virtual reality (VR), a term first introduced by computer scientist Jaron Lanier in 1989, refers to a simulation technology providing users with an immersive environmental experience [22]. VR technology integrates computer, electronic information, and simulation technologies to enable users to interact with the VR system and experience an immersive, tactile, visual, and multisensory environment via VR helmets, all-in-one devices, and other gadgets [23]. The system is known for its novelty, interactivity, and immersive qualities [24]. VR technology–based rehabilitation techniques have been widely adopted in clinical rehabilitation. There is evidence suggesting that these techniques can effectively improve cognitive function in patients with stroke [25], reduce pain levels in patients with chronic low back pain and neck pain [26,27], and alleviate anxiety and depression [28,29].

Virtual reality–based early rehabilitation intervention (VR-ERI) is an emerging ICU recovery strategy that is gaining attention and has shown preliminary application results. Using VR technology, it provides ICU patients with specific virtual environments, such as natural landscapes [30], meditation music [31], or virtual tasks (such as memory games [32]), and cognitive exercises like "catching coconuts" and "avoiding obstacles" [33]. This aims to enhance rehabilitation motivation, participation, and physiological and psychological function. Its interactive, immersive, and feedback-driven design makes VR-ERI a promising tool for personalized early rehabilitation in critically ill patients. However, current research findings are inconsistent. Some studies report that VR-ERI effectively reduces anxiety [32], while others show no significant difference [34,35]. Such discrepancies may affect the certainty and consistency of clinical decision-making, potentially hindering the technology's wider adoption in ICU early rehabilitation.

To our knowledge, researchers [36] have explored the interventional effects of VR therapy on patients' PICS through a meta-analysis, but its included population did not specifically focus on early rehabilitation strategies during the ICU period. This analysis pooled data from patients receiving VR therapy both within the ICU and during follow-up after discharge. Due to differences between these 2 populations in aspects such as intervention timing and environment, combining them in the analysis may introduce clinical heterogeneity, thereby weakening the credibility of the results. Another meta-analysis indicated that compared with traditional rehabilitation, VR-based early motor rehabilitation can improve balance, functional walking, and upper and lower limb motor function in critically ill patients [37]. However, this review only assessed motor function and did not comprehensively examine the effects on other early rehabilitation outcome indicators.

Objectives

This systematic review and meta-analysis aim to (1) characterize the types and content of VR-ERI activities and analyze the feasibility and safety of VR technology implementation in critically ill patients; (2) systematically evaluate the effectiveness of VR-ERI on physiological, psychological, and cognitive outcomes during ICU treatment and at short-term follow-up (within 3 months post-ICU discharge); and (3) provide recommendations for future research.

XSL•FO
RenderX

Methods

Study Protocol and Registration

The review protocol was registered in PROSPERO. This systematic review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement [38] to ensure transparency (Checklists 1-3).

Deviations From the PROSPERO Protocol

During implementation, the search language was limited to Chinese and English to ensure assessment accuracy, and the search cutoff date was extended to October 5, 2025, to incorporate the latest evidence. Regarding statistical analysis, due to heterogeneity among the included studies, all analyses adopted the more conservative random-effects model. To enhance statistical robustness, the analysis was switched to the R language (version 4.4.1; R Core Team), and the Hartung-Knapp-Sidik-Jonkman method was applied to correct random-effects outcomes. This method provides more conservative CIs when the number of studies is small, reducing the risk of false-positive results in the presence of high heterogeneity. Apart from the above adjustments, all other procedures remained consistent with the original protocol.

Data Sources and Search Strategy

Two independent researchers (FW and YW) conducted literature searches across 6 English-language and 4 Chinese-language databases. MEDLINE and Embase were searched simultaneously via the Ovid platform. The other English databases were searched on their respective platforms—PubMed (via National Center for Biotechnology Information), Web of Science (via Clarivate Analytics), the Cochrane Library, and PsycINFO (via ProQuest). The Chinese databases searched were China National Knowledge Infrastructure (CNKI), WanFang Data, VIP Database (VIP), and the SinoMed Service System (SinoMed). The search strategy combined Medical Subject Headings with free-text terms using Boolean operators. The search was performed in 2 stages. An initial search covered records from the inception of each database until May 11, 2025, and to incorporate the most recent evidence, the full search strategy was rerun on October 5, 2025, extending the cutoff date accordingly. All searches were limited to Chinese and English publications to align with the team's language capabilities and ensure feasibility.

All search strategies were independently designed based on the research question and were peer-reviewed and refined by experts in evidence-based medicine and critical care medicine in accordance with the PRISMA-S (Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension) checklist [39]. No published search filters were used, nor were search strategies from other systematic reviews adapted or reused. To cover the literature as comprehensively as possible, we also used the snowballing method to trace references and citations and attempted to obtain full-text articles by contacting librarians or the original authors. Furthermore, as this study focused on published efficacy evidence, we did not specifically search clinical trial registries or gray literature (including conference proceedings). The complete search strategies and procedures are detailed in Multimedia Appendix 1.

Study Selection

The rationale for formulating inclusion and exclusion criteria was based on the Population, Intervention, Comparison, Outcome, and Study Design (PICOS) framework, (1) population: adults aged ≥18 years receiving treatment during their ICU stay; (2) intervention: VR-ERI; (3) control: no treatment, usual care measures, or non–VR-based interventions; (4) outcomes: studies reporting at least 1 physical, psychological, or cognitive function outcome (such as cognition, pain, delirium, anxiety, depression, and sleep quality) or feasibility and safety outcomes during ICU treatment or after ICU transfer; and (5) study design: randomized controlled trials (RCTs) published in Chinese or English. The exclusion criteria were (1) studies that did not report original outcome measures and for which relevant outcome data could not be obtained after contacting the authors; (2) duplicate publications; and (3) case reports, letters, conference abstracts, and review articles.

Screening

In total, 2 independent researchers (FW and YW) conducted the literature screening and data extraction, followed by cross-checking. Any uncertainties were resolved by WC. First, duplicate records were automatically identified and removed using the built-in function of the EndNote X9 (Clarivate) reference management software, based on fields such as title, author, publication year, and journal name. Subsequently, the 2 researchers (FW and YW) independently examined the titles, authors, and other key information of the remaining records to further exclude any duplicates not detected by the software. Next, titles and abstracts were reviewed to exclude obviously irrelevant literature. Finally, full texts were retrieved and assessed against the eligibility criteria to determine the final set of included studies.

Data Extraction

A total of 2 independent researchers (FW and YW) performed data extraction using a uniformly developed standardized form without using automated tools. The data that were systematically extracted were (1) general information, including first author, country, and publication year; (2) participant characteristics, including disease type, sample size, and baseline demographic and clinical characteristics; and (3) intervention characteristics, including specific type of VR, content of the VR-based intervention, content of the control group intervention, intervention duration, intervention frequency, outcome measures, measurement tools, time points of data collection and raw data (such as mean, SD, median, and IQR), for outcomes reported at baseline, post intervention, or follow-up for both groups.

To manage multiple data points objectively and consistently, we predefined selection criteria following the Cochrane Handbook [40]. The core principle was to prioritize data that best aligned with the review's primary objective, had the highest validity, or greatest clinical representativeness. For multiple measures, we selected the tool that best matched the outcome definition or had the highest validity; for multiple intervention

XSL•FO
RenderX

groups, we included the primary intervention group, the most clinically representative, or the largest in sample size; for multiple control groups, we chose the "most representative usual care" group. Where data were missing, unclear, or incomplete, we contacted corresponding authors by email for clarification or missing data. If no response was received after 2 attempts, we used recommended formulas to estimate means and SDs from available data (eg, based on sample size, median, and range) [41].

Risk-of-Bias Assessment

The second author (YW) and the third author (YX) independently assessed the risk of bias of the included studies using the revised Cochrane Risk of Bias tool (RoB 2) for randomized trials [42]. The RoB 2 tool assesses five domains for randomized trials, which are (1) bias arising from the randomization process, (2) bias due to deviations from the intended interventions, (3) bias due to missing outcome data, (4) bias in the measurement of the outcome, and (5) bias in the selection of the reported result. Each domain is judged against a series of signaling questions. The overall risk of bias for a study is categorized as "low," "some concerns," or "high." A study is judged to have a low overall risk of bias if all domains are rated as "low." An overall rating of "some concerns" is given if at least 1 domain is rated as having "some concerns" and no domain is rated as "high." An overall "high" risk of bias is judged if any single domain is rated as "high" or if multiple domains are rated as having "some concerns."

Data Synthesis and Analysis

To ensure transparency in the data synthesis process, we followed a systematic decision-making procedure. First, all studies had to meet the predefined PICOS criteria and provide the raw data for the relevant outcome measures. Second, we established a rule that each study would contribute only a single data point for a given outcome. If a study contained multiple measurements or multiple group comparisons, a single dataset was selected for pooling according to the prespecified criteria, aiming to minimize heterogeneity as much as possible.

All analyses were performed using the R language (version 4.4.1) within the R Studio integrated development environment (version 2023.12). The metafor package was used for effect size pooling and forest plot generation [43], providing visual representations of each study's effect size with its 95% CI, along with the pooled overall effect size. All included outcome measures were continuous variables. We conducted a meta-analysis using change scores (from baseline to postintervention or follow-up) in means and SDs for both groups. For studies reporting only pre- and postintervention means and SDs, the mean and SD of change scores were calculated as $Mean_{change}=Mean_{post}-Mean_{pre}$ and $SD_{change}=\sqrt{(SD_{pre}^2+SD_{post}^2-2\times r\times SD_{pre}\times SD_{post})}$, where r represents the correlation coefficient between pre- and postintervention values. When correlations were not reported in the original publications, they were calculated according to the Cochrane Handbook for Systematic Reviews of Interventions. To harmonize different assessment tools and ensure consistent direction of results, we reversed the scoring for subjective sleep quality assessed by the Pittsburgh Sleep Quality Index and cognitive scores assessed by the Perceived Deficits Questionnaire for Depression, following Cochrane Handbook guidelines [40]. Consequently, higher scores in the pooled results indicate better sleep quality and cognitive function in patients. Additionally, we categorized outcome measures by assessment time points, specifically divided into (1) during ICU treatment and (2) short-term follow-up (2 wk to 3 mo after ICU discharge).

Given the variability in assessment tools and the presence of clinical and methodological heterogeneity among the included studies, the standardized mean difference (SMD) was adopted as the effect size measure for meta-analysis outcomes. The adjusted Hedges g was used to estimate the effect sizes [44,45]. Absolute SMD values of 0.2, 0.5, and 0.8 were interpreted as small, medium, and large effect sizes, respectively [46]. A random-effects model was applied to derive more conservative effect estimates. To enhance the robustness of statistical inference, particularly in reducing the risk of false-positive results when the number of studies was limited, the Hartung-Knapp-Sidik-Jonkman adjustment was used during model fitting to estimate the pooled effect size and its 95% CI [47].

To quantify between-study heterogeneity and enhance the generalizability of the results in real-world settings, we used a comprehensive assessment strategy. First, absolute heterogeneity was quantified using τ^2 (between-study variance) and its square root τ (between-study SD) [48]. τ^2 provides an absolute measure of the variation between studies, while τ, being on the same scale as the effect size, facilitates clinical interpretation. Second, the Cochran Q test was used for the hypothesis testing of effect size homogeneity, supplemented by the I^2 statistic to quantify the relative extent of heterogeneity [49], with thresholds of 25%, 50%, and 75% used to define low, moderate, and high levels, respectively [50]. Finally, for outcome measures involving at least 3 studies, 95% prediction intervals (PIs) were calculated [51] to estimate the potential range of the true effect size in future similar interventions, thereby providing more clinically informative guidance for practice.

Subgroup analyses were performed to explore potential sources of heterogeneity. A leave-one-out sensitivity analysis was conducted to assess the robustness of the pooled results. For outcome measures with fewer than 2 included studies, descriptive analyses were performed. Additionally, to assess the potential for small-study effects (such as publication bias), we performed several checks for the anxiety outcome, which had the largest number of included studies (n=8). These included a visual inspection of funnel plot symmetry, the Egger test (regression test), and the Begg test (rank correlation test). Furthermore, the trim-and-fill method was used to estimate the potential number of missing studies and to compute an adjusted effect size, thereby evaluating the potential impact of small-study effects on the pooled result. The threshold for statistical significance was set at $P\leq.05$.

Evidence Quality

To evaluate the reliability of the evidence, we used the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) framework [52] to rate the certainty of evidence for

XSL•FO
RenderX

each outcome. In addition, 2 independent raters (FW and YW) assessed the certainty of evidence using the official GRADE pro GDT software (Evidence Prime Inc), with any disagreements resolved by the corresponding author (RZ). The evidence was rated as very low, low, moderate, or high. Footnotes were provided to explain the reasons for any downgrading of the evidence, which included considerations of risk of bias, inconsistency, indirectness, imprecision, and publication bias.

Ethical Considerations

This study used literature data and did not require approval from an ethics review committee or patient consent.

Results

Search Results

A total of 3737 studies were identified from 10 databases. After removing 243 duplicate articles, the titles and abstracts of the studies were initially screened, and 58 full-text manuscripts relevant to this study were identified. Following full-text reading and quality assessment, 16 RCTs [16,17,30,32-35,53-61] were ultimately included for analysis. The specific search and inclusion process is illustrated in Figure 1, and the excluded articles after rescreening, along with the reasons, are detailed in Multimedia Appendix 2.

Figure 1. Flowchart of literature screening. CNKI: China National Knowledge Infrastructure; SinoMed: SinoMed Service System; VIP: VIP Database.

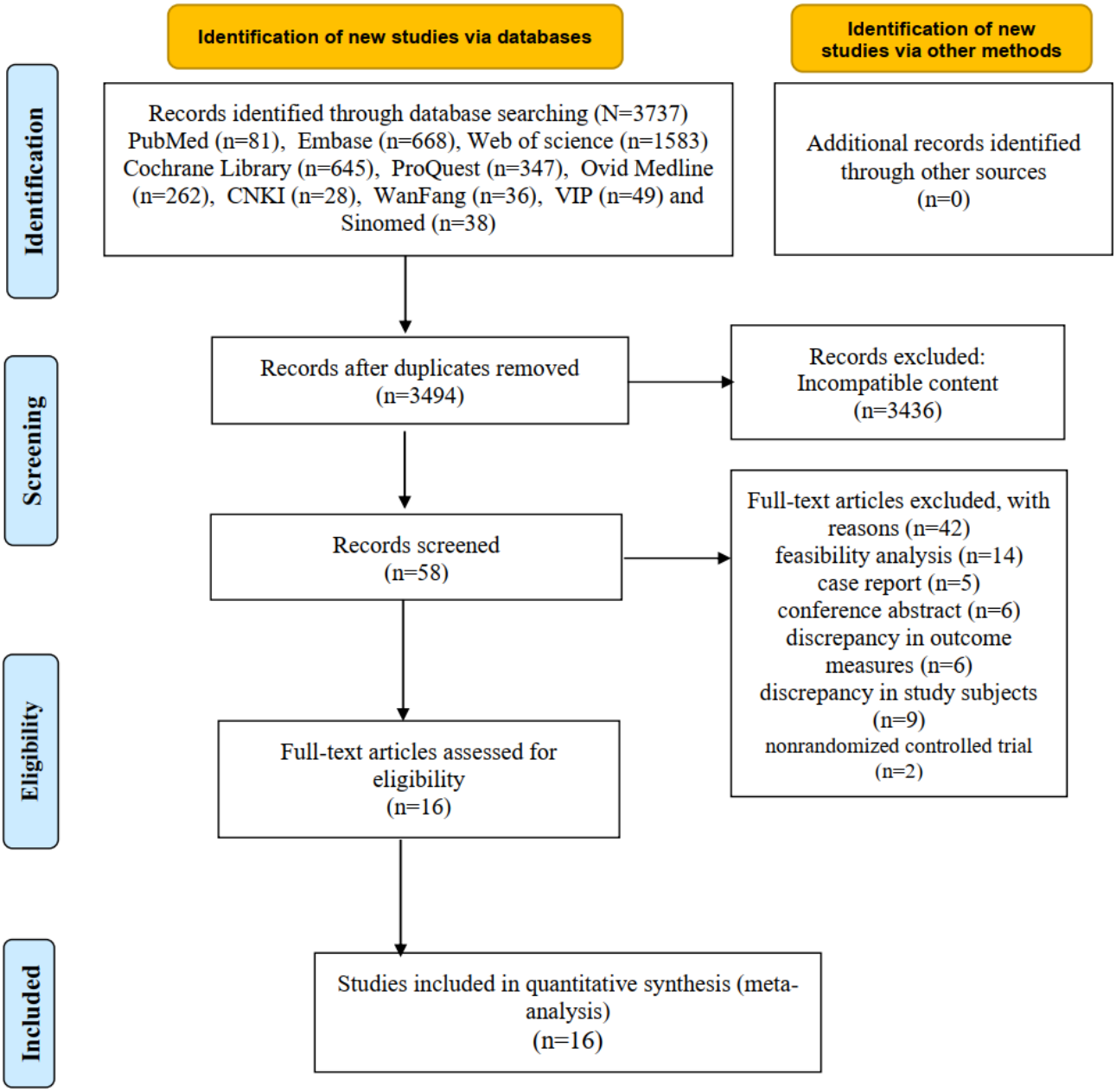

Characteristics of the Studies

Multimedia Appendix 3 summarizes the basic characteristics of the included studies. The 16 studies (published between 2020 and 2025) comprised a total of 1356 participants, with sample sizes ranging from 42 to 180. The studies originated from Brazil [53], Korea [30,54], France [17,35], Spain [16], Turkey [34], and China [32,33,55-61]. Patient conditions included cardiovascular diseases [30,35,53,55,60], postcraniotomy [33], postliver transplantation [61], postlaparoscopic sleeve

XSL•FO
RenderX

gastrectomy [34], conventional oncologic surgery [32], severe respiratory diseases [58], and postmultiple fracture surgery [56], among others. The mean age of patients was generally between 60 and 70 years. Interventions were initiated within 24 hours after ICU admission in 3 studies [35,53,59], within 6 hours in 2 studies [32,34], and on the day of ICU admission once the patient's condition was stabilized in the remaining 11 studies [16,17,30,33,54-58,60,61].

One study [16] used nonimmersive VR technology, while 15 studies [17,30,32-35,53-61] used fully immersive VR. The intervention directions included (1) early cognitive rehabilitation (training reaction and executive function through games or tasks) [16,33,55,57], (2) early motor rehabilitation (engaging in virtual activities, such as fruit cutting or cycling combined with motion capture) [53,56,58,59,61], (3) early delirium prevention (providing emotional support and cognitive training for multidimensional prevention) [32], (4) VR relaxation therapy (conducting physical and mental relaxation in natural scenes) [17,34,35], (5) VR meditation (improving sleep and mood by combining mindfulness) [30,54], and (6) VR preoperative visit (structured introduction to the ICU environment, treatment process, and rehabilitation) [60]. The intervention protocols were categorized into (1) single-session short-term intervention (15 - 30 min at key clinical time points) [17,30,34,35,60], (2) high-frequency short-term intervention during ICU stay (approximately 20 min per session, 1 - 3 times daily, lasting 3 - 7 d) [16,32,54,55,57-59], (3) medium- to long-term structured intervention (30 - 40 min per session, 5 d per wk, lasting 2 wk to 3 mo) [33,56,61], and (4) flexible individualized protocol (adjusted based on patient tolerance) [53]. The studies involved a total of 24 outcome measures, such as cognitive function, delirium, anxiety, depression, pain, sleep, and quality of life, although the assessment tools and observation windows varied across studies.

Risk of Bias

The overall risk of bias for the studies included in this review was high, with summary results presented in Figures 2 and 3, and detailed assessment results available in Multimedia Appendix 4. Furthermore, 1 study [58] was rated as "low," 5 studies [17,32,34,54,55] raised "some concerns," and 10 [16,30,33,35,53,56,57,59-61] were rated as "high." In domain 1 (randomization process), 7 studies [33-35,53,56,60,61] did not describe allocation concealment, and 1 study [32] described neither the method for generating the random sequence nor allocation concealment; all were rated as "some concerns." For domain 2 (deviations from intended interventions), 4 studies [16,32,55,57] were rated as "some concerns" due to lack of blinding of patients and implementers; 1 study [59] was rated as "high" because it lacked blinding and did not adhere to the intention-to-treat analysis principle. In domain 3 (missing outcome data), 2 studies [16,59] were rated as "high" due to a high rate of loss to follow-up and because the reasons for missing data were potentially related to health status. For domain 4 (measurement of the outcome), 11 studies [16,17,30,32,33,35,53,56,57,60,61] had issues with the lack of blinding of outcome assessors. Among these, 9 studies [16,30,33,35,53,56,57,60,61] were rated as "high" because they measured subjective patient-reported outcomes (eg, pain and anxiety), and 2 studies [17,32] were rated as "some concerns" as their outcome measures were relatively objective. In domain 5 (selection of the reported result), transparency was generally insufficient. Moreover, 4 studies [16,17,35,55] mentioned trial registration but lacked verifiable prespecified analysis descriptions for the reported results, while the remaining 11 studies [30,32-34,53,54,56,57,59-61] did not mention a preregistered protocol; consequently, all were rated as "some concerns."

Figure 2. Risk of bias graph for proportion of judgements per item in included studies.

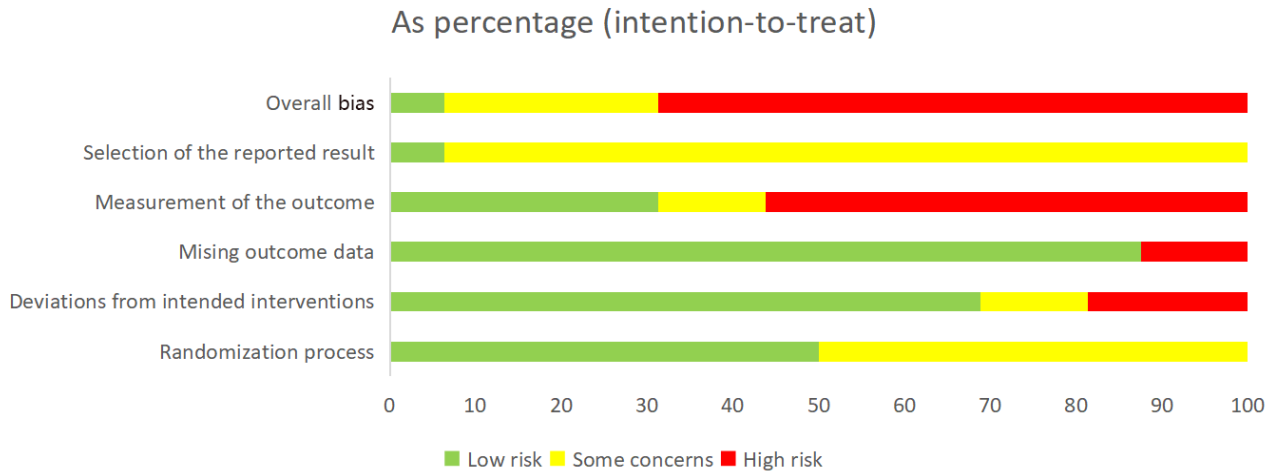

XSL•FO
RenderX

Figure 3. Risk of bias summary Items assessed in each included study [16,17,30,32-35,53-61].

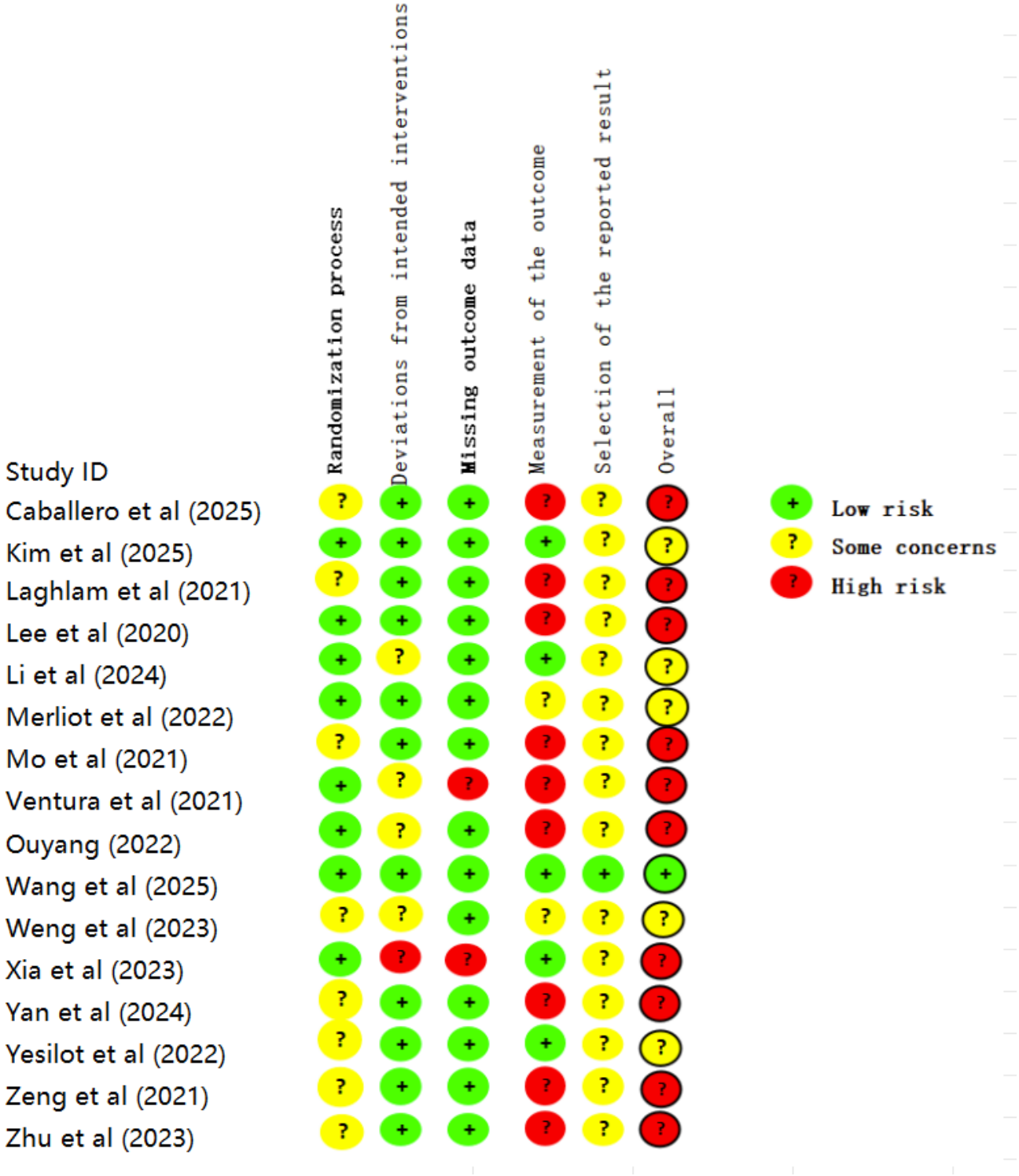

Clinical Safety and Feasibility

In total, 8 studies [16,17,35,53,55,57-59] reported on safety, 5 studies [16,55,57-59] confirmed the absence of serious adverse events but documented related reactions including dizziness [16,35,57], nausea [35,57], fatigue [16,17], anxiety [16,17], confusion [16,55], visual disturbance [17], headache or discomfort from wearing the headset [55], claustrophobia, agitated withdrawal, and dyspnea due to device displacement [17]. Furthermore, 1 study reported a significant drop in blood pressure during intervention in a patient [59], another reported a case of excessive drowsiness [16], and 1 noted that vital signs in all patients fluctuated within the normal range (±10%) during the intervention [58]. Regarding feasibility, 1 study reported a 78.3% patient recommendation rate and a 90% “good/excellent” experience rating [53]. Meanwhile, 2 studies reported that patients in the VR group demonstrated significantly better outcomes in terms of daily total exercise time, average daily out-of-bed activity time [58], and exercise adherence (with a high adherence rate of 71.74% vs 54.35% at 3 months) [56].

XSL•FO
RenderX

Results of the Meta-Analysis

Among the included trials, due to the limited number of studies, we conducted meta-analyses on outcome indicators with at least 2 studies. The specific analysis results are as follows.

Anxiety

A total of 8 studies (n=801) included anxiety as an outcome measure. According to the predefined data pooling criteria, for the study by Li et al [55], we only pooled the data assessed during ICU treatment using the State-Trait Anxiety Inventory scale. For the study by Merliot-Gailhoustet et al [17], we only pooled the data from the group receiving the DEEPSEN VR system (Saint-Didier-au-Mont-d'Or, France). All studies were analyzed using change-from-baseline values for the pooling.

The meta-analysis revealed that, in terms of the average effect, VR-ERI did not show a statistically significant effect on reducing patient anxiety (SMD −0.86, 95% CI −1.85 to 0.13). Given that higher scores on all assessment tools indicated greater anxiety, the negative point estimate suggests that VR-ERI may reduce anxiety. However, the 95% PI was wide and included the null value (−3.75 to 2.03), indicating considerable uncertainty about the intervention's effect in future settings, ranging from a substantial benefit to potential harm. Therefore, these findings require validation in further studies.

Extremely high heterogeneity was detected among the studies (I^2=95.4%, τ=1.1480, τ^2=1.3179; Cochran Q=150.82, df=7; P<.01; Figure 4). Exploratory subgroup analyses were performed based on VR intervention content, timing, patient population, and type of anxiety assessment tool, but no potential source of heterogeneity was identified. Sensitivity analysis (leave-one-out method) indicated that the pooled effect size was sensitive to the removal of individual studies, with a maximum change of 0.3478, exceeding the prespecified threshold of 0.1, suggesting that the results may not be robust. The study by Weng et al [32] had the most substantial influence on the overall effect estimate. After excluding this study, the pooled effect size remained statistically nonsignificant (SMD −0.51, 95% CI −1.17 to 0.15; 95% PI −2.31 to 1.28), and heterogeneity remained high (I^2=92.5%, τ=0.6806, τ^2=0.4632; Cochran Q=132.89, df=6; P<.01), further supporting the instability of the findings.

Figure 4. Forest plot of the meta-analysis for anxiety outcomes [16,17,32,34,35,55,57,60].

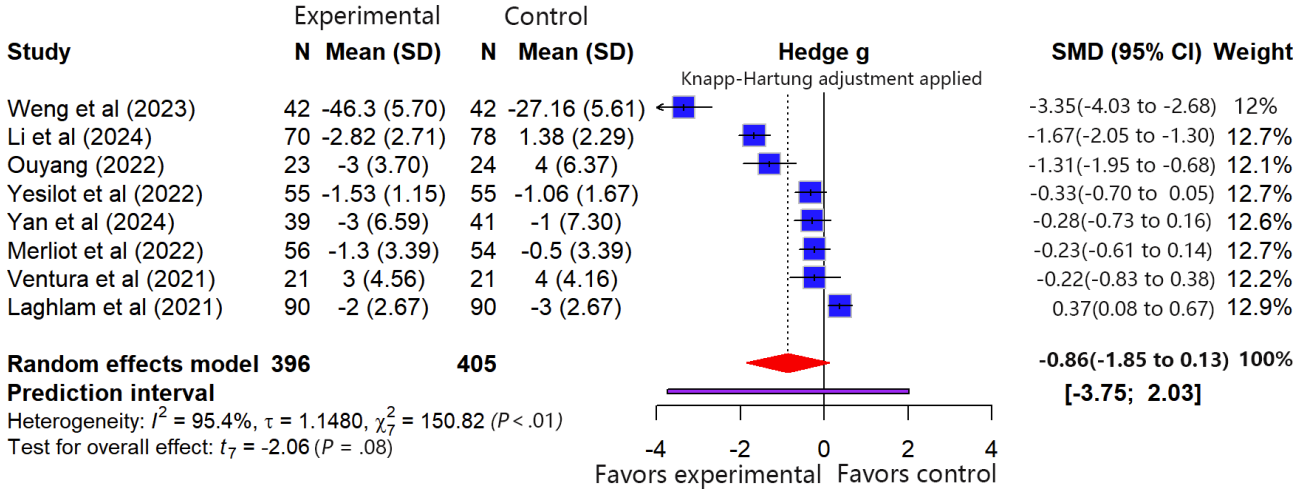

According to the GRADE framework, the certainty of evidence for anxiety was rated as "very low." The downgrading was based on (1) serious risk of bias in the included studies (4 studies [16,35,57,60] with high risk), (2) very high heterogeneity among studies, (3) imprecision of the pooled effect estimate (the 95% CI spanned both the line of null effect and the threshold for clinically meaningful benefit), and (4) potential publication bias. Although the pooled effect size suggested a potential moderate-to-large effect, an upgrade was not considered after weighing the above limitations (detailed assessment is available in Multimedia Appendix 5).

Subjective Sleep Quality (Scale Assessment)

In total, 6 studies [30,32,54,55,57,60] (n=500) assessed subjective sleep quality during ICU stay. In accordance with predefined pooling criteria, we extracted data from the second intervention day for Kim and Kang [54] and the Richards-Campbell Sleep Questionnaire (RCSQ)–measured sleep data only for Ouyang [57]. While all studies reported comparable baseline characteristics, specific data were provided by only 3 [32,55,60]. To maximize data utility, the meta-analysis was performed using postintervention outcomes.

Meta-analysis revealed considerable heterogeneity (I^2=97.5%, τ=2.4143, τ^2=5.8288; Cochran Q=198.44, df=5; P<.01), likely due to differences in populations, interventions, and assessments. Sensitivity analysis (leave-one-out method) confirmed the robustness of the results, as the maximum change in effect size after excluding any single study was 0.0347 (below the prespecified 0.1 threshold), with no specific source of heterogeneity identified. The pooled results (Figure 5) indicated a significant positive effect of VR-ERI on improving patients' subjective sleep quality compared with usual care (SMD 3.36, 95% CI 0.77-5.94). For consistent interpretation, Pittsburgh Sleep Quality Index scores were reversed so that higher scores uniformly indicated better sleep quality across all scales; thus, the positive SMD reflects an enhancing effect of VR-ERI, although more studies are needed to confirm the precision of this estimate.

XSL•FO
RenderX

Figure 5. Forest plot of the subgroup analysis for subjective sleep quality [30,32,54,55,57,60].

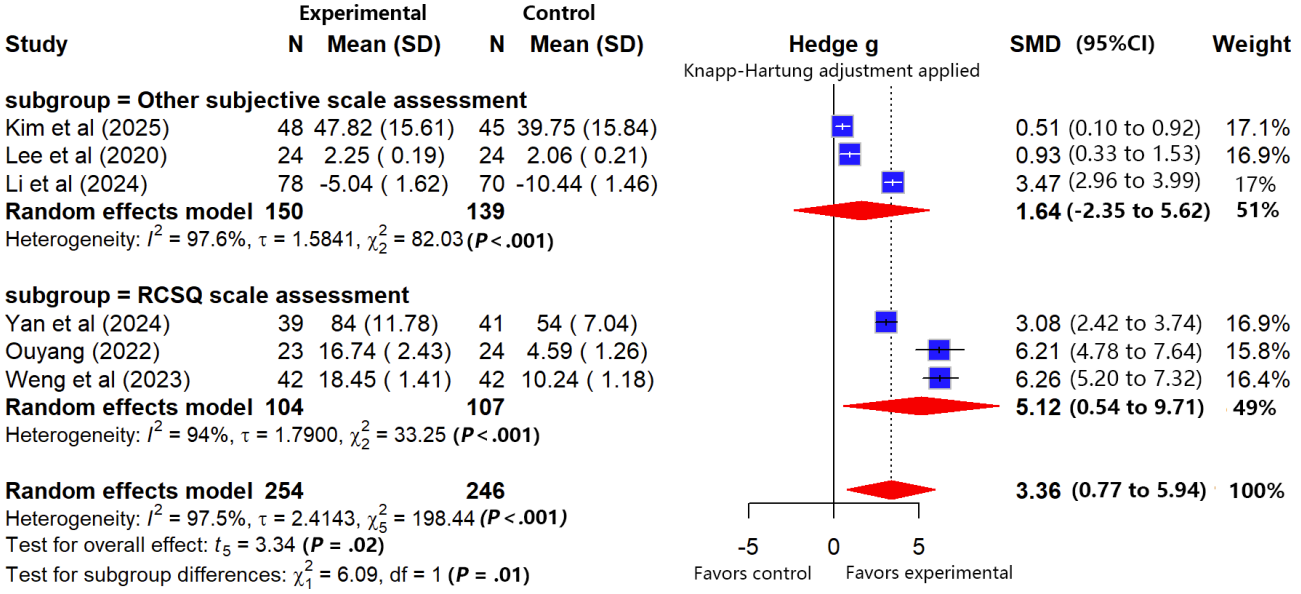

Subgroup analysis was conducted after unifying scale directions to assess how the inherent characteristics of different instruments might influence the pooled result. The analysis showed that VR-ERI did not significantly improve sleep when assessed with generic scales (SMD 1.64, 95% CI −2.35 to 5.61). In contrast, a positive effect was observed with the ICU-specific RCSQ (SMD 5.12, 95% CI 0.54 to 9.71). The between-subgroup difference was statistically significant (χ^2_1=6.1; *P*=.014), indicating that the type of assessment tool is a significant effect modifier, with the RCSQ potentially being more sensitive to VR-ERI–related sleep improvements in the ICU setting.

According to the GRADE approach, the certainty of evidence for subjective sleep quality during ICU treatment was rated as "very low." The reasons for downgrading included (1) a serious risk of bias in the included studies (3 studies [30,57,60] with high risk), (2) very high heterogeneity among studies, and (3) extreme imprecision of the pooled effect estimate (the 95% CI was extremely wide, spanning the range from moderate to very large effects). Considering all the downgrading factors above, an upgrade was not applied despite the large pooled effect size (for detailed assessment, refer to Multimedia Appendix 5).

Objective Sleep Quality (Smart Wristband Monitoring)

In total, 2 studies [30,54] (n=141) monitored objective sleep quality using smart wristbands. Key metrics included total sleep time (TST), defined as the actual duration of sleep from onset to awakening; wake after sleep onset (WASO), representing total wake time during the sleep period as an indicator of sleep fragmentation; and sleep efficiency, calculated as the percentage of TST relative to total time in bed, with higher values indicating better sleep quality. A pooled analysis was performed using change-from-baseline values, with specific results as follows.

TST

The meta-analysis indicated no heterogeneity among the studies (I^2=0%, τ=0, τ^2=0; Cochran Q=0, df=1; *P*=.95). Sensitivity analysis (leave-one-out method) indicated that the results were robust. The maximum change in the effect size after excluding any single study was 0.0375, which is below the prespecified threshold of 0.1. The meta-analysis (Figure 6) showed that, regarding the average effect, VR-ERI did not lead to a statistically significant improvement in patients' TST compared with usual care (SMD 0.08, 95% CI −0.07 to 0.22). The positive SMD suggests that the direction of the effect may favor VR-ERI, indicating an increase in TST. According to the GRADE approach, the certainty of evidence for objective TST during ICU treatment was rated as "moderate." The rating was downgraded due to a high risk of bias in 1 study [30]. No downgrading was applied for other domains, as there was no heterogeneity among studies and the pooled effect estimate was relatively precise (for detailed assessment, refer to Multimedia Appendix 5).

XSL•FO
RenderX

Figure 6. Forest plot of the pooled analysis for objective sleep quality [30,54].

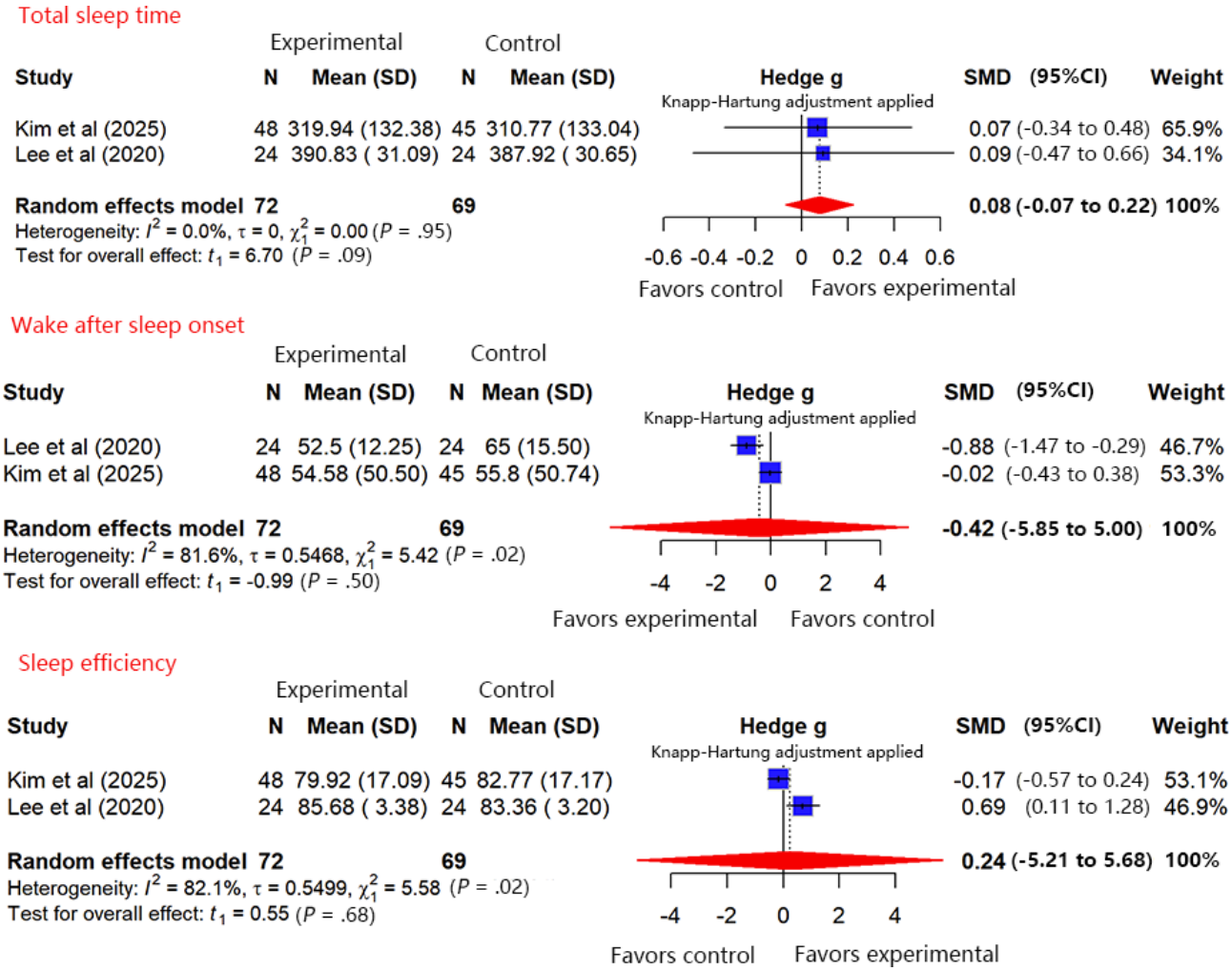

WASO

The meta-analysis indicated substantial heterogeneity among the studies (I^2=81.6%, τ=0.5468, τ²=0.2990; Cochran Q=5.42, df=1; *P*=.02). Sensitivity analysis showed that the results were robust (the maximum change in the effect size after excluding any single study was 0.0397, below the threshold of 0.1). The meta-analysis (Figure 6) showed that, regarding the average effect, VR-ERI did not lead to a statistically significant reduction in WASO compared with usual care (SMD −0.42, 95% CI −5.85 to 5.00). The negative SMD indicates that the direction of the effect points toward a reduction in WASO, signifying an improvement in sleep maintenance. According to the GRADE approach, the certainty of evidence for objective WASO during ICU treatment was rated as "very low." The rating was downgraded due to (1) a very limited number of included studies with risk of bias, (2) high heterogeneity among studies, and (3) extreme imprecision of the pooled effect estimate (the 95% CI was exceptionally wide; for detailed assessment, refer to Multimedia Appendix 5).

Sleep Efficiency

The meta-analysis indicated substantial heterogeneity among the studies (I^2=82.1%, τ=0.5499, τ²=0.3024; Cochran Q=5.58, df=1; *P*=.02). Sensitivity analysis indicated robust results (the maximum change in the effect size after excluding any single study was 0.0432, below the threshold of 0.1). Regarding the average effect, the meta-analysis (Figure 6) showed that VR-ERI did not lead to a statistically significant improvement in sleep efficiency compared with usual care (SMD 0.24, 95% CI −5.21 to 5.68). The positive SMD indicates that the direction of the effect points toward an improvement in sleep efficiency. According to the GRADE approach, the certainty of evidence for objective sleep efficiency during ICU treatment was rated as "very low." The rating was downgraded due to (1) risk of bias in the included studies, (2) substantial heterogeneity among studies, and (3) extreme imprecision of the pooled effect estimate (refer to Multimedia Appendix 5 for detailed assessment).

Pain

In total, 3 studies included pain level during ICU treatment as an outcome measure. According to the predefined data pooling criteria, for the study by Merliot-Gailhoustet et al [17], we only included the pain data from patients who received the DEEPSEN VR system. The pain outcome measure was analyzed using change-from-baseline values for the pooled analysis.

Initial meta-analysis revealed extremely high heterogeneity among the studies (I^2=95.1%, τ=0.7550, τ²=0.5700; Cochran Q=41.10, df=2; *P*<.01). Regarding the average effect, VR-ERI did not show a statistically significant difference in reducing patients' ICU-related pain compared with usual care (SMD

XSL•FO
RenderX

−0.13, 95% CI −2.06 to 1.79). Sensitivity analysis (leave-one-out method) indicated that the results may not be robust. The maximum change in the effect size after excluding a single study was 0.4062, which exceeds the prespecified threshold of 0.1. The study with the largest influence was by Laghlam et al [35]. Upon verification, this study was a noninferiority trial comparing VR therapy with another analgesic intervention. Pooling its control group with standard usual care controls could confound the pain outcome. Therefore, the study by Laghlam et al [35] was excluded, and the pooled result from the remaining 2 studies [17,34] (n=220) was used as the primary evidence for assessing the effect of VR-ERI on pain during ICU treatment. The meta-analysis of the 2 remaining studies showed low heterogeneity (I^2=23.7%, τ=0.1086, τ^2=0.0118; Cochran Q=1.31, df=1; *P*=.25; Figure 7). Regarding the average effect, VR-ERI did not demonstrate a statistically significant effect on reducing pain during ICU treatment compared with usual care (SMD −0.58, 95% CI −2.58 to 1.43). Subgroup analysis was not performed due to the limited number of studies.

Figure 7. Forest plot for pain outcomes during the intensive care unit stay [17,34].

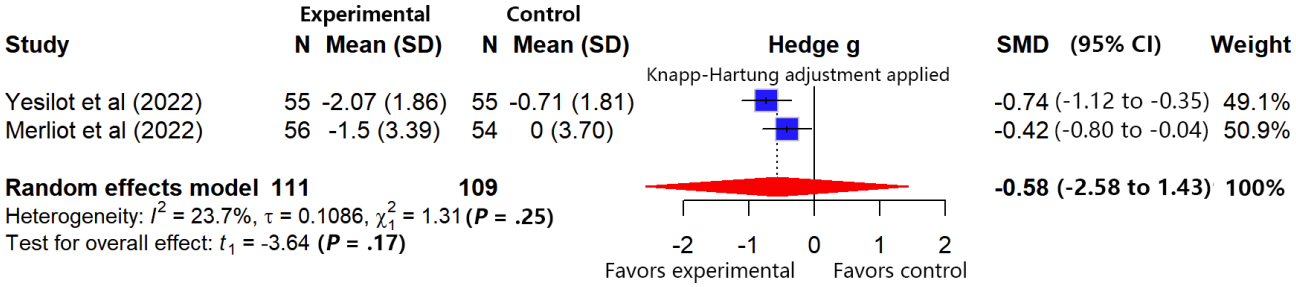

According to the GRADE criteria, the certainty of evidence for the outcome of pain during ICU treatment was rated as "low." The reasons for downgrading include (1) a high risk of bias in the included primary studies and (2) serious imprecision, as the 95% CI of the pooled effect is extremely wide and includes the line of null effect. As the heterogeneity among the remaining studies was low after the exclusion of 1 study based on clinical heterogeneity, no downgrading was applied for inconsistency (refer to Multimedia Appendix 5 for detailed assessment).

Balance

In total, 3 studies [33,56,61] (n=228) assessed patients' balance function using the Berg Balance Scale at short-term follow-up. We pooled change-from-baseline values for analysis. No heterogeneity was observed among the included studies (I^2=0%, τ=0, τ^2=0; Cochran Q=0.29, df=2; *P*=.86). The meta-analysis showed that, in terms of the average effect, VR-ERI improved patients' balance function during short-term follow-up after discharge compared with usual care (SMD 0.97, 95% CI 0.74 to 1.20). Furthermore, the 95% PI (0.37-1.58) lies entirely to the right of the line of null effect, suggesting that the positive effect of VR-ERI on balance function is likely to remain at a clinically meaningful level in most future, similar studies (Figure 8). Sensitivity analysis (leave-one-out method) indicated that the results were robust. The maximum change in the effect size after excluding any single study was 0.0461, which is below the prespecified threshold of 0.1. Subgroup analysis was not performed due to the limited number of studies.

Figure 8. Forest plot for balance outcomes [33,56,61].

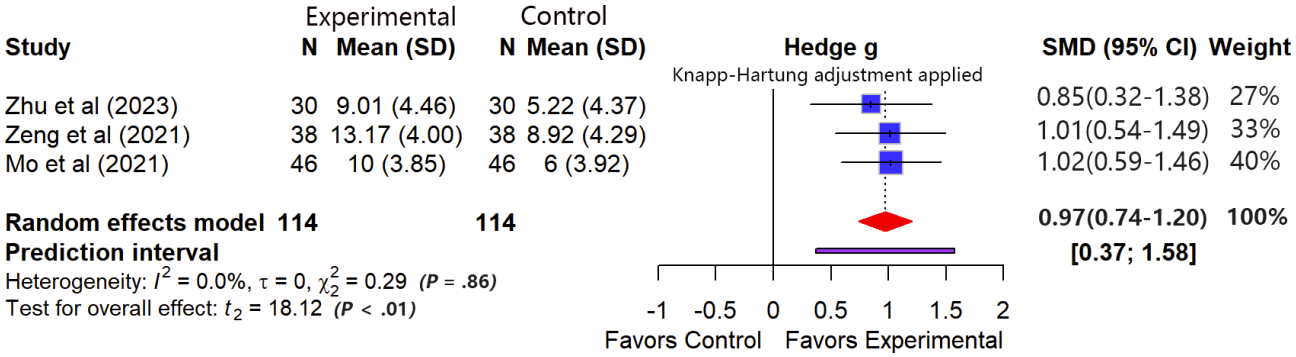

According to the GRADE approach, the certainty of evidence for balance function at short-term follow-up was rated as "moderate." The rating was downgraded because all included primary studies were assessed as having a high risk of bias. No downgrading was applied for inconsistency or imprecision, as no heterogeneity was observed among studies, and the pooled effect estimate was precise. Although the pooled effect size was large and clinically meaningful, an upgrade was not performed. This decision was made to maintain a cautious interpretation, considering that the methodological limitations commonly present in the existing studies could affect the validity of the effect estimate (refer to Multimedia Appendix 5 for the detailed evaluation process).

Cognitive Function

A total of 3 studies [33,55,57] (n=271) were included that assessed cognitive function at short-term follow-up. According to the predefined data pooling criteria, for the study by Zeng [33], we only included the results obtained using the Montreal Cognitive Assessment (MoCA) scale. All analyses were based on change-from-baseline values.

XSL•FO
RenderX

The meta-analysis (Figure 9) indicated mild heterogeneity among the studies (I^2=20.6%, τ=0.1498, τ²=0.0224; Cochran Q=2.52, df=2; *P*=.28). Sensitivity analysis (leave-one-out method) showed that the results are robust, as the maximum change in the effect size after excluding any single study was 0.0218, which is below the threshold of 0.1. The pooled results indicated that, on average, VR-ERI showed a statistically significant trend toward improving patients' cognitive function compared with usual care (SMD 0.78, 95% CI 0.16 to 1.39). However, the 95% PI (−0.16 to 1.71) crossed the line of null effect, suggesting that the beneficial effect of VR-ERI on cognitive function in most future, similar studies is unstable—it could be beneficial or show no difference and therefore requires further exploration in future research. Subgroup analysis was not performed due to the limited number of included studies.

Figure 9. Forest plot of the meta-analysis for cognitive function outcomes [33,55,57].

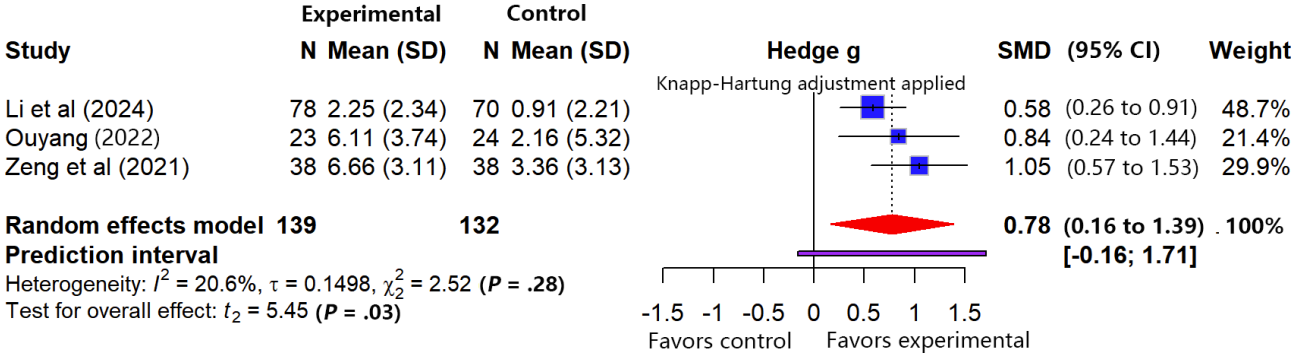

According to the GRADE framework, the quality of evidence was rated as "low." This rating was based on (1) all 3 included studies [33,55,57] were rated as having a high risk of bias, and (2) imprecision of the pooled effect estimate. Although the pooled effect size suggests a moderate effect, the methodological flaws in the studies significantly reduce our confidence in this estimate. Specific rating details are provided in Multimedia Appendix 5.

Limb Motor Function

In total, 2 studies [33,56] (n=168) assessed the improvement of patients' limb motor function during short-term follow-up using the Fugl-Meyer Assessment. As the study by Zeng [33] reported upper limb and lower limb function scores separately, we applied the weighted mean method to combine the function scores from different body parts within the same study into an overall limb motor function score. All analyses were based on change-from-baseline values.

No heterogeneity was observed between the studies (I^2=0%, τ=0, τ²=0; Cochran Q=0.55, df=1; *P*=.46). The meta-analysis (Figure 10) indicated that, based on the average effect, VR-ERI showed a trend toward improving patients' limb motor function at short-term follow-up after discharge compared to usual care (SMD 1.40, 95% CI −0.23 to 3.02), although the difference was not statistically significant and requires further verification. According to the GRADE framework, the quality of evidence was rated as "low." This rating was based on (1) both included studies had a high risk of bias, and (2) the pooled effect estimate was extremely imprecise. Although the pooled effect size suggests a potentially large effect, we did not upgrade the rating due to the very limited evidence base and the low quality of the studies (Multimedia Appendix 5).

Figure 10. Forest plot for limb motor function [33,56].

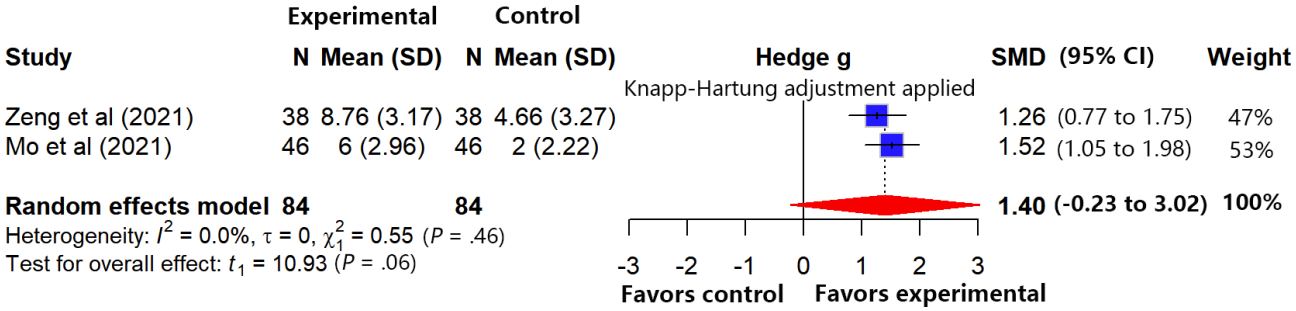

Descriptive Integration Results

Overview

We performed descriptive analysis on outcomes with fewer than 2 included studies and those unsuitable for meta-analysis due to differences in assessment methods or data collection time points. Given the limited number of studies, further validation of the intervention effects on these outcomes is currently lacking, and more high-quality RCTs are needed in the future to evaluate them.

Motor Function Outcomes (Muscle Strength, Grip Strength, Patient Engagement, and Functional Outcomes)

VR-ERI demonstrated multifaceted benefits on motor function, including improved muscle strength. In total, 2 studies supported this finding; Zhu et al [61] reported significantly higher Manual

XSL•FO
RenderX

Muscle Testing scores in the VR-ERI group versus controls (t_{58}=2.33, P=.02), while Wang et al [58] observed stronger muscle strength via the Medical Research Council scale ($F_{1,66}$=7.9, P=.006), with a maintained intervention effect over time ($F_{7,66}$=50.1, P<.001). Due to insufficient raw data, a meta-analysis was not conducted, and further studies are needed to confirm these results. In contrast, findings regarding grip strength improvement were inconsistent. Both Xia et al [59] and Wang et al [58], using a dynamometer, reported no statistically significant between-group difference in absolute grip strength post intervention. Notably, Wang et al [58] observed significant improvement over time in both groups and a significant group-time interaction, suggesting VR-ERI may influence the recovery trajectory. However, its advantage over conventional rehabilitation requires further validation.

Furthermore, VR-ERI also enhanced patient engagement. Both Xia et al [59] and Wang et al [58] reported that the intervention group had significantly longer daily exercise duration than the control group. Mo et al [56] further supported this, finding a higher proportion of patients with high exercise compliance in the VR-ERI group (33/46, 71.74%) versus the control group (25/46, 54.35%) via a self-developed questionnaire. This may be attributed to the inherent enjoyment and real-time feedback of VR-based training, although limited raw data necessitates further research. Regarding functional outcomes, VR-ERI also demonstrated positive effects. Mo et al [56] observed a higher proportion of patients attaining the highest walking ability level (Functional Ambulation Category grade 5) in the intervention group (22/46, 47.93%) versus the control group (14/46, 30.43%). Meanwhile, Xia et al [59] reported a significant reduction in ICU-acquired weakness incidence with VR-ERI (intervention: 13.5% vs control: 35.5%). Future high-quality studies are needed to confirm these effects.

Psychological Function Outcomes (Depression, Posttraumatic Stress Disorder, Somatic Symptoms, and Perceived Well-Being)

The effects of VR-ERI on psychological function were inconsistent. Navarra-Ventura et al [16] found no significant between-group differences in depression or posttraumatic stress disorder at 1 and 12 months post-ICU discharge. In contrast, Ouyang [57] reported a significant reduction in depression during ICU treatment. However, due to differing assessment timepoints between the 2 studies, meta-analysis was not performed, and these divergent conclusions require further validation. Li et al [55] observed better improvement in somatic symptoms and higher perceived well-being in the intervention group at 1 week during ICU stay, but these differences were not maintained 1 month after discharge. Due to differing assessment timepoints, meta-analysis was not performed, and the long-term psychological effects of VR-ERI require further investigation.

Cognitive Function Outcomes (Delirium)

Kim and Kang [54] used the Confusion Assessment Method (CAM) for ICU scale to evaluate the effect of VR-ERI on delirium incidence and found no significant difference between the 2 groups (12.2% vs 12.8%, χ^2_1=0.01; P=.94). In contrast, Weng et al [32] used the CAM scale to assess delirium severity and reported a statistically significant difference in CAM scores between groups (mean 2.25, SD 0.44 vs mean 6.37, SD 1.05, t_{82}=23.453; P<.05). Future high-quality studies with standardized assessment criteria are recommended to validate the intervention effects.

ICU Length of Stay and Recovery Quality

Regarding ICU length of stay, both Xia et al [59] and Wang et al [58] reported significantly shorter stays for the VR-ERI group versus controls, suggesting faster patient turnover. However, a meta-analysis was not feasible due to incomplete raw data. For quality of recovery, Zhu et al [61] (using the Postoperative Quality Recovery Scale) found significantly better recovery rates in the VR-ERI group for physiological function and activity domains. Li et al [55] noted improved mental health scores (12-Item Short Form Health Survey Mental Component Score) at 1 week post intervention, though physical health scores (12-Item Short Form Health Survey Physical Component Score) showed no difference, and no between-group differences remained significant at the 3-month follow-up.

Small-Study Effects

Although the number of included studies for all outcomes in this review did not meet the recommended threshold for conventional statistical tests of publication bias (≥10 studies), we still performed an exploratory analysis on the anxiety outcome, which had the largest number of included studies (n=8). This was done to preliminarily assess the potential impact of bias on the robustness of the conclusions.

The funnel plot (Figure 11) appeared approximately symmetrical by visual inspection. Statistical test results showed that the Egger linear regression test did not indicate significant small-study effects (t_6=−1.94; P=.10). However, the Begg rank correlation test suggested the possible presence of publication bias (z=−2.47, P=.01). To quantify the potential impact of missing studies on the pooled result, a trim-and-fill analysis was conducted. The analysis estimated that 1 potentially missing study needed to be imputed. The adjusted pooled effect size changed from the original SMD −0.862 (95% CI −1.853 to 0.129) to SMD −0.495 (95% CI −1.719 to 0.729), a change of 0.367. The decrease in the absolute value suggests that, if bias exists, its influence may have led to an overestimation of the intervention effect.

XSL•FO
RenderX

Figure 11. Funnel plot for anxiety outcomes[16,17,32,34,35,55,57,60].

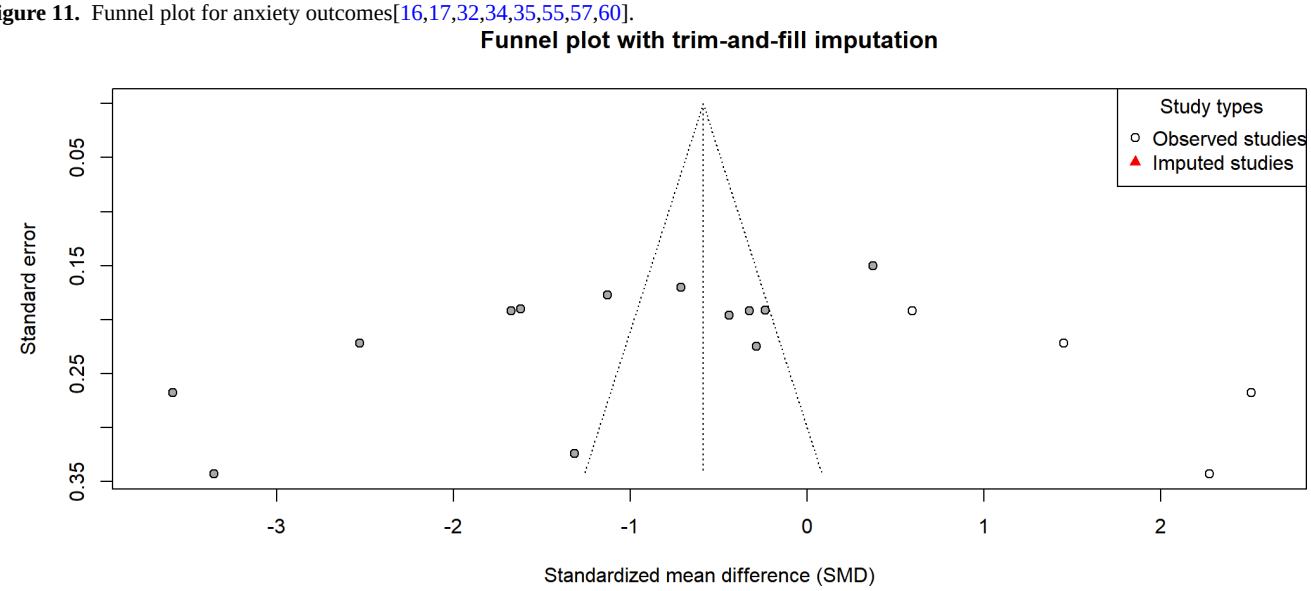

We fully acknowledge that with only 8 included studies [16,17,32,34,35,55,57,60] and substantial heterogeneity present, the statistical power of these tests is limited. Furthermore, heterogeneity itself can interfere with the interpretation of funnel plot asymmetry. Therefore, the above results should be considered exploratory. In summary, the current analysis, while suggesting that publication bias cannot be entirely ruled out, does not provide evidence from the Egger test supporting significant small-study effects. Moreover, the CI for the effect size adjusted by the trim-and-fill method remained wide and crossed the line of null effect. This indicates that, even if potential bias exists, the current conclusion regarding the anxiety outcome is stable. However, it serves as a reminder to interpret the magnitude of this effect size with caution in the future.

Discussion

Principal Results

This review systematically assessed the feasibility, safety, and rehabilitation outcomes of VR-ERI in critically ill patients across 16 RCTs [16,17,30,32-35,53-61] (n=1356). VR-ERI was well-accepted and associated with high exercise adherence. Vital signs remained within acceptable ranges during use, and no serious adverse events were reported; however, mild dizziness, nausea, and fatigue occurred in some studies, underscoring the need for standardized safety monitoring. Regarding efficacy, VR-ERI improved balance during short-term follow-up, supported by low heterogeneity, narrow PIs, and moderate-certainty evidence, indicating robust and potentially generalizable benefit. VR-ERI also shows potential to improve in-ICU anxiety and subjective sleep quality, as well as short-term cognitive and motor function, particularly sleep, when assessed with the RCSQ. However, it did not significantly improve objective sleep parameters or in-ICU pain. For most outcomes, the certainty of evidence was low or very low due to implementation bias (eg, lack of blinding), study inconsistency, and imprecision in pooled estimates. Thus, results should be interpreted cautiously, and current evidence remains insufficient for definitive conclusions. Some outcomes could not be pooled due to limited studies. Further high-quality trials are needed to confirm these findings and clarify the clinical value of VR-ERI in critical care.

VR-ERI may reduce anxiety in critically ill patients, a finding consistent with recent reviews in both critical care [36] and elective surgery [62], suggesting VR could broadly mitigate hospital-related anxiety. Stress recovery theory [63] offers a possible explanation, proposing that exposure to natural landscapes can activate the parasympathetic nervous system and promote emotional relaxation. Research confirms that biophilic elements aid stress recovery [64]. Most VR-ERI scenarios in the included studies incorporated natural views (eg, mountains and rivers), which may help block typical ICU stressors, such as constant light, noise, and isolation. However, the pooled effect in this analysis was not statistically significant, and the 95% PI crossed zero, indicating instability in the estimate. The included studies also had a high risk of bias, with substantial inconsistency and imprecision, yielding very low certainty of evidence. Thus, future rigorous RCTs with adequate sample sizes and transparent reporting are needed to clarify VR's effect on anxiety in this population.

VR-ERI shows potential in improving ICU patients' subjective sleep quality, likely due to its intervention design—2 protocols [30,54] used VR meditation and hypnosis modules to promote relaxation and distraction, enhancing perceived sleep quality [30,54]. However, objective parameters measured with smart wristbands (eg, TST and sleep efficiency) did not improve concurrently, suggesting VR may modulate sleep perception rather than alter sleep physiology. A recent network meta-analysis [65] found that simple interventions like eye masks and earplugs offer more consistent sleep benefits in critically ill patients, while the effect of VR meditation remains unestablished, underscoring VR's uncertain role in sleep intervention. Subgroup analysis indicated positive effects only when sleep was assessed with the RCSQ, not with other subjective scales. This highlights how tool selection influences outcomes, possibly because the RCSQ focuses more on immediate sleep depth and is designed specifically for critically ill patients. Future research should use standardized, population-specific sleep assessment tools to accurately capture

XSL•FO
RenderX

intervention effects. Current evidence is limited by low certainty. Future studies should develop standardized VR sleep protocols, clarify optimal dosage, and incorporate objective measures, such as polysomnography or validated wearable monitoring, to systematically evaluate VR's impact on sleep architecture, thereby building a more reliable evidence base for practice.

We also observed that VR-ERI can improve balance (measured by Berg Balance Scale) and limb motor function (measured by Fugl-Meyer Assessment) in critically ill patients during short-term follow-up. This finding is consistent with the review by He et al [37] on VR for early motor intervention in this population. By providing an immersive environment, VR compensates for the sensory and interaction deficits resulting from prolonged immobilization and isolation. It enhances sensorimotor pathways and balance-related neuroplasticity through visual-vestibular-proprioceptive integration [66]. Furthermore, VR-ERI engages both cognitive and motor functions simultaneously [67] and promotes cortical activation via feedback-driven learning and task-specific adaptation [68,69]. The engaging and gamified design of VR-ERI may also improve patient adherence to exercise, which could explain its advantage over traditional motor rehabilitation. Improvement in balance is supported by moderate-quality evidence and relatively narrow PIs, suggesting that standardized VR-ERI motor activities could be implemented early in critical care rehabilitation, with attention to personalized parameters and long-term effect tracking. However, evidence for limb motor improvement remains low in certainty and requires further rigorous validation.

However, VR-ERI did not demonstrate a statistically significant effect on reducing pain in critically ill patients during ICU treatment, a result that may be associated with several factors. The pain experience of critically ill patients is particularly complex, as frequent threats to psychological and physical integrity make their pain difficult to manage effectively [70]. Neuroimaging studies have shown that receiving VR intervention during painful stimuli can significantly reduce excitability in brain regions associated with pain processing, with a higher level of VR immersion correlating with stronger inhibitory effects [71]. Concurrently, immersive VR devices can divert attention away from nociceptive stimuli by isolating patients' audiovisual connection to the external environment [72]. Nevertheless, in this study, the reported pain scores were relatively low in both groups, which might be related to insufficient immersion of the VR experience, inadequate intervention dosage, or the widespread use of analgesic medications. This warrants further investigation in future studies.

VR-ERI shows a potential trend for improving cognitive function in short-term follow-up, although the 95% PI crosses the null effect line, indicating unstable results. This may stem from several factors. First, the pathophysiology of post-ICU cognitive impairment is complex (eg, neuroinflammation and blood-brain barrier disruption) [73]; for instance, underlying neural damage in conditions like posterior cortical atrophy may limit response to conventional interventions [74], highlighting the need for more targeted programs. Second, widely used screening tools, such as the Mini-Mental State Examination and MoCA, may lack sensitivity to detect subtle, clinically meaningful improvements in critically ill patients. The Mini-Mental State Examination, although brief and simple, has a ceiling effect and inadequately assesses executive function and complex attention [75]. The MoCA is more sensitive to mild impairment but remains a broad screening tool that may miss specific post-ICU executive deficits (eg, inhibitory control and cognitive flexibility) [76]. Therefore, future studies should develop VR modules tailored to ICU-related cognitive impairment and consider longer intervention duration. A multilevel assessment strategy is recommended, combining screening scales with computerized neuropsychological tests (eg, flanker tasks) for objective, quantitative data [77]. Integrating VR with other rehabilitation methods into a multidimensional intervention may also help achieve meaningful cognitive gains. Given the low certainty of current evidence, these findings should be interpreted with caution.

This systematic review is the first to focus on VR intervention studies conducted during the early ICU rehabilitation phase. By systematically searching ten Chinese and English databases and implementing independent dual-reviewer screening and data extraction, it strictly adhered to the PRISMA 2020 guidelines, ensuring a transparent process and reproducible results. The review comprehensively used the Cochrane RoB 2.0 tool for risk of bias assessment and the GRADE framework for evidence grading, supplemented by sensitivity analyses to verify the robustness of the findings. It summarizes and analyzes the safety and clinical feasibility of VR intervention in critically ill patients, while systematically elucidating the efficacy profile of VR on relevant outcomes during both the ICU stay and at short-term follow-up. This provides an important reference for advancing the translational application of VR technology in this patient population.

Limitations and Future Recommendations

Several limitations should be acknowledged. First, the literature search was restricted to Chinese and English publications and excluded trial registries and gray literature, potentially introducing omission and language bias. Second, the limited number of studies and small total sample size reduce statistical power, which may affect the robustness and generalizability of the pooled results. Third, substantial clinical and methodological heterogeneity existed across studies, including variations in VR immersion levels, protocol design, dosage, and outcome measures, contributing to high statistical heterogeneity. Inadequate reporting of precise dosage parameters also limited dose-response analyses. Fourth, most included studies had a high risk of bias due to issues such as unclear allocation concealment, lack of assessor blinding, and inherent challenges in blinding participants and personnel. Consequently, the certainty of evidence for most outcomes was low or very low, necessitating cautious interpretation. Finally, some outcomes were supported by only a few studies, resulting in meta-analyses with limited sample sizes and potentially unstable effect estimates. In summary, further well-designed, large-scale studies are required to validate these findings.

At the clinical level, VR-ERI can supplement traditional early rehabilitation, but widespread adoption faces systematic challenges. First, cost remains a key barrier. Although

XSL•FO
RenderX

commercial VR prices are falling, expenses may still burden some institutions or patients, with VR for early pain management costing approximately €47.48 (approximately US $51.30) per patient [78]. Policymakers should thus consider including VR in insurance reimbursement or supporting it via dedicated funding. Second, patient vulnerability limits generalizability [79]. Those who are older adults, or have vestibular dysfunction, severe cognitive impairment, or complex life-support needs may be unable to tolerate VR. Health care institutions should form multidisciplinary teams to create standardized screening protocols—focusing on vestibular and cognitive function—and implement a safety system covering pre-, intra-, and postsession phases. Implementation should be phased and individualized—starting at low intensity and adjusted gradually under therapist supervision, while training ward staff and patients in device use. Finally, before broader promotion, safety and feasibility must be validated in heterogeneous critically ill populations [14].

At the policy level, relevant departments should establish a comprehensive support framework by developing evidence-based clinical guidelines for VR rehabilitation that specify indications, contraindications, and protocols, and implementing unified device certification and quality control systems to ensure safe and standardized application. Future research should adopt a phased strategy—priorities include conducting large-scale, rigorous pragmatic trials using blinded outcome assessment, diverse control groups, objective monitoring, and long-term follow-up to minimize bias; creating personalized interventions for specific patient groups, defining a core outcome set, and using objective, standardized assessment tools; and enhancing implementation science to identify optimal rollout pathways across different resource settings—beginning with pilot projects at regional medical centers before scaling to primary care. Concurrently, interdisciplinary collaboration among engineering, rehabilitation medicine, and critical care nursing should be strengthened to jointly advance technology innovation and support the development of a more integrated early rehabilitation system for critically ill patients.

Conclusions

This review is the first to examine VR rehabilitation effectiveness specifically during early intervention in the ICU. Unlike previous reviews, we focused strictly on the in-ICU early rehabilitation period to establish an evidence base for initiating such care and to minimize the clinical heterogeneity from combining interventions across different care settings. Evidence was synthesized across physiological, psychological, and cognitive domains. Results indicate that VR-ERI shows considerable potential for improving anxiety and subjective sleep quality during ICU treatment, as well as balance, limb motor function, and cognitive function in short-term follow-up. However, effects on pain and objective sleep quality in the ICU remain unclear. Given the high risk of bias in included studies, substantial heterogeneity, and imprecise pooled effect estimates, the evidence certainty for most outcomes is low, and conclusions should be interpreted cautiously. Additionally, establishing a comprehensive safety monitoring system is essential for future clinical translation. This review offers a new perspective on early ICU rehabilitation, suggesting VR technology could serve as a nonpharmacological adjunct in this phase. However, the current evidence is insufficient to support widespread implementation. Clinical translation requires careful consideration of cost and patient suitability, and long-term benefits need validation through large-scale and rigorous research.

Acknowledgments

The authors declare that no generative artificial intelligence was used at any stage of this study.

Funding

This work was supported by the Capital Medical University Digital Intelligence Nursing Basic Clinical Collaborative Research Project (Grant SZHL23Z02) and the Beijing Hospitals Authority Youth Programme (Grant QML20230515). The former was led by principal investigator (WC), who supervised the research and reviewed the manuscript. The latter was led by the corresponding author (RZ), who was responsible for the study conception and project guidance.

Data Availability

Relevant data are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed search strategy.
[DOCX File, 48 KB - jmir_v28i1e81865_app1.docx]

Multimedia Appendix 2
Excluded literature and reasons.
[DOCX File, 69 KB - jmir_v28i1e81865_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Characteristics of included studies.
[DOCX File, 34 KB - jmir_v28i1e81865_app3.docx]

Multimedia Appendix 4
Dual Revised Cochrane Risk-of-Bias tool for randomized trials assessment process with interrater consistency.
[XLS File, 1604 KB - jmir_v28i1e81865_app4.xls]

Multimedia Appendix 5
Grading of Recommendations Assessment, Development, and Evaluation assessment results.
[DOCX File, 34 KB - jmir_v28i1e81865_app5.docx]

Checklist 1
PRISMA checklist for abstracts.
[PDF File, 119 KB - jmir_v28i1e81865_app6.pdf]

Checklist 2
PRISMA-S checklist.
[PDF File, 66 KB - jmir_v28i1e81865_app7.pdf]

Checklist 3
PRISMA checklist.
[PDF File, 321 KB - jmir_v28i1e81865_app8.pdf]

References

1. Gao S, Liang X, Pan Z, Zhang X, Zhang L. Effect size estimates of risk factors for post-intensive care syndrome: a systematic review and meta-analysis. Intensive Crit Care Nurs 2025 Apr;87:103888. [doi: 10.1016/j.iccn.2024.103888] [Medline: 39561481]
2. Fresenko LE, Rivera ZC, Parry SM, Mayer KP. Post-intensive care syndrome: physical impairments and function. Crit Care Clin 2025 Jan;41(1):1-20. [doi: 10.1016/j.ccc.2024.08.004] [Medline: 39547718]
3. Needham DM, Davidson J, Cohen H, et al. Improving long-term outcomes after discharge from intensive care unit: report from a stakeholders' conference. Crit Care Med 2012 Feb;40(2):502-509. [doi: 10.1097/CCM.0b013e318232da75] [Medline: 21946660]
4. Ho MH, Lee YW, Wang L. Estimated prevalence of post-intensive care cognitive impairment at short-term and long-term follow-ups: a proportional meta-analysis of observational studies. Ann Intensive Care 2025 Jan 10;15(1):3. [doi: 10.1186/s13613-025-01429-z] [Medline: 39792310]
5. Valsø Å, Rustøen T, Småstuen MC, et al. Occurrence and characteristics of pain after ICU discharge: a longitudinal study. Nurs Crit Care 2022 Sep;27(5):718-727. [doi: 10.1111/nicc.12701] [Medline: 34382725]
6. Fukamachi Y, Sakuramoto H, Sato T, et al. Prevalence, risk factors, and intervention of long-term sleep disturbance after intensive care unit discharge: a scoping review. Cureus 2025 Apr;17(4):e83011. [doi: 10.7759/cureus.83011] [Medline: 40421347]
7. Yoo KH, Lee J, Oh J, et al. Depression or anxiety and long-term mortality among adult survivors of intensive care unit: a population-based cohort study. Crit Care 2025 May 6;29(1):179. [doi: 10.1186/s13054-025-05381-z] [Medline: 40329374]
8. Ahn S, LaNoue M, Su H, et al. Post-intensive care syndrome and caregiver burden: a post hoc analysis of a randomized clinical trial. JAMA Netw Open 2025 Apr 1;8(4):e253443. [doi: 10.1001/jamanetworkopen.2025.3443] [Medline: 40198074]
9. Mikkelsen ME, Still M, Anderson BJ, et al. Society of Critical Care Medicine's International Consensus Conference on Prediction and Identification of Long-Term Impairments After Critical Illness. Crit Care Med 2020 Nov;48(11):1670-1679. [doi: 10.1097/CCM.0000000000004586] [Medline: 32947467]
10. Dayton K, Lindroth H, Engel HJ, et al. Creating a culture of an awake and walking intensive care unit: in-hospital strategies to mitigate post-intensive care syndrome. Crit Care Clin 2025 Jan;41(1):121-140. [doi: 10.1016/j.ccc.2024.08.002] [Medline: 39547720]
11. Tipping CJ, Harrold M, Holland A, Romero L, Nisbet T, Hodgson CL. The effects of active mobilisation and rehabilitation in ICU on mortality and function: a systematic review. Intensive Care Med 2017 Feb;43(2):171-183. [doi: 10.1007/s00134-016-4612-0] [Medline: 27864615]
12. Ding N, Zhang Z, Zhang C, et al. What is the optimum time for initiation of early mobilization in mechanically ventilated patients? A network meta-analysis. PLoS ONE 2019;14(10):e0223151. [doi: 10.1371/journal.pone.0223151] [Medline: 31589642]

XSL•FO
RenderX

13. Haghedooren E, Haghedooren R, Langer D, Gosselink R. Feasibility and safety of interactive virtual reality upper limb rehabilitation in patients with prolonged critical illness. Aust Crit Care 2024 Nov;37(6):949-956. [doi: 10.1016/j.aucc.2024.06.004] [Medline: 39054204]
14. de Vries M, Beumeler LFE, van der Meulen J, Bethlehem C, den Otter R, Boerma EC. The feasibility of virtual reality therapy for upper extremity mobilization during and after intensive care unit admission. PeerJ 2025;13:e18461. [doi: 10.7717/peerj.18461] [Medline: 39802193]
15. Godoy-González M, López-Aguilar J, Fernández-Gonzalo S, et al. Efficacy and safety of a non-immersive virtual reality-based neuropsychological intervention for cognitive stimulation and relaxation in patients with critical illness: study protocol of a randomized clinical trial (RGS-ICU). BMC Psychiatry 2024 Dec 18;24(1):917. [doi: 10.1186/s12888-024-06360-4] [Medline: 39696098]
16. Navarra-Ventura G, Gomà G, de Haro C, et al. Virtual reality-based early neurocognitive stimulation in critically ill patients: a pilot randomized clinical trial. J Pers Med 2021 Nov 29;11(12):1260. [doi: 10.3390/jpm11121260] [Medline: 34945732]
17. Merliot-Gailhoustet L, Raimbert C, Garnier O, et al. Discomfort improvement for critically ill patients using electronic relaxation devices: results of the cross-over randomized controlled trial E-CHOISIR (Electronic-CHOIce of a System for Intensive care Relaxation). Crit Care 2022 Sep 3;26(1):263. [doi: 10.1186/s13054-022-04136-4] [Medline: 36057612]
18. Alvarado Sánchez JI, Castillo Morales LM, Cardenas Bolivar YR, et al. Efficacy and safety of the early implementation of a multimodal rehabilitation program in mechanically ventilated patients: a randomized clinical trial protocol. PLoS ONE 2025;20(5):e0324335. [doi: 10.1371/journal.pone.0324335] [Medline: 40388519]
19. Matsuoka A, Yoshihiro S, Shida H, et al. Effects of mobilization within 72 h of ICU admission in critically ill patients: an updated systematic review and meta-analysis of randomized controlled trials. J Clin Med 2023 Sep 11;12(18):5888. [doi: 10.3390/jcm12185888] [Medline: 37762829]
20. Fuke R, Hifumi T, Kondo Y, et al. Early rehabilitation to prevent postintensive care syndrome in patients with critical illness: a systematic review and meta-analysis. BMJ Open 2018 May 5;8(5):e019998. [doi: 10.1136/bmjopen-2017-019998] [Medline: 29730622]
21. Xu J, Wang S, Zhang Q, Yao Y, Yu J. Effectiveness of nurse-led early mobility protocols on the outcomes of critical care patients: a systematic review and meta-analysis. Nurs Open 2025 May;12(5):e70206. [doi: 10.1002/nop2.70206] [Medline: 40371936]
22. Lanier J. Virtual Reality: CITIC Publishing Group Corporation Ltd; 2018.
23. Abrams Z. Enhancing mental health care with VR. IEEE Pulse 2022;13(5):16-20. [doi: 10.1109/MPULS.2022.3208824] [Medline: 36301813]
24. Bauer ACM, Andringa G. The potential of immersive virtual reality for cognitive training in elderly. Gerontology 2020;66(6):614-623. [doi: 10.1159/000509830] [Medline: 32906122]
25. Zhang Q, Fu Y, Lu Y, et al. Impact of virtual reality-based therapies on cognition and mental health of stroke patients: systematic review and meta-analysis. J Med Internet Res 2021 Nov 17;23(11):e31007. [doi: 10.2196/31007] [Medline: 34787571]
26. Li R, Li Y, Kong Y, et al. Virtual reality-based training in chronic low back pain: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2024 Feb 26;26:e45406. [doi: 10.2196/45406] [Medline: 38407948]
27. Guo Q, Zhang L, Gui C, et al. Virtual reality intervention for patients with neck pain: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2023 Apr 3;25:e38256. [doi: 10.2196/38256] [Medline: 37010891]
28. Schröder D, Wrona KJ, Müller F, Heinemann S, Fischer F, Dockweiler C. Impact of virtual reality applications in the treatment of anxiety disorders: a systematic review and meta-analysis of randomized-controlled trials. J Behav Ther Exp Psychiatry 2023 Dec;81:101893. [doi: 10.1016/j.jbtep.2023.101893] [Medline: 37453405]
29. Yen HY, Chiu HL. Virtual reality exergames for improving older adults' cognition and depression: a systematic review and meta-analysis of randomized control trials. J Am Med Dir Assoc 2021 May;22(5):995-1002. [doi: 10.1016/j.jamda.2021.03.009] [Medline: 33812843]
30. Lee SY, Kang J. Effect of virtual reality meditation on sleep quality of intensive care unit patients: a randomised controlled trial. Intensive Crit Care Nurs 2020 Aug;59:102849. [doi: 10.1016/j.iccn.2020.102849] [Medline: 32241625]
31. Ong TL, Ruppert MM, Akbar M, et al. Improving the intensive care patient experience with virtual reality-a feasibility study. Crit Care Explor 2020 Jun;2(6):e0122. [doi: 10.1097/CCE.0000000000000122] [Medline: 32695991]
32. Weng D, Cui M, Jiang J. Application of 5g enabled virtual reality technology in the intervention of delirium in intensive care unit patients. Chinese Journal of Integrated Traditional and Western Medicine in Intensive and Critical Care 2023;30(6):693-696. [doi: 10.3969/j.issn.1008-9691.2023.06.011]
33. Zeng Y. Application effects of cognitive behavioral intervention combined with virtual reality training in postoperative patients with severe traumatic brain injury. Henan Medical Research 2021;30(28). [doi: 10.3969/j.issn.1004-437X.2021.28.058]
34. Yesilot SB, Yeşilkuş R, Beyaz F. Use of virtual reality for reducing pain and anxiety after laparoscopic sleeve gastrectomy: a randomized controlled trial. Pain Manag Nurs 2022 Dec;23(6):826-831. [doi: 10.1016/j.pmn.2022.07.001] [Medline: 35934661]

XSL•FO
RenderX

35. Laghlam D, Naudin C, Coroyer L, et al. Virtual reality vs. Kalinox® for management of pain in intensive care unit after cardiac surgery: a randomized study. Ann Intensive Care 2021 May 13;11(1):74. [doi: 10.1186/s13613-021-00866-w] [Medline: 33983498]
36. Liu Q, Zhang Y, Chen X, Ding B, Guo X, Gai Y. Efficacy of virtual reality in alleviating post-ICU syndrome symptoms: a systematic review and meta-analysis. Nurs Crit Care 2025 Mar;30(2):e70004. [doi: 10.1111/nicc.70004] [Medline: 40070039]
37. He Y, Yang Q, Dai X, et al. Effects of virtual reality technology on early mobility in critically ill adult patients: a systematic review and meta-analysis. Front Neurol 2024;15:1469079. [doi: 10.3389/fneur.2024.1469079] [Medline: 39975851]
38. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
39. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc 2021 Apr 1;109(2):174-200. [doi: 10.5195/jmla.2021.962] [Medline: 34285662]
40. Higgins JPT, Thomas J. In: Chandler J, editor. Cochrane Handbook for Systematic Reviews of Interventions Version 65: Cochrane; 2024. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook [accessed 2025-12-04]
41. Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. BMC Med Res Methodol 2005;5(1):13. [doi: 10.1186/1471-2288-5-13]
42. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
43. Viechtbauer W. Conducting meta-analyses in r with the metafor package. J Stat Softw 2010;36(3). [doi: 10.18637/jss.v036.i03]
44. Murad MH, Wang Z, Chu H, Lin L. When continuous outcomes are measured using different scales: guide for meta-analysis and interpretation. BMJ 2019 Jan 22;364:k4817. [doi: 10.1136/bmj.k4817] [Medline: 30670455]
45. Lin L, Aloe AM. Evaluation of various estimators for standardized mean difference in meta-analysis. Stat Med 2021 Jan 30;40(2):403-426. [doi: 10.1002/sim.8781] [Medline: 33180373]
46. Braitman LE. Statistical power analysis in medical research. Ann Intern Med 1983 Aug;99(2):269-271. [doi: 10.7326/0003-4819-99-2-269] [Medline: 6349459]
47. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
48. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986 Sep;7(3):177-188. [doi: 10.1016/0197-2456(86)90046-2] [Medline: 3802833]
49. Bowden J, Tierney JF, Copas AJ, Burdett S. Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics. BMC Med Res Methodol 2011 Apr 7;11:41. [doi: 10.1186/1471-2288-11-41] [Medline: 21473747]
50. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 6;327(7414):557-560. [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
51. IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open 2016 Jul 12;6(7):e010247. [doi: 10.1136/bmjopen-2015-010247] [Medline: 27406637]
52. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926. [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
53. Caballero LG, Fraga IB, Tremea CEM, et al. Experience of heart failure patients in mobilization with virtual reality: mixed methods study. Rev Esc Enferm USP 2025;59:e20250114. [doi: 10.1590/1980-220X-REEUSP-2025-0114en] [Medline: 40955916]
54. Kim S, Kang J. Effects of virtual reality meditation on sleep and delirium in ICU patients: a randomized controlled trial. Comput Inform Nurs 2025 Apr 2;43(8):e01307. [doi: 10.1097/CIN.0000000000001307] [Medline: 40194914]
55. Li YY, Peng J, Ping YY, et al. Virtual reality-based cognitive-behavioural therapy for the treatment of anxiety in patients with acute myocardial infarction: a randomised clinical trial. Gen Psychiatr 2024;37(2):e101434. [doi: 10.1136/gpsych-2023-101434] [Medline: 38645380]
56. Mo Y, Wang H, Ye P. Application of strengthening rehabilitation nursing multidisciplinary combined with virtual reality training in severe patients with multiple fractures. China Medical Herald 2021;18(14):173-176. [doi: 10.20047/j.issn1673-7210.2021.14.042]
57. Ouyang Y. Study on Influencing Factors of Cognitive Dysfunction and Intervention of Virtual Reality Nature Therapy in Critical Patients: Fujian Medical University; 2022. [doi: 10.27020/d.cnki.gfjyu.2022.001308]
58. Wang J, Shi C, Jia Y, Xiao Q. Effectiveness of virtual reality assisted active limb movement exercises for patients in the respiratory intensive care unit: a randomized pilot study. J Rehabil Med 2025 Jun 3;57:jrm28399. [doi: 10.2340/jrm.v57.28399] [Medline: 40462528]
59. Xia Y, Pan W, Zhao Y, Chen Z, Cai S, Tang Y, et al. Application of virtual reality technology in early activities of ICU patients. Chinese Journal of Nursing 2023;58(9):1050-1055.

XSL•FO
RenderX

60. Yani Y, Wenyi T, Zhenhong F, Wenyu L, Mengyao W, Ruhui C. Improvement of sleep quality and anxiety among cardiac surgery patients by preoperative visits based on virtual reality technology. Journal of Nursing Science 2024;39(22):112-115. [doi: 10.3870/j.issn.1001-4152.2024.22.112]
61. Zhu H, Jie H, Wang X, et al. Application of virtual reality technology in early rehabilitation training of patients after liver transplantation. Nurs Res 2023;37(20):3790-3792 [FREE Full text] [Medline: 37916845]
62. Li H, Chiu PL, Efendi D, Huang H, Ko KY, Wong CL. Effects of virtual reality-based interventions on preoperative anxiety in patients undergoing elective surgery with anesthesia: systematic review and meta-analysis. J Med Internet Res 2025 Apr 30;27:e55291. [doi: 10.2196/55291] [Medline: 40305092]
63. Brosschot JF, Pieper S, Thayer JF. Expanding stress theory: prolonged activation and perseverative cognition. Psychoneuroendocrinology 2005 Nov;30(10):1043-1049. [doi: 10.1016/j.psyneuen.2005.04.008] [Medline: 15939546]
64. Yin J, Yuan J, Arfaei N, Catalano PJ, Allen JG, Spengler JD. Effects of biophilic indoor environment on stress and anxiety recovery: a between-subjects experiment in virtual reality. Environ Int 2020 Mar;136:105427. [doi: 10.1016/j.envint.2019.105427] [Medline: 31881421]
65. Huang D, Li Y, Ye J, Liu C, Shen D, Lv Y. Different nursing interventions on sleep quality among critically ill patients: a systematic review and network meta-analysis. Medicine (Baltimore) 2023 Dec 29;102(52):e36298. [doi: 10.1097/MD.0000000000036298] [Medline: 38206715]
66. Wankhede NL, Koppula S, Ballal S, et al. Virtual reality modulating dynamics of neuroplasticity: innovations in neuro-motor rehabilitation. Neuroscience 2025 Feb 6;566:97-111. [doi: 10.1016/j.neuroscience.2024.12.040] [Medline: 39722287]
67. Morone G, Spitoni GF, De Bartolo D, et al. Rehabilitative devices for a top-down approach. Expert Rev Med Devices 2019 Mar;16(3):187-195. [doi: 10.1080/17434440.2019.1574567] [Medline: 30677307]
68. Cervera MA, Soekadar SR, Ushiba J, et al. Brain-computer interfaces for post-stroke motor rehabilitation: a meta-analysis. Ann Clin Transl Neurol 2018 May;5(5):651-663. [doi: 10.1002/acn3.544] [Medline: 29761128]
69. Lavi R, Pau M, Dreyer-Alster S, Drori T, Cocco E, Kalron A. Real-time cortical activity during virtual reality practice in people with multiple sclerosis: a pilot fNIRS study. J Neuroeng Rehabil 2025 Jul 4;22(1):148. [doi: 10.1186/s12984-025-01681-1] [Medline: 40615921]
70. Leininger S. The challenges of managing pain in the intensive care unit. Crit Care Nurs Q 2022;45(4):325-331. [doi: 10.1097/CNQ.0000000000000424] [Medline: 35980794]
71. Hoffman HG, Richards TL, Van Oostrom T, et al. The analgesic effects of opioids and immersive virtual reality distraction: evidence from subjective and functional brain imaging assessments. Anesth Analg 2007 Dec;105(6):1776-1783. [doi: 10.1213/01.ane.0000270205.45146.db] [Medline: 18042882]
72. Matamala-Gomez M, Donegan T, Świdrak J. VR for pain relief. Curr Top Behav Neurosci 2023;65:309-336. [doi: 10.1007/7854_2022_402] [Medline: 36592274]
73. Pandharipande PP, Girard TD, Jackson JC, et al. Long-term cognitive impairment after critical illness. N Engl J Med 2013 Oct 3;369(14):1306-1316. [doi: 10.1056/NEJMoa1301372] [Medline: 24088092]
74. Suárez-González A, Bier N, Sauvageau H, Pelak VS, Ahmed S. Cognitive rehabilitation in posterior cortical atrophy. Pract Neurol 2025 Jan 16;25(1):40-44. [doi: 10.1136/pn-2024-004259] [Medline: 39515839]
75. Pfoh ER, Chan KS, Dinglas VD, et al. Cognitive screening among acute respiratory failure survivors: a cross-sectional evaluation of the Mini-Mental State Examination. Crit Care 2015 May 5;19(1):220. [doi: 10.1186/s13054-015-0934-5] [Medline: 25939482]
76. Fu C, Jin X, Chen B, et al. Comparison of the Mini-Mental State Examination and Montreal Cognitive Assessment executive subtests in detecting post-stroke cognitive impairment. Geriatr Gerontol Int 2017 Dec;17(12):2329-2335. [doi: 10.1111/ggi.13069] [Medline: 28675607]
77. Lin LT, Lin HA, Bai CH, Lin SF. Flanker task assessment of inhibitory control in ICU survivors: preliminary insights into cognitive impairment. Behav Brain Res 2025 Oct 18;495:115791. [doi: 10.1016/j.bbr.2025.115791] [Medline: 40907667]
78. Groenveld TD, van Dorst P, Ten Broek RPG, et al. Early health economic analysis of virtual reality therapy for pain management after surgery. Int J Surg 2025 Jun 1;111(6):3798-3809. [doi: 10.1097/JS9.0000000000002385] [Medline: 40387696]
79. Li M, Race M, Huang F, Escalon MX. Role of virtual reality to promote mobilization in the critical care setting: a narrative review. Am J Phys Med Rehabil 2025 May 1;104(5):487-494. [doi: 10.1097/PHM.0000000000002648] [Medline: 39642293]

Abbreviations

CAM: Confusion Assessment Method
CNKI: China National Knowledge Internet
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
ICU: intensive care unit
MoCA: Montreal Cognitive Assessment
PI: prediction interval
PICOS: Population, Intervention, Comparison, Outcome, and Study Design

XSL•FO
RenderX

PICS: post-intensive care syndrome
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
RCSQ: Richards-Campbell Sleep Questionnaire
RCT: randomized controlled trials
RoB 2: revised Cochrane Risk of Bias tool
SinoMed: SinoMed Service System
SMD: standardized mean difference
TST: total sleep time
VIP: VIP Database
VR: virtual reality
VR-ERI: virtual reality–based early rehabilitation intervention
WASO: wake after sleep onset

Edited by S Brini; submitted 18.Aug.2025; peer-reviewed by A Shivanna, MDP Beristain-Colorado, O Oluwole, U Ogbonna; accepted 29.Dec.2025; published 06.Mar.2026.

Please cite as:
Wu F, Wu Y, Xing Y, Cai W, Zhang R
Effectiveness of Virtual Reality–Based Early Rehabilitation Strategies on Pain, Sleep, Anxiety, Balance, Cognition, and Limb Motor Function in Adult Intensive Care Unit Patients: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e81865
URL: https://www.jmir.org/2026/1/e81865
doi:10.2196/81865

© Fei Wu, Yunting Wu, Yana Xing, Weixin Cai, Ran Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Artificial Intelligence for Medicines Information: Scoping Review of Clinical Applications and Digital Health Inequalities

Shahd Al-Arkee[1], PhD; Josephine Falade[2], MSc; Vibhu Paudyal[3], PhD

[1]Department of Life Sciences, University of Bath, Cleaverton Down, Bath, United Kingdom

[2]Research Department of Practice and Policy, UCL School of Pharmacy, University College London, London, United Kingdom

[3]Department of Health Services and Population Research, King's College London, London, United Kingdom

Corresponding Author:
Shahd Al-Arkee, PhD
Department of Life Sciences, University of Bath, Cleaverton Down, Bath, United Kingdom

Abstract

Background: Artificial intelligence (AI) has the potential to support medicines information services. However, a comprehensive mapping of its use, particularly within pharmacy practice and in the context of digital health inequalities, is lacking.

Objective: This scoping review mapped existing evidence on AI-driven medicines information, focusing on the accuracy and completeness of AI-generated content, the role of health care professionals (HCPs), particularly pharmacists, and the impact of digital health inequalities on AI adoption.

Methods: This scoping review was informed by the methodological framework proposed by Levac et al, which includes modifications to the original Arksey and O'Malley scoping review framework. A systematic search was conducted across MEDLINE (Ovid), PubMed Central, Cochrane Library, CINAHL Plus (EBSCOhost), International Pharmaceutical Abstracts (IPA), Web of Science, and Google Scholar from inception to January 2025, which served as the search cutoff date. Peer-reviewed studies in English evaluating the role of AI in medicines information across any health care settings (including patient homes) were included. The results are reported in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines.

Results: A total of 1911 citations were identified, with 14 studies meeting the inclusion criteria. AI tools showed promise in supporting medicines information services but were found to have limitations in accuracy, particularly when applied to complex clinical queries. Pharmacists were the most engaged HCPs in the evaluation of AI-generated content. Only 3 studies explored digital health inequalities in the context of AI and access to medicines information. Reported barriers included misinformation risks, regulatory gaps, and digital health inequalities, particularly infrastructure limitations and disparities in digital literacy, which affected AI adoption.

Conclusions: AI-driven tools show promise in supporting medicines information services, but concerns remain. HCPs, particularly pharmacists, play a critical role in AI evaluation and validation, yet their involvement remains ill-defined. Addressing digital health inequalities is essential for effective AI integration. Future research should focus on identifying and minimizing digital health inequalities, as well as evidence-informed AI implementation in medicines information services.

Trial Registration: OSF Registries 2byvn; https://osf.io/2byvn/overview

(J Med Internet Res 2026;28:e77747) doi:10.2196/77747

KEYWORDS

artificial intelligence; medicines information; pharmacy; digital health inequalities; scoping review

Introduction

Artificial intelligence (AI), often described as "computational intelligence" or the "science and engineering of creating intelligent machines" [1], is a rapidly evolving field focused on replicating human-like behavior in computers and related technologies [2]. A key subset of AI, machine learning (ML), enables computers to analyze vast datasets and improve predictive accuracy without explicit programming [3]. AI applications in health care range from rule-based decision systems to advanced ML tools. By identifying patterns and learning from experience, ML algorithms can assist in predicting patient outcomes and supporting clinical decision-making across various health care settings [4].

AI has recently demonstrated increasing potential in pharmacy practice, particularly in medication management. A recent systematic review analyzing multiple AI algorithms reported that technology-driven approaches can enhance medication management in primary care [5]. The growing integration of AI into health care has led to the development of various AI

XSL•FO
RenderX

interventions, including algorithms designed to support both patients and health care professionals (HCPs) in disease management [6]. Several reviews have evaluated AI beyond medicines information, encompassing interventions in disease management [7], diagnosis [8], and shared decision-making [9]. Other reviews highlighted AI's role in pharmacy practice, demonstrating its effectiveness in optimizing medication management, enhancing drug safety, and improving clinical decision support systems [10]. However, the application of AI specifically to medicines information remains underexplored. Medicines information is recognized as a key component of promoting rational use of medicines, ensuring prescribers, dispensers, and consumers have access to independent and unbiased information about medication use [11].

Despite the growing interest in AI applications, there remains a need to comprehensively map how AI is used in medicines information services. Existing research lacks a clear synthesis of applications of AI, the accuracy and completeness of AI-generated responses, and the challenges influencing its adoption. Pharmacists and other HCPs play a crucial role in evaluating and integrating AI-driven tools, yet the extent of their engagement and its implications for pharmacy practice remain unclear. Furthermore, digital health inequalities refer to disparities in access to and use of digital health technologies, which are often shaped by factors such as internet connectivity and geographic location. These may influence the equitable implementation of AI across diverse health care settings.

This scoping review aims to map the current landscape of AI in medicines information, focusing on its applications, content accuracy and completeness, and adoption within pharmacy practice.

Methods

Information Sources and Search Strategy

This scoping review was conducted following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [12], available in Checklist 1. The methodological framework proposed by Levac et al [13], which expands on the original approach by Arksey and O'Malley [14], was used to guide the review. The protocol was developed in accordance with the following steps that are (1) identifying the research questions; (2) identifying relevant studies; (3) selecting studies through a team-based approach for study selection and data extraction; (4) charting the data using a structured approach; and (5) collating, summarizing, and reporting the results. This review was guided by four research questions that are (1) What impact do AI tools have on the dissemination and accessibility of medicines information in different health care settings, particularly in pharmacy practice? (2) What is the reported accuracy and completeness of AI-generated medicines information, and what factors influence its reliability? (3) What are the barriers and facilitators to AI adoption, particularly in addressing digital health inequalities? (4) How do HCPs, particularly pharmacists, perceive and engage with AI tools for medicines information? This protocol was registered and available on the Open Science Framework website [15].

A systematic literature search was conducted across MEDLINE (Ovid), PubMed Central, Cochrane Library, CINAHL Plus (EBSCOhost), International Pharmaceutical Abstracts (IPA), Web of Science, and Google Scholar from inception to January 2025. The database searches were conducted between January 10 and 20, 2025, with January 31, 2025, set as the inclusion cutoff date. A 4-domain search strategy was used, including terms related to AI, medicines information, digital health inequalities, and the role of the pharmacy workforce, particularly pharmacists, in AI adoption. The search strategy was developed and refined by 2 reviewers in consultation with an expert from the University College London Pharmacy School library. The final search strategy is presented in Multimedia Appendix 1 and was formatted in line with PRESS (Peer Review of Electronic Search Strategies) guidelines [16].

Study Selection and Data Extraction

Studies were included if they were qualitative, quantitative, or mixed methods studies, published in English in peer-reviewed journals, and explored the role of AI in enhancing medicines information. Medicines information was defined as a key component of promoting rational use of medicines, ensuring prescribers, dispensers, and consumers have access to independent and unbiased information about medication use [11] and guided the eligibility assessment. Eligible studies involved AI-enabled tools, including rule-based expert systems and ML models. Tools lacking any AI or ML functionality, such as conventional digital databases, were excluded. Reviews, editorials, commentaries, letters, media data, gray literature (eg, reports, theses, and white papers), and conference abstracts were also excluded to ensure methodological rigor and consistency.

A 2-stage selection process was used. In the first stage, one reviewer (SA) screened titles and abstracts for relevance and removed duplicate records, and the second reviewer (VP) thoroughly assessed the titles and abstracts for inclusion. In the second stage, 2 reviewers independently evaluated the full-text studies against the prespecified eligibility criteria. Disagreements were resolved through discussion and consensus. Bibliographies of selected studies were manually searched to identify additional relevant references.

Data extraction was performed using a standardized form developed specifically for this review. Extracted data included study characteristics, AI intervention details, and key findings related to AI-generated medicines information.

Data Synthesis and Quality Assessment

Data were synthesized using a descriptive approach, following a narrative synthesis method. The evidence was mapped to the predefined categories of HCP involvement, accuracy and completeness of AI-generated content, and implications for digital health inequalities. Study characteristics were also summarized, and gaps in the literature were highlighted. A formal quality assessment of the included studies was not conducted, in line with Arksey and O'Malley methodological framework for scoping reviews [14]. However, variations in study quality and methodological limitations were described narratively to support transparency and interpretation.

XSL•FO
RenderX

Results

Search Results and Study Selection

The systematic search yielded 1911 citations, of which 447 duplicates were removed, resulting in 1464 records screened by title and abstract. Of these, 20 studies were assessed for full-text review. Six studies were excluded, and no additional records were identified through manual searching. Therefore, 14 studies met the inclusion criteria and were included in this review [17-30]. A PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart summarizing the study selection process is shown in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram depicting study selection. AI: artificial intelligence.

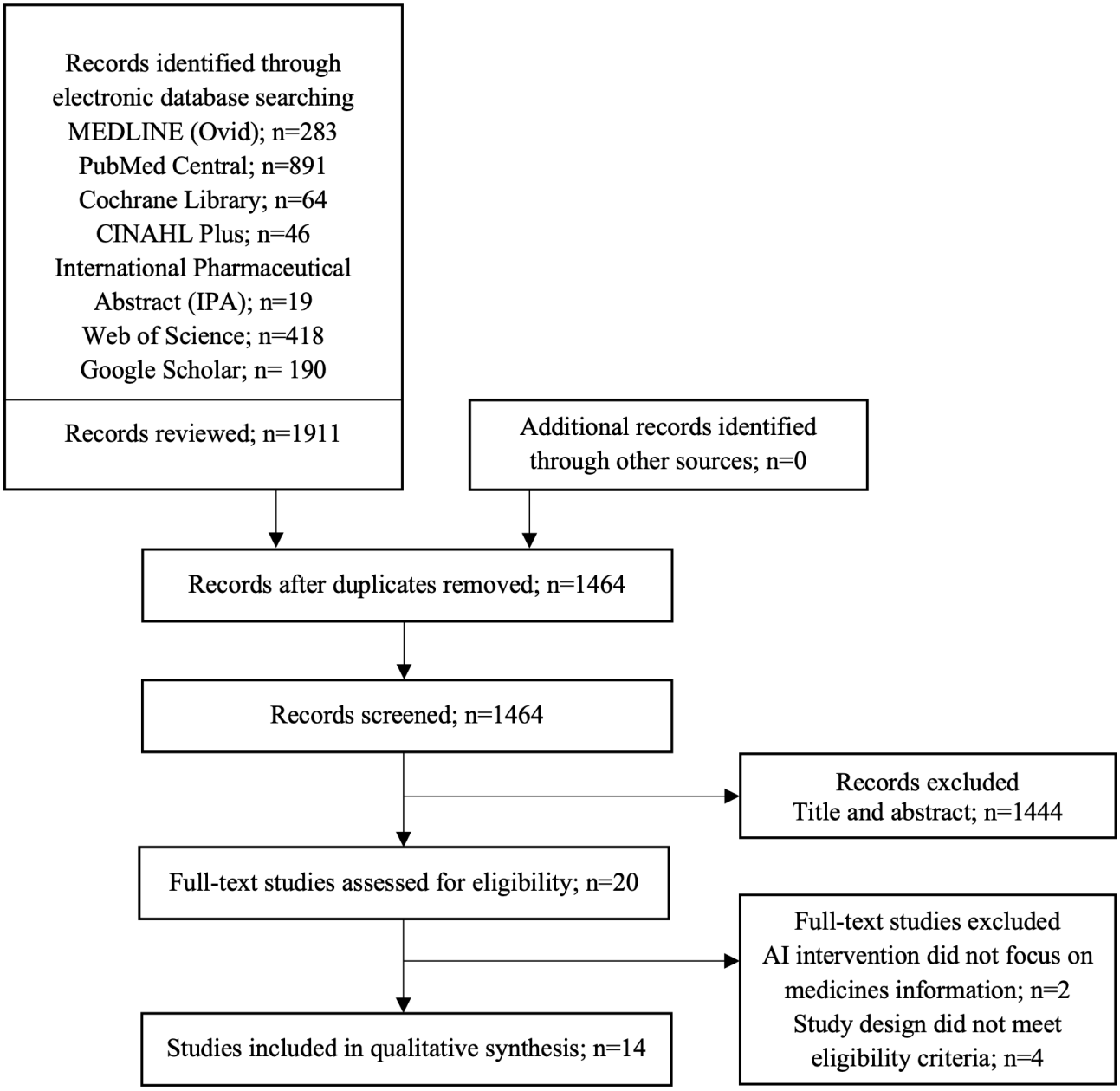

Study Characteristics and Design

The included studies were published between 2021 and 2025. Most originated from high-income economies, as classified by the World Bank, including Australia [26], Canada [21], Germany [17], the Netherlands [30], Singapore [27], and the United States [18-20,22,25,29]. In contrast, only a few studies originated from middle- and low-income countries, such as Egypt [28], Ethiopia [24], and India [23]. All studies focused on the application of AI for delivering medicines information (Table 1).

XSL•FO
RenderX

Table . Characteristics of the included studies.

Study; country	Health care setting	Data sample size	Research purpose	Reported outcomes
Andrikyan et al [17]; Germany	Online patient drug information search	500 chatbot responses (readability, completeness, and accuracy), 20 responses (safety)	Assess chatbot-generated drug advice for quality, readability, and safety concerns	• Readability • Completeness • Accuracy • Safety
Beavers et al [18]; United States	Inpatient hospital	200 real-world medication-use questions	Analyze the clinical completeness, correctness, usefulness, and safety of chatbot and medication database responses to everyday inpatient medication-use questions	• Clinical correctness • Completeness • Safety
Cornelison et al [19]; United States	AI[a]-driven medication counseling	240 chatbot responses	Evaluate chatbot accuracy and completeness in answering medication-related patient questions	• Accuracy • Completeness
Grossman et al [20]; United States	Academic drug information service	39 medication-related questions	Evaluate ChatGPT's accuracy, completeness, and relevance in drug inquiries	• Accuracy • Completeness • Satisfactory • Performance by question complexity
Laymouna et al [21]; Canada	Pharmacists in HIV care (community and hospital)	41 pharmacists from 15 municipalities	Assess pharmacists' knowledge, attitudes, and practices in HIV care and evaluate the usability of an AI-based chatbot for medicines information support	• Pharmacists' knowledge, attitudes, practices, and perceived barriers • Usability of the chatbot
Munir et al [22]; United States	Clinical pharmacy practice	32 pharmacy-based clinical questions	Assess ChatGPT's accuracy in answering pharmacy-related clinical queries	• Accuracy • Case-based questions
Ramasubramanian et al [23]; India	Not stated	462 medication dosage-related queries	Evaluate the accuracy of ChatGPT 3.5, ChatGPT 4, and Google Bard in providing medical drug dosages based on Harrison's Principles of Internal Medicine	• Accuracy
Sendekie et al [24]; Ethiopia	Community pharmacies	225 pharmacists	Investigate pharmacists' perceptions and willingness to use AI in practices, and their perceived barriers to its implementation	• Perception of AI and willingness to use AI • Barriers to AI adoption
Sheikh et al [25]; United States	Nephrology and pharmacy practice	124 nonprescription drugs and supplements assessed	Evaluate ChatGPT in drug safety for patients with kidney disease	• Accuracy • Safety
Stanceski et al [26]; Australia	Hospital discharge summaries (MIMIC-IV database)	100 chatbot responses (summaries)	Evaluate responses for safety, accuracy, and language simplification for AI-generated patient-centered discharge instructions	• Accuracy • Language simplification
Sumner et al [27]; Singapore	Outpatient clinics (hospital)	20 (10 health care providers, 10 patients)	Develop and evaluate an AI-driven nudge intervention for medication adherence	• Acceptance • Preference and concerns
Taha et al [28]; Egypt	Pharmacy practice (community and hospital pharmacies)	428 pharmacists	Explore pharmacists' perceptions, practices, and concerns regarding ChatGPT in pharmacy practice	• Perceptions on accuracy • Privacy concerns

XSL•FO
RenderX

Study; country	Health care setting	Data sample size	Research purpose	Reported outcomes
Triplett et al [29]; United States	Academic-based drug information center	84 ChatGPT responses analyzed	Assess the accuracy, completeness, and consistency of ChatGPT responses to drug information inquiries compared to drug information center responses	• Accuracy • Readability
van Nuland et al [30]; The Netherlands	Hospital-based clinical pharmacy practice	30 clinical pharmacy questions	Evaluate ChatGPT's ability to provide appropriate responses to clinical pharmacy questions and assess accuracy and consistency	• Accuracy • Completeness

[a]AI: artificial intelligence.

The included studies used a range of study designs. Eight studies used comparative and evaluation methodologies to assess AI-generated medicines information accuracy and usability [18-20,22,23,25,26,29]. Two studies used cross-sectional survey designs to explore user perceptions, knowledge, and attitudes toward AI-based tools [17,28]. Two studies adopted a human-centered design approach to assess AI-driven interventions: one evaluating AI-generated patient discharge instructions [26], and another developing an AI-driven nudge tool for medication adherence [27]. While most studies used quantitative research approaches**,** 2 studies used survey-based needs assessment designs to explore contextual factors and AI implementation considerations [21,24]. The studies were conducted across diverse health care settings, including community-based and primary care [21,24,28], secondary care (outpatients) [22,26,27], tertiary care settings [18,25,30], and web-based patient communities (online platforms without direct clinical engagement) [17,19,20,29]. One study did not specify a health care setting [23].

AI Intervention Characteristics

The AI tools evaluated across the included studies showed diverse functionalities in delivering medicines information (Table 2). Most AI systems used natural language processing (NLP) to generate medication-related responses [17-20,22,23]. Microsoft Bing Copilot analyzed user queries and provided structured drug information, while Micromedex with Watson generated text-based responses but struggled with complex clinical scenarios [17,18]. ChatGPT was widely assessed using various versions, with studies evaluating its accuracy in answering patient and academic drug inquiries, pharmacy-based clinical questions, and drug dosage recommendations [19,20,22,23]; among these, ChatGPT-4 (OpenAI) outperformed other models in a comparative evaluation [23].

Other included studies focused more on AI applications, adoption barriers, or decision-making, rather than primarily evaluating NLP-generated responses [21,24-30]. The MARVIN-Pharma chatbot supported pharmacists in HIV care, enhancing access to adherence counseling and clinical guidelines [21]. An evaluation of AI adoption in community pharmacies revealed that while AI could streamline medication management, infrastructure, and training barriers limited its potential [24]. Another study explored the utility of AI in nephrology and pharmacy practice, finding that ChatGPT sometimes misclassified supplement safety, thereby indicating the need for human oversight [25]. In hospital settings, AI-generated discharge summaries improved readability but also raised safety concerns [26]. AI-driven nudges enhanced medication adherence through personalized reminders for outpatients [27]. While AI tools improved efficiency and medication counseling, concerns persisted regarding misinformation, privacy, and patient-specific recommendations [28]. AI-generated drug information improved clarity but required pharmacist oversight, and hospital pharmacy AI responses often lacked reliability, raising concerns about clinical decision-making [29,30].

XSL•FO
RenderX

Table . Characteristics of the AI[a] interventions.

Study; country	AI name and functionality	AI potential risk	Research conclusion
Andrikyan et al [17]; Germany	Microsoft Bing Copilot: for patient drug-related queries	Incomplete or inaccurate drug information, and readability issues	• Capable of providing complete and accurate information • Not reliable for medical advice without validation
Beavers et al [18]; United States	Micromedex with Watson (IBM Watson Health): AI chatbot integrated with an evidence-based pharmacological knowledge base (MDX), using natural language processing to provide accurate, complete, and safe responses to medication-use questions	Limited scope, incomplete answers, and potential safety concerns	• Traditional drug databases outperformed AI chatbots in completeness, safety, and usefulness • Chatbot responses were clinically correct but not always complete or safe
Cornelison et al [19]; United States	ChatGPT-3.5: general AI chatbot for patient medication-related questions	Incomplete or inaccurate drug information	• Mostly accurate for answering common medication-related questions • May provide incomplete or inaccurate information
Grossman et al [20]; United States	ChatGPT 3.5: general AI chatbot for pharmaceutical medication-related questions	Incomplete or inaccurate drug information and fabricated references	• Unreliable for medication-related inquiries due to frequent inaccuracies, incomplete responses, and misleading references
Laymouna et al [21];Canada	MARVIN-Pharma: AI chatbot designed to support pharmacists in delivering evidence-based medicines information for HIV care	Usability concerns, with pharmacist engagement required	• AI could support HIV care • Needs pharmacist engagement in the chatbot's development
Munir et al [22]; United States	ChatGPT 3.5: generative AI language model designed to answer clinical questions in pharmacy practice	Misleading information, poor performance in complex cases, and safety concerns	• Limited success • Excels in calculations • Struggles with drug information and patient-specific cases
Ramasubramanian et al [23]; India	ChatGPT 3.5, ChatGPT 4, Google Bard: AI models assessed for drug dosage accuracy, based on Harrison's Principles of Internal Medicine	Accuracy varied across diseases and organ systems	• ChatGPT 4 exhibited the highest accuracy • Requires further validation before clinical application
Sendekie et al [24]; Ethiopia	Not specified: generic AI applications in pharmacy: for personalized patient care and decision support	Needs for resources, policies, training, and infrastructure	• AI is recognized as beneficial • Needs training and policy development for successful integration
Sheikh et al [25]; United States	ChatGPT 3.5 and ChatGPT 4: evaluated for assessing the safety of nonprescription medications and supplements in kidney disease, compared to Micromedex	Inaccuracy, lack of reliability, and patient safety concerns	• ChatGPT 4 has potential, but remains unreliable compared to established databases • Further AI development is needed before clinical adoption
Stanceski et al [26]; Australia	ChatGPT 3.5 (via Microsoft Azure OpenAI): AI-generated patient-friendly discharge instructions from hospital summaries	Potential medication errors	• AI tools can generate patient-centered discharge instructions • Require clinician oversight before implementation
Sumner et al [27]; Singapore	AI-driven nudge system for medication adherence, integrating reminders, social references, and automated interventions	Digital literacy barriers for older adults, and data privacy concerns	• Promising for improving medication adherence but must be tailored to user needs • Needs flexible delivery, simplified data entry, and caregiver involvement

XSL•FO
RenderX

Study; country	AI name and functionality	AI potential risk	Research conclusion
Taha et al [28]; Egypt	ChatGPT: AI-driven drug information tool assisting pharmacists with counseling, drug interactions, and clinical queries	Concerns on accuracy, data privacy, and bias	• Potential applications in pharmacy but requires accuracy improvements and regulatory guidelines
Triplett et al [29]; United States	ChatGPT 3.5: AI-generated responses for pharmaceutical inquiries, based on NLP[b] and public databases	Accuracy concerns, lack of references, requiring pharmacist validation	• ChatGPT provided clear and readable responses • Lacked accuracy, requiring expert review for reliability
van Nuland et al [30]; The Netherlands	ChatGPT 3.5: AI-generated responses for hospital pharmacy inquiries, including dosing, drug interactions, and therapeutic drug monitoring	Lack of response accuracy, inconsistent AI-generated information, risk of medication errors	• ChatGPT offered clinical advice • Inconsistent accuracy and reproducibility, making it unsuitable for independent clinical use

[a]AI: artificial intelligence.

[b]NLP: natural language processing.

Accuracy and Completeness

Accuracy was primarily evaluated based on the correctness of AI-generated responses in providing medication-related information, but only 5 studies [22,23,25,26,29] conducted a direct assessment. A study evaluated ChatGPT-3.5, ChatGPT-4, and Google Bard for drug dosage recommendations. ChatGPT-4 demonstrated the highest performance (83.77%), but accuracy varied across diseases and organ systems [23]. Another study assessing AI's classification of drug safety in patients with kidney disease found that ChatGPT-4 had higher agreement (81.4%) with Micromedex than ChatGPT-3.5 (64.5%), yet it remained unreliable for independent clinical use [25]. Similarly, a study analyzing AI-generated discharge summaries found that 18% of responses contained potentially harmful safety issues, including 6% with "hallucinated information," meaning responses that appeared plausible but were actually inaccurate and 3% introducing unprescribed medications [26]. Another study compared ChatGPT's accuracy in drug information inquiries against a drug information center, reporting an accuracy rate of 50%. ChatGPT lacked proper references and reliability for clinical use [29]. Another study identified significant limitations in ChatGPT's accuracy, particularly in patient-specific scenarios, emphasizing the need for verification before integration into pharmacy practice [22].

Five additional studies evaluated both accuracy and completeness, evaluating how well AI responses covered all essential information. Microsoft Bing Copilot achieved 100% median accuracy and completeness across 500 responses, yet experts reviewing a subset of 20 chatbot responses found that 66% were potentially harmful, with 22% classified as life-threatening [17]. Micromedex with Watson produced 85% clinically correct responses, but only 65% were complete and 71% acceptable for safety [18]. AI-generated responses to hospital pharmacy inquiries were only 26% correct and complete, while 22% were correct but incomplete, 30% partially correct, and 22% completely incorrect [30]. One study found ChatGPT-generated responses were 92.5% accurate and 80.8% complete, though 4.2% contained minor errors [19]. Another study reported that only 26% of ChatGPT's responses were satisfactory, while 38% lacked a direct response, 38% contained inaccuracies, and 41% lacked completeness [20]. A detailed grouped summary of all included studies by AI task and evaluation metric is presented in Table 3.

XSL•FO
RenderX

Table . Grouped summary by AI[a] task and metric.

AI task	Study; country	Evaluation metric	Outcome
Drug information	Andrikyan et al [17]; Germany	• Accuracy • Safety	• 100% median accuracy • 66% potentially harmful (subset safety analysis)
	Beavers et al [18]; United States	• Accuracy • Completeness • Safety	• 85% accuracy • 65% completeness • 71% safety
	Cornelison et al [19]; United States	• Accuracy • Completeness	• 92.5% accuracy • 80.8% completeness
	Grossman et al [20]; United States	• Accuracy • Completeness	• 26% satisfactory • Among unsatisfactory responses, 38% lacked accuracy and 41% lacked completeness
	Taha et al [28]; Egypt	• Perception • Concerns	• Perceived benefits (73.6%) • Concerns about privacy, bias, and accuracy
	Triplett et al [29]; United States	• Accuracy • Reference • Quality	• 50% accuracy • Poor reference quality
	van Nuland et al [30]; The Netherlands	• Accuracy • Completeness	• 26% correct and complete • 22% correct but incomplete
Dosage recommendations	Ramasubramanian et al [23]; India	• Accuracy	• 83.77% accuracy • Performance varied by disease
HIV care	Laymouna et al [21]; Canada	• Perception • Usability	• Perceived potential to support HIV care, pharmacist engagement essential
Clinical queries	Munir et al [22]; United States	• Expert judgment	• Good in calculation (100%) • Poor in complex cases (20%) • Safety concerns
System-level use	Sendekie et al [24]; Ethiopia	• Infrastructure readiness • Perception	• Positive perception • Policy and training needed for integration
Safety assessment	Sheikh et al [25]; United States	• Agreement • Safety	• 81.4% agreement (ChatGPT-4) with Micromedex • Not yet reliable for clinical use
Discharge instructions	Stanceski et al [26]; Australia	• Safety	• 18% potentially harmful • Incorrect medication content
Adherence support	Sumner et al [27]; Singapore	• Perception	• Acceptable and promising • User-centered design emphasized

[a]AI: artificial intelligence.

Digital Health Inequalities

Aspects of digital health inequalities investigated within the included studies related to AI-driven interventions to enhance health care access, support medication adherence, and provide drug information. Key barriers identified were limited infrastructure and digital literacy (Table 4).

XSL•FO
RenderX

Table . Summary of AI[a] adoption barriers by country context.

Study; country	Research setting (geography)	Barriers category	Key barriers identified	Patient-level inequities
Sendekie et al [24]; Ethiopia	LMIC[b]	Structural or institutional	• Lack of internet • Lack of AI tools • Limited training	• Infrastructure • Digital literacy
Sumner et al [27]; Singapore	HIC[c]	Personal	• Digital literacy among older adults • Privacy concerns	• Age • Data privacy concerns
Taha et al [28]; Egypt	LMIC	Personal or demographic	• Regional gaps in awareness and training • Limited accuracy of AI • Data privacy	• Geographic disparities • Low income and education levels

[a]AI: artificial intelligence.

[b]LMIC: low- and middle-income country.

[c]HIC: high-income country.

AI-driven medication adherence tools showed promise, with one study demonstrating improved adherence through personalized reminders and automated interventions, but digital literacy barriers among older adults and privacy concerns limited their effectiveness [27]. Infrastructure and accessibility challenges were key barriers to AI adoption. A study in Ethiopia found that 89.3% of surveyed pharmacists cited lack of internet availability, 88.2% lacked AI-related software and hardware, and 80.9% reported insufficient training. More than 90% emphasized the need for structured policies, better internet access, and AI-focused training to support AI integration in pharmacy services [24].

Regional disparities in AI awareness and adoption were evident. A study evaluating Egyptian pharmacists' perceptions of ChatGPT found that those in Greater Cairo showed the highest level of AI awareness, whereas those in South Upper Egypt had the lowest, reflecting geographic variations in AI exposure. Limited AI training further hindered equitable adoption and impacted AI-supported decision-making. Concerns about AI accuracy, data privacy, and bias reinforced skepticism toward AI as a standalone clinical tool, emphasizing the need for expert validation in pharmacy practice [28].

Involvement of HCPs

The involvement of HCPs in AI interventions varied across the included studies, with pharmacists being the primary group engaged. This aligns with the focus of our scoping review on their role in AI-driven medicines information. Five studies specifically involved pharmacists, either as primary users of AI [20,29] or as evaluators of AI-generated responses [22,25,30]. Three studies explored pharmacists' perceptions of AI, assessing their willingness to integrate AI into practice, perceived barriers to adoption, and concerns regarding accuracy, data privacy, and trust in AI-based recommendations [21,24,28]. These studies provided valuable insights into pharmacists' readiness to adopt AI-driven health care solutions while highlighting challenges related to infrastructure, regulatory frameworks, and professional training.

Four studies included multidisciplinary HCPs, such as physicians and nurses, in evaluating and validating AI interventions [17,18,26,27]. In contrast, 2 studies reported no direct involvement of HCPs in AI use or evaluation, as AI-generated responses were independently assessed by reviewers whose professional backgrounds were not specified [19,23].

Narrative Appraisal of Included Studies

Several studies relied on small or narrowly defined datasets [22,25], hypothetical scenarios without clinical validation [17,26], or educational and nonclinical settings [19,20,24,29], which limited their real-world applicability. Evaluations were frequently conducted internally, without independent or blinded assessors [21,30], and some lacked comparisons to human experts or assessments of clinical consequences [22,23]. In addition, poor reproducibility [30], limited topic scope [25], single-center study design [18], and nonsystematic prompt design were apparent [27,28].

Discussion

Principal Findings

This scoping review provided an overview of AI-driven tools in medicines information, focusing on HCP involvement, AI accuracy and completeness, and digital health inequalities. Some models accurately address general medication inquiries but struggle with complex clinical questions, raising concerns regarding reliability, data privacy, and usability. With respect to HCP involvement, studies highlighted the multidisciplinary engagement, particularly in medication adherence and discharge planning. Pharmacists played a key role in AI evaluation and integration. Regarding equity, the review highlighted disparities in infrastructure, digital literacy, and access to technology, which may hinder equitable implementation. One study examined digital health inequalities from the patient perspective.

Although a formal quality appraisal was not conducted, several methodological limitations were observed across the included

XSL•FO
RenderX

studies, which should be considered when interpreting the overall strength and applicability of the evidence.

Comparison With Prior Work

The included studies presented AI interventions used for medicines information, particularly in their potential applications in medication inquiries, adherence monitoring, and discharge planning. These findings align with previous reviews, which have explored the capacity of AI to support medication adherence through predictive models and real-time monitoring. AI-driven tools, including ML-driven monitoring systems, have been investigated for their ability to assess adherence patterns and identify at-risk patients; however, their effectiveness varies based on implementation and patient engagement [31]. Similarly, ML and statistical models have been explored for their potential in optimizing discharge planning by predicting patient outcomes, which may help reduce readmission rates and improve hospital efficiency [32]. Despite these promising applications, challenges remain regarding the accuracy, data privacy, and usability of AI-generated outputs. A recent systematic review highlighted that while AI has been explored in various aspects of medication use, its application in supporting clinical decision-making and complex, patient-specific inquiries remains limited [33]. Although AI models often demonstrate high accuracy when handling standard medication-related questions, their performance tends to decline in more complex clinical contexts, reinforcing the importance of human oversight. This review builds on these findings by specifically examining how AI is being applied to support medicines information services and evaluating the quality and focus of these emerging interventions.

The successful integration of AI into medication management and the provision of medicines information requires the active involvement of HCPs, particularly pharmacists, physicians, and nurses. The included studies indicate that pharmacists are key stakeholders in the adoption, evaluation, validation, and clinical implementation of AI tools. Similarly, a systematic review highlighted the potential of AI to enhance decision-making, emphasizing the need for expert validation to ensure clinical relevance and safety [34]. Furthermore, multidisciplinary engagement is essential for successful AI adoption; involving HCPs in AI development and validation improves usability and fosters trust. A narrative review identified key factors influencing trust in medical AI—such as explainability, transparency, and usability—reinforcing the importance of clinician involvement in ensuring effective decision-making [35].

Challenges remain in AI implementation, particularly in relation to digital health inequalities. Some included studies highlighted regional disparities in AI adoption, with pharmacists working in urban areas showing greater familiarity with AI tools compared to their counterparts working in rural settings [24,28]. While these studies provide valuable insights into pharmacists' access to and perceptions of digital tools, only one study specifically addressed digital health barriers from the patient perspective [27]. This highlights a key research gap in understanding how AI-driven pharmacy interventions affect patients, particularly vulnerable populations. Furthermore, limited AI-related training and infrastructure barriers, particularly in low-resource settings, hinder AI adoption [24], whereas the study from a high-income country emphasized individual-level concerns [27]. These findings align with broader research highlighting technological gaps and digital literacy disparities as significant barriers to AI adoption in medicines information services [36].

A key challenge in AI-generated medicines information involves potential risks related to data privacy, regulatory gaps, and ethical concerns. Two included studies reported inaccuracies or inconsistencies in AI-generated responses, raising concerns about their reliability and implications for patient safety [23,25]. This aligns with findings from a previous narrative review on AI governance, which highlighted the need for clear regulatory frameworks to support the adoption of AI in medicines information, particularly in addressing issues of bias, transparency, and accountability [37].

Strengths and Limitations

To the best of the authors' knowledge, this is the first scoping review to comprehensively map the use of AI for medicines information and evaluate its accuracy. A key strength is its systematic approach, adherence to established scoping review methodology, incorporating both qualitative and quantitative study designs.

Certain limitations should be acknowledged. The exclusion of gray literature, while consistent with our protocol and justified by the need to prioritize methodological rigor, may have introduced publication bias. Given the fast-evolving nature of AI research, new AI applications may be reported initially in non–peer-reviewed sources. Future updates of this review may benefit from a broader inclusion strategy or a targeted sensitivity search of these sources. Other limitations relate to the heterogeneity of included studies. The limited number of evaluated AI tools, small sample size, and small number of countries and within specific health care settings restrict the ability to draw generalizable conclusions. Inconsistencies were observed in how included studies defined and measured key outcomes such as "accuracy" and "completeness." Some studies assessed accuracy based on comparison to expert responses, while others used subjective scoring or evaluated only surface-level correctness. Similarly, the concept of "completeness" varied, ranging from the presence of keywords to the inclusion of all critical information. This variability complicates direct comparisons across studies and limits the ability to synthesize consistent conclusions regarding AI performance. Moreover, some application areas remain underexplored. Importantly, no formal quality appraisal of the included studies was performed, which may affect the ability to assess the robustness of individual findings—although this is consistent with established scoping review methodology.

Implications for Pharmacy Practice

This review highlights pharmacists as the most actively engaged HCPs in the development, evaluation, and implementation of AI tools for medicines information. Their clinical expertise, accessibility, and experience in medication counseling position them as key stakeholders in ensuring the safe and effective use

XSL•FO
RenderX

of AI-driven systems. Pharmacists require targeted training to critically evaluate AI-generated information, understand algorithmic limitations, and interpret outputs in clinical context. Embedding digital health and AI literacy into pharmacy education and continuing professional development will be critical to building confidence and capability in using these tools. Particular attention should be paid to pharmacists working in low-resource settings who lack access to digital health technologies.

Workflow integration also requires attention. AI systems should be incorporated seamlessly into pharmacy practice to enhance efficiency without disrupting established routines. This may involve embedding AI within electronic health records, clinical decision support systems, and patient counseling processes. Co-designing tools with pharmacists can ensure alignment with routine tasks and improve adoption.

Policy and governance considerations are equally important. Pharmacists should contribute actively to shaping standards for validation, data privacy, and accountability, especially in patient-facing contexts. Interdisciplinary collaborations will be essential as medicines information becomes increasingly digitized. Pharmacists are well-positioned to lead interdisciplinary teams in the development, use, and appraisal of AI applications and the information they generate.

Implication for Research

This review provides an overview of AI's role in medicines information; however, several gaps remain that warrant further investigation. Future research should focus on improving the accuracy and reliability of AI in handling complex clinical inquiries, as current models demonstrate inconsistencies in delivering clinically relevant recommendations. Equally important is the evaluation of the real-world effectiveness of AI in supporting medication adherence and discharge planning through personalized interventions.

Research should also assess the impact of digital health inequalities on AI adoption. In this review, only a small subset of studies explicitly examined aspects such as disparities in access to AI tools, language inclusivity, or representation of vulnerable populations. This limited evidence base restricts our ability to draw conclusions about the equity implications of AI-generated medicines information. Given that AI systems can perpetuate or even exacerbate existing health care disparities, future research should assess how factors such as access to digital technologies, digital literacy, and systemic bias influence access to and trust in AI tools, particularly in underresourced settings. Particular attention needs to be paid to inequalities that intersect with digital literacy skills, including older age, socioeconomic disadvantage, and people living in remote and rural areas. Frameworks such as the Digital Health Equity Framework [38] or the PROGRESS-Plus (Place of residence, Race/ethnicity/culture/language, Occupation, Gender/sex, Religion, Education, Socioeconomic status, and Social capital) lens [39] can provide conceptual guidance for equity-centered investigations. Future research should also explore ways to facilitate accessibility, affordability, and digital literacy skills from the perspectives of HCPs who practice in low-resource settings.

While this review identified pharmacists as key stakeholders in AI evaluation, further studies are needed to assess how AI influences clinical decision-making, workflow integration, and patient communication across various health care settings. Equally important is the integration of patient perspectives in the design and evaluation of AI tools. Despite their critical role as end users, patients were largely underrepresented in the included studies, indicating a need for more patient-centered research to ensure AI solutions are usable, trustworthy, and aligned with patient needs.

Conclusions

This scoping review highlights the early promise of AI in supporting medicines information, but its reliability, especially in complex clinical scenarios, remains uncertain. At present, AI should be viewed as a complementary tool rather than a replacement for human expertise, with HCPs playing a critical role in its evaluation and integration. Key barriers such as digital health inequalities, ethical concerns, and regulatory gaps must be addressed to enable the safe and equitable implementation. Future research should focus on identifying and mitigating digital health inequalities in this context, including intersectional disadvantages as well as improving evidence-based use of AI, expanding training for HCPs, and developing robust governance frameworks to support standardized AI integration into clinical practice.

Acknowledgments

The authors gratefully acknowledge the library and database resources provided by University College London, which supported the literature search and development of this review.

Funding

No financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Authors' Contributions

Conceptualization: SA (lead), VP (equal)
Methodology: SA (lead), VP (equal), JF (supporting)
Formal analysis: SA (lead), VP (equal)

XSL•FO
RenderX

Writing – original draft: SA (lead), VP (equal)
Writing – review & editing: SA (lead), VP (equal), JF (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy (PRESS) Format.
[PDF File, 86 KB - jmir_v28i1e77747_app1.pdf]

Checklist 1
PRISMA-ScR guidelines.
[PDF File, 61 KB - jmir_v28i1e77747_app2.pdf]

References

1. McCarthy J. What Is Artificial Intelligence: MIT Press; 2007. URL: https://www-formal.stanford.edu/jmc/whatisai.pdf
2. Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. J Family Med Prim Care 2019 Jul;8(7):2328-2331. [doi: 10.4103/jfmpc.jfmpc_440_19] [Medline: 31463251]
3. What is machine learning? IBM. URL: https://www.ibm.com/think/topics/machine-learning [accessed 2025-04-04]
4. Abbasgholizadeh Rahimi S, Légaré F, Sharma G, et al. Application of artificial intelligence in community-based primary health care: systematic scoping review and critical appraisal. J Med Internet Res 2021 Sep 3;23(9):e29839. [doi: 10.2196/29839] [Medline: 34477556]
5. Damiani G, Altamura G, Zedda M, et al. Potentiality of algorithms and artificial intelligence adoption to improve medication management in primary care: a systematic review. BMJ Open 2023 Mar 23;13(3):e065301. [doi: 10.1136/bmjopen-2022-065301] [Medline: 36958780]
6. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
7. Singareddy S, Sn VP, Jaramillo AP, et al. Artificial intelligence and its role in the management of chronic medical conditions: a systematic review. Cureus 2023 Sep;15(9):e46066. [doi: 10.7759/cureus.46066] [Medline: 37900468]
8. Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J Ambient Intell Humaniz Comput 2023;14(7):8459-8486. [doi: 10.1007/s12652-021-03612-z] [Medline: 35039756]
9. Abbasgholizadeh Rahimi S, Cwintal M, Huang Y, et al. Application of artificial intelligence in shared decision making: scoping review. JMIR Med Inform 2022 Aug 9;10(8):e36199. [doi: 10.2196/36199] [Medline: 35943793]
10. Choudhury A, Asan O. Role of artificial intelligence in patient safety outcomes: systematic literature review. JMIR Med Inform 2020 Jul 24;8(7):e18599. [doi: 10.2196/18599] [Medline: 32706688]
11. The pursuit of responsible use of medicines: sharing and learning from country experiences. World Health Organization. 2012. URL: https://www.who.int/publications/i/item/WHO-EMP-MAR-2012.3 [accessed 2025-04-17]
12. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
13. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5(1):69. [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
14. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
15. Al-Arkee S, Falade J, Paudyal V. Artificial intelligence for medicines information: scoping review protocol. URL: https://osf.io/p3tgs [accessed 2025-04-06]
16. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 Guideline Statement. J Clin Epidemiol 2016 Jul;75:40-46. [doi: 10.1016/j.jclinepi.2016.01.021] [Medline: 27005575]
17. Andrikyan W, Sametinger SM, Kosfeld F, et al. Artificial intelligence-powered chatbots in search engines: a cross-sectional study on the quality and risks of drug information for patients. BMJ Qual Saf 2025 Jan 28;34(2):100-109. [doi: 10.1136/bmjqs-2024-017476] [Medline: 39353736]
18. Beavers J, Schell RF, VanCleave H, et al. Evaluation of inpatient medication guidance from an artificial intelligence chatbot. Am J Health Syst Pharm 2023 Dec 5;80(24):1822-1829. [doi: 10.1093/ajhp/zxad193] [Medline: 37611187]
19. Cornelison BR, Erstad BL, Edwards C. Accuracy of a chatbot in answering questions that patients should ask before taking a new medication. J Am Pharm Assoc (2003) 2024;64(4):102110. [doi: 10.1016/j.japh.2024.102110] [Medline: 38670493]
20. Grossman S, Zerilli T, Nathan JP. Appropriateness of ChatGPT as a resource for medication-related questions. Br J Clin Pharmacol 2024 Oct;90(10):2691-2695. [doi: 10.1111/bcp.16212] [Medline: 39096130]

XSL•FO
RenderX

21. Laymouna M, Ma Y, Lessard D, et al. Needs-assessment for an artificial intelligence-based chatbot for pharmacists in HIV care: results from a knowledge-attitudes-practices survey. Healthcare (Basel) 2024 Aug 20;12(16):1661. [doi: 10.3390/healthcare12161661] [Medline: 39201222]
22. Munir F, Gehres A, Wai D, Song L. Evaluation of ChatGPT as a tool for answering clinical questions in pharmacy practice. J Pharm Pract 2024 Dec;37(6):1303-1310. [doi: 10.1177/08971900241256731] [Medline: 38775367]
23. Ramasubramanian S, Balaji S, Kannan T, et al. Comparative evaluation of artificial intelligence systems' accuracy in providing medical drug dosages: a methodological study. World J Methodol 2024 Dec 20;14(4):92802. [doi: 10.5662/wjm.v14.i4.92802] [Medline: 39712564]
24. Sendekie AK, Limenh LW, Abate BB, et al. Artificial intelligence in community pharmacy practice: pharmacists' perceptions, willingness to utilize, and barriers to implementation. Explor Res Clin Soc Pharm 2024 Dec;16:100542. [doi: 10.1016/j.rcsop.2024.100542] [Medline: 39687445]
25. Sheikh MS, Barreto EF, Miao J, et al. Evaluating ChatGPT's efficacy in assessing the safety of non-prescription medications and supplements in patients with kidney disease. Digit Health 2024;10:20552076241248082. [doi: 10.1177/20552076241248082] [Medline: 38638404]
26. Stanceski K, Zhong S, Zhang X, et al. The quality and safety of using generative AI to produce patient-centred discharge instructions. NPJ Digit Med 2024 Nov 20;7(1):329. [doi: 10.1038/s41746-024-01336-w] [Medline: 39567722]
27. Sumner J, Bundele A, Lim HW, Phan P, Motani M, Mukhopadhyay A. Developing an artificial intelligence-driven nudge intervention to improve medication adherence: a human-centred design approach. J Med Syst 2023 Dec 8;48(1):3. [doi: 10.1007/s10916-023-02024-0] [Medline: 38063940]
28. Taha T, Abdel-Qader DH, Alamiry KR, Fadl ZA, Alrawi A, Abdelsattar NK. Perception, concerns, and practice of ChatGPT among Egyptian pharmacists: a cross-sectional study in Egypt. BMC Health Serv Res 2024 Nov 28;24(1):1500. [doi: 10.1186/s12913-024-11815-1] [Medline: 39609697]
29. Triplett S, Ness-Engle GL, Behnen EM. A comparison of drug information question responses by a drug information center and by ChatGPT. Am J Health Syst Pharm 2025 Apr 17;82(8):448-460. [doi: 10.1093/ajhp/zxae316] [Medline: 39450858]
30. van Nuland M, Lobbezoo AFH, van de Garde EMW, et al. Assessing accuracy of ChatGPT in response to questions from day to day pharmaceutical care in hospitals. Explor Res Clin Soc Pharm 2024 Sep;15:100464. [doi: 10.1016/j.rcsop.2024.100464] [Medline: 39050145]
31. Bohlmann A, Mostafa J, Kumar M. Machine learning and medication adherence: scoping review. JMIRx Med 2021 Nov 24;2(4):e26993. [doi: 10.2196/26993] [Medline: 37725549]
32. Pahlevani M, Taghavi M, Vanberkel P. A systematic literature review of predicting patient discharges using statistical methods and machine learning. Health Care Manag Sci 2024 Sep;27(3):458-478. [doi: 10.1007/s10729-024-09682-7] [Medline: 39037567]
33. Khosravi M, Zare Z, Mojtabaeian SM, Izadi R. Artificial intelligence and decision-making in healthcare: a thematic analysis of a systematic review of reviews. Health Serv Res Manag Epidemiol 2024;11:23333928241234863. [doi: 10.1177/23333928241234863] [Medline: 38449840]
34. Ahmed MI, Spooner B, Isherwood J, Lane M, Orrock E, Dennison A. A systematic review of the barriers to the implementation of artificial intelligence in healthcare. Cureus 2023 Oct;15(10):e46454. [doi: 10.7759/cureus.46454] [Medline: 37927664]
35. Tucci V, Saary J, Doyle TE. Factors influencing trust in medical artificial intelligence for healthcare professionals: a narrative review. J Med Artif Intell 2021;5:4-4. [doi: 10.21037/jmai-21-25]
36. Zuhair V, Babar A, Ali R, et al. Exploring the impact of artificial intelligence on global health and enhancing healthcare in developing nations. J Prim Care Community Health 2024;15:21501319241245847. [doi: 10.1177/21501319241245847] [Medline: 38605668]
37. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: a narrative review. Heliyon 2024 Feb 29;10(4):e26297. [doi: 10.1016/j.heliyon.2024.e26297] [Medline: 38384518]
38. Rowland S, Brewer LC, Rosas LG. Digital health equity - a call to action for clinical and translational scientists. J Clin Transl Sci 2024;8(1):e145. [doi: 10.1017/cts.2024.564] [Medline: 39478785]
39. O'Neill J, Tabish H, Welch V, et al. Applying an equity lens to interventions: using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health. J Clin Epidemiol 2014 Jan;67(1):56-64. [doi: 10.1016/j.jclinepi.2013.08.005] [Medline: 24189091]

Abbreviations

AI: artificial intelligence
HCP: health care professional
IPA: International Pharmaceutical Abstracts
ML: machine learning
NLP: natural language processing
PRESS: Peer Review of Electronic Search Strategies

XSL•FO
RenderX

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
PROGRESS-Plus: Place of residence, Race/ethnicity/culture/language, Occupation, Gender/sex, Religion, Education, Socioeconomic status, and Social capital

Edited by A Coristine, TDA Cardoso; submitted 19.May.2025; peer-reviewed by N Acharya, P Sharma, Y Olanlokun; revised version received 27.Dec.2025; accepted 27.Dec.2025; published 06.Mar.2026.

Please cite as:
Al-Arkee S, Falade J, Paudyal V
Artificial Intelligence for Medicines Information: Scoping Review of Clinical Applications and Digital Health Inequalities
J Med Internet Res 2026;28:e77747
URL: https://www.jmir.org/2026/1/e77747
doi:10.2196/77747

© Shahd Al-Arkee, Josephine Falade, Vibhu Paudyal. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Machine Learning in Left Ventricular Hypertrophy Detection: Systematic Review and Meta-Analysis

Yilin Li[1], MD; Ke Zhao[2], MM; Jing Wu[1], BS

[1]Department of Geriatrics, The Third People's Hospital of Chengdu, 82 Qinglong Street, Qingyang District, Chengdu, Sichuan Province, China, 610031, Chengdu, Sichuan, China

[2]Department of Critical Care Medicine, Affiliated Hospital of North Sichuan Medical College, Nanchong, Sichuan, China

Corresponding Author:
Jing Wu, BS
Department of Geriatrics, The Third People's Hospital of Chengdu, 82 Qinglong Street, Qingyang District, Chengdu, Sichuan Province, China, 610031, Chengdu, Sichuan, China

Abstract

Background: In recent years, researchers have investigated machine learning (ML)–based approaches for the detection of left ventricular hypertrophy (LVH). However, the accuracy of ML in detecting LVH varies across different modeling variables and models. Systematic evidence is lacking in understanding how different ML approaches affect LVH detection accuracy.

Objective: The aim of this study is to systematically assess the diagnostic accuracy of these ML approaches to inform the development of artificial intelligence tools.

Methods: PubMed, Embase, Cochrane Library, and Web of Science were comprehensively searched up to November 12, 2025. The Prediction Model Risk of Bias Assessment Tool was used to evaluate the risk of bias. Subgroup analyses were performed based on ML model types and modeling variables (electrocardiogram [ECG], clinical features, and echocardiography). Only diagnostic 2×2 tables from validation sets were pooled for meta-analysis, with all statistical analyses performed using Stata.

Results: A total of 25 studies were included in the analysis. The performance of ML models varied with input data types and algorithms. A meta-analysis showed that ECG-based models, in comparison, exhibited a sensitivity of 0.76 (95% CI 0.66 - 0.84) and a specificity of 0.84 (95% CI 0.78 - 0.89). Echocardiography-based models had a sensitivity ranging from 0.71 to 0.94 and a specificity ranging from 0.67 to 0.96. The models based on clinical features had a sensitivity of 0.78 (95% CI 0.69 - 0.85) and a specificity of 0.71 (95% CI 0.65 - 0.76). A subgroup analysis of the ECG-based models revealed that the deep learning model produced a sensitivity of 0.71 (95% CI 0.60 - 0.80) and a specificity of 0.79 (95% CI 0.65 - 0.88).

Conclusions: ML demonstrates reasonably high accuracy in detecting LVH. However, these conclusions are derived from limited evidence. Meanwhile, the extreme heterogeneity reported in the meta-analysis requires more critical interpretation. Current conclusions regarding model accuracy should be interpreted with caution. Therefore, future research should focus on constructing high-performance ML models based on imaging data for LVH diagnosis.

(J Med Internet Res 2026;28:e76637) doi:10.2196/76637

KEYWORDS

artificial intelligence; AI; machine learning; left ventricular hypertrophy; meta-analysis; deep learning; electrocardiogram; ECG; echocardiography; cardiovascular risk

Introduction

Left ventricular hypertrophy (LVH) represents an increase in left ventricular mass driven by various cardiovascular risk factors [1], and it confers a 5 - 10 times greater risk of developing cardiovascular disease [2,3]. Accumulating evidence demonstrates that LVH is an independent predictor of cardiovascular risk [4-6]. However, the pathological and physiological alterations induced by LVH occur without overt symptoms or clinical signs [1]. Therefore, the early detection of LVH through sensitive screening methods is imperative, as it holds significant positive implications for patient health outcomes.

Currently, echocardiography (ECHO) and electrocardiogram (ECG) are the conventional diagnostic methods for detecting LVH [7,8]. While ECHO is a relatively accurate approach [9], it is costly and requires specialized equipment and trained operators, which poses challenges for early diagnosis [9,10]. Studies also show high rates of disagreement in ECHO readings among cardiologists, sometimes as much as 42.1% [11]. Cardiologists may mistakenly read ECHOs, particularly when tired, possibly resulting in misdiagnoses and misinterpretations [12,13]. Machine learning (ML) has garnered increasing

XSL•FO
RenderX

attention in the medical field over the past decade, particularly in cardiovascular medicine [14-16]. ML is important since it can discern critical features from complex datasets [17].

Given its capacity to process intricate, high-volume health data, ML has ascended as an indispensable tool for developing predictive models in medicine [17,18]. In this context, researchers have begun constructing ML-based approaches for diagnosing LVH.

Nonetheless, there are various models and modeling variables for constructing ML models [17,19,20]. Multiple ML models have been proposed for detecting LVH [1,10,21,22], yet their comparative diagnostic accuracy remains debated. Robust evidence assessing both the accuracy and real-world implementation of ML models for LVH identification remains insufficient. Therefore, a systematic review and meta-analysis was conducted in this study to evaluate the performance of ML in detecting LVH and furnish empirical support for artificial intelligence (AI) development in this field.

Methods

Study Registration

This study followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines and was prospectively registered in PROSPERO (registration ID: CRD42024617183) [23].

Eligibility Criteria

Detailed inclusion and exclusion criteria were formulated for this review. To improve the visualization effect, these standards were presented in tabular form (Textbox 1).

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Population: individuals suspected of having left ventricular hypertrophy
- Intervention: machine learning (ML) models for left ventricular hypertrophy identification
- Comparison: no diagnostic tool was used as a comparator
- Outcome: quantitative diagnostic performance metrics, including C-statistic, sensitivity, specificity, accuracy, recall, precision, confusion matrices, diagnostic 2×2 tables, F_1-score, or calibration curves
- Study design: case-control, cohort, or cross-sectional studies published in English

Exclusion criteria

- Intervention: studies that only examined differential factors without developing complete ML models
- Outcomes: studies lacking the following metrics for ML performance: C-statistic, sensitivity, specificity, accuracy, recall, precision, confusion matrices, diagnostic 2×2 tables, F_1-score, or calibration curves
- Study design: meta-analyses, reviews, guidelines, expert opinions, and unpublished conference abstracts

Data Sources and Search Strategy

PubMed, Embase, Cochrane Library, and Web of Science were systematically searched up to November 12, 2025, using both Medical Subject Headings (MeSH) terms and free-text words. Search terms included combinations of “artificial intelligence,” “deep learning,” “machine learning,” “random forest,” “support vector machine,” and “left ventricular hypertrophy.” No filters were applied to constrain studies by publication year or geographic region. The details on the search strategy are available in Multimedia Appendix 1.

Study Selection and Data Extraction

All identified records were uploaded to EndNote, and duplicates were filtered out. The titles and abstracts of the remaining studies were screened for potentially relevant studies. Full texts of the selected studies were obtained and evaluated for eligibility. An electronic extraction sheet was prepared in advance to record author information, publication year, country of the author, study design, patient sources, generation method of validation set, model types, and modeling variables. The entire screening process and data extraction were performed independently by 2 investigators, followed by cross-verification. Discrepancies were addressed through discussion with a third researcher.

Risk of Bias in Studies

The Prediction Model Risk of Bias Assessment Tool (PROBAST) was leveraged to appraise the risk of bias (RoB) in the included original studies [24]. The evaluation covers 4 domains: participants, predictor variables, outcomes, and statistical analysis, which collectively determine the overall RoB and applicability. Each domain contains specific signaling questions (2, 3, 6, and 9 questions, respectively), with 3 possible responses: “yes/probably yes” (low RoB), “no or probably no” (high RoB), and “no information” (unclear). Two researchers independently conducted the PROBAST assessments, followed by cross-checking. Any dissents were addressed through discussion with a third researcher.

Synthesis Methods

Some studies had multiple validation sets. When both an internal and an external validation set were available, only the results from the external set were retained. If an external validation set is unavailable, the results from the internal validation set with optimal performance were selected.

XSL•FO
RenderX

A meta-analysis was conducted on the validation set, leveraging diagnostic 2×2 tables. For studies without 2×2 tables, sensitivity, specificity, positive predictive value, and accuracy, combined with case numbers, were used for estimation (formulas 1 - 5). The bivariate mixed-effects model was used to pool diagnostic metrics (sensitivity, specificity, positive likelihood ratio [PLR], negative likelihood ratio [NLR], diagnostic odds ratio [DOR], and summary receiver operating characteristic [SROC]). Deek funnel plot was used for assessing publication bias [25], and Fagan nomogram was adopted for evaluating clinical utility [26]. Subgroup analyses were carried out based on modeling variables and model types. A significance threshold of $P<.05$ was established.

TP=Sensitivity×Events (1)

TN=Specificity×(sample size − Events) (2)

FN=Events-TP (3)

FP=samplesize-Events-TN (4)

OR

FP=TP/precision-TP (5)

Results

Study Selection

A total of 1382 records were searched. After removing duplicates, 1086 records remained. Then, 1019 records were excluded through title and abstract screening, leaving 67 records for full-text review. Of those, 42 reports were removed due to incomplete data, incompatible outcomes, and incompatible interventions. Ultimately, 25 studies were selected for analysis (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart for literature screening.

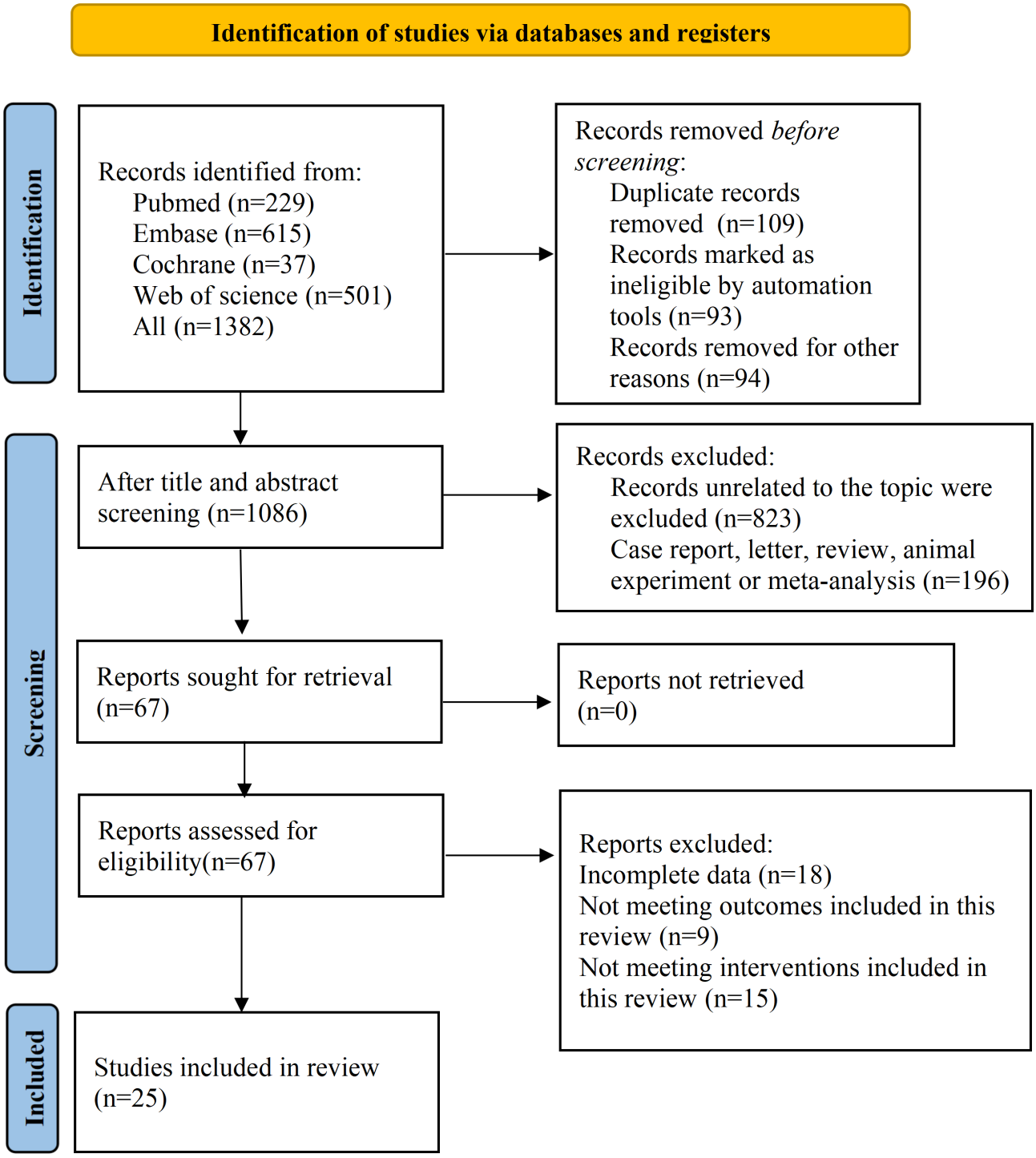

Study Characteristics

Of the 25 case-control studies included (2018 - 2025) [1,10,21,22,27-47], 19 used single-center designs, and 6 adopted multicenter approaches (including 4 studies based on databases that leveraged the UK Biobank resource). These studies were conducted across diverse regions: 10 studies in China, 6 in the United States, 2 in Japan, 2 in Korea, 1 in Mexico, 1 in the United Kingdom, 1 in Greece, 1 in Italy, and 1 in the United Arab Emirates. ECG-based models were included in 16 studies, clinical feature-based models were included in 5 studies, and ECHO-based models were included in 4 studies. Six studies performed external validation, including 5 ECG-based models and 1 clinical feature-based model. The main features of these studies are summarized in Table 1.

XSL•FO
RenderX

Table . Basic features of the included studies.

Author and publication year	Country of author	Study design	Patient source	Validation set generation method	Model type	Modeling variable	Reference standard
Ryu et al [22] (2023)	Korea	Case-control study	Single center	RS[a]	DL[b]	12-lead ECG[c]	ECHO[d]
Farhad et al [28] (2023)	United Arab Emirates	Case-control study	LVH[e] database	RS 10 folds	DL	ECHO	Clinical consensus
Zhao et al [32] (2022)	China	Case-control study	Single center	Folds RS	DL	12-lead ECG	Clinical consensus
Liu et al [33] (2022)	China	Case-control study	Multicenter	Folds EV[f]	DL	12-lead ECG	Clinical consensus
De la Garza Salazar et al [45] (2021)	Mexico	Case-control study	Single center	RS EV	DT[g]	12-lead ECG	Clinical consensus
Kwon et al [40] (2020)	Korea	Case-control study	Multicenter	RS EV	DL	12-lead ECG	Clinical consensus
Ghorbani et al [39] (2020)	United States	Case-control study	Single center	RS	DL	ECHO	Clinical consensus
Sparapani et al [43] (2019)	United States	Case-control study	The MESA[h] database	RS	DT	12-lead ECG	CMR[i]
Madani et al [44] (2018)	United States	Case-control study	Single center	RS	DL	ECHO	Clinical consensus
Yuan et al [29] (2023)	China	Case-control study	Single center	RS	LR[j]	Clinical features	Clinical consensus
Dwivedi et al [30] (2023)	United States	Case-control study	Single center	RS	RF[k]	12-lead ECG	ECHO
Yu et al [35] (2022)	China	Case-control study	Single center	RS	DL	ECHO	Clinical consensus
Wu et al [34] (2022)	China	Case-control study	Single center	EV	LR	Clinical features	Clinical consensus
Kokubo et al [36] (2022)	Japan	Case-control study	Single center	RS	DL	12-lead ECG	Clinical consensus
Ye et al [37] (2021)	China	Case-control study	Single center	10 folds[l]	LASSO[m] regression	Clinical features	Clinical consensus
Angelaki et al [38] (2021)	Greece	Case-control study	Single center	RS	RF	12-lead ECG	Clinical consensus
Lin and Liu [41] (2020)	United States	Case-control study	Single center	RS 4 folds	SVM[o]	12-lead ECG	ECHO
Tison et al [42] (2019)	United States	Case-control study	Single center	RS	GBM[n]	12-lead ECG	ECHO
Liu et al [31] (2023)	China	Case-control study	Single center	RS	ANN[p]	12-lead ECG	Clinical consensus
Zhang et al [1] (2024)	China	Case-control study	Single center	RS	LR	Clinical features	Clinical consensus
Wan et al [27] (2024)	China	Case-control study	Single center	RS	LR	Clinical features	Clinical consensus
Naderi et al [21] (2024)	United Kingdom	Case-control study	UK Biobank Database+ Single center	RS EV	SVM	12-lead ECG	CMR
Cai et al [10] (2024)	Japan	Case-control study	Single center	RS	DL	12-lead ECG	Clinical consensus
Huang et al [46] (2025)	China	Case-control study	Single center	RS	CatBoost	12-lead ECG	Clinical consensus

XSL•FO
RenderX

Author and publication year	Country of author	Study design	Patient source	Validation set generation method	Model type	Modeling variable	Reference standard
Taconné et al [47] (2025)	Italy	Case-control study	PTB-XL ECG database+ Georgia 12-lead ECG Challenge Database	10 folds EV	SVM	12-lead ECG	Clinical consensus

[a]RS: random sampling.
[b]DL: deep learning.
[c]ECG: electrocardiography.
[d]ECHO: echocardiography.
[e]LVH: left ventricular hypertrophy.
[f]EV: external validation.
[g]DT: decision tree.
[h]MESA: Multi-Ethnic Study of Atherosclerosis.
[i]CMR: cardiac magnetic resonance.
[j]LR: logistic regression.
[k]RF: random forest.
[l]10 folds: 10-fold cross-validation.
[m]LASSO: least absolute shrinkage and selection operator.
[n]GBM: gradient boosted machine.
[o]SVM: support vector machine.
[p]ANN: artificial neural network.

RoB in Studies

A high RoB was identified in 21 case-control studies (single- and multi-center) due to potential selection bias. An unclear RoB was assigned to 4 studies based on public databases. The 21 case-control studies were rated as having a high RoB due to selection bias. Due to inherent selection biases, such as the "healthy volunteer" effect commonly found in biobanks, the 4 studies based on public databases were rated as having an unclear RoB. Five case-control studies utilized clinical features for modeling and were rated as having a high RoB due to case-control influence during data collection. The remaining studies, which used image-based ML models, were assessed as having a low RoB in terms of predictors. All studies used valid outcome definitions, and the avoidance of predictor variables in model construction ensured a low RoB. For the imaging-based validation set, the accurate calculation of events per variable (EPV) proved challenging, leading to an uncertain RoB. Developers typically rely on EPV, especially EPV of 10 higher, to determine the minimum sample size required and the maximum number of candidate predictions that can be tested [48]. However, when building models based on images, the basic unit of validation data is the image, not the traditional "patient case." Therefore, we cannot calculate the sample size that meets the EPV requirements based on the number of patient cases, making it difficult to directly apply this standard in such studies. In contrast, studies using clinical features for modeling consistently met the EPV>10 criterion and thus were classified as low RoB. In 2 studies, data were segmented according to images, and patients' images were shared between the training set and the validation set, which may lead to data leakage, and thus were classified as high RoB (Figure 2).

XSL•FO
RenderX

Figure 2. Risk of bias of the included machine learning models.

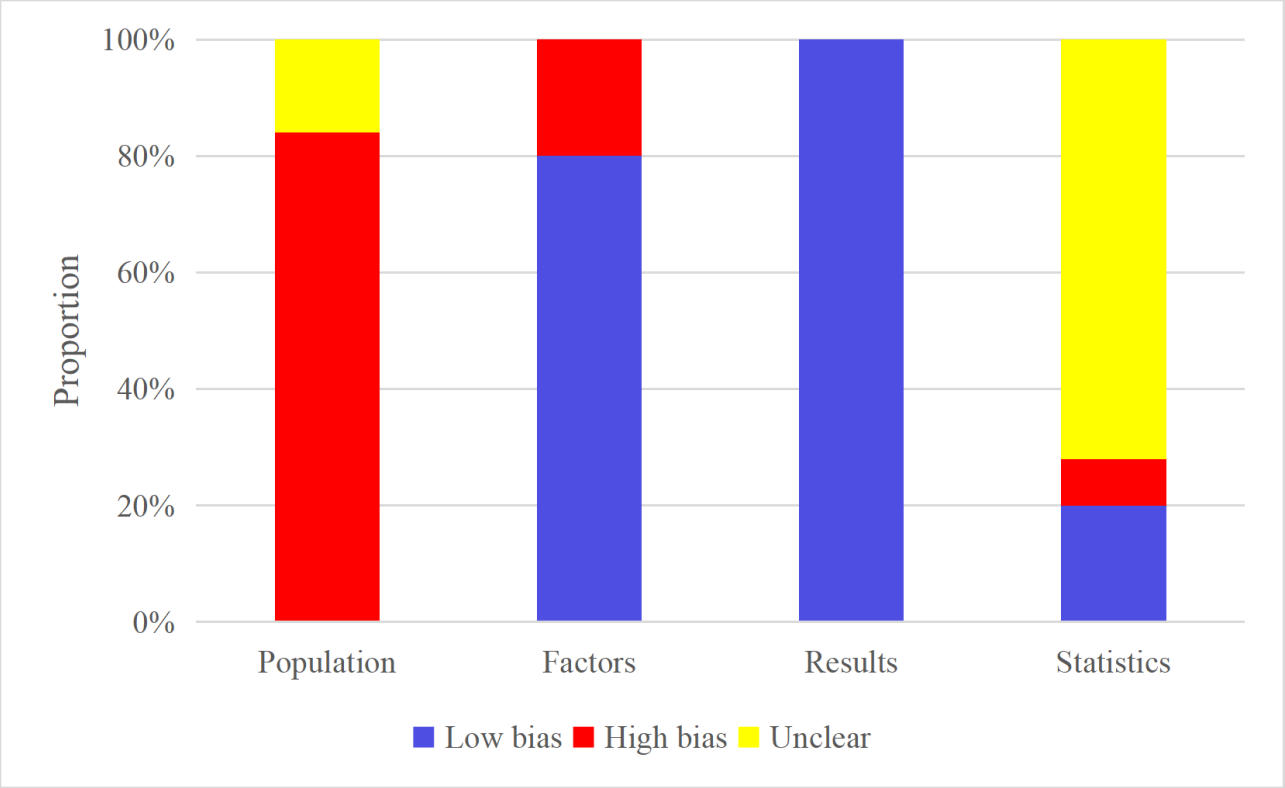

Meta-Analysis

ECG-Based ML Model for LVH Detection

In the validation set, the performance of ECG-based ML models was evaluated in 16 diagnostic 2×2 tables. The meta-analysis results demonstrated the following estimates: sensitivity 0.76 (95% CI 0.66 - 0.84), specificity 0.84 (95% CI 0.78 - 0.89), PLR 4.8 (95% CI 3.3 - 7.0), NLR 0.28 (95% CI 0.19 - 0.41), DOR 17 (95% CI 9 - 33), and SROC 0.88 (95% CI 0.75 - 0.95; Figure 3 and Figure S1 in Multimedia Appendix 2). The Deek funnel plot showed no significant publication bias among the included studies (*P*=.43; Figure S2 in Multimedia Appendix 2). Given a pretest high-risk probability of 10%, if the model result indicated LVH, the posterior probability of LVH was 35%; if the model result indicated non-LVH, the posterior probability of LVH was 3% (Figure S3 in Multimedia Appendix 2).

XSL•FO
RenderX

Figure 3. Forest plot of sensitivity and specificity for electrocardiogram-based models [10,21,22,30-33,36,38,40-43,45-47]. ANN: artificial neural network; DT: decision tree; GBM: gradient boosting machine; RF: random forest; SVM: support vector machine.

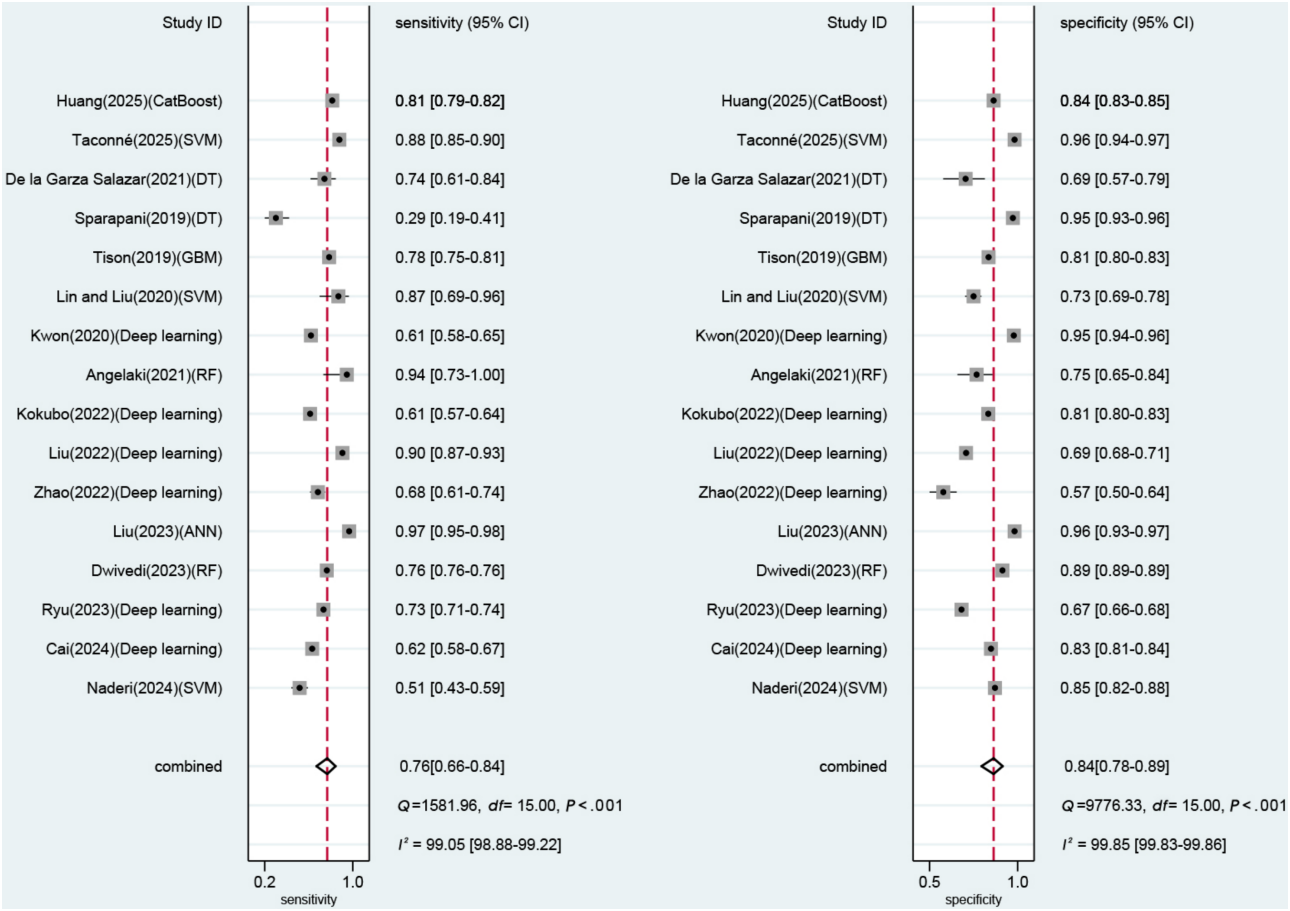

Subgroup analyses were performed across various ML approaches. The deep learning (DL) model produced a sensitivity of 0.71 (95% CI 0.60 - 0.80) and a specificity of 0.79 (95% CI 0.65 - 0.88).

Subgroup analyses were performed across various validation approaches. The results showed that internal validation produced a sensitivity of 0.77 (95% CI 0.64 - 0.86) and a specificity of 0.83 (95% CI 0.75 - 0.89), while external validation produced a sensitivity of 0.76 (95% CI 0.60 - 0.87) and a specificity of 0.87 (95% CI 0.74 - 0.94; Table 2).

Table . Subgroup analyses of electrocardiography (ECG)–based machine learning (ML) models across various validation approaches.

Methods	Machine learning approaches, n	Sen[a] (95% CI)	Spe[b] (95% CI)	PLR[c] (95% CI)	NLR[d] (95% CI)	DOR[e] (95% CI)	SROC[f] (95% CI)
IV[g]	11	0.77 (0.64 - 0.86)	0.83 (0.75 - 0.89)	4.5 (2.9 - 6.8)	0.28 (0.17 - 0.46)	16 (7-35)	0.87 (0.74 - 0.94)
EV[h]	5	0.76 (0.60 - 0.87)	0.87 (0.74 - 0.94)	5.8 (2.8 - 12.2)	0.28 (0.16 - 0.48)	21 (7-60)	0.89 (0.52 - 0.98)
Overall	16	0.76 (0.66 - 0.84)	0.84 (0.78 - 0.89)	4.8 (3.3 - 7.0)	0.28 (0.19 - 0.41)	17 (9-33)	0.88 (0.75 - 0.95)

[a]Sen: sensitivity.

[b]Spe: specificity.

[c]PLR: positive likelihood ratio.

[d]NLR: negative likelihood ratio.

[e]DOR: diagnostic odds ratio.

[f]SROC: summary receiver operating characteristic.

[g]IV: internal validation.

[h]EV: external validation.

XSL•FO
RenderX

As for internal validation, the meta-analysis results demonstrated an SROC of 0.87 (95% CI 0.74 - 0.94; Figure S4 in Multimedia Appendix 2). The Deek funnel plot showed no significant publication bias among the included studies (*P*=.30; Figure S5 in Multimedia Appendix 2). Given a pretest high-risk probability of 10%, if the model result indicated LVH, the posterior probability of LVH was 33%; if the model result indicated non-LVH, the posterior probability of LVH was 3% (Figure S6 in Multimedia Appendix 2).

As for external validation, the meta-analysis results showed an SROC of 0.89 (95% CI 0.52 - 0.98; Figure S7 in Multimedia Appendix 2). The Deek funnel plot showed no significant publication bias across the included studies (*P*=.37; Figure S8 in Multimedia Appendix 2). Given a pretest high-risk probability of 10%, if the model result indicated LVH, the posterior probability of LVH was 39%; if the model result indicated non-LVH, the posterior probability of LVH was 3% (Figure S9 in Multimedia Appendix 2).

Dwivedi et al [30] used image-level splitting, which involved sharing patient images between training and validation. This method posed a high RoB due to data leakage. Accordingly, a sensitivity analysis was conducted on the ECG-based ML model group to exclude this study. The meta-analysis results demonstrated the following estimates: sensitivity 0.76 (95% CI 0.66 - 0.85), specificity 0.84 (95% CI 0.77 - 0.89), PLR 4.7 (95% CI 3.2 - 7.0), NLR 0.28 (95% CI 0.19 - 0.42), DOR 17 (95% CI 8 - 33), and SROC 0.88 (95% CI 0.74 - 0.94).

Clinical Feature–Based ML Model for LVH Detection

In the validation set, there were 5 diagnostic 2×2 tables for evaluating the performance of ML models based on clinical features. The meta-analysis results demonstrated the following estimates: sensitivity 0.78 (95% CI 0.69 - 0.85), specificity 0.71 (95% CI 0.65 - 0.76), PLR 2.7 (95% CI 2.1 - 3.4), NLR 0.31 (95% CI 0.21 - 0.46), DOR 9 (95% CI 5 - 16), and SROC 0.79 (95% CI 0.62 - 0.90; Figure 4 and Figure S10 in Multimedia Appendix 2). The Deek funnel plot showed no significant publication bias among the included studies (*P*=.09; Figure S11 in Multimedia Appendix 2). Given a pretest high-risk probability of 10%, if the model result indicated LVH, the posterior probability of LVH was 23%; if the model result indicated non-LVH, the posterior probability of LVH was 3% (Figure S12 in Multimedia Appendix 2).

Figure 4. Forest plot of sensitivity and specificity for clinical feature-based models [1,27,29,34,37]. LASSO: least absolute shrinkage and selection operator; LR: logistic regression.

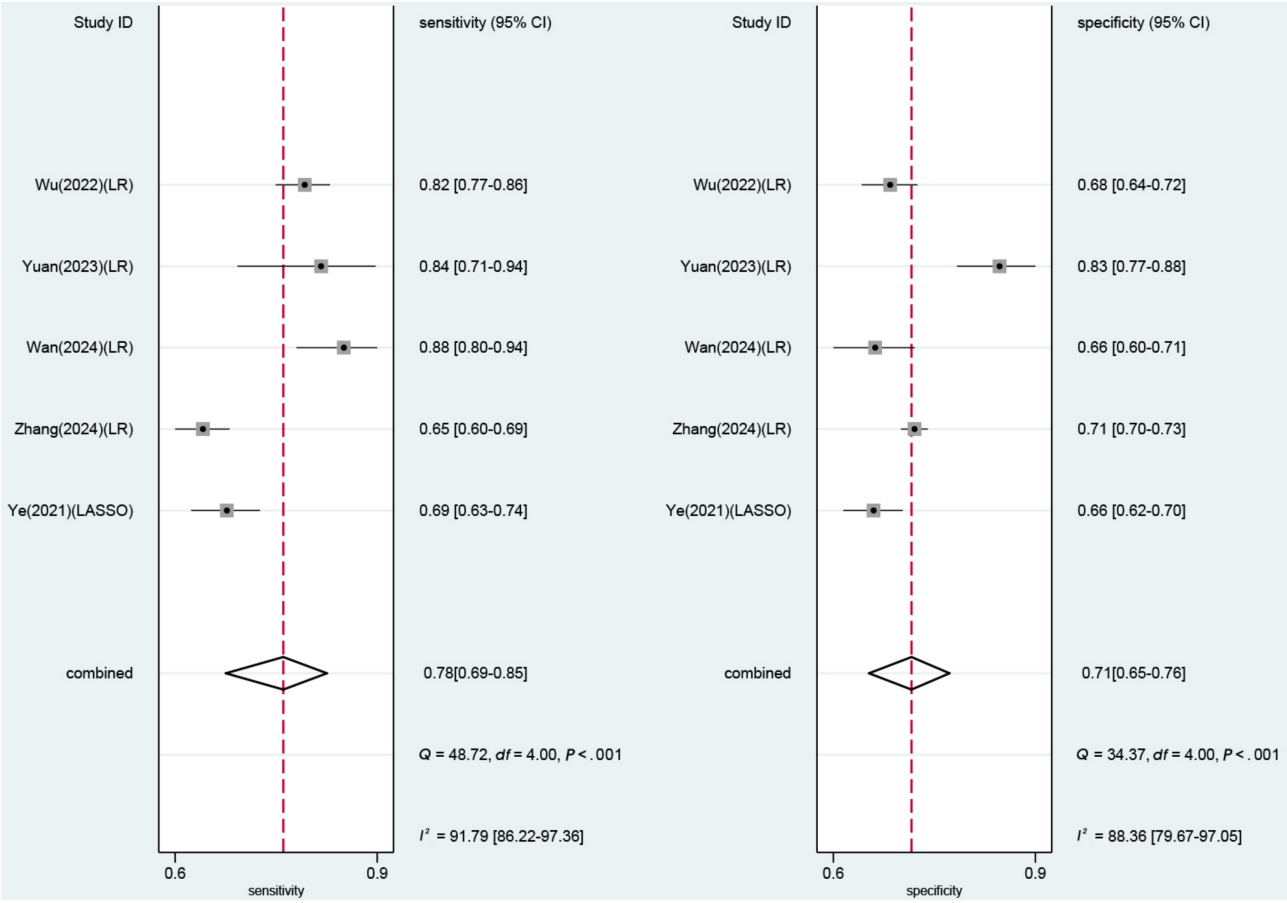

ECHO-Based ML Model for LVH Detection

Farhad et al [28] used image-level splitting, which involved sharing patient images between training and validation. This method posed a high RoB due to data leakage. Accordingly, a sensitivity analysis was conducted in the ECHO-based ML model group to exclude this study. Ghorbani et al [39] developed a DL model based on ECHO. In their validation set of 373 cases (including 142 LVH cases), the model had a pooled sensitivity of 0.71 and a specificity of 0.67. Madani et al [44] also developed a DL model using echocardiographic data. Their validation set comprised 200 cases (100 LVH cases), with the

XSL•FO
RenderX

model demonstrating a pooled sensitivity of 0.78 and a specificity of 0.96. Similarly, Yu et al [35] implemented a DL approach for ECHO. Their validation set included 150 cases (97 LVH cases), and the model yielded a pooled sensitivity of 0.94 and a specificity of 0.92.

Discussion

Principal Findings

This study examined the efficacy of 3 types of ML models for LVH detection. In the validation cohort, ECG-based models had a sensitivity of 0.76 (95% CI 0.66 - 0.84) and a specificity of 0.84 (95% CI 0.78 - 0.89). The clinical feature-based models had a sensitivity of 0.78 (95% CI 0.69 - 0.85) and a specificity of 0.71 (95% CI 0.65 - 0.76). The ECHO-based models had a sensitivity ranging from 0.71 to 0.94 and a specificity ranging from 0.67 to 0.96. These findings confirm the significant efficacy of ML in LVH detection.

Comparison to Prior Work

ML is increasingly accurate in LVH detection, which enables the identification and treatment of earlier subclinical cardiovascular diseases [49]. In clinical practice, ECG is now widely used as a screening method for LVH due to its affordability and high accessibility, and is recommended as part of standard clinical evaluation for patients with hypertension in the Chinese Hypertension Clinical Practice Guidelines [50]. ECG-based LVH detection has been a major research priority in hypertensive cardiac target organ damage studies [1]. The clinical diagnosis of LVH via 12-lead ECG relies principally on the quantitative assessment of QRS complex (Q-wave, R-wave, S-wave) voltage parameters [51,52]. Despite recent innovations in diagnostic criteria, the overall sensitivity remains suboptimal, particularly in Chinese and other East Asian populations [1,53,54]. The low sensitivity of ECG in LVH detection limits its clinical application [32]. Consistent with existing research, our results showed that ECG-based models produced a sensitivity of 0.76 (95% CI 0.66 - 0.84) and a specificity of 0.84 (95% CI 0.78 - 0.89). Although limited by its sensitivity, studies have confirmed that ECG-detected LVH is associated with adverse cardiovascular outcomes [32,55]. Given its widespread availability and low cost, ECG remains an ideal initial assessment tool for LVH screening in patients with hypertension [56]. Therefore, some researchers recommend using ECG primarily for the risk stratification and prognosis prediction of LVH [43,57].

It has also been noted that different ML methods show significant differences in their ability to identify positive or outcome events. With its powerful automatic feature extraction capabilities, DL models can directly identify complex and subtle patterns from imaging data, integrate multiple imaging variables, and improve diagnostic accuracy [58,59]. Its automated segmentation reduces the subjective bias of traditional manual segmentation in radiomics and can integrate segmentation with diagnosis, offering significant advantages [60,61]. Therefore, it is imperative to prioritize the development of automated methods based on DL. Decision tree models are highly interpretable models that generate rules by recursively splitting variables in a dataset [62]. Decision tree models add a penalty term proportional to the absolute value of the feature coefficient for feature selection and model building [63]. Multiple studies have confirmed that RF models typically exhibit higher predictive performance than other ML methods because they can model nonlinear relationships and are insensitive to overfitting [64]. RF models demonstrate significant advantages in handling imbalanced datasets [65]. RF models excel at handling nonlinear and missing data, are conducive to high-dimensional data analysis, and demonstrate strong predictive performance [66]. Support vector machine (SVM) models have high predictive accuracy [67]. For structured and well-defined data, SVM models can skillfully capture the association between biomarkers and diseases through mathematical and statistical principles, thereby achieving high diagnostic accuracy [68,69]. SVM models are known for their ability to find the optimal decision boundary in high-dimensional feature spaces, resulting in effective classification based on extracted features [70,71].

ML models that incorporated clinical features demonstrated a sensitivity of 0.78 (95% CI 0.69 - 0.85) and a specificity of 0.71 (95% CI 0.65 - 0.76) in the validation set. Unlike traditional ECG criteria, the clinical features used to construct these models encompassed demographic characteristics (eg, sex, age) and clinical factors (eg, systolic blood pressure, duration of hypertension, and abdominal obesity). The possible reasons for using these factors are as follows.

Sex constitutes a significant determinant of LVH [1,37]. Substantial evidence has established the female sex as an independent predictor. In previous predictive models, women scored 70, while men scored 0, possibly due to the older study population, where most women had reached menopause [1]. Women exhibit higher salt sensitivity than men due to female hormonal influences, which amplifies adverse renin-angiotensin-aldosterone system effects [72,73]. Regardless of race or menopausal status, women with salt-sensitive hypertension face a higher LVH risk because this sex-specific mechanism promotes aldosterone-driven endothelial dysfunction [72,73]. This association is further exacerbated by diminished cardiovascular protection due to decreased estrogen levels in postmenopausal women [37].

Age plays a pivotal role in forecasting LVH in young patients with salt-sensitive hypertension. The evidence reveals a dual pattern: (1) patients 40 years or older demonstrate a significantly higher risk than younger populations, consistent with age-related left ventricular mass increase [1,37,74], while (2) patients with early-onset hypertension exhibit more severe cardiac damage and elevated cardiovascular mortality [74,75]. Such observations suggest that the interaction between age and the timing of hypertension requires special attention.

The development of LVH is closely associated with systolic blood pressure level and duration of hypertension. Prospective studies demonstrate that each 19 mm Hg increase in systolic blood pressure elevates the LVH risk by 49% [6], while effective systolic blood pressure control markedly reduces the disease risk in old patients [76]. These findings provide crucial evidence for clinical treatment decisions. Hypertension duration, as a cumulative exposure metric, has been consistently validated in

XSL•FO
RenderX

multiple studies to be significantly associated with LVH. Shortening the course of hypertension can effectively decrease the risk of myocardial remodeling, which shows the importance of early intervention in delaying LVH progression [77].

Obesity constitutes an independent risk factor for LVH [78,79]. Waist circumference, the principal diagnostic metric for abdominal obesity, is linked with both cardiovascular and metabolic disorders [80]. The pathophysiological linkage between abdominal obesity and LVH can be attributed to the robust correlation between waist circumference and excessive visceral adipose tissue accumulation [80,81]. As a metabolically active depot, visceral adipose tissue exhibits substantially higher energy demands compared to subcutaneous fat [81]. This augmented metabolic requirement exacerbates cardiac workload in patients with hypertension, ultimately resulting in LVH [78,81]. Studies have identified waist circumference as a superior predictor of cardiovascular risk and mortality compared to BMI [80,82]. Notably, elevated systolic blood pressure and prolonged hypertension duration may synergistically accelerate concentric cardiac remodeling, while abdominal obesity plus aging compounds metabolic risk.

The complex interactions among clinical features were quantified and integrated into ML models. While previous studies have confirmed the underlying mechanism of individual factors, models combining features from diverse dimensions provide more complete predictive insights. Specifically, sex-specific risk mechanisms dynamically interact with age-related cardiovascular changes, while blood pressure's temporal cumulative effects and the spatial distribution of metabolic factors form a 3D LVH prediction network. This finding provides a theoretical foundation for developing clinical feature–based ML tools for precise prediction.

ECHO serves as the primary imaging tool for LVH identification. However, its clinical application faces multiple technical challenges, including a relatively low signal-to-noise ratio, inconsistent image quality, and suboptimal reproducibility [28,35]. These limitations are particularly pronounced in patients with obesity due to enhanced chest wall and subcutaneous tissue thickness, which significantly impede image acquisition [28,35]. ML technology offers a novel approach to overcoming these technical barriers. ML models incorporating ECHO demonstrated a sensitivity range of 0.71 to 0.94 and a specificity range of 0.67 to 0.96.

Existing hospital ECHO systems are integrated with this technology to enable automated image analysis and real-time LVH diagnosis. Without replacing existing manual verification procedures, it markedly increases diagnostic efficiency and proves particularly beneficial as a clinical decision aid for less experienced physicians [28]. In contemporary practice, LVH detection chiefly relies on ECHO, where ML proves to be powerful in image processing stages [35,83]. Conventional approaches that depend on manual segmentation and feature extraction are prone to information loss, whereas DL models incorporate image processing, texture extraction, and feature selection throughout the modeling process, which better preserves original image information [35,84,85]. All ML models based on ECHO in this study were DL models. Beyond improved diagnostic efficiency, ML models can automatically identify high-risk cases and enable cardiologists to prioritize urgent interventions, thus shortening care pathways and improving outcomes [49]. Additionally, this technique can enhance diagnostic capabilities at primary care levels, optimize the allocation of cardiovascular imaging resources, reduce unnecessary tests, and be significantly cost-effective in population screening and subclinical disease management [28].

Despite the superior image-processing capabilities of DL models, their clinical integration with ultrasound systems must overcome real-world challenges, including real-time performance, device compatibility, and patient data privacy. Furthermore, the stability of ML models across demographics needs further validation in multicenter research. Future research should prioritize enlarging and diversifying ECHO image databases and validating the actual performance of ML in the detection of LVH in prospective trials.

Meta-analysis of ECG-based ML models for detecting LVH revealed substantial heterogeneity. This restricts the statistical reliability of pooled performance metrics, preventing their interpretation as definitive generalization conclusions. Potential sources of such heterogeneity are manifold. Methodologically, substantial variation in ECG signal preprocessing, feature extraction, and used algorithms introduces fundamental differences. Clinically, diverse underlying disease profiles and severity among studied populations affect model consistency. Regional demographic variations further increase result dispersion. Consequently, pooled estimates from our research require cautious interpretation. Their value lies more in demonstrating the broad performance range and limited confirmatory nature of current evidence than in providing a stable performance expectation. Future efforts should prioritize methodological standardization and prospective multicenter validation to enhance evidence quality in this domain.

When the PROBAST was used to assess model quality, we found that the results raised some concerns, particularly in terms of its strict evaluation of the study population. We believe that some of its criteria may be overly stringent. Although the PROBAST is suitable for evaluating multivariate models for diagnosis and prediction, it is primarily based on retrospective case-control studies, with few prospective studies included [86]. This design preference often leads to a high risk of bias in assessment results, which poses a significant challenge to diagnostic model research. In the studies we included, most involved diagnostic models, and the case selection process itself might introduce bias. The PROBAST uses the EPV principle for sample size estimation, but image-based ML models face challenges in applying EPV rules for bias risk assessment. In future studies, we may explore prior sample size estimation during the research design phase, leveraging ML-based expected performance metrics, such as area under the curve and its 95% CI. Moreover, learning curve analysis serves as a powerful empirical tool. By plotting how a model's performance on both the training and validation sets changes as the sample size increases, we can intuitively determine whether the model's performance has reached a plateau, thereby establishing the optimal number of cases. In the field of DL, there exist several widely accepted rules. For instance, in complex image

XSL•FO
RenderX

classification tasks, it is generally recommended that each category requires at least several thousand or even tens of thousands of training images to achieve good generalization performance. Although these guidelines are somewhat rough, they provide a practical benchmark for assessing sample sizes in large-scale imaging studies. In addition, the PROBAST requires the verification of the consistency between model statistical weights and actual reported results, which is particularly difficult for models with low interpretability (such as SVM), as they typically do not provide the weight information required in the original validation set [24]. These issues indicate that while the PROBAST is widely used for bias risk assessment in ML research, the observed limitations arise more fundamentally from the inherent biases of the retrospective case-control design—which is applied by most of the current ML diagnostic studies—than from an inherent excessive stringency in the PROBAST criteria. Advancing this field requires constructing and validating models through prospective, multicenter, double-blinded cohort studies, thereby improving evidence quality at its source. Specifically, future work should emphasize prospective sample size estimation and learning curve analysis during development, enhance model interpretability, conduct independent external validation, and ultimately assess the real-world impact of a model on clinical workflows and patient outcomes via pragmatic effectiveness studies. Additionally, we recommend using QUADAS-AI (Quality Assessment of Diagnostic Accuracy Studies Using AI) to assess the risk of bias in included original studies for diagnostic ML tasks [87]. QUADAS-AI focuses on data sources, sample size, eligibility criteria, the rationale for splitting training, validation, test sets, imaging protocols, and preprocessing methods—rather than simply labeling all case-control studies as “high risk.” In the studies we included, a substantial number of studies clearly described the source, size, and eligibility criteria, the principles of splitting into training, validation, and test sets, as well as the imaging protocols and preprocessing methods, thereby mitigating the high risk of bias associated with the study design.

This study has high statistical heterogeneity. Even when we conducted subgroup analyses using various ML approaches and validation methods to further elucidate the sources of heterogeneity, our explanatory power remains limited. This may be due to underlying heterogeneity present during model training. We consider that the following ML processes may introduce potential heterogeneity. First, regarding image preprocessing, the included studies exhibited certain variations in their approaches. These differences may introduce significant heterogeneity. For different images, there is considerable variation in image parameters, yet there is a lack of effective discussion on how this impacts modeling. The images in the original studies exhibit significant variations across imaging protocols. Although we performed subgroup analyses based on different models, even within the same model, differences in model iterations and parameter adjustment rules persist, potentially contributing to heterogeneity.

Strengths and Limitations of the Study

Our systematic review is the first to quantitatively examine ML performance in LVH detection. While informative, several limitations should be considered when interpreting these results. First, the systematic search, while rigorous, identified solely small case-control studies, thereby narrowing the spectrum of studies for subgroup comparisons. Second, the limited number of included studies precluded meaningful comparisons between different ML methodologies. Third, while most studies completed external validation, some relied solely on random sampling for internal validation. Owing to the limited number of publications, a systematic comparison between internal and external validation approaches is infeasible. Fourth, the RoB in the included studies is high, which is inevitable given the stringent criteria for evaluating ML-based research. This observation corresponds with methodological appraisals indicating that 87% of ML models for medical prediction carry a high RoB [88]. Future studies should optimize design, implementation, reporting, and validation approaches to facilitate clinical translation of ML-based prediction models. Fifth, some studies divide data according to images, and patients’ images are shared between the training set and validation sets, which may result in data leakage. Sixth, there are different reference standards for different models. This discrepancy in reference standards indicates that any comparison of diagnostic accuracy between different model types is inherently confounded. The presented pooled estimates for each subgroup should be interpreted independently and not compared directly to each other. Seventh, the findings of this study may have been influenced to some extent by the fact that 3 papers used the 2×2 formula for estimation. The results should be interpreted with caution. Eighth, several included studies used “clinical consensus” as the reference standard for diagnosing LVH. While this consensus typically adhered to established guidelines (defining LVH as left ventricular mass index >115 g/m² in men and >95 g/m² in women [89]), original studies generally did not explicitly describe whether the consensus process was entirely independent of the outputs of evaluated ML models. Theoretically, “incorporation bias” could be introduced if the predictions of a model were known or incorporated during the adjudication of the reference standard, potentially leading to an overestimation of its diagnostic accuracy. This constitutes a key limitation for evidence interpretation in this systematic review and underscores the necessity for future original research to explicitly report reference standard independence in methodological descriptions.

Future Directions

ML is widely used in LVH research, but its clinical application still faces multiple challenges. Many models are developed based on limited samples and lack standardized processes for key steps, such as image selection, segmentation, and feature extraction, resulting in high reliance of models on the experience of developers [62,90]. This raises concerns about the robustness of the models. Significant differences in image quality across different institutions further weaken the universality of the models. Existing models often focus on accuracy and neglect population risk stratification and regional differences in medical resources, limiting their clinical utility. Considering interpretability and a lack of trust among physicians, it remains difficult to integrate ML into clinical practice [91,92]. Although some models perform well in retrospective studies, their

XSL•FO
RenderX

effectiveness and applicability in real-world clinical settings still need to be verified through rigorous prospective trials. Before clinical application, it is necessary to fully address the issues of patient data privacy protection, lack of trust among doctors due to insufficient model interpretability, and ethical considerations associated with new systems [93-95]. Additionally, we must be wary of overreliance on technological means at the expense of humanistic care. The need for systematic management and individualized treatment should not be overlooked. Technology should serve, rather than replace, patient-centered medical practices.

To promote the safe, effective, and responsible application of ML in LVH clinical practice, it is necessary to standardize the model development process, strengthen research on model generalizability and interpretability, conduct rigorous prospective validation, and focus on risk stratification and resource differences. Furthermore, it is necessary to establish a comprehensive ethical and regulatory framework covering data privacy, patient rights, algorithm transparency, and clinical workflow integration. Clear implementation guidelines are needed to ensure that technological development truly benefits patients and improves the quality of medical care.

Conclusions

This study suggests that ML demonstrates diagnostic potential in detecting LVH. However, the overall certainty of the evidence is low, primarily constrained by the limited number of included studies and substantial heterogeneity. Consequently, current conclusions about model performance should be cautiously interpreted. Future investigations should adopt prospective designs and implement standardized data acquisition and model validation protocols to develop and evaluate ML models with enhanced robustness and clinical interpretability. This will advance this technology and make it a reliable diagnostic tool.

Funding

This research was not supported by any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

All data generated or analyzed during this study are included in this published article and its multimedia appendices.

Authors' Contributions

Conceptualization: JW, KZ, YL
Formal analysis and investigation: JW, KZ, YL
Methodology: JW, KZ, YL
Supervision: JW, KZ, YL
Writing – original draft preparation: JW, KZ, YL
Writing – review and editing: JW, KZ, YL
All authors commented on previous versions of the manuscript, read and approved the final manuscript, and contributed to the study conception and design.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Literature search strategy.
[DOCX File, 20 KB - jmir_v28i1e76637_app1.docx]

Multimedia Appendix 2
Summary receiver operating characteristic curves, funnel plots, and nomograms for electrocardiogram (overall and validation analyses) and clinical features.
[DOCX File, 1410 KB - jmir_v28i1e76637_app2.docx]

Checklist 1
PRISMA flowchart.
[PDF File, 99 KB - jmir_v28i1e76637_app3.pdf]

References

1. Zhang X, He C, Lu S, et al. Construction and validation of a nomogram to predict left ventricular hypertrophy in low‐risk patients with hypertension. J Clin Hypertens 2024 Mar;26(3):274-285. [doi: 10.1111/jch.14780]
2. Đorđević DB, Koračević GP, Đorđević AD, Lović DB. Hypertension and left ventricular hypertrophy. J Hypertens 2024 Sep 1;42(9):1505-1515. [doi: 10.1097/HJH.0000000000003774] [Medline: 38747417]

XSL•FO
RenderX

3. Han Y, Li Y, Wu Z, et al. Progress in diagnosis and treatment of hypertension combined with left ventricular hypertrophy. Ann Med 2024 Dec;56(1):2405080. [doi: 10.1080/07853890.2024.2405080] [Medline: 39301864]
4. Miller RJH, Mikami Y, Heydari B, et al. Sex-specific relationships between patterns of ventricular remodelling and clinical outcomes. Eur Heart J Cardiovasc Imaging 2020 Sep 1;21(9):983-990. [doi: 10.1093/ehjci/jeaa164] [Medline: 32594163]
5. Lewis AA, Ayers CR, Selvin E, et al. Racial differences in malignant left ventricular hypertrophy and incidence of heart failure: a multicohort study. Circulation 2020 Mar 24;141(12):957-967. [doi: 10.1161/CIRCULATIONAHA.119.043628] [Medline: 31931608]
6. Cao X, Broughton ST, Waits GS, Nguyen T, Li Y, Soliman EZ. Interrelations between hypertension and electrocardiographic left ventricular hypertrophy and their associations with cardiovascular mortality. Am J Cardiol 2019 Jan 15;123(2):274-283. [doi: 10.1016/j.amjcard.2018.10.006] [Medline: 30390988]
7. Siranart N, Deepan N, Techasatian W, et al. Diagnostic accuracy of artificial intelligence in detecting left ventricular hypertrophy by electrocardiograph: a systematic review and meta-analysis. Sci Rep 2024 Jul 10;14(1):15882. [doi: 10.1038/s41598-024-66247-y] [Medline: 38987652]
8. Pedersen LR, Kristensen AMD, Petersen SS, et al. Prognostic implications of left ventricular hypertrophy diagnosed on electrocardiogram vs echocardiography. J Clin Hypertens (Greenwich) 2020 Sep;22(9):1647-1658. [doi: 10.1111/jch.13991] [Medline: 32813895]
9. Woythaler JN, Singer SL, Kwan OL, et al. Accuracy of echocardiography versus electrocardiography in detecting left ventricular hypertrophy: comparison with postmortem mass measurements. J Am Coll Cardiol 1983 Aug;2(2):305-311. [doi: 10.1016/s0735-1097(83)80167-3] [Medline: 6223063]
10. Cai C, Imai T, Hasumi E, Fujiu K. One-shot screening: utilization of a two-dimensional convolutional neural network for automatic detection of left ventricular hypertrophy using electrocardiograms. Comput Methods Programs Biomed 2024 Apr;247:108097. [doi: 10.1016/j.cmpb.2024.108097] [Medline: 38428250]
11. Spahillari A, McCormick I, Yang JX, Quinn GR, Manning WJ. On-call transthoracic echocardiographic interpretation by first year cardiology fellows: comparison with attending cardiologists. BMC Med Educ 2019 Jun 14;19(1):213. [doi: 10.1186/s12909-019-1634-7] [Medline: 31200721]
12. Rao S, Ferris TG, Hidrue MK, et al. Physician burnout, engagement and career satisfaction in a large academic medical practice. Clin Med Res 2020 Mar;18(1):3-10. [doi: 10.3121/cmr.2019.1516] [Medline: 31959669]
13. Quinn GR, Ranum D, Song E, et al. Missed diagnosis of cardiovascular disease in outpatient general medicine: insights from malpractice claims data. Jt Comm J Qual Patient Saf 2017 Oct;43(10):508-516. [doi: 10.1016/j.jcjq.2017.05.001] [Medline: 28942775]
14. Wang Z, Gu Y, Huang L, et al. Construction of machine learning diagnostic models for cardiovascular pan-disease based on blood routine and biochemical detection data. Cardiovasc Diabetol 2024 Sep 28;23(1):351. [doi: 10.1186/s12933-024-02439-0] [Medline: 39342281]
15. Qi X, Wang S, Fang C, Jia J, Lin L, Yuan T. Machine learning and SHAP value interpretation for predicting comorbidity of cardiovascular disease and cancer with dietary antioxidants. Redox Biol 2025 Feb;79:103470. [doi: 10.1016/j.redox.2024.103470] [Medline: 39700695]
16. Layton AT. AI, machine learning, and ChatGPT in hypertension. Hypertension 2024 Apr;81(4):709-716. [doi: 10.1161/HYPERTENSIONAHA.124.19468] [Medline: 38380541]
17. Ying Y, Ju R, Wang J, et al. Accuracy of machine learning in diagnosing microsatellite instability in gastric cancer: a systematic review and meta-analysis. Int J Med Inform 2025 Jan;193:105685. [doi: 10.1016/j.ijmedinf.2024.105685] [Medline: 39515046]
18. Beam AL, Kohane IS. Big data and machine learning in health care. J Am Med Assoc 2018 Apr 3;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
19. Zhao K, Zhu Y, Chen X, et al. Machine learning in hypertrophic cardiomyopathy: nonlinear model from clinical and CMR features predicting cardiovascular events. JACC Cardiovasc Imaging 2024 Aug;17(8):880-893. [doi: 10.1016/j.jcmg.2024.04.013] [Medline: 39001729]
20. You J, Guo Y, Kang JJ, et al. Development of machine learning-based models to predict 10-year risk of cardiovascular disease: a prospective cohort study. Stroke Vasc Neurol 2023 Dec 29;8(6):475-485. [doi: 10.1136/svn-2023-002332] [Medline: 37105576]
21. Naderi H, Ramírez J, van Duijvenboden S, et al. Diagnostic and prognostic value of ECG-predicted hypertension-mediated left ventricular hypertrophy using machine learning. medRxiv. Preprint posted online on Apr 23, 2024. [doi: 10.1101/2024.04.22.24306204]
22. Ryu JS, Lee S, Chu Y, Ahn MS, Park YJ, Yang S. CoAt-mixer: self-attention deep learning framework for left ventricular hypertrophy using electrocardiography. PLoS ONE 2023;18(6):e0286916. [doi: 10.1371/journal.pone.0286916] [Medline: 37289800]
23. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]

XSL•FO
RenderX

24. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019 Jan 1;170(1):W1-W33. [doi: 10.7326/M18-1377] [Medline: 30596876]
25. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol 2005 Sep;58(9):882-893. [doi: 10.1016/j.jclinepi.2005.01.016] [Medline: 16085191]
26. Fagan TJ. Letter: nomogram for Bayes's theorem. N Engl J Med 1975 Jul 31;293(5):257. [doi: 10.1056/NEJM197507312930513] [Medline: 1143310]
27. Wan J, Wang P, Liu S, Wang X, Zhou P, Yang J. Risk factors and a predictive model for left ventricular hypertrophy in young adults with salt-sensitive hypertension. J Clin Hypertens (Greenwich) 2024 Aug;26(8):933-944. [doi: 10.1111/jch.14863] [Medline: 38940286]
28. Farhad M, Masud MM, Beg A, Ahmad A, Ahmed LA, Memon S. A data-efficient zero-shot and few-shot Siamese approach for automated diagnosis of left ventricular hypertrophy. Comput Biol Med 2023 Sep;163:107129. [doi: 10.1016/j.compbiomed.2023.107129] [Medline: 37343469]
29. Yuan R, Chen J, Zhang S, Zhang X, Yu J. Establishment of an individual-specific nomogram for predicting the risk of left ventricular hypertrophy in Chinese postmenopausal hypertensive women. Medicina (Kaunas) 2023 Mar 17;59(3):598. [doi: 10.3390/medicina59030598] [Medline: 36984599]
30. Dwivedi T, Xue J, Treiman D, Dubey A, Albert D. Machine learning models of 6-lead ECGs for the interpretation of left ventricular hypertrophy (LVH). J Electrocardiol 2023;77:62-67. [doi: 10.1016/j.jelectrocard.2022.12.001] [Medline: 36641988]
31. Liu CW, Wu FH, Hu YL, et al. Left ventricular hypertrophy detection using electrocardiographic signal. Sci Rep 2023 Feb 13;13(1):2556. [doi: 10.1038/s41598-023-28325-5] [Medline: 36781924]
32. Zhao X, Huang G, Wu L, et al. Deep learning assessment of left ventricular hypertrophy based on electrocardiogram. Front Cardiovasc Med 2022;9:952089. [doi: 10.3389/fcvm.2022.952089] [Medline: 36035939]
33. Liu CM, Hsieh ME, Hu YF, et al. Artificial intelligence-enabled model for early detection of left ventricular hypertrophy and mortality prediction in young to middle-aged adults. Circ Cardiovasc Qual Outcomes 2022 Aug;15(8):e008360. [doi: 10.1161/CIRCOUTCOMES.121.008360] [Medline: 35959675]
34. Wu Z, Shi M, Wang L, Yao Y. Identification of major risk factors and non-linear effects to the development of left ventricular hypertrophy in chronic kidney disease by constructing and validation of nomograms. Front Med (Lausanne) 2022;9:914800. [doi: 10.3389/fmed.2022.914800] [Medline: 35911424]
35. Yu X, Yao X, Wu B, et al. Using deep learning method to identify left ventricular hypertrophy on echocardiography. Int J Cardiovasc Imaging 2022 Apr;38(4):759-769. [doi: 10.1007/s10554-021-02461-3] [Medline: 34757566]
36. Kokubo T, Kodera S, Sawano S, et al. Automatic detection of left ventricular dilatation and hypertrophy from electrocardiograms using deep learning. Int Heart J 2022 Sep 30;63(5):939-947. [doi: 10.1536/ihj.22-132] [Medline: 36104234]
37. Ye C, Wang T, Gong J, et al. Development of a nomogram for screening the risk of left ventricular hypertrophy in Chinese hypertensive patients. J Clin Hyperten 2021 Jun;23(6):1176-1185. [doi: 10.1111/jch.14240]
38. Angelaki E, Marketou ME, Barmparis GD, et al. Detection of abnormal left ventricular geometry in patients without cardiovascular disease through machine learning: an ECG-based approach. J Clin Hypertens (Greenwich) 2021 May;23(5):935-945. [doi: 10.1111/jch.14200] [Medline: 33507615]
39. Ghorbani A, Ouyang D, Abid A, et al. Deep learning interpretation of echocardiograms. NPJ Digit Med 2020;3(10):10. [doi: 10.1038/s41746-019-0216-8] [Medline: 31993508]
40. Kwon JM, Jeon KH, Kim HM, et al. Comparing the performance of artificial intelligence and conventional diagnosis criteria for detecting left ventricular hypertrophy using electrocardiography. Europace 2020 Mar 1;22(3):412-419. [doi: 10.1093/europace/euz324] [Medline: 31800031]
41. Lin GM, Liu K. An electrocardiographic system with anthropometrics via machine learning to screen left ventricular hypertrophy among young adults. IEEE J Transl Eng Health Med 2020;8:1800111. [doi: 10.1109/JTEHM.2020.2990073] [Medline: 32419990]
42. Tison GH, Zhang J, Delling FN, Deo RC. Automated and interpretable patient ECG profiles for disease detection, tracking, and discovery. Circ Cardiovasc Qual Outcomes 2019 Sep;12(9):e005289. [doi: 10.1161/CIRCOUTCOMES.118.005289] [Medline: 31525078]
43. Sparapani R, Dabbouseh NM, Gutterman D, et al. Detection of left ventricular hypertrophy using Bayesian additive regression trees: the MESA. J Am Heart Assoc 2019 Mar 5;8(5):e009959. [doi: 10.1161/JAHA.118.009959] [Medline: 30827132]
44. Madani A, Ong JR, Tibrewal A, Mofrad MRK. Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. NPJ Digit Med 2018;1(1):59. [doi: 10.1038/s41746-018-0065-x] [Medline: 31304338]
45. De la Garza Salazar F, Romero Ibarguengoitia ME, Azpiri López JR, González Cantú A. Optimizing ECG to detect echocardiographic left ventricular hypertrophy with computer-based ECG data and machine learning. PLoS ONE 2021;16(11):e0260661. [doi: 10.1371/journal.pone.0260661] [Medline: 34847202]

XSL•FO
RenderX

46. Huang JT, Tseng CH, Huang WM, et al. Comparison of machine learning and conventional criteria in detecting left ventricular hypertrophy and prognosis with electrocardiography. Eur Heart J Digit Health 2025 Mar;6(2):252-260. [doi: 10.1093/ehjdh/ztaf003] [Medline: 40110220]
47. Taconne M, Corino VDA, Mainardi L. An ECG-based model for left ventricular hypertrophy detection: a machine learning approach. IEEE Open J Eng Med Biol 2025;6:219-226. [doi: 10.1109/OJEMB.2024.3509379] [Medline: 39698126]
48. van Smeden M, Moons KG, de Groot JA, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. Stat Methods Med Res 2019 Aug;28(8):2455-2474. [doi: 10.1177/0962280218784726] [Medline: 29966490]
49. Duffy G, Cheng PP, Yuan N, et al. High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning. JAMA Cardiol 2022 Apr 1;7(4):386-395. [doi: 10.1001/jamacardio.2021.6059] [Medline: 35195663]
50. Liu LS, Writing Group of 2010 Chinese Guidelines for the Management of Hypertension. 2010 Chinese guidelines for the management of hypertension. Zhonghua Xin Xue Guan Bing Za Zhi 2011 Jul;39(7):579-615. [Medline: 22088239]
51. Casale PN, Devereux RB, Kligfield P, et al. Electrocardiographic detection of left ventricular hypertrophy: development and prospective validation of improved criteria. J Am Coll Cardiol 1985 Sep;6(3):572-580. [doi: 10.1016/s0735-1097(85)80115-7] [Medline: 3161926]
52. Sokolow M, Lyon TP. The ventricular complex in left ventricular hypertrophy as obtained by unipolar precordial and limb leads. Am Heart J 1949 Feb;37(2):161-186. [doi: 10.1016/0002-8703(49)90562-1] [Medline: 18107386]
53. Xia Y, Li X, Zhang H, et al. Diagnostic capability and influence factors for a new electrocardiogram criterion in the diagnosis of left ventricular hypertrophy in a Chinese population. Cardiology 2020;145(5):294-302. [doi: 10.1159/000505421] [Medline: 32289773]
54. Wang D, Xu JZ, Zhang W, et al. Performance of electrocardiographic criteria for echocardiographically diagnosed left ventricular hypertrophy in Chinese hypertensive patients. Am J Hypertens 2020 Sep 10;33(9):831-836. [doi: 10.1093/ajh/hpaa083] [Medline: 32484222]
55. Bang CN, Devereux RB, Okin PM. Regression of electrocardiographic left ventricular hypertrophy or strain is associated with lower incidence of cardiovascular morbidity and mortality in hypertensive patients independent of blood pressure reduction—a LIFE review. J Electrocardiol 2014;47(5):630-635. [doi: 10.1016/j.jelectrocard.2014.07.003] [Medline: 25052475]
56. Hancock EW, Deal BJ, Mirvis DM, et al. AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: part V: electrocardiogram changes associated with cardiac chamber hypertrophy: a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society: endorsed by the International Society for Computerized Electrocardiology. Circulation 2009 Mar 17;119(10):e251-e261. [doi: 10.1161/CIRCULATIONAHA.108.191097] [Medline: 19228820]
57. Rautaharju PM, Soliman EZ. Electrocardiographic left ventricular hypertrophy and the risk of adverse cardiovascular events: a critical appraisal. J Electrocardiol 2014;47(5):649-654. [doi: 10.1016/j.jelectrocard.2014.06.002] [Medline: 25012077]
58. Jiang Y, Yang M, Wang S, Li X, Sun Y. Emerging role of deep learning-based artificial intelligence in tumor pathology. Cancer Commun (Lond) 2020 Apr;40(4):154-166. [doi: 10.1002/cac2.12012] [Medline: 32277744]
59. Jiang X, Li J, Kan Y, et al. MRI based radiomics approach with deep learning for prediction of vessel invasion in early-stage cervical cancer. IEEE/ACM Trans Comput Biol Bioinform 2021;18(3):995-1002. [doi: 10.1109/TCBB.2019.2963867] [Medline: 31905143]
60. She L, Li Y, Wang H, et al. Imaging-based AI for predicting lymphovascular space invasion in cervical cancer: systematic review and meta-analysis. J Med Internet Res 2025 Jun 16;27:e71091. [doi: 10.2196/71091] [Medline: 40523236]
61. Zuo H, Huang B, He J, Fang L, Huang M. Machine learning approaches in high myopia: systematic review and meta-analysis. J Med Internet Res 2025 Jan 3;27:e57644. [doi: 10.2196/57644] [Medline: 39753217]
62. Zhang H, Zou P, Luo P, Jiang X. Machine learning for the early prediction of delayed cerebral ischemia in patients with subarachnoid hemorrhage: systematic review and meta-analysis. J Med Internet Res 2025 Jan 20;27:e54121. [doi: 10.2196/54121]
63. Zhu J, Yang F, Wang Y, et al. Accuracy of machine learning in discriminating Kawasaki disease and other febrile illnesses: systematic review and meta-analysis. J Med Internet Res 2024 Nov 18;26:e57641. [doi: 10.2196/57641] [Medline: 39556821]
64. Becker T, Rousseau AJ, Geubbelmans M, Burzykowski T, Valkenborg D. Decision trees and random forests. Am J Orthod Dentofacial Orthop 2023 Dec;164(6):894-897. [doi: 10.1016/j.ajodo.2023.09.011] [Medline: 38008491]
65. Liu L, Li Z, Hu Y, et al. Predictive performance of machine learning for suicide in adolescents: systematic review and meta-analysis. J Med Internet Res 2025 Jun 16;27:e73052. [doi: 10.2196/73052] [Medline: 40522723]
66. Yang J, Zeng S, Cui S, Zheng J, Wang H. Predictive modeling of acute respiratory distress syndrome using machine learning: systematic review and meta-analysis. J Med Internet Res 2025 May 13;27:e66615. [doi: 10.2196/66615] [Medline: 40359510]
67. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 2019;9(4):e1312. [doi: 10.1002/widm.1312] [Medline: 32089788]
68. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol 2019 May;20(5):e262-e273. [doi: 10.1016/S1470-2045(19)30149-4] [Medline: 31044724]

XSL•FO
RenderX

69. Parvatikar PP, Patil S, Khaparkhuntikar K, et al. Artificial intelligence: machine learning approach for screening large database and drug discovery. Antiviral Res 2023 Dec;220:105740. [doi: 10.1016/j.antiviral.2023.105740] [Medline: 37935248]
70. Tsai CA, Chang YJ. Efficient selection of Gaussian kernel SVM parameters for imbalanced data. Genes (Basel) 2023 Feb 25;14(3):583. [doi: 10.3390/genes14030583] [Medline: 36980852]
71. Park JS, Park SY, Moon JW, Kim K, Suh DI. Artificial intelligence models for pediatric lung sound analysis: systematic review and meta-analysis. J Med Internet Res 2025 Apr 18;27:e66491. [doi: 10.2196/66491] [Medline: 40249944]
72. Liao YY, Gao K, Fu BW, et al. Risk factors for electrocardiographic left ventricular hypertrophy in a young Chinese general population: the Hanzhong adolescent cohort study. BMC Cardiovasc Disord 2021 Mar 31;21(1):159. [doi: 10.1186/s12872-021-01966-y] [Medline: 33789587]
73. Tiyerili V, Müller CFH, Fung S, Panek D, Nickenig G, Becher UM. Estrogen improves vascular function via peroxisome-proliferator-activated-receptor-γ. J Mol Cell Cardiol 2012 Aug;53(2):268-276. [doi: 10.1016/j.yjmcc.2012.05.008] [Medline: 22634137]
74. Vasan RS, Song RJ, Xanthakis V, et al. Hypertension-mediated organ damage: prevalence, correlates, and prognosis in the community. Hypertension 2022 Mar;79(3):505-515. [doi: 10.1161/HYPERTENSIONAHA.121.18502] [Medline: 35138872]
75. Wang C, Yuan Y, Zheng M, et al. Association of age of onset of hypertension with cardiovascular diseases and mortality. J Am Coll Cardiol 2020 Jun 16;75(23):2921-2930. [doi: 10.1016/j.jacc.2020.04.038] [Medline: 32527401]
76. Zhou B, Li C, Shou J, Zhang Y, Wen C, Zeng C. The cumulative blood pressure load and target organ damage in patients with essential hypertension. J Clin Hypertens (Greenwich) 2020 Jun;22(6):981-990. [doi: 10.1111/jch.13875] [Medline: 32427414]
77. Carlsson AC, Ruge T, Sundström J, et al. Association between circulating endostatin, hypertension duration, and hypertensive target-organ damage. Hypertension 2013 Dec;62(6):1146-1151. [doi: 10.1161/HYPERTENSIONAHA.113.02250] [Medline: 24082055]
78. Zhang X, Li G, Zhang D, Sun Y. Influence of hypertension and global or abdominal obesity on left ventricular hypertrophy: a cross-sectional study. J Clin Hypertens (Greenwich) 2024 Dec;26(12):1342-1350. [doi: 10.1111/jch.14716] [Medline: 37794691]
79. Tsujimoto T, Kajio H. Abdominal obesity is associated with an increased risk of all-cause mortality in patients with HFpEF. J Am Coll Cardiol 2017 Dec 5;70(22):2739-2749. [doi: 10.1016/j.jacc.2017.09.1111] [Medline: 29191321]
80. Choi D, Choi S, Son JS, Oh SW, Park SM. Impact of discrepancies in general and abdominal obesity on major adverse cardiac events. J Am Heart Assoc 2019 Sep 17;8(18):e013471. [doi: 10.1161/JAHA.119.013471] [Medline: 31480883]
81. Ross R, Neeland IJ, Yamashita S, et al. Waist circumference as a vital sign in clinical practice: a Consensus Statement from the IAS and ICCR Working Group on Visceral Obesity. Nat Rev Endocrinol 2020 Mar;16(3):177-189. [doi: 10.1038/s41574-019-0310-7] [Medline: 32020062]
82. Wang Y, Howard AG, Adair LS, Wang H, Avery CL, Gordon-Larsen P. Waist circumference change is associated with blood pressure change independent of BMI change. Obesity (Silver Spring) 2020 Jan;28(1):146-153. [doi: 10.1002/oby.22638] [Medline: 31755247]
83. Wu D, Ono R, Wang S, Kobayashi Y, Sughimoto K, Liu H. Pulse wave signal-driven machine learning for identifying left ventricular enlargement in heart failure patients. Biomed Eng Online 2024 Jun 22;23(1):60. [doi: 10.1186/s12938-024-01257-5] [Medline: 38909231]
84. Lau ES, Di Achille P, Kopparapu K, et al. Deep learning-enabled assessment of left heart structure and function predicts cardiovascular outcomes. J Am Coll Cardiol 2023 Nov 14;82(20):1936-1948. [doi: 10.1016/j.jacc.2023.09.800] [Medline: 37940231]
85. Wehbe RM, Katsaggelos AK, Hammond KJ, et al. Deep learning for cardiovascular imaging: a review. JAMA Cardiol 2023 Nov 1;8(11):1089-1098. [doi: 10.1001/jamacardio.2023.3142] [Medline: 37728933]
86. Maleki F, Ovens K, Gupta R, Reinhold C, Spatz A, Forghani R. Generalizability of machine learning models: quantitative evaluation of three methodological pitfalls. Radiol Artif Intell 2023 Jan;5(1):e220028. [doi: 10.1148/ryai.220028] [Medline: 36721408]
87. Sounderajah V, Ashrafian H, Rose S, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nat Med 2021 Oct;27(10):1663-1665. [doi: 10.1038/s41591-021-01517-0] [Medline: 34635854]
88. Andaur Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ 2021 Oct 20;375:n2281. [doi: 10.1136/bmj.n2281] [Medline: 34670780]
89. Lang RM, Badano LP, Mor-Avi V, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. J Am Soc Echocardiogr 2015 Jan;28(1):1-39. [doi: 10.1016/j.echo.2014.10.003] [Medline: 25559473]
90. Wang B, Jiang B, Liu D, Zhu R. Early predictive accuracy of machine learning for hemorrhagic transformation in acute ischemic stroke: systematic review and meta-analysis. J Med Internet Res 2025 May 23;27:e71654. [doi: 10.2196/71654] [Medline: 40408765]

XSL•FO
RenderX

91. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195. [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
92. Alkhanbouli R, Matar Abdulla Almadhaani H, Alhosani F, Simsekler MCE. The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions. BMC Med Inform Decis Mak 2025 Mar 4;25(1):110. [doi: 10.1186/s12911-025-02944-6] [Medline: 40038704]
93. Aggarwal R, Farag S, Martin G, Ashrafian H, Darzi A. Patient perceptions on data sharing and applying artificial intelligence to health care data: cross-sectional survey. J Med Internet Res 2021 Aug 26;23(8):e26162. [doi: 10.2196/26162] [Medline: 34236994]
94. Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med Ethics 2021 Sep 15;22(1):122. [doi: 10.1186/s12910-021-00687-3] [Medline: 34525993]
95. Larson DB, Magnus DC, Lungren MP, Shah NH, Langlotz CP. Ethics of using and sharing clinical imaging data for artificial intelligence: a proposed framework. Radiology 2020 Jun;295(3):675-682. [doi: 10.1148/radiol.2020192536] [Medline: 32208097]

Abbreviations

AI: artificial intelligence
DL: deep learning
DOR: diagnostic odds ratio
ECG: electrocardiogram
ECHO: echocardiography
EPV: events per variable
LVH: left ventricular hypertrophy
MeSH: Medical Subject Headings
ML: machine learning
NLR: negative likelihood ratio
PLR: positive likelihood ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction Model Risk of Bias Assessment Tool
QUADAS-AI: Quality Assessment of Diagnostic Accuracy Studies Using AI
RoB: risk of bias
SROC: summary receiver operating characteristic
SVM: support vector machine

Edited by A Coristine; submitted 28.Apr.2025; peer-reviewed by N Acharya, N Shah; revised version received 28.Jan.2026; accepted 30.Jan.2026; published 27.Feb.2026.

Please cite as:
Li Y, Zhao K, Wu J
Machine Learning in Left Ventricular Hypertrophy Detection: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e76637
URL: https://www.jmir.org/2026/1/e76637
doi:10.2196/76637

© Yilin Li, Ke Zhao, Jing Wu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Accuracy of Medical Image–Based Deep Learning for Detecting Microvascular Invasion in Hepatocellular Carcinoma: Systematic Review and Meta-Analysis

Wei Feng[1*], MM; Bo Qu[2*], MM; Shuo Han[3], PhD

[1]Department of Ultrasound, the Fourth Affiliated Hospital, China Medical University, Shenyang, China

[2]Department of Urology, Jinqiu Hospital of Liaoning Province, Shenyang, China

[3]Department of Cardiology, the Fourth Affiliated Hospital, China Medical University, No. 4, Chongshan East Road, Huanggu District, Shenyang, China

[*]these authors contributed equally

Corresponding Author:
Shuo Han, PhD
Department of Cardiology, the Fourth Affiliated Hospital, China Medical University, No. 4, Chongshan East Road, Huanggu District, Shenyang, China

Abstract

Background: Hepatocellular carcinoma (HCC) is a leading cause of cancer-related mortality worldwide. Microvascular invasion (MVI) is a critical pathological indicator of postoperative recurrence and poor prognosis in patients with HCC. Some researchers have explored the diagnostic accuracy of deep learning (DL) based on various imaging modalities for MVI.

Objective: This meta-analysis aimed to systematically evaluate the preoperative diagnostic performance of DL models using medical images to predict MVI in HCC, and to investigate the impact of different imaging modalities and validation strategies on model performance and generalizability.

Methods: PubMed, Cochrane Library, Embase, and Web of Science were searched up to October 16, 2025. Studies investigating the detection of MVI in HCC using imaging-based DL techniques were eligible. Studies focusing solely on image segmentation were excluded. The Quality Assessment of Diagnostic Accuracy Studies-2 tool was used to assess risk of bias. A bivariate mixed-effects meta-analysis was performed to calculate the pooled sensitivity, specificity, and area under the summary receiver operating characteristic curve (SROC). Subgroup analyses were conducted by imaging modality and validation set generation method.

Results: This meta-analysis included 52 studies with 19,531 patients with HCC. The pooled analysis revealed that imaging-based DL models had an overall sensitivity of 0.80 (95% CI 0.78 - 0.83), a specificity of 0.82 (95% CI 0.80 - 0.85), and an SROC of 0.88 for MVI prediction. Subgroup analysis showed that models based on preoperative contrast-enhanced computed tomography performed excellently, with a sensitivity of 0.84 (95% CI 0.79 - 0.88), a specificity of 0.83 (95% CI 0.77 - 0.88), and an SROC of 0.90. These results suggest that contrast-enhanced computed tomography is the most promising noninvasive method for current clinical applications. Meanwhile, DL models using pathological sections achieved the highest diagnostic performance: a sensitivity of 0.91 (95% CI 0.87 - 0.94), a specificity of 0.90 (95% CI 0.68 - 0.97), and an SROC of 0.92. This establishes the ultimate benchmark for performance optimization for all noninvasive models. A key finding was that model performance was less consistent in independent external validation (SROC: 0.85) than in internal validation (SROC: 0.90). This discrepancy indicates that overreliance on internal validation may overestimate model efficacy and underscores the decisive role of rigorous external validation in assessing real-world generalizability.

Conclusions: This study is the first to systematically assess the use of imaging-based DL for diagnosing MVI in HCC. The results demonstrate a significant potential for these models in predicting MVI. However, their clinical applicability requires rigorous evaluation, given the scarcity of independent external validation cohorts, notable heterogeneity among them, and the observed decline in model performance. Therefore, prospective, multicenter studies following standardized reporting guidelines are a critical future direction. These studies should also focus on developing integrated algorithms that translate histopathological insights into preoperative imaging data to establish robust clinical tools.

Trial Registration: PROSPERO CRD42024613733; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024613733

(J Med Internet Res 2026;28:e82000) doi:10.2196/82000

KEYWORDS

deep learning; hepatocellular carcinoma; medical imaging; microvascular invasion; artificial intelligence

XSL•FO
RenderX

Introduction

Hepatocellular carcinoma (HCC) is the most common pathological subtype of primary liver cancer, accounting for about 90% of cases [1]. Globally, HCC is the fourth leading cause of cancer-related deaths [2]. According to recent epidemiological data, the age-standardized incidence and mortality rates of HCC are highest in Africa and the Western Pacific region. Over 70% of global HCC cases occur in Asia [1]. Despite advancements in treatment modalities for HCC, including liver transplantation, surgical resection, transarterial chemoembolization, local ablation, targeted therapy, and immunotherapy, the 5-year relative survival rate remains below 20% [3]. Even after complete surgical tumor removal, around 50% - 70% of patients with HCC experience tumor recurrence within 5 years postsurgery [4]. Consequently, HCC has become a significant oncological burden that threatens human life.

Microvascular invasion (MVI) is the pathological process by which tumor cells invade the microvascular structures of the liver tissue surrounding an HCC lesion. MVI occurs in approximately 30% - 50% of cases. It is a significant factor in HCC recurrence after surgery and is associated with poor prognoses in patients with HCC [4,5]. Studies have shown that individuals with HCC and MVI-positive status have significantly lower 5-year disease-free survival and overall survival rates than those with MVI-negative status [6, 7]. Notably, MVI status directly influences treatment strategy selection. Wide-range hepatectomy (resection of ≥ 3 liver segments) is recommended for individuals at high preoperative risk for MVI. This method has a significantly lower 5-year cumulative recurrence rate than limited resection (26.6% vs 58.3%, *P*=.040) [8]. However, other studies have found that adjuvant hepatic arterial infusion chemotherapy does not significantly improve the survival of high-risk MVI-positive individuals compared with the untreated group (*P*=.61). Nevertheless, hepatic arterial infusion chemotherapy significantly improves the prognosis of low-risk patients with MVI (*P*<.001) [9]. Furthermore, individuals with MVI undergoing radiofrequency ablation have a significantly higher recurrence risk than those undergoing radical surgery (*P*<.05) [10]. Therefore, accurately identifying MVI preoperatively is significant for formulating individualized, comprehensive treatment regimens and improving patient prognosis [11].

Currently, a definitive diagnosis of MVI relies on a postoperative pathological examination. However, this examination is subject to biases resulting from the quality of slide preparation and interobserver heterogeneity. These factors may lead to diagnostic inaccuracies. Furthermore, the absence of preoperative MVI information restricts its use in personalized treatment decisions. Therefore, developing efficient MVI auxiliary detection tools is crucial for optimizing clinical management of HCC.

While imaging examinations are crucial for evaluating MVI, predictions based on traditional imaging features rely heavily on radiologists' subjective interpretations. A systematic review and meta-analysis of 19 studies involving 1920 patients revealed that traditional contrast-enhanced features on magnetic resonance imaging (MRI) showed poor overall diagnostic performance in predicting MVI. Only peritumoral enhancement in the arterial phase exhibited moderate diagnostic accuracy. The combined efficacy of other features, such as peritumoral hypointensity in the hepatobiliary phase and irregular margins, was insufficient to meet the requirements for precise preoperative clinical assessment [12]. Recent progress in data mining techniques has accelerated the growth of radiomics. This technique assists in the analysis of imaging features (eg, shape, intensity, and texture) that are difficult for the human eye to perceive. It can overcome some of the limitations of subjectivity. However, radiomic features are mostly low- or mid-level and susceptible to noise interference. They may also not fully reflect tumor heterogeneity [13,14]. Zhang et al [15] noted that radiomics based on single-modality medical imaging is inherently limited. Due to constraints in imaging principles, such approaches can only reflect partial tumor information. Furthermore, when features are extracted using the entire tumor as the region of interest, information about intratumoral heterogeneity is inevitably lost. Additionally, these single-modality radiomic features are susceptible to image noise and variations in scanning parameters, which further compromise the model's ability to capture tumor heterogeneity comprehensively. In contrast, deep learning (DL) uses multi-layer neural networks and an end-to-end learning mode to directly extract multi-level abstract high-order features from original images. This improves the predictive performance, interpretability, and generalizability of models. DL is expected to provide a new paradigm for the preoperative, noninvasive assessment of MVI [16,17]. However, existing studies often focus on a single imaging modality or have small sample sizes. These studies lack a systematic comparison of DL model performance across different imaging modalities, which limits the interpretation of the advantages of DL in detecting MVI and poses challenges to the development or update of intelligent auxiliary diagnostic tools. Consequently, this meta-analysis was conducted to systematically evaluate the diagnostic efficacy of DL models based on medical images for MVI, as well as to explore the impact of different imaging modalities, validation strategies, and sources of heterogeneity on model performance and generalizability, to provide evidence-based support for the development or update of future intelligent auxiliary diagnostic tools.

Methods

Study Registration

This meta-analysis was prospectively registered with the PROSPERO (CRD42024613733). This systematic review and meta-analysis of diagnostic test accuracy was reported in accordance with the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Diagnostic Test Accuracy studies) [18] guidelines in Checklist 1. Due to the absence of subject information collection and its lack of impact on clinical diagnosis and treatment, ethical approval and informed consent were waived.

XSL•FO
RenderX

Eligibility Criteria

Inclusion Criteria

The inclusion criteria are as follows:

- Original research with full text published in English (including cohort, case-control, and cross-sectional studies).
- MVI status in HCC individuals had to be confirmed by histopathology or biopsy.
- Studies had to develop complete DL models based on medical images to detect MVI status in patients with HCC.
- English-language studies.

Exclusion Criteria

The exclusion criteria are as follows:

- Meta-analyses, reviews, guidelines, or expert opinions.
- Only differential factor analysis was implemented without a comprehensive DL model.
- Studies lacking outcome measures of predictive accuracy for the machine learning model, including sensitivity, C-index, accuracy, specificity, precision, F_1-score, and confusion matrix.
- Only image segmentation was performed.

Search Strategies and Data Sources

We conducted a systematic literature search in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for literature searches; completed checklist is available in Checklist 2) [19]. Relevant English-language publications were retrieved from PubMed, Web of Science, the Cochrane Library, and Embase, with the search covering all records up to October 16, 2025. The search strategy used both MeSH (Medical Subject Headings) and free-text keywords, including MeSH terms such as “Carcinoma, Hepatocellular,” “liver cell carcinoma,” and “deep learning.” Boolean operators were used to integrate MeSH terms and free-text terms, constructing tailored search queries for each database. Furthermore, the reference lists of identified review articles were manually screened to locate any additional eligible studies. No prospective study registries were searched, and no attempts were made to obtain unpublished data or to contact study authors for further information. The complete search strings for each database are provided in Table S1 in Multimedia Appendix 1.

Study Selection

Retrieved articles were imported into EndNote. After removing duplicates, the remaining articles were reviewed by title and abstract to identify initially eligible articles. Then, the full texts were downloaded and screened to determine the final eligible articles. Two researchers (WF and BQ, with 6 and 4 years of meta-analysis experience, respectively) performed the review independently. Interresearcher agreement during the literature screening was assessed using the κ coefficient (κ=0.93). Any disagreements were resolved in a consensus meeting with a third researcher (SH, with 10 years of experience in meta-analysis).

Data Extraction

Before data extraction, a standardized spreadsheet was generated. The content to be extracted included the following: publication year, patient source, author, image source, manual segmentation, number of patients with MVI, total number of patients with HCC, number of patients with MVI in the training set, number of patients with HCC in the training set, validation set generation method, number of patients with MVI in the validation set, number of patients with HCC in the validation set, confusion matrix, sensitivity, specificity, precision, and accuracy.

Two researchers (WF and BQ, with 14 and 16 years of medical experience, respectively) carried out the data extraction independently. They then cross-checked their results. Any discrepancies were resolved through consultation with a third researcher (SH, with 13 years of medical experience).

Risk of Bias in Studies

The Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool was used to evaluate the overall risk of bias (ROB) and applicability of the eligible studies. The QUADAS-2 instrument encompasses 4 domains: index test, patient selection, flow and timing, and reference standard. Each domain includes specific questions. Answers to these questions are categorized as “yes,” “no,” or “unclear.” These answers correspond to ROB ratings of “unclear,” “high,” or “low.” If all key questions within a domain received a “yes” answer, the domain was rated as having a low ROB. If any key question received a “no” answer, a potential ROB was indicated, and the researcher judged the ROB according to established guidelines. An “unclear” rating was assigned when the literature did not provide sufficient detail for the researcher to make a judgment.

Two researchers completed the QUADAS-2 assessment independently. Any discrepancies were resolved through discussion with a third researcher.

Synthesis Methods

A meta-analysis was performed using a bivariate mixed-effects model based on diagnostic 2×2 tables [20]. For studies that did not report these tables directly, we derived them from the available specificity, sensitivity, positive predictive value, accuracy, F_1-score, and case numbers. Throughout the analysis, the explicitly defined independent validation cohort from each study served as the unit of analysis. Each data point corresponded to distinct and nonoverlapping patient samples, which ensured the independence of the pooled results and mitigated potential bias from data reuse. Using the bivariate mixed-effects model, we computed the pooled estimates for specificity, sensitivity, negative likelihood ratio (LR–), positive likelihood ratio (LR+), diagnostic odds ratio (DOR), and area under the summary receiver operating characteristic (SROC) curve, along with their 95% CIs [20]. The Spearman correlation coefficient was used to evaluate the threshold effect and its contribution to between-study heterogeneity. Small-study effects were assessed using Deeks funnel plot asymmetry test. For subgroups with fewer than 10 studies, a Doi plot was used to informally assess publication bias. The degree of bias was determined based on the absolute value of the Luis

XSL•FO
RenderX

Furuya-Kanamori (LFK) index. A value less than 1 suggests minor publication bias, a value between 1 and 2 indicates moderate publication bias, and a value exceeding 2 suggests substantial publication bias. During the meta-analysis, the validation set was used, and subgroup analyses were performed according to the validation set generation method and image type. A *P* value less than .05 was considered statistically significant.

Results

Study Selection

Database retrieval yielded 11,266 articles, of which 7539 remained after duplicate removal. Following title and abstract screening, 62 articles were selected for a full-text review. The full-text review subsequently excluded 10 records: 2 non-DL studies, 3 studies with insufficient data to construct diagnostic 2×2 tables, and 5 studies that used DL solely for medical image segmentation without establishing HCC MVI prediction models. Finally, 52 articles [14,16,21-70] met the eligibility criteria. The detailed process is illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the study screening and selection process for this meta-analysis.

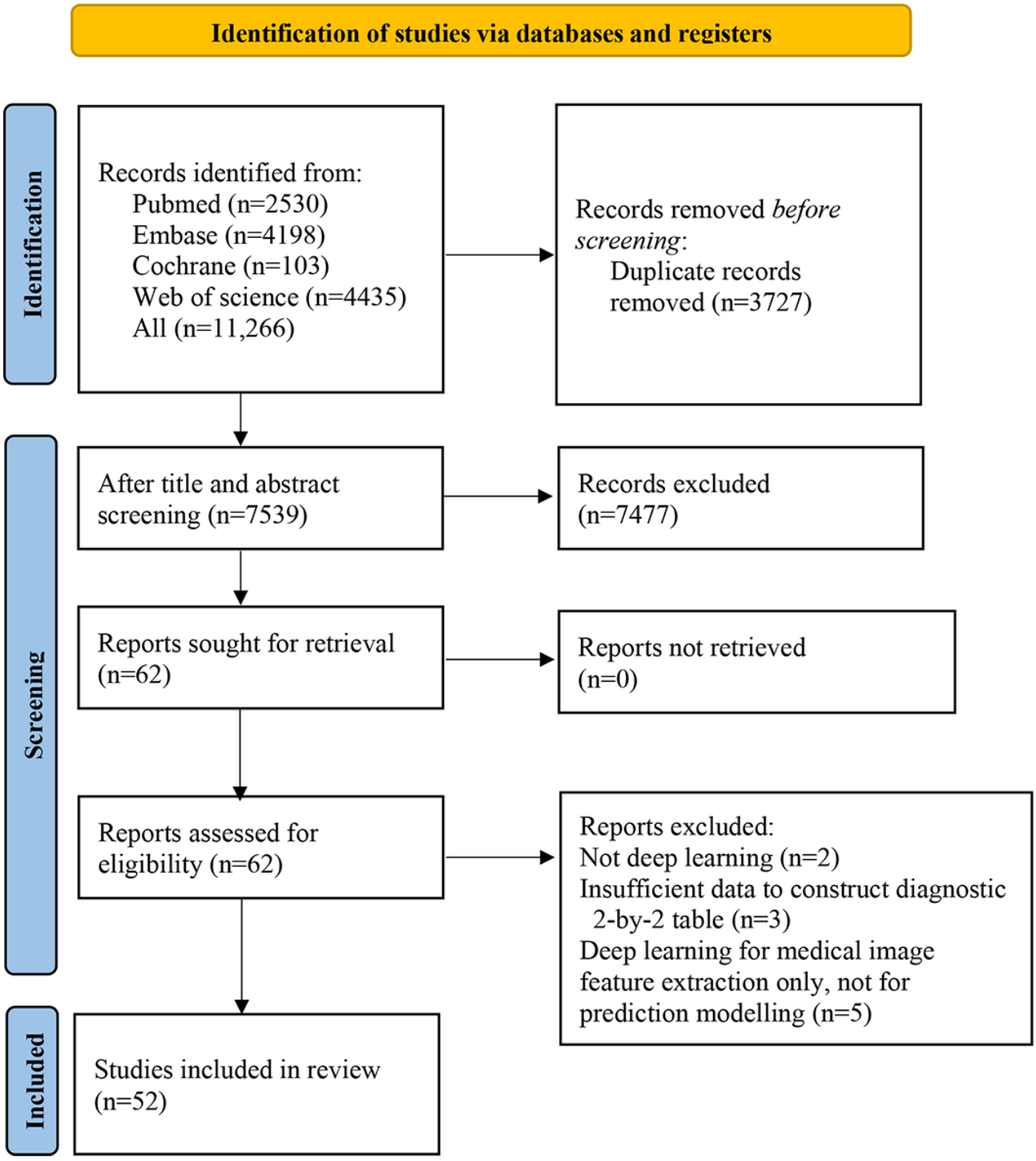

Study Characteristics

This meta-analysis included 52 studies published between 2019 and 2025. All of the studies used histopathological diagnosis as the gold standard for MVI. All 52 studies were case-control studies, encompassing 19,531 individuals with HCC, of whom 8161 were MVI cases. Regarding population source, 32 (61.5%) of the studies were single-center, 19 (36.5%) studies were multicenter, and 1 (1.9%) study was based on a registry database.

These studies primarily used single-modality imaging techniques for image modeling methods: contrast-enhanced computed tomography (CECT, n=16), contrast-enhanced magnetic resonance imaging (CEMRI, n=19), MRI (n=5), contrast-enhanced ultrasound (CEUS, n=5), and pathological sections (n=2). Five additional studies integrated multimodal imaging (computed tomography + positron emission tomography: 1, CECT +CEMRI: 4) for modeling. Regarding image segmentation methods, 38 of the 52 (73.1%) studies used manual segmentation, while the remaining studies used fully automatic segmentation (n=5) or semiautomatic segmentation (n=9).

XSL•FO
RenderX

Concerning model validation strategies, 23 studies used randomly sampled internal validation, 15 used performance through cross-validation, and 14 used external validation with independent cohorts (Table 1).

XSL•FO
RenderX

Table . Basic characteristics of the 52 eligible studies evaluating medical image-based deep learning models for preoperatively detecting microvascular invasion (MVI) in patients with hepatocellular carcinoma.

Author (year of publication)	Basic characteristics						Training versus validation cohort characteristics				
	Source of the patients[a]	Imaging modality	Segmentation	Diagnostic criteria for MVI	Patients with MVI	Sample size	Patients with MVI in the training set	Sample size in the training set	Type of validation	Patients with MVI in the validation set	Sample size in the validation set
Zhang et al (2024)[a] [21]	Single center +registration database (TCGA)[b]	Pathological section	Semiautomatic segmentation	Pathological diagnosis	576	1111	328	530	Random sampling (7:3), external validation (registration database TCGA)	Validation set 1: 143; Validation set 2: 105	Validation set 1: 223; Validation set 2: 358
Lei et al (2024) [22]	Multicenter	CECT[c]+CEMRI[d]	Semiautomatic segmentation	Pathological diagnosis	104	345	89	301	Random sampling (6:4)	15	44
Liu et al (2024) [23]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	123	265	97	211	Random sampling (8:2)	26	54
Zhou et al (2024) [24]	Multicenter	CECT	Semiautomatic segmentation	Pathological diagnosis	54	140	38	98	Random sampling (7:3)	16	42
He et al (2024) [25]	Multicenter	CECT	Semiautomatic segmentation	Pathological diagnosis	299	640	172	368	Internal validation	Validation set 1: 63; Validation set 2: 64	Validation set 1: 134; Validation set 2: 138
Zhong et al (2024) [26]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	82	173	57	120	Random sampling (7:3)	25	53
Wang et al (2024)[a] [27]	Multicenter	CEMRI	Manual segmentation	Pathological diagnosis	274	725	109	234	Random sampling (4:1); External validation (multicenter)	Validation set 1: 20; Validation set 2.1: 82; Validation set 2.2: 37; Validation set 2.3: 26	Validation set 1: 58; Validation set 2.1: 212; Validation set 2.2: 111; Validation set 2.3: 110
Yu et al (2024)[a] [28]	Multicenter	CECT	Semiautomatic segmentation	Pathological diagnosis	78	205	39	119	Random sampling (7:3); External validation (multicenter)	Validation set 1: 15; Validation set 2: 24	Validation set 1: 44; Validation set 2: 42
Ma et al (2025) [29]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	52	117	42	94	Random sampling (8:2); 10-fold cross-validation	10	23

XSL•FO
RenderX

Author (year of publication)	Basic characteristics						Training versus validation cohort characteristics				
	Source of the patients[a]	Imaging modality	Segmentation	Diagnostic criteria for MVI	Patients with MVI	Sample size	Patients with MVI in the training set	Sample size in the training set	Type of validation	Patients with MVI in the validation set	Sample size in the validation set
Zhang et al (2024)[a] [30]	Multicenter	CEUS[e]	Manual segmentation	Pathological diagnosis	219	576	175	461	Single center random sampling (8:2); External validation (multicenter)	—[f]	—
Wang et al (2023)[a] [31]	Multicenter	CECT+CEMRI	Manual segmentation	Pathological diagnosis	150	397	119	297	External validation (multicenter)	31	100
You et al (2023) [32]	Single center	CEMRI	Semiautomatic segmentation	Pathological diagnosis	70	210	56	168	Random sampling (4:1); 5-fold cross-validation	14	42
Qin et al (2023) [33]	Multicenter	CEUS	Manual segmentation	Pathological diagnosis	92	252	71	198	Random sampling (8:2)	21	54
Li et al (2023) [34]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	146	283	117	226	Random sampling (4:1); 5-fold cross-validation	29	57
Cao et al (2023) [35]	Single center	CECT	Manual segmentation	Pathological diagnosis	149	559	120	448	Random sampling (4:1)	29	111
Wang et al (2023) [36]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	109	233	76	163	Random sampling (7:3)	33	70
Ye et al (2023) [37]	Multicenter	CECT+PET[g]	Manual segmentation	Pathological diagnosis	41	100	29	70	Random sampling (7:3); 5-fold cross-validation	12	30
Xu et al (2023) [38]	Single center	CECT	Manual segmentation	Pathological diagnosis	99	305	79	244	Random sampling (8:2); 5-fold cross-validation	20	61
Deng et al (2022) [39]	Single center	CECT+CEMRI	Manual segmentation	Pathological diagnosis	44	103	35	82	Random sampling (4:1); 5-fold cross-validation	9	21
Li et al (2022) [40]	Multicenter	CECT	Manual segmentation	Pathological diagnosis	433	1116	346	892	Random sampling (4:1)	87	224

XSL•FO
RenderX

Author (year of publication)	Basic characteristics						Training versus validation cohort characteristics				
	Source of the patients[a]	Imaging modality	Segmentation	Diagnostic criteria for MVI	Patients with MVI	Sample size	Patients with MVI in the training set	Sample size in the training set	Type of validation	Patients with MVI in the validation set	Sample size in the validation set
Chen et al (2022)[a] [41]	Multicenter	Pathological section	Automatic segmentation	Pathological diagnosis	224	470	137	270	Random sampling external validation (multicenter)	Validation set 1: 43; Validation set 2: 44	Validation set 1: 80; Validation set 2: 120
Zhang et al (2022) [42]	Single center	CEUS	Manual segmentation	Pathological diagnosis	150	436	103	301	Random sampling (3:1); Internal validation	Validation set 1: 35; Validation set 2: 12	Validation set 1: 02; Validation set 2: 33
Liu et al (2022) [43]	Single center	MRI[h]	Manual segmentation	Pathological diagnosis	43	114	28	74	Random sampling	15	40
Sun et al (2022) [44]	Single center	CECT	Semiautomatic segmentation	Pathological diagnosis	134	358	77	193	Random sampling	Validation set 1: 23; Validation set 2: 34	Validation set 1: 61; Validation set 2: 104
Wang et al (2022) [45]	Single center	CECT	Automatic segmentation	Pathological diagnosis	68	138	54	110	Random sampling (8:1:1); 5-fold cross-validation	Validation set 1: 7; Validation set 2: 7	Validation set 1: 14; Validation set 2: 14
Xiao et al (2022)[a] [46]	Multicenter	CECT	Automatic segmentation	Pathological diagnosis	1103	2096	458	876	Random sampling (3:1); External validation (multicenter)	Validation set 1: 152; Validation set 2: 327; Validation set 3: 166	Validation set 1: 292; Validation set 2: 578; Validation set 3: 350
Yang et al (2022) [47]	Single center	CECT	Manual segmentation	Pathological diagnosis	36	283	25	198	Random sampling (198:85)	11	85
Sun et al (2022) [48]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	185	321	86	149	Internal validation	99	172
Dai et al (2022) [49]	Single center	CECT	Manual segmentation	Pathological diagnosis	215	400	172	320	Random sampling (80:10:10)	Validation set 1: 21; Validation set 2: 22	Validation set 1: 40; Validation set 2: 40
Zhang et al (2021) [50]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	92	237	61	158	Random sampling	31	79
Liu et al (2021)[a] [51]	Multicenter	CECT	Manual segmentation	Pathological diagnosis	135	473	68	216	Random sampling (70:30); External validation (multicenter)	Validation set 1: 28; validation set 2: 39	Validation set 1: 93; Validation set 2: 164

XSL•FO
RenderX

Author (year of publication)	Basic characteristics						Training versus validation cohort characteristics				
	Source of the patients[a]	Imaging modality	Segmentation	Diagnostic criteria for MVI	Patients with MVI	Sample size	Patients with MVI in the training set	Sample size in the training set	Type of validation	Patients with MVI in the validation set	Sample size in the validation set
Wei et al (2021)[a] [52]	Multicenter	CECT + CEMRI	Manual segmentation	Pathological diagnosis	270	750	216	635	External validation (prospective, multicenter)	54	115
Zhou et al (2021) [53]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	—	114	—	—	Random sampling	—	—
Wang et al (2021) [54]	Single center	MRI	Manual segmentation	Pathological diagnosis	43	100	24	60	Random sampling	19	40
Zeng et al (2021) [55]	Single center	MRI	Manual segmentation	Pathological diagnosis	38	98	25	64	Random sampling; 4-fold cross-validation	13	34
Gao et al (2021) [56]	Single center	MRI	Manual segmentation	Pathological diagnosis	74	225	56	168	Random sampling	18	57
Jiang et al (2020) [57]	Single center	CECT	Manual segmentation	Pathological diagnosis	220	405	176	324	Random sampling (8:2)	44	81
Song et al (2021) [58]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	225	601	174	461	Random sampling	51	140
Men et al (2019) [59]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	28	63	21	47	4-fold cross-validation	7	16
Zhou et al (2022) [60]	Single center	CECT	Manual segmentation	Pathological diagnosis	145	466	97	311	3-fold cross-validation	48	155
Chu et al (2022) [61]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	51	133	35	93	Random sampling (7:3)	16	40
Huang et al (2022) [62]	Single center	MRI	Manual segmentation	Pathological diagnosis	43	114	32	86	4-fold cross-validation	11	28
Wang et al (2025) [63]	Single center	CEUS	Semiautomatic segmentation	Pathological diagnosis	142	318	99	222	Random sampling; 5-fold cross-validation	43	96
Cen et al (2025) [64]	Single center	CECT	Semiautomatic segmentation	Pathological diagnosis	68	192	47	134	Random sampling (7:3)	21	58
Huang et al (2025) [65]	Multicenter	CEMRI	Manual segmentation	Pathological diagnosis	124	300	87	210	Random sampling; 5-fold cross-validation	37	90

XSL•FO
RenderX

Author (year of publication)	Basic characteristics						Training versus validation cohort characteristics				
	Source of the patients[a]	Imaging modality	Segmentation	Diagnostic criteria for MVI	Patients with MVI	Sample size	Patients with MVI in the training set	Sample size in the training set	Type of validation	Patients with MVI in the validation set	Sample size in the validation set
Miao et al (2025)[a] [66]	Multicenter	CECT	Manual segmentation	Pathological diagnosis	206	483	136	311	Random sampling (8:2); External validation (multicenter)	Validation set 1: 32; Validation set 2: 38	Validation set 1: 77; Validation set 2: 95
Zhu et al (2025)[a] [14]	Multicenter	CEMRI	Manual segmentation	Pathological diagnosis	120	304	90	216	External validation (multicenter, retrospective)	30	88
Dong et al (2025)[a] [67]	Single center	CEMRI	Manual segmentation	Pathological diagnosis	188	519	100	263	Random sampling (4:1); External validation (multicenter)	Validation set 1: 26; Validation set 2: 27; Validation set 3: 35	Validation set 1: 66; Validation set 2: 93; Validation set 3: 97
Zhang et al (2025) [68]	Single center	CEMRI	Automatic segmentation	Pathological diagnosis	142	270	114	216	5-fold cross-validation	28	54
Zheng et al (2025)[a] [16]	Multicenter	CEMRI	Automatic segmentation	Pathological diagnosis	292	589	154	317	Random sampling (7:3); External validation (multicenter)	Validation set 1: 52; Validation set 2: 86	Validation set 1: 106; Validation set 2: 166
Zhao et al (2025)[a] [69]	Multicenter	CEMRI	Manual segmentation	Pathological diagnosis	51	145	25	66	10-fold cross-validation; External validation (multicenter)	26	79
Qin et al (2025) [70]	Single center	CEUS	Manual segmentation	Pathological diagnosis	65	164	44	114	Random sampling (7:3); 10-fold cross-validation	21	50

[a]Studies used external validation with an independent cohort.

[b]TCGA: The Cancer Genome Atlas.

[c]CECT: contrast-enhanced computed tomography.

[d]CEMRI: contrast-enhanced magnetic resonance imaging.

[e]CEUS: contrast-enhanced ultrasound.

[f]Not available.

[g]PET: positron emission tomography.

[h]MRI: magnetic resonance imaging.

XSL•FO
RenderX

ROB in Studies

Regarding patient selection, all eligible studies included consecutive or random cases. According to the QUADAS-2 assessment criteria, this study design carries an inherent high ROB in the "Patient Selection" domain. Consequently, all included studies received a "high" rating for ROB in this domain.

Due to the adoption of supervised DL, model training was based on clear pathological outcomes. However, since DL models predict by extracting inherent image features, rather than directly relying on clinical covariates, their training process's dependence on known outcomes did not result in diagnostic information leakage. This led to a low ROB. Manual segmentation was used in 38 studies, which could have introduced operator subjectivity and led to a high ROB.

Regarding the implementation of the gold standard, all studies used histopathological diagnosis as the gold standard for MVI, ensuring the objectivity and consistency of disease classification. This indicated a low ROB in the implementation of the gold standard.

Regarding the item "the match between the conduct and interpretation of the index test and the review question" in the QUADAS-2 scale, 5 (9.6%) out of the 52 studies did not directly report specificity values. Since specificity is a key indicator for verifying the match between a DL model and a clinical question, the absence of such data may lead to an incomplete assessment of a model's diagnostic efficacy and weaken the reliability of a study's conclusions. Therefore, these 5 studies were determined to have a high ROB. The remaining 47 (90.4%) studies fully reported diagnostic performance indicators, ensuring the transparency in test conduct and interpretation, and indicating a low ROB.

There was a reasonable and appropriate time interval between imaging examinations and pathological diagnoses in all eligible studies. Therefore, there did not appear to be a significant impact on the cases' process (Figure 2).

Figure 2. Methodological quality assessment of the included studies based on the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) scale. (**A**) Summary, (**B**) Individual studies [14,16,21-70].

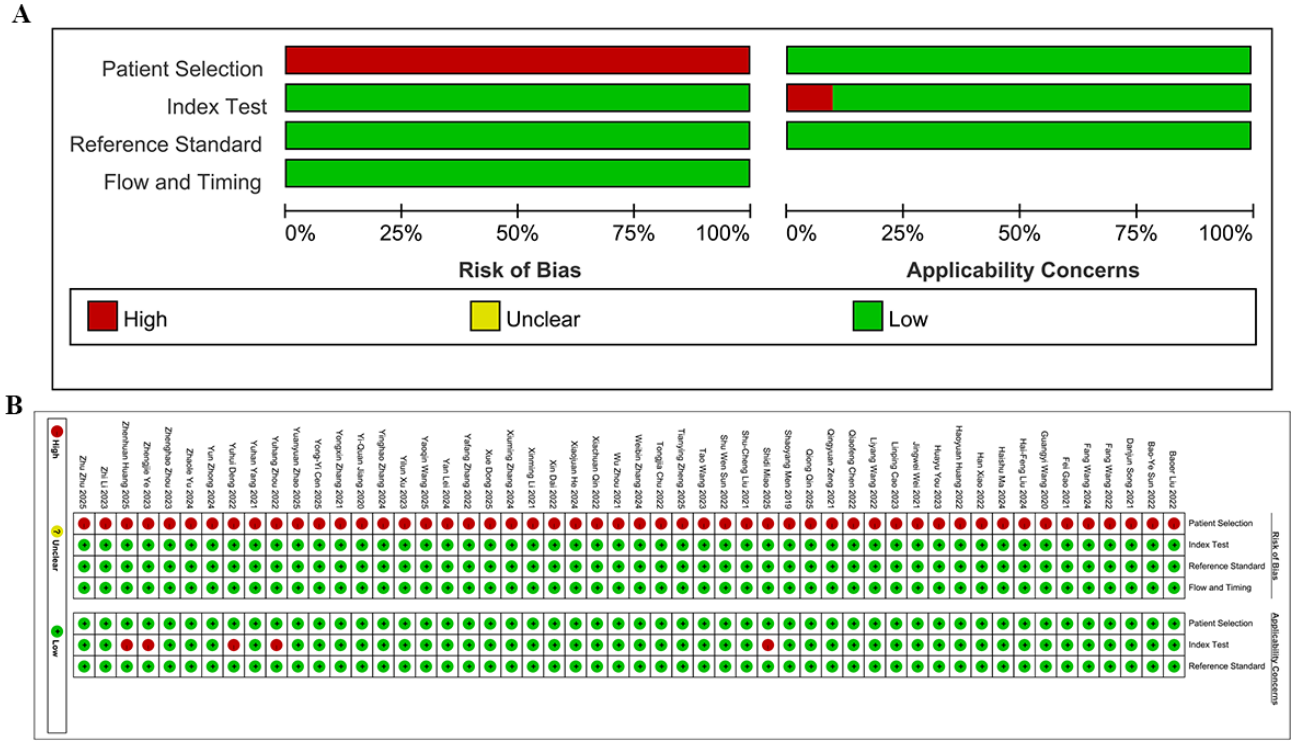

Meta-Analysis

Overall

The model's accuracy was validated using 68 diagnostic fourfold tables. The Spearman correlation coefficient was 0.02, indicating a minimal threshold effect. This effect accounted for none of the observed between-study heterogeneity. The pooled analysis revealed the following results: sensitivity 0.80 (95% CI 0.78 - 0.83, I^2=65.52%), specificity 0.82 (95% CI 0.80 - 0.85, I^2=79.13%), LR+4.6 (95% CI 3.9 - 5.3), LR- 0.24 (95% CI 0.21 - 0.27), DOR 19 (95% CI 15 - 25), and SROC 0.88 (95% CI 0.56 - 0.98; Table 2 and Figures 3 and 4).

XSL•FO
RenderX

Table . Meta-analysis results of deep learning for microvascular invasion diagnosis under different image sources and validation set generation methods.

Subgroup	N	SENS[a] (95% CI)	SPEC[b] (95% CI)	PLR[c] (95% CI)	NLR[d] (95% CI)	DOR[e] (95% CI)	SROC[f] (95% CI)	Deeks	Doi
Overall	68	0.80 (0.78-0.83)	0.82 (0.80-0.85)	4.6 (3.9-5.3)	0.24 (0.21-0.27)	19 (15-25)	0.88 (0.56-0.98)	0.77	__[g]
Validation set generation methods									
Internal validation	49	0.82 (0.79-0.85)	0.83 (0.80-0.86)	4.9 (4.1-6.0)	0.22 (0.18-0.26)	23 (17-32)	0.90 (1.00 - 0.00)	0.85	—
External validation	19	0.77 (0.72-0.82)	0.80 (0.74-0.85)	3.9 (3.0-5.0)	0.29 (0.23-0.36)	13 (9-20)	0.85 (1.00 - 0.00)	0.22	—
The number of cases in the validation set									
<100	38	0.79 (0.75-0.83)	0.83 (0.79-0.87)	4.7 (3.8-5.8)	0.25 (0.20-0.30)	19 (14-26)	0.88 (0.63 - 0.97)	0.15	—
≥100	30	0.82 (0.78-0.85)	0.81 (0.77-0.85)	4.3 (3.4-5.5)	0.23 (0.18-0.28)	19 (13-29)	0.88 (0.66 - 0.97)	0.5	—
Image sources									
CECT[h]	20	0.84 (0.79-0.88)	0.83 (0.77-0.88)	5.0 (3.6-6.9)	0.19 (0.15-0.25)	26 (16-42)	0.90 (1.00 - 0.00)	0.09	
CECT internal validation	16	0.86 (0.80-0.90)	0.85 (0.78-0.90)	5.7 (3.8-8.4)	0.17 (0.12-0.24)	33 (19-59)	0.92 (1.00 - 0.00)	0.19	—
CECT external validation	4	0.82 (0.73-0.88)	0.76 (0.64-0.85)	3.4 (2.3-5.0)	0.24 (0.17-0.34)	14 (9-23)	0.86 (1.00 - 0.00)	—	−1.14
CEUS[i]	5	0.70 (0.58-0.80)	0.88 (0.82-0.92)	5.6 (3.5-9.0)	0.34 (0.23-0.51)	17 (7-37)	0.89 (1.00 - 0.00)	—	0.46
CEMRI[j]	24	0.78 (0.73-0.83)	0.81 (0.76-0.85)	4.0 (3.2-5.0)	0.27 (0.22-0.33)	15 (10-22)	0.86 (1.00 - 0.00)	0.15	
CEMRI internal validation	16	0.81 (0.75-0.86)	0.81 (0.75-0.85)	4.3 (3.2-5.6)	0.24 (0.17-0.32)	18 (11-30)	0.88 (1.00 - 0.00)	0.17	—
CEMRI external validation	8	0.74 (0.65-0.82)	0.80 (0.71-0.87)	3.7 (2.7-5.2)	0.32 (0.24-0.43)	12 (7-18)	0.84 (1.00 - 0.00)	—	1.74
MRI[k]	5	0.76 (0.68-0.82)	0.80 (0.70-0.87)	3.7 (2.5-5.5)	0.30 (0.23-0.40)	12 (7-21)	0.84 (1.00 - 0.00)	—	3.11
Multimodal	10	0.74 (0.68-0.80)	0.79 (0.75-0.83)	3.6 (2.9-4.4)	0.33 (0.26-0.41)	11 (7-16)	0.82 (1.00 - 0.00)	0.44	
Multimodal internal validation	6	0.79 (0.69-0.87)	0.79 (0.73-0.85)	3.8 (2.8-5.3)	0.26 (0.17-0.41)	15 (7-29)	0.86 (1.00 - 0.00)	—	1.6
Multimodal external validation	4	0.71 (0.64-0.78)	0.78 (0.73-0.83)	3.3 (2.6-4.2)	0.36 (0.28-0.47)	9 (6-14)	0.82 (1.00 - 0.00)	—	1.93
Pathological Sections	4	0.91 (0.87-0.94)	0.90 (0.68-0.97)	9.2 (2.5-33.6)	0.09 (0.06-0.15)	97 (20-465)	0.92 (1.00 - 0.00)	—	3.08

[a]SENS: sensitivity.

[b]SPEC: specificity.

[c]PLR: positive likelihood ratio.

[d]NLR: negative likelihood ratio.

[e]DOR: diagnostic odds ratio.

[f]SROC: summary receiver operating characteristic.

XSL•FO
RenderX

[g]Not applicable.

[h]CECT: contrast-enhanced computed tomography.

[i]CEUS: contrast-enhanced ultrasound.

[j]CEMRI: contrast-enhanced magnetic resonance imaging.

[k]MRI: magnetic resonance imaging.

Figure 3. Meta-analysis forest plots: specificity and sensitivity of image-based deep learning in microvascular invasion diagnosis [14,16,21-70].

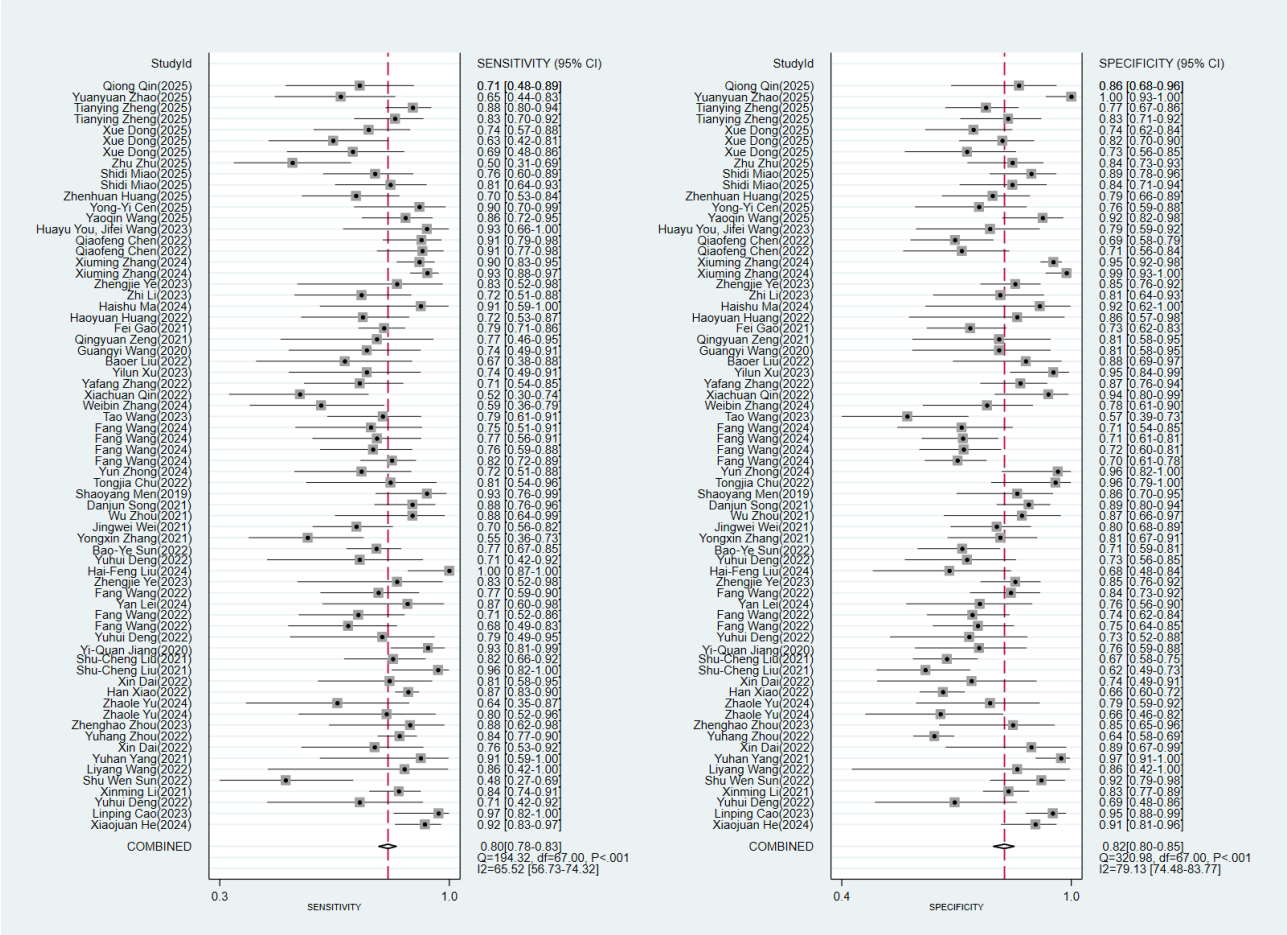

XSL•FO
RenderX

Figure 4. Meta-analysis SROC: specificity and sensitivity of image-based deep learning in microvascular invasion diagnosis. SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

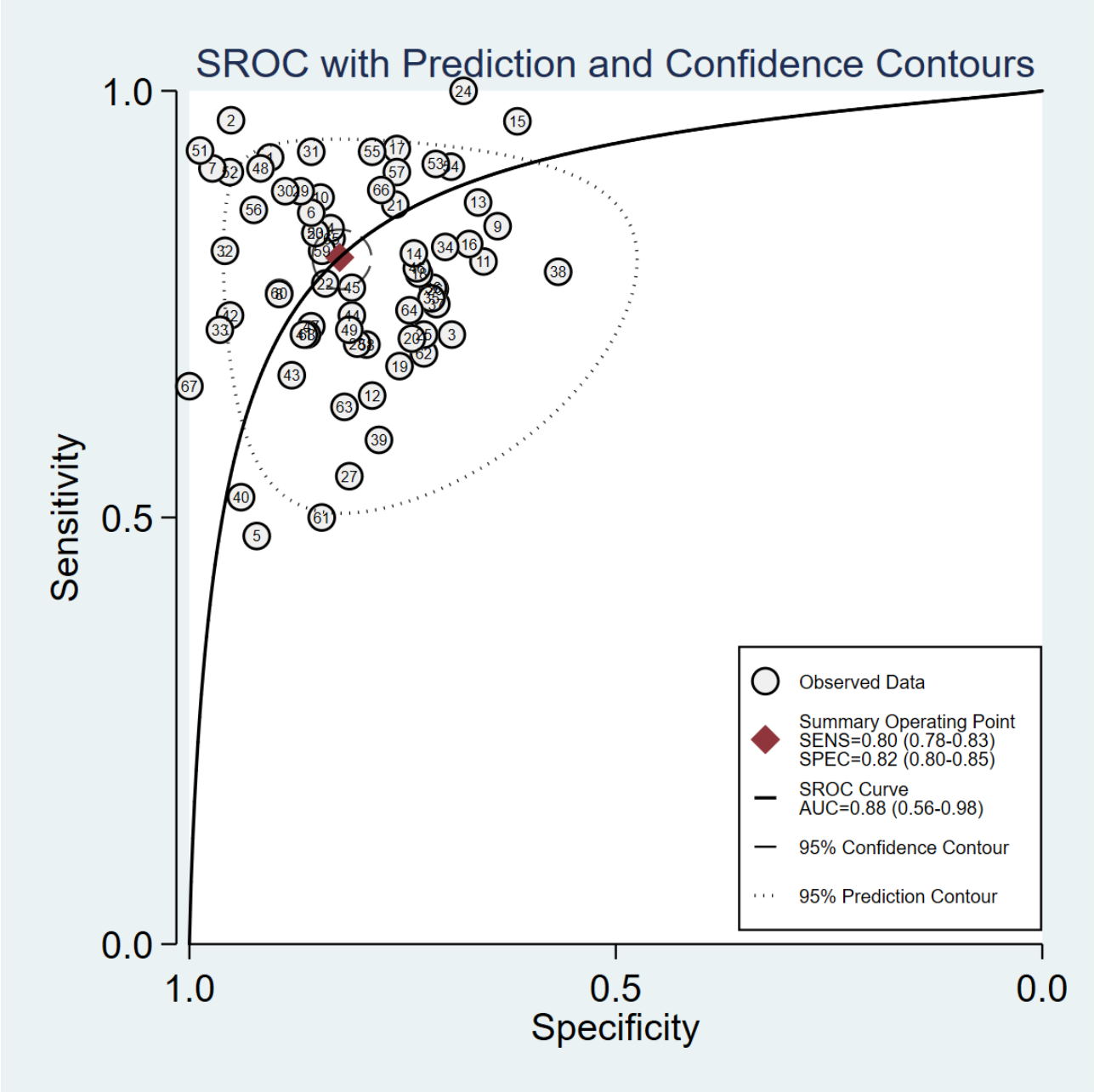

Deeks' funnel plot indicated no significant small-study effects (P=.77, Table 2 and Figure 5). When the prior probability of MVI was set to 40%, the posterior probabilities corresponding to positive and negative DL model detection results were 75% and 14%, respectively. Fagan's nomogram analysis showed that a positive detection result increased the posterior probability by 35% compared with the prior probability, whereas a negative detection result decreased it by 26% (Figure 6).

XSL•FO
RenderX

Figure 5. Deeks' funnel plot from meta-analysis of the specificity and sensitivity of image-based deep learning in microvascular invasion diagnosis.

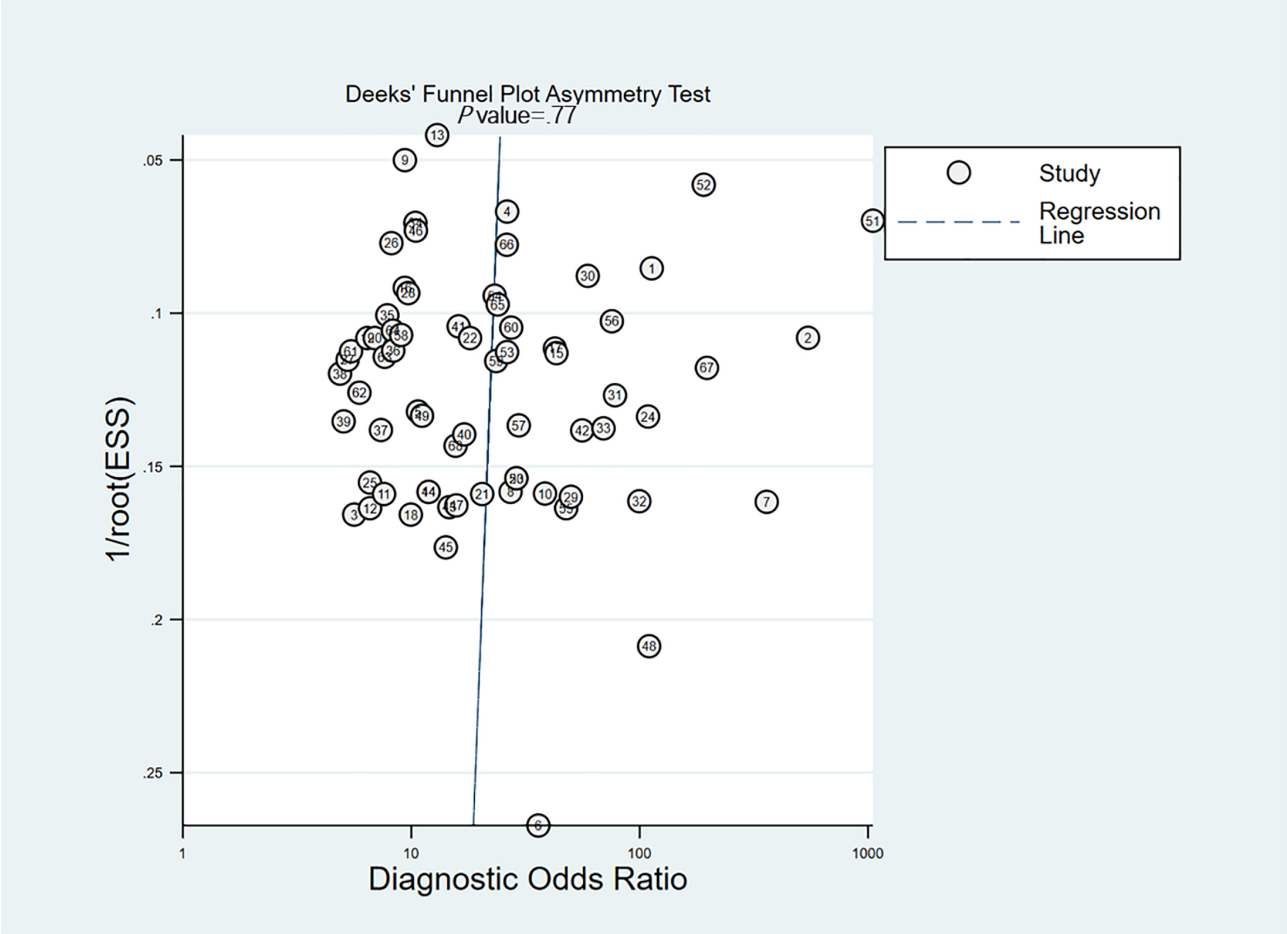

XSL•FO
RenderX

Figure 6. Fagan's nomogram from meta-analysis of the specificity and sensitivity of image-based deep learning in microvascular invasion diagnosis.

XSL•FO
RenderX

Subgroup Analysis by Validation Set Generation Method

Internal Validation

The model's accuracy was validated using 49 diagnostic fourfold tables in internal validation. The Spearman correlation coefficient was 0.04, suggesting a minimal threshold effect. This effect accounted for none of the observed between-study heterogeneity. The pooled analysis yielded the following results: sensitivity 0.82 (95% CI 0.79 - 0.85, I^2=61.27%), specificity 0.83 (95% CI 0.80 - 0.86, I^2=72.22%), LR+4.9 (95% CI 4.1 - 6.0), LR–0.22 (95% CI 0.18 - 0.26), DOR 23 (95% CI 17 - 32), and SROC 0.90 (95% CI 1.00 - 0.00; Table 2 and Figures S1A,B in Multimedia Appendix 2).

Deeks' funnel plot illustrated no significant small-study effects (*P*=.85, Table 2 and Figure S1C in Multimedia Appendix 2). When the prior probability of MVI was set to 40%, the corresponding posterior probabilities for positive and negative DL model detection results were 77% and 13%, respectively. Fagan's nomogram analysis showed that a positive detection result increased the posterior probability by 37%, compared with the prior probability; a negative detection result decreased the posterior probability by 27% (Figure S1D in Multimedia Appendix 2).

External Validation

The model's accuracy was validated using 19 diagnostic fourfold tables in external validation. The Spearman correlation coefficient was −0.15, indicating a minimal threshold effect. This effect accounted for 2% of the observed between-study heterogeneity. The pooled analysis yielded the following results: sensitivity 0.77 (95% CI 0.72 - 0.82, I^2=73.40%), specificity 0.80 (95% CI 0.74 - 0.85, I^2=83.19%), LR+3.9 (95% CI 3.0 - 5.0), LR–0.29 (95% CI 0.23 - 0.36), DOR 13 (95% CI 9 - 20), and SROC 0.85 (95% CI 1.00 - 0.00; Table 2 and Figures S2A,B in Multimedia Appendix 3).

Deeks' funnel plot revealed no significant small-study effects (*P*=.22, Table 2 and Figure S2C in Multimedia Appendix 3). When the prior probability of MVI was set to 39%, the corresponding posterior probabilities for positive and negative DL model detection results were 71% and 15%, respectively. Fagan's nomogram analysis showed that a positive detection result increased the posterior probability by 32% compared with the prior probability, whereas a negative detection result decreased it by 24% (Figure S2D inMultimedia Appendix 3).

Validation Set Size < 100 Cases

Among studies with a validation set size <100 cases, the model's accuracy was validated using 38 diagnostic fourfold tables. The Spearman correlation coefficient was −0.44, suggesting a minimal threshold effect. This effect accounted for 19% of the observed between-study heterogeneity. The pooled analysis yielded the following results: sensitivity 0.79 (95% CI 0.75 - 0.83, I^2=56.49%), specificity 0.83 (95% CI 0.79 - 0.87, I^2=67.03%), LR+4.7 (95% CI 3.8 - 5.8), LR−0.25 (95% CI 0.20 - 0.30), DOR 19 (95% CI 14 - 26), and SROC 0.88 (95% CI 0.63 - 0.97; Table 2 and Figure S3A,B in Multimedia Appendix 4).

Deeks' funnel plot revealed no significant small-study effects (*P*=.15, Table 2 and Figure S3C in Multimedia Appendix 4). When the prior probability of MVI was set to 41%, the corresponding posterior probabilities for positive and negative DL model detection results were 77% and 15%, respectively. Fagan's nomogram analysis displayed that a positive detection result increased the posterior probability by 36% compared with the prior probability, whereas a negative detection result decreased it by 26% (Figure S3D in Multimedia Appendix 4).

Validation Set Size ≥ 100 Cases

Among studies with a validation set size ≥100 cases, the model's accuracy was validated using 30 diagnostic fourfold tables. The Spearman correlation coefficient was 0.47, indicating a minimal threshold effect. This effect accounted for 22% of the observed between-study heterogeneity. The pooled analysis yielded the following results: sensitivity 0.82 (95% CI 0.78 - 0.85, I^2=69.35%), specificity 0.81 (95% CI 0.77 - 0.85, I^2=85.58%), LR+4.3 (95% CI 3.4 - 5.5), LR−0.23 (95% CI 0.18 - 0.28), DOR 19 (95% CI 13 - 29), and SROC 0.88 (95% CI 0.66 - 0.97; Table 2 and Figure S4A,B in Multimedia Appendix 5).

Deeks' funnel plot revealed no significant small-study effects (*P*=.50; Table 2 and Figure S4C in Multimedia Appendix 5). When the prior probability of MVI was set to 39%, the corresponding posterior probabilities for positive and negative DL model detection results were 73% and 13%, respectively. Fagan's nomogram analysis showed that a positive detection result increased the posterior probability by 34% compared with the prior probability, whereas a negative detection result decreased it by 26% (Figure S4D in Multimedia Appendix 5).

Subgroup Analysis by Image Source

DL Based on CECT

In CECT, the model's accuracy was validated using 20 diagnostic fourfold tables. The Spearman correlation coefficient was −0.28, suggesting a minimal threshold effect. This effect accounted for 8% of the observed between-study heterogeneity. The pooled analysis revealed the following results: sensitivity 0.84 (95% CI 0.79 - 0.88, I^2=71.67%), specificity 0.83 (95% CI 0.77 - 0.88, I^2=87.84%), LR+5.0 (95% CI 3.6 - 6.9), LR−0.19 (95% CI 0.15 - 0.25), DOR 26 (95% CI 16 - 42), and SROC 0.90 (95% CI 1.00 - 0.00; Table 2 and Figure S5 in Multimedia Appendix 6). Deeks funnel plot revealed no significant small-study effects (*P*=.09, Table 2).

Among CECT-based models with internal validation, the accuracy was validated using 16 diagnostic fourfold tables. The Spearman correlation coefficient was −0.26, indicating a minimal threshold effect. This effect accounted for 7% of the observed between-study heterogeneity. The pooled analysis yielded the following results: sensitivity 0.86 (95% CI 0.80 - 0.90, I^2=71.85%), specificity 0.85 (95% CI 0.78 - 0.90, I^2=88.74%), LR+5.7 (95% CI 3.8 - 8.4), LR−0.17 (95% CI 0.12 - 0.24), DOR 33 (95% CI 19 - 59), and SROC 0.92 (95% CI 1.00 - 0.00; Table 2 and Multimedia Appendix 7). Deeks' funnel plot revealed no significant small-study effects (*P*=.19, Table 2).

XSL•FO
RenderX

Among CECT-based models with external validation, the accuracy was validated using four diagnostic fourfold tables. The pooled analysis yielded the following results: sensitivity 0.82 (95% CI 0.73 - 0.88, I^2=62.47%), specificity 0.76 (95% CI 0.64 - 0.85, I^2=77.88%), LR+3.4 (95% CI 2.3 - 5.0), LR− 0.24 (95% CI 0.17 - 0.34), DOR 14 (95% CI 9 - 23), and SROC 0.86 (95% CI 1.00 - 0.00; Table 2 and Multimedia Appendix 7). Subsequent Doi plot analysis showed moderate publication bias among the included studies (LFK index=−1.14, Table 2).

DL Based on CEUS

In CEUS, the model's accuracy was validated using 5 diagnostic fourfold tables. The pooled analysis yielded the following results: sensitivity 0.70 (95% CI 0.58 - 0.80, I^2=58.75%), specificity 0.88 (95% CI 0.82 - 0.92, I^2=30.18%), LR+5.6 (95% CI 3.5 - 9.0), LR−0.34 (95% CI 0.23 - 0.51), DOR 17 (95% CI 7 - 37), and SROC 0.89 (95% CI 1.00 - 0.00; Table 2 and Multimedia Appendix 8). Analysis using the Doi plot revealed minimal publication bias among the included studies (LFK index=0.46, Table 2).

DL Based on CEMRI

In CEMRI, the model's accuracy was validated using 24 diagnostic fourfold tables. The Spearman correlation coefficient was −0.22, suggesting a minimal threshold effect. This effect accounted for 5% of the observed between-study heterogeneity. The pooled analysis yielded the following results: sensitivity 0.78 (95% CI 0.73 - 0.83, I^2=61.70%), specificity 0.81 (95% CI 0.76 - 0.85, I^2=65.14), LR+4.0 (95% CI 3.2 - 5.0), LR− 0.27 (95% CI 0.22 - 0.33), DOR 15 (95% CI 10 - 22), and SROC 0.86 (95% CI 1.00 - 0.00; Table 2 and Multimedia Appendix 9). Deeks funnel plot revealed no significant small-study effects (P=0.15, Table 2).

Among CEMRI models with internal validation, the accuracy was validated using 16 diagnostic fourfold tables. The Spearman correlation coefficient was 0.17, indicating a minimal threshold effect. This effect accounted for 3% of the observed between-study heterogeneity. The pooled analysis yielded the following results: sensitivity 0.81 (95% CI 0.75 - 0.86, I^2=57.28%), specificity 0.81 (95% CI 0.75 - 0.85, I^2=61.32), LR+4.3 (95% CI 3.2 - 5.6), LR− 0.24 (95% CI 0.17 - 0.32), DOR 18 (95% CI 11 - 30), and SROC 0.88 (95% CI 1.00 - 0.00; Table 2 and Figure S9 in Multimedia Appendix 10). Deeks funnel plot revealed no significant small-study effects (P=.17, Table 2).

Among CEMRI models with external validation, the accuracy was validated using 8 diagnostic fourfold tables. The Spearman correlation coefficient was −0.72, suggesting a significant threshold effect. This effect accounted for 52% of the observed between-study heterogeneity. The pooled analysis yielded the following results: sensitivity 0.74 (95% CI 0.65 - 0.82, I^2=70.90%), specificity 0.80 (95% CI 0.71 - 0.87, I^2=72.55%), LR+3.7 (95% CI 2.7 - 5.2), LR−0.32 (95% CI 0.24 - 0.43), DOR 12 (95% CI 7 - 18), and SROC 0.84 (95% CI 1.00 - 0.00; Table 2 and Figure S9 in Multimedia Appendix 10). Further Doi plot analysis revealed moderate publication bias among the included studies (LFK index=1.74, Table 2).

DL Based on MRI

In MRI, the model's accuracy was validated using 5 diagnostic fourfold tables. The pooled analysis revealed the following results: sensitivity 0.76 (95% CI 0.68 - 0.82, I^2=0.00%), specificity 0.80 (95% CI 0.70 - 0.87, I^2=0.00%), LR+3.7 (95% CI 2.5 - 5.5), LR−0.30 (95% CI 0.23 - 0.40), DOR 12 (95% CI 7 - 21), and SROC, 0.84 (95% CI 1.00-0.00; Table 2 and Multimedia Appendix 11). Analysis using the Doi plot revealed substantial publication bias among the included studies (LFK index=3.11, Table 2).

DL Based on Multimodal Imaging

In multimodal medical images, the model's accuracy was validated using 10 diagnostic fourfold tables. The pooled analysis yielded the following results: sensitivity 0.74 (95% CI 0.68 - 0.80, I^2=0.00%), specificity 0.79 (95% CI 0.75 - 0.83, I^2=11.32%), LR+3.6 (95% CI 2.9 - 4.4), LR− 0.33 (95% CI 0.26 - 0.41), DOR 11 (95% CI 7 - 16), and SROC 0.82 (95% CI 1.00 - 0.00; Table 2 and Multimedia Appendix 12). Deeks' funnel plot revealed no significant small-study effects (P=.44, Table 2).

In multimodal internal validation, the model's accuracy was validated using 6 diagnostic fourfold tables. The pooled analysis revealed the following results: sensitivity 0.79 (95% CI 0.69 - 0.87, I^2=0.00%), specificity 0.79 (95% CI 0.73 - 0.85, I^2=31.69%), LR+3.8 (95% CI 2.8 - 5.3), LR− 0.26 (95% CI 0.17 - 0.41), DOR 15 (95% CI 7 - 29), and SROC 0.86 (95% CI 1.00 - 0.00; Table 2 and Multimedia Appendix 13). Analysis using the Doi plot revealed moderate publication bias among the included studies (LFK index=1.6; Table 2).

In multimodal external validation, the model's accuracy was validated using 4 diagnostic fourfold tables. The pooled analysis showed the following results: sensitivity 0.71 (95% CI 0.64 - 0.78, I^2=0.00%), specificity 0.78 (95% CI 0.73 - 0.83, I^2=0.00%), LR+3.3 (95% CI 2.6 - 4.2), LR− 0.36 (95% CI 0.28 - 0.47), DOR 9 (95% CI 6 - 14), and SROC 0.82 (95% CI 1.00 - 0.00; Table 2 and Multimedia Appendix 13). Subsequent Doi plot analysis indicated moderate publication bias (LFK index=1.93; Table 2).

DL Based on Pathological Sections

In pathological sections, the model's accuracy was validated using 4 diagnostic fourfold tables. The pooled analysis revealed the following results: sensitivity 0.91 (95% CI 0.87 - 0.94, I^2=0.00%), specificity 0.90 (95% CI 0.68 - 0.97, I^2=94.83%), LR+9.2 (95% CI 2.5 - 33.6), LR− 0.09 (95% CI 0.06 - 0.15), DOR 97 (95% CI 20 - 465), and SROC 0.92 (95% CI 1.00 - 0.00; Table 2 and Multimedia Appendix 14). Further Doi plot analysis revealed substantial publication bias (LFK index=3.08; Table 2).

Discussion

Summary of the Main Findings

The current meta-analysis revealed that the modeling methods used for DL-based prediction of HCC MVI primarily used

XSL•FO
RenderX

CECT, CEUS, CEMRI, MRI, multimodal imaging, and pathological image techniques. DL models based on medical imaging showed favorable overall diagnostic performance in predicting HCC MVI, with a pooled sensitivity of 0.80 (95% CI 0.78 - 0.83) and specificity of 0.82 (95% CI 0.80 - 0.85). Further analysis across imaging modalities revealed that CECT-based models achieved the highest diagnostic efficacy, showing a sensitivity of 0.84 (95% CI 0.79 - 0.88) and specificity of 0.83 (95% CI 0.77 - 0.88). Models based on CEUS exhibited particularly high specificity (0.88, 95% CI 0.82 - 0.92). Furthermore, models using pathological slides, considered the diagnostic reference standard, attained the highest overall performance, with a sensitivity of 0.91 (95% CI 0.87 - 0.94) and specificity of 0.90 (95% CI 0.68 - 0.97). Therefore, these models appear promising as a diagnostic approach for MVI in HCC.

Comparison With Other Previous Reviews

This study also noted that some researchers have discussed the use of machine learning for MVI in HCC. Xiao et al [71] and Liang et al [72], who focused on MRI and ultrasound radiomics, respectively, validated the predictive potential of single modalities, with pooled area under the curves (AUCs) of 0.87 and 0.81, respectively. However, their analyses included a limited number of studies and focused only on a single imaging modality, limiting the generalizability of their conclusions. Li et al [73] integrated multiple imaging modalities across 22 studies (involving 4129 participants), reporting a pooled AUC of 0.90 for radiomic models. However, their analysis primarily incorporated traditional machine learning models and did not stratify performance based on the validation set generation method. This limitation may lead to an overly optimistic assessment of model generalizability.

Unlike previous systematic reviews that focused on traditional machine learning models, our meta-analysis focuses on the value of DL algorithms for diagnosing HCC MVI. Methodologically, this study provides a deeper exploration of the sources of performance heterogeneity through subgroup analyses based on the validation set generation method and image source. A particular methodological strength is the inclusion of pathological sections, the diagnostic gold standard, as a benchmark for performance optimization. This meta-analysis synthesized data from 52 studies involving 19,531 patients and provided more robust and reliable conclusions than those from analyses with smaller sample sizes.

Influence of Imaging Modalities on DL

The imaging modalities used to construct DL models can be categorized into 2 main types: noninvasive and invasive. This study showed that the properties of different modalities and their ability to extract biological features directly affected the diagnostic efficacy of corresponding models.

Among noninvasive imaging modalities, CECT can effectively capture tumor heterogeneity, enhancement patterns, and peritumoral microenvironment changes due to its high spatial resolution and multi-phase dynamic imaging capabilities. These features support the superior diagnostic performance of DL models (AUC=0.90) [74,75]. Models developed from CEUS achieved commendable performance (AUC=0.89) and exhibited a high pooled specificity (0.88). This finding suggests that CEUS's real-time hemodynamic properties may be valuable in ruling out MVI-negative cases. However, its comparatively lower sensitivity (0.70) concurrently suggests ongoing challenges in consistently identifying MVI-positive features. Integrating Sonazoid-based functional imaging with conventional ultrasound characteristics and serum markers in the future may be a promising way to improve performance [76]. CEMRI has unique advantages in depicting tumor boundaries and detecting subtle changes in the peritumoral liver parenchyma due to its excellent soft tissue contrast [77]. The model's performance (sensitivity: 0.78, specificity: 0.81, AUC=0.86) surpasses the assessment that relies on traditional visual features, as reported by Wu et al [78] (sensitivity: 0.55, specificity: 0.87, AUC=0.80). However, its widespread adoption is limited by long examination times, high costs, and dependence on equipment. Nonenhanced MRI models also demonstrate predictive potential (sensitivity: 0.76, specificity: 0.80, AUC=0.84). Notably, multimodal fusion models, which are designed to integrate complementary information, have not demonstrated significant advantages in the studies included (AUC=0.82). This may be due to the simplified fusion strategies or data heterogeneity.

Among invasive imaging modalities, DL models based on pathological sections demonstrated the highest diagnostic efficacy in this study (AUC=0.92). This highlights the gold-standard status of pathology in evaluating MVI. However, the inherent invasiveness of pathological examination precludes its use in preoperative decision-making. Therefore, a central challenge for future research is to effectively transfer and integrate the gold-standard-level diagnostic insights from pathological sections into preoperative, noninvasive imaging. This objective requires leveraging advanced methodologies, such as transfer learning, cross-modal fusion, and generative models. Augmenting existing advantageous modalities, such as CECT, with these advancements could ultimately pave the way for a clinically viable preoperative "virtual biopsy."

Image Segmentation

Accurate image segmentation is essential for building reliable DL models. However, the literature incorporated in this meta-analysis primarily relies on manual segmentation. This method can introduce subjective bias (38/52, 73.1%). While advanced network architectures have achieved expert-level precision in segmenting normal liver tissue, as demonstrated by Dice coefficients ranging from 0.968 to 0.982 [79,80], comparative data from the Liver Tumor Segmentation challenge reveal significant discrepancies. Specifically, the accuracy of liver tumor segmentation (Dice: 0.739) is considerably lower than that of liver parenchyma segmentation (Dice: 0.963) [81]. This discrepancy primarily stems from the heterogeneity of HCC lesions, suboptimal image contrast, and a lack of high-quality annotated data. Concurrently, selecting the appropriate segmentation strategy is critical. While 3D segmentation comprehensively captures spatial heterogeneity, the more clinically feasible 2D approach sacrifices substantial volumetric information [82]. Given the demand for submillimeter-level precision in MVI prediction, there are 2

XSL•FO
RenderX

key technological pathways for enhancing model stability. First, exploring segmentation paradigms that require fewer annotations, such as weakly or self-supervised learning, can reduce dependency on annotations. Second, developing novel network architectures designed specifically to address HCC heterogeneity and boundary ambiguity is equally crucial.

Validation Set Generation Method

The rigor of validation strategies is paramount for evaluating the real-world generalizability of DL models in predicting MVI of HCC [83,84]. The present analysis reveals that, despite exemplary performance during internal validation, consistent performance declines emerge in independent external validation cohorts. This finding clearly shows that relying too much on internal validation can lead to overestimating a model's true efficacy [85]. Concern regarding validation strategies is not unique to DL research. A recent meta-analysis [86] that focused on MRI-based radiomics for predicting HCC recurrence and MVI similarly concluded that the current predominance of internal validation results in an overestimation of model generalizability as well. Consequently, any proclaimed superior performance may substantially diminish when confronting real-world heterogeneity, if the evaluation framework remains confined to internal validation, irrespective of the underlying algorithm—be it DL or radiomics. Therefore, promoting rigorous external validation and establishing standardized, cross-institutional imaging protocols are essential steps toward reliable clinical translation in this field.

Heterogeneity Analysis

There is substantial heterogeneity among the included studies. Notably, high levels of heterogeneity persist within subgroups, even after stratifying analyses by imaging modality and validation strategy. This observation objectively reflects the inherent complexity of artificial intelligence (AI)–based medical imaging research and is a common challenge for meta-analyses in this field. The heterogeneity primarily stems from 3 levels. Technically, variations in critical parameters, including imaging equipment, magnetic field strength, and slice thickness, directly influence image texture and quality. These variations are a significant technical source of variability in model performance. Methodologically, diversity in study design, validation strategies, segmentation techniques, and DL network architectures introduces additional variation in model construction and performance interpretation. Clinically, differences in patient populations regarding geographic distribution, underlying liver disease etiology, and disease stage may also affect model performance and generalizability. While this heterogeneity limits the direct interpretability of the pooled results to some extent, it accurately reflects the diversity in methodology and clinical practice within the field. Future investigations should adhere to the Findable, Accessible, Interoperable, and Reusable principles, providing detailed reporting of imaging acquisition parameters, model architectures, and training specifics. Such comprehensive reporting will facilitate the in-depth exploration of heterogeneity sources via methods such as meta-regression. This will promote the identification of key influencing factors and the standardization of methodologies.

Methods for ROB Assessment

This systematic review used a composite strategy to assess the ROB. First, all included studies were rigorously evaluated according to the QUADAS-2 guidelines. The results showed that all primary studies were rated as having a high ROB in the "Patient Selection" domain of QUADAS-2 due to the widespread use of the retrospective case-control design. While this outcome aligns with the QUADAS-2 assessment principles, it also reveals a limitation of the tool when evaluating machine learning-based diagnostic studies that use retrospective data. The tool struggles to differentiate nuances in data construction quality among studies.

To conduct a more granular assessment of data-level bias risk, a supplemental analysis was performed using items from the Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence (QUADAS-AI) tool targeting "Study Participant Selection." The QUADAS-AI tool provides more detailed criteria for this dimension, including an explicit description of data source, size, and quality characteristics, use of open-source datasets, a clear rationale for splitting data into training, validation, and test sets, performance of image preprocessing, and provision of scanner model information. The analysis using QUADAS-AI items showed that all 52 studies appropriately described the source, size, and quality of the input data and clearly defined the patient inclusion criteria. Among these studies, only one used an open-source dataset. All studies provided a rationale for the data split. Image preprocessing was performed in all studies. However, 11 studies did not report the scanner model used for image acquisition (Table S2 in Multimedia Appendix 15 [14,16,21-70]).

Advantages and Limitations

A primary strength of this research is that it is the first large-scale, systematic meta-analysis to evaluate medical imaging-based DL models for predicting MVI in HCC. The analysis included 52 studies with 19,531 patients, providing the field with comprehensive evidence. The analytical process strictly adhered to the PRISMA guidelines, and bias risk was evaluated using QUADAS-2. These measures ensured methodological rigor and transparency. However, several limitations warrant consideration. First**,** the training cohorts in most primary studies were small. Only 3 studies had a sample size greater than 1000 patients. The robustness of DL models depends heavily on large volumes of high-quality data. Therefore, restricted training sample sizes are a potential methodological limitation. This likely contributes to overfitting in some models, which is probably an internal reason for the observed performance degradation during external validation. This finding underscores the fundamental importance of acquiring large-scale, high-quality datasets to enhance model generalizability in the development of current DL models. Second, the 95% CIs for the pooled AUCs and for most subgroup analyses were exceptionally wide. This constrained the interpretation of the result precision to some extent. This primarily stems from the significant heterogeneity among the included studies. The limited number of studies in subgroup analyses exacerbated data sparsity. Third, Doi plot analyses for some subgroups indicated moderate to substantial publication

XSL•FO
RenderX

bias. These analyses showed that the pooled results for these subgroups may be influenced by unpublished negative studies, which poses a risk of overestimating diagnostic performance. Fourth, the rigor of the validation strategies needs to be improved. Most studies relied on internal validation. Only a few conducted stringent external validation. This reliance may introduce an optimistic bias into the overall assessment of the models' real-world generalizability. While the overall performance on external validation sets was discussed, the limited quantity of external validation data prevented a more in-depth subgroup analysis of validation strategies by different imaging modalities. Fifth, the vast majority of primary studies inadequately reported model calibration metrics or details about network complexity. This omission hinders a quantitative, systematic evaluation of predictive reliability and overfitting risk at the review level and impacts the comprehensive assessment of clinical applicability. This reflects a general deficiency in the transparency of methodological reporting within the current research landscape.

Clinical Implications and Future Perspectives

This study indicates that medical imaging-based DL models, particularly those using preoperative CECT, demonstrate promising diagnostic performance in predicting MVI of HCC. These models have the potential to assist in personalized surgical planning. However, translating them into clinical practice faces multiple challenges. One primary issue is establishing clinical decision thresholds. While this analysis quantified predictive probabilities via Fagan nomograms, there is currently a lack of evidence-based guidelines defining "at what predicted probability of MVI the surgical margin should be adjusted." Future work must integrate clinical outcome data and empirically explore the net benefit of different thresholds using methods such as decision curve analysis. Second, these models' generalizability and deployment feasibility need urgent enhancement. Performance degradation during external validation suggests susceptibility to variations in imaging protocols, equipment, and patient populations. Furthermore, model interpretability is crucial for gaining clinical trust. This necessitates developing transparent methods for presenting decision rationale.

To bridge the gap between "high performance" and "high utility," future efforts must focus on 3 interconnected levels. At the research level, the focus must shift from model construction to rigorous, prospective, multicenter validation to unequivocally assess generalizability. At the algorithmic level, it is essential to explore cross-modal information fusion, especially using transfer learning, to bring "gold standard"–level diagnostic insights from histopathological sections to preoperative, noninvasive imaging. At the clinical level, establishing standardized, cross-institutional imaging protocols and developing decision support systems that integrate seamlessly into clinical workflows is imperative. This integrated approach is vital for reliably translating technology from innovation to tangible patient benefit.

Conclusions

This systematic review and meta-analysis demonstrate that medical imaging-based DL models, especially those leveraging preoperative CECT, hold significant promise for the noninvasive preoperative prediction of MVI in HCC. Unlike previous reviews that focused on radiomics or single imaging modalities, this study conducted a comprehensive comparison across multiple modalities. The study also emphasizes the critical role of external validation in the real-world generalizability of a model. However, substantial heterogeneity across studies and the performance degradation observed during independent external validation suggest that their generalizability to the real world must be confirmed through more rigorous study designs. Consequently, future research should prioritize establishing model robustness via prospective, multicenter external validation, coupled with efforts to standardize methodologies and improve reporting transparency. A critical step toward reliable clinical translation and achieving the ultimate goal of a "virtual biopsy" is developing algorithms that can translate pathology-grade diagnostic insights into preoperative, noninvasive imaging.

Funding

This study was funded by the General Project of the Liaoning Provincial Department of Education (grant number JYTMS20230109) and the Liaoning Provincial Science and Technology Program Joint Plan (Natural Science Foundation General Program) (grant number 2025-MSLH-742). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors contributed to the study conception and design. WF and BQ prepared the original draft of the manuscript and contributed to the methodology, formal analysis, and investigation. SH was responsible for conceptualization and funding acquisition and provided supervision. SH also reviewed and edited the manuscript. All authors commented on previous versions of the manuscript and also read and approved the final manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Flow diagram illustrating the search strategy.
[DOCX File, 22 KB - jmir_v28i1e82000_app1.docx]

Multimedia Appendix 2
(A) Meta-analysis forest plots: specificity and sensitivity of image-based deep learning (DL) in microvascular invasion (MVI) diagnosis in internal validation; (B) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of image-based DL in MVI diagnosis in internal validation; (C) Deeks funnel plot from meta-analysis of specificity and sensitivity of image-based DL in MVI diagnosis in internal validation; (D) Fagan nomogram from meta-analysis of specificity and sensitivity of image-based DL in MVI diagnosis in internal validation.
[PNG File, 606 KB - jmir_v28i1e82000_app2.png]

Multimedia Appendix 3
(A) Meta-analysis forest plots: specificity and sensitivity of image-based deep learning (DL) in microvascular invasion (MVI) diagnosis in external validation; (B) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of image-based DL in MVI diagnosis in external validation; (C) Deeks funnel plot from meta-analysis of specificity and sensitivity of image-based DL in MVI diagnosis in external validation; (D) Fagan nomogram from meta-analysis of specificity and sensitivity of image-based DL in MVI diagnosis in external validation.
[PNG File, 402 KB - jmir_v28i1e82000_app3.png]

Multimedia Appendix 4
(A) Meta-analysis forest plots: specificity and sensitivity of image-based deep learning (DL) in microvascular invasion (MVI) diagnosis in Validation set size of <100 cases; (B) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of image-based DL in MVI diagnosis in Validation set size < 100 cases; (C) Deeks funnel plot from meta-analysis of specificity and sensitivity of image-based DL in MVI diagnosis in the validation set size of <100 cases; (D) Fagan nomogram from meta-analysis of specificity and sensitivity of image-based DL in MVI diagnosis in the validation set size of <100 cases.
[PNG File, 519 KB - jmir_v28i1e82000_app4.png]

Multimedia Appendix 5
(A) Meta-analysis forest plots: specificity and sensitivity of image-based deep learning (DL) in microvascular invasion (MVI) diagnosis in Validation set size of ≥100 cases; (B) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of image-based DL in MVI diagnosis in Validation set size of ≥100 cases; (C) Deeks funnel plot from meta-analysis of specificity and sensitivity of image-based DL in MVI diagnosis in the validation set size ≥100 cases; (D) Fagan nomogram from meta-analysis of specificity and sensitivity of image-based DL in MVI diagnosis in Validation set size of ≥100 cases.
[PNG File, 528 KB - jmir_v28i1e82000_app5.png]

Multimedia Appendix 6
(A) Meta-analysis forest plots: specificity and sensitivity of contrast-enhanced computed tomography (CECT)–based deep learning (DL) in microvascular invasion (MVI) diagnosis; (B) Meta-analysis SROC: specificity and sensitivity of CECT-based DL in MVI diagnosis.
[PNG File, 320 KB - jmir_v28i1e82000_app6.png]

Multimedia Appendix 7
(A) Meta-analysis forest plots: specificity and sensitivity of contrast-enhanced computed tomography (CECT)–based deep learning (DL) in microvascular invasion (MVI) diagnosis in internal validation; (B) Meta-analysis SROC: specificity and sensitivity of CECT-based DL in MVI diagnosis in internal validation; (C) Meta-analysis forest plots: specificity and sensitivity of CECT-based DL in MVI diagnosis in external validation; (D) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of CECT-based DL in MVI diagnosis in external validation.
[PNG File, 394 KB - jmir_v28i1e82000_app7.png]

Multimedia Appendix 8
(A) Meta-analysis forest plots: specificity and sensitivity of contrast-enhanced ultrasound (CEUS)–based deep learning (DL) in microvascular invasion (MVI) diagnosis; (B) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of CEUS-based DL in MVI diagnosis.
[PNG File, 172 KB - jmir_v28i1e82000_app8.png]

Multimedia Appendix 9

XSL•FO
RenderX

(A) Meta-analysis forest plots: specificity and sensitivity of contrast-enhanced magnetic resonance imaging (CEMRI)–based deep learning (DL) in microvascular invasion (MVI) diagnosis; (B) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of CEMRI-based DL in MVI diagnosis.
[PNG File, 374 KB - jmir_v28i1e82000_app9.png]

Multimedia Appendix 10
(A) Meta-analysis forest plots: specificity and sensitivity of contrast-enhanced magnetic resonance imaging (CEMRI)–based deep learning (DL) in microvascular invasion (MVI) diagnosis in internal validation; (B) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of CEMRI-based DL in MVI diagnosis in internal validation; (C) Meta-analysis forest plots: specificity and sensitivity of CEMRI-based DL in MVI diagnosis in external validation; (D) Meta-analysis SROC: specificity and sensitivity of CEMRI-based DL in MVI diagnosis in external validation.
[PNG File, 424 KB - jmir_v28i1e82000_app10.png]

Multimedia Appendix 11
(A) Meta-analysis forest plots: specificity and sensitivity of magnetic resonance imaging (MRI)–based deep learning (DL) in microvascular imaging (MVI) diagnosis; (B) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of MRI-based DL in MVI diagnosis.
[PNG File, 171 KB - jmir_v28i1e82000_app11.png]

Multimedia Appendix 12
(A) Meta-analysis forest plots: specificity and sensitivity of multimodal imaging-based deep learning (DL) in microvascular invasion (MVI) diagnosis; (B) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of multimodal imaging-based DL in MVI diagnosis.
[PNG File, 222 KB - jmir_v28i1e82000_app12.png]

Multimedia Appendix 13
(A) Meta-analysis forest plots: specificity and sensitivity of multimodal imaging-based deep learning (DL) in microvascular invasion (MVI) diagnosis in internal validation; (B) Meta-analysis summary receiver operating characteristic (SROC): specificity and sensitivity of multimodal imaging-based DL in MVI diagnosis in internal validation; (C) Meta-analysis forest plots: specificity and sensitivity of multimodal imaging-based DL in MVI diagnosis in external validation; (D) Meta-analysis SROC: specificity and sensitivity of multimodal imaging-based DL in MVI diagnosis in external validation.
[PNG File, 313 KB - jmir_v28i1e82000_app13.png]

Multimedia Appendix 14
(A) Meta-analysis forest plots: specificity and sensitivity of pathological sections-based deep learning (DL) in microvascular invasion (MVI) diagnosis; (B) Meta-analysis summary receiver operating characteristic: specificity and sensitivity of pathological sections-based DL in MVI diagnosis.
[PNG File, 163 KB - jmir_v28i1e82000_app14.png]

Multimedia Appendix 15
Summary of bias risk for each study included in the paper according to the QUADAS-AI (artificial intelligence–specific quality assessment of diagnostic accuracy studies) domains.
[PDF File, 203 KB - jmir_v28i1e82000_app15.pdf]

Checklist 1
PRISMA-DTA checklist.
[DOCX File, 32 KB - jmir_v28i1e82000_app16.docx]

Checklist 2
PRISMA-S checklist.
[DOCX File, 19 KB - jmir_v28i1e82000_app17.docx]

References

1. Hwang SY, Danpanichkul P, Agopian V, et al. Hepatocellular carcinoma: updates on epidemiology, surveillance, diagnosis and treatment. Clin Mol Hepatol 2025 Feb;31(Suppl):S228-S254. [doi: 10.3350/cmh.2024.0824] [Medline: 39722614]
2. Brown ZJ, Tsilimigras DI, Ruff SM, et al. Management of hepatocellular carcinoma: a review. JAMA Surg 2023 Apr 1;158(4):410-420. [doi: 10.1001/jamasurg.2022.7989] [Medline: 36790767]

XSL•FO
RenderX

3. Vogel A, Meyer T, Sapisochin G, Salem R, Saborowski A. Hepatocellular carcinoma. Lancet 2022 Oct 15;400(10360):1345-1362. [doi: 10.1016/S0140-6736(22)01200-4] [Medline: 36084663]
4. Zhang ZH, Jiang C, Qiang ZY, et al. Role of microvascular invasion in early recurrence of hepatocellular carcinoma after liver resection: a literature review. Asian J Surg 2024 May;47(5):2138-2143. [doi: 10.1016/j.asjsur.2024.02.115] [Medline: 38443255]
5. Fuster-Anglada C, Mauro E, Ferrer-Fàbrega J, et al. Histological predictors of aggressive recurrence of hepatocellular carcinoma after liver resection. J Hepatol 2024 Dec;81(6):995-1004. [doi: 10.1016/j.jhep.2024.06.018] [Medline: 38925272]
6. Li SH, Mei J, Cheng Y, et al. Postoperative adjuvant hepatic arterial infusion chemotherapy with FOLFOX in hepatocellular carcinoma with microvascular invasion: a multicenter, phase III, randomized study. JCO 2023 Apr 1;41(10):1898-1908. [doi: 10.1200/JCO.22.01142]
7. Kang I, Jang M, Lee JG, et al. Subclassification of microscopic vascular invasion in hepatocellular carcinoma. Ann Surg 2021;274(6):e1170-e1178. [doi: 10.1097/SLA.0000000000003781]
8. Kim NR, Bae H, Hwang HS, et al. Preoperative prediction of microvascular invasion with gadoxetic acid-enhanced magnetic resonance imaging in patients with single hepatocellular carcinoma: the implication of surgical decision on the extent of liver resection. Liver Cancer 2024 Apr 24;13(2):181-192. [doi: 10.1159/000531786]
9. Ma L, Zhang C, Wen Y, et al. Imaging-based surrogate classification for risk stratification of hepatocellular carcinoma with microvascular invasion to adjuvant hepatic arterial infusion chemotherapy: a multicenter retrospective study. Int J Surg 2025 Jan 1;111(1):872-883. [doi: 10.1097/JS9.0000000000001903]
10. Lee S, Kang TW, Song KD, et al. Effect of microvascular invasion risk on early recurrence of hepatocellular carcinoma after surgery and radiofrequency ablation. Ann Surg 2021;273(3):564-571. [doi: 10.1097/SLA.0000000000003268]
11. Xu XF, Diao YK, Zeng YY, et al. Association of severity in the grading of microvascular invasion with long-term oncological prognosis after liver resection for early-stage hepatocellular carcinoma: a multicenter retrospective cohort study from a hepatitis B virus-endemic area. Int J Surg 2023 Apr 1;109(4):841-849. [doi: 10.1097/JS9.0000000000000325] [Medline: 36974673]
12. Wang L, Jia M, Wen X, Shen J, Yang H. Diagnostic value of magnetic resonance imaging features of microvascular invasion in hepatocellular carcinoma: a meta-analysis. Diagn Interv Radiol 2022 Sep;28(5):428-440. [doi: 10.5152/dir.2022.21108] [Medline: 36218148]
13. Xie XY, Chen R. Research progress of MRI-based radiomics in hepatocellular carcinoma. Front Oncol 2025;15:1420599. [doi: 10.3389/fonc.2025.1420599] [Medline: 39980543]
14. Zhu Z, Wu K, Lu J, et al. Gd-EOB-DTPA-enhanced MRI radiomics and deep learning models to predict microvascular invasion in hepatocellular carcinoma: a multicenter study. BMC Med Imaging 2025 Mar 31;25(1):40165094. [doi: 10.1186/s12880-025-01646-9]
15. Zhang L, Wang Y, Peng Z, et al. The progress of multimodal imaging combination and subregion based radiomics research of cancers. Int J Biol Sci 2022;18(8):3458-3469. [doi: 10.7150/ijbs.71046]
16. Zheng T, Zhu Y, Jiang H, et al. MRI-based topology deep learning model for noninvasive prediction of microvascular invasion and assisting prognostic stratification in HCC. Liver Int 2025 Mar;45(3):e16205. [doi: 10.1111/liv.16205] [Medline: 39992060]
17. Zhu Y, Wang P, Wang B, et al. Accelerated multi-b-value DWI using deep learning reconstruction: image quality improvement and microvascular invasion prediction in BCLC stage A hepatocellular carcinoma. Acad Radiol 2025 Jul;32(7):3924-3937. [doi: 10.1016/j.acra.2025.01.043]
18. McInnes MDF, Moher D, Thombs BD, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies. JAMA 2018 Jan 23;319(4):388. [doi: 10.1001/jama.2017.19163]
19. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc 2021 Apr 1;109(2):174-200. [doi: 10.5195/jmla.2021.962] [Medline: 34285662]
20. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005 Oct;58(10):982-990. [doi: 10.1016/j.jclinepi.2005.02.022]
21. Zhang X, Yu X, Liang W, et al. Deep learning-based accurate diagnosis and quantitative evaluation of microvascular invasion in hepatocellular carcinoma on whole-slide histopathology images. Cancer Med 2024 Mar;13(5):e7104. [doi: 10.1002/cam4.7104] [Medline: 38488408]
22. Lei Y, Feng B, Wan M, et al. Predicting microvascular invasion in hepatocellular carcinoma with a CT- and MRI-based multimodal deep learning model. Abdom Radiol 2024 May;49(5):1397-1410. [doi: 10.1007/s00261-024-04202-1]
23. Liu HF, Wang M, Lu YJ, et al. CEMRI-based quantification of intratumoral heterogeneity for predicting aggressive characteristics of hepatocellular carcinoma using habitat analysis: comparison and combination of deep learning. Acad Radiol 2024 Jun;31(6):2346-2355. [doi: 10.1016/j.acra.2023.11.024] [Medline: 38057182]
24. Zhou Z, Xia T, Zhang T, et al. Prediction of preoperative microvascular invasion by dynamic radiomic analysis based on contrast-enhanced computed tomography. Abdom Radiol 2024 Feb;49(2):611-624. [doi: 10.1007/s00261-023-04102-w]

XSL•FO
RenderX

25. He X, Xu Y, Zhou C, et al. Prediction of microvascular invasion and pathological differentiation of hepatocellular carcinoma based on a deep learning model. Eur J Radiol 2024 Mar;172:111348. [doi: 10.1016/j.ejrad.2024.111348] [Medline: 111348]
26. Zhong Y, Chen L, Ding F, Ou W, Zhang X, Weng S. Assessing microvascular invasion in HBV-related hepatocellular carcinoma: an online interactive nomogram integrating inflammatory markers, radiomics, and convolutional neural networks. Front Oncol 2024;14:1401095. [doi: 10.3389/fonc.2024.1401095] [Medline: 39351352]
27. Wang F, Zhan G, Chen QQ, et al. Multitask deep learning for prediction of microvascular invasion and recurrence-free survival in hepatocellular carcinoma based on MRI images. Liver Int 2024 Jun;44(6):1351-1362. [doi: 10.1111/liv.15870] [Medline: 38436551]
28. Yu Z, Liu Y, Dai X, Cui E, Cui J, Ma C. Enhancing preoperative diagnosis of microvascular invasion in hepatocellular carcinoma: domain-adaptation fusion of multi-phase CT images. Front Oncol 2024;14:1332188. [doi: 10.3389/fonc.2024.1332188] [Medline: 38333689]
29. Ma H, Wang L, Sun L, et al. Preoperative prediction of microvascular invasion in hepatocellular carcinoma from multi-sequence magnetic resonance imaging based on deep fusion representation learning. IEEE J Biomed Health Inform 2025 May;29(5):3259-3271. [doi: 10.1109/JBHI.2024.3451331] [Medline: 39196745]
30. Zhang W, Guo Q, Zhu Y, et al. Cross-institutional evaluation of deep learning and radiomics models in predicting microvascular invasion in hepatocellular carcinoma: validity, robustness, and ultrasound modality efficacy comparison. Cancer Imaging 2024 Oct 22;24(1):142. [doi: 10.1186/s40644-024-00790-9] [Medline: 39438929]
31. Wang F, Chen Q, Chen Y, et al. A novel multimodal deep learning model for preoperative prediction of microvascular invasion and outcome in hepatocellular carcinoma. Eur J Surg Oncol 2023 Jan;49(1):156-164. [doi: 10.1016/j.ejso.2022.08.036] [Medline: 36333180]
32. You H, Wang J, Ma R, et al. Clinical interpretability of deep learning for predicting microvascular invasion in hepatocellular carcinoma by using attention mechanism. Bioengineering (Basel) 2023 Aug 9;10(8):948. [doi: 10.3390/bioengineering10080948] [Medline: 37627833]
33. Qin X, Zhu J, Tu Z, Ma Q, Tang J, Zhang C. Contrast-enhanced ultrasound with deep learning with attention mechanisms for predicting microvascular invasion in single hepatocellular carcinoma. Acad Radiol 2023 Sep;30 Suppl 1:S73-S80. [doi: 10.1016/j.acra.2022.12.005] [Medline: 36567144]
34. Li Z, Wang Y, Zhu Y, et al. Modality-based attention and dual-stream multiple instance convolutional neural network for predicting microvascular invasion of hepatocellular carcinoma. Front Oncol 2023;13:1195110. [doi: 10.3389/fonc.2023.1195110] [Medline: 37434971]
35. Cao L, Wang Q, Hong J, et al. MVI-TR: a transformer-based deep learning model with contrast-enhanced CT for preoperative prediction of microvascular invasion in hepatocellular carcinoma. Cancers (Basel) 2023 Feb 28;15(5):1538. [doi: 10.3390/cancers15051538] [Medline: 36900327]
36. Wang T, Li Z, Yu H, et al. Prediction of microvascular invasion in hepatocellular carcinoma based on preoperative Gd-EOB-DTPA-enhanced MRI: comparison of predictive performance among 2D, 2D-expansion and 3D deep learning models. Front Oncol 2023;13:987781. [doi: 10.3389/fonc.2023.987781] [Medline: 36816963]
37. Ye Z, Zhang J, Wu N, et al. PET-guided attention for prediction of microvascular invasion in preoperative hepatocellular carcinoma on PET/CT. Ann Nucl Med 2023 Apr;37(4):238-245. [doi: 10.1007/s12149-023-01822-3] [Medline: 36723705]
38. Xu Y, Chen Y, Wu J, Pan J, Liao C, Su H. The utility of vision transformer in preoperatively predicting microvascular invasion status of hepatocellular carcinoma. HPB (Oxford) 2023 May;25(5):533-542. [doi: 10.1016/j.hpb.2023.01.015]
39. Deng Y, Jia X, Yu G, et al. Can a proposed double branch multimodality-contribution-aware TripNet improve the prediction performance of the microvascular invasion of hepatocellular carcinoma based on small samples? Front Oncol 2022;12:1035775. [doi: 10.3389/fonc.2022.1035775] [Medline: 36387069]
40. Li X, Qi Z, Du H, et al. Deep convolutional neural network for preoperative prediction of microvascular invasion and clinical outcomes in patients with HCCs. Eur Radiol 2022 Feb;32(2):771-782. [doi: 10.1007/s00330-021-08198-w]
41. Chen Q, Xiao H, Gu Y, et al. Deep learning for evaluation of microvascular invasion in hepatocellular carcinoma from tumor areas of histology images. Hepatol Int 2022 Jun;16(3):590-602. [doi: 10.1007/s12072-022-10323-w]
42. Zhang Y, Wei Q, Huang Y, et al. Deep learning of liver contrast-enhanced ultrasound to predict microvascular invasion and prognosis in hepatocellular carcinoma. Front Oncol 2022;12:878061. [doi: 10.3389/fonc.2022.878061] [Medline: 35875110]
43. Liu B, Zeng Q, Huang J, et al. IVIM using convolutional neural networks predicts microvascular invasion in HCC. Eur Radiol 2022 Oct;32(10):7185-7195. [doi: 10.1007/s00330-022-08927-9]
44. Sun SW, Xu X, Liu QP, et al. LiSNet: an artificial intelligence - based tool for liver imaging staging of hepatocellular carcinoma aggressiveness. Med Phys Mex Symp Med Phys 2022 Nov;49(11):6903-6913. [doi: 10.1002/mp.15972]
45. Wang L, Wu M, Li R, Xu X, Zhu C, Feng X. MVI-Mind: a novel deep-learning strategy using computed tomography (CT)-based radiomics for end-to-end high efficiency prediction of microvascular invasion in hepatocellular carcinoma. Cancers (Basel) 2022;14(12):2956. [doi: 10.3390/cancers14122956]
46. Xiao H, Guo Y, Zhou Q, et al. Prediction of microvascular invasion in hepatocellular carcinoma with expert-inspiration and skeleton sharing deep learning. Liver Int 2022 Jun;42(6):1423-1431. [doi: 10.1111/liv.15254] [Medline: 35319151]

XSL•FO
RenderX

47. Yang Y, Zhou Y, Zhou C, Ma X. Deep learning radiomics based on contrast enhanced computed tomography predicts microvascular invasion and survival outcome in early stage hepatocellular carcinoma. Eur J Surg Oncol 2022 May;48(5):1068-1077. [doi: 10.1016/j.ejso.2021.11.120]
48. Sun BY, Gu PY, Guan RY, et al. Deep-learning-based analysis of preoperative MRI predicts microvascular invasion and outcome in hepatocellular carcinoma. World J Surg Oncol 2022 Jun 8;20(1):189. [doi: 10.1186/s12957-022-02645-8] [Medline: 35676669]
49. Dai X, Cao S, Guo Y, et al. Preoperative identification of microvascular invasion in hepatocellular carcinoma based on multi-modal and multi response convolutional neural network. Presented at: Fourteenth International Conference on Graphics and Image Processing (ICGIP 2022); Oct 21-23, 2022. [doi: 10.1117/12.2680693]
50. Zhang Y, Lv X, Qiu J, et al. Deep learning with 3D convolutional neural network for noninvasive prediction of microvascular invasion in hepatocellular carcinoma. Magn Reson Imaging 2021 Jul;54(1):134-143. [doi: 10.1002/jmri.27538]
51. Liu SC, Lai J, Huang JY, et al. Predicting microvascular invasion in hepatocellular carcinoma: a deep learning model validated across hospitals. Cancer Imaging 2021 Oct 9;21(1):56. [doi: 10.1186/s40644-021-00425-3] [Medline: 34627393]
52. Wei J, Jiang H, Zeng M, et al. Prediction of microvascular invasion in hepatocellular carcinoma via deep learning: a multi-center and prospective validation study. Cancers (Basel) 2021 May 14;13(10):2368. [doi: 10.3390/cancers13102368]
53. Zhou W, Jian W, Cen X, et al. Prediction of microvascular invasion of hepatocellular carcinoma based on contrast-enhanced MR and 3D convolutional neural networks. Front Oncol 2021;11:588010. [doi: 10.3389/fonc.2021.588010] [Medline: 33854959]
54. Wang G, Jian W, Cen X, et al. Prediction of microvascular invasion of hepatocellular carcinoma based on preoperative diffusion-weighted MR using deep learning. Acad Radiol 2021 Nov;28:S118-S127. [doi: 10.1016/j.acra.2020.11.014]
55. Zeng Q, Liu B, Xu Y, Zhou W. An attention-based deep learning model for predicting microvascular invasion of hepatocellular carcinoma using an intra-voxel incoherent motion model of diffusion-weighted magnetic resonance imaging. Phys Med Biol 2021 Sep 21;66(18):185019. [doi: 10.1088/1361-6560/ac22db]
56. Gao F, Qiao K, Yan B, et al. Hybrid network with difference degree and attention mechanism combined with radiomics (H-DARnet) for MVI prediction in HCC. Magn Reson Imaging 2021 Nov;83:27-40. [doi: 10.1016/j.mri.2021.06.018] [Medline: 34147593]
57. Jiang YQ, Cao SE, Cao S, et al. Preoperative identification of microvascular invasion in hepatocellular carcinoma by XGBoost and deep learning. J Cancer Res Clin Oncol 2021 Mar;147(3):821-833. [doi: 10.1007/s00432-020-03366-9] [Medline: 32852634]
58. Song D, Wang Y, Wang W, et al. Using deep learning to predict microvascular invasion in hepatocellular carcinoma based on dynamic contrast-enhanced MRI combined with clinical parameters. J Cancer Res Clin Oncol 2021 Dec;147(12):3757-3767. [doi: 10.1007/s00432-021-03617-3] [Medline: 33839938]
59. Men S, Ju H, Zhang L, Zhou W. Prediction of microvascular invasion of hepatocellar carcinoma with contrast-enhanced MR using 3D CNN and LSTM. Presented at: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI); Oct 8-11, 2019. [doi: 10.1109/ISBI.2019.8759582]
60. Zhou Y, Sun SW, Liu QP, Xu X, Zhang Y, Zhang YD. TED: two-stage expert-guided interpretable diagnosis framework for microvascular invasion in hepatocellular carcinoma. Med Image Anal 2022 Nov;82:102575. [doi: 10.1016/j.media.2022.102575] [Medline: 36063747]
61. Chu T, Zhao C, Zhang J, et al. Application of a convolutional neural network for multitask learning to simultaneously predict microvascular invasion and vessels that encapsulate tumor clusters in hepatocellular carcinoma. Ann Surg Oncol 2022 Oct;29(11):6774-6783. [doi: 10.1245/s10434-022-12000-6] [Medline: 35754067]
62. Huang H, Liu B, Zhang L, Xu Y, Zhou W, editors. Transformer Based Multi-Task Deep Learning with Intravoxel Incoherent Motion Model Fitting for Microvascular Invasion Prediction of Hepatocellular Carcinoma: Springer Nature Switzerland; 2022. [doi: 10.1007/978-3-031-16449-1_26]
63. Wang Y, Xie W, Li C, et al. Automated microvascular invasion prediction of hepatocellular carcinoma via deep relation reasoning from dynamic contrast-enhanced ultrasound. Comput Med Imaging Graph 2025 Sep;124:102606. [doi: 10.1016/j.compmedimag.2025.102606] [Medline: 40680384]
64. Cen YY, Nong HY, Huang XX, et al. Computed tomography-based deep learning and multi-instance learning for predicting microvascular invasion and prognosis in hepatocellular carcinoma. World J Gastroenterol 2025 Aug 14;31(30):109186. [doi: 10.3748/wjg.v31.i30.109186] [Medline: 40933208]
65. Huang Z, Huang W, Jiang L, et al. Decision fusion model for predicting microvascular invasion in hepatocellular carcinoma based on multi-MR habitat imaging and machine-learning classifiers. Acad Radiol 2025 Apr;32(4):1971-1980. [doi: 10.1016/j.acra.2024.10.007] [Medline: 39472207]
66. Miao S, Sun M, Li X, et al. Deep learning-based prediction of microvascular invasion and survival outcomes in hepatocellular carcinoma using dual-phase CT imaging of tumors and lesser omental adipose: a multicenter study. Acad Radiol 2025 Oct;32(10):5789-5801. [doi: 10.1016/j.acra.2025.07.015]
67. Dong X, Jia X, Zhang W, et al. Interpretable and generalizable deep learning model for preoperative assessment of microvascular invasion and outcome in hepatocellular carcinoma based on MRI: a multicenter study. Insights Imaging 2025 Jul 3;16(1):151. [doi: 10.1186/s13244-025-02035-0] [Medline: 40610844]

XSL•FO
RenderX

68. Zhang Y, Liu H, Zhu L, et al. Modality-aware distillation network for microvascular invasion prediction of hepatocellular carcinoma from MRI images. IEEE Trans Biomed Eng 2025 Jun;72(6):1825-1836. [doi: 10.1109/TBME.2024.3523921]
69. Zhao Y, Huang X, Sun M, et al. Predicting microvascular invasion plus cytokeratin 19 expression positivity in hepatocellular carcinoma based on EOB-MRI using multitask deep learning. Hepatoma Res 2025;10:12. [doi: 10.20517/2394-5079.2024.143]
70. Qin Q, Pang J, Li J, et al. Transformer model based on Sonazoid contrast-enhanced ultrasound for microvascular invasion prediction in hepatocellular carcinoma. Med Phys 2025 Jul;52(7):e17895. [doi: 10.1002/mp.17895] [Medline: 40384312]
71. Xiao Q, Zhu W, Tang H, Zhou L. Ultrasound radiomics in the prediction of microvascular invasion in hepatocellular carcinoma: a systematic review and meta-analysis. Heliyon 2023 Jun;9(6):e16997. [doi: 10.1016/j.heliyon.2023.e16997]
72. Liang G, Yu W, Liu S, et al. The diagnostic performance of radiomics-based MRI in predicting microvascular invasion in hepatocellular carcinoma: a meta-analysis. Front Oncol 2022;12:960944. [doi: 10.3389/fonc.2022.960944] [Medline: 36798691]
73. Li L, Wu C, Huang Y, Chen J, Ye D, Su Z. Radiomics for the preoperative evaluation of microvascular invasion in hepatocellular carcinoma: a meta-analysis. Front Oncol 2022;12:831996. [doi: 10.3389/fonc.2022.831996] [Medline: 35463303]
74. He K, Liu X, Shahzad R, et al. Advanced deep learning approach to automatically segment malignant tumors and ablation zone in the liver with contrast-enhanced CT. Front Oncol 2021;11:669437. [doi: 10.3389/fonc.2021.669437] [Medline: 34336661]
75. Kinoshita M, Ueda D, Matsumoto T, et al. Deep learning model based on contrast-enhanced computed tomography imaging to predict postoperative early recurrence after the curative resection of a solitary hepatocellular carcinoma. Cancers (Basel) 2023 Apr 4;15(7):2140. [doi: 10.3390/cancers15072140] [Medline: 37046801]
76. Lu D, Wang LF, Han H, et al. Prediction of microvascular invasion in hepatocellular carcinoma with conventional ultrasound, Sonazoid-enhanced ultrasound, and biochemical indicator: a multicenter study. Insights Imaging 2024 Oct 28;15(1):261. [doi: 10.1186/s13244-024-01743-3] [Medline: 39466459]
77. Yan M, Zhang X, Zhang B, et al. Deep learning nomogram based on Gd-EOB-DTPA MRI for predicting early recurrence in hepatocellular carcinoma after hepatectomy. Eur Radiol 2023 Jul;33(7):4949-4961. [doi: 10.1007/s00330-023-09419-0] [Medline: 36786905]
78. Wu Y, Zhu M, Liu Y, Cao X, Zhang G, Yin L. Peritumoral imaging manifestations on Gd-EOB-DTPA-enhanced MRI for preoperative prediction of microvascular invasion in hepatocellular carcinoma: a systematic review and meta-analysis. Front Oncol 2022;12:907076. [doi: 10.3389/fonc.2022.907076] [Medline: 35814461]
79. Li C, Tan Y, Chen W, et al. Attention unet++: a nested attention-aware u-net for liver CT image segmentation. 2020 Presented at: 2020 IEEE International Conference on Image Processing (ICIP); Oct 25-28, 2020; Abu Dhabi, United Arab Emirates p. 345-349. [doi: 10.1109/ICIP40778.2020.9190761]
80. Huang H, Lin L, Tong R, et al. UNet 3+: a full-scale connected unet for medical image segmentation. Presented at: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 4-8, 2020; Barcelona, Spain p. 1055-1059. [doi: 10.1109/ICASSP40776.2020.9053405]
81. Bilic P, Christ P, Li HB, et al. The liver tumor segmentation benchmark (LiTS). Med Image Anal 2023 Feb;84:102680. [doi: 10.1016/j.media.2022.102680] [Medline: 36481607]
82. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology 2016 Feb;278(2):563-577. [doi: 10.1148/radiol.2015151169]
83. Riley RD, Archer L, Snell KIE, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ 2024 Jan 15;384:e074820. [doi: 10.1136/bmj-2023-074820]
84. Collins GS, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ 2024 Jan 8;384:e074819. [doi: 10.1136/bmj-2023-074819]
85. Strandberg R, Jepsen P, Hagström H. Developing and validating clinical prediction models in hepatology - an overview for clinicians. J Hepatol 2024 Jul;81(1):149-162. [doi: 10.1016/j.jhep.2024.03.030] [Medline: 38531493]
86. Maung ST, Tangkijvanich P, Chaiteerakij R. Magnetic resonance imaging radiomics for predicting hepatocellular carcinoma recurrence following resection or ablation: a systematic review and meta-analysis. Abdom Radiol (NY) 2025 Sep 22. [doi: 10.1007/s00261-025-05184-4] [Medline: 40981987]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
CECT: contrast-enhanced computed tomography
CEMRI: contrast-enhanced magnetic resonance imaging
CEUS: contrast-enhanced ultrasound
DL: deep learning
DOR: diagnostic odds ratio
HCC: hepatocellular carcinoma

XSL•FO
RenderX

LFK: Luis Furuya-Kanamori
LR: likelihood ratio
MeSH: Medical Subject Headings
MRI: magnetic resonance imaging
MVI: microvascular invasion
PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Diagnostic Test Accuracy
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for literature searches
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2
QUADAS-AI: Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence
ROB: risk of bias
SROC: summary receiver operating characteristic

Edited by S Brini; submitted 07.Aug.2025; peer-reviewed by L Zhou, ST Maung; accepted 08.Jan.2026; published 02.Mar.2026.

Please cite as:
Feng W, Qu B, Han S
Accuracy of Medical Image–Based Deep Learning for Detecting Microvascular Invasion in Hepatocellular Carcinoma: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e82000
URL: https://www.jmir.org/2026/1/e82000
doi:10.2196/82000

© Wei Feng, Bo Qu, Shuo Han. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Predictive Value of Machine Learning for Postoperative Delirium in Cardiac Surgery: Systematic Review and Meta-Analysis

Yi Guo[1], MS; Hong Xu[2], BS; Ankui Wang[1], BS; Mingming Zhang[1], MS; Shuai Zhang[1], BS; Peng Xie[3], MS

[1]Department of Anesthesiology, Norinco General Hospital, Xi'an, China

[2]Department of Operating Room, Norinco General Hospital, Xi'an, China

[3]Department of Hepatobiliary and Vascular Surgery, Norinco General Hospital, No. 12, Zhangba East Road, Yanta District, Xi'an, China

Corresponding Author:
Peng Xie, MS
Department of Hepatobiliary and Vascular Surgery, Norinco General Hospital, No. 12, Zhangba East Road, Yanta District, Xi'an, China

Abstract

Background: Postoperative delirium (POD) following cardiac surgery is a severe complication, and early identification of delirium risk remains a challenge in clinical practice. While machine learning (ML) has garnered increasing attention in health care applications, effective early prediction tools remain limited in current clinical practice. Recent investigations have explored the effectiveness of ML-based methods for identifying the risk of POD in patients undergoing cardiac surgery. However, more evidence is required to validate the feasibility of these methods.

Objectives: This study aims to ascertain the performance of ML in identifying the risk of POD following cardiac surgery, providing evidence for the development or updating of future ML-based assessment tools.

Methods: A comprehensive literature search was conducted across 4 databases—PubMed, the Cochrane Library, Embase, and Web of Science—through August 30, 2024, to identify studies investigating individual POD risk prediction using ML approaches and nomograms. The risk of bias of the models in the included studies was assessed leveraging the Prediction Model Bias Risk Assessment Tool. Subgroup analyses were performed based on datasets, validation methods, study types, risk of bias, and model types.

Results: The analysis incorporated 28 original studies comprising 80,143 patients undergoing cardiac surgery, of whom 6326 developed POD. Meta-analysis revealed that, in validation datasets, the c-index, sensitivity, and specificity for delirium prediction reached 0.805 (95% CI 0.759 - 0.852), 0.72 (95% CI 0.65 - 0.79), and 0.78 (95% CI 0.71 - 0.83), respectively. Logistic regression was the primary modeling method. In validation datasets, the c-index, sensitivity, and specificity reached 0.773 (95% CI 0.724 - 0.823), 0.73 (95% CI 0.64 - 0.80), and 0.70 (95% CI 0.65 - 0.74), respectively.

Conclusions: ML-based prediction tools for POD following cardiac surgery demonstrate promising performance. However, the limited number of studies and validation approaches necessitate cautious interpretation of these findings. Future multicenter studies are warranted to develop more robust ML-based prediction tools, enabling precise risk stratification and targeted preventive interventions for POD.

Trial Registration: PROSPERO CRD42024588522; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024588522

(J Med Internet Res 2026;28:e72304) doi:10.2196/72304

KEYWORDS

cardiac surgery; delirium; machine learning; predictive model; systematic review

Introduction

The number of deaths from cardiovascular disease far exceeds those from cancer [1], and cardiac surgery is often required for its treatment. Common types of cardiac surgery include congenital heart disease surgery, coronary artery bypass grafting, cardiac valve surgery, and ascending aorta and arch surgery. Although cardiac surgery has benefitted a large number of patients, postoperative complications remain a significant clinical challenge. Common complications include excessive bleeding, refractory shock, postoperative cardiac arrest, neurologic injury, respiratory failure, and acute respiratory distress syndrome [2], as well as gastrointestinal complications [3]. Delirium, characterized by acute changes in mental status with confusion and impaired attention, is one of the most common and serious neurological complications following cardiac surgery, with reported incidence rates exceeding 50% [4,5]. A study by Salluh et al [6] found that delirium correlates with higher in-hospital mortality rates, extended hospital stays, and postdischarge cognitive impairment. Therefore, early

XSL•FO
RenderX

prediction of delirium risk and the development of effective, targeted preventive strategies are of considerable clinical importance.

Currently, there is no universally recognized or effective tool for predicting postoperative delirium (POD) in cardiac surgery. Machine learning (ML) can integrate high-dimensional data to develop and validate clinical prediction tools for disease prediction, diagnosis, treatment response, prognosis, and adverse events. Several researchers have constructed predictive models for POD using ML techniques [7-9]; however, the accuracy of these models requires further validation. In the Shining Cai study [7], prediction models for postcardiac surgery delirium were summarized, but the number of included studies was limited, possibly due to the search strategy. Both Lee et al [8] and Chen et al [9] focused on POD prediction models and demonstrated promising predictive value. However, the study by Lee et al [8] specifically focused on delirium occurring in the intensive care unit after cardiac surgery, whereas Chen et al [9] examined POD without specifying cardiac surgery. Therefore, this study aims to evaluate the effectiveness of ML-based predictive models for delirium after cardiac surgery and to inform the future development or refinement of simplified clinical scoring tools.

Methods

Study Registration

This study adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines and was registered with PROSPERO (registration number CRD42024588522).

Inclusion and Exclusion Criteria

Eligibility criteria were defined before literature screening (Table 1).

Table . Eligibility criteria for original studies included in this systematic review.

Items	Inclusion criteria	Exclusion criteria
P (Population)	Patients who underwent cardiac surgery, including CABG[a], OPCABG[b], valve surgery, ascending aorta and arch surgery, and minimally invasive procedures such as TAVR[c]	Studies that did not clearly distinguish cardiac surgery from other surgical procedures
I (Intervention)	Studies that developed a complete predictive model for POD[d] after cardiac surgery	Studies limited to risk factor identification without a complete ML[e] model, or studies that assessed only univariate predictive accuracy
C (Control)	None	None
O (Outcomes)	Evaluation of ML model performance using metrics such as c-index, sensitivity (Sen), specificity (Spe), accuracy, precision, diagnostic contingency tables, or F_1-score	Studies lacking outcome measures for model accuracy
S (Study design)	Case-control, cohort, or cross-sectional studies published in English	Studies evaluating only established scales (for studies using overlapping datasets, only the study with the largest sample size was retained)

[a]CABG: coronary artery bypass graft.

[b]OPCABG: off-pump coronary artery bypass grafting.

[c]TAVR: transcatheter aortic valve replacement.

[d]POD: postoperative delirium.

[e]ML: machine learning.

Data Sources and Search Strategy

A comprehensive literature search was executed in 4 databases: PubMed, the Cochrane Library, Embase, and Web of Science, through August 30, 2024. Both Medical Subject Headings and free-text terms were used for the search. No restrictions were imposed on regions and year of publication (Table S1 in Multimedia Appendix 1).

Study Selection and Data Extraction

The retrieved records were imported into EndNote, and after removing duplicates, titles and abstracts were reviewed. Full texts of potentially relevant articles were obtained and checked. To reduce population heterogeneity, this review focused exclusively on cardiac surgery. Because some relevant studies did not explicitly mention cardiac surgery in the title or abstract, the term “cardiac surgery” was not included as a search keyword; instead, noncardiac surgical studies were excluded manually during the screening process.

A standardized digital form was leveraged for data collection. Extracted information included title, DOI, authors, year of publication, study design, patient source, type of cardiac surgery, diagnostic criteria for delirium, number of patients with delirium, total number of cardiac surgery patients, number of delirium patients in both training and validation sets, validation set formation method, total number of patients in the validation set, missing data and methods for handling missing data, variable screening methods, modeling methods, and predictive factors.

XSL•FO
RenderX

Two researchers independently reviewed the literature and cross-checked the extracted data, with a third investigator resolving any disagreements.

Risk of Bias in Included Studies

Risk of bias in the included studies was appraised using the Prediction Model Risk of Bias Assessment Tool (PROBAST) [10], which included questions across domains including participants, predictors, outcomes, and statistical analysis. These domains reflected the overall risk of bias and applicability of the studies. Specific questions within domains had 3 possible responses: yes/likely, no/likely not, and no information. If the risk of bias in all domains was low for a study, the study was marked as low risk. If the risk of bias in at least a domain was high, the study was graded to have a high risk of bias. Two researchers independently performed and cross-checked the risk of bias assessments, with a third researcher resolving any disagreements.

Statistical Analysis

A meta-analysis was performed to evaluate the overall performance of ML models using the c-index. For studies lacking CIs or SEs of c-index, SEs were estimated according to the study by Debray et al [11]. The heterogeneity between studies was quantified by leveraging the I^2 statistic. If I^2 was greater than 50%, a random-effects model was utilized. If I^2 was less than 50%, a fixed-effects model was applied. Funnel plots were employed to detect publication bias of the c-index. The Egger test was applied for statistical assessment of publication bias.

A bivariate mixed-effects model was used to synthesize sensitivity and specificity data. For studies without diagnostic contingency tables, these were reconstructed using sensitivity, specificity, precision, and case numbers for calculations. Subgroup analyses were performed based on validation methods, study types, risk of bias, and model types. The statistical computation was conducted using Stata version 15.0.

Results

Study Selection

The initial search yielded 5868 articles. After the removal of 1981 duplicates, 3760 studies were excluded based on titles and abstracts because they were unrelated to the research theme or inconsistent with the original study design. A total of 127 studies underwent full-text review. Of these, 10 unpublished conference abstracts, 25 studies limited to risk factor analysis without construction of a complete prediction model, 62 studies that failed to clearly distinguish cardiac surgery from noncardiac surgery, and 2 studies with overlapping datasets were excluded. Eventually, 28 studies [12-39] were included in the final analysis (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the study selection process.

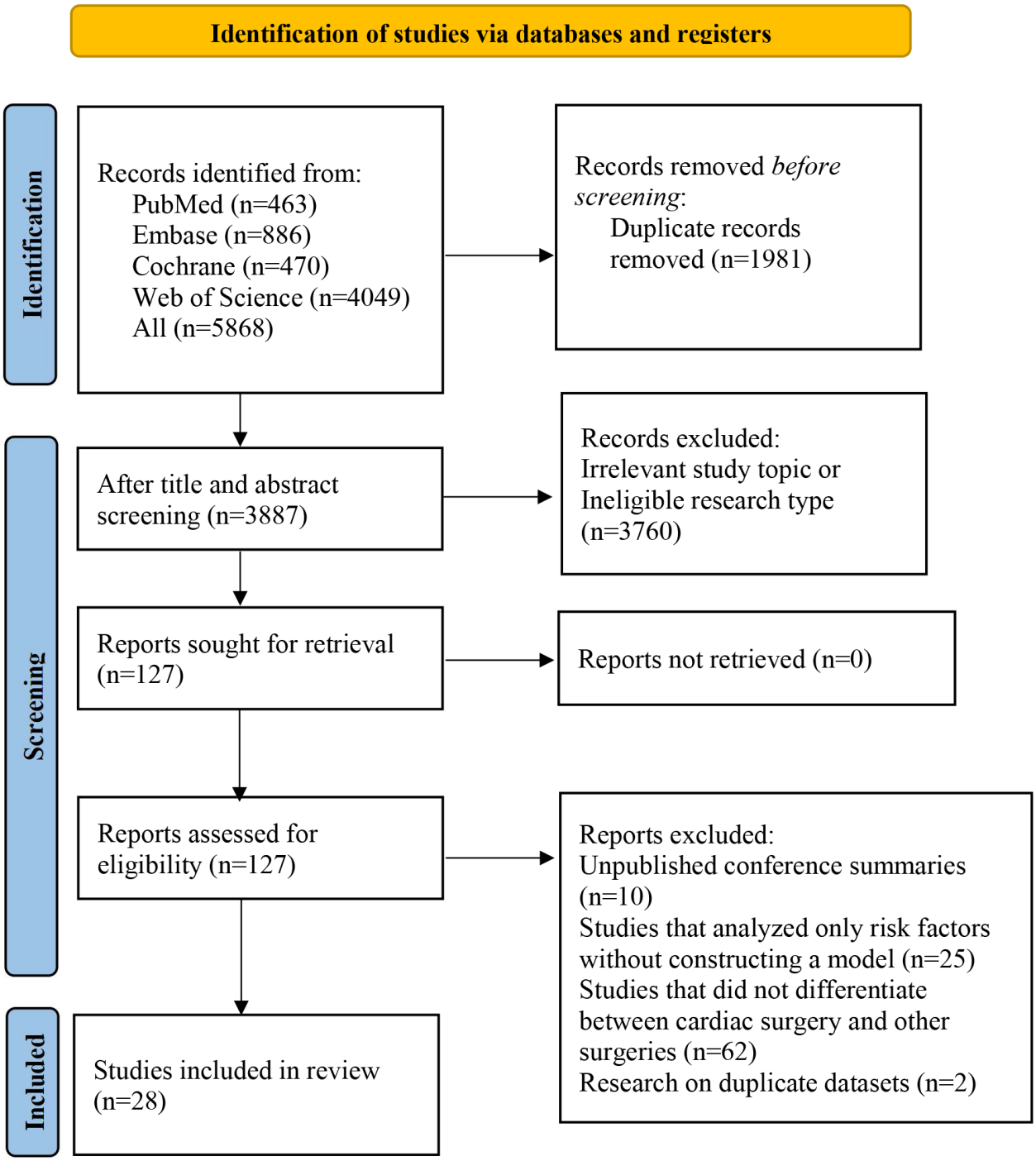

Study Characteristics

The analysis included 28 publications from 2012 to 2024. The studies collectively involved 80,143 patients undergoing cardiac surgery, of whom 6326 developed delirium. These studies included 12 retrospective cohort studies, 14 prospective cohort studies, and 2 retrospective (for model development) and prospective (for model validation) cohort studies. These studies spanned 12 countries, with 13 studies originating from China. Five studies were multicenter, one was based on a registry database, and 23 studies were single-center.

The studies included coronary artery bypass grafting, valve surgery, congenital heart surgery (patients aged ≤8 y), cardiac tumor resection, aortic arch repairs, and type A aortic dissection treated with open-heart surgery. Most studies described multiple types of cardiac surgery, whereas only a few studies focused on a single surgery type. All 28 studies clearly outlined the diagnostic methods or steps for identifying delirium. To minimize population heterogeneity, the 2 pediatric studies and the remaining 26 adult studies were analyzed separately. Among the 26 adult studies, 21 included a validation set and described its generation method, comprising 1 external validation, 4 temporal validations, 11 random split validations, 3 k-fold

XSL•FO
RenderX

cross-validations, and 2 bootstrap resampling validations. One study performed temporal validation but did not report validation set results due to inadequate sample size. To avoid sample size inflation, only the best-performing model from each study was selected, resulting in 9 model types, with logistic regression models being the most prevalent (Table S2 in Multimedia Appendix 2) [12-39].

In addition, the variables used for modeling were primarily clinical features. The modeling variables are presented in Table S3 in Multimedia Appendix 3.

Risk of Bias Assessment

Risk of bias was assessed for the 28 models from the 28 studies. In the participant domain, some studies had only validated a single model, which contributed to potential bias. Twelve models were from retrospective cohort studies, resulting in a high risk of bias in participant selection. Given that all studies constructed models based on clinical features, the retrospective cohort models may introduce bias in assessing clinical characteristics, thereby increasing the risk of bias in the prediction factors as well. All included studies used established diagnostic criteria for delirium, and no prediction factors that required delirium assessment were incorporated into the modeling process, indicating low bias risk in outcome assessment.

In the predictors domain, only one study was rated as high risk of bias due to the incorporation of outcome information during predictor assessment, whereas all other studies demonstrated low risk of bias.

In the outcomes domain, none of the studies explicitly reported blinding; consequently, all models were rated as having unclear risk of bias for this domain.

Fifteen models had a high risk of bias from insufficient events per variable (EPV <20) or lack of independent validation. Twenty-three models showed a high risk of bias in missing data handling, including direct deletion of missing values. Two studies showed unclear risk of bias due to inadequate description of predictor selection methods. Seventeen models demonstrated a high risk of bias for inadequate assessment of model fit or overfitting [40]. Overall, 26 models were identified as high-risk and 2 low-risk prediction models in statistical analysis (Figure 2).

Figure 2. Risk of bias of the included predictive models assessed with the Prediction Model Risk of Bias Assessment Tool.

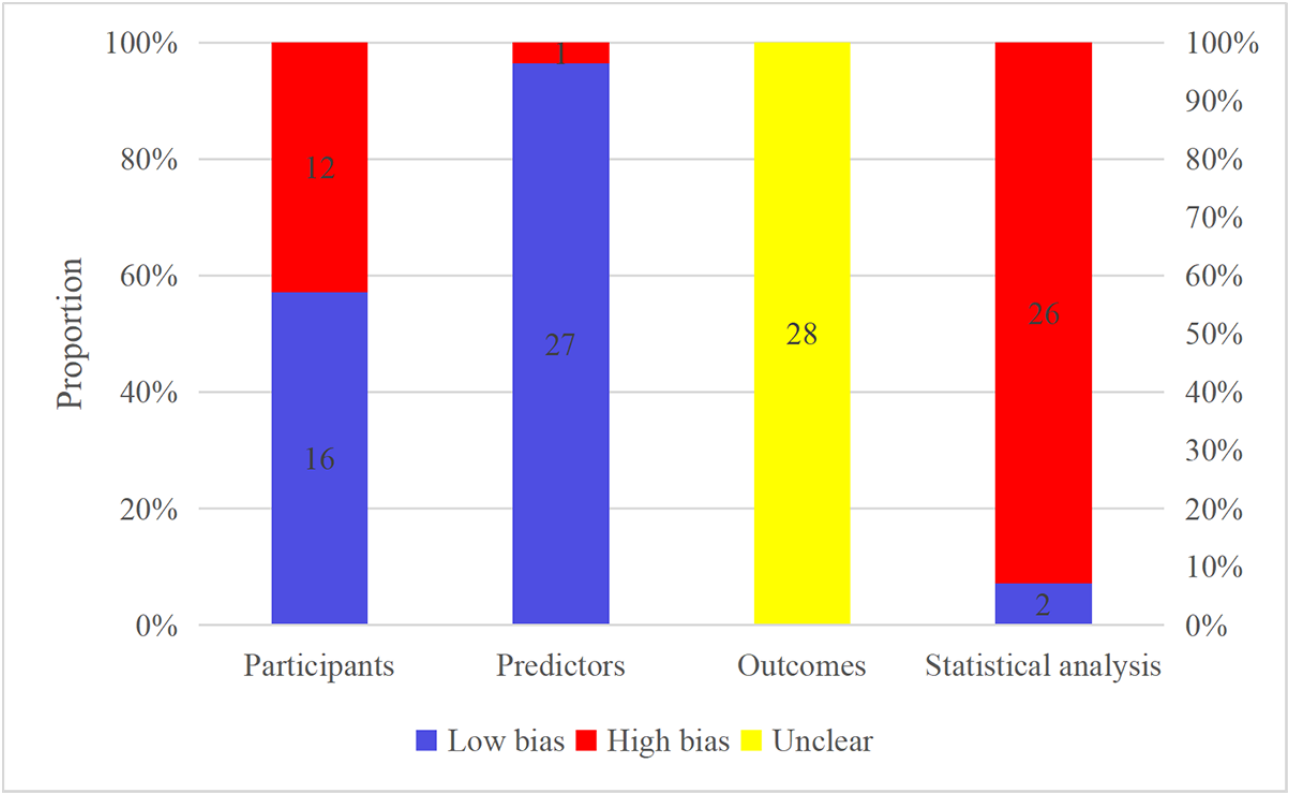

Meta-Analysis

C-Index

Because training set data were unavailable from 9 studies, only 17 prediction models for POD following cardiac surgery were reported in the training set. A random-effects model was leveraged, and the combined c-index was 0.815 (95% CI 0.754 - 0.876). For validation, 20 models were reported for validating the predictive accuracy of POD following cardiac surgery. The pooled c-index, also derived from a random-effects model, was 0.805 (95% CI 0.759 - 0.852) (Table 2).

XSL•FO
RenderX

Table . Meta-analysis results of c-index in the included studies for machine learning prediction of postoperative delirium following cardiac surgery in adults.

Subgroups	Training set			Validation set		
	Studies, n	c-index (95% CI)	I^2 (%)	Studies, n	c-index (95% CI)	I^2 (%)
Validation methods						
Random split validation	7	0.769 (0.692 - 0.847)	97.6	11	0.791 (0.722 - 0.860)	98.4
K-fold cross-validation	__[a]	—	—	3	0.845 (0.769 - 0.921)	73.9
Bootstrap resampling validation	—	—	—	2	0.805 (0.745 - 0.866)	47.1
External validation	1	0.740 (0.665 - 0.815)	—	1	0.750 (0.667 - 0.833)	—
Temporal validation	4	0.932 (0.895 - 0.969)	94.2	3	0.848 (0.807 - 0.889)	16.8
No validation	5	0.801 (0.744 - 0.857)	74.8	—	—	—
Study types						
Retrospective cohort study	6	0.819 (0.696 - 0.943)	99.5	10	0.803 (0.734 - 0.871)	98.4
Prospective cohort study	9	0.809 (0.758 - 0.860)	88.7	8	0.817 (0.762 - 0.872)	82.6
Retrospective and prospective cohort study	2	0.832 (0.599 - 1.065)	98.7	2	0.789 (0.621 - 0.958)	93.8
Risk of bias (statistic analysis)						
Low	1	0.740 (0.665 - 0.815)	—	1	0.750 (0.667 - 0.833)	—
High	16	0.819 (0.756 - 0.882)	98.9	19	0.808 (0.761 - 0.856)	97.4
Model types						
ANN[b]	2	0.867 (0.744 - 0.989)	98.5	2	0.781 (0.742 - 0.820)	0
DT[c]	—	—	—	1	0.930 (0.885 - 0.975)	—
KNN[d]	—	—	—	1	0.901 (0.866 - 0.935)	—
LASSO[e]	1	0.741 (0.678 - 0.804)	—	1	0.640 (0.531 - 0.749)	—
LR[f]	13	0.801 (0.715 - 0.887)	98.9	11	0.773 (0.724 - 0.823)	91.1
RF[g]	—	—	—	1	0.920 (0.915 - 0.925)	—
SVM[h]	—	—	—	1	0.941 (0.919 - 0.963)	—
XGBoost[i]	—	—	—	1	0.760 (0.651 - 0.869)	—
LGBM[j]	1	0.950 (0.935 - 0.965)	—	1	0.877 (0.808 - 0.946)	—
Overall	17	0.815 (0.754 - 0.876)	98.8	20	0.805 (0.759 - 0.852)	97.3

XSL•FO
RenderX

[a]Not applicable.

[b]ANN**:** artificial neural network.

[c]DT: decision tree.

[d]KNN: k-nearest neighbor.

[e]LASSO: least absolute shrinkage and selection operator.

[f]LR: logistic regression.

[g]RF: random forest.

[h]SVM: support vector machine.

[i]XGBoost: extreme gradient boosting.

[j]LightGBM: light gradient boosting machine.

Subgroup analysis results are also presented in Table 2. Of note, Subgroup analysis by model type indicated that most models were based on logistic regression. This approach achieved a c-index of 0.801 (95% CI 0.715 - 0.887) in the training set and 0.773 (95% CI 0.724 - 0.823) in the validation set. These values were lower than the overall c-index, suggesting that the performance of logistic regression models may be less effective than non-logistic regression models.

Notably, subgroup analyses revealed that heterogeneity remained extremely high across most subgroups and overall (I^2>90%), indicating substantial unexplained variability among performance estimates. This extreme heterogeneity suggests that while pooled c-indices were calculated, their clinical interpretability is limited; these values should be considered a broad approximation of possible outcomes rather than precise summary effect estimates.

Sensitivity and Specificity

In the subgroup analysis by model types, diagnostic 2 × 2 tables from 13 models were available in the training set, and 18 models were available in the validation set. A bivariate random-effects model was leveraged. The majority of models were based on logistic regression, with pooled sensitivity and specificity of 0.75 (95% CI 0.62 - 0.85) and 0.79 (95% CI 0.72 - 0.85) in the training set, and 0.73 (95% CI 0.64 - 0.80) and 0.70 (95% CI 0.65 - 0.74) in the validation set. Sensitivity and specificity results for other subgroups are presented in Table 3.

XSL•FO
RenderX

Table . Meta-analysis results of sensitivity and specificity in the included studies for machine learning prediction of postoperative delirium following cardiac surgery in adults.[a]

Subgroups	Training set			Validation set		
	Studies, n	Sensitivity (95% CI) or range	Specificity (95% CI) or range	Studies, n	Sensitivity (95% CI) or range	Specificity (95% CI) or range
Verification methods						
Random split validation	7	0.68 (0.56 - 0.78)	0.77 (0.69 - 0.83)	10	0.72 (0.62 - 0.80)	0.76 (0.67 - 0.82)
K-fold cross-validation	—[b]	—	—	3	0.67 - 0.86	0.72 - 0.91
Bootstrap resampling validation	—	—	—	2	0.71 - 0.91	0.67 - 0.68
External validation						
Temporal validation	3	0.72 - 0.97	0.80 - 0.93	3	0.35 - 0.79	0.68 - 0.96
No validation	3	0.66 - 0.81	0.66 - 0.82	—	—	—
Study types						
Retrospective cohort study	5	0.80 (0.52 - 0.93)	0.82 (0.72 - 0.89)	10	0.74 (0.66 - 0.80)	0.77 (0.69 - 0.84)
Prospective cohort study	7	0.76 (0.70 - 0.82)	0.78 (0.72 - 0.84)	6	0.77 (0.67 - 0.85)	0.70 (0.65 - 0.75)
Retrospective and prospective cohort study	1	0.64	0.64	2	0.35 - 0.53	0.77 - 0.96
Risk of bias (statistic analysis)						
Low	—	—	—	—	—	—
High	13	0.76 (0.66 - 0.84)	0.79 (0.73 - 0.84)	18	0.72 (0.65 - 0.79)	0.78 (0.71 - 0.83)
Model types						
ANN[c]	2	0.72 - 0.91	0.72 - 0.86	2	0.68 - 0.73	0.73 - 0.78
KNN[d]	—	—	—	1	0.67	0.91
LASSO[e]	1	0.72	0.73	1	0.60	0.62
LR[f]	10	0.75 (0.62 - 0.85)	0.79 (0.72 - 0.85)	10	0.73 (0.64 - 0.80)	0.70 (0.65 - 0.74)
RF[g]	—	—	—	1	0.86	0.92
SVM[h]	—	—	—	1	0.91	0.87
XGBoost[i]	—	—	—	1	0.67	0.79
LGBM[j]	—	—	—	1	0.35	0.96
Overall	13	0.76 (0.66 - 0.84)	0.79 (0.73 - 0.84)	18	0.72 (0.65 - 0.79)	0.78 (0.71 - 0.83)

[a]For subgroups with fewer than 4 studies (n<4), quantitative pooling was not performed; values presented represent the range from original studies. For subgroups with 1 study (n=1), the presented values are single numbers.

[b]Not applicable.

[c]ANN: artificial neural network.

[d]KNN: k-nearest neighbor.

[e]LASSO: least absolute shrinkage and selection operator.

[f]LR: logistic regression.

[g]RF: random forest.

XSL•FO
RenderX

[h]SVM: support vector machine.

[i]XGBoost: extreme gradient boosting.

[j]LightGBM: light gradient boosting machine.

The pooled sensitivity and specificity were 0.76 (95% CI 0.66 - 0.84) and 0.79 (95% CI 0.73 - 0.84) in the training set and 0.72 (95% CI 0.65 - 0.79) and 0.78 (95% CI 0.71 - 0.83) in the validation set, respectively (Table 3).

Publication Bias Analysis

Funnel plots were generated separately for the training and validation sets to assess publication bias. The results indicated no significant publication bias among the included studies in either the training (*P* value for Egger test was .65; Figure 3) or validation set (*P* value for Egger test was .59; Figure 4).

Figure 3. Publication bias analysis in the training set.

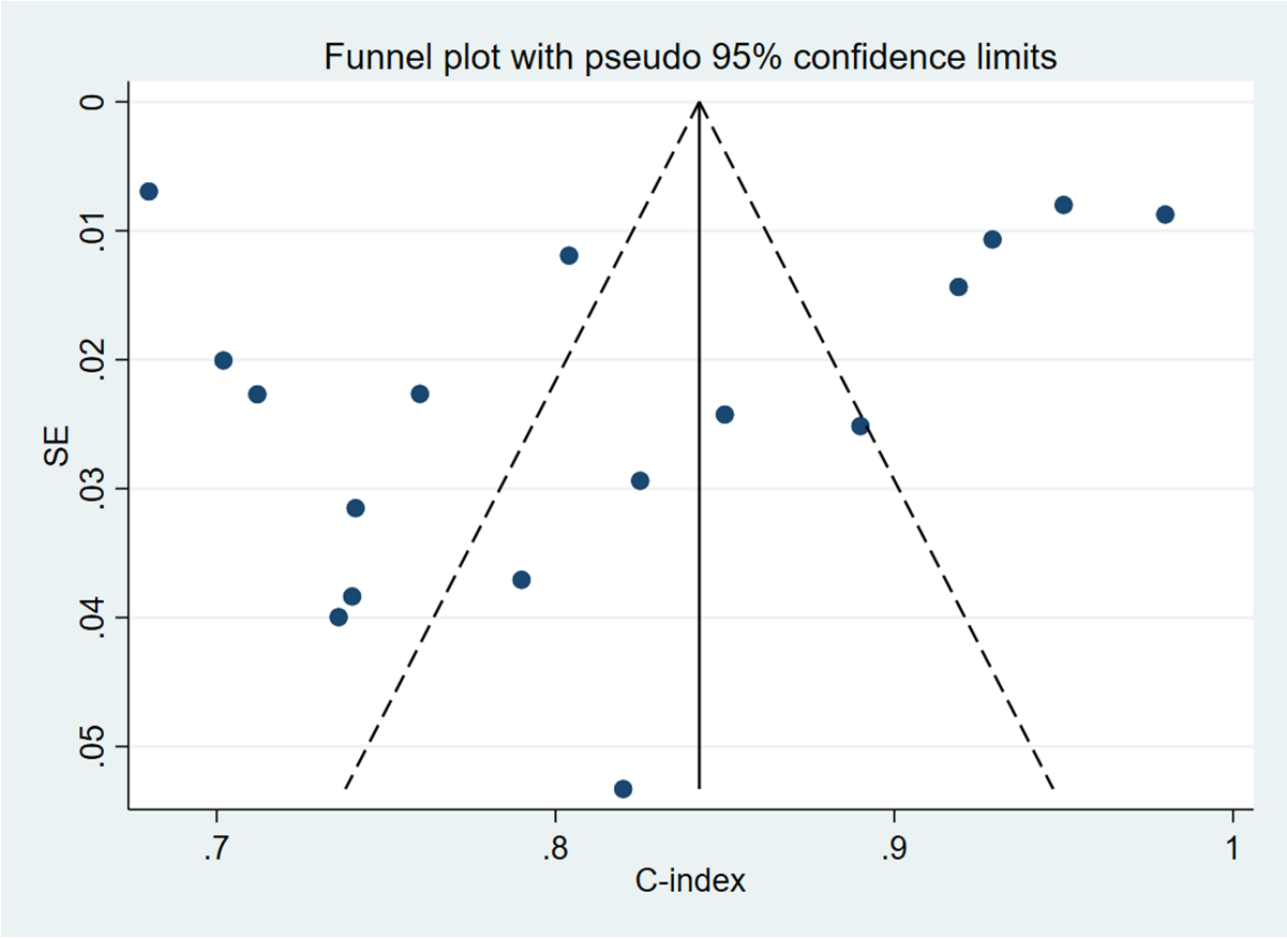

XSL•FO
RenderX

Figure 4. Publication bias analysis in the validation set.

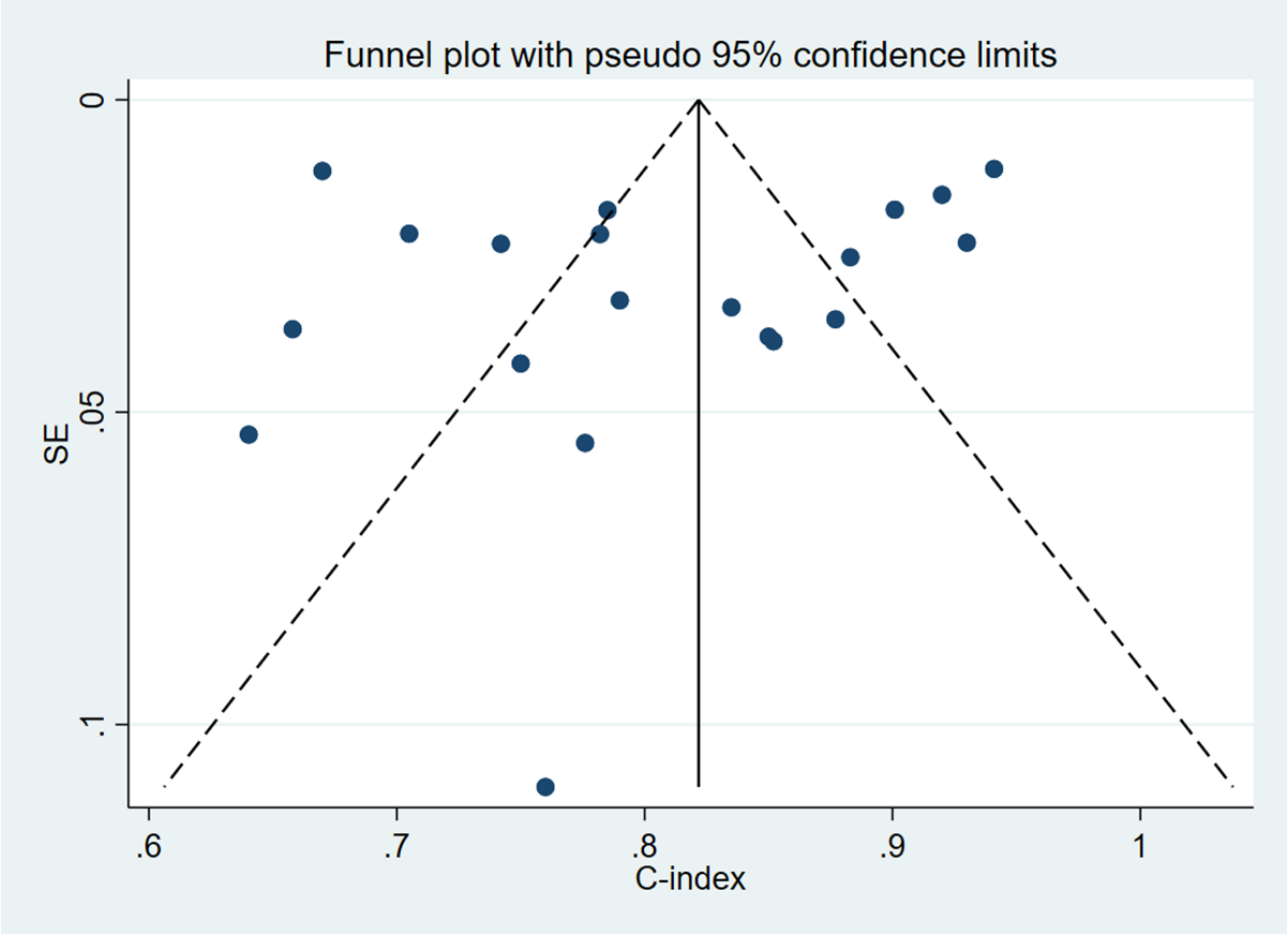

Studies for Prediction of POD Following Cardiac Surgery in Children

Two studies focusing on pediatric patients for predicting POD following cardiac surgery were ultimately included. Both were single-center studies that employed least absolute shrinkage and selection operator regression combined with multivariable logistic regression for variable selection and constructed nomogram models. The validation sets utilized random split validation and temporal validation, respectively. A meta-analysis was performed to synthesize model performance. In the validation set, the c-index, sensitivity, and specificity were 0.845 (95% CI 0.789 - 0.900), 0.84, and 0.80, respectively.

Discussion

Main Findings

This study revealed growing research interest in early prediction of POD following cardiac surgery, with ML models demonstrating promising effectiveness. However, the majority of existing models are characterized by substantial heterogeneity and high risk of bias. Logistic regression remained the predominant modeling method, showing favorable performance.

Comparison With Previous Reviews

Previous studies have reviewed early prediction methods for POD following cardiac surgery. For instance, a study by Cai et al [7] reviewed the use of ML in predicting POD in cardiac surgery. However, the included studies were limited, as the search strategy did not specifically account for the types of cardiac surgery. Furthermore, their study only provided the range of c-index values for the training set (0.74 - 0.91) and external validation set (0.54 - 0.90) without a quantitative systematic review. This limitation hindered a thorough interpretation of the value of ML in predicting POD in cardiac surgery.

In contrast, a study by Lee et al [8] summarized 3 well-established scoring tools for predicting POD in cardiac surgery. Their meta-analysis showed a pooled c-index of 0.62 (95% CI 0.58 - 0.66), while the international recalibrated PREDELIRIC model demonstrated a c-index of 0.75 (95% CI 0.72 - 0.79) in external validation. While these tools are well-established, they do not address the performance of ML. Our study has found that ML models appear to outperform these traditional scoring tools. Moreover, a review by Chen et al [9] examined POD in adults but did not specifically focus on cardiac surgery. The ML models included in their review demonstrated a c-index of 0.792 (95% CI 0.769 - 0.816) for predicting POD following cardiac surgery. Our study, based on a more comprehensive body of evidence, systematically summarizes the predictive value of ML for POD following cardiac surgery, with results from validation sets demonstrating strong predictive performance.

Previous studies have reviewed the main factors predicting POD in cardiac surgery. Koster et al [41] identified predisposing risk

XSL•FO
RenderX

factors such as atrial fibrillation, cognitive impairment, depression, stroke history, advanced age, and peripheral vascular disease. The frequency statistics in Table S3 in Multimedia Appendix 3 of the present study show that age, creatinine level, cardiopulmonary bypass duration, Mini-Mental State Examination score, and left ventricular ejection fraction are the most frequently used modeling variables in existing studies. The differences between the risk factors highlighted in this review and those reported in prior models may reflect variations in study design or data sources. Due to significant differences in modeling methods and population characteristics among the included studies, this review did not conduct formal predictive efficacy analysis to determine key predictive variables. The actual predictive value of these high-frequency variables requires further verification through more standardized designed studies.

Challenges in Clinical Translation

The studies incorporated in this review involve diverse ML models, with some observed differences in the predictive performance of these models, prompting us to summarize their predictive value for POD following cardiac surgery. The study has found that logistic regression is currently the predominant predictive model, as it allows for the construction of widely applicable or simpler predictive nomograms. These nomograms provide excellent interpretability of the relationships between clinical indicators and delirium. Although the interpretability of predictive models is a crucial measure in clinical research, we found that the accuracy of logistic regression is lower than that of several other ML models. Other models are less interpretable, which hinders their further application. Additionally, literature on these other models remains relatively scarce. Therefore, future studies should develop ML models with potentially higher accuracy, enabling earlier identification of delirium risk in cardiac surgery patients and the implementation of preventive measures.

Among included studies, the PROBAST assessment revealed high risk of bias. In the 28 included studies, high risk of bias primarily stemmed from participant selection bias (eg, some studies did not explicitly exclude patients with severe neurological comorbidities), inadequate handling of missing data, and lack of external validation or reliance solely on internal validation with small samples. These methodological limitations impose certain constraints on the interpretation of our findings: although the pooled results reflect the overall trends observed across existing studies, the actual clinical reliability of ML models warrants cautious evaluation due to methodological deficiencies in the original studies. These findings also offer important implications for future research: subsequent studies should rigorously adhere to the PROBAST assessment criteria and optimize study designs—such as refining participant selection processes, strengthening data quality control, and conducting multicenter external validation—to enhance the reliability and generalizability of ML models.

Although the predictive value of ML for POD following cardiac surgery was summarized and demonstrated relatively promising results, substantial heterogeneity cannot be ignored. This challenge is commonly faced by current meta-analyses on ML. To better explain heterogeneity, subgroup analyses were conducted based on data sets (training sets and validation sets), validation methods, study types, bias risk, and model types. Substantial heterogeneity remained among different subgroups. This heterogeneity may result from the diverse modeling variables in the model training sets. Even within the same type of model, differences may exist in model parameters and selected modeling variables. Medical conditions in different countries also seem to have potential impacts. Given these potential influences, a random-effects model was adopted for pooling. Overall, due to the impact of heterogeneity, these results require cautious interpretation. In clinical applications, comprehensive judgment based on specific population characteristics and diagnosis and treatment scenarios is necessary to avoid direct application.

Limitations

This study has several limitations. First, despite comprehensive systematic searching, the limited number of included studies may restrict result interpretation, particularly for models based on few studies. Second, original studies did not differentiate between cardiac surgery types and delirium occurrence locations (ICU vs general wards), precluding more detailed analysis. Third, most studies utilized internal validation without independent external validation, potentially limiting generalizability. Fourth, due to language search barriers, only English-language studies were included, potentially introducing bias and limiting result interpretation. Fifth, included populations showed limitations regarding source, sample size, and modeling method standardization, creating clinical application challenges without effective model stratification analysis. Sixth, this study can only reflect the variable usage trends in existing studies and cannot directly infer the predictive importance of variables. This may also result in certain established risk factors (eg, gender) appearing less frequently than their recognized importance in general clinical research would suggest. Such patterns reflect the variable selection outcomes of the ML models employed rather than the biological significance of these factors and therefore warrant cautious interpretation. Seventh, inherent limitations in the included studies—predominantly single-center, retrospective designs lacking rigorous external validation—contributed to the high risk of bias in the models, suggesting that the pooled predictive performance may be overestimated. The generalizability of these models requires further validation through prospective, multicenter studies. Therefore, future research should standardize model construction methods, enhance model transparency, and reduce bias risk to broaden clinical application. Additionally, our search strategy primarily focused on “nomogram” and specific ML terminology, potentially missing studies that reported traditional prediction models using only terms such as “logistic regression” or “multivariate analysis,” which may have introduced search bias. Although “nomogram” is a common presentation format for logistic regression models, future systematic reviews should consider more comprehensive search terms.

Conclusions

This systematic review suggests that ML-based prediction tools for POD following cardiac surgery appear to demonstrate promising performance. However, the majority of existing

XSL•FO
RenderX

models carry high risk of bias and lack rigorous external validation, potentially resulting in overly optimistic performance estimates. Current evidence is substantially constrained by extreme heterogeneity and low methodological quality. Future research should therefore standardize ML model development workflows, prioritize prospective designs, adhere to prediction model reporting standards, conduct extensive external validation, and perform comparative model studies to establish the actual clinical utility and feasibility of these tools in real-world settings.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Authors' Contributions

Writing - original draft preparation: YG
Writing - review and editing: YG
Conceptualization: PX
Methodology: PX
Formal analysis and investigation: SZ
Funding acquisition: AW
Resources: HX
Supervision: MZ
All authors contributed to the study conception and design and authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Literature search strategy.
[DOCX File, 18 KB - jmir_v28i1e72304_app1.docx]

Multimedia Appendix 2
Basic information of included studies.
[DOCX File, 41 KB - jmir_v28i1e72304_app2.docx]

Multimedia Appendix 3
Modeling variables in included studies.
[DOCX File, 26 KB - jmir_v28i1e72304_app3.docx]

Checklist 1
PRISMA 2020 checklist.
[PDF File, 101 KB - jmir_v28i1e72304_app4.pdf]

References

1. Timmis A, Vardas P, Townsend N, et al. European Society of Cardiology: cardiovascular disease statistics 2021. Eur Heart J 2022 Feb 22;43(8):716-799. [doi: 10.1093/eurheartj/ehab892] [Medline: 35016208]
2. Stephens RS, Whitman GJR. Postoperative critical care of the adult cardiac surgical patient: part II: procedure-specific considerations, management of complications, and quality improvement. Crit Care Med 2015 Sep;43(9):1995-2014. [doi: 10.1097/CCM.0000000000001171] [Medline: 26136101]
3. Andersson B, Nilsson J, Brandt J, Höglund P, Andersson R. Gastrointestinal complications after cardiac surgery. Br J Surg 2005 Mar;92(3):326-333. [doi: 10.1002/bjs.4823] [Medline: 15672438]
4. Newman MF, Kirchner JL, Phillips-Bute B, et al. Longitudinal assessment of neurocognitive function after coronary-artery bypass surgery. N Engl J Med 2001 Feb 8;344(6):395-402. [doi: 10.1056/NEJM200102083440601] [Medline: 11172175]
5. Saczynski JS, Marcantonio ER, Quach L, et al. Cognitive trajectories after postoperative delirium. N Engl J Med 2012 Jul 5;367(1):30-39. [doi: 10.1056/NEJMoa1112923] [Medline: 22762316]

XSL•FO
RenderX

6. Salluh JIF, Wang H, Schneider EB, et al. Outcome of delirium in critically ill patients: systematic review and meta-analysis. BMJ 2015 Jun 3;350:h2538. [doi: 10.1136/bmj.h2538] [Medline: 26041151]
7. Cai S, Li J, Gao J, Pan W, Zhang Y. Prediction models for postoperative delirium after cardiac surgery: systematic review and critical appraisal. Int J Nurs Stud 2022 Dec;136:104340. [doi: 10.1016/j.ijnurstu.2022.104340] [Medline: 36208541]
8. Lee A, Mu JL, Joynt GM, et al. Risk prediction models for delirium in the intensive care unit after cardiac surgery: a systematic review and independent external validation. Br J Anaesth 2017 Mar 1;118(3):391-399. [doi: 10.1093/bja/aew476] [Medline: 28186224]
9. Chen H, Yu D, Zhang J, Li J. Machine learning for prediction of postoperative delirium in adult patients: a systematic review and meta-analysis. Clin Ther 2024 Dec;46(12):1069-1081. [doi: 10.1016/j.clinthera.2024.09.013] [Medline: 39395856]
10. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019 Jan 1;170(1):51-58. [doi: 10.7326/M18-1376] [Medline: 30596875]
11. Debray TP, Damen JA, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Stat Methods Med Res 2019 Sep;28(9):2768-2786. [doi: 10.1177/0962280218785504] [Medline: 30032705]
12. Hatano Y, Narumoto J, Shibata K, et al. White-matter hyperintensities predict delirium after cardiac surgery. Am J Geriatr Psychiatry 2013 Oct;21(10):938-945. [doi: 10.1016/j.jagp.2013.01.061] [Medline: 24029014]
13. Mao D, Fu L, Zhang W. Risk factors and nomogram model of postoperative delirium in children with congenital heart disease: a single-center prospective study. Pediatr Cardiol 2024 Jan;45(1):68-80. [doi: 10.1007/s00246-023-03297-5] [Medline: 37741935]
14. Han C, Kim HI, Soh S, Choi JW, Song JW, Yoon D. Machine learning with clinical and intraoperative biosignal data for predicting postoperative delirium after cardiac surgery. iScience 2024 Jun 21;27(6):109932. [doi: 10.1016/j.isci.2024.109932] [Medline: 38799563]
15. Zhao X, Li J, Xie X, et al. Online interpretable dynamic prediction models for postoperative delirium after cardiac surgery under cardiopulmonary bypass developed based on machine learning algorithms: a retrospective cohort study. J Psychosom Res 2024 Jan;176:111553. [doi: 10.1016/j.jpsychores.2023.111553] [Medline: 37995429]
16. Lin N, Lv M, Li S, Xiang Y, Li J, Xu H. A nomogram for predicting postoperative delirium in pediatric patients following cardiopulmonary bypass: a prospective observational study. Intensive Crit Care Nurs 2024 Aug;83:103717. [doi: 10.1016/j.iccn.2024.103717] [Medline: 38692080]
17. de la Varga-Martínez O, Gómez-Pesquera E, Muñoz-Moreno MF, et al. Development and validation of a delirium risk prediction preoperative model for cardiac surgery patients (DELIPRECAS): an observational multicentre study. J Clin Anesth 2021 May;69:110158. [doi: 10.1016/j.jclinane.2020.110158] [Medline: 33296785]
18. Zhang Y, Ren M, Zhai W, Han J, Guo Z. Construction and validation of a risk prediction model for postoperative delirium in patients with off‐pump coronary artery bypass grafting. J Thorac Dis 2024 Jun;16(6):3944-3955. [doi: 10.21037/jtd-24-578] [Medline: 38983165]
19. Nagata C, Hata M, Miyazaki Y, et al. Development of postoperative delirium prediction models in patients undergoing cardiovascular surgery using machine learning algorithms. Sci Rep 2023 Nov 30;13(1):21090. [doi: 10.1038/s41598-023-48418-5] [Medline: 38036664]
20. Lin JL, Zheng GZ, Chen LW, Luo ZR. A nomogram model for assessing predictors and prognosis of postoperative delirium in patients receiving acute type A aortic dissection surgery. BMC Cardiovasc Disord 2023 Feb 7;23(1):72. [doi: 10.1186/s12872-023-03111-3] [Medline: 36750929]
21. Hata M, Miyazaki Y, Nagata C, et al. Predicting postoperative delirium after cardiovascular surgeries from preoperative portable electroencephalography oscillations. Front Psychiatry 2023;14:1287607. [doi: 10.3389/fpsyt.2023.1287607] [Medline: 38034919]
22. Xu Y, Meng Y, Qian X, et al. Prediction model for delirium in patients with cardiovascular surgery: development and validation. J Cardiothorac Surg 2022 Oct 1;17(1):247. [doi: 10.1186/s13019-022-02005-3] [Medline: 36183105]
23. Segernäs A, Skoog J, Ahlgren Andersson E, Almerud Österberg S, Thulesius H, Zachrisson H. Prediction of postoperative delirium after cardiac surgery with a quick test of cognitive speed, mini-mental state examination and hospital anxiety and depression scale. Clin Interv Aging 2022;17:359-368. [doi: 10.2147/CIA.S350195] [Medline: 35400995]
24. Cai S, Cui H, Pan W, Li J, Lin X, Zhang Y. Two-stage prediction model for postoperative delirium in patients in the intensive care unit after cardiac surgery. Eur J Cardiothorac Surg 2022 Dec 2;63(1):ezac573. [doi: 10.1093/ejcts/ezac573] [Medline: 36579859]
25. He J, Ling Q, Chen Y. Construction and application of a model for predicting the risk of delirium in postoperative patients with type a aortic dissection. Front Surg 2021;8:772675. [doi: 10.3389/fsurg.2021.772675] [Medline: 34869569]
26. Kotfis K, Ślozowska J, Safranow K, Szylińska A, Listewnik M. The practical use of white cell inflammatory biomarkers in prediction of postoperative delirium after cardiac surgery. Brain Sci 2019 Nov 2;9(11):308. [doi: 10.3390/brainsci9110308] [Medline: 31684066]
27. Ten Broeke M, Koster S, Konings T, Hensens AG, van der Palen J. Can we predict a delirium after cardiac surgery? A validation study of a delirium risk checklist. Eur J Cardiovasc Nurs 2018 Mar;17(3):255-261. [doi: 10.1177/1474515117733365] [Medline: 28980478]

XSL•FO
RenderX

28. Price CC, Garvan C, Hizel LP, Lopez MG, Billings FT 4th. Delayed recall and working memory MMSE domains predict delirium following cardiac surgery. J Alzheimers Dis 2017;59(3):1027-1035. [doi: 10.3233/JAD-170380] [Medline: 28697572]
29. Koster S, Hensens AG, Schuurmans MJ, van der Palen J. Prediction of delirium after cardiac surgery and the use of a risk checklist. Eur J Cardiovasc Nurs 2013 Jun;12(3):284-292. [doi: 10.1177/1474515112450244] [Medline: 22694810]
30. Li Q, Li J, Chen J, et al. A machine learning-based prediction model for postoperative delirium in cardiac valve surgery using electronic health records. BMC Cardiovasc Disord 2024 Jan 18;24(1):56. [doi: 10.1186/s12872-024-03723-3] [Medline: 38238677]
31. Tian Y, Ji B, Diao X, et al. Dynamic predictive scores for cardiac surgery-associated agitated delirium: a single-center retrospective observational study. J Cardiothorac Surg 2023 Jul 6;18(1):219. [doi: 10.1186/s13019-023-02339-6] [Medline: 37415226]
32. Li X, Cheng W, Zhang J, Li D, Wang F, Cui N. Early alteration of peripheral blood lymphocyte subsets as a risk factor for delirium in critically ill patients after cardiac surgery: a prospective observational study. Front Aging Neurosci 2022;14:950188. [doi: 10.3389/fnagi.2022.950188]
33. Andrási TB, Talipov I, Dinges G, Arndt C, Rastan AJ. Risk factors for postoperative delirium after cardiac surgical procedures with cardioplegic arrest. Eur J Cardiothorac Surg 2022 Jun 15;62(1):ezab570. [doi: 10.1093/ejcts/ezab570] [Medline: 35037042]
34. Rudolph JL, Jones RN, Levkoff SE, et al. Derivation and validation of a preoperative prediction rule for delirium after cardiac surgery. Circulation 2009 Jan 20;119(2):229-236. [doi: 10.1161/CIRCULATIONAHA.108.795260] [Medline: 19118253]
35. Wang Y, Shen B, Zhu C, et al. Unveiling the nexus of postoperative fever and delirium in cardiac surgery: identifying predictors for enhanced patient care. Front Cardiovasc Med 2023;10:1237055. [doi: 10.3389/fcvm.2023.1237055]
36. Yang T, Yang H, Liu Y, et al. Postoperative delirium prediction after cardiac surgery using machine learning models. Comput Biol Med 2024 Feb;169:107818. [doi: 10.1016/j.compbiomed.2023.107818] [Medline: 38134752]
37. Sadlonova M, Hansen N, Esselmann H, et al. Preoperative delirium risk screening in patients undergoing a cardiac surgery: results from the prospective observational FINDERI study. Am J Geriatr Psychiatry 2024 Jul;32(7):835-851. [doi: 10.1016/j.jagp.2023.12.017] [Medline: 38228452]
38. Mufti HN, Hirsch GM, Abidi SR, Abidi SSR. Exploiting machine learning algorithms and methods for the prediction of agitated delirium after cardiac surgery: models development and validation study. JMIR Med Inform 2019 Oct 23;7(4):e14993. [doi: 10.2196/14993] [Medline: 31558433]
39. Lapp L, Roper M, Kavanagh K, Schraag S. Predicting the onset of delirium on hourly basis in an intensive care unit following cardiac surgery. Presented at: 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS); Jul 21-23, 2022; Shenzen, China p. 234-239. [doi: 10.1109/CBMS55023.2022.00048]
40. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019 Jan 1;170(1):W1-W33. [doi: 10.7326/M18-1377] [Medline: 30596876]
41. Koster S, Hensens AG, Schuurmans MJ, van der Palen J. Risk factors of delirium after cardiac surgery: a systematic review. Eur J Cardiovasc Nurs 2011 Dec;10(4):197-204. [doi: 10.1016/j.ejcnurse.2010.09.001] [Medline: 20870463]

Abbreviations

ML: machine learning
POD: postoperative delirium
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction Model Risk of Bias Assessment Tool

Edited by A Coristine; submitted 07.Feb.2025; peer-reviewed by A Jafarizadeh, J Zeng; accepted 12.Jan.2026; published 23.Feb.2026.

Please cite as:
Guo Y, Xu H, Wang A, Zhang M, Zhang S, Xie P
The Predictive Value of Machine Learning for Postoperative Delirium in Cardiac Surgery: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e72304
URL: https://www.jmir.org/2026/1/e72304
doi:10.2196/72304

© Yi Guo, Hong Xu, Ankui Wang, Mingming Zhang, Shuai Zhang, Peng Xie. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Feb.2026. This is an open-access article distributed under the terms of the Creative

XSL•FO
RenderX

Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Securing Federated Learning With Blockchain in the Medical Field: Systematic Literature Review

Xudong Wang[1*], MD; Yi Xie[2*], MD; Xiaoliang Chen[1,3*], MD; Jiaming Yang[1], MD; Ruiyuan Li[1], MD; Weihang Gao[1], MD; Zineng Yan[1], MD; Hong Zhou[1], MD; Zhewei Ye[1], Prof Dr Med

[1]Wuhan Union Hospital, 1277 Jiefang Avenue, Wuhan, Hubei Province, China

[2]Renmin Hospital of Wuhan University, Wuhan, Hubei Province, China

[3]Ningxia Hui Autonomous Region Peoples Hospital, Yinchuan, Ningxia, China

[*]these authors contributed equally

Corresponding Author:
Zhewei Ye, Prof Dr Med
Wuhan Union Hospital, 1277 Jiefang Avenue, Wuhan, Hubei Province, China

Abstract

Background: The exponential growth of medical data and advancements in artificial intelligence (AI) have accelerated the development of data-driven health care. However, the secure and efficient sharing of sensitive medical data across institutions remains a major challenge due to privacy concerns, data silos, and regulatory restrictions. Traditional centralized systems are prone to data breaches and single points of failure, while existing privacy-preserving techniques face high computational and communication costs.

Objective: This study aims to provide a comprehensive review of the recent advances in blockchain-based federated learning (BCFL) within the medical field. By exploring the synergistic integration of federated learning and blockchain, this review evaluates how BCFL enhances data security, supports privacy-preserving cross-institutional collaboration, and facilitates practical applications in health care, including medical data sharing, Internet of Medical Things, public health surveillance, and telemedicine.

Methods: We conducted a systematic literature review using databases such as PubMed, IEEE Xplore, Web of Science, and Google Scholar. Boolean logic and domain-specific keywords were used to retrieve studies from 2018 to 2025. After automated deduplication and multistage manual screening, over 100 high-quality papers were included. These works cover BCFL's theoretical foundations, system architectures, application domains, limitations, and future directions.

Results: BCFL frameworks combine the decentralized trust and auditability of blockchain with the privacy-preserving collaborative learning capabilities of federated learning. This integration mitigates risks such as model tampering, data leakage, and a lack of incentives in federated systems. Applications span across cross-institutional medical data sharing, Internet of Medical Things, epidemic forecasting, and telemedicine. Architectures including fully coupled, flexibly coupled, and loosely coupled models offer varying trade-offs between efficiency, scalability, and security.

Conclusions: BCFL represents a transformative paradigm for secure, collaborative, and privacy-preserving medical AI. By combining decentralized trust, incentive-driven participation, and privacy-enhancing machine learning, BCFL paves the way for next-generation smart health care systems. Despite current technical and practical challenges, BCFL demonstrates strong potential to support precision medicine, global health data collaboration, and large-scale AI deployment in health care.

(J Med Internet Res 2026;28:e79052) doi:10.2196/79052

KEYWORDS

blockchain; federated learning; health care; health data; review; Internet of Medical Things; COVID-19; IoMT

Introduction

Background

With the continuous advancement of medical informatization, the volume of medical big data is growing exponentially, offering significant opportunities for the application of artificial intelligence (AI) in health care, particularly in areas such as assisted diagnosis, personalized treatment, and disease prediction. Especially in recent years, advances in computing power and algorithmic innovation have made machine learning (ML) models a cornerstone of medical intelligence, with their efficient training and optimization relying heavily on large-scale, high-quality datasets from multiple sources [1-3]. However, the acquisition and sharing of medical data face significant obstacles, including privacy concerns, data security risks, and regulatory constraints [4].

In the digital medicine era, the secure exchange and control of sensitive health information have become central concerns in

XSL•FO
RenderX

modern health care systems [5,6]. Meanwhile, strict legal and regulatory requirements—such as HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation)—must be met when handling these personal data [7]. Additionally, the competition between different medical centers and hospitals has led to data often remaining siloed [8]. These challenges not only impede the effective integration of multisource health care data but also hinder the translation of ML models from theoretical research to clinical practice.

While conventional centralized data storage architectures can partially facilitate model development, their dependence on centralized infrastructures has become increasingly problematic, rendering them vulnerable to single-point failures and malicious cyberattacks [9,10]. Consequently, establishing secure frameworks for cross-institutional data sharing and intelligent processing—while rigorously protecting data security and patient confidentiality—has emerged as a pivotal challenge impeding progress in medical AI. The fundamental tension between aggregating health care data for scientific progress and preserving individual privacy and data security has spurred the development of novel computational approaches. Federated learning (FL), as an emerging ML paradigm, addresses part of this problem by allowing institutions to collaboratively train models without exchanging raw data [11,12]. However, in practice, the deployment of FL in health care faces practical obstacles: the integrity and authenticity of model updates may be subject to security threats, including malicious attacks, client-side data tampering, and model forgery—all of which can reduce the accuracy of the global model. Furthermore, at present, there is a lack of reliable incentive structures for continuous participation, with limited auditability. The heterogeneity between institutions and edge devices undermines the integration and generalization of the model. Blockchain technology, with its decentralized architecture, immutable ledger system, and transparent traceability, has become a potential solution to address these limitations of logical reasoning [2,13,14]. When the two are integrated, blockchain can provide a verifiable source for model contributions, automated and transparent incentive mechanisms [15,16], tamper-proof logs for auditing, and a governance layer that supports cross-institutional workflows [17]. Therefore, the combined paradigm of blockchain-based federated learning (BCFL) is expected to become a practical approach to coordinating data privacy, trust, and collaborative intelligence in medicine.

Although research on FL and blockchain has accelerated in recent years, existing reviews exhibit clear gaps in focus and depth, limiting their value for clinical researchers and multidisciplinary audiences. The main shortcomings can be summarized as follows. First, most reviews examine either FL or blockchain technology in isolation, without providing a systematic analysis of how these two approaches can be integrated to address concrete challenges in health care. Second, prior reviews tend to emphasize algorithmic and technical details, while offering limited discussion of how BCFL can be adapted to real-world medical scenarios—such as cross-hospital electronic health record (EHR) integration, collaborative training of medical imaging models, Internet of Medical Things (IoMT) device coordination, and epidemic surveillance. Finally, few reviews adequately address the challenges, regulatory considerations, and future development trends of BCFL within the constraints of modern health care governance frameworks.

To address these gaps, this review makes three key contributions. First, it provides a medical demand–oriented, systematic classification of BCFL frameworks and outlines their typical workflows. Second, it maps different BCFL architectures to representative health care application scenarios, clarifying their practical relevance. Third, it analyzes the technical, regulatory, and implementation challenges that currently hinder BCFL adoption and identifies promising directions for future research, providing evidence-based insights for clinical translation and decision-making in the health care and biomedical research communities.

Objective

Although previous studies have analyzed BCFL frameworks from the perspectives of technology and commercial applications, the review of their practical application scenarios in health care remains limited. Especially in the era of AI, with the explosive growth of medical data, the application of large models in medical practice has encountered some development obstacles, bringing some new ideas for the application prospects of BCFL in health care. This study investigates the synergistic integration of FL and blockchain technology, elucidating their combined architectural framework and operational mechanisms. We demonstrate how this technological convergence enables secure, privacy-preserving health care data sharing and collaborative model development across distributed health care institutions. Although blockchain-enhanced FL has recently emerged as a promising approach in medical research, the field remains nascent, constrained by technical limitations, unresolved privacy issues, and implementation barriers. Furthermore, we present a systematic review of current advancements in blockchain-assisted FL for medical applications and propose actionable research directions to overcome existing challenges.

Key contributions of this work include:

1. This study systematically reviews the theoretical foundations of FL and blockchain technology and elaborates on their potential and advantages in the medical field.
2. A comprehensive taxonomy of existing integration frameworks, accompanied by a mechanistic analysis of bidirectional benefits: how blockchain enhances FL security and trustworthiness, and how FL expands blockchain's utility in distributed computing scenarios.
3. To summarize recent advancements and practical applications of BCFL across key health care domains, including cross-institutional medical data sharing, IoMT, public health surveillance, and telemedicine.
4. The current technological limitations and application challenges are examined, and key future research directions are proposed to address these gaps and advance real-world implementation.

By synthesizing current research and offering a structured analytical framework, we aim to provide a comprehensive reference for medical personnel, researchers, and health care

XSL•FO
RenderX

policymakers. Ultimately, it fosters the development of trustworthy, privacy-preserving, and collaborative AI systems that can support precision medicine and smart health care in a decentralized digital era.

Methods

Overview

This review systematically summarizes recent advances in BCFL within the medical field, following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist (Checklist 1). To ensure methodological rigor and reproducibility, we adopted a transparent multistep process including comprehensive literature search, independent screening, quality assessment, and evidence synthesis.

Search Strategy

The literature data were primarily retrieved from several prominent academic databases, including PubMed, IEEE Xplore, Web of Science, and Google Scholar. To ensure the timeliness of the research, the review focuses on literature published between January 2018 and February 2025, while also incorporating some early seminal studies to trace the theoretical development and technological evolution of BCFL in the medical field.

After that, the search strategy uses Boolean logic operators to formulate a comprehensive formula, for example, ("blockchain") AND ("federated learning") AND ("medical" OR "healthcare"). To enhance retrieval efficiency and encompass interdisciplinary intersections, the search terms are further expanded, such as "(blockchain-enabled federated learning)," "(distributed machine learning)," "(decentralization)," "(Internet of Medical Things (IoMT)," "(telemedicine)," "(EMR)," and "(epidemics)" are incorporated to ensure literature comprehensiveness. The exact search string is as follows: ("blockchain" OR "distributed ledger technology") AND ("federated learning" OR "collaborative learning" OR "distributed machine learning") AND ("healthcare" OR "medical" OR "clinical" OR "EMR" OR "(epidemics)" OR "IoMT" OR "telemedicine"). Gray literature (conference proceedings and preprints from arXiv) was screened manually and included only if it contained original data or detailed technical methodology relevant to BCFL.

Inclusion and Exclusion Criteria

To ensure relevance and academic rigor, this review establishes strict inclusion and exclusion criteria. The inclusion criteria are as follows: (1) the studies must involve blockchain and FL technologies and explore their medical applications; (2) the literature published in peer-reviewed journals and reviews indexed in the Science Citation Index or Social Sciences Citation Index or in top-tier international conferences (eg, IEEE and Association for Computing Machinery) and seminal papers with >30 citations were included regardless of publication venue; and (3) reported theoretical frameworks, system architectures, empirical evaluations, or case studies. The exclusion criteria then include (1) studies focusing solely on blockchain or FL without medical applications, (2) editorial and opinion articles that lack technical details or empirical validation, and (3) duplicate reports or low-quality publications from non–peer-reviewed sources.

Literature Screening and Processing

The above search strategy initially retrieved 2547 documents. After automatic deduplication by EndNote, 1327 were retained. Subsequently, two independent reviewers (XW and XC) manually screened all the literature in two stages based on the inclusion and exclusion criteria. First, based on the title and abstract, high-quality reviews, papers from top journals and conferences, and literature that clearly explained the application of BCFL in the health care field were selected, totaling 319 articles. In the second stage, further review was conducted. Through full-text reading, studies lacking empirical verification, technical details, or experimental data were eliminated, and ultimately 111 high-quality documents were retained. In addition, to ensure the comprehensiveness of the literature, this review also referred to the latest review papers and included the core research results cited therein to avoid missing key progress. Detailed literature retrieval strategies can be found in Multimedia Appendix 1. After the screening process, the reviewers evaluated the quality of the literature. Since the corpus of this review mainly comes from the cross-disciplinary research of medicine, engineering, and computer science, the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) framework commonly used in clinical evidence-based studies was not applicable. We used the Systematic Literature Review quality checklist based on the Kitchenham principle [18] to score each of the 12 indicators of the included studies, including reproducibility, method transparency, evaluation design, data description, experimental validity, attack/privacy discussion, and reproducibility, item by item (1/0.5/0). Two independent reviewers (XW and XC) assessed each document. When there were significant differences in the scores (the total score difference of a single document was greater than 2 or there were differences in key items), a third senior reviewer (YX) arbitrated. Ultimately, the studies were classified into three quality grades based on the total score: high (≥9), medium (5-8), and low (<5). When writing summaries and conclusions, priority was given to citing high-quality research. Multimedia Appendix 2 [1-130] provides complete assessment criteria, scoring rules, and statistics on peer-to-peer agreements.

Results

Study Findings

Finally, this review is founded on over 100 strictly selected papers, encompassing theoretical research, technical architecture, application scenarios, challenges, and future trends. These studies provide a robust academic foundation for an in-depth exploration of BCFL applications in medicine. A PRISMA flow diagram illustrates the systematic selection process (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the systematic review phases. FL: federated learning.

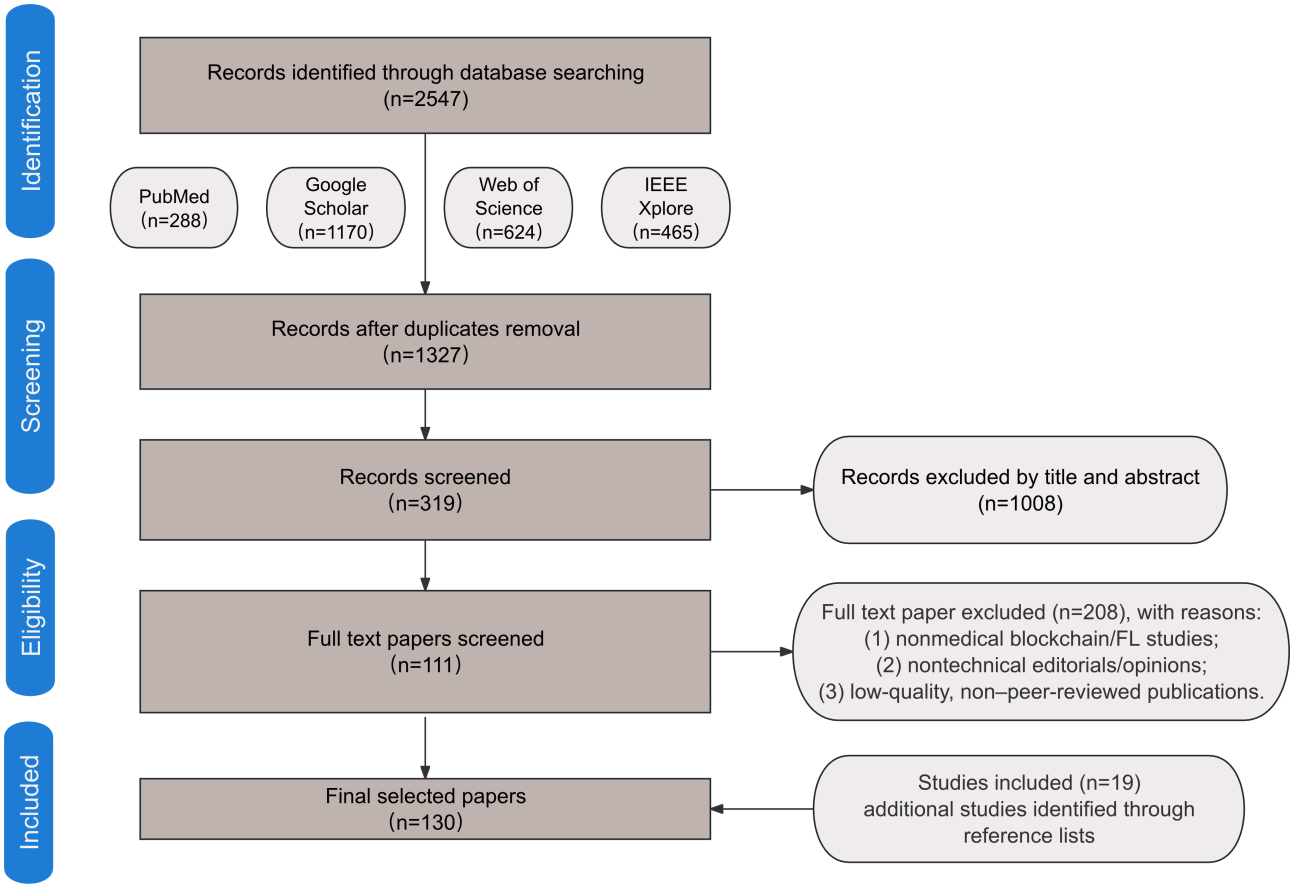

Current Status of Artificial Intelligence in Medicine

In the context of the rapid advancements in AI, the performance of ML models heavily depends on access to large volumes of high-quality data. However, the health care sector has long struggled with data sharing due to concerns over privacy, security, and regulatory compliance. The prevalence of "data silos" in health care impedes the advancement of AI and hinders the translation of research findings into practical clinical applications.

Traditional centralized ML approaches require aggregating raw data from multiple sources to a central server for training. While this approach allows for extensive data usage, it still presents significant challenges in practical implementation. First, data stored and transmitted in a centralized system is susceptible to network attacks, which can lead to sensitive information theft or tampering, posing a serious privacy risk. Second, centralized storage faces significant compliance challenges, especially under stringent privacy regulations such as GDPR [7] and HIPAA, which further restrict the cross-organizational exchange of health care data. Moreover, health care organizations, research institutions, and enterprises are often reluctant to share critical data due to competitive concerns and resource protection, thus further exacerbating data silos and hindering cross-organizational collaboration in building high-quality ML models.

Federated Learning in Health Care

In this context, the limitations of traditional algorithms are becoming increasingly apparent, driving researchers to explore innovative solutions. In 2016, Google introduced the concept of federated learning [19], a distributed and collaborative ML paradigm. At its core, FL enables multiple data holders to collaboratively develop a global model by training locally and exchanging model parameters without sharing raw data. This paradigm is well-suited for scenarios with strict privacy requirements and decentralized data that cannot be centrally stored, such as health care, finance, and smart cities [20-23]. This approach not only effectively enhances data privacy protection but also overcomes the limitation of data silos, ushering in a new era of privacy-preserving collaborative learning.

FL operates on the principles of distributed model training and global parameter aggregation. One of its core algorithms is the Federated Averaging (FedAvg) algorithm. As illustrated in Figure 2, the fundamental process can be summarized as follows [24,25]. The first step is that a central coordination server initializes the global model and distributes it to all participating clients (eg, hospitals or mobile devices). Then each client trains the model locally using its private dataset and transmits the updated model parameters (or gradients) back to the central server in an encrypted form. Following this, the central server collects and aggregates the model parameters uploaded by all clients and updates the global model based on a predefined aggregation algorithm (eg, FedAvg). The updated model is then

XSL•FO
RenderX

redistributed to the clients for the next training round. This process is iteratively performed over multiple communication rounds until the global model converges or predefined performance metrics are met.

Figure 2. Federated learning architecture and workflow in the medical field.

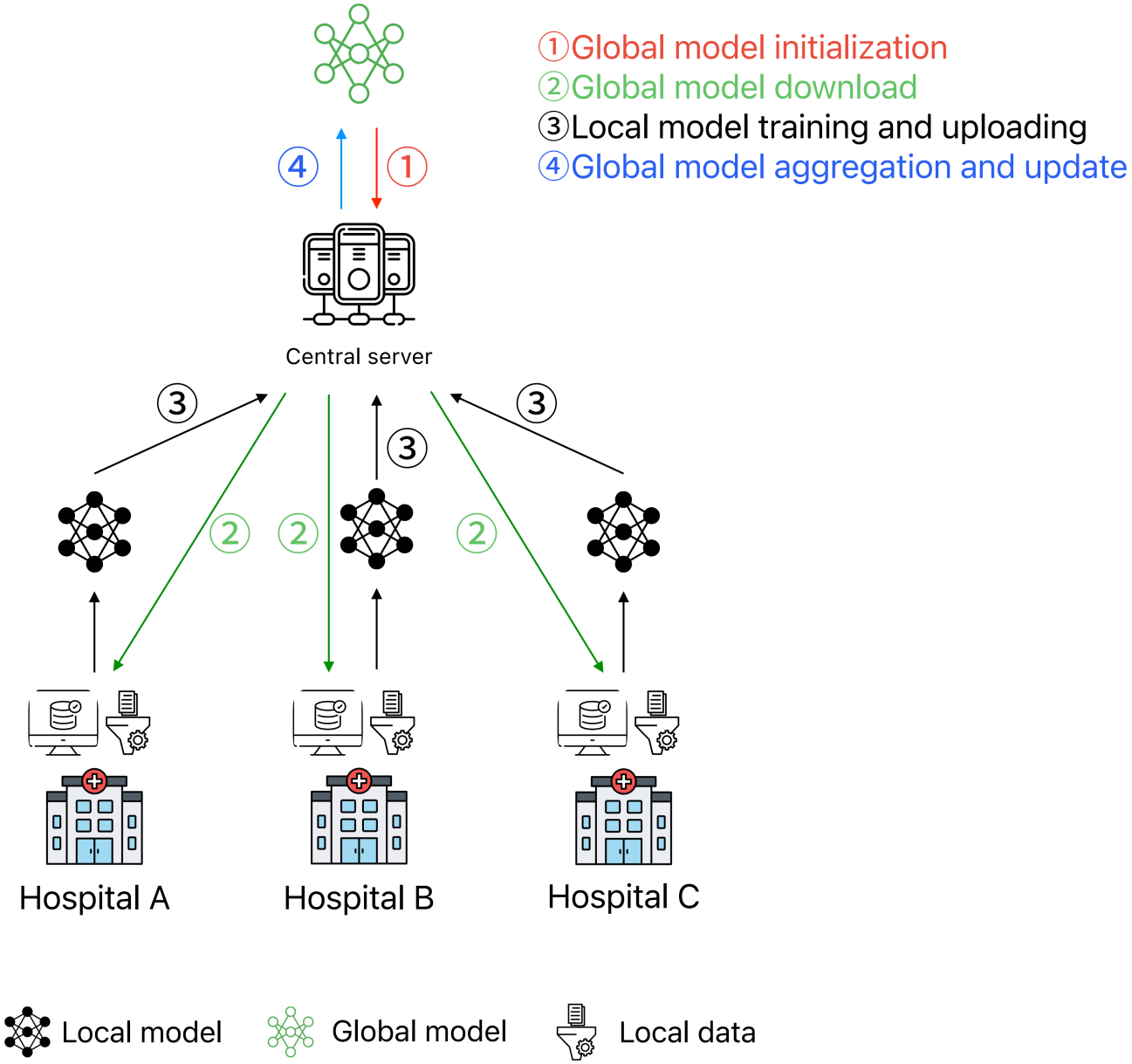

Based on the above operating principle of FL, its unique technical characteristics give it different advantages in the health care field: if it can use large-scale, diverse, and geographically distributed datasets without compromising patient privacy, this decentralized approach not only mitigates the impact of data silos, which are common in hospitals and research institutions, but also captures a broader range of features by leveraging heterogeneous real datasets, ultimately achieving an AI model with higher efficiency, robustness, and accuracy. Therefore, since its proposal, FL has shown great application potential in the field of health care [26]. For instance, FL has been applied in drug research, allowing pharmaceutical companies to leverage shared algorithmic models to accelerate drug discovery while avoiding direct data sharing [27]. Additionally, FL-based collaborative protocols, involving multiple hospitals and cloud servers, have been developed for EHR analytics [28]. Moreover, FL has been investigated for predicting hospitalizations in cardiac patients [29]. FL has also been widely studied in medical imaging applications, such as prostate cancer detection, brain tumor segmentation [30], and MRI analysis for Alzheimer and Parkinson diseases. Moreover, FL-based approaches have been proposed for detecting coronavirus infections [31,32].

Despite the promise of FL in medical applications, its practical application still faces the following challenges: a central concern is its dependence on a centralized server for model coordination and parameter aggregation. However, the central server can be a source of a single point of failure and remains susceptible to man-in-the-middle attacks [33,34]. Moreover, as an increasing number of local devices simultaneously transmit model parameters to the central server, it places significant pressure on server bandwidth and scalability, thereby elevating the risk of network congestion [8,35,36]. Besides, although FL avoids direct sharing of raw data, its training process may still

XSL•FO
RenderX

inadvertently expose sensitive information through model parameters. For instance, via gradient inversion attacks [37], adversaries can reconstruct sensitive training data from shared model parameters. The risk is further heightened by the presence of malicious participants, who may compromise the integrity of FL by uploading falsified or adversarial local training updates, such as poisoning attacks [38]. These attacks degrade the reliability of the global model, reducing its accuracy and overall utility. Additionally, the lack of a robust incentive mechanism poses a practical barrier to adoption. FL implicitly assumes that all participants willingly contribute data and computational resources without direct compensation. However, this assumption is difficult to uphold in real-world scenarios. The absence of a fair and transparent incentive mechanism may diminish participants' motivation to contribute high-quality updates, ultimately degrading system performance [39-41]. Moreover, some participants may receive rewards without actively contributing data, leading to unfair financial compensation.

When FL is applied to the medical domain, heterogeneity across systems, data, and distributions presents a fundamental challenge to model reliability and practical deployment. System heterogeneity arises from disparities in computing power, memory, network stability, and energy availability among participating devices [42]. This is particularly critical in resource-constrained environments, where limited hardware capabilities and intermittent connectivity can hinder local training and delay model updates. Data heterogeneity further complicates collaboration, as medical institutions often store data in different formats and standards, leading to integration difficulties [43,44]. Variations in data quality—such as incomplete records or inconsistent annotations—and significant differences in dataset sizes across institutions can distort the learning process and reduce model performance. Most critically, distribution heterogeneity, or the non-independent and identically distributed (non-IID) data problem, undermines generalizability. Institutions serve distinct patient populations, resulting in divergent feature and label distributions; for example, a model trained predominantly on "healthy" samples may struggle with accurate predictions when exposed to datasets with a higher prevalence of "diseased" samples, reducing its generalizability.

These advantages and limitations together underscore the urgent need for complementary technologies to enhance trust, security, and coordination in FL workflows. Blockchain, with its decentralized, tamper-resistant, and auditable infrastructure, offers a promising solution to many of these pain points. In the following section, we explore how blockchain can be integrated with FL to build more robust, transparent, and privacy-preserving frameworks for medical applications.

Blockchain in Health Care

Introduced in 2008 as the foundational technology behind the Bitcoin system, it was originally designed to solve trust-related issues in digital currency transactions [45]. Essentially, it functions as a distributed digital ledger, where transaction records are maintained and shared across all participants through a peer-to-peer network. Unlike traditional centralized systems, blockchain eliminates reliance on a trusted third party, ensuring decentralized data storage [46].

The implementation of blockchain technology relies on multiple technical layers and core components. To fully understand the principles of blockchain, it is essential to analyze its specific components. The fundamental unit of blockchain is the "block," comprising two primary components: the block header and the block body, which is shown in Figure 3. The block header contains metadata, including the hash value of the preceding block (which links the current block to its predecessor), timestamps, random numbers, and block version numbers. This metadata establishes the linkage between blocks. The block body holds smart contracts and actual data, including transaction records. These blocks are cryptographically linked in chronological order, forming an immutable chain structure. Each block is closely linked to its predecessor via a hash value, ensuring that any tampering with a single block will alter the hash values of all subsequent blocks. This change is detectable and will be rejected by the entire network, thereby preserving data authenticity and integrity.

Figure 3. A schematic diagram of the blockchain structure.

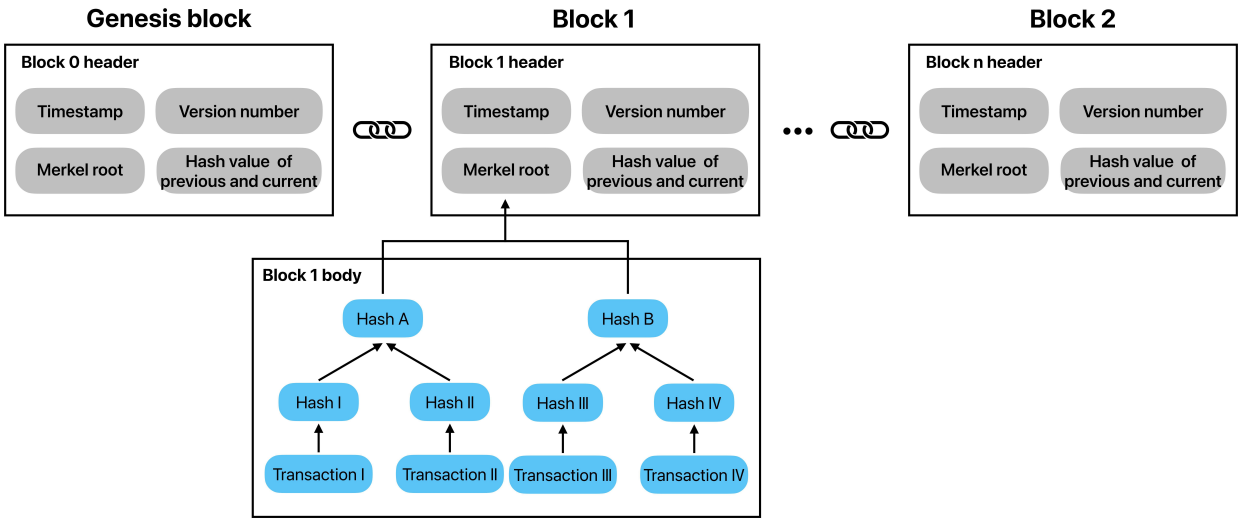

XSL•FO
RenderX

Furthermore, to fully understand the potential of blockchain in such applications, it is essential to examine its underlying architecture and systematically analyze its functional layers (shown in Figure 4).

Figure 4. The system architecture of blockchain. P2P: Peer to Peer; PBFT: Practical Byzantine Fault Tolerance; PoQ: Proof of Quality; PoS: Proof of Stake; PoW: Proof of Work; SCM: Supply Chain Management.

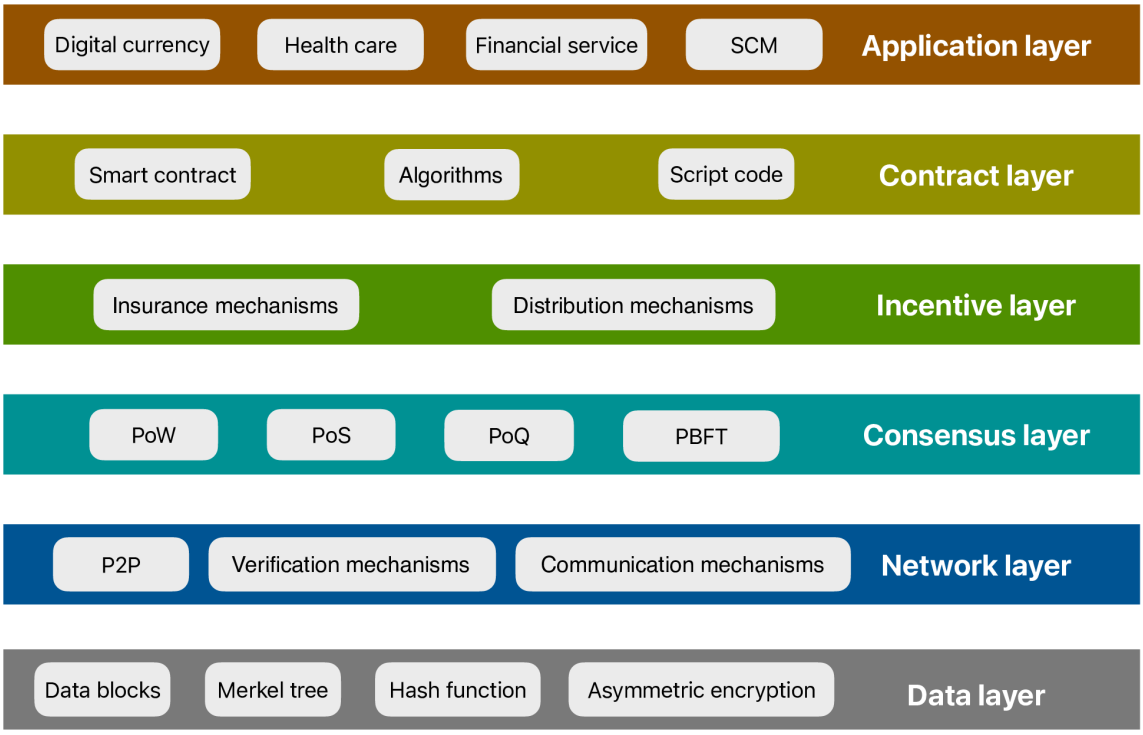

Fundamentally, blockchain systems are structured into 6 tightly interwoven layers, each addressing specific functional needs. At the core lies the data layer, which serves as the foundation for storing and organizing transaction records. The data layer organizes and secures transaction records through cryptographic techniques. This layer not only guarantees data integrity and immutability but also significantly reduces storage overhead, which is critical for handling the growing volume of medical data [47]. Above this, the network layer enables efficient and reliable information exchange across decentralized nodes via peer-to-peer communication protocols [47]. In the context of high-stakes environments like health care, securing consensus across distributed and often mutually untrusting participants is the responsibility of the consensus layer. By establishing agreement on the validity and chronological order of transactions, consensus mechanisms prevent malicious interference and ensure the reliability of shared records [48,49]. The widely adopted algorithms include Proof of Work (PoW), Proof of Stake (PoS), and Practical Byzantine Fault Tolerance (PBFT) [16,50,51]. To sustain long-term network participation, the incentive layer introduces economic mechanisms that reward nodes for validating transactions and generating new blocks. This is especially important in medical applications, where incentivized behavior facilitates the uploading of more high-quality medical data, while the authenticity and availability of data directly affects patient outcomes. Building upon these foundations, the contract layer operationalizes complex protocols through smart contracts, which autonomously execute predefined rules and transactions without the need for third-party intermediaries [52]. In the health care sector, this translates into streamlined workflows for processes such as insurance claims, consent management, and secure data access, offering enhanced transparency, reduced administrative overhead, and minimized human error. Finally, the application layer serves as the interface between users and the blockchain, bridging technical functionality with real-world applications. Across industries, this layer has fueled innovations in financial services, supply chain management, and beyond [53,54]. In health care, it enables secure sharing of EHRs, protecting patient privacy, and supporting collaborative research efforts across institutions [55]. By facilitating trusted interactions in environments where data sensitivity and security are paramount, blockchain's multilayered architecture offers an indispensable foundation for integrating advanced paradigms such as FL, thereby unlocking new possibilities for privacy-preserving, distributed medical intelligence.

Through the analysis of the structure and underlying architecture of blockchain, we can better understand that blockchain has unique characteristics that distinguish it from traditional systems, including decentralization, immutability, transparency, traceability, security, anonymity, and high availability [8,13].

These characteristics are particularly valuable in the field of health care. For instance, decentralization eliminates the inherent single point of failure in centralized medical record systems, thereby enhancing the system's resilience and reducing its

XSL•FO
RenderX

vulnerability to cyber attacks. This feature is particularly suitable for scenarios with high trust requirements, such as medical record management and supply chain supervision. Immutability and traceability not only ensure that clinical data, diagnostic results, and patient consent records cannot be altered or deleted, guaranteeing the auditing of medical data, but also are crucial for tracking the source of drugs and ensuring the integrity of the supply chain, thereby reducing the risk of counterfeit drugs. Furthermore, blockchain offers a transparent and privacy-protecting environment that enables authorized clinicians, researchers, and insurance companies to verify data sources without exposing sensitive patient information, thereby enhancing accountability in medical research and drug development. The security of blockchain stems from smart contracts, consensus mechanisms, etc, maintaining transaction integrity and preventing malicious nodes from arbitrarily altering the block addition process [56,57].

As illustrated in Multimedia Appendix 3, a comparison of the technical characteristics of blockchain and FL is presented. Simultaneously, given these characteristics, blockchain facilitates trust establishment among untrusted participants in wireless networks [58,59]. Consequently, blockchain demonstrates great potential across various domains, including cryptocurrencies, health care, and the Internet of Things (IoT) [45,60,61]. In health care, blockchain technology enables the storage and verification of IoT data within patients' electronic medical records (EMRs), clinical trials, and sensors, granting patients control over their own medical data [62-64]. During AI-driven sample learning, medical data from various institutions—including x-rays, CT scans, MRI reports, and pathological examinations—are securely stored on the blockchain. Predefined entity and event annotation platforms facilitate data labeling within each institution, followed by model training on in-hospital servers [65,66]. Kordestain et al [67] proposed HapiChain, a telemedicine platform built on a patient-centered blockchain infrastructure, ensuring the security of remote consultations between patients and doctors. In addition, decentralized blockchain solutions, such as Drug-ledger and Med-ledger, have been proposed to enhance traceability and security in the pharmaceutical supply chain [68,69].

Given blockchain's significant advantages in decentralization, privacy protection, data immutability, incentives, and automation, its attributes fit well with FL's requirements for secure data sharing and distributed modeling. Blockchain can serve as a secure and reliable collaborative infrastructure for FL, addressing challenges such as trust deficiency, data integrity, and transparency [4,22,70]. Conversely, the decentralized data processing mechanism of FL can compensate for blockchain's limitations in scalability and computational efficiency. Therefore, integrating blockchain with FL has the potential not only to address their respective challenges but also to unlock new application scenarios and possibilities. The following section examines the necessity and feasibility of integrating blockchain with FL.

Integration of Blockchain and Federated Learning in Health Care

Overview

BCFL has emerged as a promising solution in health care, fostering a balance between privacy protection and data collaboration while unlocking new opportunities for data-driven health care innovation [71,72]. The following section analyzes the complementary strengths and performance of integrating blockchain with FL, illustrating how blockchain mitigates the limitations of FL (as illustrated in Multimedia Appendix 3) and how FL benefits blockchain.

Blockchain Empowers Federated Learning in Health Care

Decentralization Mitigates Single Points of Failure and Scalability Bottlenecks

The traditional architecture of FL primarily depends on a central server to manage and coordinate participants. This architecture is susceptible to single points of failure and can also result in bandwidth and computational resource bottlenecks as the number of clients increases [70,73,74]. In contrast, blockchain's decentralized architecture eliminates reliance on a central server by leveraging a distributed network, allowing automated data collaboration across multiple nodes and effectively mitigating the risk of single points of failure [13]. In this context, temporary aggregators are selected based on blockchain's consensus mechanisms (eg, PoW or PoS) [8,36]. Moreover, blockchain's Byzantine fault tolerance, which enables the dynamic management of unreliable nodes through consensus mechanisms, further enhances the stability of FL in large-scale distributed environments [75].

Incentives to Enhance Participant Motivation

Traditional FL systems often lack effective incentives, especially in environments where resources are unevenly distributed. High-performing participants may struggle to sustain long-term contributions due to the absence of direct benefits [76,77]. Blockchain's built-in incentive mechanisms address this challenge by rewarding data contributors, validators, and maintainers through tokens or other financial instruments, introducing an economic driver for FL [78]. This incentive model encourages the contribution of high-quality data while discouraging unreliable participation, ultimately improving the overall performance and stability of FL models. For instance, Weng et al [79] proposed an incentive mechanism designed to promote collaboration in training deep learning models. The mechanism introduces two key concepts: compatibility and activity. Compatibility ensures that each participant receives optimized rewards based on their contribution, while activity incentivizes participants to update the local model and aggregate the global model actively. Upon each global model update, rewards are distributed to local devices and miners based on individual contributions. Similarly, Kang et al [77] introduced a reputation-based incentive model to measure client trustworthiness. By leveraging blockchain's immutability, the system ensures distributed reputation management and evaluates participants based on model quality and computational contributions.

XSL•FO
RenderX

Privacy Protection and Attack Resistance Enhancement

Although FL protects raw data privacy, it remains vulnerable to adversarial threats such as poisoning attacks and Byzantine attacks, which can mislead model training and hinder convergence [36,80]. Blockchain strengthens FL security through its immutability and tamper-proof nature. Its authentication mechanisms detect and exclude malicious nodes, ensuring that only authorized participants can access FL data, thereby enhancing privacy protection. Moreover, blockchain's cryptographic techniques and anonymity mechanisms reduce the risks of background knowledge attacks and collusion attacks [78]. Additionally, its consensus mechanism ensures data consistency across all nodes while using sophisticated algorithms to prevent malicious nodes from compromising the network [81,82]. For example, in medical data collaboration, blockchain records the training processes and contributions of each participating hospital, ensuring both data integrity and privacy while mitigating risks of data contamination and adversarial attacks. Shayan et al [83] proposed a multi-Krum consensus mechanism to counter poisoning attacks by electing a validation peer committee that filters out malicious model updates. Similarly, Chen et al [84] used a blockchain-based validation voting mechanism, where nodes vote on model update validity and remove malicious devices based on consensus results.

Transparency and Auditability

Blockchain's transparency and auditability effectively address the challenges of trust deficits and compliance difficulties in FL. By leveraging blockchain's transparent data-sharing mechanisms, FL participants can verify the source and integrity of health care data and model updates in real time, ensuring fairness and reliability in contributions [72]. Additionally, blockchain's immutable records establish an accountability framework for FL [33,85]. These records facilitate anomaly detection and responsibility attribution throughout the model training process, strengthening compliance and governance mechanisms. This transparent and auditable nature not only fosters trust in collaborative learning but also provides a technological foundation for regulatory compliance.

Automated Management With Smart Contracts

Smart contracts enable automated execution of key processes in FL, including model update sharing, model update validation, and global model aggregation. By enforcing predefined rules, smart contracts eliminate human intervention, ensuring an unbiased and tamper-proof process. Moreover, they can dynamically allocate resources and rewards through conditional triggering mechanisms—such as when a model reaches an expected accuracy or when a node successfully completes a specific task [86,87]. This automation enhances the autonomy and reliability of FL while ensuring transparency and fairness through open code logic. By reducing administrative overhead and mitigating trust concerns, smart contracts introduce a novel and efficient approach for managing decentralized collaborative learning [33].

How Federated Learning Can Benefit Blockchain

Enhancing Blockchain Consensus Efficiency

The blockchain consensus mechanism, while ensuring network security and data consistency, is often associated with substantial computational costs and energy consumption, a challenge that is particularly serious in the PoW mechanism [8,88]. PoW relies on miners solving complex hashing operations to compete for block generation, necessitating the continuous operation of high-performance computing hardware, which in turn results in substantial global energy consumption. Studies indicate that the annual energy consumption of the Bitcoin network is comparable to that of a small- to medium-sized country. This highly inefficient competition leads to an enormous waste of computational resources—only the first miner to discover a valid hash can package the transaction and claim the reward, rendering all other computational efforts futile. Furthermore, the excessive energy consumption of PoW constrains blockchain's sustainable adoption in critical domains such as health care data management and edge computing, necessitating the development of more energy-efficient consensus optimization strategies. In this context, FL offers a novel approach to optimizing blockchain consensus mechanisms. Integrating the blockchain consensus process with FL allows miners to contribute to model training while competing for block validation, effectively repurposing computational resources that would otherwise be wasted. This approach not only reduces energy consumption but also enhances computational resource efficiency, rendering the consensus process more practically valuable [89].

Facilitating Cross-Chain Data Collaboration

As blockchain applications continue to expand, the demand for data exchange across different blockchains has grown, making cross-chain technology a crucial solution to addressing "data silos" across various domains. FL and its variants, such as federated transfer learning, can establish a unified model collaboration framework across different blockchain networks, enabling privacy-preserving data sharing and joint modeling. By maintaining a shared ML model, disparate blockchains can collaborate while preserving autonomy and privacy, thereby facilitating cross-chain applications in finance, health care, and other sectors [90].

Enhancing Blockchain Scalability

Blockchain faces storage and computational bottlenecks when handling large-scale data. FL, by adopting a local training model that eliminates the need to upload raw data to the blockchain, significantly reduces on-chain storage demands. Additionally, FL alleviates blockchain's computational burden by distributing processing tasks among participating nodes, thereby providing a scalable foundation for large-scale collaboration.

Architectural Frameworks for Integrating Blockchain and Federated Learning

A BCFL typically adopts one of three architectural paradigms: fully coupled, flexibly coupled, and loosely coupled architectures [41,85,91]. These architectures differ in terms of the degree of coupling between blockchain nodes and FL clients,

XSL•FO
RenderX

each offering unique characteristics in function allocation, resource usage, and system structure.

Fully Coupled BCFL

The fully coupled architecture represents a highly integrated design, wherein FL clients simultaneously function as blockchain nodes, assuming dual roles. Consequently, each node is responsible for local model training, update validation, global model aggregation, and new block generation. These tasks are executed on a single node, fostering a fully decentralized collaborative model [36,74,79,92].

Global model aggregation can be carried out either by selected nodes or collaboratively by all nodes, depending on the network's design strategy. Moreover, the blockchain's distributed ledger not only records local model updates but also stores global models and other relevant information generated during training, ensuring data integrity and traceability.

Flexibly Coupled BCFL

The flexibly coupled architecture achieves higher design flexibility by separating FL clients from blockchain nodes. In this architecture, FL clients primarily handle local data collection and model training, whereas blockchain nodes are responsible for validating model updates, storing the global model, and maintaining the ledger [2,40,93]. The blockchain can aggregate global models via selected nodes, which typically possess superior computing resources and reliability, thereby reducing resource consumption and enhancing system efficiency. Alternatively, aggregation can be performed collaboratively by all nodes, achieving full decentralization and mitigating the risk of a single point of failure.

This architecture significantly lowers the resource requirements for FL clients, allowing them to function in different network environments while preserving blockchain's inherent advantages in data security and transparency. Due to its high adaptability, this architecture has become a preferred choice for large-scale distributed systems, such as health care data sharing and cross-organizational collaboration.

Loosely Coupled BCFL

The loosely coupled architecture further weakens the coupling between blockchain nodes and FL clients by optimizing functional allocation. FL clients primarily perform local model training and upload updates to the blockchain for validation, whereas the blockchain handles authentication, model update validation, and participant reputation management.

In this architecture, the blockchain does not store model updates but instead records only reputation-related data. A reputation mechanism is implemented as a key criterion for assessing participant reliability, thereby incentivizing them to contribute high-quality data and updates [77,94,95]. This design enhances system scalability by alleviating storage pressure on the ledger while ensuring the trustworthiness of participant behavior.

Workflow in BCFL

Overview

In BCFL systems, the flexibly coupled architecture has emerged as the predominant choice for real-world applications due to its optimal balance of efficiency and adaptability. By separating FL clients from blockchain nodes, this architecture allows them to operate on different networks and devices, thereby reducing system communication overhead and latency. Additionally, it alleviates the computational and storage burden on client devices while preserving key advantages such as data privacy protection and blockchain-based verification, ultimately achieving an optimal balance between efficiency and privacy. Leveraging these advantages, the flexibly coupled architecture has demonstrated significant potential in practical applications, including medical data sharing and cross-organizational collaboration.

As illustrated in Figure 5, the following section focuses on the specific workflows of mainstream BCFL frameworks, analyzing their distinct advantages in practical applications.

XSL•FO
RenderX

Figure 5. Flexibly coupled blockchain-based federated learning architecture and workflow. gRPC API: Google Remote Procedure Calls – Application Programming Interface; REST-API: Representational State Transfer – Application Programming Interface.

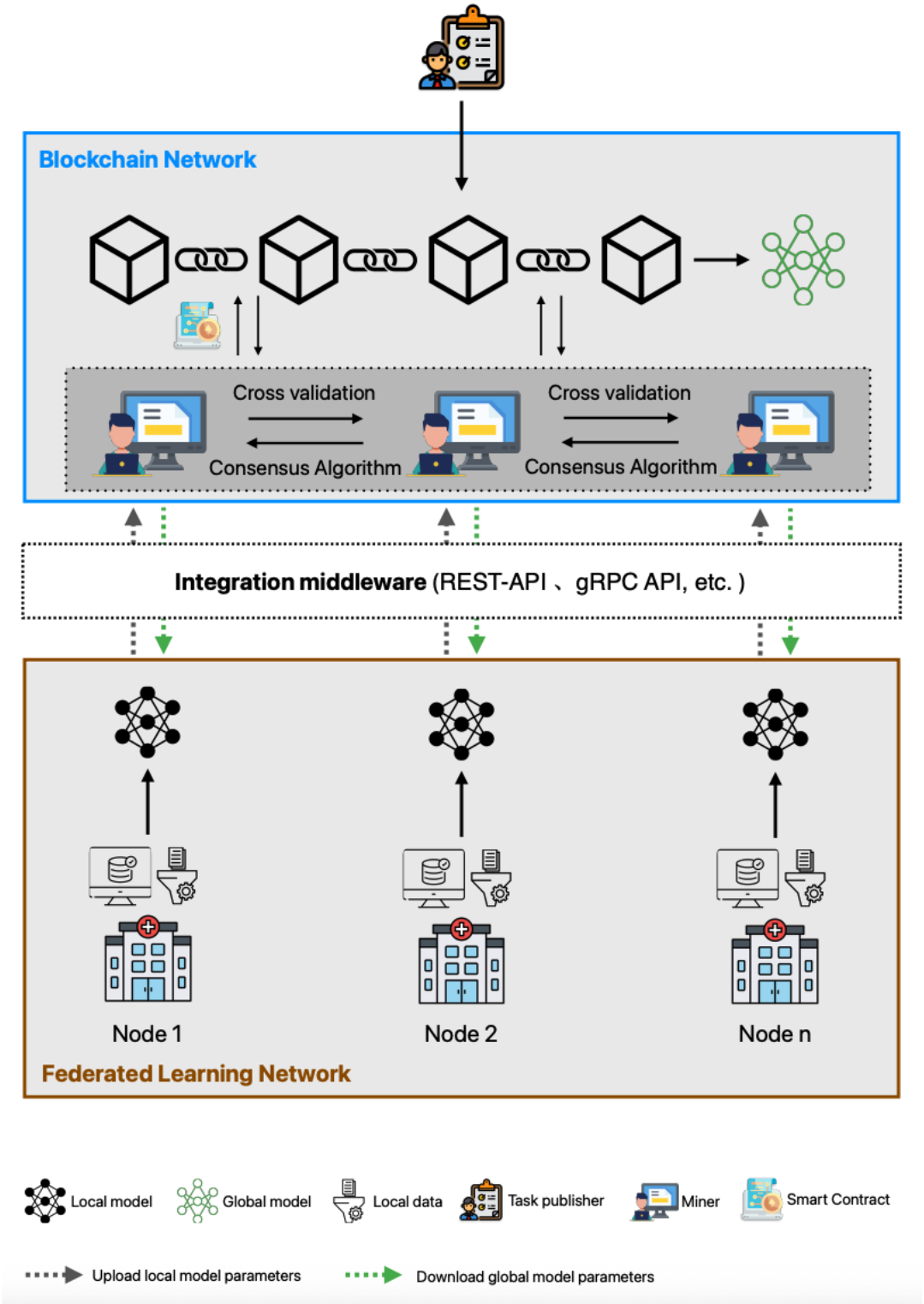

Task Release

Task initiators release FL tasks and requirements on the blockchain, specifying details such as data volume and type, hardware specifications, and the number of training rounds. Leveraging blockchain's transparency and decentralization, this process ensures fair and open task distribution while fostering participant trust.

Local Model Training and Update Transmission

Each FL client downloads the initial global model from the blockchain, after which it preprocesses local data, extracts features, and uses this data for model training, subsequently generating local model updates. These updates are then transmitted to the blockchain network in encrypted form.

XSL•FO
RenderX

It is important to note that, in the flexibly coupled architecture, FL clients and blockchain nodes operate within different networks and systems, each with clearly defined responsibilities. Therefore, this architecture heavily relies on integrated middleware, which serves as a communication bridge and coordinator between the two components. In a research study, Lamken et al [96] used REST-API (Representational State Transfer – Application Programming Interface) for communication with the Hyperledger Fabric blockchain, facilitating the recording and incentivization of gradient uploads. Additionally, the Remote Procedure Calls (RPC) protocol developed by Google, known as the gRPC API, facilitates data exchange between FL clients and the Ethereum blockchain network [33,96].

Blockchain Node Verification Update

Blockchain nodes (ie, miners) verify the uploaded model updates using a predefined validation mechanism. Concurrently, miners exchange their validated local model updates with each other. A consensus algorithm guarantees that only validated updates contribute to the global model aggregation.

Global Model Aggregation

Subsequently, the blockchain selects an interim leader among its nodes through a consensus mechanism. The selected node(s) then collect verified model updates and aggregate them to construct the global model [36]. The flexibly coupled architecture enables this process to be executed by selected nodes or collectively by all nodes, thereby opening up the possibility of optimizing efficiency across various scenarios.

New Block Generation and Model Storage

Validated model updates and global models are packaged by selected blockchain nodes to generate new blocks. Upon adding the block header information, the legitimacy of the block is verified through a consensus mechanism among the nodes.

Distributed Ledger Update

The newly generated blocks are broadcast across the entire network, and all blockchain nodes update their local ledgers after verification. This process ensures the transparency and traceability of the global model and its associated information throughout the network.

Reward Distribution and Incentives

The system allocates rewards, such as cryptocurrency or reputation scores, based on client performance. This incentive mechanism not only motivates participants to contribute high-quality updates but also deters malicious behavior, thereby enhancing the accuracy and reliability of the model.

Global Model Download

After the training is completed, all participating clients can download a newly generated block containing the updated global model parameters from the blockchain. Clients can then independently decide whether to participate in the next training round based on their specific needs. This mechanism enhances both system flexibility and participant autonomy.

As illustrated in Table 1, a comparison of BCFL integration architectures is presented. In future practical applications, the selection of a specific architecture must be carefully evaluated based on scenario requirements, resource constraints, and design objectives to achieve optimal collaboration and technical performance.

XSL•FO
RenderX

Table . Comparison of blockchain-based federated learning integration architectures.

Architecture type	Characteristics	Advantages	Disadvantages	Applicable scenarios
Fully coupled BCFL[a]	• High integration: FL[b] clients and blockchain nodes are fully merged • Fully decentralized: All nodes work together through a consensus mechanism	• High transparency: All transactions and model updates are recorded on the blockchain • Strong security: Resistant to single-point failures and man-in-the-middle attacks • Strict control: Highly controlled over data and models	• High resource demand: Requires significant computational and storage resources • High network complexity: All nodes participate in the consensus mechanism, and the network complexity is high • Intensive coordination: Frequent internode communication is required	• Large-scale distributed environments: Suitable for large medical institutions and research centers • Strict control and security requirements: Scenarios involving the sharing and analysis of sensitive medical data [97,98]
Flexibly coupled BCFL	• Functional separation: FL clients operate independently from blockchain nodes • Computational offloading: Model aggregation occurs at selected nodes	• Enhanced efficiency: Optimized allocation of computing and storage resources • Greater flexibility: Can be adapted to different application scenarios • Improved scalability: Supports large-scale data sharing and collaborative learning	• Complex coordination: The responsibilities of clients and nodes are separated, and complex coordination and management mechanisms are required • Centralization risks: Use of a centralized aggregator may introduce a single point of failure • Node selection challenges: Issues such as node selection criteria and fairness are involved	• Dynamic collaboration settings: Suitable for cross-institutional medical data sharing, IoMT[c] device management [40,99]
Loosely coupled BCFL	• Minimal integration: FL clients and blockchain nodes operate independently • Lightweight blockchain: Primarily used for identity authentication and reputation management	• Reduced overhead: Reduce the operating cost of blockchain and improve system performance • Enhanced privacy: Reduce on-chain storage pressure and improve scalability • Optimized incentives: Reputation-based mechanisms encourage high-quality contributions	• Lower decentralization: May still rely on trusted central nodes for model aggregation • Data integrity risks: Blockchain does not store model updates	• Resource-constrained environments: Suitable for wearable medical devices and real-time health monitoring • Small-scale institutions: Ideal for personal mobile health applications and smaller clinics [100]

[a]BCFL: blockchain-based federated learning.

[b]FL: federated learning.

[c]IoMT: Internet of Medical Things.

BCFL in Medicine

Overview

As the demand for data-driven technologies in health care continues to grow, the BCFL framework presents significant potential due to its advantages in privacy preservation, data security, and collaborative efficiency. BCFL facilitates cross-organizational data sharing and collaborative analytics, optimizing personalized health care solutions while driving advancements in areas such as telemedicine, IoMT, and public health monitoring, which are shown in Figure 6. In the following section, we will discuss the various applications of BCFL in the medical field and analyze its key role and potential value in addressing real-world challenges.

XSL•FO
RenderX

Figure 6. Blockchain-based federated learning framework for different domains in health care. EMR: electronic medical record.

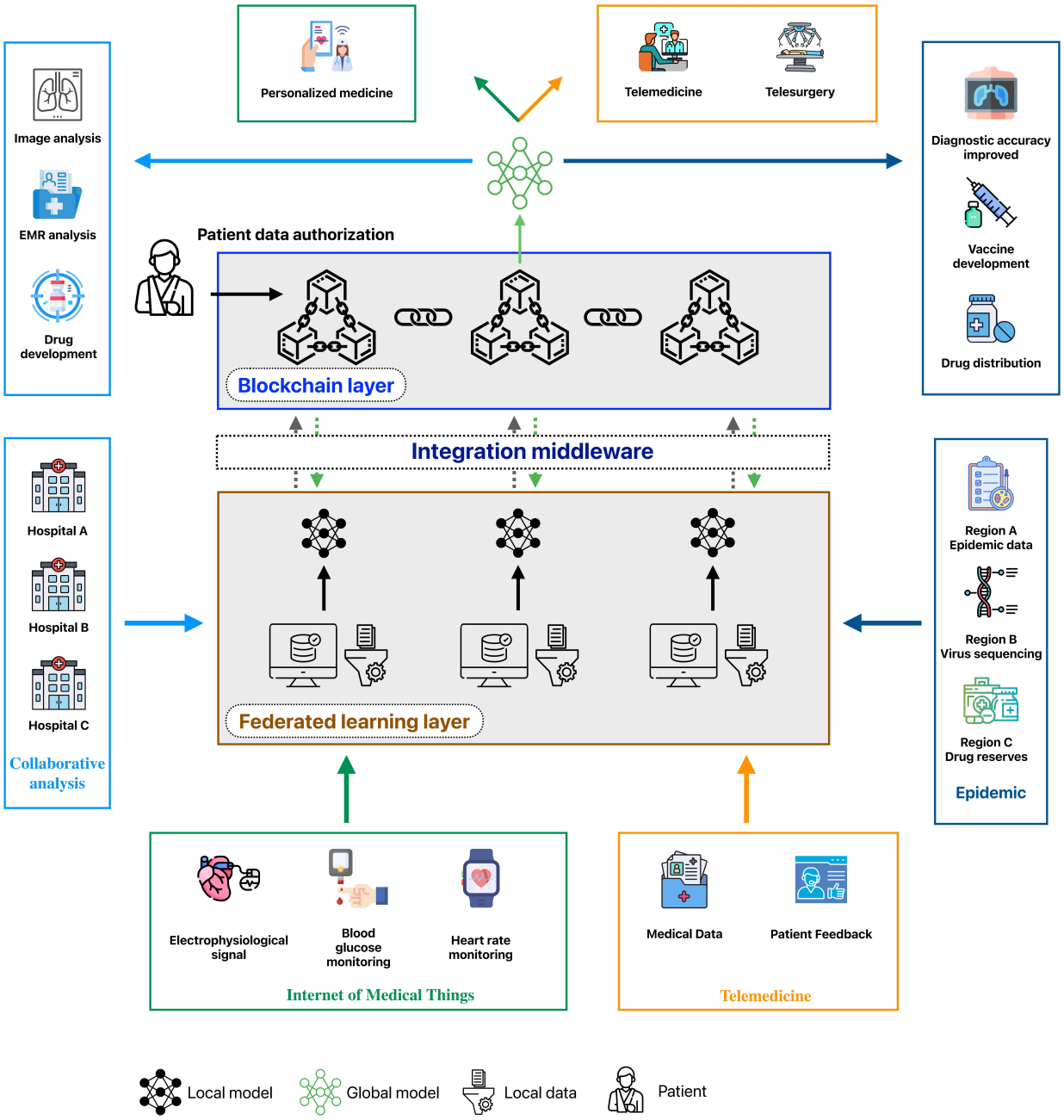

Cross-Institutional Medical Data Sharing and Collaborative Analysis

In modern health care, data serves as a crucial resource for driving innovation and enhancing treatment efficacy. However, data sharing among health care institutions is hindered by concerns over privacy, data security, and regulatory compliance. The integration of blockchain and FL offers an innovative solution for cross-institutional health care data sharing and collaborative analysis. While numerous studies have demonstrated the feasibility of BCFL in various medical domains, the strength of evidence supporting these applications varies considerably, and critical challenges remain.

Several studies have investigated BCFL in the context of chronic disease management, particularly diabetes prediction. Hasan et al [101] developed a blockchain-FL framework that reported a 15% improvement in predictive performance across multiple metrics. Although these results are encouraging, the framework relied primarily on public diabetes datasets with limited diversity, raising questions about its generalizability to heterogeneous real-world populations. Similarly, Moulahi et al [102] evaluated a BCFL model on the Pima Indians Diabetes dataset, achieving a multilayer perceptron accuracy rate of 97.11% and an average FL accuracy rate of 93.95% while protecting privacy. Yet, the reliance on small, well-characterized datasets constrains the robustness of the findings. Taken together, these studies suggest that BCFL holds promise for chronic disease prediction, but the supporting evidence remains preliminary, and large-scale multi-institutional validation is still lacking.

XSL•FO
RenderX

In the realm of the IoMT, Ramani et al [103] introduced the ODMSM-FL (Optimized Data Management and Secured Federated Learning) approach to address secure data storage and exchange using EHR datasets from HealthData.gov. This research report presents a set of numerical results on key performance indicators: transaction throughput (102.75 Kbps), data retrieval delay (64.02 ms), security (88.97%), and accuracy (86.32%). The research results highlight the great potential of ODMSM-FL in effectively addressing the urgent data management and security issues in IoMT. However, the system was evaluated under controlled experimental conditions rather than real-world clinical settings, limiting its immediate applicability. By contrast, research in medical imaging tasks, such as brain tumor segmentation, has placed greater emphasis on model accuracy and privacy preservation. For example, Kumar et al [104] proposed a permissioned blockchain-based federated framework with quality-aware model aggregation, achieving improved segmentation metrics on the BraTS 2020 dataset. Specifically, compared with the baseline method, our approach increased the Dice similarity coefficient of enhanced tumors by 1.99% and reduced the Hausdorff distance of the overall tumor by 19.08%. Although the study demonstrated methodological innovation, it still relied on benchmark imaging datasets rather than prospective clinical data, which restricts the strength of evidence regarding its clinical translatability.

Other investigations have targeted specific diagnostic applications. Heidari et al [105] designed the FBCLC-Rad (Federated Learning–Enabled Blockchain CapsNets Lung Cancer Radiologist) framework for lung cancer detection, achieving near-perfect accuracy on nodule identification tasks. This technology achieved an accuracy rate of 99.69% with the lowest classification error. While technically impressive, results derived from controlled datasets may not fully reflect the complexity of real-world diagnostic workflows. Liang et al [106] extended BCFL applications to clinical trials, where blockchain ensured data authenticity and traceability, and FL supported participant screening across organizations. This study represents an important step toward integrating BCFL into the clinical research pipeline, but it remains largely conceptual, with limited empirical validation in actual trial environments.

A growing body of work has also highlighted the integration of BCFL with EMRs to facilitate precision medicine [99,107-110]. This approach has demonstrated significant effectiveness in enhancing diagnostic accuracy, optimizing treatment planning, identifying patient subgroups for clinical trials, and accelerating the development of novel therapeutics. Within the paradigm of precision medicine, such a framework facilitates a transition from the traditional “one-size-fits-all” treatment model to a more personalized and adaptive intervention strategy. However, despite their conceptual appeal, most studies are limited to prototype frameworks or simulations and have not yet undergone prospective evaluation in clinical practice. As such, the current evidence supporting BCFL in EMR-based precision medicine remains promising but immature.

Overall, existing literature demonstrates the conceptual feasibility and technical potential of BCFL for cross-institutional health care data sharing and analysis. Nevertheless, the evidence base is uneven: studies using small public datasets provide only preliminary support, while those addressing more complex tasks such as imaging or clinical trials often lack real-world validation.

Internet of Medical Things

The IoMT is an advanced technological ecosystem that integrates internet technology with medical devices, enabling real-time data collection, exchange, and analysis to enhance clinical decision-making, disease prevention, and patient care. With the rapid advancement of the IoMT, the traditional hospital-centric model has evolved into a patient-centered health care system driven by comprehensive clinical analysis. The widespread adoption of IoMT enables individuals to conveniently monitor their health at home, thereby streamlining diagnosis and treatment while allowing patients to enjoy more efficient, personalized health care. However, despite its growing adoption, IoMT in health care is also facing significant challenges. One of the primary concerns is data privacy and security. IoMT devices collect and transmit vast amounts of sensitive data, including patient identities, insurance details, and payment information. Once these data are accessed by malicious individuals, it could lead to serious consequences. Moreover, the absence of standardized security protocols among IoMT devices exacerbates the risks of data leakage and device manipulation. Consequently, device manufacturers and health care institutions face immense pressure to ensure data privacy and regulatory compliance. Additionally, IoMT devices frequently encounter challenges such as high computational complexity, elevated costs, and communication delays due to resource limitations.

To address these limitations, several studies have explored the integration of blockchain and FL in IoMT. For instance, Rahman et al [111] proposed a lightweight hybrid FL framework that leverages blockchain to secure health data provenance and uses smart contracts to coordinate model training and trust management. While this framework demonstrates theoretical scalability and robust traceability, its evaluation was primarily conducted in simulated settings, raising concerns about its applicability in heterogeneous and large-scale real-world health care environments.

Muazu et al [112] combined BCFL with edge computing to improve resource allocation, reduce computational costs, and enhance IoMT data security. By offloading intensive computations to edge nodes, the study reported reductions in latency and energy consumption. Meanwhile, the performance of the proposed model offers a higher precision of 83% and an accuracy rate of 78%. However, the framework largely relies on linear regression as the global learning model, which—although interpretable and useful for basic clinical predictions—may not adequately capture the complexity of real-world multimodal medical data. Compared with Rahman et al [111], this work provides stronger performance evidence in terms of latency and efficiency but weaker generalizability for complex clinical prediction tasks.

Dhasaratha et al [113] extended BCFL by incorporating reinforcement learning and distributed computing to improve risk factor monitoring and COVID-19 patient prediction. The dynamic optimization enabled by reinforcement learning is a

XSL•FO
RenderX

notable strength, allowing adaptive performance improvements in evolving environments. Compared with the previous two documents, this approach offers methodological novelty but lacks equally rigorous performance benchmarking across standard IoMT metrics.

To tackle fraud detection and scheduling issues, Lakhan et al [114] introduced the FL-BETS (Federated Learning–Based Blockchain-Enabled Task Scheduling) framework, which integrates BCFL with dynamic heuristics for task scheduling across fog and cloud nodes. The experimental results show that the framework, from the initial 60:80 fraud delay ratio to a 10:10 ratio, demonstrates better performance in energy-delay trade-offs and antifraud behavior. Yet, the framework emphasizes technical efficiency rather than clinical utility, and its reliance on hard and soft scheduling constraints may limit adaptability in unpredictable medical environments. Compared with Dhasaratha et al [113], which focuses on patient-specific outcomes, this work provides stronger technical validation but weaker clinical alignment.

In the context of wearable IoMT devices, Baucas et al [115] designed a system that integrates FL with a private blockchain in a fog computing architecture to enhance privacy and adaptability. Their framework demonstrated efficiency in resource-constrained environments and produced accurate predictive models while safeguarding patient privacy. Unlike the literature above, this study directly validated its framework on wearable health care devices, thereby providing more immediate clinical relevance. However, the scalability of the approach for larger IoMT networks remains uncertain.

Overall, these studies collectively highlight the potential of BCFL in overcoming IoMT's inherent privacy, security, and performance limitations. Yet, their evidence strength varies significantly: some emphasize theoretical frameworks validated in simulations [111], while others demonstrate more robust experimental performance [112,114] or closer alignment with clinical practice [115].

Public Health Surveillance and Epidemiological Forecasting

The global outbreak of COVID-19 highlighted the limitations of existing surveillance infrastructures, particularly the inability to provide accurate, real-time epidemic monitoring. Traditional methods often struggle with the rapid spread and variability of epidemics, while stringent privacy requirements hinder effective collaboration across institutions. For instance, during global outbreaks such as COVID-19, the inability of national and regional health care organizations to efficiently integrate data has impeded comprehensive analyses of epidemic progression [116]. This phenomenon of data silos delays the formulation of precise response strategies and undermines the efficiency of vaccine distribution and health care resource allocation. Therefore, achieving efficient and secure data integration while preserving privacy has emerged as a critical challenge in public health. The BCFL framework not only integrates anonymized health data from diverse regions but also facilitates the efficient construction of predictive models for epidemic spread. This framework enables multiple health care organizations and research institutions to collaborate securely without compromising patient privacy, thereby providing robust data support for early epidemic detection, transmission trend analysis, and the formulation of intervention strategies. However, the strength of evidence supporting BCFL frameworks varies considerably across different studies, depending on data scale, validation methods, and implementation feasibility.

FedMedChain [117] represents an early attempt to address these challenges. By combining blockchain with FL and leveraging the IoMT, it enhances the trustworthiness of public health communication and mitigates risks associated with centralized data transmission. Its contribution lies in demonstrating that blockchain can ensure data transparency and tamper resistance while maintaining privacy. Nevertheless, FedMedChain was mainly verified through small-scale simulation experiments rather than real-world deployments, which limited the strength of the evidence and its direct clinical applicability.

In contrast, Kumar et al [118] explored a BCFL framework for processing heterogeneous CT images using capsule networks. This method achieves a high detection accuracy on the CC-19 dataset, and its research results include 98.68% specificity and 98% sensitivity. While the study demonstrates the feasibility of applying BCFL to medical imaging and highlights the benefits of privacy-preserving collaboration, the restricted dataset size and limited institutional diversity weaken its external validity. Compared with FedMedChain, this workplaces greater emphasis on model performance but provides weaker evidence regarding scalability and generalizability.

Durga and Poovammal [119] extended this direction by proposing the FLED-Block (Federated Learning–Ensembled Deep Learning Blockchain Model) framework, which integrates blockchain with FL for COVID-19 prediction using multisource heterogeneous CT datasets. This framework improves the classification accuracy by using capsule networks for feature extraction and extreme learning machines for efficient classification. The research results include an accuracy of 98.2%, a precision of 97.3%, and a recall rate of 96.5%. Importantly, it integrates blockchain to share model weights without the need to exchange raw data, thereby resolving privacy issues. Compared with Kumar's study [118], FLED-Block was supported by evidence from more hospitals, providing stronger validation. However, the latency of blockchain is regarded as a limitation, raising questions about its applicability in real-time clinical diagnosis. Therefore, although this framework demonstrates outstanding technical performance, its transformation in emergency health care settings remains uncertain.

Abdel-Basset et al [120] proposed the blockchain-based federated learning for pandemic diagnosis (BFLPD) framework, which stands out for its focus on system security and robustness in the context of smart cities. Unlike FedMedChain and FLED-Block, BFLPD combines more encryption technologies, including secure aggregation, homomorphic encryption (Cheon-Kim-Kim-Song scheme), and consensus mechanisms (PBFT), to mitigate malicious attacks and improve reliability. The classification accuracy of BFLPD reaches 95.14%, exceeding the benchmark set by the most advanced distributed models. In addition, this framework also demonstrated

XSL•FO
RenderX

significant precision and recall rates (95.26% and 95.77%, respectively) and a relatively high F_1-score (95.52%). Meanwhile, the authors incorporated heat map visualization, further enhancing its clinical application value. This framework provides stronger evidence than earlier works, as it addresses adversarial threats that are often overlooked in BCFL studies. Nevertheless, its reliance on complex cryptographic and consensus algorithms introduces implementation challenges, such as high computational overhead, which could hinder real-world deployment.

Telemedicine and Telesurgery

In recent years, telemedicine has experienced rapid advancements, especially in response to the global COVID-19 pandemic, which has significantly increased its role in modern health care systems. Telemedicine leverages modern information and communication technologies to facilitate medical information exchange across geographic boundaries, encompassing various applications such as remote diagnosis, remote consultation, remote treatment, and continuous health monitoring [5]. By providing on-demand, personalized health care services, telemedicine optimizes medical resource allocation, effectively addressing the challenge of unequal distribution of traditional health care resources and ensuring medical support for patients in remote or underserved areas. Despite its potential, telemedicine faces several critical challenges in practical implementation:

- Data Security and Privacy Risks: Most telemedicine systems rely on centralized cloud servers to store patient health data, making them vulnerable to single points of failure.
- Lack of Data Access Control Mechanisms: Many existing telemedicine platforms do not offer a robust data access control framework, meaning that once patient data are uploaded to the cloud, patients often lose ownership and control over their own health records.
- High Infrastructure and Computational Costs: Telemedicine demands substantial computational resources, high-speed communication networks, and specialized medical equipment, particularly for real-time diagnosis and treatment.

Building upon these advantages, recent studies have proposed generalized frameworks that integrate blockchain and FL to support secure and scalable telemedicine systems. For example, Hiwale et al [5] highlighted the importance of incorporating privacy-preserving technologies into BCFL, laying the groundwork for reliable, privacy-compliant telemedicine applications. Within such frameworks, blockchain's distributed ledger technology enables decentralized data storage, reducing the risks of single points of failure and data breaches, while ensuring transparency and traceability in data access. Simultaneously, FL enhances data privacy by enabling local model training, thereby minimizing the exposure of sensitive health information. Although valuable as a theoretical framework, this study provides limited experimental validation and thus represents weak evidence for clinical applicability. Gupta et al [121] enhanced trust between patients and providers by designing a smart contract system based on public blockchains. The framework allows patients to retain ownership and fine-grained control of health data, addressing a central shortcoming of conventional telemedicine platforms. Compared to Hiwale et al [5] conceptual work, Gupta et al [121] system offers a more concrete mechanism for authorization and data sharing. Nonetheless, its validation remains restricted to simulation environments, with no real-world deployment or clinical evaluation. As such, while it provides moderate evidence of feasibility, its generalizability remains uncertain.

With the rapid advancement of technology and the increasing improvement of medical demands, traditional telemedicine models are evolving beyond routine diagnosis and treatment. Among these advancements, telesurgery—a critical extension of telemedicine—is emerging as a transformative innovation. However, this technology imposes stringent requirements on real-time data synchronization, precise coordination of surgical equipment, and robust data security, introducing new challenges to the reliability of underlying technological infrastructures. For instance, Chaudjary et al [122] proposed a secure telesurgery system that integrates blockchain and FL with 6G communication networks and the Interplanetary File System protocol. This study demonstrated notable improvements in latency reduction, storage efficiency, and transmission reliability compared with traditional telesurgery systems. Unlike earlier works by Hiwale et al [5] and Gupta et al [121], Chaudhary et al [122] provided more systematic experimental results, suggesting stronger evidence of technical feasibility. However, these results were still derived from controlled simulations rather than real-world surgical environments, and issues such as blockchain latency and computational overhead remain unresolved. Therefore, while the study represents the strongest evidence among current works, its translation into clinical practice requires further validation.

In the past few years, BCFL's research in the health care field has shown significant growth. Multiple studies have confirmed the improvement of model performance in public datasets or experimental environments, such as enhancing the accuracy of disease prediction, strengthening image diagnostic capabilities, or improving edge device management.

However, when examined from the perspective of evidence, most of these achievements are still at the stage of simulation experiments, prototype systems, or preclinical validation. Research is usually based on controlled datasets or static data scenarios, and there is a significant gap between the model performance and the actual clinical diagnosis and treatment process. Importantly, current literature pays more attention to technical indicators (such as accuracy, Dice, and delay) rather than specific medical endpoints, such as changes in misdiagnosis rates, shortened treatment duration, or improved patient prognosis. Therefore, a direct chain of evidence has not yet been formed between the technical performance of BCFL and its actual medical value.

Furthermore, the BCFL architecture is not inherently compatible with the real deployment environment. The medical system features a complex governance structure, compliance requirements, and heterogeneous infrastructure. However, existing research often assumes node autonomy, network stability, or institutional equivalence, while neglecting key issues

XSL•FO
RenderX

such as data authorization, responsibility division, and system compatibility. Although blockchain consensus, smart contracts, and high-intensity encryption enhance security, they also bring about latency, energy consumption, and maintenance costs, which conflict with the real-time and reliability requirements of clinical practice. These mismatches have led to many frameworks performing well in experimental settings but being difficult to migrate to real medical scenarios.

Given the above limitations, relying solely on technical performance indicators cannot accurately reflect the maturity of BCFL in medical scenarios. To more systematically assess the application level of existing research, we adopted an evidence stratification strategy to categorize the existing literature from multiple dimensions such as architectural innovation, deployment environment, verification depth, clinical relevance, and potential risks. This stratification aims to reveal the gap between "technical performance" and "medical practice value," identify the most critical bottlenecks in the process from conceptual framework to clinical validation, and provide directional references for future research design. Table 2 summarizes the evidence stratification of typical BCFL studies in the health care field.

XSL•FO
RenderX

Table . Evidence stratification of blockchain-based federated learning studies in health care.

Reference	Application domain	BCFL[a] architecture/contribution	Deployment environment	Validation depth	Clinical relevance/risk	Evidence level/maturity
[101]	Cross-Institutional Medical Data Sharing (Chronic Disease)	Proposed a decentralized and privacy-preserving collaboration framework that integrates blockchain and FL[b], enhancing the predictive performance of diabetes models while ensuring data security and reducing communication overhead	Evaluation on public dataset (unspecified diabetes data)	Retrospective data validation	Population heterogeneity is not covered; it is difficult to extrapolate to real clinical patients	Level 3: Preclinical
[102]	Cross-Institutional Medical Data Sharing (Chronic Diseases)	Developed a blockchain-integrated FL mechanism to enhance IoMT[c] data privacy and improve diabetes prediction accuracy, achieving 97.11% accuracy with a multilayer perceptron model	Evaluation on public dataset (Pima Indians Diabetes)	Retrospective data validation	Relying on a small and well-defined dataset limits the robustness of the findings	Level 3: Preclinical
[103]	Cross-Institutional Medical Data Sharing (Data Management)	Introduced the ODMSM-FL[d] framework, which optimizes storage, management, and privacy protection for IoMT data, enhancing data security and system efficiency	Evaluation on public EHR[e] dataset (HealthData.gov)	Controlled experimental conditions	Data latency, human-machine device heterogeneity; The real IoMT network is uncontrollable	Level 2: Prototype validation
[104]	Cross-Institutional Medical Data Sharing (Medical Imaging)	Designed a blockchain-powered FL framework for brain tumor segmentation using 3D U-Net, achieving significant improvements in Dice similarity coefficient and Hausdorff distance	Evaluation on public benchmark dataset (BraTS 2020)	Retrospective data validation	Relying on benchmark datasets rather than prospective clinical data limits the clinical translational application	Level 2: Prototype validation
[105]	Cross-Institutional Medical Data Sharing (Medical Imaging)	Proposed the FB-CLC-Rad[f] framework, integrating CapsNets, blockchain, and FL to enhance lung cancer nodule detection accuracy in CT scans, reaching 99.69% accuracy	Evaluation on public and local dataset (Cancer Imaging Archive [CIA], Kaggle Data Science Bowl [KDSB], LUNA 16, and local datasets)	Retrospective data validation	The process of not covering the real image; lack of doctor decision-making and workflow verification	Level 2: Prototype validation

XSL•FO
RenderX

Reference	Application domain	BCFL[a] architecture/contribution	Deployment environment	Validation depth	Clinical relevance/risk	Evidence level/maturity
[106]	Cross-Institutional Medical Data Sharing (Drug Discovery)	Designed RahasakML, a decentralized blockchain-FL platform enabling multi-institutional collaboration with enhanced transparency and security in drug discovery	Theoretical analysis; limited empirical validation	Conceptual framework	Empirical verification in actual test environments is limited	Level 1: Conceptual
[107]	Cross-Institutional Medical Data Sharing (EMR[g])	Integrated FL and blockchain for cloud-based medical record recommendation systems, leveraging Hyperledger Fabric, IPFS[h], LightGBM, and N-Gram models for collaborative learning	Evaluation on public EHR dataset (not specified)	Simulated	Limited to prototype frameworks or simulations, not prospectively evaluated in clinical practice	Level 2: Prototype validation
[108]	Cross-Institutional Medical Data Sharing (EMR)	Proposed a blockchain-FL framework for EHR privacy protection, achieving 92.5% global model accuracy and 88.33% local model accuracy using a deep neural network	Evaluation on public EHR dataset (Chronic Kidney Disease [CKD] dataset [UCI Machine Learning Repository])	Retrospective data validation	Limited to prototype frameworks or simulations, not prospectively evaluated in clinical practice	Level 3: Preclinical
[109]	Cross-Institutional Medical Data Sharing (EMR)	Used lightweight encryption and FL to secure EHR data in an Ethereum test environment, reducing reliance on trusted third parties	Evaluation on public EHR dataset (Simulation in Ethereum test environment)	Simulated	Limited to prototype frameworks or simulations, not prospectively evaluated in clinical practice	Level 2: Prototype validation
[110]	Cross-Institutional Medical Data Sharing (EMR)	Combined CNN[i] and blockchain-FL to enhance EHR data security and detect abnormal user behaviors automatically	Evaluation on public EHR dataset (Python-based simulation)	Simulated	Limited to prototype frameworks or simulations, not prospectively evaluated in clinical practice	Level 2: Prototype validation
[99]	Cross-Institutional Medical Data Sharing (EMR)	Explored blockchain-FL applications in precision medicine, emphasizing diagnostic accuracy, treatment optimization, clinical trial subpopulation identification, and drug development acceleration	Evaluation on public EHR dataset (not specified)	Simulated	Limited to prototype frameworks or simulations, not prospectively evaluated in clinical practice	Level 2: Prototype validation

XSL•FO
RenderX

Reference	Application domain	BCFL[a] architecture/contribution	Deployment environment	Validation depth	Clinical relevance/risk	Evidence level/maturity
[111]	IoMT (Data Security)	Proposed a lightweight hybrid FL framework with blockchain smart contracts for edge training plan management, trust evaluation, and authentication in IoMT networks	Evaluation on public COVID-19 dataset (not specified)	Simulated	The device has heavy computational burden, high system complexity, and difficult clinical translation	Level 1: Proof-of-concept
[112]	IoMT (Data Management)	Developed a blockchain-FL system leveraging edge computing and Paillier encryption to securely manage medical resource transactions in IoMT environments	Evaluation on public dataset (unspecified diabetes data)	Retrospective data validation	The device has heavy computational burden, high system complexity, and difficult clinical translation	Level 2: Prototype validation
[113]	IoMT (Data Security)	Introduced a distributed reinforcement learning method integrating blockchain and FL for improved data privacy and security in IoMT applications	Evaluation on public COVID-19 dataset (not specified)	Simulated	The device has heavy computational burden, high system complexity, and difficult clinical translation	Level 2: Prototype validation
[114]	IoMT (Data Security)	Proposed the FL-BETS[j] framework, leveraging fog computing and blockchain to minimize energy consumption and latency while enhancing fraud detection in health care	Evaluation on Private dataset focusing on medical insurance fraud (Kaggle)	Simulated	The device has heavy computational burden, high system complexity, and difficult clinical translation	Level 1: Proof-of-concept
[115]	IoMT (Data Security)	Developed a fog computing IoT platform that integrates FL and private blockchain technology to enhance privacy protection in wearable IoMT devices	Evaluation on human activity recognition dataset (UCI[k] Machine Learning Library)	Simulated	The device has heavy computational burden, high system complexity, and difficult clinical translation	Level 2: Prototype validation
[117]	Public Health (COVID-19 Imaging)	Proposed a blockchain-FL-based IoMT architecture for COVID-19 detection and epidemic management; the architecture enhances data privacy through FL and ensures data transparency and immutability via blockchain	Evaluation on public COVID-19 dataset (Centers for Disease Control [CDC] data)	Simulated	High real-time requirements in epidemic environment; blockchain delay is not resolved	Level 1: Proof-of-concept

XSL•FO
RenderX

Reference	Application domain	BCFL[a] architecture/contribution	Deployment environment	Validation depth	Clinical relevance/risk	Evidence level/maturity
[118]	Public Health (COVID-19 Imaging)	Developed a blockchain-based FL framework for COVID-19 detection, using Capsule Networks for image segmentation and classification to enhance data privacy and model accuracy	Evaluation on Private COVID-19 dataset (CC-19)	Retrospective data validation	The limited scale and types of data restrict the generalization ability of the model	Level 3: Preclinical
[119]	Public Health (COVID-19 Imaging)	Introduced FLED-Block[l], a blockchain-based FL model integrating Capsule Networks for image feature extraction and extreme learning machines (ELM) for classification, achieving high accuracy with strong privacy protection	Evaluation on public COVID-19 dataset (CT data from multiple hospitals)	Retrospective data validation (multi-source datasets)	The computational complexity and the feasibility of actual deployment require further research	Level 3: Preclinical
[120]	Public Health (Pandemic Diagnosis)	Designed BFLPD[m], a blockchain-FL framework for epidemic diagnosis in smart cities, particularly for COVID-19; the framework ensures secure model aggregation and enhances global model integrity and efficiency	Evaluation on public ultrasound COVID-19 dataset (POCUS, ICLUS-DB, and COVIDx-US)	Retrospective data validation (multi-source datasets)	The huge computational overhead and the high resistance to actual deployment	Level 3: Preclinical
[5]	Telemedicine (Telemedicine System)	Proposed a blockchain-FL application framework for telemedicine, analyzing how these technologies improve data accessibility, security, and privacy in remote health care	Based on theoretical simulations or examples	Conceptual framework	Mainly focuses on the theoretical framework, the lack of a real-world application cases	Level 1: Proof-of-concept

XSL•FO
RenderX

Reference	Application domain	BCFL[a] architecture/contribution	Deployment environment	Validation depth	Clinical relevance/risk	Evidence level/maturity
[121]	Telemedicine (Remote Surgery System)	An intelligent remote surgery framework named BITS[n], which is based on blockchain and artificial intelligence, is proposed; this architecture integrates blockchain technologies (such as Ethereum and IPFS protocols), 6G communication networks, and federated learning (or AI[o] algorithms), aiming to enhance the security, privacy, and real-time performance of remote surgery systems	Based on theoretical simulations or examples	Simulated	Limited to a simulated environment, no actual deployment or clinical evaluation has been carried out yet	Level 1: Proof-of-concept
[122]	Telemedicine (Remote Surgery System)	Developed a remote surgery system framework leveraging blockchain and FL to enhance data security, reliability, and real-time processing; the framework integrates 6G networks and IPFS for low-latency and high-reliability data transmission	Based on theoretical simulations or examples	Conceptual framework	There is insufficient discussion on the specific implementation details of BCFL in remote surgery and a lack of clinical deployment	Level 1: Proof-of-concept

[a]BCFL: blockchain-based federated learning.

[b]FL: federated learning.

[c]IoMT: Internet of Medical Things.

[d]ODMSM-FL: Optimized Data Management and Secured Federated Learning.

[e]EHR: electronic health record.

[f]FBCLC-Rad: Federated Learning–Enabled Blockchain CapsNets Lung Cancer Radiologist.

[g]EMR: electronic medical record.

[h]IPFS: Interplanetary File System.

[i]CNN: convolutional neural network.

[j]FL-BETS: Federated Learning–Based Blockchain-Enabled Task Scheduling.

[k]UCI: University of California, Irvine.

[l]FLED-Block: Federated Learning–Ensembled Deep Learning Blockchain Model.

[m]BFLPD: blockchain-based federated learning for pandemic diagnosis.

[n]BITS: Blockchain-Driven Intelligent Scheme for Telesurgery System.

[o]AI: artificial intelligence.

As shown in the table, most BCFL studies are still focused on the concept or prototype stage, lacking multicenter real data validation and evaluation corresponding to clinical endpoint indicators. To promote the clinical application of BCFL, improvements need to be made in three aspects: (1) conduct cross-institutional and prospective validations to evaluate the model's performance in real patient populations and medical processes; (2) strike a balance among security, latency, and maintainability to avoid the unavailability caused by simply pursuing complex encryption or on-chain computing; and (3) achieve integration with medical information systems, data governance and regulatory frameworks, and deploy the system

XSL•FO
RenderX

under the premise of clearly defining data responsibilities and authorities. Only when verified under the joint constraints of clinical workflow, patient heterogeneity, and compliance requirements can BCFL gradually evolve from a conceptual technology to a usable medical solution.

Discussion

Challenge

Although BCFL shows great potential in transforming health care data sharing, its deployment in real medical environments is still highly limited. The vision of change for BCFL must be balanced with technical limitations and the specific complexity of applications (particularly the interoperability gap, high implementation costs, and unresolved scalability bottlenecks), which prevent it from being transformed from a conceptual framework into regular clinical practice.

There are many application challenges in the field of medical data sharing, including the lack of system interoperability and the absence of standardized benchmark datasets specifically designed for medical applications. Currently, various health care information systems (eg, EMR systems) use diverse system architectures, data formats, and operational standards, lacking a unified interoperability framework. Moreover, unlike conventional FL research that often leverages open-access datasets such as CIFAR or MNIST, health care data are inherently sensitive, fragmented, and institution-specific, which makes reproducibility and cross-study comparability particularly difficult.

Scalability and communication efficiency also present critical obstacles. As FL tasks expand, the number of health care data sources and the complexity of training increase significantly. However, in practical deployments, the scalability and throughput limitations of blockchain are particularly pronounced. Even in permissioned frameworks such as Hyperledger Fabric, which offer improved throughput, empirical benchmarks still report end-to-end latencies of several seconds per block under moderate workloads. Insufficient mining resources slow down block generation and verification, hindering the efficient execution of large-scale tasks [123]. Moreover, the influx of numerous participants in distributed health care environments amplifies the load on the blockchain network, while the efficiency of existing consensus mechanisms is difficult to meet the demands of health care applications [124]. When applied to BCFL, these constraints imply that each iteration of local training, parameter aggregation, and block creation may introduce cumulative delays that significantly slow model convergence. Furthermore, as the number of blockchain nodes rises, communication costs increase exponentially. The network delays and communication efficiency degradation caused by high communication costs directly impact the training speed and overall model performance.

High implementation and maintenance costs represent another practical barrier. Establishing a BCFL infrastructure requires significant upfront investment in blockchain nodes, secure servers, storage, and high-speed networking. Additionally, energy consumption associated with blockchain consensus protocols, as well as the operational costs of managing frequent model updates across institutions, may exceed the financial capacity of many health care providers, especially in resource-limited settings. Without clear evidence of cost-benefit balance, hospitals and regulators may be reluctant to adopt BCFL at scale.

Additional difficulties arise from the integration of BCFL within IoMT environments. The heterogeneity of IoMT devices results in substantial disparities in storage capacity, computational power, energy consumption, and communication capabilities. For instance, advanced hospital equipment often features powerful processors, stable power supplies, and ample storage space, whereas wearable medical devices typically operate on low-power batteries, constrained network bandwidth, and limited computational resources. This disparity in device capabilities poses a significant challenge to the deployment of FL models. Moreover, energy constraints and unstable network connections make edge medical devices prone to data transmission failures or system disconnections, ultimately resulting in end-device desynchronization. This issue not only hampers the timeliness of data uploads and model updates but may also prevent the global model from converging efficiently.

Compounding these challenges is the heterogeneity of health care data within IoMT systems. Data generated by various devices exhibit significant diversity, often displaying uneven distributions and violating the non-IID assumption. For instance, hospital A may collect dynamic ECG signals, whereas hospital B primarily acquires static medical images. Such disparities in data distribution exacerbate the complexity of model training. Moreover, variations in medical coding standards across countries and regions (eg, ICD-10 [*International Statistical Classification of Diseases, Tenth Revision*] in the United Kingdom vs ICD-10-CM [*International Classification of Diseases, Tenth Revision, Clinical Modification*] in the United States) contribute to data standard inconsistencies. Consequently, such heterogeneities complicate global model training, analysis, and evaluation, ultimately impairing the model's generalization across diverse clients [125].

Beyond the application-level challenges, BCFL also faces significant technical limitations that must be addressed to realize its full potential in medical settings. Although BCFL integrates the decentralized nature of blockchain with the "data availability without visibility" principle of FL to provide an initial level of privacy protection and enhance overall system security, it does not fully resolve privacy concerns. The sensitivity of patient data and the stringent privacy requirements in the medical field necessitate addressing a series of complex security challenges. The system remains vulnerable to various types of malicious attacks that pose significant risks to patient confidentiality and institutional trust. For instance, background knowledge attacks involve adversaries inferring sensitive information using previously known data and analyzing shared model parameters [3]. In conspiracy attacks, multiple nodes conspire to steal data features from other participants by exchanging local training information [126]. Inference attacks similarly analyze model parameter updates, and attackers can infer private details about patient data [1,43]. These threats highlight the critical need for

XSL•FO
RenderX

advanced privacy-preserving mechanisms within BCFL to ensure the safety and integrity of medical data.

While existing privacy-preserving technologies provide preliminary protections, they are often inadequate when faced with the dual demands of strong privacy and high model utility. Homomorphic encryption allows computation directly on encrypted data, preventing plaintext exposure. However, its computational inefficiency makes it unsuitable for large-scale, complex operations. Differential privacy introduces noise to model parameters to obscure individual data contributions; however, this can significantly compromise model accuracy and system performance if not carefully balanced. Secure Multiparty Computation (SMPC) offers robust data confidentiality through distributed computation, yet its reliance on frequent interaction between parties contradicts the low-interaction protocols typically favored in BCFL for efficient aggregation. These limitations underscore the urgent need for lightweight, efficient, and scalable privacy-preserving solutions specifically tailored to the BCFL context.

Another critical technical issue is the optimization of incentive mechanisms. In traditional blockchain systems, fixed token-based reward structures fail to reflect the true value of each participant's contribution [8]. This misalignment can result in low-quality nodes receiving undeserved rewards, while high-contribution nodes may become demotivated due to insufficient compensation. In BCFL, this issue is further complicated by the heterogeneity of participants, who differ in computational capabilities, data quality, and participation frequency. Resource-constrained nodes, in particular, may lack sufficient incentives to participate, ultimately affecting the quality and diversity of the global model. To address these disparities, incentive mechanisms in BCFL must go beyond simple token rewards and instead adopt dynamic, contribution-aware frameworks that account for the multidimensional nature of participant involvement. A well-designed incentive system can encourage broader and more sustained engagement, improve fairness, and enhance the overall efficiency and robustness of BCFL. Therefore, developing more sophisticated and adaptable incentive mechanisms is a pressing direction for future research.

While the immutability of blockchain and the tamper-resistant nature of smart contracts ensure data integrity and trustworthiness, these characteristics also introduce rigidity, posing challenges in dynamic health care environments [33]. Several scenarios highlight the importance for greater flexibility: when patient data are entered incorrectly or require modification, the immutable blockchain structure cannot accommodate efficiently; then patients may request the deletion or modification of their data to preserve privacy, particularly to comply with data protection regulations; moreover, in public health emergencies, ensuring timely access to accurate health care data is crucial for effective crisis management, necessitating mechanisms for controlled updates within blockchain systems. These scenarios underscore the need for greater flexibility within BCFL systems, where mechanisms must be designed to support controlled edits under predefined conditions—balancing the need for data integrity with the operational demands of evolving health care environments.

What's more, another important consideration in the medical application of BCFL is model interpretability, which directly impacts the reliability of clinical decisions, patient trust, and regulatory compliance. In medical decision-making, the ability to interpret model outcomes is essential to ensuring safety, transparency, and credibility. However, the "black-box" nature of many complex deep learning models limits their interpretability, posing significant challenges for medical applications [106]. The necessity of interpretability can be emphasized from multiple perspectives: For health care professionals, AI model predictions must be interpretable to allow physicians to understand the underlying reasoning and effectively integrate them into diagnosis, treatment planning, and patient monitoring. Moreover, interpretability enables researchers and clinicians to identify and trace the sources of bias or errors in model predictions, facilitating performance optimization and improving diagnostic accuracy and reliability. Simultaneously, in terms of patients, the widespread adoption of AI in medicine inevitably raises concerns regarding privacy and ethics. Enhancing model interpretability can build patient trust in AI-assisted diagnosis and treatment by clarifying the model's reliability and limitations, thereby mitigating concerns over "black-box" decision-making. Moreover, at the regulatory level, numerous countries and regions have mandated transparency and auditability in medical AI systems to ensure that decision-making processes align with ethical and legal standards. Furthermore, medical AI operates within a highly interdisciplinary environment, encompassing physicians, technology developers, data scientists, and other professionals. Interpretability serves as a crucial bridge for communication among experts from diverse disciplines, facilitating the effective implementation of medical AI technologies and ultimately enhancing the quality and accessibility of health care services.

Finally, summarize the limitations of the literature included in the review: first of all, the risk of prejudice is very common. Many available studies are conceptual frameworks, simulations, or small-scale case studies, rather than large-scale clinical implementations, without independent external validation. Common methodological flaws include the selective presentation of favorable performance metrics, limited or lack of adversarial and privacy attack tests, and the absence of long-term or actionable measurements (such as maintenance burden, interoperability failures, or ongoing participation rates). Unavailable code, undisclosed model/configuration details, and dependencies on nonshareable datasets often compromise reproducibility. In conclusion, these factors have created systemic uncertainties, posing the risk of overestimating feasibility and underestimating the actual deployment challenges. Second, the included studies demonstrated substantial inconsistencies in the key dimensions of system design and evaluation. Research in BCFL architecture (fully coupled, flexibly coupled, and loosely coupled), blockchain configuration (private and public), privacy countermeasures (secure aggregation, differential privacy, homomorphic encryption, and SMPC), and data sources (public benchmarks, single-center clinical records, and various IoMT streams) varies greatly. The result measurement criteria have not been standardized: some papers prioritize predictive performance, others emphasize communication or computational overhead,

XSL•FO
RenderX

and still others focus on source or incentive metrics. This heterogeneity has led to different discoveries. Furthermore, there are potential biases because many studies are conducted in a controlled environment with carefully curated datasets, which may not reflect the heterogeneity and noise of real-world medical data. Finally, due to the limited performance index reports and insufficient longitudinal validation, statistical uncertainties (confidence intervals and variability between runs) are rarely reported; large-scale, multi-institutional deployments remain uncommon; and few studies have evaluated long-term stability, scalability under large volumes of clinical data, or regulatory compliance under real-world conditions. These gaps prominently indicate the need for more robust and well-designed research, including prospective clinical trials, to verify the effectiveness, safety, and interoperability of the BCFL framework in actual health care settings.

Future Prospects

The main obstacles to data sharing and collaboration in the health care system lie in the lack of interoperability and standardization, as well as the absence of standardized benchmark datasets in medical applications in the context of FL and blockchain applications. Therefore, future research must integrate the joint efforts of the government, regulatory authorities, and industry leaders to establish unified technical standards and policy frameworks and give priority to the development of standardized BCFL benchmark datasets. These standards should encompass data formats, transmission protocols, privacy safeguards, and technology implementation guidelines to facilitate seamless data integration and collaboration across diverse health care entities. Standardized datasets should reflect the realistic heterogeneity of medical imaging, EHRs, and multimodal data streams. By providing a common reference point, these benchmarks will enable fair and transparent algorithm comparisons, promote the reproducibility of results, and accelerate the transformation of BCFL innovations into clinical validation tools. In addition, benchmark datasets can be stratified based on disease types, patterns, and clinical tasks (eg, diagnosis, prognosis, and treatment response prediction), thereby allowing for more fine-grained evaluations of system performance in different health care settings. Moreover, regulatory frameworks should prioritize patient privacy and data security, delineate the rights and responsibilities of stakeholders, and foster the compliant adoption of BCFL technology. Comprehensive policy support and technical guidance will be instrumental in mitigating data silos and enhancing the efficiency of collaborative model training.

Scalability and communication efficiency remain two of the most critical technical challenges facing BCFL, particularly in large-scale and high-frequency health care environments. To address the scalability limitations, future research should focus on several key technical optimizations. The first is off-chain computing and side-chain technology [43]. Off-chain computing enables complex training tasks to be executed off-chain, reducing the computational burden on the main blockchain. Meanwhile, side chains can independently handle task-specific transactions, alleviating congestion on the main chain. Layer 2 protocols, such as stateful channels and plasma technology, offer a viable solution by enabling faster transaction processing while preserving blockchain security [43]. These technologies enhance scalability by minimizing main-chain data storage requirements. Another crucial research direction is cross-chain technologies [42]. This approach not only decentralizes workloads and mitigates single-chain bottlenecks but also enhances system-wide parallel processing capabilities. Moreover, cross-chain FL is particularly well-suited for cross-regional and cross-organizational health care collaborations, further strengthening resource management and system security.

Regarding communication cost and efficiency, the following directions are worth exploring in depth. One promising direction is gradient compression. Using gradient compression techniques helps reduce communication overhead. For instance, Konecny et al [127] proposed that structured updates and sketch updates can significantly lower communication costs. However, this approach also introduces potential challenges, such as the loss of relevant information during compression, which may have an impact on the performance of the global model. Therefore, future research should focus on achieving an optimal balance between gradient compression and global model accuracy [128]. Lightweight consensus protocols are also another crucial research direction. Another key technological direction is the Digital Twin [14], which minimizes the need for long-distance data transmission by enabling the generation of virtual models directly on miner nodes. This approach significantly decreases communication latency and costs, making it particularly well-suited for resource-constrained health care environments.

In the field of IoMT, to address the challenges posed by heterogeneous storage, computing, and communication capabilities of medical devices and sensors, future research should prioritize optimizing resource usage. First, lightweight ML models and algorithms, such as model compression and pruning techniques [128], can be developed to alleviate the computational burden on local devices. This not only enhances device operational efficiency but also significantly reduces communication overhead. At the task allocation level, device grouping and hierarchical architectures can be leveraged to allocate computation and data aggregation tasks to devices or intermediate nodes with higher computational capacity, thereby forming a resource-optimized collaborative network.

Another critical aspect is enhancing system robustness in the face of equipment instability. To mitigate this issue, future research should focus on optimizing both failure recovery mechanisms and participant selection strategies. On the one hand, an adaptive connection protocol can be designed to enable devices to automatically rejoin the training process after a connection interruption, ensuring that the global model's convergence remains unaffected. On the other hand, an optimization strategy based on device availability and performance can be implemented to prioritize the selection of more stable devices for training. Moreover, incorporating a flexible time window would allow devices to complete tasks within a predefined period, thereby enhancing the overall system's fault tolerance and training efficiency.

Given the highly diverse and statistically heterogeneous nature of medical data in IoMT environments, achieving robust model

XSL•FO
RenderX

generalization is another priority. The prevalence of non-IID data across institutions often results in significant discrepancies between global and local models, undermining convergence and performance. Consequently, Wu and Wang [125] have proposed an optimal aggregation algorithm that dynamically adjusts the selection probability of each trainer based on the algorithm's output. However, trainer selection based on model preferences can severely compromise the generalization capability of the global model. Future research should focus on developing a highly accurate BCFL system with enhanced model generalization. One promising direction is the development of automated data normalization tools, which are capable of recognizing various dataset formats and characteristics, performing automatic transformations to ensure that the data can be directly used by model training [129].

The technical limitations faced by BCFL systems need to be well addressed in the future, and privacy protection is one of the fundamental challenges. Future research should further weigh the relationship between privacy and utility, selecting appropriate privacy-preserving techniques based on system priorities. For instance, in certain scenarios, prioritizing privacy may necessitate the adoption of computationally intensive yet more secure techniques, whereas high-performance requirements may favor more lightweight solutions. Additionally, more advanced privacy-preserving techniques can be explored. For example, zero-knowledge proof (ZKP) is a promising cryptographic method that enables a prover to demonstrate the validity of a statement to a verifier without disclosing any additional information. This ensures that the data owner can validate the accuracy of an update while keeping the original data confidential. This approach not only minimizes the trust overhead among participants but also alleviates the verification burden on clients. Future research should investigate how ZKP can be seamlessly integrated into the BCFL architecture to optimize the balance between performance and privacy, paving the way for a more efficient and secure system with enhanced privacy measures.

Another pressing issue in the development of BCFL systems is the design of fair and dynamic incentive mechanisms to ensure equitable resource allocation among contributors. On the one hand, the Shapley value-based contribution quantification method can be used to assess each participant's impact on global model performance improvement, thereby enabling a fairer incentive distribution. On the other hand, since variations in data quality directly impact model training effectiveness, data quality-driven incentives can be introduced so that participants contributing high-quality or more representative data receive greater rewards. Moreover, integrating penalty mechanisms is also another crucial research direction [85]. For instance, Cui et al [130] proposed withdrawing tokens when a trainer's behavior is identified as malicious. Similarly, Weng et al [79] suggested requiring trainers to predeposit tokens, which are forfeited upon detection of malicious activity, but the fairness of this deposit mechanism remains uncertain. Therefore, how to reasonably set the punishment rules under the premise of ensuring fairness still needs further research. These optimization strategies will contribute to building a more effective and equitable BCFL ecosystem, fostering its sustainable development in medical data sharing.

It is undeniable that token-based incentive mechanisms have been widely discussed as a promising approach to encouraging participation in the BCFL network. Although such mechanisms can enhance participation and promote fair resource allocation, they also introduce complex ethical issues. Therefore, in future research, ethical supervision should be embedded in token-based system design, and new solutions should be continuously explored.

Editable mechanisms are expected to be implemented in the near future. Future research could explore how to introduce a moderate degree of editability while ensuring data integrity and security. For instance, the modified Chameleon hash function (also known as the trapdoor hash function) enables controlled modification of blockchain data. When the trapdoor information is available, hash collisions can be efficiently identified, allowing input modifications without altering the hash output. This mechanism facilitates the correction of inaccurate or incomplete data while preserving the structural integrity of the blockchain [8].

In parallel, the interpretability of AI models remains a central concern for the deployment of BCFL in clinical practice. Trust, transparency, and accountability are all closely tied to how well clinicians and patients can understand the rationale behind AI-generated predictions. Existing mainstream interpretability tools (eg, SHAP and LIME) have limitations in the medical field, including difficulties in handling complex distributed data environments and the inability to provide clinically relevant interpretations. Future research should focus on enhancing existing tools and developing interpretability methods adapted to BCFL, such as global-local model comparison techniques, to provide more intuitive and trustworthy interpretations.

In addition to technological and application-level innovations, the future development of BCFL also needs to confront the issue that research mainly remains at the simulation, prototype, and preclinical stages, and these studies have not yet formed a direct correspondence with real medical endpoint indicators. Future work should start from real scenarios: (1) conduct prospective validations in multi-institutional, heterogeneous data and complete clinical workflows to evaluate their actual impact on diagnostic efficiency, therapeutic effect improvement, and resource usage; (2) maintain a balance between security, latency, and maintainability in system design to avoid unacceptable computing and communication costs caused by encryption and on-chain operations; and (3) at the governance level, strengthen the connection with the medical system, regulations and ethical frameworks, and clarify data ownership, responsibility attribution, and auditing mechanisms. Only when technical performance is validated in real clinical settings, system design remains practically deployable, and governance mechanisms are clearly defined, BCFL can progress from laboratory prototypes to clinically reliable infrastructure.

While the above guidance outlines strategic directions for policymakers, clinicians, and implementers, these recommendations remain high-level. To move from broad vision to actionable progress, there is a pressing need for specific

XSL•FO
RenderX

evaluative structures that can translate theories into measurable outcomes. Therefore, establishing a good evaluation framework is indispensable, which can provide a foundation for future research, allowing for comparison, verification, and ultimately integration into clinical workflows.

A unified, multidimensional evaluation framework is central to the advancement of BCFL in medicine. From a technical perspective, standardized metrics such as model accuracy, convergence speed, robustness against adversarial attacks, communication latency, and scalability across heterogeneous institutional datasets are indispensable. These indicators establish the baseline scientific validity of BCFL. Equally important afterwards are the clinical assessment criteria, which should reflect the sensitivity and specificity of the diagnosis, its universality in different patient populations and multicenter environments, reduce algorithmic bias, and bring about a tangible improvement in patient prognosis. By embedding clinical endpoints into the assessment, this framework can ensure that technological progress is in line with the real-world health care needs. Finally, there are operational indicators, which should cover interoperability with existing medical data systems, cost-effectiveness, and the long-term sustainability of deployment. By integrating these three dimensions into a unified evaluation system, researchers and clinical workers can establish benchmarks that can both horizontally compare different studies and vertically track research progress. This evaluation framework not only enhances the scientific rigor of BCFL research but also provides an evidence-based basis for decision-making sharing, thereby accelerating the transformation from experimental prototypes to clinical applications.

Conclusions

This review systematically compiles the research progress of blockchain and FL in the medical field. First, we introduce the theoretical foundations and core technical features of both technologies, analyzing how blockchain enhances the security, privacy protection, and decentralization characteristics of FL, while FL improves the computational efficiency and scalability of blockchain. In addition, we describe the three frameworks and workflows of BCFL. Next, we summarize the research progress in BCFL applications across cross-institutional health care data sharing, the IoMT, public health monitoring, and telemedicine, highlighting its practical value in privacy protection and data collaboration. Moreover, we discuss key challenges in BCFL, including computational efficiency, scalability, data privacy, and incentive mechanism design, while proposing potential solutions and future research directions. The significance of this review lies in providing a comprehensive overview of how BCFL could reshape medical data collaboration and security paradigms, thereby offering valuable insights for researchers and practitioners exploring this interdisciplinary field. However, it is important to acknowledge certain limitations of this review. This review primarily focuses on the theoretical principles and current applications of BCFL, with relatively limited exploration of specific implementation details and performance evaluations. Additionally, most existing BCFL studies rely on simulation experiments or public datasets, lacking validation with large-scale real-world medical data, which affects assessments of its practical feasibility. As such, although BCFL shows promise in supporting future intelligent diagnostics, precision medicine, and collaborative health care systems, claims about clinical readiness remain premature. Future work should focus on prospective, large-scale validation studies, interdisciplinary collaboration with health care providers, and the development of standardized evaluation protocols to ensure that BCFL solutions are clinically safe, ethically sound, and operationally feasible.

Acknowledgments

We sincerely thank the journal editor Javad Sarvestan, PhD, and the peer reviewers for their valuable feedback during this study, which has greatly enhanced our research. We used ChatGPT (OpenAI) to check and revise the vocabulary and grammar of this manuscript and polished the language in the introduction.

Funding

This research was funded by the National Natural Science Foundation of China (project no. 81974355 and 82172524), the Key Research and Development Program of Hubei Province (project no. 2021BEA161), the National Innovation Platform Cultivation Program (project no. 2020021105012440), and the Education Reform Project of Ningxia Medical University (funded by NYJY2025057).

Data Availability

All data generated or analyzed during this study are included in Multimedia Appendices 1 and 2.

Authors' Contributions

Conceptualization: XW (lead), YX (equal), XC (equal)
Data curation: XW (lead), YX (supporting), XC (supporting), JY (supporting)
Formal analysis: XW
Funding acquisition: ZY
Investigation: XW (lead), YX (supporting), XC (supporting), JY (supporting), RL (supporting)
Methodology: XW (lead), YX (supporting), XC (supporting), JY (supporting)

XSL•FO
RenderX

Project administration: XW (lead), YX (supporting), XC (supporting), JY (supporting)
Resources: XW (lead), YX (equal), XC (supporting)
Supervision: ZY (lead), YX (supporting)
Validation: YX (lead), ZY (supporting)
Visualization: XW (lead), YX (equal), XC (supporting)
Writing – original draft: XW (lead), YX (equal), XC (supporting), JY (supporting)
Writing – review & editing: XW (lead), YX (supporting), XC (supporting), ZY (supporting), JY (supporting), RL (supporting), WG (supporting), ZY (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Databases search strategy.
[DOCX File, 22 KB - jmir_v28i1e79052_app1.docx]

Multimedia Appendix 2
Scoring statistics and agreement data.
[XLSX File, 13 KB - jmir_v28i1e79052_app2.xlsx]

Multimedia Appendix 3
Additional tables.
[DOCX File, 14 KB - jmir_v28i1e79052_app3.docx]

Checklist 1
PRISMA checklist.
[DOCX File, 31 KB - jmir_v28i1e79052_app4.docx]

Checklist 2
Kitchenham SLR Quality Checklist.
[DOCX File, 17 KB - jmir_v28i1e79052_app5.docx]

References

1. Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning. In: 2019 IEEE Symposium on Security and Privacy: IEEE; 2019:739-753. [doi: 10.1109/SP.2019.00065]
2. Qu Y, Gao L, Luan TH, et al. Decentralized privacy using blockchain-enabled federated learning in fog computing. IEEE Internet Things J 2020 Jun;7(6):5171-5183. [doi: 10.1109/JIOT.2020.2977383]
3. Wang Z, Song M, Zhang Z, Song Y, Wang Q, Qi H. Beyond inferring class representatives: user-level privacy leakage from federated learning. In: IEEE INFOCOM 2019 – IEEE Conference on Computer Communications: IEEE; 2019:2512-2520. [doi: 10.1109/INFOCOM.2019.8737416]
4. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning. ACM Trans Intell Syst Technol 2019 Mar 31;10(2):1-19. [doi: 10.1145/3298981]
5. Hiwale M, Walambe R, Potdar V, Kotecha K. A systematic review of privacy-preserving methods deployed with blockchain and federated learning for the telemedicine. Healthc Anal 2023 Nov;3:100192. [doi: 10.1016/j.health.2023.100192]
6. Boland MR, Parhi P, Li L, et al. Uncovering exposures responsible for birth season – disease effects: a global study. J Am Med Inform Assoc 2018 Mar 1;25(3):275-288. [doi: 10.1093/jamia/ocx105]
7. Freire-Garabal y Núñez M. General vision of the Regulation (EU) 2016/679 of the European Parliament and of the Council of April 27, 2016. Al-Khalifa Business School 2020 Jun 21. [doi: 10.21428/18D9181C.D22AA10A]
8. Qu Y, Uddin MP, Gan C, Xiang Y, Gao L, Yearwood J. Blockchain-enabled federated learning: a survey. ACM Comput Surv 2023 Apr 30;55(4):1-35. [doi: 10.1145/3524104]
9. Uddin MA, Stranieri A, Gondal I, Balasubramanian V. A survey on the adoption of blockchain in IoT: challenges and solutions. Blockchain Res Appl 2021;2(2):100006. [doi: 10.1016/j.bcra.2021.100006]
10. Alsamhi SH, Myrzashova R, Hawbani A, et al. Federated learning meets blockchain in decentralized data sharing: healthcare use case. IEEE Internet Things J 2024 Jun 1;11(11):19602-19615. [doi: 10.1109/JIOT.2024.3367249]
11. Topaloglu MY, Morrell EM, Rajendran S, Topaloglu U. In the pursuit of privacy: the promises and predicaments of federated learning in healthcare. Front Artif Intell 2021 Oct 6;4:746497. [doi: 10.3389/frai.2021.746497]

XSL•FO
RenderX

12. Li J, Meng Y, Ma L, et al. A federated learning based privacy-preserving smart healthcare system. IEEE Trans Ind Inf 2022 Mar;18(3):2021-2031. [doi: 10.1109/TII.2021.3098010]
13. Zyskind G, Nathan O, Pentland A. Decentralizing privacy: using blockchain to protect personal data. In: 2015 IEEE Security and Privacy Workshops: IEEE; 2015:180-184. [doi: 10.1109/SPW.2015.27]
14. Liu K, Yan Z, Liang X, Kantola R, Hu C. A survey on blockchain-enabled federated learning and its prospects with digital twin. Digit Commun Netw 2024;10(2):248-264. [doi: 10.1016/j.dcan.2022.08.001]
15. Dinh TTA, Liu R, Zhang M, Chen G, Ooi BC, Wang J. Untangling blockchain: a data processing view of blockchain systems. IEEE Trans Knowl Data Eng 2018 Jul 1;30(7):1366-1385. [doi: 10.1109/TKDE.2017.2781227]
16. Gervais A, Karame GO, Wüst K, Glykantzis V, Ritzdorf H, Capkun S. On the security and performance of proof of work blockchains. In: CCS '16: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security: IEEE; 2016:3-16. [doi: 10.1145/2976749.2978341]
17. Kim H, Park J, Bennis M, Kim SL. Blockchained on-device federated learning. IEEE Commun Lett 2020 Jun;24(6):1279-1283. [doi: 10.1109/LCOMM.2019.2921755]
18. Kitchenham B, Charters S. Guidelines for performing systematic literature reviews in software engineering. : Keele University and Durham University Joint Report; 2007 Jul EBSE 2007-001 URL: https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf [accessed 2026-01-29]
19. McMahan HB, Moore E, Ramage D, Hampson S, Agüera y Arcas B. Communication-efficient learning of deep networks from decentralized data. arXiv. Preprint posted online on Jan 26, 2023. [doi: 10.48550/arXiv.1602.05629]
20. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. J Healthc Inform Res 2021;5(1):1-19. [doi: 10.1007/s41666-020-00082-4] [Medline: 33204939]
21. Chen Y, Qin X, Wang J, Yu C, Gao W. FedHealth: a federated transfer learning framework for wearable healthcare. IEEE Intell Syst 2020 Jul 1;35(4):83-93. [doi: 10.1109/MIS.2020.2988604]
22. Long G, Tan Y, Jiang J, Zhang C. Federated learning for open banking. In: Yang Q, Fan L, Yu H, editors. Federated Learning: Privacy and Incentive: Springer; 2020:240-254. [doi: 10.1007/978-3-030-63076-8_17]
23. Zheng Z, Zhou Y, Sun Y, Wang Z, Liu B, Li K. Applications of federated learning in smart cities: recent advances, taxonomy, and open challenges. Conn Sci 2022 Dec 31;34(1):1-28. [doi: 10.1080/09540091.2021.1936455]
24. Zhang C, Xie Y, Bai H, Yu B, Li W, Gao Y. A survey on federated learning. Knowl Based Syst 2021 Mar;216:106775. [doi: 10.1016/j.knosys.2021.106775]
25. Nishio T, Yonetani R. Client selection for federated learning with heterogeneous resources in mobile edge. In: 2019 IEEE International Conference on Communications: IEEE; 2019:1-7. [doi: 10.1109/ICC.2019.8761315]
26. Nguyen DC, Pham QV, Pathirana PN, et al. Federated learning for smart healthcare: a survey. ACM Comput Surv 2023 Mar 31;55(3):1-37. [doi: 10.1145/3501296]
27. Heyndrickx W, Mervin L, Morawietz T, et al. MELLODDY: cross-pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information. J Chem Inf Model 2024 Apr 8;64(7):2331-2344. [doi: 10.1021/acs.jcim.3c00799] [Medline: 37642660]
28. Hao M, Li H, Xu G, Liu Z, Chen Z. Privacy-aware and resource-saving collaborative learning for healthcare in cloud computing. In: 2020 IEEE International Conference on Communications: IEEE; 2020:1-6. [doi: 10.1109/ICC40277.2020.9148979]
29. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis I, Shi W. Federated learning of predictive models from federated electronic health records. Int J Med Inform 2018 Apr;112:59-67. [doi: 10.1016/j.ijmedinf.2018.01.007]
30. Li W, Milletarì F, Xu D, et al. Privacy-preserving federated brain tumour segmentation. In: Suk HI, Liu M, Yan P, Lian C, editors. Machine Learning in Medical Imaging: Springer; 2019:133-141. [doi: 10.1007/978-3-030-32692-0_16]
31. Zhang W, Zhou T, Lu Q, et al. Dynamic-fusion-based federated learning for COVID-19 detection. IEEE Internet Things J 2021 Nov 1;8(21):15884-15891. [doi: 10.1109/JIOT.2021.3056185] [Medline: 35663640]
32. Ulhaq A, Burmeister O. COVID-19 imaging data privacy by federated learning design: a theoretical framework. arXiv. Preprint posted online on Oct 13, 2020. [doi: 10.48550/arXiv.2010.06177]
33. Qammar A, Karim A, Ning H, Ding J. Securing federated learning with blockchain: a systematic literature review. Artif Intell Rev 2023;56(5):3951-3985. [doi: 10.1007/s10462-022-10271-9] [Medline: 36160367]
34. Feng L, Zhao Y, Guo S, Qiu X, Li W, Yu P. BAFL: a blockchain-based asynchronous federated learning framework. IEEE Trans Comput 2022 May 1;71(5):1092-1103. [doi: 10.1109/TC.2021.3072033]
35. Kim YJ, Hong CS. Blockchain-based node-aware dynamic weighting methods for improving federated learning performance. In: 2019 20th Asia-Pacific Network Operations and Management Symposium: IEEE; 2019:1-4. [doi: 10.23919/APNOMS.2019.8893114]
36. Li Y, Chen C, Liu N, Huang H, Zheng Z, Yan Q. A blockchain-based decentralized federated learning framework with committee consensus. IEEE Netw 2021 Jan;35(1):234-241. [doi: 10.1109/MNET.011.2000263]
37. Bhowmick A, Duchi J, Freudiger J, Kapoor G, Rogers R. Protection against reconstruction and its applications in private federated learning. arXiv. Preprint posted online on Jun 3, 2019. [doi: 10.48550/arXiv.1812.00984]
38. Zhang J, Chen J, Wu D, Chen B, Yu S. Poisoning attack in federated learning using generative adversarial nets. In: 2019 18th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/13th IEEE

XSL•FO
RenderX

International Conference on Big Data Science and Engineering (TrustCom/BigDataSE): IEEE; 2019:374-380. [doi: 10.1109/TrustCom/BigDataSE.2019.00057]
39. Kang J, Xiong Z, Niyato D, Yu H, Liang YC, Kim DI. Incentive design for efficient federated learning in mobile networks: a contract theory approach. In: 2019 IEEE VTS Asia Pacific Wireless Communications Symposium (APWCS): IEEE; 2019:1-5. [doi: 10.1109/VTS-APWCS.2019.8851649]
40. Passerat-Palmbach J, Farnan T, Miller R, Gross MS, Flannery HL, Gleim B. A blockchain-orchestrated federated learning architecture for healthcare consortia. arXiv. Preprint posted online on Oct 12, 2019. [doi: 10.48550/arXiv.1910.12603]
41. Majeed U, Hong CS. FLchain: federated learning via MEC-enabled blockchain network. In: 2019 20th Asia-Pacific Network Operations and Management Symposium (APNOMS): IEEE; 2019. [doi: 10.23919/APNOMS.2019.8892848]
42. Sameera KM, Nicolazzo S, Arazzi M, et al. Privacy-preserving in blockchain-based federated learning systems. Comput Commun 2024 Jun;222:38-67. [doi: 10.1016/j.comcom.2024.04.024]
43. Liu J, Chen C, Li Y, et al. Enhancing trust and privacy in distributed networks. arXiv. Preprint posted online on Mar 28, 2024. [doi: 10.48550/arXiv.2403.19178]
44. Shahsavari Y, Dambri OA, Baseri Y, Hafid AS, Makrakis D. Integration of federated learning and blockchain in healthcare: a tutorial. arXiv. Preprint posted online on Apr 15, 2024. [doi: 10.48550/arXiv.2404.10092]
45. Nakamoto S. Bitcoin: a peer-to-peer electronic cash system. SSRN. 2008 Aug 21. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3440802 [accessed 2026-02-12]
46. Zheng Z, Xie S, Dai H, Chen X, Wang H. An overview of blockchain technology: architecture, consensus, and future trends. In: 2017 IEEE International Congress on Big Data (BigData Congress): IEEE; 2017:557-564. [doi: 10.1109/BigDataCongress.2017.85]
47. Nofer M, Gomber P, Hinz O, Schiereck D. Blockchain. Bus Inf Syst Eng 2017 Jun;59(3):183-187. [doi: 10.1007/s12599-017-0467-3]
48. Garay J, Kiayias A. SoK: a consensus taxonomy in the blockchain era. In: Jarecki S, editor. Top Cryptol – CT-RSA 2020: Springer; 2020:284-318. [doi: 10.1007/978-3-030-40186-3_13]
49. Wang W, Hoang DT, Hu P, et al. A survey on consensus mechanisms and mining strategy management in blockchain networks. IEEE Access 2019;7:22328-22370. [doi: 10.1109/ACCESS.2019.2896108]
50. Sukhwani H, Martinez JM, Chang X, Trivedi KS, Rindos A. Performance modeling of PBFT consensus process for permissioned blockchain network (Hyperledger Fabric). In: 2017 IEEE 36th Symposium on Reliable Distributed Systems (SRDS): IEEE; 2017. [doi: 10.1109/SRDS.2017.36]
51. Nguyen CT, Hoang DT, Nguyen DN, Niyato D, Nguyen HT, Dutkiewicz E. Proof-of-stake consensus mechanisms for future blockchain networks: fundamentals, applications and opportunities. IEEE Access 2019;7:85727-85745. [doi: 10.1109/ACCESS.2019.2925010] [Medline: 33777591]
52. Huang H, Li K, Chen X. Blockchain‐based fair three‐party contract signing protocol for fog computing. Concurr Comput 2019 Nov 25;31(22):e4469. [doi: 10.1002/cpe.4469]
53. Wu H, Jiang S, Cao J. High-efficiency blockchain-based supply chain traceability. IEEE Trans Intell Transport Syst 2023;24(4):3748-3758. [doi: 10.1109/TITS.2022.3205445]
54. Dos Santos S, Singh J, Thulasiram RK, Kamali S, Sirico L, Loud L. A new era of blockchain-powered decentralized finance (DeFi) – a review. In: 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC): IEEE; 2022:1286-1292. [doi: 10.1109/COMPSAC54236.2022.00203]
55. Kuo TT, Kim HE, Ohno-Machado L. Blockchain distributed ledger technologies for biomedical and health care applications. J Am Med Inform Assoc 2017 Nov 1;24(6):1211-1220. [doi: 10.1093/jamia/ocx068]
56. Mohanta BK, Jena D, Panda SS, Sobhanayak S. Blockchain technology: a survey on applications and security privacy challenges. Internet Things 2019 Dec;8:100107. [doi: 10.1016/j.iot.2019.100107]
57. Huynh TT, Nguyen TD, Tan H. A survey on security and privacy issues of blockchain technology. In: 2019 International Conference on System Science and Engineering (ICSSE): IEEE; 2019:362-367. [doi: 10.1109/ICSSE.2019.8823094]
58. Dinh TN, Thai MT. AI and blockchain: a disruptive integration. Computer (Long Beach Calif) 2018 Sep;51(9):48-53. [doi: 10.1109/MC.2018.3620971]
59. Salah K, Rehman MHU, Nizamuddin N, Al-Fuqaha A. Blockchain for AI: review and open research challenges. IEEE Access 2019;7:10127-10149. [doi: 10.1109/ACCESS.2018.2890507]
60. Wang Q, Zhu X, Ni Y, Gu L, Zhu H. Blockchain for the IoT and industrial IoT: a review. Internet Things 2020 Jun;10:100081. [doi: 10.1016/j.iot.2019.100081]
61. Agbo CC, Mahmoud QH, Eklund JM. Blockchain technology in healthcare: a systematic review. Healthcare (Basel) 2019 Apr 4;7(2):56. [doi: 10.3390/healthcare7020056] [Medline: 30987333]
62. Kang M, Park E, Cho BH, Lee KS. Recent patient health monitoring platforms incorporating Internet of Things-enabled smart devices. Int Neurourol J 2018 Jul;22(Suppl 2):S76-S82. [doi: 10.5213/inj.1836144.072] [Medline: 30068069]
63. Yang G, Li C. A design of blockchain-based architecture for the security of electronic health record (EHR) systems. In: 2018 IEEE International Conference on Cloud Computing Technology and Science (CloudCom): IEEE; 2018:261-265. [doi: 10.1109/CloudCom2018.2018.00058]

XSL•FO
RenderX

64. Yuan WX, Yan B, Li W, Hao LY, Yang HM. Blockchain-based medical health record access control scheme with efficient protection mechanism and patient control. Multimed Tools Appl 2023 May;82(11):16279-16300. [doi: 10.1007/s11042-022-14023-3]
65. Zhan W, Chen CL, Weng W, Tsaur WJ, Lim ZY, Deng YY. Incentive EMR sharing system based on consortium blockchain and IPFS. Healthcare (Basel) 2022 Sep 22;10(10):1840. [doi: 10.3390/healthcare10101840]
66. Amanat A, Rizwan M, Maple C, Zikria YB, Almadhor AS, Kim SW. Blockchain and cloud computing-based secure electronic healthcare records storage and sharing. Front Public Health 2022;10:938707. [doi: 10.3389/fpubh.2022.938707] [Medline: 35928494]
67. Kordestani H, Barkaoui K, Zahran W. HapiChain: a blockchain-based framework for patient-centric telemedicine. In: 2020 IEEE 8th International Conference on Serious Games and Applications for Health (SeGAH): IEEE; 2020:1-6. [doi: 10.1109/SeGAH49190.2020.9201726]
68. Uddin M. Blockchain Medledger: hyperledger fabric enabled drug traceability system for counterfeit drugs in pharmaceutical industry. Int J Pharm 2021 Mar;597:120235. [doi: 10.1016/j.ijpharm.2021.120235]
69. Huang Y, Wu J, Long C. Drugledger: a practical blockchain system for drug traceability and regulation. In: 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData): IEEE; 2018:1137-1144. [doi: 10.1109/Cybermatics_2018.2018.00206]
70. Qu Y, Pokhrel SR, Garg S, Gao L, Xiang Y. A blockchained federated learning framework for cognitive computing in Industry 4.0 networks. IEEE Trans Ind Inf 2021 Apr;17(4):2964-2973. [doi: 10.1109/TII.2020.3007817]
71. El Rifai O, Biotteau M, Megdiche I, Ravat F, Teste O. Blockchain-based federated learning in medicine. In: Michalowski M, Moskovitch R, editors. Artificial Intelligence in Medicine: Springer; 2020:214-224. [doi: 10.1007/978-3-030-59137-3_20]
72. Aich S, Sinai NK, Kumar S, et al. Protecting personal healthcare record using blockchain & federated learning technologies. In: 2022 24th International Conference on Advanced Communication Technology (ICACT): IEEE; 2022:109-112. [doi: 10.23919/ICACT53585.2022.9728772]
73. Xu G, Zhou Z, Dong J, Zhang L, Song X. A blockchain-based federated learning scheme for data sharing in Industrial Internet of Things. IEEE Internet Things J 2023 Dec 15;10(24):21467-21478. [doi: 10.1109/JIOT.2023.3298196]
74. Lu Y, Huang X, Dai Y, Maharjan S, Zhang Y. Blockchain and federated learning for privacy-preserved data sharing in Industrial IoT. IEEE Trans Ind Inf 2020 Jun;16(6):4177-4186. [doi: 10.1109/TII.2019.2942190]
75. Lu Y, Huang X, Zhang K, Maharjan S, Zhang Y. Blockchain and federated learning for 5G Beyond. IEEE Netw 2021 Jan;35(1):219-225. [doi: 10.1109/MNET.011.1900598]
76. Fan S, Zhang H, Zeng Y, Cai W. Hybrid blockchain-based resource trading system for federated learning in edge computing. IEEE Internet Things J 2021;8(4):2252-2264. [doi: 10.1109/JIOT.2020.3028101]
77. Kang J, Xiong Z, Niyato D, Xie S, Zhang J. Incentive mechanism for reliable federated learning: a joint optimization approach to combining reputation and contract theory. IEEE Internet Things J 2019 Dec;6(6):10700-10714. [doi: 10.1109/JIOT.2019.2940820]
78. Moudoud H, Cherkaoui S, Khoukhi L. Towards a secure and reliable federated learning using blockchain. In: 2021 IEEE Global Communications Conference (GLOBECOM): IEEE; 2021:1-6. [doi: 10.1109/GLOBECOM46510.2021.9685388]
79. Weng J, Weng J, Zhang J, Li M, Zhang Y, Luo W. DeepChain: auditable and privacy-preserving deep learning with blockchain-based incentive. IEEE Trans Dependable Secure Comput 2021;18(5):2438-2455. [doi: 10.1109/TDSC.2019.2952332]
80. Zhao Y, Zhao J, Jiang L, et al. Privacy-preserving blockchain-based federated learning for IoT Devices. IEEE Internet Things J 2021 Feb;8(3):1817-1829. [doi: 10.1109/JIOT.2020.3017377] [Medline: 35783535]
81. Prokop K, Połap D, Srivastava G, Lin JCW. Blockchain-based federated learning with checksums to increase security in Internet of Things solutions. J Ambient Intell Human Comput 2023 May;14(5):4685-4694. [doi: 10.1007/s12652-022-04372-0]
82. Kalapaaking AP, Khalil I, Yi X. Blockchain-based federated learning with SMPC model verification against poisoning attack for healthcare systems. IEEE Trans Emerg Topics Comput 2024 Jan;12(1):269-280. [doi: 10.1109/TETC.2023.3268186]
83. Shayan M, Fung C, Yoon CJM, Beschastnikh I. Biscotti: a blockchain system for private and secure federated learning. IEEE Trans Parallel Distrib Syst 2021 Jul 1;32(7):1513-1525. [doi: 10.1109/TPDS.2020.3044223]
84. Chen H, Asif SA, Park J, Shen CC, Bennis M. Robust blockchained federated learning with model validation and proof-of-stake inspired consensus. arXiv. Preprint posted online on Jan 9, 2021. [doi: 10.48550/arXiv.2101.03300]
85. Wang Z, Hu Q, Xu M, Zhuang Y, Wang Y, Cheng X. A systematic survey of blockchained federated learning. arXiv. Preprint posted online on Jun 1, 2024. [doi: 10.48550/arXiv.2110.02182]
86. Behera MR, Upadhyay S, Shetty S. Federated learning using smart contracts on blockchains. arXiv. Preprint posted online on Mar 26, 2022. [doi: 10.48550/arXiv.2107.10243]
87. Awan S, Li F, Luo B, Liu M. Poster: a reliable and accountable privacy-preserving federated learning framework using the blockchain. In: CCS '19: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security: Association for Computing Machinery; 2019:2561-2563. [doi: 10.1145/3319535.3363256]

XSL•FO
RenderX

88. Shoker A. Brief announcement: sustainable blockchains through proof of exercise. In: PODC '18: Proceedings of the 2018 ACM Symposium on Principles of Distributed Computing: Association for Computing Machinery; 2018:269-271. [doi: 10.1145/3212734.3212781]
89. Qu X, Wang S, Hu Q, Cheng X. Proof of federated learning: a novel energy-recycling consensus algorithm. IEEE Trans Parallel Distrib Syst 2021 Aug 1;32(8):2074-2085. [doi: 10.1109/TPDS.2021.3056773]
90. Wang X, Garg S, Lin H, Piran M, Hu J, Hossain MS. Enabling secure authentication in Industrial IoT with transfer learning empowered blockchain. IEEE Trans Ind Inf 2021 Nov;17(11):7725-7733. [doi: 10.1109/TII.2021.3049405]
91. Zhu J, Cao J, Saxena D, Jiang S, Ferradi H. Blockchain-empowered federated learning: challenges, solutions, and future directions. ACM Comput Surv 2023 Nov 30;55(11):1-31. [doi: 10.1145/3570953]
92. Bao X, Su C, Xiong Y, Huang W, Hu Y. FLChain: a blockchain for auditable federated learning with trust and incentive. In: 2019 5th International Conference on Big Data Computing and Communications (BIGCOM): IEEE; 2019:151-159. [doi: 10.1109/BIGCOM.2019.00030]
93. Lu Y, Huang X, Zhang K, Maharjan S, Zhang Y. Blockchain empowered asynchronous federated learning for secure data sharing in Internet of Vehicles. IEEE Trans Veh Technol 2020 Apr;69(4):4298-4311. [doi: 10.1109/TVT.2020.2973651]
94. Habib ur Rehman M, Salah K, Damiani E, Svetinovic D. Towards blockchain-based reputation-aware federated learning. In: IEEE INFOCOM 2020 – IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS): IEEE; 2020:183-188. [doi: 10.1109/INFOCOMWKSHPS50562.2020.9163027]
95. Qi J, Lin F, Chen Z, Tang C, Jia R, Li M. High-quality model aggregation for blockchain-based federated learning via reputation-motivated task participation. IEEE Internet Things J 2022 Oct 1;9(19):18378-18391. [doi: 10.1109/JIOT.2022.3160425]
96. Lamken D, Wagner T, Hoiss T, et al. Design patterns and framework for blockchain integration in supply chains. In: 2021 IEEE International Conference on Blockchain and Cryptocurrency (ICBC): IEEE; 2021:1-3. [doi: 10.1109/ICBC51069.2021.9461062]
97. Farooq K, Syed HJ, Alqahtani SO, Nagmeldin W, Ibrahim AO, Gani A. Blockchain federated learning for in-home health monitoring. Electronics (Basel) 2022 Dec 28;12(1):136. [doi: 10.3390/electronics12010136]
98. Chang Y, Fang C, Sun W. A blockchain-based federated learning method for smart healthcare. Comput Intell Neurosci 2021;2021(1):4376418. [doi: 10.1155/2021/4376418] [Medline: 34868289]
99. Ali A, Al-rimy BAS, Tin TT, Altamimi SN, Qasem SN, Saeed F. Empowering precision medicine: unlocking revolutionary insights through blockchain-enabled federated learning and electronic medical records. Sensors (Basel) 2023 Aug 28;23(17):7476. [doi: 10.3390/s23177476]
100. Kumar O, Sudhakaran G, Balaji V, Nhaveen A, Sai Balakrishnan S. Securing health care data through blockchain enabled collaborative machine learning. Research Square. Preprint posted online on May 23, 2023. [doi: 10.21203/rs.3.rs-2205379/v1]
101. Hasan MR, Li Q, Saha U, Li J. Decentralized and secure collaborative framework for personalized diabetes prediction. Biomedicines 2024 Aug 21;12(8):1916. [doi: 10.3390/biomedicines12081916]
102. Moulahi W, Jdey I, Moulahi T, Alawida M, Alabdulatif A. A blockchain-based federated learning mechanism for privacy preservation of healthcare IoT data. Comput Biol Med 2023 Dec;167:107630. [doi: 10.1016/j.compbiomed.2023.107630] [Medline: 37952305]
103. Ramani R, Rosline Mary A, Edwin Raja S, Arun Shunmugam D. Optimized data management and secured federated learning in the Internet of Medical Things (IoMT) with blockchain technology. Biomed Signal Process Control 2024 Jul;93:106213. [doi: 10.1016/j.bspc.2024.106213]
104. Kumar R, Bernard CM, Ullah A, et al. Privacy-preserving blockchain-based federated learning for brain tumor segmentation. Comput Biol Med 2024 Jul;177:108646. [doi: 10.1016/j.compbiomed.2024.108646]
105. Heidari A, Javaheri D, Toumaj S, Navimipour NJ, Rezaei M, Unal M. A new lung cancer detection method based on the chest CT images using federated learning and blockchain systems. Artif Intell Med 2023 Jul;141:102572. [doi: 10.1016/j.artmed.2023.102572]
106. Liang X, Bandara E, Zhao J, Shetty S. A blockchain-empowered federated learning system and the promising use in drug discovery. In: Charles W, editor. Blockchain in Life Sciences: Springer; 2022:113-139. [doi: 10.1007/978-981-19-2976-2_6]
107. Hai T, Zhou J, Srividhya SR, Jain SK, Young P, Agrawal S. BVFLEMR: an integrated federated learning and blockchain technology for cloud-based medical records recommendation system. J Cloud Comp 2022 Dec;11(1):22. [doi: 10.1186/s13677-022-00294-6]
108. Gayathri Hegde M, Shrishti Bekal M, Deepa Shenoy P, Venugopal KR. Preserving privacy and security of electronic health records using blockchain-based federated learning (BFL) framework. In: 2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC): IEEE; 2023:853-859. [doi: 10.1109/R10-HTC57504.2023.10461823]
109. Guduri M, Chakraborty C, Maheswari U, Margala M. Blockchain-based federated learning technique for privacy preservation and security of smart electronic health records. IEEE Trans Consumer Electron 2024 Feb;70(1):2608-2617. [doi: 10.1109/TCE.2023.3315415]
110. Alzubi JA, Alzubi OA, Singh A, Ramachandran M. Cloud-IIoT-based electronic health record privacy-preserving by CNN and blockchain-enabled federated learning. IEEE Trans Ind Inf 2023 Jan;19(1):1080-1087. [doi: 10.1109/TII.2022.3189170]

XSL•FO
RenderX

111. Rahman MA, Hossain MS, Islam MS, Alrajeh NA, Muhammad G. Secure and provenance enhanced Internet of Health Things framework: a blockchain managed federated learning approach. IEEE Access 2020;8:205071-205087. [doi: 10.1109/ACCESS.2020.3037474]
112. Muazu T, Yingchi M, Muhammad AU, Ibrahim M, Samuel O, Tiwari P. IoMT: a medical resource management system using edge empowered blockchain federated learning. IEEE Trans Netw Serv Manage 2024 Feb;21(1):517-534. [doi: 10.1109/TNSM.2023.3308331]
113. Dhasaratha C, Hasan MK, Islam S, et al. Data privacy model using blockchain reinforcement federated learning approach for scalable Internet of Medical Things. CAAI Trans Intel Tech 2024. [doi: 10.1049/cit2.12287]
114. Lakhan A, Mohammed MA, Nedoma J, et al. Federated-learning based privacy preservation and fraud-enabled blockchain IoMT system for healthcare. IEEE J Biomed Health Inform 2023 Feb;27(2):664-672. [doi: 10.1109/JBHI.2022.3165945]
115. Baucas MJ, Spachos P, Plataniotis KN. Federated learning and blockchain-enabled fog-IoT platform for wearables in predictive healthcare. IEEE Trans Comput Soc Syst 2023;10(4):1732-1741. [doi: 10.1109/TCSS.2023.3235950]
116. Raisaro JL, Marino F, Troncoso-Pastoriza J, et al. SCOR: a secure international informatics infrastructure to investigate COVID-19. J Am Med Inform Assoc 2020 Nov 1;27(11):1721-1726. [doi: 10.1093/jamia/ocaa172] [Medline: 32918447]
117. Samuel O, Omojo AB, Onuja AM, et al. IoMT: a COVID-19 healthcare system driven by federated learning and blockchain. IEEE J Biomed Health Inform 2023 Feb;27(2):823-834. [doi: 10.1109/JBHI.2022.3143576]
118. Kumar R, Khan AA, Kumar J, et al. Blockchain-federated-learning and deep learning models for COVID-19 detection using CT imaging. IEEE Sensors J 2021 Jul 15;21(14):16301-16314. [doi: 10.1109/JSEN.2021.3076767]
119. Durga R, Poovammal E. FLED-Block: federated learning ensembled deep learning blockchain model for COVID-19 prediction. Front Public Health 2022 Jun 17;10:892499. [doi: 10.3389/fpubh.2022.892499]
120. Abdel-Basset M, Alrashdi I, Hawash H, Sallam K, Hameed IA. Towards efficient and trustworthy pandemic diagnosis in smart cities: a blockchain-based federated learning approach. Mathematics 2023 Jul 13;11(14):3093. [doi: 10.3390/math11143093]
121. Gupta R, Thakker U, Tanwar S, Obaidat MS, Hsiao KF. BITS: a blockchain-driven intelligent scheme for telesurgery system. In: 2020 International Conference on Computer, Information and Telecommunication Systems (CITS): IEEE; 2020:1-5. [doi: 10.1109/CITS49457.2020.9232662]
122. Chaudjary S, Kakkar R, Gupta R, Tanwar S, Agrawal S, Sharma R. Blockchain and federated learning-based security solutions for telesurgery system: a comprehensive review. Turk J Electr Eng Comput Sci 2022;30(7):2446-2488. [doi: 10.55730/1300-0632.3950]
123. Martínez Beltrán ET, Pérez MQ, Sánchez PMS, et al. Decentralized federated learning: fundamentals, state of the art, frameworks, trends, and challenges. IEEE Commun Surv Tutorials 2023;25(4):2983-3013. [doi: 10.1109/COMST.2023.3315746]
124. Wu L, Ruan W, Hu J, He Y. A survey on blockchain-based federated learning. Future Internet 2023 Dec 12;15(12):400. [doi: 10.3390/fi15120400]
125. Wu H, Wang P. Node selection toward faster convergence for federated learning on non-IID data. IEEE Trans Netw Sci Eng 2022 Sep 1;9(5):3099-3111. [doi: 10.1109/TNSE.2022.3146399]
126. Melis L, Song C, De Cristofaro E, Shmatikov V. Exploiting unintended feature leakage in collaborative learning. In: 2019 IEEE Symposium on Security and Privacy (SP): IEEE; 2019:691-706. [doi: 10.1109/SP.2019.00029]
127. Konečný J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D. Federated learning: strategies for improving communication efficiency. arXiv. Preprint posted online on Oct 30, 2017. [doi: 10.48550/arXiv.1610.05492]
128. Zhang H, Liu J, Jia J, Zhou Y, Dai H, Dou D. FedDUAP: federated learning with dynamic update and adaptive pruning using shared data on the server. arXiv. Preprint posted online on Apr 25, 2022. [doi: 10.48550/ARXIV.2204.11536]
129. Antunes RS, André da Costa C, Küderle A, Yari IA, Eskofier B. Federated learning for healthcare: systematic review and architecture proposal. ACM Trans Intell Syst Technol 2022 Aug 31;13(4):1-23. [doi: 10.1145/3501813]
130. Cui L, Su X, Ming Z, et al. CREAT: blockchain-assisted compression algorithm of federated learning for content caching in edge computing. IEEE Internet Things J 2022 Aug 15;9(16):14151-14161. [doi: 10.1109/JIOT.2020.3014370]

Abbreviations

AI: artificial intelligence
BCFL: blockchain-based federated learning
BFLPD: blockchain-based federated learning for pandemic diagnosis
EHR: electronic health record
EMR: electronic medical record
FBCLC-Rad: Federated Learning–Enabled Blockchain CapsNets Lung Cancer Radiologist
FedAvg: Federated Averaging algorithm
FL: federated learning
FL-BETS: Federated Learning–Based Blockchain-Enabled Task Scheduling
FLED-Block: Federated Learning–Ensembled Deep Learning Blockchain Model

XSL•FO
RenderX

GDPR: General Data Protection Regulation
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
gRPC API: Google Remote Procedure Calls – Application Programming Interface
HIPAA: Health Insurance Portability and Accountability Act
***ICD-10*:** *International Statistical Classification of Diseases, Tenth Revision*
***ICD-10-CM*:** *International Classification of Diseases, Tenth Revision, Clinical Modification*
IoMT: Internet of Medical Things
IoT: Internet of Things
ML: machine learning
non-IID: non-independent and identically distributed
ODMSM: Optimized Data Management and Secured Federated Learning
PBFT: Practical Byzantine Fault Tolerance
PoS: Proof of Stake
PoW: Proof of Work
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
REST-API: Representational State Transfer – Application Programming Interface
RPC: Remote Procedure Calls
SMPC: Secure Multiparty Computation
ZKP: zero-knowledge proof

Edited by J Sarvestan; submitted 14.Jun.2025; peer-reviewed by S Fitzek, S Poruran; accepted 27.Nov.2025; published 19.Feb.2026.

Please cite as:
Wang X, Xie Y, Chen X, Yang J, Li R, Gao W, Yan Z, Zhou H, Ye Z
Securing Federated Learning With Blockchain in the Medical Field: Systematic Literature Review
J Med Internet Res 2026;28:e79052
URL: https://www.jmir.org/2026/1/e79052
doi:10.2196/79052

© Xudong Wang, Yi Xie, Xiaoliang Chen, Jiaming Yang, Ruiyuan Li, Weihang Gao, Zineng Yan, Hong Zhou, Zhewei Ye. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Inequalities in the Use of eHealth Services in European Public Health Care Systems: Systematic Review of Observational Studies

Garazi Monasterio[1,2], MSc; Marcos José Fernández-López[2], MSc; Erika Valero[1,2], PhD; Unai Martin[2,3,4], PhD; Amaia Ayala-García[2,3,4], PhD

[1]Department of Nursing I, Faculty of Medicine and Nursing, University of the Basque Country (UPV/EHU), Barrio Sarriena s/n, Leioa, Spain

[2]Research Group on Social Determinants of Health and Demographic Change-OPIK, University of the Basque Country (UPV/EHU), Leioa, Spain

[3]Department of Sociology and Social Work, Faculty of Social Sciences and Communication, University of the Basque Country (UPV/EHU), Leioa, Spain

[4]Biobizkaia Health Research Institute, Barakaldo, Spain

Corresponding Author:
Garazi Monasterio, MSc
Department of Nursing I, Faculty of Medicine and Nursing, University of the Basque Country (UPV/EHU), Barrio Sarriena s/n, Leioa, Spain

Abstract

Background: European public health care systems are expanding eHealth tools such as teleconsultations, online appointment bookings, and electronic health records to improve efficiency and access to health care. However, their use depends on factors such as digital skills and internet access, which are unequally distributed across socioeconomic and demographic determinants. Most existing evidence on these inequalities is qualitative or outside universal health care systems.

Objective: This systematic review aims to synthesize quantitative evidence on social inequalities in access to and use of eHealth services within European public health care systems. Specifically, we sought to identify which social determinants were most consistently associated with unequal use of online appointment booking, teleconsultations, electronic health records, and eHealth portals, across major social determinants of health.

Methods: A systematic search was conducted across PubMed, Scopus, Web of Science, and PsycINFO for studies published in English or Spanish between 2015 and October 2025. Eligible quantitative studies included adults (≥18 years) using public health care systems in European countries. The primary outcome was differential access to or use of eHealth tools by social determinants in any level of care. Screening and data extraction were independently performed by 3 reviewers using Rayyan, resolving disagreements through consensus. Data extracted covered study design, population, eHealth tools, social determinants, and outcomes. Risk of bias was evaluated using Joanna Briggs Institute tools. Due to study heterogeneity in digital tools and inequality dimensions, results were synthesized narratively by tool type and social inequality factors. Point estimates and 95% CIs were extracted when available.

Results: Of the 2366 records retrieved, 18 observational studies met the inclusion criteria: 13 cross-sectional, 3 prevalence, 1 retrospective cohort, and 1 ecological cohort. Publication output increased from 2020 onward, mostly driven by cross-sectional studies from northern and western Europe. Findings revealed consistent social gradients in eHealth use: older adults, individuals with lower educational or socioeconomic level, ethnic minorities, and those with limited digital skills or poorer health were less likely to use eHealth tools. Most studies were rated as high quality (78%), and the remainder as moderate, heterogeneity in designs, outcomes, and populations may limit generalizability.

Conclusions: Digital transformation in European public health systems has not benefited all groups equally. This review highlights persistent social inequalities in the use of key digital health tools. While many included studies were of high quality, heterogeneity in study designs, populations, and outcomes, as well as risk of bias, limits causal inference and the direct translation of findings into policy and practice. The findings nonetheless reveal systematic patterns of exclusion that are highly relevant for policy. Emphasizing an intersectional approach and standardizing measures of digital access will be essential to develop effective, equity-focused policies that ensure inclusive digital health services for all.

Trial Registration: PROSPERO CRD420251015756; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251015756

(J Med Internet Res 2026;28:e81841) doi:10.2196/81841

XSL•FO
RenderX

KEYWORDS

eHealth; digital health; health services accessibility; delivery of health care; health inequities; health care disparities; social determinants of health; public health infrastructure

Introduction

Health systems worldwide are undergoing a strong shift toward eHealth [1,2], also referred to as digital health, telehealth, or telemedicine. This transition, long encouraged by institutions such as the World Health Organization (WHO) [3] to achieve the 2030 Sustainable Development Goals, was accelerated by the COVID-19 pandemic [4,5].

One of the earliest and most widespread applications of eHealth is the online appointment booking, teleconsultations, access to electronic health records (EHRs), and integrated digital health portals that bundle multiple services in a single interface [6,7]. These 4 modalities constitute the most common digital interaction points with health care services across European public health systems and form the conceptual foundation for evaluating digital health access and usage in this review [8]. Beyond offering advantages for health care systems and for patients [9,10], these tools have also been identified as potential levers for improving access in underserved settings, such as rural or remote areas [11].

However, there is comparatively less research on how these same tools may reinforce or even exacerbate existing social inequalities in access and usage. As Western et al [12] have conceptualized, these inequalities manifest across 3 interrelated levels of the digital divide: (1) access to internet connectivity and digital devices; (2) digital skills and literacy, trust in technology, and willingness to adopt new platforms; and (3) disparities in actual outcomes or health improvements derived from digital tool use. Only those who overcome these first 2 levels can potentially benefit, leading to the third digital health divide. Breakdowns at any of these levels can cascade to reinforce exclusion at the next, perpetuating digital and health inequities. Quantitative studies, many outside universal health care settings, point out that older adults, people with lower socioeconomic status or education [13], migrated population, ethnic minorities, rural population, or in low-skilled jobs [14], individuals with impairments [15], and other socially disadvantaged groups often lack these prerequisites [16], revealing measurable gaps in digital health use [17,18]. Consequently, digital health strategies may risk amplifying pre-existing health disparities [19]. In addition, digital divide measurement methods and indicators vary widely, from composite indexes (aggregated scores combining multiple dimensions of access and use) to simple usage metrics (such as number of logins or appointments booked) [18,20]. Yet, the absence of harmonized indicators makes it difficult to compare findings across countries, especially in health care systems with very different structures.

Within Europe's public health care systems—across all levels of care—there remains scant data quantifying inequalities in basic digital functions such as booking appointments, accessing health portals, or having teleconsultations.

Despite these concerns, public health systems across Europe (and beyond) are investing heavily in digital care models [21]. While this expansion signals progress, it also highlights the lack of robust monitoring of equity impacts. Most existing studies emphasize qualitative evidence, which provides valuable perspectives but limited measurement of scale [22]. Quantitative research is less common, and when available, it rarely disaggregates results by key social determinants, leaving important gaps in understanding how digital health inequalities unfold across vulnerable groups.

In recent years, this issue has attracted renewed policy and research attention, reflecting a growing recognition that digital transformation alone does not guarantee equitable access [23]. Several recent analyses and policy reports have documented substantial advances in national eHealth strategies and portal deployment across Europe, while noting that evaluation frameworks for equity and inclusion remain limited [24,25]. At the same time, comparative assessments by international organizations emphasize the need for harmonized indicators and monitoring systems to assess how digitalization affects different population groups within public health care [26]. This evolving evidence base reinforces the importance of systematically consolidating quantitative findings to clarify where inequalities persist and how they are measured within European public health systems.

In this context and considering that the WHO's Global Strategy on Digital Health 2020 - 2025 (extended to 2027) emphasizes that eHealth should enhance universal access, quality, efficiency, and equity [8], there is an urgent need to identify who is falling behind. To date, no systematic review has synthesized quantitative evidence on these digital inequalities across European public health care systems.

The objective of this systematic review is, therefore, to synthesize the available quantitative evidence on inequalities in access to and use of core eHealth services—including online appointment booking, teleconsultations, and EHRs—within European public health care systems between 2015 and 2025, examining how key social determinants, such as age, gender, socioeconomic status, education, migration background, and digital literacy, are associated with the use of these services.

Methods

Overview

A systematic review was conducted to identify studies examining inequalities in the use of eHealth technologies to access health care, provided by public health care systems, in relation to social determinants of health in Europe. This review was carried out in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [27] guidelines to ensure a rigorous and transparent evaluation of the review process. For details of items included in the PRISMA checklist, please see Checklist 1. The protocol was

XSL•FO
RenderX

registered in the PROSPERO database (registration number: CRD420251015756).

Information Sources and Search Strategy

A systematic search was performed in PubMed, Scopus, PsycINFO, and Web of Science up to April 2025. The search strategy combined controlled vocabulary and free-text terms related to digital health (eg, "telemedicine," "eHealth," "digital health," and "patient portal"), social inequalities (eg, "health disparities," "digital divide," "inequality," and "equity"), social determinants and user characteristics (eg, "age," "gender," "ethnic minorities," "disability," and "socioeconomic"), health care access and use (eg, "access," "use," "user," and "nonuser") and geographical terms referring to Europe and individual European countries (eg, "Europe," "United Kingdom," "Germany," and "Spain"). Truncation and Boolean operators (AND, OR, NOT) were used to combine these concepts and optimize sensitivity and specificity. The search strategy was developed collaboratively by the research team, without consultation with an information scientist. Complete search strategies and term lists are available in the Multimedia Appendix 1.

The search process and its reporting followed the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) guidelines to ensure transparency, completeness, and reproducibility of each search component [28]. In addition, a backward citation search (snowballing) was conducted from the reference lists of the included studies. The literature search was rerun on October 20, 2025, to capture any recent publications since the initial search.

Inclusion and Exclusion Criteria

Studies were eligible for inclusion if they met the following criteria: (1) study population focused on adults aged 18 years or older. Studies that included younger populations but allowed for interpretation of results in adults more than 18 years were also included; (2) assessed any type of eHealth technology designed to connect individuals to public health systems; (3) reported quantitative data related to use of digital health services provided by public health systems in association with social determinants of health; (4) were conducted in European countries, defined as member states of the European Union, the European Economic Area (Iceland, Liechtenstein, and Norway), the United Kingdom, and Switzerland. This definition was selected to ensure conceptual and structural comparability of health care systems with public health care systems coverage and established digital health infrastructure; and (5) were peer-reviewed studies published in English or Spanish between January 2015 and April 2025.

The following exclusion criteria were applied: (1) studies focusing exclusively on populations using private health care services; (2) focusing on the effects of the COVID-19 pandemic on eHealth usage; (3) focusing on health apps designed for preventive purposes, such as interventions, general health information provision, self-care, or supporting the management of specific conditions; and (4) qualitative studies, study protocols, conference abstracts, theses, editorials, opinion pieces, and systematic reviews.

Data Management and Study Selection

Study screening was conducted using the Rayyan web-based tool [29]. No automation tools were used in this process. Two researchers independently screened titles and abstracts (GM and AA), followed by full-text review of potentially eligible studies. Disagreements were resolved through a third reviewer (MF). Study authors were contacted when additional information or clarification was required.

Data Extraction

Data extraction was carried out collaboratively by 3 researchers (GM, AA, and MF) and reviewed by the other team members to ensure accuracy and consistency. The following information was extracted: authorship, year of publication, country, study design, population characteristics, type of eHealth technology, outcome, social determinants of health included, and key quantitative findings and their narrative description. In addition, eHealth tools identified in the studies were regrouped into 4 categories according to their functionality in (1) eHealth portals—which typically integrate multiple functionalities into a single tool. These include services such as prescription renewal, appointment reminders, asynchronous messaging with health care professionals, access to test results, and in some cases, teleconsultation features or visualization of EHRs; (2) EHRs—this category includes those tools that allow users to consult their medical results, diagnoses, clinical notes, and other information recorded in their health record. These functionalities may be, and often are, integrated into digital portals managed by public health services or specific hospital apps. Notable examples include platforms such as MyCare, myHealth@QEHB, or the Care Information Exchange in the United Kingdom; (3) telemedicine and remote primary care—understood as the provision of direct health care through digital channels. It includes video medical consultations, direct-to-consumer digital care models, as well as the remote provision of primary care; (4) online appointment booking. This classification enabled a more nuanced analysis of inequalities based on the specific functionality assessed in each study. To ensure clarity in the tables, this categorization is represented numerically (1-4).

Given the variability in terminology and definitions across included studies, we compiled a comparative table showing how each study defined key determinants and eHealth usage, which is provided in Multimedia Appendix 2.

Risk of Bias Assessment

Risk of bias was assessed using the appropriate critical appraisal tools developed by the Joanna Briggs Institute (JBI), selected based on each study's methodological design (eg, cross-sectional and cohort) [30]. The assessment was conducted at the study level rather than at the outcome level. This approach was chosen because the included studies were primarily observational and reported heterogeneous outcomes, making study-level assessment the most consistent and feasible strategy. The appraisal focused on key domains affecting internal validity, including participant selection, measurement of exposures and outcomes, and control of confounding factors. Studies were

XSL•FO
RenderX

classified into 3 quality categories based on the proportion of affirmative responses to the JBI appraisal items: high quality (≥70%), moderate quality (40% - 69%), and low quality (≤39%) [31]. A detailed summary of the risk of bias assessment is provided in Multimedia Appendix 3 [32-49], using the corresponding JBI Critical Appraisal Tool.

At least 2 reviewers assessed the risk of bias independently. Discrepancies were resolved through discussion and consensus, involving 3 researchers (GM, AA, and MF) during this phase.

Data Synthesis and Effect Measures

The effect measures reported in the studies varied. Some studies presented descriptive comparisons between population subgroups (such as percentages or proportions), while others reported adjusted measures of association derived from regression models, such as odds ratios (OR) with corresponding CIs. When available, information was classified according to the type of eHealth service analyzed.

In addition, some studies applied more advanced equity metrics, such as the concentration index and the horizontal inequity index. The concentration index summarizes whether use of services is disproportionately concentrated among socioeconomically advantaged or disadvantaged groups. The horizontal inequity index adjusts for differences in health care needs (eg, morbidity), isolating inequities that persist beyond expected variations in medical necessity. A few studies also used decomposition analysis to examine the contribution of individual determinants (eg, education, income, and employment) to overall inequality, and indirect standardization to account for differences in health needs when comparing groups.

Due to the heterogeneity in study designs, eHealth technologies evaluated, and inequality dimensions analyzed, a narrative synthesis was therefore conducted to summarize the findings. Results were presented in evidence tables and described narratively in the text. Studies with comparable outcomes were grouped, as explained in the “Data Extraction” section.

Results

Study Selection

The initial search identified 2366 records from 4 databases. Of these, ultimately, 18 studies [32-49] were included in the systematic review. A PRISMA 2020 flow diagram detailing the selection process is presented in Figure 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram of study selection process.

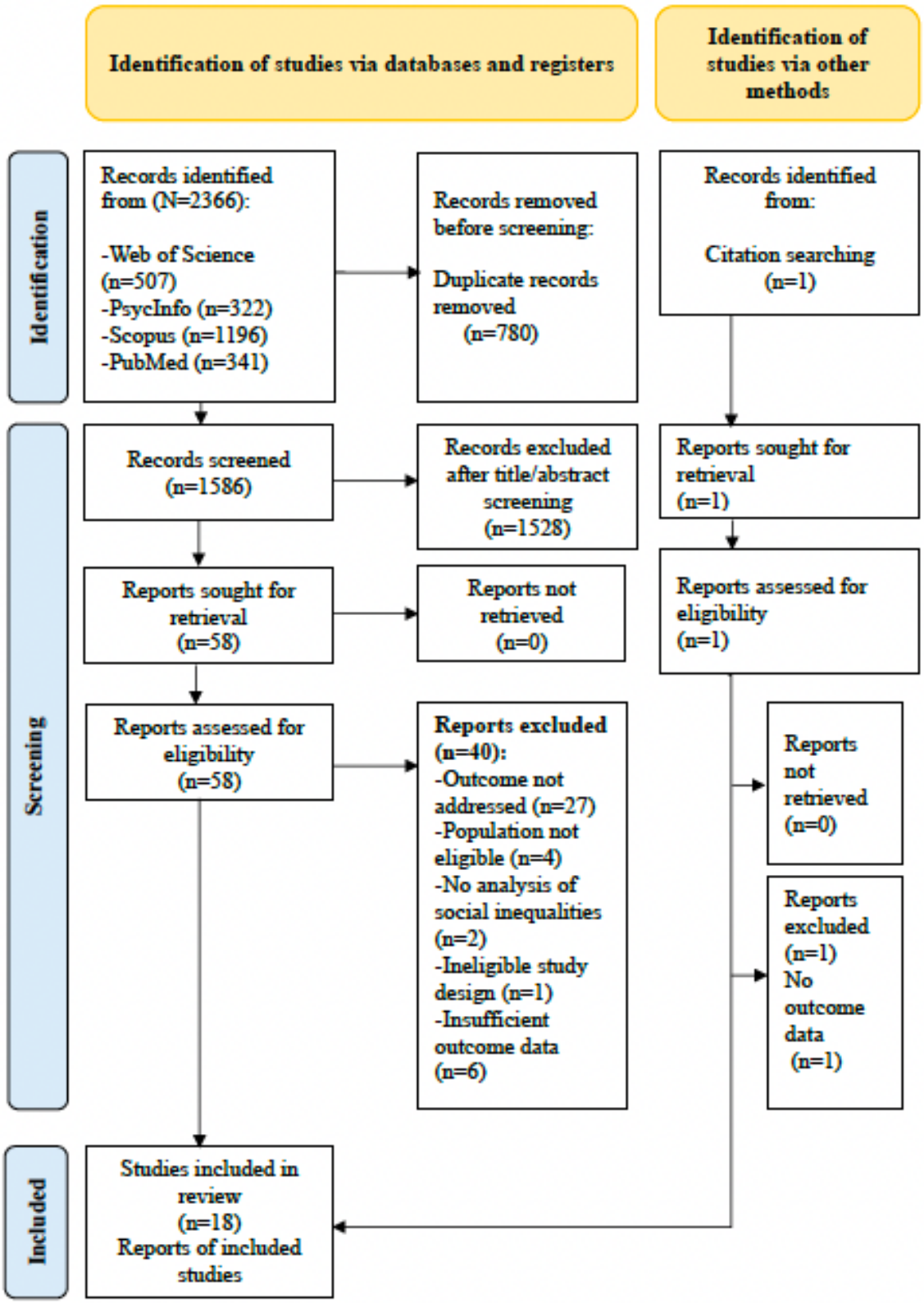

Characteristics of the Studies Included

The included studies were published between 2017 and 2025. A general upward trend in the number of publications was observed over time, with a noticeable increase in high-quality studies in the most recent years. The distribution of included studies per year according to their methodological quality category is illustrated in Figure 2, where the trend line represents the total number of studies published per year, and the green and yellow “x”’s represent the quality of the individual studies per year.

XSL•FO
RenderX

Figure 2. Publications per year by quality category.

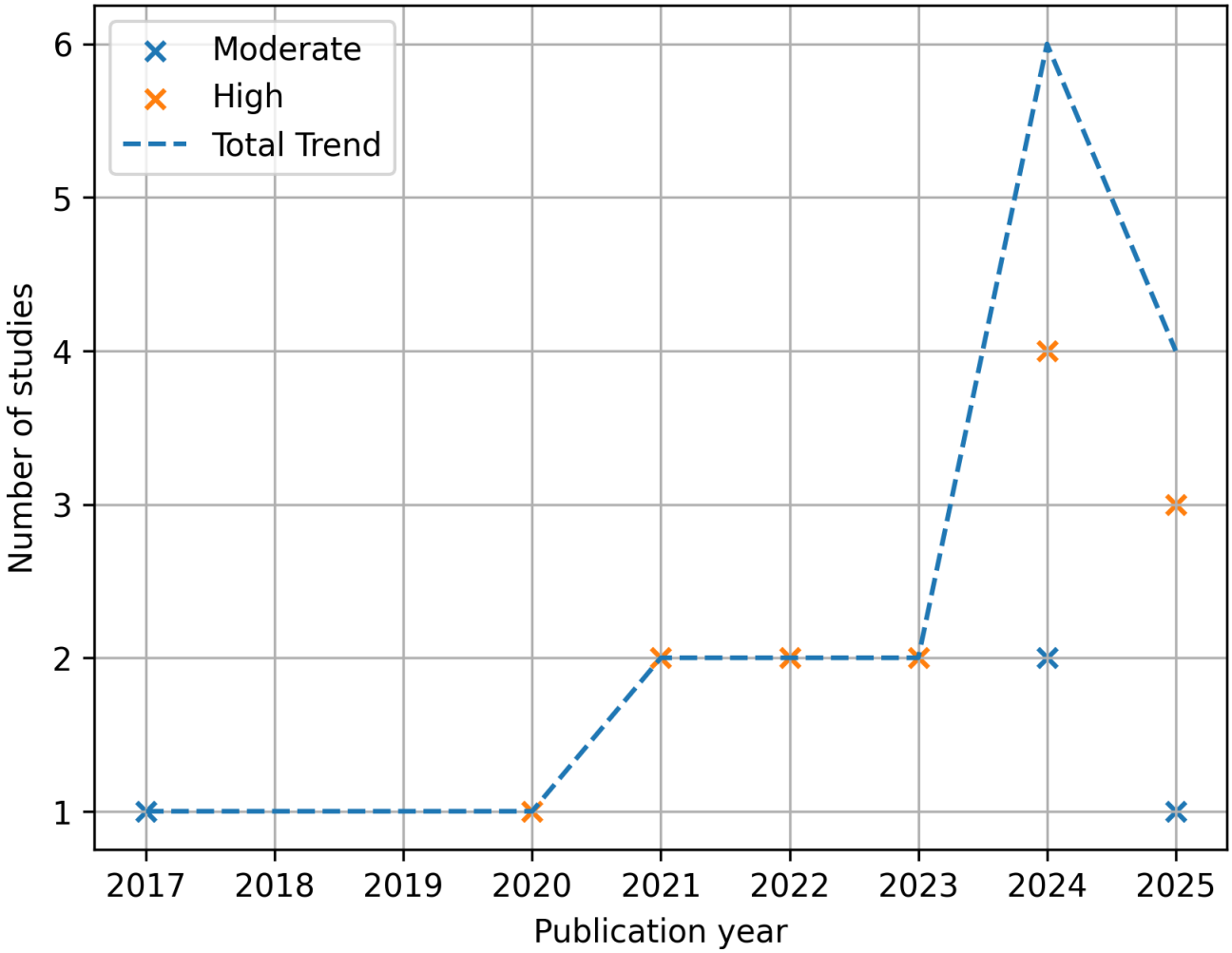

The most frequent study design was cross-sectional observational, accounting for 72% (13/18) of the studies [32-37,39-42,45,46,49], followed by prevalence studies at 17% (3/18) [38,43,47], with one retrospective cohort [44] and one ecological cohort study each representing 6% (1/18) [48]. The 18 studies [32-49] were conducted across 10 European countries, with the majority from Northern and Western Europe. Sweden had the highest number of studies, representing 28% (5/18) [33,36,37,44,45], followed by England at 22% (4/18) [35,46,48,49] and Finland at 11% (2/18) [32,41]. Other countries contributed one study each: Spain [42], Norway [47], Germany [40], Iceland [38], Denmark [43], the United Kingdom (distinct from England) [34], and one pan-European study [39]. Sample sizes ranged from under 438 to 1,991,995 participants. They varied from small studies such as Hörhammer et al [32] in Finland (n=438) to very large population-based datasets such as Dahlgren et al [33] in Sweden (n=1,991,995). Specifically, 17% (3/18) of studies had sample sizes below 1000 [32,34,35], 44% (8/18) ranged between 1000 and 10,000 [36-43], 28% (5/18) exceeded 10,000 participants [33,44-47], and 11% (2/18) did not report sample size [48,49]. Most studies focused on the general adult population aged 18 years and older.

Risk of Bias Assessment

According to the risk of bias assessment, 22% (4/18) of the studies [40,42,43,48] were classified as of moderate quality, while the remaining 78% (14/18) were rated as high quality [32-39,41,44-47,49]. No studies were considered to have low methodological quality, suggesting that the overall quality of the evidence included was generally high.

The eHealth tools analyzed were regrouped into 4 functional categories as described in the “Methods” section: (1) eHealth portals, (2) EHRs, (3) telemedicine and remote primary care, and (4) online appointment booking. Of the 18 included studies [32-49], 56% (10/18) analyzed eHealth portals [32,34,35,38-41,43,48,49], 33% (6/18) examined EHRs [37,41,42,46-48], 22% (4/18) addressed telemedicine and remote primary care [33,36,44,45], and only 11.1% (2/18) focused on online appointment booking [41,48]. Figure 3 presents the distribution of eHealth tool categories by publication year, highlighting a growing diversification in digital health functionalities over time. Studies appear more than once when assessing more than one tool usage.

XSL•FO
RenderX

Figure 3. Assessment of eHealth tools per year. EHR: electronic health record.

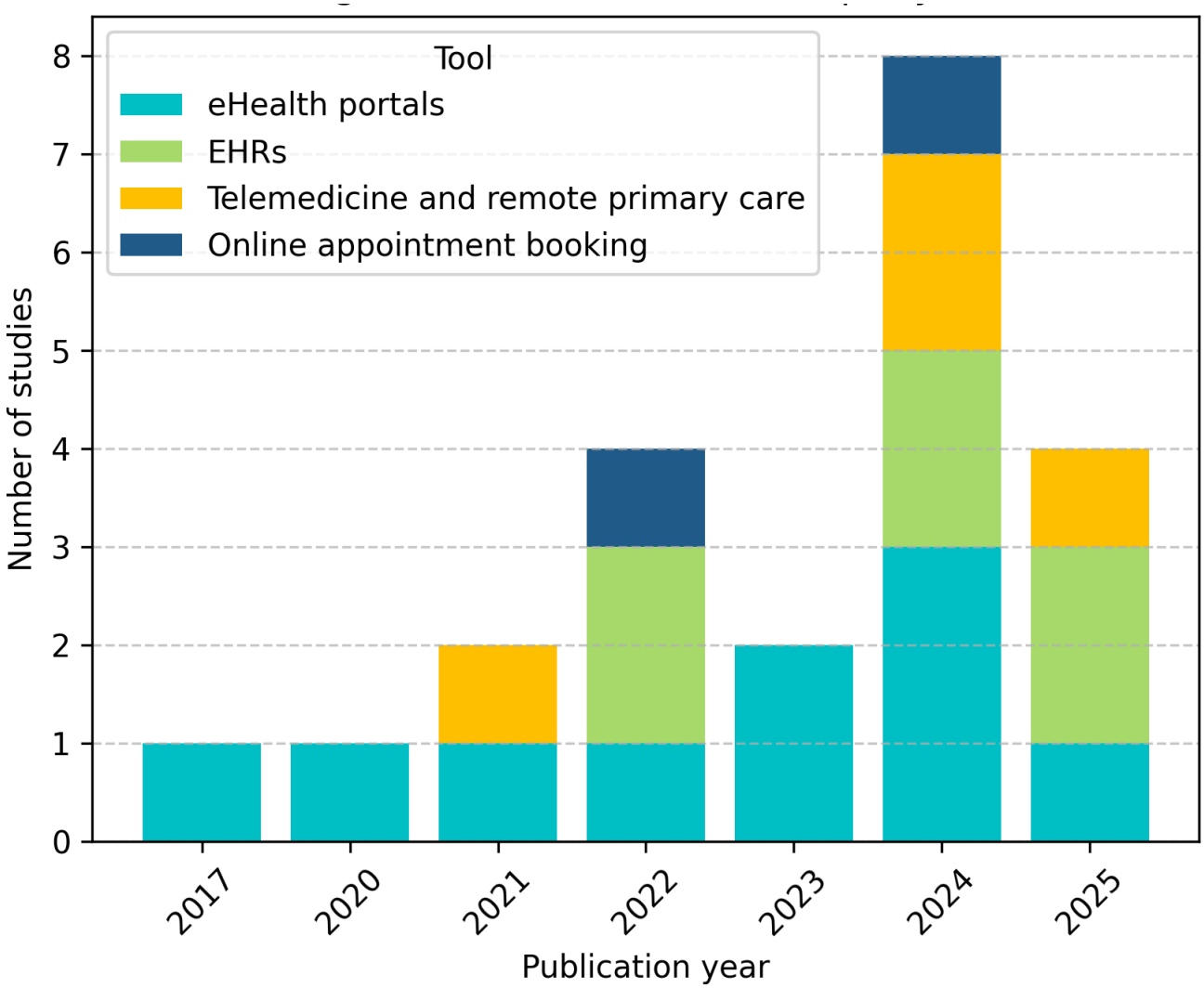

All 18 studies reviewed examined social inequalities in the use of eHealth tools, with age and gender being the most analyzed determinants [32-49]. A second major block of variables related to socioeconomic position, especially social class [39,43,44], education [33,34,36,37,42,45], employment [32,36,40,45], income [33,45], and deprivation [35,46,48,49], often using both individual-level and area-based indicators. Fewer studies addressed cultural-related factors such as ethnicity [34,46,48,49], country of birth [33,45], nationality [42], and language [35,40,46], though these were sometimes used as proxies for integration and accessibility. Determinants related to residential context (eg, urban vs rural [40,45,49] and distance to care [33]) and family structure (eg, marital status [37,39,40], household size [39], and parity [35]) were included less frequently. Health status and clinical complexity were operationalized in diverse ways, including self-rated health [34,36,37], chronic conditions [32,33,48,49], diagnoses or number of specialties involved [45,46] or resource use band [44]. Finally, several studies assessed digital competence [41,42], literacy [32,34], and internet habits [36], highlighting the role of skills and confidence in shaping access to digital health care. Figure 4 presents a heatmap summarizing the extent to which each social determinant was not only examined but also found to be associated with inequalities in the use of specific eHealth tools. Each cell displays the proportion of studies that reported significant disparities out of those that analyzed a given determinant for each tool. For example, a value of “5/5” under “gender” and “EHRs” indicates that all 5 studies assessing gender differences in the use of EHRs reported significant inequalities by this determinant. This visual summary highlights both the breadth of research on each determinant and the consistency of evidence across different tools. eHealth portals were the most frequently studied, covering a wide range of social factors; however, the findings were somewhat mixed, with inequalities observed in some areas (eg, age and education) but not others (eg, gender and ethnicity). In contrast, studies on EHRs and telemedicine and remote primary care, though less numerous, tended to yield more consistent evidence of social inequalities, particularly along the lines of age, gender, and socioeconomic status. Finally, online appointment booking was the least explored modality, with very limited evidence available.

XSL•FO
RenderX

Figure 4. Evidence of social inequalities in eHealth tool use. EHR: electronic health record.



A summary of the main characteristics and key findings of the included studies is presented in Table 1. A more detailed version is available in Multimedia Appendix 4 [32-49].

XSL•FO
RenderX

Table . Evidence table summarizing the main information of each study included in the systematic review, ordered by study design.

Author (year); Country	Design, Sample, Population, and Statistical analysis	eHealth Cat[a] and study Outcome	SDH[b]	Description of main finding of inequalities in the studies
González-Cacheda et al [42] (2025); Spain	• Cross-sectional • Population present in the Health Barometer 2022 (n=7454) • Univariate logistic regressions	(2) Use of medical records via internet	• Age • Gender • Education • Nationality • Social class • Device use	• Older adults: higher awareness (OR[c] 0.86) but lower effective use than younger and middle-aged. • No gender differences. • Higher education (OR 1.41) and digital experience (OR 1.51) associated with greater use. • Spanish nationals are more likely to use digital medical records (OR 1.45). 95% CI not reported in the primary study.
Hörhammer et al [32] (2025); Finland	• Cross-sectional • Patients from a Finnish mental health and substance abuse unit (n=438) • Univariate and multivariate logistic regressions	(1) Use of digital services	• Age • Gender • Health confidence • Employment • Duration of care	• Older adults (≥65 years; OR 0.40, 95% CI 0.21 - 0.75) less likely to use digital services. • Low health confidence is associated with reduced use (OR 0.62, 95% CI 0.45 - 0.86).
Knöchelmann et al [40] (2024); Germany	• Cross-sectional • Adults (≥ 18 years) from the HeReCa panel (n=1821) • Descriptive analysis, latent class analysis, and multinomial logistic regression	(1) Use of digitalized health care services	• Age • Gender • Education • Language • Employment • Place of residence • Marital status	• No significant associations for women (OR 0.78, 95% CI 0.56 - 1.10), age (OR 0.88, 95% CI 0.71 - 1.08), or marital status (married vs others; OR 0.77, 95% CI 0.53 - 1.13). • Higher education is linked to greater odds of rejecting participation (OR 1.23) but lower odds of active use (OR 0.68). 95% CI not reported in the primary study. • Lower educational attainment (OR 0.67, 95% CI 0.47 - 0.95) and unemployment (OR 0.57, 95% CI 0.38 - 0.84) associated with reduced active use. • Number of previous illnesses not associated with use.

XSL•FO
RenderX

Author (year); Country	Design, Sample, Population, and Statistical analysis	eHealth Cat[a] and study Outcome	SDH[b]	Description of main finding of inequalities in the studies
Söderberg et al [36]; Sweden	• Cross-sectional • Adults (≥ 18 years) residing in Sweden (n=2716) • Multivariate logistic regression	(3) Use of digital primary care	• Age • Gender • Occupation • Education • Internet habits • Self-rated health	• Younger age associated with higher likelihood of seeking digital primary care (95% CI not reported in the primary study). • University education (OR 1.41, 95% CI 1.19 - 1.67) and daily internet use (OR 3.21, 95% CI 1.30 - 7.90) linked to greater use. • Retired individuals had lower odds than those working (OR 0.68, 95% CI 0.46 - 1.02; borderline nonsignificant).
Wilkens et al (2024) [45]; Sweden	• Cross-sectional • Patients ≥18 years (n=726,087) • Concentration index and curves; decomposition analysis; indirect standardization; horizontal inequity index	(3) Use of digital primary care	• Age • Gender • Education • Morbidity • Country of birth • Geographic region • Employment	• Digital care users were younger and more often women. • Low-income patients used more office-based visits, whereas high-income patients had more digital contacts; income and employment explained most inequalities in office visits. • Pro-rich inequality in digital contacts was partly explained by higher education and being born in Sweden; digital users were generally healthier.
Muli et al (2024) [37]; Sweden	• Cross-sectional • Patients ≥18 years and guardians of minors living in Region Stockholm who had a consultation with a physician (n=3421) • Multivariate logistic regressions	(2) Access to patient-accessible electronic health records	• Age • Gender • Education • Marital status • Health status	• Younger age associated with higher likelihood of reading records (OR 0.97, 95% CI 0.95 - 0.98). • Women are more likely to have read their electronic health records (77% readers overall). • Partnered individuals (vs single; OR 0.60, 95% CI 0.44 - 0.99) and those with higher education were more likely to be readers. • No differences were observed by health status.
Zhang et al [49] (2023); England		(1) Use of National Health Service (NHS) app (NHS Digital) and a primary care portal	• Age • Deprivation • Ethnicity • Residence place • Long-term condition	

XSL•FO
RenderX

Author (year); Country	Design, Sample, Population, and Statistical analysis	eHealth Cat[a] and study Outcome	SDH[b]	Description of main finding of inequalities in the studies
	• Cross-sectional • Populations of primary care practices in National Health Service (no n reported) • Multivariate linear regression			• Practices with more male or chronically ill patients had lower usage (P<.01). • Strong socioeconomic gradient: all more deprived quintiles showed lower digital health use than the least deprived (Q1; difference = –2.047 units; P<.001). • Higher use in practices with more White and younger patients (15 - 34 years; +10.39% registration, P<.05) and in larger practices (P<.05).
Pierce et al [35] (2023); England	• Cross-sectional • Pregnant women booked into UCLH[d] for their initial antenatal appointment in February 2022, identified through the EPIC platform (n=636) • Descriptive	(1) Use of MyCare, an electronic patient portal	• Age • Parity • Language • Deprivation	• Non-users of MyCare were younger (mean age 30) and showed lower engagement with higher parity (mean parity =0.94, 1.47, 1.78 for high-, low-, and nonusers). • Women whose first language was not English were more common among non-users (48.7%), who also had lower socioeconomic status (mean SDI[e]=3.72).
Heponiemi et al [41] (2022); Finland	• Cross-sectional • Sample from the Population Register of Finland • >20 years (n=4495) • Multivariate logistic regression	(1,2, and 4) Use of online health services	• Age • Gender • Digital competence	• Use of all digital services declined sharply after age 60, except for appointment booking. • Women used online services 8% more than men. • Good digital competence strongly increased use: test results (OR 12.61, 95% CI 8.52 - 18.64), prescription renewal (OR 8.82, 95% CI 6.15 - 12.64), appointment scheduling (OR 10.91, 95% CI 7.24 - 16.44), and online appointments (OR 6.48, 95% CI 0.93 - 45.12; not significant).
Chapman et al [46] (2022); England		(2) Sign up and activation of myHealth @QEHB, a hospital-based Personal Health Record		

XSL•FO
RenderX

Author (year); Country	Design, Sample, Population, and Statistical analysis	eHealth Cat[a] and study Outcome	SDH[b]	Description of main finding of inequalities in the studies
	• Cross-sectional • Adults ≥18 years receiving hospital outpatient care (n=28,637) • Descriptive, univariate, and multivariate logistic regression		• Age • Gender • Deprivation • Ethnicity • Interpreted need • Number of hospital specialties	• Activation highest among adults aged 35 - 54 years; lower for ages 16 - 34 years (OR 0.80, 95% CI 0.70 - 0.91) and lowest for ≥75 years (OR 0.39, 95% CI 0.32 - 0.47). • Males are less likely to activate their Personal Health Record (OR 0.85, 95% CI 0.78 - 0.94) but more likely to sign up (OR 1.10, 95% CI 1.04 - 1.16); activation is 3 times higher in the least versus the most deprived areas (OR 2.99, 95% CI 2.40 - 3.71). • Asian (aOR[f] 0.61, 95% CI 0.53 - 0.71), Black (aOR 0.45, 95% CI 0.36 - 0.56), and mixed ethnic groups (aOR 0.77, 95% CI 0.60 - 0.97) less likely to activate than White patients; those not needing an interpreter had higher activation (OR 3.16, 95% CI 1.96 - 5.09). • Number of clinical specialties associated with sign-up (aOR 2.54, 95% CI 2.30 - 2.82) but not activation (aOR 0.88, 95% CI 0.79 - 0.97).
Neves et al [34] (2021); United Kingdom	• Cross-sectional • Patients >18 years of hospitals and primary care in London, registered in the Care Information Exchange (n=650) • Descriptive, univariate, and multivariate logistic regression	(1) Use of Care Information Exchange, a patient portal containing patient information	• Age • Gender • Education • Ethnicity • Health status eHEALS[g] score	• Higher education associated with increased odds of use: undergraduate and professional (aOR 1.58, 95% CI 1.04 - 2.39) and postgraduate and higher (aOR 2.38, 95% CI 1.42 - 4.02). • Greater digital literacy (>30) strongly predicted use (aOR 2.96, 95% CI 2.02 - 4.35). • Good health status linked to lower odds of use (aOR 0.58, 95% CI 0.37 - 0.91).
Dahlgren et al [33] (2021); Sweden		(3) Use of telemedicine to provide traditional primary care		

XSL•FO
RenderX

Author (year); Country	Design, Sample, Population, and Statistical analysis	eHealth Cat[a] and study Outcome	SDH[b]	Description of main finding of inequalities in the studies
	• Cross-sectional • Adults (≥18 years) residing in Stockholm County and registered with a publicly funded Primary Health Care provider (n=1,991,995) • Descriptive and multivariate logistic regression		• Age • Gender • Education • Income • Country of birth • Diagnosis of chronic conditions • Primary health care accessibility • Distance to primary health care	• Greater telemedicine use among women (OR 1.60, 95% CI 1.58 - 1.62), younger adults (19–25), those born in Sweden, and individuals with higher education and income. • Lower use among older adults (≥65 years), those born outside the EU28, with lower education or income, or with heart failure and diabetes; depression and COPD[h] or asthma predicted higher telemedicine use. • Digital care use is more concentrated in urban areas (67.5% of users).
Merkel et al [39] (2020); Europe	• Cross-sectional • Adults aged >65 years who use the internet (n=6900) • Multilevel logistic regression	(1) Use of internet-based health care services (any service, not specified)	• Age • Gender • Education • Social class • Employment • Marital status • Household size • Population density	• eHealth users were younger, better educated, and of higher social class. • Older age was associated with lower use (OR 0.97, 95% CI 0.96 - 0.98; *P*<.001) • Higher education (16 - 19 years: OR 1.43, 95% CI 1.15 - 2.79; ≥20 years: OR 1.95, 95% CI 1.54 - 2.46; *P*<.001) and higher social class (medium: OR 1.45, 95% CI 1.23 - 1.71; high: OR 2.00, 95% CI 1.53 - 2.61; *P*<.001) predicted greater eHealth use. • Living with a partner increased, and living alone reduced, eHealth use. • Urban residence (OR 1.23, 95% CI 1.02 - 1.48; *P*=.03) and higher national education participation among older adults (OR 1.06, 95% CI 1.01 - 1.13; *P*=.02) were positively associated with use.
Kharko et al [47] (2025); Norway Sweden Finland	• Prevalence • NORDeHEALTH 2022 Survey (n=27,038) • Descriptive	(2) Access to EHR[i]	• Gender	• Women less likely to have never accessed the EHR or to be first-time users. • Women are more likely to have visited the EHR >20 times (χ^2_4=57; *P*<.001).

XSL•FO
RenderX

Author (year); Country	Design, Sample, Population, and Statistical analysis	eHealth Cat[a] and study Outcome	SDH[b]	Description of main finding of inequalities in the studies
Pálsdóttir et al [38]; Iceland	• Prevalence • Adults (≥ 18 years) • Residing in Iceland (n=3000) • Descriptive	(1) Use of national digital health care system for communication with health professionals or access personal health information	• Age • Gender	• Women aged 18 - 35 years (P<.001) and 36 - 55 years (P<.10) used the digital system more frequently. • Women >56 years showed lower use compared to younger women.
Petersen et al [43] (2017); Denmark	• Prevalence • Danish adults (n=1059) • Descriptive	(1) Access and use of Sundhed.dk, a national health portal	• Education level	• Individuals with only primary education were less likely to use the Danish National Health Portal than those with higher education. • Portal use was 21% among those with primary education versus 60% among higher-educated individuals.
Eriksson et al [44] (2025); Sweden	• Retrospective cohort (registry-based) • Individuals of any age registered with Primary Health Care Centers, and having made at least 1 outpatient consultation during the study period (n=73,486) • Multivariable Logistic Regression	• (3) Use of telemedicine consultation in primary health care and EHR	• Age	• Younger adults (20-39 years) more likely to use telemedicine; older adults (60 - 79 and ≥80 years) less likely (IRR[j] 0.19, 95% CI 0.16 - 0.25). • Women showed higher telemedicine use (IRR 1.39, 95% CI 1.28 - 1.51). • Patients with high resource use (RUB 5; IRR[e] 2.67, 95% CI 0.20 - 3.54) and more EHR entries (IRR 1.50, 95% CI 1.32 - 1.72) showed higher use; no significant differences by Care Need Index.
Kc et al [48] (2024); England	• Ecological cohort • GP-registered[k] patients in practices with >200 patients between March 23, 2020, and June 27, 2022, aged ≥15, (no n reported) • Univariate negative binomial regressions	(1, 2, and 4) Use of the NHS app for registrations, log-ins, appointments booked, prescriptions, medical record views	• Age • Gender • Ethnicity • Deprivation • Health status	• Higher use in practices with more patients aged 15 - 34 years (+10.39% registration). • Lower use across more deprived areas (up to 38.84% lower registration in most vs least deprived quintile). • Use higher in larger practices and those with more White patients; lower where more male or chronically ill patients were registered.

[a]Author-defined grouping of eHealth tools categories.

[b]Social determinants of health assessed.

[c]OR: odds ratio.

XSL•FO
RenderX

[d]UCLH: University College London Hospitals.
[e]SDI: Social Deprivation Index.
[f]aOR: adjusted odds ratio.
[g]eHEALS: eHealth Literacy Scale.
[h]COPD: chronic obstructive pulmonary disease.
[i]EHR: electronic health record.
[j]IRR: incidence rate ratio.
[k]GP: general practitioner.

Overall Summary

The included studies revealed consistent social inequalities in access to telemedicine, structured along key social determinants of health and disproportionately affecting vulnerable groups.

Age

Age was the most consistently associated determinant, with lower use among adults aged 60 years and older, and higher engagement among younger and middle-aged [32,33,35-39,41,44-46,48,49].

Gender

Findings regarding gender were more heterogeneous. Several studies showed lower engagement among men across different modalities [32,33,37-39,44-48], while others found no significant differences [32,34,36,39,40].

Socioeconomic

Lower education was strongly associated with reduced use across most studies [33,34,36,37,39,42,43,45]. Similar patterns were seen for income [33,45], social class [39,42], and area deprivation [35,46,48,49]. Unemployment [36,45] was generally linked to reduced use, though findings were mixed in some contexts [32,39,40].

Migration, Ethnicity, and Language

Individuals born outside the country of study [33,42,45], belonging to ethnic minorities [46,48,49], or requiring interpreter support [35,46] were less likely to access digital health care services.

Place of Residence

In general, people living in rural or remote areas showed lower usage levels [33,45,49] compared to those in urban or more densely populated settings. However, some studies showed mixed or context-dependent patterns [39,40].

Household Composition

Results were less consistent. Some studies found lower engagement among individuals without a partner [37] or with more children [35], while others reported no significant associations [39,40].

Health Status

The findings were mixed. Chronic illness or poor self-rated health sometimes reduced engagement [33,36,45,48,49], while others observed higher usage among those with more frequent contact with the health care system [44].

Digital Skills

Across the studies that assessed this dimension, lower levels of digital competence [41], limited prior digital experience [42], lack of access to a device [36], low health literacy [34], and low confidence in using technology [32] were all associated with significantly lower use of digital health services.

In the next section, we break down these findings by type of eHealth service used.

eHealth Portal

Among the 10 studies that analyzed eHealth portals [32,34,35,38-41,43,48,49], 9 included age as a determinant [32,34,35,38-41,48,49]. Of these, 78% (6/9) found age to be a consistent factor associated with portal use [32,35,38,39,41,49]. Younger adults (15-34 years) showed higher use, and use declined after the age of 60 years [45]. Two other studies found no differences [34,46].

Regarding gender, 70% (7/10) of the studies analyzing eHealth portals included this variable [32,34,38-41,48]. Findings were mixed. Some studies found no significant differences [32,34,39,40], while others reported higher usage among women [48], and greater overall use by women [38].

Regarding socioeconomic factors, 70% (7/10) of the studies [34,35,39,40,43,48,49] examined associations between portal use and indicators such as education, income, social class, employment, and area-based deprivation. Higher education [39,40,42,45] and less deprivation [38,41,49] were consistently associated with greater use (eg, postgraduate adjusted odds ratio [aOR] 2.38, 95% CI 1.42 - 4.02) [40].

Employment status was analyzed in 30% (3/10) of studies, showing inconsistent results [32,39,40]: only one study found that being used was related to lower use (OR 0.57, 95% CI 0.38 - 0.84) [45]. The other two reported no associations [32,39].

Household composition, including size of household and marital status, was found as not associated in 2 of the 3 studies that analyzed it (2/10) [39,40]. However, Pierce et al [35] tied higher parity to lower use.

Cultural background—analyzed in 5 [34,35,40,48,49] of the 10 portal-focused studies (50%) [32,34,35,38-41,43,48,49]—showed no strong effects, with no significant differences in 2 studies [34,40]. Pierce et al [35] reported that 48.7% of nonusers were non-native compared to 24.4 % of those with high engagement; Kc et al [48] observed higher use in White patients [48]; Zhang [49] noted higher Asian patient use postadjustment.

XSL•FO
RenderX

In terms of area of residence, 2 [39,49] of the 3 (3/10) studies [39,40,49] that analyzed it found higher use among urban areas (OR 1.23, 95% CI 1.02 - 1.48) [39] and activation [49]. Knöchelmann et al [40] found no differences.

Health status was examined in 40% (4/10) of the studies [32,34,48,49], with mixed associations: some studies reported lower use among individuals with long-term conditions [48,49], while others found higher use among those with poor health [34] or no significant differences [32].

Finally, digital competence was analyzed in 30% (3/10) of the studies, where it was described as a strong predictor [32,34,41]: (OR 8.82, 95% CI 6.15 - 12.64) [41] and (aOR 2.96, 95% CI 2.02 - 4.35) for high literacy [34]. Lower health literacy was associated with lower use [32].

Electronic Health Records (EHRs)

Among the 33% (6/18) of studies [37,41,42,46-48] that analyzed EHR use, age was examined in 5, and all of them found that younger adults were more likely to use these tools [37,41,42,46,48]. Engagement declined in older adults, especially ≥75 years, with reduced registration (aOR 0.40, 95% CI 0.36 - 0.44) and activation (aOR 0.39, 95% CI 0.32 - 0.47) [46].

Gender was assessed in all 6 studies [37,41,42,46-48]. Four (67%) found higher EHR use among women [37,41,46-48]. In contrast, England reported higher male registration into the tool (aOR 1.10, 95% CI 1.04 - 1.16) but lower activation of it (aOR 0.85, 95% CI 0.78 - 0.94) [46], which was necessary for usage. González-Cacheda et al [42] found no differences.

Socioeconomic status was examined in 67% (4/6) of the studies [37,42,46,48]. All confirmed that higher levels of education increased use [37,42], and deprivation showed a clear gradient: individuals in more affluent areas were more likely to register and activate the tool [46,48].

Cultural background, assessed in 50% (3/6) of the studies [42,46,48], predicted lower activation in Asian (aOR 0.61, 95% CI 0.53 - 0.71), Black (aOR 0.45, 95% CI 0.36 - 0.56), and mixed (aOR 0.77, 95% CI 0.60 - 0.97) groups compared to White [46]. Evidence comes primarily from a large cross-sectional study conducted in England (n=28,637) [46], complemented by national datasets from England [48] and Spain [42]. Language barriers reduced use as no interpreter need increased registration (aOR 1.63, 95% CI 1.33 - 1.99) and activation (aOR 3.16, 95% CI 1.96 - 5.09) [46]. National [42] and White patients [48] had higher use.

Only one study (17%, 1/6) [37], a Swedish survey of over 3000 adults, analyzed marital status, reporting that being in a partnership was linked to higher engagement [37].

Health status findings showed mixed results among the 50% of studies (3/6): one study found no differences [37], another observed lower use in long-term conditions [48], while Chapman reported higher registration among patients with complex conditions (aOR 2.54, 95% CI 2.30 - 2.82) [46].

Finally, both studies assessing digital competence were conducted in Finland (n=4495) [41] and Spain (n=7454) [42], using large national datasets, and both found that lower digital skills were associated with reduced use of EHRs [41,42].

Telemedicine and Remote Primary Care

Among the 22% (4/18) of studies [33,36,44,45] that analyzed telemedicine and remote primary care, 100% (4/4) examined age and consistently identified younger adults as the main users, while older adults showed lower use [33,36,44,45]. In Sweden, those aged 60 - 79 years had substantially lower use (IRR [incidence rate ratio [0.45], 95% CI 0.40 - 0.50) and those ≥80 years even lower (IRR 0.19, 95% CI 0.16 - 0.25) [44].

Gender was assessed in 100% (4/4) of the studies analyzing telemedicine and remote primary care [33,36,44,45], with 75% (3/4) reporting higher usage among women [33,44,45]. In Sweden, women accounted for 65.4% of users (IRR 1.39, 95% CI 1.28 - 1.51) [44], while one study found no gender difference [36].

Socioeconomic factors were assessed in 100% (4/4) of the studies [33,36,44,45]. Higher education was associated with greater use (OR 1.41, 95% CI 1.19 - 1.67) [36], with postsecondary users showing the highest engagement (7.2%; OR 1.51, 95% CI 1.74 - 1.56) [33]. More users were observed among individuals with higher incomes (7.7% vs 4.0%; OR 2.46, 95% CI 2.38 - 2.54) [33]. Employment was also linked to usage: 85.7% of users were employed versus 14.3% unemployed [45], while one study reported lower odds among retirees (OR 0.68, 95% CI 0.46 - 1.02) [36]. One study found no differences across social class [44].

Cultural background and origin were assessed in 50% (2/4) of the studies, both of which were based on large population-based datasets [33,45]. Both studies reported lower uptake among foreign-born users compared to natives (3.2% vs 7.2%; OR 0.53, 95% CI 0.52 - 0.55) [33].

Urban residency was analyzed in 2 [33,45] of 4 studies [33,36,44,45]. One study associated it with higher use [45], while residing ≥10 km from health care services was significantly associated with higher telemedicine use (7.1%; OR≤1.11) [33].

Health status was analyzed in 100% (4/4) of the studies [33,36,44,45], showing mixed results. Better health predicted higher use (OR 1.31, 95% CI 1.01 - 1.69) [36]. In contrast, heart failure was associated with lower use (OR 0.49, 95% CI 0.40 - 0.59) [33]. Some conditions, such as depression, asthma, and chronic obstructive pulmonary disease, were linked to higher use, with greater morbidity also associated with increased engagement (IRR 2.67, 95% CI 2.02 - 3.54) [44,45].

Online Appointment Booking

Inequalities in online appointment booking were examined in 11% (2/18) of the studies [41,48], both showing that younger adults were more likely to use these systems. In Finland, use declined sharply with age [41], while in England, practices with more patients aged 15 - 34 years had slightly higher booking rates (+1.35%) [48]. Gender differences were assessed in 100% (2/2) of the studies. Both studies reported lower online booking among men [41,48].

XSL•FO
RenderX

Socioeconomic and cultural factors were analyzed in 50% (1/2) of the studies, revealing lower booking rates in practices located in more deprived areas, as well as reduced uptake among non-White populations [48].

Health status was assessed in 50% (1/2) of the studies, showing that patients with long-term conditions booked less frequently (–0.77%) [48].

Finally, digital competence was evaluated in 50% (1/2) of the studies and was strongly associated with use; patients with good digital skills were over 10 times more likely to book online (OR 10.91, 95% CI 7.24 - 16.44) [41].

Discussion

Principal Findings

Although European public health care systems are characterized by universal health care coverage, this systematic review aimed to examine inequalities in access to and use of eHealth tools across social determinants of health. The findings reveal persistent and significant social inequalities, with lower usage consistently reported among older adults, individuals with lower socioeconomic or educational status, rural populations, ethnic minorities, people with poor health, and those with limited digital competence. Findings regarding gender were more mixed, with some studies reporting slightly lower usage among men. The majority of included studies were methodologically robust, with most rated high quality and a few moderate quality according to JBI appraisal and conducted with widely varying sample sizes, which should be considered when interpreting these patterns and their potential translation into real-world health care practice. eHealth portals were the most investigated tool, but EHRs and telemedicine showed more consistent evidence of social gradients. Online appointment booking was the least studied category, limiting the strength of conclusions in this area.

Interpretation of Key Findings

Although digitalization in health care has been presented as improving access and system efficiency [1,2,50], our findings reveal a generational gap across all tool types [32,33,35-39,41,42,44-46,48,49]. Lower usage among older adults reflects not only differences in digital competence—such as less experience and lower technological confidence [51,52]—but also a lack of service adaptation to their needs and potential age-related cognitive barriers that limit effectiveness [53]. These observations, derived mostly from high-quality studies, strengthen the reliability of the age-related patterns observed. They were consistently found both in medium-sized samples [37] and in large administrative datasets [33,44], reinforcing the robustness of these patterns across study designs. At the same time, heterogeneity in measurement and study scale reduces the precision of effect estimates, particularly in smaller studies. In addition, unintuitive interfaces, low perceived usefulness, and a preference for in-person care further contribute to lower uptake [54]. While family support may partially mitigate this gap [54,55], findings are inconsistent [40,56]. This raises concerns about reduced autonomy, informal caregiver burden—often falling on women—and the exclusion of those without social support.

Gender-related patterns were less consistent. Some studies showed higher usage among women for EHRs and telemedicine, while differences in other services were weaker, albeit with a general trend toward greater women's engagement. These mixed results may reflect contextual factors such as national health care organization, cultural norms, and sample composition. Some null findings came from smaller or moderate samples [32,40], whereas large studies such as Dahlgren et al [33] consistently showed gender differences. Greater women's involvement may be interpreted in light of traditional gender roles, which assign women the responsibility for caregiving, reflected in a higher willingness to seek care, familiarity with health services, and confidence in sustained, active communication [57] as well as in more continuous use, as suggested by one of the included studies [46]. However, previous literature shows that in certain contexts, the gender gap is reversed, possibly due to broader disparities in digital engagement. Women tend to use a more limited range of digital services and engage with them less frequently and intensively than men, reflecting differences in digital familiarity and confidence [58,59].

It is well established that the socioeconomic gradient not only determines access to material resources but also shapes individuals' ability to navigate digital environments. In our review, educational attainment, income level, and residence in disadvantaged areas were systematically associated with lower use of digital health tools, in line with studies from other digitalizing systems [60]. Evidence regarding employment status was less consistent, with some studies suggesting lower use among unemployed individuals. Beyond material barriers (access to internet or digital devices), differences in digital literacy, confidence, and prior experience—the so-called third-level digital divide—further perpetuated exclusion [61,62]. Although education strongly predicted competences, structural conditions such as local digital infrastructure and social position also shaped digital engagement. In disadvantaged regions, limited digital services may affect both patients and health care professionals, potentially reinforcing existing health care disparities [63]. Given the overall rigor of the included studies, these findings appear reliable; however, their transferability to different health care contexts should be approached with caution. The WHO stresses that closing this gap requires not only investment in infrastructure, but also training and support strategies specifically targeted at the most vulnerable groups [64]. The consistency of socioeconomic gradients across both small samples [34] and very large datasets [33] strengthens confidence in these associations and suggests that they are not driven by sample-specific variability.

Beyond socioeconomic status, ethnic minorities have shown lower engagement with all the tools analyzed, even after adjusting for socioeconomic variables. Language barriers, lack of familiarity with the health care system, distrust of institutional platforms, and poor cultural adaptation of services emerge as recurring obstacles [46,65,66].

XSL•FO
RenderX

Although urban environments are generally associated with higher use of eHealth portals [39,49], one study showed greater use of video consultations in rural areas [33], consistent with previous literature [67], which suggests that digitalization compensates for limited in-person services. This pattern, particularly in studies with careful design and with large samples such as Dahlgren et al [33], suggests that geographical effects are not artifacts of small-sample instability. However, rural areas often have lower educational attainment, which can reduce digital competence [42,45]. These findings indicate that geography can both mitigate and amplify inequalities, depending on the system's capacity to adapt.

Regarding health status, some studies showed lower use among people with multimorbidity, in line with previous literature [68,69], reflecting the Inverse Care Law, where populations with greater health care needs often face reduced access [70,71]. However, other studies found higher use among people with poorer health, likely due to more frequent contact with the health care system [40,48]. These mixed results are partly attributable to methodological differences and varying study quality, with patterns more clearly observed in studies with comprehensive data and stronger design [33,37]. Variation in sample sizes also contributed: larger datasets such as Eriksson et al [44] and Dahlgren et al (N≈2 million) [33] produced more stable health-related estimates than smaller clinical samples, helping explain discrepancies in the direction and magnitude of associations. Overall, individuals with greater clinical needs may face additional barriers when poor health intersects with low education or socioeconomic status.

Although the overall methodological quality of the included studies was acceptable—with most rated high quality and a few moderate quality according to the JBI appraisal—the predominance of cross-sectional designs limits causal inference. Substantial heterogeneity in study populations, digital tools, and outcome indicators also affects comparability across studies. Furthermore, the wide variability in sample sizes introduces important differences in estimate precision, which should be taken into account when interpreting inconsistent findings across determinants. These methodological and contextual differences influence how findings can be translated into real-world practice: variations in health care organization, digital infrastructure, and population characteristics across countries may alter the magnitude and direction of observed inequalities. Therefore, interpretation and policy translation should be undertaken with caution, recognizing that local context strongly shapes the impact of the digital divide.

Limitations and Strengths

This systematic review has several limitations. One limitation is that the search strategy was not developed in collaboration with an information scientist, which might have influenced the sensitivity and comprehensiveness of the retrieved records. The conclusions may also be affected by differences in study design, sample representativeness, and reporting. Although most studies were rated as high quality according to the JBI appraisal, some moderate-quality designs introduce a potential risk of bias that should be considered when interpreting the results. The exclusion of qualitative studies—although methodologically justified to ensure comparability—may have constrained the review's ability to capture rich contextual insights, such as user experiences, perceived barriers, and cultural or behavioral factors influencing digital health use, which are crucial for understanding the mechanisms behind observed inequalities. Heterogeneity in populations, definitions of eHealth use, and measurement of social determinants limited comparability and prevented meta-analysis, requiring a narrative synthesis. Restricting the search to peer-reviewed publications in English or Spanish may have introduced language and publication bias, excluding relevant evidence published in other languages or in gray literature. Nevertheless, this criterion enhances replicability and scientific robustness. The focus on public health care systems in Europe restricts generalizability to other regions. Additionally, the studies' highly variable sample sizes, while not undermining the findings, do influence precision and should be considered when weighing the strength of evidence across determinants.

Despite these limitations, the review has several strengths. It followed PRISMA guidelines, used a comprehensive search across 4 major databases, supplemented by backward citation searching, and included independent screening and data extraction by multiple reviewers to improve validity. By synthesizing evidence across diverse European settings, this review offers an up-to-date and policy-relevant overview of eHealth-related inequalities, and its classification by type of digital service enables a more granular understanding of patterns of exclusion. No amendments were made to the registered protocol after PROSPERO registration.

These findings carry important implications for real-world health care policy and practice. By highlighting persistent inequalities in access and use of digital health tools across multiple social determinants, the review underscores the need for targeted interventions, including digital literacy programs, inclusive eHealth platform design, and support strategies for older adults, ethnic minorities, and socioeconomically disadvantaged populations. The innovative contribution of this review lies in its focus on quantitative evidence from European public health care systems, integrating results across diverse social determinants and types of eHealth tools. This provides a policy-relevant, evidence-based perspective to guide equitable digital transformation and foster inclusive health systems.

Future Research Directions

Future research should explore the long-term effects of digital exclusion on health outcomes, assess the equity impact of digital health policies, and promote participatory research with underserved communities to co-develop inclusive digital solutions. There is a clear need for intersectional frameworks to understand how intersecting inequalities shape digital health access, especially in highly vulnerable populations. Additionally, research is scarce in low-income countries, where sociodemographic and economic contexts may lead to different and sometimes contrasting patterns, for example regarding gender. Further studies are also needed to evaluate the actual impact of digital health tools on access to health care services more broadly—not only their uptake or use but also how they affect entry into the health system and care continuity.

XSL•FO
RenderX

Moreover, more attention should be paid to the underlying mechanisms driving these inequalities, to move beyond description and identify structural, institutional, or behavioral factors that generate unequal access. Finally, while the number of studies exploring social disparities in digital health access has increased, there remains a gap in longitudinal and in-depth research that can capture the complexity and dynamics of these inequities over time.

Conclusions

This systematic review shows that digital transformation in European public health systems is not equally benefiting all population groups. Across the studies analyzed, persistent social inequalities in the use of digital health tools—particularly EHRs, eHealth portals, and telemedicine and remote primary care—were observed. The most excluded from digital care services are older adults, individuals with lower socioeconomic or educational status, ethnic minorities, rural populations, people in poor health, and those with limited digital competence.

Rather than closing gaps, digitalization may replicate or deepen existing health inequalities if equity is not explicitly addressed. Ensuring inclusive design, supporting digital literacy, and engaging underserved communities are essential steps to make digital health a tool for reducing—not reinforcing—structural exclusion. Although a few studies explicitly explored interactions between determinants, the detected patterns suggest that the combination of factors such as advanced age, low socioeconomic status, migrant background, and limited digital competence tends to produce cumulative effects that significantly limit the adoption of eHealth tools. In interpreting these findings, both the heterogeneity of studies and their risk of bias should be taken into account. While these findings offer a policy-relevant perspective, the heterogeneity of study designs, populations, and outcomes, as well as the risk of bias, limits direct translation into policy and practice. By synthesizing quantitative evidence across diverse social determinants and types of digital health tools, this review not only highlights persistent inequalities but also offers a policy-relevant perspective. These findings can inform equity-oriented interventions, guide inclusive eHealth design, and support digital literacy programs, helping policymakers and health systems ensure that digital transformation benefits all population groups.

Acknowledgments

The authors gratefully acknowledge the support of the “Instituto de Salud Carlos III” and the European Union, which co-funded this study under the 2023 call for Health R&D Projects.

Generative artificial intelligence tools were used solely to assist with the review and refinement of the English language in the manuscript. No artificial intelligence tools were used for content generation, data analysis, or interpretation.

Funding

This study was funded by the “Instituto de Salud Carlos III” and co-funded by the European Union (Grant No. PI23/01063), within the framework of the 2023 call for Health R&D Projects. The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study. All data supporting the findings of this study are derived from previously published articles, which are fully referenced in the manuscript. The complete search strategies are provided in Multimedia Appendix 1, and summary data extraction tables are available in Multimedia Appendix 4. Additional extracted datasets (eg, the complete data extraction matrix) can be made available from the corresponding author upon reasonable request.

Authors' Contributions

GMG, AAG, and UM conceptualized the study. GMG, AAG, and MFL designed the data search strategy and conducted data extraction. GMG drafted the initial version of the manuscript in close collaboration with AAG. MFL revised multiple versions of the manuscript, contributing substantially to the content and wording. UM and EV reviewed the final versions of the manuscript and provided critical input to the drafting process. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete search strategies.
[DOCX File, 16 KB - jmir_v28i1e81841_app1.docx]

Multimedia Appendix 2
Study-specific definitions of determinants and eHealth use.

XSL•FO
RenderX

[XLSX File, 11 KB - jmir_v28i1e81841_app2.xlsx]

Multimedia Appendix 3
Risk of bias assessment of each study.
[DOCX File, 24 KB - jmir_v28i1e81841_app3.docx]

Multimedia Appendix 4
Extended evidence table summarizing the main information of each study included in the systematic review.
[DOCX File, 113 KB - jmir_v28i1e81841_app4.docx]

Checklist 1
PRISMA checklist.
[DOCX File, 282 KB - jmir_v28i1e81841_app5.docx]

References

1. Oh H, Rizo C, Enkin M, Jadad A, Powell J, Pagliari C. What is eHealth (3): a systematic review of published definitions. J Med Internet Res 2005 Feb 24;7(1):e1. [doi: 10.2196/jmir.7.1.e1] [Medline: 15829471]
2. Global digital health monitor. World Health Organization. 2023. URL: https://data.who.int/dashboards/gdhm/overview [accessed 2025-08-03]
3. Resolution WHA71.7: digital health. World Health Organization. 2018. URL: https://www.who.int/publications/i/item/10665-279505 [accessed 2025-12-11]
4. Special report 25/2024: digitalisation of healthcare:EU support for member states effective overall, but difficulties in using EU funds. : European Court of Auditors; 2024 URL: https://www.eca.europa.eu/ECAPublications/SR-2024-25/SR-2024-25_EN.pdf [accessed 2026-01-14]
5. Jonsson M, Johansson S, Hussain D, Gulliksen J, Gustavsson C. Development and evaluation of eHealth services regarding accessibility: scoping literature review. J Med Internet Res 2023 Aug 17;25:e45118. [doi: 10.2196/45118] [Medline: 37590050]
6. Atherton H, Eccles A, Poltawski L, Dale J, Campbell J, Abel G. Investigating patient use and experience of online appointment booking in primary care: mixed methods study. J Med Internet Res 2024 Jul 8;26:e51931. [doi: 10.2196/51931] [Medline: 38976870]
7. Giebel GD, Speckemeier C, Abels C, et al. Problems and barriers related to the use of digital health applications: scoping review. J Med Internet Res 2023 May 12;25:e43808. [doi: 10.2196/43808] [Medline: 37171838]
8. EGovernment benchmark 2022. synchronising digital governments.INSIGHT REPORT. : Publications Office of the European Union; 2022 URL: https://op.europa.eu/en/publication-detail/-/publication/cc953c44-1468-11ed-8fa0-01aa75ed71a1/language-en [accessed 2026-01-14]
9. Abernethy A, Adams L, Barrett M, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022;2022(22). [doi: 10.31478/202206e] [Medline: 36177208]
10. Global Strategy on Digital Health 2020-2025: World Health Organization; 2021.
11. Gobburi RK, Olawade DB, Olatunji GD, Kokori E, Aderinto N, David-Olawade AC. Telemedicine use in rural areas of the United Kingdom to improve access to healthcare facilities: a review of current evidence. Informatics Health 2025 Mar;2(1):41-48. [doi: 10.1016/j.infoh.2025.01.003]
12. Western MJ, Smit ES, Gültzow T, et al. Bridging the digital health divide: a narrative review of the causes, implications, and solutions for digital health inequalities. Health Psychol Behav Med 2025;13(1):2493139. [doi: 10.1080/21642850.2025.2493139] [Medline: 40276490]
13. Western MJ, Armstrong MEG, Islam I, Morgan K, Jones UF, Kelson MJ. The effectiveness of digital interventions for increasing physical activity in individuals of low socioeconomic status: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2021 Nov 9;18(1):148. [doi: 10.1186/s12966-021-01218-4] [Medline: 34753490]
14. Szinay D, Forbes CC, Busse H, DeSmet A, Smit ES, König LM. Is the uptake, engagement, and effectiveness of exclusively mobile interventions for the promotion of weight-related behaviors equal for all? A systematic review. Obes Rev 2023 Mar;24(3):e13542. [doi: 10.1111/obr.13542] [Medline: 36625062]
15. Pettersson L, Johansson S, Demmelmaier I, Gustavsson C. Disability digital divide: survey of accessibility of eHealth services as perceived by people with and without impairment. BMC Public Health 2023 Jan 27;23(1):181. [doi: 10.1186/s12889-023-15094-z] [Medline: 36707791]
16. Chae J. A comprehensive profile of those who have health-related apps. Health Educ Behav 2018 Aug;45(4):591-598. [doi: 10.1177/1090198117752784] [Medline: 29353544]
17. Ratcliff CL, Krakow M, Greenberg-Worisek A, Hesse BW. Digital health engagement in the US population: insights from the 2018 Health Information National Trends Survey. Am J Public Health 2021 Jul;111(7):1348-1351. [doi: 10.2105/AJPH.2021.306282] [Medline: 34014759]

XSL•FO
RenderX

18. Hegeman PC, Vader DT, Kamke K, El-Toukhy S. Patterns of digital health access and use among US adults: a latent class analysis. BMC Digit Health 2024;2(1):42. [doi: 10.1186/s44247-024-00100-0] [Medline: 39544227]
19. Jahnel T, Dassow HH, Gerhardus A, Schüz B. The digital rainbow: digital determinants of health inequities. Digit Health 2022;8:20552076221129093. [doi: 10.1177/20552076221129093] [Medline: 36204706]
20. Torrent-Sellens J, Díaz-Chao Á, Soler-Ramos I, Saigí-Rubió F. Modelling and predicting eHealth usage in Europe: a multidimensional approach from an online survey of 13,000 European Union internet users. J Med Internet Res 2016 Jul 22;18(7):e188. [doi: 10.2196/jmir.5605] [Medline: 27450189]
21. Wong BLH, Maaß L, Vodden A, et al. The dawn of digital public health in Europe: implications for public health policy and practice. Lancet Reg Health Eur 2022 Mar;14:100316. [doi: 10.1016/j.lanepe.2022.100316] [Medline: 35132399]
22. Wilson S, Tolley C, McArdle R, Hassan N, Slight R, Slight S. Strategies to advance digital health equity: a qualitative study with underserved groups. Lancet 2024 Nov;404:S37. [doi: 10.1016/S0140-6736(24)02004-X]
23. Woolley KE, Bright D, Ayres T, Morgan F, Little K, Davies AR. Mapping inequities in digital health technology within the World Health Organization's European Region using PROGRESS PLUS: scoping review. J Med Internet Res 2023 Apr 28;25:e44181. [doi: 10.2196/44181] [Medline: 37115613]
24. Digital health not accessible by everyone equally, new study finds. World Health Organization. 2022 Dec 21. URL: https://www.who.int/europe/news/item/21-12-2022-digital-health-not-accessible-by-everyone-equally-new-study-finds [accessed 2025-12-11]
25. Organisation for Economic Co-Operation and Development. Health at a Glance 2023: OECD Indicators: OECD Publishing; 2023. [doi: 10.1787/4dd50c09-en]
26. European Commission. Case study on the digitalisation of health (ehealth). : Luxembourg: Publications Office of the European Union; 2024 URL: https://commission.europa.eu/document/download/652a3175-d410-4608-8f0e-642049433c35_en [accessed 2025-06-26]
27. Page MJ, McKenzie JE, Bossuyt PM, et al. PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
28. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
29. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 5;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
30. Ma LL, Wang YY, Yang ZH, Huang D, Weng H, Zeng XT. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? Military Med Res 2020 Dec;7(1):7. [doi: 10.1186/s40779-020-00238-8]
31. Xu Y, Wang M, Chen D, Jiang X, Xiong Z. Inflammatory biomarkers in older adults with frailty: a systematic review and meta-analysis of cross-sectional studies. Aging Clin Exp Res 2022 May;34(5):971-987. [doi: 10.1007/s40520-021-02022-7] [Medline: 34981430]
32. Hörhammer I, Suvanto J, Kinnunen M, Kujala S. Usefulness of self-guided digital services among mental health patients: the role of health confidence and sociodemographic characteristics. Int J Med Inform 2025 Feb;194:105693. [doi: 10.1016/j.ijmedinf.2024.105693] [Medline: 39556968]
33. Dahlgren C, Dackehag M, Wändell P, Rehnberg C. Determinants for use of direct-to-consumer telemedicine consultations in primary healthcare-a registry based total population study from Stockholm, Sweden. BMC Fam Pract 2021 Jun 26;22(1):133. [doi: 10.1186/s12875-021-01481-1] [Medline: 34172009]
34. Neves AL, Smalley KR, Freise L, Harrison P, Darzi A, Mayer EK. Determinants of use of the care information exchange portal: cross-sectional study. J Med Internet Res 2021 Nov 11;23(11):e23481. [doi: 10.2196/23481] [Medline: 34762063]
35. Pierce P, Whitten M, Hillman S. The impact of digital healthcare on vulnerable pregnant women: a review of the use of the MyCare app in the maternity department at a central London tertiary unit. Front Digit Health 2023;5:1155708. [doi: 10.3389/fdgth.2023.1155708] [Medline: 37153515]
36. Söderberg D, Bonn SE, Sjöblom L, et al. Individual patient factors associated with the use of physical or digital primary care in Sweden. Telemed J E Health 2024 May;30(5):1289-1296. [doi: 10.1089/tmj.2023.0351] [Medline: 38394275]
37. Muli I, Cajander Å, Hvitfeldt H, et al. To read or not to read - a cross-sectional study of Swedish primary care patients' adoption of patient accessible electronic health records. Digit Health 2024;10:20552076241287636. [doi: 10.1177/20552076241287636] [Medline: 39381824]
38. Pálsdóttir Á. Development in the adoption of a national digital healthcare system and experience at taking new technology in use – changes from 2019 to 2022. Inf Res 2024;29(2):464-482. [doi: 10.47989/ir292841]
39. Merkel S, Hess M. The use of internet-based health and care services by elderly people in Europe and the importance of the country context: multilevel study. JMIR Aging 2020 Jun 3;3(1):e15491. [doi: 10.2196/15491] [Medline: 32490837]
40. Knöchelmann A, Healy K, Frese T, et al. User profiles in digitalized healthcare: active, potential, and rejecting — a cross-sectional study using latent class analysis. BMC Health Serv Res 2024;24(1):1083. [doi: 10.1186/s12913-024-11523-w]
41. Heponiemi T, Kaihlanen AM, Kouvonen A, Leemann L, Taipale S, Gluschkoff K. The role of age and digital competence on the use of online health and social care services: a cross-sectional population-based survey. Digit Health 2022;8:20552076221074485. [doi: 10.1177/20552076221074485] [Medline: 35111333]

XSL•FO
RenderX

42. González-Cacheda B, Outeda CC. Understanding attitudes, knowledge, and use of e-health services in the health system in Spain. J Public Health Policy 2025 Jun;46(2):295-302. [doi: 10.1057/s41271-025-00549-x] [Medline: 39827313]
43. Petersen LS, Bertelsen P. Equality challenges in the use of eHealth: selected results from a Danish citizens survey. Stud Health Technol Inform 2017;245:793-797. [Medline: 29295207]
44. Eriksson P, Randjelovic M, Thulesius H, Hammar T, Lagrosen S, Nilsson E. Differences in use of telemedicine integrated into traditional primary health care - a comparative observational study. Scand J Prim Health Care 2025 Jun;43(2):476-487. [doi: 10.1080/02813432.2025.2457542] [Medline: 39915941]
45. Wilkens J, Thulesius H, Ekman B. From office to digital primary care services: analysing income-related inequalities in utilization. Int J Equity Health 2024 Apr 30;23(1):86. [doi: 10.1186/s12939-024-02184-6] [Medline: 38689241]
46. Chapman R, Haroon S, Simms-Williams N, et al. Socioeconomic deprivation, age and language are barriers to accessing personal health records: a cross-sectional study of a large hospital-based personal health record system. BMJ Open 2022 Jan 20;12(1):e054655. [doi: 10.1136/bmjopen-2021-054655] [Medline: 35058264]
47. Kharko A, Hagström J, Simola S, Cajander Å, Blease C, Hägglund M. The gender gap in EHR experiences and preferences: results from an international cross-sectional survey of patients. Stud Health Technol Inform 2025 May 15;327:939-943. [doi: 10.3233/SHTI250510] [Medline: 40380617]
48. Kc S, Papoutsi C, Reidy C, et al. Differences in use of a patient portal across sociodemographic groups: observational study of the NHS app in England. J Med Internet Res 2024 Nov 13;26:e56320. [doi: 10.2196/56320] [Medline: 39536310]
49. Zhang J, Gallifant J, Pierce RL, et al. Quantifying digital health inequality across a national healthcare system. BMJ Health Care Inform 2023 Nov 24;30(1):e100809. [doi: 10.1136/bmjhci-2023-100809] [Medline: 38007224]
50. Al-Shorbaji N. Improving healthcare access through digital health: the use of information and communication technologies. In: Healthcare Access: IntechOpen; 2022. [doi: 10.5772/intechopen.99607]
51. Yusif S, Soar J, Hafeez-Baig A. Older people, assistive technologies, and the barriers to adoption: a systematic review. Int J Med Inform 2016 Oct;94:112-116. [doi: 10.1016/j.ijmedinf.2016.07.004] [Medline: 27573318]
52. Klaver NS, van de Klundert J, van den Broek R, Askari M. Relationship between perceived risks of using mHealth applications and the intention to use them among older adults in the Netherlands: cross-sectional study. JMIR Mhealth Uhealth 2021 Aug 30;9(8):e26845. [doi: 10.2196/26845] [Medline: 34459745]
53. Çöme O, Gökdemir Ö, Bilik Sezer B, Kasapoğlu SS, Kjær NK, Güldal D. The interplay between cognitive function and digital health literacy among older adults: implications for e-health equity and accessibility. Int J Med Inform 2025 Sep;201:105934. [doi: 10.1016/j.ijmedinf.2025.105934] [Medline: 40273597]
54. Wildenbos GA, Jaspers MWM, Schijven MP, Dusseljee-Peute LW. Mobile health for older adult patients: using an aging barriers framework to classify usability problems. Int J Med Inform 2019 Apr;124:68-77. [doi: 10.1016/j.ijmedinf.2019.01.006] [Medline: 30784429]
55. Schroeder T, Dodds L, Georgiou A, Gewald H, Siette J. Older adults and new technology: mapping review of the factors associated with older adults' intention to adopt digital technologies. JMIR Aging 2023 May 16;6:e44564. [doi: 10.2196/44564] [Medline: 37191976]
56. Garcia Reyes EP, Kelly R, Buchanan G, Waycott J. Understanding older adults' experiences with technologies for health self-management: interview study. JMIR Aging 2023 Mar 21;6:e43197. [doi: 10.2196/43197] [Medline: 36943333]
57. Jiménez Ruiz I, Moya Nicolás M. The family caregiver: the naturalized sense of obligation in women to be caregivers {Article in Spanish]. Enfermería Global . 2017;17(1):420. [doi: 10.6018/eglobal.17.1.292331]
58. Tyers-Chowdhury A, Binder G. What we know about the gender digital divide for girls: a literature review. : United Nations Children's Fund (UNICEF); 2021 URL: https://www.unicef.org/eap/media/8311/file/What%20we%20know%20about%20the%20gender%20digital%20divide%20for%20girls:%20a%20literature%20review.pdf [accessed 2025-01-12]
59. Ek S. Gender differences in health information behaviour: a Finnish population-based survey. Health Promot Int 2015 Sep;30(3):736-745. [doi: 10.1093/heapro/dat063] [Medline: 23985248]
60. Standaar L, van Tuyl L, Suijkerbuijk A, Brabers A, Friele R. Differences in eHealth access, use, and perceived benefit between different socioeconomic groups in the Dutch context: secondary cross-sectional study. JMIR Form Res 2025 Jan 7;9:e49585. [doi: 10.2196/49585] [Medline: 39773883]
61. van Deursen A, Helsper EJ. The third-level digital divide: who benefits most from being online. In: Communication and Information Technologies Annual: Emerald Group Publishing; 2015:29-52. [doi: 10.1108/S2050-206020150000010002]
62. Milanti A, Chan DNS, Parut AA, So WKW. Determinants and outcomes of eHealth literacy in healthy adults: a systematic review. PLoS One 2023;18(10):e0291229. [doi: 10.1371/journal.pone.0291229] [Medline: 37792773]
63. Alam K, Mahumud RA, Alam F, Keramat SA, Erdiaw-Kwasie MO, Sarker AR. Determinants of access to eHealth services in regional Australia. Int J Med Inform 2019 Nov;131:103960. [doi: 10.1016/j.ijmedinf.2019.103960] [Medline: 31518858]
64. Equity within digital health technology within the WHO european region: a scoping review. World Health Organization. 2022. URL: https://www.who.int/europe/publications/i/item/WHO-EURO-2022-6810-46576-67595 [accessed 2025-08-03]
65. Bergman L, Nilsson U, Dahlberg K, Jaensson M, Wångdahl J. Health literacy and e-health literacy among Arabic-speaking migrants in Sweden: a cross-sectional study. BMC Public Health 2021 Nov 25;21(1):2165. [doi: 10.1186/s12889-021-12187-5] [Medline: 34823499]

XSL•FO
RenderX

66. Whitehead L, Talevski J, Fatehi F, Beauchamp A. Barriers to and facilitators of digital health among culturally and linguistically diverse populations: qualitative systematic review. J Med Internet Res 2023 Feb 28;25:e42719. [doi: 10.2196/42719] [Medline: 36853742]
67. Ratna S, Behera KL, Renuka JR. Exploring healthcare professionals' perceptions and practices on telemedicine adoption in rural regions. Semin Med Writ Educ 2023;2:139. [doi: 10.56294/mw2023139]
68. Manning SE, Wang H, Dwibedi N, et al. Association of multimorbidity with the use of health information technology. Digit Health 2023;9:20552076231163797. [doi: 10.1177/20552076231163797] [Medline: 37124332]
69. Runz-Jørgensen SM, Schiøtz ML, Christensen U. Perceived value of eHealth among people living with multimorbidity: a qualitative study. J Comorb 2017;7(1):96-111. [doi: 10.15256/joc.2017.7.98] [Medline: 29359124]
70. Pathirana TI, Jackson CA. Socioeconomic status and multimorbidity: a systematic review and meta-analysis. Aust N Z J Public Health 2018 Apr;42(2):186-194. [doi: 10.1111/1753-6405.12762] [Medline: 29442409]
71. Tudor Hart J. The inverse care law. Lancet 1971 Feb;297(7696):405-412. [doi: 10.1016/S0140-6736(71)92410-X]

Abbreviations

EHR: electronic health record
JBI: Joanna Briggs Institute
OR: odds ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
WHO: World Health Organization

Edited by A Mavragani, S Brini; submitted 05.Aug.2025; peer-reviewed by H Maheshwari, K Triep; revised version received 24.Nov.2025; accepted 24.Nov.2025; published 09.Feb.2026.

Please cite as:
Monasterio G, Fernández-López MJ, Valero E, Martin U, Ayala-García A
Digital Inequalities in the Use of eHealth Services in European Public Health Care Systems: Systematic Review of Observational Studies
J Med Internet Res 2026;28:e81841
URL: https://www.jmir.org/2026/1/e81841
doi:10.2196/81841

© Garazi Monasterio, Marcos José Fernández-López, Erika Valero, Unai Martin, Amaia Ayala-García. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Data-Driven Guideline Adherence in Data Representation and Compliance Measurement: Scoping Review

Minh Trang Hoang[1], BMedSci (Hon); Candice Donnelly[1], PhD; Christina Igasto[1,2], MSc, PhD; Amith Shetty[1,3,4], MBBS, FACEM, PhD; Malcolm Pradhan[1], MBBS, FACHI, PhD; Tim Shaw[1], PhD

[1]Biomedical Informatics and Digital Health, School of Medical Sciences, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia

[2]Service NSW, Sydney, Australia

[3]NSW Ministry of Health, Sydney, Australia

[4]Western Sydney Local Health District, Westmead, Australia

Corresponding Author:
Minh Trang Hoang, BMedSci (Hon)
Biomedical Informatics and Digital Health, School of Medical Sciences, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia

Abstract

Background: Best practice standards aim to standardize care and improve outcomes. However, variation in clinical practice exists, and not all deviations are inappropriate. Measuring adherence to best practice standards remains challenging due to limitations in representation methods and data fidelity.

Objective: This scoping review aims to survey and synthesize the existing literature on the computable representation of guideline recommendations and to explore methods for detecting and quantifying deviations from best practice standards.

Methods: We followed the Arksey and O'Malley framework and PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines. Five databases (Ovid Medline, EMBASE, IEEE Xplore, Web of Science, and Scopus) were searched in November 2025. Studies were included if they either (1) described a computer representation of best practice standards or (2) assessed adherence to such standards using patient data, including patient data derived from electronic medical records or event logs. Screening was done using Covidence (Veritas Health Innovation). Data were extracted on representation, clinical context, data sources, adherence metrics, and modeling techniques. A narrative synthesis was conducted to identify themes.

Results: Twenty-four studies were included. Most studies were published as conference proceedings (13/24, 54%). Fourteen studies (14/24, 58%) included measurement of adherence to best practice standards. Cardiovascular conditions were the most common focus (13/24, 54%). Data sources included Health Level Seven (HL7) messages, structured electronic medical record data, event logs, and Fast Healthcare Interoperability Resources (FHIR)-transformed data. Best practice standards were formalized using Business Process Model and Notation (BPMN; 6/24, 25%), ontologies (7/24, 29%), FHIR (4/24, 17%), or hybrid approaches (4/24, 17%). The most common method for adherence measurement was rule-based alignment. Several studies incorporated weighted scoring to differentiate the severity of deviations. Process mining was used in a subset to detect sequence and timing variations. However, most models lacked contextual sensitivity and rarely incorporated patient-specific factors, such as comorbidities, patient acuity, or clinician rationale. Consequently, although deviations can be automatically identified, determining whether they were clinically warranted remained largely unresolved.

Conclusions: Despite promising advances, challenges persist in computer-interpretable representation and measuring adherence in a clinically meaningful way. Current approaches predominantly assess technical alignment rather than clinical relevance and are limited by data quality and standardization, thereby limiting real-world utility. This scoping review offers an innovative contribution by synthesizing evidence from 2 separate domains—the computable representation of best practice standards and the measurement of adherence. The findings emphasize the need for context-aware, standardized modeling and integration with clinical workflows to distinguish warranted from unwarranted deviations. Such advances are essential for scalable, transparent, and real-time adherence monitoring—ultimately driving safer, patient-centered care delivery.

(J Med Internet Res 2026;28:e79937) doi:10.2196/79937

KEYWORDS

clinical practice guidelines; best practice; guideline compliance; computer-interpretable guidelines; electronic medical records; clinical decision support

XSL•FO
RenderX

Introduction

Delivering patient care that aligns with evidence-based best practice has been associated with many benefits, including increased hospital efficiency, decreased operational costs, and improved patient outcomes [1-4], especially reduced length of stay and decreased mortality rate [4]. This is achievable through best practice standards, which are typically reflected in clinical practice guidelines (CPGs). The most cited definition refers to CPGs as "systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances" [5]. In the literature, terms such as "protocols," "practice policies," "clinical algorithms," and "clinical pathways" are sometimes used to refer to CPGs, although they differ in scope, specificity, and intended use [6]. While CPGs are narrative documents that provide overall evidence-based recommendations, clinical pathways (CPs) translate these guidelines into structured, time-sequenced clinical processes of care for a specific problem within a particular population [7]. CPs are typically developed at the institutional or regional level and are informed by international or national guidelines but tailored to local contexts [7]. Their structured, time-sequential nature may make them well-suited as benchmarks for monitoring and evaluating clinical practice.

Despite the recognized benefits and availability of CPGs and CPs, a persistent gap exists between recommended care and clinical practice delivery [8,9]. This discrepancy, often referred to as practice variations or pathway deviations, can reflect both unwarranted and warranted deviations from practice. Practice variations may indicate erroneous care resulting in adverse patient outcomes and increased health care expenses and are often referred to as unwarranted deviations. While some practice variations may be deemed warranted deviation when individualized care is provided to accommodate patient-specific needs. Distinguishing between unwarranted and warranted deviations is crucial to understanding and improving clinical practice. However, existing approaches to measuring adherence often rely on aggregated metrics or single process indicators and fail to capture the nuance and temporal complexity of clinical care.

Advancements in digital technology play a pivotal role in delivering health care and enabling systematic and data-driven monitoring of adherence to best practice standards [10]. Electronic medical records (eMRs) offer a large volume of time-stamped, structured clinical data that can be used to evaluate adherence to best practice. In addition, Clinical Decision Support Systems (CDSS) embedded within eMRs can promote adherence to best practice standards by delivering evidence-based recommendations in a variety of care aspects [11,12]. As a result, it has also led to the emergence of computer-interpretable guidelines (CIGs), which transform narrative CPGs or CPs into machine-readable, logical representations to support automated assessment of adherence.

Multiple methodologies have been developed to support the representation of best practice standards in computable formats. Guideline modeling, such as GLIF3, Asbru, SAGE, PROforma, as well as workflow and interoperability frameworks such as Business Process Model and Notation (BPMN) and Fast Healthcare Interoperability Resources (FHIR), offer different strengths in terms of expressiveness, configurability, and system integration [13-17]. In addition, Clinical Quality Language (CQL) [18], a high-level, domain-specific language used to support the creation and execution of clinical decision support rules and algorithms. However, the heterogeneity of these approaches introduces challenges in clinical implementation and usability, and cross-system interoperability [17]. Additionally, the availability of data and different interpretations of adherence measures continue to pose a challenge to the measurement of adherence to CPGs [19,20]. Nevertheless, a lack of standardized metrics and limited consideration of clinical context continue to constrain the practical utility and adoption of these methods in clinical practice.

A recent scoping review provides summaries on the role of informatics in addressing practice variations and the clinical tasks in which practice variations occur [21]. There remains a need to understand the methodologies behind the representation of best practice standards and how these representations enable the measurement of adherence to best practice. This scoping review aims to survey and synthesize the existing literature on the computable representation of guideline recommendations and to explore the extent to which these representations have been used to detect and quantify deviations from best practice. By doing so, this review highlights methodological and operational gaps in current approaches and identifies opportunities to advance adherence measurement in digital health systems.

Methods

Study Design and Framework

This scoping review was conducted following the methodology and guidelines by Arksey and O'Malley [22]. In line with scoping review conventions, no quality appraisal of the included studies was conducted, as the objective was to map the breadth of a diverse range of work.

Study Protocol

The PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) [23] was used to develop our review protocol. As a recommended practice, we used a checklist when reporting results to enhance transparency and reproducibility (eg, by sharing the exact search strategy and source databases). A complete checklist is available in Checklist 1.

Information Sources and Search Strategy

We searched 5 databases: Ovid Medline, EMBASE, IEEE Xplore, Web of Science, and Scopus. No study registries were searched. We also did not purposefully look for any online or print source. The last search was conducted in November 2025. The search strategy was developed in consultation with an experienced university librarian, without adapting previous search strategies from other literature reviews. It included combinations of controlled vocabulary and keywords related to clinical guidelines, clinician adherence, variation, and eMRs. We did not impose any time limit on obtaining a comprehensive

XSL•FO
RenderX

overview of published articles. We reported the search according to PRISMA-S (an Extension of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Statement) guidelines [24], which are available in Checklist 2. Additional articles were identified through backward searching of the reference lists in the included studies and the authors' collection. No additional studies or data were sought by contacting authors, experts, and manufacturers.

The references were imported to Covidence (Veritas Health Innovation) [25] for deduplication. During title and abstract screening, additional duplicates and irrelevant entries were manually removed. Covidence was also used for full-text screening, data extraction, and charting.

Eligibility Criteria

In this review, CPGs and CPs were considered as best practice standards as they provide evidence-based actionable recommendations. Our focus was on studies conducted in tertiary care settings, where guideline implementation and adherence monitoring are most frequently studied.

To maintain a focus on comprehensive care delivery and computable representations of best practice, we included studies that either (1) described or modeled a computer representation of best practice standards, or (2) assessed adherence to such standards using patient data, such as eMR-derived patient data or patient event logs. Studies were eligible if they captured multiple (>1) guidelines or components of care in the patient pathway. Studies focused solely on a single component, such as medication adherence—while important—were excluded, as they addressed only a single aspect of care. We also included studies that aimed to detect or quantify deviations from best practice across multiple components outlined by CPGs and CPs. Qualitative analyses (eg, interviews and ethnographic observations) that did not offer computer representations or measurable assessments of adherence were excluded. Additionally, studies that used a single process indicator as a proxy for adherence were excluded if they lacked a broader evaluation of best practice adherence.

Articles were excluded if they were not in English, had no full text available, or were not aligned with the research questions. Given the emerging nature of this topic and the evolving technological landscape, conference proceedings were also considered to capture developments in the field.

Study Selection

The retrieved articles were initially screened by applying the eligibility criteria by one reviewer (MTH). Concurrently, 5% of the total articles were randomly selected and separately screened by a second reviewer (MS). Discrepancies were discussed and resolved through consensus. The study selection process is illustrated in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

Data Charting and Data Items

A structured data charting form was developed to extract the following variables from each included study:

1. Bibliographic information (title, authors, year of publication, and publisher)
2. Study aims
3. Representation of clinical data
4. Representation of best practice standards
5. Whether deviations from clinical pathways were detected and measured
6. The mechanism used to measure adherence
7. Reported limitations

Full extraction details are available in Multimedia Appendix 1 [20,26-48].

Synthesis of the Result

Extracted data were analyzed using a narrative synthesis approach. Studies were grouped thematically based on their technical approaches to best practice representation and adherence measurement. Key features, methodological innovations, and limitations were summarized across studies. All authors contributed to iterative discussions, the refinement of the thematic structure, and the interpretation and presentation of the results.

Results

Included Articles and Metadata

For this scoping review, 24 articles [20,26-46,47,48] were included for data extraction (Figure 1). Most studies were conducted in China (6/24, 25%), the United States (5/24, 21%), and Germany (3/24, 13%). Of these, more than half (13/24, 54%) were published as conference proceedings rather than full-length journal articles, likely reflecting the technically innovative but emerging nature of this research area. Of 24 studies [20,26-46,47,48], 14 (58%) [20,26-36,46,48] included a measure of adherence to best practice standards (Figure 2). The number indicates how frequently each combination was used in the included studies. Formal knowledge representations—such as BPMN and Semantic Web Ontology Language (OWL)—combined with rule-based alignment are the most commonly used approach for representing best practice standards and measuring adherence.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram demonstrates the article search and selection process.

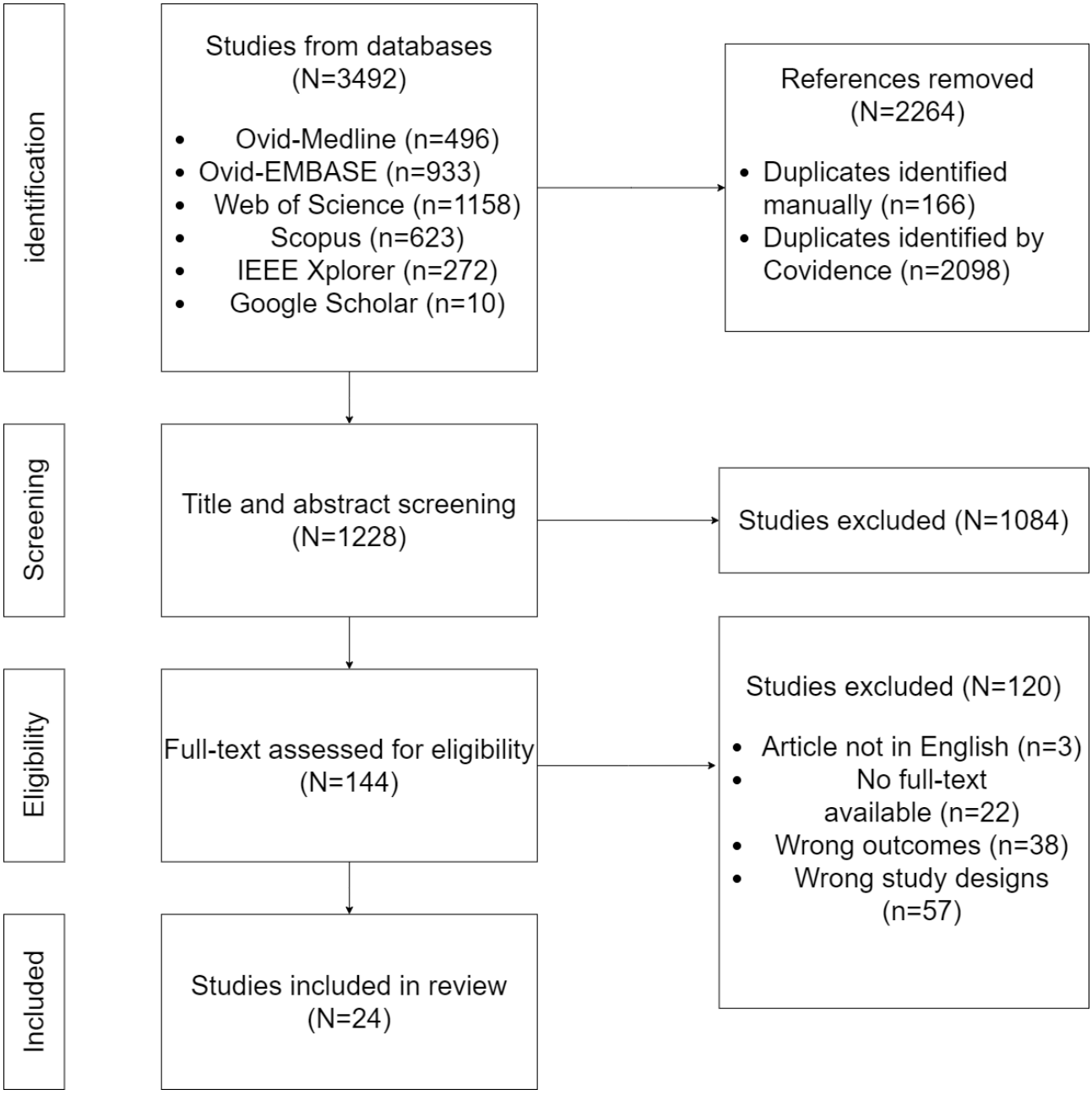

XSL•FO
RenderX

Figure 2. Map of best practice standards representation by approach to measure adherence in the included literature (N=14). BPMN: Business Process Model and Notation; FHIR: Fast Healthcare Interoperability Resources; OWL: Semantic Web Ontology Language.

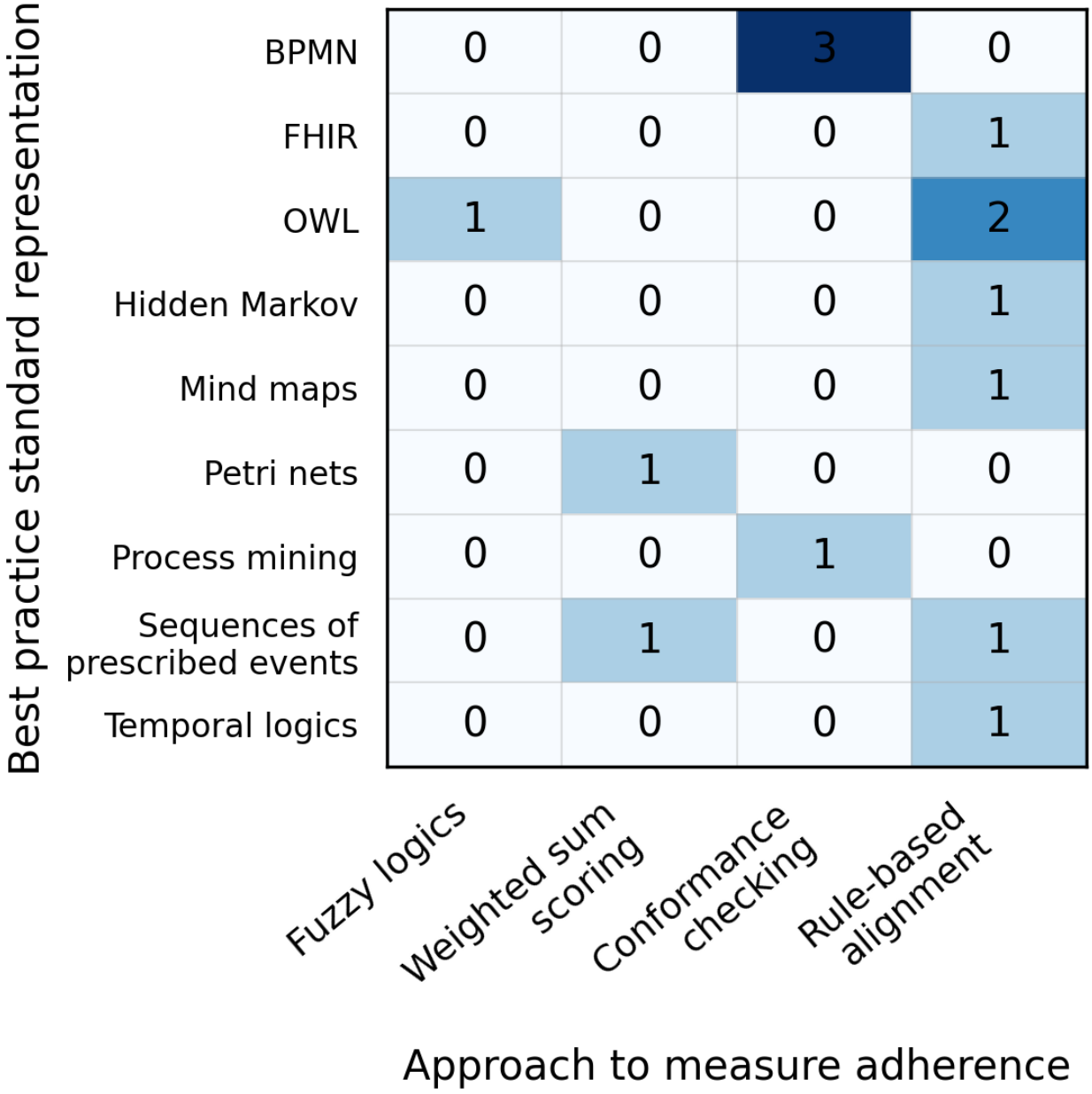

Clinical Conditions

Among the 24 included studies [20,26-46,47,48], only 1 study did not provide any clinical setting [26]. Cardiovascular disease was the most common clinical area, accounting for 54% (13/24) of the sample. Two studies applied their models across multiple clinical conditions (Table 1).

XSL•FO
RenderX

Table . Clinical conditions whose clinical practice guidelines or clinical pathways were translated to computer-interpretable guidelines in the included studies (N=24).

Clinical condition	Count
Weaning protocol in Intensive Care Unit (ICU)	1
Diabetes	1
Major joint replacement	1
Nutrition	1
Thyroid disease	1
No clinical case	1
COVID-19	2
Family planning and sexually transmitted infections	1
Cardiovascular disease	13
Mixed clinical conditions	2

Data Sources Used for Monitoring Adherence to Best Practice Standards

Various sources of clinical data were used in the included studies, including Health Level Seven (HL7) messages, structured clinical data extracted from eMR [27,28], patient events trace derived by sorting timestamped entries from eMR [20,29-31], and system-generated event logs [32-35]. Some studies transformed eMR data into standardized formats using the FHIR specification [36-38]. Conversely, several studies focused solely on the conceptual representation of best practice standards without implementation [26,39-46].

Representation of Best Practice Standards

Of the 24 studies [20,26-46,47,48], 13 [27,28,32,36,38-42,44-47] used CPGs as their reference for best practice, while the remaining studies used CPs. Various methods were used to transform CPGs and CPs to CIG format. The most used approaches were BPMN (6/24, 25%) and FHIR (4/24, 16%). For example, Yan et al [35] used BPMN to model a task-time matrix CP for unstable angina pectoris. The authors mapped terminology from eMR to pathway activities, identified key activities as milestones, and then embedded the remaining activities between those milestones to construct the process [35]. Parallel and inclusive gateways were used to represent concurrent activities, while exclusive gateways captured alternate clinical routes [35]. This BPMN representation yielded a computer-interpretable pathway model that served as a reference standard for detecting deviations in real-world care delivery [35]. Meanwhile, Lichtner et al [45] represented CPGs in FHIR by identifying the required elements and metadata from the CPGs to be captured in FHIR representation. An information model was built by organizing these elements into 3 main categories—medical relationships, metadata, and justification for the recommendations—and sorting them into 3 layers of certainty of evidence (primary clinical studies, systematic review and outcome-level assessment, and final recommendation derivation with evidence grading) [45]. Finally, the model was mapped into EBM-on-FHIR and CPG-on-FHIR profiles using resources such as PlanDefinition, EvidenceVariable, and ArtifactAssessment, thus enabling a computer-interpretable representation of CPG recommendations with underlying evidence certainty [45]. Details of the common methods used to represent CIGs are presented in Table 2.

XSL•FO
RenderX

Table . Description of the most used approaches to represent computer-interpretable guidelines.

Approach name	Included studies	Mechanism
Business Process Model and Notation (BPMN)	[32,33,35,37,44]	• Map variables extracted from eMR[a] to those components of CPGs[b] and CPs[c] • Identify key activities in CPGs and CPs to act as milestones that connect different phases of the care process • Add other care activities in between key activities • Activities are sequenced based on the clinical logic: whether they occur simultaneously or exclusively, and the order in which activities occur first
Ontology (OWL)	[26,27,30,43,44]	• Clinical concepts are defined as classes • A clinical concept can have object properties which describe the relationship between clinical concepts and data properties which record its quantitative aspect • eMR variables are then mapped to the newly created ontology to create instances of classes • Decision rules are created based on the if-then logic to encompass the conditions as outlined in CPGs and CPs
FHIR	[36,37,40,45]	• Identify the contents and metadata of CPGs and CPs to be captured in the FHIR-based[d] representation • Information model is built to understand the clinical concepts and relationships between them • Components of CPGs and CPs are mapped to the corresponding FHIR resources • Specifically, the PlanDefinition resource is used to model the care processes as outlined in CPGs and CPs. Meanwhile, the ActivityDefinition resource represents each care activity within the whole care process

[a]eMR: electronic medical record.

[b]CPG: clinical practice guideline.

[c]CP: clinical pathway.

[d]FHIR: Fast Healthcare Interoperability Resources.

Interestingly, some studies used hybrid approaches to represent best practice standards. For example, Dos Santos Leandro et al [37] used BPMN and FHIR to model stroke care CPs, first structuring the care process with BPMN, then mapping the components to corresponding FHIR resources to enable semantic interoperability. Whereas 2 studies [27,47] defined their best practice standards as linear sequences of prescribed actions without using formal modeling tools.

In addition to these core approaches, other approaches used to transform CPGs and CPs to CIGs included Petri Net [29], a combination of Natural Rule Language (NRL) and Object Constraint Language (OCL) [47], fuzzy logics [38], process mining [34], decision tree [28,39], hidden Markov model [31], and the formal representation language PROforma [41,42]. These studies thus demonstrate the use of a diverse range of formalisms to represent best practice standards; however, some of them lacked clinical integration, as observed in studies such as [39,41,42].

Approaches to Measuring Adherence

The included studies used a range of techniques to detect and quantify deviations from best practices standards. The most commonly used approach involved rule-based alignment, wherein individual patient records were assessed against predefined care steps derived from CPGs and CPs [20,27-29,36,46,48]. In this approach, CPG and CP components are translated into computable "Event-Condition-Action" (ECA) constraints, each constraint representing a specific care step. Individual patient records were then systematically aligned with these predefined constraints using structured data extracted from eMRs. For example, Kaur et al [27] translated the nutrition recommendations into "rules" that dictate actions based on specific conditions or criteria using logical expressions or attributes (day of life, birth weight, and current health

XSL•FO
RenderX

conditions). Similarly, Lichtner et al [36] used guideline recommendations in a FHIR-based declarative format and built executable ECA rules based on the criteria specified in the recommendations and then executed against standardized patient data to determine whether the patients' data aligned with the recommendations. This rule-based alignment enables automated, transparent, and scalable assessment of adherence to best practice standards across large patient cohorts.

Beyond rule-based alignment, studies [20,29,48] incorporated weighting mechanisms to reflect the clinical significance of different types of deviations. Instead of treating all deviations from best practice standards as equally significant, these weighted scoring systems distinguished between addition, omission, and substitution of care activities and assigned differential weights. Implementation typically involves differentiating between types of activities that mismatch the prescribed activities in CPGs and CPs and those observed in patient data. In this context, omissions refer to activities specified in CPGs and CPs but not observed in the patient event trace, while additions refer to activities observed in the patient event trace that are not prescribed in CPGs and CPs.

By using weighting mechanisms, the adherence score reflects not only alignment with the referenced CPGs and CPs but also the severity of deviations. For example, van de Klundert et al [20] developed and tested dynamic programming formulations to match patient trajectories to the prescribed CPs, allowing partial adherence and assigning varying penalties to deviations based on their severity. Konrad et al [48] not only assigned equal weights to the addition and omission of activities compared to what was prescribed in the CP, but also to whether patient outcomes remained within target ranges during specified time intervals. Meanwhile, Zema et al [29] did not consider weighting patient outcomes, but they emphasized documentation quality in their Clinical Pathway Deviation Index (CPDI). CPDI was the weighted sum of 5 indicators: addition, omission, variation in activities sequence and duration, and missing in documentation. While the least weight was assigned to missing in documentation, equivalent weights were assigned to the other indicators.

Process mining techniques were used in a subset of studies to reconstruct care processes from event logs or event traces. Authors in several studies [29,30,32] used conformance checking and an optimal alignment algorithm to detect deviations. Deviations were identified by misalignment between observed activities and expected pathways, with some studies applying A* based algorithm—a graph traversal and path-finding algorithm—to determine the shortest or most efficient match between the observed and expected activities.

Unlike traditional health service research, which relies on single-point indicators, these studies adopted a process-level alignment approach. Rather than evaluating isolated components of CPGs and CPs, they aimed to align entire sequences of care activities with components extracted from best practice standards. This approach captures both what and when actions were performed, thus reflecting the temporal flow of care delivery. For example, Yan et al [26] described the modeling of the relative temporal relations between 2 clinical activities. They used temporal rules expressed in the Semantic Web Rule Language (SWRL) to describe the 30-minute interval between administering intravenous antibiotics and the cesarean delivery treatment intervention, as shown in Textbox 1 [26]. By analyzing these sequences, researchers can capture deviations in timing and sequencing, providing a more holistic and dynamic assessment of adherence to best practice standards.

Textbox 1. Example of modeling the relative temporal relations between clinical activities using a Semantic Web Ontology Language (OWL).

begins(IVPuchAntibiotics, ?t1) ^

begins(CesareanDelivery, ?t2) ^ *swrlb:*

subtractDateTimesYieldingDayTimeDuration(?t, ?t2, ?t1)

⇒*swrlb:greaterThan (?t,T30M).*

Despite these methodological innovations, most studies focused on structural or process-level compliance and did not consider contextual factors such as patient conditions, comorbidities, or clinician intent. As a result, the distinction between warranted deviations and unwarranted deviations remained largely unresolved across the included studies.

A wide variety of methods were used to represent patient data, including system-generated event logs, patient event traces, and FHIR. Similarly, CIG models were constructed using BPMN, ontology frameworks, FHIR, or hybrid combinations of these. These models aim to enable structured and machine-readable, potentially interoperable standard-based guideline representation exchange format [36].

While not all studies detailed explicit methods for measuring adherence, the most common approach was rule-based alignment, wherein individual patient records were assessed against predefined care steps derived from CPGs and CPs. Some studies incorporated weighting mechanisms to differentiate the clinical significance of deviations, reflecting the recognition that not all deviations from recommendations are negative.

Three principal insights were identified. First, a diverse range of formalism types was used to represent best practice standards, though few were implemented in a clinically integrated fashion.

Discussion

Principal Findings

This scoping review explored existing literature on the computable representations of best practice standards and methods for detecting and quantifying deviations from best practice using data extracted from eMRs. We identified 24 studies [20,26-46,47,48], a majority of which were published as conference proceedings. Cardiovascular disease was the most common clinical domain, most likely due to the extensive evidence available of CPGs in this clinical specialty [49].

XSL•FO
RenderX

Second, approaches to measuring adherence varied significantly, ranging from rule-based alignment to more nuanced methods involving weighted scoring and process mining. Finally, the effectiveness of adherence measurement was closely linked to the quality and structure of the underlying data sources, which varied across studies.

Together, these findings demonstrate both the progress made and the challenges that remain in operationalizing and measuring adherence to best practice using eMR-derived data. The results of this study highlight the need for standardized, context-sensitive, and interoperable systems that can support meaningful adherence monitoring to best practice standards as well as more quality and peer-reviewed research in the area.

Representing Best Practice: Challenges in Formalization

Translating narrative CPGs and CPs into structured, computer-interpretable formats is inherently complex. The narrative language used in CPGs is often characterized by ambiguity, conditional logic, and implicit assumptions, making it difficult to convert into electronic formats without loss of fidelity. Many studies in this review used hierarchical decomposition of narrative text into modular components combined with temporal logic and decision rules to form the basis of CIGs. However, even when the logic is accurately encoded, clinicians may find it challenging to validate outputs from these systems as CIG languages are not always accessible for clinicians [50].

Moreover, the global implementation of these systems is often constrained by their design. Many of them were developed as standalone tools that do not communicate with each other or with existing health information systems [51]. Additionally, these systems are also highly tailored to specific use cases and cannot be easily modified or maintained, except by their original authors, making them unsustainable for broader adoption [17,52,53].

BPMN has the advantage of being a graphical representation that can visualize workflow processes, as demonstrated in the included studies (Table 2). In addition, a systematic review has also reinforced BPMN's potential to standardize and redesign clinical processes [54]. However, there are certain limitations to using BPMN to represent best practice standards. Despite being excellent at modeling workflow, BPMN is primarily designed to model generic business processes and may not be able to capture the complexity and dynamics of health care processes [55,56]. Implementing BPMN in modeling clinical pathways requires a considerable investment in finance, time, and human resources [57]. At the same time, its technical complexity also poses a barrier to adoption among clinicians, especially those without formal training in informatics.

These limitations emphasize the need for models that balance clinical expressiveness with technical interoperability and usability, ensuring sustainability, scalability, and the ability to provide meaningful insights into clinical practice.

Measuring Adherence: Methodological Tensions

Approaches to measuring adherence to best practice standards varied significantly, reflecting a wide methodological spectrum. Traditional approaches to measuring adherence to best practice standards often rely on external, aggregated metrics, such as length of stay, cost, infection rate, mortality, etc. Although these indicators can provide valuable insights at a system level, they offer little information about the underlying process of care, as well as patient-level information [58]. Such measures are insufficient to determine whether care was delivered in accordance with the sequence, timing, and logic prescribed by best practice standards.

Several studies in this review sought to overcome this barrier by aligning process-level components with guideline-specified components and patient data. These approaches thus provide a more granular view of whether the intended sequence and logic of care delivery were followed. Despite their methodological sophistication, most of these models did not integrate contextual data such as patient comorbidities, preferences, or clinical urgency.

This gap presents persistent tension. While current models are effective at detecting technical deviation, they remain limited in their ability to distinguish between clinically necessary deviations from those that are undesirable. The future approach to adherence measurement must evolve to be more patient-centered and context-aware.

Limitations of Process Mining and Data Source Fidelity

Clinical decision-making is primarily guided by a patient's health status and context, which are often outside the stated space that the traditional process mining techniques rely on [59]. Many process mining techniques can tell us which treatment behaviors are performed and in which order in accordance with established best practice standards [60]. These methods are attractive due to their scalability and ability to visualize workflow variation. However, their utility in clinical settings is often limited by the eMR data structure and granularity [61].

The effectiveness of process mining techniques is often hindered by the inconsistent, incomplete, or unstructured nature of eMR data. A recent scoping systematic review shows that process mining cannot distinguish the differences in deviations due to the complex nature of clinical data [62]. In addition, the authors mentioned that process mining becomes more beneficial as the input data become more structured, which is usually not the case, especially for data extracted from eMR [62].

Consequently, the accuracy of deviation detection in process mining is heavily dependent on data fidelity. Even when the result produced by process mining is technically significant, it may lack clinical relevance to clinicians.

Since clinical decision-making is primarily driven by patient-specific factors, many of which are not captured in structured eMR fields, process mining risks oversimplifying the complex care delivery process. Without integration of outcome data, clinician notes, or patient-specific modifiers,

XSL•FO
RenderX

these methods offer only a partial view of adherence and may not resonate with clinical end users.

Recent Approach Using FHIR for Interoperability and Real-Time Monitoring of Adherence to Best Practice Standards

Among the most promising recent developments in this domain is the adoption of FHIR. As an emerging interoperability standard, FHIR aims to advance the messaging standard to enhance the exchange and interpretation of information across disparate and heterogeneous systems [63]. In the included studies, FHIR was used to represent patient data in a format that supports automated adherence analysis. Therefore, the use of FHIR not only provides concise structural data input but also promotes the reusability of clinical data for research and accelerates data availability.

A recent scoping review, conducted by Tabari et al [64], provided a summary of the tools and techniques used to represent the FHIR-based data model, demonstrating the promising use of FHIR to facilitate the integration, transformation, and analysis of clinical data. Indeed, in this scoping review, as shown in Table 2, FHIR was used to represent best practice standards rather than solely represent patient data. FHIR was demonstrated to successfully translate guideline recommendations into a standardized, computer-interpretable format, coupled with a rating system that can assess the level of certainty of evidence [45]. The authors also performed a study on automating the integration of guideline representation with real-time clinical data. They found a promising possibility of real-time monitoring adherence to guideline recommendations [36]. In addition, a recent publication demonstrates that it is possible to set up a trustworthy process to develop computer-interpretable representations of quality indicators for future automated quality management systems [65]. As a result, advancements in adopting FHIR to monitor adherence to best practice standards demonstrate the possibility of a real-time adherence monitoring system that not only detects deviations but also weights them according to the certainty of the supporting evidence.

Future Approach to Move Toward a More Meaningful Evaluation of Appropriate Care Delivery

Given the limitations of eMR data in measuring adherence to best practice standards, further research should aim to improve the applicability of eMR data for this purpose. This could be addressed by assessing the quality of eMR data characteristics—accuracy, completeness, reliability, relevance, and timelines—before evaluating adherence to CPGs and CPs. Furthermore, it could be beneficial to investigate the use of large language model technologies for the representation of CPGs and CPs and the measurement of adherence to best practice standards.

Despite these methodological innovations, most studies primarily focus on structural adherence, that is, whether the documented care sequence matched the prescribed components of CPGs and CPs, rather than on whether it was clinically appropriate in the specific care context. Consequently, while deviations from best practice standards can be detected with high accuracy, the crucial distinction between warranted deviations, which reflect high-quality patient-centered care, and unwarranted deviations, which can result in patient harm and unnecessary use of health care resources, remained largely unresolved.

Future studies can address this limitation by accounting for patient characteristics, outcomes, and clinician intent. For instance, detected deviations could be reclassified based on patient-specific conditions, allowing for outcome-informed validation of adherence measurement. By combining rule-based alignment and contextual reasoning, future approaches can advance beyond static measurement of adherence to best practice standards toward a more meaningful evaluation of appropriate care delivery. Additionally, future research should consider other potential factors influencing adherence to best practice standards, such as the structure of medical billing items, which could lead to deviations from the CPGs and CPs.

This scoping review makes several novel contributions to the existing literature by providing a comprehensive mapping of the computable representation of best practice standards and how adherence is measured—2 domains that have often been examined separately. Unlike previous reviews that focus only on either the technical approach or clinical evaluation, this scoping review synthesizes evidence from both perspectives and demonstrates how computer-interpretable format representation enables automated adherence assessment. It thus establishes a conceptual link between the modeling framework used for computable representation and the analytical methods used for evaluating practice variation. Notably, the findings promote the shift toward dynamic, computerized real-time monitoring of best practice standards, which can serve as a foundation for health care systems aiming to leverage eMR data for quality improvement. This also highlights the potential for a scalable, transparent, and clinically meaningful system for monitoring best practices to reduce unwarranted variation, strengthen guideline implementation, and ultimately improve patient care.

Limitations

This scoping review has some limitations, including the restriction to English-language publications and databases indexed up to November 2025, which may have excluded relevant literature, especially from non–English-speaking countries. We also did not register this study protocol on PROSPERO. In addition, due to the heterogeneity of study designs, a formal quality appraisal was not performed. We also focused on the quantitative measurement of adherence to best practice standards. Other works had been done on investigating the deviations from best practice using qualitative approaches, which were not included and analyzed in this scoping review. However, the qualitative type of studies often focus on providing insights into the nature of deviations, as demonstrated in the studies done by Kahol et al [66], which may not fit well with our aims, as our study focuses on quantitative measurements. We also explored studies that proposed conceptual models rather than real-world implementation, limiting the generalizability of findings. Finally, we sourced studies with data extracted from

XSL•FO
RenderX

eMR without considering additional data sources which again could provide the context of guideline deviations.

Conclusions

This scoping review offers an innovative contribution by surveying and synthesizing the existing literature that highlights the significant progress made in the formalization of best practice standards and the development of quantitative measures for adherence monitoring. However, significant gaps remain. The representation of best practice standards is heterogeneous and often disconnected from clinical workflows. Adherence measurement tends to prioritize detectability over clinical relevance. Although various data sources have been leveraged for these approaches, many remain incomplete or underused.

The increasing adoption of FHIR highlights the potential to improve eMR-based adherence measurement; however, technical and operational barriers must be addressed to achieve these goals. To realize the full potential of eMR-based adherence monitoring, future work must focus on clinically meaningful modeling, standardized yet flexible adherence metrics, and real-time, interoperable systems that integrate seamlessly with care delivery. Such advances will be critical to supporting evidence-based, adaptive, and patient-centered health care systems.

Acknowledgments

This research is supported by the Digital Health Cooperative Research Centre Limited (DHCRC). DHCRC is funded under the Australian Commonwealth's Cooperative Research Centres (CRC) Program. Author MTH holds a higher degree of research scholarship funded through a DHCRC project grant. The authors would like to thank Malhar Soni for screening the articles for this scoping review.

Funding

The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

All extraction data from the included studies are presented in Multimedia Appendix 1. All other data for title and abstract screening and full-text screening are available upon request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed properties of included studies.
[DOCX File, 32 KB - jmir_v28i1e79937_app1.docx]

Checklist 1
PRISMA checklist for scoping review.
[DOCX File, 111 KB - jmir_v28i1e79937_app2.docx]

Checklist 2
PRISMA-S checklist search strategies and search strings.
[DOCX File, 26 KB - jmir_v28i1e79937_app3.docx]

References

1. Barbieri A, Vanhaecht K, Van Herck P, et al. Effects of clinical pathways in the joint replacement: a meta-analysis. BMC Med 2009 Jul 1;7(1-11):32. [doi: 10.1186/1741-7015-7-32] [Medline: 19570193]
2. Brusco NK, Sykes K, Cheng AC, et al. A state-wide implementation of a whole of hospital sepsis pathway with a mortality based cost-effectiveness analysis from a healthcare sector perspective. PLOS Glob Public Health 2023;3(5):e0000687. [doi: 10.1371/journal.pgph.0000687] [Medline: 37205639]
3. Thursky K, Lingaratnam S, Jayarajan J, et al. Implementation of a whole of hospital sepsis clinical pathway in a cancer hospital: impact on sepsis management, outcomes and costs. BMJ Open Qual 2018;7(3):e000355. [doi: 10.1136/bmjoq-2018-000355] [Medline: 30019016]
4. Connor L, Dean J, McNett M, et al. Evidence-based practice improves patient outcomes and healthcare system return on investment: findings from a scoping review. Worldviews Evid Based Nurs 2023 Feb;20(1):6-15. [doi: 10.1111/wvn.12621] [Medline: 36751881]
5. Field MJ, Lohr KN, editors. Clinical practice guidelines. In: Directions for a New Program: National Academies Press; 1990. [doi: 10.17226/1626] [Medline: 25144032]

XSL•FO
RenderX

6. A guide to the development, implementation and evaluation of clinical practice guidelines. National Health and Medical Research Council. 1999. URL: https://www.nhmrc.gov.au/sites/default/files/images/a-guide-to-the-development-and-evaluation-of-clinical-practice-guidelines.pdf [accessed 2026-01-15]
7. Rotter T, Jong RB, Lacko SE, Ronellenfitsch U, Kinsman L, editors. Clinical pathways as a quality strategy. In: Improving Healthcare Quality in Europe 2019. URL: https://www.ncbi.nlm.nih.gov/books/NBK549262 [accessed 2026-01-15]
8. Hay MC, Weisner TS, Subramanian S, Duan N, Niedzinski EJ, Kravitz RL. Harnessing experience: exploring the gap between evidence-based medicine and clinical practice. J Eval Clin Pract 2008 Oct;14(5):707-713. [doi: 10.1111/j.1365-2753.2008.01009.x] [Medline: 19018899]
9. Kennedy PJ, Leathley CM, Hughes CF. Clinical practice variation. Med J Aust 2010 Oct 18;193(S8):S97-S99. [doi: 10.5694/j.1326-5377.2010.tb04021.x] [Medline: 20955142]
10. Campanella P, Lovato E, Marone C, et al. The impact of electronic health records on healthcare quality: a systematic review and meta-analysis. Eur J Public Health 2016 Feb;26(1):60-64. [doi: 10.1093/eurpub/ckv122] [Medline: 26136462]
11. Jaspers MWM, Smeulers M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. J Am Med Inform Assoc 2011 May 1;18(3):327-334. [doi: 10.1136/amiajnl-2011-000094] [Medline: 21422100]
12. Garg AX, Adhikari NKJ, McDonald H, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005 Mar 9;293(10):1223-1238. [doi: 10.1001/jama.293.10.1223] [Medline: 15755945]
13. Boxwala AA, Peleg M, Tu S, et al. GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines. J Biomed Inform 2004 Jun;37(3):147-161. [doi: 10.1016/j.jbi.2004.04.002] [Medline: 15196480]
14. Seyfang A, Miksch S, Marcos M. Combining diagnosis and treatment using ASBRU. Int J Med Inform 2002 Dec 18;68(1-3):49-57. [doi: 10.1016/s1386-5056(02)00064-3] [Medline: 12467790]
15. Tu SW, Campbell JR, Glasgow J, et al. The SAGE Guideline Model: achievements and overview. J Am Med Inform Assoc 2007;14(5):589-598. [doi: 10.1197/jamia.M2399] [Medline: 17600098]
16. Riaño D, Peleg M, Ten Teije A. Ten years of knowledge representation for health care (2009-2018): topics, trends, and challenges. Artif Intell Med 2019 Sep;100:101713. [doi: 10.1016/j.artmed.2019.101713] [Medline: 31607346]
17. Peleg M. Computer-interpretable clinical guidelines: a methodological review. J Biomed Inform 2013 Aug;46(4):744-763. [doi: 10.1016/j.jbi.2013.06.009] [Medline: 23806274]
18. Clinical Quality Language specification. CQL. URL: https://cql.hl7.org [accessed 2026-01-15]
19. Kardas P, Aguilar-Palacio I, Almada M, et al. The need to develop standard measures of patient adherence for big data: viewpoint. J Med Internet Res 2020 Aug 27;22(8):e18150. [doi: 10.2196/18150] [Medline: 32663138]
20. van de Klundert J, Gorissen P, Zeemering S. Measuring clinical pathway adherence. J Biomed Inform 2010 Dec;43(6):861-872. [doi: 10.1016/j.jbi.2010.08.002] [Medline: 20696277]
21. Sohn S, Moon S, Prokop LJ, Montori VM, Fan JW. A scoping review of medical practice variation research within the informatics literature. Int J Med Inform 2022 Sep;165:104833. [doi: 10.1016/j.ijmedinf.2022.104833] [Medline: 35868231]
22. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
23. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
24. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
25. Covidence systematic review software, Veritas Health Innovation. Covidence. 2025. URL: www.covidence.org [accessed 2026-01-15]
26. Yan Y, Zhibin J, Dong Y, Gang D. A semantics-based clinical pathway workflow and variance management framework. Presented at: 2008 IEEE International Conference on Service Operations and Logistics, and Informatics; Oct 12-15, 2008. [doi: 10.1109/SOLI.2008.4686499]
27. Kaur R, Jain M, McAdams RM, et al. An ontology and rule-based clinical decision support system for personalized nutrition recommendations in the neonatal intensive care unit. IEEE Access 2023;11:142433-142446. [doi: 10.1109/ACCESS.2023.3341403]
28. Yu HW, Hussain M, Afzal M, et al. Use of mind maps and iterative decision trees to develop a guideline-based clinical decision support system for routine surgical practice: case study in thyroid nodules. J Am Med Inform Assoc 2019 Jun 1;26(6):524-536. [doi: 10.1093/jamia/ocz001] [Medline: 31087071]
29. Zema M, Rosati S, Duran Carvajal JE, Balestra G. CPDI: an index for measuring deviations in clinical pathways. Annu Int Conf IEEE Eng Med Biol Soc 2015 Aug;2015:1385-1388. [doi: 10.1109/EMBC.2015.7318627] [Medline: 26736527]
30. Huang ZX, Bao YR, Dong W, Lu XD, Duan HL. Online treatment compliance checking for clinical pathways. J Med Syst 2014 Oct;38(10):123. [doi: 10.1007/s10916-014-0123-0] [Medline: 25149871]
31. Li X, Liu H, Zhang S, et al. Automatic variance analysis of multistage care pathways. Stud Health Technol Inform 2014;205:715-719. [doi: 10.3233/978-1-61499-432-9-715] [Medline: 25160280]

XSL•FO
RenderX

32. Xu H, Yan H, Pang J, Nan S, Yang X, Zhao D. Evaluating the relative value of care interventions based on clinical pathway variation detection and propensity score. Presented at: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 16-19, 2020; Seoul, Korea (South p. 1184-1187 URL: https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=9312958 [doi: 10.1109/BIBM49941.2020.9313546]
33. Yan H, Lu X, Van Gorp P, et al. On accurate, automated and insightful deviation analysis of clinical protocols. Presented at: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 3-6, 2018; Madrid, Spain p. 1130-1134. [doi: 10.1109/BIBM.2018.8621133]
34. Noshad M, Rose CC, Chen JH. Signal from the noise: a mixed graphical and quantitative process mining approach to evaluate care pathways applied to emergency stroke care. J Biomed Inform 2022 Mar;127(no):104004. [doi: 10.1016/j.jbi.2022.104004] [Medline: 35085813]
35. Yan H, Van Gorp P, Kaymak U, et al. Variance analysis in Task-Time matrix clinical pathways. Presented at: 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI); Orland, FL, USA p. 253-256. [doi: 10.1109/BHI.2017.7897253]
36. Lichtner G, Spies C, Jurth C, et al. Automated monitoring of adherence to evidenced-based clinical guideline recommendations: design and implementation study. J Med Internet Res 2023 May 4;25:e41177. [doi: 10.2196/41177] [Medline: 36996044]
37. Dos Santos Leandro G, Moro CMC, Cruz-Correia RJ, Portela Santos EA. FHIR implementation guide for stroke: a dual focus on the patient's clinical pathway and value-based healthcare. Int J Med Inform 2024 Oct;190:105525. [doi: 10.1016/j.ijmedinf.2024.105525] [Medline: 39033722]
38. Minutolo A, Esposito M, Pietro G, Soonchunhyang U, Icar Cnr I, Networking V, et al. Fuzzy on FHIR: a decision support service for healthcare applications. Presented at: International Conference on P2P, Parallel, Grid, Cloud and Internet Computing; Republic of Korea p. 163-172. [doi: 10.1007/978-3-319-49109-7_16]
39. Ugon A, Bouzid AIH, Jaulent MC, Favre M, Duclos C, Jobez E, et al. Building a knowledge-based tool for auto-assessing the cardiovascular risk. Stud Health Technol Inform 2018;247:735-739. [Medline: 29678058]
40. Shivers J, Amlung J, Ratanaprayul N, Rhodes B, Biondich P. Enhancing narrative clinical guidance with computer-readable artifacts: authoring FHIR implementation guides based on WHO recommendations. J Biomed Inform 2021 Oct;122:103891. [doi: 10.1016/j.jbi.2021.103891] [Medline: 34450285]
41. Oliveira T, Neves J, Costa A, Novais P, Neves J, et al. An interpretable guideline model to handle incomplete information. In: Omatu S, De Paz Santana J, González S, Molina J, Bernardos A, Rodríguez J, editors. Distributed Computing and Artificial Intelligence Advances in Intelligent and Soft Computing: Springer, Berlin, Heidelberg; 2012, Vol. 151. URL: https://link.springer.com/chapter/10.1007/978-3-642-28765-7_52#citeas [accessed 2026-01-27] [doi: 10.1007/978-3-642-28765-7_52]
42. Martínez-Salvador B, Marcos M, Palau P, Domínguez Mafé E. A model-driven transformation approach for the modelling of processes in clinical practice guidelines. Artif Intell Med 2023 Mar;137:102495. [doi: 10.1016/j.artmed.2023.102495] [Medline: 36868689]
43. Hu Z, Li JS, Yu HY, Zhang XG, Suzuki M, Araki K. Modeling of clinical pathways based on ontology. Presented at: 2009 IEEE International Symposium on IT in Medicine & Education; Aug 14-16, 2009. [doi: 10.1109/ITIME.2009.5236262]
44. Madhusanka S, Walisadeera A, Dantanarayana G, Goonetillake J, Ginige A. An ontological clinical decision support system based on clinical guidelines for diabetes patients in Sri Lanka. Healthcare (Basel) 2020 Dec 18;8(4):573. [doi: 10.3390/healthcare8040573] [Medline: 33352875]
45. Lichtner G, Alper BS, Jurth C, et al. Representation of evidence-based clinical practice guideline recommendations on FHIR. J Biomed Inform 2023 Mar;139(104305):104305. [doi: 10.1016/j.jbi.2023.104305] [Medline: 36738871]
46. Autexier S, Bawadekji M, Hutter D, Wolters R. Supporting clinical guidelines using DL-temporal reasoning. Presented at: 2013 10th International Conference & Expo on Emerging Technologies for a Smarter World (CEWIT); Oct 21-22, 2013; Melville, NY p. 1-6. [doi: 10.1109/CEWIT.2013.6713766]
47. Farkash A, Timm JTE, Waks Z. A model-driven approach to clinical practice guidelines representation and evaluation using standards. Stud Health Technol Inform 2013;192:200-204. [Medline: 23920544]
48. Konrad R, Tulu B, Lawley M. Monitoring adherence to evidence-based practices: a method to utilize HL7 messages from hospital information systems. Appl Clin Inform 2013;4(1):126-143. [doi: 10.4338/ACI-2012-06-RA-0026] [Medline: 23650493]
49. Writing Committee Members, Virani SS, Newby LK, et al. 2023 AHA/ACC/ACCP/ASPC/NLA/PCNA Guideline for the management of patients with chronic coronary disease: a report of the American Heart Association/American College of Cardiology Joint Committee on Clinical Practice Guidelines. J Am Coll Cardiol 2023 Aug 29;82(9):833-955. [doi: 10.1016/j.jacc.2023.04.003] [Medline: 37480922]
50. Martínez-Salvador B, Marcos M. Supporting the refinement of clinical process models to computer-interpretable guideline models. Bus Inf Syst Eng 2016 Oct;58(5):355-366. [doi: 10.1007/s12599-016-0443-3]
51. Wong D, Peek N. Does not compute: challenges and solutions in managing computable biomedical knowledge. BMJ Health Care Inform 2020 Jul;27(2):e100123. [doi: 10.1136/bmjhci-2019-100123] [Medline: 32723852]

XSL•FO
RenderX

52. De Clercq P, Kaiser K, Hasman A. Computer-interpretable guideline formalisms. Stud Health Technol Inform 2008;139:22-43. [Medline: 18806319]
53. Iglesias N, Juarez JM, Campos M. Business Process Model and Notation and openEHR task planning for clinical pathway standards in infections: critical analysis. J Med Internet Res 2022 Sep 15;24(9):e29927. [doi: 10.2196/29927] [Medline: 36107480]
54. De Ramón Fernández A, Ruiz Fernández D, Sabuco García Y. Business Process Management for optimizing clinical processes: a systematic literature review. Health Informatics J 2020 Jun;26(2):1305-1320. [doi: 10.1177/1460458219877092] [Medline: 31581880]
55. Rebuge Á, Ferreira DR. Business process analysis in healthcare environments: a methodology based on process mining. Inf Syst 2012 Apr;37(2):99-116. [doi: 10.1016/j.is.2011.01.003]
56. Pufahl L, Zerbato F, Weber B, Weber I. BPMN in healthcare: challenges and best practices. Inf Syst 2022 Jul;107:102013. [doi: 10.1016/j.is.2022.102013]
57. Scheuerlein H, Rauchfuss F, Dittmar Y, et al. New methods for clinical pathways-Business Process Modeling Notation (BPMN) and Tangible Business Process Modeling (t.BPM). Langenbecks Arch Surg 2012 Jun;397(5):755-761. [doi: 10.1007/s00423-012-0914-z] [Medline: 22362053]
58. van der Lubbe SCC, Goh LH, Kontopantelis E, Tam WW, Valderas JM. Measuring guideline concordance via electronic health records: a new model for estimating concordance scores. BMJ Qual Saf 2025;PMID:bmjqs-b2025. [doi: 10.1136/bmjqs-2025-018494]
59. Grüger J, Geyer T, Kuhn M, Braun SA. Eighted violations in alignment-based conformance checking. In: Bergmann R, editor. Process Mining Workshop: Springer. [doi: 10.1007/978-3-031-27815-0_21]
60. Munoz-Gama J, Martin N, Fernandez-Llatas C, Johnson OA, Sepúlveda M. Innovative informatics methods for process mining in health care. J Biomed Inform 2022 Oct;134:104203. [doi: 10.1016/j.jbi.2022.104203] [Medline: 36113758]
61. Augusto A, Deitz T, Faux N, Manski-Nankervis JA, Capurro D. Process mining-driven analysis of COVID-19's impact on vaccination patterns. J Biomed Inform 2022 Jun;130:104081. [doi: 10.1016/j.jbi.2022.104081] [Medline: 35525400]
62. Oliart E, Rojas E, Capurro D. Are we ready for conformance checking in healthcare? Measuring adherence to clinical guidelines: a scoping systematic literature review. J Biomed Inform 2022 Jun;130:104076. [doi: 10.1016/j.jbi.2022.104076] [Medline: 35525401]
63. Stan O, Miclea L. Local EHR management based on FHIR. Presented at: 2018 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR); May 24-26, 2018. [doi: 10.1109/AQTR.2018.8402719]
64. Tabari P, Costagliola G, De Rosa M, Boeker M. State-of-the-art Fast Healthcare Interoperability Resources (FHIR)-based data model and structure implementations: systematic scoping review. JMIR Med Inform 2024 Sep 24;12(1):e58445. [doi: 10.2196/58445] [Medline: 39316433]
65. von Dincklage F, Bublitz VK, Kumpf O, et al. Computer-interpretable quality indicators for intensive care medicine: development and validation study. J Med Internet Res 2025 Sep 26;27:e77077. [doi: 10.2196/77077] [Medline: 41004802]
66. Kahol K, Vankipuram M, Patel VL, Smith ML. Deviations from protocol in a complex trauma environment: errors or innovations? J Biomed Inform 2011 Jun;44(3):425-431. [doi: 10.1016/j.jbi.2011.04.003] [Medline: 21496496]

Abbreviations

BPMN: Business Process Model and Notation
CDSS: Clinical Decision Support Systems
CIG: computer-interpretable guideline
CP: clinical pathway
CPDI: Clinical Pathway Deviation Index
CPG: clinical practice guideline
CQL: Clinical Quality Language
ECA: Event-Condition-Action
eMR: electronic medical record
FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level Seven
NRL: Natural Rule Language
OCL: Object Constraint Language
OWL: Semantic Web Ontology Language
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: An Extension of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Statement
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
SWRL: Semantic Web Rule Language

XSL•FO
RenderX

Edited by S Brini; submitted 01.Jul.2025; peer-reviewed by FV Dincklage, G Lichtner, X Huang; accepted 27.Nov.2025; published 09.Feb.2026.

Please cite as:
Hoang MT, Donnelly C, Igasto C, Shetty A, Pradhan M, Shaw T
Data-Driven Guideline Adherence in Data Representation and Compliance Measurement: Scoping Review
J Med Internet Res 2026;28:e79937
URL: https://www.jmir.org/2026/1/e79937
doi:10.2196/79937

© Minh Trang Hoang, Candice Donnelly, Christina Igasto, Amith Shetty, Malcolm Pradhan, Tim Shaw. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effects of Digital Health Interventions to Promote Safer Sex Behaviors Among Youth: Systematic Review and Bayesian Network Meta-Analysis

Yiran Zhu[1*], MSc; Wenwen Peng[1*], MSc; Die Hu[1], BN; Edmond Pui Hang Choi[2], PhD; Maritta Anneli Välimäki[3,4], PhD; Ci Zhang[1], PhD; Xianhong Li[1,5], PhD

[1]Xiangya School of Nursing, Central South University, 172 Tongzipo Road, Changsha, China

[2]School of Nursing, University of Hong Kong, Hong Kong SAR, China (Hong Kong)

[3]School of Public Health, Faculty of Medicine, University of Helsinki, Helsinki, Finland

[4]Helsinki University Hospital, Nursing Research Center (NRC), Helsinki, Finland

[5]JBI Xiangya Research Centre for Evidence-based Healthcare Innovation, Central South University, Changsha, China

[*]these authors contributed equally

Corresponding Author:
Xianhong Li, PhD
Xiangya School of Nursing, Central South University, 172 Tongzipo Road, Changsha, China

Abstract

Background: Youth aged 15 - 24 years carry a disproportionate HIV/sexually transmitted infections (STIs) burden. In recent years, different modalities of digital health interventions (DHIs) have been explored to promote safer sex behaviors among youth, but their comparative effectiveness across modalities and relative to nondigital interventions (NDIs) remains unclear.

Objective: This study aimed to compare DHI modalities on safer sex behaviors and HIV/STI incidence, rank modalities using Bayesian network meta-analysis (NMA), and position their effectiveness relative to NDIs.

Methods: A systematic review and Bayesian NMA of randomized controlled trials were conducted by comprehensively searching PubMed, EMBASE, Web of Science, and Cochrane Library (inception to November 2025). Eligible studies were those that enrolled youth aged 15 - 24 years and evaluated mobile app-based intervention, telecommunication-based intervention (TCI), static web-based intervention (SWI), or interactive online-based intervention (IOI)—with an NDI or another DHI. Primary outcomes were condom use at last sexual contact, consistent condom use, and proportion of condom use. Secondary outcomes included condom use self-efficacy, number of sexual partners, and STI incidence (including HIV). Risk of bias was assessed with the Cochrane Risk of Bias 2 tool, and certainty of evidence with GRADE/CINeMA (Confidence in NMA). Bayesian random-effects NMAs estimated odds ratios (ORs) with 95% credible intervals (CrIs), and complementary frequentist NMAs provided 95% CIs and 95% prediction intervals.

Results: Twenty-four randomized controlled trials (20,134 participants) were included, forming treatment networks across 5 intervention types. TCI was the only intervention that significantly improved condom use at last sex compared with NDI (OR 1.13, 95% CrI 1.02 - 1.26). For consistent condom use, SWI and IOI outperformed TCI (SWI vs TCI: OR 1.77, 95% CrI 1.03 - 3.06; IOI vs TCI: OR 1.68, 95% CrI 1.02 - 2.76). For the proportion of condom use, IOI outperformed SWI (OR 1.34, 95% CrI 1.01 - 1.80), and mobile app-based intervention ranked highest in probability rankings, though estimates lacked precision. For STI incidence, NDI was associated with fewer STIs than SWI (OR 0.61, 95% CrI 0.46 - 0.82).

Conclusions: This is the first NMA to compare the effectiveness of DHIs on condom use and HIV/STI outcomes among youth populations. It demonstrates that the impact of DHIs on HIV prevention varies substantially by intervention modality and outcome type. While TCI demonstrates the most consistent improvement in condom use at last sex, SWI and IOI may be more effective for promoting consistent condom use, though estimates remain imprecise. However, wide prediction intervals and low-certainty evidence suggest that self-reported behavioral changes may not translate into reductions in HIV/STI incidents without integration with offline services and broader structural support. Future trials might consider including standardized outcome indicators and longer follow-up to generate more precise estimates of the effectiveness of DHIs and guide generalization of youth-centered digital HIV/STIs prevention.

Trial Registration: PROSPERO CRD42024527317; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024527317

(J Med Internet Res 2026;28:e87071) doi:10.2196/87071

XSL•FO
RenderX

KEYWORDS

digital health intervention; HIV prevention; safer sex behavior; youth; mobile health; telecommunication-based intervention; web-based intervention; network meta-analysis; mHealth

Introduction

Adolescents and young adults aged 15 - 24 years, defined as "youth" by the United Nations [1-3], are disproportionately affected by HIV around the globe. This age range is widely used in international health research and reporting, which allows comparability across studies and alignment with global HIV surveillance data. Alarmingly, in 2023, youth in this age group comprised nearly one-third of the 3600 daily new HIV infections recorded worldwide. Youth are especially vulnerable to HIV due to high rates of unprotected sex, inconsistent condom use, and co-occurring risk behaviors such as alcohol and drug use [4,5]. Still, a significant proportion of youth around the world lack access to accurate and age-appropriate information on sexual and reproductive health, rendering them susceptible to misinformation, psychological distress, and engagement in high-risk sexual behaviors [6]. To address this public crisis, scalable evidence-based health interventions targeting safer sex practices should be prioritized in this vulnerable population [7].

Digital health interventions (DHIs) have emerged as a promising strategy for health promotion in recent years [8]. Digital health, conceptualized as an umbrella term by the World Health Organization [9], refers to the use of digital and wireless platforms to facilitate health care delivery or health interventions, including but not limited to electronic health, mobile health, telehealth, and artificial intelligence-based applications. On the other hand, with the growing accessibility of smartphones and internet services among youth, digital technologies have become a dominant force to shape their sexual behaviors [10,11]. It is more convenient for young people to meet sexual partners, including casual, one-night, and anonymous partners, through web-based platforms, dating apps, and social networking sites [12,13], which further increases the likelihood of having unprotected sex frequency [14]. While digital technologies have facilitated riskier sexual behaviors among youth, they also create opportunities for DHIs that leverage young people's existing online engagement patterns and preferences [15-19].

Accumulating evidence suggests that DHIs can improve HIV-related knowledge, risk perception, prevention intentions, and behavioral outcomes among youth [20], with the types of DHIs including mobile apps, text messaging, online videos, social media platforms, and interactive websites [21-23]. A stage-based computer-delivered intervention, for example, targeting heterosexual young men demonstrated significant improvements in condom use intention and subsequent condom use behavior [24]. Similarly, a study evaluating a social media-based intervention via Facebook reported a 23% increase in condom use and a 54% reduction in chlamydia incidence among adolescents [25]. In contrast, a large (randomized controlled trial (RCT) delivering sexual health promotion via SMS and email enhanced sexually transmitted infection (STI) knowledge and testing uptake, particularly among women, but showed no significant impact on condom use [26]. Another study reported that intervention based on a peer-led safer-sex Facebook group for Chinese college students found no significant change in contraceptive use intention or frequency [27]. Similarly, a social media-based crowdsourced HIV testing intervention among youth did not increase facility-based HIV testing, condom use, or syphilis testing [28]. Therefore, different types of DHIs may differentially affect sexual health outcomes, yet existing trials rarely distinguish the relative effectiveness of each DHI modality. Clarifying which intervention types are most effective for specific behavioral and biological outcomes is essential for optimizing digital HIV prevention strategies among youth [7,29].

In addition, several systematic reviews (SRs) have synthesized the evidence on DHIs for HIV prevention, but important limitations remain. Some SRs are purely descriptive, lacking quantitative synthesis [22,30-32]. Other reviews have focused narrowly on specific DHI types (eg, social media or telehealth) [33,34], or have failed to examine key behavioral outcomes like condom use [35,36]. In addition, traditional meta-analyses are constrained to pairwise comparisons [37], leaving uncertainty about which types of DHIs are most effective in head-to-head comparisons [30]. To address these knowledge gaps, we conducted a SR and network meta-analysis (NMA) to evaluate and compare the effectiveness of different DHIs in promoting safer sex behaviors among youth. The study aimed to: (1) identify the most effective types of DHIs in promoting safe sex among youth; (2) construct a network-based ranking of intervention effectiveness; and (3) inform the design of scalable, evidence-based digital health programs for HIV prevention among youth.

Methods

Overview

This SR and NMA follow the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA-NMA) guidelines [38]. The completed PRISMA-NMA checklist is provided in Checklist 1. The protocol for this study has been registered with PROSPERO (CRD42024527317).

Eligibility Criteria

Types of Population

Studies were eligible for inclusion if they involved participants aged 15 - 24 years or if at least half of the participants were within this age range. Those that did not report, or for which data could not be extracted, for this specific age group were excluded.

Types of Interventions and Comparison

The interventions included in this review were DHIs, which were defined in accordance with the World Health Organization's broad definition of digital health technologies [9]. For the purpose of this review, the included DHIs were further classified into four mutually exclusive categories based

XSL•FO
RenderX

on their delivery modes and characteristics: (1) mobile app-based interventions (MAIs), (2) telecommunication-based interventions (TCI), (3) static web-based interventions (SWIs), and (4) interactive online-based interventions. This operational classification was developed to reflect the interventions identified in the included studies and to avoid overlap across categories. A detailed description of common DHI subcategories was provided in Table 1; the subcategories of the DHIs were based on a previous SR [30]. The control group received nondigital interventions (NDI), referring to traditional approaches without digital technology, such as face-to-face counseling, printed materials, or group sessions.

Table . Subcategories of digital health interventions (DHIs) and abbreviations used to classify interventions evaluated.

Subcategories of DHIs	Abbreviation	Description
Mobile app–based interventions	MAI	Programs delivered primarily via dedicated software apps installed on smartphones or tablets, leveraging device features (eg, notifications, sensors, data storage) to provide interactive content, personalized feedback, tracking, and behavior change support.
Telecommunication-based intervention	TCI	Interventions using traditional telecommunication methods such as SMS text messages or telephone calls. These interventions typically involve sending reminders, educational messages, or conducting counseling via phone communication without the need for internet-based platforms.
Static web-based interventions	SWI	Interventions provided through websites that offer static, noninteractive content. This may include informational pages, downloadable resources, or educational materials without features for user engagement or real-time feedback.
Interactive online-based interventions	IOI	Interventions delivered via web-based platforms or websites that enable user interaction, such as quizzes, tailored feedback, chatbots, or real-time communication with health professionals. These platforms actively engage users to enhance learning and behavior change.

Any appropriate comparator group was included, such as usual care, placebo, no intervention, waitlist, attention control, or different DHIs. To reduce inconsistency among trials, we excluded studies that combined non-DHIs with DHIs, unless the distinction between the intervention and control groups lay solely in the DHIs. In multi-arm trials, intervention arms representing the same modality without meaningful differences in content, intensity, or delivery were considered a single treatment node for eligibility purposes and later combined analytically to avoid double-counting.

Outcomes

The primary outcomes were specific condom use behaviors, defined as follows:

1. Condom Use Rate in the Last Sexual Contact: The percentage of individuals reporting condom use during their most recent penetrative sexual act [39].
2. Consistent Condom Use Rate: The percentage of individuals reporting consistent condom use during all their penetrative sexual acts over the recall period specified in each study [40].
3. Proportion of Condom Use: The overall proportion of sexual acts in which a condom was used, calculated as the total number of times a condom was used divided by the total number of sexual acts. Unlike the consistent condom use rate, which measures whether individuals always use a condom, this indicator captures the frequency of condom use across all reported sexual encounters, allowing for partial or occasional use [41].

The secondary outcomes included (1) self-efficacy for condom use, measured by the overall mean score on a validated condom use self-efficacy scale, such as Brafford and Beck's [42] condom use self-efficacy scale, Lawrance et al's [43] self-efficacy for HIV prevention scale and others, (2) number of sexual partners, and (3) the incidence rate of STIs (including HIV). Because follow-up length varied substantially across trials, we included studies reporting at least one postintervention follow-up outcome and extracted the longest follow-up time point for synthesis to enhance comparability [44,45].

Types of Studies

Only RCTs were included, including crossover trials and cluster-randomized trials. Studies using nonrandomized, quasi-experimental, observational, or qualitative designs were excluded. Only peer-reviewed articles published in English were eligible, as non-English or non-peer-reviewed sources lack sufficient methodological detail for reliable data extraction and risk-of-bias assessment.

Search Strategy

The search strategy was developed and reported in accordance with the PRISMA-S guideline [46]. Searches for RCTs were conducted in PubMed (including MEDLINE), EMBASE, Web of Science, and the Cochrane Library. Searches were performed

XSL•FO
RenderX

through the native interfaces of each database (PubMed via NCBI, EMBASE via Elsevier, Web of Science via Clarivate, and Cochrane Library via Wiley). In addition, the reference lists of relevant SRs were checked to ensure that no eligible trials were missed.

The search terms were formulated according to the PICOs framework, including participants or populations, interventions, outcomes, and types of research design. Both Medical Subject Headings and free-text terms were included as appropriate. Boolean operators ("AND," "OR") were used to combine search terms, and database-specific search techniques such as truncation, phrase marks, and wildcards were applied. The complete search terms and algorithm were provided in Multimedia Appendix 1, and search strategies for the other databases were adapted accordingly. The search was designed and executed by 2 reviewers (YZ and WP), who were trained in SR methodology, and the strategy was cross-checked for completeness and accuracy.

The literature search was initially conducted on June 13, 2024 and was last updated on November 15, 2025. All retrieved records were imported into EndNote X9 for citation management, and duplicates were removed using both automated and manual deduplication. Additional search methods included checking the reference lists of the included studies or relevant SRs. Gray literature was also searched via Google Scholar, OpenGrey, and ProQuest Dissertations. The search was limited to studies published in English due to resource constraints for translation.

Selection Procedure and Data Extraction

Two reviewers (YZ and DH) independently screened the titles and abstracts against predefined protocol criteria. Full texts were retrieved for all potentially eligible studies. When multiple articles were identified from the same randomized controlled trial, the most recent or most comprehensive publication was retained for data extraction. Earlier reports were used to supplement missing information on study design, intervention details, or outcomes when necessary. Any discrepancies between the 2 reviewers were resolved by discussion. If disagreements persisted, a third reviewer (WP) was invited for adjudication. At the title and abstract screening stage, we excluded 14,788 records that clearly did not meet the eligibility criteria, most commonly because of wrong study design (eg, cross-sectional surveys, qualitative studies, reviews, protocols), wrong population (non-youth samples), ineligible intervention or comparator (ie, the difference between study arms did not lie in the use of a DHI), or an unrelated topic.

Data were extracted using a standardized and piloted form. Extracted variables included: first author, publication year, recruitment region, participant characteristics (mean age, SD, gender distribution), type of intervention and comparator, sample size per arm, intervention duration, and reported endpoints. Detailed characteristics of the DHIs were also extracted to facilitate subcategorization (Table 1). Further, outcomes and corresponding measurement methods were recorded, such as self-reported condom use and validated self-efficacy scales.

When outcome data were incomplete or unclear, study authors were contacted by email for clarification; trials with essential missing data were excluded from the quantitative synthesis and documented in Multimedia Appendix 2. All eligible studies were included in the SR, and only studies with usable and connected outcome data were included in the NMA.

Risk of Bias Assessment

The methodological quality of the included studies was independently assessed by 2 reviewers (YZ and WP), with disagreements resolved by a third reviewer (CZ). Risk of bias was evaluated using version 2 of the Cochrane risk of bias 2 tool (RoB 2) for randomized trials [47]. For each domain, studies were rated as having "low risk," "some concerns," or "high risk" of bias according to the Cochrane Handbook (version 6.5) [48]. Domain-level risk of bias judgments for each trial are summarized in Figure 1, with extended graphs and contribution matrices provided in Multimedia Appendix 3. The reference list of included studies is provided in Multimedia Appendix 4 [26,49-71].

XSL•FO
RenderX

Figure 1. Risk of bias assessment of included randomized controlled trials using Cochrane Risk of Bias 2 tool [26,49-71].

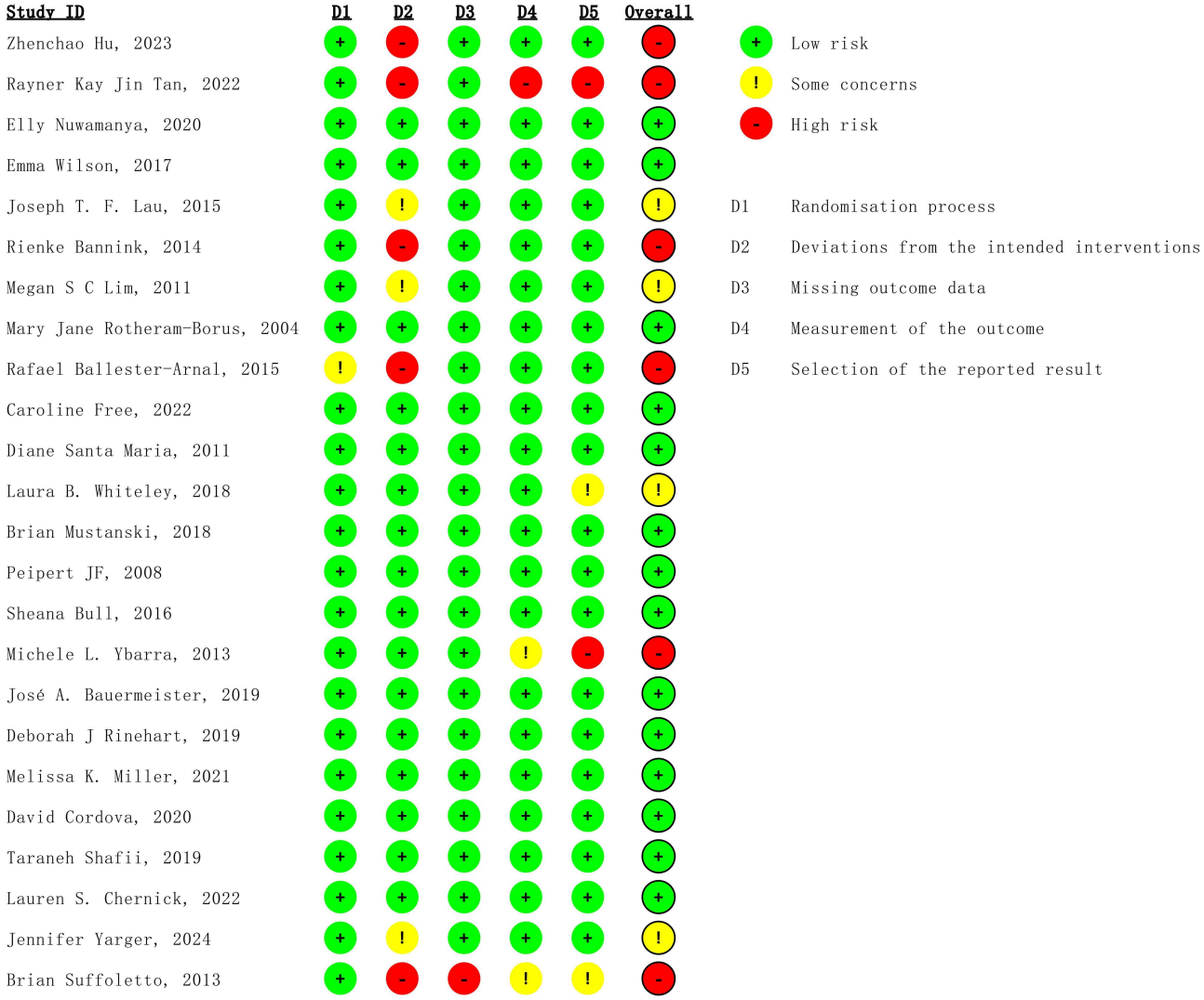

Because blinding was generally not feasible for these nonpharmacological interventions, many trials were judged at high risk of bias in the domain of deviations from intended interventions [72,73]. As this limitation was expected and unlikely to influence objectively measured outcomes, we did not consider this domain when grading the certainty of evidence. The overall certainty of evidence was determined using the CINeMA (Confidence in NMA) web application, which is based on the GRADE framework [74,75]. In addition, we constructed GRADE "Summary of Findings" tables using the official template provided by the GRADE Working Group to summarize the key relative and absolute effects and certainty ratings for the primary outcomes (Multimedia Appendix 3).

RoB 2 assessments were incorporated into the interpretation of NMA findings and into the GRADE/CINeMA evaluation of certainty, but they were not used to weight studies in the statistical synthesis.

Statistical Analysis

Geometry of the Evidence Network

We examined the geometry of the treatment network by mapping each trial arm to one of the predefined intervention nodes and summarizing the pattern of direct comparisons. A network plot was generated to visually depict the evidence base, with node size proportional to the number of randomized participants and edge thickness reflecting the number of trials informing each comparison. We further assessed potential network-related biases by identifying sparse nodes, single-study comparisons, and imbalance in the distribution of direct evidence.

Model Specification and Synthesis Methods

Model convergence was assessed through Markov Chain Monte Carlo diagnostics, including the Gelman-Rubin potential scale reduction factor and inspection of leverage plots. The number of adaptation iterations, burn-in period, and total iterations were set to ensure adequate mixing and convergence. Effect estimates were expressed as pooled odds ratios (ORs) and 95% credible intervals (CrIs), which served as the primary summary measure for all dichotomous outcomes.

Bayesian NMAs were conducted using the R package *BUGSnet* to compare the effectiveness of 4 subcategories of DHIs and control groups. Binomial likelihood models with a logit link function were specified, and both fixed-effect and random-effects consistency models were fitted. Given anticipated clinical heterogeneity, random-effects models were treated as primary, with fixed-effect models used in sensitivity analyses. Noninformative priors were assigned to treatment

XSL•FO
RenderX

effects and heterogeneity parameters to minimize prior influence. Model fit and parsimony were evaluated using the deviance information criterion (DIC), with lower values indicating better fit.

To evaluate the transitivity assumption, we compared mean age, sex distribution, intervention intensity, and follow-up duration across treatment comparisons. We further restricted inclusion to trials in which ≥50% of participants were aged 15 - 24 years, excluded trials in which nondigital components were offered only to one arm, and extracted outcomes at the longest reported follow-up to harmonize follow-up time. These design and population characteristics showed broadly overlapping ranges across interventions and no systematic differences between comparisons, so transitivity was judged plausible. Forest plots of posterior ORs with 95% CrIs from the Bayesian consistency model were generated to summarize the magnitude and uncertainty of estimated treatment effects.

To complement these Bayesian estimates and quantify uncertainty in effects that might be observed in new settings, we also performed frequentist random-effects NMAs using the netmeta package in R, specifying NDI as the reference group [76]. For each outcome, we estimated ORs and 95% CIs for each intervention versus NDI and derived 95% prediction intervals (PIs) by combining the average treatment effect with between-study heterogeneity, in line with recent recommendations that NMAs should routinely report PIs when heterogeneity is present [77].

Assessment of Inconsistency and Heterogeneity

Consistency between direct and indirect evidence was assessed by comparing the DIC between consistency and inconsistency models. A substantially lower DIC in the consistency model indicated acceptable agreement between sources of evidence. Due to the limited number of included studies, we did not formally investigate small-study effects (eg, using funnel plots or Egger's regression test), which are typically used to explore potential publication bias as one of several possible explanations for such effects. Selective outcome reporting could not be formally assessed due to insufficient reporting in the included trials; however, the potential for selective reporting was considered when interpreting the cumulative evidence. Because the number of studies informing most comparisons was limited, local inconsistency (eg, node-splitting) could not be reliably assessed; in the presence of any potential inconsistency, we planned to explore differences in study characteristics and reassess the plausibility of the transitivity assumption.

Handling of Multi-Arm Trials and Node Merging

For multi-arm trials, if 2 or more arms delivered essentially the same intervention category (eg, different versions of the same SWI content without meaningful variation in delivery or timing), we merged these arms by summing the number of events and participants. This ensured that each intervention was represented by a single node in the network and avoided duplicate contributions from the same trial [78].

Ranking of Interventions

Ranking probabilities and surface under the cumulative ranking curve (SUCRA) were computed to summarize the relative effectiveness of each intervention across the posterior distribution. Rankograms and cumulative ranking plots were used to visualize intervention hierarchies, and league tables and heatmaps were generated to present pairwise comparisons and their relative effect estimates. No additional analyses, such as sensitivity analyses, subgroup analyses, or meta-regression, were conducted because the limited number of studies and the sparse network geometry did not permit reliable implementation of these methods. All statistical analyses were conducted using R (version 4.3.2; R Foundation for Statistical Computing) with the gemtc, BUGSnet, and netmeta packages.

Results

Description of Included Studies

From a total of 25,659 records initially retrieved, 24 RCTs published between 2004 and 2024 were included in the final analysis (Figure 2). These studies were conducted across 8 diverse countries, predominantly in the United States (n=14), with the remainder from China (n=2), the United Kingdom (n=2), Uganda (n=2), Singapore (n=1), the Netherlands (n=1), Australia (n=1), and Spain (n=1). The trials collectively enrolled 20,134 participants (range 50 - 6248; mean 838.9, SD 1358.2), with a mean age of 19.5 years. Overall, 10,228 participants (53.4%) were male, although sex composition varied substantially—some studies enrolled only males [49-51], only females [52,53], or mixed populations. Of the 24 included studies, 21 studies were two-arm, and 3 studies were multi-arm. Across the included studies, 6 trials evaluated TCI, 8 assessed interactive online-based intervention (IOI), 6 examined MAI, and 8 investigated SWI. The total number of intervention approaches (n=28) exceeded the number of included studies (n=24) because several trials directly compared 2 or more active interventions (eg, IOI vs SWI) without including a conventional control group. Intervention durations ranged from brief sessions lasting 10 - 20 minutes up to 12 months. Follow-up periods were heterogeneous, spanning from immediate postintervention assessments to 24 months. Most studies reported outcomes at 3 - 6 months, while only a few provided longer-term follow-up beyond 12 months. Five studies performed analyses for more than one time point. To enhance consistency and reduce potential bias associated with short-term variability, we extracted outcomes at the longest follow-up time point, thereby facilitating a more comprehensive evaluation of the intervention's long-term effectiveness [44,45].

XSL•FO
RenderX

Figure 2. The flow diagram of the literatureease clarify th selection process for randomized controlled trials of digital health interventions included in this review.

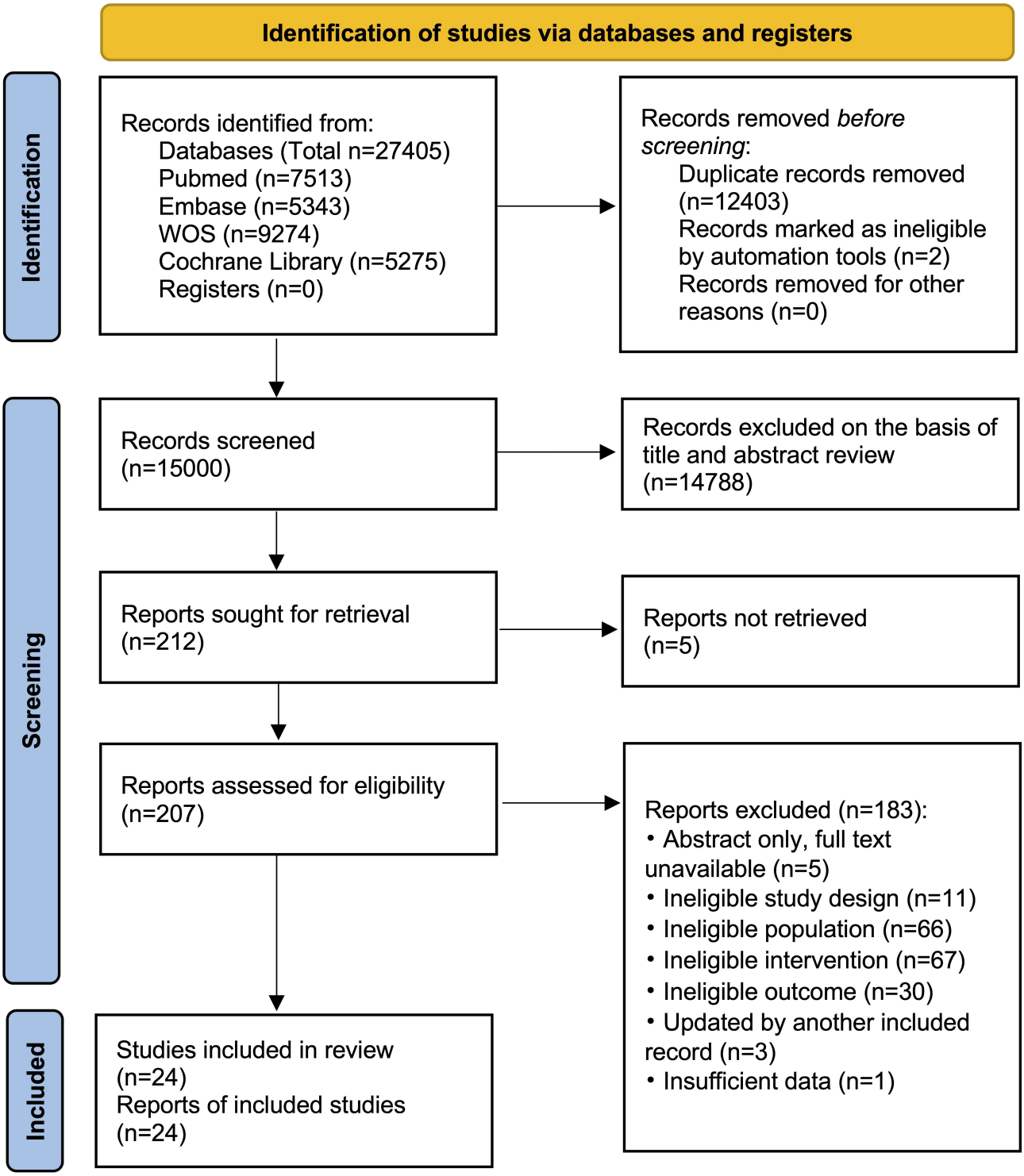

For condom use at last sexual contact, 4 out of 7 studies reported ORs greater than 1, suggesting a possible beneficial effect of the interventions; however, only one trial [54] showed a clear statistical significance (OR 1.13, 95% CI 1.01 - 1.25). Regarding consistent condom use, 6 of 11 studies showed ORs above 1, with substantial heterogeneity; one study reported a very large effect (OR 4.55, 95% CI 1.15 - 17.95) [55]. For the proportion of condom use, 5 out of 6 trials reported ORs above 1, suggesting a tendency towards higher condom use in the intervention groups; however, CIs were wide and often included the null (overall OR range 0.48 - 2.43), indicating that the evidence for this outcome is imprecise. Finally, for STIs incidence, including HIV, effects varied substantially across 7 trials, with ORs ranging from 0.53 to 2.10. Four of the 7 trials had point estimates below 1 [49,50,53,79], and 3 trials had 95% CI that excluded 1 [49,56,57], indicating heterogeneous and partly conflicting evidence for this outcome. These results summarized the observed effects across trials, highlighting that effect estimates varied considerably across outcomes and

XSL•FO
RenderX

studies. The characteristics and effect estimates of the included studies were summarized in Table 2.

XSL•FO
RenderX

Table . Study characteristics and effect estimates of the included trials.

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	Endpoints	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
Hu et al [57]	China	RCT[a]	16.09 (0.84)	—[b]	1691 (53.67)	"You and Me	—	Online education sessions; Internet-based educational platform; cartoon videos; PowerPoint slides	IOI[c] / NDI[d]	1760/1391	45 min sessions/8 wk	End of intervention / 12 wk	①, ④	• 0.72 (0.60-0.86) ① • 1.42 (1.16-1.74) ④
Tan et al [50]	Singapore	RCT	23.90 (2.98)	—	300 (100.00)	People Like Us (PLU) web drama video series	Predetermined theory of behavior change	Web-based	SWI / NDI	150/150	6 videos, each about 10 min in length/1 wk	6 mo	②, ④	• 1.22 (0.99-1.51) ② • 0.82 (0.67-1.00) ④
Nuwamanya et al [58]	Uganda	RCT	21.00 (2.00)	—	407 (36.60)	MPA-SRH (mobile phone apps-sexual reproductive health)	—	Mobile app	MAI / NDI	556/556	Within 6 mo, open for use to the intervention group.	End of intervention	①	• 1.23 (0.99-1.53)

XSL•FO
RenderX

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	Endpoints	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
Wilson et al [56]	The United Kingdom	RCT	23.00 (3.55)	—	846 (41.01)	SH:24 website	—	Website, online service	SWI[e] / NDI	1031/1032	All participants were free to use any other sexual health services or interventions during the trial period/Median =28.8 d (participant-dependent duration)	6 wk	④	• 2.10 (1.40)
Lau et al [51]	China	RCT	—	—	396 (100.00)		—	Online videos (Sent Intervention package email）	SWI / NDI	261/135	10 - 20 min	3 mo	③	• 1.10 (0.97)

XSL•FO
RenderX

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	Endpoints	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
						Online intervention based on STD-related cognitions involving videos (SC)/online intervention based on both STD-related cognitions and emotions (eg, fear) involving fear-arousing imagery and videos (SCFI)								
Bannink et al [59]	Netherlands	RCT	15.81 (0.68)	—	446 (54.00)	The E-health4U Intervention	—	Web-based tailored messages	IOI / NDI	392/434	1 mo	4 mo	②	•
Lim et al [26]	Australia	RCT	—	19	417 (41.99)	Email and SMS to a group of young people (intervention gro	—	SMS, email	TCI[f] / NDI	507/486		end of intervention (6 mo) / end of intervention (12 mo)	②	•

XSL•FO
RenderX

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	Endpoints	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
											SMS messages were sent every 3 - 4 wk (a total of 14) emails were sent less than monthly (a total of 8)/12 mo			
Rotheram-Borus et al [60]	The United States	RCT	—	23	136 (77.71)	Telephone intervention	—	Telephone sessions	TCI / NDI	59/116	2 h/6 wk	15 mo	②	• 0.76 (0.49-1.19)
Ballester-Arnal et al [61]	Spain	RCT	20.90 (1.90)	—	60 (25.00)	Fear induction group/website group	Information-Motivation-Behavioral Skills (IMB) Model	Video, music/website; computer based	SWI / NDI	Estimated 34 - 35 participants per group (total n=239; balanced allocation across 7 groups)	1 h	1 months / 4 mo	③	• 1.54 (0.89-2.64)
Free et al [54]	The United Kingdom	Parallel group RCT	20.35 (2.1)	—	2162 (34.60)	text messaging intervention (safetxt)	COM-B (capability, opportunity, motivation, and behavior) model	Delivered by text messages to improve safer sex behaviors	TCI / NDI	312/331/25		4 weeks / 12 mo	①, ④	• 1.13 (1.01-1.25) ① • 1.11 (0.98-1.25) ④

XSL•FO
RenderX

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	Endpoints	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
											Day 1‐3: 4 items per day Day 4‐28: 1‐2 items per day Second month: 2‐3 items per week Month 3‐12: 2‐5 items per month/12 mo			
Santa Maria et al [62]	The United States	Pilot RCT	21.20 (2.10)	—	56 (52.34)	MYRID (Motivating Youth to Reduce Infection and Disconnection)	Information-Motivation-Behavioral Skills (IMB) Model	Smartphone-based Just-in-Time Adaptive Intervention (JITAI) Android smartphones with unlimited data were provided to participants	MAI[g] / NDI	48/49		end of intervention	②	•

XSL•FO
RenderX

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	End-points	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
											Participants will receive a customized message for each EMA assessment completed.• Weeks 1 - 2: 3 times per day• Weeks 3 - 4: 2 times per day• Weeks 5 - 6: 1 time per day• Each EMA took 1 - 5 min to complete/6 wk			
Whiteley et al [55]	The United States	Pilot RCT	18.60 (2.30)	—	37 (61.67)	Online HIV/STI[h] prevention intervention	Information-Motivation-Behavioral Skills (IMB) Model		SWI / NDI	31/29		3 mo	②	• 45 (1-19)

XSL•FO
RenderX

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	End-points	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
								Publicly available websites and YouTube videos related to HIV/STI prevention Content accessible via computer, smartphone, or tablet			Twice per week for 4 wk, each contained 2 - 3 website or video links Total exposure: 8 emails with up to 19 different resources/4 wk			
Mustanski et al [49]	The United States	RCT	23.82[i]	—	901 (100.00)	"Keep It Up!"	Information-Motivation-Behavioral Skills (IMB) model	Fully online (eHealth intervention), delivered via computers and tablets (not available on mobile phones).	IOI / SWI	445/456	Each session lasted about 1 h./≥3 d for core intervention (3 sessions ≥24 h apart); booster sessions at 3 and 6 mo	12 mo	③, ④	• 1.34 (0.[illegible]-1.80) ③ • 0.56 (0.[illegible]-0.89) ④
Peipert et al [53]	The United States	RCT	—	22	0 (0)	Project PROTECT	Transtheoretical (TTM) model of behavior change	Computer-based multimedia program	IOI / SWI	272/270	Three computer-based sessions /80 d	24 mo	②, ④	• 0.99 (0.[illegible]-1.38) ② • 0.96 (0.[illegible]-1.53) ④
Bull et al [63]	The United States	RCT	14.94 (1.08)	—	415 (48.71)		Integrated Theory of mHealth	Text messages based on social media	TCI / NDI	436/416	5 and 7 messages weekly/25 wk	end of intervention	③	• 1.13 (0.[illegible]-3.53)

XSL•FO
RenderX

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	End-points	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
						Teen Outreach Program (TOP)+ text message program: Youth All Engaged (YAE!)								
Ybarra et al [64]	Uganda	RCT	16.10 (1.40)	—	307 (83.88)	CyberSenga	Information-Motivation-Behavioral Skills (IMB) model	Website based; computer-based	IOI / NDI	183/183	One module per week, a total of 5 modules need to be completed./5 wk	3 mo	②	• 0.91 (0.16)
Bauermeister et al [65]	The United States	Pilot RCT	21.67 (1.81)	—	123 (100.00)	myDEx(My Desires & Expectations)	The dual processing cognitive-emotional decision-making framework	Online-delivered HIV prevention interventions; App	MAI / SWI	95/28		end of intervention	③	• 2.43 (1.50)

XSL•FO
RenderX

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	Endpoints	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
											myDEx intervention includes 6 personalized online courses, completed within 3 mo, with participants logging in an average of about 5 times, and a total conversation volume of about 7 times/3 mo			
Rinehart et al [66]	The United States	Pilot RCT	15.90 (1.60)	—	0 (0)	Texts for Sexual Health Education and Empowerment (t4she)	Health belief model	Text message	TCI / NDI	122/122	58 automated messages sent over 12 wk/12 wk	3 months / 6 mo	①	• 1.40 (0.66-2.99)
Miller et al [67]	The United States	Pilot RCT	16.90 (1.00)	—	26 (28.57)	Sex-Health	Theory of Planned Behavior to inform intervention content and the Social Ecological Model		IOI / NDI	44/47	25 min	6 mo	①	• 0.50 (0.20-1.30)

XSL•FO
RenderX

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	End-points	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
								A tablet-based, interactive intervention: The educator used a tablet to deliver the intervention, intermittently sharing the screen with the participant.						
Cordova et al [68]	The United States	Pilot RCT	18.82 (2.1)	—	4 (8.00)	Storytelling 4 Empowerment (S4E)	An ecodevelopment and empowerment framework	Multilevel Mobile Health App	MAI / NDI	25/25	Complete 3 interactive modules at once /30 - 45 min	1 mo	③	• 0.48 (0.56)
Shafii et al [69]	The United States	Pilot RCT	21[i]	—	176 (64.71)	e-KISS	Information-Motivation-Behavioral Skills (IMB) model	An interactive computer-based intervention; video	IOI / NDI	130/142	15 - 20 min, a single, one-time interactive /15 - 20 min	2 mo	④	• 0.53 (0.10)
Chernick et al [52]	The United States	Pilot RCT	17.74 (1.27)	—	0 (0)			Multimedia text messaging; video; mobile-based	IOI / NDI	72/74		3 mo	①, ②	• 0.95 (0.21) ① • 1.53 (0.42) ②

XSL•FO
RenderX

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	End-points	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
						Dr. Erica (Emergency Room Interventions to improve the Care of Adolescents)	Intervention mapping, a program-planning framework; the Social Cognitive Theory and Motivational Interviewing				A minimum of 56 and maximum of 121 texts, with additional texts sent based on keywords./3 mo			
Yarger et al [70]	The United States	A Cluster Randomized Trial	15.7[i]	—	340 (40.72)	In the Know—an in-person, group-based sexual health education program integrating digital technologies,	—	Technology-based; mobile-based; app	MAI / NDI	348/487	1.5 h/4 wk	3 mo	②	• 0.86 (0.[illegible]-1.27)
Suffoletto et al [71]	The United States	Pilot RCT	21.44 (2.04)	—	0 (0)	SMS program	The Health Belief Model; the Information Motivation Behavior model	Text message	TCI / NDI	23/29	Once a week, send a text message at noon every Sunday/12 wk	3 mo	①, ②	• 1.86 (0.[illegible]-7.12) ① • 1.60 (0.[illegible]-6.96) ②

[a]RCT: randomized controlled trial.

[b]Not available.

[c]IOI: interactive online-based intervention.

[d]NDI: nondigital intervention.

[e]SWI: static web-based intervention.

[f]TCI: telecommunication-based intervention.

XSL•FO
RenderX

[g]MAI: mobile app-based intervention.

[h]STI: sexually transmitted infection.

[i]The included studies did not report SD values for these mean estimates, and the SDs cannot be derived from the available information.

Risk-of-bias assessments using the RoB 2 tool are summarized in Figure 1. Overall, most trials were judged to be at low risk of bias for the randomization process, outcome measurement, and selection of the reported result. However, a substantial minority of studies had some concerns or high risk of bias in at least one domain, most frequently for deviations from the intended interventions and missing outcome data. Consequently, several trials were rated as having some concerns or a high overall risk of bias.

Although condom use self-efficacy and number of sexual partners were prespecified as secondary outcomes, too few studies reported these measures to allow meta-analysis. Only one trial evaluated condom use self-efficacy. In Rinehart et al [66], this construct was assessed using 3 items developed within the Health Belief Model (range 0 - 12; Cronbach α=0.72). At the 3rd month, the intervention group reported significantly higher self-efficacy scores than the control group (7.38 vs 6.68; *P*=.04), but this difference was no longer significant at the 6th month (7.39 vs 6.99; *P*=.20). Two trials reported on the number of sexual partners. In Shafii et al [79], participants in the intervention arm reported a 29% reduction in the number of sexual partners at follow-up, although the effect did not reach statistical significance (IRR 0.71, 95% CI 0.50 - 1.03, *P*=.07). Changes in the control group were not reported. By contrast, Free et al [54] examined the proportion of participants reporting 2 or more sexual partners over 12 months. At one year, this outcome was reported by 56.9% of intervention participants compared with 54.8% of controls (OR 1.11, 95% CI 1.00 - 1.24, *P*=.06). Overall, the evidence on the impact of digital interventions on the number of sexual partners remains limited and inconsistent.

Results of Network Meta-Analysis

Overview

A total of 24 RCTs were included to evaluate the comparative effectiveness among 5 intervention types—4 DHIs (TCI, IOI, MAI, and SWI) and NDI—across the four analyzable outcomes: (1) condom use at last sexual contact, (2) consistent condom use, (3) overall proportion of condom use, and (4) incidence of STIs (including HIV). The remaining 2 outcomes of self-efficacy were excluded due to insufficient network connectivity. The network structures for each outcome were shown in Figure 3, where the thickness of the lines was proportional to the number of comparisons, and the size of the nodes reflected the number of studies involving each intervention. Across outcomes, the treatment network was dominated by comparisons of each DHI category versus NDI, whereas head-to-head trials comparing different DHIs were rare. Several DHI-DHI contrasts and some STI outcomes were informed by only one or two small trials, and self-efficacy outcomes formed disconnected subnetworks. Thus, the network geometry was relatively sparse and heavily anchored on NDI, implying that several treatment rankings rely mainly on indirect evidence. The indirect comparative effectiveness of DHIs was summarized in Multimedia Appendix 5. Forest plots of posterior ORs with 95% CrIs for each intervention versus NDI across all 4 outcomes are provided in Multimedia Appendix 6.

XSL•FO
RenderX

Figure 3. Network structure diagrams for randomized controlled trials of digital health interventions among youth, by outcome: (**A**) Condom use rate in the last sexual contact; (**B**) Consistent condom use rate; (**C**) Proportion of condom use; (**D**) The incidence rate of sexually transmitted infections (including HIV). The thicknesses of the lines were proportional to the number of comparisons; the diameters of the circles were proportional to the number of treatments. IOI: interactive online-based intervention; MAI: mobile app-based intervention; NDI: nondigital intervention; SWI: static web-based intervention; TCI: telecommunication-based intervention.

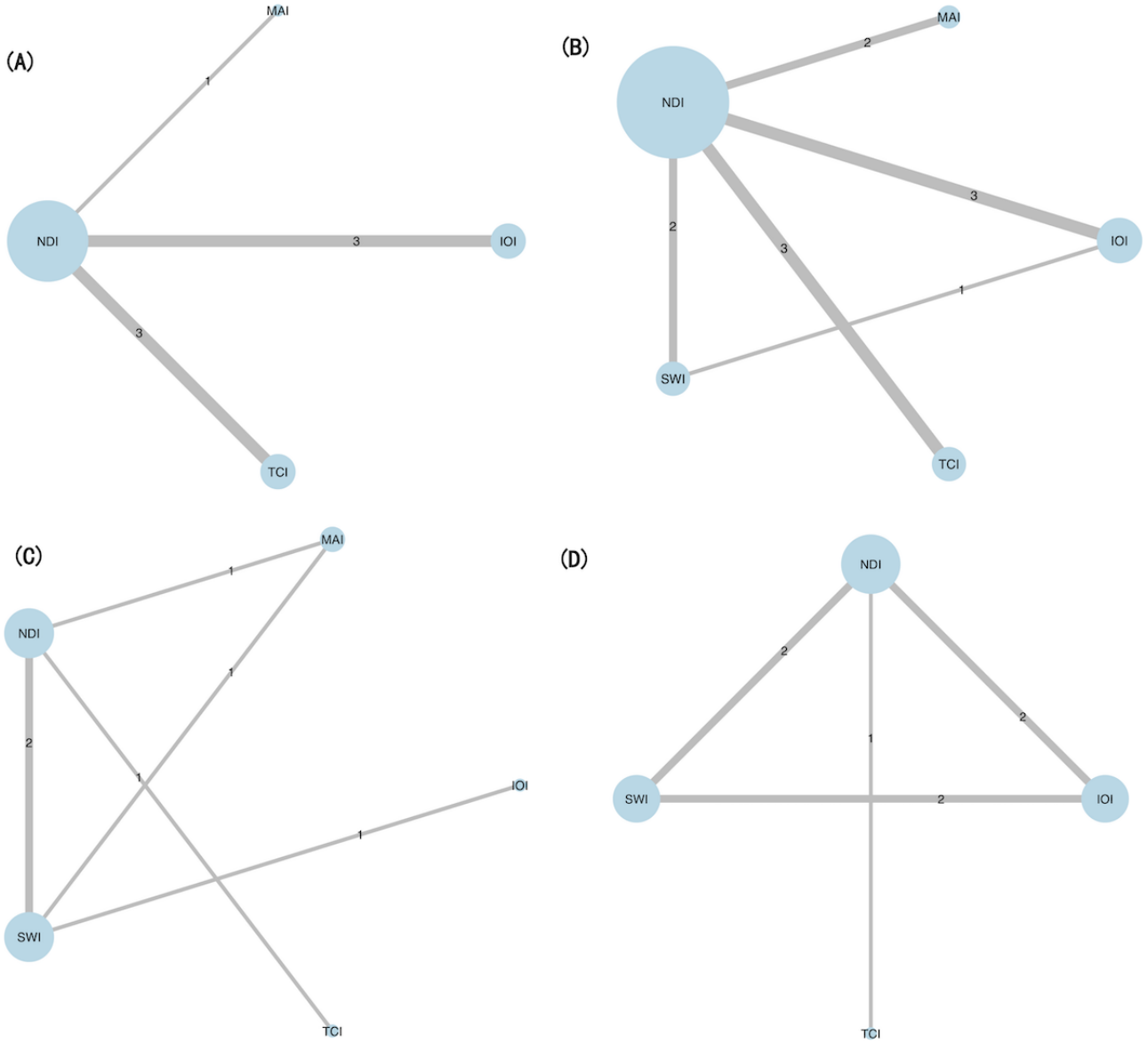

In the complementary frequentist random-effects NMAs, point estimates and 95% CIs for each intervention versus NDI were broadly consistent with the Bayesian results (Multimedia Appendix 7). Across all 4 outcomes, 95% PIs were noticeably wider than the corresponding CIs and frequently included the null value, even when the average effects suggested benefit. For example, for condom use at last sexual contact and for consistent condom use, TCI, IOI, and MAI tended to favor improved condom use versus NDI, but their PIs indicated that future trials conducted in different settings could plausibly observe smaller benefits or no clear difference from NDI. Similar patterns were observed for the proportion of condom-protected acts and for STI incidence, highlighting that between-study heterogeneity and contextual differences may lead to substantial variability in the effects realized in new populations.

Condom Use Rate in the Last Sexual Contact

Seven studies involving 4 DHIs with a total of 10,285 participants were included in the analysis of condom use at last sexual contact. The random-effects consistency model was selected based on model fit, as it showed comparable DIC and residual deviance values to the inconsistency model, indicating no substantial inconsistency. Among the interventions, only TCI showed a statistically significant improvement compared with NDI (OR 1.13, 95% CrI 1.02 - 1.26). Although MAI had the highest SUCRA value (83.44%) and was most likely to rank first (65.61%), its effect was not statistically significant. The rank probabilities for all interventions were summarized in Table 3A and illustrated in Figures 4 and 5, showing the descending order of MAI, TCI, NDI, and IOI. As shown in Multimedia Appendix 6, TCI was the only intervention with its 95% CrI entirely to the right of the line of no effect, suggesting a modest but relatively certain increase in condom use at last

XSL•FO
RenderX

sex compared with NDI. IOI and MAI showed point estimates on either side of 1 with wide CrIs, indicating no clear difference from NDI.

Table . Rank probabilities and surface under the cumulative ranking curve (SUCRA) values for digital health intervention categories in the network meta-analysis of randomized controlled trials assessing sexual health outcomes among youth.

Rank	IOI	MAI	NDI	SWI	TCI
(A) rank probability of condom use rate in the last sexual contact					
1	5.76	65.61	0.05	—[a]	28.58
2	12.02	87.34	8.58	—	92.05
3	20.34	97.38	82.46	—	99.82
4	100	100	100	—	100
SUCRA	12.71	83.44	30.36	—	73.48
(B) Rank probability of consistent condom use rate					
1	33.07	4.59	0.54	61.09	0.7
2	90.77	10.95	3.35	92.76	2.16
3	97.78	41.8	52.43	97.71	10.27
4	99.59	77.38	94.65	99.5	28.87
5	100	99.99	99.99	100	100
SUCRA	80.3	33.68	37.74	87.77	10.5
(C) Rank probability of proportion of condom use					
1	12.6	69.18	0.8	0.04	17.39
2	65.17	88.68	6.82	2.82	36.53
3	92.3	94.87	24.8	37.75	50.31
4	99.3	97.89	65.6	74.13	63.11
5	100	100	100	100	100
SUCRA	67.34	87.66	24.5	28.68	41.84
(D) Rank probability of the incidence rate of STIs (including HIV)					
1	0.38	—	91.78	0.04	7.81
2	7	—	99.95	0.43	92.63
3	95.79	—	100	4.5	99.72
4	100	—	100	100	100
SUCRA	34.39	—	97.25	1.66	66.72

[a]Not available.

XSL•FO
RenderX

Figure 4. Rank of probabilities of digital health intervention categories in the network meta-analysis of randomized controlled trials assessing sexual health outcomes among youth: (**A**) Condom use rate in the last sexual contact; (**B**) Consistent condom use rate; (**C**) Proportion of condom use; (**D**) The incidence rate of sexually transmitted infections (including HIV). Stacked bars show the probability that each digital health intervention category occupies each possible rank (from best to worst).

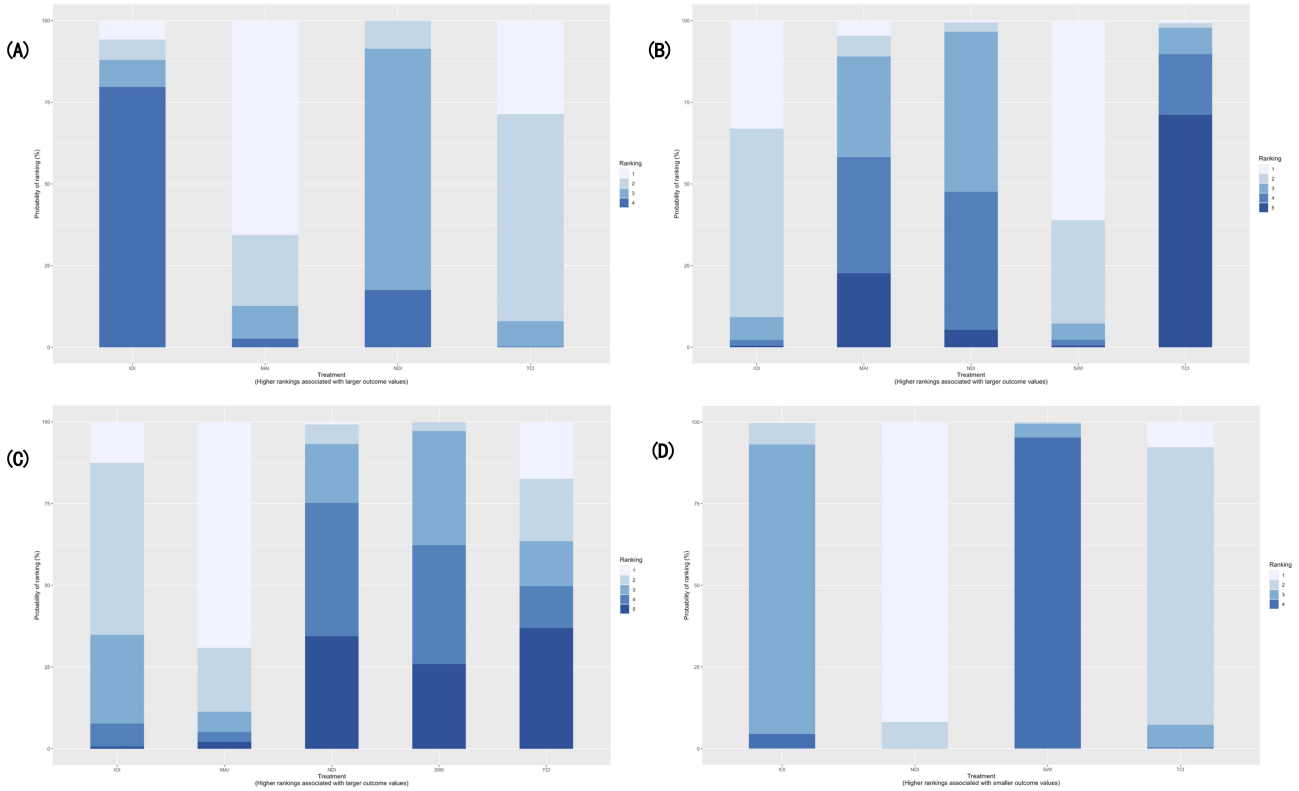

Figure 5. Cumulative rank plot for digital health intervention categories in the network meta-analysis of randomized controlled trials assessing sexual health outcomes among youth: (**A**) Condom use rate in the last sexual contact; (**B**) Consistent condom use rate; (**C**) Proportion of condom use; (**D**) The incidence rate of sexually transmitted infections (including HIV). Lines represent the cumulative probability of each digital health intervention category being ranked at or above each position.

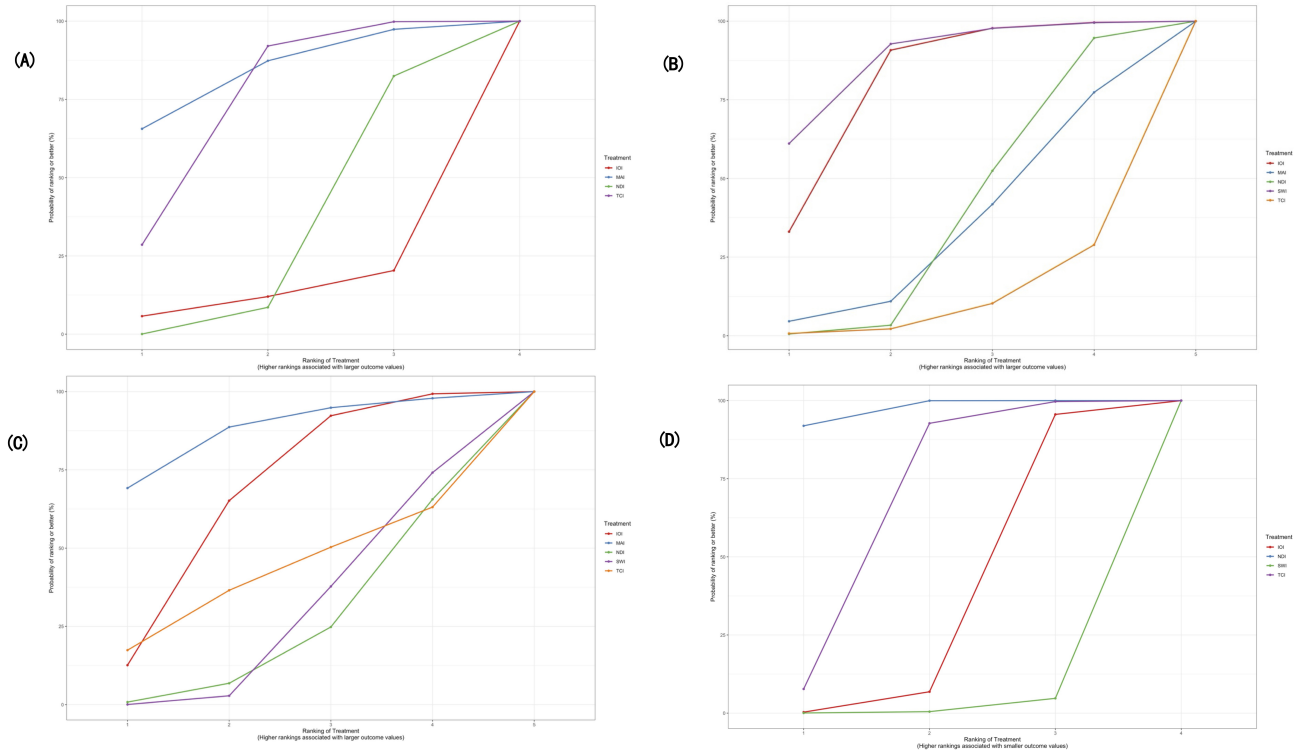

XSL•FO
RenderX

Consistent Condom Use Rate

Eleven studies involving all 5 DHIs with a total of 4881 participants were included in the analysis of consistent condom use. A random-effects consistency model was selected based on a slightly better model fit (DIC=39.49 vs 41.17) and no substantial evidence of inconsistency. Both SWI and IOI were significantly more effective than TCI (SWI vs TCI: OR 1.77, 95% CrI 1.03 - 3.06; IOI vs TCI: OR 1.68, 95% CrI 1.02 - 2.76). SWI had the highest probability of being the most effective intervention (SUCRA=87.77%), followed by IOI (80.3%). TCI ranked the lowest (10.5%), while MAI and NDI showed moderate effectiveness. The ranking probabilities were summarized in Table 3B and visualized in Figures 3B and 4B. Multimedia Appendix 6 shows that IOI and SWI had posterior ORs above 1, implying a tendency toward improved consistent condom use, whereas MAI and TCI showed ORs close to or below 1. However, all CrIs crossed 1, suggesting that these differences were uncertain.

Proportion of Condom Use

Six studies involving 5 DHIs with a total of 2048 participants were included in the analysis of the proportion of condom use. Given the slightly lower DIC (22.45 vs 22.83) and similar model complexity and fit, the consistency model was deemed preferable. Only IOI showed a statistically significant improvement compared with SWI (OR 1.34, 95% CrI 1.01 - 1.80). Regarding the ranking probabilities (Table 3C), MAI had the highest probability of being the most effective intervention (SUCRA=87.66%), followed by IOI (SUCRA=67.34%). In contrast, NDI (SUCRA=24.5%) and SWI (SUCRA=28.68%) ranked relatively low. The distribution of rank probabilities was presented in Table 3C and visualized in Figures 3C and 4C. In Appendix 4C, all interventions showed ORs >1 relative to NDI, with MAI and IOI having the largest point estimates. Nevertheless, the CrIs were wide and crossed 1, indicating that although the direction of effect generally favored digital interventions, the precision of the estimates was limited.

The Incidence Rate of STIs (Including HIV)

Seven studies involving 4 DHIs and a total of 14,966 participants were included in the analysis of STIs incidence. NDI was significantly more effective than IOI (OR 0.78, 95% CrI 0.65 - 0.93) and SWI (OR 0.61, 95% CrI 0.46 - 0.82), while TCI also showed a significant advantage over SWI (OR 0.67, 95% CrI 0.49 - 0.92). Based on rank probabilities and SUCRA values, NDI had the highest likelihood of being the most effective intervention (91.78% probability of ranking first; SUCRA=97.25%), followed by TCI (7.81%; SUCRA=66.72%). IOI and SWI had considerably lower rankings, with SUCRA values of 34.39% and 1.66%, respectively. Table 3D summarizes the intervention rankings, and Figures 3D and 4D illustrate the rank probability and cumulative ranking plots. Multimedia Appendix 6 displays ORs >1 for IOI, SWI, and TCI compared with NDI, and all CrIs lie entirely to the right of 1. This pattern suggests that these digital interventions were associated with equal or higher STI incidence, with SWI showing the highest point estimate.

Consistency and Visualization

For all 4 outcomes, consistency between direct and indirect comparisons was assessed by comparing the DIC values between the consistency and inconsistency models. In all cases, the DIC differences were less than 5, indicating no evidence of global inconsistency (Multimedia Appendix 8). In complementary frequentist random-effects NMAs conducted with the netmeta package in R, we estimated ORs with 95% CIs and 95% PIs for each intervention versus NDI for all 4 outcomes (Multimedia Appendix 7). Across outcomes, PIs were wider than the corresponding CIs and frequently crossed the null, indicating substantial uncertainty in the effects that might be observed in future implementation settings despite the direction of the average effects.

Strength of Evidence

All of these enrolled studies were RCTs, and the quality of evidence was evaluated by the Cochrane Handbook and graded each potential source of bias as low, high, or some concerns; the details were displayed in Figure 1. We assessed confidence in the results of the NMA using the CINeMA framework. Of the 4 outcomes analyzed, only “consistent condom use” met the criteria for CINeMA assessment. The remaining outcomes were excluded due to insufficient numbers of studies or disconnected network structures. For consistent condom use, certainty of evidence was mainly downgraded for within-study bias and imprecision, resulting in overall ratings of “low” to “very low” confidence. Among the 10 comparisons, 1 (10%) was rated as “very low” and 9 (90%) as “low” certainty. For condom use at last sex, the proportion of condom-protected acts, and STI incidence, the certainty of evidence is less well characterized, but given the sparse data, risk of bias, and wide intervals, these estimates should similarly be interpreted as low certainty. A detailed summary of risk-of-bias judgements, CINeMA assessments, and the corresponding GRADE “Summary of Findings” information for each primary outcome is provided in Figure 1 and Table 4, with full GRADE “Summary of Findings” tables formatted according to the GRADE Working Group template available in Multimedia Appendix 3 to aid interpretation of the magnitude and certainty of the main comparisons.

XSL•FO
RenderX

Table . Summary of confidence in the evidence for consistent condom use, assessed using Confidence in Network Meta-Analysis and GRADE.

Comparison	Number of studies	Within-study bias	Reporting bias	Indirectness	Imprecision	Heterogeneity	Incoherence	Overall confidence rating (GRADE)	Reasons for downgrading
IOI:NDI[a, b]	3	Major concerns	Low risk	No concerns	Major concerns	No concerns	No concerns	Very low	Within-study bias and imprecision
IOI:SWI[c]	1	No concerns	Low risk	No concerns	Major concerns	No concerns	No concerns	Low	Imprecision only
MAI:NDI[d]	2	Some concerns	Low risk	No concerns	Major concerns	No concerns	No concerns	Low	Within-study bias and imprecision
NDI:SWI	2	Some concerns	Low risk	No concerns	Major concerns	No concerns	No concerns	Low	Within-study bias and imprecision
NDI:TCI[e]	3	Some concerns	Low risk	No concerns	Major concerns	No concerns	No concerns	Low	Within-study bias and imprecision

[a]IOI: interactive online-based intervention.

[b]NDI: nondigital intervention.

[c]SWI: static web-based intervention.

[d]MAI: mobile app-based intervention.

[e]TCI: telecommunication-based intervention.

Discussion

Principal Findings

To the best of our knowledge, this is the first NMA to systematically evaluate the comparative effectiveness of different modalities of DHIs on promoting safer sexual behaviors among youth. By simultaneously examining 4 distinct digital modalities and 3 behavioral outcomes, and by incorporating STI incidence (including HIV) as a biological endpoint, this review expands the current evidence base and clarifies which intervention types are better suited for immediate versus sustained behavior change, and highlights the gap between improvements in self-reported safer behaviors and reductions in biological HIV/STI infection. Drawing upon data from 24 RCTs across diverse contexts, our findings offer comprehensive insights to inform future development, optimization, and implementation of DHIs in HIV/STIs prevention, underscore the need for designing multimodal, context-aware digital interventions that integrate behavioral support with access to testing and care services, and provide practical considerations for policymakers, program designers, and digital platform developers who seek to tailor DHIs to youth populations.

Across outcomes, between-study heterogeneity and statistical inconsistency were generally low to moderate, and the network satisfied the assumptions of transitivity and global consistency. However, most comparisons were informed by a small number of trials, many of which had some concerns or a high risk of bias in at least one RoB 2 domain. The complementary frequentist NMAs showed that 95% PIs were typically wide and often crossed the null, even when average effects appeared beneficial. Consistent condom use was the only outcome that met CINeMA requirements, and all network comparisons for this outcome were rated as having low or very low certainty. Together, these features suggest that our estimates reflect uncertain average effects rather than precise predictions for specific programs or settings. These patterns of risk of bias, particularly deviations from intended interventions and missing outcome data, may have led to overestimation of some intervention effects or increased uncertainty in the network estimates and contributed to downgrading the certainty of evidence in our GRADE/CINeMA assessment.

In assessing condom use at the last sexual encounter, TCI emerged as the only approach showing statistically significant improvement compared with NDI. This finding aligns with prior studies reporting absolute increases in condom use among participants receiving SMS or phone-based reminders [80]. The relatively stronger performance of TCI may be attributable to its simplicity and immediacy, directly prompting protective behaviors without requiring advanced digital literacy or prolonged engagement [81]. Previous evidence has also highlighted TCI as one of the more acceptable and widely used forms of digital intervention among young people [30]. Thus, TCIs may offer an immediate behavioral benefit, especially for outcomes tied to the most recent sexual event. Interestingly, while MAI ranked highest in SUCRA probability, its effect did not reach significance, suggesting inconsistency between ranking and statistical evidence. This discrepancy could stem from limited trial numbers, variability in app engagement, or short

XSL•FO
RenderX

intervention duration, which may have constrained the power to detect statistical changes. However, this apparent benefit of TCI is based on a few trials with some concerns or high risk of bias, and the wide PIs suggest that similar effects may not be consistently achievable in all implementation settings.

In this study, consistent condom use is improved more effectively by SWI and IOI than by TCI, with SWI ranking the highest. This finding aligns with prior evidence indicating that web-based and online interventions contributed to a substantial proportion of effective digital interventions [21,30]. First, that SWI outperformed IOI may seem counterintuitive, given the absence of interactive features. However, SWI could deliver standardized, theory-based content in a less distracting, user-driven format, allowing youth to process key prevention messages at their own pace. Moreover, while the IOIs included in this review are mostly delivered via computers or tablets, SWIs—though static—are often accessible on smartphones or distributed through popular platforms such as Facebook or WeChat with text, images, or videos [82]. This accessibility and portability may explain why SWI demonstrates stronger effects on consistent condom use. Second, TCI was less effective than both SWI and IOI for improving consistent condom use. Previous studies have suggested that TCI, especially SMS-based reminders, remain controversial in terms of their long-term effectiveness for HIV prevention behaviors [83]. Therefore, although TCI can effectively prompt immediate behaviors, its brief and repetitive messages may lack the depth and reinforcement required to sustain consistent condom use over time. Nevertheless, most trials contributing to this outcome had some concerns or high risk of bias, and CINeMA rated the certainty of these network estimates as low to very low, so the apparent superiority of SWI and IOI should be considered tentative.

When examining the proportion of condom use, IOI demonstrated a significant advantage over SWI, indicating the added value of interactive engagement. This aligns with prior evidence showing that increases in condom use were significantly associated with the use of tailored strategies [21], feedback provision, and guided navigation in digital interventions. Unlike static websites, IOIs integrated tailored feedback, quizzes, or real-time support, which represent core behavior change techniques such as personalized feedback, problem-solving, and self-regulation strategies [84]. These core behavioral change techniques are particularly effective in influencing situational decisions and negotiations during sexual encounters, thereby enhancing the overall proportion of condom use [85,86]. In other words, consistent condom use reflects the internalization of long-term protective norms, and it could be reinforced by standardized and less distracting formats like SWI, while the proportion of condom use is more sensitive to moment-to-moment decision-making. This divergence in findings—SWI being more effective for consistent use, while IOI excels in overall proportion—highlights that different behavioral outcomes may respond to distinct mechanisms of action. Besides, the evolution of technology-based intervention modes has gradually expanded from web-based formats to SMS and social media [87-89]. This trend further supports the notion that while static formats may effectively reinforce long-term norms, interactive platforms provide additional advantages for immediate behavioral decisions during sexual encounters. Yet the CrIs and PIs for this outcome were wide and frequently included the null, indicating considerable heterogeneity and imprecision and implying that any average benefit in the proportion of condom-protected acts may not translate into clear improvements in every context.

Unlike the behavioral outcomes, the effectiveness of DHIs on reducing STI infection reveals a different pattern: NDI and TCI performed more favorably, while IOI and SWI ranked lowest. This finding contrasts with prior studies showing that digital interventions can improve HIV/STI care engagement, such as testing uptake or service use [90-92], highlighting that improvements in care engagement do not necessarily translate into reductions in biological outcomes like STI incidence. Several factors may explain this inconsistency. First, STI incidence represents a distal biological endpoint that may require longer follow-up to capture meaningful reductions, and improvements in self-reported behaviors, such as condom use, may not directly translate into biological protection due to reporting bias or inconsistent application in high-risk contexts. Second, reductions in STI incidence depend not only on safer behaviors but also on timely testing, treatment, and linkage to care. However, evidence shows that stigma related to gender identity, socioeconomic status, race, and ethnicity can delay care-seeking and discourage individuals from accessing necessary services [93]. Nondigital approaches, such as community outreach or peer education programs, often combine behavioral education with direct access to services such as STI testing, treatment linkage, and ongoing support from trained staff or peers, which can directly impact biological outcomes. In contrast, many DHIs focus primarily on education and motivation, without providing structured access to testing or clinical care. Additionally, NDIs may facilitate stronger trust and engagement through in-person interactions, which can overcome barriers related to stigma, confidentiality concerns, or digital literacy limitations. Therefore, while DHIs can effectively change self-reported behaviors, the integrated, multi-component structure of NDIs may explain their relative advantage in reducing actual STI incidence (including HIV). Besides, in this study, TCI is the only digital intervention showing a relatively favorable effect on reducing STI incidence, second only to NDI. Prior studies have demonstrated that TCI can facilitate access to HIV prevention services for youth and achieve high patient and provider satisfaction [93,94]. By providing a comfortable, judgment-free platform, telemedicine may be particularly preferred by marginalized populations, especially transgender youth [95,96]. Given the small and heterogeneous evidence base, important risks of bias in several trials, and wide PIs, these findings on STI incidence should be regarded as exploratory and hypothesis-generating rather than definitive.

Our use of PIs further illustrates the extent to which the observed benefits of DHIs may vary across settings. For most comparisons, the 95% PIs were wide and often crossed the null value, even where the corresponding credible or CIs suggested modest advantages over NDI. This pattern indicates that, although certain DHI modalities tend to improve condom use

XSL•FO
RenderX

on average, implementation in new populations or health systems may yield smaller effects or no clear benefit, underscoring the need for careful adaptation, monitoring, and evaluation when scaling up digital prevention programs.

Implications

These findings have several implications. First, the differential effectiveness of DHIs suggests tailoring intervention types to targeted outcomes. TCI showed immediate benefits for condom use at last sexual contact and a relative advantage for STI incidence, indicating that simple, low-burden interventions may prompt rapid behavior change and reach marginalized youth. In contrast, SWI and IOI were more effective for consistent and habitual condom use, highlighting the value of self-paced content and behavior change techniques that enhance motivation and self-regulation. Multi-modal approaches combining these strengths may maximize overall effectiveness. Second, DHIs alone may be insufficient to reduce STI incidence. Biological outcomes require longer follow-up, and self-reported behavior change does not always translate to infection reduction. Integrating DHIs with offline services, such as condom distribution, PrEP promotion, routine testing, and clinical linkage, is likely necessary to achieve meaningful improvements. Third, these results have implications for intervention design and digital health policy. When targeting youth populations, accessibility, acceptability, and engagement should be prioritized. For example, interventions delivered via mobile platforms or telecommunication may overcome barriers related to stigma or limited digital literacy, while interactive online content can leverage BCTs to support skill acquisition and habitual behavior change. Intervention planners should also consider the balance between immediacy and sustainability of effects: brief, repeated prompts may drive immediate behavior, whereas structured, self-paced content may reinforce long-term habits.

Limitations

Several limitations of this study should be noted. First, substantial clinical and methodological heterogeneity across trials—including differences in intervention intensity and duration, digital platforms, follow-up periods, and outcome definitions—may have contributed to between-study heterogeneity and could challenge the transitivity assumption underpinning some indirect comparisons in the NMA. Second, we restricted inclusion to randomized, prevention-focused DHIs among HIV-negative or status-unknown youth and excluded nonrandomized studies and trials targeting HIV-positive adolescents; the findings may therefore not generalize to these subgroups or to broader digital programs. Third, all behavioral outcomes relied on self-report, and secondary outcomes such as self-efficacy and number of sexual partners were too sparsely and inconsistently reported to be synthesized, limiting our ability to evaluate the broader psychosocial impact of DHIs beyond condom use. Fourth, the number of trials assessing certain interventions, particularly MAI, was limited, which may reduce statistical power and the precision of effect estimates. Fifth, we were unable to formally assess small-study or publication bias, and selective nonpublication or outcome reporting cannot be ruled out. Finally, most network comparisons were rated as low or very low certainty because of within-study bias and imprecision, and PIs, estimated using standard random-effects methods in netmeta, were wide; the true comparative effects may therefore differ meaningfully from our estimates, underscoring the need for rigorous, adequately powered RCTs with standardized outcomes and longer follow-up.

Conclusions

In conclusion, this NMA provides the most comprehensive synthesis to date on the comparative effectiveness of DHIs in promoting safer sexual behaviors among youth. This NMA highlights that the effectiveness of DHIs for HIV prevention among youth depends on both intervention modality and targeted outcomes. While DHIs can enhance knowledge and protective behaviors, their impact on biological endpoints remains limited without integration with offline services and broader structural support. Tailoring interventions to behavioral targets, engagement strategies, and contextual factors is essential to maximize their potential in promoting youth sexual health. We hope that these results will inform the design of youth-centered digital prevention programs, guide clinicians and educators in selecting appropriate modalities, and support policymakers and guideline developers in integrating digital strategies into national HIV prevention frameworks. Future studies should focus on the specific characteristics of patients to provide personalized estimates of comparative effectiveness and individualized predictions regarding the probability of response to treatment and of side effects.

Acknowledgments

YZ and WP contributed equally to this work and are co-first authors. The authors thank all the reviewers for their assistance and support. A generative artificial intelligence tool (ChatGPT, OpenAI) was used solely for language refinement. All artificial intelligence–assisted edits were reviewed, verified, and approved by the authors, who take full responsibility for the final manuscript.

Funding

This research is funded by the China Medical Board (grant number 22- ‐ 465), Hainan Provincial Department of Science and Technology (grant number ZDYF2024SHFZ042), and the National Nature Science Foundation of China (grant number 72574244). The funding sources had no involvement in the study design, data collection, analysis, interpretation of results, or manuscript writing.

XSL•FO
RenderX

Data Availability

The datasets analyzed during this study were derived from previously published studies and are available in the respective articles. All data extracted and analyzed for this systematic review and Bayesian network meta-analysis are included in this published article and its multimedia appendices.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search String.
[DOCX File, 27 KB - jmir_v28i1e87071_app1.docx]

Multimedia Appendix 2
Reference list of excluded studies.
[DOCX File, 50 KB - jmir_v28i1e87071_app2.docx]

Multimedia Appendix 3
Risk of bias, CINeMA assessments, and GRADE Summary of Findings tables.
[DOCX File, 210 KB - jmir_v28i1e87071_app3.docx]

Multimedia Appendix 4
Reference list of included studies.
[DOCX File, 22 KB - jmir_v28i1e87071_app4.docx]

Multimedia Appendix 5
Indirect comparative effectiveness DHIs.
[DOCX File, 19 KB - jmir_v28i1e87071_app5.docx]

Multimedia Appendix 6
Forest plots of posterior ORs with 95.
[DOCX File, 109 KB - jmir_v28i1e87071_app6.docx]

Multimedia Appendix 7
Random‐effects network meta‐analysis of digital health interventions versus non‐digital interventions.
[DOCX File, 19 KB - jmir_v28i1e87071_app7.docx]

Multimedia Appendix 8
Model fit diagnostics for consistency and inconsistency models in the network meta.
[DOCX File, 1102 KB - jmir_v28i1e87071_app8.docx]

Checklist 1
PRISMA-NMA checklist.
[DOCX File, 166 KB - jmir_v28i1e87071_app9.docx]

References

1. Youth. United Nations. URL: https://www.un.org/en/global-issues/youth [accessed 2025-11-01]
2. Rotheram-Borus MJ, Davis E, Rezai R. Stopping the rise of HIV among adolescents globally. Curr Opin Pediatr 2018 Feb;30(1):131-136. [doi: 10.1097/MOP.0000000000000580] [Medline: 29315110]
3. Zhu Q, Wang Y, Liu J, et al. Identifying major drivers of incident HIV infection using recent infection testing algorithms (RITAs) to precisely inform targeted prevention. Int J Infect Dis 2020 Dec;101:131-137. [doi: 10.1016/j.ijid.2020.09.1421] [Medline: 32987184]
4. Barrett M, Laris BA, Anderson P, et al. Condom use and error experience among young adolescents: implications for classroom instruction. Health Promot Pract 2021 May;22(3):313-317. [doi: 10.1177/1524839920935431] [Medline: 32536213]

XSL•FO
RenderX

5. Zhang X, Tang W, Li Y, et al. The HIV/AIDS epidemic among young people in China between 2005 and 2012: results of a spatial temporal analysis. HIV Med 2017 Mar;18(3):141-150. [doi: 10.1111/hiv.12408] [Medline: 27552922]
6. Digitizing sexual health information in Kenya and Peru. World Health Organization. URL: https://www.who.int/news-room/feature-stories/detail/digitizing-sexual-health-information-in-kenya-and-peru [accessed 2025-11-01]
7. Brasileiro J, Queiroz A, Hightow-Weidman LB, Muessig KE. Implementation strategies for Digital HIV prevention and care interventions for youth: a scoping review. Curr HIV/AIDS Rep 2025 Mar 13;22(1):23. [doi: 10.1007/s11904-025-00732-5] [Medline: 40080278]
8. Guiding optimal development and use of digital health towards improved health outcomes. World Health Organization. URL: https://www.who.int/westernpacific/activities/guiding-optimal-development-and-use-of-digital-health-towards-improved-health-outcomes [accessed 2025-11-01]
9. Global Strategy on Digital Health 2020-2025: World Health Organization; 2021. URL: https://www.who.int/publications/i/item/9789240020924 [accessed 2025-03-12]
10. Jennings Mayo-Wilson L, Glass NE, Labrique A, et al. Feasibility of assessing economic and sexual risk behaviors using text message surveys in African-American young adults experiencing homelessness and unemployment: single-group study. JMIR Form Res 2020 Jul 17;4(7):e14833. [doi: 10.2196/14833] [Medline: 32706656]
11. Ippoliti NB, L'Engle K. Meet us on the phone: mobile phone programs for adolescent sexual and reproductive health in low-to-middle income countries. Reprod Health 2017 Jan 17;14(1):11. [doi: 10.1186/s12978-016-0276-z] [Medline: 28095855]
12. Castro Á, Barrada JR, Ramos-Villagrasa PJ, Fernández-Del-Río E. Profiling dating apps users: sociodemographic and personality characteristics. Int J Environ Res Public Health 2020 May 22;17(10):3653. [doi: 10.3390/ijerph17103653] [Medline: 32455986]
13. Watchirs Smith L, Guy R, Degenhardt L, et al. Meeting sexual partners through internet sites and smartphone apps in Australia: national representative study. J Med Internet Res 2018 Dec 18;20(12):e10683. [doi: 10.2196/10683] [Medline: 30563809]
14. Smith LW, Liu B, Degenhardt L, et al. Is sexual content in new media linked to sexual risk behaviour in young people? A systematic review and meta-analysis. Sex Health 2016 Nov 28;13(6):501-515. [doi: 10.1071/SH16037]
15. Cornelius JB, Whitaker-Brown C, Neely T, Kennedy A, Okoro F. Mobile phone, social media usage, and perceptions of delivering a social media safer sex intervention for adolescents: results from two countries. Adolesc Health Med Ther 2019;10:29-37. [doi: 10.2147/AHMT.S185041] [Medline: 31118855]
16. Burns JC, Chakraborty S, Saint Arnault D. Social media preference and condom use behaviors: an analysis of digital spaces with young African American males. Health Educ Behav 2021 Apr;48(2):190-198. [doi: 10.1177/1090198121993043] [Medline: 33703958]
17. Bailey J, Mann S, Wayal S, Abraham C, Murray E. Digital media interventions for sexual health promotion-opportunities and challenges: a great way to reach people, particularly those at increased risk of sexual ill health. BMJ 2015 Mar 3;350:h1099. [doi: 10.1136/bmj.h1099] [Medline: 25736806]
18. Berendes S, Gubijev A, French R, Hickson FCI, Free C. A qualitative study exploring experiences of the safetxt digital health intervention to reduce sexually transmitted infections in young people in the UK. BMJ Open 2023 Oct 24;13(10):e072701. [doi: 10.1136/bmjopen-2023-072701] [Medline: 37879678]
19. Eleuteri S, Toso M. How the smartphone apps can improve your sexual wellbeing. Int J Impot Res 2024 Nov;36(7):781-785. [doi: 10.1038/s41443-023-00730-4] [Medline: 37414872]
20. Bailey JV, Wayal S, Aicken CRH, et al. Interactive digital interventions for prevention of sexually transmitted HIV. AIDS 2021 Mar 15;35(4):643-653. [doi: 10.1097/QAD.0000000000002780] [Medline: 33259345]
21. Xin M, Viswanath K, Li AYC, et al. The effectiveness of electronic health interventions for promoting HIV-preventive behaviors among men who have sex with men: meta-analysis based on an integrative framework of design and implementation features. J Med Internet Res 2020 May 25;22(5):e15977. [doi: 10.2196/15977] [Medline: 32449685]
22. Wadham E, Green C, Debattista J, Somerset S, Sav A. New digital media interventions for sexual health promotion among young people: a systematic review. Sex Health 2019 Apr;16(2):101-123. [doi: 10.1071/SH18127] [Medline: 30819326]
23. Javidi H, Widman L, Lipsey N, Brasileiro J, Javidi F, Jhala A. Redeveloping a digital sexual health intervention for adolescents to allow for broader dissemination: implications for HIV and STD prevention. AIDS Educ Prev 2021 Apr;33(2):89-102. [doi: 10.1521/aeap.2021.33.2.89] [Medline: 33821678]
24. Carvalho T, Alvarez MJ, Pereira C, Schwarzer R. Stage-based computer-delivered interventions to increase condom use in young men. Int J Sex Health 2016 Apr 2;28(2):176-186. [doi: 10.1080/19317611.2016.1158764]
25. Jones K, Baldwin KA, Lewis PR. The potential influence of a social media intervention on risky sexual behavior and chlamydia incidence. J Community Health Nurs 2012;29(2):106-120. [doi: 10.1080/07370016.2012.670579] [Medline: 22536914]
26. Lim MSC, Hocking JS, Aitken CK, et al. Impact of text and email messaging on the sexual health of young people: a randomised controlled trial. J Epidemiol Community Health 2012 Jan;66(1):69-74. [doi: 10.1136/jech.2009.100396] [Medline: 21415232]

XSL•FO
RenderX

27. Sun WH, Wong CKH, Wong WCW, Peer-Led A. A peer-led, social media-delivered, safer sex intervention for Chinese college students: randomized controlled trial. J Med Internet Res 2017 Aug 9;19(8):e284. [doi: 10.2196/jmir.7403] [Medline: 28793980]
28. Tang W, Wei C, Cao B, et al. Crowdsourcing to expand HIV testing among men who have sex with men in China: a closed cohort stepped wedge cluster randomized controlled trial. PLoS Med 2018 Aug;15(8):e1002645. [doi: 10.1371/journal.pmed.1002645] [Medline: 30153265]
29. Evans WD, Ulasevich A, Hatheway M, Deperthes B. Systematic review of peer-reviewed literature on global condom promotion programs. Int J Environ Res Public Health 2020 Mar 27;17(7):2262. [doi: 10.3390/ijerph17072262] [Medline: 32230929]
30. Sewak A, Yousef M, Deshpande S, Seydel T, Hashemi N. The effectiveness of digital sexual health interventions for young adults: a systematic literature review (2010-2020). HEALTH Promot Int 2023 Feb 1;38(1):daac104. [doi: 10.1093/heapro/daac104] [Medline: 36757346]
31. Guse K, Levine D, Martins S, et al. Interventions using new digital media to improve adolescent sexual health: a systematic review. J Adolesc Health 2012 Dec;51(6):535-543. [doi: 10.1016/j.jadohealth.2012.03.014] [Medline: 23174462]
32. Nwaozuru U, Obiezu-Umeh C, Shato T, et al. Mobile health interventions for HIV/STI prevention among youth in low- and middle-income countries (LMICs): a systematic review of studies reporting implementation outcomes. Implement Sci Commun 2021 Nov 6;2(1):126. [doi: 10.1186/s43058-021-00230-w] [Medline: 34742357]
33. Saragih ID, Imanuel Tonapa S, Porta CM, Lee BO. Effects of telehealth interventions for adolescent sexual health: a systematic review and meta-analysis of randomized controlled studies. J Telemed Telecare 2024 Feb;30(2):201-214. [doi: 10.1177/1357633X211047762]
34. Teadt S, Burns JC, Montgomery TM, Darbes L. African American adolescents and young adults, new media, and sexual health: scoping review. JMIR Mhealth Uhealth 2020 Oct 5;8(10):e19459. [doi: 10.2196/19459] [Medline: 33016890]
35. Goldstein M, Archary M, Adong J, et al. Systematic review of mHealth interventions for adolescent and young adult HIV prevention and the adolescent HIV continuum of care in low to middle income countries. AIDS Behav 2023 May;27(Suppl 1):94-115. [doi: 10.1007/s10461-022-03840-0] [Medline: 36322217]
36. Widman L, Nesi J, Kamke K, Choukas-Bradley S, Stewart JL. Technology-based interventions to reduce sexually transmitted infections and unintended pregnancy among youth. J Adolesc Health 2018 Jun;62(6):651-660. [doi: 10.1016/j.jadohealth.2018.02.007] [Medline: 29784112]
37. Buti J, Glenny AM, Worthington HV, Nieri M, Baccini M. Network meta-analysis of randomised controlled trials: direct and indirect treatment comparisons. Eur J Oral Implantol 2011;4(1):55-62. [Medline: 21594220]
38. Hutton B, Salanti G, Caldwell DM, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Ann Intern Med 2015 Jun 2;162(11):777-784. [doi: 10.7326/M14-2385] [Medline: 26030634]
39. Younge SN, Salazar LF, Crosby RF, DiClemente RJ, Wingood GM, Rose E. Condom use at last sex as a proxy for other measures of condom use: is it good enough? Adolescence 2008;43(172):927-931. [Medline: 19149154]
40. Pallonen UE, Williams ML, Timpson SC, Bowen A, Ross MW. Personal and partner measures in stages of consistent condom use among African-American heterosexual crack cocaine smokers. AIDS Care 2008 Feb;20(2):205-213. [doi: 10.1080/09540120701513669] [Medline: 18293131]
41. Slaymaker E, Zaba B. Measurement of condom use as a risk factor for HIV infection. Reprod Health Matters 2003 Nov;11(22):174-184. [doi: 10.1016/s0968-8080(03)02299-7] [Medline: 14708408]
42. Brafford LJ, Beck KH. Development and validation of a condom self-efficacy scale for college students. J Am Coll Health 1991 Mar;39(5):219-225. [doi: 10.1080/07448481.1991.9936238] [Medline: 1783705]
43. Lawrance L, Levy SR, Rubinson L. Self-efficacy and AIDS prevention for pregnant teens. J Sch Health 1990 Jan;60(1):19-24. [doi: 10.1111/j.1746-1561.1990.tb04771.x] [Medline: 2299814]
44. Goldkuhle M, Bender R, Akl EA, et al. GRADE guidelines: 29. rating the certainty in time-to-event outcomes-study limitations due to censoring of participants with missing data in intervention studies. J Clin Epidemiol 2021 Jan;129:126-137. [doi: 10.1016/j.jclinepi.2020.09.017] [Medline: 33007458]
45. Renson A, Hudgens MG, Keil AP, Zivich PN, Aiello AE. Identifying and estimating effects of sustained interventions under parallel trends assumptions. Biometrics 2023 Dec;79(4):2998-3009. [doi: 10.1111/biom.13862] [Medline: 36989497]
46. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
47. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898. [doi: 10.1136/bmj.l4898]
48. Cochrane Handbook for Systematic Reviews of Interventions: Cochrane; 2024. URL: https://www.cochrane.org/handbook [accessed 2025-02-05]
49. Mustanski B, Parsons JT, Sullivan PS, Madkins K, Rosenberg E, Swann G. Biomedical and behavioral outcomes of keep it up!: an eHealth HIV prevention program RCT. Am J Prev Med 2018 Aug;55(2):151-158. [doi: 10.1016/j.amepre.2018.04.026] [Medline: 29937115]

XSL•FO
RenderX

50. Tan RKJ, Koh WL, Le D, et al. Effect of a popular web drama video series on HIV and other sexually transmitted infection testing among gay, bisexual, and other men who have sex with men in Singapore: community-based, pragmatic, randomized controlled trial. J Med Internet Res 2022 May 6;24(5):e31401. [doi: 10.2196/31401] [Medline: 35522470]
51. Lau JTF, Lee AL, Tse WS, et al. A randomized control trial for evaluating efficacies of two online cognitive interventions with and without fear-appeal imagery approaches in preventing unprotected anal sex among Chinese men who have sex with men. AIDS Behav 2016 Sep;20(9):1851-1862. [doi: 10.1007/s10461-015-1263-z]
52. Chernick LS, Santelli J, Stockwell MS, et al. A multi-media digital intervention to improve the sexual and reproductive health of female adolescent emergency department patients. Acad Emerg Med 2022 Mar;29(3):308-316. [doi: 10.1111/acem.14411] [Medline: 34738284]
53. Peipert JF, Redding CA, Blume JD, et al. Tailored intervention to increase dual-contraceptive method use: a randomized trial to reduce unintended pregnancies and sexually transmitted infections. Am J Obstet Gynecol 2008 Jun;198(6):630. [doi: 10.1016/j.ajog.2008.01.038] [Medline: 18395692]
54. Free C, Palmer MJ, McCarthy OL, et al. Effectiveness of a behavioural intervention delivered by text messages (safetxt) on sexually transmitted reinfections in people aged 16-24 years: randomised controlled trial. BMJ 2022 Sep 28;378(378):e070351. [doi: 10.1136/bmj-2022-070351] [Medline: 36170988]
55. Whiteley LB, Brown LK, Curtis V, Ryoo HJ, Beausoleil N. Publicly available internet content as a HIV/STI prevention intervention for urban youth. J Prim Prev 2018 Aug;39(4):361-370. [doi: 10.1007/s10935-018-0514-y] [Medline: 30003459]
56. Wilson E, Free C, Morris TP, et al. Internet-accessed sexually transmitted infection (e-STI) testing and results service: a randomised, single-blind, controlled trial. PLoS Med 2017 Dec;14(12):e1002479. [doi: 10.1371/journal.pmed.1002479] [Medline: 29281628]
57. Hu Z, Fu Y, Wang X, et al. Effects of sexuality education on sexual knowledge, sexual attitudes, and sexual behaviors of youths in China: a cluster-randomized controlled trial. J Adolesc Health 2023 Apr;72(4):607-615. [doi: 10.1016/j.jadohealth.2022.11.006] [Medline: 36604206]
58. Nuwamanya E, Nalwanga R, Nuwasiima A, et al. Effectiveness of a mobile phone application to increase access to sexual and reproductive health information, goods, and services among university students in Uganda: a randomized controlled trial. Contracept Reprod Med 2020 Oct 31;5(1):31. [doi: 10.1186/s40834-020-00134-5] [Medline: 33292724]
59. Bannink R, Broeren S, Joosten-van Zwanenburg E, van As E, van de Looij-Jansen P, Raat H. Effectiveness of a Web-based tailored intervention (E-health4Uth) and consultation to promote adolescents' health: randomized controlled trial. J Med Internet Res ;16(5). [doi: 10.2196/jmir.3163]
60. Rotheram-Borus MJ, Swendeman D, Comulada WS, Weiss RE, Lee M, Lightfoot M. Prevention for substance-using HIV-positive young people. J Acquir Immune Defic Syndr 2004;37(Supplement 2):S68-S77. [doi: 10.1097/01.qai.0000140604.57478.67]
61. Ballester-Arnal R, Gil-Llario MD, Giménez-García C, Kalichman SC. What works well in HIV prevention among Spanish young people? An analysis of differential effectiveness among six intervention techniques. AIDS Behav 2015 Jul;19(7):1157-1169. [doi: 10.1007/s10461-014-0863-3] [Medline: 25085080]
62. Santa Maria D, Padhye N, Businelle M, et al. Efficacy of a just-in-time adaptive intervention to promote HIV risk reduction behaviors among young adults experiencing homelessness: pilot randomized controlled trial. J Med Internet Res 2021 Jul 6;23(7):e26704. [doi: 10.2196/26704] [Medline: 34255679]
63. Bull S, Devine S, Schmiege SJ, Pickard L, Campbell J, Shlay JC. Text messaging, teen outreach program, and sexual health behavior: a cluster randomized trial. Am J Public Health 2016 Sep;106(S1):S117-S124. [doi: 10.2105/AJPH.2016.303363]
64. Ybarra ML, Bull SS, Prescott TL, Korchmaros JD, Bangsberg DR, Kiwanuka JP. Adolescent abstinence and unprotected sex in CyberSenga, an internet-based HIV prevention program: randomized clinical trial of efficacy. PLoS ONE 2013;8(8):e70083. [doi: 10.1371/journal.pone.0070083] [Medline: 23967069]
65. Bauermeister JA, Tingler RC, Demers M, et al. Acceptability and preliminary efficacy of an online HIV prevention intervention for single young men who have sex with men seeking partners online: the myDEx project. AIDS Behav 2019 Nov;23(11):3064-3077. [doi: 10.1007/s10461-019-02426-7] [Medline: 30762190]
66. Rinehart DJ, Leslie S, Durfee MJ, et al. Acceptability and efficacy of a sexual health texting intervention designed to support adolescent females. Acad Pediatr 2020;20(4):475-484. [doi: 10.1016/j.acap.2019.09.004] [Medline: 31560971]
67. Miller MK, Catley D, Adams A, et al. Brief motivational intervention to improve adolescent sexual health service uptake: a pilot randomized controlled trial in the emergency department. J Pediatr 2021 Oct;237:250-257. [doi: 10.1016/j.jpeds.2021.06.007] [Medline: 34144031]
68. Cordova D, Munoz-Velazquez J, Mendoza Lua F, et al. Pilot study of a multilevel mobile health app for substance use, sexual risk behaviors, and testing for sexually transmitted infections and HIV among youth: randomized controlled trial. JMIR Mhealth Uhealth 2020 Mar 17;8(3):e16251. [doi: 10.2196/16251] [Medline: 32181747]
69. Shafii T, Benson SK, Morrison DM, Hughes JP, Golden MR, Holmes KK. Results from e-KISS: electronic-KIOSK intervention for safer sex: a pilot randomized controlled trial of an interactive computer-based intervention for sexual health in adolescents and young adults. PLOS ONE 2019;14(1):e0209064. [doi: 10.1371/journal.pone.0209064] [Medline: 30673710]

XSL•FO
RenderX

70. Yarger J, Gutmann-Gonzalez A, Borgen N, Romero J, Decker MJ. In the know: a cluster randomized trial of an in-person sexual health education program integrating digital technologies for adolescents. J Adolesc Health 2024 May;74(5):1019-1025. [doi: 10.1016/j.jadohealth.2023.12.012] [Medline: 38323966]
71. Suffoletto B, Akers A, McGinnis KA, Calabria J, Wiesenfeld HC, Clark DB. A sex risk reduction text-message program for young adult females discharged from the emergency department. J Adolesc Health 2013 Sep;53(3):387-393. [doi: 10.1016/j.jadohealth.2013.04.006] [Medline: 23707402]
72. Stoffers JM, Völlm BA, Rücker G, Timmer A, Huband N, Lieb K. Psychological therapies for people with borderline personality disorder. Cochrane Database Syst Rev 2012 Aug 15;2012(8):CD005652. [doi: 10.1002/14651858.CD005652.pub2] [Medline: 22895952]
73. Zhao T, Tang C, Yan H, Lu Q, Guo M, Wang H. Comparative efficacy and acceptability of non-pharmacological interventions for depression in people living with HIV: a systematic review and network meta-analysis. Int J Nurs Stud 2023 Apr;140:104452. [doi: 10.1016/j.ijnurstu.2023.104452] [Medline: 36821952]
74. Nikolakopoulou A, Higgins JPT, Papakonstantinou T, et al. CINeMA: an approach for assessing confidence in the results of a network meta-analysis. PLoS Med 2020 Apr;17(4):e1003082. [doi: 10.1371/journal.pmed.1003082] [Medline: 32243458]
75. Puhan MA, Schünemann HJ, Murad MH, et al. A GRADE working group approach for rating the quality of treatment effect estimates from network meta-analysis. BMJ 2014 Sep 24;349:g5630. [doi: 10.1136/bmj.g5630] [Medline: 25252733]
76. Balduzzi S, Ruecker G, Nikolakopoulou OS. Guido: netmeta: an R package for network meta-analysis using frequentist methods. J Stat Softw 2023;106(2):1-40. [doi: 10.18637/jss.v106.i02]
77. Noma H, Hamura Y, Sugasawa S, Furukawa TA. Improved methods to construct prediction intervals for network meta-analysis. Res Synth Methods 2023 Nov;14(6):794-806. [doi: 10.1002/jrsm.1651] [Medline: 37399809]
78. Chaimani A, Higgins JPT, Mavridis D, Spyridonos P, Salanti G. Graphical tools for network meta-analysis in STATA. PLoS ONE 2013;8(10):e76654. [doi: 10.1371/journal.pone.0076654] [Medline: 24098547]
79. Shafii T, Benson SK, Morrison DM, Hughes JP, Golden MR, Holmes KK. Results from e-KISS: electronic-KIOSK ts and young adults. PLoS ONE 2019;14(1):e0209064. [doi: 10.1371/journal.pone.0209064] [Medline: 30673710]
80. Huang W, Stegmueller D, Ong JJ, et al. Technology-based HIV prevention interventions for men who have sex with men: systematic review and meta-analysis. J Med Internet Res 2025 Apr 28;27:e63111. [doi: 10.2196/63111] [Medline: 40293786]
81. Alkhaldi G, Modrow K, Hamilton F, Pal K, Ross J, Murray E. Promoting engagement with a digital health intervention (HeLP-diabetes) using email and text message prompts: mixed-methods study. Interact J Med Res 2017 Aug 22;6(2):e14. [doi: 10.2196/ijmr.6952] [Medline: 28829328]
82. Plantin JC, de Seta G. WeChat as infrastructure: the techno-nationalist shaping of Chinese digital platforms. Chinese Journal of Communication 2019 Jul 3;12(3):257-273. [doi: 10.1080/17544750.2019.1572633]
83. Fischer AE, Hanif H, Stocks JB, et al. Mobile health intervention tools promoting HIV pre-exposure prophylaxis among adolescent girls and young women in Sub-Saharan Africa: scoping review. JMIR Mhealth Uhealth 2025 Jun 20;13:e60819. [doi: 10.2196/60819] [Medline: 40540732]
84. Mo PH, Xie L, Lee TC, Li AYC. Use of behavior change techniques in digital HIV prevention programs for adolescents and young people: systematic review. JMIR Public Health Surveill 2025 Apr 28;11:e59519. [doi: 10.2196/59519] [Medline: 40293783]
85. Kwasnicka D, Dombrowski SU, White M, Sniehotta F. Theoretical explanations for maintenance of behaviour change: a systematic review of behaviour theories. Health Psychol Rev 2016 Sep;10(3):277-296. [doi: 10.1080/17437199.2016.1151372] [Medline: 26854092]
86. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. J Med Internet Res 2017 Jun 29;19(6):e232. [doi: 10.2196/jmir.7126] [Medline: 28663162]
87. Cao B, Gupta S, Wang J, et al. Social media interventions to promote HIV testing, linkage, adherence, and retention: systematic review and meta-analysis. J Med Internet Res 2017 Nov 24;19(11):e394. [doi: 10.2196/jmir.7997] [Medline: 29175811]
88. Nguyen LH, Tran BX, Rocha LEC, et al. A systematic review of ehealth interventions addressing HIV/STI prevention among men who have sex with men. AIDS Behav 2019 Sep;23(9):2253-2272. [doi: 10.1007/s10461-019-02626-1] [Medline: 31401741]
89. Muessig KE, Nekkanti M, Bauermeister J, Bull S, Hightow-Weidman LB. A systematic review of recent smartphone, internet and web 2.0 interventions to address the HIV continuum of care. Curr HIV/AIDS Rep 2015 Mar;12(1):173-190. [doi: 10.1007/s11904-014-0239-3] [Medline: 25626718]
90. McNeill IM, Borland R, Abraham C. Digital health interventions providing behavioral assessment and goal prioritization support: scoping review. J Med Internet Res 2025 Aug 28;27:e68112. [doi: 10.2196/68112] [Medline: 40875981]
91. Wantland DJ, Portillo CJ, Holzemer WL, Slaughter R, McGhee EM. The effectiveness of web-based vs. non-web-based interventions: a meta-analysis of behavioral change outcomes. J Med Internet Res 2004 Nov 10;6(4):e40. [doi: 10.2196/jmir.6.4.e40] [Medline: 15631964]
92. Noar SM, Black HG, Pierce LB. Efficacy of computer technology-based HIV prevention interventions: a meta-analysis. AIDS 2009 Jan 2;23(1):107-115. [doi: 10.1097/QAD.0b013e32831c5500] [Medline: 19050392]

XSL•FO
RenderX

93. Radix AE, Bond K, Carneiro PB, Restar A. Transgender individuals and digital health. Curr HIV/AIDS Rep 2022 Dec;19(6):592-599. [doi: 10.1007/s11904-022-00629-7] [Medline: 36136217]
94. Waad A. Caring for our community: telehealth interventions as a promising practice for addressing population health disparities of LGBTQ+ communities in health care settings. Delaware Journal of Public Health 2019;5(3):12-15. [doi: 10.32481/djph.2019.06.005]
95. Homkham N, Manojai N, Patpeerapong P, et al. A comparative study of transgender women accessing HIV testing via face-to-face and telemedicine services in Chiang Mai, Thailand during the COVID-19 pandemic and their risk of being HIV-positive. BMC Public Health 2023 Nov 4;23(1):2161. [doi: 10.1186/s12889-023-17124-2] [Medline: 37925430]
96. Atkinson E, Galinkala P, Campos-Castillo C. Telehealth use in 2022 among US adults by sexual orientation. Am J Manag Care 2024 Jan 1;30(1):e19-e25 [FREE Full text] [doi: 10.37765/ajmc.2024.89490] [Medline: 38271570]

Abbreviations

CINeMA: Confidence in Network Meta-Analysis
CrI: credible interval
DHI: digital health intervention
DIC: deviance information criterion
IOI: interactive online-based intervention
MAI: mobile app-based intervention
NDI: nondigital intervention
NMA: network meta-analysis
OR: odds ratio
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyzes
PRISMA-NMA: PRISMA extension for network meta-analyzes
RCT: randomized controlled trial
RoB 2: Cochrane Risk of Bias 2 tool
SR: systematic review
STI: sexually transmitted infection
SUCRA: surface under the cumulative ranking curve
SWI: static web-based intervention
TCI: telecommunication-based intervention

Edited by S Brini; submitted 04.Nov.2025; peer-reviewed by C Aladeokin, JR de-Moya-Romero; accepted 09.Dec.2025; published 04.Feb.2026.

Please cite as:
Zhu Y, Peng W, Hu D, Choi EPH, Välimäki MA, Zhang C, Li X
Effects of Digital Health Interventions to Promote Safer Sex Behaviors Among Youth: Systematic Review and Bayesian Network Meta-Analysis
J Med Internet Res 2026;28:e87071
URL: https://www.jmir.org/2026/1/e87071
doi:10.2196/87071

© Yiran Zhu, Wenwen Peng, Die Hu, Edmond Pui Hang Choi, Maritta Anneli Välimäki, Ci Zhang, Xianhong Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Characterization of Models for Identifying Physical and Cognitive Frailty in Older Adults With Diabetes: Systematic Review and Meta-Analysis

Xia Wang[1*], MS; Shujie Meng[1*], MS; Xiang Xiao[2], BS; Liu Lu[3]; Hongyan Chen[1], MS; Yong Li[4], MEd; Rong Zhang[4]; Qiwu Jiang[5], BM; Shan Liu[6], MS; Ru Gao[6], PhD

[1]School of Basic Medical Sciences and School of Nursing, Chengdu University, No. 2025, Chengluo Avenue, Chengdu, China

[2]School of Nursing, Yibin Vocational College of Medicine and Health, Yibin, China

[3]Nursing Department, The Fourth People's Hospital of Yibin, Yibin, China

[4]Rehabilitation College, Sichuan Health Rehabilitation Vocational College, Zigong, China

[5]School of Rehabilitation and Wellness, Yibin Vocational College of Medicine and Health, Yibin, China

[6]Nursing Department, Wenjiang District People's Hospital, No.86 Kangtai Road Wenjiang District, Chengdu, Sichuan Province, China

[*]these authors contributed equally

Corresponding Author:
Ru Gao, PhD
Nursing Department, Wenjiang District People's Hospital, No.86 Kangtai Road Wenjiang District, Chengdu, Sichuan Province, China

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2026/1/e92717

Abstract

Background: Physical frailty and cognitive frailty are increasingly recognized as critical geriatric syndromes among older adults with diabetes, contributing to adverse outcomes such as disability, hospitalization, and mortality. Early identification of individuals at high risk is therefore essential for timely prevention and intervention. Although a growing number of prediction models have been developed for this population, evidence regarding their methodological rigor, predictive performance, and generalizability remains fragmented.

Objective: This study aims to evaluate and characterize existing models for detecting or predicting physical frailty and cognitive frailty in older adults with diabetes.

Methods: PubMed, Embase, Web of Science, China National Knowledge Infrastructure (CNKI), Wanfang, and VIP databases were searched from their inception to December 2025. Retrospective, cross-sectional, and prospective studies that developed or validated models predicting frailty or cognitive frailty in older adults with diabetes were included. The Prediction Model Study Risk Of Bias Assessment Tool (PROBAST) was used to assess risk of bias and applicability. Random effects meta-analyses using the Hartung-Knapp-Sidik-Jonkman method were conducted to synthesize model performance, including the pooled area under the receiver operating characteristic curve (AUC). Heterogeneity was explored through subgroup and sensitivity analyses. Small study effects were evaluated using funnel plots, the Egger test, and the Deeks funnel plot asymmetry test.

Results: A total of 24 studies comprising 32 diagnostic models were included. The overall pooled analysis demonstrated an AUC of 0.851 (95% CI 0.820 - 0.882) with a 95% prediction interval of 0.710 - 0.992, sensitivity of 0.810 (95% CI 0.740 - 0.850), and specificity of 0.850 (95% CI 0.810 - 0.890). Statistical comparisons in the modeling approach revealed that logistic regression models achieved a significantly higher pooled AUC (0.850) compared with machine learning models (0.785; *P*=.003). Similarly, retrospective studies demonstrated superior performance, with an AUC of 0.900 compared with 0.843 for cross-sectional studies (*P*=.03). Conversely, no significant differences were observed across subgroups stratified by data source (*P*=.42), patient characteristics (*P*=.77), validation methods (*P*=.16), or specific outcomes (*P*=.94). The most common predictors identified were depression, age, and regular exercise; however, all included studies were assessed as having a high risk of bias.

Conclusions: To our knowledge, this review provides the first comprehensive synthesis of models for risk stratification of physical frailty and cognitive frailty in older adults with diabetes. The findings indicate that existing models demonstrate satisfactory discrimination; specifically, CIs confirmed a robust average effect, while prediction intervals suggested that performance in future settings, though variable, is likely to remain acceptable. However, clinical utility is currently constrained by high risk of bias and

XSL•FO
RenderX

limited external validation. Future research must prioritize rigorous, prospective, multicenter studies adhering to standard reporting guidelines (eg, TRIPOD [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis]) to establish valid, generalizable, and clinically actionable prognostic instruments.

Trial Registration: PROSPERO CRD420251019308; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251019308

(J Med Internet Res 2026;28:e84617) doi:10.2196/84617

KEYWORDS

prediction model; diabetes; frailty; systematic review; meta-analysis

Introduction

Background

Diabetes mellitus has evolved into one of the most critical global public health challenges of the 21st century. According to the International Diabetes Federation (IDF), approximately 589 million adults were living with diabetes globally in 2024, with projections indicating this number could rise to 853 million by 2050 [1]. In the absence of optimal management, patients with diabetes are predisposed to micro- and macrovascular complications that significantly shorten life expectancy [2,3]. Recent data from the Global Burden of Disease study indicate that the prevalence of diabetes increases with age, reaching 24.4% among individuals aged ≥75 years [4]. Older adults are especially vulnerable to diabetes-related complications due to greater medical complexity and a higher likelihood of frailty compared with younger populations [5].

Frailty is regarded as a consequence of the decline in function and reserve of multiple organs with age, particularly involving the neuromuscular, endocrine, and immune systems [6]. Notably, frailty is particularly prevalent among patients with diabetes, with reported prevalence rates ranging from 10.4% to 20.8% across different studies [7-10]. Furthermore, previous studies indicated that individuals with diabetes have an approximately 1.6-fold higher risk of developing frailty than those without diabetes [11]. However, frailty frequently co-occurs with cognitive impairment [12]; their simultaneous presence is termed cognitive frailty, a distinct clinical entity that represents a crucial subtype of frailty requiring specific attention.

Once established, frailty typically follows a progressive trajectory, increasing the likelihood of adverse clinical outcomes such as falls, incontinence, rapid functional decline, pressure ulcers, and delirium [13-17]. In addition to these risks, frailty is linked to higher rates of hospitalization, emergency department visits, prolonged inpatient stays, and mortality [18,19]. Of particular concern is that the coexistence of physical and cognitive impairment further amplifies these risks, leading to greater adverse outcomes [20,21].

Evidence suggests a bidirectional relationship between diabetes and frailty, often creating a cycle where each condition exacerbates the other [22]. The presence of physical or cognitive frailty introduces significant complexity to diabetes management [23]. In frail patients, physiological deterioration and multi-organ dysfunction fundamentally alter the pharmacokinetics of antihyperglycemic agents [24]. Specifically, sarcopenia, increased adiposity, and compromised renal or hepatic clearance heighten the susceptibility to adverse drug events, such as hypoglycemia and unintended weight loss. Additionally, the decreased caloric intake typical of this population further aggravates the risk of hypoglycemia and hinders recovery from hypoglycemic events [25,26].

In recent years, physical frailty and cognitive frailty have been increasingly conceptualized as dynamic and potentially preventable or reversible conditions, especially when identified at an early stage [27,28]. In patients with diabetes, nonpharmacological interventions—including structured physical activity, nutritional optimization, and multimodal strategies—have demonstrated potential benefits for mitigating frailty progression. Consequently, early identification of individuals at high risk has become a cornerstone of effective prevention and management strategies. To this end, diagnostic and prognostic models designed to detect physical or cognitive frailty integrate multiple demographic, clinical, and psychosocial factors to estimate an individual's risk profile. These models serve to support health care professionals with stratifying risk, facilitating timely and targeted interventions, and optimizing the allocation of health care resources.

However, the clinical application of these models may be hindered due to insufficient evidence regarding their performance, risk of bias, and applicability in routine practice. Although individual studies exist, no systematic review has yet comprehensively evaluated these models for both physical frailty and cognitive frailty in older adults with diabetes. Therefore, it is essential to conduct a systematic review that thoroughly assesses the methodological quality and clinical applicability of existing models.

Objectives

The aim of this systematic review and meta-analysis was to evaluate the methodological quality and clinical utility of existing models designed for the identification or prediction of physical frailty and cognitive frailty in older adults with diabetes. The specific aims included the following: (1) to determine the characteristics and most frequent predictors of risk prediction models developed for physical frailty and cognitive frailty in this population; (2) to analyze the methodological limitations and risk of bias of these models using the Prediction Model Study Risk Of Bias Assessment Tool (PROBAST); and (3) to investigate the pooled predictive performance of these tools to assess their potential for real-world clinical implementation.

XSL•FO
RenderX

Methods

Search Strategy and Selection Criteria

This systematic review and meta-analysis was registered on PROSPERO (CRD420251019308). The study followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) expanded checklist [29] and the PRISMA extension for diagnostic test accuracy (PRISMA-DTA) [30], while the literature search was conducted and reported in accordance with PRISMA-S (PRISMA Search) [31]. A comprehensive literature search was conducted across the PubMed, EMBASE, Web of Science, China National Knowledge Infrastructure (CNKI), Wanfang, and VIP databases, covering records from database inception to December 2025. We developed the search strategy based on the PITROS (Participants, Index Test, Target Conditions, Reference Standard, Outcomes, Settings) framework (Table S1 in Multimedia Appendix 1). The strategy combined Medical Subject Headings (MeSH) with free-text terms (Table S2 in Multimedia Appendix 1). The search strategy was independently evaluated by another librarian in accordance with the PRESS (Peer Review of Electronic Search Strategies) guidelines. In addition, references of relevant studies, guidelines, and reviews were manually searched, and citation tracking was performed using the Web of Science database to identify other relevant studies. No study registries were searched.

In clinical settings, prediction encompasses both diagnostic models (estimating the probability of a particular condition being present) and prognostic models (forecasting the likelihood of future outcomes) [32,33]. This review included all primary studies describing the development and/or validation of prediction models, tools, or scores for estimating the risk of physical frailty or cognitive frailty in older adults with diabetes. The inclusion criteria were (1) participant age ≥60 years and presence of diabetes, including diabetes only and diabetes with other comorbidities or complications; (2) research content involving the construction of a predictive model for identifying physical frailty or cognitive frailty in individuals with diabetes; (3) retrospective studies, cross-sectional studies, and prospective studies; and (4) published in English or Chinese. The exclusion criteria were (1) duplicate publications; (2) reviews, case reports, or conference abstracts; (3) literature that could not be obtained from the original text; (4) literature that could not provide valid data; and (5) studies in which the model contained only a single predictor.

Data Extraction

This study used the reference management software EndNote X9 to identify and remove duplicate records. We then eliminated literature unrelated to the research topic by screening titles and abstracts. Finally, the full texts were reviewed to identify studies satisfying both the inclusion and exclusion criteria. Upon completion of the literature screening process, a data extraction form was devised in accordance with the Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies (CHARMS) [34]. The contents included (1) basic information of the study including author, year of publication, study location, study design, and source of data used; (2) patient characteristics including sample size and patient diagnosis; (3) number of predictors, predictor type, most important predictors, and predictor screening methods; (4) model characteristics including outcome indicators, modeling methods, model verification methods, and missing data processing methods; (5) model presentation; and (6) model performance as measured using the area under the curve (AUC), sensitivity, and specificity. In studies where multiple models were developed and the best-performing model was explicitly reported, we included the best model in the analysis. For studies that reported multiple models without specifying a preferred one, we selected the model with the highest AUC to represent the study. We prioritized internally validated estimates, resorting to development performance or external validation data only when internal estimates were unavailable. For studies with unclear or incomplete data, attempts were made to contact the corresponding authors. To ensure the consistency and accuracy of the final data, two researchers (XW and SM) independently extracted the data, and the extracted results were compared and checked. Inconsistencies were resolved through discussion and consultation, and a third researcher (RG) was asked to assist in judgment when necessary.

Quality Assessment

Two reviewers independently used PROBAST [35] to appraise the risk of bias and the applicability of each included study. Disagreements were resolved by discussion or by consulting a third reviewer. PROBAST evaluates 4 domains: participants, predictors, outcome, and analysis. Each domain is rated as high, unclear, or low risk of bias. The applicability evaluation focuses on the 3 areas of research (subjects, predictors, and results), and its evaluation process is similar to the risk of bias assessment.

Statistical Analysis

Meta-analyses were conducted using Stata 18 (specifically the midas and metan commands) and R version 4.3.2 (the metafor package). Using AUC values derived from models, we calculated the pooled AUC and produced an AUC forest plot. An AUC below 0.7 signified inadequate discrimination, an AUC ranging from 0.7 to 0.8 denoted moderate discrimination, and an AUC exceeding 0.8 suggested excellent discrimination [36]. Additionally, models that reported sample sizes and sensitivity and specificity were extracted. The true positive (TP), false positive (FP), false negative (FN), and true negative (TN) for each model were calculated using the formulas sensitivity=TP/(TP+ FN) and specificity=TN/(FP+ TN). Pooled sensitivity and pooled specificity were computed based on TP, FP, FN, and TN, and corresponding forest plots of sensitivity and specificity were constructed. Subsequently, a summary receiver operating characteristic curve was generated. The degree of heterogeneity across the models under consideration was evaluated using the *Q* test and measured using the I^2 statistic (where an I^2 value <25% signifies low heterogeneity, between 25% and 50% indicates moderate heterogeneity, and >50% denotes high heterogeneity) [37]. To account for between-study heterogeneity and provide more robust variance estimation, we used the Hartung-Knapp-Sidik-Jonkman method for a random effects meta-analysis on the logit scale [38]. For studies with significant heterogeneity, subgroup analyses, sensitivity

XSL•FO
RenderX

analyses, or only descriptive analyses were performed. The presence of small study effects was evaluated using the Egger test, funnel plots, and the Deeks funnel plot [39]. A *P* value <.05 was deemed to indicate statistical significance.

Results

Study Selection

The initial search identified a total of 4873 records. After removing duplicates, 3124 records remained. Following title and abstract screening, 88 studies were selected for full-text review. We could not retrieve 4 studies. During the full-text assessment, 18 studies were excluded because the study population did not have diabetes mellitus. Additionally, 16 studies were excluded because the outcome variable was not physical frailty or cognitive frailty; 16 were excluded because they were nonoriginal studies, such as reviews and meta-analyses; and 10 were excluded due to an inappropriate study design. Ultimately, 24 studies [40-63] were included in this meta-analysis. The literature screening process and results are illustrated in Figure 1.

Figure 1. Literature screening flow chart. CNKI: China National Knowledge Infrastructure.

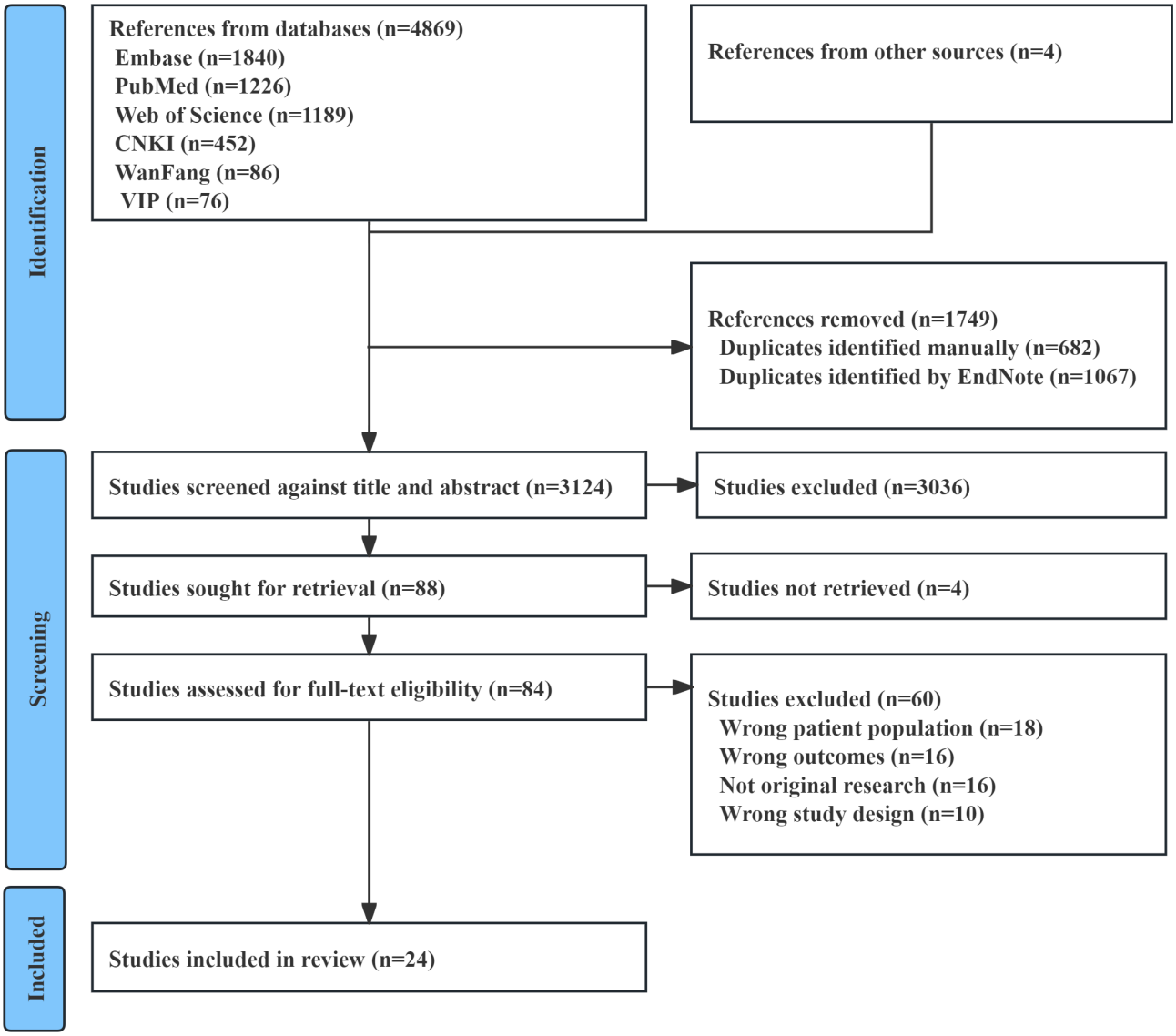

Characteristics of Included Studies

The specific characteristics of the included studies are detailed in Tables 1 and 2. A total of 24 studies reported 32 diagnostic models for physical frailty and cognitive frailty in older adults with diabetes. The publication years of these papers spanned from 2023 to 2025. All the studies were conducted in China. The number of participants in the included studies ranged from 152 to 1436. The prevalence of frailty varied from 10.1% to 51.2%, while that of cognitive frailty ranged from 20.3% to 62.1%. Regarding study design, 21 studies [41-46,48-60,62,63] used a cross-sectional design, while 3 were retrospective studies [40,47,61]. In addition, 18 studies [41-48,50,51,53-59,63] were conducted at single centers, whereas 6 studies [40,49,52,60-62] were multicenter studies. In terms of the target population, 7 studies [40-42,57,58,60,61] included all types of diabetes, 14 studies [43,45,47-51,53-56,59,62,63] focused specifically on patients with type 2 diabetes, and 3 studies [44,46,52] included patients with diabetes and other comorbidities or complications. Physical frailty was the primary outcome in 13 studies [40-42,44,45,51,52,55,57,58,60-62], while the remaining 11 studies [43,46-50,53,54,56,59,63] investigated cognitive frailty.

XSL•FO
RenderX

For frailty screening, 22 studies [40-51,53-57,59-63] used the Fatigue, Resistance, Ambulation, Illnesses, and Loss of Weight (FRAIL) scale, while 2 studies [52,58] used the Tilburg Frailty Indicator. To evaluate cognitive function in the 11 cognitive frailty studies, the Montreal Cognitive Assessment was the predominant tool used in 10 studies [43,46-50,53,54,59,63], whereas the Mini-Mental State Examination was used in only 1 study [56]. Regarding data handling, missing data were not reported in 9 studies [42,43,45-47,49,51,56,59], 3 studies [40,57,62] used imputation methods, and 12 studies [41,44,48,50,52-55,58,60,61,63] excluded participants with missing data. Continuous variables were maintained as continuous in 14 studies [40,42-44,47,51,52,55-59,61,62] and transformed into categorical variables in 10 studies [41,45,46,48-50,53,54,60,63]. The majority of models were presented as nomograms: Specifically, 13 studies [40,43,45,48,50,51,54-57,59,61,63] presented results solely as nomograms, 4 studies [41,42,44,46] provided both nomograms and full equations, 1 study [52] presented a nomogram and risk chart, and 1 study [58] presented a nomogram and decision tree. Additionally, 3 studies [53,60,62] developed risk sum scores, and 2 studies [47,49] provided logistic regression (LR) equations.

XSL•FO
RenderX

Table . Overview of the basic characteristics of included studies (n=24) identifying physical and cognitive frailty in older adults with diabetes.

Author	Year	Outcome	Definition of outcome	Sample size, n	Event rate, n (%)	Modeling algorithms used	Population	Internal validation	Predictors, n	Model presentation
Wu et al [62]	2025	Frailty	FRAIL[a]	509	148 (29.1)	LR[b], SVM[c], GBM[d], RF[e], CatBoost	T2DM[f]	Cross-validation	7	Sum score
Xiao et al [61]	2025	Frailty	FRAIL	1107	113 (10.2)	LR	Diabetes	Random split	7	Nomogram
Wang et al [55]	2025	Frailty	FRAIL	152	47 (31. 1)	LR	T2DM	Random split	4	Nomogram
Du et al [45]	2024	Frailty	FRAIL	458	83 (18.1)	LR	T2DM	Bootstrap	8	Nomogram
Tang et al [51]	2024	Frailty	FRAIL	566	213 (37.6)	LR	T2DM	Random split	6	Nomogram
Wang et al [52]	2024	Frailty	TFI[g]	491	216 (44.0)	LR, NN[h]	Diabetes with diabetic foot	Random split	8	Nomogram and risk chart
Dang [42]	2024	Frailty	FRAIL	360	115 (32.0)	LR	Diabetes	Random split (+ external)	5	Nomogram and full equation
Xi [57]	2024	Frailty	FRAIL	338	130 (38.5)	LR	Diabetes	Random split	6	Nomogram
Cheng [41]	2024	Frailty	FRAIL	317	118 (37.2)	LR	Diabetes	Bootstrap	5	Nomogram and full equation
Yin [58]	2024	Frailty	TFI	379	194 (51.2)	LR, DT[i]	Diabetes	Bootstrap	8	Nomogram and decision tree
Zheng [60]	2024	Frailty	FRAIL	380	112 (29.5)	RF, SVM, KNN[j]	Diabetes	Cross-validation	7	Sum score
Bu et al [40]	2023	Frailty	FRAIL	1436	145 (10.1)	LR	Diabetes	Random split	7	Nomogram
Dong et al [44]	2023	Frailty	FRAIL	485	211 (43.5)	LR	Diabetes with diabetic retinopathy	Bootstrap	7	Nomogram and full equation
Ma et al [49]	2025	Cognitive frailty	FRAIL, MoCA[k]	253	76 (30.0)	LR	T2DM	None	5	Full equation
Wang et al [53]	2025	Cognitive frailty	FRAIL, MoCA	202	80 (39.6)	DT	T2DM	Random split	11	Sum score
Liang et al [46]	2024	Cognitive frailty	FRAIL, MoCA	265	93 (35.1)	LR	Diabetes with COPD[l]	Random split (+ external)	7	Nomogram and full equation
Yu and Yu [63]	2024	Cognitive frailty	FRAIL, MoCA	430	132 (30.7)	LR	T2DM	Bootstrap (+ external)	7	Nomogram
Zhang et al [59]	2024	Cognitive frailty	FRAIL, MoCA	215	66 (30.7)	LR	T2DM	Bootstrap	5	Nomogram
Liu [47]	2024	Cognitive frailty	FRAIL, MoCA	220	137 (62.1)	LR	T2DM	Random split	4	Full equation
Deng et al [43]	2023	Cognitive frailty	FRAIL, MoCA	315	87 (27.6)	LR	T2DM	Random split	6	Nomogram

XSL•FO
RenderX

Author	Year	Outcome	Definition of outcome	Sample size, n	Event rate, n (%)	Modeling algorithms used	Population	Internal validation	Predictors, n	Model presentation
Wang and Xu [54]	2023	Cognitive frailty	FRAIL, MoCA	262	85 (32.4)	LR	T2DM	Bootstrap	8	Nomogram
Wang et al [56]	2023	Cognitive frailty	FRAIL, MMSE[m]	321	85 (26.5)	LR	T2DM	Bootstrap	5	Nomogram
Liu [48]	2023	Cognitive frailty	FRAIL, MoCA	483	98 (20.3)	LR	T2DM	Random split	6	Nomogram
Meng [50]	2022	Cognitive frailty	FRAIL, MoCA	508	117 (23.0)	LR	T2DM	Bootstrap (+ external)	6	Nomogram

[a]FRAIL: Fatigue, Resistance, Ambulation, Illness, and Loss of Weight scale.

[b]LR: logistic regression.

[c]SVM: support vector machine.

[d]GBM: gradient boosting machine.

[e]RF: random forest.

[f]T2DM: type 2 diabetes mellitus.

[g]TFI: Tilburg Frailty Indicator

[h]NN: neural network.

[i]DT: decision tree.

[j]KNN: k-nearest neighbors.

[k]MoCA: Montreal Cognitive Assessment.

[l]COPD: chronic obstructive pulmonary disease.

[m]MMSE: Mini-Mental State Examination.

XSL•FO
RenderX

Table . Methodological and clinical characteristics of included studies (n=24) identifying physical frailty and cognitive frailty in older adults with diabetes.

Characteristic	Studies, n (%)
Study design	
Retrospective studies	3 (13)
Cross-sectional study	21 (88)
Source of data used	
Single center	18 (75)
Multicenter	6 (25)
Missing data handling	
Not reported	9 (38)
Exclusion	12 (50)
Imputation	3 (13)
Handling of continuous data	
Continuous	14 (58)
Categorical or dichotomous	10 (42)
Feature selection	
Univariate analysis	6 (25)
Multivariate analysis	7 (29)
Univariate analysis and multivariate analysis	11 (46)
Calibration method	
Hosmer-Lemeshow test	2 (8)
Calibration plot	4 (17)
Hosmer-Lemeshow test and calibration plot	14 (58)
None	4 (17)
Validation method	
Internal validation	19 (79)
External validation and internal validation	4 (17)
None	1 (4)

Characteristics of Included Prediction Models

Regarding modeling methods, among the 32 models included, LR was the most commonly used algorithm. LR analyses were used in 22 models, while 10 models used machine learning (ML) techniques, including random forest (n=2), support vector machines (n=2), decision trees (n=2), k-nearest neighbors (n=1), CatBoost (n=1), gradient boosting machine (n=1), and neural networks (n=1). Model discrimination was reported for all models, with AUC values ranging from 0.703 to 0.983 (Table S3 in Multimedia Appendix 1). Specificity and sensitivity were reported in 17 studies [40-42,44,46-50,52,53,56-58,60-62] involving 25 models. Specifically, sensitivity ranged from 0.102 to 0.955, and specificity varied from 0.505 to 0.990. However, model calibration was not reported in 4 studies [52,53,60,62]. *P* values from both the Hosmer-Lemeshow test and calibration plots were used in 14 studies [40-44,48,50,51,54-59], 2 studies [47,49] used *P* values only from the Hosmer-Lemeshow test, and 4 studies [45,46,61,63] used only calibration plots. Regarding model validation, 1 study [49] developed models without validation, and 19 studies [40,41,43-45,47,48,51-62] conducted only internal validation without external validation.

Features

All features covered a wide range of factors, including sociodemographic characteristics, lifestyle factors, health-related factors, mental health status, laboratory test indicators, and anthropometric measurements. A total of 33 features were involved in the studies. The number of features incorporated into each study varied from 4 to 11. Among the features, the 5 most frequently occurring were depression, age, regular exercise, social activity, and duration of diabetes. The frequency distribution of all features is illustrated in Figure 2.

XSL•FO
RenderX

Figure 2. Frequency of predictors in the included studies. GFR: glomerular filtration rate; HbA_{1c}: glycated hemoglobin; L/A: ratio of serum leptin to adiponectin.

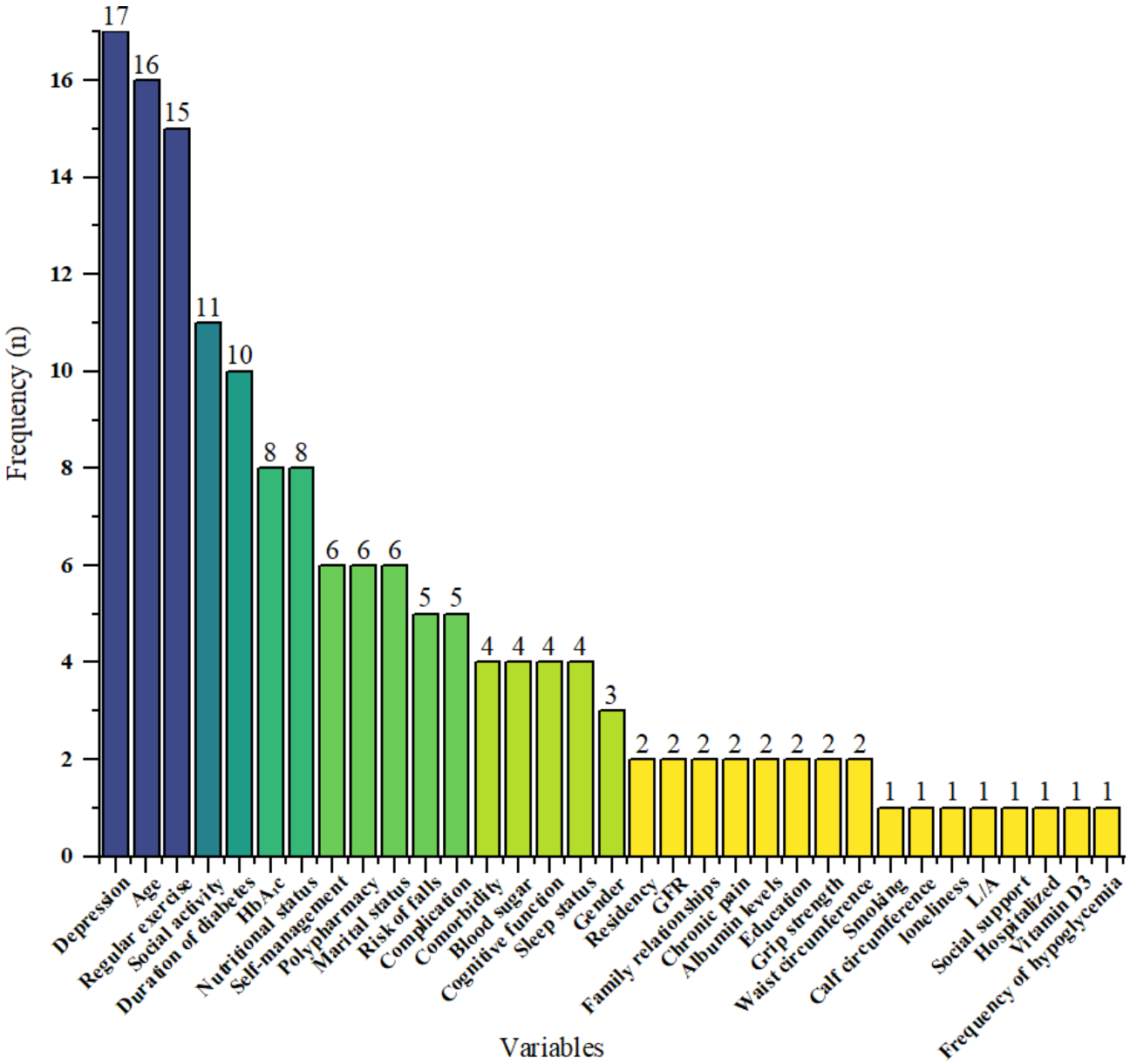

Quality Assessment

Overview

The PROBAST tool was used to assess the risk of bias and the applicability of the included prediction model studies (Figure 3 and Table S4 in Multimedia Appendix 1). According to the established criteria, all 24 studies, which encompassed 32 models, were identified as having a high risk of bias. In terms of applicability, 13 studies [43,44,46-50,52-54,56,59,63], including 14 models, were deemed to have high concerns regarding applicability. Conversely, the remaining 11 studies [40-42,45,51,55,57,58,60-62], which included 18 models, were considered to have low concerns regarding applicability. Notably, 4 studies [52,58,60,62] included multiple models each; however, there was no difference in the quality assessment results between the models within these studies.

XSL•FO
RenderX

Figure 3. Prediction Model Study Risk Of Bias Assessment Tool (PROBAST) risk of bias (ROB) and applicability assessment for all included studies.

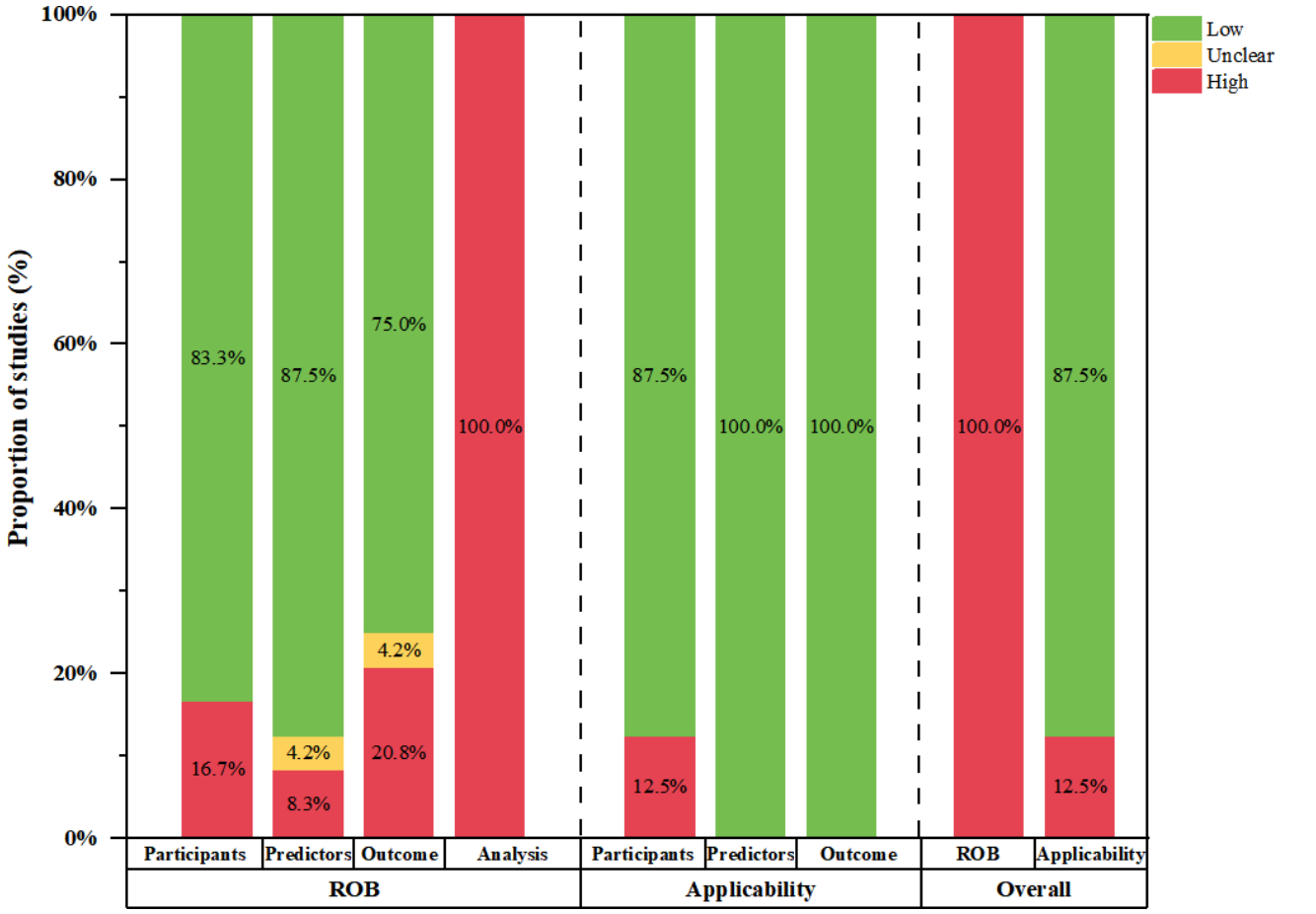

Risk of Bias Assessment

Within the participant domain, 4 studies [40,43,47,61] were recognized as exhibiting a high risk of bias. Of these, 3 studies [40,47,61] were deemed as having a high risk owing to their study designs, while the remaining study [43] was classified as such due to the exclusion of specific subgroups that could potentially alter the performance of the prediction model. In the predictor domain, 2 studies [40,63] were assessed as having a significant risk of bias due to the use of outcome information in the evaluation of predictors, 1 study [55] was rated as having an unclear risk of bias because the researchers did not report whether they used the same assessment measures when evaluating the predictors. In the outcome domain, 3 studies [40,62,63] had a significant risk of bias because the definition of outcomes included ≥1 predictor, 2 studies [52,53] were rated as having a high risk of bias due to the potentially inappropriate time interval between predictor assessment and outcome determination, and 1 study [49] was deemed to be at unclear risk of bias as they did not report information on the method of outcome classification. In the analysis domain, all studies were judged to have a high risk of bias. Current guidance recommends that studies developing predictive models achieve at least 20 events per variable (EPV). However, 13 studies [43,45,46,48,49,53-56,59-61,63] did not meet this requirement. Moreover, 10 studies [41,45,46,48-50,53,54,60,63] transformed continuous variables into categorical variables, either in part or entirely, and the authors did not report whether standard definitions were used for the categorization; 1 study [40] partially excluded participants for unreasonable reasons. Regarding the handling of missing data, 12 studies [41,42,44,48,50,52-54,58,60,61,63] directly excluded cases with missing data, while 9 studies [43,45-47,49,51,55,56,59] did not explicitly report whether data were missing. In addition, 6 studies [44,49,50,52,54,59] did not avoid selecting variables based solely on univariate analysis; 3 studies [52,53,60] did not comprehensively assess the predictive performance of their models, using only discrimination measures without calibration; 6 studies [41,44,48,50,55,57] neglected to evaluate the risk of overfitting, underfitting, or optimism that could bias the apparent performance of their predictive models; 1 study [49] developed models without validation; and 19 studies [40,41,43-45,47,48,51-62] conducted only internal validation without external validation.

Applicability Risk Assessment

In the participant domain, 3 studies [44,46,52] had a high risk of applicability concerns due to the inclusion of individuals with other comorbidities or complications.

Meta-Analysis

A random effects meta-analysis using the Hartung-Knapp-Sidik-Jonkman method was performed to evaluate the predictive performance at both the study and model levels. Regarding the analysis of the 24 included studies, the overall pooled AUC was 0.851 (95% CI 0.820 - 0.882), with a 95% prediction interval (PI) of 0.710 to 0.992 (P<.001; I^2=92.0%; Figure 4). When analyzing the 32 models, the overall

XSL•FO
RenderX

pooled AUC was 0.829 (95% CI 0.802 - 0.856), with a 95% PI of 0.686 to 0.972 (P<.001; I^2=92.5%; Figure S1 in Multimedia Appendix 1).

Figure 4. Forest plot of the random effects meta-analysis of pooled area under the curve (AUC) estimates for 29 validation models [40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62]. HKSJ: Hartung-Knapp-Sidik-Jonkman.

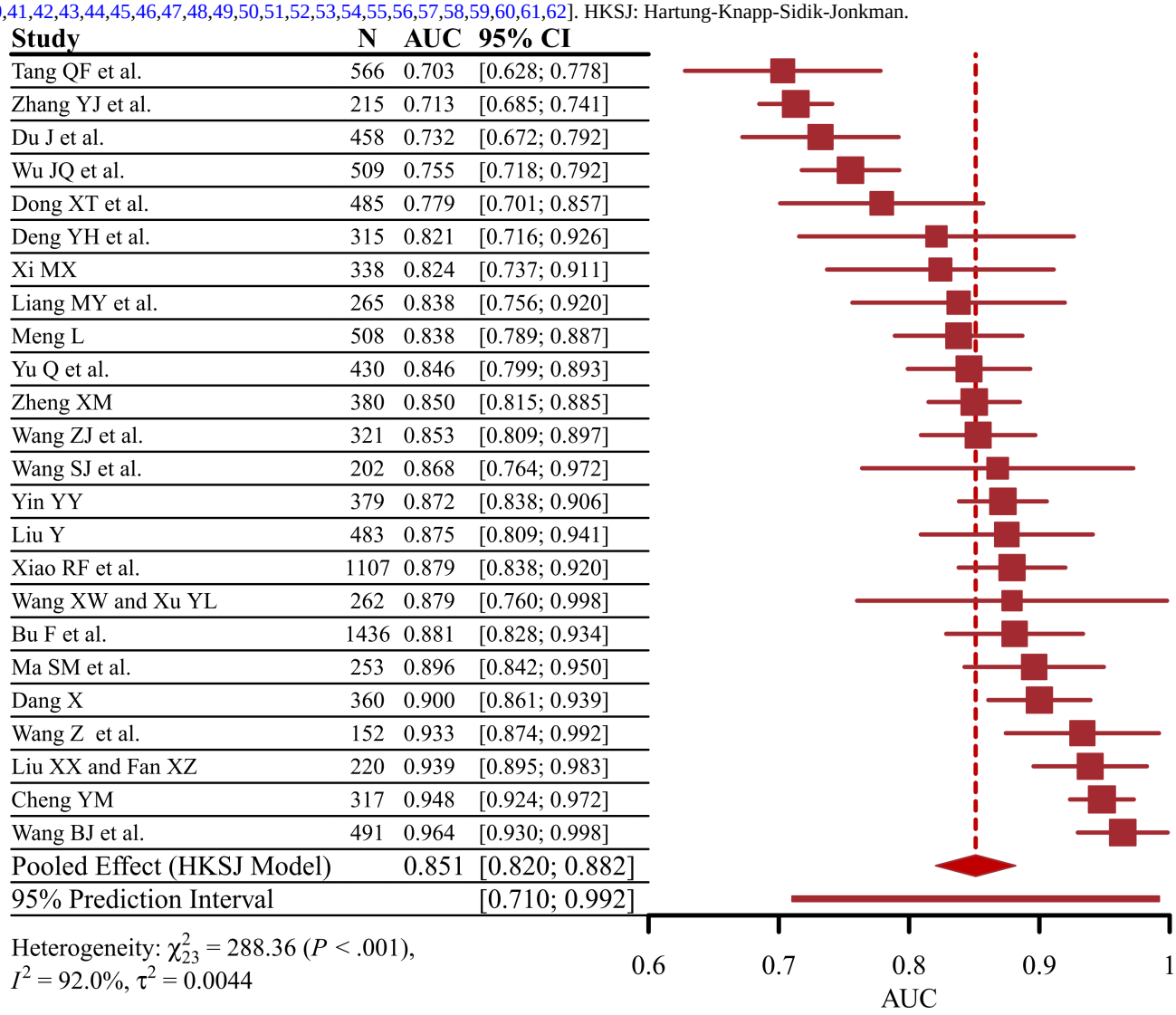

Additional data on sample size, sensitivity, and specificity were extracted from 17 studies to calculate TP, FP, FN, and TN (Table S5 in Multimedia Appendix 1). Based on these values, the pooled sensitivity was 0.810 (95% CI 0.740 - 0.850; I^2=92.26%), as illustrated in the forest plot. The pooled specificity was 0.850 (95% CI 0.810 - 0.890; I^2=92.44%), with the corresponding forest plot also presented (Figure 5B). Furthermore, a summary receiver operating characteristic curve was generated, as depicted in Figure 5A. These results indicate significant heterogeneity across the models regarding AUC, sensitivity, and specificity.

XSL•FO
RenderX

Figure 5. (A) Summary receiver operating characteristic curve and (B) forest plots of the random effects meta-analysis of the sensitivity and specificity for 22 validation models [40-42,44,46-50,52,53,56-58,60-62]. AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

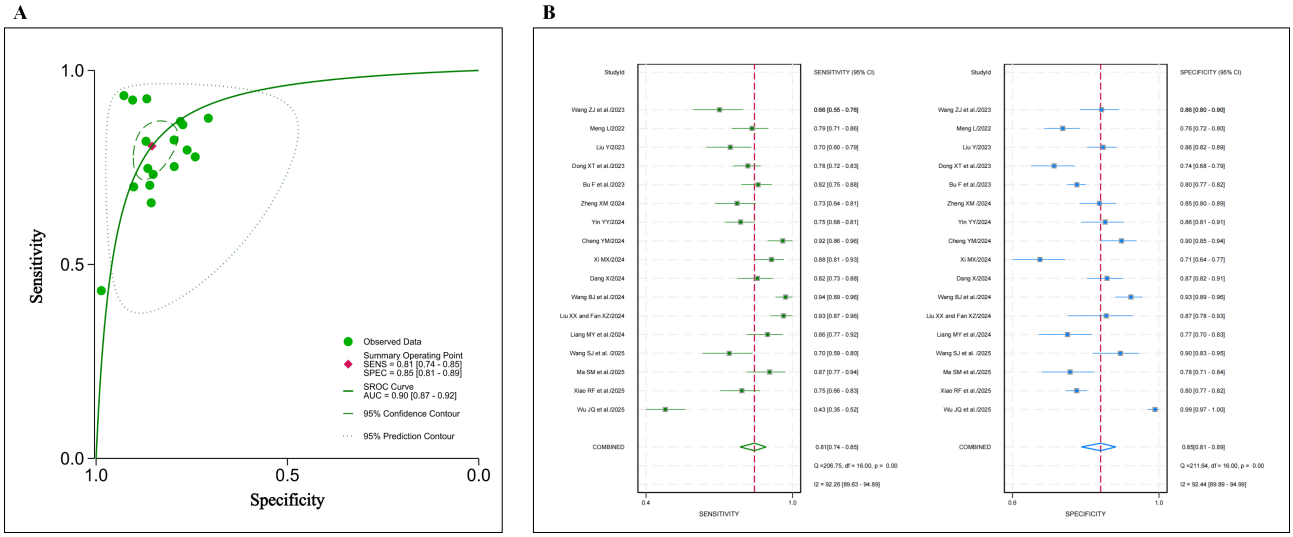

Sensitivity Analysis

Sensitivity analyses were undertaken by sequentially removing one study at a time. The point estimates obtained after excluding any single study all fell within the 95% CI of the overall effect size (Figure 6A). This indicates that the removal of any individual study did not significantly influence the pooled AUC. Therefore, the combined results were relatively stable.

Figure 6. Assessment of robustness, heterogeneity, and small study effects in the meta-analysis of 24 studies evaluating physical frailty and cognitive frailty in older adults with diabetes: (A) sensitivity analysis, (B) subgroup forest plot stratified by modeling approach (32 models), and (C) funnel plot [40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63]. AUC: area under the curve; HKSJ: Hartung-Knapp-Sidik-Jonkman; LR: logistic regression; ML: machine learning.

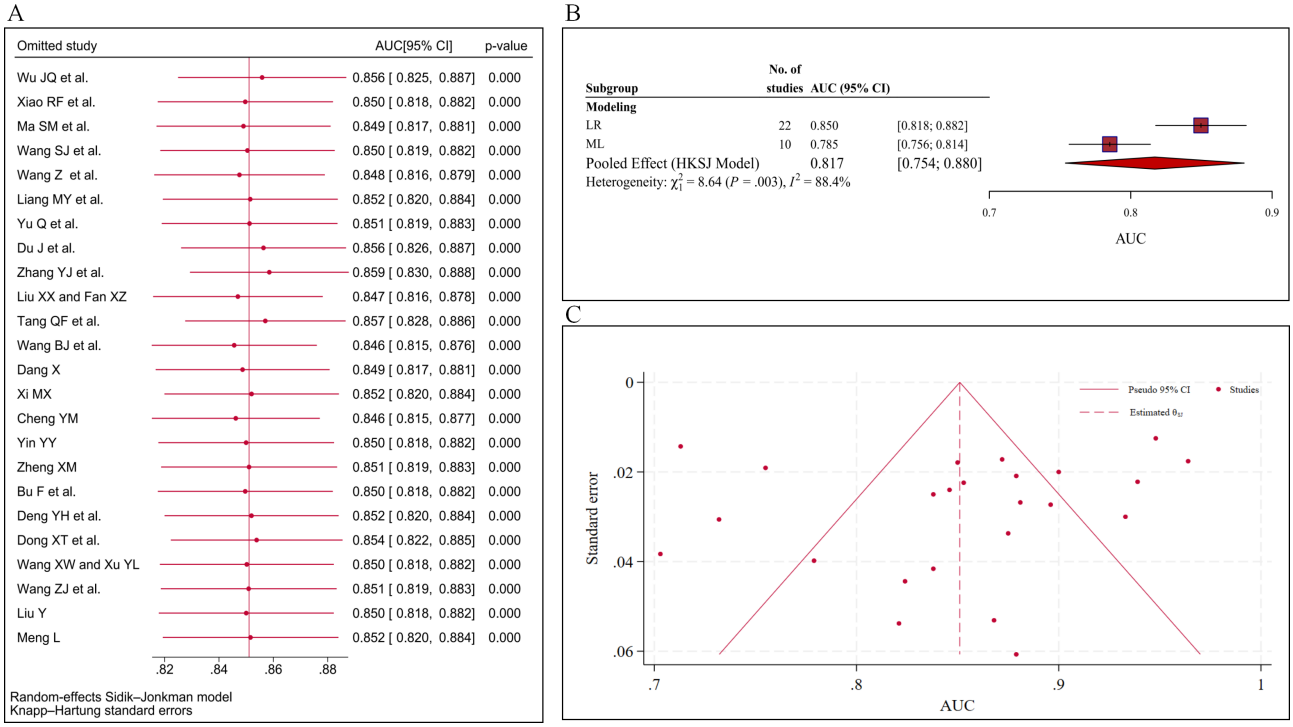

Subgroup Analysis

Subgroup analyses were performed based on modeling approach, study design, data source, population, outcome, and validation method (Table 3 and Figure 6B). Significant differences were observed in modeling approaches (*P*=.003), where LR models yielded a higher pooled AUC (0.850) compared with ML models (0.785). Similarly, study design showed significant heterogeneity (*P*=.03), with retrospective studies demonstrating superior diagnostic performance (AUC=0.900) than cross-sectional studies (AUC=0.843). However, no statistically significant differences were found across subgroups stratified by data source (*P*=.42), patient characteristics (*P*=.77), validation methods (*P*=.16), or specific outcomes (*P*=.94).

XSL•FO
RenderX

Table . Subgroup analysis of the pooled area under the curve (AUC) of studies of physical frailty and cognitive frailty in older adults with diabetes to explore potential sources of heterogeneity (n=24).

Subgroup	Studies, n	AUC (95% CI)	I^2, %	*P* value
Population			0	.77
Diabetes	21	0.849 (0.818-0.879)		
Diabetes with comorbidities or complications	3	0.866 (0.757-0.975)		
Study design			79.6	.03
Cross-sectional study	21	0.843 (0.812-0.875)		
Retrospective study	3	0.900 (0.861-0.939)		
Data source			0	.42
Multicenter	6	0.871 (0.815-0.926)		
Single center	18	0.844 (0.809-0.878)		
Outcome			0	.94
Physical frailty	13	0.851 (0.806-0.896)		
Cognitive frailty	11	0.849 (0.812-0.886)		
Validation			42.5	.16
Random split	12	0.875 (0.836-0.914)		
Bootstrap	9	0.829 (0.779-0.879)		
Cross-validation	2	0.803 (0.710-0.896)		
None	1	0.896 (0.842-0.950)		

Small Study Effects Assessment

We examined small study effects using the Egger test, funnel plots, and the Deeks funnel plot. The Egger test yielded a coefficient of −1.07 (*P*=.40), indicating no statistically significant small study effects. This result aligns with the visual inspection of the contour-enhanced funnel plot (Figure 6C), which displayed a relatively symmetrical distribution of the studies, suggesting no obvious small study effects among the included studies. Furthermore, the Deeks funnel plot asymmetry test yielded a nonsignificant *P* value of .94, providing no evidence of significant small study effects (Figure S2 in Multimedia Appendix 1).

However, these results must be interpreted with caution. Funnel plot symmetry or nonsignificant statistical tests do not definitively rule out publication bias. Conversely, asymmetry may arise from heterogeneity or methodological limitations not merely from publication bias. Therefore, the potential for publication bias cannot be entirely excluded, particularly due to limitations in study selection. Although our search strategy was designed to be global, the final pool of eligible studies consisted exclusively of research conducted in China. Additionally, the restriction of inclusion criteria to English and Chinese languages may have introduced language bias by excluding relevant data published in other languages.

Discussion

Principal Findings

To our knowledge, this is the first systematic review and meta-analysis to comprehensively evaluate the performance of prediction models for physical frailty and cognitive frailty specifically in older adults with diabetes. Our analysis included 24 studies encompassing 32 models. At the study level, the pooled AUC was 0.851 (95% CI 0.820 - 0.882), while the model-level analysis yielded a similarly high pooled AUC of 0.829 (95% CI 0.802 - 0.856). In addition, the pooled sensitivity was 0.810, and the pooled specificity was 0.850, indicating that these models demonstrated reasonable discriminative performance for identifying physical frailty and cognitive frailty in this high-risk population.

A notable finding from our subgroup analysis was the difference in performance based on the modeling approach (*P*=.003). LR models yielded a higher pooled AUC (0.850) than ML models (0.785). However, this finding should be interpreted with caution. Consistent with the systematic review by Christodoulou et al [64], we observed no consistent performance advantage of ML over LR analysis in this dataset. The apparent disparity may be attributed to the smaller number of ML studies included, heterogeneity in populations and predictors, and potential risk of bias, rather than an inherent superiority of LR. For instance, Wang et al [52] and Yin [58] developed models using both approaches and found LR performed slightly better. Ultimately, there is no “one-size-fits-all” modeling method; performance often depends on the specific data structure and clinical context. Therefore, future research should prioritize rigorous comparisons

XSL•FO
RenderX

of multiple modeling approaches—including proper hyperparameter tuning for ML—to identify the optimal strategy for specific physical frailty and cognitive frailty prediction scenarios. Additionally, we observed that retrospective studies yielded significantly higher AUC values (0.900) than cross-sectional studies (0.843; *P*=.03). This phenomenon likely stems from the inherent selection bias and better data quality control often present in retrospective cohorts, potentially leading to overoptimistic performance estimates.

The diagnostic models included in this review possess meaningful clinical implications, facilitating a shift toward the precise identification and risk stratification of physical frailty and cognitive frailty in clinical settings. Our analysis revealed that the most frequently used features—depression, age, regular exercise, social activity, and diabetes duration—provide concrete metrics for identifying these concurrent conditions. Notably, depression emerged as the most consistent and prominent feature across multiple studies, highlighting its strong correlation with both physical frailty and cognitive frailty. Existing evidence indicates a bidirectional relationship between depression and these conditions, potentially mediated through inflammatory pathways, endocrine dysregulation, and overlapping symptoms such as fatigue and psychomotor slowing [65,66]. Consequently, assessing mental health not only is vital for psychological well-being but also serves as a critical entry point for identifying patients who may already be experiencing physical frailty or cognitive frailty. Advanced age was also a predominant feature, consistent with reports by Kong et al [9] and Wang et al [67]. This association likely reflects immunosenescence, chronic inflammation, and metabolic dysregulation, which collectively contribute to sarcopenia and functional decline [68-70]. These findings suggest that age remains a fundamental stratification factor, warranting heightened clinical vigilance for physical and cognitive frailty in older cohorts. Regular exercise was identified as a powerful discriminatory feature. Physiologically, physical activity is known to reduce inflammation, preserve muscle function, and support cognitive health [71]. In the context of these diagnostic models, the absence of regular exercise serves as a robust, easily accessible clinical marker for detecting potential physical frailty and cognitive frailty. This allows health care providers to efficiently target vulnerable populations in community settings where elaborate geriatric assessments may be impractical. Similarly, lower levels of social activity emerged as a significant indicator. This suggests that social isolation often co-occurs with physical frailty and cognitive frailty, making social history a valuable component of the risk stratification process for older adults with diabetes. Finally, the duration of diabetes was a frequent feature in the included models. The strong association between longer disease duration and physical frailty and cognitive frailty likely reflects the cumulative burden of chronic hyperglycemia, insulin resistance, and related complications over time [8,26]. As diabetes duration increases, physiological and cognitive reserves decline, thereby increasing the probability of concurrent physical frailty and cognitive frailty. These insights emphasize that patients with a long history of diabetes represent a high-risk group requiring prioritized screening and comprehensive management.

Comparison With Prior Work

Previous systematic reviews and narrative summaries on physical frailty or cognitive frailty in older adults with diabetes have primarily concentrated on estimating prevalence, identifying associated risk factors, or describing frailty phenotypes, rather than systematically evaluating multivariable prediction or identification models [8,9,72,73]. Moreover, most prior reviews did not attempt a quantitative synthesis of model performance, likely due to the substantial methodological and clinical heterogeneity across studies, which is commonly encountered in the evaluation of prediction models. Consistent with this literature, we observed considerable variation among included studies in terms of study populations, definitions of physical frailty and cognitive frailty, predictor selection, model development strategies, and validation approaches. These differences reflect the evolving and fragmented nature of model development in this field and pose challenges for direct comparison across studies. Nevertheless, by conducting a meta-analysis of discrimination performance, particularly through the synthesis of the AUC, our study provides a quantitative overview of the overall performance of existing models for identifying physical frailty and cognitive frailty in older adults with diabetes.

Heterogeneity

Substantial heterogeneity was observed across the included studies (I^2>90% for sensitivity, specificity, and AUC). This is a common challenge in diagnostic meta-analyses and may be attributed to variations in study design, population characteristics, and modeling methodologies. Our subgroup analysis identified that the modeling approach (ML vs LR) and study design (retrospective vs cross-sectional) were significant sources of heterogeneity (P<.05). However, other factors such as data source (single vs multicenter) and outcome definitions (physical frailty vs cognitive frailty) did not significantly contribute to the observed variance. It is also important to note the variability in feature selection across studies. With 33 different features identified—ranging from depression and age to regular exercise—the lack of a standardized set of predictors likely contributes to the heterogeneity in model performance. This diversity reflects the multifaceted pathophysiology of frailty in diabetes but complicates the direct comparison of models.

Importantly, beyond the conventional I^2 statistic, we further quantified between-study heterogeneity using 95% PIs, which provide a clinically meaningful estimate of the expected range of model performance in future settings. Although I^2 values exceeding 90% indicate substantial relative heterogeneity, they do not convey the absolute extent to which predictive performance may vary across populations and clinical contexts. In contrast, PIs directly address this limitation by reflecting the dispersion of true effects on the original AUC scale. At the study level, the pooled AUC of 0.851 was accompanied by a 95% PI ranging from 0.710 to 0.992, indicating that, although predictive performance may vary considerably across different real-world settings, most future applications are still likely to achieve at least acceptable discrimination.

XSL•FO
RenderX

Methodological Quality and Risk of Bias

Evaluation using the PROBAST checklist indicated that all included studies exhibited a high risk of bias, predominantly in the analysis domain. Consequently, the pooled performance estimates reported in this review should be interpreted with caution, likely representing "best-case scenarios" or optimistic estimates rather than robust predictions of real-world performance.

In the participant domain, specifically regarding data sources, although retrospective designs were identified as a source of bias for a few studies, the overarching issue remains the analytical approach. We recommend using prospective data or registry data for model development in future optimization efforts to reduce the risk of bias arising from data sources. Additionally, the evaluation of model applicability indicated that certain studies included not only patients with diabetes but also those with other comorbidities or complications. These factors limited the applicability of the respective models to the general diabetes population. In the outcome domain, some studies inappropriately incorporated outcome-related information into the predictor assessment, leading to information leakage and inflated model performance. In addition, inadequate reporting regarding the consistency of predictor measurement raised concerns about reproducibility in at least one study. In the outcome domain, several studies were judged to have a high risk of bias due to problematic outcome definitions.

The analysis domain had the highest frequency of a high risk of bias, with all studies rated as high risk in this domain. According to the PROBAST assessment tool, an EPV≥20 is commonly used as a heuristic to indicate an adequate sample size for developing prediction models. In this review, 13 studies had an EPV <20, which may suggest an increased risk of bias related to model overfitting. However, relying solely on fixed rules of thumb may be insufficient; therefore, future studies should prioritize formal, model-tailored sample size calculations (eg, approaches proposed by Riley et al [74]) to ensure precise estimation and adequate statistical power. During the predictor selection process, several studies relied solely on univariate screening, which often fails to identify confounding factors and can lead to model overfitting. Therefore, predictor selection should not solely depend on univariate screening but should also be combined with clinical practice. Moreover, an increasing number of studies are using least absolute shrinkage and selection operator (LASSO) regression to handle high-dimensional data and select potential variables. By introducing a penalty term, LASSO regression reduces the estimates of extreme variables, thereby effectively enhancing the accuracy of model estimation and decreasing the likelihood of overfitting [75]. Moreover, the handling of missing data was suboptimal. Cases with missing data were directly excluded in 12 studies, which can introduce bias and reduce statistical power, while 9 studies failed to report how missing data were handled. A minority of studies used appropriate methods such as multiple imputation. Accurate data reporting and careful handling of missing observations help reduce model overfitting. It is recommended that future studies strengthen the management of missing data to ensure the integrity of the study. When dealing with continuous variables, transforming continuous data into categorical variables for modeling may lead to a significant loss of model efficacy [76]; however, 10 studies in our review performed such transformations without reporting standard definitions. Although data transformation can be considered to enhance the convenience of application for researchers during the clinical dissemination phase, it should be done with caution during development. Crucially, the majority of models lacked external validation. Although they demonstrated good discrimination in derivation cohorts, their performance remains inherently tied to their specific development settings.

Therefore, the primary implication of this review is methodological: Rather than endorsing specific existing tools for immediate clinical use, we emphasize the urgent need for better-designed research. Future studies must strictly adhere to TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guidelines and PROBAST standards—specifically ensuring adequate sample sizes, appropriate handling of continuous variables and missing data, and rigorous external validation in independent, geographically distinct populations—to develop robust and transportable prediction models.

Limitations

This study has several limitations that warrant consideration. First, the predominance of cross-sectional and retrospective designs among the included studies restricts the scope of these models to the diagnostic identification of prevalent frailty. Consequently, they function as concurrent screening tools rather than prognostic instruments for predicting future incidence, precluding the ability to infer causal relationships between predictors and outcomes. Second, according to the PROBAST assessment, all included studies exhibited a high risk of bias, particularly within the analysis domain. This methodological weakness likely results in overoptimistic performance estimates and limits the transportability of these models to diverse clinical populations. Third, substantial statistical heterogeneity (I^2) was observed, stemming from variations in study design and modeling methodologies. We calculated 95% PIs to provide a more clinically meaningful estimate of the expected range of model performance in future settings. Finally, all included studies were conducted in China, and the review was restricted to English and Chinese literature. Although this reflects the rapid emergence of this research focus within China, the lack of geographic and ethnic diversity limits the generalizability of our findings to other global populations and health care systems.

Conclusion

This review provides the first comprehensive synthesis of models for risk stratification of physical and cognitive frailty in older adults with diabetes. The findings indicate that existing models demonstrate satisfactory pooled discriminative performance. Specifically, although the CIs confirm a robust average effect, the 95% PIs indicate that the distribution of predictive performance in future real-world settings is expected to vary across different clinical contexts, yet likely remaining within an acceptable range. Nevertheless, their clinical utility is currently constrained by significant methodological limitations. Specifically, the identified models rely heavily on readily available clinical and psychosocial predictors, such as

XSL•FO
RenderX

depression, age, regular exercise, and social activity, suggesting that early risk stratification is feasible in routine practice. However, the evidence is underpinned by a pervasive high risk of bias, primarily due to analytical shortcomings, small sample sizes, and a lack of rigorous external validation. Furthermore, the predominance of cross-sectional designs and the geographic restriction of studies to China limit the generalizability of these tools to broader global populations and their ability to function as true prognostic instruments for future risk. Consequently, although current models show promise for screening and identifying prevalent physical and cognitive frailty, they are not yet sufficiently robust for widespread deployment in diverse clinical settings. Future research must pivot from developing new, redundant models to conducting robust, prospective, multicenter studies that adhere strictly to TRIPOD guidelines. Emphasis should be placed on external validation and the development of longitudinal prognostic tools to ensure reliable, transportable, and clinically actionable risk stratification for this vulnerable population.

Acknowledgments

The authors declare that generative artificial intelligence (AI) was not used in the creation of this manuscript.

Funding

This study was funded by grants from the nursing research project of Sichuan Province (number H23003) and the research project of the Science and Technology Department of Sichuan Province (number 2024ZYD0338). The funders played no part in the study design, data collection, analysis, interpretation of the results, or the writing of the manuscript.

Disclaimer

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: XW
Data curation: LL, RZ, QJ
Formal analysis: XW, XX, HC
Funding acquisition: YL, RZ, RG
Methodology: XW, SM, XX, RG
Project administration: RG
Resources: LL
Software: QJ
Writing - original draft: XW, XX
Writing - review & editing: XW, XM, XX, LL, HC, YL, RZ, QJ, SL, RG, SM

Conflicts of Interest

None declared.

Multimedia Appendix 1
Inclusion criteria, search terms, performance measures, and assessments of bias.
[DOCX File, 495 KB - jmir_v28i1e84617_app1.docx]

Checklist 1
PRISMA checklist.
[PDF File, 106 KB - jmir_v28i1e84617_app2.pdf]

References

1. Genitsaridi I, Salpea P, Salim A, et al. 11th edition of the IDF Diabetes Atlas: global, regional, and national diabetes prevalence estimates for 2024 and projections for 2050. Lancet Diabetes Endocrinol 2026 Feb;14(2):149-156. [doi: 10.1016/S2213-8587(25)00299-2] [Medline: 41412135]

XSL•FO
RenderX

2. Kobayashi E, Linden-Santangeli NJ, Chan N, et al. Longitudinal metabolic trajectories in diabetes prevention program participants reveal subgroups with varying micro- and macrovascular complication risks. Diabetes Care 2025 Oct 1;48(10):1704-1712. [doi: 10.2337/dc25-0866] [Medline: 40857122]
3. Xu X, Wang F, Liao S, Liu J, Xiao L. Tele-cognitive behavioral therapy for the treatment of diabetes-related distress in individuals With diabetes mellitus: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2025 Dec 24;27:e80476. [doi: 10.2196/80476] [Medline: 41442672]
4. Ong KL, Stafford LK, McLaughlin SA. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2023 Jul 15;402(10397):203-234. [doi: 10.1016/S0140-6736(23)01301-6] [Medline: 37356446]
5. O'Neil H, Todd A, Pearce M, Husband A. What are the consequences of over and undertreatment of type 2 diabetes mellitus in a frail population? A systematic review. Endocrinol Diabetes Metab 2024 Mar;7(2):e00470. [doi: 10.1002/edm2.470] [Medline: 38411378]
6. Clegg A, Young J, Iliffe S, Rikkert MO, Rockwood K. Frailty in elderly people. The Lancet 2013 Mar;381(9868):752-762. [doi: 10.1016/S0140-6736(12)62167-9]
7. Wu Y, Xiong T, Tan X, Chen L. Frailty and risk of microvascular complications in patients with type 2 diabetes: a population-based cohort study. BMC Med 2022 Dec 8;20(1):473. [doi: 10.1186/s12916-022-02675-9] [Medline: 36482467]
8. Hanlon P, Fauré I, Corcoran N, et al. Frailty measurement, prevalence, incidence, and clinical implications in people with diabetes: a systematic review and study-level meta-analysis. Lancet Healthy Longev 2020 Dec;1(3):e106-e116. [doi: 10.1016/S2666-7568(20)30014-3] [Medline: 33313578]
9. Kong LN, Lyu Q, Yao HY, Yang L, Chen SZ. The prevalence of frailty among community-dwelling older adults with diabetes: a meta-analysis. Int J Nurs Stud 2021 Jul;119:103952. [doi: 10.1016/j.ijnurstu.2021.103952] [Medline: 34022743]
10. Huang ST, Chen LK, Hsiao FY. Clinical impacts of frailty on 123,172 people with diabetes mellitus considering the age of onset and drugs of choice: a nationwide population-based 10-year trajectory analysis. Age Ageing 2023 Jul 1;52(7):afad128. [doi: 10.1093/ageing/afad128] [Medline: 37505989]
11. Walston J, McBurnie MA, Newman A, et al. Frailty and activation of the inflammation and coagulation systems with and without clinical comorbidities: results from the Cardiovascular Health Study. Arch Intern Med 2002 Nov 11;162(20):2333-2341. [doi: 10.1001/archinte.162.20.2333] [Medline: 12418947]
12. Zhang X, Bo Y, Li Z, et al. Association between frailty and cognitive function: a pooled analysis of three ageing cohorts. Transl Psychiatry 2025 Nov 18;15(1):486. [doi: 10.1038/s41398-025-03674-z]
13. Li CL, Stanaway FF, Lin JD, Chang HY. Frailty and health care use among community-dwelling older adults with diabetes: a population-based study. Clin Interv Aging 2018;13:2295-2300. [doi: 10.2147/CIA.S183681] [Medline: 30519011]
14. Ferri-Guerra J, Aparicio-Ugarriza R, Salguero D, et al. The association of frailty with hospitalizations and mortality among community dwelling older adults with diabetes. J Frailty Aging 2020;9(2):94-100. [doi: 10.14283/jfa.2019.31] [Medline: 32259183]
15. Wang X, Chen Z, Li Z, et al. Association between frailty and risk of fall among diabetic patients. Endocr Connect 2020 Oct;9(10):1057-1064. [doi: 10.1530/EC-20-0405] [Medline: 33112808]
16. Santulli G, Sabatelli G, Wang B, et al. Interplay between frailty and cardiometabolic disorders: from pathophysiology to clinical implications. Cardiovasc Diabetol 2025 Dec 8;25(1):1. [doi: 10.1186/s12933-025-03022-x] [Medline: 41361441]
17. Zhang RH, Wang J, Wang Y, et al. Frailty and transitions across cardiometabolic disease states: evidence from multistate models in a 16-year Chinese cohort. NPJ Aging 2025 Nov 29;12(1):4. [doi: 10.1038/s41514-025-00301-5] [Medline: 41318578]
18. Fan J, Yu C, Guo Y, et al. Frailty index and all-cause and cause-specific mortality in Chinese adults: a prospective cohort study. Lancet Public Health 2020 Dec;5(12):e650-e660. [doi: 10.1016/S2468-2667(20)30113-4] [Medline: 33271078]
19. Haapanen MJ, Jansson Sigfrids F, Ylinen A, et al. Frailty outperforms conventional risk factors for predicting complications and death in type 1 diabetes. Diabetes Res Clin Pract 2025 Dec;230:112984. [doi: 10.1016/j.diabres.2025.112984] [Medline: 41240994]
20. Guo X, Pei J, Ma Y, et al. Cognitive frailty as a predictor of future falls in older adults: a systematic review and meta-analysis. J Am Med Dir Assoc 2023 Jan;24(1):38-47. [doi: 10.1016/j.jamda.2022.10.011] [Medline: 36423679]
21. Ren M, Guo H, Guo Y, Guo W, Zhu L. The risk prediction models for cognitive frailty in the older people in China: a systematic review and meta-analysis. BMC Geriatr 2025 May 22;25(1):365. [doi: 10.1186/s12877-025-05961-2] [Medline: 40405068]
22. Si H, Zhang Y, Zhao P, et al. Bidirectional relationship between diabetes and frailty in middle-aged and older adults: a systematic review and meta-analysis. Arch Gerontol Geriatr 2025 Aug;135:105880. [doi: 10.1016/j.archger.2025.105880] [Medline: 40319625]
23. Cheng M, He M, Ning L, et al. The impact of frailty on clinical outcomes among older adults with diabetes: a systematic review and meta-analysis. Medicine (Abingdon) 2024;103(26):e38621. [doi: 10.1097/MD.0000000000038621]
24. Sinclair AJ, Pennells D, Abdelhafiz AH. Hypoglycaemic therapy in frail older people with type 2 diabetes mellitus-a choice determined by metabolic phenotype. Aging Clin Exp Res 2022 Sep;34(9):1949-1967. [doi: 10.1007/s40520-022-02142-8] [Medline: 35723859]

XSL•FO
RenderX

25. Ji CH, Huang XQ, Li Y, Muheremu A, Luo ZH, Dong ZH. The relationship between physical activity, nutritional status, and sarcopenia in community- dwelling older adults with type 2 diabetes: a cross-sectional study. BMC Geriatr 2024 Jun 7;24(1):506. [doi: 10.1186/s12877-024-05038-6] [Medline: 38849763]
26. Sinclair AJ, Abdelhafiz AH. Metabolic impact of frailty changes diabetes trajectory. Metabolites 2023 Feb 16;13(2):295. [doi: 10.3390/metabo13020295] [Medline: 36837914]
27. Zhong W, Huang W, Deng H, Qiu S, Yang Q, Jia H. A randomized controlled trial to assess the efficacy of standardized tai chi in prefrail older adults with immunosenescence: design and protocol. BMC Complement Med Ther 2025 Jan 3;25(1):1. [doi: 10.1186/s12906-024-04732-7] [Medline: 39754159]
28. Sobrinho ACDS, de Paula Venancio RC, da Silva Rodrigues G, et al. Systematic review of interventions for pre-frail and frail older adults: evidence from clinical trials on frailty levels. Arch Gerontol Geriatr 2025 Jul;134:105851. [doi: 10.1016/j.archger.2025.105851] [Medline: 40262339]
29. Page MJ, Moher D, Bossuyt PM, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021 Mar 29;372:n160. [doi: 10.1136/bmj.n160] [Medline: 33781993]
30. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: the PRISMA-DTA statement. JAMA 2018 Jan 23;319(4):388-396. [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
31. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
32. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019 Jan 1;170(1):51-58. [doi: 10.7326/M18-1376] [Medline: 30596875]
33. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015 Jan 6;162(1):W1-W73. [doi: 10.7326/M14-0698] [Medline: 25560730]
34. Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014 Oct;11(10):e1001744. [doi: 10.1371/journal.pmed.1001744] [Medline: 25314315]
35. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019 Jan 1;170(1):W1-W33. [doi: 10.7326/M18-1377] [Medline: 30596876]
36. Šimundić AM. Measures of diagnostic accuracy: basic definitions. EJIFCC 2009 Jan;19(4):203-211. [Medline: 27683318]
37. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 6;327(7414):557-560. [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
38. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
39. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634. [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
40. Bu F, Deng XH, Zhan NN, et al. Development and validation of a risk prediction model for frailty in patients with diabetes. BMC Geriatr 2023 Mar 27;23(1):172. [doi: 10.1186/s12877-023-03823-3] [Medline: 36973658]
41. Cheng YM. Construction of a predictive model for the risk of frailty in hospitalized elderly patients with diabetes mellitus [Master's thesis]. : Gannan Medical University; 2024 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1024557771.nh&uniplatform=OVERSEA&v=aCGchnfEAAyXVFPDXN9Dm_iJvJYtdsledsnnFwYOOKq65nJCZmCdm-7Y3NSbCg_U [accessed 2026-01-22] [doi: 10.27959/d.cnki.ggnyx.2024.000179]
42. Dang X. Study on the construction of a model for predicting the risk of debility in hospitalized elderly patients with diabetes mellitus [Master's thesis]. : Changchun University of Chinese Medicine; 2024 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1024551416.nh&uniplatform=OVERSEA&v=saoWsbwFh_hMUrvE2hYKMbmPvq6dTUrm6EUZP7P8Emfni9qewTW8gjWKg0liRN7b [accessed 2026-01-22] [doi: 10.26980/d.cnki.gcczc.2024.000284]
43. Deng YH, Li N, Wang Y, Xiong C, Zou X. Risk factors and prediction nomogram of cognitive frailty with diabetes in the elderly. Diabetes Metab Syndr Obes 2023;16:3175-3185. [doi: 10.2147/DMSO.S426315] [Medline: 37867632]
44. Dong XT, Wang XH, Sheng YH, Wang GP, Gu T. Construction of frailty risk prediction model in elderly patients with diabetic retinopathy. Journal of Gannan Medical University 2023;43(12):1275-1281. [doi: 10.3969/j.issn.1001-5779.2023.12.013]
45. Du J, Zhang D, Chen Y, Zhang W. Development of a prediction model for frailty among older Chinese individuals with type 2 diabetes residing in the community. Public Health Nurs 2024;41(6):1271-1280. [doi: 10.1111/phn.13377] [Medline: 39101656]

XSL•FO
RenderX

46. Liang MY, Li R, Feng L, Qian WP. Construction and verification of a risk prediction model for cognitive frailty in older patients with chronic obstructive pulmonary disease and diabetes mellitus. J Int Med Res 2024 Sep;52(9):3000605241274211. [doi: 10.1177/03000605241274211] [Medline: 39224937]
47. Liu XX, Fan XZ. Establishment and validation of a prediction model for cognitive frailty in elderly patients with type 2 diabetes mellitus. Anhui Medical Journal 2024;45(12):1543-1548. [doi: 10.3969/j.issn.1000-0399.2024.12.014]
48. Liu Y. Construction and verification of cognitive frailty risk prediction model in elderly patients with type 2 diabetes [Master's thesis]. : Tianjin University of Traditional Chinese Medicine; 2023 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1023871608.nh&uniplatform=OVERSEA&v=bMx3cCr3OPvyjITKNyqnW0JcHM8BIfzLFhPkowxsujZYch6jIjA1m7lsencsQrYF [accessed 2026-01-22] [doi: 10.27368/d.cnki.gtzyy.2023.000432]
49. Ma SM, Ni LL, Guo L, Gu LP. Construction of a risk prediction model for cognitive decline in elderly patients with diabetic foot ulcers. Psychological Monthly 2025;20(2):52-54 [FREE Full text] [doi: 10.19738/j.cnki.psy.2025.02.014]
50. Meng L. Research on the influencing factors of cognitive frailty and the construction of risk prediction model in elderly diabetic patients [Master's thesis]. : Zunyi Medical University; 2023 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1024333258.nh&uniplatform=OVERSEA&v=OQcVHZS3INx0BTBGPkZnRQy4Deev1Fhcs49_DXQJd8j1MnacoqsS81_n7RhEcusZ [accessed 2026-01-26] [doi: 10.27680/d.cnki.gzyyc.2023.000437]
51. Tang QF, Wang JJ, Zhao HY, Wang HY. Establishment and validation of a frailty risk prediction model for elderly patients with diabetes mellitus. Chinese Nursing Research 2024;38(17):3065-3071. [doi: 10.12102/j.issn.1009-6493.2024.17.009]
52. Wang BJ, Liang Q, Liu Y, Cheng Y, Zhang CM. Construction and validation of frailty risk prediction model in elderly patients with diabetic foot. Mil Nurs 2024;41(5):6-10. [doi: 10.3696/j.issn.2097-1826.2024.05.002]
53. Wang SJ, Tan TT, Wang QQ, et al. Construction of a risk prediction model for cognitive decline in elderly hospitalized patients with type 2 diabetes based on decision tree. Sichuan Medical Journal 2025;46(2):204-210. [doi: 10.16252/j.cnki.issn1004-0501-2025.02.017]
54. Wang XW, Xu YL. Prediction of cognitive decline among elderly patients with type 2 diabetes mellitus. China Preventive Medicine Journal 2023;35(12):1037-1042. [doi: 10.19485/j.cnki.issn2096-5087.2023.12.006]
55. Wang Z, Zheng HF, Liang LL. Analysis of risk factors of elderly patients with type 2 diabetes complicated with frailty and establishment of prediction model. Int J Geriatr 2025;46(2):162-168. [doi: 10.3969/j.issn.1674-7593.2025.02.007]
56. Wang ZJ, Chen JY, Li MX. Establishment of a nomogram predictive model based on serum leptin-to-adiponectin ratio for cognitive frailty in elderly patients with type 2 diabetes mellitus. Shandong Medical Journal 2023;63(34):31-37. [doi: 10.3969/j.issn.1002-266X.2023.34.007]
57. Xi MX. Construction of prediction model for frailty in hospitalized elderly patients with diabetes mellitus [Master's thesis]. : Yan'an University; 2024 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1024455798.nh&uniplatform=OVERSEA&v=S5OEfMlXAp2MuMsWXc_GQtzuxBN-77f0i0NyJbd51w8xQlUaiOwjOVtxUElm5mcS [accessed 2026-01-22] [doi: 10.27438/d.cnki.gyadu.2024.000104]
58. Yin YY. Construction and verification of frailty risk prediction model in elderly diabetic patients [Master's thesis]. : Hebei University; 2024 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1024530021.nh&uniplatform=OVERSEA&v=SApAmgAzcplPDX5BRNmZPhLMUnWViopteNW4E_KeL-FHGr3Rn323ES_7vpnrm8Sb [accessed 2026-01-22] [doi: 10.27103/d.cnki.ghebu.2024.000915]
59. Zhang YJ, Tian XF, Zhang H, Tang ZY, Zhang ZY. Influencing factors of cognitive decline in elderly patients with type 2 diabetes mellitus and hypertension and construction of nomogram model for predicting its risk. Practical Journal of Cardiac Cerebral Pneumal and Vascular Disease 2024;32(12):49-54 [FREE Full text] [doi: 10.12114/j.issn.1008-5971.2024.00.257]
60. Zheng XM. Research on the construction of frailty prediction model for elderly hospitalized patients with diabetes based on machine learning algorithm. : Yangtze University; 2024 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1024659916.nh&uniplatform=OVERSEA&v=iU31hKCfxDkcQQIIuCxhedEIj2mwaaiQ5ki4TyGPmRJytm6uY7RU4TpTMk0kFlYT [accessed 2026-01-22] [doi: 10.26981/d.cnki.gjhsc.2024.000756]
61. Xiao RF, Wang R, Xu L, Xu MJ. Risk factors for the development of frailty in Chinese elderly diabetic patients and the predictive value of the nomogram model. Journal of Lanzhou University (Medical Sciences) 2025;51(10):20-25. [doi: 10.13885/j.issn.2097-681X.2025.10.004]
62. Wu JQ, Fang SZ, Li K, et al. Construction and validation of frailty risk prediction model for hospitalized elderly patients with type 2 diabetes based on machine learning and SHAP explainabilility. Journal of Nursing (China) 2025;32(11):7-12. [doi: 10.16460/j.issn2097-6569.2025.11.007]
63. Yu Q, Yu H. Development and validation of a risk prediction model for cognitive frailty in elderly patients with type 2 diabetes mellitus. J Clin Nurs 2025 Aug;34(8):3261-3275. [doi: 10.1111/jocn.17508] [Medline: 39809596]

XSL•FO
RenderX

64. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019 Jun;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004] [Medline: 30763612]
65. Chu W, Chang SF, Ho HY, Lin HC. The relationship between depression and frailty in community-dwelling older people: a systematic review and meta-analysis of 84,351 older adults. J Nurs Scholarsh 2019 Sep;51(5):547-559. [doi: 10.1111/jnu.12501] [Medline: 31328878]
66. Borges MK, Aprahamian I, Romanini CV, et al. Depression as a determinant of frailty in late life. Aging Ment Health 2021 Dec;25(12):2279-2285. [doi: 10.1080/13607863.2020.1857689] [Medline: 33307781]
67. Wang Q, Wang J, Dai G. Prevalence, characteristics, and impact on health outcomes of frailty in elderly outpatients with diabetes: a cross-sectional study. Medicine (Abingdon) 2023;102(47):e36187. [doi: 10.1097/MD.0000000000036187]
68. Ferrucci L, Fabbri E. Inflammageing: chronic inflammation in ageing, cardiovascular disease, and frailty. Nat Rev Cardiol 2018 Sep;15(9):505-522. [doi: 10.1038/s41569-018-0064-2] [Medline: 30065258]
69. Zeng M, Li Y, Zhu Y, Sun Y. Inflammatory markers and clinical factors as key independent risk factors for frailty: a retrospective study. BMC Geriatr 2025 Jun 4;25(1):404. [doi: 10.1186/s12877-025-06033-1] [Medline: 40468183]
70. Straub RH. Interaction of the endocrine system with inflammation: a function of energy and volume regulation. Arthritis Res Ther 2014 Feb 13;16(1):203. [doi: 10.1186/ar4484] [Medline: 24524669]
71. Sutkowy P, Woźniak A, Mila-Kierzenkowska C, et al. Physical activity vs. redox balance in the brain: brain health, aging and diseases. Antioxidants (Basel) 2021 Dec 30;11(1):95. [doi: 10.3390/antiox11010095] [Medline: 35052600]
72. Wen L, Lu Y, Li X, An Y, Tan X, Chen L. Association of frailty and pre-frailty with all-cause and cardiovascular mortality in diabetes: three prospective cohorts and a meta-analysis. Ageing Res Rev 2025 Apr;106:102696. [doi: 10.1016/j.arr.2025.102696] [Medline: 39971101]
73. Lyu Q, Guan CX, Kong LN, Zhu JL. Prevalence and risk factors of cognitive frailty in community-dwelling older adults with diabetes: a systematic review and meta-analysis. Diabet Med 2023 Jan;40(1):e14935. [doi: 10.1111/dme.14935] [Medline: 35962598]
74. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ 2020 Mar 18;368:m441. [doi: 10.1136/bmj.m441] [Medline: 32188600]
75. Zarringhalam K, Degras D, Brockel C, Ziemek D. Robust phenotype prediction from gene expression data using differential shrinkage of co-regulated genes. Sci Rep 2018 Jan 19;8(1):1237. [doi: 10.1038/s41598-018-19635-0] [Medline: 29352257]
76. Rhemtulla M, Brosseau-Liard P, Savalei V. When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. Psychol Methods 2012 Sep;17(3):354-373. [doi: 10.1037/a0029315] [Medline: 22799625]

Abbreviations

AUC: area under the curve
CHARMS: Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies
CNKI: China National Knowledge Infrastructure
EPV: events per variable
FN: false negative
FP: false positive
FRAIL: Fatigue, Resistance, Ambulation, Illnesses, and Loss of Weight
IDF: International Diabetes Federation
LASSO: least absolute shrinkage and selection operator
LR: logistic regression
MeSH: Medical Subject Headings
ML: machine learning
PI: prediction interval
PITROS: Participants, Index Test, Target Conditions, Reference Standard, Outcomes, Settings
PRESS: Peer Review of Electronic Search Strategies
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction Model Study Risk Of Bias Assessment Tool
TN: true negative
TP: true positive
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

XSL•FO
RenderX

Edited by S Brini; submitted 23.Oct.2025; peer-reviewed by A Shivanna, CY Su; revised version received 04.Jan.2026; accepted 05.Jan.2026; published 29.Jan.2026.

Please cite as:
Wang X, Meng S, Xiao X, Lu L, Chen H, Li Y, Zhang R, Jiang Q, Liu S, Gao R
Characterization of Models for Identifying Physical and Cognitive Frailty in Older Adults With Diabetes: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e84617
URL: https://www.jmir.org/2026/1/e84617
doi:10.2196/84617

© Xia Wang, Shujie Meng, Xiang Xiao, Liu Lu, Hongyan Chen, Yong Li, Rong Zhang, Qiwu Jiang, Shan Liu, Ru Gao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Predictive Performance of Artificial Intelligence Algorithms for Gestational Diabetes Mellitus in Pregnant Women: Systematic Review and Meta-Analysis

Yingni Liang[1*], MM; Anran Dai[1*], MM; Meiyan Luo[2*], MM; Zhuolian Zheng[1], MM; Jiayu Shen[1], MM; Yinhua Su[1*], PhD; Zhongyu Li[1*], PhD

[1]School of Nursing, University of South China, No. 28, Changsheng West Road, Hengyang, Hunan, China

[2]Department of Obstetrics, Second Affiliated Hospital of the University of South China, Hunan, China

[*]these authors contributed equally

Corresponding Author:
Zhongyu Li, PhD
School of Nursing, University of South China, No. 28, Changsheng West Road, Hengyang, Hunan, China

Abstract

Background: Gestational diabetes mellitus (GDM) is a common complication during pregnancy, with its incidence increasing year by year. It poses numerous adverse health effects on both mothers and newborns. Accurate prediction of GDM can significantly improve patient prognosis. In recent years, artificial intelligence (AI) algorithms have been increasingly used in the construction of GDM prediction models. However, there is still no consensus on the most effective algorithm or model.

Objective: This study aimed to evaluate and compare the performance of existing GDM prediction models constructed using AI algorithms and propose strategies for enhancing model generalizability and predictive accuracy, thereby providing evidence-based insights for the development of more accurate and effective GDM prediction models.

Methods: A comprehensive search was conducted across PubMed, Web of Science, Cochrane Library, EMBASE, Scopus, and OVID, covering publications from the inception of databases to June 1, 2025, to include studies that developed or validated GDM prediction models based on AI algorithms. Study selection, data extraction, and risk of bias assessment using the Prediction Model Risk of Bias Assessment Tool were performed independently by 2 reviewers. A bivariate mixed-effects model was used to summarize sensitivity and specificity and to generate a summary receiver operating characteristic (SROC) curve, calculating area under the curve (AUC). The Hartung-Knapp-Sidik-Jonkman method was further used to adjust for the pooled sensitivity and specificity. Between-study standard deviation (τ) and variance (τ^2) were extracted from the bivariate model to quantify absolute heterogeneity. The Deek test was used to evaluate small-study effects among included studies. Additionally, subgroup analysis and meta-regression were conducted to compare the performance differences among algorithms and to explore sources of heterogeneity.

Results: Fourteen studies reported on the predictive value for AI algorithms for GDM. After adjustment with the Hartung-Knapp-Sidik-Jonkman method, the pooled sensitivity and specificity were 0.78 (95% CI 0.69 - 0.86; τ=0.15, τ^2=0.02; PI 0.47 - 1.09) and 0.85 (95% CI 0.78 - 0.92; τ=0.11, τ^2=0.01; PI 0.59 - 1.11), respectively. The SROC curve showed that the AUC for predicting GDM using AI algorithms was 0.94 (95% CI 0.92 - 0.96), indicating a strong predictive capability. Deek test (*P*=.03) and the funnel plot both showed clear asymmetry, suggesting the presence of small-study effects. Subgroup analysis showed that the random forest algorithm exhibited the highest sensitivity (0.83, 95% CI 0.74 - 0.93), while the extreme gradient boosting algorithm exhibited the highest specificity (0.82, 95% CI 0.77 - 0.87). Meta-regression further revealed an evaluation in predictive accuracy in prospective study designs (regression coefficient=2.289, *P*=.001).

Conclusions: Unlike previous narrative reviews, this systematic review innovatively provided a comparative and quantitative synthesis of AI algorithms for GDM prediction. This established an evidence-based framework to guide model selection and identified a critical evidence gap. The key implication for real-world application was the demonstrated necessity of local validation before clinical adoption. Therefore, future work should focus on large-scale, prospective validation studies to develop clinically applicable tools.

Trial Registration: PROSPERO CRD42025645913; https://www.crd.york.ac.uk/PROSPERO/view/CRD42025645913

(J Med Internet Res 2026;28:e79729) doi:10.2196/79729

XSL•FO
RenderX

KEYWORDS

gestational diabetes mellitus; artificial intelligence; prediction; meta-analysis; PRISMA; Preferred Reporting Items for Systematic Reviews and Meta-Analysis

Introduction

Gestational diabetes mellitus (GDM) is one of the most common metabolic disorders during pregnancy, characterized by glucose metabolism abnormalities that first appear during gestation [1]. The incidence of GDM has risen to 15.8% due to factors like increased childbearing age, dietary changes, and pre-pregnancy obesity [2-4]. GDM not only significantly increased the risk of adverse pregnancy outcomes for pregnant women, such as macrosomia, preterm birth, and preeclampsia, but also had a profound impact on the long-term health of their offspring, including an increased risk of developing obesity, type 2 diabetes, and other metabolic disorders in the future [5-7]. Therefore, early prediction and management of GDM could effectively reduce the incidence of GDM and its associated maternal and neonatal complications, thereby optimizing perinatal care and improving long-term health outcomes.

The emergence of artificial intelligence (AI) algorithms in medicine has opened new frontiers for predictive analytics, offering the potential to model complex, non-linear interactions within multidimensional health data [8]. In fields such as oncology, cardiology, and endocrinology, AI-driven prediction models have demonstrated superior discriminative accuracy compared to conventional statistical approaches, largely by capturing subtle patterns and interactions among risk factors that traditional methods might overlook [9-12]. This capability was particularly salient for GDM, a condition influenced by a dynamic interplay of genetic, metabolic, hormonal, and lifestyle factors [13].

Building on this general capability, the application of AI algorithms for the specific task of GDM prediction has gained considerable momentum, with primary attention to 2 domains: machine learning (ML) and deep learning (DL) [14-16]. Commonly used ML algorithms, such as random forest (RF), support vector machine, and extreme gradient boosting (XGBoost), have been applied to structured clinical and biomarker data, while DL algorithms typically use neural networks to exploit high-dimensional inputs, including eHealth records and even image-based data [17]. Despite promising reported accuracies, a critical and persistent challenge is the marked heterogeneity in model performance across different populations and settings [18-20]. The ML model developed by Gallardo et al [21], based on routine early-pregnancy examination data, showed high predictive accuracy in a particular population but performed poorly in other GDM populations due to differences in data characteristics. This discrepancy revealed a severe methodological inconsistency in these studies, such as the lack of standardized data preprocessing, non-uniform validation strategies, and incomplete reporting of performance metrics. This heterogeneity made it difficult to directly compare and integrate the results of different studies.

Consequently, although a growing body of primary studies investigating AI models for GDM prediction, the evidence in this field remained fragmented and methodologically heterogeneous. Currently, for the prediction of GDM, there was still a lack of systematic reviews and meta-analyses that could directly compare multiple AI algorithms head-to-head, quantitatively assess their cross-population applicability, and systematically examine methodological rigor. The majority of existing original studies have developed single-algorithm models and validated them only within mono-ethnic or single-center cohorts [16,17,21,22]. Consequently, clinicians lack the high-level evidence required to determine which algorithm is superior and whether reported accuracies generalize to other settings, which markedly impedes the credible clinical adoption and broader dissemination of AI-based prediction models.

To address these evidence gaps, this systematic review and meta-analysis aimed to quantitatively synthesize the predictive performance of prediction models constructed using AI algorithms across different scenarios for GDM, compare the effectiveness of different AI algorithms, and identify the key factors influencing performance. By providing a rigorous, evidence-based framework for evaluating and comparing AI prediction models in GDM, this systematic review sought to inform the future development of more robust, generalizable, and clinically actionable tools, thereby supporting efforts toward early identification, risk stratification, and personalized management of GDM.

Methods

Registration and Protocol

This systematic review adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) 2020 extended checklist, with extensions for Diagnostic Test Accuracy (PRISMA-DTA) and literature search reporting (PRISMA-S) [23-25]. The protocol was prospectively registered with PROSPERO (International Prospective Register of Systematic Reviews; ID CRD42025645913). And the registration was completed on February 13, 2025, prior to the commencement of data extraction and analysis (Checklist 1).

Information Sources and Search Strategy

A comprehensive search was conducted across 6 databases, including PubMed, Web of Science, Cochrane Library, Scopus, EMBASE, and OVID, from the inception of each database to June 1, 2025. To enhance the accuracy of the search results and avoid the omission of relevant studies, the research term developed a rigorous search strategy by combining Medical Subject Headings terms, keywords, and synonyms. No previously published search filters were applied so as to maintain a highly sensitive search strategy. Table 1 summarizes the core search concepts and representative terms. And the detailed search strategy is presented in Multimedia Appendix 1. In addition, we also reviewed the reference lists of relevant

XSL•FO
RenderX

literature, particularly systematic reviews related to the topic of this study, and conducted additional searches in the electronic databases to minimize the omission of the key literature as much as possible. All searches were conducted under the supervision of an academic librarian.

Table . Search strategy using the population, Intervention framework for artificial intelligence–based gestational diabetes mellitus prediction studies.

Concept	Key terms (PubMed example)
Population	"Gestational Diabetes Mellitus" OR "Pregnancy-induced Diabetes" OR "GDM" OR "Diabetes in Pregnancy" OR "Maternal Diabetes"
Intervention	"Artificial Intelligence" OR "Machine Learning Algorithms" OR "Deep Learning Algorithms" OR "Ensemble Learning Algorithms"

Eligibility Criteria

To screen out the original studies relevant to this systematic review from the retrieved literature, detailed inclusion and exclusion criteria were defined (Textbox 1).

Textbox 1. Inclusion and exclusion criteria

Inclusion criteria

- Studies that conducted among pregnant women with gestational diabetes mellitus (GDM) or those at risk of developing GDM.
- Studies that completely constructed one or more predictive models for predicting GDM.
- Studies that used AI algorithms for the construction of a predictive model.
- Studies published in English.

Exclusion criteria

- Reviews, meta-analysis, protocols, letters, conference abstracts, case reports, and animal studies.
- Studies on the predictive accuracy of single-factor predictors.
- Studies only conducted a risk factor analysis without constructing a predictive model.
- Studies did not include any outcome measures for assessing the predictive accuracy of the predictive model.

Selection and Data Collection Process

Following the completion of the systematic research, all records were imported into the reference management software Endnote 21. After removing duplicate records, 2 reviewers independently examined the titles and abstracts of each study. Studies not reporting AI-based predictive models were discarded. Subsequently, a thorough full-text assessment was conducted for all studies that initially met the criteria, and the reasons for excluding each study were recorded in detail. In the predesigned Excel spreadsheet, data was extracted from studies that qualified based on the inclusion criteria. The extracted information included: characteristics of the study (authors, country, publication year, study design, and sample size), characteristics of the participants (diagnostic criteria for GDM and number of GDM cases), intervention features (model development process, types of AI algorithms used, methods for handling missing data, predictors, and model validation), and study outcomes (assessment of model accuracy). In cases where the information presented in the literature was ambiguous, the researchers would proactively contact the corresponding author to acquire the relevant information. The aforementioned process was independently conducted by 2 authors. Any discrepancies were discussed and resolved with a third author.

Study Risk-of-Bias Assessment

The Prediction Model Risk of Bias Assessment Tool (PROBAST) was used to assess the risk of bias (ROB) for each study. PROBAST consisted of four domains: participants, predictors, outcomes, and analysis [26]. Based on the responses to the items provided in the PROBAST checklist, a ROB rating (high, low, or unclear) was assigned to each domain. The criteria for assessment were detailed below: (1) the overall ROB was deemed "low" when all domains were classified as "low risk"; (2) the overall ROB was considered "high" if any domain was identified as "high risk"; (3) the overall ROB was determined to be "unclear" when there was at least one domain with an "unclear" rating, while the other domains were classified as "low risk" [26]. The quality assessment was conducted by the same 2 authors who performed the study selection and data extraction. Any disagreements between the 2 authors were resolved through consultation with a third author.

Effect Measures and Synthesis Methods

Statistical analyses were performed using Stata (version 17.0; StataCorp LLC), R (version 4.2.0; R Development Core Team), and Meta DiSc (version 1.4; Clinical Biostatistics Unit) software. A bivariate mixed-effects model was used to pool sensitivity and specificity, generate a summary receiver operating characteristic (SROC) curve, and calculate area under

XSL•FO
RenderX

the curve (AUC). The Hartung-Knapp-Sidik-Jonkman method was further used to adjust the pooled estimates. All results were reported with 95% CI values. Between-study standard deviation (τ) and variance (τ^2) were extracted from the bivariate model to quantify absolute heterogeneity. And prediction intervals (PIs) were subsequently computed to estimate the range within which the true sensitivity or specificity of a future study was expected to lie, providing a clinically interpretable measure of real-world dispersion. Moreover, the Fagan nomogram was used to explore the relationship between pretest probability, likelihood ratios (LR), and post-test probability. The LR dot plot, divided into 4 quadrants based on the strength of evidence threshold, was used to determine the exclusion and confirmation of the AI model. Additionally, a bivariate boxplot was drawn to detect heterogeneity caused by threshold effects. And subgroup analysis was used to compare the predictive capabilities of different AI algorithms in GDM prediction. In line with current recommendations for interpreting heterogeneity, we quantified real-world dispersion primarily using the τ, τ^2, and calculated PIs as the key measure of practical uncertainty [27]. The I^2 statistic was considered but not emphasized, given its limited use informing the generalizability of findings compared to PIs [27]. Based on the clinical and methodological characteristics anticipated to cause heterogeneity across studies, a meta-regression analysis was used to explore and explain such heterogeneity. It aimed to uncover potential influencing factors and analyze which variables might account for variations in the effect sizes. And the Deek test was used to evaluate small-study effects among the included studies, with $P<.05$ indicating funnel-plot asymmetry.

Ethical Considerations

This systematic review and meta-analysis was conducted exclusively with published aggregate data. No individual-level or identifiable participant information was involved. Therefore, informed consent, institutional review board approval, privacy protection, and participant compensation were not applicable.

Results

Study Selection and Characteristics of Included Studies

A total of 2790 studies were retrieved from the database. After removing duplicates, the titles and abstracts of 1455 studies were reviewed, and the full texts of 116 studies were screened. Finally, 22 studies were included in this study, with 8 studies [14,15,28-35] being included in the systematic review and 14 studies being incorporated into the meta-analysis [15,16,21,22,28,36-44]. The detailed process of the literature screening is illustrated in Figure 1. The fundamental characteristics of the included studies are presented in Table 2. The included studies were conducted in 11 countries, with 12 being single-center studies [14,16,21,28,30-32,36,37,40,41,43], 10 being multicenter studies [15,22,29,33-35,38,39,42,44], 14 being retrospective studies [14,16,21,22,28-30,32,36,37,39-41,43], and 8 being prospective studies [15,31,33-35,38,42,44]. All 22 studies used ML algorithms, and 2 of them further used DL algorithms [16,42]. To evaluate the predictive performance of the models, 12 studies conducted internal validation [14,15,21,22,31,32,34,37,38,40-42], and 8 studies performed external validation [16,28-30,32,37,39,42]. Multimedia Appendix 2 provides a detailed record of the model performance parameters for each study.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) flow diagram for study selection. This figure illustrates the process of identifying, screening, and selecting studies for inclusion in the systematic review, showing the number of records at each stage and reasons for exclusions. AI: artificial intelligence; GDM: gestational diabetes mellitus.

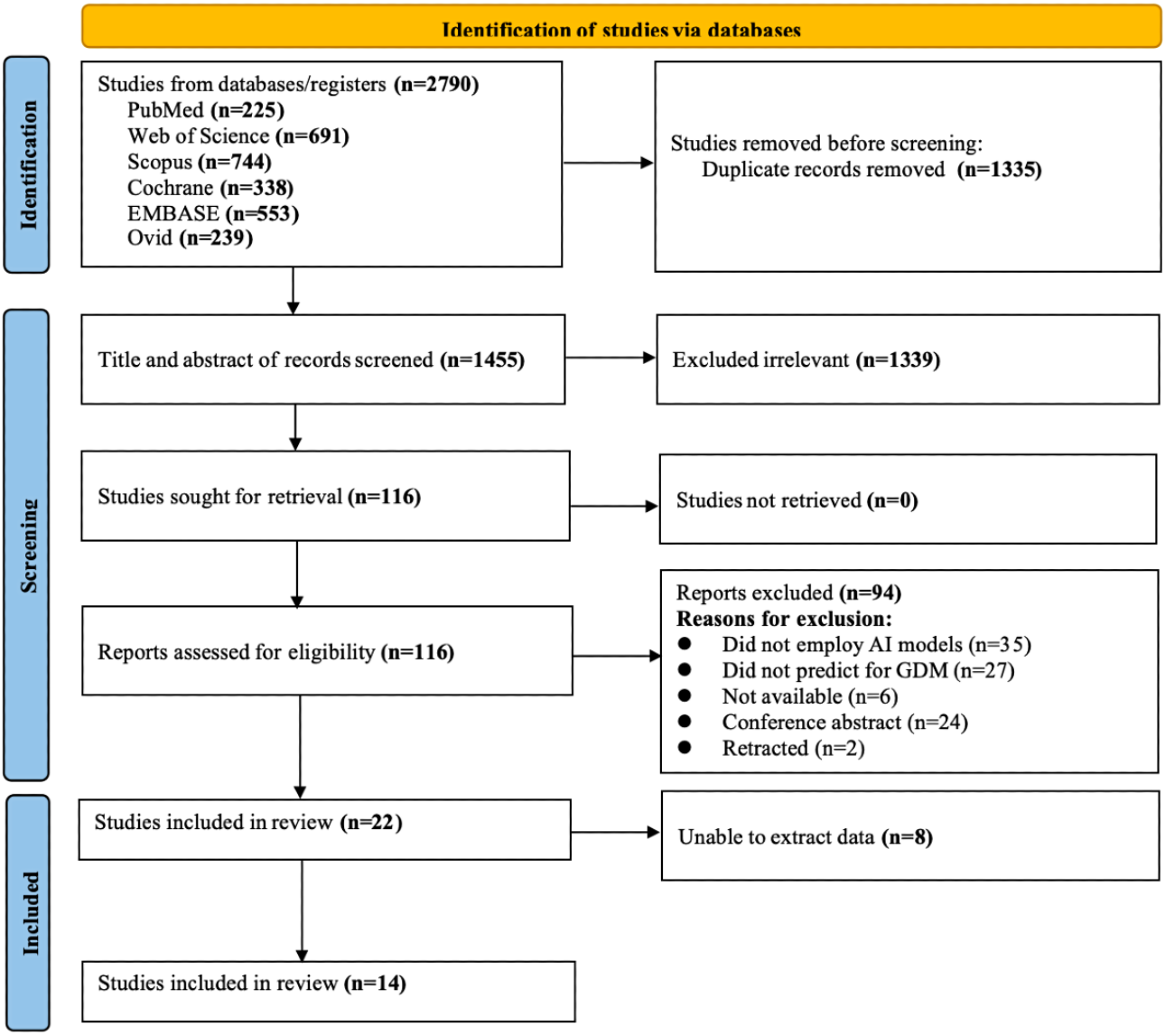

XSL•FO
RenderX

Table . General characteristics of the 22 studies included in the systematic review of artificial intelligence models for gestational diabetes mellitus prediction.

Study	Country	Study type	Single-center or multicenter	Diagnostic criteria	Sample size	Type of model
Belsti et al (2023) [22]	Australia	Retrospective	Multicenter	IADPSG[a]	48,502	ML[b]
Ali et al (2022) [33]	United Arab Emirates	Prospective	Multicenter	IADPSG	3858	ML
Wu et al (2021) [16]	China	Retrospective	Single-center	IADPSG	32,190	ML and DL[c]
Lin and Fang (2023) [36]	China	Retrospective	Single-center	IADPSG	406	ML
Ye et al (2020) [37]	China	Retrospective	Single-center	IADPSG	22,242	ML
Wang et al (2022) [30]	China	Retrospective	Single-center	IADPSG	1075	ML
Wu et al (2021) [28]	China	Retrospective	Single-center	IADPSG	17,005	ML
Wang et al (2021) [38]	China	Prospective	Multicenter	IADPSG	1139	ML
Syngelaki et al (2025) [31]	England	Prospective	Single-center	NICE[d]	41,587	ML
Donovan et al (2019) [39]	America	Retrospective	Multicenter	NIH[e]	11,56,708	ML
Kaya et al (2024) [40]	Turkey	Retrospective	Single-center	IADPSG	97	ML
Hu et al (2023) [41]	China	Retrospective	Single-center	IADPSG	735	ML
Liu et al (2022) [34]	China	Prospective	Multicenter	IADPSG	6848	ML
Lee et al (2021) [42]	Korea	Prospective	Multicenter	NIH	1443	ML and DL
Kumar et al (2022) [35]	Singapore	Prospective	Multicenter	IADPSG	222	ML
Bigdeli et al (2025) [14]	Iran	Retrospective	Single-center	NIH	743	ML
Kurt et al (2023) [15]	Turkey	Prospective	Multicenter	IADPSG	489	DL
Cubillos et al (2023) [21]	Chile	Retrospective	Single-center	IADPSG	1611	ML
Ding et al (2024) [43]	China	Retrospective	Single-center	IADPSG	554	ML
Kang et al (2023) [29]	Korea	Retrospective	Multicenter	NIH	34,387	ML
Zhao et al (2025) [32]	China	Retrospective	Single-center	IADPSG	1,03,172	ML
Liu et al (2020) [44]	China	Prospective	Multicenter	IADPSG	19,331	ML

XSL•FO
RenderX

[a]IADPSG: International Association of Diabetes and Pregnancy Study Groups.

[b]ML: machine learning.

[c]DL: deep learning.

[d]NICE: National Institute for Health and Care Excellence.

[e]NIH: National Institutes of Health.

ROB in Studies

Based on the PROBAST checklist, each study was assessed in terms of participants, predictors, outcomes, and analysis (Figure 2). The majority of studies consistently demonstrated low overall ROB and high applicability, indicating reliable methodology. However, in terms of overall ROB, 5 studies were rated as "unclear" [15,33,35,37,40]. One study was identified as having "high risk" in overall applicability due to insufficiently detailed descriptions of predictors used in model development [15]. Additionally, within the analysis domain, 2 studies were rated as "unclear" due to relatively small sample size, and this might also be one of the potential sources of bias [35,40]. In summary, most studies exhibited strong methodological quality and applicability. The detailed quality assessment of the included studies is detailed in Multimedia Appendix 3.

Figure 2. Risk assessment of the included models. This graph summarizes the methodological quality of the included prediction models, categorizing ROB across key domains to help readers assess the reliability of the evidence. ROB: risk of bias

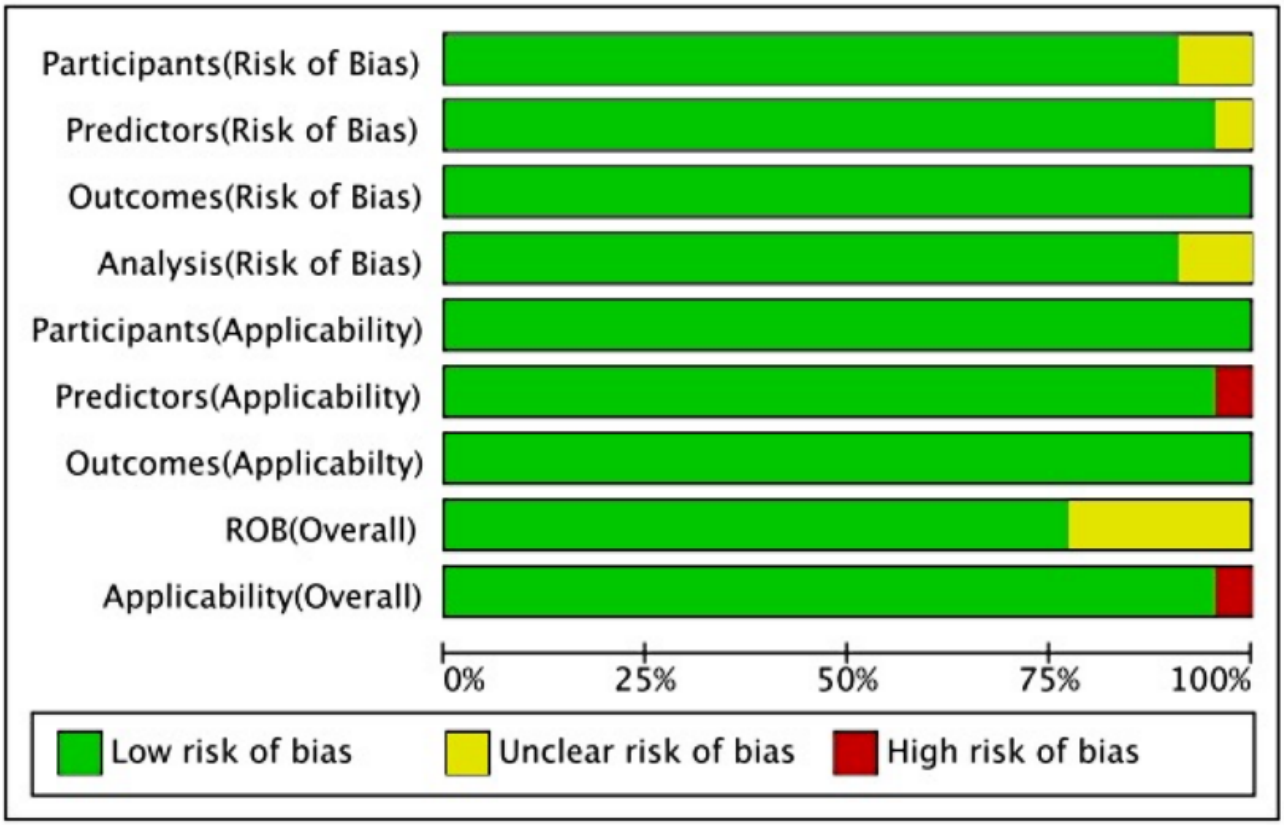

Performance of AI Algorithms for GDM

A total of 14 studies conducted on independent patient populations were included with the aim of evaluating the predictive value of AI algorithms for GDM [15,16,21,22,28,29,36-43]. Since some studies used multiple AI algorithms to construct several prediction models, this systematic review selected the model with the best performance reported in each study for meta-analysis. The pooled sensitivity was 0.78 (95% CI 0.69 - 0.86; τ=0.15, τ^2=0.02; PI 0.47 - 1.09), and specificity was 0.85 (95% CI 0.78 - 0.92; τ=0.11, τ^2=0.01; PI 0.59 - 1.11) after adjustment for the Hartung-Knapp-Sidik-Jonkman method (Figure 3). The wide PIs indicated substantial heterogeneity in real-world performance across populations, supporting the recommendation for local validation in the target population before clinical deployment. Note that the upper bounds of the PIs exceeded 1.0, specifically reaching 1.09 and 1.11. This occurred as a result of back-transformation from the logit scale and was a recognized statistical artifact, which did not indicate actual predictive performance greater than 100%.

As depicted in Figure 4, the SROC curve revealed the AUC of 0.94 (95% CI 0.92 - 0.96) for AI algorithms predicting GDM, suggesting a strong predictive capability. Furthermore, we set the pretest probability at 20% based on the pretest probability of the disease. At this level, when patients were predicted to have GDM by the AI algorithms, the true positive rate was 79%, and when the prediction was not GDM, the false negative rate was 4% (Figure 5). Moreover, the model demonstrated a positive LR of 15 and a negative LR of 0.17 (Figure 5). However, the

XSL•FO
RenderX

summary LR plot for the AI algorithms was located in the upper right quadrant (positive LR>10 and negative LR>0.1: confirmation only), and the individual studies were widely dispersed (Figure 6). The results indicated that while the prediction models built on AI algorithms generally demonstrated acceptable performance, they were not yet adequate for definitive diagnosis or exclusion of GDM. Additionally, there were notable variations in performance among the existing models.

XSL•FO
RenderX

Figure 3. Forest plots of sensitivity and specificity in 14 included studies on using artificial intelligence algorithms for predicting gestational diabetes mellitus. Each horizontal line represents the performance estimate of an individual study, with the diamond indicating the pooled result. The wide variability across studies highlights substantial heterogeneity in model performance [15,16,21,22,28,36-44]. DNN: deep neural network; GBDT: gradient-boosting decision tree; LR: logistic regression; RF: random forest; RNN-LSTM: recurrent neural network-long short-term memory; SVM: support vector machine; XGBoost: extreme-gradient boosting.

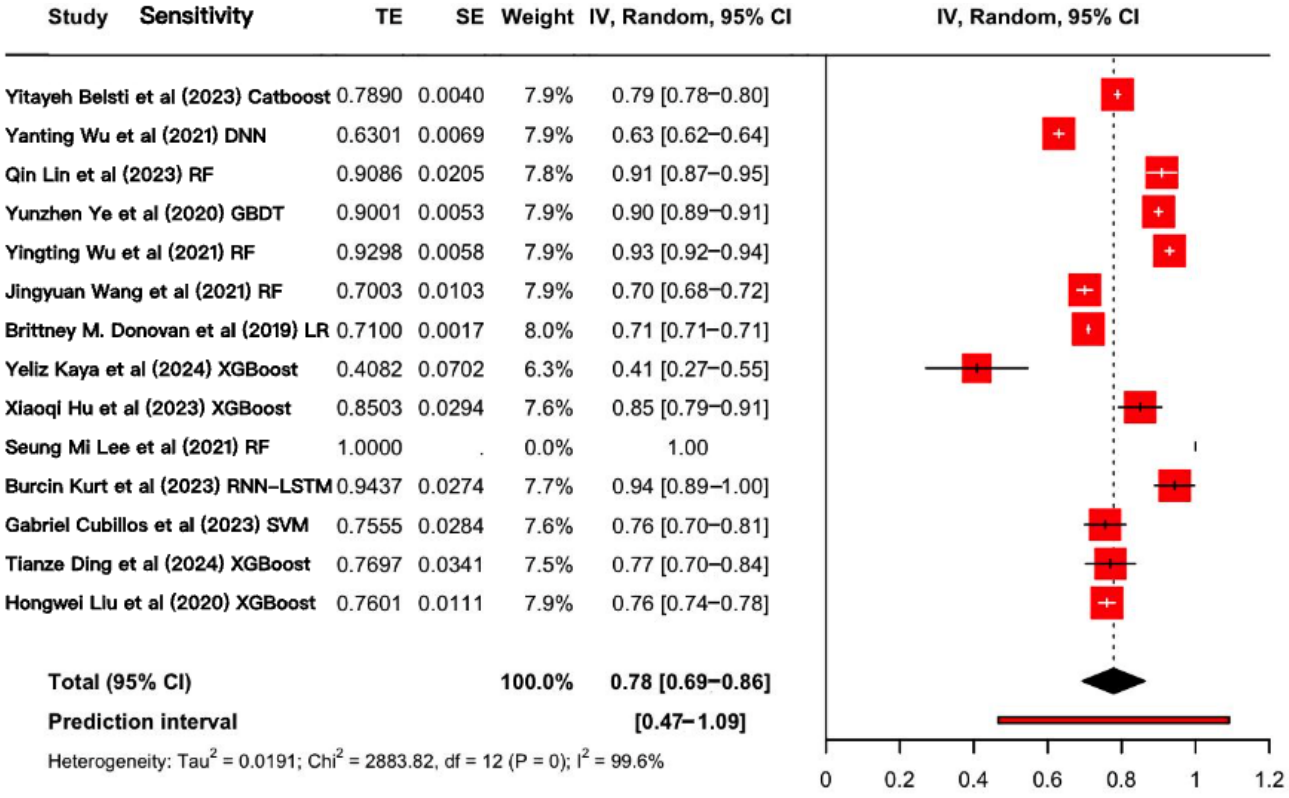

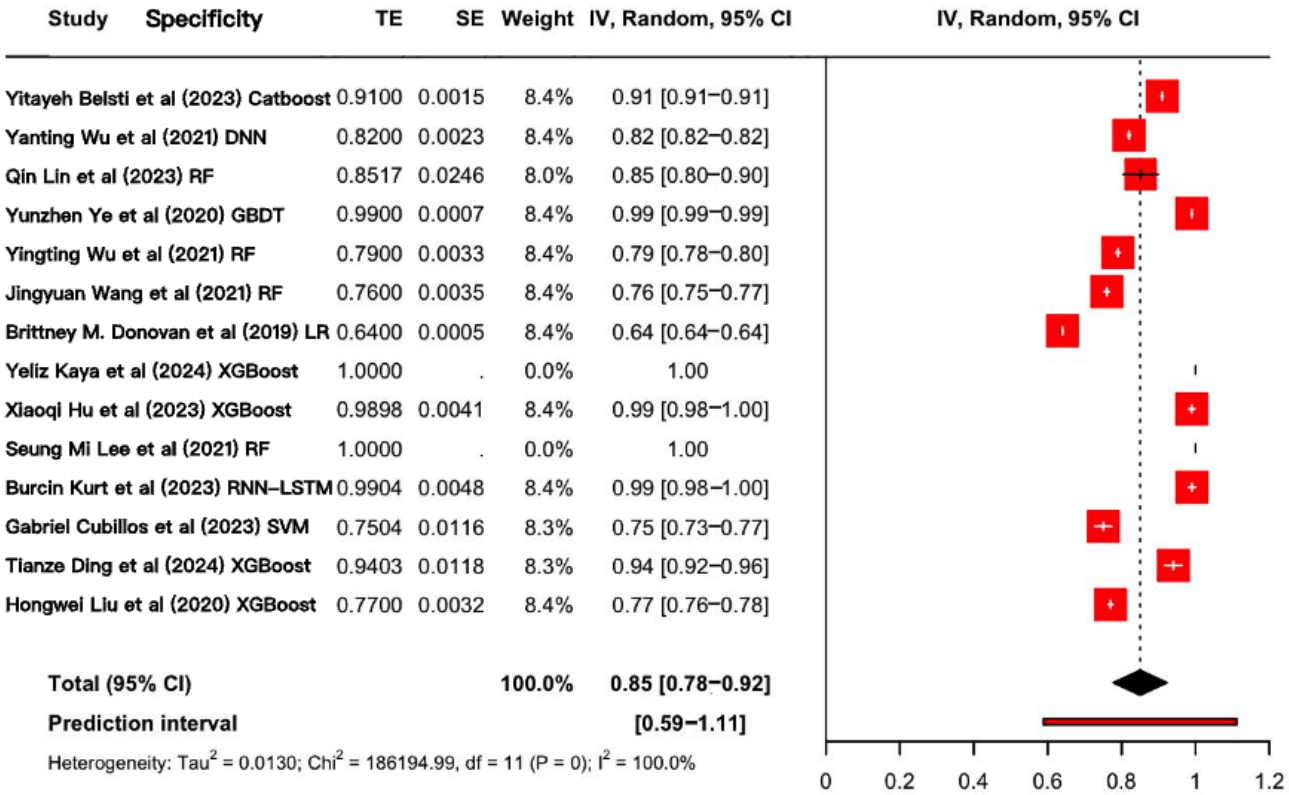

XSL•FO
RenderX

Figure 4. SROCs of included studies. This plot shows the overall diagnostic accuracy of artificial intelligence algorithms, with the curve position indicating the trade-off between sensitivity and specificity across different thresholds. The high AUC (0.87) reflects strong average discriminatory power. SROC: summary receiver operating characteristic curve; AUC: area under the curve.

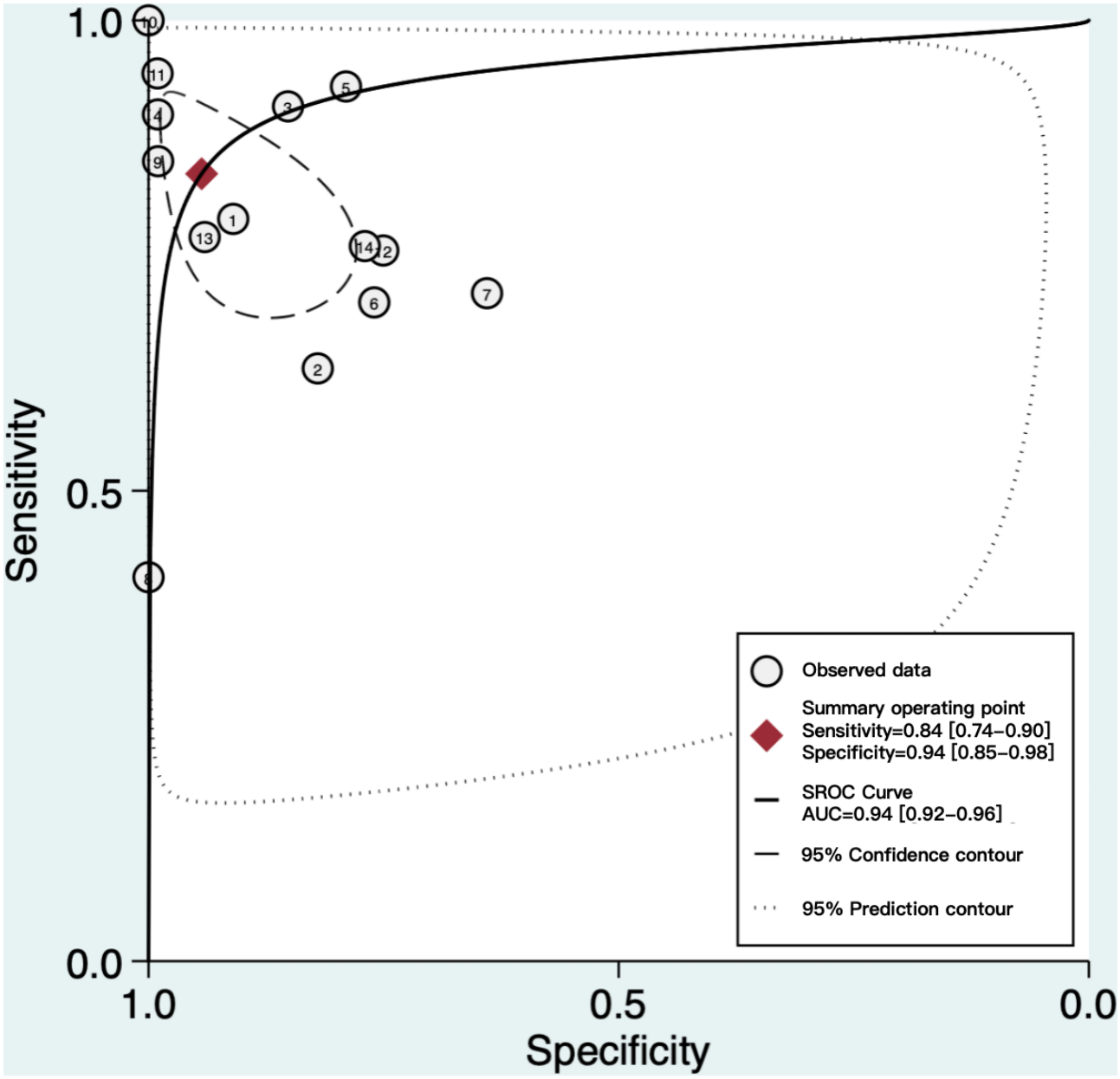

XSL•FO
RenderX

Figure 5. Fagan nomogram of artificial intelligence (AI) algorithms for predicting gestational diabetes mellitus. The first column of this nomogram represents the pretest probability, the second column represents the likelihood ratio, and the third shows the posttest probability. Interpretation: This tool helps clinicians estimate how a positive or negative test result changes the probability of gestational diabetes mellitus. The limited shift from pre to posttest probability indicates that current AI models provide only modest diagnostic value in clinical practice.

XSL•FO
RenderX

Figure 6. Likelihood ratio dot plot of artificial intelligence algorithms for predicting gestational diabetes mellitus. The position of the summary point in the upper right quadrant indicates that current artificial intelligence algorithms have confirmation but limited exclusion ability (positive likelihood ratio>10 and negative likelihood ratio>0.1), supporting their role as screening adjuncts rather than definitive diagnostic tools. LRN: likelihood ratio for a negative test; LRP: likelihood ratio for a positive test.

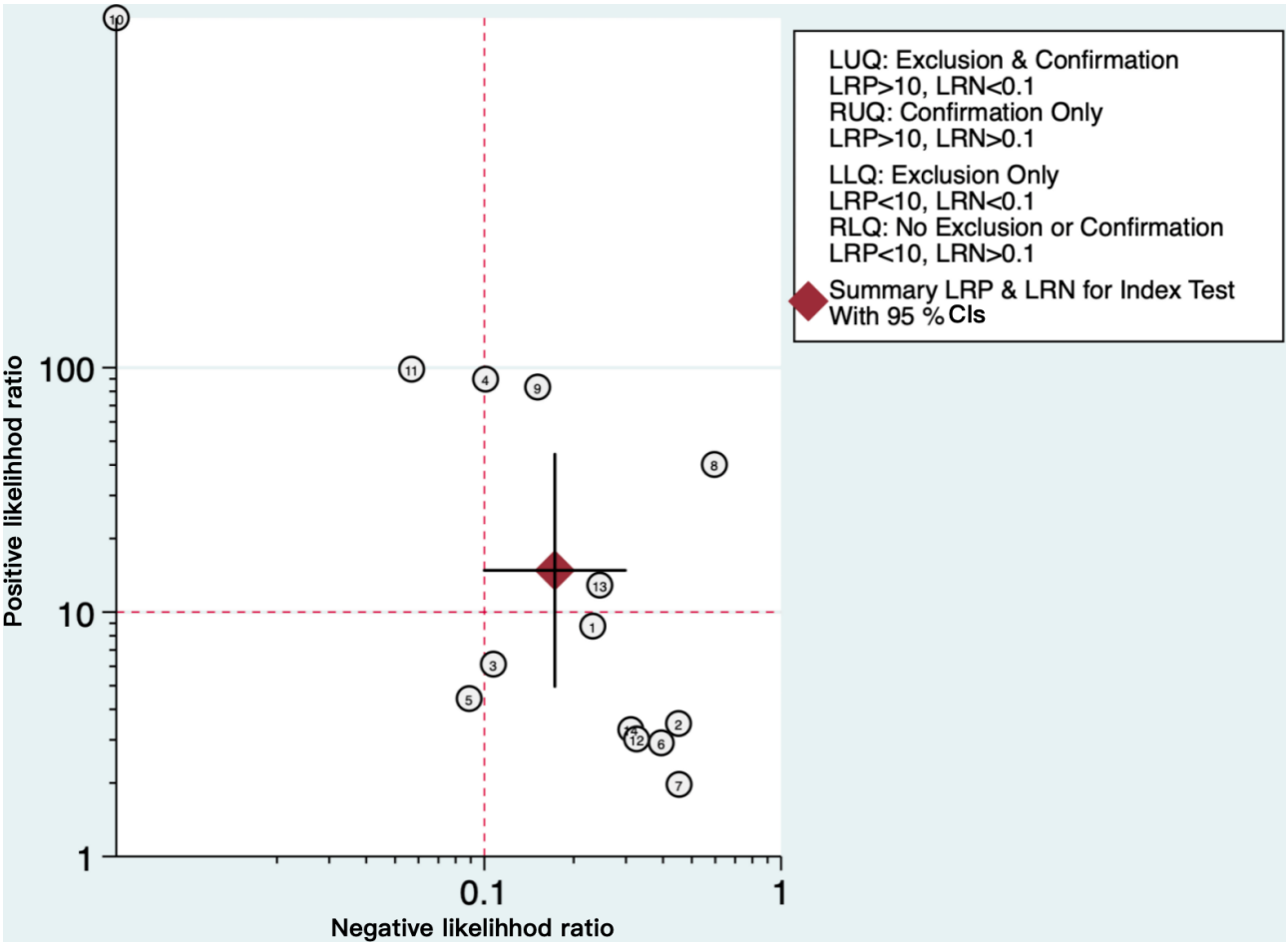

Predictors

From the models included in this systematic review, all reported predictors were systematically extracted and cataloged. The selection of key predictors for presentation and further analysis was based on three principal criteria: (1) clinical and pathophysiological relevance to GDM development, as established in prior literature and clinical guidelines; (2) frequency of reporting across the included studies, ensuring the findings were representative of common modeling practices; and (3) feasibility of meta-analytic synthesis, prioritizing variables with consistent definitions and measurements.

The consistently reported and clinically salient predictors were age, pre-pregnancy body mass index, first-trimester fast blood glucose, family history of diabetes, parity, gravidity, and history of GDM. These factors were well-recognized risk determinants in existing GDM etiological research and screening protocols. Detailed information is provided in Multimedia Appendix 4.

Assessment of Small-Study Effects

Deek test (P=.03) and the funnel plot (Figure 7) both showed clear asymmetry, suggesting the presence of small-study effects. This asymmetry might stem from publication bias, selective reporting, and methodological differences among smaller studies.

XSL•FO
RenderX

Figure 7. Deek funnel plot asymmetry test of small-study effects. The asymmetric distribution of studies suggests potential publication bias, where smaller studies reporting higher accuracy may be overrepresented in the literature.

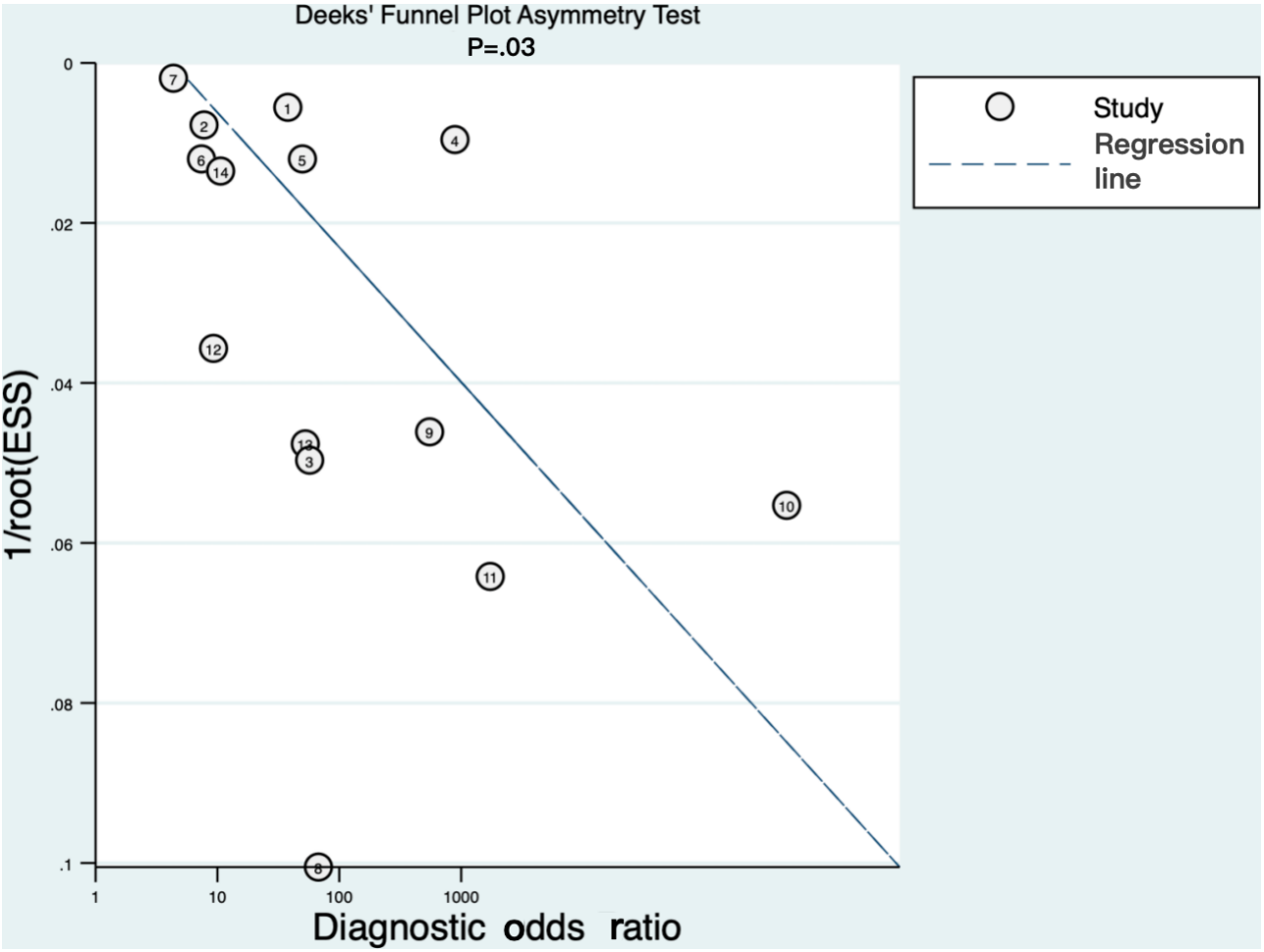

Heterogeneity Analysis

Threshold Effect Analysis

Bivariate boxplot (Figure 8) showed a positive correlation between the sensitivity and specificity of the included studies, indicating the absence of a threshold effect among the studies included in this systematic review. Moreover, some individual studies fell outside the shaded area, indicating the potential presence of heterogeneity.

XSL•FO
RenderX

Figure 8. Bivariate boxplot of threshold effect analysis. This plot evaluates whether heterogeneity in results can be explained by differences in diagnostic thresholds used across studies, with dispersed points indicating substantial variability beyond what threshold effects alone can explain.

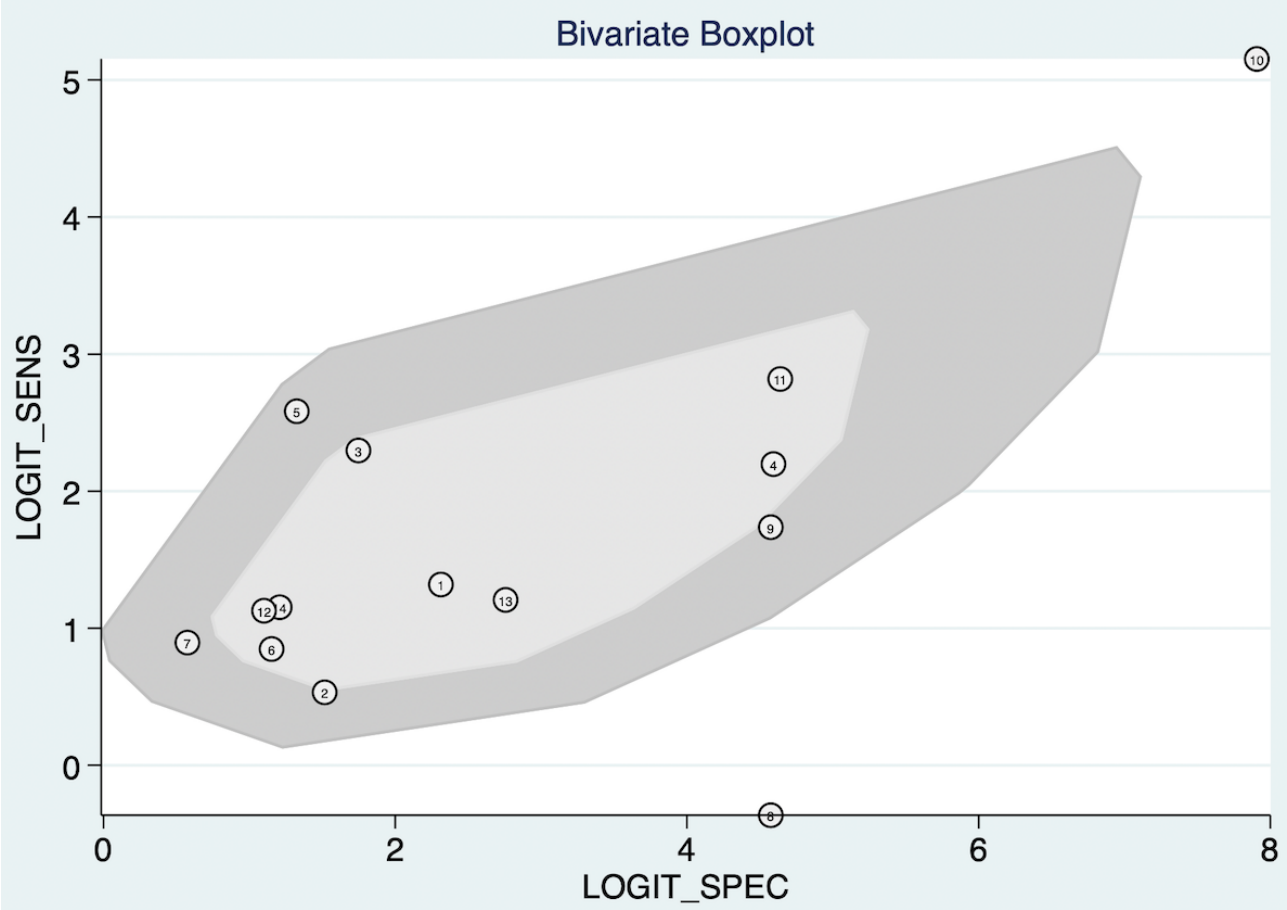

Subgroup Analysis

To evaluate the performance of prediction models constructed using various algorithms, subgroup analyses were performed on models that had been used in at least 3 studies, after first excluding 2 studies with extreme values caused by sparse data [40,42]. The performance of each algorithm was assessed using AUC, sensitivity, specificity, positive LR, negative LR, and diagnostic odds ratio (DOR). Details are presented in Table 3 and forest plots for sensitivity and specificity are shown in Figure 9. Among the subgroup models with sparse-data studies removed, the models using the RF algorithm exhibited the highest AUC, followed by those using the XGBoost algorithm, while the models using the logistic regression algorithm demonstrated the lowest AUC performance. Additionally, these models demonstrated varying performance across different metrics. The RF algorithm exhibited the highest sensitivity (0.83, 95% CI 0.74 - 0.93), while the XGBoost algorithm demonstrated the highest specificity (0.82, 95% CI 0.77 - 0.87) and DOR (49, 95% CI 11 - 211).

XSL•FO
RenderX

Table . Subgroup analysis of predictive performance across different artificial intelligence algorithms.

Models	Logistic regression	Random forest	XGBoost[a]	SVM[b]	*P* value
Number	8	4	4	4	—[c]
AUC[d]	0.75	0.87	0.86	0.78	<.001
Sensitivity (95% CI)	0.67 (0.62 - 0.72)	0.83 (0.74 - 0.93)	0.82 (0.79 - 0.85)	0.61 (0.36 - 0.86)	<.001
Specificity (95% CI)	0.72 (0.66 - 0.79)	0.80 (0.75 - 0.85)	0.82 (0.77 - 0.87)	0.80 (0.61 - 0.99)	.03
Positive LR[e] (95% CI)	2.8 (1.7 - 4.7)	4.5 (3.5 - 5.7)	10.1 (2.9 - 35.3)	4.2 (1.9 - 9.2)	<.001
Negative LR (95% CI)	0.42 (0.31 - 0.55)	0.17 (0.09 - 0.31)	0.21 (0.16 - 0.27)	0.45 (0.34 - 0.60)	<.001
DOR[f] (95% CI)	7 (3-15)	26 (12-58)	49 (11-211)	9 (5-17)	<.001

[a]XGBoost: extreme gradient boosting.

[b]SVM: support vector machine.

[c]Not applicable.

[d]AUC: area under the curve.

[e]LR: likelihood ratio.

[f]DOR: diagnostic odds ratio.

XSL•FO
RenderX

Figure 9. Forest plots of sensitivity and specificity in subgroup analysis. This forest plot presents pooled sensitivity and specificity estimates stratified by algorithm type (logistic regression, random forest, XGBoost, and SVM), allowing visual assessment of performance variability across model subgroups. The width of CI values reflects the precision of each estimate, while consistent point estimates across studies within a subgroup indicate algorithm-specific stability in diagnostic performance [16,21,22,28,36,38,39,41-44]. DL: deep learning; SVM: support vector machine; XGBoost: extreme gradient boosting.

XSL•FO
RenderX

Meta-Regression Analysis

To further explore the potential sources of heterogeneity in the performance of prediction models, a meta-regression analysis was conducted by including the study design (whether the study was conducted in Asia), study type (whether it was prospective), study design (whether it was multicenter), sample size (whether it exceeded 1000), GDM diagnostic criteria (whether it was based on IADPSG), and the timing of model use (whether it was in first trimester). Through meta-regression, we identified sources of heterogeneity among studies and evaluated their impact on diagnostic outcomes. The results indicated that study type significantly influenced heterogeneity among studies, with a trend toward increased predictive accuracy in prospective study designs (regression coefficient=2.289; *P*=.001). And the sample size had a substantial impact on the heterogeneity across studies, with predictive accuracy declining as the sample size increased (regression coefficient=−2.535; *P*=.001; Figure 10). This might reflect overfitting in small single-center datasets and greater clinical heterogeneity in large multicenter cohorts. Moreover, given the disparities among regions, the study area also served as one of the potential sources of heterogeneity (regression coefficient=−2.139, *P*=.002). The detailed procedures of the meta-regression are provided in Multimedia Appendix 5.

Figure 10. Bubble plot of meta-regression examining the association between sample size and predictive accuracy. This bubble plot visualizes the relationship between study sample size (log-transformed) and predictive accuracy (log-transformed diagnostic odds ratio) across all included studies. Each circle represents an individual study. The fitted regression line demonstrates a significant negative association, indicating that larger sample sizes tend to be associated with lower diagnostic accuracy. The plot provides an intuitive graphical confirmation of the quantitative meta-regression results, highlighting sample size as an important source of heterogeneity in model performance. log(DOR) values were truncated at ±3 for extreme cases (fp=0 or fn=0).

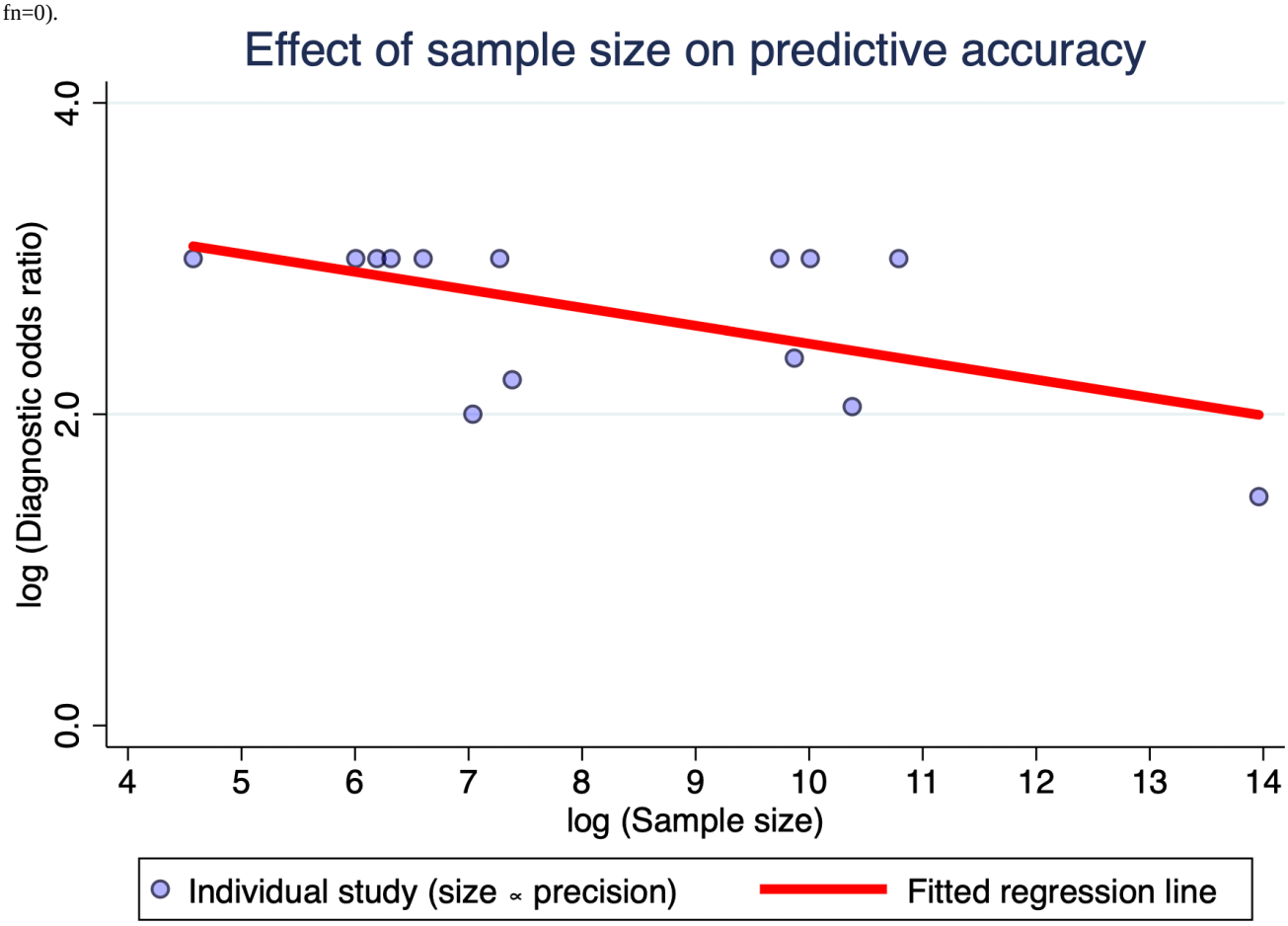

Discussion

Overview

This systematic review and meta-analysis aimed to evaluate the predictive performance of AI algorithms for GDM, compare the efficacy of different algorithms, and determine the key performance determinants. The pooled analysis revealed that AI-based models exhibited robust predictive capability for GDM prediction. However, the wide PIs revealed substantial performance heterogeneity in real-world applications, urging cautious interpretation of the currently summarized high-performance metrics. In addition, the summary LR plot and Fagan nomogram analyses indicated that existing models were insufficient to independently confirm or exclude GDM, so their present role should be positioned as an adjunct screening tool.

Consistent with the mainstream research trend, this systematic review further confirmed the dominant role of the RF algorithm in predicting GDM, which corroborated the findings of prior systematic reviews that highlighted ensemble methods for their robustness [14,45]. However, our analysis moved beyond merely confirming superiority by quantifying its extent and contrasting

XSL•FO
RenderX

it with other algorithms. Specifically, the RF algorithm performed the best in key metrics such as AUC and sensitivity, mainly because it handles the complex, non-linear relationships inherent in GDM prediction more effectively than linear models [46,47]. This was particularly relevant in clinical settings where data could be incomplete; the inherent ability of RF to handle missing values gracefully contributed to its stronger robustness when dealing with the imperfect data often presented in routine care, which was a critical practical advantage for implementation in real-world settings [47]. In contrast, the XGBoost algorithm demonstrated higher specificity, probably benefiting from its built-in regularization and feature-importance ranking, which made it more proficient at identifying true-negative cases [48,49]. It was worth noting that the 95% CI for the DOR of XGBoost was wide, reflecting marked between-study differences in sample size, event rate, or clinical heterogeneity and indicating that its actual diagnostic consistency was highly dependent on specific population characteristics and implementation settings. Notably, this systematic review identified and emphasized methodological heterogeneity as a key driver of performance disparities. Inconsistencies across studies in data preprocessing (eg, handling of missing values and feature scaling), validation strategies (eg, data split ratios and internal validation methods), and performance reporting standards significantly hindered the comparability and integrability of research outcomes. Therefore, while pursuing superior algorithms, future studies should prioritize the establishment and adherence to methodological reporting standards for the development and validation of AI-based prediction models.

To further elucidate the real-world implications of our findings, our meta-regression analysis identified several influential factors related to variations in model accuracy, providing a more nuanced understanding than simple performance pooling. Specifically, we found that prospective study design was associated with significantly higher predictive accuracy. This might be attributed to more standardized data collection procedures and better control of confounders in prospective settings, whereas retrospective studies often relied on preexisting eHealth record data, which could be heterogeneous and incomplete [50-52]. These findings aligned with the results reported by Liu et al [53], who reported that AI-based models in prospective cohort studies achieved AUC values 4% - 7% higher than those from retrospective studies. This consistency across different analyses strengthened the argument for prioritizing prospective validation designs. Additionally, we observed that studies with larger sample sizes tended to report lower accuracy estimates. This counterintuitive finding was crucial, as it likely reflected greater demographic and clinical diversity in larger cohorts, thereby reducing overfitting and offering a more realistic, generalizable performance assessment than optimistic estimates from small, homogeneous samples. This underscored that larger, more diverse studies provided a more trustworthy evidence base for clinical deployment. Similarly, studies conducted in certain geographic regions also showed systematically lower accuracy, possibly due to regional differences in diagnostic criteria, risk factor prevalence, or health care infrastructure. These findings indicated that the performance of a model depended not only on the algorithm itself but was also profoundly shaped by the environment in which it was developed and validated. This had direct implications for implementation: a model successful in one region might not translate directly to another without adaptation and local validation.

Despite the strong performance of some algorithms, AI models still faced critical barriers to clinical deployment that should be addressed to realize their potential [54]. These included the "black-box" nature leading to limited interpretability, a persistent lack of large-scale external validation in diverse populations, and the absence of standardized interfaces for integration with existing clinical workflows—especially eHealth record systems [55,56]. To overcome these barriers, future efforts should adopt a multifaceted implementation-science approach. This entails: (1) prioritizing prospective, multicenter validation studies to generate high-grade, generalizable evidence; and (2) incorporating explainable AI techniques to enhance model interpretability and foster clinician confidence. Ultimately, realizing the full potential of AI in GDM prediction requires a concerted shift from merely developing accurate algorithms to engineering clinically viable, trustworthy, and deployable solutions.

Limitations

However, several limitations exist in this systematic review and meta-analysis. First, most included studies and citations focused on East Asian populations, which might limit the generalizability of our findings to multi-ethnic or low-resource settings. External validation in diverse cohorts from Europe, North America, and Africa should therefore be needed to assess global applicability and to examine performance after feature-set simplification. Second, owing to limited application frequency, several emerging algorithms such as artificial neural networks and DL were not included in the subgroup analysis. Future studies should pay attention to the development of these emerging algorithms, verify their performance through more empirical studies, and explore their unique value in GDM prediction. Third, the Deek funnel plot asymmetry test indicated potential publication bias, suggesting that studies reporting higher performance metrics might be overrepresented. This could inflate the pooled estimates and limit generalizability. Future studies should consider preregistering protocols and sharing analysis code and datasets to improve reproducibility and reduce selective reporting.

Conclusions

This systematic review and meta-analysis confirmed the strong discriminative performance of AI models for GDM prediction. However, substantial heterogeneity, publication bias, and small-study effects currently limited their readiness for direct clinical deployment. Unlike previous narrative reviews, this study innovatively provided the first direct comparative and quantitative synthesis of multiple AI algorithms in this field. This approach filled a critical gap in existing literature by offering an evidence-based framework to guide algorithm selection, rather than merely summarizing performance metrics. The key implication for real-world application was the demonstrated need for local validation in target populations before implementation. To translate this potential into practice,

XSL•FO
RenderX

future studies must prioritize prospective, multicenter, large-scale external validations. The ultimate goal was to develop AI tools that were not only accurate but also interpretable and seamlessly integrable into clinical workflows, thereby enabling reliable AI-driven early prediction and management of GDM.

Acknowledgments

This manuscript did not use generative artificial intelligence for content creation, data analysis, or study design.

Funding

This study was supported by the National Natural Science Foundation of China (32070189), Postgraduate Scientific Research Innovation Project of Hunan Province (CX20251466), and Hunan Province Health Commission Scientific Research Project (D202314038701). The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the article.

Data Availability

All data analyzed in this study are included in this published article and its supplementary files.

Authors' Contributions

Conceptualization: YL (lead), YS (equal), and ZL (equal)
Data curation: AD (lead) and ML (supporting)
Formal analysis: YL (lead), AD (supporting), and ML (supporting)
Funding acquisition: YL, ZL, and ML
Investigation: ML (lead) and ZZ (supporting)
Methodology: YL (lead), AD (supporting), and ML (supporting)
Project administration: YL (lead) and JS (supporting)
Software: YL
Supervision: ZL (lead) and YS (supporting)
Validation: SJ (equal) and ZZ (equal)
Visualization: YL
Writing – original draft: YL (lead), AD (equal), and ML (equal)
Writing – review and editing: ZL (lead) and YS (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details of search strategy and result.
[DOCX File, 22 KB - jmir_v28i1e79729_app1.docx]

Multimedia Appendix 2
Details of model performance parameters for each study.
[DOCX File, 26 KB - jmir_v28i1e79729_app2.docx]

Multimedia Appendix 3
Result of Prediction Model Risk of Bias Assessment Tool assessment.
[DOCX File, 417 KB - jmir_v28i1e79729_app3.docx]

Multimedia Appendix 4
Predictive factors.
[DOCX File, 16 KB - jmir_v28i1e79729_app4.docx]

Multimedia Appendix 5
Result of meta-regression.
[DOCX File, 16 KB - jmir_v28i1e79729_app5.docx]

XSL•FO
RenderX

Checklist 1
PRISMA checklists.
[DOCX File, 61 KB - jmir_v28i1e79729_app6.docx]

References

1. Sweeting A, Wong J, Murphy HR, Ross GP. A clinical update on gestational diabetes mellitus. Endocr Rev 2022 Sep 26;43(5):763-793. [doi: 10.1210/endrev/bnac003] [Medline: 35041752]
2. Yang W, Liu J, Li J, et al. Interactive effects of prepregnancy overweight and gestational diabetes on macrosomia and large for gestational age: a population-based prospective cohort in Tianjin, China. Diabetes Res Clin Pract 2019 Aug;154:82-89. [doi: 10.1016/j.diabres.2019.06.014] [Medline: 31271809]
3. Gao C, Sun X, Lu L, Liu F, Yuan J. Prevalence of gestational diabetes mellitus in mainland China: a systematic review and meta-analysis. J Diabetes Investig 2019 Jan;10(1):154-162. [doi: 10.1111/jdi.12854] [Medline: 29683557]
4. Duong TL, Shahunja KM, Le M, McIntyre DH, Ward J, Mamun AA. Gestational diabetes mellitus and its impact on maternal and neonatal outcomes in Indigenous populations: a systematic review and meta-analysis. Diabetes Res Clin Pract 2025 Nov;229:112462. [doi: 10.1016/j.diabres.2025.112462] [Medline: 40947022]
5. Yang F, Liu H, Ding C. Gestational diabetes mellitus and risk of neonatal respiratory distress syndrome: a systematic review and meta-analysis. Diabetol Metab Syndr 2024 Dec 5;16(1):294. [doi: 10.1186/s13098-024-01539-x] [Medline: 39639383]
6. Zhang Y, Chen L, Ouyang Y, et al. A new classification method for gestational diabetes mellitus: a study on the relationship between abnormal blood glucose values at different time points in oral glucose tolerance test and adverse maternal and neonatal outcomes in pregnant women with gestational diabetes mellitus. AJOG Glob Rep 2024 Nov;4(4):100390. [doi: 10.1016/j.xagr.2024.100390] [Medline: 39309607]
7. Rizzo HE, Escaname EN, Alana NB, et al. Maternal diabetes and obesity influence the fetal epigenome in a largely Hispanic population. Clin Epigenetics 2020 Feb 19;12(1):34. [doi: 10.1186/s13148-020-0824-9] [Medline: 32075680]
8. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021 Jul;8(2):e188-e194. [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
9. Helm JM, Swiergosz AM, Haeberle HS, et al. Machine learning and artificial intelligence: definitions, applications, and future directions. Curr Rev Musculoskelet Med 2020 Feb;13(1):69-76. [doi: 10.1007/s12178-020-09600-8] [Medline: 31983042]
10. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol 2020 Feb 27;9(2):14. [doi: 10.1167/tvst.9.2.14] [Medline: 32704420]
11. Segar MW, Jaeger BC, Patel KV, et al. Development and validation of machine learning-based aace-specific Mmodels to predict 10-year risk of heart failure: a multicohort analysis. Circulation 2021 Jun 15;143(24):2370-2383. [doi: 10.1161/CIRCULATIONAHA.120.053134] [Medline: 33845593]
12. Luo D, Zhang J, Xie L, Liu C, Wang R. Early detection of positive urine culture in patients with urolithiasis: a machine learning model with dynamic online nomogram. Ann Med 2025 Dec;57(1):2550582. [doi: 10.1080/07853890.2025.2550582] [Medline: 40853723]
13. Ma L, Yang L, Wang Y, et al. Enhancing early gestational diabetes mellitus prediction with imputation-based machine learning framework: a comparative study on real-world clinical records. Digit Health 2025;11:20552076251352436. [doi: 10.1177/20552076251352436] [Medline: 40755962]
14. Bigdeli SK, Ghazisaedi M, Ayyoubzadeh SM, Hantoushzadeh S, Ahmadi M. Predicting gestational diabetes mellitus in the first trimester using machine learning algorithms: a cross-sectional study at a hospital fertility health center in Iran. BMC Med Inform Decis Mak 2025 Jan 3;25(1):3. [doi: 10.1186/s12911-024-02799-3] [Medline: 39754258]
15. Kurt B, Gürlek B, Keskin S, et al. Prediction of gestational diabetes using deep learning and Bayesian optimization and traditional machine learning techniques. Med Biol Eng Comput 2023 Jul;61(7):1649-1660. [doi: 10.1007/s11517-023-02800-7] [Medline: 36848010]
16. Wu YT, Zhang CJ, Mol BW, et al. Early prediction of gestational diabetes mellitus in the Chinese population via advanced machine learning. J Clin Endocrinol Metab 2021 Mar 8;106(3):e1191-e1205. [doi: 10.1210/clinem/dgaa899] [Medline: 33351102]
17. AlSaad R, Elhenidy A, Tabassum A, et al. Artificial intelligence in gestational diabetes care: a systematic review. J Diabetes Sci Technol 2025 Aug 25;25:19322968251355967. [doi: 10.1177/19322968251355967] [Medline: 40855734]
18. Kokori E, Olatunji G, Aderinto N, et al. The role of machine learning algorithms in detection of gestational diabetes; a narrative review of current evidence. Clin Diabetes Endocrinol 2024 Jun 25;10(1):18. [doi: 10.1186/s40842-024-00176-7] [Medline: 38915129]
19. Snell KI, Ensor J, Debray TP, Moons KG, Riley RD. Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures? Stat Methods Med Res 2018 Nov;27(11):3505-3522. [doi: 10.1177/0962280217705678] [Medline: 28480827]
20. Hah H, Goldin DS. How clinicians perceive artificial intelligence-assisted technologies in diagnostic decision making: mixed methods approach. J Med Internet Res 2021 Dec 16;23(12):e33540. [doi: 10.2196/33540] [Medline: 34924356]

XSL•FO
RenderX

21. Cubillos G, Monckeberg M, Plaza A, et al. Development of machine learning models to predict gestational diabetes risk in the first half of pregnancy. BMC Pregnancy Childbirth 2023 Jun 23;23(1):469. [doi: 10.1186/s12884-023-05766-4] [Medline: 37353749]
22. Belsti Y, Moran L, Du L, et al. Comparison of machine learning and conventional logistic regression-based prediction models for gestational diabetes in an ethnically diverse population; the Monash GDM Machine learning model. Int J Med Inform 2023 Nov;179:105228. [doi: 10.1016/j.ijmedinf.2023.105228] [Medline: 37774429]
23. Salameh JP, Bossuyt PM, McGrath TA, et al. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. BMJ 2020 Aug 14;370:m2632. [doi: 10.1136/bmj.m2632] [Medline: 32816740]
24. Rethlefsen ML, Page MJ. PRISMA 2020 and PRISMA-S: common questions on tracking records and the flow diagram. J Med Libr Assoc 2022 Apr 1;110(2):253-257. [doi: 10.5195/jmla.2022.1449] [Medline: 35440907]
25. Ivaldi D, Burgos M, Oltra G, Liquitay CE, Garegnani L. Adherence to PRISMA 2020 statement assessed through the expanded checklist in systematic reviews of interventions: a meta-epidemiological study. Cochrane Evid Synth Methods 2024 May;2(5):e12074. [doi: 10.1002/cesm.12074] [Medline: 40476264]
26. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019 Jan 1;170(1):W1-W33. [doi: 10.7326/M18-1377] [Medline: 30596876]
27. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
28. Wu Y, Ma S, Wang Y, et al. A risk prediction model of gestational diabetes mellitus before 16 gestational weeks in Chinese pregnant women. Diabetes Res Clin Pract 2021 Sep;179:109001. [doi: 10.1016/j.diabres.2021.109001] [Medline: 34390760]
29. Kang BS, Lee SU, Hong S, et al. Prediction of gestational diabetes mellitus in Asian women using machine learning algorithms. Sci Rep 2023 Aug 16;13(1):13356. [doi: 10.1038/s41598-023-39680-8] [Medline: 37587201]
30. Wang N, Guo H, Jing Y, et al. Development and validation of risk prediction models for gestational diabetes mellitus using four different methods. Metabolites 2022 Oct 29;12(11):1040. [doi: 10.3390/metabo12111040] [Medline: 36355123]
31. Syngelaki A, Wright A, Gomez Fernandez C, Mitsigiorgi R, Nicolaides KH. First-trimester prediction of gestational diabetes mellitus based on maternal risk factors. BJOG 2025 Jun;132(7):972-982. [doi: 10.1111/1471-0528.18110] [Medline: 40000426]
32. Zhao M, Su X, Huang L. Early gestational diabetes mellitus risk predictor using neural network with NearMiss. Gynecol Endocrinol 2025 Dec;41(1):2470317. [doi: 10.1080/09513590.2025.2470317] [Medline: 39992231]
33. Ali N, Khan W, Ahmad A, Masud MM, Adam H, Ahmed LA. Predictive modeling for the diagnosis of gestational diabetes mellitus using epidemiological data in the United Arab Emirates. Information 2022;13(10):485. [doi: 10.3390/info13100485]
34. Liu R, Zhan Y, Liu X, et al. Stacking ensemble method for gestational diabetes mellitus prediction in Chinese pregnant women: a prospective cohort study. J Healthc Eng 2022;2022:8948082. [doi: 10.1155/2022/8948082] [Medline: 36147870]
35. Kumar M, Ang LT, Png H, et al. Automated machine learning (AutoML)-derived preconception predictive risk model to guide early intervention for gestational diabetes mellitus. Int J Environ Res Public Health 2022 Jun 1;19(11):6792. [doi: 10.3390/ijerph19116792] [Medline: 35682375]
36. Lin Q, Fang ZJ. Establishment and evaluation of a risk prediction model for gestational diabetes mellitus. World J Diabetes 2023 Oct 15;14(10):1541-1550. [doi: 10.4239/wjd.v14.i10.1541] [Medline: 37970129]
37. Ye Y, Xiong Y, Zhou Q, Wu J, Li X, Xiao X. Comparison of machine learning methods and conventional logistic regressions for predicting gestational diabetes using routine clinical data: a retrospective cohort study. J Diabetes Res 2020;2020:4168340. [doi: 10.1155/2020/4168340] [Medline: 32626780]
38. Wang J, Lv B, Chen X, et al. An early model to predict the risk of gestational diabetes mellitus in the absence of blood examination indexes: application in primary health care centres. BMC Pregnancy Childbirth 2021 Dec 8;21(1):814. [doi: 10.1186/s12884-021-04295-2] [Medline: 34879850]
39. Donovan BM, Breheny PJ, Robinson JG, et al. Development and validation of a clinical model for preconception and early pregnancy risk prediction of gestational diabetes mellitus in nulliparous women. PLoS ONE 2019;14(4):e0215173. [doi: 10.1371/journal.pone.0215173] [Medline: 30978258]
40. Kaya Y, Bütün Z, Çelik Ö, Salik EA, Tahta T, Yavuz AA. The early prediction of gestational diabetes mellitus by machine learning models. BMC Pregnancy Childbirth 2024 Aug 31;24(1):574. [doi: 10.1186/s12884-024-06783-7] [Medline: 39217284]
41. Hu X, Hu X, Yu Y, Wang J. Prediction model for gestational diabetes mellitus using the XG Boost machine learning algorithm. Front Endocrinol (Lausanne) 2023;14:1105062. [doi: 10.3389/fendo.2023.1105062] [Medline: 36967760]
42. Lee SM, Hwangbo S, Norwitz ER, et al. Nonalcoholic fatty liver disease and early prediction of gestational diabetes mellitus using machine learning methods. Clin Mol Hepatol 2022 Jan;28(1):105-116. [doi: 10.3350/cmh.2021.0174] [Medline: 34649307]
43. Ding T, Liu P, Jia J, Wu H, Zhu J, Yang K. Application of machine learning algorithm incorporating dietary intake in prediction of gestational diabetes mellitus. Endocr Connect 2024 Dec 1;13(12):e240169. [doi: 10.1530/EC-24-0169] [Medline: 39393404]

XSL•FO
RenderX

44. Liu H, Li J, Leng J, et al. Machine learning risk score for prediction of gestational diabetes in early pregnancy in Tianjin, China. Diabetes Metab Res Rev 2021 Jul;37(5):e3397. [doi: 10.1002/dmrr.3397] [Medline: 32845061]
45. Kolozali S, White SL, Norris S, Fasli M, van Heerden A. Explainable early prediction of gestational diabetes biomarkers by combining medical background and wearable devices: a pilot study with a cohort group in South Africa. IEEE J Biomed Health Inform 2024 Apr;28(4):1860-1871. [doi: 10.1109/JBHI.2024.3361505] [Medline: 38345955]
46. Speiser JL. A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data. J Biomed Inform 2021 May;117:103763. [doi: 10.1016/j.jbi.2021.103763] [Medline: 33781921]
47. Uddin S, Lu H. Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data. PLoS ONE 2024;19(4):e0301541. [doi: 10.1371/journal.pone.0301541] [Medline: 38635591]
48. Wang X, Ren H, Ren J, et al. Machine learning-enabled risk prediction of chronic obstructive pulmonary disease with unbalanced data. Comput Methods Programs Biomed 2023 Mar;230:107340. [doi: 10.1016/j.cmpb.2023.107340] [Medline: 36640604]
49. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot 2013;7(21):21. [doi: 10.3389/fnbot.2013.00021] [Medline: 24409142]
50. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. Annu Rev Public Health 2016;37:61-81. [doi: 10.1146/annurev-publhealth-032315-021353] [Medline: 26667605]
51. Behman R, Bubis L, Karanicolas P. Prospective and retrospective cohort studies. In: Evidence-Based Surgery: A Guide to Understanding and Interpreting the Surgical Literature: Springer International Publishing; 2019:159-170. [doi: 10.1007/978-3-030-05120-4_16]
52. Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet 2002 Jan 19;359(9302):248-252. [doi: 10.1016/S0140-6736(02)07451-2] [Medline: 11812579]
53. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019 Oct;1(6):e271-e297. [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
54. Ioannidis JPA, Greenland S, Hlatky MA, et al. Increasing value and reducing waste in research design, conduct, and analysis. Lancet 2014 Jan 11;383(9912):166-175. [doi: 10.1016/S0140-6736(13)62227-8] [Medline: 24411645]
55. Shim JM, Shin E, Johnson TP. Self-rated health assessed by web versus mail modes in a mixed mode survey: the digital divide effect and the genuine survey mode effect. Med Care 2013 Sep;51(9):774-781. [doi: 10.1097/MLR.0b013e31829a4f92] [Medline: 23774510]
56. Lee KJ, Tilling KM, Cornish RP, et al. Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. J Clin Epidemiol 2021 Jun;134:79-88. [doi: 10.1016/j.jclinepi.2021.01.008] [Medline: 33539930]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
DL: deep learning
DOR: diagnostic odds ratio
GDM: gestational diabetes mellitus
LR: likelihood ratio
ML: machine learning
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis
PROBAST: Prediction Model Risk of Bias Assessment Tool
RF: random forest
ROB: risk of bias
SROC: summary receiver operating characteristic curve
XGBoost: extreme gradient boosting
τ: standard deviation
τ^2: standard variance

XSL•FO
RenderX

Edited by S Brini; submitted 26.Jun.2025; peer-reviewed by F Khosravi, X Guo; accepted 05.Jan.2026; published 30.Jan.2026.

Please cite as:
Liang Y, Dai A, Luo M, Zheng Z, Shen J, Su Y, Li Z
Predictive Performance of Artificial Intelligence Algorithms for Gestational Diabetes Mellitus in Pregnant Women: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e79729
URL: https://www.jmir.org/2026/1/e79729
doi:10.2196/79729

© Yingni Liang, Anran Dai, Meiyan Luo, Zhuolian Zheng, Jiayu Shen, Yinhua Su, Zhongyu Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Barriers to Digital Health Adoption in Older Adults: Scoping Review Informed by Innovation Resistance Theory

Yosefa Birati[1,2], RN, PhD; Roy Tzemah-Shahar[1], PhD

[1]The Cheryl Spencer Department of Nursing, Faculty of Social Welfare and Health Sciences, University of Haifa, Haifa, Israel

[2]The Center of Research and Study of Aging, University of Haifa, Haifa, Israel

Corresponding Author:
Yosefa Birati, RN, PhD
The Cheryl Spencer Department of Nursing, Faculty of Social Welfare and Health Sciences, University of Haifa, Haifa, Israel

Abstract

Background: The transformation of digital health technologies has reshaped health care delivery in primary care. Despite these benefits, older adults remain among the most resistant users. Traditional technology adoption models may not fully capture this reluctance, which is shaped not only by usability challenges but also by emotional, psychological, and identity-related concerns. Innovation resistance theory (IRT) offers a complementary framework focused on barriers to adoption rather than solely on facilitators.

Objective: This study aims to map and synthesize evidence on older adults' resistance to digital health in primary care through the lens of IRT, and to examine how resistance factors align with, extend, or refine IRT's functional and psychological barriers.

Methods: A scoping review with concept-driven thematic synthesis was conducted. A search for studies published between 2014 and 2025 was conducted across 5 databases: PubMed, CINAHL, Ovid Medline, Web of Science, and Scopus; the final search was completed in November 2025. Eligible studies were those that examined barriers or resistance to digital health use among adults aged 60 years and older in primary care settings. Search terms included "older adults," "digital health/eHealth," and "technology resistance." We excluded studies outside primary care and in which caregivers or health care professionals were the primary users. Data were extracted into a structured matrix and coded to the IRT domains: usage, value, risk, tradition, and image barriers. Relational integration was used to examine co-occurrence and linkages among barriers to inform the conceptual model.

Results: Seventeen studies were included, comprising 6822 participants (sample sizes ranged from 11 to 4525). Most studies were conducted in high-income Western countries, predominantly with qualitative designs, alongside mixed-methods and cross-sectional surveys. Functional barriers included usability challenges, interface complexity, and age-related impairments. Psychological resistance was linked to emotional discomfort, symbolic misalignment, and concerns about the loss of relational care. Value and risk concerns included distrust in diagnostic accuracy, privacy and data security, and skepticism about care quality. Traditional preferences for face-to-face interactions and generational digital divides reinforced image-based resistance. Interactions between barriers were identified, with low self-efficacy and technology anxiety creating feedback loops that reinforce avoidance behaviors.

Conclusions: Older adults' resistance to digital health is not simply a lack of adoption but a complex, emotionally grounded process involving functional, psychological, and identity-based barriers. This review applies IRT to primary care digital health, shifting the focus from adoption facilitators to resistance mechanisms and integrating co-occurrence patterns into a conceptual model. The synthesis reveals interacting factors of usability, self-efficacy, anxiety, trust, and legitimacy concerns that reinforce avoidance, suggesting that implementation strategies should extend beyond technical usability to rebuild trust, preserve relational care, and align digital solutions with older adults' values. Review limitations include the predominance of Western-based studies and limited longitudinal data on how resistance evolves.

(J Med Internet Res 2026;28:e75591) doi:10.2196/75591

KEYWORDS

older adults; digital health; telemedicine; technology resistance; technology adoption; primary health care; innovation resistance theory

XSL•FO
RenderX

Introduction

Background

The digital transformation of health care has been driven by the integration of telemedicine, mobile health (mHealth) applications, electronic health records, and wearable devices, which have significantly reshaped the delivery of medical services. These innovations address the limitations of traditional care models, which often struggle to meet the evolving demands of health care, particularly for aging populations in rural or underserved areas [1]. By improving access to care, supporting chronic disease management, and promoting preventive health care initiatives, digital health technologies offer promising solutions.

Notably, older adults, who often face mobility limitations, chronic illnesses, and restricted access to traditional health care services, are likely to gain substantial benefits from these digital health innovations [2,3]. However, despite the potential benefits, older adults remain among the most resistant groups to adopting these technologies [4]. This reluctance is widely documented in prior research and often attributed to multiple factors, including limited digital literacy, usability concerns, lower self-efficacy, privacy concerns, and a strong preference for in-person health care interactions. These barriers contribute to older adults' limited willingness to engage with digital health care solutions [5-7].

The persistence of this reluctance suggests that adoption-centric models may offer an incomplete explanation, highlighting the need for complementary resistance-focused frameworks. To better understand these patterns, we first reviewed established technology adoption models used in health care, clarifying their scope and limitations, and then introduced innovation resistance theory (IRT) as a complementary resistance-focused framework.

Existing Technology Adoption Models

Established theoretical models, such as the technology acceptance model (TAM) and the Unified Theory of Acceptance and Use of Technology (UTAUT), have been widely used to explain individuals' adoption and use of new technologies [8]. These models highlight factors such as perceived usefulness, ease of use, performance expectancy, and social influence as key determinants of technology adoption [9,10]. Complementary to these, Rogers' Diffusion of Innovations Theory describes how new technologies spread through populations by considering factors such as adopters' characteristics, communication channels, and social systems [11]. These frameworks have been extensively validated and remain central tools for understanding and measuring technology acceptability and usage intentions in health care.

In the context of older adults' digital health use, adoption-focused models provide valuable insights into factors associated with acceptance and initial uptake; however, prior literature suggests that older adults' persistent nonuse and resistance are also shaped by affective, psychological, and contextual factors that are not always represented as central constructs in these models [12]. For example, a scoping review by Wilson et al [13] applied UTAUT2 as an analytic framework to map barriers and facilitators to eHealth use among older adults. They identified gaps in the evidence base for certain UTAUT2 constructs (eg, habit and hedonic motivation) alongside recurring concerns related to privacy, trust, and support needs [13]. Another empirical study showed that older adults' intention to use mHealth was not explained solely by perceived ease of use and perceived usefulness, with person-related, technology-related, and contextual barriers influencing adoption [14]. Fox and Connolly further argue that research on older adults' resistance to mHealth remains limited and therefore examine how privacy concerns, trust, and risk beliefs influence willingness to adopt beyond standard adoption-model constructs [15]. Taken together, these findings suggest that complementing adoption-focused models with resistance-oriented frameworks may better capture why some older adults actively avoid digital health technologies, including perceived risks, emotional discomfort, and contextual constraints [12,16-18]. Accordingly, adoption-focused models may emphasize intention and perceived benefits, whereas nonadoption can also reflect an active decision-making process shaped by perceived risks and psychological discomfort. Therefore, we propose complementing adoption-focused theories with a resistance-oriented framework, such as IRT.

Innovation Resistance Theory (IRT) as a Conceptual Framework

IRT, introduced by Ram and Sheth [19], was developed to understand consumer resistance to marketing innovations and their behavior. Unlike models that emphasize adoption facilitators, IRT focuses on understanding why individuals hesitate or actively refuse to adopt new products, services, and ideas, even when they offer potential benefits [19]. The strength of IRT lies in its focus on perceived barriers rather than enablers, making it well-suited for populations such as older adults, where complex emotional, cognitive, and contextual factors influence nonuse. By focusing on the barriers, IRT offers a different perspective that shifts attention from the characteristics of innovations themselves to the reasons behind consumer reluctance to adopt them, especially when such adoption threatens established habits and routines or involves perceived risks [20-22]. In this view, resistance is not merely a lack of adoption but an active process that focuses on barriers to acceptance, including functional, psychological, and social resistance factors [19].

Resistance is defined as a multidimensional construct encompassing 3 dimensions: cognitive resistance, which involves individuals' appraisal of innovations and their perceived risks; affective resistance, which stems from emotional responses such as fear, frustration, or anxiety; and behavioral resistance, which manifests in actions ranging from passive disengagement to active opposition [23,24]. Within the IRT, these dimensions are further classified into functional and psychological barriers. Functional barriers include the usage barrier, which reflects the extent to which an innovation is perceived as requiring changes to established routines or habits; the value barrier, which arises when the individual perceives that the benefits of an innovation do not outweigh its costs; and the risk barrier, which represents concerns about the financial, functional, and social consequences of adopting an innovation.

XSL•FO
RenderX

Psychological barriers encompass traditional barriers, which refer to the degree to which an innovation forces an individual to accept changes that challenge cultural norms or long-standing behaviors, and image barriers, which relate to the degree to which an innovation is perceived as having an unfavorable image or negative associations [19,25]. These psychological categories often reflect deeper symbolic concerns, such as identity, generational belonging, or perceived legitimacy of digital care. This classification allows IRT to capture the multifaceted nature of resistance in older populations, particularly their emotional unease, normative preferences, and experiential distrust of digital systems. By categorizing resistance into functional and psychological barriers, IRT may provide a comprehensive framework for understanding why older adults struggle to adopt digital health solutions.

Over time, IRT has gained strong empirical support across different service and technology contexts. For example, in mobile banking research across Thailand and Taiwan, IRT barriers explained 60% - 66% of the variance in resistance intentions, with usage, value, risk, and image barriers showing statistically significant effects [21]. In a large Italian survey, Spinelli et al [26] showed that usage barriers and value-related concerns significantly reduced both actual mobile payment use and intention to adopt, whereas risk and image barriers had weaker or nonsignificant effects, and their impact varied across technology-readiness clusters [26]. Similarly, a study of Internet and mobile banking in Finland found that the value barrier was the dominant inhibitor of adoption and intention to adopt, while image and tradition barriers differentiated postponers from rejecters across seemingly similar service innovations [20].

Together, these findings demonstrate that IRT-based barriers have substantial explanatory power for resistance to digital innovations. Therefore, in this review, we apply IRT to structure the evidence on older adults' resistance toward digital health technologies and to examine whether the identified resistance factors map onto, extend, or refine the original IRT barrier categories. The aim of this scoping review was to synthesize and conceptualize evidence on older adults' resistance to digital health technologies in primary care using IRT. Specifically, we aimed to identify and categorize resistance factors into IRT functional and psychological barriers and to examine how these barriers co-occur and interact to inform a conceptual model of resistance. The review was guided by the following research questions: (1) What is known from the existing literature about older adults' resistance to using digital health technologies for monitoring and management in primary health care? (2) What are the functional (usage, value, risk) and psychological (tradition, image) IRT barriers reported across studies? (3) How do IRT barriers co-occur and link within and across studies?

Methods

Study Design

The methodology for this scoping review follows the framework proposed by Arksey and O'Malley [27], incorporating refinements by Levac et al [28], and the Joanna Briggs Institute (JBI) Reviewers' Manual [29]. We selected the scoping review approach to explore the current body of knowledge regarding older adults' resistance to digital health technologies through the lens of IRT, as it is well-suited to mapping the existing literature, identifying and interpreting patterns of functional and psychological resistance across heterogeneous study types. Within this design, our goal was to provide a theory-informed synthesis that evaluates how well IRT accounts for older adults' resistance to digital health in primary care and to identify conceptual and empirical gaps that warrant further investigation and measurement development. The reporting of this scoping review was guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) guidelines [30]. Reporting of the search methods followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension checklist (PRISMA-S) [31] to ensure transparent and complete reporting. The completed PRISMA-ScR checklist is provided in Checklist 2, and the PRISMA-S checklist is provided in Checklist 1.

Stage 1: Identifying the Research Question

The review was guided by predefined research questions (presented at the end of the Introduction section) informed by IRT and scoping review guidance.

Stage 2: Searching and Identifying Relevant Studies

A literature search was conducted across 5 major databases: PubMed, CINAHL, Ovid Medline, Web of Science, and Scopus, to identify peer-reviewed publications relevant to the research question. These databases were selected for their broad coverage of health, behavioral, and interdisciplinary studies on older adults and digital health. Each database was searched separately through its web interface, and all retrieved records were exported to Mendeley (version 1.63.0; Mendeley Reference Manager) for deduplication. Review studies were not included in this scoping review; however, their reference lists were screened to identify potentially eligible primary studies. No study registries were searched. Apart from reference-list screening, no additional sources were searched, and no citation searching was undertaken. We did not contact authors to identify additional studies, and no other search methods were used beyond those described. We did not use any previously validated search filters. Search strategies were developed specifically for this scoping review by the authors and were not peer reviewed by an independent expert before execution. We did not adapt or reuse search strategies from previous literature reviews for any substantive part of our search.

The search was carried out on December 20, 2024, and was rerun on November 18, 2025, to identify newly published studies since the initial search. The search followed the JBI PCC structure (Participants, Concept, Context) and combination of the following keywords and MeSH terms: "older adults," "elderly," phenomena of "digital health," "eHealth," "Telemedicine," and context of "primary health care," and "barriers to health technology." Boolean operators were used to combine search strings (eg, AND, OR). Title and abstract screening and full-text review were conducted by 2 independent reviewers (YB and RTS). The search strategy and keyword combination can be found in Multimedia Appendix 1. Additionally, reference lists of included studies were manually

XSL•FO
RenderX

screened to identify relevant studies not captured in the initial searches.

Stage 3: Selecting the Relevant Studies

Inclusion and Exclusion Criteria

The review included papers that met predefined inclusion and exclusion criteria aligned with the JBI PCC framework for scoping reviews (Table 1).

Table . Study eligibility criteria (Population-Concept-Context) for the scoping review.

Criteria	Inclusion	Exclusion
Participants/population	• Older adults aged 60 years and older	• Children, adolescents, and younger adults aged <60 years • Caregivers • Health care professionals
Concept (intervention)	• Use of mHealth[a] for monitoring and management • mHealth: telemedicine, mobile phone apps, smartphone apps, web-based systems • Resistance or barriers to the use of digital health technologies	• Use of mHealth telemonitoring for patients who are not adults and younger adults aged <60 years • Use of mHealth telemonitoring by caregivers or health care professionals
Context (cultural factors, geographic location, setting)	• Use of digital health technologies in primary health care	• Secondary/tertiary care: hospital inpatient wards, surgical centers • Emergency/urgent care
Type of studies	• Qualitative, quantitative, or mixed methods studies • Observational and experimental, cross-sectional, or longitudinal, randomized controlled trial or nonrandomized or noncontrolled trial, case series or case reports	• Conference abstracts, editorials, commentaries, letters to editor, essays, book chapters, and books
Language	• English	• Language other than English
Publication date	• From 2014	—[b]

[a]mHealth: mobile health.

[b]Not applicable.

The context of the review centered on the resistance to digital health within the framework of IRT. Publications addressing the use of digital health within the resistance domains of usage, value, risk, traditional, and image barriers were considered, while those focusing solely on the description of digital health adoption and facilitators were excluded. Also, no minimum sample size threshold was applied. Consistent with the objectives of a scoping review, studies were eligible regardless of their sample size to maximize coverage of designs (qualitative, quantitative, mixed methods) and contexts.

Study Selection Process

The studies were screened against the inclusion and exclusion criteria developed by the authors. The selection process followed three steps: (1) Title and abstract screening to remove irrelevant or duplicate records; (2) Full-text review based on predefined inclusion and exclusion criteria; and (3) Final inclusion based on relevance for examination of resistance to digital health technologies among older adults.

A total of 4976 records were identified through database searches, and 2387 duplicates were removed. After screening 2589 titles and abstracts, 227 full-text articles were reviewed. Two independent reviewers (YB and RTS) evaluated relevant publications for eligibility and selected qualifying publications based on the inclusion/exclusion criteria. We used a consensus-based approach, prioritizing unanimous agreement through re-evaluation of the eligibility criteria; if consensus could not be reached, a third reviewer would adjudicate. An initial pilot screening was conducted independently by both reviewers to ensure consistent interpretation of the eligibility criteria. Discrepancies identified at this stage were resolved through discussion and used to refine the criteria, resulting in full agreement during subsequent screening. A total of 17 studies met the inclusion criteria and were included in the final synthesis. A PRISMA flow diagram illustrates the selection process.

Stage 4: Charting the Data - Data Extraction and Synthesis

Two authors independently extracted data from all included studies. Data were charted using a standardized extraction form developed for this review, capturing study design, aims, population, type of digital health intervention, and resistance-related findings. Using a concept-driven thematic synthesis, findings were organized into 5 resistance categories from the IRT: usage, value, risk, tradition, and image barriers.

XSL•FO
RenderX

A structured matrix was used to map resistance dimensions across the studies. Barrier statements were first open-coded descriptively and then mapped to one IRT family using prespecified rules. Data charting was conducted by the 2 authors, and disagreements were resolved by consensus.

Stage 5: Collating, Summarizing, and Reporting the Results

Findings were organized in three layers: (1) mapping of the evidence base (study characteristics, settings, modalities), (2) concept-driven qualitative synthesis using IRT classification (usage, value, risk, tradition, and image), and (3) relational integration examined interconnections across IRT barriers. We extracted and coded barrier co-occurrences and linkages reported in the studies' results sections and participant quotations when two or more barriers were described as co-occurring or interacting. Links were considered explicit when directly stated, inferential when implied within a study's narrative context, and integrative when consistent patterns recurred across multiple studies (worked examples are provided in the Results).

Results

General Characteristics of the Studies

The database search initially identified 4976 records. After removing duplicates, screening titles and abstracts, and full papers, 17 studies were included in the final synthesis (Figure 1). The included papers represent a predominantly high-income Western countries from the United States (n=4), Sweden (n=3), the Netherlands (n=3), Canada (n=2), Finland (n=1), Norway (n=1), and the United Kingdom (n=1), with only 2 studies from non-Western settings Israel (n=1) and Indonesia (n=1). Eleven studies were qualitative, 3 studies were cross-sectional, and 4 studies were mixed methods designs. Sample sizes ranged from 11 to over 4500 participants, though qualitative samples were generally smaller and in-depth. In terms of digital health modalities, most studies focused on telemedicine or digital consultations (12/17) and patient portals or eHealth services (8/17), with comparatively few studies examining mobile apps or tablets (3/17) and wearables or remote monitoring (2/17) (Table 2).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the screening and selection process for the Scoping Review on resistance to digital health among older adults.

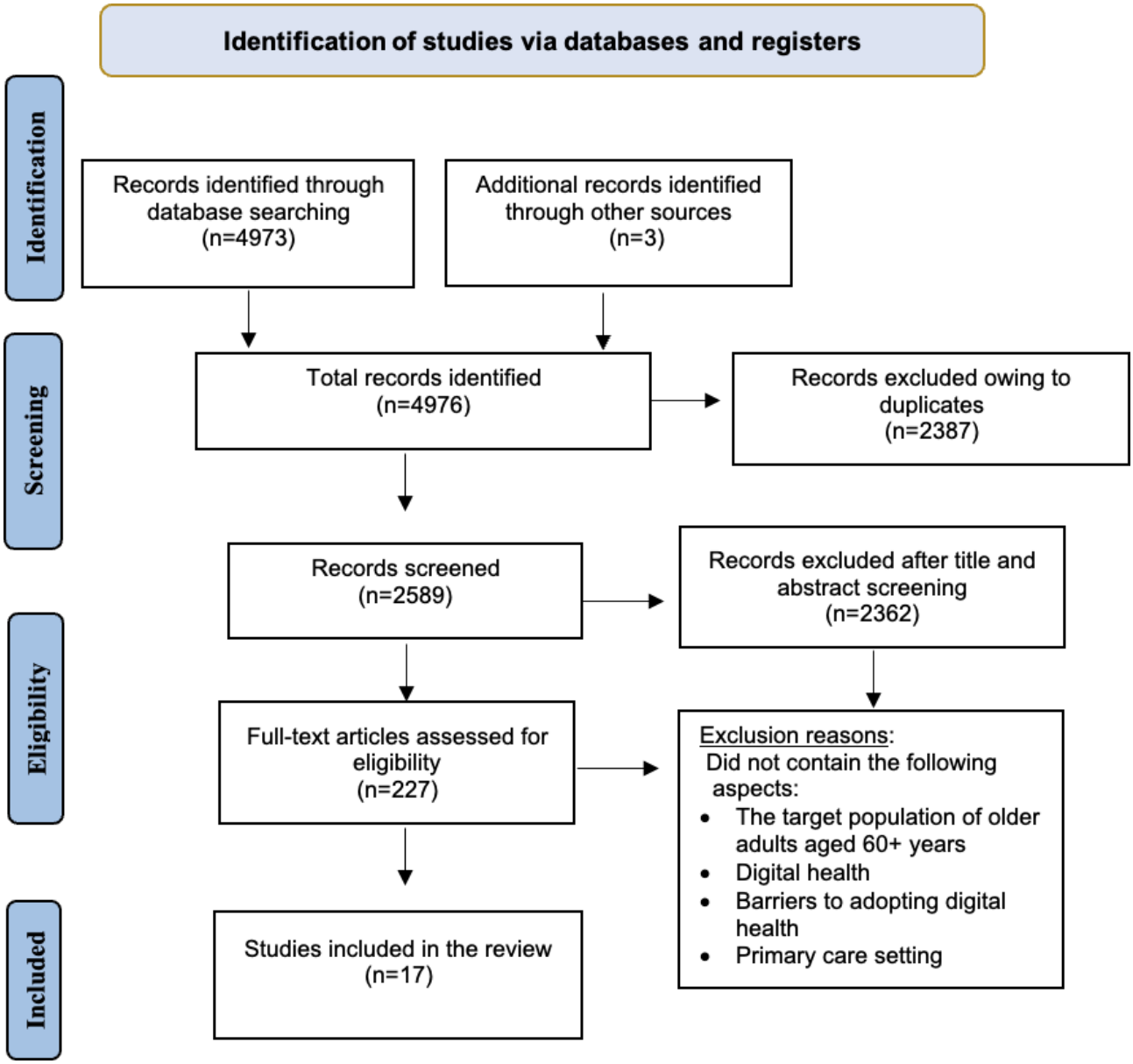

XSL•FO
RenderX

Table . Included studies on older adults' resistance to digital health technologies (n=17). This table presents the country, study design, population sample size, age range, and digital health modality.

	Study design	Aims	Study population	Digital health
Khanassov et al [32] (Canada)	Qualitative study: semistructured interviews and 3 focus groups to explore the experiences of both older adults and health care professionals	• How do older adults and health care professionals experience the use of telemedicine? • What are the facilitators and barriers to telemedicine use in the care of older adults? • What recommendations can enhance telemedicine engagement for older adults and health care professionals	• 29 older adults and health care professionals (family physicians, nurses, social workers, physiotherapists) • Age range 65 - 90 years	• Telemedicine in primary care
Vergouw et al [33] (Netherlands)	Qualitative study: semistructured and think-aloud interviews	• Identify the needs, barriers, and facilitators among community-dwelling older adults (60 years and older) with chronic health conditions in using web-based eHealth applications to support general practice services	• 19 community-dwelling older adults with at least one chronic condition • Mean age 73 (SD 5.3) years	eHealth applications for: • e-Consultation • Schedule e-Appointment • e-Prescription ordering • e-Lab results viewing • Access to e-File
Knotnerus et al [34](Netherlands)	Qualitative study: semistructured interviews thematic analysis	• Investigate the experiences of older patients (65 years and older) who use a digital health platform in general practice • Identify barriers and facilitators for using digital health • Examine whether a practice's focus on digital health influences older patients' choice to become a patient at the practice	• 18 older patients enrolled in 2 general practices • Age range 68 - 89 years	Digital health platform for: • Communicate with general practitioners • Appointment scheduling • Order repeat medications
Bhatia et al [35] (United States)	Cross-sectional multimethod study: mixed methods (Quantitative and Qualitative: close and open-ended questions)	• Understand older adults' experience with primary care telemedicine since the COVID-19 pandemic • Identify satisfaction levels and technical challenges in telemedicine use • Provide policy recommendations for the future of telemedicine services	• 208 older adults (≥65 y) who had a telemedicine visit within primary care visit • Mean age 74.4 (SD 4.4)	• Telemedicine (telephone and video visits)

XSL•FO
RenderX

	Study design	Aims	Study population	Digital health
Lam et al [36] (United States)	Cross-sectional study: data from the 2018 National Health and Aging Trends Study (NHATS)	• Assess the prevalence of telemedicine unreadiness and how older adults may be left behind in the United States when the migration to telemedicine occurred • Identify key barriers preventing the use of video-based telemedicine • Examine disparities in telemedicine access based on demographic, socioeconomic, and health-related factors	• 4525 community-dwelling older adults (≥65 y) in the United States • Mean age 79.6 (SD 6.9)	• Telemedicine (video and telephone visits)
Nymberg et al [37] (Sweden)	Qualitative research using focus group interviews thematic content analysis	• Explore older adults' beliefs, attitudes, experiences, and expectations regarding eHealth services in primary health care • Understand factors influencing adherence to eHealth tools in primary care among elderly patients • Identify barriers and facilitators affecting older adults' engagement with eHealth services	• 15 elderly patients from 3 primary health care centers in Southern Sweden, selected based on chronic disease status and medication use • Age range 65 - 80 years	eHealth services and use of the mobile phone for: • Contacting the health care system via web • Self-monitoring of chronic illnesses • Seeking medical information
van Houwelingen et al [38] (Netherlands)	A mixed method triangulation design, including a cross-sectional survey study (quantitative phase) and qualitative observations of older adults performing digital tasks in their daily lives	• Understand older adults' readiness for telehealth, particularly videoconferencing • Identify factors influencing their intention to use videoconferencing • Examine their capacities and barriers in using digital technology in daily life	• 256 participants in the survey and 15 older adults aged 65 years or older in the qualitative observations • Median (IQR) age=71 (67 - 76) years	• Telehealth, focused particularly on the use of videoconferencing for health care consultations
Laukka et al [39] (Finland)	Survey study with qualitative inductive content analysis of open-ended questions	• Investigate the preferences and needs of older adults regarding the use and development of digital health and social services • Understand how digital health and social services can be designed to more effectively meet the needs of older adults	• 1100 Finnish individuals aged 75 and older • Age range 75 - 99 years	• Telemedicine consultations • eHealth services
Rochmawati et al [40] (Indonesia)	Exploratory qualitative study using semistructured interviews, thematic analysis			• Digital health technologies (mobile apps, smartwatches) for health monitoring.

XSL•FO
RenderX

	Study design	Aims	Study population	Digital health
		• Explore the acceptance of eHealth technology among older adults in primary care • Examine perceptions, attitudes, experiences, and expectations of older people patients regarding eHealth services used in primary care	• 11 Older adults with chronic conditions (diabetes, hypertension) from a suburban primary health clinic in Indonesia • Mean age 66.9 years	
Fjellså et al [41] (Norway)	Explorative qualitative study using semistructured interviews	• To explore multimorbid older adults' experiences with participation and eHealth in care coordination with the support of general practitioners and district nurses	• 20 older adults with multimorbidity (COPD, heart failure, diabetes, and physical disabilities) receiving primary care services • Mean age 82 (range 71 - 98) years	• Patient portals to share and access information • Electronic messaging with general practitioners • Schedule appointments • Order prescriptions
Mao et al [42] (United States)	Mixed methods needs assessment (cross-sectional survey and qualitative interviews)	• Identify barriers to telemedicine video visits among older adults with differing socioeconomic backgrounds and primary spoken languages • Understand technological, cognitive, and language-related obstacles to telemedicine use • Provide recommendations to improve access and engagement with telemedicine	• 249 older adults from 2 independent living facilities • Mean age 84.6 (SD 6.6) years	• Telemedicine visits.
Frishammar et al [43] (Sweden)	Qualitative interviews and process data from a Swedish DHP provider	• To investigate adoption and usage barriers of digital health platforms among older adults • To understand how to facilitate increased adoption and usage of digital health platforms among the elderly	• 22 older adults aged ≥65 years, including both users and nonusers of digital health platforms, as well as individuals with experience in digital health development • Age range 65 - 80 years	• Video calls • Chats • Asynchronous messaging
Haimi et al [44] (Israel)	Qualitative study using semistructured interviews.	• To identify the challenges and barriers faced by the senior population when utilizing telemedicine services	• 14 elderly individuals from a primary health care clinic in Israel • Mean age=73 (range 66 - 85) years	• Telemedicine (phone and video visits) • electronic medical records prescription refills • Digital referrals • Electronic messages with the medical provider
Landgren and Cajander [45] (Sweden)	Qualitative, semistructured interviews.		• 13 participants aged >65 years • Mean age 74 years	• Digital health consultations delivered by video or chat/phone applications in primary care settings

XSL•FO
RenderX

	Study design	Aims	Study population	Digital health
		• To identify reasons for nonuse of digital health consultations among elderly in rural areas • To describe their attitudes toward technology, and possible challenges and opportunities.		
Ahmed et al [46] (United Kingdom)	Qualitative, focus group study.	• To explore the experiences, perceptions, and expectations of older adults from 3 minoritized ethnic group backgrounds regarding digitalized primary care services since the beginning of COVID-19.	• 27 participants age >65 years • Median (IQR) age=69 (66.5 - 72.5) years	• Telemedicine (phone and video visits) • Web-based services: View medical records Schedule appointments Order prescriptions
Ufholz et al [47] (United States)	Cross-sectional survey.	• To assess telemedicine preparedness of older primary care patients: internet use, device ownership, prior telemedicine experience, concerns, and perceived barriers	• 30 community-dwelling adults aged ≥65 • Age range 65 - 89 years	• Telemedicine for primary care (video/online visits)
Sproul et al [48] (Canada)	Cross-sectional survey	• To determine what technologies and apps are in current use by older adults, to explore the types of technologies and apps that may be of interest to people in this age group, to explore concerns about technologies, and to examine any age-related differences	• 266 participants aged ≥60 years • 60.2% participants were 60 - 74 years and 39.8% participants were 75 years or older	• Mobile phones • Tablets • Health-related apps

The IRT framework was used to guide the coding of extracted findings into the 5 barrier domains (usage, value, risk, tradition, and image).

The findings synthesis is presented in the following sections and summarized in Tables 3 and 4.

XSL•FO
RenderX

Table . Matrix mapping of innovation resistance theory (IRT) functional and psychological barrier domains (usage, value, risk, tradition, and image) across included studies of older adults' resistance to digital health in primary care (n=17).

	Functional barriers			Psychological barriers	
	Usage barriers	Value barriers	Risk barriers	Tradition barriers	Image barriers
Khanassov et al [32]	• Technical challenges • Symptom articulation • Technology usability	• Informality bias • Limited use perception	• Diagnostic uncertainty • Missed diagnosis concern • Technology misuse anxiety	• In-person preference	• Legitimacy gap • Unsuitable for complex care
Vergouw et al [33]	• Digital learning curve • Technology usability • Interface complexity	• Limited use perception	• Privacy and security concerns • Technology misuse anxiety	• Symptom articulation • In-person preference	• Legitimacy gap
Knotnerus et al [34]	• Technology usability • Interface complexity • Symptom articulation • Digital learning curve	• Limited use perception • Disrupted continuity of care	• Privacy and security concerns • Technology misuse anxiety	• In-person preference	• Legitimacy gap
Bhatia et al [35]	• Symptom articulation • Technology usability • Cognitive and sensory limitations • Digital learning curve	N/A[a]	• Missed diagnosis concern • Technology misuse anxiety • Missed diagnosis concern	• In-person preference	• Legitimacy gap • Unsuitable for complex care
Lam et al [36]	• Digital learning curve • Symptom articulation	N/A	N/A	• In-person preference	N/A
Nymberg et al [37]	• Digital learning curve • Tech usability. • Technology anxiety • Physical and sensory impairments	• Limited use perception	• Privacy and security concerns • Technology misuse anxiety	• In-person preference • Preference for physical documentation	• Legitimacy gap • Generational digital divide
Houwelingen et al [38]	• Digital learning curve • Technology anxiety • Self-efficacy deficit	N/A	• Technology misuse anxiety • Privacy and security concerns	N/A	N/A

XSL•FO
RenderX

	Functional barriers			Psychological barriers	
Laukka et al [39]	• Interface complexity • Self-efficacy deficit • Language and terminology complexity • Physical and sensory impairments • Technology usability	• Limited use perception	• Fraud and scam concerns • Privacy and security concerns	• In-person preference • Need for familiarity in care	• Generational digital divide • Unsuitable for complex care
Rochmawati et al [40]	• Self-efficacy deficit • Digital learning curve	• Limited use perception • Informality bias	N/A	• In-person preference • Need for familiarity in care	N/A
Fjellså et al [41]	• Technology usability • Interface complexity	• Limited use perception • Informality bias	• Diagnostic uncertainty • Missed diagnosis concern • Technology misuse anxiety	• In-person preference • Need for familiarity in care	• Generational digital divide
Mao et al [42]	• Physical and sensory impairments • Digital learning curve technical challenges language barriers • Cognitive and sensory impairments • Symptom articulation • Physical and sensory impairments • Technical challenges • Technology anxiety • Interface complexity	• Limited use perception • Limited use perception	• Diagnostic uncertainty	• In-person preference	• Unsuitable for complex care
Frishammar et al [43]	• Digital learning curve • Self-efficacy deficit • Technology anxiety	• Limited use perception • Informality bias	• Diagnostic uncertainty • Missed diagnosis concern • Privacy and security concerns	• In-person preference	• Unsuitable for complex care • Legitimacy gap
Haimi et al [44]	• Symptom articulation • Technology anxiety • Language and terminology complexity • Technical challenges • Physical and sensory impairments	N/A	• Missed diagnosis concern	• In-person preference	N/A
Landgren and Cajander [45]		• Limited use perception • Informality bias		• In-person preference	• Generational digital divide

XSL•FO
RenderX

	Functional barriers			Psychological barriers	
	• Interface complexity • Digital learning curve • Self-efficacy deficit • Technology anxiety		• Diagnostic uncertainty • Missed diagnosis concern		
Ahmed et al [46]	• Technology usability • Language and terminology complexity • Interface complexity	• Limited use perception	• Diagnostic uncertainty • Technology misuse anxiety	• In-person preference • Need for familiarity in care	N/A
Ufholz et al [47]	N/A	• Limited use perception	• Privacy and security concerns • Diagnostic uncertainty	• In-person preference	N/A
Sproul et al [48]	• Technology usability	• Limited use perception	• Privacy and security concerns	N/A	• Legitimacy gap

[a]Not applicable.

XSL•FO
RenderX

Table . Thematic categorization and definitions of digital health resistance barriers subcategories among older adults.

Category and subcategory	Definition
Usage barriers	
Symptom articulation [32,34-36,42,44,45]	Difficulty in effectively describing symptoms or raising multiple health concerns during telemedicine or digital health interactions, often due to sensory limitations, cognitive strain, or unfamiliarity with web-based communication formats
Technology usability [32-35,37,39,41,44-46,48]	Difficulties interacting with digital health tools due to poor interface design, complex navigation, multi-step login processes, or lack of age-appropriate accessibility features
Digital learning curve [33-38,40,42,43,45]	Challenges individuals face in acquiring, applying, and retaining the skills required to use digital health technologies, often due to limited prior exposure or memory-related difficulties
Interface complexity [33,34,39,41,42,45,46]	Obstacles users encounter when engaging with digital platforms due to poor design elements, confusing navigation, and unclear layouts
Technology anxiety [37,38,42,43,45]	Fear or discomfort experienced when using digital health technologies, often stemming from low confidence, mistrust in one's digital abilities, or intimidation by unfamiliar systems. This anxiety may lead to hesitation or complete avoidance, driven by concerns about making mistakes that could negatively impact one's health or care
Physical and sensory impairments [35,37,39,42]	Difficulties in using digital health technologies due to age-related sensory and motor impairments, such as reduced vision, hearing loss, or diminished fine motor control
Self-efficacy deficit [38,40,43,45]	A lack of confidence in one's ability to successfully use digital health tools or perform required technological tasks, often rooted in limited digital literacy, minimal prior experience, or insufficient training and support
Language and terminology complexity [39,42,44,46]	Difficulty using digital health tools due to complex medical, technical, or bureaucratic language, often compounded by limited proficiency in the language used by the platform
Value barriers	
Informality bias [32,40,41,43,45]	Reluctance to engage with digital health tools based on the perception that they lack legitimacy or necessity in medical care, accompanied by a belief that traditional health care methods are sufficient without digital augmentation
Limited use perception [32-34,37,39-43,45-48]	The belief that digital health tools offer little to no added value compared with traditional care methods, resulting in low motivation to adopt or engage with them
Risk barriers	
Diagnostic uncertainty [32,41-43,45,46]	Concerns about the accuracy and reliability of medical diagnosis due to the absence of physical examination, direct visual assessment, and potential miscommunication, which may increase the risk of medical errors
Missed diagnosis concern [32,35,41,43-45]	Fear that health care providers will miss critical patient information and that essential health issues may be overlooked due to the absence of physical exams, technical distractions, or miscommunication in digital health interactions
Technology misuse anxiety [32-35,37,38,41,46]	Uncertainty or fear about using digital health technologies incorrectly, driven by concerns about user error, system malfunctions, or communication failures that could negatively impact care delivery
Privacy and security concerns [33,34,37-39,43,47,48]	Concerns about the confidentiality, security, and accuracy of personal medical information in digital health care services, driven by fears of data breaches, unauthorized access, and unreliable IT systems
Tradition barrier	
In-person preference [32-37,39-47]	A strong preference for face-to-face health care interactions, rooted in trust in direct communication, perceived importance of physical examinations, and the belief that in-person care offers superior quality

XSL•FO
RenderX

Category and subcategory	Definition
Need for familiarity in care [39-41,46]	Preference for established health care routines and trusted provider relationships over digital health solutions, due to a desire for personalized care, continuity with known providers, and a reluctance to alter traditional in-person interactions
Image barrier	
Legitimacy gap [32-35,37,43,48]	Perception that digital health care is less effective and trustworthy than traditional in-person care, driven by concerns about depersonalization, bureaucratic complexity, and reduced reliability, leading to skepticism about its value and quality
Unsuitable for complex care [32,35,39,42,43,45]	Perception that digital health care services are insufficient for addressing complex medical conditions or cases requiring physical examination, due to concerns about thoroughness, accuracy, and the ability to provide a comprehensive diagnosis and care
Generational digital divide [37,39,41,45]	Perception that digital health care is designed for younger users and is difficult for older adults to adopt, due to differences in familiarity, confidence, and digital literacy

Table 3 details the barriers identified by each study, presenting a matrix that maps each study to the usage, value, risk, tradition, and image barriers. Table 4 defines each barrier subcategory and summarizes how these resistance themes were operationalized across the studies.

Functional Barriers

In the context of IRT, functional barriers refer to resistance stemming from the practical and objective attributes of the innovation itself, including its required usage, perceived value, and associated risks [19].

Usage Barriers

Usage Barriers were the most consistently reported resistance factor, found in 16 studies. Older adults face significant usage barriers to adopting digital health technologies, largely due to technical challenges, usability difficulties, and concerns about quality of care. A central theme across studies was interface complexity. Many participants described digital health platforms as confusing, unintuitive, and poorly designed. Common challenges included unclear layouts, unintuitive menus, and multi-step authentication processes requiring repetitive actions such as logging in, remembering passwords, and uploading medical documents [32-34,39,41,42,45,46]. These features increased cognitive load and made even basic digital interactions feel burdensome and prone to mistakes.

The difficulties were compounded by technology usability issues linked to age-related cognitive and sensory impairments. Older adults with a decline in vision, hearing loss, or memory difficulties and reduced fine motor skills struggled with small font sizes, poor audio quality, poorly structured information, and touchscreen sensitivity, which makes many applications inaccessible without assistance [35,37,42,44,48]. In addition, language and terminology complexity emerged as a significant obstacle. Technical jargon or unfamiliar medical terms often made it difficult for users to interpret instructions or understand the content presented on-screen, particularly among those with limited formal education or health literacy [39,42,44,46].

Another recurring issue was the digital learning curve. Older adults reported limited prior experience with digital health tools or services and found it challenging to adapt to new systems [32,34,39-43,45]. This often led to a self-efficacy deficit where individuals doubted their ability to complete digital health tasks independently. These doubts fueled hesitation and reinforced a sense of digital exclusion, leading to frustrations, avoidance behaviors, and a greater need for support before successfully adopting telemedicine tools [38-40,43,45]. Closely related to this was technology anxiety, the fear of making mistakes or causing harm through improper use, which discouraged many from engaging with telemedicine platforms.

Concerns about system reliability and uncertainty about using digital health care tools make older adults feel less confident in their technical abilities and unprepared [32,33,36,39,40,42,43,45], leading to avoidance behaviors, where they opt not to engage with digital health solutions to minimize the risk of errors [37,38,42].

Beyond usability concerns, preadoption resistance arises from changes in communication dynamics within digital health care. In contrast to traditional face-to-face consultations, which allow patients to express multiple health concerns in a single visit and rely on nonverbal cues, digital health platforms, particularly telemedicine services, alter this dynamic. Studies showed that when older adults use digital health services, they struggle to articulate their symptoms or find it difficult to understand medical terminology or provider explanations [39,45]. As a result, they hesitate to fully communicate medical concerns, whether typing them into digital platforms or discussing multiple health issues during digital visits. This contributes to a perception that digital care is less effective than in-person care [34-36], further discouraging older adults from fully embracing digital health technologies.

Value Barriers

Value barriers to adopting digital health solutions among older adults primarily stem from informality bias, the perceived lack of necessity of digital tools, concerns about care quality, and misalignment between the effectiveness of available digital health care services and patient expectations [40,43,45]. While many acknowledge that telemedicine may be appropriate for minor health issues and routine follow-ups, they often do not

XSL•FO
RenderX

view it as an adequate substitute for in-person consultations. This limited use perception is particularly strong when it comes to complex conditions that require physical examination or long-term management [32,34,41-43,45,46,48].

Skepticism about the effectiveness of remote consultations is a common concern. Many older adults feel that digital platforms fail to capture nonverbal cues, which are essential for accurate medical assessment and effective patient-provider communication. This concern is particularly pronounced among individuals managing chronic illnesses, who consider ongoing physical evaluations and in-person interactions with health care professionals to be vital components of proper care [33]. Moreover, older adults often emphasize the importance of relational continuity with their health care providers, an aspect they feel is disrupted and compromised in digital health environments. Telemedicine is frequently perceived as impersonal and transactional, lacking the trust and emotional support that typically characterize in-person visits, qualities that many older adults highly value in primary care settings [33,34,37,39]. As a result, some individuals refuse to see their providers outside of traditional clinical settings, which further reinforces resistance to digital health solutions [37,42].

Beyond concerns about quality of care, many older adults also question the necessity of digital health interventions, particularly when the current health care system meets their needs effectively [33,37]. Some dismissed telemedicine as a "solution for a nonexisting problem," believing that traditional in-person visits provide sufficient care without the added complexity of digital tools [33,37,40]. This skepticism is often exacerbated by low digital literacy or past negative experiences with digital health technology, leading many to view telemedicine and digital health apps as unnecessary, ineffective, or not worth the effort required to learn and adapt [37]. When the perceived benefits of digital health do not outweigh the effort and risks associated with adoption, resistance to these solutions remains strong.

Risk Barriers

Risk barriers to digital health adoption among older adults primarily revolve around concerns about diagnostic uncertainty and the potential of missed health issues due to the absence of physical examinations, body language, and other visual cues essential to accurate clinical assessment [32,41-46]. Many older individuals worry that the lack of hands-on evaluation in telemedicine could lead to overlooked symptoms or misinterpretations by health care providers. A prominent concern is technology misuse anxiety, which arises from fear of making errors during digital interactions. Participants described anxiety about technical distractions, errors in digital documentation, incomplete data entry, and uncertainty about whether submitted information, such as messages, forms, or test results, would be properly received and understood by their health care team [41,45,46]. These apprehensions are linked to fears of miscommunication with health care providers, incorrect medical decisions, or overlooked health conditions [32-35,37,40].

Beyond diagnostic concerns, older adults express privacy and security concerns. There is a common distrust of the integrity and security of digital health platforms [33,34,37-40,43,47,48], particularly related to fear that personal health data could be exposed to unauthorized access, fraud, or misuse. Some participants described concerns about scams that mimic legitimate digital services, increasing their reluctance to trust or engage with digital health tools [39]. This skepticism is further compounded by uncertainty around how health care institutions collect, store, and share data through electronic health records and patient portals [39].

Additionally, lack of confidence in digital skills was repeatedly cited as a major factor behind misuse anxiety. Older adults often lack confidence in their digital skills, particularly in navigating complex interfaces or troubleshooting technical issues. Common fears included accidentally deleting important information, misunderstanding medical results, or failing to complete critical health care tasks [34,40]. As a result, many preferred to avoid digital health services entirely rather than risk making mistakes that could negatively impact their care.

Another key source of resistance is the perceived loss of autonomy in health care decision-making. Some older adults expressed concerns that eHealth solutions shift decision-making control from patients to automated systems, reducing their ability to advocate for personalized care and communicate effectively with health care providers about their health care [37,41]. This fear is particularly prevalent among those unfamiliar with electronic health records or unaware of how to use digital clinical discussions.

Psychological Barriers

Psychological barriers refer to resistance stemming from subjective, cognitive, and emotional conflicts between the innovation and the individual's established traditions and self-image barriers [19].

Tradition Barriers

Traditional barriers to digital health adoption among older adults arise from long-established care routines, personal preferences for in-person interactions, and a strong need for familiarity in health care interactions. Many older adults have their health-seeking behaviors around face-to-face consultations, expressing satisfaction with traditional care models and questioning the necessity or value of digital alternatives [37,40,41]. They often perceive little incentive to switch to eHealth services when current systems already meet their expectations [37,44,45]. A central theme is the belief that *in-person interactions* offer superior quality of care, stronger provider-patient relationships, and greater emotional warmth. Digital platforms are often seen as impersonal, lacking the human touch and nonverbal communication cues that older adults consider essential for effective medical consultations [32,34-36,39,41,43,45-47]. This is especially concerning for individuals managing chronic conditions or complex health issues, where verbal-only communication may be insufficient for accurate symptom reporting and clinical assessment [32,33].

The need for familiarity in care also contributes to resistance toward digital health adoption. Many older adults prefer continuity with known health care professionals, such as physicians, nurses, or other health care professionals, and value personalized guidance and documentation, such as printed instructions or handwritten over generic digital content. Some

XSL•FO
RenderX

do not want all services to be transferred through digital platforms, especially when health care and social service issues are too complex to be handled without face-to-face contact [35,37,39,42,46].

Another common concern is the perceived legitimacy of telemedicine. Some older adults do not view phone or video consultations as valid medical encounters, describing them as informal and lacking the authority of traditional office visits [32]. This perception is heightened among individuals who had not used digital health before the COVID-19 pandemic and who experienced the rapid shift to telehealth as both disruptive and disorienting, owing to complex interfaces and limited user guidance [34]. For these individuals, digital health solutions interfere with familiar health care routines and pose significant adaptation challenges [39].

Image Barriers

Image barriers to digital health adoption among older adults arise from negative perceptions of technology, distrust in digital health solutions, and skepticism about their legitimacy and effectiveness in clinical care. Many older adults associate digital health technologies with lower quality of care and consider them as an unacceptable alternative to traditional in-person visits [35,42]. For some, these technologies are viewed as overly complex, impersonal, and rigid, contributing to a Legitimacy Gap, a perception that digital health care lacks the authenticity, reliability, and interpersonal value of conventional medical interactions [33,43,48]. This skepticism is reinforced by the belief that health care should be hands-on, personalized, and relational, the qualities they feel digital platforms fail to deliver.

Another central issue underlying this perception is the Generational Digital Divide. Many older adults view digital health tools as designed primarily for younger, digitally proficient users, and they report feeling excluded or disadvantaged by their limited experience with digital technologies [37,39-41,45]. This belief is often coupled with self-perceived technological inadequacy, where individuals feel "too old" to learn or incapable of using new systems effectively [39]. These psychological barriers are compounded by negative past encounters with health care bureaucracy or poorly designed interfaces, which foster the impression that digital health prioritizes efficiency over patient-centered care [37]. Additionally, difficulties navigating eHealth platforms often lead to a sense of powerlessness in managing their health, further alienating them from digital solutions.

Older adults also view telemedicine and digital health as unsuitable for both routine and complex care needs [32,42,43]. Many perceive these technologies as inferior to traditional, in-person medical consultations, citing concerns about their inability to provide thorough physical examinations, comprehensive assessments, and hands-on diagnostics [39]. Digital health is also associated with social isolation and reduced autonomy, as some fear that shifting toward digital health care may limit direct patient-provider interactions and diminish their role in medical decision-making [34]. This contributes to a strong preference for traditional care models, where in-person visits provide greater trust, familiarity, and perceived quality.

Evidence and Gap Map

Across the included studies, there was substantial variation in both the types of digital health technologies examined and the specific resistance factors reported. To strengthen the mapping component of this scoping review, we developed an evidence and gap map to summarize the distribution of evidence and identify gaps across digital health modalities and resistance constructs. Guided by IRT, we categorized studies by the type of digital health modality and by IRT-informed barrier subcategories derived from the extracted findings (Figure 2).

Figure 2. Evidence and gap map of digital health modalities by IRT-informed resistance subcategories in primary care among older adults. Bubble size and color intensity represent the number of included studies contributing to each intersection (n=17). IRT: innovation resistance theory.

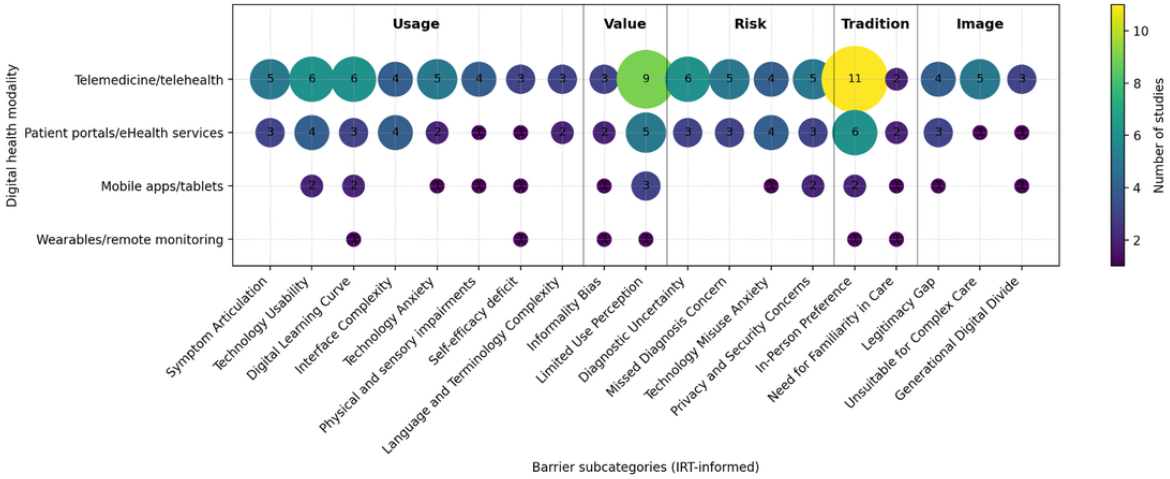

Specifically, the map highlights that evidence is concentrated in studies of telemedicine and patient portals or eHealth services, with fewer studies addressing mobile apps or tablets and minimal evidence on wearables or remote monitoring. Across modalities, frequently represented barriers included usability and interface complexity, self-efficacy and technology anxiety, and trust-related concerns such as privacy, data security, and perceived legitimacy of digital encounters. In contrast, several modalities-barrier intersections show limited or absent evidence, indicating that resistance to certain technologies, particularly wearables and app-based monitoring, remains underexplored in primary care contexts.

XSL•FO
RenderX

Conceptual Integration: Interconnected Barriers Leading to Digital Health Avoidance

Across the 17 included studies, usage barriers were the most consistently reported (16/17 studies). Risk barriers and tradition barriers were also prevalent (15/17 studies). Value barriers were common (13/17 studies), and image barriers were reported in a smaller, but still substantial subset (11/17 studies). Co-occurrence patterns were apparent across domains, and worked examples illustrate how linkages were derived. For example, one participant described limiting use to familiar functions and avoiding other features, indicating a usage barrier, accompanied by anxiety when stepping outside her comfort zone, suggesting an affective risk component and fear of making mistakes: “I never look over there, I just do everything I have learned… Outside of that, I become nervous.” [38]. In another study, a participant noted that he did not grow up with technology, indicating a usage barrier related to limited digital familiarity, and expressed a tradition barrier by preferring to arrange appointments by phone and speak with the physician face-to-face rather than use digital channels: “But we did not grow up with the computer. I would rather make a phone call to arrange an appointment and prefer to talk face-to-face to the physician” [33]. Another participant questioned the adequacy of digital encounters for a proper clinical assessment, reflecting an image or quality concern that co-occurred with a tradition-related preference for face-to-face care and an implied need for greater diagnostic assurance (risk): “I would rather that the doctor can actually touch me, examine me with a stethoscope… I also think in-person communication is sometimes better…” [42]. Together, these patterns suggest that resistance is rarely attributable to a single factor; rather, studies frequently report clusters of functional and psychological barriers that co-occur. These recurring clusters informed the relational integration step; linkages were coded as explicit when directly stated in study results or participant quotes, inferential when implied through within-study co-occurrence and narrative context, and integrative when synthesized across multiple studies showing consistent patterns.

As part of the relational integration step of our synthesis (Stage 5), we developed a conceptual model that integrates the identified barriers into an interconnected structure (Figure 3). This conceptual integration was undertaken to move beyond listing individual barriers and to summarize recurring co-occurrence patterns observed across the studies. The interconnected nature of resistance barriers creates a self-reinforcing reaction cycle that leads to avoidance behaviors among older adults. Rather than operating in isolation, functional and psychological barriers interact dynamically, compounding resistance and entrenching disengagement from digital health platforms.

Figure 3. Conceptual model of interconnected resistance barriers leading to digital health avoidance among older adults in primary care, interacting co-occurrence patterns across included studies to illustrate directional relationships and feedback loops.

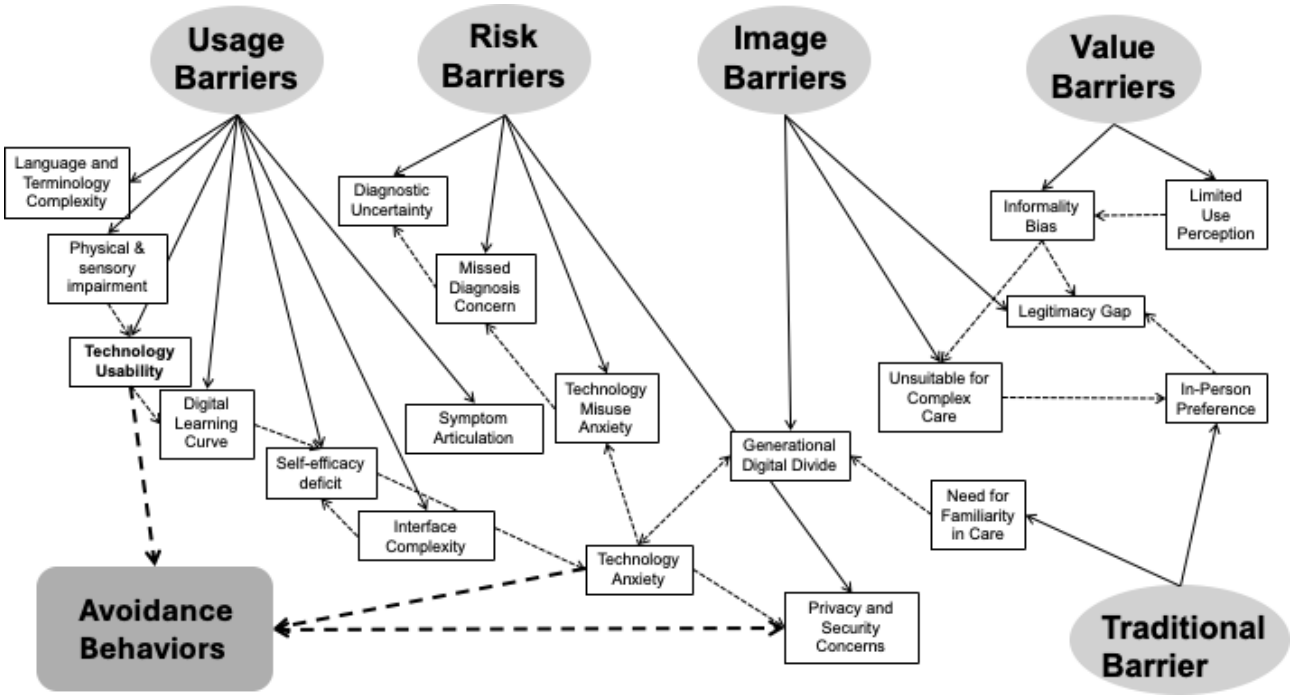

Technology usability challenges contribute to difficulties in the digital learning curve, which, along with interface complexity, results in a self-efficacy deficit and a lack of confidence in using digital health technologies. This diminished self-efficacy further fuels technology anxiety, increasing hesitation and discouraging engagement. Importantly, these usability issues do not just reduce confidence; they initiate a cascade of psychological barriers that elevate emotional discomfort and cognitive overload. Figure 3 illustrates this cascading effect: a feedback system where usability problems initiate low self-efficacy, which in turn escalates into technology anxiety. This psychological discomfort amplifies risk perceptions, including fear of misdiagnosis, privacy breaches, and technology misuse. These concerns reduce trust in digital health care solutions and reinforce avoidance behaviors. Privacy and security concerns and technology anxiety reinforce each other, creating a cycle of distrust. As the trust in the system diminishes, older adults become less likely to interact with digital platforms, which

XSL•FO
RenderX

limits exposure and impedes skill acquisition, further deepening their self-efficacy deficit. This cycle in Figure 3 is illustrated through closed feedback loops, where arrows between barriers represent how one resistance factor amplifies another (eg, Interface Complexity → Low Self-Efficacy → Technology Anxiety → Avoidance).

Traditional barriers, such as a strong preference for in-person care and the need for familiarity, also strengthen image barriers, including the legitimacy gap and the generational digital divide, further discouraging digital health adoption. As shown in Figure 3, these values-based preferences and generational perceptions reinforce internal skepticism with digital tools, particularly when technology is perceived as impersonal. The legitimacy gap reflects older adults' perception that digital tools lack the authenticity and authority of face-to-face care, while the generational divide reinforces feelings of exclusion from technologies perceived as designed for younger users. Figure 3 also highlights this convergence between identity-based resistance (eg, tradition/image) and capability-based resistance (eg, usability, anxiety). Together, these interrelated barriers form a self-reinforcing loop, where initial usability difficulties and emotional skepticism amplify resistance, which leads to withdrawal from digital health use entirely (Figure 3).

Discussion

Principal Findings

This scoping review applied the IRT to examine older adults' resistance to digital health technologies within primary care contexts. Across the included studies, we found consistent functional barriers (such as usability difficulties, interface complexity, and sensory or cognitive limitations) and recurrent psychological barriers (such as a preference for in-person care and concerns about the legitimacy of digital encounters), with value-related concerns (limited perceived benefit) and risk-related concerns (diagnostic uncertainty, privacy, and security worries) also prominent.

The findings suggest that resistance is not a static failure to adopt nor a passive disengagement, but rather a dynamic, emotionally embedded process. This process is shaped by the interaction of functional and psychological factors, including identity and value-related concerns, which do not operate in isolation but reinforce each other in feedback loops that entrench avoidance behaviors over time. The interplay between usability challenges, emotional discomfort, and value-based misalignment reflects the multifaceted nature of resistance in this population. Also, interrelationships indicated that capability-related barriers erode confidence and increase anxiety, while identity-related concerns reinforce distrust and preference for face-to-face care, together discouraging engagement. Linkages were categorized by evidentiary basis (explicit, inferential, integrative), supporting IRT as a useful framework for organizing and interpreting resistance patterns.

Functional barriers such as interface complexity, digital learning curves, and age-related sensory or cognitive limitations were among the most identified sources of resistance. However, their significance lies not only in their prevalence but in their role as catalysts: they often trigger negative psychological responses, including diminished self-efficacy, anxiety, and fear of error. These emotional reactions contributed to a broader sense of technological vulnerability and led to sustained disengagement, demonstrating how technical design and user experience are deeply interconnected.

Beyond usability, resistance was often rooted in symbolic and identity-related concerns. A preference for face-to-face interactions, generational beliefs regarding technology, and the desire for continuity with known providers were consistently linked to what can be described as symbolic distancing, a form of resistance grounded in perceived legitimacy and personal norms. Even where functionality improved, older adults continued to express skepticism, viewing digital tools as impersonal, exclusionary, or inappropriate for managing complex health needs. This suggests that emotional and symbolic dimensions may play a stronger influence on resistance than previously recognized.

These insights align with earlier theoretical work that repositions resistance as a dynamic, emotionally driven response process. The findings support an evolving theoretical perspective that frames resistance as an active process. Rather than being the inverse of adoption, resistance emerges from distinct cognitive and emotional pathways and may dominate decision-making even in the presence of positive attitudes [49]. Other research has also shown that tradition and identity-based concerns frequently outweigh usability considerations in shaping innovation rejection, particularly in service-oriented settings [20]. This review affirms that older adults' resistance is rarely due to a lack of awareness or rational evaluation alone, but rather reflects deeply embedded emotional and symbolic stances.

Breaking these loops requires targeted interventions that not only simplify interface design but also rebuild self-efficacy, trust, and the perceived legitimacy of digital care. Accordingly, programs should pair practical usability supports (eg, task simplification, assisted-digital options, scaffolded practice) with psychological strategies (eg, anxiety reduction, trust-building, culturally and linguistically responsive framing).

Comparison to Prior Work

The findings of this review both align with and challenge established models of technology acceptance. For instance, it complements the critiques of the extended UTAUT, which has been applied to prior studies involving older adults in health care settings. One study has highlighted effort expectancy, perceived usefulness, and trust in health care providers as primary predictors of adoption. While these factors remain relevant, this review suggests they are insufficient to fully account for persistent resistance observed in older populations. This resistance appears to stem not from a lack of understanding but from deeper emotional and symbolic misalignments between digital tools and the users' personal values, care routines, or generational identities [50]. In this context, resistance is not a knowledge deficit but a deliberate, emotionally grounded response to perceived risks, impersonality, or social exclusion. Our synthesis clarifies how such misalignments link to concrete pathways (eg, usability → low self-efficacy → anxiety → avoidance), adding a mechanism to prior critiques.

XSL•FO
RenderX

Reinhardt et al [51] claim in their study that resistance to innovation is not merely the opposite of adoption but a distinct phenomenon that operates through its own logic and dynamics, and thus warrants a separate theoretical approach. They proposed the concept of "adoption triggers," external events or contextual changes that interrupt entrenched resistance and enable eventual uptake. This finding aligns with the results of this review, where participants continued to resist engagement even after usability improvements, suggesting that design enhancements alone are insufficient [51]. Psychosocial catalysts such as trust in providers, alignment with identity, or significant life transitions may be necessary to shift deeply embedded resistance patterns.

Further support comes from the argument that TAM and UTAUT, widely used models, were not originally developed for health care but rather in organizational contexts. Like IRT, they were formulated outside the health domain and may require adaptation when applied in complex settings, such as digital health for older adults. In their original formulations, these models assume that perceived usefulness and ease of use directly predict technology acceptance. However, in health care, these assumptions are challenged, especially in the context of older adult users [52]. Health care studies often have to add context-specific variables such as computer anxiety, trust, or physician endorsement to increase explanatory power. This suggests that existing models may benefit from complementary perspectives that foreground resistance shaped by emotional discomfort and identity-related concerns, including symbolic dissonance around how digital health fits with older adults' roles and expectations. This review affirms the need to view resistance among older adults as socially embedded and identity-relevant, rather than reducible to issues of usability or cognitive evaluation.

Resistance constructs are not intended to replace established acceptance models such as TAM and UTAUT, but to extend them and provide a more complete account of older adults' technology use and nonuse patterns. Yu et al [53], in their research, also extend UTAUT with aging-specific variables such as perceived physical condition, self-actualization needs, and technology anxiety. Their empirical study among Chinese older adults found that while traditional UTAUT predictors (eg, performance and effort expectancy) remain significant, behavioral use was also shaped by perceived physical limitations and psychological needs for self-fulfillment. Notably, the effect of technology anxiety was nonsignificant, suggesting that usability alone does not explain resistance; rather, broader psychosocial and experiential factors must be considered [53]. These adaptations have introduced constructs such as perceived physical condition, self-actualization needs, and psychosocial well-being to better explain behavioral engagement with health care conversational agents among older adults. Our mapping complements these extensions by locating these constructs within the IRT domains and by indicating which inter-barrier links are explicitly supported by the literature.

Theoretical Implications

This review advances theory on digital health adoption and resistance among older adults in 2 main ways. First, it refines IRT for the context of aging and digital health by highlighting aging-specific resistance themes such as legitimacy gaps, generational digital divides, and anxiety about technology misuse as candidates for further conceptualization and measurement within the original IRT domains. Second, it points to resistance as a dynamic process in which these barriers interact in feedback patterns rather than operating as isolated categories. This mechanism-oriented view complements existing TAMs by underscoring that persistent nonuse reflects active, emotionally and symbolically shaped resistance, rather than merely weak adoption intentions.

Practical Implications

From a gerontechnology and age-inclusive design perspective, the IRT-based model translates the identified barriers and linkages into actionable design and implementation levers to reduce resistance among older adults in primary care. This review has important implications for digital health design, practice, and policy.

First, the disproportionate concentration of extant research within high-income Western countries necessitates a nuanced approach to global implementation, as resistance profiles are not homogenous but are contingent upon divergent socioeconomic structures, varying levels of digital literacy, and culture-specific perceptions of aging [54]. Addressing these complexities requires a paradigm shift from a reactive model, characterized by a narrow focus on technical troubleshooting and interface simplification, toward a proactive design. While mitigating interface complexity and accommodating sensory impairments remain fundamental requirements, such technical refinements in isolation are insufficient to resolve resistance that is fundamentally anchored in emotional and psychological factors. Consequently, proactive age-tech development should prioritize the alignment of digital interventions with users' long-standing traditions and the preservation of relational continuity in care [55]. By acknowledging traditional barriers and framing digital tools as seamless extensions of familiar, trusted care routines rather than disruptive innovations, developers can transition from delivering impersonal technical products to co-creating solutions that resonate with the core identities and values of older populations.

Building on the conceptual model in Figure 2, breaking the self-reinforcing cycle of resistance requires targeted interventions that address both practical usability barriers and underlying psychological resistance; focusing on interface design or digital literacy alone is unlikely to change deeply rooted patterns of nonuse. Designers need to focus not only on functionality but also on providing emotional reassurance and strengthening the perceived legitimacy and social meaning of digital care. Therefore, solutions should be co-designed with older adults not only to ensure they fit with their routines, communication styles, and cultural values, but also to directly address the specific IRT barriers identified in this review by incorporating strategies that reduce friction and promote confidence. These strategies may include simplifying high-friction tasks by using shorter flows, fewer required fields, larger tap targets, and accessible defaults. Also, designers can provide stepwise guidance and "practice mode," and offer assisted-digital options such as telephone call-back support,

XSL•FO
RenderX

shared on-screen navigation with staff, and on-site digital stations within clinics where staff can help patients complete digital tasks.

Privacy, risk perceptions, and distrust emerged as central barriers in our synthesis. Digital health platforms should incorporate trust-enhancing features, including sustained relationships with known providers, easy access to human support, and clear, simple explanations of data practices. To strengthen perceived legitimacy, systems should preserve care delivery choice (seamless switch to phone or in-person visits), display continuity cues (named clinician, photo, prior encounters), and surface concrete benefits (time saved, refill accuracy, faster appointments). Culturally and linguistically responsive content, combined with feedback that reinforces mastery, can further mitigate anxiety and improve self-efficacy, helping to disrupt the self-reinforcing loops that lead to avoidance. Together, these design-oriented recommendations translate our conceptual findings into practical guidance for technology designers and implementers seeking to reduce resistance among older adults.

Future Research Directions

Future research should investigate the temporal evolution of resistance, including how initial avoidance may shift or diminish over time, and under what conditions. There is a need to explore resistance dynamics among underrepresented populations, such as ethnic minorities, linguistically diverse groups, and individuals living in lower-resource settings. In line with Bevilacqua et al [56], emerging work on service-specific acceptance measures for older adults who developed the Robot-Era Inventory as a tailored acceptance scale for a social robotics platform, and called for customizable, context-specific tools tailored to specific technologies and services for older adults, future studies should develop and validate IRT-informed scales tailored to particular digital health modalities [56]. In addition, longitudinal and mixed-methods designs could provide deeper insight into how resistance is maintained or disrupted. Finally, the development and empirical testing of interventions grounded in IRT would help bridge the gap between theory and design strategies.

Strengths and Limitations

A key strength of this review is its structured, theory-driven synthesis across diverse empirical studies. By applying the IRT to various study designs and health care contexts, this review enhances the conceptual understanding of digital resistance among older adults. It was conducted according to best-practice guidelines for scoping reviews, which reflect established methodological standards.

Several limitations should be noted. First, the search was restricted to English-language publications, which might have excluded relevant studies published in other languages. Second, the review encompasses studies published between 2014 and 2025, a period characterized by rapid technological advancement. Improvements in device usability during this time may have influenced user experiences and patterns of resistance, potentially affecting cross-study comparability. Third, most of the included studies were conducted in high-income Western countries, and the patterns of resistance identified here may not fully capture experiences in lower-income or non-Western contexts, where digital infrastructures, health systems, and cultural norms around aging and technology may differ substantially. This concentration substantially reduces generalizability beyond high-income Western settings and limits the applicability of our findings to global contexts where digital literacy, socioeconomic factors, and cultural perceptions of aging and health care may create distinct resistance profiles. Fourth, none of the included studies reported participants' cognitive status or used standardized cognitive screening measures. As a result, we could not examine whether resistance barriers vary by cognitive integrity or distinguish attitudinal resistance from barriers related to cognitive impairment, which may influence learnability, confidence, and sustained use of digital health technologies. Finally, the proposed conceptual model has not yet been validated in practice and should be regarded as hypothesis-generating. Future research should operationalize the IRT domains and evaluate their factor structure, reliability, and predictive validity in empirical studies.

Conclusions

Applying IRT to older adults' experiences with digital health shifts the focus from "lack of readiness" or skills gaps to resistance mechanisms and how technologies are designed and integrated into primary care. Resistance emerges as an active, emotionally rooted process involving functional, psychological, and identity-based barriers to adoption, and this review integrates recurring co-occurrence patterns into a conceptual model, thereby moving beyond prior work that lists barriers in isolation. The synthesis clarifies how usability problems can undermine self-efficacy, increase technology anxiety, and amplify trust and legitimacy concerns, creating feedback loops that reinforce avoidance. Real-world implications: implementation strategies should go beyond technical usability by rebuilding emotional trust, supporting relational continuity, and aligning digital solutions with older adults' values and routines through meaningful channel choice and transparent communication about risks. In addition, IRT offers a structure for developing domain-specific measures and interventions that address usage, value, risk, tradition, and image barriers, supporting a more realistic and equitable digital transformation in primary care for aging populations.

Data Availability

The datasets generated and analyzed during this study are reported in the article and multimedia appendix.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Online search strategy.
[DOCX File, 18 KB - jmir_v28i1e75591_app1.docx]

Checklist 1
PRISMA-S Checklist.
[DOCX File, 17 KB - jmir_v28i1e75591_app2.docx]

Checklist 2
Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist.
[DOCX File, 56 KB - jmir_v28i1e75591_app3.docx]

References

1. Chen C, Ding S, Wang J. Digital health for aging populations. Nat Med 2023 Jul;29(7):1623-1630. [doi: 10.1038/s41591-023-02391-8] [Medline: 37464029]
2. Mitchell M, Kan L. Digital technology and the future of health systems. Health Syst Reform 2019;5(2):113-120. [doi: 10.1080/23288604.2019.1583040] [Medline: 30908111]
3. Stoumpos AI, Kitsios F, Talias MA. Digital transformation in healthcare: technology acceptance and its applications. Int J Environ Res Public Health 2023 Feb 15;20(4):3407. [doi: 10.3390/ijerph20043407] [Medline: 36834105]
4. Alruwaili MM, Shaban M, Elsayed Ramadan OM. Digital health interventions for promoting healthy aging: a systematic review of adoption patterns, efficacy, and user experience. Sustainability 2023 Dec 2;15(23):16503. [doi: 10.3390/su152316503]
5. Jokisch MR, Schmidt LI, Doh M, Marquard M, Wahl HW. The role of internet self-efficacy, innovativeness and technology avoidance in breadth of internet use: comparing older technology experts and non-experts. Comput Human Behav 2020 Oct;111:106408. [doi: 10.1016/j.chb.2020.106408]
6. Dhagarra D, Goswami M, Kumar G. Impact of trust and privacy concerns on technology acceptance in healthcare: an Indian perspective. Int J Med Inform 2020 Sep;141:104164. [doi: 10.1016/j.ijmedinf.2020.104164] [Medline: 32593847]
7. Zhou J, Salvendy G, Boot WR, et al. Grand challenges of smart technology for older adults. International Journal of Human–Computer Interaction 2025:1-43. [doi: 10.1080/10447318.2025.2457003]
8. Rahimi B, Nadri H, Lotfnezhad Afshar H, Timpka T. A systematic review of the technology acceptance model in health informatics. Appl Clin Inform 2018 Jul;9(3):604-634. [doi: 10.1055/s-0038-1668091] [Medline: 30112741]
9. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
10. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view1. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
11. Rogers EM. Lessons for guidelines from the diffusion of innovations. Jt Comm J Qual Improv 1995 Jul;21(7):324-328. [doi: 10.1016/s1070-3241(16)30155-9] [Medline: 7581733]
12. Lee C, Coughlin JF. PERSPECTIVE: Older adults' adoption of technology: an integrated approach to identifying determinants and barriers. J of Product Innov Manag 2015 Sep;32(5):747-759 [FREE Full text] [doi: 10.1111/jpim.12176]
13. Wilson J, Heinsch M, Betts D, Booth D, Kay-Lambkin F. Barriers and facilitators to the use of e-health by older adults: a scoping review. BMC Public Health 2021 Aug 17;21(1):1556. [doi: 10.1186/s12889-021-11623-w] [Medline: 34399716]
14. Cajita MI, Hodgson NA, Lam KW, Yoo S, Han HR. Facilitators of and barriers to mHealth adoption in older adults with heart failure. Comput Inform Nurs 2018 Aug;36(8):376-382. [doi: 10.1097/CIN.0000000000000442] [Medline: 29742549]
15. Fox G, Connolly R. Mobile health technology adoption across generations: narrowing the digital divide. Information Systems Journal 2018 Nov;28(6):995-1019 [FREE Full text] [doi: 10.1111/isj.12179]
16. Russell TG, Gillespie N, Hartley N, Theodoros D, Hill A, Gray L. Exploring the predictors of home telehealth uptake by elderly Australian healthcare consumers. J Telemed Telecare 2015 Dec;21(8):485-489. [doi: 10.1177/1357633X15606264] [Medline: 26391512]
17. Mumtaz H, Riaz MH, Wajid H, et al. Current challenges and potential solutions to the use of digital health technologies in evidence generation: a narrative review. Front Digit Health 2023;5:1203945. [doi: 10.3389/fdgth.2023.1203945] [Medline: 37840685]
18. Hoque R, Sorwar G. Understanding factors influencing the adoption of mHealth by the elderly: an extension of the UTAUT model. Int J Med Inform 2017 May;101:75-84. [doi: 10.1016/j.ijmedinf.2017.02.002]
19. Ram S, Sheth JN. Consumer resistance to innovations: the marketing problem and its solutions. J Consum Mark 1989 Feb 1;6(2):5-14. [doi: 10.1108/EUM0000000002542]
20. Laukkanen T. Consumer adoption versus rejection decisions in seemingly similar service innovations: the case of the internet and mobile banking. J Bus Res 2016 Jul;69(7):2432-2439. [doi: 10.1016/j.jbusres.2016.01.013]

XSL•FO
RenderX

21. Laukkanen P, Sinkkonen S, Laukkanen T. Consumer resistance to internet banking: postponers, opponents and rejectors. International Journal of Bank Marketing 2008 Sep 5;26(6):440-455. [doi: 10.1108/02652320810902451]
22. Kaur P, Dhir A, Ray A, Bala PK, Khalil A. Innovation resistance theory perspective on the use of food delivery applications. Journal of Enterprise Information Management 2021 Nov 11;34(6):1746-1768. [doi: 10.1108/JEIM-03-2020-0091]
23. Castro CAB, Zambaldi F, Ponchio MC. Cognitive and emotional resistance to innovations: concept and measurement. JPBM 2019 Jul 4;29(4):441-455. [doi: 10.1108/JPBM-10-2018-2092]
24. Cieslak V, Valor C. Moving beyond conventional resistance and resistors: an integrative review of employee resistance to digital transformation. Cogent Business & Management 2025 Dec 12;12(1). [doi: 10.1080/23311975.2024.2442550]
25. Alshallaqi M, Al Halbusi H, Abbas M, Alhaidan H. Resistance to innovation in low-income populations: the case of university students' resistance to using digital productivity applications. Front Psychol 2022;13:961589. [doi: 10.3389/fpsyg.2022.961589] [Medline: 36275207]
26. Spinelli G, Gastaldi L, Van Hove L, Van Droogenbroeck E. Can cluster analysis enrich the innovation resistance theory? the case of mobile payment usage in Italy. Technol Soc 2024 Dec;79:102729. [doi: 10.1016/j.techsoc.2024.102729]
27. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
28. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5(1):69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
29. The Joanna Briggs Institute reviewers' manual. : The Joanna Briggs Institute; 2015 URL: https://reben.com.br/revista/wp-content/uploads/2020/10/Scoping.pdf [accessed 2026-01-28]
30. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
31. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
32. Khanassov V, Ilali M, Ruiz AS, Rojas-Rozo L, Sourial R. Telemedicine in primary care of older adults: a qualitative study. BMC Prim Care 2024 Jul 17;25(1):259. [doi: 10.1186/s12875-024-02518-x] [Medline: 39020277]
33. Vergouw JW, Smits-Pelser H, Kars MC, et al. Needs, barriers and facilitators of older adults towards eHealth in general practice: a qualitative study. Prim Health Care Res Dev 2020 Dec 2;21:e54. [doi: 10.1017/S1463423620000547] [Medline: 33263272]
34. Knotnerus HR, Ngo HTN, Maarsingh OR, van Vugt VA. Understanding older adults' experiences with a digital health platform in general practice: qualitative interview study. JMIR Aging 2024 Aug 30;7:e59168. [doi: 10.2196/59168] [Medline: 39212599]
35. Bhatia R, Gilliam E, Aliberti G, et al. Older adults' perspectives on primary care telemedicine during the COVID-19 pandemic. J Am Geriatr Soc 2022 Dec;70(12):3480-3492. [doi: 10.1111/jgs.18035] [Medline: 36169152]
36. Lam K, Lu AD, Shi Y, Covinsky KE. Assessing telemedicine unreadiness among older adults in the united states during the COVID-19 pandemic. JAMA Intern Med 2020 Oct 1;180(10):1389-1391. [doi: 10.1001/jamainternmed.2020.2671] [Medline: 32744593]
37. Nymberg VM, Bolmsjö BB, Wolff M, Calling S, Gerward S, Sandberg M. "Having to learn this so late in our lives…" Swedish elderly patients' beliefs, experiences, attitudes and expectations of e-health in primary health care. Scand J Prim Health Care 2019 Mar;37(1):41-52. [doi: 10.1080/02813432.2019.1570612] [Medline: 30732519]
38. van Houwelingen CT, Ettema RG, Antonietti MG, Kort HS. Understanding older people's readiness for receiving telehealth: mixed-method study. J Med Internet Res 2018 Apr 6;20(4):e123. [doi: 10.2196/jmir.8407] [Medline: 29625950]
39. Laukka E, Lakoma S, Harjumaa M, et al. Older adults' preferences in the utilization of digital health and social services: a qualitative analysis of responses to open-ended questions. BMC Health Serv Res 2024 Oct 4;24(1):1184. [doi: 10.1186/s12913-024-11564-1] [Medline: 39367429]
40. Rochmawati E, Kamilah F, Iskandar AC. Acceptance of e-health technology among older people: a qualitative study. Nurs Health Sci 2022 Jun;24(2):437-446. [doi: 10.1111/nhs.12939] [Medline: 35297152]
41. Hunsbedt Fjellså HM, Husebø AML, Braut H, Mikkelsen A, Storm M. Older adults' experiences with participation and eHealth in care coordination: qualitative interview study in a primary care setting. J Particip Med 2023 Oct 2;15:e47550. [doi: 10.2196/47550] [Medline: 37782538]
42. Mao A, Tam L, Xu A, et al. Barriers to telemedicine video visits for older adults in independent living facilities: mixed methods cross-sectional needs assessment. JMIR Aging 2022 Apr 19;5(2):e34326. [doi: 10.2196/34326] [Medline: 35438648]
43. Frishammar J, Essén A, Bergström F, Ekman T. Digital health platforms for the elderly? key adoption and usage barriers and ways to address them. Technol Forecast Soc Change 2023 Apr;189:122319. [doi: 10.1016/j.techfore.2023.122319]
44. Haimi M, Goren U, Grossman Z. Barriers and challenges to telemedicine usage among the elderly population in Israel in light of the COVID-19 era: a qualitative study. Digit Health 2024;10:20552076241240235. [doi: 10.1177/20552076241240235] [Medline: 38550265]
45. Landgren S, Cajander Å. Non-use of digital health consultations among Swedish elderly living in the countryside. Front Public Health 2021;9:588583. [doi: 10.3389/fpubh.2021.588583] [Medline: 34568247]

XSL•FO
RenderX

46. Ahmed N, Hall A, Poku B, McDermott J, Astbury J, Todd C. Experiences and views of older adults of South Asian, Black African, and Caribbean backgrounds about the digitalization of primary care services since the COVID-19 pandemic: qualitative focus group study. JMIR Form Res 2024 Dec 18;8:e57580. [doi: 10.2196/57580] [Medline: 39693146]
47. Ufholz K, Sheon A, Bhargava D, Rao G. Telemedicine preparedness among older adults with chronic illness: survey of primary care patients. JMIR Form Res 2022 Jul 27;6(7):e35028. [doi: 10.2196/35028] [Medline: 35896013]
48. Sproul A, Stevens J, Richard J. Older adults' use of and interest in technology and applications for health management: a survey study. Can J Hosp Pharm 2023;76(3):209-215. [doi: 10.4212/cjhp.3261] [Medline: 37409153]
49. Claudy MC, Garcia R, O'Driscoll A. Consumer resistance to innovation—a behavioral reasoning perspective. J of the Acad Mark Sci 2015 Jul;43(4):528-544. [doi: 10.1007/s11747-014-0399-0]
50. Cimperman M, Makovec Brenčič M, Trkman P. Analyzing older users' home telehealth services acceptance behavior-applying an extended UTAUT model. Int J Med Inform 2016 Jun;90:22-31. [doi: 10.1016/j.ijmedinf.2016.03.002] [Medline: 27103194]
51. Reinhardt R, Hietschold N, Gurtner S. Overcoming consumer resistance to innovations – an analysis of adoption triggers. R & D Management 2019 Mar;49(2):139-154 [FREE Full text] [doi: 10.1111/radm.12259]
52. Ammenwerth E. Technology acceptance models in health informatics: TAM and UTAUT. Stud Health Technol Inform 2019 Jul 30;263(264):64-71. [doi: 10.3233/SHTI190111] [Medline: 31411153]
53. Yu S, Chen T. Understanding older adults' acceptance of chatbots in healthcare delivery: an extended UTAUT model. Front Public Health 2024;12:1435329. [doi: 10.3389/fpubh.2024.1435329] [Medline: 39628811]
54. Badr J, Motulsky A, Denis JL. Digital health technologies and inequalities: a scoping review of potential impacts and policy recommendations. Health Policy 2024 Aug;146:105122. [doi: 10.1016/j.healthpol.2024.105122] [Medline: 38986333]
55. Ladds E, Khan M, Moore L, Kalin A, Greenhalgh T. The impact of remote care approaches on continuity in primary care: a mixed-studies systematic review. Br J Gen Pract 2023 May;73(730):e374-e383. [doi: 10.3399/BJGP.2022.0398] [Medline: 37105731]
56. Bevilacqua R, Di Rosa M, Riccardi GR, et al. Design and development of a scale for evaluating the acceptance of social robotics for older people: the robot era inventory. Front Neurorobot 2022;16:883106. [doi: 10.3389/fnbot.2022.883106] [Medline: 35874107]

Abbreviations:

IRT: innovation resistance theory
JBI: Joanna Briggs Institute
mHealth: mobile health
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension checklist
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
TAM: technology acceptance model
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by S Brini; submitted 07.Apr.2025; peer-reviewed by G Amabili, M Pinelli; revised version received 30.Dec.2025; accepted 30.Dec.2025; published 02.Feb.2026.

Please cite as:
Birati Y, Tzemah-Shahar R
Barriers to Digital Health Adoption in Older Adults: Scoping Review Informed by Innovation Resistance Theory
J Med Internet Res 2026;28:e75591
URL: https://www.jmir.org/2026/1/e75591
doi:10.2196/75591

© Yosefa Birati, Roy Tzemah-Shahar. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Machine Learning Techniques Used for the Identification of Sociodemographic Factors Associated With Cancer: Systematic Literature Review

Liz González-Infante[1,2*], MHRM; Gaston Marquez[2,3*], PhD; Solange Parra-Soto[2,4*], PhD; Mónica Cardona-Valencia[5*], PhD; Carla Taramasco[2,6*], PhD

[1]Facultad de Ciencias Empresariales, Universidad del Bío-Bío, Andrés Bello 720, Chillán, Chile

[2]Centro para la Prevención y el Control del Cáncer, Santiago, Chile

[3]Departamento de Ciencias de la Computación y Tecnologías de la Información, Facultad de Ciencias Empresariales, Universidad del Bío-Bío, Chillan, Chile

[4]Departamento de Nutrición y Salud Pública, Facultad Ciencias de la Salud y de los Alimentos, Universidad del Bío-Bío, Chillán, Chile

[5]Departamento Ciencias de la Rehabilitación en Salud, Facultad de Ciencias de la Salud y de los Alimentos, Universidad del Bío-Bío, Chillán, Chile

[6]ITISB, Facultad de Ingeniería, Universidad Andrés Bello, Viña del Mar, Chile

[*]all authors contributed equally

Corresponding Author:
Liz González-Infante, MHRM
Facultad de Ciencias Empresariales, Universidad del Bío-Bío, Andrés Bello 720, Chillán, Chile

Abstract

Background: Cancer remains one of the foremost global causes of mortality, with nearly 10 million deaths recorded by 2020. As incidence rates rise, there is a growing interest in leveraging machine learning (ML) to enhance prediction, diagnosis, and treatment strategies. Despite these advancements, insufficient attention has been directed toward the integration of sociodemographic variables, which are crucial determinants of health equity, into ML models in oncology.

Objective: This review aims to investigate how ML techniques have been used to identify patterns of predictive association between sociodemographic factors and cancer-related outcomes. Specifically, it seeks to map current research endeavors by detailing the types of algorithms used, the sociodemographic variables examined, and the validation methodologies used.

Methods: We conducted a systematic literature review in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. Searches were executed across 6 databases, focusing on the primary studies using ML to investigate the association between sociodemographic characteristics and cancer-related outcomes. The search strategy was informed by the PICO (population, intervention, comparison, and outcome) framework, and a set of predefined inclusion criteria was used to screen the studies. The methodological quality of each included paper was assessed.

Results: Out of the 328 records examined, 19 satisfied the inclusion criteria. The majority of studies used supervised ML techniques, with random forest and extreme gradient boosting being the most commonly used. Frequently analyzed variables include age, male or female or intersex, education level, income, and geographic location. Cross-validation is the predominant method for evaluating model performance. Nevertheless, the integration of clinical and sociodemographic data is limited, and efforts toward external validation are infrequent.

Conclusions: ML holds significant potential for discerning patterns associated with the social determinants of cancer. Nevertheless, research in this domain remains fragmented and inconsistent. Future investigations should prioritize the integration of contextual factors, enhance model transparency, and bolster external validation. These measures are crucial for the development of more equitable, generalizable, and actionable ML applications in cancer care.

(J Med Internet Res 2026;28:e79187) doi:10.2196/79187

KEYWORDS

cancer; health disparities; machine learning; predictive models; social determinants of health; sociodemographic factors; systematic review

XSL•FO
RenderX

Introduction

The use of machine learning (ML) in oncology has advanced significantly over the past decade, offering new opportunities for early detection, survival prediction, and treatment personalization. Models based on techniques such as random forests (RFs), extreme gradient boosting (XGBoost), and deep neural networks have demonstrated remarkable performance across different types of cancer, fueling enthusiasm for what has been termed digital precision oncology [1]. However, most of these applications rely almost exclusively on clinical and biomedical data, limiting their ability to capture the broader social and structural factors that shape health outcomes [2]. This gap raises important concerns, as it may compromise both the external validity and the equity of ML models. In this review, we consistently use the term sociodemographic factors to refer to variables such as age, male or female or intersex, educational attainment, income, ethnicity, rurality, and access to health care. These factors conceptually overlap with the broader category of social determinants of health (SDoH), but our focus is on those variables that are typically available in clinical and research datasets and are explicitly integrated into ML models. By doing so, we ensure clarity and terminological consistency throughout the paper.

Our review focuses on the most common sociodemographic variables in clinical and research datasets, such as age, male or female or intersex, education, income, and others, reflecting the current landscape of published ML studies rather than a deliberate theoretical choice. We recognize that these indicators only capture part of the social gradient influencing cancer outcomes. Therefore, we highlight the importance of future research integrating contextual and multilevel determinants, such as neighborhood characteristics, health care infrastructure, environmental exposures, and political factors, to promote an equity-centered approach to ML applications in oncology.

In parallel, the rise of explainable artificial intelligence (AI) has highlighted the importance of transparency and interpretability in clinical settings. Tools such as Shapley Additive Explanations and local interpretable model-agnostic explanations allow health care professionals to better understand ML models by identifying which variables are most relevant in predictions and how they interact with both clinical and sociodemographic factors [3]. These advances not only strengthen trust in ML-based systems but also enhance their potential for integration into clinical practice and public health policy [4]. The convergence of explainable AI and SDoH emerges as a promising pathway toward developing fairer and more actionable models.

Nevertheless, our review of the literature reveals that although research and reviews on ML in oncology are rapidly expanding, most have concentrated on methodological, genomic, or clinical aspects without adequately addressing sociodemographic factors. This omission limits the ability of the scientific community to develop robust guidelines for implementing models across diverse contexts and health systems. Against this backdrop, this study aimed to identify, characterize, and synthesize primary research that applied ML methods to analyze sociodemographic factors associated with cancer. The objective was to address both methodological and conceptual gaps while contributing to the development of fairer and more transparent models that can inform data-driven public health strategies. We present the results of a systematic literature review (SLR) examining how ML techniques have been used to identify and interpret sociodemographic factors in cancer-related studies. Of the 328 papers screened, 19 (5.8%) met the inclusion criteria. Rather than being a limitation, this number reflects the emerging nature of the field and highlights the value of conducting an early review to consolidate initial progress, make methodological and equity-related gaps more visible, and guide future research toward a stronger integration of sociodemographic factors in ML models applied to oncology.

Methods

Research Questions

Based on the main objective, we defined the following research questions:

1. What ML techniques have been applied in studies that analyze sociodemographic data of patients with cancer to identify factors associated with the disease?
2. What sociodemographic factors have been consistently identified as relevant to the diagnosis, progression, or treatment of cancer?

Identification

The SLR was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Checklist 1), which provide a rigorous framework for ensuring transparency and reproducibility in evidence synthesis [5]. To guide the construction of the search strategy, we also adopted the PICO (population, intervention, comparison, and outcome) model, as recommended by Petersen et al [6]. This framework allowed us to clearly define the target population, specify the type of intervention (ie, application of ML techniques), and focus the outcome on the identification of relevant sociodemographic factors associated with cancer (Table 1).

XSL•FO
RenderX

Table . Keywords used in the PICO (population, intervention, comparison, and outcome) structure.

Component	Description	Keywords
Population	Studies analyzing data from patients with cancer that include sociodemographic variables. These may encompass age, male or female or intersex, socioeconomic status, education, and residence among others.	"Sociodemographic factors," "social determinants," "sociodemographic characteristics," and "socio-demographic variables"
Intervention	Application of machine learning techniques to identify and analyze sociodemographic factors associated with cancer.	"Machine learning" and "artificial intelligence"
Comparison	No previous studies with similar scope and objectives were identified as suitable comparators. This review explores a novel approach.	Not applicable
Outcome	Identification of the most relevant sociodemographic variables associated with cancer outcomes, and assessment of the predictive performance of the applied machine learning models.	"Cancer," "oncology," variable importance, model accuracy, and AUC[a]

[a]AUC: area under the curve.

The search terms were combined using the Boolean operators AND and OR to ensure comprehensive retrieval of relevant literature. The final search string was as follows:

(["sociodemographic factors" OR "socio-demographic factors" OR "sociodemographic characteristics" OR "socio-demographic characteristics" OR "social determinants" OR "sociodemographic variables" OR "socio-demographic variables") AND ("machine learning" OR "artificial intelligence") AND ("cancer" OR "oncology"])

Screening

We conducted a comprehensive literature search across 6 major databases: PubMed (n=76), ACM Digital Library (n=85), ScienceDirect (n=7), IEEE Xplore (n=1), Web of Science Core Collection (n=80), and Scopus (n=79). Searches covered the period from database inception to October 14, 2024. PubMed was selected as the primary source for biomedical and oncology research. ScienceDirect was included to capture papers published in Elsevier journals not indexed elsewhere. ACM Digital Library and IEEE Xplore were used to retrieve computer science and engineering studies, where ML methods are often first reported. Web of Science facilitated interdisciplinary retrieval and citation tracking, while Scopus provided broad multidisciplinary coverage.

All records were exported, merged, and deduplicated prior to screening. To maximize comprehensiveness and minimize selection bias, we also applied forward and backward citation chasing on included studies. Full electronic search strategies for each database are provided in Multimedia Appendix 1.

Paper Selection

Eligibility Criteria

Primary studies were screened and selected based on predefined inclusion and exclusion criteria. The specific inclusion criteria applied are summarized in Textbox 1.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Type of study: primary studies presenting original data or analysis. Quantitative studies applying machine learning techniques to analyze sociodemographic factors related to cancer, including experimental, observational (cohort, case-control, and cross-sectional), or methodological designs.
- Study area: application of machine learning in health, focused on the analysis of sociodemographic factors (eg, age, male or female or intersex, ethnicity, socioeconomic status, and health care access) and their association with any type of cancer (eg, breast, lung, prostate, and gastrointestinal).
- Machine learning techniques: use of supervised algorithms (eg, neural networks, decision trees, support vector machines, and logistic regression), unsupervised (eg, clustering), or semisupervised algorithms. Reporting of performance metrics such as accuracy, sensitivity, specificity, and receiver operating characteristic area under the curve.
- Sociodemographic factors: explicit analysis of sociodemographic variables related to cancer risk, prevalence, or progression, including age, male or female or intersex, ethnicity, income, education, occupation, geographic location, health care access, and other socioeconomic determinants.
- Publication period: studies published from 2014 onward.
- Language: publications in English or Spanish.
- Accessibility: full-text access or access to essential data and results enabling methodological evaluation.

Exclusion criteria

- Type of study: systematic reviews, narrative reviews, meta-analyses, or secondary studies.
- Study area: studies not analyzing the association between sociodemographic factors and cancer. Studies focused on other diseases (eg, diabetes and cardiovascular diseases).
- Machine learning techniques: studies relying solely on traditional statistical methods and not reporting model validation metrics.
- Sociodemographic factors: studies applying machine learning without including sociodemographic variables (eg, focused only on genetic, molecular, or biological data).
- Publication period: Studies published before 2014.
- Language: publications in other languages without available translation.
- Accessibility: abstracts or conference proceedings without access to the full paper.

Quality Assessment

The purpose of the quality assessment was to evaluate the relevance of each selected paper. Although quality assessment did not influence the selection of primary studies [7], we included it primarily to reflect the validity of the selected studies. Based on the response to each research question, we scored each paper with 2, 1, or 0 points. We then selected those papers that exceeded the 50% threshold. The studies chosen through this assessment ensure that our conclusions, drawn from the extracted data, are supported by adequately resourced evidence (Multimedia Appendix 1).

Study Selection and Resolution of Discrepancies

Each paper was independently screened by 2 reviewers according to predefined inclusion and exclusion criteria. Any disagreements regarding eligibility were addressed during consensus meetings, where reviewers jointly discussed the rationale for inclusion or exclusion. When consensus could not be reached, a third author was consulted to make the final decision. This procedure ensured transparency, reproducibility, and rigor throughout the study selection process.

Results

Overview

The SLR was conducted in accordance with the PRISMA guidelines, which provide a rigorous framework for ensuring transparency and reproducibility in evidence synthesis (Figure 1). Following the PRISMA methodology, a total of 15 primary studies published in peer-reviewed journals were identified. An additional 4 papers were included through forward snowballing, yielding a final sample of 19 studies. Among these, 58% (11/19) were conducted in the United States. Iran contributed 21% (4/19), followed by India with 11% (2/19), and South Korea with 5% (1/19). One study (5%) represented a collaborative effort between institutions in China and the United States (Table 2). The publication dates of the included studies ranged from 2018 to 2024. No eligible primary studies were found in workshop proceedings or book chapters.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the selection of primary studies for the systematic literature review. N/A: not applicable.

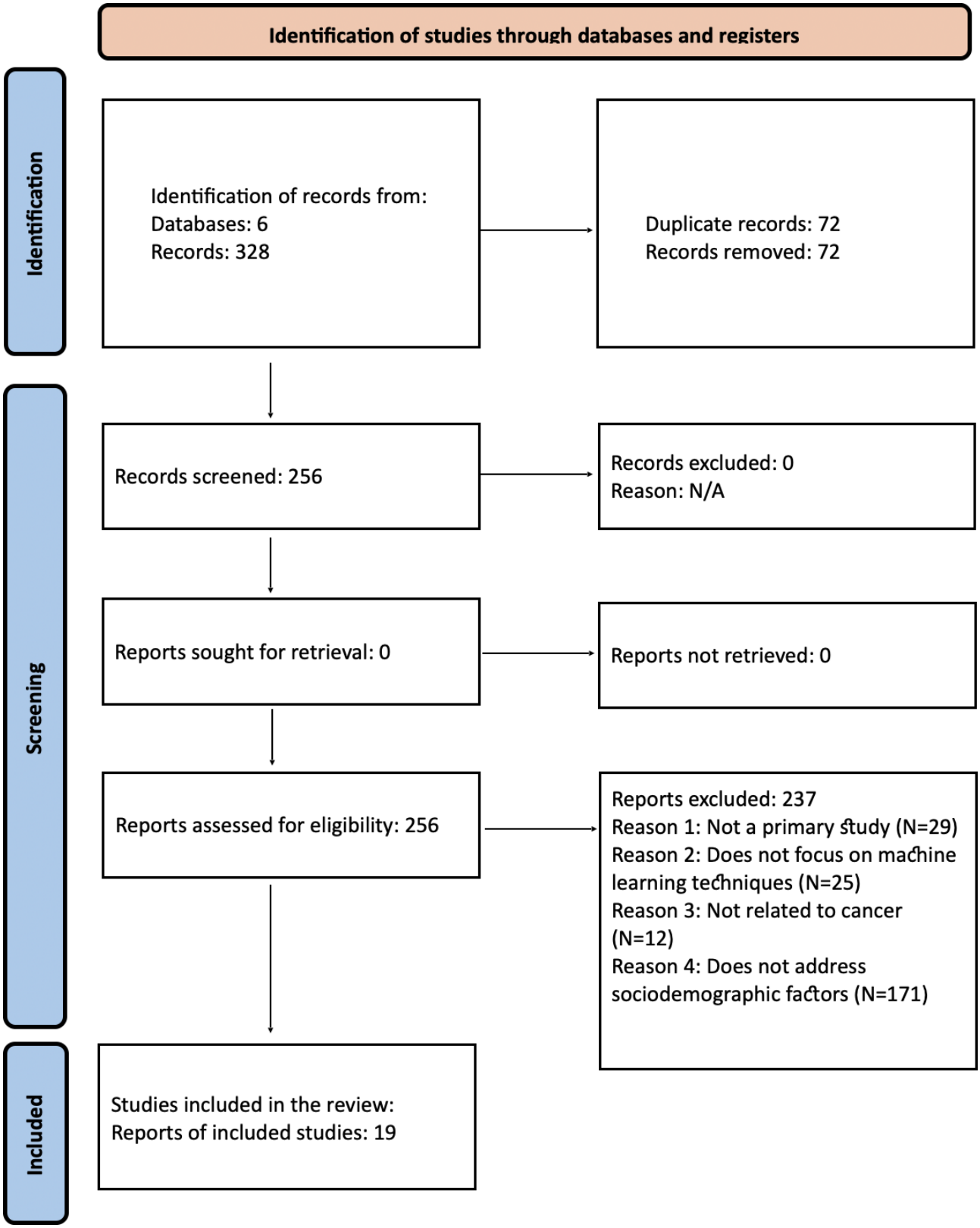

XSL•FO
RenderX

Table . Distribution of primary studies by country.

Country	Number of studies
United States	11
Iran	4
India	2
South Korea	1
China-US collaboration	1

Machine Learning Algorithms and Validation Strategies Reported

Across the studies analyzed, consistent patterns emerged in both the selection of ML algorithms and the validation methods used (Table 3).

XSL•FO
RenderX

Table . Summary of the machine learning algorithms and validation strategies reported across the 19 primary studies. Most studies applied ensemble methods such as random forest (RF) or gradient boosting, frequently combined with cross-validation schemes.

Study ID	Algorithms used	Validation strategy	Reference
S1	Lasso[a] LR[b], RF, gradient boosting, DT[c], SVM[d]	5-fold CV[e], ROC-AUC[f], accuracy, sensitivity, specificity	[8]
S2	XGBoost[g], LightGBM[h], CatBoost[i], RF, AdaBoost, Lasso regression	10-fold CV	[9]
S3	DT, RF	10-fold CV	[10]
S4	RF, artificial neural networks, bootstrap aggregating CART[j], XGBoost	10-fold CV	[11]
S5	XGBoost	10-fold CV	[12]
S6	LightGBM, XGBoost	10-fold CV	[13]
S7	RF, Neural networks, LR, XGBoost	CV, AUC, grid search	[14]
S8	RF, gradient boosting machine, SVM	5-fold CV, ROC	[15]
S9	Radiomics-signature model	No formal validation performed	[16]
S10	Multilayer perceptron, SVM, XGBoost	10-fold CV	[17]
S11	Max-p-regions, RF, Jenks natural breaks	RF VIMP[k] ranking	[18]
S12	CART, RF	Bootstrap sampling	[19]
S13	DT, RF, Boruta feature selection	Confusion matrix	[20]
S14	Bayesian additive, regression trees	Partial dependence plots, variable inclusion proportion	[21]
S15	LR, ridge classifier, SGD[l]classifier, KNN[m], DT, linear support vector classifier, support vector classifier with radial basis function kernel, Gaussian Naïve Bayes, AdaBoost classifier, RF, gradient boosting, QDA[n]	5-fold CV, LOOCV[o]	[22]
S16	Semiautomated segmentation + conditional LR	80/20 hold-out CV, ROC-AUC, Youden Index	[23]
S17	Random survival forest, Cox proportional hazards	Grid search, C-index[p]	[24]
S18	RF, SVM, gradient boosting machine	10-fold CV	[25]
S19	SVM, DT, naive Bayesian model, and KNN	10-fold CV	[26]

[a]Lasso: least absolute shrinkage and selection operator.

[b]LR: logistic regression.

[c]DT: decision tree.

[d]SVM: support vector machine.

[e]CV: cross-validation.

[f]ROC-AUC: receiver operating characteristic area under the curve.

[g]XGBoost: extreme gradient boosting.

[h]LightGBM: light gradient boosting machine.

[i]CatBoost: categorical boosting.

[j]CART: classification and regression tree.

[k]VIMP: variable importance.

[l]SGD: stochastic gradient descent.

XSL•FO
RenderX

[m]KNN: *K*-nearest neighbors.

[n]QDA: quadratic discriminant analysis.

[o]LOOCV: leave-one-out cross-validation.

[p]C-index: concordance index.

This review identified a wide array of ML algorithms applied to the analysis of sociodemographic and clinical data related to cancer. Each method presents distinct advantages and limitations, influencing its suitability depending on the specific research context and analytical goals. The most relevant algorithmic approaches are summarized below.

Tree-based methods, particularly RF, were the most frequently used, appearing in 13 of the included studies. RF is widely valued for its interpretability, robustness, and ability to process both categorical and continuous variables, making it especially well-suited to heterogeneous datasets.

Boosting techniques, such as XGBoost and light gradient boosting machine (LightGBM), featured prominently in studies aiming for high predictive accuracy. XGBoost, used in 7 studies, is noted for its computational efficiency and its capacity to manage imbalanced data, while LightGBM is often selected in contexts where large-scale data processing is prioritized.

A smaller subset of studies used Bayesian additive regression trees, which were particularly useful in modeling uncertainty and capturing complex non-linear associations. These features make Bayesian additive regression trees well-suited for analyzing disparities across ethnic and clinical subgroups.

Support vector machines (SVM) appeared in 5 studies and are recognized for their ability to handle high-dimensional data and to separate complex classes using nonlinear decision boundaries [27]. However, their performance is highly dependent on careful hyperparameter tuning, which can be challenging in the presence of large or noisy datasets [27]. Overall, SVM models remain a valuable choice for complex biomedical data when appropriately optimized and validated within diverse clinical contexts.

Artificial neural networks (ANNs) were applied in select studies and demonstrated strong performance in modeling nonlinear relationships and uncovering hidden patterns in complex datasets [28]. Despite their flexibility, the limited interpretability of ANNs often restricts their use in clinical contexts where transparency and explainability are required [28]. Their use, therefore, should be accompanied by complementary interpretability frameworks to ensure clinical reliability and trustworthiness.

Regression-based models, including the least absolute shrinkage and selection operator and ridge regression, were commonly used as baseline models or for feature selection. These methods are appreciated for their simplicity and interpretability, although they may underperform in settings involving nonlinear relationships or intricate interactions between variables [29]. Nevertheless, their transparency and ease of implementation make them a critical reference point for benchmarking more advanced ML models in oncology research.

Some studies also implemented bagged classification and regression tree models and ensemble methods such as stacking, reflecting a methodological interest in combining simplicity with predictive robustness. These strategies reduce model variance and enhance accuracy by integrating multiple base learners.

Overall, the analysis reveals a strong preference for tree-based algorithms, which offer an optimal balance between accuracy, interpretability, and adaptability to real-world clinical data. However, the choice of algorithm varied according to the nature of the dataset and the specific research objectives. More recent studies have increasingly adopted advanced methods such as boosting and neural networks, which provide enhanced predictive power but require greater expertise for interpretation and implementation.

Common Validation Methods

The reviewed studies showed a strong preference for cross-validation (CV) as the primary strategy to evaluate ML models applied to the identification of sociodemographic factors related to cancer. This approach is widely recognized for its ability to reduce overfitting and enhance the robustness of predictive performance. Several configurations of CV were used across studies, with 10-fold CV being the most commonly used. This method appeared in studies such as Dianati-Nasab et al [24], Stabellini et al [20], and Afrash et al [22], where it facilitated efficient partitioning of data into training and testing subsets, maximizing the use of available datasets.

In some cases, CV was complemented with repeated sampling to mitigate random variation and reinforce consistency. For instance, Wang et al [30] implemented repetitions alongside 10-fold CV to strengthen model reliability. A less frequently used configuration, 5-fold CV, was applied in studies like Kaushik et al [11], offering a computationally efficient alternative without substantially compromising model evaluation.

Several studies further enhanced reliability by incorporating multiple repetitions. A notable example is the work of He et al [9], who used 200 repetitions and evaluated model performance using metrics such as the concordance index and variable importance measures to ensure consistency and interpretability.

The choice of evaluation metrics reflected a balanced interest in both model discrimination and interpretability. The area under the receiver operating characteristic curve was one of the most frequently reported metrics, particularly valued for its ability to quantify discrimination capacity. It was prominently featured in studies such as Dehdar et al [19] and Niell et al [12]. Additionally, accuracy, sensitivity, and specificity were widely reported, especially in studies such as Galadima et al [25] and Lilhore et al [14], as they provided a detailed picture of false positive and false negative rates.

Some researchers adopted tailored interpretability metrics to better understand model behavior. For example, Niu et al [15] used variable inclusion proportions and partial dependence plots to explore the relative importance and marginal effect of

XSL•FO
RenderX

predictors, offering deeper insights into model mechanisms. Model optimization also played a critical role in the validation process. Techniques such as grid search were frequently used to fine-tune hyperparameters, as observed in the work of Dehdar et al [19]. In more specialized contexts, such as radiomics applications, validation using pretrained models was implemented, for example, in Dercle et al [21], focusing on metastatic colorectal cancer and highlighting the relevance of domain-specific strategies.

While most studies ensured strong internal validity, a common limitation was the lack of external validation. Although a few studies used unseen datasets or pretrained models to assess generalizability, the overall scarcity of external validation in heterogeneous populations restricts the broader applicability of findings. This underscores the importance of expanding validation practices to include more diverse datasets and real-world scenarios.

Analysis of Sociodemographic Variables

The reviewed studies demonstrate considerable variability in the types of sociodemographic variables incorporated into oncology research using ML techniques. Individual-level factors, such as age and male or female or intersex, were the most frequently included, underscoring their foundational role in the development and prognosis of various cancer types. For example, in breast cancer research, variables such as age at diagnosis and hormonal status appear consistently, as noted in the studies by Dianati-Nasab et al [24] and Niell et al [12]. Similarly, race and ethnicity were widely explored in studies addressing lung and colorectal cancer [9], highlighting disparities in health outcomes associated with these variables.

In addition to individual characteristics, several studies incorporated socioeconomic and access-related factors, which reflect broader SDH. Educational attainment and household income, often used as proxies for access to health resources and health-seeking behavior, featured prominently in studies on colorectal cancer [13] and advanced-stage breast cancer [13]. Other key access variables, such as transportation availability and type of health insurance, were also frequently considered to assess barriers to diagnosis and treatment, as shown in the works of Wang et al [30] and Afrash et al [22].

Some studies expanded their scope to include community- and environment-level variables, though these remain underrepresented overall. Galadima et al [25], for instance, investigated aspects of the built environment, such as crime rates and housing values, and their association with late-stage colorectal cancer diagnoses. Similarly, Dehdar et al [19] examined the influence of residence location, urban versus rural, on access to medical services, illustrating geographic disparities in health care delivery.

Regarding cancer types, breast cancer was the most frequently studied, followed by colorectal, lung, and gastric cancer. Research on breast cancer often focuses on the impact of delayed diagnosis and racial disparities, as seen in studies by Stabellini et al [20]. In contrast, studies on colorectal cancer emphasized socioeconomic factors and health care access, particularly in relation to late-stage detection [13,25]. Lung cancer studies primarily explored racial disparities and quality-of-life indicators in survival prediction [9,10].

A few studies adopted a broader, multicancer approach, examining sociodemographic patterns across different tumor types. For example, Stabellini et al [17] analyzed unplanned hospital readmissions in patients with solid tumors, integrating sociodemographic variables that have a direct influence on health outcomes. To provide a visual synthesis of these findings, Figure 2 presents a summary linking the ML algorithms used with the most frequently analyzed sociodemographic variables.

XSL•FO
RenderX

Figure 2. Association between machine learning techniques and sociodemographic variables. ANN: artificial neural network; BT: Bayesian tree; DT: decision tree; LASSO: least absolute shrinkage and selection operator; LightGBM: light gradient boosting machine; LR: logistic regression; RF: random forest; SVM: support vector machine; XGBoost: extreme gradient boosting.

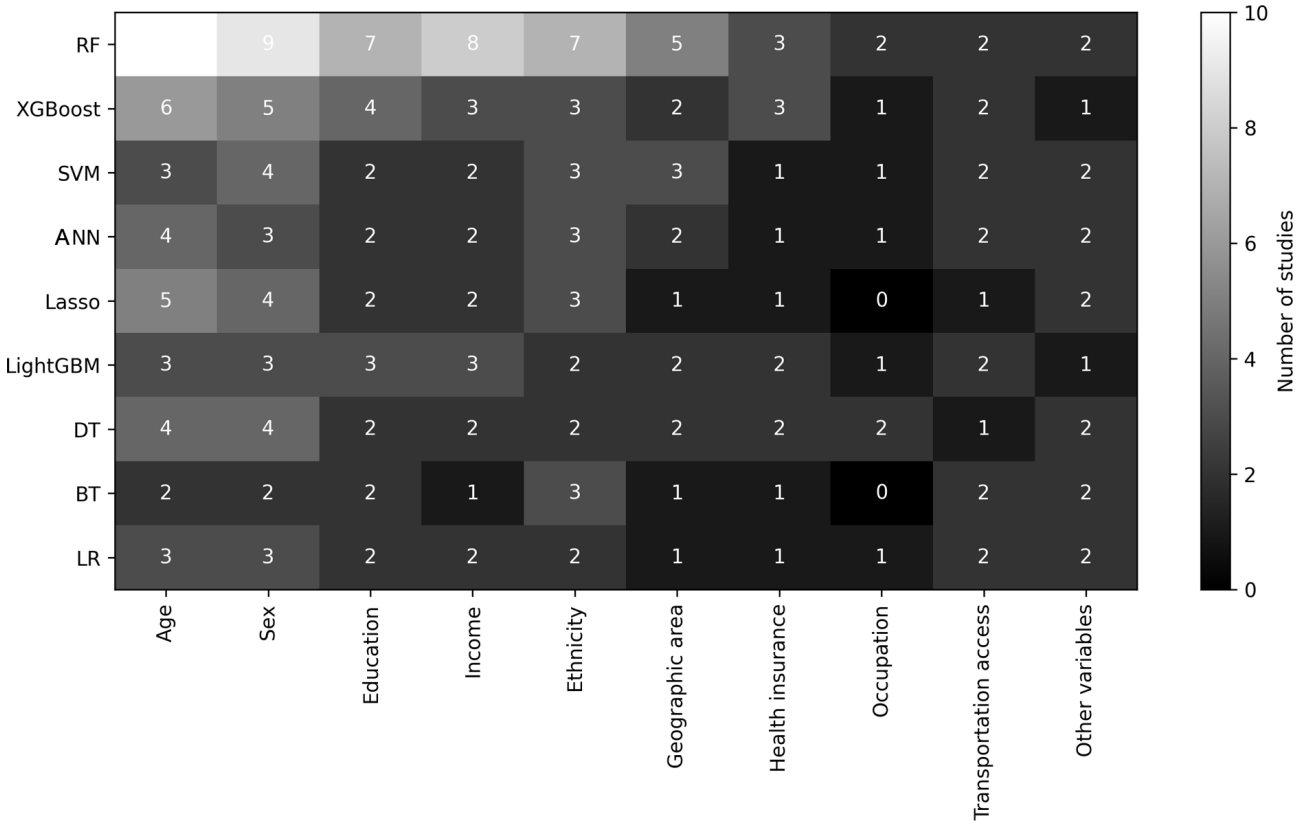

Discussion

Stratification of Findings

The reviewed studies confirm the potential of ML to identify patterns of predictive relevance of sociodemographic variables in relation to oncologic outcomes. However, the evidence remains fragmented and heterogeneous, with limited integration of contextual factors, reliance on predominantly internal validation, and little standardization in the reporting of performance and fairness. Overall, the findings suggest that ML can enhance risk stratification and the detection of disparities, but its real impact depends on methodological decisions that currently remain inconsistent.

In breast cancer, models most often prioritize age, race or ethnicity, and socioeconomic proxies to explain adverse events and late diagnosis. In colorectal cancer, income, insurance coverage, and geographic location are central for predicting advanced stage and survival. In lung cancer, studies more frequently explore ethnic disparities and quality-of-life measures associated with prognosis. This diversity suggests that the relevant set of SDoH is tumor-specific and linked to each care pathway.

Retrospective studies dominate; while they provide volume and feasibility, they limit causal inference and the ability to adapt to temporal social changes (eg, economic shocks, migration, or health system reforms). Prospective and longitudinal cohort designs would better capture the temporal variability of SDoH.

Greater interpretative weight should be placed on studies with stronger control of confounding, explicit handling of missing data, subgroup analyses, and when available, external validation. In contrast, studies with incomplete reporting of variables and opaque pipelines should be viewed as exploratory signals rather than evidence ready for implementation.

Linking Inequities and ML Limitations

When sociodemographic factors are omitted or inconsistently defined, ML models often end up reflecting pre-existing inequities in access to and quality of care instead of uncovering or addressing them. This reflection of structural disparities undermines both the external validity and the generalizability of predictive models [31,32]. Evidence from recent reviews indicates that algorithmic bias in health care typically emerges from unbalanced data representation and the absence of systematic fairness assessments, highlighting the importance of transparency and interpretability in model design [33,34]. Although variable-importance analyses can reveal which sociodemographic features most influence predictions, they fall short of explaining underlying causal mechanisms. As Prosperi et al [35] and McCradden et al [36] emphasize, achieving fairness and accountability in ML-driven health applications requires methodological and ethical frameworks that move beyond conventional supervised learning. For this reason, throughout this review, the term “associated factors” is used exclusively in a predictive, not causal, sense.

To advance the field, it is essential to standardize the reporting of sociodemographic variables including age, male or female or intersex, race or ethnicity, education, income, rurality, and

XSL•FO
RenderX

health insurance as a minimum dataset to reduce heterogeneity and enable comparability across studies. Fairness metrics, such as demographic parity, equal opportunity, and subgroup calibration, should be applied alongside conventional measures like area under the curve and accuracy to explicitly assess model performance in vulnerable populations. Routine multicenter external validation is needed, testing models across diverse geographical and socioeconomic contexts. Incorporating neighborhood-level data (eg, area-level socioeconomic indices, transportation access, and housing conditions) can provide valuable context for individual predictors. Interdisciplinary collaboration between data scientists, oncologists, public health practitioners, and experts in social science and policy should be promoted to ensure that models achieve both technical precision and equity. Finally, transparent dissemination, including open-source code and model cards documenting limitations, is crucial to strengthen reproducibility and accountability.

Principal Findings

This systematic review synthesized evidence from 19 primary studies published between 2018 and 2024 that applied ML techniques to analyze sociodemographic factors associated with cancer. The analysis revealed consistent methodological patterns, frequently used variables, and prevalent validation strategies, while also identifying key implications for both academic research and professional practice.

From a methodological perspective, there was a strong preference for tree-based algorithms, particularly RF, which was the most frequently used due to its capacity to manage heterogeneous datasets while preserving a degree of interpretability. Boosting methods, notably XGBoost and LightGBM, were also prominent, especially in studies aiming for high predictive accuracy in high-dimensional or imbalanced data contexts. Less frequently, SVMs and ANNs were used to capture complex, nonlinear relationships, typically in specialized modeling scenarios. Regression-based approaches such as the least absolute shrinkage and selection operator and Ridge regression were primarily used for feature selection or as baseline models for comparative purposes.

Across the studies, a consistent set of core sociodemographic variables was identified. The most commonly included were age, male or female or intersex, educational level, income, ethnicity, and geographic location. These factors were primarily used to predict diagnostic timelines, disparities in access to treatment, and survival outcomes. However, only a limited number of studies incorporated broader structural or contextual variables—such as neighborhood characteristics, transportation access, or housing conditions—that could enrich model performance by capturing deeper dimensions of health inequity.

In terms of validation strategies, 10-fold CV was the most frequently implemented, followed by 5-fold validation in settings with limited computational resources. Most studies relied on standard evaluation metrics such as accuracy, area under the receiver operating characteristic curve, and sensitivity or specificity, reflecting a predominant focus on internal performance. However, the use of external validation with independent datasets was rare, limiting the generalizability of findings to broader, more diverse populations and real-world clinical environments.

From an applied perspective, the findings suggest that ML holds significant promise for identifying and quantifying structural health disparities in oncology. For the academic research community, this review highlights the importance of developing models that explicitly integrate SDoH, moving beyond individual-level data to encompass contextual and systemic influences. For clinicians and policymakers, predictive models incorporating sociodemographic factors offer a valuable complement to traditional clinical assessments, enabling the early identification of at-risk populations who might otherwise be overlooked.

Taken together, these findings underscore the transformative potential of ML when applied with methodological rigor, interpretability, and an explicit commitment to equity. Advancing this field will require not only continued technical innovation, but also interdisciplinary collaboration and a deliberate focus on addressing the social and structural dimensions of cancer prevention, diagnosis, and care.

Limitations

We critically assessed potential threats to the validity of our SLR based on the Wohlin classification, which provides clear guidelines for identifying and mitigating such threats [37].

Internal validity threats involve factors that could influence the reliability and accuracy of our study outcomes. A primary concern is selection bias, potentially stemming from limitations inherent in our search strategy and inclusion criteria. To minimize this risk, we carefully defined explicit and rigorous inclusion and exclusion criteria, conducting systematic searches across multiple reputable academic databases. Despite these measures, the relatively small final sample size (N=19) remains a limitation. To further reinforce internal validity, we conducted independent cross-checking and reviews with three domain experts, ensuring consistency and reliability in the selection and evaluation of studies.

External validity threats refer to the generalizability of our findings beyond the specific studies reviewed. A significant concern here is the representativeness of the primary studies regarding the broader application of ML to sociodemographic determinants of cancer. To mitigate this threat, we engaged external experts in data science and public health to provide critical insights and feedback on our findings, enhancing the relevance and applicability across different contexts [7].

Finally, construct validity threats pertain to the accurate interpretation and generalization of results in alignment with the study objectives. The primary concern here is potential subjectivity or bias in interpreting the findings. To address this, external collaborators participated in the analysis and classification phases, providing independent perspectives that strengthened the robustness and objectivity of our conclusions.

Comparison With Prior Work

Several systematic reviews have examined the application of ML techniques in oncology, but their scope differs significantly from this study. Adeoye et al [38] evaluated ML models in

XSL•FO
RenderX

oncology settings with limited resources, identifying gaps in external validation and clinical adoption, but without providing a detailed analysis of sociodemographic variables. Hossain Raju et al [26] reviewed the use of deep learning for breast cancer risk prediction, focusing mainly on imaging and genomic data. Kumar et al [39] offered a broad overview of AI in oncology, emphasizing technical innovation rather than social determinants. Zeinali et al [40] analyzed the application of ML in predicting cancer-related symptoms, again with a focus on clinical variables.

In addition, recent editorials and reviews have highlighted the need to move toward more interpretable and explainable models. For example, Hrinivich et al [4] warned about the risks associated with the lack of interpretability in ML models in oncology, noting that reliance on opaque systems may amplify biases and weaken clinical trust. However, while these works underscore the importance of technical transparency, they do not systematically address the incorporation of sociodemographic factors into predictive cancer models.

Our review differs from previous contributions in three main ways. First, we provide a systematic synthesis of primary studies in which sociodemographic factors are explicitly integrated into ML models applied to oncological outcomes, thereby moving beyond an exclusively clinical or technical lens. Second, we critically assess methodological limitations—such as the lack of external validation, limited interpretability, and absence of fairness metrics—specifically in relation to the inclusion of sociodemographic data. Third, we connect these findings to broader discussions of equity and public health, emphasizing that neglecting social determinants may inadvertently reinforce inequalities in cancer care. By placing sociodemographic factors at the center rather than at the periphery, this review addresses an underexplored yet essential dimension of the field.

Ultimately, our findings contribute meaningfully to the growing body of literature by illustrating how ML can be leveraged to deepen our understanding of social inequalities in cancer outcomes. Rather than treating sociodemographic variables as peripheral, this study brings them to the forefront of analysis, offering a more nuanced view of how structural and contextual factors shape cancer risk, access to care, and treatment outcomes. These insights can help guide the development of more inclusive health policies and inform interventions that are responsive to the realities of diverse and historically underserved populations.

Conclusions

This review indicates that the integration of sociodemographic factors into ML models for oncology is still an emerging field, with a modest evidence base that appears to be steadily growing. Only 19 primary studies met our inclusion criteria, yet their collective findings point to the potential benefits of embedding these variables within predictive frameworks. There is some evidence to suggest that explicitly accounting for sociodemographic factors could refine predictive accuracy and fairness, although these associations remain noncausal. That said, such conclusions remain tentative, as further research is needed to substantiate these observations. Looking ahead, researchers might prioritize enhancing the transparency of these models, exploring fairness metrics, and considering how such tools align with the broader goals of health policy. Advancing these aspects could prove vital in ensuring that ML supports both precision oncology and equitable public health outcomes. It is worth noting that, although the variables examined in this review are those most frequently reported in existing datasets, future research could benefit from incorporating contextual and structural determinants to strengthen both fairness and interpretability in ML-based cancer studies (Multimedia Appendices 2 and 3).

Acknowledgments

This study was conducted as part of the Doctoral Program in Economics and Information Management at the Universidad del Bío-Bío. The authors also acknowledge the support of the Center for Cancer Control and Prevention. Additionally, all the authors thank the Ñuble Health Hub of the Universidad del Bío-Bío for their valuable support. The authors did not use generative artificial intelligence technologies (such as ChatGPT or similar tools) in the generation of this manuscript.

Funding

This research was supported by the Agencia Nacional de Investigación y Desarrollo through the Fondo de Financiamiento de Centros de Investigación en Áreas Prioritarias (grant 152220002).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Quality assessment criteria and the assignment of scores.
[DOCX File, 16 KB - jmir_v28i1e79187_app1.docx]

Multimedia Appendix 2
Primary studies by year and publication type.
[PNG File, 323 KB - jmir_v28i1e79187_app2.png]

XSL•FO
RenderX

Multimedia Appendix 3
Primary studies description.
[DOCX File, 20 KB - jmir_v28i1e79187_app3.docx]

Checklist 1
PRISMA 2020 checklist.
[PDF File, 133 KB - jmir_v28i1e79187_app4.pdf]

References

1. Zhang J, Lu Y, Zhang N, et al. Global burden of female breast cancer and its association with socioeconomic development status, 1990-2044. Cancer Rep (Hoboken) 2023 Sep;6(Suppl 1):e1827. [doi: 10.1002/cnr2.1827] [Medline: 37095062]
2. Fountzilas E, Pearce T, Baysal MA, Chakraborty A, Tsimberidou AM. Convergence of evolving artificial intelligence and machine learning techniques in precision oncology. NPJ Digit Med 2025 Jan 31;8(1):75. [doi: 10.1038/s41746-025-01471-y] [Medline: 39890986]
3. Alelyani T, Alshammari MM, Almuhanna A, Asan O. Explainable artificial intelligence in quantifying breast cancer factors: Saudi Arabia context. Healthcare (Basel) 2024 May 15;12(10):1025. [doi: 10.3390/healthcare12101025] [Medline: 38786433]
4. Hrinivich WT, Wang T, Wang C. Editorial: Interpretable and explainable machine learning models in oncology. Front Oncol 2023;13:1184428. [doi: 10.3389/fonc.2023.1184428] [Medline: 37035194]
5. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097. [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
6. Petersen K, Vakkalanka S, Kuzniarz L. Guidelines for conducting systematic mapping studies in software engineering: an update. Inf Softw Technol 2015 Aug;64:1-18. [doi: 10.1016/j.infsof.2015.03.007]
7. Kitchenham B. Procedures for performing systematic reviews (technical report TR/SE-0401). : Keele University; 2004 URL: https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf [accessed 2026-01-20]
8. Mortezagholi A, Khosravizadeh O, Menhaj MB, Shafigh Y, Kalhor R. Make intelligent of gastric cancer diagnosis error in Qazvin's medical centers: using data mining method. Asian Pac J Cancer Prev 2019 Sep 1;20(9):2607-2610. [doi: 10.31557/APJCP.2019.20.9.2607] [Medline: 31554353]
9. He J, Zhang JX, Chen CT, et al. The relative importance of clinical and socio-demographic variables in prognostic prediction in non-small cell lung cancer: a variable importance approach. Med Care 2020 May;58(5):461-467. [doi: 10.1097/MLR.0000000000001288] [Medline: 31985586]
10. Sim JA, Kim YA, Kim JH, et al. The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: applications of machine learning. Sci Rep 2020 Jul 1;10(1):10693. [doi: 10.1038/s41598-020-67604-3] [Medline: 32612283]
11. Kaushik M, Joshi RC, Kushwah AS, et al. Cytokine gene variants and socio-demographic characteristics as predictors of cervical cancer: a machine learning approach. Comput Biol Med 2021 Jul;134:104559. [doi: 10.1016/j.compbiomed.2021.104559] [Medline: 34147008]
12. Niell BL, Abdalah M, Stringfield O, et al. Quantitative measures of background parenchymal enhancement predict breast cancer risk. AJR Am J Roentgenol 2021 Jul;217(1):64-75. [doi: 10.2214/AJR.20.23804] [Medline: 32876474]
13. Dong W, Bensken WP, Kim U, Rose J, Berger NA, Koroukian SM. Phenotype discovery and geographic disparities of late-stage breast cancer diagnosis across U.S. counties: a machine learning approach. Cancer Epidemiol Biomarkers Prev 2022 Jan;31(1):66-76. [doi: 10.1158/1055-9965.EPI-21-0838] [Medline: 34697059]
14. Lilhore UK, Poongodi M, Kaur A, et al. Hybrid model for detection of cervical cancer using causal analysis and machine learning techniques. Comput Math Methods Med 2022;2022:4688327. [doi: 10.1155/2022/4688327] [Medline: 35572826]
15. Niu L, Hu L, Li Y, Liu B. Correlates of cancer prevalence across census tracts in the United States: a Bayesian machine learning approach. Spat Spatiotemporal Epidemiol 2022 Aug;42:100522. [doi: 10.1016/j.sste.2022.100522] [Medline: 35934328]
16. Stabellini N, Dmukauskas M, Bittencourt MS, et al. Social determinants of health and racial disparities in cardiac events in breast cancer. J Natl Compr Canc Netw 2023 Jul;21(7):705-714. [doi: 10.6004/jnccn.2023.7023] [Medline: 37433439]
17. Stabellini N, Nazha A, Agrawal N, et al. Thirty-day unplanned hospital readmissions in patients with cancer and the impact of social determinants of health: a machine learning approach. JCO Clin Cancer Inform 2023 Jul;7:e2200143. [doi: 10.1200/CCI.22.00143] [Medline: 37463363]
18. Stone A, Kalahiki C, Li L, Hubig N, Iuricich F, Dunn H. Evaluation of breast tumor morphologies from African American and Caucasian patients. Comput Struct Biotechnol J 2023;21:3459-3465. [doi: 10.1016/j.csbj.2023.06.019] [Medline: 38213888]
19. Dehdar S, Salimifard K, Mohammadi R, et al. Applications of different machine learning approaches in prediction of breast cancer diagnosis delay. Front Oncol 2023;13:1103369. [doi: 10.3389/fonc.2023.1103369] [Medline: 36874113]

XSL•FO
RenderX

20. Stabellini N, Cullen J, Moore JX, et al. Social determinants of health data improve the prediction of cardiac outcomes in females with breast cancer. Cancers (Basel) 2023 Sep 19;15(18):4630. [doi: 10.3390/cancers15184630] [Medline: 37760599]
21. Dercle L, Yang M, Gönen M, et al. Ethnic diversity in treatment response for colorectal cancer: proof of concept for radiomics-driven enrichment trials. Eur Radiol 2023 Dec;33(12):9254-9261. [doi: 10.1007/s00330-023-09862-z] [Medline: 37368111]
22. Afrash MR, Shafiee M, Kazemi-Arpanahi H. Establishing machine learning models to predict the early risk of gastric cancer based on lifestyle factors. BMC Gastroenterol 2023 Jan 10;23(1):6. [doi: 10.1186/s12876-022-02626-x] [Medline: 36627564]
23. Dong W, Kim U, Rose J, et al. Geographic variation and risk factor association of early versus late onset colorectal cancer. Cancers (Basel) 2023 Feb 4;15(4):1006. [doi: 10.3390/cancers15041006] [Medline: 36831350]
24. Dianati-Nasab M, Salimifard K, Mohammadi R, et al. Machine learning algorithms to uncover risk factors of breast cancer: insights from a large case-control study. Front Oncol 2024;13:1276232. [doi: 10.3389/fonc.2023.1276232] [Medline: 38425674]
25. Galadima H, Anson-Dwamena R, Johnson A, Bello G, Adunlin G, Blando J. Machine learning as a tool for early detection: a focus on late-stage colorectal cancer across socioeconomic spectrums. Cancers (Basel) 2024 Jan 26;16(3):540. [doi: 10.3390/cancers16030540] [Medline: 38339293]
26. Raju MAH, Imam T, Islam J, Al Rakin A, Nayyem MN, Uddin MS. An ontological framework for lung carcinoma prognostication via sophisticated stacking and synthetic minority oversampling techniques. Presented at: 2024 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob); Nov 28-30, 2024; Bali, Indonesia p. 125-130. [doi: 10.1109/APWiMob64015.2024.10792946]
27. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition: Springer; 2009. URL: https://hastie.su.domains/ElemStatLearn/ [accessed 2025-01-20]
28. Goodfellow I, Bengio Y, Courville A. Deep Learning: MIT Press; 2016. URL: https://www.deeplearningbook.org/ [accessed 2026-01-20]
29. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol 2005 Apr 1;67(2):301-320. [doi: 10.1111/j.1467-9868.2005.00503.x]
30. Wang Z, Kim Y, Barbosa EJM. Demographics and socioeconomic determinants of health predict continued participation in a CT lung cancer screening program. Curr Probl Diagn Radiol 2024;53(5):552-559. [doi: 10.1067/j.cpradiol.2024.04.004] [Medline: 38658287]
31. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018 Dec 18;169(12):866-872. [doi: 10.7326/M18-1990] [Medline: 30508424]
32. Colacci M, Huang YQ, Postill G, et al. Sociodemographic bias in clinical machine learning models: a scoping review of algorithmic bias instances and mechanisms. J Clin Epidemiol 2025 Feb;178:111606. [doi: 10.1016/j.jclinepi.2024.111606] [Medline: 39532254]
33. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. JAMA 2019 Dec 24;322(24):2377-2378. [doi: 10.1001/jama.2019.18058] [Medline: 31755905]
34. Ning Y, Li S, Ng YY, et al. Variable importance analysis with interpretable machine learning for fair risk prediction. PLOS Digit Health 2024 Jul;3(7):e0000542. [doi: 10.1371/journal.pdig.0000542] [Medline: 38995879]
35. Prosperi M, Guo Y, Sperrin M, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nat Mach Intell 2020;2(7):369-375. [doi: 10.1038/s42256-020-0197-y]
36. McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. Lancet Digit Health 2020 May;2(5):e221-e223. [doi: 10.1016/S2589-7500(20)30065-0] [Medline: 33328054]
37. Kitchenham B, Charters S. Guidelines for performing systematic literature reviews in software engineering. : Keele University and University of Durham; 2007 URL: https://legacyfileshare.elsevier.com/promis_misc/525444systematicreviewsguide.pdf [accessed 2025-01-20]
38. Adeoye J, Akinshipo A, Koohi-Moghadam M, Thomson P, Su YX. Construction of machine learning-based models for cancer outcomes in low and lower-middle income countries: a scoping review. Front Oncol 2022;12:976168. [doi: 10.3389/fonc.2022.976168] [Medline: 36531037]
39. Kumar Y, Gupta S, Singla R, Hu YC. A systematic review of artificial intelligence techniques in cancer prediction and diagnosis. Arch Comput Methods Eng 2022;29(4):2043-2070. [doi: 10.1007/s11831-021-09648-w] [Medline: 34602811]
40. Zeinali N, Youn N, Albashayreh A, Fan W, Gilbertson White S. Machine learning approaches to predict symptoms in people with cancer: systematic review. JMIR Cancer 2024 Mar 19;10:e52322. [doi: 10.2196/52322] [Medline: 38502171]

Abbreviations

AI: artificial intelligence
ANN: artificial neural network
LightGBM: light gradient boosting machine

XSL•FO
RenderX

ML: machine learning
PICO: population, intervention, comparison, and outcome
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RF: random forest
SDoH: social determinants of health
SLR: systematic literature review
SVM: support vector machine
XGBoost: extreme gradient boosting

Edited by J Sarvestan; submitted 17.Jun.2025; peer-reviewed by JCL Chow, SM Shaffi; revised version received 18.Oct.2025; accepted 30.Oct.2025; published 28.Jan.2026.

Please cite as:
González-Infante L, Marquez G, Parra-Soto S, Cardona-Valencia M, Taramasco C
Machine Learning Techniques Used for the Identification of Sociodemographic Factors Associated With Cancer: Systematic Literature Review
J Med Internet Res 2026;28:e79187
URL: https://www.jmir.org/2026/1/e79187
doi:10.2196/79187

© Liz González-Infante, Gaston Marquez, Solange Parra, Mónica Cardona, Carla Taramasco. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Phases of Living Evidence Synthesis Using AI: Living Evidence Synthesis (Version 1)

Xuping Song[1,2,3,4], PhD; Zhenjie Lian[1,2], MSc; Rui Wang[2,5], MSc; Ruixin Li[1,2], MSc; Zhenzhen Yang[1,2], MSc; Xufei Luo[4,6], PhD; Lei Feng[1], BSc; Zhiming Ma[1], BSc; Zhen Pu[1], BSc; Qi Wang[7], PhD; Long Ge[1,2,3,4], PhD; Caihong Li[8], PhD; Yaolong Chen[1,3,4,6], PhD; Kehu Yang[1,2,3,4], MSc; John Lavis[9], PhD

[1]School of Public Health, Lanzhou University, No. 222 South Tianshui Road, Lanzhou, Lanzhou, Gansu, China

[2]The Centre of Evidence-based Social Science, Lanzhou University, Lanzhou, Gansu, China

[3]Key Laboratory of Evidence Based Medicine & Knowledge Translation of Gansu Province, Lanzhou, Gansu, China

[4]WHO Collaborating Centre for Guideline Implementation and Knowledge Translation, Lanzhou, Gansu, China

[5]Dingxi Center for Disease Control and Prevention, Dingxi, Gansu, China

[6]Evidence-Based Medicine Center, School of Basic Medicine, Lanzhou University, Lanzhou, Gansu, China

[7]School of Public Health, University of Hong Kong, Hong Kong, China

[8]School of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu, China

[9]Department of Health Research Methods, Evidence, and Impact, McMaster Health Forum, McMaster University, Hamilton, ON, Canada

Corresponding Author:
Kehu Yang, MSc
School of Public Health, Lanzhou University, No. 222 South Tianshui Road, Lanzhou, Lanzhou, Gansu, China

Abstract

Background: Living evidence (LE) synthesis refers to the method of continuously updating systematic evidence reviews to incorporate new evidence. It has emerged to address the limitations of the traditional systematic review process, particularly the absence of or delays in publication updates. The emergence of COVID-19 accelerated the progress in the field of LE synthesis, and currently, the applications of artificial intelligence (AI) in LE synthesis are expanding rapidly. However, in which phases of LE synthesis should AI be used remains an unanswered question.

Objective: This study aims to (1) document the phases of LE synthesis where AI is used and (2) investigate whether AI improves the efficiency, accuracy, or utility of LE synthesis.

Methods: We searched Web of Science, PubMed, the Cochrane Library, Epistemonikos, the Campbell Library, IEEE Xplore, medRxiv, COVID-19 Evidence Network to support Decision-making, and McMaster Health Forum. We used Covidence to facilitate the monthly screening and extraction processes to maintain the LE synthesis process. Studies that used or developed AI or semiautomated tools in the phases of LE synthesis were included.

Results: A total of 24 studies were included, including 17 on LE syntheses, with 4 involving tool development, and 7 on living meta-analyses, with 3 involving tool development. First, a total of 34 AI or semiautomated tools were involved, comprising 12 AI tools and 22 semiautomated tools. The most frequently used AI or semiautomated tools were machine learning classifiers (n=5) and the Living Interactive Evidence synthesis platform (n=3). Second, 20 AI or semiautomated tools were used for the data extraction or collection and risk of bias assessment phase, and only 1 AI tool was used for the publication update phase. Third, 3 studies demonstrated the improvement in efficiency achieved based on time, workload, and conflict rate metrics. Nine studies applied AI or semiautomated tools in LE synthesis, obtaining a mean recall rate of 96.24%, and 6 studies achieved a mean F_1-score of 92.17%. Additionally, 8 studies reported precision values ranging from 0.2% to 100%.

Conclusions: AI and semiautomated tools primarily facilitate data extraction or collection and risk of bias assessment. The use of AI or semiautomated tools in LE synthesis improves efficiency, leading to high accuracy, recall, and F_1-scores, while precision varies across tools.

Trial Registration: OSF Registries 87tp4; https://osf.io/4fvdq/overview

(J Med Internet Res 2026;28:e76130) doi:10.2196/76130

KEYWORDS

accuracy; artificial intelligence; efficiency; living evidence synthesis; phases; semiautomated tools; utility

XSL•FO
RenderX

Introduction

Evidence synthesis refers to an approach where data across studies are identified and combined to gain a clearer understanding of a body of research [1]. There is typically a significant gap between the time when a search is performed and the time when the results are published, often exceeding a year [2]. Furthermore, only a limited number of reviews are updated once they have been published [3]. This process can result in missing evidence, potentially affecting the accuracy of the findings. The approach of living evidence (LE) synthesis has been developed to address this challenge.

The method of constantly updating a systematic synthesis of evidence to incorporate newly available evidence is known as LE [4]. Elliott et al [5] developed the basis of the LE model in 2014, which effectively incorporates and summarizes new evidence. The LE synthesis process includes 4 phases: database searching and eligibility assessment, data extraction or collection and risk of bias assessment, synthesis and analysis, and publication update [6]. It has also been adapted in areas such as network meta-analysis and guidelines. The onset of COVID-19 increased the incentive to use LE [7]. Unlike traditional evidence synthesis, which requires the redeployment of significant resources for updates, the maintenance of an LE synthesis can require more modest resources [8]. However, LE synthesis that focuses on evolving topics may have a reduced reliability compared to traditional evidence synthesis. The incorporation of artificial intelligence (AI) techniques has the potential to enhance the reliability of LE synthesis by, for example, leveraging advanced algorithms to continuously assess and filter the most relevant and high-quality evidence [9].

The field of AI, which encompasses machine learning, deep learning, natural language processing, data mining, image recognition, and computer vision, to name a few, has the potential to enhance the efficiency of LE synthesis [10,11]. In 2013, Adams et al [11] indicated that leveraging AI to automate the LE synthesis procedures could simplify the regular updating and maintenance of evidence. The development of AI systems, particularly AI based on large language models (LLMs), such as the generative pretrained transformer, has significantly advanced natural generative language systems [12]. Various AI-driven tools have been developed for different phases of LE synthesis, such as crowdsourcing and task-sharing platforms like HDAS [13]. However, the performance of the AI techniques and the phases of LE synthesis where AI is used remain unclear.

Overall, the objectives of this review are (1) to conduct a review analyzing the phases of LE synthesis that use AI and (2) to explore whether AI can improve the efficiency, accuracy, or utility of LE synthesis.

Methods

This is the first version of an LE synthesis. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 statement for living systematic reviews (PRISMA-LSR; Checklist 1) was used as a guide for reporting this LE synthesis [14]. The review has been registered in the Open Science Forum [15].

Search Strategy

We systematically searched the Web of Science, PubMed, the Cochrane Library, Epistemonikos, the Campbell Library, IEEE Xplore, medRxiv, COVID-19 Evidence Network to support Decision-making, and McMaster Health Forum for publications up to April 2, 2025. The details of the search strategy used can be found in Table S1 in Multimedia Appendix 1. We subscribed to the Web of Science, PubMed, the Cochrane Library, the Campbell Library, and IEEE Xplore for monthly dynamic updates and used Covidence to facilitate the screening and extraction processes for maintaining an LE synthesis. We plan to conduct living updates for a 12-month period (from April 2025 to April 2026). The final update is scheduled for April 2, 2026, after which we will assess whether to retire the living mode based on the following established triggers: (1) evidence on “the AI application in LE synthesis” has reached conclusiveness, (2) the topic no longer holds decision-making value for the field, (3) no new eligible studies emerge during the 12-month update period, or (4) subsequent resource or funding support is unavailable [16,17].

Inclusion and Exclusion Criteria

First, the LE synthesis includes living systematic review, living meta-analysis, living network meta-analysis, living guideline, living scoping review, living overview, living umbrella review, and living mapping. In this review, the types of included studies were classified into 2 categories based primarily on whether a meta-analysis had been performed. These categories include the LE synthesis (without a meta-analysis) and living meta-analysis (with a meta-analysis conducted).

Second, the criteria for inclusion in this review are studies that use AI or semiautomated tools in the following phases of LE synthesis: (1) database searching and eligibility assessment, (2) data extraction or collection and risk of bias assessment, (3) synthesis and analysis, or (4) publication update [6]. The LE syntheses from any field were included. In addition, studies that developed AI or semiautomated tools for LE synthesis were also included. Textbox 1 provides further details.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria for the study.

Inclusion criteria

- The studies using artificial intelligence (AI) or semiautomated tools in the following phases of living evidence (LE) synthesis: (1) database searching and eligibility assessment, (2) data extraction or collection and risk of bias assessment, (3) synthesis and analysis, or (4) publication update. A study can be any type of LE synthesis in any field, including but not limited to all scientific journals in the social sciences.
- Studies that developed AI or semiautomated tools for LE synthesis.

Exclusion criteria

- Studies that did not document the use of AI or semiautomated tools in LE synthesis.
- Protocol, commentaries, editorials, letters to the editor, and updating studies.

We excluded studies that did not document the use of AI or semiautomated tools in LE synthesis. In addition, protocols, commentaries, editorials, letters to the editor, and updating studies were also excluded, as shown in Textbox 1.

Third, AI tools are characterized by autonomous learning and end-to-end decision-making. They enable the independent execution of data collection, feature extraction, model training, and inference and generate output results without any human intervention. However, semiautomated tools incorporate human review or decision support at critical stages, using a "machine assistance and human oversight" collaborative paradigm [18,19]. Textbox 2 shows the types of AI or semiautomated tools, where AI or semiautomated tools were categorized by the application phases. First, the first segment of the AI or semiautomated tools for each phase is sourced from Bendersky et al [13]. Second, the subsequent segment is derived from the work of Khalil et al [20]. Third, for the final segment, AI or semiautomated tools were identified and summarized from relevant studies using a manual search. The AI techniques based on LLMs, such as the generative pretrained transformer, were also included.

Textbox 2. Artificial intelligence (AI) or semiautomated tools used in the 4 phases of living evidence (LE) synthesis.

Phase 1. Database searching and eligibility assessment

- Segment 1.1: Automatic, continuous database search with push notification, database aggregators (such as HDAS, Epistemonikos), notification from clinical trial registries, randomized clinical trial classifier, text mining technologies, and automatic retrieval of full-text papers
- Segment 1.2: RCT tagger, LitSuggest, Evidence mapping tool, SRA-Polyglot Search Translator, QuickClinical, HDAS, ROBOTsearch, SRA-word, frequency analyzer, The Search Refiner, Sherlock, SRA De-duplicate, Distiller, R package-rev tools, Rayyan, EPPI-reviewer, Abstrackr, SRA helper, LibSVM classifier, Bibot, Active Screener, RobotAnalyst, Swift-Review, Evidence Pipeline, JBI Sumari, EndNote, SARA, eSuRFr, ParsCit, and Citation searcher
- Segment 1.3: Natural language processing–assisted abstract screening tool, automatic text classifiers supported by deep learning–based language models, machine learning classifiers, Cochrane Crowd, Living Interactive Evidence (LIvE) synthesis platform, Cochrane RCT classifier, OpenAlex, Risklick AI, Bayesian classifier, Generative Pretrained Transformer models, and RobotReviewer LIVE

Phase 2. Data extraction or collection and risk of bias assessment

- Segment 2.1: Machine learning information-extraction systems, automated structured data extraction tools for PDFs, machine learning–assisted RoB tool, data repositories, and linked data
- Segment 2.2: RobotReviewer, DistelleR, JBI Sumari, in-house data extraction tool written in R, statistical package R, ExaCT, Revman, Raptor, ContentMine, Graph2Data, and Evidence mapping tool
- Segment 2.3: BioMart, MetaInsight COVID-19, LIvE synthesis platform, Open Science Framework (OSF), PsychOpen CAMA, and Generative Pretrained Transformer models

Phase 3. Synthesis and analysis

- Segment 3.1: Structured data extraction tools, which automatically provide data in a suitable format for statistical analysis; continuous analysis updating based on availability of structured extracted data; and statistical surveillance of key analysis results, with threshold set for potential conclusion change
- Segment 3.2: MetaPreg, MetaXL, NetMetaXL, Meta-analyst, Webplotdigitizer, Evidence mapping tool, PRISMA flow diagram generator, Evidence mapping tool, R package-rev tools
- Segment 3.3: Risklick AI, Web Source Processing Pipeline, LIvE synthesis platform, and generative pretrained transformer models

Phase 4. Publication update

- Segment 4.1: Templated reporting of some report items, automatic text generation tools for synthesis and writing, automatization in the identification of changes between LSR versions for peer review, and editorial process (such as Archie)
- Segment 4.2: Trial2rev, RevManHAL, DistelleR, SRA replicant writer, SRA-RevMan Replicant, and JBI Sumari
- Segment 4.3: Generative pretrained transformer models

XSL•FO
RenderX

Study Screening and Data Collection

Two reviewers independently screened the titles and abstracts of all selected studies, followed by a full-text review. Any disagreements regarding selection were resolved by a third researcher. Data were extracted using a predesigned Microsoft Excel sheet. Two reviewers independently extracted data from all included studies, including information such as title, first author, journal, year of publication, LE synthesis type, types of tool or technology, types of AI or semiautomated tools, phases of LE synthesis, outcomes, and so forth. Any disagreements were resolved by a third researcher. During data extraction, representative outcomes (such as means or ranges) were prioritized for synthesis, with the range of values considered subsequently when outcomes were similarly representative.

Methodological Quality Assessment

Given the lack of a standardized tool for assessing the methodological quality of AI-related studies, the 24 studies were categorized into 3 types by methodological characteristics and primary objective (diagnostic test, tool development, or—when neither applied—a general synthesis) and assessed for methodological quality using the modified version of the Quality Assessment of Diagnostic Accuracy Studies version 2 (QUADAS-2) tool, Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Textual Evidence: Narrative, and AMSTAR 2 tool. First, 10 studies were assessed with the modified version of the QUADAS-2 tool: these studies specifically assessed the application of AI in the database searching and eligibility assessment phase, which aligns with a diagnostic test accuracy (DTA) framework. We adopted the modified version of the QUADAS-2 proposed by Rashid et al [21-23]. As QUADAS-2 is designed for DTA research contexts, this framework was only applicable to those studies where one of the objectives included the application of AI in the database searching and eligibility assessment phase [21,24,25]. The core elements of QUADAS-2 were revised to adapt it to AI-related research scenarios, as follows: “patient” was replaced with “study,” “index test” with “AI,” “reference standard” with “comparator,” and “case-control design” with “DTA framework.” We also constructed a 2×2 table, categorizing studies into “included” or “excluded” based on both “AI screening results” and “reference/original systematic review (SR) screening results,” with counts denoted as a, b, c, and d, respectively. The details of the modified QUADAS-2 are provided in Table S2 in Multimedia Appendix 1. Second, 5 studies, which specifically developed AI or semiautomated tools for LE synthesis without DTA-related accuracy evaluation and were not designed as LE synthesis themselves, were assessed using the JBI Critical Appraisal Checklist for Textual Evidence: Narrative [26]. Third, 9 studies, which were designed as LE syntheses without DTA-related accuracy evaluation and not primarily focused on AI or semiautomated tool development (or tool development was only an auxiliary means), were assessed using the AMSTAR 2 tool [27,28]. The details are shown in Tables S3 and S4 in Multimedia Appendix 1. All of the included studies were evaluated independently by 2 reviewers (RL and ZY), and disagreement was resolved by a third reviewer (ZL). The LE synthesis did not involve a statistical combination of results (meta-analysis), as its aims were to document the phases of LE synthesis where AI is used and to investigate whether AI improves the efficiency, accuracy, or utility of LE synthesis. Therefore, several systematic review procedures—including sensitivity analyses, reporting bias assessment, certainty assessment, and investigations of heterogeneity—were not used.

Data Analysis

This review conducted 3 complementary analyses, as shown in Figure 1.

Figure 1. Road map for the use of artificial intelligence (AI): applications and extractable clinical outcomes across 4 phases of living evidence synthesis. LE: living evidence.

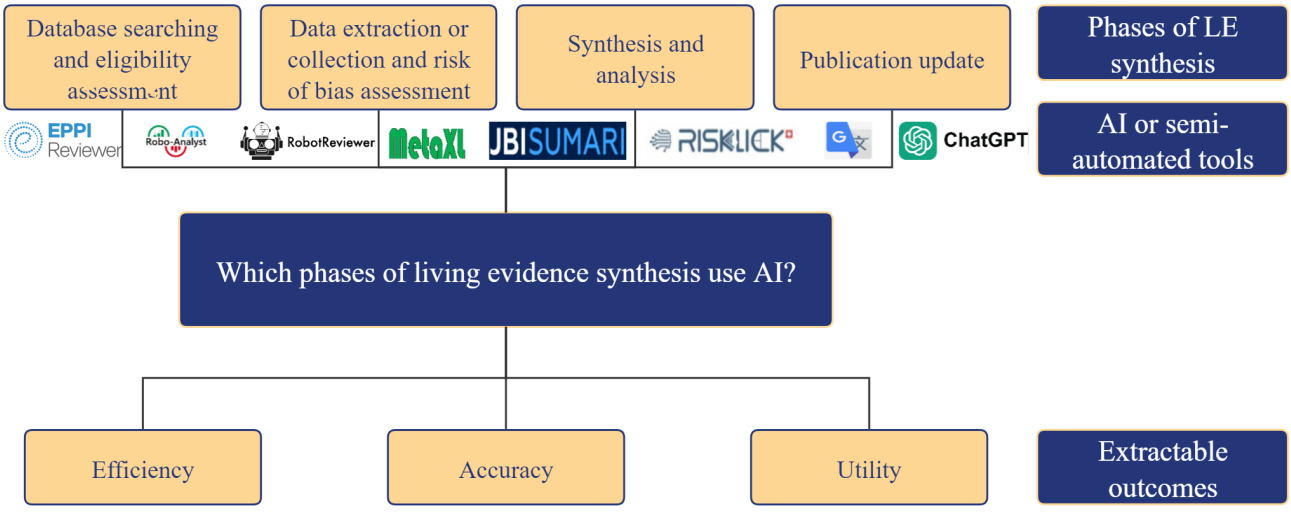

Analysis 1: Phases of LE Synthesis Utilizing AI or Semiautomated Tools

We analyzed the prevalence and distribution of AI or semiautomated tools across 4 phases of LE synthesis. Phase 1 is database searching and eligibility assessment. This process includes going through the databases, retrieving the results, importing them into the citation management software, removing any duplicate results, and assessing their eligibility individually. Phase 2 is data extraction or collection and risk of bias

XSL•FO
RenderX

assessment; once the eligibility of studies has been verified and they have been included in the review process, it becomes crucial to systematically extract and collect information about their main characteristics and results. Additionally, it is very important to assess the risk of bias associated with the conduct and methodology used in the studies. In phase 3—synthesis and analysis—the data that have been assessed to conform to the criteria are integrated, and the data are analyzed. In phase 4—publication update—after going through the aforementioned phases 1-3, sections of a review are generated based on their results, and conclusions are updated.

Analysis 2: AI or Semiautomated Tools Used in LE Synthesis

First, the types of AI or semiautomated tools applied in each LE synthesis phase were investigated. Second, the frequency of AI or semiautomated tools applied in the LE synthesis was analyzed.

Analysis 3: Primary Outcomes Investigating AI or Semiautomated Tools in LE Synthesis

The impact of applied AI or semiautomated tools in LE synthesis was analyzed across 3 outcomes [29]. First, efficiency, defined as the relationship between the time required to complete a workload and the workload itself, was evaluated to determine whether either the duration or workload was reduced with the use of AI or semiautomated tools. This outcome may be described as time reduction, workload reduction, and conflict rates with and without the tool.

Second, accuracy is used to assess performance with and without AI or semiautomated tools. It may be described as accuracy, recall, precision, F_1-score, area under the receiver operating characteristic curve, number needed to read, and study relevance. In addition, we calculated the overall mean recall and mean F_1-score using the following formula:

$$\bar{M}=1N \sum i=1NMi$$

where Mi is the representative value for study i, defined as the reported single value, if provided, or the midpoint of the reported range [L, U], calculated as (L+U)/2, if a range was provided. N is the number of studies reporting that metric [30,31].

Third, utility is used to assess whether user decisions align with those of AI or semiautomated tools, including user consistency, user satisfaction, perceived ease of use, and study quality.

Results

Search Results

Out of 9180 studies, 24 studies applied AI or semiautomated tools in LE synthesis, including 17 LE syntheses (4 developing tools) and 7 living meta-analyses (3 developing tools), as shown in Figure 2 [29,32-54]. In addition, 8 studies exclusively applied AI tools in LE synthesis, 11 studies exclusively applied semiautomated tools, and 5 studies utilized both AI and semiautomated tools. The basic characteristics of the included studies are shown in Table S5 in Multimedia Appendix 1. The details of the studies excluded at the full-text eligibility stage with reasons are shown in Table S6 in Multimedia Appendix 1 [5,9,55-75].

Figure 2. Database search flow diagram. LE: living evidence.

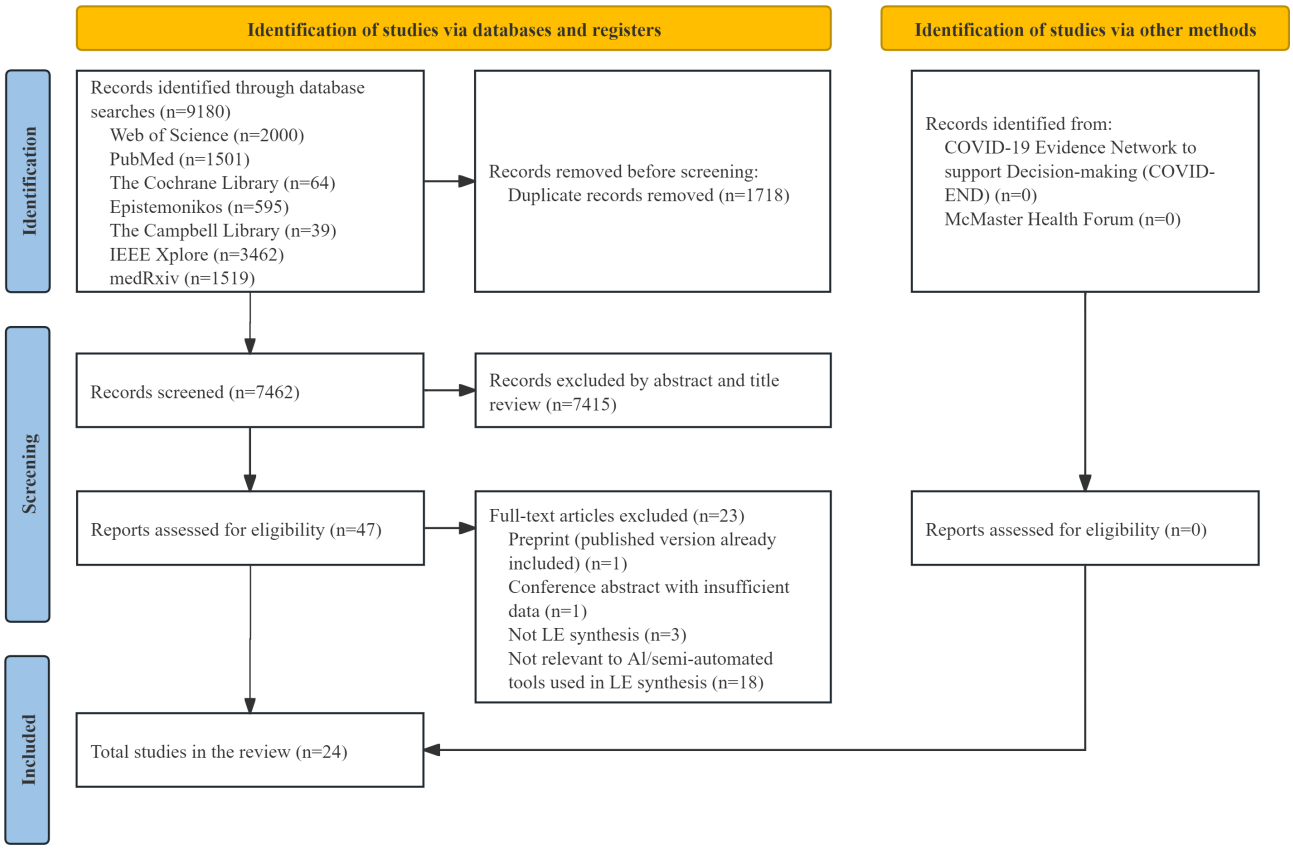

XSL•FO
RenderX

Methodological Quality of Included Studies

We conducted a methodological quality assessment of 10 studies using a revised QUADAS-2 tool within the DTA framework [29,32,35,36,42-44,51,52,54]. All studies were assessed as low-risk in the "Study selection," "Index test (AI)," and "Reference (comparator)" domains. While none of the studies specified the time interval between the task execution of AI and comparator-based analysis, all were determined as low-risk in the "Flow and timing" domain. Additionally, we did not identify any applicability concerns, as all studies were classified as low-risk in the "Applicability" domain (Table 1). Five studies were subjected to methodological quality assessment using the JBI Critical Appraisal Checklist for Textual Evidence: Narrative [41,46,48-50]. Four studies obtained a score of 5/6, with a narrative appraisal of "Exclude" owing to failure to meet the narrative classification criterion [41,46,48,49]. One study achieved a full score of 6/6 and was thus appraised as "Include" (Table S7 in Multimedia Appendix 1) [50]. In addition, we conducted a methodological quality assessment of 9 studies using AMSTAR 2 [33,34,37-40,45,47,53]. The methodological quality scores of the included studies ranged from 11 to 15. Overall, the methodological quality of eight studies [34,37-40,45,47,53] was rated as moderate, while only 1 study [33] was rated as low in methodological quality. The most common limitation was that the authors failed to provide a list of excluded studies (Table S8 in Multimedia Appendix 1).

Table . Summary of modified Quality Assessment of Diagnostic Accuracy Studies version 2 (QUADAS-2) assessments for studies using artificial intelligence (AI) or semiautomated tools in the database searching and eligibility phase of the living evidence (LE) synthesis process.

Author, year	Risk of bias				Applicability concern		
	Study selection	Index test (AI)	Reference (comparator)	Flow and timing	Study selection	Index test (AI)	Reference (comparator)
Knafou et al [32] (2023)	Low	Low	Low	Low	Low	Low	Low
Perlman-Arrow et al [29] (2023)	Low	Low	Low	Low	Low	Low	Low
Chou et al [35] (2020)	Low	Low	Low	Low	Low	Low	Low
Kamso et al [36] (2023)	Low	Low	Low	Low	Low	Low	Low
Marshall et al [42] (2023)	Low	Low	Low	Low	Low	Low	Low
Haas et al [43] (2021)	Low	Low	Low	Low	Low	Low	Low
Vaghela et al [44] (2021)	Low	Low	Low	Low	Low	Low	Low
Shemilt et al [51] (2024)	Low	Low	Low	Low	Low	Low	Low
Le-Khac et al [52] (2024)	Low	Low	Low	Low	Low	Low	Low
Hair et al [54] (2024)	Low	Low	Low	Low	Low	Low	Low

Types and Frequency of AI or Semiautomated Tools in LE Synthesis

A total of 34 AI or semiautomated tools were involved, including 12 (35.3%) AI tools and 22 (64.7%) semiautomated tools, as shown in Multimedia Appendix 2. The most frequently used AI or semiautomated tools were machine learning classifiers (n=5), followed by the Living Interactive Evidence (LIvE) synthesis platform (n=3), AD-SOLES (n=2), Covidence (n=2), and MAGICapp (n=2).

Phases of AI or Semiautomated Tools Application in LE Synthesis

There were 18 AI or semiautomated tools for database searching and eligibility assessment, 20 for data extraction or collection and risk of bias assessment, and 10 for synthesis and analysis. However, only 1 AI tool was used for publication updates. Out of all the tools, RobotReviewer LIVE can be used for all phases of LE synthesis, as shown in Textbox 3.

XSL•FO
RenderX

Textbox 3. Types of artificial intelligence (AI) or semiautomated tools applications in the 4 phases of living evidence (LE) synthesis.

Phase 1. Database searching and eligibility assessment

- LIvE platform, automatic text classifiers, machine learning ensemble classifier, Natural language processing–assisted abstract screening tool, machine learning classifiers, machine learning, PICO annotators, STAR tool, AD-SOLES, Covidence, rcrossref, openalexR, RISmed, RobotReviewer LIVE, Risklick AI, metaCOVID application, supervised text classification models, and text mining techniques

Phase 2. Data extraction or collection and risk of bias assessment

- LIvE platform, web-based interactive app, open-source living systematic review application, Covidence, AD-SOLES, Google Refine tool, script, REDASA, RobotReviewer LIVE, Risklick AI, Metainsight COVID-19, metaCOVID application, information extraction techniques, EndNote, semiautomated model, supervised text classification models, text mining techniques, GPT-4-turbo, Claude-3-Opus, and EPPI-Reviewer

Phase 3. Synthesis and analysis

- LIvE platform, MAGICapp, Trial sequential analysis (TSA) software, AD-SOLES, ODDPub, RobotReviewer LIVE, script, Metainsight COVID-19, metaCOVID application, and Dynameta

Phase 4. Publication update

- RobotReviewer LIVE

Impact of AI or Semiautomated Tools on LE Synthesis

Overview

A total of 10 (41.7%) studies reported on the impact of AI or semiautomated tools on LE synthesis in terms of efficiency, accuracy, or utility in the database searching and eligibility phase or the data extraction or collection and risk of bias assessment phase. Table 2 provides a description of the outcome metrics in the included studies.

XSL•FO
RenderX

Table . Summary of the indicator terms for outcome metrics in the included studies.

Metrics	Explanation
Efficiency	
Time	AI[a] or semiautomated tools were used to save time. Only 2 (8.3%) studies reported on time saving [29,35]. Specifically, Perlman-Arrow et al [29] reported a 45.9% reduction in screening time per abstract in the database searching and eligibility phase. Chou et al [35] estimated the time saving ranged from 2.0 to 13.2 hours in the database searching and eligibility phase.
Workload	Two (8.3%) studies reported on workload metrics related to the use of AI or semiautomated tools [29,42]. Perlman-Arrow et al [29] reported that the semiautomated tool completed 68% of the workload in the database searching and eligibility phase. Marshall et al [42] found that manual screening had an efficiency rate of 23% in obtaining 31 abstracts, whereas AI achieved a rate of 55%, demonstrating an efficiency improvement of approximately 140% in the database searching and eligibility phase.
Conflict rates with and without the tool	The efficiency of abstract screening decreases as the number of conflicting votes increases [29]. Perlman-Arrow et al [29] reported a reduction in conflict rates from 8.32% to 3.64% with the use of semiautomated tool in the database searching and eligibility phase.
Accuracy[b]	
Precision	Precision refers to the ratio of accurately categorized documents among all the documents that the model assigns to a particular class [32]. Eight (33.3%) studies reported on precision [29,32,35,42,43,51,53,54]. 1. Khan et al [53] reported a precision rate of even 100% using AI in the data extraction or collection and risk of bias assessment phase. 2. Perlman-Arrow et al [29] and Haas et al [43] reported precision rates of 92.10% and 96.07%, respectively, using AI or semiautomated tools in the database searching and eligibility phase. 3. Hair et al [54] reported that the average precision rate using AI is about 84.5% in the database searching and eligibility phase. 4. Shemilt et al [51] reported a precision rate of 50% - 86% using AI in the database searching and eligibility phase. 5. Marshall et al [42] reported a precision rate of 55% using AI in the database searching and eligibility phase. 6. Knafou et al [32] reported a precision rate of only 29.69% using AI in the database searching and eligibility phase. 7. However, Chou et al [35] reported a precision rate of only 0.2% - 8% using AI in the database searching and eligibility phase.
Recall[c]	Recall (also known as sensitivity) refers to the fraction of positive documents that have been accurately identified among all documents for the specified class [32]. Nine (37.5%) studies reported on recall [29,32,35,36,42,43,51,53,54]. All studies reported recall rates in excess of 87%. The average value was about 96.24%. 1. Perlman-Arrow et al [29], Chou et al [35], and Marshall et al [42] found recall rates of even 100% using AI or semiautomated tools in the database searching and eligibility phase. 2. Knafou et al [32], Haas et al [43], and Kamso et al [36] reported a recall rate of 89%, 99.25% and 99.3%, respectively, using AI in the database searching and eligibility phase. 3. Shemilt et al [51] reported a recall rate of 94% - 99% using AI in the database searching and eligibility phase. 4. Khan et al [53] reported a recall rate of 92% - 96% using AI in the data extraction or collection and risk of bias assessment phase. 5. Hair et al [54] reported that the average sensitivity rate using AI is about 95.1% in the database searching and eligibility phase.

XSL•FO
RenderX

Metrics	Explanation
F_1-score[c]	F_1-score refers to the balanced harmonic average between the model precision and recall [32]. Six (25%) studies reported on F_1-score [29,32,43,52-54]. All studies reported F_1-score between 80.47% and 99% after using AI. The average value was about 92.17%. 1. Knafou et al [32], Perlman-Arrow et al [29], and Haas et al [43] reported an F_1-score of 89.2%, 92.6%, and 97.59%, respectively, using AI or semiautomated tools in the database searching and eligibility phase. 2. Le-Khac et al [52] reported an F_1-score of 87% using AI in the data extraction or collection and risk of bias assessment phase. 3. Khan et al [53] reported F_1-scores between 96% and 98% after using AI in the data extraction or collection and risk of bias assessment phase. 4. Hair et al [54] reported that the average F_1-score using AI is about 89.6% in the database searching and eligibility phase.
Area under the receiver operating characteristic curve (AUC-ROC)	AUC-ROC calculates the area under the curve between the true positive rate and the false positive rate [32]. Knafou et al [32] reported higher AUC-ROC performance using AI in the database searching and eligibility phase and had an AUC-ROC performance of 94.25% - 94.77%.
Number needed to read (NNR)	NNR refers to the total number of literature considered within the search divided by the number of literature included from the search [35]. Only 2 (8.3%) studies reported on NNR [29,35]. Perlman-Arrow et al [29] reported an NNR between 1.086 and 1.125 after using a semiautomated tool in the database searching and eligibility phase. Chou et al [35] reported an NNR between 15 and 100 after using AI in the database searching and eligibility phase.
Article relevance	Vaghela et al [44] reported on studies included after searching using AI, and 50.49% were considered relevant to the query in the database searching and eligibility phase.
Utility	
User satisfaction	Perlman-Arrow et al [29] reported that the average satisfaction of users with the tool reached 4.2/5 in the database searching and eligibility phase.
Consistency	Kamso et al [36] reported that consistency in the use of AI between 2 reviewers was assessed using percentage agreement and Kappa scores, revealing a range of percentage agreement from 79.0% to 96.0%, and a variation in Kappa scores from moderate (0.40) to substantial (0.63) in the database searching and eligibility phase.
Article quality	Vaghela et al [44] reported that 64.53% of the included studies possess reliable quality in the database searching and eligibility phase.

[a]AI: artificial intelligence.

[b]Kamso et al [36] achieved an accuracy ranging from 75.9% to 96.9% in research classification using AI in the database searching and eligibility phase. Khan et al [53] reported that the collaborative large language models' accuracy, based on concordant responses in the prompt set, reached 99% in the data extraction or collection and risk of bias assessment phase.

[c]The overall mean recall (96.24%) and F_1-score (92.17%) are the simple averages of study-level values from Table S5 in Multimedia Appendix 1. For studies reporting a range, the midpoint was used as the study-level value.

Efficiency Enhancements Through AI or Semiautomated Tools in LE Synthesis

Three studies showed improved efficiency in the database searching and eligibility phase in terms of 3 indicator terms. A total of 2 (8.3%) studies [29,35] reported on time saving with AI or semiautomated tools, 2 (8.3%) studies [29,42] reported on workload metrics related to the use of AI or semiautomated tools, and 1 study [29] reported a reduction in conflict rates with the use of semiautomated tool, which consequently increases the efficiency.

Accuracy Improvements With AI or Semiautomated Tools in LE Synthesis

A total of 9 and 6 studies that applied AI or semiautomated tools in LE synthesis reported a mean recall rate and a mean F_1-score of 96.24% and 92.17%, respectively. While Khan et al [53] reported a precision rate of even 100% achieved using AI in the data extraction or collection and risk of bias assessment phase. However, in 7 studies, the reported precision rates varied significantly, ranging from 0.2% to 96.07% in the database searching and eligibility phase.

XSL•FO
RenderX

Utility of AI or Semiautomated Tools in LE Synthesis

Three studies reported on the utility of AI or semiautomated tools in the database searching and eligibility phase of LE synthesis, including user satisfaction, consistency, and study quality. Consistency in the use of AI between 2 reviewers was assessed using percentage agreement and Kappa scores [36].

Discussion

Principal Findings

AI or semiautomated tools are actively used to facilitate the process of LE synthesis. We conducted this review to identify the phases of LE synthesis that use AI and explore whether AI can improve the efficiency, accuracy, or utility of LE synthesis.

AI or semiautomated tools have been increasingly used in LE synthesis, particularly in living systematic review. This review discovered that AI or semiautomated tools are most commonly used for data extraction or collection and risk of bias assessment. However, only a few studies have addressed the use of AI or semiautomated systems for publication updates, highlighting the need for further development in this phase.

Diverse types of AI or semiautomated tools were identified in this study. These include the LIvE synthesis platform, AD-SOLES, metaCOVID application, and RobotReviewer LIVE, which are utilized in multiple phases of LE synthesis, indicating their versatility and potential for wider adoption [37,39,40,42,47,54]. The most frequently used AI or semiautomated tools were machine learning classifiers, the LIvE synthesis platform, Covidence, AD-SOLES, and MAGICapp. Furthermore, the rapid rise of AI tools involving LLM types, such as GPT-4-turbo and Claude-3-Opus, has led to their use in LE synthesis. These tools can be suitable for application in multiple or even all phases of LE synthesis, especially in the publication update phase. The application of LLMs to further enhance the efficiency, accuracy, and utility of LE synthesis remains a key focus for researchers and practitioners.

Governments worldwide, particularly those in leading AI nations such as China, the United States, Germany, the United Kingdom, France, and Canada, are especially emphasizing the transformative impact of AI on research and decision-making processes [76,77]. Funding from various sources, including the Economic and Social Research Council, reflects a strong financial commitment to advancing AI technologies in evidence synthesis. Furthermore, a growing number of AI guidance and organizations are emerging to embrace the opportunity that AI has taken in producing LE synthesis. For example, Responsible AI in Evidence SynthEsis has provided recommendations for the main roles of responsible AI in the evidence synthesis ecosystem that are involved in responsible AI use [78]. Furthermore, organizations such as ALIVE aim to improve societal outcomes by producing and utilizing timely, trustworthy, and affordable evidence.

Challenges remain in the application of AI in LE synthesis. Machine learning classifiers suffer from low precision and varying efficiency across different topics [35]. As an example, RobotReviewer LIVE faces challenges in performance variability for complex reviews, limited study types, and data source constraints [42]. Therefore, further research aimed at enhancing the adaptability and stability of AI across various research areas is urgently needed. In addition, ethical issues, data protection measures, and transparency in AI-driven LE synthesis are also key challenges that need to be addressed [79]. At the ethical level, AI is prone to selection bias due to the skewness of its training data, which impairs the inclusivity of evidence, and the mechanism of responsibility attribution remains unclear [80]. Data protection is another area that faces challenges, as research data required for AI training often contain sensitive information, and existing anonymization technologies cannot fully avoid the risk of privacy breaches [81]. Cost considerations in the implementation of AI tools, including initial investment, ongoing operational costs, training expenses, and requirements for hardware and software resources also constitute a significant issue [82].

Policymaking involves judgment, making it more of an art than a science, whereas science is primarily driven by evidence and shapes evidence-informed policymaking [83]. Study has indicated that relying solely on systematic reviews for policymaking is far from sufficient; instead, policymakers need to obtain a more diverse range of synthesized evidence to underpin decision-making [84]. The LE synthesis, especially by incorporating AI into evidence production, can deliver updated evidence to facilitate evidence-informed policymaking. AI could revolutionize policymaking by facilitating ongoing assessments, ensuring that the policies remain aligned with the latest evidence and evolve in response to new information as it emerges [2,5,85]. Furthermore, AI enables policymakers to continuously monitor and assess policies throughout their lifecycle, which allows adaptation to shifting circumstances and evolving societal needs in real time [86]. Furthermore, the advancement of AI capabilities, particularly through LLMs, adds a deeper analytical layer; LLMs can provide nuanced insights and help predict future research directions relevant to policymaking [87]. The application of AI in LE synthesis could transform policy decision-making, advancing policy formulation for policymakers.

Recent advances in AI provide researchers with new transformative capabilities [79]. Van Dijk et al [88] indicated that AI tools are a promising innovation in the current practice of systematic evaluation, and researchers have reported positive experiences with these tools. The use of AI enhances efficiency by significantly reducing researchers' time and workload [2,89]. Manion et al [90] indicated that natural language processing could enhance accuracy and reduce errors through a "human-in-the-loop" approach. The application of AI in LE synthesis has considerably benefited researchers, significantly enhancing their research capabilities.

This LE synthesis will retain its living mode beyond the present publication, consistent with the methodology. This decision is based on two key considerations: (1) the predefined retirement triggers have not been triggered and (2) the Safe and Responsible Use of AI Working Group (Working Group 3) and the Methods & Process Innovation Working Group (Working Group 4) of the Evidence Synthesis Infrastructure Collaborative will benefit from the continuous updates from this LE synthesis to support their future research initiatives [91-94].

XSL•FO
RenderX

Future Research Directions

In the above discussion, we have suggested the advancement of future work across multiple dimensions. From a technical point of view, efforts are needed to address limitations of existing AI tools, such as inadequate precision and poor adaptability, while deepening research into the LLM applications in the publication update phase of LE synthesis. In the realm of ethics and data governance, it is essential to establish responsibility attribution mechanisms and cross-regulatory data governance frameworks, as well as enhance evidence inclusivity and mitigate privacy risks through algorithmic optimization. Methodologically, we recommend the establishment of a standardized evaluation system for AI applications and refining research design and quality assessment protocols to strengthen the evidence base.

Strengths and Limitations

The strengths of this review include the following: (1) it systematically analyzes the types of AI and semiautomated tools used across the 4 phases of LE synthesis and (2) it provides insights into the opportunities and challenges of using AI or semiautomated tools in LE synthesis regarding efficiency, accuracy, and utility. However, this review still has a few limitations. First, study screening was based on whether the studies reported on the tools used in LE synthesis. Second, studies that did not document the use of AI or semiautomated tools in LE synthesis were excluded from this review, which may introduce bias. Third, the focus of our search strategy on "living evidence" terminology may have excluded studies describing AI tools for review updates that used different terminology.

Conclusion

Researchers are actively utilizing various AI and semiautomated tools in LE synthesis, primarily for data extraction or collection and risk of bias assessment, while their application in updating publications remains limited. The use of AI or semiautomated tools in LE synthesis improves efficiency in the database searching and eligibility phase and accuracy in the database searching and eligibility phase, as well as in the data extraction or collection and risk of bias assessment phase. The AI or semiautomated tools demonstrate high accuracy, recall, and F_1-scores, while precision varies across tools. AI or semiautomated tools also demonstrate good performance in terms of utility in the database searching and eligibility phase.

Funding

This work was supported by the Fundamental Research Funds for the Central Universities (lzujbky-2025 - 15) and Gansu Provincial Center for Disease Control and Prevention Research Program (GSJKKY2025-02).

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

XS designed the study, analyzed the data, and drafted the manuscript. ZL designed the study, analyzed the data, drafted the manuscript, and evaluated the quality of included studies. RW, QW, and XL developed the research design. RL and ZY drafted the manuscript and evaluated the quality of the included studies. LF, ZM, and ZP were in charge of data curation. CL, LG, YC, KY, and JL critically reviewed and revised the manuscript. All authors critically revised the study for important intellectual content and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies, included and excluded study information, and methodological quality assessment methods and results.
[DOCX File, 53 KB - jmir_v28i1e76130_app1.docx]

Multimedia Appendix 2
Frequency of artificial intelligence (AI) or semiautomated tools use.
[PNG File, 139 KB - jmir_v28i1e76130_app2.png]

Checklist 1
PRISMA-LSR checklist.
[DOCX File, 29 KB - jmir_v28i1e76130_app3.docx]

References

1. Elliott J, Lawrence R, Minx JC, et al. Decision makers need constantly updated evidence synthesis. Nature 2021 Dec;600(7889):383-385. [doi: 10.1038/d41586-021-03690-1] [Medline: 34912079]

XSL•FO
RenderX

2. Sampson M, Shojania KG, Garritty C, Horsley T, Ocampo M, Moher D. Systematic reviews can be produced and published faster. J Clin Epidemiol 2008 Jun;61(6):531-536. [doi: 10.1016/j.jclinepi.2008.02.004] [Medline: 18471656]
3. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med 2007 Aug 21;147(4):224-233. [doi: 10.7326/0003-4819-147-4-200708210-00179] [Medline: 17638714]
4. Turner T, Lavis JN, Grimshaw JM, Green S, Elliott J. Living evidence and adaptive policy: perfect partners? Health Res Policy Syst 2023 Dec 18;21(1):135. [doi: 10.1186/s12961-023-01085-4] [Medline: 38111030]
5. Elliott JH, Turner T, Clavisi O, et al. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. PLoS Med 2014 Feb;11(2):e1001603. [doi: 10.1371/journal.pmed.1001603] [Medline: 24558353]
6. Thomas J, Noel-Storr A, Marshall I, et al. Living systematic reviews: 2. Combining human and machine effort. J Clin Epidemiol 2017 Nov;91:31-37. [doi: 10.1016/j.jclinepi.2017.08.011] [Medline: 28912003]
7. Tendal B, Vogel JP, McDonald S, et al. Weekly updates of national living evidence-based guidelines: methods for the Australian living guidelines for care of people with COVID-19. J Clin Epidemiol 2021 Mar;131:11-21. [doi: 10.1016/j.jclinepi.2020.11.005] [Medline: 33188858]
8. Elliott JH, Synnot A, Turner T, et al. Living systematic review: 1. Introduction-the why, what, when, and how. J Clin Epidemiol 2017 Nov;91:23-30. [doi: 10.1016/j.jclinepi.2017.08.010] [Medline: 28912002]
9. Schmidt L, Sinyor M, Webb RT, et al. A narrative review of recent tools and innovations toward automating living systematic reviews and evidence syntheses. Z Evid Fortbild Qual Gesundhwes 2023 Sep;181:65-75. [doi: 10.1016/j.zefq.2023.06.007] [Medline: 37596160]
10. Yang Y, Qin J, Lei J, Liu Y. Research status and challenges on the sustainable development of artificial intelligence courses from a global perspective. Sustainability 2023;15(12):9335. [doi: 10.3390/su15129335]
11. Adams CE, Polzmacher S, Wolff A. Systematic reviews: work that needs to be done and not to be done. J Evid Based Med 2013 Nov;6(4):232-235. [doi: 10.1111/jebm.12072] [Medline: 24325416]
12. Suárez A, Jiménez J, Llorente de Pedro M, et al. Beyond the scalpel: assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. Comput Struct Biotechnol J 2024 Dec;24:46-52. [doi: 10.1016/j.csbj.2023.11.058] [Medline: 38162955]
13. Bendersky J, Auladell-Rispau A, Urrútia G, Rojas-Reyes MX. Methods for developing and reporting living evidence synthesis. J Clin Epidemiol 2022 Dec;152:89-100. [doi: 10.1016/j.jclinepi.2022.09.020] [Medline: 36220626]
14. Akl EA, Khabsa J, Iannizzi C, et al. Extension of the PRISMA 2020 statement for living systematic reviews (PRISMA-LSR): checklist and explanation. BMJ 2024 Nov 19;387:e079183. [doi: 10.1136/bmj-2024-079183] [Medline: 39562017]
15. Which phases of living evidence synthesis use artificial intelligence (AI)? an living evidence synthesis. OSF. URL: https://doi.org/10.17605/OSF.IO/4FVDQ [accessed 2026-01-16]
16. Murad MH, Wang Z, Chu H, et al. Proposed triggers for retiring a living systematic review. BMJ Evid Based Med 2023 Oct;28(5):348-352. [doi: 10.1136/bmjebm-2022-112100] [Medline: 36889900]
17. Cochrane Living Systematic Reviews Network. Guidance for the production and publication of Cochrane living systematic reviews: Cochrane Reviews in living mode. : Cochrane; 2019 URL: https://community.cochrane.org/sites/default/files/uploads/inline-files/Transform/201912_LSR_Revised_Guidance.pdf [accessed 2026-01-13]
18. Kozhakhmetova A, Mamyrbayev A, Zhidebekkyzy A, Bilan S. Assessing the impact of artificial intelligence on project efficiency enhancement. Knowl Perform Manag 2024;8(2):109-126. [doi: 10.21511/kpm.08(2).2024.09]
19. Røhl UBU. Automated, administrative decision-making and good administration: friends, foes or complete strangers [Dissertation]. : Aalborg Universitetsforlag; 2022 URL: https://vbn.aau.dk/ws/files/549540893/PHD_UBUR.pdf [accessed 2026-01-19]
20. Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. J Clin Epidemiol 2022 Apr;144:22-42. [doi: 10.1016/j.jclinepi.2021.12.005] [Medline: 34896236]
21. Rashid M, Yi CS, Lawin S, et al. MT14 role of generative artificial intelligence in assisting systematic review process in health research: a systematic review. Value Health 2025 Jul;28(6):S268. [doi: 10.1016/j.jval.2025.04.1124] [Medline: 40848037]
22. Validity assessment tools for evidence synthesis: your one-stop-shop. Latitudes Network. URL: https://www.latitudes-network.org [accessed 2026-01-13]
23. Equator Network - Enhancing the QUAlity and Transparency Of health Research. URL: http://equator-network.org [accessed 2025-08-06]
24. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
25. Santini A, Man A, Voidăzan S. Accuracy of diagnostic tests. J Crit Care Med (Targu Mures) 2021 Jul;7(3):241-248. [doi: 10.2478/jccm-2021-0022] [Medline: 34722928]
26. McArthur A, Klugarova J, Yan H, Florescu S. Chapter 4: systematic reviews of text and opinion. In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis JBI: JBI; 2020. [doi: 10.46658/JBIRM-17-04]

XSL•FO
RenderX

27. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017 Sep 21;358:j4008. [doi: 10.1136/bmj.j4008] [Medline: 28935701]
28. Welsh EJ, Normansell RA, Cates CJ. Assessing the methodological quality of systematic reviews. NPJ Prim Care Respir Med 2015 Mar 19;25:15019. [doi: 10.1038/npjpcrm.2015.19] [Medline: 25789911]
29. Perlman-Arrow S, Loo N, Bobrovitz N, Yan T, Arora RK. A real-world evaluation of the implementation of NLP technology in abstract screening of a systematic review. Res Synth Methods 2023 Jul;14(4):608-621. [doi: 10.1002/jrsm.1636] [Medline: 37230483]
30. Higgins JPT, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions Version 6.5: Cochrane; 2024. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current [accessed 2026-01-13]
31. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ 2020 Jan 16;368:l6890. [doi: 10.1136/bmj.l6890] [Medline: 31948937]
32. Knafou J, Haas Q, Borissov N, et al. Ensemble of deep learning language models to support the creation of living systematic reviews for the COVID-19 literature. Syst Rev 2023 Jun 5;12(1):94. [doi: 10.1186/s13643-023-02247-9] [Medline: 37277872]
33. Matl S, Brosig R, Baust M, Navab N, Demirci S. Vascular image registration techniques: a living review. Med Image Anal 2017 Jan;35:1-17. [doi: 10.1016/j.media.2016.05.005] [Medline: 27294558]
34. Schmidt L, Finnerty Mutlu AN, Elmore R, Olorisade BK, Thomas J, Higgins JPT. Data extraction methods for systematic review (semi)automation: update of a living systematic review. F1000Res 2021;10:401. [doi: 10.12688/f1000research.51117.3] [Medline: 34408850]
35. Chou R, Dana T, Shetty KD. Testing a Machine Learning Tool for Facilitating Living Systematic Reviews of Chronic Pain Treatments: Agency for Healthcare Research and Quality; 2020. [Medline: 33175480]
36. Kamso MM, Pardo JP, Whittle SL, et al. Crowd-sourcing and automation facilitated the identification and classification of randomized controlled trials in a living review. J Clin Epidemiol 2023 Dec;164:1-8. [doi: 10.1016/j.jclinepi.2023.10.007] [Medline: 37865299]
37. Riaz IB, Sipra Q, Naqvi SAA, et al. Quantifying absolute benefit for adjuvant treatment options in renal cell carcinoma: a living interactive systematic review and network meta-analysis. Crit Rev Oncol Hematol 2022 Jul;175:103706. [doi: 10.1016/j.critrevonc.2022.103706] [Medline: 35537621]
38. Butler AR, Hartmann-Boyce J, Livingstone-Banks J, Turner T, Lindson N. Optimizing process and methods for a living systematic review: 30 search updates and three review updates later. J Clin Epidemiol 2024 Feb;166:111231. [doi: 10.1016/j.jclinepi.2023.111231] [Medline: 38043829]
39. Riaz IB, He H, Ryu AJ, et al. A living, interactive systematic review and network meta-analysis of first-line treatment of metastatic renal cell carcinoma. Eur Urol 2021 Dec;80(6):712-723. [doi: 10.1016/j.eururo.2021.03.016] [Medline: 33824031]
40. Riaz IB, Siddiqi R, Islam M, et al. Adjuvant tyrosine kinase inhibitors in renal cell carcinoma: a concluded living systematic review and meta-analysis. JCO Clin Cancer Inform 2021 May;5:588-599. [doi: 10.1200/CCI.21.00035] [Medline: 34043431]
41. Hair K, Wilson E, Wong C, Tsang A, Macleod M, Bannach-Brown A. Systematic online living evidence summaries: emerging tools to accelerate evidence synthesis. Clin Sci (Lond) 2023 May 31;137(10):773-784. [doi: 10.1042/CS20220494] [Medline: 37219941]
42. Marshall IJ, Trikalinos TA, Soboczenski F, et al. In a pilot study, automated real-time systematic review updates were feasible, accurate, and work-saving. J Clin Epidemiol 2023 Jan;153:26-33. [doi: 10.1016/j.jclinepi.2022.08.013] [Medline: 36150548]
43. Haas Q, Alvarez DV, Borissov N, et al. Utilizing artificial intelligence to manage COVID-19 scientific evidence torrent with Risklick AI: a critical tool for pharmacology and therapy development. Pharmacology 2021;106(5-6):244-253. [doi: 10.1159/000515908] [Medline: 33910199]
44. Vaghela U, Rabinowicz S, Bratsos P, et al. Using a secure, continually updating, web source processing pipeline to support the real-time data synthesis and analysis of scientific literature: development and validation study. J Med Internet Res 2021 May 6;23(5):e25714. [doi: 10.2196/25714] [Medline: 33835932]
45. Karakülah G, Suner A, Adlassnig KP, Samwald M. A data-driven living review for pharmacogenomic decision support in cancer treatment. Stud Health Technol Inform 2012;180:688-692. [doi: 10.3233/978-1-61499-101-4-688] [Medline: 22874279]
46. Xin Y, Nevill CR, Nevill J, et al. Feasibility study for interactive reporting of network meta-analysis: experiences from the development of the Metainsight COVID-19 app for stakeholder exploration, re-analysis and sensitivity analysis from living systematic reviews. BMC Med Res Methodol 2022 Jan 22;22(1):26. [doi: 10.1186/s12874-022-01507-x] [Medline: 35065603]
47. Evrenoglou T, Boutron I, Seitidis G, Ghosn L, Chaimani A. metaCOVID: a web-application for living meta-analyses of COVID-19 trials. Res Synth Methods 2023 May;14(3):479-488. [doi: 10.1002/jrsm.1627] [Medline: 36772980]
48. Kaiser K, Miksch S. Versioning computer-interpretable guidelines: semi-automatic modeling of “living guidelines” using an information extraction method. Artif Intell Med 2009 May;46(1):55-66. [doi: 10.1016/j.artmed.2008.08.009] [Medline: 18950994]
49. Skinner G, Cooke R, Keum J, et al. Dynameta: a dynamic platform for ecological meta-analyses in R Shiny. SoftwareX 2023 Jul;23:101439. [doi: 10.1016/j.softx.2023.101439]

XSL•FO
RenderX

50. McDonald S, Hill K, Li HZ, Turner T. Evidence surveillance for a living clinical guideline: case study of the Australian stroke guidelines. Health Info Libr J 2023 Nov 9. [doi: 10.1111/hir.12515] [Medline: 37942888]
51. Shemilt I, Arno A, Thomas J, et al. Cost-effectiveness of Microsoft Academic Graph with machine learning for automated study identification in a living map of coronavirus disease 2019 (COVID-19) research. Wellcome Open Res 2024;6:210. [doi: 10.12688/wellcomeopenres.17141.2] [Medline: 38686019]
52. Le-Khac UN, Bolton M, Boxall NJ, Wallace SMN, George Y. Living review framework for better policy design and management of hazardous waste in Australia. Sci Total Environ 2024 May 10;924:171556. [doi: 10.1016/j.scitotenv.2024.171556] [Medline: 38458450]
53. Khan MA, Ayub U, Naqvi SAA, et al. Collaborative large language models for automated data extraction in living systematic reviews. J Am Med Inform Assoc 2025 Apr 1;32(4):638-647. [doi: 10.1093/jamia/ocae325] [Medline: 39836495]
54. Hair K, Wilson E, Maksym O, Macleod MR, Sena ES. A systematic online living evidence summary of experimental Alzheimer's disease research. J Neurosci Methods 2024 Sep;409:110209. [doi: 10.1016/j.jneumeth.2024.110209] [Medline: 38964475]
55. Grbin L, Nichols P, Russell F, Fuller-Tyszkiewicz M, Olsson CA. The development of a living knowledge system and implications for future systematic searching. J Aust Libr Inf Assoc 2022 Jul 3;71(3):275-292. [doi: 10.1080/24750158.2022.2087954]
56. Hearnden J, Dudoit K, Kim E, Tremblay G, Forsythe A. PMU118 use of computer-assisted methods to realize the concept of a living systematic review via an online platform. Value Health 2019 Nov;22:S729. [doi: 10.1016/j.jval.2019.09.1736]
57. Evrenoglou T, Boutron I, Chaimani A. metaCOVID: an R-Shiny application for living meta-analyses of COVID-19 trials. medRxiv. Preprint posted online on Sep 10, 2021. [doi: 10.1101/2021.09.07.21263207]
58. Stoll A, Wilms L, Ziegele M. Developing an incivility dictionary for German online discussions—a semi-automated approach combining human and artificial knowledge. Commun Methods Meas 2023 Apr 3;17(2):131-149. [doi: 10.1080/19312458.2023.2166028]
59. Meza N, Pérez-Bracchiglione J, Pérez I, et al. Angiotensin-converting-enzyme inhibitors and angiotensin II receptor blockers for COVID-19: a living systematic review of randomized clinical trials. Medwave 2021 Mar 3;21(2):e8105. [doi: 10.5867/medwave.2021.02.8105] [Medline: 33830976]
60. Verdejo C, Vergara-Merino L, Meza N, et al. Macrolides for the treatment of COVID-19: a living, systematic review. Medwave 2020 Dec 14;20(11):e8074. [doi: 10.5867/medwave.2020.11.8073] [Medline: 33361755]
61. Baladia E, Pizarro AB, Ortiz-Muñoz L, Rada G. Vitamin C for COVID-19: a living systematic review. Medwave 2020 Jul 28;20(6):e7978. [doi: 10.5867/medwave.2020.06.7978] [Medline: 32759894]
62. Rada G, Corbalán J, Rojas P, COVID-19 L·OVE Working Group. Cell-based therapies for COVID-19: a living, systematic review. Medwave 2020 Dec 17;20(11):e8079. [doi: 10.5867/medwave.2020.11.8078] [Medline: 33382060]
63. Gates M, Elliott SA, Gates A, et al. LOCATE: a prospective evaluation of the value of leveraging ongoing citation acquisition techniques for living evidence syntheses. Syst Rev 2021 Apr 19;10(1):116. [doi: 10.1186/s13643-021-01665-x] [Medline: 33875014]
64. Piechotta V, Iannizzi C, Chai KL, et al. Convalescent plasma or hyperimmune immunoglobulin for people with COVID-19: a living systematic review. Cochrane Database Syst Rev 2021 May 20;5(5):CD013600. [doi: 10.1002/14651858.CD013600.pub4] [Medline: 34013969]
65. Verdugo-Paiva F, Acuña MP, Solá I, Rada G, COVID-19 L·OVE Working Group. Remdesivir for the treatment of COVID-19: a living systematic review. Medwave 2020 Dec 9;20(11):e8080. [doi: 10.5867/medwave.2020.11.8080] [Medline: 33361753]
66. Shackelford GE, Martin PA, Hood ASC, Christie AP, Kulinskaya E, Sutherland WJ. Dynamic meta-analysis: a method of using global evidence for local decision making. BMC Biol 2021 Feb 17;19(1):33. [doi: 10.1186/s12915-021-00974-w] [Medline: 33596922]
67. Verdugo-Paiva F, Izcovich A, Ragusa M, Rada G. Lopinavir-ritonavir for COVID-19: a living systematic review. Medwave 2020 Jul 15;20(6):e7967. [doi: 10.5867/medwave.2020.06.7966] [Medline: 32678815]
68. Iannizzi C, Chai KL, Piechotta V, et al. Convalescent plasma for people with COVID-19: a living systematic review. Cochrane Database Syst Rev 2023 Feb 1;2(2):CD013600. [doi: 10.1002/14651858.CD013600.pub5] [Medline: 36734509]
69. Sommer I, Ledinger D, Thaler K, et al. Outpatient treatment of confirmed COVID-19: a living, rapid evidence review for the American College of Physicians (Version 2). Ann Intern Med 2023 Oct;176(10):1377-1385. [doi: 10.7326/M23-1626] [Medline: 37722115]
70. Verdugo-Paiva F, Vergara C, Ávila C, et al. COVID-19 living overview of evidence repository is highly comprehensive and can be used as a single source for COVID-19 studies. J Clin Epidemiol 2022 Sep;149:195-202. [doi: 10.1016/j.jclinepi.2022.05.001] [Medline: 35597369]
71. Paul D, Chakdar D, Saha S, Mathew J. Online research topic modeling and recommendation utilizing multiview autoencoder-based approach. IEEE Trans Comput Soc Syst 2024;11(1):1013-1022. [doi: 10.1109/TCSS.2023.3253502]
72. Vergara-Merino L, Verdejo C, Carrasco C, Vargas-Peirano M. Living systematic review: new inputs and challenges. Medwave 2020 Dec 23;20(11):e8092. [doi: 10.5867/medwave.2020.11.8092] [Medline: 33382391]
73. Elbers S, Wittink H, Kaiser U, et al. Living systematic reviews in rehabilitation science can improve evidence-based healthcare. Syst Rev 2021 Dec 7;10(1):309. [doi: 10.1186/s13643-021-01857-5] [Medline: 34876231]

XSL•FO
RenderX

74. Elvidge J, Hopkin G, Narayanan N, Nicholls D, Dawoud D. Diagnostics and treatments of COVID-19: two-year update to a living systematic review of economic evaluations. Front Pharmacol 2023;14:1291164. [doi: 10.3389/fphar.2023.1291164] [Medline: 38035028]
75. Winters M, Lyng KD, Holden S, et al. Infographic. Comparative effectiveness of treatments for patellofemoral pain: a living systematic review with network meta-analysis. Br J Sports Med 2021 Nov;55(22):1311-1312. [doi: 10.1136/bjsports-2021-104360] [Medline: 34244170]
76. Katanandov SL, Kovalev AA. Technological development of modern states: artificial intelligence in public administration. State and Municipal Management Scholar Notes 2023 Mar;1(1):174-182. [doi: 10.22394/2079-1690-2023-1-1-174-182]
77. Medaglia R, Gil-Garcia JR, Pardo TA. Artificial intelligence in government: taking stock and moving forward. Soc Sci Comput Rev 2021;41(1):123-140. [doi: 10.1177/08944393211034087]
78. Thomas J, Flemyng E, Noel-Storr A, et al. Responsible AI in Evidence SynthEsis (RAISE): guidance and recommendations. : Open Science Framework; 2024 URL: https://osf.io/fwaud/files/cn7x4 [accessed 2026-01-13]
79. Filetti S, Fenza G, Gallo A. Research design and writing of scholarly articles: new artificial intelligence tools available for researchers. Endocrine 2024 Sep;85(3):1104-1116. [doi: 10.1007/s12020-024-03977-z] [Medline: 39085566]
80. Ramnani S. Exploring ethical considerations of artificial intelligence in educational settings: an examination of bias, privacy, and accountability. International Journal of Novel Research and Development 2024;9(2):b173-b191. [doi: 10.1729/Journal.37869]
81. Li Y, Shao S, He Y, et al. Rethinking data protection in the (generative) artificial intelligence era. arXiv. Preprint posted online on Jul 3, 2025. [doi: 10.48550/arXiv.2507.03034]
82. Umeh II, Umeh KC. A comparative analysis of AI system development tools for improved outcomes. International Journal of Sustainability Management and Information Technologies 2025;11:1-20. [doi: 10.11648/j.ijsmit.20251101.11]
83. Ramírez G. Improving the health of populations—evidence for policy and practice action. J Evid Based Med 2009 Nov;2(4):216-219. [doi: 10.1111/j.1756-5391.2009.01044.x] [Medline: 21349019]
84. Manson H. Systematic reviews are not enough: policymakers need a greater variety of synthesized evidence. J Clin Epidemiol 2016 May;73:11-14. [doi: 10.1016/j.jclinepi.2015.08.032] [Medline: 26912122]
85. Berger-Tal O, Wong BBM, Adams CA, et al. Leveraging AI to improve evidence synthesis in conservation. Trends Ecol Evol 2024 Jun;39(6):548-557. [doi: 10.1016/j.tree.2024.04.007] [Medline: 38796352]
86. Jacob S. Artificial intelligence and the future of evaluation: from augmented to automated evaluation. Digit Gov Res Pract 2025 Mar 31;6(1):1-10. [doi: 10.1145/3696009]
87. Head CB, Jasper P, McConnachie M, Raftree L, Higdon G. Large language model applications for evaluation: opportunities and ethical implications. New Dir Eval 2023 Jun;2023(178-179):33-46. [doi: 10.1002/ev.20556]
88. van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, van der Palen J, Doggen CJM, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open 2023 Jul 7;13(7):e072254. [doi: 10.1136/bmjopen-2023-072254] [Medline: 37419641]
89. Thomas IN, Roche P, Grêt-Regamey A. Harnessing artificial intelligence for efficient systematic reviews: a case study in ecosystem condition indicators. Ecol Inform 2024 Nov;83:102819. [doi: 10.1016/j.ecoinf.2024.102819]
90. Manion FJ, Du J, Wang D, et al. Accelerating evidence synthesis in observational studies: development of a living natural language processing-assisted intelligent systematic literature review system. JMIR Med Inform 2024 Oct 23;12:e54653. [doi: 10.2196/54653] [Medline: 39441204]
91. Glanville J. The role of AI tools in developing search strategies and identifying evidence for systematic reviews. : Evidence Synthesis Ireland; 2025 URL: https://evidencesynthesisireland.ie/wp-content/uploads/2025/05/ESI-2025-AI-tools-Julie-Glanville.pdf [accessed 2026-01-13]
92. New Joint AI Methods Group: guiding responsible use of AI in evidence synthesis. Joanna Briggs Institute (JBI). URL: https://jbi.global/news/article/new-joint-ai-methods-group [accessed 2026-01-13]
93. Roadmap. Evidence Synthesis Infrastructure Collaborative. 2024. URL: https://evidencesynthesis.atlassian.net/wiki/spaces/ESE/pages/344817670/English [accessed 2026-01-13]
94. Scotcher S. Evidence Synthesis Infrastructure Collaborative. European Evaluation Society. 2025. URL: https://europeanevaluation.org/events/evidence-synthesis-infrastructure-collaborative [accessed 2026-01-13]

Abbreviations

AI: artificial intelligence
DTA: diagnostic test accuracy
JBI: Joanna Briggs Institute
LE: living evidence
LIvE: Living Interactive Evidence
LLM: large language model
PRISMA-LSR : Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 statement for living systematic reviews

XSL•FO
RenderX

QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies version 2

Edited by A Coristine; submitted 17.Apr.2025; peer-reviewed by C Bhagat, E Ting; accepted 15.Dec.2025; published 27.Jan.2026.

Please cite as:
Song X, Lian Z, Wang R, Li R, Yang Z, Luo X, Feng L, Ma Z, Pu Z, Wang Q, Ge L, Li C, Chen Y, Yang K, Lavis J
The Phases of Living Evidence Synthesis Using AI: Living Evidence Synthesis (Version 1)
J Med Internet Res 2026;28:e76130
URL: https://www.jmir.org/2026/1/e76130
doi:10.2196/76130

© Xuping Song, Zhenjie Lian, Rui Wang, Ruixin Li, Zhenzhen Yang, Xufei Luo, Lei Feng, Zhiming Ma, Zhen Pu, Qi Wang, Long Ge, Caihong Li, Yaolong Chen, Kehu Yang, John Lavis. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Patient Benefits in the Context of Sepsis-Related AI-Based Clinical Decision Support Systems: Scoping Review

Pascal Raszke[1], MA; Godwin Denk Giebel[1], Dr rer medic; Jürgen Wasem[1], Prof Dr; Michael Adamzik[2], Prof Dr Med; Hartmuth Nowak[3], PD, Dr med; Lars Palmowski[2], Dr med; Philipp Heinz[4], Dipl-Kfm; Nina Timmesfeld[5], Prof Dr; Marianne Tokic[5], MSc; Frank Martin Brunkhorst[6], Prof Dr Med; Nikola Blase[1], Dr med

[1]Institute for Health Care Management and Research, University of Duisburg-Essen, Thea-Leymann-Str. 9, Essen, Germany

[2]Department of Anesthesiology, Intensive Care Medicine and Pain Therapy, Ruhr University Bochum, Knappschaft Kliniken University Hospital Bochum, Bochum, Germany

[3]Department of Anesthesiology, Intensive Care Medicine and Pain Therapy, Center for Artificial Intelligence, Medical Informatics and Data Science, Ruhr-University Bochum, Knappschaft Kliniken University Hospital Bochum, Bochum, Germany

[4]Knappschaft Kliniken GmbH, Recklinghausen, Germany

[5]Department of Medical Informatics, Biometry and Epidemiology, Ruhr University Bochum, Bochum, Germany

[6]Institute of Infectious Diseases and Infection Control, Jena University Hospital, Jena, Germany

Corresponding Author:
Pascal Raszke, MA
Institute for Health Care Management and Research, University of Duisburg-Essen, Thea-Leymann-Str. 9, Essen, Germany

Abstract

Background: Global digitalization continues to advance, extending its influence into medicine and health care systems worldwide. In recent years, substantial advancements have been made in the research and development of artificial intelligence (AI), raising questions about its potential in medicine. The integration and application of AI in intensive care medicine, particularly in sepsis treatment, presents significant potential for advancing patient outcomes and enhancing patient-relevant benefits. However, a comprehensive and systematic overview of the full spectrum of patient-relevant benefits associated with AI-based clinical decision support systems (CDSS) remains lacking.

Objective: This scoping review aimed to identify and categorize evidence on patient-relevant benefits of AI-based CDSS in sepsis care.

Methods: Systematic research was conducted in 4 electronic databases: MEDLINE via PubMed, Embase, the ACM Digital Library, and IEEE Xplore. In addition, a comprehensive search on the websites of relevant international organizations, along with a citation search of the included articles, was conducted. Articles were included if they (1) focused on sepsis and (2) described patient-relevant benefits of AI-based CDSS. Articles published between January 1, 2008, and March 2, 2023, were considered for inclusion. Study selection was performed independently by 2 reviewers. The manuscript was drafted in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist. The analysis of the included articles was conducted using the program MAXQDA (VERBI Software GmbH), with systemization finalized in a consensus workshop.

Results: A total of 3368 records were identified across the 4 databases, of which 24 met the inclusion criteria and were included in the scoping review. The additional search on international websites and in reference lists identified 6 more relevant articles, resulting in 30 included studies. Of these, 20 were quantitative, comprising 7 prospective and 13 retrospective designs. In addition, 1 qualitative study, 1 mixed methods study, 6 review articles, and 2 articles from institutional websites were included. Patient-relevant benefits were systematized in six main categories: (1) prediction, (2) earlier treatment and prioritization of high-risk patients, (3) individualized therapy, (4) improved patient outcomes (including improved Sequential Organ Failure Assessment score, reduced length of stay, and reduced mortality), (5) general improvements in care, and (6) reduced readmission rate.

Conclusions: This scoping review underscores the potential of AI-based CDSS to positively impact patient-relevant benefits, particularly in sepsis care, where they demonstrate considerable promise for improving intensive care. However, the majority of the identified studies rely on retrospective database analyses. Future research should focus on validating these findings through prospective studies.

(J Med Internet Res 2026;28:e76772) doi:10.2196/76772

XSL•FO
RenderX

KEYWORDS

medical informatics; artificial intelligence; machine learning; computational intelligence; clinical decision support systems; CDSS; decision support; sepsis; bloodstream infection

Introduction

The treatment of infectious diseases has historically resulted in medical progress, exemplified by antibiotics and vaccines. Despite all medical advances, infections remain a major global cause of morbidity and mortality [1,2]. Sepsis, defined as "life-threatening organ dysfunction caused by a dysregulated host response to infection" [3], remains among the top contributors to worldwide mortality. It accounts for 30% - 50% of all hospital deaths in high-income countries, such as the United States [1], and approximately 11 million annual deaths worldwide [2]. Sepsis is a heterogeneous syndrome with variable phenotypes and outcomes. Thus, the interpretation of initial symptoms can be difficult for health care providers [3,4].

The effectiveness and accuracy of established rule-based scoring systems used for the assessment of patients in the intensive care unit (ICU), such as the systemic inflammatory response syndrome (SIRS) criteria, which were historically of importance, the sequential organ failure assessment (SOFA) score or the quickSOFA (qSOFA) score, the acute physiology and chronic health evaluation II (APACHE II) score, or the national early warning score 2 (NEWS2), is limited. This is partly because these scoring systems are not always specifically developed for sepsis patients and are therefore of limited use to health care providers in this context [5,6]. Nevertheless, timely identification and treatment are crucial to enhance patient outcomes [7-9], as untreated sepsis can progress to septic shock, exacerbating the patient's condition [10] and leading to multiple organ failure, which carries an even higher mortality rate than sepsis itself [11].

This is where recent developments in artificial intelligence (AI) become particularly relevant, as they are considered to hold substantial potential for improving sepsis diagnostics. Especially machine learning (ML), a branch of AI, has the ability to rapidly analyze vast amounts of data, exceeding human capacity to process. By evaluating numerous data points, ML can derive conclusions and recognize correlations that a human health care provider would be incapable of identifying. This is why ML is well-suited as a technological foundation for clinical decision support systems (CDSS), particularly in the complex clinical picture of sepsis [3]. The use of ML in the development of CDSS can make the sepsis diagnosis more reliable, with the prospect of long-term improvements in patient outcomes. Machine learning algorithms (MLAs) demonstrated potential to enhance patient-relevant benefits in distinct studies. Documented benefits include, for example, reductions in sepsis-related mortality and the average hospital length of stay (LOS). Additionally, MLAs facilitate earlier interventions, such as the timely administration of antibiotics [12-14].

Despite the high clinical relevance of sepsis and significant advancements in both the availability of digital patient data and in the field of ML, the real-world application of AI-based CDSS remains negligible. The majority of these algorithms remain in the prototype phase, with deployment limited to a single hospital or a single hospital operator. This gap is highlighted by an analysis of the Food and Drug Administration's database of medical devices using AI or ML. As of April 2025, none of the over 1000 listed products are specifically dedicated to intensive care [15], the medical field at the forefront of sepsis treatment. This illustrates the discrepancy between technological progress and its real-world implementation in the critical care environment. For AI-based CDSS to be successfully implemented in clinical practice, it is a necessary prerequisite that they demonstrate tangible added value. Accordingly, patient-relevant benefits should constitute a primary focus.

The research objective of the present study differs from those of previous scoping reviews on AI-based CDSS in sepsis care. Certain reviews focused specifically on neonatal [16] or pediatric [17] sepsis, whereas others concentrated on tasks for which MLAs were designed—such as risk assessment, treatment planning, or process support—and thus focused on the process of medical service delivery rather than on actual patient-relevant benefits [18] or on the actual design of the CDSS and its intended users [19]. Importantly, none of the aforementioned reviews [16-19] focused exclusively on patient-relevant benefits. Furthermore, several existing scoping reviews used narrow methodological approaches, for example, being restricted to a single ML method [17] or considering only antibiotic treatment of sepsis [16]. To the authors' knowledge, no other scoping review has explicitly examined the patient-relevant benefits of AI-based CDSS in the context of sepsis while applying a broad and exploratory methodological approach, without restrictions regarding the ML methods used or the types of patient-relevant benefits assessed. Accordingly, the objective of the present study is to identify patient-relevant benefits of AI-based CDSS in sepsis care compared with the current standard of care, thereby addressing this research gap, as patient-relevant benefits constitute a meaningful benchmark for evaluating the value of any medical innovation. In this context, a taxonomy of benefits comprising 6 main categories has been developed.

This scoping review was conducted within the framework of the KI@work (User-Oriented Requirements for AI-Based Clinical Decision Support Systems) project, which is funded by the German Federal Joint Committee (funding code: 01VSF22050). The research project is led by the Institute for Health Care Management and Research at the University of Duisburg-Essen. Consortium partners include the Department of Anesthesiology, Intensive Care Medicine and Pain Therapy at the University Hospital Knappschaftskrankenhaus Bochum, the Knappschaft Kliniken GmbH, the Department of Medical Informatics, Biometry and Epidemiology at the Ruhr University Bochum and the German Sepsis Society. This scoping review addressed 2 additional research questions. However, to ensure a coherent presentation of the findings, this article focuses exclusively on patient-relevant benefits.

XSL•FO
RenderX

Methods

Overview

This scoping review is based on the methodology framework of the Joanna Briggs Manual for evidence synthesis [20], a further development of the work of Arksey and O'Malley [21] and Levac et al [22]. The review process followed the five stages originally described by Arksey and O'Malley: (1) identifying the research question, (2) identifying relevant studies, (3) study selection, (4) charting the data, and (5) collating, summarizing, and reporting the results [21]. The manuscript was prepared according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist by Tricco et al [23] (Checklist 1). As scoping reviews encompass a broad range of study types in order to present a comprehensive overview of the research field [20-22], comparability between studies is limited. Consequently, no formal quality appraisal was conducted. Although no distinct protocol for the scoping review was published, the methodology was described in detail in a protocol for the overarching multimethod research project [24].

Search Strategy

The development of the search strategy commenced with an initial limited search in MEDLINE via PubMed and Embase to identify relevant search terms. Subsequently, the identified terms were discussed in recurring team discussions within the consortium. In a third step, the consented search terms were combined into search queries.

The electronic databases MEDLINE via PubMed, Embase, as well as the ACM Digital Library and IEEE Xplore, were searched for relevant literature on March 2, 2023. The databases were selected to ensure that the interdisciplinary research question could be adequately addressed from both a medical and a computer science perspective. The search string was developed using the PCC (population=persons with or at risk of sepsis, concept=CDSS, and context=AI) framework. The MEDLINE via PubMed search string was quality-assured by the chief librarian at the library of the University Medical Centre Essen before the database search was conducted. The other 3 search strings were developed based on the same quality assurance principles as the MEDLINE via PubMed search query. The individual search terms were limited to occurrences in title, abstract, and keyword searches but were supplemented by indexing terms (MeSH and Emtree) and truncations. The final search strategies for each database can be found in Multimedia Appendices 1-4.

In agreement with ML experts (NT, HN), the search was limited to articles published in the last 15 years. Further explanation for the time restriction is provided in the discussion of this article. The search was restricted to English and German. In cases of missing full texts, the interlibrary loan service of the University of Duisburg-Essen was used. If that approach was not successful, the reviewers contacted the respective authors of the papers of interest. The identified citations were imported into the reference management program Endnote 20 (Clarivate Analytics).

In addition to the systematic search of electronic databases, a structured search for gray literature (eg, working papers and guidelines) from various governmental and nongovernmental stakeholders was conducted via their websites. The selection of countries included in the search was based on the results of the Bertelsmann #SmartHealthSystem study, which examined the degree of digitalization of various health care systems in 2018. It was assumed that the prospect of identifying information on AI-based CDSS would be particularly high in countries with highly digitalized health care systems. According to the Bertelsmann study, this applies to the health care systems of Canada, Denmark, Estonia, Israel, and Spain. In addition, 3 large economies—Germany, the United Kingdom, and the United States—were included in the structured research. Alongside institutional websites from these countries, websites of relevant international stakeholders were also examined. These included the World Health Organization (WHO) and the Organisation for Economic Co-operation and Development (OECD), as well as websites of international sepsis, intensive care, and medical informatics associations. Further information about included websites can be found in Multimedia Appendix 5. To supplement further evidence, reference lists of articles identified through the systematic and structured search were screened, and the cited articles were subsequently assessed for eligibility. If eligible, the referenced articles were included in the scoping review.

Eligibility Criteria

Exploratory research and internal discussions contributed to the development of inclusion and exclusion criteria, which were refined iteratively during the initial stages of the research process. The search strategy was designed to address 3 different research questions. Studies were considered for inclusion if they described (1) patient-relevant benefits of AI-based CDSS in the context of sepsis as well as (2) problems in their development, implementation, or application, or (3) suggestions for improving these processes. Patient-relevant benefits were identified entirely exploratively and categorized independently of existing frameworks, allowing AI-based CDSS benefits to be classified without reliance on established definitions or patient-relevant endpoints. This approach provides a comprehensive and complete overview of the potential benefits of this emerging technology, without constraining the findings of this paper to predefined frameworks and definitions. Patient-relevant benefits were defined as the positive impact of an intervention on patients, irrespective of whether these comprise general qualitative observations or specific, measurable quantitative endpoints. Specific inclusion and exclusion criteria were developed for each research question to ensure a tailored approach to the unique scope of each question. AI was defined as ML-based algorithms that operate as a "black box" for the user (physician or caregiver), meaning their output is not directly interpretable for health care providers. Consequently, all ML-based technologies developed through data-driven training and sufficiently complex to preclude full comprehension by the user were eligible for inclusion. In contrast, rule-based algorithms, such as those relying on SIRS or SOFA criteria, did not meet this definition and were therefore excluded in this review. Moreover, earlier diagnosis facilitated by AI-based

XSL•FO
RenderX

CDSS was not considered a patient-relevant benefit, as earlier diagnosis itself has no impact on patient outcomes. It is the interventions that follow an earlier diagnosis—such as increased attention by health care providers to patients developing sepsis or earlier initiation of treatment—that positively influence patient-relevant benefits. Accordingly, these parameters are pertinent to the scope of this review. Articles were selected regardless of the research method used. The inclusion criteria are presented in Textbox 1.

Exclusion criteria for this review were not answering the research question, an exclusively technical description of the algorithms developed, or exclusively mathematical approaches not providing evidence for patient-relevant benefit. In addition, articles were excluded if they focused only on the evaluation of binary classifiers such as sensitivity, specificity, positive predictive value, or negative predictive value, as the superiority of AI-based algorithms over rule-based scores was considered a prerequisite for such systems. AI-based CDSS developed exclusively for neonates and/or children or for animals were also not included, because (1) the treatment of neonatal or pediatric sepsis patients differs significantly from the treatment of adult patients [25-27] and (2) the focus of the study is on human sepsis. Articles published before 2008 were also excluded, as were those written in languages other than English or German. Research protocols, conference abstracts, letters to the editor, and articles that were only expressions of opinions were also excluded. The exclusion criteria are listed in Textbox 1.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Articles focusing on sepsis and
- Involving AI-based CDSS, that
  - Describe patient-relevant benefits, or
  - Describe problems with development, implementation, or application, or
  - Describe strategies for success

Exclusion criteria

- Exclusively technical description of systems, or
- Focus on description of the evaluation of binary classifiers, or
- Articles describing AI-based CDSS for neonates and children or animals, or
- Not addressing any of the research questions in more detail, or
- Research protocols, conference abstracts, theses, letters to the editor, or expression of opinions, or
- Article published before 2008, or
- Language other than English or German

Evidence Screening, Selection, and Data Extraction

After identification and deletion of duplicates, title and abstract screening was conducted independently by 2 reviewers (PR and GDG) to decide whether an article was eligible for full-text screening. In a second step, the same 2 reviewers conducted a full-text screening of the included articles against the inclusion and exclusion criteria. In case of disagreement between the 2 reviewers during step 2 of the screening process, other members of the study team (NB, HN, and NT) were involved to decide whether an article was eligible for inclusion.

MAXQDA (VERBI Software GmbH) software was used to identify and tag relevant content in the included articles and to precategorize the patient-relevant benefit categories (PR) using an inductive coding approach. The preliminary categories were discussed and further refined in an in-person workshop based on the affinity mapping technique (PR, NB, and GDG). For this purpose, all relevant text passages were printed as snippets and physically assigned to the respective preliminary categories before being refined and finalized during the workshop. Each assignment was discussed in detail until full consensus among all 3 team members was reached. The results of the workshop were subsequently digitalized in Microsoft Excel. In addition to the patient-relevant benefits of AI-based CDSS, metadata, such as participating authors, year of publication, country of study, database for MLA, or study type, were extracted and summarized (see Multimedia Appendix 6).

Analysis and Presentation of Results

The results of the included studies were summarized descriptively, and analysis was conducted to derive implications for policy, practice, and research. The patient-relevant benefits were grouped into 6 main categories. The main categories were presented in tabular form in an Excel file and diagrammatically. The patient-relevant benefit categories are presented chronologically in Multimedia Appendix 7.

XSL•FO
RenderX

Results

Selection of Sources of Evidence

Selection Process

In the systematic search, a total of 3368 titles and abstracts were retrieved. After removing 850 duplicates, 2518 articles remained for screening (Figure 1). Of these, 141 articles were screened for full text, and 39 met the inclusion criteria. Among these, 24 provided statements on patient-relevant benefits [28-51]. In addition, reference lists of the articles identified through the systematic search were analyzed, resulting in the identification of 5 additional articles, 2 of which reported information on patient-relevant benefits [52,53]. A complementary search of institutional websites led to the inclusion of 5 additional articles, 4 of which contained relevant information on patient-relevant benefits [54-57]. In total, 30 articles were included in the scoping review about patient-relevant benefits. The full-text screening process, including a detailed account of the reasons for exclusion, is presented in Multimedia Appendix 8.

Figure 1. Flow diagram illustrating the selection process of evidence. CDSS: clinical decision support system.

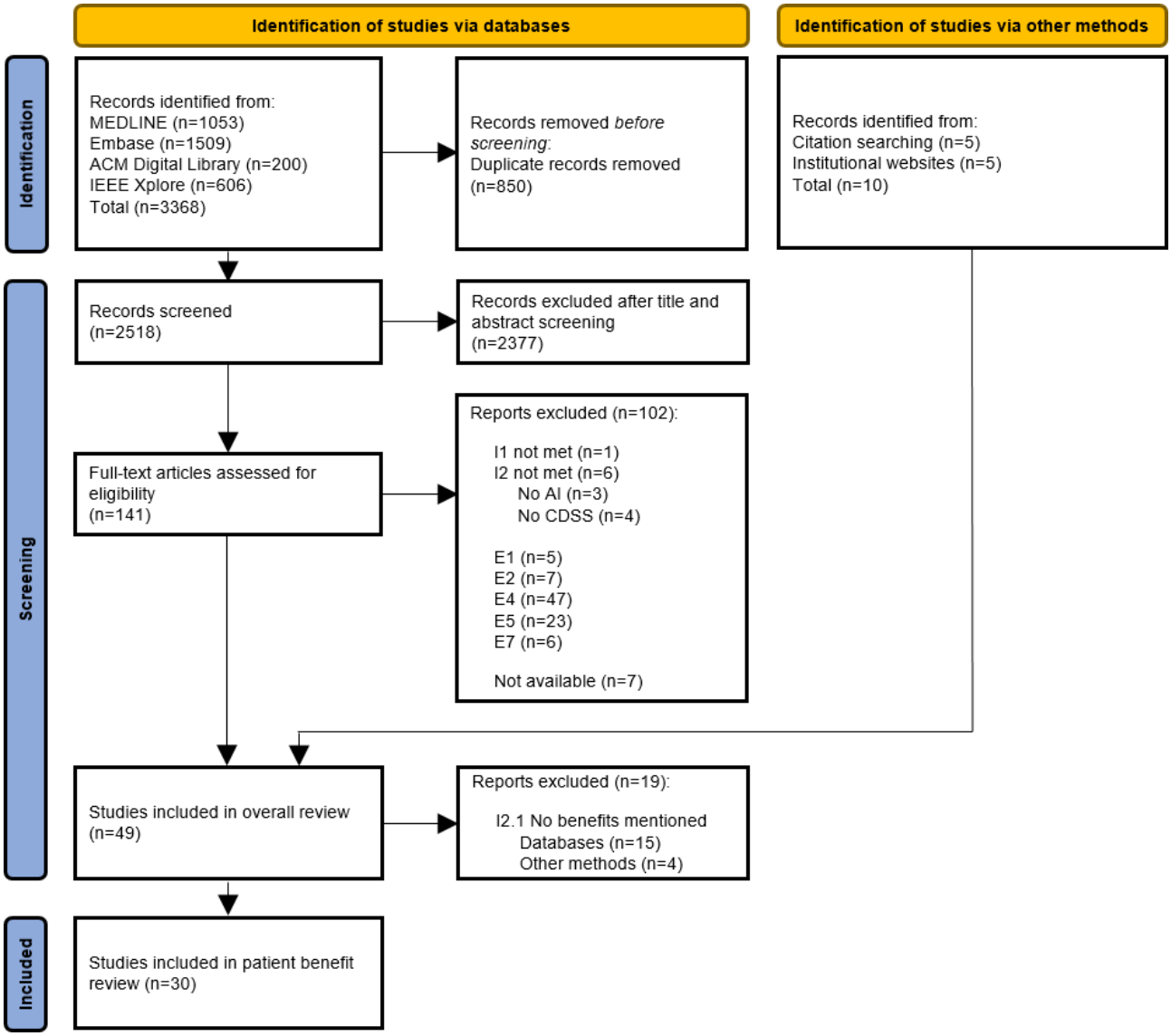

Included Studies

Of the 30 articles included, 16 originated from North America, all of which are from the United States (53.3%) [28,30-32,34,36-38,42,43,45,49,52,53,56,57]. Seven articles stem from Europe (23.3%); 3 from the Netherlands (10%) [46,48,50], 2 from Spain (6.6%) [33,44], 1 from Austria (3.3%) [29], and 1 from the United Kingdom (3.3%) [39]. Five articles are from Asia (16.7%), including 2 each from China [41,51] and Taiwan [40,54] (6.6% each), and 1 from Singapore (3.3%) [35]. There is 1 article from Australia (3.3%) [55] and 1 article from South America (Brazil) (3.3%) [47].

The study designs used in the included articles cover a wide range. Overall, 20 quantitative articles were identified. Of these, 7 used a prospective study design, of which 2 are multicenter studies [28,31] and 5 are single-center studies [34,43,49,52,54]. Thirteen of the quantitative studies used a retrospective approach, comprising 6 research database studies [29,30,39,40,46,53] and 7 electronic health record database studies [33,35,41,42,45,47,50]. In addition to the quantitative

XSL•FO
RenderX

articles, 1 article used a qualitative approach [37] and another applied a mixed methods approach [36]. Additionally, 6 review articles [32,38,44,48,51,55] and 2 articles from news sections of institutional websites were identified [56,57]. All articles are listed in Multimedia Appendix 6.

Synthesis of Results

In total, 6 main categories of patient-relevant benefit were identified. The 6 main categories identified reflect the patient pathway from pretreatment to posttreatment period. They include (1) prediction, (2) earlier treatment and prioritization of high-risk patients, (3) individualized therapy (which encompasses patient-centered care), (4) improved patient outcomes (which includes improved SOFA score, reduced length of stay, and reduced mortality), (5) general improvements in care, and (6) reduced readmission rate (see Figure 2). Multimedia Appendix 7 gives a detailed overview of the benefit categories addressed in each study.

Figure 2. Patient benefit categories related to artificial intelligence-based clinical decision support systems.

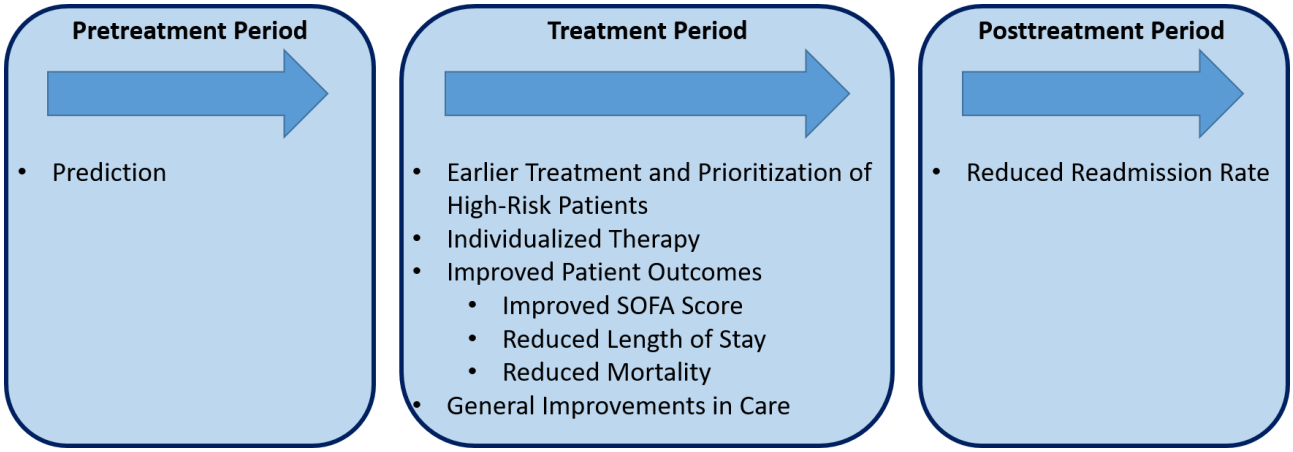

Pretreatment Period (Prediction)

Prediction of sepsis, septic shock, or sepsis-related organ dysfunction was addressed in 12 articles [31,33,35,37,38,40,42,45,48,51,53,56], comprising 7 quantitative studies (1 prospective [31] and 6 retrospective [33,35,40,42,45,53]), 1 qualitative study [37], 3 reviews [38,48,51], and 1 institutional news report [56]. The MLAs identified in this review indicate predictive capacity [48], which may be further optimized through algorithm fine-tuning [45]. Findings suggest that these models may predict sepsis between 4 and 48 hours prior to its onset [35,38,40,51,56], even before significant changes in vital or laboratory parameters become apparent [40]. MLAs were reported to support the identification of appropriate preventive measures [33]. Such predictions may have the potential to improve patient outcomes by providing timely warning of sepsis onset [31,37]. Beyond sepsis, the studies also reported the prediction of septic shock, with MLA predictions occurring between 4 and 7 hours before the onset of septic shock [42,51]. Compared with traditional rule-based routine screening protocols, predictive MLAs demonstrated superior early warning performance, identifying 58.6% more patients before organ dysfunction [53] and potentially contributing to a reduction in septic shock incidence [38]. Moreover, MLAs were shown to predict sepsis-related organ dysfunction approximately 7.5 hours earlier than rule-based routine screening protocols [53].

Treatment Period

Earlier Treatment and Prioritization of High-Risk Patients

Earlier treatment facilitated by MLAs was reported in 10 articles [28,31,33,36,37,40,49,52,53,57], including 7 quantitative studies (4 prospective [28,31,49,52] and 3 retrospective [33,40,53]), 1 mixed methods study [36], 1 qualitative study [37], and 1 institutional news report [57]. One retrospective study reported that the used MLA enabled earlier treatment up to 40 hours before the onset of sepsis [40], while another indicated that the use of ML may reduce the time to treatment, not providing a specific time reduction [33]. Earlier treatment was reported to enable intervention before or during clinical deterioration [53] and potentially prevent sepsis progression [52]. It may allow for early identification and intervention of patients at high risk for severe sepsis prior to clinical onset [31]. Additionally, the literature highlighted early identification and control of the pathogen causing sepsis [52]. Detecting patients before the onset of septic shock may facilitate earlier clinical assessment, diagnostic testing, therapeutic interventions, and transfer to appropriate levels of care [53]. Ultimately, earlier treatment may improve patient outcomes [37] and lead to an alteration in the prevalence of septic shock through timely intervention by health care providers from 5.3% in the control group to 1.5% in the experimental group (–71.7%) of the corresponding study [49]. MLAs were also associated with shorter times to obtain blood cultures (0.98 - 2.79 hours) [49,57], fluid administration (1.05 hours) [57] and earlier administration or adjustment of antibiotics (0.55 - 2.76) [28,36,49,57]. A positive correlation between timely evaluation of MLA alerts and quicker administration of antibiotics was reported, as earlier evaluation of alerts leads to faster use of antibiotics [36]. Finally, a quantitative prospective study reported that MLAs allow prioritization of high-risk patients, with the targeted real-time early warning system (TREWS) identifying who is most likely to benefit from timely treatment [28].

XSL•FO
RenderX

Individualized Therapy

MLAs were reported to support individualized, patient-centered therapy in 11 articles [29,30,33,39,41,44-47,50,54], comprising 10 quantitative studies (1 prospective [54] and 9 retrospective [29,30,33,39,41,45-47,50]) and 1 review [44]. Four main approaches for individualization were identified: (1) subgroup analyses and clustering, (2) optimized substance administration, (3) personalized nursing care, and (4) general statements. Three articles reported subgroup analyses and clustering of patients [41,45,46], which may enable hospitals to provide targeted treatments tailored to the specific needs of defined subgroups [45] and classify patients according to their diverging mortality risk due to factors such as fluid overload or norepinephrine overdose. Such classification might support the development of tailored resuscitation strategies for patients with septic shock [41]. Furthermore, subgroup analyses applied to populations with differing disease severity and progression allow MLAs to adjust the intensity of therapy [46]. The use of MLAs for optimal substance administration was reported in 6 articles [29,33,39,41,50,54]. Applications include personalized antibiotic dosing [33], faster adjustment to the most effective antibiotics, and drug resistance prediction. The comprehensive Intelligent Antimicrobial System demonstrated potentially faster drug resistance prediction times compared with conventional methods, requiring 39.8 hours for carbapenem-resistant Klebsiella pneumonia and 40.9 hours for methicillin-resistant Staphylococcus aureus, compared with 99.5 and 106.4 hours, respectively, using traditional methods [54]. MLA use was also associated with an 8% reduction in antibiotic resistance [50], may shorten the time to antimicrobial resistance detection by 37 hours [54], and was reported to reduce the duration of antibiotic treatment [50]. Furthermore, MLAs may support physicians in selecting appropriate antibiotic therapy [54], with AI-based antibiotic stewardship linked to decreased *Clostridium difficile* infections [50]. Beyond antibiotics, MLAs have demonstrated utility in optimizing dosing strategies for norepinephrine [41], vasopressors [39], corticosteroids [29], and fluid volume management [41]. MLAs also reported to enhance nursing competence and support more evidence-based, personalized nursing care [47]. General statements on individualized therapy were identified in 4 articles [30,33,44,54], including personalized treatment to support physicians in diagnosing and managing bacteremia [33], facilitation of shared decision-making through preoperative discussions [30], improved physician adherence [44], and more precise treatment tailored to individual patients [54].

Improved Patient Outcomes

Improved SOFA Score

Improved SOFA scores associated with the application of and timely response to MLAs were reported in 1 quantitative prospective study [28]. The SOFA score, the predominant measure for assessing the severity of organ dysfunction, is closely linked to the probability of mortality, with a higher score indicating an increased probability of death [3,58]. Using the TREWS algorithm, Adams et al [28] reported a SOFA score progression of –0.8 in their intervention group, compared to –0.4 in the control group. The article highlights a disproportionate reduction in the SOFA score for high-risk patients compared to nonhigh-risk patients. Additionally, timely evaluation and confirmation of the TREWS alerts is associated with improvements in SOFA score progression.

Reduced Length of Stay

Seven quantitative articles reported reductions in LOS [28,31,43,49,50,52,54], including 6 prospective [28,31,43,49,52,54] and 1 retrospective study [50]. Based on the identified literature, a distinction can be drawn between (1) specific reductions, reported in absolute or relative terms [28,31,43,49,52,54], and (2) general statements without precise quantification [28,49,50]. Reported specific reductions in hospital LOS ranged from 0.43 to 8.1 days [28,31,43,49,52], corresponding to decreases of 12.84%-45.25% [31,43,49,52]. Reported reductions in ICU LOS varied between 2.09 and 10.5 days [49,50]. One study also highlighted that shorter ICU stays may contribute to an overall reduction in hospital LOS, although LOS on the general ward increased by 2.4 days [50]. Another article reported a potential annual reduction of 1100 days in emergency department stays and the prevention of 34 ICU stays associated with MLA usage in the examined hospital [54]. General statements indicated a disproportionate, though not statistically significant, reduction in LOS among high-risk patients as well as reduced LOS when MLA-generated alarms were evaluated and confirmed timely [28]. AI-based antibiotic stewardship was also associated with shorter LOS [50] and MLAs were reported to significantly shorten hospital LOS compared to rule-based systems [49]. Furthermore, 1 study suggested that timely physician responses to MLA-generated alerts may contribute to reduced LOS [28].

Reduced Mortality

A reduction in mortality was reported in 14 articles [28,29,31-33,39,43,48,49,52,54-57], including 9 quantitative studies (6 prospective [28,31,43,49,52,54] and 3 retrospective [29,33,39]) as well as 3 reviews [32,48,55] and 2 institutional news reports [56,57]. Reported mortality reductions varied in type and presentation, encompassing (1) specific quantitative statements, expressed in relative or absolute terms [28,29,31,32,43,49,52,54,56,57], and (2) general statements without numerical specifications [28,29,33,39,48,49,54,55]. Relative reductions of mortality ranged from 13.19% to 74.94% [28,31,43,49,52,56,57], whereas absolute reductions ranged from 1.33% points to 26.4% points [28,29,31,43,49,52,54]. One study reported an increase in absolute survival rate of 11.7% and 23.7%, depending on the type of bacteria responsible for the sepsis [54]. Two articles provided reductions in natural numbers; one projected 22 potentially preventable annual deaths in the emergency department of the China Medical University Hospital [54], while another estimated several thousand preventable deaths in the United States alone [32]. General statements suggested that MLAs may disproportionately reduce mortality among high-risk patient cohorts, particularly when outputs are promptly evaluated and confirmed by physicians [28]. Improved survival rates may also be linked to the use of MLA-guided antibiotic recommendations [54] and the application of the 3PM (predictive, preventive, and personalized medicine) principles [33]. MLAs are associated with lower mortality compared to traditional physician assessments [29,39]

XSL•FO
RenderX

and predictions generated by rule-based tools [49]. Additionally, literature provided general statements, offering limited informational depth and indicating that the use of ML may contribute to reduced mortality [29,39,48,49,55].

General Improvements in Care

Eight articles reported improvements in care associated with MLAs [31,34,38,43,47-49,54], including 6 quantitative studies (5 prospective [31,34,43,49,54] and 1 retrospective [47]) as well as 2 reviews [38,48]. Reported benefits can be divided into 2 domains: (1) statements related to time and (2) statements on patient care enhancements. One study reported a reduced duration of septic shock [48]. Within the patient care enhancement category, MLAs were described as posing no risk to patients and offering potential benefits to patients and health care providers [31], reducing events of clinical deterioration [38], improving care accuracy [47,54], and increasing sepsis awareness among physicians [43,49]. Physicians and nurses also reported perceived improvements in care [34].

Posttreatment Period (Reduced Readmission Rate)

Predictive AI-based CDSS were associated with reduced 30-day readmission rates, as reported in 2 quantitative prospective studies [31,43]. In Burdick et al [31], implementation of an MLA reduced the 30-day readmissions from 36.4% to 28.12%, representing a 22.74% reduction compared to the baseline period. McCoy and Das [43] reported a decline from 46.19% (188/407) during the preimplementation baseline period to 29.8% (100/336) in a first postimplementation period and further to 25.2% (96/381) in a second postimplementation period. In a subsequent steady-state period, the 30-day readmission rate was further reduced to 7.84% (16/204). Across all surveyed months after implementation, the 30-day readmission rate was 23.03%, representing a 50.14% reduction in the sepsis-related 30-day readmission rate.

Discussion

Principal Findings

This scoping review presents the evidence on the patient-relevant benefits of AI-based CDSS in sepsis care. All articles focusing on sepsis and presenting the influence of AI-based CDSS on patient-relevant benefits, identified through the comprehensive search strategy, were included. In total, 30 articles were identified and integrated into the review. Investigating the literature, there is a number of AI-based CDSS for sepsis treatment developed in the past or currently under development. However, research typically has no or only limited reference to patient-relevant benefits and (1) mostly focuses on problems and/or success strategies [32,55,59-61] and/or (2) is indication-independent [59]. To the best of the authors' knowledge, this represents the first scoping review on this specific topic.

The findings of this scoping review, systematized into the 6 main categories, (1) prediction, (2) earlier treatment and prioritization of high-risk patients, (3) individualized therapy (which encompasses patient-centered care), (4) improved patient outcomes (which includes improved SOFA score, reduced length of stay, and reduced mortality), (5) general improvements in care, and (6) reduced readmission rate, underscore the potential patient-relevant benefits of AI-based CDSS in sepsis care across the entire inpatient pathway. The literature indicates that MLAs can potentially predict sepsis before its clinical onset [35,38,40,51,56]. Additionally, septic shock [42,51] and sepsis-related organ dysfunction [53] may be predicted in advance. These predictive capabilities can contribute to reducing the incidence of septic shock [38] and supporting decreased mortality rates among sepsis patients [12]. Sepsis prediction may facilitate timely treatment initiation through the use of MLAs [40]. This was associated with improved patient outcomes and a decreased prevalence of septic shock [49]. Furthermore, individualized therapy can potentially have a positive impact on patient-relevant benefits by reducing the time to treatment or LOS for each individual patient [33]. Moreover, the disproportionate reduction in the SOFA score through the use of ML compared to a control group whose treatment was not supported by MLAs should be mentioned. According to the Sepsis-3 definition, the level of the SOFA score positively correlates with the probability of death [3], and a SOFA score of ≥2 points corresponds to a mortality risk of over 10% in hospitalized patients outside the ICU [25]. The TREWS algorithm presented by Adams et al was able to reduce the SOFA score by 0.8 points, while a reduction of only 0.4 points was observed in the control group. Accordingly, the use of this MLA may contribute to the reduction in mortality. In general, the usage of MLAs was associated with a mortality reduction of up to 74.94% [28,31,43,49,52,56,57], with faster response times being associated with greater reductions in mortality [28]. This demonstrates the medical potential of ML in the treatment of sepsis, particularly when clinical recommendations are accepted and promptly implemented by physicians. With approximately 11 million deaths annually from sepsis according to the WHO [2], a corresponding reduction in mortality could translate into a substantial global health impact. Additionally, MLAs were linked to reduced hospital LOS [31,43,49,52] and ICU LOS [49,50]. Beyond their predictive capabilities, facilitation of timely treatment, mortality, and LOS reductions, AI-based CDSS in sepsis care provide further patient benefits, including shortened duration of septic shock [48], reduced antibiotic resistance, and reduced duration of antibiotic treatment [50]. MLAs may also contribute to a reduction of events of clinical deterioration [38] and increased physician awareness of sepsis [43,49]. Finally, the literature indicates that AI-based CDSS in sepsis care can contribute to reducing hospital readmission rates [43], further demonstrating their potential to improve patient-relevant benefits.

Comparison With Prior Work

While 6 reviews were included in this work, they primarily focused on other topics and predominantly used less systematic approaches [32,38,44,48,51,55]. Among the included reviews, 4 adopted a narrative review methodology [32,44,48,51]. By design, this approach is inherently less systematic than systematic reviews or scoping reviews, and this was evident in the search and selection process of the included narrative reviews. Two relied exclusively on limited, nonsystematic keyword searches, one using 8 keywords across 4 search engines [44] and another restricted to 3 keywords in a single database

XSL•FO
RenderX

[32]. Moreover, the review conducted by Ferreira et al [32] focused primarily on problems and success strategies related to AI-based CDSS, thereby addressing a different thematic focus than the present scoping review. Another narrative review applied a brief and partial search string without predefined inclusion and exclusion criteria and was limited to a single database [51], representing considerable methodological limitations relative to the present comprehensive scoping review. The narrative review conducted by Schinkel et al [48] adopted a more systematic approach, using a predefined search string and assessing the clinical value of AI-based systems by evaluating the AUROC as a criterion for article selection. While methodologically more robust, this review nonetheless differed from the present article, as it primarily evaluated the advantage of MLAs over rule-based scores, reflecting the status quo using a binary classifier. The advantage of MLAs over rule-based scores was considered a prerequisite for AI-based CDSS in the present study. With the exception of 1 review, where a manual search of reference lists was conducted [32], none of the narrative reviews [44,48,51] undertook a comprehensive search for gray literature or an analysis of the reference lists. Furthermore, only 1 narrative review reported a screening process conducted by 2 independent reviewers [48], whereas the other 3 reviews did not provide methodological detail [32,44,51]. By contrast, the present scoping review implemented a rigorous screening process with 2 independent reviewers to enhance objectivity, reliability, and reproducibility. Beyond these narrative reviews, 1 study followed an integrative review approach, explicitly focusing on predictive algorithms and embedding this narrow focus within a brief predefined search string [38]. In contrast, the present exploratory scoping review aimed to inductively derive patient benefit categories associated with AI-based CDSS in sepsis care. This integrative review relied on a single reviewer for screening [38], representing a methodological limitation in comparison with the dual-reviewer approach of the present scoping review. Finally, 1 systematic review included in this study used a largely rigorous and systematic methodology, with the notable exception of a gray literature search, which was not reported. In addition, this systematic review focused primarily on problems and success strategies [55], thereby diverging from the present scoping reviews' explicit focus on patient-relevant benefits. In sum, the present scoping review can be clearly distinguished from the included reviews both methodologically and thematically. By applying a comprehensive, exploratory design, centered on patient-relevant benefits, it makes a substantive and valuable contribution to closing the research gap regarding patient-relevant benefits of AI-based CDSS in sepsis care.

Implications and Recommendations

Patient-relevant benefits identified in the literature are not sufficient to ensure successful implementation of AI-based CDSS. Equally critical is the acceptance of the underlying technology by health care providers and their belief that its use possesses tangible benefits. The unified theory of acceptance and use of technology (UTAUT) provides a framework to understand factors influencing behavioral intention and use behavior using four constructs: (1) performance expectancy, (2) effort expectancy, (3) social influence, and (4) facilitating conditions. In this context, effective design of AI-based CDSS should ensure that providers perceive the system as both beneficial and easy to use, corresponding to the first 2 constructs of the UTAUT. Specifically, (1) users should believe that using AI-based CDSS enhances gains in job performance, and (2) the system is intuitive and easy to operate. Equally important are contextual factors: health care providers should perceive that (3) important others endorse system use, and (4) organizational and technical infrastructure is in place to support usage [61]. A meta-analysis by Dingel et al [62] applying the UTAUT to health care practitioners' intention to use AI-enabled CDSS confirms that implementation must address not only technical and organizational aspects but also psychological and social factors, particularly fostering user trust. Successful implementation of AI-based CDSS therefore depends only partly on system performance; it is largely contingent on user attitudes and framework conditions.

Beyond the 4 UTAUT constructs, specific barriers [63] and facilitators [64] must be considered when evaluating AI-based CDSS. A nuanced understanding of these factors is essential to accurately evaluate the potential impact of AI-based CDSS on sepsis care. The current evidence demonstrates a pronounced lack of prospective studies investigating the optimal integration of such systems [29]. This paucity of implementation-oriented research, coupled with limited clinician acceptance [37] and insufficient knowledge of AI among health care providers [65], constitutes a substantial barrier to clinical adoption. Concurrently, extant literature highlights pivotal facilitators, emphasizing the importance of prioritizing research on effective integration strategies [38]. For instance, low acceptance may be mitigated by involving health care providers directly in the design and development of CDSS [66-68], while targeted training and educational programs could address knowledge gaps among service providers and enhance trust in this technology [37,67]. These factors must therefore be carefully considered by all stakeholders involved in implementation (eg, caregivers, physicians, and researchers) before real-world adoption can occur. For clinicians, the findings provide insights into realistic benefits, current limitations, and evidence gaps that may guide expectations in clinical decision-making. For researchers, this review underscores the importance of conducting prospective studies and fostering user-centered development to ensure that CDSS effectively translate into clinical practice. In addition, although patient-relevant benefits—and not only measurable patient-relevant outcomes—have been investigated, the findings may contribute to the development of a consistent set of generic patient-relevant outcomes, as proposed by Kersting et al [69]. This could, in turn, facilitate a shared understanding and enhance comparability across studies targeting patient-relevant outcomes, particularly given the absence of a clear, widely accepted definition and standardized criteria for selecting such outcomes.

Strengths and Limitations

This scoping review was conducted by an interdisciplinary team comprising computer scientists, physicians, statisticians, and (health) economists. This diverse expertise facilitated a comprehensive examination of all relevant aspects across these fields, ensuring a thorough evaluation of the reviewed literature.

XSL•FO
RenderX

To address the interdisciplinary research question comprehensively, 4 databases focusing on medicine and informatics were included in the review. In addition, the structured search for gray literature targeted various institutions in 8 countries, including each country's Ministry of Health, diverse sepsis and intensive care associations for each country, and diverse health informatics associations for each country. The 8 countries were selected based on two criteria: (1) having highly digitalized health care systems and/or (2) holding the status of industrialized nations. These countries were presumed to have a higher likelihood of using AI-based systems. Furthermore, the search encompassed internationally active stakeholders, such as the WHO and OECD, alongside globally active health informatics organizations and sepsis and critical care associations.

Despite all efforts, this scoping review is not free of limitations. Given the exploratory nature of the methodology, publication bias must be considered a potential limitation [70]. Studies in which AI-based CDSS do not demonstrate improvements in patient-relevant benefits compared with conventional scores may not be submitted in peer-reviewed journals, potentially leading to an overestimation of their true patient benefit. Although no formal risk of bias assessment was conducted, the included studies demonstrated considerable heterogeneity in design. Moreover, 65% (13/20) of the quantitative studies relied solely on retrospective methodologies, in which evidence of patient benefits was demonstrated only theoretically. Consequently, the findings of this scoping review should be interpreted with caution, as the reported effects may be overestimated in the context of real-world care. The overall strength of evidence was limited by the predominance of retrospective study designs and the theoretical nature of reported benefits. In contrast, the included prospective studies provided more robust support for the identified benefit categories. Importantly, each benefit category has been substantiated in prospective studies, thereby affirming its validity in real-world clinical contexts rather than solely theoretically in retrospective or descriptive studies (Multimedia Appendix 9). Detailed information on the study designs of all included articles is provided in Multimedia Appendix 6. Multimedia Appendix 9 summarizes benefit categories identified across the respective study designs. Furthermore, the comparability of the reported MLA performance across articles is limited due to varying definitions of sepsis (eg, different causative pathogens, divergent sepsis definitions, and variations in the examined indications such as sepsis, septic shock, or sepsis-related organ dysfunction). The same limitation applies to the databases used for training and validation, which differed substantially in size. In addition, no assessment of the applied MLA methods was conducted, nor was the level of maturity of the individual MLAs explicitly considered. Furthermore, due to the heterogeneity of the included studies, no formal quality assessment was conducted. Rather, the present review was designed to exploratively map and comparatively present the entirety of available evidence in order to identify research gaps, without imposing methodological restrictions on the literature to be included [20-22]. Finally, a methodological limitation should be noted: Research conducted on institutional websites could only be partially conducted for Estonia, Denmark, and Spain due to language restrictions (English and German), as some stakeholder websites were available exclusively in the respective national languages. The utilization of translation tools was deliberately avoided, as the inclusion of material that none of the authors could fully comprehend and critically appraise in the original language was considered methodologically inappropriate.

The search restriction of 15 years should not be considered a limitation. The inclusion period was defined in consultation with ML experts (NT, HN), and algorithms developed prior to the review period (January 1, 2008-March 2, 2023) were predominantly anticipated to be (1) rule-based systems and/or nonblack-box systems for the users. Both types of algorithms are outside the scope of this review. Moreover, an initial limited search in the databases MEDLINE via PubMed and Embase, which accounted for approximately 75% of the screened literature (Figure 1), indicated that only a marginal proportion of articles relevant to the research question were published before 2008. Consequently, the time restriction is unlikely to have affected the identification of relevant literature.

Conclusion

The findings of this scoping review highlight the considerable medical relevance of AI-based CDSS in sepsis care. These systems offer benefits across the entire patient care pathway, from early detection and risk stratification to individualized therapy and various improved outcomes. AI-based CDSS has shown the ability to predict sepsis, septic shock, and sepsis-related organ dysfunction, enabling earlier initiation of treatment, prioritization of high-risk patients, and tailored therapeutic strategies. In addition to supporting earlier and more targeted interventions, AI-based CDSS contribute to better clinical outcomes, including improved SOFA scores, reduced LOS both in general wards and ICUs, and lower mortality rates. They may also help reduce readmission rates among sepsis patients, further enhancing long-term care quality. With their transformative potential, AI-based CDSS could fundamentally improve the global management of sepsis. However, further research is needed to optimize the development, implementation, and clinical application of these systems to maximize patient benefits and further improve outcomes for sepsis patients in the future. This is particularly important given the highly heterogeneous evidence base, with a substantial proportion of studies relying on retrospective data, as the results of the included studies cannot be directly generalized or applied without caution.

Acknowledgments

The authors acknowledge the support of the Open Access Publication Fund of the University of Duisburg-Essen and thank Mrs Katrin Wibker for her assistance in quality control during the development of the search string for the scoping review.

XSL•FO
RenderX

Funding

This research is part of a wider research project (Multimethod Research Project on Requirements for Artificial Intelligence-Based Clinical Decision Support Systems Using the Medical Example of Sepsis (="“KI@work"")) funded by the German Federal Joint Committee (funding code: 01VSF22050 –KI@work). The funders had no influence in the study design, conduct of the study, or the decision to publish or prepare the manuscript.

Authors' Contributions

Conceptualization: PR (lead), GDG (equal), NB (equal), MA (supporting), HN (supporting), LP (supporting), PH (supporting), NT (supporting), MT (supporting), FMB (supporting), JW (supporting)
Data curation: PR (lead), GDG (equal)
Formal analysis: PR (lead), GDG (supporting), NB (supporting)
Funding acquisition: GDG, JW, MA, HN, NT, MT, FMB, NB
Investigation: PR (lead), GDG (equal)
Methodology: PR (lead), GDG (supporting), NB (supporting), NT (supporting), HN (supporting)
Project administration: NB
Supervision: NB
Validation: PR (lead), GDG (equal), NB (equal)
Visualization: PR
Writing – original draft: PR
Writing – review & editing: PR (lead), GDG (supporting), NB (supporting), MA (supporting), HN (supporting), LP (supporting), PH (supporting), NT (supporting), MT (supporting), FMB (supporting), JW (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy – Medline via PubMed.
[DOCX File, 14 KB - jmir_v28i1e76772_app1.docx]

Multimedia Appendix 2
Search strategy – Embase.
[DOCX File, 14 KB - jmir_v28i1e76772_app2.docx]

Multimedia Appendix 3
Search strategy – ACM Digital Library.
[DOCX File, 14 KB - jmir_v28i1e76772_app3.docx]

Multimedia Appendix 4
Search strategy – IEEE Xplore.
[DOCX File, 15 KB - jmir_v28i1e76772_app4.docx]

Multimedia Appendix 5
Included websites in structured search.
[DOCX File, 20 KB - jmir_v28i1e76772_app5.docx]

Multimedia Appendix 6
Overview of included articles.
[DOCX File, 39 KB - jmir_v28i1e76772_app6.docx]

Multimedia Appendix 7
Patient benefits mentioned in articles.
[DOCX File, 15 KB - jmir_v28i1e76772_app7.docx]

Multimedia Appendix 8
Screened articles.
[DOCX File, 34 KB - jmir_v28i1e76772_app8.docx]

XSL•FO
RenderX

Multimedia Appendix 9
Used methods per benefit category.
[DOCX File, 13 KB - jmir_v28i1e76772_app9.docx]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 108 KB - jmir_v28i1e76772_app10.docx]

References

1. Rhee C, Jones TM, Hamad Y, et al. Prevalence, underlying causes, and preventability of sepsis-associated mortality in US acute care hospitals. JAMA Netw Open 2019 Feb 1;2(2):e187571. [doi: 10.1001/jamanetworkopen.2018.7571] [Medline: 30768188]
2. Global report on the epidemiology and burden of sepsis current evidence, identifying gaps and future directions. World Health Organization. 2020. URL: https://iris.who.int/server/api/core/bitstreams/d4ce3613-bf94-4205-85c8-14f3fc0609db/content [accessed 2025-12-04]
3. Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 2016 Feb 23;315(8):801-810. [doi: 10.1001/jama.2016.0287] [Medline: 26903338]
4. Ericson O, Hjelmgren J, Sjövall F, Söderberg J, Persson I. The potential cost and cost-effectiveness impact of using a machine learning algorithm for early detection of sepsis in intensive care units in Sweden. J Health Econ Outcomes Res 2022;9(1):101-110. [doi: 10.36469/jheor.2022.33951] [Medline: 35620451]
5. Arriaga-Pizano LA, Gonzalez-Olvera MA, Ferat-Osorio EA, et al. Accurate diagnosis of sepsis using a neural network: Pilot study using routine clinical variables. Comput Methods Programs Biomed 2021 Oct;210:106366. [doi: 10.1016/j.cmpb.2021.106366] [Medline: 34500141]
6. Vincent JL, Martin GS, Levy MM. qSOFA does not replace SIRS in the definition of sepsis. Crit Care 2016 Jul 17;20(1):210. [doi: 10.1186/s13054-016-1389-z] [Medline: 27423462]
7. Dellinger RP, Levy MM, Rhodes A, et al. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012. Crit Care Med 2013 Feb;41(2):580-637. [doi: 10.1097/CCM.0b013e31827e83af] [Medline: 23353941]
8. Kumar A, Roberts D, Wood KE, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Crit Care Med 2006 Jun;34(6):1589-1596. [doi: 10.1097/01.CCM.0000217961.75225.E9] [Medline: 16625125]
9. Brunkhorst FM, Weigand MA, Pletz M, et al. S3 Guideline Sepsis-prevention, diagnosis, therapy, and aftercare: long version. Med Klin Intensivmed Notfmed 2020 May;115(Suppl 2):37-109. [doi: 10.1007/s00063-020-00685-0] [Medline: 32356041]
10. Sakr Y, Jaschinski U, Wittebole X, et al. Sepsis in intensive care unit patients: worldwide data from the intensive care over nations audit. Open Forum Infect Dis 2018 Dec;5(12):ofy313. [doi: 10.1093/ofid/ofy313] [Medline: 30555852]
11. Arwyn-Jones J, Brent AJ. Sepsis. Surgery (Oxford) 2019 Jan;37(1):1-8. [doi: 10.1016/j.mpsur.2018.11.007]
12. Boussina A, Shashikumar SP, Malhotra A, et al. Impact of a deep learning sepsis prediction model on quality of care and survival. NPJ Digit Med 2024 Jan 23;7(1):14. [doi: 10.1038/s41746-023-00986-6] [Medline: 38263386]
13. Alanazi A, Aldakhil L, Aldhoayan M, Aldosari B. Machine learning for early prediction of sepsis in intensive care unit (ICU) patients. Med Bogota Colomb 2023 Jul 9;59(7):1276. [doi: 10.3390/medicina59071276]
14. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med 2018 Apr;46(4):547-553. [doi: 10.1097/CCM.0000000000002936] [Medline: 29286945]
15. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. US Food and Drug Administration. 2025. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices [accessed 2025-04-29]
16. O'Sullivan C, Tsai DHT, Wu ICY, et al. Machine learning applications on neonatal sepsis treatment: a scoping review. BMC Infect Dis 2023 Jun 29;23(1):441. [doi: 10.1186/s12879-023-08409-3] [Medline: 37386442]
17. Heneghan JA, Walker SB, Fawcett A, et al. The pediatric data science and analytics subgroup of the pediatric acute lung injury and sepsis investigators network: use of supervised machine learning applications in pediatric critical care medicine research. Pediatr Crit Care Med 2024 Apr 1;25(4):364-374. [doi: 10.1097/PCC.0000000000003425] [Medline: 38059732]
18. Susanto AP, Lyell D, Widyantoro B, Berkovsky S, Magrabi F. Effects of machine learning-based clinical decision support systems on decision-making, care delivery, and patient outcomes: a scoping review. J Am Med Inform Assoc 2023 Nov 17;30(12):2050-2063. [doi: 10.1093/jamia/ocad180] [Medline: 37647865]
19. Wan YKJ, Wright MC, McFarland MM, et al. Information displays for automated surveillance algorithms of in-hospital patient deterioration: a scoping review. J Am Med Inform Assoc 2023 Dec 22;31(1):256-273. [doi: 10.1093/jamia/ocad203] [Medline: 37847664]

XSL•FO
RenderX

20. Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. JBI Manual for Evidence Synthesis JBI 2024. [doi: 10.46658/JBIMES-24-01]
21. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
22. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5(5):69. [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
23. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
24. Raszke P, Giebel GD, Abels C, et al. User-oriented requirements for artificial intelligence-based clinical decision support systems in sepsis: protocol for a multimethod research project. JMIR Res Protoc 2025 Jan 30;14:e62704. [doi: 10.2196/62704] [Medline: 39883929]
25. S3-leitlinie: sepsis – prävention, diagnose, therapie und nachsorge – update 2025 [Website in German]. AWMF-register der leitlinien der wissenschaftlichen medizinischen fachgesellschaften. 2025. URL: https://register.awmf.org/assets/guidelines/079-001l_S3_Sepsis-Praevention-Diagnose-Therapie-Nachsorge_2025-07.pdf [accessed 2025-12-12]
26. S2k-leitlinie 024-025 „sepsis bei kindern jenseits der neonatalperiode [Website in German]. AWMF-Register der Leitlinien der wissenschaftlichen medizinischen Fachgesellschaften. 2015. URL: https://register.awmf.org/assets/guidelines/024-025l_S2k_Sepsis_nach_Neonatalperiode_2016-04-abgelaufen.pdf [accessed 2025-01-20]
27. S2k-leitlinie 024-008: bakterielle infektionen bei neugeborenen [Website in German]. AWMF-Register der Leitlinien der wissenschaftlichen medizinischen Fachgesellschaften. 2021. URL: https://register.awmf.org/assets/guidelines/024-008l_S2k_Bakterielle_Infektionen_Neugeborene_2021-03-abgelaufne.pdf [accessed 2025-01-20]
28. Adams R, Henry KE, Sridharan A, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med 2022 Jul;28(7):1455-1460. [doi: 10.1038/s41591-022-01894-0] [Medline: 35864252]
29. Bologheanu R, Kapral L, Laxar D, et al. Development of a reinforcement learning algorithm to optimize corticosteroid therapy in critically ill patients with sepsis. J Clin Med 2023 Feb 14;12(4):1513. [doi: 10.3390/jcm12041513] [Medline: 36836046]
30. Bunn C, Kulshrestha S, Boyda J, et al. Application of machine learning to the prediction of postoperative sepsis after appendectomy. Surgery 2021 Mar;169(3):671-677. [doi: 10.1016/j.surg.2020.07.045] [Medline: 32951903]
31. Burdick H, Pino E, Gabel-Comeau D, et al. Effect of a sepsis prediction algorithm on patient mortality, length of stay and readmission: a prospective multicentre clinical outcomes evaluation of real-world patient data from US hospitals. BMJ Health Care Inform 2020 Apr;27(1):e100109. [doi: 10.1136/bmjhci-2019-100109] [Medline: 32354696]
32. Ferreira LD, McCants D, Velamuri S. Using machine learning for process improvement in sepsis management. J Healthc Qual Res 2023;38(5):304-311. [doi: 10.1016/j.jhqr.2022.09.006] [Medline: 36319584]
33. Garnica O, Gómez D, Ramos V, Hidalgo JI, Ruiz-Giardín JM. Diagnosing hospital bacteraemia in the framework of predictive, preventive and personalised medicine using electronic health records and machine learning classifiers. EPMA J 2021 Sep;12(3):365-381. [doi: 10.1007/s13167-021-00252-3] [Medline: 34484472]
34. Ginestra JC, Giannini HM, Schweickert WD, et al. Clinician perception of a machine learning-based early warning system designed to predict severe sepsis and septic shock. Crit Care Med 2019 Nov;47(11):1477-1484. [doi: 10.1097/CCM.0000000000003803] [Medline: 31135500]
35. Goh KH, Wang L, Yeow AYK, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun 2021 Jan 29;12(1):711. [doi: 10.1038/s41467-021-20910-4] [Medline: 33514699]
36. Henry KE, Adams R, Parent C, et al. Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. Nat Med 2022 Jul;28(7):1447-1454. [doi: 10.1038/s41591-022-01895-z] [Medline: 35864251]
37. Joshi M, Mecklai K, Rozenblum R, Samal L. Implementation approaches and barriers for rule-based and machine learning-based sepsis risk prediction tools: a qualitative study. JAMIA Open 2022 Jul;5(2):ooac022. [doi: 10.1093/jamiaopen/ooac022] [Medline: 35474719]
38. Kausch SL, Moorman JR, Lake DE, Keim-Malpass J. Physiological machine learning models for prediction of sepsis in hospitalized adults: an integrative review. Intensive Crit Care Nurs 2021 Aug;65:103035. [doi: 10.1016/j.iccn.2021.103035] [Medline: 33875337]
39. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 2018 Nov;24(11):1716-1720. [doi: 10.1038/s41591-018-0213-5] [Medline: 30349085]
40. Kuo YY, Huang ST, Chiu HW. Applying artificial neural network for early detection of sepsis with intentionally preserved highly missing real-world data for simulating clinical situation. BMC Med Inform Decis Mak 2021 Oct 22;21(1):290. [doi: 10.1186/s12911-021-01653-0] [Medline: 34686163]
41. Ma P, Liu J, Shen F, et al. Individualized resuscitation strategy for septic shock formalized by finite mixture modeling and dynamic treatment regimen. Crit Care 2021 Jul 12;25(1):243. [doi: 10.1186/s13054-021-03682-7]

XSL•FO
RenderX

42. Mao Q, Jay M, Hoffman JL, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. BMJ Open 2018 Jan 26;8(1):e017833. [doi: 10.1136/bmjopen-2017-017833] [Medline: 29374661]
43. McCoy A, Das R. Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. BMJ Open Qual 2017;6(2):e000158. [doi: 10.1136/bmjoq-2017-000158] [Medline: 29450295]
44. Ocampo-Quintero N, Vidal-Cortés P, del Río Carbajo L, Fdez-Riverola F, Reboiro-Jato M, Glez-Peña D. Enhancing sepsis management through machine learning techniques: a review. Medicina Intensiva (English Edition) 2022 Mar;46(3):140-156. [doi: 10.1016/j.medine.2020.04.015]
45. Rogers P, Boussina AE, Shashikumar SP, Wardi G, Longhurst CA, Nemati S. Optimizing the implementation of clinical predictive models to minimize national costs: sepsis case study. J Med Internet Res 2023 Feb 13;25:e43486. [doi: 10.2196/43486] [Medline: 36780203]
46. Roggeveen L, El Hassouni A, Ahrendt J, et al. Transatlantic transferability of a new reinforcement learning model for optimizing haemodynamic treatment for critically ill patients with sepsis. Artif Intell Med 2021 Feb;112:102003. [doi: 10.1016/j.artmed.2020.102003] [Medline: 33581824]
47. Scherer JDS, Pereira JS, Debastiani MS, Bica CG. Beyond technology: can artificial intelligence support clinical decisions in the prediction of sepsis? Rev Bras Enferm 2022;75(5):e20210586. [doi: 10.1590/0034-7167-2021-0586] [Medline: 35584427]
48. Schinkel M, Paranjape K, Nannan Panday RS, Skyttberg N, Nanayakkara PWB. Clinical applications of artificial intelligence in sepsis: a narrative review. Comput Biol Med 2019 Dec;115:103488. [doi: 10.1016/j.compbiomed.2019.103488] [Medline: 31634699]
49. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ Open Respir Res 2017;4(1):e000234. [doi: 10.1136/bmjresp-2017-000234] [Medline: 29435343]
50. Voermans AM, Mewes JC, Broyles MR, Steuten LMG. Cost-effectiveness analysis of a procalcitonin-guided decision algorithm for antibiotic stewardship using real-world U.S. hospital data. OMICS 2019 Oct;23(10):508-515. [doi: 10.1089/omi.2019.0113] [Medline: 31509068]
51. Wu M, Du X, Gu R, Wei J. Artificial intelligence for clinical decision support in sepsis. Front Med (Lausanne) 2021;8:665464. [doi: 10.3389/fmed.2021.665464] [Medline: 34055839]
52. Burdick H, Pino E, Gabel-Comeau D, et al. Evaluating a sepsis prediction machine learning algorithm in the emergency department and intensive care unit: a before and after comparative study. Clinical Trials. . [doi: 10.1101/224014]
53. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med 2015 Aug 5;7(299):299ra122. [doi: 10.1126/scitranslmed.aab3719] [Medline: 26246167]
54. Hsu KC, Cho DY, Hsueh PR, Yu J, Sun PR. Development of a comprehensive intelligent antimicrobial system: an epochal, fast, and digitally precise prediction of therapeutic antibiotics case study. China Medical University Hospital, Taiwan 2023 [FREE Full text]
55. van der Vegt AH, Scott IA, Dermawan K, Schnetler RJ, Kalke VR, Lane PJ. Deployment of machine learning algorithms to predict sepsis: systematic review and application of the SALIENT clinical AI implementation framework. J Am Med Inform Assoc 2023 Jun 20;30(7):1349-1361. [doi: 10.1093/jamia/ocad075] [Medline: 37172264]
56. Dascena pledges support for sepsis allianc clinical community. Sepsis alliance. 2021. URL: https://www.sepsis.org/news/dascena-pledges-support-for-sepsis-alliance-clinical-community/ [accessed 2025-01-20]
57. Saving Lives With Enhanced Sepsis Treatment Protocol. Healthcare information and management systems society (HIMSS). 2020. URL: https://oregon.himss.org/news/saving-lives-enhanced-sepsis-treatment-protocol [accessed 2025-01-20]
58. Vincent JL, de Mendonca A, Cantraine F, et al. Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units. Crit Care Med 1998 Nov;26(11):1793-1800. [doi: 10.1097/00003246-199811000-00016]
59. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17. [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
60. Moor M, Rieck B, Horn M, Jutzeler CR, Borgwardt K. Early prediction of sepsis in the ICU using machine learning: a systematic review. Front Med (Lausanne) 2021;8:607952. [doi: 10.3389/fmed.2021.607952] [Medline: 34124082]
61. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view1. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
62. Dingel J, Kleine AK, Cecil J, Sigl AL, Lermer E, Gaube S. Predictors of health care practitioners' intention to use AI-enabled clinical decision support systems: meta-analysis based on the unified theory of acceptance and use of technology. J Med Internet Res 2024 Aug 5;26:e57224. [doi: 10.2196/57224] [Medline: 39102675]
63. Giebel GD, Raszke P, Nowak H, et al. Problems and barriers related to the use of AI-based clinical decision support systems: interview study. J Med Internet Res 2025 Feb 3;27:e63377. [doi: 10.2196/63377] [Medline: 39899342]

XSL•FO
RenderX

64. Giebel GD, Raszke P, Nowak H, et al. Improving AI-based clinical decision support systems and their integration into care from the perspective of experts: interview study among different stakeholders. JMIR Med Inform 2025 Jul 7;13:e69688. [doi: 10.2196/69688] [Medline: 40623684]
65. Sandhu S, Lin AL, Brajer N, et al. Integrating a machine learning system into clinical workflows: qualitative study. J Med Internet Res 2020 Nov 19;22(11):e22421. [doi: 10.2196/22421] [Medline: 33211015]
66. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Manage Sci 1989 Aug;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
67. Manetti S, Cumetti M, De Benedictis A, Lettieri E. Adoption of novel biomarker test parameters with machine learning-based algorithms for the early detection of sepsis in hospital practice. J Nurs Manag 2022 Nov;30(8):3754-3764. [doi: 10.1111/jonm.13807] [Medline: 36125938]
68. Sendak M, Elish MC, Gao M, et al. "The human body is a black box": supporting clinical decision-making with deep learning. 2020 Presented at: FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; Jan 27-30, 2020; Barcelona, Spain p. 99-109 URL: https://dl.acm.org/doi/10.1145/3351095.3372827 [accessed 2026-01-02]
69. Kersting C, Kneer M, Barzel A. Patient-relevant outcomes: what are we talking about? A scoping review to improve conceptual clarity. BMC Health Serv Res 2020 Jun 29;20(1):596. [doi: 10.1186/s12913-020-05442-9] [Medline: 32600321]
70. Nair AS. Publication bias - importance of studies with negative results!. Indian J Anaesth 2019 Jun;63(6):505-507. [doi: 10.4103/ija.IJA_142_19] [Medline: 31263309]

Abbreviations

3PM: predictive, preventive, and personalized medicine
AI: artificial intelligence
APACHE II: Acute Physiology and Chronic Health Evaluation II
CDSS: clinical decision support system
ICU: intensive care unit
LOS: length of stay
ML: machine learning
MLA: machine learning algorithm
NEWS2: National Early Warning Score 2
OECD: Organisation for Economic Co-operation and Development
PCC: Population, Concept, Context
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
qSOFA score: Quick Sequential Organ Failure Assessment Score
SIRS: systemic inflammatory response syndrome
SOFA score: Sequential Organ Failure Assessment Score
TREWS: targeted real-time early warning system
UTAUT: unified theory of acceptance and use of technology
WHO: World Health Organization

Edited by J Sarvestan; submitted 30.Apr.2025; peer-reviewed by AG Gad, A Staffini, S Meister; revised version received 25.Nov.2025; accepted 26.Nov.2025; published 26.Jan.2026.

Please cite as:
Raszke P, Giebel GD, Wasem J, Adamzik M, Nowak H, Palmowski L, Heinz P, Timmesfeld N, Tokic M, Brunkhorst FM, Blase N
Patient Benefits in the Context of Sepsis-Related AI-Based Clinical Decision Support Systems: Scoping Review
J Med Internet Res 2026;28:e76772
URL: https://www.jmir.org/2026/1/e76772
doi:10.2196/76772

© Pascal Raszke, Godwin Denk Giebel, Jürgen Wasem, Michael Adamzik, Hartmuth Nowak, Lars Palmowski, Philipp Heinz, Nina Timmesfeld, Marianne Tokic, Frank Martin Brunkhorst, Nikola Blase. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Diagnostic Performance of Deep Learning and Radiomics in Extracranial Carotid Plaque Detection: Systematic Review and Meta-Analysis

Lingjie Ju[1*], MM; Yongsheng Guo[1*], MM; Haiyong Guo[2], MM; Ruijuan Liu[1], MM; Yiyang Wang[3], MM; Siyu Wang[3], MD; Na Ma[3], MD; Junhong Ren[1], MM

[1]Department of Sonography, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, 1 Dahua Road, Dongdan, Dongcheng District, Beijing, China

[2]Department of Sonography, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Peking University Health Science Center, Beijing, China

[3]Department of Sonography, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Junhong Ren, MM
Department of Sonography, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, 1 Dahua Road, Dongdan, Dongcheng District, Beijing, China

Abstract

Background: Artificial intelligence–enhanced imaging techniques have demonstrated promising diagnostic potential for carotid plaques, a key cardiovascular and cerebrovascular risk factor. However, previous studies did not systematically synthesize their diagnostic accuracy.

Objective: This study aimed to quantitatively explore the diagnostic efficacy of deep learning (DL) and radiomics for extracranial carotid plaques and establish a standardized framework for improving plaque detection.

Methods: We searched the PubMed, Embase, Cochrane, Web of Science, and Institute of Electrical and Electronics Engineers databases to identify studies involving the use of radiomics or DL models to diagnose extracranial carotid artery plaques from inception up to September 24, 2025. The quality of the studies was determined using Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence (QUADAS-AI). A meta-analysis was conducted using StataMP (version 17.0; StataCorp) with a bivariate mixed-effects model to calculate pooled sensitivity and specificity, generate summary receiver operating characteristic (SROC) curves, assess Cochran *Q* statistic and I^2-based heterogeneity, and conduct subgroup analyses and regression analysis.

Results: Among 40 studies comprising 17,246 patients, 34 integrated independent test sets or validation sets in the quantitative statistical analysis. Among them, 24 focused on DL models, 10 on machine learning models based on radiomics. The combined sensitivity, specificity, and area under the SROC curve were 0.88 (95% CI 0.85 - 0.91; P<.001; I^2=93.58%), 0.89 (95% CI 0.85 - 0.92; P<.001; I^2=91.38%), and 0.95 (95% CI 0.92 - 0.96), respectively. Compared with the machine learning models based on radiomics algorithms, DL models achieved comparable improvements in specificity and area under the SROC curve. It was observed that transfer learning and a large sample size enhanced the diagnostic performance of models. Models used to identify plaque stability and presence had similar diagnostic performances, both of which were more effective in identifying symptomatic plaque models. A total of 7 studies demonstrated that the models that combined clinical features exhibited comparable diagnostic capability to pure DL and radiomics models. Additionally, 7 studies performed external validation, obtaining lower diagnostic performance than in testing groups. Limited regression analysis failed to identify significant sources of heterogeneity, and the limited number of eligible studies restricted more comprehensive subgroup analyses. The high heterogeneity in the study results may be due to different scanning parameters, model architecture, image segmentation, and algorithms.

Conclusions: Radiomics algorithms and DL models can effectively diagnose extracranial carotid plaque. However, there are concerns regarding irregularities in research design and the absence of multicenter studies and external validation. Future research should aim to reduce bias risk and enhance the generalizability and clinical orientation of the models.

(J Med Internet Res 2026;28:e77092) doi:10.2196/77092

KEYWORDS

extracranial carotid plaque; deep learning; radiomics; systematic review; diagnosis

XSL•FO
RenderX

Introduction

Extracranial carotid plaques are biomarkers of coronary artery disease and cerebral ischemic events, including ischemic heart disease and stroke. The global prevalence of carotid plaques among individuals aged 30 - 79 years is estimated at 21.1% (n=815.76 million) in 2020. This high prevalence reflects a growing global burden of cardiovascular and cerebrovascular diseases, posing a significant challenge to public health systems [1]. Therefore, early detection and management of carotid plaque can potentially reduce the risk of stroke and cardiovascular events [2-4], and thus, effective detection and classification technologies need to be prioritized.

Imaging methods for carotid plaque imaging, such as ultrasound, computed tomography angiography (CTA), magnetic resonance imaging (MRI), and digital subtraction angiography, facilitate detection, stenosis assessment, and plaque composition analysis [5]. Conventional ultrasound is the first-line screening method [6]. Studies show that periapical radiographs (PRs) can serve as a supplementary screening tool, demonstrating a 50% concordance with ultrasound or CTA [7-9]. Current imaging primarily identifies high-risk features, such as plaque neovascularity, lipid-rich necrotic cores, thin fibrous caps, and intraplaque hemorrhage plaque ulceration [4,10]. Among them, the contrast-enhanced ultrasound or superb microvascular imaging can accurately quantify neovascularization and correlates well with histopathology [11-14], offering rapid, noninvasive, and reliable quantification [15]. It is proficient in vascular imaging and ulcer detection [16], as well as stenosis assessment [17], but it faces challenges with small lipid cores and thin fibrous caps [18]. MRI remains the gold standard for assessing plaque composition, particularly for identifying lipid cores and intraplaque hemorrhage [19]. While digital subtraction angiography is the reference standard, its invasive nature limits its application. Notably, the accuracy of these diagnostic techniques largely relies on the expertise of imaging or clinical physicians, which causes inconsistencies in the assessment results of carotid atherosclerotic plaques—particularly in measuring carotid intima-media thickness, characterizing intraplaque components, and evaluating fibrous cap integrity.

The radiomics algorithms and deep learning (DL) models have demonstrated significant potential in medical image analysis [20]. Radiomics is a quantitative medical imaging analysis approach that aims to transform high-dimensional image features (such as texture heterogeneity, spatial topological relationships, and intensity distribution) into quantifiable digital biomarkers, thereby providing objective evidence to guide clinical decision-making. However, the characteristic dimensionality of radiomics data often far exceeds sample sizes, which renders the traditional statistical methods inadequate [21]. Machine learning (ML), with the potential to process large-scale, high-dimensional data and uncover deep correlations among these complex features [22]. Combining radiomics with ML to develop an ML model using radiomics can enhance the diagnostic performance of AI in large and complex datasets, exceeding the performance of models constructed through traditional statistical methods.

DL is also one of the important subbranches of artificial intelligence, which can automatically learn and layer from raw data without manual design of features, ultimately generating predictions via an output layer [23]. DL-driven image generation techniques have demonstrated remarkable effectiveness in cross-modality imaging and synthesis tasks across various sequences within the same modality. With the rapid development of computer technology, ML models based on radiomics and DL models based on radiomics have become important tools for cardiovascular disease research. Current evidence suggests that these methods can significantly improve the quantitative assessment accuracy of atherosclerotic plaque progression and enhance the diagnostic and predictive power of major adverse cardiovascular events [24-26]. In recent years, research on the application of these methods in the fields of plaque diagnosis, stability assessment, and symptomatic plaque identification has increased significantly. Although these advancements have significantly improved the diagnosis of carotid plaques, variations in data dependency and imaging configurations among different models create inconsistencies in diagnostic accuracy. Moreover, these models may become overly specialized in common imaging configurations, even when using radiomics data from identical sources. Currently, systematic evaluations of its clinical validity remain limited.

Therefore, this systematic review comprehensively assesses the applications of ML models based on radiomics algorithms and DL models in carotid plaques, while highlighting gray areas in the available literature.

Methods

Study Registration

The study was performed in line with the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy Studies) guidelines [27] and PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) standards [28,29] and was registered on the International Prospective Register of Systematic Reviews (PROSPERO CRD42025638492).

Data Sources and Search Strategy

Relevant articles were searched on PubMed, Embase, Web of Science, Cochrane Library, and Institute of Electrical and Electronics Engineers (IEEE) databases, focusing on English-language articles published up to September 24, 2025. The literature search was based on the PIO (population, intervention, and outcomes) principles: “P” represents carotid artery disease, carotid plaques, or atherosclerosis populations; “I” represents radiomics or DL as interventions; and “O” represents the outcomes of diagnosis and their subordinates and other keywords. Furthermore, we manually analyzed the reference lists of all included articles to identify additional relevant publications. The complete search strategy is outlined in Table S1 in Multimedia Appendix 1. The EndNote 20 software (Clarivate Analytics) was used to manage the included studies.

XSL•FO
RenderX

Eligibility Criteria

Inclusion Criteria

The inclusion criteria included:

1. Studies on patients with extracranial carotid plaques that aimed to detect or distinguish between unstable and symptomatic plaques, among other factors.
2. Studies using radiomics algorithms or DL models based on medical imaging techniques, such as ultrasound, CTA, or MRI, to diagnose carotid plaques.
3. Studies reported the diagnostic performance metrics, including confusion matrix, 2×2 diagnostic tables, accuracy, sensitivity, specificity, receiver operating characteristic (ROC) curves, F_1-score, precision, recall, etc.
4. Those that adopted the following designs: prospective or retrospective cohorts, diagnostic accuracy trials, model development or validation studies, and comparative studies (eg, AI models vs AI models combined with clinical features).
5. Only studies published in English and with extractable quantitative data were deemed eligible.

Exclusion Criteria

The exclusion criteria excluded:

1. Studies involving nonhuman subjects (animal experiments or in vitro models), those that explored intracranial or coronary plaques, enrolled pediatric populations (<18 years), or reported only generalized atherosclerosis without plaque-specific criteria (focal intima-media thickness ≥1.5 mm) or specific diagnostic metrics;
2. Those that did not adopt well-defined deep learning models or radiomics algorithms, focused only on image segmentation or texture analysis without diagnostic validation, or reported predictive models without providing a clear diagnostic relevance.
3. Studies that lacked a validated reference standard.
4. Studies that did not report diagnostic performance.
5. Informal publication types (eg, reviews, letters to the editor, editorials, and conference abstracts).
6. Studies that did not report validation or test sets.

Screening of Articles and Data Extraction

In the initial screening, duplicates were excluded followed by reading of full texts, and data were entered into a predefined extraction table, which included surnames of authors, source of data, publication year, algorithm architecture, type of internal validation, availability of open access data, external verification status, reference standard, transfer learning application, number of cases for training, test, internal, or external validation, study design, sample size, mean or median age, inclusion criteria, and model evaluation metrics. The contingency tables are derived from the models explicitly identified by the original authors as the best-performing ones. Data from external validation sets were prioritized. If there were no external validation set in the original studies, data from internal validation sets were used. If neither was available, the contingency tables corresponding to the test sets were selected. This process was performed by two researchers (LJ and YG), working independently, and any differences were resolved through discussion with a third researcher (HG).

Quality Assessment

Two blinded investigators (LJ and YG) systematically assessed the quality of studies using the Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence (QUADAS-AI) tool. Specifically, they evaluated the risk of bias and applicability concerns across 4 domains: flow and timing, reference standard, index test, and participant selection. Although the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) is extensively applied to assess the quality of diagnostic accuracy studies [30], it does not address the specific methodological choices, result analyses, and measurements related to diagnostic studies using AI. To address this gap, QUADAS-AI was developed as a consensus-based tool to aid readers in systematically examining the risk of bias and the usability of AI-related diagnostic accuracy studies (Table S6 in Multimedia Appendix 1) [31], thereby improving the quality assessment process [32,33]. Any evaluation discrepancies were resolved by a third investigator (HG).

Statistical Analysis

A meta-analysis was performed using STATA/MP software (version 17.0; Stata Corporation) with a bivariate random-effects model. For meta-analyses of the diagnostic accuracy of AI-based models, bivariate mixed-effects models can account for both within-study variability (random effects) and between-study heterogeneity (fixed effects), ensuring the robustness of the pooled estimates [34]. A contingency table was generated using data from the included literature, and then we calculated metrics such as the number of cases, the Youden index, sensitivity, specificity, and recall. The diagnostic efficacy of radiomics algorithms and DL models in evaluating carotid plaque was determined using a summary receiver operating characteristic (SROC) curve and area under the curve (AUC; 0.7≤AUC<0.8 fair; 0.8≤AUC<0.9 good; and AUC≥0.9 excellent). Publication bias was explored using Deeks funnel plot asymmetry test. The Fagan nomogram was developed to determine clinically pertinent posttest probabilities (P-post) and likelihood ratios (LRs). LRs were determined by comparing the probability of test results between diseased and nondiseased groups. The pretest probability was subsequently adjusted based on test results and LRs to obtain P-post [35]. The Cochran Q (P≤.05) and I^2 statistic were used to explore heterogeneity among the included studies, and regression analysis was conducted to assess sources of heterogeneity. I^2≤50% indicated mild heterogeneity, 50%<I^2<75% reflected moderate heterogeneity, and I^2≥75% indicated high heterogeneity.

The subgroup analysis encompassed the following factors: (1) model type (DL or ML model), (2) medical imaging modalities (PRs, ultrasound, MRI, or CTA), (3) application of transfer learning, (4) characteristics of carotid plaques (presence vs absence, stable vs vulnerable, and symptomatic vs asymptomatic), (5) comparison of the most effective ML model based on radiomics algorithm and DL models using the same dataset and clinicians' diagnoses, (6) different types of datasets (testing and validation), (7) low and high or unclear risk of bias

XSL•FO
RenderX

studies, (8) different sample sizes of model, and (9) models with different research designs (multicenter studies and single-center studies). To identify the sources of heterogeneity associated with nonthreshold effects, meta-regression was performed using the above-mentioned covariates.

Sensitivity analysis was performed to assess the stability of the results by several steps: (1) excluding specific articles one by one to determine the stability of the results, (2) excluding studies with extremely large sample sizes (N≥500; n=7 studies), (3) excluding studies with extremely small sample sizes (N≤50; n=4 studies), and (4) excluding studies with extreme effect sizes (sensitivity or specificity>0.95 or <0.7; n=11 studies).

Results

Study Selection

We obtained 5834 studies in the initial analysis, of which 1233 were excluded for duplication or redundancy. After screening titles and abstracts, 4507 publications were eliminated. After the full texts of the 94 articles were read, 40 studies were eligible for meta-analysis. The PRISMA flow diagram of the study showing the selection process is presented in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of study selection. IEEE: Institute of Electrical and Electronics Engineers.

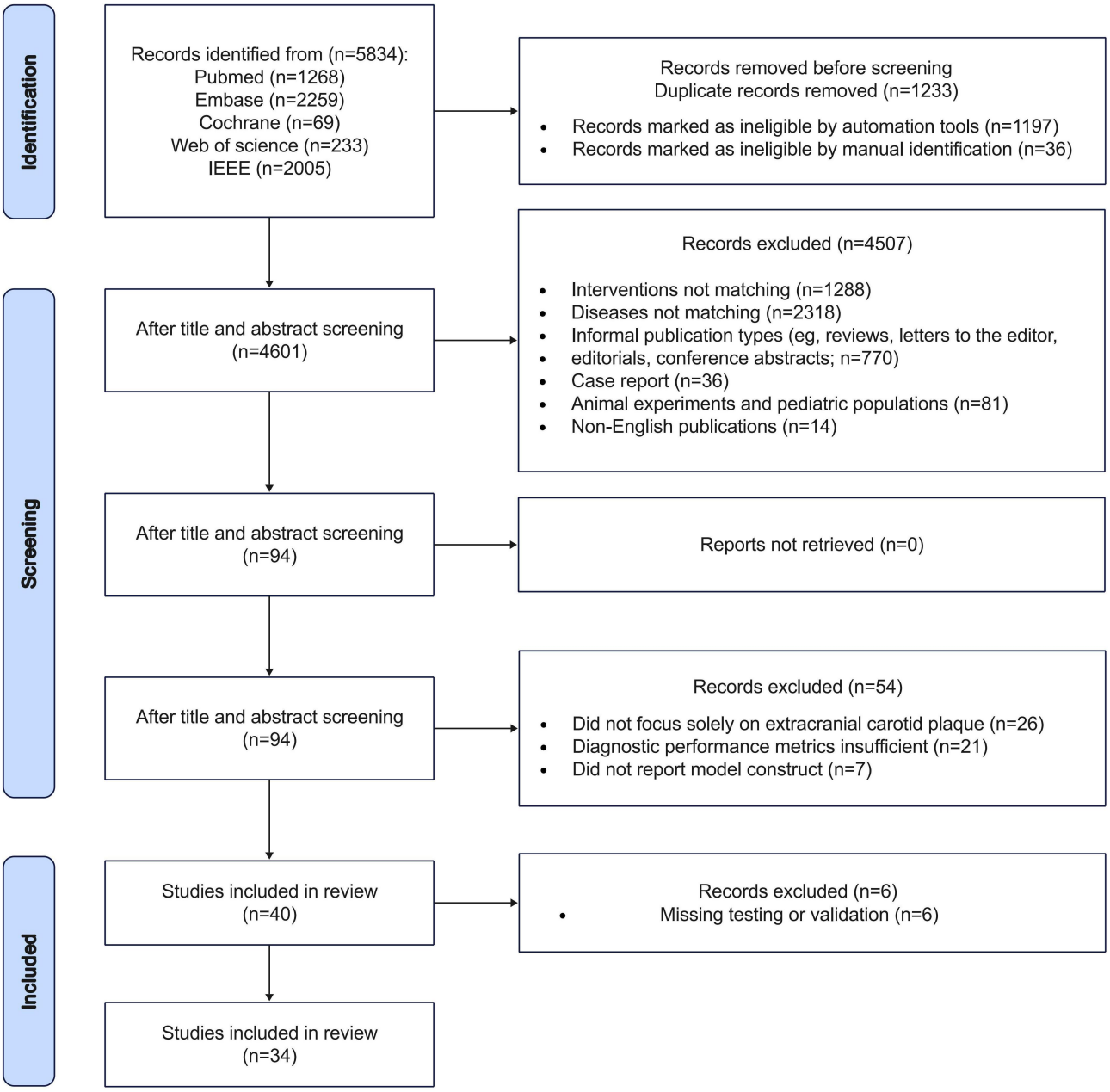

Study Characteristics

Among the 40 studies that fulfilled the systematic review's inclusion criteria, 34 provided sufficient quantitative data (contingency tables from validation or test sets) eligible for incorporation into the meta-analysis. The detailed characteristics of all 40 eligible studies are summarized in Tables S3 and S4 in Multimedia Appendix 1, while all subsequent quantitative

XSL•FO
RenderX

analyses were conducted based on the 34 studies with available quantitative data. Overall, 34 studies were included [36-69], among which 9 were multicenter studies [41,43,45,49,57,63-65,69], 3 used public databases [37,40,53], 13 provided open access to the data [37,40,45,48-50,53,57,59,63-66]. A total of 12 studies conducted internal validation [38,39,41,42,44,47,48,57,61,64,69,70] to confirm the reproducibility of the model development process and prevent overfitting. In addition, 7 studies conducted external validation [41,50,57,60,63,64,69] to assess the model's transportability and generalizability using unused datasets. Only 1 study conducted a comparative analysis of the diagnostic performance of DL models with that of clinicians [57]. The medical imaging modalities included PRs (n=5), ultrasound (n=16), MRI (n=5), and CTA (n=8). The core features of the 34 studies are presented in Tables 1 and 2, with further details provided in Tables S2 and S3 in Multimedia Appendix 1.

XSL•FO
RenderX

Table . Data sources, indicators, and algorithms of included studies.

Study, year	Data source				Validation type
	Source of data	Number of cases for training, test, internal, or external	Data range	Labels	
Su et al [51], 2023	China	322; 138; NR[a]; NR	NR	Stable or vulnerable plaque	No
Zhang et al [70], 2024	China	4064; NR; 1016; NR	NR	Stable or vulnerable plaque	Internal validation
Zhou et al [44], 2024	China	751; 261; 258; NR	NR	Stable or vulnerable plaque	Internal validation
Zhang et al [58], 2021	China	121; 41; NR; NR	NR	Symptomatic or asymptomatic	No
Zhai et al [41], 2024	NR	240; NR; 60; 100	January 2017-January 2022	Normal or abnormal	External validation
Yoo et al [39], 2024	South Korea	388; 130; 130; NR	2009 - 2022	Normal or abnormal	Internal validation
Xu et al [56], 2022	NR	NR	NR	Stable or vulnerable plaque	No
Xie et al [47], 2023	China	264; 75; 38; NR	2020 - 2021	Stable or vulnerable plaque	Internal validation
Wei et al [62], 2024	China	2725; 554; NR; NR	NR	Normal or abnormal	No
Ganitidis et al [60], 2021	Greece	46; 10; 18; NR	NR	Symptomatic or asymptomatic	Internal validation
Shi et al [50], 2023	China	134; 33; NR; NR	October 2019-July 2022	Symptomatic or asymptomatic	No
Gui et al [49], 2023	China	84; 20; NR; NR	NR	Symptomatic or asymptomatic	No
Ali et al [71], 2024	Italy	336; 84; NR; NR	NR	Symptomatic or asymptomatic	No
Amitay et al [48], 2023	Israel	371; 144; 144; NR	2016 - 2021	Normal or abnormal	Internal validation
Ayoub et al [72], 2023	China	136; 150; 69; NR	NR	Stable or vulnerable plaque	Internal validation
Cilla et al [55], 2022	Italy	NR	October 2015-October 2019	Stable or vulnerable plaque	No
Guang et al [57], 2021	China	136; NR; 69; NR	September 2017-September 2018	Stable or vulnerable plaque	Internal validation
He et al [69], 2024	China	3088; NR; 772; 1564	January 2021-March 2023	Normal or abnormal; stable or vulnerable plaque	Internal and external validation
Latha et al [73], 2021	India	NR	NR	Normal or abnormal	No
Ma et al [59], 2021	China	1169; 294; NR; NR	NR	A total of 3 types (echo-rich, intermediate, and echolucent)	No
Pisu et al [43], 2024	Italy	163; 106; NR; NR	March 2013-October 2019	Symptomatic or asymptomatic	No
Wang et al [74], 2024	China	154; 39; NR; NR	January 1, 2018-December 31, 2021	Symptomatic or asymptomatic	No
Gago et al [53], 2022	Spain	NR	2007 - 2010	Normal or abnormal	No
Omarov et al [40], 2024	The United Kingdom	577; 103; NR; NR	NR	Normal or abnormal	No
Wang et al [45], 2023	China	2619; 1122; NR; NR	NR	Stable or vulnerable plaque	No

XSL•FO
RenderX

Study, year	Data source				Validation type
	Source of data	Number of cases for training, test, internal, or external	Data range	Labels	
Vinayahalingam et al [42], 2024	Germany	280; 37; 37; NR	NR	Normal or abnormal	No
Singh et al [37], 2024	Cyprus; The United Kingdom; NR	3088; 772; NR; NR	NR	Stable or vulnerable plaque	No
Shan et al [46], 2023	China	52; 22; NR; NR	January 2018-December 2021	Stable or vulnerable plaque	No
Li et al [38], 2024	NR	4546; 1471; 1019; NR	NR	Normal or abnormal	Internal validation
Jain et al [54], 2021	NR	682; 76; NR; NR	July 2009-September 2010	Stable or vulnerable plaque	No
Molinari et al [36], 2018	Italy	NR	2004 - 2010	Symptomatic or asymptomatic	No
Kats et al [61], 2019	Israel	1946; 7; 12; NR	NR	Normal or abnormal	Internal validation
Chen et al [52], 2022	China	81; 34; NR; NR	July 2015-May 2021	Symptomatic or asymptomatic	No
Zhao et al [63], 2025	China	317; NR; NR; 328	January 2018-December 2023 (Center 1); Jan 2022-December 2023 (Center 2,3)	Symptomatic or asymptomatic	External validation
Hu et al [64], 2025	China	213; NR; 93; 110	January 2018-May 2023 (Center 1); January 2020-May 2023 (Center 2)	Symptomatic or asymptomatic	Internal and external validation
Li et al [75], 2025	China	2069; 887; NR; NR	October 2021-January 2022	normal or abnormal	No
Yu et al [66], 2025	China	146; 63; NR; NR	April 2022-August 2023	HIPs[b] or NHIPs[c]	No
Liapi et al [65], 2025	Cyprus, The United Kingdom, and Greece	168; 46; 22; NR	NR	Symptomatic or asymptomatic	Internal validation
Kuwada et al [67], 2025	Japan	Training and validation data: 500; Test data: 80	2008 - 2023	Normal or abnormal	No
Lao et al [68], 2025	China	76; 31; NR; NR	January 2017-October 2022	Stable or vulnerable plaque	No

[a]NR: not reported.

[b]HIP: highly inflammatory plaque.

[c]NHIP: non–highly inflammatory plaque.

XSL•FO
RenderX

Table . Data sources, indicators, and algorithms of all studies.

Study, year	Indicator definition		Algorithm		
	Device	Exclusion of poor quality cases	Algorithm architecture	ML[a] or DL[b]	Transfer learning applied
Su et al [51], 2023	Ultrasound	NR[c]	Inception V3; VGG-16[d]	DL	No
Zhang et al [70], 2024	Ultrasound	NR	Fusion-SSL	DL	No
Zhou et al [44], 2024	Ultrasound	NR	Tri-Correcting	DL	No
Zhang et al [58], 2021	MRI[e]	Yes	LASSO[f] MRI-based model (HRPMM[g])	ML models based on radiomics algorithms[h] (LASSO algorithm)	No
Zhai et al [41], 2024	CT	Yes	3D-UNet; ResUNet	DL	No
Yoo et al [39], 2024	PRs	Yes	CACSNet	DL	Yes
Xu et al [56], 2022	Ultrasound	NR	Multi-feature fusion method	DL	No
Xie et al [47], 2023	Ultrasound	NR	CPTV[i]	DL	No
Wei et al [62], 2024	Ultrasound	Yes	BETU[j]	DL	Yes
Ganitidis et al [60], 2021	Ultrasound	NR	CNNs[k]	DL	No
Shi et al [50], 2023	CT[l] and MRI	Yes	LASSO regression	ML models based on radiomics algorithms (LASSO algorithm)	No
Gui et al [49], 2023	MRI	Yes	3D-SE-DenseNet121[m]; ANOVA_spearman_LASSO and MLP[n]	ML models based on radiomics algorithms (LASSO, ANOVA_LASSO and ANOVA_spearman_LASSO) and DL	No
Ali et al [71], 2024	Ultrasound	No	CAROTIDNet[o]	DL	No
Amitay et al [48], 2023	PRs	Yes	InceptionResNetV2 (minimum-maximum)	DL	Yes
Ayoub et al [72], 2023	MRI	NR	HViT[p]	DL	No
Cilla et al [55], 2022	CT	Yes	SVM RBF[q] kernel	ML models based radiomics algorithms (logistic regression [LR]), support vector machine (SVM), and CART[r]	No
Guang et al [57], 2021	Ultrasound	Yes	DL-DCCP[s]	DL	Yes
He et al [69], 2024	Ultrasound	Yes	BCNN[t]-ResNet[u]	DL	No
Latha et al [73], 2021	Ultrasound	NR	CART; logistic regression; random forest; CNN; Mobilenet; Capsulenet	ML models based radiomics algorithms (CART, logistic regression, and random forest algorithm) and DL	Yes
Ma et al [59], 2021	Ultrasound	NR	MSP[v]-VGG	DL	Yes
Pisu et al [43], 2024	CT	Yes	GB-GAM[w]	ML models based radiomics algorithms (NR)	No
Wang et al [74], 2024	CT	Yes	SR[x]	DL	Yes
Gago et al [53], 2022	Ultrasound	NR	End-to-end framework	DL	No

XSL•FO
RenderX

Study, year	Indicator definition		Algorithm		
	Device	Exclusion of poor quality cases	Algorithm architecture	ML[a] or DL[b]	Transfer learning applied
Omarov et al [40], 2024	Ultrasound	Yes	YOLOv8[y]	DL	Yes
Wang et al [45], 2023	MRI	Yes	ResNet-50	DL	Yes
Vinayahalingam et al [42], 2024	PRs[z]	Yes	Faster R-CNN[aa] with Swin Transformer (Swin-T)	DL	Yes
Singh et al [37], 2024	Ultrasound	Yes	GoogLeNet[ab]	ML models based on radiomics algorithms (SVM algorithms) and DL	Yes
Shan et al [46], 2023	CT and ultrasound	Yes	LR[ac]; SVM[ad]; RF[ae]; LGBM[af]; daBoost; XGBoost[ag]; MLP	ML models based on radiomics algorithms (Pyradiomics package in Python software)	Yes
Li et al [38], 2024	Ultrasound	NR	U-Net; CNN	DL	No
Jain et al [54], 2022	Ultrasound	NR	SegNet-UNet[ah]	DL	No
Molinari et al [36], 2018	Ultrasound	NR	SVM	ML models based on radiomics algorithms (BEMD[ai])	No
Kats et al [61], 2019	PRs	NR	Faster R-CNN	DL	No
Chen et al [52], 2022	MRI	Yes	LASSO	ML models based on radiomics algorithms (mRMR[aj] algorithm and LASSO algorithm)	No
Zhao et al [63], 2025	CTA[ak]	Yes	XGBoost	ML models based on radiomics algorithms (XGBoost)	No
Hu et al [64], 2025	CTA	Yes	LASSO regression; SVM; logistic regression	ML models based on radiomics algorithms (LASSO algorithm) and classifier (SVM)	No
Li et al [75], 2025	Ultrasound	NR	XGBoost; RF; LASSO regression	ML models based on radiomics algorithms (XGBoost, RF, LASSO regression)	No
Yu et al [66], 2025	MRI	Yes	Plaque-R model; PVAT-R[al] model; ensemble model	ML models based on radiomics algorithms (LASSO algorithm) and ensemble learning	No
Liapi et al [65], 2025	Ultrasound	NR	Xception	DL	Yes
Kuwada et al [67], 2025	Ultrasound	NR	GoogLeNet; YOLOv7	DL	No

XSL•FO
RenderX

Study, year	Indicator definition		Algorithm		
	Device	Exclusion of poor quality cases	Algorithm architecture	ML[a] or DL[b]	Transfer learning applied
Lao et al [68], 2025	CTA	Yes	mRMR algorithm; LASSO regression	ML models based on radiomics algorithms (mRMR algorithm; LASSO algorithm)	No

[a]ML: machine learning.

[b]DL: deep learning.

[c]NR: not reported.

[d]VCG: VGG visual geometry group network.

[e]MRI: magnetic resonance imaging.

[f]LASSO: least absolute shrinkage and selection operator.

[g]HRPMM: high-risk plaque MRI-based model.

[h]Definition of ML models based on radiomics algorithms and deep learning (DL): ML models based on radiomics algorithms are models that rely on artificially designed features (such as texture and shape features) and use traditional algorithms (such as random forest, support vector machine, logistic regression, etc) to complete classification, without the need for DL algorithms to be in the core task. The DL model was defined as a model that automatically extracts features and completes classification through neural networks (such as convolutional neural network, ResNet, etc), regardless of whether the input contains a small number of artificial features, as long as the core task relies on the DL algorithm.

[i]CPTV: classification of plaque by tracking videos.

[j]BETU: be easy to use.

[k]CNN: convolutional neural network.

[l]CT: computed tomography.

[m]3D-SE-DenseNet121: 3D squeeze-and-excitation DenseNet with 121 layers.

[n]MLP: multilayer perceptron.

[o]CAROTIDNet: carotid symptomatic/asymptomatic plaque detection network.

[p]HViT: hybrid vision transformer.

[q]SVM RBF: kernel support vector machine with radial basis function kernel.

[r]CART: classification and regression tree.

[s]DL-DCCP: deep learning-based detection and classification of carotid plaque.

[t]BCNN: bilinear convolutional neural network.

[u]ResNet: deep residual network.

[v]MSP: multilevel strip pooling.

[w]GB-GAM: gradient-boosting generalized additive model.

[x]SR: super resolution.

[y]YOLOv8: you only look once version 8.

[z]PR: panoramic radiograph.

[aa]Faster R-CNN: faster region-based convolutional network.

[ab]GoogLeNet: Google network.

[ac]LR: logistic regression.

[ad]SVM: support vector machine.

[ae]RF: random forest.

[af]LGBM: light gradient boosting machine.

[ag]XGBoost: extreme gradient boosting.

[ah]SegNet-UNet: segmentation network-UNet.

[ai]BEMD: bidimensional empirical mode decomposition.

[aj]mRMR: minimum redundancy maximum relevance.

[ak]CTA: computed tomography angiography.

[al]PVAT: perivascular adipose tissue.

Meta-Analysis of Diagnostic Performance

Synthesized Results

The meta-analysis revealed pooled sensitivity, specificity, and an area under the SROC curve (SROC AUC) of 0.88 (95% CI 0.85 - 0.91; I^2=93.58%; P<.001; in Multimedia Appendix 2 [36-69]), 0.89 (95% CI 0.85 - 0.92; I^2=91.38%; P<.001; in Multimedia Appendix 2 [36-69]), and 0.95 (95% CI 0.92 - 0.96) for all 34 studies (Figure 2A); 0.88 (95% CI 0.84 - 0.92; I^2=93.70%; P<.001; Multimedia Appendix 3 [36-69]), 0.91

XSL•FO
RenderX

(95% CI 0.86 - 0.94; I^2=95.55%; *P*<.001; Multimedia Appendix 3 [36-69]), and 0.95 (95% CI 0.93 - 0.97) for all DL models (Figure 2B); 0.89 (95% CI 0.82 - 0.93; I^2=90.20%; *P*<.001; Multimedia Appendix 3 [36-69]), 0.83 (95% CI 0.76 - 0.88; I^2=78.92%; *P*<.001; Multimedia Appendix 3 [36-69]), and 0.92 (95% CI 0.89 - 0.94) for all ML models based on radiomics algorithms (Figure 2C), respectively. Notably, some studies used multiple diagnostic models; however, the diagnostic accuracy of certain models was not thoroughly assessed.

Figure 2. Receiver operating characteristic curves based on the overall performance of different algorithms. (A) All studies included in the meta-analysis (34 studies with 34 tables). (B) Deep learning (DL) models (24 studies with 24 tables). (C) Machine learning (ML) models based on radiomics algorithms (10 studies with 10 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

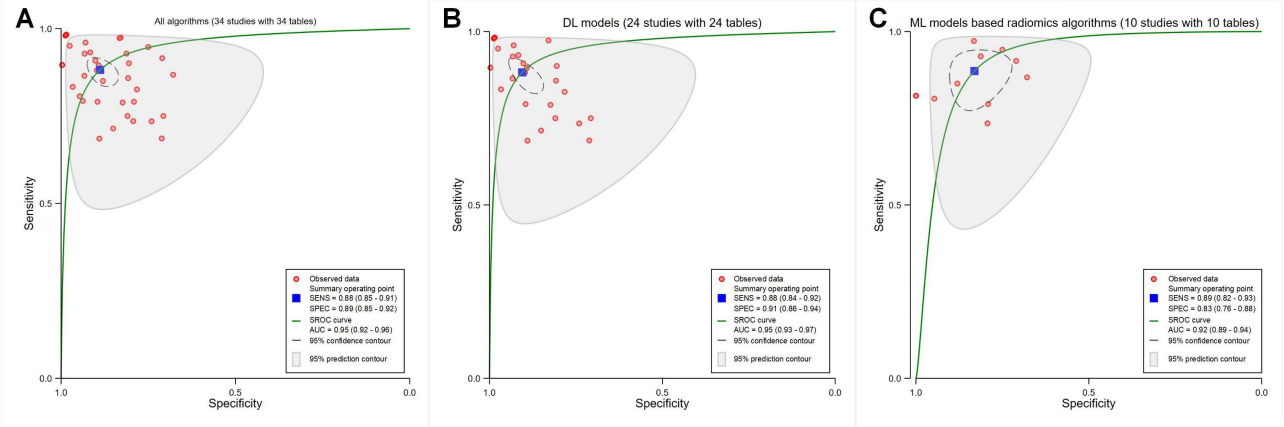

Subgroup Analysis

Medical Imaging Modalities

The pooled sensitivity, specificity, and SROC AUC were 0.91 (95% CI 0.80 - 0.96), 0.93 (95% CI 0.84 - 0.97), and 0.97 (95% CI 0.95 - 0.98) for the 5 studies using PRs (*P*<.001; with 5 contingency tables; Figure 3A); 0.89 (95% CI 0.84 - 0.93), 0.90 (95% CI 0.84 - 0.94), and 0.95 (95% CI 0.93 - 0.97) for the 16 studies using ultrasound images (*P*<.001with 16 contingency tables; Figure 3B); 0.87 (95% CI 0.87 - 0.92), 0.87 (95% CI 0.76 - 0.93), and 0.93 (95% CI 0.91 - 0.95) for the 5 studies using MRI images (*P*<.001; with 5 contingency tables; Figure 3C); 0.83 (95% CI 0.76 - 0.88), 0.83 (95% CI 0.75 - 0.89), and 0.90 (95% CI 0.87 - 0.92) for the 8 studies using CTA images (*P*<.001; with 8 contingency tables; Figure 3D), respectively. In addition, we conducted subgroup analyses using the same imaging modality based on differentiation. However, only subgroups of identifying the presence and stability of plaque had sufficient data for the ultrasound modality to perform statistical analyses and obtain pooled diagnostic performance metrics (Table S5 in Multimedia Appendix 1). The pooled sensitivity, specificity, and SROC AUC were 0.88 (95% CI 0.72 - 0.96), 0.91 (95% CI 0.80 - 0.96), and 0.95 (95% CI 0.93 - 0.97) for determining the presence of plaques (*P*<.001; with 5 contingency tables; Figure 3E), 0.90 (95% CI 0.84 - 0.94), 0.92 (95% CI 0.83 - 0.96), and 0.96 (95% CI 0.94 - 0.97) for distinguishing the stability of plaques (*P*<.001; with 8 contingency tables; Figure 3F).

XSL•FO
RenderX

Figure 3. Receiver operating characteristic curves for different medical imaging modalities. (A) Periapical radiographs (PRs) imaging models (5 studies with 5 tables). (B) Ultrasound imaging models (16 studies with 22 tables). (C) Magnetic resonance imaging (MRI) models (5 studies with 7 tables). (D) Computed tomography angiography (CTA) models (8 studies with 10 tables). (E) Models based on ultrasound modality for detecting the presence of carotid plaque (5 studies with 5 tables). (F) Models based on ultrasound modality for distinguishing the stability of carotid plaques (8 studies with 8 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

XSL•FO
RenderX

Use of Transfer Learning

The pooled sensitivity, specificity, and SROC AUC were 0.92 (95% CI 0.87 - 0.95), 0.93 (95% CI 0.88 - 0.96), and 0.97 (95% CI 0.95 - 0.96) for the 10 studies using transfer learning (*P*<.001; with 10 contingency tables; Figure 4A) and 0.86 (95% CI 0.82 - 0.90), 0.86 (95% CI 0.81 - 0.90), and 0.93 (95% CI 0.90 - 0.95) for the 24 studies without transfer learning (*P*<.001; with 24 contingency tables; Figure 4B), respectively.

Figure 4. Receiver operating characteristic curves demonstrating transfer learning application. (A) Models using transfer learning (10 studies with 10 tables). (B) Models without transfer learning (24 studies with 24 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

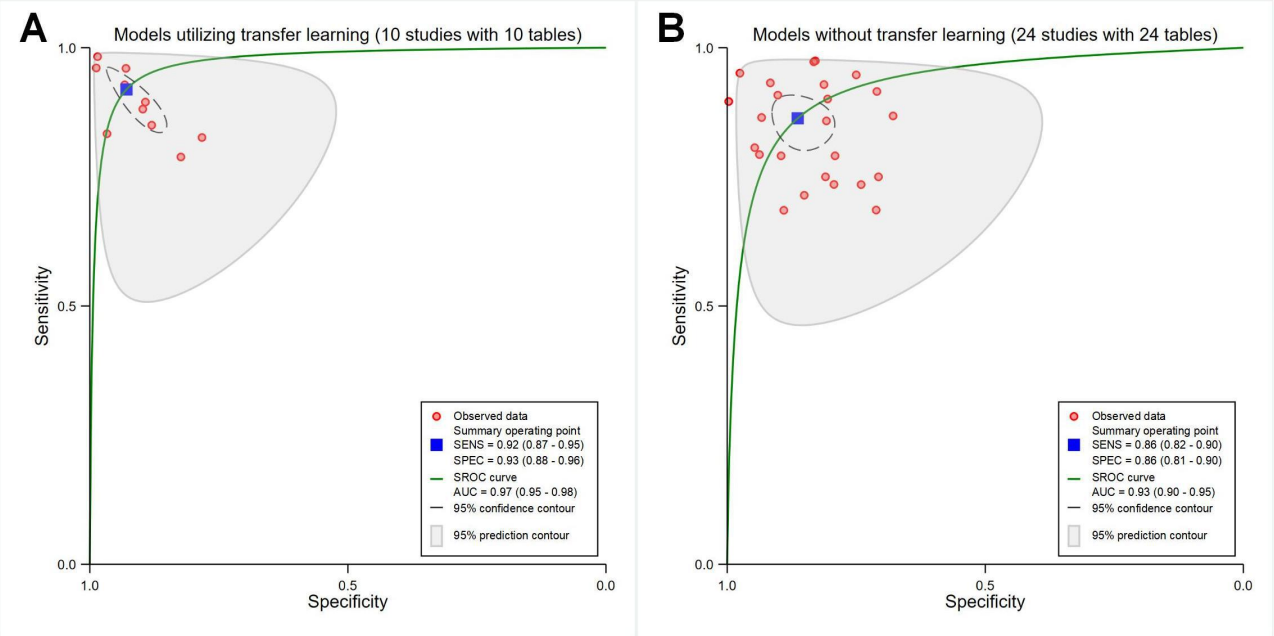

Carotid Plaque Type

The pooled sensitivity, specificity, and AUC were 0.89 (95% CI 0.81 - 0.94), 0.91 (95% CI 0.86 - 0.95), and 0.96 (95% CI 0.94 - 0.97) for the 11 studies identifying the presence or absence of carotid plaques (*P*<.001; with 11 contingency tables; Figure 5A); 0.90 (95% CI 0.85 - 0.94), 0.91 (95% CI 0.85 - 0.95), and 0.96 (95% CI 0.94 - 0.97) for the 12 studies identifying stable or vulnerable carotid plaques (*P*<.001; with 12 contingency tables), respectively (Figure 5B); and 0.86 (95% CI 0.78 - 0.91), 0.81 (95% CI 0.74 - 0.87), and 0.90 (95% CI 0.87 - 0.92) for the 10 studies identifying symptomatic or asymptomatic plaques (*P*<.001; with 10 contingency tables; Figure 5C), respectively.

Figure 5. Receiver operating characteristic curves for different carotid plaque types. (A) Presence versus absence of carotid plaques (11 studies with 11 tables). (B) Stable versus vulnerable carotid plaques (12 studies with 12 tables). (C) Symptomatic versus asymptomatic carotid plaques (10 studies with 10 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

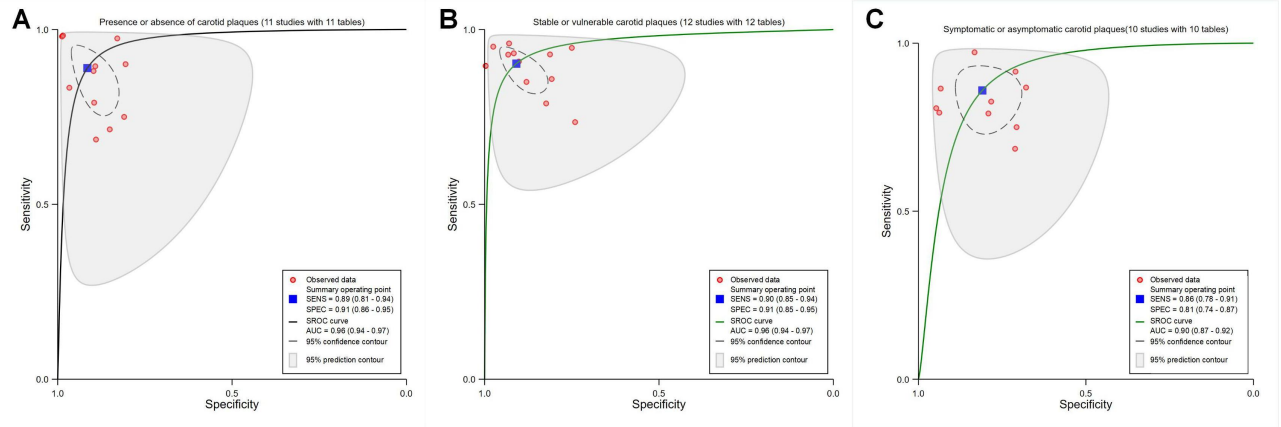

Pure Artificial Intelligence Models Versus Models Constructed by Combining Clinical Features

The pooled sensitivity, specificity, and SROC AUC were 0.82 (95% CI 0.74 - 0.88), 0.74 (95% CI 0.69 - 0.79), and 0.77 (95% CI 0.73 - 0.80) for the 7 studies involving pure artificial intelligence models meeting the inclusion criteria (*P*<.001; with 7 contingency tables; Figure 6A) and 0.85 (95% CI 0.76 - 0.92), 0.75 (95% CI 0.70 - 0.80), and 0.77 (95% CI 0.73 - 0.81) for models constructed by combining clinical features (*P*<.001; with 7 contingency tables; Figure 6B), respectively.

XSL•FO
RenderX

Figure 6. Receiver operating characteristic curves showing the diagnostic performance of pure artificial intelligence models or models constructed by combining clinical features. (A) Artificial intelligence models (7 studies with 7 tables). (B) Combined models (7 studies with 7 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

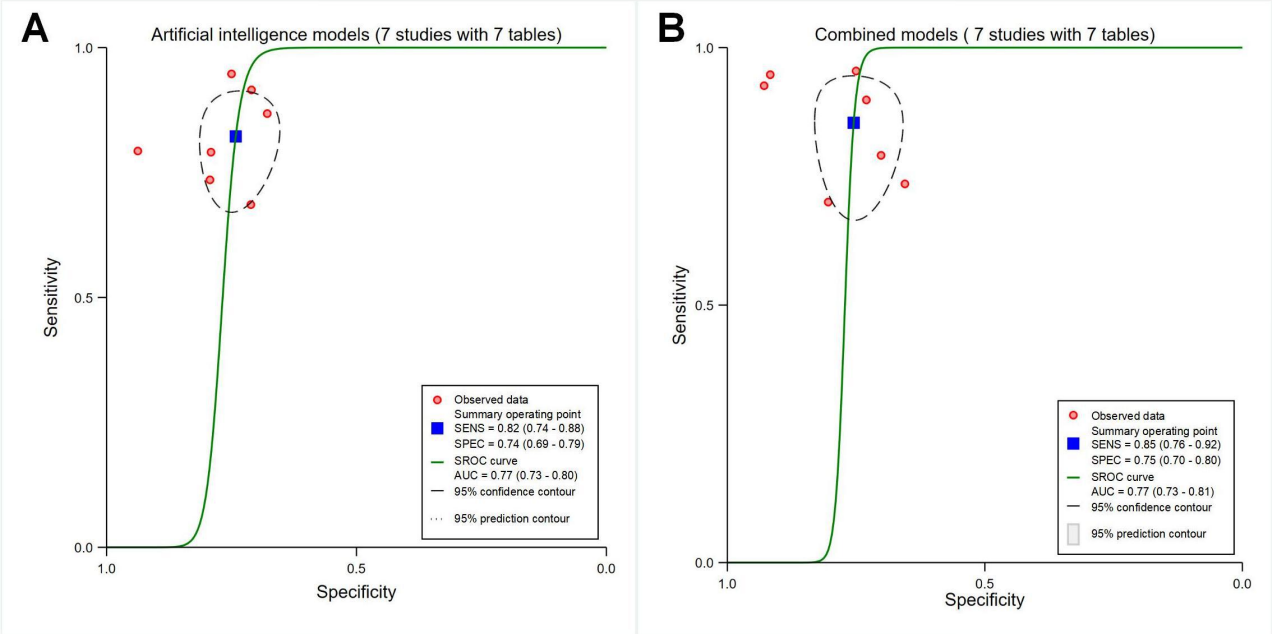

Different Sets of Datasets

The pooled sensitivity, specificity, and AUC were 0.90 (95% CI 0.87 - 0.93), 0.91 (95% CI 0.87 - 0.93), and 0.96 (95% CI 0.94 - 0.97) for testing sets (P<.001; with 27 contingency tables; Figure 7A); 0.78 (95% CI 0.71 - 0.83), 0.80 (95% CI 0.73 - 0.86), and 0.86 (95% CI 0.82 - 0.88) for external validation sets (P<.001; with 7 contingency tables; Figure 7B), respectively.

Figure 7. Receiver operating characteristic curves showing different sets of datasets. (A) Testing (27 studies with 27 tables). (B) External validation (7 studies with 7 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

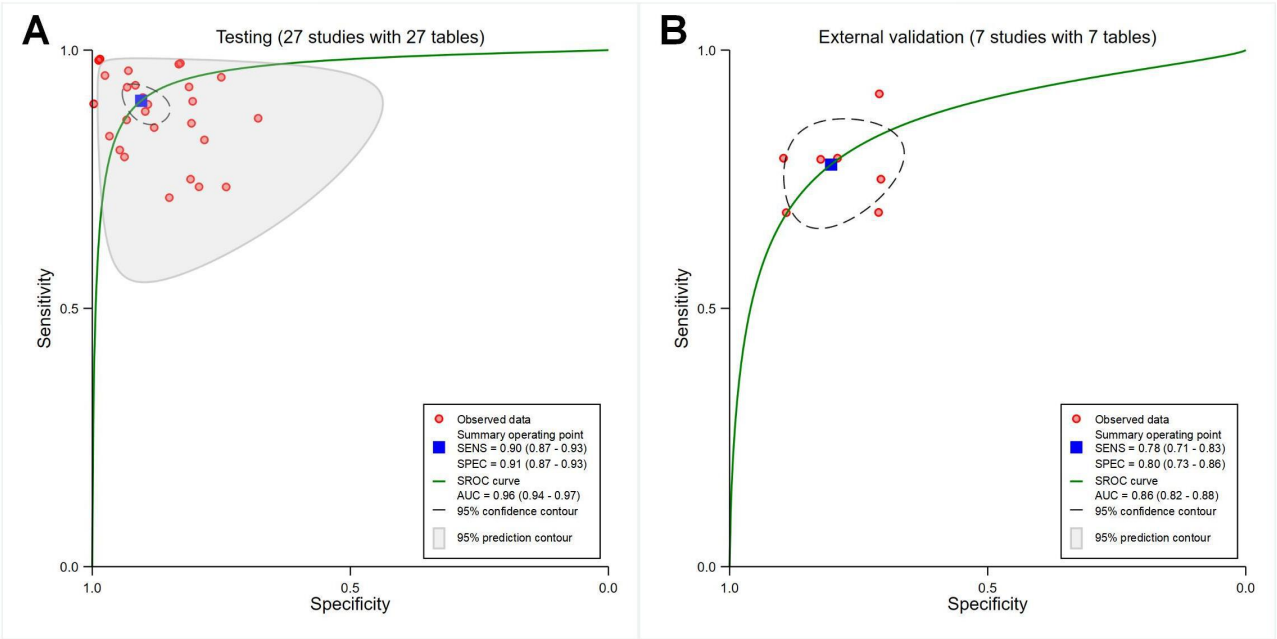

Low and High or Unclear Risk of Bias Studies

The pooled sensitivity, specificity, and AUC were 0.80 (95% CI 0.73 - 0.85), 0.80 (95% CI 0.71 - 0.87), and 0.86 (95% CI 0.83 - 0.89) for studies with a low risk of bias (P<.001; with 5 contingency tables; Figure 8A), and 0.89 (95% CI 0.86 - 0.92), 0.90 (95% CI 0.86 - 0.93), and 0.95 (95% CI 0.93 - 0.97) for studies with a high or unclear risk of bias (P<.001; with 29 contingency tables; Figure 8B), respectively.

XSL•FO
RenderX

Figure 8. Receiver operating characteristic curves showing studies with different risk of bias. (A) Studies with a low risk of bias (5 studies with 5 tables). (B) Studies with a high/unclear risk of bias (29 studies with 29 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

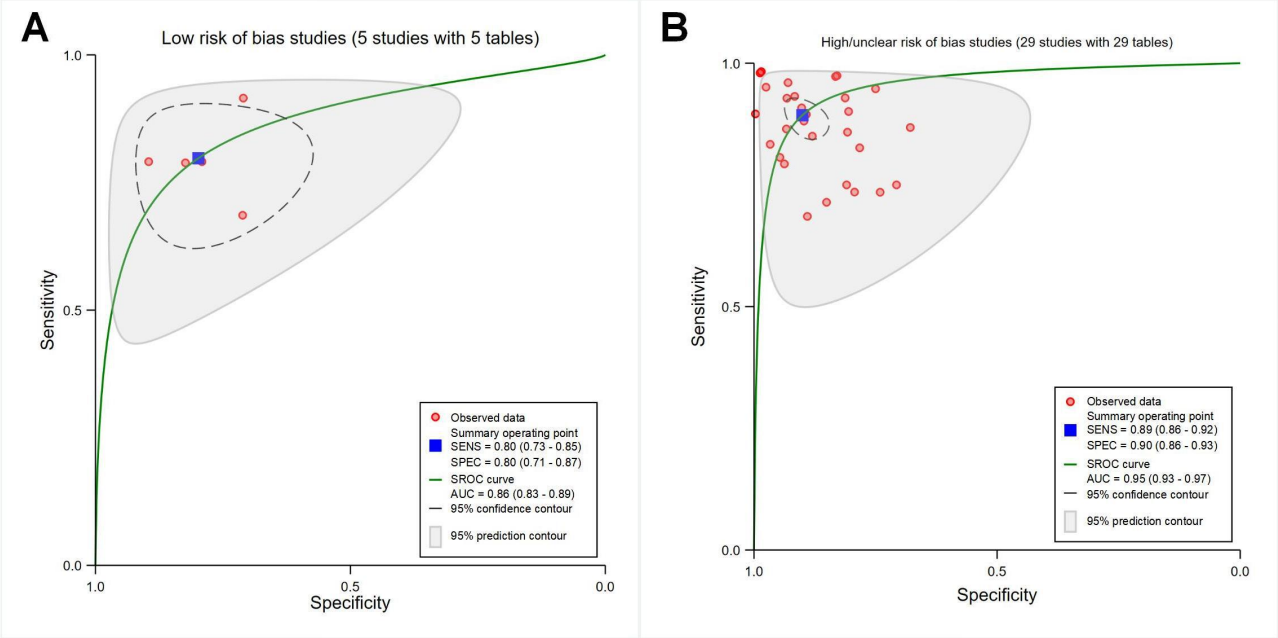

Different Sample Sizes of Model

The pooled sensitivity, specificity, and AUC were 0.91 (95% CI 0.86 - 0.94), 0.92 (95% CI 0.87 - 0.95), and 0.97 (95% CI 0.95 - 0.98) for sample size≥200 (*P*<.001; with 14 contingency tables) (Figure 9A), and 0.85 (95% CI 0.80 - 0.88), 0.86 (95% CI 0.80 - 0.90), and 0.91 (95% CI 0.89 - 0.94) for sample size<200 (*P*<.001; with 20 contingency tables; Figure 9B), respectively.

Figure 9. Receiver operating characteristic curves showing different sample sizes of model. (A) Sample size ≥200 (14 studies with 14 tables). (B) Sample size <200 (20 studies with 20 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

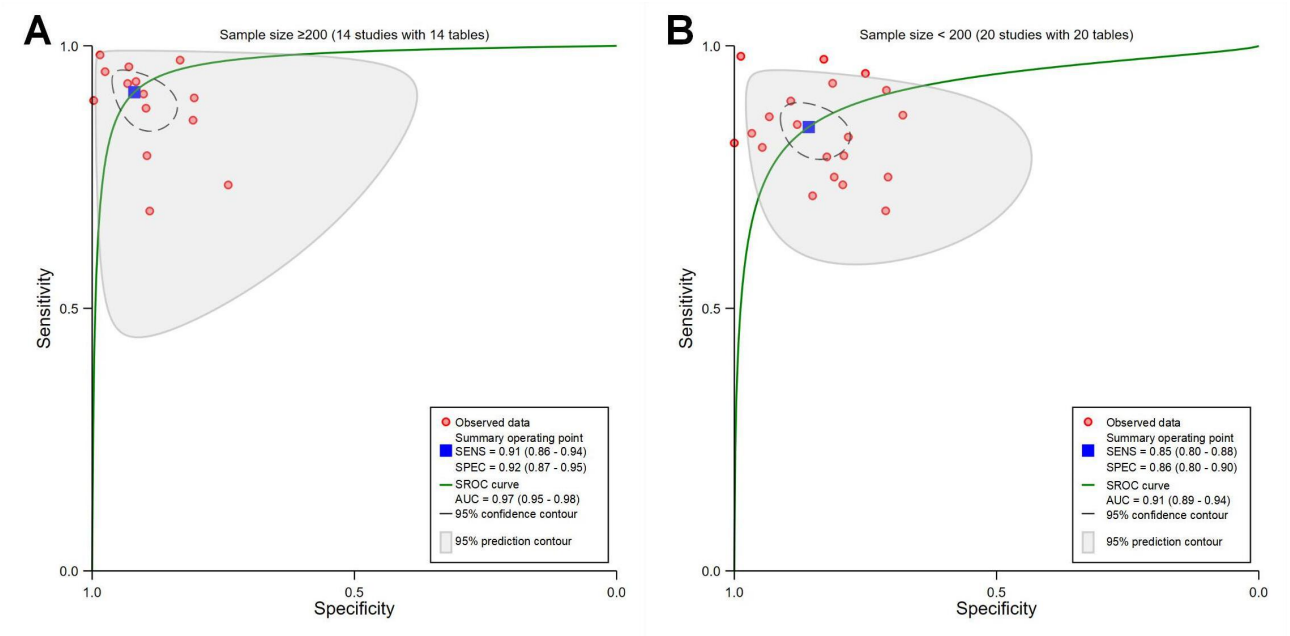

Models With Different Research Designs (Multicenter Studies and Single-Center Studies)

The pooled sensitivity, specificity, and AUC were 0.84 (95% CI 0.77 - 0.89), 0.87 (95% CI 0.81 - 0.91), and 0.92 (95% CI 0.90 - 0.94) for multicenter studies (*P*<.001; with 9 contingency tables; Figure 10A), and 0.89 (95% CI 0.84 - 0.92), 0.89 (95% CI 0.84 - 0.93), and 0.95 (95% CI 0.93 - 0.97) for single-center studies (*P*<.001; with 22 contingency tables; Figure 10B), respectively.

XSL•FO
RenderX

Figure 10. Receiver operating characteristic curves showing models with different research designs. (A) Multicenter studies (9 studies with 9 tables). (B) Single-center studies (22 studies with 22 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

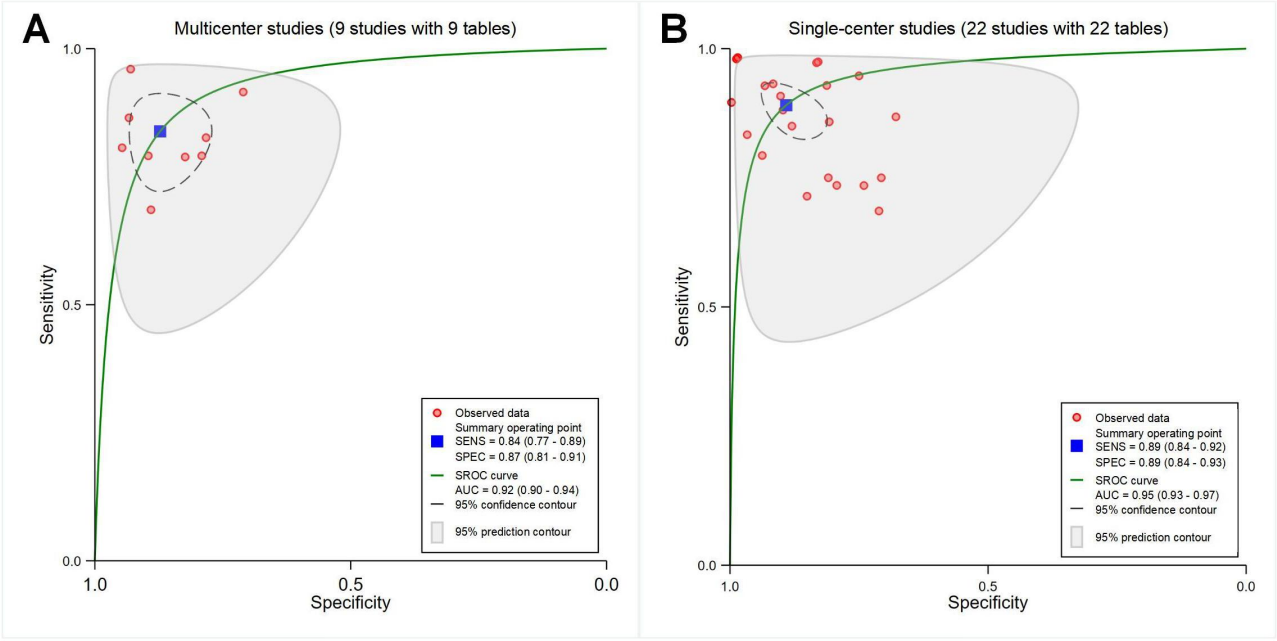

Heterogeneity Analysis and Meta-Regression Analysis

The Cochran Q test was used to indicate the presence of heterogeneity among subgroups (significance level $P \leq .05$) [15]. The I^2 index was used to assess the extent of heterogeneity among studies [15], revealing high sensitivity (I^2=93.58%) and specificity (I^2=91.38%; Multimedia Appendix 2). The Deek funnel plot asymmetry test, with P=.21, indicated no apparent publication bias (Multimedia Appendix 4). Subgroup analyses were performed using the random-effects models to identify the potential sources of heterogeneity, particularly when I^2 exceeded 50% [16]. Results were as follows:

1. AI model for carotid plaques: Both ML models based on radiomics algorithms and DL models exhibited high sensitivity, with an I^2 of 90.20% and 93.70%, and high specificity, with an I^2 of 78.92% and 95.55%, suggesting high performance and significant heterogeneity (Multimedia Appendix 3 [36-69]).
2. Medical imaging modalities: the sensitivity and specificity for PRs (sensitivity I^2=82.28%; specificity I^2=79.16%; Multimedia Appendix 5 [36-69]) and ultrasound (sensitivity I^2=96.92%; specificity I^2=94.98%; Multimedia Appendix 5 [36-69]). The sensitivity and specificity for MRI (sensitivity I^2=71.57%; specificity I^2=73.21%; Multimedia Appendix 5 [36-69]) and the sensitivity for CTA (I^2=56.80%) displayed moderate heterogeneity (Multimedia Appendix 5 [36-69]). The specificity of CTA (I^2=83.79%) was high (Multimedia Appendix 5 [36-69]). In the ultrasound modality, the sensitivity and specificity for determining the presence of plaques (sensitivity I^2=96.78%; specificity I^2=97.97%; Multimedia Appendix 5 [36-69]) and distinguishing the stability of plaques (sensitivity I^2=97.01%; sensitivity I^2=94.43%; Multimedia Appendix 5 [36-69]) were high.
3. Use of transfer learning: the specificity for models using transfer learning (specificity I^2=74.85%; Multimedia Appendix 6 [36-69]) displayed moderate heterogeneity. The sensitivity for models using transfer learning (sensitivity I^2=79.84%; Multimedia Appendix 6 [36-69]) and the sensitivity and specificity for the models without transfer learning (sensitivity I^2=94.12%; specificity I^2=87.35%; Multimedia Appendix 6 [36-69]) were high.
4. Carotid plaque type: all plaque types showed higher sensitivity and specificity; presence or absence of plaques (sensitivity I^2=94.08%; specificity I^2=97.60%; part A in Multimedia Appendix 7 [36-69]), stable or vulnerable plaques with (sensitivity I^2=95.19%; specificity I^2=91.29%; part B in Multimedia Appendix 7 [36-69]), and symptomatic or asymptomatic plaques (sensitivity I^2=93.28%; specificity I^2=84.67%; part C in Multimedia Appendix 7 [36-69]).
5. Both pure AI models and combined clinical features models did not exhibit high heterogeneity for AI models (sensitivity I^2=62.97%; specificity I^2=2.41%; part B in Multimedia Appendix 8 [50,52,58,63,64,66,68]) and combined models (sensitivity I^2=69.77%; specificity I^2=40.08%) for combined models (part A in Multimedia Appendix 8 [50,52,58,63,64,66,68]).
6. Different sets of datasets: both testing (sensitivity I^2=94.23%; specificity I^2=93.45%; part A in Multimedia Appendix 9 [36-69]) and external validation (specificity I^2=84.42%; part B in Multimedia Appendix 9 [36-69]) were high heterogeneity, except the sensitivity for external validation (I^2=66.67%; part B in Multimedia Appendix 9 [36-69]).

XSL•FO
RenderX

7. Different risk of bias studies: the sensitivity and specificity for high or unclear risk of bias studies (sensitivity I^2=94.61%; specificity I^2=92.59%; part B in Multimedia Appendix 10 [36-69]) and the specificity for low risk of bias studies (I^2=87.10%) were high (part A in Multimedia Appendix 10 [36-69]). The sensitivity for low risk of bias studies (I^2=62.20%) was moderate (part A in Multimedia Appendix 10 [36-69]).
8. Different sample sizes of model: The sensitivity and specificity for sample size ≥200 (sensitivity I^2=97.91%; specificity I^2=97.40%; part A in Multimedia Appendix 11 [36-69]) and the specificity for sample size <200 (I^2=78.02%; part B in Multimedia Appendix 11 [36-69]) were high. The sensitivity for sample size <200 (I^2=60.64%) was moderate (part B in Multimedia Appendix 11 [36-69]).
9. Models with different research designs: The sensitivity and specificity for multicenter studies (sensitivity I^2=81.36%; specificity I^2=80.24%; part A in Multimedia Appendix 12 [36,38,39,41-52,54-69]) and single-center studies (sensitivity I^2=95.07 %; specificity I^2=90.63%) were high (part B in Multimedia Appendix 12 [36,38,39,41-52,54-69]).

The meta-regression did not explore the factors contributing to heterogeneity (parts A-I in Multimedia Appendix 13 [36-69]). The results of all subgroups are depicted in Table S4 in Multimedia Appendix 1. The Fagan nomogram was used to evaluate the diagnostic performance of ML models based on radiomics algorithms and DL models for carotid plaques. The results showed a P-post of 89% and 12% for the positive and negative tests, respectively (Multimedia Appendix 14).

Sensitivity Analysis

Excluding the specific studies did not significantly change our research results (Table S7-S8 in Multimedia Appendix 1).

Quality Assessment

The quality of the 34 studies was evaluated using the QUADAS-AI tool (Multimedia Appendix 15). The QUADAS-AI specifically evaluates bias risk and applicability concerns in AI studies. Here, we observed that most studies had significant bias or applicability concerns, particularly regarding the selection of patients and index test. In the “patient selection” domain, 20 studies were classified as either high-risk or indeterminate due to reliance on closed-access data or failure to present the rationale and breakdown of its training, validation, and test sets. Only 7 externally validated studies were classified as low-risk in the “index test” category, while others showed elevated risks due to a lack of validation. In the “reference standard” assessment, the reference standard of all studies could be used to classify the target condition correctly. For the “flow and timing” assessment, 10 studies showed indeterminate risks due to insufficient justification for the timing between index and reference tests. Additionally, 20 studies presented significant concerns regarding applicability in the “patient selection” domain, receiving unclear ratings. In the “index test” domain, 7 studies were rated as having low applicability, while all studies received low applicability ratings in the “Reference Standard” domain.

Discussion

Principal Findings

This study represents the first systematic evaluation of ML models based on radiomics and DL models for the characterization of extracranial carotid plaques. Both approaches demonstrated robust diagnostic performance, with high SROC values of 0.95 and 0.92, respectively, highlighting their promising potential for clinical application in plaque detection and risk stratification.

Initially, the SP and SROC AUC of DL models were improved compared to ML models based on radiomics (0.91 vs 0.83; 0.95 vs 0.92), while their sensitivity was similar to that of ML (0.88). Moreover, we observed that radiomics and DL models used to identify the presence of plaques and stable plaques had similar diagnostic capabilities (SROC 0.96, 95% CI 0.94 - 0.97), and both were effective in identifying symptomatic plaques (SROC 0.90, 95% CI 0.87 - 0.92). Notably, these differences may not be simply due to model performance, but could result from a combination of different clinical objectives (simple exclusion diagnosis or differentiation of specific cases), imaging variations, and model techniques. By using knowledge gained from previous tasks, transfer learning enhances model performance on new datasets and minimizes data requirements. It has been successfully applied in various areas of cardiovascular disease to boost the performance of models [2,76,77]. In subgroup analyses, transfer learning significantly enhances model performance in data-limited scenarios and prevents overfitting. Large sample sizes can minimize sampling bias, decrease overfitting, and enhance the stability and reproducibility of the models. Moreover, we performed more detailed subgroup analyses based on the same imaging modality. Only the type of plaques in the ultrasound modality had sufficient data to perform statistical analysis and obtain summary diagnostic efficacy indicators. Results showed that ultrasound-based models have demonstrated excellent and similar performance in detecting the presence of plaques and assessing their stability. Considering the differences in equipment characteristics, patient demographics, and study design, these findings should be interpreted with caution. Nevertheless, these results provide valuable insights into the efficacy of radiomics algorithms and DL models in the diagnosis of carotid plaque.

Analysis of the Main Aspects

This meta-analysis demonstrates that radiomics-based models and DL models can diagnose extracranial carotid plaque, but the advantages of DL models in specificity and SROC should be interpreted with caution. A review of the included studies revealed that, among the 24 investigations using DL models, 20 primarily focused on plaque characterization (11 on the detection of plaques and 9 on plaque stability). Of these, 13 studies used ultrasound imaging to identify plaque-specific features such as echogenicity, morphology, and composition. In contrast, among the 10 studies using radiomics-based ML models, 6 were dedicated to identifying symptomatic plaques, predominantly using MRI (n=2) and CTA (n=3). The accuracy of symptomatic plaque identification was influenced not only

XSL•FO
RenderX

by intrinsic imaging characteristics but also by clinical indicators, including plaque rupture, thrombus formation, and the occurrence of cerebral hypoperfusion. The tasks were more complex, and model training seemed to focus on reducing false negatives to lower the risk of adverse outcomes such as stroke. In addition, traditional ML algorithms may rely on manual preprocessing and struggle to capture other subtle differences (such as the presence of tiny thrombi or fibrous cap thickness), which may introduce variability and additional costs. In contrast, the DL models (particularly convolutional neural networks) do not rely on artificially designed features; instead, they can directly process raw medical images, automatically filter noise, and automatically extract more meaningful image features (eg, slight echo attenuation behind plaques, differences in vascular wall elasticity, etc) [78]. It can also analyze the preset artificial extraction features, conduct independent learning, and uncover potential rules, thereby addressing the aforementioned challenges [23,79]. It is worth noting that a mismatch in the number of studies may also affect the interpretation of the results. Therefore, these differences may not be simply due to model performance, but could also be caused by multiple factors, which need to be further investigated.

Besides, the "black box" nature of AI algorithms, particularly DL models, raises concerns about the transparency and reliability of decision-making. Of the 34 studies reviewed, only 2 used explainable DL models, achieving an accuracy of 98.2% [37,65]. The explainable AI (XAI) approach leverages visualization techniques, feature attribution analysis, and both global and local explanations to clarify how models derive predictions from input data. By enhancing transparency, XAI fosters greater trust among medical professionals, strengthens model reliability and accountability, and helps mitigate concerns related to opaque decision-making [80]. The integration of XAI in medicine not only represents a technological advancement but also ensures safe, efficient, and robust medical decision-making, which needs to be further investigated. To realize this potential, a clinically oriented XAI implementation framework needs to be developed. First, the reporting criteria for interpretable techniques (including clinical applicability evaluation and operational guidelines) should be standardized to lower the threshold for physician use. Second, the design of algorithms should be optimized through collaborative efforts of medical professionals and engineers to improve the specificity of feature attribution methods based on real clinical needs. Further clinical validation studies are needed to evaluate the practical utility of XAI across diverse diagnostic settings—such as varying regions, hospital levels, and clinician experience—and to determine its true value in supporting clinical decision-making beyond algorithmic performance [28]. Furthermore, incomplete disclosure of model development processes in reports, selective presentation of results by investigators, and heterogeneity in diagnostic standard implementation across practitioners with different levels of experience may decrease the reliability and generalizability of findings. Therefore, we recommend the formulation of standardized imaging protocols, reporting procedures, and quality control measures for carotid plaque assessment and advocate for the establishment of specialized AI reporting guidelines for cardiovascular diseases.

Advances in imaging technology have now largely met the diagnostic requirements of current clinical practice, and current guidelines place heavy reliance on imaging tests for carotid plaque assessment. Among the 34 included studies, 27 constructed diagnostic models based only on imaging data. However, this should not be interpreted as rendering other clinical parameters irrelevant. Multidimensional diagnostic models combined with clinical features have been shown to achieve good diagnostic performance in identifying various diseases, such as pancreatic ductal adenocarcinoma [81], HCC recurrence after liver transplantation [82], hemorrhagic brain metastases [83], malignant BI-RADS 4 breast masses [84], and others. In our study, the diagnostic performance of combined models did not slightly improve, which may be due to the small sample size or some features could not provide more diagnostic information (for example, Hu et al [2] constructed a model relying only on indirect perivascular adipose tissue radiomic features and clinical features to identify symptomatic plaques, lacking direct imaging features). Considering this evidence, we strongly recommend that future research should aim to not only systematically incorporate laboratory tests, medical history, and other clinical parameters to develop multidimensional diagnostic models, but also to summarize the most meaningful features for specific types of plaques. This could address the limitations in current studies regarding single imaging modalities. This will also improve the precise classification of carotid plaques and personalized risk assessment.

This meta-analysis identified significant heterogeneity, while meta-regression and subgroup regression analysis did not identify the source, primarily attributable to the intrinsic challenges in regulating all potential confounding factors. Different imaging techniques can affect model performance based on the type of images used (static images vs dynamic videos), the equipment, and the operators. Guang et al [57] used a contrast-enhanced ultrasound video-based DL model to evaluate the diagnostic efficacy of a new carotid network structure for assessing carotid plaques, whereas other ultrasound studies consistently used static images. The sequence of MRI scans also influences diagnostic outcomes. Zhang et al [58] reported that a model incorporating a combination of T1-weighted, T2-weighted, dynamic contrast-enhanced, and postcontrast (POST) MRI sequences achieved a higher AUC for identifying high-risk carotid plaques compared to models using individual sequences or partial combinations. This enhanced performance is attributed to the complementary nature of these imaging sequences, each capturing distinct pathophysiological characteristics of the plaque, thereby improving diagnostic accuracy when used in combination. PRs have limited resolution, only detecting calcified components of carotid plaques and missing features such as lipid-rich necrotic cores or thin or ruptured fibrous caps. There are also notable differences in model architecture. Yoo et al [39] found performance variations among different convolutional neural network architectures within the CACSNet framework on the same dataset. Gui et al [49] compared multiple DL models (eg, 3D-DenseNet, 3D-SE-DenseNet) with 9 ML algorithms (including Decision Tree, Random Forest, SVM, etc) using identical datasets. They found that DL models generally performed better across key metrics like AUC and accuracy,

XSL•FO
RenderX

with significant performance differences between and within the two model types. These suggest that scanning parameters, model architectures, image segmentation, and algorithms may explain the heterogeneity in the research results. However, the small number of studies limits our ability to perform comprehensive subgroup analyses, which need to be further investigated.

The use of AI has significantly promoted the diagnosis of carotid plaque; however, its application requires cautious evaluation. Only 9 studies were multicenter (most used external validation), with diagnostic performance lower than single-center studies. Most studies (n=29) had a high risk of bias due to a lack of open-source data and external validation and failure to present the rationale and breakdown of its sets, which led to overestimation of the research results and affected the reproducibility and generalizability of the findings. Similar issues have been noted in previous reports, highlighting a broader deficiency in rigorous research standards within the field [85-87]. Furthermore, the contingency tables mostly come from the testing sets. Although the testing set achieved the best diagnostic performance, it had higher data quality or similar data distribution to the training, or overfitting noise, resulting in inaccurate performance estimation, and strong regularization may also decrease its performance, ultimately undermining clinical confidence in these models.

This study has certain clinical significance. We conducted an in-depth literature review and methodological quality evaluation, presenting the most current and comprehensive systematic review of AI-based diagnostic approaches for assessing carotid plaque. The findings reveal that AI technology shows considerable potential for diagnosing carotid plaque, but the findings need to be further validated by conducting more rigorous external validation using large-scale, high-quality independent datasets.

Limitations

This study has several limitations. First, the heterogeneity in model architectures and validation methods across studies prevents definitive conclusions regarding the most effective AI approaches. Second, many studies lack multicenter external validation, leading to a high risk of bias. The model overfitting and clinical applicability need to be carefully evaluated. Third, meta-regression and subgroup analysis did not identify the sources of high heterogeneity that existed in most of the included studies. We hypothesize that this heterogeneity may be caused by scanning parameters, model architectures, image segmentation, and algorithms. However, the overly scattered distribution of subgroups due to the limited number of studies restricts more in-depth subgroup analyses. Finally, although the Deeks test did not show significant publication bias, the included studies may have intentionally unreported negative results and omitted potentially relevant non-English literature.

Future studies should use a more comprehensive analytical methodology based on the current model. Researchers should strictly follow regulatory norms and standardized operating procedures. Prospective and multicenter studies and additional external validation are warranted to enhance the robustness and generalizability of the existing models. In the future, researchers should perform independent systematic reviews on specific subtopics—such as imaging modalities, lesion types, or model architectures—to facilitate targeted evaluations of AI performance across distinct clinical scenarios. In addition, studies on imaging modalities such as CT and MRI are advocated to generate more data, conduct subgroup analyses, and clarify the optimal matching of modality, plaque type, and algorithm. Future efforts should focus on identifying more meaningful features and building and evaluating the diagnostic performance of multidimensional diagnostic models. In parallel, establishing clinically oriented, XAI frameworks will be essential for enhancing transparency.

Conclusions

Current findings indicate that radiomics algorithms and DL models can effectively diagnose extracranial carotid plaque. However, the irregularities in research design and the lack of multicenter studies and external validation limit the robustness of the present findings. Future research should aim to reduce bias risk and enhance the generalizability and clinical orientation of the models.

Acknowledgments

The manuscript was written without the use of ChatGPT or other generative language models.

Funding

The conduct of this study, the writing of the manuscript, and its publication did not receive any external financial support or grants from any public, commercial, or nonprofit entities.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: LJ (lead), JR (equal)
Methodology: LJ (lead), YG (equal), HG (supporting)
Data curation: LJ (lead), YG (supporting)
Investigation: LJ (lead), YG (supporting), HG (supporting)

XSL•FO
RenderX

Software: LJ (lead), RL (equal), YW (supporting)
Supervision: JR (lead), NM (supporting)
Validation: LJ (lead), RL (supporting), YW (supporting)
Visualization: LJ (lead), YG (supporting), HG (supporting)
Writing – original draft: LJ (lead), YG (supporting)
Writing – review & editing: LJ (lead), RL (supporting), YW (supporting), SW (supporting), NM (supporting), JR (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete supplementary data tables for the systematic review and meta-analysis.
[DOC File, 259 KB - jmir_v28i1e77092_app1.doc]

Multimedia Appendix 2
Combined diagnostic performance estimates from the included studies (34 studies with 34 tables) [36-69].
[PNG File, 381 KB - jmir_v28i1e77092_app2.png]

Multimedia Appendix 3
Sensitivity and specificity of deep learning (DL) versus radiomics-based machine learning (ML) models. (A) DL models (24 studies with 24 tables). (B) ML models based on radiomics algorithms (10 studies with 10 tables) [36-69].
[JPG File, 601 KB - jmir_v28i1e77092_app3.jpg]

Multimedia Appendix 4
Publication bias.
[PNG File, 73 KB - jmir_v28i1e77092_app4.png]

Multimedia Appendix 5
Sensitivity and specificity for different medical imaging modalities. (A) Models based on periapical radiographs (PRs) imaging (5 studies with 5 tables). (B) Models based on ultrasound imaging (16 studies with 16 tables). (C) Models based on magnetic resonance imaging (MRI) imaging (5 studies with 5 tables). (D) Models based on computed tomography angiography (CTA) imaging (8 studies with 8 tables). (E) Models based on ultrasound modality for detecting the presence of carotid plaque (5 studies with 5 tables). (F) Models based on ultrasound modality for distinguishing the stability of carotid plaques (8 studies with 8 tables) [36-69].
[JPG File, 879 KB - jmir_v28i1e77092_app5.jpg]

Multimedia Appendix 6
Sensitivity and specificity of models for using transfer learning or not. (A) Models using transfer learning (10 studies with 10 tables). (B) Models without transfer learning (24 studies with 24 tables) [36-69].
[JPG File, 584 KB - jmir_v28i1e77092_app6.jpg]

Multimedia Appendix 7
Sensitivity and specificity for different diagnostic tasks. (A) Presence or absence of carotid plaques (11 studies with 11 tables). (B) Stable or vulnerable carotid plaques (12 studies with 12 tables). (C) Symptomatic or asymptomatic carotid plaques (10 studies with 11 tables) [36-69].
[JPG File, 743 KB - jmir_v28i1e77092_app7.jpg]

Multimedia Appendix 8
Sensitivity and specificity of pure artificial intelligence models or models constructed by combining clinical features for carotid plaques. (A) Combined models (7 studies with 7 tables). (B) Artificial intelligence models (7 studies with 7 tables) [50,52,58,63,64,66,68].
[JPG File, 416 KB - jmir_v28i1e77092_app8.jpg]

Multimedia Appendix 9
Sensitivity and specificity for different dataset types. (A) Testing (27 studies with 27 tables). (B) External validation (7 studies with 7 tables) [36-69].

XSL•FO
RenderX

[JPG File, 617 KB - jmir_v28i1e77092_app9.jpg]

Multimedia Appendix 10
Sensitivity and specificity for different risk of-bias studies. (A) Low risk of bias studies (5 studies with 5 tables). (B) High/unclear risk of bias studies (29 studies with 29 tables) [36-69].
[JPG File, 631 KB - jmir_v28i1e77092_app10.jpg]

Multimedia Appendix 11
Sensitivity and specificity of study using different sample sizes. (A) Sample size ≥200 (14 studies with 14 tables). (B) Sample size <200 (20 studies with 20 tables) [36-69].
[JPG File, 583 KB - jmir_v28i1e77092_app11.jpg]

Multimedia Appendix 12
Sensitivity and specificity for different research designs. (A) Multicenter studies (9 studies with 9 tables). (B) Single-center studies (22 studies with 22 tables) [36,38,39,41-52,54-69].
[JPG File, 578 KB - jmir_v28i1e77092_app12.jpg]

Multimedia Appendix 13
Exploration of potential sources of heterogeneity across multiple variables. (A) Different algorithms (deep learning models or machine learning models based radiomics algorithms). (B) Different medical imaging modalities. (C) Different carotid plaque type. (D) Utilizing transfer learning or not. (E) Different sets of datasets. (F) Different sample size. (G) Single-center or multicenter studies. (H) Pure artificial intelligence (AI) models or models constructed by combining clinical features. (I) Different risk of bias studies [36-69].
[JPG File, 1753 KB - jmir_v28i1e77092_app13.jpg]

Multimedia Appendix 14
The Fagan nomogram assesses the diagnostic ability of radiomics and deep learning to carotid plaques.
[PNG File, 186 KB - jmir_v28i1e77092_app14.png]

Multimedia Appendix 15
Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence (QUADAS-AI) for the assessment of the methodological qualities of all the enrolled studies.
[PNG File, 240 KB - jmir_v28i1e77092_app15.png]

Checklist 1
PRISMA 2020 checklist.
[PDF File, 99 KB - jmir_v28i1e77092_app16.pdf]

References

1. Song P, Fang Z, Wang H, et al. Global and regional prevalence, burden, and risk factors for carotid atherosclerosis: a systematic review, meta-analysis, and modelling study. Lancet Glob Health 2020 May;8(5):e721-e729. [doi: 10.1016/S2214-109X(20)30117-0] [Medline: 32353319]
2. Evans NR, Bhakta S, Chowdhury MM, Markus H, Warburton E. Management of carotid atherosclerosis in stroke. Pract Neurol 2024 Sep 13;24(5):382-386. [doi: 10.1136/pn-2023-003918] [Medline: 38589215]
3. Oushy SH, Essibayi MA, Savastano LE, Lanzino G. Carotid artery revascularization: endarterectomy versus endovascular therapy. J Neurosurg Sci 2021 Jun;65(3):322-326. [doi: 10.23736/S0390-5616.20.05207-8] [Medline: 33297612]
4. Kopczak A, Schindler A, Bayer-Karpinska A, et al. Complicated carotid artery plaques as a cause of cryptogenic stroke. J Am Coll Cardiol 2020 Nov 10;76(19):2212-2222. [doi: 10.1016/j.jacc.2020.09.532] [Medline: 33153580]
5. Abbott AL, Paraskevas KI, Kakkos SK, et al. Systematic review of guidelines for the management of asymptomatic and symptomatic carotid stenosisa. Stroke 2015 Nov;46(11):3288-3301. [doi: 10.1161/STROKEAHA.115.003390] [Medline: 26451020]
6. Brinjikji W, Huston J, Rabinstein AA, Kim GM, Lerman A, Lanzino G. Contemporary carotid imaging: from degree of stenosis to plaque vulnerability. JNS 2016 Jan;124(1):27-42. [doi: 10.3171/2015.1.JNS142452]
7. Paju S, Pietiäinen M, Liljestrand JM, et al. Carotid artery calcification in panoramic radiographs associates with oral infections and mortality. Int Endodontic J 2021 Apr;54(4):638-638. [doi: 10.1111/iej.13451]

XSL•FO
RenderX

8. Bengtsson VW, Persson GR, Renvert S. Assessment of carotid calcifications on panoramic radiographs in relation to other used methods and relationship to periodontitis and stroke: a literature review. Acta Odontol Scand 2014 Aug;72(6):401-412. [doi: 10.3109/00016357.2013.847489] [Medline: 24432815]
9. Schroder AGD, de Araujo CM, Guariza-Filho O, Flores-Mir C, de Luca Canto G, Porporatti AL. Diagnostic accuracy of panoramic radiography in the detection of calcified carotid artery atheroma: a meta-analysis. Clin Oral Investig 2019 May;23(5):2021-2040. [doi: 10.1007/s00784-019-02880-6] [Medline: 30923911]
10. Zeng P, Zhang Q, Liang X, Zhang M, Luo D, Chen Z. Progress of ultrasound techniques in the evaluation of carotid vulnerable plaque neovascularization. Cerebrovasc Dis 2024;53(4):479-487. [doi: 10.1159/000534372] [Medline: 37812915]
11. Guo Y, Wang X, Wang L, et al. The value of superb microvascular imaging and contrast-enhanced ultrasound for the evaluation of neovascularization in carotid artery plaques. Acad Radiol 2023 Mar;30(3):403-411. [doi: 10.1016/j.acra.2022.08.001] [Medline: 36123231]
12. Chen X, Wang H, Jiang Y, et al. Neovascularization in carotid atherosclerotic plaques can be effectively evaluated by superb microvascular imaging (SMI): Initial experience. Vasc Med 2020 Aug;25(4):328-333. [doi: 10.1177/1358863X20909992] [Medline: 32303154]
13. Li C, He W, Guo D, et al. Quantification of carotid plaque neovascularization using contrast-enhanced ultrasound with histopathologic validation. Ultrasound Med Biol 2014 Aug;40(8):1827-1833. [doi: 10.1016/j.ultrasmedbio.2014.02.010] [Medline: 24798387]
14. Hoogi A, Adam D, Hoffman A, Kerner H, Reisner S, Gaitini D. Carotid plaque vulnerability: quantification of neovascularization on contrast-enhanced ultrasound with histopathologic correlation. AJR Am J Roentgenol 2011 Feb;196(2):431-436. [doi: 10.2214/AJR.10.4522] [Medline: 21257897]
15. Zamani M, Skagen K, Scott H, Russell D, Skjelland M. Advanced ultrasound methods in assessment of carotid plaque instability: a prospective multimodal study. BMC Neurol 2020 Jan 29;20(1):39. [doi: 10.1186/s12883-020-1620-z] [Medline: 31996153]
16. Saba L, Caddeo G, Sanfilippo R, Montisci R, Mallarini G. Efficacy and sensitivity of axial scans and different reconstruction methods in the study of the ulcerated carotid plaque using multidetector-row CT angiography: comparison with surgical results. AJNR Am J Neuroradiol 2007 Apr;28(4):716-723. [Medline: 17416828]
17. Horev A, Honig A, Cohen JE, et al. Overestimation of carotid stenosis on CTA - real world experience. J Clin Neurosci 2021 Mar;85:36-40. [doi: 10.1016/j.jocn.2020.12.018] [Medline: 33581787]
18. Baradaran H, Gupta A. Carotid vessel wall imaging on CTA. AJNR Am J Neuroradiol 2020 Mar;41(3):380-386. [doi: 10.3174/ajnr.A6403] [Medline: 32029468]
19. Saba L, Yuan C, Hatsukami TS, et al. Carotid artery wall imaging: perspective and guidelines from the ASNR vessel wall imaging study group and expert consensus recommendations of the American Society of Neuroradiology. AJNR Am J Neuroradiol 2018 Feb;39(2):E9-E31. [doi: 10.3174/ajnr.A5488] [Medline: 29326139]
20. Xu HL, Gong TT, Song XJ, et al. Artificial intelligence performance in image-based cancer identification: umbrella review of systematic reviews. J Med Internet Res 2025 Apr 1;27:e53567. [doi: 10.2196/53567] [Medline: 40167239]
21. Scicolone R, Vacca S, Pisu F, et al. Radiomics and artificial intelligence: general notions and applications in the carotid vulnerable plaque. Eur J Radiol 2024 Jul;176:111497. [doi: 10.1016/j.ejrad.2024.111497] [Medline: 38749095]
22. Koçak B. Key concepts, common pitfalls, and best practices in artificial intelligence and machine learning: focus on radiomics. Diagn Interv Radiol 2022 Sep;28(5):450-462. [doi: 10.5152/dir.2022.211297] [Medline: 36218149]
23. Zhou T, Cheng Q, Lu H, Li Q, Zhang X, Qiu S. Deep learning methods for medical image fusion: a review. Comput Biol Med 2023 Jun;160:106959. [doi: 10.1016/j.compbiomed.2023.106959] [Medline: 37141652]
24. Ma Y, Li M, Wu H. The machine learning models in major cardiovascular adverse events prediction based on coronary computed tomography angiography: systematic review. J Med Internet Res 2025 Jun 13;27:e68872. [doi: 10.2196/68872] [Medline: 40513092]
25. Oikonomou EK, Williams MC, Kotanidis CP, et al. A novel machine learning-derived radiotranscriptomic signature of perivascular fat improves cardiac risk prediction using coronary CT angiography. Eur Heart J 2019 Nov 14;40(43):3529-3543. [doi: 10.1093/eurheartj/ehz592] [Medline: 31504423]
26. Pan J, Huang Q, Zhu J, et al. Prediction of plaque progression using different machine learning models of pericoronary adipose tissue radiomics based on coronary computed tomography angiography. Eur J Radiol Open 2025 Jun;14:100638. [doi: 10.1016/j.ejro.2025.100638] [Medline: 40034660]
27. Cohen JF, Deeks JJ, Hooft L, et al. Preferred reporting items for journal and conference abstracts of systematic reviews and meta-analyses of diagnostic test accuracy studies (PRISMA-DTA for Abstracts): checklist, explanation, and elaboration. BMJ 2021 Mar 15;372:n265. [doi: 10.1136/bmj.n265] [Medline: 33722791]
28. Bayor AA, Li J, Yang IA, Varnfield M. Designing Clinical Decision Support Systems (CDSS)-A user-centered lens of the design characteristics, challenges, and implications: systematic review. J Med Internet Res 2025 Jun 20;27:e63733. [doi: 10.2196/63733] [Medline: 40540451]
29. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: the PRISMA-DTA statement. JAMA 2018 Jan 23;319(4):388-396. [doi: 10.1001/jama.2017.19163] [Medline: 29362800]

XSL•FO
RenderX

30. Lee J, Mulder F, Leeflang M, Wolff R, Whiting P, Bossuyt PM. QUAPAS: an adaptation of the QUADAS-2 tool to assess prognostic accuracy studies. Ann Intern Med 2022 Jul;175(7):1010-1018. [doi: 10.7326/M22-0276] [Medline: 35696685]
31. Sounderajah V, Ashrafian H, Rose S, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nat Med 2021 Oct;27(10):1663-1665. [doi: 10.1038/s41591-021-01517-0] [Medline: 34635854]
32. Guni A, Sounderajah V, Whiting P, Bossuyt P, Darzi A, Ashrafian H. Revised tool for the Quality Assessment of Diagnostic Accuracy Studies Using AI (QUADAS-AI): protocol for a qualitative study. JMIR Res Protoc 2024 Sep 18;13:e58202. [doi: 10.2196/58202] [Medline: 39293047]
33. Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. NPJ Digit Med 2020;3:126. [doi: 10.1038/s41746-020-00333-z] [Medline: 33043150]
34. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005 Oct;58(10):982-990. [doi: 10.1016/j.jclinepi.2005.02.022] [Medline: 16168343]
35. Samawi H, Alsharman M, Keko M, Kersey J. Post-test diagnostic accuracy measures under tree ordering of disease classes. Stat Med 2023 Dec 10;42(28):5135-5159. [doi: 10.1002/sim.9905] [Medline: 37720999]
36. Molinari F, Raghavendra U, Gudigar A, Meiburger KM, Rajendra Acharya U. An efficient data mining framework for the characterization of symptomatic and asymptomatic carotid plaque using bidimensional empirical mode decomposition technique. Med Biol Eng Comput 2018 Sep;56(9):1579-1593. [doi: 10.1007/s11517-018-1792-5] [Medline: 29473126]
37. Singh S, Jain PK, Sharma N, Pohit M, Roy S. Atherosclerotic plaque classification in carotid ultrasound images using machine learning and explainable deep learning. Intelligent Medicine 2024 May;4(2):83-95. [doi: 10.1016/j.imed.2023.05.003]
38. Li J, Huang Y, Song S, et al. Automatic diagnosis of carotid atherosclerosis using a portable freehand 3-D ultrasound imaging system. IEEE Trans Ultrason Ferroelectr Freq Control 2024 Feb;71(2):266-279. [doi: 10.1109/TUFFC.2023.3345740] [Medline: 38127609]
39. Yoo SW, Yang S, Kim JE, et al. CACSNet for automatic robust classification and segmentation of carotid artery calcification on panoramic radiographs using a cascaded deep learning network. Sci Rep 2024 Jun 17;14(1):13894. [doi: 10.1038/s41598-024-64265-4] [Medline: 38886356]
40. Omarov M, Zhang L, Jorshery SD, et al. Automated deep learning-based detection of early atherosclerotic plaques in carotid ultrasound imaging. medRxiv 2025 Sep 3:2024.10.17.24315675. [doi: 10.1101/2024.10.17.24315675] [Medline: 39484270]
41. Zhai D, Liu R, Liu Y, et al. Deep learning-based fully automatic screening of carotid artery plaques in computed tomography angiography: a multicenter study. Clin Radiol 2024 Aug;79(8):e994-e1002. [doi: 10.1016/j.crad.2024.04.015] [Medline: 38789330]
42. Vinayahalingam S, van Nistelrooij N, Xi T, et al. Detection of carotid plaques on panoramic radiographs using deep learning. J Dent 2024 Dec;151:105432. [doi: 10.1016/j.jdent.2024.105432] [Medline: 39461583]
43. Pisu F, Williamson BJ, Nardi V, et al. Machine learning detects symptomatic plaques in patients with carotid atherosclerosis on CT angiography. Circ Cardiovasc Imaging 2024 Jun;17(6):e016274. [doi: 10.1161/CIRCIMAGING.123.016274] [Medline: 38889214]
44. Zhou R, Gan W, Wang F, Yang Z, Huang Z, Gan H. Tri-correcting: label noise correction via triple CNN ensemble for carotid plaque ultrasound image classification. Biomed Signal Process Control 2024 May;91:105981. [doi: 10.1016/j.bspc.2024.105981]
45. Wang Y, Cai C, Du YM, et al. Assessment of stroke risk using MRI-VPD with automatic segmentation of carotid plaques and classification of plaque properties based on deep learning. J Radiat Res Appl Sci 2023 Sep;16(3):100630. [doi: 10.1016/j.jrras.2023.100630]
46. Shan D, Wang S, Wang J, et al. Computed tomography angiography-based radiomics model for predicting carotid atherosclerotic plaque vulnerability. Front Neurol 2023;14:1151326. [doi: 10.3389/fneur.2023.1151326] [Medline: 37396779]
47. Xie J, Li Y, Xu X, et al. CPTV: classification by tracking of carotid plaque in ultrasound videos. Comput Med Imaging Graph 2023 Mar;104:102175. [doi: 10.1016/j.compmedimag.2022.102175] [Medline: 36630795]
48. Amitay M, Barnett-Itzhaki Z, Sudri S, et al. Deep convolution neural network for screening carotid calcification in dental panoramic radiographs. PLOS Digit Health 2023 Apr;2(4):e0000081. [doi: 10.1371/journal.pdig.0000081] [Medline: 37043433]
49. Gui C, Cao C, Zhang X, Zhang J, Ni G, Ming D. Radiomics and artificial neural networks modelling for identification of high-risk carotid plaques. Front Cardiovasc Med 2023;10:1173769. [doi: 10.3389/fcvm.2023.1173769] [Medline: 37485276]
50. Shi J, Sun Y, Hou J, et al. Radiomics signatures of carotid plaque on computed tomography angiography: an approach to identify symptomatic plaques. Clin Neuroradiol 2023 Dec;33(4):931-941. [doi: 10.1007/s00062-023-01289-9] [Medline: 37195452]
51. Su SS, Li LY, Wang Y, Li YZ. Stroke risk prediction by color Doppler ultrasound of carotid artery-based deep learning using Inception V3 and VGG-16. Front Neurol 2023;14:1111906. [doi: 10.3389/fneur.2023.1111906] [Medline: 36864909]
52. Chen S, Liu C, Chen X, Liu WV, Ma L, Zha Y. A radiomics approach to assess high risk carotid plaques: a non-invasive imaging biomarker, retrospective study. Front Neurol 2022;13:35350403. [doi: 10.3389/fneur.2022.788652]

XSL•FO
RenderX

53. Gago L, Vila MDM, Grau M, Remeseiro B, Igual L. An end-to-end framework for intima media measurement and atherosclerotic plaque detection in the carotid artery. Comput Methods Programs Biomed 2022 Aug;223:106954. [doi: 10.1016/j.cmpb.2022.106954] [Medline: 35777216]
54. Jain PK, Sharma N, Saba L, et al. Automated deep learning-based paradigm for high-risk plaque detection in B-mode common carotid ultrasound scans: an asymptomatic Japanese cohort study. Int Angiol 2022 Feb;41(1):9-23. [doi: 10.23736/S0392-9590.21.04771-4] [Medline: 34825801]
55. Cilla S, Macchia G, Lenkowicz J, et al. CT angiography-based radiomics as a tool for carotid plaque characterization: a pilot study. Radiol Med 2022 Jul;127(7):743-753. [doi: 10.1007/s11547-022-01505-5] [Medline: 35680773]
56. Xu X, Huang L, Wu R, et al. Multi-feature fusion method for identifying carotid artery vulnerable plaque. IRBM 2022 Aug;43(4):272-278. [doi: 10.1016/j.irbm.2021.07.004]
57. Guang Y, He W, Ning B, et al. Deep learning-based carotid plaque vulnerability classification with multicentre contrast-enhanced ultrasound video: a comparative diagnostic study. BMJ Open 2021 Aug 27;11(8):e047528. [doi: 10.1136/bmjopen-2020-047528] [Medline: 34452961]
58. Zhang R, Zhang Q, Ji A, et al. Identification of high-risk carotid plaque with MRI-based radiomics and machine learning. Eur Radiol 2021 May;31(5):3116-3126. [doi: 10.1007/s00330-020-07361-z] [Medline: 33068185]
59. Ma W, Cheng X, Xu X, et al. Multilevel strip pooling-based convolutional neural network for the classification of carotid plaque echogenicity. Comput Math Methods Med 2021;2021:3425893. [doi: 10.1155/2021/3425893] [Medline: 34457035]
60. Ganitidis T, Athanasiou M, Dalakleidi K, Melanitis N, Golemati S, Nikita KS. Stratification of carotid atheromatous plaque using interpretable deep learning methods on B-mode ultrasound images. Annu Int Conf IEEE Eng Med Biol Soc 2021 Nov;2021:3902-3905. [doi: 10.1109/EMBC46164.2021.9630402] [Medline: 34892085]
61. Kats L, Vered M, Zlotogorski-Hurvitz A, Harpaz I. Atherosclerotic carotid plaque on panoramic radiographs: neural network detection. Int J Comput Dent 2019;22(2):163-169. [Medline: 31134222]
62. Wei Y, Yang B, Wei L, et al. Real-time carotid plaque recognition from dynamic ultrasound videos based on artificial neural network. Ultraschall Med 2024 Oct;45(5):493-500. [doi: 10.1055/a-2180-8405] [Medline: 38113893]
63. Zhao T, Lin G, Chen W, et al. Predicting symptomatic carotid artery plaques with radiomics-based carotid perivascular adipose tissue characteristics: a multicenter, multiclassifier study. BMC Med Imaging 2025 Aug 19;25(1):337. [doi: 10.1186/s12880-025-01876-x] [Medline: 40830841]
64. Hu W, Lin G, Chen W, et al. Radiomics based on dual-energy CT virtual monoenergetic images to identify symptomatic carotid plaques: a multicenter study. Sci Rep 2025 Mar 26;15(1):10415. [doi: 10.1038/s41598-025-92855-3] [Medline: 40140428]
65. Liapi GD, Loizou CP, Griffin M, Pattichis CS, Nicolaides A, Kyriacou E. Transfer learning with class activation maps in compositions driving plaque classification in carotid ultrasound. Front Digit Health 2025;7:1484231. [doi: 10.3389/fdgth.2025.1484231] [Medline: 40704367]
66. Yu F, Li X, Zhang Y, et al. MRI ensemble model of plaque and perivascular adipose tissue as PET-equivalent for identifying carotid atherosclerotic inflammation. EJNMMI Res 2025 Aug 6;15(1):103. [doi: 10.1186/s13550-025-01293-9] [Medline: 40768105]
67. Kuwada C, Mitsuya Y, Fukuda M, et al. Area detection improves the person-based performance of a deep learning system for classifying the presence of carotid artery calcifications on panoramic radiographs. Oral Radiol 2025 Jul 22;2025(1-10). [doi: 10.1007/s11282-025-00843-0] [Medline: 40694246]
68. Lao Q, Zhou R, Wu Y, et al. Predicting vulnerability status of carotid plaques using CTA-based quantitative analysis. J Cardiovasc Pharmacol 2025 Mar 1;85(3):217-224. [doi: 10.1097/FJC.0000000000001664] [Medline: 39739382]
69. He L, Yang Z, Wang Y, et al. A deep learning algorithm to identify carotid plaques and assess their stability. Front Artif Intell 2024;7:1321884. [doi: 10.3389/frai.2024.1321884] [Medline: 38952409]
70. Zhang Y, Gan H, Wang F, et al. A self-supervised fusion network for carotid plaque ultrasound image classification. Math Biosci Eng 2024 Jan 31;21(2):3110-3128. [doi: 10.3934/mbe.2024138] [Medline: 38454721]
71. Ali T, Pathan S, Salvi M, Meiburger KM, Molinari F, Acharya UR. CAROTIDNet: a novel carotid symptomatic/asymptomatic plaque detection system using CNN-based tangent optimization algorithm in B-mode ultrasound images. IEEE Access 2024;12:73970-73979. [doi: 10.1109/ACCESS.2024.3404023]
72. Ayoub M, Liao Z, Li L, Wong KKL. HViT: hybrid vision inspired transformer for the assessment of carotid artery plaque by addressing the cross-modality domain adaptation problem in MRI. Comput Med Imaging Graph 2023 Oct;109:102295. [doi: 10.1016/j.compmedimag.2023.102295] [Medline: 37717365]
73. Latha S, Muthu P, Lai KW, Khalil A, Dhanalakshmi S. Performance analysis of machine learning and deep learning architectures on early stroke detection using carotid artery ultrasound images. Front Aging Neurosci 2021;13:828214. [doi: 10.3389/fnagi.2021.828214] [Medline: 35153728]
74. Wang L, Guo T, Wang L, et al. Improving radiomic modeling for the identification of symptomatic carotid atherosclerotic plaques using deep learning-based 3D super-resolution CT angiography. Heliyon 2024 Apr 30;10(8):e29331. [doi: 10.1016/j.heliyon.2024.e29331] [Medline: 38644848]
75. Li YC, Zhang TR, Zhang F, et al. Development and validation of a carotid plaque risk prediction model for coal miners. Front Cardiovasc Med 2025;12:1490961. [doi: 10.3389/fcvm.2025.1490961] [Medline: 40416817]

XSL•FO
RenderX

76. Weimann K, Conrad TOF. Transfer learning for ECG classification. Sci Rep 2021 Mar 4;11(1):5251. [doi: 10.1038/s41598-021-84374-8] [Medline: 33664343]
77. Chiu IM, Cheng JY, Chen TY, et al. Using deep transfer learning to detect hyperkalemia from ambulatory electrocardiogram monitors in intensive care units: personalized medicine approach. J Med Internet Res 2022 Dec 5;24(12):e41163. [doi: 10.2196/41163] [Medline: 36469396]
78. van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal 2022 Jul;79:102470. [doi: 10.1016/j.media.2022.102470] [Medline: 35576821]
79. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol 2020 Feb 27;9(2):14. [doi: 10.1167/tvst.9.2.14] [Medline: 32704420]
80. de Vries BM, Zwezerijnen GJC, Burchell GL, van Velden FHP, Menke-van der Houven van Oordt CW, Boellaard R. Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. Front Med (Lausanne) 2023;10:1180773. [doi: 10.3389/fmed.2023.1180773] [Medline: 37250654]
81. Huang Y, Zhang H, Ding Q, et al. Comparison of multiple machine learning models for predicting prognosis of pancreatic ductal adenocarcinoma based on contrast-enhanced CT radiomics and clinical features. Front Oncol 2024;14:1419297. [doi: 10.3389/fonc.2024.1419297] [Medline: 39605884]
82. Schindler P, von Beauvais P, Hoffmann E, et al. Combining radiomics and imaging biomarkers with clinical variables for the prediction of HCC recurrence after liver transplantation. Liver Transpl 2025 Oct 1;31(10):1226-1237. [doi: 10.1097/LVT.0000000000000603] [Medline: 40100771]
83. Cui L, Yu L, Shao S, et al. Improving differentiation of hemorrhagic brain metastases from non-neoplastic hematomas using radiomics and clinical feature fusion. Neuroradiology 2025 Jun;67(6):1455-1468. [doi: 10.1007/s00234-025-03590-5] [Medline: 40131431]
84. Zhang Q, Gao J, Agyekum EA, et al. A combined clinical-ultrasound radiomics model for differentiating benign and malignant BI-RADS category 4 breast masses. Am J Transl Res 2025;17(8):6370-6380. [doi: 10.62347/SBKU2090] [Medline: 40950304]
85. Fu Y, Huang Z, Deng X, et al. Artificial intelligence in lymphoma histopathology: systematic review. J Med Internet Res 2025 Feb 14;27:e62851. [doi: 10.2196/62851] [Medline: 39951716]
86. Wu Y, Chao J, Bao M, Zhang N. Predictive value of machine learning on fracture risk in osteoporosis: a systematic review and meta-analysis. BMJ Open 2023 Dec;13(12):e071430. [doi: 10.1136/bmjopen-2022-071430]
87. Mäkitie AA, Alabi RO, Ng SP, et al. Artificial intelligence in head and neck cancer: a systematic review of systematic reviews. Adv Ther 2023 Aug;40(8):3360-3380. [doi: 10.1007/s12325-023-02527-9] [Medline: 37291378]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
BI-RADS: Breast Imaging Reporting and Data System
CTA: computed tomography angiography
DL: deep learning
IEEE: Institute of Electrical and Electronics Engineers
LR: likelihood ratio
ML: machine learning
MRI: magnetic resonance imaging
P-post: posttest probability
PR: periapical radiograph
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy Studies
PROSPERO: International Prospective Register of Systematic Reviews
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2
QUADAS-AI: Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence
ROC: receiver operating characteristic
SROC: summary receiver operating characteristic
SROC AUC: area under the summary receiver operating characteristic curve
XAI: explainable AI

XSL•FO
RenderX

Edited by A Coristine; submitted 07.May.2025; peer-reviewed by MA Ashoobi, R Orozco, Z Fang; accepted 17.Nov.2025; published 22.Jan.2026.

Please cite as:
Ju L, Guo Y, Guo H, Liu R, Wang Y, Wang S, Ma N, Ren J
Diagnostic Performance of Deep Learning and Radiomics in Extracranial Carotid Plaque Detection: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e77092
URL: https://www.jmir.org/2026/1/e77092
doi:10.2196/77092

© Lingjie Ju, Yongsheng Guo, Haiyong Guo, Ruijuan Liu, Yiyang Wang, Siyu Wang, Na Ma, Junhong Ren. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 22.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evidence for Digital Health Tools Designed to Support the Triage of Musculoskeletal Conditions in Primary, Urgent, and Emergency Care Settings: Scoping Review

Linda K Truong[1,2], PT, PhD; James G Wrightson[3], PhD; Raphaël Vincent[4,5,6], PT, MSc; Eunice Lui[1], MPH; Jamon L Couch[7,8], PT; Ellen Wang[2,8], MSc; Cobie Starcevich[9,10], PT; Dean Giustini[11], PhD; Alex Haagaard[12], MA; Elena Lopatina[13,14], MD, PhD; Niels van Berkel[15], PhD; Michael Skovdal Rathleff[16,17], PT, PhD; Clare L Ardern[1,2,7], PT, PhD

[1]Centre for Aging SMART, University of British Columbia, Vancouver, BC, Canada

[10]Physiotherapy Department, Rockingham General Hospital, South Metropolitan Health Service, Perth, Australia

[11]Biomedical Branch Library, University of British Columbia, Vancouver, BC, Canada

[12]Pain BC, Vancouver, BC, Canada

[13]University of Calgary, Calgary, AB, Canada

[14]Primary Care Alberta, Calgary, AB, Canada

[15]Department of Computer Science, Aalborg University, Aalborg, Denmark

[16]Department of Health Science and Technology, Faculty of Medicine, Aalborg University, Aalborg, Denmark

[17]Center for General Practice at Aalborg University, Aalborg, Denmark

[2]Department of Physical Therapy, Faculty of Medicine, University of British Columbia, 13737 96 Avenue, Surrey, BC, Canada

[3]Department of Family Practice, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

[4]School of Rehabilitation, Faculty of Medicine, Université de Montréal, Montreal, QC, Canada

[5]Hôpital Maisonneuve-Rosemont Research Center, Université de Montréal Affiliated Research Center, Montreal, QC, Canada

[6]Centre for Interdisciplinary Research in Rehabilitation of Greater Montreal, Institut Universitaire sur la Réadaptation en Déficience Physique de Montréal, Montreal, QC, Canada

[7]La Trobe Sport and Exercise Medicine Research Centre, La Trobe University, Melbourne, Australia

[8]Arthritis Research Canada, Vancouver, BC, Canada

[9]School of Allied Health, Faculty of Health Sciences, Curtin University, Perth, Australia

Corresponding Author:
Linda K Truong, PT, PhD
Centre for Aging SMART, University of British Columbia, Vancouver, BC, Canada

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2026/1/e92722

Abstract

Background: The digital health research field is growing rapidly, and a summary of the available digital tools for triaging musculoskeletal conditions is needed. Effective and safe digital triage tools for musculoskeletal conditions could support patients and clinicians in making informed care decisions and may contribute to reducing emergency department overcrowding and health care costs.

Objective: The aim of the study is to identify and describe digital health tools for use by adults to triage musculoskeletal conditions across primary, urgent, or emergency care settings.

Methods: Our scoping review was conducted following the Johanna Briggs Institute recommendations for scoping reviews and Arksey and O'Malley's framework. Systematic searches in MEDLINE (OVID), Embase (OVID), CENTRAL (OVID), CINAHL (EBSCO), Compendex, Web of Science, OpenGrey, Google Scholar, arXiv, medRxiv, and an extensive gray literature search were conducted with a librarian scientist from inception to September 18, 2025. Studies had to recruit adults (aged 18 years and older) with musculoskeletal conditions that identified a digital health tool designed to triage or diagnose in primary, urgent, or emergency care settings and report primary data to be included. In total, 2 reviewer pairs independently screened abstracts and full-text papers. Relevant data were extracted in duplicate, and results were summarized descriptively.

XSL•FO
RenderX

Results: The search yielded 5695 records, and we screened 189 full-text papers. In total, 34 studies (n=37,509 patients) met the inclusion criteria. The most common musculoskeletal conditions reported were rheumatoid or inflammatory arthritis (13/34, 38%). In total, 19 (19/34, 56%) studies reported on symptom checkers, 13 (13/34, 38%) studies on triage or diagnosis tools, and 2 (2/34, 6%) were studies of diagnostic predictor tools. There were 16 unique digital health tools. A total of 2 tools were built for triaging musculoskeletal conditions and were not publicly available outside the UK National Health Service. Most tools were generic tools designed to screen for general health problems, including musculoskeletal conditions. The most common approach to evaluating performance (eg, accuracy) of the tools was to compare the concordance of the tool to a clinician diagnosis or triage recommendation. Sensitivity and specificity ranged from 39% to 91% and 23% to 80%, respectively. The reported accuracy of the included tools ranged from 33% to 98%.

Conclusions: Musculoskeletal conditions remain a blind spot for people designing, implementing, and evaluating digital health for triage: few tools were specifically designed for musculoskeletal conditions, and most existing tools performed poorly when applied to musculoskeletal populations. We recommend health systems and clinicians use a multimodal approach, integrating both digital health tools and clinical decision-making to safely triage and diagnose until a more robust tool for musculoskeletal conditions is available. Future tool developers need to use transparent, standardized processes that prioritize tool safety, clinical value, and trustworthiness when designing for clinicians and patients.

(J Med Internet Res 2026;28:e81578) doi:10.2196/81578

KEYWORDS

musculoskeletal; digital health; emergency department; health system; triage; PRISMA; Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Introduction

Musculoskeletal conditions are one of the largest contributors to the global burden of disease and the sixth largest contributor to disability worldwide [1]. The global forecast predicts that the burden of musculoskeletal conditions will more than double in the decades between 2020 and 2050 [1]. Most musculoskeletal conditions can be effectively managed proactively in a primary care setting, not in the emergency department (ED) [2]. Our recent analysis of epidemiological data revealed that approximately 1 in 10 ED visits were related to musculoskeletal issues, and 6 in 10 of these cases could have been appropriately managed outside the ED [3]. This indicates a need to re-evaluate how people with musculoskeletal conditions access health care.

Effective and efficient triage processes are needed to help patients navigate health systems and find timely and high-quality care, avoiding inappropriate use of the ED [4]. The idea of triage was first applied in military settings to help allocate resources and timely care for the wounded [4]. In today's context, triage is often considered in the ED or the first point of contact for clinicians to help prioritize who needs attention first when patients present to the ED [4]. In regard to triaging musculoskeletal conditions, triage is often conducted by tele-triage, paper-based triage, and face-to-face triage [5]. However, patients and musculoskeletal experts have reported that these approaches are inefficient and ineffective in moving patients through the health system [5].

There is an increasing trend toward the use of digital triage, such as online symptom checkers by patients, to make an informed decision on the next and best course of action for their current problem [6-8]. More recently, the World Health Organization has launched a global strategy on digital health to help improve the health and well-being of all humans [9]. This includes defining digital health as "the use of information and communications technology in support of health and health-related fields," which encompasses eHealth, mobile health (mHealth), advanced computer sciences, such as big data and artificial intelligence (AI) or machine learning, and the broad scope of telehealth and telemedicine [10]. Digital health tools have the potential to tackle overcrowding in the ED and primary care settings by guiding patients to alternative services for musculoskeletal care that may be just as effective as the ED.

In the last decade, there has been a shift toward integrating digital health tools, such as symptom checkers, into the health system, as reflected in the volume of reviews evaluating such tools [6,8,11-14]. With the proliferation and widespread adoption of generative AI (eg, large language models [LLMs]), the public seems accepting of using digital health tools like AI to provide guidance and diagnoses for health conditions. This is despite generative AI not being specifically designed for health care use [15]. These findings reflect the growing demand for the integration of digital technology into health care.

Despite the advancement of digital health tools, there are no reviews currently available that have studied digital health tools for diagnosing and triaging musculoskeletal conditions [8,11-14]. Understanding the available tools, including their performance (eg, accuracy), will help researchers, policymakers, and clinicians tailor future digital health technologies for musculoskeletal conditions and make informed decisions about how to implement technology in health systems. Helping patients find the "right care at the right time" for musculoskeletal conditions may help reduce burden on the health system and allow the ED to do what it was created for: provide life-saving care.

The primary objective of this review was to identify and describe available digital health tools that can triage and diagnose musculoskeletal conditions in primary, urgent, and emergency settings. The secondary objective was to summarize the performance and accuracy of digital health tools.

XSL•FO
RenderX

Methods

Overview

This scoping review was conducted in accordance with the Johanna Briggs Institute methodology for scoping reviews [16,17] and reported following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews; Checklist 1) and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Search) [18,19]. We were guided by Arksey and O'Malley's [20] framework with the additions of Levac et al [21]. A protocol was prospectively registered on the Open Science Framework (OSF). Amendments to the protocol were updated and uploaded to the OSF [22].

Search Strategy

An electronic search was conducted in 6 databases (MEDLINE [OVID], Embase [OVID], CENTRAL [OVID], CINAHL [EBSCO], Compendex, and Web of Science) and 4 gray literature sites (OpenGrey, GoogleScholar, arXiv, and medRxiv) with the aid of a biomedical librarian and information specialist. Our search strategy was not peer-reviewed but was tested through an iterative process by the biomedical librarian to ensure that search strategies returned identified seed papers. Initial search strategies were adapted from previously published work [11,14]. We searched databases and gray literature from inception to September 18, 2025. Table 1 provides the population-concept-context framework for our search strategy and illustrates how we operationalized our search. The full MEDLINE search strategy is outlined in Multimedia Appendix 1, and all other search strategies are uploaded and available on OSF [22]. To supplement the search, we screened the reference lists of relevant reviews and included records. We also searched the Cochrane Database of Systematic Reviews, PROSPERO, OSF, and *JBI Evidence Synthesis* to identify any active systematic or scoping reviews on the topic. The search approach for identifying gray literature is detailed in Multimedia Appendix 2.

Table . Population-concept-context (P-C-C) framework.

P-C-C	Definition	Keywords
Population (people with musculoskeletal pain)	Acute (traumatic) or chronic injury related to muscles, bones, joints, tendons, or ligaments problems that cause regional or generalized pain[23] and musculoskeletal disease or conditions as defined by the Global Burden of Disease (rheumatoid arthritis, osteoarthritis, low back pain, neck pain, and gout) [1].	(Musculoskeletal or MSK) injur* or pain* or tear* or ligament* or sprain* or strain* or gout or arthritis or rheumatic arthritis
Concept (digital health)	"The use of information and communications technology in support of health and health-related fields" [10]. Digital health captures eHealth, mobile health, advanced computer sciences, such as big data and AI[a], and the broad scope of telehealth and telemedicine [10].	telemedic* or telehealth or teletriag* or telecon-sult* or telecare or tele-care or virtual medicine or virtual care or virtual triage or digital health or digital tool or digital care or digital health technology or AI or artificial intelligence or deep learning or machine learning
Context (triage)	Triage guides the distribution of medical resources to patients when there is a scarcity of health care resources and often refers to a process to allocate, ration, or prioritize patient treatment and is considered first point-of-contact care [4]. Triage can be done by a clinician, patient, or technology (eg, AI) and may involve patients' self-assessment.	self-refer* or self-assess* or self-access* or tele-triage or triage or diagnosis or decision making or symptom checker*

[a]AI: artificial intelligence.

Inclusion Criteria

Studies of adults (aged 18 years and older) with musculoskeletal conditions (≥25% of the sample had to be musculoskeletal-related) that identified and reported a digital health tool designed specifically to triage or diagnose in primary, urgent, or emergency care settings were included. Textbox 1 describes the inclusion and exclusion criteria. We excluded studies that evaluated the effectiveness of virtual assessments. We also excluded studies that used digital health tools for secondary diagnoses (ie, patient had already seen a practitioner and given a diagnosis), as the tools were typically used to manage symptoms and not for primary triage or diagnosis.

XSL•FO
RenderX

Textbox 1. Overview of study selection criteria.

Inclusion criteria

- Adult participants (≥18 years) with a primary complaint of a musculoskeletal condition
- Sample has ≥25% musculoskeletal conditions
- Identifies and reports a digital health tool designed specifically for triage or diagnosis in primary care, urgent care, or emergency settings

Exclusion criteria

- Not English language
- Nonhuman data (eg, vignettes or simulated clinical cases)
- Study design (not original data, eg, review, opinion paper, commentaries, and guidelines)
- Not adult population (all participants aged at least 18 years)
- Not related to a digital health tool (instrument testing or replication or validation studies of clinician assessment to virtual assessment were excluded, no wearable or technology testing was excluded)

Study Selection

Records were collated and uploaded into EndNote (version 20.3; Clarivate Analytics), and duplicates were removed before uploading to Covidence (Veritas Health Innovation) for screening. Pairs of independent reviewers screened all records by title and abstracts, and a third reviewer (CLA) resolved any discrepancies if consensus could not be reached.

At the full-text stage, we first conducted pilot screening, where all reviewers assessed the same 5 full texts. If major discrepancies were identified, we met to review and discuss how to apply the screening criteria in a standardized manner. All full-text papers were reviewed by pairs of independent reviewers. Reasons for exclusion during full-text screening were recorded. Any disagreements between the reviewers at each stage of the selection process were resolved through consensus or by an additional reviewer (CLA) as required.

Data Extraction

Data were extracted from included records independently by pairs of reviewers using a custom data extraction tool designed in Microsoft Excel by the research team. Any disagreements were resolved via consensus. We extracted the following details where available: study characteristics (author, country, sample size, and study aim), participants' demographics (sex, age, and musculoskeletal pain or diagnosis), type of digital tool (name, purpose of tool, and target users), design and development process, platform of tool, tool delivery (eg, clinician or patient self-access), context or care setting in which the tool was used, assessment of performance or accuracy results, and key findings relevant to the review question. Where relevant, authors were contacted once via email to request missing data and to clarify details about the digital health tool.

Data Synthesis

Studies were summarized by study characteristics, digital health tool details, and performance assessment. Descriptive data were summarized as proportions when appropriate. Digital health tools were classified according to the World Health Organization digital health category [10] (eHealth, mHealth, and AI or machine learning), function (ie, triage, diagnosis, or both), care setting in which the tool was used (ie, ED, primary or urgent care, or mixed), research setting (ie, urban, rural, or both), how the tool was administered (eg, self-access or clinician-delivered), technology interface (eg, web-based or app), and intended user (patient-facing, clinician-facing, or both).

Digital tools were rated using the technology readiness level (TRL) and associated technology stage by the first author (LKT) and verified by a second rater [24]. The TRL ranges from 1 to 9, with 1 representing tool conception and 9 representing that the tool is ready to be used in real-world settings [24]. We classified TRL across technology stages (fundamental research, research and development, pilot and demonstration, early adoption, and commercially available) [24].

When available, the performance of the digital health tool was reported by identifying appropriate triage referrals or recommendations or diagnoses compared to a reference standard (eg, physician diagnosis). Measures of performance included diagnostic test accuracy (area under the receiver operator characteristic curve, sensitivity, specificity, positive predictive value, negative predictive value, and likelihood ratio) or reliability measures (internal consistency, test-retest reliability, and intra- or interrater reliability).

Deviations From Protocol

We sought to use the continuous active learning on Covidence to support title and abstract screening [25]. During attempts to calibrate the algorithm, the algorithm did not perform well in identifying relevant papers for this review. We were not confident in screening titles and abstracts with only 1 human reviewer (LKT). Instead, all titles and abstracts and full text records were screened in duplicate by 2 independent human reviewers.

As scoping reviews are an iterative process and aim to assess and evaluate the available evidence, a broad research question often results in a highly sensitive search and less specific records. We made pragmatic decisions and minor amendments to the selection criteria as the review progressed. At the full-text screening stage, studies including a general population had to report ≥25% of the sample being musculoskeletal-related to be included. This cutoff was determined based on studies that

XSL•FO
RenderX

indicated the prevalence of musculoskeletal presentations in ED was approximately 25% [26,27].

Results

Overview

The titles and abstracts of 5695 unique records were screened, and 189 papers were reviewed in full (Figure 1). In total, 34 studies met the inclusion criteria (n=37,509 participants across 33 studies, 12,470/37,509, 33% female). The median age was 50 (range 18-91) years. One study did not report the sample size. Sex and age data distribution were missing in 12 and 13 studies, respectively. A list of the studies that were excluded at the full-text stage is presented in Multimedia Appendix 3. Figure 2 charts the data of all 34 studies included by condition, publication year, and sample size of the study, if reported.

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) flowchart.

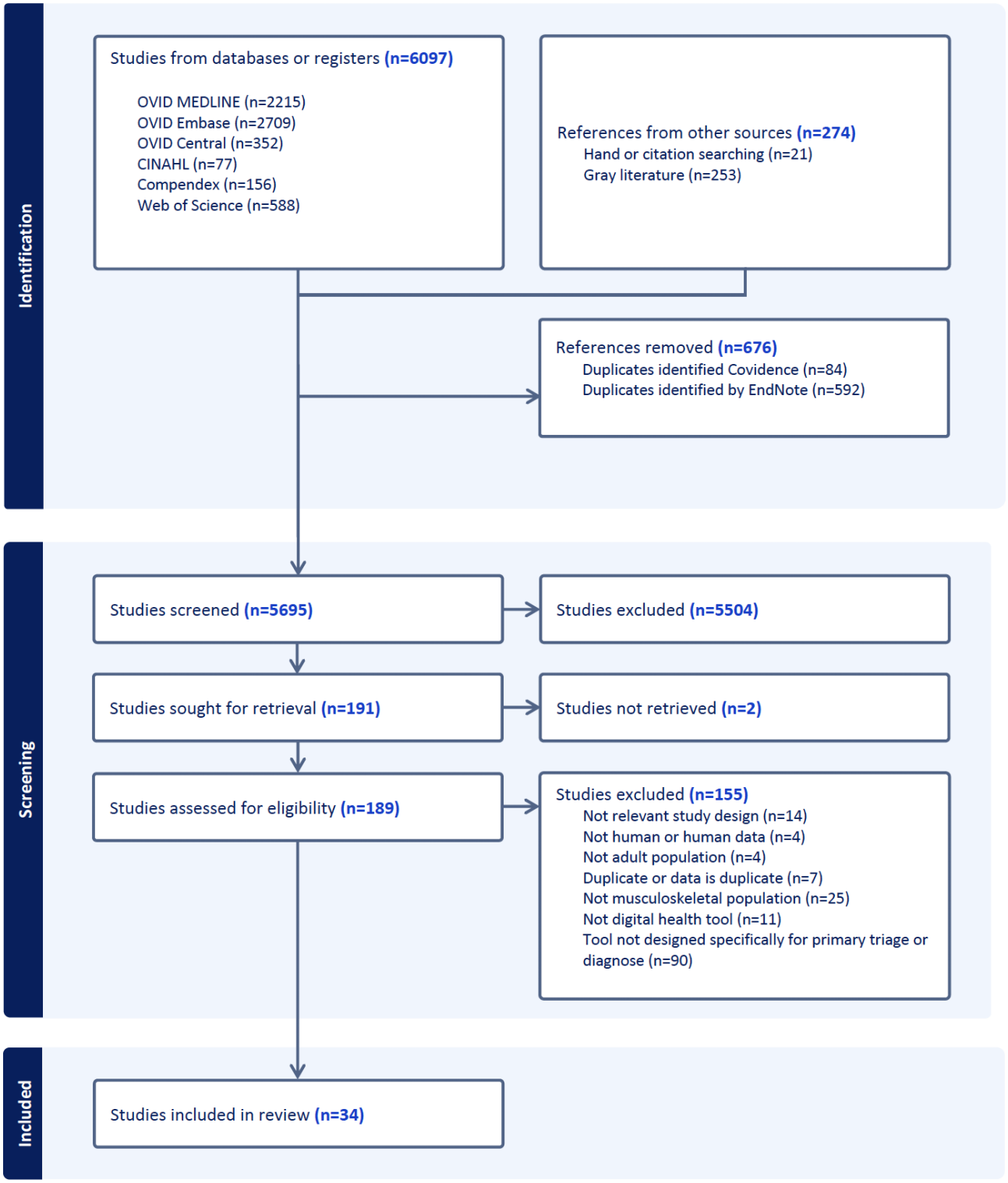

XSL•FO
RenderX

Figure 2. Sources of data charted by publication year and condition.

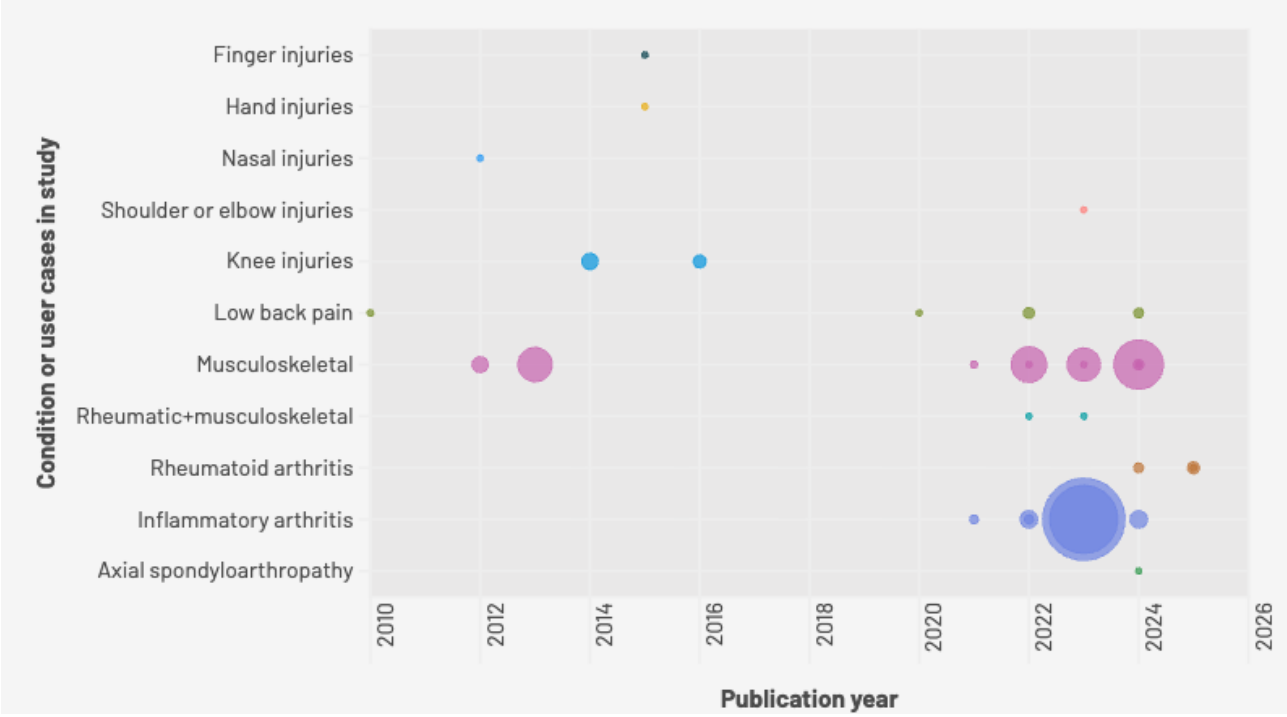

Study Characteristics

In total, 30 of 34 (88%) studies focused on primarily musculoskeletal conditions, and 4 of 34 (12%) studies focused on general populations with a subset of the sample being musculoskeletal conditions. Full study details can be found in Multimedia Appendix 4 [28-61].

In total, 25 of 34 (74%) studies were peer-reviewed, 7 of 34 (21%) records were conference abstracts, and 2 of 34 (6%) were industry case reports. Studies were published between 2010 and 2025. Cross-sectional (10/34, 29%) studies were most common, followed by nonrandomized or quasi-experimental studies (8/34, 24%), randomized controlled trials (5/34, 15%), retrospective cohort (4/34, 12%), mixed or multimethods (3/34), prospective observational cohort (2/34), and case report (2/34, 6%). All studies were conducted in high-income countries (n=1 country not reported), with the majority being from Europe (21/34, 62%) or the United States (7/34, 21%).

Digital health tools were evaluated across various care settings, including ED or urgent care (6/34, 18%), physician-led primary care (6/34, 18%), physiotherapist-led primary care (1/34, 3%), patient self-access (19/34, 56%), and mixed (eg, primary care and ED) settings (2/34, 6%).

Tool Identification and Characteristics

Overview

Inflammatory arthritis (eg, rheumatoid arthritis, gout, and spondyloarthropathy) and arthritis-related conditions were the most common (13/34, 38%) musculoskeletal conditions studied, followed by generic musculoskeletal conditions (11/34, 32%). Others tested digital tools for low back pain (4/34, 12%), knee (2/34, 6%), finger or hand (2/34, 6%), shoulder (1/34, 3%) conditions, and nasal fractures (1/34, 3%).

We identified 16 unique digital health tools (Table 2). In total, 7 studies did not report the name of the digital health tool that was studied or studied a bespoke tool designed for the study where we could not extract the name of the tool. Of the 34 tools, 13 (38%) tools reported that they were designed to “diagnose and triage,” 12 (35%) tools were designed to diagnose, and 9 (26%) tools were designed to triage (Table 2). Only 2 tools (Phio [32,39] and Digital Assessment Routing Tool [DART] [51,52]) were designed specifically to triage musculoskeletal conditions. Overall, 3 tools (ADA [37,41,46-49], Buoy Health [33], and WebMD Symptom Checker [40]) were generic health tools that had integrated algorithms to screen musculoskeletal conditions (among other conditions). One tool used OpenAI or LLMs (ChatGPT [29,34,49]) and reported algorithms capable of diagnosing musculoskeletal conditions. Five tools (Rheumatic? [43,45,53,56], Rheport [46-48], RheumConnect [60], ReumAI [38], and Bechterew-check [41]) were condition-specific (ie, designed for rheumatological or inflammatory conditions) with capabilities to differentiate these conditions from other types of musculoskeletal conditions. Two tools were joint-specific (Therapha for low back pain [28] and Virtual Knee Doc for acute knee injuries [30,31]).

XSL•FO
RenderX

Table . Characteristics of digital health tools, summarizing purpose, use, technology development, and availability.

Digital health tool name	Purpose of tool[a]	Type of tool[b], intended user, and access level	Tool format	Digital health category	Technology readiness level[c]	Technology readiness assessment[d]	Tool processes[e]	Available to public
ADA [37,41,46-49]	Diagnosis and triage	Symptom checker (patient; self-access)	App	mHealth[f], AI[g] or ML[h]	9	Commercially available	AI	Yes
Bechterew-check [41]	Diagnosis	Symptom checker (patient; self-access)	Web-based	mHealth	7	Pilot and demonstration	Clinical or decision support pathway	Yes (only in German)
Buoy Health [33]	Diagnosis and triage	Symptom checker (patient; self-access)	Web-based or app	eHealth, mHealth, AI or ML	9	Commercially available	AI	Yes
ChatGPT [29,34,49]	Diagnosis	Diagnostic predictor (patient; self-access)	Web-based	eHealth, AI or ML	9	Commercially available	AI	Yes
Digital Assessment Routing Tool [51,52]	Triage	Digital triage[i] (patient; self-access)	App	mHealth	8	Research and development	Clinical or decision support pathway	No
Phio [32,39]	Diagnosis and triage	Symptom checker[i] (patient; self-access)	App	mHealth, AI or ML	9	Commercially available	AI	No, Proprietary
Phone camera [42] (any built-in camera)	Triage	Tele-triage (clinician; clinician-administered)	Phone	mHealth	6	Pilot and demonstration	Clinical or decision support pathway	Yes
PhysioDirect [44]	Diagnosis and triage	Tele-triage (clinician; clinician-administered)	Phone	eHealth	6	Pilot and demonstration	Clinical or decision support pathway	No
Rheumatic? [43,45,53,56]	Diagnosis and triage	Symptom checker (patient; self-access)	Web-based	eHealth	8	Pilot and demonstration	Clinical or decision support pathway	Yes
Rheport[[37,46-48]	Diagnosis	Symptom checker (patient; self-access)	Web-based	eHealth	9	Early adoption	Clinical or decision support pathway	Yes (only in German)
Therapha [28]	Diagnosis and triage	Digital triage (clinician; clinician-administered)	Web-based	eHealth, AI or ML	9	Commercially available	Clinical or decision support pathway	No (propriety)
TriageXpert Dual Purpose [50]	Triage	Tele-triage (clinician; clinician-administered)	Phone	eHealth	7	Commercially available	Clinical or decision support pathway	No
RheumConnect [60]	Triage	Symptom checker (patient; self-access)	Web-based chatbot	eHealth, AI or ML	6	Pilot and demonstration	AI	No
ReumAI [38]	Triage	Tele-triage (clinician; clinician-administered)	Phone	eHealth, AI or ML	6	Pilot and demonstration	AI	No

XSL•FO
RenderX

Digital health tool name	Purpose of tool[a]	Type of tool[b], intended user, and access level	Tool format	Digital health category	Technology readiness level[c]	Technology readiness assessment[d]	Tool processes[e]	Available to public
Virtual Knee Doc [30,31]	Diagnosis	Symptom checker (patient; self-access)	Web-based	eHealth	6	Pilot and demonstration	Clinical or decision support pathway	No
WebMD Symptom Checker [40]	Diagnosis	Symptom checker (patient; self-access)	Web-based	eHealth, AI or ML	9	Commercially available	AI	Yes

[a]Triage: provide next steps for care based on symptoms and urgency, may provide preliminary diagnoses but not the objective of the tool. Diagnosis: provide a preliminary diagnosis based on symptoms, which aids to direct next steps in care.

[b]Type of tool: symptom checker: tool designed for patients to enter their symptom data; tele-triage: tool designed to triage using telephone interface; digital triage: tool designed to triage using eHealth or mHealth interface; diagnostic predictor: tool designed to use data to predict diagnosis or triage pathway.

[c]Based on Innovation Canada Technology Readiness Level: rated on a scale of 1-9, where 1=tool conception and 9=tool ready for real-world settings.

[d]Based on Innovation Canada Technology Readiness Stages: fundamental research, research and development, pilot and demonstration, early adoption, commercially available.

[e]Tool processes: AI: tool that uses big data to assign probability to allow for computer-driven decision-making; clinical decision support pathway: predefined decision tree or rule-based algorithms that support clinical decision-making.

[f]mHealth: mobile health.

[g]AI: artificial intelligence.

[h]ML: machine learning.

[i]Self-access within the UK National Health Service.

Intended Users

Most studies (24/34, 71%) reported on digital health tools designed for use by patients, 10 (10/34, 29%) studies targeted tools at clinicians. In total, 19 (19/34, 56%) studies reported on tools that were symptom checkers and were designed to be patient-facing. A total of 10 (29%) studies reported on clinician-facing tools for triage, diagnosis, or diagnostic prediction.

Patient-Facing

We classified ADA, Buoy Health, ChatGPT, Bechterew-check, Rheport, Rheumatic?, RheumConnect, Virtual Knee Doc, DART, and Phio as tools for patients. All used an app or a web-based interface. ADA, Buoy Health, and ChatGPT were tools used for generic health purposes, while the others were designed for specific groups of conditions (ie, rheumatological or musculoskeletal conditions). DART and Phio were designed to integrate with the UK National Health Service. Patients who used DART or Phio had their results forwarded to a primary care team or physiotherapist.

Clinician-Facing

In total, 7 tools were identified for clinicians. Therepha is a clinical decision support system designed for physiotherapists to diagnose and triage low back pain and was piloted in the ED [28]. ReumAI uses tele-triage where a nonphysician staff uses AI-guided telephone interviews to identify diagnoses and potential previsit tests [38]. Triage Xpert Dual Purpose [50] and PhysioDirect [58] were triage tools designed for implementation within specific health systems. One study examined triage of nasal fractures using a built-in camera to triage to the right hospital setting [42], and another used tele-triage to assess whether the ED could be avoided altogether for finger injuries [36]. Most clinician-facing triage tools used tele-triage (ie, phone call) as their interface, except for Therepha, which was a web-based tool. In total, 2 studies leveraged large datasets and AI to predict diagnoses, with 1 study using ChatGPT [29,60].

Performance and Usability

Of the 34 studies identified, 19 (56%) evaluated the performance of the digital health tool (Table 3). The most common definition and method to determine performance was measuring concordance to a clinician diagnosis or recommended triage pathway, often by evaluating sensitivity and specificity. The performance of these tools varied widely and was partly dependent on the context in which they were being used (eg, ED or primary care). Sensitivity ranged from 39% to 91%, and specificity from 23% to 80% [28,30,31,37,41,45,46,48]. The methods for measuring accuracy were poorly reported, often in the form of proportion of correct triage or diagnoses. Reported accuracy ranged from 33% to 98% across 12 unique tools (n=16 studies) [28,29,34,36-38,41,45,49,51,59,61]. The accuracy of tools used by patients in tertiary settings (eg, seeking care from a specialist such as an orthopedic surgeon or rheumatologist) was reported as higher than the accuracy of tools used in primary care settings.

For studies that compared digital health tools against each other, ADA was the most common tool used for comparison. When compared to rheumatologists and medical students, ADA was superior to clinician's diagnosis of rheumatic and nonrheumatic conditions, ChatGPT, and Bechterew-check [37,41,49]. ADA was comparable to Rheport for diagnostic accuracy of rheumatic conditions [48]. ChatGPT performed similar to experienced

XSL•FO
RenderX

rheumatologists for potential diagnostic accuracy for rheumatic conditions [49].

In total, 4 tools were available in multiple languages (ADA, ChatGPT, Rheumatic?, and WebMD Symptom Checker), and 8 tools were accessible to the public; however, 2 were designed for German speakers (Table 3). Based on the TRL, we classified 6 tools as being at the commercially available stage (ADA, Buoy Health, ChatGPT, Phio, Therapha, and WebMD Symptom Checker).

Figure 3 provides a visualization comparing TRL and performance evaluation for the identified digital health tools (Only 15 of the 16 identified digital tools are reported in this figure. The "phone built in camera" was not graphed.). If the tool did not complete a performance evaluation of the tool, a 0 was given for reported performance (ie, accuracy) on Figure 3. Despite some tools being commercially available, there was a discrepancy in reported performance findings for musculoskeletal conditions.

XSL•FO
RenderX

Table . Performance statistics of digital health tools.

Digital health tool and authors	Performance of tool evaluated (yes or no)	Definition used to define tool performance	Condition evaluated	Methods to evaluate performance	Sensitivity (%)	Specificity (%)	Accuracy of tool[a]	Other findings reported (%) (95% CI)
ADA[b]								
Knitza et al (2021) [46]	Yes	Concordance with physician diagnosis	Rheumatic	Sensitivity or specificity, PPV[c], NPV[d]	43	64	NR[e]	PPV 37 (26-48), NPV 69 (60-80)
Knitza et al (2024) [48]	Yes	Concordance with physician diagnosis	Rheumatic	Sensitivity or specificity, PPV, NPV	52	68	NR	PPV or NPV varied depending on whether ADA or Rheport was used first
Graf et al (2022) [37]	Yes	Concordance with identified diagnosis from clinical trial	Rheumatic	Sensitivity or specificity, accuracy	71	64	54% accurately diagnosed same condition	NR
Hannah et al (2024) [41]	Yes	Concordance with discharge summary report	Rheumatic	Sensitivity or specificity, accuracy	39	78	58% accurately diagnosed same condition	NR
Krusche et al (2024) [49]	Yes	Concordance with physician diagnosis	Rheumatic	Proportion	NR	NR	65% accurate for all cases; 71% accurate for cases with IRDs[f]; 61% accurate for non-IRD cases	NR
Bechterew-check								
Hannah et al (2024) [41]	Yes	Concordance with discharge summary report	Axial spondyloarthropathy	Sensitivity or specificity, accuracy	41	53	47% accurately diagnosed same condition	NR
Bespoke tool (no name reported)								
Demmelmaier et al (2010) [35]	No	NT[g]	Low back pain	NT	NT	NT	NT	NR
Martin and Payne (2020) [54]	No	NT	Low back pain	NT	NT	NT	NT	NR
Phillips et al (2012) [55]	No	NT	MSK[h]	NT	NT	NT	NT	NR
Ryan and Grinbergs (2024) [57]	No	NT	MSK	NT	NT	NT	NT	NR
Trivedi et al (2024) [61]	Yes	Concordance with nurse triage	MSK	Proportion	NR	NR	63% accurately triage	NR
Soin et al (2022) [59]	Yes	Concordance with physician diagnosis	Low back pain	NR	NR	NR	72% software predicted correct diagnosis	NR

XSL•FO
RenderX

Digital health tool and authors		Performance of tool evaluated (yes or no)	Definition used to define tool performance	Condition evaluated	Methods to evaluate performance	Sensitivity (%)	Specificity (%)	Accuracy of tool[a]	Other findings reported (%) (95% CI)
Buoy Health									
	Carmona et al (2022) [33]	No	NT	Generic MSK	NT	NT	NT	NT	NR
ChatGPT									
	Badsha et al (2024) [29]	Yes	Concordance with physician diagnosis	Rheumatic	NR	NR	NR	98% accurate with rheumatologist diagnosis	NR
	Daher et al (2023) [34]	Yes	Concordance with physician diagnosis	Shoulder or elbow injuries	NR	NR	NR	93% accurate with surgeon diagnosis; 83% accurate with surgeon management	NR
	Krusche et al (2024) [49]	Yes	Concordance with physician diagnosis	Rheumatic	Proportion	NR	NR	35% accurate for all cases; 71% accurate for cases with IRDs; 15% accurate for non-IRD cases	NR
Digital Assessment Routing Tool (DART)									
	Lowe et al (2022) [51]	Yes	Concordance with physiotherapist expert	MSK	Proportion	NR	NR	84% DART matched physiotherapist	NR
	Lowe et al (2024) [52]	Yes	Concordance with physiotherapist expert	MSK	Intraclass correlation coefficient (ICC)	NR	NR	NR	ICC 0.37 (0.16 - 0.55)
Phio									
	Bond et al (2024) [32]	No	NT	NT	NT	NT	NT	NT	NR
	Gymer et al (2023) [39]	No	NT	NT	NT	NT	NT	NT	NR
Phone Camera									
	Hara et al (2015) [42]	Yes	Accuracy of triage recommendations	Finger injuries	NR	NR	NR	NR	NR
PhysioDirect									
	Kelly et al (2021) [44]	No	NT	NT	NT	NT	NT	NT	NR
Rheport									
	Knitza et al (2021) [46]	Yes	Concordance with physician diagnosis	Rheumatic	Sensitivity or specificity, PPV, NPV	54	52	NR	PPV 35 (25-47); NPV 70 (58-79)

XSL•FO
RenderX

Digital health tool and authors		Performance of tool evaluated (yes or no)	Definition used to define tool performance	Condition evaluated	Methods to evaluate performance	Sensitivity (%)	Specificity (%)	Accuracy of tool[a]	Other findings reported (%) (95% CI)
	Knitza et al (2024) [48]	Yes	Concordance with physician diagnosis	Rheumatic	Sensitivity or specificity, PPV, NPV	62	47	NR	PPV or NPV varied depending on whether ADA or Rheport was used first
Rheumatic?[i]									
	Knevel et al (2022) [45]	Yes	Concordance with physician diagnosis/treatment recommendation	Rheumatic	Sensitivity or specificity, AUC-ROC[j]	67	72	AUC-ROC 75 (95% CI 62 - 89)	NR
	Qin et al (2024) [56]	No	NT	Rheumatic	NT	NT	NT	NT	NR
	Lundberg et al (2023) [53]	No	NT	Rheumatic	NT	NT	NT	NT	NR
	Jakobi et al (2025) [43]	No	NT	Rheumatic	NT	NT	NT	NT	NR
ReumAI									
	Gómez-Centeno et al (2025) [38]	Yes	Concordance with physician diagnosis	Rheumatic	NR	NR	NR	53% accurate with rheumatologists	NR
RheumConnect									
	Tan et al (2023) [60]	No	NT	Rheumatic	NT	NT	NT	NT	NT
Therapha									
	Badahman et al (2024) [28]	Yes	Concordance with MRI findings	Low back pain	Sensitivity or specificity, PPV, NPV, ROC[k]	88	80	ROC 0.84 (95% CI 0.6 - 1.0; *P*=.001)	PPV 99 (25-47); NPV 27 (58-79)
Triage Xpert Dual Purpose									
	Li et al (2023) [50]	No	NT	MSK	NT	NT	NT	NT	NT
Virtual Knee Doc									
	Bisson et al (2014) [30]	Yes	Concordance with physician diagnosis	Knee injuries	Sensitivity or specificity	89	27	NR	NT
	Bisson et al (2016) [31]	Yes	Concordance with physician diagnosis	Knee injuries	Sensitivity or specificity	91	23	NR	NT
WebMD Symptom Checker									

XSL•FO
RenderX

Digital health tool and authors	Performance of tool evaluated (yes or no)	Definition used to define tool performance	Condition evaluated	Methods to evaluate performance	Sensitivity (%)	Specificity (%)	Accuracy of tool[a]	Other findings reported (%) (95% CI)
Hageman et al (2015) [40]	Yes	Concordance with physician diagnosis	Hand injuries	Proportion	NR	NR	33% accurate with hand surgeon diagnosis	NT

[a]Reported values from the study and not interpretation of authors.

[b]Findings reported from ADA diagnosis 1 (D1) in study.

[c]PPV: positive predictive value.

[d]NPV: negative predictive value.

[e]NR: not reported.

[f]IRD: inflammatory rheumatic disease.

[g]NT: not tested.

[h]MSK: musculoskeletal.

[i]Findings reported from dataset A in study.

[j]AUC-ROC: area under the receiver operating curve.

[k]ROC: receiver operating curve.

Figure 3. Visualization comparing TRL and the highest reported performance evaluation across identified digital health tools. NR: not reported; TRL: technology readiness level.

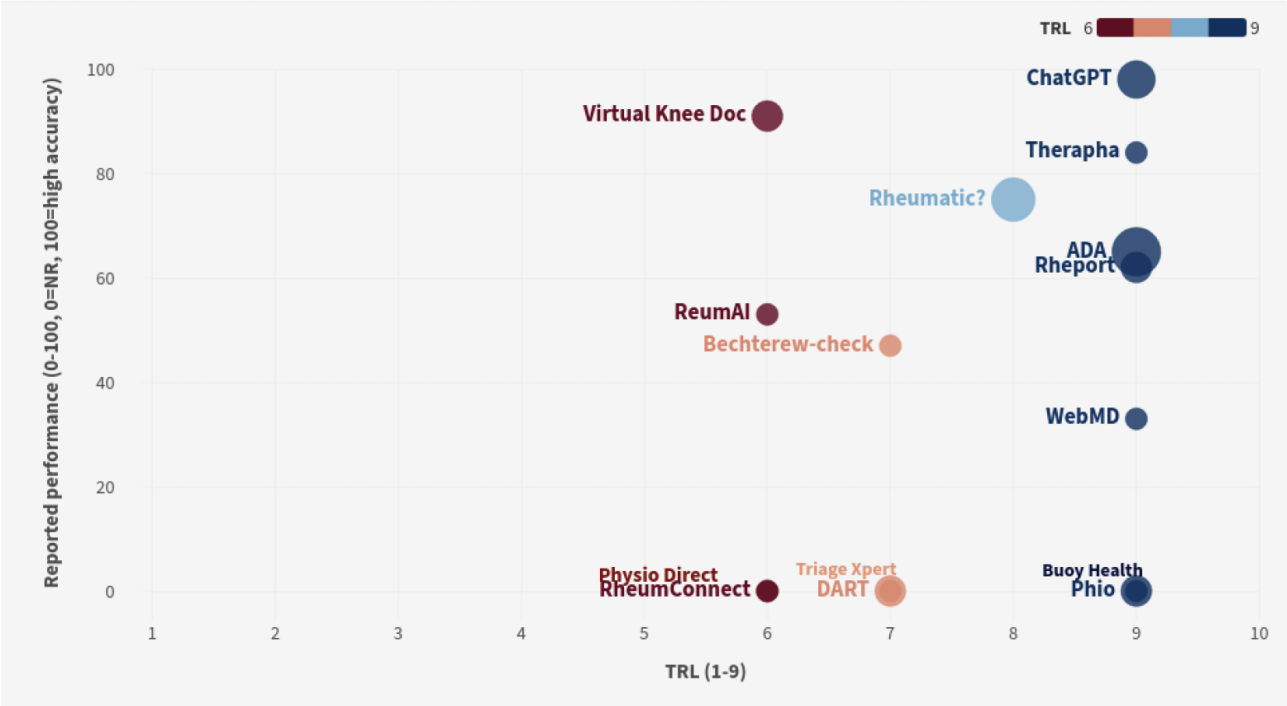

Discussion

Principal Findings

We aimed to identify and describe the available tools for triaging and diagnosing musculoskeletal conditions in primary, urgent, and emergency settings. Based on a synthesis of 34 studies and data from 16 different digital health tools, there were no digital health tools with sufficient evidence to support effective triage and diagnosis of musculoskeletal conditions across these settings. Approximately half of these tools were available to the public. Not all tools were available in English, with 2 tools only available in German (Bechterew-check and Rheport). The most frequently studied digital health tool was ADA (n=5), followed by Rheumatic? (n=4), then ChatGPT (n=3). Only 2 tools (DART and Phio) were purposely developed for screening musculoskeletal conditions. Both tools are not currently available outside of the UK's National Health Service. We were surprised to find so few digital health tools targeting musculoskeletal conditions, given the substantial global burden of musculoskeletal conditions [1]. Notably, rheumatological or inflammatory arthritis was the most prevalent musculoskeletal condition studied, despite low back pain being the most common musculoskeletal condition seen in ED and primary care settings [62]. We identified 4 studies that included digital health tools

XSL•FO
RenderX

targeting low back pain, but only 1 of these reported which tool was used (Therapha). Our findings reflect the discordance of research across digital health technology and the current health landscape. Many tools were inaccessible or not designed for practical use in managing musculoskeletal pain, the most burdensome conditions seen in primary care.

Our secondary objective was to summarize the performance and accuracy of the included digital health tools. Approximately 50% of the studies evaluated the performance of a digital health tool. Apart from ChatGPT, most generic digital health tools (eg, ADA and WebMD Symptom Checker) reported poor accuracy (often less than 50% accuracy in identifying the correct diagnosis compared to clinicians) for musculoskeletal conditions [37,40,48]. Despite the use of ChatGPT by the public as a symptom checker [15], ChatGPT's accuracy for diagnosing musculoskeletal and rheumatic conditions was variable, ranging from 33% to 98% [29,34,49]. We suggest that further research is needed before considering ChatGPT as an accurate diagnostic or screening tool. Tools that were designed to diagnose peripheral or spinal musculoskeletal conditions (eg, low back pain or knee injuries) appear to be more promising with high sensitivity (88% - 91%) [28,31]. Finally, tools designed specifically to triage (rather than diagnose) musculoskeletal conditions (ie, Phio and DART) demonstrated the best performance. Recent findings published on DART and Phio indicate that these tools have high agreement (>90%) with expert physiotherapist recommendations on next care pathways [63,64]. However, the heterogeneity across evaluation methods highlights the importance of standardized development and evaluation frameworks to ensure that digital triage tools for musculoskeletal conditions are accurate, transparent, and safe before integrated into clinical settings and workflows.

Not Yet Ready for Prime Time

One of the key findings of our review is that some tools are commercially available and integrated into health systems for musculoskeletal screening without robust methodological evaluation or reporting. Premature implementation raises concerns, particularly given the risk of misdirecting patients or delaying appropriate care. Before being adopted at scale, digital triage tools must demonstrate value in real-world settings and meet minimum standards for safety, accuracy, and usability. However, many studies evaluating these tools lack transparent reporting, making it difficult to assess how performance claims were derived. Studies reporting high accuracy of their digital tools often had poor transparency or a lack of details on their tool evaluation. We suggest caution with interpreting these digital health tools as ready for public use without further evaluation. It is also unclear how tools that use LLMs operate. These networks are often termed "black boxes" due to the inability to explain how these systems achieve their output [65].

Our findings have been confirmed by a recent study that evaluated diagnostic accuracy and clinical reasoning using 6 different generative AI (LLMs) for rheumatic diagnoses [66]. Despite the LLMs reporting high diagnostic accuracy (~80%), all models reported subpar clinical reasoning quality (eg, explaining reasons for supporting diagnoses) [66]. These findings underscore the importance of digital health tools requiring both high diagnostic accuracy alongside transparent algorithms to help to explain the logic behind the tool's decision. To improve transparency and enable reproducibility, it is important to establish standards for incorporating ethical AI in digital health. Without transparency in how tools were developed, or in the algorithms used, it is unclear whether the tools are safe for the public to use.

The only digital health tool with robust evaluation of its performance was the generic health app, ADA, which is a Conformité Européenne–certified medical product [46,48]. ADA's performance was inconsistent across the studies, and ADA correctly identified the musculoskeletal condition or triage option in fewer than half of the cases [37,41,46,48,49]. Condition-specific digital health tools (Rheport [46,48], Rheumatic? [45], Virtual Knee Doc [30,31], Therapha [28], and ReumAI [38]) performed slightly better. The reported accuracy was higher in these tools, especially if these tools were implemented in tertiary care settings (outside of the ED or primary care). We are not aware of an acceptable threshold for performance (ie, accuracy) for digital health tools. However, we recommend implementing tools that are at least more accurate than flipping a coin and provide consistent results across different study contexts or musculoskeletal conditions.

AI-driven tools, like ADA or ChatGPT, may perform better than clinician decision support systems or physicians or rheumatologists in diagnosing rheumatic conditions [49,67]. Integrating digital health tools in tandem with other nonspecialist professions (eg, general practitioners and allied health professionals) could help guide patients to their next care steps as they wait for specialists (eg, rheumatologists) or avoid unnecessary visits to specialists or other care providers. AI-driven tools that have included diagnostic findings (eg, imaging, clinical symptoms or signs, and bloodwork) have superior diagnostic accuracy to other AI models [41,67,68]. Until robust stand-alone digital health tools are developed (ie, a symptom checker that can be used independently by patients), combining digital health tools and clinician feedback may be the best method to streamline diagnosis and care in complex cases while providing timely care for common musculoskeletal conditions.

Several frameworks for evaluating digital health tools have been proposed [69]. A recent scoping review identified 12 key domains—ranging from tool description and content to safety, clinical effectiveness, and efficacy—across 95 frameworks that developers and researchers can draw on [69]. However, the heterogeneity reflects a broader challenge: many digital health tools span multiple categories (eg, eHealth or mHealth tools incorporating AI), making classification inconsistent and evaluation difficult. Advancing this field requires standardized terminology, harmonized testing and evaluation frameworks, and clear reporting guidelines—crucial steps to ensure both progress and patient safety.

Why Would a Digital Health Tool Do a Poor Job at Screening Musculoskeletal Conditions?

Through the process of screening studies for inclusion into our review, we found definitions of musculoskeletal conditions that were vague and varied widely. Definitions of "musculoskeletal"

XSL•FO
RenderX

are often limited to orthopedic conditions or pain related to musculoskeletal structures [1,23]. However, musculoskeletal conditions are a complex category involving heterogeneous conditions, such as rheumatological or inflammatory arthritis or gout, that are not typically grouped as musculoskeletal in clinical practice. We relied on a broad definition to capture specific musculoskeletal conditions (eg, rheumatological conditions, arthritis, and gout) and pain related to musculoskeletal structures (eg, sprains and strains).

There is nuance in how triage would be conducted for acute versus chronic musculoskeletal conditions, including screening questions related to condition pathophysiology, subjective history, pattern of symptoms, and disability (eg, red flags), which might explain some of the variability in performance metrics of different digital tools [70]. Early diagnosis and treatment planning is often iterative for those with musculoskeletal conditions and varies depending on the condition. For example, targeted medication plays a vital role in managing rheumatological conditions [71], whereas some orthopedic conditions are managed with exercise and minimal pharmacological interventions [2]. This complexity will impact triage algorithms by influencing treatment recommendations (eg, who the patient should see) and timing of care (eg, urgent or wait-and-see). Therefore, tools that have high accuracy (ie, good performance) for triaging and diagnosing general health conditions may not necessarily have the same effectiveness when applied to musculoskeletal conditions.

Digital health tools may perform poorly at screening because of user error relating to symptom data entry and patient interaction with the tool. One solution to this is adding more key information (eg, diagnostic tests) to an AI-driven model to improve the diagnostic accuracy of the model [67]. We also suggest future work to involve patient end users to develop and refine digital health tools. Most digital health tool algorithms are derived from clinicians' clinical reasoning, which may not follow the same thought process as a patient. In a recent qualitative study exploring how patients should be engaged in AI application to health care, patients felt that the priorities of researchers, particularly for AI tools, were to improve efficiency and effectiveness of care [72]. In contrast, patients were more interested in using AI to address issues related to accessing health care [72]. Patients should be involved early in the design and development phases to enhance the usability and understandability of digital health tools. However, patient perspectives are often included only after the digital health tool is designed. We argue that engaging patients early in the development process, such as developing the AI algorithms, may yield more acceptable and usable digital health tools.

It is unlikely that a "one size fits all" digital health tool can effectively diagnose and triage all musculoskeletal conditions. Most patient-facing tools in our review were web- or app-based tools in the form of generic symptom checkers. ChatGPT has an accessible interface and is relatively easy to use [15]. Clinician-facing tools may benefit from greater complexity or condition specificity, depending on the context in which the tools will be implemented. Instead of an either-or—general or condition-specific—we advocate for designers to consider their design goal (ie, triage or diagnosis) and intended user (ie, patient or clinician), which may improve accuracy in digital health tools for musculoskeletal conditions.

Move (Relatively) Fast, and Try Not to Break Things

The field of digital health is growing and changing rapidly. Many health systems have been forced to move toward implementing digital health, particularly AI-driven tools, without being afforded adequate time and resources to consider safety, effectiveness, or downstream consequences [13]. This may be in part due to social and political imperatives to set key performance (productivity) indicators, transition of health care services, and drive toward greater and faster innovation. We suggest that such a climate could be dangerous for health care, especially if digital health implementation continues without adequate evidence, as our findings highlight.

There is a place for digital health triage tools used by patients and clinicians in the current health care context. Self-referral and symptom checkers can be effective for musculoskeletal conditions and to support patients' access to care, particularly when patients do not have a consistent primary care team or provider [11]. Acute care clinics using a self-referral form found that patients with musculoskeletal conditions were accurate at self-referring, used less health care, and incurred fewer costs [73]. Emerging evidence also indicates that patients are using LLMs such as ChatGPT to make health care decisions, and it appears that the general public is accepting of using AI for health care advice and psychological support [15]. However, more research is needed to ensure that patients presenting with musculoskeletal conditions have a safe, accurate, and well-designed tool to direct them to the best care for their situation. Digital health tools also need to be designed to suit diverse populations, including those with low health literacy and limited digital literacy.

Future Considerations and Clinical Implications

While there is a breadth of studies available for digital health and digital triage, we identified the following knowledge gaps: (1) reporting and transparency on digital health tool development must improve, (2) evaluating digital health tools needs a standard approach, (3) studying the accuracy of triage recommendations requires robust prospective studies, and (4) implementing musculoskeletal-focused digital health tools for first point-of-contact care requires attention.

Despite the absence of digital health tools for triage of musculoskeletal conditions, we are aware of other tools in development, such as SupportPrim [74], which might fill some of the knowledge gaps for health care providers. Our findings do not provide conclusive evidence to support using digital health tools to accurately screen musculoskeletal conditions in many health settings. We recommend that clinicians use these digital health tools as an adjunct to help guide patients, particularly when used as a symptom checker, but to still defer to sound clinical judgment and help patients understand the limitations of the tools.

Limitations

Although we used a thorough search of published and unpublished data, it is possible that we have missed relevant

XSL•FO
RenderX

digital health tools or papers. We set a sample threshold of at least 25% of the sample population with musculoskeletal conditions, and this may have resulted in us missing some studies (eg, studies that were just below the threshold were excluded). The threshold was intended to maximize external validity [26,27]. Our goal was to identify tools that were primarily designed to triage or diagnose (vs manage) musculoskeletal conditions. Therefore, we excluded studies and tools that were designed for self-management, even if they included a symptom checker. This led us to exclude studies that used tools for secondary triage or diagnosis (ie, used by patients who had a diagnosis or had already been seen in a primary or emergency setting) as we wanted to capture tools that could be used at the first point-of-contact. We identified some potential musculoskeletal-specific digital health tools that could be used for secondary triage or diagnosis (Multimedia Appendix 5). While we attempted to report on the performance and accuracy of the tools identified, some tools pooled data from the entire population (ie, not musculoskeletal only). Therefore, the findings may under- or overestimate the accuracy of the tool for musculoskeletal conditions. This again points to the need to design musculoskeletal-specific tools and carefully evaluate their performance.

Conclusions

The rapid growth of AI and digital health solutions is transforming health care systems worldwide, with increasing interest in automating triage and diagnosis. However, our review shows that musculoskeletal conditions remain a blind spot: few tools were specifically designed for this purpose, and most performed poorly when applied to musculoskeletal populations. Despite commercial availability and implementation in some settings, the evidence base was weak, and tool performance was inconsistent and opaque. Health systems and clinicians should exercise caution before integrating these tools into care pathways. Musculoskeletal-specific digital tools developed through transparent, standardized processes are urgently needed to ensure safety, clinical value, and trustworthiness.

Acknowledgments

Generative artificial intelligence was not used to draft any portion of this manuscript.

Funding

LKT is a Mitacs Elevate Fellow and a 2025 Health Research BC Research Trainee recipient and is funded by Health Research BC (RT-2025-04847)

Data Availability

The datasets generated or analyzed during this study are available in the Open Science Framework repository [22].

Authors' Contributions

Conceptualization: LKT, CLA, JGW (equal)
Data curation (database searching): DG (lead), LKT, CLA (supporting)
Investigation: LKT, JGW, RV, EL, JLC, EW, CS, CLA
Methodology: LKT, CLA (equal)
Formal analysis: LKT (lead), CLA, JGW, RV (supporting)
Project administration: LKT (lead), CLA (supporting)
Visualization: LKT (lead), JGW, CLA (supporting)
Writing—original draft: LKT (lead), CLA (supporting)
Writing—review and editing: All authors

Conflicts of Interest

None declared.

Multimedia Appendix 1
OVID MEDLINE full search strategy.
[DOCX File, 49 KB - jmir_v28i1e81578_app1.docx]

Multimedia Appendix 2
Gray literature search strategy and results.
[DOCX File, 27 KB - jmir_v28i1e81578_app2.docx]

Multimedia Appendix 3
Studies excluded at full-text stage.
[DOCX File, 73 KB - jmir_v28i1e81578_app3.docx]

XSL•FO
RenderX

Multimedia Appendix 4
Characteristics of included studies, summarizing design, demographics, and digital tool features.
[DOCX File, 57 KB - jmir_v28i1e81578_app4.docx]

Multimedia Appendix 5
Digital health tools for secondary triage or diagnosis.
[DOCX File, 25 KB - jmir_v28i1e81578_app5.docx]

Checklist 1
PRISMA-ScR checklist.
[PDF File, 249 KB - jmir_v28i1e81578_app6.pdf]

References

1. Gill TK, Mittinty MM, March LM, et al. Global, regional, and national burden of other musculoskeletal disorders, 1990–2020, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. Lancet Rheumatol 2023 Nov;5(11):e670-e682. [doi: 10.1016/S2665-9913(23)00232-1]
2. Lin I, Wiles L, Waller R, et al. What does best practice care for musculoskeletal pain look like? Eleven consistent recommendations from high-quality clinical practice guidelines: systematic review. Br J Sports Med 2020 Jan;54(2):79-86. [doi: 10.1136/bjsports-2018-099878] [Medline: 30826805]
3. Wrightson J, Truong LK, Haagaard A, Ardern CL. Estimating the prevalence of low acuity musculoskeletal pain in the emergency department. Presented at: Canadian for Health Services and Policy Research (CHSPR) 2025 Conference Abstract; Mar 3-4, 2025.
4. Iserson KV, Moskop JC. Triage in medicine, part I: concept, history, and types. Ann Emerg Med 2007 Mar;49(3):275-281. [doi: 10.1016/j.annemergmed.2006.05.019] [Medline: 17141139]
5. Joseph C, Morrissey D, Abdur-Rahman M, Hussenbux A, Barton C. Musculoskeletal triage: a mixed methods study, integrating systematic review with expert and patient perspectives. Physiotherapy 2014 Dec;100(4):277-289. [doi: 10.1016/j.physio.2014.03.007] [Medline: 25242531]
6. Erku D, Khatri R, Endalamaw A, et al. Digital health interventions to improve access to and quality of primary health care services: a scoping review. Int J Environ Res Public Health 2023 Sep 28;20(19):19. [doi: 10.3390/ijerph20196854] [Medline: 37835125]
7. Ibrahim MS, Mohamed Yusoff H, Abu Bakar YI, Thwe Aung MM, Abas MI, Ramli RA. Digital health for quality healthcare: a systematic mapping of review studies. Digit Health 2022;8:20552076221085810. [doi: 10.1177/20552076221085810] [Medline: 35340904]
8. Chambers D, Cantrell AJ, Johnson M, et al. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. BMJ Open 2019 Aug 1;9(8):e027743. [doi: 10.1136/bmjopen-2018-027743] [Medline: 31375610]
9. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2025-10-24]
10. Recommendations on digital interventions for health system strengthening—executive summary. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/WHO-RHR-19.8 [accessed 2025-01-05]
11. Babatunde OO, Bishop A, Cottrell E, et al. A systematic review and evidence synthesis of non-medical triage, self-referral and direct access services for patients with musculoskeletal pain. PLOS ONE 2020;15(7):e0235364. [doi: 10.1371/journal.pone.0235364] [Medline: 32628696]
12. Pairon A, Philips H, Verhoeven V. A scoping review on the use and usefulness of online symptom checkers and triage systems: how to proceed? Front Med (Lausanne) 2022;9:1040926. [doi: 10.3389/fmed.2022.1040926] [Medline: 36687416]
13. Tyler S, Olis M, Aust N, et al. Use of artificial intelligence in triage in hospital emergency departments: a scoping review. Cureus 2024 May;16(5):e59906. [doi: 10.7759/cureus.59906] [Medline: 38854295]
14. Wallace W, Chan C, Chidambaram S, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. NPJ Digit Med 2022 Aug 17;5(1):118. [doi: 10.1038/s41746-022-00667-w] [Medline: 35977992]
15. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023 May 17;10:e47564. [doi: 10.2196/47564] [Medline: 37195756]
16. Peters MDJ, Godfrey C, McInerney P, et al. Best practice guidance and reporting items for the development of scoping review protocols. JBI Evid Synth 2022 Apr 1;20(4):953-968. [doi: 10.11124/JBIES-21-00242] [Medline: 35102103]
17. Pollock D, Peters MDJ, Khalil H, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evid Synth 2023 Mar 1;21(3):520-532. [doi: 10.11124/JBIES-22-00123] [Medline: 36081365]
18. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]

XSL•FO
RenderX

19. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
20. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
21. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69. [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
22. Truong LK, Lui E, Wrightson J, et al. Digital health tools used to triage musculoskeletal pain in primary, urgent and emergency settings: a scoping review. Open Science Framework. 2024. URL: https://osf.io/y5rp7/overview [accessed 2026-01-05]
23. Smith E, Hoy DG, Cross M, et al. The global burden of other musculoskeletal disorders: estimates from the Global Burden of Disease 2010 study. Ann Rheum Dis 2014 Aug;73(8):1462-1469. [doi: 10.1136/annrheumdis-2013-204680] [Medline: 24590181]
24. Technology Readiness Level (TRL) Assessment Tool. Government of Canada. 2021. URL: https://ised-isde.canada.ca/site/clean-growth-hub/en/technology-readiness-level-trl-assessment-tool [accessed 2025-05-24]
25. Walton A. December 2022—title and abstract screening using machine learning. Covidence. 2022. URL: https://www.covidence.org/blog/release-notes-december-2022-machine-learning [accessed 2024-07-11]
26. Bird S, Thompson C, Williams KE. Primary contact physiotherapy services reduce waiting and treatment times for patients presenting with musculoskeletal conditions in Australian emergency departments: an observational study. J Physiother 2016 Oct;62(4):209-214. [doi: 10.1016/j.jphys.2016.08.005] [Medline: 27637771]
27. Gagnon R, Perreault K, Berthelot S, et al. Direct-access physiotherapy to help manage patients with musculoskeletal disorders in an emergency department: results of a randomized controlled trial. Acad Emerg Med 2021 Aug;28(8):848-858. [doi: 10.1111/acem.14237] [Medline: 33617696]
28. Badahman F, Alsobhi M, Alzahrani A, et al. Validating the accuracy of a patient-facing clinical decision support system in predicting lumbar disc herniation: diagnostic accuracy study. Diagnostics (Basel) 2024 Aug 26;14(17):1870. [doi: 10.3390/diagnostics14171870] [Medline: 39272655]
29. Badsha HM, Khan B, Harifi G, J A, Raman S. AB1488 is the future of rheumatology here? A study of a proprietary rule engine and artificial intelligence GPT4 (AI GPT4) for initial evaluation of rheumatology cases. Ann Rheum Dis 2024 Jun;83(Suppl 1):2112. [doi: 10.1136/annrheumdis-2024-eular.1942] [Medline: 644868572]
30. Bisson LJ, Komm JT, Bernas GA, et al. Accuracy of a computer-based diagnostic program for ambulatory patients with knee pain. Am J Sports Med 2014 Oct;42(10):2371-2376. [doi: 10.1177/0363546514541654] [Medline: 25073597]
31. Bisson LJ, Komm JT, Bernas GA, et al. How accurate are patients at diagnosing the cause of their knee pain with the help of a web-based symptom checker? Orthop J Sports Med 2016 Feb;4(2):2325967116630286. [doi: 10.1177/2325967116630286] [Medline: 26962542]
32. Bond C, Guard M, Grinbergs P. Case report: digital musculoskeletal triage and rehabilitation tools enhance accessibility, user experience and outcomes in mechanical knee pain. Physiotherapy 2024 Jun;123(Suppl 1):e115. [doi: 10.1016/j.physio.2024.04.143]
33. Carmona KA, Chittamuru D, Kravitz RL, Ramondt S, Ramírez AS. Health information seeking from an intelligent web-based symptom checker: cross-sectional questionnaire study. J Med Internet Res 2022 Aug 19;24(8):e36322. [doi: 10.2196/36322] [Medline: 35984690]
34. Daher M, Koa J, Boufadel P, Singh J, Fares MY, Abboud JA. Breaking barriers: can ChatGPT compete with a shoulder and elbow specialist in diagnosis and management? JSES Int 2023 Nov;7(6):2534-2541. [doi: 10.1016/j.jseint.2023.07.018] [Medline: 37969495]
35. Demmelmaier I, Denison E, Lindberg P, Asenlöf P. Physiotherapists' telephone consultations regarding back pain: a method to analyze screening of risk factors. Physiother Theory Pract 2010 Oct;26(7):468-475. [doi: 10.3109/09593980903433938] [Medline: 20649497]
36. Dias L, Maughan E, Kisha A, Moorthy R. Telephone triage in the management of patients with nasal injuries [Abstract]. Clin Otolaryngol 2012;37:45. [doi: 10.1111/j.1749-4486.2012.02517.x] [Medline: 71023181]
37. Gräf M, Knitza J, Leipe J, et al. Comparison of physician and artificial intelligence-based symptom checker diagnostic accuracy. Rheumatol Int 2022 Dec;42(12):2167-2176. [doi: 10.1007/s00296-022-05202-4] [Medline: 36087130]
38. Gómez-Centeno A, Sabaris-Vilas M, Garcia-Sancho F, Segura-Sanchez J. POS0883 Optimizing rheumatology consultations with artificial intelligence: insights from the ReumAI pilot study. Ann Rheum Dis 2025 Jun;84(Suppl 1):1018. [doi: 10.1016/j.ard.2025.06.238]
39. Gymer M, Guard M, Grinbergs P. A case report: digital musculoskeletal triage and rehabilitation tools improve outcomes and offer a positive experience for lower back pain. JMIR Bioinform Biotechnol. Preprint posted online on Oct 20, 2022. [doi: 10.2196/preprints.43686]
40. Hageman M, Anderson J, Blok R, Bossen JKJ, Ring D. Internet self-diagnosis in hand surgery. HAND (N Y) 2015 Sep;10(3):565-569. [doi: 10.1007/s11552-014-9707-x] [Medline: 26330798]

XSL•FO
RenderX

41. Hannah L, von Sophie R, Gabriella RM, et al. Stepwise asynchronous telehealth assessment of patients with suspected axial spondyloarthritis: results from a pilot study. Rheumatol Int 2024 Jan;44(1):173-180. [doi: 10.1007/s00296-023-05360-z] [Medline: 37316631]
42. Hara T, Nishizuka T, Yamamoto M, Iwatsuki K, Natsume T, Hirata H. Teletriage for patients with traumatic finger injury directing emergency medical transportation services to appropriate hospitals: a pilot project in Nagoya City, Japan. Injury 2015 Jul;46(7):1349-1353. [doi: 10.1016/j.injury.2015.02.022] [Medline: 25799472]
43. Jakobi S, Boy K, Wagner M, et al. Rheumatic? A diagnostic decision support tool for individuals suspecting rheumatic diseases: mixed-methods usability and acceptability study. BMC Rheumatol 2025 May 23;9(1):59. [doi: 10.1186/s41927-025-00507-w] [Medline: 40410901]
44. Kelly M, Higgins A, Murphy A, McCreesh K. A telephone assessment and advice service within an ED physiotherapy clinic: a single-site quality improvement cohort study. Arch Physiother 2021 Feb 8;11(1):4. [doi: 10.1186/s40945-020-00098-4] [Medline: 33550990]
45. Knevel R, Knitza J, Hensvold A, et al. Rheumatic?—A digital diagnostic decision support tool for individuals suspecting rheumatic diseases: a multicenter pilot validation study. Front Med 2022;9. [doi: 10.3389/fmed.2022.774945] [Medline: 2016520483]
46. Knitza J, Mohn J, Bergmann C, et al. Accuracy, patient-perceived usability, and acceptance of two symptom checkers (Ada and Rheport) in rheumatology: interim results from a randomized controlled crossover trial. Arthritis Res Ther 2021 Apr 13;23(1):112. [doi: 10.1186/s13075-021-02498-8] [Medline: 33849654]
47. Knitza J, Muehlensiepen F, Ignatyev Y, et al. Patient's perception of digital symptom assessment technologies in rheumatology: results from a multicentre study. Front Public Health 2022;10. [doi: 10.3389/fpubh.2022.844669] [Medline: 637486321]
48. Knitza J, Tascilar K, Fuchs F, et al. Diagnostic accuracy of a mobile AI-based symptom checker and a web-based self-referral tool in rheumatology: multicenter randomized controlled trial. J Med Internet Res 2024 Jul 23;26:e55542. [doi: 10.2196/55542] [Medline: 39042425]
49. Krusche M, Callhoff J, Knitza J, Ruffer N. Diagnostic accuracy of a large language model in rheumatology: comparison of physician and ChatGPT-4. Rheumatol Int 2024 Feb;44(2):303-306. [doi: 10.1007/s00296-023-05464-6] [Medline: 37742280]
50. Li KY, Kim PS, Thariath J, Wong ES, Barkham J, Kocher KE. Standard nurse phone triage versus tele-emergency care pilot on Veteran use of in-person acute care: an instrumental variable analysis. Acad Emerg Med 2023 Apr;30(4):310-320. [doi: 10.1111/acem.14681] [Medline: 36757685]
51. Lowe C, Browne M, Marsh W, Morrissey D. Usability testing of a digital assessment routing tool for musculoskeletal disorders: iterative, convergent mixed methods study. J Med Internet Res 2022 Aug 30;24(8):e38352. [doi: 10.2196/38352] [Medline: 36040787]
52. Lowe C, Sephton R, Marsh W, Morrissey D. Evaluation of a musculoskeletal Digital Assessment Routing Tool (DART): crossover noninferiority randomized pilot trial. JMIR Form Res 2024 Jul 30;8:e56715. [doi: 10.2196/56715] [Medline: 39078682]
53. Lundberg K, Qin L, Aulin C, van Spil WE, Maurits MP, Knevel R. Population-based user-perceived experience of Rheumatic?: a novel digital symptom-checker in rheumatology. RMD Open 2023 Apr;9(2):e002974. [doi: 10.1136/rmdopen-2022-002974] [Medline: 37094982]
54. Martin MJ, Payne KM. Using digital technology and user-centred design to develop a physiotherapy self-referral service for back pain. Physiotherapy 2020 May;107(Suppl 1):e139-e140. [doi: 10.1016/j.physio.2020.03.203]
55. Phillips CJ, Phillips Nee Buck R, Main CJ, et al. The cost effectiveness of NHS physiotherapy support for occupational health (OH) services. BMC Musculoskelet Disord 2012 Feb 23;13(1):29. [doi: 10.1186/1471-2474-13-29] [Medline: 22361319]
56. Qin L, Zegers F, Selani D, et al. Differentiation of immune mediated versus non immune mediated rheumatic diseases by online symptom checker in real-world patients—multiple diagnoses and particularly fibromyalgia is a stumbling block. Ann Rheum Dis 2024 Jun;83(Suppl 1):2082-2083. [doi: 10.1136/annrheumdis-2024-eular.5438] [Medline: 644868620]
57. Ryan K, Grinbergs P. Demographic analysis of users of a musculoskeletal physiotherapy self-referral digital triage tool in Bromley. Physiotherapy 2024 Jun;123:e210-e211. [doi: 10.1016/j.physio.2024.04.263]
58. Salisbury C, Montgomery AA, Hollinghurst S, et al. Effectiveness of PhysioDirect telephone assessment and advice services for patients with musculoskeletal problems: pragmatic randomised controlled trial. BMJ 2013 Jan 29;346(7893):f43. [doi: 10.1136/bmj.f43] [Medline: 23360891]
59. Soin A, Hirschbeck M, Verdon M, Manchikanti L. A pilot study implementing a machine learning algorithm to use artificial intelligence to diagnose spinal conditions. Pain Physician 2022 Mar;25(2):171-178. [Medline: 35322974]
60. Tan T, Santosa A, Roslan N, Li J. The development of an AI-based conversational agent for screening of rheumatic diseases [Abstract]. Int J Rheum Dis 2023;26(9). [doi: 10.1111/1756-185X.14505]
61. Trivedi SV, Batta R, Henao-Romero N, Mondal P, Wilson T, Stempien J. A comparison of self-triage tools to nurse driven triage in the emergency department. PLOS ONE 2024;19(8):e0297321. [doi: 10.1371/journal.pone.0297321] [Medline: 39196994]

XSL•FO
RenderX

62. Edwards J, Hayden J, Asbridge M, Gregoire B, Magee K. Prevalence of low back pain in emergency settings: a systematic review and meta-analysis. BMC Musculoskelet Disord 2017 Apr 4;18(1):143. [doi: 10.1186/s12891-017-1511-7] [Medline: 28376873]
63. Lowe C, Atherton L, Lloyd P, Waters A, Morrissey D. Improving safety, efficiency, cost, and satisfaction across a musculoskeletal pathway using the digital assessment routing tool for triage: quality improvement study. J Med Internet Res 2025 Apr 25;27:e67269. [doi: 10.2196/67269] [Medline: 40279646]
64. Burgess R, Tucker K, Smithson R, Dimbleby P, Casey C. Optimising musculoskeletal patient flow through digital triage and supported self-management: a service evaluation set within community musculoskeletal care. Musculoskelet Care 2024 Dec;22(4):e70013. [doi: 10.1002/msc.70013] [Medline: 39625285]
65. Xu H, Shuttleworth KMJ. Medical artificial intelligence and the black box problem: a view based on the ethical principle of “do no harm”. Intell Med 2024 Feb;4(1):52-57. [doi: 10.1016/j.imed.2023.08.001]
66. Mruthyunjaya P, Verma S, Agarwal A, Maharana U, Mandal M, Ahmed S. Right diagnoses but wrong reasoning: current large-language model-based agentic frameworks have flawed clinical reasoning despite high diagnostic accuracy. The Lancet. Preprint posted online on Jul 9, 2025. [doi: 10.2139/ssrn.5339074]
67. Kremer P, Schiebisch H, Lechner F, et al. Comparative analysis of large language models and traditional diagnostic decision support systems for rare rheumatic disease identification. EULAR Rheumatol Open 2025 Jun;1(2):51-59. [doi: 10.1016/j.ero.2025.04.007]
68. Zegers F, Qin L, Selani D, et al. POS1131 prediction models for rheumatic diseases: from clinical simplicity to data-driven complexity with patient-reported symptoms for an online symptom checker. Ann Rheum Dis 2025 Jun;84:1211-1212. [doi: 10.1016/j.ard.2025.06.481] [Medline: 19279015]
69. Segur-Ferrer J, Moltó-Puigmartí C, Pastells-Peiró R, Vivanco-Hidalgo RM. Methodological frameworks and dimensions to be considered in digital health technology assessment: scoping review and thematic analysis. J Med Internet Res 2024 Apr 10;26:e48694. [doi: 10.2196/48694] [Medline: 38598288]
70. Triage guidelines for orthopaedic optimisation pathway (based on musculoskeletal (MSK) referral) V6.0. South East London Integrated Care Board. 2024. URL: https://www.selondonics.org/wp-content/uploads/MSK-Triage-Guidelines-v6-Final.pdf [accessed 2025-07-14]
71. Fraenkel L, Bathon JM, England BR, et al. 2021 American College of Rheumatology Guideline for the treatment of rheumatoid arthritis. Arthritis Care Res (Hoboken) 2021 Jul;73(7):924-939. [doi: 10.1002/acr.24596] [Medline: 34101387]
72. Adus S, Macklin J, Pinto A. Exploring patient perspectives on how they can and should be engaged in the development of artificial intelligence (AI) applications in health care. BMC Health Serv Res 2023 Oct 26;23(1):1163. [doi: 10.1186/s12913-023-10098-2] [Medline: 37884940]
73. Lau BHF, Lafave MR, Mohtadi NG, Butterwick DJ. Utilization and cost of a new model of care for managing acute knee injuries: the Calgary Acute Knee Injury Clinic. BMC Health Serv Res 2012 Dec 5;12(1):445. [doi: 10.1186/1472-6963-12-445] [Medline: 23216946]
74. Lervik LCN, Vasseljen O, Austad B, et al. SupportPrim—a computerized clinical decision support system for stratified care for patients with musculoskeletal pain complaints in general practice: study protocol for a randomized controlled trial. Trials 2023 Apr 11;24(1):267. [doi: 10.1186/s13063-023-07272-6] [Medline: 37041631]

Abbreviations

AI: artificial intelligence
DART: Digital Assessment Routing Tool
ED: emergency department
LLM: large language model
mHealth: mobile health
OSF: Open Science Framework
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Search
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
TRL: technology readiness level

XSL•FO
RenderX

Edited by S Brini; submitted 31.Jul.2025; peer-reviewed by J Knitza; revised version received 08.Dec.2025; accepted 09.Dec.2025; published 14.Jan.2026.

Please cite as:
Truong LK, Wrightson JG, Vincent R, Lui E, Couch JL, Wang E, Starcevich C, Giustini D, Haagaard A, Lopatina E, van Berkel N, Rathleff MS, Ardern CL
Evidence for Digital Health Tools Designed to Support the Triage of Musculoskeletal Conditions in Primary, Urgent, and Emergency Care Settings: Scoping Review
J Med Internet Res 2026;28:e81578
URL: https://www.jmir.org/2026/1/e81578
doi:10.2196/81578

© Linda K Truong, James G Wrightson, Raphaël Vincent, Eunice Lui, Jamon L Couch, Ellen Wang, Cobie Starcevich, Dean Giustini, Alex Haagaard, Elena Lopatina, Niels van Berkel, Michael Skovdal Rathleff, Clare L Ardern. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Diagnostic Value of Image-Based Machine Learning for Osteoporosis: Systematic Review and Meta-Analysis

Rui Zhao[1], MM; Haolin Yang[1], MM; Yangbo Li[1], MM; Xiaoyun Li[1], MM; Zhijie Yang[1], MM; Yanping Lin[2], MM; Jiachun Huang[2], MD; Lei Wan[2], MD; Hongxing Huang[2], MD

[1]The Third Clinical College of Medicine, Guangzhou University of Traditional Chinese Medicine, Guangzhou, China

[2]The Third Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou University of Chinese Medicine, 261 Longxi Avenue, Liwan District, Guangzhou, China

Corresponding Author:
Hongxing Huang, MD
The Third Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou University of Chinese Medicine, 261 Longxi Avenue, Liwan District, Guangzhou, China

Abstract

Background: Osteoporosis (OP) is projected to be a major issue significantly impacting the well-being of middle-aged and old populations. Machine learning (ML) and deep learning (DL) models developed based on medical imaging have enhanced clinicians' diagnostic accuracy and work efficiency. However, the diagnostic performance of different types of medical imaging for OP has not been systematically assessed.

Objective: By summarizing related literature, this study aims to elucidate the role of DL models based on different medical imaging modalities in OP detection.

Methods: PubMed, Embase, the Cochrane Library, and Web of Science were systematically searched for studies using ML for the diagnosis of OP based on medical imaging. The final search was conducted on May 16, 2024. The risk of bias in the included studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 tool. A bivariate mixed-effects model was applied to perform meta-analyses of sensitivity (SEN) and specificity (SPC), stratified by imaging modality (x-ray, computed tomography [CT], magnetic resonance imaging [MRI]). In addition, subgroup analyses were carried out based on the type of ML algorithm, the method of validation dataset generation, and the anatomical site of assessment.

Results: A total of 60 studies comprising 66,195 participants were encompassed in this systematic review and meta-analysis. Among these, 22 studies used x-ray imaging, 37 applied CT imaging, and 3 used MRI for ML-based OP diagnosis. For x-ray–based models, the pooled SEN and SPC for studies focusing on the appendicular skeleton were 0.97 (95% CI 0.83 - 0.99) and 0.90 (95% CI 0.75 - 0.96), respectively. For studies using the mandible as the target site, SEN and SPC were 0.94 (95% CI 0.89 - 0.97) and 0.80 (95% CI 0.56 - 0.93), respectively. For those focusing on the lumbar spine, the pooled SEN and SPC were 0.87 (95% CI 0.77 - 0.93) and 0.82 (95% CI 0.75 - 0.87), respectively. For CT-based models, studies targeting the hip joint reported a pooled SEN and SPC of 0.87 (95% CI 0.83 - 0.90) and 0.92 (95% CI 0.81 - 0.96), respectively. For the thoracic spine, SEN and SPC were 0.91 (95% CI 0.86 - 0.94) and 0.94 (95% CI 0.92 - 0.95), respectively, while for the lumbar spine, they were 0.91 (95% CI 0.87 - 0.94) and 0.92 (95% CI 0.86 - 0.95), respectively.

Conclusions: ML based on medical imaging demonstrates high diagnosis accuracy for OP, particularly DL models using x-ray and CT modalities. However, this study included only a limited number of original studies using MRI-based ML, and there remains a lack of adequate external validation across studies, which poses interpretative limitations. Future research should aim to develop artificial intelligence tools with broader applicability and enhanced diagnostic precision.

(J Med Internet Res 2026;28:e75965) doi:10.2196/75965

KEYWORDS

osteoporosis; machine learning; artificial intelligence; systematic review; diagnostic imaging

Introduction

Osteoporosis (OP), a metabolic disorder, features a systemic reduction in bone mass and impaired bone microarchitecture and elevates the risk of fragility fractures. As the most prevalent chronic metabolic bone disease, it is strongly associated with advancing age, posing significant health threats. However, due to its insidious onset, prolonged disease course, and challenges in treatment, public awareness and attention toward OP prevention and management remain insufficient [1,2]. With the emerging global trend of population aging, OP is projected to become a major issue adversely affecting the quality of life of

XSL•FO
RenderX

middle-aged and older people. Epidemiological studies estimate that by 2050, the global population at high risk of fractures will surge to 6.26 million from 1.66 million in 1990 [3]. This escalation imposes immense social pressures and substantial economic burdens on early OP screening, prevention, and treatment.

At present, a variety of diagnostic methods are available for the clinical assessment of OP. Among them, dual-energy x-ray absorptiometry (DXA) for measuring *T* scores recommended by the World Health Organization is regarded as the authoritative and standardized technique [4]. Although DXA is widely used, it is unable to assess whole-body skeletal, fat, and lean mass, which restricts its utility in the routine diagnosis or evaluation of OP [5]. Moreover, due to disparities in socioeconomic development across different regions worldwide, DXA is not accessible in underdeveloped countries and regions. Therefore, some high-risk populations, such as postmenopausal women and older adults, are not detected and untreated. Medical imaging is crucial in clinical diagnosis and treatment. However, the hidden features within imaging techniques including x-rays, computed tomography (CT), and magnetic resonance imaging (MRI) are often overlooked due to low spatial resolution and high contrast resolution [6].

In the 1980s, computer-aided diagnosis (CAD) systems were developed to deeply interpret key features in medical images, providing radiologists with valuable insights into image interpretation [7]. Currently, CAD tools primarily include traditional machine learning (ML) models built on explainable clinical features and deep learning (DL) models developed using pathological or nuclear medicine images. They assist clinicians in disease diagnosis and prognostic prediction. Increasing evidence has demonstrated the utility of CAD in diagnosing conditions such as autism [8], pulmonary embolism [9], breast cancer [10], and bone metastases [11]. DL approaches based on medical imaging have attracted substantial research interest. Against this backdrop, ML models based on various imaging modalities such as x-rays, CT, and MRI have been constructed to diagnose OP [12]. However, the diagnostic performance of various imaging methods in OP is not supported by systematic evidence. This hindered the application of artificial intelligence (AI)–based CAD tools in OP and posed challenges for further systematic development.

Therefore, our study seeks to provide a comprehensive review of DL research in the diagnosis of OP based on medical imaging modalities, including x-ray, CT, and MRI. Furthermore, this study aims to analyze and evaluate the feasibility and accuracy of AI-driven DL in enhancing the screening and diagnostic rates of OP, thereby offering robust support for the prevention and management of the disease.

Methods

Study Registration

This study followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines and was prospectively registered on PROSPERO (CRD42024567736).

Eligibility Criteria

The eligible studies were (1) case-control, cohort, or cross-sectional studies; (2) papers with comprehensively developed image-based DL models for OP diagnosis; and (3) English publications. The following studies were excluded: (1) studies that only developed traditional ML models, (2) those that performed image segmentation without a complete DL model, and (3) those lacking outcome measures for evaluating the DL model's accuracy. Outcome measures must include at least 1 of the following: c-statistic, sensitivity (SEN), specificity (SPC), accuracy, recall, precision, confusion matrix, F_1-score, or calibration curve.

Data Sources and Search Strategy

PubMed, Cochrane, Embase, and Web of Science databases were thoroughly retrieved up to May 16, 2024. Both MeSH and free-text terms were used without restrictions on geographic location or study type. The search strategy is detailed in Multimedia Appendix 1.

Study Selection

The retrieved literature was uploaded to EndNote (Thomson Corporation), and duplicates were ostracized. Titles and abstracts were reviewed to identify potentially eligible studies. Full-text papers were subsequently screened to determine the eligible ones. Two researchers (RZ and HY) independently conducted the literature screening and cross-checked their results. Dissents were addressed by a third researcher (YL).

Data Extraction

The eligible papers were imported into EndNote, and data extraction was performed. A standard electronic data extraction form was developed beforehand to capture the following information: title, DOI, first author, publication year, author's country, study type, patient source, OP diagnosis criteria, medical imaging, background population, gender, age, use of image segmentation, number of OP cases, total cases, number of OP and total cases in the training or validation set, validation set generation method, model type, and comparison with clinical practitioners. Data were independently extracted by 2 researchers (RZ and HY), followed by cross-checking. There was a high level of agreement between the 2 reviewers in the screening process (Cohen κ=0.879). In cases of disagreement, a third reviewer (YL) would assist in addressing it.

Risk of Bias in Studies

The bias of risk in the eligible studies was assessed via Quality Assessment of Diagnostic Accuracy Studies-2, a tool for evaluating the collation risk of bias and clinical applicability of original diagnostic studies [13]. Quality Assessment of Diagnostic Accuracy Studies-2 covers 5 domains: case selection, trials to be evaluated, reference standard, case flow, and progress, with each involving a few specific questions. The answer of "Yes," "No," or "Uncertain" corresponds to a low, high, or uncertain risk of bias. The risk of bias was deemed low if all of the landmark questions within a range were answered with "Yes"; if one of the informative questions was answered with "No," bias may exist, and the evaluators must determine the risk of bias in line with the established guidelines. The risk

XSL•FO
RenderX

of bias must be judged by the evaluation authors as per the established criteria. An unclear risk indicated that the studies reported sufficient details. Therefore, evaluators could not make a definitive judgment.

The risk of bias in studies was independently conducted by 2 researchers (RZ and HY), followed by cross-checking. If any dissent arose, a third researcher (YL) would assist in addressing it.

Synthesis Methods

The meta-analysis was carried out via a bivariate mixed-effects model based on diagnostic 2×2 contingency tables. However, some of the original studies did not provide complete 2×2 diagnostic data. In such cases, the necessary information was derived using SEN, SPC, positive predictive value, negative predictive value, and accuracy, in conjunction with the corresponding sample sizes. The meta-analysis reported pooled estimates of SEN, SPC, positive likelihood ratio (PLR), negative likelihood ratio (NLR), diagnostic odds ratio (DOR), and the summary receiver operating characteristic (SROC) curve along with their corresponding 95% CIs. Publication bias across studies was assessed through Deeks' funnel plot, while the clinical utility of the predictive models was evaluated via Fagan's nomogram. Subgroup analyses were performed based on imaging modality (x-ray, CT, and MRI), modeling approach (traditional ML vs DL), and validation strategy. It is important to note that the bivariate mixed-effects model requires a minimum of four 2×2 diagnostic tables. As the ML models based on MRI images only provided 3 such tables, a narrative synthesis was performed for this subgroup instead. A 2-sided *P* value of <.05 denoted statistical significance.

Results

Study Registration

A total of 3427 papers were retrieved, including 685 from PubMed, 15 from Cochrane, 1942 from Embase, and 785 papers from Web of Science. Among them, 639 papers were duplicates and were excluded. After the title and abstract review, 2587 studies unrelated to the study topic were removed. Full texts of the rest were subsequently reviewed. In total, 23 conference abstracts without full-text publications and 68 that did not include medical imaging in the modeling process were ostracized. Ultimately, 60 studies were included in the analysis (Figure 1) [14-73]. This study was conducted in accordance with the PRISMA 2020 checklist (Checklist 1).

XSL•FO
RenderX

Figure 1. Literature screening process. RCT: randomized controlled trial.

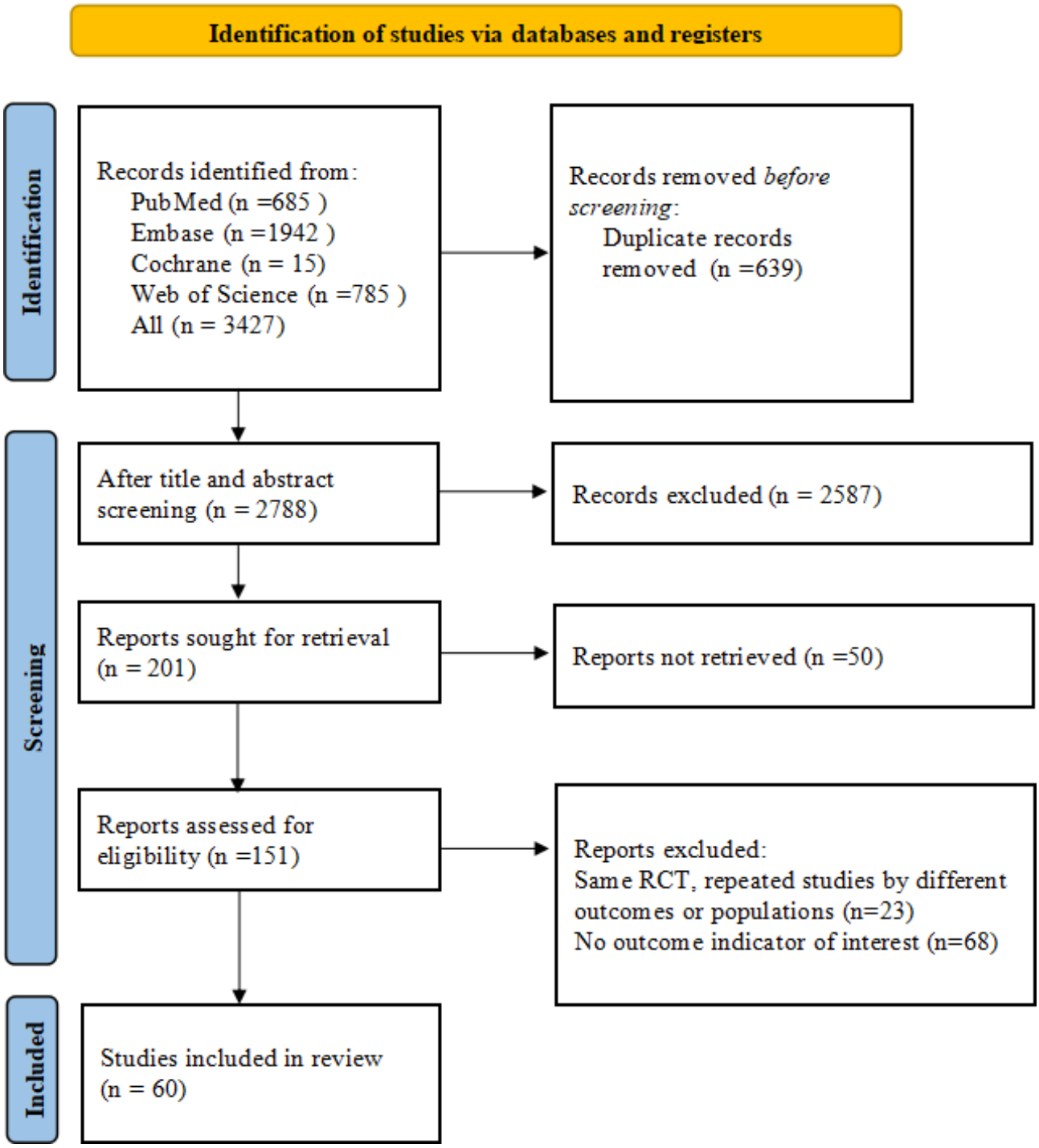

Study Characteristics

Among the 60 studies included in our analysis, 55 were case-control studies [14-21,25-35,37-67,69-73], and 5 were cohort studies [22-24,36,68]. These studies were predominantly published between 2012 and 2024 and involved 66,195 cases. These studies were published in 11 countries, including China (n=27), South Korea (n=10), the United States (n=8), and Japan (n=5) [15-17,20,22,23,26-33,35-48,50,51,53-60,62-73]. A smaller number of studies were from India (n=3), Saudi Arabia (n=2), Jordan (n=1), Latvia (n=1), Malaysia (n=1), Poland (n=1), and Switzerland (n=1) [14,18,19,21,24,25,34,39,52,61]. In total, 57 studies reported patient sources, of which 42 were single-center studies [14,15,17,19-23,26,29,31,32,36-42,47,48,50-60,62,64-67,69-73], 12 were multicenter studies [16,27,28,30,33,34,43-45,61,63,68], and 4 used database sources [24,41,46,49]. In terms of OP diagnosis, 47 studies explicitly provided diagnosis criteria [16,17,20-23,25,27-33,35-42,44,45,47,50-58,60-64,66-73]. Regarding medical imaging, 37 studies developed CT-based imaging models [15,16,18,19,21-23,25,26,29,31-33,38-41,43-45,47,48,50-54,56-58,65,67-69,72,73], 22 developed x-ray–based models [14,17,20,24,27,28,30,34-37,42,46,49,55,59-63,70,71], and 3 focused on MRI-based models [45,64,66]. Concerning the population, 4 studies specifically examined postmenopausal

XSL•FO
RenderX

women [16,35,36,52,71], and 1 study focused on men aged 50 years and older [16]. In terms of image processing, 48 studies used manual segmentation techniques to define regions of interest for analysis [14-19,21-23,25-36,38,39,41,42,44,45,47,49,52-57,59-66,68,70-73], while 12 did not define regions of interest [20,24,37,40,43,46,48,50,51,58,67,69]. Regarding the skeletal parts, 33 studies used lumbar vertebrae images [15-17,20,22,23,26,29-32,36,38-41,45,47,48,50,53-55,57,62,64-68,70,71,73], 9 used thoracic vertebrae images [23,25,31,38,44,48,53,55,56], 10 used hip images (including femoral neck, femoral head, and pelvis) [18,21,33,36,57,62,68,70,72], 7 used mandible images [17,36,42,52,60,63,71], and 10 used images of limb bones [14,28,34,35,43,51,58,59,61,69]. Regarding the generation of validation sets, 35 studies adopted random sampling [15-17,20-23,25,27,29-32,36,38,40-42,44,48,50-53,56,57,60-62,64,65,67,69,71,72], 12 used K-fold cross-validation [14,24,34,37,39,46,49,58,59,66,70,73], and 5 applied external validation [28,33,45,54,63]. In total, 9 studies compared their results with the screening results of clinicians [22,25,28,45-47,60,62,65]. In terms of model construction, 32 built DL models [18,21,23-31,33,37-42,44-47,52-56,59-62,66], and 28 constructed ML models [14-17,19,20,22,32,34-36,43,48-51,57,58,63-65,67-73] (Multimedia Appendix 2).

Risk of Bias in Studies

In all eligible studies, consecutive cases were included. Although most studies were case-control studies, 32 developed DL models, with variables derived from medical images. Therefore, these studies demonstrated a low risk of bias in case selection. In total, 28 studies applied ML models, where the process of variable generation might be influenced by the case-control study design, thereby leading to a higher risk of bias. Since this research is a meta-analysis of ML, whether or not the reference standards for OP diagnosis are known does not affect the results. Additionally, the criteria for determining positive results were pre-established, indicating that the trials under evaluation posed a low risk of bias. The implementation of a reference standard for OP diagnosis was considered reasonable, thus introducing a low risk of bias. Furthermore, there was a proper time interval between the trial and reference standard, and all patients in a given study followed the same diagnosis rules, with no cases omitted. Therefore, there was a low risk of bias in clinical applicability [14-73] (Figure 2 and Multimedia Appendix 3).

XSL•FO
RenderX

Figure 2. Risk of bias plot [14-73].

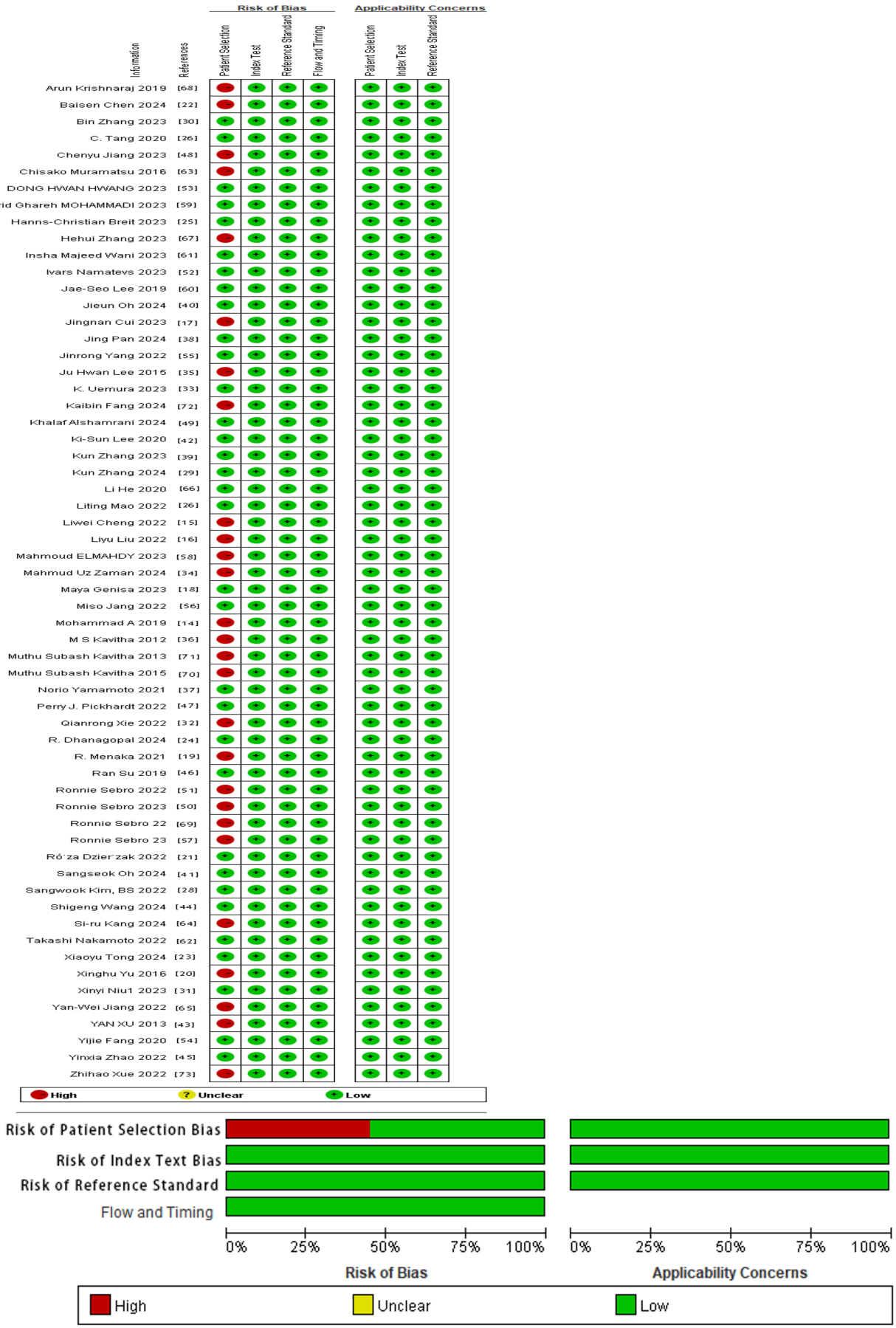

XSL•FO
RenderX

Meta-Analysis

ML Based on X-Ray

Synthesized Results

This validation set comprised 24 diagnostic 4-fold tables, which were used to verify the ML models based on x-rays for OP diagnosis. The results were summarized through the bivariate mixed-effects model. The pooled SEN, SPC, PLR, NLR, DOR, and SROC curves were 0.92 (95% CI 0.88 - 0.94), 0.83 (95% CI 0.76 - 0.88), 5.4 (95% CI 3.8 - 7.6), 0.10 (95% CI 0.07 - 0.15), 54 (95% CI 28 - 105), and 0.94 (95% CI 0.92 - 0.96), respectively (Figures 3 and 4) [14,17,20,24,27,28,30,34-37,42,46,49,55,59-63,70,71]. There was no discernible publication bias in the studies according to Deeks' funnel plot (Figure 5). In the included study participants, approximately 48.44% (n=6429) had OP. Assuming this as the prior probability, if the ML prediction result was OP, the actual probability of OP was .83. If the ML prediction result was non-OP, the actual probability of non-OP was .92 (Figure 6).

Figure 3. Forest plot of sensitivity and specificity for x-ray–based machine learning for diagnosing osteoporosis [14,17,20,24,27,28,30,34-36,42,46,49,56,60,62,63,70,71].

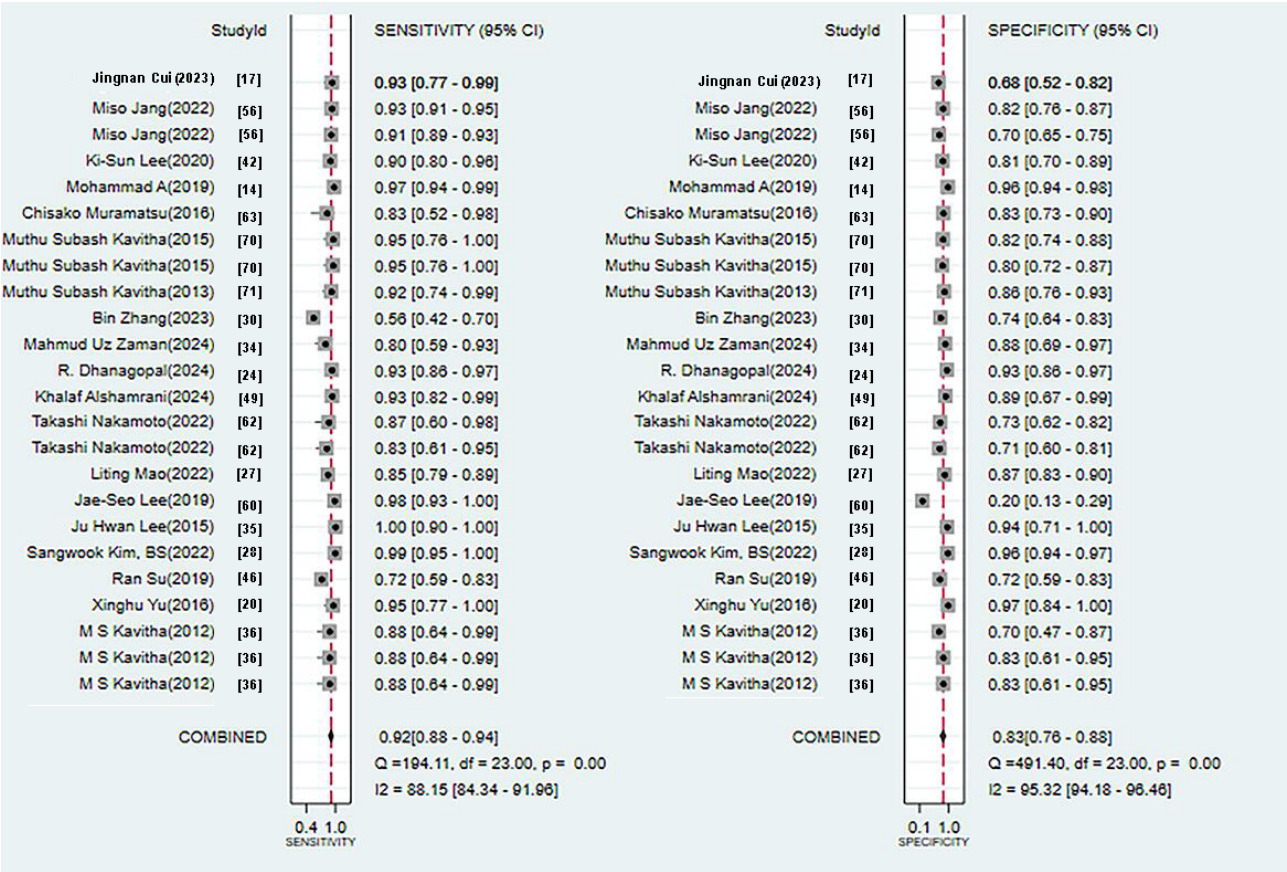

XSL•FO
RenderX

Figure 4. 95% Confidence contour for x-ray. AUC: Area underthe curve;SEN: Sensitivity; SPC: Specificity;SROC: Summary receiver operating characteristic.

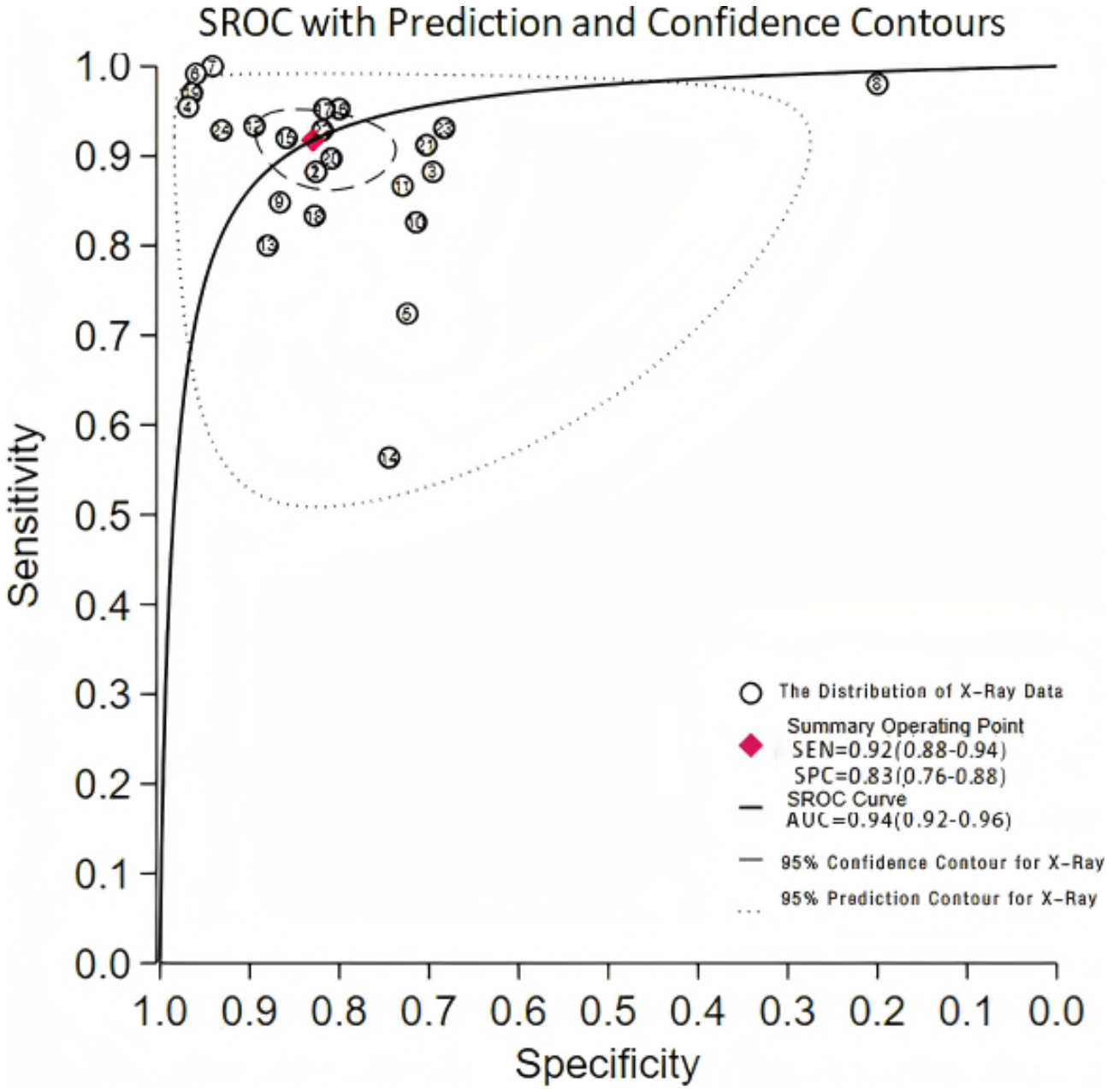

XSL•FO
RenderX

Figure 5. Diagnostic odds ratio for x-ray. ESS:Effective sample size.

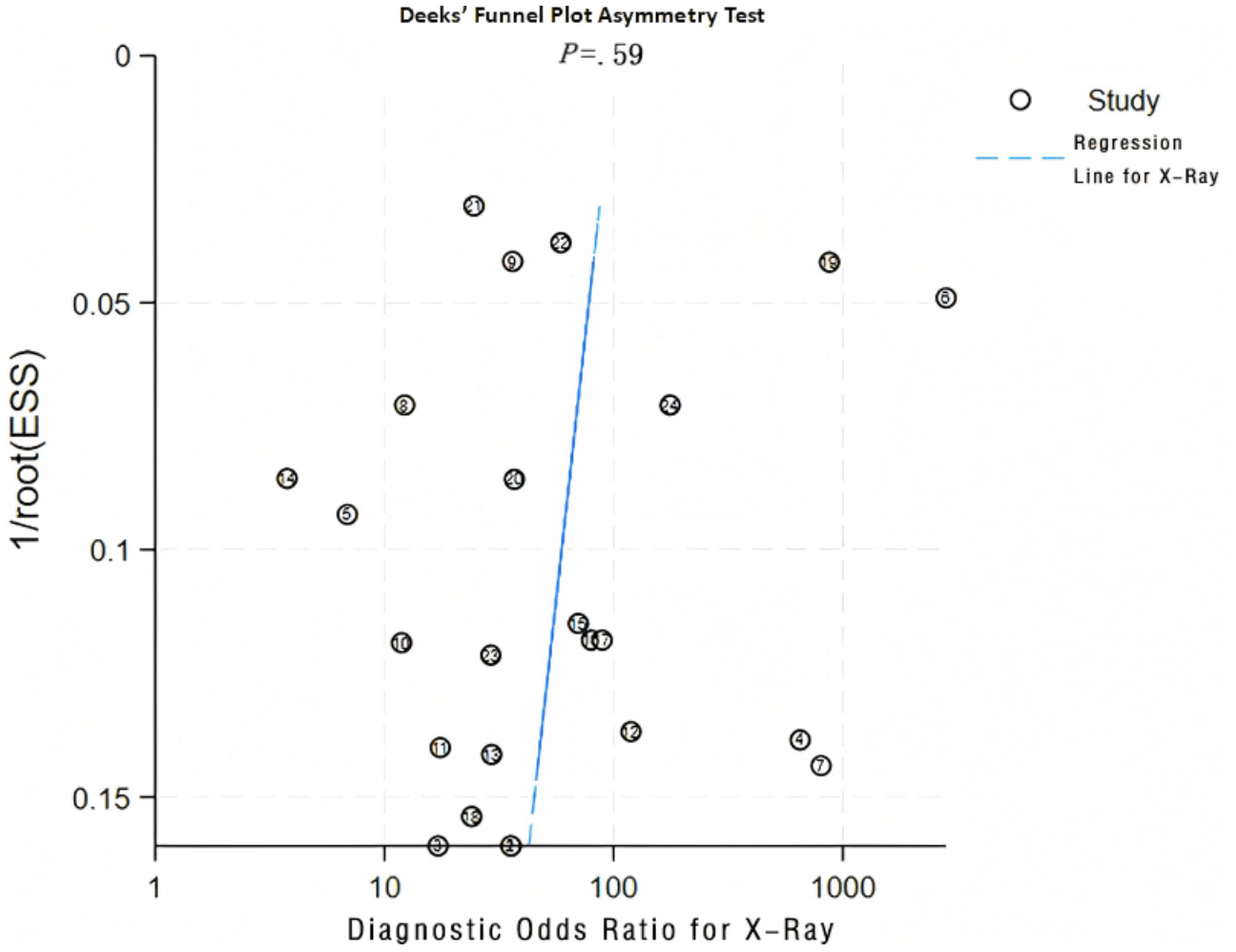

XSL•FO
RenderX

Figure 6. Pretest probability for x-ray. LR:Likelihood ratio.

XSL•FO
RenderX

Subgroup Analysis: Types of ML

Deep Learning

The validation set included 9 diagnostic 4-fold tables to assess the performance of DL models based on x-ray images for OP diagnosis. The results summarized from the bivariate mixed-effects model showed that SEN, SPC, PLR, NLR, DOR, and the SROC curve were 0.90 (95% CI 0.79 - 0.95), 0.79 (95% CI 0.62 - 0.89), 4.2 (95% CI 2.2 - 8.0), 0.13 (95% CI 0.06 - 0.29), 32 (95% CI 9 - 107), and 0.92 (95% CI 0.89 - 0.94), respectively (Figures S1 and S2 in Multimedia Appendix 4). Deeks' funnel plot revealed no marked publication bias (Figure S3 in Multimedia Appendix 4). In the encompassed studies, approximately 30% (n=2556) of the participants had OP. Therefore, assuming this as the prior probability, if the result from ML indicated OP, the actual probability of OP was .64. If the ML result indicated non-OP, the actual probability of non-OP was .95 (Figure S4 in Multimedia Appendix 4).

Traditional ML

The validation set encompassed 15 diagnostic 4-fold tables for validating the traditional ML models based on x-ray imaging for OP diagnosis. The bivariate mixed-effects model was leveraged. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.93 (95% CI 0.92 - 0.95), 0.85 (95% CI 0.79 - 0.89), 6.0 (95% CI 4.3 - 8.5), 0.08 (95% CI 0.06 - 0.10), 78 (95% CI 44 - 139), and 0.96 (95% CI 0.94 - 0.97), respectively (Figures S5 and S6 in Multimedia Appendix 4). Deeks' funnel plot showed no significant publication bias in studies (Figure S7 in Multimedia Appendix 4). Approximately 61% (n=3863) of the participants had OP. Therefore, assuming this as the prior probability, if the result from ML indicated OP, the actual probability of OP was .90. If the ML result indicated non-OP, the actual probability of non-OP was .81 (Figure S8 in Multimedia Appendix 4).

Generation Method of the Validation Set

K-Fold Cross-Validation

Among models constructed through x-ray for OP diagnosis, 7 diagnostic 4-fold tables used the K-fold cross-validation to generate the validation set. The results summarized by the bivariate mixed-effects model demonstrated that the SEN, SPC, PLR, NLR, DOR, and SROC curve were 0.90 (95% CI 0.83 - 0.95), 0.87 (95% CI 0.79 - 0.93), 7.2 (95% CI 4.0 - 12.7), 0.11 (95% CI 0.06 - 0.21), 64 (95% CI 20 - 204), and 0.95 (95% CI 0.93 - 0.96), respectively (Figures S9 and S10 in Multimedia Appendix 4). Deeks' funnel plot did not exhibit significant publication bias (Figure S11 in Multimedia Appendix 4). Among the participants in our included studies, approximately 42% (n=1287) had OP. Therefore, assuming this as the prior probability, if the ML result indicated OP, the actual probability of OP was .84. If the ML result indicated non-OP, the actual probability of non-OP was .92 (Figure S12 in Multimedia Appendix 4).

Random Sampling

In total, 14 diagnostic 4-fold tables used the random sampling method to generate the validation set. The results summarized by the bivariate mixed-effects model showed that the pooled SEN, SPC, PLR, NLR, DOR, and SROC curve were 0.90 (95% CI 0.84 - 0.93), 0.76 (95% CI 0.67 - 0.84), 3.8 (95% CI 2.7 - 5.4), 0.14 (95% CI 0.09 - 0.20), 28 (95% CI 16 - 48), and 0.91 (95% CI 0.88 - 0.93), respectively (Figures S13 and S14 in Multimedia Appendix 4). Significant publication bias was not noted in Deeks' funnel plot (Figure S15 in Multimedia Appendix 4). Among the participants in our included studies, approximately 60% (n=4049) had OP. Therefore, assuming this as the prior probability, if the ML result indicated OP, the actual probability of OP was .85. If the ML result showed non-OP, the actual probability of non-OP was .83 (Figure S16 in Multimedia Appendix 4).

Examination Parts

Limbs

In the OP diagnosis models constructed based on x-rays, 4 diagnostic 4-fold tables focused on the limb bones. The results summarized by the bivariate mixed-effects model showed a SEN of 0.97 (95% CI 0.83 - 0.99), SPC of 0.90 (95% CI 0.75 - 0.96), PLR of 9.6 (95% CI 3.5 - 25.9), NLR of 0.03 (95% CI 0.01 - 0.22), DOR of 277 (95% CI 20 - 3783), and the SROC curve of 0.98 (95% CI 0.96 - 0.99; Figures S17 and S18 in Multimedia Appendix 4). Deeks' funnel plot did not demonstrate significant publication bias (Figure S19 in Multimedia Appendix 4). In the included study participants, the proportion of OP cases was approximately 19% (n=1114). Assuming this as the prior probability, if the ML result indicated OP, the actual probability of OP was .69. If the ML result indicated non-OP, the actual probability of non-OP was <.001 (Figure S20 in Multimedia Appendix 4).

Mandible

In total, 6 diagnostic 4-fold tables focused on the mandible. The bivariate mixed-effects model was used. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.94 (95% CI 0.89 - 0.97), 0.80 (95% CI 0.56 - 0.93), 4.8 (95% CI 1.9 - 12.1), 0.07 (95% CI 0.04 - 0.14), 69 (95% CI 20 - 241), and 0.96 (95% CI 0.94 - 0.97), respectively (Figures S21 and S22 in Multimedia Appendix 4). Deeks' funnel plot indicated no significant publication bias (Figure S23 in Multimedia Appendix 4). In all included study participants, the proportion of OP cases was approximately 42% (n=1153). Assuming this as the prior probability, if the ML result indicated OP, the actual probability of OP was .78. If the ML result indicated non-OP, the actual probability of non-OP was .95 (Figure S24 in Multimedia Appendix 4).

Lumbar Vertebrae

In total, 8 diagnostic 4-fold tables focused on the lumbar vertebrae. The bivariate mixed-effects model was used to summarize data. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.87 (95% CI 0.77 - 0.93), 0.82 (95% CI 0.75 - 0.87), 4.8 (95% CI 3.4 - 6.7), 0.16 (95% CI 0.08 - 0.30), 31 (95% CI 12 - 77), and 0.90 (95% CI 0.87 - 0.92), respectively (Figures S25 and S26 in Multimedia Appendix 4). Significant publication bias was not found in Deeks' funnel plot (Figure S27 in Multimedia Appendix 4). In the included study participants, the proportion of OP cases was approximately 32% (n=1281). Assuming this as the prior probability, if the ML

XSL•FO
RenderX

result indicated OP, the actual probability of having OP was .69. If the ML result indicated non-OP, the actual probability of non-OP was .93 (Figure S28 in Multimedia Appendix 4).

ML Based on CT

Synthesized Results

The validation set consisted of 24 diagnostic 4-fold tables for validating CT-based ML models for diagnosing OP. The bivariate mixed-effects model was used to pool data. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.91 (95% CI 0.89 - 0.93), 0.92 (95% CI 0.89 - 0.94), 11.6 (95% CI 8.5 - 9.7), 0.09 (95% CI 0.07 - 0.12), 123 (95% CI 80 - 90), and 0.97 (95% CI 0.53 - 1.00), respectively (Figures 7 and 8) [15,16,18,19,21-23,25,26,29,31-33,38-41,43-45,47,48,50-54,56-58,65,67-69,72,73]. According to Deeks' funnel plot, there was no significant publication bias (Figure 9). Among the included research participants, the proportion of individuals with OP was approximately 50% (n=10,995). Therefore, assuming this as the prior probability, if the ML models predicted OP, the actual probability of OP was .92. If the ML models predicted no OP, the actual probability of non-OP was .91 (Figure 10).

Figure 7. Forestplot of sensitivity and specificity for computed tomography-based machine learning for diagnosing osteoporosis [15,16,18,19,21-23,25,26,29,31-33,38-41,43-45,47,48,50-54,56-58,65,67-69,72,73].

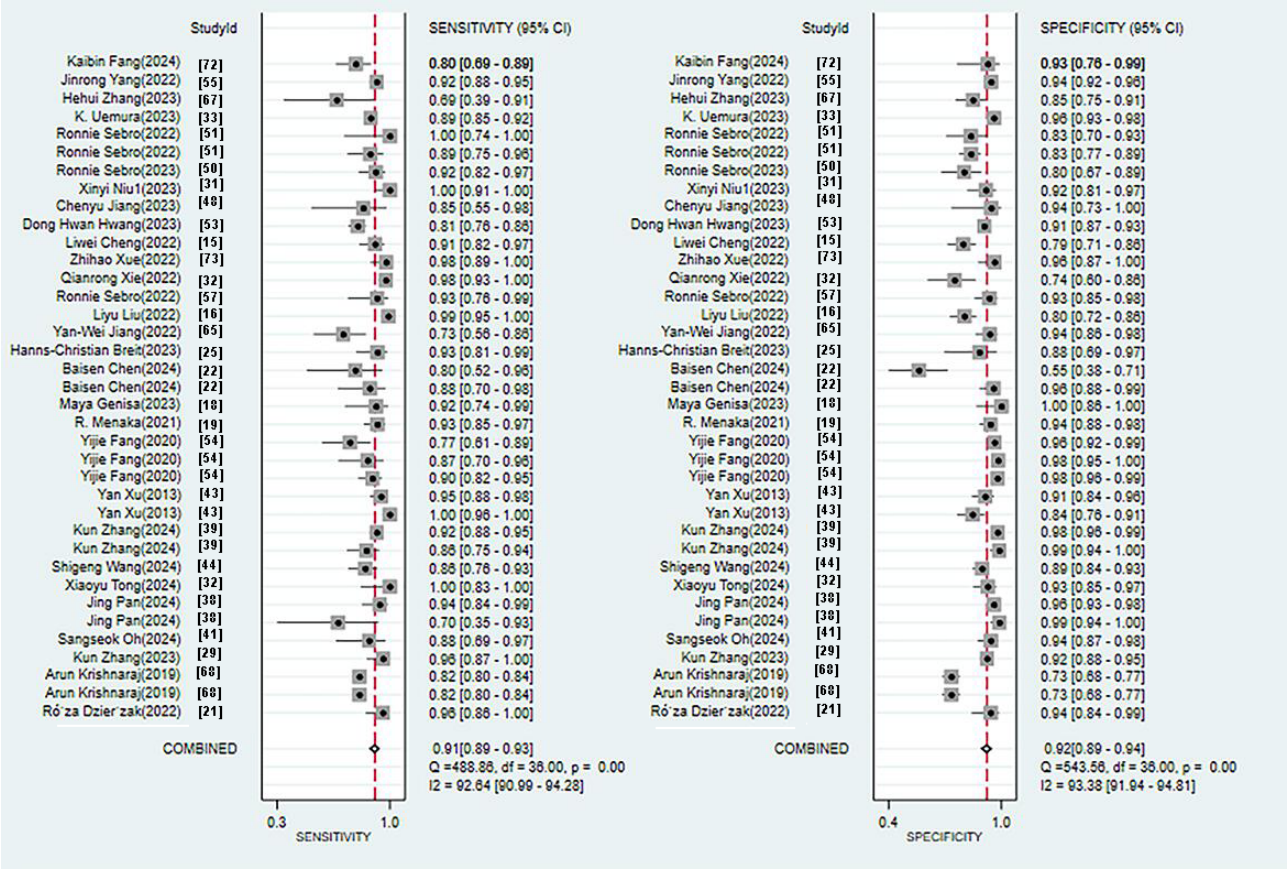

XSL•FO
RenderX

Figure 8. 95% Confidence contour for CT.AUC:Area under the curve;CT:Computed tomography;SEN: Sensitivity; SPC: Specificity;SROC: Summary receiver operating characteristic.

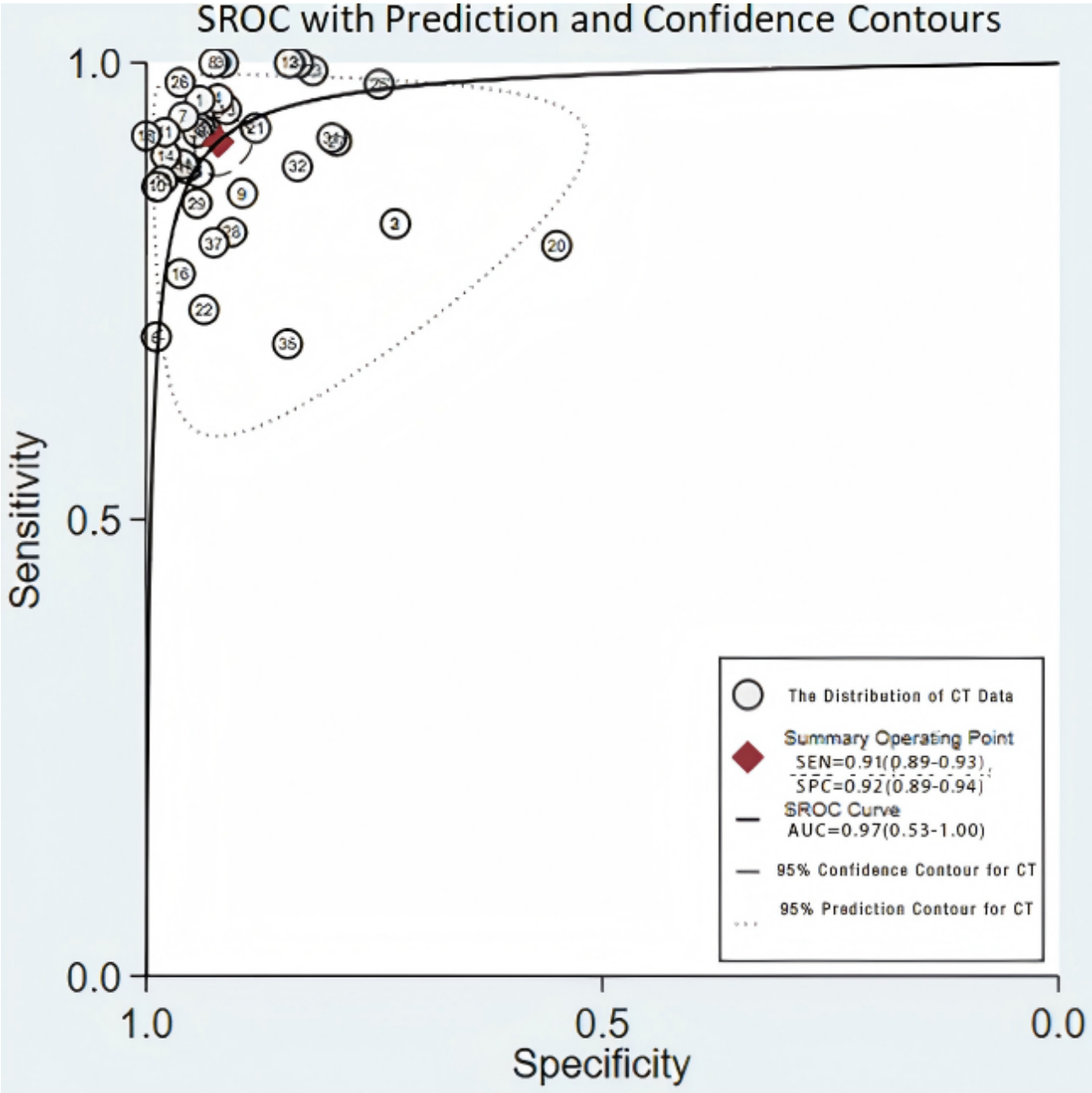

XSL•FO
RenderX

Figure 9. Diagnostic odds ratio for CT .CT: Computed tomography;ESS:Effective sample size.

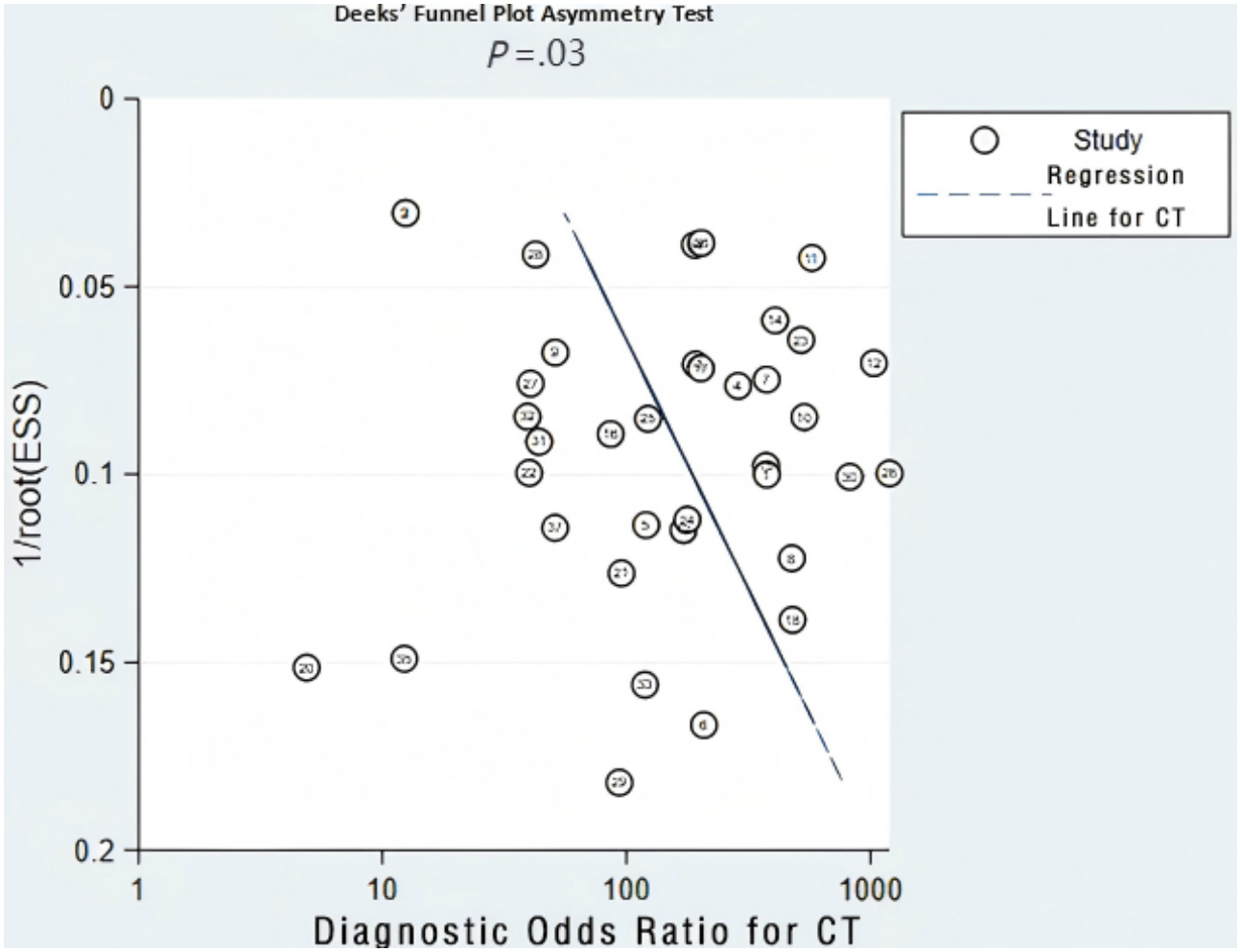

XSL•FO
RenderX

Figure 10. Pretest probability for CT. CT:Computed tomography;LR:Likelihood ratio.

XSL•FO
RenderX

Subgroup Analysis: Types of ML

Deep Learning

In the validation set, there were 15 diagnostic 4-fold tables for validating the CT-based DL models for diagnosing OP. The bivariate mixed-effects model was used. The pooled SEN, SPC, PLR, NLR, DOR, and SROC curve were 0.91 (95% CI 0.88 - 0.94), 0.94 (95% CI 0.92 - 0.96), 16.3 (95% CI 11.9 - 22.3), 0.09 (95% CI 0.06 - 0.13), 178 (95% CI 106 - 299), and 0.98 (95% CI 0.96 - 0.99), respectively (Figures S29 and S30 in Multimedia Appendix 4). Deeks' funnel plot exhibited no marked publication bias (Figure S31 in Multimedia Appendix 4). Among the research participants included, the proportion of individuals with OP was approximately 32% (n=3197). Therefore, assuming this as the prior probability, if the ML models predicted OP, the actual probability of OP was .88. If the ML models predicted no OP, the actual probability of non-OP was .96 (Figure S32 in Multimedia Appendix 4).

Traditional ML

In the validation set, there were 15 diagnostic 4-fold tables for validating traditional ML models based on CT for diagnosing OP. The bivariate mixed-effects model was used. The pooled SEN, SPC, PLR, NLR, DOR, and SROC curve were 0.92 (95% CI 0.88 - 0.95), 0.85 (95% CI 0.77 - 0.90), 6.1 (95% CI 4.0 - 9.4), 0.09 (95% CI 0.06 - 0.15), 67 (95% CI 35 - 128), and 0.95 (95% CI 0.93 - 0.97), respectively (Figures S33 and S34 in Multimedia Appendix 4). Deeks' funnel plot did not show notable publication bias (Figure S35 in Multimedia Appendix 4). Among the research participants, the proportion of individuals with OP was approximately 60% (n=6486). Therefore, assuming this as the prior probability, if the ML models predicted OP, the actual probability of OP was .90. If the ML models predicted no OP, the actual probability of non-OP was .88 (Figure S36 in Multimedia Appendix 4).

Validation Set Generation Method

External Validation

In the OP diagnosis models constructed based on CT, validation sets for 5 diagnostic 4-fold tables were generated through external validation. The bivariate mixed-effects model was leveraged to pool data. The pooled SEN, SPC, PLR, NLR, DOR, and SROC curve were 0.88 (95% CI 0.85 - 0.91), 0.97 (95% CI 0.96 - 0.98), 28.4 (95% CI 20.4 - 39.7), 0.12 (95% CI 0.10 - 0.16), 229 (95% CI 148 - 355), and 0.98 (95% CI 0.96 - 0.99), respectively (Figures S37 and S38 in Multimedia Appendix 4). Deeks' funnel plot indicated no discernible publication bias (Figure S39 in Multimedia Appendix 4). Among the included study participants, approximately 31% (n=1590) had OP. Assuming this as the prior probability, if ML predicted OP, the actual probability of OP was .93. Conversely, if ML predicted non-OP, the actual probability of non-OP was .95 (Figure S40 in Multimedia Appendix 4).

Random Sampling

Validation sets for 24 diagnostic 4-fold tables were generated using the random sampling method. The bivariate mixed-effects model was leveraged. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.91 (95% CI 0.87 - 0.94), 0.90 (95% CI 0.85 - 0.94), 9.4 (95% CI 6.0 - 14.9), 0.10 (95% CI 0.07 - 0.14), 96 (95% CI 57 - 161), and 0.96 (95% CI 0.94 - 0.97), respectively (Figures S41 and S42 in Multimedia Appendix 4). Deeks' funnel plot presented no significant publication bias (Figure S43 in Multimedia Appendix 4). Among the included study participants, approximately 36% (n=4175) had OP. Given this as the prior probability, when the ML models predicted OP, the actual probability of OP was .84. On the other hand, when the ML models predicted non-OP, the actual probability of non-OP was .95 (Figure S44 in Multimedia Appendix 4).

Examination Parts

Hip Joint

In the OP diagnostic models constructed based on CT, 6 diagnostic 4-fold tables focused on the hip joint. The bivariate mixed-effects model was used. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.87 (95% CI 0.83 - 0.90), 0.92 (95% CI 0.81 - 0.96), 10.4 (95% CI 4.4 - 24.7), 0.14 (95% CI 0.10 - 0.19), 76 (95% CI 24 - 239), and 0.92 (95% CI 0.90 - 0.94), respectively (Figures S45 and S46 in Multimedia Appendix 4). Deeks' funnel plot did not show any marked publication bias (Figure S47 in Multimedia Appendix 4). Among the included study participants, approximately 69% (n=2719) had OP. Assuming this as the prior probability, if ML predicted OP, the actual probability of OP was .96. If the models predicted non-OP, the actual probability of OP was .77 (Figure S48 in Multimedia Appendix 4).

Thoracic Vertebrae

In total, 9 diagnostic 4-fold tables focused on the thoracic vertebrae. The bivariate mixed-effects model was leveraged to pool data. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.91 (95% CI 0.86 - 0.94), 0.94 (95% CI 0.92 - 0.95), 14.4 (95% CI 10.7 - 19.3), 0.10 (95% CI 0.06 - 0.15), 150 (95% CI 75 - 300), and 0.97 (95% CI 0.95 - 0.98), respectively (Figures S49 and S50 in Multimedia Appendix 4). Significant publication bias was not observed in Deeks' funnel plot (Figure S51 in Multimedia Appendix 4). Among the encompassed study participants, approximately 29% (n=2523) had OP. Assuming this as the prior probability, if ML predicted OP, the actual probability of OP was .85. If ML predicted non-OP, the actual probability of non-OP was .96 (Figure S52 in Multimedia Appendix 4).

Lumbar Vertebrae

For OP diagnostic models using the lumbar vertebrae as the target part, 26 diagnostic 4-fold tables were analyzed. The bivariate mixed-effects model yielded a SEN of 0.91 (95% CI 0.87 - 0.94), SPC of 0.92 (95% CI 0.86 - 0.95), PLR of 10.7 (95% CI 6.7 - 17.2), NLR of 0.10 (95% CI 0.07 - 0.14), DOR of 110 (95% CI 63 - 191), and SROC curve of 0.96 (95% CI 0.94 - 0.98; Figures S53 and S54 in Multimedia Appendix 4). Deeks' funnel plot did not reflect discernible publication bias (Figure S55 in Multimedia Appendix 4). Among the encompassed study participants, approximately 42% (n=7327) had OP. Assuming this as the prior probability, if ML predicted OP, the probability of actual OP was .89. Conversely, if ML predicted non-OP, the actual likelihood of non-OP was .93 (Figure S56 in Multimedia Appendix 4).

XSL•FO
RenderX

ML Based on MRI

Only 3 studies have constructed diagnostic models for OP based on MRI [45,64,66], all of which used the lumbar vertebrae as the examination part. Due to the limited number of studies of this type and the substantial heterogeneity noted in the meta-analysis, the conclusions drawn lack sufficient reference significance. Therefore, this study presents only a narrative analysis of this part.

Among these, 2 studies used traditional ML models, while 1 used a DL model. The SEN of these models was 0.857, 0.872, and 0.892, and the SPC was 0.944, 0.688, and 0.892, respectively. The validation strategies used in these studies included external validation, K-fold cross-validation, and random sampling.

Discussion

Main Findings of This Study

Medical imaging is an indispensable tool in the diagnosis, treatment, and management of OP. Conventional imaging methods such as x-ray, CT, and MRI are pivotal clinically.

X-ray imaging enables clinicians to visually assess reductions in vertebral height, cortical bone thickness, and morphological changes in appendicular and mandibular bones, thus screening OP. However, as DXA is updated and improved, clinicians can more accurately know bone mineral density and structural parameters of the lumbar vertebrae and hip, thereby facilitating the diagnosis of OP. The World Health Organization has designated DXA as the gold standard for determining bone mineral density and diagnosing postmenopausal OP [74,75]. However, in low-resource environments and economically underdeveloped regions, the clinical application of DXA is limited due to factors such as insufficient medical knowledge and constrained health care infrastructure. In contrast, AI tools have the potential to maximize the extraction of clinically relevant information from various medical images, thereby enabling the early identification of the population with OP or low bone mass. This significantly supports the early prevention, diagnosis, and management of the disease.

Advantages of Different Imaging Modalities in the Diagnosis of OP

This shift has diminished the application of x-rays in quantitative analysis for OP. Nevertheless, ML and DL models have improved the diagnostic performance of x-ray imaging, providing significant impetus for its broader clinical application. CT, with its high resolution, enables clinicians to observe cortical and trabecular bone integrity, offering distinct advantages in evaluating spinal OP and changes in trabecular bone volume ratios in the hip [76]. Conventional CT generates images by measuring differences in the linear attenuation coefficients of x-ray beams as they pass through various biological tissues. However, when tissues possess similar densities, such as calcium and bone, conventional CT often yields comparable Hounsfield unit values due to the use of a single x-ray energy spectrum, limiting its ability to differentiate between such tissues. In contrast, spectral CT imaging, which is based on tissue-specific photoelectric effect weighting, offers enhanced resolution in distinguishing fine bone microarchitecture. This technological advancement holds significant potential for improving the diagnostic accuracy of OP. MRI is highly efficient in assessing bone microarchitecture [77]. However, MRI is not the first choice to detect OP because of its high cost, extended scan times, and obstacles faced by patients with metallic implants or claustrophobia. Our database search corroborated that most studies have focused on x-ray and CT imaging, while comparatively fewer have investigated MRI. Nevertheless, existing evidence supports the robust diagnostic performance of ML models based on imaging data. For example, the pooled SEN and SPC of ML models based on x-ray for OP diagnosis were 0.92 (95% CI 0.88 - 0.94) and 0.83 (95% CI 0.76 - 0.88), respectively. Similarly, ML models developed via CT achieved SEN and SPC of 0.91 (95% CI 0.89 - 0.93) and 0.92 (95% CI 0.89 - 0.94), respectively. These findings demonstrate the high accuracy of x-ray and CT in OP diagnosis. In addition, quantitative ultrasound is another commonly used modality for OP detection. Quantitative ultrasound relies on 2 primary parameters: speed of sound and broadband ultrasound attenuation, which assess the ability of ultrasound waves to propagate through bone both horizontally and longitudinally [78]. In summary, diverse imaging modalities and bone types provide flexible and enriched diagnostic options for OP. Furthermore, this variety brings ample opportunities for the development of advanced ML models tailored to different imaging techniques.

Status Quo of Research on ML

With advances in computer science, numerous researchers have sought to use these techniques in the prevention and treatment of OP. Compared with clinicians, who visually observe positive imaging features, AI-assisted tools significantly improve the efficiency and accuracy of diagnosing OP [46,60]. In addition, Yang et al [79] developed an ML-based predictive model using data from surveys on risk factors for OP, which is highly prospective for early screening and treating OP in the Hong Kong population. Similarly, ML models based on community health examinations and serum bone turnover markers have demonstrated a high area under the receiver operating characteristic curve, F_1-scores, and accuracy [80,81]. These findings highlight the efficiency of ML in the diagnosis and management of OP.

Mechanism of Image-Based ML

Image-based ML can broadly be categorized into traditional ML and DL. Traditional ML involves dividing data into a training set for model development and a test set for model validation. Through processes such as image segmentation, texture extraction, and feature selection, traditional ML models are constructed for predicting outcome events. However, the process of texture feature extraction and selection carries a significant risk of data loss. In contrast, DL incorporates feature extraction directly into the training process, thereby maximizing the retention of meaningful information within the image data. Convolutional neural networks, as a representative DL approach, can simultaneously extract and select features across multiple hidden layers to accomplish classification tasks. Moreover, DL-based models can correct image blurring in panoramic

XSL•FO
RenderX

x-rays caused by patient mispositioning and mitigate the impact of metal artifacts in CT images on feature extraction [82-84]. This study further demonstrates that ML models based on x-ray and CT outperform traditional ML models, suggesting that DL is more accurate than traditional ML approaches. Image analysis using DL can leverage AI to develop more efficient and user-friendly image interpretation tools, providing valuable insights into the development of medical imaging software.

The Impact of Validation Set Generation Methods on ML Performance

Validation methods are critical metrics for assessing the performance of ML models. These methods can be categorized into external validation and internal validation. Internal validation can be further subdivided into random sampling, leave-one-out validation, and K-fold cross-validation. External validation, which can accurately reflect the clinical applicability of ML, is widely preferred by researchers. In contrast, internal validation typically generates validation sets via random methods, which inherently carries a risk of similarity in features and distribution trends between the validation and training sets. This issue is prominent in image-based studies, where the application of ML is restricted in medical research due to the high similarity in images and parameters between internal validation sets. Although external validation offers a superior means of assessing model performance, conducting such validation requires access to independent research cohorts and often entails consideration of factors such as periods, geographical regions, populations, and health care institutions. These requirements inevitably lead to substantial increases in both the time and financial costs of research. This perspective provides an objective explanation for the limited external validation in this study.

Advantages and Limitations

This study is the first to summarize the evidence of the application of ML based on various imaging modalities in the diagnosis of OP. This study provides theoretical support for the subsequent development of clinical scoring systems and medical software. However, our research has the following limitations: first, despite a substantial number of included studies, only a small number of studies on MRI were encompassed in view of the practicability in clinical work. Therefore, in future research, our emphasis will be put on meta-analyses involving MRI studies, aiming to evaluate the utility of ML in the diagnosis of OP through medical imaging. As a result, only a narrative review was performed, without a direct evaluation of its diagnostic performance. Most of the included studies rely primarily on internal validation, with insufficient external validation, which imposes certain limitations on the interpretability and generalizability of our findings. This study encompassed only English publications, with the majority of research originating from countries where AI is more widely applied. In addition, the external validation conducted in this study was limited, constituting an objective constraint that may have influenced the outcomes of the meta-analysis. Future studies will endeavor to comprehensively incorporate globally available literature to enhance the authority and generalizability of the conclusions.

Conclusions

Image-based ML, particularly DL based on x-ray and CT images, is highly accurate in the diagnosis of OP. Future focus should be placed on developing AI-based software to expand its clinical applicability and enhance diagnostic precision.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (82274551) and the Guangdong Provincial Fundamental and Applied Basic Research Fund Program (2023B1515230001). All authors of this study hereby solemnly declare that no forms of artificial intelligence tools were used at any stage of the research process to ensure the originality and independence of this manuscript.

Data Availability

The datasets used and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

RZ was responsible for the formal analysis, data curation, and original manuscript writing of all the experiments. HY was responsible for software operation and data management. Y Li and XL were responsible for the formal analysis. ZY was in charge of the methodology. Y Lin and JH were responsible for the methodology and data validation. LW was in charge of project management. HH was responsible for project design, project management, and fund acquisition. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Literature search strategy.
[DOCX File, 25 KB - jmir_v28i1e75965_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Supplementary data sheet.
[XLSX File, 20 KB - jmir_v28i1e75965_app2.xlsx]

Multimedia Appendix 3
Quality Assessment of Diagnostic Accuracy Studies assessment process for included studies.
[XLSX File, 15 KB - jmir_v28i1e75965_app3.xlsx]

Multimedia Appendix 4
Supplementary materials.
[DOC File, 56031 KB - jmir_v28i1e75965_app4.doc]

Checklist 1
PRISMA checklist.
[PDF File, 99 KB - jmir_v28i1e75965_app5.pdf]

References

1. Consensus development conference: diagnosis, prophylaxis, and treatment of osteoporosis. Am J Med 1993 Jun;94(6):646-650. [doi: 10.1016/0002-9343(93)90218-E]
2. Wang Y, Tao Y, Hyman ME, Li J, Chen Y. Osteoporosis in china. Osteoporos Int 2009 Oct;20(10):1651-1662. [doi: 10.1007/s00198-009-0925-y] [Medline: 19415374]
3. Clynes MA, Harvey NC, Curtis EM, Fuggle NR, Dennison EM, Cooper C. The epidemiology of osteoporosis. Br Med Bull 2020 May 15;133(1):105-117. [doi: 10.1093/bmb/ldaa005] [Medline: 32282039]
4. Assessment of fracture risk and its application to screening for postmenopausal osteoporosis. Report of a WHO Study Group. World Health Organ Tech Rep Ser 1994;843(1-129):1-129. [Medline: 7941614]
5. Kanis JA, Cooper C, Rizzoli R, Reginster JY, on behalf of the Scientific Advisory Board of the European Society for Clinical and Economic Aspects of Osteoporosis (ESCEO) and the Committees of Scientific Advisors and National Societies of the International Osteoporosis Foundation (IOF). European guidance for the diagnosis and management of osteoporosis in postmenopausal women. Osteoporos Int 2019 Jan 18;30(1):3-44. [doi: 10.1007/s00198-018-4704-5]
6. Erickson BJ. Basic artificial intelligence techniques: machine learning and deep learning. Radiol Clin North Am 2021 Nov;59(6):933-940. [doi: 10.1016/j.rcl.2021.06.004] [Medline: 34689878]
7. Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep learning in medical image analysis. Adv Exp Med Biol 2020;1213(3-21):3-21. [doi: 10.1007/978-3-030-33128-3_1] [Medline: 32030660]
8. Oliveira JS, Franco FO, Revers MC, et al. Computer-aided autism diagnosis based on visual attention models using eye tracking. Sci Rep 2021 May 12;11(1):10131. [doi: 10.1038/s41598-021-89023-8] [Medline: 33980874]
9. Khan M, Shah PM, Khan IA, et al. IoMT-enabled computer-aided diagnosis of pulmonary embolism from computed tomography scans using deep learning. Sensors (Basel) 2023 Jan 28;23(3):1471. [doi: 10.3390/s23031471] [Medline: 36772510]
10. Loizidou K, Skouroumouni G, Nikolaou C, Pitris C. A review of computer-aided breast cancer diagnosis using sequential mammograms. Tomography 2022 Dec 6;8(6):2874-2892. [doi: 10.3390/tomography8060241] [Medline: 36548533]
11. Ceranka J, Wuts J, Chiabai O, Lecouvet F, Vandemeulebroucke J. Computer-aided diagnosis of skeletal metastases in multi-parametric whole-body MRI. Comput Methods Programs Biomed 2023 Dec;242:107811. [doi: 10.1016/j.cmpb.2023.107811] [Medline: 37742486]
12. Hong N, Cho SW, Shin S, et al. Deep-learning-based detection of vertebral fracture and osteoporosis using lateral spine X-ray radiography. J Bone Miner Res 2020 Dec 1;38(6):887-895. [doi: 10.1002/jbmr.4814]
13. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
14. Alzubaidi MA, Otoom M. A comprehensive study on feature types for osteoporosis classification in dental panoramic radiographs. Comput Methods Programs Biomed 2020 May;188(105301):105301. [doi: 10.1016/j.cmpb.2019.105301] [Medline: 31911333]
15. Cheng L, Cai F, Xu M, Liu P, Liao J, Zong S. A diagnostic approach integrated multimodal radiomics with machine learning models based on lumbar spine CT and X-ray for osteoporosis. J Bone Miner Metab 2023 Nov;41(6):877-889. [doi: 10.1007/s00774-023-01469-0] [Medline: 37898574]
16. Liu L, Si M, Ma H, et al. A hierarchical opportunistic screening model for osteoporosis using machine learning applied to clinical data and CT images. BMC Bioinformatics 2022 Feb 10;23(1):63. [doi: 10.1186/s12859-022-04596-z] [Medline: 35144529]

XSL•FO
RenderX

17. Cui J, Liu CL, Jennane R, Ai S, Dai K, Tsai TY. A highly generalized classifier for osteoporosis radiography based on multiscale fractal, lacunarity, and entropy distributions. Front Bioeng Biotechnol 2023;11:1054991. [doi: 10.3389/fbioe.2023.1054991] [Medline: 37274169]
18. Genisa M, Abdullah JY, Yusoff BM, Arief EM, Hermana M, Utomo CP. Adopting signal processing technique for osteoporosis detection based on CT scan image. Appl Sci (Basel) 2023;13(8):5094. [doi: 10.3390/app13085094]
19. Menaka R, Ramesh R, Dhanagopal R. Aggregation of region-based and boundary-based knowledge biased segmentation for osteoporosis detection from X-ray, dual X-ray and CT images. Curr Med Imaging 2021;17(2):288-295. [doi: 10.2174/1573405616999200730175526] [Medline: 32748751]
20. Yu X, Ye C, Xiang L. Application of artificial neural network in the diagnostic system of osteoporosis. Neurocomputing 2016 Nov;214:376-381. [doi: 10.1016/j.neucom.2016.06.023]
21. Dzierżak R, Omiotek Z. Application of deep convolutional neural networks in the diagnosis of osteoporosis. Sensors (Basel) 2022 Oct 26;22(21):8189. [doi: 10.3390/s22218189] [Medline: 36365886]
22. Chen B, Cui J, Li C, et al. Application of radiomics model based on lumbar computed tomography in diagnosis of elderly osteoporosis. J Orthop Res 2024 Jun;42(6):1356-1368. [doi: 10.1002/jor.25789] [Medline: 38245854]
23. Tong X, Wang S, Zhang J, Fan Y, Liu Y, Wei W. Automatic osteoporosis screening system using radiomics and deep learning from low-dose chest CT images. Bioengineering (Basel) 2024 Jan 2;11(1):50. [doi: 10.3390/bioengineering11010050] [Medline: 38247927]
24. Dhanagopal R, Menaka R, Suresh Kumar R, Vasanth Raj PT, Debrah EL, Pradeep K. Channel-boosted and transfer learning convolutional neural network-based osteoporosis detection from CT scan, dual X-ray, and X-ray images. J Healthc Eng 2024;2024(3733705):3733705. [doi: 10.1155/2024/3733705] [Medline: 38223259]
25. Breit HC, Varga-Szemes A, Schoepf UJ, et al. CNN-based evaluation of bone density improves diagnostic performance to detect osteopenia and osteoporosis in patients with non-contrast chest CT examinations. Eur J Radiol 2023 Apr;161:110728. [doi: 10.1016/j.ejrad.2023.110728] [Medline: 36773426]
26. Tang C, Zhang W, Li H, et al. CNN-based qualitative detection of bone mineral density via diagnostic CT slices for osteoporosis screening. Osteoporos Int 2021 May;32(5):971-979. [doi: 10.1007/s00198-020-05673-w]
27. Mao L, Xia Z, Pan L, et al. Deep learning for screening primary osteopenia and osteoporosis using spine radiographs and patient clinical covariates in a Chinese population. Front Endocrinol (Lausanne) 2022;13(971877):971877. [doi: 10.3389/fendo.2022.971877] [Medline: 36176468]
28. Kim S, Kim BR, Chae HD, et al. Deep radiomics-based approach to the diagnosis of osteoporosis using hip radiographs. Radiol Artif Intell 2022 Jul;4(4):e210212. [doi: 10.1148/ryai.210212] [Medline: 35923378]
29. Zhang K, Lin PC, Pan J, et al. DeepmdQCT: a multitask network with domain invariant features and comprehensive attention mechanism for quantitative computer tomography diagnosis of osteoporosis. Comput Biol Med 2024 Mar;170:107916. [doi: 10.1016/j.compbiomed.2023.107916] [Medline: 38237237]
30. Zhang B, Chen Z, Yan R, et al. Development and validation of a feature-based broad-learning system for opportunistic osteoporosis screening using lumbar spine radiographs. Acad Radiol 2024 Jan;31(1):84-92. [doi: 10.1016/j.acra.2023.07.002]
31. Niu X, Huang Y, Li X, et al. Development and validation of a fully automated system using deep learning for opportunistic osteoporosis screening using low-dose computed tomography scans. Quant Imaging Med Surg 2023 Aug 1;13(8):5294-5305. [doi: 10.21037/qims-22-1438] [Medline: 37581046]
32. Xie Q, Chen Y, Hu Y, et al. Development and validation of a machine learning-derived radiomics model for diagnosis of osteoporosis and osteopenia using quantitative computed tomography. BMC Med Imaging 2022 Aug 8;22(1):140. [doi: 10.1186/s12880-022-00868-5] [Medline: 35941568]
33. Uemura K, Otake Y, Takashima K, et al. Development and validation of an open-source tool for opportunistic screening of osteoporosis from hip CT images. Bone Joint Res 2023 Sep 20;12(9):590-597. [doi: 10.1302/2046-3758.129.BJR-2023-0115.R1] [Medline: 37728034]
34. Zaman MU, Alam MK, Alqhtani NR, et al. RETRACTED ARTICLE: Diagnosing osteoporosis using deep neural networkassisted optical image processing method. Opt Quant Electron 2024 Mar;56(3). [doi: 10.1007/s11082-023-06031-w]
35. Lee JH, Hwang YN, Park SY, Jeong JH, Kim SM. Diagnosis of osteoporosis by quantification of trabecular microarchitectures from hip radiographs using artificial neural networks. J Comp Theo Nano 2015 Jul 1;12(7):1115-1120. [doi: 10.1166/jctn.2015.3859]
36. Kavitha MS, Asano A, Taguchi A, Kurita T, Sanada M. Diagnosis of osteoporosis from dental panoramic radiographs using the support vector machine method in a computer-aided system. BMC Med Imaging 2012 Jan 16;12(1):22248480. [doi: 10.1186/1471-2342-12-1] [Medline: 22248480]
37. Yamamoto N, Sukegawa S, Yamashita K, et al. Effect of patient clinical variables in osteoporosis classification using hip X-rays in deep learning analysis. Medicina (Kaunas) 2021 Aug 20;57(8):846. [doi: 10.3390/medicina57080846] [Medline: 34441052]
38. Pan J, Lin PC, Gong SC, et al. Effectiveness of opportunistic osteoporosis screening on chest CT using the DCNN model. BMC Musculoskelet Disord 2024 Feb 27;25(1):176. [doi: 10.1186/s12891-024-07297-1] [Medline: 38413868]
39. Zhang K, Lin P, Pan J, et al. End to end multitask joint learning model for osteoporosis classification in CT images. Comput Intell Neurosci 2023;2023:3018320. [doi: 10.1155/2023/3018320] [Medline: 36970245]

XSL•FO
RenderX

40. Oh J, Kim B, Oh G, Hwangbo Y, Ye JC. End-to-end semi-supervised opportunistic osteoporosis screening using computed tomography. Endocrinol Metab (Seoul) 2024 Jun;39(3):500-510. [doi: 10.3803/EnM.2023.1860] [Medline: 38721637]
41. Oh S, Kang WY, Park H, et al. Evaluation of deep learning-based quantitative computed tomography for opportunistic osteoporosis screening. Sci Rep 2024 Jan 5;14(1):363. [doi: 10.1038/s41598-023-45824-7] [Medline: 38182616]
42. Lee KS, Jung SK, Ryu JJ, Shin SW, Choi J. Evaluation of transfer learning with deep convolutional neural networks for screening osteoporosis in dental panoramic radiographs. J Clin Med 2020 Feb 1;9(2):392. [doi: 10.3390/jcm9020392] [Medline: 32024114]
43. Xu Y, Li D, Chen Q, Fan Y. Full supervised learning for osteoporosis diagnosis using micro‐CT images. Microsc Res Tech 2013 Apr;76(4):333-341. [doi: 10.1002/jemt.22171]
44. Wang S, Tong X, Cheng Q, et al. Fully automated deep learning system for osteoporosis screening using chest computed tomography images. Quant Imaging Med Surg 2024 Apr 3;14(4):2816-2827. [doi: 10.21037/qims-23-1617] [Medline: 38617137]
45. Zhao Y, Zhao T, Chen S, et al. Fully automated radiomic screening pipeline for osteoporosis and abnormal bone density with a deep learning-based segmentation using a short lumbar mDixon sequence. Quant Imaging Med Surg 2022 Feb;12(2):1198-1213. [doi: 10.21037/qims-21-587] [Medline: 35111616]
46. Su R, Liu T, Sun C, Jin Q, Jennane R, Wei L. Fusing convolutional neural network features with hand-crafted features for osteoporosis diagnoses. Neurocomputing 2020 Apr;385:300-309. [doi: 10.1016/j.neucom.2019.12.083]
47. Pickhardt PJ, Nguyen T, Perez AA, et al. Improved CT-based osteoporosis assessment with a fully automated deep learning tool. Radiol Artif Intell 2022 Sep;4(5):e220042. [doi: 10.1148/ryai.220042] [Medline: 36204542]
48. Jiang C, Jin D, Ni M, Zhang Y, Yuan H. Influence of image reconstruction kernel on computed tomography-based finite element analysis in the clinical opportunistic screening of osteoporosis—a preliminary result. Front Endocrinol 2023;14(1076990):36936156. [doi: 10.3389/fendo.2023.1076990]
49. Alshamrani K, Alshamrani HA. Lossless compression-based detection of osteoporosis using bone X-ray imaging. J Xray Sci Technol 2024;32(2):475-491. [doi: 10.3233/XST-230238] [Medline: 38393881]
50. Sebro R, Elmahdy M. Machine learning for opportunistic screening for osteoporosis and osteopenia using knee CT scans. Can Assoc Radiol J 2023 Nov;74(4):676-687. [doi: 10.1177/08465371231164743] [Medline: 36960893]
51. Sebro R, De la Garza-Ramos C. Machine learning for opportunistic screening for osteoporosis from CT scans of the wrist and forearm. Diagnostics (Basel) 2022 Mar 11;12(3):691. [doi: 10.3390/diagnostics12030691] [Medline: 35328244]
52. Namatevs I, Nikulins A, Edelmers E, et al. Modular neural networks for osteoporosis detection in mandibular cone-beam computed tomography scans. Tomography 2023 Sep 22;9(5):1772-1786. [doi: 10.3390/tomography9050141] [Medline: 37888733]
53. Hwang DH, Bak SH, Ha TJ, Kim Y, Kim WJ, Choi HS. Multi-view computed tomography network for osteoporosis classification. IEEE Access 2023;11:22297-22306. [doi: 10.1109/ACCESS.2023.3252361]
54. Fang Y, Li W, Chen X, et al. Opportunistic osteoporosis screening in multi-detector CT images using deep convolutional neural networks. Eur Radiol 2021 Apr;31(4):1831-1842. [doi: 10.1007/s00330-020-07312-8]
55. Yang J, Liao M, Wang Y, et al. Opportunistic osteoporosis screening using chest CT with artificial intelligence. Osteoporos Int 2022 Dec;33(12):2547-2561. [doi: 10.1007/s00198-022-06491-y]
56. Jang M, Kim M, Bae SJ, Lee SH, Koh JM, Kim N. Opportunistic osteoporosis screening using chest radiographs with deep learning: development and external validation with a cohort dataset. J Bone Miner Res 2022 Feb;37(2):369-377. [doi: 10.1002/jbmr.4477] [Medline: 34812546]
57. Sebro R, De la Garza-Ramos C. Opportunistic screening for osteoporosis and osteopenia from CT scans of the abdomen and pelvis using machine learning. Eur Radiol 2023 Mar;33(3):1812-1823. [doi: 10.1007/s00330-022-09136-0] [Medline: 36166085]
58. Elmahdy M, Sebro R. Opportunistic screening for osteoporosis using CT scans of the knee: a pilot study. Stud Health Technol Inform 2023 May 18;302:909-910. [doi: 10.3233/SHTI230305] [Medline: 37203533]
59. Mohammadi FG, Sebro R. Opportunistic screening for osteoporosis using hand radiographs: a preliminary study. Stud Health Technol Inform 2023 May 18;302:911-912. [doi: 10.3233/SHTI230306] [Medline: 37203534]
60. Lee JS, Adhikari S, Liu L, Jeong HG, Kim H, Yoon SJ. Osteoporosis detection in panoramic radiographs using a deep convolutional neural network-based computer-assisted diagnosis system: a preliminary study. Dentomaxillofac Radiol 2019 Jan;48(1):20170344. [doi: 10.1259/dmfr.20170344] [Medline: 30004241]
61. Wani IM, Arora S. Osteoporosis diagnosis in knee X-rays by transfer learning based on convolution neural network. Multimed Tools Appl 2023;82(9):14193-14217. [doi: 10.1007/s11042-022-13911-y] [Medline: 36185321]
62. Nakamoto T, Taguchi A, Kakimoto N. Osteoporosis screening support system from panoramic radiographs using deep learning by convolutional neural network. Dentomaxillofac Radiol 2022 Sep 1;51(6):20220135. [doi: 10.1259/dmfr.20220135] [Medline: 35816516]
63. Muramatsu C, Horiba K, Hayashi T, et al. Quantitative assessment of mandibular cortical erosion on dental panoramic radiographs for screening osteoporosis. Int J Comput Assist Radiol Surg 2016 Nov;11(11):2021-2032. [doi: 10.1007/s11548-016-1438-8] [Medline: 27289239]

XSL•FO
RenderX

64. Kang SR, Wang K. Radiomic nomogram based on lumbar spine magnetic resonance images to diagnose osteoporosis. Acta Radiol 2024 Aug;65(8):950-958. [doi: 10.1177/02841851241242052]
65. Jiang YW, Xu XJ, Wang R, Chen CM. Radiomics analysis based on lumbar spine CT to detect osteoporosis. Eur Radiol 2022 Nov;32(11):8019-8026. [doi: 10.1007/s00330-022-08805-4] [Medline: 35499565]
66. He L, Liu Z, Liu C, et al. Radiomics based on lumbar spine magnetic resonance imaging to detect osteoporosis. Acad Radiol 2021 Jun;28(6):e165-e171. [doi: 10.1016/j.acra.2020.03.046] [Medline: 32386949]
67. Zhang H, Wei W, Qian B, et al. Screening for osteoporosis based on IQon spectral CT virtual low monoenergetic images: comparison with conventional 120 kVp images. Heliyon 2023 Oct;9(10):e20750. [doi: 10.1016/j.heliyon.2023.e20750]
68. Krishnaraj A, Barrett S, Bregman-Amitai O, et al. Simulating dual-energy X-ray absorptiometry in CT using deep-learning segmentation cascade. J Am Coll Radiol 2019 Oct;16(10):1473-1479. [doi: 10.1016/j.jacr.2019.02.033]
69. Sebro R, De la Garza-Ramos C. Support vector machines are superior to principal components analysis for selecting the optimal bones' CT attenuations for opportunistic screening for osteoporosis using CT scans of the foot or ankle. Osteoporos Sarcopenia 2022 Sep;8(3):112-122. [doi: 10.1016/j.afos.2022.09.002] [Medline: 36268496]
70. Kavitha MS, An SY, An CH, et al. Texture analysis of mandibular cortical bone on digital dental panoramic radiographs for the diagnosis of osteoporosis in Korean women. Oral Surg Oral Med Oral Pathol Oral Radiol 2015 Mar;119(3):346-356. [doi: 10.1016/j.oooo.2014.11.009] [Medline: 25600978]
71. Kavitha MS, Asano A, Taguchi A, Heo MS. The combination of a histogram-based clustering algorithm and support vector machine for the diagnosis of osteoporosis. Imaging Sci Dent 2013;43(3):153. [doi: 10.5624/isd.2013.43.3.153]
72. Fang K, Zheng X, Lin X, Dai Z. Unveiling osteoporosis through radiomics analysis of hip CT imaging. Acad Radiol 2024 Mar;31(3):1003-1013. [doi: 10.1016/j.acra.2023.10.009]
73. Xue Z, Huo J, Sun X, et al. Using radiomic features of lumbar spine CT images to differentiate osteoporosis from normal bone density. BMC Musculoskelet Disord 2022 Dec;23(1):35395769. [doi: 10.1186/s12891-022-05309-6]
74. Chen M, Gerges M, Raynor WY, et al. State of the art imaging of osteoporosis. Semin Nucl Med 2024 May;54(3):415-426. [doi: 10.1053/j.semnuclmed.2023.10.008]
75. El Maghraoui A, Roux C. DXA scanning in clinical practice. QJM 2008 Aug;101(8):605-617. [doi: 10.1093/qjmed/hcn022] [Medline: 18334497]
76. Kessenich CR. Diagnostic imaging and biochemical markers of bone turnover. Nurs Clin North Am 2001 Sep;36(3):409-416. [Medline: 11532656]
77. Sollmann N, Löffler MT, Kronthaler S, et al. MRI‐based quantitative osteoporosis imaging at the spine and femur. Magn Reson Imaging 2021 Jul;54(1):12-35. [doi: 10.1002/jmri.27260]
78. Oei L, Koromani F, Rivadeneira F, Zillikens MC, Oei EHG. Quantitative imaging methods in osteoporosis. Quant Imaging Med Surg 2016 Dec;6(6):680-698. [doi: 10.21037/qims.2016.12.13] [Medline: 28090446]
79. Yang Q, Cheng H, Qin J, et al. A machine learning-based Preclinical Osteoporosis Screening Tool (POST): model development and validation study. JMIR Aging 2023 Nov 8;6:e46791. [doi: 10.2196/46791] [Medline: 37986117]
80. Baik SM, Kwon HJ, Kim Y, Lee J, Park YH, Park DJ. Machine learning model for osteoporosis diagnosis based on bone turnover markers. Health Informatics J 2024;30(3):39115269. [doi: 10.1177/14604582241270778] [Medline: 39115269]
81. Ou Yang WY, Lai CC, Tsou MT, Hwang LC. Development of machine learning models for prediction of osteoporosis from clinical health examination data. Int J Environ Res Public Health 2021 Jul 18;18(14):7635. [doi: 10.3390/ijerph18147635] [Medline: 34300086]
82. Putra RH, Doi C, Yoda N, Astuti ER, Sasaki K. Current applications and development of artificial intelligence for digital dental radiography. Dentomaxillofac Radiol 2022 Jan 1;51(1):20210197. [doi: 10.1259/dmfr.20210197] [Medline: 34233515]
83. Du X, Chen Y, Zhao J, Xi Y. A convolutional neural network based auto-positioning method for dental arch in rotational panoramic radiography. Annu Int Conf IEEE Eng Med Biol Soc 2018 Jul;2018(2615-8):2615-2618. [doi: 10.1109/EMBC.2018.8512732] [Medline: 30440944]
84. Liang K, Zhang L, Yang H, Yang Y, Chen Z, Xing Y. Metal artifact reduction for practical dental computed tomography by improving interpolation-based reconstruction with deep learning. Med Phys 2019 Dec;46(12):e823-e834. [doi: 10.1002/mp.13644] [Medline: 31811792]

Abbreviations

AI: artificial intelligence
CAD: computer-aided diagnosis
CT: computed tomography
DL: deep learning
DOR: diagnostic odds ratio
DXA: dual-energy x-ray absorptiometry
ML: machine learning
MRI: magnetic resonance imaging
NLR: negative likelihood ratio

XSL•FO
RenderX

OP: osteoporosis
PLR: positive likelihood ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SEN: sensitivity
SPC: specificity
SROC: summary receiver operating characteristic

Edited by J Sarvestan; submitted 14.Apr.2025; peer-reviewed by M He, N Nanthasamroeng; revised version received 03.Jun.2025; accepted 04.Jun.2025; published 16.Jan.2026.

Please cite as:
Zhao R, Yang H, Li Y, Li X, Yang Z, Lin Y, Huang J, Wan L, Huang H
The Diagnostic Value of Image-Based Machine Learning for Osteoporosis: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e75965
URL: https://www.jmir.org/2026/1/e75965
doi:10.2196/75965

© Rui Zhao, Haolin Yang, Yangbo Li, Xiaoyun Li, Zhijie Yang, Yanping Lin, Jiachun Huang, Lei Wan, Hongxing Huang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Application of Mobile Health in Self-Management Among Patients Undergoing Dialysis: Scoping Review

Qin Xu, MNS; Yulin Xu, MNS; Xiaoqin Liu, MNS; Xiaolin Ma, BSN

Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, No. 1095 Jiefang Avenue, Wuhan, Hubei, China

Corresponding Author:
Xiaoqin Liu, MNS
Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, No. 1095 Jiefang Avenue, Wuhan, Hubei, China

Abstract

Background: The incidence of end-stage renal disease continues to rise annually, with dialysis currently serving as the primary replacement therapy. The effectiveness of dialysis treatment and patients' quality of life are highly dependent on their self-management. Mobile health (mHealth), which provides real-time medical support through portable devices, has become an essential tool for assisting patients undergoing dialysis in optimizing their self-management.

Objective: This study aimed to systematically explore the core elements of self-management in patients undergoing dialysis and clarify the primary applications of mHealth, including types of mHealth, relevant theories and models, mHealth-based interventions, and evaluation indicators.

Methods: This study was guided by Arksey and O'Malley's methodology, PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews), and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension). Databases, such as PubMed, Embase, CINAHL, PsycINFO, and Web of Science, were systematically searched from January 2010 until October 2025. The participants included in this study were patients undergoing dialysis, and the study design must incorporate quantitative research. Published protocols, reviews, editorials, conference papers, books, and non-English studies were excluded. The Mixed Methods Appraisal Tool was used to evaluate the quality of the included studies. Quantitative studies were extracted, mapped, and summarized. The results were collated and synthesized using a structured spreadsheet.

Results: Out of 1483 relevant studies, this scoping review ultimately selected 34 studies involving 2068 patients undergoing dialysis. Self-management among patients undergoing dialysis in this study included 6 major areas, including self-monitoring, diet and fluid management, medication management, disease-related knowledge, exercise management, and psychological management. Most studies used a single app (n=22) for management of patients undergoing dialysis, followed by 2 or more online interventions (n=6) and a remote patient monitoring system (n=3). The mHealth-based interventions in this study focused on self-monitoring, dietary and fluid management, and medication management. The transtheoretical model and stages of change (n=5), self-efficacy theory (n=4), and social cognitive theory (n=4) were the most commonly used theories. Among the evaluation indicators, interdialytic weight gain (n=12), serum potassium (n=14), serum phosphorus (n=20), and serum albumin (n=14) were the most commonly used objective indicators. Subjective indicators were assessed using scales, primarily covering adherence (n=17), self-efficacy (n=14), quality of life (n=12), knowledge (n=9), and diet and nutrition (n=9).

Conclusions: Although mHealth holds promise for improving self-management and outcomes among patients undergoing dialysis, there remains significant room for advancement. Future research in this field should focus on enhancing adaptive software development, deeply integrating artificial intelligence technologies, addressing the needs of special populations, and establishing a standardized self-management evaluation system. Our findings not only provide a theoretical framework for optimizing clinical management strategies for patients undergoing dialysis but also offer targeted guidance and practical insights for the subsequent development of apps.

(J Med Internet Res 2026;28:e76880) doi:10.2196/76880

KEYWORDS

kidney; self-management; mHealth; hemodialysis; peritoneal dialysis; mobile health

Introduction

End-stage renal disease (ESRD) refers to the end stage of various chronic kidney diseases (CKDs). The global prevalence of patients with renal failure receiving dialysis treatment continues to rise, with the latest estimate reaching 823 per million population [1,2]. Although kidney transplantation is the treatment of choice for patients with ESRD, the majority of

XSL•FO
RenderX

patients still rely on dialysis due to the shortage of donor kidneys [3]. Hemodialysis and peritoneal dialysis (PD) are the 2 most common types of dialysis. Although hemodialysis and PD have significantly improved survival rates among patients with ESRD [1,4], the invasive and long-term treatments also substantially increase the risk of dialysis-related complications or infections. Studies have shown that comorbidities, such as hypertension, diabetes mellitus, hyperkalemia, and hyperphosphatemia, and cardiovascular diseases are common in patients undergoing dialysis [1,5]. The quality of life and survival rate of patients undergoing dialysis also decline with increasing dialysis duration [5,6]. The quality of life and survival outcomes of patients undergoing dialysis are closely related to the quality of dialysis treatment, which in turn is directly dependent on the level of the patient's self-management [7].

Self-management encompasses multiple aspects of health management. According to Lorig et al [8], self-management involved medical management (special dietary adherence and medication adherence), role management, and emotion management. In patients undergoing dialysis, self-management refers to whether the patients perform self-monitoring, strict control of diet (sodium, potassium, phosphorus, and other micronutrients) and fluid intake, regular medication administration, and prevention and management of complications. Patients undergoing dialysis can reduce negative symptoms and improve the quality of life through self-management behaviors [7].

Disease management of patients undergoing dialysis is a difficult point in the current medical work. The results of several studies have shown that the overall self-management level of patients undergoing dialysis was low [9,10]. Patients undergoing dialysis generally lacked knowledge of the disease, and their self-management behaviors, such as dietary control, fluid intake, and treatment adherence, fell short of standards [10]. The proposal of mobile health (mHealth) provides new ideas and methods for the remote management of patients undergoing dialysis. mHealth is a medical and public health service initiative based on mobile communications technology delivered through mobile phones, monitoring devices, personal digital assistant devices, and other wireless devices [11]. Because of its unique convenience, it has significant advantages in monitoring diseases, controlling symptoms, and promoting healthy behaviors.

There is a growing body of research on the role of mHealth in improving self-management in patients undergoing dialysis [12-15]. Most studies concentrate on developing mobile apps specifically designed for patients undergoing dialysis, that is, specialized software tools running on mobile devices. Additionally, telephones and SMS text messaging are also commonly used tools. Despite the growing interest in mHealth, evidence on mHealth-based self-management among patients undergoing dialysis remains limited. Therefore, this review aims to provide an overview of the use of mHealth in the self-management of patients undergoing dialysis, examine existing interventions, and summarize existing evaluation tools. The goal is to empower patients undergoing dialysis through mHealth, improve their self-management to enhance prognosis, and provide a practical reference for subsequent app development.

Methods

Overview

A scoping review based on the 5-stage methodological framework of Arksey and O'Malley [16], involving (1) identifying the research question, (2) identifying relevant studies, (3) study selection, (4) charting the data, and (5) collating, summarizing, and reporting the results [16]. The PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) [17] and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) guidelines [18] (Checklist 1) were used as the protocol for this study. The study quality assessment was conducted using the 2018 version of the Mixed Methods Appraisal Tool, a tool specifically designed to evaluate the quality of qualitative, quantitative, and mixed methods studies [19].

Stage 1: Identifying the Research Questions

The study population was adult patients on dialysis (both hemodialysis and PD), and the type of intervention was a service using mHealth. The primary objective of this review was to explore the use of mHealth in the self-management of patients undergoing dialysis. The research questions were based on an initial literature search and were refined during discussions in the research team.

In accordance with the overall objectives of this review, we have refined the research questions as follows:

1. What are the types of mHealth in the included studies?
2. What are the attributes of self-management in the included studies?
3. What are the interventions for mHealth-based self-management?
4. What are the evaluation tools for self-management in the included studies?

Stage 2: Identifying Relevant Studies

A systematic literature search was conducted in PubMed (National Center for Biotechnology Information), Embase (Ovid), CINAHL (EBSCOhost), PsycINFO (EBSCOhost), and Web of Science (Clarivate Analytics) to identify studies relevant to the research objectives. We limited the search to studies that were published from January 2010 to October 2025. First, we identified the search strategy and search terms through group discussion. Then, a presearch was conducted in the database using the search strategy and search terms, and the search strategy and search terms were adjusted according to the search results. Subsequently, formal searches were conducted in 5 databases using the identified search strategies and search terms. Two keywords "mHealth" and "dialysis" were used in combination to cover the 2 main concepts of the research question. The specific search strategy and updated search methods can be found in Multimedia Appendix 1. The reference lists of all eligible studies were examined to identify any potentially relevant studies.

XSL•FO
RenderX

Stage 3: Study Selection

Included studies were required to fulfill the following criteria. (1) Participants: adult patients (age≥18 years) receiving long-term dialysis with no gender restrictions; (2) Concept: studies were included if they addressed mHealth and self-management. Self-management is the focus of this scoping review. Among patients undergoing dialysis, we defined self-management as knowledge, diet and fluids, dialysis treatments, medications, dialysis access, exercise, and psychology. mHealth refers to health and medical services (including remote monitoring, health education, online counseling, etc) that are delivered using mobile devices (eg, smartphones and tablets); (3) Context: dialysis occurs at home or in a hospital. Eligible study designs must include quantitative research. We excluded published research protocols, reviews, editorials, conference papers, books, and non-English studies. Finally, if the full text cannot be obtained, the study will be excluded.

EndNote (Clarivate Analytics) software was used to identify duplicates and manage literature. The literature was screened by 2 trained reviewers (QX and YX). In the first stage, 2 reviewers independently reviewed the titles and abstracts of studies based on inclusion and exclusion criteria. Then, the 2 reviewers continued to independently screen the full text. In the second stage, 2 reviewers assessed the quality of the included literature based on the Mixed Methods Appraisal Tool. When 2 reviewers disagree, a third reviewer (Xiaoqin Liu) will join the discussion until all reviewers reach consensus, ensuring the rigor of the selection process.

Stage 4: Charting the Data

The research team worked together to develop a data chart to guide the extraction of key information from each study. Descriptive chart information includes (1) a general description of the study, such as first author and year, country, study design, patient population, and purpose of the study; and (2) intervention-specific information, including type of mHealth, primary function, method of implementation, intervention time, and evaluation tools.

Stage 5: Collating, Summarizing, and Reporting the Results

The research team summarized the data iteratively. Descriptive analyses were used to summarize the types of mHealth and self-management evaluation tools, while thematic content analyses were used to summarize the attributes of self-management and the content of mHealth-based self-management. First, codes were developed and applied to analyze the data. Coded segments of the data chart were then created with color-coded quotations, and the coding results were summarized in an Excel (Microsoft Corp) sheet. The Excel sheet was sorted by code and density. Key themes were extracted by analyzing the studies in an overall iterative comparison.

Ethical Considerations

Ethical approval was not needed for this review.

Results

Basic Characteristics of the Included Studies

We retrieved 1483 records from PsychlNFO (n=67), Web of Science (n=516), PubMed (n=478), CINAHL (n=171), and Embase (n=251). In total, 359 studies were excluded due to duplication, and 913 studies were excluded after reading the title and abstract. There were still 211 studies that needed to be read in full. After reading the full text, a total of 34 studies [12-15,20-49] were included in this scoping review (Figure 1).

A total of 34 studies [12-15,20-49] were included in this study, involving 2068 patients undergoing dialysis. Among the included studies, 26 studies [12,13,20-26,28-44] involved patients undergoing hemodialysis, 6 studies [14,15,27,45-47] focused on patients undergoing PD, and 2 studies [48,49] encompassed patients undergoing both hemodialysis and PD. In terms of the regional distribution of the included studies, most of the studies were concentrated in Asia and North America. Among the 34 studies, Korea was dominated with 8 (23.53%) studies [13,15,23,27-29,35,46], followed by 5 (14.71%) studies [25,38,40,44,48] in the United States, 3 (8.82%) studies [12,24,47] in China, 3 (8.82%) studies [21,31,39] in Iran, 3 (8.82%) studies [26,30,49] in Australia, and 3 (8.82%) studies [14,37,42] in Thailand. Around 2 (5.88%) studies [20,22] in the Netherlands, 2 (5.88%) studies [32,45] in Japan, and 2 (5.88%) studies [34,43] in Indonesia each contributed to 2 studies, while 1 (2.94%) study [41] in Malaysia, 1 (2.94%) study [33] in Turkey, and 1 (2.94%) study [36] in Brazil each contributed 1 study. The studies were published mainly between 2019 and 2025, with a total of 31 studies [12-15,20-24,26-31,33-47,49]. Three studies were published in 2011, 2013, and 2017 (Figure 2A). The results of the quality assessment indicated that most studies demonstrated good quality. Specific details of the quality assessment can be found in Multimedia Appendix 2. Table 1 provides an overview of the key characteristics of the included studies. Additionally, Multimedia Appendix 3 presents the intervention type, core intervention contents, and evaluation indicators for each study.

XSL•FO
RenderX

Figure 1. Flowchart outlining the search process for studies across databases, following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines.

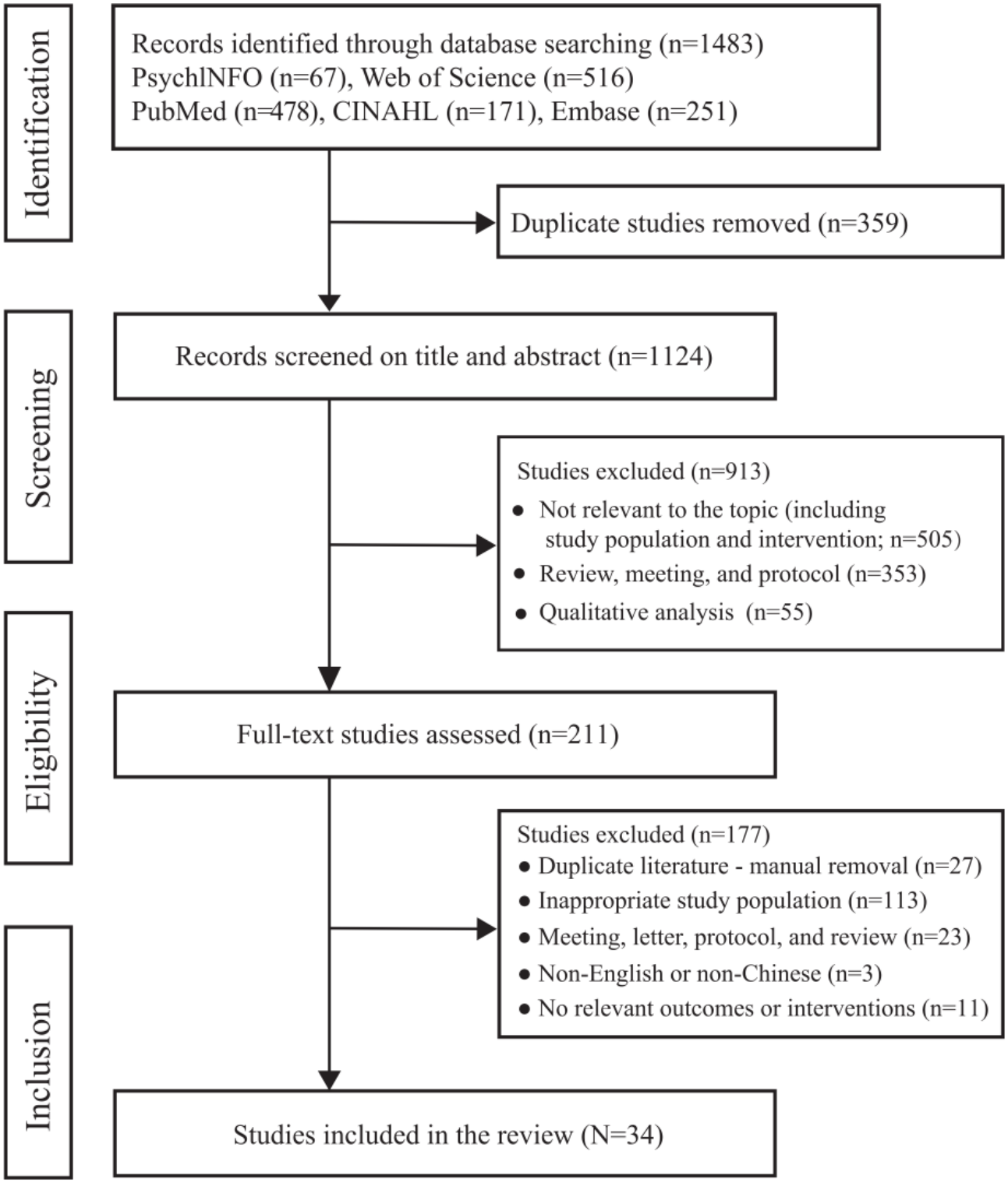

XSL•FO
RenderX

Figure 2. (A) The trend analysis of publication dates in the included studies. (B) The specific information on theories and models used to guide the content of mHealth self-management interventions. HD: hemodialysis; mHealth: mobile health; PD: peritoneal dialysis.

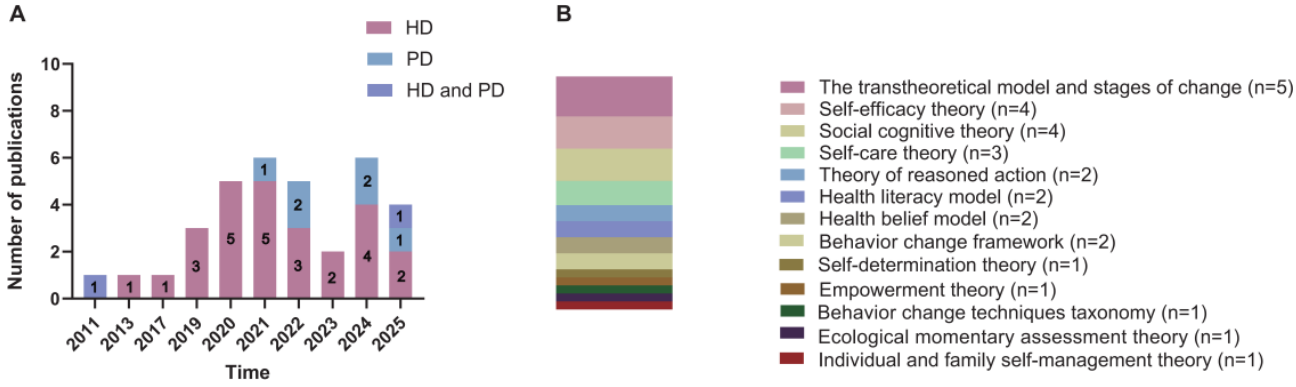

XSL•FO
RenderX

Table . Summary of basic characteristics of the included studies.

Author and year	Country	Aim of the study	Study design	Population	Total sample (experimental/control group)
Saadatifar et al [39], 2022	Iran	To explore the impact of mHealth[a] on treatment adherence in patients undergoing HD[b].	Quasi-experimental study	HD	80 (40/40)
Ren et al [24], 2019	China	To explore the impact of a WeChat-based health education program on the self-management of patients undergoing HD.	A longitudinal experimental intervention study	HD	85 (49/36)
Park and Kim [23], 2019	South Korea	To evaluate the impact of a program based on app and SMS text messaging for patients undergoing HD.	Quasi-experimental study	HD	84 (42/42)
Pack and Lee [35], 2021	South Korea	To develop a mobile app for dietary management and evaluate the impact on patients undergoing HD.	RCT[c]	HD	75 (37/38)
Fakih El Khoury et al [22], 2020	Netherlands	To evaluate the efficacy of the app-based dietary intervention for patients undergoing HD.	A pilot study: a before-and-after study	HD	23
Hanifi et al [31], 2019	Iran	To assess the impact of counseling and follow-up phone calls on patients undergoing HD.	Quasi-experimental study	HD	86 (43/43)
Cho and Park [28], 2020	South Korea	To assess the impact of a tablet-based self-management program on patients undergoing HD.	Quasi-experimental study	HD	46 (23/23)
Chiang et al [12], 2021	China	To evaluate whether the Assisted Care Program in the app can help patients better control their dietary phosphorus intake.	Quasi-experimental study	HD	60 (30/30)
Zwi et al [26], 2022	Australia	To evaluate the feasibility of the app and its impact on patients undergoing HD.	A mixed methods study	HD	61
Welch et al [25], 2013	United States	To evaluate the impact of mobile programs on diet and fluid intake in patients undergoing HD.	A pilot study	HD	33 (16/17)
Thongsunti et al [42], 2024	Thailand	To evaluate the effectiveness of telemedicine-based management of hyperphosphatemia in patients undergoing HD.	RCT	HD	80 (40/40)

XSL•FO
RenderX

Author and year	Country	Aim of the study	Study design	Population	Total sample (experimental/control group)
Chung et al [29], 2024	South Korea	To assess the impact of adaptive nutrition and education on patients undergoing HD using real electronic medical record data.	A decentralized clinical trial	HD	153 (42/40/34/37)[d]
Dawson et al [30], 2021	Australia	To assess the impact of SMS text messaging on patients undergoing HD.	A randomized feasibility study	HD	115 (78/37)
Fakih El Khoury et al [20], 2021	Netherlands	To assess the efficacy of an intervention using the app on phosphorus.	A pilot study: a before-and-after study	HD	23
Hosseini et al [21], 2023	Iran	To determine the effect of an app on self-efficacy and self-care of patients undergoing HD.	Quasi-experimental study (longitudinal single-group study)	HD	54
Min and Park [13], 2020	South Korea	To assess the impact of a mobile app–based self-management support program on older adults undergoing HD.	Quasi-experimental study	HD	56 (28/28)
Mollaoğlu et al [33], 2024	Turkey	To evaluate the impact of education and art therapy through a telemedicine approach in patients undergoing HD.	RCT	HD	60 (30/30)
Rocco et al [38], 2023	United States	To evaluate the impact of an app on self-monitoring of daily fluids in patients undergoing HD.	A pilot study: a before-and-after study	HD	18
St-Jules et al [40], 2021	United States	To assess the feasibility and acceptability of mHealth for managing hyperphosphatemia in patients undergoing HD.	A feasibility trial	HD	40 (13/14/13)[e]
Teong et al [41], 2022	Malaysia	To evaluate the effectiveness of an app for nutritional management in patients undergoing HD.	RCT	HD	66 (33/33)
Pungchompoo et al [37], 2024	Thailand	To evaluate the impact of a home telemedicine model on older patients undergoing HD.	A mixed methods study	HD	54 (24/30)
Nursalam et al [34], 2020	Indonesia	To evaluate the impact of the app on improving fluid restriction adherence in patients undergoing HD.	A mixed methods study	HD	60 (30/30)
Hayashil et al [32], 2017	Japan	To evaluate the usefulness of the self-management support system for self-monitoring in patients undergoing HD.	A pilot study	HD	18 (8/10)

XSL•FO
RenderX

Author and year	Country	Aim of the study	Study design	Population	Total sample (experimental/control group)
Pinto et al [36], 2020	Brazil	To evaluate the impact of the app on fluid restriction and dietary control in patients undergoing HD.	A randomized, single-center, self-controlled study	HD	48
Andriati [43], 2025	Indonesia	To evaluate the impact of the app on adherence and renal function of patients undergoing HD.	Quasi-experimental study	HD	55
Taguiam [44], 2025	United States	To evaluate the impact of the app on fluid intake management and body weight in patients undergoing HD.	A mixed methods study	HD	18
Lee and Kang [15], 2024	South Korea	Using a mobile instant messaging tool to customize diet plans for patients undergoing PD[f] and evaluate outcomes.	Quasi-experimental study	PD	43 (21/22)
Chae and Kim [27], 2024	South Korea	To develop the app for improved self-management and evaluate its impact on patients undergoing PD.	RCT	PD	53 (27/26)
Uchiyama et al [45], 2022	Japan	To assess the impact of using a remote patient monitoring system on patients undergoing PD.	A randomized crossover controlled trial	PD	15
Jung et al [46], 2021	South Korea	To evaluate the impact of remote patient monitoring on automated patients undergoing PD.	RCT	PD	50 (28/22)
Lukkanalikitkul et al [14], 2022	Thailand	To evaluate the availability and impact on patients undergoing PD for app.	User-centered design study	PD	9
Zeng et al [47], 2025	China	To evaluate the effectiveness of a PD management system in improving adherence and clinical outcomes.	A retrospective cohort study	PD	127
Stark et al [48], 2011	United States	To assess the effectiveness of a PDA[g]-based app for dietary management in patients undergoing dialysis.	RCT	PD, HD	HD:19 (9/10); PD: 21 (11/10)

XSL•FO
RenderX

Author and year	Country	Aim of the study	Study design	Population	Total sample (experimental/control group)
Beer et al [49], 2025	Australia	To evaluate the effectiveness of the app in controlling serum phosphorus levels in patients undergoing dialysis.	RCT	PD, HD	180 (90/90)

[a]mHealth: mobile health.

[b]HD: hemodialysis.

[c]RCT: randomized controlled trial.

[d]The study divided participants into four groups: (1) control (n=42), (2) education intervention (n=40), (3) meal intervention (n=34), and (4) education and meal interventions (n=37).

[e]The grouping is set up as follows: (1) educational videos and handouts (Education; n=13), (2) education intervention plus mobile self-monitoring with email feedback (Monitoring; n=14), or (3) education and monitoring interventions plus social cognitive theory-based behavioral videos (Combined; n=13).

[f]PD: peritoneal dialysis.

[g]PDA: personal digital assistant.

Theories or Models Involved

Among the studies included, 14 [14,15,29,31-33,36,37,39,41,43,45-47] did not explicitly mention the theories or models involved. Of these studies, 5 [14,15,45-47] were in patients undergoing PD, and 9 [29,31-33,36,37,39,41,43] involved patients undergoing hemodialysis. Nine [13,20-27] studies combined 2 or more theories or models, and 11 [12,28,30,34,35,38,40,42,44,48,49] studies involved only 1 theory or model.

Among mHealth-based self-management interventions for patients undergoing dialysis, the most frequently incorporated theories included the transtheoretical model (TTM) and stages of change (n=5 [20,22,24,26,42]), self-efficacy theory (n=4 [12,21,24,35]), social cognitive theory (n=4 [25,27,40,48]), and Orem's theory of self-care (n=3 [21,23,28]). Most studies have focused their attention on theories related to behavior change (such as the TTM, the theory of reasoned action [20,22], the health belief model [13,34], behavior change frameworks [30,49], and the behavioral change techniques taxonomy [38]). Additionally, Bandura self-efficacy theory and Oren self-care theory were often used in combination with other theories. Accordingly, we visualized the theories or models involved in this study (Figure 2B).

Attributes of Self-Management in Patients Undergoing Dialysis

We categorized the self-management included in this study into six main themes, which were (1) self-monitoring, (2) diet and fluid management, (3) medication management, (4) exercise management, (5) psychological management, and (6) disease-related knowledge. The components of self-management for patients undergoing dialysis and their corresponding explanations are detailed in Figure 3.

XSL•FO
RenderX

Figure 3. Components of self-management for patients undergoing dialysis. AVF: arteriovenous fistula; AVG: arteriovenous graft; IDWG: intradialytic weight gain; iPTH: intact parathyroid hormone; PTH: parathyroid hormone.

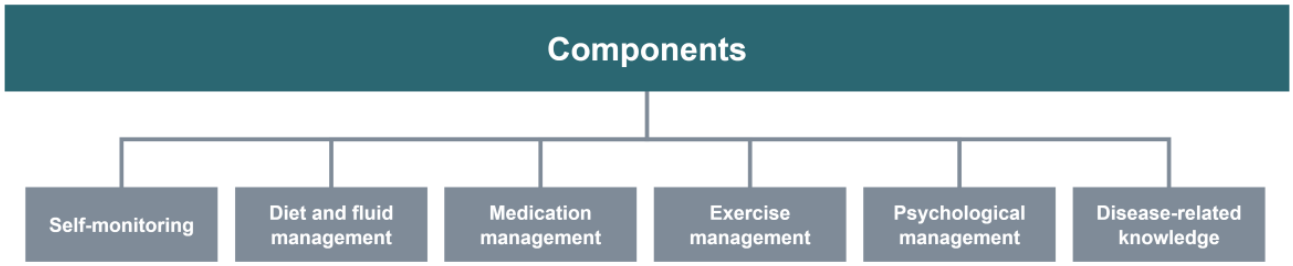

Components	Content	Meaning
Self-monitoring	Weight	Morning and evening weights, pre- and postdialysis weights, dry weights, IDWG, and BMI.
	Vital signs	Pay closer attention to changes in patients' blood pressure before and after dialysis.
	Body fluid	24-hour fluid intake and urine output.
	Symptom	Edema, chest tightness, or dyspnea etc.
	Dialysis indicators	Dialysis duration, number of dialyses, ultrafiltration volume, and adequacy of dialysis.
	Vascular access	The appearance, tremor, and murmur of the AVF or AVG.
	Laboratory tests	Blood glucose, serum potassium, serum sodium, serum phosphorus, serum calcium, serum albumin, iPTH, PTH, alkaline phosphatase, and hemoglobin.
Diet and fluid management	Diet	Control the intake of phosphorus, sodium, potassium, protein, and total calories in food.
	Fluid	Drink water up to the target values.
Medication management	Medication adherence	Accuracy of dosage and timing of administration.
	Drug reactions	Regularly record symptoms and adverse reactions after taking medication.
Exercise management	Exercise	Select the type, duration, and intensity of exercise based on their physical condition.
Psychological management	Psychology	Reduce negative emotions (depression and anxiety) and enhance emotional management skills.
Disease-related knowledge	Knowledge	Knowledge of the disease includes renal function, principles and treatment of dialysis, maintenance of dialysis access (eg, AVF, peritoneal dialysis tubing), prevention of complications, management of diet and fluids, effects and adverse reactions of medications, exercise and rehabilitation, and emotional adjustment.

mHealth-Based Self-Management Intervention Program

The mHealth-based self-management intervention program is focused on the type of mHealth, the content of the mHealth-based intervention, and the duration of the intervention.

The types of mHealth mentioned in this study included app (based on mobile device, tablet personal computer, or personal digital assistant), a remote patient monitoring system, SMS text messaging, and telephones. Apps were divided into 2 categories, such as dialysis-specific software and instant messaging software (eg, WeChat [Tencent Holdings Limited], Line [LY Corporation], Facebook [Meta Platforms, Inc], WhatsApp [Meta Platforms, Inc], and KakaoTalk [Kakao Corp]). Two studies [29,30] intervened via SMS text messaging only, 1 study [31] via telephone only, 4 studies [15,24,33,42] via instant messaging software only, and 18 studies [12-14,20-22,25,26,28,34-36,38,41,43,44,47,49] via dialysis-specific software only. Six studies [23,27,37,39,40,48] used 2 or more online interventions. There were only 3 studies [32,45,46] based on a remote patient monitoring system, 1 [32] for patients undergoing hemodialysis and the remaining 2 [45,46] for patients undergoing PD.

This scope review categorized the app's content into the following dimensions, including intelligent education hub, full-dimensional monitoring system, accurate nutritional management, behavioral interventions (fluid, exercise, and medication adherence), intelligent reminders and alerts, doctor-patient collaboration network, and social support system. Remote patient monitoring systems placed more emphasis on remote monitoring, alarms, and dynamic interventions for patients undergoing dialysis. The details could be found in Table 2. Interventions delivered via apps and remote patient monitoring systems were more comprehensive and satisfactory than those delivered via phone and SMS text messaging, even though they showed similar functionalities in certain aspects. Figure 4 illustrates the gap map of core self-management interventions across different mHealth categories for patients undergoing dialysis (blue circles indicate studies with hemodialysis as the research participants; purple circles indicate

XSL•FO
RenderX

studies with peritoneal dialysis as the research participants; red circles indicate studies with hemodialysis and peritoneal dialysis as research participants).

XSL•FO
RenderX

Table . Type of mobile health and mobile health–based interventions to improve self-management in patients undergoing dialysis.

Classification of mHealth[a]	Numbers	Function and intervention contents
App		

Dialysis-specific software	24	• Intelligent education hub
		  • Disease knowledge base: kidney function, dialysis principles, complication prevention, medications, and lifestyle guidance presented through animated videos, podcasts, manuals, and charts
		  • Operating system training: training for new patients (including equipment operation instruction) and periodic knowledge reinforcement.
		• Full-dimensional monitoring system
		  • Automatic acquisition and manual entry of data: physiological indicators, dialysis parameters, symptom logs, and medication records
		  • Visualization analysis: trend analysis and correlation analysis.
		• Accurate nutritional management
		  • Database support: 500+kinds of kidney disease-specific food nutrients, nutrition calculator (real-time display of phosphorus, sodium, potassium, and protein)
		  • Recording function: barcode scanning to enter food, manual recording of intake
		  • Analysis and feedback: nutritional value calculation for each meal, health scoring system (weekly or monthly summary), electrolyte excess warning (sodium, potassium, and phosphorus)
		  • Behavioral interventions: personalized dynamic recipe recommendations (adjusted based on lab data), daily water intake limits (residual urine volume algorithm)
		• Behavioral interventions (fluid, exercise, and medication adherence)
		  • Personalized goal management: goal setting, progress tracking (instant values + trend charts + health scores)
		  • Behavior shaping tools: badge reward mechanism (continuous recording of achievements)
		• Intelligent reminders and alerts
		  • Treatment reminder: dialysis time and follow-up appointment
		  • Medication reminder: medication time, dosage, and drug interaction
		  • Early warning system: set thresholds to trigger abnormal alerts and notify patients and health care at the same time.
		• Doctor-patient collaboration network
		  • Prescription cloud adjustment, test result synchronization (direct connection to electronic medical records), and equipment interconnection.
		• Social support system
		  • Patient community: experience sharing (recipes or exercise programs)
		  • Family linkage: caregiver collaborative recording function and family health data sharing.
		• System Infrastructure
		  • System compatibility: mobile (iOS/Android)+Web+PDA[b] compatibility, offline data caching, and synchronization
		  • Intelligent devices: support joint NFC[c], OCR[d], and PDA terminals
		  • Multimodal recording: support voice recording, image recognition, and manual supplementation.

XSL•FO
RenderX

Classification of mHealth[a]	Numbers	Function and intervention contents
Instant messaging software	5	• Knowledge delivery and personalized guidance: basic health manuals, video educational resources, and customized content delivery • Dietary interventions and guidance • Psychological intervention: structured facilitation (motivational interviewing), psychological facilitation (drawing healing-experts' video guidance), and emotional support (full psychological status tracking) • Instant interaction: online Q&A[e] with health care, regular phone consultations (psychological support and health guidance), and videoconference support (group discussions or personalized guidance).

A remote patient monitoring system	3	• Intelligent monitoring hub
		  • Device interconnection: bidirectional communication with automatic peritoneal dialysis devices through a cloud platform
		  • Full-dimensional data collection: real-time access to dialysis parameters, device alarm logs, and patient physiological indicators.
		• Remote dynamic intervention
		  • Intelligent alarms: yellow and red alarms, instantly triggering phone interventions;
		  • Prescription cloud adjustment: physicians remotely optimize automated peritoneal dialysis prescription parameters based on real-time data
		  • Physician-patient collaboration platform: support real-time treatment issues through system messages and phone calls.

Telephones	4	• Health education • Treatment adherence assessment • Personalized dietary recommendations • Psychological support • Problem solving
SMS text messaging	4	• Personalized health education • Regular collection of patients' feedback • Regular medication reminders • Positive motivational text messages

[a]mHealth: mobile health.

[b]PDA: personal digital assistant.

[c]NFC: near field communication.

[d]OCR: optical character recognition.

[e]Q&A: question-and-answer.

XSL•FO
RenderX

Figure 4. Gap map of core self-management interventions across different mobile health (mHealth) categories for patients undergoing dialysis.

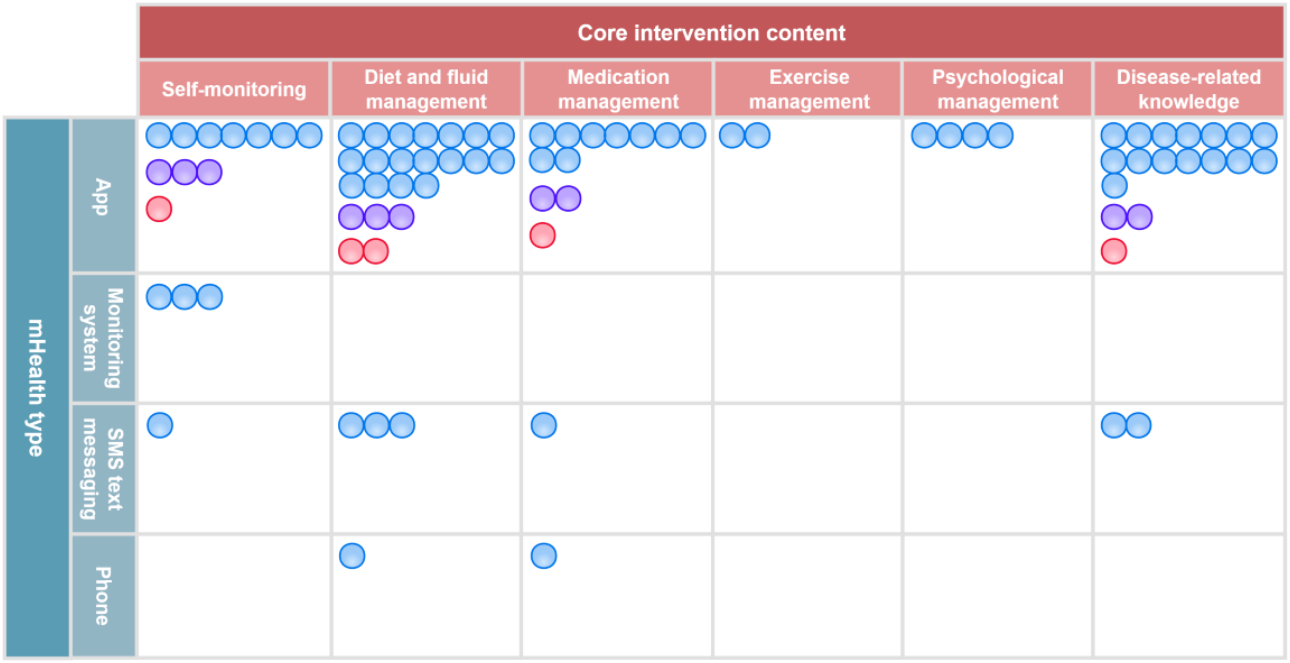

The duration of the intervention ranged from 2 to 24 weeks. The included studies preferred interventions for 3 months [12,24,36,39,43,49], 6 months [21,30,37,40], 12 weeks [26,31,33,41], 2 weeks [20,22,32], 8 weeks [23,29,35].

Key Indicators and Evaluation Tools on Self-Management

Indicators for assessing self-management in patients undergoing dialysis were divided into 2 categories, namely subjective and objective indicators. The subjective indicators mainly involved scales, such as measuring self-management, adherence, self-efficacy, literacy, depression, anxiety, perceived benefits, and quality of life of patients undergoing dialysis. Objective indicators included weight, blood pressure, and laboratory tests. Of these, intradialytic weight gain, serum albumin, serum potassium, and serum phosphorus were the most mentioned indicators in the included studies. Detailed information is provided in Table 3. Figure 5 illustrates the gap map of self-management assessment indicators across different mHealth categories for patients undergoing dialysis (blue circles indicate studies with hemodialysis as the research participants; purple circles indicate studies with peritoneal dialysis as the research participants; red circles indicate studies with hemodialysis and peritoneal dialysis as research participants).

XSL•FO
RenderX

Table . Mobile health–based indicators for assessing self-management in patients undergoing dialysis.

Items	Numbers	Tools
Self-management	4	• The self-management scale for patients undergoing hemodialysis [24,50] • The self-care performance questionnaire [21] • The Perceived Medical Condition Self-Management Scale (PKDSMS) [51] • The Patient Activation Measure-13 (PAM-13) [52]
Adherence	17	• The End-Stage Renal Disease Adherence Questionnaire (ESRD-AQ) [53] • Hemodialysis Compliance Questionnaire [43] • The Simplified Medication Adherence Questionnaire (SMAQ) [54] • The Dialysis Diet and Fluid Nonadherence Questionnaire (DDFQ) [55] • The Modified Morisky Scale (MMS) • The Compliance of Patient Role Behavior Tool [23] • The Sick-role Behavior Adherence [13] • The Adherence Questionnaire (self-developed or revised version)
Self-efficacy	14	• The 15-item dietary self-efficacy questionnaire [56] • The decision self-efficacy scale [57] • The 11-item Fluid Self-Efficacy Scale (FS-ES) [25] • The 6-item Chronic Disease Self-Efficacy Scale (CDSES) [24] • The Self-Efficacy for Appropriate Medication Use Scale (SEAMS) [58] • The strategies used by people to promote health • The Self-Efficacy Scale (self-developed or revised version) [59]
Perceived benefits	2	• The Benefits of Sodium Adherence (BSA) [25,60] • The 9-item Benefits of Fluid Adherence Scale [25,61]
Literacy	2	• The Media Health Literacy Questionnaire (MeHLit) [62] • The Health Literacy Questionnaire (HLQ) [63]
Knowledge	9	• Self-developed or revised knowledge-related questionnaires
Quality of life	12	• Kidney Disease Quality of Life Instrument-Short Form (KDQOL-SF) [64] • The Kidney Disease Quality of Life (KDQOL-36) [65] • Short Form 36 (SF-36) • EuroQol Five Dimensions Questionnaire (EQ-5D) • Health-Related Quality of Life (HRQoL) • 9-item Thai Health Status Assessment Instrument (9-THAI) [66]

XSL•FO
RenderX

Items	Numbers	Tools
Diet and nutrition	9	• 24-hour dietary recall method • 3-day dietary recall method • App automatic calculation • Food Frequency Questionnaires (FFQ) • Malnutrition Inflammation Score (MIS) • Healthy Eating Index (HEI-20)
Psychosocial	3	• Beck Depression Inventory (BDI) • Beck Anxiety Inventory (BAI)
Weight, dry weight	5	• Medical equipment
IDWG[a]	12	• A calculation formula
Blood pressure	3	• Medical equipment
Blood urea nitrogen	4	• Blood sample collection
Creatinine	5	• Blood sample collection
Urea	2	• Blood sample collection
KT/V, dialysis adequacy	5	• A calculation formula
Serum albumin	14	• Blood sample collection
Serum sodium	2	• Blood sample collection
Serum calcium	5	• Blood sample collection
Serum potassium	14	• Blood sample collection
Serum phosphorus	20	• Blood sample collection
Serum aluminum	1	• Blood sample collection
Serum iron	1	• Blood sample collection
Hemoglobin	6	• Blood sample collection
Glycosylated hemoglobin	1	• Blood sample collection
Hematocrit	1	• Blood sample collection
Bicarbonate	1	• Blood sample collection
Alkaline phosphatase	1	• Blood sample collection
iPTH[b]	2	• Blood sample collection
PTH[c]	4	• Blood sample collection
Brain natriuretic peptide	1	• Blood sample collection
C-reactive protein	1	• Blood sample collection

[a]IDWG: intradialytic weight gain.

[b]iPTH: intact parathyroid hormone.

[c]PTH: parathyroid hormone.

XSL•FO
RenderX

Figure 5. Gap map of self-management assessment indicators across different mobile health (mHealth) categories for patients undergoing dialysis.

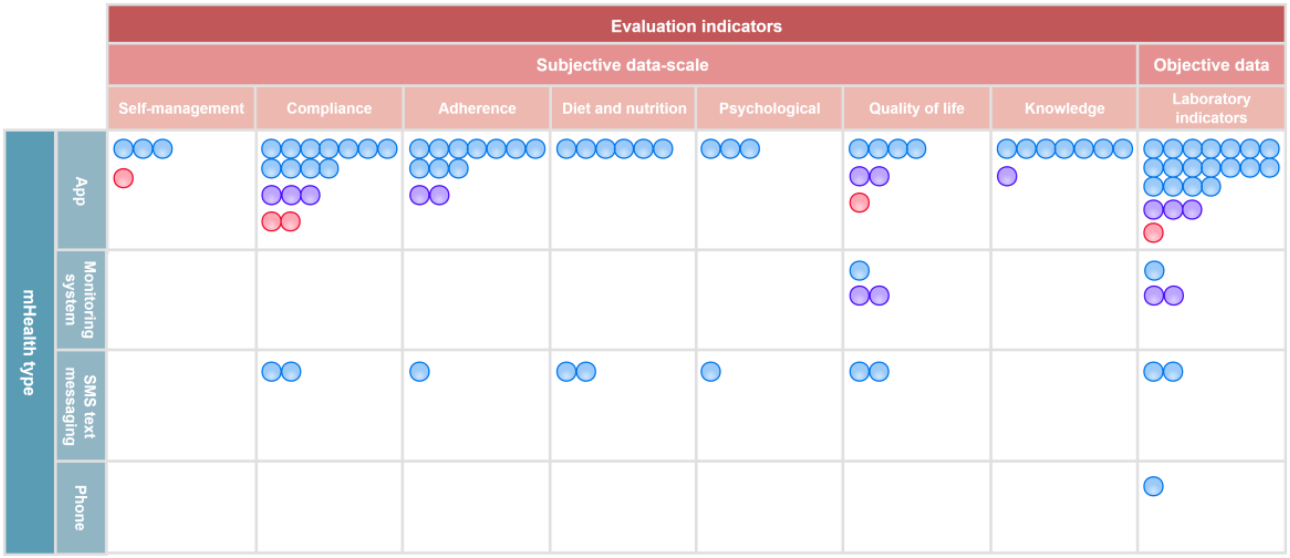

Discussion

Principal Findings

This scoping review provides the first comprehensive analysis and summary of self-management content for patients undergoing dialysis, types of mHealth, relevant theories and models, the content of mHealth-based interventions, and methods for evaluating their effectiveness. Most studies indicate that mHealth-based interventions significantly improve self-management of patients undergoing dialysis and enhance patient outcomes [23,27,37,39,40,48].

Compared to face-to-face training, mHealth-based interventions have a greater impact on patient adherence and laboratory outcomes [67]. In this study, apps and remote monitoring systems were more common. SMS text messaging and phone calls could be used as an aid to promote self-management in patients undergoing dialysis [68]. Due to differences in mHealth-based intervention content and evaluation indicators, quantitative analysis of apps, remote monitoring systems, phones, and SMS text messaging is extremely challenging. Overall, the vast majority of apps demonstrate strong potential for improving medication adherence, enhancing care efficiency, and increasing patient satisfaction and treatment outcomes [69,70].

Special Considerations for Specific Groups Among Patients Undergoing Dialysis

Currently, there are relatively few apps available on the market for patients undergoing dialysis. Existing research has identified 12 Android apps, 11 iOS apps, and 5 dual-platform apps closely associated with kidney disease [71]. In addition, middle-aged and older people are less receptive to apps than younger people, which is particularly reflected in lower usage rates and more negative attitudes toward apps [72]. Since middle-aged and older patients are less receptive to new things and have declining eyesight, the app should be designed to meet the special needs and usage habits of middle-aged and older patients. A survey of requirements for the app in middle-aged and older patients with CKD based on the Kano model showed that the app first had to protect privacy, followed by simplifying the data entry process, simplifying in-app navigation, and simplifying the function and number of buttons [73]. Navigation for beginners, appropriate text size, using color to distinguish different options or icons, and ensuring all buttons maintain consistency in size, labels, and spacing are also details that should be considered during software design [73]. It is also worth noting that there are also barriers that prevent the promotion of apps, such as information barriers, trustworthiness, security, compatibility, complexity, time constraints, and low mHealth literacy [68]. Therefore, in future research, user profiles can be incorporated based on patient age, dialysis modality, and digital literacy levels to automatically match interfaces [74]. During initial use, an interactive tutorial should be set up to guide patients. Beyond traditional touchscreen operations, additional features, such as voice commands, remote assistance, and offline mode, can be added to accommodate the diverse needs and usage habits of patients undergoing dialysis.

Key Areas for Self-Management in Patients Undergoing Dialysis Using mHealth

In terms of app self-management content, CKD disease-related knowledge, symptom management, medication management, provision of health insurance information, diet management, exercise guidance, and psychosocial support may be the content they need more (arranged according to the patient's needs) [73]. However, among the apps for patients undergoing dialysis in this study, the provision of health insurance information was what they lacked. Moreover, the app in this study focused more on disease knowledge, self-monitoring, diet, and medication management and lacked sufficient attention to exercise and psychological support. Compared to self-monitoring, dietary self-management becomes significantly more challenging for patients undergoing dialysis. Strict dietary restrictions, dynamic adjustments to meal plans (based on laboratory indicators), the need for specialized knowledge, lack of external supervision (eg, hospital and home), and the influence of long-standing dietary habits and psychological factors further compound the complexity of implementing dietary management for patients undergoing dialysis. Currently, mHealth's promotion offers

XSL•FO
RenderX

multiple solutions for dietary management among patients undergoing dialysis. The current app can automatically calculate the calories, protein, sodium, phosphorus, and potassium of foods consumed by patients undergoing dialysis, typically achieved through a nutritional database and barcode scanning functionality [25]. A study has also monitored the diet of patients undergoing dialysis by sending photos of food through an app and having them evaluated by the expert [15]. Food analysis and feedback also include the health system score, which is an easy-to-understand health score calculated by the app based on the phosphorus-to-protein ratio of the food [12]. Health care professionals provide personalized advice through dietary records and laboratory tests to promote changes in dietary behavior in patients undergoing dialysis.

Among the included studies, there were fewer studies for patients undergoing PD compared to patients undergoing hemodialysis. This may be related to the late start of PD. Self-management is an important influence on the quality of life and outcomes of patients undergoing PD [75]. Patients undergoing PD who undergo inappropriate operations face a higher risk of developing peritonitis [76]. If patients undergoing PD experience a technical failure, they must be treated with hemodialysis. Compared with PD, long-term regular hemodialysis significantly impacts patients' quality of life, hinders their social reintegration, and increases the risk of complications [77,78]. Therefore, for patients undergoing PD, self-management is necessary for them to master. The focus of self-management in patients undergoing PD is not the same as in patients undergoing hemodialysis. As patients undergoing PD need to perform peritoneal dialysis-related operations at home, training in relevant knowledge and operational skills is particularly important [76]. The training of skills mainly includes aseptic operation, change of peritoneal fluid and peritoneal dialysis catheter, and outlet care [79]. Patients undergoing PD also need to learn the calculation of ultrafiltration [79]. All of these can be learned and managed remotely based on mHealth. In the event of an emergency (eg, contamination or disconnection of the peritoneal dialysis tubing), patients should immediately contact a health care professional for on-site guidance. Thus, establishing emergency contact channels within the app is particularly crucial.

Theoretical or Model Support for mHealth Intervention Programs in Patients Undergoing Dialysis

Many theories and models have been used to guide practice in the self-management of patients undergoing dialysis. TTM, self-efficacy theory, and social cognitive theory are the most commonly used theories [12,20,40,42]. Currently, most studies use a single theory as their guiding framework, with fewer adopting 2 or more theories for guidance. Taking the TTM as an example, it emphasizes guiding patients through staged behavioral shifts but overlooks individual differences (such as cultural background and cognitive level) [20,22,24,26,42]. Self-care theory, on the other hand, emphasizes personalized care needs and can be combined with TTM. Additionally, the maintenance phase of TTM is prone to behavioral relapse, particularly among patients requiring lifelong treatment (eg, dialysis). A single theory cannot sustainably motivate patients. Integrating other theories (such as the PERMA [Positive Emotion, Engagement, Relationship, Meaning and Accomplishment] model [80]) during this phase can further reinforce patients' behavior. Therefore, future research should focus more on integrating multiple theories rather than relying on a single one. In addition, goal-setting theory, the information-motivation-behavioral skills model, the chronic care model, the ecological model of health behavior, and the theory of planned behavior can also be applied in subsequent research.

Implications for Future Research

Future research on self-management software for patients undergoing dialysis will focus on integrating the strengths of existing tools and promoting their synergistic use. Software development should break down barriers between tools, such as deeply integrating the real-time data collection capabilities of remote patient monitoring systems, the personalized analysis functions of intelligent decision support systems, and the instant communication advantages of phone or SMS text messaging, to provide patients with more comprehensive self-management support.

A combination of artificial intelligence can be considered for use in the remote management of patients undergoing dialysis (eg, wearable devices). Smart wristbands can monitor data, such as heart rate, blood pressure, activity, and sleep [81] and automatically transmit the data to the app for analysis and storage. Some researchers have implemented data linkage through apps using near field communication and optical character recognition. Data from measuring devices, such as sphygmomanometers and weight scales, can be automatically transmitted to the app via near field communication [14]. Alternatively, the numbers from the sphygmomanometer can be captured using the phone's camera and imported into the app [14]. These not only improve the efficiency of patients undergoing dialysis but also increase the accuracy of the records. In conclusion, there is still a lot of room to explore the use of artificial intelligence in combination with mHealth in patients undergoing dialysis.

Different studies have different insights into the evaluation criteria for self-management. Most of the studies used objective indicators as one of the evaluation criteria for self-management. For the evaluation of knowledge of patients undergoing dialysis, most studies have assessed it using self-developed questionnaires and lacked uniform criteria for judging. Various scales are currently available for assessing self-management, adherence, self-efficacy, and quality of life. The use of these assessment tools varies across different studies. Existing research lacks a gold standard for evaluating self-management in patients undergoing dialysis. Therefore, there is a need to standardize the criteria for evaluating self-management in patients undergoing dialysis in future studies.

This study provides guidance for the development of subsequent dialysis-related software for patients, including the design of functional modules, user experience optimization, and the integration of clinical indicators and assessment tools. Our findings will enhance the ability of patients undergoing dialysis to self-manage their health at home, improving both the effectiveness of their dialysis treatment and their quality of life.

XSL•FO
RenderX

In the future design of the app, attention should also be paid to the usage needs of special groups, strengthening adaptive software development, deeply integrating artificial intelligence technology, and establishing standardized self-management evaluation criteria.

Limitation

This scoping review also has some limitations. First, the broad scope of scoping reviews and the complexity of search strategies may lead to the omission of relevant studies. These challenges persist despite strict adherence to the PRISMA-ScR guidelines for greater rigor and transparency. Second, self-management assessment tools incorporate numerous subjective and objective indicators, which complicate data integration and comparability. Third, publication bias was one of the limitations of this study. This review only covered relevant studies published in English. Most of the included studies were limited to Asia and North America, which may have resulted in limited global generalizability. Future studies should use standardized, integrated measures to improve consistency and reliability. Additionally, these shortcomings can be remedied by improving search strategies, expanding database coverage, and removing language restrictions to include a more diverse patient population from different continents.

Conclusion

This study conducted a scoping review of the existing literature on mHealth-based self-management among patients undergoing dialysis. Although mHealth holds potential advantages for self-management in patients undergoing dialysis, it has not been widely adopted and integrated into standard renal care, requiring further optimization and refinement. This study provides theoretical and practical guidance for subsequent research, helping to enhance self-management of patients undergoing dialysis and improve their quality of life, ultimately offering insights for digital transformation in chronic disease management.

Multimedia Appendix 1
Search strategy
[DOCX File, 17 KB - jmir_v28i1e76880_app1.docx]
Multimedia Appendix 2
Critical appraisal of the selected studies using the Mixed Methods Appraisal Tool (MMAT).
[XLSX File, 72 KB - jmir_v28i1e76880_app2.xlsx] Multimedia Appendix 3
Types of mobile health, core intervention content, intervention time, and evaluation indicators for self-management in patients undergoing dialysis.
[DOCX File, 45 KB - jmir_v28i1e76880_app3.docx] Checklist 1
PRISMA-ScR and PRISMA-S checklists.
[PDF File, 315 KB - jmir_v28i1e76880_app4.pdf]

Acknowledgments

We are grateful to the editors and reviewers who helped with this study.

Funding

This work was supported by grants from the Tongji Hospital Research Fund Project (2024D10, 2022D22). The funder participated in the research design, data collection, analysis, interpretation, and manuscript writing.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Conflicts of Interest

None declared.

References

1. See E, Ethier I, Cho Y, et al. Dialysis outcomes across countries and regions: a global perspective from the International Society of Nephrology Global Kidney Health Atlas study. Kidney Int Rep 2024 Aug;9(8):2410-2419. [doi: 10.1016/j.ekir.2024.05.014] [Medline: 39156158]
2. Bello AK, Okpechi IG, Levin A, et al. An update on the global disparities in kidney disease burden and care across world countries and regions. Lancet Glob Health 2024 Mar;12(3):e382-e395. [doi: 10.1016/S2214-109X(23)00570-3] [Medline: 38365413]
3. Bastani B. The present and future of transplant organ shortage: some potential remedies. J Nephrol 2020 Apr;33(2):277-288. [doi: 10.1007/s40620-019-00634-x] [Medline: 31399908]
4. He Z, Hou H, Zhang D, et al. Effects of dialysis modality choice on the survival of end-stage renal disease patients in southern China: a retrospective cohort study. BMC Nephrol 2020 Sep 24;21(1):412. [doi: 10.1186/s12882-020-02070-7] [Medline: 32972378]
5. Chisavu L, Mihaescu A, Bob F, et al. Trends in mortality and comorbidities in hemodialysis patients between 2012 and 2017 in an East-European Country: a retrospective study. Int Urol Nephrol 2023 Oct;55(10):2579-2587. [doi: 10.1007/s11255-023-03549-6] [Medline: 36917413]

XSL•FO
RenderX

6. Al-Mansouri A, Al-Ali FS, Hamad AI, et al. Assessment of treatment burden and its impact on quality of life in dialysis-dependent and pre-dialysis chronic kidney disease patients. Res Social Adm Pharm 2021 Nov;17(11):1937-1944. [doi: 10.1016/j.sapharm.2021.02.010] [Medline: 33612446]
7. Pretto CR, Winkelmann ER, Hildebrandt LM, Barbosa DA, Colet CDF, Stumm EMF. Quality of life of chronic kidney patients on hemodialysis and related factors. Rev Lat Am Enfermagem 2020;28:e3327. [doi: 10.1590/1518-8345.3641.3327] [Medline: 32696925]
8. Lorig KR, Sobel DS, Ritter PL, Laurent D, Hobbs M. Effect of a self-management program on patients with chronic disease. Eff Clin Pract 2001;4(6):256-262. [Medline: 11769298]
9. Almutary H, Tayyib N. Evaluating self-efficacy among patients undergoing dialysis therapy. Nurs Rep 2021 Mar 23;11(1):195-201. [doi: 10.3390/nursrep11010019] [Medline: 34968324]
10. Escudero-Lopez M, Martinez-Andres M, Marcilla-Toribio I, Moratalla-Cebrian ML, Perez-Moreno A, Bartolome-Gutierrez R. Barriers and facilitators in self-care and management of chronic kidney disease in dialysis patients: a systematic review of qualitative studies. J Clin Nurs 2024 Oct;33(10):3815-3830. [doi: 10.1111/jocn.17193] [Medline: 38716807]
11. Yang Y, Chen H, Qazi H, Morita PP. Intervention and evaluation of mobile health technologies in management of patients undergoing chronic dialysis: scoping review. JMIR Mhealth Uhealth 2020 Apr 3;8(4):e15549. [doi: 10.2196/15549] [Medline: 32242823]
12. Chiang YC, Chang YP, Lin SC, et al. Effects of individualized dietary phosphate control program with a smartphone application in hemodialysis patients in Taiwan in China. Biol Res Nurs 2021 Jul;23(3):375-381. [doi: 10.1177/1099800420975504] [Medline: 33251815]
13. Min Y, Park M. Effects of a mobile-app-based self-management support program for elderly hemodialysis patients. Healthc Inform Res 2020 Apr;26(2):93-103. [doi: 10.4258/hir.2020.26.2.93] [Medline: 32547806]
14. Lukkanalikitkul E, Kongpetch S, Chotmongkol W, et al. Optimization of the chronic kidney disease-peritoneal dialysis app to improve care for patients on peritoneal dialysis in northeast Thailand: user-centered design study. JMIR Form Res 2022 Jul 6;6(7):e37291. [doi: 10.2196/37291] [Medline: 35793137]
15. Lee HJ, Kang HY. Effects of a customized diet education program using a mobile instant messenger for people undergoing peritoneal dialysis: a feasibility test. Asian Nurs Res (Korean Soc Nurs Sci) 2024 Oct;18(4):367-376. [doi: 10.1016/j.anr.2024.09.007] [Medline: 39284546]
16. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
17. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
18. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc 2021 Jan;109(2):34285662. [doi: 10.5195/jmla.2021.962]
19. Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018;34(4):285-291. [doi: 10.3233/EFI-180221]
20. Fakih El Khoury C, Crutzen R, Schols JM, Halfens RJ, Karavetian M. Adequate management of phosphorus in patients undergoing hemodialysis using a dietary smartphone app: prospective pilot study. JMIR Form Res 2021 Jun 1;5(6):e17858. [doi: 10.2196/17858] [Medline: 34061034]
21. Hosseini A, Jackson AC, Chegini N, et al. The effect of an educational app on hemodialysis patients' self-efficacy and self-care: A quasi-experimental longitudinal study. Chronic Illn 2023 Jun;19(2):383-394. [doi: 10.1177/17423953211073365] [Medline: 35179394]
22. Fakih El Khoury C, Crutzen R, Schols J, Halfens RJG, Karavetian M. A dietary mobile app for patients undergoing hemodialysis: prospective pilot study to improve dietary intakes. J Med Internet Res 2020 Jul 20;22(7):e17817. [doi: 10.2196/17817] [Medline: 32706698]
23. Park OL, Kim SR. Integrated self-management program effects on hemodialysis patients: a quasi-experimental study. Jpn J Nurs Sci 2019 Oct;16(4):396-406. [doi: 10.1111/jjns.12249] [Medline: 30669185]
24. Ren Q, Lian M, Liu Y, Thomas-Hawkins C, Zhu L, Shen Q. Effects of a transtheoretical model-based WeChat health education programme on self-management among haemodialysis patients: a longitudinal experimental intervention study. J Adv Nurs 2019 Dec;75(12):3554-3565. [doi: 10.1111/jan.14182] [Medline: 31441525]
25. Welch JL, Astroth KS, Perkins SM, et al. Using a mobile application to self-monitor diet and fluid intake among adults receiving hemodialysis. Res Nurs Health 2013 Jun;36(3):284-298. [doi: 10.1002/nur.21539] [Medline: 23512869]
26. Zwi S, Isautier J, Webster AC, et al. A feasibility study of a best practice health literacy app for Australian adults with chronic kidney disease. PEC Innov 2022 Dec;1:100047. [doi: 10.1016/j.pecinn.2022.100047] [Medline: 37213763]
27. Chae YJ, Kim HS. Effects of a mobile application on improving self-management of adult patients receiving peritoneal dialysis: a randomized controlled trial. Jpn J Nurs Sci 2024 Jan;21(1):e12555. [doi: 10.1111/jjns.12555] [Medline: 37589209]
28. Cho H, Park S. The effects of self‐performance management video program on patients receiving hemodialysis. Jpn J Nurs Sci 2020 Apr;17(2):e12303. [doi: 10.1111/jjns.12303] [Medline: 31746139]

XSL•FO
RenderX

29. Chung MK, Kim DH, Park JI, et al. Adaptive nutrition intervention stabilizes serum phosphorus levels in hemodialysis patients: a multicenter decentralized clinical trial using real-world data. J Ren Nutr 2024 Jan;34(1):47-57. [doi: 10.1053/j.jrn.2023.07.004] [Medline: 37586668]
30. Dawson J, Campbell KL, Craig JC, et al. A text messaging intervention for dietary behaviors for people receiving maintenance hemodialysis: a feasibility study of KIDNEYTEXT. Am J Kidney Dis 2021 Jul;78(1):85-95. [doi: 10.1053/j.ajkd.2020.11.015] [Medline: 33421456]
31. Hanifi N, Ezzat LS, Dinmohammadi M. Effect of consultation and follow-up phone calls on biochemical indicators and intradialytic weight gain in patients undergoing hemodialysis. Oman Med J 2019 Mar;34(2):137-146. [doi: 10.5001/omj.2019.26] [Medline: 30918608]
32. Hayashi A, Yamaguchi S, Waki K, et al. Testing the feasibility and usability of a novel smartphone-based self-management support system for dialysis patients: a pilot study. JMIR Res Protoc 2017 Apr 20;6(4):e63. [doi: 10.2196/resprot.7105] [Medline: 28428168]
33. Mollaoğlu M, Candan F, Solmaz G, Mollaoğlu S, Başer E, Yanmiş S. The effect of education and art therapy with telehealth method on diet-fluid restriction and anxiety in hemodialysis patients during the COVID-19 pandemic. Altern Ther Health Med 2024 Jul;30(7):58-64. [Medline: 39110042]
34. Nursalam N, Kurniawati ND, Putri IRP, Priyantini D. Automatic reminder for fluids management on confidence and compliance with fluid restrictions in hemodialysis patients. Sys Rev Pharm 2020 Jun 1;11(5):226-233. [doi: 10.31838/srp.2020.5.34]
35. Pack S, Lee J. Randomised controlled trial of a smartphone application-based dietary self-management program on haemodialysis patients. J Clin Nurs 2021 Mar;30(5-6):840-848. [doi: 10.1111/jocn.15627] [Medline: 33377565]
36. Pinto LCS, Andrade MC, Chaves RO, et al. Development and validation of an application for follow-up of patients undergoing dialysis: NefroPortátil. J Ren Nutr 2020 Jul;30(4):e51-e57. [doi: 10.1053/j.jrn.2019.03.082] [Medline: 32081517]
37. Pungchompoo W, Parinyachitta S, Pungchompoo S, Udomkhwamsuk W, Suwan P. The feasibility of integrating a home telehealth model for older persons living with hemodialysis. BMC Geriatr 2024 Apr 27;24(1):378. [doi: 10.1186/s12877-024-04981-8] [Medline: 38671357]
38. Rocco MV, Rigaud M, Ertel C, Russell G, Zemdegs J, Vecchio M. Fluid intake management in maintenance hemodialysis using a smartphone-based application: a pilot study. Kidney Med 2023 Sep;5(9):100703. [doi: 10.1016/j.xkme.2023.100703] [Medline: 37663954]
39. Saadatifar B, Sharifi S, Faghihi H, Sadeghi Googhary N. Effect of mHealth training on treatment adherence in hemodialysis patients. Med Surg Nurs J 2022;11(3):1-8. [doi: 10.5812/msnj-134851]
40. St-Jules DE, Woolf K, Goldfarb DS, et al. Feasibility and acceptability of mHealth interventions for managing hyperphosphatemia in patients undergoing hemodialysis. J Ren Nutr 2021 Jul;31(4):403-410. [doi: 10.1053/j.jrn.2020.07.009] [Medline: 33160812]
41. Teong LF, Khor BH, Ng HM, et al. Effectiveness of a nutritional mobile application for management of hyperphosphatemia in patients on hemodialysis: a multicenter open-label randomized clinical trial. J Pers Med 2022 Jun 12;12(6):961. [doi: 10.3390/jpm12060961] [Medline: 35743746]
42. Thongsunti A, Silpakit C, Rattananupong T, Kittanamongkolchai W, Sumethpimolchai W, Lohsoonthorn V. Effect of a transtheoretical model-based intervention and motivational interviewing on hyperphosphatemia management via telehealth (TMT program) among hemodialysis patients during the COVID-19 pandemic. Front Public Health 2024;12:1361778. [doi: 10.3389/fpubh.2024.1361778] [Medline: 39668955]
43. Andriati R. Innovation of the M-Nursing application to improve compliance with haemodialysis therapy and urea-creatinine levels in patients with chronic kidney disease. Malays J Nurs 2025;17(1) [FREE Full text] [doi: 10.31674/mjn.2025.v17i01.011]
44. Taguiam DJ. Optimizing fluid intake management in adult hemodialysis patients: the impact of the H2Overload mobile health app on interdialytic weight gain [Dissertation]. : William Paterson University; 2025 URL: http://hdl.handle.net/20.500.12164/3510 [accessed 2025-12-06]
45. Uchiyama K, Morimoto K, Washida N, et al. Effects of a remote patient monitoring system for patients on automated peritoneal dialysis: a randomized crossover controlled trial. Int Urol Nephrol 2022 Oct;54(10):2673-2681. [doi: 10.1007/s11255-022-03178-5]
46. Jung HY, Jeon Y, Kim YS, et al. Outcomes of remote patient monitoring for automated peritoneal dialysis: a randomized controlled trial. Nephron 2021;145(6):702-710. [doi: 10.1159/000518364] [Medline: 34515160]
47. Zeng Y, Yin Y, Deng J, et al. Effectiveness of a smart management system in improving adherence and clinical outcomes of patients receiving peritoneal dialysis: a retrospective cohort analysis. BMC Nurs 2025 Jul 7;24(1):860. [doi: 10.1186/s12912-025-03506-x] [Medline: 40624644]
48. Stark S, Snetselaar L, Piraino B, et al. Personal digital assistant-based self-monitoring adherence rates in 2 dialysis dietary intervention pilot studies: BalanceWise-HD and BalanceWise-PD. J Ren Nutr 2011 Nov;21(6):492-498. [doi: 10.1053/j.jrn.2010.10.026] [Medline: 21420316]
49. Beer J, Jacques A, Lambert K, Lim W, Howell M, Boudville N. TELEnutrition and KIdNey hEalth Study: protocol for a randomised controlled trial comparing the effect of digital health to standard care on serum phoSphate control in patients

XSL•FO
RenderX

on dIalysiS (TeleKinesis Study). BMJ Open 2025 May 2;15(5):e096381. [doi: 10.1136/bmjopen-2024-096381] [Medline: 40316358]
50. Song Y. Developing and testing of the hemodialysis self-management instrument. : Kaohsiung Medical University; 2009.
51. Wild MG, Wallston KA, Green JA, et al. The Perceived Medical Condition Self-Management Scale can be applied to patients with chronic kidney disease. Kidney Int 2017 Oct;92(4):972-978. [doi: 10.1016/j.kint.2017.03.018] [Medline: 28528132]
52. Hibbard JH, Mahoney ER, Stockard J, Tusler M. Development and testing of a short form of the patient activation measure. Health Serv Res 2005 Dec;40(6 Pt 1):1918-1930. [doi: 10.1111/j.1475-6773.2005.00438.x] [Medline: 16336556]
53. Kim Y, Evangelista LS, Phillips LR, Pavlish C, Kopple JD. The End-Stage Renal Disease Adherence Questionnaire (ESRD-AQ): testing the psychometric properties in patients receiving in-center hemodialysis. Nephrol Nurs J 2010;37(4):377-393. [Medline: 20830945]
54. Ortega Suárez FJ, Sánchez Plumed J, Pérez Valentín MA, et al. Validation on the simplified medication adherence questionnaire (SMAQ) in renal transplant patients on tacrolimus. Nefrologia 2011;31(6):690-696. [doi: 10.3265/Nefrologia.pre2011.Aug.10973] [Medline: 22130285]
55. Vlaminck H, Maes B, Jacobs A, Reyntjens S, Evers G. The dialysis diet and fluid non-adherence questionnaire: validity testing of a self-report instrument for clinical practice. J Clin Nurs 2001 Sep;10(5):707-715. [doi: 10.1046/j.1365-2702.2001.00537.x] [Medline: 11822521]
56. Seo AR, Park KS, Kim BK, Kim YL, Choi JY. A validation of dietary self-efficacy questionnaire in hemodialysis patients. Korean J Health Promot ;12(1):22-30 [FREE Full text]
57. O'Connor AM. User Manual - Decision Self-Efficacy Scale. Ottawa Hospital Research Institute. 1995. URL: https://decisionaid.ohri.ca/docs/develop/User_Manuals/UM_Decision_SelfEfficacy.pdf [accessed 2025-12-06]
58. Risser J, Jacobson TA, Kripalani S. Development and psychometric evaluation of the Self-efficacy for Appropriate Medication Use Scale (SEAMS) in low-literacy patients with chronic disease. J Nurs Meas 2007;15(3):203-219. [doi: 10.1891/106137407783095757] [Medline: 18232619]
59. Song MR, Kim MJ, Lee ME, Lee IB, Shu MR. A study on the correlation between self-efficacy and self-care in hemodialysis patients. J Korean Acad Nurs 1999;29(3):563. [doi: 10.4040/jkan.1999.29.3.563]
60. Welch JL, Bennett SJ, Delp RL, Agarwal R. Benefits of and barriers to dietary sodium adherence. West J Nurs Res 2006 Mar;28(2):162-180. [doi: 10.1177/0193945905282323] [Medline: 16513918]
61. Welch JL. Hemodialysis patient beliefs by stage of fluid adherence. Res Nurs Health 2001 Apr;24(2):105-112. [doi: 10.1002/nur.1013] [Medline: 11353458]
62. Nazarnia M, Zarei F, Rozbahani N. Development and psychometric properties of a tool to assess Media Health Literacy (MeHLit). BMC Public Health 2022 Oct 1;22(1):1839. [doi: 10.1186/s12889-022-14221-6] [Medline: 36180875]
63. Osborne RH, Batterham RW, Elsworth GR, Hawkins M, Buchbinder R. The grounded psychometric development and initial validation of the Health Literacy Questionnaire (HLQ). BMC Public Health 2013 Jul 16;13:658. [doi: 10.1186/1471-2458-13-658] [Medline: 23855504]
64. Hays RD, Kallich JD, Mapes DL, Coons SJ, Carter WB. Development of the kidney disease quality of life (KDQOL) instrument. Qual Life Res 1994 Oct;3(5):329-338. [doi: 10.1007/BF00451725] [Medline: 7841967]
65. Ricardo AC, Hacker E, Lora CM, et al. Validation of the Kidney Disease Quality of Life Short Form 36 (KDQOL-36) US Spanish and English versions in a cohort of Hispanics with chronic kidney disease. Ethn Dis 2013;23(2):202-209. [Medline: 23530302]
66. Cheawchanwattana A, Limwattananon C, Gross C, et al. The validity of A new practical quality of life measure in patients on renal replacement therapy. J Med Assoc Thai 2006 Aug;89 Suppl 2:S207-S217. [Medline: 17044474]
67. Torabikhah M, Farsi Z, Sajadi SA. Comparing the effects of mHealth app use and face-to-face training on the clinical and laboratory parameters of dietary and fluid intake adherence in hemodialysis patients: a randomized clinical trial. BMC Nephrol 2023 Jun 29;24(1):194. [doi: 10.1186/s12882-023-03246-7] [Medline: 37386428]
68. Shen H, van der Kleij R, van der Boog PJM, et al. Digital tools/eHealth to support CKD self-management: a qualitative study of perceptions, attitudes and needs of patients and health care professionals in China. Int J Med Inform 2022 Sep;165:104811. [doi: 10.1016/j.ijmedinf.2022.104811] [Medline: 35753175]
69. Paneerselvam GS, Lua PL, Chooi WH, Rehman IU, Goh KW, Ming LC. Effectiveness of mobile apps in improving medication adherence among chronic kidney disease patients: systematic review. J Med Internet Res 2025 Apr 16;27:e53144. [doi: 10.2196/53144] [Medline: 40239197]
70. Chao SM, Pan CK, Wang ML, Fang YW, Chen SF. Functionality and usability of mHealth apps in patients with peritoneal dialysis: a systematic review. Healthcare (Basel) 2024 Mar 5;12(5):593. [doi: 10.3390/healthcare12050593] [Medline: 38470704]
71. Singh K, Diamantidis CJ, Ramani S, et al. Patients' and Nephrologists' evaluation of patient-facing smartphone apps for CKD. Clin J Am Soc Nephrol 2019 Apr 5;14(4):523-529. [doi: 10.2215/CJN.10370818] [Medline: 30898873]
72. Baer NR, Vietzke J, Schenk L. Middle-aged and older adults' acceptance of mobile nutrition and fitness apps: a systematic mixed studies review. PLoS ONE 2022;17(12):e0278879. [doi: 10.1371/journal.pone.0278879] [Medline: 36520839]

XSL•FO
RenderX

73. Yan Y, Liu M, Duan DF, Yan LJ, Li L, Ma DY. Demand analysis of self-management mobile health applications for middle-aged and older patients with chronic kidney disease based on the Kano model. Nephron 2025;149(3):166-177. [doi: 10.1159/000541729] [Medline: 39396506]
74. Gosetto L, Falquet G, Ehrler F. Personalizing mobile applications for health based on user profiles: a preference matrix from a scoping review. PLOS Digit Health 2025 Aug;4(8):e0000978. [doi: 10.1371/journal.pdig.0000978] [Medline: 40828800]
75. Huang Y, Li S, Lu X, Chen W, Zhang Y. The effect of self-management on patients with chronic diseases: a systematic review and meta-analysis. Healthcare (Basel) 2024 Oct 29;12(21):2151. [doi: 10.3390/healthcare12212151] [Medline: 39517362]
76. Jaelani TR, Ibrahim K, Jonny J, et al. Peritoneal dialysis patient training program to enhance independence and prevent complications: a scoping review. Int J Nephrol Renovasc Dis 2023;16:207-222. [doi: 10.2147/IJNRD.S414447] [Medline: 37720493]
77. Ethier I, Hayat A, Pei J, et al. Peritoneal dialysis versus haemodialysis for people commencing dialysis. Cochrane Database Syst Rev 2024 Jun 20;6(6):CD013800. [doi: 10.1002/14651858.CD013800.pub2] [Medline: 38899545]
78. Chuasuwan A, Pooripussarakul S, Thakkinstian A, Ingsathit A, Pattanaprateep O. Comparisons of quality of life between patients underwent peritoneal dialysis and hemodialysis: a systematic review and meta-analysis. Health Qual Life Outcomes 2020 Jun 18;18(1):191. [doi: 10.1186/s12955-020-01449-2] [Medline: 32552800]
79. Figueiredo AE, Bernardini J, Bowes E, et al. A syllabus for teaching peritoneal dialysis to patients and caregivers. Perit Dial Int 2016;36(6):592-605. [doi: 10.3747/pdi.2015.00277] [Medline: 26917664]
80. Madeson M. Seligman's PERMA+ model explained: a theory of wellbeing. PositivePsychology.com. 2017 Feb 24. URL: https://positivepsychology.com/perma-model/ [accessed 2025-12-16]
81. Stauss M, Htay H, Kooman JP, Lindsay T, Woywodt A. Wearables in Nephrology: Fanciful Gadgetry or Prêt-à-Porter? Sensors (Basel) 2023 Jan 26;23(3):1361. [doi: 10.3390/s23031361]

Abbreviations

CKD: chronic kidney diseases
ESRD: end-stage renal disease
mHealth: mobile health
PD: peritoneal dialysis
PERMA: Positive Emotion, Engagement, Relationship, Meaning and Accomplishment
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
TTM: transtheoretical model

Edited by S Brini; submitted 03.May.2025; peer-reviewed by D Iryawati, Y Shi; accepted 17.Nov.2025; published 02.Jan.2026.

Please cite as:
Xu Q, Xu Y, Liu X, Ma X
The Application of Mobile Health in Self-Management Among Patients Undergoing Dialysis: Scoping Review
J Med Internet Res 2026;28:e76880
URL: https://www.jmir.org/2026/1/e76880
doi:10.2196/76880

© Qin Xu, Yulin Xu, Xiaoqin Liu, Xiaolin Ma. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Accuracy of Deep Learning in Diagnosing Chronic Obstructive Pulmonary Disease: Systematic Review and Meta-Analysis

Hui Yang[1,2]; Yijiu Wu[3]; Tong Wu[1]; Jingyan Ji[2]; Sitao Lei[2]; Weibin Xu[1]

[1]School of Management, Guizhou University, Huaxi District, Guiyang, Guizhou, China

[2]School of Forensic Medicine, Guizhou Medical University, Guiyang, Guizhou, China

[3]School of Anesthesiology, Guizhou Medical University, Guiyang, Guizhou, China

Corresponding Author:
Weibin Xu
School of Management, Guizhou University, Huaxi District, Guiyang, Guizhou, China

Abstract

Background: Chronic obstructive pulmonary disease (COPD) is a common chronic lung disease. Deep learning (DL), a data-driven machine learning approach, has gained attention in clinical practice, particularly for diagnosing COPD and grading its severity. However, systematic evidence of its diagnostic and grading accuracy remains limited, posing challenges for developing intelligent diagnostic tools.

Objective: This study aimed to systematically estimate the accuracy of DL models for diagnosing and grading COPD, providing up-to-date evidence for the design and clinical implementation of intelligent detection tools.

Methods: The Cochrane Library, Embase, Web of Science, and PubMed were systematically searched for studies on DL for diagnosing COPD and grading its severity published up to November 1, 2025. Risk of bias was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 tool. Subgroup analyses by the validation set generation method and imaging data source were conducted, and meta-analyses were performed on the validation sets. For binary outcomes, diagnostic 2×2 tables were synthesized using a bivariate mixed effects model; for multiclass outcomes, accuracy estimates were pooled using random-effects models.

Results: In total, 56 studies comprising 886,753 participants were included. Inputs were computed tomography (CT) imaging (n=30), breath sounds or audio (n=12), conventional chest X-ray (n=2), X-ray film (n=2), and other modalities (n=10), including pulmonary function indices or curves or physiological waveforms, electrocardiograms, volumetric capnography maps, radiogenetic data, and clinical scores. For binary classification of COPD, DL models yielded a pooled sensitivity of 0.87 (95% CI 0.85 - 0.90), specificity of 0.88 (95% CI 0.84 - 0.92), diagnostic odds ratio (DOR) of 52 (95% CI 30 - 88), and the area under the summary receiver operating characteristic curve (AUC) of 0.93. For CT-based DL models, pooled sensitivity was 0.86 (95% CI 0.84 - 0.89), specificity was 0.87 (95% CI 0.82 - 0.90), DOR was 42 (95% CI 26 - 68), and AUC was 0.92. For respiratory sound–based models, sensitivity was 0.91 (95% CI 0.84 - 0.95), specificity was 0.96 (95% CI 0.91 - 0.98), DOR was 237 (95% CI 78 - 723), and AUC was 0.98. In multiclass classification, the DL models showed limited accuracy in discriminating Global Initiative for Chronic Obstructive Lung Disease (GOLD) stages: GOLD stage 0 (84.2%, 95% CI 60.5% - 98.2%), stage 1 (61.7%, 95% CI 40.7% - 80.8%), stage 2 (67.9%, 95% CI 37.6% - 91.7%), stage 3 (70.8%, 95% CI 16.3% - 100%), and stage 4 (70.8%, 95% CI 16.3% - 100%).

Conclusions: This study is the first systematic synthesis of DL applications for COPD detection and GOLD staging. DL models based on CT images and breath sounds show high accuracy for binary COPD detection, whereas multiclass GOLD grading remains concerning. These findings support the development and updating of artificial intelligence−assisted COPD screening tools; however, substantial heterogeneity and limited external validation warrant cautious interpretation. Future reproducible multicenter studies with standardized reporting are needed.

Trial Registration: PROSPERO CRD420251114195; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251114195

(J Med Internet Res 2026;28:e83459) doi:10.2196/83459

KEYWORDS

chronic obstructive pulmonary disease; COPD; deep learning; diagnosis; Global Initiative for Chronic Obstructive Lung Disease; GOLD grading; meta-analysis

XSL•FO
RenderX

Introduction

Chronic obstructive pulmonary disease (COPD) is a prevalent chronic respiratory illness characterized by persistent airflow limitation. It is irreversible and progressively worsens over time, severely affecting patients' quality of life and life expectancy [1]. According to the latest World Health Organization report, COPD is the fourth leading cause of death worldwide, responsible for over 3 million deaths each year, leading to a disproportionate burden in low- and middle-income countries [2]. China accounts for about one-quarter of the global burden of COPD, with an estimated 99.9 million people affected and a prevalence of 13.7% among adults aged ≥40 years [3]. Acute exacerbations are pivotal events in COPD, causing hospital admission and increasing the risk of mortality. The 5-year mortality rate after exacerbation is about 50% after hospitalization [4]. A real-world multicenter prospective cohort study in Japan has reported a 5-year survival rate of 85.4% among COPD patients, whereas those with very severe airflow limitation have a reduced 5-year survival rate of 66.1% [5]. Consequently, COPD not only represents a significant public health issue worldwide but has also become one of the main causes of disability and death.

In clinical practice, the gold standard for diagnosing COPD is pulmonary function testing (PFT), which primarily quantifies expiratory airflow limitation. Based on the Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines, COPD is defined as the ratio of forced expiratory volume in 1 second to forced vital capacity <0.70 (or the lower limit of normal for individuals of the same age, sex, and height), measured prior to and following bronchodilator use [6]. However, it is challenging to implement PFT. It requires specialized spirometry equipment and trained personnel, and participants must repeatedly perform forceful exhalation maneuvers. Older adults or severely ill patients often produce false-negative results due to insufficient effort. In addition, the procedure may induce coughing, dizziness, or other discomforts and poses a risk of cross-infection under pandemic conditions or in poorly controlled environments. These factors limit the application of PFT in community and primary care settings [7]. Thus, relying solely on conventional PFT is insufficient for screening COPD. Developing simpler, non-invasive, and more scalable auxiliary diagnostic methods for early detection of COPD is, therefore, imperative.

In recent years, deep learning (DL) has attracted significant attention in clinical practice. DL is a complex neural network framework. Common DL models include convolutional neural networks, residual networks, densely connected networks, inception networks, and vision transformer models [8]. These models excel at feature extraction and classification, allowing the automatic learning of high-level semantic information from large datasets, thereby markedly improving the precision and efficiency of image processing and signal analysis [9]. Although PFT is recognized as the gold standard for the auxiliary diagnosis of COPD, researchers often employ chest imaging (including computed tomography [CT] scans and X-rays) or respiratory sounds to develop DL-based alternative or complementary tools for improving diagnostic efficiency and convenience. However, these traditional methods heavily rely on researchers' prior knowledge, and variations in diagnostic criteria and annotation practices across different teams result in significant heterogeneity, affecting the reproducibility and generalizability of diagnostic outcomes [10,11]. In this context, some studies have used DL for the automatic diagnosis of COPD, such as DL-based chest X-ray (CXR), for the classification of COPD [10] and DL-based cough sound signal analysis [11]. Nevertheless, systematic evidence of the actual performance and comparative advantages of different DL frameworks in the diagnosis of COPD is lacking.

Therefore, we conducted a systematic review and meta-analysis of diagnostic test accuracy studies on DL models for COPD. Our first objective was to describe the diagnostic performance of these models for identifying COPD across different data sources (such as CT images and respiratory sounds) in both internal and external validation sets. Our second objective was to assess the performance of DL models in classifying the severity of COPD, particularly GOLD stages. We hypothesized that DL models would show good accuracy for the diagnosis of COPD, whereas their performance for staging COPD would be more variable and less stable.

Methods

Study Registration

This systematic review and diagnostic test accuracy meta-analysis was conducted and reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement and the PRISMA-DTA (Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies) extension, and the search methods were reported following PRISMA-S (Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension) [12,13]. The PRISMA-S checklist is provided in Checklist 1. The protocol was prospectively registered in PROSPERO (International Prospective Register of Systematic Reviews; CRD420251114195 [14]).

Eligibility Criteria

The inclusion criteria were as follows: (1) original research that developed a DL model for diagnosing COPD or classifying COPD severity; (2) studies reported at least one of the following outcome measures for appraising the accuracy of DL model: concordance index, the receiver operating characteristic curve, specificity, sensitivity, precision rate, accuracy, recall rate, calibration curve, F_1-score, or confusion matrix; and (3) studies published in English.

Exclusion criteria were as follows: (1) conference abstracts without full-text publication; (2) studies limited to traditional machine learning, without the development of DL models; and (3) studies applying DL solely for image segmentation, without developing models for the diagnosis or classification of COPD. Although a very small number of the included studies may have used data from the same public database, we still included these studies because their DL models incorporated comparable

XSL•FO
RenderX

experimental designs, which helped us better understand the diagnostic performance of DL models for COPD.

Data Sources and Search Strategy

The search methods and reporting were guided by PRISMA-S [12]. Embase, Web of Science, the Cochrane Library, and PubMed were systematically searched from database inception to November 1, 2025. The search strategy was designed by combining medical subject headings and free-text keywords. To maximize the retrieval of relevant studies, no restrictions were applied on language or geographic location. The complete search strategies are provided in Table S1 in Multimedia Appendix 1.

We screened the reference lists of the included studies and relevant reviews; we did not search gray literature or conference proceedings and did not contact authors for additional data. No published search filters were used. Search strategies were developed de novo and were not adapted or reused from prior reviews. We did not conduct a formal peer review of the search strategy.

Study Selection

The retrieved studies were imported into EndNote. Duplicates were automatically and manually removed. Subsequently, the titles and abstracts of the remaining articles were independently reviewed by 2 authors (YH and YW) to identify potentially eligible studies. The full texts of these studies were then assessed to identify eligible studies. Any disagreements at any stage were resolved through discussion with a third reviewer (TW).

Data Extraction

Before data extraction, a standardized extraction form was developed. The collected data encompassed study title, publication year, DOI, country, authors, patient source, study design, task type, COPD diagnostic criteria, imaging modality used for modeling, number of COPD cases, total number of cases, number of COPD cases in the training set, total number of cases in the training set, method for validation set generation, external validation, number of COPD cases in the validation set, total number of cases in the validation set, and comparison with clinicians (yes or no). Two reviewers (HY and TW) independently extracted the data, followed by cross-checking. Any disagreements were addressed through consultation with a third reviewer (YW).

Risk of Bias in Studies

The QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2) tool was utilized to appraise the risk of bias (RoB) of the selected studies. The assessment covered 4 domains: reference standard, index test, patient selection, as well as flow and timing. Each domain included several specific questions, which were answered by “Yes (low RoB),” “No (high RoB),” or “Unclear (RoB uncertain).” The overall RoB for each domain was categorized as low, high, or unclear. The RoB assessment was independently performed by 2 reviewers (YW and TW), and disagreements were addressed through discussion with a third reviewer (HY).

Synthesis Methods

For binary classification tasks, a bivariate mixed effects model was used to pool diagnostic 2×2 contingency tables for DL for the diagnosis of COPD. In studies without complete contingency tables, specificity, sensitivity, negative and positive predictive values, accuracy, and the number of cases were used to estimate the contingency table. Sensitivity, specificity, negative likelihood ratio (NLR), positive likelihood ratio (PLR), diagnostic odds ratio (DOR), and the summary receiver operating characteristic curve with corresponding 95% CIs were pooled. Deeks’ funnel plot was applied to examine the small-study effects of the selected original studies, and clinical applicability was assessed through nomograms. Subgroup analyses by modality (CT, respiratory sounds, or CXR) were performed. All meta-analyses were based on validation set data. If a study reported multiple validation cohorts, each independent validation cohort was included in the analysis separately. If multiple models were evaluated on the same validation cohort, only 1 estimate (ie, the primary and final model reported) was extracted to avoid the nonindependence of the data.

For multiclass classification tasks, the accuracy across different severity grades was pooled. When the reported accuracy approached 99%, a double arcsine transformation was applied before meta-analysis. During the meta-analysis, we utilized the Hartung-Knapp-Sidik-Jonkman modified method [15]. Due to the potential heterogeneity, the 95% prediction intervals for the summary estimates were calculated using the confidence distribution approach proposed by Nagashima et al [16]. All analyses were carried out using STATA (version 15.0; StataCorp LLC) or R (version 4.4.3; R Foundation for Statistical Computing).

Results

Study Selection

Overall, 5194 records were retrieved from databases. After excluding 1958 duplicates, we removed 1695 studies unrelated to the study topic and 492 studies for other reasons. The titles and abstracts of 1049 studies were checked. Among them, 969 studies were removed due to irrelevant or unsuitable study design. The full texts of 80 articles were assessed for eligibility, among which 24 ineligible studies were further excluded. Ultimately, 56 studies [10,17-71] were included (Figure 1).

XSL•FO
RenderX

Figure 1. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart for the systematic review and diagnostic test accuracy meta-analysis, showing the entire process from the database search in Embase, Web of Science, Cochrane Library, and PubMed up to November 1, 2025, to the literature screening and final inclusion of 56 studies on chronic obstructive pulmonary disease (COPD) in adults.

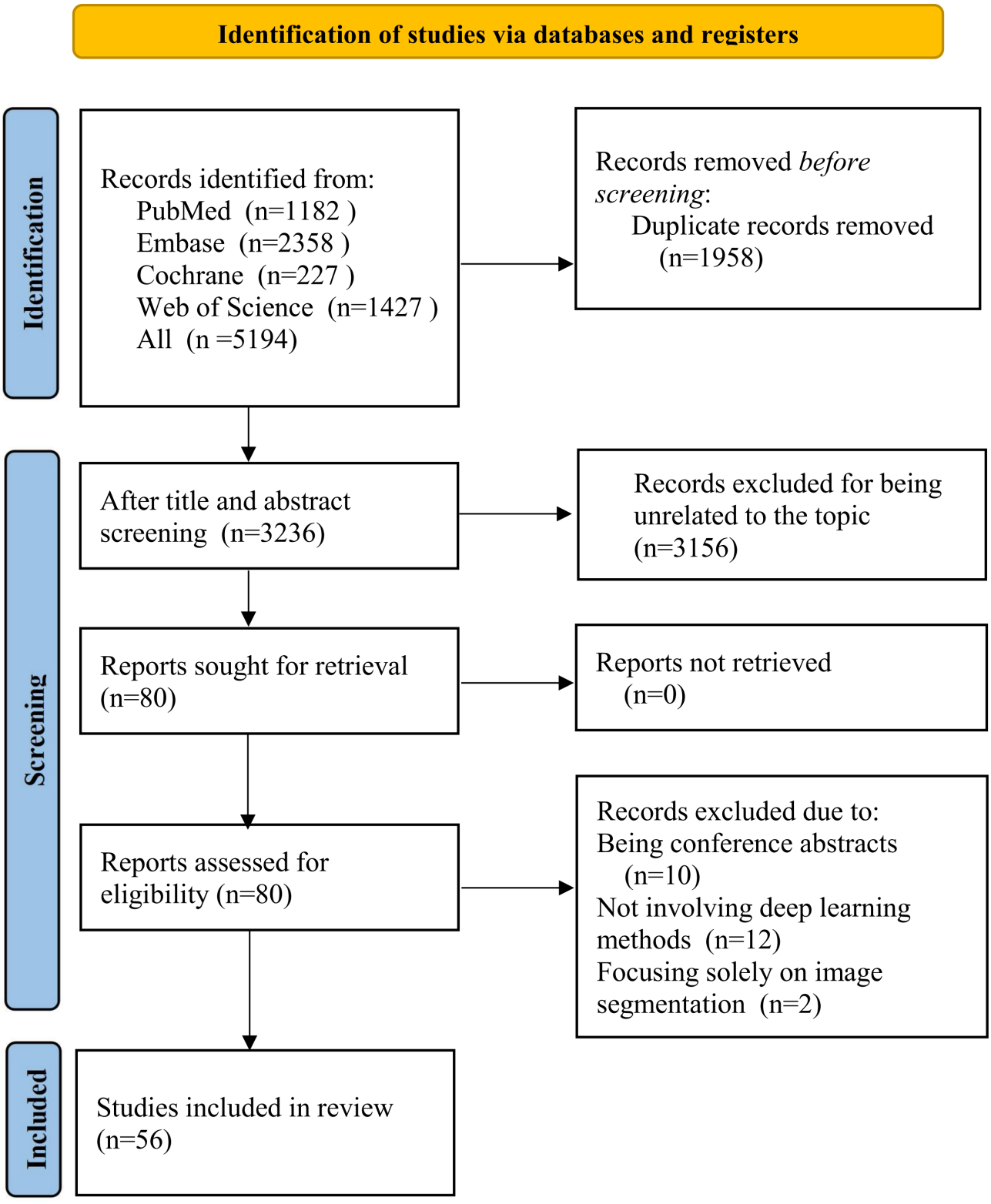

Study Characteristics

The 56 selected studies were published between 2019 and 2025 across 14 countries, with the majority conducted in China (n=21) and the United States (n=11). In terms of study design, there were 39 cohort studies (including retrospective cohort studies), 16 case-control studies, and 1 retrospective cross-sectional diagnostic study. Most datasets were derived from single-center (n=16) or multi-center (n=31) studies, while 9 studies utilized registry databases. Regarding task types, 23 studies focused solely on diagnosis, 17 studies solely on classification, and 16 studies on both diagnosis and classification (out of 16). All studies clearly reported the diagnostic criteria for COPD. The variables of the models primarily came from CT images (30

XSL•FO
RenderX

studies) and breath sound or audio data (12 studies); 4 studies used CXRs (including X-ray films in 2 studies); the remaining 10 studies used other input data (eg, pulmonary function indicators or curves or waveforms, electrocardiograms, volumetric carbon dioxide monitoring, clinical data, imaging-genetic data, or CT-based scores). The total number of cases was 886,753, with 272,881 in the validation sets and 1,352,782 in the training sets. The methods for generating the validation set were categorized as follows: only cross-validation used in 22 studies; only internal validation in 20 studies; external validation in 9 studies; a combination of internal and external validation in 3 studies; and a combination of cross-validation, internal validation, and external validation in 1 study. One study did not report its validation strategy (1 study; Table S2 in Multimedia Appendix 1).

RoB in Studies

In the patient selection domain, all studies employed consecutive or random case selection and applied appropriate exclusion criteria, thereby avoiding including inappropriate cases; therefore, RoB was judged to be low in this domain. For the index test domain, the included studies generally applied supervised DL methods with clearly defined decision rules, and RoB was judged to be mostly low. Regarding the reference standard, all studies used appropriate diagnostic criteria capable of effectively distinguishing COPD and its severity; however, if a study did not explicitly report whether the reference standard assessment was performed blinded to the index test, we rated this item as unclear, leading to an overall judgment of unclear RoB in the reference standard domain for those studies. For the flow and timing domain, RoB was generally low, although incomplete reporting of participant flow and timing resulted in some unclear judgments. In terms of applicability, patient selection was largely consistent with the review question, while a subset of studies raised applicability concerns related to the index test, the reference standard, or both. In addition, some studies reported only summary performance metrics (eg, accuracy) without complete 2×2 contingency tables, which limited transparency for evidence synthesis and introduced uncertainty when reconstructing contingency tables (Figures 2 and 3).

XSL•FO
RenderX

Figure 2. Detailed QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2) risk-of-bias assessment process for the included 56 diagnostic accuracy studies on deep learning (DL) models for chronic obstructive pulmonary disease (COPD) [10,17-19,21,23,24,26-29,31-35,37-41,43-61,63-70,72].

XSL•FO
RenderX

Figure 3. Summary of QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2) risk-of-bias and applicability assessments for the 56 diagnostic accuracy studies on deep learning (DL) models for chronic obstructive pulmonary disease (COPD).

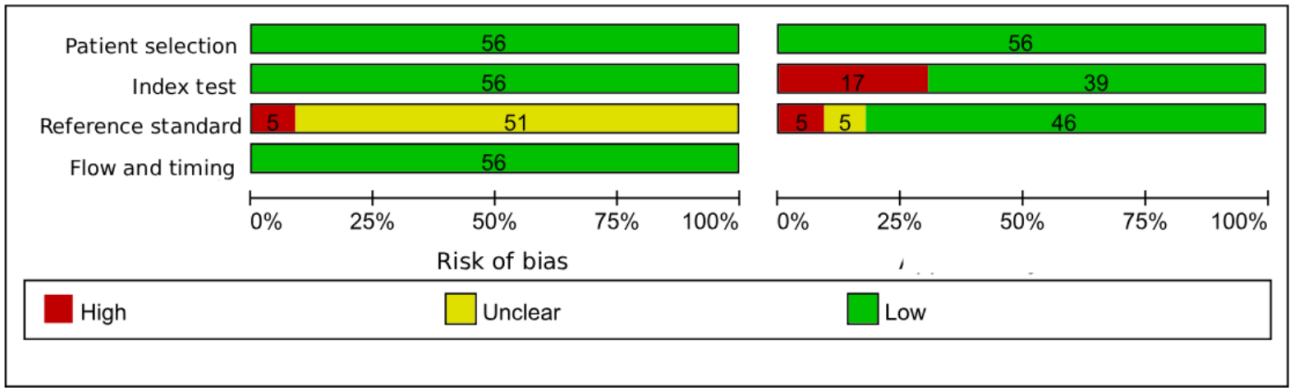

Meta-Analysis of Binary Classification Tasks

Overall

A total of 43 diagnostic 2×2 contingency tables were synthesized to appraise the diagnostic accuracy of DL models for COPD. The pooled results demonstrated that the DL models yielded a sensitivity of 0.87 (95% CI: 0.85 - 0.90), specificity of 0.88 (95% CI 0.84 - 0.92), PLR of 7.4 (95% CI 5.2 - 10.5), NLR of 0.14 (95% CI 0.11 - 0.18), DOR of 52 (95% CI 30 - 88), and the area under the summary receiver operating characteristic curve (AUC) of 0.93 (95% CI 0.18 - 1.00; Figures 4 and 5). Deeks' funnel plot demonstrated no significant small-study effects (P=.08; Figure 6). Assuming a pretest probability of 25%, the posttest probability rose to about 71% for a positive result and decreased to about 5% for a negative result, suggesting the potential clinical value of the models in the screening and diagnosis of COPD (Figure 7).

Figure 4. Forest plots of sensitivity and specificity of deep learning (DL) model for binary classification diagnosis of chronic obstructive pulmonary disease (COPD), summarizing 2×2 contingency table results from 43 validation cohorts in 14 countries from 2019 to 2025 [10,11,17,18,20,22,23,25-34,38,40,41,43-45,48-51,53,54,56,61,63,67,68,70,73].

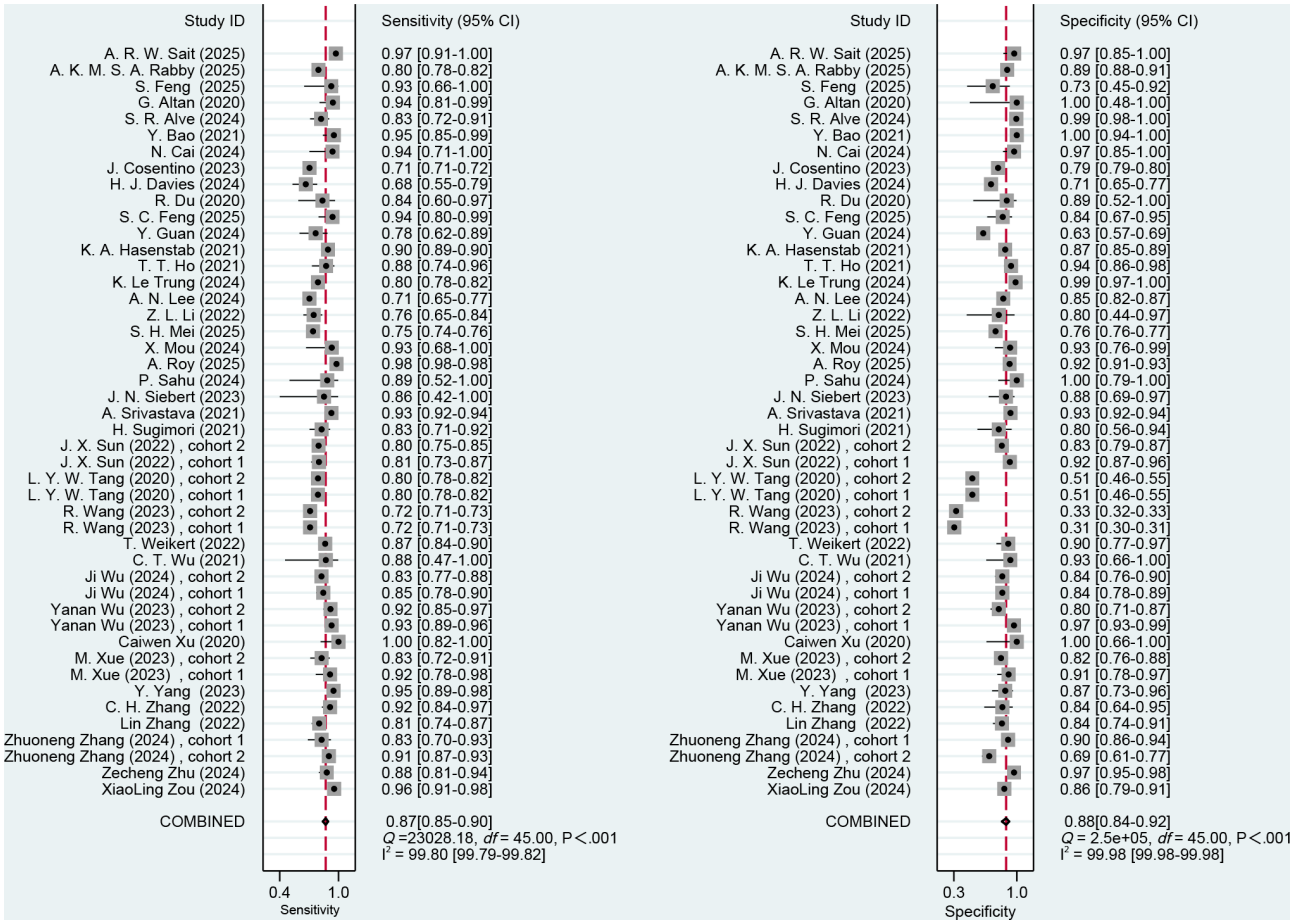

XSL•FO
RenderX

Figure 5. Summary receiver operating characteristic (SROC) curve for the meta-analysis of deep learning (DL) for the diagnosis of chronic obstructive pulmonary disease (COPD) in the validation sets. AUC: area under the summary receiver operating characteristic curve; SENS: sensitivity; SPEC: specificity.

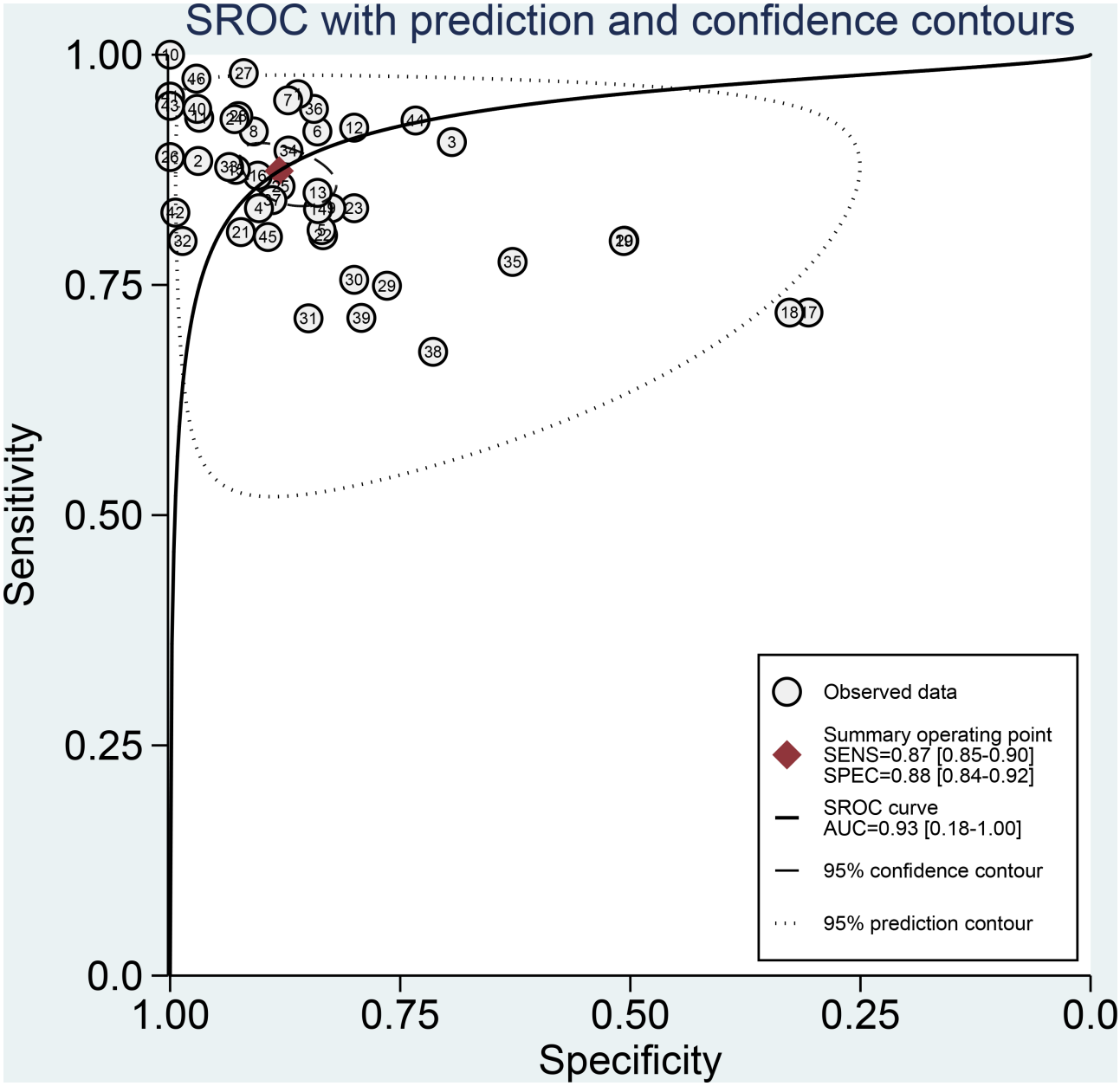

XSL•FO
RenderX

Figure 6. Deeks funnel plot of deep learning (DL) model in the binary classification diagnosis of chronic obstructive pulmonary disease (COPD), assessing publication bias and small-sample effect based on 43 validation cohorts. ESS: effective sample size.

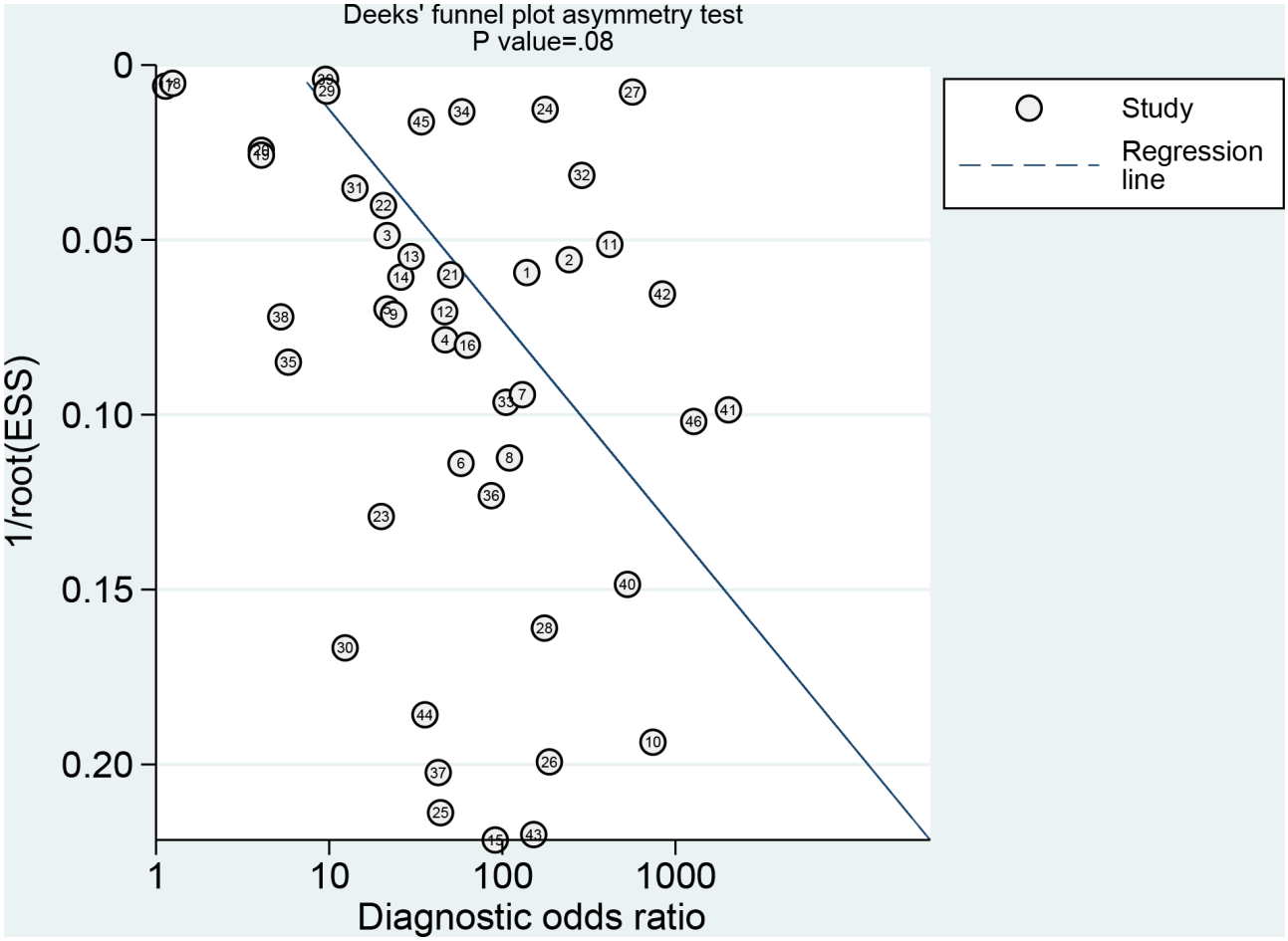

XSL•FO
RenderX

Figure 7. The Fagan cursor plot for deep learning (DL) model in the binary classification diagnosis of chronic obstructive pulmonary disease (COPD), providing the posterior probabilities of positive and negative results with a 25% predetection probability in individuals with suspected COPD based on the pooled likelihood ratio (43 validation cohorts).

XSL•FO
RenderX

DL Based on CT Images

A total of 30 contingency tables were included. The pooled sensitivity of the models was 0.86 (95% CI 0.84 - 0.89), specificity was 0.87 (95% CI 0.82 - 0.90), PLR was 6.6 (95% CI 4.8 - 9.1), NLR was 0.15 (95% CI 0.12 - 0.19), DOR was 42 (95% CI 26 - 68), and AUC was 0.92 (95% CI 0.90 - 0.94; Figures S1-S2 in Multimedia Appendix 1). Deeks' test indicated potential small-study effects (*P*=.02; Figure S3 in Multimedia Appendix 1). Assuming a pretest probability of 25%, the posttest probability rose to 69% for a positive result and decreased to 5% for a negative result (Figure S4 in Multimedia Appendix 1).

Among these, 24 contingency tables were derived from internal validation sets. The pooled sensitivity was 0.86 (95% CI 0.83 - 0.89), specificity was 0.88 (95% CI 0.83 - 0.92), PLR was 7.4 (95% CI 5.0 - 10.9), NLR was 0.16 (95% CI 0.12 - 0.20), DOR was 48 (95% CI 26 - 86), and AUC was 0.93 (95% CI 0.90 - 0.95; Figures S5–S6 in Multimedia Appendix 1). Deeks' funnel plot demonstrated small-study effects (*P*=.04; Figure S7 in Multimedia Appendix 1). Assuming a pretest probability of 25%, the posttest probability increased to about 71% for a positive result and decreased to about 5% for a negative result (Figure S8 in Multimedia Appendix 1). Among studies on CT-based DL, most models incorporated lung parenchymal attenuation patterns related to emphysema. Some additionally incorporated airway and bronchial wall morphology; gas trapping and small-airway abnormalities on inspiratory or expiratory CT; or combined radiomics features of lung parenchyma, airways, and pulmonary vessels.

A total of 8 contingency tables originated from external validation sets. The pooled sensitivity was 0.87 (95% CI 0.82 - 0.90), specificity was 0.83 (95% CI 0.72 - 0.90), PLR was 5.1 (95% CI 2.9 - 8.8), NLR was 0.16 (95% CI 0.11 - 0.23), DOR was 31 (95% CI 14 - 72), and AUC was 0.91 (95% CI 1.00 - 0.00; Figures S9-S10 in Multimedia Appendix 1). Deeks' funnel plot demonstrated no small-study effects (*P*=.06; Figure S11 in Multimedia Appendix 1). Assuming a pretest probability of 25%, the posttest probability rose to about 63% for a positive result and decreased to about 5% for a negative result (Figure S12 in Multimedia Appendix 1).

To further evaluate potential small-study effects, we additionally stratified the CT-based validation cohorts by the number of COPD cases in the validation set. A total of 15 cohorts were classified as a small-sample subgroup (COPD cases <100) and 17 cohorts as a large-sample subgroup (COPD cases ≥100). In the small-sample subgroup, the pooled sensitivity and specificity were 0.89 (95% CI 0.84 - 0.92) and 0.89 (95% CI 0.83 - 0.94), respectively, with an AUC of 0.94 (95% CI 0.92 - 0.96). In the large-sample subgroup, the pooled sensitivity and specificity were slightly lower at 0.85 (95% CI 0.82 - 0.88) and 0.85 (95% CI 0.78 - 0.90), respectively, with an AUC of 0.91 (95% CI 0.88 - 0.93; Figures S13-S16 in Multimedia Appendix 1). Assuming a pretest probability of 25%, the Fagan nomograms indicated that the posttest probability increased to 74% for a positive DL result in the small-sample studies and 65% in the large-sample studies, while it reduced to 4% and 6% for a negative result, respectively (Figures S17 and S18 in Multimedia Appendix 1). Deeks' funnel plot asymmetry tests for the small- and large-sample subgroups were not statistically significant (*P*=.34 and *P*=.15, respectively; Figures S19 and S20 in Multimedia Appendix 1), suggesting no strong evidence of small-study effects. However, given the consistently higher point estimates in the small-sample subgroup, some degree of small-study effects cannot be completely ruled out.

DL Based on Respiratory Sounds

A total of 10 contingency tables were included. The pooled sensitivity was 0.91 (95% CI 0.84 - 0.95), specificity was 0.96 (95% CI 0.91 - 0.98), PLR was 22.1 (95% CI 9.5 - 51.5), NLR was 0.09 (95% CI 0.05 - 0.18), DOR was 237 (95% CI 78 - 723), and AUC was 0.98 (95% CI 0.96 - 0.99; Figures S21-S22 in Multimedia Appendix 1). Deeks' funnel plot demonstrated no small-study effects (*P*=.32; Figure S23 in Multimedia Appendix 1). With a pretest probability of 25%, the posttest probability rose to about 88% following a positive result and decreased to about 3% following a negative result (Figure S24 in Multimedia Appendix 1). For respiratory sound–based DL models, lung sounds were recorded using electronic or digital stethoscopes at standard chest auscultation sites or obtained from open respiratory sound databases (eg, RespiratoryDatabase@TR and other multichannel lung sound datasets) and analyzed as single- or multichannel signals.

Summary of DL Based on CXR

Only 2 included studies evaluated DL models based on CXR for the diagnosis of COPD. In a multicenter study, Zou et al [10] constructed a DL model integrating CXR images and clinical parameters. This model achieved favorable performance in internal validation with a sensitivity of 0.96 and a specificity of 0.86. Conversely, Wang et al [29] constructed a model solely based on CXR images. Their model yielded a sensitivity of 0.72 and specificity of 0.31 in the MIMIC-CXR internal validation set and a sensitivity of 0.72 and specificity of 0.33 in the Emory-CXR external validation set. These findings suggest that combining clinical parameters with imaging data may substantially enhance diagnostic performance, whereas single-image models exhibit limited specificity.

Summary of DL Based on Externally Applied Airway Resistance

In the study by Davies [54], a physical simulation device was utilized to generate surrogate data for training a DL model. Tubes of varying diameters (3 - 25 mm) were installed in the respiratory tract of healthy participants to independently modulate inspiratory and expiratory resistance, thereby simulating COPD-related obstruction. Based on the generated photoplethysmography signals, a 1D convolutional neural network achieved an AUC of 0.75 in the binary classification of COPD and healthy controls. The accuracy of the model reached 40% - 88% for real COPD cases, with a 14% misdiagnosis rate in healthy participants. This approach may offer a low-cost alternative for data-scarce scenarios, particularly suitable for screening with wearable devices in primary care. However, since dynamic resistance simulation was limited, and the sample size for validation was small (only 4 patients), the model needs to be further optimized.

XSL•FO
RenderX

Multiclass DL for COPD Grading

A total of 6 studies [10,22,31,32,40,44] developed DL models for GOLD grading of COPD (multiclass classification). Among these studies, 5 developed models based on CT images, while Zou et al [10] used CXR images for modeling. Most studies applied different GOLD classification strategies. Several studies [10,31,40,44] implemented 5-class classification (GOLD 0 - 4). In another analysis by Zou [10], a 3-class strategy was applied (GOLD 0, GOLD 1 - 2, GOLD 3 - 4). Sugimori [32] and Yang [22] employed a 4-class strategy (GOLD 0, 1, 2, 3 - 4).

Overall analysis indicated considerable differences in the accuracy of the DL models for identifying each GOLD stage, reflecting substantial heterogeneity in model performance. The pooled results based on a random-effects model were as follows: the diagnostic accuracy was 0.842 (95% CI 0.605 - 0.982) for GOLD 0, 0.617 (95% CI 0.407 - 0.808) for GOLD 1, 0.679 (95% CI 0.376 - 0.917) for GOLD 2, 0.708 (95% CI 0.163 - 1.000) for GOLD 3, and 0.708 (95% CI 0.163 - 1.000) for GOLD 4 (Figure S25 in Multimedia Appendix 1). These findings demonstrated that the DL models were unstable in the identification of mild (GOLD 1) and very severe (GOLD 4) stages. Given the wide CIs, the diagnostic accuracy was still limited.

Discussion

Summary of the Main Findings

Current DL models for detecting COPD are primarily constructed based on CT imaging and respiratory sound data. The tasks are generally divided into binary and multiclass classifications. Our findings suggested that in binary classification tasks, the CT-based models performed well in internal validation cohorts, with a pooled sensitivity of 0.86 (95% CI 0.83 - 0.89) and specificity of 0.88 (95% CI 0.83 - 0.92). The models based on respiratory sounds yielded a sensitivity of 0.91 (95% CI 0.84 - 0.95) and specificity of 0.96 (95% CI 0.91 - 0.98), indicating a strong exclusion ability.

In multiclass classification tasks, the included studies mainly focused on the staging of GOLD. Overall analysis demonstrated that the DL models were unstable for discriminating between different GOLD stages. This finding supports our hypothesis that compared with binary diagnosis, the accuracy and reliability of the DL models for staging COPD still need to be improved.

Comparison With Previous Reviews

Prior studies have examined the application of CT and respiratory sounds in the diagnosis of COPD. The systematic review and network meta-analysis carried out by Balasubramanian et al [74] focuses on the diagnostic performance of CT-guided transthoracic biopsy or fine-needle aspiration in lung diseases, particularly lung cancer. Their study included 363 studies involving 79,519 patients and reported a pooled sensitivity of 88.9% but did not address the use of CT in the diagnosis of COPD. In addition, Arts et al [75] have evaluated the use of respiratory sounds for diagnosing acute pulmonary diseases. Their results demonstrate that respiratory sounds have a sensitivity of 37% (95% CI 30% - 47%) and specificity of 89% (95% CI 85% - 92%) for diagnosing COPD, based on approximately 12 relevant studies [75]. Willer et al [73] have examined the performance of X-ray dark-field imaging in detecting and evaluating emphysema in patients with COPD. Their study includes 77 patients and reports that this imaging modality exhibits high diagnostic performance for emphysema (correlation coefficient ρ=0.62, P<.0001) and is closely associated with microstructural changes in the lung. These findings suggest that dark-field chest imaging may be a rapid, low-dose, and sensitive tool for the screening and assessment of COPD. However, their study does not evaluate the diagnostic accuracy of conventional CXR for COPD.

In contrast, this meta-analysis reported higher diagnostic performance of the DL models based on CT imaging and respiratory sounds. The pooled results demonstrated that the DL models based on CT yielded a sensitivity of 0.86 (95% CI 0.84 - 0.89) and specificity of 0.87 (95% CI 0.82 - 0.90), while respiratory sound–based models yielded a sensitivity of 0.91 (95% CI 0.84 - 0.95) and specificity of 0.96 (95% CI 0.91 - 0.98). These results suggest that DL approaches might outperform traditional diagnostic methods. Earlier research has also investigated the role of artificial intelligence (AI) in COPD diagnosis. For instance, Wu et al [72] examined the potential of machine learning and DL in the detection, staging, and quantitative analysis of COPD using CT imaging. However, their review does not clearly differentiate between machine learning and DL, nor does it discuss in depth the advantages and limitations of image-based AI models for the diagnosis of COPD.

This study found that the included studies on DL for diagnosing COPD focused mainly on CT imaging, respiratory sounds, CXR, and externally applied airway resistance. Among these, CT, respiratory sounds, and CXR were the most frequently used data sources for model development and carried distinct clinical implications. Chest CT exerts a crucial role in diagnosing and phenotyping COPD, as it can identify structural abnormalities, such as airway narrowing and emphysema, and is recommended by current clinical guidelines. Our findings demonstrated that the CT-based DL models offered excellent specificity and sensitivity for the diagnosis of COPD, suggesting their potential as auxiliary diagnostic tools in clinical practice. The DL models based on respiratory sounds, as a non-invasive and portable modality, also had good diagnostic performance, particularly with high specificity, indicating potential value in primary screening. In contrast, the number of studies using CXR remains limited, and the existing evidence is insufficient to determine the stability and generalizability of CXR-based DL models for diagnosing COPD. It should be validated in the future. Moreover, although a few preliminary studies have explored the use of externally applied airway resistance to generate model inputs, the number of studies remains small, and reproducible, generalizable evidence is lacking. Thus, future studies are required to assess the utility and reliability of this approach in clinical practice.

Despite the promise of AI in the diagnosis of COPD, significant challenges need to be addressed before widespread clinical application, particularly in explainability and data integration. Although current research demonstrates encouraging diagnostic performance, a substantial gap persists between theoretical

XSL•FO
RenderX

development and real-world application. First, most included studies did not thoroughly examine how variations in imaging protocols, such as scan parameters or reconstruction algorithms, influence image features and the performance of DL models. Hence, a systematic evaluation of these factors is lacking. Second, as complex neural network frameworks, DL models rely on large-scale training datasets to improve robustness. However, most included studies developed models using limited samples, with only a few utilizing large datasets. The scarcity of data represents a core bottleneck in model development, constraining the generalizability of the models. Future studies should incorporate richer and more diverse imaging data. Third, the current model evaluation primarily relied on internal validation techniques, such as random sampling, cross-validation, or bootstrap methods. While internal validation sets share similar distributions with training data and often yield favorable results, they do not accurately reflect the generalizability of the models on heterogeneous datasets. Models should be rigorously externally validated before real-world application, particularly across institutions and using datasets obtained under different imaging protocols. Studies based on high-quality external validation remain scarce, and substantial differences in imaging protocols make it challenging to interpret model performance in external validation.

In clinical research and practice, grading disease severity is as crucial as diagnosing COPD. The widely applied GOLD classification, which stratifies COPD into 5 grades (0 and 1‐4), reflects significant differences in clinical presentation, treatment strategies, and prognosis of COPD. Achieving early and precise grading is therefore of high clinical relevance. However, only 6 studies have attempted to develop DL models for grading the severity of COPD, providing limited evidence. These studies indicate that DL models generally perform suboptimally in multiclass classification tasks, with particularly low accuracy for GOLD 1, GOLD 2, and GOLD 4. These models achieve relatively higher accuracy only for GOLD 0 and GOLD 3, exceeding 70%. Nevertheless, their stability still needs to be enhanced. This suggests that multiclass classification itself represents a technical challenge for DL models. Moreover, under the current dataset size, label distribution, and model architecture, stable differentiation across all GOLD grades remains difficult. Future research should aim to enhance the discriminative ability of models, incorporate richer imaging data, and integrate clinical information to optimize training strategies, ultimately developing more accurate and adaptable intelligent tools for grading the severity of COPD to support clinical decision-making.

Strengths and Limitations of the Study

This meta-analysis systematically assessed the performance of DL in the detection of COPD for the first time, providing evidence to support the development of intelligent diagnostic tools. The findings indicate that DL models hold substantial potential for improving diagnostic accuracy, particularly through noninvasive and nonintrusive detection methods. This study provides valuable insights. However, some limitations must be noted. First, although a systematic literature search was carried out, the number of studies focusing on respiratory sounds remained relatively small. As respiratory sound analysis is an emerging diagnostic approach, the number and diversity of relevant studies remain far below those of CT imaging, which may limit a comprehensive assessment of this method. Second, most included studies relied primarily on internal validation, and only relatively few studies performed external validation. Although internal validation can provide some indication of diagnostic accuracy, limitations in sample size and validation methods may compromise the generalizability of the results. To further confirm the clinical utility of DL models, future studies should perform external validation. Third, research on the severity of COPD was relatively scarce, and some studies employed differing grading strategies. These variations may affect the reliability of classification models and the generalizability of their findings. Thus, this finding should be cautiously interpreted.

Heterogeneity and Clinical Applicability of DL Models

Although subgroup analyses were performed to explore the source of heterogeneity, significant heterogeneity still existed among the subgroups. This heterogeneity may stem from differences in DL frameworks used in different studies, such as 2D or 3D convolutional neural networks, multiview networks, multi-instance learning, and late fusion. The included studies used diverse DL models, which differed in network structure, input format, and parameter settings. Consequently, their model training and validation methods may also differ. Therefore, these differences in structure and parameters can lead to potential heterogeneity, which is a common challenge in current meta-analyses of DL models.

From the perspective of clinical practicality, DL still holds significant advantages over traditional radiomics. Traditional radiomics typically requires manual or semiautomatic image segmentation, followed by the extraction of a limited number of manual features, such as texture. An original image is compressed into a small number of quantitative features, then to a machine learning model. This multistep process is time-consuming, highly dependent on the researcher's experience, and may lose some original image information during dimensionality reduction and feature selection. DL, on the other hand, can directly train models end-to-end based on labeled (or segmented) images without additional feature engineering. It can preserve lesion-related image information to the greatest extent, potentially improving model performance and reducing manual operations and time costs. Therefore, given the relatively ideal diagnostic and grading accuracy of DL models, it is hoped that AI-assisted diagnostic DL tools should be developed to support, rather than replace, clinicians in screening and assessing the severity of COPD.

Future Perspectives

Most current studies are based on relatively limited imaging datasets and rely mainly on internal validation. Thus, the reported accuracy may not fully reflect the generalizability of models. Given substantial between-study heterogeneity and limited external validation, these findings should be interpreted cautiously. Future research should improve and update these DL models by using larger, multicenter imaging datasets from different geographical regions and scanners, and by

XSL•FO
RenderX

incorporating robust external validation and more rigorous model development strategies.

To our knowledge, this is the first systematic synthesis to quantify the diagnostic and grading performance of DL models across major data sources (eg, CT imaging and respiratory sounds), showing promising accuracy for binary COPD detection but suboptimal and less stable performance for multiclass GOLD staging.

In summary, our comprehensive study on DL provides an evidence base for guiding the development and external validation of AI-assisted screening tools for COPD, especially given the insufficient application of spirometry.

Conclusions

This study observed that DL models achieved promising accuracy in the detection of COPD. The models performed particularly well in binary classification tasks, exhibiting high sensitivity and specificity. However, its accuracy was suboptimal in multiclass tasks for grading the severity of GOLD. In addition, research on respiratory sound analysis and multiclass classification of COPD severity is still limited. Given the substantial heterogeneity and limited external validation, these results should be interpreted cautiously. Thus, future research should integrate larger and more diverse imaging datasets, particularly including images from different racial populations, to develop more robust and generalizable intelligent diagnostic tools. This approach would not only enhance the generalizability of models but also improve the accuracy of diagnosing COPD across diverse patient groups.

Funding

This research was not supported by any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: HY
Data curation: YW
Methodology: HY
Software: HY, JJ, SL
Supervision: TW
Validation: JJ, SL
Writing – original draft preparation: HY, YW
Writing – reviewing and editing: WX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials (search strategies, study characteristics, and additional analyses).
[DOCX File, 36906 KB - jmir_v28i1e83459_app1.docx]

Checklist 1
PRISMA-S checklist.
[PDF File, 66 KB - jmir_v28i1e83459_app2.pdf]

References

1. Agustí A, Celli BR, Criner GJ, et al. Global Initiative for Chronic Obstructive Lung Disease 2023 Report: GOLD executive summary. Eur Respir J 2023 Apr;61(4):2300239. [doi: 10.1183/13993003.00239-2023] [Medline: 36858443]
2. Al Wachami N, Guennouni M, Iderdar Y, et al. Estimating the global prevalence of chronic obstructive pulmonary disease (COPD): a systematic review and meta-analysis. BMC Public Health 2024 Jan 25;24(1):297. [doi: 10.1186/s12889-024-17686-9] [Medline: 38273271]
3. Xu J, Ji Z, Zhang P, Chen T, Xie Y, Li J. Disease burden of COPD in the Chinese population: a systematic review. Ther Adv Respir Dis 2023;17:17534666231218899. [doi: 10.1177/17534666231218899] [Medline: 38146618]
4. Agustí A, Celli BR, Criner GJ, et al. Global Initiative for Chronic Obstructive Lung Disease 2023 Report: GOLD executive summary. Am J Respir Crit Care Med 2023 Apr 1;207(7):819-837. [doi: 10.1164/rccm.202301-0106PP] [Medline: 36856433]

XSL•FO
RenderX

5. Takano T, Tsubouchi K, Hamada N, et al. Update of prognosis and characteristics of chronic obstructive pulmonary disease in a real-world setting: a 5-year follow-up analysis of a multi-institutional registry. BMC Pulm Med 2024 Nov 6;24(1):556. [doi: 10.1186/s12890-024-03347-5] [Medline: 39506773]
6. Singh D, Stockley R, Anzueto A, et al. GOLD Science Committee recommendations for the use of pre- and post-bronchodilator spirometry for the diagnosis of COPD. Eur Respir J 2025 Feb;65(2):2401603. [doi: 10.1183/13993003.01603-2024] [Medline: 39638416]
7. Baldomero AK, Kunisaki KM, Bangerter A, et al. Beyond access: factors associated with spirometry underutilization among patients with a diagnosis of COPD in urban tertiary care centers. Chronic Obstr Pulm Dis 2022 Oct 26;9(4):538-548. [doi: 10.15326/jcopdf.2022.0303] [Medline: 36040836]
8. Han K, Wang Y, Chen H, et al. A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell 2023 Jan;45(1):87-110. [doi: 10.1109/TPAMI.2022.3152247] [Medline: 35180075]
9. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017 Dec;42:60-88. [doi: 10.1016/j.media.2017.07.005] [Medline: 28778026]
10. Zou X, Ren Y, Yang H, et al. Screening and staging of chronic obstructive pulmonary disease with deep learning based on chest X-ray images and clinical parameters. BMC Pulm Med 2024 Mar 26;24(1):153. [doi: 10.1186/s12890-024-02945-7] [Medline: 38532368]
11. Zhang P, Swaminathan A, Uddin AA. Pulmonary disease detection and classification in patient respiratory audio files using long short-term memory neural networks. Front Med (Lausanne) 2023;10:1269784. [doi: 10.3389/fmed.2023.1269784] [Medline: 38020156]
12. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
13. McInnes MDF, Moher D, Thombs BD, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA 2018 Jan 23;319(4):388-396. [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
14. Yang H, Wu Y, Wu T. Accuracy of deep learning in diagnosing COPD: a systematic review and meta-analysis. Centre for Reviews and Dissemination. URL: https://www.crd.york.ac.uk/PROSPERO/view/CRD420251114195 [accessed 2025-12-25]
15. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
16. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019 Jun;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
17. Zhu Z, Zhao S, Li J, et al. Development and application of a deep learning-based comprehensive early diagnostic model for chronic obstructive pulmonary disease. Respir Res 2024 Apr 18;25(1):167. [doi: 10.1186/s12931-024-02793-3] [Medline: 38637823]
18. Zhang Z, Wu F, Zhou Y, et al. Detection of chronic obstructive pulmonary disease with deep learning using inspiratory and expiratory chest computed tomography and clinical information. J Thorac Dis 2024 Sep 30;16(9):6101-6111. [doi: 10.21037/jtd-24-367] [Medline: 39444883]
19. Zhang L, Jiang B, Wisselink HJ, Vliegenthart R, Xie X. COPD identification and grading based on deep learning of lung parenchyma and bronchial wall in chest CT images. Br J Radiol 2022 May 1;95(1133):20210637. [doi: 10.1259/bjr.20210637] [Medline: 35143286]
20. Zhang C, Liu J, Cao L, et al. Deep learning-based computed tomography features in evaluating early screening and risk factors for chronic obstructive pulmonary disease. Contrast Media Mol Imaging 2022;2022(1):5951418. [doi: 10.1155/2022/5951418] [Medline: 36051929]
21. Ying J, Dutta J, Guo N, et al. Classification of exacerbation frequency in the COPDGene Cohort Using Deep Learning With Deep Belief Networks. IEEE J Biomed Health Inform 2020 Jun;24(6):1805-1813. [doi: 10.1109/JBHI.2016.2642944] [Medline: 28026794]
22. Yang Y, Zeng N, Chen Z, et al. Multi-layer perceptron classifier with the proposed combined feature vector of 3D CNN Features and lung radiomics features for COPD stage classification. J Healthc Eng 2023;2023:3715603. [doi: 10.1155/2023/3715603] [Medline: 37953910]
23. Xue M, Jia S, Chen L, Huang H, Yu L, Zhu W. CT-based COPD identification using multiple instance learning with two-stage attention. Comput Methods Programs Biomed 2023 Mar;230. [doi: 10.1016/j.cmpb.2023.107356] [Medline: 36682106]
24. Xu C, Qi S, Feng J, et al. DCT-MIL: deep CNN transferred multiple instance learning for COPD identification using CT images. Phys Med Biol 2020 Jul 22;65(14). [doi: 10.1088/1361-6560/ab857d] [Medline: 32235077]
25. Wu Y, Du R, Feng J, et al. Deep CNN for COPD identification by multi-view snapshot integration of 3D airway tree and lung field. Biomed Signal Process Control 2023 Jan;79:104162. [doi: 10.1016/j.bspc.2022.104162]
26. Wu J, Lu Y, Dong S, Wu L, Shen X. Predicting COPD exacerbations based on quantitative CT analysis: an external validation study. Front Med (Lausanne) 2024;11:1370917. [doi: 10.3389/fmed.2024.1370917] [Medline: 38933101]

XSL•FO
RenderX

27. Wu CT, Li GH, Huang CT, et al. Acute exacerbation of a chronic obstructive pulmonary disease prediction system using wearable device data, machine learning, and deep learning: development and cohort study. JMIR Mhealth Uhealth 2021 May 6;9(5):e22591. [doi: 10.2196/22591] [Medline: 33955840]
28. Weikert T, Friebe L, Wilder-Smith A, et al. Automated quantification of airway wall thickness on chest CT using retina U-Nets—performance evaluation and application to a large cohort of chest CTs of COPD patients. Eur J Radiol 2022 Oct;155:110460. [doi: 10.1016/j.ejrad.2022.110460] [Medline: 35963191]
29. Wang R, Chen LC, Moukheiber L, et al. Enabling chronic obstructive pulmonary disease diagnosis through chest X-rays: a multi-site and multi-modality study. Int J Med Inform 2023 Oct;178:105211. [doi: 10.1016/j.ijmedinf.2023.105211] [Medline: 37690225]
30. Tang LYW, Coxson HO, Lam S, Leipsic J, Tam RC, Sin DD. Towards large-scale case-finding: training and validation of residual networks for detection of chronic obstructive pulmonary disease using low-dose CT. Lancet Digit Health 2020 May;2(5):e259-e267. [doi: 10.1016/S2589-7500(20)30064-9] [Medline: 33328058]
31. Sun J, Liao X, Yan Y, et al. Detection and staging of chronic obstructive pulmonary disease using a computed tomography-based weakly supervised deep learning approach. Eur Radiol 2022 Aug;32(8):5319-5329. [doi: 10.1007/s00330-022-08632-7] [Medline: 35201409]
32. Sugimori H, Shimizu K, Makita H, Suzuki M, Konno S. A comparative evaluation of computed tomography images for the classification of spirometric severity of the chronic obstructive pulmonary disease with deep learning. Diagnostics (Basel) 2021 May 21;11(6):929. [doi: 10.3390/diagnostics11060929] [Medline: 34064240]
33. Srivastava A, Jain S, Miranda R, Patil S, Pandya S, Kotecha K. Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease. PeerJ Comput Sci 2021;7:e369. [doi: 10.7717/peerj-cs.369] [Medline: 33817019]
34. Siebert JN, Hartley MA, Courvoisier DS, et al. Deep learning diagnostic and severity-stratification for interstitial lung diseases and chronic obstructive pulmonary disease in digital lung auscultations and ultrasonography: clinical protocol for an observational case-control study. BMC Pulm Med 2023 Jun 2;23(1):191. [doi: 10.1186/s12890-022-02255-w] [Medline: 37264374]
35. Sharma J, Vaid A, Nadkarni G, Kraft M. Diagnosis of chronic obstructive pulmonary disease using deep-learning on electrocardiograms. 2024 May Presented at: American Thoracic Society 2024 International Conference; May 17-22, 2024. [doi: 10.1164/ajrccm-conference.2024.209.1_MeetingAbstracts.A4522]
36. Seastedt KP, Litchman T, Moukheiber L, et al. Predicting chronic obstructive pulmonary disease from chest x-rays using deep learning. 2022 May Presented at: American Thoracic Society 2022 International Conference; May 13-18, 2022. [doi: 10.1164/ajrccm-conference.2022.205.1_MeetingAbstracts.A1078]
37. Sahu P, Kumar S, Behera AK. SOUNDNet: leveraging deep learning for the severity classification of chronic obstructive pulmonary disease based on lung sound analysis. Presented at: 2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT); Jul 12-14, 2024. [doi: 10.1109/CONECCT62155.2024.10677193]
38. Roy A, Gyanchandani B, Oza A, Singh A. TriSpectraKAN: a novel approach for COPD detection via lung sound analysis. Sci Rep 2025 Feb 21;15(1):6296. [doi: 10.1038/s41598-024-82781-1] [Medline: 39984500]
39. Nallanthighal VS, Harma A, Strik H. Detection of COPD exacerbation from speech: comparison of acoustic features and deep learning based speech breathing models. Presented at: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 23-27, 2022. [doi: 10.1109/ICASSP43922.2022.9747785]
40. Mou X, Wang P, Sun J, et al. A novel approach for the detection and severity grading of chronic obstructive pulmonary disease based on transformed volumetric capnography. Bioengineering (Basel) 2024 May 23;11(6):530. [doi: 10.3390/bioengineering11060530] [Medline: 38927766]
41. Mei S, Li X, Zhou Y, et al. Deep learning for detecting and early predicting chronic obstructive pulmonary disease from spirogram time series. NPJ Syst Biol Appl 2025 Feb 15;11(1):18. [doi: 10.1038/s41540-025-00489-y] [Medline: 39955293]
42. Makimoto K, Hague CJ, Hogg JC, Bourbeau J, Tan WC, Kirby M. Chronic obstructive pulmonary disease classification using sex-specific machine learning models and quantitative computed tomography imaging. 2024 May Presented at: American Thoracic Society 2024 International Conference; May 17-22, 2024. [doi: 10.1164/ajrccm-conference.2024.209.1_MeetingAbstracts.A5121]
43. Li Z, Huang K, Liu L, Zhang Z. Early detection of COPD based on graph convolutional network and small and weakly labeled data. Med Biol Eng Comput 2022 Aug;60(8):2321-2333. [doi: 10.1007/s11517-022-02589-x] [Medline: 35750976]
44. Lee AN, Hsiao A, Hasenstab KA. Evaluating the cumulative benefit of inspiratory CT, expiratory CT, and clinical data for COPD diagnosis and staging through deep learning. Radiol Cardiothorac Imaging 2024 Dec;6(6):e240005. [doi: 10.1148/ryct.240005] [Medline: 39665633]
45. Le Trung K, Nguyen Anh P, Han TT. A novel method in COPD diagnosing using respiratory signal generation based on CycleGAN and machine learning. Comput Methods Biomech Biomed Engin 2025 Jul;28(9):1538-1553. [doi: 10.1080/10255842.2024.2329938] [Medline: 38551327]
46. Doroodgar Jorshery S, Chandra J, Walia AS, et al. Leveraging deep learning of chest radiograph images to identify individuals at high risk for chronic obstructive pulmonary disease. medRxiv. 2024 Nov 15 p. 2024-11. [doi: 10.1101/2024.11.14.24317055] [Medline: 39606360]

XSL•FO
RenderX

47. Iturrioz Campo M, Nardelli P, Koc S, Bradford K, Krishnamurthy AK, San Jose Estepar R. COPD stage classification using deep learning on NHLBI biodata catalyst. 2022 May Presented at: American Thoracic Society 2022 International Conference; May 13-18, 2022. [doi: 10.1164/ajrccm-conference.2022.205.1_MeetingAbstracts.A1087]
48. Ho TT, Kim T, Kim WJ, et al. A 3D-CNN model with CT-based parametric response mapping for classifying COPD subjects. Sci Rep 2021 Jan 8;11(1):34. [doi: 10.1038/s41598-020-79336-5] [Medline: 33420092]
49. Hasenstab KA, Yuan N, Retson T, et al. Automated CT staging of chronic obstructive pulmonary disease severity for predicting disease progression and mortality with a deep learning convolutional neural network. Radiol Cardiothorac Imaging 2021 Apr;3(2):e200477. [doi: 10.1148/ryct.2021200477] [Medline: 33969307]
50. Guan Y, Zhang D, Zhou X, et al. Comparison of deep-learning and radiomics-based machine-learning methods for the identification of chronic obstructive pulmonary disease on low-dose computed tomography images. Quant Imaging Med Surg 2024 Mar 15;14(3):2485-2498. [doi: 10.21037/qims-23-1307] [Medline: 38545077]
51. Feng S, Zhang R, Zhang W, et al. Predicting acute exacerbation phenotype in chronic obstructive pulmonary disease patients using VGG-16 deep learning. Respiration 2025;104(1):1-14. [doi: 10.1159/000540383] [Medline: 39047695]
52. El Boueiz AR, Dy JG, Ross JC, et al. Deep learning prediction of COPD progression using enriched densitometry phenotypes. 2019 May Presented at: American Thoracic Society 2019 International Conference; May 17-22, 2019. [doi: 10.1164/ajrccm-conference.2019.199.1_MeetingAbstracts.A4054]
53. Du R, Qi S, Feng J, et al. Identification of COPD from multi-view snapshots of 3D lung airway tree via deep CNN. IEEE Access 2020;8:38907-38919. [doi: 10.1109/ACCESS.2020.2974617]
54. Davies HJ, Hammour G, Xiao H, et al. Physically meaningful surrogate data for COPD. IEEE Open J Eng Med Biol 2024;5:148-156. [doi: 10.1109/OJEMB.2024.3360688] [Medline: 38487098]
55. D Almeida S, Norajitra T, Lüth CT, et al. How do deep-learning models generalize across populations? Cross-ethnicity generalization of COPD detection. Insights Imaging 2024 Aug 7;15(1):198. [doi: 10.1186/s13244-024-01781-x] [Medline: 39112910]
56. Cosentino J, Behsaz B, Alipanahi B, et al. Inference of chronic obstructive pulmonary disease with deep learning on raw spirograms identifies new genetic loci and improves risk models. Nat Genet 2023 May;55(5):787-795. [doi: 10.1038/s41588-023-01372-4]
57. Christina Dally E, Banu Rekha B. Automated chronic obstructive pulmonary disease (COPD) detection and classification using Mayfly optimization with deep belief network model. Biomed Signal Process Control 2024 Oct;96:106488. [doi: 10.1016/j.bspc.2024.106488]
58. Chen J, Xu Z, Sun L, et al. Deep learning integration of chest computed tomography imaging and gene expression identifies novel aspects of COPD. Chronic Obstr Pulm Dis 2023 Oct 26;10(4):355-368. [doi: 10.15326/jcopdf.2023.0399] [Medline: 37413999]
59. Chaudhary MFA, Awan HA, Gerard SE, et al. Deep learning estimation of small airways disease from inspiratory chest CT is associated with FEV1 decline in COPD. medRxiv. Preprint posted online on Sep 11, 2024. [doi: 10.1101/2024.09.10.24313079] [Medline: 39314974]
60. Cai N, Xie Y, Cai Z, Liang Y, Zhou Y, Wang P. Deep learning assisted diagnosis of chronic obstructive pulmonary disease based on a local-to-global framework. Electronics (Basel) 2024;13(22):4443. [doi: 10.3390/electronics13224443]
61. Bao Y, Al Makady Y, Mahmoodi S. Automatic diagnosis of COPD in lung CT images based on multi-view DCNN. 2021 Presented at: 10th International Conference on Pattern Recognition Applications and Methods; Feb 4-6, 2021. [doi: 10.5220/0010296805710578]
62. Awan HA, Chaudhary MFA, Gerard SE, et al. Deep residual convolutional network predicts future severe exacerbations of COPD in SPIROMICS. 2023 May Presented at: American Thoracic Society 2023 International Conference; May 19-24, 2023. [doi: 10.1164/ajrccm-conference.2023.207.1_MeetingAbstracts.A3318]
63. Alve SR, Mahmud MZ, Islam S, Khan MM. Chronic obstructive pulmonary disease prediction using deep convolutional network. medRxiv. Preprint posted online on Dec 24, 2024. [doi: 10.1101/2024.12.22.24319500]
64. Altan G, Kutlu Y, Gökçen A. Chronic obstructive pulmonary disease severity analysis using deep learning on multi-channel lung sounds. Turk J Elec Eng & Comp Sci 2020;28(5):2979-2996. [doi: 10.3906/elk-2004-68]
65. Almeida SD, Norajitra T, Lüth CT, et al. Prediction of disease severity in COPD: a deep learning approach for anomaly-based quantitative assessment of chest CT. Eur Radiol 2024 Jul;34(7):4379-4392. [doi: 10.1007/s00330-023-10540-3] [Medline: 38150075]
66. Dorosti T, Schultheiss M, Hofmann F, et al. Optimizing convolutional neural networks for chronic obstructive pulmonary disease detection in clinical computed tomography imaging. Comput Biol Med 2025 Feb;185:109533. [doi: 10.1016/j.compbiomed.2024.109533] [Medline: 39705795]
67. Feng S, Zhang W, Zhang R, et al. The identification and severity staging of chronic obstructive pulmonary disease using quantitative CT parameters, radiomics features, and deep learning features. Respiration 2025 Sep 25;25:1-13. [doi: 10.1159/000548595] [Medline: 40996946]
68. Azad Rabby AS, Chaudhary MFA, Saha P, et al. Light convolutional neural network to detect chronic obstructive pulmonary disease (COPDxNET): a multicenter model development and external validation study. medRxiv. Preprint posted online on Aug 1, 2025. [doi: 10.1101/2025.07.30.25332459] [Medline: 40766163]

XSL•FO
RenderX

69. Rezvanjou S, Moslemi A, Peterson S, et al. Classifying chronic obstructive pulmonary disease status using computed tomography imaging and convolutional neural networks: comparison of model input image types and training data severity. J Med Imag (Bellingham) 2025;12(3):034502. [doi: 10.1117/1.JMI.12.3.034502] [Medline: 40415865]
70. Rahaman Wahab Sait A, Kumar Dutta A, Ahmed Shaikh M. Optimized Kolmogorov–Arnold networks-driven chronic obstructive pulmonary disease detection model. IEEE Access 2025;13:162947-162960. [doi: 10.1109/ACCESS.2025.3610633]
71. Sahu P, Prasad P, Verma VP, Kumar S. Deep learning framework for early diagnosis of COPD and respiratory diseases using lung sound analysis. 2025 Presented at: International Conference on Big Data Analytics; Dec 17-20, 2024; Hyderabad, India p. 295-304. [doi: 10.1007/978-3-031-81821-9_17]
72. Wu Y, Xia S, Liang Z, Chen R, Qi S. Artificial intelligence in COPD CT images: identification, staging, and quantitation. Respir Res 2024 Aug 22;25(1):319. [doi: 10.1186/s12931-024-02913-z] [Medline: 39174978]
73. Willer K, Fingerle AA, Noichl W, et al. X-ray dark-field chest imaging for detection and quantification of emphysema in patients with chronic obstructive pulmonary disease: a diagnostic accuracy study. Lancet Digit Health 2021 Nov;3(11):e733-e744. [doi: 10.1016/S2589-7500(21)00146-1] [Medline: 34711378]
74. Balasubramanian P, Abia-Trujillo D, Barrios-Ruiz A, et al. Diagnostic yield and safety of diagnostic techniques for pulmonary lesions: systematic review, meta-analysis and network meta-analysis. Eur Respir Rev 2024 Jul;33(173):240046. [doi: 10.1183/16000617.0046-2024] [Medline: 39293856]
75. Arts L, Lim EHT, van de Ven PM, Heunks L, Tuinman PR. The diagnostic accuracy of lung auscultation in adult patients with acute pulmonary pathologies: a meta-analysis. Sci Rep 2020 Apr 30;10(1):7347. [doi: 10.1038/s41598-020-64405-6] [Medline: 32355210]

Abbreviations

AI: artificial intelligence
AUC: area under the summary receiver operating characteristic curve
COPD: chronic obstructive pulmonary disease
CT: computed tomography
CXR: chest X-ray
DL: deep learning
DOR: diagnostic odds ratio
GOLD: Global Initiative for Chronic Obstructive Lung Disease
NLR: negative likelihood ratio
PFT: pulmonary function testing
PLR: positive likelihood ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension
PROSPERO: International Prospective Register of Systematic Reviews
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2
RoB: risk of bias

Edited by S Brini; submitted 03.Sep.2025; peer-reviewed by I Yang, Y Liao; revised version received 09.Dec.2025; accepted 09.Dec.2025; published 14.Jan.2026.

Please cite as:
Yang H, Wu Y, Wu T, Ji J, Lei S, Xu W
Accuracy of Deep Learning in Diagnosing Chronic Obstructive Pulmonary Disease: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e83459
URL: https://www.jmir.org/2026/1/e83459
doi:10.2196/83459

© Hui Yang, Yijiu Wu, Tong Wu, Jingyan Ji, Sitao Lei, Weibin Xu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessment of the Diagnostic Performance and Clinical Impact of AI in Hepatic Steatosis: Systematic Review and Meta-Analysis

Jiamei Song*, MM; Dan Liu*, MS; Jitong Li, MM; Haoru Cong, MM; Ruixue Deng, MM; Yihan Lu, MBBS; Jiayi Sun, MBBS; Jingzhou Zhang, MD

College of Traditional Chinese Medicine, Changchun University of Chinese Medicine, No 1035, Boshuo Road Jingyue National High-Tech Industrial Development Zone Changchun City, Changchun, Jilin, China

*these authors contributed equally

Corresponding Author:
Jingzhou Zhang, MD
College of Traditional Chinese Medicine, Changchun University of Chinese Medicine, No 1035, Boshuo Road Jingyue National High-Tech Industrial Development Zone Changchun City, Changchun, Jilin, China

Abstract

Background: The global rise of metabolic associated fatty liver disease reflects the urgent need for accurate, noninvasive diagnostic approaches. The invasive nature of liver biopsy and the limited sensitivity of ultrasound in detecting early steatosis highlight a critical diagnostic gap. Artificial intelligence (AI) has emerged as a transformative tool, enabling the automated detection and grading of hepatic steatosis (HS) from medical imaging data.

Objective: This review aims to quantitatively evaluate the diagnostic performance of AI models for HS, explore sources of interstudy heterogeneity, and provide an appraisal of their clinical applicability, translational potential, and the major barriers impeding widespread implementation.

Methods: PubMed, Cochrane Library, Embase, Web of Science, and IEEE Xplore databases were searched until September 24, 2025. Studies using AI for HS diagnosis, meeting predefined PIRT (Patient Selection, Index Test, Reference Standard, Flow and Timing) framework and providing extractable data were included. Diagnostic performance indicators, including sensitivity, specificity, and the area under the summary receiver operating characteristic curve (AUC), were extracted and quantitatively synthesized. Meta-analyses were conducted using a bivariate random effects model. The methodological quality and risk of bias were evaluated using the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) tool. Heterogeneity was assessed through the I^2 statistic, bivariate box plots, 95% PIs, and threshold effect analysis. Clinical applicability was examined using the Fagan nomogram and likelihood ratio tests.

Results: A total of 36 eligible studies were identified, of which 33 (comprising 36 cohorts) were included in the subgroup analyses. Results demonstrated excellent diagnostic accuracy of AI models, with a summary sensitivity of 0.95 (95% CI 0.93-0.96), specificity of 0.93 (95% CI 0.91-0.94), and an AUC of 0.98 (95% CI 0.96-0.99). Clinical applicability analysis (positive likelihood ratio >10; negative likelihood ratio <0.1) supported AI's strong potential for both confirming and excluding HS. However, substantial heterogeneity was observed across studies (I^2 >75%). According to QUADAS-2, a high risk of bias, particularly in the Patient Selection domain (44.4%), may have contributed to the overestimation of real-world performance. Subgroup analyses showed that deep learning models significantly outperformed traditional machine learning approaches (AUC: 0.98 vs 0.94). Models using ultrasound or histopathology references, retrospective designs, transfer learning, and public datasets achieved the highest accuracy (AUC 0.98-0.99) but contributed to interstudy heterogeneity.

Conclusions: AI demonstrates remarkable potential for noninvasive screening and assessment of HS, especially in primary care. Nonetheless, clinical translation remains limited by performance variability, retrospective designs, lack of external validation, practical barriers such as data privacy and workflow integration. Future studies should prioritize prospective multicenter trials and standardized external validation to bridge the gap between current evidence and clinical application. The key innovation of this review lies in establishing a unified, modality-agnostic analytical framework that integrates evidence beyond single-modality evaluations.

Trial Registration: PROSPERO CRD420251046862; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251046862

(J Med Internet Res 2026;28:e78310) doi:10.2196/78310

KEYWORDS

artificial intelligence; AI; diagnostic performance; hepatic steatosis; meta-analysis; clinical impact

XSL•FO
RenderX

Introduction

Metabolic-associated fatty liver disease (MAFLD) has emerged as one of the most prevalent chronic liver diseases worldwide, with its pathophysiology intrinsically linked to metabolic syndrome. Affected individuals frequently exhibit concomitant metabolic abnormalities such as central obesity, type 2 diabetes mellitus, and insulin resistance. The disease spectrum of MAFLD represents a continuum ranging from simple hepatic steatosis (HS) to metabolic dysfunction-associated steatohepatitis (MASH), which may progress to hepatic fibrosis, cirrhosis, or hepatocellular carcinoma (HCC) [1]. Therefore, MAFLD is a significant and growing global public health threat [2].

In 2020, an international consensus panel proposed renaming "non-alcoholic fatty liver disease" to "MAFLD" to better reflect its metabolic foundation [3,4]. To ensure the present study focuses on the diagnostic performance of artificial intelligence (AI) for the core pathological feature of HS and to enhance its clinical generalizability, including to populations with mixed etiologies such as concomitant metabolic disorders and alcohol use, we adopted the broader term "MAFLD." This terminology aligns more closely with real-world clinical practice and provides a consistent framework for AI model training and validation. Epidemiological data estimate the global prevalence of MAFLD at approximately 38%, with substantial regional variation, the highest burdens observed in Latin America, the Middle East, and North Africa [5]. The disease is also increasingly recognized among pediatric and adolescent populations, particularly in individuals with obesity, where prevalence rates have been reported to range from 7% to 14% or higher [6].

Nevertheless, the reported MAFLD prevalence varies markedly across studies, from 5% to 46% [7], reflecting considerable heterogeneity. First, diagnostic methodologies differ. Although liver biopsy remains the histopathological gold standard, its invasiveness limits clinical use, shifting reliance toward multimodal imaging. Noninvasive modalities such as ultrasound and computed tomography (CT) are widely used due to accessibility and low cost, but they lack precision in quantifying HS. Quantitative imaging techniques, including magnetic resonance imaging–proton density fat fraction (MRI-PDFF) [8], controlled attenuation parameter-based transient elastography, and noninvasive analysis [9], offer superior accuracy but are constrained by cost and limited availability. Clinical prediction models such as the Fatty Liver Index [10], Hepatic Steatosis Index [11], and Liver Fat Equation [12] enable noninvasive diagnosis through integration of anthropometric and biochemical parameters. Nevertheless, they remain vulnerable to measurement variability and lack use for longitudinal monitoring. Second, the sensitivity of existing diagnostic modalities in detecting early-stage steatosis (hepatic fat content <5%) remains suboptimal. Conventional ultrasound, in particular, has a high false-negative rate when hepatic fat content falls below 20% [13], leading to underdiagnosis and misdiagnosis in subclinical populations.

Such diagnostic inaccuracies carry serious clinical implications. Patients erroneously classified as having "simple MASLD" but who also exhibit alcohol use disorder have been shown to experience mortality risks exceeding those of individuals with typical alcoholic liver disease [14]. As MAFLD incidence rises globally, associated cirrhosis and HCC cases are also increasing. Failure to achieve early and accurate diagnosis forfeits the therapeutic window during the reversible steatosis stage, allowing progression to MASH and fibrosis. Notably, MAFLD-related HCC may arise in noncirrhotic livers [15], challenging conventional surveillance strategies that primarily target cirrhotic patients. Moreover, MAFLD is an established independent risk factor for cardiovascular disease [16]. This elevated cardiovascular risk persists throughout the disease course and remains heightened even following liver transplantation [17], underscoring the necessity of lifelong risk management.

Recent advances in AI have revolutionized medical image analysis, and hepatology has been no exception. AI-based approaches have demonstrated strong diagnostic performance across multiple hepatic pathologies. For instance, Meng et al [18] developed a VGGNet-based multistage fibrosis classifier, achieving high accuracy across 3 fibrosis grades. Wang et al [19] introduced the Explainable Diagnosis Recommender intelligent diagnostic system, which uses deep learning (DL) to automatically detect hepatic echinococcosis and cysts from CT scans. Xiao et al [20] proposed a ResNet-101-based multimodal model that classified 6 hepatobiliary diseases using slit-lamp and fundus images, outperforming clinicians of varying experience levels. Calderaro et al [21] used a DL model to reclassify combined hepatocellular-cholangiocarcinoma into pure HCC or intrahepatic cholangiocarcinoma with high sensitivity and specificity, yielding predictions consistent with clinical and molecular profiles. Specifically for HS assessment, Yang et al [22] developed a 2-stage DL model that classified four steatosis grades with an overall accuracy of 76.3% and an area under the summary receiver operating characteristic (SROC) curve (AUC) of 0.88, surpassing traditional clinical indices. Similarly, Wang et al [23] employed DL to quantify hepatic fat content by inferring proton density fat fraction (PDFF) from routine T1-weighted magnetic resonance imaging (MRI) images, surpassing the performance of the conventional 2-point Dixon fat-fraction model.

Despite these promising developments, the application of AI in the diagnosis and grading of MAFLD or HS remains at an early stage [24]. Existing systematic reviews have primarily assessed AI performance within individual imaging modalities. A critical gap remains: a comprehensive evaluation of AI's overall diagnostic efficacy across diverse imaging platforms and a systematic analysis of the key technical and methodological determinants of performance are still lacking.

Therefore, this study, for the first time, uses a bivariate mixed effects model [25] to systematically assess the overall diagnostic performance of AI in imaging-based detection of HS. The primary objectives are: (1) to quantitatively determine the aggregate diagnostic accuracy of AI models in identifying HS; (2) to comprehensively explore the sources of heterogeneity, with particular emphasis on the influence of factors such as

XSL•FO
RenderX

algorithm type, reference standard, imaging modality, study design, and data accessibility; and (3) to evaluate the clinical applicability and translational potential of AI-based diagnostic systems, while identifying major barriers to their broad clinical adoption. Through these aims, the present study seeks to generate robust, high-level evidence that transcends the limitations of individual analytical approaches, thereby providing meaningful guidance for future research and clinical practice.

Methods

Research Design and Clinical Questions

This study protocol was registered with the PROSPERO International Prospective Register of Systematic Reviews (Registration: CRD420251046862). The research was conducted as per the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Diagnostic Test Accuracy) guidelines (Checklist 1) [26].

Search Strategy

This systematic search was independently designed and conducted by two researchers as per the PRISMA-DTA (Checklist 1) [26]. PubMed, the Cochrane Library, Embase, Web of Science, and IEEE Xplore were retrieved until September 24, 2025. The search strategy was structured around three core concepts: (1) the disease ("HS," "non-alcoholic fatty liver disease," and "MAFLD"); (2) the technology ("AI," "machine learning [ML]," "DL"); and (3) the diagnostic context ("diagnosis," "detection").

Keywords within each conceptual category were combined via the OR operator (eg, AI OR DL OR ML), whereas keywords across different categories were linked using the AND operator (eg, AI AND MAFLD AND diagnosis).

The authors of the identified studies were not contacted. Reference lists of all included studies were manually reviewed to identify any additional eligible publications. No restrictions on language or publication date were applied at the database level to maximize search sensitivity. However, non-English records were excluded during subsequent screening. Gray literature, preprints, and unpublished studies were not systematically searched. This decision was made a priori to focus on peer-reviewed, full-text articles that had undergone editorial review, thereby ensuring baseline methodological quality and the availability of sufficient details for data extraction. The complete, reproducible search strings for all databases are provided in Table S1 in Multimedia Appendix 1.

Screening Process

Two independent reviewers initially screened all retrieved titles and abstracts. After removing duplicate records, studies were deemed eligible for inclusion if they met the following criteria:

- Study content: the research conformed to the predefined PIRT (Patient Selection, Index Test, Reference Standard, Flow and Timing) framework:
  - Patient Selection (P): patients undergoing abdominal imaging or pathological examination for HS assessment.
  - Index Test (I): AI models based on DL or ML, using input images derived from ultrasound, CT, MRI, or pathology.
  - Reference Standard (R): defined by the original studies, including MRI-PDFF, liver biopsy pathology, or expert-graded ultrasound. These reference standards reflect real-world diagnostic diversity and were recognized as potential sources of heterogeneity.
  - Target Condition (T): Diagnosis and grading of HS according to the thresholds and criteria adopted in the included studies, allowing cross-comparison of AI performance across varying diagnostic definitions.
- Data availability: studies had to provide diagnostic contingency data, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), or sufficient information to derive diagnostic performance metrics, such as AUC with 95% CIs, sensitivity, specificity, accuracy, and predictive values.

Exclusion criteria were: (1) language: non-English publications; (2) study type: letters, conference abstracts, reviews, or academic papers lacking original data; (3) study subjects: animal or nonhuman research, bioinformatics-based analyses, and predictive modeling studies focused on indices, risks, or associations rather than diagnosis; and (4) data sufficiency: studies without key contingency data or insufficient information to calculate diagnostic performance metrics.

Data Extraction

Two researchers independently extracted data based on the following domains: (1) study characteristics: first author, publication year, site of data collection, and duration of the study period; (2) study population: total sample size, and demographic characteristics (mean or median age); (3) methodological parameters: accessibility of clinical sample data, diagnostic reference standard, and validation strategy; (4) algorithmic architecture: type of algorithm, classifier employed, and application of transfer learning (TL); and (5) diagnostic efficacy: raw contingency table data, and aggregated diagnostic performance metrics.

Diagnostic Performance Evaluation and Quality Assessment

Pooled estimates of sensitivity, specificity, and AUC, together with their 95% CIs, were presented using forest plots. Heterogeneity was quantified using the I^2 statistic. The AUC was designated as the primary indicator for overall diagnostic accuracy, as it integrates performance across all thresholds and remains unaffected by any single cut-off point. A SROC curve was constructed following an assessment of the threshold effect using the Spearman correlation coefficient between the logit of sensitivity and the logit of (1-specificity). Heterogeneity and its implications were further visualized via 95% PIs and bivariate boxplots. Potential small-study effects were evaluated using the Deeks funnel plot asymmetry test. Additional diagnostic indicators, including the diagnostic odds ratio (DOR), positive likelihood ratio (LRP), and negative likelihood ratio (LRN), were calculated. Clinical applicability was further examined

XSL•FO
RenderX

using a Fagan nomogram, while the distribution of likelihood ratios across studies was illustrated via scatterplots.

Two investigators assessed the risk of bias via the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) checklist (Checklist 2) in Rev-Man in terms of PIRT framework [27]. The QUADAS-2 tool, recommended by the Cochrane collaboration, was used to assess the methodological quality and risk of bias of included diagnostic accuracy studies.

Quality Assurance and Dispute Resolution

All screening, data extraction, and quality assessment procedures were independently conducted by 2 reviewers. Any discrepancies were first resolved through discussion to reach a consensus. When consensus was not achieved, a third senior investigator adjudicated the disagreement to make the final determination. This multilevel process ensured that all extracted data represented unanimous agreement within the research team.

Subgroup Analysis

The following independent meta-analyses were conducted:

1. AI type (ML versus DL): DL models were defined as those based on multilayer artificial neural networks, such as convolutional neural networks and recurrent neural networks (RNNs). This category included all studies that explicitly reported using DL or identified architectures such as VGG, ResNet, U-Net, DenseNet, or Inception. ML models referred to traditional algorithms that learn from data without relying primarily on deep neural architectures, including support vector machines (SVM), random forests, decision trees, and logistic regression. To explore the potential influence of different algorithmic approaches on diagnostic performance.
2. Reference standards for steatosis grading (MRI-PDFF, liver histopathology, or ultrasound): to determine whether there was a performance gap between models based on noninvasive imaging and those based on the pathological "gold standard."
3. Imaging modality (ultrasound, CT, or histopathology): to assess how differences in imaging principles, invasiveness, and the diagnostic information scale (macroscopic versus microscopic) affected model performance.
4. Application of TL: TL was used when a study explicitly reported the use of a model pretrained on a large-scale dataset (eg, ImageNet) as the initial framework for feature extraction or model fine-tuning. To evaluate whether this specific technique could improve model performance in small-sample medical datasets.
5. Study design (single-center versus multicenter): to assess the generalizability of models across different data distributions.
6. Study type (prospective vs retrospective): to explore the temporal relationship between data collection and model development and to evaluate the potential impact of selection bias on performance assessment.
7. Data accessibility: to evaluate the effect of study reproducibility and transparency on research outcomes.

Data Analysis

Given the substantial heterogeneity observed among included studies with respect to patient populations, imaging devices, and AI algorithms, a bivariate mixed effects model was used to derive more accurate and reliable pooled estimates [25]. To ensure the robustness of the meta-analytic results, quantitative synthesis (eg, subgroup analysis) was performed only when at least 3 independent studies, defined as studies conducted by different authors, using distinct experimental protocols, or involving separate participant cohorts, were available. Multiple effect estimates from the same publication were included when they originated from distinct participant cohorts (eg, multicenter datasets or independent validation sets). When multiple model outputs were reported, only the best-performing model or that validated using an independent dataset was retained. Subgroup analyses were not conducted when fewer than three independent studies were available for a given subgroup. All statistical analyses and visualizations were performed via Stata MP 18 (StataCorp LLC). A 2-tailed P value <.05 denoted statistical significance.

Results

Included Study Description

As of September 24, 2025, 2536 articles were retrieved. After removing 864 duplicates, the titles and abstracts of the rest were screened as per the predefined eligibility criteria, resulting in the exclusion of 1596 articles. Specifically, 9 were non-English publications, 884 were of other types, 673 involved inappropriate study subjects, and 30 used unsuitable research methods. The full texts of the remaining 76 articles were subsequently reviewed. Seventeen studies were excluded for incomplete data, 7 for being of other types, 7 for inappropriate methodologies, and 9 for being inaccessible. Ultimately, 36 studies were included in the final analysis (Figure 1). The characteristics of the included studies are summarized in Table 1, and the results of the subgroup analyses are presented in Table 2.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Search Extension) flowchart depicting the study selection process for the systematic review of artificial intelligence in diagnosing hepatic steatosis.

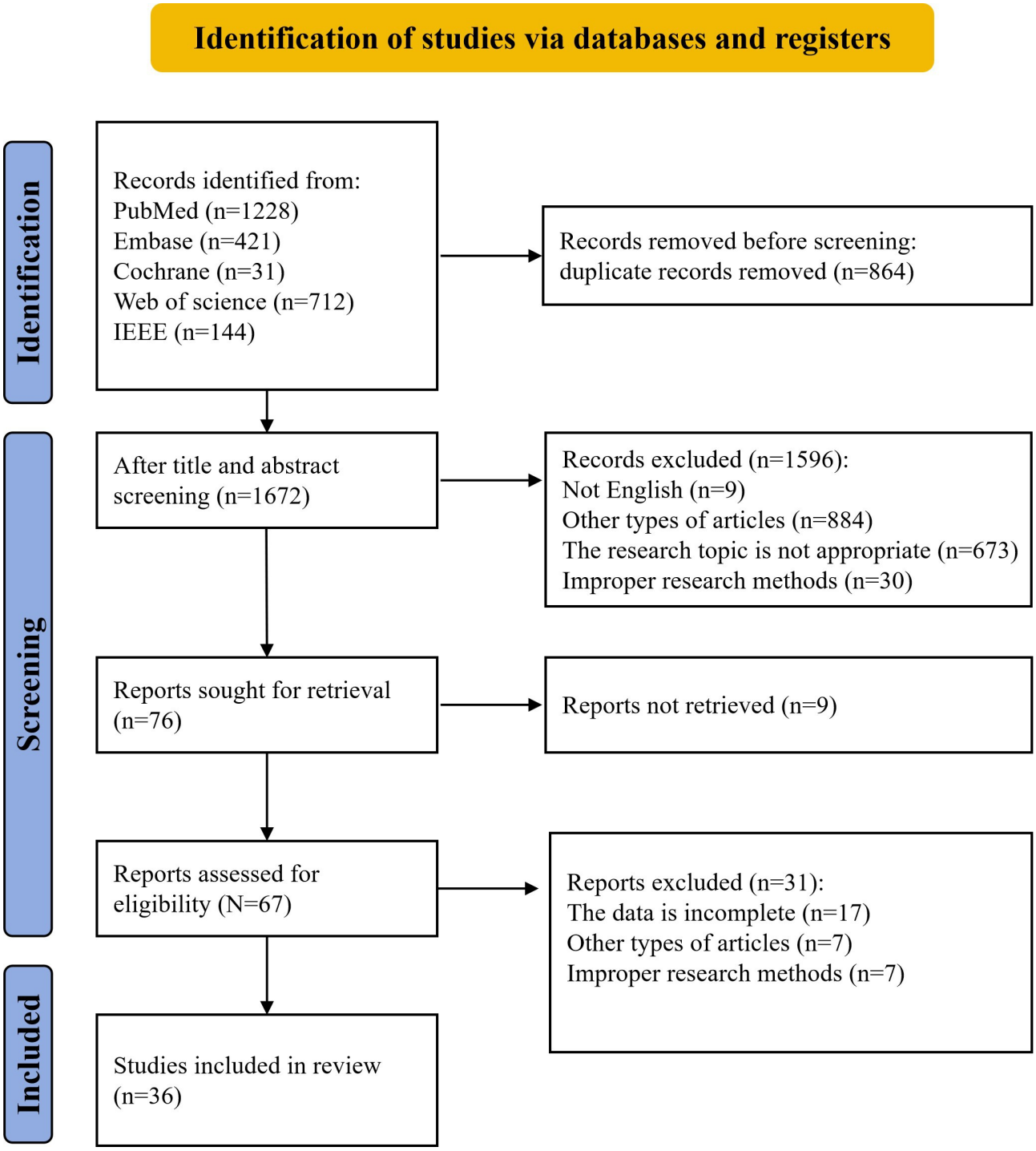

XSL•FO
RenderX

Table . Summary of study characteristics in the systematic review of artificial intelligence (AI)-assisted hepatic steatosis (HS) detection (N=36).

Number	Study	Device	DL[a] or ML[b]	TL[c]	Source of data	Data range	Sample size	Age, years	Open access data	Research type
1	Yang et al (2022) [28]	MRI[d]	DL	Yes	A hospital in Beijing	Jun - Jul 2020	50	• NR[e]	No	• Retrospective study • Single-center
2	Acharya et al (2016) [29]	Ultrasound	DL	NR	The University of Malaya Medical Centre, Malaysia	NR	100	• NR	No	• Retrospective study • Single-center
3	Jeon et al (2023) [30]	Ultrasound	DL	NR	Seoul National University Hospital	Jul 2020-Jun 2021	173	• mean (SD): 51 (14) • range: 19-74	No	• Prospective study • Single-center
4	Neogi et al (2018) [31]	Ultrasound	DL	NR	Chittaranjan National Cancer Hospital	NR (a span of 6 months)	51	• NR	No	• Retrospective study • Single-center
5	Chen et al (2020) [32]	Ultrasound	DL	NR	Chang Gung Memorial Hospital in Taiwan	2017 - 2020	205	• mean (SD): 55 (11.6)	No	• Retrospective study • Single-center
6	Dubois et al (2019) [33]	Ultrasound	DL	NR	Rennes University Hospital	Jun 2017 - Aug 2018	53	• median (IQR): 61 (28–72)	No	• Prospective study • Single-center
7	Shi et al (2019) [34]	Ultrasound	ML	NR	Shanghai Public Health Clinical Center	NR	60	• range: 19 - 69	No	• Retrospective study • Single-center
8	Jesper et al (2020) [35]	Ultrasound	ML	NR	Erlangen University Hospital	Oct 2018 - Sep 2019	27	• mean (SD): 50 (17)	No	• Prospective study • Single-center

XSL•FO
RenderX

Number	Study	Device	DL[a] or ML[b]	TL[c]	Source of data	Data range	Sample size	Age, years	Open access data	Research type
9	McHenry et al (2020) [36]	MRI	DL	NR	Dallas County	2000-2002; 2007-2009	2139	• median: 44	No	• Prospective study • Single-center
10	Roy et al (2021) [37]	Pathology	DL	NR	Children's Hospital of Atlanta and Emory University	2014 - 2016	36	• mean (SD): 14.9 (2.59)	Yes	• Retrospective study • Multicenter
11	Sun et al (2020) [38]	Pathology	DL	Yes	Washington University School of Medicine Transplant Pathology Service	Apr 2015 - Sep 2016	91	• NR	No	• Retrospective study • Multicenter
12	Chou et al (2021) [39]	Ultrasound	DL	Yes	Taipei Medical University Hospital	2016 - 2018	2070	• NR	No	• Retrospective study • Single-center
13	Constantinescu et al (2021) [40]	Ultrasound	DL	Yes	Outpatient clinic of a private healthcare network	NR	60	• range: 18 - 92	No	• Prospective study • Single-center
14	Pérez-Sanz et al (2021) [41]	Pathology	DL	NR	University Clinical Hospital Virgen de la Arrixaca-Biomedical Research Institute of Murcia	NR	20	• NR	No	• Retrospective study • Single-center
15	Pickhardt et al (2020) [42]	CT	DL	NR	A single academic medical center	Feb 2010 - Jan 2017	1204	• mean (SD): 45.2 (12.4)	No	• Retrospective study • Single-center
16	Rhyou et al (2021) [43]	Ultrasound	DL	Yes	Samsung Medical Center and Byra Dataset	NR	NR	• NR	Yes	• Retrospective study • Multicenter
17		Ultrasound	ML	NR			82		No	

XSL•FO
RenderX

Number	Study	Device	DL[a] or ML[b]	TL[c]	Source of data	Data range	Sample size	Age, years	Open access data	Research type
	Destrempes et al (2022) [44]				Center Hospitalier de l'Universite´ de Montre´al and McGill University Health Center	Oct 2014 - Sep 2018		• mean (SD): 56 (12) • range: 23-78		• Retrospective study • Multicenter
18	Alshagathrh et al (2023) [45]	Ultrasound	DL	Yes	University of Warsaw, Poland	NR	55	• mean (SD): 40.1 (9.1)	Yes	• Retrospective study • Single-center
19	Podder et al (2023) [46]	Pathology	DL	Yes	Open Science Framework	NR	NR	• NR	Yes	• Retrospective study • Single-center
20	Ibrahim et al (2023) [47]	Ultrasound	DL	Yes	Beijing You'an Hospital in Beijing, China, and the National Hepatology and Tropical Medicine Research Institute in Cairo, Egypt	NR	478	• mean (SD): 40.97 (10.61)	No	• Prospective study • Single-center
21	Yao et al (2023) [48]	Ultrasound	DL	Yes	Byra dataset and the Health Service Center in the Chenghua District of Chengdu	2020 - 2022	1320	• NR	Yes	• Retrospective study • Multicenter
22	Byra et al (2018) [49]	Ultrasound	DL	Yes	Medical University of Warsaw, Poland	NR	55	• mean (SD): 40.1 (9.1)	Yes	• Prospective study • Multicenter
23	Torgersen et al (2024) [50]	CT	DL	NR	Philadelphia VA[f] Medical Center	01 Jan 2010 - 30 Dec 2017	120	• 61.1 (55.3 - 64.6)	No	• Retrospective study • Single-center
24		Ultrasound	DL	NR		NR	131		No	

XSL•FO
RenderX

Number	Study	Device	DL[a] or ML[b]	TL[c]	Source of data	Data range	Sample size	Age, years	Open access data	Research type
	Wang et al (2023) [51]				Chang Gung Memorial Hospital, Taiwan			• Mean (age range) • [illegible] • [illegible] • [illegible]		• Prospective study • Single-center
25	Jeon et al (2024) [52]	CT	ML	NR	Institute of Radiation Medicine, Seoul National University Medical Research Center, Seoul National University Hospital, Seoul, Korea	Dec 2018 - Dec 2021	252	• mean: 37.3 • range: 18-64	No	• Retrospective study • Single-center
26	Piella et al (2024) [53]	Mobile phones	ML	NR	Vall d'Hebron University Hospital	NR	192	• median (IQR): 62 (50.25 - 71.75)	No	• Retrospective study • Single-center
27	Yoo et al (2024) [54]	CT	DL	NR	Radiologic database in our institution	Jan 2017 - Jun 2021	362	• mean (SD): 37.3 (11.5) • range: 18-65	No	• Retrospective study • Single-center
28	Zhang et al (2024) [55]	CT	DL	NR		NR	986	• NR	Yes	• Retrospective study • Multi-center

XSL•FO
RenderX

Number	Study	Device	DL[a] or ML[b]	TL[c]	Source of data	Data range	Sample size	Age, years	Open access data	Research type
					LIDC-IDRI, NSCLC-Lung1, RIDER, VES-SEL12, MIDRC-RICORD, COVID-19-Italy, and COVID-19-China					
29	Cherchi et al (2021) [56]	Pathology	DL	NR	The University Hospital of Udine	Jan 2018 - May 2019	33	• NR	No	• Retrospective study • Single-center
30	Wu et al (2023) [57]	Ultrasound	DL	NR	Chang Gung Memorial Hospital	NR	276	• NR	No	• Retrospective study • Single-center
31	Drazinos et al (2025) [58]	Ultrasound	DL	NR	The University of Texas MD Anderson Cancer Center	Jan 2018 - Jan 2019	112	• mean (SD): 51 (16.13)	No	• Retrospective study • Single-center
32	Kaffas et al (2025) [59]	Ultrasound	DL	NR	NR	01 Jan 2010 - 01 Jan 2022	403	• median (IQR): 53 (40 - 66)	No	• Retrospective study • Single-center
33	Kim et al (2025) [60]	CT	DL	NR	Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea	2001 - 2016	3620	• mean (SD): 31.7 (9.4)	No	• Retrospective study • Single-center
34	Zhang et al (2025) [61]	CT	DL	NR	The First Affiliated Hospital of Zhengzhou University	Jul 2022 - May 2023	840	• mean (SD): 49.1 (11.5)	Yes	• Retrospective study • Single-center
35		CT	DL	NR		NR	1740		Yes	

XSL•FO
RenderX

Number	Study	Device	DL[a] or ML[b]	TL[c]	Source of data	Data range	Sample size	Age, years	Open access data	Research type
	Derstine et al (2025) [62]				Michigan Medicine; VA			• mean (SD): 43.1 (12.8)		• Retrospective study • Single-center
36	Corso et al (2024) [63]	Ultrasound	DL	NR	The University of Texas MD Anderson Cancer Center	NR	186	• mean (SD): 51.95 (13.4)	No	• Prospective study • Multicenter

[a]DL: deep learning.

[b]ML: machine learning.

[c]TL: transfer learning.

[d]MRI: magnetic resonance imaging.

[e]NR: No report.

[f]VA: Veterans Administration.

XSL•FO
RenderX

Table . Summary of artificial intelligence (AI) performance for diagnosing hepatic steatosis (HS) based on different subgroups (number of studies or cohorts; pooled sensitivity; specificity; area under the curve [AUC]; heterogeneity [I^2, %]; Spearman correlation coefficient [for threshold effect]; and publication bias such as posttest probability positive or negative, positive likelihood ratio [LRP], and negative likelihood ratio [LRN]).

Subgroup and subgroup analysis (studies/datasets)	Pooled sensitivity (95% CI); I^2 (%)	Pooled specificity (95% CI); I^2 (%)	Summary AUC (95% CI)	Spearman correlation coefficient (*P* value)	Publication bias (*P* value)	Posttest probability positive (%)/ negative (%)	LRP	LRN
AI type								
DL[a] (29/32)	0.96 (0.94 - 0.98); 95.87	0.94 (0.91 - 0.95); 97.69	0.98 (0.97 - 0.99)	0.21 (*P*=.05)	.46	94/4	>10	<0.1
ML[b] (4/4)	0.87 (0.78 - 0.93); 32.63	0.88 (0.80 - 0.93); 63.45	0.94 (0.91 - 0.96)	–1 (*P*=.99)	.49	88/13	<10	>0.1
Reference standard								
MRI-PDFF[c] (7/7)	0.92 (0.86 - 0.95); 95.70	0.91 (0.86 - 0.94); 98.43	0.97 (0.95 - 0.98)	–0.36 (*P*=.13)	.15	91/8	>10	<0.1
Pathology (13/14)	0.97 (0.92 - 0.99); 97.86	0.92 (0.86 - 0.95); 85.60	0.98 (0.96 - 0.99)	0.12 (*P*=.02)	.29	92/3	>10	<0.1
Ultrasound (6/6)	0.98 (0.90 - 1.00); 94.32	0.96 (0.94 - 0.98); 85.23	0.98 (0.96 - 0.99)	1 (*P*=.99)	.21	96/2	>10	<0.1
Imaging modality								
Ultrasound (20/22)	0.96 (0.93 - 0.98); 94.86	0.93 (0.90 - 0.96); 96.16	0.98 (0.97 - 0.99)	0.21 (*P*=.4)	.50	94/4	>10	<0.1
CT[d] (8/9)	0.93 (0.86 - 0.96); 94.43	0.93 (0.87 - 0.96); 94.76	0.97 (0.95 - 0.98)	0.20 (*P*=.04)	.24	93/7	>10	<0.1
Pathology (4/4)	0.98 (0.91 - 1.00); 79.53	0.96 (0.86 - 0.99); 0.00	0.99 (0.98 - 0.99)	–1 (*P*=.99)	.00	96/2	>10	<0.1
TL[e]								
Used (9/9)	0.99 (0.96 - 1.00); 95.36	0.93 (0.88 - 0.97); 93.80	0.99 (0.98 - 1.00)	0.2 (*P*=.04)	.77	94/1	>10	<0.1
Not used (24/27)	0.93 (0.90 - 0.96); 84.22	0.93 (0.90 - 0.95); 96.71	0.98 (0.96 - 0.99)	0.22 (*P*=.05)	.53	93/7	>10	<0.1
Study design								
Single-center (25/27)	0.94 (0.91 - 0.96); 94.44	0.93 (0.91 - 0.95); 97.33	0.98 (0.96 - 0.99)	0.25 (*P*=.06)	.98	93/6	>10	<0.1
Multicenter (8/9)	0.99 (0.94 - 1.00); 95.26	0.92 (0.85 - 0.96); 82.33	0.97 (0.96 - 0.99)	0.47 (*P*=.22)	.30	92/1	>10	<0.1
Study type								
Retrospective (25/26)	0.95 (0.92 - 0.97); 96.58	0.95 (0.92 - 0.97); 98.15	0.98 (0.97 - 0.99)	0.39 (*P*=.15)	.53	95/5	>10	<0.1

XSL•FO
RenderX

Subgroup and subgroup analysis (studies/datasets)	Pooled sensitivity (95% CI); I^2 (%)	Pooled specificity (95% CI); I^2 (%)	Summary AUC (95% CI)	Spearman correlation coefficient (*P* value)	Publication bias (*P* value)	Posttest probability positive (%)/ negative (%)	LRP	LRN
Prospective (8/10)	0.97 (0.92 - 0.99); 82.76	0.87 (0.84 - 0.89); 53.89	0.90 (0.87 - 0.92)	1 (*P*=.99)	.87	88/4	<10	<0.1
Data availability								
Available (9/10)	0.99 (0.96 - 1.00); 97.06	0.95 (0.92 - 0.97); 73.17	0.99 (0.97 - 0.99)	−0.5 (*P*=.25)	.19	96/1	>10	<0.1
Unavailable (24/26)	0.92 (0.89 - 0.95); 81.89	0.92 (0.89 - 0.94); 96.62	0.97 (0.95 - 0.98)	0.09 (*P*=.01)	.30	92/8	>10	<0.1

[a]DL: deep learning.

[b]ML: machine learning.

[c]MRI-PDFF: magnetic resonance imaging–proton density fat fraction.

[d]CT: computed tomography.

[e]TL: transfer learning

Diagnostic Performance and Heterogeneity

Of the 36 included studies, 33 (comprising 36 datasets) satisfied the criteria for subgroup analysis. The pooled results (Figure 2) demonstrated a summary sensitivity of 0.96 (95% CI 0.93 - 0.97), a specificity of 0.93 (95% CI 0.91 - 0.95), and an AUC of 0.98 (95% CI 0.96 - 0.99), indicating excellent diagnostic discrimination by the AI models. Substantial heterogeneity was observed across studies (I^2>75%). The Spearman correlation coefficient (0.21, *P*=.05) suggested that threshold effects contributed minimally to overall heterogeneity. The broad 95% PI, however, indicated that differences in diagnostic thresholds were a major source of variability. No significant small-study effects were identified (*P*=.65). In terms of clinical applicability, at a pretest probability of 50%, a positive AI result increased the posttest probability to 93%, whereas a negative result reduced it to 4%. Likelihood ratio scattergram analysis confirmed that the pooled estimates were located within the “confirm and exclude” quadrant (LRP >10 and LRN <0.1), underscoring the strong clinical value of AI for both confirming and excluding HS.

XSL•FO
RenderX

Figure 2. Diagnostic performance of artificial intelligence (AI) models for hepatic steatosis (HS) detection across 33 studies comprising 36 datasets [28-35,37,39-52,54-63]. (A) Forest plots illustrating sensitivity and specificity for the subgroup of AI applications across 33 studies with 36 datasets. (B) Summary receiver operating characteristic (SROC) curve depicting diagnostic performance of AI across 33 studies with 36 datasets, with corresponding 95% CIs. The 95% prediction region reflects the expected range of true sensitivity and specificity in future studies. (C) Bivariate boxplot illustrating the distribution and heterogeneity of AI performance across 33 studies with 36 datasets. (D) The Deeks funnel plot for evaluation of potential publication bias. (E) The Fagan nomogram depicting posttest probabilities. (F) Clinical application plot showing positive likelihood ratio (LRP) and negative likelihood ratio (LRN). LLQ: lower-left quadrant; LUQ: upper-left quadrant; RLQ: lower-right quadrant; RUQ: upper-right quadrant; SROC: summary receiver operating characteristic.

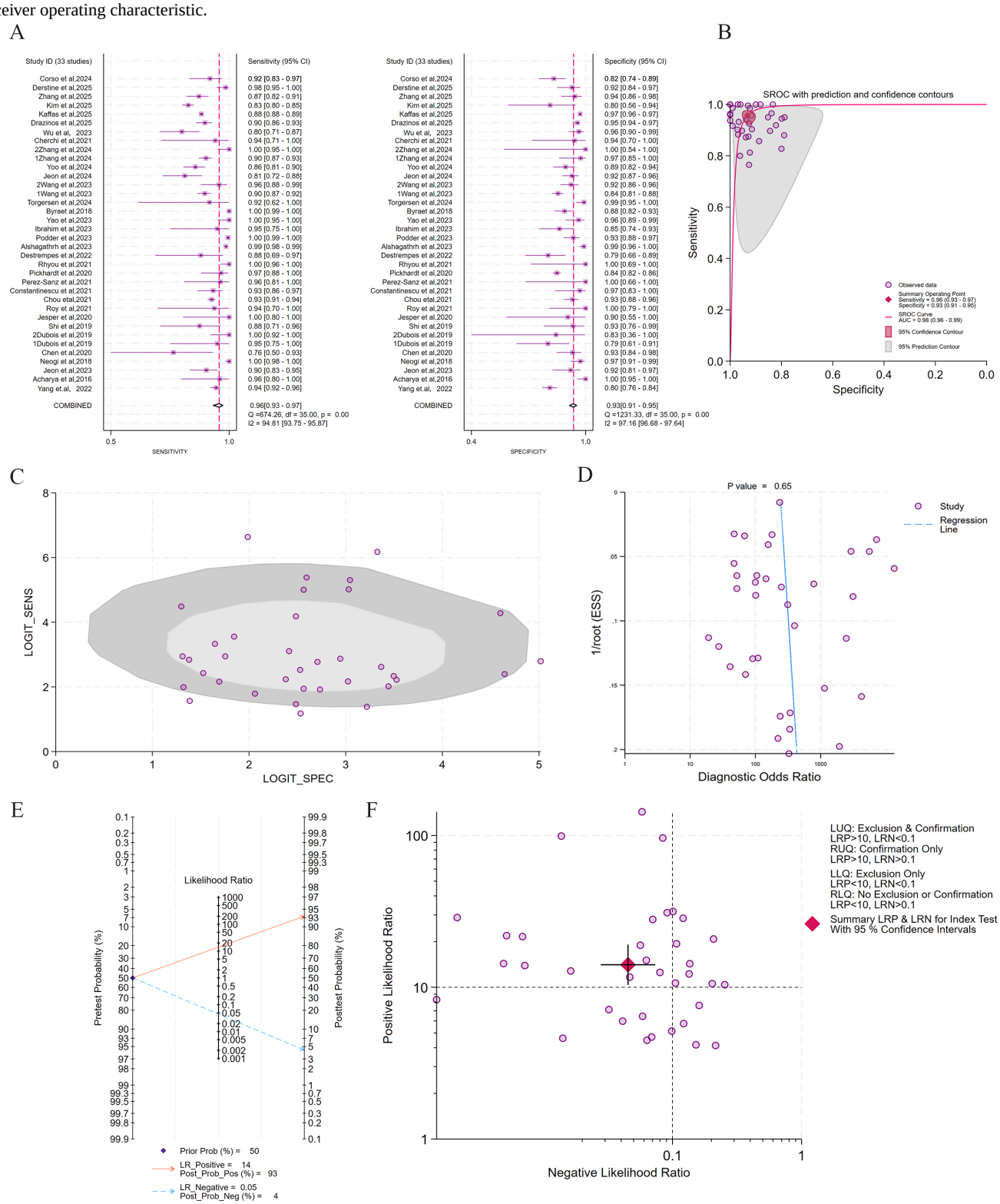

Risk of Bias Assessment

The QUADAS-2 quality assessment (Figure 3) revealed that 44% (16/36) of studies exhibited a high risk of bias in the patient selection domain, primarily due to selection bias, limited representativeness of study populations, and incomplete reporting of key clinical parameters. In the Index test domain, 6% (2/36) of studies were rated as high risk, largely attributable

XSL•FO
RenderX

to the absence of image quality control, subjective elements during image processing, and nonstandardized training or validation procedures. In the reference standard domain, 17% (6/36) of studies demonstrated a high risk of bias, most commonly due to deviations from gold-standard reference methods, unclear blinding procedures, or incomplete pathological sampling information. The flow and timing domain exhibited unclear risk in 36% (14/36) of studies, often due to insufficient reporting on patient inclusion pathways and the interval between image acquisition and diagnostic confirmation. These methodological limitations may contribute to an overestimation of AI model performance in real-world clinical practice.

Figure 3. Risk of bias assessment of the 36 included studies on artificial intelligence–based hepatic steatosis diagnosis using the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) tool [28-63].

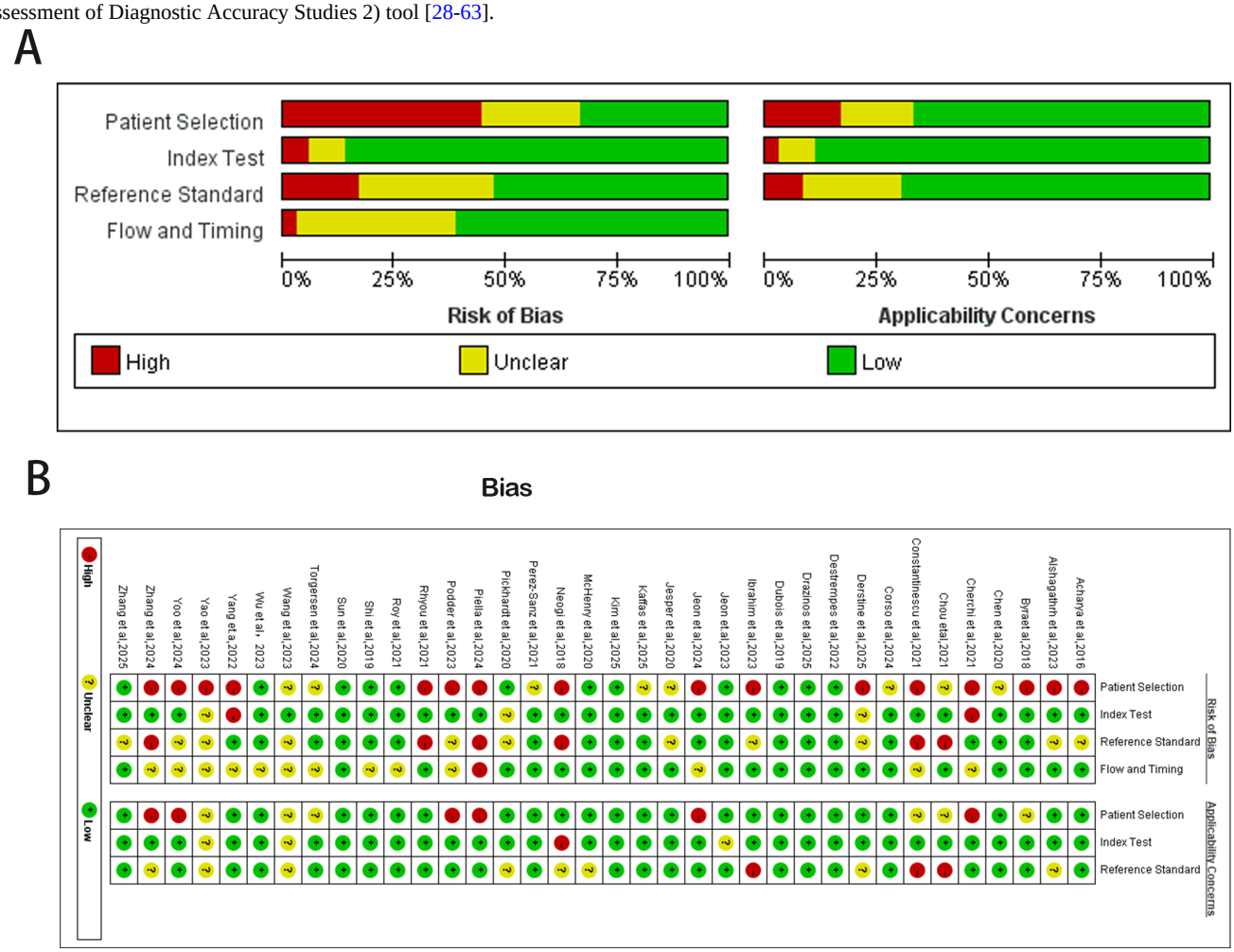

Subgroup Meta-Analyses

Subgroup analyses were conducted for 7 key variables (Figures S1-S16 in Multimedia Appendix 1), with diagnostic performance interpreted relative to clinical applicability thresholds (LRP >10 for strong rule-in capability; LRN <0.1 for strong rule-out capability).

Algorithm Type

As shown in Figure 4 A-B, the DL models demonstrated significantly higher diagnostic accuracy than ML models (AUC: 0.98 vs 0.94), exhibiting strong rule-in and rule-out performance. However, DL models displayed pronounced heterogeneity (I^2 >95%), likely influenced by threshold effects (Spearman=0.21, P=.05), suggesting that these findings should be generalized with caution. ML models showed lower heterogeneity (sensitivity I^2=32.63%; specificity I^2=63.45%) but weaker discriminatory power (LRP <10, LRN >0.1).

XSL•FO
RenderX

Figure 4. Diagnostic performance stratified by algorithm type for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for deep learning (DL) algorithms in 29 studies [28-33,37,39-43,45-51,54-63] comprising 32 datasets; (B) SROC curve for machine learning (ML) algorithms in 4 studies [34,35,44,52] comprising 4 datasets.

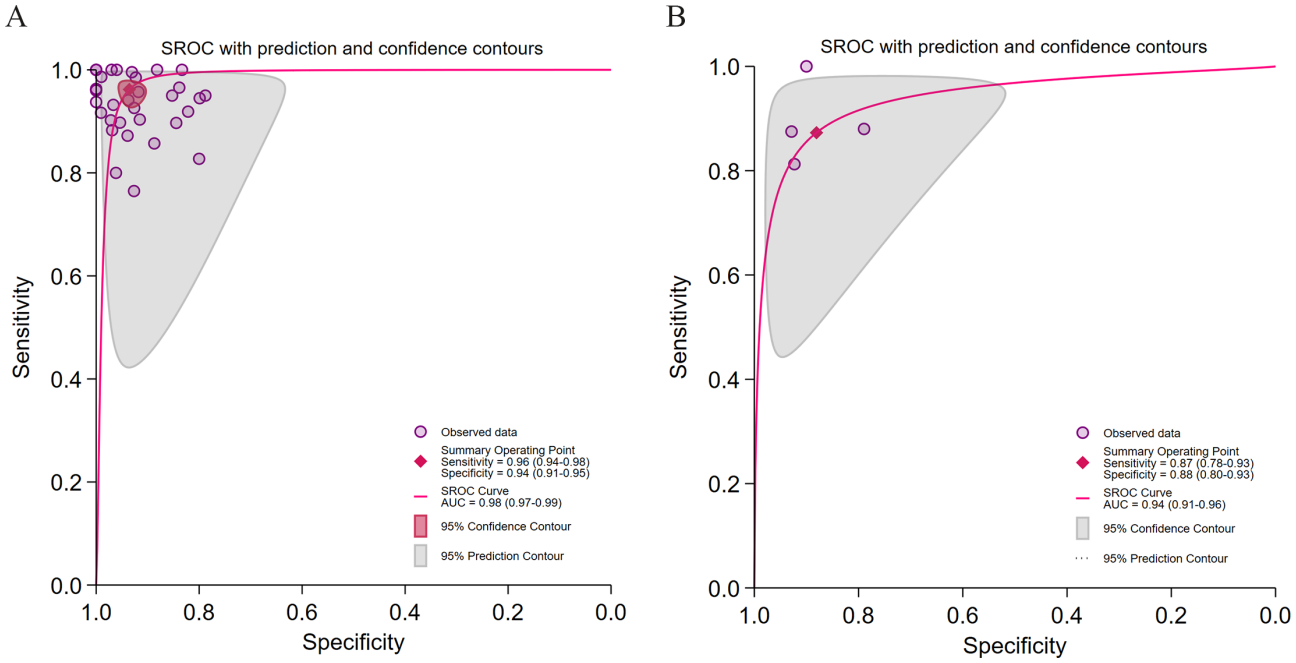

Reference Standard

As shown in Figure 5 A-C, studies using ultrasound and histopathology as reference standards achieved comparable AUCs (both were 0.98) with ideal likelihood ratios, although ultrasound-based models exhibited higher pooled sensitivity and specificity (0.98 and 0.96, respectively) than histopathology-based models (0.97 and 0.92). The ultrasound subgroup showed a perfect threshold effect (Spearman=1; *P*=.99), indicating well-defined diagnostic criteria that may be subjectively constrained. The histopathology subgroup exhibited a minimal threshold effect (Spearman=0.12; *P*=.02), suggesting that interstudy variations in sample handling and scoring could significantly influence model performance. The MRI-PDFF subgroup achieved a comparable AUC (0.97) but demonstrated very high heterogeneity (I^2 >95%), limiting result stability.

Figure 5. Diagnostic performance stratified by reference standard for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for magnetic resonance imaging–proton density fat fraction (MRI-PDFF) in 7 studies [30,42,52,54,59,62,63] comprising 7 datasets; (B) SROC curve for pathology in 13 studies [28,32,33,35,41,43-46,49,56,57,60] comprising 14 datasets; (C) SROC curve for ultrasound in 6 studies [29,31,39,40,48,58] comprising 6 datasets.

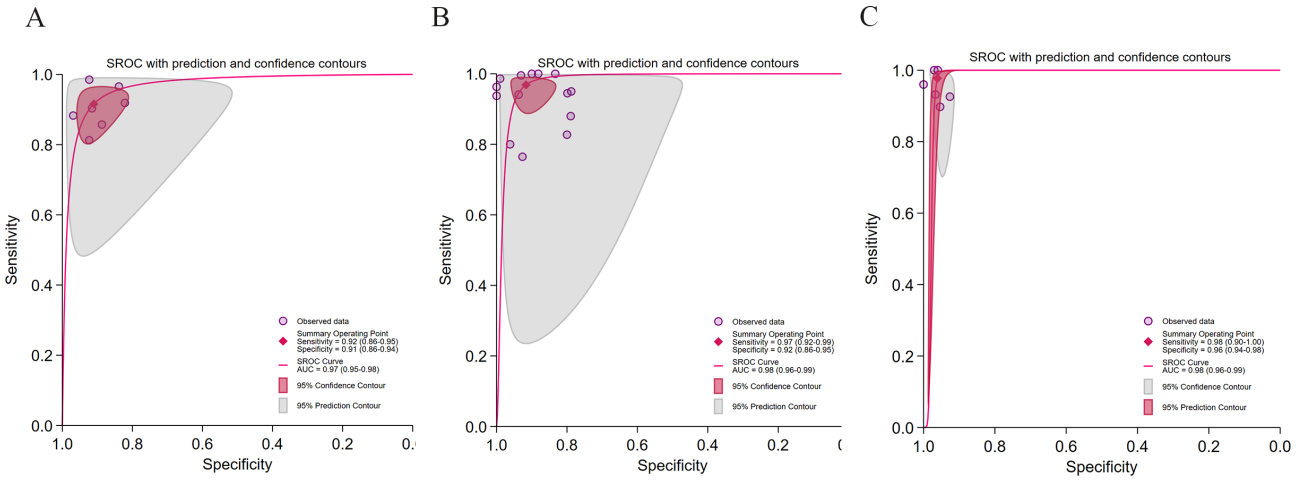

Imaging Modality

As shown in Figure 6 A-C, histopathology-based models achieved the highest diagnostic performance (AUC=0.99) with no detectable heterogeneity, suggesting robust and consistent results. However, significant publication bias was identified (*P*<.001), implying potential preferential publication of high-performing studies. AI models based on ultrasound and CT achieved comparable accuracy (AUC: 0.98 vs 0.97), though both exhibited marked heterogeneity (I^2 >94%). Only the CT subgroup showed a negligible threshold effect (Spearman=0.20; *P*=.04).

XSL•FO
RenderX

Figure 6. Diagnostic performance stratified by imaging modality for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for ultrasound imaging in 20 studies [29-35,39,40,43-45,47-49,51,57-59,63] comprising 22 datasets; (B) SROC curve for computed tomography (CT) imaging in 8 studies [42,50,52,54,55,60-62] comprising 9 datasets; (C) SROC curve for pathology imaging in 4 studies [37,41,46,56] comprising 4 datasets.

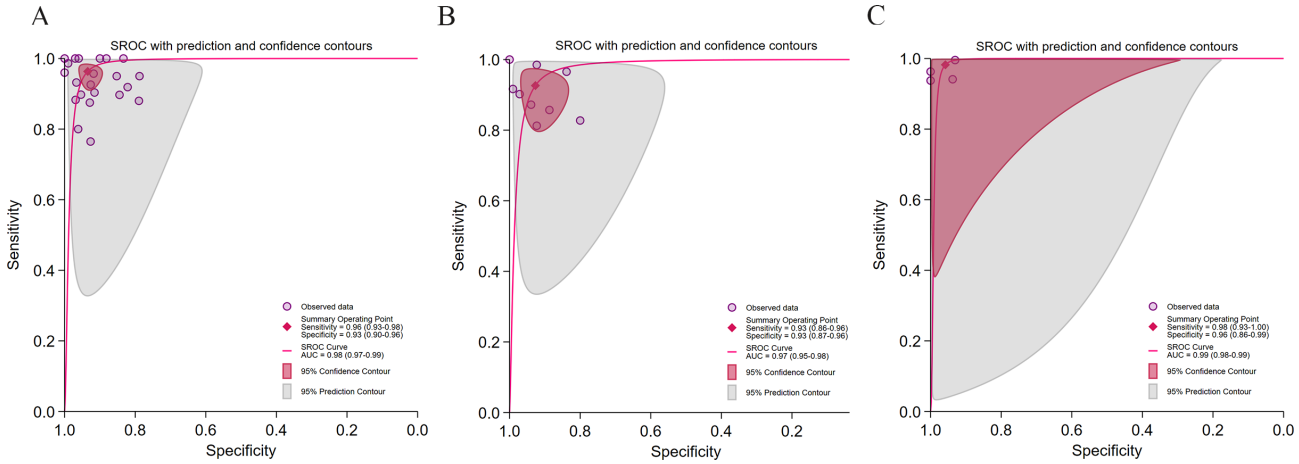

Application of TL

As shown in the figure Figure 7A-B models employing TL achieved higher pooled sensitivity (0.99 vs 0.93) and stronger rule-out capability (LRN: 0.01 vs 0.07). No significant publication bias was detected in either subgroup (P >.05). Nevertheless, both demonstrated considerable heterogeneity (I^2 >84%) and mild threshold effects (Spearman=0.20 vs 0.22; P=.04 vs .05), reflecting the influence of interdomain data discrepancies.

Figure 7. Diagnostic performance of transfer learning (TL) for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for studies employing TL in 9 studies [28,39,40,43,45-49] comprising 9 datasets; (B) SROC curve for studies not employing TL in 24 studies [29-35,37,41,42,44,50-52,54-63] comprising 26 datasets.

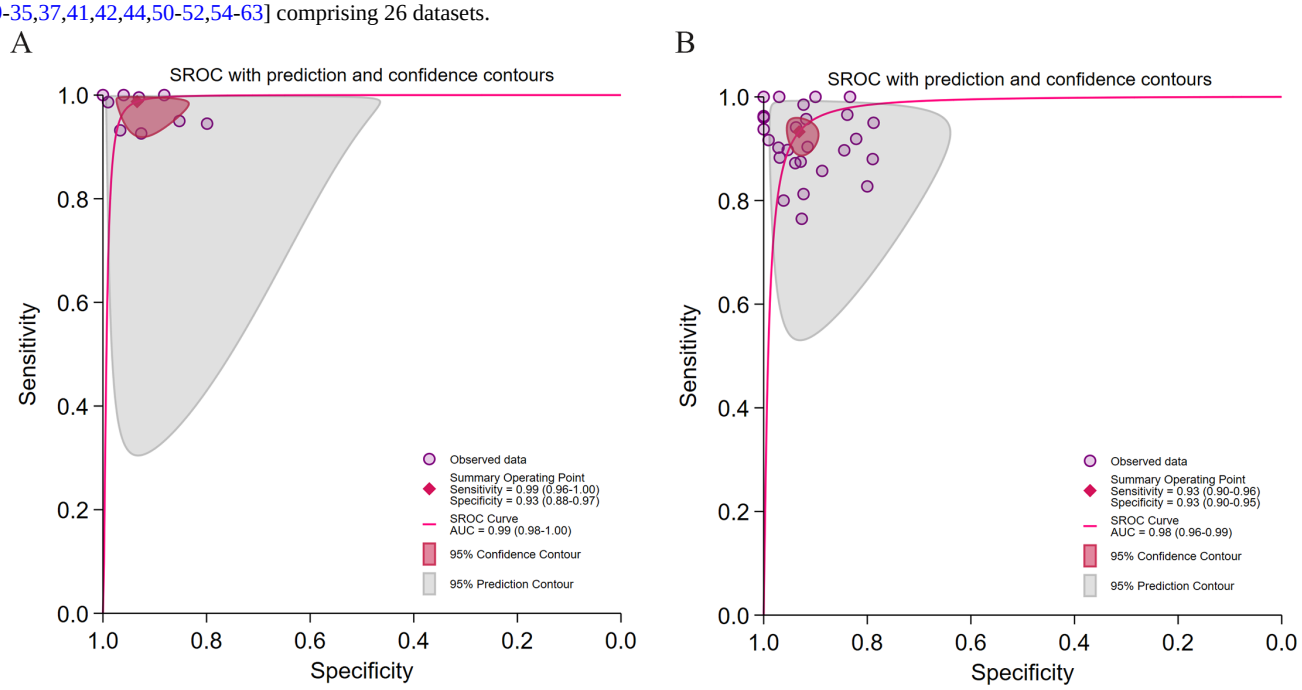

Study Design

As shown in Figure 8A-B, multicenter studies demonstrated superior sensitivity (0.99 vs 0.94) and lower heterogeneity (I^2=82.33%), indicating greater generalizability and stronger rule-out potential (LRN: 0.01 vs 0.06). In contrast, single-center studies exhibited marginally higher specificity (0.93 vs 0.92) but very high heterogeneity (I^2 >94%), suggesting limited external validity.

XSL•FO
RenderX

Figure 8. Diagnostic performance of different research designs for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for single-center studies in 25 studies [28-35,39-42,45-47,50-52,54,56-61] with 26 datasets; (B) SROC curve for multicenter studies in 8 studies [37,43,44,48,49,55,62,63] with 9 datasets.

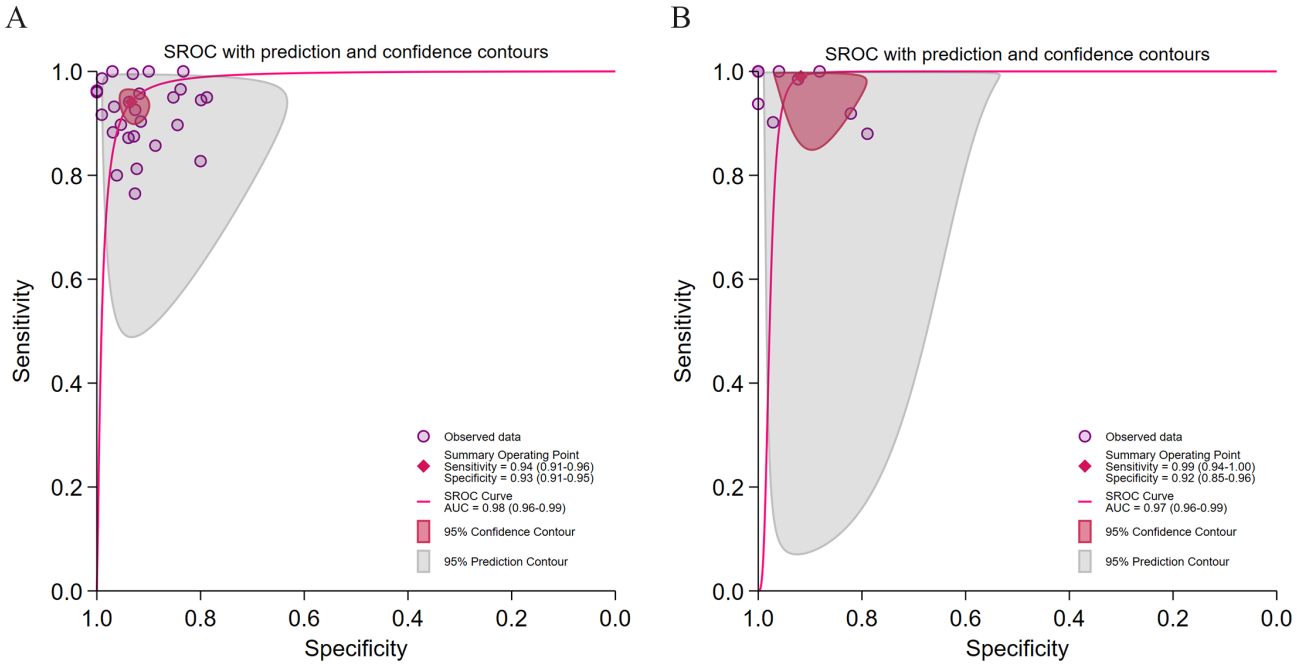

Study Type

As shown in Figure 9 A-B, retrospective studies achieved higher overall accuracy (AUC: 0.98 versus 0.90) and stronger rule-in ability (LRP: 9.5 versus 8.8), though with significant heterogeneity (I^2 >96%). Prospective studies, which can better reflect clinical reality, were affected by a perfect threshold effect (Spearman=1) and exhibited weaker rule-in performance (LRP<10).

Figure 9. Diagnostic performance of different research types for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for retrospective studies in 25 [28,29,31,32,34,37,41-46,48,50,52,54-62] studies with 26 datasets; (B) SROC curve for prospective studies in 8 studies with 9 datasets. AUC: area under curve.

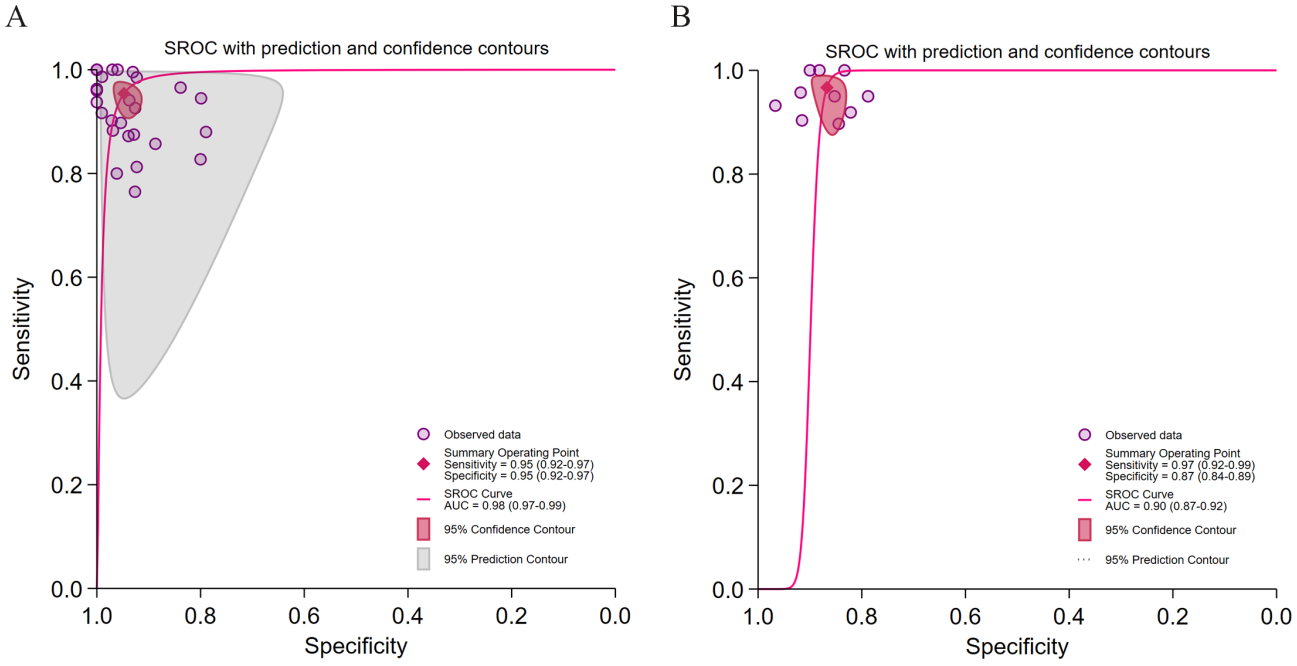

Data Accessibility

As shown in Figure 10A-B studies using publicly available datasets (n=9) achieved superior diagnostic accuracy (AUC=0.99) and stronger clinical applicability (LRP=9.6; LRN=0.01). In contrast, studies using nonpublic data performed comparably (AUC=0.97) but showed a significant threshold effect (Spearman=0.09; *P*=.01), indicating reduced result stability.

XSL•FO
RenderX

Figure 10. Diagnostic performance of data availability for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for studies with available data in 9 studies [37,43,45,46,48,49,55,61,62] with 10 datasets; (B) SROC curve for studies with unavailable data in 24 studies [28-35,38-42,44,47,50-52,54,56-60,63] with 25 datasets.

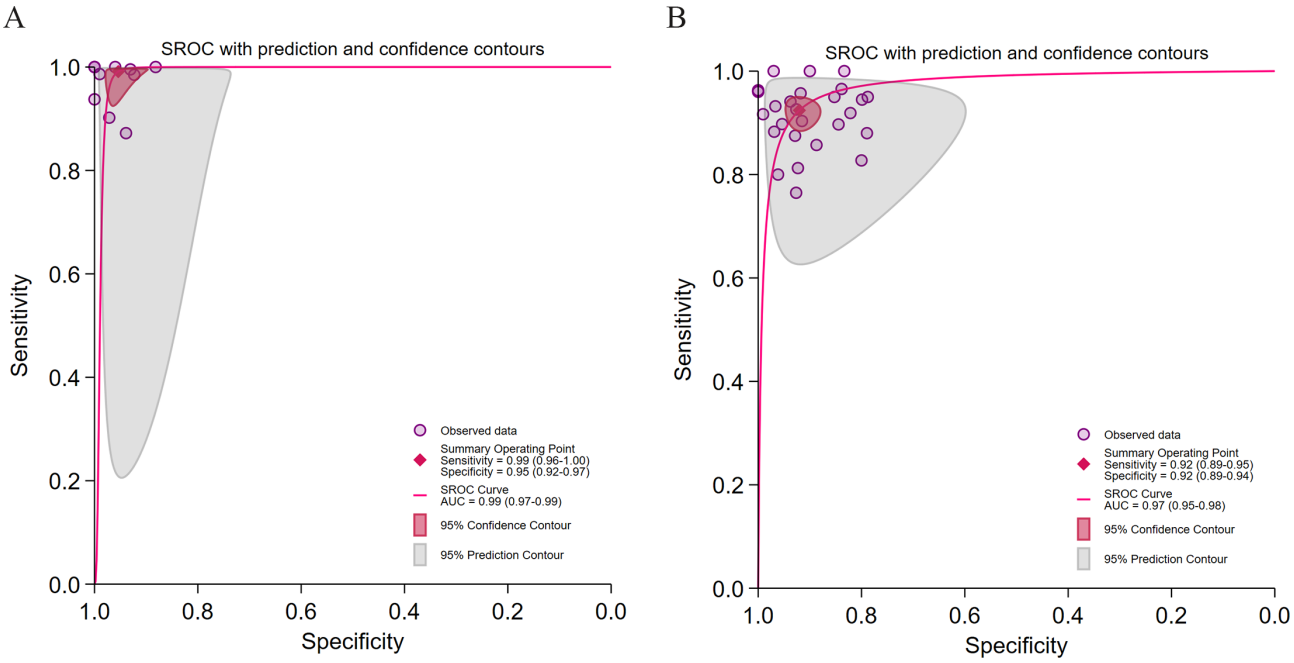

Subgroup definitions are detailed in the methods section. Heterogeneity was categorized as follows: I^2<50% low heterogeneity, 50% - 75% moderate heterogeneity, and >75% high heterogeneity.

Discussion

Principal Findings

This meta-analysis of 36 studies demonstrates the superior diagnostic performance of AI in identifying HS, yielding a pooled AUC of 0.98, surpassing that of conventional ultrasound, CT, and MRI, whose pooled AUCs were 0.93, 0.975, and 0.97, respectively [64-66]. These findings underscore AI's potential to overcome the inherent physical constraints of individual imaging techniques, thereby establishing it as a versatile and adaptive diagnostic approach. This capability carries considerable clinical significance, providing a strong rationale for further meta-analyses dedicated to AI-based diagnostic technologies and informing the development of flexible, context-specific clinical applications. By optimizing the use of the most accessible and cost-effective diagnostic resources, AI could markedly broaden the availability of early HS screening across diverse health care settings.

Although the encouraging performance of AI in diagnosing HS is promising, interpretation of these findings must be tempered by a critical appraisal of the methodological limitations underlying this research. Our analysis revealed substantial heterogeneity (I^2>75%) and a high overall risk of bias among included studies, particularly within the patient selection domain, where 44% (16/36) were judged to be at high risk. A major limitation lies in the predominance of retrospective, single-center designs (25/36, 69%). Such studies typically develop and validate models within controlled, idealized data environments, meaning that reported metrics may reflect "best-case scenarios" rather than true clinical performance across diverse devices, operators, and patient populations in routine practice. Moreover, independent external validation and multicenter prospective trials remain notably scarce, severely limiting the assessment of these models' generalizability. Therefore, while existing evidence underscores AI's considerable potential in HS diagnosis, the current body of research remains insufficient to justify its widespread clinical adoption. Bridging the translational gap from high-performing algorithms to reliable, universally applicable clinical tools thus remains a substantial challenge.

Subgroup analyses provide valuable insights for optimizing AI model design and informing clinical integration. DL-based models demonstrate exceptionally high specificity in diagnosing HS, offering a distinct clinical advantage by reducing unnecessary liver biopsies. However, these models require higher-quality data annotation and greater computational resources. Model performance was also closely associated with the reference standard and imaging modality used. Systems using histopathological images as input achieved the highest diagnostic accuracy. Nevertheless, their clinical applicability is restricted by procedural invasiveness and sampling error [67]. The AI-assisted whole-slide analysis model proposed by Roy et al [37] improves quantitative consistency but faces practical barriers related to cost and patient acceptance.

The choice of imaging modality inherently involves trade-offs between diagnostic accuracy, accessibility, and cost. MRI-PDFF, while providing precise, noninvasive quantification, is affected by confounders such as iron overload, edema, and concurrent pathologies [7], and its high cost limits use in primary care. Ultrasound remains the most accessible and economical option but suffers from operator dependency, reduced sensitivity for mild steatosis, limited penetration in obese individuals [68,69], and suboptimal accuracy in detecting fibrosis [70]. AI integration could mitigate these limitations by standardizing

XSL•FO
RenderX

acquisition and interpretation, though at the cost of increased system complexity and computational demand. CT achieves a sensitivity and specificity of 0.93 but is constrained by ionizing radiation exposure and potential interference from iodine-based contrast agents [71]. The dual-energy CT 3D nnU-Net model developed by Yoo et al [54] achieved an AUC of 0.97 for distinguishing steatotic from normal tissue, yet its clinical application is constrained by limited equipment availability.

To optimize model performance and data use, TL has been widely adopted, an especially valuable strategy given the substantial costs associated with medical data annotation. Nonetheless, the effectiveness of TL depends critically on the degree of similarity between the source and target domains; substantial domain discrepancies may lead to "negative transfer," as illustrated by sensitivity variations of up to 10% in the Inception-v3 model reported by Constantinescu et al [40]. To address this limitation, emerging approaches such as adversarial domain adaptation frameworks have achieved near-human classification accuracy on heterogeneous MRI datasets [72]. Similarly, hybrid pretraining strategies [73] and federated learning techniques have reached up to 99% of the performance attained through centralized training [74]. These approaches enhance model robustness while effectively addressing data privacy and heterogeneity.

Beyond algorithmic optimization, the real-world implementation of AI is profoundly influenced by study design and data governance. Retrospective studies, which constituted the majority (25/36, 69%) of the included reports, demonstrated significantly higher performance than prospective studies (AUC: 0.98 vs 0.94), likely reflecting the high-quality and well-curated imaging data typically available in retrospective cohorts. In contrast, prospective designs more faithfully capture real-world clinical workflows but are inherently subject to operational variability, such as inconsistent imaging protocols and unpredictable patient factors, thereby leading to attenuated performance.

Furthermore, data governance and accessibility are pivotal in determining model generalizability. Multicenter collaborations and data sharing can improve generalizability and reproducibility, though they require standardized imaging protocols, increased logistical coordination, and greater resource investment, posing feasibility challenges in resource-limited settings. Moreover, access to medical data for AI development remains hindered by privacy regulations, institutional policies, and technical interoperability barriers. Privacy-preserving strategies, such as federated learning, offer promising solutions by enabling multi-institutional collaboration without direct data exchange, albeit at the cost of increased computational demands and system complexity. It should also be noted that publicly available datasets may not fully represent the clinical heterogeneity encountered in real-world practice, thereby introducing potential selection bias. These factors, while critical for improving AI performance, also contribute substantially to heterogeneity, underscoring the necessity of comprehensive external validation and context-specific adaptation before large-scale clinical implementation.

Expanding Role in HS Management

The use of AI extends beyond diagnostic precision to encompass the comprehensive management of HS. Accumulating evidence indicates that AI not only enables accurate quantification of hepatic fat but also integrates radiomic, pathological, and clinical data to facilitate fibrosis staging, predict HCC risk, assess posttransplant survival, and stratify cardiovascular complications. For instance, a VGG16-based ultrasound model outperformed human interpretation in classifying borderline cases [75]. The integration of macrogenomic sequencing with ML has proven effective for the differential diagnosis of HS in obese pediatric populations [76]. Similarly, an ML model based on MRI-derived liver fat quantification markedly improved diagnostic accuracy for liver fibrosis [77]. AI-powered digital pathology platforms reduce the inherent subjectivity of conventional histological assessment [78], while DL-based radiomics facilitates the identification of critical pathological features such as microvascular invasion [79]. A DL algorithm demonstrated 99% accuracy in predicting postliver transplantation survival [80]. In the context of MAFLD-related complications, AI algorithms have been employed to accurately identify affected patients from electronic health records, revealing type 2 diabetes mellitus as a significant predictor of all-cause mortality (hazard ratio: 1.36) [81]. Moreover, a dual model combining tongue imaging with clinical indicators achieved precise prediction of coronary heart disease risk among patients with fatty liver [82]. The foregoing advances signal a diagnostic paradigm shift in HS management from a traditional "liver-centric" approach towards a "patient-centric" model of multi-system risk management, paving the way for early intervention and personalized therapy.

In summary, the advantages of AI in HS diagnosis are threefold as follows:

1. Enhanced early detection: DL models can detect subclinical pathological alterations, including hepatic fat infiltration below 5%, thereby reducing diagnostic subjectivity and improving reproducibility [43,45,83].
2. Standardized quantitative analysis: End-to-end, pixel-level segmentation enables automated calculation of HS, minimizing reliance on manual interpretation and potentially substituting for histopathological assessment in resource-constrained settings.
3. Longitudinal predictive modeling: The integration of time-series radiomic and metabolomic features facilitates the construction of individualized models predicting cirrhosis progression and MAFLD onset within 3 years, providing actionable insights for precision treatment planning.

Challenges and a Phased Implementation Framework

Despite its promising outlook, the widespread clinical adoption of AI in HS management faces multiple challenges. Technically, data heterogeneity, stemming from variations in imaging quality [84], scanner types, and reference standard thresholds, impedes the development of universally robust and generalizable models. Many high-performing algorithms are derived from single-center, retrospective datasets (eg, Yang et al [22], n=50, Beijing) with limited demographic diversity, thereby

XSL•FO
RenderX

compromising their external validity and real-world applicability. Moreover, most existing models primarily focus on imaging biomarkers for fat quantification without adequately elucidating the complex pathophysiological interplay among steatosis, metabolic comorbidities, and fibrosis, limiting both clinical interpretability and holistic disease assessment.

From a clinical integration perspective, the transition from algorithmic development to real-world deployment necessitates careful consideration of workflow compatibility, device dependency, and cost-effectiveness. Lightweight AI models hold promise for incorporation into primary care ultrasound systems, facilitating large-scale population screening, whereas more advanced MRI- or CT-based models may be more appropriately implemented in tertiary medical centers. The overarching objective is seamless integration into existing clinical workflows, ensuring that AI serves as an assistive, rather than disruptive, technology that streamlines radiological practice, conserves clinician time, and enhances diagnostic efficiency [85]. Furthermore, issues concerning data privacy [86], algorithmic bias [87], and accountability [88] lack clear regulatory frameworks.

From a global health perspective, the clinical use of AI in HS diagnosis varies according to resource availability. To promote both efficiency and equity in HS diagnosis and management, a phased implementation framework is proposed:

1. Tiered deployment in specific scenarios: in resource-limited settings, lightweight AI systems can be paired with portable ultrasound to enable cost-effective community screening and early detection. Suspected cases may then be referred to higher-level hospitals for precise stratified diagnosis (eg, MRI-PDFF), thereby optimizing resource allocation and minimizing unnecessary biopsies. In high-resource environments, AI-driven automated image processing facilitates accurate fat quantification and disease staging, forming a synergistic diagnostic–therapeutic feedback loop.
2. Establish cross-institutional collaborative data platforms: the adoption of federated learning and related technologies can enhance data diversity while ensuring privacy protection. Such approaches enable robust model development based on heterogeneous real-world data, mitigate model bias and validation gaps, eliminate the need for centralized storage of sensitive information, and provide the foundation for scalable, privacy-preserving deployment.
3. Transition from standalone tools to integrated management platforms: the ultimate objective is to advance AI from a single-function diagnostic aid to a comprehensive, multi-task management system. By synchronously quantifying steatosis, assessing fibrosis, and evaluating inflammatory markers through multimodal data integration. Incorporating these outputs directly into clinical decision-making workflows, AI could evolve from diagnostic assistance to intelligent, holistic disease management.

Limitations in the Literature

Several limitations warrant cautious interpretation. First, considerable methodological and clinical heterogeneity was observed across the included studies, constraining the reliability of the conclusions. Despite extensive subgroup analyses, variability arising from differences in patient characteristics, imaging equipment, and diagnostic thresholds could not be fully addressed. This residual heterogeneity undermines the robustness of pooled estimates and suggests the influence of unmeasured factors affecting AI performance.

Second, the analysis was limited by methodological shortcomings inherent in the primary studies. Inadequate reporting of key patient characteristics hindered subgroup analyses by disease etiology, particularly distinguishing pure MAFLD from mixed forms, a critical gap given the potential impact of comorbidities on diagnostic accuracy. Furthermore, wide variation in AI architectures and the limited number of comparable models precluded meaningful comparisons across technical approaches, leaving the effect of architectural design on diagnostic performance unclear.

Third, the generalizability and real-world applicability of the findings remain limited. Most studies were retrospective, single-center designs prone to selection bias, with scarce external or temporal validation. Thus, the high-performance metrics reported may represent an idealized best-case scenario rather than outcomes achievable in prospective clinical settings.

Additionally, although our restriction to peer-reviewed full-text publications ensured a baseline level of methodological rigor, the exclusion of relevant preprints and gray literature may have introduced publication bias. Such selective inclusion likely favored studies reporting positive outcomes, potentially leading to overestimated performance measures. Moreover, key practical factors, such as computational burden, workflow integration, and technical expertise, could not be quantitatively evaluated, despite their importance for real-world implementation.

Conclusions

This meta-analysis highlights the substantial diagnostic potential of AI, particularly DL, in assessing HS. Its key contribution lies in establishing a unified, imaging-modality-independent analytical framework that provides comprehensive evidence beyond the constraints of individual imaging techniques. Nonetheless, these results reflect technical promise rather than confirmed clinical use. The translation from high-performing algorithms to reliable clinical tools remains hindered by performance heterogeneity, retrospective study designs, and insufficient external validation. While the technological foundation of AI in HS is encouraging, clinical maturity has yet to be achieved. Bridging this translational gap will require prospective multicenter studies, standardized reporting protocols, and rigorous external validation. Ultimately, successful clinical adoption will depend on demonstrating not only algorithmic robustness but also tangible improvements in patient outcomes and workflow efficiency across real-world health care settings.

XSL•FO
RenderX

Funding

This work was supported by the Science and Technology Research Program of Jilin Provincial Department of Education (No. JKH20250664KJ). The sponsor participated in the preliminary design phase and guided the selection of research methodologies. However, the sponsor did not participate in subsequent data collection, analysis, interpretation of results, or the preparation of this manuscript.

Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: JS
Data curation: JS and JL
Formal analysis: YL
Funding acquisition: DL
Investigation: JL and HC
Methodology: DL
Project administration: JZ
Resources: JZ
Software: JS
Supervision: DL and RD
Validation: JS
Visualization: HC
Writing – original draft: JS
Writing – review & editing: JZ
All authors commented on a previous version of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Retrieval strategy, and Forest plots, bivariate boxplots, Deeks' funnel plots, Fagan's nomogram plots, and clinical application plots for each subgroup analysis.
[DOCX File, 9782 KB - jmir_v28i1e78310_app1.docx]

Checklist 1
PRISMA-DTA checklist.
[PDF File, 138 KB - jmir_v28i1e78310_app2.pdf]

Checklist 2
QUADAS-2 checklist.
[XLSX File, 31 KB - jmir_v28i1e78310_app3.xlsx]

References

1. Unalp-Arida A, Ruhl CE. Prevalence of metabolic dysfunction-associated steatotic liver disease and fibrosis defined by liver elastography in the United States using national health and nutrition examination survey 2017-March 2020 and August 2021-August 2023 data. Hepatology 2025 Nov 1;82(5):1256-1273. [doi: 10.1097/HEP.0000000000001211] [Medline: 39724144]
2. Riazi K, Azhari H, Charette JH, et al. The prevalence and incidence of NAFLD worldwide: a systematic review and meta-analysis. Lancet Gastroenterol Hepatol 2022 Sep;7(9):851-861. [doi: 10.1016/S2468-1253(22)00165-0] [Medline: 35798021]
3. Eslam M, Newsome PN, Sarin SK, et al. A new definition for metabolic dysfunction-associated fatty liver disease: an international expert consensus statement. J Hepatol 2020 Jul;73(1):202-209. [doi: 10.1016/j.jhep.2020.03.039] [Medline: 32278004]
4. Gofton C, Upendran Y, Zheng MH, George J. MAFLD: how is it different from NAFLD? Clin Mol Hepatol 2023 Feb;29(Suppl):S17-S31. [doi: 10.3350/cmh.2022.0367] [Medline: 36443926]

XSL•FO
RenderX

5. Younossi ZM, Golabi P, Paik JM, Henry A, Van Dongen C, Henry L. The global epidemiology of nonalcoholic fatty liver disease (NAFLD) and nonalcoholic steatohepatitis (NASH): a systematic review. Hepatology 2023 Apr 1;77(4):1335-1347. [doi: 10.1097/HEP.0000000000000004] [Medline: 36626630]
6. Younossi ZM, Kalligeros M, Henry L. Epidemiology of metabolic dysfunction-associated steatotic liver disease. Clin Mol Hepatol 2025 Feb;31(Suppl):S32-S50. [doi: 10.3350/cmh.2024.0431] [Medline: 39159948]
7. Wong VWS, Ekstedt M, Wong GLH, Hagström H. Changing epidemiology, global trends and implications for outcomes of NAFLD. J Hepatol 2023 Sep;79(3):842-852. [doi: 10.1016/j.jhep.2023.04.036] [Medline: 37169151]
8. Starekova J, Hernando D, Pickhardt PJ, Reeder SB. Quantification of liver fat content with CT and MRI: state of the art. Radiology 2021 Nov;301(2):250-262. [doi: 10.1148/radiol.2021204288] [Medline: 34546125]
9. Leporq B, Ratiney H, Pilleul F, Beuf O. Liver fat volume fraction quantification with fat and water T1 and T 2* estimation and accounting for NMR multiple components in patients with chronic liver disease at 1.5 and 3.0 T. Eur Radiol 2013 Aug;23(8):2175-2186. [doi: 10.1007/s00330-013-2826-x] [Medline: 23588583]
10. Bedogni G, Bellentani S, Miglioli L, et al. The fatty liver index: a simple and accurate predictor of hepatic steatosis in the general population. BMC Gastroenterol 2006 Nov 2;6:33. [doi: 10.1186/1471-230X-6-33] [Medline: 17081293]
11. Lee JH, Kim D, Kim HJ, et al. Hepatic steatosis index: a simple screening tool reflecting nonalcoholic fatty liver disease. Dig Liver Dis 2010 Jul;42(7):503-508. [doi: 10.1016/j.dld.2009.08.002] [Medline: 19766548]
12. Kotronen A, Peltonen M, Hakkarainen A, et al. Prediction of non-alcoholic fatty liver disease and liver fat using metabolic and genetic factors. Gastroenterol 2009 Sep;137(3):865-872. [doi: 10.1053/j.gastro.2009.06.005] [Medline: 19524579]
13. Dasarathy S, Dasarathy J, Khiyami A, Joseph R, Lopez R, McCullough AJ. Validity of real time ultrasound in the diagnosis of hepatic steatosis: a prospective study. J Hepatol 2009 Dec;51(6):1061-1067. [doi: 10.1016/j.jhep.2009.09.001] [Medline: 19846234]
14. Cho Y. Hidden burden of alcohol use disorder in MASLD and MetALD: clinical and nomenclatural implications. Gut Liver 2025 Sep 15;19(5):637-638. [doi: 10.5009/gnl250414] [Medline: 40947954]
15. Lin H, Zhang X, Li G, Wong GLH, Wong VWS. Epidemiology and clinical outcomes of metabolic (dysfunction)-associated fatty liver disease. J Clin Transl Hepatol 2021 Dec 28;9(6):972-982. [doi: 10.14218/JCTH.2021.00201] [Medline: 34966660]
16. Stefan N, Yki-Järvinen H, Neuschwander-Tetri BA. Metabolic dysfunction-associated steatotic liver disease: heterogeneous pathomechanisms and effectiveness of metabolism-based treatment. Lancet Diabetes Endocrinol 2025 Feb;13(2):134-148. [doi: 10.1016/S2213-8587(24)00318-8] [Medline: 39681121]
17. Soldera J, Corso LL, Rech MM, et al. Predicting major adverse cardiovascular events after orthotopic liver transplantation using a supervised machine learning model: a cohort study. World J Hepatol 2024 Feb 27;16(2):193-210. [doi: 10.4254/wjh.v16.i2.193] [Medline: 38495288]
18. Meng D, Zhang L, Cao G, Cao W, Zhang G, Hu B. Liver fibrosis classification based on transfer learning and FCNet for ultrasound images. IEEE Access 2017;5:1-1. [doi: 10.1109/ACCESS.2017.2689058]
19. Wang Z, Bian H, Li J, et al. Detection and subtyping of hepatic echinococcosis from plain CT images with deep learning: a retrospective, multicentre study. Lancet Digit Health 2023 Nov;5(11):e754-e762. [doi: 10.1016/S2589-7500(23)00136-X] [Medline: 37770335]
20. Xiao W, Huang X, Wang JH, et al. Screening and identifying hepatobiliary diseases through deep learning using ocular images: a prospective, multicentre study. Lancet Digit Health 2021 Feb;3(2):e88-e97. [doi: 10.1016/S2589-7500(20)30288-0] [Medline: 33509389]
21. Calderaro J, Ghaffari Laleh N, Zeng Q, et al. Deep learning-based phenotyping reclassifies combined hepatocellular-cholangiocarcinoma. Nat Commun 2023 Dec 14;14(1):8290. [doi: 10.1038/s41467-023-43749-3] [Medline: 38092727]
22. Yang Y, Liu J, Sun C, et al. Nonalcoholic fatty liver disease (NAFLD) detection and deep learning in a Chinese community-based population. Eur Radiol 2023 Aug;33(8):5894-5906. [doi: 10.1007/s00330-023-09515-1] [Medline: 36892645]
23. Wang K, Cunha GM, Hasenstab K, et al. Deep learning for inference of hepatic proton sensity fat fraction from T1-weighted in-phase and opposed-phase MRI: retrospective analysis of population-based trial data. AJR Am J Roentgenol 2023 Nov;221(5):620-631. [doi: 10.2214/AJR.23.29607] [Medline: 37466189]
24. Zhou LQ, Wang JY, Yu SY, et al. Artificial intelligence in medical imaging of the liver. World J Gastroenterol 2019 Feb 14;25(6):672-682. [doi: 10.3748/wjg.v25.i6.672] [Medline: 30783371]
25. Reitsma JB, Glas AS, Rutjes AWS, Scholten R, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005 Oct;58(10):982-990. [doi: 10.1016/j.jclinepi.2005.02.022] [Medline: 16168343]
26. Cohen JF, Deeks JJ, Hooft L, et al. Preferred reporting items for journal and conference abstracts of systematic reviews and meta-analyses of diagnostic test accuracy studies (PRISMA-DTA for Abstracts): checklist, explanation, and elaboration. BMJ 2021 Mar 15;372:n265. [doi: 10.1136/bmj.n265] [Medline: 33722791]
27. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]

XSL•FO
RenderX

28. Yang R, Zhou Y, Liu W, Shang H. Study on the grading model of hepatic steatosis based on improved DenseNet. J Healthc Eng 2022;2022:9601470. [doi: 10.1155/2022/9601470] [Medline: 35340251]
29. Acharya UR, Fujita H, Bhat S, et al. Decision support system for fatty liver disease using GIST descriptors extracted from ultrasound images. Inf Fusion 2016 May;29:32-39. [doi: 10.1016/j.inffus.2015.09.006]
30. Jeon SK, Lee JM, Joo I, Yoon JH, Lee G. Two-dimensional convolutional neural network using quantitative US for noninvasive assessment of hepatic steatosis in NAFLD. Radiology 2023 Apr;307(1):e221510. [doi: 10.1148/radiol.221510] [Medline: 36594835]
31. Neogi N, Adhikari A, Roy M. Use of a novel set of features based on texture anisotropy for identification of liver steatosis from ultrasound images: a simple method. Multimed Tools Appl 2019 May;78(9):11105-11127. [doi: 10.1007/s11042-018-6675-0]
32. Chen JR, Chao YP, Tsai YW, et al. Clinical value of information entropy compared with deep learning for ultrasound grading of hepatic steatosis. Entropy (Basel) 2020 Sep 9;22(9):1006. [doi: 10.3390/e22091006] [Medline: 33286775]
33. Dubois M, Ronot M, Houssel-Debry P, et al. Performance of B-mode ratio and 2D shear wave elastography for the detection and quantification of hepatic steatosis and fibrosis after liver transplantation. Eur J Gastroenterol Hepatol 2020 Feb;32(2):222-230. [doi: 10.1097/MEG.0000000000001500] [Medline: 31464783]
34. Shi X, Ye W, Liu F, et al. Ultrasonic liver steatosis quantification by a learning-based acoustic model from a novel shear wave sequence. Biomed Eng Online 2019 Dec 21;18(1):121. [doi: 10.1186/s12938-019-0742-2] [Medline: 31864367]
35. Jesper D, Klett D, Schellhaas B, et al. Ultrasound-based attenuation imaging for the non-invasive quantification of liver fat - a pilot study on feasibility and inter-observer variability. IEEE J Transl Eng Health Med 2020;8:1800409. [doi: 10.1109/JTEHM.2020.3001488] [Medline: 32617199]
36. McHenry S, Park Y, Browning JD, Sayuk G, Davidson NO. Dallas steatosis index identifies patients with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol 2020 Aug;18(9):2073-2080. [doi: 10.1016/j.cgh.2020.01.020] [Medline: 31982611]
37. Roy M, Wang F, Vo H, et al. Deep-learning-based accurate hepatic steatosis quantification for histological assessment of liver biopsies. Lab Invest 2020 Oct;100(10):1367-1383. [doi: 10.1038/s41374-020-0463-y] [Medline: 32661341]
38. Sun L, Marsh JN, Matlock MK, et al. Deep learning quantification of percent steatosis in donor liver biopsy frozen sections. EBioMedicine 2020 Oct;60:103029. [doi: 10.1016/j.ebiom.2020.103029] [Medline: 32980688]
39. Chou TH, Yeh HJ, Chang CC, et al. Deep learning for abdominal ultrasound: a computer-aided diagnostic system for the severity of fatty liver. J Chin Med Assoc 2021 Sep 1;84(9):842-850. [doi: 10.1097/JCMA.0000000000000585] [Medline: 34282076]
40. Constantinescu EC, Udri toiu AL, Udri toiu C, et al. Transfer learning with pre-trained deep convolutional neural networks for the automatic assessment of liver steatosis in ultrasound images. Med Ultrason 2021 May 20;23(2):135-139. [doi: 10.11152/mu-2746] [Medline: 33626114]
41. Pérez-Sanz F, Riquelme-Pérez M, Martínez-Barba E, et al. Efficiency of machine learning algorithms for the determination of macrovesicular steatosis in frozen sections stained with sudan to evaluate the quality of the graft in liver transplantation. Sensors (Basel) 2021 Mar 12;21(6):1993. [doi: 10.3390/s21061993] [Medline: 33808978]
42. Pickhardt PJ, Blake GM, Graffy PM, et al. Liver steatosis categorization on contrast-enhanced CT using a fully automated deep learning volumetric segmentation tool: evaluation in 1204 healthy adults using unenhanced CT as a reference standard. AJR Am J Roentgenol 2021 Aug;217(2):359-367. [doi: 10.2214/AJR.20.24415] [Medline: 32936018]
43. Rhyou SY, Yoo JC. Cascaded deep learning neural network for automated liver steatosis diagnosis using ultrasound images. Sensors (Basel) 2021 Aug 5;21(16):16. [doi: 10.3390/s21165304] [Medline: 34450746]
44. Destrempes F, Gesnik M, Chayer B, et al. Quantitative ultrasound, elastography, and machine learning for assessment of steatosis, inflammation, and fibrosis in chronic liver disease. PLoS One 2022;17(1):e0262291. [doi: 10.1371/journal.pone.0262291] [Medline: 35085294]
45. Alshagathrh FM, Musleh S, Alzubaidi M, Schneider J, Househ MS. Efficient detection of hepatic steatosis in ultrasound images using convolutional neural networks: a comparative study. Trait du Signa 2023 Oct 30;40(5):1781-1794. [doi: 10.18280/ts.400501]
46. Podder S, Mallick A, Das S, Sau K, Roy A. Accurate diagnosis of liver diseases through the application of deep convolutional neural network on biopsy images. . 2023(4) p. 453-481. [doi: 10.3934/biophy.2023026]
47. Ibrahim MN, Blázquez-García R, Lightstone A, et al. Automated fatty liver disease detection in point-of-care ultrasound b-mode images. J Med Imaging (Bellingham) 2023 May;10(3):034505. [doi: 10.1117/1.JMI.10.3.034505] [Medline: 37284231]
48. Yao Y, Zhang Z, Peng B, Tang J. Bio-inspired network for diagnosing liver steatosis in ultrasound images. Bioengineering (Basel) 2023 Jun 26;10(7):768. [doi: 10.3390/bioengineering10070768] [Medline: 37508795]
49. Byra M, Styczynski G, Szmigielski C, et al. Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. Int J Comput Assist Radiol Surg 2018 Dec;13(12):1895-1903. [doi: 10.1007/s11548-018-1843-2] [Medline: 30094778]

XSL•FO
RenderX

50. Torgersen J, Akers S, Huo Y, et al. Performance of an automated deep learning algorithm to identify hepatic steatosis within noncontrast computed tomography scans among people with and without HIV. Pharmacoepidemiol Drug Saf 2023 Oct;32(10):1121-1130. [doi: 10.1002/pds.5648] [Medline: 37276449]
51. Wang Q, Lai MW, Bin G, et al. MBR-Net: a multi-branch residual network based on ultrasound backscattered signals for characterizing pediatric hepatic steatosis. Ultrasonics 2023 Dec;135:107093. [doi: 10.1016/j.ultras.2023.107093] [Medline: 37482038]
52. Jeon SK, Joo I, Park J, Yoo J. Automated hepatic steatosis assessment on dual-energy CT-derived virtual non-contrast images through fully-automated 3D organ segmentation. Radiol Med 2024 Jul;129(7):967-976. [doi: 10.1007/s11547-024-01833-8] [Medline: 38869829]
53. Piella G, Farré N, Esono D, et al. LiverColor: an artificial intelligence platform for liver graft assessment. Diagnostics (Basel) 2024 Jul 31;14(15):15. [doi: 10.3390/diagnostics14151654] [Medline: 39125531]
54. Yoo J, Joo I, Jeon SK, Park J, Yoon SH. Utilizing fully-automated 3D organ segmentation for hepatic steatosis assessment with CT attenuation-based parameters. Eur Radiol 2024 Sep;34(9):6205-6213. [doi: 10.1007/s00330-024-10660-4] [Medline: 38393403]
55. Zhang Z, Li G, Wang Z, et al. Deep-learning segmentation to select liver parenchyma for categorizing hepatic steatosis on multinational chest CT. Sci Rep 2024;14(1):11987. [doi: 10.1038/s41598-024-62887-2]
56. Cherchi V, Mea VD, Terrosu G, et al. Assessment of hepatic steatosis based on needle biopsy images from deceased donor livers. Clin Transplant 2022 Mar;36(3):e14557. [doi: 10.1111/ctr.14557] [Medline: 34890087]
57. Wu X, Lv K, Wu S, Tai DI, Tsui PH, Zhou Z. Parallelized ultrasound homodyned-K imaging based on a generalized artificial neural network estimator. Ultrasonics 2023 Jul;132:106987. [doi: 10.1016/j.ultras.2023.106987] [Medline: 36958066]
58. Drazinos P, Gatos I, Katsakiori PF, et al. Comparison of deep learning schemes in grading non-alcoholic fatty liver disease using b-mode ultrasound hepatorenal window images with liver biopsy as the gold standard. Phys Med 2025 Jan;129:104862. [doi: 10.1016/j.ejmp.2024.104862] [Medline: 39626614]
59. Kaffas AE, Bhatraju KC, Vo-Phamhi JM, et al. Development of a deep learning model for classification of hepatic steatosis from clinical standard ultrasound. Ultrasound Med Biol 2025 Feb;51(2):242-249. [doi: 10.1016/j.ultrasmedbio.2024.09.020] [Medline: 39537545]
60. Kim HY, Lee KJ, Lee SS, et al. Diagnosis of moderate-to-severe hepatic steatosis using deep learning-based automated attenuation measurements on contrast-enhanced CT. Abdom Radiol (NY) 2025 Sep;50(9):4139-4147. [doi: 10.1007/s00261-025-04872-5] [Medline: 40095018]
61. Zhang H, Liu J, Su D, et al. Diagnostic of fatty liver using radiomics and deep learning models on non-contrast abdominal CT. PLoS ONE 2025;20(2):e0310938. [doi: 10.1371/journal.pone.0310938] [Medline: 39946425]
62. Derstine BA, Holcombe SA, Chen VL, et al. Quantification of hepatic steatosis on post-contrast computed tomography scans using artificial intelligence tools. Abdom Radiol 2025. [doi: 10.1007/s00261-025-05137-x]
63. Del Corso G, Pascali MA, Caudai C, et al. ANN uncertainty estimates in assessing fatty liver content from ultrasound data. Comput Struct Biotechnol J 2024 Dec;24:603-610. [doi: 10.1016/j.csbj.2024.09.021] [Medline: 39421530]
64. Hernaez R, Lazo M, Bonekamp S, et al. Diagnostic accuracy and reliability of ultrasonography for the detection of fatty liver: a meta-analysis. Hepatology 2011 Sep 2;54(3):1082-1090. [doi: 10.1002/hep.24452] [Medline: 21618575]
65. Park HJ, Kim KW, Kwon HJ, et al. CT-based visual grading system for assessment of hepatic steatosis: diagnostic performance and interobserver agreement. Hepatol Int 2022 Oct;16(5):1075-1084. [doi: 10.1007/s12072-022-10373-0] [Medline: 35789473]
66. Azizi N, Naghibi H, Shakiba M, et al. Evaluation of MRI proton density fat fraction in hepatic steatosis: a systematic review and meta-analysis. Eur Radiol 2025 Apr;35(4):1794-1807. [doi: 10.1007/s00330-024-11001-1] [Medline: 39254718]
67. Tapper EB, Lok ASF. Use of liver imaging and biopsy in clinical practice. N Engl J Med 2017 Aug 24;377(8):756-768. [doi: 10.1056/NEJMra1610570] [Medline: 28834467]
68. Petroff D, Blank V, Newsome PN, et al. Assessment of hepatic steatosis by controlled attenuation parameter using the M and XL probes: an individual patient data meta-analysis. Lancet Gastroenterol Hepatol 2021 Mar;6(3):185-198. [doi: 10.1016/S2468-1253(20)30357-5] [Medline: 33460567]
69. Siddiqui MS, Vuppalanchi R, Van Natta ML, et al. Vibration-controlled transient elastography to assess fibrosis and steatosis in patients with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol 2019 Jan;17(1):156-163. [doi: 10.1016/j.cgh.2018.04.043] [Medline: 29705261]
70. Hepburn MJ, Vos JA, Fillman EP, Lawitz EJ. The accuracy of the report of hepatic steatosis on ultrasonography in patients infected with hepatitis C in a clinical setting: a retrospective observational study. BMC Gastroenterol 2005 Apr 13;5:14. [doi: 10.1186/1471-230X-5-14] [Medline: 15829009]
71. Fischer MA, Reiner CS, Raptis D, et al. Quantification of liver iron content with CT-added value of dual-energy. Eur Radiol 2011 Aug;21(8):1727-1732. [doi: 10.1007/s00330-011-2119-1] [Medline: 21472472]
72. Loizillon S, Bottani S, Maire A, et al. Automatic quality control of brain 3D FLAIR MRIs for a clinical data warehouse. Med Image Anal 2025 Jul;103:103617. [doi: 10.1016/j.media.2025.103617] [Medline: 40344945]

XSL•FO
RenderX

73. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. Nat Commun 2024 Jan 22;15(1):654. [doi: 10.1038/s41467-024-44824-z] [Medline: 38253604]
74. Sheller MJ, Edwards B, Reina GA, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep 2020 Jul 28;10(1):12598. [doi: 10.1038/s41598-020-69250-1] [Medline: 32724046]
75. Vianna P, Calce SI, Boustros P, et al. Comparison of radiologists and deep learning for US grading of hepatic steatosis. Radiology 2023 Oct;309(1):e230659. [doi: 10.1148/radiol.230659] [Medline: 37787678]
76. Zöggeler T, Kavallar AM, Pollio AR, et al. Meta-analysis of shotgun sequencing of gut microbiota in obese children with MASLD or MASH. Gut Microbes 2025 Dec;17(1):2508951. [doi: 10.1080/19490976.2025.2508951] [Medline: 40396204]
77. Hou M, Zhu Y, Zhou H, et al. Innovative machine learning approach for liver fibrosis and disease severity evaluation in MAFLD patients using MRI fat content analysis. Clin Exp Med 2025 Aug 5;25(1):275. [doi: 10.1007/s10238-025-01818-5] [Medline: 40762814]
78. Wei Y, Yang M, Zhang M, et al. Focal liver lesion diagnosis with deep learning and multistage CT imaging. Nat Commun 2024;15(1):7040. [doi: 10.1038/s41467-024-51260-6] [Medline: 39147767]
79. Dunn N, Verma N, Dunn W. Artificial Intelligence for predictive diagnostics, prognosis, and decision support in MASLD, hepatocellular carcinoma, and digital pathology. J Clin Exp Hepatol 2026;16(1):103184. [doi: 10.1016/j.jceh.2025.103184] [Medline: 41127419]
80. Raji CG, Chandra SSV, Gracious N, Pillai YR, Sasidharan A. Advanced prognostic modeling with deep learning: assessing long-term outcomes in liver transplant recipients from deceased and living donors. J Transl Med 2025 Feb 16;23(1):188. [doi: 10.1186/s12967-025-06183-1] [Medline: 39956905]
81. Guillot J, Williams CYK, Azzam S, et al. Risk prediction in patients with metabolic dysfunction-associated steatohepatitis using natural language processing. Gastro Hep Adv 2025;4(9):100701. [doi: 10.1016/j.gastha.2025.100701] [Medline: 40688387]
82. Zhang J, Feng S, Xue J, et al. AI-driven multimodal fusion of tongue images and clinical indicators for identifying MAFLD patients at risk of coronary artery disease: an exploratory study. ILIVER 2025 Sep;4(3):100181. [doi: 10.1016/j.iliver.2025.100181] [Medline: 41054419]
83. Li B, Tai DI, Yan K, et al. Accurate and generalizable quantitative scoring of liver steatosis from ultrasound images via scalable deep learning. World J Gastroenterol 2022 Jun 14;28(22):2494-2508. [doi: 10.3748/wjg.v28.i22.2494] [Medline: 35979264]
84. Determination of signal-to-noise ratio (SNR) in diagnostic magnetic resonance imaging. : NEMA; 2021 URL: https://www.nema.org/Standards/view/Determination-of-Signal-to-Noise-Ratio-in-Diagnostic-Magnetic-Resonance-Imaging [accessed 2021-03-31]
85. Lee B, Kramer P, Sandri S, et al. Early recalls and clinical validation gaps in artificial intelligence-enabled medical devices. JAMA Health Forum 2025 Aug 1;6(8):e253172. [doi: 10.1001/jamahealthforum.2025.3172] [Medline: 40844774]
86. Li YH, Li YL, Wei MY, Li GY. Innovation and challenges of artificial intelligence technology in personalized healthcare. Sci Rep 2024 Aug 16;14(1):18994. [doi: 10.1038/s41598-024-70073-7] [Medline: 39152194]
87. Bhandari M, Zeffiro T, Reddiboina M. Artificial intelligence and robotic surgery: current perspective and future directions. Curr Opin Urol 2020 Jan;30(1):48-54. [doi: 10.1097/MOU.0000000000000692] [Medline: 31724999]
88. Deshmukh AD, Wagner JK. FDA draft guidelines for AI and the need for ethical frameworks. JAMA Pediatr 2025 Sep 1;179(9):937-938. [doi: 10.1001/jamapediatrics.2025.1979] [Medline: 40622691]

Abbreviations

AI: artificial intelligence
AUC: area under curve
DL: deep learning
HCC: hepatocellular carcinoma
HS: hepatic steatosis
LRN: negative likelihood ratio
LRP: positive likelihood ratio
MAFLD: metabolic dysfunction associated fatty liver disease
ML: machine learning
MRI: magnetic resonance imaging
MRI-PDFF: magnetic resonance imaging-proton density fat fraction
PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Diagnostic Test Accuracy
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search Extension
SROC: summary receiver operating characteristic
TL: transfer learning

XSL•FO
RenderX

Edited by S Brini; submitted 30.May.2025; peer-reviewed by J Soldera, KH Lin, MBR Mahmoud; revised version received 11.Nov.2025; accepted 11.Nov.2025; published 13.Jan.2026.

Please cite as:
Song J, Liu D, Li J, Cong H, Deng R, Lu Y, Sun J, Zhang J
Assessment of the Diagnostic Performance and Clinical Impact of AI in Hepatic Steatosis: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e78310
URL: https://www.jmir.org/2026/1/e78310
doi:10.2196/78310

© Jiamei Song, Dan Liu, Jitong Li, Haoru Cong, Ruixue Deng, Yihan Lu, Jiayi Sun, Jingzhou Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI for Detecting and Predicting Postpartum Depression: Scoping Review

Mais Alkhateeb[1,2,3], PhD; Ajisha Nayeem[2]; Arfan Ahmed[2], PhD; Mohammed Alsahli[4], PhD; Javaid Sheikh[2], MD; Alaa Abd-Alrazaq[2], PhD

[1]College of Education and Art, Lusail University, Doha, Qatar

[2]AI Center for Precision Health, Weill Cornell Medicine-Qatar, Doha, Qatar

[3]Computer Science and Engineering, Hamad bin Khalifa University, Qatar Foundation, Doha, Qatar

[4]Health Informatics Department, College of Health Science, Saudi Electronic University, Riyadh, Saudi Arabia

Corresponding Author:
Mais Alkhateeb, PhD
College of Education and Art, Lusail University, Doha, Qatar

Abstract

Background: Postpartum depression (PPD) affects up to 20% of mothers globally. Early detection is vital for better outcomes, yet screening lacks scalability and predictive power. Artificial intelligence (AI)—through machine learning, deep learning, and natural language processing—enhances the early identification of mothers at risk with greater accuracy.

Objective: This study aims to systematically map the existing literature on AI-based methods for detecting and predicting PPD.

Methods: This scoping review was conducted in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines. We included empirical studies that applied AI techniques to detect or predict PPD and were published in peer-reviewed journals, conference proceedings, or dissertations. Studies were excluded if they were nonempirical (eg, reviews, editorials, and abstracts), not published in English, focused on general perinatal mental health without a specific emphasis on PPD, or used AI solely for monitoring or treatment rather than prediction or detection. We systematically searched 8 databases—MEDLINE, Embase, PsycINFO, CINAHL, Scopus, IEEE Xplore, ACM Digital Library, and Google Scholar—from inception through February 28, 2025. The search strategy was supplemented by backward and forward reference screening and biweekly alerts to capture newly published studies. Two independent (M [Alkhateeb] and A [Nayeem])reviewers (M [Alkhateeb] and A [Nayeem]) screened the retrieved studies, with disagreements resolved by a third reviewer (AA [Alrazaq]). Data were extracted by 2 independent reviewers using a standardized extraction form capturing study characteristics, AI model types, data sources, features, preprocessing, validation strategies, and performance metrics. A formal risk-of-bias assessment was not performed due to the scoping nature of the review. All extracted data were synthesized narratively.

Results: Out of 503 retrieved studies, 65 met the inclusion criteria. The United States contributed the largest proportion of studies (18/65, 27.7%). The highest number of publications occurred in 2024 (17/65, 26%). Most included studies were journal articles (46/65, 71%). Short-term postpartum outcomes (≤12 weeks) were most frequently assessed (20/65, 30.8%). Most included studies (52/65, 80%) applied AI models for predicting PPD, while 14 of 65 (22%) studies used them for detection. Sociodemographic data were most frequently used (49/65, 75.4%), followed by psychological data (44/65, 68%) and obstetric data (35/65, 55%). Data preprocessing mostly relied on basic scaling (51/65, 79%) and some missing data imputation (29/65, 44.6%). Machine learning dominated (57/65, 87.7%), especially random forest, support vector machines, and logistic regression. Internal validation (k-fold, hold-out) was standard, while external validation was scarce. Ensemble-based boosting models consistently demonstrated superior performance across key metrics, highlighting their potential for accurate and scalable PPD prediction. Current studies suffer from limited sample sizes, geographic bias, lack of standardized feature sets, minimal external validation, and inconsistent reporting of comprehensive model metrics.

Conclusions: This scoping review analyzes 65 studies on AI in PPD, highlighting dominant use of classical machine learning, limited deep learning adoption, underuse of advanced preprocessing, inconsistent validation, and reliance on structured, unimodal data—mainly sociodemographic, clinical, and obstetric features.

(J Med Internet Res 2026;28:e77376) doi:10.2196/77376

KEYWORDS

postpartum depression; maternal mental health; perinatal depression; artificial intelligence; machine learning; deep learning; natural language processing; prediction models; computer-aided diagnosis

XSL•FO
RenderX

Introduction

Background

Postpartum depression (PPD) is a common mental health issue affecting new mothers after they give birth. Its salient features include feelings of enduring melancholy, losing interest in hobbies and their everyday life (and, potentially, their baby), and reduced feelings of pleasure in activities.

Traditionally referred to as the "baby blues," PPD is a profound and serious condition undermining the activities of daily life and the psychosocial well-being of mothers. It affects up to a fifth of new mothers worldwide, albeit it is often undiagnosed; in any case, it is a major concern for public health [1].

Postpartum care is vital to ensure the best outcomes for both neonates and mothers. This includes creating a supportive environment with health-promoting activities and breastfeeding encouragement. It must also address each mother's individual mental health needs. [2].

Worldwide, studies of perinatal mental health have noted that PPD is increasingly evident [3]. According to the estimations of the National Institute of Mental Health, up to 15% of all women who experience pregnancy also experience related depression, whether during or after pregnancy. Prevalence is typically higher (ie, 18% - 25%) in low- and middle-income countries, where it is associated with socioeconomic issues and health care resource availability and access, as well as sociocultural factors [4].

The great variety in PPD prevalence underscores the requirement for efficacious strategies for screening women and delivering interventions catering to various needs. The mainstay for PPD detection now is dependent on women reporting classic symptoms and completing self-reported tools, of which the Edinburgh Postnatal Depression Scale (EPDS) [5] and the Patient Health Questionnaire-9 (PHQ-9) [6] are common and effective. However, such tools for screening women during and after pregnancy are typically not administered consistently. For instance, women are often screened once during the early stages of pregnancy, such as the second trimester. However, the same tools are rarely reapplied in later or postpartum periods [7,8].

Furthermore, the tools detect current depression, with no scope to anticipate future risk (based on current symptoms and feelings) [9]. Prediction and early diagnosis of PPD remain challenging, largely because qualitative narrative data are difficult to interpret and integrate alongside quantitative clinical metrics.

A detailed professional analysis is necessary to interpret data appropriately, which is costly, time-consuming, and potentially subjective [10]. PPD prevention and treatment interventions require improved screening solutions that can be delivered during early pregnancy and throughout the pregnancy journey and postpartum period.

Artificial intelligence (AI) can potentially address this impasse, with its capability to handle and process vast volumes of complex, high-dimensional, nonlinear data. Using machine learning (ML), large language models, and natural language processing (NLP) techniques, AI can detect subtle patterns inherent within data that could otherwise evade human analysis [10,11].

AI can enhance prediction accuracy by incorporating diverse data sources. These include electronic health records (EHRs), diagnostic indicators, self-reported feelings, and behavioral cues gathered from digital platforms, with appropriate safeguards [1,12]. Such possibilities render AI a highly useful clinical tool, offering real-time decision-making input for digital care delivery.

Research Problem and Aim

Many studies have developed AI models for detecting and predicting PPD, yet these studies offer fragmented insights into the full potential of AI methodologies. Several previous reviews attempted to summarize these insights [10,13-19], but they have notable limitations. Specifically, some prior reviews were traditional narrative reviews rather than systematic or scoping reviews and thus lacked rigorous, structured methodologies [13,14,16].

In addition, many earlier reviews used narrow search queries or omitted critical databases (eg, PsycINFO, ACM Digital Library, IEEE Xplore, Scopus, and Embase), potentially excluding relevant studies [10,13-19]. Furthermore, past reviews often broadly addressed general depression or women's mental health instead of specifically targeting PPD, limiting their direct relevance [10,15]. Also, the bibliographic searches of previous reviews mostly concluded before September 2022, omitting recent advancements in AI methodologies and multimodal data integration techniques. Importantly, most prior research emphasized traditional clinical and survey-based data, neglecting innovative data sources such as social media and wearable sensors. These novel data sources represent a promising opportunity to enhance AI model accuracy and predictive capabilities for PPD [10,13,14,16-19].

The primary aim of this review is to map the landscape of AI methodologies used in PPD detection and prediction and to identify key research trends, methodological features, and evidence gaps. Specifically, this review is guided by the following research subquestions:

- What types of AI models have been used to detect or predict PPD, and how do they differ in approach and complexity?
- What data modalities (eg, structured, unstructured, physiological, and digital) have been used in these studies?
- How have studies handled model development processes such as feature selection, validation, and interpretability?
- What are the key limitations, challenges, and future opportunities for applying AI to PPD detection in real-world clinical and community settings?

By addressing these questions, this review provides a structured, up-to-date, and integrative overview of AI in postpartum mental health—highlighting opportunities for innovation, responsible deployment, and policy translation in maternal care.

XSL•FO
RenderX

Methods

We conducted a scoping review in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines (see PRISMA-ScR checklist). The following sections detail the specific methods we used in this review.

Search Strategy

All-inclusive searches were done across the following 8 major electronic databases on November 18, 2024, to determine relevant studies: MEDLINE (via Ovid), PsycINFO (via Ovid), Embase (via Ovid), CINAHL (via EBSCO), IEEE Xplore, ACM Digital Library, Scopus, and Google Scholar. To keep our search up to date, we set up an automatic biweekly search alert for 24 weeks, ending on February 28, 2025. Given the vast number of results generated by Google Scholar, we focused on the first 100 results (10 pages), as they are ranked by relevance. In addition to database searches, we expanded our review by manually screening reference lists of included studies (backward reference checking) and identifying studies that cited them (forward reference checking). We also collected additional papers through automatic email alerts. To ensure that the search query was well-structured and effective, 3 experts in digital mental health were consulted and previous relevant literature was reviewed. Two main categories of terms were included in the final search query: AI-related terms (eg, artificial intelligence, machine learning, and deep learning) and PPD-related terms (eg, postpartum depression, postpartum depression, and postnatal depression). A detailed search query used for each database is shown in Multimedia Appendix 1.

Study Eligibility Criteria

This scoping review targeted studies that specifically applied AI to the detection or prediction of PPD. Eligible studies were empirical in nature, used AI methodologies, and were published in peer-reviewed journals, dissertations, or conference proceedings. There were no restrictions regarding publication year, country of origin, data type, study design, population, or outcome type.

Exclusion criteria encompassed nonempirical works such as reviews, abstracts, commentaries, and proposals, as well as studies lacking a specific focus on PPD (eg, addressing broader maternal or perinatal mental health). Studies that used AI solely for managing or monitoring PPD, rather than detecting or predicting it, were also excluded. In addition, only those papers published in English were considered.

Study Selection

The study selection process in this review involved 3 main steps. First, we used EndNote to remove any duplicate studies from our search results. Then, the titles and abstracts of the remaining studies were screened to determine their relevance. For studies that passed this initial screening, a full-text review was conducted, during which the entire paper, including any supplementary materials, was carefully read. To ensure accuracy, 2 independent reviewers (MA and AN) conducted the study selection process. In cases of disagreement during title or abstract screening or full-text review, a third reviewer (AAA) was consulted to resolve the conflict. In addition, we calculated Cohen κ statistic to assess interreviewer agreement, which yielded a value of 0.78-0.83 by title or abstract screening or full-text review—indicating a high level of consistency and reliability in the data selection process [20].

Data Extraction

To ensure a structured and consistent approach to data extraction, we developed a data extraction form, which was pilot-tested using 5 selected studies before full implementation. This form was designed to capture key details related to the study characteristics, datasets, features, and AI methodologies. The finalized data extraction form used in this review is shown in Multimedia Appendix 2. Two independent reviewers (MA and AN) used Microsoft Excel to extract data systematically. Any discrepancies between them were resolved through discussion.

Data Synthesis

We analyzed the extracted data using a narrative synthesis approach, summarizing key findings in descriptive text and tables to provide a clear overview of the research. First, we outlined the basic details of each study, including the year of publication and the country where the research was conducted. Subsequently, we characterized the datasets underpinning AI model development, cataloged the AI methodologies used in each study, and detailed the feature attributes used in model construction. To keep the process structured and ensure accuracy, we used Microsoft Excel to organize and synthesize the extracted data efficiently.

Results

Search Results

As illustrated in Figure 1, a total of 503 records were retrieved through searches across 9 databases: Ovid MEDLINE (n=64), Embase (n=48), PsycINFO (n=26), CINAHL (n=22), IEEE Xplore (n=16), ACM Digital Library (n=2), Scopus (n=145), Web of Science (n=80), and Google Scholar (n=100, limited to the top 100 results ranked by relevance). After removing 272 duplicate records using reference management software, 231 unique reports remained for screening. After reviewing the titles and abstracts, 145 records were excluded. The full texts of the remaining 86 records were retrieved for further assessment. Of these, 7 full-text papers were not available. After evaluating the 79 available full-text papers, 17 studies were excluded for the following reasons: they did not use AI (n=7); did not focus on PPD (n=1); were not journal papers, conference papers, or dissertations (n=8); or were not written in English (n=1). Three additional relevant studies were identified through both backward and forward reference list screening. Ultimately, 65 studies were included in this review [21-85].

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram illustrating the study selection process. AI: artificial intelligence; PPD: postpartum depression.

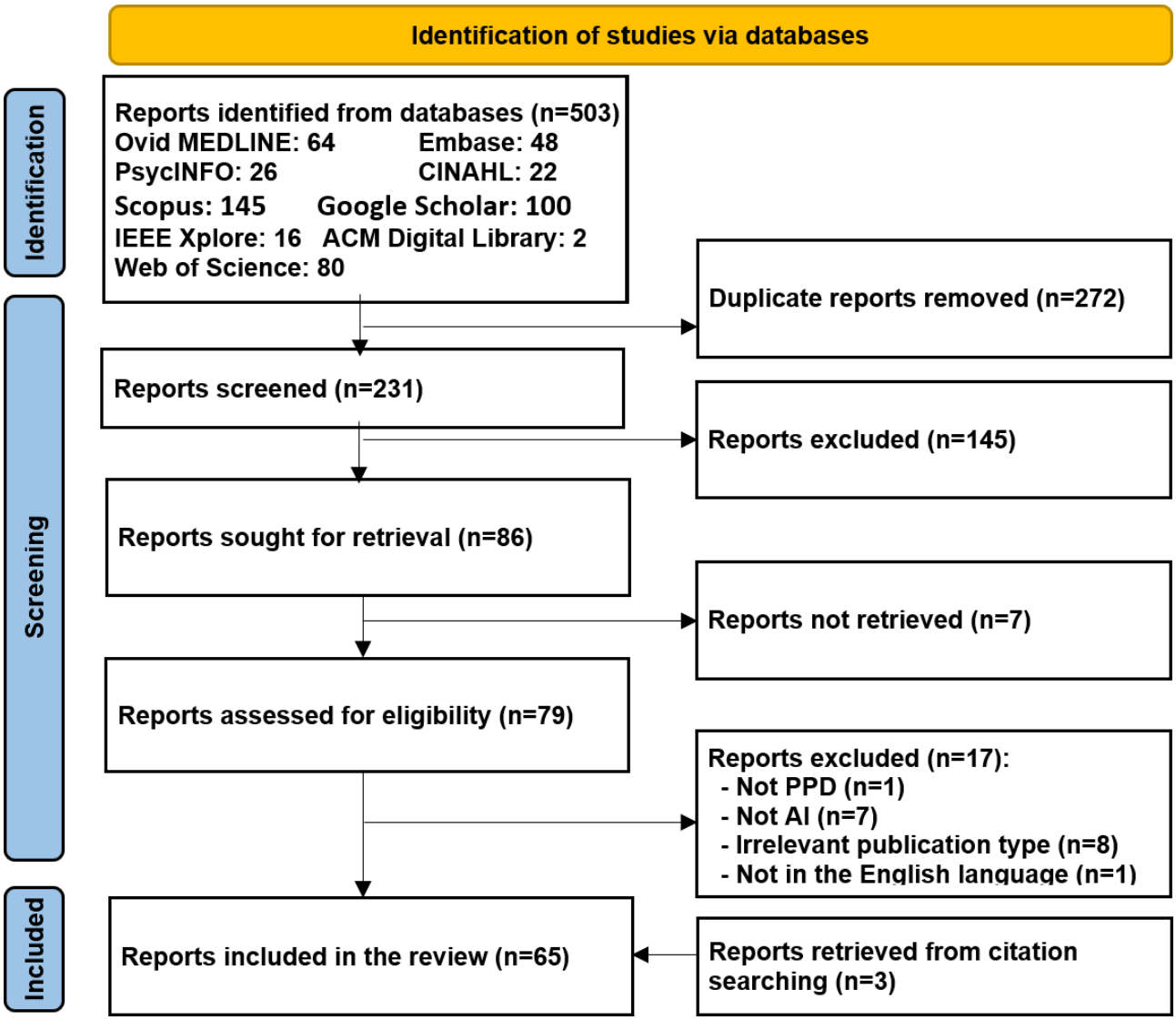

Characteristics of the Included Studies

As shown in Table 1, the included studies were published between 2009 and 2025, with the highest number of publications occurring in 2024 (26.1%). Regarding publication types, the majority were journal papers (70.8%). The United States contributed the largest proportion of studies (27.7%), followed by China (15.4%) and Bangladesh (13.9%). This review included 39 out of 65 (60%) retrospective studies and 27 out of 65 (41.5%) prospective studies. The number of participants in the included studies ranged from 11 to 573,634, with a mean of 18,187.4 (SD 72,933.8). Participant distribution was as follows: 23 studies had fewer than 500 participants, 27 studies included between 500 and 5000 participants, 8 studies had between 5001 and 50,000 participants, and 7 studies involved more than 50,000 participants. Among the 26 studies that reported mean participant age, values ranged from 26 to 44.5 years, with an overall mean average of 31.08 (SD 3.42) years. See Multimedia Appendix 3 for overall information about included studies.

XSL•FO
RenderX

Table . Characteristics of the included studies.

Key aspects	Studies	References
Year of publication, n/N (%)		
2025	3/65 (4.6)	[30,46,77]
2024	17/65 (26.1)	[21,24,35,37,40,41,43,45,51,52,65-67,69,78,83,85]
2023	14/65 (21.5)	[28,32,33,42,44,50,56,63,68,72-74,79,80]
2022	10/65 (15.4)	[34,36,39,47,58-62,81]
2021	5/65 (7.7)	[22,26,48,55,84]
2020	6/65 (9.2)	[23,31,57,64,71,82]
2019	4/65 (6.2)	[25,29,49,76]
2018	3/65 (4.6)	[27,54,75]
≤2017	3/65 (4.6)	[38,53,70]
Publication type, n/N (%)		
Journal paper	46/65 (70.8)	[22,23,25,28-31,33,34,37-44,46,47,49-52,54,55,57,60,62-66,67-70,74,76-85]
Conference paper	18/65 (27.7)	[21,24,26,27,32,35,45,48,53,56,58,59,61,66,71-73,75]
Dissertation	1/65 (1.5)	[36]
Country of publication, n/N (%)		
United States	18/65 (27.7)	[24,25,31,37,39,41,43,51,53,55,57,62,64,74,76,80,83,84]
China	10/65 (15.4)	[27,42,63,69,75,77-79,82,85]
Bangladesh	9/65 (13.9)	[21,35,40,45,52,61,65,67,73]
India	6/65 (9.2)	[32-34,48,56,66]
Japan	3/65 (4.6)	[46,47,81]
United Kingdom	2/65 (3.1)	[44,71]
Spain	2/65 (3.1)	[38,70]
Sri Lanka	2/65 (3.1)	[58,59]
Others[a]	13/65 (1 each) (20)	[22,23,26,28-30,36,49,50,54,60,68,72]
Research design[b], n/N (%)		
Retrospective	39/65 (60)	[21,23-27,30,32,35,37,40-45,47-49,51-56,60,61,64-66,71-76,80,83,84]
Prospective	27/65 (41.5)	[22,28,29,31,33,34,36,38,39,46,50,51,57-59,62,63,67-70,77-79,81,82,85]
Number of participants		
Mean (SD)	18,187.4 (72,933.8)	[21-85]
Range	11 - 573,634	[21-85]
<500	23/65 (35.4)	[24,26-28,30,31,33,34,37,48,49,53,54,57,61,66,68,71,72,77,79,81,85]
500 - 5000	27/65 (41.5)	[21,22,25,29,35,38-42,45,50,52,58-60,62,63,65,67,69,70,73,75,78,80,82]
5001 - 50,000	8/65 (12.3)	[32,36,47,51,56,64,74,76]
>50,000	7/65 (10.8)	[23,43,44,46,55,83,84]
Mean age (years)[c]		
Mean (SD)	31.08 (3.42)	[21-85]
Range	25.99 - 44.5	[21-85]

[a]Others include Australia, Brazil, Indonesia, Italy, Mexico, Nigeria, Norway, Pakistan, Palestine, Portugal, Saudi Arabia, Slovenia, and Sweden.

[b]The number of studies does not add up as one study used both retrospective and prospective designs.

[c]Mean age not reported: 39 studies (60%).

XSL•FO
RenderX

Characteristics of Datasets

As depicted in Table 2, the average dataset size was 37,338.5 (SD 160,309.6), with a range spanning from 16 to 1,170,446. The dataset size fell between 500 and 5000 in 28 out of 65 (43.1%) studies. Most studies (49/65, 75.4%) used closed-source data, while the remaining (16/65, 24.6%) relied on open-source datasets. The studies included various data formats, including textual, tabular, audio, video, and images. The majority (57/65, 87.7%) used a single data format (unimodal), while the remaining studies (8/65, 12.3%) integrated multiple data formats (multimodal). The survey was the most common data collection approach (50/65, 76.9% of studies), followed by data sourced from EHRs (25/65, 38.5% of studies). Most studies (45/65, 69.2%) were conducted in health care settings. Regarding the timing of outcome measurement for PPD, 20 out of 65 (30.8%) studies assessed outcomes within 12 weeks of delivery (short term), 14 out of 65 (21.5%) studies between 12 and 36 weeks (medium term), and 17 out of 65 (26.2%) studies after more than 36 weeks (long term). The most common reference standard used for labeling the data (outcomes) was the EPDS (37/65, 56.9% of studies). Further details on the characteristics of the datasets used in the included studies are provided in Multimedia Appendix 4.

XSL•FO
RenderX

Table . Characteristics of datasets used in the included studies.

Data summary	Studies, n (%)	References
Dataset size[a]		
Mean (SD)	37,338.5 (160,309.6)	[21-85]
Range	16 - 1,170,446	[21-85]
Dataset size categories, n/N (%)		
<500	20/65 (30.8)	[24,26,28,30,33,34,37,48,49,53,54,57,61,66,68,71,72,79,81,85]
500 - 5000	28/65 (43.1)	[21,22,25,29,31,35,38-42,45,50,52,58-60,62,63,65,67,69,70,73,77,78,80,82]
5001 - 50,000	8/65 (12.3)	[32,36,47,56,64,74-76]
>50,000	9/65 (13.9)	[23,27,43,44,46,51,55,83,84]
Data source		
Closed	49/65 (75.4)	[22-27,29-32,36-39,42-44,46-51,53-55,57-63,68-72,74-79,81-85]
Open	16/65 (24.6)	[21,28,33-35,40,41,45,52,56,64-67,73,80]
Data format[b]		
Unimodal	57/65 (87.7)	[21-26,29,30,32,34-40,42-50,52,54-66,68-76,79-85]
Multimodal	8/65 (12.3)	[27,28,31,33,41,51,53,67]
Data collection methodology[c]		
Survey	50/65 (76.9)	[21,22,24-28,30,31,33-36,38,40-42,44-46,50-54,56-66,67-73,75,78-82,85]
EHRs[d]	25/65 (38.5)	[23,37,41-44,46,47,49,50,55,67-70,73,74,76-79,81,83-85]
Social media	8/65 (12.3)	[27,29,32,33,51,53,66,67]
Sensor-based	5/65 (7.7)	[37-39,44,49]
Laboratory-based data	2/65 (3.1)	[79,81]
Setting[c]		
Health care	45/65 (69.2)	[21-23,28,30,33-35,37-43,45-47,49-52,54,55,57-59,63,65,67-71,73-79,82-85]
Community	18/65 (27.7)	[25,27,29,31,32,36,48,51,53,56,60-62,64,67,72,80,81]
Academic	5/65 (7.7)	[24,26,37,44,66]
Outcome measurement timing (weeks)[e]		
Short term (<12)	20/65 (30.8)	[34,38,42,46,47,50,52,55,57,62,63,67,68,71,74,77,78,81,82,85]
Medium term (12 - 36)	14/65 (21.5)	[36,46,50,57-59,62,63,70,72,78,80,81,85]
Long term (>36)	17/65 (26.2)	[22-24,36,37,43,48,53,57,63,64,69,76,80,83-85]
Reference standard		
EPDS[f]	37/65 (56.9)	[22,24,26-30,34,38-40,42,46,47,50,52,53,55-59,61-63,67-72,76-78,80-82]
ICD[g]	8/65 (12.3)	[2,23,29,41,42,44,49,76,83]
PHQ[h]	7/65 (10.8)	[3,34,43,60,61,64,67]
PDSS[i]	3/65 (4.6)	[4,25,34,48]
PPDS[j]	2/65 (3.1)	[5,32,67]

[a]Mean (SD) is calculated.

[b]The number of studies does not add up as some studies used multiple data collection methodologies.

[c]The number of studies does not add up as some studies are conducted in multiple settings.

[d]EHRs: electronic health records.

[e]The number of studies does not add up as the timing of outcome measurement varied across studies. Outcome measurement timing not reported: 27 (41.5%).

[f]EPDS: Edinburgh Postnatal Depression Scale.

XSL•FO
RenderX

[g]*ICD*: *International Classification of Diseases*.

[h]PHQ: Patient Health Questionnaire.

[i]PDSS: Postpartum Depression Screening Scale.

[j]PPDS: Postpartum Depression Scale.

Characteristics of Preprocessing Techniques

Table 3 summarizes the most frequently used preprocessing techniques identified across the reviewed studies. Across the reviewed literature, feature transformation overwhelmingly dominates preprocessing: Min-Max scaling and Z score standardization appear in 78.5% (51/65) of studies. In contrast, missing data strategies remain underutilized—only 44.6% (29/65) of papers applied any form of imputation, leaving 33.8% (22/65) to rely on case deletion or ignore the issue entirely. Class imbalance remedies are similarly rare: just 4.6% (3/65) of studies used stratified resampling or SMOTE variants, while cost-sensitive learning appeared in only 6.2% (4/65).

Categorical encoding methods vary in popularity: label encoding leads at 29.2% (19/65), one-hot encoding in 13.9% (9/65), binary encoding in 9.2% (6/65), dummy encoding in 6.2% (4/65), and target encoding in a mere 3.1% (2/65). For feature selection, tree-based importance (Gini impurity) featured in 18.5% (12/65) of studies and Pearson correlation filtering in 12.3%, with recursive feature elimination (5/65, 7.7%), information-gain ratio (6.2%), and SHAP-based methods (4/65, 6.2%) trailing behind.

Finally, dimensionality reduction and specialized feature extraction remain fringe techniques: sequential floating forward selection was used in only 7.7% (5/65) of papers and principal component analysis (PCA) in 6.2% (4/65), text vectorization methods (eg, N-grams, TF-IDF) in 4.6% (3/65), domain-specific statistical features in 3.1% (2/65), and acoustic‐signal processing (MFCC) in just 1 study (1/65, 1.5%). For a comprehensive overview of dataset characteristics used in the studies, refer to Multimedia Appendix 5.

XSL•FO
RenderX

Table . Characteristics of datasets used in the included studies.

Preprocessing techniques	Studies, n/N (%)	References
Dimensionality reduction techniques		
Sequential floating forward selection and SHAP[a]	5/65 (7.7)	[51,56,71,75,82]
Principal component analysis	4 /65 (6.2)	[36,52,77,81]
Others (each one 1)[b]	1/65 (1.5)	[33,57,64,66,67,72]
Feature extraction[c]		
Psycholinguistic and N-gram text vectorization	3/65 (4.6)	[29,33]
Domain-specific statistical	2/65 (3.1)	[51,56]
Acoustic signal feature extraction	1/65 (1.5)	[31]
Feature selection		
Regularization of the model		
Pearson correlation	8/65 (12.3)	[33,36,41,44,45,48,54,61]
Spearman rank filtering	2/65 (3.1)	[58,59]
Chi-square independence test	2/65 (3.1)	[21,65]
Cox proportional hazards and Kaplan-Meier survival analysis	1/65 (1.5)	[24]
Wrapper and tree-based selection		
Gini importance/mean decrease in impurity	12/65 (18.5)	[22,23,25,40,46,52,71,73,82]
Recursive feature elimination with cross-validation	5/65 (7.7)	[30,40,64,73,74]
Entropy-based information gain ratio	4/65 (6.2)	[25,30,41,46]
SHAP value-based importance via differential evolution	4/65 (6.2)	[37,43,62,63]
Other metaheuristic and ensemble filters[d]	1/65 (1.5)	[64,71,75,84]
Encoding approaches		
Label encoding	19/65 (29.2)	[23-25,29,31,33,38,42,48,50,51,53,61,63,64,66-68,73]
One-hot encoding	9/65 (13.9)	[26,27,32,36,41,43,52,56,65]
Binary encoding	6/65 (9.2)	[22,34,40,50,74,76]
Dummy encoding	4/65 (6.2)	[47,70,78,84]
Target encoding	2/65 (3.1)	[41,59]
Handling unbalanced data		
Manual resampling	15/65 (23.1)	[23,29,31,47,50,52,53,56,58,60,61,64,70,71,73]
Class weighting and cost-sensitive	4/65 (6.2)	[55,63,69]
Random oversampling (eg, SMOTE[e] variants)	3/65 (4.6)	[36,56,76]
Handling missing data		
Statistical imputation (mean/median/KNN[f]/MICE[g])	29/65 (44.6)	[22-24,31,33,34,36,38,41,43,44,46,51-54,58,60,62,65-67,70,73,78,80,82,84,85]
Complete case analysis (listwise deletion)	22/65 (33.8)	[23,27-29,34,36-38,40,45,47,50,51,54,64,74-79,81]
Feature transformation techniques		
Min–max scaling and Z-score standardization	51/65 (78.5)	[22-24,29-32,34,36-38,40,41,43,49-52,54,56,58,59,67-71,74-76,78-82,84,85]
Text tokenization, lemmatization, and stop-word removal	23/65 (35.4)	[26,27,29,32,33,36,40,41,47,50-52,64,65,67,72]

XSL•FO
RenderX

Preprocessing techniques	Studies, n/N (%)	References
Log/power transforms (Box-Cox, Yeo-Johnson)	9/65 (13.8)	[25,31,40,43,45,46,66,70]
Polynomial and interaction feature generation	9/65 (13.8)	[25,31,40,43,45,46,66,70]

[a]SHAP: SHapley Additive exPlanations to interpret model.

[b]Others include linear discriminant analysis (LDA), t-SNE, latent semantic analysis (LSA), latent Dirichlet allocation (LDA), spatial feature extraction, and relief algorithm.

[c]Psycholinguistic text vectorization includes N-gram characteristics; linguistic inquiry and word count for emotion, cognition, social content; LDA topics; and TF-IDF. Acoustic signal feature extraction includes MFCC, spectral contrast, and chroma.

[d]Others include bagging-based selection-by-filter methods, sequential floating forward selection, sequential forward selection, relief algorithm, and Boruta algorithm.

[e]SMOTE: Synthetic Minority Oversampling Technique.

[f]KNN: K-Nearest neighbor algorithm.

[g]MICE: Multiple Imputation by Chained Equations to handle missing data.

Characteristics of Features Used in Included Studies

The reviewed studies incorporated 9 distinct categories of data in AI model development (Table 4). Sociodemographic data were most frequently used (49/65, 75.4% of studies), followed by psychological data (44/65, 67.7% of studies), obstetric data (36/65, 55.4% of studies), and behavioral data (23/65, 35.4% of studies). The number of features used varied significantly across studies, ranging from 2 to 988, with a mean average of 44.88 (SD 129.72). Notably, nearly two-thirds of the studies (43/65, 66.2%) used fewer than 26 features.

Within each data type, the most commonly used individual features were age for sociodemographic data (37/65, 56.9% of studies), mode of delivery for obstetric data (15/65, 23.1% of studies), maternal anxiety for psychological data (13/65, 20% of studies), breastfeeding status for behavioral data (11/65, 16.9% of studies), linguistic inquiry and word count (LIWC) features—such as positive emotions (“happy”), cognitive processes (“think”), and personal pronouns (“I” and “we”)—for linguistic data (11/65, 16.9% of studies), metabolic pathways and circulating markers for biomarker data (4/65, 6.2% of studies), newborn gender for neonatal data (11/65, 16.9% of studies), hypertensive disorders for medical history data (11/65, 16.9% of studies), and tweet metadata for sensor-based data (3/65, 4.6% of studies). Additional characteristics of the datasets used in the reviewed studies are shown in Multimedia Appendix 6.

XSL•FO
RenderX

Table . Characteristics of features used in the included studies.

Features characteristics	Studies, n/N (%)	References
Data type[a]		
Sociodemographic	49/65 (75.4)	[21-24,28,30-32,34-36,38,40-50,52,53,55,56,58,66,67-71,74,76-78,80,82-85]
Psychological	44/65 (67.7)	[21-24,26,28-30,34-36,38,40-46,48,50-54,57,60,66,67,70-74,76,78,80,82-84]
Obstetric	36/65 (55.4)	[22-24,30,34,36,41,43,44,46-49,51,55-61,63,64,67-71,74,76-78,80,83-85]
Behavioral	23/65 (35.4)	[22,25,32,34,36,42-44,47,50,51,53,55,56,58-61,63,64,67,80,82]
Medical history	17/65 (26.2)	[22,23,43,46,47,49,56,63,67,68,71,76,78,80,83-85]
Neonatal	16/65 (24.6)	[22,28,30,38,47,50,53,56,61,63,64,70-72,78,85]
Linguistic	9/65 (13.9)	[26,27,29,31,33,39,66,67,75]
Biomarkers	7/65 (10.8)	[46,57,68,69,77,79,81]
Sensor-based	5/65 (7.7)	[37,44,51,60,66]
Number of features[b]		
Mean (SD)	44.9 (129.7)	[21-85]
Range	2 - 988	[21-85]
Feature range		
≤25	43/65 (66.2)	[21,23-29,31-33,35-38,40-42,45,49-52,54-60,65-71,73,74,78,79,82,85]
26 - 50	16/65 (24.6)	[22,30,34,43,47,48,53,61,63,64,72,76,77,80,83,84]
>50	6/65 (9.2)	[39,44,46,62,75,81]
Data input features sociodemographic data[a]		
Age	37/65 (56.9)	[21-24,28,30-32,34,35,38,40,41,43,45,47-51,53,55,56,61,63,66,68,71,74,78,80,82,85]
Education level	21/65 (32.3)	[22,24,28,30,32,34,36,41,46,48,58-61,63,64,68,74,80,82,85]
Marital status	20/65 (30.8)	[22,30,34,41,43,46-48,53,58,59,61,63,64,68,70,76,83-85]
Monthly income	13/65 (20)	[30,34,36,38,41,50,60,61,64,70,78,82,85]
Employment status	11/65 (16.9)	[22,28,30,43,46,48,61,63,64,70,85]
Obstetric data[a]		
Mode of delivery	15/65 (23.1)	[30,32,34,41,47,48,55,61,68,74,78,80,83-85]
Parity	11/65 (16.9)	[22,30,36,43,46,63,64,68,78,80,85]
Gestational age	9/65 (13.9)	[24,30,34,47,49,61,68,71,78]
Gravida	7/65 (10.8)	[24,30,49,60,68,83,84]
Obstetric complications	6/65 (9.2)	[23,34,43,61,69,85]
Psychological data[a]		
Maternal anxiety	13/65 (20)	[21,25,27,35,40,45,48,52,62,65,71,76,83]
Depression history	12/65 (18.5)	[22,30,34,41,43,44,48,53,55,63,69,82]
Feeling of guilt	9/65 (13.9)	[21,25,27,35,40,45,52,65,73]
Feeling sad	8/65 (12.3)	[21,27,35,40,45,52,54,65]
Sleeping disorders	7/65 (10.8)	[25,27,46,47,53,54,62]
Behavioral data[a]		
Breastfeeding status	11/65 (16.9)	[23,28,34,47,48,53,56,61,64,78,85]
Problems bonding with baby	9/65 (13.9)	[21,35,40,45,48,52,61,65,73]
Planned pregnancy	8/65 (12.3)	[32,34,48,53,61,74,80,85]
Smoking status	7/65 (10.8)	[22,23,46,47,60,63,64]
Alcohol use	6/65 (9.2)	[22,36,46,47,63,64]

XSL•FO
RenderX

Features characteristics	Studies, n/N (%)	References
Linguistic data[a]		
LIWC[c] features	11/65 (16.9)	[27]
Speech and acoustic	8/65 (12.3)	[31]
Emotional and cognitive expression	4/65 (6.2)	[33]
Language models	2/65 (3.1)	[39]
Tweet attributes language	2/65 (3.1)	[66]
Biomarker data[a]		
Metabolic pathway	4/65 (6.2)	[81]
Circulating biomarkers	4/65 (6.2)	[46,57,68]
Neurological	3/65 (4.6)	[69,79]
Protein-related	2/65 (3.1)	[77]
Genetic/epigenetic	1/65 (1.5)	[57]
Neonatal data[a]		
Newborn gender	11/65 (16.9)	[30,31,38,47,50,58,59,61,64,70,85]
Birth weight	8/65 (12.3)	[30,34,41,47,64,71,77,78]
Preterm birth	6/65 (9.2)	[43,48,58,59,76,77]
Health of baby	4/65 (6.2)	[28,30,53,85]
Apgar scores	3/65 (4.6)	[47,71,78]
Medical history[a]		
Hypertension disorders	111/65 (16.9)	[22,43,47,49,56,63,76,78,80,83,84]
Gestational diabetes	44/65 (6.2)	[43,49,78,80]
Migraine	44/65 (6.2)	[22,63,83,84]
Preeclampsia	33/65 (4.6)	[43,83,84]
Hypothyroidism	33/65 (4.6)	[83-85]
Sensor-based[b]		
Tweet metadata	33/65 (4.6)	[66]
Activity intensity	33/65 (4.6)	[37,60]
Calories burned	11/65 (1.5)	[37]
Heart rate	11/65 (1.5)	[37]

[a]The number of studies does not add up as certain features are reported in multiple studies within each category, resulting in repeated counts.

[b]The sensor-based category includes only 4 features.

[c]LIWC: linguistic inquiry and word count.

Characteristics of AI Techniques

As shown in Table 5, the most included studies (52/65, 80%) used AI models for predicting PPD (ie, identifying women at risk of developing PPD in the future), while 14 out of 65 (21.5%) studies used them for detection (ie, identifying whether a woman is currently experiencing PPD). Most studies leveraged ML techniques (57/65, 87.7%), whereas DL techniques were applied in 11 out of 65 (16.9%) studies. The predominant application of AI models was in classification tasks (eg, identifying the presence, absence, or severity level of PPD). In contrast, 5 out of 65 (7.7%) studies used AI models for regression tasks (eg, detecting EPDS score). Various AI algorithms were used in the included studies, with random forest (RF) being the most common (29/65, 44.6%), followed by support vector machine (26/65, 4%) and logistic regression (LogR) (23/65, 35.4%).

The most frequently used optimization strategy among the included studies was stochastic gradient descent (9/65, 13.9%), followed by Adam (7/65, 10.8%) and learning rate scheduling (6/65, 9.2%). The most applied regularization and model stabilization techniques were L1/L2 regularization (9/65, 13.9%) and grid search (9/65, 13.9%). To validate AI model performance, both k-fold cross-validation and holdout validation were the most widely adopted approaches (32/65, 49.2%).

XSL•FO
RenderX

Accuracy was the most reported performance metric (49/65, 75.4%), followed by sensitivity (48/65, 73.9%) and area under the curve (AUC) (41/65, 63.1%). Additional characteristics of the Model of Characteristics of AI Techniques used in the reviewed studies are shown in Multimedia Appendix 7.

XSL•FO
RenderX

Table . Characteristics of artificial intelligence techniques used in the included studies.

Characteristics	Studies, n/N (%)	References
AI[a] algorithm aim		
Prediction	52/65 (80)	[21-27,29,30,32-36,39-47,49,50,52-55,57-60,62,64-67,72-74,76-78,80,85]
Detection	14/65 (21.5)	[28,31,37,38,48,51,56,61,63,69-71,75,79]
AI category		
Machine learning	57/65 (87.7)	[21-25,29-39,41-62,64,66-70,73-85]
Deep learning	11/65 (16.9)	[26-28,32,40,41,58,59,63,65,71]
Transfer learning	3/65 (4.6)	[26,40,41]
Natural language processing	3/65 (4.6)	[33,39,72]
Reinforcement learning	2/65 (3.1)	[63,69]
Problem solving approach		
Classification	60/65 (92.3%)	[21-35,37-43,45-53,55-67,69-76,78-85]
Regression	5/65 (7.7%)	[36,44,54,68,77]
AI model type		
Ensemble methods		
Bagging		
Random forest	29/65 (44.6)	[21,24,33,34,37,41,43,47,48,51,52,55,56,58,61,64,65,71,73,74,76,78,80,82,84,85]
Bagging	1/65 (1.5)	[48]
Extreme random trees	1/65 (1.5)	[73]
Extra trees	1/65 (1.5)	[22]
Boosting		
XGBoost	15/65 (23.1)	[21,24,30,36,41,43,45,55,61,65,68,76,80,84,85]
Gradient boosting	10/65 (15.4)	[22,23,30,51,56,60-62,68,73]
AdaBoost	9/65 (13.9)	[33,41,45,48,53,56,64,65,73]
CatBoost	6/65 (9.2)	[35,41,45,65,68,73]
LightGBM	4/65 (6.2)	[35,45,65,68]
Stacking		
Stacking ensemble	1/65 (1.5)	[21]
Stacking model	1/65 (1.5)	[52]
Nested stacking	1/65 (1.5)	[73]
Neural networks		
Multilayer perceptron	11/65 (16.9)	[43,52,56,67,70,84]
Recurrent neural	6/65 (9.2)	[26,27,40,41,66]
Convolutional neural	5/65 (7.7)	[28,31,36,64,80]
Natural language processing	1/65 (1.5)	[72]
Classification models		
Support vector machine	26/65 (40)	[21,26,29,32,33,37,38,43,47,49-51,53,56-61,64,75,76,78,79,82,85]
Decision tree	18/65 (27.7)	[21,24,25,30,41,46,48,49,51-53,56,60,65,76,78,84,85]
K-Nearest neighbors	9/65 (13.9)	[21,37,49,51,52,56,64,65,78]
Recursive partitioning	1/65 (1.5)	[64]
Probabilistic classification		
Logistic regression	23/65 (35.4)	[21,23,29,33,38,42-44,47,50,53,55,56,61,64,65,70,74,76,77,83-85]
Naive Bayes	7/65 (10.8)	[32,34,38,50,53,60,64]

XSL•FO
RenderX

Characteristics	Studies, n/N (%)	References
Linear regression models		
Ridge regression	6/65 (9.2)	[22,34,44,47,52,68]
LASSO regression	5/65 (7.7)	[22,34,39,44,77]
Elastic net	5/65 (7.7)	[23,36,44,47,68]
Support vector regression	2/65 (3)	[68]
Kernel regression	1/65 (1.55)	[68]
Optimization strategies (gradient-based optimization)		
Stochastic gradient descent	9/65 (13.9)	[27-29,32,36,40,56,64,66]
Adam	7/65 (10.8)	[27,32,36,41,48,56,66]
Learning rate scheduling	6/65 (9.2)	[23,27,29,32,41,56]
AdamW	2/65 (3.1)	[28,40]
Cosine Annealing	2/65 (3.1)	[40,48]
Momentum	1/65 (1.5)	[36]
Regularization and model stabilization		
L1/l2 regularization	9/65 (13.9)	[23,28,29,36,42,44,76,77,83]
Grid search	4/65 (6.2)	[43,65,79,81]
Dropout	8/65 (12.3)	[27,36,40,41,48,56,59,64]
Batch normalization	3/65 (4.6)	[36,40,41]
Early stopping	2/65 (3.1)	[28,36]
Weight decay	2/65 (3.1)	[36,56]
Osprey optimization	1/65 (1.5)	[67]
Validation techniques		
K-fold cross-validation	32/65 (49.2)	[22,23,25,29,30,35,37,39,43,47-49,52,53,61-65,68,69,71,73-77,79,82-84]
Holdout validation	32/65 (49.2)	[21,24,26-29,31-34,36,38-43,45-47,50,51,53-56,58-60,70,78,81]
Leave-one-out cross-validation	1/65 (1.5)	[57]
Nested cross-validation	1/65 (1.5)	[44]
ML[b] performance measures		
Accuracy	49/65 (75.4)	[21,22,24,26-35,38,40,41,43-46,48,49,52,54-67,69-71,73,74,76-80,82,85]
Sensitivity	48/65 (73.9)	[21,22,24-26,29-35,37,38,40-46,49-65,67,70,71,73,75,76,79,82-84]
AUC[c]	41/65 (63.1)	[22-25,31,34,37-51,53,55-57,59-61,64,70,74-84]
Precision	36/65 (55.4)	[21,22,24,26,29-35,37,40,41,43-45,50,53,56,57,59-62,64,65,67,73,78,82-85]
Specificity	23/65 (35.4)	[22,30,31,34,37,38,42,44-46,48,51,55,63,64,70,71,73,75,76,79,82,84]
Geometric mean	7/65 (10.8)	[38,50,51,61,62,69,70]
Negative predictive value	7/65 (10.8)	[22,32,34,44,45,82,84]
F_1-score	5/65 (7.7)	[21,35,58,69,73]
Root mean squared error/mean squared error	3/65 (4.6)	[36,58,68]

[a]AI: artificial intelligence.

[b]ML: machine learning.

[c]AUC: area under the curve, ROC-AUC (receiver operating characteristic) that plots true-positive rate against false-positive rate at different threshold settings.

Among the AI models evaluated in Table 6, ensemble methods emerged as the top performers, with an average accuracy of 93.4%, an F_1-score of 92.5%, and an AUC of 89.4%. Among gradient boosting techniques, CatBoost achieved the highest

XSL•FO
RenderX

AUC of 98.6%, alongside robust accuracy and F_1-score metrics. LightGBM also demonstrated strong performance, recording 92.6% accuracy, an F_1-score of 87.8%, and an AUC of 91.1%, highlighting its scalability and effectiveness. XGBoost delivered competitive results, with an accuracy of 89.1% and an AUC of 86.8%. Convolutional neural networks (CNNs) showed excellent performance as well—particularly in accuracy (92%) and F_1-score (95.1%)—although they were evaluated in a smaller subset of studies.

Traditional tree-based models, including RFs and recursive partitioning, also showed moderate to strong performance. RFs achieved an average accuracy of 80.5% and an AUC of 82.4%, while broader tree-based classifiers averaged 82.8% accuracy and 82.6% AUC. Recursive partitioning, however, showed lower accuracy (71.8%) and AUC (74.7%) across the few studies assessed.

Across all models included, the mean performance was 81.7% (SD 11.05) for accuracy, 80.51% (SD 15.44) for F_1-score, and 81.0% (SD 12.0) for AUC. Collectively, these findings underscore the strong predictive capabilities of DL architectures and ensemble-based approaches—especially boosting models—in detecting PPD, consistently outperforming conventional ML algorithms across most evaluation metrics or detailed information on performance metrics (accuracy, F_1-score, and AUC) (Multimedia Appendix 8).

XSL•FO
RenderX

Table . Accuracy, F_1-score, and area under the curve of artificial intelligence models used in postpartum depression prediction.

Metrics[a]	Accuracy			F_1-score			AUC[b]		
Model	Studies, n	Mean (SD)	Range	Studies, n	Mean (SD)	Range	Studies, n	Mean (SD)	Range
Random forest	23	80.5 (9.2)	59-96	19	80.9 (13.7)	39.3-95	25	82.4 (9.2)	65.1-98
Decision trees	16	82.8 (9.1)	68.8-98.1	12	83.9 (10.5)	66-98.58	13	82.6 (7.9)	69-97.6
XGBoost	13	89.1 (10.3)	67.6-100	8	91.8 (11.4)	66-100	7	86.8 (11.1)	73-100
LightGBM	6	92.6 (11)	70.2-98.4	6	87.8 (22.7)	41.6-98.6	4	91.1 (12.3)	72.7-98
AdaBoost	7	78.7 (9.6)	66-94	6	75.4 (11.6)	58-89	6	78.5 (5.7)	69-85.7
CatBoost	5	93.6 (9.5)	77-99.46	5	90.5 (11.1)	72-99.1	3	98.6 (0.5)	98-99
Gradient Boosting	8	79.2 (9.5)	67-79	7	76.7 (15.6)	45-92	10	86.8 (10.4)	70-97.3
Linear regressions	8	70.9 (4.9)	67-79	2	76.5 (0.7)	76-77	14	77.2 (5.3)	67-87
Logistic regressions	16	75.3 (7.3)	65.5-94.3	9	72.1 (16)	38.2-94.1	21	81.9 (9.6)	69.6-97
Naive Bayes	9	74 (6.6)	67.5-86.4	18	73.4 (13.2)	56-88.7	10	76.8 (7.1)	65.6-92
SVMs[c]	18	80 (8.4)	64-94.9	13	77.5 (14.1)	42.2 - 94	18	78.7 (6.9)	64.2-90.3
KNNs[d]	8	79.5 (13)	61.5 - 97	8	78.3 (13.9)	57 - 97	7	79.3 (9.5)	61.5-88.2
Neural networks	3	79.6 (14.4)	65-93.75	4	80.1 (15.1)	65.1-95.2	6	64.8 (24.8)	31.2-90.8
MLPs[e]	9	81.7 (8.3)	68-92	3	73.6 (24)	40.6-91.7	12	74.9 (17.3)	31.2-91.2
ANNs[f]	6	85.3 (10.8)	70.7-97.1	3	85.9 (12.3)	71.7-93	3	70.6 (6.3)	66-77.79
CNNs[g]	5	92 (8.1)	77.3-100	4	95.1 (3.9)	91.1-100	N/A[h]	N/A	N/A
Reinforcement learning	3	85.4 (4.1)	81-89.07	3	84.2 (4.7)	79.2-88.4	7	86.6 (3.8)	83-90.66
Ensemble models	9	93.4 (10.8)	65-99.84	9	92.5 (15.3)	52-99.21	7	89.4 (16)	56.5-98.95
Overall	174	81.7 (11.1)	59-100	141	80.5 (15.4)	38.2-100	176	81.0 (12)	31.2-100

[a]Models are grouped into tree-based, boosting, probabilistic, traditional machine learning, neural networks, and ensembles. Metrics are mean (SD). Study counts refer to the number of models reporting accuracy, F_1-score, or AUC—not the number of references.

[b]AUC:area under the curve.

[c]SVMs: Support Vector Machines.

[d]KNNs: K-Nearest Neighbor algorithm.

[e]MLPs: multilayer perceptrons.

[f]ANNs: Artificial Neural Network.

[g]CNNs: Convolutional Neural Networks.

[h]N/A: not applicable.

Discussion

Principal Findings

This scoping review examines the evolving application of AI in PPD research, with approximately 80% of studies prioritizing early prediction over detection. This reflects a growing awareness of AI's potential to enable proactive mental health interventions.

ML algorithms dominated (87.7%), suggesting a preference for structured data handling and model interpretability. Classical models such as RF (44.6%), LogR (35.4%), and XGBoost (23.1%) were especially prevalent, likely due to their ease of implementation, strong performance on tabular datasets, and alignment with the interpretability demands in health care as these models are well suited for structured and tabular clinical data (eg, demographics, EPDS scores, and EHR) and offer high interpretability, a core requirement in health care settings for clinical transparency and trust. In contrast, DL

XSL•FO
RenderX

approaches—while more capable of handling complex, high-dimensional inputs—were used in only 16.9% of studies, indicating underutilization of architectures such as convolutional or recurrent neural networks (RNNs). This aligns with the nature of DL methods (CNNs, RNNs, and transformers) that require large, high-dimensional datasets (eg, text, electroencephalogram, and sensor signals), which were rare in the reviewed studies; also, DL models are less interpretable, a major limitation in mental health where explainability is critical for clinician adoption. NLP and reinforcement learning (RL) were rarely used, despite their potential for analyzing unstructured clinical notes and dynamic decision-making, respectively. This aligns with the fact that NLP is suitable for analyzing unstructured clinical notes, patient narratives, or social media data, and few studies had access to such datasets. In addition, RL's strength in dynamic decision-making (eg, treatment adjustment over time) is difficult to apply in static, retrospective datasets common in PPD research.

More than 90% of studies focused on classification tasks—categorizing individuals as at risk or not for PPD—while only a few adopted regression models to estimate continuous risk levels. Although classification supports clinical decision-making, regression can offer more granular risk assessments, useful for personalized interventions and monitoring symptom trajectories. This aligns with real-world clinical workflows—binary screening tools are common. However, the underuse of regression models (predicting continuous risk) limits personalization and longitudinal risk tracking.

The optimization strategies such as Adam or stochastic gradient descent optimizers, learning rate scheduling, and dropout were rarely reported across the literature. This may reflect that most studies used classical ML, which does not require such parameters or limited technical expertise or a focus on classical methods over deep architectures. Regularization practices such as L1/L2 penalties, batch normalization, and early stopping were underused, despite their importance for model generalizability and performance stabilization. These techniques help prevent overfitting and improve generalizability. Their absence may reflect limited ML maturity or reliance on default model settings without tuning. Moreover, model validation techniques varied considerably. Although nearly half of the reviewed studies used k-fold cross-validation or holdout validation, external validation was seldom implemented. External validation requires access to independent datasets, which are often unavailable due to privacy constraints in mental health and raising concerns about generalizability. Performance evaluation also lacked consistency, with accuracy (75.4%) and sensitivity (73.9%) reported most frequently, while key metrics such as specificity, AUC, and F_1-score were less commonly disclosed, which are essential for imbalanced datasets such as PPD.

By considering the results of performance evaluation across accuracy, F_1-score, and AUC, they indicate that ensemble models, especially boosting techniques such as CatBoost, LightGBM, and XGBoost, consistently outperformed other AI methods in predicting PPD. Their high accuracy and AUC reflect strong generalization and robustness, owing to their ability to iteratively correct misclassifications and capture complex, nonlinear patterns—particularly valuable in noisy, imbalanced health care datasets.

CatBoost led with an AUC of 98.6%, benefiting from its advanced handling of categorical variables and built-in overfitting control, making it highly suited for structured health data. LightGBM followed closely, offering high accuracy (92.6%) and efficiency due to its gradient-based sampling and fast training, making it ideal for large-scale or real-time applications. XGBoost also performed competitively (89.1% accuracy and 86.8%) and remains popular for its transparency and feature importance tools.

In contrast, traditional decision tree–based classifiers such as RFs and generic tree–based models achieved moderate performance, with accuracy ranging from 80.5% to 82.8% and AUC values near 82%. While interpretable and computationally efficient, these models lacked the advanced learning mechanisms of boosting methods. Recursive partitioning methods, evaluated in only 2 studies, performed the weakest among tree-based approaches.

Overall, the mean model performance—accuracy: 81.7% (SD 11.05), F_1-score: 80.51% (SD 15.44), and AUC: 81.0% (SD 12.0)—shows variability likely due to differences in datasets, preprocessing, and validation strategies. This underscores the need for standardized evaluation and external validation to ensure model reproducibility and clinical reliability. Also, this suggests limited awareness or inconsistent standards in performance reporting and hinders meaningful comparisons and meta-analysis across studies.

This inconsistency complicates comparative assessments across models and highlights the need for standardized evaluation frameworks.

This scoping review offers the first comprehensive synthesis of both foundational and advanced preprocessing techniques used in AI-driven PPD studies. The high prevalence of basic normalization methods—such as Min-Max scaling and Z-score standardization, reported in 78.5% of studies—demonstrates a broad consensus on the need to standardize input features, particularly in ML models sensitive to feature magnitude. These practices are foundational for ensuring convergence stability and improving model performance, especially in algorithms such as LogR and k-nearest neighbors. However, more advanced preprocessing techniques were markedly underutilized, limiting the full potential of AI in PPD prediction.

For example, although missing data are ubiquitous in real-world health care datasets, only 44.6% of studies applied imputation techniques to address it. The remaining studies either dropped missing values or excluded incomplete cases—approaches that risk reducing sample size and introducing systematic bias, particularly in psychiatric populations where follow-up and self-report compliance can vary. Likewise, class imbalance, a well-documented issue in mental health data (eg, more controls than PPD cases), was insufficiently addressed: only 23.1% of studies used resampling methods such as SMOTE, and an even smaller fraction (6.2%) incorporated cost-sensitive learning,

XSL•FO
RenderX

which could improve model fairness and reduce false negatives—an important consideration in screening contexts.

In terms of categorical variable processing, label encoding (29.2%) and one-hot encoding (13.9%) were commonly used. While these methods are simple to implement, they may introduce ordinal bias or dimensionality explosion, respectively. More efficient encoding schemes (eg, target encoding and frequency encoding) that better preserve categorical relationships were rarely used, reflecting either limited awareness or concerns over interpretability.

Advanced feature engineering and selection techniques were also underexploited. Tree-based feature selection was used in only 18.5% of studies, despite its use in identifying nonlinear relationships and reducing overfitting. Dimensionality reduction methods such as PCA (6.2%) and interpretability tools such as SHAP (6.2%) were seldom implemented, limiting transparency and the ability to uncover key risk factors. Furthermore, multimodal or unstructured data processing techniques—including text or acoustic feature extraction—were applied in fewer than 5% of studies, despite their relevance in analyzing patient interviews, social media posts, or voice biomarkers.

In summary, while basic preprocessing steps have become standard practice, the limited adoption of more sophisticated strategies reflects a missed opportunity to enhance model robustness, generalizability, and interpretability. These gaps underscore the need for broader methodological literacy and the integration of more nuanced preprocessing pipelines tailored to the complexity and heterogeneity of PPD data.

Geographically, North America led the research landscape, with the United States contributing the largest share—just less than two-thirds. Asia also featured prominently, especially China, Bangladesh, and India. This wide participation demonstrates the global relevance and flexibility of AI solutions in diverse health care systems. However, it also reveals disparities in research capacity, underscoring the need for more contributions from underrepresented regions to ensure global equity. Bangladesh's notable presence is largely due to the use of public datasets, showing how open access data can significantly influence research output. The included studies span from 2009 to 2025, with publication volume rising steadily. A quarter of the studies were published in 2024 alone, reflecting growing global interest, better access to digital health data, and advances in AI. The lower count from 2025 is likely due to early-year data collection. Most studies appeared in peer-reviewed journals (more than two-thirds), while fewer were conference papers (less than one-third), and only 1 was a dissertation—illustrating both the academic rigor and fast-evolving nature of this field. Sample sizes varied widely, ranging from 11 to 5,73,634 participants (mean 18,187.4), yet nearly half of the studies had fewer than 5000 participants, raising concerns about generalizability and model performance. Among the 26 studies reporting participant age, the average was 31.08 years—consistent with the typical childbearing population—although the lack of demographic transparency in many studies limits comparability and clinical applicability. Among studies reporting participant age, the mean average was 31.08 (SD 3.42) years, aligning with the typical reproductive age. However, more than half of the studies did not report age, limiting comparability and model applicability across age groups.

Sample sizes varied greatly—from less than a dozen to more than half a million—with most studies enrolling fewer than 5000 participants. Smaller studies often used surveys or interviews, while larger ones relied on national registries. This underscores the importance of large datasets, especially for training DL models. Closed-source datasets dominated, with only about 25% of studies using open data. This limits reproducibility and hinders collaboration. Expanding open access datasets and standardized repositories would improve transparency and accelerate innovation. Most studies were retrospective, drawing on accessible surveys and EHR data. While more than two-thirds used surveys and many depended on structured clinical inputs, a recent shift toward prospective designs reflects growing interest in real-time, high-quality data for AI validation. Use of social media and sensor data is emerging, indicating a move toward passive, continuous monitoring. However, objective biomarkers—such as hormonal, genetic, or neuroimaging data—were underutilized, appearing in only 2 studies, underscoring a missed opportunity for clinical robustness. Moreover, nearly 90% of studies used unimodal inputs (eg, surveys and EHRs), with few incorporating multimodal data such as text, audio, or imaging. This limits the ability to capture the complex biopsychosocial nature of PPD. Research predominantly occurred in health care settings, reflecting strong clinical relevance. Around one-third took place in communities and fewer than 10% in academic contexts. Expanding into diverse settings could improve the inclusivity and generalizability of AI-based PPD interventions. Assessment timing for PPD varied widely, with 12-month evaluations being most common. Studies spanned short-, medium-, and long-term intervals, yet more than 40% failed to specify timing—revealing a major gap in methodological transparency. This variability underscores both evolving perspectives on PPD progression and the need for standardized follow-up periods to enhance comparability, reproducibility, and clinical relevance of AI models. The EPDS is the most widely used reference in PPD research, followed by *International Classification of Diseases* (*ICD*) codes and PHQ-9, highlighting its strong validation in postpartum populations. However, inconsistent use of diagnostic tools across studies hampers comparability and a unified understanding of PPD. While AI models show promise using varied data sources, few are benchmarked against tools such as EPDS or PHQ-9—limiting assessments of their real-world clinical use.

Feature counts varied widely (2-988), with an average of 44.9; nearly two-thirds of studies used fewer than 25 features, indicating a preference for simplicity and interpretability. These findings underscore the need for broader adoption of sophisticated feature selection and dimensionality reduction techniques (eg, PCA, SHAP, and recursive elimination) to enhance predictive performance and clinical relevance. The results of this scoping review underscore the central role of sociodemographic features, which were the most frequently used across included studies on PPD. Variables such as age

XSL•FO
RenderX

(56.9%), education level (32.3%), and marital status (30.8%) were among the most common predictors, highlighting the consistent reliance on structured patient-reported or administrative health records. Obstetric features were the second most common group, particularly mode of delivery (23.1%), parity (16.9%), and gestational age (13.9%). These findings align with literature suggesting that birth experience and maternal clinical history offer critical information in predicting PPD onset. Psychological indicators such as maternal anxiety (20%) and history of depression (18.5%) were also well represented. Their inclusion reflects growing interest in integrating mental health history and current affective symptoms into predictive frameworks. Similarly, behavioral factors such as breastfeeding status and bonding issues were frequently used to enhance emotional and functional contextualization of risk. Less frequently used were linguistic features (eg, LIWC metrics and emotional expression) and biomarkers (eg, epigenetic markers and neurological proteins), suggesting a growing but underutilized frontier. Notably, sensor-derived features (eg, tweet metadata, wearable-derived activity, or heart rate) appeared only in a handful of studies, despite the increasing ubiquity of digital health data. This spectrum of feature types illustrates a multidomain data integration trend, particularly among recent studies that incorporate EHRs, digital behavioral traces, and physiological data to enhance model robustness and precision.

Comparison With Previous Reviews

The findings of this review are broadly consistent with earlier literature, including reviews by Kwok et al [10], Fazraningtyas et al [15], Qi et al [17], Saqib et al [18], Fazraningtyas et al [30], and Acharya et al [13]. These prior studies similarly identified a reliance on retrospective study designs, structured demographic and clinical features (eg, age, parity, and psychiatric history), and traditional ML models such as RFs, support vector machines, and LogR. Most were applied to survey-based or EHR-derived datasets, reflecting the accessibility and interpretability of structured data in maternal mental health contexts.

However, this scoping review extends the current literature in several important ways. First, our review is not focused narrowly only on model types or performance metrics; it even systematically maps the entire AI modeling pipeline, from data characteristics and preprocessing techniques to model training, optimization, and validation strategies. For example, while earlier studies acknowledged preprocessing in general terms, our analysis quantifies the usage of basic methods (eg, scaling and label encoding) and highlights the underuse of advanced techniques such as SMOTE, SHAP, recursive feature elimination, and cost-sensitive learning.

Second, this review identifies the limited adoption of advanced AI methodologies, including DL, transformer-based NLP, and transfer learning, despite their growing success in related health care domains. While Fazraningtyas et al [15] and Qi et al [17] recognized these tools conceptually, few studies in their datasets applied them operationally to PPD detection tasks—an observation confirmed and quantified by our analysis.

Third, our review offers a granular classification of more than 45 features across 9 thematic domains, revealing persistent dependence on sociodemographic and self-reported data. This pattern, while accessible and interpretable, introduces potential biases and limits the generalizability of models. In contrast, passive and objective inputs—such as biosensors, electroencephalogram, speech signals, or real-time behavioral metrics—remain substantially underused, despite their promise for early and noninvasive detection of PPD.

Fourth, unlike previous studies that typically summarized trends descriptively, this review visualizes and quantitatively tracks the growth of literature over time, the global distribution of research output, and the evolution of study design types, using structured frameworks and stacked visualizations. For instance, we highlight that Bangladesh's growing presence is largely driven by the reuse of public datasets, illustrating how open data democratizes research participation.

Finally, this review distinguishes itself by its methodological scope and rigor. It includes studies published through February 2025 across 8 multidisciplinary databases, covering both prospective and retrospective designs, a wide range of countries, and diverse data sources. This comprehensive coverage enables a more nuanced understanding of current capabilities and persistent gaps in AI-based maternal mental health research. In particular, our audit of model regularization, hyperparameter tuning, and evaluation practices offers insight into areas often overlooked by earlier reviews. Taken together, these contributions provide a stronger foundation for the development of transparent, reproducible, and clinically relevant AI tools in PPD research—addressing both methodological blind spots and equity concerns raised in prior literature.

Implication and Further Works

To significantly advance the field of AI-driven PPD research, several key strategic priorities should be addressed. These priorities focus on improving methodological rigor, inclusivity of data, clinical applicability, and ethical implementation.

First, enhance the integration of multimodal and objective data sources. Currently, research predominantly relies on traditional sociodemographic and self-reported survey data. There is an urgent need to incorporate underutilized modalities such as linguistic data (eg, LIWC, sentiment analysis, and acoustic speech patterns), biosignals (eg, heart rate variability and activity monitoring), wearable technology outputs, and biological biomarkers (eg, hormonal, metabolic, genetic, or epigenetic markers). Leveraging these richer data types can significantly enhance the accuracy, personalization, and early detection capabilities of predictive models.

Second, expand and prioritize sharing of open access datasets. Only about 25% of studies included in our review used publicly available datasets, highlighting a substantial barrier to reproducibility, benchmarking, and international collaboration. Developing standardized, large-scale, anonymized datasets should become a priority. Techniques such as federated learning could facilitate collaborative research across different institutions while maintaining data privacy and security.

XSL•FO
RenderX

Third, increase the adoption of longitudinal and prospective research designs. Most existing AI models for PPD prediction are based on retrospective or immediate postpartum data, limiting their ability to capture evolving symptom patterns over time. Incorporating longitudinal data collection into future studies is essential to better understand symptom progression, delayed onset, and relapse scenarios, thus enhancing the clinical relevance and predictive accuracy of AI models.

Fourth, advance multimodal fusion frameworks. Given the complexity of PPD, future models must systematically integrate structured data (eg, EHRs) and unstructured inputs (eg, text, audio, and sensor signals). Developing robust multimodal fusion approaches that effectively combine diverse data sources will significantly enhance model interpretability, clinical effectiveness, and predictive power.

Fifth, standardize preprocessing and feature engineering pipelines. Variability and incomplete reporting in preprocessing methods currently limit model comparability and reproducibility. Adopting standardized protocols for data preprocessing—including feature extraction, transformation techniques, and class imbalance adjustments (eg, SMOTE and cost-sensitive learning)—is necessary. Transparent reporting of these processes should be enforced to enhance scientific rigor and validation.

Sixth, emphasize model explainability and ethical AI practices. Transparent and interpretable AI models are crucial for clinical adoption. Few studies currently apply advanced explainability methods such as SHAP, Local Interpretable Model-Agnostic Explanations, or counterfactual analyses. Integrating these interpretability techniques into AI pipelines will facilitate clinician trust and understanding. Moreover, ethical considerations—such as minimizing algorithmic bias (such as balanced datasets and resampling to correct imbalances in training data)—are considered in few studies, but preventing potential harms from false positives, safeguarding patient autonomy, and ensuring cultural sensitivity should be systematically addressed in all AI model developments.

Seventh, standardize evaluation and reporting metrics. While accuracy is often prioritized, metrics such as specificity, AUC, F_1-score, and precision must be consistently reported to enable comprehensive evaluation and meaningful comparisons across studies. Furthermore, systematic reviews and meta-analyses are required to identify existing methodological inconsistencies, biases, and underrepresented findings to refine future AI-based research approaches.

Eighth, shift from subjective screening tools to objective validation measures. Current studies heavily rely on subjective instruments (eg, EPDS, PHQ, and *ICD* codes), which, despite validation, vary significantly across studies. Future research should validate AI models using objective clinical measures such as physiological indicators and behavioral markers, thus improving reliability and facilitating clinical implementation.

Ninth, the superior performance of ensemble models—particularly boosting techniques such as CatBoost, LightGBM, and XGBoost—suggests that they are promising candidates for clinical implementation in PPD screening. Their ability to consistently achieve high accuracy, F_1-score, and AUC underscores their robustness in handling structured health data, including demographic and clinical features. Given the strong results of CatBoost in handling categorical variables and LightGBM's efficiency in large-scale settings, future research should prioritize evaluating these models in real-world clinical workflows and mobile health platforms, where scalability and interpretability are critical. In addition, since performance varied across studies due to differences in data characteristics and preprocessing strategies, future work should aim to establish benchmark datasets and standardized pipelines for fair comparison.

Efforts should also be made to assess model performance across different population subgroups, ensuring that these algorithms do not inadvertently introduce or amplify bias. Finally, comparative studies should continue to assess whether boosting models maintain their advantage as datasets grow and diversify, particularly in longitudinal or multisite contexts.

Finally, foster global and interdisciplinary collaboration. PPD research remains unevenly distributed globally. Encouraging cross-regional and interdisciplinary collaboration—particularly with underrepresented regions and diverse professional backgrounds such as computer science, psychiatry, public health, and ethics—will foster equitable research practices and drive innovation in maternal mental health care globally.

Limitations

Despite the comprehensive nature of this scoping review, several limitations should be acknowledged; first, we limited our inclusion to studies published in English. This language restriction may have resulted in the exclusion of relevant research published in other languages, particularly from non–English-speaking countries where maternal mental health may be a pressing issue.

Second, we prioritized peer-reviewed and indexed literature from 8 major databases and limited Google Scholar results to the first 100 entries ranked by relevance. Consequently, gray literature, including government reports, dissertations beyond ProQuest, and nonindexed conference proceedings, may have been underrepresented.

Third, our review focused specifically on studies using AI techniques for detection or prediction of PPD. As a result, studies that applied AI for monitoring, treatment delivery, or resource allocation in maternal mental health were excluded, which narrows the scope of applicability.

Finally, we did not conduct a quantitative meta-analysis or risk of bias assessment, as these are typically outside the scope of scoping reviews. Consequently, while we mapped methodological patterns and gaps, we did not evaluate effect sizes, statistical heterogeneity, or study-level quality in a standardized manner.

Conclusions

This scoping review comprehensively maps the application of AI in PPD research, analyzing 87 studies published between 2009 and 2025. The review identifies a predominant emphasis on early prediction (~80%) over detection, with ML

XSL•FO
RenderX

methods—particularly RF (44.6%), LogR (35.4%), and XGBoost (23.1%)—used in 87.7% of studies. These models were favored for their compatibility with structured clinical data and interpretability. DL approaches, including CNNs and RNNs, were underutilized (16.9%), reflecting data limitations and interpretability concerns. NLP and RL were rarely applied, mirroring limited access to unstructured or sequential data sources.

More than 90% of studies focused on classification tasks, aligning with standard clinical workflows, while regression approaches remained limited. Basic preprocessing practices, such as normalization, were widely adopted (78.5%), but advanced strategies—such as imputation (44.6%), resampling (23.1%), cost-sensitive learning (6.2%), and feature selection techniques such as PCA or SHAP—were inconsistently applied. Most models lacked detailed reporting of optimization strategies or regularization methods, and only half used internal validation. External validation was rarely reported, complicating model comparability.

The comparative analysis between accuracy, F_1-score, and AUC confirms that ensemble learning approaches, particularly boosting algorithms such as CatBoost and LightGBM, consistently outperform traditional models in predicting PPD, achieving superior accuracy, F_1-scores, and AUC values across studies.

Geographic trends showed research dominance by North America, particularly the United States, with notable contributions from Asia, driven by access to public datasets. Most studies used retrospective designs and unimodal inputs—mainly survey or EHR data—while multimodal and objective data (eg, biomarkers and sensor data) were rarely incorporated. Assessment timing, feature selection, and dataset transparency varied widely. Sociodemographic and obstetric features were the most frequently used predictors, while linguistic, behavioral, and physiological data were underrepresented. This review offers the first detailed synthesis of preprocessing workflows and feature domains in PPD-AI research, underscoring both progress and methodological gaps across the literature.

Acknowledgments

The authors declare the use of generative artificial intelligence in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing. The GAI tool used was ChatGPT-4. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by all authors. All intellectual content, study design, data collection, analysis, and final interpretations are the sole responsibility of the authors. Multimedia Appendix 9 shows prompts and responses used. No custom code or mathematical algorithm was used in this study.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

AN developed the protocol with guidance from and under the supervision of AAA. AAA searched the electronic databases and conducted backward and forward reference list checking. The study selection, data extraction, and data synthesis process were carried out by MA and AN under the supervision of AAA. MA wrote results, methods, discussion, and conclusion sections. M Alsahli wrote the background section. The paper was revised critically for important intellectual content by all authors under the supervision of AAA. All authors approved the manuscript for publication and agreed to be accountable for all aspects of the work.

Conflicts of Interest

AA-A is an associate editor of *JMIR Nursing* at the time of publication. All other authors declare no conflict of interest.

Multimedia Appendix 1
Search strategy.
[DOCX File, 26 KB - jmir_v28i1e77376_app1.docx]

Multimedia Appendix 2
Data extraction form.
[DOCX File, 24 KB - jmir_v28i1e77376_app2.docx]

Multimedia Appendix 3

XSL•FO
RenderX

Characteristics of each included study.
[DOCX File, 38 KB - jmir_v28i1e77376_app3.docx]

Multimedia Appendix 4
Characteristics of the dataset.
[DOCX File, 78 KB - jmir_v28i1e77376_app4.docx]

Multimedia Appendix 5
Characteristics of preprocessing.
[DOCX File, 33 KB - jmir_v28i1e77376_app5.docx]

Multimedia Appendix 6
Feature characteristics.
[DOCX File, 34 KB - jmir_v28i1e77376_app6.docx]

Multimedia Appendix 7
Characteristics of artificial intelligence.
[DOCX File, 40 KB - jmir_v28i1e77376_app7.docx]

Multimedia Appendix 8
Characteristics of machine learning performances.
[DOCX File, 101 KB - jmir_v28i1e77376_app8.docx]

Multimedia Appendix 9
Prompts and responses.
[DOCX File, 40 KB - jmir_v28i1e77376_app9.docx]

Checklist 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[DOCX File, 110 KB - jmir_v28i1e77376_app10.docx]

References

1. Saharoy R, Potdukhe A, Wanjari M, Taksande AB. Postpartum depression and maternal care: exploring the complex effects on mothers and infants. Cureus 2023 Jul;15(7):e41381. [doi: 10.7759/cureus.41381] [Medline: 37546054]
2. Lopez-Gonzalez DM, Kopparapu AK. Postpartum care of the new mother. In: StatPearls: StatPearls Publishing; 2025. [Medline: 33351433]
3. Howard LM, Khalifeh H. Perinatal mental health: a review of progress and challenges. World Psychiatry 2020 Oct;19(3):313-327. [doi: 10.1002/wps.20769] [Medline: 32931106]
4. Roddy Mitchell A, Gordon H, Lindquist A, et al. Prevalence of perinatal depression in low- and middle-income countries. JAMA Psychiatry 2023 May 1;80(5):425. [doi: 10.1001/jamapsychiatry.2023.0069]
5. Levis B, Negeri Z, Sun Y, Benedetti A, Thombs BD, DEPRESsion Screening Data (DEPRESSD) EPDS Group. Accuracy of the Edinburgh Postnatal Depression Scale (EPDS) for screening to detect major depression among pregnant and postpartum women: systematic review and meta-analysis of individual participant data. BMJ 2020 Nov 11;371:m4022. [doi: 10.1136/bmj.m4022] [Medline: 33177069]
6. Levis B, Benedetti A, Thombs BD, DEPRESsion Screening Data (DEPRESSD) Collaboration. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. BMJ 2019 Apr 9;365:l1476. [doi: 10.1136/bmj.l1476] [Medline: 30967483]
7. Long MM, Cramer RJ, Bennington L, et al. Perinatal depression screening rates, correlates, and treatment recommendations in an obstetric population. Fam Syst Health 2020 Dec;38(4):369-379. [doi: 10.1037/fsh0000531] [Medline: 33119369]
8. Sidebottom A, Vacquier M, LaRusso E, Erickson D, Hardeman R. Perinatal depression screening practices in a large health system: identifying current state and assessing opportunities to provide more equitable care. Arch Womens Ment Health 2021 Feb;24(1):133-144. [doi: 10.1007/s00737-020-01035-x] [Medline: 32372299]
9. Cox J. Use and misuse of the Edinburgh Postnatal Depression Scale (EPDS): a ten point “survival analysis”. Arch Womens Ment Health 2017 Dec;20(6):789-790. [doi: 10.1007/s00737-017-0789-7] [Medline: 29101480]
10. Kwok WH, Zhang Y, Wang G. Artificial intelligence in perinatal mental health research: a scoping review. Comput Biol Med 2024 Jul;177:108685. [doi: 10.1016/j.compbiomed.2024.108685] [Medline: 38838557]

XSL•FO
RenderX

11. Mapari SA, Shrivastava D, Dave A, et al. Revolutionizing maternal health: the role of artificial intelligence in enhancing care and accessibility. Cureus 2024;16(9):e69555. [doi: 10.7759/cureus.69555]
12. Turchioe MR, Hermann A, Benda NC. Recentering responsible and explainable artificial intelligence research on patients: implications in perinatal psychiatry. Front Psychiatry 2023;14:1321265. [doi: 10.3389/fpsyt.2023.1321265] [Medline: 38304402]
13. Acharya A, Ramesh R, Fathima T, Lakhani T, S SK. Clinical tools to detect postpartum depression based on machine learning and EEG: a review. Presented at: 2023 2nd International Conference on Computational Systems and Communication (ICCSC); Mar 3-4, 2023. [doi: 10.1109/ICCSC56913.2023.10142970]
14. Cellini P, Pigoni A, Delvecchio G, Moltrasio C, Brambilla P. Machine learning in the prediction of postpartum depression: a review. J Affect Disord 2022 Jul;309:350-357. [doi: 10.1016/j.jad.2022.04.093]
15. Fazraningtyas WA, Rahmatullah B, Dwi Salmarini D, Arrieya Ariffin S, Ismail A. Recent advancements in postpartum depression prediction through machine learning approaches: a systematic review. Bull EEI 2024;13(4):2729-2737. [doi: 10.11591/eei.v13i4.7185]
16. Kimwomi G, Mgala M, Mwakondo F, Kimeto P. Machine learning prediction models for postpartum depression, a review of literature. IJCATR 2022;11(06):205-212. [doi: 10.7753/IJCATR1106.1005]
17. Qi W, Wang Y, Li C, et al. Predictive models for predicting the risk of maternal postpartum depression: a systematic review and evaluation. J Affect Disord 2023 Jul 15;333:107-120. [doi: 10.1016/j.jad.2023.04.026] [Medline: 37084958]
18. Saqib K, Khan AF, Butt ZA. Machine learning methods for predicting postpartum depression: scoping review. JMIR Ment Health 2021 Nov 24;8(11):e29838. [doi: 10.2196/29838] [Medline: 34822337]
19. Zhong M, Zhang H, Yu C, Jiang J, Duan X. Application of machine learning in predicting the risk of postpartum depression: a systematic review. J Affect Disord 2022 Dec 1;318:364-379. [doi: 10.1016/j.jad.2022.08.070] [Medline: 36055532]
20. Cohen JM. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46. [doi: 10.1177/001316446002000104]
21. Sharma Y, Jain V, Tarwani S. Ensemble machine learning model for predicting postpartum depression disorder. 2024 Presented at: 2024 IEEE Region 10 Symposium (TENSYMP); Mar 3-4, 2023; New Delhi, India p. 1-6. [doi: 10.1109/TENSYMP61132.2024.10752305]
22. Andersson S, Bathula DR, Iliadis SI, Walter M, Skalkidou A. Predicting women with depressive symptoms postpartum with machine learning methods. Sci Rep 2021 Apr 12;11(1):7877. [doi: 10.1038/s41598-021-86368-y] [Medline: 33846362]
23. Betts KS, Kisely S, Alati R. Predicting postpartum psychiatric admission using a machine learning approach. J Psychiatr Res 2020 Nov;130:35-40. [doi: 10.1016/j.jpsychires.2020.07.002] [Medline: 32771679]
24. Sona Ajay C, Juliet S, Alex A. Machine learning and survival analysis models for postpartum depression: a comprehensive risk factor analysis. 2024 Presented at: 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS); Mar 14-15, 2024; Coimbatore, India p. 1340-1345. [doi: 10.1109/ICACCS60874.2024.10717207]
25. Cai M, Wang Y, Luo Q, Wei G. Factor analysis of the prediction of the Postpartum Depression Screening Scale. Int J Environ Res Public Health 2019 Dec 10;16(24):5025. [doi: 10.3390/ijerph16245025] [Medline: 31835547]
26. Carneiro MB, Moreira MWL, Pereira SSL, Gallindo EL, Rodrigues J. Recommender system for postpartum depression monitoring based on sentiment analysis. Presented at: 2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM); Mar 1-2, 2020; Shenzhen, China p. 1-6. [doi: 10.1109/HEALTHCOM49281.2021.9398922]
27. Chen Y, Zhou B, Zhang W, Gong W, Sun G. Sentiment analysis based on deep learning and its application in screening for perinatal depression. Presented at: 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC); Jun 18-21, 2018; Guangzhou, China p. 451-456. [doi: 10.1109/DSC.2018.00073]
28. Fanos V, Dessì A, Deledda L, et al. Postpartum depression screening through artificial intelligence: preliminary data through the Talking About algorithm. J Pediatr Neonatal Individualized Med 2023;12(2):e120222. [doi: 10.7363/120222]
29. Fatima I, Abbasi BUD, Khan S, Al‐Saeed M, Ahmad HF, Mumtaz R. Prediction of postpartum depression using machine learning techniques from social media text. Expert Syst 2019 Aug;36(4). [doi: 10.1111/exsy.12409]
30. Fazraningtyas WA, Rahmatullah B, Naparin H, Basit M, Razak NA. A predictive model for postpartum depression: ensemble learning strategies in machine learning. Indones J Electrical Eng Comput Sci 2025;37(1):443. [doi: 10.11591/ijeecs.v37.i1.pp443-451]
31. Gabrieli G, Bornstein MH, Manian N, Esposito G. Assessing mothers' postpartum depression from their infants' cry vocalizations. Behav Sci (Basel) 2020;10(2):55. [doi: 10.3390/bs10020055]
32. Gopalakrishnan A, et al. A combined attribute extraction method for detecting postpartum depression using social media. Presented at: Health Information Science: 12th International Conference, HIS 2023; Oct 23-24, 2023; Melbourne, VIC, Australia p. 17-29. [doi: 10.1007/978-981-99-7108-4_2]
33. Gopalakrishnan A, Gururajan R, Venkataraman R, et al. Attribute Selection Hybrid Network Model for risk factors analysis of postpartum depression using social media. Brain Inform 2023 Oct 31;10(1):28. [doi: 10.1186/s40708-023-00206-7] [Medline: 37906324]
34. Gopalakrishnan A, Venkataraman R, Gururajan R, Zhou X, Zhu G. Predicting women with postpartum depression symptoms using machine learning techniques. Mathematics 2022;10(23):4570. [doi: 10.3390/math10234570]

XSL•FO
RenderX

35. Gupta V, Tripathi S, Singh D, Bansal A. Predictive algorithms for early postpartum depression detection: CatBoost vs. LightGBM. Presented at: 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO); Oct 14-15, 2024; Noida, India p. 1-4. [doi: 10.1109/ICRITO61523.2024.10522300]
36. Horgen ML. A machine learning approach to understanding depression and anxiety in new mothers. Predicting symptom levels using population-based registry data from a large Norwegian prospective study [Master's Thesis]. : Department of Physics, Faculty of Mathematics and Natural Sciences, University of Oslo; 2020 URL: https://github.com/marialinea/predicting-depression-and-anxiety-moba
37. Hurwitz E, Butzin-Dozier Z, Master H, et al. Harnessing consumer wearable digital biomarkers for individualized recognition of postpartum depression using the All of Us Research Program data set: cross-sectional study. JMIR Mhealth Uhealth 2024 May 2;12:e54622. [doi: 10.2196/54622] [Medline: 38696234]
38. Jiménez-Serrano S, Tortajada S, García-Gómez JM. A mobile health application to predict postpartum depression based on machine learning. Telemed J E Health 2015 Jul;21(7):567-574. [doi: 10.1089/tmj.2014.0113] [Medline: 25734829]
39. Krishnamurti T, Allen K, Hayani L, Rodriguez S, Davis AL. Identification of maternal depression risk from natural language collected in a mobile health app. Procedia Comput Sci 2022;206:132-140. [doi: 10.1016/j.procs.2022.09.092] [Medline: 36712815]
40. Lilhore UK, Dalal S, Faujdar N, et al. Unveiling the prevalence and risk factors of early stage postpartum depression: a hybrid deep learning approach. Multimed Tools Appl 2024;83(26):68281-68315. [doi: 10.1007/s11042-024-18182-3]
41. Lilhore UK, Dalal S, Varshney N, et al. Prevalence and risk factors analysis of postpartum depression at early stage using hybrid deep learning model. Sci Rep 2024 Feb 24;14(1):4533. [doi: 10.1038/s41598-024-54927-8] [Medline: 38402249]
42. Liu H, Dai A, Zhou Z, et al. An optimization for postpartum depression risk assessment and preventive intervention strategy based machine learning approaches. J Affect Disord 2023 May;328:163-174. [doi: 10.1016/j.jad.2023.02.028]
43. Liu Y, Joly R, Reading Turchioe M, et al. Preparing for the bedside-optimizing a postpartum depression risk prediction model for clinical implementation in a health system. J Am Med Inform Assoc 2024 May 20;31(6):1258-1267. [doi: 10.1093/jamia/ocae056] [Medline: 38531676]
44. Lyall LM, Sangha N, Zhu X, et al. Subjective and objective sleep and circadian parameters as predictors of depression-related outcomes: a machine learning approach in UK Biobank. J Affect Disord 2023 Aug 15;335:83-94. [doi: 10.1016/j.jad.2023.04.138] [Medline: 37156273]
45. Marshad I, Islam M, Samin AM, Shomyo M, Nishat MM, Faisal F. Optimizing maternal mental health: a study on boosting algorithms for suicidal tendencies prediction in postpartum depression. Presented at: 2024 International Conference on Inventive Computation Technologies (ICICT); Apr 24-26, 2024; Lalitpur, Nepal p. 1077-1081. [doi: 10.1109/ICICT60155.2024.10544607]
46. Matsumura K, Hamazaki K, Kasamatsu H, Tsuchida A, Inadera H. Decision tree learning for predicting chronic postpartum depression in the Japan Environment and Children's Study. J Affect Disord 2025 Jan 15;369:643-652. [doi: 10.1016/j.jad.2024.10.034] [Medline: 39389121]
47. Matsuo S, Ushida T, Emoto R, et al. Machine learning prediction models for postpartum depression: a multicenter study in Japan. J Obstet Gynaecol Res 2022 Jul;48(7):1775-1785. [doi: 10.1111/jog.15266] [Medline: 35438215]
48. Mazumder P, Baruah S. A community based study for early detection of postpartum depression using improved data mining techniques. 2021 Presented at: 2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS); Dec 16-18, 2021; Bangalore, India p. 1-7. [doi: 10.1109/CSITSS54238.2021.9682941]
49. Moreira MWL, Rodrigues J, Kumar N, Saleem K, Illin IV. Postpartum depression prediction through pregnancy data analysis for emotion-aware smart systems. Information Fusion 2019 May;47:23-31. [doi: 10.1016/j.inffus.2018.07.001]
50. Mustafa N. Use of m-health application to figure out post-natal depression, an evidence-based study. JAMMR 2023;35(24):81-90. [doi: 10.9734/jammr/2023/v35i245326]
51. Myneni S, Zingg A, Singh T, et al. Digital health technologies for high-risk pregnancy management: three case studies using Digilego framework. JAMIA Open 2024 Apr;7(1):ooae022. [doi: 10.1093/jamiaopen/ooae022] [Medline: 38455839]
52. Nasim S, Sami Al-Shamayleh A, Thalji N, et al. Novel meta learning approach for detecting postpartum depression disorder using questionnaire data. IEEE Access 2024;12:101247-101259. [doi: 10.1109/ACCESS.2024.3427685]
53. Natarajan S, Prabhakar A, Ramanan N, Bagilone A, Siek K, Connelly K. Boosting for postpartum depression prediction. 2017 Presented at: 2017 IEEE/ACM International Conference on Connected Health; Jul 17-19, 2017; Philadelphia, PA p. 232-240. [doi: 10.1109/CHASE.2017.82]
54. Osubor VI, Egwali AO. A neuro fuzzy approach for the diagnosis of postpartum depression disorder. Iran J Comput Sci 2018 Dec;1(4):217-225. [doi: 10.1007/s42044-018-0021-6]
55. Park Y, Hu J, Singh M, et al. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. JAMA Netw Open 2021 Apr 1;4(4):e213909. [doi: 10.1001/jamanetworkopen.2021.3909]
56. Paul A, Pragada SD, Murthy DN, Shruthi MLJ, Gurugopinath S. Performance comparison of machine learning techniques for early detection of postpartum depression using PRAMS dataset. 2023 Presented at: 2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN); Dec 22-23, 2023; Bangkok, Thailand p. 310-315. [doi: 10.1109/CICN59264.2023.10402321]

XSL•FO
RenderX

57. Payne JL, Osborne LM, Cox O, et al. DNA methylation biomarkers prospectively predict both antenatal and postpartum depression. Psychiatry Res 2020 Mar;285:112711. [doi: 10.1016/j.psychres.2019.112711] [Medline: 31843207]
58. Prabhashwaree T, Wagarachchi NM. Predicting mothers with postpartum depression using machine learning approaches. 2022 Presented at: 2022 International Research Conference on Smart Computing and Systems Engineering (SCSE); Sep 1, 2022; Colombo, Sri Lanka p. 28-34. [doi: 10.1109/SCSE56529.2022.9905186]
59. Prabhashwaree T, Wagarachchi NM. Towards machine learning approaches for predicting risk level of postpartum depression. 2022 Presented at: 2022 6th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI); Dec 1-2, 2022; Colombo, Sri Lanka p. 1-6. [doi: 10.1109/SLAAI-ICAI56923.2022.10002477]
60. Qasrawi R, Amro M, VicunaPolo S, et al. Machine learning techniques for predicting depression and anxiety in pregnant and postpartum women during the COVID-19 pandemic: a cross-sectional regional study. F1000Res 2022;11:390. [doi: 10.12688/f1000research.110090.1] [Medline: 36111217]
61. Raisa JF, Kaiser MS, Mahmud M. A machine learning approach for early detection of postpartum depression in bangladesh. 2022 Presented at: Brain Informatics: 15th International Conference, BI 2022; Jul 15-17, 2022; Padua, Italy p. 241-252. [doi: 10.1007/978-3-031-15037-1_20]
62. Reps JM, Wilcox M, McGee BA, Leonte M, LaCross L, Wildenhaus K. Development of multivariable models to predict perinatal depression before and after delivery using patient reported survey responses at weeks 4-10 of pregnancy. BMC Pregnancy Childbirth 2022 May 26;22(1):442. [doi: 10.1186/s12884-022-04741-9] [Medline: 35619056]
63. Shen S, Qi S, Luo H. Automatic model for postpartum depression identification using deep reinforcement learning and differential evolution algorithm. IJACSA 2023;14(11):154-166. [doi: 10.14569/IJACSA.2023.0141115]
64. Shin D, Lee KJ, Adeluwa T, Hur J. Machine learning-based predictive modeling of postpartum depression. J Clin Med 2020 Sep 8;9(9):2899. [doi: 10.3390/jcm9092899] [Medline: 32911726]
65. Shivaprasad S, Chadaga K, Sampathila N, Prabhu S, Chadaga P R, K S S. Explainable machine learning methods to predict postpartum depression risk. Systems Science & Control Engineering 2024 Dec 31;12(1). [doi: 10.1080/21642583.2024.2427033]
66. Srivatsav P, Nanthini S. Detecting early markers of post-partum depression in new mothers: an efficient LSTM-CNN approach compared to logistic regression. 2024 Presented at: 2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT); Mar 15-16, 2024; Kottayam, India p. 1-6. [doi: 10.1109/ICITIIT61487.2024.10580321]
67. Suganthi D, Geetha A. Predicting postpartum depression with aid of social media texts using optimized machine learning model. IJIES 2024;17(3):417-427. [doi: 10.22266/ijies2024.0630.33]
68. Susič D, Bombač Tavčar L, Lučovnik M, Hrobat H, Gornik L, Gradišek A. Wellbeing forecasting in postpartum anemia patients. Healthcare (Basel) 2023 Jun 9;11(12):12. [doi: 10.3390/healthcare11121694] [Medline: 37372812]
69. Tang Y, Huang T, Yin X. Postpartum depression identification: integrating mutual learning-based artificial bee colony and proximal policy optimization for enhanced diagnostic precision. IJACSA 2024;15(6):332-347. [doi: 10.14569/IJACSA.2024.0150636]
70. Tortajada S, García-Gomez JM, Vicente J, et al. Prediction of postpartum depression using multilayer perceptrons and pruning. Methods Inf Med 2009;48(3):291-298. [doi: 10.3414/ME0562] [Medline: 19387507]
71. Valavani E, Doudesis D, Kourtesis I, et al. Data-driven insights towards risk assessment of postpartum depression. Presented at: Special Session on Mining Self-reported Outcome Measures, Clinical Assessments, and Non-invasive Sensor Data Towards Facilitating Diagnosis, Longitudinal Monitoring, and Treatment; Feb 24-26, 2020. [doi: 10.5220/0009369303820389]
72. Valdeolivar-Hernandez LI, Flores Quijano ME, Echeverria-Arjonilla JC, Perez-Gonzalez J, Piña-Ramírez O. Towards breastfeeding self-efficacy and postpartum depression estimation based on analysis of free-speech interviews through natural language processing. Presented at: 18th International Symposium on Medical Information Processing and Analysis (SIPAIM 2022); Nov 9-11, 2022. [doi: 10.1117/12.2669883]
73. Wagay FA. Comparing ensemble techniques for postpartum depression detection: a comprehensive analysis. Presented at: 2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI); Dec 29-30, 2023. [doi: 10.1109/ICAIIHI57871.2023.10489773]
74. Wakefield C, Frasch MG. Predicting patients requiring treatment for depression in the postpartum period using common electronic medical record data available antepartum. AJPM Focus 2023 Sep;2(3):100100. [doi: 10.1016/j.focus.2023.100100]
75. Wang J, Sui X, Hu B, Flint J, et al. Detecting postpartum depression in depressed people by speech features. In: Zu Q, Hu B, editors. Human Centered Computing: Springer, Cham; 2017, Vol. 10745. [doi: 10.1007/978-3-319-74521-3_46]
76. Wang S, Pathak J, Zhang Y. Using electronic health records and machine learning to predict postpartum depression. Stud Health Technol Inform 2019 Aug 21;264:888-892. [doi: 10.3233/SHTI190351] [Medline: 31438052]
77. Wang S, Xu R, Li G, Liu S, Zhu J, Gao P. A plasma proteomics-based model for identifying the risk of postpartum depression using machine learning. J Proteome Res 2025 Feb 7;24(2):824-833. [doi: 10.1021/acs.jproteome.4c00826] [Medline: 39772732]
78. Wang Y, Yan P, Wang G, et al. Trajectory on postpartum depression of Chinese women and the risk prediction models: a machine-learning based three-wave follow-up research. J Affect Disord 2024 Nov;365:185-192. [doi: 10.1016/j.jad.2024.08.074]

XSL•FO
RenderX

79. Xu J, Yu H, Lv H, et al. Consistent functional abnormalities in patients with postpartum depression. Behav Brain Res 2023 Jul 26;450:114467. [doi: 10.1016/j.bbr.2023.114467] [Medline: 37146719]
80. Xu W, Sampson M. Prenatal and childbirth risk factors of postpartum pain and depression: a machine learning approach. Matern Child Health J 2023 Feb;27(2):286-296. [doi: 10.1007/s10995-022-03532-0] [Medline: 36526882]
81. Yu Z, Matsukawa N, Saigusa D, et al. Plasma metabolic disturbances during pregnancy and postpartum in women with depression. iScience 2022 Dec 22;25(12):105666. [doi: 10.1016/j.isci.2022.105666] [Medline: 36505921]
82. Zhang W, Liu H, Silenzio VMB, Qiu P, Gong W. Machine learning models for the prediction of postpartum depression: application and comparison based on a cohort study. JMIR Med Inform 2020 Apr 30;8(4):e15516. [doi: 10.2196/15516] [Medline: 32352387]
83. Zhang Y, Joly R, Beecy AN, et al. Implementation of a machine learning risk prediction model for postpartum depression in the electronic health records. AMIA Jt Summits Transl Sci Proc 2024;2024:1057-1066. [Medline: 39444417]
84. Zhang Y, Wang S, Hermann A, Joly R, Pathak J. Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women. J Affect Disord 2021 Jan;279:1-8. [doi: 10.1016/j.jad.2020.09.113]
85. Zhu J, Ye Y, Liu X, et al. The incidence and risk factors of depression across six time points in the perinatal period: a prospective study in China. Front Med 2024;11:1407034. [doi: 10.3389/fmed.2024.1407034]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
CNN: convolutional neural network
DL: deep learning
EHR: electronic health record
EPDS: Edinburgh Postnatal Depression Scale
***ICD*:** *International Classification of Diseases*
LIWC: linguistic inquiry and word count
LogR: logistic regression
ML: machine learning
PCA: principal component analysis
PHQ-9: Patient Health Questionnaire-9
PPD: postpartum depression
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RF: random forest
RL: reinforcement learning
RNN: recurrent neural network

Edited by A Coristine, T Leung; submitted 12.May.2025; peer-reviewed by S Tedesco, YA Moaiad; revised version received 24.Jul.2025; accepted 31.Aug.2025; published 08.Jan.2026.

Please cite as:
Alkhateeb M, Nayeem A, Ahmed A, Alsahli M, Sheikh J, Abd-Alrazaq A
AI for Detecting and Predicting Postpartum Depression: Scoping Review
J Med Internet Res 2026;28:e77376
URL: https://www.jmir.org/2026/1/e77376
doi:10.2196/77376

© Mais Alkhateeb, Ajisha Nayeem, Arfan Ahmed, Mohammed Alsahli, Javaid Sheikh, Alaa Abd-Alrazaq. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Face-to-Face Versus Digital, Telephone-Delivered, and Self-Help Cognitive Behavioral Therapy for Irritable Bowel Syndrome: Systematic Review and Bayesian Indirect Treatment Comparison Meta-Analysis

Qing-Feng Tao[1*], MD; Can Hua[2*], MD; Xiao Zhuo[3], MD; Jian-Jiao Mou[1], MD; Chao-Rong Xie[1], MD; Yu-Xin Zhang[1], MD; Bei Lv[1], MD; Xin-Ying Niu[1], MD; Min Chen[3], MD, PhD; Hui Zheng[1], MD, PhD

[1]Acupuncture and Tuina School, Chengdu University of Traditional Chinese Medicine, No.1166 Liutai Avenue, Wenjiang District, Chengdu, China

[2]Department of Traditional Chinese Medicine, Dazhou Dachuan District People's Hospital (Dazhou Third People's Hospital), Dazhou, China

[3]Department of Colorectal Diseases, Hospital of Chengdu University of Traditional Chinese Medicine, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Hui Zheng, MD, PhD
Acupuncture and Tuina School, Chengdu University of Traditional Chinese Medicine, No.1166 Liutai Avenue, Wenjiang District, Chengdu, China

Abstract

Background: Cognitive behavioral therapy (CBT) is recommended for irritable bowel syndrome (IBS). However, it remains unclear whether face-to-face CBT is as effective as digital, self-help, or telephone-delivered CBT for IBS.

Objective: This study aimed to estimate the relative effects of face-to-face CBT compared with digital, telephone-delivered, and self-help CBT for IBS and to assess whether there are adequate effective sample sizes to support the findings.

Methods: Ovid MEDLINE, Embase, and the Cochrane Library were searched up to September 27, 2025. Randomized controlled trials of face-to-face, digital, self-help, or telephone-delivered CBT for IBS in adults were included. The primary outcome was the IBS symptom severity scale. The secondary outcomes were IBS quality of life and abdominal pain intensity. A Bayesian random effects model was used for the meta-analysis. The effective and required sample sizes were calculated to estimate whether the sample sizes were adequate. The certainty of evidence was evaluated using the Confidence in Network Meta-Analysis Framework. The risk of bias of included studies was assessed using the Cochrane Collaboration's risk of bias tool (version 2).

Results: We analyzed 22 studies involving 3161 participants. The number of participants ranged between 28 and 558. The mean (SD) age of participants was 37.2 (10.6) years, and 78.6% (2485/3161) were women. These randomized controlled trials were published between 2003 and 2025. We found that face-to-face CBT had similar effects compared with digital CBT (mean difference [MD] −0.89, 95% credible interval [CrI] −20.78 to 18.73), self-help CBT (MD −1.73, 95% CrI −21.03 to 17.80), and telephone-delivered CBT (MD −0.76, 95% CrI −20.86 to 19.38) in improving IBS symptom severity scale scores. The comparison between face-to-face CBT and self-help CBT had sufficient effective sample sizes (375/140), whereas the effective sample sizes for comparisons with digital CBT (347/729) and telephone-delivered CBT (140/627) were insufficient. The certainty of evidence was moderate to low. Similarly, in improving quality of life and abdominal pain intensity, face-to-face CBT showed equal effect compared with digital and self-help CBT, with insufficient sample sizes and low to very low evidence certainty.

Conclusions: This is the first Bayesian meta-analysis to incorporate effective and required sample size calculations for comparisons among CBT modalities in IBS. We analyzed continuous data of the outcomes. Meanwhile, we computed the effective and required sample sizes, thereby quantifying the informational adequacy of each comparison. Our Bayesian meta-analysis demonstrated significant potential for digital, self-help, and telephone-delivered CBT for patients with IBS, but the effective sample sizes of most comparisons were inadequate. Digital, self-help, and telephone-delivered CBT can serve as important options for managing IBS in clinical practice. Given high heterogeneity, high risk of bias, and inadequate effective sample sizes, more high-quality studies are warranted.

Trial Registration: OSF Registries OSF.IO/ZW2HQ; https://doi.org/10.17605/OSF.IO/ZW2HQ

(J Med Internet Res 2026;28:e75833) doi:10.2196/75833

KEYWORDS

cognitive behavioral therapy; digital CBT; irritable bowel syndrome; indirect treatment comparison; meta-analysis

XSL•FO
RenderX

Introduction

Rationale

Irritable bowel syndrome (IBS) is a prevalent disorder of brain-gut interaction marked by recurrent abdominal pain and altered bowel habits [1]. It affects 5% to 10% of the global population [2], significantly reducing quality of life, productivity, and mental health [3-6]. IBS causes a huge burden with years lived with disability at 627 per 100,000 [7]. The condition imposes substantial economic burdens, with direct costs approximately US $1 billion and indirect costs reaching US $50 million [8].

Research has shown the effectiveness of brain-gut behavioral therapies in improving IBS symptoms and quality of life [9-11]. Cognitive behavioral therapy (CBT), which integrates cognitive and behavioral techniques, is one such approach for alleviating symptoms [12]. Previous randomized controlled trials (RCTs) and meta-analyses have demonstrated that face-to-face CBT effectively relieves gastrointestinal symptoms and enhances quality of life [13-15]. However, despite the robust evidence supporting face-to-face CBT, its implementation is limited by the need for skilled mental health providers and the time and financial burden it imposes on patients [16,17]. In regions with limited CBT availability, accessing face-to-face sessions can be challenging. Consequently, there is a need for effective, accessible, and cost-effective alternative treatments.

The advent of the internet and digital technologies has provided a promising solution to the challenges associated with traditional face-to-face CBT for IBS. Digital CBT, which incorporates interactive programs based on cognitive behavioral models specific to IBS, has been increasingly integrated into medical practice [18,19]. Additionally, telephone-delivered CBT and self-help CBT have emerged as alternative delivery methods for IBS treatment. While RCTs have explored the clinical efficacy of these digital and remote CBT approaches, most have compared them to usual care or waiting lists [19,20], with only a few studies directly comparing them to face-to-face CBT [18,21,22]. As a result, the relative effectiveness of face-to-face CBT versus digital, telephone-delivered, and self-help CBT in managing IBS remains unclear.

Black et al [23] have conducted a network meta-analysis to evaluate the efficacy of psychological therapies for IBS. They found that face-to-face, digital, telephone-delivered, and self-help CBT were all efficacious for global IBS symptoms or abdominal pain, but none was superior to the others. Similarly, Goodoory et al [15] have explored the brain-gut behavioral treatments for abdominal pain in patients with IBS. Their results also found that face-to-face, digital, telephone-delivered, and self-help CBT were all effective for overall abdominal pain, with no one approach superior to another. However, both studies dichotomized outcomes—abdominal pain or global IBS symptoms were classified as *improved* or *not improved*—a strategy that can result in loss of information, reduced statistical power, and an increased risk of false-positive findings [24]. Therefore, the relative effectiveness of these distinct CBT modalities for IBS management requires further validation.

Moreover, studies using active controls require a large number of participants to achieve sufficient statistical power, counteracting inherent expectation biases and biases introduced by blinding [16]. Previous meta-analyses reported comparative effect sizes without evaluating whether the sample sizes were sufficient to yield stable estimates.

Indirect treatment comparison (ITC) allows us to evaluate the relative effectiveness of different interventions by fully using existing studies when there is no or insufficient direct evidence [25]. In addition, calculating the effective sample sizes and required sample sizes for each comparison can further validate the robustness of the findings [26].

Objectives

Consequently, we conducted a Bayesian ITC meta-analysis with two primary objectives: (1) to estimate the relative effect of face-to-face CBT versus digital, telephone-delivered, and self-help CBT for IBS; and (2) to assess whether there are adequate effective sample sizes to support the findings.

Methods

Ethical Considerations

Our study was a systematic review and ITC meta-analysis of published studies and was exempt from review and approval by the research ethics committee. Ethical approval and consent to participate were acquired by each included study.

Protocol and Registration

The systematic review was designed, conducted, and reported following the PRISMA-NMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Network Meta-Analyses) guidelines [27] and the PRISMA-S (Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension) guidelines (Checklist 1) [28]. The systematic review had been previously registered on the Open Science Framework [29].

Eligibility Criteria

The inclusion criteria were as follows. (1) participants—adults (aged ≥18 y) diagnosed with IBS according to the Rome or Manning criteria; (2) interventions—at least one CBT intervention, categorized into four distinct delivery modalities by referring to a previous study [30]: (i) face-to-face CBT (therapist-guided, in-person sessions conducted individually or in groups), (ii) digital CBT (therapist guided, provided via web-based platforms or mobile apps), (iii) telephone-delivered CBT (therapist guided, administered via telephone), and (iv) self-help CBT (therapist guided or unguided, structured written or web-based self-help materials that patients use to implement CBT techniques independently; the definitions of different CBT interventions are presented in Table S4 in Multimedia Appendix 1); (3) comparisons—placebo, waiting list, usual care, or other active treatments; (4) outcomes—at least one of the following outcomes: IBS symptom severity scale (IBS-SSS), IBS quality of life (IBS-QOL), and abdominal pain intensity (API); and (4) study design—a parallel-design RCT or a crossover-design RCT with available first-period data.

XSL•FO
RenderX

The exclusion criteria were as follows: (1) full-text articles unavailable and (2) duplicate articles.

Information Sources and Search

MEDLINE, Embase, and the Cochrane Library were systematically searched from inception to September 29, 2024. A supplementary search was conducted on September 27, 2025. No single platform was used to search multiple databases. Medline was accessed via Ovid, whereas Embase and the Cochrane Library were searched through their respective official websites. There was no language restriction. Search terms encompassed *cognitive behavioral therapy* and *irritable bowel syndrome*. The search strategies are detailed in Table S1 - S3 in Multimedia Appendix 1. Our search strategies were not informed by previous reviews and were not peer-reviewed. Simultaneously, clinicaltrials.gov was also searched for potentially eligible studies. In addition, the reference lists of previous meta-analyses were screened for any eligible studies [15,23]. We did not search other online resources, browse websites, contact authors or experts, reach out to manufacturers, or use any other methods to obtain additional literature.

Study Selection

Search results were imported into Zotero (version 7.0.8; Corporation for Digital Scholarship). Duplicate records were removed manually. Then, two reviewers (QF-T and CH) independently scanned the title and abstract. Subsequently, they assessed the full texts of potential RCTs for eligibility. Any disagreements were resolved through discussion, with a third reviewer (HZ) consulted if necessary.

Outcome Assessments

The primary outcome was the change in IBS-SSS score. The secondary outcomes were the changes in IBS-QOL score and API score (measured by a visual analog scale or a gastrointestinal symptom diary for abdominal pain). Outcomes were evaluated at the end of treatment.

Data Collection Process and Data Items

A reviewer (QF-T) extracted data using a standardized form, and the data were checked by the second reviewer (CH) independently. The extracted items included characteristics of the included RCTs, specifics of the intervention and control groups, and outcomes data. Data extraction for the meta-analysis used within-group change scores and their corresponding SDs for continuous outcomes. Disagreements were addressed through discussion and were consulted with a third reviewer (HZ). For missing data, we attempted to obtain them by reviewing previous studies.

Risk of Bias Assessment

Risk of bias was assessed by two reviewers (JJ-M and XZ) independently using the Cochrane Collaboration's risk of bias tool (version 2) for randomized trials [31]. By evaluating the following five components of the problem, that is, randomization process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported result, each included RCT was assessed as having low risk, some concerns, or high risk of bias.

Statistical Analysis

Geometry of the Network

A network plot was generated in which the nodes represent the interventions and the edges represent the direct comparisons from RCTs. The size of each node was proportional to the number of participants receiving the intervention, and the thickness of the edges reflected the number of studies contributing to each direct comparison.

Summary Measures

As all the outcomes were continuous, the pooled effect size of the Bayesian analysis was estimated as mean difference (MD) with 95% credible intervals (CrIs). The effect size is the pooled estimate from a network meta-analysis that combines both direct and indirect evidence for each intervention. In addition, the surface under the cumulative rank curve (SUCRA) values were also calculated to determine the relative ranking of each intervention, with higher SUCRA values signifying a more favorable ranking for the intervention.

Planned Methods of Analysis

This systematic review was performed under a Bayesian framework with vague priors, which facilitates the integration of existing evidence with new data to update estimates of treatment effects and provide superior handling of uncertainty in small sample studies compared to traditional approaches. We assessed the fit of the random effects (REs) and fixed effects (FEs) models by examining their posterior total residual deviance, the deviance information criterion (DIC), and the number of unconstrained data points. A model was considered to have a better fit if it exhibited a posterior total residual deviance closer to the number of unconstrained data points and a lower DIC.

When two or more arms received the same intervention, we pooled their participants, means, and SD. Meanwhile, we pooled mindfulness, stress management, and education as an alternative psychotherapy. On the basis of the actual delivery methods used in the study, we categorized these interventions separately as alternative face-to-face psychotherapy, alternative self-help psychotherapy, or alternative digital psychotherapy. τ^2 was used to estimate the heterogeneity among RCTs, with a τ^2 of more than 0.36 indicating significant heterogeneity [23].

Assessment of Inconsistency

An unrelated mean effects model was fit by drawing the dev-dev plots to assess the inconsistency globally, and node-splitting was fit to assess the inconsistency locally for each potentially inconsistent comparison in turn.

Additional Analyses

Two sensitivity analyses were conducted to calculate the robustness of the results: (1) excluding the RCT that was rated as high risk of bias and (2) using frequentist framework as the statistical method; To explore the source of heterogeneity, we performed subgroup analyses according to treatment duration (<8 wk or ≥8 wk), the delivery format of face-to-face CBT (individual vs group), and the guidance level of self-help CBT (guided vs unguided) using the primary outcome data.

XSL•FO
RenderX

All the analyses were conducted in R version 4.3.1. The Bayesian analysis was performed using the *multinma* package, and the frequentist analysis used the *netmeta* package.

The Certainty of Evidence

The certainty of evidence of each outcome was graded using a web application—the Confidence in Network Meta-Analysis Framework [32]. This approach evaluated six domains, such as within-study bias, reporting bias, indirectness, imprecision, heterogeneity, and incoherence. The certainty of evidence of each outcome was graded as high, moderate, low, or very low.

The Estimate of Effective Sample Sizes and Required Sample Sizes

The effective number of trials and the effective sample sizes for the ITC were calculated using the method developed by Thorlund and Mills [26]. The effective number of trials represents the number of trials required in an ITC to achieve a comparable level of power and precision to that of a single direct head-to-head trial. The effective sample sizes denote the number of participants in the comparison that would provide the same degree and strength of evidence as that provided in an RCT. Additionally, the required sample sizes were estimated [33]. In this systematic review, a noninferiority design with 1:1 randomization was used to estimate the required sample sizes. If the effective sample sizes reach the required sample sizes, it would demonstrate the robustness of the findings; otherwise, further research is needed to confirm the findings.

Results

Study Selection

The initial database and registry searches identified 955 articles (Figure 1). After screening the titles and abstracts of 568 articles, 473 were excluded. Following a full-text review of 88 articles, 22 eligible RCTs were included [13,14,18-22,34-48].

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the literature search and study selection. IBS: irritable bowel syndrome; RCT: randomized controlled trial.

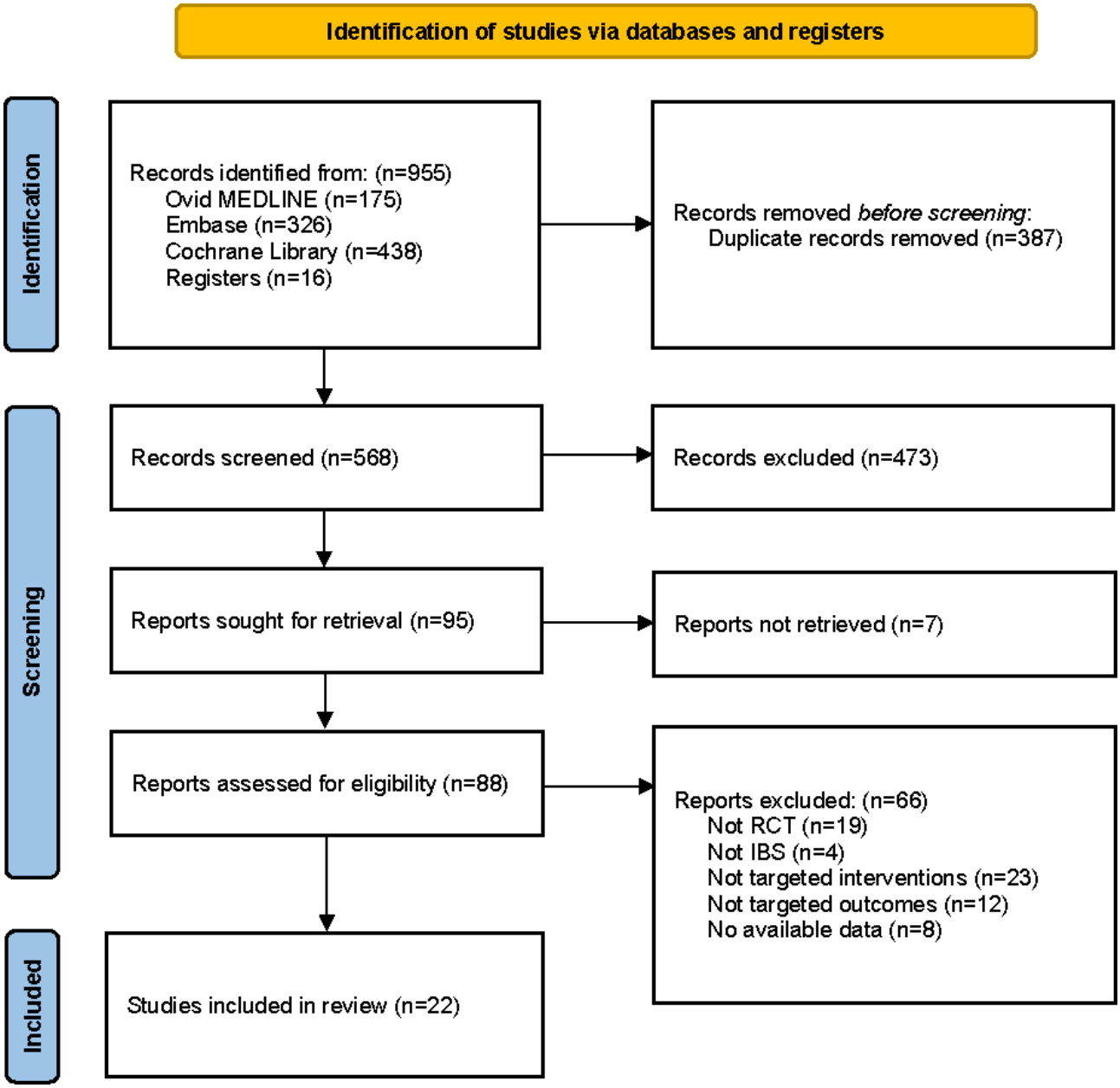

Study Characteristics

The characteristics of the eligible RCTs are presented in Table 1. A total of 3161 participants were enrolled in these RCTs, and the number of participants ranged between 28 and 558. The mean (SD) age of participants was 37.2 (10.6) years, and 78.6% were women. The RCTs included in our study were published between 2003 and 2025. Among these RCTs, 8 were from the United States (36.4%), 5 (22.7%) were from Sweden, 2 (9.1%) each from the United Kingdom, Korea, and Iran, and 1 (4.5%) each from the Netherlands, Canada, and Japan. In terms of intervention types, 7 (31.8%) RCTs reported face-to-face CBT, 6 (27.3%) RCTs reported digital CBT, 4 (18.2%) RCTs reported self-help CBT, 2 (9.1%) RCTs assessed both digital CBT and self-help CBT, 2 (9.1%) RCTs evaluated both self-help CBT and face-to-face CBT, and another (4.5%) examined both digital CBT and telephone-delivered CBT. All studies adopted the Rome criteria as the diagnostic standard. Two RCTs specifically focused on diarrhea-predominant IBS, while the remaining studies included all IBS subtypes.

XSL•FO
RenderX

Table . Characteristics of the included studies.

Study, year	Country	Study design	Disease	Diagnostic criteria	Interventions	Treatment duration (wk)	Population [sample size (n), female (%), mean ages (y)]	Results
Dehkordi and Solati, 2017 [34]	Iran	RCT[a]	IBS-D[b]	Rome III	Face-to-face CBT[c] versus usual care	8	64, 63, 33.7	IBS-QOL[d]: face-to-face CBT>usual care
Everitt et al, 2019 [35]	UK	RCT	IBS[e]	Rome III	Digital CBT versus telephone-delivered CBT versus usual care	48	558, 76, 43.1	IBS-SSS[f]: digital CBT=telephone-delivered CBT>usual care
Haghayegh et al, 2011 [36]	Iran	RCT	IBS-D	Rome II	Face-to-face CBT versus waiting list	8	32, 46, N/A[g]	IBS-QOL: face-to-face CBT>waiting list
Hunt et al, 2009 [38]	USA	RCT	IBS	Rome II	Digital CBT versus waiting list	6	54, 81, 38.5	IBS-QOL: digital CBT>waiting list
Hunt et al, 2015 [37]	USA	RCT	IBS	Rome III	Self-help CBT versus waiting list	6	60, 83, 36	IBS-QOL: self-help CBT>waiting list
Hunt et al, 2021 [20]	USA	RCT	IBS	Rome III	Self-help CBT versus waiting list	8	121, 75, 32	IBS-QOL: self-help CBT>waiting list
Hunt et al, 2025 [48]	USA	RCT	IBS	Rome IV	Self-help CBT versus alternative self-help psychotherapy	8	267, 72.3, 36.6	IBS-QOL: self-help CBT>alternative self-help psychotherapy
Jang et al, 2014 [39]	Korea	RCT	IBS	Rome III	Face-to-face CBT versus waiting list	8	90, 100, 21.6	IBS-QOL: face-to-face CBT>waiting list
Kennedy et al, 2005 [40]	UK	RCT	IBS	Rome I	Face-to-face CBT versus usual care	12	149, N/A, N/A	IBS-SSS: face-to-face CBT>usual care
Kikuchi et al, 2022 [13]	Japan	RCT	IBS	Rome III	Digital CBT versus face-to-face CBT versus usual care	10	114, 63, 39.7	IBS-SSS, IBS-QOL: digital CBT=face-to-face CBT>usual care

XSL•FO
RenderX

Study, year	Country	Study design	Disease	Diagnostic criteria	Interventions	Treatment duration (wk)	Population [sample size (n), female (%), mean ages (y)]	Results
Lackner et al, 2007 [14]	USA	RCT	IBS	Rome II	Face-to-face CBT versus alternative face-to-face psychotherapy versus waiting list	10	147, 82, 49.9	IBS-QOL: face-to-face CBT=alternative face-to-face psychotherapy=waiting list
Lackner et al, 2008 [21]	USA	RCT	IBS	Rome II	Face-to-face CBT versus self-help CBT versus waiting list	10	75, 86, 46.6	IBS-SSS, IBS-QOL: face-to-face CBT=self-help CBT>waiting list
Lackner et al, 2018 [22]	USA	RCT	IBS	Rome III	Face-to-face CBT versus self-help CBT versus alternative face-to-face psychotherapy	10	436, 80, 41.4	IBS-SSS: Face-to-face CBT=self-help CBT=alternative face-to-face psychotherapy
Lindfors 2020 [19]	Sweden	RCT	IBS	Rome IV	Digital CBT versus face-to-face CBT	3	141, 80, 37	IBS-SSS, IBS-QOL, API[h]: Digital CBT=face-to-face CBT
Ljótsson et al, 2010 [44]	Sweden	RCT	IBS	Rome III	Digital CBT versus waiting list	10	86, 85, 34.6	IBS-QOL, API: Digital CBT>waiting list
Ljótsson et al, 2011a [42]	Sweden	RCT	IBS	Rome III	Digital CBT versus alternative digital psychotherapy	10	195, 79, 38.9	IBS-QOL, API: Digital CBT=alternative digital psychotherapy
Ljótsson et al, 2011b [[41]	Sweden	RCT	IBS	Rome III	Digital CBT versus waiting list	10	61, 74, 34.9	IBS-QOL: Digital CBT>waiting list
Ljótsson et al, 2014 [43]	Sweden	RCT	IBS	Rome III	Digital CBT versus alternative digital psychotherapy	10	311, 80, 42.4	IBS-QOL: Digital CBT=alternative digital psychotherapy
Oerlemans et al, 2011 [19]	Netherlands	RCT	IBS	Rome III	Digital CBT versus usual care	4	76, 84, 38.3	IBS-QOL, API: Digital CBT>usual care

XSL•FO
RenderX

Study, year	Country	Study design	Disease	Diagnostic criteria	Interventions	Treatment duration (wk)	Population [sample size (n), female (%), mean ages (y)]	Results
Owusu et al, 2021 [45]	USA	RCT	IBS	Rome IV	Self-help CBT versus waiting list	12	36, 78, 39.2	IBS-SSS: self-help CBT>waiting list
Tkachuk et al, 2003 [46]	Canada	RCT	IBS	Rome II	Face-to-face CBT versus waiting list	9	28, 96, 39.5	API: face-to-face CBT=waiting list
Yang et al, 2022 [47]	Korea	RCT	IBS	Rome III	Face-to-face CBT versus waiting list	4	60, 88, 20.5	IBS-SSS, IBS-QOL: face-to-face CBT>waiting list

[a]RCT: randomized controlled trial.

[b]IBS-D: diarrhea-predominant irritable bowel syndrome.

[c]CBT: cognitive behavioral therapy.

[d]IBS-QOL: irritable bowel syndrome quality of life.

[e]IBS: irritable bowel syndrome.

[f]IBS-SSS: irritable bowel syndrome symptom severity scale.

[g]N/A: not available.

[h]API: abdominal pain intensity.

Risk of Bias

Risk of bias assessment is presented in Figure 2. Of the RCTs, 5 (22.7%) were adjudicated to be at high risk of bias, and 17 (77.3%) were adjudicated to have some concerns.

XSL•FO
RenderX

Figure 2. Risk of bias of included randomized controlled trials [13,14,18-22,34-48].

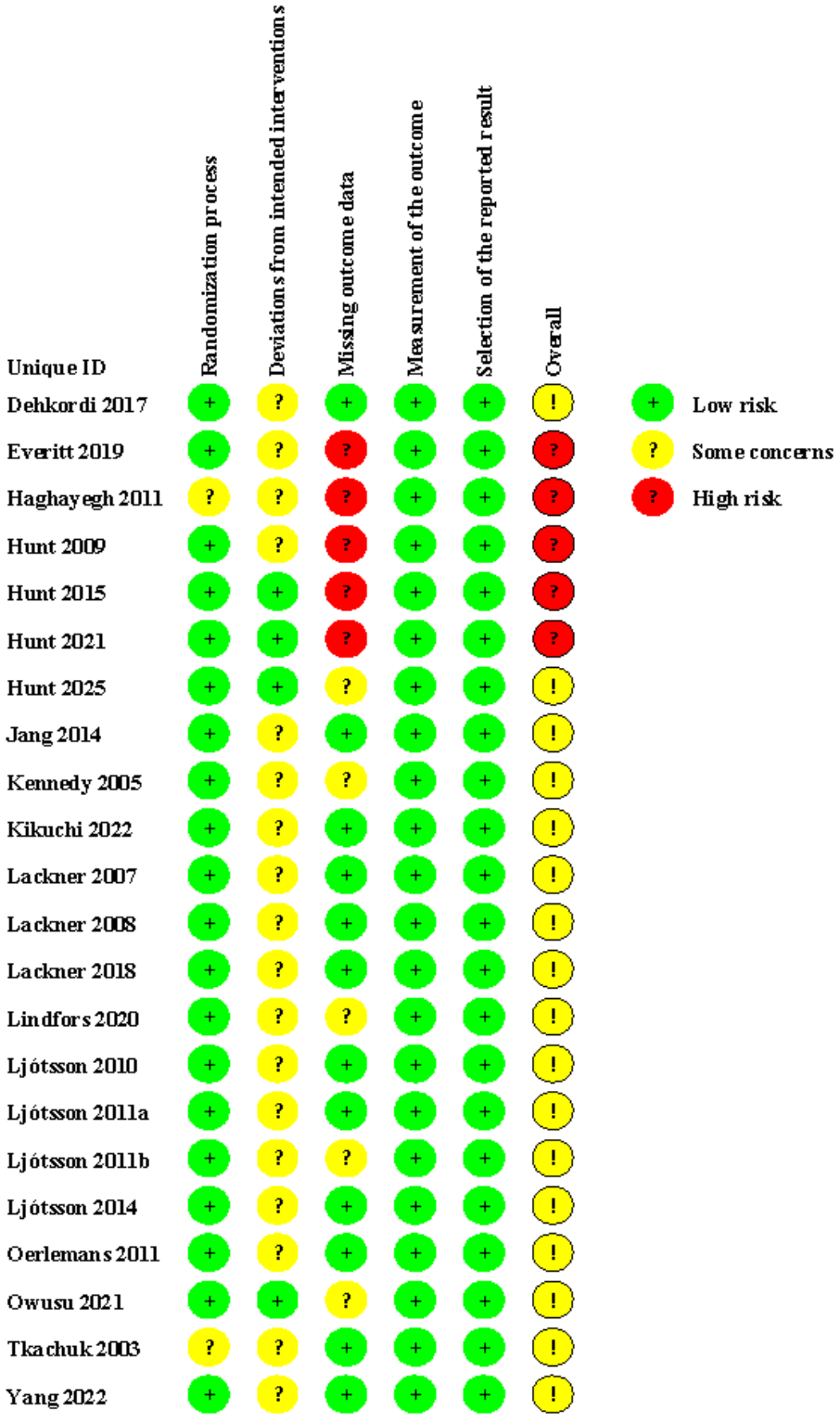

Exploration for Model Fit and Inconsistency

The results of the total posterior residuals deviance, the number of data points, and the DIC for both RE and FE models across all outcomes are presented in Table S5 in Multimedia Appendix 1. The RE model demonstrated a residual deviance closer to the number of data points and exhibited a lower DIC. Moreover, dev-dev plots indicated that all points were approximately aligned with the equality line, suggesting no evidence of global inconsistency (Figure S1 in Multimedia Appendix 1). Furthermore, the results of the node-splitting analysis indicated no local inconsistency (Figure S2 in Multimedia Appendix 1). Consequently, RE consistency models were selected for conducting the analyses.

Irritable Bowel Syndrome Symptom Severity Scale

A total of 8 RCTs, comprising 1562 participants, contributed data to the outcome analysis. The network plot is presented in

XSL•FO
RenderX

Figure S3 in Multimedia Appendix 1. Bayesian model results indicated that face-to-face CBT had similar effects compared to digital, telephone-delivered, and self-help CBT in reducing IBS-SSS (self-help CBT: MD −1.73, 95% CrI −21.03 to 17.80; digital CBT: MD −0.89, 95% CrI −20.78 to 18.73; telephone-delivered CBT: MD −0.76, 95% CrI −20.86 to 19.38; τ^2=24.61; Figure 3). The certainty of evidence was rated as moderate to low (Table S6 in Multimedia Appendix 1). The SUCRA rankings suggested that self-help CBT was the most effective (SUCRA 0.56, Table S7 in Multimedia Appendix 1). Furthermore, sensitivity analyses, which excluded high-risk RCTs (Figure S4 in Multimedia Appendix 1) and used a frequentist method (Figure S5 in Multimedia Appendix 1), yielded similar results, indicating the stability of the findings.

Figure 3. Effect of comparison between face-to-face CBT and other types of CBT for irritable bowel syndrome symptom severity scale [13,18,21,22,35,40,45,47]. The black vertical line corresponds to 0. CBT: cognitive behavioral therapy; CrI: credible interval; MD: mean difference.

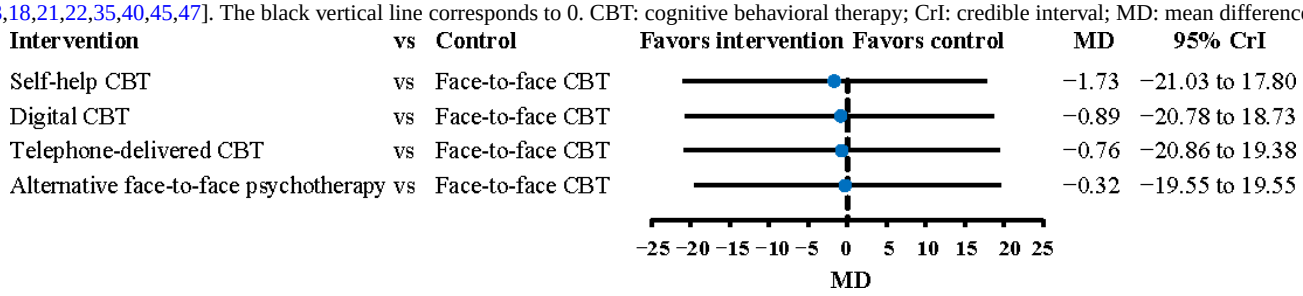

The results of the effective sample sizes and the required sample sizes are presented in Table 2 and Multimedia Appendix 2. Adequate effective sample sizes were observed for the comparison between face-to-face and self-help CBT (375/140, 267.9%; Table 2 and Multimedia Appendix 2). However, for the other comparisons, the effective sample sizes were insufficient, suggesting that more studies are needed in the future (digital CBT: 347/729, 47.6%; telephone-delivered CBT: 140/627, 22.3%; Table 2 and Multimedia Appendix 2).

Table . The effective sample sizes and required sample sizes estimation of outcomes.

Comparison (vs face-to-face CBT[a])	Effective number of trials (n/N)	Effective head-to-head sample sizes, n	Effective indirect sample sizes, n	Total effective sample sizes, n	Required sample sizes, n	Information fraction (n/N,%)	Network meta-analysis, MD[b] estimate (95% credible intervals)
IBS-SSS[c]							
Digital CBT	4/8	195	152	347	729	347/729, 47.6	−0.89 (−20.78 to 18.73)
Self-help CBT	4/8	339	36	375	140	375/140, 267.9	−1.73 (−21.03 to 17.80)
Telephone-delivered CBT	3/9	N/A[d]	140	140	627	140/627, 22.3	−0.76 (−20.86 to 19.38)
IBS-QOL[e]							
Digital CBT	8/9	195	107	302	3550	302/3550, 8.5	−4.29 (−19.10 to 10.59)
Self-help CBT	8/9	48	123	171	4610	171/4610, 3.7	−2.82 (−18.83 to 13.76)
API[f]							
Digital CBT	2/12	141	21	162	3764	162/3764, 4.3	−1.14 (−19.69 to 16.49)

[a]CBT: cognitive behavioral therapy.

[b]MD: mean difference.

[c]IBS-SSS: irritable bowel syndrome symptom severity scale.

[d]N/A: not applicable.

[e]IBS-QOL: irritable bowel syndrome quality of life.

[f]API: abdominal pain intensity.

Irritable Bowel Syndrome Quality of Life

A total of 17 RCTs, involving 1798 participants, contributed data to the outcome analysis. The network plot is presented in Figure S6 in Multimedia Appendix 1. Face-to-face CBT was found to be equally effective as digital CBT and self-help CBT in enhancing the quality of life for patients with IBS (digital CBT: MD −4.29, 95% CrI −19.10 to 10.59; self-help CBT: MD

XSL•FO
RenderX

−2.82, 95% CrI −18.83 to 13.76; τ^2=15.93; Figure 4). The Confidence in Network Meta-Analysis Framework evidence rating was very low (Table S6 in Multimedia Appendix 1). Digital CBT ranked higher than face-to-face CBT, with a SUCRA value of 0.71 (Table S7 in Multimedia Appendix 1). Sensitivity analyses confirmed the stability of these results after excluding high-risk RCTs (Figure S7 in Multimedia Appendix 1) and using a frequentist method (Figure S8 in Multimedia Appendix 1).

Figure 4. Effect of comparison between face-to-face CBT and other types of CBT for irritable bowel syndrome quality of life [13,14,18-21,34,36-39,41-44,47,48]. The black vertical line corresponds to 0. CBT: cognitive behavioral therapy; CrI: credible interval; MD: mean difference.

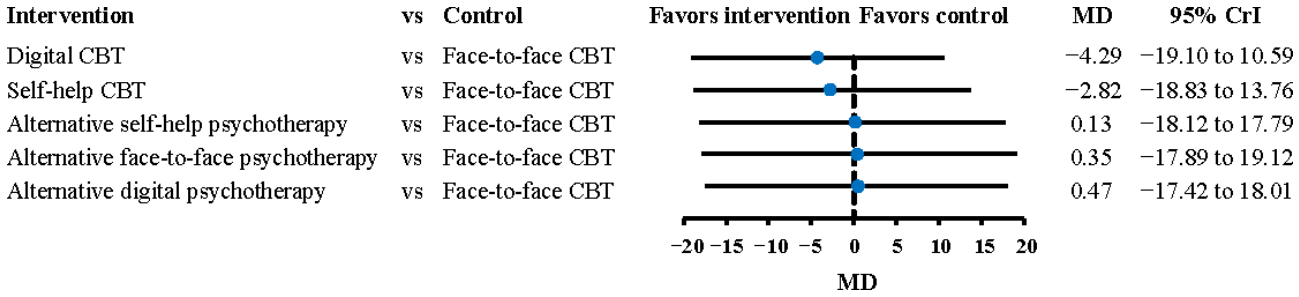

The effective sample sizes did not meet the required sample sizes, suggesting that more studies are warranted (digital CBT: 302/3550, 8.5%; self-help CBT: 171/4610, 3.7%; Table 2 and Multimedia Appendix 2).

Abdominal Pain Intensity

Four RCTs, involving 327 participants, contributed data to the outcome analysis. The network plot is presented in Figure S9 in Multimedia Appendix 1. Digital CBT showed similar effects compared with face-to-face CBT in reducing the API, with very low certainty of evidence (MD −1.14, 95% CrI −19.69 to 16.49, τ^2=23.66; Figure S10 in Multimedia Appendix 1, Table S6 in Multimedia Appendix 1). The SUCRA rankings indicated that digital CBT was ranked higher than face-to-face CBT (SUCRA 0.54 vs 0.49, Table S7 in Multimedia Appendix 1). The result was consistent with the sensitivity analysis using a frequentist method (Figure S11 in Multimedia Appendix 1).

As presented in Table 2 and Multimedia Appendix 2, the effective sample sizes did not meet the required sample sizes in the comparison between face-to-face CBT and digital CBT (141/3764, 4.3%), suggesting that more studies are needed.

Results of Subgroup Analysis

The results of the subgroup analysis of the primary outcome are presented in Figure S13 - S15 in Multimedia Appendix 1. We found that the duration of treatment (Figure S12 in Multimedia Appendix 1), the delivery method of face-to-face CBT (Figure S13 in Multimedia Appendix 1), or the guidance level of self-help CBT (Figure S14 in Multimedia Appendix 1) had no significant influence on the heterogeneity. Additionally, subgroup analysis suggested that the treatment duration (<8 wk or ≥8 wk; Figure S12 in Multimedia Appendix 1), the delivery method of face-to-face CBT (individual or group face-to-face CBT, Figure S13 in Multimedia Appendix 1), or the guidance level of self-help CBT (guided or unguided self-help CBT, Figure S14 in Multimedia Appendix 1) did not significantly modify the effect on global IBS symptoms.

Discussion

Summary of Evidence

In this systematic review, we assessed the relative effect of different delivery methods of CBT for patients with IBS and assessed the effective sample sizes and required sample sizes of each finding to evaluate the robustness. We included 22 RCTs involving 3161 adults with IBS. We found that face-to-face CBT presented a similar effect compared with self-help CBT in improving global IBS symptoms, with sufficient effective sample sizes. Face-to-face CBT showed an equal effect compared with digital CBT and telephone-delivered CBT in improving global IBS symptoms; however, the effective sample sizes were less than the required sample sizes. For quality of life and API, there were similar effects between face-to-face CBT and digital and self-help CBT, but the effective sample sizes were insufficient.

Face-to-face CBT is the primary mode of CBT delivery and is recognized as an effective intervention for various conditions, including major depressive disorder, insomnia, and IBS [1,49,50]. Our results showed that compared with face-to-face CBT, there were slight differences between digital, telephone-delivered, and self-help CBT in IBS-SSS, yet none reached the minimal clinically important difference value of 50 [51]. The similar effects might be explained by the fact that digital, telephone-delivered, and self-help CBT represent modifications of traditional face-to-face CBT in terms of delivery method. However, the core principles of CBT remain consistent across these modalities, ensuring efficacy while enhancing convenience and accessibility. Our findings are consistent with those of previous meta-analyses by Black et al [23] and Goodoory et al [15].

Compared to previous meta-analyses [15,23], their studies primarily used binary outcome measures and exclusively used frequentist analysis methods. In contrast, our study analyzed continuous outcomes for the IBS-SSS, IBS-QOL, and API. We used a Bayesian model for the primary analysis and a frequentist model for sensitivity analysis to confirm the stability of our findings. Additionally, our study is the first to incorporate effective and required sample size calculations for ITC within

XSL•FO
RenderX

CBT for IBS. We conducted a novel assessment of effective and required sample sizes, which has not been previously undertaken in CBT studies for IBS. Our results indicate that there are sufficient effective sample sizes to support that there is a similar effect between face-to-face CBT and self-help CBT in improving global IBS symptoms. However, other comparisons exhibited insufficient effective sample sizes.

Implication for Practice and Research

For clinical practice, the social and economic impact of IBS necessitates the demonstration of both clinical effectiveness and innovative direct-to-patient delivery systems to enhance patient access to appropriate therapeutic interventions [22]. Our findings indicate that digital, self-help, and telephone-delivered CBT have similar effects compared with face-to-face CBT. Meanwhile, a previous study has shown that a 10-week digital CBT treatment can reduce direct medical costs by US $358 and indirect costs by US $5014 [52]. Therefore, these alternative delivery methods deserve consideration in the clinical management of IBS. Notably, digital CBT has considerable potential for managing IBS because of its effectiveness [18], accessibility [38], and cost-effectiveness [52,53]. Additionally, our research found that both individualized face-to-face CBT and group face-to-face CBT demonstrated a similar effect in relieving the overall symptoms of patients with IBS compared to other treatments. Therefore, clinicians may choose between individual CBT and group CBT based on resource availability and patient preference in clinical practice.

Furthermore, for research, although our analysis suggests comparable effectiveness between digital CBT and face-to-face CBT, the effective sample size (n=347) represents only 47.6% of the required sample size (n=729) for this comparison. This indicates that while the current evidence is promising, implementation of digital CBT as a substitute for traditional face-to-face delivery in clinical practice would necessitate monitoring of larger patient cohorts to definitively confirm therapeutic equivalence and ensure consistent clinical outcomes.

Limitations

Several limitations should be considered when interpreting our findings. At the study and outcome level, first, none of the included studies were rated as low risk of bias, primarily due to the difficulty of blinding in RCTs of CBT. Moreover, five studies (22.7%) were rated as high risk of bias; hence, we excluded RCTs with high risk of bias for sensitivity analysis and found that the results were consistent with the main analysis. Second, there was significant heterogeneity among the included studies; therefore, we conducted subgroup analyses on treatment duration, the delivery method of face-to-face CBT, and the guidance level of self-help CBT. These analyses revealed that these factors did not significantly influence the heterogeneity. Initially, we intended to explore whether the source of heterogeneity was related to disease duration, disease subtype, and gender through subgroup analyses. However, this analysis could not be conducted due to the limited number of included studies, which is also attributed to the scarcity of high-quality RCTs on CBT for IBS, further emphasizing the need for more high-quality RCTs.

At the review level, first, we searched only Ovid MEDLINE, Embase, Cochrane Library, and ClinicalTrials.gov. We did not pursue additional literature via other online resources, correspondence with authors or experts, contact with manufacturers, or any other means; thus, some literature may have been missed. Nevertheless, we screened the reference lists of previous meta-analyses for any missed studies. Second, our search strategies did not consult an information scientist, which is another limitation of our study. Third, when estimating effective sample sizes, we used an estimation method that was not adjusted for heterogeneity, which may have overestimated the effective sample sizes [26]. Nonetheless, our results indicated that even if the effective sample sizes were overestimated, they still fell short of the required sample sizes for most comparisons. This finding further underscores the necessity for additional RCTs to validate our results in the future.

Conclusions

To our knowledge, this is the first Bayesian meta-analysis to incorporate effective and required sample size calculations for indirect treatment comparisons among CBT modalities in IBS. Compared with previous meta-analyses, we analyzed continuous outcomes of global symptoms, IBS-QOL, and API. For each comparison, we computed the effective sample size and the required sample size, thereby quantifying the informational adequacy. The meta-analysis suggested that digital, self-help, and telephone-delivered CBT had similar effects on global IBS symptoms, quality of life, and API as face-to-face CBT in patients with IBS. Therefore, digital CBT, self-help CBT, and telephone-delivered CBT can serve as important methods for managing IBS in clinical practice. However, our findings suggested that the effective sample sizes of most comparisons were insufficient. Given high heterogeneity, high risk of bias, and inadequate effective sample sizes, more high-quality studies are warranted in the future.

Acknowledgments

No generative artificial intelligence tools were used in any stage of the research, data analysis, or writing of this manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (2474658) and Chengdu Municipal Health Commission (WXLH202402020). These sponsors did not participate in the design of the study, the analyzed and interpretation the data, and the decision process to submit the manuscript for publication.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: HZ, MC
Data curation: QF-T, CH
Formal analysis: QF-T, CH
Funding acquisition: HZ, MC
Investigation: QF-T, CH
Methodology: QF-T, CH, XZ, JJ-M, CR-X, YX-Z
Project administration: MC
Software: QF-T, CH
Supervision: HZ
Validation: CH
Visualization: QF-T, CH, XZ, JJ-M
Writing – original draft: QF-T, CH
Writing – review & editing: XZ, JJ-M, CR-X, YX-Z, BL, XY-N, MC, HZ
All authors had full access to all the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis. All authors reviewed and approved the final manuscript.
QF-T and CH: co-first authors
HZ and MC are co-corresponding authors. MC can be contacted at Department of Colorectal Diseases, Hospital of Chengdu University of Traditional Chinese Medicine, No.39 Shi-er-qiao Road, Jinniu District, Chengdu 610000, China, cm@cdutcm.edu.cn

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies, definition of different cognitive behavioral therapy, model fit, certainty of evidence, consistency assessment, value of surface under the cumulative rank curve, network diagram, sensitivity analysis, and subgroup analysis.
[DOCX File, 3683 KB - jmir_v28i1e75833_app1.docx]

Multimedia Appendix 2
Assessment of effective number of trials, effective sample sizes, and required sample sizes.
[DOCX File, 2857 KB - jmir_v28i1e75833_app2.docx]

Checklist 1
PRISMA checklist.
[PDF File, 147 KB - jmir_v28i1e75833_app3.pdf]

References

1. Mearin F, Lacy BE, Chang L, et al. Bowel disorders. Gastroenterology 2016 Feb 18:S0016-5085(16)00222-5. [doi: 10.1053/j.gastro.2016.02.031] [Medline: 27144627]
2. Oka P, Parr H, Barberio B, Black CJ, Savarino EV, Ford AC. Global prevalence of irritable bowel syndrome according to Rome III or IV criteria: a systematic review and meta-analysis. Lancet Gastroenterol Hepatol 2020 Oct;5(10):908-917. [doi: 10.1016/S2468-1253(20)30217-X] [Medline: 32702295]
3. Peery AF, Crockett SD, Murphy CC, et al. Burden and cost of gastrointestinal, liver, and pancreatic diseases in the United States: update 2021. Gastroenterology 2022 Feb;162(2):621-644. [doi: 10.1053/j.gastro.2021.10.017] [Medline: 34678215]
4. Black CJ, Ford AC. Global burden of irritable bowel syndrome: trends, predictions and risk factors. Nat Rev Gastroenterol Hepatol 2020 Aug;17(8):473-486. [doi: 10.1038/s41575-020-0286-8] [Medline: 32296140]
5. Frändemark Å, Törnblom H, Jakobsson S, Simrén M. Work productivity and activity impairment in irritable bowel syndrome (IBS): a multifaceted problem. Am J Gastroenterol 2018 Oct;113(10):1540-1549. [doi: 10.1038/s41395-018-0262-x] [Medline: 30254230]
6. Kutschke J, Harris JR, Bengtson MB. The relationships between IBS and perceptions of physical and mental health-a Norwegian twin study. BMC Gastroenterol 2022 May 28;22(1):266. [doi: 10.1186/s12876-022-02340-8] [Medline: 35643443]
7. Porter CK, Thura N, Riddle MS. Quantifying the incidence and burden of postinfectious enteric sequelae. Mil Med 2013 Apr;178(4):452-469. [doi: 10.7205/MILMED-D-12-00510] [Medline: 23707833]

XSL•FO
RenderX

8. Everhart JE, Ruhl CE. Burden of digestive diseases in the United States part II: lower gastrointestinal diseases. Gastroenterology 2009 Mar;136(3):741-754. [doi: 10.1053/j.gastro.2009.01.015] [Medline: 19166855]
9. Henrich JF, Gjelsvik B, Surawy C, Evans E, Martin M. A randomized clinical trial of mindfulness-based cognitive therapy for women with irritable bowel syndrome-Effects and mechanisms. J Consult Clin Psychol 2020 Apr;88(4):295-310. [doi: 10.1037/ccp0000483] [Medline: 32134291]
10. Berry SK, Berry R, Recker D, Botbyl J, Pun L, Chey WD. A randomized parallel-group study of digital gut-directed hypnotherapy vs muscle relaxation for irritable bowel syndrome. Clin Gastroenterol Hepatol 2023 Nov;21(12):3152-3159. [doi: 10.1016/j.cgh.2023.06.015] [Medline: 37391055]
11. Gaylord SA, Palsson OS, Garland EL, et al. Mindfulness training reduces the severity of irritable bowel syndrome in women: results of a randomized controlled trial. Am J Gastroenterol 2011 Sep;106(9):1678-1688. [doi: 10.1038/ajg.2011.184] [Medline: 21691341]
12. Hayes SC, Hofmann SG. The third wave of cognitive behavioral therapy and the rise of process-based care. World Psychiatry 2017 Oct;16(3):245-246. [doi: 10.1002/wps.20442] [Medline: 28941087]
13. Kikuchi S, Oe Y, Ito Y, et al. Group cognitive-behavioral therapy with interoceptive exposure for drug-refractory irritable bowel syndrome: a randomized controlled trial. Am J Gastroenterol 2022 Apr 1;117(4):668-677. [doi: 10.14309/ajg.0000000000001664] [Medline: 35103022]
14. Lackner JM, Jaccard J, Krasner SS, Katz LA, Gudleski GD, Blanchard EB. How does cognitive behavior therapy for irritable bowel syndrome work? A mediational analysis of a randomized clinical trial. Gastroenterology 2007 Aug;133(2):433-444. [doi: 10.1053/j.gastro.2007.05.014] [Medline: 17681164]
15. Goodoory VC, Khasawneh M, Thakur ER, et al. Effect of brain-gut behavioral treatments on abdominal pain in irritable bowel syndrome: systematic review and network meta-analysis. Gastroenterology 2024 Oct;167(5):934-943. [doi: 10.1053/j.gastro.2024.05.010] [Medline: 38777133]
16. Staudacher HM, Black CJ, Teasdale SB, Mikocka-Walus A, Keefer L. Irritable bowel syndrome and mental health comorbidity-approach to multidisciplinary management. Nat Rev Gastroenterol Hepatol 2023 Sep;20(9):582-596. [doi: 10.1038/s41575-023-00794-z] [Medline: 37268741]
17. Camilleri M, Dilmaghani S. Update on treatment of abdominal pain in irritable bowel syndrome: a narrative review. Pharmacol Ther 2023 May;245:108400. [doi: 10.1016/j.pharmthera.2023.108400] [Medline: 37001737]
18. Lindfors P, Axelsson E, Engstrand K, et al. Online education is non-inferior to group education for irritable bowel syndrome: a randomized trial and patient preference trial. Clin Gastroenterol Hepatol 2021 Apr;19(4):743-751. [doi: 10.1016/j.cgh.2020.04.005] [Medline: 32289541]
19. Oerlemans S, van Cranenburgh O, Herremans PJ, Spreeuwenberg P, van Dulmen S. Intervening on cognitions and behavior in irritable bowel syndrome: a feasibility trial using PDAs. J Psychosom Res 2011 Mar;70(3):267-277. [doi: 10.1016/j.jpsychores.2010.09.018] [Medline: 21334498]
20. Hunt M, Miguez S, Dukas B, Onwude O, White S. Efficacy of zemedy, a mobile digital therapeutic for the self-management of irritable bowel syndrome: crossover randomized controlled trial. JMIR Mhealth Uhealth 2021 May 20;9(5):e26152. [doi: 10.2196/26152] [Medline: 33872182]
21. Lackner JM, Jaccard J, Krasner SS, Katz LA, Gudleski GD, Holroyd K. Self-administered cognitive behavior therapy for moderate to severe irritable bowel syndrome: clinical efficacy, tolerability, feasibility. Clin Gastroenterol Hepatol 2008 Aug;6(8):899-906. [doi: 10.1016/j.cgh.2008.03.004] [Medline: 18524691]
22. Lackner JM, Jaccard J, Keefer L, et al. Improvement in gastrointestinal symptoms after cognitive behavior therapy for refractory irritable bowel syndrome. Gastroenterology 2018 Jul;155(1):47-57. [doi: 10.1053/j.gastro.2018.03.063] [Medline: 29702118]
23. Black CJ, Thakur ER, Houghton LA, Quigley EMM, Moayyedi P, Ford AC. Efficacy of psychological therapies for irritable bowel syndrome: systematic review and network meta-analysis. Gut 2020 Aug;69(8):1441-1451. [doi: 10.1136/gutjnl-2020-321191] [Medline: 32276950]
24. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ 2006 May 6;332(7549):1080. [doi: 10.1136/bmj.332.7549.1080] [Medline: 16675816]
25. Song F, Altman DG, Glenny AM, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. BMJ 2003 Mar 1;326(7387):472. [doi: 10.1136/bmj.326.7387.472] [Medline: 12609941]
26. Thorlund K, Mills EJ. Sample size and power considerations in network meta-analysis. Syst Rev 2012 Sep 19;1:41. [doi: 10.1186/2046-4053-1-41] [Medline: 22992327]
27. Hutton B, Salanti G, Caldwell DM, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Ann Intern Med 2015 Jun 2;162(11):777-784. [doi: 10.7326/M14-2385] [Medline: 26030634]
28. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
29. OSF. Digital versus face-to-face, telephone-delivered, and self-helped cognitive behavioral therapy for irritable bowel syndrome: an indirect treatment comparison meta-analysis. URL: https://osf.io/zw2hq/overview

XSL•FO
RenderX

30. Gao Y, Ge L, Liu M, et al. Comparative efficacy and acceptability of cognitive behavioral therapy delivery formats for insomnia in adults: a systematic review and network meta-analysis. Sleep Med Rev 2022 Aug;64:101648. [doi: 10.1016/j.smrv.2022.101648] [Medline: 35759820]
31. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
32. Nikolakopoulou A, Higgins JPT, Papakonstantinou T, et al. CINeMA: an approach for assessing confidence in the results of a network meta-analysis. PLoS Med 2020 Apr;17(4):e1003082. [doi: 10.1371/journal.pmed.1003082] [Medline: 32243458]
33. Zhong B. How to calculate sample size in randomized controlled trial? J Thorac Dis 2009 Dec;1(1):51-54. [Medline: 22263004]
34. Dehkordi AH, Solati K. The effects of cognitive behavioral therapy and drug therapy on quality of life and symptoms of patients with irritable bowel syndrome. J Adv Pharm Technol Res 2017;8(2):67-72. [doi: 10.4103/japtr.JAPTR-170-16]
35. Everitt HA, Landau S, O'Reilly G, et al. Assessing telephone-delivered cognitive-behavioural therapy (CBT) and web-delivered CBT versus treatment as usual in irritable bowel syndrome (ACTIB): a multicentre randomised trial. Gut 2019 Sep;68(9):1613-1623. [doi: 10.1136/gutjnl-2018-317805] [Medline: 30971419]
36. Haghayegh SA, Kalantari M, Molavi H, Talebi M, Alemohamad J. The efficacy of cognitive-behavior group therapy on health-related quality of life, health anxiety and depression in patients with diarrhea-predominant irritable bowel syndrome. Pak J Med Sci 2011;27(4):749-753. [doi: 10.1002/central/CN-00850913/full]
37. Hunt MG, Ertel E, Coello JA, Rodriguez L. Empirical support for a self-help treatment for IBS. Cogn Ther Res 2015 Apr;39(2):215-227. [doi: 10.1007/s10608-014-9647-3]
38. Hunt MG, Moshier S, Milonova M. Brief cognitive-behavioral internet therapy for irritable bowel syndrome. Behav Res Ther 2009 Sep;47(9):797-802. [doi: 10.1016/j.brat.2009.05.002] [Medline: 19570525]
39. Jang AL, Hwang SK, Kim DU. The effects of cognitive behavioral therapy in female nursing students with irritable bowel syndrome: a randomized trial. Eur J Gastroenterol Hepatol 2014 Aug;26(8):918-926. [doi: 10.1097/MEG.0000000000000140] [Medline: 24999797]
40. Kennedy T, Jones R, Darnley S, Seed P, Wessely S, Chalder T. Cognitive behaviour therapy in addition to antispasmodic treatment for irritable bowel syndrome in primary care: randomised controlled trial. BMJ 2005 Aug 20;331(7514):435. [doi: 10.1136/bmj.38545.505764.06] [Medline: 16093252]
41. Ljótsson B, Andersson G, Andersson E, et al. Acceptability, effectiveness, and cost-effectiveness of internet-based exposure treatment for irritable bowel syndrome in a clinical sample: a randomized controlled trial. BMC Gastroenterol 2011 Oct 12;11:110. [doi: 10.1186/1471-230X-11-110] [Medline: 21992655]
42. Ljótsson B, Hedman E, Andersson E, et al. Internet-delivered exposure-based treatment vs. stress management for irritable bowel syndrome: a randomized trial. Am J Gastroenterol 2011 Aug;106(8):1481-1491. [doi: 10.1038/ajg.2011.139] [Medline: 21537360]
43. Ljótsson B, Hesser H, Andersson E, et al. Provoking symptoms to relieve symptoms: a randomized controlled dismantling study of exposure therapy in irritable bowel syndrome. Behav Res Ther 2014 Apr;55:27-39. [doi: 10.1016/j.brat.2014.01.007] [Medline: 24584055]
44. Ljótsson B, Falk L, Vesterlund AW, et al. Internet-delivered exposure and mindfulness based therapy for irritable bowel syndrome--a randomized controlled trial. Behav Res Ther 2010 Jun;48(6):531-539. [doi: 10.1016/j.brat.2010.03.003] [Medline: 20362976]
45. Owusu JT, Sibelli A, Moss-Morris R, van Tilburg MAL, Levy RL, Oser M. A pilot feasibility study of an unguided, internet-delivered cognitive behavioral therapy program for irritable bowel syndrome. Neurogastroenterol Motil 2021 Nov;33(11):e14108. [doi: 10.1111/nmo.14108] [Medline: 33745228]
46. Tkachuk GA, Graff LA, Martin GL, Bernstein CN. Randomized controlled trial of cognitive–behavioral group therapy for irritable bowel syndrome in a medical setting. J Clin Psychol Med Settings 2003 Mar;10(1):57-69. [doi: 10.1023/A:1022809914863]
47. Yang YY, Jun S. The effects of cognitive behavioral therapy for insomnia among college students with irritable bowel syndrome: a randomized controlled trial. Int J Environ Res Public Health 2022 Oct 29;19(21):14174. [doi: 10.3390/ijerph192114174] [Medline: 36361052]
48. Hunt M, Dalvie A, Ipek S, Glinski S, Macks R. Efficacy of a CBT self-help app (Zemedy) versus an education, relaxation, and mindfulness app for IBS: results from post-treatment, 3-month, and 6-month follow-up. J Clin Gastroenterol 2025 Mar 12. [doi: 10.1097/MCG.0000000000002164] [Medline: 40071822]
49. Otte C, Gold SM, Penninx BW, et al. Major depressive disorder. Nat Rev Dis Primers 2016 Sep 15;2:16065. [doi: 10.1038/nrdp.2016.65] [Medline: 27629598]
50. Perlis ML, Posner D, Riemann D, Bastien CH, Teel J, Thase M. Insomnia. Lancet 2022 Sep 24;400(10357):1047-1060. [doi: 10.1016/S0140-6736(22)00879-0] [Medline: 36115372]
51. Francis CY, Morris J, Whorwell PJ. The irritable bowel severity scoring system: a simple method of monitoring irritable bowel syndrome and its progress. Aliment Pharmacol Ther 1997 Apr;11(2):395-402. [doi: 10.1046/j.1365-2036.1997.142318000.x] [Medline: 9146781]

XSL•FO
RenderX

52. Andersson E, Ljótsson B, Smit F, et al. Cost-effectiveness of internet-based cognitive behavior therapy for irritable bowel syndrome: results from a randomized controlled trial. BMC Public Health 2011 Apr 7;11(1):215. [doi: 10.1186/1471-2458-11-215] [Medline: 21473754]
53. Wallén H, Lindfors P, Andersson E, et al. Return on investment of internet delivered exposure therapy for irritable bowel syndrome: a randomized controlled trial. BMC Gastroenterol 2021 Jul 13;21(1):289. [doi: 10.1186/s12876-021-01867-6] [Medline: 34256715]

Abbreviations

API: abdominal pain intensity
CBT: cognitive behavioral therapy
CrI: credible interval
DIC: Deviance information criteria
FE: fixed effect
IBS: irritable bowel syndrome
IBS-QOL: irritable bowel syndrome quality of life
IBS-SSS: irritable bowel syndrome symptom severity scale
ITC: indirect treatment comparison
MD: mean difference
PRISMA-NMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Network Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension
RCT: randomized controlled trial
RE: random effect
SUCRA: Surface under the cumulative rank curve

Edited by S Brini; submitted 11.Apr.2025; peer-reviewed by E Sahker, P Sangvatanakul; revised version received 29.Nov.2025; accepted 01.Dec.2025; published 08.Jan.2026.

Please cite as:
Tao QF, Hua C, Zhuo X, Mou JJ, Xie CR, Zhang YX, Lv B, Niu XY, Chen M, Zheng H
Face-to-Face Versus Digital, Telephone-Delivered, and Self-Help Cognitive Behavioral Therapy for Irritable Bowel Syndrome: Systematic Review and Bayesian Indirect Treatment Comparison Meta-Analysis
J Med Internet Res 2026;28:e75833
URL: https://www.jmir.org/2026/1/e75833
doi:10.2196/75833

© Qing-Feng Tao, Can Hua, Xiao Zhuo, Jian-Jiao Mou, Chao-Rong Xie, Yu-Xin Zhang, Bei Lv, Xin-Ying Niu, Min Chen, Hui Zheng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Interventions Targeting Healthy and Sustainable Eating Behavior: Systematic Review and Meta-Analysis

Käbi Vanwinkelen, MA; Bram Spruyt, MA; Tim Smits, Prof Dr

Media, Information and Persuasion Lab, Department of Communication Science, KU Leuven, Parkstraat 45, Leuven, Belgium

Corresponding Author:
Käbi Vanwinkelen, MA
Media, Information and Persuasion Lab, Department of Communication Science, KU Leuven, Parkstraat 45, Leuven, Belgium

Abstract

Background: Current food consumption patterns contribute to the rising prevalence of obesity and noncommunicable diseases and exacerbate environmental degradation. Digital media offer promising opportunities to promote healthier and more sustainable eating; yet, evidence regarding their effectiveness remains fragmented.

Objective: The aim of this systematic review and meta-analysis is (1) to evaluate the effectiveness of digital interventions in improving healthy and sustainable food consumption and (2) to identify which participant and intervention characteristics are associated with greater effectiveness.

Methods: A systematic search was conducted in January 2024 and repeated in September 2025 across Web of Science, Embase, and Scopus, supplemented with forward and backward reference searching. Eligible studies were those with a quasi-experimental or longitudinal design evaluating digital interventions targeting nonclinical populations, with the aim of increasing plant-based food consumption or reducing animal-based food intake. Risk of bias was assessed using the Cochrane risk-of-bias tool. Included interventions were coded for behavior change techniques using the Behavior Change Taxonomy version 1. A random-effects meta-analysis with robust variance estimation was performed, and moderator analyses were conducted with participant and intervention characteristics.

Results: Eligibility screening led to the inclusion of 52 papers published between 2004 and 2025, with 24,652 participants in total. The meta-analysis revealed a small but statistically significant positive effect of digital interventions on food consumption outcomes (d=0.33, 95% CI 0.25 - 0.42; P<.001). However, substantial heterogeneity (I^2=86%, 95% prediction interval −0.21 to 0.87) indicates considerable variation in effectiveness across intervention characteristics. A moderator analysis showed no significant difference in effectiveness (P=.53) between interventions aimed at reducing meat consumption (d=0.38, 95% CI 0.20 - 0.57; P<.001) and those promoting plant-based eating (d=0.33, 95% CI 0.23 - 0.42; P<.001). Although digital interventions had the strongest effects among young adults (d=0.46, 95% CI 0.30 - 0.61; P<.001), age-related differences were not statistically significant. Intervention effectiveness differed significantly by platform (P=.03), with social media interventions (d=0.65, 95% CI 0.41 - 0.90; P<.001) yielding stronger effects than other modalities. Incorporating prompts or cues significantly enhanced effectiveness (d=0.58 vs d=0.30; P=.04). Although not statistically significant, interventions including social support or behavioral comparison (both d=0.39; P<.001) yielded larger effects. Few studies included adolescents or individuals from lower socioeconomic backgrounds.

Conclusions: This review underscores the innovative potential of digital interventions in improving eating behavior, highlighting how effectiveness varies by intervention design. Social media emerge as particularly promising, likely due to their unique social and interactive features. By pinpointing the contexts and types of digital interventions that most effectively promote plant-based eating, this study provides timely guidance for researchers and practitioners in increasingly digitalized food environments. Nonetheless, more high-quality studies are needed to confirm these insights and address the critical gap among adolescents and low socioeconomic groups.

Trial Registration: PROSPERO CRD42023487955; https://www.crd.york.ac.uk/PROSPERO/view/CRD42023487955

(J Med Internet Res 2026;28:e80821) doi:10.2196/80821

KEYWORDS

diet; healthy eating; digital media; social media; best practices; internet-based intervention; nutrition intervention; digital interventions; systematic review; meta-analysis

XSL•FO
RenderX

Introduction

Food plays a pivotal role in both human and planetary health. Current food consumption patterns are driving the steep increase in obesity and noncommunicable diseases such as diabetes or cancers [1]. Simultaneously, contemporary dietary habits contribute significantly to greenhouse gas emissions, deforestation, and water scarcity, thereby exacerbating environmental degradation [2]. As high intake of animal-based foods plays a substantial role herein, the EAT-Lancet Commission emphasizes the urgency of a worldwide shift to healthy and sustainable diets mainly characterized by a variety of plant-based foods (eg, fruits and vegetables, whole grains, and legumes) and low quantities of animal-based foods [2,3]. Adherence to the EAT-Lancet diet is associated with a lower risk of diabetes, cardiovascular disease, and cancer-related mortality while reducing greenhouse gas emissions by 29% [4-6]. The most profound health and environmental benefits can be achieved by reducing meat consumption and increasing the intake of fruit, vegetables, and legumes [7-9]. Nevertheless, global meat consumption has risen substantially over the past 5 decades and is projected to continue increasing, while intake of fruits, vegetables, and legumes remains inadequate [10-13]. Hence, it is crucial to explore effective strategies for promoting plant-based dietary patterns.

The widespread use of digital media and their capacity to influence consumption patterns through the promotion of unhealthy foods have sparked growing interest among researchers in leveraging them for health interventions [14,15]. Digital interventions can be delivered through a variety of platforms, including mobile apps, SMS text messaging, websites, and perhaps, most notably, social media. While digital intervention studies often still rely on custom-built or more traditional platforms, health researchers are increasingly exploring the potential of social media for dietary interventions and health promotion campaigns [16-19]. These developments highlight the importance of evaluating and comparing the effectiveness of different digital platforms for interventions [20,21].

To achieve dietary behavior change, interventions incorporate behavior change techniques (BCTs) [22]. BCTs are the smallest identifiable components of an intervention that can independently influence behavior [23,24]. Strategically implementing BCTs has been consistently emphasized as essential for developing effective interventions [22,25]. However, it remains unclear which techniques are most effective in improving healthy and sustainable eating and which combinations of BCTs and digital intervention platforms enhance the intervention's effectiveness. Hence, systematically reviewing and identifying the BCTs most strongly associated with the intervention's effects could enhance the design of future programs targeting food consumption [23].

Although digital interventions targeting eating behavior are gaining popularity, prior reviews report mixed conclusions regarding their effectiveness [18,19,26-29]. Notably, many of these reviews were narrow in scope: some focused on specific food categories, such as sugar-sweetened beverages or vegetables [19,30,31], targeted specific age groups [19,32,33], or selectively included specific digital platforms [18,19,33,34]. Moreover, previous systematic reviews that summarize the literature on digital interventions often overlook social media and focus on research-created platforms or (now) outdated digital tools [16,17,19,35]. Systematic reviews that do consider social media use limited keywords in their search strategy or focus exclusively on social media platforms, missing the opportunity to compare traditional digital interventions with social media interventions [18,28,36,37]. In addition, few reviews assess which BCTs are most successful in targeting dietary change through digital interventions, nor do they explore whether effectiveness differs across age groups or socioeconomic groups [38]. This is critical, as individuals with lower socioeconomic status (SES) are disproportionately affected by poor diets, and the interventions that succeed in high SES populations may not be equally effective for low-income groups [19,39].

While narrow reviews are valuable for exploring targeted questions in depth, there is a need for a comprehensive synthesis that compares intervention strategies across different digital media to gain a more profound understanding of the key components that contribute to a successful digital intervention. Therefore, the aim of this systematic review and meta-analysis is (1) to evaluate the overall effectiveness of digital interventions in terms of increased healthy or sustainable eating behavior and (2) to identify when interventions are (more) effective by considering the behavioral goal orientation (prevention vs promotion approach), intervention characteristics (ie, the digital platform and BCTs), and participant characteristics (ie, age and SES).

Methods

The reporting of this systematic review is in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) guidelines (Checklist 1) [40,41], and its protocol was prospectively registered in PROSPERO (CRD42023487955). This study was a secondary analysis of published literature and did not require ethics approval.

Data Sources and Search Strategy

Systematic searches were conducted on January 19, 2024, in Web of Science core collection, Embase, and Scopus, with an updated search run on September 23, 2025, to retrieve newly published papers. An experienced academic librarian was consulted to optimize the search strategy and to ensure that database-specific operators were used correctly. The final search string was limited to peer-reviewed papers published in English and contained keywords related to digital media, interventions, and healthy and sustainable eating behavior (Multimedia Appendix 1). No date restrictions or other published search filters were applied. The reference lists of all included papers were manually searched, and a citation search using Web of Science was conducted. Authors of all included studies were contacted for additional relevant papers. No other online resources or study registries were searched.

XSL•FO
RenderX

Eligibility Criteria

Studies were included if they met the following PICO-based criteria. Reviews, dissertations, books, and conference papers were excluded.

Population

Studies had to recruit a nonclinical population of participants. Studies exclusively including individuals with specific health conditions or specific nutritional needs (eg, diabetes, pregnancy, and eating disorders) were excluded. Similarly, populations primarily composed of individuals with overweight or obesity were not considered because of differences in appetite [42].

Intervention

Eligible studies had to test a digital intervention designed to promote healthy and sustainable eating behavior, specifically by increasing plant-based food consumption or reducing intake of animal-based products. An intervention was defined as "a coordinated set of activities designed to change specified behavior patterns" [43], thus excluding interventions with a one-time exposure, as these rarely allow for sustained behavioral changes. Multicomponent interventions with an offline element were not included, as effects (or lack thereof) cannot be specifically attributed to the digital component.

Comparison

Studies were required to include a form of baseline comparison or control group that could either be inactive (ie, no intervention and waitlist) or active (ie, alternative intervention). Studies had to use a quasi-experimental (eg, randomized controlled trial [RCT]) or longitudinal design.

Outcomes

Eligible studies had to provide an outcome measure related to eating behavior, such as food consumption, intention, choice, or purchase. These outcomes could be related to intake of plant-based foods (eg, fruits, vegetables, and legumes) and animal-based foods (eg, red or processed meat) or adherence to a healthy and sustainable diet (eg, Mediterranean diet).

Data Extraction

KV and BS independently extracted relevant data using the preregistered extraction book, covering study characteristics (eg, authors, year of publication, and country of origin), study design (eg, comparator and research question), population (eg, age, gender, and SES indicators), intervention characteristics (eg, digital medium, duration, BCTs, and approach), outcomes (eg, behavioral outcomes, metrics, and covariates), and results (eg, analysis technique and statistical results). The 3 reviewers collaborated closely to resolve any discrepancies during the data extraction process.

Coding of BCTs

BCTs that were not explicitly reported in the included studies were coded by 2 reviewers (KV and BS) using the Behavior Change Technique Taxonomy version 1 (BCTTv1) [24]. One reviewer (KV) coded all studies, and a second reviewer (BS) randomly double-coded 5, achieving 90% agreement. Both reviewers were certified coders, as they completed the online taxonomy training ("BCTTv1 Online Training"). Discrepancies were resolved via discussion between the 2 reviewers. For studies with multiple intervention groups (IG), BCTs were coded separately for each active arm. When insufficient detail was provided to code the BCTs, related documents (eg, protocols) were consulted.

Risk of Bias Assessment

Included studies were appraised for study quality using the Cochrane risk-of-bias tool [44]. For RCTs, the Revised RoB 2 tool was used; for cluster RCTs, an adapted version of RoB 2; and for nonrandomized studies (NRS), ROBINS-I [45-47]. The risk of bias assessment was conducted with consideration of best practices in food and communication research. While self-reported measures typically carry a moderate risk of bias, this was not considered problematic, as self-reporting is often the most appropriate or even the only feasible method for measuring eating behavior.

Risk of bias assessments were independently conducted by 2 reviewers (KV and BS), and discrepancies were resolved through consultation with a third reviewer (TS). One paper included 2 distinct experimental studies [48], resulting in a total of 53 separate studies evaluated for risk of bias. The majority of RCTs, including all 3 cluster RCTs, had a moderate risk of bias, primarily due to the use of self-report measures or lack of prespecified analysis plan (31/39, 79%; Figure 1) [49]. In total, 5 RCTs were rated as high risk and 3 as low risk. NRS were rated as moderate (8/14, 57%) or high (6/14, 43%) risk (Figure 2), commonly due to baseline confounding and issues similar to those in the RCTs (see Multimedia Appendix 2 [48,50-99] for a detailed overview of the risk of bias assessments).

Figure 1. Visualization of the risk of bias of randomized controlled trials (n=36).

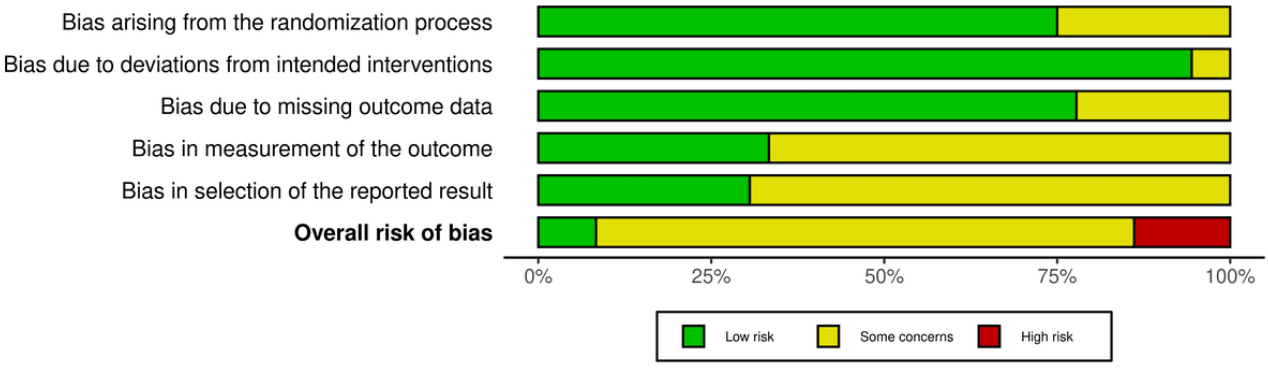

XSL•FO
RenderX

Figure 2. Visualization of the risk of bias of nonrandomized studies (n=14).

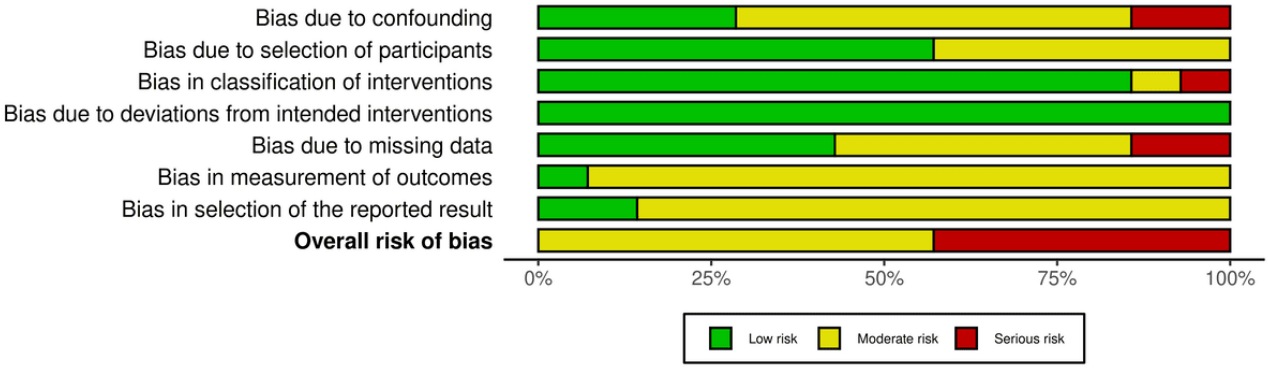

Data Synthesis

To narratively summarize the findings, IGs with differing content were treated as distinct, and identical interventions across studies were grouped [ie, 50-53]. A study was considered to have a consistent effect if it demonstrated a significant improvement in food consumption (ie, within-subjects effect) or if this change was significant compared to a control group (ie, between-subjects effect). For quantitative synthesis, standardized mean differences (Cohen *d*) were calculated. For pre- and postdesigns, Cohen *dav* was estimated, which is based on the mean difference and average SD of both sets of observations [100,101]. For studies that included both an intervention and control group, we calculated the effect size based on the mean pre- and postchange in the treatment group minus the mean pre- and postchange in the control group, divided by the pooled pretest SD [102]. This estimation has proven to be the most recommended effect size for repeated-measure designs in terms of bias and precision [103]. The variance of this effect size was calculated in R (R Foundation for Statistical Computing) with equation 25 of Morris [103], assuming a standard pre- to postintervention correlation of r=0.50. As an estimate of ρ (ie, the correlation of effect sizes) is required to develop efficient weights, several sensitivity analyses were conducted with correlation values of r=0.10 and r=0.90, which led to the standard size of r=0.5 [104]. Missing data were requested from authors and obtained for 2 papers.

A random-effects meta-analysis was performed, as we aimed to estimate the mean of a distribution of effects, and we anticipated heterogeneity in true effect sizes across studies due to variations in study characteristics [105]. Positive effect sizes reflect improvements in healthy or sustainable food intake (ie, increased plant-based consumption or reduced meat consumption), whereas negative values reflect declines. Effect sizes were dependent, as studies often contained multiple IGs compared to a common control, multiple types of food outcomes, or multiple follow-ups. Including effect sizes from the same study in a single model creates complications due to statistical dependence, which violates the assumption of independent sampling errors. To account for this dependency, robust variance estimation was performed in R with the packages *metafor*, *meta*, and *ClubSandwich* [106,107]. This method offers a robust solution to handle dependency, even when the nature of the dependence structure is unknown, by grouping effect sizes based on commonalities (ie, hierarchical clustering) and accounting for the correlation of sampling errors [107].

In total, 6 outliers and 1 intervention with inconsistent results were excluded, leaving 41 papers containing 57 interventions, with 82 effect sizes. Given that NRS represent an important part of the evidence base in digital intervention research and the lack of clear consensus on how to best integrate different types of study designs into meta-analyses [108,109], both RCTs and NRS were included in the meta-analyses. To assess the robustness of the findings, we conducted sensitivity analyses excluding all NRS. The modified method of Hartung-Knapp-Sidik-Jonkman was applied for greater accuracy [110]. Heterogeneity was evaluated using the I^2 statistic and prediction intervals (PIs). Although I^2 is a commonly reported metric, researchers have noted that it may not be the most informative indicator of heterogeneity; PIs are often preferred, as they reflect the distribution of true effects [111,112]. Funnel plots and Egger test were applied to examine small-study effects; however, the interpretation of funnel plots should be approached with caution due to their known limitations [113]. To identify the conditions under which digital interventions differ in effectiveness, we conducted a series of exploratory moderator (ie, subgroup) analyses. More specifically, subgroup analyses were conducted to examine (1) the stability of effects (ie, postintervention vs follow-up), as well as variations in effectiveness based on (2) behavioral goal orientation, (3) participant age, (4) digital medium, and (5) the presence of each distinct BCT cluster. Each moderator was examined in a separate model to avoid overfitting. Moderator analyses were also conducted for individual BCTs, but results are reported only in Multimedia Appendix 3 due to the limited number of studies per BCT and their similarity to the findings from BCT cluster analyses. More information on the meta-analysis can be found in Multimedia Appendix 4 [50,51,78,96,100-107,114].

Results

Study Selection

The database searches identified 9318 records in total. Following the removal of duplicates through the SR Accelerator software (Bond University) [115], 4765 papers were uploaded to Zotero for eligibility screening. Titles and abstracts were screened independently by the first author (KV) and 2 graduate students, followed by full-text screening. Papers that were commonly

XSL•FO
RenderX

included by all screeners received a definitive inclusion. Discrepancies were resolved through discussion with 2 researchers (BS and TS) who had not been involved in the initial screening process. The final database of included papers was checked again by 1 reviewer (BS) to ensure that each paper fulfilled eligibility criteria. The updated search conducted in September 2025 led to the identification of 2283 additional records. After deduplication, 1603 records remained and underwent the same screening procedure, resulting in 6 additional papers being included. This process resulted in 42 papers, with 10 additional studies identified through searching methods, totaling 52 papers. Figure 3 shows the study selection process and provides a more detailed summary of each screening stage, and Figure 4 provides an overview of the updated search.

XSL•FO
RenderX

Figure 3. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study inclusion.

XSL•FO
RenderX

Figure 4. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study inclusion for updated search (September 23, 2025).

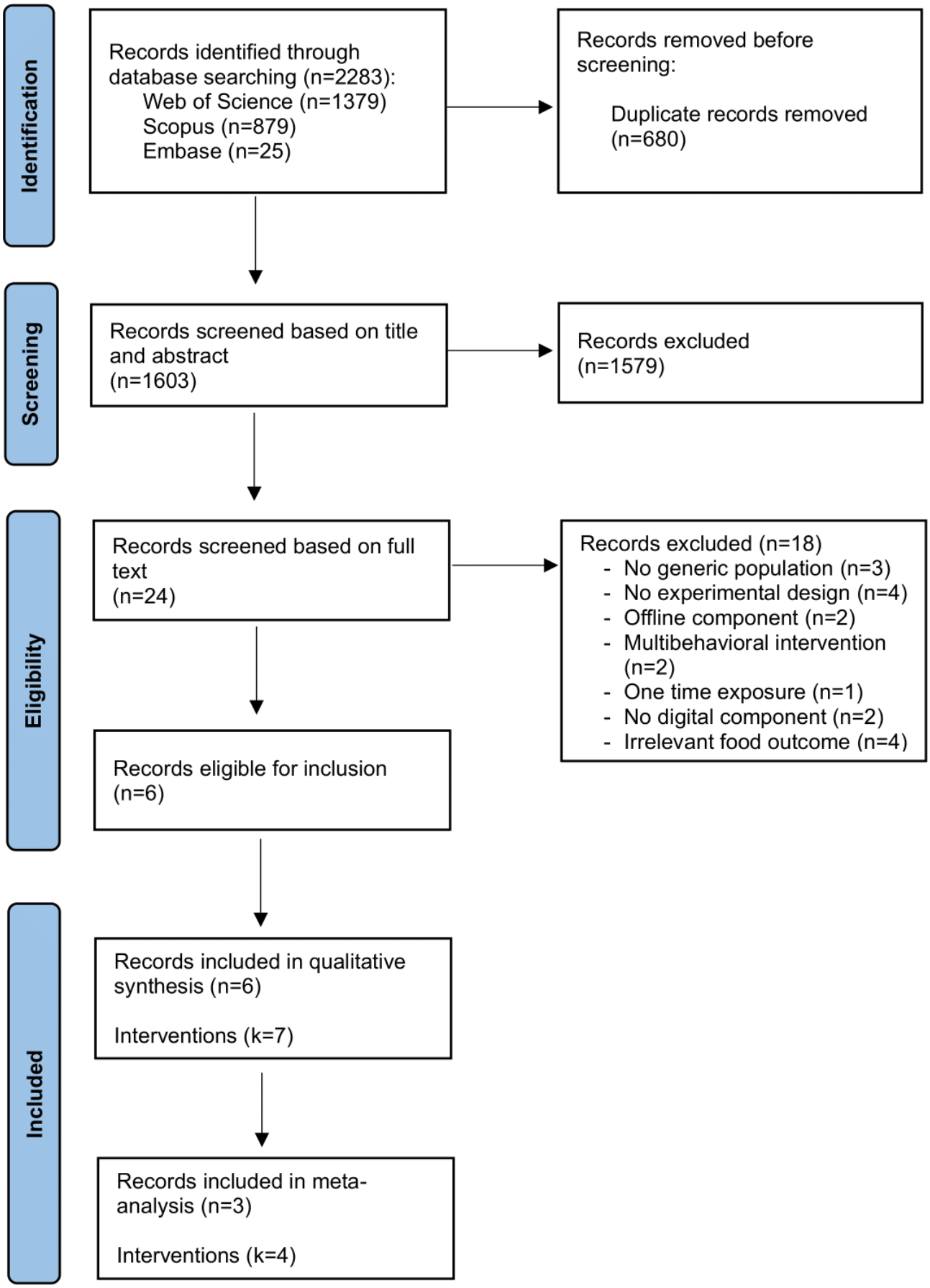

XSL•FO
RenderX

Intervention Characteristics

Of the 52 papers that met the inclusion criteria, 1 paper contained 2 separate experiments [48], while 2 pairs of papers reported findings from the same study conducted among the same sample [54-57]. Therefore, this review contained 51 unique studies, of which most were RCTs (n=38). The included studies were published across a wide range of disciplines, including public health, medical, communication, and psychological journals, reflecting the interdisciplinary nature of research on digital interventions and eating behavior. Most studies were conducted in developed countries, with the majority in the United States (n=10), followed by Italy (n=9), the United Kingdom (n=6), Australia (n=5), and other Western countries such as Denmark (n=2) and Belgium (n=2). Only 4 studies took place in a developing country, specifically in Mexico (n=2), China (n=1), and Saudi Arabia (n=1). All studies were published between 2004 and 2025 and were mostly conducted among adults (n=24), young adults (n=16), or adolescents (n=6). Some focused on children (n=1) or child-parent dyads (n=3), while 1 study did not disclose on population type or age [58]. Sample sizes varied considerably from 28 to 5062 participants at baseline and 24 to 1788 participants after the intervention. Altogether, the studies included 24,652 participants at baseline. More information on the study characteristics can be found in Multimedia Appendix 5 [48,50-99].

Socioeconomic characteristics were inconsistently reported, with only 56% (29/52) of the included studies providing relevant data. The most commonly used indicators were educational level (24/29, 83%), income (8/29, 28%), and occupation or working status (5/29, 17%). One study used food assistance as a proxy for income [54], and 3 incorporated area-based indicators [59-61]. Nearly all studies that included indicators were conducted among adults, and only 2 studies reported socioeconomic characteristics for adolescents. The majority demonstrated a predominance of participants with higher educational attainment, with some reporting up to 95% having tertiary education or International Standard Classification of Education levels 3 to 8. Similarly, among studies reporting income data, the proportion of participants in the lowest income category was small, ranging from 3.5% to 21%, with the exception of 1 study that reported an equal distribution between low- and middle-income groups [62]. While some studies controlled for socioeconomic indicators, few included them as moderators or created subgroups. Only 4 studies explicitly examined the role of SES in intervention effectiveness, of which 3 found significant effects on food intake among lower SES participants [56,57,62]. Additionally, Lim et al [63] found that income correlated with increased legume intake and decreased intention to consume animal-based foods. However, due to the limited and inconsistent reporting of SES as well as the underrepresentation of individuals from lower socioeconomic backgrounds, the planned moderator analysis based on SES was not feasible.

Across the 52 papers, a total of 69 digital interventions (k) were assessed. Intervention duration and intensity varied widely, spanning from 1 week to 6 months. While some were self-paced, most delivered content at intervals ranging from once a week to twice daily. The majority focused on promoting fruit and vegetable intake (FVI; 61%) as an indicator of healthy and sustainable eating behavior. A smaller proportion addressed the reduction of meat consumption (19%) or a broader healthy and sustainable diet (20%; eg, Mediterranean diet). Most interventions used SMS text messaging (41%) or websites (25%), followed by social media (13%), mobile apps (9%), games (7%), and emails (4%).

Use of BCTs

In total, 53 unique BCTs were identified belonging to 15 of the 16 hierarchical clusters of the BCTTv1. The most frequently used BCT clusters were “goals and planning” (65%), “natural consequences” (64%), “feedback and monitoring” (52%), and “comparison of behavior” (41%). The number of BCTs per intervention ranged from 1 to 21, with an average of 6 (SD 4.5). More information on the prevalence of each BCT can be found in Multimedia Appendix 6.

Narrative Summary of Findings by Digital Medium

SMS Text Messaging Interventions

Among the 28 SMS text messaging interventions, 22 (79%) showed significant effects. Most SMS text messaging interventions incorporated BCTs that aimed to inform participants about the consequences of unhealthy or unsustainable eating (BCT cluster 5; 86%). Some facilitated goal setting and planning (BCT cluster 1; 50%) or provided feedback and supported self-monitoring (BCT cluster 2; 25%). SMS text messaging interventions that draw on BCTs from clusters 5 and 2 appeared particularly promising, with significant effects in 79% and 71% of cases, respectively. Those based on cluster 1 were less consistently effective, with just over half (57%) yielding significant outcomes.

In total, 6 of 8 (75%) interventions targeting adults reported effects on eating behavior [59,64-66]. For instance, Carfora and Catellani [64] showed that a 2-week SMS text messaging intervention significantly increased legume intake and reduced meat consumption compared to a passive control group. The most effective messages leveraged dynamic norms that highlight an increase in people engaging in healthy and sustainable eating behavior. In contrast, 2 interventions aiming to improve adults’ FVI were unsuccessful [59,66]. Both studies might have had an insufficient intervention dose to foster behavior change, as participants only received messages 2 times per week or 5 times per month.

Among young adults (aged 18‐30 years), 16 SMS text messaging interventions were evaluated. Nearly all interventions (88%) demonstrated significant positive impacts on eating behavior, particularly in reducing meat intake. In total, 12 interventions decreased red or processed meat consumption, with effects lasting up to 8 weeks after the intervention [67-72]. Incorporating dynamic norms into SMS text messages about the environmental impact of meat seemed to enlarge the effects [71], which aligns with findings among adults [64]. Although most interventions successfully targeted meat reduction, those focusing on increasing intake of plant-based foods (k=4) showed mixed results [63,73,74].

XSL•FO
RenderX

A smaller subset of SMS text messaging interventions focused on adolescents (aged 14 - 19 years; k=4), with 75% reporting significant effects [75,76,116]. One study reported significant differences in FVI change between intervention and control, but this was primarily due to a sharp postintervention decline in the control group rather than an increase in FVI in the IG [75]. On the contrary, Pedersen et al [76] did not find significant improvements in FVI compared to the control group, likely due to low participant engagement, as positive changes were observed only among those who sent more than 50% of the SMS text messages.

Web-Based Interventions

Studies assessing the effectiveness of web-based interventions (k=16) reported limited significant results [50,51,56-62,77-83]: 7 (44%) interventions showed mixed effects due to inconsistent effects across time points or eating outcomes, and 5 found no significant effects. Almost all web-based interventions included BCTs related to goal setting and planning (BCT cluster 1; 94%) and feedback and monitoring (BCT cluster 2; 81%). Other commonly used techniques targeted knowledge (BCT cluster 4; 63%), comparison of behavior (BCT cluster 6; 56%), and social support (BCT cluster 3; 50%). Incorporating feedback and monitoring was effective in 85% of cases, and 80% of the interventions targeting knowledge demonstrated effects.

Among the 14 interventions targeting adults, the most promising were the web-based self-regulation interventions. The study of Plaete et al [50] reported significant improvements in FVI, but the feasibility study only noted an effect for fruit consumption [51]. Similarly, Frie et al [52] and Stewart et al [53] examined the effects of a self-regulation intervention on meat intake, observing reductions in meat consumption 1 week after the intervention but not at 1-month follow-up. Other interventions among adults showed limited success, as they either only had significant effects during the intervention that did not persist afterward [75] or because intake increased only for specific subgroups of the sample [62,81]. In total, 4 web-based interventions did not improve adults' eating behavior compared to the control group [56,57,60,78].

Only 2 interventions targeted the eating behavior of younger age groups. Chamberland et al [61] tested the impact of a web-based school intervention on adolescents aged 14 to 16 years, while Røed et al [80] developed a website for parents that focused on creating a healthy food environment to indirectly improve children's FVI. Both studies showed significant postintervention effects on FVI, but the effects did not persist at 3- to 6-month follow-up.

Social Media Interventions

In total, 9 interventions tested the effectiveness of a social media intervention, with 5 reporting significant effects. The most frequently incorporated BCT clusters were goals and planning (BCT cluster 1; 67%), social support (BCT cluster 3; 44%), natural consequences (BCT cluster 5; 56%), and comparison of behavior (BCT cluster 6; 56%). BCT clusters 3 and 6 appear most promising in social media intervention, as 75% and 80% of the interventions using them demonstrated significant effects.

In total, 4 interventions were conducted among young adults, 4 among adults, and 1 did not report the target group of the intervention. In 1 study of Kilb et al [48], young adults participated as dyads in an intervention, in which senders were asked to post about fruit and vegetables on Facebook, and network members were exposed to these messages. Neither senders nor network members significantly increased their FVI compared to control dyads. In contrast, the second study of this paper found that both private and public self-monitoring via social media increased FVI [48]. Similarly, Meng et al [84] showed that group-based self-tracking on a (researcher-developed) social network website led to a greater increase in FVI compared with individual self-tracking. A recent study of Hawkins et al [85] highlighted that mere exposure to healthy food content on Instagram can improve young adults' FVI.

Among adults, an intervention with support groups reported a significant increase in FVI during the intervention, but these effects were not maintained at follow-up [86]. In Ng et al [87], participants' FVI increased after completing a 4-week intervention containing recipes and videos delivered via Facebook. Carreño Enciso et al [88] tested an educational intervention delivered via Instagram or Facebook but found no significant effects on adherence to the Mediterranean diet. Weber and Nigg [58] tested an intervention containing motivational YouTube videos related to healthy eating. No changes were observed in FVI, which may be due to the low intervention dose (ie, only 6 exposures) and the requirement for participants to actively expose themselves to the intervention.

Mobile App Interventions

Mobile apps were used in 7 interventions. All incorporated feedback and monitoring techniques (BCT cluster 2), while 5 interventions also applied BCTs related to goals and planning (BCT cluster 1), and 4 addressed the consequences of unhealthy eating (BCT cluster 5). However, the effectiveness of these techniques within mobile app interventions appears limited, as only 40% - 57% of the interventions incorporating these clusters reported significant effects.

In total, 3 of 7 app-based interventions reported significant effects. Hendrie et al [89] tested an app that included different sections with recipes and feedback in a real-world setting and found an increase in vegetable intake. The PersuHabit app of Vázquez-Paz et al [90] effectively targeted parents in order to promote FVI among young children, and Liu et al [91] found that a food evaluation app significantly improved the animal-to-plant food ratio in employees' lunches. These studies were conducted among adults, while the 4 interventions showing no or mixed results targeted (pre or late) adolescents (aged 9 - 18 years) [92-94].

Interventions Using Games or Emails

While apps appear to be more effective among adults, games tend to yield better results in child populations. The FoodRateMaster game of Espinosa-Curiel et al [95] significantly increased FVI intake among children. Thompson et al [96] targeted both parents and children to improve children's FVI and found that games were effective in improving children's

XSL•FO
RenderX

diets, but only when they contained action planning. One study tested an intervention featuring 3 games among adults but found no improvement in their FVI [97]. However, it is important to note that this was one of the oldest studies included in this review; therefore, the gaming experience may have differed from those of more recent studies. Since 6 BCT clusters were applied in over 80% of the game-based interventions, it is difficult to explore the most used and promising BCTs in gaming interventions.

A limited number of studies tested email interventions (k=3). Rompotis et al [74] found that habit-based messages providing strategies to strengthen the automaticity of FVI were more effective in improving young adults' fruit intake compared to general nutrition information, regardless of whether the messages were sent via texting or email. Block et al [98] reported similar results for a worksite email intervention among adults. One study did not find significant effects on young adults' FVI [99], which could be explained by the limited intervention dose (30 days with emails every 3 days), compared to the 2 other interventions, which lasted 8 to 12 weeks [74,98]. All of the emailing interventions included BCTs related to goals and planning (BCT cluster 1).

Effectiveness of Digital Interventions

Overview

Of the 52 included studies, 41 provided sufficient data for inclusion in the meta-analysis. The results show a pooled effect size of 0.33 (95% CI 0.25 - 0.42; P<.001), indicating that digital interventions on average have a small, positive effect on healthy and sustainable eating behavior. The forest plot in Figure 5 presents the effect size for each intervention with its 95% CI. Outlier analysis identified 19 potential outliers; however, these were fairly uniformly distributed. Both the graphical representation and sensitivity analysis gave no clear indication of bias (Multimedia Appendix 7). Visual inspection of the funnel plot for small-study effects suggested limited asymmetry (Multimedia Appendix 7), which was supported by the nonsignificant value of the Egger test (P=.18). After removing high risk-of-bias studies, the analysis yielded similar results (d=0.36, 95% CI 0.27 - 0.45; P<.001). The sensitivity analysis excluding 8 effect sizes from NRS yielded results highly consistent with the primary analysis (d=0.32, 95% CI 0.23 - 0.41; P<.001), indicating that the findings are robust to study design. However, heterogeneity statistics revealed substantial heterogeneity (Q_{81}=559.40; P<.001; τ=0.16; τ^2=0.03; I^2=86%), which was supported by the 95% PI, which ranged from −0.21 to 0.87. Given the substantial heterogeneity, we cannot be confident that the positive effect is robust. True effects may vary considerably in settings, with both negative effects as well as strong positive effects being possible. Moderator analyses with categorical variables (ie, subgroup analyses) were conducted to explore between-study variance and provide a more nuanced understanding of the effectiveness of digital interventions. Forest plots for the subgroup analyses can be found in Figure 6.

XSL•FO
RenderX

Figure 5. Forest plot of standardized mean differences (Cohen *d*) for the effect of digital interventions on healthy and sustainable eating (ie, increased plant-based intake or reduced meat intake; for study details, see Multimedia Appendix 5) [48,51-56,58-60,62-71,73-77,79,80,82,84,85,87-90,92-94,96,98,99,116]. IG: intervention group; PI: prediction interval; SES: socioeconomic status; SMD: standardized mean difference.

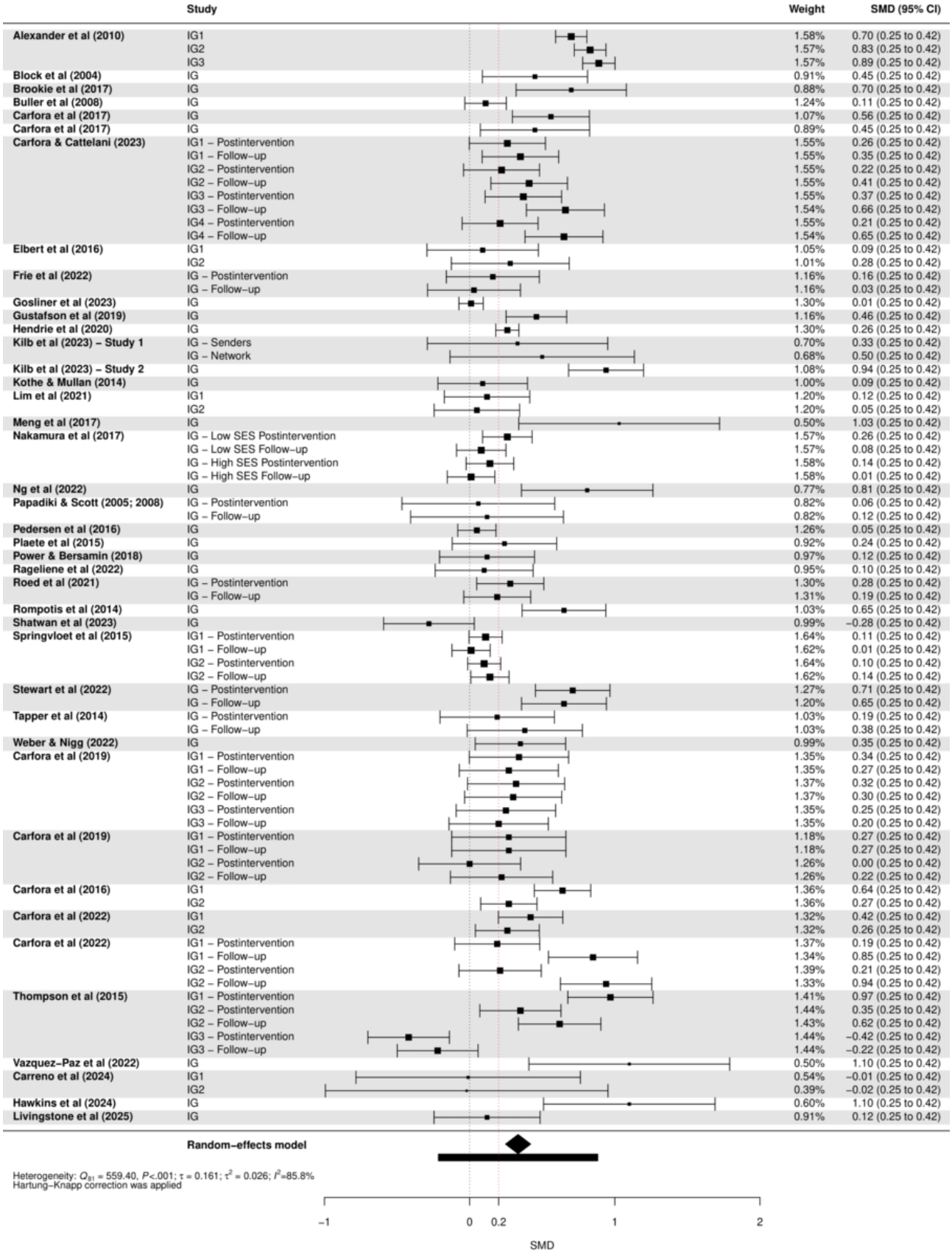

XSL•FO
RenderX

Figure 6. Forest plot of the pooled standardized mean differences (Cohen *d*) for each subgroup of the moderator analyses. SMD: standardized mean difference.

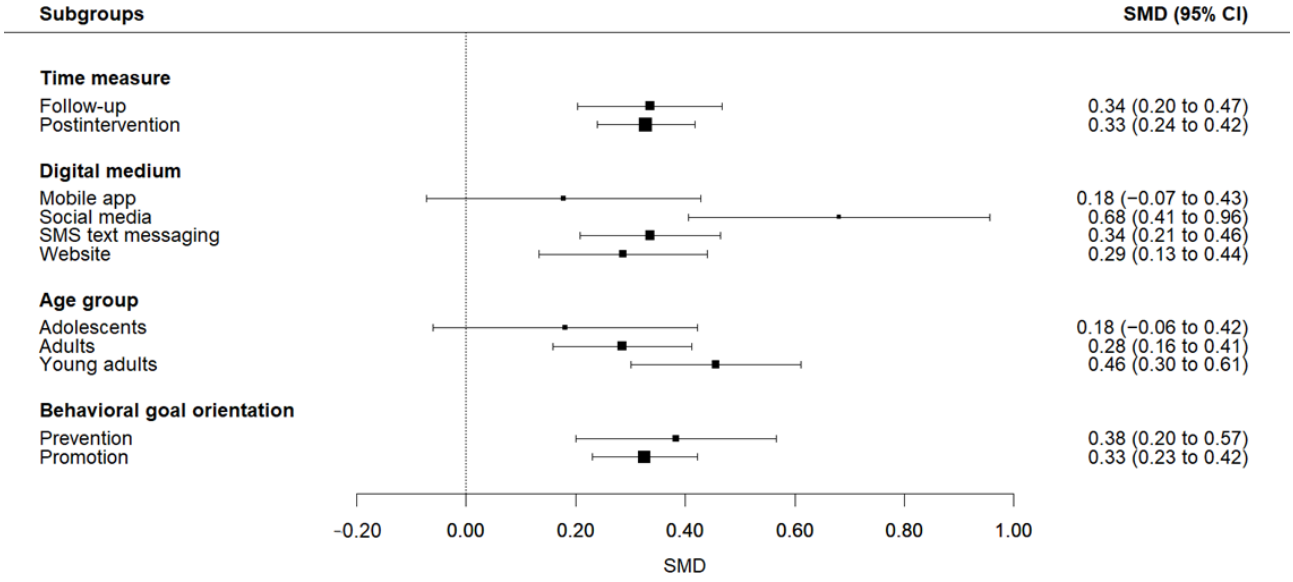

Postintervention and Follow-Up Effects

A meta-analysis with time as covariate yielded similar effect sizes for outcomes measured during or immediately after the intervention (d=0.33, 95% CI 0.24 - 0.42; P<.001; I^2=65%, 95% PI −0.17 to 0.85) and outcomes measured 1 to 6 months after the intervention (d=0.34, 95% CI 0.20 - 0.47; P<.001; I^2=77%, 95% PI −0.29 to 0.84). The test for subgroup differences shows that there are no significant differences between postintervention and follow-up effects (Q_M=0.01, df=1; P=.91). These results held when analyses were restricted to RCTs only (Multimedia Appendix 7). This suggests that, on average, the effectiveness of the digital interventions was quite stable over time, with no meaningful difference in effect between the postintervention and follow-up measures.

Behavioral Goal Orientation

The analysis with behavioral goal orientation as a moderator assessed whether promotion-focused and prevention-focused interventions differed significantly in effectiveness. Digital interventions with a promotion focus (k=47), aimed at encouraging intake of plant-based foods, yielded a statistically significant, small pooled effect size of d=0.33 (95% CI 0.23 - 0.42; P<.001). Heterogeneity was still substantial in this subgroup (I^2=70%, 95% PI −0.23 to 0.88). Prevention-focused interventions (k=12), all of which targeted reductions in red or processed meat consumption, demonstrated a slightly larger pooled effect size of d=0.38 (95% CI 0.20 - 0.57; P<.001) and lower heterogeneity (I^2=64%, 95% PI −0.08 to 0.85). The effect difference between promotion and prevention studies was not significant (Q_M=0.29, df=1; P=.59). The sensitivity analysis restricted to RCTs yielded comparable findings (Multimedia Appendix 7).

Age Group

Including participant age group as a moderator in the meta-analysis revealed that digital interventions had the largest effects among young adults (Table 1). In contrast, the smallest pooled effect size was found among adolescents; however, this finding is based on only 6 interventions. The test of moderators did not indicate a significant difference in pooled effect sizes across age groups (Q_M=4.51, df=2; P=.10). Effect estimates remained consistent when the analysis was restricted to RCTs (Multimedia Appendix 7).

Table . Meta-analysis for the effect of digital interventions with age group as moderator.

Age group	*k*	*d*	95% CI	I^2 (%)	95% PI[a]
Adults	26	0.28[b]	0.16 to 0.41	64	−0.19 to 0.75
Young adults	19	0.46[b]	0.30 to 0.61	65	−0.10 to 1.00
Adolescents	6	0.18	−0.06 to 0.42	86	−0.45 to 0.84

[a]PI: prediction interval.

[b]P<.001.

Digital Medium

The analysis with type of digital medium as moderator was significant (Q_M=8.49, df=3; P=.03), indicating that the mode of delivery modified the pooled effect. Interventions delivered via social media or SMS text messages yielded the largest effect sizes (Table 2). Despite the low number of social media interventions, pairwise comparison showed a significant subgroup difference with higher effectiveness for social media than for SMS text messaging (z=2.23; P=.03), website (z=2.56;

XSL•FO
RenderX

P=.01), and mobile app interventions (*z*=2.64; *P*=.01). The heterogeneity score for social media interventions was moderate, and the positive PI suggests that future social media interventions are likely to yield positive effects, although the lower bound is close to 0 (I^2=47%, 95% PI 0.01 - 1.31). Exclusion of NRS did not meaningfully alter the results, except that the subgroup difference between social media and SMS text messaging interventions was no longer statistically significant (*z*=1.72; *P*=.09). Additionally, the sensitivity analysis for mobile app interventions could not be conducted due to an insufficient number of studies (Multimedia Appendix 7).

Table . Meta-analysis for the effect of digital interventions with digital medium as moderator.

Digital medium	*k*	*d*	95% CI	I^2 (%)	95% PI[a]
SMS text messaging	24	0.34[b]	0.21 to 0.47	62	−0.05 to 0.73
Website	13	0.28[b]	0.13 to 0.43	77	−0.25 to 0.80
Social media	8	0.65[b]	0.41 to 0.90	47	0.01 to 1.31
Mobile app	5	0.18	−0.07 to 0.43	72	−0.53 to 0.93

[a]PI: prediction interval.
[b]*P*<.001.

BCT Clusters

Meta-analyses with the presence of each BCT cluster found the largest standardized mean difference for interventions incorporating prompts or cues (BCT cluster 7), with a pooled effect size of *d*=0.62 (95% CI 0.33 - 0.91; *P*<.001). The test of moderators indicated that interventions including this BCT cluster were significantly associated with a greater improvement in eating behavior than those without this cluster (*P*=.04). Interventions including social support (cluster 3) and comparison of behavior (cluster 6) demonstrated higher pooled effect sizes (*d*=0.39) compared to those that did not include these BCTs (*d*=0.31 and *d*=0.29, respectively). Although these differences were not statistically significant (*P*=.40 and *P*=.17, respectively), the numerically higher effect sizes suggest that these BCTs may still contribute positively to intervention outcomes. The lack of significance could be attributed to the limited number of studies incorporating these BCT clusters, underscoring the need for further research.

Several BCT clusters showed similar effect sizes regardless of whether they were present in interventions or not. For instance, feedback and monitoring (cluster 2), repetition and substitution (cluster 8), and antecedents (cluster 12) had comparable pooled effect sizes in both subgroups, suggesting no clear added value of including these BCTs (Table 3). Interventions that incorporated the BCT clusters “shaping knowledge” (cluster 4) and “natural consequences” (cluster 5) showed smaller pooled effect sizes compared to those in which these clusters were absent; however, these differences were not statistically significant (*P*=.07 and *P*=.35, respectively). Overall, restricting these moderator analyses to RCTs produced similar results (Multimedia Appendix 7). Moderator analyses were also conducted for individual BCTs, which generated similar findings (Multimedia Appendix 3).

XSL•FO
RenderX

Table . Meta-analysis for the effect of digital interventions with behavior change technique (BCT) cluster as moderator.

BCT cluster	BCT cluster present			BCT cluster absent			Test of moderators
	k	*d*	95% CI	*k*	*d*	95% CI	*P* value
1. Goals and planning	37	0.31[a]	0.19 to 0.43	20	0.37[a]	0.24 to 0.49	.56
2. Feedback and monitoring	32	0.33[a]	0.20 to 0.47	25	0.33[a]	0.22 to 0.44	>.99
3. Social support	16	0.39[a]	0.21 to 0.56	41	0.31[a]	0.22 to 0.40	.40
4. Shaping knowledge	27	0.25[b]	0.10 to 0.39	30	0.40[a]	0.29 to 0.52	.07
5. Natural consequences	39	0.30[a]	0.22 to 0.39	18	0.39[b]	0.18 to 0.59	.35
6. Comparison of behavior	27	0.39[a]	0.27 to 0.51	30	0.29[a]	0.18 to 0.39	.17
7. Associations	6	0.58[a]	−0.10 to 1.25	51	0.30[a]	0.23 to 0.38	.04
8. Repetition and substitution	12	0.30[b]	0.15 to 0.45	45	0.34[a]	0.24 to 0.44	.64
9. Comparison of outcomes	9	0.38	−0.06 to 0.83	48	0.32[a]	0.24 to 0.41	.60
10. Reward and threat	13	0.36[b]	0.13 to 0.60	44	0.32[a]	0.23 to 0.42	.72
12. Antecedents	6	0.33[b]	0.11 to 0.55	51	0.33[a]	0.24 to 0.42	.97
13. Identity	10	0.19[c]	0.07 to 0.30	47	0.36[a]	0.27 to 0.46	.09
14. Scheduled consequences	4	0.29	−0.74 to 1.32	53	0.34[a]	0.25 to 0.42	.78
15. Self-belief	7	0.29	−0.09 to 0.67	50	0.34[a]	0.24 to 0.43	.75

[a]*P*<.001.

[b]*P*<.01.

[c]*P*<.05.

Discussion

Principal Findings

While the meta-analysis found a moderate and statistically significant overall effect of digital interventions on healthy and sustainable food consumption, the substantial heterogeneity observed across studies suggests that these interventions do not have a consistent effect across all populations and contexts. Additionally, included studies mostly had a moderate risk of bias, and together with the inconsistency in effects, this may reduce the certainty of the evidence regarding the overall effectiveness of digital interventions. Both the narrative review and the moderator analyses showed that the effectiveness of digital interventions varied by multiple characteristics and settings.

Regarding the food outcomes, interventions preventing animal-based food consumption and those promoting plant-based food intake both yielded a significant, moderate effect. While interventions with a prevention focus yielded a slightly larger pooled effect size, moderator analyses showed no significant difference with interventions promoting plant-based food intake. This contrasts with prior research, suggesting that a promotion focus is more effective in encouraging healthy dietary patterns [26,117]. However, this might be explained by a difference in the food consumption studied. For example, the studies reviewed by Zlatevska et al [117] primarily targeted discretionary items such as sugary drinks and snacks, not meat consumption. Only 1 of the included interventions in this review combined a prevention and promotion strategy and found sustained effects at follow-up [64], highlighting a need for future research to test more integrative approaches that align closely to the EAT-Lancet diet.

Both the narrative synthesis and moderator analyses show that the effectiveness of digital interventions varies by modality. SMS text messaging interventions had a moderate effect on healthy and sustainable food intake, with over 75% demonstrating significant results in the narrative review. Despite the small number of social media interventions, they had the strongest effects on eating behavior, being significantly associated with larger effect sizes than other modes of delivery. While prior research associates social media use with unhealthy eating habits [118,119], these findings suggest that social media

XSL•FO
RenderX

can also be used to positively influence dietary behavior. Since many consumers engage with food content on social media, platforms such as Instagram or Facebook hold great potential to promote healthier and more sustainable eating [120,121]. Social media platforms present unique environments in which users can form large social networks, allowing them to seek information about others' behavior and receive positive reinforcements for their own behavior. Although the evidence base remains limited, social media interventions may promote stronger behavior change for several reasons: they expose users to peers' behaviors (social modeling), leverage participants' existing engagement with the platform to sustain intervention exposure and retention, and offer interactive features that facilitate social support [122,123]. Notably, none of the social media interventions targeted adolescents, despite this being a key demographic due to their high engagement with social media and high exposure to unhealthy food marketing [124,125].

These findings suggest that interventions delivered through accessible and familiar platforms (ie, SMS text messaging or social media) tend to be more effective than interventions requiring more intentional user engagement (ie, stand-alone websites or apps). This difference also links to the amount of agentic demand (ie, the degree to which participants are required to engage with the content in order to achieve the intended outcome) [126,127], though this requires further empirical testing. The common academic practice of developing digital interventions via research-created platforms, often used only briefly and actively throughout the study period, implies a misalignment with real-world digital behavior [35]. To enhance ecological validity and long-term behavior change, future studies could focus on accessible interventions using existing, familiar platforms [126,128]. While social media interventions were significantly more effective than interventions via other platforms, they also present specific challenges for researchers. As opposed to more controlled intervention environments such as apps or websites, social media constitute open and dynamic spaces that are saturated with unhealthy food marketing and health misinformation [129]. This may dilute or even counteract intervention effects and can lead to nutrition confusion [130].

A wide range of BCTs was incorporated across the interventions, and the choice of BCT varied depending on the type of digital medium used. BCTs targeting goals and planning were most frequently applied across all interventions, particularly in web-based interventions. Knowledge-related techniques, such as providing instructions, were frequently applied, especially in web-based interventions or games. Communicating the consequences of unhealthy eating behavior was also a commonly used cluster of techniques, mainly in SMS text messaging interventions. Yet, these information-based BCTs were not the most effective ones. Although definitive conclusions on the most effective BCT cluster cannot be made due to the limited number of studies incorporating certain clusters, moderator analyses [1] indicated that interventions incorporating prompts or cues (BCT cluster 7) were significantly more effective than those that did not, yielding a large effect size. Additionally, interventions targeting social mechanisms through social support (cluster 3) or comparison of behavior (cluster 6) demonstrated larger effects, though not significantly larger than those not including these BCTs. The narrative synthesis further suggests that social norm communication is an effective strategy, as several studies with significant effects incorporated descriptive norms to influence behavior. These findings suggest that future interventions may benefit from shifting emphasis away from purely informational strategies (eg, raising awareness or increasing knowledge) and instead testing techniques that leverage social influence, peer dynamics, and cues. Previous research on health interventions also suggests that providing social support strengthens the impact of digital interventions [131-133], and that norm communication can gradually reshape individuals' perceptions of others' behavior, prompting them to align with these evolving norms [134-136].

Finally, while age-related differences in the effect size of the interventions were not significant, moderator analyses yielded the strongest effect size for young adults. Due to the limited number of studies targeting adolescents, conclusions for this age group cannot be drawn. Regarding SES, most studies included SES indicators only as control variables rather than focusing on targeted recruitment, resulting in the underrepresentation of lower SES individuals. This limited our ability to conduct subgroup analyses for this population. This lack of research is concerning, given that food intake is socially structured, with unhealthy and unsustainable diets being more prevalent among lower SES individuals [137-139]. While financial constraints and educational disparities contribute to these patterns, modifiable psychosocial factors (eg, attitudes, literacy, desired identity, or social norms) also play a critical role [140-142]. Despite potential challenges related to digital access and literacy, it is crucial to reach lower SES populations in the current digital era to help mitigate health inequalities related to dietary behavior, especially adolescents, given that they are highly active on digital platforms. Leveraging platforms such as social media may help embed interventions within their daily routines and reduce barriers to participation. Future research should conduct both targeted testing of interventions within lower SES groups and exploration of SES moderation effects to understand intervention effectiveness across socioeconomic groups.

This review reveals some major gaps in the literature. First, individuals from lower socioeconomic backgrounds remain underrepresented in research testing digital interventions. None of the included studies used subjective measures of SES (ie, perceptions of own social status) [2] despite evidence that perceived SES is a strong predictor of health (behavior) even after controlling for objective indicators [143-145]. Future research should therefore not only systematically measure SES but also incorporate subjective indicators more frequently [146]. Second, while the included interventions targeted various age groups, few studies specifically focused on adolescents. The handful of studies that included adolescents as participant group primarily tested web-based or mobile app interventions, which were among the least effective digital modalities. Both the lack of research as well as the prominent focus on websites or apps may explain the smaller effect size for this demographic. Future research should prioritize low-agentic interventions for adolescents of different socioeconomic backgrounds, particularly through social media, as adolescents not only increasingly use

XSL•FO
RenderX

these platforms, but research also shows that exposure to social media food content influences their eating behavior [118,124]. Finally, most studies focused on specific food groups, particularly fruits and vegetables, framing them as key components of a healthy diet. These findings are similar to previous systematic reviews on interventions aiming to improve food consumption [33,147]. While fruits and vegetables are indeed a critical part of a healthy and sustainable diet, interventions should also highlight other important components, such as legumes and whole grain products. Furthermore, rather than solely focusing on reducing meat consumption, interventions can promote replacing meat with whole-food plant-based alternatives, as substituting certain products is often more achievable than complete elimination [18,64].

Implications

By identifying the conditions in which digital interventions are most effective for encouraging plant-based eating, this study offers timely guidance for practitioners in a rapidly evolving digital landscape. The results of this review suggest that digital interventions can effectively be implemented to support people in transitioning to healthier and more sustainable eating behavior. Given the variability in effectiveness following intervention characteristics, health care providers and practitioners should consider selecting digital platforms that best fit the preferences, literacy, and lifestyle of the target group. Social media may be particularly useful for delivering low-threshold interventions in community settings, targeting mechanisms of social influence for behavior change. Practitioners are encouraged to adopt evidence-based BCTs that leverage social influence (ie, social support and demonstration of behavior) and thus shift emphasis from purely informational strategies (eg, raising awareness or increasing knowledge). Collaboration between behavioral scientists and health care professionals can help ensure that interventions are evidence-based and aligned with participants' needs. Special attention is needed for the participation of underrepresented groups, more specifically, adolescents and individuals with lower SES. Future research should use specific strategies to recruit and engage these vulnerable populations, for example, through school-based programs or partnerships with community organizations. Moreover, it is important to ensure that intervention materials are accessible and inclusive to help reduce disparities in participation and retention. One way to ensure this is by actively involving members of the vulnerable group or trusted intermediaries in the design of the intervention [148].

Limitations

Despite using extensive search strategies, relevant research may have been overlooked if published in languages other than English or if they were inaccessible through the selected databases. The majority of included studies demonstrated moderate risk of bias, with measurement of outcome bias being the most prevalent concern due to reliance on self-reported dietary measures rather than objective assessments. This may lead to overestimation of intervention effects, as self-report measures are susceptible to social desirability bias and recall error. Future research should prioritize objective outcome measures to strengthen evidence for real-world applicability. In total, 11 studies could not be included in the meta-analysis due to insufficient data for effect size calculation. Additionally, the inconsistency in measuring SES and the scarcity of research involving participants from lower SES backgrounds hindered our ability to assess the differential effects of digital interventions across SES groups. Moreover, the small number of studies contributing data to certain moderators may have limited the statistical power to detect potential moderation effects [149] and restricted our ability to assess their combined influence. Ideally, a sufficiently large number of studies would be available to allow the modeling of joint effects, rather than examining each moderator separately. Furthermore, most evidence was derived from studies with a moderate risk of bias, and substantial heterogeneity was present. Although the subgroup analyses explained some of the observed variation, these factors may still affect the overall certainty of the evidence. We did not perform a formal Grading Recommendations Assessment, Development, and Evaluation assessment to systematically rate the certainty of the evidence, which limits our ability to evaluate the confidence in our findings [150].

Another potential limitation is the inclusion of both RCTs and NRS in the meta-analysis. The appropriateness of combining different study designs remains debated in the methodological literature, with studies having inconsistent conclusions about the risks of pooling RCTs and NRS (eg, achieving statistical significance only after inclusion of NRS) [109,151,152]. We retained NRS in our primary analyses for 2 reasons. First, sensitivity analyses excluding NRS demonstrated that our findings were robust, with effect estimates, CIs, and heterogeneity statistics remaining consistent. Second, NRS are increasingly recognized as valuable sources of evidence, also within digital intervention research, as ecological validity and real-world application are important criteria. In this context, NRS can provide complementary evidence alongside RCTs [153,154]. However, the pooled estimates should therefore be interpreted with awareness of some methodological considerations. RCTs and NRS are susceptible to different types of bias [151,152]. RCTs face risks related to randomization, allocation concealment, and blinding, while NRS are prone to confounding and selection bias, which is reflected in the different risk-of-bias assessment tools used for each design. The pooled estimates, therefore, represent weighted averages that incorporate these different types of bias, which may complicate direct translation of the results to practice [151,152]. Moreover, these study designs provide different types of evidence: RCTs answer questions of effectiveness under controlled conditions, while NRS can, for instance, address generalizability or real-world performance [151,153]. Consequently, the primary findings should be interpreted as reflecting the overall evidence base for digital interventions, rather than as a pure estimate of intervention effectiveness derived solely from controlled trials. Importantly, the consistency of the findings in the sensitivity analyses restricted to RCTs supports the robustness of the study's main conclusions regarding intervention effectiveness.

Other important limitations are with regard to the BCTs. Categorizing intervention components into BCTs based on study descriptions and protocols proved challenging, as studies used varying terminology and levels of detail on the intervention

XSL•FO
RenderX

content. Similar issues have been highlighted in previous systematic reviews [36,146]. Therefore, despite the use of a coding manual and following a training in BCT taxonomy, the presence of some techniques may have been overlooked. To enhance the replicability of interventions and accuracy of coding BCT presence for meta-analyses, future research should describe intervention content in greater detail and more systematically by following taxonomies or reporting guidelines [24,155]. It is also essential to recognize that BCTs are not exhaustive; they possess a certain level of superficiality. While BCTs can be implemented in various ways, the content was generalized to a certain type of BCT, potentially masking variations in their implementation. For instance, informing participants about the components of a healthy diet or providing personalized recipes both fall under the same BCT category (4.1 Instruction on how to perform the behavior) but represent distinct approaches that may yield different effects. Moreover, some characteristics of persuasive communication are overlooked by focusing on BCTs (eg, framing of information).

Conclusions

This review provides a comprehensive overview of digital interventions, suggesting that digital interventions can effectively improve eating behavior, though their success varies by intervention design and population targeted. Social media emerge as particularly promising, likely due to their unique social and interactive features. Importantly, the evidence base mainly consists of studies with a moderate risk of bias, highlighting the need for more high-quality studies to confirm current results. Moreover, the meta-analytic results have a broad PI, indicating that while the average effect is positive, individual interventions may range from highly effective to potentially ineffective, depending on context and design. To our knowledge, it is one of the first reviews to systematically code the characteristics of digital interventions, including their mode of delivery (ie, digital medium), content (ie, BCTs), behavioral goal orientation (prevention vs promotion), and targeted demographic (ie, age and SES), and to link these with the intervention effect size. Despite our aim to explore effects specifically among low SES groups, the limited available research restricted our ability to conduct subgroup analyses for this population. Our findings offer valuable insights for practitioners and researchers interested in leveraging digital media for behavior change by providing an evidence base on the contexts and types of digital interventions that most effectively promote plant-based eating.

Acknowledgments

The authors would like to thank Emma Sageot and Steffi Bellekens for their support with the title, abstract, and full-text screening. The authors would also like to express their appreciation to Veerle Tuerlinckx, who helped optimize the search string.

Funding

This work was supported by the FEAST project funded by the European Research Executive Agency (grant 101060536) and the HashTagToFork project funded by Internal Funds KU Leuven (grant C2M/23/007). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Data Availability

All data generated or analyzed during this study are included in Multimedia Appendix 5.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search string.
[DOCX File, 22 KB - jmir_v28i1e80821_app1.docx]

Multimedia Appendix 2
Risk of bias.
[PDF File, 174 KB - jmir_v28i1e80821_app2.pdf]

Multimedia Appendix 3
Moderator analyses—behavior change techniques.
[DOCX File, 27 KB - jmir_v28i1e80821_app3.docx]

Multimedia Appendix 4
Calculation of effect sizes and meta-analysis.
[DOCX File, 21 KB - jmir_v28i1e80821_app4.docx]

XSL•FO
RenderX

Multimedia Appendix 5
Study characteristics.
[XLSX File, 27 KB - jmir_v28i1e80821_app5.xlsx]

Multimedia Appendix 6
Prevalence of behavior change techniques.
[DOCX File, 24 KB - jmir_v28i1e80821_app6.docx]

Multimedia Appendix 7
Sensitivity analyses.
[PDF File, 215 KB - jmir_v28i1e80821_app7.pdf]

Checklist 1
PRISMA and PRISMA-S checklist.
[DOCX File, 280 KB - jmir_v28i1e80821_app8.docx]

References

1. Sun X, Yon DK, Nguyen TT, et al. Dietary and other lifestyle factors and their influence on non-communicable diseases in the Western Pacific region. Lancet Reg Health West Pac 2024 Feb;43:100842. [doi: 10.1016/j.lanwpc.2023.100842] [Medline: 38456094]
2. Willett W, Rockström J, Loken B, et al. Food in the anthropocene: the EAT–Lancet Commission on healthy diets from sustainable food systems. Lancet 2019 Feb;393(10170):447-492. [doi: 10.1016/S0140-6736(18)31788-4]
3. Clark M, Tilman D. Comparative analysis of environmental impacts of agricultural production systems, agricultural input efficiency, and food choice. Environ Res Lett 2017 Jun 1;12(6):064016. [doi: 10.1088/1748-9326/aa6cd5] [Medline: 37829169]
4. Bui LP, Pham TT, Wang F, et al. Planetary Health Diet Index and risk of total and cause-specific mortality in three prospective cohorts. Am J Clin Nutr 2024 Jul;120(1):80-91. [doi: 10.1016/j.ajcnut.2024.03.019] [Medline: 38960579]
5. Karavasiloglou N, Thompson AS, Pestoni G, et al. Adherence to the EAT-Lancet reference diet is associated with a reduced risk of incident cancer and all-cause mortality in UK adults. One Earth 2023 Dec 15;6(12):1726-1734. [doi: 10.1016/j.oneear.2023.11.002] [Medline: 38130482]
6. Knuppel A, Papier K, Key TJ, Travis RC. EAT-Lancet score and major health outcomes: the EPIC-Oxford study. Lancet 2019 Jul;394(10194):213-214. [doi: 10.1016/S0140-6736(19)31236-X]
7. Wallace TC, Bailey RL, Blumberg JB, et al. Fruits, vegetables, and health: a comprehensive narrative, umbrella review of the science and recommendations for enhanced public policy to improve intake. Crit Rev Food Sci Nutr 2020;60(13):2174-2211. [doi: 10.1080/10408398.2019.1632258] [Medline: 31267783]
8. Westhoek H, Lesschen JP, Rood T, et al. Food choices, health and environment: effects of cutting Europe's meat and dairy intake. Glob Environ Change 2014 May;26:196-205. [doi: 10.1016/j.gloenvcha.2014.02.004]
9. Yanni AE, Iakovidi S, Vasilikopoulou E, Karathanos VT. Legumes: a vehicle for transition to sustainability. Nutrients 2023 Dec 27;16(1):98. [doi: 10.3390/nu16010098] [Medline: 38201928]
10. Daily consumption of fruit and vegetables by sex, age and educational attainment level. Eurostat. 2022 Apr 4. URL: https://ec.europa.eu/eurostat/databrowser/view/HLTH_EHIS_FV3E__custom_1588514/default/table?lang=en [accessed 2025-12-17]
11. Gibbs J, Cappuccio FP. Plant-based dietary patterns for human and planetary health. Nutrients 2022 Apr 13;14(8):1614. [doi: 10.3390/nu14081614] [Medline: 35458176]
12. Hughes J, Pearson E, Grafenauer S. Legumes—a comprehensive exploration of global food-based dietary guidelines and consumption. Nutrients 2022 Jul 27;14(15):3080. [doi: 10.3390/nu14153080] [Medline: 35956258]
13. Vranken L, Avermaete T, Petalios D, Mathijs E. Curbing global meat consumption: emerging evidence of a second nutrition transition. Environ Sci Policy 2014 May;39:95-106. [doi: 10.1016/j.envsci.2014.02.009]
14. Evans WD, Abroms LC, Broniatowski D, et al. Digital media for behavior change: review of an emerging field of study. Int J Environ Res Public Health 2022 Jul 26;19(15):9129. [doi: 10.3390/ijerph19159129] [Medline: 35897494]
15. Folkvord F. The promotion of healthy foods: a review of the literature and theoretical framework. In: The Psychology of Food Marketing and Overeating: Routledge; 2019. [doi: 10.4324/9780429274404-8]
16. Beck Silva KB, Miranda Pereira E, Santana MD, Costa PRF, Silva RDC. Effects of computer-based interventions on food consumption and anthropometric parameters of adolescents: a systematic review and metanalysis. Crit Rev Food Sci Nutr 2024;64(6):1617-1631. [doi: 10.1080/10408398.2022.2118227] [Medline: 36062829]
17. Chen Y, Perez-Cueto FJA, Giboreau A, Mavridis I, Hartwell H. The promotion of eating behaviour change through digital interventions. IJERPH 2020;17(20):7488. [doi: 10.3390/ijerph17207488]

XSL•FO
RenderX

18. Hsu MSH, Rouf A, Allman-Farinelli M. Effectiveness and behavioral mechanisms of social media interventions for positive nutrition behaviors in adolescents: a systematic review. J Adolesc Health 2018 Nov;63(5):531-545. [doi: 10.1016/j.jadohealth.2018.06.009] [Medline: 30197198]
19. Livingstone KM, Rawstorn JC, Partridge SR, et al. Digital behaviour change interventions to increase vegetable intake in adults: a systematic review. Int J Behav Nutr Phys Act 2023 Mar 27;20(1):36. [doi: 10.1186/s12966-023-01439-9] [Medline: 36973716]
20. Al-Dhahir I, Reijnders T, Faber JS, et al. The barriers and facilitators of eHealth-based lifestyle intervention programs for people with a low socioeconomic status: scoping review. J Med Internet Res 2022 Aug 24;24(8):e34229. [doi: 10.2196/34229] [Medline: 36001380]
21. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010 Feb 17;12(1):e4. [doi: 10.2196/jmir.1376] [Medline: 20164043]
22. Michie S, West R, Sheals K, Godinho CA. Evaluating the effectiveness of behavior change techniques in health-related behavior: a scoping review of methods used. Transl Behav Med 2018 Mar 1;8(2):212-224. [doi: 10.1093/tbm/ibx019] [Medline: 29381786]
23. Carey RN, Connell LE, Johnston M, et al. Behavior change techniques and their mechanisms of action: a synthesis of links described in published intervention literature. Ann Behav Med 2018;10. [doi: 10.1093/abm/kay078]
24. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
25. Michie S, Abraham C, Whittington C, McAteer J, Gupta S. Effective techniques in healthy eating and physical activity interventions: a meta-regression. Health Psychol 2009 Nov;28(6):690-701. [doi: 10.1037/a0016136] [Medline: 19916637]
26. Kachwaha S, Kim SS, Das JK, et al. Behavior change interventions to address unhealthy food consumption: a scoping review. Curr Dev Nutr 2024 Mar;8(3):102104. [doi: 10.1016/j.cdnut.2024.102104] [Medline: 38482184]
27. Hedin B, Katzeff C, Eriksson E, Pargman D. A systematic review of digital behaviour change interventions for more sustainable food consumption. Sustainability 2019;11(9):2638. [doi: 10.3390/su11092638]
28. Klassen KM, Douglass CH, Brennan L, Truby H, Lim MSC. Social media use for nutrition outcomes in young adults: a mixed-methods systematic review. Int J Behav Nutr Phys Act 2018 Jul 24;15(1):70. [doi: 10.1186/s12966-018-0696-y] [Medline: 30041699]
29. Prowse R, Carsley S. Digital interventions to promote healthy eating in children: umbrella review. JMIR Pediatr Parent 2021 Nov 25;4(4):e30160. [doi: 10.2196/30160] [Medline: 34842561]
30. Vargas-Garcia EJ, Evans CEL, Prestwich A, Sykes-Muskett BJ, Hooson J, Cade JE. Interventions to reduce consumption of sugar-sweetened beverages or increase water intake: evidence from a systematic review and meta-analysis. Obes Rev 2017 Nov;18(11):1350-1363. [doi: 10.1111/obr.12580] [Medline: 28721697]
31. Wolfenden L, Barnes C, Lane C, et al. Consolidating evidence on the effectiveness of interventions promoting fruit and vegetable consumption: an umbrella review. Int J Behav Nutr Phys Act 2021 Jan 11;18(1):11. [doi: 10.1186/s12966-020-01046-y] [Medline: 33430879]
32. Ashton LM, Sharkey T, Whatnall MC, et al. Effectiveness of interventions and behaviour change techniques for improving dietary intake in young adults: a systematic review and meta-analysis of RCTs. Nutrients 2019 Apr 11;11(4):825. [doi: 10.3390/nu11040825] [Medline: 30979065]
33. Ghammachi N, Dharmayani PNA, Mihrshahi S, Ronto R. Investigating web-based nutrition education interventions for promoting sustainable and healthy diets in young adults: a systematic literature review. Int J Environ Res Public Health 2022 Feb 1;19(3):1691. [doi: 10.3390/ijerph19031691] [Medline: 35162714]
34. Curtin E, Green R, Brown KA, et al. The effectiveness of mobile app-based interventions in facilitating behaviour change towards healthier and more sustainable diets: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2025 Sep 30;22(1):122. [doi: 10.1186/s12966-025-01823-7] [Medline: 41035009]
35. Hingle M, Patrick H. There are thousands of apps for that: navigating mobile technology for nutrition education and behavior. J Nutr Educ Behav 2016 Mar;48(3):213-218. [doi: 10.1016/j.jneb.2015.12.009] [Medline: 26965099]
36. Ronteltap A, Bukman AJ, Nagelhout GE, et al. Digital health interventions to improve eating behaviour of people with a lower socioeconomic position: a scoping review of behaviour change techniques. BMC Nutr 2022 Dec 8;8(1):145. [doi: 10.1186/s40795-022-00635-3] [Medline: 36482430]
37. Tang H, Spreckley M, van Sluijs E, Ahern AL, Smith AD. The impact of social media interventions on eating behaviours and diet in adolescents and young adults: a mixed methods systematic review protocol. BMJ Open 2024 Apr;14(4):e083465. [doi: 10.1136/bmjopen-2023-083465]
38. Michie S, Jochelson K, Markham WA, Bridle C. Low-income groups and behaviour change interventions: a review of intervention content, effectiveness and theoretical frameworks. J Epidemiol Community Health 2009 Aug;63(8):610-622. [doi: 10.1136/jech.2008.078725] [Medline: 19386612]

XSL•FO
RenderX

39. Karimi N, Opie R, Crawford D, O'Connell S, Ball K. Digitally delivered interventions to improve nutrition behaviors among resource-poor and ethnic minority groups with type 2 diabetes: systematic review. J Med Internet Res 2023;26:e42595. [doi: 10.2196/42595] [Medline: 37490331]
40. Page MJ, Moher D, Bossuyt PM, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021 Mar 29;372:n160. [doi: 10.1136/bmj.n160] [Medline: 33781993]
41. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
42. Dykes J, Brunner EJ, Martikainen PT, Wardle J. Socioeconomic gradient in body size and obesity among women: the role of dietary restraint, disinhibition and hunger in the Whitehall II study. Int J Obes 2004 Feb;28(2):262-268. [doi: 10.1038/sj.ijo.0802523]
43. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Dec;6(1). [doi: 10.1186/1748-5908-6-42]
44. Cochrane handbook for systematic reviews of interventions (current version). Cochrane. URL: https://training.cochrane.org/handbook/current [accessed 2025-12-17]
45. Eldridge S, Campbell MK, Campbell MJ, et al. A revised Cochrane risk of bias tool for randomized trials (RoB 2). Additional considerations for cluster-randomized trials (RoB 2 CRT). 2021 Mar 18. URL: https://drive.google.com/file/d/1yDQtDkrp68_8kJiIUdbongK99sx7RFI-/view [accessed 2025-12-27]
46. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct 12;355:i4919. [doi: 10.1136/bmj.i4919] [Medline: 27733354]
47. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
48. Kilb M, Giese H, Mata J. How eating-related social media postings influence healthy eating in senders and network members: two field experiments with intensive longitudinal data. Appetite 2023 Mar 1;182:106430. [doi: 10.1016/j.appet.2022.106430] [Medline: 36549365]
49. McGuinness LA, Higgins JPT. Risk-of-bias VISualization (robvis): an R package and Shiny web app for visualizing risk-of-bias assessments. Res Synth Methods 2021 Jan;12(1):55-61. [doi: 10.1002/jrsm.1411] [Medline: 32336025]
50. Plaete J, Crombez G, Van der Mispel C, Verloigne M, Van Stappen V, De Bourdeaudhuij I. Effect of the web-based intervention MyPlan 1.0 on self-reported fruit and vegetable intake in adults who visit general practice: a quasi-experimental trial. J Med Internet Res 2016 Feb 29;18(2):e47. [doi: 10.2196/jmir.5252] [Medline: 26929095]
51. Plaete J, De Bourdeaudhuij I, Verloigne M, Crombez G. Acceptability, feasibility and effectiveness of an eHealth behaviour intervention using self-regulation: "MyPlan". Patient Educ Couns 2015 Jul 26;98(12):1617-1624. [doi: 10.1016/j.pec.2015.07.014] [Medline: 26277282]
52. Frie K, Stewart C, Piernas C, Cook B, Jebb SA. Effectiveness of an Online Programme to Tackle Individual's Meat Intake through SElf-regulation (OPTIMISE): a randomised controlled trial. Eur J Nutr 2022 Aug;61(5):2615-2626. [doi: 10.1007/s00394-022-02828-9] [Medline: 35244757]
53. Stewart C, Piernas C, Frie K, Cook B, Jebb SA. Evaluation of OPTIMISE (Online Programme to Tackle Individual's Meat Intake Through Self-regulation): cohort study. J Med Internet Res 2022 Dec 12;24(12):e37389. [doi: 10.2196/37389] [Medline: 36508245]
54. Papadaki A, Scott JA. The Mediterranean eating in Scotland experience project: evaluation of an internet-based intervention promoting the Mediterranean diet. Br J Nutr 2005 Aug;94(2):290-298. [doi: 10.1079/bjn20051476] [Medline: 16115365]
55. Papadaki A, Scott JA. Follow-up of a web-based tailored intervention promoting the Mediterranean diet in Scotland. Patient Educ Couns 2008 Nov;73(2):256-263. [doi: 10.1016/j.pec.2008.05.030] [Medline: 18640000]
56. Springvloet L, Lechner L, de Vries H, Candel MJJM, Oenema A. Short- and medium-term efficacy of a web-based computer-tailored nutrition education intervention for adults including cognitive and environmental feedback: randomized controlled trial. J Med Internet Res 2015 Jan 19;17(1):e23. [doi: 10.2196/jmir.3837] [Medline: 25599828]
57. Springvloet L, Lechner L, de Vries H, Oenema A. Long-term efficacy of a web-based computer-tailored nutrition education intervention for adults including cognitive and environmental feedback: a randomized controlled trial. BMC Public Health 2015 Apr 12;15(1):372. [doi: 10.1186/s12889-015-1707-4] [Medline: 25887891]
58. Weber J, Nigg CR. Promoting fruit and vegetable consumption during the COVID-19 pandemic—SportStudisMoveYou (SSMY): a randomized controlled trial. AIMS Public Health 2022;9(4):690-702. [doi: 10.3934/publichealth.2022048] [Medline: 36636149]
59. Power JM, Bersamin A. A text messaging intervention (Txt4HappyKids) to promote fruit and vegetable intake among families with young children: pilot study. JMIR Form Res 2018 Jul 6;2(2):e13. [doi: 10.2196/formative.8544] [Medline: 30684412]
60. Buller DB, Woodall WG, Zimmerman DE, et al. Randomized trial on the 5 a day, the Rio Grande Way Website, a web-based program to improve fruit and vegetable consumption in rural communities. J Health Commun 2008;13(3):230-249. [doi: 10.1080/10810730801985285] [Medline: 18569356]

XSL•FO
RenderX

61. Chamberland K, Sanchez M, Panahi S, Provencher V, Gagnon J, Drapeau V. The impact of an innovative web-based school nutrition intervention to increase fruits and vegetables and milk and alternatives in adolescents: a clustered randomized trial. Int J Behav Nutr Phys Act 2017 Oct 16;14(1):140. [doi: 10.1186/s12966-017-0595-7] [Medline: 29037203]
62. Nakamura S, Inayama T, Harada K, Arao T. Reduction in vegetable intake disparities with a web-based nutrition education intervention among lower-income adults in Japan: randomized controlled trial. J Med Internet Res 2017 Nov 24;19(11):e377. [doi: 10.2196/jmir.8031] [Medline: 29175810]
63. Lim TJ, Okine RN, Kershaw JC. Health- or environment-focused text messages as a potential strategy to increase plant-based eating among young adults: an exploratory study. Foods 2021 Dec 19;10(12):3147. [doi: 10.3390/foods10123147] [Medline: 34945698]
64. Carfora V, Catellani P. Legumes or meat? The effectiveness of recommendation messages towards a plant-based diet depends on people's identification with flexitarians. Nutrients 2023;15(1):15. [doi: 10.3390/nu15010015]
65. Carfora V, Morandi M, Catellani P. The effect of message framing in promoting the Mediterranean diet: the moderating role of eating self-efficacy. Foods 2022 May 17;11(10):1454. [doi: 10.3390/foods11101454] [Medline: 35627024]
66. Gosliner W, Felix C, Strochlic R, et al. Feasibility and response to the San Diego County, California, Supplemental Nutrition Assistance Program (SNAP) agency sending food and nutrition text messages to all participants: quasi-experimental web-based survey pilot study. J Med Internet Res 2023 Apr 19;25:e41021. [doi: 10.2196/41021] [Medline: 37074786]
67. Carfora V, Bertolotti M, Catellani P. Informational and emotional daily messages to reduce red and processed meat consumption. Appetite 2019 Oct 1;141:104331. [doi: 10.1016/j.appet.2019.104331] [Medline: 31276710]
68. Carfora V, Caso D, Conner M. Correlational study and randomised controlled trial for understanding and changing red meat consumption: the role of eating identities. Soc Sci Med 2017 Feb;175:244-252. [doi: 10.1016/j.socscimed.2017.01.005]
69. Carfora V, Caso D, Conner M. Randomised controlled trial of a text messaging intervention for reducing processed meat consumption: the mediating roles of anticipated regret and intention. Appetite 2017 Oct 1;117:152-160. [doi: 10.1016/j.appet.2017.06.025] [Medline: 28651971]
70. Carfora V, Catellani P, Caso D, Conner M. How to reduce red and processed meat consumption by daily text messages targeting environment or health benefits. J Environ Psychol 2019 Oct;65:101319. [doi: 10.1016/j.jenvp.2019.101319]
71. Carfora V, Zeiske N, van der Werff E, Steg L, Catellani P. Adding dynamic norm to environmental information in messages promoting the reduction of meat consumption. Environ Commun 2022 Oct 3;16(7):900-919. [doi: 10.1080/17524032.2022.2062019]
72. Wolstenholme E, Poortinga W, Whitmarsh L. Two birds, one stone: the effectiveness of health and environmental messages to reduce meat consumption and encourage pro-environmental behavioral spillover. Front Psychol 2020;11:577111. [doi: 10.3389/fpsyg.2020.577111] [Medline: 33117243]
73. Brookie KL, Mainvil LA, Carr AC, Vissers MCM, Conner TS. The development and effectiveness of an ecological momentary intervention to increase daily fruit and vegetable consumption in low-consuming young adults. Appetite 2017 Jan 1;108:32-41. [doi: 10.1016/j.appet.2016.09.015] [Medline: 27642037]
74. Rompotis CJ, Grove JR, Byrne SM. Benefits of habit‐based informational interventions: a randomised controlled trial of fruit and vegetable consumption. Aust N Z J Public Health 2014 Jun;38(3):247-252. [doi: 10.1111/1753-6405.12232]
75. Gustafson A, Jilcott Pitts SB, McQuerry K, Babtunde O, Mullins J. A mentor-led text-messaging intervention increases intake of fruits and vegetables and goal setting for healthier dietary consumption among rural adolescents in Kentucky and North Carolina, 2017. Nutrients 2019 Mar 11;11(3):593. [doi: 10.3390/nu11030593] [Medline: 30862118]
76. Pedersen S, Grønhøj A, Thøgersen J. Texting your way to healthier eating? Effects of participating in a feedback intervention using text messaging on adolescents' fruit and vegetable intake. Health Educ Res 2016 Apr;31(2):171-184. [doi: 10.1093/her/cyv104] [Medline: 26850061]
77. Alexander GL, McClure JB, Calvi JH, et al. A randomized clinical trial evaluating online interventions to improve fruit and vegetable consumption. Am J Public Health 2010 Feb;100(2):319-326. [doi: 10.2105/AJPH.2008.154468] [Medline: 20019315]
78. Dumas AA, Lemieux S, Lapointe A, Provencher V, Robitaille J, Desroches S. Effects of an evidence-informed healthy eating blog on dietary intakes and food-related behaviors of mothers of preschool- and school-aged children: a randomized controlled trial. J Acad Nutr Diet 2020 Jan;120(1):53-68. [doi: 10.1016/j.jand.2019.05.016] [Medline: 31519466]
79. Tapper K, Jiga-Boy G, Maio GR, Haddock G, Lewis M. Development and preliminary evaluation of an internet-based healthy eating program: randomized controlled trial. J Med Internet Res 2014 Oct 10;16(10):e231. [doi: 10.2196/jmir.3534] [Medline: 25305376]
80. Røed M, Medin AC, Vik FN, et al. Effect of a parent-focused eHealth intervention on children's fruit, vegetable, and discretionary food intake (Food4toddlers): randomized controlled trial. J Med Internet Res 2021 Feb 16;23(2):e18311. [doi: 10.2196/18311] [Medline: 33591279]
81. Eckert KF, Agostinelli J, Laila A, et al. Feasibility, acceptability, and preliminary impact of "Supper Heroes", a family-based sustainable diet intervention. Appetite 2025 Feb 1;206:107849. [doi: 10.1016/j.appet.2025.107849] [Medline: 39788349]
82. Livingstone KM, Rawstorn JC, Partridge SR, et al. Feasibility of a co-designed and personalised intervention to improve vegetable intake in rural-dwelling young adults. Int J Behav Nutr Phys Act 2025;22(1):97. [doi: 10.1186/s12966-025-01796-7]

XSL•FO
RenderX

83. Ricci M, Devecchi A, Migliavada R, Piochi M, Torri L. Effect of demographic characteristics and personality traits on eating patterns in the context of dietary intervention: the EATMED Case Study. Int J Environ Res Public Health 2025 Jul 10;22(7):1095. [doi: 10.3390/ijerph22071095] [Medline: 40724162]
84. Meng J, Peng W, Shin SY, Chung M. Online self-tracking groups to increase fruit and vegetable intake: a small-scale study on mechanisms of group effect on behavior change. J Med Internet Res 2017 Mar 6;19(3):e63. [doi: 10.2196/jmir.6537] [Medline: 28264793]
85. Hawkins L, Farrow C, Clayton M, Thomas JM. Can social media be used to increase fruit and vegetable consumption? A pilot intervention study. Digit Health 2024 Jan;10. [doi: 10.1177/20552076241241262]
86. Inauen J, Bolger N, Shrout PE, et al. Using smartphone‐based support groups to promote healthy eating in daily life: a randomised trial. Applied Psych Health Well 2017 Nov;9(3):303-323. [doi: 10.1111/aphw.12093]
87. Ng AH, ElGhattis Y, Biesiekierski JR, Moschonis G. Assessing the effectiveness of a 4‐week online intervention on food literacy and fruit and vegetable consumption in Australian adults: the online MedDiet challenge. Health Social Care Comm 2022 Nov;30. [doi: 10.1111/hsc.13909]
88. Carreño Enciso L, de Mateo Silleras B, de la Cruz Marcos S, Redondo Del Río P. Social media for nutrition education—a randomized controlled trial to promote fruit and vegetable intake in a university setting: "The University of Valladolid Community Eats Healthy" Study. Nutrients 2024 Apr 26;16(9):1308. [doi: 10.3390/nu16091308] [Medline: 38732555]
89. Hendrie GA, Hussain MS, Brindal E, James-Martin G, Williams G, Crook A. Impact of a mobile phone app to increase vegetable consumption and variety in adults: large-scale community cohort study. JMIR Mhealth Uhealth 2019;8(4):e14726. [doi: 10.2196/14726] [Medline: 31486407]
90. Vázquez-Paz AM, Michel-Nava RM, Delgado-Pérez EE, Lares-Michel M, Espinosa-Curiel IE. Parents' mHealth app for promoting healthy eating behaviors in children: feasibility, acceptability, and pilot study. J Med Syst 2022 Sep 16;46(11):70. [doi: 10.1007/s10916-022-01860-w] [Medline: 36109423]
91. Liu H, Feng J, Shi Z, et al. Effects of a novel applet-based personalized dietary intervention on dietary intakes: a randomized controlled trial in a real-world scenario. Nutrients 2024;16(4):565. [doi: 10.3390/nu16040565]
92. Elbert SP, Dijkstra A, Oenema A. A mobile phone app intervention targeting fruit and vegetable consumption: the efficacy of textual and auditory tailored health information tested in a randomized controlled trial. J Med Internet Res 2016 Jun 10;18(6):e147. [doi: 10.2196/jmir.5056] [Medline: 27287823]
93. Ragelienė T, Aschemann-Witzel J, Grønhøj A. Efficacy of a smartphone application-based intervention for encouraging children's healthy eating in Denmark. Health Promot Int 2022 Feb 17;37(1). [doi: 10.1093/heapro/daab081]
94. Shatwan IM, Alhefani RS, Bukhari MF, et al. Effects of a smartphone app on fruit and vegetable consumption among Saudi adolescents: randomized controlled trial. JMIR Pediatr Parent 2023 Feb 9;6:e43160. [doi: 10.2196/43160] [Medline: 36757770]
95. Espinosa-Curiel IE, Pozas-Bogarin EE, Lozano-Salas JL, Martínez-Miranda J, Delgado-Pérez EE, Estrada-Zamarron LS. Nutritional education and promotion of healthy eating behaviors among Mexican children through video games: design and pilot test of FoodRateMaster. JMIR Serious Games 2020 Apr 13;8(2):e16431. [doi: 10.2196/16431] [Medline: 32281539]
96. Thompson D, Bhatt R, Vazquez I, et al. Creating action plans in a serious video game increases and maintains child fruit-vegetable intake: a randomized controlled trial. Int J Behav Nutr Phys Act 2015 Dec;12(1). [doi: 10.1186/s12966-015-0199-z]
97. Buller MK, Kane IL, Dunn AL, Edwards EJ, Buller DB, Liu X. Marketing fruit and vegetable intake with interactive games on the internet. Soc Mar Q 2009 Mar;15(1_suppl):136-154. [doi: 10.1080/15245000903038316]
98. Block G, Block T, Wakimoto P, Block CH. Demonstration of an e-mailed worksite nutrition intervention program. Prev Chronic Dis 2004 Oct;1(4):A06. [Medline: 15670437]
99. Kothe EJ, Mullan BA. A randomised controlled trial of a theory of planned behaviour to increase fruit and vegetable consumption. Fresh facts. Appetite 2014 Jul;78:68-75. [doi: 10.1016/j.appet.2014.03.006] [Medline: 24656949]
100. Cumming G. Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis: Routledge; 2013. [doi: 10.4324/9780203807002]
101. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 2013 Nov 26;4:863. [doi: 10.3389/fpsyg.2013.00863] [Medline: 24324449]
102. Lenhard W. Computation of different effect sizes like d, f, r and transformation of different effect sizes. Psychometrica. URL: https://www.psychometrica.de/effect_size.html [accessed 2025-12-17]
103. Morris SB. Estimating effect sizes from pretest-posttest-control group designs. Organ Res Methods 2008 Apr;11(2):364-386. [doi: 10.1177/1094428106291059]
104. Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. Res Synth Methods 2010 Jan;1(1):39-65. [doi: 10.1002/jrsm.5] [Medline: 26056092]
105. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods 2010 Apr;1(2):97-111. [doi: 10.1002/jrsm.12] [Medline: 26061376]
106. Harrer M, Cuijpers P, Furukawa TA, Ebert DD. Doing Meta-Analysis with R: A Hands-on Guide: Chapman & Hall; 2021. URL: https://www.routledge.com/Doing-Meta-Analysis-with-R-A-Hands-On-Guide/Harrer-Cuijpers-Furukawa-Ebert/p/book/9780367610074 [accessed 2025-12-27]

XSL•FO
RenderX

107. Pustejovsky JE, Tipton E. Meta-analysis with robust variance estimation: expanding the range of working models. Prev Sci 2022 Apr;23(3):425-438. [doi: 10.1007/s11121-021-01246-3] [Medline: 33961175]
108. Sarri G, Patorno E, Yuan H, et al. Framework for the synthesis of non-randomised studies and randomised controlled trials: a guidance on conducting a systematic review and meta-analysis for healthcare decision making. BMJ Evid Based Med 2022 Apr;27(2):109-119. [doi: 10.1136/bmjebm-2020-111493] [Medline: 33298465]
109. Yao M, Mei F, Ma Y, et al. Including non-randomized studies of interventions in meta-analyses of randomized controlled trials changed the estimates in more than a third of the studies: evidence from an empirical analysis. J Clin Epidemiol 2025 Jul;183:111815. [doi: 10.1016/j.jclinepi.2025.111815] [Medline: 40334718]
110. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14(1):25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
111. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
112. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta‐analysis: I2 is not an absolute measure of heterogeneity. Res Synth Methods 2017;8(1):5-18. [doi: 10.1002/jrsm.1230] [Medline: 28058794]
113. Lau J, Ioannidis JPA, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. BMJ 2006 Sep 16;333(7568):597-600. [doi: 10.1136/bmj.333.7568.597]
114. Hirst RJ, Cragg L, Allen HA. Vision dominates audition in adults but not children: a meta-analysis of the Colavita effect. Neurosci Biobehav Rev 2018 Nov;94:286-301. [doi: 10.1016/j.neubiorev.2018.07.012] [Medline: 30048672]
115. Forbes C, Greenwood H, Carter M, Clark J. Automation of duplicate record detection for systematic reviews: deduplicator. Syst Rev 2024 Aug 2;13(1):206. [doi: 10.1186/s13643-024-02619-9] [Medline: 39095913]
116. Carfora V, Caso D, Conner M. Randomized controlled trial of a messaging intervention to increase fruit and vegetable intake in adolescents: affective versus instrumental messages. Br J Health Psychol 2016 Nov;21(4):937-955. [doi: 10.1111/bjhp.12208] [Medline: 27374753]
117. Zlatevska N, Barton B, Dubelaar C, Hohberger J. Navigating through nutrition labeling effects: a second-order meta-analysis. J Public Policy Mark 2024 Jan;43(1):76-94. [doi: 10.1177/07439156231158115]
118. Qutteina Y, Hallez L, Raedschelders M, De Backer C, Smits T. Food for teens: how social media is associated with adolescent eating outcomes. Public Health Nutr 2022 Feb;25(2):290-302. [doi: 10.1017/S1368980021003116] [Medline: 34325764]
119. Ventura V, Cavaliere A, Iannò B. #Socialfood: virtuous or vicious? A systematic review. Trends Food Sci Technol 2021 Apr;110:674-686. [doi: 10.1016/j.tifs.2021.02.018]
120. Cuykx I, Decorte P, Teunissen L, et al. The magic is in the mix: a uses and gratifications approach to the cross-media use of food-related media content. Food Cult Soc 2024 Aug 7;27(4):1146-1170. [doi: 10.1080/15528014.2023.2263705]
121. Decorte P, Teunissen L, Cuykx I, et al. Media and personal socialization agents toward emerging adults' recipe choices: a cluster approach. J Foodserv Bus Res 2025:1-31. [doi: 10.1080/15378020.2024.2448642]
122. Laranjo L, Arguel A, Neves AL, et al. The influence of social networking sites on health behavior change: a systematic review and meta-analysis. J Am Med Inform Assoc 2015 Jan;22(1):243-256. [doi: 10.1136/amiajnl-2014-002841] [Medline: 25005606]
123. Laranjo L. Social media and health behavior change. In: Syed-Abdul S, Gabarron E, Lau AYS, editors. Participatory Health Through Social Media: Elsevier eBooks; 2016:83-111. [doi: 10.1016/B978-0-12-809269-9.00006-2]
124. Vogels EA, Gelles-Watnick R, Massarat N. Teens, social media and technology 2022. Pew Research Center. 2022. URL: https://www.pewresearch.org/internet/2022/08/10/teens-social-media-and-technology-2022/ [accessed 2025-12-27]
125. Qutteina Y, Hallez L, Mennes N, De Backer C, Smits T. What do adolescents see on social media? A diary study of food marketing images on social media. Front Psychol 2019;10:2637. [doi: 10.3389/fpsyg.2019.02637] [Medline: 31824391]
126. Adams J, Mytton O, White M, Monsivais P. Why are some population interventions for diet and obesity more equitable and effective than others? The role of individual agency. PLoS Med 2016 Apr;13(4):e1001990. [doi: 10.1371/journal.pmed.1001990] [Medline: 27046234]
127. Garrott K, Ogilvie D, Panter J, et al. Explaining differential socioeconomic effects in population health interventions: development and application of a new tool to classify intervention agentic demand. Lancet 2023 Nov;402:S3. [doi: 10.1016/S0140-6736(23)02056-1]
128. Heron KE, Smyth JM. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. Br J Health Psychol 2010 Feb;15(Pt 1):1-39. [doi: 10.1348/135910709X466063] [Medline: 19646331]
129. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2020;23(1):e17187. [doi: 10.2196/17187]
130. Nagler RH. Adverse outcomes associated with media exposure to contradictory nutrition messages. J Health Commun 2014;19(1):24-40. [doi: 10.1080/10810730.2013.798384] [Medline: 24117281]
131. Mair JL, Salamanca-Sanabria A, Augsburger M, et al. Effective behavior change techniques in digital health interventions for the prevention or management of noncommunicable diseases: an umbrella review. Ann Behav Med 2023 Sep 13;57(10):817-835. [doi: 10.1093/abm/kaad041] [Medline: 37625030]

XSL•FO
RenderX

132. Ortiz R, Massar RE, McMacken M, Albert SL. Stronger together than apart: the role of social support in adopting a healthy plant-based eating pattern. Appetite 2024 Jul 1;198:107341. [doi: 10.1016/j.appet.2024.107341] [Medline: 38599245]
133. Seo DC, Niu J. Evaluation of internet-based interventions on waist circumference reduction: a meta-analysis. J Med Internet Res 2015 Jul 21;17(7):e181. [doi: 10.2196/jmir.3921] [Medline: 26199208]
134. Hawkins LK, Farrow C, Thomas JM. Do perceived norms of social media users' eating habits and preferences predict our own food consumption and BMI? Appetite 2020 Jun 1;149:104611. [doi: 10.1016/j.appet.2020.104611] [Medline: 31958481]
135. Robinson E, Fleming A, Higgs S. Prompting healthier eating: testing the use of health and social norm based messages. Health Psychol 2014 Sep;33(9):1057-1064. [doi: 10.1037/a0034213] [Medline: 24295025]
136. Stok FM, de Ridder DTD, de Vet E, de Wit JBF. Don't tell me what I should do, but what others do: the influence of descriptive and injunctive peer norms on fruit consumption in adolescents. Br J Health Psychol 2014 Feb;19(1):52-64. [doi: 10.1111/bjhp.12030] [Medline: 23406475]
137. Darmon N, Drewnowski A. Does social class predict diet quality? Am J Clin Nutr 2008 May;87(5):1107-1117. [doi: 10.1093/ajcn/87.5.1107]
138. De Irala-Estévez J, Groth M, Johansson L, Oltersdorf U, Prättälä R, Martínez-González M. A systematic review of socio-economic differences in food habits in Europe: consumption of fruit and vegetables. Eur J Clin Nutr 2000 Sep 1;54(9):706-714. [doi: 10.1038/sj.ejcn.1601080]
139. Malo JS, Schafer MH, Stull AJ. Healthy eating in life course context: asymmetric implications of socioeconomic origins and destinations. Soc Sci Med 2025 May;372:117936. [doi: 10.1016/j.socscimed.2025.117936]
140. Chan EY, Zlatevska N. Jerkies, tacos, and burgers: subjective socioeconomic status and meat preference. Appetite 2019 Jan 1;132:257-266. [doi: 10.1016/j.appet.2018.08.027] [Medline: 30172366]
141. Sawyer ADM, van Lenthe F, Kamphuis CBM, et al. Dynamics of the complex food environment underlying dietary intake in low-income groups: a systems map of associations extracted from a systematic umbrella literature review. Int J Behav Nutr Phys Act 2021 Dec;18(1). [doi: 10.1186/s12966-021-01164-1]
142. van der Heijden A, Te Molder H, Jager G, Mulder BC. Healthy eating beliefs and the meaning of food in populations with a low socioeconomic position: a scoping review. Appetite 2021 Jun 1;161:105135. [doi: 10.1016/j.appet.2021.105135] [Medline: 33493606]
143. Adler NE, Epel ES, Castellazzo G, Ickovics JR. Relationship of subjective and objective social status with psychological and physiological functioning: preliminary data in healthy, White women. Health Psychol 2000;19(6):586-592. [doi: 10.1037//0278-6133.19.6.586]
144. Cheon BK, Hong YY. Mere experience of low subjective socioeconomic status stimulates appetite and food intake. Proc Natl Acad Sci USA 2017 Jan 3;114(1):72-77. [doi: 10.1073/pnas.1607330114]
145. D'Hooge L, Achterberg P, Reeskens T. Mind over matter. The impact of subjective social status on health outcomes and health behaviors. PLoS ONE 2018;13(9):e0202489. [doi: 10.1371/journal.pone.0202489] [Medline: 30183731]
146. van den Bekerom L, van Gestel LC, Schoones JW, Bussemaker J, Adriaanse MA. Health behavior interventions among people with lower socio-economic position: a scoping review of behavior change techniques and effectiveness. Health Psychol Behav Med 2024;12(1):2365931. [doi: 10.1080/21642850.2024.2365931] [Medline: 38903803]
147. Taufik D, Verain MCD, Bouwman EP, Reinders MJ. Determinants of real-life behavioural interventions to stimulate more plant-based and less animal-based diets: a systematic review. Trends Food Sci Technol 2019 Nov;93:281-303. [doi: 10.1016/j.tifs.2019.09.019]
148. Eppes EV, Augustyn M, Gross SM, Vernon P, Caulfield LE, Paige DM. Engagement with and acceptability of digital media platforms for use in improving health behaviors among vulnerable families: systematic review. J Med Internet Res 2022;25:e40934. [doi: 10.2196/40934]
149. Richardson M, Garner P, Donegan S. Interpretation of subgroup analyses in systematic reviews: a tutorial. Clin Epidemiol Glob Health 2019 Jun;7(2):192-198. [doi: 10.1016/j.cegh.2018.05.005]
150. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926. [doi: 10.1136/bmj.39489.470347.AD]
151. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev 2014;2014(4):MR000034. [doi: 10.1002/14651858.MR000034.pub2]
152. Golder S, Loke YK, Bland M. Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: methodological overview. PLoS Med 2011 May;8(5):e1001026. [doi: 10.1371/journal.pmed.1001026] [Medline: 21559325]
153. Cuello-Garcia CA, Santesso N, Morgan RL, et al. GRADE guidance 24 optimizing the integration of randomized and non-randomized studies of interventions in evidence syntheses and health guidelines. J Clin Epidemiol 2022 Feb;142:200-208. [doi: 10.1016/j.jclinepi.2021.11.026]
154. Zhou Y, Yao M, Mei F, et al. Integrating randomized controlled trials and non-randomized studies of interventions to assess the effect of rare events: a Bayesian re-analysis of two meta-analyses. BMC Med Res Methodol 2024 Sep 27;24(1):219. [doi: 10.1186/s12874-024-02347-7] [Medline: 39333867]

XSL•FO
RenderX

155. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348(mar07 3):g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]

Abbreviations

BCT: behavior change technique
BCTTv1: Behavior Change Technique Taxonomy version 1
FVI: fruit and vegetable intake
IG: intervention group
NRS: nonrandomized study
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
RCT: randomized controlled trial
SES: socioeconomic status

Edited by S Brini; submitted 17.Jul.2025; peer-reviewed by B Lam, ME Heidari, N Maye, O Oyetunji, S Narayan; revised version received 01.Dec.2025; accepted 02.Dec.2025; published 08.Jan.2026.

Please cite as:
Vanwinkelen K, Spruyt B, Smits T
Digital Interventions Targeting Healthy and Sustainable Eating Behavior: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e80821
URL: https://www.jmir.org/2026/1/e80821
doi:10.2196/80821

© Käbi Vanwinkelen, Bram Spruyt, Tim Smits. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Diet-Related Health Recommender Systems for Patients With Chronic Health Conditions: Scoping Review

Xiaolan Dong[1*], MD; Bei Yun[2*], MD; Anni Pakarinen[1,3], PhD; Zhuting Zheng[2], MSc; Hao Niu[2], PhD; Tian Jin[2], MD; Changrong Yuan[2], PhD; Jingting Wang[4], PhD

[1]Department of Nursing Science, University of Turku, Turku, Finland

[2]School of Nursing, Fudan University, Shanghai, China

[3]Faculty of Nursing and Physiotherapy, University of Lleida, Lleida, Spain

[4]School of Nursing, Naval Medical University, 800 Xiangyin Road, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Jingting Wang, PhD
School of Nursing, Naval Medical University, 800 Xiangyin Road, Shanghai, China

Abstract

Background: Diet-related Health Recommender Systems (HRSs) have gained attention for their potential to provide personalized dietary guidance, particularly for patients with chronic conditions. However, studies on diet-related HRSs in health care are relatively limited.

Objective: This scoping review aims to present the state of current research on diet-related HRSs for patients with chronic health conditions, identify existing gaps, and suggest future research directions.

Methods: The scoping review was conducted following the Arksey and O'Malley framework and was reported in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines. The literature search was conducted in October 2024 across 6 English databases (PubMed, Medline, Embase, Web of Science Core Collection, IEEE Xplore, and CINAHL) and 4 Chinese databases (SinoMed, CNKI, Wanfang, and VIP). Studies focusing on diet-related HRSs for patients with chronic conditions were included.

Results: Fifteen studies published between 2010 and 2024 from 9 countries were included. Diet-related HRSs mainly target adults with chronic diseases, with 9 systems (60%) including users with diabetes and 6 (40%) including users with hypertension. Nine studies (60%) described functional structures, which were categorized into 4 components: user information, food or diet recommendations, knowledge and decision support, and data management with additional functions. Recommended content was categorized into 5 types: food (n=6, 40%), recipes (n=4, 26.67%), diet plans or meal plans (n=3, 20%), recipes and food (n=1, 6.67%), and meals (n=1, 6.67%). Recommendation methods included constraint-based (n=6, 40%), focusing on patients' dietary restrictions; preference-based (n=5, 33.33%), considering patients' food preferences; and hybrid (n=4, 26.67%), combining both approaches. Of all recommendation technologies, most studies (n=13, 86.67%) applied hybrid approaches, enabling more robust personalization. For the data used for training, 13 studies (86.67%) explicitly mentioned the data sources, and 10 studies' (66.67%) data came from professional organizations and websites. The recommendation process followed a structured workflow. Twelve studies (80%) evaluated diet-related HRSs using either online or offline methods, while accuracy (n=9, 60%) has been the most common evaluation criterion. However, no studies went deeper into how these systems affected users' dietary behaviors over time.

Conclusions: Diet-related HRSs have the potential to deliver personalized dietary support for patients with chronic diseases, but current systems show key gaps. Future development must adopt user-centered design, provide practical and actionable dietary guidance, and use hybrid recommendation techniques to increase precision and clinical relevance. Standardized evaluation methods and real-world, long-term studies are essential to evaluate the impact of diet-related HRSs on dietary behavior and health outcomes. Addressing these needs will enable diet-related HRSs to become reliable tools for chronic disease management and patient-centered care.

Trial Registration: International Platform of Registered Systematic Review and Meta-analysis Protocols INPLASY202550081; https://inplasy.com/wp-content/uploads/2025/05/INPLASY-Protocol-7837.pdf

(J Med Internet Res 2026;28:e77726) doi:10.2196/77726

XSL•FO
RenderX

KEYWORDS

Health Recommender System; Diet-related Health Recommender System; chronic health conditions; diet; scoping review; PRISMA; HRS

Introduction

Dietary management is crucial to overall health, especially in the context of global trends where poor dietary habits have become a significant factor contributing to weight-related issues [1]. Statistics from the latest Global Nutrition Report indicated alarming rates of overweight and obesity among adults, with 40.8% of adult (18 years or older) women and 40.4% of adult men affected. Conversely, 9.1% of adult women and 8.1 % of adult men were underweight [2]. Improper dietary management not only affects body weight but also contributes to the development of diabetes [3,4], hypertension [5], cardiovascular diseases [6], chronic kidney disease [7], and inflammatory bowel disease [8], among others.

The Scientific Research Report on Dietary Guidelines for Chinese Residents (2021) highlighted that the overweight and obesity rates among children younger than 6 years old and those aged 6 - 17 years were 10.4% and 19.0%, respectively [9]. Among residents aged 18 years and older, the overweight rate was 34.3% and the obesity rate was 16.4%, with 50.7% of adults being overweight or obese [9]. Overweight and obesity are significant risk factors for cardiovascular diseases, diabetes, cancer, and other chronic health conditions [9].

Poor dietary habits contribute to both the onset and progression of these chronic conditions. Therefore, scientific and effective diet management is essential for ensuring proper nutrient intake, which directly enhances individual immunity, halts disease progression, impacts health status, and supports recovery from chronic health conditions, thereby improving overall health outcomes [7,10,11]. Moreover, good diet management regulates sleep and mood, reduces fatigue, and comprehensively enhances overall health [12,13]. In February 2024, China's National Health Commission released the "Dietary Guidelines for Adults with Hyperuricemia and Gout (2024 Edition)," "Dietary Guidelines for Adult Obesity (2024 Edition)," "Dietary Guidelines for Childhood and Adolescent Obesity (2024 Edition)," and "Dietary Guidelines for Adults with Chronic Kidney Disease (2024 Edition)" [14]. These guidelines aim to prevent and control the occurrence and progression of chronic diseases among the Chinese population through dietary management.

Patients with chronic health conditions often need to adjust their diets based on their specific health status to manage health conditions effectively and enhance the overall quality of life. When patients are required to follow dietary restrictions due to their chronic health condition, it is important for the patient to distinguish whether certain foods are permissible, ensure that the cooking methods meet the essential requirements, evaluate whether portion sizes are appropriate, and strive to maintain a balanced diet to reduce the risk of malnutrition and other issues caused by dietary limitations [15]. Therefore, these patients require targeted and personalized guidance to help them implement scientific and feasible dietary management practices.

Recommender Systems (RSs) are software tools that provide suggestions or recommendations of items to users, such as products, services, information, or content, based on their preferences, interests, and past behaviors [16], which have been used in several domains such as e-commerce, e-learning, e-tourism, or eHealth. To generate recommendations, RSs commonly rely on a set of foundational recommendation technologies. Collaborative filtering (CF) is one of the most widely applied recommendation technologies, which provides recommendations for users by using the known preferences of other users with similar behaviors [17]. In contrast, content-based (CB) methods recommend items similar to those a given user has previously liked [18]. Unlike CF and CB, which rely on large historical rating datasets, knowledge-based (KB) approaches produce suggestions by leveraging domain knowledge, expert rules, and explicit user constraints [19]. Knowledge graph (KG)-driven recommendation further enhances personalization by representing domain knowledge in graph structures to generate precise suggestions [20]. Finally, hybrid recommendation (HyR) integrates 2 or more strategies to maximize their strengths and mitigate individual limitations [21].

Health Recommender Systems (HRSs) are also one of the RS's important application scenarios. HRSs offer the potential to motivate and engage users to change their behavior by sharing better choices and practical knowledge based on observed user behavior [22]. HRSs have been applied in health care in recent years, in areas such as mental health [23], hearing aid usage [24], health education [25], physical activity [26], and diet-related health [27].

Diet-related HRSs are software tools that use personalized data to provide tailored food recommendations from a wide range of options [28]. Diet-related HRSs present promising solutions to the issues of information overload and limited food choices, which contribute significantly to diet-related health problems [29]. By using personalized data, diet-related HRSs offer tailored food recommendations that take into account users' taste preferences, dietary needs, and medical conditions [30]. Such systems can filter and sort food options [27], fostering a better understanding of dietary health and increasing user engagement [31]. Additionally, diet-related HRSs can enhance dietary outcomes by providing nutrition assessments and offering healthier meal plans and recipes [28].

While diet-related HRSs have been explored within the field of information and communication technology, research on their application in chronic disease dietary management is still limited, and there is a lack of comprehensive literature reviews in this area. To address these gaps, this study adopted a scoping review methodology to review the research status quo of diet-related HRSs for individuals with chronic conditions. In this review, diet-related HRSs are defined as systems specifically designed for patients with chronic health conditions that require disease-related dietary restrictions. The target populations,

XSL•FO
RenderX

function structures, recommendation content, implementation of recommendation features, and evaluation of the diet-related HRSs were analyzed to provide a reference for health care researchers seeking to design more effective and user-friendly systems.

Methods

Overview

Scoping reviews assess the extent of the research, range, and nature, identify gaps, determine systematic review value, and disseminate research findings [32]. A scoping review methodology is used to systematically map available research on the broad, complex, and emerging research question [32]. In emerging research fields, a lack of randomized controlled trials may impede formal systematic reviews or meta-analyses. Scoping studies can address diverse questions and incorporate a range of study designs, making them ideal to complement clinical trial findings. The scoping review was conducted following the Arksey and O'Malley framework [32] and was reported in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines [33] and was structured according to the following outlined steps.

Stage 1: Identification of the Research Question

The research question was identified after an initial review of the literature and a discussion within our research team. The key research questions that guided the review were as follows:

1. What is the current state of research on diet-related HRSs for patients with chronic health conditions?
2. What are the target users for diet-related HRSs for patients with chronic health conditions?
3. What are the function structures in diet-related HRSs for patients with chronic health conditions?
4. What types of recommendation content are provided by diet-related HRSs for patients with chronic health conditions?
5. How are recommendation features implemented in diet-related HRSs for patients with chronic health conditions?
6. How are diet-related HRSs for patients with chronic health conditions evaluated?

Stage 2: Identifying Relevant Studies

Before searching, the research team worked collaboratively in making decisions about inclusion and exclusion criteria and in planning the initial search strategies to comprehensively identify the relevant literature. The final search strategies were developed with the assistance of a health science librarian (Multimedia Appendix 1). The literature search was performed by 3 researchers in October 2024, covering 6 English databases (PubMed, Medline, Embase, Web of Science Core Collection, IEEE Xplore, and CINAHL) and 4 Chinese databases (SinoMed, CNKI, Wanfang, and VIPC). Relevant studies were searched from January 2010 to October 2024 to answer the above research questions (1-6). An example of the search strategy performed in Web of Science Core Collection is presented here: (recommender system* OR hybrid recommendation* OR collaborative filtering OR content based recommendation* OR recommendation* system* OR knowledge based recommendation*) AND (recipe* OR diet* OR food OR eat* OR nutrition*). As the search proceeded, additional terms were suggested by experts to potentially modify the question, but new research did not result in additional data, and the question did not change. Duplicate references were filtered out using EndNote. In addition to database search, hand searching was conducted by screening the reference lists of all included articles and relevant review papers to identify additional eligible studies. We also manually checked key journals in the field and conference proceedings to ensure comprehensive coverage.

Stage 3: Study Selection

All studies searched were independently assessed by 2 authors (XD and BY) based on the inclusion criteria listed below, and discrepancies were verified by a third author (JW). Studies were eligible for inclusion in this review if all the following criteria were met: (1) the recommended information was related to at least one of the following: food, meal plan, diet plan, and recipe; (2) the study applied personalized recommendation strategy; (3) the recommendations were generated using algorithmic and technological methods; (4) the study population comprised patients with chronic health conditions; and (5) the study was published in a peer-reviewed journal or conference proceeding. The exclusion criteria included: (1) the recommendation unrelated to human health (eg, study focusing on animal health [34]), (2) studies reporting the same RSs were considered duplicates; only the most recent or most comprehensive publication was retained (eg, the latest published study [35] was included, while the earlier one [36] was excluded), and (3) full-text articles not in English or Chinese language.

Stage 4: Charting the Data

Decisions regarding the information to be recorded from the primary studies were made through discussions within the research group. Subsequently, a structured chart was developed to collate, summarize, and share the extracted data. A descriptive-analytical narrative approach was used to extract and chart the data from the selected articles [32,37]. Three researchers (XD, BY, and ZZ) independently extracted the data and performed coding, with the other two authors (HN and JW) verifying accuracy. All discrepancies were resolved by consensus. The following details were documented for each included study to answer the research questions: (1) nationality and publication year (Multimedia Appendix 2); (2) target users; (3) function structures; (4) recommendation content; (5) recommendation method, recommendation technology, data of training set, and recommendation process; and (6) evaluation method, evaluation criteria, test set, or evaluation sample size.

Stage 5: Collating, Summarizing, and Reporting the Results

The scoping review methodology aimed to summarize the breadth and depth of the existing literature. At this stage, an overview of the characteristics of all included articles was collated, summarized, and reported. Initially, a basic numerical summary of the studies, including the extent, nature, and distribution of the articles, was presented. As this was a scoping

XSL•FO
RenderX

review, the critical appraisal of the quality of the included studies was not conducted. However, efforts were made to map the diversity and variety of diet-related HRSs based on factors such as the characteristics of target users, function structure, recommendation content, and other factors. This process facilitated researchers in reaching conclusions about the key characteristics of research in this field and provided insights for future studies.

Results

Overview

In total, 4492 published studies were identified in the searching process (Figure 1). EndNote X9 was used to exclude 350 duplicates, and 4142 studies were excluded based on a review of their titles and abstracts. The remaining 208 studies were searched for full text. Ultimately, 193 were excluded based on the exclusion criteria, and 15 studies [35,38-51] were included and analyzed.

Figure 1. Flow diagram according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines.

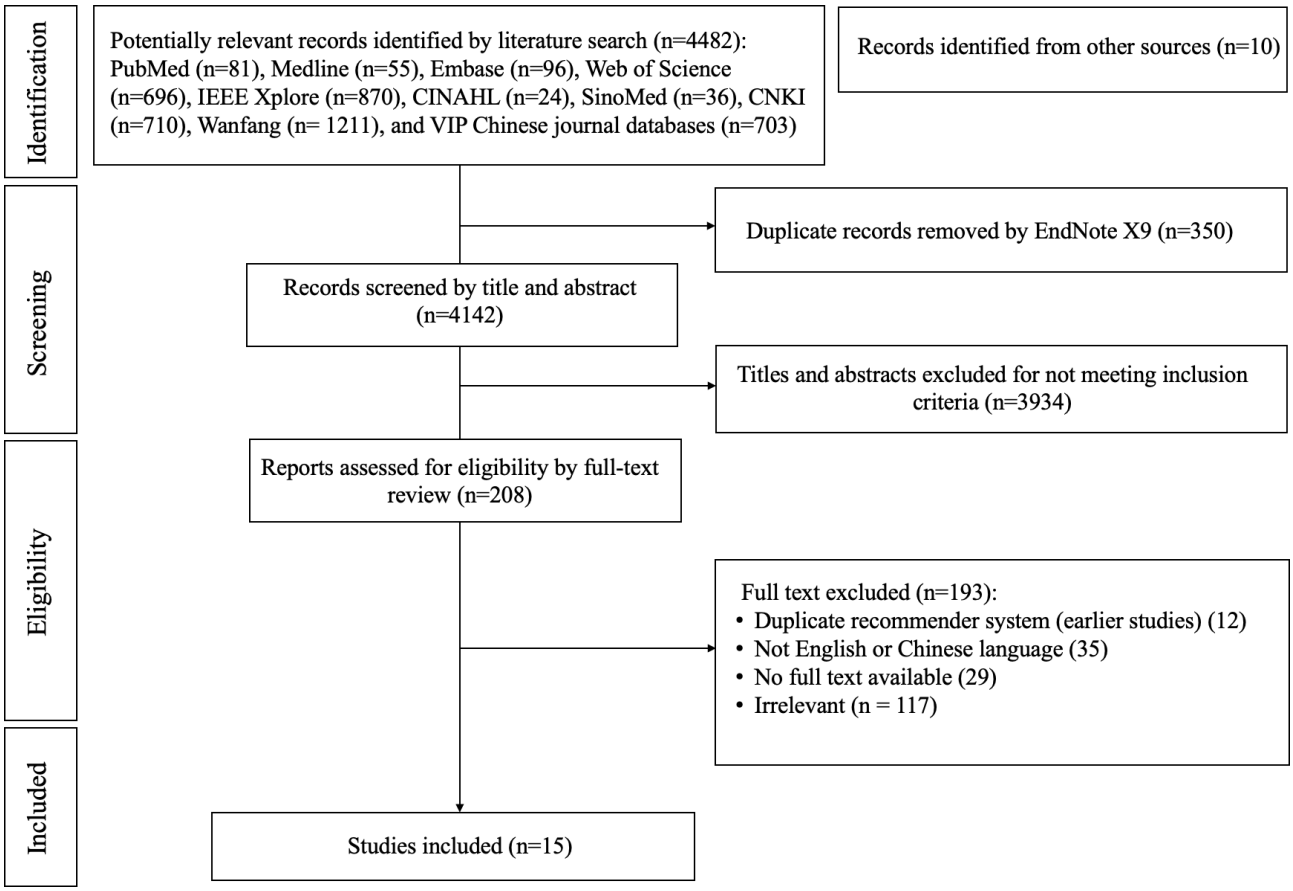

The Current State of Research on Diet-Related HRSs for Patients with Chronic Health Conditions

The publication years and the countries of the included studies are shown in Multimedia Appendix 2. The studies were published between 2010 and 2024, and there has not been a noticeable increase in their publication over the past 5 years. The articles originated from 9 countries (based on the first author's affiliations), with the top 3 being China (n=5, 33.33%) [39,40,43,44,49], India (n=2, 13.33%) [41,47], and Pakistan (n=2, 13.33%) [38,48].

Target Users for Diet-Related HRSs for Patients with Chronic Health Conditions

Among the 15 included studies [35,38-51], only 2 reported the age of the target users, which were 18 - 80 years [50] and older than 65 years [39], respectively. The target users and patients' chronic health conditions are shown in Table 1. Diet-related HRSs primarily target adults with chronic diseases. Among chronic diseases, diabetes and hypertension were the most commonly targeted by diet-related HRSs, with 9 systems (60%) including users with diabetes and 6 (40%) including users with hypertension.

XSL•FO
RenderX

Table . The characteristics of target users for diet-related Health Recommender Systems.

Variable	Studies, n (%)	References
Target users		
Patients with a single disease	6 (40)	[38,40,41,44,46,51]
Patients with complex diseases	4 (26.67)	[39,43,47,48]
Both healthy people and patients	5 (33.33)	[35,42,45,49,50]
Patients' chronic health conditions		
Diabetes	9 (60)	[35,38,42,45-49,51]
Hypertension	6 (40)	[35,40,45,47-49]
Cancer	3 (20)	[41,42,44]
Obesity	3 (20)	[42,47,50]
Kidney disease	3 (20)	[35,47,48]
High cholesterol	2 (13.33)	[47,49]
Metabolic syndrome	1 (6.67)	[43]
Osteoporosis	1 (6.67)	[42]
Iron deficiency	1 (6.67)	[48]
Cardiovascular diseases	1 (6.67)	[42]
Chronic dental problems	1 (6.67)	[42]
Unspecified geriatric diseases	1 (6.67)	[39]

The Function Structures in Diet-Related HRSs for Patients with Chronic Health Conditions

Among the 15 studies included [35,38-51], 9 (60%) [35,38-45] studies described function structures. The function structures described in these studies varied, but certain recurring components were consistently noted, which could be summarized into 4 major components: user information, food or diet recommendations, knowledge and decision support, and data management and additional functions. Detailed information is shown in Table 2.

XSL•FO
RenderX

Table . The function structures in diet-related Health Recommender Systems.

Variable	References
User information	
Basic user information	
Personal/user profile	[35,38,39]
User information	[40]
History	[35,41,42]
Personal health information	
Physical activity	[38]
Health management report	[43]
Diseases	[44]
Symptoms	[44]
Diagnoses	[41]
Food or diet recommendations	
Food filter and security	[35]
Food recommendations or suggestions	[38,45]
Weekly meal plans	[42]
Diet guides	[39,40]
Knowledge and decision support	
Chronic kidney disease calculator	[35]
Food	[38]
Nutrient search engines	[45]
News and related sites	[45]
Ingredients	[44]
Nutritional expert knowledge	[39]
Clinical practice and nutrition guidelines	[39]
Knowledge base	[39]
Diet charts	[41]
User ratings	[42]
Data management and additional functions	
Data entry	[38]
Data management	[40]
Reminders	[35]
Real-time interaction	[43]
Settings	[40]
Changing details	[41]

The Types of Recommendation Content Provided by Diet-Related HRSs for Patients with Chronic Health Conditions

The recommended content was divided into 5 categories. Detailed information is shown in Table 3.

XSL•FO
RenderX

Table . The types of recommendation content provided by the diet-related Health Recommender Systems.

Recommended content	Studies, n (%)	References
Food	6 (40)	[35,41,46-49]
Recipe	4 (26.67)	[38,42,44,45]
Diet plan or meal plan	3 (20)	[40,43,50]
Both recipe and food	1 (6.67)	[51]
Meal	1 (6.67)	[39]

The Implementation of Recommendation Features in Diet-Related HRSs for Patients with Chronic Health Conditions

Table 4 presents detailed information on the implementation of recommendation features in the reviewed diet-related HRSs. Recommendation methods were inductively classified into 3 categories: constraint-based, preference-based, and hybrid. In this review, constraint-based refers to recommendations based on the patient's condition-related dietary restrictions, while preference-based refers to recommendations based on the patient's dietary preferences. Hybrid refers to methods that incorporate both constraint-based and preference-based approaches. Among the 15 included studies [35,38-51], 6 (40%) used constraint-based methods [41,43,46,47,49,51], 5 (33.33%) applied preference-based methods [39,40,44,45,50], and 4 (26.67%) adopted hybrid methods [35,38,42,48].

XSL•FO
RenderX

Table . The implementation of recommendation features in diet-related Health Recommender Systems.

Author [reference], publication year	Recommendation method	Recommendation technology	Data of the training set	Recommendation process
Phanich et al [46], 2010	Constraint-based	Hybrid recommendation	Data were selected from the nutrition division, Ministry of Public Health, to form the training dataset (n=290[a])	• The dataset was grouped by the system and categorized according to food characteristics and nutrition for diabetes (normal food, limited food, and avoidable food). • Features were extracted by nutrient ranking. • Food clustering was analyzed. • Relevant food items were ranked based on the minimum distance, extracted, and recommended in ascending order from the ranking.
Arwan et al [51], 2013	Constraint-based	Hybrid recommendation	Data were selected from some references of foods and enriched with the information from nutrition experts to form the training dataset (n=NR[b]).	• The food ontology and calorie food ontology were developed. • The rule for classifying or grouping data categories within the ontology was developed. • The food menu search feature was built. • An experiment was conducted to test whether the system could recommend correctly.
Faiz et al [38], 2014	Preference-based and constraint-based	Hybrid recommendation	Data were selected from standard and well-known resources such as the US Department of Agriculture and MyFitnessPal to form the training dataset (n=NR[b]).	• The domain ontologies (personal health profile, food, and diseases) were built. • The ontologies were integrated and rule terms were defined. • The diet was recommended to the user at the predefined time.
Ting et al [45], 2014	Preference-based	Hybrid recommendation	Data were selected from the Food Composition Database and Japan Preventive Association of Lifestyle-related Disease to form the training dataset (n=NR[b]).	• A query with the user data was sent to the system. • Recipes were obtained from the database through a search. • Recipes that did not suit the current health status or match the user's preferences were filtered out. • Five recipes were recommended from the filtered set.

XSL•FO
RenderX

Author [reference], publication year	Recommendation method	Recommendation technology	Data of the training set	Recommendation process
Chen et al [49], 2015	Constraint-based	Hybrid recommendation	Data were collected from the Taiwan Area Food Nutrition Database website published by the US Food and Drug Administration to form the training dataset (n=NR[b]).	• The system collected personal information and classified the user's nutrients. • The patient's dietary records were imported into the nutrition expert knowledge system. • The personal information was imported into the personal disease ontology. • The nutrient data were analyzed by the expert knowledge system. • Suitable foods for the user were inferred.
Tseng et al [43], 2015	Constraint-based	Hybrid recommendation	__[c]	• The patient's vital signs were collected and transmitted. • The risk was evaluated and reported. • A diet plan was generated and recommended to the patient.
Elsweiler et al [50], 2015	Preference-based	Hybrid recommendation	Data were selected from a self-created food portal website, including users (n=148[d]) and recipes (n=957[a]) to form the training dataset.	• The user's nutritional requirements were established. • Recipe ratings were estimated based on each profile. • Recipes were combined. • Recipes were established if the combination met the requirements.
Rehman et al [48], 2017	Preference-based and constraint-based	Hybrid recommendation	Data were selected from the official website of the Composition of Foods Integrated Dataset (CoFID) to form the training dataset (n=3400[a]).	• A cloud-based food recommender system, named Diet-Right, was developed based on users' pathological reports. • The ant colony algorithm was used to generate an optimal food list. • Suitable foods were recommended according to the values of pathological reports.
Agapito et al [35], 2018	Preference-based and constraint-based	Hybrid recommendation	__[c]	

XSL•FO
RenderX

Author [reference], publication year	Recommendation method	Recommendation technology	Data of the training set	Recommendation process
				• The user profile was created by giving specific questions about clinical parameters. • All changes made by the user were saved, allowing the data to be used for monitoring the user's health status. • After the user was profiled, typical foods that could be consumed by the user were recommended.
Rathi et al [41], 2019	Constraint-based	Hybrid recommendation	Data were selected from the UC Irvine Machine Learning Repository (including the Liver Patient Dataset, Heart Disease Dataset, Diabetes Dataset, Breast Cancer Dataset, and Thyroid Dataset).	• The dataset was collected and preprocessed from different sources. • The system was built to detect diseases and recommend diets accordingly.
Manoharan et al [47], 2020	Constraint-based	Hybrid recommendation	Data were collected from 50 patients through the internet and hospitals to form the training dataset (n=50[d]).	• Data were collected from the internet and hospitals. • Features were sorted, preprocessed, and encoded, and the food data were segregated based on similarities. • Food was recommended. • The system was trained, tested, and cross-validated.
Qi et al [40], 2021	Preference-based	Hybrid recommendation	Data were selected from hospital meal history records and split into training and testing sets in a 7:3 ratio (n=718[a]).	• A MySQL database was used to implement data dictionary management. • The rule extraction module was implemented with the knowledge base management system. • A meal plan was generated and recommended.
Tang et al [44], 2023	Preference-based	Knowledge graph	Data were crawled from recipe websites and manually extracted from textbooks to form the training dataset (n=NR[b]).	• The embedded representation of items was enhanced through message-passing and update functions on node features. • The influence of time on users' taste preferences was considered. • Long Short-Term Memory (LSTM) networks were introduced to dynamically adjust users' personal taste preferences.

XSL•FO
RenderX

Author [reference], publication year	Recommendation method	Recommendation technology	Data of the training set	Recommendation process
Zioutos et al [42], 2024	Preference-based and constraint-based	Hybrid recommendation	Data were selected from a large real-world dataset of recipes to form the training dataset (n=2,774,676[a]).	• The user's health history was analyzed. • Similar users were identified. • Personalized recipe recommendations were provided. • Dynamically adaptable adjustments were made.
Xu et al [39], 2024	Preference-based	Knowledge graph	Data were selected from a community survey (n=96[d]) and a website of Chinese cuisine recipes and eating history (n=180[a]) to form the training dataset.	• The FoodKG was constructed. • User profiles related to older adults' dietary behaviors were built. • Personalized meal recommendation algorithms were developed, including candidate dish generation and combo meal recommendations.

[a]The sample size of subjects including a recipe, food, and diet plan or meal plan.

[b]NR: not reported.

[c]Not available.

[d]The sample size of the population.

While the recommendation technologies were coded according to terms reported in the studies. The recommendation technology included: KG (n=2, 13.33%) [39,44] and HyR (n=13, 86.67%) [35,38,40-43,45-51].

For the data of the training set, 13 studies (86.67%) [38-42,44-51] explicitly mentioned the data sources of the training set, while the other 2 studies (13.33%) [35,43] did not specify this information. The data sources used in the reviewed studies varied widely and can be categorized into 4 main types: authoritative government and institutional databases (n=5, 33.33%) [38,45,46,48,49], academic databases and publicly available datasets (n=2, 13.33%) [39,41], expert and hospital data (n=3, 20%) [40,47,51], and recipe and user-generated data (n=3, 20%) [42,44,50].

The recommendation process followed a structured workflow, integrating advanced technologies to generate more accurate, adaptive, and context-aware dietary recommendations. The structured workflow included user profiling, integration of structured knowledge (such as food databases or ontologies), personalized filtering and matching based on health conditions and preferences, and ranking of suitable options for final recommendation. While this core process was consistent, 6 systems incorporated advanced technologies to enable more accurate, adaptable, and context-aware dietary recommendations, such as ontology-based reasoning [51], optimization algorithms (eg, ant colony [48]), dynamic modeling using Long Short-Term Memory networks [44], expert rule systems [39,49], and the use of KGs such as FoodKG [39].

The Evaluation of Diet-Related HRSs for Patients with Chronic Health Conditions

Among the 15 included studies [35,38-51], 12 studies (80%) [35,39-44,46-49,51] evaluated the diet-related HRSs, while the remaining 3 studies (20%) [38,45,50] did not conduct an evaluation. Table 5 presents detailed information on the evaluation criteria, evaluation method, and test set or evaluation sample size of the 12 studies [35,39-44,46-49,51].

XSL•FO
RenderX

Table . The evaluation of diet-related Health Recommender Systems (n=12).

Author [reference], publication year	Evaluation criteria	Evaluation method	Test set or evaluation sample size
Phanich et al [46], 2010	Acceptance, usability, and accuracy	Online	Nutritionists (n=NR[a])
Arwan et al [51], 2013	Accuracy	Offline	Ontology patient instances (n=30[b])
Chen et al [49], 2015	Accuracy	Online	One older adult from a silver-haired home and dietitians (n=NR[a])
Tseng et al [43], 2015	Feasibility	Online	Patients (n=NR[a])
Rehman et al [48], 2017	Accuracy	Offline	Same as the training set
Agapito et al [35], 2018	Accuracy and usability	Online	Patients with CKD[c] (n=20[b]); healthy people (n=20[b])
Rathi et al [41], 2019	Accuracy	Offline	Same as the training set. The dataset was split into training:testing 75:25 (n=9326[d])
Manoharan et al [47], 2020	Accuracy	Offline	Same as the training set
Qi et al [40], 2021	Efficiency and satisfaction	Online	Patients (n=37[b]) and nutritionists (n=3[b])
Tang et al [44], 2023	Accuracy	Offline	Same as the training set
Zioutos et al [42], 2024	Accuracy and usability	Online	40 existing users of the food.com website: patients (n=20[b]) and healthy people (n=20[b])
Xu et al [39], 2024	Effectiveness	Online	Community-dwelling older adults (n=96[b]): tracked group (n=34[b]) and an untracked group (n=62[b]); A total of 91 participants (94.79%) were diagnosed with chronic conditions

[a]NR: not reported.

[b]The sample size of the population.

[c]CKD: chronic kidney disease.

[d]The sample size of subjects including a recipe, food, and diet plan or meal plan.

The evaluation criteria were extracted directly from the included studies, including: accuracy (n=9) [35,41,42,44,46-49,51] refers to the extent to which the system's predictions of patients' preferences or nutritional needs match their actual dietary preferences or intake; usability (n=3) [35,42,46] refers to the ease with which patients can, with little or no assistance, successfully operate the system to receive recommendations, navigate, and interact with its interface; acceptance (n=1) [46] refers to the degree to which patients are willing to receive and use the dietary recommendations provided by the system; satisfaction (n=1) [40] refers to patients' subjective sense of contentment or evaluation of the overall usefulness and experience of the recommendations, measured through participants' comprehensive comparison of system-generated meal plans with manually designed ones; efficiency (n=1) [40] refers to the amount of time, steps, or cognitive effort required for patients to obtain appropriate dietary recommendations from the system; feasibility (n=1) [43] refers to the practicality of deploying the recommendation system in real-life settings; and effectiveness (n=1) [39] refers to the actual positive impact of the system in real-life contexts on patients' dietary behavior changes or health outcomes. Four studies [35,40,42,46] included more than one evaluation criterion.

The evaluation methods included: online evaluation (n=7) [35,39,40,42,43,46,49] and offline evaluation (n=5) [41,44,47,48,51]. Among the 7 studies [35,38,42,44,45,48,51] that used online evaluation, 6 studies [35,39,40,42,43,49] evaluated the diet-related HRSs with target users of patients, while 3 studies evaluated the HRSs that relied on experts, such as nutritionists [40,46] and dietitians [49]; their main roles were to validate the recommendation results. Four of the 5 offline evaluation studies [39,40,43,46,47] used the same dataset for training and evaluation [41,44,47,48].

Discussion

Principal Findings

This scoping review revealed that diet-related HRSs for individuals with chronic conditions were still in their early stages, with limited patient-specific designs and significant room for improvement. The review highlighted substantial gaps in target users, system functions, recommendation content, recommendation feature implementation, and evaluation approaches, which need to be addressed to support the

XSL•FO
RenderX

development of more effective and patient-centered diet-related HRSs.

Comparison with Prior Work

Target Users of Diet-Related HRSs

The target users of diet-related HRSs were mainly patients with single or complex chronic health conditions. They not only need dietary recommendations tailored to their preferences [39,41,42,44,48,50], but most importantly, their health conditions [38,40,45,46,51], highlighting the need for more disease-specific solutions. Age groups are another factor that needs to be considered, for example, older adults often face challenges such as limited mobility, cognitive decline, and changes in appetite or food preferences [40,42], while sick children and their caregivers face unique dietary needs due to treatment-related restrictions and appetite loss [52,53]. Therefore, balancing disease-related requirements with personal needs is fundamental to developing high-quality diet-related HRS. It is also essential to support caregivers in preparing meals that are nutritionally appropriate and medically compliant.

Function Structures of Diet-Related HRSs

The user information component enables diet-related HRSs to collect basic demographic and health data, which supports the generation of tailored dietary recommendations [35,38-40]. By incorporating dynamic inputs such as food tracking and activity logs, the system can further provide personalized, context-sensitive advice with timely feedback and adjustments based on patients' health progress [43,54]. Knowledge and decision support functions served as a critical layer in enhancing the intelligence and reliability. For example, by integrating chronic kidney disease calculators [35], nutrient search engines [45], and expert knowledge bases [39], these systems move beyond simple food suggestions to provide evidence-based recommendations. This clarified how the nutritional rationale, which potentially enhances patients' understanding and trust, long-term adherence, and sustainable dietary behavior change [44,45]. The accuracy and relevance of food or diet recommendations depend on the robustness of the algorithm used to process the data and the diversity of food or recipes options integrated into the system [55-57]. Data management and additional interactive functions have enhanced the usability and clinical relevance of diet-related HRSs and supported user engagement, empowering patients to actively and continuously manage their dietary self-care.

Recommendation Content of Diet-Related HRSs

Patients' needs for the content of diet-related HRSs are multidimensional. Beyond simple food lists, an effective diet-related HRS must provide actionable guidance on cooking methods, ingredient combinations, and personalized meal plans [39,42-44,47,48,50]. Well-structured recipes emerge as a key feature, as they help users visualize meal preparation and improve understanding of the nutritional rationale behind recommendations [44]. Such transparency fosters engagement and adherence, particularly when supported by nutrition professionals [58,59]. Additionally, offering varied recipe options allows patients to personalize their meal plans according to their tastes, dietary restrictions, and the availability of ingredients [39]. This flexibility caters to users with complex health conditions and empowers them with a sense of control, which is vital for sustaining long-term dietary management [42].

Implementation of Recommendation Features in Diet-Related HRSs

Recommendation Methods

In diet-related HRSs, hybrid approaches appear particularly valuable for balancing clinical appropriateness with individual preferences to promote long-term adherence. In addition, adaptability is an essential feature, allowing systems to respond to changing health conditions, behaviors, and user needs. As patients' health status and preferences evolve, systems such as "SHARE" demonstrate the value of dynamic updates that allow users to refine their meal plans, making the system more user-centered and responsive [42]. However, a critical gap across the reviewed studies is the insufficient consideration of medication-food interactions. For patients with chronic conditions, clinically significant interactions such as warfarin-vitamin K [60], angiotensin-converting–enzyme inhibitors-potassium [61], or metformin-alcohol [62], carry substantial risks. Future diet-related HRSs should adopt hybrid approaches that integrate drug-nutrient knowledge and dynamic monitoring with preference learning, ensuring recommendations that are both personalized and clinically safe.

Recommendation Technology

Appropriate recommendation technology can transform users' vague needs into clear ones and filter out irrelevant information [63]. HyR that combines CF, CB, and KB methods is particularly effective. These methods overcome traditional limitations such as cold start and data sparsity by integrating multiple information sources, including nutritional knowledge and disease-specific data [64]. Notably, KB and KG-driven algorithms were especially well-suited for diet-related HRSs. By incorporating expert knowledge and medical guidelines, they ensure scientific accuracy, clinical reliability, and alignment with patients' dietary management goals, resulting in more comprehensive and robust systems [65].

Data Sources

The effectiveness of diet-related HRSs largely depends on the quality and appropriateness of data sources. Patient data mainly come from personal input, including medical records, treatment history, and sociodemographic information [35,38-45,47-49]. Incorporating dietary behavior data (eg, meal logs, food purchase records, and subjective reports) and clinical data (eg, diagnosis, treatment, and disease stage) helps ensure that dietary recommendations are both personalized and medically appropriate [66,67]. Although patient-specific data are essential for personalized recommendations [35,44], sensitive medical and behavioral data entail serious privacy, security, and ethical risks. Given patients' fragile health status and dietary restrictions, dietary data from professional resources, such as the Nutrition Division of the Ministry of Public Health (MOPH) [46], the official website of the composition of foods integrated dataset (CoFID) [48], and the Japan Preventive Association of Lifestyle-related Disease (JPALD) [45], are vital for ensuring the accuracy of recommendations. In contrast, nonprofessional

XSL•FO
RenderX

sources such as general cooking websites are not recommended in clinical contexts due to their limited reliability, particularly for patients with complex conditions.

Implementation Process

The Association for Computing Machinery has issued guidelines for designing and evaluating RSs [68]. However, no widely accepted standards have been established for implementing HRSs, especially diet-related HRSs. This gap highlights challenges arising from disease diversity, complex food preferences, and context-dependent factors that are difficult to acquire or simulate [28]. The reviewed systems generally followed a similar workflow, involving user profiling, structured knowledge integration, personalized filtering, and ranking, reflecting a growing understanding of how to align dietary recommendations with individual health needs and preferences. The integration of advanced recommendation technologies, such as ontology-based reasoning [51], optimization algorithms (eg, ant colony methods [48]), Long Short-Term Memory–based dynamic modeling [44], and KGs [39], has advanced more intelligent and context-aware systems. However, the adoption of these technologies remains limited, and few studies have systematically evaluated their impact on recommendation quality or patient outcomes.

Evaluation of Diet-Related HRSs

Evaluation Criteria

In diet-related HRSs, accuracy has been the primary evaluation criterion, typically evaluated by prediction score accuracy, prediction score correlation, classification accuracy, and sorting accuracy [69]. However, the rationality of the recommendations has received limited attention. None of the 15 reviewed studies [35,38-51] evaluated the scientific soundness of food or recipe recommendations. While health care providers were commonly involved in evaluating the recommendation results, official institutional input on content was rare, revealing a major gap in ensuring the credibility and reliability of recommendations.

Beyond accuracy and content rationality, the effectiveness of diet-related HRSs is closely tied to their ability to promote healthy eating behaviors. The key goal of the diet-related HRSs is to empower users to make informed dietary choices and motivate healthier behaviors [70]. Existing evaluations of behavior change have primarily examined improvements in diet quality and diversity, measured by the China Elderly Dietary Guidelines Index and the dietary diversity score [50]. Although user engagement and personalized feedback were recognized as essential for sustaining behavior change, few studies evaluated long-term effects on patients' dietary behaviors.

Moreover, clinical trials in real-world health care settings remain lacking; even the study [39] involving patients was an observational cohort study rather than a clinical trial, limiting statistical power and generalizability. Additionally, no studies examined the cost-effectiveness of implementing diet-related HRSs in health care settings, which is crucial for adoption in resource-constrained healthcare systems.

Evaluation Methods and Process

The evaluation of diet-related HRSs can be carried out using various methods, each with its advantages and limitations. Online evaluations, which assess system performance through real-time user feedback or surveys, provide valuable insights into user interactions but often involve higher costs [22]. Offline evaluations in controlled settings are more feasible but have limited external validity due to their inability to reflect real-world usage [22,71]. Emerging approaches such as online A or B testing enable rapid, cost-effective comparisons of different system versions, helping predict broader performance even with small samples [72,73].

Diet-related HRS evaluations generally follow a staged approach from development to postdeployment. During development, stakeholder involvement in usability evaluation is critical to ensure user-centered design [74]. After development, accurate assessments verify that recommendations are reliable and relevant to target users [35,41,42,44,46-49,51]. Postdevelopment, gathering user feedback on effectiveness and satisfaction is essential for refining the system and improving its responsiveness to user needs [35,39,40,42,43,46]. These stages form a comprehensive evaluation process ensuring the reliability and usability of diet-related HRSs.

Although several reviews have examined HRSs, few have focused specifically on diet-related HRSs. One scoping review of 36 studies identified only 8 diet-focused HRSs [75]. A systematic review of 73 studies identified 26 nutrition-related systems, most of which targeted populations without strict medical dietary requirements, thereby revealing the lack of disease-specific syntheses [22]. Existing diet-related HRS reviews have primarily addressed technical aspects, such as semantic interoperability [76], explainable artificial intelligence [77], and data-driven personalization methods [78]. They offer useful technical insights but rarely address disease-specific dietary needs or the related clinical and safety challenges. In contrast, our review focuses on diet-related HRSs for patients with chronic conditions. It highlights key requirements such as clinical data integration, dynamic health monitoring, tailored recipe-level recommendations, and persistent gaps in evaluating behavior change and clinical outcomes. These disease-focused insights extend prior work and inform the development of more personalized and clinically reliable diet-related HRSs.

Future Directions

Future research should prioritize developing diet-related HRSs based on a user-centered design framework. These systems must be tailored to patients' specific needs with a deep understanding of their cognitive abilities, digital literacy, cultural dietary preferences, and daily routines. Diet-related HRSs need to address clinical challenges through contextually relevant functional designs, customized for different patient groups, from older adults with multimorbidity to pediatric patients requiring caregiver-mediated support. Practical usability features, such as adaptive reminders, contextualized recipe guidance, and dynamic feedback loops, are essential for real-world effectiveness.

XSL•FO
RenderX

To support patients' sustained dietary behavior change, recommendations should move beyond static food lists toward evidence-based, actionable guidance using practical tools such as portion guides, ingredient lists, and meal preparation videos [79]. Future systems should adopt hybrid, adaptive recommendation mechanisms and integrate with electronic health records (EHRs) and mobile apps to enhance personalization, interpretability, clinical relevance, and engagement, which is essential for sustaining long-term adherence to dietary changes. Additionally, incorporating behavior change theories, such as Social Cognitive Theory [80], Theory of Planned Behavior [81], and the Fogg Behavior Model [82], can strengthen the dietary intervention design. Gamification elements informed by these theories, such as rewards and rankings, may further motivate users and increase adherence [83].

When managing sensitive health data, especially when integrating with EHRs and personal medical records, the systems must implement robust data governance, including secure storage, encryption, role-based access, and transparent consent. Compliance with regulations such as the US Health Insurance Portability and Accountability Act (HIPAA) [84], the European Union's General Data Protection Regulation (GDPR) [85], and national standards are essential to safeguard patient privacy and foster user trust. Finally, standardized evaluation methods and robust frameworks, combining clinical trials, behavior change evaluation, and usability evaluation, are essential to verify effectiveness, accuracy, and long-term impact in real-world settings, while ensuring cost-effectiveness and compliance (eg, US Food and Drug Administration approval and European Conformity marking) for safe and scalable healthcare implementation.

Implication for Clinical Practice

For clinicians and dietitians, diet-related HRSs can serve as supportive tools to deliver personalized, evidence-based, and adaptive dietary guidance aligned with patients' medical conditions, treatment regimens, cultural preferences, and daily routines. By incorporating medication-food interaction checks and actionable educational content, these systems can enhance patient safety, dietary adherence, and long-term health management. Future diet-related HRSs could integrate seamlessly into clinical workflows and demonstrate measurable health outcomes. Embedding diet-related HRSs within EHRs would ensure that dietary recommendations are consistent with medications, lab results, and care plans. By linking dietary adherence to physiological indicators such as hemoglobin A_{1c}, blood pressure, or lipid levels, systems can enable real-time feedback and timely clinical intervention. Beyond accuracy, future evaluations should predefine clinical end points (eg, hemoglobin A_{1c} reduction or blood pressure control) and validate effectiveness through randomized controlled trials and real-world studies.

Limitations

This study has several limitations. First, only studies published from January 2010 to October 2024 were included, which may not reflect recent developments. Second, only English and Chinese literature were reviewed, excluding studies in other languages. Third, the focus was primarily on chronic health conditions, with limited exploration of diet-related HRSs for acute conditions or general wellness, suggesting a need for broader research in future studies. In addition, the included studies themselves demonstrated certain methodological limitations, such as small sample sizes, reliance on offline testing, and limited evaluation of actual behavior change. These issues reflect not only the limitations of this review but also highlight gaps in the existing research that warrant more rigorous clinical evaluation.

Conclusions

Diet-related HRSs can offer personalized dietary recommendations for patients with chronic conditions. The analysis reveals significant gaps, demanding that future systems be grounded in user-centered design to meet patient-specific needs. These systems must recommend more practical and actionable dietary guidance. The adoption of hybrid recommendation techniques can enhance the precision and clinical relevance of dietary recommendations. Establishing standardized evaluation metrics and conducting real-world studies with long-term follow-up will be essential. This will verify the system's ability to positively change dietary behaviors and improve clinical outcomes. Addressing these issues will transform diet-related HRSs into trustworthy and impactful tools for managing chronic diseases and delivering patient-centered care.

Acknowledgments

We would like to thank Liping Fu for her assistance and guidance in developing the search strategy.

Generative artificial intelligence and artificial intelligence–assisted technologies in the writing process: during the preparation of this work, the authors used ChatGPT 4.0 to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Funding

This study was supported by the National Natural Science Foundation of China (grant number 72374204) and Soft Science Research Project of Shanghai (25692108400).

XSL•FO
RenderX

Authors' Contributions

Conceptualization: XD, JW
Data curation: XD, BY, HN, TJ, JW
Formal analysis: XD, BY, ZZ, JW
Project administration: XD, JW
Supervision: AP, CY, JW
Visualization: XD
Writing – original draft: XD, BY, ZZ, JW
Writing – review & editing: XD, BY, AP, JW

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[PDF File, 850 KB - jmir_v28i1e77726_app1.pdf]

Multimedia Appendix 2
The publication years and countries of the included studies.
[PNG File, 82 KB - jmir_v28i1e77726_app2.png]

Checklist 1
PRISMA-ScR Checklist.
[PDF File, 114 KB - jmir_v28i1e77726_app3.pdf]

References

1. Murray CJL, Aravkin AY, Zheng P, et al. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct;396(10258):1223-1249. [doi: 10.1016/S0140-6736(20)30752-2]
2. The Global Nutrition Report. URL: https://globalnutritionreport.org/ [accessed 2025-12-31]
3. Ong KL, Stafford LK, McLaughlin SA, et al. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2023 Jul;402(10397):203-234. [doi: 10.1016/S0140-6736(23)01301-6]
4. Magkos F, Hjorth MF, Astrup A. Diet and exercise in the prevention and treatment of type 2 diabetes mellitus. Nat Rev Endocrinol 2020 Oct;16(10):545-555. [doi: 10.1038/s41574-020-0381-5] [Medline: 32690918]
5. Filippou CD, Tsioufis CP, Thomopoulos CG, et al. Dietary Approaches to Stop Hypertension (DASH) diet and blood pressure reduction in adults with and without hypertension: a systematic review and meta-analysis of randomized controlled trials. Adv Nutr 2020 Sep 1;11(5):1150-1160. [doi: 10.1093/advances/nmaa041] [Medline: 32330233]
6. Zhuang P, Wang F, Yao J, et al. Unhealthy plant-based diet is associated with a higher cardiovascular disease risk in patients with prediabetes and diabetes: a large-scale population-based study. BMC Med 2024 Oct 23;22(1):485. [doi: 10.1186/s12916-024-03683-7] [Medline: 39443972]
7. Lockwood G, Davey L, McFarlane C, Gray NA, Wright HH. Factors influencing meal provision and dietary support behaviour of caregivers of people with chronic kidney disease: a cross-sectional study. Nutrients 2024 Oct 14;16(20):3479. [doi: 10.3390/nu16203479] [Medline: 39458474]
8. Xiong H, Zhang X, Zeng H, Xie S, Yi S. Experience of diet in patients with inflammatory bowel disease: a thematic synthesis of qualitative studies. J Clin Nurs 2024 Aug;33(8):3283-3293. [doi: 10.1111/jocn.17186] [Medline: 38661241]
9. Scientific research report on dietary guidelines for Chinese residents [Website in Chinese]. : National Institute for Nutrition and Health, Chinese Center for Disease Control and Prevention; 2021 URL: https://www.chinanutri.cn/yyjkzxpt/yyjkkpzx/yytsg/zgjm/202103/t20210311_224598.html [accessed 2025-12-31]
10. Dahl C, Crichton M, Jenkins J, et al. Evidence for dietary fibre modification in the recovery and prevention of reoccurrence of acute, uncomplicated diverticulitis: a systematic literature review. Nutrients 2018 Jan 27;10(2):137. [doi: 10.3390/nu10020137] [Medline: 29382074]
11. Collins N, Belkaid Y. Control of immunity via nutritional interventions. Immunity 2022 Feb 8;55(2):210-223. [doi: 10.1016/j.immuni.2022.01.004] [Medline: 35139351]
12. Dietary Guidelines for Americans: US Department of Health and Human Services and US Department of Agriculture; 2015. URL: https://odphp.health.gov/sites/default/files/2019-09/2015-2020_Dietary_Guidelines.pdf [accessed 2025-11-05]

XSL•FO
RenderX

13. Flaskerud JH. Mood and food. Issues Ment Health Nurs 2015 Apr;36(4):307-310. [doi: 10.3109/01612840.2014.962677] [Medline: 25988966]
14. Notice on the issuance of dietary guidelines for the prevention and management of hyperuricemia and gout in adults (2024 edition) and three other dietary guidelines [Report in Chinese]. : National Health Commission of the People's Republic of China; 2024 URL: https://www.nhc.gov.cn/wjw/c100378/202402/afd5dda4bd6745fda10aad8d43a16369/files/1732845129551_37712.pdf [accessed 2025-12-31]
15. Rines J, Daley K, Loo S, et al. A patient-led, peer-to-peer qualitative study on the psychosocial relationship between young adults with inflammatory bowel disease and food. Health Expect 2022 Aug;25(4):1486-1497. [doi: 10.1111/hex.13488] [Medline: 35383400]
16. Singh PK, Pramanik PKD, Dey AK, Choudhury P. Recommender systems: an overview, research trends, and future directions. Int J Bus Syst Res 2021;15(1):14. [doi: 10.1504/IJBSR.2021.111753]
17. Su X, Khoshgoftaar TM. A survey of collaborative filtering techniques. Adv Artif Intell 2009 Oct 27;2009(1):1-19. [doi: 10.1155/2009/421425] [Medline: 41092928]
18. Lops P, de Gemmis M, Semeraro G. Content-based recommender systems: state of the art and trends. In: Recommender Systems Handbook: Springer; 2011:73-105. [doi: 10.1007/978-0-387-85820-3_3]
19. Burke R. Knowledge-based recommender systems. Encycl Libr Inf Syst 2000;69(Supplement 32) [FREE Full text]
20. Hou S, Wei D. Research on knowledge graph-based recommender systems. Presented at: 2023 3rd International Symposium on Computer Technology and Information Science (ISCTIS); Jul 7-9, 2023. [doi: 10.1109/ISCTIS58954.2023.10213083]
21. Muñoz EG, Parraga-Alava J, Meza J, Proaño Morales JJ, Ventura S. Housing fuzzy recommender system: a systematic literature review. Heliyon 2024 Mar 15;10(5):e26444. [doi: 10.1016/j.heliyon.2024.e26444] [Medline: 38439861]
22. De Croon R, Van Houdt L, Htun NN, Štiglic G, Vanden Abeele V, Verbert K. Health Recommender Systems: systematic review. J Med Internet Res 2021 Jun 29;23(6):e18035. [doi: 10.2196/18035] [Medline: 34185014]
23. Siriaraya P, Suzuki K, Nakajima S. Utilizing collaborative filtering to recommend opportunities for positive affect in daily life. HealthRecSys@ RecSys 2019 [FREE Full text]
24. Pasta A, Petersen MK, Jensen KJ, Larsen JE. Rethinking hearing aids as recommender systems. CEUR Workshop Proc 2019:11-17 [FREE Full text]
25. Sanchez Bocanegra CL, Sevillano Ramos JL, Rizo C, Civit A, Fernandez-Luque L. HealthRecSys: a semantic content-based recommender system to complement health videos. BMC Med Inform Decis Mak 2017 May 15;17(1):63. [doi: 10.1186/s12911-017-0431-7] [Medline: 28506225]
26. Agu E, Claypool M. Cypress: a cyber-physical recommender system to discover smartphone exergame enjoyment. Presented at: Proceedings of the ACM workshop on engendering health with recommender systems; Sep 15-16, 2016 URL: https://web.cs.wpi.edu/~claypool/papers/cypress-recsys/paper.pdf [accessed 2025-12-31]
27. Chavan P, Thoms B, Isaacs J. A recommender system for healthy food choices: building a hybrid model for recipe recommendations using big data sets. In: Chavan P, Thoms B, Isaacs J, editors. Presented at: Hawaii International Conference on System Sciences; Jan 5-8, 2021. [doi: 10.24251/HICSS.2021.458]
28. Mahajan P, Kaur PD. A systematic literature review of food recommender systems. SN Comput Sci 2024;5(1):1-26. [doi: 10.1007/s42979-023-02537-y]
29. Trattner C, Elsweiler D. Food recommender systems: important contributions, challenges and future research directions. arXiv. Preprint posted online on Nov 7, 2017. [doi: 10.48550/arXiv.1711.02760]
30. Islam T, Joyita AR, Alam M, Mehedi Hassan M, Hassan M, Gravina R. Human-behavior-based personalized meal recommendation and menu planning social system. IEEE Trans Comput Soc Syst 2022;10(4):2099-2110. [doi: 10.1109/TCSS.2022.3213506]
31. Vandeputte J, Herold P, Kuslii M, et al. Principles and validations of an artificial intelligence-based recommender system suggesting acceptable food changes. J Nutr 2023 Feb;153(2):598-604. [doi: 10.1016/j.tjnut.2022.12.022] [Medline: 36894251]
32. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
33. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
34. Song HS, Kim YA. A dog food recommendation system based on nutrient suitability. Expert Systems 2021 Mar;38(2). [doi: 10.1111/exsy.12623]
35. Agapito G, Simeoni M, Calabrese B, et al. DIETOS: a dietary recommender system for chronic diseases monitoring and management. Comput Methods Programs Biomed 2018 Jan;153:93-104. [doi: 10.1016/j.cmpb.2017.10.014] [Medline: 29157465]
36. Agapito G, Calabrese B, Guzzi PH, et al. DIETOS: a recommender system for adaptive diet monitoring and personalized food suggestion. Presented at: 2016 IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob); Oct 17-19, 2016. [doi: 10.1109/WiMOB.2016.7763190]
37. Wang J, Yuan C. Evidence-based scoping reviews: an introduction to the methodology. J Nurs Sci 2017;32(8):103-105 [FREE Full text] [doi: 10.3870/j.issn.1001-4152.2017.08.103]

XSL•FO
RenderX

38. Faiz I, Mukhtar H, Khan S. An integrated approach of diet and exercise recommendations for diabetes patients. Presented at: 2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom 2014); Oct 15-18, 2014. [doi: 10.1109/HealthCom.2014.7001899]
39. Xu Z, Gu Y, Xu X, et al. Developing a personalized meal recommendation system for Chinese older adults: observational cohort study. JMIR Form Res 2024 May 30;8:e52170. [doi: 10.2196/52170] [Medline: 38814702]
40. Qi M, Chen W, Li D, Yang F. Design of a knowledge-based nutritional meal recommendation system for patients. China Med Equip 2021;18(1). [doi: 10.3969/J.ISSN.1672-8270.2021.01.026]
41. Rathi M, Pareek V. Mobile based healthcare tool an integrated disease prediction & recommendation system. Int J Knowl Syst Sci 2019;10(1):38-62. [doi: 10.4018/IJKSS.2019010103]
42. Zioutos K, Kondylakis H, Stefanidis K. Healthy personalized recipe recommendations for weekly meal planning. Computers 2024;13(1):1. [doi: 10.3390/computers13010001]
43. Tseng JCC, Lin BH, Lin YF, et al. An interactive healthcare system with personalized diet and exercise guideline recommendation. Presented at: 2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI); Nov 20-22, 2015. [doi: 10.1109/TAAI.2015.7407106]
44. Tang J, Huang B, Xie M. Anticancer recipe recommendation based on cancer dietary knowledge graph. Eur J Cancer Care (Engl) 2023 Oct 18;2023:1-13. [doi: 10.1155/2023/8816960]
45. Ting YH, Zhao Q, Chen RC. Dietary recommendation based on recipe ontology. Presented at: 2014 IEEE 6th International Conference on Awareness Science and Technology (iCAST); Oct 29-31, 2014. [doi: 10.1109/ICAwST.2014.6981829]
46. Phanich M, Pholkul P, Phimoltares S. Food recommendation system using clustering analysis for diabetic patients. Presented at: 2010 International Conference on Information Science and Applications; Apr 21-23, 2010. [doi: 10.1109/ICISA.2010.5480416]
47. Manoharan D. Patient diet recommendation system using K clique and deep learning classifiers. J Artif Intell Capsule Netw 2020;2(2):121-130. [doi: 10.36548/jaicn.2020.2.005]
48. Rehman F, Khalid O, Haq N, Khan AUR, Bilal K, Madani S. Diet-Right: a smart food recommendation system. KSII TIIS 2017;11(6):2910-2925. [doi: 10.3837/tiis.2017.06.006]
49. Chen RC, Ting YH, Chen JK, Lo YW. The nutrients of chronic diet recommended based on domain ontology and decision tree. Presented at: 2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI); Dec 2-4, 2015. [doi: 10.1109/TAAI.2015.7407127]
50. Elsweiler D, Harvey M. Towards automatic meal plan recommendations for balanced nutrition. Presented at: RecSys '15; Sep 16-20, 2015; Vienna, Austria p. 313-316. [doi: 10.1145/2792838.2799665]
51. Arwan A, Sidiq M, Priyambadha B, Kristianto H, Sarno R. Ontology and semantic matching diabetic food recommendations. Presented at: 2013 International Conference on Information Technology and Electrical Engineering (ICITEE); Oct 7-8, 2013. [doi: 10.1109/ICITEED.2013.6676233]
52. Wang Y, Wang J, Yu S, Zhou F, Yuan C. Qualitative study on the needs of parents of children with newly diagnosed acute lymphoblastic leukemia. J Nurs PLA China 2016;33(4):6-10 [FREE Full text]
53. van den Brink M, Ter Hedde MM, van den Heuvel E, Tissing WJE, Havermans RC. The impact of changes in taste, smell, and eating behavior in children with cancer undergoing chemotherapy: a qualitative study. Front Nutr 2022;9:984101. [doi: 10.3389/fnut.2022.984101] [Medline: 36245523]
54. Zenun Franco R, Fallaize R, Weech M, Hwang F, Lovegrove JA. Effectiveness of web-based personalized nutrition advice for adults using the eNutri web app: evidence from the EatWellUK randomized controlled trial. J Med Internet Res 2022 Apr 25;24(4):e29088. [doi: 10.2196/29088] [Medline: 35468093]
55. Celis-Morales C, Livingstone KM, Marsaux CF, et al. Effect of personalized nutrition on health-related behaviour change: evidence from the Food4Me European randomized controlled trial. Int J Epidemiol 2017 Apr 1;46(2):578-588. [doi: 10.1093/ije/dyw186] [Medline: 27524815]
56. Ordovas JM, Ferguson LR, Tai ES, Mathers JC. Personalised nutrition and health. BMJ 2018 Jun 13;361:bmj.k2173. [doi: 10.1136/bmj.k2173] [Medline: 29898881]
57. Melese A. Food and restaurant recommendation system using hybrid filtering mechanism. North Am Acad Res (NAAR) 2021;4(4):268-281. [doi: 10.5281/zenodo.4712849]
58. Worthington A, Coffey T, Gillies K, Roy R, Braakhuis A. Exploring how researchers consider nutrition trial design and participant adherence: a theory-based analysis. Front Nutr 2024;11:1457708. [doi: 10.3389/fnut.2024.1457708] [Medline: 39742103]
59. Mattei J, Alfonso C. Strategies for healthy eating promotion and behavioral change perceived as effective by nutrition professionals: a mixed-methods study. Front Nutr 2020;7:114. [doi: 10.3389/fnut.2020.00114] [Medline: 32923451]
60. Holbrook AM, Pereira JA, Labiris R, et al. Systematic overview of warfarin and its drug and food interactions. Arch Intern Med 2005 May 23;165(10):1095-1106. [doi: 10.1001/archinte.165.10.1095] [Medline: 15911722]
61. Palmer BF. Managing hyperkalemia caused by inhibitors of the renin-angiotensin-aldosterone system. N Engl J Med 2004 Aug 5;351(6):585-592. [doi: 10.1056/NEJMra035279] [Medline: 15295051]
62. Bailey CJ, Turner RC. Metformin. N Engl J Med 1996 Feb 29;334(9):574-579. [doi: 10.1056/NEJM199602293340906] [Medline: 8569826]

XSL•FO
RenderX

63. Joy J, Pillai RVG. Review and classification of content recommenders in E-learning environment. J King Saud Univ, Comput Inf Sci 2022 Oct;34(9):7670-7685. [doi: 10.1016/j.jksuci.2021.06.009]
64. Ayemowa MO, Ibrahim R, Bena YA. A systematic review of the literature on deep learning approaches for cross-domain recommender systems. Decision Analytics Journal 2024 Dec;13:100518. [doi: 10.1016/j.dajour.2024.100518]
65. Abhari S. A systematic review of nutrition recommendation systems: with focus on technical aspects. J Biomed Phys Eng 2019 Dec;9(6). [doi: 10.31661/JBPE.V0I0.1248]
66. Arakaki M, Li L, Kaneko T, et al. Personalized nutritional therapy based on blood data analysis for malaise patients. Nutrients 2021 Oct 18;13(10):3641. [doi: 10.3390/nu13103641] [Medline: 34684641]
67. Shim JS, Oh K, Kim HC. Dietary assessment methods in epidemiologic studies. Epidemiol Health 2014;36:e2014009. [doi: 10.4178/epih/e2014009] [Medline: 25078382]
68. Zangerle E, Bauer C. Evaluating recommender systems: survey and framework. ACM Comput Surv 2023 Aug 31;55(8):1-38. [doi: 10.1145/3556536]
69. Zhu Y, Lyu L. Review of evaluation metrics for recommender systems. J Univ Electron Sci Technol China 2012;41(2):163-175. [doi: 10.3969/j.issn.1001-0548.2012.02.001]
70. Wang J, Dong X, Jin T, Yuan C. Research progress and prospects of health recommendation systems. J Med Intell 2024;45(1):70-76 [FREE Full text]
71. Rossetti M, Stella F, Zanker M. Contrasting offline and online results when evaluating recommendation algorithms. Presented at: RecSys '16; Sep 15-19, 2016. [doi: 10.1145/2959100.2959176]
72. Kohavi R, Longbotham R. Online controlled experiments and a/b testing. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning and Data Mining: Springer; 2017:922-929. [doi: 10.1007/978-1-4899-7687-1_891]
73. Georgiev GZ. Statistical Methods in Online A/B Testing: Statistics for Data-Driven Business Decisions and Risk Management in e-Commerce: Georgi Z: Georgiev; 2019.
74. Sun Y, Zhou J, Ji M, Pei L, Wang Z. Development and evaluation of health recommender systems: systematic scoping review and evidence mapping. J Med Internet Res 2023 Jan 19;25:e38184. [doi: 10.2196/38184] [Medline: 36656630]
75. Ananthakrishnan A, Milne-Ives M, Cong C, et al. The evaluation of health recommender systems: a scoping review. Int J Med Inform 2025 Mar;195:105697. [doi: 10.1016/j.ijmedinf.2024.105697] [Medline: 39608231]
76. Xhani D, Sedrakyan G, Gavai A, Guizzardi R, van Hillegersberg J. The role and applications of semantic interoperability tools and eXplainable AI in the development of smart food systems: findings from a systematic literature review. Intell Syst Appl 2025 Sep;27:200547. [doi: 10.1016/j.iswa.2025.200547]
77. Kalu KA, Ataguba G, Onifade O, Orji F, Giweli N, Orji R. Application of artificial intelligence technologies as an intervention for promoting healthy eating and nutrition in older adults: a systematic literature review. Nutrients 2025 Oct 14;17(20):3223. [doi: 10.3390/nu17203223] [Medline: 41156474]
78. Tsolakidis D, Gymnopoulos LP, Dimitropoulos K. Artificial intelligence and machine learning technologies for personalized nutrition: a review. Informatics (MDPI) 2024;11(3):62. [doi: 10.3390/informatics11030062]
79. Anderson HL, Moore JE, Millar BC. Comparison of innovative communication approaches in nutrition to promote and improve health literacy. Ulster Med J 2022 May;91(2):85-91. [Medline: 35722219]
80. Locke EA, Bandura A. Social foundations of thought and action: a social-cognitive view. Acad Manag Rev 1987 Jan;12(1):169. [doi: 10.2307/258004]
81. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
82. Fogg BJ. Persuasive technology: using computers to change what we think and do. Ubiquity 2002;2002. [doi: 10.1145/764008.763957]
83. Wu X, Xin X. The application of gamification design in persuading users' behavior change [Article in Chinese]. Packag Eng 2017;38(20):194-198 [FREE Full text]
84. Cole LJ, Fleisher LD. Update on HIPAA privacy: are you ready? Genet Med 2003;5(3):183-186. [doi: 10.1097/01.GIM.0000068625.72823.86] [Medline: 12792427]
85. Voigt P, von dem Bussche A. The EU General Data Protection Regulation (GDPR): A Practical Guide: Springer Publishing Company, Incorporated; 2017. [doi: 10.1007/978-3-319-57959-7]

Abbreviations

CB: content-based
CF: collaborative filtering
CoFID: composition of foods integrated dataset
EHR: electronic health record
GDPR: General Data Protection Regulation
HIPAA: Health Insurance Portability and Accountability Act
HRS: Health Recommender System
HyR: hybrid recommendation

XSL•FO
RenderX

JPALD: Japan Preventive Association of Lifestyle-related Disease
KB: knowledge-based
KG: knowledge graph
MOPH: Ministry of Public Health
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RS: Recommender System

Edited by N Cahill; submitted 19.May.2025; peer-reviewed by A Berihun, C Okolue, T Ojo; accepted 01.Dec.2025; published 14.Jan.2026.

Please cite as:
Dong X, Yun B, Pakarinen A, Zheng Z, Niu H, Jin T, Yuan C, Wang J
Diet-Related Health Recommender Systems for Patients With Chronic Health Conditions: Scoping Review
J Med Internet Res 2026;28:e77726
URL: https://www.jmir.org/2026/1/e77726
doi:10.2196/77726

© Xiaolan Dong, Bei Yun, Anni Pakarinen, Zhuting Zheng, Hao Niu, Tian Jin, Changrong Yuan, Jingting Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI-Supported Digital Microscopy Diagnostics in Primary Health Care Laboratories: Scoping Review

Joar von Bahr[1,2,3], MD; Antti Suutala[3], MSc; Vinod Diwan[1], MD, PhD; Andreas Mårtensson[2,4], MD, PhD; Johan Lundin[1,3*], MD, PhD; Nina Linder[2,3*], MD, PhD

[1]Department of Global Public Health, Karolinska Institutet, Tomtebodavägen 18 A, Solna, Sweden

[2]Global Health and Migration Unit, Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden

[3]Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland

[4]Department of Infectious Diseases, Uppsala University Hospital, Uppsala, Sweden

[*]these authors contributed equally

Corresponding Author:
Joar von Bahr, MD
Department of Global Public Health, Karolinska Institutet, Tomtebodavägen 18 A, Solna, Sweden

Abstract

Background: Digital microscopy combined with artificial intelligence (AI) is increasingly being implemented in health care, predominantly in advanced laboratory settings. However, AI-supported digital microscopy could be especially advantageous in primary health care settings, since such methods could improve access to diagnostics via automation and a decreased need for experts on-site. To our knowledge, no scoping or systematic review has previously examined the use of AI-supported digital microscopy in primary health care laboratories, and a scoping review could guide future research by providing insights into the challenges of implementing these novel methods.

Objective: This scoping review aimed to map published peer-reviewed studies on AI-supported digital microscopy in primary health care laboratories to generate an overview of the subject.

Methods: A systematic search of the databases PubMed, Web of Science, Embase, and IEEE was conducted on October 2, 2024. The inclusion criteria in the scoping review were based on 3 concepts: using digital microscopy, AI, and comparison of the results with a standard diagnostic system, and 1 context, being performed in primary health care laboratories. Additional inclusion criteria were peer-reviewed diagnostic accuracy studies published in English, performed on humans and achieving a sample-level diagnosis. The study selection and data extraction were performed by 2 independent researchers (JVB and AS), and cases of disagreement were resolved through discussion with a third researcher (NL). The methodology is in accordance with the Joanna Briggs Institute methodology for scoping reviews.

Results: A total of 3403 papers were screened during the paper identification process, of which 22 (0.6%) were included in the scoping review. The samples analyzed were as follows: blood (n=12) for blood cell and malaria detection, urine (n=4) for urinalysis and parasite detection, cytology of atypical oral (n=1) and cervical cells (n=2), stool (n=2) for parasite detection, and sputum (n=1) for ferning patterns indicating inflammation. Both conventional (n=15) and specifically developed methods (n=7) were used in sample preparation. The AI-supported digital microscopy achieved comparable diagnostic accuracy to the reference standard for complete blood counts, malaria detection, identification of stool and genitourinary parasites, screening for oral and cervical cellular atypia, detection of pulmonary inflammation, and urinalysis. Furthermore, AI-supported digital microscopy achieved higher sensitivity than manual microscopy in 6/7 (85.7%) studies that used a reference standard that allowed for this comparison.

Conclusions: AI-supported digital microscopy achieved comparable diagnostic accuracy to the reference standard for diagnosing multiple targets in primary health care laboratories and may be particularly advantageous for improving diagnostic sensitivity. With further research addressing challenges such as scalability and cost-effectiveness, AI-supported digital microscopy could improve access to diagnostics, especially in expert-scarce and resource-limited settings.

International Registered Report Identifier (IRRID): RR2-10.2196/58149

(J Med Internet Res 2026;28:e78500) doi:10.2196/78500

KEYWORDS

AI; artificial intelligence; convolutional neural network; deep learning; diagnosis; digital diagnostics; machine learning; pathology; primary health care; whole slide images

XSL•FO
RenderX

Introduction

Background

Artificial intelligence (AI) in the form of machine learning has successfully been applied to image-based diagnostics within several medical fields [1]. In parallel, manual microscopy remains a cornerstone of diagnostic practice in resource-limited settings and at the primary health care (PHC) level due to its low cost, versatility, and ability to provide direct visualization of pathogens and cellular changes. It is widely used for the diagnosis of infectious diseases such as malaria and intestinal parasitic infections, as well as for full blood counts and analysis of cervical and oral cytological samples and fine needle aspirates [2]. Despite its usefulness and broad applicability, microscopy is highly dependent on the availability of trained personnel and adequate infrastructure, which are often limited in such settings, leading to variability in diagnostic quality and coverage [3]. These limitations have motivated the development of AI-driven approaches, where deep learning methods can assist or automate microscopy-based diagnostics to improve accuracy and accessibility. Deep learning approaches, particularly convolutional neural networks (CNNs) and vision transformers, have become the dominant architectures for image classification and interpretation in medical imaging [4]. CNNs extract visual features, enabling recognition of complex structures such as cells, pathogens, and tissue patterns, while vision transformers can capture contextual relationships between distant structures [4,5].

Leveraging these methods for AI-based microscopy within laboratory workflows has the potential to automate processes, increase productivity, and improve diagnostic accuracy [6]. Multiple AI-based diagnostic systems have been approved for clinical use, for example, for cervical cancer screening and prostate cancer diagnostics [6-8]. Most of these AI-based diagnostic systems depend on expensive, high-end digital imaging instruments and require advanced laboratory infrastructure and are therefore not feasible for use in PHC laboratories [6,7]. However, the development of less expensive, portable digital microscope scanners has enabled research on the use of AI-supported diagnostic systems suitable for PHC laboratories [9-11].

A PHC laboratory, also known as a tier 1 laboratory, can be defined as a laboratory primarily serving outpatients by providing point-of-care (POC) tests and manual microscopy of specimens with simple preparations. An additional responsibility is preparing fine needle aspirations and other simple tissue specimens that are later dispatched to a tier 2 laboratory in a first-level hospital for analysis. The PHC laboratories work with a small budget compared with more advanced laboratories and are generally managed by a laboratory technician supervised by a pathologist from a distance [2].

The World Health Organization has emphasized the importance of providing diagnostics near the patient to enhance the accuracy and timeliness of diagnoses, improve clinical decision-making, and reduce the risk of diagnostic errors [12]. The implementation of AI-supported digital microscopy could help address these challenges at PHC laboratories. To begin with, since PHC laboratories lack access to pathology expertise, application of AI could enable more analyses on-site, consequently increasing both the availability and speed of diagnostics [2,13]. Increased speed and access to diagnostics through AI and telemedicine could reduce health inequities by strengthening diagnostic capacity, particularly in low- and middle-income countries (LMICs) and also in sparsely populated regions of high-income countries [11,14,15]. In addition, a systematic review showed that the implementation of AI-supported diagnostics for microscopy increased the effectiveness of laboratory personnel [6]. Although there is a global shortage of microscopy experts, the shortage of these specialists is more severe in LMICs; therefore, AI-supported digital microscopy may be especially advantageous in strengthening health systems and reducing the diagnostic gaps in these settings [11,16].

There are several diseases where AI-supported digital microscopy diagnostics in PHC laboratories could be advantageous, and studies have been performed on, for example, screening of oral and cervical cancer as well as targeting parasitic infections, such as schistosomiasis and infections caused by soil-transmitted helminths [9,17-19]. Although the targeted diseases differed in these studies, researchers often encountered similar challenges due to commonalities in the methodologies applied, and a review mapping these challenges could provide valuable insights.

A preliminary search of the databases PubMed and Cochrane was performed to investigate whether any scoping or systematic review had been performed on AI-supported digital microscopy in PHC laboratories. A few related reviews were found. One systematic review of AI diagnostics for oral cancer [20] overlaps to some extent with our review; however, since it focuses on a single disease, it does not provide an overview of the development of AI-supported digital microscopy in PHC laboratories. Another systematic review evaluating the application of AI to whole slide images of tissue samples stained with hematoxylin and eosin was also identified [21]. This paper presents the current state of knowledge on AI implementation in pathology within high-end laboratories.

While these reviews are similar to this scoping review, they do not provide an overview of which diseases have been investigated in AI-supported digital microscopy and the disease-agnostic challenges faced in PHC laboratories. Furthermore, the development of more affordable scanners and improved AI, along with persistent workforce and resource constraints, makes a scoping review timely. A scoping review performed on AI-supported digital microscopy in PHC laboratories would, therefore, provide a valuable overview of the subject and collate knowledge that could guide future implementation.

This scoping review aimed to systematically review published peer-reviewed studies that have been performed related to AI-supported digital microscopy in PHC laboratories and specifically address the following questions: (1) In which diseases and for which conditions and targets has AI-based microscopy been applied for diagnostics within PHC laboratories? (2) What methods have been used in acquiring microscopy images to train and analyze AI models for

XSL•FO
RenderX

diagnostics? (3) What AI models and training approaches have been applied? (4) How has the AI-supported diagnostic system performed compared with expert microscopists with regard to diagnostic accuracy?

Review Question

What peer-reviewed studies have been published on implementing AI-supported digital microscopy in PHC laboratories? What methods have been used, what issues have been faced, and what results have been achieved?

Methods

Study Design

The scoping review was conducted in accordance with the Joanna Briggs Institute methodology for scoping reviews updated in 2020 [22]. A PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) checklist is included [23]. A protocol was initially published in the Open Science Framework and later in the peer-reviewed journal JMIR Research Protocols [24,25]. The inclusion and exclusion criteria are shown in Table 1.

Table . Inclusion and exclusion criteria for identified studies.

Study characteristic	Inclusion criteria	Exclusion criteria
Language	• English	• Non-English
Study design	• Published peer-reviewed studies • Diagnostic test accuracy studies	• Non–peer reviewed studies • Not diagnostic test accuracy studies
Population	• Humans	• Studies performed on animals
Concept	• AI[a] techniques applied as a diagnostic tool on microscopy • Final slide-level diagnosis was performed and compared with a standard microscopist • Outcome valuable for clinicians	• Studies that applied AI models on images not conventionally analyzed in microscopy • No final slide diagnosis
Context	• Performed at primary health care laboratory (tier 1 laboratory) • No pathologist needed on site • Samples such as stool, urine, blood, cytology smears, and fine needle aspirations of superficial tissue (eg, from breast lumps) prepared with simple methods	• Studies performed in an advanced laboratory setting

[a]AI: artificial intelligence.

Eligibility Criteria

Participants

This scoping review considered studies on human participants. No exclusion was made based on age, sex, economic status, or nationality.

Concept

The studies included in this scoping review fulfilled 3 concept criteria. First, the studies needed to be performed on images obtained with an imaging instrument built to automatically capture microscopy sample areas large enough for diagnostic purposes. Furthermore, the imaging instrument used must be operated in a way that does not require human expertise to determine what areas of the slide should be captured. Microscopy was defined as deploying a light source, optical lenses, and a digital camera to acquire a magnified image of a biological sample, generating an image conventionally interpreted by a microscopist.

Second, the studies needed to use AI when analyzing the microscopy images. AI was defined as a computer system that is trained to perform a task that typically requires human intelligence. No exclusion was made based on the architecture of the AI model or the dataset used for training. This analysis of the microscopy images could be performed on-site or in a remote cloud environment.

Third, the studies needed to compare the AI-supported diagnostic system with a standard diagnostic system. A diagnostic system was defined as all the steps included in the diagnostic process, from sample collection to the acquisition of results. The result needed to be sufficient to reach a diagnosis at the subject level.

Context

The included studies needed to be performed in a PHC laboratory setting. To be defined as a PHC laboratory, also known as a tier 1 laboratory, the laboratory needed to fulfill 2 criteria. First, regarding staffing, the laboratory must be run by a laboratory technician, not requiring a pathologist on-site. Second, the sample preparations could not exceed the capabilities of a PHC laboratory. Acceptable samples collected included stool, urine, blood, cytology smears, and fine needle aspirations of superficial and easily accessible tissues (eg, from breast lumps and superficial lymph nodes). The sample staining procedure must be possible to perform manually without advanced laboratory equipment such as a microtome or tissue

XSL•FO
RenderX

processor [2]. Sample procedures that fulfill these criteria include Kato-Katz thick stool smears, blood smears, centrifuged urine samples, Papanicolaou-stained cervical or oral smears, and hematoxylin and eosin–stained fine needle cytology smears [2]. Since the context of PHC laboratories in this scoping review is based on human medicine, the exclusion criteria and initial search strategy were changed to exclude veterinary medicine, which was included in the initial protocol published on Open Science Framework [24]. This adjustment was made before submitting the protocol to JMIR Research Protocols to focus the scoping review specifically on challenges in implementing AI-supported microscopy in human health care [25].

Types of Sources

All types of diagnostic test accuracy studies were included. Because data collection in diagnostic test accuracy studies can be both retrospective and prospective, studies using either approach were included. In addition, studies using both paired and random designs for reference standards were included [26]. The included studies had to be published in English.

Search Strategy

The search strategy was designed to identify peer-reviewed published papers. An initial limited search of PubMed and Cochrane was undertaken to identify papers on the topic. Search blocks were created for the final search based on terms used in the identified papers. The search blocks were developed to find papers containing the 2 concepts, microscopy and AI, as well as the context specification of being in a PHC setting, with 1 block created for each. The databases searched were PubMed, Web of Science, Embase, and IEEE, and a detailed description of the search strategy is given in Multimedia Appendix 1. The search was performed on October 2, 2024. The reference lists and all the papers citing the included papers were gathered through the SpiderCite tool on December 3, 2024, and included in the review process [27].

Study Selection

Following the search, all identified papers were compiled in a reference management software system, Zotero (version 6.0.20, Digital Scholar; January 13, 2023, opensource) and duplicates removed. Following the pilot test, titles and abstracts were screened by 2 independent reviewers (JvB and AS) for assessment against the inclusion and exclusion criteria using Covidence systematic review software (Veritas Health Innovation, 2024) [28]. During this step, the Cohen κ agreement was 0.59. All disagreements between JvB and AS were resolved by NL, who provided the deciding vote and could consult the other screeners for their rationale. Thereafter, the full texts of the remaining papers were assessed in detail against the inclusion criteria by 2 independent reviewers (JvB and AS). During full text screening, the Cohen κ agreement was 0.75 for the database search and 0.36 for the citation search. Two reasons caused 17 out of 21 disagreements in the citation search and were resolved through discussions between JvB, AS, JL, and NL. The first issue concerned whether urine analyzers such as Iris iQ200 or Sysmex UF-100 fulfilled the PHC criteria: it was concluded that they did not, as these devices perform advanced sample preprocessing within the machine [29]. The second issue concerned whether handcrafted feature classification qualified as AI: it was concluded that it did not, as it does not involve AI training. With these 2 issues resolved, the citation search had a Cohen κ agreement of 0.79.

Data Charting and Synthesis

Data were extracted from the studies included in the scoping review by 2 reviewers (JvB and AS) using a data extraction tool developed with Covidence systematic review software. The predeveloped extraction tool can be found in Multimedia Appendix 2. Initially, the extraction was performed by JvB. Afterward, the extracted information was checked by AS. All disagreements were resolved through discussion between JvB and AS. When questions arose regarding an original paper, the corresponding author of that manuscript was contacted. The findings are presented narratively and additionally in a table format based on the extraction tool. The information from the extraction tool was split into 3 tables and 1 figure to increase readability. The figure contains a simplified overview of the studies, the first table summarizes the process from sample collection to scanning, the second table shows information on the AI analysis pipeline and training data, and the third table reports the study outcomes. Based on the information extracted to the tables, a narrative description was written to provide an overview of the mapped information. The studies were grouped based on the sample type investigated and the disease targeted as per the first objective of the study: to map in which diseases and for which conditions and targets has AI-based microscopy been applied for diagnostics within PHC laboratories.

Critical Appraisal of Results

The QUADAS-2 tool was applied to investigate the bias of the included studies. This tool was developed to assess the risk of bias for diagnostic accuracy studies in 4 areas: patient selection, index test, reference standard, and flow and timing [30]. The results are shown in the “Results” section, and the form used can be found in Multimedia Appendix 3.

Results

Overview

In total, 3403 papers were screened during the paper identification process, of which 22 (0.6%) were included in the scoping review. The results of the search and the study inclusion process are reported in full in a PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) flow diagram (Figure 1) [31].

XSL•FO
RenderX

Figure 1. Flowchart for study inclusion.

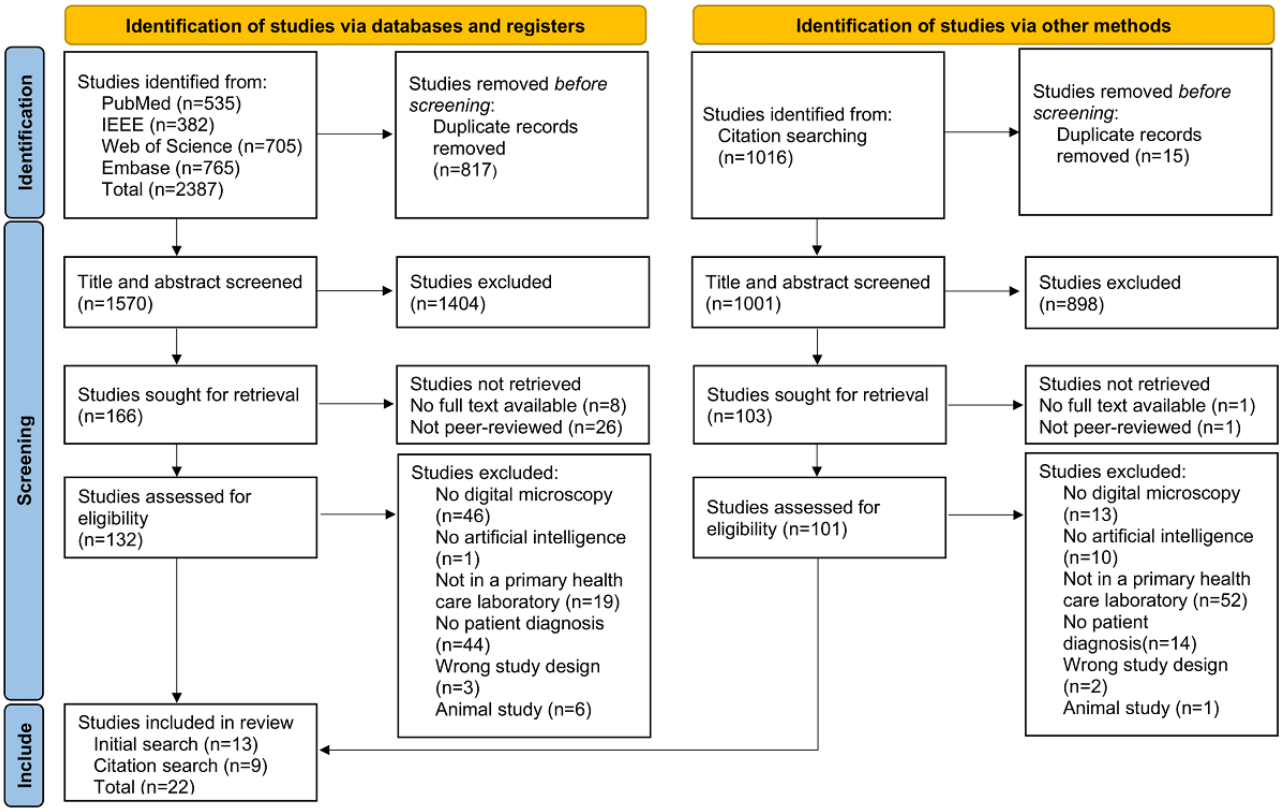

The oldest included study was published in 2014, while the remaining studies were published in 2018 or later, with 9 out of the 22 (40.9%) studies published in 2024. The papers were published in 15 different journals with the most common being *Malaria Journal* (n=4) and *PLOS One* (n=4). The most analyzed samples were blood (n=12), followed by urine (n=4), cytology (n=3), stool (n=2), and sputum (n=1). Different parasites (malaria, intestinal, or genitourinary parasites) were the most common targets (n=13), followed by blood cells (n=4), atypical cervical cells (n=2), atypical oral cells (n=1), and urine particles, such as cells (n=1) and crystalline ferning patterns in sputum (n=1). Detection of these targets was used for multiple diseases and conditions; complete blood counts (CBCs) and urinalysis were used for both organ-specific and systemic diseases, parasite detection for corresponding infections, atypical cells for screening and detection of cancer, and ferning patterns for identifying pulmonary inflammation in patients with COVID-19. An overview of all studies is shown in Figure 2.

XSL•FO
RenderX

Figure 2. Overview of the included studies. AI better than microscopy: White = No comparison, Yes = Higher/same sensitivity and specificity, Mix = Higher sensitivity and lower specificity, and No = Lower sensitivity and specificity. Number of samples in the test set: k=1000. Conditional formatting was applied to all numerical values, with high values shaded green and low values shaded yellow. AI: artificial intelligence; CBC: complete blood count; D: Downey cells; F: Ferning patterns indicative of inflammation; S: sputum; U: urinalysis.

Study Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Samples	Blood												Cytology			S	Stool		Urine			
Targets	CBC			D	Malaria parasite								Cell atypia			F	Parasite		U	Parasite		
Conventional prep.				✓		✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
Custom-built scanner						✓										✓	✓		✓	✓	✓	✓
Single AI algorithm		✓	✓	✓		✓					✓			✓		✓			✓	✓	✓	✓
Human verification		✓	✓		✓			✓			✓											
Reference: microscopy				✓			✓		✓		✓	✓	✓	✓				✓		✓	✓	
Sensitivity %				57	90		95	83	91	72	95	87		100	89	94	86	80	81	87	94	63
Specificity %				100	96		91	94	76	85	100	100		78	100	96	100	95	88	49	94	79
Pearson correlation	0.94	0.94	0.91			0.90																
AI better than microscopy					Mix			Mix		Mix					Yes		Yes		No			Mix
No. samples in test set	679	550	450	31	190	27	488	1.2k	2.3k	400	31	35	307	361	30	160	73	792	240	487	65	339

Sample Preparation and Scanning

Out of the 22 included studies, 12 relied solely on manual preparation methods and 3 used centrifuges. The remaining studies (n=7) used cartridges that simplified and eliminated manual steps. Both in-house–built and commercially available scanners such as Grundium, MiLab, and Motic EasyScan GO were used. The lowest numerical aperture used was 0.1 and the highest was 1.4. Several scanners used both autofocus algorithms and z-stacking to avoid out-of-focus areas. In the 5 studies reporting the time from sample collection to diagnosis using AI-supported digital microscopy, it was 20 - 40 minutes, but there was also a study reporting that it took more than 50 minutes for the scanning and AI analysis (Table 2) and a more detailed table in Multimedia Appendix 4.

XSL•FO
RenderX

Table . Time for analysis and sample processing for the included studies.

Study	Sample	Target	Sample preparation	Sample scanning	Time for analysis
Bachar et al (2021) [32]	Blood	CBC[a]	Cartridge with 2 stains	No retrievable magnification and resolution	No retrievable information
Gasparin et al (2023) [33]	Blood	CBC	Dual-chamber cartridge with 2 stains	No retrievable magnification and resolution	Total: 30 - 40 minutes
Gasparin et al (2022) [34]	Blood	CBC	Dual-chamber cartridge with 2 stains	No retrievable magnification and resolution	Total: 30 - 40 minutes
Akisin et al (2023) [35]	Blood	Downey cells	Manual blood smears stained with May-Grünwald and Giemsa	100× with oil immersion	No retrievable information
Hamid et al (2024) [36]	Blood	Malaria parasites	Cartridge with Giemsa staining	A resolution similar to 50× microscopy [37]	Total: <30 minutes
Holmström et al (2020) [38]	Blood	Malaria parasites	Manual blood smears stained with DAPI[b]	A resolution of 0.9 µm	No retrievable information
Bae et al (2024) [37]	Blood	Malaria parasites	Cartridge with Giemsa staining	A resolution similar to 50×	Total: <30 minutes [36,39]; Scanning: 7 - 10 minutes
Ewnetu et al (2024) [39]	Blood	Malaria parasites	Cartridge with Giemsa staining	A resolution similar to 50× [37]	Total: circa 20 minutes
Das et al (2022) [40]	Blood	Malaria parasites	Manual blood smears stained with Giemsa	40× (NA[c] 0.75)	Scanning and AI[d] analysis: 20 - 30 minutes
Torres et al (2018) [41]	Blood	Malaria parasites	Manual blood smears stained with Giemsa	100× with oil immersion (NA 1.25)	No retrievable information
Linder et al (2014) [42]	Blood	Malaria parasites	Thin blood smears stained with Giemsa	63× with oil immersion (NA 1.4)	No retrievable information
Horning et al (2021) [43]	Blood	Malaria parasites	Manual blood smears stained with Giemsa	40x (NA 0.75)	Scanning and AI analysis 54 minutes
Stegmüller et al (2024) [44]	Cervical cytology	Cellular atypia	SurePath procedure with Papanicolaou stain	40× (NA 0.75)	No retrievable information
Holmström et al (2021) [9]	Cervical cytology	Cellular atypia	Conventional Papanicolaou smears	20× (NA 0.4)	Scanning: 5 - 10 minutes; uploading 10 - 40 minutes
Sunny et al (2019) [19]	Oral cytology	Cellular atypia	Manual liquid-based cytology, stained with H&E[e] [45]	20× (NA 0.4) [45]	AI analysis: 10 minutes
Ghaderinia et al (2024) [46]	Sputum	Ferning patterns	Sedimented unstained sputum samples	40× magnification	No retrievable information
Soares et al (2024) [47]	Stool	Intestinal parasites	Fecal samples with centrifugation, flotation, and sedimentation [48]	No retrievable magnification and resolution	AI analysis: circa 3 minutes
Lundin et al (2024) [49]	Stool	Soil-transmitted helminths	Kato-Katz thick smears	20× (NA 0.4)	Scanning 5 - 10; uploading 10 - 20 minutes; AI analysis 5 minutes
Sahu et al (2024) [50]	Urine	Urinalysis	Cartridge that concentrates the urine through 5 minutes of sedimentation	40× (NA 0.65)	No retrievable information
Meulah et al (2022) [51]	Urine	Schistosoma	A membrane capturing filtered urine particles	4× (NA 0.1)	Scanning: 12 minutes; AI analysis: 5 minutes

XSL•FO
RenderX

Study	Sample	Target	Sample preparation	Sample scanning	Time for analysis
Oyibo et al (2022) [52]	Urine	Schistosoma	A membrane capturing filtered urine particles	4× (NA 0.1)	Scanning: 12 minutes; AI analysis: 10 - 12 minutes
Meulah et al (2024) [53]	Urine	Schistosoma	A membrane capturing filtered urine particles	4× (NA 0.1)	Scanning and AI analysis: 25 minutes

[a]CBC: complete blood count.

[b]DAPI: 4',6-diamidino-2-phenylindole.

[c]NA: numerical aperture.

[d]AI: artificial intelligence.

[e]H&E: hematoxylin and eosin.

Training Data and AI Analysis Pipeline

For training AI models, most studies used in-house collected and annotated datasets of varying sizes; some had hundreds of target objects in their dataset, whereas others had hundreds of thousands. Many studies reported using pretrained neural networks with different datasets such as COCOtrain2017 and ImageNet for training [47,53]. One study used unlabeled data from their collection for unsupervised pretraining and incorporated publicly available datasets [44].

The AI analysis pipeline for all included studies can be summarized as follows: a digitized microscopy sample was provided as input, fields-of-view (FOVs) were analyzed, FOV results were aggregated to produce a slide-level diagnosis, and this diagnosis served as the output (Figure 3). The digitized sample used as input could consist of either whole-slide images or multiple FOVs captured from the physical slide. The FOV analysis involved both the identification and the classification of specific targets; however, not all studies used this first identification of suspicious FOVs (regions of interest).

Figure 3. Visualization of the artificial intelligence analysis pipeline. As an illustrative case, the pipeline is applied to a digitized fecal smear with *Ascaris lumbricoides* parasite eggs.

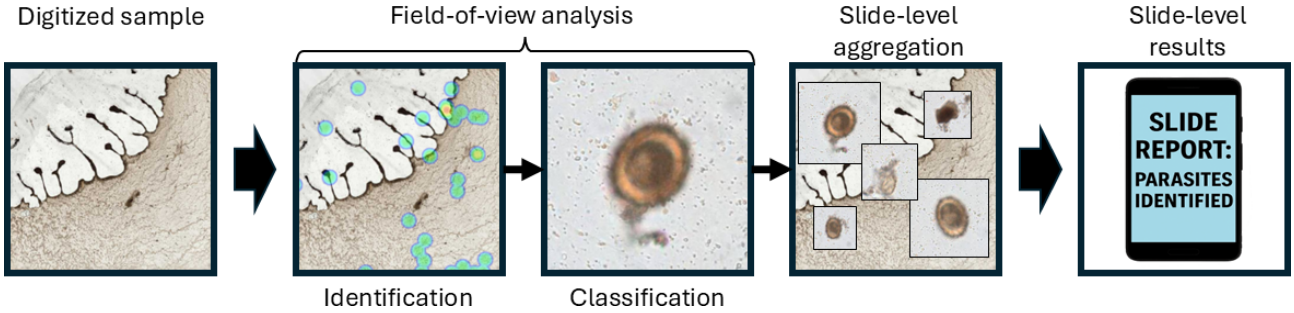

An initial detection of suspicious FOVs was described in half of the studies; when this was applied, algorithmic approaches, shallow CNNs, or support vector machines (SVMs) were used. The purpose of performing this initial identification of FOVs of interest was to reduce the number of FOVs that needed to be analyzed by more computationally expensive AI algorithms. An additional advantage of performing an initial detection of suspicious FOVs was that the FOVs were more homogeneous in content and quality, which can improve the accuracy of the AI classifier. All studies used AI for the FOV classification step, predominantly CNNs, except for the oldest study that used an SVM [42]. One approach was to use multiple classification steps, for example, by using SVM or shallow CNNs to classify targets and then reclassifying those with higher uncertainty with deeper CNNs [37,47].

To achieve the slide-level diagnosis from the FOV analysis results, multiple methods were used; for example, classifying slides with any number of positive targets as positive, using different cutoffs based on confidence or number of findings, or using AI-based methods such as SVMs and multiple instance learning (Table 3).

XSL•FO
RenderX

Table . Artificial intelligence model training and architecture for the included studies.

Study	Sample and target	Samples in training set	AI[a] model architecture and training
Bachar et al (2021) [32]	Blood and CBC[b]	No retrievable information	AI model with separate pipelines for platelets, RBCs[c], and WBCs[d] (1) algorithmically identifies candidates, and (2) candidates categorized by specialized CNNs[e] and machine learning algorithms
Gasparin et al (2023) [33]	Blood and CBC	Expert-verified training data gathered throughout development; no further retrievable information	AI model with CNN architecture using the YOLO[f] framework
Gasparin et al (2022) [34]	Blood and CBC	Expert-verified training data gathered throughout development; no further retrievable information [33]	AI model with CNN architecture using the YOLO framework
Akisin et al (2023) [35]	Blood and Downey cells	15,885 expert-annotated WBCs containing 172 Downey cells	AI model with YOLOv4-tiny-based framework with spatial attention using average and maximum pooling along the channel axis
Hamid et al (2024) [36]	Blood and malaria parasites	No retrievable information	AI model with (1) U-Net segmenting RBCs, (2) a 3-layer CNN removing normal RBCs, (3) a 23-layer CNN for detecting parasites, and (4) 1 positive object sufficient for slide positivity [37]
Holmström et al (2020) [38]	Blood and malaria parasites	25 thin blood smears with annotated trophozoites (n=5059) and other fluorescence signals (n=856)	AI model with (1) Circle Hough Transform identifying RBCs, (2) fluorescence signals from within the detected RBCs are used, and (3) RBCs with fluorescence signals were analyzed with a CNN (GoogLeNet)
Bae et al (2024) [37]	Blood and malaria parasites	No retrievable information	AI model with (1) U-Net segmenting RBCs, (2) a 3-layer CNN removing normal RBCs, (3) a 23-layer CNN for detecting parasites, and (4) 1 positive object sufficient for slide positivity
Ewnetu et al (2024) [39]	Blood and malaria parasites	No retrievable information	AI model with (1) U-Net segmenting RBCs, (2) a 3-layer CNN removing normal RBCs, (3) a 23-layer CNN for detecting parasites, and (4) 1 positive object sufficient for slide positivity [37]
Das et al (2022) [40]	Blood and malaria parasites	Subset of 1452 blood samples and 956,531 annotated parasite objects [54]	AI model analyzes only thick region with (1) potential parasites identified through dynamic thresholding and SVM[g], (2) CNN (VGG[h] architecture) classifies parasites, and (3) a predetermined threshold decides slide positivity [54]
Torres et al (2018) [41]	Blood and malaria parasites	Approximately 150 high-quality thick films with 75,000 parasites	AI model for thick region with (1) local thresholding and low-cost methods to identify potential parasites, (2) CNNs (VGG architecture) classifies parasites and stage, and (3) number and confidence of parasites determine slide-level diagnosis
Linder et al (2014) [42]	Blood and malaria parasites	A training set (n=10) with parasites (n=8329) and a validation set (n=6) parasites (n=569)	AI model with (1) thresholding algorithm segments potential parasites, and (2) mathematical feature extraction and classification with SVM

XSL•FO
RenderX

Study	Sample and target	Samples in training set	AI[a] model architecture and training
Horning et al (2021) [43]	Blood and malaria parasites	Thick model: Subset of 1452 blood samples with 956,531 parasite objects [54]. Thin model: 798 blood samples with more than 92,000 parasites [55]. Tuning slides: 48 slides	Separate AI models for thin and thick regions: Thick-AI model: (1) potential parasites identified through dynamic thresholding and SVM. (2) CNNs (VGG architecture) classify [54]. Thin AI model: (1) potential parasites detected with a gradient-boosted tree classifier. (2) CNNs for classifying parasite stages [55]
Stegmüller et al (2024) [44]	Cervical cytology and cellular atypia	A stratified 4-fold split approach to partition the 307 slides with 1228 tile-level annotations into training, validation, and test sets; 2 public datasets also used	AI model with (1) CNN (ResNet-50) with self-supervised training (DINO) and then supervised training with cell pasting, and (2) 8 most suspicious tiles used for slide classification with multiple instance learning (CLAM)
Holmström et al (2021) [9]	Cervical cytology and cellular atypia	350 WSIs[i] were used for training with 16,133 annotations made by a pathologist	AI model with (1) a CNN that segments slide into high- and low-grade atypia, and (2) a threshold that decides slide positivity
Sunny et al (2019) [19]	Oral cytology and cellular atypia	252 atypical and 280 normal cell images annotated (90% for training and 10% for validation)	AI model with (1) cells segmented to single cells, (2) a CNN (Inception V3) used for classification, and (3) cut-offs and SVMs based on percentage and mean score of atypical cells and mean cell score for slide diagnosis
Ghaderinia et al (2024) [46]	Sputum and ferning patterns (inflammation)	650 images (520 training and 130 validation) derived from 70 participants	AI model with (1) a CNN (EfficientNet-B0); and (2) CNN output used to classify sample
Soares et al (2024) [47]	Stool and intestinal parasites (both helminths and protozoans)	51,919 images containing 12,225 annotations of 15 parasite species (ranging from 83 to 3297 per species) [56]	AI model with (1) classification with extracted features and probabilistic SVM, and (2) uncertain objects analyzed with a CNN (Vgg-16) [56]
Lundin et al (2024) [49]	Stool and soil-transmitted helminths	388 samples with 15,058 annotations: *Ascaris lumbricoides* (n=2299), *Trichuris trichiura* (n=2727), hookworm (n=552), and artifacts (n=9480)	AI model with (1) YOLOv2 used to detect potential parasites, (2) a CNN (ResNet50) used for classification, and (3) 1 parasite sufficient for slide positivity
Sahu et al (2024) [50]	Urine and urinalysis	A dataset annotated by a pathologist	AI model with (1) a single CNN (YOLOX) to detect objects, and (2) object counts used to grade slide in tiers of positivity
Meulah et al (2022) [51]	Urine and Schistosoma	Both spiked laboratory samples and 33 field samples [17]	AI model with (1) a CNN segmentation model (U-Net architecture) [17]
Oyibo et al (2022) [52]	Urine and Schistosoma	17,799 annotated *Schistosoma haematobium* eggs in 2997 FOV images; dataset split into 80% training and 20% validation set	AI model with (1) a CNN (DeepLabv3-MobileNetV3), (2) egg-shaped ellipses fitted to segmented regions for counting, and (3) 1 parasite fulfilling criteria sufficient for slide positivity

XSL•FO
RenderX

Study	Sample and target	Samples in training set	AI[a] model architecture and training
Meulah et al (2024) [53]	Urine and Schistosoma	17,799 annotated *S. haematobium* eggs in 2997 FOV[j] images; dataset split into 80% training and 20% validation set [52]	AI model with (1) a CNN (DeepLabv3-MobileNetV3), (2) egg-shaped ellipses fitted to segmented regions for counting, and (3) 1 parasite fulfilling criteria sufficient for slide positivity [52]

[a]AI: artificial intelligence.
[b]CBC: complete blood count.
[c]RBCs: red blood cells.
[d]WBCs: white blood cells.
[e]CNNs: convolutional neural networks.
[f]YOLO: You Only Look Once.
[g]SVM: support vector machine.
[h]VGG: Visual Geometry Group.
[i]WSIs: whole slide images.
[j]FOV: field of view.

Study Outcomes

When the study outcomes were mapped, differences were observed in the reference standards, study sizes, and performance metrics used. For the 3 studies investigating CBCs, the Pearson correlation coefficient was compared with high-end analyzers and was above 0.9 for all cells except basophils, where the value ranged from 0.6 to 0.8 in all studies [32-34]. Nine studies used manual microscopy of the same samples as the reference standard and reported results with sensitivity and specificity (for malaria, soil-transmitted helminths, Schistosoma, and cervical cell atypia). Across these 9 studies, all reported a sensitivity and specificity of at least 80% except for 1 with a lower sensitivity of 57% [35] and 3 with lower specificity (75.6%, 78.4%, and 48.9%) [9,40,51]. One study included results with and without human expert verification: human verification of AI model findings increased specificity by 29.5% but conversely led to a sensitivity decrease of 0.9% for malaria detection [36]. Seven studies used reference standards such as polymerase chain reaction (PCR) or histology and included comparisons between AI-supported digital microscopy and manual microscopy; in 4 of these 7 studies, a higher sensitivity but lower specificity was reported for AI-supported digital microscopy. Of the remaining 3 studies, 1 evaluated urinalysis, in which manual analysis had higher sensitivity, specificity, or both across all targets [50], 1 for intestinal parasites where the AI had higher sensitivity and the same specificity [47], and 1 for oral atypia where the AI had both higher sensitivity and specificity [19]. The number of samples included in the diagnostic evaluations ranged from 27 to 2250. Most studies (n=15) achieved a low risk for bias according to QUADAS-2; however, some studies either lacked the information needed to properly evaluate bias or had methodological issues (n=7) (Table 4).

XSL•FO
RenderX

Table . Results for the included studies.

Study	Sample and target	Human verification	Outcome	Manual microscopy	Number of samples	Reference standard	QUADAS-2[a]
Bachar et al (2021) [32]	Blood and CBC[b]	No	*r*[c]≥0.94 (except basophils=0.6)	NR[d]	679	Hematology analyzer	Low
Gasparin et al (2023) [33]	Blood and CBC	Yes	*r*≥0.94 (except eosinophils/basophils=0.81)	NR	550	Hematology analyzer	Low
Gasparin et al (2022) [34]	Blood and CBC	Yes	*r*≥0.91 (except eosinophils/basophils=0.80)	NR	450	Hematology analyzer	Low
4: Akisin et al (2023) [35]	Blood and Downey cells	No	Se[e] 57%, Sp[f] 100%	NR	31	Manual microscopy	Mostly low
Hamid et al (2024) [36]	Blood and malaria	Yes	Se 90.2%, Sp 96.2%	Se 89.3%, Sp 100%	190	PCR[g]	Low
Holmström et al (2020) [38]	Blood and malaria	No	*r*=0.90 for parasite counts	NR	27	PCR	Mostly low
Bae et al (2024) [37]	Blood and malaria	No	Se 95.1%, Sp 91.4%	NR	488	Microscopy and RDTs[h]	Low
Ewnetu et al (2024) [39]	Blood and malaria	Yes	Se 83% - 93.9%, Sp 94% - 97.6%	Se 67% - 69.9%, Sp 97% - 98.7%	1165	PCR	Low
Das et al (2022) [40]	Blood and malaria	No	Se 91.1%, Sp 75.6%	NR	2250	Microscopy	Low
Torres et al (2018) [41]	Blood and malaria	No	Site 1: Se 72%, Sp 85% Site 2: Se 52%, Sp 70%	Site 1: Se 68%, Sp 100% Site 2: Se 42%, Sp 97%	Site 1: 400 Site 2: 300	PCR	Low
Linder et al (2014) [42]	Blood and malaria	Yes	Se 95%, Sp 100%	NR	31	Microscopy	Low
Horning et al (2021) [43]	Blood and malaria	No	Se 86.7%, Sp 100%	NR	35	Microscopy	Low
Stegmüller et al (2024) [44]	Cervical cytology and cellular atypia	No	Mean area under curve 77.5	NR	307 (4-fold split)	Microscopy	Low
Holmström et al (2021) [9]	Cervical cytology and cellular atypia	No	Se 100%, Sp 78.4%	NR	361	Microscopy	Mostly low
Sunny et al (2019) [19]	Oral cytology and cellular atypia	No	Se 89%, Sp 100%	Se 59%, Sp 67%	30	Histology	Low
Ghaderinia et al (2024) [46]	Sputum and fern-ing patterns (inflammation)	No	Se 94.3%, Sp 95.9%	NR	160	CT[i]	Mostly low
Soares et al (2024) [47]	Stool and intestinal parasites (both helminths and protozoans)	No	Se 86%, Sp 100%	Se 81%, Sp 100%	73	Manual and AI[j] microscopy	Mostly low
Lundin et al (2024) [49]	Stool and soil-transmitted helminths	No	Se 76.4% - 91.9% Sp 89.7% - 98.2%	NR	792	Microscopy	Low
Sahu et al (2024) [50]	Urine and urinalysis	No	Se ≥81% except for bacteria (76%) and casts (71%), Sp ≥88%	Se ≥94% and Sp ≥93%	240	Microscopy	Mostly low

XSL•FO
RenderX

Study	Sample and target	Human verification	Outcome	Manual microscopy	Number of samples	Reference standard	QUADAS-2[a]
Meulah et al (2022) [51]	Urine and Schistosoma	No	Se 87.3%, Sp 48.9%	NR	487	Microscopy	Low
Oyibo et al (2022) [52]	Urine and Schistosoma	No	Se 93.8%, Sp 93.9%	NR	65	Microscopy	Mostly low
Meulah et al (2024) [53]	Urine and Schistosoma	No	Se 62.9%, Sp 78.8 %	Se 61.9%, Sp 96.4%	339	PCR and particle lateral flow test	Low

[a]QUADAS-2: Quality assessment of diagnostic accuracy studies 2.

[b]CBC: complete blood count.

[c]*r*: Pearson correlation coefficient.

[d]NR: Not reported.

[e]Se: sensitivity.

[f]Sp: specificity.

[g]PCR: polymerase chain reaction.

[h]RDT: rapid diagnostic test.

[i]CT: computed tomographic scan.

[j]AI: artificial intelligence.

Discussion

Summary

This scoping review included 22 publications deploying AI-supported digital microscopy in PHC laboratories for multiple targets, published in 15 different journals. These studies fulfilled the concepts of using AI and digital microscopy to achieve a slide-level diagnosis in PHC laboratories. The number of included studies was low, given the extensive research on AI in medical imaging. The exclusion of 58 papers due to the absence of sample-level diagnoses and of 71 papers due to not being conducted in PHC laboratories suggests that most research has focused on target detection or advanced laboratory settings, rather than evaluating end-to-end diagnostic systems for PHC use. This is notable, given the potential benefits of such technologies in PHC laboratories. However, 9 of the 22 included studies were published in 2024 indicating an upward trend in studies focused on AI-supported digital microscopy at the PHC level.

The studies targeting specific diseases primarily focused on conditions that disproportionally affect vulnerable populations. The results from the included studies in this scoping review indicate that AI-supported digital microscopy can achieve accuracy comparable to that of standard microscopy for malaria, intestinal parasites, cell atypia, and urinalysis; to that of computed tomography for detecting pulmonary inflammation in patients with COVID-19; and to that of conventional hematology analyzers for CBC. Diagnostic accuracy comparable to the reference standard was defined as sensitivity and specificity of >80% or a Pearson correlation of >0.90. The reported results also indicate that AI-supported digital microscopy could be particularly advantageous for increasing sensitivity, as 6 out of 7 (85.7%) studies comparing it with manual microscopy reported higher sensitivity for AI-supported digital microscopy. Furthermore, the objective of the scoping review was to map target-agnostic challenges and solutions regarding sample preparation, scanning, AI methods, and human integration and discuss future implications for AI-supported digital microscopy in PHC laboratories.

Sample Preparation

Variability in target morphology and artifacts may reduce AI performance and can be introduced in all steps from sample collection to scanning. Manual steps in sample preparation are prone to introducing variability, and all included studies involved such steps, with 12 relying on entirely manual preparation. Decreased specificity due to sample variability was observed in one study, where poorly prepared samples led to the introduction of artifacts [41], and in another study where synthetically prepared samples in the training dataset lacked artifacts present in real-world samples [51]. Sensitivity can also be affected by variability in preparation, as demonstrated in one study on soil-transmitted helminths [49]. This indicates that variability introduced during sample preparation may be a major hurdle when developing AI-supported digital microscopy for PHC laboratories as more steps are performed manually.

There are possible solutions to sample variability. For example, improving consistency through good laboratory practices and standard operating procedures is one way to minimize sample variability; however, this requires system-specific training for personnel, good laboratory infrastructure, and quality controls, which might reduce the feasibility of implementation in PHC laboratories. The use of equipment such as cartridges to limit manual steps is another approach to minimize variability [33,36,50]. This may lower the demands on personnel; however, using disease-specific consumables may introduce issues, for example, increased costs. Another potential approach to minimize variability is to simplify sample preparation, for example, by removing staining or smearing steps [57]. Although this could reduce variability, it may also lead to a loss of valuable diagnostic information, in turn decreasing the AI model performance.

XSL•FO
RenderX

Scanning

The scanner needs to capture sufficient information to allow AI model classification of targets, but scanning large sample areas at high magnification is time-consuming. The scanning time can be decreased by analyzing a smaller sample area. However, this can lead to a reduced ability to detect low-density targets, highlighting a trade-off between faster diagnostics and high sensitivity for these cases. This is exemplified by one solution for malaria, where clinicians are able to increase the area analyzed to detect low-density infections [39]. Another solution to decrease scanning time is to use a lower magnification than what is conventionally used by microscopists: this approach achieved diagnostic accuracy comparable to manual microscopy for cytology [19], malaria [36], and parasitic infections [49,53]. Nonetheless, the use of lower magnification could result in information loss that reduces the AI model performance.

Training Data

All studies that specified the AI-training methods used variants of supervised learning which require annotated data. Annotating data is time-consuming and requires digitized samples that are rarely produced in PHC laboratories due to limited access to scanners. Therefore, many studies had to collect and annotate their own datasets rather than access existing data. One study also used laboratory-enriched samples to increase the number of targets [51]. In some cases, certain targets were underrepresented in the dataset, which caused the AI models to perform poorly on those [32,43], emphasizing the challenges of limited training data. To overcome limited datasets, approaches, such as using data augmentation, publicly available datasets, CNNs with pretrained weights, and unsupervised learning, were deployed [38,44,52]. Although there are many ways to limit the effect of small datasets, the improved diagnostic performance in studies, iteratively collecting larger datasets, highlights that insufficient training data remain a limiting factor when developing AI-supported digital microscopy for PHC laboratories [17,33,34,52]. Larger studies and collaborations that allow data sharing could provide solutions to the issue of limited training data.

AI Analysis Pipeline

The AI analysis pipelines used can be broadly divided into 3 main steps: FOV identification, FOV classification, and aggregation for sample-level diagnosis (Figure 3). Given that the analysis of a single sample took more than 30 minutes in some studies and that access to graphics processing units may be limited in PHC laboratories, efficiency becomes important. One strategy to minimize computational demands is to combine identification and classification, as implemented in the You Only Look Once framework, which uses a single CNN [34,50,58]. Another strategy is to first identify targets using fast and computationally efficient methods and subsequently feed the suspicious FOVs into more computationally intensive algorithms for classification. Using an initial object identification step may also enhance the uniformity of the data entering the classification stage, which may be particularly beneficial due to the variability in manually prepared samples that are commonly used in PHC laboratories [19].

For the third step, slide-level classification, different approaches were used: slides were classified as positive if a single positive target was detected; others applied cutoffs to reduce noise and false positives. In addition, certain studies used methods such as SVMs [19] or multiple instance learning [44] to aggregate slide-level results. While these methods may improve classification, they carry a risk of overfitting, especially since the number of training samples at the slide level is much smaller than at the object level.

Manual Verification

One study investigated AI-supported digital microscopy with and without human verification. In the study, human verification was performed on targets initially classified as positive by the AI models, which led to a 0.9% drop in sensitivity but a 29.5% increase in specificity [36]. This demonstrates that, with human intelligence, AI errors can be identified and removed without a substantial loss of sensitivity. This is in line with the high specificity presented in studies using human verification, which all showed specificity of >90% [39,42]. Expanding human verification to include borderline cases classified as negative may also be used to reduce false negatives and increase sensitivity.

Reported Diagnostic Performance

The reported diagnostic performance of the studies included in the scoping review indicates that AI-supported digital microscopy may achieve comparable diagnostic accuracy in PHC laboratories; however, it is important to account for methodological choices when interpreting the results and due to the heterogeneity in study designs, comparisons between studies and diseases become challenging. One example of methodological choices is that most studies used manual microscopy as the reference standard. Since microscopy itself is an imperfect diagnostic test, it can affect the performance of the index test and may result in over- or underestimation of the diagnostic accuracy of AI-supported microscopy. Two studies argued that this limitation may have reduced the apparent diagnostic accuracy of AI-supported digital microscopy [49,50]. Another aspect to consider is the number of samples on which the method was evaluated. For example, the study in which the AI-supported digital microscopy had higher sensitivity and specificity than manual microscopy analyzed 30 samples [19]. Generally, the QUADAS-2 tool indicated a low risk of bias. However, it did not capture the issue of AI models being trained on samples from the same collection, which is a potential source of poor generalizability for AI. This can occur even when the detection algorithms are trained on different datasets, for example, when thresholds or rules for deriving slide-level diagnoses are developed using the same slides on which diagnostic performance is later evaluated, leading to inflated estimates of diagnostic accuracy. However, some studies avoided training on the data from the same collection, included more than 100 test samples, had low risk of QUADAS-2 bias, and used a more advanced reference standard and still achieved comparable or better results than manual microscopy [33,36,39,53].

XSL•FO
RenderX

Limitations

A limitation of the extraction process was the lack of consistent terminology used in the field. This was exemplified in the search block aimed at identifying PHC. Terms such as “low-cost” and “PHC” were included but not “remote,” which was used to describe one study that fulfilled the inclusion criteria [59]. Another limitation was the broad definition of PHC laboratories adopted from Fleming et al, which led to the inclusion of studies using relatively advanced methods, such as oil immersion scanning at 100× magnification and specially designed cartridges for sample preparation [2,33,41,42]. These methods may be difficult to implement in some PHC laboratories, but to achieve a more comprehensive overview of the field, the inclusion of these studies was deemed advantageous [25]. A third limitation stems from the lack of standardized methodological descriptions in the included studies. In some cases, key information, such as scanner magnification, was missing or reported inconsistently across studies, which complicated comparisons in the scoping review.

Steps Needed to Achieve Clinical Implementation

In this scoping review, we identified hurdles that were shared across several studies and that must be overcome before implementing AI-supported digital microscopy. Many developers have recognized a need to iteratively improve their AI-supported digital microscopy; thus, a framework that enables continuous improvements might be advantageous for supporting the development of more accurate AI models. This requires health policy guidelines and frameworks that give details on how these processes should be conducted [60]. Another hurdle in the implementation of AI-supported digital microscopy is cost. The development of lower-cost scanners has reduced expenses; however, most commercially available scanners remain more expensive than traditional microscopes [38,61]. Microscopes typically function reliably over long periods, and scanners may need comparable longevity for AI-supported digital microscopy to be cost-effective. One potential solution to this is modular scanner construction, which may improve its lifespan through component updates and thereby its sustainability. Some systems in this review were developed for specific diseases, which increases the cost of implementing them in PHC laboratories, as multiple systems would have to be acquired to replace microscopes. To make it economically feasible to implement AI-supported digital microscopy, it may, therefore, be necessary to adopt a multipurpose approach where systems are developed for multiple diseases. Some studies show that scanners can digitize different samples and similar approaches can be applied to different diseases [9,37,49]. Systems developed for specific diseases may instead be useful in large screening programs or epidemiological surveys, for example, for soil-transmitted helminths, malaria, or cancer screening. Cost-effectiveness trials could provide guidance for the feasibility of AI-supported digital microscopy and evaluate single-disease systems against multipurpose platforms.

Potential Implications for PHC Diagnostics

AI-supported digital microscopy has potential advantages compared with manual microscopy in PHC laboratories. First, it could improve diagnostic accuracy, especially sensitivity, and this may be further enhanced by incorporating human verification [39,53]. Second, it might increase the access and timeliness of diagnostics by allowing diagnostic procedures to be performed at the POC in PHC laboratories and eliminate the need to send samples elsewhere for analysis [9,19]. Third, it could alleviate the workload of personnel through task shifting. This could increase the productivity of experts and thereby access to image-based diagnosis [36,39].

Other POC diagnostic technologies, including rapid diagnostic tests (RDTs) and PCR-based methods, provide alternative means of diagnosing some conditions discussed in this scoping review [62,63]. One included study compared AI-supported digital microscopy with another POC-diagnostic method: a malaria study comparing it with RDTs [39]. The AI-supported digital microscopy had higher sensitivity and specificity than RDTs for some malaria species and settings but lower in others. One review proposed that AI-supported digital microscopy holds more promise than RDTs for malaria diagnosis [64]. However, other studies highlight the possibilities of other methods, such as RDTs and PCR, for POC diagnostics [62].

Implementation of AI-supported digital microscopy could strengthen health systems and increase health equity, particularly where resources are limited such as in scarcely populated areas and LMICs. This may be possible since, in addition to comparable diagnostic accuracy to microscopy, slides can be scanned and analyzed within approximately 30 minutes for multiple diseases [34,39,53]. As the process eliminates the need for microscopy expertise on-site, it could enable timely and accurate diagnostics in PHC laboratories that currently lack this capacity: even if manual verification is required, it can be performed remotely. Moreover, by decentralizing diagnostics, it may reduce referrals to higher-tier health care facilities, alleviating their work and minimizing the risk of referral-related dropouts [12].

Knowledge Gaps and Research Priorities

This scoping review identified evidence of the feasibility of AI-supported digital microscopy for multiple targets in PHC laboratories. Drawing on the evidence mapped here, future research should prioritize studying scalable and robust systems that can be transferred and implemented in new laboratories and settings. Achieving scalability requires research into AI-supported digital microscopy with an end-to-end perspective, where everything from sample preparation, scanning, and AI analysis until the final diagnosis is accounted for and easily reproducible. Adding to this, research that examines how predeveloped AI-based systems are transferred and implemented in new clinical settings would provide valuable insights into real-world robustness, which was done by some of the included studies. To enable this kind of research, large multisite collaborations are important, which could be facilitated by improved health policy guidelines and frameworks, as well as initiatives led by key stakeholders including governments and nongovernmental organizations. Furthermore, important research priorities include assessing cost-effectiveness and exploring perceived barriers to implementation among patients and health care professionals. Finally, the scoping review’s screening process identified additional potential applications for

XSL•FO
RenderX

AI-supported digital microscopy, including tuberculosis, other parasitic diseases, respiratory cytology, sperm motility, and sickle cell anemia, which warrant further investigation in PHC settings [65-69].

Conclusions

This scoping review identified 22 studies deploying AI-supported digital microscopy in PHC laboratories. For multiple diagnostic purposes, AI-supported digital microscopy achieved comparable results to the reference standard and could be particularly advantageous for increasing sensitivity in diagnosis. Further research is needed on challenges such as generalizability, scalability, and cost-effectiveness. Such evidence is critical to stimulate product development, enable regulatory approval, and support reimbursement and adoption by health care authorities. If the methods can be demonstrated to be feasible in real-life clinical PHC settings, translated into medical device products, and carefully integrated into health care systems, they are likely to improve access to diagnostics, particularly in LMICs and scarcely populated regions.

Acknowledgments

The authors acknowledge the support of the Karolinska Institute's Library for guidance in developing a satisfactory search strategy. The authors declare the use of generative artificial intelligence in the research and writing process. According to the GAIDeT taxonomy (2025) [70], the following tasks were delegated to GAI tools under full human supervision: (1) Proofreading and editing, and (2) adapting and adjusting the tone and style. Paperpal Preflight and ChatGPT 4, 4.5, and 5 (GPT, OpenAI's large-scale language generation model) were used for these purposes.

Funding

This study was funded by the Erling-Persson Foundation and Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. In addition, it was supported by the Swedish Research Council, Finska Läkaresällskapet r.f., Wilhelm och Else Stockmanns stiftelse r.f and Medicinska Understödsföreningen Liv och Hälsa r.f. No funder had any role in the study design or decision to submit the paper for publication.

Data Availability

All data generated within this scoping review are available in the study, such as search details.

Authors' Contributions

JVB conceived and designed the study, wrote the manuscript, developed the search strategy, screened the studies for eligibility, and performed the data extraction. AS screened the studies for eligibility, performed the data extraction, and assisted in writing the manuscript. JL conceived and designed the study, assisted in writing the manuscript, and developing the search strategy. NL conceived and designed the study, assisted in writing the manuscript, developing the search strategy, and screened the studies for eligibility. VD conceived and designed the study and assisted in writing the manuscript. AM assisted in designing the study and in writing the manuscript.

Conflicts of Interest

JL reported receiving personal fees from Aiforia Technologies Oy and serving as cofounder and co-owner of Aiforia Technologies Oy outside the submitted work. JL and AS reported having a patent for Mobile Microscope pending (no. WO2017037334A1, the invention is related to the use of fluorescence imaging filters combined with inexpensive plastic lenses, and all rights are with the University of Helsinki) and JL having a patent for a slide holder for an optical microscope pending (no. WO2015185805A1; related to motorization of regular microscopes). All other authors have no conflicts of interest to declare.

Multimedia Appendix 1
Search strategy.
[DOCX File, 14 KB - jmir_v28i1e78500_app1.docx]

Multimedia Appendix 2
Original data extraction tool.
[DOCX File, 14 KB - jmir_v28i1e78500_app2.docx]

Multimedia Appendix 3
Quadas-2 tool.
[DOCX File, 40 KB - jmir_v28i1e78500_app3.docx]

Multimedia Appendix 4

XSL•FO
RenderX

Sample collection and scanning.
[DOCX File, 45 KB - jmir_v28i1e78500_app4.docx]

Checklist 1
PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File, 118 KB - jmir_v28i1e78500_app5.pdf]

References

1. Pinto-Coelho L. How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications. Bioengineering (Basel) 2023 Dec 18;10(12):1435. [doi: 10.3390/bioengineering10121435] [Medline: 38136026]
2. Fleming KA, Naidoo M, Wilson M, et al. High-quality diagnosis: an essential pathology package. In: Jamison DT, Gelband H, Horton S, editors. Disease Control Priorities: Improving Health and Reducing Poverty, 3rd edition: International Bank for Reconstruction and Development/World Bank; 2017. [doi: 10.1596/978-1-4648-0527-1_ch11]
3. Wilson ML, Fleming KA, Kuti MA, Looi LM, Lago N, Ru K. Access to pathology and laboratory medicine services: a crucial gap. Lancet 2018 May 12;391(10133):1927-1938. [doi: 10.1016/S0140-6736(18)30458-6] [Medline: 29550029]
4. Shamshad F, Khan S, Zamir SW, et al. Transformers in medical imaging: a survey. Med Image Anal 2023 Aug;88:102802. [doi: 10.1016/j.media.2023.102802] [Medline: 37315483]
5. Mohammed FA, Tune KK, Assefa BG, Jett M, Muhie S. Medical image classifications using convolutional neural networks: a survey of current methods and statistical modeling of the literature. MAKE 2024;6(1):699-736. [doi: 10.3390/make6010033]
6. Rezende MT, Bianchi AGC, Carneiro CM. Cervical cancer: automation of Pap test screening. Diagn Cytopathol 2021 Apr;49(4):559-574. [doi: 10.1002/dc.24708] [Medline: 33548162]
7. da Silva LM, Pereira EM, Salles PG, et al. Independent real-world application of a clinical-grade automated prostate cancer detection system. J Pathol 2021 Jun;254(2):147-158. [doi: 10.1002/path.5662] [Medline: 33904171]
8. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): a comparative analysis. Lancet Digit Health 2021 Mar;3(3):e195-e203. [doi: 10.1016/S2589-7500(20)30292-2] [Medline: 33478929]
9. Holmström O, Linder N, Kaingu H, et al. Point-of-care digital cytology with artificial intelligence for cervical cancer screening in a resource-limited setting. JAMA Netw Open 2021 Mar 1;4(3):e211740. [doi: 10.1001/jamanetworkopen.2021.1740] [Medline: 33729503]
10. Ward P, Dahlberg P, Lagatie O, et al. Affordable artificial intelligence-based digital pathology for neglected tropical diseases: a proof-of-concept for the detection of soil-transmitted helminths and Schistosoma mansoni eggs in Kato-Katz stool thick smears. PLoS Negl Trop Dis 2022 Jun;16(6):e0010500. [doi: 10.1371/journal.pntd.0010500] [Medline: 35714140]
11. Linder N, Nyirenda D, Mårtensson A, Kaingu H, Ngasala B, Lundin J. AI supported diagnostic innovations for impact in global women's health. BMJ 2025 Oct 10;391:e086009. [doi: 10.1136/bmj-2025-086009] [Medline: 41073085]
12. Diagnostic errors. World Health Organization. 2016. URL: https://iris.who.int/handle/10665/252410 [accessed 2025-11-27]
13. Bogoch II, Lundin J, Lo NC, Andrews JR. Mobile phone and handheld microscopes for public health applications. Lancet Public Health 2017 Aug;2(8):e355. [doi: 10.1016/S2468-2667(17)30120-2]
14. Galvan P, Ortellado J, Rivas R, Grossling B, Hilario E. PP84 developing the network for the future of healthcare through telemedicine-driven diagnostic innovation. Int J Technol Assess Health Care 2024 Dec;40(S1):S89-S89. [doi: 10.1017/S0266462324002538]
15. Reschke P, Gruenewald LD, Koch V, et al. Radiology access in rural Germany: a nationwide survey on outpatient imaging and teleradiology. Diagnostics (Basel) 2025 Apr 10;15(8):962. [doi: 10.3390/diagnostics15080962] [Medline: 40310336]
16. Bychkov A, Fukuoka J. Evaluation of the global supply of pathologists [abstract]. In: Modern Pathology: Springer Nature; 2022, Vol. 35. [doi: 10.1038/s41379-022-01050-6]
17. Oyibo P, Jujjavarapu S, Meulah B, et al. Schistoscope: an automated microscope with artificial intelligence for detection of Schistosoma haematobium eggs in resource-limited settings. Micromachines (Basel) 2022 Apr 19;13(5):643. [doi: 10.3390/mi13050643] [Medline: 35630110]
18. Holmström O, Linder N, Ngasala B, et al. Point-of-care mobile digital microscopy and deep learning for the detection of soil-transmitted helminths and Schistosoma haematobium. Glob Health Action 2017 Jun;10(sup3):1337325. [doi: 10.1080/16549716.2017.1337325] [Medline: 28838305]
19. Sunny S, Baby A, James BL, et al. A smart tele-cytology point-of-care platform for oral cancer screening. PLoS ONE 2019;14(11):e0224885. [doi: 10.1371/journal.pone.0224885] [Medline: 31730638]
20. Khanagar SB, Naik S, Al Kheraif AA, et al. Application and performance of artificial intelligence technology in oral cancer diagnosis and prediction of prognosis: a systematic review. Diagnostics (Basel) 2021;11(6):1004. [doi: 10.3390/diagnostics11061004]
21. Rodriguez JPM, Rodriguez R, Silva VWK, et al. Artificial intelligence as a tool for diagnosis in digital pathology whole slide images: a systematic review. J Pathol Inform 2022;13:100138. [doi: 10.1016/j.jpi.2022.100138] [Medline: 36268059]

XSL•FO
RenderX

22. Peters MDJ, Godfrey C, McInerney P, Munn Z, Tricco AC, Khalil H. Chapter 11: scoping reviews (2020 version). In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis: JBI; 2020. [doi: 10.46658/JBIRM-20-01]
23. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
24. von Bahr J, Diwan V, Mårtensson A, Linder N, Lundin J. Artificial intelligence-supported digital microscopy diagnostics in primary health care laboratories: a scoping review protocol. OSF. 2024 Mar 7. URL: https://osf.io/yz67t/overview [accessed 2025-12-10]
25. von Bahr J, Diwan V, Mårtensson A, Linder N, Lundin J. AI-supported digital microscopy diagnostics in primary health care laboratories: protocol for a scoping review. JMIR Res Protoc 2024 Nov 1;13:e58149. [doi: 10.2196/58149] [Medline: 39486020]
26. Yang B, Olsen M, Vali Y, et al. Study designs for comparative diagnostic test accuracy: a methodological review and classification scheme. J Clin Epidemiol 2021 Oct;138:128-138. [doi: 10.1016/j.jclinepi.2021.04.013]
27. Clark J, Glasziou P, Del Mar C, Bannach-Brown A, Stehlik P, Scott AM. A full systematic review was completed in 2 weeks using automation tools: a case study. J Clin Epidemiol 2020 May;121:81-90. [doi: 10.1016/j.jclinepi.2020.01.008] [Medline: 32004673]
28. Covidence.: Veritas Health Innovation URL: https://www.covidence.org/ [accessed 2025-12-07]
29. Chien TI, Kao JT, Liu HL, et al. Urine sediment examination: a comparison of automated urinalysis systems and manual microscopy. Clinica Chimica Acta 2007 Sep;384(1-2):28-34. [doi: 10.1016/j.cca.2007.05.012]
30. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
31. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. [doi: 10.1136/bmj.n71]
32. Bachar N, Benbassat D, Brailovsky D, et al. An artificial intelligence-assisted diagnostic platform for rapid near-patient hematology. Am J Hematol 2021 Oct 1;96(10):1264-1274. [doi: 10.1002/ajh.26295] [Medline: 34264525]
33. Gasparin AT, Araujo CIF, Cardoso MR, et al. Hilab system device in an oncological hospital: a new clinical approach for point of care CBC test, supported by the internet of things and machine learning. Diagnostics (Basel) 2023 May 11;13(10):1695. [doi: 10.3390/diagnostics13101695] [Medline: 37238184]
34. Gasparin AT, Araujo CIF, Schmitt P, et al. Hilab system, a new point-of-care hematology analyzer supported by the Internet of Things and artificial intelligence. Sci Rep 2022 Jun 21;12(1):10409. [doi: 10.1038/s41598-022-13913-8] [Medline: 35729182]
35. Ardicoglu Akisin Y, Akar N, Burkay Cotelı M. Decision support system for the classification of Downey cells as a pre-diagnostic tool. Turkish Journal of Biochemistry 2024 Jan 2;48(6):634-640. [doi: 10.1515/tjb-2023-0035]
36. Hamid MMA, Mohamed AO, Mohammed FO, et al. Diagnostic accuracy of an automated microscope solution (miLab™) in detecting malaria parasites in symptomatic patients at point-of-care in Sudan: a case-control study. Malar J 2024 Jun 28;23(1):200. [doi: 10.1186/s12936-024-05029-3] [Medline: 38943203]
37. Bae CY, Shin YM, Kim M, et al. Embedded-deep-learning-based sample-to-answer device for on-site malaria diagnosis. Front Bioeng Biotechnol 2024;12:1392269. [doi: 10.3389/fbioe.2024.1392269] [Medline: 39100623]
38. Holmström O, Stenman S, Suutala A, et al. A novel deep learning-based point-of-care diagnostic method for detecting Plasmodium falciparum with fluorescence digital microscopy. PLoS One 2020;15(11):e0242355. [doi: 10.1371/journal.pone.0242355] [Medline: 33201905]
39. Ewnetu Y, Badu K, Carlier L, et al. A digital microscope for the diagnosis of Plasmodium falciparum and Plasmodium vivax, including P. falciparum with hrp2/hrp3 deletion. PLOS Glob Public Health 2024;4(5):e0003091. [doi: 10.1371/journal.pgph.0003091] [Medline: 38768243]
40. Das D, Vongpromek R, Assawariyathipat T, et al. Field evaluation of the diagnostic performance of EasyScan GO: a digital malaria microscopy device based on machine-learning. Malar J 2022 Apr 12;21(1):122. [doi: 10.1186/s12936-022-04146-1] [Medline: 35413904]
41. Torres K, Bachman CM, Delahunt CB, et al. Automated microscopy for routine malaria diagnosis: a field comparison on Giemsa-stained blood films in Peru. Malar J 2018 Sep 25;17(1):339. [doi: 10.1186/s12936-018-2493-0] [Medline: 30253764]
42. Linder N, Turkki R, Walliander M, et al. A malaria diagnostic tool based on computer vision screening and visualization of Plasmodium falciparum candidate areas in digitized blood smears. PLOS One 2014;9(8):e104855. [doi: 10.1371/journal.pone.0104855] [Medline: 25144549]
43. Horning MP, Delahunt CB, Bachman CM, et al. Performance of a fully-automated system on a WHO malaria microscopy evaluation slide set. Malar J 2021 Feb 25;20(1):110. [doi: 10.1186/s12936-021-03631-3] [Medline: 33632222]
44. Stegmüller T, Abbet C, Bozorgtabar B, et al. Self-supervised learning-based cervical cytology for the triage of HPV-positive women in resource-limited settings and low-data regime. Comput Biol Med 2024 Feb;169:107809. [doi: 10.1016/j.compbiomed.2023.107809] [Medline: 38113684]
45. Skandarajah A, Sunny SP, Gurpur P, et al. Mobile microscopy as a screening tool for oral cancer in India: a pilot study. PLoS One 2017;12(11):e0188440. [doi: 10.1371/journal.pone.0188440] [Medline: 29176904]

XSL•FO
RenderX

46. Ghaderinia M, Abadijoo H, Mahdavian A, et al. Smartphone-based device for point-of-care diagnostics of pulmonary inflammation using convolutional neural networks (CNNs). Sci Rep 2024 Mar 22;14(1):6912. [doi: 10.1038/s41598-024-54939-4] [Medline: 38519489]
47. Soares FA, Suzuki CTN, Sabadini E, et al. Laboratory validation of the automated diagnosis of intestinal parasites via fecal sample processing for the recovery of intestinal parasites through the dissolved air flotation technique. Parasit Vectors 2024 Aug 30;17(1):368. [doi: 10.1186/s13071-024-06434-y] [Medline: 39215369]
48. Carvalho JD, Santos BD, Gomes JF, et al. TF‐Test modified: new diagnostic tool for human enteroparasitosis. Clinical Laboratory Analysis 2016 Jul;30(4):293-300. [doi: 10.1002/jcla.21854]
49. Lundin J, Suutala A, Holmström O, et al. Diagnosis of soil-transmitted helminth infections with digital mobile microscopy and artificial intelligence in a resource-limited setting. PLoS Negl Trop Dis 2024 Apr;18(4):e0012041. [doi: 10.1371/journal.pntd.0012041] [Medline: 38602896]
50. Sahu A, Kandaswamy S, Singh DV, et al. AI driven lab-on-chip cartridge for automated urinalysis. SLAS Technol 2024 Jun;29(3):100137. [doi: 10.1016/j.slast.2024.100137] [Medline: 38657705]
51. Meulah B, Oyibo P, Bengtson M, et al. Performance evaluation of the Schistoscope 5.0 for (semi-)automated digital detection and quantification of Schistosoma haematobium eggs in urine: a field-based study in Nigeria. Am J Trop Med Hyg 2022 Nov 14;107(5):1047-1054. [doi: 10.4269/ajtmh.22-0276] [Medline: 36252803]
52. Oyibo P, Meulah B, Bengtson M, et al. Two-stage automated diagnosis framework for urogenital schistosomiasis in microscopy images from low-resource settings. J Med Imag 2023;10(4):044005. [doi: 10.1117/1.JMI.10.4.044005]
53. Meulah B, Oyibo P, Hoekstra PT, et al. Validation of artificial intelligence-based digital microscopy for automated detection of Schistosoma haematobium eggs in urine in Gabon. PLoS Negl Trop Dis 2024 Feb;18(2):e0011967. [doi: 10.1371/journal.pntd.0011967] [Medline: 38394298]
54. Mehanian C, Jaiswal M, Delahunt C, et al. Computer-automated malaria diagnosis and quantitation using convolutional neural networks. Presented at: 2017 IEEE International Conference on Computer Vision Workshop (ICCVW); Oct 22-29, 2017; Venice, Italy p. 125. [doi: 10.1109/ICCVW.2017.22]
55. Delahunt CB, Jaiswal MS, Horning MP, et al. Fully-automated patient-level malaria assessment on field-prepared thin blood film microscopy images, including supplementary information. arXiv. Preprint posted online on Sep 11, 2022. [doi: 10.48550/arXiv.1908.01901]
56. Osaku D, Cuba CF, Suzuki CTN, Gomes JF, Falcão AX. Automated diagnosis of intestinal parasites: a new hybrid approach and its benefits. Comput Biol Med 2020 Aug;123:103917. [doi: 10.1016/j.compbiomed.2020.103917] [Medline: 32768052]
57. Rivenson Y, Wang H, Wei Z, et al. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. Nat Biomed Eng 2019 Jun;3(6):466-477. [doi: 10.1038/s41551-019-0362-y] [Medline: 31142829]
58. Ali ML, Zhang Z. The YOLO Framework: a comprehensive review of evolution, applications, and benchmarks in object detection. Computers 2024;13(12):336. [doi: 10.3390/computers13120336]
59. Cure-Bolt N, Perez F, Broadfield LA, et al. Artificial intelligence-based digital pathology for the detection and quantification of soil-transmitted helminths eggs. PLoS Negl Trop Dis 2024 Sep;18(9):e0012492. [doi: 10.1371/journal.pntd.0012492] [Medline: 39348405]
60. Regulatory considerations on artificial intelligence for health. World Health Organization. 2023. URL: https://www.who.int/publications/i/item/9789240078871 [accessed 2025-11-27]
61. Patel A, Balis UGJ, Cheng J, et al. Contemporary whole slide imaging devices and their applications within the modern pathology department: a selected hardware review. J Pathol Inform 2021;12:50. [doi: 10.4103/jpi.jpi_66_21] [Medline: 35070479]
62. Ofori B, Twum S, Nkansah Yeboah S, Ansah F, Amofa Nketia Sarpong K. Towards the development of cost-effective point-of-care diagnostic tools for poverty-related infectious diseases in sub-Saharan Africa. PeerJ 2024;12:e17198. [doi: 10.7717/peerj.17198] [Medline: 38915381]
63. Gupta R, Gupta S. Point-of-care tests for human papillomavirus detection in uterine cervical samples: a review of advances in resource-constrained settings. Indian J Med Res 2023 Nov 1;158(5 & amp; amp;6):509-521. [doi: 10.4103/ijmr.ijmr_1143_23] [Medline: 38236008]
64. Coro F, De Maria C, Mangano VD, Ahluwalia A. Technologies for the point-of-care diagnosis of malaria: a scoping review. Infect Dis Poverty 2025 Jun 23;14(1):54. [doi: 10.1186/s40249-025-01329-1] [Medline: 40551195]
65. Mandal S, Das D, Udutalapally V. mSickle: sickle cell identification through gradient evaluation and smartphone microscopy. J Ambient Intell Human Comput 2023 Oct;14(10):13319-13331. [doi: 10.1007/s12652-022-03786-0]
66. Aulia S, Suksmono AB, Mengko TR, Alisjahbana B. A novel digitized microscopic images of ZN-stained sputum smear and its classification based on IUATLD grades. IEEE Access 2024;12:51364-51380. [doi: 10.1109/ACCESS.2024.3386208]
67. Roberts J, Flenaugh E, Oprea-Ilies G. AI aided rapid on site evaluation of respiratory cytology. J Am Soc Cytopathol 2021 Sep;10(5):S2. [doi: 10.1016/j.jasc.2021.07.131]
68. Kanakasabapathy MK, Sadasivam M, Singh A, et al. An automated smartphone-based diagnostic assay for point-of-care semen analysis. Sci Transl Med 2017 Mar 22;9(382):eaai7863. [doi: 10.1126/scitranslmed.aai7863]
69. D'Ambrosio MV, Bakalar M, Bennuru S, et al. Point-of-care quantification of blood-borne filarial parasites with a mobile phone microscope. Sci Transl Med 2015 May 6;7(286):286re4. [doi: 10.1126/scitranslmed.aaa3480]

XSL•FO
RenderX

70. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2025 Aug 8:1-27. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]

Abbreviations

AI: artificial intelligence
CBC: complete blood count
CNN: convolutional neural network
FDA: US Food and Drug Administration
FOV: fields-of-view
LMICs: low- and middle-income countries
PCR: polymerase chain reaction
PHC: primary health care
POC: point of care
PRISMA: Preferred Reporting Items for Systematic reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2
RDT: rapid diagnostic test
SVM: support vector machine
WHO: World Health Organization

Edited by J Sarvestan; submitted 05.Jun.2025; peer-reviewed by HH Kim, SM Shaffi; revised version received 16.Nov.2025; accepted 17.Nov.2025; published 05.Jan.2026.

Please cite as:
von Bahr J, Suutala A, Diwan V, Mårtensson A, Lundin J, Linder N
AI-Supported Digital Microscopy Diagnostics in Primary Health Care Laboratories: Scoping Review
J Med Internet Res 2026;28:e78500
URL: https://www.jmir.org/2026/1/e78500
doi:10.2196/78500

© Joar von Bahr, Antti Suutala, Vinod Diwan, Andreas Mårtensson, Johan Lundin, Nina Linder. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effectiveness of Machine Learning in Detecting Vessels Encapsulating Tumor Clusters in Hepatocellular Carcinoma: Systematic Review and Meta-Analysis

Huili Shui[1,2,3], MD; Wenyu Wu[1,2,3], MD; Zhenming Xie[1,2,3], MD; Bing Yang[1,2,3], MD; Jia Deng[1,2,3], MD; Dongxin Tang[1,2,3], MD

[1]Guizhou University of Traditional Chinese Medicine, Guiyang, Guizhou Province, China

[2]Clinical Medical Research Center, The First Affiliated Hospital of Guizhou University of Traditional Chinese Medicine, Guiyang, Guizhou Province, China

[3]Guizhou Traditional Chinese Medicine Oncology Heritage and Technology Innovation Talent Base, Guiyang, Guizhou Province, China

Corresponding Author:
Dongxin Tang, MD
Guizhou University of Traditional Chinese Medicine
71 Baoshan North Road, Yunyan District
Guiyang, Guizhou Province
China
Phone: 86 18608511204
Email: hemingankang@sina.com

Abstract

Background: Vessels encapsulating tumor clusters (VETC) are significantly associated with poor prognosis in hepatocellular carcinoma (HCC). However, identifying VETC early remains challenging. Recently, machine learning has shown promise for VETC detection, but their diagnostic accuracy lacks systematic validation.

Objective: This meta-analysis aimed to systematically evaluate the diagnostic accuracy of machine learning models for detecting VETC in patients with HCC.

Methods: The Cochrane Library, Embase, Web of Science, and PubMed were searched up to June 21, 2025. Eligible studies focused on machine learning models for HCC VETC diagnosis. Studies that merely analyzed risk factors or lacked outcome measures were excluded. The Prediction Model Risk of Bias Assessment Tool was used to evaluate the risk of bias. A bivariate mixed-effects model was used for a meta-analysis based on 2×2 diagnostic tables. Subgroup analyses were performed according to modeling variables (nonradiomic vs radiomic features) and model types (traditional machine learning vs deep learning).

Results: This meta-analysis included 31 studies comprising 6755 patients with HCC (2699 VETC-positive). Nineteen studies used machine learning models based on nonradiomic features, and 12 used radiomic features (including deep learning). In the validation set, the nonradiomic model demonstrated a pooled sensitivity of 0.72 (95% CI 0.66-0.78), specificity of 0.74 (95% CI 0.68-0.80), and an area under the summary receiver operating characteristic curve (SROC AUC) of 0.80 (95% CI 0.76-0.83). The radiomic model showed sensitivity of 0.81 (95% CI 0.73-0.87), specificity of 0.73 (95% CI 0.67-0.79), and SROC AUC of 0.84 (95% CI 0.80-0.87). Traditional machine learning achieved sensitivity of 0.84 (95% CI 0.71-0.92), specificity of 0.75 (95% CI 0.67-0.81), and SROC AUC of 0.83 (95% CI 0.80-0.86). Deep learning exhibited sensitivity of 0.77 (95% CI 0.69-0.84), specificity of 0.70 (95% CI 0.59-0.79), and SROC AUC of 0.81 (95% CI 0.77-0.85).

Conclusions: This meta-analysis is the first to quantitatively assess the efficacy of machine learning models in HCC VETC diagnosis, addressing an evidence gap in this field. Unlike previous descriptive reviews, this analysis provides the first quantitative evidence revealing the potential value of machine learning in detecting HCC VETC. The findings provide a foundation for developing and refining subsequent intelligent detection tools. Despite their promising prospects, machine learning models have not yet reached the maturity required for clinical translation, owing to methodological heterogeneity, limited validation, and a high risk of bias. Future research should focus on conducting multicenter, large-sample, standardized, prospective studies to advance clinical translation.

Trial Registration: PROSPERO CRD420251084894; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251084894

(J Med Internet Res 2026;28:e82839) doi:10.2196/82839

XSL•FO
RenderX

KEYWORDS

carcinoma; hepatocellular carcinoma; hepatocellular; machine learning; radiomics; sensitivity; specificity; vessels encapsulating tumor clusters; VETC

Introduction

Liver cancer is the sixth most frequently diagnosed malignancy and the third leading cause of cancer-related mortality worldwide. In 2022, there were approximately 865,000 new cases of liver cancer and 757,948 related deaths. Hepatocellular carcinoma (HCC) accounts for 75% to 85% of primary liver cancers. Higher incidence and mortality rates are predominantly observed in developing regions, including Mongolia, Cambodia, Laos, Thailand, Vietnam, and Egypt [1]. Consequently, HCC has become a significant global oncological concern. For early-stage HCC, curative interventions such as surgical resection, liver transplantation, and ablation are recommended. For intermediate stages, locoregional therapies are typically used, while systemic treatment is preferred for individuals with a significant intrahepatic tumor burden. Advanced HCC is primarily managed with immune checkpoint inhibitors [2]. While these treatments have prolonged the survival of some patients, others still have a poor prognosis, even after undergoing the same treatment regimen. Patients with HCC have an overall 5-year survival rate of less than 20% [3]. Adding to the clinical challenge, the postoperative recurrence rate for HCC is high, at around 70%, even after curative resection. This persistent risk is a primary factor in unfavorable long-term patient prognosis [4]. Several factors have been associated with poor HCC prognosis, including microvascular invasion [5], the macrotrabecular-massive subtype [6], and the coexpression of Ki-67 and cytokeratin 19 [7]. Thus, identifying the key factors that drive poor prognosis in HCC is crucial.

Recently, there has been an increase in attention directed toward a distinct microvascular pattern known as vessels encapsulating tumor clusters (VETC). First described by Fang et al [8] in 2015, VETC refers to a vascular network surrounding tumor clusters in a spiderweb configuration. This pattern has been characterized as an independent vascular morphology that is distinct from epithelial-mesenchymal transition. It facilitates the release of entire tumor clusters into the bloodstream and enables a noninvasive metastatic mechanism in HCC. Research indicates that the prevalence of VETC correlates with tumor stage and aggressiveness. VETC occurs in approximately 30%-40% of patients undergoing resection, 50%-55% of individuals with postresection recurrence, and up to 76% of patients with unresectable disease receiving liver transplants [9]. A positive VETC status substantially influences the long-term prognosis of patients with HCC [10]. Recent studies show that patients with VETC have significantly shorter overall and disease-free survival than patients without VETC [11]. The presence of VETC has been established as a robust predictor of aggressive HCC behavior [10]. A meta-analysis by Wang et al [12] further confirmed VETC as a significant predictor of overall survival and tumor recurrence, supporting its role as an effective prognostic biomarker. Furthermore, multiparameter prognostic models that incorporate VETC status demonstrate superior predictive capacity for disease-free and overall survival in patients with HCC compared to the conventional tumor-node-metastasis staging system. These models facilitate personalized temporal survival estimation and have the potential to enhance clinical decision-making regarding surveillance management and therapeutic strategies [13,14]. Concurrently, VETC status is valuable in guiding systemic therapy selection and predicting treatment response in HCC. Notably, patients with VETC experience greater survival benefits from therapies including sorafenib [15], lenvatinib [16], and transarterial chemoembolization [17] than their VETC-negative counterparts do. These observations suggest that VETC-based stratified treatment strategies may optimize patient outcomes further, providing an evidence base for clinical decision-making [9]. Therefore, early detection of VETC status is clinically relevant for improving HCC prognosis. Currently, a definitive diagnosis of VETC relies on histopathological examination of biopsy or resected tissue specimens. However, this approach has several limitations. Technical challenges include dependence on tumor size and needle gauge, as well as variability among clinicians and pathologists. Procedural risks encompass hemorrhage, seeding metastasis, sampling error, and uncertainty in tumor characterization [18]. Thus, noninvasive methods for identifying VETC status in HCC are urgently needed to circumvent the limitations of tissue acquisition. Recent advancements in image processing and artificial intelligence have sparked growing interest in clinical oncology in the development of predictive models that integrate computed tomography, magnetic resonance imaging (MRI), and contrast-enhanced ultrasound with machine learning algorithms. This methodology is increasingly being explored for the noninvasive diagnosis of HCC VETC [19-21]. Several studies have explored the potential for directly diagnosing VETC in HCC using images alone [22,23]. Furthermore, machine learning models that incorporate clinical and imaging features have been developed to noninvasively predict VETC status [20,24]. Despite these promising findings, there is a lack of systematic evidence substantiating the efficacy of machine learning–based approaches for VETC detection in HCC. This lack of evidence poses a significant challenge to the development and improvement of artificial intelligence–assisted diagnostic tools. To address this deficiency, this systematic review and meta-analysis were conducted to summarize the performance of machine learning in noninvasively detecting VETC in HCC. The aim is to provide evidence-based support for developing and optimizing future intelligent diagnostic tools.

Methods

Study Registration

This meta-analysis was conducted in strict accordance with the PRISMA (Preferred Reporting Items for a Systematic Review and Meta-Analysis; checklist provided in Multimedia Appendix 1) Diagnostic Test Accuracy Studies guidelines [25] and was prospectively registered with the International Prospective Register of Systematic Reviews (CRD420251084894).

XSL•FO
RenderX

Eligibility Criteria

Textbox 1 presents the eligibility criteria for studies.

Textbox 1. Eligibility criteria.

Inclusion criteria

- Participants diagnosed with hepatocellular carcinoma
- Cohort, case-control, or cross-sectional studies
- Studies that developed machine learning models for the diagnosis vessels encapsulating tumor clusters
- Publications reported in English

Exclusion criteria

- Reviews, guidelines, expert opinions, or conference abstracts
- Studies that only performed risk factor analyses without constructing machine learning models
- Studies lacking key metrics for assessing the accuracy of machine learning models
- Studies reporting only univariable predictive performance

Data Sources and Search Strategy

According to the PRISMA search guidelines, the PubMed, Embase, Cochrane Library, and Web of Science databases were searched up to June 21, 2025. The search combined Medical Subject Headings and free-text terms, with no restrictions on language, country, or publication date. The search strategy was developed independently for this analysis. It was not adapted from existing systematic reviews, nor did it incorporate additional information sources or use search filters. The strategy did not undergo peer review before its execution, and no updates were made to the search following the initial retrieval. Based on the existing literature, we manually examined the reference lists of selected studies and relevant reviews to identify additional articles. Conference proceedings were excluded, and no attempts were made to contact authors for additional information [26]. Details are presented in Table S1 in Multimedia Appendix 2.

Study Selection

All retrieved articles were imported into EndNote (version 21; Clarivate) to remove duplicates. Two researchers (HS and ZX) screened the titles and abstracts of the articles independently and excluded the irrelevant ones. Subsequently, the full texts of potentially eligible studies were acquired and assessed for final inclusion. The researchers then cross-checked their results. Any discrepancies were resolved through discussion or adjudication by a third researcher (WW).

Data Extraction

Prior to data extraction, a standardized spreadsheet was developed. The extracted data included the following: first author, number of VETC cases, patient source, total sample size, study design, detection method, number of VETC cases in the training set, total training set size, country, method of validation set generation, model type, publication year, total validation set size, number of VETC cases in the validation set, and modeling variables.

Risk of Bias

The Prediction Model Risk of Bias Assessment Tool was applied to evaluate the risk of bias across four domains: participants, predictors, analysis, and outcome. Each domain contained 2-9 signaling questions, which could be answered as “yes or probably yes,” “no or probably no,” or “no information.” Domain-specific judgments were categorized as low, high, or unclear risk of bias. A domain was judged as having a low risk of bias if all signaling questions were answered “yes or probably yes”; a high risk of bias if at least one was answered “no or probably no”; or an unclear risk of bias if at least 1 was answered “no information” while all others were answered “yes or probably yes.” Two researchers (HS and ZX) conducted the assessment independently. They then cross-checked their results. Any disagreements were settled by consensus or arbitration by a third researcher (WW).

Synthesis Methods

A bivariate random-effects model was used for the meta-analysis based on available 2×2 diagnostic tables (either reported directly or reconstructed from reported performance metrics and sample size). The following pooled estimates were derived with their corresponding 95% CIs, sensitivity, specificity, positive likelihood ratio (LR+), negative likelihood ratio (LR–), diagnostic odds ratio (DOR), and the area under the summary receiver operating characteristic curve (SROC AUC). Deeks’ funnel plot was used to evaluate small-study effects. Fagan’s nomogram was applied to evaluate clinical applicability. Subgroup analyses were conducted according to modeling variables (nonradiomic vs radiomic features) and model type (traditional machine learning vs deep learning). A *P* value of <.05 indicated statistical significance. Stata (version 15.1; StataCorp LLC) was used for all meta-analyses.

Results

Study Selection

The database search yielded 302 potentially relevant articles. Of these, 177 duplicates were excluded (117 identified by software and 60 manually). After screening the titles and

XSL•FO
RenderX

abstracts, 89 articles unrelated to the topic were removed. The full texts of the remaining 36 articles were assessed for eligibility. Among them, 5 records were excluded; 3 because they did not develop machine learning models, and 2 because they were conference abstracts without full-text publication. Ultimately, 31 eligible studies were included [20-24,27-52]. The specific process is depicted in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for a Systematic Review and Meta-Analysis) flow diagram of the selection process for studies applying machine learning to detect hepatocellular carcinoma vessels encapsulating tumor clusters.

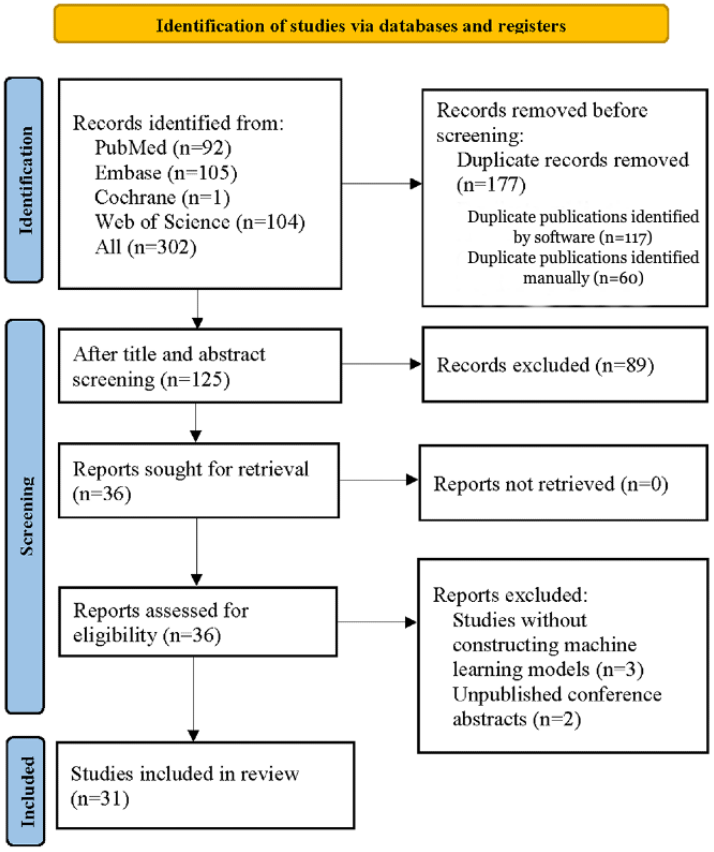

Study Characteristics

The 31 studies were published between 2021 and 2025. All were conducted in China and Japan. Of these, 8 studies used a case-control design, and 23 used a cohort design. Patient data were derived from multiple centers in 10 studies and from single centers in 21 studies. A total of 6755 participants with HCC were included, 2699 of whom were identified as VETC-positive. Regarding detection methods, 1 study used radiomic features based on computed tomography, 6 studies used MRI-based radiomics, 5 studies used deep learning, and 19 studies used traditional machine learning. The training sets collectively comprised 4411 participants with HCC, including 1714 with VETC. Internal validation was conducted in 14 studies, external validation in 3 studies, and both in 7 studies. The validation sets encompassed 2344 participants with HCC, 955 of whom were VETC-positive. The prediction models incorporated machine learning (n=5), logistic regression (n=24), least absolute shrinkage and selection operator regression (n=1), and random forest (n=1). Detailed characteristics are illustrated in Tables S2-S4 in Multimedia Appendix 2.

Risk of Bias

The Prediction Model Risk of Bias Assessment Tool was applied across 4 domains to assess the overall risk of bias. First, 8 of the 31 eligible studies in the participants domain used a case-control design, which introduced a high risk of bias due to potential differences in data sources and patient selection. Second, case-control studies were judged to carry a high risk of bias in the predictors domain because predictor assessment was influenced by knowledge of the outcome. Third, in the outcome domain, VETC status was consistently defined and confirmed via histopathological examination. Since the outcome definition, measurement, and classification were independent of predictor assessment and participant selection, this domain was assessed as having a low risk of bias. Fourth, in the analysis domain, 14 studies were judged to have a high risk of bias due to an insufficient sample size (including an events-per-variable ratio of <10 in model development, a validation set size of <100, or an absence of external validation). A total of 12 studies were rated as having an unclear risk of bias due to an inability to calculate the events-per-variable ratio. One study provided no explanation for missing values and was therefore judged to be at high risk of bias regarding missing data. Concerning model validation, 6 studies relied solely on random data splitting without cross-validation or mediator effect testing, resulting in a high risk of bias. Overall, 10 studies did not report the validation method used and were categorized as having an unclear risk of bias. Detailed assessment results are shown in Figure 2.

XSL•FO
RenderX

Figure 2. Risk of bias assessment for the included primary studies using the Prediction Model Risk of Bias Assessment Tool.

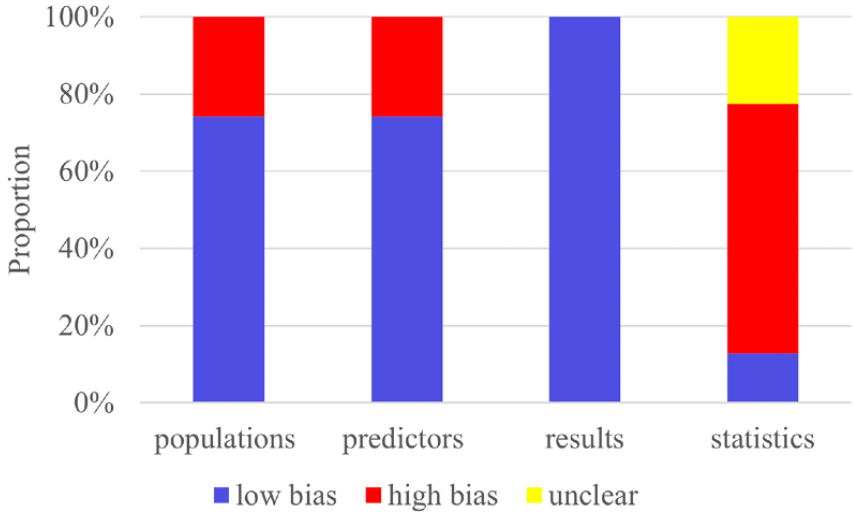

Meta-Analysis

Training Set-Overall

A total of 27 models from the training sets provided 2×2 diagnostic tables, with a 39% VETC-positive proportion. The pooled estimates were as follows: sensitivity 0.77 (95% CI 0.72-0.82), specificity 0.83 (95% CI 0.78-0.87), LR+ 4.5 (95% CI 3.5-5.8), LR– 0.27 (95% CI 0.22-0.34), DOR 16 (95% CI 11-24), and SROC AUC 0.87 (95% CI 0.84-0.89; Figures S1 and S2 in Multimedia Appendix 2). No significant small-study effect was illustrated via Deeks' funnel plot (*P*=.70; Figure S3 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 75%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 85% (Figure S4 in Multimedia Appendix 2).

Training Set-Nonradiomic Features

A total of 18 nonradiomic models from the training sets provided 2×2 diagnostic tables, with 38% representing VETC-positive cases. The pooled estimates were as follows: sensitivity 0.74 (95% CI 0.67-0.79), specificity 0.81 (95% CI 0.77-0.85), LR+ 3.9 (95% CI 3.2-4.7), LR– 0.33 (95% CI 0.27-0.40), DOR 12 (95% CI 9-16), and SROC AUC 0.85 (95% CI 0.81-0.88; Figures S5 and S6 in Multimedia Appendix 2). No significant small-study effect was detected via Deeks' funnel plot (*P*=.46; Figure S7 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 72%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 82% (Figure S8 in Multimedia Appendix 2).

Training Set-Radiomic Features

A total of 9 radiomic models from the training set provided 2×2 diagnostic tables, with a VETC-positive rate of 40%. The pooled estimates were as follows: sensitivity 0.83 (95% CI 0.75-0.90), specificity 0.86 (95% CI 0.71-0.94), LR+ 6.0 (95% CI 2.6-13.5), LR– 0.19 (95% CI 0.11-0.32), DOR 31 (95% CI 9-106), and SROC AUC 0.91 (95% CI 0.88-0.93; Figures S9 and S10 in Multimedia Appendix 2). No significant small-study effect was observed via Deeks' funnel plot (*P*=.40; Figure S11 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 80%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 89% (Figure S12 in Multimedia Appendix 2).

Of these, 6 traditional machine learning models provided 2×2 diagnostic tables, with a VETC-positive rate of 39%. The pooled estimates were as follows: sensitivity 0.88 (95% CI 0.70-0.96), specificity 0.85 (95% CI 0.67-0.94), LR+ 5.7 (95% CI 2.2-15.2), LR– 0.14 (95% CI 0.04-0.45), DOR 40 (95% CI 5-326), and SROC AUC 0.93 (95% CI 0.90-0.95; Figures S13 and S14 in Multimedia Appendix 2). No significant small-study effect was found via Deeks' funnel plot (*P*=.78; Figure S15 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 79%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 91% (Figure S16 in Multimedia Appendix 2).

Only 3 deep learning studies reported 2×2 diagnostic tables. Yu et al [35] developed an MRI-based deep learning model with a sensitivity of 0.87, a specificity of 0.54, and an area under the receiver operating characteristic curve (ROC AUC) of 0.83 (95% CI 0.83-0.84). Xu et al [49] reported a contrast-enhanced ultrasound–based model with sensitivity of 0.75, specificity of 0.92, and ROC AUC of 0.92 (95% CI 0.88-0.96). Yang et al [48] developed an MRI-based model with a sensitivity of 0.71, a specificity of 0.97, and an ROC AUC of 0.90 (95% CI 0.85-0.95).

Validation Set-Overall

A total of 27 models in the validation set provided complete 2×2 diagnostic tables, with a VETC-positive proportion of 41%. The pooled estimates were as follows: sensitivity 0.77 (95% CI 0.72-0.81), specificity 0.74 (95% CI 0.69-0.78), LR+ 2.9 (95% CI 2.5-3.3), LR– 0.32 (95% CI 0.26-0.38), DOR 9 (95% CI 7-12), and SROC AUC 0.82 (95% CI 0.78-0.85; Figures S17

XSL•FO
RenderX

and S18 in Multimedia Appendix 2). Deeks' funnel plot demonstrated no significant small-study effects (*P*=.09; Figure S19 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 66%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 83% (Figure S20 in Multimedia Appendix 2).

Validation Set-Nonradiomic Features

A total of 12 nonradiomic models in the validation set provided 2×2 diagnostic tables, with a VETC-positive proportion of 40%. The pooled estimates were as follows: sensitivity 0.72 (95% CI 0.66-0.78), specificity 0.74 (95% CI 0.68-0.80), LR+ 2.8 (95% CI 2.3-3.5), LR– 0.37 (95% CI 0.31-0.45), DOR 8 (95% CI 6-10), and SROC AUC 0.80 (95% CI 0.76-0.83; Figures S21 and S22 in Multimedia Appendix 2). No significant small-study effect was detected via Deeks' funnel plot (*P*=.98; Figure S23 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 65%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 80% (Figure S24 in Multimedia Appendix 2).

Validation Set-Radiomic Features

A total of 15 radiomic models in the validation set provided 2×2 diagnostic tables, with a VETC-positive rate of 41%. The pooled estimates were as follows: sensitivity 0.81 (95% CI 0.73-0.87), specificity 0.73 (95% CI 0.67-0.79), LR+ 3.0 (95% CI 2.5-3.7), LR– 0.26 (95% CI 0.19-0.36), DOR 12 (95% CI 8-17), and SROC AUC 0.84 (95% CI 0.80-0.87; Figures S25 and S26 in Multimedia Appendix 2). No significant small-study effect was observed via Deeks' funnel plot (*P*=.11; Figure S27 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 67%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 85% (Figure S28 in Multimedia Appendix 2).

Of these, 9 traditional machine learning models provided 2×2 diagnostic tables, with a VETC-positive rate of 41%. The pooled estimates were as follows: sensitivity 0.84 (95% CI 0.71-0.92), specificity 0.75 (95% CI 0.67-0.81), LR+ 3.3 (95% CI 2.6-4.3), LR– 0.21 (95% CI 0.11-0.39), DOR 16 (95% CI 8-32), and SROC AUC 0.83 (95% CI 0.80-0.86; Figure S29 and S30 in Multimedia Appendix 2). No significant small-study effect was shown via Deeks' funnel plot (*P*=.37; Figures S31 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 69%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 88% (Figure S32 in Multimedia Appendix 2).

Additionally, 6 deep learning models reported 2×2 diagnostic tables, with a VETC-positive proportion of 41%. The pooled estimates were as follows: sensitivity 0.77 (95% CI 0.69-0.84), specificity 0.70 (95% CI 0.59-0.79), LR+ 2.6 (95% CI 1.9-3.5), LR– 0.32 (95% CI 0.24-0.43), DOR 8 (95% CI 5-13), and SROC AUC 0.81 (95% CI 0.77-0.85; Figures S33 and S34 in Multimedia Appendix 2). Deeks' funnel plot suggested significant small-study effects (*P*=.04; Figure S35 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 63%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 82% (Figure S36 in Multimedia Appendix 2).

Discussion

Summary of Main Findings

This meta-analysis demonstrates that developing prediction models based on machine learning to detect HCC VETC status appears to be a feasible approach. Currently, these models are primarily constructed using nonradiomic and radiomic features. For nonradiomic machine learning models in the validation set, the pooled estimates were 0.72 (95% CI 0.66-0.78) for sensitivity and 0.74 (95% CI 0.68-0.80) for specificity. For radiomic machine learning models, the estimates were a sensitivity of 0.81 (95% CI 0.73-0.87) and a specificity of 0.73 (95% CI 0.67-0.79). For traditional machine learning models, the estimates were a sensitivity of 0.84 (95% CI 0.71-0.92) and a specificity of 0.75 (95% CI 0.67-0.81). For deep learning models, the estimates were a sensitivity of 0.77 (95% CI 0.69-0.84) and a specificity of 0.70 (95% CI 0.59-0.79).

Comparison With Previous Reviews

Previous research by Hyungjin Rhee et al [53] reviewed the angiodynamic changes in multistep HCC carcinogenesis. They introduced the typical pathological, clinical, and imaging features of HCC VETC and provided detailed guidance for VETC diagnosis. However, their study focused primarily on describing pathological mechanisms and typical features, lacking a quantitative assessment of different diagnostic methods. Ken Liu et al [9] investigated various methods for diagnosing VETC, including histopathology, imaging, and laboratory tests. They suggested that VETC could be predicted radiologically. While their research provided a comprehensive analysis of various diagnostic approaches, they did not quantitatively compare the sensitivity and specificity of different diagnostic methods. This omission limited a thorough evaluation of VETC diagnostic accuracy. Miaomiao Wang et al [54] explored the potential of machine learning in HCC VETC detection through a literature review and provided guidance for the auxiliary VETC diagnosis. While their review demonstrated the potential applications of machine learning in VETC detection, it lacked a direct comparison of different types of machine learning models, making it difficult to assess these models' actual application value in clinical practice. This study summarized nonradiomic (clinical features, image features, etc) and radiomic prediction models, and the diagnosis of current HCC VETC status appears to be an ideal noninvasive detection scheme that provides specific guidance for clinicians.

This study found that the model variables used to detect HCC VETC include both nonradiomic features (clinical features, image features, etc) and radiomic features. The clinical features

XSL•FO
RenderX

primarily consist of alpha-fetoprotein, carbohydrate antigen 19-9, aspartate aminotransferase, and indirect bilirubin. Image features mainly comprise intratumoral necrosis, low signal intensity around the tumor in the hepatobiliary phase, the tumor-to-liver signal intensity ratio on the hepatobiliary phase, and the tumor-to-liver apparent diffusion coefficient ratio. Various studies used different modeling variables. Most studies did not quantitatively present the association of modeling variables with VETC. Thus, a further summary of such correlations was not performed. Recently, radiomics has advanced the development and application of prediction models by converting images into repeatable quantitative data. Prediction models based on radiomic features have demonstrated significant clinical value in diagnosing and treating HCC. Studies have shown that radiomic features are effective in predicting HCC microvascular invasion [5], early recurrence [55], and Ki-67 and cytokeratin 19 expression [7].

In this meta-analysis, only a limited number of studies explored the diagnostic performance of radiomics for HCC VETC. While the studies demonstrated promising results, radiomics still faces significant challenges in practical application. For example, the quality of the image appears to change under different image parameters. Most studies in this meta-analysis did not discuss how such changes in image features affect radiomics results. Additionally, image segmentation is primarily divided into manual and deep learning automatic segmentation. The studies included in this meta-analysis primarily used manual segmentation. However, manual segmentation may be affected by the segmenter's prior knowledge. Although some researchers have attempted to summarize its repeatability through independent segmentation by multiple people, it is difficult to avoid the influence of the segmenter's experience on the region-of-interest area division. Therefore, future studies should consider developing and promoting a standardized radiomics analysis process manual to improve research repeatability. Many studies have demonstrated that models combining radiomics, clinical features, and imaging features perform better in disease diagnosis and prognosis prediction [56]. In this study, relatively few studies attempted to construct prediction models using a combination of clinical features and radiomics. Therefore, an effective quantitative analysis of the advantages of a combined model was difficult to perform. Future studies should explore and verify the value of radiomic models constructed from clinical features and imaging features in improving the diagnostic accuracy of HCC VETC.

The prediction models used in this study primarily encompassed logistic regression, random forest, deep learning, and least absolute shrinkage and selection operator regression. Due to the interpretability of its parameters, logistic regression allows for the development of simple and intuitive nomograms in clinical practice and appears to be favored by many researchers [57-59]. However, the interpretability of other machine learning models, such as random forest, support vector machines, and XGBoost, depends on analyses like Shapley additive explanations. Using them in clinical practice requires developing plugins, which increases the complexity of the application process [60-62]. Thus, from the perspectives of clinical simplicity and interpretability, logistic regression has relatively ideal advantages. Nonetheless, in many cases, logistic regression's predictive accuracy often appears no better than that of traditional machine learning models, such as random forest [46,63]. In radiomics, the core advantage of deep learning lies in its ability to efficiently process image data for disease diagnosis and prognosis prediction [64,65]. Relatively few studies in the radiomic feature literature included in this meta-analysis addressed deep learning models. Initial evidence suggested that deep learning models did not perform significantly better than traditional machine learning models. The primary reasons for this include the following. First, the study only incorporated 6 deep learning research projects, which is a relatively small sample size. Deep learning models typically require large-scale datasets to leverage their full advantages. Second, most studies lacked external validation, leaving the generalizability of the models inadequately tested. Third, variations in image acquisition parameters and quality across different research centers suggest that the design of deep learning model architectures and hyperparameter optimization may not yet be optimal. Therefore, future research developing intelligent tools to detect HCC VETC should attempt to integrate multicenter, large-sample medical image data to construct deep learning models for training and validation.

Advantages and Limitations

This meta-analysis is the first comprehensive summary of the performance of machine learning models in diagnosing HCC VETC. It provides evidence-based support for the subsequent development or updating of artificial intelligence systems. However, this study also has the following limitations. First, all 31 eligible studies originated from East Asia, and most relied primarily on internal validation. The lack of multicenter, multiethnic validation limited the assessment of the models' generalizability. Second, the best prediction model from each article was extracted, which covered a narrow range of machine learning types. The differences between different machine learning methods were not described. Third, the modeling variables were diverse. They were only presented without a quantitative description of their association with HCC VETC. Future research should adopt more transparent and interpretable modeling approaches to identify efficient predictors. Fourth, although deep learning can efficiently process image data, it does not have a significant advantage over traditional machine learning-based radiomics. However, the literature is limited, and the interpretation of the results may be subject to certain limitations. Fifth, HCC VETC is a novel mode of microvascular metastasis that has been proposed in recent years, and the associated research is in its initial stage. The positive definition has not yet been standardized.

Conclusions

This meta-analysis is the first to provide a systematic and quantitative assessment of machine learning for diagnosing HCC VETC, thereby addressing an evidence gap in this field. Unlike previous reviews, this study provides a quantitative evaluation of diagnostic performance. The findings demonstrate the feasibility and clinical potential of using machine learning to determine VETC status in patients with HCC. Notably, radiomics-based models exhibited significantly better

XSL•FO
RenderX

performance than nonradiomic models. While deep learning efficiently processes image data in radiomics, its performance is not significantly better than traditional machine learning-based radiomics. Despite their promising prospects, machine learning models have not yet reached the maturity required for clinical translation, owing to methodological heterogeneity, limited validation, and a high risk of bias. Future research should focus on conducting multicenter, large-sample, standardized, prospective studies to develop intelligent detection tools with higher performance. Validating the models across multiple regions and ethnic populations is essential to ensure their generalizability. This will ultimately enable the effective translation of research into clinical applications.

Funding

This research was supported by the Guizhou Provincial Engineering Research Center for Medical Transformation of Traditional Chinese Medicine and Ethnic Medicine in Cancer Prevention and Treatment (Qian-Jiao-Ji, 2023, No. 037), the Guizhou Provincial Science and Technology Program (Qian-Ke-He Platform Talents, 2020, No. 5013), the Talent Base for Traditional Chinese Medicine Oncology Inheritance and Technological Innovation of Guizhou Province (Qian-Ren-Ling-Fa, 2018, No. 3), and the Guizhou High-Level Innovative Talent Training Program ("Hundred-Level" Talents; Qian-Ke-He Ren-Cai, 2016, No. 4032).

The funding agency for this study did not participate in any aspect of the research design, data collection, analysis, interpretation of results, or manuscript preparation.

Data Availability

All data used in this study are included in the main text and supplementary materials. For further information or clarification, please contact the corresponding researchers.

Authors' Contributions

Conceptualization: HS
Methodology: HS, WW
Formal analysis: HS, JD, BY, WW
Investigation: HS
Data curation: HS, ZX, WW
Visualization: JD, BY
Writing—Original Draft: HS, JD, BY
Writing—Review & Editing: HS, DT
Supervision: DT
Project administration: DT
Funding acquisition: DT
All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-S checklist.
[DOCX File , 23 KB - jmir_v28i1e82839_app1.docx]

Multimedia Appendix 2
Supplementary documents.
[DOCX File , 36193 KB - jmir_v28i1e82839_app2.docx]

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263 [FREE Full text] [doi: 10.3322/caac.21834] [Medline: 38572751]
2. Hwang SY, Danpanichkul P, Agopian V, Mehta N, Parikh ND, Abou-Alfa GK, et al. Hepatocellular carcinoma: updates on epidemiology, surveillance, diagnosis and treatment. Clin Mol Hepatol 2025;31(Suppl):S228-S254 [FREE Full text] [doi: 10.3350/cmh.2024.0824] [Medline: 39722614]
3. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin 2022;72(1):7-33 [FREE Full text] [doi: 10.3322/caac.21708] [Medline: 35020204]

XSL•FO
RenderX

4. Herrero A, Toubert C, Bedoya JU, Assenat E, Guiu B, Panaro F, et al. Management of hepatocellular carcinoma recurrence after liver surgery and thermal ablations: state of the art and future perspectives. Hepatobiliary Surg Nutr 2024;13(1):71-88 [FREE Full text] [doi: 10.21037/hbsn-22-579] [Medline: 38322198]
5. Li L, Wu C, Huang Y, Chen J, Ye D, Su Z. Radiomics for the preoperative evaluation of microvascular invasion in hepatocellular carcinoma: a meta-analysis. Front Oncol 2022;12:831996 [FREE Full text] [doi: 10.3389/fonc.2022.831996] [Medline: 35463303]
6. Ziol M, Poté N, Amaddeo G, Laurent A, Nault J, Oberti F, et al. Macrotrabecular-massive hepatocellular carcinoma: a distinctive histological subtype with clinical relevance. Hepatology 2018;68(1):103-112. [doi: 10.1002/hep.29762] [Medline: 29281854]
7. Zhou L, Chen Y, Li Y, Wu C, Xue C, Wang X. Diagnostic value of radiomics in predicting Ki-67 and cytokeratin 19 expression in hepatocellular carcinoma: a systematic review and meta-analysis. Front Oncol 2023;13:1323534 [FREE Full text] [doi: 10.3389/fonc.2023.1323534] [Medline: 38234405]
8. Fang J, Zhou H, Zhang C, Shang L, Zhang L, Xu J, et al. A novel vascular pattern promotes metastasis of hepatocellular carcinoma in an epithelial-mesenchymal transition-independent manner. Hepatology 2015;62(2):452-454. [doi: 10.1002/hep.27760] [Medline: 25711742]
9. Liu K, Dennis C, Prince DS, Marsh-Wakefield F, Santhakumar C, Gamble JR, et al. Vessels that encapsulate tumour clusters vascular pattern in hepatocellular carcinoma. JHEP Rep 2023;5(8):100792. [doi: 10.1016/j.jhepr.2023.100792] [Medline: 37456680]
10. Renne SL, Woo HY, Allegra S, Rudini N, Yano H, Donadon M, et al. Vessels encapsulating tumor clusters (VETC) is a powerful predictor of aggressive hepatocellular carcinoma. Hepatology 2020;71(1):183-195. [doi: 10.1002/hep.30814] [Medline: 31206715]
11. Xu D, Li R, Shu C, Li Y, Tao R, Chen Y, et al. Association between vessels encapsulating tumor clusters and circulating tumor cells in hepatocellular carcinoma: clinical evidence and risk model development. Int J Med Sci 2025;22(12):2944-2955 [FREE Full text] [doi: 10.7150/ijms.111025] [Medline: 40657396]
12. Wang M, Cao L, Wang Y, Huang H, Tian X, Lei J. The prognostic value of vessels encapsulating tumor clusters (VETC) in patients with hepatocellular carcinoma: a systematic review and meta-analysis. Clin Transl Oncol 2024;26(8):2037-2046. [doi: 10.1007/s12094-024-03427-2] [Medline: 38523240]
13. Xiong S, Wang C, Zhang M, Yang X, Yun J, Liu L. A multi-parametric prognostic model based on clinicopathologic features: vessels encapsulating tumor clusters and hepatic plates predict overall survival in hepatocellular carcinoma patients. J Transl Med 2024;22(1):472 [FREE Full text] [doi: 10.1186/s12967-024-05296-3] [Medline: 38762511]
14. Wu M, Xiao Y, Wang Y, Deng L, Wang X, An T. Establishment of a clinical model based on vessels encapsulating tumour clusters that could efficiently predict recurrence of patients with hepatocellular carcinoma after curative hepatectomy. Pathology 2025;57(3):320-327. [doi: 10.1016/j.pathol.2024.08.014] [Medline: 39668071]
15. Fang J, Xu L, Shang L, Pan C, Ding J, Tang Y, et al. Vessels that encapsulate tumor clusters (VETC) pattern is a predictor of sorafenib benefit in patients with hepatocellular carcinoma. Hepatology 2019;70(3):824-839. [doi: 10.1002/hep.30366] [Medline: 30506570]
16. Zhang P, Ono A, Fujii Y, Hayes CN, Tamura Y, Miura R, et al. The presence of vessels encapsulating tumor clusters is associated with an immunosuppressive tumor microenvironment in hepatocellular carcinoma. Int J Cancer 2022;151(12):2278-2290. [doi: 10.1002/ijc.34247] [Medline: 36054900]
17. Wang J, Li X, Tang H, Fang R, Song J, Feng Y, et al. Vessels that encapsulate tumor clusters (VETC) pattern predicts the efficacy of adjuvant TACE in hepatocellular carcinoma. J Cancer Res Clin Oncol 2023;149(8):4163-4172. [doi: 10.1007/s00432-022-04323-4] [Medline: 36050540]
18. European Association for the Study of the Liver. EASL clinical practice guidelines on the management of hepatocellular carcinoma. J Hepatol 2025;82(2):315-374. [doi: 10.1016/j.jhep.2024.08.028] [Medline: 39690085]
19. Wei Y, Huang S, Huang L, Pei W, Zuo Y, Liao H. CT-based radiomics features combined with AFP for predicting vessels encapsulating tumor clusters and prognosis of hepatocellular carcinoma. J Hepatocell Carcinoma 2025;12:2069-2081 [FREE Full text] [doi: 10.2147/JHC.S542092] [Medline: 40969196]
20. Che F, Gao F, Li Q, Ren W, Tang H, Zaina G, et al. Fractal analysis based on Gd-EOB-DTPA-enhanced MRI for prediction of vessels that encapsulate tumor clusters in patients with hepatocellular carcinoma. Int J Surg 2025;111(7):4389-4399. [doi: 10.1097/JS9.0000000000002547] [Medline: 40441719]
21. Xu W, Huang B, Zhang R, Zhong X, Zhou W, Zhuang S, et al. Diagnostic and prognostic ability of contrast-enhanced unltrasound and biomarkers in hepatocellular carcinoma subtypes. Ultrasound Med Biol 2024;50(4):617-626. [doi: 10.1016/j.ultrasmedbio.2024.01.007] [Medline: 38281888]
22. Li C, Wen Y, Xie J, Chen Q, Dang Y, Zhang H, et al. Preoperative prediction of VETC in hepatocellular carcinoma using non-Gaussian diffusion-weighted imaging at high b values: a pilot study. Front Oncol 2023;13:1167209 [FREE Full text] [doi: 10.3389/fonc.2023.1167209] [Medline: 37305565]
23. Wang M, Cao L, Wang Y, Huang H, Cao S, Tian X, et al. Prediction of vessels encapsulating tumor clusters pattern and prognosis of hepatocellular carcinoma based on preoperative gadolinium-ethoxybenzyl-diethylenetriaminepentaacetic acid

XSL•FO
RenderX

magnetic resonance imaging. J Gastrointest Surg 2024;28(4):442-450. [doi: 10.1016/j.gassur.2024.02.004] [Medline: 38583894]
24. Qu Q, Liu Z, Lu M, Xu L, Zhang J, Liu M, et al. Preoperative gadoxetic acid-enhanced MRI features for evaluation of vessels encapsulating tumor clusters and microvascular invasion in hepatocellular carcinoma: creating nomograms for risk assessment. J Magn Reson Imaging 2024;60(3):1094-1110. [doi: 10.1002/jmri.29187] [Medline: 38116997]
25. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, the PRISMA-DTA Group, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA 2018;319(4):388-396 [FREE Full text] [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
26. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc 2021;109(2):174-200 [FREE Full text] [doi: 10.5195/jmla.2021.962] [Medline: 34285662]
27. Chen H, Dong H, He R, Gu M, Zhao X, Song K, et al. Optimizing predictions: improved performance of preoperative gadobenate-enhanced MRI hepatobiliary phase features in predicting vessels encapsulating tumor clusters in hepatocellular carcinoma-a multicenter study. Abdom Radiol (NY) 2024;49(10):3412-3426. [doi: 10.1007/s00261-024-04283-y] [Medline: 38713432]
28. Li M, Zhang G, Li J, Ren Y, Jin X, Ke Q, et al. Intravoxel incoherent motion improves the accuracy of preoperative prediction of vessels encapsulating tumor clusters in hepatocellular carcinoma. J Hepatocell Carcinoma 2025;12:1177-1190 [FREE Full text] [doi: 10.2147/JHC.S519223] [Medline: 40524872]
29. Chu T, Zhao C, Zhang J, Duan K, Li M, Zhang T, et al. Application of a convolutional neural network for multitask learning to simultaneously predict microvascular invasion and vessels that encapsulate tumor clusters in hepatocellular carcinoma. Ann Surg Oncol 2022;29(11):6774-6783 [FREE Full text] [doi: 10.1245/s10434-022-12000-6] [Medline: 35754067]
30. Chen F, Du M, Qi X, Bian L, Wu D, Zhang S, et al. Nomogram estimating vessels encapsulating tumor clusters in hepatocellular carcinoma from preoperative gadoxetate disodium-enhanced MRI. J Magn Reson Imaging 2023;57(6):1893-1905. [doi: 10.1002/jmri.28488] [Medline: 36259347]
31. Chen H, He R, Gu M, Zhao X, Song K, Zou W, et al. Nomogram prediction of vessels encapsulating tumor clusters in small hepatocellular carcinoma ≤3 cm based on enhanced magnetic resonance imaging. World J Gastrointest Oncol 2024;16(5):1808-1820 [FREE Full text] [doi: 10.4251/wjgo.v16.i5.1808] [Medline: 38764811]
32. Fan Y, Yu Y, Wang X, Hu M, Du M, Guo L, et al. Texture analysis based on Gd-EOB-DTPA-enhanced MRI for identifying vessels encapsulating tumor clusters (VETC)-positive hepatocellular carcinoma. J Hepatocell Carcinoma 2021;8:349-359 [FREE Full text] [doi: 10.2147/JHC.S293755] [Medline: 33981636]
33. Zhang C, Zhong H, Zhao F, Ma Z, Dai Z, Pang G. Preoperatively predicting vessels encapsulating tumor clusters in hepatocellular carcinoma: machine learning model based on contrast-enhanced computed tomography. World J Gastrointest Oncol 2024;16(3):857-874 [FREE Full text] [doi: 10.4251/wjgo.v16.i3.857] [Medline: 38577448]
34. Fan Y, Yu Y, Hu M, Wang X, Du M, Guo L, et al. Imaging features based on Gd-EOB-DTPA-enhanced MRI for predicting vessels encapsulating tumor clusters (VETC) in patients with hepatocellular carcinoma. Br J Radiol 2021;94(1119):20200950 [FREE Full text] [doi: 10.1259/bjr.20200950] [Medline: 33417489]
35. Yu Y, Cao L, Shen B, Du M, Gu W, Gu C, et al. Deep learning radiopathomics models based on contrast-enhanced MRI and pathologic imaging for predicting vessels encapsulating tumor clusters and prognosis in hepatocellular carcinoma. Radiol Imaging Cancer 2025;7(2):e240213. [doi: 10.1148/rycan.240213] [Medline: 40084948]
36. Wang F, Numata K, Funaoka A, Liu X, Kumamoto T, Takeda K, et al. Establishment of nomogram prediction model of contrast-enhanced ultrasound and Gd-EOB-DTPA-enhanced MRI for vessels encapsulating tumor clusters pattern of hepatocellular carcinoma. Biosci Trends 2024;18(3):277-288 [FREE Full text] [doi: 10.5582/bst.2024.01112] [Medline: 38866488]
37. Ding Q, Deng X, Huang J, Zhang R, Liu T, Wang J, et al. Application value of enhanced CT imaging features in predicting vessels encapsulating tumor clusters (VETC) positivity in hepatocellular carcinoma. Curr Med Imaging 2025;21:e15734056361565. [doi: 10.2174/0115734056361565250530050624] [Medline: 40511650]
38. Zhang J, Liu M, Qu Q, Lu M, Liu Z, Yan Z, et al. Radiomics analysis of gadoxetic acid-enhanced MRI for evaluating vessels encapsulating tumour clusters in hepatocellular carcinoma. Front Oncol 2024;14:1422119 [FREE Full text] [doi: 10.3389/fonc.2024.1422119] [Medline: 39193385]
39. Feng Z, Li H, Zhao H, Jiang Y, Liu Q, Chen Q, et al. Preoperative CT for characterization of aggressive macrotrabecular-massive subtype and vessels that encapsulate tumor clusters pattern in hepatocellular carcinoma. Radiology 2021;300(1):219-229. [doi: 10.1148/radiol.2021203614] [Medline: 33973839]
40. Yu Y, Liang X, Hou G, Chen X, Hou W, Hou H, et al. Spectral computed tomography parameters for predicting vessels encapsulating tumor clusters (VETC) pattern in hepatocellular carcinoma: a pilot study. Quant Imaging Med Surg 2025;15(4):3285-3297 [FREE Full text] [doi: 10.21037/qims-24-2077] [Medline: 40235763]
41. Matsuda K, Ueno A, Tsuzaki J, Kurebayashi Y, Masugi Y, Yamazaki K, et al. Vessels encapsulating tumor clusters contribute to the intratumor heterogeneity of HCC on Gd-EOB-DTPA-enhanced MRI. Hepatol Commun 2025;9(1):e0593 [FREE Full text] [doi: 10.1097/HC9.0000000000000593] [Medline: 39670871]

XSL•FO
RenderX

42. Yang J, Dong X, Jin S, Wang S, Wang Y, Zhang L, et al. Radiomics model of dynamic contrast-enhanced mri for evaluating vessels encapsulating tumor clusters and microvascular invasion in hepatocellular carcinoma. Acad Radiol 2025;32(1):146-156 [FREE Full text] [doi: 10.1016/j.acra.2024.07.007] [Medline: 39025700]
43. Pan J, Huang H, Zhang S, Zhu Y, Zhang Y, Wang M, et al. Intraindividual comparison of CT and MRI for predicting vessels encapsulating tumor clusters in hepatocellular carcinoma. Eur Radiol 2025;35(1):61-72. [doi: 10.1007/s00330-024-10944-9] [Medline: 38992109]
44. Chernyak V. Editorial for "Deep learning radiomics model of dynamic contrast-enhanced MRI for evaluating vessels encapsulating tumor clusters and prognosis in hepatocellular carcinoma". J Magn Reson Imaging 2024;59(1):120-121. [doi: 10.1002/jmri.28775] [Medline: 37165916]
45. Li Z, Song W, Zhang J, Li Q, Song Z, Ren X, et al. Identification of vessels encapsulating tumor clusters in solitary hepatocellular carcinoma via imaging biomarkers in preoperative contrast-enhanced magnetic resonance imaging. Quant Imaging Med Surg 2024;14(12):8586-8600 [FREE Full text] [doi: 10.21037/qims-24-315] [Medline: 39698687]
46. Yu Y, Fan Y, Wang X, Zhu M, Hu M, Shi C, et al. Gd-EOB-DTPA-enhanced MRI radiomics to predict vessels encapsulating tumor clusters (VETC) and patient prognosis in hepatocellular carcinoma. Eur Radiol 2022;32(2):959-970. [doi: 10.1007/s00330-021-08250-9] [Medline: 34480625]
47. Guan R, Lin W, Zou J, Mei J, Wen Y, Lu L, et al. Development and validation of a novel nomogram for predicting vessels that encapsulate tumor cluster in hepatocellular carcinoma. Cancer Control 2022;29:10732748221102820 [FREE Full text] [doi: 10.1177/10732748221102820] [Medline: 35609265]
48. Yang J, Dong X, Wang F, Jin S, Zhang B, Zhang H, et al. A deep learning model based on MRI for prediction of vessels encapsulating tumour clusters and prognosis in hepatocellular carcinoma. Abdom Radiol (NY) 2024;49(4):1074-1083. [doi: 10.1007/s00261-023-04141-3] [Medline: 38175256]
49. Xu W, Zhang H, Zhang R, Zhong X, Li X, Zhou W, et al. Deep learning model based on contrast-enhanced ultrasound for predicting vessels encapsulating tumor clusters in hepatocellular carcinoma. Eur Radiol 2025;35(2):989-1000. [doi: 10.1007/s00330-024-10985-0] [Medline: 39066894]
50. Meng X, Qu X, Guo Y, Qi X, Bian L, Wu D, et al. Validation of proposed imaging criteria for estimating vessels encapsulating tumor clusters in hepatocellular carcinoma at CT and gadoxetic acid-enhanced MRI. Eur J Radiol 2025;183:111936. [doi: 10.1016/j.ejrad.2025.111936] [Medline: 39848126]
51. Wang Y, Wang M, Cao L, Huang H, Cao S, Tian X, et al. A nomogram for preoperative prediction of vessels encapsulating tumor clusters (VETC) pattern and prognosis of hepatocellular carcinoma. Am J Surg 2024;234:172-178. [doi: 10.1016/j.amjsurg.2024.05.004] [Medline: 38755026]
52. Ruan L, Yu J, Lu X, Numata K, Zhang D, Liu X, et al. A nomogram based on features of ultrasonography and contrast-enhanced CT to predict vessels encapsulating tumor clusters pattern of hepatocellular carcinoma. Ultrasound Med Biol 2024;50(12):1919-1929. [doi: 10.1016/j.ultrasmedbio.2024.08.020] [Medline: 39289116]
53. Rhee H, Park YN, Choi J. Advances in understanding hepatocellular carcinoma vasculature: implications for diagnosis, prognostication, and treatment. Korean J Radiol 2024;25(10):887-901 [FREE Full text] [doi: 10.3348/kjr.2024.0307] [Medline: 39344546]
54. Wang M, Wang Y, Shen Y, Cao L, Yan R, Lei J. Role of vessels encapsulating tumor clusters patterns in hepatocellular carcinoma: a literature review. Eur J Gastroenterol Hepatol 2025. [doi: 10.1097/MEG.0000000000003032] [Medline: 40631491]
55. Lu M, Wang C, Zhuo Y, Gou J, Li Y, Li J, et al. Preoperative prediction power of radiomics and non-radiomics methods based on MRI for early recurrence in hepatocellular carcinoma: a systemic review and meta-analysis. Abdom Radiol (NY) 2024;49(10):3397-3411. [doi: 10.1007/s00261-024-04356-y] [Medline: 38704783]
56. Wang B, Jiang B, Liu D, Zhu R. Early predictive accuracy of machine learning for hemorrhagic transformation in acute ischemic stroke: systematic review and meta-analysis. J Med Internet Res 2025;27:e71654 [FREE Full text] [doi: 10.2196/71654] [Medline: 40408765]
57. Qi Y, Yang S, Li J, Xing H, Su Q, Wang S, et al. Development and validation of a nomogram to predict impacted ureteral stones via machine learning. Minerva Urol Nephrol 2024;76(6):736-747 [FREE Full text] [doi: 10.23736/S2724-6051.24.05856-7] [Medline: 39093225]
58. Wu C, Zhu S, Wang Q, Xu Y, Mo X, Xu W, et al. Development, validation, and visualization of a novel nomogram to predict depression risk in patients with stroke. J Affect Disord 2024;365:351-358. [doi: 10.1016/j.jad.2024.08.105] [Medline: 39173927]
59. Bertens LCM, Moons KGM, Rutten FH, van Mourik Y, Hoes AW, Reitsma JB. A nomogram was developed to enhance the use of multinomial logistic regression modeling in diagnostic research. J Clin Epidemiol 2016;71:51-57. [doi: 10.1016/j.jclinepi.2015.10.016] [Medline: 26577433]
60. Nwanosike EM, Conway BR, Merchant HA, Hasan SS. Potential applications and performance of machine learning techniques and algorithms in clinical practice: a systematic review. Int J Med Inform 2022;159:104679. [doi: 10.1016/j.ijmedinf.2021.104679] [Medline: 34990939]

XSL•FO
RenderX

61. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020;20(1):310 [FREE Full text] [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
62. Qi X, Wang S, Fang C, Jia J, Lin L, Yuan T. Machine learning and SHAP value interpretation for predicting comorbidity of cardiovascular disease and cancer with dietary antioxidants. Redox Biol 2025;79:103470 [FREE Full text] [doi: 10.1016/j.redox.2024.103470] [Medline: 39700695]
63. Li D, Lu H, Wu J, Chen H, Shen M, Tong B, et al. Development of machine learning models for predicting depressive symptoms in knee osteoarthritis patients. Sci Rep 2024;14(1):28603 [FREE Full text] [doi: 10.1038/s41598-024-79601-x] [Medline: 39562701]
64. Zhou T, Cheng Q, Lu H, Li Q, Zhang X, Qiu S. Deep learning methods for medical image fusion: a review. Comput Biol Med 2023;160:106959. [doi: 10.1016/j.compbiomed.2023.106959] [Medline: 37141652]
65. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol 2019;20(5):e262-e273. [doi: 10.1016/S1470-2045(19)30149-4] [Medline: 31044724]

Abbreviations

DOR: diagnostic odds ratio
HCC: hepatocellular carcinoma
LR–: negative likelihood ratio
LR+: positive likelihood ratio
MRI: magnetic resonance imaging
PRISMA: Preferred Reporting Items for a Systematic Review and Meta-Analysis
ROC AUC: area under the receiver operating characteristic curve
SROC AUC: area under the summary receiver operating characteristic curve
VETC: vessels encapsulating tumor clusters

Edited by S Brini; submitted 22.Aug.2025; peer-reviewed by W Wang, X Liang; comments to author 31.Oct.2025; accepted 03.Dec.2025; published 14.Jan.2026.

Please cite as:
Shui H, Wu W, Xie Z, Yang B, Deng J, Tang D
Effectiveness of Machine Learning in Detecting Vessels Encapsulating Tumor Clusters in Hepatocellular Carcinoma: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e82839
URL: https://www.jmir.org/2026/1/e82839
doi:10.2196/82839
PMID:

©Huili Shui, Wenyu Wu, Zhenming Xie, Bing Yang, Jia Deng, Dongxin Tang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Machine Learning Prediction Models for Preeclampsia: Systematic Review and Meta-Analysis

Lu Liu[1,2], MM; Qixuan Zhu[3], BEng; Yichi Zong[2], MM; Xueyuan Chen[2], MM; Wei Zhang[2], MM; Jun Wang[4], PhD

[1]School of Public Health, China Medical University, Shenyang City, Liaoning Province, China

[2]Department of the Obstetrics and Gynecology, Shengjing Hospital Affiliated to China Medical University, Shenyang City, Liaoning Province China, China

[3]School of Engineering, University of Pennsylvania, Philadelphia, PA, United States

[4]Department of the Obstetrics and Gynecology, Shengjing Hospital Affiliated to China Medical University, Shenyang City, Liaoning Province, China

Corresponding Author:
Jun Wang, PhD
Department of the Obstetrics and Gynecology
Shengjing Hospital Affiliated to China Medical University
Heping District, No 36 Sanhao Street
Shenyang City, Liaoning Province, 110000
China
Phone: 86 18940254480
Email: wangj1@sj-hospital.org

Abstract

Background: Preeclampsia is a severe hypertensive disorder with rising global prevalence. While machine learning (ML) models for predicting preeclampsia are increasingly published, existing evidence shows high heterogeneity, and the distinction between internal performance and external transferability remains unclear.

Objective: This study aims to evaluate the performance of ML models in predicting preeclampsia through a systematic review and meta-analysis, while also exploring their potential clinical application value, in order to specifically enhance the quality of future research and the predictive capability of the models.

Methods: Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines and PROSPERO registration, we searched PubMed, Web of Science, IEEE Xplore, and CNKI (China National Knowledge Infrastructure) for studies published through February 2025. We included studies using ML to predict preeclampsia in pregnant women. Bias was assessed using PROBAST (Prediction model Risk of Bias Assessment Tool). We calculated summary estimates using random-effects models and, crucially, computed 95% prediction intervals (PIs) to estimate performance in future clinical settings. Subgroup and meta-regression analyses were conducted to explore heterogeneity.

Results: In total, 26 studies comprising 31 ML models were included. While the pooled area under the receiver operating characteristic curve was high at 0.91 (95% CI 0.87-0.92), extreme heterogeneity was observed (I^2>99%). The 95% PI for sensitivity was wide (0.32-0.96), indicating that in some external settings, sensitivity could drop to 32%. Only 6 studies conducted external validation; in these, the pooled sensitivity decreased to 0.68, with a PI of 0.25-0.94.Subgroup analysis suggested that models incorporating laboratory biomarkers and neural networks outperformed others, though CIs overlapped.

Conclusions: Current evidence suggests that a high area under the curve in ML models is more likely to reflect the “performance” of the model on the internal development dataset rather than its universal “effectiveness” and clinical utility in independent, diverse populations. The apparent performance exhibits significant contextual dependence. Future studies should conduct multicenter, prospective external validation and recalibration research to enhance transferability and reliability.

Trial Registration: PROSPERO CRD420251005830;https://www.crd.york.ac.uk/PROSPERO/view/CRD420251005830

(J Med Internet Res 2026;28:e78714) doi:10.2196/78714

KEYWORDS

preeclampsia; machine learning; predictive models; meta-analysis; artificial intelligence; computer-assisted diagnosis

XSL•FO
RenderX

Introduction

Preeclampsia is a pregnancy-related hypertensive condition marked by the development of high blood pressure and protein in the urine after 20 weeks of gestation. Due to its multiple etiologies and complex pathogenesis, it poses significant risks to both maternal and perinatal health [1]. This specific condition negatively impacts maternal health and can also lead to serious complications for the fetus, including placental abruption and restricted fetal growth. According to global statistics, the incidence of preeclampsia ranges from 3% to 9%, with even higher rates observed in certain high-risk populations [2]. Furthermore, preeclampsia is one of the leading causes of maternal mortality worldwide, particularly in low- and middle-income countries. The prevalence of preeclampsia in China has increased from 5.79% in 2005 to 9.5% in 2019 [3], further underscoring the urgent need for early screening and management. To date, the etiology and pathogenesis of preeclampsia remain incompletely understood, and effective treatment measures are lacking. Consequently, early detection and enhanced management are essential clinical strategies.

Understanding the epidemiological characteristics of preeclampsia is essential for developing effective public health strategies. In the study of preeclampsia, traditional statistical methods primarily emphasize linear models and hypothesis testing, which are effective in uncovering singular relationships between variables. However, the pathological mechanisms underlying preeclampsia are highly complex, involving multiple interacting factors, and traditional methods may face limitations when addressing nonlinear and high-dimensional data. In contrast, machine learning (ML) technology has shown considerable promise in this domain.

A subset of artificial intelligence (AI), ML is a technology that allows computers to independently learn from data and make decisions or predictions using algorithms and models. Its application in clinical settings can effectively prevent and manage diseases. Currently, the usage of ML to develop predictive models for preeclampsia is becoming increasingly prevalent. For instance, Sylvain et al [4] noted that the implementation of ML methods has significantly improved the prediction accuracy of high-risk pregnancies, offering a novel perspective for the early identification of preeclampsia. Furthermore, Ranjbar et al [5] indicated that ML-based models surpass traditional regression models in predicting the incidence of preeclampsia. The multidimensional optimization capabilities of these models allow them to account for interactions among various clinical features and biomarkers, thereby enhancing diagnostic accuracy.

By leveraging ML, researchers can explore both linear and nonlinear relationships, as well as uncover deep-seated features and patterns within the data. This method establishes a scientific foundation for the prompt recognition and intervention of preeclampsia.

Compared with prior systematic reviews and protocols on pregnancy outcomes or preeclampsia, the incremental contributions of this study are as follows: (1) we prespecified and implemented subgroup analyses by outcome definition, gestational window, data source, and validation type to avoid indiscriminate pooling across highly heterogeneous models and populations; (2) we treated area under the curve (AUC) as the primary summary measure and applied robust univariate random-effects models (Hartung-Knapp-Sidik-Jonkman method) to pool sensitivity and specificity separately, accompanied by 95% prediction intervals (PIs) to estimate future performance; and 3) we clearly separated performance in internal vs external validation and documented whether decision-curve analysis was conducted. Taken together, these methodological enhancements aim to provide more interpretable evidence about where deployment may be appropriate and where it remains premature.

Methods

Research Design

This research was carried out in alignment with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 standards [6] (Multimedia Appendix 1 [7]). Specific details regarding the search keywords can be found in Textbox S1 of the Multimedia Appendix 2. Before the study began, the protocol received approval and was registered with the PROSPERO under the reference number CRD420251005830.

Literature Search Strategy

Comprehensive searches were executed in several prestigious databases, including PubMed, Web of Science, IEEE Xplore, and the CNKI (China National Knowledge Infrastructure). These searches focused on locating scholarly papers that were published in either English or Chinese. The time frame for this search encompassed works published until February 2025, ensuring that the most recent and relevant literature was included in the investigation. The search strategy was developed based on the PICO (Population, Intervention, Comparison, and Outcome) framework. In this study, “P” denotes the population with PE, “I” refers to ML methods as the intervention, “C” indicates the gold standard for comparison, and “O” encompasses outcomes, such as sensitivity, specificity, and accuracy for prediction and diagnosis (Table S1 in Multimedia Appendix 2). Additionally, the reference lists from each identified study underwent a manual review to uncover further relevant research. Zotero (Center for History and New Media at George Mason University) was used to organize the studies and remove any duplicates.

The study’s inclusion criteria were formulated to guarantee the rigor and relevance of the research. The criteria encompassed (1) research papers published in English or Chinese; (2) investigations involving pregnant women from the general population that explicitly defined the diagnosis of preeclampsia; (3) studies that used ML models for predicting preeclampsia, along with a thorough explanation of these models; and (4) investigations that showcased the performance of the ML models, offering adequate data to determine both sensitivity and specificity. These criteria aimed to strengthen the validity of the results and ensure a thorough assessment of the existing literature.

XSL•FO
RenderX

The exclusion criteria for this study are as follows: (1) studies that solely investigated risk factors without developing a predictive model; (2) papers published in languages other than English or of types other than original research, such as reports and reviews; (3) duplicate publications; (4) studies that included 2 or fewer predictors in the constructed model; and (5) studies for which the full text was not accessible.

Literature Screening and Data Extraction

Five researchers (LL, QZ, YZ, XC, and WZ) meticulously followed the established inclusion and exclusion criteria to screen the titles and abstracts of the literature. Studies that met these criteria advanced to the full-text reading phase, where all relevant studies were reviewed. Each article underwent a minimum of 2 rounds of screening. Both the title and abstract screening, as well as the full-text reading, were conducted independently by the 2 researchers (LL and QZ). In instances of disagreement between them, another researcher (JW) made the final decision.

In total, 26 studies [8-33] were chosen for analysis. Data extraction was independently performed by 2 researchers (LL and QZ) following the standardized protocol established by the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis), as outlined in the existing literature [34]. Data collected from each study included the following: (1) demographic details, such as the country of data collection, the study setting, the source of the data, the design of the study, and the definition of outcomes; (2) methods for data partitioning, feature selection algorithms, types of ML prediction models, model validation, and applications; (3) results of predictions, which involved accuracy, sensitivity, specificity, and the AUC; and (4) sources of funding and the approval of ethics. This study extracted sensitivity and specificity data from each research report, all based on the "optimal threshold" set in the respective original studies. This research did not standardize or adjust for the differences in thresholds among the various studies.

Bias and Applicability Assessment

Overview

We used PROBAST (Prediction Model Risk of Bias Assessment Tool) as the primary instrument to preserve comparability with prior preeclampsia meta-analyses (for detailed information, see Multimedia Appendix 3). Because many included studies predate PROBAST-AI and lack AI-specific reporting (eg, leakage safeguards, hyperparameter tuning, calibration, and thresholds), a full PROBAST-AI assessment would be dominated by underreporting rather than demonstrated bias. The PROBAST [35] was used to assess the risk of bias in the included studies across 4 domains, namely participants, predictors, outcomes, and analysis. Additionally, applicability assessments were conducted for the domains of population, predictors, and outcomes. Two researchers (LL and QZ) independently reviewed the studies, undergoing consistency training based on a preprepared and trialed scoring manual. The discrepancies were resolved through discussion, and if necessary, a third researcher (JW) acted as an adjudicator.

Bias Assessment

For all questions within a category, if the answers are "yes" or "possibly," the category is assessed as low risk. Conversely, if any answer is "no" or "possibly not," the category is classified as high risk. In cases where there is insufficient information, the category is deemed unclear. The overall risk of bias in the study is determined according to the PROBAST guidelines: (1) if all 4 domains are assessed as low risk, the overall risk of the study is low; (2) if one or more domains are assessed as high risk, the overall risk of the study is high; and (3) if one or more domains are assessed as unclear (and there are no high-risk domains), the overall risk of the study is unclear.

Applicability Assessment

The evaluation encompasses 3 categories, including study object, predictor, and outcome. Each category is assessed based on 3 levels of applicability, namely good applicability, poor applicability, and unclear applicability. If all 3 assessments are classified as good, the overall applicability is determined to be good. Conversely, if any one assessment is classified as poor, the overall applicability is deemed poor. In cases where one assessment is unclear while the other two are good, the overall applicability is classified as unclear.

Statistical Analysis

The methods described in the guidelines for conducting systematic reviews and meta-analyses concerning the performance of prediction models, along with previous meta-analyses of such models, indicate that the concordance index of a model is similar to the AUC [36]. This index indicates the diagnostic or prognostic discrimination ability, categorized as none (AUC≤0.6), poor (0.6<AUC<0.7), moderate (0.7<AUC<0.8), good (0.8<AUC<0.9), or excellent (0.9<AUC<1). Model calibration acts as an indicator of how well the model fits the data by evaluating the alignment between the actual and forecasted results, while also demonstrating the model's reliability via calibration graphs. Additionally, the diagnostic odds ratio (DOR) is calculated using the following formula:

$$DOR=PLR/NLR$$

In this study, we use the positive likelihood ratio (PLR) and the negative likelihood ratio (NLR) to evaluate the predictive performance of our model for preeclampsia. The equations used to calculate PLR and NLR express the frequency of preeclampsia in individuals who are predicted by the model to have preeclampsia compared to those who are predicted not to have preeclampsia:

$$PLR=Sensitivity/(1-Specificity)$$

$$NLR=(1-Sensitivity)/Specificity$$

Considering the diversity in populations, predictors, and algorithms across the included ML models, our objective was to generalize findings to broader clinical contexts. Therefore, following the recommendation of Borenstein et al [37], we a priori selected the random-effects model for all meta-analyses, irrespective of the magnitude of statistical heterogeneity (I^2). Specifically, we used the more robust

XSL•FO
RenderX

Hartung-Knapp-Sidik-Jonkman (HKSJ) method for final pooled estimates and interval calculations to ensure the robustness of statistical inferences [38]. The ML models included in this study exhibited substantial variations in sample size and population characteristics, with the I^2 statistic often approaching 100% in larger samples, potentially limiting their ability to effectively distinguish the actual clinical impact of heterogeneity. Therefore, in addition to reporting the 95% CI for pooled effect sizes, this study further calculated the 95% PI. Unlike CIs, which only reflect the precision of the average effect, PIs estimate the expected range of performance when the model is applied in a new, similar clinical setting in the future. This approach provides a more intuitive assessment of the model's clinical applicability and transferability [38]. Since the Meta-DiSC software (The developer is the clinical biostatistics team at Ramón y Cajal Hospital) cannot calculate PIs, we used the *meta* package (version 7.0) [39] in R software (R Foundation for Statistical Computing; version 4.4.2) with the HKSJ method to compute 95% PIs for area under the receiver operating characteristic curve (AUROC), sensitivity, and specificity. For AUROC values without reported SEs, we estimated them based on sample size using the Hanley & McNeil [40] method. External validation is regarded as the "gold standard" for assessing the transportability of models. Therefore, a separate evaluation of the performance of models that use external validation is conducted. Subsequently, the 4 predictive models with the highest and lowest values were excluded to conduct a sensitivity analysis aimed at evaluating the impact of outliers on the sensitivity and specificity of the summary. To reduce conceptual heterogeneity and enhance the interpretability of results, stratification is performed along the following dimensions: sample size (less than 2000 and greater than or equal to 2000); data source (electronic medical records; laboratory biomarkers; omics or imaging; mixed); gestational age window (early pregnancy; midpregnancy and late pregnancy or specific gestational weeks); and validation methods (internal validation and external validation); ML models (logistic regression [LR] and nonlogistic regression), followed by more detailed subgroup analysis (LR, extreme gradient boosting [XGBoost], random forest [RF], and support vector machine [SVM]) based on nonlogistic regression; types of predictive variables (demographic information; biological genetic markers; laboratory tests; demographic information and laboratory tests); and the number of predictive variables (less than 10 and greater than or equal to 10). Handling of missing data (extraction and synthesis). For each study, we recorded how missing data were handled and classified methods into 5 categories, namely listwise deletion, single-value imputation (eg, mean and median), multiple imputation, other (eg, random subset iterations), and not reported. When multiple approaches were mentioned, we coded the method used for the primary model. We summarize the overall distribution in the results of "Inclusion of Study Characteristics in the Paper" and discuss implications for comparability and generalizability. Subgroup analyses will be conducted on the included studies to evaluate the performance of ML methods in predicting preeclampsia across different clinical scenarios. Subgroup Analysis discusses the capabilities of different ML algorithms in predicting preeclampsia. Additionally, meta-regression was used to investigate the sources of heterogeneity. Given the extreme heterogeneity (I^2>99%) observed across studies and the lack of standardized threshold reporting (eg, fixed false-positive rates), hierarchical or bivariate models often fail to converge or yield unstable estimates. Therefore, we prioritized univariate random-effects models using the HKSJ adjustment for pooling sensitivity and specificity separately. This method is demonstrated to provide more robust coverage probabilities for CIs in the presence of substantial heterogeneity compared to standard DerSimonian-Laird [41] methods.

Results

Literature Screening

After removing duplicate entries, a total of 284 papers were evaluated. Of these, 284 papers were evaluated through abstract screening, which was subsequently followed by a full-text evaluation of 88 papers. This process culminated in the identification of 26 papers [8-33] that satisfied the overall inclusion criteria. The literature screening procedure and its outcomes are depicted in the related Figure 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for study selection. CNKI: China National Knowledge Infrastructure; PE: preeclampsia.

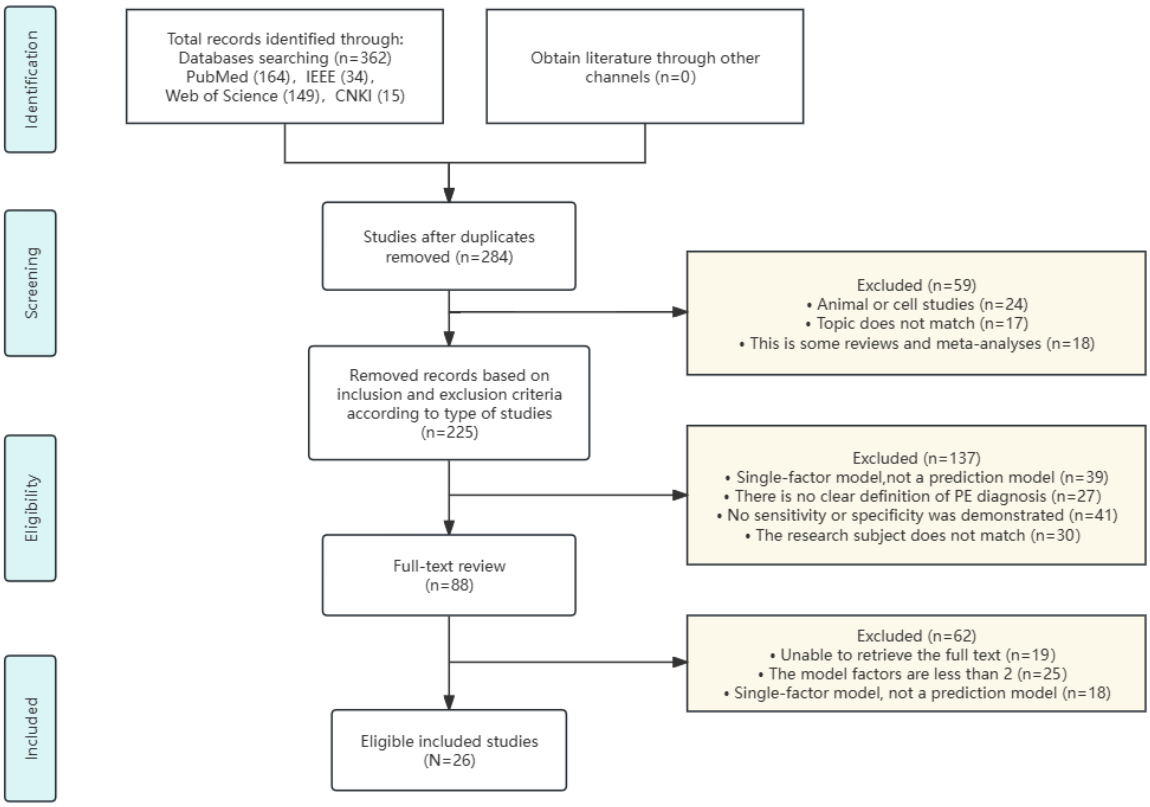

Inclusion of Study Characteristics in the Paper

The literature included in this study spans from 2019 to 2025 and consists of 23 English papers [8-10,12-26,28-30,32,33] and 3 Chinese papers [11,27,31]. When a study presented more than 2 models, the top 2 models demonstrating the best performance were selected based on a comprehensive evaluation of metrics, such as AUC, sensitivity, and specificity, culminating in the inclusion of 31 models from 26 papers [8-33]. The data sources for ML predominantly consisted of clinical electronic health records, community research cohorts, and self-administered questionnaires. The overall sample sizes in the studies examined showed considerable variation, fluctuating between 53 and 62,562 cases, while the count of predictors in the ultimate models ranged from 3 to 50. Among all the studies, 20 [8,11,13,14,16-18,20-32] conducted internal validation, while 6 [8,21,24,28,30,32] performed external validation. To assess model performance, the AUC, sensitivity, and specificity emerged as the most frequently used metrics. Among the 26 studies [8-33] reviewed, 5 (19.2%) studies [8,17,25,26,33] were prospective cohort studies, 17 (65.4%) studies [10,11,13-16,18,19,21-24,27-29,31,32] were retrospective cohort studies, 2 (7.7%) studies [9,20] were case-control studies, 1 (3.8%) study [30] was a retrospective case-control study, and 1 (3.8%) study [12]was a multicenter study. Regarding model approaches, of the 31 models included, 3 were LR. Among the remaining 28 models, there were 5 RF, 4 XGBoost, 4 Elastic-net, 3 neural network (NN), 3 SVM, 2 light gradient boosting, 2 AdaBoost, 1 k-nearest neighbor, 1 Naive Bayes, 1 stochastic gradient boosting, 1 CatBoost, and 1 voting classifier. In terms of handling missing data, 8 studies [11,18,22,24-27,29] opted to delete cases with missing data, 7 studies [9,12,14-16,19,23] used mean imputation to address the missing values, 3 studies [13,17,31] used multiple imputation techniques, 1 study [21] implemented random selection of data subsets for multiple iterative analyses, while the remaining 7 studies [8,10,20,28,30,32,33] did not explicitly report the presence of missing values. Such variation limits comparability and external transportability of performance metrics and increases uncertainty around calibration and threshold transfer. The specific details of the models are presented in Table 1.

XSL•FO
RenderX

Table 1. Construction of the risk prediction model for preeclampsia.

Literature and modeling method	Model performance			Sample size (modeling; internal validation; external validation)	Missing data		Predictors
	AUC[a]	Sensitivity	Specificity		Quantity (PCS[b])	Handling method	
Ansbacher et al [8]							
FfNN[c]	0.816	0.533	0.9	30437/10000/20352	—[d]	—	10 predictors: maternal age, maternal weight, maternal height, interpregnancy interval, ethnicity, medical history (such as chronic hypertension, diabetes, etc), uterine artery pulsatility index, mean arterial pressure, placental growth factor, and pregnancy-associated plasma protein-A.
Araújo et al [9]							
LGB[e]	0.9	0.95	0.79	132/—/—	—	Mean imputation	3 predictors: neutrophil count, mean corpuscular hemoglobin, and aggregate index of systemic inflammation.
Chen et al [10]							
SVM[f]	0.88	0.87	0.76	166/—/—	—	—	7 predictors: IL-17, IL-21, IL-22, IL-10, transforming growth factor-β, placental alkaline phosphatase, and lysosome-associated membrane protein 3.
Chen et al [11]							
CB[g]	0.983	0.8881	0.9848	1325/398/—	—	Delete	18 predictors: BMI, systolic blood pressure, diastolic blood pressure, number of pregnancies, mean corpuscular hemoglobin concentration, bacteria (urinalysis), glycocholic acid, high-density lipoprotein, potassium, sodium, phosphorus, uric acid, urine protein, creatinine, direct bilirubin, low-density lipoprotein, gestational age≥34 weeks, and family history of hypertension.
Giménez et al [12]				597/—/—	—	Mean imputation	6 predictors: gestational age, history of chronic hypertension, Soluble FMS-like Tyrosine Kinase-1, placental growth factor, N-terminal pro-brain natriuretic peptide, and uric acid.
PTB-RF[h]	0.901	0.796	0.91				
RF[i]	0.941	0.775	0.949				
Jhee et al [13]							
SGB[j]	0.924	0.603	0.991	7704/3302/—	25	Multiple Imputation	14 predictors: systolic blood pressure, serum urea nitrogen, serum creatinine, platelet count, serum potassium level, white blood cell count, serum calcium level, and urinary protein.
Kaya et al [14]							
XGBoost[k]	0.767	0.6	0.833	53/20/—	—	Mean imputation	8 predictors: maternal age, BMI, smoking status, history of diabetes, history of gestational diabetes, mean arterial pressure, and history of previous preeclampsia.
Kovacheva et al [15]				1125/—/—	—	Mean imputation	7 predictors: maternal age, BMI, systolic blood pressure, diastolic blood pressure, uric acid, history of kidney disease, and SBP PRS[m].

XSL•FO
RenderX

Literature and modeling method	Model performance			Sample size (modeling; internal validation; external validation)	Missing data		Predictors
	AUC[a]	Sensitivity	Specificity		Quantity (PCS[b])	Handling method	
LR[l]	0.83	0.85	0.66				
XGBoost	0.91	0.96	0.44				
Li et al [16]							
XGBoost	0.955	0.789	0.93	3759/191/—	—	Mean imputation	38 predictors: maternal age, BMI, mean blood pressure, abdominal circumference, gravidity, parity, history of preeclampsia, history of previous cesarean section, interpregnancy interval, primipara, multiple gestation, assisted reproductive technology, heart disease, pregestational diabetes, thyroid disease, kidney disease, autoimmune disease, mental illness, uterine fibroids, adenomyosis, uterine malformation, history of epilepsy, family history of hypertension, hemoglobin, white blood cell count, platelet count, creatinine, fasting blood glucose, total cholesterol, high-density lipoprotein, low-density lipoprotein, total protein, albumin, bile acids, uric acid, total bilirubin, direct bilirubin, and gamma-glutamyl transferase.
Li et al [17]							
VC[n]	0.831	0.77	0.769	3715/929/—		Multiple Imputation	16 predictors: maternal age, height, prepregnancy weight, primiparity, mode of conception, family history, smoking status, history of preeclampsia, history of chronic hypertension, history of chronic kidney disease, history of diabetes, history of systemic lupus erythematosus/antiphospholipid syndrome, mean arterial pressure, uterine artery pulsatility index, pregnancy-associated placental protein a, and placental growth factor.
Lv et al [18]							
XGBoost	0.963	0.917	0.894	832/208/—	—	Delete	6 predictors: prepregnancy BMI, gravidity, mean arterial pressure, smoking, alpha-fetoprotein, and conception method.
Marić et al [19]							

XSL•FO
RenderX

Literature and modeling method	Model performance			Sample size (modeling; internal validation; external validation)	Missing data		Predictors
	AUC[a]	Sensitivity	Specificity		Quantity (PCS[b])	Handling method	
EN[o]	0.79	0.452	0.919	5245/—/—	—	Mean imputation	55 predictors: maternal age, height, weight, ethnicity, number of fetuses, mean systolic blood pressure, mean diastolic blood pressure, maximum systolic blood pressure, maximum diastolic blood pressure, history of preeclampsia, chronic hypertension, type 1 and type 1 diabetes, gestational diabetes, obesity, assisted reproductive technology, diagnosis of autoimmune diseases, kidney disease, anemia, antiphospholipid syndrome, sexually transmitted diseases, hyperemesis gravidarum, headache, migraine, poor obstetric history, high-risk pregnancy, protein and glucose in urine, platelet count, red blood cells, white blood cells, creatinine, hemoglobin, hematocrit, monocytes, lymphocytes, eosinophils, neutrophils, basophils, Rh blood type, gastric acid, rubella, chickenpox, hepatitis B virus, syphilis, gonorrhea, aspirin, nifedipine, aldomet, labetalol, insulin, glyburide, prednisone, azathioprine, Plaquenil, heparin, levothyroxine, doxylamine, and acyclovir.
Melinte-Popescu et al [20]							
NB[p]	0.98	0.963	0.964	163/70/—	—	—	14 predictors: age, BMI, smoking status, interpregnancy interval, use of assisted reproductive technology, pregestational diabetes, chronic hypertension, history of kidney disease, personal or family history of preeclampsia, placental growth factor, pregnancy-associated plasma protein A, placental protein 13, uterine artery pulsatility index, and mean arterial pressure.
Munchel et al [21]							
AB[q]	0.964	0.88	0.92	113/11/448	—	Randomly select a subset of data for multiple iterative analyses.	49 predictors circulating transcripts in blood: immunomodulatory, fetal development, angiogenesis, and extracellular matrix remodeling.
Roque et al [22]							
LR	0.976	0.9	0.951	35706/8927/—	—	Delete	11 predictors: platelet count, white blood cell count, lymphocyte percentage, monocyte percentage, red blood cell count, red cell distribution width, platelet distribution width, band neutrophil percentage, red cell distribution width, hematocrit, and maternal age.
Sandström et al [23]							

XSL•FO
RenderX

Literature and modeling method	Model performance			Sample size (modeling; internal validation; external validation)	Missing data		Predictors
	AUC[a]	Sensitivity	Specificity		Quantity (PCS[b])	Handling method	
LR	0.67	0.282	0.9	62562/6256/—	—	Mean imputation	36 predictors: gestational age at first visit, maternal age, BMI, mean arterial pressure, capillary blood glucose level, urine protein, hemoglobin level, history of miscarriage, history of ectopic pregnancy, history of infertility treatment, family status, country of birth, smoking history, smoking status at registration, use of snuff in the first trimester of pregnancy, use of snuff during pregnancy, alcohol consumption in the 3 months before registration, alcohol consumption habits at the time of pregnancy registration, family history of preeclampsia, infertility, family history of hypertension, previous diabetes, chronic hypertension, chronic kidney disease, cardiovascular disease, endocrine disease, history of thrombosis, history of mental illness, history of epilepsy, Crohn/ulcerative colitis, lung disease or asthma, hepatitis, gynecological disease or surgery, recurrent urinary tract infections, and blood type.
Sufriyana et al [24]							
RF	0.86	0.7	0.89	23201/20975/GEV[r]:1322, TEV[s]: 90	301	Delete	13 predictors: age, family role, parity, type of work, infectious diseases, endocrine, nutritional and metabolic diseases, circulatory system diseases, immune-related diseases, ophthalmic diseases, urogenital diseases, skin and subcutaneous tissue–related diseases, breast-related diseases, digestive system–related diseases, and skin-related diseases.
Tiruneh et al [25]							
RF	0.84	0.76	0.79	33767/14475/—	66	Delete	13 predictors: maternal age, ethnicity, prepregnancy/early pregnancy BMI, history of preeclampsia in previous pregnancies, primiparity, history of gestational diabetes, pre-existing hypertension, diabetes, family history of hypertension and diabetes, family history of preeclampsia, renal disease, smoking history, and polycystic ovary syndrome.
Torres et al [26]				1068/914/—	78	Delete	13 predictors: placental growth factor, mean arterial pressure, uterine artery pulsatility index, BMI, antiphospholipid syndrome, previous preeclampsia, previous diabetes, smoking status, natural conception, Other drug use (such as cocaine and heroin), systemic lupus erythematosus, chronic hypertension, and maternal age.
all-EN	0.778	0.501	0.9				
EPE-EN[t]	0.963	0.882	0.9				
PPE-EN[u]	0.897	0.765	0.9				
Wang et al [27]							
KNN[v]	0.9	0.7142	0.926	516/172/—	—	Delete	7 predictors: urine protein, urine conductivity, alkaline phosphatase, serum uric acid, lactate dehydrogenase, mean corpuscular hemoglobin concentration, and amylase.

XSL•FO
RenderX

Literature and modeling method	Model performance			Sample size (modeling; internal validation; external validation)	Missing data		Predictors
	AUC[a]	Sensitivity	Specificity		Quantity (PCS[b])	Handling method	
Wang et al [28]							
AB	0.8775	0.7271	0.9	25709/77713/1760	—	—	20 predictors: maternal age, maternal BMI, regularity of maternal menstrual cycle, vomiting and nausea during pregnancy, previous miscarriages, preterm births, history of hypertension during pregnancy, hypertension, diabetes, chronic hypertension, history of drug allergies, maternal smoking history, previous delivery history, nutritional status during pregnancy, maternal ethnic background, history of hypertension, history of diabetes, glycated hemoglobin, and albumin.
Xue et al [29]							
SVM	0.93	0.67	0.999	800/160/—	—	Delete	50 predictors: diabetes mellitus, thrombotic diseases, systemic lupus erythematosus, antiphospholipid syndrome, renal diseases, assisted reproductive technology, obstructive sleep apnea syndrome, prepregnancy BMI>30 kg/m², age>35 years, multiple pregnancy, primipara, history of eclampsia or preeclampsia, Albumin, Alanine aminotransferase, Aspartate aminotransferase, Alkaline phosphatase, Complement C1q, Calcium, Creatinine, C-reactive protein, Cystatin C, Gamma-glutamyl transferase, Globulin, Triglycerides, Total cholesterol, High-density lipoprotein cholesterol, Low-density lipoprotein cholesterol, Lipoprotein(a), Apolipoprotein A1, Apolipoprotein B, Small dense low-density lipoprotein, Total protein, Total bile acid, Total bilirubin, Direct bilirubin, Uric acid, Urea, Phosphorus, Absolute Lymphocyte count, Absolute neutrophil count, Platelet count, NEU/LYM ratio, PLT/LYM ratio, Prothrombin time, Prothrombin activity, Activated partial thromboplastin time, Fibrinogen, D-Dimer, Fibrin degradation products, Thrombin time.
Yu et al [30]							
RF	0.96	0.87	0.91	404/1384/899	—	—	12 predictors: maternal age, BMI, parity, medical history (chronic hypertension, preeclampsia, systemic lupus erythematosus, antiphospholipid syndrome), mode of conception; cfDNA profile indicators: Fos-related antigen 2 (FOSL2), calcium/calmodulin-dependent protein kinase kinase 2 (CAMKK2), G1/S-specific cyclin-D1 (CCND1), Inositol 1,4,5-trisphosphate receptor type 1 (ITPR1), Protein kinase A catalytic subunit beta (PRKACB), Protein Wnt-7b (WNT7B), Voltage-dependent L-type calcium channel subunit beta-2(CACNB2), Nuclear respiratory factor 1 (NRF1), Fms-related tyrosine kinase 3 ligand (FLT3LG), Epidermal growth factor (EGF).
Zheng et al [31]							

XSL•FO
RenderX

Literature and modeling method	Model performance			Sample size (modeling; internal validation; external validation)	Missing data		Predictors
	AUC[a]	Sensitivity	Specificity		Quantity (PCS[b])	Handling method	
LGB	0.964	0.849	0.927	1609/483/—	—	Multiple imputation	12 predictors: urine specific gravity, uric acid, mean corpuscular hemoglobin concentration, globulin, platelet distribution width, potassium ion, age, family history of hypertension, systolic blood pressure, diastolic blood pressure, pulse, and gestational age≥34 weeks.
Zhou et al [32]				432/197/288	—	—	19 predictors: mRNA markers: Albumin, Fibrinogen Alpha Chain, Leptin, Insulin-Like Growth Factor Binding Protein 5, Alpha-1 Antitrypsin, S100 Calcium Binding Protein A9, Apolipoprotein A1, Thyroid Stimulating Hormone Beta Subunit, miRNA markers: MIR130A, MIR144, MIR19B1, MIR215, MIR376C, MIR27A, MIR106A, MIR33A, Inc ENA markers**:** Macrophage Migration Inhibitory Factor, Assisted Reproductive Technology, Mean Arterial Pressure.
AvNN[w]	0.91	0.63	0.93				
SVM	0.93	0.47	0.99				
Zhou et al [33]							
CNN[x]	0.883	0.722	0.934	1138/—/—	—	—	8 predictors: Retinal fundus image score, Prepregnancy BMI, maternal age, chronic hypertension, diabetes, history of gestational hypertension or preeclampsia, assisted reproductive technology, and autoimmune diseases.

[a]AUC: area under the curve.
[b]PCS: pieces.
[c]FfNN: feed-forward neural network.
[d]not reported.
[e]LGB: light gradient boosting.
[f]SVM: support vector machine.
[g]CB: CatBoost.
[h]PTB-RF: Premature birth - Random Forest.
[i]RF: random forest.
[f]KNN: k-nearest neighbor.
[j]SGB: stochastic gradient boosting.
[k]XGBoost: extreme gradient boosting.
[l]LR: logistic regression.
[m]SBP PRS: systolic blood pressure polygenic risk score.
[n]VC: Voting Classifier.
[o]EN: Elastic-net.
[p]NB: Naive Bayes.
[q]AB: AdaBoost.
[r]GEV: geographic external validation
[s]TEV: temporal external validation
[t]EPE-EN: early onset of preeclampsia Elastic-net.
[u]PPE-EN: Premature birth of preeclampsia Elastic-net.
[v]KNN: k-nearest neighbor.
[w]AvNN: Average Neural Network.
[x]CNN: Convolutional Neural Networks.

XSL•FO
RenderX

Research Quality

We evaluated the potential for bias and the relevance of the prediction models based on the PROBAST checklist, examining a total of 26 [8-33] studies. Among these, 3 (12%) studies [9,11,16] in the participant domain exhibited unclear risk of bias, primarily due to their case-control design, which is inherently associated with a higher risk of selection bias. In the predictor domain, 1 (4%) study [21] was identified as having unclear risk of bias because it used C-RNA transcriptome assays that depend on transcriptome enrichment and high-throughput sequencing, methods that are not typically used in routine clinical testing. In the analysis of bias domains, 8 (31%) studies [9,10,14,16,21,29,32,33] demonstrated unclear risk of bias, mainly due to insufficient sample sizes, unclear methodologies for addressing missing data, and uncertainties regarding the management of overfitting risks. Furthermore, 1 (4%) study [22] was classified with a high risk of bias as all data were sourced from a single hospital, despite the volume of data, failing to represent a multicenter or stratified analysis. Overall, the bias risk was determined to be unclear for 9 (35%) studies [9-11,14,16,21,29,32,33]. The applicability ratings were moderate for 4 (15%) studies [10,11,21,33], high for 1（4%) study [22], and low for the remaining studies [8,9,12-20,23-32], as detailed in Table 2. For the remaining details, see Table S2 in the Multimedia Appendix 2.

XSL•FO
RenderX

Table 2. Risk of bias and applicability assessment using PROBAST (Prediction Model Risk of Bias Assessment Tool).

Study and year	ROB[a]				Overall bias rating	Overall applicability rating	External validation
	Participants	Predictors	Outcome	Analysis			
Ansbacher et al [8], 2022	Low	Low	Low	Low	Low	Low	Yes
Araújo et al [9], 2024	Unclear	Low	Low	Unclear	Unclear	Low	No
Chen et al [10], 2022	Low	Low	Low	Unclear	Unclear	Unclear	No
Chen et al [11], 2023	Unclear	Low	Low	Low	Unclear	Unclear	No
Garrido-Giménez et al [12], 2023	Low	Low	Low	Low	Low	Low	No
Jhee et al [13], 2019	Low	Low	Low	Low	Low	Low	No
Kaya et al [14], 2024	Low	Low	Low	Unclear	Unclear	Low	No
Kovacheva et al [15], 2023	Low	Low	Low	Low	Low	Low	No
Li et al [16], 2021	Unclear	Low	Low	Unclear	Unclear	Low	No
Li et al [17], 2024	Low	Low	Low	Low	Low	Low	No
Lv et al [18], 2025	Low	Low	Low	Low	Low	Low	No
Marić et al [19], 2020	Low	Low	Low	Low	Low	Low	No
Melinte-Popescu et al [20], 2023	Low	Low	Low	Low	Low	Low	No
Munchel et al [21], 2020	Low	Unclear	Low	Unclear	Unclear	Unclear	Yes
Roque et al [22], 2024	Low	Low	Low	High	Low	High	No
Sandström et al [23], 2019	Low	Low	Low	Low	Low	Low	No
Sufriyana et al [24], 2020	Low	Low	Low	Low	Low	Low	Yes
Tiruneh et al [25], 2024	Low	Low	Low	Low	Low	Low	No
Torres et al [26], 2024	Low	Low	Low	Low	Low	Low	No
Wang et al [27], 2022	Low	Low	Low	Low	Low	Low	No
Wang et al [28], 2024	Low	Low	Low	Low	Low	Low	Yes
Xue et al [29], 2023	Low	Low	Low	Unclear	Unclear	Low	No
Yu et al [30], 2024	Low	Low	Low	Low	Low	Low	Yes
Zheng et al [31], 2021	Low	Low	Low	Low	Low	Low	No
Zhou et al [32], 2024	Low	Low	Low	Unclear	Unclear	Low	Yes
Zhou et al [33], 2023	Low	Low	Low	Unclear	Unclear	Unclear	No

[a]ROB: risk of bias.

The Performance of ML Models in Preeclampsia Prediction

A total of 26 (31 models) studies [8-33] were included. While the pooled estimates demonstrated high average discriminative potential of ML models, substantial between-study heterogeneity was observed, indicating significant context-dependency of model performance. The overall pooled AUROC was 0.91 (95% CI 0.87-0.92; Figure 2). However, its 95% PI ranged from 0.75 to 1.00, suggesting that AUC might decrease to 0.75 in some external validation settings. The pooled sensitivity was 0.81 (95% CI 0.70-0.83; P<.001; I^2=99.6%) In the Figure 3 [8-33], the first author of each study is listed along the Y-axis, the circles represent the point estimates of sensitivity for each model, with the size of the circles being proportional to the weight of the study; the horizontal lines indicate their 95% CIs. The letter Q represents the intersection point of the SROC curve with the inverse diagonal line where “Sensitivity = Specificity.” The diamonds represent the aggregated sensitivity estimates of the models, with their width corresponding to the 95% CI of the aggregated values. The vertical red dashed line represents the 95% CI of the pooled sensitivity. However, this only

XSL•FO
RenderX

represents an average level; the wide 95% PI of 0.32-0.96] reveals potential clinical risks. In certain specific studies or future applications, the sensitivity may be as low as 32%, indicating a substantial risk of missed diagnoses. Similarly, although the pooled specificity was 0.88 (95% CI 0.84-0.94; P<.001; I^2=99.7%; Figure 4 [8-33]), its PI across different contexts was 0.49-0.99, demonstrating a similar lack of consistency in specificity. The other summary metrics were as follows: DOR was 37.67 (95% CI 23.46-60.48); PLR was 8.52 (95% CI 6.43-11.29); NLR was 0.24 (95% CI 0.18-0.34). Additionally, we calculated the Spearman correlation coefficient between the log of sensitivity and the log of (1-specificity), which yielded a result of 0.254 (P=.17), indicating no significant threshold effect in the included studies. This suggests that the observed high heterogeneity (as well as the broad PIs mentioned above) primarily stems from nonthreshold factors (such as differences in predictor selection or population characteristics), rather than merely from variations in cutoff value selection.

Figure 2. Summary Receiver Operating Characteristic (SROC) plot illustrating the dispersion of study results. AUC: area under the curve; SROC: Summary Receiver Operating Characteristic.

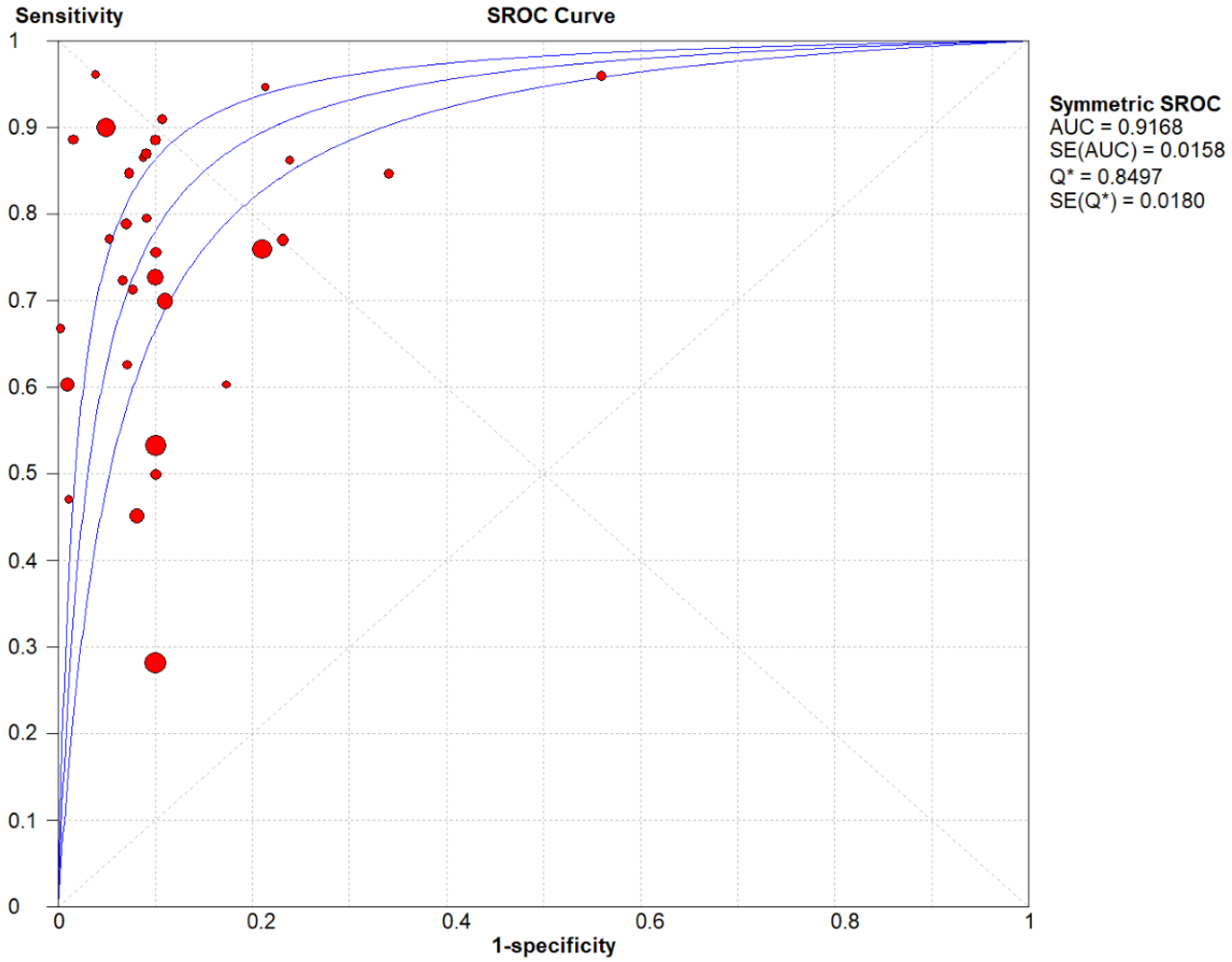

XSL•FO
RenderX

Figure 3. Overall sensitivity of machine learning models for the prediction of preeclampsia [13-17, 19-39].

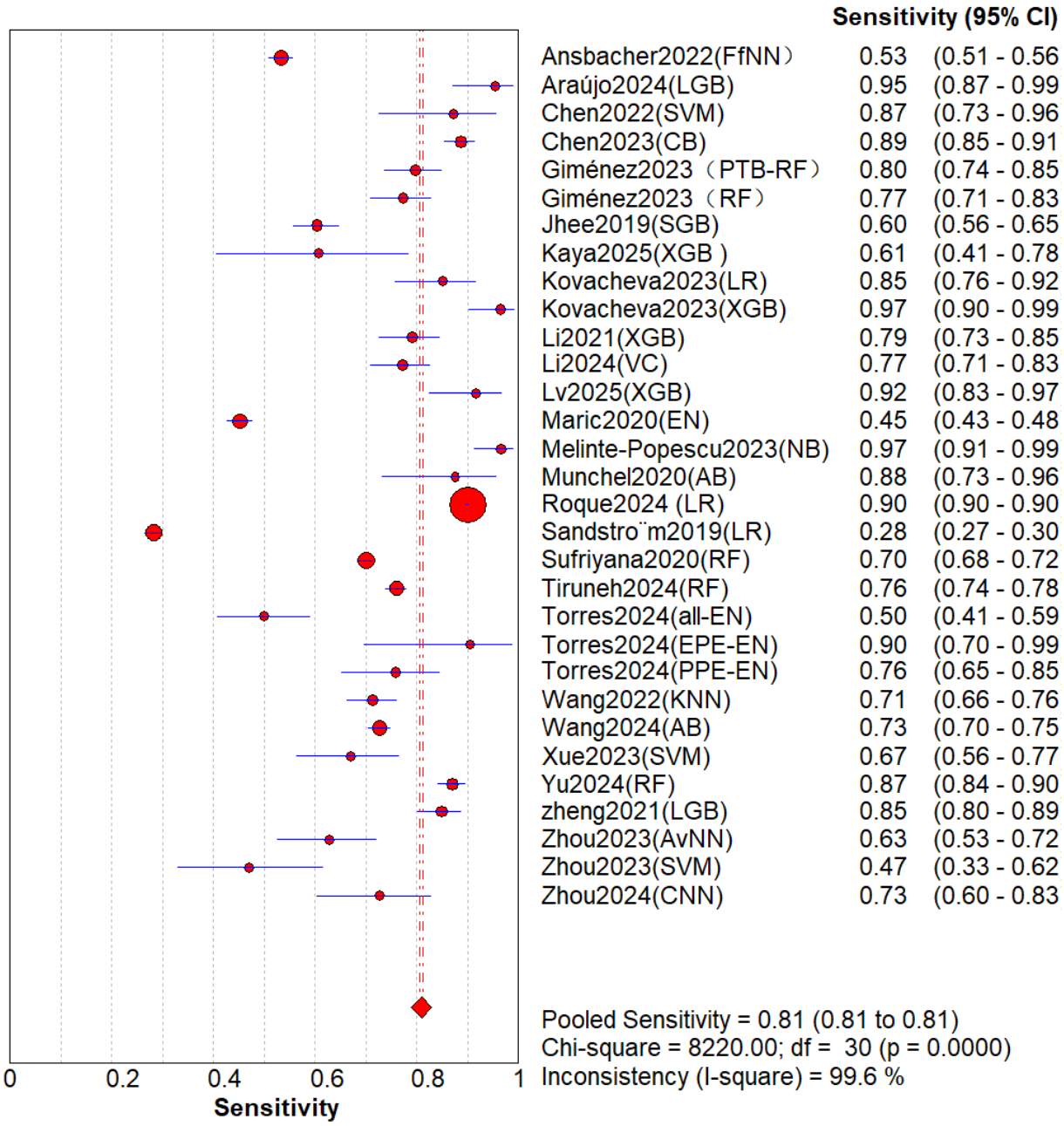

XSL•FO
RenderX

Figure 4. Overall summary specificity of machine learning for the prediction of preeclampsia. [13-17, 19-39].

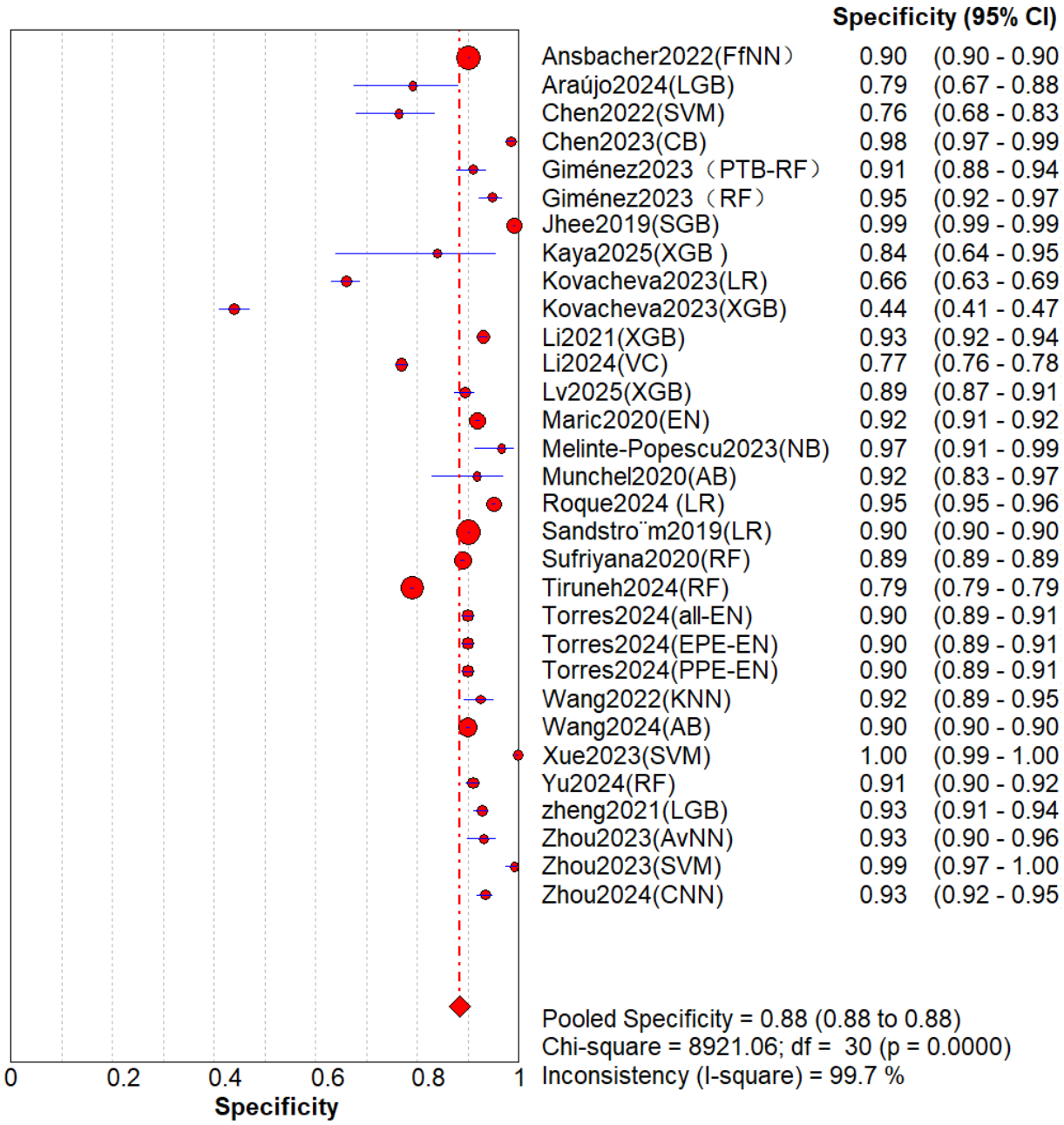

Performance Analysis of External Validation Models

A total of 6 (comprising 7 models) studies [8,17,24,28,30,32] underwent external validation. The analysis revealed that when applied to independent external populations, the models exhibited performance decline with persistent high heterogeneity. Specifically, the pooled AUC was 0.91 (95% CI 0.85-0.95; Figure 5). However, its 95% PI was 0.76-1.00, indicating that the model's discriminative ability might be suboptimal in certain external settings. The pooled sensitivity significantly decreased to 0.68 (95% CI 0.54-0.83; P<.001; I^2=99.6%; Figure 6 [8,21,24,28,30,32]), with a 95% PI of 0.25-0.94. The lower limit of 0.25 indicates that in the worst-case external validation scenario, the model may miss 75% (23/31) of patients, posing an extremely high risk of missed diagnosis. The pooled specificity was 0.90 (95% CI 0.86-0.96; P<.001; I^2=99.7%; Figure 7 [8,21,24,28,30,32]), with a 95% PI of 0.62-0.99. Other indicators included: DOR of 28.21 (95% CI 18.10-43.98; I^2=97.6%); PLR of 7.51; NLR of 0.32. The decrease in sensitivity (from 0.81 in the primary analysis to 0.68) and the extremely low limit of the PI (0.25) strongly confirmed the limited transportability of the model across populations, indicating that direct clinical application requires extreme caution.

XSL•FO
RenderX

Figure 5. Summary Receiver Operating Characteristic (SROC) plot for external validation models. AUC: area under the curve; SROC: Summary Receiver Operating Characteristic.

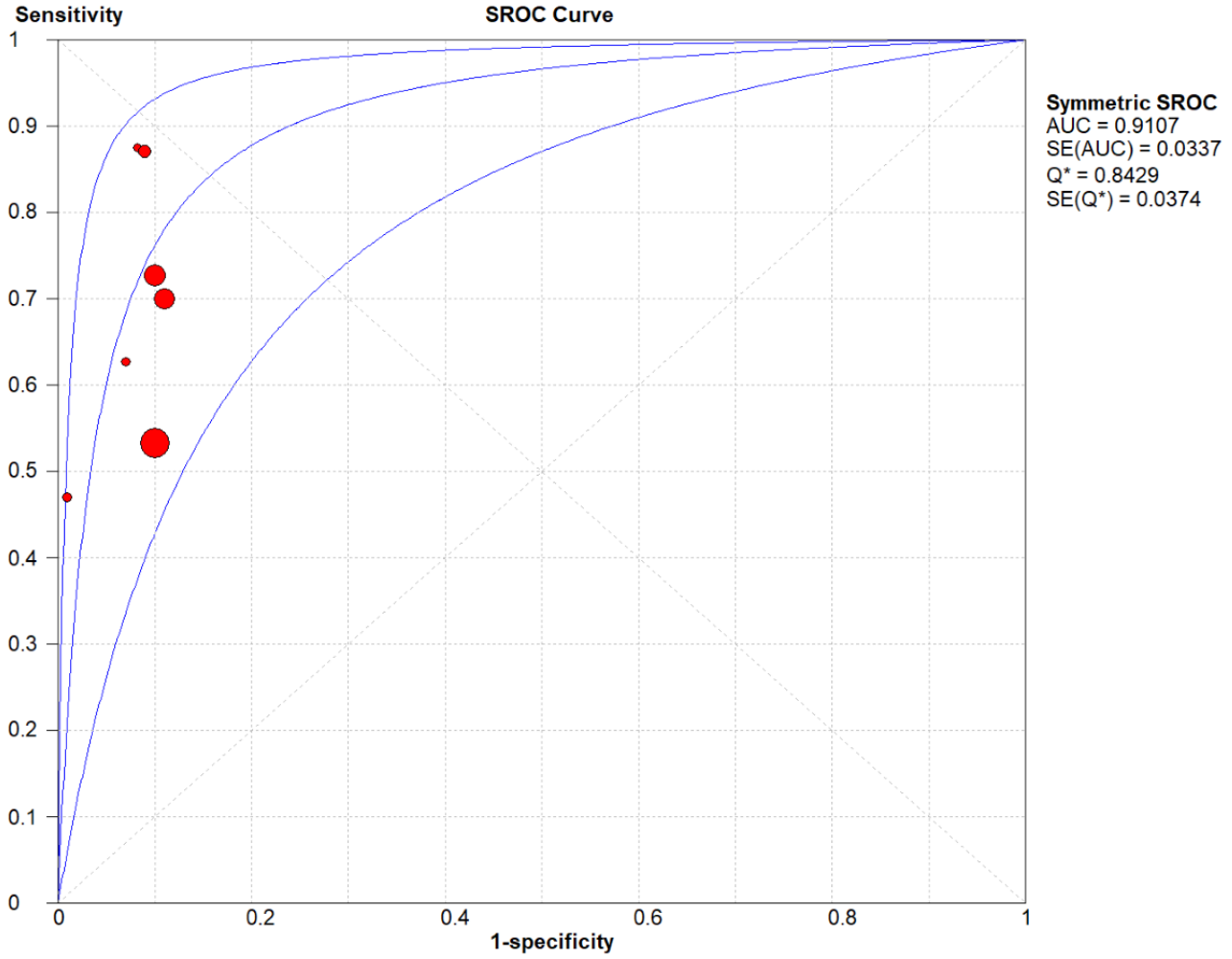

Figure 6. Summary sensitivity of machine learning models for predicting preeclampsia based on external validation [13,27,30,34,36,38].

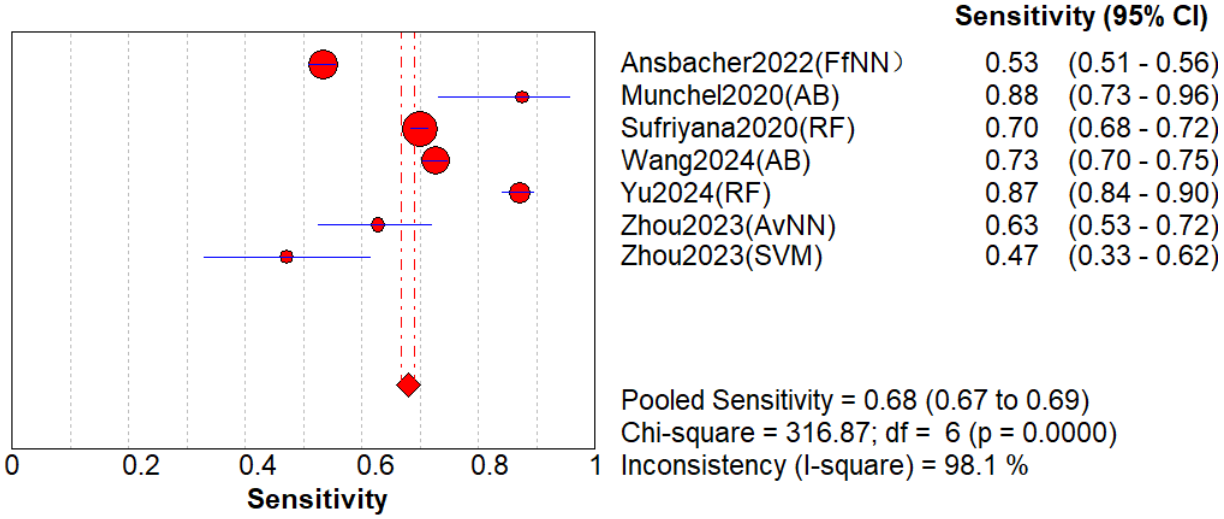

XSL•FO
RenderX

Figure 7. Overall summary specificity of machine learning for predicting preeclampsia [13,27,30,34,36,38].

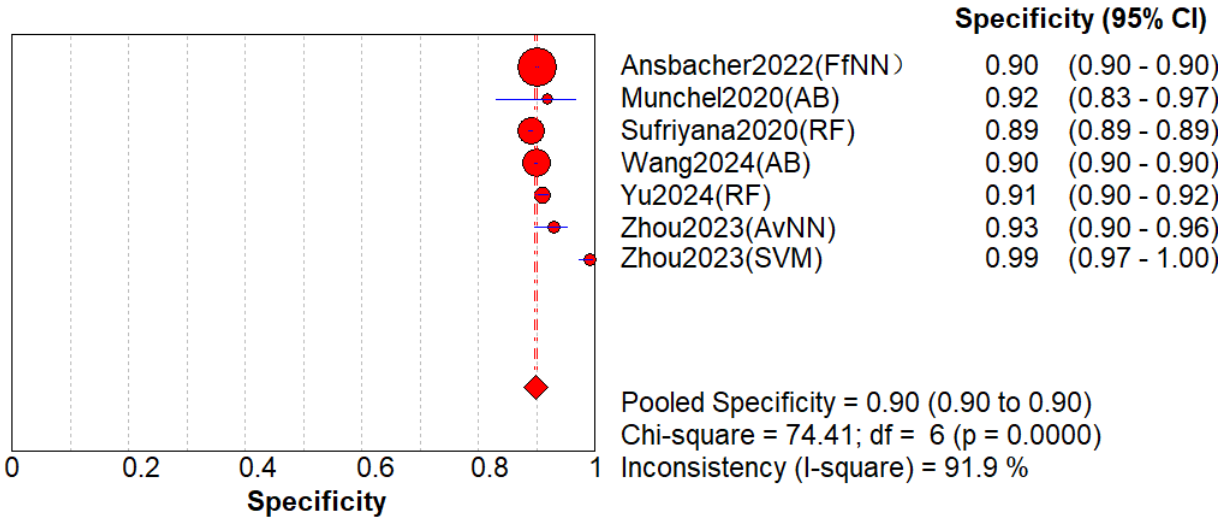

Sensitivity Analysis

After conducting a sensitivity analysis excluding case-control studies in a leave-one-domain-out with 4 (15%) models, the overall summary AUROC is 0.9109 (95% CI 0.8642-0.9390). The summary sensitivity estimate derived from the random-effects meta-analysis is 0.81 (95% CI 0.70-0.83; P<.001; I^2=99.7%), and the summary specificity is 0.88 (95% CI 0.84-0.94; P<.001; I^2=99.7%), as detailed in Figure 8 [8-33]. Consequently, it was concluded that the pooled estimates remained unaffected by the exclusion of outlier values. With an AUC>0.8, the model demonstrated good discriminative ability, but an I^2>75% indicated substantial heterogeneity within most subgroups. To address this issue and gain deeper insights, we undertook a subgroup analysis to investigate the potential sources of this heterogeneity across the studies that were included in our review. Accordingly, we do not interpret a single pooled estimate as “average clinical performance” and instead prioritize subgroup results. In addition, to eliminate the impact of multiple models (derived from the same population) within a single study on statistical independence (unit-of-analysis error), we conducted additional sensitivity analyses by retaining only the model with the highest AUROC from each study (N=26). The results showed that the pooled sensitivity after deduplication was 0.81 (95% CI 0.73-0.87), specificity was 0.88 (95% CI 0.83-0.91), and AUROC was 0.90 (95% CI 0.87-0.93). The above results were highly consistent with the primary analysis (N=31), with no significant differences observed in the CIs, indicating that incorporating different models from the same study did not lead to inflated results or underestimated variance. Therefore, we retained all models in the primary analysis to demonstrate the performance differences among various predictor combinations.

Figure 8. Forest plots of diagnostic performance [13-17, 19-39].

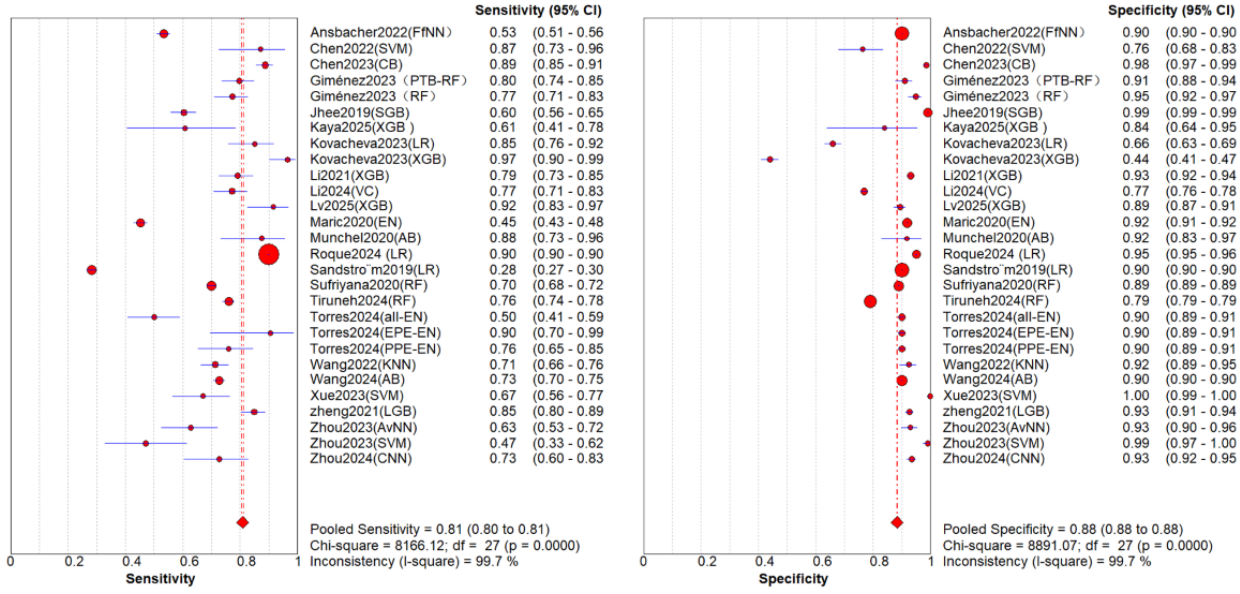

XSL•FO
RenderX

Subgroup Analysis

The comparative results of the subgroup analysis on preeclampsia prediction performance are presented in Table 3; types of ML models, forest plots are shown in Figures S1-S22 in Multimedia Appendix 2. The comparison between subgroups was determined by examining whether the 95% CI of the AUC overlapped. Nonoverlapping intervals indicated statistical significance while overlapping intervals indicated no statistical significance. Data were derived from electronic health records, high-throughput omics, and hybrid sources. Subgroup analysis indicated that models based on hybrid data demonstrated superior performance, followed by those using electronic health records and high-throughput omics. However, considerable heterogeneity was observed, and the 95% CIs extensively overlapped across the 3 data types, suggesting no statistically significant differences among them. The "pregnancy window" refers to the index timing window during which predictors were collected or model discrimination was performed. Models constructed using third-trimester data showed better performance with low heterogeneity. Nonetheless, overlapping 95% CIs across models indicated no statistically significant differences among pregnancy window subgroups. Regarding validation strategies, internally validated models outperformed externally validated ones, albeit with high heterogeneity. Subgroup analysis revealed overlapping 95% CIs between the 2 validation types, implying that the difference was not statistically significant. Regarding sample size, the subgroup analysis results showed that models with smaller sample sizes outperformed those with larger sample sizes, exhibiting lower heterogeneity. However, since the 95% CI overlapped, the differences between sample size subgroups were not statistically significant. Regarding the adopted model, nonlogistic regression prediction models outperformed logistic regression prediction models. Further analysis was conducted on nonlogistic regression models with 3 or more instances in each model category, revealing that neural networks exhibited the best predictive performance with an AUC of 0.9966 (95% CI 0.9772-1.0000) and the lowest heterogeneity. The difference in model performance was statistically significant when compared to elastic net models, but not statistically significant when compared to other models. Regarding the type of predictive variables, prediction models constructed solely using laboratory test indicators achieved the highest predictive performance with an AUC of 0.9463 (95% CI 0.9097-0.9820) and the lowest heterogeneity. Nevertheless, when compared to models built with alternative indicators, the difference in performance was not statistically significant. For the number of predictor variables used in model building, models with 10 or more variables exhibited higher predictive performance with an AUC of 0.9204 (95% CI 0.8671-0.9737), but the difference was not statistically significant compared to models with fewer than 10 variables.

XSL•FO
RenderX

Table 3. Subgroup analysis results.

Grouping	Number of prediction models (PCS[a])	AUC[b] (95% CI)	I^2 (%)	*P* value
Entire study	31	0.9168 (0.891-0.950)	99.6	<.001
Sample size				
<2000	16	0.9361 (0.9079-0.9643)	90.9	<.001
≥2000	15	0.9109 (0.8501-0.9717)	99.8	<.001
Data source				
Mixed	14	0.9154 (0.8713-0.9595)	99.6	<.001
EHR[c]	12	0.9126 (0.8430-0.982)	99.4	<.001
Omics	4	0.9406 (0.8898-0.9914)	95.3	<.001
Pregnancy window				
Early	10	0.9406 (0.7853-1.0000)	95.2	<.001
Mid	4	0.9304 (0.8965-0.9643)	77.2	.004
Late	3	0.9665 (0.9314-1.0000)	71.4	.03
Specific	14	0.9138 (0.8805-0.9471)	99.1	<.001
Machine learning model				
Logistic regression	3	0.9044 (0.6857-1.0000)	100.0	<.001
Nonlogistic regression	28	0.9171 (0.8871-0.9471)	97.6	<.001
RF[d]	5	0.8917 (0.7950-0.9884)	95.8	<.001
SVM[e]	3	0.9068 (0.7623-1.0000)	88.2	<.001
XGBoost[f]	4	0.9177 (0.8500-0.9854)	89.9	<.001
EN[g]	4	0.9419 (0.9125-0.9713)	93.9	<.001
NN[h]	3	0.9966 (0.9772-1.0000)	84.7	.001
Predictor variable type				
Demographic information	10	0.8754 (0.8315-0.9193)	99.4	<.001
Biological genetic marker	3	0.9300 (0.8375-1.0000)	96.8	<.001
Demographic information and laboratory tests	13	0.9275 (0.8665-0.9885)	98.4	<.001
Laboratory testing	5	0.9463 (0.9097-0.9820)	95.8	<.001
Number of predictor variables				
<10	10	0.9124 (0.8855-0.9393)	86.6	<.001
≥10	21	0.9196 (0.8665-0.9727)	99.8	<.001

[a]PCS: piece.

[b]AUC: area under the curve.

[c]EHR: electronic health record.

[d]RF: random forest.

[e]SVM: support vector machine.

[f]XGBoost: extreme gradient boosting.

[g]EN: elastic network.

[h]NN: neural network.

Meta-Regression Analysis

Due to the significant heterogeneity observed among the studies, a meta-regression analysis was conducted. The meta-analysis focused on various factors, including sample size, country of publication, type of ML model, year of publication, study design, study quality, and predictors, as detailed in Table 4. Variables were systematically removed based on the magnitude of their *P* values, and separate meta-regression analyses were performed

XSL•FO
RenderX

for each variable. The results indicated that the source of heterogeneity among the studies was primarily associated with the research quality, as illustrated in Table 5.

Table 4. Meta-regression analysis.

Variable	β coefficient (SE)	*P* value	RDOR[a](95% CI)
Constant	3.547 (1.3356)	.01	—[b]
Sample size	1.075 (0.5388)	.06	0.34 (0.11-1.05)
Country	–0.322 (0.4741)	.50	0.72 (0.27-1.94)
ML[c] method	0.588 (0.7387)	.43	1.80 (0.39-8.37)
Year	0.007 (0.4578)	.99	1.01 (0.39-2.61)
Design	–1.435 (0.8047)	.09	0.24 (0.04-1.27)
Quality	0.672 (0.4076)	.11	1.96 (0.84-4.57)
Predictive	0.773 (0.6075)	.22	2.17 (0.61-7.67)
Validation type	–0.318 (0.4797)	.51	0.73 (0.27-1.97)

[a]RDOR: relative diagnostic odds ratio.

[b]Not applicable.

[c]ML: machine learning.

Table 5. Meta-regression analysis after excluding *P* values from largest to smallest.

Variable	β coefficient (SE)	*P* value	RDOR[a] (95% CI)
Constant	2.398 (0.5879)	<.001	—[b]
Quality	0.800 (0.3951)	.05	2.23 (0.99-5.00)

[a]RDOR: relative diagnostic odds ratio.

[b]Not applicable.

Discussion

Principal Findings

This systematic review identified 31 ML models for preeclampsia prediction. Our primary finding highlights a critical paradox. While models demonstrate high average discriminative potential (pooled AUROC 0.91), they exhibit extreme heterogeneity (I^2>99%) and limited transportability. The wide 95% PI for sensitivity (0.32-0.96) warns that a model performing perfectly in development may miss nearly 70% of cases when applied to a new population. This "context dependence" is further confirmed by the performance drop in external validation studies (pooled sensitivity of 0.68), suggesting that current high AUROCs largely reflect internal fit rather than universal clinical effectiveness.

To investigate the sources contributing to this heterogeneity (as well as the wide PIs), our subgroup analysis revealed several key factors. In the subgroup analysis of all 31 models, we observed that their predictive performance was better when the sample size was small (less than 2000 cases), which contradicts the conventional understanding that "larger sample sizes lead to better predictive performance" [42]. The analysis may be significantly influenced by confounding factors, such as study design (eg, case-control studies) and research type—especially considering the very high AUC of the elastic net (AUC=0.963 for Torres et al [26]; AUC=0.96 for Yu et al [30]). Therefore, careful discernment is required, and one should not hastily interpret this as indicating superior predictive performance of models with smaller sample sizes. Regarding predictor types, laboratory test indicators exhibit superior predictive performance, as the core pathological mechanisms of preeclampsia include placental perfusion disorders, endothelial dysfunction, oxidative stress, and inflammatory responses [43]. Laboratory indicators can directly reflect pathological states, while demographic information provides only indirect risk assessments.

Among the ML models analyzed in this study, including RF, SVM, NN, and Elastic-net, the NN model demonstrated the highest predictive performance (AUC=0.99, 95% CI 0.98-1.00), surpassing traditional ML methods, such as LR, RF, and extreme gradient boosting. This analysis may be attributed to the complex etiology of preeclampsia, a pregnancy complication characterized by multiple pathological processes. The intricate, multidimensional interactions inherent in preeclampsia are challenging to capture comprehensively using linear models. In contrast, NN models are well-equipped to model nonlinear relationships and higher-order variable interactions, which more accurately reflect the pathological characteristics of preeclampsia [44]. Compared to traditional methods, NN can automatically extract features and assign weights to input variables without the need for extensive manual variable screening, demonstrating particular advantages in handling high-dimensional data [45]. Moreover, NN models can integrate

XSL•FO
RenderX

multisource heterogeneous data, such as demographic information, laboratory indicators, and biological genetic markers, thereby adapting to the increasingly complex trends in clinical data.

Higher predictive performance is observed when the number of predictors is equal to or greater than 10. This indicates that using a greater number of predictors helps to more comprehensively reflect disease status, significantly enhancing the model's predictive performance. This is especially true for nonlinear algorithms, which are better equipped to capture interaction effects and underlying patterns.

Nonstandardized handling of missing data means that AUC; concordance index and calibration may not be directly comparable across studies; in particular, listwise deletion or simple imputation combined with restricted case-mix and threshold tuning can inflate discrimination and understate uncertainty. We therefore recommend at minimum (1) transparent reporting of missingness (overall and by variable) and the primary imputation strategy; (2) preferential use of multiple imputation or model-based methods, with minimal recalibration (slope and Brier) and decision-curve analysis during external validation; and (3) reporting confusion matrices under fixed thresholds and top-N% triage plus subgroup robustness (GA window; outcome definitions and sites) to enhance interpretability for clinical and digital health use.

Strengths and Limitations

First, regarding methodological rigor and transparency, we strictly adhered to the PRISMA guidelines for reporting, and the research protocol has been preregistered in the international prospective systematic review registry PROSPERO (CRD420251005830). This ensures that the research objectives and methods are predetermined, thereby minimizing reporting bias. Second, concerning the comprehensiveness of the literature search, our search strategy exhibits significant interdisciplinary characteristics. We not only searched mainstream medical databases such as PubMed and CNKI, but also included IEEE Xplore and Web of Science to ensure a comprehensive capture of ML models published in the fields of engineering technology and computer science. This is critical for a topic that bridges clinical medicine and artificial intelligence, avoiding potential omissions of models that might occur if only medical databases were searched. Third, regarding the reliability of data processing, the entire process of literature screening and data extraction in this study was conducted independently by 2 researchers, with any discrepancies resolved through discussion or by involving a third researcher as an adjudicator. This "dual review" process is considered the gold standard for systematic reviews, ensuring the accuracy of data extraction. Fourth, in terms of the professionalism of quality assessment, we used the PROBAST tool, which is currently recommended by international authorities and specifically designed for predictive model research, rather than traditional diagnostic test evaluation tools, such as QUADAS-2 (Whiting and colleagues [46]). PROBAST enables us to thoroughly assess the risk of bias and applicability of the models across 4 key domains, including participants, predictive factors, outcomes, and analysis, which is more in-depth and relevant than previous reviews. Finally, regarding the prudence of analysis, this study recognizes the common pitfall of "performance overestimation" in meta-analyses of predictive models. Therefore, we clearly identified models lacking external validation and conducted an independent meta-analysis of studies that reported external validation. This approach allowed us to more accurately assess the transportability of the models in real-world applications, leading to the conclusion that they are "highly context-dependent," which is a more cautious and clinically realistic interpretation, avoiding overinterpretation of the aggregated AUROC.

Our study has several limitations that should be considered when interpreting the findings. First, and most critically, is the issue of threshold heterogeneity and optimistic bias. As detailed in the "Methods" section, the performance metrics were synthesized from study-specific "optimal thresholds." This precluded the use of threshold-independent summary measures from a bivariate model and means our pooled sensitivity and specificity are likely inflated compared to what would be achieved with a prespecified, clinically relevant cutoff. The wide PIs we report are, in part, a quantification of this inflation risk. Future primary studies should report performance at multiple, clinically justified thresholds to facilitate more meaningful meta-analysis. Second, related to the above, our statistical synthesis approach was necessitated by the data characteristics. The extreme heterogeneity and lack of threshold standardization made the preferred bivariate modeling approach unfeasible. While our use of univariate HKSJ models with PIs is a robust alternative that honestly communicates uncertainty, it does not model the correlation between sensitivity and specificity. Our subgroup and meta-regression analyses help explore sources of heterogeneity, but residual confounding is likely. Third, our search, though comprehensive, may have missed studies in other languages or in nonindexed repositories. Furthermore, we did not formally assess for publication bias using funnel plots or statistical tests, as these methods are less established and interpretable for diagnostic accuracy data with high heterogeneity. Therefore, our results may be influenced by the preferential publication of studies with positive or high-performance results.

Clinical Significance

The methodological choices in this meta-analysis directly inform its central message. The decision to extract data at study-specific "optimal thresholds" inherently captures the optimistic bias prevalent in ML model development. The strikingly wide 95% PI for sensitivity (0.32-0.96), calculated from these potentially inflated estimates, therefore represents a conservative and realistic warning. The true performance in a new setting, after necessary recalibration to a local threshold, could fall to clinically unacceptable levels. This finding powerfully reinforces the principle that external validation is not a mere formality but a fundamental requirement to bridge the gap between algorithmic promise and clinical utility.

Clinical implementation of these models requires a shift from "universal application" to "local adaptation." Given the wide PIs, hospitals should not adopt published models directly. Instead, we recommend a workflow of local validation and recalibration. Future research should prioritize multicenter

XSL•FO
RenderX

external validation over developing new models. Where data sharing is restricted, federated learning offers a promising pathway to train robust models across diverse populations without compromising privacy.

Conclusions

In summary, ML models demonstrate promising potential for predicting preeclampsia, rather than serving as ready-made universal solutions. While pooled analyses indicate high discriminative performance, the substantial heterogeneity (I^2>99%) and wide 95% PIs (sensitivity 0.32-0.96) reveal significant instability in model performance across different clinical contexts. This "context dependency" was further corroborated in external validation analyses. When applied to independent populations, the model not only exhibited decreased aggregate sensitivity but also the lower bound of its PI dropped to 0.25, quantifying the substantial transplantation risk encountered in cross-center applications. Current evidence therefore supports considering ML as a potential screening adjunct, but does not yet justify its use as a universal clinical diagnostic tool. Future research should shift focus from solely pursuing new models with high AUC values to conducting rigorous multicenter external validation and recalibration of existing models, in order to establish their applicable boundaries within real-world clinical pathways.

Acknowledgments

During the preparation of this work, the authors used Gemini (Google) to assist in refining the English language and structure of the manuscript, as well as to generate R code for the statistical analysis (specifically for the Hartung-Knapp-Sidik-Jonkman method and prediction intervals). After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This work was supported by the grants from Liaoning Provincial Science and Technology Program Joint Initiative (Key Research and Development Program Project) and The general project of the Department of Education in Liaoning Province (JYTMS20230103).

Authors' Contributions

Methodology: LL
Formal analysis: LL
Investigation: LL, QZ, YZ, XC, WZ
Resources: QZ
Supervision: JW
Writing–original draft: LL

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA checklist.
[DOCX File , 18 KB - jmir_v28i1e78714_app1.docx]

Multimedia Appendix 2
Search strategies, baseline characteristics of included studies, and additional forest plots for subgroup analyses.
[DOCX File , 15347 KB - jmir_v28i1e78714_app2.docx]

Multimedia Appendix 3
Checklists from the Prediction Model Risk of Bias Assessment Tool (PROBAST).
[DOCX File , 24 KB - jmir_v28i1e78714_app3.docx]

Multimedia Appendix 4
Workflow of the meta-analysis on machine learning models for preeclampsia prediction.
[PNG File , 593 KB - jmir_v28i1e78714_app4.png]

References

XSL•FO
RenderX

1. Mol BWJ, Roberts CT, Thangaratinam S, Magee LA, de Groot CJM, Hofmeyr GJ. Thangaratinam S, Magee LA, De Groot CJM, Hofmeyr GJ. Pre-eclampsia. Lancet. Lancet 2016;387(10022):999-1011. [doi: 10.1016/S0140-6736(15)00070-7] [Medline: 26342729]
2. Roberts JM. Preeclampsia epidemiology(ies) and pathophysiology(ies). Best Pract Res Clin Obstet Gynaecol 2024;94:102480. [doi: 10.1016/j.bpobgyn.2024.102480] [Medline: 38490067]
3. Liu Y, Li N, Li Z, Zhang L, Li H, Zhang Y, et al. Impact of gestational hypertension and preeclampsia on fetal gender: a large prospective cohort study in China. Pregnancy Hypertens 2019;18:132-136. [doi: 10.1016/j.preghy.2019.09.020] [Medline: 31610399]
4. Sylvain MH, Nyabyenda EC, Uwase M, Komezusenge I, Ndikumana F, Ngaruye I. Prediction of adverse pregnancy outcomes using machine learning techniques: evidence from analysis of electronic medical records data in Rwanda. BMC Med Inform Decis Mak 2025;25(1):76 [FREE Full text] [doi: 10.1186/s12911-025-02921-z] [Medline: 39939998]
5. Ranjbar A, Taeidi E, Mehrnoush V, Roozbeh N, Darsareh F. Machine learning models for predicting pre-eclampsia: a systematic review protocol. BMJ Open 2023;13(9):e074705. [doi: 10.1136/bmjopen-2023-074705] [Medline: 37696628]
6. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009;339:b2700 [FREE Full text] [doi: 10.1136/bmj.b2700] [Medline: 19622552]
7. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
8. Ansbacher-Feldman Z, Syngelaki A, Meiri H, Cirkin R, Nicolaides KH, Louzoun Y. Machine-learning-based prediction of pre-eclampsia using first-trimester maternal characteristics and biomarkers. Ultrasound Obstet Gynecol 2022;60(6):739-745 [FREE Full text] [doi: 10.1002/uog.26105] [Medline: 36454636]
9. Araújo DC, de Macedo AA, Veloso AA, Alpoim PN, Gomes KB, Carvalho M, et al. Complete blood count as a biomarker for preeclampsia with severe features diagnosis: a machine learning approach. BMC Pregnancy Childbirth 2024 Oct 01;24(1):628 [FREE Full text] [doi: 10.1186/s12884-024-06821-4] [Medline: 39354367]
10. Chen H, Wang X, Lu K, Yu C, Su M, Kang L, et al. Maternal Th17/treg cytokines and small extracellular vesicles in plasma as potential biomarkers for preeclampsia. Int J Med Sci 2022;19(11):1672-1679 [FREE Full text] [doi: 10.7150/ijms.71047] [Medline: 36237987]
11. Chen Z, Chen Z, Huang Z, Cao Y, Wang H, Wang P. Construction and evaluation of machine learning-based predictive models for early-onset preeclampsia. Journal of Xuzhou Medical University 2023;43(8):571-576 [FREE Full text] [doi: 10.3969/j.issn.2096-3882.2023.08.005]
12. Garrido-Giménez C, Cruz-Lemini M, Álvarez FV, Nan MN, Carretero F, Fernández-Oliva A, et al. Predictive model for preeclampsia combining sFlt-1, PlGF, NT-proBNP, and uric acid as biomarkers. J Clin Med 2023;12(2):431 [FREE Full text] [doi: 10.3390/jcm12020431] [Medline: 36675361]
13. Jhee JH, Lee S, Park Y, Lee SE, Kim YA, Kang SW, et al. Prediction model development of late-onset preeclampsia using machine learning-based methods. PLoS One 2019 Aug;14(8):e0221202. [doi: 10.1371/journal.pone.0221202] [Medline: 31442238]
14. Kaya Y, Bütün Z, Çelik Ö, Salik EA, Tahta T. Risk assessment for preeclampsia in the preconception period based on maternal clinical history via machine learning methods. J Clin Med 2024;14(1):155 [FREE Full text] [doi: 10.3390/jcm14010155] [Medline: 39797241]
15. Kovacheva VP, Eberhard BW, Cohen RY, Maher M, Saxena R, Gray KJ. Preeclampsia prediction using machine learning and polygenic risk scores from clinical and genetic risk factors in early and late pregnancies. Hypertension 2024 Mar;81(2):264-272. [doi: 10.1161/HYPERTENSIONAHA.123.21053] [Medline: 37901968]
16. Li Y, Shen X, Yang C, Cao Z, Du R, Yu M, et al. Novelelectronic health records applied for prediction of pre-eclampsia: machine-learning algorithms. Pregnancy Hypertens 2021;26:102-109 [FREE Full text] [doi: 10.1016/j.preghy.2021.10.006] [Medline: 34739939]
17. Li T, Xu M, Wang Y, Wang Y, Tang H, Duan H, et al. Prediction model of preeclampsia using machine learning based methods: a population based cohort study in China. Front Endocrinol (Lausanne) 2024;15:1345573 [FREE Full text] [doi: 10.3389/fendo.2024.1345573] [Medline: 38919479]
18. Lv B, Wang G, Pan Y, Yuan G, Wei L. Construction and evaluation of machine learning-based predictive models for early-onset preeclampsia. Pregnancy Hypertens 2025;39:101198. [doi: 10.1016/j.preghy.2025.101198] [Medline: 39889366]
19. Marić I, Tsur A, Aghaeepour N, Montanari A, Stevenson DK, Shaw GM, et al. Early prediction of preeclampsia via machine learning. Am J Obstet Gynecol MFM 2020;2(2):100100. [doi: 10.1016/j.ajogmf.2020.100100] [Medline: 33345966]
20. Melinte-Popescu A, Vasilache I, Socolov D, Melinte-Popescu M. Predictive performance of machine learning-based methods for the prediction of preeclampsia-a prospective study. J Clin Med 2023;12(2):418 [FREE Full text] [doi: 10.3390/jcm12020418] [Medline: 36675347]
21. Munchel S, Rohrback S, Randise-Hinchliff C, Kinnings S, Deshmukh S, Alla N, et al. Circulating transcripts in maternal blood reflect a molecular signature of early-onset preeclampsia. Sci Transl Med 2020;12(550):eaaz0131. [doi: 10.1126/scitranslmed.aaz0131] [Medline: 32611681]

XSL•FO
RenderX

22. Roque ACA, Huamanzana J, Mauricio D. Forecasting risk pregnancies in peru using machine learning. In: IEEE.: IEEE; 2024 Presented at: 10th International Conference on Optimization and Applications (ICOA); Oct 17-18, 2024; Almeria, Spain p. 1-6 URL: https://ieeexplore.ieee.org/document/10753982 [doi: 10.1109/ICOA62581.2024.10753982]
23. Sandström A, Snowden JM, Höijer J, Bottai M, Wikström AK. Clinical risk assessment in early pregnancy for preeclampsia in nulliparous women: a population based cohort study. PLoS One 2019;14(11):e0225716 [FREE Full text] [doi: 10.1371/journal.pone.0225716] [Medline: 31774875]
24. Sufriyana H, Wu Y, Su EC. Artificial intelligence-assisted prediction of preeclampsia: development and external validation of a nationwide health insurance dataset of the BPJS Kesehatan in Indonesia. EBioMedicine 2020;54:102710 [FREE Full text] [doi: 10.1016/j.ebiom.2020.102710] [Medline: 32283530]
25. Tiruneh SA, Rolnik DL, Teede HJ, Enticott J. Prediction of pre-eclampsia with machine learning approaches: leveraging important information from routinely collected data. Int J Med Inform 2024;192:105645 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105645] [Medline: 39393122]
26. Torres-Torres J, Villafan-Bernal JR, Martinez-Portilla RJ, Hidalgo-Carrera JA, Estrada-Gutierrez G, Adalid-Martinez-Cisneros R, et al. Performance of machine-learning approach for prediction of pre-eclampsia in a middle-income country. Ultrasound Obstet Gynecol 2024 Mar;63(3):350-357 [FREE Full text] [doi: 10.1002/uog.27510] [Medline: 37774112]
27. Wang H, Wu Y, Guo Y, et al. Construction of a risk prediction model for pre-eclampsia based on routine clinical laboratory indicators. J Clin Lab 2022;40(10):731-736 [FREE Full text]
28. Wang L, Ma Y, Bi W, Meng C, Liang X, Wu H, et al. An early screening model for preeclampsia: utilizing zero-cost maternal predictors exclusively. Hypertens Res 2024;47(4):1051-1062 [FREE Full text] [doi: 10.1038/s41440-023-01573-8] [Medline: 38326453]
29. Xue Y, Yang N, Gu X, Wang Y, Zhang H, Jia K. Risk Prediction Model of Early-Onset Preeclampsia Based on Risk Factors and Routine Laboratory Indicators. Life (Basel) 2023;13(8) [FREE Full text] [doi: 10.3390/life13081648] [Medline: 37629504]
30. Yu Y, Xu W, Zhang S, Feng S, Feng F, Dai J, et al. Non-invasive prediction of preeclampsia using the maternal plasma cell-free DNA profile and clinical risk factors. Front Med (Lausanne) 2024;11:1254467 [FREE Full text] [doi: 10.3389/fmed.2024.1254467] [Medline: 38695016]
31. Zheng J, Zhu R, Yan Y, Zhou Y, Luo Y. Construction of a prediction model for preeclampsia based on machine learning algorithm. Front Endocrinol (Lausanne) 2022;47(8):802-808 [FREE Full text]
32. Zhou S, Li J, Yang W, Xue P, Yin Y, Wang Y, et al. Noninvasive preeclampsia prediction using plasma cell-free RNA signatures. Am J Obstet Gynecol 2023;229(5):553.e1-553.e16 [FREE Full text] [doi: 10.1016/j.ajog.2023.05.015] [Medline: 37211139]
33. Zhou T, Gu S, Shao F, Li P, Wu Y, Xiong J, et al. Prediction of preeclampsia from retinal fundus images via deep learning in singleton pregnancies: a prospective cohort study. J Hypertens 2024;42(4):701-710 [FREE Full text] [doi: 10.1097/HJH.0000000000003658] [Medline: 38230614]
34. Collins GS, Reitsma JB, Altman DG, Moons KGM, TRIPOD Group. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. The TRIPOD Group. Circulation 2015;131(2):211-219 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.114.014508] [Medline: 25561516]
35. Chen R, Wang SF, Zhou JC, Sun F, Wei WW, Zhan SY. [Introduction of the Prediction model Risk Of Bias ASsessment Tool: a tool to assess risk of bias and applicability of prediction model studies]. Zhonghua Liu Xing Bing Xue Za Zhi 2020;41(5):776-781. [doi: 10.3760/cma.j.cn112338-20190805-00580] [Medline: 32447924]
36. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010 Jan;21(1):128-138 [FREE Full text] [doi: 10.1097/EDE.0b013e3181c30fb2] [Medline: 20010215]
37. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: I is not an absolute measure of heterogeneity. Res Synth Methods 2017 Mar;8(1):5-18. [doi: 10.1002/jrsm.1230] [Medline: 28058794]
38. IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open 2016;6(7):e010247 [FREE Full text] [doi: 10.1136/bmjopen-2015-010247] [Medline: 27406637]
39. Plana MN, Arevalo-Rodriguez I, Fernández-García S, Soto J, Fabregate M, Pérez T, et al. Meta-DiSc 2.0: a web application for meta-analysis of diagnostic test accuracy data. BMC Med Res Methodol 2022;22(1):306 [FREE Full text] [doi: 10.1186/s12874-022-01788-2] [Medline: 36443653]
40. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983 Sep;148(3):839-843. [doi: 10.1148/radiology.148.3.6878708] [Medline: 6878708]
41. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
42. Deo RC. Machine learning in medicine. Circulation 2015;132(20):1920-1930 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.115.001593] [Medline: 26572668]
43. Jung E, Romero R, Yeo L, Gomez-Lopez N, Chaemsaithong P, Jaovisidha A, et al. The etiology of preeclampsia. Am J Obstet Gynecol 2022;226(2S):S844-S866 [FREE Full text] [doi: 10.1016/j.ajog.2021.11.1356] [Medline: 35177222]

XSL•FO
RenderX

44. Can a neural network model using common clinical data predict preeclampsia? The ObG Project. 2024. URL: https://www.obgproject.com/2024/11/15/can-a-neural-network-model-using-common-clinical-data-predict-preeclampsia/ [accessed 2025-04-26]
45. Zhao Z, Liu X, Guan Y, Li C, Wang Z. Exploring the potential of cell-free RNA and Pyramid Scene Parsing Network for early preeclampsia screening. BMC Pregnancy Childbirth 2025 Apr 14;25(1):445 [FREE Full text] [doi: 10.1186/s12884-025-07503-5] [Medline: 40229739]
46. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536 [FREE Full text] [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
AUROC: area under the receiver operating characteristic curve
CNKI: China National Knowledge Infrastructure
DOR: diagnostic odds ratio
HKSJ: Hartung-Knapp-Sidik-Jonkman
LR: logistic regression
ML: machine learning
NLR: negative likelihood ratio
NN: neural network
PI: prediction interval
PICO: Population, Intervention, Comparison, and Outcome
PLR: positive likelihood ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction Risk of Bias Assessment Tool
RF: random forest
SVM: support vector machine
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
XGBoost: extreme gradient boosting

Edited by J Sarvestan; submitted 08.Jun.2025; peer-reviewed by T Shi, P Wu; comments to author 01.Sep.2025; revised version received 20.Dec.2025; accepted 20.Dec.2025; published 19.Jan.2026.

Please cite as:
Liu L, Zhu Q, Zong Y, Chen X, Zhang W, Wang J
Machine Learning Prediction Models for Preeclampsia: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e78714
URL: https://www.jmir.org/2026/1/e78714
doi:10.2196/78714
PMID:

©Lu Liu, Qixuan Zhu, Yichi Zong, Xueyuan Chen, Wei Zhang, Jun Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Communication Strategies to Promote Patient Engagement in Telemedicine: Systematic Review

Yangna Hu[1], BA, MA, MEd; Cindy Sing Bik Ngai[1], BA, MPhil, PhD; Rui Jiang[1], BA, MA, DALS

Department of Language Science and Technology, Faculty of Humanities, Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)

Corresponding Author:
Cindy Sing Bik Ngai, BA, MPhil, PhD
Department of Language Science and Technology
Faculty of Humanities
Hong Kong Polytechnic University
PolyU Hung Hom Bay Campus, 8 Hung Lok Road
Hung Hom, Kowloon
Hong Kong, 999077
China (Hong Kong)
Phone: 852 27665111
Email: cindy.sb.ngai@polyu.edu.hk

Abstract

Background: The rapid growth of telemedicine offers convenience, flexibility, and accessibility for patients to have health care services worldwide. To succeed in telemedicine, health care practitioners and telemedicine tools must engage patients through effective communication. However, a research gap exists in understanding the communication strategies used in telemedicine and how they effectively engage patients.

Objective: This study aims to identify communication strategies influencing patient engagement in telemedicine with provider-patient interactions, as well as how included studies evaluate patient engagement through a systematic review.

Methods: We searched the literature comprehensively using 6 databases, Web of Science, PubMed, Scopus, MEDLINE, CINAHL, and Embase, from inception to October 2025. We included empirical, English-language studies that examined communication strategies affecting patient engagement in telemedicine with provider-patient interactions. Studies lacking actual patients or provider-patient interactions in telemedicine were excluded. We used content analysis to identify texts that were related to Theme 1: the communication strategies affecting patient engagement, and Theme 2: evaluation of patient engagement. Coded texts were analyzed to develop subthemes and themes of identified communication strategies. Methods for evaluating patient engagement were summarized. A narrative synthesis was conducted because of heterogeneity across study design and outcomes. We used the Mixed Methods Appraisal Tool to assess the quality of research included in this study.

Results: This study systematically reviewed 34 peer-reviewed articles, revealing 3 overarching themes of effective communication strategies that enhance patient engagement: interpersonal communication strategies, with 6 subthemes (building relationships, supportive attitude, interactive dialogic loop, nonverbal communication, professionalism and accuracy, and tailored communication); team-level communication strategies, with 3 subthemes (training and preparation, teamwork and care coordination, and cultural and linguistic sensitivity); and system-level communication strategies, with 3 subthemes (usefulness of information, ease of use, and data privacy and security). We also found that included studies predominantly used qualitative research methods, such as semistructured interviews and focus groups, to collect patient engagement data.

Conclusions: This review provides an innovative synthesis of communication strategies that promote patient engagement in telemedicine by integrating interpersonal (micro), team (meso), and system-level (macro) perspectives. Unlike previous reviews that focused on single aspects or levels of communication, this study offers a holistic framework that advances theoretical understanding of how multilevel communication strategies collectively shape patient engagement. Practically, the findings offer actionable guidance for health care professionals, telemedicine developers, and policymakers seeking to enhance the quality and sustainability of telemedicine services. In real-world settings, the identified strategies can inform professional training, platform design, and policy development to support patient-centered digital care. This review is the first to systematically bring together communication strategies for patient engagement in telemedicine across all 3 levels. Future research should build on this framework by developing and validating quantitative measures of patient engagement and examining the relationships between communication strategies and telemedicine outcomes.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e85456) doi:10.2196/85456

KEYWORDS

communication strategies; patient engagement; telemedicine; health care services; provider-patient interactions

Introduction

Background

Digitally accessed health care has accelerated globally, prompted not only by the advancement of communication technologies but also by the increasing demand for accessible and efficient care delivery [1,2]. Consequently, the global use of telemedicine services has grown substantially, with an estimated compound annual growth rate of around 24% between 2022 and 2032 [3]. Telemedicine involves the delivery of health care services via the use of ICTs to engage health care providers (HCPs), patients, and caregivers, and improve health care outcomes [4-6]. It offers convenience and flexibility for both patients and providers and reduces medical service costs and patient wait times [7-11]. Furthermore, it significantly contributes to medical resource allocation, improving patient access and helping health care departments in low-resource settings address resource shortages [6,12-14]. A study analyzed telemedicine consultations in a university-based outpatient telemedicine program and found that the average savings per consultation were 278 miles, 245 minutes, and US $156 [15]. Suzuki and colleagues' study [16] used principal component analysis and cluster analysis to identify countries in Asia and Africa with high potential for telemedicine development, such as Algeria, Egypt, Morocco, and Indonesia. It concluded that telemedicine could address the scarcity of medical resources in these countries.

Despite the great potential of telemedicine to enhance health care accessibility, its adoption remains relatively limited [12,17]. Studies reported that although there are over 300,000 mobile health (mHealth) apps, the user adoption of mHealth apps is low [18,19]. In China, statistics show that telemedicine services account for only 2% of total outpatient services, indicating the underuse of telemedicine services [10]. Except for technology-specific barriers [17,20], a significant factor contributing to this issue is the insufficient communication between patients and service providers, especially on telemedicine platforms where patients or users must initially visit to use these services [21]. Rosler [22] argues that intentional communication skills and tactics can overcome potential barriers to patient engagement within telemedicine and increase patients' connection with providers. Similarly, Fernández Coves and colleagues' study [21] revealed that established means of communication were the most prominent facilitators between patients and service providers at the organizational level of telemedicine adoption in primary care settings.

To succeed on telemedicine platforms, HCPs must effectively engage patients by addressing their needs and preferences [23]. Patient engagement refers to the multidimensional experiences that patients engage with their health management, including cognitive (think), emotional (feel), and behavioral (act) subdimensions of enactment [24,25]. Patient engagement is often used interchangeably with patient activation [26], a concept that focuses on the scenario where patients develop an incremental attitude and have cognitive and behavioral participation in their day-to-day health management [25,27,28]. While there are overlaps between these two concepts, patient engagement is seen as a more holistic consideration, which also includes the psychological involvement during patients' health management situations [25]. In telemedicine settings, patient engagement has been reported to be positively related to high levels of patient satisfaction, improved patient-provider relationships, and increased involvement in health care management [29-33]. For example, in a review study focusing on patient engagement in using hypertension telemedicine tools, Khanijahani et al [34] found that patients' engagement levels were associated with blood pressure reduction levels, their performance in follow-up consultations, and their interests in recording and monitoring their health data.

Despite the many benefits of patient engagement in telemedicine, current studies pay scant attention to the communication strategies used on telemedicine platforms and how they effectively engage users [23,35]. Costa and Serra [36] conducted one of the few review studies examining how communication influences patient engagement in telemedicine contexts. They found that effective communication serves as a cornerstone for improving patient adherence to treatment, whereas communication barriers, such as language barriers, can hinder patient participation in their own care. However, their review primarily focused on reviewing the general role of communication rather than identifying specific effective communication strategies, and it was limited to the field of chronic wound management. Understanding communication strategies is crucial for maximizing the potential of telemedicine, as effective communication in telemedicine is an essential prerequisite for its success, which not only fosters initial engagement but also maintains trust and cooperation and ensures the continued participation of telemedicine [37]. Specifically, communication in telemedicine with access to HCPs is argued to have high potential to stimulate patient engagement [38,39], which remains a favorable way to improve health care outcomes in telemedicine [40-42].

Objectives

Given the rapid growth of telemedicine in health care service delivery and the increasing significance of communication strategies for patient engagement in telemedicine systems [23,35,37], this paper aims to identify the communication strategies promoting patient engagement in telemedicine with HCP-patient interactions by conducting a systematic review of the existing telemedicine studies to explore the effective communication strategies discussed. As such, we propose the following research questions (RQs) to guide our study:

RQ1: What communication strategies have been found or hypothesized to contribute to patient engagement on telemedicine platforms with HCP-patient interactions?

XSL•FO
RenderX

RQ2: How has patient engagement in telemedicine been evaluated in the selected literature?

By synthesizing existing research on crucial communication strategies that enhance patient engagement in telemedicine, this review endeavors to provide HCPs, policymakers, telemedicine tool developers, and researchers with insights to inform the development of more effective telehealth strategies and policies.

Methods

Overview

This study was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [43]. We registered this systematic review on PROSPERO (International Prospective Register of Systematic Reviews; CRD420251053245). This study has been revised and updated from the originally registered PROSPERO protocol to incorporate methodological and reporting improvements based on editorial feedback.

Eligibility Criteria

We included studies if (1) they involved telemedicine using ICTs to deliver health care services, (2) they studied telemedicine tools including HCP-patient interactions, (3) they examined communication strategies influencing patient engagement, (4) they involved real patients or clinical populations who actively engaged with telemedicine, (5) they were peer-reviewed empirical studies, (6) they were published in English, and (7) they were available with full texts.

Articles were excluded if they did not include HCP-patient interactions and only included patients' health care management functions or health care education information in the telemedicine tool. We excluded studies that used standardized, virtual, or fictional patients without actual patient use with the telemedicine platform, as well as studies that focused on improving patient involvement and engagement in health care research. During the screening process, we excluded articles that were not empirical studies and were not published in a peer-reviewed journal, such as conference papers, editorial notes, and book chapters.

Search Strategy

We applied the PRISMA-S (Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension; Multimedia Appendix 1) to guide our search strategy [44] and searched Web of Science, PubMed, Scopus, MEDLINE (via EbscoHost), CINAHL (via EbscoHost), and Embase for relevant studies because these databases ensure that researchers can find comprehensive studies in a wide range of disciplines, including medicine, public health, and social sciences [45-49]. Two experienced librarians specializing in health, social science, and humanities provided professional consultation to help refine and enhance our search strategy. We summarized and searched key terms of "telemedicine," "patient engagement," and "HCP-patient interaction" in the title or abstract, or keywords as shown in Textbox 1. The search strategy combined these three concept blocks using Boolean operators (search strategy: Category 1 AND Category 2 AND Category 3). Apart from using three groups of key terms to identify relevant literature, no language or other restrictions were applied to the search, which was completed on October 31, 2025. The full research strategies applied to the 6 databases are summarized in Multimedia Appendix 2.

Textbox 1. Key terms and search strategy for studies on communication strategies influencing patient engagement in telemedicine involving health care provider (HCP)–patient interactions.

Category 1: telemedicine

eHealth OR e-health OR "electronic health" OR e-consultation OR econsultation* OR e-therapy OR mHealth OR "mobile health" OR telecare OR "tele care" OR telecardiology OR teleconsultation* OR teledentistry OR teledermatology OR telediagnosis OR telehealth OR "tele intensive care" OR "tele ICU" OR telemedicine OR telemonitoring OR telenephrology OR teleneurology OR telenursing OR telepathology OR telepharmacy OR telepsychiatry OR teleradiology OR teleradiotherapy OR telerehabilitation* OR tele-referral* OR "tele referral*" OR telesurgery OR teletherapy OR "virtual care" OR "remote care" OR "virtual medicine" OR "remote rehabilitation*" or "virtual rehabilitation*"

Category 2: patient engagement

"patient activation" OR "patient-centeredness" OR "patient engagement" OR "patient involvement" OR "patient participation"

Category 3: HCP-patient interaction

consultation* OR "online consultation*" OR "video consultation*" OR "video visit*" OR "virtual visit*" OR "remote visit*" OR "televisit*" OR "virtual appointment*" OR "remote appointment*" OR "clinician-patient interaction*" OR "clinician-patient communication*" OR "doctor-patient interaction*" OR "doctor-patient communication" OR "provider-patient interaction*" OR "provider-patient communication" OR "patient-provider interaction*" OR "patient-provider communication" OR "healthcare professional-patient communication" OR "healthcare professional-patient interaction*" OR "HCP-patient interaction*" OR "HCP-patient communication"

Selection Process

A total of 3 authors participated in the selection process. After removing the duplicates, the first reviewer (YH) and the second reviewer (RJ) independently screened all titles and abstracts for eligibility. Any discrepancies regarding study eligibility were resolved through discussion with a third reviewer (CSBN), who served as the adjudicator and made the final decision. During the full-text screening phase, the first reviewer (YH) and second reviewer (RJ) independently assessed all studies, and any disagreements were again resolved in consultation with the third reviewer (CSBN).

Data Collection Process

After the selection process, 2 reviewers (YH and RJ) independently extracted data from each included study using a standardized data extraction table [50] developed for this review. The extraction form was piloted on 7 studies to ensure clarity

XSL•FO
RenderX

and consistency. Extracted data included reference, study setting, country, type and number of participants, recruitment and sampling of participants, participant characteristics, enrollment time, telemedicine type, communication strategies influencing patient engagement, and patient engagement measures. Any discrepancies between reviewers were resolved through discussion. The data extraction table is presented in Multimedia Appendix 3 [39,51-83].

Study Outcomes

The primary outcome domains for this review were (1) communication strategies influencing patient engagement in telemedicine, and (2) methods used to evaluate patient engagement. Communication strategies were defined as any provider-, team-, or system-level communicative actions or decisions intended to enhance communicative effectiveness or compensate for communicative barriers [84-86], thereby shaping patients' cognitive, emotional, or behavioral engagement [25] during telemedicine encounters. Patient engagement measure was defined as any qualitative or quantitative approaches used to assess patients' cognitive, emotional, or behavioral engagement in telemedicine. All results that were compatible with these outcome domains were extracted regardless of the time frame of measurement.

The secondary outcomes extracted from each study included reference information, study setting, country, type and number of participants, recruitment and sampling of participants, participant characteristics, enrollment time, and telemedicine type. The extracted information provided contextual information necessary for interpreting outcome variability across studies.

Quality Assessment

The critical appraisal tool, Mixed Methods Appraisal Tool (MMAT), was used to assess the quality of research included in this study [87]. This tool provides a flexible framework for appraising qualitative, quantitative, and mixed methods studies included in a systematic review [87]. The first reviewer (YH) and the second reviewer (RJ) appraised all the included studies in quality assessment independently, and any disagreements were discussed and resolved with the third reviewer (CSBN) [88]. The product of the quality assessment can be found in the Methodological Quality subsection in the Results section.

Synthesis Methods

We conducted a deductive and inductive qualitative content analysis [89-91] to identify and analyze words, phrases, and texts extracted in the critical primary outcome domain, that is, the communication strategies influencing patient engagement. The extracted content was then examined through thematic analysis to develop sub-themes and overarching themes representing different types of communication strategies. Approaches used to assess patient engagement were also summarized.

An initial codebook for coding the primary outcome domains was developed based on 10 included studies, and new codes were added inductively as the analysis progressed. Multiple coding approaches were applied to ensure comprehensive analysis, since multicoding helps to reveal patterns and associations within the data, providing deeper insights [92,93]. The coding was conducted by two researchers, both with backgrounds in health communication and content analysis methodologies. The first coder (YH) and the second coder (RJ) performed 20% of the initial coding independently. The intercoder reliability was calculated using Cohen κ. The resulting $\kappa=0.82$ indicated almost perfect agreement [94]. The first coder (YH) then coded the rest of the included articles. Finally, the third coder (CSBN) reviewed a portion (4/34, 11.76%) of studies to further assess coding accuracy and ensure consistency. Any discrepancies were discussed and resolved through consensus.

A meta-analysis was not performed due to substantial methodological and contextual heterogeneity across studies. Meta-analysis requires sufficient homogeneity in study design, population, intervention, and outcome measures to ensure meaningful comparability of effect estimates [95]. Given the wide variation in health care contexts, forms of telemedicine, research methods, participant groups, as well as the limited number of comparable quantitative findings in the included research, a narrative synthesis was conducted instead. Consequently, quantitative effect measures (eg, risk ratios, odds ratios, and mean differences), methods to explore statistical heterogeneity (eg, subgroup analysis and meta-regression), sensitivity analyses, assessment of reporting bias due to missing results, and certainty or confidence assessment were not performed, as this review did not aim to statistically pool outcomes across studies. This synthesis approach emphasized thematic patterns in communication strategies and their reported influence on patient engagement.

Results

Study Selection and Study Characteristics

In total, 1726 articles were retrieved from 6 identified databases: Web of Science (n=269), PubMed (n=240), Scopus (n=663), MEDLINE (n=147), CINAHL (n=52), Embase (n=355). These studies were published between 1998 and 2025. After removing 857 duplicates, 869 studies remained to review titles and abstracts, and 126 studies were identified as potentially relevant documents. After the full-text review, 34 studies [39,51-83] were included in this systematic review (Figure 1). Included studies were published between 2015 and 2025, with 28/34 (82.35%) articles published after 2020, reflecting a growing scholarly focus on communication processes within rapidly evolving telemedicine practices. A list of included studies is provided in Multimedia Appendix 3 [39,51-83], and Table 1 presents primary outcomes of data extraction [50].

XSL•FO
RenderX

Figure 1. Flowchart of the literature search and screening process for studies on communication strategies influencing patient engagement in telemedicine involving health care provider–patient interactions (1998-2025).

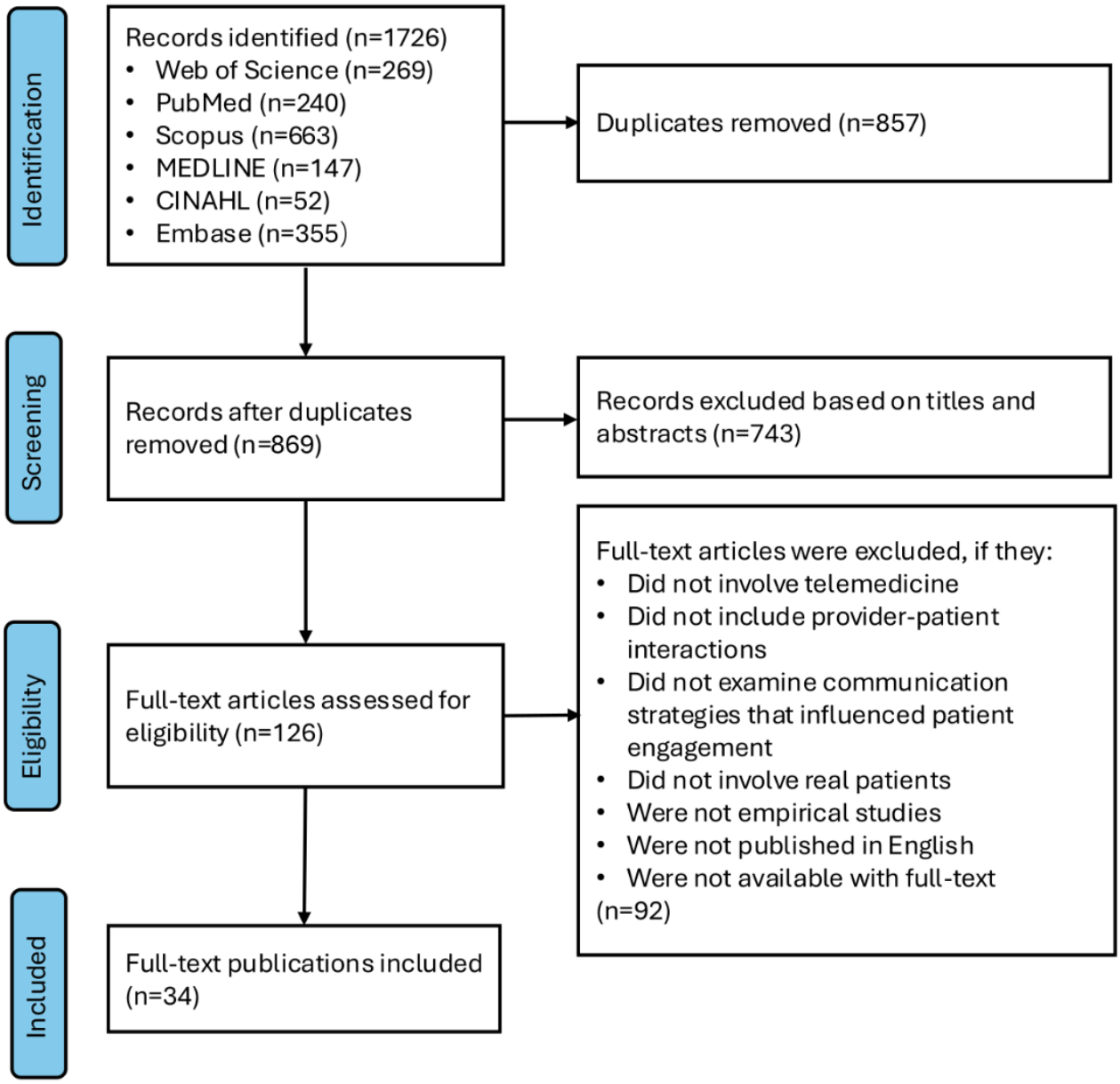

XSL•FO
RenderX

Table 1. Primary outcomes of data extraction on communication strategies influencing patient engagement and patient engagement measures.

Reference, year	Communication strategies influencing patient engagement	Patient engagement measures
Ackerman et al [51], 2020	• Trust-based communication between patients and primary care clinicians; • Using understandable language to provide clear explanations and updates about eConsult decisions.	• No standardized measure; patient engagement was reflected qualitatively via patient accounts of their acceptability of eConsult and feeling involved in care decisions.
Alpert et al [52], 2022	• Using a sincere, empathetic tone and plain language to communicate with patients; • Offering emotional support; • Encouraging patients' participation by using open-ended questions, validating patient input, and fostering dialogue; • Responding promptly to convey accessibility and approachability.	• No standardized measure; patient engagement was reflected qualitatively via clinician accounts of patients' participation, emotional responsiveness, and message interactivity.
Bavngaard et al [53], 2023	• Use of visuality in surroundings, such as showing the medicine bottles, facilitated communication; • Nonverbal communication through gaze direction and smartphone positioning signaled attentiveness and engagement; • Patients' gaze disengagement was interpreted as cognitive engagement in decision-making; • Showing rapport by permitting gaze disengagement from patients.	• No standardized measure; patient engagement was operationalized through the observation and thematic analysis of eight video-recorded consultations, focusing on exploring patients' verbal and non-verbal actions, including attending, contributing, clarifying, and signaling attentiveness.
Björndell [54], 2021	• Listening to patients' thoughts, concerns, and requests; • Guiding and trusting patients in self-examination during video consultations.	• No standardized measure; patient engagement was reflected qualitatively via physicians' accounts of patients' active participation in the consultation process, and patient involvement in decision-making.
Breton et al [55], 2021	• Using visual cues, such as seeing patients' facial expressions, during video visits to enhance communication; • Avoiding the issue of reduced confidentiality of consultations, such as conducting consultations with patients during the patient's grocery time.	• No standardized measure; patient engagement was reflected qualitatively via physicians' perceptions of patients' access, participation, and responsiveness during telemedicine consultations, including comfort, comprehension, follow-up adherence, and involvement in decision-making.
Brodar et al [56], 2022	• Teamwork between departments, including joint virtual visits, interdisciplinary "warm handoffs" between endocrinologists and psychology staff during virtual visits, educating the importance and relevance of consultation and care, and sharing important documents in electronic health records; • Encouraging the provider team to use creative and interactive methods to engage patients, such as playing an online game, using the Zoom Whiteboard feature, and sharing the screen to review materials; • Ensuring staff training in telemedicine.	• No standardized measure; patient engagement was recorded based on the psychosocial screener completion and consultation rates, as well as reflected qualitatively via team members' feedback about patient acceptability of online consults and patient participation.
Caffery et al [57], 2017	• Confusion around issues such as medical liability, privacy, and storage of images was identified as a barrier to patient engagement; • Communication issues, such as a language barrier, between the clinicians and patients hindered engagement.	• No standardized measure; patient engagement was reflected qualitatively via practitioners' perceptions of patient satisfaction, participation in teleconsultations, and continuity of care.
Davoust et al [58], 2025	• Building rapport and trust through open, honest communication; • Visual connections with providers; • Providing tailored communication, such as flexibility in visit modalities to accommodate patient preferences.	• No standardized measure; patient engagement was assessed qualitatively through participants' narratives about their experiences and perceived patient involvement in care.

XSL•FO
RenderX

Reference, year	Communication strategies influencing patient engagement	Patient engagement measures
Day et al [59], 2025	• A consistent, thorough, and mechanistic consultation structure helped engage patients; • Usefulness of information, such as appointment and treatment reminders, increased engagement.	• No standardized measure; patient engagement was reflected qualitatively through semistructured interviews.
Dong et al [60], 2023	• Established clinician-patient relationships influenced engagement.	• No standardized measure; patient engagement was assessed through self-report surveys on patients' engagement with tele-mental health sessions.
Esayed et al [61], 2025	• Building rapport through prior in-person contact and avoiding impersonal communication in telemedicine; • Providing interactive communication through facilitating dialogue and openness.	• No standardized measure; patient engagement was reflected qualitatively through patients' perceptions of their preference for telecare.
Gibson et al [62], 2016	• Interactive communication, such as speaking directly with consultants and getting feedback from them.	• No standardized measure; patient engagement was reflected qualitatively through patient accounts of their involvement in the process of teleconsultations and decision-making.
Grens et al [63], 2022	• Concerns about missing nonverbal cues in video consultations; • Concerns about impersonal telemedicine visits.	• No standardized measure; patient engagement was reflected qualitatively through participant accounts of their involvement in the process of teleconsultations.
Grove et al [64], 2023	• Providing feedback on patient-reported outcomes enhanced patient engagement; • The opportunity to initiate dialogue with providers; • Taking patients seriously and making them feel seen.	• No standardized measure; patient engagement was reflected qualitatively based on patients' perceptions and experiences, such as feelings of support, understanding of their condition, willingness to share information, and overall satisfaction with the remote follow-up.
Higa et al [65], 2021	• Tailoring communication style to meet patient preference; • Building trust-based relationships with patients, such as encouraging patients through text messaging.	• No standardized measure; patient engagement was reflected qualitatively based on multiple data sources, including participants' feedback interview data, answers to open-ended survey questions, the lead researcher's participant observations, and field notes from group meetings, telehealth sessions, and informal interactions with participants, text messages, emails, etc. Engagement outcomes included improvements in diabetes knowledge, frequency of blood glucose monitoring, self-care behaviors, and hemoglobin A1c levels.
Islind et al [66], 2019	• Interactive dialogic loop based on text and links shared via a text chat; • Explaining the reason why health care providers shifted sight and lost eye contact due to screen changes; • Understanding, acknowledging, caring, and trusting patients.	• No standardized measure; patient engagement was reflected qualitatively based on interview data and the researcher's observations.
James et al [67], 2021	• Concerns about missing nonverbal cues during teleconsultations; • Considering patients' multicultural backgrounds and allowing them to bring interpreters to facilitate communication.	• No standardized measure; patient engagement was reflected qualitatively based on nurses' perceptions of patients' need to be seen and respected with cultural sensitivity.
Jensen et al [68], 2023	• Establishing relationships with patients to engage in meaningful conversations.	• No standardized measure; patient engagement was reflected qualitatively through patient accounts of their engagement in care.
Jethwa et al [69], 2022	• Establishing relationships between patients and providers to engage in meaningful conversations; • Having trust and building rapport; • Concerns from patients who do not speak English as a first language; • Ensuring clarity in layman's terms; • Being emphatic when communicating with patients.	• No standardized measure; patient engagement was reflected qualitatively through answers to open-ended questions in a questionnaire, collecting patient preferences for telemedicine.

XSL•FO
RenderX

Reference, year	Communication strategies influencing patient engagement	Patient engagement measures
Jung et al [70], 2023	• Increasing interactions with patients to enhance both patient and staff engagement.	• No standardized measure; patient engagement was primarily observed through participation in daily symptom reporting via mobile/web apps and nurse call follow-ups.
Moore et al [71], 2022	• Prompt responses from providers to show care; • Maintaining established, ongoing patient-provider relationships to foster trust; • Provider knowledge and support regarding portal features; • Useful functions, such as written records to facilitate communication and engagement; • The user-friendly design impacted patients' decisions about how or to what extent they used the portal; • Concerns about the security of the portal.	• No standardized measure; patient engagement was assessed qualitatively through patients' perceptions of their willingness to use telemedicine tools and their preferences for these tools.
Morrison et al [72], 2021	• Ease of use regarding Near Me facilitated continued use of this tool.	• No standardized measure; patient engagement was assessed through feedback collected via surveys, informal verbal feedback during appointments, and participation in improvement cycles, contributing to iterative service refinement.
Olayiwola et al [73], 2018	• Establishing a trust-based provider-patient relationship; • Ensuring responsibilities and roles between clinicians were clearly communicated to patients; • Providing patients with clear explanations of referral processes and allowing communication for clarifications; • Coordination and communication between health care departments; • Cultural-linguistic alignment facilitated acceptance of the electronic consultation and referral; • Potential security and confidentiality concerns hindered engagement.	• No standardized measure; patient engagement was assessed qualitatively through focus groups and survey responses, focusing on perceptions, preferences, and attitudes toward involvement in electronic referral processes.
Osmundsen et al [74], 2015	• Increased knowledge and understanding of patients' disease improved patient engagement.	• No standardized measure; patient engagement was assessed qualitatively through participant questionnaires and interviews, focusing on perceptions of care involvement.
Rodkjær et al [75], 2022	• Using the information patients provide to increase patient engagement and focus on patients' needs.	• No standardized measure; patient engagement was assessed qualitatively through participant questionnaires and interviews, focusing on their perceptions of patient involvement in remote care.
Scruton et al [76], 2025	• Smooth communication between multiple health care providers; • Forming trusting and strong physician-patient relationships; • Giving patients time to process information and ask questions; • Providing emotional support; • Including useful functions or information, such as designing straightforward processes to obtain information, support, and care.	• No standardized measure; patient engagement was assessed qualitatively through patient perceptions of engagement, specifically feeling cared for and their desire to continue virtual options post pandemic.
Selick et al [77], 2023	• Using visual aids and assistive communication tools, choosing appropriate modalities (video over telephone) to support visual and nonverbal cues; • Using nonverbal communication, including body language and facial expressions, to support patient comprehension; • Establishing connections and building trusting relationships with providers.	• No standardized measure; patient engagement was assessed qualitatively based on participant reports of participation, comfort, and involvement during virtual encounters.

XSL•FO
RenderX

Reference, year	Communication strategies influencing patient engagement	Patient engagement measures
Spiess et al [78], 2023	• Concerns about "virtual inhibition," such as missing nonverbal cues and expressing empathy virtually to engage patients from the perspective of providers; • Using artwork to start a conversation and connect with patients meaningfully, helping them feel safe for self-disclosure.	• No standardized measure; patient engagement was primarily assessed through providers' perceptions of patient participation, such as self-disclosure, during virtual visits.
Trondsen et al [79], 2018	• Facilitating immediacy of assessment through real-time visual and verbal interaction; • Building trusting relationships; • Providing a sense of access to the "real" expert (psychiatrist), making patients feel seen and heard and invited to decision-making; • Showing professionalism in clearly justifying and clarifying assumptions and expectations.	• No standardized measure; patient engagement was qualitatively assessed based on participants' perceptions of patient involvement, the sense of being seen and heard, and the feeling of being involved in decision-making during video consultations.
Van Middelaar et al [39], 2018	• Building trusting relationships; • Providing useful information or functions, such as personal reminder and the measurement functionality; • User-friendliness design, such as the clear layout; • Timely and adequate response; • Using a positive and personal tone to motivate the use of telemedicine tools.	• No standardized measure; patient engagement was assessed qualitatively through interview themes addressing initial and sustained use and perceived usability.
Wei and Mao [80], 2023	• Using small talk; • Establishing doctor-patient connections.	• No standardized measure; patient engagement was assessed qualitatively through the analysis of patients' interactional behavior, such as initiation, avoidance, refusal, and topic shifting in the conversation excerpts.
White et al [81], 2024	• Asking questions and encouraging patient participation; • Providing clear explanations and checking for understanding; • Using visual aids; • Interactive communication, such as screen sharing (for video consultations) and sending links or additional resources, during the consultation; • Clarifying information and summarizing key points, engaging patients with health knowledge; • Using small talk to build rapport; • Building trusting relationships.	• No standardized measure; patient engagement was assessed by using multiple research methods, including discourse analysis and conversational analysis to study telehealth consultation recordings, interviewing patients and providers, and conducting patient surveys by asking patients to rate the engagement questions.
Wood et al [82], 2021	• Concerns about diminished rapport from clinicians.	• No standardized measure; patient engagement was primarily assessed qualitatively through participants' perceptions of engagement during telehealth visits.
Zainal et al [83], 2024	• Clear explanation of medical conditions and treatments, and doctors' efficiency was appreciated; • Maintaining eye contact during consultations (valued but not essential in telehealth); • Empathy and respectful communication, and doctors' abilities to address patient concerns patiently and compassionately.	• No standardized measure; patient engagement was assessed qualitatively based on participants' perceptions of patient involvement and participation during consultations.

Methodological Quality

The methodological quality assessment using the MMAT indicated generally high quality across the 34 included studies [39,51-83]. Of these, 22 used qualitative designs and 12 used mixed methods approaches. All studies presented clear research questions or objectives, and the collected data were appropriate for addressing them. Overall, the included studies demonstrated a low risk of bias. A summary of the quality assessment is provided in Table 2.

XSL•FO
RenderX

Table 2. Quality assessment of included studies on communication strategies influencing patient engagement in telemedicine with health care provider–patient interactions using the Mixed Methods Appraisal Tool.

Reference	Year of Publication	All studies		Qualitative studies					Mixed methods				
		S1[a]	S2[b]	1.1[c]	1.2[d]	1.3[e]	1.4[f]	1.5[g]	5.1[h]	5.2[i]	5.3[j]	5.4[k]	5.5[l]
Ackerman et al [51]	2020	✓	✓	✓	✓	✓	✓	✓					
Alpert et al [52]	2022	✓	✓	✓	✓	✓	✓	✓					
Bavngaard et al [53]	2023	✓	✓	✓	✓	✓	✓	✓					
Björndell et al [54]	2021	✓	✓	✓	✓	✓	✓	✓					
Breton et al [55]	2021	✓	✓	✓	✓	✓	✓	✓					
Brodar et al [56]	2022	✓	✓						✓	✓	✓	✓	✓
Caffery et al [57]	2017	✓	✓	✓	✓	✓	✓	✓					
Davoust et al [58]	2025	✓	✓	✓	✓	✓	✓	✓					
Day et al [59]	2025	✓	✓						✓	✓	✓	C[m]	✓
Dong et al [60]	2023	✓	✓						✓	✓	✓	C	✓
Esayed et al [61]	2025	✓	✓	✓	✓	✓	✓	✓					
Gibson et al [62]	2016	✓	✓	✓	✓	✓	✓	✓					
Grens et al [63]	2022	✓	✓						✓	✓	✓	✓	✓
Grove et al [64]	2023	✓	✓	✓	✓	✓	✓	✓					
Higa et al [65]	2021	✓	✓						✓	✓	✓	✓	✓
Islind et al [66]	2025	✓	✓	✓	✓	✓	✓	✓					
James et al [67]	2021	✓	✓	✓	✓	✓	✓	✓					
Jensen et al [68]	2023	✓	✓	✓	✓	✓	✓	✓					
Jethwa et al [69]	2023	✓	✓	✓	✓	✓	✓	✓					
Jung et al [70]	2022	✓	✓						✓	✓	✓	✓	✓
Moore et al [71]	2022	✓	✓	✓	✓	✓	✓	✓					
Morrison et al [72]	2021	✓	✓						✓	✓	✓	✓	✓
Olayiwola et al [73]	2018	✓	✓						✓	✓	✓	✓	✓
Osmundsen et al [74]	2015	✓	✓	✓	✓	✓	✓	✓					
Rodkjær et al [75]	2022	✓	✓						✓	✓	✓	✓	✓
Scruton et al [76]	2023	✓	✓	✓	✓	✓	✓	✓					
Selick et al [77]	2025	✓	✓						✓	✓	✓	✓	✓
Spiess et al [78]	2023	✓	✓	✓	✓	✓	✓	✓					
Trondsen et al [79]	2018	✓	✓	✓	✓	✓	✓	✓					
Van Middelaar et al [39]	2018	✓	✓	✓	✓	✓	✓	✓					
Wei and Mao [80]	2023	✓	✓	✓	✓	✓	✓	✓					
White et al [81]	2024	✓	✓						✓	✓	✓	✓	✓
Wood et al [82]	2021	✓	✓						✓	✓	✓	✓	✓
Zainal et al [83]	2024	✓	✓	✓	✓	✓	✓	✓					

[a]S1: Are there clear research questions?

[b]S2: Do the collected data allow addressing the research questions?

[c]1.1: Is the qualitative approach appropriate to answer the research question?

[d]1.2: Are the qualitative data collection methods adequate to address the research question?

[e]1.3: Are the findings adequately derived from the data?

XSL•FO
RenderX

[f]1.4: Is the interpretation of results sufficiently substantiated by data?

[g]1.5: Is there coherence between qualitative data sources, collection, analysis, and interpretation?

[h]5.1: Is there an adequate rationale for using a mixed methods design to address the research question?

[i]5.2: Are the different components of the study effectively integrated to answer the research question?

[j]5.3: Are the outputs of the integration of qualitative and quantitative components adequately interpreted?

[k]5.4: Are divergences and inconsistencies between quantitative and qualitative results adequately addressed?

[l]5.5: Do the different components of the study adhere to the quality criteria of each tradition of the methods involved?

[m]C: Can't tell.

Results of Syntheses

Based on 34 studies [39,51-83] included in this review, 3 themes of communication strategies were identified as associated with patient engagement: "interpersonal communication strategies," "team-level communication strategies," and "system-level communication strategies." Most studies used qualitative methods, including semistructured interviews and focus groups, to collect information about patient engagement. Studies also used mixed methods to collect patient engagement data, such as combining telemedicine tool use data with patients' qualitative feedback, to understand patient engagement.

Communication Strategies to Promote Patient Engagement in Telemedicine

Based on content analysis of included studies, three synthetic constructs were identified and synthesized; that is, interpersonal communication strategies, team-level communication strategies, and system-level communication strategies, which covered micro-, meso-, and macrolevels of communication strategies to enhance patient engagement in the environment of telemedicine. We developed Figure 2 to illustrate the conceptual framework, with the subsequent content explaining how the 3 levels of communication strategies contribute to patient engagement in telemedicine.

Figure 2. Communication strategies promoting patient engagement in telemedicine, identified in studies involving health care provider–patient interactions across various clinical contexts (2015-2025).

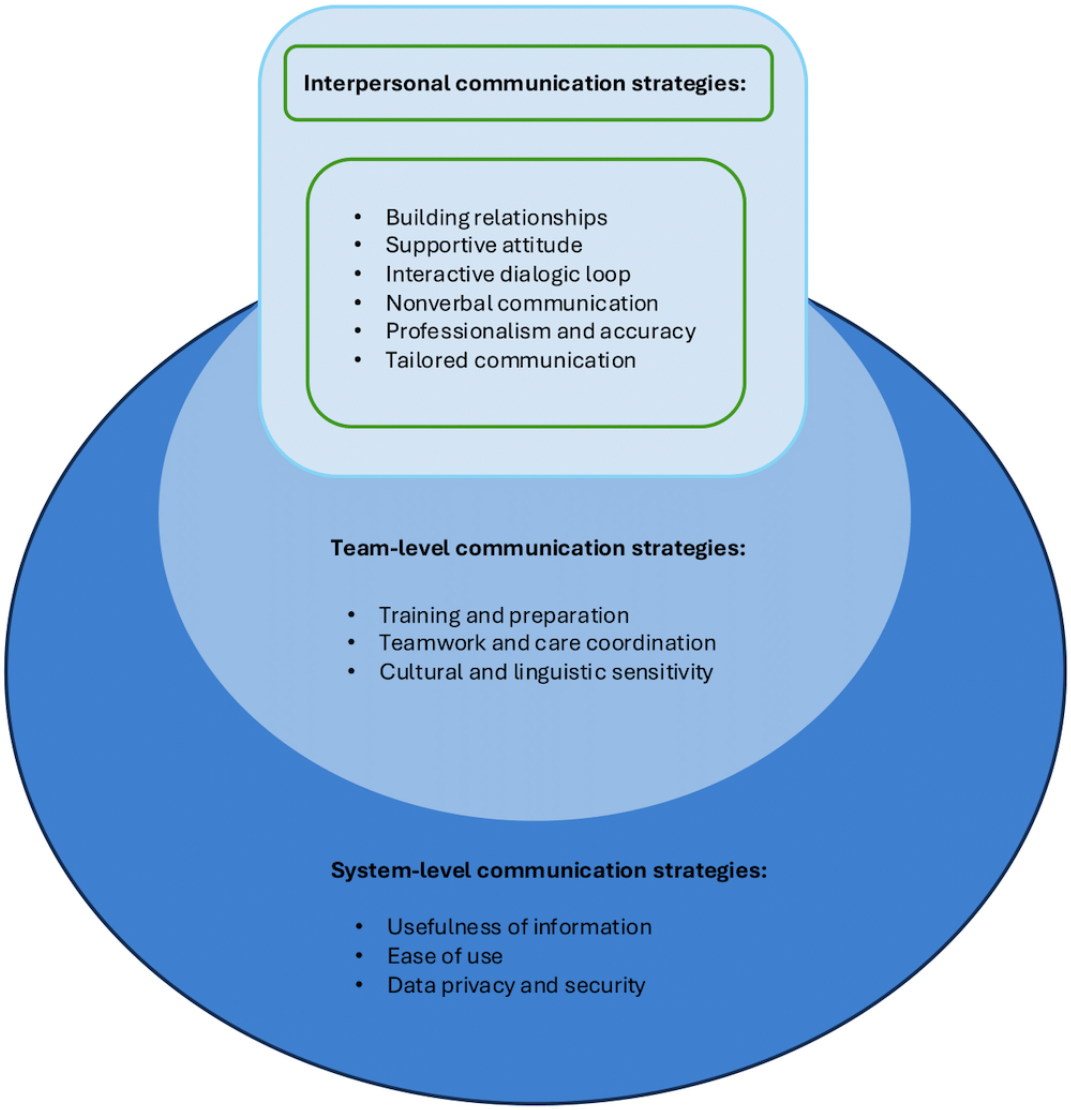

XSL•FO
RenderX

Microlevel: Interpersonal Communication Strategies

At the microlevel, included studies presented prominent interpersonal communication strategies in direct HCP-patient interactions that could improve patient engagement in telemedicine. Specifically, we synthesized 6 subthemes at this level, including building relationships, supportive attitude, interactive dialogic loop, nonverbal communication, professionalism and accuracy, and tailored communication.

The majority of included studies argued that building relationships between HCPs and patients positively impacted patient engagement on diverse telemedicine platforms. Studies found that when patients developed positive and trusting relationships with clinicians, nurses, or other HCPs, they were more likely to accept telemedicine, engage in meaningful conversations with providers, and complete consultation tasks [51,68,71,79,81]. Interpersonal relationship was not only the prerequisite for patients to share their health behaviors, medical concerns, and potential goals [39,78], but also the necessary condition to sustain engagement with telemedicine tools [39]. On the contrary, without established HCP-patient relationships, patients might have concerns about impersonal telemedicine visits [61,63]. Positive interpersonal relationships with HCPs could be built through previous in-person visits [61], visual cues during teleconsultations, such as seeing patients' facial expressions [54,55].

The second subtheme in interpersonal communication strategy is supportive attitude. During teleconsultations, providers were expected to demonstrate a supportive and sincere approach to enhance patient engagement [39,52]. When patients discussed their health behaviors, providers needed to take them seriously, actively listen, understand their concerns, and acknowledge their challenges [54,64,66]. Effective communication to engage patients also involved incorporating "emotional content," such as showing care [71,79], expressing empathy [52,83], and praising patients for positive health behaviors [76]. Such supportive attitudes and actions enabled patients to perceive rapport and genuine support from HCPs [52,79,82], which in turn encouraged greater participation in teleconsultations.

An interactive dialogic loop between HCPs and patients was identified as a crucial component of interpersonal communication strategies that enhanced patient engagement in telemedicine. Direct two-way communication with providers not only strengthened patients' cognitive engagement, such as improving their understanding of disease and increasing access to health knowledge [62,74,81], but also promoted behavioral engagement by encouraging active participation in treatment [52]. During teleconsultations, providers were expected to use a range of communication skills to sustain dialogue and foster engagement [52,61]. These included having small talk [80,81], finding common topics such as artwork to start a conversation and connect with patients [78], explaining the underlying causes of symptoms in detail [52,62], asking open-ended questions [52,53], checking patients' understanding [81], giving patients time to ask questions [76,81], and using chat functions to share screens and links for interactive exchanges [66,81]. In asynchronous communication, prompt and adequate responses to patient messages were essential for stimulating patient engagement, as patients felt reassured by sufficient access to HCPs [39,52]. Conversely, delays or lack of responses often led patients to discontinue platform use [39,64]. Across both synchronous and asynchronous consultations, clear and accessible communication in lay terms was consistently reported to encourage dialogues and strengthen provider-patient interactions [51,52,69].

Nonverbal communication was also found to play a critical role in patient engagement in telemedicine [77]. Studies noted that patients were concerned about the lack of nonverbal cues, such as being able to see what doctors were doing during telephone consultations [67] or missing body language during video consultations [63]. Islind et al [66] and Bavngaard et al [53] further highlighted the role of eye gaze in shaping patient engagement during teleconsultations. Islind et al [66] emphasized that explaining the reason why HCPs shifted their gaze or lost eye contact, often due to screen changes, was important for sustaining engagement. On the other hand, Bavngaard et al [53] underscored the value of allowing flexibility in patients' gaze directionality and even acknowledging momentary gaze disengagement, as brief breaks in eye contact could signal thoughtful and active involvement during consultations. They also highlighted that leveraging visual elements in the surroundings, such as showing the medicine bottles to convey accurate information, could facilitate patients' active participation [53]. Taken together, body language, eye gaze, and the use of visual objects were identified as key nonverbal communication strategies associated with patient engagement.

Within interpersonal communication strategies in telemedicine, patients emphasized the importance of both professionalism and accuracy, as well as tailored communication from HCPs. Zainal et al [83] found that although patients appreciated eye contact during teleconsultations, they placed great value on providers' efficiency and accuracy in communication to avoid errors. Conversely, when providers failed to justify or clearly clarify assumptions and expectations during teleconsultations, patient disengagement was evident [79]. Higa et al [65] highlighted that adapting communication according to patients' individual preferences was crucial for sustaining their engagement. For instance, while some patients responded positively to providers who gave nurturing and encouraging suggestions, others preferred a strict and relentless communication style. Similarly, Davoust et al [58] found that although patients valued a trusting relationship and positive rapport, their levels of comfort varied. Therefore, offering patients flexible options and implementing tailored approaches in telemedicine are essential to accommodate individual preferences and needs.

Mesolevel: Team-Level Communication Strategies

Included studies in this review also presented how communication strategies used by health care teams and organizations could influence patient engagement. A total of 3 subthemes, that is, training and preparation, teamwork and care coordination, and cultural and linguistic sensitivity in health care teams, were synthesized from the mesolevel of communication strategies in telemedicine.

XSL•FO
RenderX

Training and preparation in HCP teams was identified as crucial for patient engagement in telemedicine [56,73]. Patients who experienced difficulty in sustaining attention or "Zoom fatigue" during a remote visit might reduce engagement. To solve this issue, health care organizations should ensure that providers receive communication training in telemedicine, such as using the screen-sharing function to engage patients and playing an online game [56]. Members in provider teams should prepare and provide consistent and clear explanations of the teleconsultation process with patients to have their questions answered, which was reported to impact patients' acceptance of telemedicine tools [73]. Importantly, preparation in HCP teams extended beyond communication training to necessary patient education, particularly around confidentiality. Patients needed guidance on when and how to participate in teleconsultations appropriately, such as avoiding virtual meetings while at the grocery store or driving, so as to maintain privacy and reduce distraction and disengagement [55].

Teamwork and care coordination were identified as essential to influence patients' acceptance and use of telemedicine when they received care from multiple providers. Olayiwola et al [73] reported that clearly defined responsibilities and effective coordination among clinicians were prerequisites for patient acceptance of telemedicine. Similarly, Brodar et al [56] found that teamwork across departments and HCPs, such as joint virtual visits, warm handoff through visit summaries, and sharing key information in electronic health records, helped ensure continuity of care and strengthened patient engagement. Conversely, poor communication among multiple HCPs undermined continuity and reduced care quality, leaving patients feeling neglected and less willing to engage in teleconsultations [76].

For patients from multicultural backgrounds, cultural and linguistic sensitivity within health care teams was crucial to alleviating concerns about using telemedicine [57,67,69,73]. Teams needed to recognize potential cultural and language barriers, particularly when providers interacted with patients who were nonnative English speakers [57,69]. In such cases, involving interpreters during teleconsultations was recommended to help overcome these barriers and support patient engagement [67].

Macrolevel: System-Level Communication Strategies

In addition to identifying communication strategies involving individual HCPs and their teams, this review also examined system-level strategies within telemedicine that influenced patient engagement. A total of 3 key subthemes were identified within this category: usefulness of information, ease of use, and data privacy and security.

Patients reported that the perceived usefulness of information provided by telemedicine platforms, such as self-management tools, personal reminders, access to relevant health information, and a written record function that helped them recall providers' guidance and details from HCP-patient communication, facilitated their engagement [39,71,76]. Ease of use was another critical system-level factor influencing patients' adoption and continued use of telemedicine [71,72]. Platforms with a clear and simple layout and user-friendly features increased acceptability [39,71,72], whereas barriers, such as login difficulties, navigation challenges, or app freezing, discouraged patients from ongoing use and reduced the likelihood of recommending telemedicine tools [71].

Additionally, scholars reported that patients were sometimes hesitant to use telemedicine tools due to concerns about data privacy and security [57,71,73]. Given the sensitive nature of personal health information, some patients expressed worry about how their data were stored and protected [57,71]. Therefore, ensuring secure handling and safeguarding patient information on telemedicine platforms is essential to building trust and encouraging patient engagement.

Evaluation of Patient Engagement

The overwhelming majority of included studies (31/34, 91.18%) used qualitative methods, such as observations, one-on-one interviews, focused groups, asking open-ended questions, and collecting qualitative feedback, to investigate patient engagement from patients and HCPs. Researchers collected qualitative data about patient acceptability of telemedicine, user engagement, patient participation, attention during consultation, and involvement in decision making to evaluate patient engagement. For example, Bavngaard et al [53] conducted a qualitative observational study analyzing 8 video-recorded HCP-patient consultations to explore patient participation during teleconsultations. Van Middelaar et al [39] used semistructured interviews to investigate 20 patients' engagement experience on an online cardiovascular risk management tool. Olayiwola et al [73] collected patient engagement data from both patient focus groups and HCPs' perceptions about patient engagement from their open-ended feedback in an online survey.

Three studies [56,60,81] used mixed methods to evaluate patient engagement. Brodar et al [56] combined quantitative components, that is, health screener completion rate and consultation rate as indicators of engagement, with a qualitative component, that is, participants' feedback through open-ended responses and comments about their telehealth experiences. In Dong and colleagues' [60] telemental health study, patient engagement was measured through quantitative survey items, such as provider-reported ratings of patient engagement, as well as qualitative feedback from providers' open-ended responses describing types of patients that engaged or disengaged in tele-mental health services. White et al [81] used multiple research methods to evaluate patient engagement, including using discourse analysis and conversational analysis to study telehealth consultation recordings, interviewing patients and HCPs, and conducting patient surveys by asking patients to rate the engagement questions, which related to the patient's ability and comfort in communicating and participating in their care from the Telehealth Usability Questionnaire.

Discussion

Principal Findings

The objective of this systematic review was to identify communication strategies that influence patient engagement in telemedicine with the function of HCP-patient interactions. A total of 34 peer-reviewed studies were analyzed, revealing 3

XSL•FO
RenderX

overarching themes of effective communication strategies that enhance patient engagement: interpersonal communication strategies, with 6 subthemes (building relationships, supportive attitude, interactive dialogic loop, nonverbal communication, professionalism and accuracy, and tailored communication); team-level communication strategies, with 3 subthemes (training and preparation, teamwork and care coordination, and cultural and linguistic sensitivity); and system-level communication strategies, with 3 subthemes (usefulness of information, ease of use, and data privacy and security). Furthermore, this review found that qualitative research methods were the most commonly employed approach for assessing patient engagement in the included studies.

Implications Across Micro-, Meso-, and Macrolevel Communication Strategies

At the microlevel, interpersonal communication strategies between HCPs and patients emerged as a cornerstone of enhancing patient engagement in telemedicine. This finding is consistent with previous health care research. For example, Ngai et al [89] highlighted that communication strategies such as maintaining an interactive dialogic loop and demonstrating empathy during two-way HCP-patient communication were crucial for engaging users in health care settings. Similarly, Kwame and Petrucka [96] advanced a patient-centered model, arguing that person-centered communication fosters effective communication and contributes to positive health outcomes. Their model emphasized building meaningful relationships with patients, recognizing their concerns and needs, encouraging self-expression, explaining health conditions and care plans clearly, and engaging in empathetic communication—all of which align with the subthemes of interpersonal communication strategies identified in this review. These insights reinforce the approach of patient-centered communication. Rather than focusing solely on completing consultation tasks, HCPs should view patients as unique individuals with distinct care needs and as collaborators in the care process [65,66,83,96]. Such an approach facilitates effective communication and, ultimately, strengthens patient engagement in telemedicine.

This review identified communication strategies applied not only during synchronous or asynchronous consultations, but also in the form of adequate preparation, particularly at the team level. At the mesolevel, 3 key team-level communication strategies were identified, that is, training and preparation, teamwork and care coordination, and cultural and linguistic sensitivity, which resonate with relational coordination theory [97] and cultural competence model [98]. The relational coordination theory is widely discussed in organizational communication, which emphasizes shared goals, shared knowledge, and mutual respect among team members [97]. This aligns with evidence showing that coordinated teamwork, including team-level communication training in the environment of telemedicine, consistent and clear explanations of the teleconsultation processes, warm handoffs, and joint virtual visits, improved telemedicine acceptance and sustained patient engagement [56,73].

In addition, cultural and linguistic sensitivity emerged as a crucial dimension of team-level communication, consistent with the cultural competence model, which proposes a model of care that includes cultural awareness, knowledge, skills, encounters, and desire [98]. This framework underscores the importance of understanding patients' unique cultural backgrounds and needs, adapting communication styles, addressing language barriers, and involving interpreters where necessary to ensure equitable access and rapport with diverse patient populations [57,67,69,98]. Collectively, these strategies at the team level illustrate that patient engagement in telemedicine is not only an outcome of interpersonal interactions but also the product of well-prepared, well-coordinated, and culturally responsive health care teams.

The identified system-level communication strategies align with previous research on health-related communication on patient engagement. For example, many health communication studies have validated that providing useful content could improve the engagement of the targeted audience [89,99-101]. In addition, Xie and colleagues' [102] and Vasiloglou and colleagues' [103] studies reported that ease of use was a critical reason for users to choose a health app. The identified subthemes of usefulness of information and ease of use at the macrolevel resonate with the technology acceptance model, a leading model in technology acceptance, which argues that users' perceived usefulness and ease of use are primary factors influencing their adoption of new technologies [104].

Moreover, data privacy and security emerged as a critical system-level communication strategy in this review. Given the highly private and sensitive nature of health care data, it is understandable that some patients were reluctant to adopt telemedicine tools due to concerns about confidentiality [105,106]. To address these concerns, telemedicine developers must prioritize robust data protection measures. Suggested strategies include implementing an authentication mechanism [107] and providing patient telehealth "drop-in" kiosks with devices and soundproof space [82].

Advancing the Evaluation of Patient Engagement in Telemedicine

It is surprising to find that the included studies in this review predominantly used qualitative methods, such as semistructured interviews and qualitative feedback, to collect data about patient engagement. Research primarily using quantitative measurements of patient engagement was missing from the included studies. Although 3 studies [56,60,81] used surveys to collect participants' ratings of patient engagement-related items, none of the included studies measured patient engagement in the sense of quantifying engagement through standardized scales. In other words, the quantitative assessment tools for evaluating patient engagement were not unified and standardized. This might be due to a significant lack of clarity regarding the definition and conceptualization of patient engagement, as evidenced by the plethora of terms frequently used interchangeably in this field, as well as the lack of assessment instruments [25].

Not identified in this review, but in a worldwide context, the Patient Activation Measurement (PAM) scale [27] is one of the few assessment scales that have been used to evaluate patient engagement in telemedicine [40,108-110]. The PAM scale was

XSL•FO
RenderX

developed to quantify patients' knowledge, skills, and confidence in managing their health [27,111]. However, although the concepts of patient engagement and patient activation overlap, they differ in their conceptual breadth [25]. As discussed earlier, patient engagement represents a multidimensional psychosocial process in which individuals' cognitive, emotional, and behavioral actions collectively shape how they manage their health. In contrast, patient activation primarily emphasizes the cognitive and behavioral components of this process [25,31]. As such, the PAM scale could not capture the holistic nature of patient engagement. Another widely accepted patient engagement scale is the 5-item Patient Health Engagement (PHE) scale developed by Graffigna and colleagues [25]. The PHE scale assesses patients' perceived readiness for cognitive, emotional, and behavioral engagement. However, none of the studies included in this review used this instrument. In addition, patient engagement has been measured in previous research using other standardized tools, such as the observing patient involvement in decision making (OPTION) scale for measuring patient involvement [112], the Perceived Involvement in Care Scale [113], and the Patient Participation Scale [114], none of which were applied in the included studies. Nevertheless, these existing instruments hold potential for integration or adaptation to enable more consistent evaluation of engagement outcomes in future telemedicine research. We summarized available standardized tools for assessing patient engagement and their potential adaptations to telemedicine in Multimedia Appendix 4.

Limitations and Future Directions

This review has some limitations to note: First, it only included telemedicine studies with HCP-patient interactions. Although telemedicine tools with interactive support from providers have great potential to engage patients [38,39], other studies on telemedicine platforms that focus on patient education, health data management, or the dissemination of health-related information may also incorporate additional effective communication strategies that enhance patient engagement, which can be explored in future reviews. Second, the review did not include gray literature, which may have led to the omission of recent developments or emerging trends in the field first reported at conferences. Incorporating conference proceedings in future review could provide a more comprehensive and up-to-date understanding of the field. Third, this review only included peer-reviewed articles published in English, which may have excluded important research published in other languages that explored telemedicine in various contexts. Despite these limitations, this review serves as a foundational step in the field. It is hoped that future research will address these deficits by exploring the topic more comprehensively.

Future research can explore the following directions in studying effective communication strategies for promoting patient engagement with telemedicine tools. First, researchers should further clarify what patient engagement is by providing a rigorous conceptualization and exploring the dimensions of patient engagement, particularly in the telemedicine environment. Currently, studies have tested and collected data on usability, patient acceptability, patient participation, health condition management, and so on, to understand patient engagement. However, what the components of patient engagement are and how to measure them scientifically remain unclear. In addition to using explorative qualitative methods to ask questions about patients' attitudes and preferences toward telemedicine tools, validated assessment instruments for patient engagement in this field are expected to be developed. Second, future studies should examine and validate the relationships between 12 subthemes across the 3 overarching communication strategy themes identified in this review and patient engagement. Such efforts could contribute to the development of an integrated communication framework that fosters patient engagement with telemedicine tools. In particular, future studies may explore and empirically test the connections between specific communication subthemes and different dimensions of patient engagement. Third, future work can build on this study by exploring additional telemedicine contexts beyond HCP-patient interactions, integrating grey literature and conference proceedings, and including non-English publications to capture more comprehensive evidence, emerging trends, and broader cultural perspectives on communication strategies influencing patient engagement.

Conclusion

This systematic review underscores the critical role of various communication strategies in enhancing patient engagement in telemedicine with HCP-patient interactions. A total of 3 themes of communication strategies, namely interpersonal (micro), team (meso), and system (macro) level communication strategies, with 12 subthemes, were identified as important factors influencing patient engagement. This review offers an innovative and pioneering effort to systematically synthesize communication strategies that promote patient engagement in telemedicine. Unlike previous reviews that focused on isolated aspects or levels of communication, our review uniquely integrates strategies across all three levels to provide a holistic and comprehensive framework. Theoretically, it advances understanding of how micro-, meso-, and macrolevel communication strategies collectively influence patient engagement, filling a critical gap in existing literature. Practically, it provides actionable guidance for telemedicine developers, health care professionals, and policymakers. The identified strategies offer a comprehensive framework for improving the quality and sustainability of telemedicine practices. In real-world terms, these insights can inform training programs for health care professionals, guide platform design, and support policy initiatives that promote equitable, patient-centered digital care. We also found that the majority of included studies used qualitative research methods to assess patient engagement. Future studies can further explore, validate, and test quantitative methods to evaluate patient engagement and the relationships between different communication strategies and patient engagement in telemedicine.

XSL•FO
RenderX

Acknowledgments

This research was funded by Hong Kong Polytechnic University for the article processing fee (APF) payment support. The authors appreciate librarians (Queennie Ip and Emily Wu) for their professional assistance in developing the search strategies for this systematic review.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary files.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA 2020_S checklist.
[DOCX File , 58 KB - jmir_v28i1e85456_app1.docx]

Multimedia Appendix 2
Databases and search strategies for studies on communication strategies influencing patient engagement in telemedicine involving healthcare provider-patient interactions.
[DOCX File , 17 KB - jmir_v28i1e85456_app2.docx]

Multimedia Appendix 3
Data extraction table for included studies on communication strategies influencing patient engagement in telemedicine involving healthcare provider-patient interactions.
[XLSX File (Microsoft Excel File), 31 KB - jmir_v28i1e85456_app3.xlsx]

Multimedia Appendix 4
Summary of standardized instruments for measuring patient engagement and their relevance and potential to be adapted in telemedicine research.
[DOCX File , 21 KB - jmir_v28i1e85456_app4.docx]

References

1. Greenhalgh T, Wherton J, Shaw S, Morrison C. Video consultations for COVID-19. BMJ 2020;368:m998. [doi: 10.1136/bmj.m998] [Medline: 32165352]
2. Yeung AWK, Torkamani A, Butte AJ, Glicksberg BS, Schuller B, Rodriguez B, et al. The promise of digital healthcare technologies. Front Public Health 2023;11:1196596 [FREE Full text] [doi: 10.3389/fpubh.2023.1196596] [Medline: 37822534]
3. Lee NK, Kim JS. Status and trends of the digital healthcare industry. Healthc Inform Res 2024;30(3):172-183 [FREE Full text] [doi: 10.4258/hir.2024.30.3.172] [Medline: 39160777]
4. Cui F, He X, Zhai Y, Lyu M, Shi J, Sun D, et al. Application of telemedicine services based on a regional telemedicine platform in China From 2014 to 2020: longitudinal trend analysis. J Med Internet Res 2021;23(7):e28009 [FREE Full text] [doi: 10.2196/28009] [Medline: 34255686]
5. Kruse CS, Williams K, Bohls J, Shamsi W. Telemedicine and health policy: A systematic review. Health Policy and Technology 2021;10(1):209-229. [doi: 10.1016/j.hlpt.2020.10.006]
6. Telemedicine: Opportunities and developments in member state. WHO. 2010. URL: https://www.afro.who.int/publications/telemedicine-opportunities-and-developments-member-state [accessed 2025-07-10]
7. Caiata-Zufferey M, Abraham A, Sommerhalder K, Schulz PJ. Online health information seeking in the context of the medical consultation in Switzerland. Qual Health Res 2010;20(8):1050-1061. [doi: 10.1177/1049732310368404] [Medline: 20442347]
8. Moffatt JJ, Eley DS. Barriers to the up-take of telemedicine in Australia - a view from providers. RRH 2011;11(1):116-121. [doi: 10.22605/rrh1581]
9. Noceda AVG, Acierto LMM, Bertiz MCC, Dionisio DEH, Laurito CBL, Sanchez GAT, et al. Patient satisfaction with telemedicine in the Philippines during the COVID-19 pandemic: a mixed methods study. BMC Health Serv Res 2023;23(1):277 [FREE Full text] [doi: 10.1186/s12913-023-09127-x] [Medline: 36949479]
10. Wu X, Yang Y, Li Y, Li Y, Li H, Lyu Y, et al. Using theory-based frameworks to identify barriers and enablers of physicians' telemedicine adoption and develop intervention strategies in China: multicenter qualitative study. J Med Internet Res 2025;27:e73412 [FREE Full text] [doi: 10.2196/73412] [Medline: 40920450]

XSL•FO
RenderX

11. Craig A, Lawford H, Miller M, Chen-Cao L, Woods L, Liaw S, et al. Use of technology to support health care providers delivering care in low- and lower-middle-income countries: systematic umbrella review. J Med Internet Res 2025;27:e66288 [FREE Full text] [doi: 10.2196/66288] [Medline: 40533075]
12. Wootton R, Bonnardot L. Telemedicine in low-resource settings. Front Public Health 2015;3:3 [FREE Full text] [doi: 10.3389/fpubh.2015.00003] [Medline: 25654074]
13. Xiong D, Zhao L. Research on credit evaluation of mobile medical APP interactive online consultation service-take Haodaifu APP online payment service as an example. J Phys Conf Ser 2017;910(1):012055 [FREE Full text] [doi: 10.1088/1742-6596/910/1/012055]
14. Bonnechère B, Kossi O, Mapinduzi J, Panda J, Rintala A, Guidetti S, et al. Mobile health solutions: an opportunity for rehabilitation in low- and middle income countries? Front Public Health 2022;10:1072322. [doi: 10.3389/fpubh.2022.1072322] [Medline: 36761328]
15. Dullet NW, Geraghty EM, Kaufman T, Kissee JL, King J, Dharmar M, et al. Impact of a university-based outpatient telemedicine program on time savings, travel costs, and environmental pollutants. Value Health 2017;20(4):542-546 [FREE Full text] [doi: 10.1016/j.jval.2017.01.014] [Medline: 28407995]
16. Suzuki T, Hotta J, Kuwabara T, Yamashina H, Ishikawa T, Tani Y, et al. Possibility of introducing telemedicine services in Asian and African countries. Health Policy and Technology 2020;9(1):13-22. [doi: 10.1016/j.hlpt.2020.01.006]
17. Du Y, Zhou Q, Cheng W, Zhang Z, Hoelzer S, Liang Y, et al. Factors influencing adoption and use of telemedicine services in rural areas of China: mixed methods study. JMIR Public Health Surveill 2022;8(12):e40771 [FREE Full text] [doi: 10.2196/40771] [Medline: 36563026]
18. Giebel GD, Abels C, Plescher F, Speckemeier C, Schrader NF, Börchers K, et al. Problems and barriers related to the use of mhealth apps from the perspective of patients: focus group and interview study. J Med Internet Res 2024;26:e49982 [FREE Full text] [doi: 10.2196/49982] [Medline: 38652508]
19. An Q, Kelley MM, Hanners A, Yen P. Sustainable development for mobile health apps using the human-centered design process. JMIR Form Res 2023;7:e45694 [FREE Full text] [doi: 10.2196/45694] [Medline: 37624639]
20. Ftouni R, AlJardali B, Hamdanieh M, Ftouni L, Salem N. Challenges of telemedicine during the COVID-19 pandemic: a systematic review. BMC Med Inform Decis Mak 2022;22(1):207 [FREE Full text] [doi: 10.1186/s12911-022-01952-0] [Medline: 35922817]
21. Fernández Coves A, Yeung KHT, van der Putten IM, Nelson EAS. Teleconsultation adoption since COVID-19: comparison of barriers and facilitators in primary care settings in Hong Kong and the Netherlands. Health Policy 2022;126(10):933-944 [FREE Full text] [doi: 10.1016/j.healthpol.2022.07.012] [Medline: 36050194]
22. Rosler G. Pediatric telehealth experiences: myths and truths about video visits from a parent. J Patient Exp 2020;7(6):836-838 [FREE Full text] [doi: 10.1177/2374373520932724] [Medline: 33457506]
23. Bertoncello C, Colucci M, Baldovin T, Buja A, Baldo V. How does it work? Factors involved in telemedicine home-interventions effectiveness: A review of reviews. PLoS One 2018;13(11):e0207332 [FREE Full text] [doi: 10.1371/journal.pone.0207332] [Medline: 30440004]
24. Liu Z, Brandon-Jones A, Vasilakis C. Unpacking patient engagement in remote consultation. IJOPM 2024;44(13):157-194. [doi: 10.1108/ijopm-03-2023-0188]
25. Graffigna G, Barello S, Bonanomi A, Lozza E. Measuring patient engagement: development and psychometric properties of the patient health engagement (PHE) scale. Front Psychol 2015;6:274 [FREE Full text] [doi: 10.3389/fpsyg.2015.00274] [Medline: 25870566]
26. Harrington RL, Hanna ML, Oehrlein EM, Camp R, Wheeler R, Cooblall C, et al. Defining patient engagement in research: results of a systematic review and analysis: report of the ISPOR patient-centered special interest group. Value Health 2020;23(6):677-688. [doi: 10.1016/j.jval.2020.01.019] [Medline: 32540224]
27. Hibbard JH, Mahoney ER, Stockard J, Tusler M. Development and testing of a short form of the patient activation measure. Health Serv Res 2005;40(6 Pt 1):1918-1930 [FREE Full text] [doi: 10.1111/j.1475-6773.2005.00438.x] [Medline: 16336556]
28. Menichetti J, Libreri C, Lozza E, Graffigna G. Giving patients a starring role in their own care: a bibliometric analysis of the on-going literature debate. Health Expect 2016;19(3):516-526 [FREE Full text] [doi: 10.1111/hex.12299] [Medline: 25369557]
29. Kruse CS, Krowski N, Rodriguez B, Tran L, Vela J, Brooks M. Telehealth and patient satisfaction: a systematic review and narrative analysis. BMJ Open 2017;7(8):e016242 [FREE Full text] [doi: 10.1136/bmjopen-2017-016242] [Medline: 28775188]
30. Agha Z, Schapira RM, Laud PW, McNutt G, Roter DL. Patient satisfaction with physician-patient communication during telemedicine. Telemed J E Health 2009;15(9):830-839. [doi: 10.1089/tmj.2009.0030] [Medline: 19919189]
31. Barello S, Triberti S, Graffigna G, Libreri C, Serino S, Hibbard J, et al. eHealth for patient engagement: a systematic review. Front Psychol 2015;6:2013 [FREE Full text] [doi: 10.3389/fpsyg.2015.02013] [Medline: 26779108]
32. Inglis SC, Clark RA, Dierckx R, Prieto-Merino D, Cleland JGF. Structured telephone support or non-invasive telemonitoring for patients with heart failure. Cochrane Database Syst Rev 2015;2015(10):CD007228 [FREE Full text] [doi: 10.1002/14651858.CD007228.pub3] [Medline: 26517969]

XSL•FO
RenderX

33. Vicente MA, Fernández C, Guilabert M, Carrillo I, Martín-Delgado J, Mira JJ. Patient engagement using telemedicine in primary care during COVID-19 pandemic: a trial study. Int J Environ Res Public Health 2022;19(22):14682 [FREE Full text] [doi: 10.3390/ijerph192214682] [Medline: 36429402]
34. Khanijahani A, Akinci N, Quitiquit E. A systematic review of the role of telemedicine in blood pressure control: focus on patient engagement. Curr Hypertens Rep 2022;24(7):247-258. [doi: 10.1007/s11906-022-01186-5] [Medline: 35412188]
35. Meyer MA. COVID-19 pandemic accelerates need to improve online patient engagement practices to enhance patient experience. J Patient Exp 2020;7(5):657-664 [FREE Full text] [doi: 10.1177/2374373520959486] [Medline: 33294595]
36. Costa D, Serra R. The role of communication in managing chronic lower limb wounds. J Multidiscip Healthc 2025;18:3685-3708 [FREE Full text] [doi: 10.2147/JMDH.S533416] [Medline: 40589782]
37. Talal AH, Sofikitou EM, Jaanimägi U, Zeremski M, Tobin JN, Markatou M. A framework for patient-centered telemedicine: application and lessons learned from vulnerable populations. J Biomed Inform 2020;112:103622 [FREE Full text] [doi: 10.1016/j.jbi.2020.103622] [Medline: 33186707]
38. Cingi C, Yorgancioglu A, Cingi CC, Oguzulgen K, Muluk NB, Ulusoy S, et al. The "physician on call patient engagement trial" (POPET): measuring the impact of a mobile patient engagement application on health outcomes and quality of life in allergic rhinitis and asthma patients. Int Forum Allergy Rhinol 2015;5(6):487-497. [doi: 10.1002/alr.21468] [Medline: 25856270]
39. van Middelaar T, Beishuizen CRL, Guillemont J, Barbera M, Richard E, Moll van Charante EP, HATICE consortium. Engaging older people in an internet platform for cardiovascular risk self-management: a qualitative study among Dutch HATICE participants. BMJ Open 2018;8(1):e019683 [FREE Full text] [doi: 10.1136/bmjopen-2017-019683] [Medline: 29358447]
40. Clarke AL, Roscoe J, Appleton R, Parashar D, Muthuswamy R, Khan O, et al. Promoting integrated care in prostate cancer through online prostate cancer-specific holistic needs assessment: a feasibility study in primary care. Support Care Cancer 2020;28(4):1817-1827 [FREE Full text] [doi: 10.1007/s00520-019-04967-y] [Medline: 31338642]
41. Crotty BH, Hyun N, Polovneff A, Dong Y, Decker MC, Mortensen N, et al. Analysis of clinician and patient factors and completion of telemedicine appointments using video. JAMA Netw Open 2021;4(11):e2132917 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.32917] [Medline: 34735013]
42. Habbash F, Rabeeah A, Huwaidi Z, Abuobaidah H, Alqabbat J, Hayyan F, et al. Telemedicine in non-communicable chronic diseases care during the COVID-19 pandemic: exploring patients' perspectives. Front Public Health 2023;11:1270069 [FREE Full text] [doi: 10.3389/fpubh.2023.1270069] [Medline: 37818295]
43. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
44. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
45. Niyomyart A, Ruksakulpiwat S, Benjasirisan C, Phianhasin L, Nigussie K, Thorngthip S, et al. Current status of barriers to mhealth access among patients with stroke and steps toward the digital health era: systematic review. JMIR Mhealth Uhealth 2024;12:e54511 [FREE Full text] [doi: 10.2196/54511] [Medline: 39173152]
46. Chen S, Niu M, Ngai CSB. What is the next step of ICT development? The changes of ICT use in promoting elderly healthcare access: a systematic literature review. Heliyon 2024;10(3):e25197 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e25197] [Medline: 38371988]
47. Hu Y, Ngai CSB, Chen S. Automated approaches to screening developmental language disorder: a comprehensive review and future prospects. J Speech Lang Hear Res 2025;68(5):2478-2498. [doi: 10.1044/2025_JSLHR-24-00488] [Medline: 40228046]
48. Lu W, Ngai CSB, Yi L. A bibliometric review of constituents, themes, and trends in online medical consultation research. Health Commun 2024;39(2):229-243. [doi: 10.1080/10410236.2022.2163108] [Medline: 36581497]
49. Bjerkan J, Kane B, Uhrenfeldt L, Veie M, Fossum M. Citizen-patient involvement in the development of mhealth technology: protocol for a systematic scoping review. JMIR Res Protoc 2020;9(8):e16781 [FREE Full text] [doi: 10.2196/16781] [Medline: 32857061]
50. Li T, Higgins J, Deeks J. Chapter 5: collecting data. In: Cochrane Handbook for Systematic Reviews of Interventions Version 65. London: Wiley; 2024.
51. Ackerman SL, Dowdell K, Clebak KT, Quinn M, Shipman SA. Patients assess an econsult model's acceptability at 5 us academic medical centers. Ann Fam Med 2020;18(1):35-41 [FREE Full text] [doi: 10.1370/afm.2487] [Medline: 31937531]
52. Alpert JM, Hampton CN, Raisa A, Markham MJ, Bylund CL. Integrating patient-centeredness into online patient-clinician communication: a qualitative analysis of clinicians' secure messaging usage. Support Care Cancer 2022;30(12):9851-9857 [FREE Full text] [doi: 10.1007/s00520-022-07408-5] [Medline: 36260178]
53. Bavngaard MV, Lüchau EC, Hvidt EA, Grønning A. Exploring patient participation during video consultations: A qualitative study. Digit Health 2023;9:20552076231180682 [FREE Full text] [doi: 10.1177/20552076231180682] [Medline: 37325071]

XSL•FO
RenderX

54. Björndell C, Premberg Å. Physicians' experiences of video consultation with patients at a public virtual primary care clinic: a qualitative interview study. Scand J Prim Health Care 2021;39(1):67-76 [FREE Full text] [doi: 10.1080/02813432.2021.1882082] [Medline: 33650941]
55. Breton M, Sullivan EE, Deville-Stoetzel N, McKinstry D, DePuccio M, Sriharan A, et al. Telehealth challenges during COVID-19 as reported by primary healthcare physicians in Quebec and Massachusetts. BMC Fam Pract 2021;22(1):192 [FREE Full text] [doi: 10.1186/s12875-021-01543-4] [Medline: 34563113]
56. Brodar KE, Hong N, Liddle M, Hernandez L, Waks J, Sanchez J, et al. Transitioning to telehealth services in a pediatric diabetes clinic during COVID-19: an interdisciplinary quality improvement initiative. J Clin Psychol Med Settings 2022;29(4):727-738 [FREE Full text] [doi: 10.1007/s10880-021-09830-z] [Medline: 34708318]
57. Caffery LJ, Taylor M, North JB, Smith AC. Tele-orthopaedics: a snapshot of services in Australia. J Telemed Telecare 2017;23(10):835-841. [doi: 10.1177/1357633X17732800] [Medline: 28950754]
58. Davoust M, Bazzi AR, Blakemore S, Blodgett J, Cheng A, Fielman S, et al. Patient and clinician experiences with the implementation of telemedicine and related adaptations in office-based buprenorphine treatment during the COVID-19 pandemic: a qualitative study. Addict Sci Clin Pract 2025;20(1):21 [FREE Full text] [doi: 10.1186/s13722-025-00536-3] [Medline: 40033433]
59. Day S, Rae C, McOwan A, Wilkins R, Gray A, Harvey A, et al. Patient evaluation of Klick, a technology-enabled, nurse-delivered HIV outpatient pathway. HIV Med 2025;26(1):128-139. [doi: 10.1111/hiv.13710] [Medline: 39279040]
60. Dong F, Jumper MBE, Becker-Haimes EM, Vatza C, Miao L, Conroy C, et al. Tele-mental health transitions for Pennsylvania coordinated specialty care programs for early psychosis during the COVID-19 pandemic. Psychiatr Q 2023;94(2):89-102 [FREE Full text] [doi: 10.1007/s11126-023-10015-0] [Medline: 36820952]
61. Esayed S, Kim E, Sung HC, Al-Seraji A, Adeyemo S, Troutt H, et al. Hybrid telemedicine and in-person care for kidney transplant follow-up: a qualitative study. Clin Transplant 2025;39(2):e70106 [FREE Full text] [doi: 10.1111/ctr.70106] [Medline: 39945199]
62. Gibson J, Lightbody E, McLoughlin A, McAdam J, Gibson A, Day E, et al. 'It was like he was in the room with us': patients' and carers' perspectives of telemedicine in acute stroke. Health Expect 2016;19(1):98-111 [FREE Full text] [doi: 10.1111/hex.12333] [Medline: 25581591]
63. Grens H, de Bruin JP, Huppelschoten A, Kremer JAM. Fertility workup with video consultation during the COVID-19 pandemic: pilot quantitative and qualitative study. JMIR Form Res 2022;6(2):e32000. [doi: 10.2196/32000] [Medline: 34936981]
64. Grove BE, Valen Schougaard LM, Ivarsen P, Hjollund NH, de Thurah A, Mejdahl CT. Remote follow-up based on patient-reported outcomes in patients with chronic kidney disease: A qualitative study of patient perspectives. PLoS One 2023;18(2):e0281393 [FREE Full text] [doi: 10.1371/journal.pone.0281393] [Medline: 36763600]
65. Higa C, Davidson EJ, Loos JR. Integrating family and friend support, information technology, and diabetes education in community-centric diabetes self-management. J Am Med Inform Assoc 2021;28(2):261-275 [FREE Full text] [doi: 10.1093/jamia/ocaa223] [Medline: 33164074]
66. Islind AS, Snis UL, Lindroth T, Lundin J, Cerna K, Steineck G. The virtual clinic: two-sided affordances in consultation practice. Comput Supported Coop Work 2019;28(3-4):435-468. [doi: 10.1007/s10606-019-09350-3]
67. James S, Ashley C, Williams A, Desborough J, Mcinnes S, Calma K, et al. Experiences of Australian primary healthcare nurses in using telehealth during COVID-19: a qualitative study. BMJ Open 2021;11(8):e049095 [FREE Full text] [doi: 10.1136/bmjopen-2021-049095] [Medline: 34362804]
68. Jensen AL, Schougaard LMV, Laurberg T, Hansen TK, Lomborg K. Flexible patient-reported outcome-based telehealth follow-up for type 1 diabetes: A qualitative study. Scand J Caring Sci 2023;37(3):662-676. [doi: 10.1111/scs.13154] [Medline: 36775917]
69. Jethwa H, Brooke M, Parkinson A, Dures E, Gullick NJ. Patients' perspectives of telemedicine appointments for psoriatic arthritis during the COVID-19 pandemic: results of a patient-driven pilot survey. BMC Rheumatol 2022;6(1):13 [FREE Full text] [doi: 10.1186/s41927-021-00242-y] [Medline: 35189975]
70. Jung OS, Graetz I, Dorner SC, Hayden EM. Implementing a COVID-19 virtual observation unit in emergency medicine: frontline clinician and staff experiences. Med Care Res Rev 2023;80(1):79-91 [FREE Full text] [doi: 10.1177/10775587221108750] [Medline: 35815570]
71. Moore A, Chavez C, Fisher MP. Factors enhancing trust in electronic communication among patients from an internal medicine clinic: qualitative results of the RECEPT study. J Gen Intern Med 2022;37(12):3121-3127 [FREE Full text] [doi: 10.1007/s11606-021-07345-9] [Medline: 35048293]
72. Morrison C, Beattie M, Wherton J, Stark C, Anderson J, Hunter-Rowe C, et al. Testing and implementing video consulting for outpatient appointments: using quality improvement system thinking and codesign principles. BMJ Open Qual 2021;10(1):e001259 [FREE Full text] [doi: 10.1136/bmjoq-2020-001259] [Medline: 33674346]
73. Olayiwola JN, Knox M, Dubé K, Lu EC, Woldeyesus T, James IE, et al. Understanding the potential for patient engagement in electronic consultation and referral systems: lessons from one safety net system. Health Serv Res 2018;53(4):2483-2502 [FREE Full text] [doi: 10.1111/1475-6773.12776] [Medline: 28940495]

XSL•FO
RenderX

74. Osmundsen TC, Andreassen Jaatun EA, Heggem GF, Kulseng BE. Service innovation from the edges: enhanced by telemedicine decision support. Pers Ubiquit Comput 2015;19(3-4):699-708. [doi: 10.1007/s00779-015-0857-9]
75. Rodkjær L, Jeppesen M, Schougaard L. Management of cystic fibrosis during COVID-19: patient reported outcomes based remote follow-up among CF patients in Denmark - A feasibility study. J Cyst Fibros 2022;21(2):e106-e112 [FREE Full text] [doi: 10.1016/j.jcf.2021.10.010] [Medline: 34785157]
76. Scruton S, Wong G, Babinski S, Squires LR, Berlin A, Easley J, et al. Optimizing virtual follow-up care: realist evaluation of experiences and perspectives of patients with breast and prostate cancer. J Med Internet Res 2025;27:e65148 [FREE Full text] [doi: 10.2196/65148] [Medline: 39752659]
77. Selick A, Durbin J, Hamdani Y, Rayner J, Lunsky Y. "Can you hear me now?": a qualitative exploration of communication quality in virtual primary care encounters for patients with intellectual and developmental disabilities. BMC Prim Care 2023;24(1):105 [FREE Full text] [doi: 10.1186/s12875-023-02055-z] [Medline: 37081380]
78. Spiess ST, Gardner E, Turner C, Galt A, Fortenberry K, Ho T, et al. We cannot put this genie back in the bottle: qualitative interview study among family medicine providers about their experiences with virtual visits during the COVID-19 pandemic. J Med Internet Res 2023;25:e43877 [FREE Full text] [doi: 10.2196/43877] [Medline: 37651162]
79. Trondsen MV, Tjora A, Broom A, Scambler G. The symbolic affordances of a video-mediated gaze in emergency psychiatry. Soc Sci Med 2018;197:87-94. [doi: 10.1016/j.socscimed.2017.11.056] [Medline: 29222999]
80. Wei S, Mao Y. Small talk is a big deal: a discursive analysis of online off-topic doctor-patient interaction in traditional Chinese medicine. Soc Sci Med 2023;317:115632. [doi: 10.1016/j.socscimed.2022.115632] [Medline: 36584441]
81. White SJ, Nguyen AD, Roger P, Tse T, Cartmill JA, Hatem S, et al. Tailoring communication practices to support effective delivery of telehealth in general practice. BMC Prim Care 2024;25(1):232. [doi: 10.1186/s12875-024-02441-1] [Medline: 38937674]
82. Wood SM, Pickel J, Phillips AW, Baber K, Chuo J, Maleki P, et al. Acceptability, feasibility, and quality of telehealth for adolescent health care delivery during the COVID-19 pandemic: cross-sectional study of patient and family experiences. JMIR Pediatr Parent 2021;4(4):e32708 [FREE Full text] [doi: 10.2196/32708] [Medline: 34779782]
83. Zainal H, Hui XX, Thumboo J, Fong W, Yong FK. Patients' expectations of doctors' clinical competencies in the digital health care era: qualitative semistructured interview study among patients. JMIR Hum Factors 2024;11:e51972 [FREE Full text] [doi: 10.2196/51972] [Medline: 39190915]
84. Canale M. From communicative competence to communicative language pedagogy. In: Language and Communication. London: Longman Press Publishing; 1983:2-27.
85. Porter ME. What is strategy? Harvard business review 1996;74(6):61-78 [FREE Full text]
86. Porter ME, Lee TH. Why strategy matters now. N Engl J Med 2015;372(18):1681-1684. [doi: 10.1056/NEJMp1502419] [Medline: 25923546]
87. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The mixed methods appraisal tool (MMAT) version 2018 for information professionals and researchers. EFI 2018;34(4):285-291. [doi: 10.3233/EFI-180221]
88. Draganidis A, Fernando AN, West ML, Sharp G. Social media delivered mental health campaigns and public service announcements: A systematic literature review of public engagement and help-seeking behaviours. Soc Sci Med 2024;359:117231 [FREE Full text] [doi: 10.1016/j.socscimed.2024.117231] [Medline: 39278158]
89. Ngai CSB, Singh RG, Lu W, Koon AC. Grappling with the COVID-19 health crisis: content analysis of communication strategies and their effects on public engagement on social media. J Med Internet Res 2020;22(8):e21360 [FREE Full text] [doi: 10.2196/21360] [Medline: 32750013]
90. Ngai CSB, Singh RG, Yao L. Impact of COVID-19 vaccine misinformation on social media virality: content analysis of message themes and writing strategies. J Med Internet Res 2022;24(7):e37806 [FREE Full text] [doi: 10.2196/37806] [Medline: 35731969]
91. Hsieh H, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
92. Elliott V. Thinking about the coding process in qualitative data analysis. TQR 2018. [doi: 10.46743/2160-3715/2018.3560]
93. Ritchie J, Spencer L. Qualitative data analysis for applied policy research. In: Analyzing Qualitative Data Taylor & Francis e-Library. Thousand Oaks, CA: Sage Publications, Inc; 2002:173-194.
94. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310]
95. Borenstein M, Hedges LV, Higgins J, Rothstein H. When does it make sense to perform a meta'analysis? In: Introduction to meta?analysis. New Jerssey: John Wiley & Sons, Ltd; 2009:357-364.
96. Kwame A, Petrucka PM. A literature-based study of patient-centered care and communication in nurse-patient interactions: barriers, facilitators, and the way forward. BMC Nurs 2021;20(1):158 [FREE Full text] [doi: 10.1186/s12912-021-00684-2] [Medline: 34479560]
97. Gittell JH, Ali HN. The theory of relational coordination. In: Relational Analytics. United Kingdom: Routledge; 2021:16-38.
98. Campinha-Bacote J. The process of cultural competence in the delivery of healthcare services: a model of care. J Transcult Nurs 2002;13(3):181-4; discussion 200. [doi: 10.1177/10459602013003003] [Medline: 12113146]

XSL•FO
RenderX

99. Ngai CSB, Singh RG, Lu W. Exploring drivers for public engagement in social media communication with medical social influencers in China. PLoS One 2020;15(10):e0240303 [FREE Full text] [doi: 10.1371/journal.pone.0240303] [Medline: 33027269]
100. Sungur H, Yılmaz NG, Chan BMC, van den Muijsenbergh METC, van Weert JCM, Schouten BC. Development and evaluation of a digital intervention for fulfilling the needs of older migrant patients with cancer: user-centered design approach. J Med Internet Res 2020;22(10):e21238 [FREE Full text] [doi: 10.2196/21238] [Medline: 33104008]
101. Crafoord M, Fjell M, Sundberg K, Nilsson M, Langius-Eklöf A. Engagement in an interactive app for symptom self-management during treatment in patients with breast or prostate cancer: mixed methods study. J Med Internet Res 2020;22(8):e17058 [FREE Full text] [doi: 10.2196/17058] [Medline: 32663140]
102. Xie Z, Nacioglu A, Or C. Prevalence, demographic correlates, and perceived impacts of mobile health app use amongst Chinese adults: cross-sectional survey study. JMIR Mhealth Uhealth 2018;6(4):e103 [FREE Full text] [doi: 10.2196/mhealth.9002] [Medline: 29699971]
103. Vasiloglou MF, Christodoulidis S, Reber E, Stathopoulou T, Lu Y, Stanga Z, et al. Nutrients 2020;12(8):2214 [FREE Full text] [doi: 10.3390/nu12082214] [Medline: 32722339]
104. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Management Science 1989;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
105. Alzahrani A, Gay V, Alturki R. Exploring Saudi individuals' perspectives and needs to design a hypertension management mobile technology solution: qualitative study. Int J Environ Res Public Health 2022;19(19):12956 [FREE Full text] [doi: 10.3390/ijerph191912956] [Medline: 36232254]
106. Chu D, Lessard D, Laymouna MA, Engler K, Schuster T, Ma Y, et al. Understanding the risks and benefits of a patient portal configured for hiv care: patient and healthcare professional perspectives. J Pers Med 2022;12(2):314. [doi: 10.3390/jpm12020314] [Medline: 35207803]
107. Toni E, Pirnejad H, Makhdoomi K, Mivefroshan A, Niazkhani Z. Patient empowerment through a user-centered design of an electronic personal health record: a qualitative study of user requirements in chronic kidney disease. BMC Med Inform Decis Mak 2021;21(1):329 [FREE Full text] [doi: 10.1186/s12911-021-01689-2] [Medline: 34819050]
108. Chaudhry T, Ormandy P, Vasilica C. Using FLO text-messages to enhance health behaviours and self-management of long-term conditions in South-Asian patients. Digit Health 2024;10:20552076241242558 [FREE Full text] [doi: 10.1177/20552076241242558] [Medline: 38708186]
109. Jiang Y, Hwang M, Cho Y, Friese CR, Hawley ST, Manojlovich M, et al. The acceptance and use of digital technologies for self-reporting medication safety events after care transitions to home in patients with cancer: survey study. J Med Internet Res 2024;26:e47685 [FREE Full text] [doi: 10.2196/47685] [Medline: 38457204]
110. Shewchuk B, Green LA, Barber T, Miller J, Teare S, Campbell-Scherer D, et al. Patients' use of mobile health for self-management of knee osteoarthritis: results of a 6-week pilot study. JMIR Form Res 2021;5(11):e30495 [FREE Full text] [doi: 10.2196/30495] [Medline: 34842526]
111. Hibbard JH, Stockard J, Mahoney ER, Tusler M. Development of the patient activation measure (PAM): conceptualizing and measuring activation in patients and consumers. Health Serv Res 2004;39(4 Pt 1):1005-1026 [FREE Full text] [doi: 10.1111/j.1475-6773.2004.00269.x] [Medline: 15230939]
112. Elwyn G, Edwards A, Wensing M, Hood K, Atwell C, Grol R. Shared decision making: developing the OPTION scale for measuring patient involvement. Qual Saf Health Care 2003;12(2):93-99 [FREE Full text] [doi: 10.1136/qhc.12.2.93] [Medline: 12679504]
113. Lerman CE, Brody DS, Caputo GC, Smith DG, Lazaro CG, Wolfson HG. Patients' perceived involvement in care scale: relationship to attitudes about illness and medical care. J Gen Intern Med 1990;5(1):29-33. [doi: 10.1007/BF02602306] [Medline: 2299426]
114. Song M, Kim M. Development and validation of a patient participation scale. J Adv Nurs 2023;79(6):2393-2403. [doi: 10.1111/jan.15593] [Medline: 36814372]

Abbreviations

HCP: health care provider
mHealth: mobile health
MMAT: Mixed Methods Appraisal Tool
OPTION: observing patient involvement in decision making
PAM: Patient Activation Measurement
PHE: 5-item Patient Health Engagement
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension
PROSPERO: International Prospective Register of Systematic Reviews
RQ: research question

XSL•FO
RenderX

Edited by S Brini; submitted 08.Oct.2025; peer-reviewed by Y Li, E Afarikumah; comments to author 24.Oct.2025; revised version received 12.Dec.2025; accepted 22.Dec.2025; published 21.Jan.2026.

Please cite as:
Hu Y, Ngai CSB, Jiang R
Communication Strategies to Promote Patient Engagement in Telemedicine: Systematic Review
J Med Internet Res 2026;28:e85456
URL: https://www.jmir.org/2026/1/e85456
doi:10.2196/85456
PMID:

©Yangna Hu, Cindy Sing Bik Ngai, Rui Jiang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Products, Performance, and Technological Development of Ambulatory Oxygen Therapy Devices: Scoping Review

Shohei Kawachi[1,2*], BPT, PhD; Mariana Hoffman[1,3*], PhD; Lorena Romero[4], BA, MBIT; Magnus Ekström[5], MD, PhD; Jerry A Krishnan[6], MD, PhD; Anne E Holland[1,3,7*], BAppSc, PhD

[1]Respiratory Research@Alfred, School of Translational Medicine, Monash University, Melbourne, Victoria, Australia

[2]Department of Rehabilitation, Shinshu University Hospital, Matsumoto, Nagano, Japan

[3]Institute for Breathing and Sleep, Melbourne, Victoria, Australia

[4]Ian Potter Library, The Alfred Hospital, Melbourne, Victoria, Australia

[5]Respiratory Medicine, Allergology and Palliative Medicine, Department of Clinical Sciences Lund, Lund University, Lund, Sweden

[6]Breathe Chicago Center, University of Illinois Chicago, Chicago, IL, United States

[7]Departments of Physiotherapy and Respiratory Medicine, Alfred Health, Melbourne, Victoria, Australia

[*]these authors contributed equally

Corresponding Author:
Shohei Kawachi, BPT, PhD
Respiratory Research@Alfred
School of Translational Medicine
Monash University
Level 6 The Alfred Centre, 99 Commercial Rd
Melbourne, Victoria, 3004
Australia
Phone: 81 762652948
Email: shohei.kawachi@gmail.com

Abstract

Background: Ambulatory oxygen therapy is prescribed for patients with chronic lung diseases who experience exertional hypoxemia. However, available devices may not adequately meet user requirements, and their performance characteristics are heterogeneous.

Objective: This study aims to identify devices available for delivery of ambulatory oxygen therapy, the technologies that they use to generate oxygen, the performance characteristics of each device, and the development status.

Methods: We used medical and engineering databases to identify peer-reviewed papers (eg, MEDLINE, IEEE). Gray literature was used to identify additional descriptions of ambulatory oxygen devices in military medicine, space exploration, or patents. The last search was conducted in September 2025. Documents that described a device that can deliver oxygen in an ambulatory context (defined as weighing less than 10 kg) and were written in English were included. Search results were screened for inclusion by 2 independent reviewers. Data were synthesized by descriptively mapping the performance of each product, the technology used, and the development status of emerging technologies.

Results: From 9702 records identified, a total of 166 met eligibility criteria (106 scientific publications and 60 gray literature). We identified 33 portable oxygen concentrators (POCs; 29 commercially available), 10 oxygen cylinders, and 6 portable liquid oxygen (LOX) devices. The POC products showed a trade-off between portability and oxygen delivery capacity (maximum flow rate ranging from 2.0 to 6.0 L/min; device weight ranging from 1.0 to 9.1 kg). Pressure swing adsorption with zeolite was the most common oxygen generation technology in POCs on the market. The mean maximum continuous operating time of POCs was 3.8 hours. Two prototype POCs (maximum flow rate of 4-6 L/min and device weight of 8-9 kg) were developed for space exploration using modified adsorbents. LOX devices were the lightest and had the longest continuous operating time. Innovations in delivery included the downsizing of a POC by using nanozeolite as an adsorbent and pulse oximeter oxygen saturation (SpO_2)–targeted automatic titration of oxygen delivery based on the user's SpO_2.

Conclusions: This scoping review is the first study to integrate medical, engineering, and gray literature on ambulatory oxygen devices and their development. Although prior literature has narratively explained the products and technologies, no previous research has systematically investigated them. This review showed that POCs available to consumers may not meet the needs of

XSL•FO
RenderX

patients in terms of flow rate, portability, and operating time. LOX devices offered superior performance but are limited by high costs. Limitations of this review include the difficulty of comparing product performance across oxygen delivery settings and that the records were largely obtained from English-language sources. Innovation in ambulatory oxygen technology has been limited over the past decade, highlighting urgent need for research and development of new lightweight devices with higher oxygen delivery.

Trial Registration: OSF Registries 10.17605/OSF.IO/QS7FX; https://osf.io/qs7fx

(J Med Internet Res 2026;28:e81077) doi:10.2196/81077

KEYWORDS

ambulatory oxygen therapy; automatic titration of oxygen; home oxygen therapy; liquid oxygen; medical device innovation; oxygen cylinder; portable oxygen concentrator; pressure swing adsorption; zeolite

Introduction

Background

Supplemental oxygen may be prescribed to correct severe hypoxemia at rest, with exertion, and/or during sleep. Ambulatory oxygen therapy is defined as the use of supplemental oxygen during exertion [1]. The American Thoracic Society guideline conditionally recommends prescribing ambulatory oxygen therapy for patients who have severe exertional room air hypoxemia [1], based on acute improvements in exercise capacity and modest evidence of improvements in health-related quality of life [2,3]. Portable oxygen concentrators (POCs), oxygen cylinders, and liquid oxygen (LOX) can be used to deliver ambulatory oxygen therapy [1]. Oxygen cylinders for ambulatory use can deliver continuous oxygen flow up to 6 L per minute but run out quickly, especially when used at high flow rates, and need to be refilled. POCs supply concentrated oxygen by removing nitrogen from the air as long as they have a power source; however, they may provide oxygen only intermittently (“pulse flow” triggered by inspiration) and may have limited battery life. In continuous-flow settings, oxygen is delivered throughout both inhalation and exhalation, resulting in substantial oxygen waste during exhalation. Pulse-flow settings were developed to minimize this waste by supplying oxygen only during inspiration, thereby conserving both oxygen and battery power [4]. However, if the pulse is not synchronized well with inspiration, a portion of the oxygen bolus may be exhaled before reaching the alveoli [5]. LOX supplies oxygen by gradually evaporating LOX at cryogenic temperatures and can deliver higher flow rates for longer periods. However, LOX devices are costly, and availability may be limited [6].

In previous studies, users of portable oxygen devices reported a lack of physically manageable portable systems, a lack of devices capable of delivering higher oxygen-flow rates (>3 L/min), and an inability to leave their homes for more than 2-4 hours due to lack of reliable and enduring ambulatory oxygen supply [7-9]. In addition, users may face large out-of-pocket expenses for the ongoing costs of oxygen equipment [7]. Given these limitations, people using ambulatory oxygen reported that it was important to have a variety of device options that allowed a choice based on their individual needs [9]. However, there is little information on the types of devices available, the performance of each device, and the costs of ambulatory oxygen therapy devices [7]. It is possible that technologies to deliver ambulatory oxygen therapy are available in other fields (eg, military medicine) that are not yet available as commercial ambulatory devices.

A scoping review was deemed most appropriate to investigate the current status of portable oxygen devices, as the concepts of interest (technology, performance characteristics, and cost) extend beyond the medical field and may be published outside traditional peer-reviewed medical literature (eg, patent documents, technical reports) [10]. This scoping review aimed to identify the range of available portable oxygen devices, the technologies that are used to generate oxygen by the devices, the performance of each device, and the costs associated with its use. In addition, given the limitations of current portable oxygen devices, it is clinically important to clarify the status of technological innovation. Some patients require high oxygen flow rates that exceed the capacity of the currently available portable oxygen devices [11].

Furthermore, it is recognized that the weight of portable oxygen systems may limit their usability in clinical practice [12]. This may reduce adherence and contribute to inconsistent evidence for efficacy in exertional hypoxemia [11,13]. Therefore, we also aimed to identify innovations in the design of ambulatory oxygen therapy devices, such as improvements in weight or flow rates, and pulse oximeter oxygen saturation (SpO_2)–targeted automatic titration of oxygen delivery, which may not be available commercially [14].

Objectives

This scoping review aimed to identify the range of available portable oxygen devices, the technologies that are used to generate oxygen by the devices, the performance of each device, and the costs associated with its use. We also aimed to identify innovations in the design of ambulatory oxygen therapy devices, such as improvements in weight or flow rates and SpO_2-targeted automatic titration of oxygen delivery, which may not be available commercially.

Methods

Overview

This scoping review was conducted according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guideline (Multimedia Appendices 1-3; [15]), and the Joanna Briggs Institute (JBI) Manual for Evidence Synthesis [16]. The

XSL•FO
RenderX

protocol was registered prospectively with the Open Science Framework on May 29, 2024.

Eligibility Criteria

We included English language documents that described a device that can deliver oxygen in an ambulatory context, including scientific papers (including review articles) and gray literature (conference proceedings, patent documents, company websites, technical reports, and government documents). We excluded documents that only described stationary equipment, high-flow nasal oxygen therapy, hyperbaric oxygen therapy, positive pressure ventilation systems, heliotherapy, short-burst oxygen therapy, and wall-based high-flow nasal oxygen therapy, as well as those limited to over-the-counter medical devices.

Search Strategy

An initial search was performed on May 29, 2024, in MEDLINE (Ovid) and IEEE Xplore. After the initial search, the text words in the titles and abstracts of the retrieved articles and the index terms used to describe the articles were analyzed to refine the search terms. For the second search using updated terms, we searched scientific papers in 5 databases (MEDLINE [Ovid], Embase [Ovid], SCOPUS, CENTRAL [Wiley], and IEEE Xplore) using the search strings created with index terms (MeSH [Medical Subject Headings] terms or IEEE terms). Search strings are shown in Table S1 in Multimedia Appendix 4.

For searching gray literature, the following three strategies were used based on the guideline for gray literature and a previous study [17,18]: (1) targeted website browsing and searching, (2) gray literature database search, and (3) search engine searching. Specifically, technical reports, white papers including military medicine, and other gray literature on portable oxygen devices (manuals, company websites on product performance, etc) were searched using the appropriate websites respectively (National Technical Reports Library, World Health Organization, Defense Technical Information Center, International Health technology Assessment Database, and Google Advanced). As complex search strings were not allowed in some gray literature databases, we adjusted the search strings to match the database functionality. Gray literature search strings are also shown in Table S1 in Multimedia Appendix 4. Google Advanced searches were run on corporate, government, and military domains (.com, .gov, and .mil) to review the top 100 results from each. The latest patents of inventions and developments for the period January 1, 2022, to June 25, 2024, were searched in the World Intellectual Property Organization (WIPO) database using adjusted search strings. The search was repeated on September 29, 2025, for CENTRAL, MEDLINE (Ovid), and Embase (Ovid), on September 25, 2025, for all other databases including Scopus, IEEE Xplore, and gray literature sources, and on September 29, 2025, for WIPO.

A third search consisted of hand searching reference lists of all selected articles and review papers to identify any additional articles not found by the first two search methods. Additionally, Google Advanced searches were performed to obtain missing product performance information. The search period in each database was 2004 to the present to capture devices that were currently or recently available. We reported database searches and literature selection according to PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Search extension) [19].

Selection of Studies

Scientific papers were imported from the databases into the Covidence platform [20], and gray literature search results were imported into a spreadsheet. In the Covidence platform, duplicate records were automatically removed. Two independent reviewers (SK and MH) screened titles and abstracts of scientific papers for eligibility. Studies that met the inclusion criteria, or for which eligibility was unclear, underwent full-text review. Gray literature records were also screened for eligibility by 2 independent reviewers (SK and MH). Disagreements in study selection were resolved by consensus or by consulting a third reviewer (AEH). The reasons for exclusion at the full-text stage are reported.

Outcomes of Interest

Outcome selection was guided by patient priorities for oxygen devices, identified in previous studies and are shown in Table 1 [7-9].

XSL•FO
RenderX

Table 1. Outcomes of interest in the scoping review of ambulatory oxygen therapy devices extracted from scientific and gray literature published between 2004 and 2025.

Category	Outcome
Portable oxygen devices and their performance characteristics	• Type of device (POCs[a], portable oxygen cylinders, portable LOX[b]) • Product name • Manufacturer • Release date • Size of device • Weight of device (and battery weight where relevant) • Pulse or continuous flow • Method of transporting the device • Range of flow rates in the pulse flow setting • Range of flow rates in continuous flow setting • Pulse-dose bolus volume in maximum pulse flow setting (as much as possible at a respiratory rate of 20 breaths per minute) • Maximum continuous operating time (in the case of POC, battery duration, and in the case of oxygen cylinders or LOX, the duration until the oxygen stored in the tank runs out) • Concentration of supplied oxygen[c] • Operating noisec • Battery recharge time • Trigger sensitivity • Technology used to deliver oxygen (eg, nature of the sorbent in POC) • Product availability to consumers • FDA[d] and EMA[e] approved or equivalent • FAA[f] approved for air travel or equivalent • Frequency of product failures • Remote control function for changing flow rate
Cost of portable oxygen devices	• Description of publicly available supplier costs for each device (eg, purchase and rental costs of equipment, electricity costs for its use, etc)[g]

[a]POC: portable oxygen concentrator.

[b]LOX: liquid oxygen.

[c]Concentration of supplied oxygen and operating noise were included as they are important for the International Organization for Standardization (ISO) requirements to be approved as a medical device through the United States Food and Drug Administration and European Medicines Agency [21].

[d]FDA: Food and Drug Administration.

[e]EMA: European Medicines Agency.

[f]FAA: Federal Aviation Administration.

[g]The cost information, only officially disclosed by the company of each product, was extracted.

Data Charting Process

Data from the included sources of evidence were charted using a custom-designed form (SK and MH). Two review authors (SK and MH) independently charted the data from the eligible studies. Disagreements regarding the data charting between authors were resolved by discussion. If consensus could not be reached, a third author (AEH) reviewed the study and arbitrated.

The data chart included: types of publication (scientific paper, patent documents, technical report, etc), types of content (products, performances, costs, etc), author, year of publication, country where the information originated, information related to the products and performances outlined, and information related to the costs.

Risk of Bias (Quality Assessment)

Scoping reviews are conducted to provide an overview of the existing evidence regardless of methodological quality or risk of bias [22]. Therefore, the included sources of evidence are typically not critically appraised for scoping reviews. As such, we did not undertake a quality assessment of the included sources of evidence.

Data Synthesis

To synthesize data on products and their performance, we prioritized information published by the product company (including user manual), scientific papers, white papers, technical reports, and websites, in that order, if there was different information on the same item regarding one product. If multiple products were identified as having the same product name but differing only in numbers, the newest product was checked at the official website, and only the newest one was extracted (eg, Eclipse 1/2/3/4/5 Caire, Ball Ground, United States). As stationary oxygen concentrators are generally considered to weigh 10 kg or more, and POCs range from 1-9 kg [4], we defined ambulatory oxygen devices in this study as those weighing less than 10 kg and excluded heavier devices. Performance characteristics, the technology used, and the current status of development were reported descriptively. Costs for ambulatory oxygen therapy were mapped descriptively by country and type of device (portable oxygen cylinders, POCs,

XSL•FO
RenderX

and portable LOX). Descriptive statistics were used to describe all data. Categorical data were presented as frequency and percentages. In addition to descriptive synthesis, we developed a gap map to describe areas in need of future development. The gap map addressed the degree to which key patient needs for ambulatory oxygen devices (oxygen flow rate, device weight, operating time, and auto-titration) are met by characteristics of current products or those in development. Patients' needs extracted from previous studies were used as the framework [7-9].

Results

Principal Findings

A total of 9702 records were identified from scientific databases (medical database: n=3720; engineering database: n=133; multidisciplinary database: n=1842), and 4007 records from gray literature. After removing duplicates, a total of 3028 records from scientific databases and 4007 records from gray literature sources were screened (Figure 1). Of these, a total of 166 records related to ambulatory oxygen therapy devices were finally included in this review (n=106 from scientific databases; n=60 from gray literature sources). Among the 166 included records, a total of 81 (48.8%) originated from North America, 37 (22.3%) from Europe, 31 (18.7%) from Asia, 11 (6.6%) from Oceania, 2 (1.2%) from Africa, and 2 (1.2%) from South America (Table S2 in Multimedia Appendix 4). The scientific literature included 85 original research articles, 20 review articles (of which 1 was a systematic review and 19 were narrative reviews), and 1 clinical guideline. The gray literature comprised 7 technical reports, 3 white papers, 3 clinical trial registrations, 14 government-related websites (eg, military, energy, and space exploration), 4 company websites, 1 charitable organization website, and 28 patents (Tables 2 and 3). Although we identified 3 chemical oxygen generators in the scientific literature, they were excluded because chemical oxygen generators exhaust in 30 minutes or less and their output cannot be adjusted, making these devices unsuitable for delivery of ambulatory oxygen in clinical care [23].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) flow diagram showing the identification, screening, and inclusion of studies in this scoping review of ambulatory oxygen therapy devices. The search included scientific and gray literature from 2004 to 2025 across medical, engineering, military, and space exploration fields.

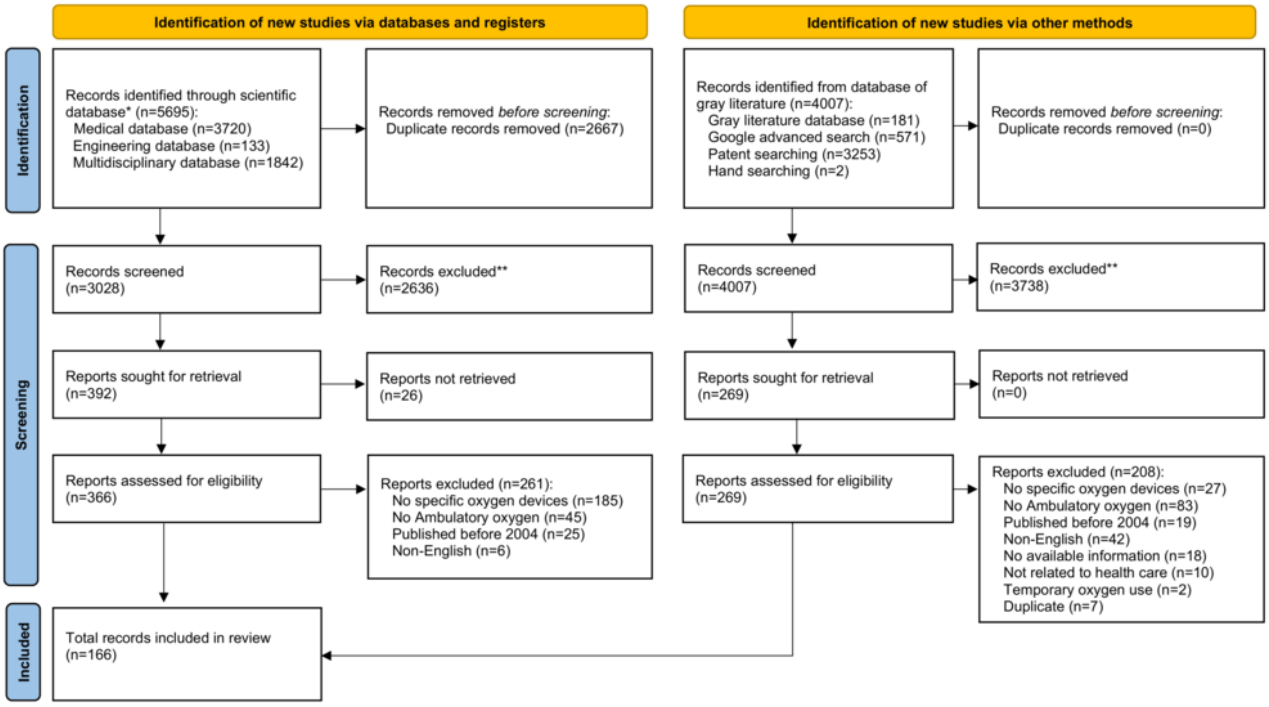

XSL•FO
RenderX

Table 2. Summary of literature on ambulatory oxygen therapy devices (n=166). Types of sources identified in this scoping review include scientific articles, technical reports, patents, government and military documents, and company websites published between 2004 and 2025.

Type of literature	Records, n (%)
Scientific literature (n=106)	
Original paper	85 (51)
Review paper	20 (12)
Systematic	1 (1)
Narrative	19 (11)
Guideline	1 (1)
Gray literature (n=60)	
Technical report	7 (4)
White paper	3 (2)
Clinical trial registry	3 (2)
Government website	14 (9)
Military	8 (5)
Energy	5 (3)
Space exploration	1 (1)
Company website	4 (2)
Charitable organization	1 (1)
Patent	28 (17)

Table 3. Summary of content from eligible literature (n=166) included in this scoping review of ambulatory oxygen therapy devices. Information sources span medicine, engineering, energy, military, and space exploration fields and describe portable oxygen concentrators, oxygen cylinders, and liquid oxygen systems, including innovations in portability, oxygen generation, and automatic titration technologies (2004-2025).

Type of contents	Records, n
Source of information	
Medicine	87
Engineering	63
Military	10
Energy	5
Space exploration	1
Device	
Oxygen concentrator (including devices not on market)	156
Oxygen cylinders	50
Liquid oxygen	16
New technology	
Downsize and portability enhancement	13
Improving or optimizing the oxygen generation process	15
Remote control (ie, automatic titration)	17
Other	12

Portable Oxygen Devices

We identified 33 different POCs in this study, of which 29 were on the market, 1 had discontinued production, and 3 were prototypes under development. The summary of identified POCs is shown in Table 4. The weight of the POCs products ranged from 1.0 to 9.1 kg, and size ranged from 15.7 × 11.7 × 6.1 cm (756 cm^3) to 51.3 × 27.7 × 20.3 cm (28,847 cm^3). Oxygen delivery capacity varied widely, with maximum flow rate in the continuous flow setting ranging from 2.0 to 6.0 L/min and maximum pulse-dose bolus volume of pulse flow setting ranging from 17.3 to 90 mL of oxygen. The pulse-dose bolus volume

XSL•FO
RenderX

varies widely across products, even within the same pulse flow setting (Figure 2). A trade-off between portability and oxygen delivery was shown in the identified POCs (Figure 3, Table S3 in Multimedia Appendix 4). Smaller POC products tended to be transported using bags, while larger POC products were transported using a dedicated cart (Table S3 in Multimedia Appendix 4). Smaller POC products also tended to have only pulse flow settings. The most recent POCs were the OxLife Liberty2 (25.4 × 22.9 × 8.9 cm), released in 2024, and the DISCOV-R (23.6 × 10.4 × 25.2 cm), released in 2023, which were capable of delivering oxygen in the continuous flow setting. This contrasts with earlier POCs of the same size (eg, Inogen Rove 6, 18.3 × 8.3 × 20.5 cm with single battery), which only had a pulse flow setting (Table S3 in Multimedia Appendix 4). The mean maximum continuous operating time in pulse flow setting was 3.8 hours, with a range of 1.3 to 6.3 hours, depending on flow settings and battery option (Table 4 and Figure 4). The maximum operating time tended to be shortened by half in the continuous flow setting compared to the same flow rate in the pulse flow setting in terms of equivalent volume (Table 4 and Table S3 in Multimedia Appendix 4). There were also POC products, such as the Simply Go mini, which extended the continuous operating time from 2 hours 40 minutes to 5 hours with an additional battery (Table S3 in Multimedia Appendix 4).

Pressure swing adsorption (PSA), vacuum pressure swing adsorption (VPSA), and vacuum pressure cycle (VPC) were the technologies used for oxygen generation (Table 4, Table S4 in Multimedia Appendix 4). Zeolites were identified as the adsorbent used in POCs. Air is composed of approximately 80% nitrogen, and PSA is a method to generate high concentrations of oxygen by selectively adsorbing nitrogen from air. In PSA, air pressurized by a compressor is sent through an adsorption column with zeolite to adsorb nitrogen and increase oxygen concentration. Then, the pressure is reduced to release the nitrogen. This adsorption and desorption cycle process is alternated in the multiple adsorption columns to continuously produce highly concentrated oxygen [24]. VPSA and VPC are adsorption methods that use a vacuum pump in combination with or instead of a compressor to reduce the pressure in the column [25]. We identified 2 POC prototypes and 1 POC product that were under development. The 4-SLPM prototype developed by TDA Research (Wheat Ridge, United States) weighed approximately 7.8 kg and measured 28 × 25 × 18 cm (Table S3 in Multimedia Appendix 4). The 4-SLPM has a maximum oxygen flow rate of 4 L/min in continuous flow setting, greater than most POCs on the market of comparable weight. The 4-SLPM used high lithium-exchanged X (LiLSX) as the adsorbent. The other PSA-based prototype, co-developed by NASA and Chart Industries (Ball Ground, United States), weighed less than 8.2 kg with a size of 35.6 × 30.5 × 20.3 cm and was capable of delivering up to 6 L/min in the continuous flow setting. There was no available information on commercially available POC innovations, such as remote control of flow settings in POCs. A POC product under development named JUNO (Roam Technologies Pty Ltd, Carlton NSW, Australia) was identified. According to the company's product information, JUNO is an ultraportable, tankless oxygen system currently under clinical development. Despite being designed to be carried with one hand, it has a continuous flow setting ranging from 1-3 L/min with a concentration of 91% [26]. Although the measurement conditions and detailed specification are not publicly disclosed (Table S3 in Multimedia Appendix 4), the recent patent on downsizing of POCs by Roam Technologies Pty Ltd, has been identified. The patent adopts a miniaturized PSA architecture with a compact arrangement of the adsorbent layer and internal gas pathway structure to shorten flow paths and minimize dead space [27].

Oxygen cylinders and LOX are summarized in Tables 5 and 6. Table 5 and Tables S3, S4, and S5 (shown in Multimedia Appendix 4) are also provided as xlsx files in Multimedia Appendix 5. The weight of cylinders ranged from 0.3-6.5 kg, and the size ranged from 14.9 × 6.4 cm to 86.5 × 10.2 cm. The maximum continuous operating time varied from 0.3 to 5.0 hours at 2 L/min in the continuous flow setting. The continuous operating time per continuous flow setting was comparable to the POC. (Figure 4). Smaller cylinders were transported using a shoulder bag, whereas larger cylinders required trolleys or hand-drawn carts for transport. Conventional cylinders were generally filled at approximately 150 bar, whereas the Ultra Lightweight Cylinder Oxygen System (IOSVR) uses a high-pressure filling of 300 bar. This allows nearly double the amount of oxygen to be filled compared with a comparably sized type M7 cylinder (Table 6). IOSVR is made possible by the high-strength aluminum alloy (L7X), which is reinforced with carbon fiber wrapping technology [28]. LOX products weighed from 1.6-3.7 kg (filled) and had gaseous oxygen capacities ranging from 275.0-1058 L (Table 6). Many oxygen cylinders require a pressure regulator, which is attached to the cylinder's top and works like a tap, allowing the safe adjustment of oxygen flow rate provided, in L/min [4]. In addition, some regulators support a pulse-dose delivery mode, which can extend the operating duration compared with continuous flow [4,29].

XSL•FO
RenderX

Table 4. Summary of identified portable oxygen concentrators (n=33) included in this scoping review of ambulatory oxygen therapy devices. Extracted characteristics include device weight and dimensions, flow settings, oxygen output, operating noise, trigger sensitivity, oxygen delivery technology, regulatory approval, and market availability based on records published between 2004 and 2025.

Variables	Statistic	Notes
Device volume (cm^3), mean (range)	8342.4 (755.9-28,846.5)	n=32
Weight (kg), mean (range)	3.8 (1.0-9.1)	n=16 (including battery); n=16 (unclear included)
Transport method, n	15	n=30; some devices fall into multiple categories
Carrying bag	2	
Carry case	2	
Backpack	6	
Cart	3	
Shoulder strap	2	
Waist	1	
Carry-on baggage	3	
PF[a] and CF[b] setting, n		
Only PF	19	N/A[c]
Both CF and PF	11	N/A
Not reported	2	N/A
Maximum number of PF setting, mean (range)	5 (1-10)	n=29
Maximum flow rates in CF setting (L/min), mean (range)	2.9 (2.0-6.0)	n=12 (only devices with CF settings)
Maximum pulse-dose bolus in PF setting (mL), mean (range)	53.7 (17.3-90.0)	n=15 (Only those at 20 bpm)
Maximum continuous operating time in PF setting (h)[d], mean (range)	3.8 (1.3-6.3)	n=25; depends on settings or respiratory rate
Maximum continuous operating time in CF setting (h)[d], mean (range)	1.9 (0.6-4.5)	n=9; depends on settings or respiratory rate
Concentration of supplied oxygen (%)[d], mean (range)	90.4 (82.0-96.0)	n=29; depends on settings or respiratory rate
Operating noise (dBA[e])[d], mean (range)	44.3 (35.0-59.0)	n=28; depends on settings or respiratory rate
Maximum battery recharge time (h), mean (range)	4.4 (1.5-12.8)	n=27
Trigger sensitivity (cm H_2O), mean (range)	–0.24 (–0.50 to –0.05)	n=23
Nature of the sorbent, n		
Molecular sieve	14	N/A
Zeolite	3	N/A
High lithium-exchanged X	1	N/A
Not reported	14	N/A
Technology used, n		
Pressure swing adsorption	7	N/A
Vacuum pressure swing adsorption	1	N/A
Vacuum pressure cycle	1	N/A
Not reported	23	N/A
Market availability, n		
On the market	29	N/A
Discontinued	1	N/A
Prototype	2	N/A

XSL•FO
RenderX

Variables	Statistic	Notes
FDA[f], n		
Approved	19	N/A
Not reported	13	N/A
FAA[g], n		
Approved	28	N/A
Not reported	4	N/A

[a]PF: pulse flow.

[b]CF: continuous flow.

[c]N/A: not applicable.

[d]JUNO (Roam Technologies Pty Ltd, Carlton NSW, Australia) was excluded from this analysis due to a significant lack of information.

[e]dBA: A-weighted decibels.

[f]FDA: Food and Drug Administration.

[g]FAA: Federal Aviation Administration.

Figure 2. Scatter plot showing the relationship between maximum pulse flow setting and pulse-dose bolus volume (mL) in portable oxygen concentrators identified in this scoping review of ambulatory oxygen therapy devices. Data were extracted from publicly available scientific and gray literature published between 2004 and 2025. Only products with publicly available data on pulse-dose bolus volume at 20 breaths per minute at the maximum pulse flow setting are shown.

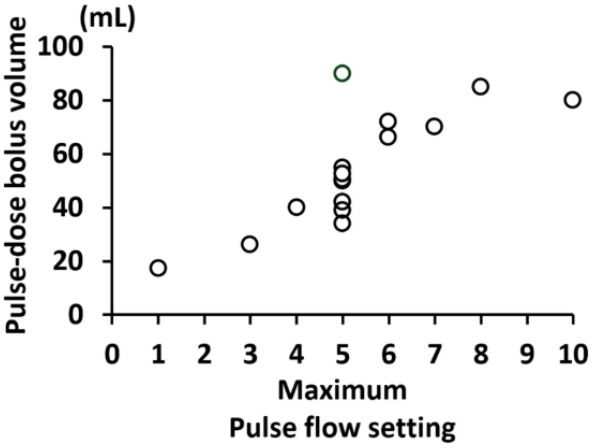

Figure 3. Scatter plot showing the relationship between product weight and pulse-dose bolus volume (mL) at the maximum pulse flow setting in portable oxygen concentrators identified in this scoping review of ambulatory oxygen therapy devices. Data were extracted from Lpublicly available scientific and gray literature published between 2004 and 2025. Only products with publicly available data on pulse-dose bolus volume at 20 breaths per minute at the maximum pulse flow setting are shown.

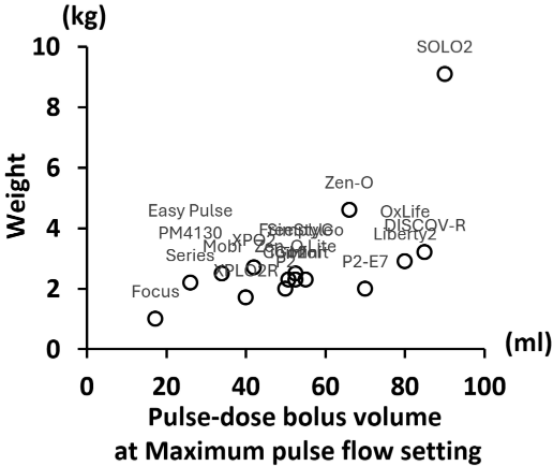

XSL•FO
RenderX

Figure 4. Continuous operating time at different continuous flow settings for (A) portable oxygen concentrators, (B) oxygen cylinders, and (C) liquid oxygen devices identified in this scoping review of ambulatory oxygen therapy devices. Data were extracted from publicly available product manuals and literature published between 2004 and 2025. Only products with publicly available data on continuous operating time for each continuous flow setting are shown.

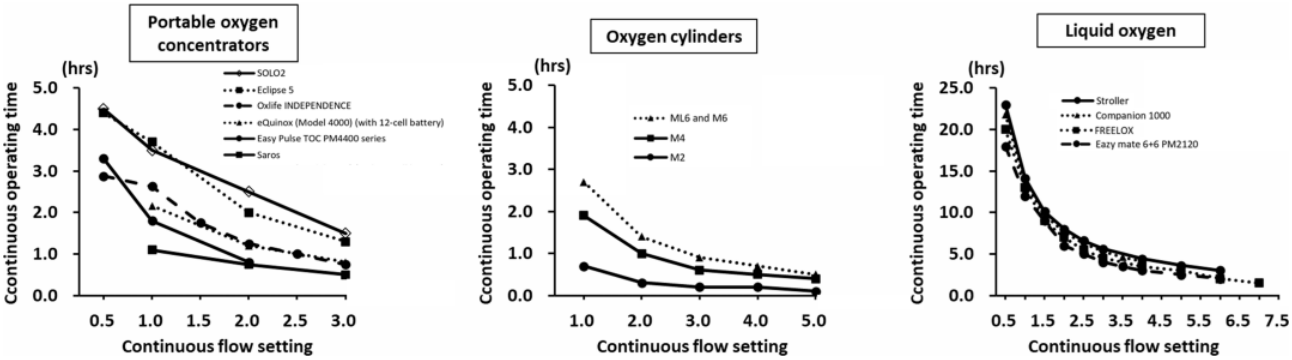

Table 5. Summary of identified oxygen cylinders included in this scoping review of ambulatory oxygen therapy devices. Data were extracted from patents, manuals, and technological reports published between 2004 and 2025 and include cylinder size, weight, oxygen capacity, nominal pressure, maximum operating time, and transport methods (1 bar=100 kPa).

Type	Size (cm)	Weight (kg)	Content (L)	Nominal pressure (bar)	Maximum continuous operating time	Transport method
M2	14.9 × 6.4	0.34	45	153[a]	0.3 h at 2 L/min[a]	Not reported
M4	22.1 × 8.1	1.0	113	153[a]	1 h at 2 L/min[a]	Not reported
ML6	20.0 × 11.1	1.3	170	139[a]	1.4 h at 2 L/min[a]	Shoulder bag
M6	30.0 × 8.2	1.3	170	153[a]	1.4 h at 2 L/min[b]	Shoulder bag
M7	23.1 × 11.1	1.5	198	139[a]	1.7 h at 2 L/min[a]	Shoulder bag
IOSVR	34.2 × 8.5	1.7	612	300[§]	8 h at 2 L/min with a standard conserver[c]	Handheld or bag
M9	27.7 × 11.1	1.7	255	139[a]	2.1 h at 2 L/min[a]	Shoulder bag
D	53.5 × 10.2	3.9	340	140[a]	3.6 h at 2 L/min[a]	Shoulder bag
CH/C	51.5 × 11.7	4.2	470	163 or 250[a]	Not reported	Medical trolleys or carry bags
E	86.5 × 10.2	6.5	680	137[a]	5 h at 2 L/min[c]	Hand-drawn cylinder cart

[a]Data obtained from nonmanufacturer websites.

[b]Data reported in the technical reports.

[c]Data reported in scientific reports.

LOX had a longer continuous operating time compared to POCs and cylinders, with a mean maximum continuous operating time of approximately 7.4 hours even at 2 L/min of continuous flow setting (Figure 4). There was no available information on the pulse-dose bolus volume of pulse flow setting in LOX.

XSL•FO
RenderX

Table 6. Summary of identified liquid oxygen systems included in this scoping review of ambulatory oxygen therapy devices. Data were extracted from manuals, patents, and technical reports published between 2004 and 2025 and include weight, oxygen capacity, flow settings, maximum operating duration, and regulatory approval.

Variables	Statistic	Notes
Device volume (cm^3), mean (range)	7889.2 (3650.2-13879.1)	n=6
Weight (kg), mean (range)	3.00 (1.6-3.7)	n=6
Transport method, n		Some devices fall into multiple categories
Backpack	3	
Cart	3	
Carry bag	1	
Belt pack	3	
Handle	1	
Gaseous oxygen capacity (L), mean (range)	699.4 (275.0-1058.0)	N/A[a]
PF[b] or CF[c] setting, n		
Only PF	3	N/A
Both CF and PF	3	N/A
Maximum number of PF setting (n), mean (range)	5 (4-6)	n=3
Maximum flow rates in CF setting (L/min), mean (range)	4.3 (0.8-7)	n=6
Maximum continuous operating time in CF setting (h), mean (range)	7.4 (6.0-9.0)	n=5 (only those at 2 L/min of CF setting)
FDA[d], n		
Approved	5	N/A
Not approved	1	N/A

[a]N/A: not applicable.

[b]PF, pulse flow.

[c]CF, continuous flow.

[d]FDA, Food and Drug Administration.

Developments and New Technologies

Improving Portability

A patent was published in 2022 regarding the use of the vacuum swing adsorption method instead of the conventional PSA method in POCs to generate oxygen more efficiently [30]. By replacing the compressor in the PSA method with a fan or other air-moving device, air at atmospheric pressure instead of high pressure is passed through the zeolite adsorbent to adsorb nitrogen, and then the pressure is reduced to a vacuum level, which improves the efficiency of nitrogen desorption, resulting in improved oxygen generation rates [25]. This low-pressure operation makes it possible to use a small vacuum pump without a compressor in PSA, thus reducing the weight of the device and power consumption. Specifically, the POC in this patent had a device weight of 1.3 kg and a maximum oxygen flow rate of 1.5 L/min in the continuous flow setting. In addition, the POC in this patent uses LiLSX, which has recently been developed with improved selectivity and adsorption capacity for nitrogen [25,30,31]. A technology has been incorporated to maximize the efficiency of oxygen generation by monitoring the work of the zeolite adsorbent and the flow rate between the inlet and outlet of the adsorption column in real time to accurately detect the point when the adsorbent is saturated with nitrogen during the adsorption process (breakthrough point) and switch the cycle just before the saturated nitrogen is mixed with the concentrated oxygen. This allows oxygen to be separated without waste and maintains high oxygen purity while reducing the size of the device [30]. Another patent was published in 2023 for the design of a nasal-wearable POC entitled “Device and Method of Generating an Enriched Gas Within a Nasal Vestibule” [32]. However, few details were provided on the technical solutions used to achieve this small POC.

Improving or Optimizing the Oxygen Generation Process

A 2-bed rapid pressure swing adsorption (RPSA) system as an oxygen generation process was described in a scientific paper in 2021 [33]. The adsorption and desorption cycles performed in multiple adsorption columns in a conventional PSA system take a certain amount of time, which limits the amount of oxygen delivery. In this study, the cycle time was significantly reduced while maintaining a high oxygen concentration by optimizing the adsorbent and parameters such as pressure and cycle time related to adsorption. As a result, high oxygen

XSL•FO
RenderX

productivity was observed despite the much lower ratio of adsorbent volume required (bed size factor) [33]. Other scientific papers have shown that sensitivity analysis modelling of PSA systems has led to optimization of operating conditions such as adsorption pressure, cycle time and adsorption bed size, leading to their applicability as oxygen supply infrastructure in small medical facilities and in high-humidity environments [34,35].

Remote Control

A scientific report for SpO_2-targeted automatic titration of oxygen delivery during activities of daily living (ADL) testing based on the user's SpO_2 was identified [36]. This double-blinded randomized crossover trial evaluated the effect of SpO_2-targeted automatic titration during ADL testing in 31 patients with chronic obstructive pulmonary disease (COPD) on long-term oxygen therapy. In this trial, an oxygen cylinder was connected to the closed-loop device (O2matic, Herlev, Denmark), and this device was placed on a rollator during ADL testing. In the active arm, automatic titration of oxygen flow rate (0-8 L/min) was set to aim at keeping an SpO_2 target range of 90%-94% and those adjustments were done every second based on the average SpO_2 for the last 15 seconds. In the control arm, the flow rate was kept fixed according to each patient's medical prescription. As a result, automatic titration increased flow rates by triple and reduced ADL completion time, improved dyspnea, and reduced the number of events of severe desaturation compared to the control arm. Another technology for SpO_2-targeted adjustment of oxygen delivery based on the user's SpO_2 was identified [37]. An "intelligent oxygen concentrator" for patients with COPD has been developed as a system that measures physical activity levels and automatically adjusts oxygen delivery according to machine learning (eg, decision tree, probabilistic neuronal network, logistic regression) [38]. This machine learning model was trained using patient activity data to discriminate the patient's activity level (sedentary, light activity, and moderate activity). In a pilot study of 5 patients with COPD with long-term oxygen therapy, machine learning was used to automatically adjust the flow setting of pulse flow to match each patient's physical activity level. As a result, the cumulative time of SpO_2 below 90% when walking the circuit course was significantly reduced compared to manual changes [39].

Other Developments and New Technologies

A patent was identified for a new oxygen concentration control technique that delivers oxygen at low to moderate concentrations (eg, 30%-50%) by adjusting the pressure during the PSA cycle [40]. It was mentioned that the importance of this low-concentration oxygen delivery in clinical practice is unclear, but this patent could lead to reduced energy consumption of the POC, which has implications for extending battery life. In addition, a number of technologies for oxygen delivery control systems using sensors have been reported in recent years. One patent described sensors that use light to detect the timing of inhalation and valves that open and close electromagnetically at that timing to inject oxygen for a very short time (microbursts) [41]. This allows oxygen to be delivered with less waste of gas. In addition, a new delivery method was reported in which oxygen is generated and stored in advance and released in immediate response to inhalation [42].

Cost

There were no available records of the cost of POCs officially disclosed by POC companies. Two identified scientific papers referred to POC prices of approximately US $2000 and US $3000 as retail costs to end users [43,44]. There were also no records of specific costs for LOX and oxygen cylinders. However, an identified article reported that LOX was approximately 4 times as expensive as standard oxygen therapy using POCs or portable cylinders [45]. Another identified article indicated that LOX systems entail higher service costs due to the need for regular home refills [46].

Gap Map

The gap map (Figure 5) showed that only a limited range of oxygen cylinders met patient needs for high flow rates (>3 L/min) and light weight (<2.5 kg). Although current POCs did not meet this need, a one-hand carry device capable of delivering 3 L/min was under clinical development. In all device types, few products met patient needs for continuous operating time (>5-6 hours). For auto-titration, SpO_2-targeted automatic titration was available for POCs and cylinders as an interface, and the development of automatic titration capabilities embedded into the device itself has been reported only in POCs.

Figure 5. Gap map of patient needs and device availability in ambulatory oxygen therapy. The figure illustrates the extent to which current devices address 3 major patient needs: green indicates needs are met ("available"), yellow indicates limited devices meet patient needs ("limited availability"), orange indicates no current devices but technologies are under development ("in development"), and red indicates no available devices and no technologies under development ("not available").

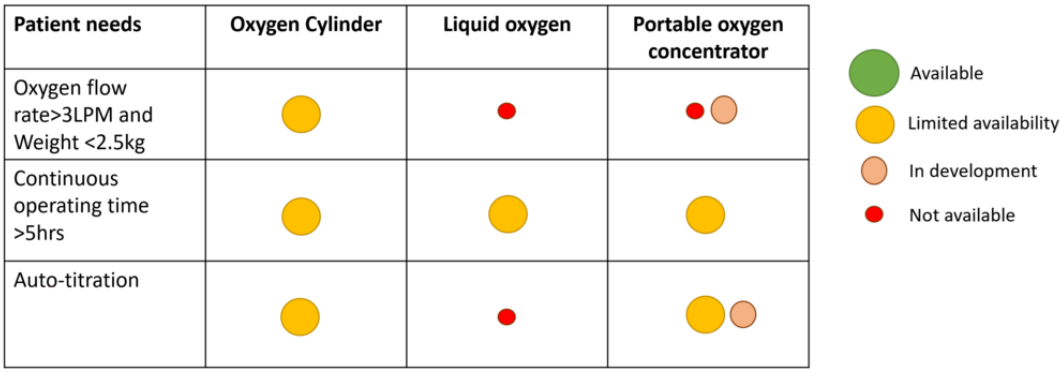

XSL•FO
RenderX

Discussion

Overview

This scoping review aimed to identify the range of portable oxygen devices, their oxygen generation technologies, performance characteristics, innovations in the design of ambulatory oxygen therapy devices, and cost. We identified 33 POCs, 10 oxygen cylinders, and 6 portable LOX systems. There was a trade-off between the portability (weight) and oxygen delivery capacity of POCs. The mean maximum continuous operating time of POCs was 3.8 hours. PSA with zeolite was the most common oxygen generation technology in POCs. Two POC prototypes were identified that had better oxygen delivery capacity despite being the same size as POCs on the market. LOX was the most portable and had the best continuous operating time among the devices. In terms of development and innovation, the downsizing of POCs using nanozeolite as the adsorbent, improving flow rate using RPSA and SpO_2-targeted automated oxygen titration based on to patient's SpO_2 and physical activity were identified. Costs were rarely disclosed by manufacturers, and published data indicated that LOX is substantially more expensive than POCs or cylinders.

This study showed that high-flow oxygen delivery of 3 L/min in the continuous flow setting in POCs is only possible from commercially available POCs weighing more than 7.2 kg (eg eQuinox, Table S3 in Multimedia Appendix 4), which need to be transported on a cart. This may be a burden for patients requiring high flow rates. Furthermore, the short continuous operating time of the POCs identified in this study may also limit daily activities and community engagement. Patients have previously identified a desirable operating time for portable oxygen devices of 5-6 hours [9]. In the identified POC products, the mean maximum continuous operating time was up to 3.8 hours in the pulse flow setting. As higher flow rates shorten the continuous operating time, it is likely that fewer devices will meet patients' needs. Although an additional battery can be used to double the operating time (eg, Simply Go), carrying the additional battery may be a burden for the patient. Furthermore, it should be noted that the numerical setting of a pulse flow does not correspond to the continuous flow rate and that the bolus volume varies between products even within the same pulse flow rate setting [47].

Among the devices, LOX had the best portability and continuous operating time, but the complicated maintenance was an obstacle. Previous studies have recognized LOX as a device with both better portability and continuous operating time [8], and have reported that LOX use is associated with better quality of life and physical activity compared to POCs and cylinders [48,49]. On the other hand, LOX requires refilling of liquid oxygen by a supplier at least 2-3 times a month, which imposes a high cost on the supplier and user [4,45]. As a result, the number of suppliers servicing LOX has been declining in the United States [8].

It was suggested that the trade-off between portability and oxygen delivery capacity is improving based on the latest products and prototypes identified. However, even in the latest POC products (DISCOV-R released in 2023 and OxLife Liberty2 released in 2024), the maximum oxygen flow rate of the continuous flow setting was still 2 L/min (Table S3 in Multimedia Appendix 4). JUNO, which is designed to be carried with one hand, is under clinical development and has a continuous flow setting ranging from 1-3 L/min with a concentration of 91% [26], and may be an innovative device, as the PSA Prototype was published in 2015 and the 4-SLPM in 2018 [50,51]. In terms of technological development, the nanozeolite (LilSX) had been of interest because of its use in 4-SLPM prototypes, patents, and basic research [30,52]. Although nanozeolites are characterized by higher nitrogen adsorption capacity, they are more expensive than conventional zeolites and have challenges such as long-term structural stability and degradation [53,54]. Although RPSA systems can improve oxygen productivity by shortening adsorption–desorption cycles, their application to POCs remains challenging because no clinical application studies have yet been reported [52]. Several recent clinical trials and patents described systems capable of adjusting flow rates based on the user's SpO_2 and physical activity. In a clinical trial with a small sample size, automated titration improved functional capacity in ADL, dyspnea, and the number of severe hypoxemia events [37]. However, these technologies have barriers, including sensor reliability or inaccuracy even during motion, requiring further development and clinical trials [37,55]. Identified patents in this study should be carefully considered because approximately half of the patents in general are not commercialized and launched on the market [56].

Limitations

Limitations of this scoping review include the following. First, it is difficult to compare performance such as pulse-dose bolus volumes and continuous operating time between POCs because they depend on the number of breaths and device settings [57]. Second, this scoping review included only information published in English, which may have underestimated information on devices developed and used in non–English-speaking countries. Third, the pulse-dose bolus volume of POC at maximum pulse flow setting (as much as possible at a respiratory rate of 20 breaths per minute) extracted in this study does not exactly indicate the product's oxygen delivery capacity. POC products vary in their algorithms for converting pulse-bolus volume and flow rate in response to changes in respiratory rate [57]. Fourth, available evidence regarding product performance relied almost on manufacturer-reported specifications, and no available evidence on device lifespan, failure rates, or long-term durability was available. There was no information available on the costs of POC products disclosed by manufacturers in this study. Lastly, the geographical distribution of the searched literature was biased toward the United States, as some gray literature was retrieved using domains such as ".gov," ".mil," and ".org" in Google Advanced Search.

Conclusions

This scoping review is the first study to integrate medical, engineering, and gray literature evidence on ambulatory oxygen devices and map current ambulatory oxygen devices and their development status. Although prior literature has narratively described products and technologies for ambulatory oxygen

XSL•FO
RenderX

therapy, no previous research has systematically investigated these products and new technologies. We identified 33 POCs, 10 oxygen cylinders, and 6 LOX systems. We showed the performance limitations of current devices and gaps in technological development in ambulatory oxygen therapy and suggested directions for future research and development. Specifically, POCs available to consumers may not meet patients' needs in terms of oxygen flow rate, portability, and operating time. LOX offered the best performance in terms of operating time and portability among the devices but is restricted by high costs and declining availability. Although POCs are the most widely developed devices, technological innovation to achieve high oxygen flow rates, better portability, and longer continuous operating time has been limited since the POC prototype published in 2015. Collaboration among device developers, researchers, health professionals, and patients is urgently required to develop new lightweight devices with greater oxygen delivery capacity. It is essential to incorporate consumer input from the early stages of design and testing to ensure that future portable oxygen devices meet patients' needs.

Acknowledgments

Generative artificial intelligence (ChatGPT) was used solely to assist in summarizing complex patent descriptions during the data extraction process. All outputs were reviewed and verified by the authors to ensure accuracy and to avoid hallucinations.

Funding

The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available in the Open Science Framework Storage from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: SK (lead), MH (equal), AEH (equal), ME (supporting), JAK (supporting)
Data curation: SK (lead), MH (equal), LR (supporting), AEH (supporting)
Formal analysis: SK (lead), MH (equal), LR (supporting), AEH (supporting)
Funding acquisition: AEH (lead), SK (equal)
Investigation: SK (lead), MH (equal), AEH (equal)
Methodology: SK (lead), MH (equal), AEH (equal)
Project administration: AEH (lead), SK (equal), MH (equal), ME (supporting), JAK (supporting)
Resources: SK (lead), MH (equal), AEH (equal)
Supervision: SK (lead), MH (equal), AEH (equal)
Validation: SK (lead), MH (equal), AEH (equal)
Visualization: SK (lead), MH (equal), AEH (equal)
Writing – original draft: SK (lead), AEH (equal), MH (supporting), ME (supporting), JAK (supporting)
Writing – review & editing: SK (lead), AEH (equal), MH (supporting), ME (supporting), JAK (supporting)

Conflicts of Interest

Unrelated to this work, 2 authors (JAK and ME) have relationships with Inogen and ResMed, respectively, which are manufacturers of portable oxygen concentrators (POCs). JAK reports receiving personal fees for Inogen to serve as a consultant on POCs, and ME reports receiving a research grant from ResMed. JAK also has received researchs grant from the National Institutes of Health, Patient-Centered Outcomes Research Institute, American Lung Association, COPD Foundation, and BioVie, and personal fees from AstraZeneca, DynaMed/EBSCO, Genentech, MedImmune, and Verona Pharmaceuticals. ME declares no conflicts of interest related to this work. Unrelated to this work, ME also has received personal fees from AstraZeneca, Boehringer Ingelheim, Novartis, and Roche. MH and AEH report receiving a grant from the National Health and Medical Research Council–Australia (GNT1139953). MH and AH report nonfinancial support from BOC Australia and Air Liquide Healthcare related to their research on ambulatory oxygen therapy for pulmonary fibrosis (PFOX trial).

Multimedia Appendix 1
PRISMA-ScR-Fillable-Checklist_revised.
[DOCX File , 86 KB - jmir_v28i1e81077_app1.docx]

Multimedia Appendix 2
PRISMA 2020 for Abstracts Checklist.
[DOCX File , 265 KB - jmir_v28i1e81077_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist.
[DOCX File , 21 KB - jmir_v28i1e81077_app3.docx]

Multimedia Appendix 4
Additional material.
[DOCX File , 106 KB - jmir_v28i1e81077_app4.docx]

Multimedia Appendix 5
The performance characteristics of each identified products.
[XLSX File (Microsoft Excel File), 25 KB - jmir_v28i1e81077_app5.xlsx]

References

1. Jacobs SS, Krishnan JA, Lederer DJ, Ghazipura M, Hossain T, Tan AM. Home oxygen therapy for adults with chronic lung disease: an official American Thoracic Society clinical practice guideline. Am J Respir Crit Care Med 2020;202(10):e121-e141 [FREE Full text] [doi: 10.1164/rccm.202009-3608ST] [Medline: 33185464]
2. Lacasse Y, Lecours R, Pelletier C, Bégin R, Maltais F. Randomised trial of ambulatory oxygen in oxygen-dependent COPD. Eur Respir J 2005;25(6):1032-1038 [FREE Full text] [doi: 10.1183/09031936.05.00113504] [Medline: 15929958]
3. Jarosch I, Gloeckl R, Damm E, Schwedhelm A, Buhrow D, Jerrentrup A. Short-term effects of supplemental oxygen on 6-min walk test outcomes in patients with COPD: a randomized, placebo-controlled, single-blind, crossover trial. Chest 2017;151(4):795-803. [doi: 10.1016/j.chest.2016.11.044] [Medline: 27940278]
4. Hardavella G, Karampinis I, Frille A, Sreter K, Rousalova I. Oxygen devices and delivery systems. Breathe (Sheff) 2019;15(3):e108-e116 [FREE Full text] [doi: 10.1183/20734735.0204-2019] [Medline: 31777573]
5. Chen J, Katz I, Pichelin M, Zhu K, Caillibotte G, Noga M. Comparison of pulsed versus continuous oxygen delivery using realistic adult nasal airway replicas. COPD 2017;Volume 12:2559-2571. [doi: 10.2147/copd.s141976]
6. Melani AS, Sestini P, Rottoli P. Home oxygen therapy: re-thinking the role of devices. Expert Rev Clin Pharmacol 2018;11(3):279-289. [doi: 10.1080/17512433.2018.1421457] [Medline: 29272974]
7. Tikellis G, Hoffman M, Mellerick C, Burge AT, Holland AE. Barriers to and facilitators of the use of oxygen therapy in people living with an interstitial lung disease: a systematic review of qualitative evidence. Eur Respir Rev 2023;32(169):230066 [FREE Full text] [doi: 10.1183/16000617.0066-2023] [Medline: 37611946]
8. Jacobs SS, Lederer DJ, Garvey CM, Hernandez C, Lindell KO, McLaughlin S. Optimizing home oxygen therapy: an official American Thoracic Society workshop report. Ann Am Thorac Soc 2018;15(12):1369-1381. [doi: 10.1513/AnnalsATS.201809-627WS] [Medline: 30499721]
9. Jacobs SS, Lindell KO, Collins EG, Garvey CM, Hernandez C, McLaughlin S. Patient perceptions of the adequacy of supplemental oxygen therapy: results of the American Thoracic Society Nursing Assembly Oxygen Working Group Survey. Ann Am Thorac Soc 2018;15(1):24-32. [doi: 10.1513/AnnalsATS.201703-209OC] [Medline: 29048941]
10. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
11. Johannson KA, Pendharkar SR, Mathison K, Fell CD, Guenette JA, Kalluri M. Supplemental oxygen in interstitial lung disease: an art in need of science. Ann Am Thorac Soc 2017;14(9):1373-1377. [doi: 10.1513/AnnalsATS.201702-137OI] [Medline: 28644693]
12. Ejiofor SI, Bayliss S, Gassamma A, Turner AM. Ambulatory oxygen for exercise-induced desaturation and dyspnea in chronic obstructive pulmonary disease (COPD): systematic review and meta-analysis. Chronic Obstr Pulm Dis 2016;3(1):419-434 [FREE Full text] [doi: 10.15326/jcopdf.3.1.2015.0146] [Medline: 28848863]
13. Long-Term Oxygen Treatment Trial Research Group, Albert RK, Au DH, Blackford AL, Casaburi R, Cooper JA. A randomized trial of long-term oxygen for COPD with moderate desaturation. N Engl J Med 2016;375(17):1617-1627 [FREE Full text] [doi: 10.1056/NEJMoa1604344] [Medline: 27783918]
14. Sanchez-Morillo D, Muñoz-Zara P, Lara-Doña A, Leon-Jimenez A. Automated home oxygen delivery for patients with COPD and respiratory failure: a new approach. Sensors (Basel) 2020;20(4):1178 [FREE Full text] [doi: 10.3390/s20041178] [Medline: 32093418]
15. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
16. JBI Manual for Evidence Synthesis.: JBI; 2024. URL: https://synthesismanual.jbi.global/ [accessed 2025-05-12]
17. Landerdahl Stridsberg S, Richardson MX, Redekop K, Ehn M, Wamala Andersson S. Gray literature in evaluating effectiveness in digital health and health and welfare technology: a source worth considering. J Med Internet Res 2022;24(3):e29307 [FREE Full text] [doi: 10.2196/29307] [Medline: 35319479]

XSL•FO
RenderX

18. Godin K, Stapleton J, Kirkpatrick SI, Hanning RM, Leatherdale ST. Applying systematic review search methods to the grey literature: a case study examining guidelines for school-based breakfast programs in Canada. Syst Rev 2015;4:138 [FREE Full text] [doi: 10.1186/s13643-015-0125-0] [Medline: 26494010]
19. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala A, Moher D, Page M. PRISMA-S: an extension to the PRISMA Statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
20. covidence. URL: https://www.covidence.org/ [accessed 2025-12-28]
21. WHO technical specifications for oxygen concentrators. World Health Organization. URL: https://www.who.int/publications/i/item/9789241509886 [accessed 2025-01-09]
22. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc 2015;13(3):141-146. [doi: 10.1097/XEB.0000000000000050] [Medline: 26134548]
23. Blakeman TC, Rodriquez D, Britton TJ, Johannigman JA, Petro MC, Branson RD. Evaluation of oxygen concentrators and chemical oxygen generators at altitude and temperature extremes. Mil Med 2016;181(5 Suppl):160-168. [doi: 10.7205/MILMED-D-15-00130] [Medline: 27168568]
24. Yadav VK, Choudhary N, Inwati GK, Rai A, Singh B, Solanki B. Recent trends in the nanozeolites-based oxygen concentrators and their application in respiratory disorders. Front Med (Lausanne) 2023;10:1147373 [FREE Full text] [doi: 10.3389/fmed.2023.1147373] [Medline: 37181347]
25. Ackley MW. Medical oxygen concentrators: a review of progress in air separation technology. Adsorption 2019;25(8):1437-1474. [doi: 10.1007/s10450-019-00155-w]
26. Tankless Oxygen Live Life Without Limits.: Roam Technologies URL: https://www.roamtech.ai/ [accessed 2025-12-24]
27. Shan-Shan WE, Sisi Z, Jay RF, Jeremy TK, Natalie EH, Gavin DM. Oxygen concentrator. WIPO.: Roam Technologies Pty Ltd; 2024. URL: https://patentscope2.wipo.int/search/en/detail.jsf?docId=US451301825 [accessed 2025-12-24]
28. Bruton A, Sinclair I, Arnold E, Hepples W, Kay F, Maisey G. The design and development of a new light-weight portable oxygen system. J Med Devices 2012;6(3):031007. [doi: 10.1115/1.4007180]
29. Palwai A, Skowronski M, Coreno A, Drummond C, McFadden ER. Critical comparisons of the clinical performance of oxygen-conserving devices. Am J Respir Crit Care Med 2010;181(10):1061-1071 [FREE Full text] [doi: 10.1164/rccm.200910-1638OC] [Medline: 20133925]
30. Jagger TW, Van Brunt NP, Kivisto JA, Lonnes PB. Removable cartridge for oxygen concentrator. WIPO. 2022. URL: https://patentscope2.wipo.int/search/en/detail.jsf?docId=US105472432&_cid=JP2-MIE7CN-34520-1 [accessed 2025-12-24]
31. Arora A, Hasan MMF. Flexible oxygen concentrators for medical applications. Sci Rep 2021;11(1):14317 [FREE Full text] [doi: 10.1038/s41598-021-93796-3] [Medline: 34253838]
32. Assani KD. Device and method of generating an enriched gas within a nasal vestibule. WIPO.: Worldwide Health Innovations LLC; 2023. URL: https://patentscope2.wipo.int/search/en/detail.jsf?docId=US321199054&_cid=JP1-MHZW3X-13119-1 [accessed 2025-12-24]
33. Qadir S, Li D, Gu Y, Yuan Z, Zhao Y, Wang S. Experimental and numerical analysis on the enhanced separation performance of a medical oxygen concentrator through two-bed rapid pressure swing adsorption. Ind Eng Chem Res 2021;60(16):5903-5913. [doi: 10.1021/acs.iecr.1c00420.s001]
34. Benkirane L, Samid A, Chafik T. Small-scale medical oxygen production unit using PSA technology: modeling and sensitivity analysis. J Med Eng Technol 2023;47(6):321-335. [doi: 10.1080/03091902.2024.2331693] [Medline: 38626001]
35. Prayoga GA, Husni E, Damar Jaya S. Design of an embedded controller and optimal algorithm of PSA for a novel medical oxygen concentrator. Int J Electr Eng Inform 2023;15(2):220-239 [FREE Full text]
36. Kofod LM, Hansen EF, Brocki BC, Kristensen MT, Roberts NB, Westerdahl E. Optimised oxygenation improves functional capacity during daily activities in patients with COPD on long-term oxygen therapy: a randomised crossover trial. Thorax 2025;80(11):803-809 [FREE Full text] [doi: 10.1136/thorax-2024-221883] [Medline: 40473413]
37. Prayoga GA. Design and implementation system of mobile oxygen concentrator and telemedicine for comprehensive treatment of SpO2. Int J Adv Technol Eng Explor 2023;10(106). [doi: 10.19101/ijatee.2023.10101547]
38. Lara-Doña A, Sánchez-Morillo D, Pérez-Morales M, Fernández-Granero M, León-Jiménez A. A prototype of intelligent portable oxygen concentrator for patients with COPD under oxygen therapy. 2019 Presented at: MEDICON 2019, IFMBE Proceedings; September 26-28, 2019; Porto. [doi: 10.1007/978-3-030-31635-8_55]
39. Sanchez-Morillo D, Muñoz-Zara P, Lara-Doña A, Leon-Jimenez A. Automated home oxygen delivery for patients with COPD and respiratory failure: a new approach. Sensors (Basel) 2020;20(4):1178 [FREE Full text] [doi: 10.3390/s20041178] [Medline: 32093418]
40. Warren J. Efficient enriched oxygen airflow systems and methods. Wearair Ventures Inc.: WIPO; 2022. URL: https://patentscope2.wipo.int/search/en/detail.jsf?docId=US390988244 [accessed 2025-11-15]
41. James JRJ, Zuzana EM, inventor SHL. Pulsed oxygen system and process. Seabeck Holdings LLC.: JUSTIA Patents; 2022. URL: https://patents.justia.com/patent/20220347418 [accessed 2025-11-15]
42. Wang Q, Liu Y. Integrated oxygen supply device. Telesair Inc. 2022. URL: https://patents.google.com/patent/US11517702B1/en?oq=+US11517702+-+Integrated+oxygen+supply+device [accessed 2025-11-15]

XSL•FO
RenderX

43. Mapel DW, Robinson SB, Lydick E. A comparison of health-care costs in patients with chronic obstructive pulmonary disease using lightweight portable oxygen systems versus traditional compressed-oxygen systems. Respir Care 2008;53(9):1169-1175. [Medline: 18718034]
44. Casaburi R, Hess M, Porszasz J, Clark W, Diesem R, Tal-Singer R, et al. Evaluation of over-the-counter portable oxygen concentrators utilizing a metabolic simulator. Respir Care 2023;68(4):445-451 [FREE Full text] [doi: 10.4187/respcare.10495] [Medline: 36400446]
45. Law S. Liquid Oxygen Therapy at Home. 2005. URL: https://database.inahta.org/article/4059 [accessed 2025-03-08]
46. Dunne PJ. The clinical impact of new long-term oxygen therapy technology. Respiratory care 2009;54(8):1100-1111 [FREE Full text] [Medline: 19650950]
47. Dunne PJ. Long-term oxygen therapy (LTOT) revisited: in defense of non-delivery LTOT technology. Rev Port Pneumol 2012;18(4):155-157 [FREE Full text] [doi: 10.1016/j.rppneu.2012.02.011] [Medline: 22575635]
48. Andersson A, Ström K, Brodin H, Alton M, Boman G, Jakobsson P. Domiciliary liquid oxygen versus concentrator treatment in chronic hypoxaemia: a cost-utility analysis. Eur Respir J 1998;12(6):1284-1289 [FREE Full text] [doi: 10.1183/09031936.98.12061284] [Medline: 9877478]
49. Su C, Lee C, Chen H, Feng L, Lin H, Chiang L. Comparison of domiciliary oxygen using liquid oxygen and concentrator in northern Taiwan. J Formos Med Assoc 2014;113(1):23-32 [FREE Full text] [doi: 10.1016/j.jfma.2012.03.013] [Medline: 24445009]
50. Alptekin G. Low power medical oxygen concentrators for space missions. 2018 Presented at: 48th International Conference on Environmental Systems; July 8-12, 2018; Albuquerque.
51. Gilkey KM, Olson SL. Evaluation of the Oxygen Concentrator Prototypes: Pressure Swing Adsorption Prototype and Electrochemical Prototype. 2015. URL: https://ntrs.nasa.gov/api/citations/20150011038/downloads/20150011038.pdf [accessed 2025-12-24]
52. Qadir S, Li D, Gu Y, Yuan Z, Zhao Y, Wang S. Experimental and numerical analysis on the enhanced separation performance of a medical oxygen concentrator through two-bed rapid pressure swing adsorption. Ind Eng Chem Res 2021;60(16):5903-5913. [doi: 10.1021/acs.iecr.1c00420]
53. Pan M, Omar H, Rohani S. Application of nanosize zeolite molecular sieves for medical oxygen concentration. Nanomaterials (Basel) 2017;7(8):195 [FREE Full text] [doi: 10.3390/nano7080195] [Medline: 28757586]
54. Yadav VK, Choudhary N, Inwati GK, Rai A, Singh B, Solanki B. Recent trends in the nanozeolites-based oxygen concentrators and their application in respiratory disorders. Front Med (Lausanne) 2023;10:1147373 [FREE Full text] [doi: 10.3389/fmed.2023.1147373] [Medline: 37181347]
55. Mondal A, Dutta D, Chanda N, Mandal N, Mandal S. RESPIPulse: machine learning assisted sensory device for pulsed mode delivery of oxygen bolus using surface electromyography (sEMG) signals. Sens Actuators A Phys 2024;369:115121. [doi: 10.1016/j.sna.2024.115121]
56. Svensson R. The scientific output of a database on commercialized patents. IFN Working Paper No. 1349.: Research Institute of Industrial Economics; 2020. URL: https://www.ifn.se/media/unzdc0gw/wp1349.pdf [accessed 2025-01-09]
57. Chatburn RL, Williams TJ. Performance comparison of 4 portable oxygen concentrators. Respir Care 2010;55(4):433-442. [Medline: 20406511]

Abbreviations

ADL: activities of daily living
COPD: chronic obstructive pulmonary disease
IOSVR: Ultra Lightweight Cylinder Oxygen System
JBI: Joanna Briggs Institute
LiLSX: high lithium-exchanged X
LOX: liquid oxygen
MeSH: Medical Subject Headings
POC: portable oxygen concentrator
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Search extension
PRISMA-ScR: Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews
PSA: pressure swing adsorption
RPSA: rapid pressure swing adsorption
SpO_2: pulse oximeter oxygen saturation
VPC: vacuum pressure cycle
VPSA: vacuum pressure swing adsorption
WIPO: World Intellectual Property Organization

XSL•FO
RenderX

Edited by S Brini; submitted 22.Jul.2025; peer-reviewed by MW Hess, S Narayan; comments to author 24.Sep.2025; accepted 22.Dec.2025; published 27.Jan.2026.

Please cite as:
Kawachi S, Hoffman M, Romero L, Ekström M, Krishnan JA, Holland AE
Products, Performance, and Technological Development of Ambulatory Oxygen Therapy Devices: Scoping Review
J Med Internet Res 2026;28:e81077
URL: https://www.jmir.org/2026/1/e81077
doi:10.2196/81077
PMID:41429116

©Shohei Kawachi, Mariana Hoffman, Lorena Romero, Magnus Ekström, Jerry A Krishnan, Anne E Holland. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Development and Use of AI Chatbots for Health Behavior Change: Scoping Review

Lingyi Fu[1], MS; Ryan Burns[1], PhD; Yuhuan Xie[1], MS; Jincheng Shen[2], PhD; Shandian Zhe[3], PhD; Paul Estabrooks[1], PhD; Yang Bai[1], PhD

[1]Department of Health and Kinesiology, College of Health, University of Utah, Salt Lake City, UT, United States

[2]Department of Internal Medicine, School of Medicine, University of Utah, Salt Lake City, UT, United States

[3]Kahlert School of Computing, University of Utah, Salt Lake City, UT, United States

Corresponding Author:
Yang Bai, PhD
Department of Health and Kinesiology
College of Health
University of Utah
HPER-North
Room 204
Salt Lake City, UT, 84102
United States
Phone: 1 8015870482
Email: Yang.Bai@utah.edu

Abstract

Background: Artificial intelligence (AI) chatbots are technologies that facilitate human-computer interaction through communication in a natural language format. By increasing cost-effectiveness, interaction, autonomy, personalization, and support, mobile health interventions can benefit health behavior change and make it more natural and intuitive.

Objective: This study aimed to provide an up-to-date and practical overview of how text-based AI chatbots are designed, developed, and evaluated across 8 health behaviors, including their roles, theoretical foundations, health behavior change techniques, technology development workflow, and performance validation framework.

Methods: In accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) framework, relevant studies published before March 2024 were identified from 9 bibliographic databases (ie, PubMed, CINAHL, MEDLINE, Embase, Web of Science, Scopus, APA PsycINFO, IEEE Xplore, and ACM Digital Library). Two stages (ie, title and abstract screening followed by full-text screening) were conducted to screen the eligibility of the papers via Covidence software. Finally, we extracted the data via Microsoft Excel software and used a narrative approach, content analysis, and evidence map to synthesize the reported results.

Results: Our systematic search initially identified 10,508 publications, 43 of which met our inclusion criteria. AI chatbots primarily served 2 main roles: routine coach (27/43, 62.79%) and on-demand assistant (12/43, 27.91%), while 4 studies (4/43, 9.30%) integrated both roles. Frameworks like cognitive behavioral therapy (13/24, 54.17%) and behavior change techniques, such as goal setting, feedback and monitoring, and social support, guided the development of theory-driven AI chatbots. Noncode platforms (eg, Google Dialogflow and IBM Watson) integrated with social messaging platforms (eg, Facebook Messenger) were commonly used to develop AI chatbots (23/43, 53.49%). AI chatbots have been evaluated across 4 domains: technical performance (17/43, 39.53%), usability (17/43, 39.53%), engagement (37/43, 86.05%), and health behavior change (33/43, 76.74%). Evidence for health behavior changes remains exploratory but promising. Among 33 studies with 120 comparisons, 81.67% (98/120) showed positive outcomes, though only 35.83% (43/120) had moderate or larger effects (Hedges g or odds ratio or Cohen d>0.5). Most involved nonclinical (36/43, 83.72%) and adults (23/43, 53.49%), and a few were randomized controlled trials (14/43, 32.56%). Benefits were mainly seen in physical activity, smoking cessation, stress management, and diet, with limited evidence for other behaviors. Findings were inconsistent regarding the influence of long-term effects, intervention duration, modality, and engagement on health behavior change outcomes.

XSL•FO
RenderX

Conclusions: The exploratory synthesis provides a roadmap for developing and evaluating AI chatbots in health behavior change, highlighting the need for further research on cost, implementation outcomes, and underexplored behaviors such as sleep, weight management, sedentary behavior, and alcohol use.

(J Med Internet Res 2026;28:e79677) doi:10.2196/79677

KEYWORDS

conversational agent; machine learning; health behavior change; physical activity; diet; sleep

Introduction

Health-related behavior (hereafter referred to as "health behaviors") is essential for improving population health worldwide [1]. Engaging in health-promoting behaviors, such as having a healthy diet, getting adequate physical activity (PA) and sleep, and avoiding health-risk behaviors, such as smoking, can substantially reduce chronic disease and all-cause mortality risk [1,2] as well as benefit mental health [3]. Despite these benefits, unhealthy behaviors remain a significant public health concern and place a substantial burden on health care systems [4]. Health coaching is one suggested intervention for promoting health behavior changes [5,6]. Health coaching is defined as "the practice of health education and health promotion within a coaching context, to increase the well-being of individuals and to facilitate the achievement of their health-related goals" [7]. Most health coaching interventions are delivered by humans in various ways, such as face-to-face, telephonically, or via email [5]. Effective communication between coaches and users can support user-centered care and shared decision-making [8]. However, human health coaching interventions are limited in their ability to reach everyone in need of support because of a lack of coaching practitioners, resources in resource-limited communities, and barriers for individuals in accessing coaching support, such as low income [9]. From the implementer's perspective, human health coaching often lacks consistent data collection, continuous monitoring, scalability, and long-term sustainability [10]. Therefore, finding resource-efficient, cost-effective, and easily implementable strategies to promote health behavior change can be helpful to alleviate an already burdened health care system.

A chatbot is a computer program designed to respond to conversational or informational replies with verbal (audio-based chatbot) or written (text-based chatbot) messages from users [11]. This technology can be another type of resource for delivering health coaching, complemented by traditional human health coaching [9]. Chatbots can be developed with and without artificial intelligence (AI) algorithms. Most prior chatbots for health behavior change were developed without AI algorithms ("Non-AI chatbots") used in template-, rule-, or retrieval-based dialogue systems [12]. These chatbots responded to users by selecting from a predefined list, allowing for a high degree of researcher control but lacking the conversational flexibility and personalization typically offered by human coaches [9,13]. Recently, with advances in AI in the health care field, some chatbots have been developed using AI algorithms [14], such as reinforcement learning, deep neural networks, and random forest. These AI chatbots can communicate with users in natural language [15-20], offering personalized support, multimodal reasoning, and greater conversational flexibility and intuitiveness [21]. In particular, AI chatbots can support health behavior change in various ways. Research has shown that motivational interviewing (MI)–based AI chatbot tends to be perceived as more empathetic and trustworthy than the directed intervention, and it significantly raises the participants' self-efficacy to overcome barriers and positively impacts intrinsic motivation and PA levels [22]. Additionally, AI chatbots can also help alleviate stress by enhancing perceived supportiveness through the provision of emotional support [23]. Although the benefits are numerous, several concerns remain, including privacy and security [16]; limited empathy, affect, and emotional support [24]; low engagement; and challenges in monitoring intervention fidelity. Therefore, it is important to provide evidence to address these concerns and ensure the efficacy and safety of chatbot interventions.

Previous review studies have reported varying evidence on the application of AI chatbots for health behavior change. For example, a meta-analysis demonstrated strong efficacy of chatbot-based interventions in increasing physical activity, fruit and vegetable consumption, and sleep duration and quality. The analysis also showed that the effects varied by intervention duration, intervention modality (chatbot-only vs multicomponent interventions), and chatbot characteristics (text-based vs audio-based and AI-driven vs non–AI-driven chatbots) [12]. Other systematic reviews have summarized the outcome in addition to efficacy, including engagement, acceptability, satisfaction, and safety [25], as well as feasibility, usability, and intervention characteristics [10]. However, there remains a lack of research exploring key topics, such as the role of AI chatbots in behavior change, the health behavior change techniques (BCTs) adopted by AI chatbots, comprehensive technology frameworks for chatbot development, and frameworks for performance validation. Therefore, to address this gap and complement existing review studies [10,12,25], it is necessary to conduct a scoping review to provide a comprehensive overview of existing research for both scholars and practitioners in this field. This scoping review aimed to provide an up-to-date and practical examination of the design (ie, roles, theories, and health BCTs), development (technology workflow), and use (ie, performance validation tool) of text-based AI chatbots for 8 health behaviors, including PA, diet, sleep, weight management, sedentary, stress management, smoking cessation, and alcohol. In particular, 4 specific research questions were proposed based on the indications from the scoping reviews [26]:

- Question 1: What are the most commonly targeted health behaviors in text-based AI chatbots?

XSL•FO
RenderX

- Question 2: What roles, theoretical foundations, and BCTs are applied in text-based AI chatbots, supporting health behavior change interventions?
- Question 3: What technologies are used to develop text-based AI chatbots for health behavior change?
- Question 4: How to validate text-based AI chatbot performance in health behavior change?
  - Question 4.1: What measures are used to assess technical performance, usability, engagement, and cost?
  - Question 4.2: What are the health behavior change outcomes?

Methods

Protocol and Registration

The scoping review process was designed following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analysis extension for Scoping Reviews) framework [27]. The PRISMA-ScR checklist is reported in Multimedia Appendix 1. The study has been registered in the Open Science Framework [28], the most common platform for deposit protocols for scoping reviews.

Search Strategy

Search Resources

Relevant studies published before March 2024 were identified through 9 bibliographic databases, including 4 widely used health science databases (ie, PubMed, CINAHL, MEDLINE, and Embase), 2 multidisciplinary databases (ie, Web of Science and Scopus), 1 behavior and social science database (ie, APA PsycINFO), and 2 technology databases (ie, IEEE Xplore and ACM Digital Library). The last search date for each database is March 13, 2024.

Search Terms

Our search strategy incorporated terms related to both the intervention (AI chatbots) and the outcomes (health behaviors; Table S1 in Multimedia Appendix 2). Given the 6 pillars of lifestyle medicine [29], this study focused on 8 health behaviors, including PA, diet, sleep, weight management, sedentary behavior, stress management, smoking cessation, and alcohol. We identified synonyms for AI chatbots and various health behaviors by searching for relevant terms in dictionaries and references. We subsequently generated search syntaxes adapted to the specific requirements of each database. The example of search syntax for PubMed is as follows: "(((lifestyle*[tiab] OR tobac*[tiab] OR cigarette*[tiab] OR cigar* [tiab] OR vap*[tiab] OR smok*[tiab] OR nico*[tiab] OR sleep*[tiab] OR bedtime[tiab] OR nap[tiab] OR insomnia[tiab] OR physical activ*[tiab] OR sport*[tiab] OR exercise*[tiab] OR diet*[tiab] OR nutriti*[tiab] OR eating [tiab] OR food*[tiab] OR appetite*[tiab] OR *weight*[tiab] OR obes*[tiab] OR sedentar*[tiab] OR screen time [tiab] OR stress* [tiab]) AND (Chatbot*[tiab] OR chat-bot*[tiab] OR chat bot*[tiab] OR chat robot*[tiab] OR virtual robot*[tiab] OR voice-bot[tiab] OR social bot*[tiab] OR social robot*[tiab] OR infobot*[tiab] OR health bot*[tiab] OR smartbot*[tiab] OR conversational bot*[tiab] OR artificial intelligence chatbot*[tiab] OR Ai agent*[tiab] OR conversational agent*[tiab] OR dialogue agent*[tiab] OR dialog agent* [tiab] OR interactive agent*[tiab] OR virtual agent*[tiab] OR automated agent*[tiab] OR relational agent*[tiab] OR AI assist*[tiab] OR conversational assistant*[tiab] OR digital assist*[tiab] OR intelligent assist*[tiab] OR virtual assist*[tiab] OR smart assist*[tiab] OR voice assist*[tiab] OR speech assist*[tiab] OR virtual health assist*[tiab] OR dialogue agent*[tiab] OR dialog agent*[tiab] OR AI advisor*[tiab] OR virtual advisor*[tiab] OR animated advisor*[tiab] OR smart advisor*[tiab] OR AI avatar*[tiab] OR virtual avatar*[tiab] OR animated avatar*[tiab] OR smart avatar*[tiab] OR AI coach*[tiab] OR virtual coach*[tiab] OR smart coach*[tiab] OR animated coach*[tiab] OR artificial conversation entit*[tiab] OR Assistance technolog*[tiab] OR conversational AI[tiab] OR conversational interface*[tiab] OR conversational system*[tiab] OR Dialog system*[tiab] OR dialogue system*[tiab] OR natural language interface*[tiab] OR automated conversation[tiab] OR virtual conversation[tiab] OR chatGPT[tiab])) AND (eng[la])) NOT (Systematic review[pt] OR meta-analysis[pt] OR review[pt])." Finally, the metadata of the identified papers were imported into the Covidence platform to eliminate duplication and screening.

Study Eligibility Criteria

Table S2 in Multimedia Appendix 2 outlines the study eligibility criteria for the study and publication characteristics. The study characteristics were designed based on the PICOS framework, including population, intervention, comparison, outcome, and study type [30]. The publication characteristics included publication date, language, and publication status (eg, full online).

Study Selection

The selection process had 2 stages: title and abstract screening, followed by full-text screening. Both were conducted independently by 2 reviewers (LF and YX), with conflicts resolved by a third reviewer (YB). We used Cohen κ metric to evaluate interrater agreement [31]. The reviewers achieved substantial agreement in stage 1, with a κ measure of 0.83, which is greater than the cutoff point of 0.81 [31]. The screening process and calculation of Cohen κ value were performed via Covidence software.

Study Quality Assessment

Given the diversity of study designs, we used the Mixed Methods Appraisal Tool (MMAT) to assess methodological quality [32] (Table S3 in Multimedia Appendix 2 [16-20,23,24,33-68]). The MMAT is a 21-item checklist covering 5 research designs: qualitative, quantitative randomized controlled trials (RCTs), quantitative nonrandomized studies, quantitative descriptive studies, and mixed methods studies. Interrater reliability for the MMAT has been reported to range from moderate to perfect [69]. Two reviewers (LF and YX) independently evaluated each paper, and any disagreements were resolved through discussion. In general, all studies clearly stated their research questions, and the collected data were sufficient to address them.

Assigning an overall numerical score based on MMAT ratings is discouraged, as a single number cannot capture specific

XSL•FO
RenderX

methodological issues [32]. Therefore, we presented detailed ratings for each criterion. All eligible studies were included in this review, regardless of their MMAT ratings, as excluding studies solely on the basis of low methodological quality is not recommended [32,70].

Data Items

We designed the initial elements of the charting form based on the research questions and the mobile health evidence reporting and assessment checklist [71]. The items were primarily developed by 1 investigator (LF) and subsequently verified by another investigator (YB). The items were refined throughout the process, resulting in a final set of 22 elements (Table S4 in Multimedia Appendix 2).

Charting Process and Data Synthesis

Regarding data charting, 1 reviewer (LF) was primarily responsible for data extraction. This process involved 2 stages: the first stage focused on extraction, and the second stage on confirmation and supplementation. Two additional reviewers (Conrad Ma and Shreya Sanghvi) then independently validated the extracted data. Clear instructions for data validation were provided to these reviewers. Any discrepancies were resolved through discussion among the 3 reviewers. Microsoft Excel software was used for data charting processes. Finally, we used a narrative approach, content analysis, and evidence map to synthesize the reported results. In particular, we conducted a deductive coding process to map each chatbot function to the existing BCT Taxonomy (version 1) [72], enabling cross-study comparisons. For the AI chatbot validation framework in health behavior change, we applied a combined deductive-inductive approach to map the measures from each study onto 5 domains (ie, technical performance, health behavior change, usability, cost, and engagement), drawing from the digital health scorecard framework [73] and engagement framework [74].

Results

Study Selection

Figure 1 summarizes the process of selecting the studies via the Covidence software (Veritas Health Innovation). A total of 10,508 studies were returned after searching the databases. After duplication removal and title and abstract screening, 229 studies remained. In total, 40 studies remained after the full-text screening phase, with 189 studies removed for the following reasons: not involving an AI chatbot (n=98), not being a health behavior change intervention or implementation study (n=81), or lacking sufficient information (n=10). We included 3 additional studies after forward and backward reference checking, bringing the total to 43 included studies.

Figure 1. Study selection process. AI: artificial intelligence.

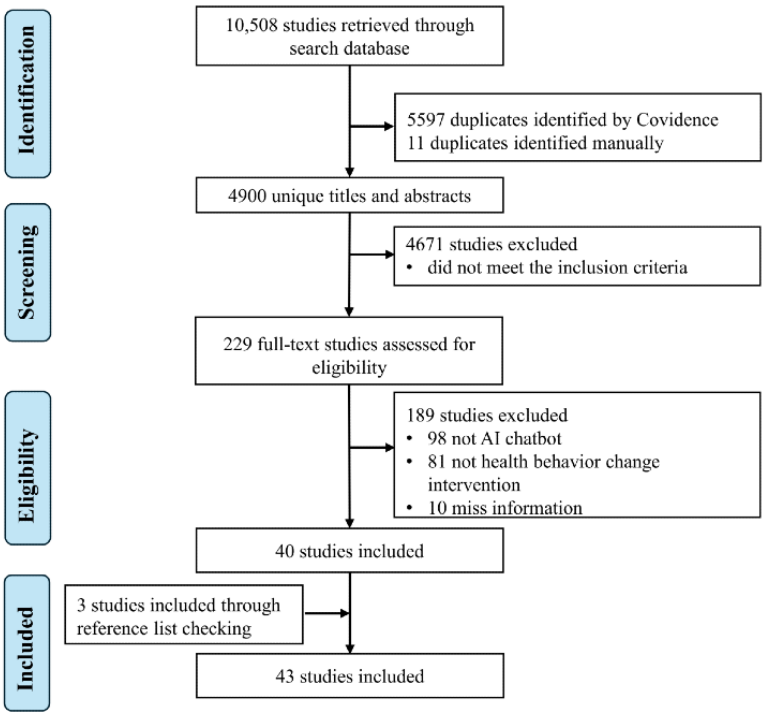

Study Overview

Overview

An overview of the studies included is presented in Table S5 in Multimedia Appendix 2 [16-20,23,24,33-68]. We further synthesized the findings across studies. Table S6 in Multimedia Appendix 2 [16-20,23,24,33-68] summarizes the publication characteristics of the studies included in the review. They were published between 2018 and 2024, with most published in 2023 (11/43, 25.58%). The studies were conducted across 15 countries or regions, with most of them conducted in Western countries (32/43, 74.42%), especially the United States (9/32, 28.13%), followed by the Netherlands (5/32, 15.63%), Italy (5/32, 15.63%), and the United Kingdom (5/32, 15.63%). Most research papers were published in journals (33/43, 76.74%), with the largest number published in *JMIR mHealth and uHealth* (6/33, 18.18%).

XSL•FO
RenderX

Methodological and Participant Characteristics

Table S7 in Multimedia Appendix 2 [16-20,23,24,33-68] summarizes the methodological and participant characteristics. Studies were classified based on the study types from Hong et al [32], including 14 RCTs, 15 quantitative nonrandomized trials (non-RCTs), 8 quantitative descriptive studies, 2 qualitative studies, and 4 mixed methods studies. The number of participants ranged from 7 [33] to 57,214 [34], with a primary focus on young and middle adulthood (aged 19-64 years; 23/43, 53.49%) based on age categories of Lindemann et al [75]. Most studies targeted nonclinical populations (36/43, 83.72%), such as individuals who are physically inactive, those with unhealthy diets, smokers, substance users, students (middle school, high school, and university), workers, and vulnerable groups, including low-income English- and Spanish-speaking individuals and residents of health professional shortage areas. In total, 7 studies targeted clinical populations (7/43, 16.28%), such as patients with colorectal cancer [35], patients with celiac [36], patients with cardiovascular problem [37], survivors of cancer [38], population with clinical eating disorder [39], as well as children [40] and youths [41] with obesity.

Intervention Characteristics

Table S7 in Multimedia Appendix 2 also provides a summary of the intervention characteristics. Regarding chatbot use, most studies used AI chatbots for 1-time use (10/43, 23.26%), followed by 8 weeks (7/43, 16.28%), 4 weeks (6/43, 13.95%), 12 weeks (5/43, 11.63%), 1 week (5/43, 11.63%), and 2 weeks (3/43, 6.98%). A small number of studies ranged from 3 days [42], 9 days [43], 3 weeks [44], 6 weeks [20], 24 weeks [19,45] to 48 weeks [46]. In terms of intervention modalities, most studies used a text-based chatbot alone (27/43, 62.79%; Table S5 [Study description] in Multimedia Appendix 2). Among them, some studies compared chatbots with other interventions, such as usual care by humans [19] and virtual humans or human teletherapists [24]. Design-focused studies explored personalization differences [47], reward structures [45], and emotional support and self-disclosure [23] to enhance chatbots' performance. In addition, a subset of studies (16/43, 37.21%) integrated AI chatbots with other components as part of multicomponent interventions, including digital tools, human interaction, and chatbot-implemented app functions (Figure 2 and Table S5 in Multimedia Appendix 2). Digital tools included exergames [37], the MedLiPal website [48], and wearable activity trackers (WATs). In particular, several studies (4/7, 57.14% [20,35,37,38]) technically integrated WATs with AI chatbots, enabling automatic data sharing. In others (3/7, 42.86% [46,48,49]), participants used WATs for self-monitoring to achieve goals set by the chatbots. AI chatbots were also integrated with human-delivered interventions, such as chemotherapy sessions [35], weight management programs in hospitals [41], physicians' clinic visits [45], the StudentBodies web-based program [39], family-based lifestyle modification programs [40], and remote traditional therapy [50,51]. Furthermore, several studies implemented AI chatbots within stand-alone applications that incorporated additional supportive features. These included a calendar and scoreboard [34], an online self-help library and e-book collection [33], an online smoking cessation diary [45], and self-care practice tools [18].

Figure 2. Architecture for the development of AI chatbots. This is a comprehensive picture derived from all the selected studies. Not every study reported the details of each module. AI: artificial intelligence; API: application programming interface.

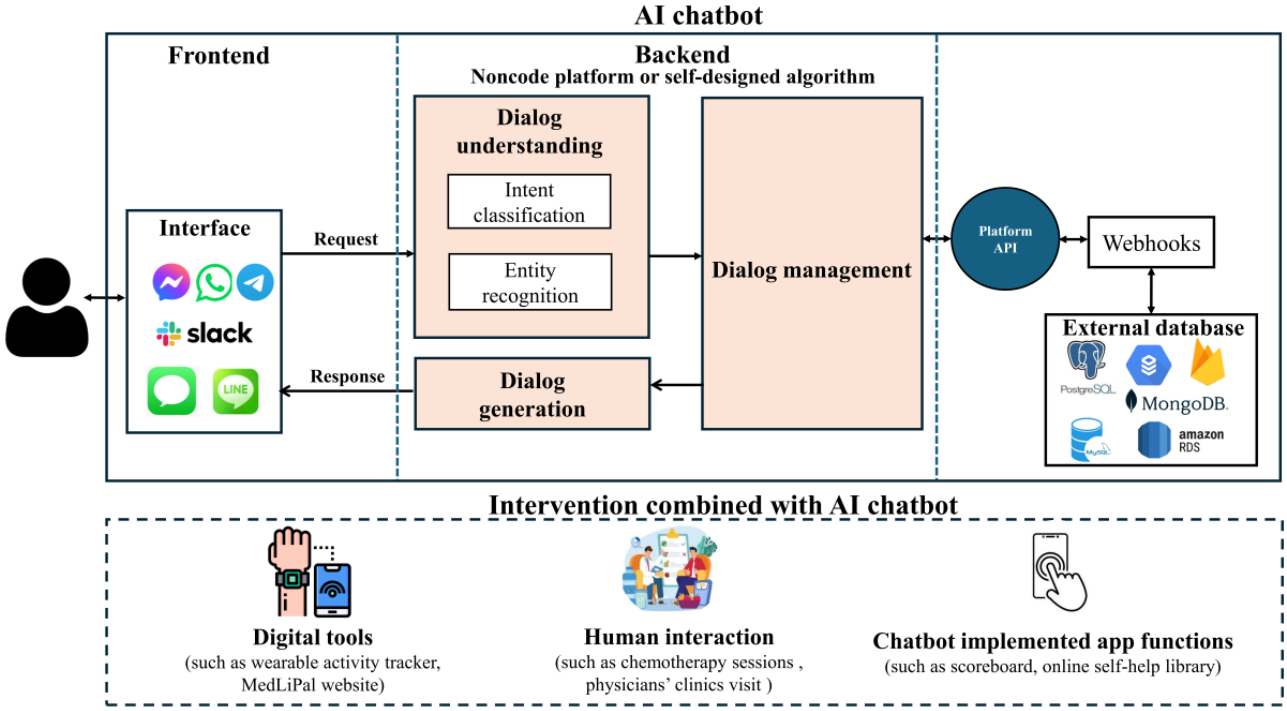

XSL•FO
RenderX

Question 1: What Are the Most Commonly Targeted Health Behaviors in Text-Based AI Chatbots?

Most studies focused on one health behavior (OHB; 29/43, 67.44%), and 14 studies addressed multiple health behaviors (MHBs; 14/43, 32.56%). PA was widely explored (18/43 with 8 OHB and 10 MHB), followed by stress management (16/43 with 10 OHB and 6 MHB), diet (11/43 with 1 OHB and 10 MHB), smoking cessation (7/43 with 6 OHB and 1 MHB), sleep (7/43 with 1 OHB and 6 MHB), weight management (6/43 with 2 OHB and 4 MHB), alcohol (1/43 OHB), and sedentary (1/43 MHB; Table 1).

Table 1. Targeted health behaviors (N=43).

Target behavior	References	Values, n (%)
Single behaviors (n=29, 67.44%)		
Stress management	[23,24,42,50-56]	10 (23.26)
Physical activity (PA)	[16,20,37,38,44,47,57,58]	8 (18.60)
Smoking cessation	[19,34,43,45,59,60]	6 (13.95)
Weight management	[17,41]	2 (4.65)
Diet	[36]	1 (2.33)
Sleep	[61]	1 (2.33)
Alcohol	[62]	1 (2.33)
Multiple behaviors (n=14, 32.56%)		
Diet and PA	[35,48,49,63]	4 (9.30)
Diet and weight management	[39,64]	2 (4.65)
Diet, PA, sleep, and stress management	[65,66]	2 (4.65)
Diet, PA, sleep, and weight management	[40]	1 (2.33)
Diet, PA, sleep, stress management, and sedentary	[67]	1 (2.33)
Sleep and stress management	[18,33]	2 (4.65)
PA and stress management	[68]	1 (2.33)
PA, smoking cessation, and weight management	[46]	1 (2.33)

Question 2: What Roles, Theoretical Foundations, and Behavior Change Techniques Are Applied in Text-Based AI Chatbots, Supporting Health Behavior Change Interventions?

Overview

We classified the AI chatbot for health behavior change into 2 roles and summarized theoretical foundations as well as corresponding functionalities (Table 2).

XSL•FO
RenderX

Table 2. Artificial intelligence chatbot roles, theoretical foundations, and health behavioral change techniques[a].

Paper	Theoretical foundation	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Routine coach and on-demand assistant (n=4)																	
Davis et al (2020) [49]	NR[b]	✓[c]	✓	✓	✓			✓									
Hassoon et al (2021) [38]	NR	✓	✓								✓						
Maher et al (2020) [48]	NR	✓	✓	✓	✓			✓									
Perski et al (2019) [34]	Mohr's model of supportive accountability					✓		✓									
Category sum	—[d]	3	3	2	2	1	0	3	0	0	1	0	0	0	0	0	0
Routine coach (n=27)																	
Albino de Queiroz et al (2023) [35]	CBT[e]		✓					✓									
Albers et al (2023) [47]	NR	✓				✓	✓										✓
Almusharraf et al (2020) [59]	MI[f]	✓			✓					✓							
Brown et al (2023) [60]	MI	✓		✓				✓		✓							
Cameron et al (2018) [33]	NR		✓	✓	✓												
Catellani et al (2023) [37]	Multiple theory 1[g]					✓											
Daley et al (2020) [52]	CBT and positive psychology		✓	✓	✓												
Dhinagaran et al (2021) [67]	COM-B[h]			✓	✓												
Figueroa et al (2021) [16]	Multiple theory 2[i]	✓															
Fitzsimmons - Craft et al (2022) [39]	CBT		✓	✓	✓				✓					✓			
Medeiros et al (2022) [42]	Emotion self-regulation			✓	✓											✓	
Legaspi et al (2022) [18]	NR	✓	✓		✓			✓						✓			
Karhiy et al (2023) [24]	NR				✓												
Meng and Dai (2021) [23]	NR		✓	✓													
Moore et al (2024) [57]	COM-B and the Theoretical Domains Framework	✓	✓		✓				✓							✓	
Piao et al (2020) [44]	HFM[j]	✓	✓					✓	✓		✓						
Piao et al (2020) [58]	HFM	✓	✓					✓	✓		✓						
Rahmanti et al (2022) [64]	COM-B	✓	✓	✓	✓	✓		✓			✓						
Sia et al (2021) [66]	NR	✓		✓	✓			✓			✓					✓	
Sun et al (2023) [68]	MI and graded exercise therapy	✓			✓												
Albers et al (2023) [43]	NR	✓	✓	✓	✓		✓	✓		✓	✓						
Aarts et al (2022) [61]	NR		✓														
Holmes et al (2019) [17]	NR		✓	✓													
Griol et al (2022) [63]	NR		✓	✓	✓	✓											
De Nieva et al (2020) [53]	CBT	✓	✓		✓				✓								
Carrasco-Hernandez et al (2020) [46]	CBT and MI		✓	✓		✓										✓	
Masaki et al (2019) [45]	NR	✓	✓	✓	✓												
Category sum	—	14	17	14	16	5	2	8	5	3	5	0	0	2	0	4	1
On-demand assistant (n=12)																	
Stephens et al (2019) [41]	Multiple theory 3[k]	✓	✓	✓													
Danieli et al (2021) [50]	CBT		✓	✓	✓												
Danieli et al (2022) [51]	CBT		✓	✓	✓												

XSL•FO
RenderX

Paper	Theoretical foundation	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Fadhil et al (2019) [65]	CBT		✓	✓	✓			✓									
Larizza et al (2023) [40]	NR	✓	✓	✓	✓												
Olano-Espinosa et al (2022) [19]	5A clinic practice guideline			✓	✓				✓		✓						
Alghamdi et al (2021) [36]	Multiple theory 4[l]		✓								✓	✓					
To et al (2021) [20]	COM-B	✓	✓	✓		✓		✓	✓		✓					✓	
Durden et al (2023) [54]	CBT	✓	✓		✓				✓								
Forman-Hoffman et al (2023) [55]	CBT	✓	✓	✓	✓				✓								
Hoffman et al (2023) [56]	CBT	✓	✓	✓	✓				✓								
Prochaska et al (2021) [62]	CBT	✓	✓	✓	✓				✓								
Category sum	—	7	11	10	9	1	0	2	6	0	3	1	0	0	0	1	0
Total number	—	24	31	26	27	7	2	13	11	3	9	1	0	2	0	5	1

[a]A: goals and planning; B: feedback and monitoring; C: social support; D: shaping knowledge; E: natural consequences; F: comparison of behavior; G: associations; H: repetition and substitution; I: comparison of outcomes; J: reward and threat; K: regulation; L: antecedents; M: identity; N: scheduled consequences; O: self-belief; P: covert learning.

[b]NR: not reported.

[c]✓: The study reported the results of this theme or subtheme.

[d]Not applicable.

[e]CBT: cognitive behavioral therapy.

[f]MI: motivational interviewing.

[g]Multiple theory 1: elaboration likelihood model, the self-regulatory model of message framing, the regulatory focus theory, and theories of emotions.

[h]COM-B: capability, opportunity, motivation—behavior.

[i]Multiple theory 2: MI, behavioral activation, acceptance and commitment therapy, and solution-focused brief therapy.

[j]HFM: habit formation model.

[k]Multiple theory 3: CBT, MI, and emotionally focused therapy.

[l]Multiple theory 4: chronic-disease extended model extending from the health belief model, the theory of planned behavior, diffusion of innovation theory, social norms theory, and the transtheoretical model.

Roles: Routine Coach and On-Demand Assistant

We classified the chatbot into 2 roles based on the intervention dosage (ie, use frequency and duration per interaction), specifically as a routine coach and an on-demand assistant. AI chatbots mostly played 1 role, such as routine coach (27/43, 62.79%) and on-demand assistant (12/43, 27.91%), whereas 4 studies integrated 2 roles (4/43, 9.30%; Table 2). Specifically, routine coaches delivered support in a defined use frequency and duration per interaction, such as 4 times per week, and each focused on 1 of 4 targeted health behaviors [67]. In contrast, on-demand assistants offer support with flexible frequency and intensity, allowing patients to contact the chatbot anytime and anywhere and determine the duration and frequency of interactions themselves [19].

Theoretical Foundation

Most studies incorporated a theoretical foundation to guide chatbot design strategies (28/43, 65.12%), with 20 studies applying a single theory and 8 studies using an integrated theoretical approach (Table 2 and Table S8 in Multimedia Appendix 2 [16-20,23,24,33-68]). Cognitive behavioral therapy (CBT; 13/28, 46.43%), MI (6/28, 21.53%), and the capability, opportunity, motivation—behavior framework (4/28, 14.29%) were primarily used, either individually or in combination with other theories.

CBT is a form of psychotherapy based on the concept that people's thinking influences their emotions and behaviors [76]. MI is a client-centered, directive therapeutic style to discover the client's own motivation for making changes by guiding clients to explore and resolve ambivalence [50,77]. Finally, capability, opportunity, motivation—behavior is a framework for understanding and changing behavior that posits capability, opportunity, and motivation.

Behavior Change Techniques

Because the functions of AI chatbots varied across studies, we mapped them onto the existing BCT taxonomy (version 1) [72] to enable cross-study comparisons (Table 2). The taxonomy included 16 themes and 93 techniques. We conducted a deductive coding process to code each chatbot function based on these 93 techniques and labeled a cluster if at least 1 technique within it was used. For example, if a chatbot provided automated, tailored feedback on reports and behavioral activity [35], we coded this as the technique "2.7 Feedback on outcome(s) of behavior," which falls under the "Feedback and Monitoring (B)" cluster. Accordingly, we indicated in the table that the study used at least 1 technique within that cluster. The coding was conducted by 1 primary extractor (LF) and validated by 2 additional reviewers (Conrad Ma and Shreya Sanghvi). Any discrepancies were resolved through discussion among all 3 reviewers. The interrater reliability was 81.70%.

XSL•FO
RenderX

Across all included studies, the most frequently applied clusters were goals and planning (A, 24/43, 55.81%), feedback and monitoring (B, 31/43, 72.09%), social support (C, 26/43, 60.47%), and shape knowledge (D, 27/43, 62.79%). No study used the techniques of antecedent (L) and scheduled consequences (N). This pattern shows that AI chatbots currently prioritize conversational, scalable, and digitally applicable techniques, while environmental restructuring and reinforcement-based strategies remain underused due to resource demands and limited environmental control.

BCTs by Roles

Goals and planning, social support, and shaping knowledge were commonly used in both routine coach and on-demand assistant (Table 2). The AI chatbot supported users in setting behavior goals and creating action plans [16,44,47,58], as well as reviewing behavior goals using historical data [20]. Additionally, AI chatbots provided social support (emotional or unspecified) in various ways, including personalized motivational dialogue [46], free-form responses [60], nurse contact information [33], and expressive elements such as sending emojis, icons, GIFs, gamification [52], and images [67]. Other strategies included providing a crisis hotline [36] and a 24-hour on-call [19]. Finally, shaping knowledge primarily involved providing clear instructions on how to perform the behavior [18,24,33,38,42,57,66,68]. For example, chatbots prompted users to share stressful experiences and then offered personalized suggestions, such as planning their day or reframing their mindset [18,42]. The main difference between routine coaches and on-demand assistants was the use of feedback and monitoring techniques (B), applied by 91.67% (11/12) of on-demand assistants and 62.96% (17/27) of routine coaches. This may be explained by the characteristics of on-demand assistants, which allow for continuous interaction, real-time behavior tracking, immediate feedback delivery, and progress reinforcement at any time. Routine coaches typically delivered feedback based on 1-time interactions, such as mood reflection [52], behavioral reporting [17,63], sleep diary [61], and gratitude journaling [53]. In contrast, on-demand assistants provided personalized feedback through active monitoring of health behavior level [20,40,65], mood status [41,55,56,62], and overall health behavior change progress [19,35,36].

Question 3: What Technologies Are Used to Develop Text-Based AI Chatbots for Health Behavior Change?

Workflow of AI Chatbots

Figure 2 summarizes the workflow of AI chatbots, including the frontend module, backend module, and external service module. The process generally involves the following steps: (1) the user sends messages through the frontend interface (eg, social messaging platform, web-based interface, or SMS); (2) the frontend interface forwards the message to the backend (eg, noncode platforms and self-designed algorithms) hosted on the research team's server; and (3) the backend system processes the messages, including dialogue understanding, management, and generation. The function of dialogue understanding is to extract meaning from the user input, such as intents and entities. Dialogue management involves domain-specific knowledge for tailoring replies. The text generation provides output to the user [78]; (4) the generated replies are sent and delivered back to the user through the frontend interface. If the selected intent requires additional operations, such as retrieving data from the external database, the platform sends a request to the webhook through its application programming interface (API). The webhook then processes the intents and returns a response to the platform, which subsequently delivers it to the user in an understandable format. Researchers can create custom webhooks to handle more complex intents.

Based on the development workflow, we classified AI chatbot technologies into 3 types: noncode platforms (23/43, 53.49%), self-designed algorithms (8/43, 18.60%), and stand-alone apps (12/43, 27.91%). These approaches can inform and guide researchers in developing AI chatbots for future use (Table S9 in Multimedia Appendix 2 [16-20,23,24,33-68]).

Noncode Platform

Noncode platforms offer built-in AI algorithms, allowing researchers to develop chatbots without writing code. Among the 23 studies using noncode platforms, the most commonly used platforms were Google Dialogflow (8/23, 34.78%) and IBM Watson (8/23, 34.78%), followed by X2AI (2/23, 8.70%) and Chatfuel (2/23, 8.70%). These AI chatbots were primarily accessed through social messaging platforms, such as Facebook Messenger [23,35,42,67], Slack [48,49], and SMS [16,39,41]. Several studies connected the chatbot to external systems to customize historical data through APIs. Examples include Google Cloud Functions and databases [35], PostgreSQL [40], and MongoDB [16,17,42,63].

Self-Designed Algorithms

Self-designed algorithms require researchers to develop their own AI models, allowing for greater customization of chatbot capabilities. Among 8 studies, some applied Bayes' theorem to assess individual needs and used natural language processing to enhance the chatbot's understanding and response capabilities [19]. Other approaches included combining GPT-2 XL with natural language processing for dialogue understanding and generation [60], deep reinforcement learning [37,43], supervised goal-based models [50,51], self-learning algorithms [46], and the Microsoft Bot Framework [33]. These custom backend systems were integrated with frontend interfaces through APIs, including Telegram [19], web-based interfaces, and SMS text messaging [38], as well as connected with external databases, such as MySQL database [33].

Stand-Alone Applications

Stand-alone apps encapsulate the entire workflow of AI chatbots, as illustrated in Figure 2. In total, 12 studies used stand-alone apps, such as Woebot [53-56,62], Wysa [18], Vitalk [52], Smoke-Free [34], and PsyMe [37]. Each app offered unique features tailored to specific functions. For example, Woebot emphasized helping users develop emotional regulation skills and support mood monitoring and management through conversations. Wysa was not only an AI chatbot but also a comprehensive mental health app, offering additional features such as access to a human talk therapist, stress management techniques, a journal for gratitude, and an international distress signal feature to seek help.

XSL•FO
RenderX

Question 4: How to Validate Text-Based AI Chatbot Performance in Health Behavior Change?

Performance Validation Framework

We mapped the measurements from all included studies onto the digital health scorecard framework [73] and engagement framework [74] to identify existing evidence and current gaps in validating the performance of AI chatbots in health behavior change. The digital health scorecard framework included 4 domains: technical, clinical, usability, and cost. Technical refers to testing if the AI chatbots actually perform to their self-proclaimed functionality with accuracy and precision. Clinical was operationalized as the critical appraisal of evidence demonstrating whether the AI chatbots impact the defined health behavior change outcomes. Usability refers to how easily an AI chatbot can be used for its intended purpose and the minimal effort required to complete tasks. Cost refers to the price for user access, technology lifecycle expenses, and integration costs within clinical workflows. However, these domains primarily assess how well a chatbot performs in influencing changes in health behaviors. Therefore, we incorporated engagement into the framework to capture user engagement between human-AI interaction, reflecting motivational and relational aspects that the other domains do not address. Engagement with digital behavior change interventions includes (1) the extent of use, such as amount, frequency, duration, and depth; and (2) a subjective experience characterized by attention, interest, and affect [74].

In terms of the coding process, first, we conducted deductive coding to map the measures from each study onto 5 domains: technical performance, health behavior change, usability, cost, and engagement. For example, if a study measured "technical feedback from users, ease of use, ease of learning, perceived usefulness, and satisfaction," we mapped "technical feedback from users" to the technical domain; "ease of use" and "ease of learning" to the usability domain; and "perceived usefulness" and "satisfaction" to the engagement domain, based on the definitions of each domain [65]. This process was conducted by 1 primary extractor (LF) and independently validated by 2 additional reviewers (Conrad Ma and Shreya Sanghvi). Any discrepancies were resolved through discussion among all 3 reviewers. The interrater reliability was 90.18%. In the next step, within each domain, we conducted inductive coding to group similar measurements and identify representative metrics, as shown in Figure 3.

Across 43 studies, 17 assessed technical performance, 33 evaluated health behavior change outcomes, 17 examined usability outcomes, and 37 measured engagement outcomes. None of the studies reported cost-related evidence (Figure 3).

Figure 3. Artificial intelligence chatbots on the health behavior change validation framework. MVPA: moderate-to-vigorous physical activity; PA: physical activity.

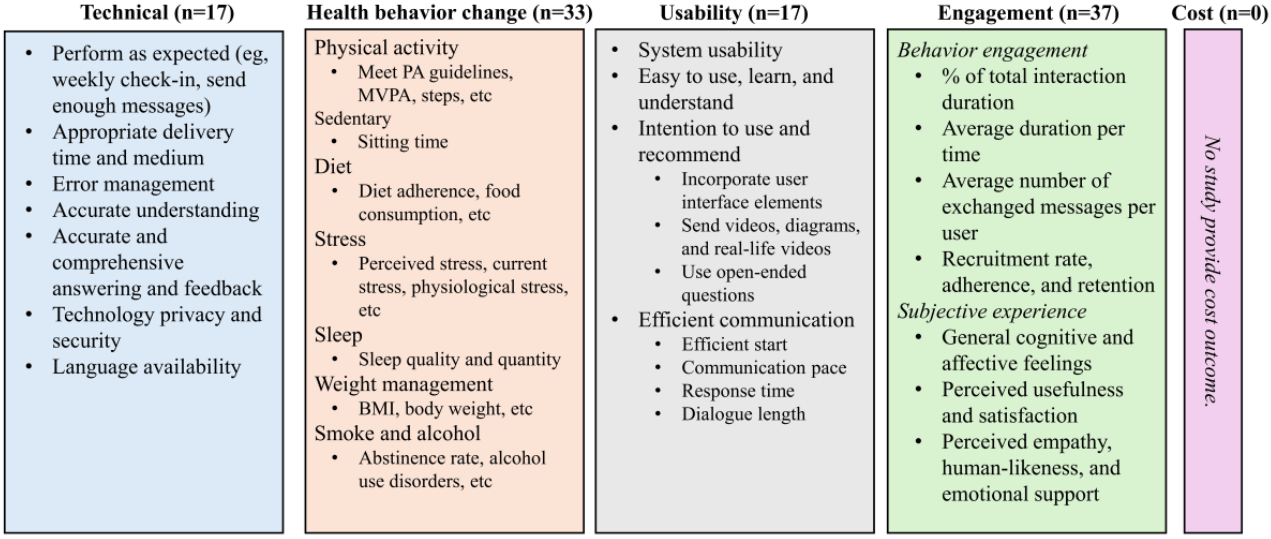

Question 4.1: What Measures Are Used to Assess Technical Performance, Usability, and Engagement?

Technical

In Figure 3, the technical performance of AI chatbots for health behavior change was evaluated across several aspects. The first metric was performance as intended, such as supporting weekly check-ins [49] and delivering a sufficient number of messages [20]. Delivery time and medium were evaluated to determine whether chatbots provided information promptly and through appropriate channels [67]. Error management focused on how effectively chatbots handled unexpected issues [17,20,33]. Several studies also assessed the chatbot's ability to accurately understand user input [33,46,49,53,63,65,66], as well as the provision of accurate and comprehensive information and feedback [40,63,68]. Language availability emerged as a key metric influencing chatbot performance, including the need for additional language options [63] and maintaining language consistency to generate appropriate, user-aligned responses [61,66]. Studies also highlighted that the use of local languages enhanced human connection and personalization [65], while simple and clear language improved user interaction and accessibility [67]. Finally, privacy and security concerns related to the technology were also important metrics to consider when adopting AI chatbots for user interventions [16].

XSL•FO
RenderX

Usability

In terms of usability, 4 studies measured general usability using the System Usability Scale, which includes items such as "I thought the system was easy to use." In total, 3 of those reported above-average industry scores (>68), including 88.2 [33], 84.8 [17], and 79.6 [35], while 1 reported a below-average score of 61.6 [20]. Beyond the System Usability Scale, some studies included items measuring ease of use, ease of learning, and ease of understanding [16,17,35,44,64,65,67,68], as well as intention to use [16,57,58,67,68] and recommendations to others [62]. Usability was also assessed through efficient communication, including smooth onboarding [17,33], suitable interaction pace and response time [17,33,57,64,67], and appropriate conversation length to maintain user engagement [61]. Researchers also identified several features that enhance usability, including human touch and user interface elements [65]; using multimedia such as videos, diagrams, and real-life examples; posing open-ended questions [57]; and allowing free-text input for communication [43].

Engagement

In Figure 3, behavioral engagement refers to engagement intensity and longevity. Most studies reported moderate interaction duration, including 50% [55], 62.5% [35], 66.78% [45] of the total intervention period. The average interaction duration per time was typically less than 30 minutes, such as 5.1 (SD 7.4) minutes [41], 12.5 (SD 15.62) minutes [59], and 21.3 (SD 14.0) minutes [20]. The average number of exchanged messages per user throughout the entire interventions varied between 245.1 [52] and 547.3 [59]. In addition, engagement at different stages was measured through recruitment (how many new individuals are added to a project within a specific time frame), adherence (the extent to which a person's behavior corresponds with agreed-upon recommendations from a care provider [79,80]), and retention (the extent to which the participants completed the study). The recruitment rate ranged from a high of 82% among inactive community-dwelling adults aged 45-75 years [48] to lower rates of 60% among young adults with eating disorders aged 18-30 years [39] and 55.1% among healthy adults aged 21 years and older [67]. Additionally, low adherence (<70% [48]) was reported in 2 studies, with participants completing an average of 63% (6.9/11) of weekly check-ins in one study [48] and 61% (6.7/11) in another [49]. A slight decline in weekly adherence was also observed, decreasing from 77% in week 1 to 69% in week 4 [62]. In contrast, higher retention rates were reported in the other 2 studies (>70% [48]), including 90% [48] and 93% [67].

In terms of subjective experience, several trials reported positive attitudes and acceptance [43,47,53,57], as well as feelings of low frustration, enjoyment [59], interesting [37], attractiveness, stimulation, novelty [17,57], and openness [18]. In addition, other metrics, such as feelings of helpfulness [39,41,51,59] and satisfaction [16,51,65,67], were also commonly reported. Furthermore, several studies highlighted negative perceptions of relational quality, such as complications [16], lack of empathy [18,60], limited human likeness [16,18,66], low affective support [53,57,66], robotic or unfriendly [17], lack of authenticity [47], and low motivational, as well as low emotional support [23].

Question 4.2: What Are the Health Behavior Change Outcomes?

Behavior Change Outcome Overview

Figure 3 illustrates the primary health behavior change outcomes, and Table S10 in Multimedia Appendix 2 [18-20,23,24,34-39,41-43,45-58,60,62,66-68] provides exploratory findings on the efficacy of AI chatbots for each outcome. In total, 33 of the included studies reported health behavior–related outcomes, yielding a total of 120 comparisons. To quantify the magnitude of change across interventions or pre- and postassessments, effect sizes were expressed as either Hedges *g*, odds ratios (ORs), or Cohen *d*. Hedges *g* was calculated when means and SDs were available, whereas ORs were used when only categorical data were reported. Cohen *d* from the original study was used when insufficient information was available to calculate Hedges *g*. Studies reporting Cohen *d* are indicated in Table S10 in Multimedia Appendix 2. According to Cohen conventions, a medium effect of 0.5 is visible to the naked eye of a careful observer [81].

Positive Changes

Among 33 studies with 120 comparisons, 81.67% (98/120) reported positive changes in promoting health behaviors. Positive changes refer to either statistically significant or nonsignificant improvements. However, only 35.83% (43/120) of these comparisons demonstrated observable positive changes with a moderate or larger effect size (Hedges *g* or OR or Cohen *d*>0.5). Moreover, it should be noted that most positive findings were observed in PA, smoking, stress management, and diet, indicating the need for more evidence in weight management, sleep, alcohol use, and sedentary behavior. Additionally, only a small portion of studies were RCTs (14/33, 42.42%), and the populations were primarily nonclinical adults (21/33, 63.64%).

Effectiveness in Real-World Settings

Only 4 of 33 (12.12%) studies evaluated the effectiveness of AI chatbots in real-world settings, all of which were non-RCTs focusing on stress reduction. Among them, the strongest clinically significant decrease in stress was *g*=–0.90 (95% CI –0.97 to –0.83) [52]. When comparing effectiveness by location, use patterns, and emotional status, studies found no significant differences between medically underserved areas and nonmedically underserved areas (t_{253}=0.30; *P*=.77; *d*=0.04, 95% CI –0.23 to 0.30) or between mental health provider shortage areas and nonmental health provider shortage areas (t_{253}=–1.39; *P*=.17; *d*=–0.18, 95% CI –0.44 to 0.07) [55]. Efficient users, those with lower behavioral engagement but stronger therapeutic alliance, achieved greater stress reductions (*g*=–0.60, 95% CI –0.86 to –0.33) than typical users (*g*=–0.25, 95% CI –0.47 to –0.03) and early users (*g*=–0.44, 95% CI –0.71 to –0.17) [56]. Participants with elevated mood symptoms at baseline experienced the greatest stress reduction (*g*=–0.68, 95% CI –0.93 to –0.44) compared with those with low mood symptoms (*g*=–0.28, 95% CI –0.53 to –0.02) [54].

Long-Term Efficacy

A total of 5 (5/33, 15.15%) studies evaluated follow-up efficacy after the intervention, including smoking cessation [45,60], stress [50,51], and diet-related outcomes [39]. Continuous

XSL•FO
RenderX

smoking-related improvements were observed. One study reported confidence (g=0.56, 95% CI 0.27-0.84), importance (g=0.24, 95% CI –0.03 to 0.52), and readiness to quit smoking (g=0.17, 95% CI –0.10 to 0.45) from baseline to 1-week follow-up compared to a single postsession measurement [60]. There was also strong and sustained smoking cessation across multiple follow-up points, with large effects at 12 weeks (g=1.40, 95% CI 1.03-1.78), 24 weeks (g=1.74, 95% CI 1.31-2.17), and 52 weeks (g=1.24, 95% CI 0.88-1.61) compared to 9 weeks after the intervention [45]. However, 2 RCTs did not find a consistent reduction in stress up to 12 weeks after an 8-week intervention in either the chatbot-only group [51] or the multicomponent intervention integrated with the chatbot group [50,51]. Similarly, for nonclinical eating disorder symptoms, the effect size between the intervention and control groups declined over time after the 4-week intervention (12-week: g=–0.41, 95% CI –0.63 to –0.20; 24-week: g=–0.20, 95% CI –0.41 to 0.01) [39].

Intervention Duration

Intervention duration appears to be an important factor influencing the efficacy of AI chatbots on health behavior change (4/33, 12.12%). A pre- and poststudy found that longer intervention duration (>6 weeks) yielded small but additional benefits across multiple behaviors among middle-aged and older adults, including weight loss (6 weeks: g=–0.06; 12 weeks: g=–0.07), waist circumference (6 weeks: g=–0.06; 12 weeks: g=–0.13), diet adherence (6 weeks: g=2.04; 12 weeks: g=2.06), and PA (6 weeks: g=0.32; 12 weeks: g=0.39) [48]. Similarly, another RCT reported that the percentage of participants increasing their metabolic equivalent of task scores rose from mid-intervention (g=–0.14, 95% CI –1.35 to 1.07) to the end of the 48-week intervention (g=0.06, 95% CI –0.87 to 0.98) [46]. Differences in smoking cessation outcomes related to intervention duration were also observed across 2 RCTs. A longer intervention duration of 48 weeks was associated with higher odds of biochemically validated abstinence in the chatbot group compared with the control group (OR 1.01, 95% CI 0.18-1.84) [46], whereas a shorter 24-week duration showed lower odds (OR 0.84, 95% CI 0.31-1.37) [19].

Intervention Modalities

In total, 9 of 33 (27.27%) studies have examined the impact of chatbot modalities on health behavior change outcomes, with all of them being RCTs. First, text-based chatbots performed worse than other modalities, such as video-based chatbots, virtual humans, and human coaches. A text-based AI chatbot showed a smaller increase in PA (g=0.35, 95% CI –0.39 to 1.10) compared with a video-based chatbot (g=1.14, 95% CI 0.34-1.94) after a 4-week intervention [38]. Similarly, the text-based chatbot demonstrated the smallest effect size in stress reduction (Cohen d=0.36) compared with the virtual human (Cohen d=0.52) and teletherapy (Cohen d=0.54) groups [24]. A text-based chatbot-only intervention (g=–0.34, 95% CI –1.33 to 0.65) also performed worse in reducing stress compared with traditional therapy (g=–0.71, 95% CI –1.44 to 0.03) [51]. Furthermore, participants who believed that they were interacting with a bot experienced a smaller reduction in stress than those who knew they were interacting with a human [42]. In addition, multicomponent interventions that combined text-based chatbots performed better than traditional therapy or other digital tools alone. For example, traditional therapy plus an AI chatbot led to greater stress reduction after an 8-week intervention compared with traditional therapy alone [50,51]. Similarly, combining psychopharmacological therapy with a digital therapeutic solution including an AI chatbot produced better stress reduction than psychopharmacological therapy alone (g=0.13, 95% CI –0.28 to 0.53) [46]. Finally, regarding chatbot design features, chatbots incorporating cues and intrinsic or extrinsic rewards significantly increased PA compared with a control chatbot without intrinsic rewards (t_{104}=2.12; P=.04) [58]. Personalized examples were linked to a significant increase in motivation (g=0.98, 95% CI 0.60-1.36) but a significant decrease in self-efficacy for PA engagement (g=–2.57, 95% CI –3.22 to –1.92) [47].

Engagement

Engagement, encompassing both behavioral engagement and subjective experience, emerged as a significant factor in promoting health behavior change (6/33, 18.18%). Most studies found that strong engagement was associated with positive outcomes, with the exception of one study [52]. High engagers (≥8 weekly check-ins) demonstrated greater increases in PA (high: g=0.65, 95% CI –0.14 to 1.44; low: g=0.51, 95% CI –0.15 to 1.18) but lower improvements in diet adherence (high: g=2.60, 95% CI 1.55-3.64; low: g=3.66, 95% CI 2.54-4.78) compared with low engagers [49]. Efficient engagers, those with lower behavioral engagement but stronger therapeutic alliance, had significantly greater reductions in stress than other user groups (g=–0.60, 95% CI –0.86 to –0.33) [56]. Intensive users (>4 contacts and >30 minutes of total interaction time) achieved higher quit rates than nonintensive users in both the chatbot intervention group (g=1.12, 95% CI 0.55-1.68) and usual care group (g=0.52, 95% CI 0.02-1.02) [19]. In addition, a causal mediation analysis explained that higher message involvement positively influenced PA intention through increased feelings of calmness (β=.07; P=.003) and greater hope (β=.44; P<.001) [37]. Finally, subjective feelings, such as the emotional support provided by AI chatbots, significantly reduced perceived stress through perceived supportiveness, underscoring the importance of subjective engagement experiences [23].

Sleep, Alcohol, Sedentary, and Weight Management

There were a small number of studies examining the efficacy of AI chatbots on sleep (n=2), weight management (n=4), alcohol use (n=1), and sedentary behavior (n=1). All of these were non-RCTs, except for the study by Carrasco-Hernandez et al [46]. First, there was no consistent evidence that AI chatbots effectively improved sleep quality or sleep quantity. One study found no significant effects on sleep quality (g=0.02, 95% CI –0.34 to 0.38) or sleep duration, with the proportion of short sleepers increasing by 6% after a 4-week intervention [67]. In contrast, another study reported a modest 3% improvement in sleep quality after a 1-week intervention [66]. Additionally, weight management appeared to be more challenging to change through chatbot interventions. A pre- and poststudy observed only small effects on weight loss (g=–0.07, 95% CI –0.57 to 0.43) and waist circumference reduction (g=–0.13, 95% CI –0.63 to 0.37) after 12 weeks among

XSL•FO
RenderX

middle-aged and older adults [48]. Similarly, other studies found no significant changes in BMI at 6-week postintervention (g=–0.01, 95% CI –0.27 to 0.24) [20], 24-week mid-intervention (g=–0.05, 95% CI –0.45 to 0.35) [46], and 48-week postintervention (g=0.13, 95% CI –0.28 to 0.53) [46]. Furthermore, the findings for alcohol use and sedentary behavior were relatively positive, showing a significant reduction in alcohol use disorder symptoms (g=–0.42, 95% CI –0.81 to –0.03) [62] and a 32 minutes per day decrease in sitting time [67].

Discussion

Principal Findings

The rapid advancement of AI and increased computational power have significantly expanded the potential applications and advantages of AI chatbots in facilitating health behavior change. This study aimed to provide an up-to-date overview of AI chatbot applications in this domain, along with practical guidance for their development and implementation. Consistent with prior research reviews [10,25,82], PA has emerged as a prominent focus area. This might be due to the need for a scalable intervention to solve the pandemic PA problems [83]. Health behavior change chatbots were classified as routine coaches (predefined frequency and intensity) and on-demand assistants (no specific frequency and intensity). Routine coaching offers a low-cost alternative that can supplement human therapists by providing guidance during their unavailable periods. On-demand assistants allow users to self-monitor and provide timely feedback. The 2 roles address key limitations of conventional interventions by providing more timely, low-cost, and personalized support while also reducing the resource burden on the traditional health care system. Considering theoretical foundations, most AI chatbots have been developed based on CBT and use BCTs such as goal setting and planning, feedback and monitoring, social support, and shaping knowledge. Notably, compared with routine couches, on-demand support chatbots rely more heavily on CBT as well as feedback and monitoring techniques. To achieve these functions, 3 main approaches have been used to develop AI chatbots: noncode platforms, self-designed algorithms, and stand-alone applications. Most studies used noncode platforms, such as Google Dialogflow and IBM Watson, which were then integrated into popular social messaging interfaces, including Facebook Messenger. These noncode platforms are particularly feasible for health behavior researchers who might lack programming expertise. Thus, it significantly improved accessibility and promoted wider adoption of chatbot interventions in health behavior change (across 262 health care centers [19] and up to 57,214 participants [34]).

We refined the validated digital framework [73] by adding engagement elements [74]. The updated framework, which includes technical, health behavior change, usability, engagement, and cost, captures all major measures of assessing the performance of AI chatbots in supporting health behavior change. The findings revealed a notable gap in cost-related evidence and highlighted the need for standardized approaches to calculate a global performance score. Such a standardized benchmark would help distinguish between high- and low-performing AI chatbots and enable cross-study comparisons. Moreover, the exploratory efficacy findings indicated that, although existing studies generally show positive effects of AI chatbots on health behavior change, evidence supporting clinically observable outcomes remains limited. Additionally, most studies have been conducted with nonclinical adult populations (aged 19-64 years), using nonrandomized or short-term trials (≤4 weeks), and have primarily focused on PA, stress management, smoking, and diet. Therefore, researchers should be cautious when applying these findings to clinical settings.

There is also a lack of evidence on the effectiveness of AI chatbots in real-world settings and their long-term efficacy in supporting health behavior change. With regard to intervention design and efficacy, a recent meta-analysis found no significant differences in chatbot effectiveness for increasing moderate-to-vigorous PA, daily steps, or fruit and vegetable consumption by intervention duration or intervention components [12]. In contrast, our exploration scoping review identified consistent findings that the longer intervention duration provides additional benefits across multiple behaviors, such as PA, diet, stress management, and weight management. Multicomponent interventions appeared more effective for stress management and food intake than chatbot-only interventions, though findings for PA were inconsistent. Regarding chatbot modalities, a previous meta-analysis reported that text-based chatbots were more efficacious than audio-based chatbots for fruit and vegetable consumption [12]. In contrast, our exploratory scoping review consistently found that text-based chatbots did not outperform other modalities, such as audio-based chatbots, human therapy, and virtual humans, in terms of PA and perceived stress management. This confirmed the statements that AI chatbots are not intended to replace health care professionals or provide treatment, but rather to complement existing care [52]. The inconsistent findings between this scoping review and the previous meta-analysis [12] underscore the need for additional systematic reviews and meta-analyses to provide more up-to-date and definitive conclusions. The exploratory findings also showed that higher engagement with AI chatbots was associated with greater improvements in health behavior outcomes, including increased PA, better diet adherence, lower perceived stress, and higher quit rates. These findings support previous research, indicating that engagement is a key factor in promoting health behavior change [83]. Finally, the minimal effects on weight management outcomes found in this scoping review were consistent with findings from the broader digital health intervention literature [84,85]. This likely reflects the physiological constraints of weight loss [46] and the fact that most chatbot interventions have targeted activity-related outcomes rather than weight outcomes. Despite these insightful findings, researchers should interpret this conclusion with caution, as it is exploratory and drawn from a broad scoping review rather than a rigorous systematic review and meta-analysis. Moreover, the evidence is based on small and fragmented samples across diverse health behaviors, which limits the strength of conclusions for any single behavior. Future systematic reviews and meta-analyses

XSL•FO
RenderX

covering a wider range of health behaviors are needed to provide stronger and more definitive evidence.

Implications

Practical Implications

AI chatbot shows benefits in promoting health behaviors among nonclinical adult populations, including PA, smoking, stress management, and diet. The chatbot can be strategically leveraged to facilitate health behavior change either as a stand-alone tool or by integrating it into existing programs, serving 2 primary roles: routine coaching and on-demand assistance. Establishing a clear distinction between these roles is critical for determining the appropriate frequency, intensity, and structure of user use. Moreover, researchers can design AI chatbot functionalities based on the synthesized evidence from health behavior change theories and BCTs identified in this scoping review. However, the most effective functionalities remain to be fully explored, and the underlying mechanisms are not yet well understood. Additionally, an accessible approach for health behavior scientists is to use no-code platforms (eg, IBM Watson and Google Dialogflow) or consumer-facing applications (eg, Woebot and Wysa) to develop and deploy AI chatbots for health behavior change interventions. Engagement is a critical factor that requires careful consideration, given the well-documented challenges of sustaining long-term engagement in AI chatbot interventions [34,39]. To address this issue, researchers should develop strategic approaches to maintain user engagement throughout the intervention period. Such strategies may include ensuring high response quality, optimizing interaction length [39], and incorporating visual elements, such as icons and graphs, to enhance user experience and promote sustained participation. It should also be noted that designing such chatbots requires careful consideration of participant characteristics (eg, age, gender, and clinical vs nonclinical populations) and contextual factors (eg, socioeconomic status, digital literacy, cultural norms, and technological environments) to ensure relevance and accessibility, thereby enhancing long-term engagement and achieving targeted outcomes. Finally, future research should incorporate a comprehensive set of evaluation measures encompassing 5 key domains, including technical performance, usability, health behavior changes outcomes, user engagement, and cost, to enable a more rigorous and holistic validation of AI chatbot efficacy.

Research Implications

This review summarized only the BCTs and theoretical foundations, underscoring the need for future research to identify the most influential BCTs and to examine how specific techniques (eg, rewards and graphical feedback) influence health behavior change outcomes within particular theoretical frameworks. In terms of technologies, most studies rely heavily on noncode platforms and conventional AI models. This approach might result in limited natural language communication capabilities and several well-documented issues, including insufficient human-like interaction, a lack of affect, empathy, and emotional support. To address these challenges, future research should consider integrating more advanced AI algorithms, such as generative models (eg, GPTs). Examining whether variations in these technologies influence the overall performance of AI chatbots is also important [86]. Additionally, most of the studies included in this review were conducted among Western, educated, industrialized, rich, and democratic populations [87] and nonclinical healthy adults. This limits the generalizability of the findings and practical guidance, as AI chatbot performance may be moderated by factors such as age differences [18,61], digital literacy, app familiarity, linguistic and cultural differences [16], as well as underserved settings [55]. Therefore, future research should focus on designing AI chatbots tailored to diverse demographic groups (eg, clinical populations, youths, and older adults) and contextual factors (eg, digital health equity) to achieve better outcomes across a broader range of populations.

Regarding AI chatbot validation outcomes, more evidence is needed on cost, weight management, sleep, sedentary behavior, and alcohol use. Additionally, more RCTs involving diverse populations, including younger and older adults, clinical populations, and individuals from varied social, economic, and cultural backgrounds, are needed to provide stronger and more comprehensive evidence. There is also a need to establish a gold standard to standardize scoring across different framework domains, including technical, usability, health behavior change, engagement, and cost. For example, a benchmark can be used to determine that when ≥75% of people think the chatbot is useful, it can be regarded as high accuracy (10/10). This enables the aggregation of individual domain assessments into a Global Digital Health Score, which can help validate the quality of AI chatbots and identify effective digital solutions. It can also highlight areas for improvement and inform stakeholders about potential gaps prior to product deployment. In addition to AI chatbot intervention outcomes, implementation outcomes such as reach, adoption, cost-effectiveness, fidelity, maintenance, scalability, and effectiveness also need to be explored. This would enhance the practical relevance of AI chatbots for digital health practitioners, supporting their implementation in real-world settings and improving scalability. Furthermore, more systematic reviews and meta-analyses need to explore the influence of intervention duration, multicomponent designs, and dose-response factors (eg, duration, frequency, and intensity) on AI chatbot performance, particularly given the variations across different health behaviors. Finally, the associations among different measures within the 5 clusters, including technical, usability, health behavior change, engagement, and cost, require further investigation. For example, usability, measured by willingness to continue, was associated with motivation to engage in activities and smoking quitter self-identity [43]. This can help optimize chatbots to better align with user needs, ultimately leading to improved health behavior change outcomes.

Strengths and Limitations

Strengths

This scoping review contributed to previous research in 5 key ways [10,12,25]. First, unlike prior reviews that focused on narrow behavioral domains, this study encompassed a comprehensive range of health behaviors, including PA, diet, sleep, weight management, sedentary behavior, stress

XSL•FO
RenderX

management, smoking, and alcohol consumption. We also used an extensive search strategy incorporating synonyms and part-of-speech variations. This approach yielded a substantially larger pool of eligible papers, providing a more holistic understanding of AI chatbots' role in health behavior change. Second, this review presents a detailed technology workflow for developing AI chatbots, which spans from frontend interfaces, backend architecture, and integration with external systems. By presenting this framework, we offered health behavior researchers, particularly those without computer science expertise, clear guidance on the technical foundations of chatbot implementation. Third, we classified AI chatbots on the predefined frequency and intensity, offering practical insights for researchers who sought to integrate this technology into health behavior change intervention studies. Fourth, we mapped chatbot functionalities onto the health BCT framework to help practitioners select appropriate BCTs for AI chatbots. Finally, we refined the digital validation tools by incorporating engagement measures, providing future intervention studies with clearer guidance on assessing chatbot performance comprehensively.

Limitations and Future Studies

Several limitations should be noted. First, we excluded studies that integrated audio-based chatbots, embodied conversational agents, humanoid coach virtual reality, augmented reality virtual coach, therapeutic robots, etc. This occurred because our focus was on the communication characteristics of AI chatbots in health behavior change rather than visual, action, or simulated environments. These additional characteristics add another layer of complexity to the deployment of AI chatbots. Future reviews could explore different AI chatbots that include these technologies. Second, we limited our study to publications in English, which might exclude relevant chatbots developed in other languages. Future reviews could consider including studies published in other languages. We also strongly encourage researchers conducting studies in non-English–speaking countries to publish their findings in English to enable cross-cultural comparisons. Finally, we included all types of studies to provide a more comprehensive synthesis, even though some were of relatively low quality. However, the heterogeneity in study designs and methodologies may limit the comparability of findings and the overall strength of the conclusions. We encourage future systematic reviews and meta-analyses to draw more robust insights by focusing on high-quality studies only.

Conclusions

This scoping review offers a comprehensive synthesis of AI chatbots as health behavior change interventions. The analysis revealed that PA was mostly targeted. When designing an AI chatbot, it is important to clearly define its roles (ie, routine coach or on-demand assistant or the combination) as well as to specify its theoretical foundation (eg, CBT), BCTs (eg, goals and planning), and technology workflow (eg, Google Dialogflow integrated with Facebook Messenger). The performance of AI chatbots can be evaluated across 5 clusters: technical, health behavior change, usability, engagement, and cost. Future studies should explore more on cost, sleep, weight management, sedentary behavior, and alcohol use to provide more comprehensive evidence. Additionally, they should also examine implementation outcomes to enhance the scalability of AI chatbot interventions. Moreover, rigorous RCTs in diverse populations are needed to generate robust and generalizable findings. Finally, the sustainability of AI chatbot effects on health behavior change along with factors such as intervention duration, modality, and engagement (eg, use duration, frequency, and intensity), as well as the interactions among the 5 evaluation clusters, warrant further exploration.

Acknowledgments

The authors especially thank Conrad Ma and Shreya Sanghvi for validating the data charting and synthesis for this scoping review. There is "nonhuman" assistance, such as machine learning, artificial intelligence, language models, and similar technology or tools, for the purposes of information gathering, analysis, content creation, manuscript writing, and editing.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Authors' Contributions

Conceptualization: LF (lead), YB (supporting)
Methodology: LF (lead), RB (supporting), YB (supporting)
Data curation: LF (lead), YX (supporting)
Formal analysis: LF (lead)
Writing—original draft: LF (lead)
Writing—review and editing: LF (lead), RB (supporting), YX (supporting), JS (supporting), SZ (supporting), PE (supporting), YB (supporting)

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
PRISMA-ScR checklist.
[DOCX File , 42 KB - jmir_v28i1e79677_app1.docx]

Multimedia Appendix 2
Detailed information on search terms, study eligibility criteria, quality assessment, data extraction items, overview of included studies, study description, characteristics of included studies, theoretical foundation, techniques for developing artificial intelligence chatbots, and efficacy and effectiveness of artificial intelligence chatbots on health behavior change.
[DOCX File , 954 KB - jmir_v28i1e79677_app2.docx]

References

1. Glanz K, Rimer BK, Viswanath K. Health Behavior: Theory, Research, and Practice. 5th Edition. San Francisco, CA: Jossey-Bass; 2015.
2. Rippe JM. Lifestyle medicine: the health promoting power of daily habits and practices. Am J Lifestyle Med 2018;12(6):499-512 [FREE Full text] [doi: 10.1177/1559827618785554] [Medline: 30783405]
3. Merlo G, Vela A. Mental health in lifestyle medicine: a call to action. Am J Lifestyle Med 2022;16(1):7-20 [FREE Full text] [doi: 10.1177/15598276211013313] [Medline: 35185421]
4. Cecchini M, Sassi F, Lauer JA, Lee YY, Guajardo-Barron V, Chisholm D. Tackling of unhealthy diets, physical inactivity, and obesity: health effects and cost-effectiveness. Lancet 2010;376(9754):1775-1784. [doi: 10.1016/S0140-6736(10)61514-0] [Medline: 21074255]
5. Olsen JM, Nesbitt BJ. Health coaching to improve healthy lifestyle behaviors: an integrative review. Am J Health Promot 2010;25(1):e1-e12. [doi: 10.4278/ajhp.090313-LIT-101] [Medline: 20809820]
6. Kivelä K, Elo S, Kyngäs H, Kääriäinen M. The effects of health coaching on adult patients with chronic diseases: a systematic review. Patient Educ Couns 2014;97(2):147-157. [doi: 10.1016/j.pec.2014.07.026] [Medline: 25127667]
7. Palmer S, Tubbs I, Whybrow A. Health coaching to facilitate the promotion of healthy behaviour and achievement of health-related goals. Int J Health Promot Educ 2003;41(3):91-93. [doi: 10.1080/14635240.2003.10806231]
8. Barry MJ, Edgman-Levitan S. Shared decision making—pinnacle of patient-centered care. N Engl J Med 2012;366(9):780-781. [doi: 10.1056/NEJMp1109283] [Medline: 22375967]
9. Mitchell EG, Maimone R, Cassells A, Tobin JN, Davidson P, Smaldone AM, et al. Automated vs. human health coaching: exploring participant and practitioner experiences. Proc ACM Hum Comput Interact 2021;5(CSCW1):1-37 [FREE Full text] [doi: 10.1145/3449173] [Medline: 36304916]
10. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: Systematic review. J Med Internet Res 2023;25:e40789 [FREE Full text] [doi: 10.2196/40789] [Medline: 36826990]
11. Chatbot. Dictionary.com. URL: https://www.dictionary.com/browse/chatbot [accessed 2024-06-25]
12. Singh B, Olds T, Brinsley J, Dumuid D, Virgara R, Matricciani L, et al. Systematic review and meta-analysis of the effectiveness of chatbots on lifestyle behaviours. NPJ Digit Med 2023;6(1):118 [FREE Full text] [doi: 10.1038/s41746-023-00856-1] [Medline: 37353578]
13. Rutjes H, Willemsen MC, Jsselsteijn IWA. Beyond behavior: the coach's perspective on technology in health coaching. 2019 Presented at: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; May 4-9, 2019; Glasgow, Scotland, United Kingdom p. 1-14. [doi: https://doi.org/10.1145/3290605.3300900]
14. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019;6(2):94-98 [FREE Full text] [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
15. Adamopoulou E, Moussiades L. An overview of chatbot technology. Artif Intell Appl Innov 2020;584:373-383. [doi: 10.1007/978-3-030-49186-4_31]
16. Figueroa CA, Luo TC, Jacobo A, Munoz A, Manuel M, Chan D, et al. Conversational physical activity coaches for Spanish and English speaking women: a user design study. Front Digit Health 2021;3:747153 [FREE Full text] [doi: 10.3389/fdgth.2021.747153] [Medline: 34713207]
17. Holmes S, Moorhead A, Bond R, Zheng H, Coates V, Mctear M. Usability testing of a healthcare chatbot: can we use conventional methods to assess conversational user interfaces? 2019 Presented at: Proceedings of the 31st European Conference on Cognitive Ergonomics; September 10-13, 2019; Belfast, United Kingdom p. 207-214. [doi: 10.1145/3335082.3335094]
18. Legaspi CM, Pacana TR, Loja K, Sing C, Ong E. User perception of Wysa as a mental well-being support tool during the COVID-19 pandemic. 2022 Presented at: Asian HCI Symposium '22. ACM; April 29-May 5, 2022; New Orleans, LA, United States p. 52-57. [doi: 10.1145/3516492.3559064]
19. Olano-Espinosa E, Avila-Tomas JF, Minue-Lorenzo C, Matilla-Pardo B, Serrano Serrano ME, Martinez-Suberviola FJ, et al. Effectiveness of a conversational chatbot (Dejal@bot) for the adult population to quit smoking: pragmatic, multicenter, controlled, randomized clinical trial in primary care. JMIR Mhealth Uhealth 2022;10(6):e34273 [FREE Full text] [doi: 10.2196/34273] [Medline: 35759328]

XSL•FO
RenderX

20. To QG, Green C, Vandelanotte C. Feasibility, usability, and effectiveness of a machine learning-based physical activity chatbot: quasi-experimental study. JMIR Mhealth Uhealth 2021;9(11):e28577 [FREE Full text] [doi: 10.2196/28577] [Medline: 34842552]
21. Jörke M, Sapkota S, Warkenthien L. Supporting physical activity behavior change with LLM-based conversational agents. ArXiv Preprint posted online on May 9, 2024. [doi: 10.48550/arXiv.2405.06061]
22. Chauvin R, Clavel C, Ravenet B, Sabouret N. A virtual coach with more or less empathy: impact on older adults' engagement to exercise. 2023 Presented at: 23rd International Conference on Intelligent Virtual Agents (ACM IVA 2023); September 19-22, 2023; Würzburg, Germany. [doi: 10.1145/3570945.3607338]
23. Meng J, Dai YN. Emotional support from AI chatbots: should a supportive partner self-disclose or not? J Comput-Mediat Commun 2021;26(4):207-222. [doi: 10.1093/jcmc/zmab005]
24. Karhiy M, Sagar M, Antoni M, Loveys K, Broadbent E. Mindfulness based stress reduction: a randomised trial of a virtual human, teletherapy, and a chatbot. 2023 Presented at: 2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW); September 10-13, 2023; Cambridge, MA, United States p. 1-7. [doi: 10.1109/aciiw59127.2023.10388195]
25. Oh YJ, Zhang J, Fang M, Fukuoka Y. A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. Int J Behav Nutr Phys Act 2021;18(1):160 [FREE Full text] [doi: 10.1186/s12966-021-01224-6] [Medline: 34895247]
26. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018;18(1):143 [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
27. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
28. Fu L, Xie Y, Bai Y. The development and use of artificial intelligence chatbots for health behavior change—a scoping review. Open Science Framework. 2024. URL: https://osf.io/wcspx/overview [accessed 2026-01-04]
29. Gray ID, Kross AR, Renfrew ME, Wood P. Precision medicine in lifestyle medicine: the way of the future? Am J Lifestyle Med 2019;14(2):169-186 [FREE Full text] [doi: 10.1177/1559827619834527] [Medline: 32231483]
30. Methley AM, Campbell S, Chew-Graham C, McNally R, Cheraghi-Sohi S. PICO, PICOS and SPIDER: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. BMC Health Serv Res 2014;14:579 [FREE Full text] [doi: 10.1186/s12913-014-0579-0] [Medline: 25413154]
31. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012:276-282. [doi: 10.11613/BM.2012.031]
32. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018;34(4):285-291. [doi: 10.3233/EFI-180221]
33. Cameron G, Cameron D, Megaw G. Assessing the usability of a chatbot for mental health care. In: Lecture Notes in Computer Science. 2018 Presented at: INSCI 2018 International Workshop; October 24-26, 2018; St Petersburg, Russia p. 121-132. [doi: 10.1007/978-3-030-17705-8_11]
34. Perski O, Crane D, Beard E, Brown J. Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. Digit Health 2019;5:2055207619880676 [FREE Full text] [doi: 10.1177/2055207619880676] [Medline: 31620306]
35. Albino de Queiroz D, Silva Passarello R, Veloso de Moura Fé V, Rossini A, Folchini da Silveira E, Aparecida Isquierdo Fonseca de Queiroz E, et al. A wearable chatbot-based model for monitoring colorectal cancer patients in the active phase of treatment. Healthc Anal 2023;4:100257. [doi: 10.1016/j.health.2023.100257]
36. Alghamdi E, Alnanih R. Chatbot design for a healthy life to celiac patients: a study according to a new behavior change model. Int J Adv Comput Sci Appl 2021;12(10):698-707. [doi: 10.14569/ijacsa.2021.0121077]
37. Catellani P, Biella M, Carfora V, Nardone A, Brischigiaro L, Manera MR, et al. A theory-based and data-driven approach to promoting physical activity through message-based interventions. Front Psychol 2023;14:1200304 [FREE Full text] [doi: 10.3389/fpsyg.2023.1200304] [Medline: 37575427]
38. Hassoon A, Baig Y, Naiman DQ, Celentano DD, Lansey D, Stearns V, et al. Randomized trial of two artificial intelligence coaching interventions to increase physical activity in cancer survivors. NPJ Digit Med 2021;4(1):168 [FREE Full text] [doi: 10.1038/s41746-021-00539-9] [Medline: 34887491]
39. Fitzsimmons-Craft EE, Chan WW, Smith AC, Firebaugh M, Fowler LA, Topooco N, et al. Effectiveness of a chatbot for eating disorders prevention: a randomized clinical trial. Int J Eat Disord 2022;55(3):343-353. [doi: 10.1002/eat.23662] [Medline: 35274362]
40. Larizza C, Bosoni P, Quaglini S, Chasseur M, Bevolo V, Zuccotti G, et al. V-care: an application to support lifestyle improvement in children with obesity. Int J Med Inform 2023;177:105140. [doi: 10.1016/j.ijmedinf.2023.105140] [Medline: 37463558]
41. Stephens TN, Joerin A, Rauws M, Werk LN. Feasibility of pediatric obesity and prediabetes treatment support through Tess, the AI behavioral coaching chatbot. Transl Behav Med 2019;9(3):440-447. [doi: 10.1093/tbm/ibz043] [Medline: 31094445]

XSL•FO
RenderX

42. Medeiros L, Bosse T, Gerritsen C. Can a chatbot comfort humans? Studying the impact of a supportive chatbot on users' self-perceived stress. IEEE Trans Human-Mach Syst 2022;52(3):343-353. [doi: 10.1109/THMS.2021.3113643]
43. Albers N, Neerincx MA, Aretz NL, Ali M, Ekinci A, Brinkman WP. Attitudes toward a virtual smoking cessation coach: relationship and willingness to continue. In: Meschtscherjakov A, Midden C, Ham J, editors. Persuasive Technology. PERSUASIVE 2023. Lecture Notes in Computer Science, vol 13832. Switzerland: Springer; 2023:265-274.
44. Piao M, Kim J, Ryu H, Lee H. Development and usability evaluation of a healthy lifestyle coaching chatbot using a habit formation model. Healthc Inform Res 2020;26(4):255-264 [FREE Full text] [doi: 10.4258/hir.2020.26.4.255] [Medline: 33190459]
45. Masaki K, Tateno H, Kameyama N, Morino E, Watanabe R, Sekine K, et al. Impact of a novel smartphone app (CureApp smoking cessation) on nicotine dependence: prospective single-arm interventional pilot study. JMIR Mhealth Uhealth 2019;7(2):e12694 [FREE Full text] [doi: 10.2196/12694] [Medline: 30777848]
46. Carrasco-Hernandez L, Jódar-Sánchez F, Núñez-Benjumea F, Moreno Conde J, Mesa González M, Civit-Balcells A, et al. A mobile health solution complementing psychopharmacology-supported smoking cessation: randomized controlled trial. JMIR Mhealth Uhealth 2020;8(4):e17530 [FREE Full text] [doi: 10.2196/17530] [Medline: 32338624]
47. Albers N, Hizli B, Scheltinga BL, Meijer E, Brinkman WP. Setting physical activity goals with a virtual coach: Vicarious experiences, personalization and acceptance. J Med Syst 2023;47(1):15 [FREE Full text] [doi: 10.1007/s10916-022-01899-9] [Medline: 36710276]
48. Maher CA, Davis CR, Curtis RG, Short CE, Murphy KJ. A physical activity and diet program delivered by artificially intelligent virtual health coach: proof-of-concept study. JMIR Mhealth Uhealth 2020;8(7):e17558 [FREE Full text] [doi: 10.2196/17558] [Medline: 32673246]
49. Davis CR, Murphy KJ, Curtis RG, Maher CA. A process evaluation examining the performance, adherence, and acceptability of a physical activity and diet artificial intelligence virtual health assistant. Int J Environ Res Public Health 2020;17(23):9137 [FREE Full text] [doi: 10.3390/ijerph17239137] [Medline: 33297456]
50. Danieli M, Ciulli T, Mousavi SM, Riccardi G. A conversational artificial intelligence agent for a mental health care app: evaluation study of its participatory design. JMIR Form Res 2021;5(12):e30053 [FREE Full text] [doi: 10.2196/30053] [Medline: 34855607]
51. Danieli M, Ciulli T, Mousavi SM, Silvestri G, Barbato S, Di Natale L, et al. Assessing the impact of conversational artificial intelligence in the treatment of stress and anxiety in aging adults: randomized controlled trial. JMIR Ment Health 2022;9(9):e38067 [FREE Full text] [doi: 10.2196/38067] [Medline: 36149730]
52. Daley K, Hungerbuehler I, Cavanagh K, Claro HG, Swinton PA, Kapps M. Preliminary evaluation of the engagement and effectiveness of a mental health chatbot. Front Digit Health 2020;2:576361 [FREE Full text] [doi: 10.3389/fdgth.2020.576361] [Medline: 34713049]
53. De Nieva JO, Joaquin JA, Tan CB, Marc Te RK, Ong E. Investigating students' use of a mental health chatbot to alleviate academic stress. 2021 Presented at: 6th International ACM In-Cooperation HCI and UX Conference; October 21-23, 2020; Jakarta & Bandung, Indonesia p. 1-10. [doi: 10.1145/3431656.3431657]
54. Durden E, Pirner MC, Rapoport SJ, Williams A, Robinson A, Forman-Hoffman VL. Changes in stress, burnout, and resilience associated with an 8-week intervention with relational agent "Woebot". Internet Interv 2023;33:100637 [FREE Full text] [doi: 10.1016/j.invent.2023.100637] [Medline: 37635948]
55. Forman-Hoffman VL, Pirner MC, Flom M, Kirvin-Quamme A, Durden E, Kissinger JA, et al. Engagement, satisfaction, and mental health outcomes across different residential subgroup users of a digital mental health relational agent: exploratory single-arm study. JMIR Form Res 2023;7:e46473 [FREE Full text] [doi: 10.2196/46473] [Medline: 37756047]
56. Hoffman V, Flom M, Mariano TY, Chiauzzi E, Williams A, Kirvin-Quamme A, et al. User engagement clusters of an 8-week digital mental health intervention guided by a relational agent (Woebot): exploratory study. J Med Internet Res 2023;25:e47198 [FREE Full text] [doi: 10.2196/47198] [Medline: 37831490]
57. Moore R, Al-Tamimi AK, Freeman E. Investigating the potential of a conversational agent (Phyllis) to support adolescent health and overcome barriers to physical activity: co-design study. JMIR Form Res 2024;8:e51571 [FREE Full text] [doi: 10.2196/51571] [Medline: 38294857]
58. Piao M, Ryu H, Lee H, Kim J. Use of the healthy lifestyle coaching chatbot app to promote stair-climbing habits among office workers: exploratory randomized controlled trial. JMIR Mhealth Uhealth 2020;8(5):e15085 [FREE Full text] [doi: 10.2196/15085] [Medline: 32427114]
59. Almusharraf F, Rose J, Selby P. Engaging unmotivated smokers to move toward quitting: design of motivational interviewing-based chatbot through iterative interactions. J Med Internet Res 2020;22(11):e20251 [FREE Full text] [doi: 10.2196/20251] [Medline: 33141095]
60. Brown A, Kumar AT, Melamed O, Ahmed I, Wang YH, Deza A, et al. A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking: iterative development study. JMIR Ment Health 2023;10:e49132 [FREE Full text] [doi: 10.2196/49132] [Medline: 37847539]
61. Aarts T, Markopoulos P, Giling L, Vacaretu T, Pillen S. Snoozy: a chatbot-based sleep diary for children aged eight to twelve. 2022 Presented at: IDC '22: Interaction Design and Children; June 27-30, 2022; Braga, Portugal p. 297-307. [doi: https://doi.org/10.1145/3501712.3529718]

XSL•FO
RenderX

62. Prochaska JJ, Vogel EA, Chieng A, Kendra M, Baiocchi M, Pajarito S, et al. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. J Med Internet Res 2021;23(3):e24850 [FREE Full text] [doi: 10.2196/24850] [Medline: 33755028]
63. Griol D, Callejas Z, Fernandez-Martinez F, Esposito A. An application of conversational systems to promote healthy lifestyle habits. 2022 Presented at: 2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech); September 12-15, 2022; Falerna, Italy p. 1-6. [doi: 10.1109/dasc/picom/cbdcom/cy55231.2022.9927835]
64. Rahmanti AR, Yang H, Bintoro BS, Nursetyo AA, Muhtar MS, Syed-Abdul S, et al. SlimMe, a chatbot with artificial empathy for personal weight management: system design and finding. Front Nutr 2022;9:870775 [FREE Full text] [doi: 10.3389/fnut.2022.870775] [Medline: 35811989]
65. Fadhil A, AbuRa'ed A. OlloBot—towards a text-based Arabic health conversational agent: evaluation and results. 2019 Presented at: Proceedings—Natural Language Processing in a Deep Learning World; September 2-4, 2019; Shoumen, Bulgaria p. 295-303. [doi: 10.26615/978-954-452-056-4_034]
66. Sia DE, Yu MJ, Daliva JL, Montenegro J, Ong E. Investigating the acceptability and perceived effectiveness of a chatbot in helping students assess their well-being. 2021 Presented at: Asian CHI Symposium 2021; May 8-13, 2021; Yokohama, Japan p. 34-40. [doi: 10.1145/3429360.3468177]
67. Dhinagaran DA, Sathish T, Soong A, Theng Y, Best J, Tudor Car L. Conversational agent for healthy lifestyle behavior change: web-based feasibility study. JMIR Form Res 2021;5(12):e27956 [FREE Full text] [doi: 10.2196/27956] [Medline: 34870611]
68. Sun X, Casula D, Navaratnam A. Virtual support for real-world movement: using chatbots to overcome barriers to physical activity. 2023 Presented at: 2nd International Conference on Hybrid Human-Artificial Intelligence; June 26-30, 2023; Munich, Germany p. 201-214. [doi: 10.3233/FAIA230084]
69. Pace R, Pluye P, Bartlett G, Macaulay AC, Salsberg J, Jagosh J, et al. Testing the reliability and efficiency of the pilot Mixed Methods Appraisal Tool (MMAT) for systematic mixed studies review. Int J Nurs Stud 2012;49(1):47-53. [doi: 10.1016/j.ijnurstu.2011.07.002] [Medline: 21835406]
70. Luo TC, Aguilera A, Lyles CR, Figueroa CA. Promoting physical activity through conversational agents: mixed methods systematic review. J Med Internet Res 2021;23(9):e25486. [doi: 10.2196/25486] [Medline: 34519653]
71. Lefevre AE, Agarwal S, Zeller K, Vasudevan L, Healey K, Tamrat T, et al. Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. World Health Organization. 2016. URL: https://www.who.int/publications/i/item/9789241511766 [accessed 2024-03-08]
72. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
73. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. NPJ Digit Med 2019;2(1):38 [FREE Full text] [doi: 10.1038/s41746-019-0111-3] [Medline: 31304384]
74. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
75. Lindemann EA, Chen ES, Wang Y, Skube SJ, Melton GB. Representation of social history factors across age groups: a topic analysis of free-text social documentation. AMIA Annu Symp Proc 2017:1169-1178 [FREE Full text] [Medline: 29854185]
76. Beck JS. Cognitive behavior therapy. 3rd Edition. In: Basics and Beyond. New York, NY: Guilford Publications; 2020.
77. Miller WR, Rollnick S. Motivational Interviewing: Helping People Change. New York, NY: Guilford Press; 2012.
78. Safi Z, Abd-Alrazaq A, Khalifa M, Househ M. Technical aspects of developing chatbots for medical applications: scoping review. J Med Internet Res 2020;22(12):e19127 [FREE Full text] [doi: 10.2196/19127] [Medline: 33337337]
79. Persch AC, Page SJ. Protocol development, treatment fidelity, adherence to treatment, and quality control. Am J Occup Ther 2013;67(2):146-153 [FREE Full text] [doi: 10.5014/ajot.2013.006213] [Medline: 23433268]
80. Salema NEM, Elliott RA, Glazebrook C. A systematic review of adherence-enhancing interventions in adolescents taking long-term medicines. J Adolesc Health 2011;49(5):455-466. [doi: 10.1016/j.jadohealth.2011.02.010] [Medline: 22018559]
81. Sullivan GM, Feinn R. Using effect size—or why the P value is not enough. J Grad Med Educ 2012;4(3):279-282 [FREE Full text] [doi: 10.4300/JGME-D-12-00156.1] [Medline: 23997866]
82. Chew HSJ. The use of artificial intelligence-based conversational agents (chatbots) for weight loss: scoping review and practical recommendations. JMIR Med Inform 2022;10(4):e32578 [FREE Full text] [doi: 10.2196/32578] [Medline: 35416791]
83. Kohl HW, Craig CL, Lambert EV, Inoue S, Alkandari JR, Leetongin G, et al. The pandemic of physical inactivity: global action for public health. Lancet 2012;380(9838):294-305 [FREE Full text] [doi: 10.1016/S0140-6736(12)60898-8] [Medline: 22818941]

XSL•FO
RenderX

84. Beleigoli AM, Andrade AQ, Cançado AG, Paulo MN, Diniz MDFH, Ribeiro AL. Web-based digital health interventions for weight loss and lifestyle habit changes in overweight and obese adults: systematic review and meta-analysis. J Med Internet Res 2019;21(1):e298 [FREE Full text] [doi: 10.2196/jmir.9609] [Medline: 30622090]
85. Cavero-Redondo I, Martinez-Vizcaino V, Fernandez-Rodriguez R, Saz-Lara A, Pascual-Morena C, Álvarez-Bueno C. Effect of behavioral weight management interventions using lifestyle mHealth self-monitoring on weight loss: a systematic review and meta-analysis. Nutrients 2020;12(7):1977 [FREE Full text] [doi: 10.3390/nu12071977] [Medline: 32635174]
86. Albers N, Neerincx MA, Brinkman WP. Addressing people's current and future states in a reinforcement learning algorithm for persuading to quit smoking and to be physically active. PLoS One 2022;17(12):e0277295 [FREE Full text] [doi: 10.1371/journal.pone.0277295] [Medline: 36454782]
87. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? Behav Brain Sci 2010;33(2-3):61-83 [FREE Full text] [doi: 10.1017/s0140525x0999152x]

Abbreviations

AI: artificial intelligence
API: application programming interface
BCT: behavior change technique
CBT: cognitive behavioral therapy
MHB: multiple health behavior
MI: motivational interviewing
MMAT: Mixed Methods Appraisal Tool
OHB: one health behavior
OR: odds ratio
PA: physical activity
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RCT: randomized controlled trial
WAT: wearable activity tracker

Edited by J Sarvestan; submitted 25.Jun.2025; peer-reviewed by N Acharya, S Hooker; comments to author 29.Jul.2025; revised version received 09.Dec.2025; accepted 10.Dec.2025; published 28.Jan.2026.

Please cite as:
Fu L, Burns R, Xie Y, Shen J, Zhe S, Estabrooks P, Bai Y
The Development and Use of AI Chatbots for Health Behavior Change: Scoping Review
J Med Internet Res 2026;28:e79677
URL: https://www.jmir.org/2026/1/e79677
doi:10.2196/79677
PMID:

©Lingyi Fu, Ryan Burns, Yuhuan Xie, Jincheng Shen, Shandian Zhe, Paul Estabrooks, Yang Bai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Efficacy of Brain-Computer Interface Therapy for Upper Limb Rehabilitation in Chronic Stroke: Systematic Review and Meta-Analysis of Randomized Controlled Trials

HongJie Chen[1,2], MSc; GuoJun Yun[1], MD

[1]Department of Rehabilitation Medicine, Shenzhen Children's Hospital, Shenzhen City, Guangdong Province, China

[2]College of Rehabilitation Medicine, Jiamusi University, Jiamusi, Heilongjiang, China

Corresponding Author:
GuoJun Yun, MD
Department of Rehabilitation Medicine
Shenzhen Children's Hospital
7019 Yitian Road, Futian District
Shenzhen City, Guangdong Province, 518026
China
Phone: 86 18938691526
Email: 103872187@qq.com

Abstract

Background: Over 50% of people with chronic stroke experience persistent upper limb dysfunction. Brain-computer interface (BCI) therapy, creating a sensorimotor loop via neural feedback, is a promising alternative; yet, its optimal application remains unclear.

Objective: This meta-analysis evaluates BCI's efficacy on motor function, tone, and activities of daily living (ADL) in chronic stroke and identifies optimal feedback modalities and intervention parameters.

Methods: We systematically searched Cochrane Library, Embase, PubMed, Scopus, Web of Science, and Wanfang Data from inception to October 2025 for randomized controlled trials (RCTs) comparing BCI-based training to control interventions in adults with chronic stroke. Primary outcomes were upper limb motor function (Fugl-Meyer Assessment for upper extremity [FMA-UE], Action Research Arm Test [ARAT]), muscle tone (Modified Ashworth Scale [MAS]), and ADL (Modified Barthel Index [MBI], Motor Activity Log [MAL]). Screening, data extraction, and risk-of-bias assessment were performed independently. Meta-analysis used a random-effects model with Hartung-Knapp-Sidik-Jonkman adjustment. Pooled mean differences (MDs) with 95% CIs and 95% prediction intervals (PIs) were calculated. Subgroup analyses examined feedback modalities, intervention intensity, and follow-up effects. Sensitivity analysis was also conducted.

Results: From 3529 records, 21 RCTs (650 participants) were included. BCI training significantly improved motor function (FMA-UE: MD 2.50, 95% CI 0.60-4.40; *P*=.01; 95% PI –2.52 to 7.22) and ADL performance (MBI: MD 8.38, 95% CI 2.23-14.53; *P*=.02; 95% PI –3.92 to 20.53; MAL: MD 2.09, 95% CI 0.42-3.76; *P*=.03; 95% PI –0.69 to 4.54). No significant effects were observed for fine motor skills (ARAT: MD 0.18, 95% CI –0.27 to 0.62; *P*=.30; 95% PI –3.64 to 3.99) or muscle tone (MAS: MD –0.48, 95% CI –1 to 0.03; *P*=.06; 95% PI –1.27 to 0.35). Subgroup analyses revealed that BCI-functional electrical stimulation (FES) yielded the greatest improvement in motor recovery (FMA-UE: MD 5, 95% CI 1.86-8.13; *P*=.01). The optimal intervention protocol was identified as 30-minute sessions, administered 4-5 times per week over 2 weeks (total of 10-12 sessions). However, benefits were not sustained at follow-up.

Conclusions: Low- to moderate-certainty evidence suggests that BCI training, particularly the BCI-FES paradigm, can improve upper limb motor function and ADL in people with chronic stroke on average. However, wide prediction intervals indicate the effect may vary substantially across settings, ranging from negligible to beneficial. Subgroup analyses suggested a potential optimal protocol of 30-minute sessions, 4-5 times per week for 2 weeks, but these findings are limited by the small number of studies in each subgroup and the high risk of bias in several included trials. Therefore, this proposed protocol should be viewed as preliminary and requires validation in future, high-quality RCTs. Future research should also focus on identifying patient subgroups most likely to benefit and on strategies to sustain long-term gains.

Trial Registration: PROSPERO CRD420251063808; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251063808

XSL•FO
RenderX

(J Med Internet Res 2026;28:e79132) doi:10.2196/79132

KEYWORDS

brain-computer interface; chronic disease; upper limb function; stroke; rehabilitation; meta-analysis

Introduction

Stroke, a neurological disorder resulting from cerebrovascular rupture or obstruction, leads to motor, speech, and cognitive impairments, consequently compromising performance in activities of daily living (ADL) [1]. Upper limb motor dysfunction represents one of the most prevalent sequelae of stroke [2,3], affecting a substantial proportion of survivors. Contemporary evidence indicates conventional rehabilitation strategies achieve optimal therapeutic gains predominantly within the first 6 months poststroke [4,5]. However, a substantial proportion of patients miss this critical window. For those in the chronic phase, current interventions demonstrate limited efficacy [6]. Notably, comparative studies reveal that stand-alone exoskeleton-assisted training or functional electrical stimulation (FES) provides comparable therapeutic benefits to conventional rehabilitation in chronic stroke cohorts [7]. Given the persistent, unmet rehabilitation needs in this population, brain-computer interface (BCI)–based training merits serious consideration. By enabling heightened patient engagement in volitional motor tasks [8], BCI represents a paradigm-shifting approach with significant potential to enhance recovery trajectories.

BCI technology establishes a direct communication pathway between the brain and an external device, bypassing damaged neural pathways. Fundamentally, BCI systems acquire and decode characteristic patterns of neural activity associated with user intent, such as motor imagery [9]. These decoded signals are then translated into commands to operate external feedback devices. BCI training represents an emerging neurorehabilitation technology based on this principle. This approach collects and decodes characteristic brain activity patterns, translating them into computerized commands to operate external feedback devices. These devices include FES [10], robotic exoskeletons [11], and visual feedback systems [12]. BCI systems are broadly categorized as invasive or noninvasive based on signal acquisition methodology. Due to safety concerns and practical limitations associated with invasive techniques [13], noninvasive BCIs are predominantly favored in current rehabilitation practice [14]. Primary noninvasive signal acquisition modalities encompass electroencephalography (EEG), magnetoencephalography, functional near-infrared spectroscopy (fNIRS), and functional magnetic resonance imaging. Among these, EEG stands as the predominant modality for signal acquisition in clinical rehabilitation settings [15]. The level of control exerted over external devices is contingent upon the specific neural signal source used. Integrating diverse neural signals within BCI frameworks enables more refined and efficient operation of feedback apparatus [16]. In contrast to conventional rehabilitation methods, the BCI paradigm establishes a “central-peripheral-central” closed-loop model. This approach holds promise for facilitating more timely movement adjustments and compensation strategies in people who have experienced a stroke [17], potentially offering greater alignment with personalized rehabilitation requirements.

Since the initial report on EEG-based BCI training in 2009 [18], numerous studies have demonstrated the efficacy of BCI interventions for improving upper limb motor outcomes in people who have experienced a stroke, including muscle strength, motor function, and ADL [12,19,20]. While some meta-analyses have addressed stroke stages in subgroup analyses, recent systematic reviews and meta-analyses primarily focus on comparing the magnitude of improvement across different stroke stages for a single primary outcome measure. For instance, subgroup analyses in studies by Xie et al [21] and Yang et al [22] solely compared BCI efficacy on upper limb function across stroke stages, without specifically evaluating the long-term therapeutic effects of BCI in people with chronic stroke. Furthermore, the influence of critical clinical parameters—such as intervention duration and session frequency—remains inadequately examined. A systematic analysis of BCI-based training protocols optimized for chronic stroke populations has yet to be conducted. Addressing these aspects is crucial for developing tailored rehabilitation protocols to enhance upper limb recovery, ADL performance, and overall quality of life in people with chronic stroke.

Therefore, this meta-analysis was conducted with three specific aims:

1. To systematically evaluate the efficacy of BCI-based training on upper limb motor function, muscle tone, and ADL exclusively in people with chronic stroke;
2. To perform in-depth subgroup analyses of key moderating factors—including BCI feedback modalities, intervention intensity parameters (session duration, frequency, and total sessions), and follow-up effects—which have been inadequately addressed in prior syntheses;
3. To attempt to propose optimized BCI intervention protocols tailored to the chronic stroke population based on current evidence.

By addressing these gaps, this review seeks to provide clearer guidance for clinical practice and future research in BCI-based stroke rehabilitation.

Methods

Protocol and Registration

This systematic review was registered with PROSPERO (International Prospective Register of Systematic Reviews), the international systematic review database, bearing identifier CRD420251063808. This meta-analysis followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines published in 2020 [23] (Multimedia Appendix 1).

XSL•FO
RenderX

Search Strategy

The systematic literature search was designed, conducted, and reported in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses – Search Extension) guideline [24]. An experienced information specialist (HJC) developed the search strategy in collaboration with the review team.

We executed a comprehensive search across 6 electronic databases from their inception until October 16, 2025, including PubMed (via the National Library of Medicine), Embase (via Ovid), Cochrane Central Register of Controlled Trials (via Wiley), Scopus (via Elsevier), Web of Science Core Collection (via Clarivate Analytics), and Wanfang Data. These platforms were selected to ensure extensive coverage of both international and Chinese literature. All databases were searched individually; no multidatabase searching was performed on a single platform.

The search strategy used a combination of controlled vocabulary (eg, MeSH [Medical Subject Headings] in PubMed and Emtree in Embase) and keywords related to the core concepts of "brain-computer interfaces," "stroke," "upper extremity," and "rehabilitation." The complete search strategies for all databases are provided in Multimedia Appendix 2. To maximize sensitivity, no restrictions were placed on language, publication date, or study design during the search. Similarly, no published search filters were used, and the strategy was developed de novo for this review, not adapted from prior work.

To enhance the robustness of the search, the PubMed strategy underwent peer review by an information specialist prior to execution, following the Peer Review of Electronic Search Strategies (PRESS) guideline framework [25]. Additionally, in attempts to obtain missing or incomplete data, we contacted the corresponding authors of studies via email.

Beyond the methods above, we did not systematically search study registries, websites, or gray literature, nor did we use citation searching.

The total number of records retrieved from each database is documented in the PRISMA flow diagram. All identified records were imported into EndNote X9 (Clarivate) for management, where duplicates were removed using the software's automated deduplication feature, followed by a manual verification process conducted independently by 2 reviewers (HJC and GJY).

Inclusion criteria comprised (1) population: adults (>18 years) diagnosed with chronic stroke (>6 months poststroke), exhibiting stable vital signs and alert consciousness; (2) intervention: receiving any form of BCI-based training; (3) control: receiving either sham BCI interventions or conventional rehabilitation therapy (eg, physical therapy, occupational therapy, and treadmill training); (4) outcomes: assessment using the Fugl-Meyer Assessment for upper extremity (FMA-UE), Action Research Arm Test (ARAT), Modified Ashworth Scale (MAS), Modified Barthel Index (MBI), and Motor Activity Log (MAL); (5) study design: randomized controlled trials (RCTs) published in English or Chinese.

Exclusion criteria were (1) nonprimary research publications (eg, reviews, meta-analyses, systematic reviews, and conference abstracts); (2) studies reporting outcome measures inconsistent with those prespecified for analysis; and (3) publications with incomplete or irretrievable data.

Selection Process

All records identified through the database searching were imported into EndNote X9 (Clarivate Analytics) for management. The total number of records retrieved from each database and information source was documented. Duplicates were removed using a 2-step process: first, automatically using EndNote's built-in deduplication feature, followed by a manual verification conducted independently by 2 investigators (HJC and GJY). The screening process was then carried out in 2 stages; first, titles and abstracts were screened against the inclusion criteria; second, the full texts of potentially eligible records were retrieved and assessed. Both investigators independently evaluated the studies at each stage. Any disagreements regarding study selection were resolved through discussion.

Data Extraction

Two investigators (HJC and GJY) independently reviewed the full texts of included studies. Data extraction was performed independently by both reviewers, capturing key details, including first author's name, the age of the participants, time after stroke, the number of participants by different hemiplegic sides and stroke types, publication year, BCI signal acquisition method, feedback device, sample size, intervention details, intervention duration, outcome measures, and follow-up period. For studies reporting outcomes solely as median and IQR, these values were converted to estimated mean and SD using the methods recommended by the Cochrane Handbook [26]: mean ≈ median; SD ≈ IQR / 1.35 [9]. Any discrepancies between reviewers were initially resolved through discussion. Persistent disagreements were resolved through discussion.

Quality Assessment

According to the PRISMA guidelines [23], the risk of bias assessment was conducted for the included RCTs. Two investigators (HJC and GJY) independently assessed the methodological quality and risk of bias for included studies using the Cochrane Risk of Bias tool (RoB 2.0) [27]. The RoB 2.0 evaluates five domains: (1) bias arising from the randomization process; (2) bias due to deviations from intended interventions; (3) bias due to missing outcome data; (4) bias in outcome measurement; and (5) bias in selection of the reported result. Studies were rated as superior if all domains were judged at low risk, indicating minimal bias concerns. Studies with some domains at low risk but others raising concerns were rated good, reflecting a moderate risk of bias. Studies where no domain achieved low risk were rated poor, indicating substantial bias concerns. Disagreements between assessors were resolved through discussion.

Additionally, the overall quality of evidence for each primary outcome was assessed using the GRADE (Grading of Recommendations Assessment, Development and Evaluation) framework [28]. Two reviewers (HJC and GJY) independently evaluated the evidence quality for the following outcomes. The GRADE approach considers 5 domains, including risk of bias,

XSL•FO
RenderX

inconsistency, indirectness, imprecision, and publication bias. Evidence quality was categorized as high, moderate, low, or very low. Any discrepancies in ratings were resolved through discussion.

Statistical Analyses

Statistical analysis was performed using Stata 18 (StataCorp LLC) software. Data were pooled using mean differences (MDs) with 95% CIs to assess BCI training efficacy. The extent of heterogeneity was quantified using the tau-square (τ^2) statistic, which estimates the variance of true effect sizes across studies. The I^2 statistic is reported as a supplementary measure, representing the percentage of total variability in effect estimates that is due to heterogeneity rather than sampling error [29,30]. The Cochran Q test (chi-square test) was used to test the null hypothesis of homogeneity; a significant *P* value ($P<.05$) was taken as evidence of the presence of heterogeneity [31]. Given the anticipated clinical and methodological diversity among the included studies, all meta-analyses were performed using the random-effects model. To ensure more robust and conservative estimates, especially given the varying number of studies included in different analyses, we applied the Hartung-Knapp-Sidik-Jonkman adjustment for calculating the 95% CIs around the pooled MDs [32]. This model provides a more conservative estimate of the effect size and its CI by accounting for both within-study and between-study variability. To quantify the implications of heterogeneity and estimate the range within which the true effect size is likely to fall in future similar scenarios, we calculated 95% prediction intervals (PIs) for the meta-analyses of all outcome measures included in this systematic review so as to ensure the reliability of the estimation of between-study variance [33].

The outcome measures of our meta-analysis included upper limb motor function, muscle tone, and ADL. The primary outcome measure was the FMA-UE, which is a commonly used indicator for clinical assessment of upper limb dysfunction in people who have experienced a stroke [34]. Specifically, the ARAT served as the assessment scale for upper limb function, the MAS was used for muscle tone evaluation, and both the MBI and MAL were used to assess ADL in people who have experienced a stroke.

Subgroup analyses were performed to assess the influence of the following potential factors on upper limb functional outcomes:

1. Feedback modality: BCI-FES, BCI-robot, or BCI-visual feedback;
2. Intervention intensity: stratified by session duration (20 min vs 30 min vs 60 min), weekly frequency (2-3 days vs 4-5 days), total intervention duration (2 weeks vs 4-5 weeks vs 8 weeks), and total number of sessions (10-12 times vs 20-24 times);
3. Follow-up time point: short-term (≤3 months) vs long-term (>3 months).

The thresholds for these subgroup classifications were defined based on the most frequently reported intervention parameters across the included studies and common dosing regimens in prior clinical research [35,36]. This categorization allows for a direct comparison of the different training intensities most commonly encountered in the current evidence base.

Sensitivity analysis was performed to investigate potential sources of heterogeneity. Assessment of small-study effects was performed using funnel plots and the Egger test for outcomes that included a sufficient number of studies (typically n>10). A *P* value <.05 in the Egger test was considered indicative of potential small-study effects [37].

Results

Search Results

The systematic literature search yielded a total of 3529 records from the 6 databases Cochrane Library (n=1173), Embase (n=622), PubMed (n=378), Scopus (n=398), Web of Science (n=546), and Wanfang (n=412). Following the deduplication process as prespecified in the methods, 1534 duplicate publications were removed. The remaining 1995 unique records underwent initial title and abstract screening. Subsequently, 152 articles were selected for full-text assessment, from which 21 studies met the predefined inclusion criteria and were included in the final meta-analysis. The study selection process is detailed in the PRISMA flow diagram (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the study selection process.

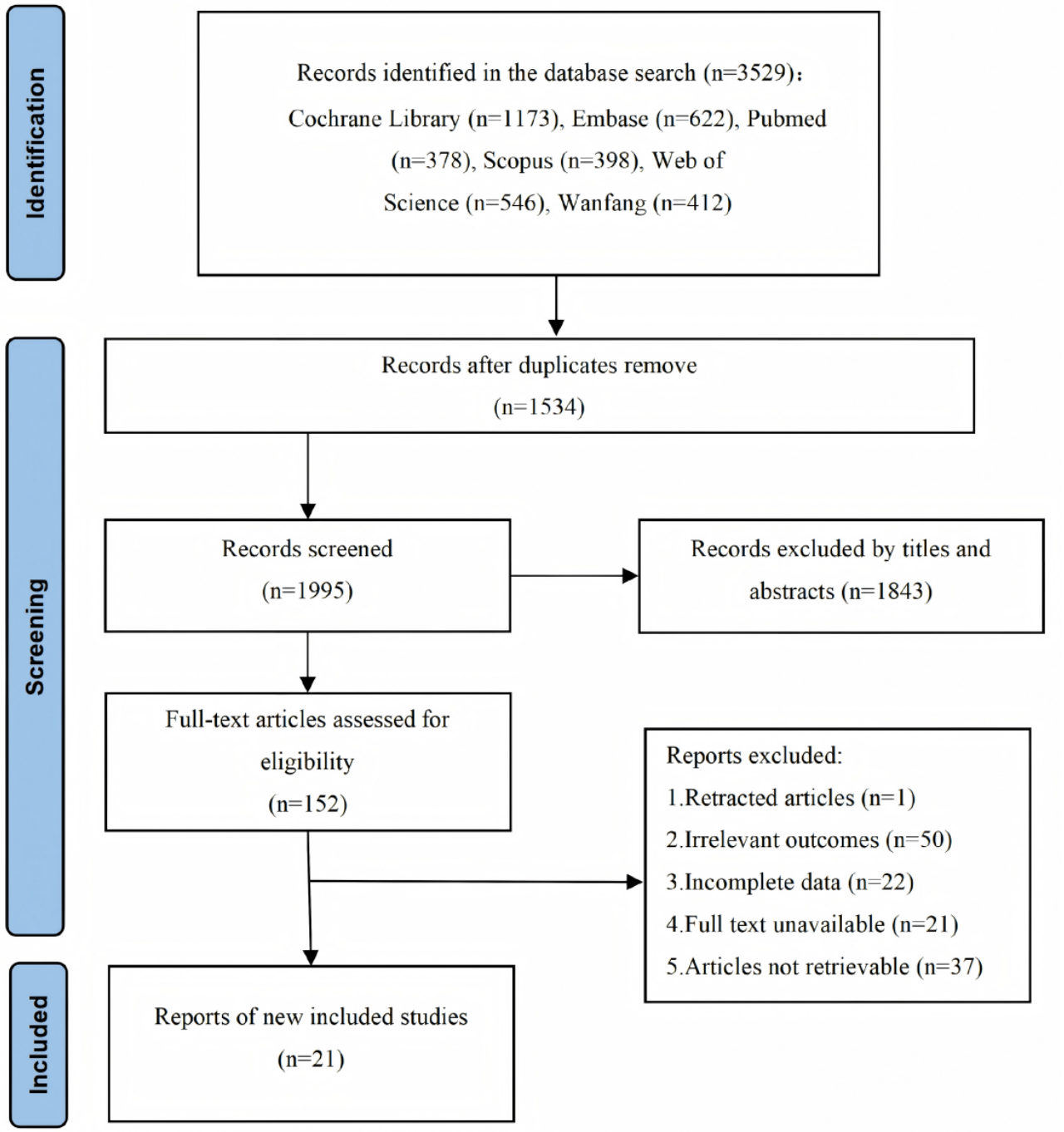

Quality Evaluation

The methodological quality of the 21 included RCTs was assessed using the revised RoB 2. Based on this assessment, 6 RCTs were classified as “Superior,” 6 as “Good,” and 9 as “Poor.” In several studies, the substantial differences in interventions between the experimental and control groups made blinding of participants and intervention providers not feasible. Consequently, these studies were judged to carry a high risk of bias. The specific evaluation results are shown in Figure 2 [11,19,35,38-55].

XSL•FO
RenderX

Figure 2. Risk of bias assessment for included randomized controlled trials (RoB 2.0) [11,19,35,38-55].

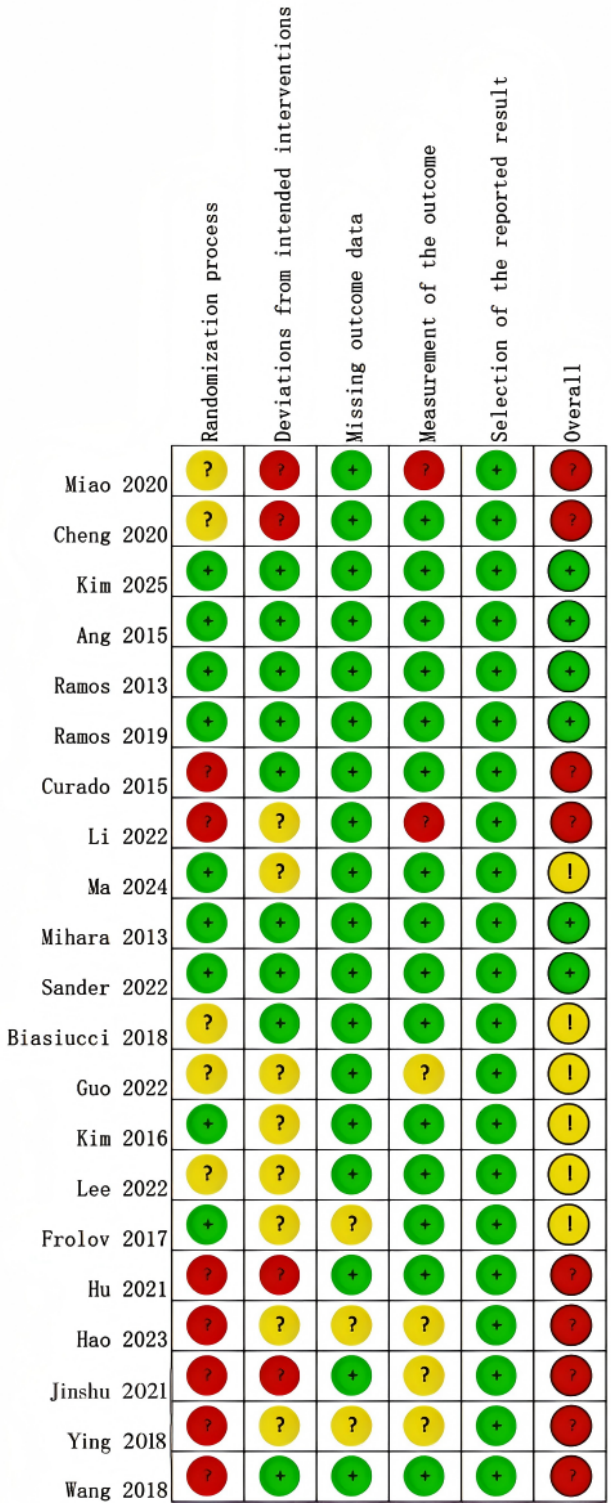

The overall quality of evidence for the primary outcomes, as assessed by the GRADE approach, ranged from low to moderate. The summary of findings and the detailed GRADE evidence profile for each outcome are available in Multimedia Appendix 3.

Characteristics of the Included Literature

A total of 21 studies [11,19,35,38-55], published between 2013 and 2025, were included in this meta-analysis. These studies comprised 337 participants in experimental groups and 313 in control groups. Individual intervention sessions ranged from 20 minutes to 120 minutes, while the total intervention duration varied from 3 days to 10 weeks, encompassing 6-70 sessions in total. Follow-up assessments were performed in 8 studies [11,19,39,42,45-47,55]. Regarding the BCI feedback modality, FES was used in 7 studies [19,38,40,48,49,52,54], exoskeleton devices in 10 studies [11,39,41-44,47,50,53,55], and visual feedback in 4 studies [35,45,46,51]. Specific methods for

XSL•FO
RenderX

random sequence generation were described in 15 studies [11,19,35,40-42,44-46,48-50,52-54]. Allocation concealment was implemented in 8 studies [11,19,40-42,45,46,48], and outcome assessors were blinded in 15 studies [11,19,35,39-43,45,46,48-51,55]. The detailed characteristics of the included studies are presented in Multimedia Appendix 4.

Results of Meta-Analysis

Effect of BCI on Overall Motor Function (FMA-UE)

As illustrated in Figure 3, sixteen [11,19,35,38,40,41,43-49,51,54,55] studies reported the effects of BCI training on FMA-UE scores in people with chronic stroke. The meta-analysis demonstrated a statistically significant improvement in FMA-UE scores following BCI training (MD 2.50, 95% CI 0.60-4.40; *P*=.01). The test for heterogeneity was not statistically significant (Q=18.72; *P*=.23), and the I^2 was 45.18%. The estimated between-study variance was tau²=6.06. The 95% PI was (–2.52 to 7.22), indicating that the effect of BCI on FMA-UE in a future similar study could range from a clinically irrelevant decline of 2.52 points to a clinically meaningful improvement of 7.22 points.

Figure 3. Forest plot for the effect of brain-computer interface (BCI) training on Fugl-Meyer Assessment for upper extremity (FMA-UE) scores [11,19,35,38,40,41,43-49,51,54,55].

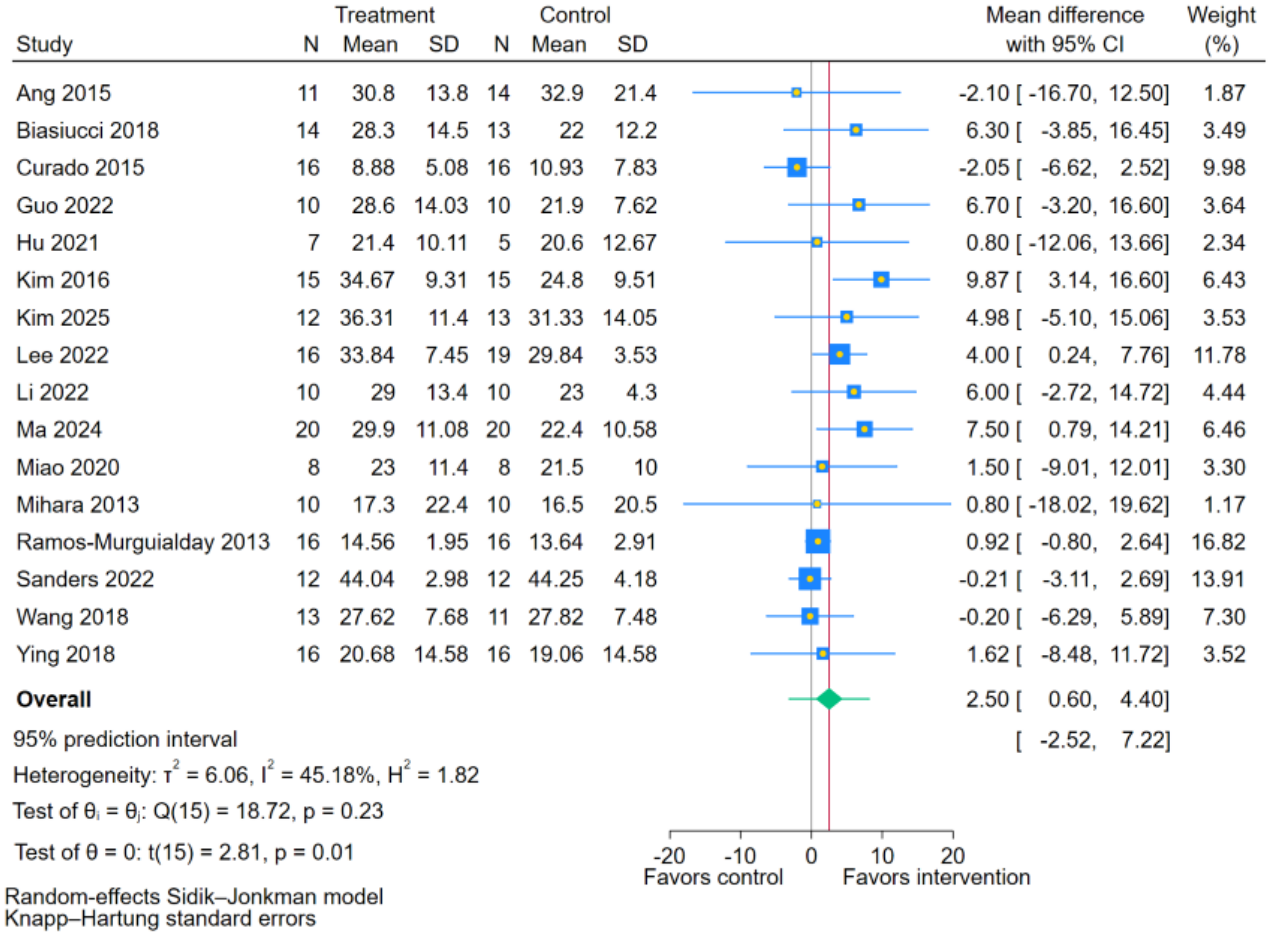

Effect of BCI on Fine Motor Skills (ARAT)

As depicted in Figure 4, four [39,45,50,51] studies evaluated the impact of BCI training on ARAT scores in people with chronic stroke. The meta-analysis revealed no statistically significant difference in ARAT scores between the intervention and control groups (MD=0.18, 95% CI –0.27 to 0.62; *P*=.30). The test for homogeneity was not statistically significant (Q=0.04; *P*=.99). The estimated between-study variance was tau²=0. The I^2 statistic was 0.03%. Furthermore, the 95% PI was (–3.64 to 3.99). This wide interval, which spans both clinically negligible negative and positive effects, underscores the considerable uncertainty regarding the true effect of BCI on fine motor skills and indicates that in future settings, the outcome could range from a slight worsening to a modest improvement.

XSL•FO
RenderX

Figure 4. Forest plot for the effect of brain-computer interface (BCI) training on Action Research Arm Test (ARAT) scores [39,45,50,51].

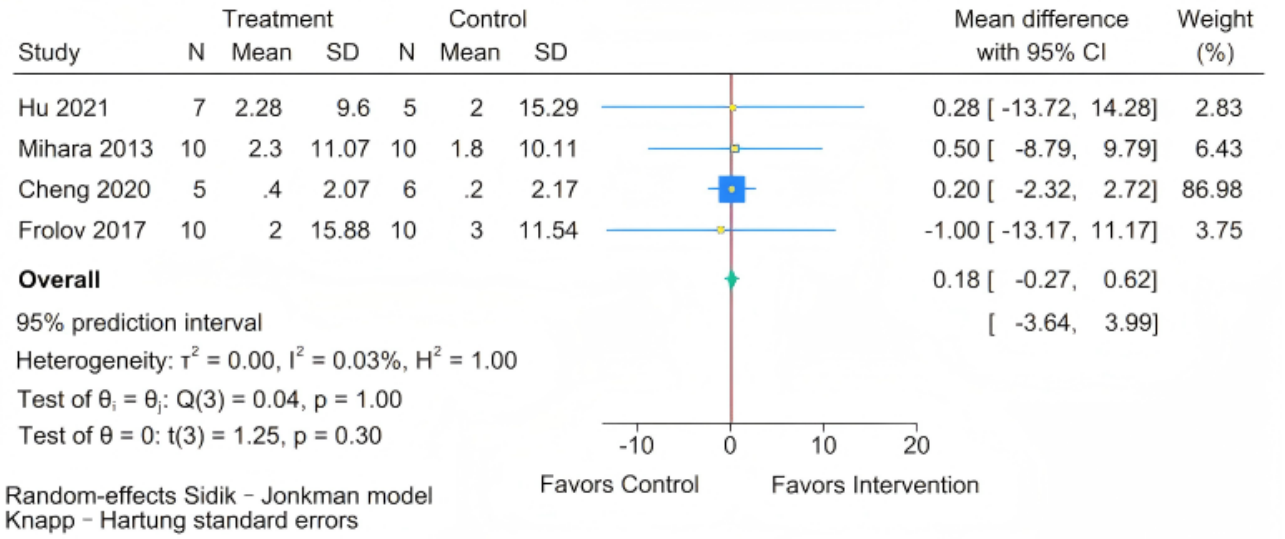

Effect of BCI on Muscle Tone (MAS)

As presented in Figure 5, four [19,41,47,52] studies assessed the effect of BCI training on upper limb muscle tone in people with chronic stroke. The meta-analysis demonstrated no statistically significant difference in muscle tone outcomes between the BCI and control groups (MD –0.48, 95% CI –1 to 0.03; *P*=.06). The test for homogeneity was not statistically significant (Q=4.69; *P*=.20). The estimated between-study variance was tau²=0.05. The I^2 statistic was 42.28%. However, the 95% PI was (–1.27 to 0.35), offering a more nuanced interpretation. As lower scores on the MAS indicate reduced spasticity, this interval suggests that in future clinical settings, the effect of BCI on muscle tone is predicted to range from a small but potentially meaningful reduction to a negligible change.

Figure 5. Forest plot for the effect of brain-computer interface (BCI) training on Modified Ashworth Scale (MAS) scores [19,41,47,52].

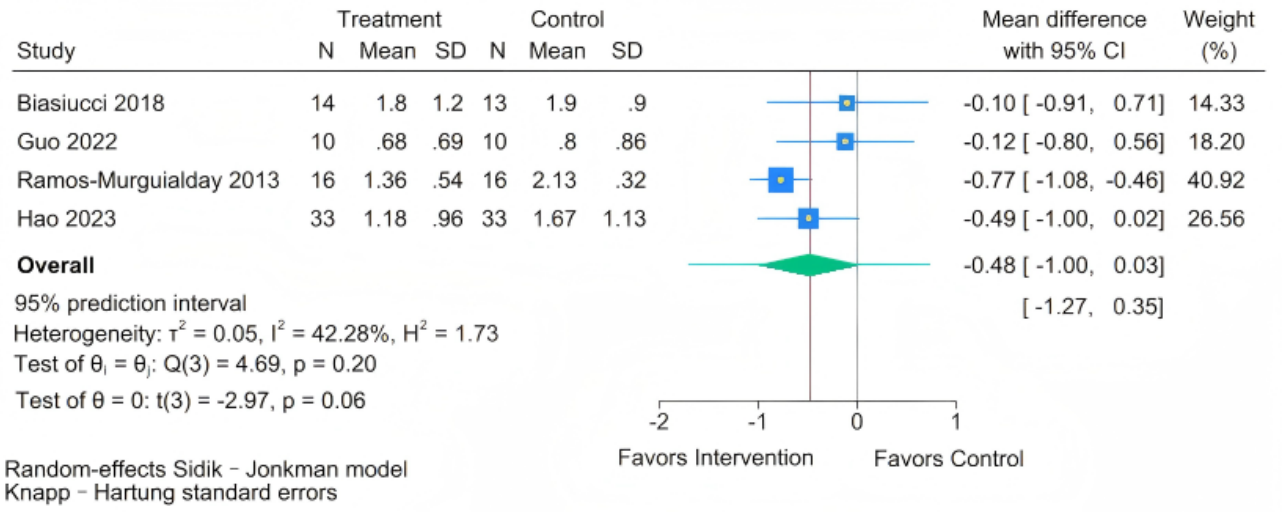

Effects on ADL

Effect of BCI on Activities of Daily Living (MBI)

As illustrated in Figure 6, six [48,49,51-54] studies evaluated the effects of BCI training on MBI scores in people with chronic stroke. The meta-analysis demonstrated a statistically significant improvement in MBI scores following BCI intervention (MD 8.38, 95% CI 2.23-14.53; *P*=.02). The test for homogeneity was not statistically significant (Q=9.85; *P*=.08). However, the I^2 statistic was 63.63%, and the estimated between-study variance was tau²=22.55, indicating substantial heterogeneity in the magnitude of the effect across studies. The 95% PI was (–3.92 to 20.53), suggesting substantial uncertainty in the magnitude of the effect across different clinical settings, with effects potentially ranging from negligible to substantially beneficial.

XSL•FO
RenderX

Figure 6. Forest plot for the effect of brain-computer interface training on Modified Barthel Index (MBI) scores [48,49,51-54].

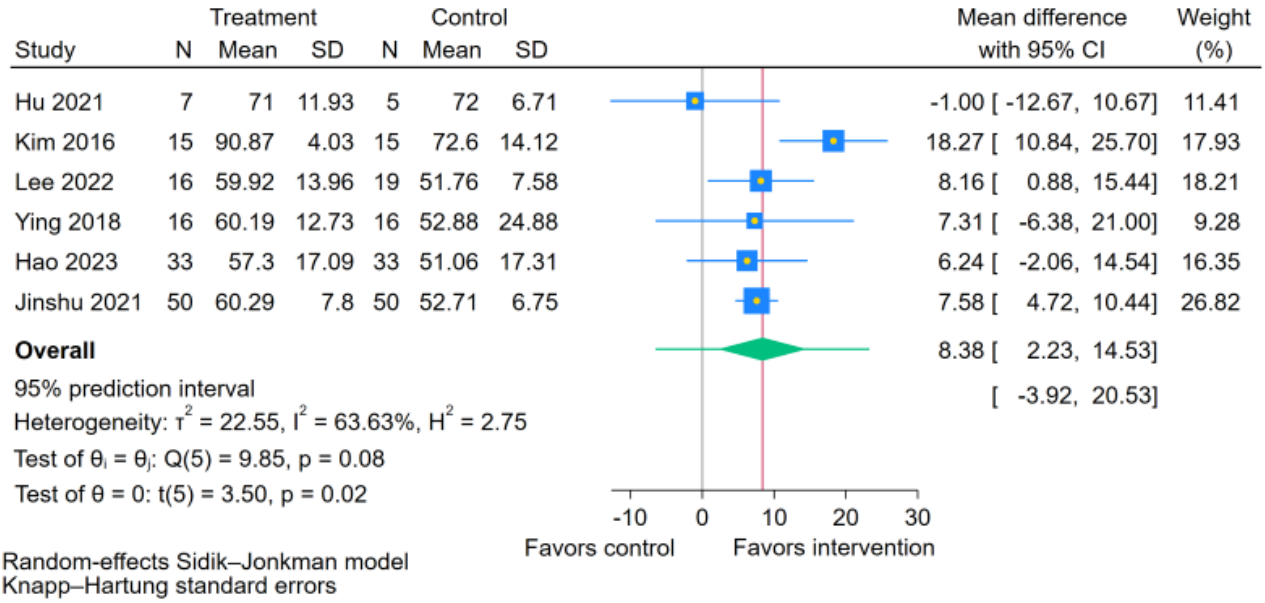

Effect of BCI on Self-Reported Arm Use (MAL)

As shown in Figure 7, five [41,42,45,48,49] studies examined the effect of BCI training on MAL scores in people with chronic stroke. The meta-analysis showed a statistically significant improvement in MAL scores following BCI intervention (MD 2.09, 95% CI 0.42-3.76; *P*=.03). The test for homogeneity was not statistically significant (Q=3.86; *P*=.42). The estimated between-study variance was tau²=1.43. The I^2 statistic was 45.7%. The 95% PI was (–0.69 to 4.54). This indicates that while the average effect is positive, the true effect in a new setting could range from a negligible or slightly negative impact to a substantial improvement in patient-perceived arm use during daily activities. The fact that the majority of the interval lies above zero strengthens the evidence for a likely beneficial effect, albeit of variable magnitude.

Figure 7. Forest plot for the effect of brain-computer interface (BCI) training on Motor Activity Log (MAL) scores [41,42,45,48,49].

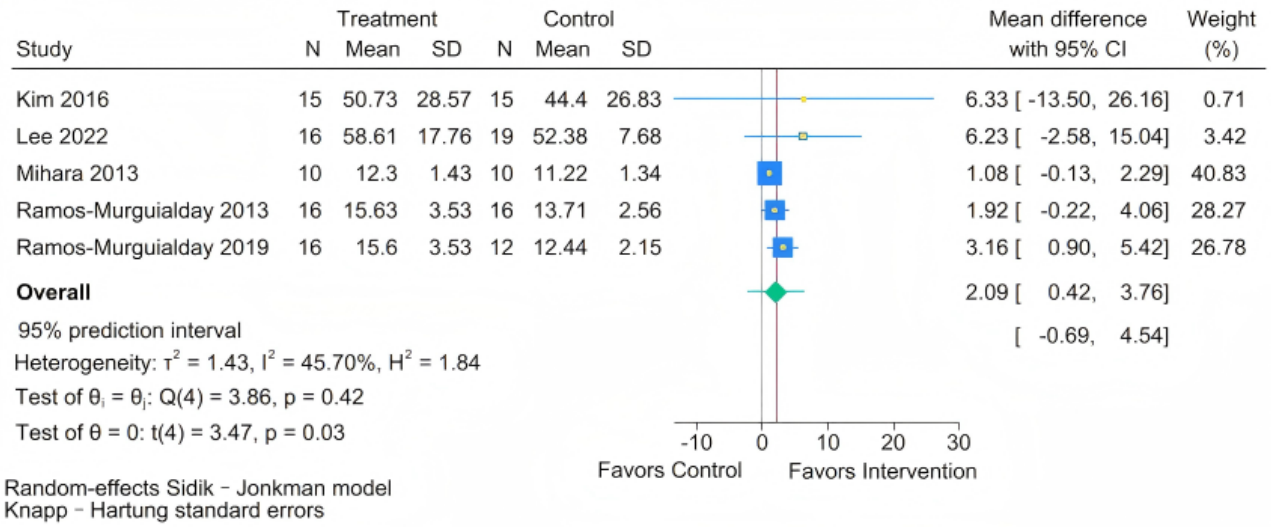

Subgroup Analysis: Feedback

In the subgroup analysis stratified by BCI feedback modality (Figure 8 [11,19,35,38,40,41,43-49,51,54,55]), the test for subgroup differences indicated no statistically significant difference between the modalities (*P*=.21). However, within-subgroup analyses revealed that only the BCI-FES paradigm showed a significant improvement in FMA-UE scores compared to routine rehabilitation therapy (RRT: MD 5, 95% CI 1.86-8.13; *P*=.01). In contrast, neither the BCI-Exoskeleton (MD 0.92, 95% CI –2.31 to 4.15; *P*=.50) nor the BCI-Visual (Beijing Intelligent Brain Science and Technology Co, Ltd) feedback (MD 2.01, 95% CI –4.14 to 8.16; *P*=.37) subgroups demonstrated significant effects. Although the differences between feedback modalities were not statistically significant, BCI-FES may be associated with greater motor recovery relative to RRT than the other modalities.

XSL•FO
RenderX

Figure 8. Subgroup analysis of Fugl-Meyer Assessment for upper extremity (FMA-UE) scores by brain-computer interface (BCI) feedback modality [11,19,35,38,40,41,43-49,51,54,55].

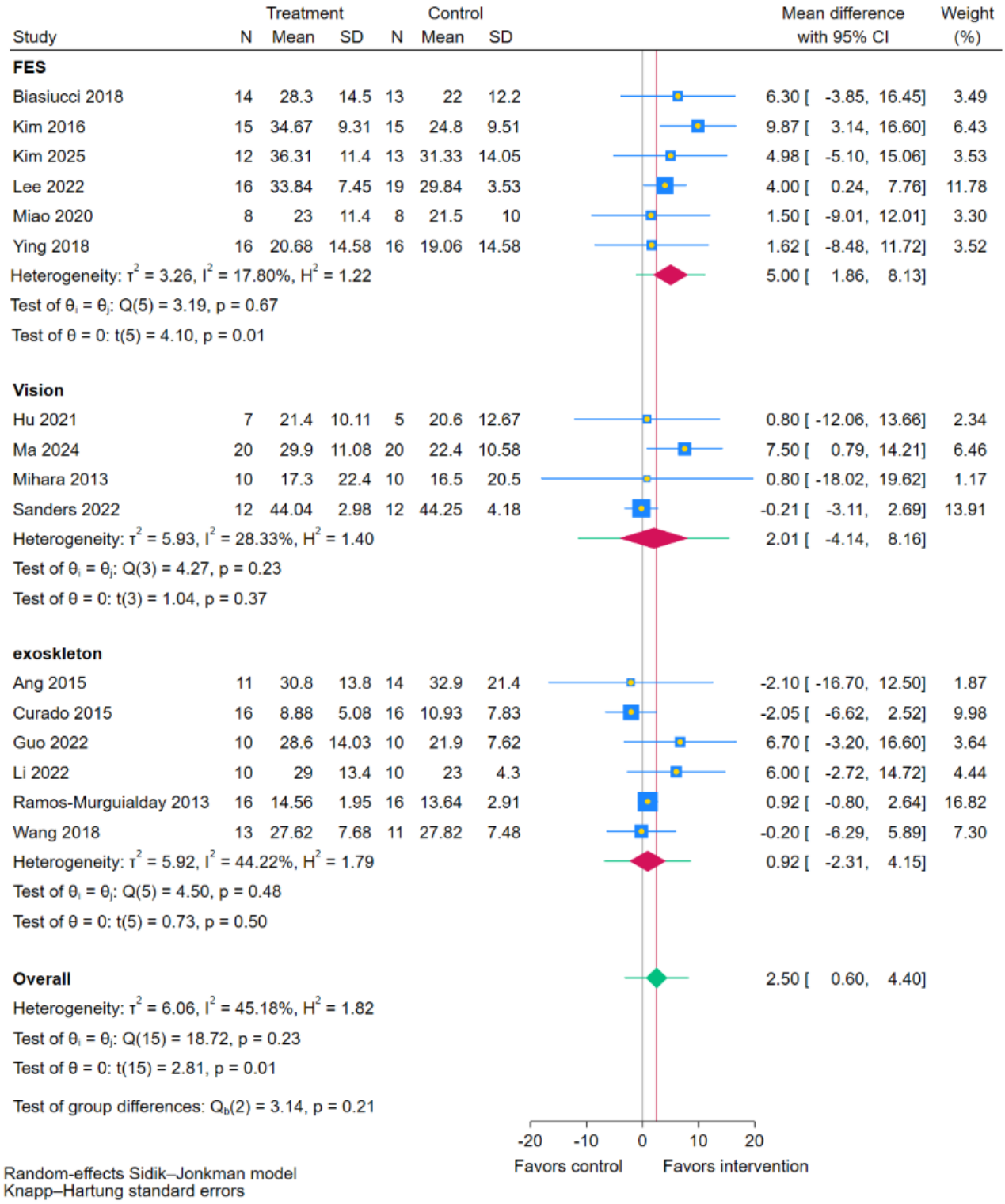

Intervention Intensity: Session Duration

Subgroup analysis based on single-session intervention duration (Figure 9 [11,19,35,38,40,41,43,44,47-49,51,54,55]) demonstrated no statistically significant difference between session duration subgroups (*P*=.16). Within the subgroups, a regimen of 30-minute sessions elicited a significant improvement in FMA-UE scores compared to RRT (MD 4.36, 95% CI 0.28-8.44; *P*=.04), whereas sessions lasting 20 minutes (MD –0.07, 95% CI –1.45 to 1.32; *P*=.85) or 60 minutes (MD 1.90, 95% CI –1.36 to 5.17; *P*=.20) did not. These results suggest that while the difference between session durations was not statistically significant, a 30-minute session may be associated with optimal outcomes.

XSL•FO
RenderX

Figure 9. Subgroup analyses of Fugl-Meyer Assessment for upper extremity (FMA-UE) scores by session duration [11,19,35,38,40,41,43-49,51,54,55].

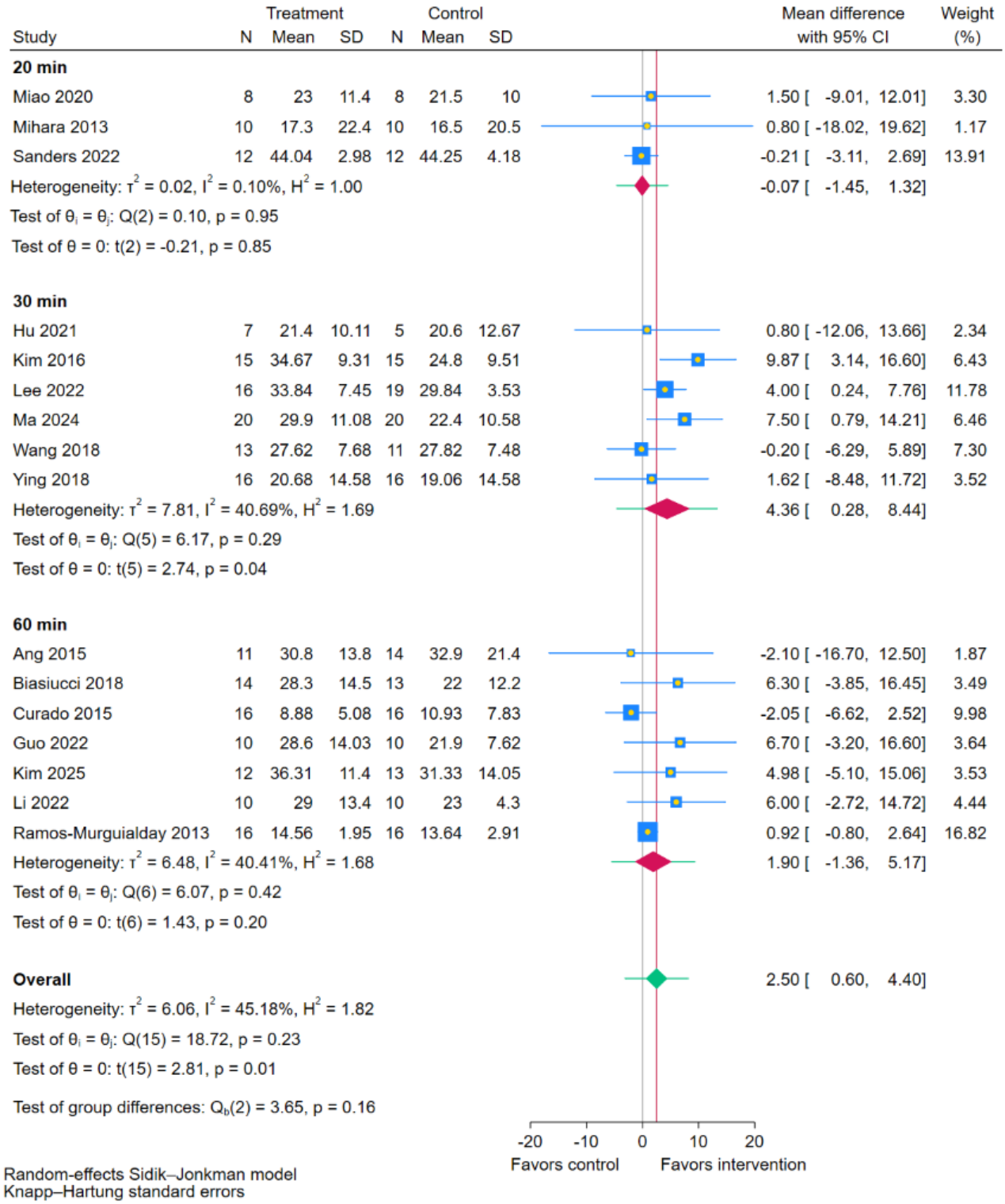

Training Sessions per Week

Subgroup analysis based on weekly intervention frequency (Figure 10 [11,19,35,38,40,41,43,44,47-49,51,54,55]) found no statistically significant difference between the frequency subgroups (*P*=.22). The higher-frequency regimen (4-5 sessions per week) was associated with a significant improvement in FMA-UE scores compared to RRT (MD 3.20, 95% CI 0.42-5.97; *P*=.03). The lower-frequency regimen (2-3 sessions per week) did not show a significant effect (MD 0.61, 95% CI –1.62 to 2.85; *P*=.51). This indicates that while the difference between weekly frequencies was not statistically significant, a higher frequency of 4-5 sessions per week may be linked to better motor recovery.

XSL•FO
RenderX

Figure 10. Subgroup analyses of Fugl-Meyer Assessment for upper extremity (FMA-UE) scores by training sessions per week [11,19,35,38,40,41,43-49,51,54,55].

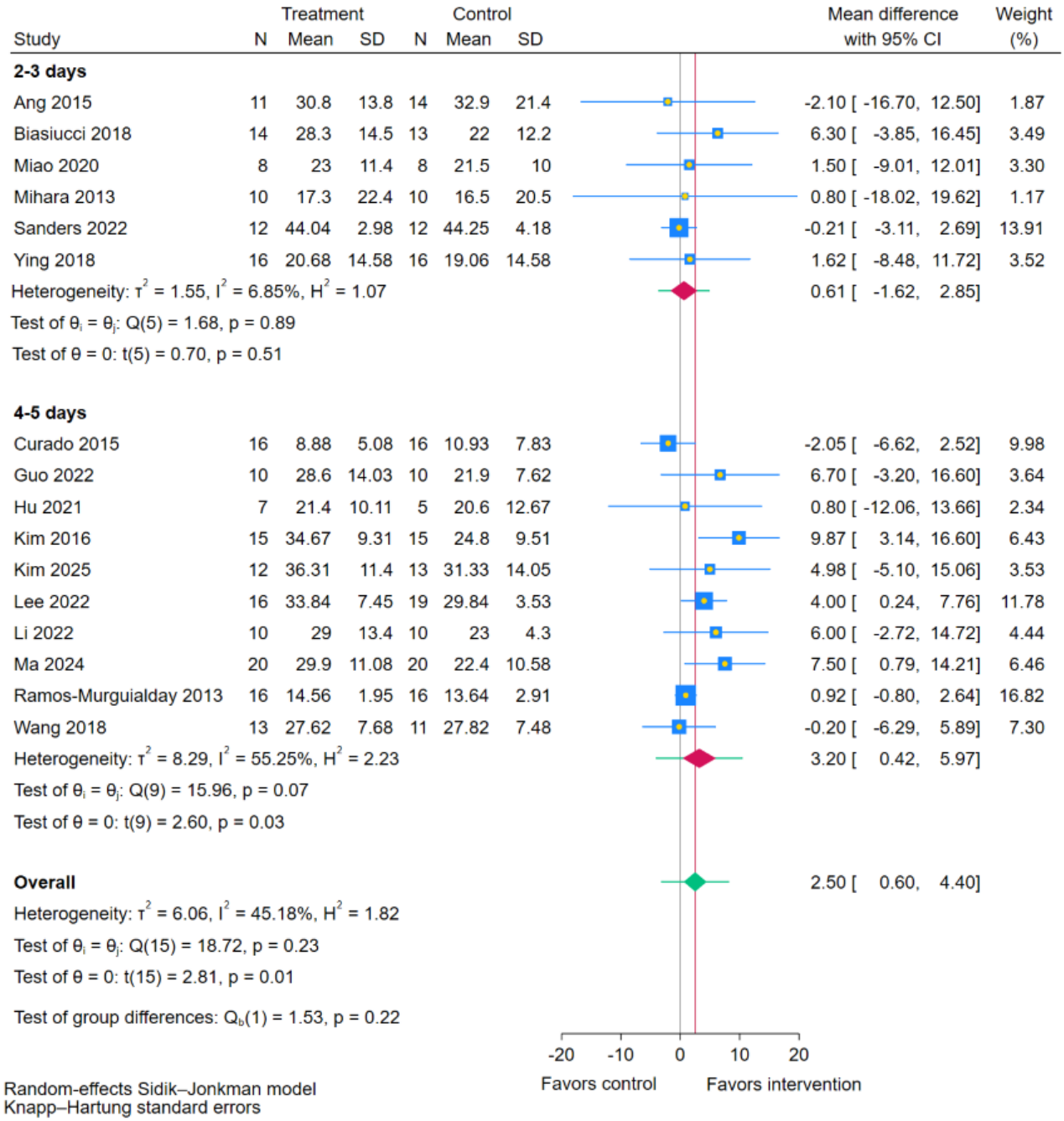

Duration of Intervention

Subgroup analysis based on the total intervention duration (Figure 11 [11,19,35,38,40,41,43-45,47-49,51,54,55]) showed no statistically significant difference between the duration subgroups (*P*=.39). Within the subgroups, a shorter-duration regimen of 2 weeks elicited a significant improvement in FMA-UE scores compared to RRT (MD 6.67, 95% CI 1.04-12.31; *P*=.04). Interventions lasting 4-5 weeks (MD 2.46, 95% CI 0.01-4.92; *P*=.05) or 8 weeks (MD 1.62, 95% CI –8.48 to 11.72) did not show significant effects. This suggests that although the difference between intervention durations was not statistically significant, a shorter, more concentrated 2-week period may be associated with superior efficacy.

XSL•FO
RenderX

Figure 11. Subgroup analyses of Fugl-Meyer Assessment for upper extremity (FMA-UE) scores by duration of intervention [11,19,35,38,40,41,43-45,47-49,51,54,55].

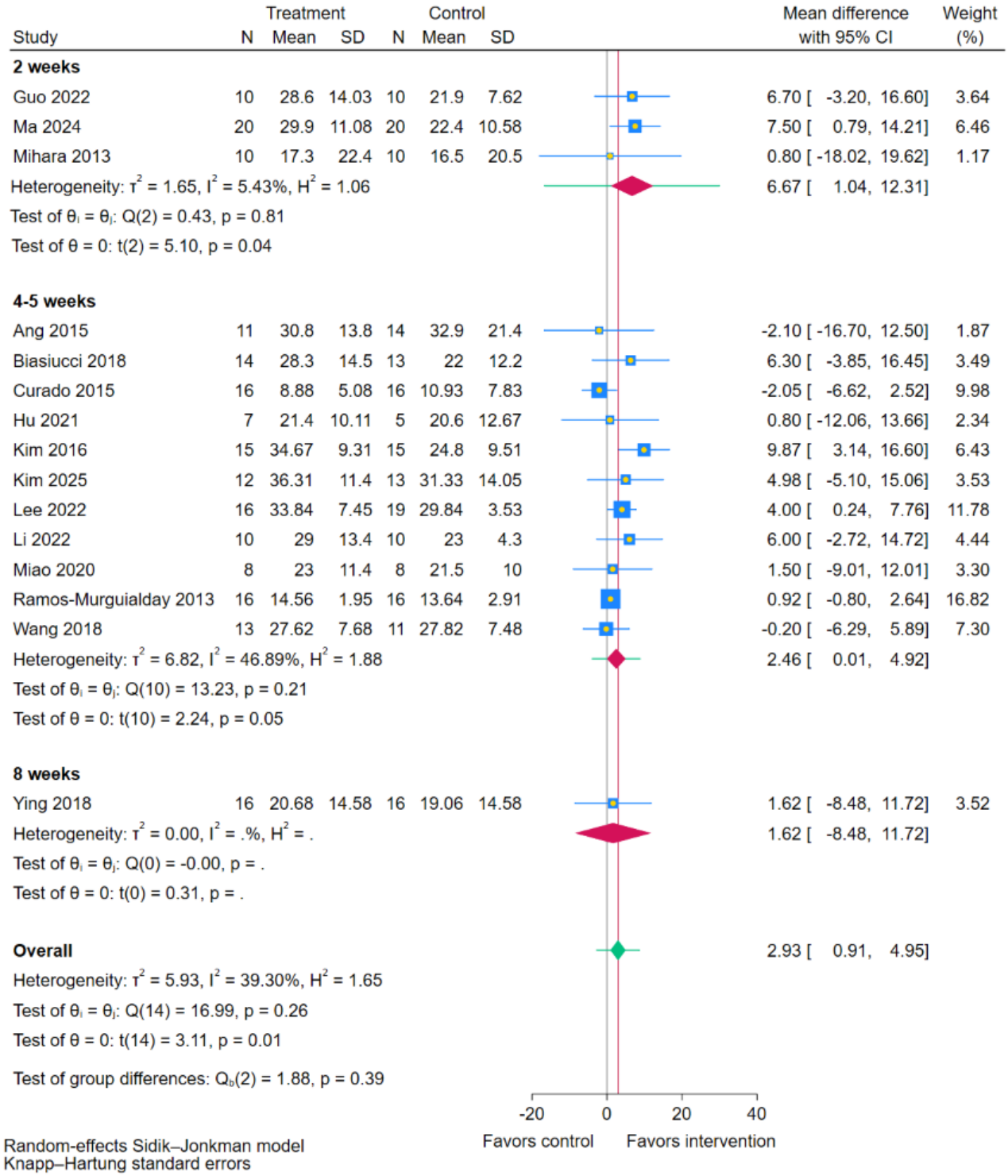

Total Number of Sessions

Subgroup analysis stratified by the total number of intervention sessions (Figure 12 [11,19,35,38,40,41,43,44,47-49,51,54,55]) indicated no statistically significant difference between the session count subgroups (*P*=.32). A lower total session count (10-12 sessions) was associated with a significant improvement in FMA-UE scores compared to RRT (MD 5.16, 95% CI 0.76-9.56; *P*=.03). A higher session count (20-24 sessions) did not demonstrate a significant effect (MD=2.40, 95% CI –0.39 to 5.20; *P*=.08). These findings imply that while the difference between total session numbers was not statistically significant, a protocol comprising 10-12 sessions may be linked to more favorable outcomes.

XSL•FO
RenderX

Figure 12. Subgroup analyses of Fugl-Meyer Assessment for upper extremity (FMA-UE) scores by total number of sessions [11,19,35,38,40,41,43,44,47-49,51,54,55].

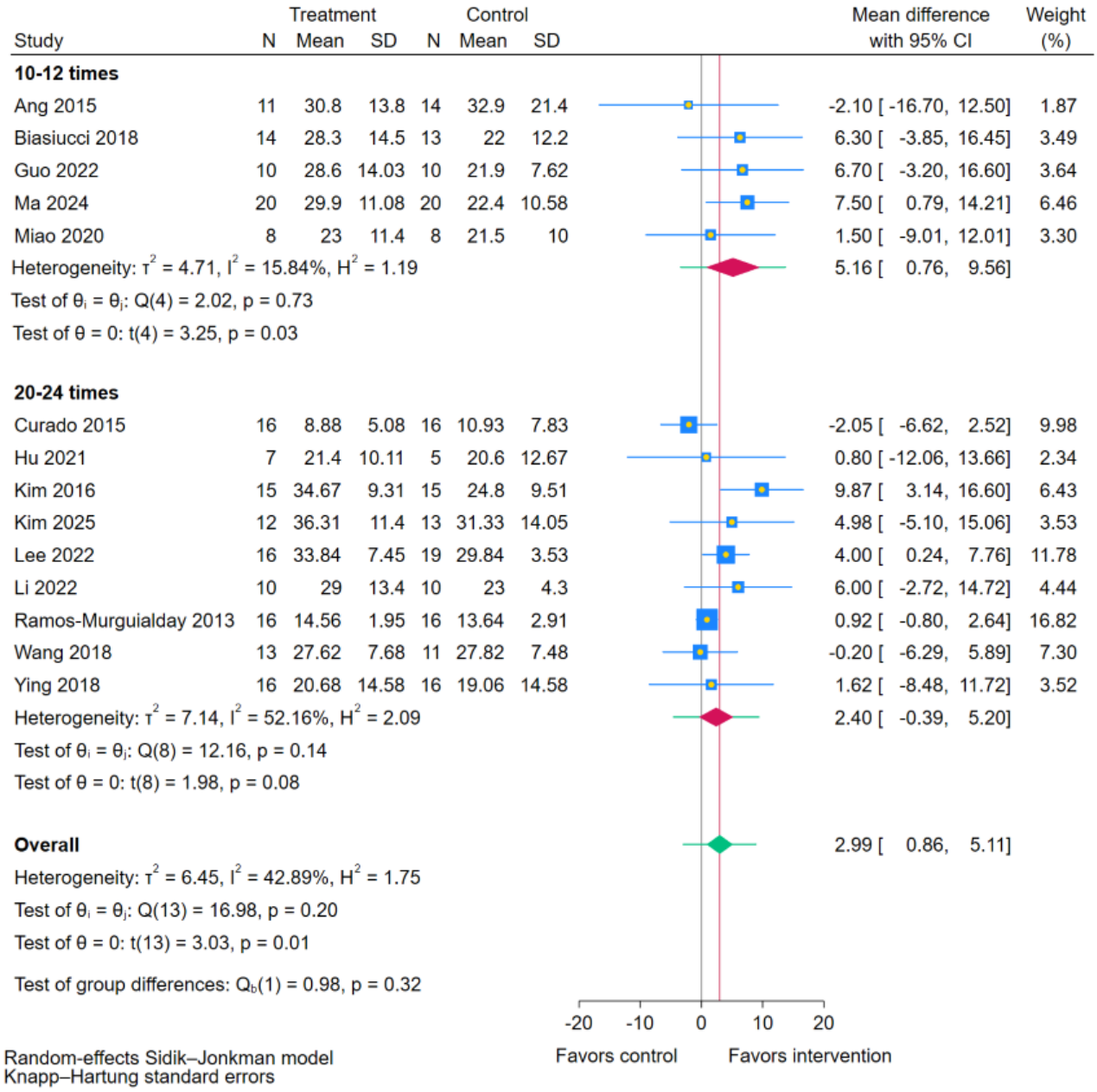

Follow-Up

Analysis of follow-up outcomes (Figure 13 [11,19,41,45,47,55]) revealed no significant differences in long-term FMA-UE improvement between BCI-based training and RRT at any follow-up interval, whether assessed at short-to-medium term (≤3 months; MD 3.23, 95% CI –8.75 to 15.22; *P*=.37) or long-term (>3 months; MD 1.19, 95% CI –4.06 to 6.43; *P*=.43).

XSL•FO
RenderX

Figure 13. Subgroup analysis of Fugl-Meyer Assessment for upper extremity (FMA-UE) scores by follow-up period [11,19,41,45,47,55].

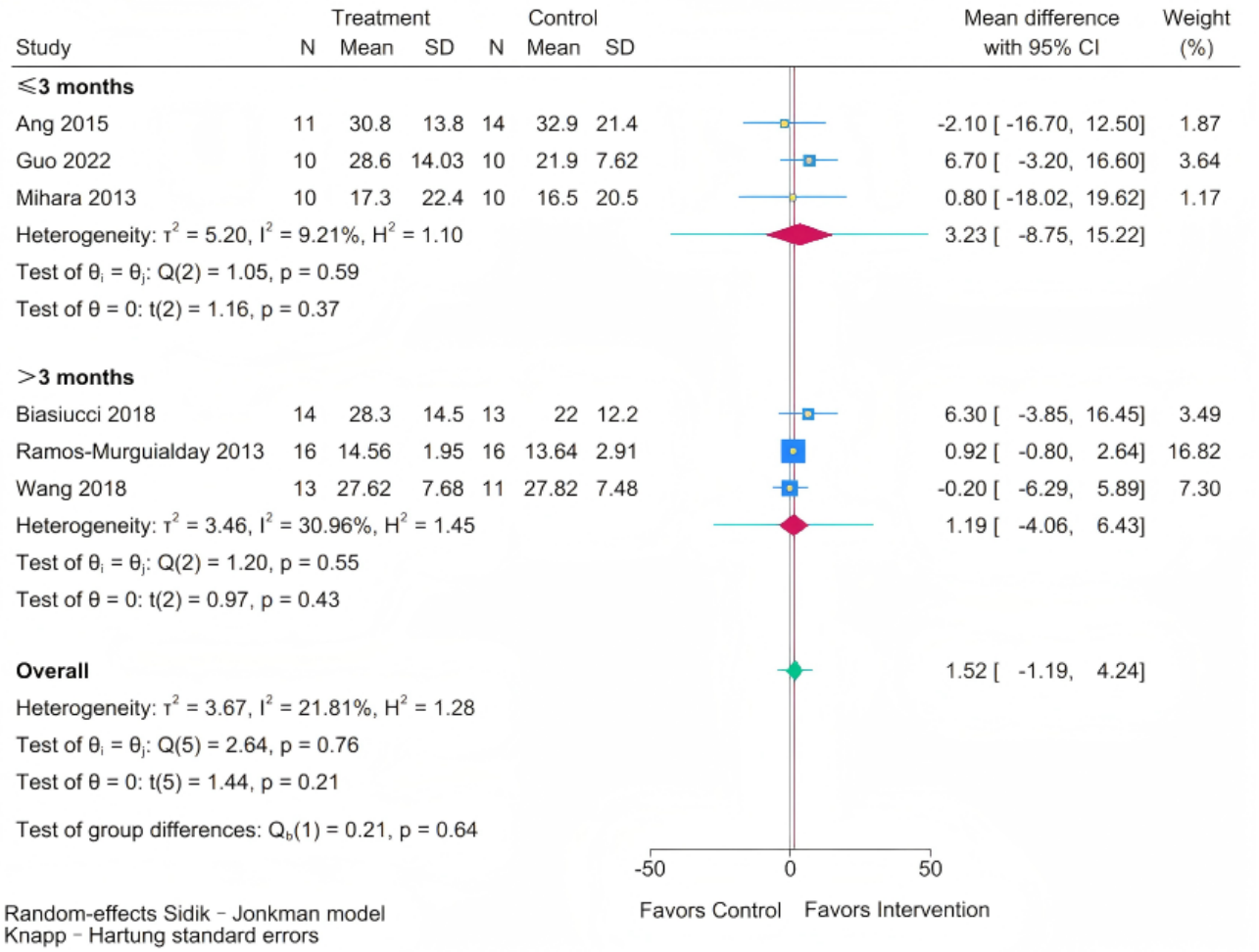

Sensitivity Analysis

A leave-one-out sensitivity analysis was performed to assess the robustness of the pooled MBI results and to investigate the influence of individual studies on the substantial observed heterogeneity (initial I^2=63.63%; τ^2=22.55).

As shown in Figure 14 [49,51-54], the sequential exclusion of each study revealed that the findings were robust overall. However, the exclusion of a single study—Kim et al [48]—led to a marked reduction in heterogeneity, with the I^2 statistic decreasing from 63.63% to 22.66% and the tau² value dropping from 22.55 to 4.03. Crucially, the pooled estimate remained statistically significant and the CI narrowed, indicating increased precision (MD 6.77, 95% CI 3.45-10.09; P<.001), compared to the original analysis (MD 8.38, 95% CI 2.23-14.53; P=.02). This suggests that while the study by Kim et al [48] was a major contributor to the statistical heterogeneity, the conclusion that BCI training improves ADL is robust.

XSL•FO
RenderX

Figure 14. Sensitivity analysis for the effect of brain-computer interface (BCI) training on Modified Barthel Index (MBI) scores [49,51-54].

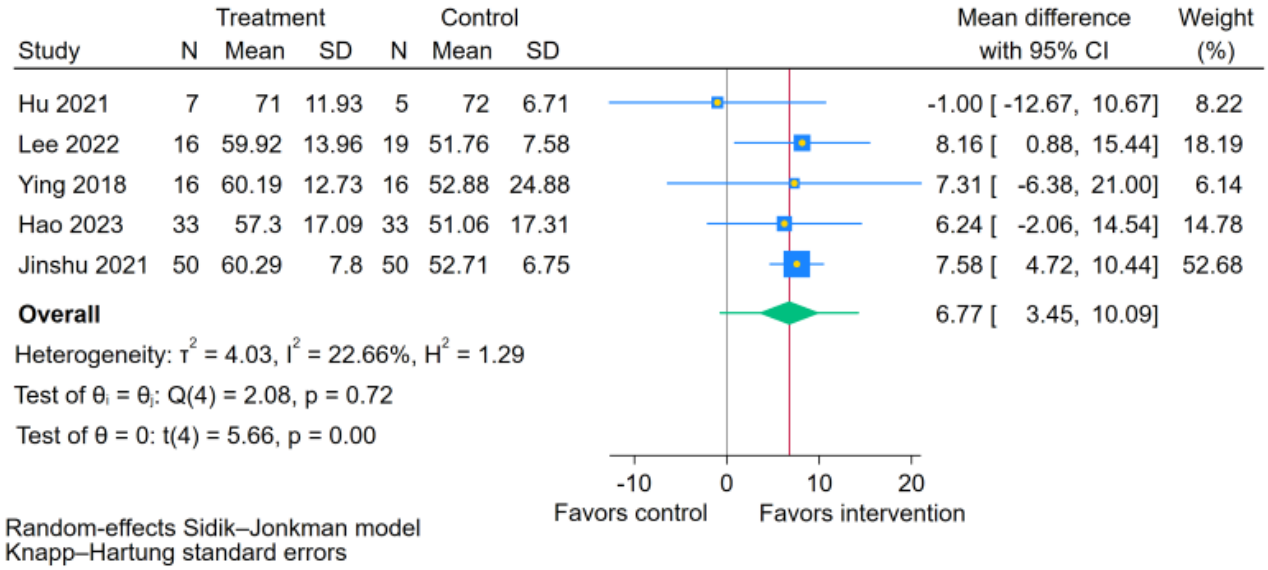

Small-Study Effects Analysis

Within this analysis, only the FMA-UE outcome pooled a sufficient number of studies (n>10) for the assessment of small-study effects. Accordingly, a funnel plot was constructed for the FMA-UE outcome (Figure 15). Visual inspection revealed no substantial asymmetry. Furthermore, the Egger regression test yielded a nonsignificant result (*P*=.20), suggesting no strong evidence for small-study effects for this outcome.

Figure 15. Funnel plot assessing small-study effects for the Fugl-Meyer Assessment for upper extremity (FMA-UE) outcome.

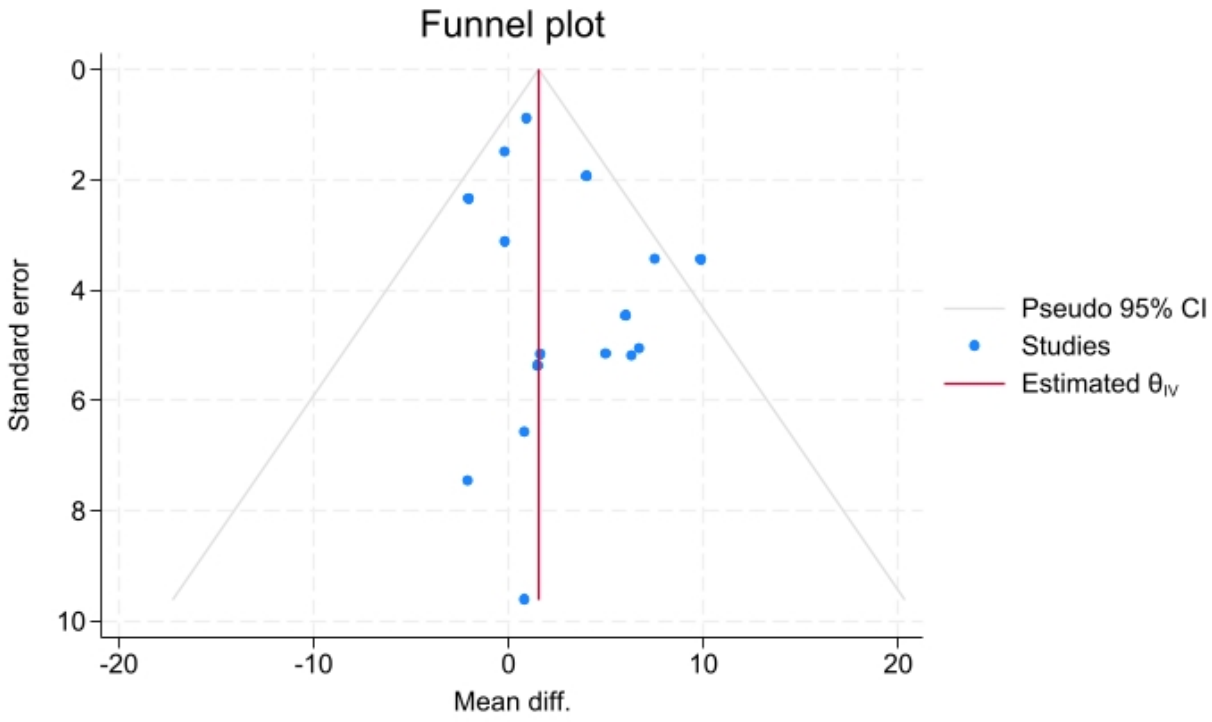

Discussion

Overview

This meta-analysis differs from the general efficacy evaluation of BCI in the treatment of chronic stroke. By conducting an in-depth analysis of how treatment parameters influence prognosis, it has deepened the understanding of this therapeutic approach. While the overall benefits of BCI for upper limb motor function and ADLs are confirmed, our meta-analysis attempts to offer greater clinical utility by identifying the specific feedback modality (BCI-FES) and a distinct, time-efficient training schedule associated with optimal recovery. These findings contribute to optimizing BCI intervention protocols and promoting their clinical

XSL•FO
RenderX

implementation, advancing the translation of this technology between experimental applications and real-world clinical practice.

Summary of Main Findings

The meta-analysis findings demonstrate that BCI training yields benefits in improving overall upper limb motor function and ADLs in patients with chronic stroke. Specifically, BCI interventions led to significant improvements in the primary outcome of motor impairment (FMA-UE) and in patient-reported and performance-based measures of daily function (MBI and MAL). In contrast, BCI training did not yield significantly superior effects on fine motor function (ARAT) or muscle tone (MAS) compared to control interventions.

Clinical Significance and Heterogeneity of Effects

Building on previous meta-analyses [9,56], our findings reveal a statistically significant incremental improvement in FMA-UE scores over conventional therapy. While this gain alone falls below the minimal clinically important difference threshold of approximately 5 points [57], it represents a meaningful augmentation to the foundational improvements from standard care. This is particularly noteworthy given the significant enhancements in MBI and MAL scores. This convergence suggests that BCI's closed-loop methodology, integrating central neural signals with peripheral feedback, may provide a synergistic effect that more effectively translates into functional gains in daily activities.

Furthermore, while the 95% CIs for the FMA-UE, MBI, and MAL confirm a positive average benefit of BCI over control interventions, the prediction intervals (FMA-UE: –2.52 to 7.22; MBI: –3.92 to 20.53; MAL: –0.69 to 4.54) reveal a more complex scenario. These intervals suggest that in future clinical settings, the effect of BCI compared to RRT could range from negligible or even slightly adverse to substantial improvements meeting the minimal clinically important difference. This indicates that current BCI training may be insufficient to yield reliable therapeutic effects for all patients in the chronic phase. The considerable heterogeneity observed suggests that the efficacy of BCI training is not uniform and may be influenced by individual patient conditions or differences in BCI treatment protocols. In line with this, a study by Guo et al [47] also emphasizes that future research should prioritize identifying patient subgroups most likely to benefit from this therapy, specifically, those with effect sizes at the upper end of the prediction interval, such as patients retaining partial integrity of the corticospinal tract.

Lack of Effect on Fine Motor Control and Spasticity

The absence of significant improvement in fine motor function, as measured by the ARAT, and in muscle tone, assessed by the MAS, warrants further mechanistic consideration. The ARAT primarily evaluates distal upper limb functions, such as grasp, grip, and pinch. The nonsignificant findings may stem from a fundamental limitation of current BCI paradigms, which often decode neural correlates of gross motor imagery (eg, whole-arm reaching or hand opening/closing) rather than the finely graded, individuated movements required for dexterous tasks. The feedback provided, particularly via exoskeleton or visual modalities, may lack the specificity necessary to engage and reinforce the delicate cortical representations of the hand.

The PI for the ARAT (–3.64-3.99) provides a deeper perspective on this null result. This interval is not only symmetrically distributed around the null effect but also entirely excludes the possibility of any large, clinically meaningful positive effects. However, given that only 6 of the included studies were rated as having a low risk of bias and the GRADE assessment indicated moderate-quality evidence for the ARAT outcome, this conclusion must be interpreted with caution.

Consequently, clinicians should be cautious about prioritizing the improvement of fine motor function as a primary goal when applying current mainstream BCI paradigms. Future optimization of BCI systems should focus on enhancing the decoding resolution of finer motor intentions, potentially by using high-density EEG or hybrid BCI approaches, and by integrating hand-specific training adjuncts such as virtual reality environments with object manipulation or wearable devices providing tactile or proprioceptive feedback to the distal limb.

Similarly, the lack of a significant effect of BCI on muscle tone suggests that its primary mechanism of action likely involves facilitating active motor control and cortical reorganization, rather than directly modulating the spinal reflex pathways underlying spasticity. In chronic stroke, hypertonia is often well-established, necessitating targeted interventions. While BCI can promote Hebbian plasticity through the associative pairing of motor intention and movement execution, this effect may be insufficient to reverse impaired supraspinal inhibitory control over the spinal motor pool.

Nevertheless, the 95% PI for the MAS (–1.27 to 0.35) provides valuable clinical insight. As lower scores indicate reduced spasticity, this interval—spanning from "no change" to "improvement"—suggests that BCI therapy is unlikely to exacerbate spasticity in future applications. Moreover, its lower bound of –1.27 indicates that under specific conditions, such as when combined with certain forms of FES, BCI may yield meaningful reductions in muscle tone. This potentially "non-harmful" profile, particularly when considered alongside the moderate quality of the existing evidence, represents an important factor for clinical decision-making and offers a preliminary rationale for exploring BCI as a component of comprehensive spasticity management protocols.

Superiority of BCI-FES and the Role of Feedback Modality

Subgroup analyses revealed that only the BCI-FES paradigm demonstrated significantly greater improvement in FMA-UE compared to control. This superiority can be explained through the lens of neuroplasticity and sensorimotor integration. The BCI-FES paradigm creates a closed-loop system that tightly couples motor intention with peripheral afferent feedback. According to Hebbian learning principles [58], which posit that "neurons that fire together, wire together," the synchronous activation of the motor cortex (during attempted movement imagery) and the somatosensory cortex (via FES-induced limb movement and proprioceptive input) strengthens the synaptic

XSL•FO
RenderX

connections within the sensorimotor network, aligning with findings from the meta-analysis by Li et al [9].

The lack of significant benefit with BCI-Exoskeleton may relate to factors such as device complexity, comfort limitations, and suboptimal anatomical fit. Furthermore, excessive robotic assistance could potentially reduce patient engagement and diminish the crucial training effect driven by active neural effort [59]. The negative results with BCI-Vision indicate that visual feedback alone, in the absence of concomitant somatosensory input and actual limb movement, may have limited efficacy in driving neural reorganization and functional recovery, particularly in patients with chronic deficits or more severe functional impairments.

Sustainability of Benefits and the Need for Maintenance

Subgroup analysis based on follow-up duration revealed no significant sustained advantage of BCI over control after the active treatment phase ceased. This implies that the functional gains may lack long-term stability without ongoing application.

In chronic patients, BCI may effectively induce neuroplasticity during the intensive training period, potentially by reinforcing specific neural pathways or facilitating compensatory mechanisms. However, these newly formed connections or patterns might lack sufficient stability or robustness. Following intervention cessation, without ongoing functional application or specific maintenance training, these acquired neural adaptations may gradually regress or weaken. Conversely, standard RRT protocols often inherently incorporate recommendations for continued activity.

These findings highlight the necessity for systematic maintenance strategies—such as telerehabilitation, behavioral incentive programs, or continued use of assistive technologies—to be integrated postintervention. Addressing this challenge of sustained efficacy represents a crucial future research direction.

Toward an Optimal and Efficient Intervention Protocol

Although tests for subgroup differences were not statistically significant, the within-subgroup comparisons revealed a consistent and clinically meaningful pattern favoring specific parameters. The most robust finding is the efficacy of a short-term, high-density protocol.

This paradigm—comprising sessions of approximately 30 minutes each, delivered 4-5 times per week over a total of 10-12 sessions (approximately 2 weeks)—yielded optimal FMA-UE outcomes compared to control interventions. While broader intensity subgroup comparisons (eg, session duration or total weeks) were not statistically significant, this specific, condensed protocol was the only intensity paradigm that consistently demonstrated a significant within-subgroup effect. This finding underscores the potential primacy of training density—the concentration of practice within a shorter timeframe—over the total intervention duration [50]. Furthermore, a higher frequency of 4-5 sessions per week proved superior to regimens of 2-3 sessions per week, indicating that more frequent exposure facilitates sharper motor patterns and stronger memory traces [60].

Notably, completing 10-12 sessions within 2 weeks was more effective than protocols delivering 20-24 sessions over 4-5 weeks. This observation strongly supports the established concept that maximizing neuroplasticity in people with chronic stroke often requires intensive, repetitive, and focused training to overcome neural inhibition and promote synaptic strengthening [61].

Therefore, based on the current evidence, a protocol of 30-minute sessions, administered 4-5 times per week over 2 weeks (totaling 10-12 sessions), emerges as a promising and efficient model for BCI intervention. It is crucial to emphasize that this proposal is not intended as a definitive guideline but rather highlights a potentially optimal treatment paradigm derived from the existing, albeit limited, data. This model warrants prioritization and validation in future rigorous research.

Research Prospects

Although subgroup analyses favor short-term, high-frequency protocols, this intensity may be insufficient to induce lasting neuroplastic reorganization. This observation aligns with the dissipation of functional gains at follow-up. Future high-quality RCTs are required to (1) delineate the dose-response relationship of BCI training, (2) analyze the synergistic effects of BCI combined with complementary therapies to optimize rehabilitation protocols, and (3) evaluate efficacy differentials based on lesion characteristics and upper-limb impairment severity in patients with chronic disease, thereby identifying responsive subpopulations. These investigations aim to inform evidence-based rehabilitation strategies and research priorities for stroke recovery.

Limitations

Several limitations warrant consideration in this meta-analysis. First, while the included 21 studies underwent rigorous quality assessment using the Cochrane RoB 2 tool, only 6 were rated as having “Low” risk of bias. This relatively low proportion of high-quality studies may limit the robustness of our findings. Second, insufficient studies (fewer than 5) were included in some subgroup analyses, potentially compromising the reliability of conclusions drawn for those specific comparisons. Third, the subgroup analyses, particularly for intervention intensity, were likely underpowered to detect statistically significant differences between parameters due to the limited number of studies in each category. Therefore, the identified optimal protocol should be viewed as the most evidence-based recommendation from the current data, rather than a definitively proven superior approach. Future research should prioritize incorporating a greater number of high-quality RCTs. Furthermore, greater emphasis is needed on exploring the impact of BCI training intervention intensity and focusing on outcomes such as improvements in muscle tone and standardized assessments like the ARAT. Investigating these areas represents promising avenues for advancing stroke rehabilitation.

XSL•FO
RenderX

Conclusion

Low- to moderate-certainty evidence suggests that BCI training, particularly the BCI-FES paradigm, can improve upper limb motor function and ADL in people with chronic stroke on average. However, wide prediction intervals indicate the effect may vary substantially across settings, ranging from negligible to beneficial. Subgroup analyses suggested a potential optimal protocol of 30-minute sessions, 4-5 times per week for 2 weeks, but these findings are limited by the small number of studies in each subgroup and the high risk of bias in several included trials. Therefore, this proposed protocol should be viewed as preliminary and requires validation in future, high-quality RCTs. Future research should also focus on identifying patient subgroups most likely to benefit and on strategies to sustain long-term gains.

Data Availability

The datasets generated or analyzed during this systematic review are available from the corresponding author on reasonable request.

Authors' Contributions

HJC and GJY were responsible for conceptualization. HJC completed database searching. HJC and GJY completed article screening, data extraction, and critical appraisal. HJC was responsible for data curation and analysis. HJC and GJY contributed to data interpretation. HJC wrote the original draft. All authors contributed to reviewing and editing the final manuscript and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA checklist.
[DOCX File , 30 KB - jmir_v28i1e79132_app1.docx]

Multimedia Appendix 2
Search strategy.
[DOCX File , 19 KB - jmir_v28i1e79132_app2.docx]

Multimedia Appendix 3
GRADE assessment.
[DOCX File , 36 KB - jmir_v28i1e79132_app3.docx]

Multimedia Appendix 4
The characteristics of included literatures.
[DOCX File , 31 KB - jmir_v28i1e79132_app4.docx]

References

1. Gibbons EM, Thomson AN, de Noronha M, Joseph S. Are virtual reality technologies effective in improving lower limb outcomes for patients following stroke - a systematic review with meta-analysis. Top Stroke Rehabil 2016;23(6):440-457. [doi: 10.1080/10749357.2016.1183349] [Medline: 27237336]
2. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019. Lancet Psychiatry 2022;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
3. Ahmed I, Mustafaoglu R, Erhan B. The effects of low-intensity resistance training with blood flow restriction versus traditional resistance exercise on lower extremity muscle strength and motor functionin ischemic stroke survivors: a randomized controlled trial. Top Stroke Rehabil 2024;31(4):418-429. [doi: 10.1080/10749357.2023.2259170] [Medline: 37724785]
4. O'Dell MW, Jaywant A, Frantz M, Patel R, Kwong E, Wen K, et al. Changes in the activity measure for post-acute care domains in persons with stroke during the first year after discharge from inpatient rehabilitation. Arch Phys Med Rehabil 2021;102(4):645-655. [doi: 10.1016/j.apmr.2020.11.020] [Medline: 33440132]
5. Heller A, Wade DT, Wood VA, Sunderland A, Hewer RL, Ward E. Arm function after stroke: measurement and recovery over the first three months. J Neurol Neurosurg Psychiatry 1987;50(6):714-719 [FREE Full text] [doi: 10.1136/jnnp.50.6.714] [Medline: 3612152]

XSL•FO
RenderX

6. Duncan PW, Goldstein LB, Horner RD, Landsman PB, Samsa GP, Matchar DB. Similar motor recovery of upper and lower extremities after stroke. Stroke 1994;25(6):1181-1188. [doi: 10.1161/01.str.25.6.1181] [Medline: 8202977]
7. McCabe J, Monkiewicz M, Holcomb J, Pundik S, Daly JJ. Comparison of robotics, functional electrical stimulation, and motor learning methods for treatment of persistent upper extremity dysfunction after stroke: a randomized controlled trial. Arch Phys Med Rehabil 2015;96(6):981-990 [FREE Full text] [doi: 10.1016/j.apmr.2014.10.022] [Medline: 25461822]
8. Remsik A, Young B, Vermilyea R, Kiekhoefer L, Abrams J, Evander Elmore S, et al. A review of the progression and future implications of brain-computer interface therapies for restoration of distal upper extremity motor function after stroke. Expert Rev Med Devices 2016;13(5):445-454. [doi: 10.1080/17434440.2016.1174572] [Medline: 27112213]
9. Li D, Li R, Song Y, Qin W, Sun G, Liu Y, et al. Effects of brain-computer interface based training on post-stroke upper-limb rehabilitation: a meta-analysis. J Neuroeng Rehabil 2025;22(1):44 [FREE Full text] [doi: 10.1186/s12984-025-01588-x] [Medline: 40033447]
10. Li M, Liu Y, Wu Y, Liu S, Jia J, Zhang L. Neurophysiological substrates of stroke patients with motor imagery-based brain-computer interface training. Int J Neurosci 2014;124(6):403-415. [doi: 10.3109/00207454.2013.850082] [Medline: 24079396]
11. Ang KK, Chua KSG, Phua KS, Wang C, Chin ZY, Kuah CWK, et al. A randomized controlled trial of EEG-based motor imagery brain-computer interface robotic rehabilitation for stroke. Clin EEG Neurosci 2015;46(4):310-320. [doi: 10.1177/1550059414522229] [Medline: 24756025]
12. Pichiorri F, Morone G, Petti M, Toppi J, Pisotta I, Molinari M, et al. Brain-computer interface boosts motor imagery practice during stroke recovery. Ann Neurol 2015;77(5):851-865. [doi: 10.1002/ana.24390] [Medline: 25712802]
13. Miller KJ, Schalk G, Fetz EE, den Nijs M, Ojemann JG, Rao RPN. Cortical activity during motor execution, motor imagery, and imagery-based online feedback. Proc Natl Acad Sci U S A 2010;107(9):4430-4435 [FREE Full text] [doi: 10.1073/pnas.0913697107] [Medline: 20160084]
14. Casimo K, Weaver KE, Wander J, Ojemann JG. BCI use and its relation to adaptation in cortical networks. IEEE Trans Neural Syst Rehabil Eng 2017;25(10):1697-1704 [FREE Full text] [doi: 10.1109/TNSRE.2017.2681963] [Medline: 28320670]
15. Lebedev MA, Nicolelis MAL. Brain-machine interfaces: from basic science to neuroprostheses and neurorehabilitation. Physiol Rev 2017;97(2):767-837 [FREE Full text] [doi: 10.1152/physrev.00027.2016] [Medline: 28275048]
16. Ma T, Li H, Deng L, Yang H, Lv X, Li P, et al. The hybrid BCI system for movement control by combining motor imagery and moving onset visual evoked potential. J Neural Eng 2017;14(2):026015. [doi: 10.1088/1741-2552/aa5d5f] [Medline: 28145274]
17. Cervera MA, Soekadar SR, Ushiba J, Millán JDR, Liu M, Birbaumer N, et al. Brain-computer interfaces for post-stroke motor rehabilitation: a meta-analysis. Ann Clin Transl Neurol 2018;5(5):651-663 [FREE Full text] [doi: 10.1002/acn3.544] [Medline: 29761128]
18. Daly JJ, Cheng R, Rogers J, Litinas K, Hrovat K, Dohring M. Feasibility of a new application of noninvasive brain computer interface (BCI): a case study of training for recovery of volitional motor control after stroke. J Neurol Phys Ther 2009;33(4):203-211. [doi: 10.1097/NPT.0b013e3181c1fc0b] [Medline: 20208465]
19. Biasiucci A, Leeb R, Iturrate I, Perdikis S, Al-Khodairy A, Corbet T, et al. Brain-actuated functional electrical stimulation elicits lasting arm motor recovery after stroke. Nat Commun 2018;9(1):2421 [FREE Full text] [doi: 10.1038/s41467-018-04673-z] [Medline: 29925890]
20. Liu X, Zhang W, Li W, Zhang S, Lv P, Yin Y. Effects of motor imagery based brain-computer interface on upper limb function and attention in stroke patients with hemiplegia: a randomized controlled trial. BMC Neurol 2023;23(1):136 [FREE Full text] [doi: 10.1186/s12883-023-03150-5] [Medline: 37003976]
21. Xie Y, Yang Y, Jiang H, Duan X, Gu L, Qing W, et al. Brain-machine interface-based training for improving upper extremity function after stroke: a meta-analysis of randomized controlled trials. Front Neurosci 2022;16:949575 [FREE Full text] [doi: 10.3389/fnins.2022.949575] [Medline: 35992923]
22. Peng Y, Wang J, Liu Z, Zhong L, Wen X, Wang P, et al. The application of brain-computer interface in upper limb dysfunction after stroke: a systematic review and meta-analysis of randomized controlled trials. Front Hum Neurosci 2022;16:798883 [FREE Full text] [doi: 10.3389/fnhum.2022.798883] [Medline: 35422693]
23. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
24. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
25. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc 2021;109(2):174-200 [FREE Full text] [doi: 10.5195/jmla.2021.962] [Medline: 34285662]
26. Cumpston M, Li T, Page MJ, Chandler J, Welch VA, Higgins JP, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane handbook for systematic reviews of interventions. Cochrane Database Syst Rev 2019;10(10):ED000142. [doi: 10.1002/14651858.ED000142] [Medline: 31643080]

XSL•FO
RenderX

27. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
28. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336(7650):924-926 [FREE Full text] [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
29. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
30. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: I is not an absolute measure of heterogeneity. Res Synth Methods 2017;8(1):5-18. [doi: 10.1002/jrsm.1230] [Medline: 28058794]
31. Michaelis R, Tang V, Wagner JL, Modi AC, LaFrance WC, Goldstein LH, et al. Cochrane systematic review and meta-analysis of the impact of psychological treatments for people with epilepsy on health-related quality of life. Epilepsia 2018;59(2):315-332 [FREE Full text] [doi: 10.1111/epi.13989] [Medline: 29313968]
32. IntHout J, Ioannidis JPA, Borm GF. The hartung-knapp-sidik-jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard dersimonian-laird method. BMC Med Res Methodol 2014;14:25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
33. Gomes C, Geels J, Debray TPA, Malekzadeh A, Asselbergs FW, Linschoten M. Risk prediction models for cancer therapy related cardiac dysfunction in patients with cancer and cancer survivors: systematic review and meta-analysis. BMJ 2025;390:e084062 [FREE Full text] [doi: 10.1136/bmj-2025-084062] [Medline: 40987514]
34. Fugl-Meyer AR, Jääskö L, Leyman I, Olsson S, Steglind S. The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. Scand J Rehabil Med 1975;7(1):13-31. [Medline: 1135616]
35. Ma Z, Wu J, Cao Z, Hua X, Zheng M, Xing X, et al. Motor imagery-based brain-computer interface rehabilitation programs enhance upper extremity performance and cortical activation in stroke patients. J Neuroeng Rehabil 2024;21(1):91 [FREE Full text] [doi: 10.1186/s12984-024-01387-w] [Medline: 38812014]
36. Lu R, Pang Z, Gao T, He Z, Hu Y, Zhuang J, et al. Multisensory BCI promotes motor recovery via high-order network-mediated interhemispheric integration in chronic stroke. BMC Med 2025;23(1):380. [doi: 10.1186/s12916-025-04214-8] [Medline: 40598460]
37. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
38. Miao Y, Chen S, Zhang X, Jin J, Xu R, Daly I, et al. BCI-based rehabilitation on the stroke in sequela stage. Neural Plast 2020;2020:8882764 [FREE Full text] [doi: 10.1155/2020/8882764] [Medline: 33414824]
39. Cheng N, Phua KS, Lai HS, Tam PK, Tang KY, Cheng KK, et al. Brain-computer interface-based soft robotic glove rehabilitation for stroke. IEEE Trans Biomed Eng 2020;67(12):3339-3351. [doi: 10.1109/TBME.2020.2984003] [Medline: 32248089]
40. Kim MS, Park H, Kwon I, An K, Kim H, Park G, et al. Efficacy of brain-computer interface training with motor imagery-contingent feedback in improving upper limb function and neuroplasticity among persons with chronic stroke: a double-blinded, parallel-group, randomized controlled trial. J Neuroeng Rehabil 2025;22(1):1 [FREE Full text] [doi: 10.1186/s12984-024-01535-2] [Medline: 39757218]
41. Ramos-Murguialday A, Broetz D, Rea M, Läer L, Yilmaz O, Brasil FL, et al. Brain-machine interface in chronic stroke rehabilitation: a controlled study. Ann Neurol 2013;74(1):100-108 [FREE Full text] [doi: 10.1002/ana.23879] [Medline: 23494615]
42. Ramos-Murguialday A, Curado MR, Broetz D, Yilmaz, Brasil FL, Liberati G, et al. Brain-machine interface in chronic stroke: randomized trial long-term follow-up. Neurorehabil Neural Repair 2019;33(3):188-198. [doi: 10.1177/1545968319827573] [Medline: 30722727]
43. Curado MR, Cossio EG, Broetz D, Agostini M, Cho W, Brasil FL, et al. Residual upper arm motor function primes innervation of paretic forearm muscles in chronic stroke after brain-machine interface (BMI) training. PLoS One 2015;10(10):e0140161 [FREE Full text] [doi: 10.1371/journal.pone.0140161] [Medline: 26495971]
44. Li X, Wang J, Cao X, Huang W, Hu Y. Soft robotic glove with alpha band brain computer interface for post-stroke hand function rehabilitation. : IEEE; 2022 Presented at: 14th Biomedical Engineering International Conference (BMEiCON); 2022 November 10-13; Songkhla, Thailand. [doi: 10.1109/bmeicon56653.2022.10012103]
45. Mihara M, Hattori N, Hatakenaka M, Yagura H, Kawano T, Hino T, et al. Near-infrared spectroscopy-mediated neurofeedback enhances efficacy of motor imagery-based training in poststroke victims: a pilot study. Stroke 2013;44(4):1091-1098. [doi: 10.1161/STROKEAHA.111.674507] [Medline: 23404723]
46. Sanders Z, Fleming MK, Smejka T, Marzolla MC, Zich C, Rieger SW, et al. Self-modulation of motor cortex activity after stroke: a randomized controlled trial. Brain 2022;145(10):3391-3404 [FREE Full text] [doi: 10.1093/brain/awac239] [Medline: 35960166]
47. Guo N, Wang X, Duanmu D, Huang X, Li X, Fan Y, et al. SSVEP-based brain computer interface controlled soft robotic glove for post-stroke hand function rehabilitation. IEEE Trans Neural Syst Rehabil Eng 2022;30:1737-1744. [doi: 10.1109/TNSRE.2022.3185262] [Medline: 35731756]

XSL•FO
RenderX

48. Kim T, Kim S, Lee B. Effects of action observational training plus brain-computer interface-based functional electrical stimulation on paretic arm motor recovery in patient with stroke: a randomized controlled trial. Occup Ther Int 2016;23(1):39-47 [FREE Full text] [doi: 10.1002/oti.1403] [Medline: 26301519]
49. Lee S, Kim SS, Lee B. Action observation training and brain-computer interface controlled functional electrical stimulation enhance upper extremity performance and cortical activation in patients with stroke: a randomized controlled trial. Physiother Theory Pract 2022;38(9):1126-1134. [doi: 10.1080/09593985.2020.1831114] [Medline: 33026895]
50. Frolov AA, Mokienko O, Lyukmanov R, Biryukova E, Kotov S, Turbina L, et al. Post-stroke rehabilitation training with a motor-imagery-based brain-computer interface (bci)-controlled hand exoskeleton: a randomized controlled multicenter trial. Front Neurosci 2017;11:400 [FREE Full text] [doi: 10.3389/fnins.2017.00400] [Medline: 28775677]
51. Hu Y, Gao T, Li J, Tao J, Bai Y, Lu R. Motor imagery-based brain-computer interface combined with multimodal feedback to promote upper limb motor function after stroke: a preliminary study. Evid Based Complement Alternat Med 2021;2021:1116126 [FREE Full text] [doi: 10.1155/2021/1116126] [Medline: 34777531]
52. Hao M, Fang Q, Wu B, Liu L, Tang H, Tian F, et al. Rehabilitation effect of intelligent rehabilitation training system on hemiplegic limb spasms after stroke. Open Life Sci 2023;18(1):20220724 [FREE Full text] [doi: 10.1515/biol-2022-0724] [Medline: 37791058]
53. Jinshu Z, Mingming W, Yuan Z, Xuanxiang S. Clinical research on rehabilitation treatment of hemiplegia after stroke by rehabilitation robot based on brain-computer interaction technology. J External Ther Tradit Chin Med 2021;30(3):3-5. [doi: 10.3969/j.issn.1006-978X.2021.03.001]
54. Ying X, Yanyun J, Jie J, Xiaomei W. Efficacy observation of brain - computer interface combined with functional electrical stimulation training on upper limb function and cognition in elderly stroke patients. Chin J Geriatr Heart Brain Vessel Dis 2018;20(9):988-990. [doi: 10.3969/j.issn.1009-0126.2018.09.023]
55. Wang X, Wong W, Sun R, Chu WC, Tong K. Differentiated effects of robot hand training with and without neural guidance on neuroplasticity patterns in chronic stroke. Front Neurol 2018;9:810 [FREE Full text] [doi: 10.3389/fneur.2018.00810] [Medline: 30349505]
56. Ren C, Li X, Gao Q, Pan M, Wang J, Yang F, et al. The effect of brain-computer interface controlled functional electrical stimulation training on rehabilitation of upper limb after stroke: a systematic review and meta-analysis. Front Hum Neurosci 2024;18:1438095 [FREE Full text] [doi: 10.3389/fnhum.2024.1438095] [Medline: 39391265]
57. Hung J, Wu W, Chen Y, Pong Y, Chang K. Predictors of clinically important improvements in motor function and daily use of affected arm after a botulinum toxin a injection in patients with chronic stroke. Toxins (Basel) 2021;14(1):13 [FREE Full text] [doi: 10.3390/toxins14010013] [Medline: 35050990]
58. Lazari A, Salvan P, Cottaar M, Papp D, Rushworth MFS, Johansen-Berg H. Hebbian activity-dependent plasticity in white matter. Cell Rep 2022;39(11):110951 [FREE Full text] [doi: 10.1016/j.celrep.2022.110951] [Medline: 35705046]
59. Veerbeek JM, Langbroek-Amersfoort AC, van Wegen EEH, Meskers CGM, Kwakkel G. Effects of robot-assisted therapy for the upper limb after stroke. Neurorehabil Neural Repair 2017;31(2):107-121. [doi: 10.1177/1545968316666957] [Medline: 27597165]
60. Kruse A, Suica Z, Taeymans J, Schuster-Amft C. Effect of brain-computer interface training based on non-invasive electroencephalography using motor imagery on functional recovery after stroke - a systematic review and meta-analysis. BMC Neurol 2020;20(1):385 [FREE Full text] [doi: 10.1186/s12883-020-01960-5] [Medline: 33092554]
61. Park H, Kim S, Winstein CJ, Gordon J, Schweighofer N. Short-duration and intensive training improves long-term reaching performance in individuals with chronic stroke. Neurorehabil Neural Repair 2016;30(6):551-561 [FREE Full text] [doi: 10.1177/1545968315606990] [Medline: 26405046]

Abbreviations

ADL: activities of daily living
ARAT: Action Research Arm Test
BCI: brain-computer interface
EEG: electroencephalography
FES: functional electrical stimulation
FMA-UE: Fugl-Meyer Assessment for upper extremity
fNIRS: functional Near-Infrared Spectroscopy
GRADE: Grading of Recommendations Assessment, Development and Evaluation
MAL: Motor Activity Log
MAS: Modified Ashworth Scale
MBI: Modified Barthel Index
MD: mean difference
MeSH: Medical Subject Headings
PI: prediction interval
PRESS: Peer Review of Electronic Search Strategies

XSL•FO
RenderX

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses – Search Extension
PROSPERO: International Prospective Register of Systematic Reviews
RCT: randomized controlled trial
ROB: Risk of Bias tool
RRT: routine rehabilitation therapy

Edited by S Brini; submitted 16.Jun.2025; peer-reviewed by J Liang, Y Fan; comments to author 14.Oct.2025; revised version received 30.Nov.2025; accepted 04.Dec.2025; published 28.Jan.2026.

Please cite as:
Chen H, Yun G
Efficacy of Brain-Computer Interface Therapy for Upper Limb Rehabilitation in Chronic Stroke: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e79132
URL: https://www.jmir.org/2026/1/e79132
doi:10.2196/79132
PMID:

©HongJie Chen, GuoJun Yun. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Behavioral Determinants and Effectiveness of Digital Behavior Change Interventions for the Prevention of Sexually Transmitted Infections and HIV: Overview of Systematic Reviews

Giuliano Duarte-Anselmi[1,2*], MPH; Susana Sanduvete-Chaves[3*], PhD; Salvador Chacón-Moscoso[3,4*], PhD; Daniel López-Arenas[3*], MPR

[1]Facultad de Psicología, Universitat de Barcelona, Barcelona, Spain

[2]Facultad de Ciencias Médicas, Universidad de Santiago de Chile, Santiago, Chile

[3]Departamento de Psicología Experimental, Facultad de Psicología, Universidad de Sevilla, Sevilla, Spain

[4]Universidad Autónoma de Chile, Santiago, Chile

[*]all authors contributed equally

Corresponding Author:
Susana Sanduvete-Chaves, PhD
Departamento de Psicología Experimental
Facultad de Psicología
Universidad de Sevilla
Campus Ramón y Cajal. C/ Camilo José Cela, s/n
Sevilla, 41018
Spain
Phone: 34 954557676
Email: sussancha@us.es

Abstract

Background: Unsafe sexual practices remain a major contributor to global morbidity, premature mortality, and health care burden. More than 1 million people acquire a sexually transmitted infection (STI) daily, including HIV. Although biomedical innovations such as pre-exposure prophylaxis have expanded prevention options, consistent condom use and regular HIV and STI testing remain essential behavioral strategies. Adherence to these behaviors remains uneven, underscoring the need for complementary digital and behavioral approaches. Digital behavior change interventions (DBCIs), technology-based programs designed to support health-related behavior change, offer scalable and personalized tools for safer-sex promotion. However, evidence regarding their behavioral components and effectiveness remains fragmented across systematic reviews (SRs).

Objective: This study aims to synthesize and critically appraise evidence on the effectiveness of DBCIs for preventing STIs and HIV, and to identify which behavior change techniques (BCTs) and theoretical domains framework (TDF) have been used to improve safe-sex behaviors.

Methods: A search was conducted in MEDLINE, Cochrane Database of SRs, Epistemonikos, and PsycINFO for all publications up to November 12, 2025, without language or date restrictions. Eligible SRs examined DBCIs targeting STI and HIV prevention or reduction of risky sexual behaviors. Two reviewers (GDA and DLA) independently screened, extracted data, and appraised methodological quality using the AMSTAR-2 tool. The reporting followed the PRIOR (Preferred Reporting Items for Overviews of Reviews) and PRISMA-S (Preferred Reporting Items for SRs and Meta-Analyses Literature Search Extension) recommendations.

Results: Overall, 23 SRs, comprising 514 primary studies and 129,481 participants, met the inclusion criteria. Most interventions were SMS-based, mobile app–based, or web-delivered. Digital interventions consistently improved STI and HIV testing uptake and engagement with sexual health services. Evidence for condom use and biological outcomes was mixed. Improvements in cognitive determinants, such as HIV-related knowledge, motivation, and self-efficacy, were frequently reported. Only 4 reviews explicitly applied BCT or TDF taxonomies, identifying goal setting, feedback on behavior, and prompts and cues as commonly used techniques. Research predominantly originated from high-income settings, with limited evidence from low- and middle-income countries and minimal reporting of sex- or gender-disaggregated outcomes.

Conclusions: DBCIs show promise for strengthening STI/HIV prevention, particularly by increasing testing behaviors and supporting cognitive determinants of risk reduction. However, sustained condom use and biological outcomes remain inconsistent,

XSL•FO
RenderX

and reporting of behavioral mechanisms is limited. This overview is the first to integrate effectiveness evidence with a systematic, mechanism-focused mapping of BCTs and TDF constructs, providing an innovation not present in earlier reviews. Clarifying which active components of digital interventions are most consistently linked to beneficial outcomes offers concrete guidance for designing culturally tailored, theory-driven, and equity-focused digital strategies. These insights have direct implications for researchers, clinicians, and policymakers seeking to develop digital prevention programs that more effectively address behavioral determinants of STI and HIV risk.

Trial Registration: PROSPERO CRD42023485887; https://www.crd.york.ac.uk/PROSPERO/view/CRD42023485887

International Registered Report Identifier (IRRID): RR2-10.5867/medwave.2025.02.3020

(J Med Internet Res 2026;28:e74201) doi:10.2196/74201

KEYWORDS

behavioral change; behavioral design; sexually transmitted diseases; HIV; digital behavior change intervention (DBCI)

Introduction

Unsafe sexual practices are major contributors to global morbidity and premature mortality, representing one of the leading behavioral risk factors worldwide [1,2]. Among young people, these practices significantly increase the risk of sexually transmitted infections (STIs) including syphilis, gonorrhea, chlamydia, and HIV, as well as human papillomavirus (HPV)–related cancers [3,4]. A recent report by the World Health Organization highlights an alarming decrease in condom use among adolescents in Europe, leading to higher rates of unprotected sex and, consequently, an increased risk of STIs, particularly among adolescents from low-income families [5]. Although biomedical innovations such as pre-exposure prophylaxis (PrEP) have expanded prevention options, consistent condom use and regular HIV and STI testing remain essential behavioral strategies for reducing infection risk [6]. Together, these measures form a complementary prevention framework; yet, adherence remains uneven across populations. HIV and AIDS continues to be a leading cause of death globally, with more than 1 million people acquiring an STI daily and nearly 39 million living with HIV [7,8]. Behavioral determinants, including motivation, self-regulation, risk perception, social and cultural norms, and structural barriers, shape whether individuals engage in STI and HIV prevention behaviors, yet they are often insufficiently addressed or poorly defined in existing prevention strategies.

Given the scale of these challenges, effective prevention strategies must emphasize sociocultural and behavioral changes, such as increasing awareness, reducing stigma, and promoting safe sex practices like consistent condom use and regular STI and HIV testing [9-11]. Widespread access to the Internet and mobile phones presents a unique opportunity to leverage digital interventions as private and effective methods for improving sexual health, particularly in regions with varying levels of literacy [11-14].

Recent evidence underscores the growing use of digital technologies in HIV and STI prevention. A 2024 umbrella review found that eHealth interventions, ranging from mobile apps and websites to telemedicine and social media programs, were generally effective in supporting HIV prevention, testing, and clinical management, although the methodological quality of many reviews was low [15]. However, the Shi et al [15] review included both prevention and treatment interventions, whereas the present overview focuses exclusively on preventive strategies and their behavioral mechanisms. Evidence also suggests that the inclusion of behavior change techniques (BCTs) in digital tools enhances user engagement and intervention effectiveness [16]. Additionally, recent work has highlighted the expanding role of interactive digital tools in partner notification and sexual health engagement [17]. Together, these findings demonstrate the rapid evolution of digital health approaches and emphasize the need to systematically map their behavioral components to guide the design of effective prevention programs. However, most existing systematic reviews (SRs) and umbrella reviews describe the effects of digital interventions without examining their mechanisms of action, that is, the specific theoretical pathways and BCTs through which interventions influence prevention behaviors. Without identifying these mechanisms of action, it is difficult to understand why some digital programs succeed while others do not, and which components should be replicated or scaled.

Digital interventions can be broadly defined as health-promoting programs delivered through digital platforms such as websites, mobile apps, text messaging, or social media [18]. Within this broad category, digital behavior change interventions (DBCIs) are those that explicitly incorporate theoretical frameworks and structured BCTs to influence health-related behaviors [19]. In other words, while all DBCIs are digital interventions, not all digital interventions qualify as DBCIs. This conceptual distinction underpins our search strategy and synthesis approach, focusing on interventions that use digital delivery to achieve behavioral outcomes through identifiable active ingredients. These definitions are provided upfront to reduce conceptual ambiguity, as emphasized by recent critiques in digital behavior change research.

DBCIs offer multiple advantages over traditional prevention approaches: they can deliver tailored, interactive, and adaptive content; are cost-effective and scalable; and can integrate technological features such as automated feedback and passive sensing [20-22]. However, information alone is insufficient to drive behavior change: integrating BCTs to these digital platforms is essential for achieving meaningful health outcomes [23,24].

XSL•FO
RenderX

In this context, the theoretical domains framework (TDF) is presented as a widely used proposal to systematically identify barriers and facilitators to change specific behaviors, helping design more effective interventions in health, among other fields. It is useful to understand why people do or do not do something, and highlight factors needing intervention [25,26]. It synthesizes 33 behavior change theories into 14 core domains, such as cognitive (knowledge, skills, beliefs about capabilities and consequences), affective and emotional (emotions and reinforcement), social and environmental (social and professional role, social influences, environmental context, and resources), and beliefs and intentions (optimism, intentions, goals, memory, attention, decision processes, behavioral regulation) [27].

Given the volume of SRs assessing digital interventions for STI and HIV prevention, an overview of reviews enables synthesis and comparison across multiple bodies of evidence rather than relying on a single set of primary studies. This approach provides a broader understanding of intervention effectiveness, identifying patterns, strengths, methodological gaps, and avenues for future research [28,29]. Despite their promise, many digital interventions lack clear descriptions of the BCTs and theoretical domains they use, hindering replicability and practical translation [30,31]. Identifying the most effective BCTs, especially those that successfully promote safe sex and reduce STI and HIV transmission, is crucial for public health initiatives [32-35]. To date, no overview has systematically integrated BCTs, theoretical domains, and prevention outcomes to produce a mechanism-focused synthesis of digital interventions for STI and HIV prevention. Existing reviews also provide limited up-to-date evidence and do not incorporate studies published through 2025. Addressing this gap is essential for identifying which behavioral components drive meaningful changes in prevention behaviors and for informing the development of digital strategies that are theoretically grounded, culturally responsive, and scalable.

Accordingly, this overview aims to synthesize current evidence on the use of BCTs and the TDF in digital interventions designed to prevent STIs and HIV. By examining how these behavioral components are implemented and how they influence prevention-related outcomes, this research seeks to inform the development of more effective, theory-driven digital interventions and strengthen future public health strategies. This overview therefore provides not only an updated assessment of the evidence but also a behavioral mapping that has been largely absent from previous syntheses. It further advances the field by incorporating SRs published through 2025, which have not yet been integrated in any prior synthesis.

Methods

Study Design

This study is an overview of SRs and adheres to the Cochrane Handbook for SRs of Interventions [29] and the PRIOR (Preferred Reporting Items for Overviews of Reviews) statement [36]. The PRIOR checklist is reported in Multimedia Appendix 1 and the Sex and Gender Equity in Research (SAGER) guidelines [37], in Multimedia Appendix 2 (section 2). In addition, the search strategy and reporting follow the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) [38] to ensure full transparency and reproducibility of search methods (see Multimedia Appendix 3).

Protocol and Registration

The protocol for this overview was prospectively registered in PROSPERO (CRD42023485887) on December 5, 2023 (Multimedia Appendix 2 section 3), and later published in full [39]. The methods adhered to the predefined protocol and incorporated SAGER guidance where applicable. Reporting also followed PRISMA 2020 recommendations for SRs and overviews [40]. The full protocol can be downloaded from the Open Science Framework [41,42].

Patient and Public Involvement

Neither patients nor the public were involved in designing or conducting this study. Therefore, no ethical approval was required for this overview. The analyzed data were open access.

Eligibility Criteria

The eligibility criteria were reported in detail in our protocol [39]. The inclusion criteria for this overview were based on the population, intervention, comparison, outcome, study type (PICOS) framework (Table 1).

XSL•FO
RenderX

Table 1. Eligibility criteria for elements of a comprehensive search strategy.

Element	Inclusion and exclusion criteria
Population	• Included: SRs[a] that have evaluated the effect of digital behavior change interventions in any population and that have described BCTs[b], mechanisms of action, or any behavioral model or framework that takes into account how the digital intervention influences the behavior change used to reduce the risk or prevent the transmission of STIs[b], including HIV. • Excluded: studies that did not focus on prevention or that focused on treatment and adherence to antiretroviral therapy and self-care of people living with HIV were excluded, as the focus of the question research is risk reduction and prevention of STIs and HIV.
Intervention	• Included: SRs that evaluated digital and mobile health behavior change interventions focusing on modifying unsafe sexual behaviors or preventing STIs and HIV, that is, interventions carried out using a digital or mobile platform as a direct interface with the participants. • Excluded: studies that did not report the use of digital intervention, that is, did not incorporate digital technology such as smartphones, computers, tablets, multimedia, and social networks.
Comparator	• As this was an overview of SRs, a comparator or control group was not an inclusion criterion for this study. However, given our interest in effectiveness of interventions and BCTs, we included reviews where evidence from primary experimental studies with an appropriate comparator was available.
Outcome	• Included: during the selection process, we do not consider concrete results as an inclusion criterion. All studies that assessed short- or long-term behavior change concerning the following primary outcomes were included: reduction in risky sexual behaviors, such as condom use (last sexual encounter, frequency, consistency) and increased STI and HIV testing; and prevention (vaccination against HPV[d] and hepatitis A and B, and HIV pre-exposure prophylaxis). • As secondary outcomes, the use of behavior change theories or techniques was analyzed, using standardized classifications if available, such as the taxonomy of BCTs. • Excluded: any study that did not include a primary and/or secondary outcome.
Study design	• Included: SRs only • Excluded: studies other than SRs (eg, primary studies, commentary articles, and conferences) were excluded.

[a]SR: systematic review.

[b]BCT: behavior change technique.

[c]STI: sexually transmitted infection.

[d]HPV: human papillomavirus.

Summary of inclusion and exclusion criteria used to identify SRs of digital behavior change interventions for the prevention of STIs and HIV. Criteria were defined following the Cochrane Handbook and the PRIOR guideline. All SRs were operationally defined as secondary research including primary clinical studies with explicit search strategies in ≥2 databases [29,43].

Information Sources

In this overview, a comprehensive search strategy was used, leveraging multiple key sources. Primary searches were conducted on the leading international databases MEDLINE (via PubMed), the Cochrane Database of SRs, Epistemonikos, and PsycINFO. Each database was searched independently using its native platform; no multidatabase platform (eg, EBSCOhost and Ovid) was used. To enhance the scope of our search, we conducted supplementary searches to identify any studies potentially overlooked by the primary search strategy or absent from the indexed databases. These supplementary efforts included (1) meticulous manual reviews of the references cited in the included studies, (2) examination of related SRs that shared at least one study with the reviews included in our study, and (3) additional records identified through various websites (n=7), organizations (n=1), citation searching (n=321), references provided by authors (n=1), and consultations with experts (n=28; Figure 1). Citation searching involved both backward citation screening (reference list checking) and forward citation tracking using Google Scholar. No additional search methods such as automated alerts, web-scraping tools, or application programming interface–based search retrieval were used.

XSL•FO
RenderX

Figure 1. Study selection flow diagram for the overview of systematic reviews on digital behavior change interventions (DBCIs) for STI/HIV prevention, following the PRIOR reporting guideline. PRIOR: Preferred Reporting Items for Overviews of Reviews.

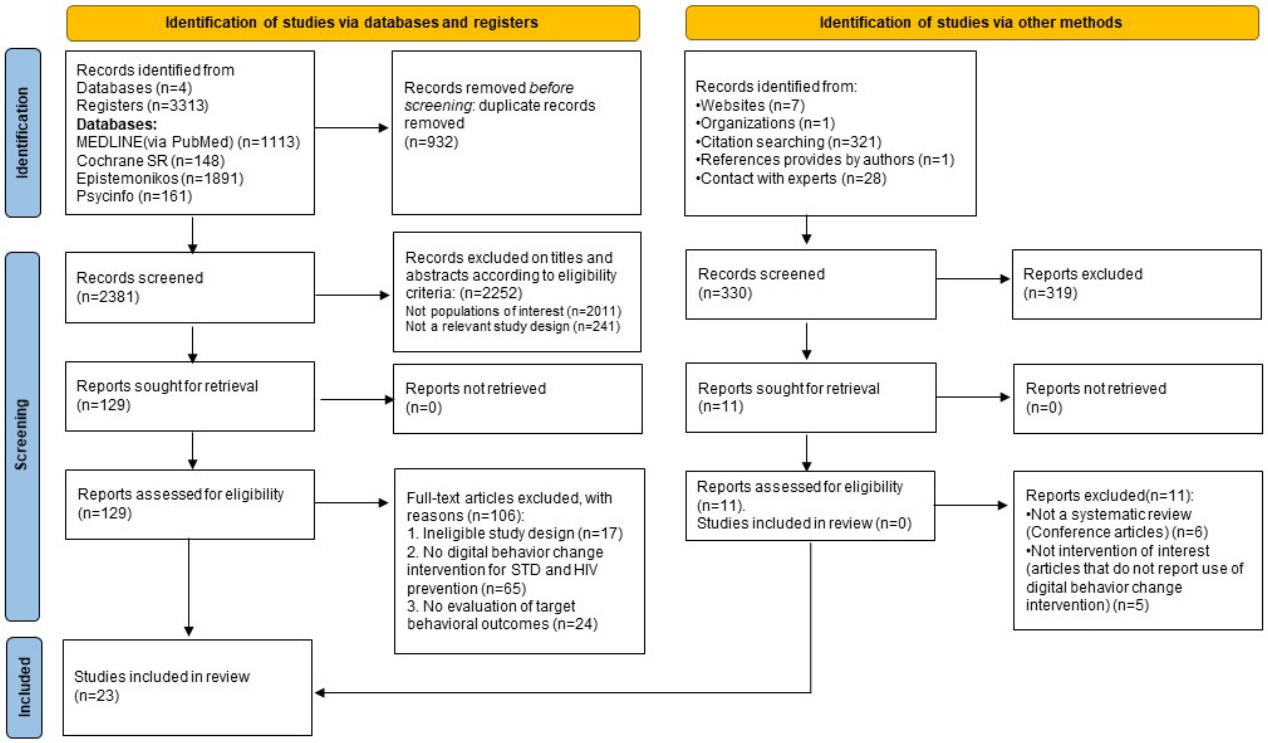

Since this overview synthesizes published SRs rather than primary studies, study and trial registries (eg, ClinicalTrials.gov, International Clinical Trials Registry Platform) were not searched. Each source was rigorously scrutinized, with the date of the last search or consultation carefully documented to ensure the currency and relevance of the findings. Database-specific yields were MEDLINE (via PubMed; n=1113), Cochrane Library (n=148), PsycINFO (n=161), and Epistemonikos (n=1891), as detailed in Multimedia Appendix 2. Detailed methodologies and search strategies are available in Multimedia Appendix 2 (section 4).

Search Strategy

The electronic search strategy (sections 5 and 6 in Multimedia Appendix 2) was developed and conducted under the supervision of an experienced librarian. The strategy did not undergo formal PRESS (Peer Review of Electronic Search Strategies) peer review. The search strategy was newly developed for this overview and was not adapted from any previous review. The first author performed the electronic search from the databases' inception up to November 12, 2025, with no restrictions on publication date, language, or country of origin, following PRISMA-S recommendations to ensure transparent and reproducible reporting of search methods. No methodological search filters (eg, SR filters, randomized trial filters, and human-only filters) were applied beyond the predefined eligibility criteria. An earlier search completed on August 31, 2024, was rerun and updated in accordance with PRISMA-S guidance [38], to capture the most recent SRs before resubmission. Additionally, we manually searched the bibliographies of relevant reviews and the articles initially retrieved. Letters were also sent to authors and experts identified in the included and excluded studies during the screening stage to identify additional eligible studies. The eligibility criteria are listed in Table 1. Full search dates for each database, including the initial search (August 31, 2024) and rerun (November 12, 2025), are reported in Multimedia Appendix 2. Full search strategies for all databases, copied verbatim as executed, are provided in Multimedia Appendix 2.

Study Selection Process

Following deduplication and a pilot test of the inclusion and exclusion criteria, 2 independent reviewers (GDA and DLA) screened all titles, abstracts, and full-text articles for eligibility without knowledge of each other's decisions. For the deduplication process, all records retrieved from database searches and supplementary sources were imported into Collaboratron (Epistemonikos). The software's automated similarity-detection algorithm was used to identify duplicate entries, followed by manual verification by 2 independent reviewers (GDA and DLA) to ensure accuracy. This hybrid deduplication approach, combining automated and manual procedures, aligns with PRISMA-S recommendations for transparent management of search records. Records from electronic and bibliographic searches were stored and full text screening was conducted using Collaboratron by Epistemonikos [44]. Differences between the 2 reviewers (GDA and DLA) were resolved through discussion, and a third reviewer (SSC) was consulted when necessary. The list of studies excluded after the full text review, along with reasons for exclusions, is detailed in section 7 in Multimedia Appendix 2.

We reached out to 28 authors (up to 3 email attempts) to request additional information, particularly on gray literature such as conference presentations and reports (see the list of experts in

XSL•FO
RenderX

the Contacting Experts section of section 4 in Multimedia Appendix 2).

The interrater reliability was assessed using Cohen kappa coefficient [45]. Two independent reviewers (GDA and DLA) evaluated the full-texts (section 8 in Multimedia Appendix 2).

Data Collection Process

We developed a data extraction tool in Microsoft Excel to obtain various study data recommended by Cochrane [46,47]. The data extracted from the SRs selected for the study was tested and calibrated by the team (section 9 in Multimedia Appendix 2). For this purpose, 1 author (GDA) created the spreadsheet and then extracted the data from 1 SR. Subsequently, 2 authors (GDA and DLA) independently extracted data from 3 SRs, and all authors provided feedback on whether the data elements were complete, and the extracted data were unambiguous. Once a consensus was reached through discussion, 1 author (GDA) created a data extraction manual for the spreadsheet, which can be found in Multimedia Appendix 2 (section 10).

Data Items

Data items included SR characteristics, PICOS criteria, and variables related to DBCI (Textbox 1).

XSL•FO
RenderX

Textbox 1. Data items in this overview of systematic reviews.

Data items

- Bibliographic information (author, year of publication, title, and aim of the SR).

Population characteristics:

- Participants (total number of participants included in the studies)
- Population age
- Specific population (men who have sex with men, lesbian, gay, bisexual, transgender, queer, and other sexual orientations and gender identities (LGBTQI+) individuals, people with a diagnosis of sexually transmitted disease [STD] without HIV)

Study characteristics

- Total number of studies included in the review
- Number of randomized controlled trials (RCTs) included
- Type of studies (only RCTs, only non-RCTs, including experimental nonrandomized and observational studies-, both RCTs and non-RCTs)
- Review period (range of publication years of the primary studies included in each systematic review, reflecting the temporal coverage of the synthesized evidence)
- Period or specific date range of the literature search (years during which databases were searched by the authors of each SR);
- Country or geographic location of the studies

Intervention details

- Target population
- Target behavior
- Explicit mention or extraction of theoretical frameworks used.
- Behavioral outcomes (condom use, frequency of unprotected sexual intercourse, number of sexual partners, STD and HIV testing, uptake of medical male circumcision, HIV counseling, vaccination)
- Cognitive outcomes (self-efficacy, STD and HIV-related knowledge, attitudes toward condom use, and pre-exposure prophylaxis awareness)
- Biological outcomes (acquisition of HIV or sexually transmitted infection)

Intervention acceptability and feasibility

- Acceptability: participants' acceptance of the intervention
- Practicability: ease of implementation in the real world
- Effectiveness: achievement of the intervention objectives
- Affordability: cost-effectiveness of the intervention
- Spill-over effects: unintended consequences
- Equity: impact on health equity

Technology and delivery methods

- Mobile devices, desktop computers, digital billboards, wearable accessories, digital objects, and projection and holograms
- Mode of delivery of digital contents (audio calls and messages, video calls and messages, text and instant messages, emails, audio broadcasts and podcasts, websites and computer programs and apps, eBooks, virtual or augmented reality, artificial intelligence; for example, use of artificial intelligence–based chatbots to promote safe sex or other sexual behaviors)

Other characteristics

- Number of primary studies included and overlap, tools used to assess the risk of bias in primary studies, whether a meta-analysis was conducted, and certainty of evidence (Grading of Recommendations, Assessment, Development, and Evaluation)

Quality Appraisal of the SRs

We performed critical appraisals of SRs using AMSTAR-2 [48], as outlined in our published protocol [39]. AMSTAR-2 consists of 16 items that assess the thoroughness of various aspects of a SR, such as the preparation process, literature search, study selection, data extraction, and analysis, as well as potential biases (eg, risk of bias, publication bias, or funding sources). Based on the type and number of weaknesses identified (ie, unmet items), the reviews were assigned a confidence rating:

XSL•FO
RenderX

high, moderate, low, or critically low. Two authors (GDA and DLA) independently assessed all SRs using a spreadsheet (Microsoft Excel 2010) and reached consensus through discussion (section 11 in Multimedia Appendix 2).

Overlap in Primary Studies Included in Reviews

To ensure the accuracy of the primary study outcome data and avoid overlap, we checked whether the included SRs shared overlapping primary studies. This was done by creating a citation matrix and calculating the overall corrected covered area (CCA) using Graphical Representation of Overlap for Overviews (GROOVE) [49]. The CCA quantifies the degree of overlap between primary studies included across SRs, calculated as the number of repeated primary studies (numerator) divided by the product of the total number of unique primary studies and the number of reviews, minus the total number of unique studies (denominator). The resulting value represents the proportion of shared evidence between reviews. A CCA of 0% to 5% indicates a slight overlap, 6% to 10% a moderate overlap, 11% to 15% a high overlap, and greater than 15% a very high overlap."

Data Synthesis Methods

Overview

Across the included SRs, the types of control or comparison groups varied substantially. In most cases, DBCIs were compared against nondigital or usual-care conditions, such as standard health education, printed materials, or no intervention controls. A smaller number of reviews included comparators that were themselves digital but lacked explicit behavior change components (eg, informational websites or SMS reminders without BCTs). This heterogeneity makes it difficult to disentangle whether observed effects are attributable to the digital delivery mode, the behavioral content, or both. Therefore, comparator conditions reported by each review were documented, and findings were interpreted with caution, emphasizing the combined influence of digital and behavioral mechanisms [50].

Outcome Definitions

Behavioral outcomes were classified according to the definitions provided in the included SRs. "STI and HIV prevention behaviors" encompassed all behavioral actions aimed at reducing infection risk, including but not limited to condom use, STI and HIV testing, vaccination, and adherence to treatment. "Safe sex behaviors" referred specifically to sexual practices such as consistent condom use, partner reduction, and negotiation of safer sex. This hierarchical approach was adopted to maintain consistency with the terminology used in the original reviews while avoiding redundancy between overlapping categories. A clarifying note was also added to Table 2 indicating that "safe sex behaviors" represent a subset of "STI and HIV prevention behaviors.

XSL•FO
RenderX

Table 2. Main characteristics of systematic reviews published between 2014 and 2015. “Safe sex behaviors” are a subset of “sexually transmitted infection and HIV prevention behaviors. The table summarizes study design, population, intervention features, behavioral and cognitive outcomes, and country income level based on World Bank classification.

Characteristics (N=23[a])	Results
SR[b] design	
Number of studies included	514[c]
Participants (total)	129,481
Range of years of the studies included	(2014-2025)
Place (country: geographic location)	
High-income countries	204
Upper-middle-income countries	22
Lower-middle-income countries	17
Only RCT[d] (n=23)	9
Only non-RCT (includes nonrandomized experimental and observational studies) (n=23)	2
RCT and non-RCT (n=23)	12
Total RCTs (N=514)	410
Population type (n=23)	
Adolescents (population age 10 to 19 years)	14
Youth (20 to 29 years)	17
Adults (>29 years)	15
Men who have sex with men	14
LGBTQI+[e] individuals	10
People diagnosed with an STD[f] (without HIV)	10
Intervention characteristics (n=23)	
Target of behavior (prevention of STI[g] and HIV)	19
Target behavior (safe sex)	16
Target behavior (STI and HIV testing)	17
Target behavior (treatment-related, ie, attend the appointment or getting treatment for an STI)	10
Environment: schools	10
Environment: universities (where it was implemented)	4
Environment: health centers (where it was implemented)	10
Explicit framework of theoretical domains of behavior change from the studies included in the SR	10
Description of behavior change techniques according to BCTTv1	2
Description according to Theoretical Domains Framework (TDF)	1
Behavioral outcomes (n=23)	
Condom use (internal or external)	19
Frequency in unprotected sexual intercourse	11
Number of sexual partners	7
STD and HIV testing	18
Uptake of medical male circumcision	4
HIV counseling	8
Get vaccinated against (VPH- HEP A y B)	7
Cognitive outcomes (mediators of prevention) (n=23)	

XSL•FO
RenderX

Characteristics (N=23[a])	Results
Self-efficacy	10
STD- and HIV-related knowledge	10
Attitudes toward condom use	8
Pre-exposure prophylaxis awareness	4
Biological outcomes	
HIV or STI acquisition (n=23)	8
Technology delivered (n=23)	
Mobile device	19
Desktop computer	12
Digital billboard	0
Wearable (clothing and accessory)	0
Digital object	0
Projection and hologram	0
Digital content type (n=23)	
Audio call and message	10
Video call and message	14
Text and instant message	18
Email	9
Video game	6
Audio broadcast and podcast	1
Website, computer, program and app	18
eBook	1
Virtual or augmented reality	2
Other descriptions	
Artificial intelligence-based chatbots for promoting safe sex or other sexual behaviors	0
Meta-analysis and certainty of evidence (n=23)	
Meta-analysis	7
Certainty of evidence (GRADE)[h]	4
APEASE[i] (n=23)	
Acceptability	6
Practicability	0
Effectiveness	19
Affordability	4
Spill-over effects	3

XSL•FO
RenderX

Characteristics (N=23[a])	Results
Equity	2

[a]Total number of SRs included in this overview.

[b]SR: systematic review.

[c]Total number of primary studies included in the SRs of this overview.

[d]RCT: randomized controlled trial.

[e]LGBTQI+: lesbian, gay, bisexual, transgender, queer, intersex, and plus.

[f]STD: sextually transmitted disease.

[g]STI: sexually transmitted infection.

[h]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.

[i]APEASE: Acceptability, Practicability, Effectiveness, Affordability, Spill-over effects, and Equity.

A written description was generated on the results of each SR on digital interventions for the prevention of STIs and HIV, and a table detailed the characteristics and outcomes of each review. The extracted data were synthesized using a predefined and team-approved template, identifying common themes and mapping them according to the stated objectives (section 9 in Multimedia Appendix 2).

Subgroup analyses were done to evaluate various factors, including the aim of the SR, the target population, behavioral outcomes, cognitive outcomes, biological outcomes, and the type of digital content and intervention. The mode of delivery (MoD) framework by Marques et al [51] was used to categorize interventions into types such as text and instant messages, video calls and messages, websites, computer programs and apps, emails, video games, audio broadcasts and podcasts, e-books, and virtual or augmented reality.

The effectiveness of digital interventions for STI and HIV prevention was assessed based on SRs that explicitly described the use of BCTs or the TDF. Subsequently, the impact of BCTs on each specific outcome was analyzed following the approach of Michie et al [52,53], allowing for the identification of the most effective techniques. Findings were synthesized into structured tables to visualize the impact of digital interventions across behavioral, cognitive, and biological domains.

Finally, a 5-level classification system was applied: "▼" indicated strong evidence of a negative effect; "O," mixed or null evidence; "O−" and "O+," a negative or positive effect with limited evidence; and "▲," strong evidence of a positive effect. The strength of evidence was determined according to three criteria: (1) the number of SRs reporting consistent results in the same direction, (2) the methodological quality of these reviews based on AMSTAR-2, and (3) the presence of meta-analytic data when available. Strong evidence ("▲" or "▼") required consistent findings across at least 2 high-quality reviews or meta-analyses, while limited evidence ("O−" or "O+") was assigned when findings were reported in only one review or when quality or consistency was lower. Mixed or null evidence ("O") indicated conflicting or inconclusive findings. This approach aligns with recent overviews applying the same evidence-grading framework for BCTs [53,54].

This system aims to provide clear guidance on which BCTs are most effective in digital interventions for STI and HIV prevention. The BCTs were coded according to the BCT Taxonomy version 1 (BCTTv1) [31,55], in studies that reported interventions using TDF, while the original descriptions were retained for studies that explicitly reported BCTs. This methodological approach ensured that conclusions were systematic, evidence-based, and aligned with established behavior change frameworks.

The SAGER [37] guidelines were used to ensure the consideration of sex and gender variables during the data extraction. These guidelines aim to prevent bias and improve the relevance and validity of findings by promoting the clear distinction between "sex" (biological differences) and "gender" (social and cultural factors), and its purpose is for these distinctions to be accurately reflected in study design, data analysis, and results reporting (sections 2 and 12 in Multimedia Appendix 2).

Results

Overview

From the 3643 records identified (3313 from electronic databases and 330 from bibliographic sources), 129 full-text articles were assessed for eligibility. Of these, 106 were excluded, leading to the inclusion of 23 SRs in this overview [56-78] (Figure 1). The interrater reliability for full-text screening of the initial 122 studies was robust (κ=0.88). An updated search conducted on November 12 yielded 378 additional records, of which 7 were assessed in full-text and 4 met the inclusion criteria. Additional studies were identified through supplementary methods, including website searches, citation searching, and expert consultations, ensuring a comprehensive review (section 4 in in Multimedia Appendix 2).

Characteristics of SRs

The 23 SRs included in the overview were published between 2014 and 2025, and covered 514 primary studies, including 410 randomized controlled trials (RCTs). Of these reviews, 9 focused exclusively on RCTs [56-58,66,69,70,75,77,78], 2 on non-RCTs [72,73], and 12 included both study types [59-65,67,68,71,74,76]. These reviews encompassed 129,481 participants, with individual studies ranging from 2662 to 27,704 participants. A detailed appraisal of methodological quality and overlap across the included SRs is presented at the end of this section to contextualize confidence in the synthesized evidence. The targeted populations included adolescents (n=14), youth

XSL•FO
RenderX

(n=17), adults (n=15), men who have sex with men (n=14), LGBTQ+ individuals (n=10), and people diagnosed with sexually transmitted diseases other than HIV (n=10; sections 13, 14a, and 14b in Multimedia Appendix 2).

Most studies were conducted in high-income countries, including the United States, Portugal, and Chile, followed by upper-middle-income countries such as China and South Africa, and lower-middle–income countries such as India and Kenya. Country income levels were classified according to the World Bank Country and Lending Groups (FY2025, Atlas method, USD), which use Gross National Income per capita as the defining criterion. Classifications were verified as of November 10, 2025, based on the latest publicly available dataset [79]. All data are reported, and detailed characteristics of the included SRs are presented in section 13 in in Multimedia Appendix 2. The synthesis of study characteristics is summarized in Table 2.

SAGER Application

The SAGER guidelines revealed that most studies did not include sex- or gender-disaggregated data, and significant gender differences were generally not reported. Only one study, Kamitani et al [63], mentioned transgender participants, but without a detailed analysis (section 12 in in Multimedia Appendix 2).

Characteristics of the Interventions

Overview

Across the included SRs, digital interventions targeted multiple prevention-related behaviors, including STI and HIV prevention, safer sex practices, and engagement with testing and sexual health services. Most SRs focused on STI and HIV prevention behaviors (19/23, 82.6%) safe-sex promotion (16/23, 69.6%), and STI and HIV testing (17/23, 73.9%). Additionally, 10 SRs (43.5%) incorporated strategies to enhance treatment adherence, such as attending medical appointments or completing syphilis treatment (Table 2 and section 15a in in Multimedia Appendix 2).

The research study settings varied, with health centers being the most common (10/23, 43.5%), followed by schools (10/23, 43.5%), while universities (4/23, 17.4%) were the least frequent. Regarding the use of theoretical frameworks, 10 SRs (43.5%) applied a behavior change framework, yet only 2 SRs (8.7%) explicitly described techniques based on the BCTTv1, and just one (4.3%) used the TDF (section 15a in in Multimedia Appendix 2).

In terms of behavioral outcomes, out of 23 SRs analyzed, condom use was assessed in 19 reviews (82.6%), while STI and HIV testing was reported in 18 reviews (78.3%). Other relevant outcomes included the frequency of unprotected sexual intercourse (47.8%), the number of sexual partners (30.4%), and vaccination against HPV or hepatitis (30.4%). Finally, the most analyzed cognitive outcomes were self-efficacy (43.5%), STI and HIV-related knowledge (43.5%), and attitudes toward condom use (34.8%), whereas PrEP awareness was examined in only 4 studies (17.4%; section 15b and 15c in in Multimedia Appendix 2).

MoD

According to the MoD classification by Marques et al [51], most interventions were delivered via mobile devices (n=19) and desktop computers (n=12), with text and instant messaging being the most common digital content type (n=18). Other content types included video calls and messages (n=14), emails (n=9), and video games (n=6). Notably, no studies used emerging delivery methods such as digital billboards, wearable accessories, digital objects, or projection and holograms (Table 2; section 15d in in Multimedia Appendix 2).

Theoretical Frameworks and BCTs

Of the 23 SRs analyzed, 10 (43.5%) incorporated behavioral theories, with the information-motivation-behavioral skills model (10 of 71 framework mentions, 14.3%), health belief model, and social cognitive theory being the most common. However, 13 reviews (56.5%) lacked any theoretical framework, reflecting inconsistent application of behavior change science. Only 4 (17.4%) reviews explicitly reported the identification or coding of BCTs or TDF [56,58,59,78] (section 15e in in Multimedia Appendix 2).

Overall, there was limited variability in the reporting of theoretical and behavioral frameworks across the 23 SRs. Explicit descriptions of BCTs or TDF mapping were rare, with most reviews indicating “not reported.” This pattern reflects heterogeneity in reporting practices across digital interventions for STI and HIV prevention and highlights that only a minority of reviews provided systematic or detailed descriptions of behavioral frameworks.

Effectiveness and Implementation of BCTs in Digital Interventions: Subset Analysis

This subsection focuses specifically on the subset of SRs (4 out of 23) that explicitly identified, coded, or analyzed BCTs within digital interventions for STI and HIV prevention. These reviews (Bailey et al [56]; Burns et al [58]; Clarke et al [59]; and Mo et al [78]) provided sufficient methodological detail to enable comparison of BCT use, frequency, and effectiveness. The remaining reviews, which did not report BCT coding or implementation frameworks, are synthesized in the previous sections that address broader behavioral, cognitive, and biological outcomes. This clarification ensures transparency and maintains consistency with the overview’s comprehensive scope.

DBCIs for STI and HIV prevention have incorporated various BCTs; however, their explicit classification using standardized frameworks such as the BCTTv1 or TDF remains limited. Among the studies analyzed, Burns et al [58] and Clarke et al [59] reported interventions explicitly coded using BCTTv1, whereas Bailey et al [56] used domains associated with TDF to describe behavioral determinants. Mo et al [78] expanded the evidence base by systematically identifying and mapping BCTs across digital HIV prevention interventions for adolescents and young people (Table 3).

XSL•FO
RenderX

Table 3. Summary of behavioral determinants, frequently reported behavioral change techniques (BCTs), and observed effectiveness of digital interventions for the prevention of sexually transmitted infections (STIs) and HIV based on 4 reviews (Bailey et al [56]; Burns et al [58]; Clarke et al [59]; and Mo et al [78]). Includes intervention design, target population, mode of delivery, and AMSTAR-2 quality rating. BCT codes correspond to Behavior Change Technique Taxonomy version 1.

Reference and review type	Review period	Mode of delivery	N Primary studies and study designs	Target population and number of participants	Target of behavior	Intervention effectiveness and observed behavior change	Determinants of behavior	Most frequently used BCTs and components	Statistical significance and effect size (*P* value, Cohen *d*, r)	AMSTAR 2 Rating
Bailey et al [56] SR[a] and meta-analysis	Searches: from 2014 to June 2017 Publication of primary studies included: 1991-2017	Web-based programs, mobile apps, online modules	31 studies RCTs[b]	Population: young people, men who have sex with men (MSM), HIV-positive people, at-risk adults, African American women. Total: 11,293 participants - IDI vs. minimal intervention: 10,423 participants - IDI vs. face-to-face intervention: 870 participants	Prevention of HIV and other STIs, promotion of safe sex behaviors, adherence to testing and treatment. Condom use, partner reduction, and safe sex negotiation	Increased HIV-related knowledge (moderate effect) - small improvement in behavioral intention - Positive effect on HIV prevention behaviors - No clear impact on self-efficacy - No significant effect on biological outcomes (STI and HIV acquisition, viral load)	Goals, behavioral regulation, knowledge, emotion, optimism, beliefs about capabilities	Goal setting (1.1), commitment (1.9), feedback on behavior (2.2), biofeedback (2.6), information about antecedents (4.2), re-attribution (4.3), information about health consequences (5.1), salience of consequences (5.2), verbal persuasion (15.1), self-talk (15.4)	HIV-related knowledge: SMD[c]=0.56 (95% CI 0.33-0.80) - HIV prevention self-efficacy: SMD=0.13 (95% CI 0.00-0.27) - HIV prevention intention: SMD=0.16 (95% CI 0.06-0.26) - HIV prevention behaviors: OR[d] 1.28 (95% CI 1.04-1.57) - Biological outcomes (STI and HIV acquisition, viral load): OR 1.48 (95% CI 0.96-2.28), *P*=.08 (not significant)	High No critical flaws One noncritical weakness (no information on funding sources)
Burns et al [58] SR of RCTs	Searches: January 1999 and July 2014 Publication of primary studies included:: 2006-2014	Mobile phone-based interventions, SMS reminders, mobile apps, video messages	10 studies RCTs	Population: General population, at-risk adults, young people, (MSM) Total: 16,773 participants	Promotion of sexual health services uptake, reduction of risky sexual behaviors, reduction of recall bias in self-reported sexual activity	Two trials showed significant increases in clinic attendance with SMS reminders. - One trial improved sexual health knowledge. - No trials showed significant increases in condom use. - One trial found mobile technology acceptable for sexual health data collection	Goals, intentions, behavioral regulation, knowledge, social influences, environmental context and resources, reinforcement	Goal Setting (1.1), feedback on Behavior (2.2), information about health consequences (5.1), demonstration of behavior (6.1), social comparison (6.2), prompts/cues (7.1), material incentives (10.1)	Clinic attendance: SMS reminders significantly increased attendance (RR[e] 0.86, 95% CI 0.74-1.00) - Chlamydia retesting: SMS reminders increased retesting (RR 4.5, 95% CI 1.05-19.22) - HIV testing uptake: no significant effect (RR 0.94, 95% CI 0.81-1.09) - Sexual health knowledge: SMS improved knowledge (RR 1.75, 95% CI 1.11-2.77) - Condom use: no significant changes (RR 0.87, 95% CI 0.62-1.24)	Moderate No critical flaws More than one noncritical weakness (no information on funding sources, publication bias reported but not discussed)

XSL•FO
RenderX

Reference and review type	Review period	Mode of delivery	N Primary studies and study designs	Target population and number of participants	Target of behavior	Intervention effectiveness and observed behavior change	Determinants of behavior	Most frequently used BCTs and components	Statistical significance and effect size (*P* value, Cohen *d*, r)	AMSTAR 2 Rating
Clarke et al [59] SR	Searches: 1 January 2000 to 1 September 2021. Publication of primary studies included:: 2011-2018	Digital interventions including SMS (text messages), social media, and app-based messaging	13 studies (RCTs=5 Non RCTs before-after=8)	Population: Adolescents, youth, adults, MSM, LGBTQI+[f]. Total: not reported	Increasing attendance at scheduled sexual health appointments	Behavioral interventions increased attendance at scheduled sexual health appointments. Text messages were the most frequently used MoD[g]. Some interventions were effective, while others had mixed results	Beliefs about consequences, environmental context and resources, emotion, reinforcement, social influence, optimism	Credible source (9.1), prompts/cues (7.1), social support (3.2), social reward (10.4), self-incentive (10.5), restructuring of the environment (12.1, 12.2), focus on past success (15.3), vicarious consequences (16.3)	Some interventions significantly increased attendance while others had mixed results	Critically Low More than one critical flaw (no justification for excluding individual studies, no consideration of bias when interpreting results) More than one noncritical weakness (study selection not done in duplicate, no information on funding sources)
Mo et al [78] SR	Searches: January 2008 to November 2024. Publication of primary studies included: 2008-2023	Mobile apps, SMS text messaging, web-based modules, computer-based digital game (IYG-Tech), online educational platforms	34 studies (RCTs, quasi-experimental, and observational designs)	Population: adolescents and young people (10-29 y). Total: Not reported	Prevention of HIV; promotion of safer sex practices; HIV testing; risk reduction behaviors	Narrative synthesis indicated consistent improvements in cognitive determinants (HIV-related knowledge, self-efficacy, perceived risk). Small, inconsistent effects were reported for condom use. Some interventions showed increases in HIV testing motivation or intentions, but behavioral outcomes were heterogeneously measured and rarely pooled	Knowledge, beliefs about consequences, behavioral skills, self-efficacy, environmental context and resources, motivation	Information about health consequences (5.1), feedback on behavior (2.2), prompts/cues (7.1), goal setting (1.1), problem solving (1.2), demonstration of behavior (6.1), social support (3.1), self-monitoring (2.3)	No pooled effect sizes reported. Several primary studies demonstrated significant improvements in HIV knowledge and self-efficacy but effects could not be meta-analyzed	Critically Low More than one critical flaw (no justification for excluding individual studies, no consideration of bias when interpreting results) More than one noncritical weakness (study selection not done in duplicate, no information on funding sources)

[a]SR: systematic review.

[b]RCT: randomized controlled trial.

[c]SMD: standardized mean difference.

[d]OR: odds ratio.

[e]RR: relative risk.

[f]LGBTQI+: lesbian, gay, bisexual, transgender, queer, and other sexual orientations and gender identities.

[g]MoD: mode of delivery.

XSL•FO
RenderX

In total, 26 BCTs were reported. The most frequently used in these interventions included goal setting (1.1), feedback on behavior (2.2), and the use of prompts and cues (7.1), commonly delivered through mobile apps, text messaging (SMS), online modules, and digital games. Other approaches included commitment strategies (1.9), biofeedback (2.6), social support (3.1, 3.2), and providing information about health consequences (5.1). These techniques targeted key behavioral determinants such as knowledge, behavioral regulation, social influence, and reinforcement strategies, aiming to enhance self-efficacy, motivation, and risk awareness (Table 3).

Table 3 summarizes behavioral determinants, commonly used BCTs, and intervention effectiveness. Figure 2 synthesizes the impact of specific BCTs across behavioral, cognitive, and biological outcomes. Findings on BCT effectiveness were drawn from the 3 SRs that reported effect estimates (Bailey et al [56]; Burns et al [58]; Clarke et al [59]). Mo et al [78] contributed descriptive evidence on BCT implementation but did not report quantitative effect estimates. This section also highlights how BCTs were implemented across SMS-based strategies, mobile apps, online platforms, and digital learning modules.

Figure 2. Summary of the effectiveness of behavior change techniques in digital interventions for sexually transmitted infection and HIV prevention.

	Author	AMSTAR-2 Rating	1.1 Goal setting (behavior)	1.9 Commitment	2.2 Feedback on behavior	2.6 Biofeedback	4.2 Information about antecedents	4.3 Re-attribution	5.1 Information about health consequences	5.2 Salience consequences	7.1 Prompts/Cues	9.1 Credible source	10.1 Material incentive (behavior)	10.4 Social reward	10.5 Self-Incentive	11.2 Reduce negative emotions
Overall effect			▲		▲		▲		O		▲	O	▲	O+	O+	O+
Behavioral outcome																
Condom use	Bailey 2021	🟢	▲	▲	▲	-	▲	▲	▲	-	-	-	-	-	-	-
	Burns 2016	🟡	O	-	-	-	-	-	O	-	O	-	-	-	-	-
Sexual health services uptake	Burns 2016	🟡	O+	-	▲	-	-	-	▲	-	▲	-	▲	-	-	-
STI and HIV testing	Burns 2016	🟡	O	O+	▲	-	-	-	▲	-	▲	-	O	-	-	-
	Clarke 2022	🔴	-	-	-	-	-	-	O	-	▲	O	O	O+	O+	O+
Cognitive outcome																
HIV knowledge	Bailey 2021	🟢	▲	▲	▲	▲	-	-	▲	-	-	-	-	-	-	-
HIV prevention self-efficacy	Bailey 2021	🟢	▲	▲	▲	-	-	-	▲	-	-	-	-	-	-	-
HIV prevention intentions	Bailey 2021	🟢	O+	O+	O+	-	-	-	O	-	-	-	-	-	-	-
Attitudes toward condom use	Bailey 2021	🟢	O+	O+	O+	-	-	-	O	▲	-	-	-	-	-	-
Biological outcome																
STI and HIV acquisition	Bailey 2021	🟢	O+	O+	O	-	-	-	▲	-	-	-	-	-	-	-

Behavioral Outcomes

Condom Use

Digital interventions showed mixed effectiveness. Bailey et al [56] found a significant increase in condom use (odds ratio [OR] 1.28, 95% CI 1.04-1.57, ▲ evidence), supporting the role of goal setting (1.1), commitment (1.9), and feedback (2.2). Conversely, Burns et al [58], reported no significant effect (relative risk [RR] 0.87, 95% CI 0.62-1.24, O evidence), suggesting that effectiveness varied by population and intervention design.

Regarding mobile apps for risk awareness [56], digital interventions focusing on health consequences (5.1) significantly improved HIV-related knowledge (standardized mean difference [SMD]=0.56, 95% CI 0.33-0.80), reinforcing the role of interactive digital tools in promoting condom use behaviors.

XSL•FO
RenderX

Sexual Health Services Uptake

To ensure conceptual clarity, in this overview STI and HIV testing refers to diagnostic testing behaviors, whereas sexual health services uptake encompasses broader engagement with preventive or clinical services, including retesting when reported as part of general service use [58].

Mixed results (O+ to ▲) were found regarding the effectiveness of goal setting (1.1) and Feedback (2.2) in increasing engagement with sexual health services. SMS reminders significantly increased clinic attendance (RR 0.86, 95% CI 0.74-1) and chlamydia retesting rates (RR 4.5, 95% CI 1.05-19.22) [58]. SMS reminders providing health-related information (5.1) and material incentives (10.1) were effective in some cases, particularly for STI retesting, though results varied across populations.

Regarding SMS feedback and goal setting [58], participants received weekly SMS messages inquiring about risky sexual behaviors, followed by feedback and goal-setting prompts to encourage health service use.

STI/HIV Testing

Evidence from Burns et al [58] was mixed (O to ▲), with Feedback (2.2) and prompts and cues (7.1) being somewhat effective in encouraging STI testing. Clarke et al [59] found that personalized SMS reminders significantly improved STI retesting rates (56% vs. 33%; *P*<.01). However, material incentives (BCT 10.1) yielded mixed effects (O), as financial rewards increased short-term attendance but did not sustain engagement.

Regarding personalized SMS reminders and incentives [59], tailored SMS messages encouraged STI retesting, with financial incentives increasing attendance (29.17% in the incentive group vs. 0% in the control group).

Cognitive Outcomes

HIV-Related Knowledge

Digital interventions incorporating Information about Health Consequences (5.1) and Biofeedback (2.6) significantly improved HIV-related knowledge (SMD=0.56, 95% CI 0.33-0.80, ▲ evidence). These interventions relied on mobile apps to enhance risk awareness through educational modules and interactive tools [56].

HIV Prevention Self-Efficacy

Strong evidence (▲) supported Goal Setting (1.1), Commitment (1.9), and Feedback (2.2) in improving self-efficacy for HIV prevention. Self-monitoring and digital education modules were key to reinforcing behavior change [56].

HIV Prevention Intentions and Attitudes Toward Condom Use

For both outcomes, evidence was weaker (O+ to O), indicating that Goal Setting (1.1) and Feedback (2.2) may contribute to positive attitudes toward condom use, but their long-term impact remains uncertain (SMD=0.16, 95% CI 0.06-0.26). Health Consequences (5.1) showed inconsistent effects, suggesting that behavior change may require additional reinforcement. [56].

Regarding mobile apps for risk awareness [56]: digital interventions focusing on HIV prevention knowledge were designed to improve risk awareness and motivation, though long-term behavior adoption remained a challenge.

Complementary evidence was provided by Mo et al [78], who systematically reviewed digital HIV prevention interventions for adolescents and young adults and identified a broad set of BCTs mapped across the included programs. Although this review did not quantify behavioral or biological outcomes associated with specific BCTs, it reported consistent improvements in key cognitive determinants—particularly HIV-related knowledge, prevention self-efficacy, and attitudes toward condom use. These findings reinforce the role of information-based strategies (eg, providing information about health consequences, 5.1), feedback mechanisms (2.2), and prompts and cues (7.1) as foundational techniques supporting cognitive readiness for behavior change in digital sexual health interventions.

Biological Outcomes: STI/HIV Acquisition

Weak positive evidence (O+ to O) was assigned to goal setting (1.1), commitment (1.9), and feedback (2.2), with ▲ (strong evidence) for health consequences (5.1). Despite improvements in knowledge and behavioral determinants, these interventions did not significantly reduce STI and HIV acquisition rates (OR 1.48, 95% CI 0.96-2.28; *P*=.08) [56].

In Figure 2, summary of the effectiveness of individual BCTs reported in the 3 reviews included in the BCT subset analysis. Outcomes are grouped into behavioral (eg, condom use, STI and HIV testing), cognitive (eg, knowledge, self-efficacy), and biological domains. Effect strength was graded using a 5-level evidence classification system adapted from Michie et al. (2018) [9] and Mair et al. (2023) [53]: "▲" denotes strong positive. "O" denotes mixed or null, "O+/O−" denotes limited positive or negative, and "▼" denotes strong negative. AMSTAR-2 quality ratings are indicated by colored circles as follows: Green = High, Yellow = Moderate, Orange = Low, Red = Critical Low Classification of effectiveness from this systematic reviews and meta-analyses. ▲ = Positive effect of BCT based on good evidence such as subgroup or regression analyses. O+ = Positive effect of BCT based on low evidence such as frequency of individual BCTs within effective interventions. O = Mixed evidence or no effect. O− = Negative effect of BCT based on low level of evidence such as frequency of individual BCTs within effective interventions. ▼ = Negative effect of BCT based on good evidence such as subgroup or regression analyses.

Quality Appraisal of the SRs

Confidence levels were determined according to the AMSTAR-2 criteria described in the Methods section, based on the number and severity of critical and noncritical weaknesses. Overall, confidence in the results was high in 21.7% (5/23) of the SRs, moderate in 13% (3/23), low in 17.4% (4/23), and critically low in 47.8% (11/23). The most common weaknesses identified were the absence of reported funding sources for primary studies (22/23, 95.7%), the lack of a full list of excluded studies (13/23, 56.5%), and the omission of a review protocol (10/23, 43.5%; section 16 in in Multimedia Appendix 2).

XSL•FO
RenderX

Overlap in Primary Studies Cited in Reviews

The overlap assessment revealed a minimal overlap in the 321 primary studies, most of which were cited only once in the 23 SRs, with a CCA of 2.70. Comparisons between pairs of SRs indicated that 82.6% (209 out of 253) had a low overlap (<5%), 8.7% (22 out of 253) had a moderate overlap (5% to <10%), 4% (10 out of 253) had a high overlap (10% to <15%), and 4.7% (12 out of 253) had a very high overlap (≥15%).

Given that 4 reviews (Bailey et al [56]; Burns et al [58]; Clarke et al [59]; and Mo et al [78]) provided explicit coding or mapping of BCTs, these were analyzed as a separate methodological subset (Table 3). The overlap assessment indicated consistently slight overlap (<5%) among them, confirming that this BCT-focused evidence base draws on distinct sets of primary studies. These findings are illustrated in sections 17 and 18 in in Multimedia Appendix 2.

Discussion

Principal Results

This overview synthesized evidence from 23 SRs to evaluate the behavioral determinants, theoretical foundations, and effectiveness of DBCIs for preventing STIs and HIV. Consistent with the study aims, the findings indicate that DBCIs, particularly SMS reminders, mobile apps, and interactive web-based programs, can enhance engagement with sexual health services and increase STI and HIV testing. Cognitive determinants, such as HIV-related knowledge, motivation, and self-efficacy, also improved across numerous interventions. The most frequently identified and effective BCTs included goal setting, feedback on behavior, and prompts and cues. However, the high heterogeneity of intervention formats and the inconsistent reporting of BCTs across reviews limit the ability to draw definitive conclusions regarding the independent contribution of specific components. Taken together, these findings offer a clear synthesis of behavioral mechanisms across digital interventions and strengthen confidence in the patterns observed across reviews.

Only 4 reviews, Bailey et al [56], Burns et al [58], Clarke et al [59], and Mo et al [78], explicitly coded or mapped BCTs using standardized frameworks (BCTTv1 or TDF), allowing for a focused subset analysis. For the remaining reviews, behavioral mechanisms could only be inferred from narrative descriptions. The GROOVE overlap analysis confirmed minimal overlap across the 321 primary studies, indicating that the synthesized evidence draws from distinct and independent datasets. This methodological combination strengthens the validity of the synthesized findings and responds directly to recent calls for more mechanism-oriented evidence synthesis in digital health [80].

Interpretation of Findings

Overall, DBCIs show considerable promise in supporting STI and HIV prevention, particularly through timely reminders, personalized feedback, and interactive educational content. SMS-based interventions consistently improved clinic attendance and STI retesting, reinforcing the well-established role of prompts and cues in facilitating preventive behaviors. Web-based programs and mobile apps contributed to enhanced knowledge, motivation, and behavioral skills, aligning with established behavioral models emphasizing the importance of cognitive determinants. This overview highlights that cognitive determinants represent the most consistent pathways of change in DBCIs, clarifying how digital components influence prevention behaviors.

Behavioral outcomes such as condom use demonstrated mixed effects: Bailey et al [56] reported significant improvements, whereas Burns et al [58] found no notable impact. These differences likely reflect variations in populations, intervention content, MoDs, and follow-up durations, as well as the extent to which interventions incorporated explicit BCTs or theoretical frameworks. Findings from Mo et al [78] further highlight the inconsistency in how interventions implement and report behavior change strategies, particularly among adolescent populations. These discrepancies underscore that digital modalities may be more effective when they target motivational and self-regulatory determinants rather than complex interpersonal behaviors, a distinction that has been underexplored in prior reviews.

The limited and inconsistent application of behavioral theory across reviews is a central issue. Nearly half of the reviews did not reference any behavioral framework, despite strong evidence from broader health literature showing that theoretically grounded digital interventions are more effective and more interpretable. This gap constrains the field's ability to identify mechanisms of action and optimize intervention design. By systematically examining these gaps, our overview provides conceptual clarity on how limited theoretical integration constrains interpretability, scalability, and optimization of digital interventions. Addressing these gaps is essential for moving the field toward theory-driven digital prevention, where mechanisms of action are explicitly linked to intervention components and outcomes.

The updated search added 4 recent SRs [75-78], which expand the evidence base with up-to-date findings on digital PrEP adherence support, smartphone-based HIV prevention tools, and BCT-coded adolescent interventions. However, their conclusions mirror earlier patterns: digital tools consistently improve testing and cognitive outcomes, whereas sustained behavioral and biological impacts remain inconsistent. By incorporating the most recent evidence available, including 4 SRs published in 2024-2025, this overview offers a timely and comprehensive picture of current digital prevention strategies and how they engage (or fail to engage) key behavioral mechanisms.

Comparison With Prior Work

Our findings align with prior work demonstrating that digital interventions can effectively promote preventive health behaviors when grounded in behavioral theory and equipped with active engagement strategies. Studies by Simoni et al [23], Albarracín et al [34], and Thomas Craig et al [24] underscore the importance of behavioral mechanisms, self-regulation processes, and contextual tailoring, elements only partially reflected across the reviews included in this overview. Unlike previous syntheses, this overview integrates behavioral

XSL•FO
RenderX

mechanisms, intervention content, and methodological quality to provide a more coherent understanding of how and why DBCIs work. This integration allows for a more granular understanding of intervention pathways than outcome-only syntheses can provide.

A key divergence from other health domains (eg, mental health and chronic disease management) is the limited integration of emerging digital technologies, such as AI-driven personalization, conversational agents, and virtual or augmented reality, in STI and HIV prevention. Similarly, sex- and gender-disaggregated analyses remain rare, with only 1 review explicitly addressing transgender populations. This raises concerns about equity and generalizability in digital sexual health research. This highlights a broader limitation in the STI and HIV prevention literature: the field has been slower than other digital health domains to adopt advanced technologies and rigorous behavioral frameworks, reducing its capacity to generate equitable and generalizable impact. Future digital prevention efforts must bridge this technological and conceptual gap to ensure that innovation translates into equitable public health impact. Extending beyond earlier reviews, this overview examines not only whether digital interventions work but also how and why they work, through explicit mapping of behavioral mechanisms across reviews.

Strengths and Limitations

This overview offers the most comprehensive behavioral synthesis to date of digital interventions for STI and HIV prevention, following the PRISMA-S (Multimedia Appendix 3). The use of AMSTAR-2 enabled robust appraisal of methodological quality, while GROOVE allowed quantification of overlap across primary studies. Together, the minimal overlap and explicit behavioral coding enhance confidence in the synthesized evidence.

However, several limitations should be acknowledged. First, conclusions depend on the quality and reporting of the included SRs, nearly half of which were rated critically low by AMSTAR-2. Second, BCT coding was absent or insufficient in most reviews, restricting the depth of mechanistic synthesis. Third, high intervention heterogeneity precluded meta-analytic pooling and limits comparability. Finally, reliance on SRs means that relevant primary studies not captured in those reviews may have been missed.

Despite these limitations, the overview provides a clear, methodologically grounded synthesis of how digital tools contribute to STI and HIV prevention and where future research should focus. This multilayered approach responds to recent calls for more rigorous, mechanism-oriented evidence synthesis capable of informing real-world decision-making and aligns with emerging frameworks such as evidence-based X, which emphasize the integration of mechanisms, context, and methodological rigor in digital health research [80].

Conclusions

DBCIs represent a promising and scalable strategy for strengthening STI and HIV prevention, particularly for improving testing behaviors and key cognitive determinants. Interventions incorporating goal setting, feedback on behavior, and prompts and cues show the most consistent positive effects, whereas outcomes related to condom use and biological measures remain mixed. The inconsistent application and reporting of behavioral theory across reviews limits both interpretability and scalability.

Building on prior work, this overview provides a novel contribution by integrating effectiveness evidence with a systematic mapping of BCTs and theoretical mechanisms, a perspective that has been largely absent from earlier syntheses. By identifying which active components are most consistently associated with beneficial outcomes and highlighting persistent reporting and methodological gaps, this study offers actionable guidance for designing next-generation digital interventions. These findings, which incorporate evidence updated through 2025, have pragmatic real-world implications for researchers, clinicians, and policymakers seeking to develop scalable, culturally responsive, and equity-focused digital prevention programs that meaningfully address the behavioral pathways that drive STI and HIV risk across diverse populations. Collectively, these insights offer a pathway for accelerating the development of digital public health tools that are both evidence-based and behaviorally informed.

Funding

This work was funded by the ANID, the Chilean government, the National Scientific and Technological Development Fund (FONDECYT) research project number 1250316. The funders had no role in the study design; data collection, analysis, or interpretation; the drafting of the report; or the decision to submit the paper for publication.

Data Availability

Data sharing is not applicable to this article as no new datasets were generated or analyzed during this study. All data used in this overview were extracted from previously published systematic reviews and are included in the manuscript and its supplementary materials.

Acknowledgments

The authors acknowledge the University of Santiago of Chile for granting the sabbatical under University Decree No. 372 and Memorandum No. 43663 in support of GD's PhD dissertation. This work would not have been possible without the help of the PhD Program in the Psychology of Communication and Change at Universitat de Barcelona. The authors would also like to thank the Spanish Ministry of Science, Innovation, and Universities (MICIU/AEI/10.13039/501100011033) under research project PID2020-115486GB-I00 for their support. No generative artificial intelligence tools (such as ChatGPT or other large language models) were used to write, edit, or analyze any portion of this manuscript.

XSL•FO
RenderX

Authors' Contributions

The study design and literature search strategy were developed by GD, SSC, and SCM. Articles were located, identified, and evaluated by GD and DL, who also check the data extraction. All authors were responsible for the initial dimension and criteria list. The manuscript was drafted and edited by GD with close revision and feedback from SSC and SCM as well as review and feedback from the other coauthors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Reporting guideline for overviews of reviews of healthcare interventions: development of the PRIOR statement.
[DOCX File , 18 KB - jmir_v28i1e74201_app1.docx]

Multimedia Appendix 2
Behavioral Determinants and Effectiveness of Digital Behavior Change Interventions for STI/HIV Prevention: An Overview of Systematic Reviews.
[DOCX File , 1900 KB - jmir_v28i1e74201_app2.docx]

Multimedia Appendix 3
PRISMA-S Checklist for Reporting Literature Searches.
[DOCX File , 16 KB - jmir_v28i1e74201_app3.docx]

References

1. Murray C, Aravkin AY, Zheng P, Abbafati C, Abbas KM, Abbasi-Kangevari M, et al. Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020;396(10258):1223-1249 [FREE Full text] [doi: 10.1016/S0140-6736(20)30752-2] [Medline: 33069327]
2. Fishbein M, Triandis H, Kanfer F, Becker M, Middlestadt S, Eichler A. Factors influencing behavior and behavior change. In: Handbook of Health Psychology. Mahwah, NY: Erlbaum; 2001.
3. Tran KB, Lang JJ, Compton K, Xu R, Acheson AR, Henrikson HJ, et al. The global burden of cancer attributable to risk factors, 2010-19: a systematic analysis for the global burden of disease study 2019. Lancet 2022;400(10352):563-591 [FREE Full text] [doi: 10.1016/S0140-6736(22)01438-6] [Medline: 35988567]
4. Zheng Y, Yu Q, Lin Y, Zhou Y, Lan L, Yang S, et al. Global burden and trends of sexually transmitted infections from 1990 to 2019: an observational trend study. Lancet Infect Dis 2022;22(4):541-551. [doi: 10.1016/S1473-3099(21)00448-5] [Medline: 34942091]
5. Alarming decline in adolescent condom use, increased risk of sexually transmitted infections and unintended pregnancies, reveals new WHO report. World Health Organization. URL: https://tinyurl.com/r49f98yc [accessed 2024-09-03]
6. Eisingerich AB, Wheelock A, Gomez GB, Garnett GP, Dybul MR, Piot PK. Attitudes and acceptance of oral and parenteral HIV preexposure prophylaxis among potential user groups: a multinational study. PLoS One 2012;7(1):e28238 [FREE Full text] [doi: 10.1371/journal.pone.0028238] [Medline: 22247757]
7. Infecciones de transmisión sexual (ITS) [Web page in Spanish]. World Health Organization. 2024. URL: https://www.who.int/es/news-room/fact-sheets/detail/sexually-transmitted-infections-(stis) [accessed 2024-08-31]
8. Ritchie H, Spooner F, Roser M. Causes of death. Our World in Data Internet. URL: https://ourworldindata.org/causes-of-death [accessed 2022-11-30]
9. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
10. Johnson BT, Michie S, Snyder LB. Effects of behavioral intervention content on HIV prevention outcomes: a meta-review of meta-analyses. J Acquir Immune Defic Syndr 2014;66 Suppl 3(Suppl 3):S259-S270 [FREE Full text] [doi: 10.1097/QAI.0000000000000235] [Medline: 25007195]
11. Duarte-Anselmi G, Okan Y, Johnston M, Ortiz L, Villalobos Dintrans P, Armayones M. Introduction to behavioral science and its practical applications in public health. Medwave 2025;25(1):e3017-e3017. [doi: 10.5867/medwave.2025.01.3017] [Medline: 39836869]
12. Michie S, Johnston M. Theories and techniques of behaviour change: developing a cumulative science of behaviour change. Health Psychol Rev 2012;6(1):1-6. [doi: 10.1080/17437199.2012.654964]
13. How to prevent STIs. U.S. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/sti/prevention/index.html [accessed 2024-09-04]

XSL•FO
RenderX

14. Duarte-Anselmi G, Crane SM, Ruiz MA, Dintrans PV. Behavioral science meets public health: a scoping review of the fogg behavior model in behavior change interventions. BMC Public Health 2025;25(1):3468 [FREE Full text] [doi: 10.1186/s12889-025-24525-y] [Medline: 41088011]
15. Shi H, Du J, Jin G, Yang H, Guo H, Yuan G, et al. Effectiveness of eHealth interventions for HIV prevention, testing and management: an umbrella review. Int J STD AIDS 2024;35(10):752-774. [doi: 10.1177/09564624241252457] [Medline: 38733263]
16. Milne-Ives M, Homer SR, Andrade J, Meinert E. Potential associations between behavior change techniques and engagement with mobile health apps: a systematic review. Front Psychol 2023;14:1227443 [FREE Full text] [doi: 10.3389/fpsyg.2023.1227443] [Medline: 37794916]
17. Woodward C, Bloch S, McInnes-Dean A, Lloyd KC, McLeod J, Saunders J, et al. Digital interventions for STI and HIV partner notification: a scoping review. Sex Transm Infect 2024;100(4):242-250 [FREE Full text] [doi: 10.1136/sextrans-2023-056097] [Medline: 38754986]
18. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
19. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. J Med Internet Res 2017;19(6):e232 [FREE Full text] [doi: 10.2196/jmir.7126] [Medline: 28663162]
20. Hollands GJ, Shemilt I, Marteau TM, Jebb SA, Kelly MP, Nakamura R, et al. Altering micro-environments to change population health behaviour: towards an evidence base for choice architecture interventions. BMC Public Health 2013;13:1218 [FREE Full text] [doi: 10.1186/1471-2458-13-1218] [Medline: 24359583]
21. Hekler EB, Michie S, Pavel M, Rivera DE, Collins LM, Jimison HB, et al. Advancing models and theories for digital behavior change interventions. Am J Prev Med 2016;51(5):825-832 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.013] [Medline: 27745682]
22. West R, Michie S. Guide to Development and Evaluation of Digital Behaviour Change Interventions in Healthcare. United Kingdom: Silverback Publishing; 2016.
23. Simoni JM, Ronen K, Aunon FM. Health behavior theory to enhance ehealth intervention research in HIV: rationale and review. Curr HIV/AIDS Rep 2018;15(6):423-430 [FREE Full text] [doi: 10.1007/s11904-018-0418-8] [Medline: 30511186]
24. Thomas Craig KJ, Morgan LC, Chen C, Michie S, Fusco N, Snowdon JL, et al. Systematic review of context-aware digital behavior change interventions to improve health. Transl Behav Med 2021;11(5):1037-1048 [FREE Full text] [doi: 10.1093/tbm/ibaa099] [Medline: 33085767]
25. Cane J, O'Connor D, Michie S. Validation of the theoretical domains framework for use in behaviour change and implementation research. Implement Sci 2012;7:37 [FREE Full text] [doi: 10.1186/1748-5908-7-37] [Medline: 22530986]
26. Atkins L, Francis J, Islam R, O'Connor D, Patey A, Ivers N, et al. A guide to using the theoretical domains framework of behaviour change to investigate implementation problems. Implement Sci 2017;12(1):77 [FREE Full text] [doi: 10.1186/s13012-017-0605-9] [Medline: 28637486]
27. Michie S, Atkins L, West R. The Behaviour Change Wheel Book - A Guide To Designing Interventions. United Kingdom: Silverback Publishing; 2014.
28. Smith V, Devane D, Begley CM, Clarke M. Methodology in conducting a systematic review of systematic reviews of healthcare interventions. BMC Med Res Methodol 2011;11(1):15 [FREE Full text] [doi: 10.1186/1471-2288-11-15] [Medline: 21291558]
29. Pollock M, Fernandes RM, Becker LA, Pieper D, Hartling L. Chapter V: overviews of reviews. In: Cochrane Handbook for Systematic Reviews of Interventions version 6.3. Chichester, UK: John Wiley & Sons, Ltd; 2022.
30. Glasziou P, Meats E, Heneghan C, Shepperd S. What is missing from descriptions of treatment in trials and reviews? BMJ 2008;336(7659):1472-1474 [FREE Full text] [doi: 10.1136/bmj.39590.732037.47] [Medline: 18583680]
31. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
32. Cornet VP, Holden RJ. Systematic review of smartphone-based passive sensing for health and wellbeing. J Biomed Inform 2018;77:120-132 [FREE Full text] [doi: 10.1016/j.jbi.2017.12.008] [Medline: 29248628]
33. Armayones Ruiz M, Pinto EL, Figueroa O, Robles N, Prehn DL, Riquelme FV, et al. Barriers and facilitators for safe sex behaviors in students from universidad de santiago de Chile (USACH) through the COM-B model. BMC Public Health 2023;23(1):677 [FREE Full text] [doi: 10.1186/s12889-023-15489-y] [Medline: 37041528]
34. Albarracín D, Fayaz-Farkhad B, Granados Samayoa JA. Determinants of behaviour and their efficacy as targets of behavioural change interventions. Nat Rev Psychol 2024;3(6):377-392. [doi: 10.1038/s44159-024-00305-0] [Medline: 40909436]
35. Duarte-Anselmi G, Leiva-Pinto E, Vanegas-López J, Thomas-Lange J. Experiences and perceptions on sexuality, risk and STI/HIV prevention campaigns by university students. Designing a digital intervention. Cien Saude Colet 2022;27(3):909-920 [FREE Full text] [doi: 10.1590/1413-81232022273.05372021] [Medline: 35293468]

XSL•FO
RenderX

36. Gates M, Gates A, Pieper D, Fernandes RM, Tricco AC, Moher D, et al. Reporting guideline for overviews of reviews of healthcare interventions: development of the PRIOR statement. BMJ 2022;378:e070849 [FREE Full text] [doi: 10.1136/bmj-2022-070849] [Medline: 35944924]
37. Heidari S, Babor TF, De Castro P, Tort S, Curno M. Sex and gender equity in research: rationale for the SAGER guidelines and recommended use. Res Integr Peer Rev 2016;1:2 [FREE Full text] [doi: 10.1186/s41073-016-0007-6] [Medline: 29451543]
38. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
39. Duarte G, Sanduvete-Chaves S, López-Arenas D, Chacón-Moscoso S. Digital strategies and behavior change techniques for preventing sexually transmitted infections: protocol for an overview of systematic reviews. Medwave 2025;25(2):e3020. [doi: 10.5867/medwave.2025.02.3020] [Medline: 40063926]
40. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009;339:b2535 [FREE Full text] [doi: 10.1136/bmj.b2535] [Medline: 19622551]
41. Duarte G, Sanduvete-Chaves S, Chacón-Moscoso S. Specifying the content of digital interventions for prevention of sexually transmitted infections /HIV using the behaviour change technique taxonomy v1 (BCTTv1): protocol for an overview of systematic reviews. PROSPERO International prospective register of systematic reviews Internet. URL: https://www.crd.york.ac.uk/PROSPERO/view/CRD42023485887 [accessed 2024-07-18]
42. Duarte Anselmi G. Digital behaviour change interventions to prevent sexually transmitted infections (STIs) including HIV: evidence reviews and integrated report on the quantitative and qualitative evidence. Open Science Framework. 2024. URL: https://osf.io/u9gz7/ [accessed 2024-04-12]
43. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions. Cochrane. 2022. URL: https://handbook-5-1.cochrane.org/ [accessed 2022-11-24]
44. Collaboratron™ [Software]. Epistemonikos. URL: https://www.epistemonikos.org/en/documents/647cdbf675135cca583bcfeca7dadc0560f80e02 [accessed 2024-05-06]
45. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005;37:360-363. [Medline: 15883903]
46. Ryan R, Synnot A, Hill MP. Cochrane Consumers and Communication. Data extraction template. Australia: La Trobe University; 2018.
47. Ryan R, Hill S. Supporting implementation of Cochrane methods in complex communication reviews: resources developed and lessons learned for editorial practice and policy. Health Res Policy Syst 2019;17(1):32 [FREE Full text] [doi: 10.1186/s12961-019-0435-0] [Medline: 30922338]
48. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017;358:j4008 [FREE Full text] [doi: 10.1136/bmj.j4008] [Medline: 28935701]
49. Pérez-Bracchiglione J, Meza N, Bangdiwala SI, Niño de Guzmán E, Urrútia G, Bonfill X, et al. Graphical representation of overlap for OVErviews: GROOVE tool. Res Synth Methods 2022;13(3):381-388. [doi: 10.1002/jrsm.1557] [Medline: 35278030]
50. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014;348(mar07 3):g1687-g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]
51. Marques MM, Carey RN, Norris E, Evans F, Finnerty AN, Hastings J, et al. Delivering behaviour change interventions: development of a mode of delivery ontology. Wellcome Open Res 2020;5:125. [doi: 10.12688/wellcomeopenres.15906.1]
52. Michie S, West R, Sheals K, Godinho CA. Evaluating the effectiveness of behavior change techniques in health-related behavior: a scoping review of methods used. Transl Behav Med 2018;8(2):212-224 [FREE Full text] [doi: 10.1093/tbm/ibx019] [Medline: 29381786]
53. Mair JL, Salamanca-Sanabria A, Augsburger M, Frese BF, Abend S, Jakob R, et al. Effective behavior change techniques in digital health interventions for the prevention or management of noncommunicable diseases: an umbrella review. Ann Behav Med 2023;57(10):817-835 [FREE Full text] [doi: 10.1093/abm/kaad041] [Medline: 37625030]
54. Duarte-Anselmi G, de Oliveira Barbosa AA, Ferrari G, Fernandes ACP, Araneda-Flores J, González DA, et al. Overview of systematic reviews on behavioral determinants of physical activity and healthy eating in schoolchildren. Sci Rep 2025;15(1):35379 [FREE Full text] [doi: 10.1038/s41598-025-19332-9] [Medline: 41068188]
55. Castro O, Fajardo G, Johnston M, Laroze D, Leiva-Pinto E, Figueroa O, et al. Translating the behaviour change technique taxonomy version 1 into Spanish: methodology and validation. Wellcome Open Res 2024;9:298 [FREE Full text] [doi: 10.12688/wellcomeopenres.21388.1] [Medline: 39323609]
56. Bailey JV, Wayal S, Aicken CRH, Webster R, Mercer CH, Nazareth I, et al. Interactive digital interventions for prevention of sexually transmitted HIV. AIDS 2021;35(4):643-653 [FREE Full text] [doi: 10.1097/QAD.0000000000002780] [Medline: 33259345]

XSL•FO
RenderX

57. Berendes S, Gubijev A, McCarthy OL, Palmer MJ, Wilson E, Free C. Sexual health interventions delivered to participants by mobile technology: a systematic review and meta-analysis of randomised controlled trials. Sex Transm Infect 2021;97(3):190-200. [doi: 10.1136/sextrans-2020-054853] [Medline: 33452130]
58. Burns K, Keating P, Free C. A systematic review of randomised control trials of sexual health interventions delivered by mobile technologies. BMC Public Health 2016;16(1):778 [FREE Full text] [doi: 10.1186/s12889-016-3408-z] [Medline: 27514851]
59. Clarke R, Heath G, Ross JDC, Farrow C. Increasing attendance at pre-booked sexual health consultations: a systematic review. Sex Health 2022;19(4):236-247. [doi: 10.1071/SH21245] [Medline: 35615776]
60. Conserve DF, Jennings L, Aguiar C, Shin G, Handler L, Maman S. Systematic review of mobile health behavioural interventions to improve uptake of HIV testing for vulnerable and key populations. J Telemed Telecare 2017;23(2):347-359 [FREE Full text] [doi: 10.1177/1357633X16639186] [Medline: 27056905]
61. Ilskens K, Wrona KJ, Dockweiler C, Fischer F. An evidence map on serious games in preventing sexually transmitted infections among adolescents: systematic review about outcome categories investigated in primary studies. JMIR Serious Games 2022;10(1):e30526 [FREE Full text] [doi: 10.2196/30526] [Medline: 35107438]
62. Jones K, Eathington P, Baldwin K, Sipsma H. The impact of health education transmitted via social media or text messaging on adolescent and young adult risky sexual behavior: a systematic review of the literature. Sex Transm Dis 2014;41(7):413-419. [doi: 10.1097/OLQ.0000000000000146] [Medline: 24922099]
63. Kamitani E, Peng Y, Hopkins D, Higa DH, Mullins MM, Community Preventive Services Task Force. A community guide systematic review: digital HIV Pre-exposure prophylaxis interventions. Am J Prev Med 2024;67(2):303-310. [doi: 10.1016/j.amepre.2024.02.009] [Medline: 38367928]
64. Khuwaja SS, Peck JL. Increasing HPV vaccination rates using text reminders: an integrative review of the literature. J Pediatr Health Care 2022;36(4):310-320. [doi: 10.1016/j.pedhc.2022.02.001] [Medline: 35288016]
65. Knight R, Karamouzian M, Salway T, Gilbert M, Shoveller J. Online interventions to address HIV and other sexually transmitted and blood-borne infections among young gay, bisexual and other men who have sex with men: a systematic review. J Int AIDS Soc 2017;20(3):e25017 [FREE Full text] [doi: 10.1002/jia2.25017] [Medline: 29091340]
66. Manby L, Aicken C, Delgrange M, Bailey JV. Effectiveness of eHealth interventions for HIV prevention and management in sub-saharan Africa: systematic review and meta-analyses. AIDS Behav 2022;26(2):457-469 [FREE Full text] [doi: 10.1007/s10461-021-03402-w] [Medline: 34427813]
67. Nguyen LH, Tran BX, Rocha LEC, Nguyen HLT, Yang C, Latkin CA, et al. A systematic review of ehealth interventions addressing HIV/STI prevention among men who have sex with men. AIDS Behav 2019;23(9):2253-2272 [FREE Full text] [doi: 10.1007/s10461-019-02626-1] [Medline: 31401741]
68. Ou L, Chen AC, Amresh A. The effectiveness of mhealth interventions targeting parents and youth in human papillomavirus vaccination: systematic review. JMIR Pediatr Parent 2023;6:e47334 [FREE Full text] [doi: 10.2196/47334] [Medline: 37988155]
69. Palmer M, Henschke N, Villanueva G, Maayan N, Bergman H, Glenton C, et al. Targeted client communication via mobile devices for improving sexual and reproductive health. Cochrane Database Syst Rev 2020;8(8):CD013680 [FREE Full text] [doi: 10.1002/14651858.CD013680] [Medline: 32779730]
70. Saragih ID, Imanuel Tonapa S, Porta CM, Lee B. Effects of telehealth interventions for adolescent sexual health: a systematic review and meta-analysis of randomized controlled studies. J Telemed Telecare 2024;30(2):201-214. [doi: 10.1177/1357633X211047762] [Medline: 34903065]
71. Sewak A, Yousef M, Deshpande S, Seydel T, Hashemi N. The effectiveness of digital sexual health interventions for young adults: a systematic literature review (2010-2020). Health Promot Int 2023;38(1):38. [doi: 10.1093/heapro/daac104] [Medline: 36757346]
72. Veronese V, Ryan KE, Hughes C, Lim MS, Pedrana A, Stoové M. Using digital communication technology to increase HIV testing among men who have sex with men and transgender women: systematic review and meta-analysis. J Med Internet Res 2020;22(7):e14230 [FREE Full text] [doi: 10.2196/14230] [Medline: 32720902]
73. Schnall R, Travers J, Rojas M, Carballo-Diéguez A. eHealth interventions for HIV prevention in high-risk men who have sex with men: a systematic review. J Med Internet Res 2014;16(5):e134 [FREE Full text] [doi: 10.2196/jmir.3393] [Medline: 24862459]
74. Xin M, Viswanath K, Li AY, Cao W, Hu Y, Lau JT, et al. The effectiveness of electronic health interventions for promoting HIV-preventive behaviors among men who have sex with men: meta-analysis based on an integrative framework of design and implementation features. J Med Internet Res 2020;22(5):e15977 [FREE Full text] [doi: 10.2196/15977] [Medline: 32449685]
75. Du J, Jin G, Zhang H, Don O, Shi H, Wang S, et al. Effectiveness of digital health interventions in promoting the pre-exposure prophylaxis (PrEP) care continuum among men who have sex with men (MSM): a systematic review of randomized controlled trials. Curr HIV/AIDS Rep 2025;22(1):25. [doi: 10.1007/s11904-025-00733-4] [Medline: 40100510]
76. Huang W, Stegmueller D, Ong JJ, Wirtz SS, Ning K, Wang Y, et al. Technology-based HIV prevention interventions for men who have sex with men: systematic review and meta-analysis. J Med Internet Res 2025;27:e63111 [FREE Full text] [doi: 10.2196/63111] [Medline: 40293786]

XSL•FO
RenderX

77. Li F, Xie C, Xiang F. Impact of mHealth on enhancing pre-exposure prophylaxis adherence and strengthening the HIV prevention cascade among key populations: a systematic review and meta-analysis. Front Public Health 2025;13:1600773 [FREE Full text] [doi: 10.3389/fpubh.2025.1600773] [Medline: 40642253]
78. Mo PK, Xie L, Lee TC, Li AYC. Use of behavior change techniques in digital HIV prevention programs for adolescents and young people: systematic review. JMIR Public Health Surveill 2025;11:e59519 [FREE Full text] [doi: 10.2196/59519] [Medline: 40293783]
79. World bank country and lending groups. World Bank Data Help Desk. URL: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups [accessed 2025-11-08]
80. Stewart G. Is realising evidence-based X the future of evidence synthesis? Campbell Syst Rev 2025;21(2):e70037. [doi: 10.1002/cl2.70037] [Medline: 40124218]

Abbreviations

BCT: behavior change technique
BCTTv1: behavior change technique taxonomy version 1
CCA: corrected covered area
DBCI: digital behavior change intervention
GROOVE: Graphical Representation of Overlap for Overviews
HPV: human papillomavirus
MoD: mode of delivery
OR: odds ratio
PICOS: Population, Intervention, Comparison, Outcome, Study type
PrEP: pre-exposure prophylaxis
PRESS: Peer Review of Electronic Search Strategies
PRIOR: Preferred Reporting Items for Overviews of Reviews
PRISMA-S: Preferred Reporting Items for SRs and Meta-Analyses Literature Search Extension
RCT: randomized controlled trial
RR: relative risk
SAGER: Sex and Gender Equity in Research
SMD: standardized mean difference
SR: systematic review
STD: sexually transmitted disease
STI: sexually transmitted infection
TDF: theoretical domains framework

Edited by S Brini; submitted 17.Apr.2025; peer-reviewed by C Hao, A Eisingerich; comments to author 31.Oct.2025; accepted 21.Dec.2025; published 29.Jan.2026.

Please cite as:
Duarte-Anselmi G, Sanduvete-Chaves S, Chacón-Moscoso S, López-Arenas D
Behavioral Determinants and Effectiveness of Digital Behavior Change Interventions for the Prevention of Sexually Transmitted Infections and HIV: Overview of Systematic Reviews
J Med Internet Res 2026;28:e74201
URL: https://www.jmir.org/2026/1/e74201
doi:10.2196/74201
PMID:

©Giuliano Duarte-Anselmi, Susana Sanduvete-Chaves, Salvador Chacón-Moscoso, Daniel López-Arenas. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effect of Digital Health Interventions on College Students' Lifestyle Behaviors: Systematic Review

Qingyuan Zhou[1*], BEd; Jiajun Jiang[1*], BEd; Zhihua Yin[1], PhD; Ruishi Fan[1], BEd

College of Physical Education and Health, East China Normal University, Shanghai, China

*these authors contributed equally

Corresponding Author:
Zhihua Yin, PhD
College of Physical Education and Health
East China Normal University
500 Dongchuan Road
Shanghai, 200241
China
Phone: 86 1 512 104 3880
Email: yzhkj86888@sina.com

Abstract

Background: College students undergo a critical transition from adolescence to adulthood, during which lifestyle behaviors such as physical activity, sedentary behavior, diet, and sleep are key determinants of long-term health. Digital health interventions (DHIs) are increasingly recognized as a promising strategy for improving these behaviors among college students.

Objective: This systematic review aims to evaluate the effectiveness and applicability of DHIs targeting lifestyle behaviors among college students by analyzing intervention objectives, modalities, functionalities, outcomes, and other key characteristics.

Methods: In accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines, multiple scientific databases, including Scopus, Web of Science, PubMed, MEDLINE, PsycINFO, SPORTDiscus, ProQuest Central, APA PsycArticles, ERIC, and Academic Search Premier, were searched for studies published between January 2010 and December 2025 (initial search: August 5, 2025; updated search: December 27, 2025). The inclusion criteria were original empirical studies on DHIs targeting lifestyle behaviors (physical activity, sedentary behavior, diet, and sleep) among college students, published in English. Studies focusing on nondigital interventions, lacking sufficient methodological details, or not reporting lifestyle behavior–related outcomes were excluded. Quality assessment was conducted in 2 stages: all studies were first evaluated using the Mixed Methods Appraisal Tool (2018 version), followed by Risk of Bias 2 for randomized controlled trials and Joanna Briggs Institute critical appraisal tools for nonrandomized studies. A narrative synthesis was used to present and synthesize the findings.

Results: A total of 2998 records were retrieved, of which 46 publications met the inclusion criteria. These included 30 (65%) studies related to physical activity, 26 (57%) studies to diet, 10 (22%) studies related to sedentary behavior, and 6 (13%) studies related to sleep. This review enabled an examination of the effects of DHIs on college students' lifestyle behaviors. DHIs primarily used mobile apps, web-based platforms, and mobile communication technologies, with core functionalities such as education, guidance, monitoring, and prompting. DHIs were more effective in improving physical activity and diet; however, evidence for reducing sedentary behavior and improving sleep remained limited. Of the 46 studies, 31 (67%) reported positive effects, with larger sample sizes and intervention durations of 8-16 weeks being associated with more favorable outcomes.

Conclusions: This review focuses on college students, addressing a gap in the literature that often centers on general adult populations. Unlike previous reviews that focus on a single behavior, this study integrates multiple lifestyle behaviors and evaluates DHIs across diverse modalities and functionalities. These contributions help refine future DHIs for college students and inform health promotion strategies in higher education. Although DHIs show potential for improving lifestyle behaviors, evidence of their long-term effectiveness remains limited. Future interventions should prioritize multibehavior integration, interactivity, and population-differentiated design to enhance precision, sustainability, and equity. This study has several limitations, including issues related to sample representativeness, intervention refinement, and methodological rigor.

Trial Registration: PROSPERO CRD420251119078; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251119078

(J Med Internet Res 2026;28:e82192) doi:10.2196/82192

XSL•FO
RenderX

KEYWORDS

digital health interventions; college students; lifestyle behaviors; systematic review; physical activity; sedentary behavior; diet; sleep

Introduction

College students are in a critical developmental stage characterized by the transition from adolescence to adulthood, during which they encounter multiple challenges, including increased academic demands and evolving social roles. Evidence suggests that college students often exhibit insufficient self-management capacity related to healthy lifestyle behaviors [1], with inadequate physical activity, prolonged sedentary behavior, irregular diet patterns, and sleep disturbances being particularly prevalent. Previous research has demonstrated that health behaviors established during this developmental period tend to exhibit substantial stability and continuity over time [2]. The adoption of unhealthy lifestyle behaviors during this stage has been shown to significantly increase the risk of chronic diseases, depression, and anxiety later in adulthood [3,4]. Therefore, the implementation of early and effective interventions targeting these 4 key lifestyle behaviors among college students is of substantial public health significance [1].

With the rapid advancement of digital technologies and the widespread adoption of smart devices, digital health interventions (DHIs) have emerged as an innovative approach to health promotion and are increasingly recognized as an important means of improving lifestyle behaviors among college students [5,6]. Particularly in the post–COVID-19 era, DHIs have demonstrated greater adaptability and broader application potential than traditional face-to-face health intervention models [7-9]. In recent years, a growing body of empirical evidence has shown that DHIs are effective in promoting physical activity among college students [10-12], reducing sedentary time [13,14], improving diet behaviors [15,16], and enhancing sleep quality [17,18]. These interventions—encompassing mobile apps, wearable devices, online platforms, and social media—offer several advantages, including low cost, high scalability, and a high degree of personalization [19,20], and have been shown to enhance user engagement and facilitate sustained behavior change [21,22]. Concurrently, advancements in emerging technologies, such as artificial intelligence, continue to drive the refinement of DHI implementation strategies and further enhance intervention effectiveness [23].

However, the existing body of research on DHIs targeting lifestyle behaviors among college students remains subject to several limitations. On the one hand, the majority of original intervention studies have focused on single lifestyle behaviors or specific technological modalities, with a relative lack of comprehensive designs that integrate multiple behaviors and intervention approaches. At the same time, key intervention dimensions—such as functional characteristics, intervention duration, participant demographics, and adherence—have yet to reach unified standards or methodological consensus [24,25]. On the other hand, existing systematic reviews and meta-analyses in this field also demonstrate limitations in terms of specificity and methodological rigor. First, systematic syntheses that specifically target the college student population remain relatively scarce, with insufficient attention paid to lifestyle behaviors such as sedentary behavior and sleep. Second, existing analyses have not adequately synthesized the combined effects of multiple lifestyle behaviors across diverse DHI intervention formats [26].

In light of the current research context and identified limitations, this review is guided by the following research questions: (1) What is the current state of the literature on DHIs targeting 4 key lifestyle behaviors among college students (physical activity, sedentary behavior, diet, and sleep)? (2) What are the specific implementation strategies and modalities of DHIs addressing these behaviors? (3) To what extent are DHIs effective in influencing these 4 target lifestyle behaviors among college students? Through a comprehensive synthesis and analysis of relevant primary research evidence, this review will explicitly consider the characteristics of college students as “digital natives” [27]. The review will systematically examine the forms, functions, and key components of different DHIs, and comprehensively evaluate their effects on the 4 target behaviors that are closely related to college student health. This review aims to clarify the applicability and effectiveness of DHIs within this population, thereby providing evidence-based recommendations for optimizing DHI tools, informing health promotion strategies in higher education settings, and guiding future research.

Methods

Search Strategy

This systematic review was prospectively registered in PROSPERO (International Prospective Register of Systematic Reviews) on August 4, 2025 (registration number: CRD420251119078), and the reporting of the review findings adheres to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines (see Multimedia Appendix 1). A comprehensive literature search was conducted across 10 major English-language electronic databases, including Scopus, Web of Science, PubMed, ProQuest Central, and 6 databases accessed via the EBSCOhost platform (MEDLINE, PsycINFO, SPORTDiscus, APA PsycArticles, ERIC, and Academic Search Premier), with Google Scholar used as a supplementary search source. In addition, the reference lists of relevant articles were screened to identify potentially missed studies. The initial search was completed on August 5, 2025, covering studies published between January 1, 2010, and June 1, 2025, for primary study identification, and an updated search was conducted on December 27, 2025, to capture studies published within the most recent 6 months; the same search strategy was applied consistently across both searches. No published search filters were used, and the search strategy was neither adapted from nor reused, in whole or in part, from previous reviews. The search strategy was initially developed by the authors and subsequently peer reviewed by an experienced searcher with

XSL•FO
RenderX

expertise in scientific information retrieval. Beyond these approaches, no study registries were searched, no purposeful searching or browsing (eg, table of contents screening, print conference proceedings, or website searches) was conducted, and no additional information was sought by contacting authors, experts, manufacturers, or other relevant parties.

The literature search strategy was systematically developed in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Search Extension) guidelines to ensure transparency and reproducibility of the search process. The strategy combined Medical Subject Headings terms with free-text terms and constructed keyword combinations around 3 core concepts: (1) intervention formats (eg, digital health, digital intervention, eHealth, and mobile health [mHealth]); (2) target behaviors (eg, physical activity, sedentary behavior, sleep, and diet); and (3) study populations (eg, university students, college students, and undergraduate students). Boolean operators (AND and OR) were applied to balance search sensitivity and specificity. Using Scopus as an example, the search query was as follows: TITLE-ABS-KEY (“digital health” OR “eHealth” OR “mHealth” OR “mobile health” OR “digital intervention” OR “health app”) AND TITLE-ABS-KEY (“college students” OR “university students” OR “undergraduate students” OR “young adults”) AND TITLE-ABS-KEY (“lifestyle behavior” OR “health behavior” OR “physical activity” OR “exercise” OR “diet” OR “nutrition” OR “sleep” OR “sedentary behavior”). The complete English-language search terms used across all databases are provided in Multimedia Appendix 2. To ensure comprehensive coverage, no geographical restrictions were applied during the literature search, allowing for the inclusion of relevant studies from diverse global regions.

Inclusion and Exclusion Criteria

The study inclusion criteria were developed in accordance with the PICOS (Population-Intervention-Comparison-Outcome -Study Design) framework, as outlined in Textbox 1.

The exclusion criteria were as follows: (1) The study population was not explicitly identified as “college students,” “university students,” or individuals enrolled in higher education institutions. (2) The nondigital components constituted the majority of the intervention (≥50%), or the study relied solely on wearable devices for passive behavioral monitoring without incorporating feedback mechanisms or active intervention strategies. (3) The study did not implement a behavioral intervention, or the intervention description lacked sufficient detail to determine its content and implementation procedures. (4) Studies that did not report any lifestyle behavior–related outcome measures. (5) Conference abstracts, theses, unpublished manuscripts, and other forms of gray literature. (6) Full-text articles were unavailable, or the publication was not in English.

Textbox 1. Study inclusion criteria.

1. Population

- Participants were required to be aged ≥18 years and explicitly identified as “college students,” “university students,” or “young adults enrolled in higher education.”

2. Intervention

- Studies were required to evaluate at least one health intervention primarily delivered through digital health technologies and targeting lifestyle-related behaviors. Digital health interventions included, but were not limited to, mobile apps, web-based platforms, SMS text message reminders, online courses, virtual coaches, digital gamification strategies, social media, and other eHealth/mHealth tools.

3. Comparison

- The presence of a control group was not mandatory; all original studies reporting intervention effects were eligible for inclusion.

4. Outcomes

- The primary outcomes included lifestyle behavior indicators, specifically physical activity, sedentary behavior, diet, and sleep.
- Secondary outcomes included physical and mental health indicators, such as weight, waist circumference, and self-efficacy.

5. Study design

- Original empirical studies targeting 1 or more of the 4 lifestyle behavior domains among college students and implementing digital health interventions were included.
- No restrictions were placed on study design; however, intervention content, participant characteristics, and relevant outcome measures were required to be clearly reported.

Study Selection

All retrieved records were imported into EndNote 20 (Clarivate Plc) reference management software for duplicate removal and standardized record numbering. Subsequently, 2 reviewers (QYZ and JJJ) independently screened titles and abstracts for initial eligibility. Records that passed the initial screening were subjected to full-text assessment to determine final eligibility for inclusion. To ensure standardization and consistency in the screening process, all reviewers received standardized training on the predefined inclusion and exclusion criteria. Interrater reliability between the 2 reviewers was assessed using the Cohen κ coefficient, yielding a value of 0.86, which indicates a high level of screening agreement. In cases of disagreement regarding

XSL•FO
RenderX

individual records, a third reviewer (ZHY) was consulted to facilitate discussion and achieve a final consensus.

Data Extraction and Synthesis

To ensure standardization and consistency in the data extraction process, the research team developed a structured data extraction form in advance, covering the study title, first author, publication year, study region, study design, intervention population characteristics, intervention protocol characteristics, outcome measures, intervention effectiveness, and study conclusions. The data extraction form was pilot-tested using 5 studies to assess its feasibility. During the formal data extraction process, 2 reviewers (QYZ and JJJ) independently extracted the data. In cases of missing data or discrepancies in interpretation, a third reviewer (ZHY) was consulted to resolve disagreements. The final extracted data were consolidated into a standardized table and are presented in Multimedia Appendix 3.

Quality Assessment

All included studies were initially assessed for methodological quality using the Mixed Methods Appraisal Tool (MMAT, 2018 version) to obtain an overall preliminary appraisal of study quality. The MMAT is designed to evaluate 5 categories of study designs: qualitative research (QR), quantitative randomized controlled trials (QRCTs), quantitative nonrandomized studies (QNRSs), quantitative descriptive studies (QDSs), and mixed methods studies (MMSs), each comprising 5 appraisal criteria [28]. To enhance specificity and methodological rigor, the Risk of Bias 2 (RoB 2) tool was further applied to assess the risk of bias in QRCTs. For all other study designs, the Joanna Briggs Institute (JBI) critical appraisal tools were applied. This 2-stage quality assessment approach was intended to balance breadth and depth in methodological evaluation. Quality assessments were conducted independently by 2 reviewers (QYZ and JJJ), with discrepancies resolved through discussion. To ensure consistency, both reviewers received standardized training on the MMAT, RoB 2, and JBI critical appraisal tools and completed pilot scoring exercises before the formal assessment.

Results

Screening and Inclusion Results

Search and Screening Results

In this study, a total of 2998 records were retrieved from 10 major English-language databases. After deduplication and initial title and abstract screening, 273 articles were selected for full-text review. Based on the predefined exclusion criteria, exclusions were made for the following reasons: nonuniversity samples (n=68); interventions not primarily digital-based (n=20); wearable devices only, without active intervention components (n=5); absence of behavioral interventions (n=83); lack of relevant behavioral outcomes (n=35); and protocol or abstract only (n=19). Additionally, 3 more articles were identified through manual reference tracing of relevant review papers. Ultimately, 46 publications met the inclusion criteria and were included in the final analysis, as depicted in Figure 1.

Figure 1. PRISMA flowchart of the study selection process.

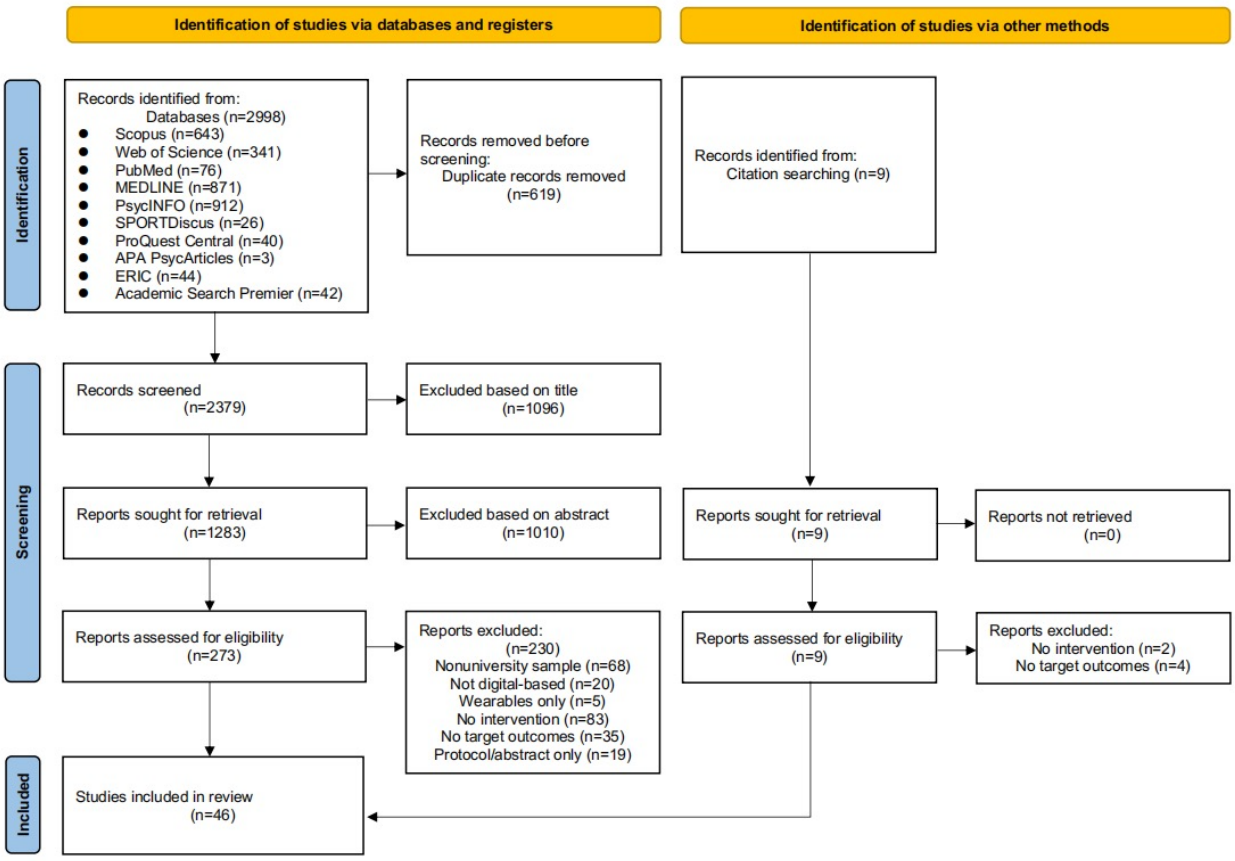

XSL•FO
RenderX

Quality Assessment Results

Following 2 rounds of quality assessment, the first-round MMAT evaluation indicated that the 46 included studies demonstrated an overall high level of methodological quality. Specifically, 28 (61%) studies were rated as high quality, 14 (30%) as moderate quality, and 4 (9%) as low quality (see Figure 2; also see [10,11,13-15,17,20,26,29-66]). Major methodological concerns identified during the assessment were primarily concentrated in MMAT items C4 and C5. Item C4 was primarily related to the implementation of blinding procedures, the adequacy of outcome interpretation, and the control of risk of bias, whereas item C5 reflected issues such as insufficient intervention adherence and the lack of rigorous statistical analyses. In the second round of assessment, the RoB 2 tool was applied to evaluate 30 QRCTs, indicating that the primary sources of bias were related to outcome measurement, deviations from intended interventions, and the handling of missing outcome data (see Figure 3; see also [13,14,17,29-31,33-35,37,38,42-48,50-52,54,56,58,59,61-65]). Concurrently, the JBI critical appraisal of the remaining 16 studies indicated that key factors influencing study quality primarily included sample representativeness, intervention adherence, and the objectivity of outcome measurement (see Figure 4; see also [10,11,15,20,26,32,36,39-41,49,53,55,57,60,66]).

The emergence of these methodological issues can be primarily attributed to 2 factors. On the one hand, the behavioral nature of DHIs makes the implementation of blinding inherently challenging, and several key behavioral outcomes rely on participant self-report measures. On the other hand, relatively high dropout rates associated with DHIs contribute to issues such as low intervention adherence and elevated loss-to-follow-up rates in some studies. When combined with insufficiently rigorous statistical analyses, these challenges may result in suboptimal handling of missing data or deviations from intended interventions. Although studies rated as moderate to low quality constitute a notable proportion of the included literature, it is important to recognize that many of their methodological limitations are closely related to the inherent characteristics of DHIs. Moreover, many of these studies primarily aimed to explore the feasibility and applicability of DHIs rather than to provide definitive evidence of intervention efficacy. Therefore, these studies retain substantial value for informing future research and intervention development. Given these considerations, no studies were excluded from this review solely based on methodological quality. Instead, all eligible studies were included, and findings from risk-of-bias assessments were systematically incorporated into the narrative synthesis. This approach allows for a comprehensive presentation of the current evidence landscape while explicitly identifying both its strengths and limitations.

XSL•FO
RenderX

Figure 2. Quality assessment results of the Mixed Methods Appraisal Tool.

Num.	Study	Study design	Screening Q1	Screening Q2	C1	C2	C3	C4	C5	Results
1	Hebden et al [35]	Quantitative randomized controlled trial	YES	YES						High
2	Nour et al [66]	Quantitative descriptive study	YES	YES						High
3	Schweitzer et al [54]	Quantitative randomized controlled trial	YES	YES						High
4	Allman-Farinelli et al [29]	Quantitative randomized controlled trial	YES	YES						High
5	Walsh et al [45]	Quantitative randomized controlled trial	YES	YES						High
6	Hutchesson et al [53]	Quantitative nonrandomized study	YES	YES						Moderate
7	O'Brien and Palfai [42]	Quantitative randomized controlled trial	YES	YES						High
8	Partridge et al [30]	Quantitative randomized controlled trial	YES	YES						High
9	Cotton and Prapavessis [43]	Quantitative randomized controlled trial	YES	YES						High
10	Morris et al [33]	Quantitative randomized controlled trial	YES	YES						Low
11	Xian et al [60]	Quantitative nonrandomized study	YES	YES						Moderate
12	Sarcona et al [55]	Quantitative descriptive study	YES	YES						High
13	Ashton et al [46]	Quantitative randomized controlled trial	YES	YES						High
14	Chung et al [49]	Quantitative nonrandomized study	YES	YES						Moderate
15	Inauen et al [34]	Quantitative randomized controlled trial	YES	YES						High
16	Hershner and O'Brien [38]	Quantitative randomized controlled trial	YES	YES						High
17	Fitzsimmons-Craft et al [32]	Quantitative nonrandomized study	YES	YES						Moderate
18	Whatnall et al [56]	Quantitative randomized controlled trial	YES	YES						High
19	Nour et al [50]	Quantitative randomized controlled trial	YES	YES						Moderate
20	Lee and Park [20]	Quantitative nonrandomized study	YES	YES						Moderate
21	Napolitano et al [36]	Quantitative nonrandomized study	YES	YES						High
22	Roure et al [51]	Quantitative randomized controlled trial	YES	YES						High
23	Napolitano et al [59]	Quantitative randomized controlled trial	YES	YES						High
24	Hahn et al [63]	Quantitative randomized controlled trial	YES	YES						High
25	Figueroa et al [44]	Quantitative randomized controlled trial	YES	YES						Moderate
26	Stork et al [61]	Quantitative randomized controlled trial	YES	YES						High
27	Muntaner-Mas et al [31]	Quantitative randomized controlled trial	YES	YES						High
28	Pope and Gao [14]	Quantitative randomized controlled trial	YES	YES						Moderate
29	Smith and Volkwyn [39]	Quantitative descriptive study	YES	YES						High
30	Al-Nawaiseh et al [62]	Quantitative randomized controlled trial	YES	YES						High
31	Cantisano et al [41]	Quantitative nonrandomized study	YES	YES						Moderate
32	Haslam et al [65]	Quantitative randomized controlled trial	YES	YES						Low
33	Belogianni et al [58]	Quantitative randomized controlled trial	YES	YES						Moderate
34	Kellner et al [52]	Quantitative randomized controlled trial	YES	YES						High
35	Floyd and Vargas [37]	Quantitative randomized controlled trial	YES	YES						Low
36	Kaneda et al [47]	Quantitative randomized controlled trial	YES	YES						High
37	Åsberg et al [26]	Qualitative research	YES	YES						High
38	Rajan and Muthunarayanan [11]	Quantitative descriptive study	YES	YES						High
39	Khatri and Sharma [40]	Quantitative nonrandomized study	YES	YES						Moderate
40	Malloy et al [48]	Quantitative randomized controlled trial	YES	YES						High
41	Kim et al [17]	Quantitative randomized controlled trial	YES	YES						High
42	Wittmar et al [57]	Mixed methods study	YES	YES						Moderate
43	Andargeery and EI-Rafey [13]	Quantitative randomized controlled trial	YES	YES						High
44	Olatona et al [15]	Quantitative nonrandomized study	YES	YES						Moderate
45	Fucito et al [64]	Quantitative randomized controlled trial	YES	YES						Low
46	Gao et al [10]	Quantitative nonrandomized study	YES	YES						Moderate

XSL•FO
RenderX

Figure 3. Quality assessment results of the Risk of Bias 2 tool.



XSL•FO
RenderX

Figure 4. Quality assessment results of Joanna Briggs Institute critical appraisal tools. N/A: not applicable.

Study	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Overall appraisal
Checklist for quasi-experimental studies 2023 (9 questions)											
Hutchesson et al [53]	Yes	No	N/A	N/A	Yes	Yes	Yes	No	Yes	N/A	Include
Xian et al [60]	Yes	No	N/A	N/A	Yes	Yes	Yes	Unclear	Yes	N/A	Include
Chung et al [49]	Yes	No	N/A	N/A	Yes	Yes	Yes	Unclear	Yes	N/A	Include
Fitzsimmons-Craft et al [32]	Yes	No	N/A	N/A	Unclear	Yes	Yes	No	Yes	N/A	Include
Lee and Park [20]	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	N/A	Include
Napolitano et al [36]	Yes	No	N/A	N/A	Yes	Yes	Unclear	Yes	Yes	N/A	Include
Cantisano et al [41]	Yes	N/A	Unclear	Yes	No	Yes	N/A	No	Yes	N/A	Include
Khatri and Sharma [40]	Yes	No	Yes	Yes	Yes	N/A	Yes	Yes	Yes	N/A	Include
Olatona et al [15]	Yes	N/A	Unclear	Yes	Yes	Yes	N/A	No	Yes	N/A	Include
Gao et al [10]	N/A	N/A	Yes	Unclear	No	Yes	Yes	Yes	Yes	N/A	Include
Checklist for analytical cross sectional studies 2020 (8 questions)											
Nour et al [66]	Yes	Yes	Yes	Yes	Unclear	Unclear	Yes	Yes	N/A	N/A	Include
Sarcona et al [55]	Yes	Yes	No	Unclear	Yes	Yes	Unclear	Yes	N/A	N/A	Include
Smith and Volkwyn [39]	Yes	Yes	Unclear	Yes	No	No	Unclear	Yes	N/A	N/A	Include
Rajan and Muthunarayanan [11]	Yes	Yes	Yes	Yes	Yes	No	Unclear	Yes	N/A	N/A	Include
Wittmar et al [57]	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	Yes	N/A	N/A	Include
Checklist for qualitative research 2020 (10 questions)											
Åsberg et al [26]	No	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Include
Wittmar et al [57]	Unclear	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Include

Study design	Checklist	Study	Total questions	"N/A" count	"Yes" count	Compliance rate (%)	Quality grade	Overall appraisal
Quantitative nonrandomized study (n=10)	Checklist for quasi-experimental studies 2023	Hutchesson et al [53]	9	2	5	71.4%	Moderate	Include
		Xian et al [60]	9	2	5	71.4%	Moderate	Include
		Chung et al [49]	9	2	5	71.4%	Moderate	Include
		Fitzsimmons-Craft et al [32]	9	2	4	57.1%	Low	Include
		Lee and Park [20]	9	0	7	77.8%	High	Include
		Napolitano et al [36]	9	2	5	71.4%	Moderate	Include
		Cantisano et al [41]	9	2	4	57.1%	Low	Include
		Khatri and Sharma [40]	9	1	7	87.5%	High	Include
		Olatona et al [15]	9	2	5	71.4%	Moderate	Include
		Gao et al [10]	9	2	5	71.4%	Moderate	Include
Quantitative descriptive study (n=4)	Checklist for analytical cross sectional studies 2020	Nour et al [66]	8	0	6	75.0%	High	Include
		Sarcona et al [55]	8	0	5	62.5%	Moderate	Include
		Smith and Volkwyn [39]	8	0	4	50.0%	Moderate	Include
		Rajan and Muthunarayanan [11]	8	0	7	87.5%	High	Include
Qualitative research (n=1)	Checklist for qualitative research 2020	Åsberg et al [26]	10	0	7	70.0%	Moderate	Include
mixed methods study (n=1)	Checklist for analytical cross sectional studies 2020 + Checklist for qualitative research 2020	Wittmar et al [57]	18	0	14	77.8%	High	Include

Data Extraction Results

This review included a total of 46 studies. The basic characteristics of the included studies are summarized in Table 1, with detailed data extraction results provided in Multimedia Appendix 3. Given the substantial heterogeneity among the included studies with respect to study design, target behaviors, intervention formats, core functions, and primary outcome measures, as well as variations in methodological quality, a meta-analysis was not conducted. Instead, a comprehensive analysis was performed using descriptive synthesis and comparative approaches. By systematically organizing and describing key characteristics of DHIs—including intervention targets, participant characteristics, sample sizes, formats, functions, durations, outcomes, and effects—this review delineates the overall patterns and heterogeneity within the field. Specific details are elaborated in the subsequent sections and illustrated through relevant tables, charts, and figures.

XSL•FO
RenderX

Table 1. Summary of data extraction from included studies.

Study design and relevant studies	Total completed, N	Participant age (years), mean (SD)	Intervention(s)	Target behavior(s)	Function	Effectiveness
Quantitative randomized controlled trial						
Hebden et al [35]	46	22.8 (4.6)	SMS text messages, emails, smartphone apps, and internet forums	• Physical activity • Sedentary behavior • Diet	• Prompting • Education • Guidance	Limited[a]
Schweitzer et al [54]	106	19.7 (0.73)	Email	• Physical activity • Diet	• Guidance • Education • Prompting	Yes[b]
Allman-Farinelli et al [29]	202	27.7 (4.9)	Coaching calls, SMS text messages, emails, apps, and downloadable website resources	• Physical activity • Diet	• Guidance • Prompting • Education	Yes
Walsh et al [45]	55	20.55 (2.07)	Smartphone app	• Physical activity	• Monitoring • Feedback	Yes
O'Brien and Palfai [42]	148	19.24 (1.16)	Web and SMS text messages	• Diet	• Education • Prompting • Guidance	Limited
Partridge et al [30]	248	27.0 (4.0)	Coaching calls, SMS text messages, emails, smartphone apps, and website	• Physical activity • Diet	• Education • Guidance	Yes
Cotten and Prapavessis [43]	56	21.19 (4.19)	SMS text messages	• Sedentary behavior	• Prompting • Guidance	Limited
Morris et al [33]	112	20.5 (1.95)	Internet	• Sleep	• Education • Guidance	Yes
Ashton et al [46]	47	22.1 (2.0)	Website, wearable device, and Facebook support group	• Physical activity • Diet	• Guidance • Education • Interaction	Limited
Inauen et al [34]	141	27.5 (8.6)	App	• Diet	• Interaction • Monitoring	Limited
Hershner and O'Brien [38]	358	21.9 (4.1)	Website	• Sleep	• Education	Yes
Whatnall et al [56]	90	22.4 (4.0)	Website	• Diet	• Education • Guidance	Limited
Nour et al [50]	47	24.8 (3.4)	Self-monitoring app, gamified app, and social media (Facebook)	• Diet	• Monitoring • Interaction	Limited
Roure et al [51]	60	20.8 (1.3)	Exergame	• Physical activity	• Immersion • Engagement	Yes
Napolitano et al [59]	283	23.3 (4.4)	Facebook and SMS text messages	• Physical activity • Diet	• Prompting • Feedback	Yes
Hahn et al [63]	192	20.2 (2.4)	App	• Physical activity • Diet	• Monitoring	No[c]
Figueroa et al [44]	93	20.2 (2.47)	App and SMS text messages	• Physical activity	• Prompting • Feedback • Monitoring	Yes

XSL•FO
RenderX

Study design and relevant studies	Total completed, N	Participant age (years), mean (SD)	Intervention(s)	Target behavior(s)	Function	Effectiveness
Stork et al [61]	46	24.0 (5.0)	App	• Physical activity	• Guidance • Monitoring	Yes
Muntaner-Mas et al [31]	66	23.1 (4.0)	App	• Physical activity	• Guidance	Yes
Pope and Gao [14]	42	21.6 (NR)[d]	App and Facebook	• Physical activity • Sedentary behavior	• Monitoring • Education • Prompting	Yes
Al-Nawaiseh et al [62]	114	21.1 (2.2)	App	• Physical activity	• Monitoring • Feedback	Yes
Haslam et al [65]	141	21.7 (2.0)	Website	• Diet	• Feedback • Education	No
Belogianni et al [58]	65	23.01 (3.82)	Website	• Physical activity • Sedentary behavior • Diet	• Education • Immersion	No
Kellner et al [52]	34	22.31 (2.59)	SMS text messages	• Sedentary behavior	• Prompting	Yes
Floyd and Vargas [37]	55	19.9 (0.97)	App	• Sleep	• Guidance • Education	Yes
Kaneda et al [47]	46	20.8 (1.2)	E-learning and exercise video	• Physical activity • Sedentary behavior	• Education • Guidance	No
Malloy et al [48]	46	21.34 (2.02)	Social media	• Physical activity • Diet	• Education • Prompting • Guidance	Limited
Kim et al [17]	60	21.9 (1.43)	Virtual reality	• Sleep	• Immersion • Guidance	Yes
Andargeery and El-Rafey [13]	220	19.97 (2.61)	Mobile health tools and videos	• Physical activity • Diet • Sleep	• Education • Guidance • Monitoring	Yes
Fucito et al [64]	98	21.16 (1.75)	Wearable devices, website, and smartphone	• Sleep	• Monitoring • Guidance • Feedback	Yes
Quantitative nonrandomized study						

XSL•FO
RenderX

Study design and relevant studies	Total completed, N	Participant age (years), mean (SD)	Intervention(s)	Target behavior(s)	Function	Effectiveness
Hutchesson et al [53]	12	22.8 (3.2)	Website, emails, online forum, smartphone app, and SMS text messages	• Physical activity • Sedentary behavior • Diet	• Feedback • Education • Interaction	Yes
Xian et al [60]	167	25.0 (4.0)	Reality game	• Physical activity	• Immersion • Prompting	Yes
Chung et al [49]	12	19.8 (1.0)	Fitbit, Twitter, and gamification	• Physical activity • Diet	• Monitoring • Interaction • Prompting	Yes
Fitzsimmons-Craft et al [32]	2454	22.89 (6.59)	Online platforms	• Diet	• Screening • Guidance	Yes
Lee and Park [20]	59	22.0 (2.0)	Apps and wearable devices	• Physical activity • Diet	• Monitoring • Guidance	Yes
Napolitano et al [36]	20	18.3 (0.72)	Digital learning modules	• Physical activity • Sedentary behavior • Diet	• Monitoring • Feedback	Limited
Cantisano et al [41]	16	20.69 (1.74)	eHealth tools	• Physical activity • Diet	• Education	Limited
Khatri and Sharma [40]	500	20.74 (1.77)	App	• Sedentary behavior	• Monitoring • Feedback • Guidance	Yes
Olatona et al [15]	1182	Unclear	Social media	• Diet	• Education • Guidance	Yes
Gao et al [10]	456	21.5 (1.4)	Artificial intelligence–powered gamification	• Physical activity	• Interaction	Yes
Quantitative descriptive study						
Nour et al [66]	401	27.7 (4.9)	Telephone, website, smartphone app, and SMS text messages	• Diet	• Guidance • Education	Yes
Sarcona et al [55]	230	22.0 (3.0)	Mobile health apps	• Physical activity • Diet	• Monitoring	Yes
Smith and Volkwyn [39]	192	22.7 (3.7)	App	• Physical activity	• Monitoring • Feedback	Yes
Rajan and Muthunarayanan [11]	680	23.82 (1.62)	App	• Physical activity • Diet	• Monitoring • Education • Screening	Yes
Qualitative research						
Åsberg et al [26]	50	31.3 (6.4)	SMS text messages	• Physical activity • Sedentary behavior • Diet	• Guidance • Education • Feedback	Limited
Mixed methods study						

XSL•FO
RenderX

Study design and relevant studies	Total completed, N	Participant age (years), mean (SD)	Intervention(s)	Target behavior(s)	Function	Effectiveness
Wittmar et al [57]	142	24.0 (4.0)	Web application	• Physical activity	• Education • Interaction	Yes

[a]Limited: limited evidence of effectiveness, based on reported effect measures, CIs, and authors' conclusions (see Multimedia Appendix 3).

[b]Yes: evidence of effectiveness, based on reported effect measures, CIs, and authors' conclusions (see Multimedia Appendix 3).

[c]No: no evidence of effectiveness, based on reported effect measures and authors' conclusions (see Multimedia Appendix 3).

[d]NR: not reported.

In terms of annual distribution (see Figure 5), the number of studies during the early period (2014-2015) was low, with only 1 publication per year. Since 2016, the number of publications increased markedly, reaching a first minor peak in 2016 (n=8), possibly associated with the rapid adoption of smartphones and mobile apps among college students. From 2017 to 2020, the number of studies fluctuated between 1 and 5 annually, maintaining an overall moderate level. The number increased again and stabilized in 2021-2022, declined slightly in 2023, reached a second peak in 2024 (n=8), and remained high in 2025 (n=4). Publications from the last 5 years accounted for more than half of all studies identified.

Figure 5. Annual and cumulative publication counts of the included studies.

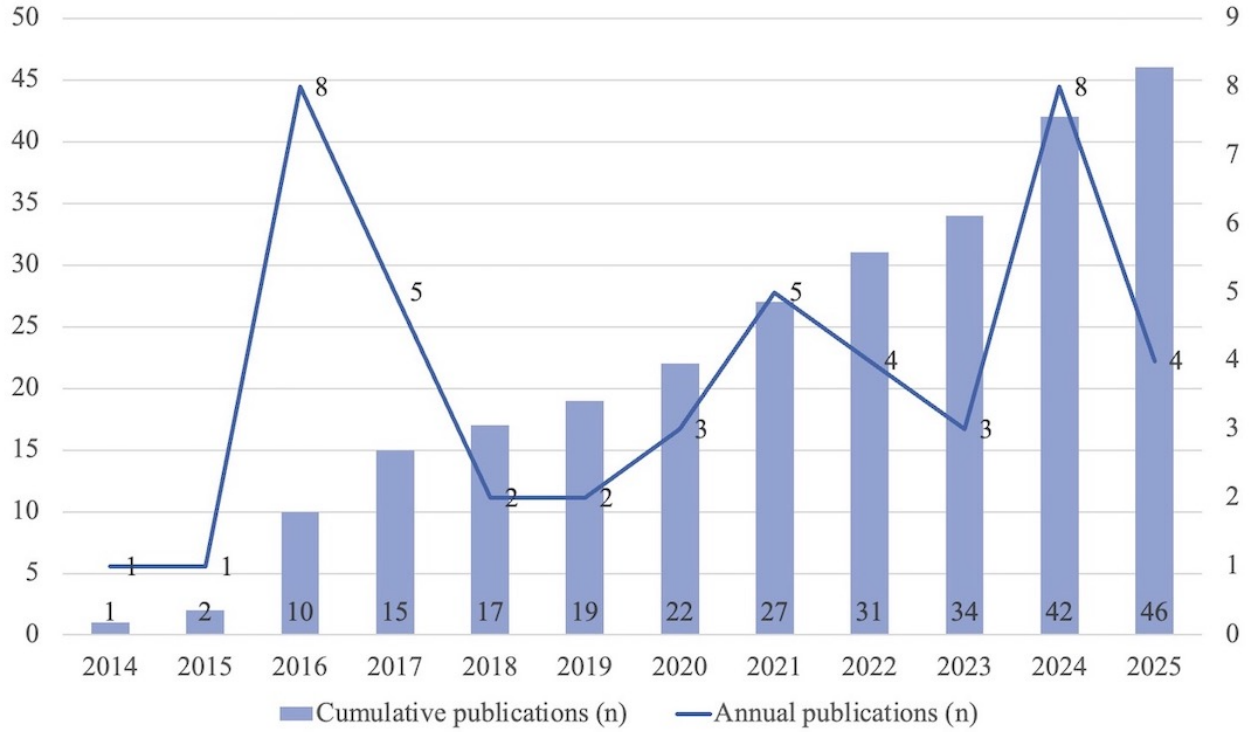

The regional and country distribution of the included studies demonstrates a clear geographical concentration. At the regional level, most studies were conducted in North America (n=18, 39%), followed by Oceania (n=10, 22%) and Europe (n=9, 20%). Asia accounted for 6 (13%) studies, while Africa contributed the smallest share with 3 (7%) studies. At the country level, the United States recorded the highest number of publications (n=15, 33%), followed by Australia (n=9, 20%). The United Kingdom, Germany, Canada, South Korea, and India each contributed 2 studies. The remaining countries were represented by a single study, indicating a relatively dispersed distribution beyond the leading contributors.

The distribution of study design types among the included studies exhibited a clear structural pattern. The largest proportion comprised QRCTs (n=30, 65%). This was followed by QNRSs (n=10) and QDSs (n=4), which were primarily used for exploratory analyses and descriptive accounts of phenomena. By contrast, QR and MMS were represented by only 1 article each, accounting for less than 2% of the total. Overall, DHI studies addressing college students' lifestyle behaviors are predominantly quantitative, with a marked preference for QRCTs.

With respect to ethical compliance, all included studies adhered to relevant ethical guidelines, with all 46 (100%) explicitly reporting informed consent procedures and ethics committee approval or review status. Regarding privacy protection and data security, 24 (52%) studies explicitly reported the implementation of protective measures, including secure server storage compliant with data safety standards, encrypted data transmission, data deidentification, and strict access control mechanisms. With respect to adverse events and intervention-related risks, no serious adverse events were

XSL•FO
RenderX

reported across the included studies. Only a small number of studies reported minor negative issues related to technology use, such as fluctuations in intervention engagement, higher dropout rates, or reduced compliance attributable to participants' competing academic or personal commitments. No health risks were identified that were directly attributable to the DHIs.

Intervention Design and Implementation Results

Intervention Objectives

Among the intervention objectives examined in the included studies, 30 addressed physical activity, 26 addressed diet, 10 targeted sedentary behavior, and 6 targeted sleep. Single-behavior interventions accounted for a large proportion of the studies; however, multibehavior crossover interventions were also substantial, with combined physical activity and diet interventions being the most common (n=18). Notably, physical activity was both the most frequent single-behavior intervention target and the primary entry point for multibehavior combined interventions, whereas sleep was relatively underemphasized in intervention design.

Intervention Participants

Based on the PROGRESS-Plus (Place of Residence, Race/Ethnicity, Occupation, Gender/Sex, Religion, Education, Socioeconomic Status, Plus Other Relevant Factors) framework, a synthesis of sociodemographic characteristics from 46 DHI studies identified 10 primary participant categories (see Multimedia Appendix 3), including health status (n=46), age (n=45), gender/sex (n=45), education (n=41), occupation (n=39), place of residence (n=36), race/ethnicity (n=28), socioeconomic status (n=14), social capital (n=8), and religion (n=1). The analysis revealed the following: (1) All participants were college students, predominantly aged 18-30 years, which is consistent with typical college student demographics and showed no substantial deviation across studies. (2) Most interventions targeted students with generally healthy status, whereas 14 out of 46 (30%) focused on subpopulations with specific health risks or special needs, such as overweight or obesity, sleep disorders, psychological stress, or disordered eating behaviors. (3) Gender/sex distribution was relatively balanced across studies, whereas education and occupation exhibited limited variability owing to the homogeneity of the study population. (4) By contrast, PROGRESS-Plus dimensions such as race/ethnicity, socioeconomic status, social capital, and religion received notably limited attention, with a lack of systematic analysis from a health equity perspective.

Intervention Sample

The sample sizes of the included studies varied considerably. Histograms indicated that most studies had sample sizes concentrated below 200 participants, with a median of approximately 95, whereas a few studies had small (<50) or extremely large (>400) samples. As shown in Figure 6, box-and-whisker plots further revealed an uneven distribution with long-tailed characteristics. Variations in sample size were closely associated with study design. Rigorous QRCTs typically require larger samples to ensure statistical power and therefore tend to employ medium- to large-scale sample sizes. By contrast, QDSs and QR are more inclined toward small-sample explorations, sometimes recruiting only a few dozen participants, and are more susceptible to selection bias.

Figure 6. Sample size distribution of the included studies.

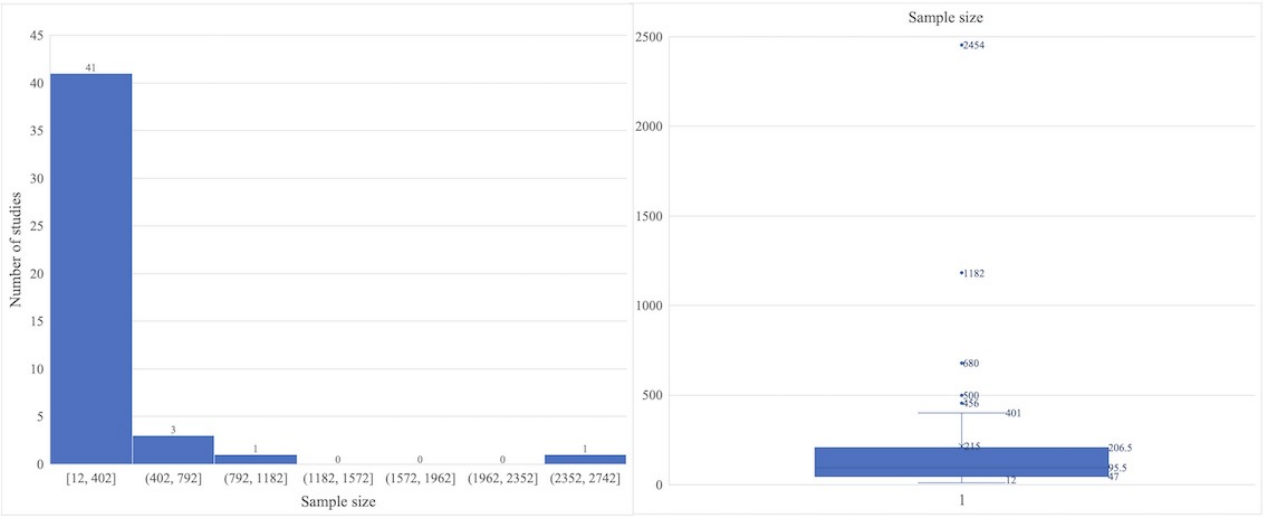

Intervention Modalities

The intervention formats in the included studies fell into 3 main categories. The first category, *single*, referred to interventions employing only 1 digital health technology (n=29), such as mobile apps. The second category, *multiple*, involved combining multiple digital health technologies within the same intervention (n=10). For example, the TXT2BFiT program integrated phone calls, websites, apps, and SMS text messaging simultaneously to achieve intervention goals. The third category, *combined* (n=7), compared the effectiveness of different combinations of digital health technologies, such as a "web-based nutrition intervention only" versus a "web-based intervention combined with daily SMS text message reminders." Regarding the types of intervention technologies, these could be categorized into 7 groups: (1) mobile apps, used 21 times; (2) web-based platforms, including websites (13 times), online forums (3 times), and digital learning or eHealth tools (4 times); (3) mobile

XSL•FO
RenderX

communications, including SMS text messages (11 times), emails (5 times), and phone calls (3 times); (4) social media (7 times); (5) wearable devices (4 times); (6) gamification and multimedia, including gamification and exergames (5 times), videos (2 times), and virtual reality (1 time); and (7) intelligent technologies, represented only by artificial intelligence (1 time). Overall, mobile apps and web-based platforms were the most frequently used technologies.

Intervention Functionalities

The technological functions of the DHIs included in this review exhibited distinct patterns of emphasis. Educational and guidance-related functions predominated across most interventions, followed by monitoring and prompting functions; by contrast, feedback and interactive functions were used less frequently, while immersive, screening, and engagement-related functions were rarely incorporated. Coding these interventions using the Behavior Change Technique Taxonomy version 1 (BCTTv1) indicated that the most frequently employed techniques were "4.1 Instruction on how to perform the behavior" and "5.1 Information about health consequences," suggesting that current DHIs primarily emphasize foundational behavioral support functions. Further frequency analysis of BCT coding among effective intervention studies (see Table 2) showed that BCTTv1 codes 4.1 (16/87, 18%), 5.1 (14/87, 16%), and 2.3 (13/87, 15%) constituted the core set of techniques, collectively accounting for nearly half of all techniques used in effective interventions.

Table 2. Frequency distribution of codes in effective intervention studies (N=87).

Behavior Change Technique Taxonomy version 1 code	Description	Frequency, n (%)
4.1	Instruction on how to perform behavior	16 (18)
5.1	Information about health consequences	14 (16)
2.3	Self-monitoring of behavior	13 (15)
2.2	Feedback on behavior	8 (9)
7.1	Prompts/cues	8 (9)
6.1	Demonstration of behavior	4 (5)
2.1	Monitoring by others (no feedback)	3 (3)
3.1	Social support (unspecified)	3 (3)
5.3	Social/environmental consequences	3 (3)
6.2	Social comparison	3 (3)
12.1	Restructuring physical environment	3 (3)
1.2	Problem solving	2 (2)
1.1	Goal setting (behavior)	1 (1)
1.6	Discrepancy between current behavior and goal	1 (1)
2.4	Self-monitoring of outcomes	1 (1)
2.6	Biofeedback	1 (1)
2.7	Feedback on outcomes	1 (1)
5.6	Emotional consequences	1 (1)
9.1	Credible source	1 (1)

Intervention Duration

The duration of interventions varied considerably across the included studies (see Figure 7; see also [10,11,13-15,17,20,26,29-66]), with the majority concentrated in the short- to medium-term range (1-16 weeks). Studies involving long-term interventions (>16 weeks) were relatively scarce, with only 4 studies identified. Among these studies, most incorporated follow-up periods, and medium- to long-term interventions were typically associated with more systematic follow-up protocols. With respect to study design, randomized controlled trials predominantly employed interventions of medium duration (8-16 weeks). Among the QDSs (n=4) and MMS (n=1) analyzed, some studies employed longer intervention durations to observe behavioral maintenance; however, these accounted for a relatively small proportion of the evidence base. Subgroup analysis demonstrated a progressive increase in the proportion of studies classified as "effective" with increasing intervention duration (see Table 3): 2 out of 4 (50.0%) for ultra-short-term (<1 week), 10 out of 16 (63%) for short-term (>1 and <8 weeks), 12 out of 18 (67%) for medium-term (8-16 weeks), and 3 out of 4 (75%) for long-term (>16 weeks). Notably, medium-duration interventions (8-16 weeks) not only represented the largest proportion of the existing evidence but also demonstrated both a relatively high "effective" rate (12/18, 67%) and a low "ineffective" rate (1/18, 6%). These findings indicate that current DHI research remains skewed toward short- and medium-term interventions, with the

XSL•FO
RenderX

8-16-week category standing out in terms of evidence volume and the apparent stability of intervention effects.

Figure 7. Chart of intervention duration and follow-up duration.

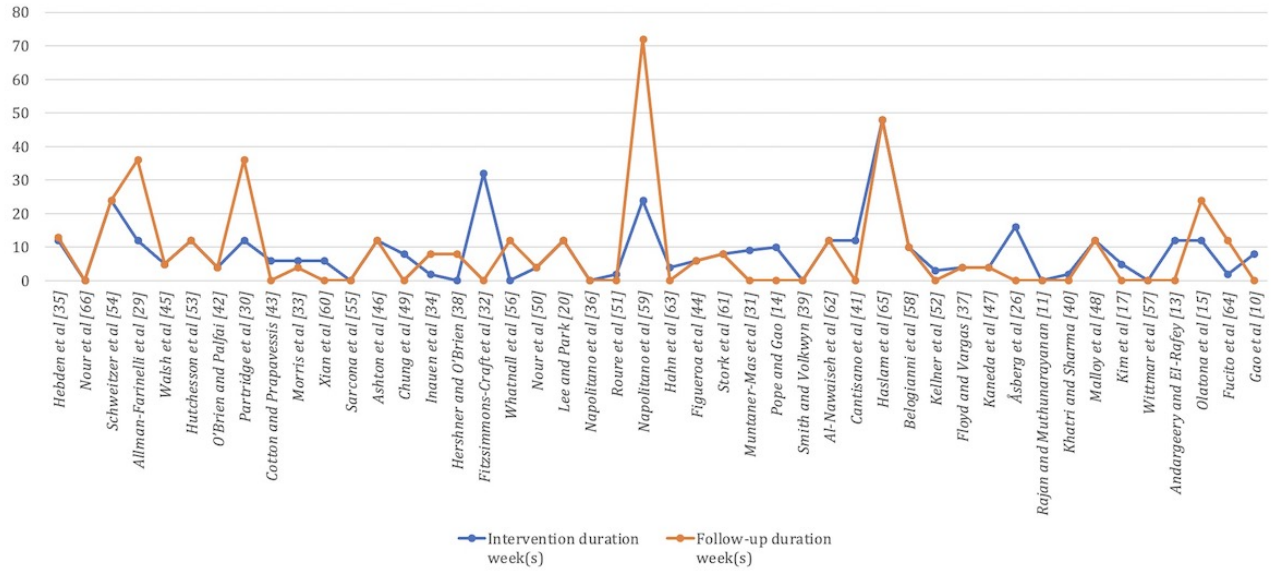

Table 3. Subgroup analysis of intervention duration.

Duration group (weeks)	Number, n	Effective (yes), n (%)	Limited effect, n (%)	Not effective (no), n (%)
Ultrashort (≤1)	4	2 (50.0)	2 (50.0)	0 (0)
Short (>1 and <8)	16	10 (63)	4 (25)	2 (13)
Medium (8-16)	18	12 (67)	5 (28)	1 (6)
Long (>16)	4	3 (75)	0 (0)	1 (25)
Subtotal (analyzed)	42	27 (64)	11 (26)	4 (10)
Excluded: not reported	4	N/A[a]	N/A	N/A

[a]N/A: not applicable.

Intervention Outcomes

As a result of substantial heterogeneity among the included studies with respect to outcome measurement instruments, outcome definitions, and assessment time points, it was not feasible to define a unified primary outcome or to conduct a statistically valid meta-analysis. Accordingly, this review adopted a descriptive synthesis framework to summarize and integrate the relevant outcomes. The outcome metrics in the included studies were classified into 2 main categories. The primary outcomes focused on lifestyle behaviors, including physical activity (eg, activity level, step count, and activity intensity), sedentary behaviors (eg, total sedentary time and frequency of breaks from sitting or resting), diet (eg, dietary quality; intake of fruits, vegetables, and sugar-sweetened beverages; energy intake; and nutritional knowledge), and sleep (eg, sleep quality, duration, efficiency, and severity of insomnia). These indicators directly reflect changes in core health behaviors resulting from the intervention and serve as a key basis for evaluating its effectiveness. Secondary outcomes, serving as supplementary indicators, were more diverse and encompassed physical health status and psychosocial dimensions, such as weight and body composition (eg, weight, BMI, waist circumference, and body fat percentage), physical fitness indicators (eg, flexibility, muscle strength, and cardiorespiratory fitness), cardiometabolic indicators (eg, blood pressure, blood glucose, and blood lipid profiles), and psychological and self-perception measures (eg, self-efficacy, body image, and life satisfaction). Overall, current studies remain primarily focused on primary outcomes, while secondary outcomes have expanded but continue to exhibit limited coverage.

Intervention Effectiveness

Based on the reported effect measure types, effect estimates, confidence levels (%), and CIs across the included studies, together with a comprehensive assessment of the authors' conclusions (see Multimedia Appendix 3), the results indicated that 31 (67%) studies demonstrated evidence of intervention effectiveness, suggesting that DHIs are generally associated with positive outcomes in improving lifestyle behaviors among college students. Four studies reported no statistically significant effects, with limitations primarily attributed to small sample sizes or short intervention durations. The remaining 11 studies demonstrated limited effectiveness, with improvements observed only in selected secondary outcomes or during short-term follow-up periods.

XSL•FO
RenderX

Based on a comprehensive assessment of each behavioral domain using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework, the certainty of evidence for the physical activity and diet domains was rated as "moderate," whereas the evidence for the sedentary behavior and sleep domains was rated as "low." With respect to evidence credibility, this review indicates a moderate level of confidence in the overall estimate that DHIs are effective in improving lifestyle behaviors among college students. The certainty of evidence in some domains was downgraded due to methodological limitations in the existing primary studies, including small sample sizes, challenges in implementing blinding, and inconsistencies in outcome assessment tools. Nevertheless, these GRADE assessments provide an accurate reflection of the current state of the evidence and its overall strength for DHIs among college students, thereby offering valuable guidance for interpretation and future research.

Discussion

Principal Findings

Discussion on Current Research Status

In terms of temporal trends, research on DHIs targeting college students' lifestyle behaviors has gradually emerged since 2014, expanded rapidly after 2016, and reached a peak in the past 5 years [6]. This trend has been driven primarily by 4 categories of factors. First, technological advances have laid a solid foundation for DHIs, with the proliferation of smartphones, wearable devices, and app ecosystems significantly enhancing their accessibility and operability [67]. Second, conceptual advancements have accelerated theoretical and methodological innovations in DHIs, underscoring their distinctive advantages in facilitating behavioral improvement [68]. Third, demand has increased substantially, particularly during the COVID-19 pandemic, with DHIs gaining broad recognition as viable alternatives when traditional approaches were constrained [14]. Fourth, resource investment has continued to expand, with funding, supportive policy frameworks, and interdisciplinary collaboration creating a favorable environment for research. Overall, future research is expected to shift from assessing short-term feasibility to evaluating long-term effectiveness, scalability, and the capacity to accommodate personalization [29,30].

In terms of spatial distribution, research on DHIs is predominantly concentrated in high-income countries, particularly in the United States and Australia. This concentration is primarily driven by a combination of technological infrastructure, research resources, and supportive policy environments. On the one hand, North America and Oceania initiated mHealth development relatively early, benefiting from substantial technological and financial advantages [31]. On the other hand, colleges in these regions generally possess mature health promotion systems and well-established ethical review mechanisms, facilitating the implementation of intervention trials. In addition, higher levels of health awareness and greater digital acceptance in Western cultures further contribute to this pattern. However, the generalizability of these findings may be limited when extrapolated to low- and middle-income countries. For example, resource-constrained settings may encounter infrastructural and hardware-related barriers, such as uneven network coverage and low rates of digital device ownership. Furthermore, substantial cross-cultural variations exist in perceptions of privacy, the role of family involvement, and prevailing health communication practices. In the future, cross-cultural validation and localized adaptation of DHIs should be strengthened [69,70], particularly in resource-constrained settings. Moreover, the development of low-cost, low-threshold DHI models should be explored to advance global health equity [11,71].

In terms of population structure, current research on DHIs has predominantly focused on generally healthy college students, a focus attributable to this group's modifiable health behaviors and susceptibility to environmental influences. However, some studies have extended to special populations, including college students with overweight or obesity, individuals at risk for eating disorders [15,32], and students experiencing sleep disorders or psychological stress [33,34]. This differentiation strategy is partly motivated by the fact that special populations face higher health risks, thereby increasing the potential benefits and clinical value of interventions [72,73]. It also aligns with the need for precision interventions and stratified management. However, existing research has not yet sufficiently examined variations in engagement levels among students from diverse sociodemographic backgrounds. Limited attention to factors such as socioeconomic status and access to digital devices may result in disproportionate benefits for students with greater financial or digital resources, while those experiencing economic constraints or limited device access may be marginalized in the intervention process. Accordingly, future research is likely to advance along 2 complementary directions: first, continuing large-scale studies targeting general undergraduate populations to assess the generalizability of interventions; and second, strengthening targeted interventions for high-risk groups while prioritizing the reduction of participation barriers among students from diverse backgrounds [74]. Such efforts may drive the development of DHIs toward greater refinement, equity, and personalization.

Discussion on Intervention Implementation

In terms of intervention objectives, physical activity and diet are the 2 lifestyle behavior categories receiving the most research attention [35], whereas sedentary behavior and sleep are relatively underrepresented. Both single-behavior and multibehavior combined interventions coexist. This pattern is primarily influenced by several factors. First, physical activity and diet are directly associated with weight management, energy balance, and metabolic health—core variables that affect college students' physical fitness and chronic disease risk [36]. Related measures (eg, step count, energy intake) are more easily quantifiable and standardized, making them more likely targets for intervention. By contrast, sedentary behavior and sleep, despite their recognized importance [37], pose technical and operational challenges for DHIs, including measurement complexity and delayed feedback on intervention effects [38], contributing to a relative paucity of research [39,40]. Current DHIs demonstrate limited effectiveness in reducing sedentary time among college students [75], whereas sleep interventions,

XSL•FO
RenderX

although promising, remain understudied and predominantly focus on insomnia relief [18,76]. Second, intervention strategies reflect researchers' assessment of behavioral variability: physical activity and diet exhibit a wide window for controllability and modification, whereas sedentary behavior typically occurs in academic or leisure contexts, complicating immediate adjustment via a single technique. The prevalence of single-behavior interventions is attributed to their suitability for early exploratory phases, allowing easier control of variables and validation of intervention effects. Conversely, the increase in multibehavior interventions reflects the aggregation of lifestyle risks among college students, which complicates achieving sufficient health benefits through changes in a single behavior. Notably, the combination of physical activity and diet is the most frequent, reflecting the necessity for integrated interventions targeting weight management and energy metabolism [41]. Overall, future research is expected to increasingly adopt multibehavior approaches, integrating behavioral science theories and technological tools to develop synergistic interventions that address the complexity of lifestyle risks.

In terms of intervention modalities, an evolutionary trend is evident, progressing from single to multiple formats and from low to high levels of interaction, driven by the combined forces of technological advancement, user demand, and intervention science. In early studies, SMS text messages and emails were the predominant forms of DHIs [42], owing to their low technological threshold, ease of deployment, and minimal cost, which made them suitable for rapid implementation in resource-limited contexts [43,44]. However, these approaches primarily involved 1-way information delivery, lacked personalization and real-time interaction, and were insufficient in maintaining user engagement. With the widespread adoption of smartphones and the maturation of the app ecosystem, mobile apps have gradually become the mainstream form of DHIs. These apps are highly integrated and interactive, capable of incorporating multiple functions such as goal-setting, feedback, reminders, and data tracking [13], aligning with college students' high-frequency mobile usage habits and significantly enhancing the intervention experience and engagement [45]. Web platforms retain advantages in scalability but are somewhat less user-friendly and less effective in delivering push notifications compared with apps [46,47]. In addition, the integration of social media and wearable devices increases the interactivity and contextual adaptability of DHIs [48], further enhancing behavioral monitoring and the provision of immediate feedback [49-51]. Future trends are expected to emphasize technological convergence and intelligent development. On the one hand, combinations of multiple formats (eg, apps, social platforms, and gamification) will become increasingly prevalent to address the multidimensional needs of behavioral interventions [52]. On the other hand, personalized interventions leveraging artificial intelligence, virtual coaching, and immersive experiences (eg, augmented reality/virtual reality) are anticipated to emerge as key research directions [17,53], shifting DHIs from being information-driven to experience-driven and ultimately facilitating sustained behavior change.

In terms of intervention functions, current DHIs are predominantly characterized by education, guidance, monitoring, and prompting components [54], indicating that these interventions primarily emphasize information delivery and basic behavior management. This design approach is partly driven by the substantial demand for health knowledge and skills among college students, with education and guidance functions facilitating improvements in cognition and self-efficacy. Concurrently, monitoring and prompting functions leverage technology to enable data recording and behavioral reinforcement, thereby promoting the initiation of target behaviors in the short term [55]. However, high-engagement features such as feedback, interaction, and gamification-based incentives remain underutilized [56,57], suggesting that DHIs often lack deep personalization and social support components [58], which may be a critical factor limiting long-term user engagement and intervention effectiveness. In terms of future trends, the convergence of behavior change theories (eg, Capability-Opportunity-Motivation-Behavior [COM-B], behavioral economics) with intelligent algorithmic applications is expected to drive the evolution of DHI functionality toward greater personalization, interactivity, and emotional engagement [59]. For example, artificial intelligence–driven real-time feedback could enhance intervention adaptability, virtual communities could strengthen social support, and gamification mechanisms coupled with reward systems could foster intrinsic motivation. Such advancements are likely to not only increase intervention engagement but also substantially improve behavioral maintenance, fostering a gradual shift from information delivery–oriented DHIs to approaches that place greater emphasis on user experience and social interaction.

Discussion on Intervention Effectiveness

Overall, 31 of 46 (67%) studies reported effective outcomes (yes), indicating the high feasibility and considerable potential of DHIs in improving the lifestyle behaviors of college students [60-62,77]. However, a subset of studies yielded insignificant (no) or limited (limited) effects, which can be examined from several analytical dimensions. First, insufficient refinement and lack of theoretical underpinning in intervention design represent key factors constraining effectiveness. In several cases, interventions lacked explicit theoretical frameworks for behavior change, relying predominantly on information delivery. Such approaches often failed to sufficiently stimulate participant motivation or reinforce behavior maintenance, leading to short-term gains that were difficult to sustain [26]. Second, existing intervention studies generally lack robust validation of long-term effects. Most studies are limited to durations of 8-16 weeks and include insufficient follow-up, which constrains the ability to verify the sustainability and stability of behavioral changes [63]. As a result, the long-term value and durability of DHIs remain difficult to assess adequately. Third, intervention effectiveness appears to be strongly influenced by participant adherence. Analyses of engagement-related metrics indicate that higher levels of user engagement, compliance, and intervention consistency are generally associated with more favorable behavioral and clinical outcomes. By contrast, studies characterized by high dropout rates often rely predominantly on 1-way information delivery, with limited opportunities for

XSL•FO
RenderX

feedback and interaction. Fourth, the type of target behavior and associated measurement challenges also contribute to these outcomes. Compared with physical activity, the intervention effects on diet, sedentary behavior, and sleep were more vulnerable to external environmental influences (eg, academic workload, dietary contexts), and measurement tools relied predominantly on self-reporting, thereby increasing bias and uncertainty. Taken together, variations in intervention design, technological application, and behavioral characteristics collectively contribute to the substantial heterogeneity observed in intervention outcomes [64].

To gain a deeper understanding of variations in intervention effectiveness, the COM-B framework can be applied as a systematic analytical tool [78]. (1) Within the "Capability" dimension, most interventions primarily enhanced college students' health-related knowledge through educational content and guidance materials. Examples included the provision of diet guidelines, exercise plans, and sleep regulation strategies designed to increase participants' awareness of the importance of healthy behaviors. However, these improvements often remained at the cognitive level, with limited emphasis on the development of practical behavioral skills. Specific components, such as diet substitution options, situational coping strategies, or flexible exercise planning, were frequently absent. In addition, some studies did not provide adequate support for data interpretation, which limited participants' ability to translate behavioral monitoring data into actionable steps [79]. (2) Within the "Opportunity" dimension, DHIs generally rely on virtual platforms to create enabling behavioral conditions, such as goal tracking, reminder functions, and online resource sharing, which may theoretically reduce psychological barriers to behavior enactment. However, the structuring of opportunities within real-world contexts remains insufficiently optimized. Some interventions do not adequately account for the distinctive time pressures and contextual constraints experienced by college students on campus. For example, strategies aimed at reducing sedentary behavior often remain limited to generic standing reminders, without adaptation to classroom environments or common study spaces, thereby constraining opportunities for sustained behavior change. Furthermore, although some interventions attempt to incorporate social support mechanisms (eg, community interactions or peer challenges), the depth and quality of participant engagement are generally limited. These interactions frequently involve 1-way information transmission, with limited capacity to foster emotional connection or effective behavioral modeling. (3) Within the "Motivation" dimension, existing interventions primarily emphasize the stimulation of extrinsic motivation through short-term incentives, such as point-based rewards and task completion reminders. While such strategies may promote initial engagement, they generally lack mechanisms for the sustained cultivation of intrinsic motivation. Specifically, many interventions have not effectively supported college students in developing a sense of self-worth derived from continued engagement in healthy behaviors. In addition, strategies aimed at enhancing positive emotional experiences are rarely incorporated. For example, gamification designs often remain confined to superficial point-based systems, with limited capacity to stimulate participants' sense of exploration, mastery, or accomplishment. Additionally, insufficient personalization of feedback appears to substantially constrain the maintenance of motivation over time. Participants often receive generic informational messages rather than timely, individualized feedback closely aligned with their actual behavioral performance.

In summary, current DHIs predominantly adopt a "technology-driven" or "utility-oriented" design logic, with a primary emphasis on functional implementation and surface-level engagement metrics. Because of the limited integration of behavior change theory, such interventions tend to exhibit constrained effectiveness in sustaining long-term outcomes. By contrast, theory-driven interventions—such as those grounded in the COM-B framework—extend beyond short-term behavior initiation, emphasizing the synergistic development and dynamic support of capability, opportunity, and motivation. Through structured and phased behavioral support strategies, such interventions may facilitate the establishment of enduring foundations for sustained change across cognitive, skill-based, environmental, and emotional dimensions [10]. As a result, long-term behavior maintenance may become more attainable [65]. Future research should further position behavior change theory as a central guiding principle in intervention design, moving beyond the view of technology as a standalone tool and instead embedding it organically within support systems centered on behavior change mechanisms.

Strengths and Limitations

This study is among the first English-language reviews to systematically integrate multiple forms of DHIs and multiple lifestyle behavior domains within a core population of college students, and it presents the following strengths. First, the study design strictly adheres to PRISMA 2020 and was preregistered on PROSPERO. The systematic search spanned 10 major international databases, ensuring the comprehensiveness and representativeness of the evidence base. Second, by focusing on college students as "digital natives," this study systematically analyzes intervention characteristics across 4 health behavior domains—physical activity, sedentary behavior, diet, and sleep—thereby addressing limitations of prior reviews that emphasized a single behavior or tool. Third, drawing on the COM-B framework, this study examines the mechanisms of DHIs across the Capability, Opportunity, and Motivation dimensions and identifies key bottlenecks in intervention strategies—such as limited technological functionality, suboptimal ecological adaptability, and insufficient motivational activation—thereby providing both theoretical support and practical guidance for the future design and optimization of DHIs for college students.

Although this review endeavored to incorporate the existing literature as comprehensively as possible, several limitations remain. First, the geographical distribution of the included studies was uneven, with a heavy concentration in high-income countries—particularly North America and Australia—which constrains the global applicability of the findings; specifically, their generalizability to college students in low- and middle-income countries requires empirical verification. Second, many studies employed small samples, short intervention durations, and limited follow-up, and some lacked robust control

XSL•FO
RenderX

groups or adequate randomization, thereby weakening the stability of effect estimates and the strength of causal inference. In addition, DHIs were often relatively homogeneous, with limited multidimensional interactivity and personalization; blinding procedures were difficult to implement; and risks of bias arose in adherence assessment and outcome measurement. Therefore, future research should strengthen sample representativeness, enhance intervention refinement, and improve methodological rigor to increase the external validity and practical utility of the findings.

Implications and Recommendations

Recommendations for Policy and Practice

To fully realize the potential of DHIs while ensuring the sustainability and broad accessibility of intervention effects, systematic improvements in policy design and implementation pathways are required. First, college students should be explicitly incorporated into national and regional digital health strategies to facilitate a shift from traditional health education toward integrated digital platforms, and higher-education institutions should be encouraged to develop or adopt scientifically grounded, standardized tools with clearly articulated mechanisms of action. Second, localized development of intervention content and functionality should be supported, with attention to adaptability across behavioral domains, cultural contexts, and student needs, thereby advancing refined, human-centered design with respect to technological thresholds, data security, and personalized recommendations. Third, intervention practice should strengthen students' active engagement and establish feedback-driven, behavior-reinforcing, and peer-support mechanisms to enhance sustained use and intrinsic motivation. In parallel, cross-departmental cooperation mechanisms should be established at the college level, and health interventions should be embedded within curricula, psychological support systems, and campus service resources to form a synergistic support network. Finally, at the policy level, ethical oversight and effectiveness evaluation of DHIs programs should be strengthened, and an evidence-based evaluation framework for DHIs should be established to ensure fairer, more adaptable, and more effective interventions for college students.

Recommendations for Future Research

This study indicates that current research on DHIs for college students remains constrained by unrepresentative samples, single-focus intervention content, and unclear technological mechanisms; future work should be refined and deepened in the following respects. First, geographical and cultural diversity should be expanded, prioritizing studies from low- and middle-income countries, varied higher-education institution types, and diverse social groups to enhance the external validity of the findings. Second, the design and evaluation of multibehavior-integrated interventions should be strengthened by moving beyond single-behavior paradigms and examining behavioral synergies and optimal combinations of intervention components. Third, higher-quality study designs—such as QRCTs, MMSs, and long-term follow-up—should be employed to strengthen causal inference and the sustainability of intervention effects. Fourth, theoretical development and empirical testing of intervention mechanisms should be strengthened by grounding analyses in behavior change theory to clarify how technology enhances Capability, Opportunity, and Motivation, and to advance DHIs from merely providing technical functions to creating a supportive ecosystem conducive to sustained behavior change. Finally, future studies should emphasize the assessment of intervention equity, systematically account for potential moderators such as gender, socioeconomic status, and psychological status, and identify subgroups with limited responsiveness, thereby providing a robust evidence base for constructing a more inclusive and adaptive DHI model for college students.

Conclusions

This review addresses a gap in the literature by focusing specifically on college students, a group often overlooked in research that typically centers on broader adult populations. Unlike prior reviews that mainly examine a single lifestyle behavior, this study adopts a more holistic approach by integrating multiple behaviors and evaluating a range of DHIs with diverse modalities and functionalities. These findings provide valuable insights for refining future DHIs targeting college students and contribute to the development of more effective health promotion strategies in higher education. Although DHIs show potential for improving lifestyle behaviors, their long-term effectiveness remains uncertain. Current interventions face several limitations, including a narrow behavioral focus, basic technological functionality, and limited adaptability to diverse contexts, all of which may restrict long-term engagement and personalized responsiveness. Moreover, many interventions do not fully account for variations in resource access and individual behavior change pathways, potentially limiting their applicability and equity. Future research should prioritize integrating multiple behaviors, enhancing user engagement, improving contextual adaptability, and expanding technological accessibility. Long-term studies and equity-focused evaluations are essential for strengthening the evidence base and ensuring the sustainability and inclusivity of health behavior change among college students.

Acknowledgments

We are grateful to all participants in this study for their time, effort, and dedication. The authors declare that no generative artificial intelligence was used in the preparation of this manuscript.

XSL•FO
RenderX

Funding

This research was funded by the Post-Funded Project of China National Social Science Foundation (grant 25FTYB012), the General Project of Shanghai Philosophy and Social Science Foundation (grant 2025BTY002), and the Youth Project of Shanghai Eastern Talent Plan (grant QNJY2025162).

Data Availability

Data are provided within the manuscript or multimedia appendices; data/materials can be shared upon reasonable request to the corresponding author (ZHY).

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 28 KB - jmir_v28i1e82192_app1.docx]

Multimedia Appendix 2
Search strategy used in this review.
[DOCX File , 19 KB - jmir_v28i1e82192_app2.docx]

Multimedia Appendix 3
Comprehensive data extraction table of included studies.
[XLSX File (Microsoft Excel File), 44 KB - jmir_v28i1e82192_app3.xlsx]

References

1. Peng S, Yuan F, Othman AT, Zhou X, Shen G, Liang J. The effectiveness of e-Health interventions promoting physical activity and reducing sedentary behavior in college students: a systematic review and meta-analysis of randomized controlled trials. Int J Environ Res Public Health 2022 Dec 25;20(1):318-318 [FREE Full text] [doi: 10.3390/ijerph20010318] [Medline: 36612643]
2. Huang T, Zheng K, Li S, Yang Y, Kong L, Zhao Y. Screen-based sedentary behaviors but not total sedentary time are associated with anxiety among college students. Front Public Health 2022 Oct 20;10:994612-994612 [FREE Full text] [doi: 10.3389/fpubh.2022.994612] [Medline: 36339232]
3. Feng J, Lau PWC, Shi L, Huang WY. Longitudinal shifts in movement behaviors during the COVID-19 pandemic: relations to posttraumatic stress disorder among university students. Int J Environ Res Public Health 2022 Oct 18;19(20):13449-13449 [FREE Full text] [doi: 10.3390/ijerph192013449] [Medline: 36294027]
4. Barbosa BCR, Menezes-Júnior LAAD, de Paula W, Chagas CMDS, Machado EL, de Freitas ED, et al. Sedentary behavior is associated with the mental health of university students during the Covid-19 pandemic, and not practicing physical activity accentuates its adverse effects: cross-sectional study. BMC Public Health 2024 Jul 11;24(1):1860 [FREE Full text] [doi: 10.1186/s12889-024-19345-5] [Medline: 38992623]
5. Bi S, Yuan J, Wang Y, Zhang W, Zhang L, Zhang Y, et al. Effectiveness of digital health interventions in promoting physical activity among college students: systematic review and meta-analysis. J Med Internet Res 2024 Nov 20;26:e51714-e51714 [FREE Full text] [doi: 10.2196/51714] [Medline: 39566049]
6. Taylor ME, Liu M, Abelson S, Eisenberg D, Lipson SK, Schueller SM. The reach, effectiveness, adoption, implementation, and maintenance of digital mental health interventions for college students: a systematic review. Curr Psychiatry Rep 2024 Oct 11;26(12):683-693 [FREE Full text] [doi: 10.1007/s11920-024-01545-w] [Medline: 39392547]
7. Gunasekeran DV, Tseng RMWW, Tham Y, Wong TY. Applications of digital health for public health responses to COVID-19: a systematic scoping review of artificial intelligence, telehealth and related technologies. NPJ Digit Med 2021 Feb 26;4(1):40-40 [FREE Full text] [doi: 10.1038/s41746-021-00412-9] [Medline: 33637833]
8. Keesara S, Jonas A, Schulman K. Covid-19 and health care's digital revolution. N Engl J Med 2020 Jun 04;382(23):e82. [doi: 10.1056/NEJMp2005835] [Medline: 32240581]
9. Tan SY, Sumner J, Wang Y, Wenjun Yip A. A systematic review of the impacts of remote patient monitoring (RPM) interventions on safety, adherence, quality-of-life and cost-related outcomes. NPJ Digit Med 2024 Jul 18;7(1):192-192 [FREE Full text] [doi: 10.1038/s41746-024-01182-w] [Medline: 39025937]
10. Gao Y, Zhang J, He Z, Zhou Z. Feasibility and usability of an artificial intelligence-powered gamification intervention for enhancing physical activity among college students: quasi-experimental study. JMIR Serious Games 2025 Mar 24;13:e65498-e65498 [FREE Full text] [doi: 10.2196/65498] [Medline: 40127464]

XSL•FO
RenderX

11. Rajan R, Muthunarayanan L. A comparative study on the prevalence of lifestyle risk factors among users and non-users of any health-based mobile app among university students in Chennai. Cureus 2024 Mar 15;16(3):e56203-e56203 [FREE Full text] [doi: 10.7759/cureus.56203] [Medline: 38618332]
12. Remskar M, Ainsworth B, Maynard OM, Malkowski OS, Birch A, Burd AK, et al. Getting active through mindfulness: randomised controlled trial of a digital mindfulness-based intervention promoting physical activity engagement and enjoyment. Mental Health and Physical Activity 2025 Mar 08;28:100680-100680. [doi: 10.1016/j.mhpa.2025.100680]
13. Andargeery SY, El-Rafey DS. A randomized controlled trial of the effectiveness of the mHealth program in improving the lifestyle of nursing students. Sci Rep 2025 Mar 21;15(1):9765-9765 [FREE Full text] [doi: 10.1038/s41598-024-80982-2] [Medline: 40118869]
14. Pope ZC, Gao Z. Feasibility of smartphone application- and social media-based intervention on college students' health outcomes: a pilot randomized trial. J Am Coll Health 2022 Jan 09;70(1):89-98. [doi: 10.1080/07448481.2020.1726925] [Medline: 32150514]
15. Olatona FA, Van Onselen A, Kolanisi U. The effect of online nutrition and lifestyle education on body image dissatisfaction, body mass index, and disordered eating among female university undergraduate students in Lagos, Nigeria. South African Journal of Clinical Nutrition 2025 Mar 05;38(1):15-22. [doi: 10.1080/16070658.2024.2439752]
16. Seid A, Fufa DD, Bitew ZW. The use of internet-based smartphone apps consistently improved consumers' healthy eating behaviors: a systematic review of randomized controlled trials. Front Digit Health 2024 Jan 12;6:1282570-1282570 [FREE Full text] [doi: 10.3389/fdgth.2024.1282570] [Medline: 38283582]
17. Kim K, Hur M, Kim W. Effects of virtual reality (VR)-based meditation on sleep quality, stress, and autonomic nervous system balance in nursing students. Healthcare (Basel) 2024 Aug 08;12(16):1581-1581 [FREE Full text] [doi: 10.3390/healthcare12161581] [Medline: 39201140]
18. Lu Y, Lin H, Tsai P. Effects of digital sleep interventions on sleep among college students and young adults: systematic review and meta-analysis. J Med Internet Res 2025 May 12;27:e69657-e69657 [FREE Full text] [doi: 10.2196/69657] [Medline: 40354636]
19. Kechagias EP, Papadopoulos GA, Rokai I. Evaluating the impact of digital health interventions on workplace health outcomes: a systematic review. Administrative Sciences 2024 Jun 20;14(6):131-131. [doi: 10.3390/admsci14060131]
20. Lee S, Park J. Systematic review and meta analysis of standalone digital behavior change interventions on physical activity. NPJ Digit Med 2025 Jul 14;8(1):436-436 [FREE Full text] [doi: 10.1038/s41746-025-01827-4] [Medline: 40659761]
21. Milne-Ives M, Homer S, Andrade J, Meinert E. Associations between behavior change techniques and engagement with mobile health apps: protocol for a systematic review. JMIR Res Protoc 2022 Mar 29;11(3):e35172-e35172 [FREE Full text] [doi: 10.2196/35172] [Medline: 35348460]
22. Milne-Ives M, Homer SR, Andrade J, Meinert E. Potential associations between behavior change techniques and engagement with mobile health apps: a systematic review. Front Psychol 2023 Sep 18;14:1227443-1227443 [FREE Full text] [doi: 10.3389/fpsyg.2023.1227443] [Medline: 37794916]
23. Gabarron E, Larbi D, Rivera-Romero O, Denecke K. Human factors in AI-driven digital solutions for increasing physical activity: scoping review. JMIR Hum Factors 2024 Jul 03;11:e55964-e55964 [FREE Full text] [doi: 10.2196/55964] [Medline: 38959064]
24. Ferrari M, Allan S, Arnold C, Eleftheriadis D, Alvarez-Jimenez M, Gumley A, et al. Digital interventions for psychological well-being in university students: systematic review and meta-analysis. J Med Internet Res 2022 Sep 28;24(9):e39686-e39686 [FREE Full text] [doi: 10.2196/39686] [Medline: 36169988]
25. Matos Fialho PM, Wenig V, Heumann E, Müller M, Stock C, Pischke CR. Digital public health interventions for the promotion of mental well-being and health behaviors among university students: a rapid review. BMC Public Health 2025 Jul 18;25(1):2500-2500 [FREE Full text] [doi: 10.1186/s12889-025-23669-1] [Medline: 40681968]
26. Åsberg K, Eldh AC, Löf M, Bendtsen M. "Simply complicated": uncovering the processes of lifestyle behavior change among college and university students with access to a digital multiple lifestyle intervention. Digit Health 2024 Apr 09;10:20552076241245905 [FREE Full text] [doi: 10.1177/20552076241245905] [Medline: 38601184]
27. Ta A, Salgin N, Demir M. Real-time stress monitoring, detection, and management in college students: a wearable technology and machine-learning approach. arXiv. Preprint posted online on May 21, 2025 [FREE Full text] [doi: 10.48550/arXiv.2505.15974]
28. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. EFI 2018 Dec 18;34(4):285-291. [doi: 10.3233/EFI-180221]
29. Allman-Farinelli M, Partridge SR, McGeechan K, Balestracci K, Hebden L, Wong A, et al. A mobile health lifestyle program for prevention of weight gain in young adults (TXT2BFiT): nine-month outcomes of a randomized controlled trial. JMIR Mhealth Uhealth 2016 Jun 22;4(2):e78-e78 [FREE Full text] [doi: 10.2196/mhealth.5768] [Medline: 27335237]
30. Partridge SR, McGeechan K, Bauman A, Phongsavan P, Allman-Farinelli M. Improved eating behaviours mediate weight gain prevention of young adults: moderation and mediation results of a randomised controlled trial of TXT2BFiT, mHealth program. Int J Behav Nutr Phys Act 2016 Apr 02;13:44-44 [FREE Full text] [doi: 10.1186/s12966-016-0368-8] [Medline: 27039178]

XSL•FO
RenderX

31. Muntaner-Mas A, Sanchez-Azanza VA, Ortega FB, Vidal-Conti J, Borràs PA, Cantallops J, et al. The effects of a physical activity intervention based on a fatness and fitness smartphone app for university students. Health Informatics J 2021;27(1):1460458220987275-1460458220987275 [FREE Full text] [doi: 10.1177/1460458220987275] [Medline: 33446036]
32. Fitzsimmons-Craft EE, Firebaugh M, Graham AK, Eichen DM, Monterubio GE, Balantekin KN, et al. State-wide university implementation of an online platform for eating disorders screening and intervention. Psychol Serv 2018 Nov 08;16(2):239-249 [FREE Full text] [doi: 10.1037/ser0000264] [Medline: 30407047]
33. Morris J, Firkins A, Millings A, Mohr C, Redford P, Rowe A. Internet-delivered cognitive behavior therapy for anxiety and insomnia in a higher education context. Anxiety Stress Coping 2015 Jul 20;29(4):415-431. [doi: 10.1080/10615806.2015.1058924] [Medline: 26079158]
34. Inauen J, Bolger N, Shrout PE, Stadler G, Amrein M, Rackow P, et al. Using smartphone-based support groups to promote healthy eating in daily life: a randomised trial. Appl Psychol Health Well Being 2017 Sep 25;9(3):303-323. [doi: 10.1111/aphw.12093] [Medline: 28948690]
35. Hebden L, Cook A, van DPHP, King L, Bauman A, Allman-Farinelli M. A mobile health intervention for weight management among young adults: a pilot randomised controlled trial. J Hum Nutr Diet 2013 Aug 29;27(4):322-332. [doi: 10.1111/jhn.12155] [Medline: 23992038]
36. Napolitano MA, Lynch SB, Mavredes M, Shambon B, Posey L. Evaluating an interactive digital intervention for college weight gain prevention. J Nutr Educ Behav 2020 May 23;52(9):890-897. [doi: 10.1016/j.jneb.2020.04.007] [Medline: 32456988]
37. Floyd V, Vargas I. Evaluating a mobile application based intervention for insomnia in college students: a preliminary study. J Am Coll Health 2024 Nov 08;73(9):3540-3548. [doi: 10.1080/07448481.2024.2423225] [Medline: 39514819]
38. Hershner S, O'Brien LM. The impact of a randomized sleep education intervention for college students. J Clin Sleep Med 2018 Mar 15;14(3):337-347 [FREE Full text] [doi: 10.5664/jcsm.6974] [Medline: 29510791]
39. Smith L, Volkwyn C. Promotion of physical activity through health applications among students of selected universities: a preliminary study. Cent Eur J Sport Sci Med 2022 Feb 01;38(2):5-12. [doi: 10.18276/cej.2022.2-01]
40. Khatri S, Sharma R. Effective management of sedentary behavior among Indian university students: an empirical exploration into health-related behavior. J Educ Health Promot 2024 Apr 29;13:131-131 [FREE Full text] [doi: 10.4103/jehp.jehp_1489_23] [Medline: 38784278]
41. Cantisano LM, Gonzalez-Soltero R, Blanco-Fernández A, Belando-Pedreño N. ePSICONUT: An e-health programme to improve emotional health and lifestyle in university students. Int J Environ Res Public Health 2022 Jul 28;19(15):9253-9253 [FREE Full text] [doi: 10.3390/ijerph19159253] [Medline: 35954601]
42. O'Brien LM, Palfai TP. Efficacy of a brief web-based intervention with and without SMS to enhance healthy eating behaviors among university students. Eat Behav 2016 Sep 01;23:104-109. [doi: 10.1016/j.eatbeh.2016.08.012] [Medline: 27619174]
43. Cotten E, Prapavessis H. Increasing nonsedentary behaviors in university students using text messages: randomized controlled trial. JMIR Mhealth Uhealth 2016 Aug 19;4(3):e99 [FREE Full text] [doi: 10.2196/mhealth.5411] [Medline: 27543317]
44. Figueroa CA, Deliu N, Chakraborty B, Modiri A, Xu J, Aggarwal J, et al. Daily motivational text messages to promote physical activity in university students: results from a microrandomized trial. Ann Behav Med 2022 Feb 11;56(2):212-218. [doi: 10.1093/abm/kaab028] [Medline: 33871015]
45. Walsh JC, Corbett T, Hogan M, Duggan J, McNamara A. An mHealth intervention using a smartphone app to increase walking behavior in young adults: a pilot study. JMIR Mhealth Uhealth 2016 Sep 22;4(3):e109 [FREE Full text] [doi: 10.2196/mhealth.5227] [Medline: 27658677]
46. Ashton LM, Morgan PJ, Hutchesson MJ, Rollo ME, Collins CE. Feasibility and preliminary efficacy of the 'HEYMAN' healthy lifestyle program for young men: a pilot randomised controlled trial. Nutr J 2017 Jan 13;16(1):2 [FREE Full text] [doi: 10.1186/s12937-017-0227-8] [Medline: 28086890]
47. Kaneda K, Maeda N, Fukui K, Tashiro T, Komiya M, Urabe Y. Impact of the simultaneous distribution of e-learning and exercise videos on the health literacy and lifestyle of college students during the COVID-19 pandemic: a randomized controlled trial. J Phys Ther Sci 2024 Nov 01;36(11):703-710 [FREE Full text] [doi: 10.1589/jpts.36.703] [Medline: 39493688]
48. Malloy JA, Partridge SR, Kemper JA, Braakhuis A, Roy R. Feasibility and preliminary efficacy of co-designed and co-created healthy lifestyle social media intervention programme the daily health coach for young women: a pilot randomised controlled trial. Nutrients 2024 Dec 18;16(24):4364 [FREE Full text] [doi: 10.3390/nu16244364] [Medline: 39770984]
49. Chung AE, Skinner AC, Hasty SE, Perrin EM. Tweeting to health: a novel mHealth intervention using Fitbits and Twitter to foster healthy lifestyles. Clin Pediatr (Phila) 2016 Jul 16:26-32. [doi: 10.1177/0009922816653385] [Medline: 27317609]
50. Nour M, Chen J, Allman-Farinelli M. Young adults' engagement with a self-monitoring app for vegetable intake and the impact of social media and gamification: feasibility study. JMIR Form Res 2019 May 10;3(2):e13324 [FREE Full text] [doi: 10.2196/13324] [Medline: 31094322]
51. Roure C, Pasco D, Benoît N, Deldicque L. Impact of a design-based bike exergame on young adults' physical activity metrics and situational interest. Res Q Exerc Sport 2019 Nov 13;91(2):309-315. [doi: 10.1080/02701367.2019.1665621] [Medline: 31718499]

XSL•FO
RenderX

52. Kellner M, Dold C, Lohkamp M. Objectively assessing the effect of a messenger-based intervention to reduce sedentary behavior in university students: a pilot study. J Prev 2023 May 12;44(5):521-534 [FREE Full text] [doi: 10.1007/s10935-023-00735-1] [Medline: 37171555]
53. Hutchesson MJ, Morgan PJ, Callister R, Pranata I, Skinner G, Collins CE. Be positive be healthe: development and implementation of a targeted e-Health weight loss program for young women. Telemed J E Health 2015 Dec 24;22(6):519-528. [doi: 10.1089/tmj.2015.0085] [Medline: 26701611]
54. Schweitzer AL, Ross JT, Klein CJ, Lei KY, Mackey ER. An electronic wellness program to improve diet and exercise in college students: a pilot study. JMIR Res Protoc 2016 Feb 29;5(1):e29 [FREE Full text] [doi: 10.2196/resprot.4855] [Medline: 26929118]
55. Sarcona A, Kovacs L, Wright J, Williams C. Differences in eating behavior, physical activity, and health-related lifestyle choices between users and nonusers of mobile health apps. American Journal of Health Education 2017 Jul 11;48(5):298-305. [doi: 10.1080/19325037.2017.1335630]
56. Whatnall MC, Patterson AJ, Chiu S, Oldmeadow C, Hutchesson MJ. Feasibility and preliminary efficacy of the Eating Advice to Students (EATS) brief web-based nutrition intervention for young adult university students: a pilot randomized controlled trial. Nutrients 2019 Apr 23;11(4):905 [FREE Full text] [doi: 10.3390/nu11040905] [Medline: 31018565]
57. Wittmar S, Frankenstein T, Timm V, Frei P, Kurpiers N, Wölwer S, et al. User experience with a personalized mHealth service for physical activity promotion in university students: mixed methods study. JMIR Form Res 2025 Mar 28;9:e64384 [FREE Full text] [doi: 10.2196/64384] [Medline: 40153787]
58. Belogianni K, Ooms A, Lykou A, Nikoletou D, Jayne Moir H. An online game-based intervention using quizzes to improve nutrition and physical activity outcomes among university students. Health Education Journal 2023 Jun 09;82(6):636-650. [doi: 10.1177/00178969231179032]
59. Napolitano MA, Whiteley JA, Mavredes M, Tjaden AH, Simmens S, Hayman LL, et al. Effect of tailoring on weight loss among young adults receiving digital interventions: an 18 month randomized controlled trial. Transl Behav Med 2021 Apr 26;11(4):970-980 [FREE Full text] [doi: 10.1093/tbm/ibab017] [Medline: 33739422]
60. Xian Y, Xu H, Xu H, Liang L, Hernandez AF, Wang TY, et al. An initial evaluation of the impact of Pokémon GO on physical activity. J Am Heart Assoc 2017 May 16;6(5):e005341 [FREE Full text] [doi: 10.1161/JAHA.116.005341] [Medline: 28512111]
61. Stork MJ, Bell EG, Jung ME. Examining the impact of a mobile health app on functional movement and physical fitness: pilot pragmatic randomized controlled trial. JMIR Mhealth Uhealth 2021 May 28;9(5):e24076 [FREE Full text] [doi: 10.2196/24076] [Medline: 34047704]
62. Al-Nawaiseh HK, McIntosh WA, McKyer LJ. An-m-Health intervention using smartphone app to improve physical activity in college students: a randomized controlled trial. Int J Environ Res Public Health 2022 Jun 13;19(12):7228 [FREE Full text] [doi: 10.3390/ijerph19127228] [Medline: 35742477]
63. Hahn SL, Kaciroti N, Eisenberg D, Weeks HM, Bauer KW, Sonneville KR. Introducing dietary self-monitoring to undergraduate women via a calorie counting app has no effect on mental health or health behaviors: results from a randomized controlled trial. J Acad Nutr Diet 2021 Aug 20;121(12):2377-2388 [FREE Full text] [doi: 10.1016/j.jand.2021.06.311] [Medline: 34427188]
64. Fucito LM, Ash GI, Wu R, Pittman B, Barnett NP, Li CR, et al. Wearable intervention for alcohol use risk and sleep in young adults: a randomized clinical trial. JAMA Netw Open 2025 May 01;8(5):e2513167. [doi: 10.1001/jamanetworkopen.2025.13167] [Medline: 40445615]
65. Haslam RL, Baldwin JN, Pezdirc K, Truby H, Attia J, Hutchesson MJ, et al. Efficacy of technology-based personalised feedback on diet quality in young Australian adults: results for the advice, ideas and motivation for my eating (Aim4Me) randomised controlled trial. Public Health Nutr 2023 Feb 09;26(6):1293-1305. [doi: 10.1017/S1368980023000253] [Medline: 36755380]
66. Nour MM, McGeechan K, Wong AT, Partridge SR, Balestracci K, Roy R, et al. Diet quality of young adults enrolling in TXT2BFiT, a mobile phone-based healthy lifestyle intervention. JMIR Res Protoc 2015 May 27;4(2):e60 [FREE Full text] [doi: 10.2196/resprot.4484] [Medline: 26018723]
67. Oliveira C, Pereira A, Vagos P, Nóbrega C, Gonçalves J, Afonso B. Effectiveness of mobile app-based psychological interventions for college students: a systematic review of the literature. Front Psychol 2021 May 11;12:647606-647606 [FREE Full text] [doi: 10.3389/fpsyg.2021.647606] [Medline: 34045994]
68. Konlan KD, Ibrahim ZA, Lee J, Lee H. The inclusion of implementation outcomes in digital health interventions for young adults: a scoping review. Digit Health 2025 Mar 28;11:20552076251330194 [FREE Full text] [doi: 10.1177/20552076251330194] [Medline: 40162162]
69. Spanhel K, Balci S, Feldhahn F, Bengel J, Baumeister H, Sander LB. Cultural adaptation of internet- and mobile-based interventions for mental disorders: a systematic review. NPJ Digit Med 2021 Aug 25;4(1):128-128 [FREE Full text] [doi: 10.1038/s41746-021-00498-1] [Medline: 34433875]
70. Whitehead L, Talevski J, Fatehi F, Beauchamp A. Barriers to and facilitators of digital health among culturally and linguistically diverse populations: qualitative systematic review. J Med Internet Res 2023 Feb 28;25:e42719-e42719 [FREE Full text] [doi: 10.2196/42719] [Medline: 36853742]

XSL•FO
RenderX

71. Kim J, Aryee LMD, Bang H, Prajogo S, Choi YK, Hoch JS, et al. Effectiveness of digital mental health tools to reduce depressive and anxiety symptoms in low- and middle-income countries: systematic review and meta-analysis. JMIR Ment Health 2023 Mar 20;10:e43066-e43066 [FREE Full text] [doi: 10.2196/43066] [Medline: 36939820]
72. Hayes JF, Darling KE, Tomashek H, Elwy AR, Wing RR. Behavioral weight loss interventions in college health centers: a qualitative analysis of barriers and facilitators to implementation. Obes Sci Pract 2024 Dec 14;10(6):e70021-e70021. [doi: 10.1002/osp4.70021] [Medline: 39544506]
73. Hernández-Jaña S, Huber-Pérez T, Palma-Leal X, Guerrero-Ibacache P, Campos-Nuñez V, Zavala-Crichton JP, et al. Effect of a single nutritional intervention previous to a critical period of fat gain in university students with overweight and obesity: a randomized controlled trial. Int J Environ Res Public Health 2020 Jul 16;17(14):5149-5149 [FREE Full text] [doi: 10.3390/ijerph17145149] [Medline: 32708831]
74. Berman AH, Topooco N, Lindfors P, Bendtsen M, Lindner P, Molander O, et al. Transdiagnostic and tailored internet intervention to improve mental health among university students: research protocol for a randomized controlled trial. Trials 2024 Mar 01;25(1):158-158 [FREE Full text] [doi: 10.1186/s13063-024-07986-1] [Medline: 38429834]
75. Iwakura M, Ozeki C, Jung S, Yamazaki T, Miki T, Nohara M, et al. An umbrella review of efficacy of digital health interventions for workers. NPJ Digit Med 2025 Apr 14;8(1):207-207 [FREE Full text] [doi: 10.1038/s41746-025-01578-2] [Medline: 40229460]
76. Vollert B, Müller L, Jacobi C, Trockel M, Beintner I. Effectiveness of an app-based short intervention to improve sleep: randomized controlled trial. JMIR Ment Health 2023 Mar 21;10:e39052-e39052 [FREE Full text] [doi: 10.2196/39052] [Medline: 36943337]
77. Lee J, Kang M, Lee S. Effects of the e-Motivate4Change program on metabolic syndrome in young adults using health apps and wearable devices: quasi-experimental study. J Med Internet Res 2020 Jul 30;22(7):e17031 [FREE Full text] [doi: 10.2196/17031] [Medline: 32729838]
78. Daryabeygi-Khotbehsara R, Dunstan DW, Shariful Islam SM, Rhodes RE, Hojjatinia S, Abdelrazek M, et al. A control system model of capability-opportunity-motivation and behaviour (COM-B) framework for sedentary and physical activity behaviours. Digit Health 2024 Jun 07;10:20552076241255658 [FREE Full text] [doi: 10.1177/20552076241255658] [Medline: 38854921]
79. Paterson S, Dawes H, Winward C, Bartram E, Dodds E, McKinon J, et al. Use of the Capability, Opportunity and Motivation Behaviour model (COM-B) to understand interventions to support physical activity behaviour in people with stroke: an overview of reviews. Clin Rehabil 2024 Jan 09;38(4):543-557 [FREE Full text] [doi: 10.1177/02692155231224365] [Medline: 38192225]

Abbreviations

BCTTv1: Behavior Change Technique Taxonomy version 1
COM-B: Capability-Opportunity-Motivation-Behavior
DHI: digital health intervention
GRADE: Grading of Recommendations Assessment, Development and Evaluation
JBI: Joanna Briggs Institute
mHealth: mobile health
MMAT: Mixed Methods Appraisal Tool
MMS: mixed methods study
PICOS: Population-Intervention-Comparison-Outcome-Study Design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Search Extension
PROGRESS-Plus: Place of Residence, Race/Ethnicity, Occupation, Gender/Sex, Religion, Education, Socioeconomic Status, Plus Other Relevant Factors
PROSPERO: International Prospective Register of Systematic Reviews
QDS: quantitative descriptive study
QNRS: quantitative nonrandomized study
QR: qualitative research
QRCT: quantitative randomized controlled trial
RoB 2: Risk of Bias 2

XSL•FO
RenderX

Edited by S Brini; submitted 11.Aug.2025; peer-reviewed by GLP Bodagala, F Owoseje, Y Elsanousi, M Ekwueme, H Kitiabi; comments to author 27.Oct.2025; accepted 05.Jan.2026; published 04.Feb.2026.

Please cite as:
Zhou Q, Jiang J, Yin Z, Fan R
Effect of Digital Health Interventions on College Students' Lifestyle Behaviors: Systematic Review
J Med Internet Res 2026;28:e82192
URL: https://www.jmir.org/2026/1/e82192
doi:10.2196/82192
PMID:

©Qingyuan Zhou, Jiajun Jiang, Zhihua Yin, Ruishi Fan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Key Components and Barriers in Web-Based Suicide Prevention Gatekeeper Training: Systematic Narrative Review

Olivier Ferlatte[1,2], PhD; Emmanuelle Gareau[1,2], MSc; Keven Lee[1,2], PhD; Kinda Wassef[1,2], MPH; John Lindsay Oliffe[3,4], PhD; Hannah Kia[5], PhD; Brock Dumville[6], MPH

[1]École de santé publique de l'Université de Montréal, Montréal, QC, Canada

[2]Centre de recherche en santé publique, Montréal, QC, Canada

[3]School of Nursing, University of British Columbia, Vancouver, BC, Canada

[4]Department of Nursing, University of Melbourne, Melbourne, Australia

[5]School of Social Work, University of British Columbia, Vancouver, QC, Canada

[6]Suicide Prevention Centre of Montreal, Montréal, QC, Canada

Corresponding Author:
Olivier Ferlatte, PhD
École de santé publique de l'Université de Montréal
7101 Av. du Parc
Montréal, QC, H3N 1X9
Canada
Phone: 1 514 343 6111 ext 44090
Email: olivier.ferlatte@umontreal.ca

Abstract

Background: Gatekeeper training programs (GTPs) are a key component of contemporary suicide prevention strategies, equipping community members and non–mental health professionals with the skills to identify, engage with, and refer individuals at risk of suicide. Increasingly, these programs are delivered via the web, offering a compelling alternative to in-person training through greater scalability, flexibility, and cost-effectiveness. However, little consensus exists regarding the design, modes of delivery, and implementation strategies of web-based GTPs. Further, there is a limited understanding of which components affect their usability and engagement.

Objective: This systematic narrative review aims to identify the key components—including facilitators and barriers—of web-based GTPs.

Methods: We systematically searched web-based databases (CINAHL, Embase, MEDLINE, PsycINFO, and Web of Science) to identify peer-reviewed articles published between 2000 and 2025 that involved web-based GTPs. After screening, 59 studies met the inclusion criteria and were analyzed using content analysis to identify key components and barriers affecting the delivery and receipt of web-based GTPs.

Results: Results were organized under 3 categories: design, content, and pedagogy. Key design considerations emphasized accessibility for diverse learning styles and digital literacy levels, customizability for different user groups, privacy protection, and the long-term sustainability of training content and delivery platforms. Core training content covered four domains: (1) suicide-related knowledge (eg, prevalence, myths, and at-risk groups), (2) gatekeeping skills (eg, understanding risk factors, recognizing warning signs, problem-solving and safety planning), (3) resource awareness (eg, available local resources and referral procedures), and (4) general mental health education (eg, mental fitness, mindfulness, and self-care strategies for gatekeepers). In terms of pedagogy, the reviewed studies used a wide range of strategies that comprised interactive learning activities (eg, simulation, practice exercises), periodic knowledge checks (eg, quizzes), and reinforcement mechanisms (eg, booster sessions). Additionally, fostering a sense of community (eg, online support spaces or discussion forums) and promoting trainees' autonomy (eg, self-paced training) were highlighted as key components of training delivery.

Conclusions: Web-based GTPs represent a promising avenue for expanding access to suicide prevention training. Their effectiveness may be strengthened through the integration of frameworks tailored to web-based learning environments, as well as interactive and user-centered design elements that support learning and retention. Future research should examine the acceptability, feasibility, and sustainability of these programs, while also refining their adaptation for diverse populations. In this regard, co-design approaches could facilitate the tailoring of such programs to the needs and specificities of their target populations.

XSL•FO
RenderX

Overall, enhancing the design and delivery of web-based GTPs may ultimately improve their contribution to suicide prevention efforts.

(J Med Internet Res 2026;28:e81572) doi:10.2196/81572

KEYWORDS

narrative review; suicide prevention; suicide; online; web-based; gatekeeper training

Introduction

Suicide is a critical public health issue worldwide, with significant social, emotional, and economic impacts on individuals, families, and communities [1]. As one of the leading causes of preventable death [2], suicide requires multifaceted approaches including awareness raising, reducing mental illness stigmas, and enhancing timely interventions [3]. Among these strategies, gatekeeper training programs (GTPs) have emerged as a cornerstone in equipping individuals to identify, approach, and support those at risk of suicide [4]. Gatekeepers are non–mental health professionals who may have contact with individuals at risk of suicide (ie, educators, parents, peers, or other community members), and are trained to recognize warning signs, initiate conversations about suicide, and connect individuals to appropriate professional help [1,5]. These programs, which are endorsed by the World Health Organization (WHO) [6], have demonstrated effectiveness in various settings and populations [7], highlighting their potential key role in suicide prevention.

Traditionally, GTPs have been delivered through in-person workshops and seminars [5], offering opportunities for direct interaction, role-playing, and immediate feedback. However, advances in technology and the increasing digitization of health education have led to the development and adoption of web-based GTPs [8,9]. Web-based formats provide significant advantages, including scalability, accessibility for geographically dispersed participants, and the ability to tailor training to diverse populations [7,10,11]. These programs can not only overcome logistical barriers (eg, physical attendance or availability of training) [12] but can also be particularly valuable for populations where confidentiality and anonymity are essential, such as stigmatized communities, including migrants and lesbian, gay, bisexual, trans, queer, and other sexual and gender minorities (LGBTQ+) populations [13]. Indeed, confidentiality can encourage more meaningful engagement with sensitive topics such as mental health and suicide among communities already affected by cumulative stigmas [14]. In addition, web-based training allows participants to learn at their own pace, accommodating busy schedules and varying levels of prior knowledge [15].

Recent evaluations indicate that web-based and in-person GTPs have similar effectiveness [8,9,11,14]. However, there are implementation challenges for web-based programs, including limited internet access, technological difficulties, varying levels of digital literacy, and user engagement issues [15]. Moreover, the design and content of web-based training are critical, and poorly structured or overly generic programs may fail to meet the complex and varied needs of participants [16].

Nonetheless, recent advancements in interactive technologies present unique opportunities to innovate and scale suicide prevention with GTPs. However, effectively harnessing web-based technologies requires a clear understanding of both the factors that contribute to program success and the challenges that hinder implementation. This systematic narrative review synthesizes the current evidence for web-based GTPs, addressing two key questions: (1) What are the key components of promising and successful web-based suicide prevention GTPs? (2) What are the barriers to delivery and usability of these programs? In summarizing the existing literature, this review provides recommendations for the development, implementation, and scaling of web-based GTPs, with the overall goal of contributing to global efforts for reducing suicide.

Methods

Overview

This systematic narrative review was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Multimedia Appendix 1) [17] and has been registered with PROSPERO (Registration ID: CRD42023462414). Since this study is a review of published, peer-reviewed articles, ethical approval was not required.

Search Strategy

A comprehensive search strategy was developed and conducted under the guidance of a specialized university librarian. The strategy combined 3 main concepts: suicide prevention, gatekeeping intervention, and web-based training. The full search strategy is available in Multimedia Appendix 2. The search was initially launched on October 1, 2023, and conducted across 5 databases: CINAHL, Embase, MEDLINE, PsycINFO, and Web of Science. Additionally, we searched the first 20 pages of Google Scholar, as results beyond the first 20 pages were not related to our 3 main concepts. The reference lists of all subsequently included articles were manually reviewed to identify further relevant sources. To ensure the inclusion of the most recent literature, the complete search strategy was relaunched on June 19, 2025.

Eligibility Criteria

We included peer-reviewed research articles on web-based GTPs, published in English or in French from 2000 onwards. We included all types of studies and designs, except for articles that did not discuss the web-based components of the training and research protocols. Articles focusing on general mental health programs not specific to suicide prevention were excluded. In addition, articles targeting mental health professionals (eg, psychotherapists, psychiatrists, and mental health nurse practitioners) were excluded, as their educational

XSL•FO
RenderX

background and professional experiences could result in significantly different training components, facilitators, and barriers compared with training programs for laypersons [18]. Articles targeting health care professionals outside specialized mental health—such as pharmacists, pharmacy staff, health care lecturers, and police officers—were eligible for inclusion. Studies published before 2000 were also excluded due to significant technological advancements since that time.

Screening and Study Selection

Using Covidence, EG, KW, and KL independently screened all titles and abstracts. Any records deemed potentially relevant by at least one reviewer were then subjected to a full-text examination, which followed the same independent screening process. In case of disagreement after full-text screening, EG, KW, and KL discussed their interpretations of the articles until they reached consensus. If consensus could not be achieved, OF served as a third-party mediator.

Data Extraction

Data extraction was conducted using Covidence. KW and EG independently extracted the following descriptive data items:

1. Methodological information: Geographical context, study goals, research approach, and study design.
2. Gatekeeper training program information: Program name, training setting, format, if training was adapted from an existing program, development process, training supporting platform, training objectives, training description, organizing institution, trainees' characteristics, number of trainees, components of training, topics covered by training, effectiveness of the training, target population of the gatekeeping intervention.

Conflicts in data extraction were resolved jointly by KW and EG using the Covidence Consensus tool. These data were used to characterize the included studies and inform Table 1, but were not analyzed further in the identification of barriers or facilitators. The analysis of facilitators and barriers was conducted independently of the data extraction process and is detailed in the following section.

XSL•FO
RenderX

Table 1. A brief overview of the characteristics of the training programs included in the systematic narrative review of web-based gatekeeper training programs (GTPs).

Author	Name	Components of training	Topics covered[a,b]			
			Information about suicide	Information about being a gatekeeper	Information about resources/referrals	Information about general mental health
Afsharnejad et al [19]	Talk-to-me MOOC[c]	Videos, quiz/tests	✓	✓		
Albritton et al [20]	Be present	Videos, e-tool box, homework assignments, social media posting	✓			
Bartgis and Albright [21]	Kognito gatekeeper simulations	Role-play/simulation		✓	✓	
Brown et al [22]	Indigenous network suicide intervention skills training (INSIST) program	N/S[d]	N/S	N/S	N/S	N/S
Bryant et al [23]	Kogito "At Risk Primary Care"	Role-play/simulation		✓	✓	
Canady [24]	Signs matter: early detection	Role-play/simulation quiz/tests, resources	✓			
Canady [24]	At-risk for high school educators	Role-play/simulation quiz/tests, resources	✓	✓	✓	
Canady [24]	At-risk for middle school educators	Role-play/simulation quiz/tests, resources	✓	✓	✓	
Carpenter et al [25]	Online veteran administration's (VA) suicide prevention GTP (S.A.V.E./SAVE [signs, ask, validate, encourage/expedite])	Videos, role-play/simulation	✓	✓		
Caughlan et al [26]	Mind4Health	Videos, readings, resources		✓	✓	
Cohen et al [27]	Israeli gatekeeper training	Role-play/simulation	✓		✓	
Colder Carras et al [28]	Stack-Up overwatch program	N/S	N/S	N/S	N/S	N/S
Coleman et al [29]	Kognito at risk for college students	Role-play/simulation		✓	✓	
Colucci et al [30]	Suicide first aid guidelines training	Videos, quiz/tests, infographics, homework assignments, reflective journaling	✓	✓	✓	
Ghoncheh et al [8]	MHO	PowerPoint presentation, audio features, graphs, quiz/tests, reading material, discussion board	✓	✓	✓	
Ghoncheh et al [8]	Children and family court advisory and support service program	N/S	N/S	N/S	N/S	N/S
Ghoncheh et al [8]	Question persuade and respond (QPR) online gatekeeper training	PowerPoint presentation, role-play/simulation, quiz/tests, reading material, videos, audio features	N/S	N/S	N/S	N/S
Ghoncheh et al [8]	Hollywood homeless youth partnership (HHYP) program	PowerPoint presentation, audio features, quiz/tests	N/S	N/S	N/S	N/S
Ghoncheh et al [8]	In the line of duty	Videos, audio features	N/S	N/S	N/S	N/S
Ghoncheh et al [15]	MHO	PowerPoint presentation, quiz/tests, discussion board	✓	✓	✓	

XSL•FO
RenderX

Author	Name	Components of training	Topics covered[a,b]			
			Information about suicide	Information about being a gatekeeper	Information about resources/referrals	Information about general mental health
Hawley et al [31]	Not applicable	Didactic content, video and audio clips, and reflection questions	✓	✓	✓	✓
Hill and McCray [32]	The Texas ask about suicide to save a life (AS + K?) suicide GTP	N/S	✓	✓	✓	
Hill et al [33]	ASK about suicide to save a life (AS + K?)	Videos	✓	✓		
Hofmann et al [34]	COPS (coping with suicide)	Videos, reading material, worksheets, quiz/tests	✓	✓	✓	
Hofmann and Wagner [35]	N/S	Videos, audio plays, manual	✓	✓	✓	✓
Holmes et al [14]	Start	N/S	N/S	N/S	N/S	N/S
Kawashima et al [36]	N/S	PowerPoint presentation, videos	✓	✓	✓	
Kimbrel et al [37]	Safety planning intervention (SPI)	PowerPoint presentation, videos, role-play/simulation, reading material, worksheets		✓		
Kingi-Ulu'av et al [38]	LifeKeepers booster session	Readings	✓	✓	✓	
Kingi-Ulu'av et al [7]	QPR	N/S	N/S	N/S	N/S	N/S
Kingi-Ulu'av et al [7]	MHO	N/S	N/S	N/S	N/S	N/S
Kingi-Ulu'av et al [7]	I CARE	N/S	N/S	N/S	N/S	N/S
Kingi-Ulu'av et al [7]	Act on FACTS: making educators partners (MEP)	N/S	N/S	N/S	N/S	N/S
Kingi-Ulu'av et al [7]	Kognito gatekeeper simulations	N/S	N/S	N/S	N/S	N/S
Kreuze and Ruggiero [39]	Kognito at-risk for high school educators	Videos, role-play/simulation		✓	✓	
Kreuze and Ruggiero [39]	QPR	Videos	✓	✓	✓	
Kreuze and Ruggiero [39]	MEP in youth suicide prevention: ACT on FACTS	Videos, role-play/simulation, audio features	✓	✓	✓	
Kreuze et al [10]	QPR	Videos, testimonials, narration, bulleted lists, mnemonics, pocket cards, role-play/simulation, self - audit checklist	✓	✓	✓	
Kreuze et al [10]	MEP in youth suicide prevention	Videos lectures, expert content, conversations example, role-play, testimonies, activities related to videos	✓	✓	✓	
Lamis et al [40]	MEP in youth suicide prevention: ACT on FACTS	Lecture, question and answers, digital vignettes/interactive activities	✓	✓	✓	
Lancaster et al [9]	Web-based QPR	Videos, text, pictures, audio features	N/S	N/S	N/S	N/S
Lee-Tauler et al [41]	Chaplains-CARE online program	Didactic lectures, videos, reading material, quiz/tests, interactive activities	✓	✓		✓

XSL•FO
RenderX

Author	Name	Components of training	Topics covered[a,b]			
			Information about suicide	Information about being a gatekeeper	Information about resources/referrals	Information about general mental health
Liu et al [42][e]	Various	Various	N/S			
MacDonald Hart et al [43]	Suicide intervention first aid (SIFA)	Didactic lectures, interactive, discussions, role-play/simulation exercise, skills practice	✓	✓	✓	
Manning and Van Deusen [44]	Western Michigan University suicide prevention program online course	Videos, photographs, and graphics	✓	✓	✓	
Marley et al [45]	Pharm-SAVES training	N/S		✓	✓	
McKay et al [46]	Living works start	Videos, reading material	✓	✓	✓	
Mirick [47]	SOS (signs of suicide) for school staff	Role-play/simulation with both child and adolescent	✓	✓	✓	
Mishkind et al [48]	VitalCog: suicide prevention in the workplace (formerly known as Working Minds)	Videos, role-play/simulation, group discussion, workbook	✓	✓	✓	
Osteen et al [49]	QPR for law enforcement	N/S		✓		
Perepezko et al [50]	Stack-Up overwatch program	N/S	N/S	N/S	N/S	N/S
Pilbrow et al [51]	Advanced suicide prevention training for pharmacists	Video, role-play/simulation, digital workbook, group discussion	✓	✓	✓	
Postuvan et al [52]	IAlive (iˇZiv in Slovenian)	Videos lectures, animated examples, interactive images/graphics with pop-ups	✓	✓		
Quinnett [53]	QPR pathfinder training	Video, role-play/simulation, reading	✓	✓	✓	✓
Reifegerste et al [54]	Help for relatives	Videos, audio-recordings, manual, text content	✓	✓	✓	✓
Rein et al [55]	Kognito	Role-play/simulation		✓		
Robinson-Link et al [56]	Kognito	Role-play/simulation	✓	✓	✓	
Roslan et al [57]	Online advanced C.A.R.E suicide prevention GTP (AdCARE)	Role-play/simulation	N/S	N/S	N/S	N/S
Ross et al [58]	Suicide prevention for college student gatekeepers program	Role-play/simulation, skills practice, discussions, peer cofacilitation	✓	✓		
Ross et al [59]	Suicide prevention for college student gatekeepers program	Role-play/simulation, skills practice, discussions	✓	✓		
Schmeckenbecher et al[e] [60]	Various	N/S	N/S	N/S	N/S	N/S
Seabury [61]	Crisis counseling: I Am Chipper!	Interactive PowerPoint presentation, videos, role-play/simulation, reading material	✓	✓		
Seabury [61]	Suicide assessment: rube farmer	Interactive PowerPoint presentation, videos, role-play/simulation, reading material	✓	✓		

XSL•FO
RenderX

Author	Name	Components of training	Topics covered[a,b]			
			Information about suicide	Information about being a gatekeeper	Information about resources/referrals	Information about general mental health
Seabury [62]	Crisis counseling: I Am Chipper!	Interactive PowerPoint presentation, videos, role-play/simulation, reading material, quiz/tests	✓	✓		
Seabury [62]	Suicide assessment: rube farmer	Interactive PowerPoint presentation, videos, role-play/simulation, reading material, quiz/tests	✓	✓		
ShantaBridges et al [63]	Suicide prevention and awareness for depression	N/S	✓	✓	✓	
Smith-Millman et al [64]	Kognito	Role-play/simulation		✓	✓	
Stone et al [65]	Youth suicide prevention: an introduction to gatekeeping	PowerPoint presentation, quiz/tests, resources, worksheet, audio files	✓	✓	✓	✓
Stover et al [66]	Pharm - SAVES	Videos, reading material, resources	✓	✓	✓	
Sun et al [67]	Chinese life gatekeeper training program	Videos, role-play/simulation, contextual understanding, group discussion, Q and A session	✓	✓	✓	
Teo et al [68]	VA S.A.V.E.	Videos	✓		✓	
Teo et al [69]	VA S.A.V.E.	Videos, vignettes	✓	✓	✓	
Timmons-Mitchell et al [70]	Kognito At-Risk for Middle School Educators	Role-play/simulation	✓	✓	✓	
Wislocki et al [71]	Multiple (n=506)[f]	Videos	✓	✓	✓	

[a]We categorized topics in four categories: (1) Information about suicide (including information about suicide prevention, suicidal or self-injury behaviors, suicide myths, suicide prevalence and statistics, risk factors for suicide, protective factors against suicide, and signs of mental distress/suicidal ideation/warning signs), (2) Information about being a gatekeeper (including intervention skills and identification of at-risk individuals), (3) Information about resources and referrals, and (4) Information about general mental health (including mental fitness and self-care).

[b]We exclusively reported the topics explicitly mentioned in the article, but we acknowledge that the training programs might cover additional topics not mentioned in the article.

[c]MOOC: mass open online course.

[d]N/S: not specified.

[e]These systematic reviews and meta-analyses are included for thoroughness; however, data from the primary studies were not re-extracted in this table since most were already included individually, and the remaining did not meet our eligibility criteria.

[f]This scoping review included 506 training videos. For the sake of conciseness, we did not include the program names in this table.

Data Analysis

We used an inductive content analysis approach [72,73] to identify the barriers and facilitators of web-based GTPs. Unlike the data extraction process, which aimed to summarize descriptive characteristics of the studies, the content analysis involved a separate, in-depth examination of the full texts of the included articles. EG and KW used Dedoose (version 9.0.107; Socio Cultural Research Consultants, LLC) to independently and inductively code full-text PDFs of the articles to reduce bias in interpretation. Coding began with line-by-line open coding, assigning initial code to relevant segments of text without predetermined categories. After 10% of the articles were coded, the two coders met to compare their coding frameworks and ensure consistency. Joined by KL, they then proceeded to independently code the remaining articles. This process allowed for a systematic and thorough examination of the data and minimized the risk of missing important nuances. Once all articles were coded, EG aggregated codes into higher-order headings, which were reviewed by the research team to develop broader categories through the process referred to as abstraction [72].

Results

Results of the Search

Our search yielded 4662 entries, from which 1365 duplicates were removed. After title and abstract screening, 3143 articles were excluded, and 95 articles were further removed after full-text screening (Figure 1). The final sample included 59 articles published between 2003 and 2025.

XSL•FO
RenderX

Figure 1. PRISMA flow diagram.

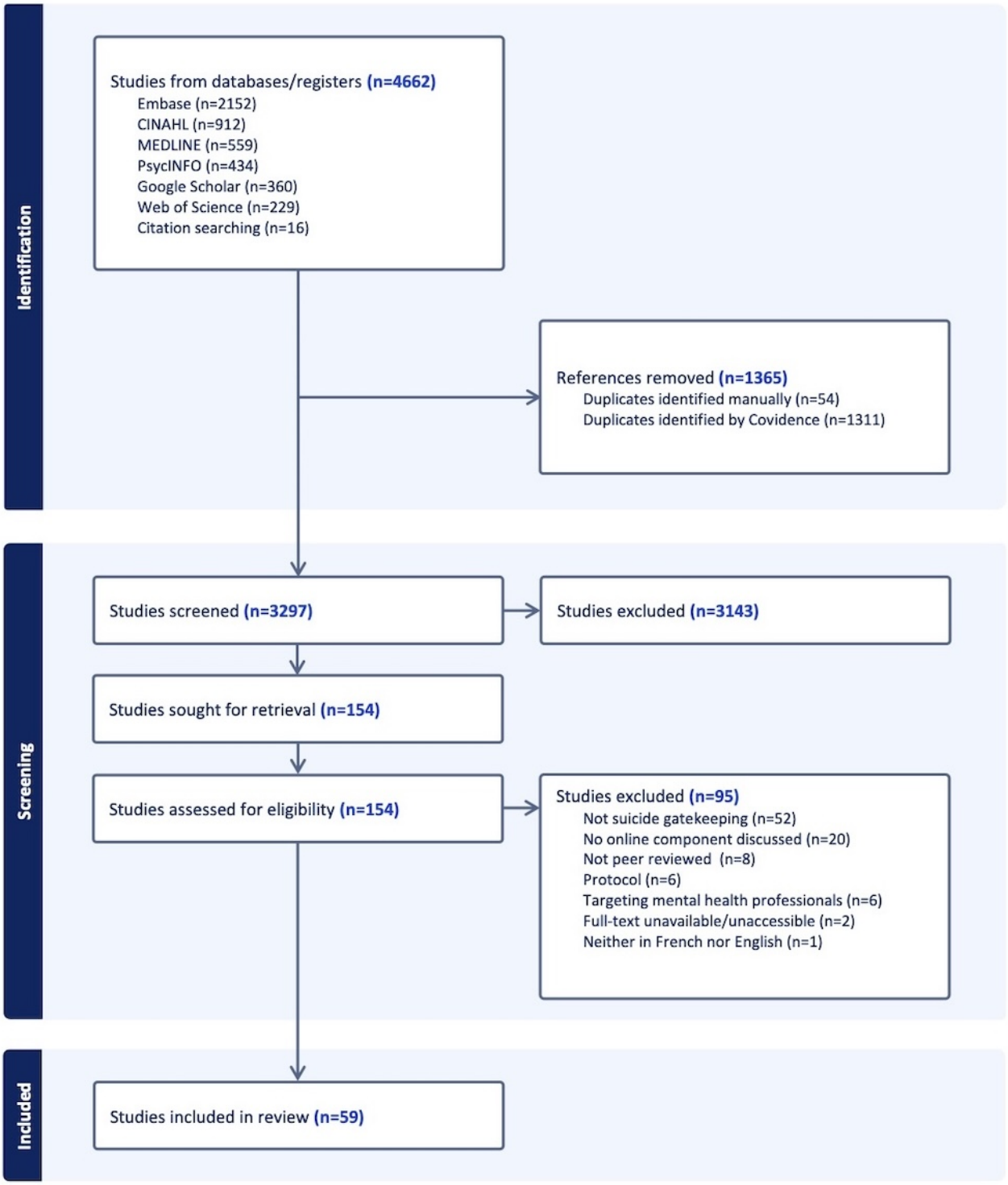

Reports Included in the Synthesis

Among the 59 articles reviewed, 36 used quantitative methodologies, including 11 randomized controlled trials. Additionally, 10 articles discussed multiple training [7,8,10,24,39,42,60-62,71]. The most common training programs discussed were Kognito (n=7); S.A.V.E./SAVE (signs, ask, validate, and encourage/expedite) or pharm-SAVE (n=5); and question, persuade, refer (n=4; excluding counts from review articles). The program settings varied, with schools/universities (n=21), military or law enforcement environments (n=6), and health care/ clinical settings (n=4) being the most common. Training duration ranged widely, from 20 minutes [25] to 32 hours [50] or multiple weeks [26,63,65]. Most training programs were aimed at youth or students (n=13), school staff (n=9), health care providers (n=6), or parents/caregivers (n=3). A brief overview of the characteristics of the GTPs included in this review is presented in Table 1, while a complete version is provided in Multimedia Appendix 3.

XSL•FO
RenderX

Design

We identified several critical design elements for web-based GTPs. Foremost, authors recommend that programs be grounded in theory and evidence-based practices [23,26,31,43,52,54,67,71]. For instance, Timmons-Mitchell et al [70] recommended evidence-based techniques such as motivational interviewing, and Afsharnejad et al [19] emphasized that mental health education principles, specifically the PERMA framework, should be used to enhance the efficacy of training. In addition, it was recommended that programs align with existing available suicide prevention initiatives to leverage and integrate resources [27,53,58]. This design specificity underlines web-based GTPs as a connector between community members and existing prevention efforts rather than being a stand-alone solution [22,43,54]. Indeed, it was emphasized by several authors that web-based GTPs should complement and be purposefully integrated rather than aspire to replace other suicide prevention initiatives [22,30,40,54,56]. To ensure usefulness and promote uptake, several programs integrated a co-design approach, involving interdisciplinary experts (including website developers), stakeholders, and potential end users participating in the design and implementation processes to better tailor the training programs [19,22,26,28,34,37,50,66-68,71]. In addition, the need for adequate financial support to sustain, maintain, and improve the programs to ensure the integration of new knowledge and best practices was mentioned [8,21].

Accessibility was also an essential design element. Authors highlighted that technical issues, such as glitches in program software, can hinder accessibility by disrupting the flow of the training [10,34,41,61]. Despite assumptions of widespread internet access, real-world barriers such as low bandwidth, limited availability of internet or technologies (eg, computers, smartphones), and insufficient media skills persist, potentially limiting users' access [10,21,22,30,35,61]. Authors suggested that the training platforms should be more accessible by reducing the bandwidth requirements and ensuring 24/7 availability [8,53]. They also emphasized simplifying navigation [54] through avoiding registration [52], the use of specific software (eg, Adobe), or requiring specific web browsers [10]. Further, the variability of individual trainee competencies and preferences in terms of technology use necessitates that training platforms be user-friendly and ideally use technologies familiar to learners for a better user experience [22,48,52,57,70]. Brown et al [22] recommended the adaptability of training programs across multiple technological devices (eg, computer, smartphones, and tablets) and suggested incorporating diverse web-based supports (mobile app, Facebook group, messenger feature, etc). Accessibility was also contingent on delivering content through diverse formats, including audio, video, and text, and ensuring these vehicles could be adjusted to suit individual learning speeds and accommodate trainees with disabilities and learning or attention deficits. Practically, this means adding audio components to written material [8] with a sufficiently large font [10], being able to control the speed of audio and video content [30], and adding embedded text explanations and captions to videos [66]. Multiple studies [10,54,61] suggested adding simple, nongeneric visual components to text as well as color-coding sections to create a visual learning structure, avoiding long sentences to describe concepts and prioritizing bullet points or synthesized information. Authors also highlighted the importance of accessibility and program duration, noting that web-based training programs should be as brief as possible while still meeting learning objectives [10,20,25,34,37,41,44,46,54,57,69]. For example, trainees in the study by Carpenter et al [25] preferred the theoretical part of the training to last 30 minutes or less, and trainees in the study by Reifegerste et al [54] rated videos averaging 7-8 minutes to be too lengthy. Still, Liu et al [42] recommend that training sessions last more than 2 hours to induce a significant change in trainees' attitudes and behaviors. Last, Carpenter et al [25] and Wislocki et al [71] also emphasized the importance of cost for the training programs, and making them financially feasible for a wide range of trainees.

Customizability was another critical element wherein the program design needed to be adaptable to meet trainees' unique needs, personal characteristics, and professional backgrounds [8,22,26,27,31,35,41,43,53,54,57,64,71]. This approach comprised providing personalized feedback tailored to each trainee's progress [10,21,57] and developing customized learning experiences for trainees based on their prior gatekeeping and suicide prevention experience and knowledge [8,22,32,37,54]. This included adding complementary modules, supplemental material, and additional content or offering flexible options for practice sessions [10,47]. Moreover, the customization of web-based training should encompass a variety of gatekeeping examples, scenarios, and practice exercises, thereby facilitating the alignment of content with diverse trainee preferences [10,25,37,41,66,70]. Furthermore, to facilitate customization that accommodates trainees' variable schedules and pacing, authors recommend enhancing program flexibility through multiple learning modes and having the option to segment learning sessions [8,10,31,35,48,70].

Another design consideration for web-based training was standardization. Emphasized was the importance of implementing mechanisms to ensure consistency in the delivery of training content [31]. In web-based synchronous programs led by human instructors, standardization largely depends on the instructor's expertise and familiarity with the material and the target audience, which was not specified in the reviewed articles. While instructors can adapt to trainees' needs and pace, maintaining consistency remains crucial, especially given potential fatigue, which can affect delivery. In asynchronous programs, standardization was built into the training itself. To support this, authors have recommended incorporating elements such as standardized scenarios, automated feedback systems, and prerecorded videos to ensure uniformity across sessions [21,62]. These standardization mechanisms promote consistency and create a sense of psychological safety for trainees, as they reduce the risk of judgment from instructors or peers [21,62,70]. Bartgis et al [21] have recommended standardization in delivery mechanisms, including web-based role-play modalities using avatars, to ensure consistency in content delivery, regardless of trainees' personal characteristics.

XSL•FO
RenderX

Another design consideration was privacy [10,22,28,42,46]. Reifegerste et al [54] noted that web-based formats may offer a greater sense of anonymity compared with in-person settings. While complete anonymity cannot be guaranteed in web-based training programs, Lancaster et al [9], Caughlan et al [26], and Reifegerste et al [54] argue that it remains an important feature, as it may reduce trainees' anxiety, thereby encouraging more open engagement with sensitive topics. Some authors raised concerns about privacy risks associated with using third-party internet platforms. For example, Brown et al [22] noted that while using a private Facebook group can foster support and connections among trainees, it may also pose risks to participants' privacy and confidentiality. Further, confidentiality concerns extended beyond the training itself to include program evaluation. Thus, it is essential that all data collected during the training, including during role-playing exercises, is stored on a secure server [21].

The final key element to consider in the design phase was sustainability. Specifically, maintaining and updating technological aspects were significant challenges for the use and longevity of web-based GTPs [8]. Brown et al [22] suggested that program designers allocate resources to ensure the maintenance and updates of the training in the initial development plan. For example, they specified that while web-based forums can be found valuable among trainees, they require substantial effort in terms of moderation and maintenance, which can be resource-intensive. Ghoncheh et al [15] emphasized the importance of developing a sustainability plan that minimizes maintenance and cost while preserving the integrity of the program.

Content

A consistent content recommendation was the use of clear and concise terminology and vocabulary familiar to trainees [10,51,54,69]. In programs designed for nonprofessionals, avoiding technical terminologies or clinical jargon, which may act as a barrier, was a consistent recommendation [10,37,54]. Authors recommended balancing testimonials ("emotional content") and practical information ("informative content") to better engage trainees and support the destigmatization of suicide prevention [10,54]. In addition, the importance of activities that help trainees feel comfortable using the word "suicide" [25] and support them in discussing suicidality in a nonstigmatizing manner was emphasized [19,26,27,31,57]. At the same time, caution was advised against including graphic details of suicide death [57] or downplaying the gravity of the topic [47]. Authors of the included studies recommended providing a manageable amount of clear, straightforward, and easy-to-follow information to avoid overwhelming trainees [10,69] while still offering sufficiently rich content [47,54].

In terms of didactic topics to cover, trainees endorsed four different topics: (1) Information about suicide including suicide definition, epidemiology, statistics and prevalence [10,25,27,31,43,52,66], and suicide myths and beliefs [10,25,32,52]; contextual factors connected with suicide [26,30], content about at-risk subgroups [27,30,35,52,54], legal requirements or policies [10,25], and broader community concerns regarding suicide prevention [22,58]; (2) Information about how to be a gatekeeper including how to identify at-risk individuals [22] and the differentiation between warning signs and risk factors according to various settings [10,27,30,31,35,43,52,54,58,69,71], clear and memorable steps to follow for gatekeeping interventions [10,25,51,52,66], ways to initiate conversations about suicide [25,26,35,51,52,58,69], what to say and topics to avoid when talking with suicidal individuals [41,47,54,69], safety planning and problem-solving [35,41,71], including with dealing with nonreceptive individuals [50], and follow-up strategies postintervention and postvention care [22,41,66,71]; (3) Information about the broad range of existing services including referral guidance [10,22,25,27,30,35,42,43,45,50,52,54,69,71] according to existing local resources [10,21,25,26,66]; and (4) Information about general mental health for gatekeepers [42,54] including mental fitness [19] or mindfulness [41], the challenges of being a [22,25,30,31] and gatekeepers' well-being and self-care strategies [22,35,41,57-60]. Importantly, these topics should be adapted to trainees' backgrounds and accompanied by relevant examples to which trainees could relate [10,25,41,42,45,51,66] as well as the specific context of the intervention [10,22,30,43,47]. The importance of ensuring that the content was culturally relevant to the intended audience was described as a factor bolstering the effectiveness and the inclusivity of training programs [7,21,22,26,27,30,38,42,53,63,67]. Notably, 4 studies underscored the limited diversity in training, pointing to a lack of content and examples specifically addressing the needs of LGBTQ+ populations and women [28,41,54,71]. In contrast, topics perceived as less relevant by trainees included procedural guidance on reporting suicide cases, professional assessment practices that were not directly applicable to their roles or lived experiences, and research data from unrelated contexts [22,30].

In addition to the topics above, several core skills and competencies to be acquired by trainees were highlighted. These include the ability to establish a strong rapport and build a trusting relationship with individuals experiencing suicidal ideation [22,30]. Brown et al [22] and Hawley et al [31] emphasized the inclusion of interpersonal "soft skills"—including active listening, compassion, patience, and nonjudgmental attitudes—as foundational elements of effective training. Similarly, Bartgis et al [21] underscored the importance of motivational interviewing skills, including the use of open-ended questions, providing affirmation, reflective listening, and summarizing. The development of advocacy skills was also identified as a particularly important skill when supporting individuals living in marginalizing conditions who face significant structural barriers to services [22].

Pedagogy

Several pedagogical elements were identified as critical for the effective delivery of web-based gatekeeper training. First, the inclusion of interactive learning activities—including role-plays, hands-on activities, practice, and scenario-based exercises—was consistently emphasized as essential to skills development and enhancing learning outcomes [8,10,24,25,36,41,47,53,57,63,64]. Trainees highly valued these activities [10,22,23,25,26,30,37,41,45,51,62,69]. For example, participants in 4 studies [10,41,47,51] emphasized the importance of having

XSL•FO
RenderX

more time allocated for practice during the web-based training through hands-on exercises and interactive learning opportunities to enhance motivation and information retention. Ross et al [59] also suggested having small group sizes to increase participation and knowledge gains in the case of synchronous training. Participants in 5 studies [10,25,47,57,66] endorsed role-plays that were concrete, realistic, relatable, and applicable to their contexts. Although some authors noted challenges for implementing role-plays via web [15,30,51], Seabury [62] and Liu et al [42] encouraged leveraging the use of innovative technologies. Examples of such technologies include the use of avatars [21,23,29,55,60,64], virtual reality [60], interactive videos or video demonstrations [41,51,66], live videoconferencing role-play practice sessions [10], or mathematical behavioral models and algorithms to create realistic simulations where trainees can practice gatekeeper skills [70]. Knowledge checks, such as tests and quizzes, were identified as another vital component of interactive learning [8,10,34,41], providing trainees with immediate feedback on their understanding of the material covered [61]. However, some authors cautioned that web-based training might lack the interactive and adaptable learning environments synonymous with traditional in-person training programs where instructors could respond to trainees in real-time [62,71]. Offering clarifications and real-time feedback dispersed throughout the training, including via knowledge checks, could address this limitation [8,10,41,62].

A second key pedagogical element was the autonomy for trainees regarding the pace and style of learning [8,9,19,21,34,35,62]. Inversely, several authors cautioned against over-reliance on trainees' intrinsic motivation [9,20], which can limit engagement [10,19] and even increase attrition [65]. Thus, authors emphasized that it is crucial to implement ongoing guidance, learner incentives (eg, raffle for a gift certificate [44]), and motivational strategies (eg, email reminders to finish the training [37]), given that intrinsic motivation may not be sufficient to complete the training [19,41,63]. Such mechanisms could take the form of a time limit for training completion [20] or reminders to encourage trainees to complete the training [37,41].

Building a community of practice emerged as another key recommendation. Unlike face-to-face training, web-based training programs often limit direct interactions with trainers as well as with other trainees. This lack of human interactions was mentioned as hindering skills practice and development [19,21]. Afsharnejad et al [19] highlighted the importance of fostering emotional connection with trainees. Establishing a community of practice, as suggested by other authors, could support connectivity by enabling trainees to access expert insights and feedback [8,41,53,57]. Some also highlighted that such communities of practice could provide trainees with opportunities to be paired with gatekeeping buddies for peer support [22], offering continuing networking and debriefing opportunities [22,38], offering support [59], and contributing to shared learning [51]. Examples of implemented approaches include messaging platforms [57], moderated forums with discussion boards and threads [8,65], chatrooms [28], and digital coaches providing direct and personalized feedback [21].

Finally, reinforcement strategies were recommended to consolidate skills and knowledge acquired during web-based training. This could include incorporating multiple repetitions of the learning material throughout the training [8]. While repetition was mentioned to maximize learning, participants in the studies by Kimbrel et al [37] and Kreuze et al [10] indicated that repetition was irritating, distracting, and even useless. Other recommendations included follow-up training sessions spaced in time to reinforce and sustain knowledge [31,42,51], training refresher sessions [15,21,22,37,42,51,53,56,69]—although Kingi-Ulu'ave et al [38] concluded that passive boosters were not impacting knowledge retention and trainees' self-efficacy—and complementary material like digital workbooks to write notes, take-home training summaries, and complementary resources [10,22,34,35,41,51].

Overall, Kreuze et al [10] recommended having a variety of teaching and evaluating strategies to address diverse learning styles and needs, promote critical thinking, and incorporate learning across the cognitive, affective, and psychomotor domains (eg, through role-plays).

An integrated list of dos and don'ts, organized from the results, is provided as a one-page checklist in Multimedia Appendix 4 [8-10,14,15,19-54,56-71,74].

Discussion

Principal Findings

This narrative review systematically synthesized findings drawn from published literature focused on web-based suicide prevention GTPs to make recommendations for key design, content, and delivery. While there is increasing enthusiasm for web-based education broadly [75,76], including in the field of suicide prevention [9,77], there is a lack of clear guidance on best practices for the development and implementation of these programs [14]. The findings of this review contribute to addressing this knowledge gap by detailing critical elements across 3 key areas: design, content, and pedagogy. These insights offer a conceptual foundation for future research and offer practical guidance for the development and implementation of effective web-based gatekeeper training initiatives.

A central finding of this review is the necessity of grounding program design in evidence and theory, both in terms of the GTP content and its pedagogy. Integrating evidence-based practices in its content not only enhances the credibility of gatekeeper training but is also essential for ensuring its effectiveness. Yet, this is not without challenges. Evidence-based suicide prevention initiatives remain in early stages of translation [78], and a key critique of GTPs is the frequent lack of rigorous evaluation that can demonstrate their effectiveness [79]. In terms of pedagogy, most programs to date draw from social cognitive theory [80] and the theory of planned behavior [81], which highlight psychosocial determinants of behavior change and draw attention to contextual factors that influence how new behaviors are learned and sustained. However, these theories focus on individual behaviors and may fall short of accounting for the specific affordances and constraints of digital learning environments [82]. The inclusion

XSL•FO
RenderX

of instructional design theories tailored to web-based modalities—such as the community of inquiry framework for web-based learning [83] and the cognitive theory of multimedia learning [84]—could significantly strengthen the pedagogical foundation of these programs, taking full advantage of web-based platforms.

Moreover, research and existing guidelines suggest that effective gatekeeper training should be designed and situated within a broader, multilevel suicide prevention strategy [1]. This aligns with recommendations from the WHO [1] and multiple reviews of national suicide prevention strategies [85-87], which recognize the potential of GTPs and their limitations as a stand-alone intervention (Zalsman et al [79]). For example, GTPs are not meant to replace suicide specific interventions delivered by trained mental health professionals. In addition, the effectiveness of GTPs may be limited in environments where suicide is highly stigmatized [88]. Indeed, stigma has consistently been identified as a key factor contributing to gatekeepers' reluctance to intervene [4]. It can also play a complicit role in silencing individuals with suicidal thoughts [12,89]. The threat of stigma is especially relevant for marginalized populations, including LGBTQ+ and racialized populations, who already experience significant barriers to safe and appropriate mental health care [90,91]. Moreover, while these programs can help connect individuals experiencing suicidal thoughts to appropriate support, there is little evidence on how they foster hope or promote a meaningful life [12,92]. In addition to building core gatekeeper competencies—identifying warning signs, talking about suicide, referring—future GTPs should incorporate upstream approaches that actively promote mental health well-being and reduce stigma.

To better tailor GTPs to current needs, several articles recommended co-design approaches in the development of web-based GTPs. Co-design approaches have been described in the health literature as an effective approach to enhance program relevance and user engagement by involving end users in the development process [93,94]. However, as identified by Qasim et al [95], important knowledge gaps remain on how to best mobilize co-design approaches in maximizing outcomes. A useful starting point to optimize the inclusion of end users' insights and needs in the co-design process could be to draw from participatory and community-based research principles [96]. Despite the espoused widespread use of participatory and community-based methods with vulnerable populations, there is still very little consensus on best practices [97,98]. Nevertheless, emerging key components such as fostering collaborative spaces [99,100], building capacities [98,101], and the balancing of power [102-104] could inform co-design practices of future GTPs.

In terms of content and pedagogy, this review reinforces the need for flexibility and adaptability to accommodate the varied needs and learning styles of trainees, aligning with broader findings on adult education and training programs [105]. Customizability—including tailoring scenarios, examples, and resources to specific user groups—has been shown to enhance engagement and practical application in education [106]. Dreier et al [16] suggest that customization of the training should also include the "desired degree of confrontation"—that is, text, images, videos, testimonials, representation of suicidality—not only to improve the engagement and satisfaction of the learners, but also to provide a sense of agency regarding the possible emotional distress in reaction to sensitive content. Interactive elements like role-plays, simulations, and real-time feedback are frequently cited as best practices in the education literature and are particularly effective in improving skill retention and building confidence [107-109]. While the adaptability and flexibility of web-based training have been extensively acknowledged as a strength in logistical terms (eg, pace and access) and range of possible content [110], they have often been criticized for their lack of interpersonal interactions, a key component to enhancing learning outcomes [111]. Among the reviewed articles, incorporating ways to foster a sense of connection and community among trainees in a web-based format (eg, creating or mobilizing already existing peer support networks and collaborative learning opportunities) has been emphasized and is consistent with research on social learning theory [112].

As a possible alternative to actual interpersonal interaction through the building of peer networks for collaborative learning, some authors have suggested the use of avatars or models—whose machine learning algorithm was not specified in the current reviewed articles—that would provide in vivo feedback in response to the trainees' interactions. Such use of recent technologies provides an interesting avenue for interactive learning when actual human interaction is not possible. Although technologies based on artificial intelligence (AI) have been adopted in education—whether through students' initiative or within the actual teaching curriculum—the absence of guidelines and best practices limits their adoption. In their systematic review of the use of AI in higher education, Ouyang et al [113] discussed the potential of AI technologies for improving the learner's experiences in terms of engagement and providing accurate predictions (ie, real-time feedback). Yet, they further claim that the mere use of AI technologies does not necessarily lead to positive educational outcomes, emphasizing the importance of integrative frameworks or theories to support and enhance their ethical use. Another key ethical consideration in the use of AI technologies and open-source machine learning models is related to issues of confidentiality and privacy [114,115]. Such precautions become even more important in the context of vulnerable and sensitive topics such as suicide. Concerns about the privacy of information may also limit trainees in their willingness to participate in real-time AI-based role-play or feedback sessions. To mitigate risks and barriers to engagement, GTPs should explicitly disclose their use of such technologies, ensure secure data storage (eg, encrypted servers and restricted access), and implement clear protocols for handling data from role-play or training recordings [116].

Key barriers to the successful implementation of web-based GTPs for suicide prevention were identified in this review, many of which are inherent to web-based education in general. In addition to the lack of interpersonal and customizable human interactions, accessibility—while a strength—remains a significant challenge in terms of limited internet bandwidth, lack of access to the necessary technology, and varying levels

XSL•FO
RenderX

of digital literacy impeding program reach and effectiveness. While the digital divide is narrowing globally [117], it continues to be a barrier for some populations due to persisting structural inequities flowing to and from the social determinants of health, including income, geographic location (ie, underdeveloped areas), and educational disparities [117,118]. Further, technical difficulties, including platform glitches and inefficiencies, not only hinder user experience and engagement but also the retention of learnings [119].

Sustainability is a critical challenge for web-based suicide GTPs. Without dedicated long-term funding, government, or commercial support, updates to keep content relevant and evidence-based are not possible. In fact, some of the interventions reviewed in this study could not be accessed at the time of writing, suggesting potential challenges to sustainability or ongoing availability. Ensuring long-term success requires more than just the initial development and implementation of an intervention. Sustainability should be taken into consideration in the initial steps of program development, given that it demands ongoing engagement, adaptation to evolving needs, and seamless integration into broader suicide prevention frameworks. The successful implementation of GTPs and their sustainability could benefit from network interventions at the public health level that focus attention on forging and solidifying intersectoral partnerships towards a more distributive model of responsibilities and accountability for suicide prevention [120].

Future Directions for Research on Web-Based GTPs

Multiple calls have been made for more robust research into the evaluation of GTPs [4,79,121], and while we agree with this need, we also emphasize that understanding the most effective ways to deliver GTPs in web-based formats is important. Future studies should explore questions such as: "What are the most effective web-based approaches for training gatekeepers?" "What specific features of web-based training (eg, interactive modules, live simulations, peer discussions) can enhance skill acquisition and retention?" "How does web-based GTPs impact participants' ability to recognize and respond to suicidal crises compared to in-person training?" Additionally, process evaluations and user feedback are essential to assess the quality of training delivery, user engagement, and practical application of skills learned via the web. Such evaluations can help to better understand implementation barriers and facilitators to improve the overall design of web-based GTPs. Another aspect to consider in the evaluation and design of GTPs is the impact and best delivery options of refreshers and booster sessions. While there is growing attention in tailoring GTPs to the needs of at-risk populations (eg, indigenous youth, migrant youth, gamers, and veteran as seen in this review), future research should also explore how web-based formats can better reach underrepresented and marginalized populations—such as LGBTQ+, Black, Indigenous and Racialized communities, or geographically isolated communities—more effectively and assess the training's scalability and sustainability. Mixed methods studies and comparative trials are critical to advancing knowledge about how web-based GTPs can contribute to comprehensive suicide prevention strategies.

Limitations

This review is limited by its focus on peer-reviewed articles published in English and French, potentially excluding relevant studies in other languages or formats (gray literature), which may introduce selection bias. The heterogeneity of included studies, such as variations in design, target populations, and program settings, made it difficult to draw generalizable conclusions. Many studies lacked rigorous evaluation methods, limiting our understanding of the impacts of the GTPs included. Additionally, insufficient reporting on user feedback, contextual factors (such as cultural and technological differences), and the implementation process restricts the applicability of findings across diverse settings. Notably, none of the studies reviewed specifically addressed our research question, which is, "what works" and "does not work" in web-based GTP. This gap underscores the need for further research that directly explores effective strategies to deliver web-based suicide prevention training to provide definitive guidance and improve training outcomes.

Conclusions

Web-based GTPs hold significant promise for suicide prevention, yet much of the existing research has focused on determining whether these programs are effective [7], with evidence inconclusive as stand-alone interventions [88,122,123]. To advance the field, it is crucial to enhance evaluation efforts [123] while simultaneously exploring what works best in web-based formats for building gatekeeper skills. While there is a call for randomized controlled trials to build evidence of suicide prevention interventions such as GTPs to prove their efficacy, the next logical step in the evaluation and development of effective web-based GTPs should focus attention on the process of their implementation, including acceptability and feasibility. This review highlights a wide range of key considerations regarding web-based pedagogy that require closer attention to determine which might represent best practices to implement in the development of future GTPs. This involves a commitment to continuous improvement by leveraging technological advancements to enhance accessibility, engagement, and adaptability for diverse populations. By aligning program design with the latest innovations, including emerging tools such as AI [124,125], and rigorously evaluating their impact, web-based GTPs can evolve into a vital component of comprehensive suicide prevention strategies. Ultimately, these efforts will not only improve the quality and reach of interventions but also have the power to save lives.

Acknowledgments

The authors would like to thank Sylvie Fontaine and Travis Salway for their valuable assistance with the development of the search strategy. Generative artificial intelligence assistance (ChatGPT, OpenAI) was used solely to proofread and improve the grammar and clarity of text drafted entirely by the authors. The tool was not used to generate content, analyze or interpret data,

XSL•FO
RenderX

or produce tables or figures. All artificial intelligence–assisted edits were reviewed and verified by the authors, who remain fully responsible for the scientific integrity of the manuscript.

Funding

This work was supported by the Canadian Institutes of Health Research (CIHR-grant number: 183605). OF is supported by a salary award from the Fonds de Recherche du Québec.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

Authors' Contributions

OF contributed to the conceptualization of the study and was responsible for writing the original draft, as well as reviewing and editing the manuscript. OF also provided resources, acquired funding, and supervised the project. EG contributed to the investigation, methodology, data curation, and formal analysis, and participated in writing the original draft and reviewing and editing the manuscript. KL contributed to writing the original draft and to reviewing and editing the manuscript. KW contributed to formal analysis, software development, validation, and reviewing and editing the manuscript. JLO, HK, and BD each contributed to reviewing and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA checklist.
[PDF File (Adobe PDF File), 322 KB - jmir_v28i1e81572_app1.pdf]

Multimedia Appendix 2
Search strategy.
[DOCX File , 34 KB - jmir_v28i1e81572_app2.docx]

Multimedia Appendix 3
Complete table of characteristics of the online gatekeeper training programs included in the systematic narrative review.
[DOCX File , 90 KB - jmir_v28i1e81572_app3.docx]

Multimedia Appendix 4
Online gatekeeper training programs (GTP): Do’s and Don’ts checklist.
[DOCX File , 18 KB - jmir_v28i1e81572_app4.docx]

References

1. Preventing suicide: a global imperative. World Health Organization. 2014. URL: https://www.who.int/publications/i/item/9789241564779 [accessed 2025-12-13]
2. Suicide worldwide in 2019: global health estimates. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240026643 [accessed 2025-12-13]
3. Caine ED, Reed J, Hindman J, Quinlan K. Comprehensive, integrated approaches to suicide prevention: practical guidance. Inj Prev 2018 Jun;24(Suppl 1):i38-i45. [doi: 10.1136/injuryprev-2017-042366] [Medline: 29263088]
4. Burnette C, Ramchand R, Ayer L. Gatekeeper Training for Suicide Prevention: A Theoretical Model and Review of the Empirical Literature. Rand Health Q 2015 Jul 15;5(1):16 [FREE Full text] [Medline: 28083369]
5. Isaac M, Elias B, Katz LY, Belik S, Deane FP, Enns MW, Swampy Cree Suicide Prevention Team. Gatekeeper training as a preventative intervention for suicide: a systematic review. Can J Psychiatry 2009 Apr;54(4):260-268. [doi: 10.1177/070674370905400407] [Medline: 19321032]
6. Public health action for the prevention of suicide: a framework. World Health Organization. 2012 Jan 01. URL: https://www.who.int/publications/i/item/9789241503570 [accessed 2025-12-13]
7. Kingi-Uluave D, Taufa N, Tuesday R, Cargo T, Stasiak K, Merry S, et al. A review of systematic reviews: gatekeeper training for suicide prevention with a focus on effectiveness and findings. Arch Suicide Res 2025;29(2):329-346 [FREE Full text] [doi: 10.1080/13811118.2024.2358411] [Medline: 38884349]
8. Ghoncheh R, Koot HM, Kerkhof AJFM. Suicide prevention e-learning modules designed for gatekeepers: a descriptive review. Crisis 2014;35(3):176-185. [doi: 10.1027/0227-5910/a000249] [Medline: 24901058]

XSL•FO
RenderX

9. Lancaster PG, Moore JT, Putter SE, Chen PY, Cigularov KP, Baker A, et al. Feasibility of a web-based gatekeeper training: implications for suicide prevention. Suicide Life Threat Behav 2014;44(5):510-523. [doi: 10.1111/sltb.12086] [Medline: 24571612]
10. Kreuze E, York J, Lamis DA, Jenkins C, Quinnett P, Mueller M, et al. Gatekeeper training for youth suicide prevention: a mixed method comparative analysis of two online programs. Psychol Sch 2024;62(2):492-511. [doi: 10.1002/pits.23335]
11. Torok M, Calear AL, Smart A, Nicolopoulos A, Wong Q. Preventing adolescent suicide: a systematic review of the effectiveness and change mechanisms of suicide prevention gatekeeping training programs for teachers and parents. J Adolesc 2019;73:100-112. [doi: 10.1016/j.adolescence.2019.04.005] [Medline: 31054373]
12. Baril A. Undoing Suicidism: A Trans, Queer, Crip Approach to Rethinking (Assisted) Suicide. Philadelphia, Pennsylvania: Temple University Press; 2023.
13. Ferlatte O, Salway T, Oliffe JL, Kia H, Rice S, Morgan J, et al. Sexual and gender minorities' readiness and interest in supporting peers experiencing suicide-related behaviors. Crisis 2020;41(4):273-279. [doi: 10.1027/0227-5910/a000632] [Medline: 31657638]
14. Holmes G, Clacy A, Hamilton A, Kõlves K. Online versus in-person gatekeeper suicide prevention training: comparison in a community sample. J Ment Health 2024;33(5):605-612 [FREE Full text] [doi: 10.1080/09638237.2024.2332811] [Medline: 38602188]
15. Ghoncheh R, Gould MS, Twisk JW, Kerkhof AJ, Koot HM. Efficacy of adolescent suicide prevention E-learning modules for gatekeepers: a randomized controlled trial. JMIR Ment Health 2016;3(1):e8 [FREE Full text] [doi: 10.2196/mental.4614] [Medline: 26825006]
16. Dreier M, Ludwig J, Härter M, von dem Knesebeck O, Rezvani F, Baumgardt J, et al. Evaluation of an online suicide prevention program to improve suicide literacy and to reduce suicide stigma: a mixed methods study. PLoS One 2023;18(4):e0284944 [FREE Full text] [doi: 10.1371/journal.pone.0284944] [Medline: 37115766]
17. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015;4(1):1 [FREE Full text] [doi: 10.1186/2046-4053-4-1] [Medline: 25554246]
18. Holmes G, Clacy A, Hermens DF, Lagopoulos J. The long-term efficacy of suicide prevention gatekeeper training: a systematic review. Arch Suicide Res 2021;25(2):177-207. [doi: 10.1080/13811118.2019.1690608] [Medline: 31809659]
19. Afsharnejad B, Milbourn B, Hayden-Evans M, Baker-Young E, Black MH, Thompson C, et al. The efficacy of the "Talk-to-Me" suicide prevention and mental health education program for tertiary students: a crossover randomised control trial. Eur Child Adolesc Psychiatry 2023;32(12):2477-2489 [FREE Full text] [doi: 10.1007/s00787-022-02094-4] [Medline: 36194311]
20. Albritton T, Ford KL, Elsbernd K, Santodomingo M, Juzang I, Weddington P, et al. Implementing a peer advocate mental health digital intervention program for Ohio youth: descriptive pilot study. JMIR Ment Health 2021;8(4):e24605 [FREE Full text] [doi: 10.2196/24605] [Medline: 33890868]
21. Bartgis J, Albright G. Online role-play simulations with emotionally responsive avatars for the early detection of native youth psychological distress, including depression and suicidal ideation. Am Indian Alsk Native Ment Health Res 2016;23(2):1-27. [doi: 10.5820/aian.2302.2016.1] [Medline: 27115130]
22. Brown K, Toombs M, Nasir B, Kisely S, Ranmuthugala G, Brennan-Olsen SL, et al. How can mobile applications support suicide prevention gatekeepers in Australian indigenous communities? Soc Sci Med 2020;258:113015. [doi: 10.1016/j.socscimed.2020.113015] [Medline: 32559573]
23. Bryant R, Vogt M, Miller C. Effects of an avatar-based simulation on family nurse practitioner students' self-evaluated suicide prevention knowledge and confidence. Nurs Educ Perspect 2025;46(3):179-181. [doi: 10.1097/01.NEP.0000000000001321] [Medline: 39167360]
24. Canady VA. National, local efforts train educators on suicide prevention, MH support. Mental Health Weekly 2015 Jun;25(22):1-7. [doi: 10.1002/mhw.30206]
25. Carpenter DM, Roberts CA, Lavigne JE, Cross WF. Gatekeeper training needs of community pharmacy staff. Suicide Life Threat Behav 2021;51(2):220-228. [doi: 10.1111/sltb.12697] [Medline: 33876495]
26. Caughlan C, Kakuska A, Manthei J, Galvin L, Martinez A, Kelley A, et al. Mind4Health: decolonizing gatekeeper trainings using a culturally relevant text message intervention. Front Public Health 2024;12:1397640 [FREE Full text] [doi: 10.3389/fpubh.2024.1397640] [Medline: 39286750]
27. Cohen E, Pomerance Y, Touati Ohayon L, Brunstein Klomek A. Efficacy of suicide prevention gatekeeper training in Israel: exploring diverse at-risk populations, gender differences, and comparisons between online and in-person training. Death Stud 2025:1-9 [FREE Full text] [doi: 10.1080/07481187.2025.2510482] [Medline: 40442900]
28. Colder Carras M, Bergendahl M, Labrique AB. Community case study: stack up's overwatch program, an online suicide prevention and peer support program for video gamers. Front Psychol 2021;12:575224 [FREE Full text] [doi: 10.3389/fpsyg.2021.575224] [Medline: 33776826]
29. Coleman D, Black N, Ng J, Blumenthal E. Kognito's avatar-based suicide prevention training for college students: results of a randomized controlled trial and a naturalistic evaluation. Suicide Life Threat Behav 2019;49(6):1735-1745. [doi: 10.1111/sltb.12550] [Medline: 30957909]

XSL•FO
RenderX

30. Colucci E, Jaroudy S, Rossmann M. Piloting of a suicide first aid gatekeeper training (online) for children and young people in conflict affected areas in Syria. Int Rev Psychiatry 2022;34(6):640-648 [FREE Full text] [doi: 10.1080/09540261.2022.2100245] [Medline: 36695207]
31. Hawley SR, Skinner T, Young M, St Romain T, Provines J. Suicide prevention across the community: evaluation of mental health training for multiple gatekeeper groups. Kans J Med 2024;17(6):127-132. [doi: 10.17161/kjm.vol17.22524] [Medline: 39758535]
32. Hill RM, McCray CL. Suicide-related stigma and social responsibility moderate the effects of an online suicide prevention gatekeeper training program. Arch Suicide Res 2024;28(2):706-715. [doi: 10.1080/13811118.2023.2199802] [Medline: 37209132]
33. Hill RM, Picou P, Hussain Z, Vieyra BA, Perkins KM. Randomized Controlled Trial of an Online Suicide Prevention Gatekeeper Training Program. Göttingen, Germany: Hogrefe Publishing; 2024:57-64.
34. Hofmann L, Glaesmer H, Przyrembel M, Wagner B. An evaluation of a suicide prevention e-learning program for police officers (COPS): improvement in knowledge and competence. Front Psychiatry 2021;12:770277 [FREE Full text] [doi: 10.3389/fpsyt.2021.770277] [Medline: 34966307]
35. Hofmann L, Wagner B. Efficacy of an online gatekeeper program for relatives of men at risk of suicide - a randomized controlled trial. BMC Public Health 2024;24(1):2693 [FREE Full text] [doi: 10.1186/s12889-024-20193-6] [Medline: 39358752]
36. Kawashima D, Koga Y, Yoshioka M. Feasibility of brief online gatekeeper training for Japanese university students: a randomized controlled trial. Death Stud 2023;47(5):531-540. [doi: 10.1080/07481187.2022.2101076] [Medline: 35855580]
37. Kimbrel NA, Aho NA, Neal LC, Bernes SA, Beaver TA, Hertzberg JA, et al. Development and implementation of web-based safety planning intervention training for firefighter peer support specialists. Crisis 2024;45(2):108-117. [doi: 10.1027/0227-5910/a000924] [Medline: 37727969]
38. Kingi-Ulu'ave D, Framptom C, Cargo T, Stasiak K, Hetrick S. Evaluating the effectiveness of a post-training enhancement to the lifekeepers suicide prevention gatekeeper training. Crisis 2025;46(3):157-165. [doi: 10.1027/0227-5910/a001001] [Medline: 40235266]
39. Kreuze E, Ruggiero KJ. Technology-oriented suicide prevention interventions for adolescents and adolescent gatekeepers: a qualitative review. Adolescent Res Rev 2017;3(2):219-233. [doi: 10.1007/s40894-017-0060-5]
40. Lamis DA, Underwood M, D'Amore N. Outcomes of a suicide prevention gatekeeper training program among school personnel. Crisis 2017;38(2):89-99. [doi: 10.1027/0227-5910/a000414] [Medline: 27561223]
41. Lee-Tauler SY, Grammer J, LaCroix JM, Walsh AK, Clark SE, Holloway KJ, et al. Pilot evaluation of the online 'Chaplains-CARE' program: enhancing skills for united states military suicide intervention practices and care. J Relig Health 2023;62(6):3856-3873. [doi: 10.1007/s10943-023-01882-9] [Medline: 37612485]
42. Liu H, Zheng C, Cao Y, Zeng F, Chen H, Gao W. Gatekeeper training for suicide prevention: a systematic review and meta-analysis of randomized controlled trials. BMC Public Health 2025;25(1):1206 [FREE Full text] [doi: 10.1186/s12889-025-21736-1] [Medline: 40165179]
43. Hart SM, Colucci E, Marzano L. Evaluating suicide prevention gatekeeper training designed to identify and support people from asylum-seeking and refugee backgrounds. BMC Public Health 2024;24(1):2959 [FREE Full text] [doi: 10.1186/s12889-024-20304-3] [Medline: 39455999]
44. Manning J, Vandeusen K. Suicide prevention in the dot com era: technological aspects of a university suicide prevention program. J Am Coll Health 2011;59(5):431-433. [doi: 10.1080/07448480903540507] [Medline: 21500064]
45. Marley G, Lavigne JE, Cross W, Gamble A, Zhang Z, Carpenter DM. Comparing three methods to assess learning outcomes for a suicide prevention training program for pharmacy staff. PEC Innov 2024;5:100348 [FREE Full text] [doi: 10.1016/j.pecinn.2024.100348] [Medline: 39444544]
46. McKay S, Byrne S, Clarke A, Lamblin M, Veresova M, Robinson J. Parent education for responding to and supporting youth with suicidal thoughts (PERSYST): An evaluation of an online gatekeeper training program with Australian parents. Int J Environ Res Public Health 2022;19(9):5025 [FREE Full text] [doi: 10.3390/ijerph19095025] [Medline: 35564419]
47. Mirick RG. Acceptability and feasibility of a brief online suicide prevention training for school staff. J Technol Hum Serv 2025;43(1):34-48. [doi: 10.1080/15228835.2024.2447692]
48. Mishkind MC, Yannacone A, Lopez A, Jortberg BT, Sherrill A, Mescher T. Virtual versus in-person suicide prevention training in the workplace: evaluation of the vitalcog program. J Technol Behav Sci 2023:1-8 [FREE Full text] [doi: 10.1007/s41347-023-00301-w] [Medline: 36742417]
49. Osteen PJ, Ohme K, Morris R, Arciniegas J, Frey JJ, Woods M, et al. Suicide intervention training with law enforcement officers. Suicide Life Threat Behav 2021;51(4):785-794. [doi: 10.1111/sltb.12763] [Medline: 33998030]
50. Perepezko K, Bergendahl M, Kunz C, Labrique A, Carras M, Colder Carras M. "Instead, You're Going to a Friend": evaluation of a community-developed, peer-delivered online crisis prevention intervention. Psychiatr Serv 2024;75(12):1267-1275. [doi: 10.1176/appi.ps.20230233] [Medline: 39054853]
51. Pilbrow S, Staniland L, Uren HV, Shand F, McGoldrick J, Thorp E, et al. Evaluation of an online advanced suicide prevention training for pharmacists. Int J Clin Pharm 2023;45(5):1203-1211 [FREE Full text] [doi: 10.1007/s11096-023-01636-3] [Medline: 37702959]

XSL•FO
RenderX

52. Poštuvan V, Gomboc V, Čopič Pucihar K, Kljun M, Vičič J, Tančič Grum A, et al. Development and evaluation of online suicide preventive tool ialive to increase competences in engaging with a suicidal person. Crisis 2024;45(3):187-196. [doi: 10.1027/0227-5910/a000934] [Medline: 38140805]
53. Quinnett PG. The certified QPR pathfinder training program: A description of a novel public health gatekeeper training program to mitigate suicidal ideation and suicide deaths. J Prev (2022) 2023;44(6):813-824 [FREE Full text] [doi: 10.1007/s10935-023-00748-w] [Medline: 37740846]
54. Reifegerste D, Wagner AJM, Huber L, Fastuca M. Formative evaluation of suicide prevention websites for men: qualitative study with men at risk of suicide and with potential gatekeepers. JMIR Form Res 2025;9:e59829 [FREE Full text] [doi: 10.2196/59829] [Medline: 40009838]
55. Rein BA, McNeil DW, Hayes AR, Hawkins TA, Ng HM, Yura CA. Evaluation of an avatar-based training program to promote suicide prevention awareness in a college setting. J Am Coll Health 2018;66(5):401-411. [doi: 10.1080/07448481.2018.1432626] [Medline: 29461940]
56. Robinson-Link N, Hoover S, Bernstein L, Lever N, Maton K, Wilcox H. Is gatekeeper training enough for suicide prevention? School Mental Health 2019;12(2):239-249. [doi: 10.1007/s12310-019-09345-x]
57. Roslan AF, Pheh KS, Mahadevan R, Bujang SM, Subramaniam P, Yahya HF, et al. Effectiveness of online advanced C.A.R.E suicide prevention gatekeeper training program among healthcare lecturers and workers in national university of Malaysia: a pilot study. Front Psychiatry 2023;14:1009754 [FREE Full text] [doi: 10.3389/fpsyt.2023.1009754] [Medline: 36741120]
58. Ross SG, Pazienza R, Rosa JD. The suicide prevention for college student (SPCS) gatekeepers program. Crisis 2024;45(1):41-47. [doi: 10.1027/0227-5910/a000914] [Medline: 37322902]
59. Ross SG, Pazienza R, Rosa JD. The suicide prevention for college students (SPCS) gatekeepers program: comparing in-person and online training outcomes. J Am Coll Health 2025;73(9):3322-3325. [doi: 10.1080/07448481.2024.2423237] [Medline: 39514812]
60. Schmeckenbecher J, Lentner S, Emilian CA, Plener PL, Baran A, Kapusta ND. E-learning as a tool of suicide prevention training: a meta-analysis and systematic review. Death Stud 2024;48(9):962-974. [doi: 10.1080/07481187.2023.2297058] [Medline: 38133538]
61. Seabury B. On-line, computer-based, interactive simulations: bridging classroom and field. J Technol Hum Serv 2003;22(1):29-48. [doi: 10.1300/j017v22n01_04]
62. Seabury BA. An evaluation of on-line, interactive tutorials designed to teach practice concepts. J Teach Soc Work 2005;25(1-2):103-115. [doi: 10.1300/j067v25n01_07]
63. Shanta Bridges L, Sharma M, Lee JHSH, Bennett R, Buxbaum SG, Reese-Smith J. Using the PRECEDE-PROCEED model for an online peer-to-peer suicide prevention and awareness for depression (SPAD) intervention among African American college students: experimental study. Health Promot Perspect 2018;8(1):15-24 [FREE Full text] [doi: 10.15171/hpp.2018.02] [Medline: 29423358]
64. Smith-Millman M, Bernstein L, Link N, Hoover S, Lever N. Effectiveness of an online suicide prevention program for college faculty and students. J Am Coll Health 2022;70(5):1457-1464. [doi: 10.1080/07448481.2020.1804389] [Medline: 32813627]
65. Stone DM, Barber CW, Potter L. Public health training online: the national center for suicide prevention training. Am J Prev Med 2005;29(5 Suppl 2):247-251. [doi: 10.1016/j.amepre.2005.08.019] [Medline: 16376726]
66. Stover AN, Lavigne JE, Shook A, MacAllister C, Cross WF, Carpenter DM. Development of the pharm-SAVES educational module for gatekeeper suicide prevention training for community pharmacy staff. Health Expect 2023;26(3):1246-1254 [FREE Full text] [doi: 10.1111/hex.13741] [Medline: 36852881]
67. Sun Y, Zhang Q, Wu W, Lin J, Sun S, An J, et al. Efficacy of a localized caregiver gatekeeper training program for suicide prevention among Chinese adolescents: a pilot study. Asian J Psychiatr 2025;109:104555. [doi: 10.1016/j.ajp.2025.104555] [Medline: 40449414]
68. Teo AR, Call AA, Hooker ER, Fong C, Karras E, Dobscha SK. Feasibility of recruitment and retention in a remote trial of gatekeeper training for close supports of military veterans: mixed methods study. Contemp Clin Trials Commun 2022;30:100993 [FREE Full text] [doi: 10.1016/j.conctc.2022.100993] [Medline: 36159001]
69. Teo AR, Hooker ER, Call AA, Dobscha SK, Gamble S, Cross WF, et al. Brief video training for suicide prevention in veterans: a randomized controlled trial of VA S.A.V.E. Suicide Life Threat Behav 2024;54(1):154-166 [FREE Full text] [doi: 10.1111/sltb.13028] [Medline: 38095049]
70. Timmons-Mitchell J, Albright G, McMillan J, Shockley K, Cho S. Virtual role-play: middle school educators addressing student mental health. Health Behav Policy Rev 2019;6(6):546-557. [doi: 10.14485/hbpr.6.6.1]
71. Wislocki K, Jager-Hyman S, Brady M, Weiss M, Schaechter T, Khazanov G, et al. Freely available training videos for suicide prevention: scoping review. JMIR Ment Health 2023;10:e48404 [FREE Full text] [doi: 10.2196/48404] [Medline: 37921847]
72. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
73. Mikkonen K, Kääriäinen M. Content Analysis in Systematic Reviews. Cham: Springer; 2020:105-115.

XSL•FO
RenderX

74. Kingi-Uluave D, Taufa N, Tuesday R, Cargo T, Stasiak K, Merry S, et al. A review of systematic reviews: gatekeeper training for suicide prevention with a focus on effectiveness and findings. Arch Suicide Res 2025;29(2):329-346 [FREE Full text] [doi: 10.1080/13811118.2024.2358411] [Medline: 38884349]
75. Martin F, Sun T, Westine CD. A systematic review of research on online teaching and learning from 2009 to 2018. Comput Educ 2020;159:104009 [FREE Full text] [doi: 10.1016/j.compedu.2020.104009] [Medline: 32921895]
76. Greenhow C, Graham CR, Koehler MJ. Foundations of online learning: challenges and opportunities. Educ Psychol 2022;57(3):131-147. [doi: 10.1080/00461520.2022.2090364]
77. Kingi-Ulu'ave D, Framptom C, Cargo T, Stasiak K, Hetrick S. Evaluating the impact and cultural relevance of lifekeepers gatekeeper training across three training modalities. Crisis 2024;45(6):417-424. [doi: 10.1027/0227-5910/a000977] [Medline: 39545404]
78. Melhem N, Moutier CY, Brent DA. Implementing evidence-based suicide prevention strategies for greatest impact. Focus (Am Psychiatr Publ) 2023;21(2):117-128 [FREE Full text] [doi: 10.1176/appi.focus.20220078] [Medline: 37201145]
79. Zalsman G, Hawton K, Wasserman D, van Heeringen K, Arensman E, Sarchiapone M, et al. Suicide prevention strategies revisited: 10-year systematic review. Lancet Psychiatry 2016;3(7):646-659. [doi: 10.1016/S2215-0366(16)30030-X] [Medline: 27289303]
80. Bandura A. Social cognitive theory: an agentic perspective. Annu Rev Psychol 2001;52:1-26. [doi: 10.1146/annurev.psych.52.1.1] [Medline: 11148297]
81. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
82. Brieger E, Arghode V, McLean G. Connecting theory and practice: reviewing six learning theories to inform online instruction. Eur J Train Dev 2020;44(4/5):321-339. [doi: 10.1108/ejtd-07-2019-0116]
83. Shea P, Richardson J, Swan K. Building bridges to advance the community of inquiry framework for online learning. Educ Psychol 2022;57(3):148-161. [doi: 10.1080/00461520.2022.2089989]
84. Mayer RE. The past, present, and future of the cognitive theory of multimedia learning. Educ Psychol Rev 2024;36(1):8. [doi: 10.1007/s10648-023-09842-1]
85. van der Feltz-Cornelis CM, Sarchiapone M, Postuvan V, Volker D, Roskar S, Grum AT, et al. Best practice elements of multilevel suicide prevention strategies: a review of systematic reviews. Crisis 2011;32(6):319-333 [FREE Full text] [doi: 10.1027/0227-5910/a000109] [Medline: 21945840]
86. Beautrais A, Fergusson D, Coggan C, Collings C, Doughty C, Ellis P, et al. Effective strategies for suicide prevention in New Zealand: a review of the evidence. N Z Med J 2007;120(1251):U2459. [Medline: 17384687]
87. Mann JJ, Apter A, Bertolote J, Beautrais A, Currier D, Haas A, et al. Suicide prevention strategies: a systematic review. JAMA 2005;294(16):2064-2074. [doi: 10.1001/jama.294.16.2064] [Medline: 16249421]
88. Spafford SG, Silverman MM, Gutierrez PM. What is known about suicide prevention gatekeeper training and directions for future research. Suicide Life Threat Behav 2025;55(1):e13130. [doi: 10.1111/sltb.13130] [Medline: 39351789]
89. Creighton G, Oliffe J, Ogrodniczuk J, Frank B. "You've Gotta Be That Tough Crust Exterior Man": depression and suicide in rural-based men. Qual Health Res 2017;27(12):1882-1891. [doi: 10.1177/1049732317718148] [Medline: 28936927]
90. Farvid P, Vance TA, Klein SL, Nikiforova Y, Rubin LR, Lopez FG. The health and wellbeing of transgender and gender non-conforming people of colour in the United States: a systematic literature search and review. J Community Appl Soc Psychol 2021;31(6):703-731. [doi: 10.1002/casp.2555]
91. Clark KD, Lunn MR, Bosse JD, Sevelius JM, Dawson-Rose C, Weiss SJ, et al. Societal stigma and mistreatment in healthcare among gender minority people: a cross-sectional study. Int J Equity Health 2023;22(1):162 [FREE Full text] [doi: 10.1186/s12939-023-01975-7] [Medline: 37620832]
92. Wexler L, White J, Trainor B. Why an alternative to suicide prevention gatekeeper training is needed for rural indigenous communities: presenting an empowering community storytelling approach. Crit Public Health 2015;25(2):205-217 [FREE Full text] [doi: 10.1080/09581596.2014.904039] [Medline: 36779086]
93. Arnold D, Glässel A, Böttger T, Sarma N, Bethmann A, Narimani P. "What Do You Need? What Are You Experiencing?" relationship building and power dynamics in participatory research projects: critical self-reflections of researchers. Int J Environ Res Public Health 2022;19(15):9336 [FREE Full text] [doi: 10.3390/ijerph19159336] [Medline: 35954692]
94. Mulvale A, Miatello A, Hackett C, Mulvale G. Applying experience-based co-design with vulnerable populations: lessons from a systematic review of methods to involve patients, families and service providers in child and youth mental health service improvement. Patient Exp J 2016;3(1):117-129. [doi: 10.35680/2372-0247.1104]
95. Qasim R, Farooqui WA, Rahman A, Haroon R, Saleem M, Rafique M, et al. Community centred co-design methodology for designing and implementing socio-behavioural interventions to counter COVID-19 related misinformation among marginalized population living in the squatter settlements of Karachi, Pakistan: a methodology paper. BMC Proc 2023;17(Suppl 7):15 [FREE Full text] [doi: 10.1186/s12919-023-00265-y] [Medline: 37438805]
96. Kral MJ, Kidd S. Community-based participatory research and community empowerment for suicide prevention. In: A Positive Psychological Approach to Suicide : Theory, Research, and Prevention. Cham: Springer International Publishing; 2018:285-299.

XSL•FO
RenderX

97. Chudyk AM, Stoddard R, Duhamel TA, Patient Engagement in Research Partners, Schultz ASH. Future directions for patient engagement in research: a participatory workshop with Canadian patient partners and academic researchers. Health Res Policy Syst 2024;22(1):24 [FREE Full text] [doi: 10.1186/s12961-024-01106-w] [Medline: 38350974]
98. Thomas E, Benjamin-Thomas TE, Sithambaram A, Shankar J, Chen S. Participatory action research among people with serious mental illness: a scoping review. Qual Health Res 2024;34(1-2):3-19 [FREE Full text] [doi: 10.1177/10497323231208111] [Medline: 37929751]
99. Wallerstein N, Muhammad M, Sanchez-Youngman S, Rodriguez Espinosa P, Avila M, Baker EA, et al. Power dynamics in community-based participatory research: a multiple-case study analysis of partnering contexts, histories, and practices. Health Educ Behav 2019;46(1_suppl):19S-32S. [doi: 10.1177/1090198119852998] [Medline: 31549557]
100. Haarmans M, Nazroo J, Kapadia D, Maxwell C, Osahan S, Edant J, et al. The practice of participatory action research: complicity, power and prestige in dialogue with the 'racialised mad'. Sociol Health Illn 2022;44 Suppl 1(Suppl 1):106-123 [FREE Full text] [doi: 10.1111/1467-9566.13517] [Medline: 36001350]
101. Brush BL, Mentz G, Jensen M, Jacobs B, Saylor KM, Rowe Z, et al. Success in long-standing community-based participatory research (CBPR) partnerships: a scoping literature review. Health Educ Behav 2020;47(4):556-568 [FREE Full text] [doi: 10.1177/1090198119882989] [Medline: 31619072]
102. Park M, Lee K, Estein O. “That was sexy” to “hope for change”: resituating an ethics of power in the capacity to aspire in public and patient-involved research. In: Meaningful and Safe: The Ethics and Ethical Implications of Patient and Public Involvement in Health and Medical Research. Suffolk, UK: Ethics Press; 2024:89-119.
103. Popay J, Whitehead M, Ponsford R, Egan M, Mead R. Power, control, communities and health inequalities I: theories, concepts and analytical frameworks. Health Promot Int 2021;36(5):1253-1263 [FREE Full text] [doi: 10.1093/heapro/daaa133] [Medline: 33382890]
104. Rose D, Kalathil J. Power, privilege and knowledge: the untenable promise of co-production in mental "Health". Front Sociol 2019;4:57 [FREE Full text] [doi: 10.3389/fsoc.2019.00057] [Medline: 33869380]
105. Flexible adult learning provision: what it is, why it matters, and how to make it work. OECD. 2023. URL: https://www.oecd.org/content/dam/oecd/en/topics/policy-sub-issues/adult-learning/booklet-flexibility-2023.pdf [accessed 2025-12-13]
106. Aladar A, Ilg D, King J, Hajja A. EduKona: A customizable, mobile-friendly platform for enhanced educational engagement and collaboration. : IEEE; 2024 Presented at: IEEE Frontiers in Education Conference (FIE); 2024 October 13-16; Washington, DC, USA p. 13-16. [doi: 10.1109/fie61694.2024.10892814]
107. Chernikova O, Heitzmann N, Stadler M, Holzberger D, Seidel T, Fischer F. Simulation-based learning in higher education: a meta-analysis. Rev Educ Res 2020;90(4):499-541. [doi: 10.3102/0034654320933544]
108. Elendu C, Amaechi DC, Okatta AU, Amaechi EC, Elendu TC, Ezeh CP, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024;103(27):e38813 [FREE Full text] [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
109. Lavanya KM, Somu LK, Mishra SK. Effectiveness of scenario-based roleplay as a method of teaching soft skills for undergraduate medical students. Int J Appl Basic Med Res 2024;14(2):78-84 [FREE Full text] [doi: 10.4103/ijabmr.ijabmr_431_23] [Medline: 38912358]
110. Salloum SA, Salloum A, Alfaisal R, Basiouni A, Shaalan K. Predicting student adaptability to online education using machine learning. In: Breaking Barriers with Generative Intelligence Using GI to Improve Human Education and Well-Being. Cham: Springer Nature Switzerland; 2024.
111. Mojtahedzadeh R, Hasanvand S, Mohammadi A, Malmir S, Vatankhah M. Students' experience of interpersonal interactions quality in e-Learning: a qualitative research. PLoS One 2024;19(3):e0298079 [FREE Full text] [doi: 10.1371/journal.pone.0298079] [Medline: 38530814]
112. Hill JR, Song L, West RE. Social learning theory and web-based learning environments: a review of research and discussion of implications. Am J Distance Educ 2009;23(2):88-103. [doi: 10.1080/08923640902857713]
113. Ouyang F, Zheng L, Jiao P. Artificial intelligence in online higher education: a systematic review of empirical research from 2011 to 2020. Educ Inf Technol 2022;27(6):7893-7925. [doi: 10.1007/s10639-022-10925-9]
114. Nel JM. The The implications of artificial intelligence in online education: a critical examination in the context of philosophy and ethics. In: Towards a Hybrid, Flexible and Socially Engaged Higher Education. Cham: Springer Nature Switzerland; 2024.
115. Mougiakou E, Papadimitriou S, Chrysafiadi K, Virvou M. Artificial intelligence in educational games and consent under general data protection regulation. Int Decis Technol 2025;19(2):670-686. [doi: 10.1177/18724981251322884]
116. Holmes G, Tang B, Gupta S, Venkatesh S, Christensen H, Whitton A. Applications of large language models in the field of suicide prevention: scoping review. J Med Internet Res 2025;27:e63126 [FREE Full text] [doi: 10.2196/63126] [Medline: 39847414]
117. Buchholz BA, DeHart J, Moorman G. Digital citizenship during a global pandemic: moving beyond digital literacy. J Adolesc Adult Lit 2020;64(1):11-17 [FREE Full text] [doi: 10.1002/jaal.1076] [Medline: 32834710]
118. Withers EM. The Digital Divide and Health: Examining Digital Access as a Social Determinant of Health. Downtown Portland: Portland State University; 2021.

XSL•FO
RenderX

119. Sitzmann T, Ely K, Bell BS, Bauer KN. The effects of technical difficulties on learning and attrition during online training. J Exp Psychol Appl 2010;16(3):281-292. [doi: 10.1037/a0019968] [Medline: 20853987]
120. Heeren T, Ward C, Sewell D, Ashida S. Applying network analysis to assess the development and sustainability of multi-sector coalitions. PLoS One 2022;17(10):e0276114 [FREE Full text] [doi: 10.1371/journal.pone.0276114] [Medline: 36256640]
121. Hawgood J, Woodward A, Quinnett P, Leo DD. Gatekeeper Training and Minimum Standards Competency: Essentials for the Suicide Prevention Workforce. Newburyport, Massachusetts: Hogrefe Publishing; 2022:516-522.
122. Morton M, Wang S, Tse K, Chung C, Bergmans Y, Ceniti A, et al. Gatekeeper training for friends and family of individuals at risk of suicide: a systematic review. J Community Psychol 2021;49(6):1838-1871. [doi: 10.1002/jcop.22624] [Medline: 34125969]
123. Spafford SG, McWhirter Boisen MR, Tanner-Smith EE, Rodriguez G, Muruthi JR, Seeley JR. The effects of suicide prevention gatekeeper training on behavioral intention and intervention behavior: a systematic review and meta-analysis. Prev Sci 2024;25(6):978-988. [doi: 10.1007/s11121-024-01710-w] [Medline: 39023720]
124. Atmakuru A, Shahini A, Chakraborty S, Seoni S, Salvi M, Hafeez-Baig A, et al. Artificial intelligence-based suicide prevention and prediction: a systematic review (2019–2023). Inf Fusion 2025;114:102673. [doi: 10.1016/j.inffus.2024.102673]
125. Li X, Chen F, Ma L. Exploring the potential of artificial intelligence in adolescent suicide prevention: current applications, challenges, and future directions. Psychiatry 2024;87(1):7-20. [doi: 10.1080/00332747.2023.2291945] [Medline: 38227496]

Abbreviations

AI: artificial intelligence
GTP: gatekeeper training program
LGBTQ+: lesbian, gay, bisexual, trans, queer, and other sexual and gender minorities
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
S.A.V.E./SAVE: signs, ask, validate, encourage/expedite
WHO: World Health Organization

Edited by A Stone; submitted 31.Jul.2025; peer-reviewed by A Shivanna, G Holmes; comments to author 20.Aug.2025; revised version received 25.Nov.2025; accepted 25.Nov.2025; published 05.Feb.2026.

Please cite as:
Ferlatte O, Gareau E, Lee K, Wassef K, Oliffe JL, Kia H, Dumville B
Key Components and Barriers in Web-Based Suicide Prevention Gatekeeper Training: Systematic Narrative Review
J Med Internet Res 2026;28:e81572
URL: https://www.jmir.org/2026/1/e81572
doi:10.2196/81572
PMID:

©Olivier Ferlatte, Emmanuelle Gareau, Keven Lee, Kinda Wassef, John Lindsay Oliffe, Hannah Kia, Brock Dumville. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Interventions That Use Highly Visual Social Media Platforms to Tackle Unhealthy Body Image in Adolescents and Young Adults: Systematic Review of Randomized Controlled Trials and Quasi-Experimental Studies

Anthony Copez-Lonzoy[1,2], MSc; Janina Bazalar-Palacios[3], MSc; Nikol Mayo-Puchoc[4], BS; Gustavo Calderon de la Cruz[5], MSc; Liliana Cruz-Ausejo[6], MSc; Evelyn Goicochea-Ríos[7], PhD; Juan Carlos Bazo-Alvarez[8,9], PhD

[1]Bibliometrics Unit, Universidad San Ignacio de Loyola, Lima, Lima Province, Peru

[2]Research Unit, PSYCOPERU Peruvian Institute for Psychological and Psychosocial Research, Ancash, Chimbote, Peru

[3]Facultad de Ciencias de la Salud, Universidad Tecnológica del Perú, Lima, Lima Province, Peru

[4]Mental Health Unit, Instituto Peruano de Orientación Psicológica, Lima, Peru

[5]School of Psychology, Peruvian University of Applied Sciences, Lima, Peru

[6]School Medicine, National University of San Marcos, Lima, Peru

[7]Grupo MedFam, Escuela de Medicina, Universidad César Vallejo, Trujillo, Peru

[8]Research Department of Primary Care and Population Health, University College London, London, United Kingdom

[9]Grupo MedFam, Escuela de Psicología, Universidad Cesar Vallejo, Trujillo, Peru

Corresponding Author:
Juan Carlos Bazo-Alvarez, PhD
Research Department of Primary Care and Population Health
University College London
Gower Street
London, WC1E 6BT
United Kingdom
Phone: 1 +44 7376 076260
Email: juan.alvarez.16@ucl.ac.uk

Abstract

Background: Highly visual social media (HVSM) platforms such as Facebook (Meta Platforms, Inc), Instagram (Meta Platforms, Inc), TikTok (ByteDance Ltd), and Snapchat (Snap Inc) have become central to the digital lives of adolescents and young adults. While these platforms have been linked to body dissatisfaction, they are also increasingly used as vehicles for health promotion. However, the evidence on interventions delivered through HVSM to address body image issues remains fragmented.

Objective: This review aimed to synthesize available evidence on interventions using HVSM platforms to reduce negative body image in adolescents and young adults.

Methods: We conducted a systematic search across 5 electronic databases (Scopus, MEDLINE, APA Psynet, Embase, and Web of Science) for studies published between January 2012 and October 2025. Eligible studies included experimental or quasi-experimental designs evaluating the effect of an HVSM-based intervention on body image outcomes in individuals aged 13 to 35 years. Risk of bias was assessed using the Risk of Bias Tool 2.0 (Cochrane) and was conducted independently by 2 researchers.

Results: Eight studies met the inclusion criteria with 4975 participants (2612 in intervention groups and 2363 in control groups). Most studies were conducted in high-income countries and had predominantly female participants. The interventions varied widely in format, duration, and theoretical basis. Microinterventions, brief interactive strategies such as gamified chatbots or short videos, were the most common and had moderate effects. Stimulus-based interventions using content with a positive body image or that did not focus on appearance were also identified, achieving moderate effects (η_p^2<.07), as well as combined approaches that integrated digital and face-to-face components to reduce negative body image (P<.001). The use and functionality of interventions using social media platforms were also compared by gender.

XSL•FO
RenderX

Conclusions: Body image management platforms offer an emerging avenue for implementing body image interventions in adolescents and young adults. While current evidence suggests modest benefits, the high heterogeneity among presentation formats and the variability in duration make comparisons between these studies difficult. This review synthesizes social media–delivered interventions for body image disturbance, going beyond broader digital approaches centered on websites or apps. It identifies cross-platform, putative mechanisms of action and common intervention formats, highlighting the potential of brief interventions for scalable reach and user empowerment via content curation. These findings define targets for optimization and underscore the need for platform safeguards and supportive policy and regulatory frameworks to enable safe real-world implementation, particularly for adolescents.

(J Med Internet Res 2026;28:e80141) doi:10.2196/80141

KEYWORDS

adolescent; body image; digital intervention; microinterventions; social media; systematic review; young adult

Introduction

Body image (BI) is a person's perception of their physical appearance and the potential positive or negative thoughts about self-worth based on that perception [1]. The problems associated with this negative perception of BI can affect various levels of health, such as eating disorders, risky behaviors, and excessive dieting [2]. This generates a public health problem affecting diverse age and gender populations [3]. Among the various mental health problems, negative BI remains a major concern for young people, especially adolescent girls [4]. According to research, between 20% and 40% of adolescent girls experience severe body dissatisfaction, which often persists or worsens during adulthood [4,5]. However, this prevalence of negative BI can reach 35% to 81% in adolescent girls and 16% to 55% in adolescent boys in low- and middle-income countries [5].

There is a strong correlation between BI and self-esteem. Adolescents and young adults frequently develop erroneous self-perceptions as a result of unrealistic physique ideals promoted by social media platforms [6]. According to research, excessive use of social media increases body dissatisfaction, promotes unfavorable comparisons, and lowers self-esteem [7]. This combination of low self-esteem and distorted BI increases the risk of developing eating disorders such as anorexia and bulimia nervosa [8]. Systematic reviews show that body dissatisfaction predicts the beginning of eating disorders and is linked to long-term mental health difficulties such as sadness and anxiety [9].

Adolescents and young adults are among the most active social media users, with more than 90% reporting frequent use of at least one site such as Instagram, TikTok, or Snapchat (Snap Inc) [10]. Globally, social media users reach 60% of the population, with adolescents and young adults frequently having accounts on various platforms and switching between them depending on age, cultural influences, and geographic location [11]. These platforms are intentionally designed to increase user engagement with features such as tailored content feeds, targeted notifications, and tools for modifying and sharing multimedia material [12]. Highly visual social media (HVSM) platforms, such as Instagram and TikTok, have transformed the digital environment by emphasizing photos and videos, making visual appeal an important driver of user retention [13,14]. These systems' algorithms continuously adjust content delivery to match user preferences, increasing interactivity and encouraging long-term engagement [15]. This dynamic produces a digital environment that captivates adolescents and young adults while also influencing their actions and beliefs, especially those connected to self-image and mental health [16].

Though highly visually appealing social media (HVSM) platforms are designed to enhance user engagement and dependency, they also have great potential as tools for mental health screening and intervention among teenagers and young adults. Recent research shows how feasible and successful digital mental health interventions through social media (eg, including HVSM platforms) [17-20]. These types of interventions overcome the accessibility and affordability limitations of traditional mental health services [21]. They can be administered at scale, are cost-effective to implement, and are easy to train, improving access to mental health services for individuals, thereby reducing the workload of health personnel [22].

Despite these developments, there are still important gaps in methodically summarizing treatments aimed at poor BI using HVSM. For example, little data exist on which particular platforms best handle BI issues and the special characteristics of interventions catered to these platforms [23]. Nevertheless, some alternatives use affordances of HVSM, such as specific algorithms, visual material, and interaction, to assess and minimize negative concerns about BI among adolescents and young adults. Nonetheless, further investigation is required to find best practices for using HVSM-based therapies, so that they are accessible, culturally relevant, and able to generate quantifiable changes in BI and general mental health [24,25], ultimately leading to the creation of creative and scalable solutions for urgent problems.

We aim to summarize evidence on the effectiveness of interventions that use HVSM platforms to tackle unhealthy BI in adolescents and young adults.

Methods

Overview

The systematic review was conducted according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Search Extension) guidelines [26-28] (Multimedia Appendices 1 and 2). The study protocol was published in open access [29].

XSL•FO
RenderX

Eligibility Criteria

Only studies included (1) adolescent and/or young adult populations (only groups with an age range from 13 to 35 years were included), (2) randomized clinical trials or quasi-experimental studies that evaluated BI interventions, (3) studies that fully or partially used social media as an intervention (eg, Facebook [Meta Platforms, Inc], Instagram [Meta Platforms, Inc], and TikTok [ByteDance Ltd)], or other interactive social media), and (4) studies that evaluated outcomes directly related to BI (eg, body dissatisfaction, body satisfaction, self-esteem, disordered eating, and appearance satisfaction). Studies that meet any of these criteria were excluded (1) studies that are not original journal articles (letters, book chapters, systematic reviews, and meta-analyses), (2) studies that cannot be considered interventions (eg, without pre or postevaluation and random assignment), or that are comparison conditions (eg, stimuli only), (3) studies that include a higher age range than established (eg, university students more than 35 years old at the time of the intervention), (4) studies that use social media to build the intervention, and (5) studies before 2012 (the first study that used social media as part of the intervention dates back to 2012) [30].

Information Sources

The initial search was conducted during 2024 and the final search was conducted up to October 10, 2025. Five databases were consulted (Scopus, Embase, Web of Science, Medline, and PsycINFO) in their entirety (ie, only studies that are listed in the databases). Only the information published in the studies was taken into account. We examined MeSH (Medical Subject Headings) terms for relevance and frequency of potentially included studies. We used validated filters to locate adolescent-based and young adult–based HVSM interventions where appropriate (eg, title, keyword, or abstract). AC-L and JB-P conducted citation tracking of initially included studies included in Scopus, identifying previously excluded studies and some of the studies included in the final review. Throughout the search process, standardized terms (eg, MeSH for MEDLINE) and Boolean operators (AND, OR) were applied to ensure a replicable process for each database.

Search Strategy

The following search strategy was used, based on the literature on BI and social media interventions in adolescents and young adults. In addition, publication year filters were included (eg, pubyear >2011). We include search strategies on BI, Mahon and Seekis [31], and adapt them to the needs of this study. The search strategies were peer reviewed by 2 team members (ACL and JBP) and an external expert (senior information specialist) prior to execution. All identified fields (eg, social media AND intervention) were unified to generate the search strategy. BI terms (appearance OR “body anxiety” OR “body attitude” OR “body attitudes” OR “body checking” OR “body concern*” OR “body esteem” OR “body evaluation” OR “body satisfaction” OR “body dissatisfaction” OR “body image” OR “body functionality” OR “body surveillance” OR “body shame” OR “body positive” OR “body positivity” OR “body acceptance” OR “body pride” OR “body preoccupation” OR “body awareness” OR “body anxiety” OR “body appreciation” OR “body neutrality” OR “physical self-perception” OR “physical self-perception” OR “physical self-concept” OR “physical self-concept” OR “shape concern” OR “shape concerns” OR “weight concern*” OR “shape satisfaction” OR “shape dissatisfaction” OR interocep* OR “weight dissatisfaction” OR “weight satisfaction” OR “body ideal*” OR “appearance ideal*” OR muscularity OR “thin ideal” OR “physical appearance” OR muscular* OR “social physique anxiety” OR “self-objectification” OR self-objectification OR “objectified body consciousness” OR thinness OR “cognitive restraint” OR “body identity” OR “body schema” OR “body representation*” OR “self-esteem” OR “eating disorders”), social media terms (internet OR “social media” OR “social network site*” OR instagram OR facebook OR “tik tok” OR “TikTok” OR twitter OR reddit OR telegram OR whapsapp OR weibo OR snapchat OR youtube OR pinterest), population terms (Child* OR adolescent* OR young OR youth OR teen* OR college OR universit* OR “Young Adult*”) randomized controlled trial (RCT) and quasi-experimental terms (intervention* OR “controlled trial” OR “clinical trial” OR “major clinical trial” OR “quasi-experimental” OR “quasi experimental”).

We limited the results to studies from 2012 onward, as this year saw the proliferation of large-scale studies on social media (eg, Facebook, Instagram, TikTok, or other HVSMs) [13]. The search strategy for each database is in Zenodo [32] or refer to Multimedia Appendix 3.

Selection Process and Data Collection

After eliminating duplicates, 2 researchers (ACL and JBP) independently assessed the titles and abstracts of all articles based on established criteria using Rayyan software (Qatar Computing Research Institute) [33]. We removed results that did not provide sufficient information from the full-text review. We repeated this process for the full-text phase. Data were extracted using a digital form predesigned by the research team. The form contains information on (1) general characteristics of the studies: first author, journal, country, study design (eg, RCT and quasi-experimental), participants or target population (Table 1); (2) social media: social media components and features used for the intervention or evaluation; (3) interventions: type of intervention, level of intervention (eg, country and hospital), duration of intervention, methodological details (eg, details of the design for the effectiveness evaluation, sample size, and instruments), effectiveness estimate and/or its estimation (eg, mean difference); and (4) assessment tools: construct or object to be assessed (eg, image, video, chatbot, or others), self-reports used one or more dimensions [34,35] (scales or inventories; Table 2).

XSL•FO
RenderX

Table 1. Description of intervention characteristics and highly visual social media platform features in 8 studies [24,25,36-41] to improve body image.a

Author (year)	Country	Body image outcomes	Social media platform	Affordances	Duration	Delivered by	Intervention description
Sampson et al [25] (2020)	United Kingdom	Body and facial dissatisfaction	Instagram	Photos	2 weeks	Researchers	They explored the assigned images on specific individual devices (eg, iPads), comparing the viewing of idealized content (eg, attractive smiles) and neutral images in Instagram users for 5 minutes.
Seekis et al [40] (2020)	Australia	Body dissatisfaction, drive for thinness, body appreciation, and self-compassion	Facebook	Facebook group	2 weeks	Researchers: workshop and Facebook group moderator	Attendance at 50-minute in-person workshops (6 workshops) held for 2 days, with an average of 12 participants each. These workshops are based on 6 strategies from the Mindful Self-Compassion manual, delivered as psychoeducation on self-compassion for body image. In addition, discussions are held in a private Facebook group (3 times per 2 weeks) to share experiences and apply what was learned in the workshop.
Fioravanti et al [38] (2023)	Italy	Body satisfaction, appearance comparison	Instagram	Images	28 days	Researchers	Researchers created Instagram profiles based on "body positivity," "encompassing fitspiration," and "neutral content" images, creating images for each profile type (approximately 112 images). Participants were randomly assigned to each group to follow their respective profiles. Content was delivered daily (1 post + 3 Instagram stories) for 28 days. In addition, participants received 2 messages per day asking questions about the content they received on Instagram.
Garbett et al [41] (2023)	Indonesia	Body satisfaction, attitudes toward appearance	Facebook and Instagram	Video	30 minutes	Self-guided	Self-guided web-based activities on 6 sequential videos (~ 5 minutes each) on body image issues related to Indonesian youth, complemented by social media activities.
Pilot et al [39] (2023)	United States	Body dissatisfaction	Instagram	Images	Not clear	Researchers	Progressive viewing of 20 pairs of fitspiration images on Instagram contrasted with reading and generating a list of 12 personal values and identifying the most important value.
Lewis-Smith et al [37] (2023)	United States	Body satisfaction, acceptance of appearance diversity and appearance-related internalized racism	YouTube	Video	5 minutes	Researchers	A theoretically grounded microintervention in the form of a dramatic miniseries (first episode only) aimed at improving body satisfaction, internalized appearance-related racism, and acceptance of appearance diversity (specifically, acceptance of larger-bodied people of Black people) among Black and non-Black adolescents (American Indian and Alaska Native, Asian, Native Hawaiian and Other Pacific Islands or White) in the United States.
Matheson et al [24] (2023)	Brazil	Body satisfaction, positive and negative affect	Facebook	Chatbot	3 days	Avatar	A chatbot operated through Facebook Messenger that provided microinterventions for improving body image and mental well-being based on psychoeducational techniques, interactive activities, and coping strategies.
Fardouly et al [36] (2023)	Australia	Body dissatisfaction and appreciation, self-objectification, appearance comparison tendency, and body activism	Facebook	Facebook group	14 days	Social media	Microintervention created in private Facebook groups was based on the exposure of body-positive content. Two different conditions were administered (positive and neutral body awareness). Participants were dosed with posts of content relevant to each condition for 14 consecutive days (3 times a day).

[a]Affordances = set of tools, characteristics, and options of social media platforms for interaction, creation, management, and sharing between users (eg, photos, private or open groups, text, video, etc).

XSL•FO
RenderX

Table 2. Essential characteristics and key findings from randomized controlled trials on the effectiveness of social media interventions to improve body image[a].

Author (year)	Sample size	Age	Study design	Follow-up	Comparison group	Body images measures	Key findings	Did the intervention achieve its goals?
Sampson et al [25] (2020)	132	Mean 20.50 (2.21), range 18-30 years	Randomized controlled trial	N/A[b]	Yes, images of neutral nature	BSS[c]; FSS[d]; SACS[e]; SDI[f]	Exposure to "ideal" facial images on social media decreases facial satisfaction (P<.001).	Partially
Seekis et al [40] (2020)	76	Mean 18.04 (0.90), range 17-21 years	Cluster randomized trial	1-3 months	Yes, waitlist	Subscale EDI[g]; SAAS[h]; UPACS[i]; BAS-2[j]; SCS-SF[k]	The combined strategy (workshop + Facebook group) based on self-compassion reduced body dissatisfaction (β=–.74; P<.001).	Yes
Garbett et al [41] (2023)	1847	Intervention mean 16.94 (SD 1.40) years; control: mean 17 (SD 1.41), range 15-19 years	Randomized controlled trial	1 month	Yes, waitlist	BESAA[l]; SATAQ-3[m]; PANAS-C[n]	Warna-Warni Waktu is an effective eHealth intervention to reduce body dissatisfaction among adolescent girls and young women (P=.005, η^2_p=0.05).	Yes
Lewis-Smith et al [37] (2023)	686	Mean 15.72 (SD 1.60), range 13-18 years	Parallel randomized controlled trial	N/A	Yes, non–body image–related video	VAS[o]; subscale IROS[p]	No changes in body satisfaction were observed. The changes identified were only in conditions of racism and acceptance of diversity (P<.001; Cohen d>0.17).	No
Matheson et al [24] (2023)	1715	Range 13-18 years (52.54% girls)	Randomized controlled trial	1 week; 1 month	Yes, assessment-only control	BESAA; PANAS; BIS-ES[q]	Body self-esteem identified improvements in positive (Cohen d=0.13; P=.02) and negative (Cohen d=0.11; P=.047) appearance.	Yes
Fardouly et al [36] (2023)	159	Mean 20.92 (SD 2.21), 18-25 years	Parallel randomized controlled trial	4 weeks	Yes, assessment-only control	Subscale EDI; BAS-2; PANAS-SF[r]; SOBBS[s]; PACS[t]; SAM[u] modified	Body dissatisfaction was reduced in the body positive Facebook group condition (P=.03; η_p^2=0.05) with small effect.	Yes
Pilot et al [39] (2023)	238	Mean 19.89 (SD 1.25), 18-22 years	Randomized controlled trial	N/A	Yes, travel images	Subscale ED-EQ[v]; WREQ[w]; VAS	The intervention based on Values Affirmation + Fitspiration does not show significant changes in body image (P=.73) or in negative mood (P=.60) compared to the control.	No

[a]Relevant results for BI = The information was assessed by the team as "no" (no favorable results for body image), "partially" (favorable results for some body image outcomes), and "yes" (favorable results for body image outcomes).

[b]N/A: not applicable.

[c]BSS: Body Satisfaction Scale.

[d]FSS: Facial Satisfaction Scale.

[e]SACS: State Appearance Comparison Scale.

[f]SDI: Self-Discrepancy Index.

[g]EDI-3: Eating Disorder Inventory.

[h]SAAS: Social Appearance Anxiety Scale.

[i]UPACS: Upward Physical Appearance Comparison Scale.

[j]BAS-2: Body Appreciation Scale-2.

[k]SCS-SF: Self-Compassion Scale–Short-Form.

[l]BESAA: Body Esteem Scale for Adolescents and Adults.

[m]SATAQ-3 Sociocultural Attitudes Towards Appearance Questionnaire.

[n]PANAS-C: Positive and Negative Affect Schedule for Children.

[o]VAS: Visual Analogue Scales.

XSL•FO
RenderX

[p]IROS: Internalized Racial Oppression Scale (Alteration of Physical Appearance and Hair Change subscales).

[q]BISES: Body Image Self-Efficacy Scale.

[r]PANAS-SF: Positive and Negative Affect Scale-Short Form.

[s]SOBBS: Self-Objectification Beliefs and Behaviors Scale.

[t]PACS: Physical Appearance Comparison Scale.

[u]SAM: modified version of the social activism measure.

[v]EDEQ: Eating Disorder Examination Questionnaire (weight or shape subscales).

[w]WREQ: Weight-related eating questionnaire.

Risk of Bias

Methodological quality was assessed using the Risk of Bias Tool (RoB 2; Cochrane Collaboration) established by the Cochrane System of Interventions Assessment Handbook [42]. This tool evaluates 5 domains: the randomization process, deviations from planned interventions, missing data, outcome measurement, and selection of reported outcomes. We classified the judgment rating for each domain as "yes," "probably yes," "no," "probably no," and "unreported." Furthermore, the combination of the 5 domains provides an overall score that classifies the risk as "low," "high," or "some concerns." Two reviewers (AC-L and JB-P) conducted independent assessments of the studies included in this paper. We resolved any disagreements through discussions with a third reviewer (JCB-A).

Synthesis Methods

The information was systematically synthesized into tables containing descriptive information on the selected results. Tables were generated for the descriptive results and separated by specific design type (eg, RCT or quasi-experimental). Figures were generated based on the specific characteristics of the included studies and the methodological quality assessment by domain and overall (risk of bias). The information was grouped taking into account 4 intervention conditions that use social media as microinterventions, stimulus comparison, combined interventions, and self-guided videos. A narrative synthesis was chosen due to two critical conditions that prevented a meta-analysis: (1) the wide heterogeneity previously identified in the studies such as the different populations (eg, adolescents, young adults, or both), social media platforms (Facebook, Instagram, YouTube, or combined), affordances (ie, photos, images, videos, Facebook groups, post or chatbots), duration of interventions (eg, 5 minutes, 3 days, and up to 14 days) and follow-up (1 week to 1 month) and (2) the small number of studies included (k=3) per condition assessed (eg, body satisfaction or dissatisfaction) would not allow a subgroup analysis to understand the possible source of heterogeneity that could affect clinical decision-making as mentioned in the Cochrane guide [43].

Ethical Considerations

The research ethics committee from Universidad Cesar Vallejo, with a code of 100-CEI-EPM-UCV-2022, reviewed and approved this study. No additional permissions were required as the data were based on articles published in different databases accessible to the authors (AC-L, JB-P, and JCB-A).

Results

Study Selection

A total of 2571 references were identified across all databases reviewed in the literature search. After identifying duplicates, 1761 articles were retained for review by title and abstract. Sixty-one full-text articles were included. Of these 61 articles, 53 were excluded, the following exclusion characteristics were outside the age range (3/53, 5.7%) were not an intervention (4/53, 7.5%), did not use social media (4/53, 7.5%) did not have RCTs or quasi-experimental (40/53, 75.5%) and the outcome was not related to BI (2/53, 3.8%) leaving 8 articles for this review (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analysis) flow diagram describing the literature review process. RCT: randomized controlled trial.

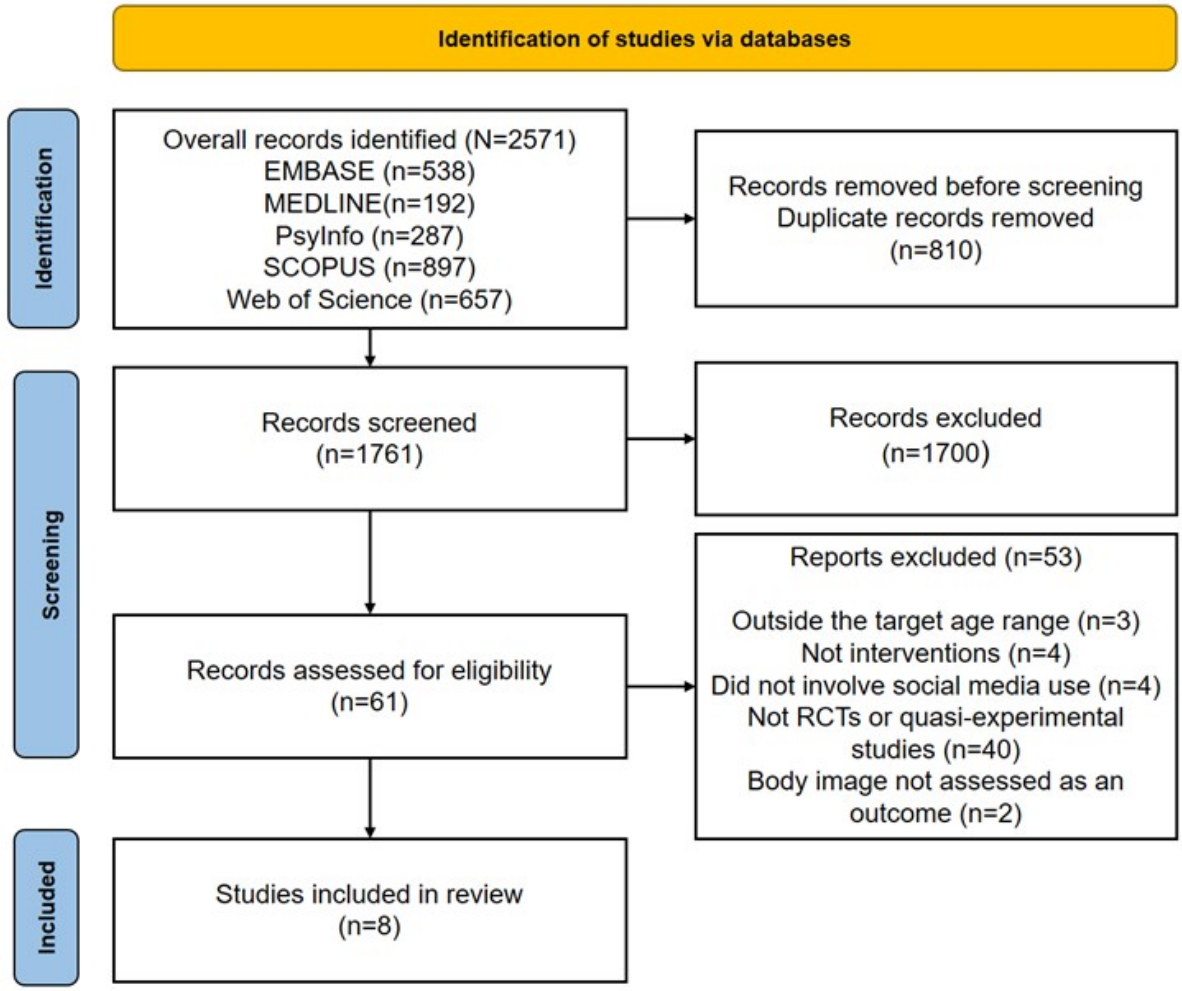

Descriptive Characteristics of the Studies

Table 1 shows an overview of the studies identified by social media year, country, and social media use. All identified studies were published between 2020 and 2023, with a majority based in the American Continent (Brazil and the United States). The majority of studies were conducted on female participants (6/8, 75%), and only 2 studies were conducted for mixed-gender groups [24,25]. The main platform for intervention (partial or total) was Facebook (4/8, 50%). The intervention duration ranged from 5 minutes to 28 days, and in most cases, the intervention was delivered by the researchers (6/8, 75%). The characteristics of the identified interventions were microinterventions to improve body satisfaction (n=3) [24,36,37], combined interventions (workshop + online group; n=1) [40], stimulus-based interventions with the presentation of socially "ideal" and "attractive" images (n=3) [25,38,39], and a self-guided model based on sequential videos to combat appearance-related pressures (n=1) [41].

Characteristics of the Interventions

The total number of study participants across all studies was 4975 (approximately 17.7% boys [24,25] and 0.3% others [13]) with a sample size range from 76 to 1847 participants per study. The age range of participants was 13 to 30 years. There were 7 RCTs and 1 quasi-experimental study. Furthermore, 50% (4/8) of the studies included follow-up, ranging from 1 week to 3 months [24,36,40,41]. Control group conditions varied, including placebo (3/8, 42.8%) [25,37,39], waitlist (2/8, 28.6%) [40,41], and assessment-only control (2/8, 28.6%) [24,36] (Table 2; Figure 2). The instruments to assess BI were very diverse, including the visual analogical scale (n=3) [37-39], positive or negative affect (n=2) [24,41], and subscales of the eating disorders inventory (n=3) [36,39,40] (Table S1 in Multimedia Appendix 4) [24,25,36-41].

XSL•FO
RenderX

Figure 2. Graphical summary of the main characteristics and outcomes identified.

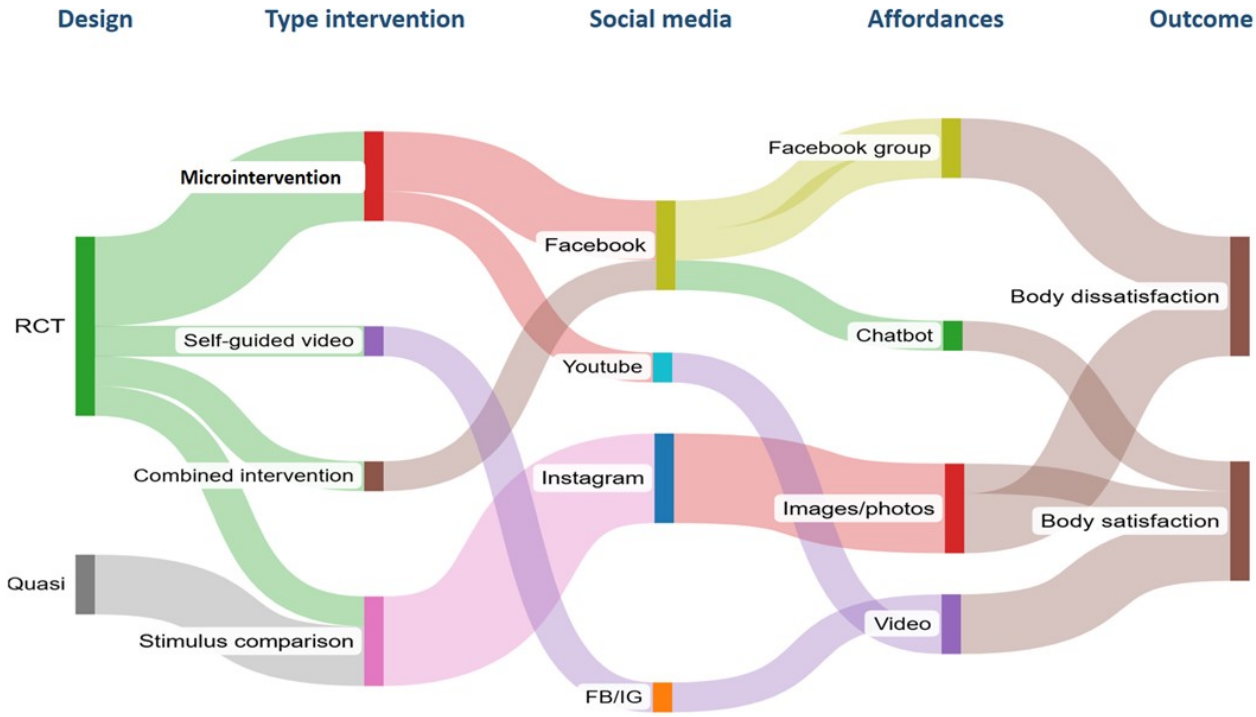

The Effectiveness of Microintervention for Improving BI

Microinterventions, due to their ease of administration (ie, single or multiple), can generate an effective and positive sequence on the immediate impact of specific symptoms (eg, body satisfaction or dissatisfaction) [44]. The effectiveness of these interventions for improving BI is centered on reducing body comparison, developing strategies to mitigate criticism of media beauty standards, and cognitive reevaluation of body-related thoughts [45,46]. These types of interventions show positive effects when using gamified and interactive self-guided models or edutainment strategies for healthy behavior change [47,48]. Three studies included microinterventions to improve body satisfaction and/or dissatisfaction [24,36,37]. They offered easily understandable content to adolescents and young adults on a single occasion or over a specific period, the latter usually daily, with different degrees of success. For example, African American girls did not show any improvement in body satisfaction after watching content on social media for comparison and pressure in a drama miniseries format (P=.11) [37]. In another study, gamified chatbot-based models did improve the body self-esteem of boys and girls, although with a small effect size (P<.02) [24]. A third study exposed adolescents and young adults to body-positive content on Facebook in different presentations (eg, access to nonidealized social beauty content presented in photos, images, external links, or access to other social media) in specific Facebook groups, resulting in reduced body dissatisfaction (P=.03) and reduced tendency to compare appearance (P<.01) [36].

The Effectiveness of Comparative Stimulus Interventions

Three studies assessed how specific social media features (eg, posting photos or images) can change some aspects of BI in adolescents and young adult women [25,38,39]. Exposure to idealized images (eg, attractive smiles of Instagram users) was found to reduce facial satisfaction (P<.01) [25]. In another study, creating profiles on social media platforms (eg, Instagram) with body-positive content increased positive mood and body satisfaction (ie, with the neutral image group) [38]. Strategies based on identifying self-affirming values (ie, selecting and reflecting on the most important value presented in the intervention) did not produce significant changes in BI and mood (P>.60) [39].

The Effectiveness of a Combined Intervention

One study combined a workshop with group interactions on Facebook [40]. The brief workshop was focused on self-compassion competencies. Then, workshop participants shared previous workshop experiences and learnings in a private Facebook group. The evaluation showed a favorable reduction in body dissatisfaction (P<.01).

The Effectiveness of Self-Guided Video Activities

One study used a self-guided educational model consisting of short, sequential videos. These videos were shared on social media (Facebook or Instagram) to explore BI risk factors among young Indonesian women [41]. These activities were effective in reducing body dissatisfaction in girls and young women (P<.01), although with a small effect size (η^2=0.05).

XSL•FO
RenderX

Efficacy of HVSM Interventions Comparison Between Adolescents and Young Adults

The studies targeting adolescents were mostly conducted in the Americas (ie, the United States or Brazil) with a variety of HVSM administration (eg, Facebook, Instagram, YouTube, or a combination of both HVSM) [24,37]. Overall, the time of the interventions in the adolescent group was shorter (ie, from 5 minutes to 3 days) compared to young adults (14 to 28 days) [24,25,36-40]. Only the adolescent group used video and chatbot interventions. Two of the 3 interventions for adolescents did improve BI, with small effects for reducing body dissatisfaction or improving body self-esteem (Cohen d=0.13) [24,41], while the interventions that achieved positive effects on BI in young adults focused on reducing body dissatisfaction.

Risk of Bias

A summary of the risk of bias assessment of the RCT studies is visible in Figure 3. Three studies did not report enough details on randomization [25,37,39,40], leading to internal validity uncertainty. One study raised concerns about bias resulting from potential overestimated intervention effects [25,39]. All studies presented low risk in terms of missing data, outcome measurement validity, and outcome selection [24,25,36,37,39,40,41].

Figure 3. Risk of bias graph for randomized controlled trials.



Discussion

Principal Findings

This is the first systematic review of interventions that use HVSM platforms to tackle unhealthy BI in adolescents and young adults. While the number of available studies remains limited, the data suggest that these interventions—despite their variability—can support modest improvements in body-related outcomes. We classified these programs into 4 types: microinterventions, interventions based on comparative stimulus, combined interventions, and self-guided video activities. All intervention types showed improvements in BI, albeit with small effect sizes. Some risk of bias was detected, particularly a lack of randomization in half of the studies, which can compromise the validity of intervention effect estimates. Overall, the studies identified in the systematic review addressed both the increase in body satisfaction and the reduction in dissatisfaction.

Microinterventions can be effective under certain conditions, such as clarity in the actions directed to the target audience (eg, formative or interventionist), a combination of specific strategies (ie, how and when to implement the actions), including interactive strategies (eg, completing activities). Chatbot-based microintervention techniques (eg, social media content management, cognitive-behavioral models, and psychoeducational techniques) showed positive effects on body satisfaction [24]. Microinterventions using gamification for BI may be advantageous due to their narrow applicability (ie, even up to a single session), short duration (eg, videos less than 5 minutes long), and greater specificity (eg, adolescents in specific age ranges) [47,49]. Brief experiences followed by quick feedback—typical in social media—can produce rewarding experiences that promote engagement and adherence [50,51]. Similarly, microinterventions using a sequenced dosage of body-positive content (eg, photo posts, image posts, and links) in social media groups (eg, Facebook groups) were also beneficial in improving body satisfaction [36]. Thus, repeated and sequenced exposure to content that promotes body acceptance or is aesthetically neutral on a daily basis (eg, ~2 weeks and at least once a day) could reduce the internalization of beauty ideals and improve BI, acceptance, and body diversity by counteracting the effect of feeds of idealized images [52].

Exposure to positive content on social media could increase cognitive flexibility regarding the internalization of socially acceptable beauty standards [53]. Even viewing neutral content

XSL•FO
RenderX

(eg, travel, landscapes, art, etc) can motivate young women to decentralize their BI (ie, interests beyond appearance) [54].

However, a microintervention based solely on a single online dramatization episode did not reduce body dissatisfaction [37]. This may be due to the indirect and unrealistic narrative (ie, presented in a single short video) of body dissatisfaction-related messages, as opposed to strategies based on successive videos (eg, miniseries or sequential videos) that can provide feedback on specific and direct topics [55]. The lack of human interaction may decrease the appeal of these interventions, especially for adolescents and young adults [12,47]. Likewise, having main characters that resemble thin-ideal standards (eg, in online series produced for an intervention) may encourage internalized appearance comparison, generating the indirect effect of body dissatisfaction [52].

Comparison stimuli based on positive images or photographs (eg, more realistic bodies) or neutral content (eg, nature, landscapes, or trips) can reduce the effects of body dissatisfaction [25,38,39]. Exposure to positive visual content can significantly reduce the evaluative determinants of body dissatisfaction (eg, increased stress and decreased subjective well-being) [56]. Neutral content (ie, related to nature) can generate subjective benefits through the sensation and appreciation of a healthy body [57,58]. However, images with thinness standards or fitspiration have been shown to increase thoughts directed toward social comparison with people subjectively perceived as superior (eg, artists, actors, influencers, or other stereotypes), increasing body dissatisfaction [59].

Exposure to sequential educational series (ie, more than one chapter) that included interactive activities (eg, completing a comic or identifying the message of a specific episode) reduced body dissatisfaction [41]. The use of entertainment series (eg, drama series designed with educational content) may have a positive effect on attitude change [60]. Thus, the familiarity that the characters can generate and the interaction with the quality of the storyline (ie, the anticipatory effect) play a crucial role in the incorporation of the desired messages about thought control and the elimination of personal comparisons that promote risk factors in body dissatisfaction [61].

The use of Facebook groups in combination with self-compassion workshops shows favorable results in reducing body dissatisfaction [40]. Combined strategies of conscious self-compassion through healthy self-talk, positive changes in self-worth, and physical warmth foster tolerance of one's own physical appearance [62,63]. Added to the online component (eg, Facebook group), they can reduce appearance comparison ideals and body concerns. A combined intervention that includes active participation in Facebook groups (eg, reading others' experiences and compassionate writing) may increase the benefits of joint feedback and how this can help recognize negative emotions related to corporality [64,65].

Despite the variability of the interventions, they share common elements such as (1) repeated dose and sequence (eg, microinterventions), (2) specific conceptual focus, that is, they must be centered on outcomes close to BI, and (3) adaptable, with plausible and culturally clear messages about body acceptance. Exposure to positive images on social media (eg, inspirational images that do not conform to hegemonic beauty standards) can foster self-improvement motives related to body satisfaction [38,66]. Furthermore, self-compassionate writing as a core component in interventions using private social media (eg, Facebook) can enhance interventions or workshops aimed at reducing body-related concerns [40,65]. Sequenced videos (eg, entertainment-education) primarily rely on psychoeducational components (eg, appearance-based comparisons) and elements of cognitive behavioral therapy, such as cognitive dissonance (ie, redirecting thoughts about physical changes), combined with interactive models that can generate motivation and adherence [24,41].

Our review identified differential effectiveness in BI interventions that align with the activities and needs of adolescents and young adults groups with a majority of women. In adolescent populations, interactive resources (eg, videos or game strategies) that incorporate cognitive-behavioral and coping components can help manage external validation, regulate feelings of BI self-worth, and address peer influence [24,41]. In contrast, young adults respond effectively to positive visual content (ie, photos, posts, or links), generating positive upward comparison (ie, motivation for self-improvement) [67]. Furthermore, the use of private groups can facilitate self-compassionate feedback to reduce preoccupation with rigid beauty ideals [65]. This suggests the need to develop differentiated strategies for both channels and content to optimize the effectiveness of interventions.

Overall, there is a methodological limitation in half of the studies, which may be considered unreliable results. Specifically, the studies by Sampson et al [25], Seekis et al [40], Pilot and Stutts [39], and Lewis-Smith et al [37] presented some risk of bias due to a lack of randomization and 2 studies also included issues regarding deviations from planned interventions [25,39]. Regardless of the intervention resource administered in HVSM, selection bias (D1) could exist, generating differences in potentially important prognostic characteristics (eg, related to baseline BI) [68,69]. This lack of comparability to overestimate the potential effects (eg, in a combined intervention workshop + Facebook group) means that the observed benefits could be attributed to the pre-existing characteristics of the participants, rather than the intervention itself [68]. Additionally, implementation bias (D2) could introduce further biases, such as potential contamination of the control group (eg, in interventions using HVSM), thus reducing the actual effects of the intervention group [69]. Despite this, another group of studies achieved a low risk of bias and identified high feasibility and acceptability with favorable results for HVSM interventions in different age groups (eg, adolescents and young adults) and intervention formats (ie, video, chat, and visual content) [24,36,41]. This suggests that platforms such as Facebook and Instagram (or others) could be useful for improving BI once the methodological limitations identified in this study are overcome, potentially providing a strong rationale for future interventions.

While other reviews have explored various digital interventions (eg, apps or websites), this research is innovative because it focuses exclusively on social media as a core component of interventions to improve BI. To our knowledge, this is the first review to synthesize evidence specifically on social media–based

XSL•FO
RenderX

interventions for BI disturbance (rather than broader digital or eHealth approaches), thereby expanding our understanding of the potential use of social media as an intervention delivery channel and active therapeutic ingredient. This review extends prior work by moving beyond platform descriptions to examine how interventions operate, identifying putative mechanisms of action used regardless of platform (eg, social comparison, cognitive dissonance, self-compassion, and behavioral modeling) and the intervention format used (eg, video-based content, stimulus comparison, and microinterventions). The contributions to the field include evidence suggesting that brief interventions can still be effective and the potential for large-scale reach and scalability. Although studies of social media–based interventions remain limited, our findings highlight clear targets for optimization that future intervention developers can incorporate. In real-world terms, social media–based tools could be embedded within platforms or delivered alongside routine social media use to help prevent or reduce symptoms of BI disturbance and counterbalance exposure to unhealthy beauty standards among groups at higher risk, including adolescents [24,70]. Realizing this potential may require platform-level design changes, clearer safeguarding standards, and supportive policy and regulatory frameworks, alongside partnerships with civil society organizations working to protect young people's mental health [15].

Strengths and Limitations

This study offers the first systematic evaluation of BI interventions implemented through HVSM platforms among youth and emerging adults. Furthermore, our study identifies the content and digital affordances that can improve body satisfaction in this vulnerable population. However, in the search strategy, other terms that encompass the full meaning of BI may not have been captured. To mitigate this, the research team comprehensively mapped search terms that could potentially capture BI-related words. The small number of studies identified precluded a meta-analysis, which would help to provide cumulative evidence for informed recommendations and clinical decision-making. Nevertheless, our study is useful in summarizing the updated information on this type of cost-effective intervention with a high adherence rate in improving mental health problems.

Conclusions

Interventions delivered through HVSM platforms to improve BI still require further development and research due to the preliminary evidence of their findings. Some formats—especially those that are interactive, brief, and visually engaging—appear to foster small but meaningful changes in how adolescents and young adults perceive their appearance. The identified methodological heterogeneity implies difficulty in obtaining a single estimate, which currently prevents a formal comparison of effects across included studies. This suggests that the interventions were designed exclusively for specific contexts and may not generate the same expected effects. However, this variety of intervention formats allows us to identify functional principles and adaptable conditions related to the dosage and mechanisms of healthy BI improvement, as well as the contextual sensitivity of success or failure in specific populations (eg, adolescents and young adults). Despite this challenge, social media remains a space with enormous reach and influence. If thoughtfully designed, interventions embedded in these platforms could become an important part of broader efforts to address BI issues and promote psychological well-being in vulnerable populations.

Acknowledgments

This review is independent research supported by the National Institute for Health and Care Research Applied Research Collaboration (ARC) North Thames. The views expressed in this publication are those of the author(s) and not necessarily those of the National Institute for Health and Care Research or the Department of Health and Social Care. The authors did not use generative artificial intelligence in the research and writing process.

Funding

This study was partially funded by Universidad Cesar Vallejo (project number 894). JCB-A was funded by the National Institute for Health Research (NIHR) 3 Research Schools' Mental Health Programme (Grant Reference Number: MHF012).

Data Availability

The protocol generated or analyzed during this study is available in Zenodo [71].

Authors' Contributions

AC-L, JB-P, EG-R, and JCB-A conceived the study idea and prepared and published the protocol. AC-L and JB-P performed the search, review, and synthesis supervised by JCB-A. NMP, LCA, and GCDC carried out both phases of the extraction process. AC-L and JB-P drafted the manuscript, receiving iterative edits and feedback from JCB-A and EG-R. All authors provided final approval for the submitted version. AC-L is the guarantor for this submission.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
PRISMA checklist.
[PDF File (Adobe PDF File), 41 KB - jmir_v28i1e80141_app1.pdf]

Multimedia Appendix 2
PRISMA-S checklist.
[PDF File (Adobe PDF File), 472 KB - jmir_v28i1e80141_app2.pdf]

Multimedia Appendix 3
Search strategy.
[DOCX File , 16 KB - jmir_v28i1e80141_app3.docx]

Multimedia Appendix 4
Measurement tools.
[DOCX File , 14 KB - jmir_v28i1e80141_app4.docx]

References

1. Grogan S. Body Image: Understanding Body Dissatisfaction in Men, Women and Children. 4th ed. London: Routledge; 2021.
2. Hong W, Ahmad NS. Body image dissatisfaction, depression and anxiety in a sample of college students. Open Psychol J 2024;17(1):e18743501303541. [doi: 10.2174/0118743501303541240521104133]
3. Mallaram GK, Sharma P, Kattula D, Singh S, Pavuluru P. Body image perception, eating disorder behavior, self-esteem and quality of life: a cross-sectional study among female medical students. J Eat Disord 2023;11(1):225 [FREE Full text] [doi: 10.1186/s40337-023-00945-2] [Medline: 38102717]
4. McLean SA, Rodgers RF, Slater A, Jarman HK, Gordon CS, Paxton SJ. Clinically significant body dissatisfaction: prevalence and association with depressive symptoms in adolescent boys and girls. Eur Child Adolesc Psychiatry 2022;31(12):1921-1932. [doi: 10.1007/s00787-021-01824-4] [Medline: 34129105]
5. Moehlecke M, Blume CA, Cureau FV, Kieling C, Schaan BD. Self-perceived body image, dissatisfaction with body weight and nutritional status of Brazilian adolescents: a nationwide study. J Pediatr (Rio J) 2020;96(1):76-83 [FREE Full text] [doi: 10.1016/j.jped.2018.07.006] [Medline: 30098939]
6. Jiménez-García AM, Arias N, Hontanaya EP, Sanz A, García-Velasco O. Impact of body-positive social media content on body image perception. J Eat Disord 2025;13(1):153 [FREE Full text] [doi: 10.1186/s40337-025-01286-y] [Medline: 40713881]
7. Holland G, Tiggemann M. A systematic review of the impact of the use of social networking sites on body image and disordered eating outcomes. Body Image 2016;17:100-110. [doi: 10.1016/j.bodyim.2016.02.008] [Medline: 26995158]
8. Karakuş Aydos Y, Dövencioğlu D, Karlı Oğuz K, Özdemir P, Pehlivantürk Kızılkan M, Kanbur N, et al. Neural correlates of distorted body images in adolescent girls with anorexia nervosa: How is it different from major depressive disorder? J Neuropsychol 2024;18(1):154-172. [doi: 10.1111/jnp.12340]
9. Stice E. Risk and maintenance factors for eating pathology: a meta-analytic review. Psychol Bull 2002;128(5):825-848. [doi: 10.1037/0033-2909.128.5.825] [Medline: 12206196]
10. Anderson M, Auxier B. Social media use in 2021. Pew Research Center. URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ [accessed 2025-02-19]
11. Social media. Statista. URL: https://www.statista.com/topics/1164/social-networks/ [accessed 2025-02-19]
12. Kruzan KP, Williams KD, Meyerhoff J, Yoo DW, O'Dwyer LC, De Choudhury M, et al. Social media-based interventions for adolescent and young adult mental health: a scoping review. Internet Interv 2022;30:100578. [doi: 10.1016/j.invent.2022.100578]
13. Sherlock M, Wagstaff DL. Exploring the relationship between frequency of Instagram use, exposure to idealized images, and psychological well-being in women. Psychol Pop Media Cult 2019;8(4):482-490. [doi: 10.1037/ppm0000182]
14. Herriman Z, Taylor AM, Roberts RM. Interventions to reduce the negative impact of online highly visual social networking site use on mental health outcomes: a scoping review. Psychol Pop Media 2024;13(1):111-139. [doi: 10.1037/ppm0000455]
15. Harriger JA, Evans JA, Thompson JK, Tylka TL. The dangers of the rabbit hole: reflections on social media as a portal into a distorted world of edited bodies and eating disorder risk and the role of algorithms. Body Image 2022;41:292-297. [doi: 10.1016/j.bodyim.2022.03.007] [Medline: 35378338]
16. Ridout B, Campbell A. The use of social networking sites in mental health interventions for young people: systematic review. J Med Internet Res 2018;20(12):e12244 [FREE Full text] [doi: 10.2196/12244] [Medline: 30563811]
17. Valentine L, McEnery C, D'Alfonso S, Phillips J, Bailey E, Alvarez-Jimenez M. Harnessing the potential of social media to develop the next generation of digital health treatments in youth mental health. Curr Treat Options Psych 2019;6(4):325-336. [doi: 10.1007/s40501-019-00184-w]

XSL•FO
RenderX

18. Ballesteros-Holmes J, Teague M, Derouin A. Decreasing social media use through motivational interviewing: a pediatric primary care quality improvement project. J Pediatr Health Care 2024;38(6):903-911. [doi: 10.1016/j.pedhc.2024.05.006] [Medline: 39365218]
19. Yu DJ, Wing YK, Li TMH, Chan NY. The impact of social media use on sleep and mental health in youth: a scoping review. Curr Psychiatry Rep 2024;26(3):104-119 [FREE Full text] [doi: 10.1007/s11920-024-01481-9] [Medline: 38329569]
20. Valentine L, McEnery C, O'Sullivan S, Gleeson J, Bendall S, Alvarez-Jimenez M. Young people's experience of a long-term social media-based intervention for first-episode psychosis: qualitative analysis. J Med Internet Res 2020;22(6):e17570 [FREE Full text] [doi: 10.2196/17570] [Medline: 32384056]
21. Marcu G, Ondersma SJ, Spiller AN, Broderick BM, Kadri R, Buis LR. The perceived benefits of digital interventions for behavioral health: qualitative interview study. J Med Internet Res 2022;24(3):e34300 [FREE Full text] [doi: 10.2196/34300] [Medline: 35353045]
22. Zhang Q, Huang Z, Sui Y, Lin F, Guan H, Li L, et al. Social-media-based mental health interventions: meta-analysis of randomized controlled trials. J Med Internet Res 2025;27:e67953 [FREE Full text] [doi: 10.2196/67953] [Medline: 40811685]
23. Marengo D, Longobardi C, Fabris M, Settanni M. Highly-visual social media and internalizing symptoms in adolescence: the mediating role of body image concerns. Comput Hum Behav 2018;82:63-69. [doi: 10.1016/j.chb.2018.01.003]
24. Matheson EL, Smith HG, Amaral ACS, Meireles JFF, Almeida MC, Linardon J, et al. Using chatbot technology to improve brazilian adolescents' body image and mental health at scale: randomized controlled trial. JMIR Mhealth Uhealth 2023;11:e39934 [FREE Full text] [doi: 10.2196/39934] [Medline: 37335604]
25. Sampson A, Jeremiah HG, Andiappan M, Newton JT. The effect of viewing idealised smile images versus nature images via social media on immediate facial satisfaction in young adults: a randomised controlled trial. J Orthod 2020;47(1):55-64. [doi: 10.1177/1465312519899664]
26. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
27. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39. [doi: 10.1186/s13643-020-01542-z]
28. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021 Mar 29;372:n160 [FREE Full text] [doi: 10.1136/bmj.n160] [Medline: 33781993]
29. Copez-Lonzoy A, Palacios J, Goicochea-Ríos E, Bazo-Alvarez J. Interventions that use highly visual social media platforms to tackle unhealthy body image in adolescents and young adults: a systematic review protocol. BMJ Open 2025;15(7):e095133. [doi: 10.1136/bmjopen-2024-095133]
30. Cavallo DN, Tate DF, Ries AV, Brown JD, DeVellis RF, Ammerman AS. A social media-based physical activity intervention: a randomized controlled trial. Am J Prev Med 2012;43(5):527-532 [FREE Full text] [doi: 10.1016/j.amepre.2012.07.019] [Medline: 23079176]
31. Mahon C, Seekis V. Systematic review of digital interventions for adolescent and young adult women's body image. Front Glob Womens Health 2022;3:832805. [doi: 10.3389/fgwh.2022.832805]
32. body image search strategy. Zenodo. URL: https://zenodo.org/records/18111594 [accessed 2026-01-20]
33. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
34. Clark LA, Watson D. Constructing validity: new developments in creating objective measuring instruments. Psychol Assess 2019;31(12):1412-1427 [FREE Full text] [doi: 10.1037/pas0000626] [Medline: 30896212]
35. Gorrall BK, Curtis J, Little TD, Panko P. Innovations in measurement: visual analog scales and retrospective pretest self-report designs. Actualidades en Psicología 2016;30(120):1-6 [FREE Full text] [doi: 10.15517/ap.v30i120.22932]
36. Fardouly J, Slater A, Parnell J, Diedrichs PC. Can following body positive or appearance neutral Facebook pages improve young women's body image and mood? Testing novel social media micro-interventions. Body Image 2023;44:136-147. [doi: 10.1016/j.bodyim.2022.12.008] [Medline: 36608433]
37. Lewis-Smith H, Pegram G, White P, Ward LM, Diedrichs PC. A short-form drama series created for the digital media environment: a randomised controlled trial exploring effects on girls' body satisfaction, acceptance of appearance diversity, and appearance-related internalised racism. Body Image 2023;47:101610. [doi: 10.1016/j.bodyim.2023.08.002] [Medline: 37659248]
38. Fioravanti G, Svicher A, Ceragioli G, Bruni V, Casale S. Examining the impact of daily exposure to body-positive and fitspiration Instagram content on young women’s mood and body image: an intensive longitudinal study. New Media Soc 2021;25(12):3266-3288. [doi: 10.1177/14614448211038904]
39. Pilot IG, Stutts LA. The impact of a values affirmation intervention on body dissatisfaction and negative mood in women exposed to fitspiration. Body Image 2023;44:36-42. [doi: 10.1016/j.bodyim.2022.11.005] [Medline: 36455512]
40. Seekis V, Bradley GL, Duffy AL. Does a Facebook-enhanced mindful self-compassion intervention improve body image? An evaluation study. Body Image 2020;34:259-269. [doi: 10.1016/j.bodyim.2020.07.006] [Medline: 32717627]

XSL•FO
RenderX

41. Garbett KM, Haywood S, Craddock N, Gentili C, Nasution K, Saraswati LA, et al. Evaluating the efficacy of a social media-based intervention (Warna-Warni Waktu) to improve body image among young indonesian women: parallel randomized controlled trial. J Med Internet Res 2023;25:e42499 [FREE Full text] [doi: 10.2196/42499] [Medline: 37010911]
42. Chandler J, Cumpston M, Li T, Page M, Welch V. Cochrane Handbook for Systematic Reviews of Interventions. In: Higgins JPT, Thomas J. editors. Newark. University of Bristol (UK) and Thomas at UCL Institute of Education (UK): John Wiley & Sons; 2019.
43. Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, editors. Cochrane Handbook for Systematic Reviews of Interventions. 2nd Edition. Chichester (UK): John Wiley & Sons; 2019.
44. Persson DR, Ramasawmy M, Khan N. Framework for microintervention software technology in digital health: critical interpretive synthesis. J Med Internet Res 2025;27(1):e72658.
45. Atkinson MJ, Diedrichs PC. Examining the efficacy of video-based microinterventions for improving risk and protective factors for disordered eating among young adult women. Int J Eat Disord 2021;54(5):708-720. [doi: 10.1002/eat.23460] [Medline: 33415790]
46. Cohen R, Irwin L, Newton-John T, Slater A. #bodypositivity: a content analysis of body positive accounts on Instagram. Body Image 2019;29:47-57. [doi: 10.1016/j.bodyim.2019.02.007] [Medline: 30831334]
47. Baumel A, Fleming T, Schueller SM. Digital micro interventions for behavioral and mental health gains: core components and conceptualization of digital micro intervention care. J Med Internet Res 2020;22(10):e20631 [FREE Full text] [doi: 10.2196/20631] [Medline: 33118946]
48. Yim SH, Schmidt U. Experiences of computer-based and conventional self-help interventions for eating disorders: a systematic review and meta-synthesis of qualitative research. Int J Eat Disord 2019;52(10):1108-1124. [doi: 10.1002/eat.23142] [Medline: 31343088]
49. Sardi L, Idri A, Fernández-Alemán JL. A systematic review of gamification in e-Health. J Biomed Inform 2017;71:31-48 [FREE Full text] [doi: 10.1016/j.jbi.2017.05.011] [Medline: 28536062]
50. Corepal R, Best P, O'Neill R, Tully MA, Edwards M, Jago R, et al. Exploring the use of a gamified intervention for encouraging physical activity in adolescents: a qualitative longitudinal study in Northern Ireland. BMJ Open 2018;8(4):e019663 [FREE Full text] [doi: 10.1136/bmjopen-2017-019663] [Medline: 29678971]
51. Kristoffersen M, Johnson C, Atkinson MJ. Feasibility and acceptability of video-based microinterventions for eating disorder prevention among adolescents in secondary schools. Int J Eat Disord 2022;55(11):1496-1505 [FREE Full text] [doi: 10.1002/eat.23781] [Medline: 35851686]
52. Rodgers RF, McLean SA, Paxton SJ. Longitudinal relationships among internalization of the media ideal, peer social comparison, and body dissatisfaction: implications for the tripartite influence model. Dev Psychol 2015;51(5):706-713. [doi: 10.1037/dev0000013] [Medline: 25751099]
53. Cowles E, Guest E, Slater A. Imagery versus captions: The effect of body positive Instagram content on young women's mood and body image. Body Image 2023;44:120-130. [doi: 10.1016/j.bodyim.2022.12.004] [Medline: 36563473]
54. Rutter H, Campoverde C, Hoang T, Goldberg SF, Berenson KR. Self-compassion and women's experience of social media content portraying body positivity and appearance ideals. Psychol Pop Media 2024;13(1):12-22. [doi: 10.1037/ppm0000453]
55. Selensky JC, Carels RA. Weight stigma and media: An examination of the effect of advertising campaigns on weight bias, internalized weight bias, self-esteem, body image, and affect. Body Image 2021;36:95-106. [doi: 10.1016/j.bodyim.2020.10.008] [Medline: 33217717]
56. Bratman GN, Olvera‐Alvarez HA, Gross JJ. The affective benefits of nature exposure. Social & Personality Psych 2021;15(8):e12630. [doi: 10.1111/spc3.12630]
57. Swami V, Barron D, Todd J, Horne G, Furnham A. Nature exposure and positive body image: (re-)examining the mediating roles of connectedness to nature and trait mindfulness. Body Image 2020;34:201-208 [FREE Full text] [doi: 10.1016/j.bodyim.2020.06.004] [Medline: 32604023]
58. Naor L, Mayseless O. The therapeutic value of experiencing spirituality in nature. Spiritual Clin Pract 2020;7(2):114-133. [doi: 10.1037/scp0000204]
59. Tiggemann M, Zaccardo M. 'Strong is the new skinny': a content analysis of #fitspiration images on Instagram. J Health Psychol 2018;23(8):1003-1011. [doi: 10.1177/1359105316639436]
60. Singhal A, Cody M, Rogers E, Sabido M. Entertainment-Education and Social Change: History, Research, and Practice. New York: Routledge; 2003.
61. Shen F, Han J(. Effectiveness of entertainment education in communicating health information: a systematic review. Asian J Commun 2014 Jul 15;24(6):605-616 [FREE Full text] [doi: 10.1080/01292986.2014.927895]
62. Nightingale BA, Cassin SE. Self-Compassion may have benefits for body image among women with a higher body mass index and internalized weight bias. Healthcare (Basel) 2023;11(7):970 [FREE Full text] [doi: 10.3390/healthcare11070970] [Medline: 37046897]
63. Cadena SBR, Iyer A, Webb TL, Millings A. Understanding the relationship between self‐compassion and body dissatisfaction. J Applied Social Pyschol 2024;54(11):669-682. [doi: 10.1111/jasp.13065]

XSL•FO
RenderX

64. Barbeau K, Guertin C, Boileau K, Pelletier L. The effects of self-compassion and self-esteem writing interventions on women's valuation of weight management goals, body appreciation, and eating behaviors. Psychol Women Q 2021;46(1):82-98. [doi: 10.1177/03616843211013465]
65. Stern NG, Engeln R. Self-Compassionate writing exercises increase college women's body satisfaction. Psychol Women Q 2018;42(3):326-341. [doi: 10.1177/0361684318773356]
66. Cohen R, Fardouly J, Newton-John T, Slater A. #BoPo on Instagram: an experimental investigation of the effects of viewing body positive content on young women's mood and body image. New Media Soc 2019;21(7):1546-1564. [doi: 10.1177/1461444819826530]
67. Festinger L. A theory of social comparison processes. Hum Relat 1954;7(2):117-140. [doi: 10.1177/001872675400700202]
68. Nguyen HQ, Cuenco D, Wolpin S, Benditt J, Carrieri-Kohlman V. Methodological considerations in evaluating eHealth interventions. Can J Nurs Res 2007;39(1):116-134. [Medline: 17450708]
69. Moore THM, Higgins JPT, Dwan K. Ten tips for successful assessment of risk of bias in randomized trials using the RoB 2 tool: early lessons from Cochrane. Cochrane Evid Synth Methods 2023;1(10):e12031. [doi: 10.1002/cesm.12031]
70. Shah J, DePietro B, D'Adamo L, Firebaugh M, Laing O, Fowler LA, et al. Development and usability testing of a chatbot to promote mental health services use among individuals with eating disorders following screening. Int J Eat Disord 2022;55(9):1229-1244 [FREE Full text] [doi: 10.1002/eat.23798] [Medline: 36056648]
71. highly visual social media platforms to tackle unhealthy body image in adolescents and young adults: a systematic review protocol. Zenodo. URL: https://doi.org/10.5281/zenodo.15660997 [accessed 2026-01-20]

Abbreviations

BI: body image
HVSM: highly visual social media
MeSH: Medical Subject Headings
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Search Extension
RCT: randomized controlled trial

Edited by S Brini; submitted 04.Jul.2025; peer-reviewed by J Willoughby, B Jiménez-Alonso; comments to author 07.Oct.2025; accepted 11.Jan.2026; published 09.Feb.2026.

Please cite as:
Copez-Lonzoy A, Bazalar-Palacios J, Mayo-Puchoc N, Calderon de la Cruz G, Cruz-Ausejo L, Goicochea-Ríos E, Bazo-Alvarez JC Interventions That Use Highly Visual Social Media Platforms to Tackle Unhealthy Body Image in Adolescents and Young Adults: Systematic Review of Randomized Controlled Trials and Quasi-Experimental Studies
J Med Internet Res 2026;28:e80141
URL: https://www.jmir.org/2026/1/e80141
doi:10.2196/80141
PMID:

©Anthony Copez-Lonzoy, Janina Bazalar-Palacios, Nikol Mayo-Puchoc, Gustavo Calderon de la Cruz, Liliana Cruz-Ausejo, Evelyn Goicochea-Ríos, Juan Carlos Bazo-Alvarez. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Text-Based Depression Estimation Using Machine Learning With Standard Labels: Systematic Review and Meta-Analysis

Shengming Zhang[1], MSc, BSc; Chaohai Zhang[1], BSc; Jiaxin Zhang[1,2], PhD

[1]School of Automation and Intelligent Manufacturing, Southern University of Science and Technology, Shenzhen, Guangdong, China

[2]Guangdong Provincial Key Laboratory of Fully Actuated System Control Theory and Technology, School of Automation and Intelligent Manufacturing, Southern University of Science and Technology, Shenzhen, Guangdong, China

Corresponding Author:
Jiaxin Zhang, PhD
School of Automation and Intelligent Manufacturing
Southern University of Science and Technology
1088 Xueyuan Avenue, Nanshan District
Shenzhen, Guangdong, 518055
China
Phone: 86 13416141690
Email: zhangjx@sustech.edu.cn

Abstract

Background: Depression affects people's daily lives and even leads to suicidal behavior. Text-based depression estimation using natural language processing has emerged as a feasible approach for early mental health screening. However, most existing reviews often included studies with weak depression labels, which affected the reliability of the results and further limited the practical application of the automatic depression estimation models.

Objective: This review aimed to evaluate the predictive performance of text-based depression models that used standard labels, and to identify text resources, text representation, model architecture, annotation source, and reporting quality contributing to performance heterogeneity.

Methods: Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines, we systematically searched 4 main databases (PubMed, Scopus, IEEE Xplore, and Web of Science) for studies published between 2014 and 2025. The eligible studies were included: machine learning models were developed based on the text generated by the participants and used validated scales or clinical diagnoses as depression labels. Pooled effect sizes (r) were calculated using random-effects meta-analysis with Hartung-Knapp-Sidik-Jonkman correction, and subgroup and meta-regression analyses were conducted to explore potential moderators.

Results: We scanned 3067 articles and finally filtered 15 models from 11 studies for the meta-analysis. The overall pooled effect size was 0.605 (95% CI 0.498-0.693), indicating a large strength of association. Subgroup analyses showed that models using embedding-based text representations achieved higher performance than those using traditional features (r=0.741, 95% CI 0.648-0.812 vs r=0.514, 95% CI 0.385-0.623; P<.001 for subgroup difference), and deep learning architectures outperformed shallow models (r=0.731, 95% CI 0.660-0.789 vs r=0.486, 95% CI 0.352-0.599; P<.001). Models trained with clinician diagnoses also outperformed better than those relying on self-report scales (r=0.688, 95% CI 0.554-0.787 vs r=0.500, 95% CI 0.340-0.631; P=.03). Reporting quality was positively associated with model performance (β=0.085, 95% CI 0.050-0.119; P<.001). Begg–Mazumdar and Egger tests provided no evidence of small-study effects. Begg–Mazumdar test (Kendall τ=0.17143, P=.37) and the Egger test (t_{14}=1.13401, 2-tailed P=.28) indicated no evidence of small-study effects.

Conclusions: Text-based depression estimation models trained with standard depression labels demonstrate solid predictive performance, with embedding features, deep model architectures, and clinician diagnosis labels showing significantly higher performance. Transparent reporting is also positively associated with model performance. This study highlights the importance of standard labels, feature representation, and reporting quality for improving model reliability. Unlike prior reviews that included weak or heterogeneous depression labels, this study offers more clinically reliable and comparable evidence. Moreover, this review provides clearer methodological guidance for developing more consistent and practically informative text-based depression screening models.

Trial Registration: PROSPERO CRD420251056902; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251056902

XSL•FO
RenderX

(J Med Internet Res 2026;28:e82686) doi:10.2196/82686

KEYWORDS

depression; natural language processing; standard labels; text; TRIPOD; Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

Introduction

Background

Depression, a type of mental disorder, is clinically characterized by persistent and significant low mood [1]. Individuals with severe depression often experience impaired social functioning and may even exhibit suicidal behaviors [2,3]. Currently, the diagnosis of depression relies on face-to-face consultation by psychiatrists and refers to the patients' self-reports. Although the clinical diagnosis by psychiatrists is regarded as the gold standard of depression detection [4], due to the time-constrained consultations and subjective information in psychiatry [5], the lack of clinical resources has prevented the gold standard from being widely used. To alleviate clinical burden, standardized self-reported scales have been developed based on psychiatrists' clinical experience [6], and several of them have been demonstrated to have comparable performance to clinical diagnosis [7,8]. However, these tools partially address the shortage of clinical resources; limitations such as subjective bias and time-consuming assessments remain [9-11]. With the development of artificial intelligence technologies, the limitations of traditional depression diagnosis methods facilitate the emergence of automatic depression estimation (ADE) models based on various multimodal data sources [12,13].

Among these multimodal data, text-based features have become a popular target in depression estimation studies [14-17]. Unlike other features used in ADE models [18-20], text (language) serves as a natural medium of human communication, conveying not only semantic content but also affective states [21,22]. The ability to capture emotional cues from text aligns closely with the depression diagnostic and may provide references for estimation [23]. For example, Cariola et al [24] found that mothers with depression used more first-person singular pronouns and present-focused words in mother-child dialogues, possibly reflecting heightened self-focus and introspection. Another study demonstrated that the ratio of negative to positive language played a vital role in establishing emotional tone [25]. Additionally, text data can be sourced from various contexts, including social media posts [26,27], SMS messages [28], chat logs [17], clinical transcripts [29], and even electronic health records [30]. Wang et al [27] achieved 91% classification accuracy based on social media posts, while Liu et al [28] reported an area under the curve of 76% on SMS messages. These studies achieved notable performance and supported the feasibility of text-based ADEs. However, differences in the reported contents (eg, text sources, model construction and validation, and depression labels) across studies may lead to substantial variability in model performance, thereby affecting the reliability of the results and the further practical application.

The quality of training data is critical to model performance, and the difference in text sources affects the performance of the ADE models [31]. The datasets of existing studies could be broadly categorized into two types: (1) public datasets, such as the Distress Analysis Interview Corpus – Wizard-of-Oz (DAIC-WOZ) [32], the Audio/Visual Emotion Challenge 2014 dataset [33], the Extended Audio-Transcript Depression Corpus [34], and the Chinese Multimodal Depression Corpus [35], which are collected under standardized protocols and provide a unified benchmark for comparing algorithm performance. For example, text-only models trained on DAIC-WOZ have shown gradual improvements in recent years [36,37]. The others belong to (2) self-constructed datasets, which are more diverse compared to public datasets in collection, and are usually used for exploring model feasibility in specific populations or contexts rather than comparing model algorithms. They include transcripts of clinical interviews or therapeutic dialogues, personal essays, diaries, questionnaires, etc. (see the previous paragraph for details). These datasets tend to better reflect natural language use in real-world contexts, which enhances the practical application ability of the ADE models [38]. For example, Cheng et al [39] analyzed social media (Weibo [Sina Corporation]) texts to predict depression and found that linguistic and cultural factors can affect the model performance. While Dogrucu et al [40] used text data via smartphones for developing Moodable, an Android-based depression sensing application developed at Worcester Polytechnic Institute, and validated this application for depression estimation. Overall, these examples illustrated the importance of self-constructed datasets, while the differences in text sources and contexts may contribute to heterogeneity in model performance.

After determining the dataset, model construction and validation are the core stages of ADE studies. Previous studies mainly relied on statistical methods to manually extract linguistic features. Rude et al [41] noted that individuals with depression frequently used negations and first-person singular pronouns, while another research examined the frequency of affective words such as "anger" or "sadness" to derive participants' emotion [42]. However, with the rapid development of natural language processing (NLP), embedding-based representations (eg, Word2Vec [43], Global Vector [44]) and pretrained models (eg, BERT [Bidirectional Encoder Representations from Transformers] [45], GPT [46]) have enhanced the semantic modeling capacity of ADE models. Niu et al [47] applied graph attention networks to capture hierarchical contextual semantics, and some studies used BERT to encode transcribed speech into context-aware sentence embeddings [48,49]. In model architecture, traditional shallow models, such as support vector machines, random forests, and decision trees, were commonly used in previous works [50], while deep learning models have also performed well in the ADE tasks recently [51-53]. Dinkel et al [51] built a multitask model based on a bidirectional gated recurrent unit, achieving an F_1-score of 0.84 on DAIC-WOZ. Martinez et al [52] further improved classification performance using RoBERTa. Furthermore, the validation strategies also vary: while single-holdout testing is commonly used when

XSL•FO
RenderX

datasets are large [54-56], small-sample medical texts often use repeated bootstrapping [57] or k-fold cross-validation [30,58] to ensure robustness. Overall, differences in feature representation, model architecture, and validation strategies all contribute to potential differences in model performance and generalization.

Another potential factor affecting the model performance is the quality of depression labels. Clinical diagnosis by certified psychiatrists is widely considered the gold standard of depression [1,59]. However, clinical diagnosis is costly and subject to the experience of the psychiatrists [60]. As a substitute, many text-based ADE studies used standardized self-report depression scales. Common scales include the Patient Health Questionnaire, 8 or 9 items (PHQ-8 or 9) [61], Beck Depression Inventory-II (BDI-II) [62], Zung Self-Rating Depression Scale (SDS) [63], 21-item Depression, Anxiety, and Stress Scale (DASS-21) [64], and Center for Epidemiological Studies Depression Scale (CES-D) [65]. The ADE tasks can be equal to predicting the self-reported scores. For example, Li et al [66] achieved an F_1-score of 78.3% on the DAIC-WOZ dataset using PHQ-8, and another study reported a classification accuracy of 69% using PHQ-9 labels [67]. However, structured scales demonstrate good reliability and validity, but their use is often limited by ethical constraints and annotation costs. Consequently, many studies resort to weak depression labels, such as keyword matching [68,69], sentiment analysis tools [70], or self-declared depressive status from social media [71]. These approaches facilitate large-scale data collection and exploration, but there may be deviations in the reliability and validity of practical applications. For example, keyword-based labels may ignore semantic context [72], and sentiment analysis tools may confuse sadness with clinical depression [73]. Therefore, increasing emphasis has been placed on adopting gold-standard labels for model training and validation [16,31,74]. In this study, we included only those studies that use either clinical diagnosis or the PHQ-9 scale as depression labels: Other scales, such as BDI-II and CES-D, though commonly used for screening, differ in target populations and sensitivity, which may impair cross-study comparability [75]. The potential impact of annotation source on model performance is examined in our Results section.

In recent years, several systematic reviews have investigated NLP-based depression estimation. Mao et al [31] provided a multimodal overview covering facial expression, speech, and text models, highlighting the lack of model transparency and limited practical application as critical challenges. Some reviews focus on specific methods. For example, Yao et al [16] summarized depression-related studies using social media text and found that machine learning (ML) and statistical analysis methods were more prevalent. Tahir et al [76] provided a comprehensive review of ML and deep learning approaches for ADE tasks based on social media data. Nanggala et al [77] conducted a systematic review of the performance of the transformer structure in ADE tasks and noted the competitive advantages of the text modalities. Many studies have examined the use of social media text for depression estimation, analyzing feature design and training strategies [16,78]. While these previous works provided valuable insights into this field, 2 major limitations remain: First, most reviews concentrate on specific text sources (eg, social media and public datasets), which may not generalize well to practical applications [79]. Second, annotation sources are often overlooked, and weak depression labels, such as self-declared or keyword matching, are often confused without discussing the impact of their differences from standard labels on model performance [72]. Therefore, this study uses label quality as a core inclusion criterion. We systematically assessed performance and sources of heterogeneity in text-based ADE models.

Research Aims and Structure

This review aims to evaluate the performance of text-based depression estimation models that use standard labels and to identify the potential moderators that may account for performance differences. We specifically analyzed variations in text sources, model architectures, validation strategies, annotation standards, as well as the quality of study reporting, to examine their potential influence on depression predictive performance. The structure of this review is as follows: the Introduction section outlines the background and research aims, the Methods section presents the study selection criteria, data extraction, and meta-analysis protocol, the Results section presents the outcomes of the meta-analysis, and the Discussion and Conclusion section interprets the findings and provides implications for future research and practical applications.

Methods

Study Design

This systematic review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) 2020 and the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) for reporting literature searches guidelines [80] (Multimedia Appendix 1). We also registered our review on PROSPERO (CRD20251056902). There were no deviations from the registered protocol.

Information Sources and Search Strategy

To include as many studies of text-based ADE as possible at the beginning, we systematically searched 4 datasets: Scopus, IEEE Xplore, Web of Science, and PubMed. The search strategy followed the PRISMA-S reporting extension [81] and was developed based on common model architectures, text features, and outcome formats used in ADE research. The search strategy included combinations of the following terms: (“depression” OR “depressive” OR “clinical depression” OR “major depressive disorder” OR “MDD”) AND (“assessment” OR “measur*” OR “diagnos*” OR “predict*” OR “estimat*”) AND (“automated” OR “automatic” OR “AI” OR “machine learning” OR “deep learning” OR “large language model” OR “natural language processing” OR “NLP”) AND (“text” OR “linguistic analys*” OR “sentiment analys*” OR “semantic analys*” OR “lexical feature*” OR “transcribed speech” OR “written” OR “textual*”). The initial search was conducted on February 24, 2025, covering studies published from January 2014 onward, and was updated on December 3, 2025. Full database-specific search strings are provided in Multimedia Appendix 2. Additional eligible studies

XSL•FO
RenderX

were identified by manually searching the reference lists of the included studies and previous reviews.

Eligibility Criteria

We included studies that fulfill the following four criteria: (1) used texts to investigate depression, (2) used ML to establish an ADE model, (3) reported the data collection process and the depression label sources, and (4) reported the metrics of the model for effect size calculation. Our exclusion criteria are the studies that (1) did not use the standard depression label (details in paragraph 5 of the introduction); (2) were not written in English; (3) published before 2024 and having fewer citations than years of publish were excluded (eg, articles published in 2021 with 3 citations, or articles published in 2019 with 5 citations would be included) [31], to ensure that full-text screening focused on studies that had been acknowledged and referenced within the field; and (4) were developed exclusively on public datasets, as these models often iterate rapidly, which could introduce potential bias.

Data Collection Process

Two independent researchers (SZ and JZ) excluded duplicates from the relevant studies retrieved, then screened and selected the relevant studies by inclusion and exclusion criteria. Disagreements were solved through discussion with a third author (CZ) when necessary. The relevant studies were evaluated through title and abstract screening. Thereafter, the remaining studies were screened through full texts to identify those for further analysis. Studies were managed using EndNote and Microsoft Excel. No automation tools were used.

Data Items

Two researchers (SZ and JZ) independently coded the included studies in Microsoft Excel. The following information was extracted from each article: (1) study characteristics (first author and publication year), (2) annotation source (clinician diagnosis and self-report scales [PHQ-9]), (3) population characteristics (sample size and positive rate), (4) modeling strategy (text sources [eg, interviews, social media, writing tasks, and diaries], text representation [eg, RoBERTa, transformer, TF-IDF, LIWC, and Empath], model architecture [deep learning vs shallow ML] and validation strategies [eg, cross-validation, hold-out, and external validation]), and (5) predictor values [sensitivity, specificity, F_1-score or data used to impute these values]. Furthermore, for studies that used different algorithms based on the same text dataset, only the model with the highest F_1-score was included in the meta-analysis. If a study reported results based on different text sources (eg, clinical transcripts vs written documents), the best-performing model from each dataset was included separately. For multimodal studies, only the performance metrics from the text modality were included.

In addition, we assessed the reporting quality of each included study using the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) tool [82]. TRIPOD was adapted to be more specific to NLP studies: The modifications were based on previously published research of NLP studies [83,84]. Two independent researchers (SZ and JZ) independently assessed each study based on TRIPOD items; disagreements were resolved through a third author (CZ).

Effect Measures

For each study, the sensitivity, specificity, and sample sizes were used to calculate the effect size (ES). Specifically, the odds ratio (OR) was first computed by sensitivity and specificity, followed by a log transformation to obtain log (OR) [85]. The SE of log (OR) was then calculated based on the sensitivity, specificity, and sample size. Log (OR) and SE (log OR) were converted into ES by using Comprehensive Meta-Analysis (version 4; Biostat Inc). To direct comparison of model performance across studies, Pearson correlation coefficient (r), transformed via Fisher z for analysis and subsequently back-transformed for interpretation, was selected as the ES indicator [86]. Furthermore, the Pearson correlation coefficient contributes to a general understanding of the relationship between text-based models' predictions and true labels. Values of r around 0.1, 0.3, and 0.5 are generally interpreted as indicating small, moderate, and large effect sizes, respectively. These thresholds have been widely applied in psychological and behavioral research, including prior reviews of depression prediction models [78,87].

Data Synthesis and Analysis

Meta-analysis was initially performed using Comprehensive Meta-Analysis (version 4; Biostat Inc). Between-study heterogeneity was assessed using the chi-square Q statistic, which tests the null hypothesis of homogeneity [88]. The extent of heterogeneity was further described using the inconsistency index (I^2), along with the between-study variance (τ^2) and its SD (τ) [89]. In addition, 95% prediction intervals were reported to reflect the expected range of effects in future studies [90]. We adopted the random-effects model to assess the heterogeneity among studies and estimate the pooled effect size from each study. Given the limited number of included studies, the Hartung-Knapp-Sidik-Jonkman (HKSJ) method was applied to adjust the SEs [91]. Because CMA 4.0 does not implement the HKSJ procedure, we reestimated the pooled effects using the Meta package in R (version 4.3.2), and the 95% CI values reported in the main text were derived from the HKSJ-adjusted models.

To explore potential sources of heterogeneity, we performed a moderator analysis with a random effects model [92]. According to the extracted information from each study, four predefined groups were included: (1) text representation (embedding-based vs traditional features), (2) annotation source (clinician diagnosis vs self-reported scale), (3) model architecture (deep vs shallow), and (4) text source (documentation vs transcribed speech). For each subgroup, pooled ES (r) were estimated under a random-effects model, and between-group differences were assessed using the Q-test for heterogeneity. In addition, univariate meta-regression analyses were conducted to examine the potential moderating effects of 3 continuous covariates: reporting quality (TRIPOD score), sample size (log-transformed), and positive rate. Each covariate was analyzed separately with a random-effects model, and an analog R^2 indicated its explanatory power for study variance. Finally,

XSL•FO
RenderX

we performed a sensitivity analysis using the leave-one-out method to assess the influence of each study on the overall results. We also conducted a cumulative meta-analysis to explore whether more recent studies have contributed to increased consistency.

Small-Study Effects

We assessed potential small-study effects, including the possibility of publication bias, by visually inspecting a funnel plot of the SE Fisher *z* [93]. Statistical evaluation of funnel plot asymmetry was conducted using the Begg–Mazumdar rank correlation test and the Egger regression test, with a *P* value <.05 indicating significant small-study effects [94]. In addition, we applied the Duval and Tweedie trim-and-fill procedure to explore the potential impact of missing studies on the pooled effect size [95].

Quality and Certainty Assessment

The methodological risk of bias of each included study was independently assessed by 2 authors (SZ and CZ) across five domains relevant to text-based ML research: (1) participant selection, (2) outcome labeling, (3) text acquisition and preprocessing, (4) model development and validation, and (5) reporting completeness. Each domain was rated as "low risk," "unclear risk," or "high risk" and an overall judgment was derived accordingly. The certainty of evidence for the primary pooled effect was evaluated using the GRADE (Grading of Recommendations Assessment, Development and Evaluation) framework [96], considering risk of bias, inconsistency, indirectness, imprecision, and publication bias.

Results

Study Selection Process

The flow diagram of studies screening and selection is presented in Figure 1. The initial database search was conducted on February 24, 2025, and was updated on December 3, 2025.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram of studies included in this review.

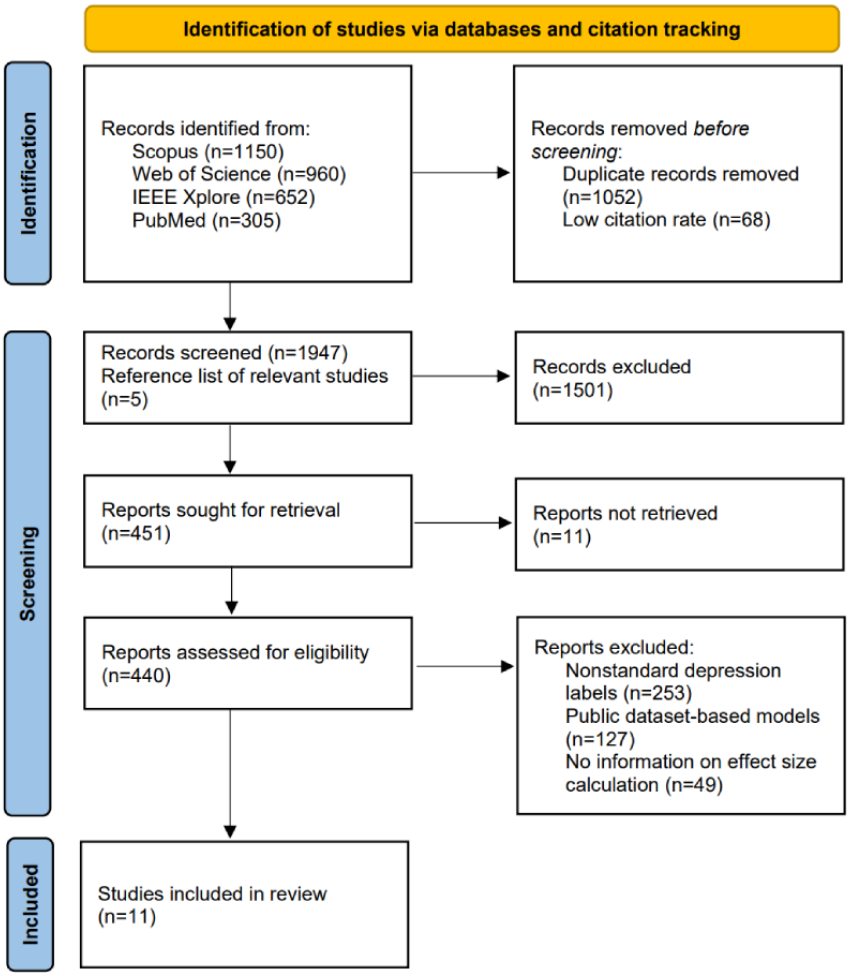

A total of 3067 records were identified through database searching, including Web of Science (n=960), PubMed (n=305), Scopus (n=1150), and IEEE Xplore (n=652). After removing 1052 duplicates and 68 low-citation records, 1947 records remained for title and abstract screening, together with 5 additional records identified through manual citation tracking. Among them, 1501 records were excluded because they were not relevant to text-based depression estimation, including that the model inputs were not text data and no quantitative results were reported. The remaining 451 records were included in the full-text screening, and 11 studies that were inaccessible or retracted were first excluded. Next, we excluded studies that

XSL•FO
RenderX

used nonstandard depression labels (n=253), studies based on public datasets (n=127), and studies lacking sufficient information to compute effect sizes (n=49). Finally, 11 studies [17,24,26,67,97-103] were included in the meta-analysis.

Characteristics of Included Studies

A total of 11 studies were included in the final meta-analysis [17,24,26,67,97-103], contributing 15 independent text-based depression estimation models. A summary of included studies and model characteristics is presented in Table 1, and the detailed data extraction table is provided in Multimedia Appendix 3. The sample size ranged from 77 to 749 participants, with a median of 110. Positive rates (proportion of participants labeled as depressed) varied substantially across studies, ranging from 9.2% to 77.9%, reflecting differences in population selection and depression labeling strategies.

Table 1. Summary of study and model characteristics included in the meta-analysis (n=15).

Author (year)	Sample size[a]	Positive rate	TRIPOD	Text representation	Annotation source	Model architecture	Text source	Sensitivity	Specificity
Geraci et al (2017) [97]	366	0.243	17	Traditional features	Clinician diagnosis	Deep	Documentation	0.94	0.68
Ricard et al (2018) [26]	749	0.0921	13	Traditional features	Self-report scale	Shallow	Documentation	0.57	0.77
Tlachac et al (2020) [98]	162	0.3395	12	Traditional features	Self-report scale	Shallow	Documentation	0.93	0.63
Zhao et al (2021) [67]	110	0.4	11	Traditional features	Self-report scale	Shallow	Documentation	0.33	0.86
Zhao et al (2021) [67]	114	0.614	11	Traditional features	Self-report scale	Shallow	Documentation	0.81	0.53
Zhao et al (2021) [67]	341	0.463	11	Traditional features	Self-report scale	Shallow	Documentation	0.51	0.86
Shin et al (2022) [99]	166	0.5	16	Traditional features	Clinician diagnosis	Shallow	Transcribed speech	0.7	0.97
Cariola et al (2022) [24]	140	0.514	9	Traditional features	Clinician diagnosis	Shallow	Transcribed speech	0.68	0.66
Munthuli et al (2023) [100]	80	0.5	15	Embedding-based	Clinician diagnosis	Deep	Transcribed speech	0.83	0.85
Munthuli et al (2023) [100]	80	0.5	15	Embedding-based	Clinician diagnosis	Deep	Transcribed speech	0.88	0.93
Munthuli et al (2023) [100]	80	0.5	15	Embedding-based	Clinician diagnosis	Deep	Transcribed speech	0.9	0.83
Tlachac et al (2023) [101]	88	0.602	13	Traditional features	Self-report scale	Shallow	Documentation	0.79	0.74
Jihoon et al (2024) [102]	77	0.779	17	Traditional features	Clinician diagnosis	Shallow	Transcribed speech	0.96	0.25
Shin et al (2024) [17]	91	0.171	14	Embedding-based	Self-report scale	Deep	Documentation	0.929	0.761
Xu et al (2025) [103]	100	0.55	19	Embedding-based	Clinical diagnosis	Deep	Transcribed speech	0.955	0.933

[a]Validation strategies include k-fold cross-validation, leave-one-out, and nested validation.

Sample size refers to the total number of participants included in each model.

Positive rate refers to the proportion of participants labeled as depressed.

TRIPOD score was based on a modified 27-item checklist adapted for NLP-based studies.

Text representations include traditional features (eg, Lexical, TF-IDF, LWIC, and Emotional Dictionary) and

XSL•FO
RenderX

embedding-based features (eg, BERT, Word2Vec, and RoBERTa).

Annotation source indicates whether the depression label was derived from clinician diagnosis or self-report scales (Patient Health Questionnaire-9, PHQ-9).

Model architecture refers to shallow learning (eg, SVM, Logistic Regression, and Random Forests) versus deep learning (eg, GRU, BERT, Transformers).

Text sources were classified as documentation (eg, Writing, Social media, and Messages) or transcribed speech (eg, Clinical Interviews and Psychological Conversations).

Regarding some characteristics of the modeling, 8 (53.3%) models used clinician diagnosis as depression labels, and the remaining 7 (46.7%) models relied on self-report scales such as the PHQ-9. In total, 10 (66.7%) models used traditional features to text representation, such as TF-IDF, LIWC, or emotional dictionaries, while 5 (33.3%) models used embedding-based techniques, including BERT and Word2Vec. 9 (60%) models used shallow learning models (eg, SVM, random forest, and logistic regression), and the remaining 6 (40%) models adopted deep learning architectures (eg, GRU and BERT-based classifiers). Text sources were classified as either documentation (n=8, 53.3%) or transcribed speech (n=7, 46.7%), the latter often derived from clinical interviews or therapeutic dialogues. Validation strategies also varied: 11 models implemented some form of k-fold cross-validation, while others applied nested or leave-group-out cross-validation.

The reporting quality (mean 13.9, SD 2.8; range 9-19) of all included models was assessed using an adapted 21-item TRIPOD checklist for NLP-based predictive modeling. Detailed scoring criteria and individual scores are provided in Multimedia Appendix 4.

Results of Meta-Analysis

A random-effects model was applied to synthesize the effect sizes of 15 text-based depression estimation models extracted from 11 eligible studies [17,24,26,67,97-103]. In studies reporting multiple models, each model was derived from a distinct text dataset with nonoverlapping participant samples and was therefore treated as a point estimate [104]. The pooled effect size (*r*) was 0.605 (95% CI 0.498-0.693) [87], indicating an overall high predictive ability in models using standard depression labels (Figure 2). Substantial between-model heterogeneity was observed (Q=99.02, P<.001; I^2=85.9%). The estimated between-study variance was τ^2=0.062 (τ=0.249). The 95% prediction interval ranged from 0.140 to 0.851, indicating considerable variability in the magnitude of effects expected across future studies. The forest plot revealed that each model achieved significant correlations (P<.05), the effect sizes varied across models (ranging from 0.292 to 0.776), further supporting the observed heterogeneity. The HKSJ-adjusted forest plot generated in R is provided in Multimedia Appendix 5. This variation highlights the influence of study-level characteristics such as model architecture, annotation source, and text representation on model performance, which were further analyzed in subsequent moderator and meta-regression analyses.

Figure 2. Forest plot presenting the pooled effect size (correlation r) of machine-learning models for text-based depression estimation trained with gold-standard labels [17,24,26,67,97-103]. For studies appearing more than once in the forest plot (eg, Zhao et al, 2021 [67]; Munthuli et al, 2023 [100]), indicate models trained and evaluated on independent text datasets derived from nonoverlapping participant samples.

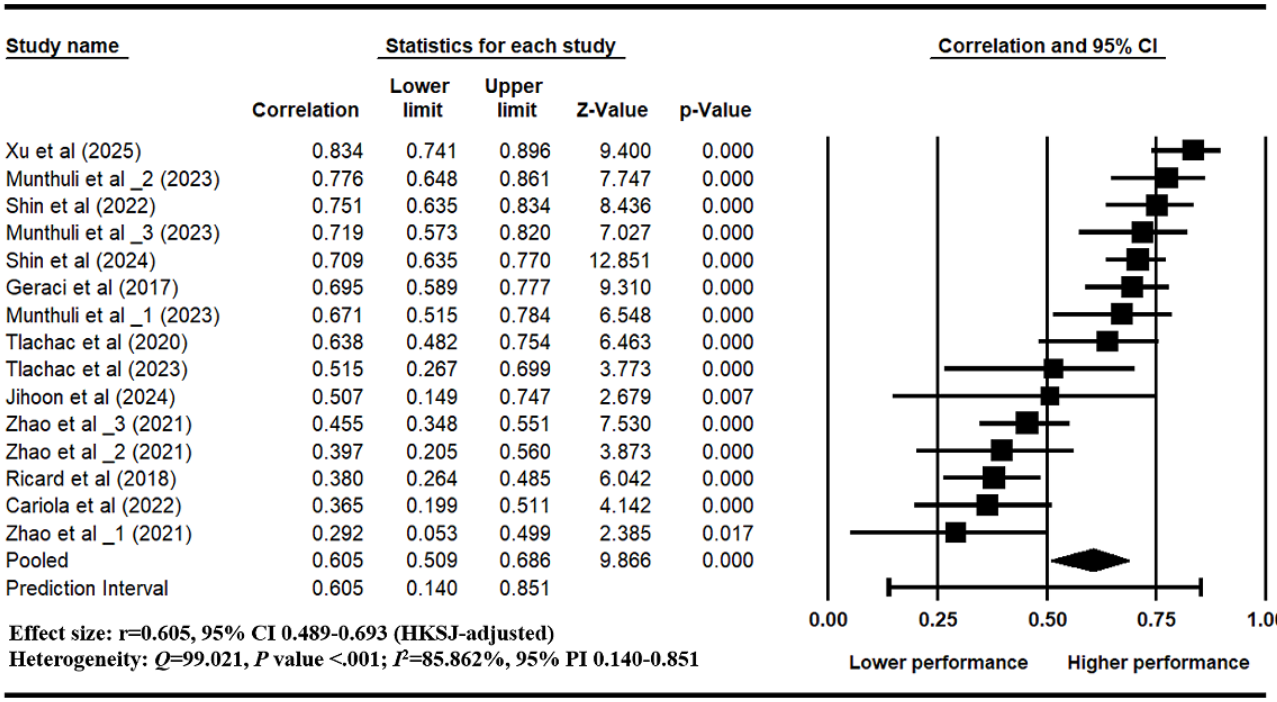

Results of Subgroup Analysis for Moderators

Subgroup analyses were conducted to examine the moderating effects of text representation, annotation source, model architecture, and text source (Table 2). Significant moderation effects were observed for text representation (Q=16.47, P<.001) and model architecture (Q=22.60, P<.001). Models using embedding-based features yielded higher performance (*r*=0.741,

XSL•FO
RenderX

95% CI 0.648-0.812) compared with those using traditional features (r=0.514, 95% CI 0.385-0.623). Deep learning architectures also outperformed shallow models (r=0.731, 95% CI 0.660-0.789 vs r=0.486, 95% CI 0.352-0.599). Annotation source showed a statistically significant moderation effect (Q=5.00, P=.03), with models using clinician diagnoses achieving higher pooled performance (r=0.688, 95% CI 0.554-0.787) than those using self-report scales (r=0.500, 95% CI 0.340-0.631). Text source did not reach statistical significance (Q=3.00, P=.08), although models using transcribed speech tended to outperform those using documentation (r=0.687 vs 0.529). Forest plots of these subgroup analyses are presented in Figure 3, and HKSJ-adjusted subgroup plots generated in R are provided in Multimedia Appendix 5.

Table 2. Subgroup analyses examining the influence of text representation, annotation source, model architecture, and text source on model performance in text-based depression estimation.

Moderators	Models, n (%)	Point estimation (95% CI)	Q-value (df)	P value
Text representation	15 (100)	—[a]	16.472 (1)	<.001
Embedding-based	5 (33.3)	0.741 (0.648-0.812)	—[a]	<.001
Traditional features	10 (66.7)	0.514 (0.385-0.623)	—*	<.001
Annotation source	15 (100)	—[a]	4.996 (1)	.03
Clinician diagnosis	8 (53.3)	0.688 (0.554-0.787)	—[a]	<.001
Self-report scale	7 (46.7)	0.500 (0.340-0.631)	—[a]	<.001
Model architecture	15 (100)	—[a]	22.595 (1)	<.001
Deep	6 (40)	0.731 (0.660-0.789)	—[a]	<.001
Shallow	9 (60)	0.486 (0.352-0.599)	—[a]	<.001
Text source	15 (100)	—[a]	3.003 (1)	.08
Documentation	8 (53.3)	0.529 (0.381-0.650)	—[a]	<.001
Transcribed speech	7 (46.7)	0.687 (0.521-0.803)	—[a]	<.001

[a]Not applicable.

XSL•FO
RenderX

Figure 3. Forest plots of subgroup analyses by (A) text representation, (B) annotation source, (C) model architecture, and (D) text source generated using CMA4.0 [17, 24, 26, 67, 97, 98, 99, 100, 101, 102, 103]. High-resolution versions of all subgroup forest plots are provided in Multimedia Appendix 5.

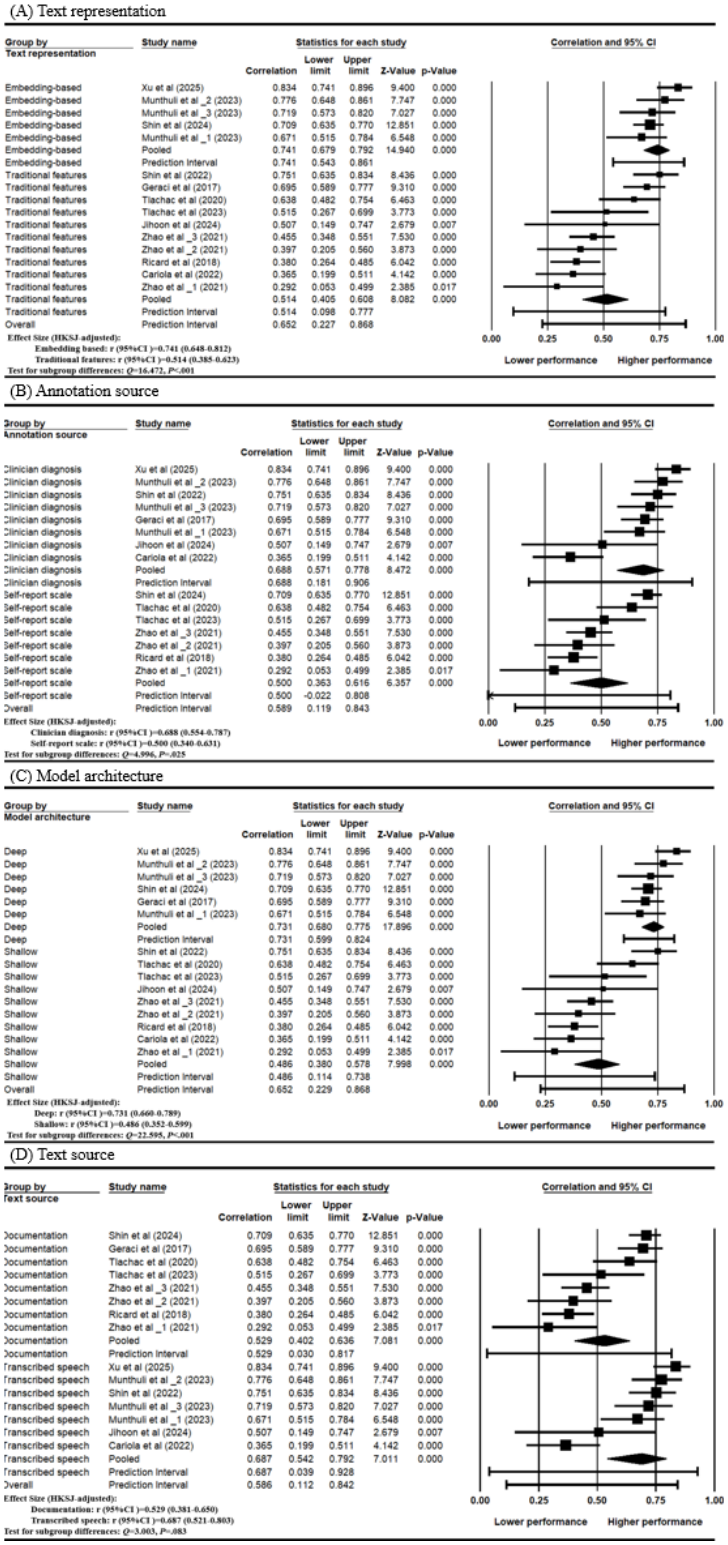

Univariable meta-regression analysis was used to test 3 continuous moderators: TRIPOD score, the positive rates, and the log-transformed sample size (log n). All meta-regressions were conducted using Fisher *z*-transformed correlation coefficients under a random-effects model. As shown in Table 3, only the TRIPOD score was significantly associated with model performance (β=0.085, 95% CI 0.050-0.119; *P*<.001), indicating a positive association between reporting quality and effect size. This model explained 71% of the between-study variance (R^2 analog=0.71). In contrast, neither the positive rate (β=–0.027, 95% CI –0.830 to 0.884; *P*=.95) nor the sample size (β=–0.012, 95% CI –0.214 to 0.190; *P*=.91) showed significant associations with effect size. Corresponding regression plots

XSL•FO
RenderX

are included in Multimedia Appendix 6 (CMA-based) and Multimedia Appendix 5 (HKSJ-adjusted).

Table 3. Univariable meta-regression assessing reporting quality (TRIPOD), positive rate, and sample size as moderators of model performance.

Moderators	N	β (95% CI)	SE	R^2	*P* value
TRIPOD[a] NLP[b]	15	0.085 (0.050-0.119)	0.018	0.71	<.001
Positive Rate	15	–0.027 (–0.830 to 0.884)	0.437	<0.2	.95
Log n	15	–0.012 (–0.214-0.190)	0.103	<0.2	.91

[a]TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

[b]NLP: natural language processing.

Risk of Bias and Certainty Assessment

Risk-of-bias assessment showed that most included models demonstrated low methodological concerns. Of the 15 models, 7 were rated as low risk, 6 as unclear risk, and 2 as high risk, with the latter typically related to concerns in outcome measurement or reporting (Multimedia Appendix 7). The overall certainty of evidence for the primary pooled effect was rated as moderate under the GRADE framework. This rating was driven mainly by the substantial heterogeneity observed across models, while concerns regarding imprecision and publication bias were minimal. Detailed domain ratings and justification for GRADE decisions are provided in Supplementary Table S4 (Multimedia Appendix 7).

Sensitivity Analysis and Cumulative Analysis

Leave-one-out analysis indicated that the pooled effect size remained stable across all iterations (r=0.605, 95% CI 0.498-0.693), suggesting that no single study had a disproportionate impact on the overall result (Figure 4). Cumulative meta-analysis showed that the effect estimates gradually stabilized over time, indicating consistency across publication years (Figure 5).

Figure 4. Leave-one-out sensitivity analysis [17,24,26,67,97-103].

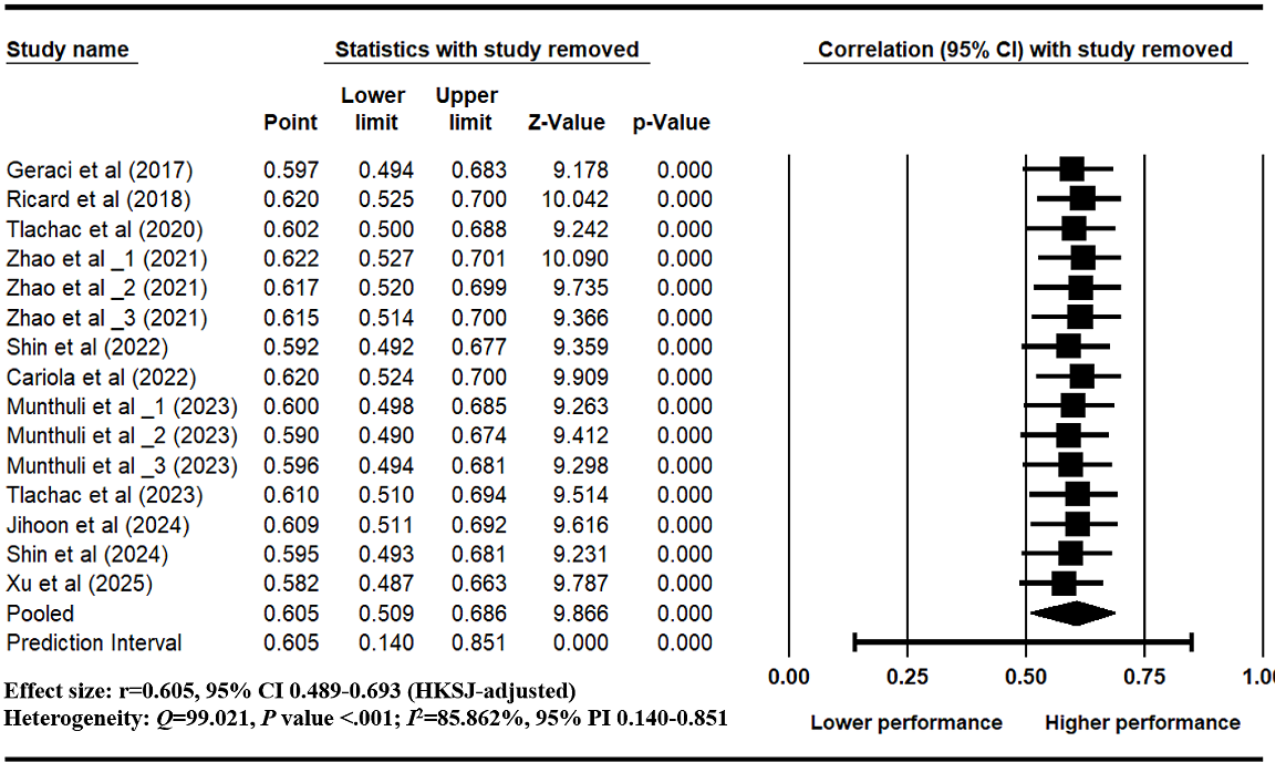

XSL•FO
RenderX

Figure 5. Cumulative meta-analysis [17,24,26,67,97-103].

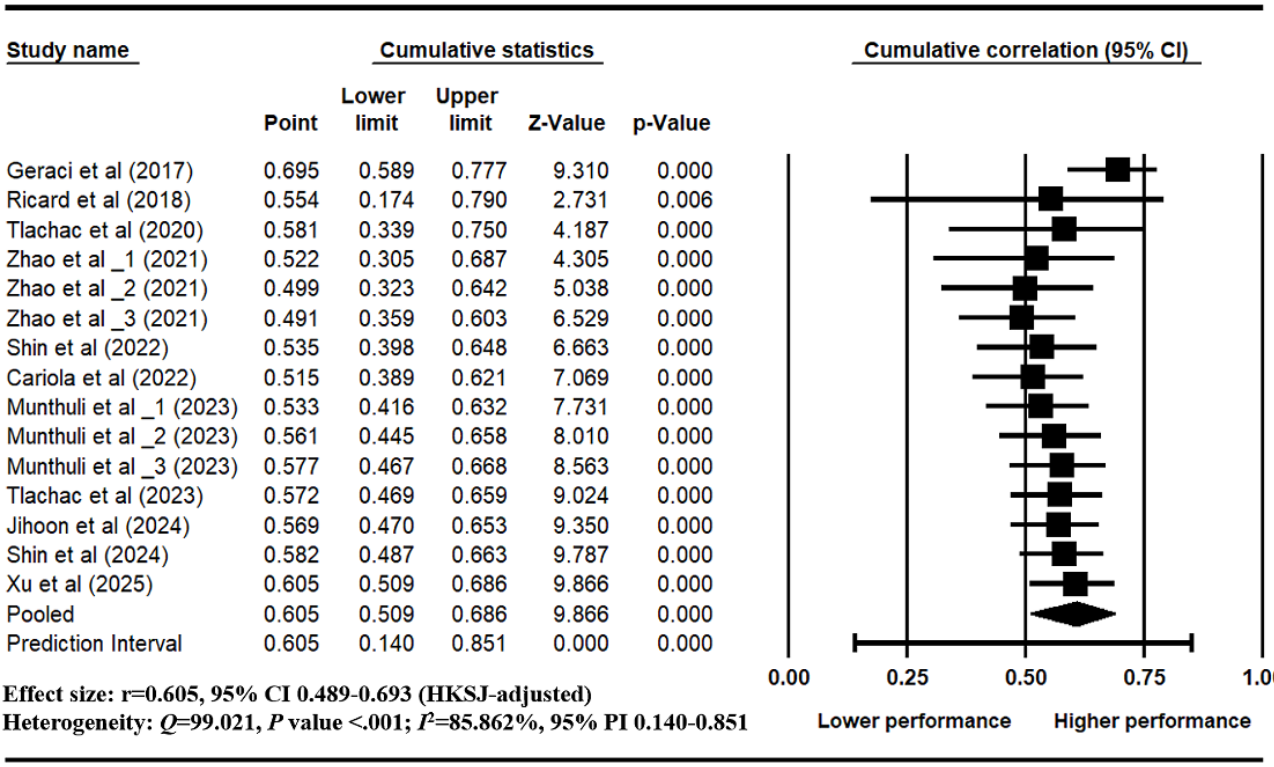

Small-Study Effects

The funnel plot (Figure 6) showed a generally symmetrical distribution of effect sizes. Both the Begg–Mazumdar test (Kendall τ=0.17143, *P*=.37) and the Egger regression test (t_{14}=1.13401, 2-tailed *P*=.28) were nonsignificant, indicating no evidence of small-study effects. The trim-and-fill procedure did not impute any additional studies, and the adjusted pooled effect was identical to the original estimate.

Figure 6. Funnel plot.

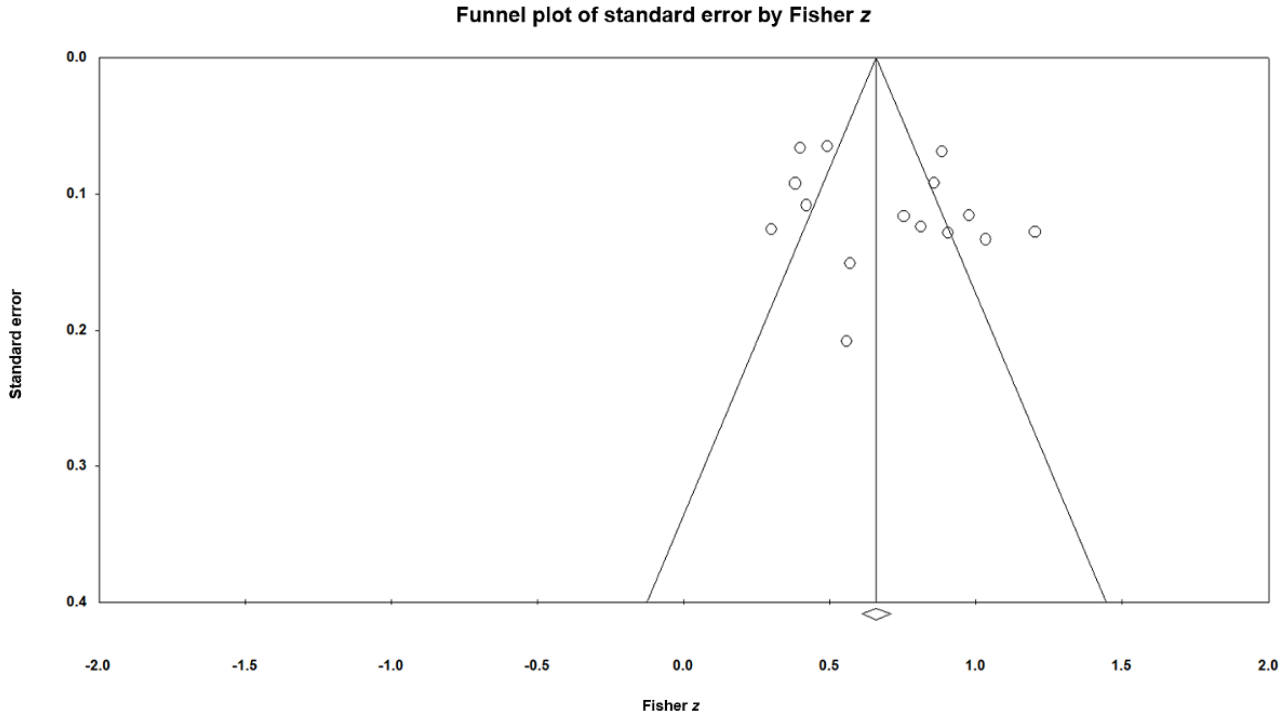

XSL•FO
RenderX

Discussion

Principal Findings

This study evaluated the performance of text-based depression estimation models that were trained and validated using standard depression labels. Our findings indicate that text-based ADE models demonstrate promising predictive performance. A previous review study mentioned the importance of a uniform standard label when evaluating the consistency of model performance across studies [78]. In this study, we focused on the studies using clinical diagnoses or PHQ-9 and excluded weak depression labels datasets (eg, keyword matching and self-declared), thereby enhancing the clinical reliability of the results [72,73]. These findings support the potential of text, especially that in transcribed speech and personal documentation, as informative signals for depression detection [23-25]. Nonetheless, the high between-study heterogeneity implies the differences in study-level characteristics, including modeling approaches, corpus sources, and reporting quality.

For the text representation and model architecture, the results of subgroup analyses showed that models using embedding-based text representations achieved higher performance than those using traditional features, and deep learning architectures outperformed shallow models. Consistent results have been reported on commonly used public benchmarks [105,106]. On the DAIC-WOZ dataset, the high-performing models often use embedding-based representation and deep model architecture [37,107]. This result aligns with the development of context-aware models like BERT [45] and GPT [46], which have superior semantic representation capabilities. A previous review study on emotion detection also reported similar conclusions [22]. However, the traditional linguistic features remain critical in the field of ADEs [16]. Linguistic features can be extracted from depressive texts by topic modeling techniques [108], which provide insights into underlying cognitive and emotional states and contribute to clinical understanding [109,110]. Overall, embedding-based and deep model architectures are superior in prediction performance, and the traditional features and shallow models remain valuable for enhancing interpretability and depression understanding.

For the annotation sources, the results of subgroup analyses indicated that models trained with clinician diagnoses also outperformed better than those relying on self-report scales. This supported the view expressed by Mao et al [31] that self-reported measures may not align with clinical diagnoses, potentially leading to inconsistency in model performance. Another study also highlighted the heterogeneity of depression estimation caused by subjectivity [111]. In addition, the results indicate that while models trained with clinically validated scales like PHQ-9 remain usable, their performance is generally lower than models trained with clinician diagnoses. For the text sources, models using transcribed speech were slightly higher than those using personal documentation, but this difference did not reach a statistically meaningful level. This may reflect the more information and affective cues embedded in spoken language [38], but some studies have pointed out that there are also noises such as stop words that are not directly related to emotion [112,113]. Overall, these results suggest that the annotation source demonstrated a determinative impact on model performance, with clinician diagnoses outperforming PHQ-9, whereas the text source showed only directional effects. We suggest constructing the ADE models based on the gold standard and natural environment-related text data. However, when clinician diagnoses are unavailable, validated self-report scales represent feasible alternatives.

It is worth noting that there is a significant positive correlation between the model report quality (TRIPOD score) and the model performance. This indicates that studies with higher report quality exhibit better performance in the ADE model. The TRIPOD checklist has been previously adopted in other automatic estimation domains [114,115]. This result further emphasized the necessity for comprehensive reporting of study-level characteristics, including modeling approaches and corpus sources, to enhance reliability and validity in ADE models [82]. In contrast, sample size and positive rates were not significantly associated with model performance. It is consistent with the previous review studies' results [76,78], suggesting that in the ADE tasks, the data quality, annotation, and methodological transparency may be more crucial than the quantity of the samples. This finding underscores the importance of rigorous methodological reporting for ADE studies.

Taken together, the findings of this meta-analysis highlight several methodological and reporting considerations that are relevant for the development and evaluation of text-based ADEs. Specifically, models using embedding-based features and deep architectures are superior in predictive performance, suggesting that future studies could prioritize richer text representations combined with more sophisticated model architectures. Adopting standardized depression labels, particularly clinician diagnoses or widely validated scales, may further facilitate comparability across studies. Notably, we found that there is a significant positive correlation between the TRIPOD score and the model performance. This indicates that improving adherence to established reporting guidelines could enhance the transparency and reproducibility of ADE models. The overall certainty of evidence was rated as moderate using the GRADE approach, reflecting acceptable confidence in the pooled findings despite substantial between-study heterogeneity.

Strengths and Limitations

The primary strength of this review lies in its strict focus on text-based ADE models trained and validated using standard depression labels. By excluding studies based on weak or inconsistent annotation sources, such as keyword matching or self-declared labels, we reduced label-related heterogeneity and improved the clinical interpretability of the synthesized findings. In addition, this meta-analysis systematically examined multiple methodological sources of heterogeneity, including text representation, model architecture, annotation source, and reporting quality, using a unified analytical framework combining subgroup analyses and meta-regression.

Nevertheless, this study also has some limitations. First, our inclusion criteria strictly require the use of standard depression labels. Although this enhanced the reliability of the labels, it

XSL•FO
RenderX

also reduced the number of eligible studies (57.5% of the literature was excluded in the full-text assessment). Some of the latest studies were also not included, including large language models [116] and the prompt learning strategy [36]. Studies on conversational agents with NLP-based automatic depression detection were also not included [117]. Second, although the Egger test indicated no significant publication bias, the exclusion of non-English and low-visibility studies raises the possibility of study omission. Future work may cautiously broaden the inclusion scope to better capture emerging methodological trends, including hybrid approaches that integrate text with other modalities, with the aim of improving the practical applicability of ADE models.

Implications

The results of this review indicate that multiple methodological factors, including text source, text representation, model architecture, annotation source, and reporting quality, influence the performance of text-based depression estimation models. These findings not only provide methodological guidance for future text-based ADE research but also offer important insights into the sources of variation observed in existing studies. To our knowledge, this study is among the first systematic reviews to explicitly restrict inclusion to models trained and validated using standard depression labels. Previous systematic reviews in this field have largely focused on aspects such as model architecture or the use of social media data, with little systematic attention given to the impact of depression label quality. The inclusion of weak depression labels in prior research may have affected the reliability of reported results and further limited their practical applicability. By treating standard depression labels as a core inclusion criterion, this review demonstrates that annotation source is one of the key determinants of model performance. This observation suggests that performance differences across studies may reflect variations in labeling standards and reporting quality, rather than differences in intrinsic model capability alone. More broadly, the present findings support a shift in text-based ADE research toward evaluation frameworks that emphasize label rigor, transparent reporting, and methodological consistency, thereby enabling more meaningful comparison across studies and ultimately facilitating the clinical translation of ADE models.

Conclusions

In summary, this systematic review and meta-analysis surveyed the last decade of text-based ADE using standard depression labels. The overall pooled effect size (r=0.605) suggests that ADE models perform well and have the potential for practical application. However, substantial heterogeneity across studies was observed. Models using embedding-based features and deep architectures generally achieved superior performance, whereas the influence of annotation source and text source was comparatively limited. Models trained with clinical diagnoses and transcribed speech tended to outperform those using self-report scales and documentation, though the difference was not statistically significant. Moreover, reporting quality, as assessed by TRIPOD, was positively associated with performance, highlighting the need for transparent reporting. Future studies should consider richer text representations, standardized labels, integration with other modalities, and transparent reporting to enhance the reliability and practical applicability of ADE models. In our future work, we will also aim to capture emerging approaches and cautiously broaden the inclusion scope, thereby providing a more comprehensive view of the ADE field.

Acknowledgments

The authors would like to thank the editor and reviewers for their constructive comments, which strengthened the quality of this manuscript. The authors declare the use of generative artificial intelligence (GenAI) to assist with language editing, proofreading, and improving clarity of expression. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing, and adapting and adjusting emotional tone. The GenAI tool used was ChatGPT (OpenAI). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Data Availability

All data and search strategies are available in the additional files accompanying this article. Further details in this study are available from the corresponding author upon reasonable request.

Funding

This study was funded by the Second Batch of the 2023 Ministry of Education of the People's Republic of China Industry–University Collaborative Education Program (Program number 230905329052215; Kingfar-CES "Human Factors and Ergonomics" Program); the Guangdong Provincial Key Laboratory of Fully Actuated System Control Theory and Technology (Grant number 2024B1212010002); the Startup Fund of Southern University of Science and Technology (grant numbers Y01966117 and Y01966223).

Authors' Contributions

SZ contributed to the conceptualization, literature screening, data extraction, meta-analysis, and writing of the original draft. CZ contributed to the refinement of the review methodology, quality assessment, preparation of supplementary materials, and revision

XSL•FO
RenderX

of the manuscript. JZ provided guidance on methodological design, supervised the review process, and contributed to conceptualization, data extraction, and manuscript revision. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA 2020 checklist.
[DOCX File , 274 KB - jmir_v28i1e82686_app1.docx]

Multimedia Appendix 2
Search strategies.
[DOCX File , 17 KB - jmir_v28i1e82686_app2.docx]

Multimedia Appendix 3
Detailed data extraction table of included studies.
[DOCX File , 27 KB - jmir_v28i1e82686_app3.docx]

Multimedia Appendix 4
TRIPOD-AI_checklist.
[DOCX File , 86 KB - jmir_v28i1e82686_app4.docx]

Multimedia Appendix 5
Hartung–Knapp–Sidik–Jonkman corrected analyses using the R meta package.
[DOCX File , 1457 KB - jmir_v28i1e82686_app5.docx]

Multimedia Appendix 6
CMA4.0-based subgroup and meta-regression analyses.
[DOCX File , 1513 KB - jmir_v28i1e82686_app6.docx]

Multimedia Appendix 7
Risk of Bias and GRADE Assessment Tables.
[DOCX File , 141 KB - jmir_v28i1e82686_app7.docx]

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. In: Diagnostic and Statistical Manual of Mental Disorders (DSM-5-TR). United States: American Psychiatric Association Publishing DSM-5, 5th ed; 2013.
2. Tolentino JC, Schmidt SL. DSM-5 criteria and depression severity: implications for clinical practice. Front Psychiatry 2018;9:450 [FREE Full text] [doi: 10.3389/fpsyt.2018.00450] [Medline: 30333763]
3. Hawton K, Casañas I C, Haw C, Saunders K. Risk factors for suicide in individuals with depression: a systematic review. J Affect Disord 2013;147(1-3):17-28 [FREE Full text] [doi: 10.1016/j.jad.2013.01.004] [Medline: 23411024]
4. Gabbard GO, Crisp-Han H. The early career psychiatrist and the psychotherapeutic identity. Academic Psychiatry 2017;41(1):30-34. [doi: 10.1007/s40596-016-0627-7] [Medline: 27882522]
5. Beck A, Alford B. Depression: Causes and Treatment, 2nd ed. US: University of Pennsylvania Press; 2009.
6. Breedvelt JJF, Zamperoni V, South E, Uphoff EP, Gilbody S, Bockting CLH, et al. A systematic review of mental health measurement scales for evaluating the effects of mental health prevention interventions. Eur J Public Health 2020;30(3):539-545. [doi: 10.1093/eurpub/ckz233] [Medline: 32236548]
7. Maust D, Cristancho M, Gray L, Rushing S, Tjoa C, Thase M. In: Aminoff MJ, Boller F, Swaab D, editors. Chapter 13 - Psychiatric rating scales. Cambridge University Press, Springer: Indian Psychiatric Society; 2012.
8. Nuevo R, Lehtinen V, Reyna-Liberato PM, Ayuso-Mateos JL. Usefulness of the beck depression inventory as a screening method for depression among the general population of finland. Scand J Public Health 2009;37(1):28-34. [doi: 10.1177/1403494808097169] [Medline: 19141552]
9. Stockings E, Degenhardt L, Lee Y, Mihalopoulos C, Liu A, Hobbs M, et al. Symptom screening scales for detecting major depressive disorder in children and adolescents: a systematic review and meta-analysis of reliability, validity and diagnostic utility. Journal of Affective Disorders 2015;174:447-463 [FREE Full text] [doi: 10.1016/j.jad.2014.11.061] [Medline: 25553406]

XSL•FO
RenderX

10. Gilbody S, Sheldon T, House A. Screening and case-finding instruments for depression: a meta-analysis. Canadian Medical Association Journal 2008;178(8):997-1003 [FREE Full text] [doi: 10.1503/cmaj.070281] [Medline: 18390942]
11. Ren Y, Yang H, Browning C, Thomas S, Liu M. Performance of screening tools in detecting major depressive disorder among patients with coronary heart disease: a systematic review. Med Sci Monit 2015;21:646-653 [FREE Full text] [doi: 10.12659/MSM.892537] [Medline: 25725615]
12. Ray A, Bhardwaj A, Malik Y, Singh S, Gupta R. Artificial intelligence and Psychiatry: an overview. Asian J Psychiatr 2022;70:103021 [FREE Full text] [doi: 10.1016/j.ajp.2022.103021] [Medline: 35219978]
13. Sun J. Artificial intelligence in psychiatry research, diagnosis, and therapy. Asian Journal of Psychiatry 2023:103705 [FREE Full text] [doi: 10.1016/j.ajp.2023.103705]
14. Wu C, Kuo C, Su C, Wang S. Using text mining to extract depressive symptoms and to validate the diagnosis of major depressive disorder from electronic health records. Journal of Affective Disorders 2020:617-623 [FREE Full text] [doi: 10.1016/j.jad.2019.09.044]
15. Yao X, Yu G, Tang J, Zhang J. Extracting depressive symptoms and their associations from an online depression community. Computers in Human Behavior 2021 Jul;120:106734 [FREE Full text] [doi: 10.1016/j.chb.2021.106734]
16. Yao S, Wang F, Chen J, Lu Q. Utilizing health-related text on social media for depression research: themes and methods. Library Hi Tech 2023;43(1):274-294. [doi: 10.1108/lht-02-2023-0076]
17. Shin D, Kim H, Lee S, Cho Y, Jung W. Using large language models to detect depression from user-generated diary text data as a novel approach in digital mental health screening: instrument validation study. J Med Internet Res 2024;26:e54617 [FREE Full text] [doi: 10.2196/54617] [Medline: 39292502]
18. Haque A, Guo M, Miner A, Fei-Fei L. Measuring depression symptom severity from spoken language and 3D facial expressions. arXiv 2018 Nov 21. [doi: 10.48550/arXiv.1811.08592]
19. Zhang B, Wei D, Yan G, Lei T, Cai H, Yang Z. Feature-level fusion based on spatial-temporal of pervasive EEG for depression recognition. Comput Methods Programs Biomed 2022;226:107113. [doi: 10.1016/j.cmpb.2022.107113] [Medline: 36103735]
20. Zhang Z, Huang P, Li S, Liu Z, Zhang J, Li Y, et al. Neural mechanisms underlying the processing of emotional stimuli in individuals with depression: an ALE meta-analysis study. Psychiatry Res 2022;313:114598. [doi: 10.1016/j.psychres.2022.114598] [Medline: 35544984]
21. Berger J, Packard G. Using natural language processing to understand people and culture. American Psychologist 2022;77(4):525-537. [doi: 10.1037/amp0000882] [Medline: 34914405]
22. Maruf AA, Khanam F, Haque MM, Jiyad ZM, Mridha MF, Aung Z. Challenges and opportunities of text-based emotion detection: a survey. IEEE Access 2024;12:18416-18450. [doi: 10.1109/access.2024.3356357]
23. Jackson J, Watts J, List J, Drabble R, Lindquist K. From text to thought: how analyzing language can advance psychological science. PsyArXiv 2020 Jul 01. [doi: 10.31234/osf.io/nws35]
24. Cariola L, Hinduja S, Bilalpur M, Sheeber L, Allen N, Morency L, et al. Language use in mother-adolescent dyadic interaction: preliminary results. Int Conf Affect Comput Intell Interact Workshops 2022;2022. [doi: 10.1109/acii55700.2022.9953886] [Medline: 39161704]
25. Olson G, Damme K, Cowan HR, Alliende LM, Mittal VA. Emotional tone in clinical high risk for psychosis: novel insights from a natural language analysis approach. Front Psychiatry 2024;15:1389597 [FREE Full text] [doi: 10.3389/fpsyt.2024.1389597] [Medline: 38803678]
26. Ricard B, Marsch L, Crosier B, Hassanpour S. Exploring the utility of community-generated social media content for detecting depression: an analytical study on instagram. JMIR Preprints 2018 Aug 06 [FREE Full text] [doi: 10.2196/preprints.11817]
27. Wang L, Zhang Y, Zhou B, Cao S, Hu K, Tan Y. Automatic depression prediction via cross-modal attention-based multi-modal fusion in social networks. Comput Electr Eng 2024;118:109413 [FREE Full text] [doi: 10.1016/j.compeleceng.2024.109413]
28. Liu T, Meyerhoff J, Eichstaedt J, Karr C, Kaiser S, Kording K, et al. The relationship between text message sentiment and self-reported depression. J Affect Disord 2022;302:7-14 [FREE Full text] [doi: 10.1016/j.jad.2021.12.048] [Medline: 34963643]
29. Tao Y, Yang M, Shen H, Yang Z, Weng Z, Hu B. Classifying anxiety and depression through LLMs virtual interactions: a case study with ChatGPT. 2023 Presented at: Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 5-8, 2023; Istanbul, Turkiye. [doi: 10.1109/bibm58861.2023.10385305]
30. Feng W, Wu H, Ma H, Yin Y, Tao Z, Lu S, et al. Deep learning based prediction of depression and anxiety in patients with type 2 diabetes mellitus using regional electronic health records. Int J Med Inform 2025;196:105801 [FREE Full text] [doi: 10.1016/j.ijmedinf.2025.105801] [Medline: 39889672]
31. Mao K, Wu Y, Chen J. A systematic review on automated clinical depression diagnosis. Npj Ment Health Res 2023;2(1):20 [FREE Full text] [doi: 10.1038/s44184-023-00040-z] [Medline: 38609509]
32. Gratch J, Arstein R, Lucas G, Stratou G, Morency L. The distress analysis interview corpus of human and computer interviews. 2014 Presented at: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); May 28-30, 2014; Paris, France.

XSL•FO
RenderX

33. Valstar M, Schuller B, Smith K, Almaev T, Eyben F, Krajewski J, et al. AVEC 2014: 3D dimensional affect and depression recognition challenge. 2014 Presented at: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge; Nov 7, 2014; Orlando Florida p. 3. [doi: 10.1145/2661806.2661807]
34. Shen Y, Yang H, Lin L. Automatic depression detection: an emotional audio-textual corpus and a Gru/Bilstm-based model. 2022 Presented at: Proceedings of the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 23-27, 2022; Singapore, Singapore. [doi: 10.1109/icassp43922.2022.9746569]
35. Zou B, Han J, Wang Y, Liu R, Zhao S, Feng L, et al. Semi-structural interview-based Chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders. IEEE Trans. Affective Comput 2023;14(4):2823-2838. [doi: 10.1109/taffc.2022.3181210]
36. Guo Y, Liu J, Wang L, Qin W, Hao S, Hong R. A prompt-based topic-modeling method for depression detection on low-resource data. IEEE Trans Comput Soc Syst 2024;11(1):1430-1439. [doi: 10.1109/tcss.2023.3260080]
37. Senn S, Tlachac M, Flores R, Rundensteiner E. Ensembles of BERT for depression classification. 2022 Presented at: Proceedings of the 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); May 11-15, 2022; New York City p. 11-15. [doi: 10.1109/embc48229.2022.9871120]
38. Sikström S, Valavičiūtė I, Kuusela I, Evors N. Question-based computational language approach outperforms rating scales in quantifying emotional states. Commun Psychol 2024;2(1):45. [doi: 10.1038/s44271-024-00097-2] [Medline: 39242812]
39. Cheng Q, Li T, Kwok C, Zhu T, Yip P. Assessing suicide risk and emotional distress in Chinese social media: a text mining and machine learning study. J Med Internet Res 2017;19(7):e243 [FREE Full text] [doi: 10.2196/jmir.7276] [Medline: 28694239]
40. Dogrucu A, Perucic A, Isaro A, Ball D, Toto E, Rundensteiner E, et al. Moodable: on feasibility of instantaneous depression assessment using machine learning on voice samples with retrospectively harvested smartphone and social media data. Smart Health 2020;17:100118 [FREE Full text] [doi: 10.1016/j.smhl.2020.100118]
41. Rude S, Gortner E, Pennebaker J. Language use of depressed and depression-vulnerable college students. Cognition & Emotion 2004;18(8):1121-1133. [doi: 10.1080/02699930441000030]
42. Gong Y, Poellabauer C. Topic modeling based multi-modal depression detection. 2017 Presented at: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge; Oct 23, 2017; Mountain View California USA. [doi: 10.1145/3133944.3133945]
43. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. Computer Science 2020. [doi: 10.3126/jiee.v3i1.34327]
44. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. 2014 Presented at: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); Oct 26-28, 2014; Doha, Qatar. [doi: 10.3115/v1/d14-1162]
45. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv 2019 May 24. [doi: 10.48550/arXiv.1810.04805]
46. Radford A, Narasimhan K. Improving Language Understanding by Generative Pre-Training. San Francisco, California: OpenAI; 2018.
47. Niu M, Chen K, Chen Q, Yang L. HCAG: a hierarchical context-aware graph attention model for depression detection. 2021 Presented at: Proceedings of the ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP);; June 06-11, 2021; Toronto, ON, Canada p. 6-11. [doi: 10.1109/icassp39728.2021.9413486]
48. Makiuchi M, Warnita T, Uto K, Shinoda K. Multimodal fusion of BERT-CNN and gated CNN representations for depression detection. : Association for Computing Machinery; 2019 Presented at: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop; Nice, France: Association for Computing Machinery; Oct 21, 2019; Nice, France p. 55-63. [doi: 10.1145/3347320.3357694]
49. Fan W, He Z, Xing X, Cai B, Lu W. Multi-modality depression detection via multi-scale temporal dilated CNNs. : Association for Computing Machinery; 2019 Presented at: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop; Oct 19, 2019; Nice, France p. 73-80. [doi: 10.1145/3347320.3357695]
50. Yang L, Jiang D, He L, Pei E, Oveneke M, Sahli H. Decision tree based depression classification from audio video and language information. : Association for Computing Machinery; 2016 Presented at: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge; Amsterdam, The Netherlands: Association for Computing Machinery; Oct 16, 2016; Amsterdam, The Netherlands. [doi: 10.1145/2988257.2988269]
51. Dinkel H, Wu M, Yu K. Text-based depression detection: what triggers an alert. arXiv 2019 Apr 08. [doi: 10.48550/arXiv.1904.05154]
52. Martinez AE, Cuadrado J, Peña D, Martinez SJ, Puertas E. Automated depression detection in text data. Leveraging lexical features, phonesthemes embedding, and RoBERTa transformer model? 2023.
53. Wu Y, Liu Z, Yuan J, Chen B, Cai H, Liu L. PIE: a personalized information embedded model for text-based depression detection. Inf Process Manag 2024:103830 [FREE Full text] [doi: 10.1016/j.ipm.2024.103830]
54. Dai H, Su C, Lee Y, Zhang Y, Wang C, Kuo C, et al. Deep learning-based natural language processing for screening psychiatric patients. Front Psychiatry 2020;11:533949 [FREE Full text] [doi: 10.3389/fpsyt.2020.533949] [Medline: 33584354]

XSL•FO
RenderX

55. Jarvers I, Ecker A, Donabauer P, Kampa K, Weißenbacher M, Schleicher D, et al. M.I.N.I.-KID interviews with adolescents: a corpus-based language analysis of adolescents with depressive disorders and the possibilities of continuation using Chat GPT. Front Psychiatry 2024;15:1425820 [FREE Full text] [doi: 10.3389/fpsyt.2024.1425820] [Medline: 39748903]
56. Zhang Z, Zhang S, Ni D, Wei Z, Yang K, Jin S, et al. Multimodal sensing for depression risk detection: integrating audio, video, and text data. Sensors (Basel) 2024;24(12):3714 [FREE Full text] [doi: 10.3390/s24123714] [Medline: 38931497]
57. He Z. Prediction mechanism of depression tendency among college students under computer intelligent systems. J Intell Syst 2024;33(1). [doi: 10.1515/jisys-2023-0209]
58. Liu Z, Wu Y, Zhang H, Li G, Ding Z, Hu B. Stimulus-response patterns: the key to giving generalizability to text-based depression detection models. IEEE J Biomed Health Inform 2024;28(8):4925-4936. [doi: 10.1109/JBHI.2024.3393244] [Medline: 38656850]
59. He L, Niu M, Tiwari P, Marttinen P, Su R, Jiang J, et al. Deep learning for depression recognition with audiovisual cues: a review. Inf Fusion 2022;80:56-86 [FREE Full text] [doi: 10.1016/j.inffus.2021.10.012]
60. Freedman R, Lewis DA, Michels R, Pine DS, Schultz SK, Tamminga CA, et al. The initial field trials of DSM-5: new blooms and old thorns. Am J Psychiatry 2013;170(1):1-5. [doi: 10.1176/appi.ajp.2012.12091189] [Medline: 23288382]
61. Kroenke K, Spitzer R. The PHQ-9: A new depression diagnostic and severity measure. San Francisco, California: US: SLACK; 2002.
62. Beck AT, Steer RA, Ball R, Ranieri WF. Comparison of beck depression inventories -IA and -II in psychiatric outpatients. J Pers Assess 1996;67(3):588-597. [doi: 10.1207/s15327752jpa6703_13] [Medline: 8991972]
63. Zung W. A self-rating depression scale. Arch Gen Psychiatry 1965;12:63-70. [doi: 10.1001/archpsyc.1965.01720310065008] [Medline: 14221692]
64. Lovibond S, Lovibond P. Manual for the Depression Anxiety Stress Scales. Australia: S. H. Lovibond, Peter F. Lovibond; 1995.
65. Radloff L. The CES-D Scale: a self-report depression scale for research in the general population. US: Sage Publications; 1977:385-401.
66. Li M, Xu H, Liu W, Liu J. Bidirectional LSTM and attention for depression detection on clinical interview transcripts. 2022 Presented at: Proceedings of the 2022 IEEE 10th International Conference on Information, Communication and Networks (ICICN); Aug 23-24, 2022; Zhangye, China. [doi: 10.1109/icicn56848.2022.10006532]
67. Zhao T, Tlachac M. Optimization with tree ensembles to improve depression screening on textual datasets. IEEE Transactions on Affective Computing 2025;16(2):573-585. [doi: 10.1109/taffc.2024.3442557]
68. Deshpande M, Rao V. Depression detection using emotion artificial intelligence. 2018 Presented at: Proceedings of the 2017 International Conference on Intelligent Sustainable Systems (ICISS); June 21, 2018; Palladam, India p. 858-862. [doi: 10.1109/iss1.2017.8389299]
69. Amanat A, Rizwan M, Javed AR, Abdelhaq M, Alsaqour R, Pandya S, et al. Deep learning for depression detection from textual data. Electronics 2022;11(5):676. [doi: 10.3390/electronics11050676]
70. Raza G, Butt Z, Latif S, Wahid A. Sentiment analysis on COVID tweets: an experimental analysis on the impact of count vectorizer and TF-IDF on sentiment predictions using deep learning models. 2021 Presented at: Proceedings of the 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2); May 20-21, 2021; Islamabad, Pakistan. [doi: 10.1109/icodt252288.2021.9441508]
71. Kour H, Gupta MK. Predicting the language of depression from multivariate twitter data using a feature‐rich hybrid deep learning model. Concurr Comput 2022 Jul 22;34(24):107363. [doi: 10.1002/cpe.7224]
72. Milintsevich K, Sirts K. Your model is not predicting depression well and that is why: a case study of PRIMATE dataset. arXiv 2024 Mar 01. [doi: 10.48550/arXiv.2403.00438]
73. Arias J, Williams C, Raghvani R, Aghajani M, Baez S, Belzung C, et al. The neuroscience of sadness: a multidisciplinary synthesis and collaborative review. Neurosci Biobehav Rev 2020;111:199-228 [FREE Full text] [doi: 10.1016/j.neubiorev.2020.01.006] [Medline: 32001274]
74. Kaushik P, Bansal K, Kumar Y, Changela A. Mental disorders prognosis and predictions using artificial intelligence techniques: a comprehensive study. SN Comput Sci 2024 Nov 14;5(8):1048. [doi: 10.1007/s42979-024-03416-w]
75. Levis B, Benedetti A, Ioannidis J, Sun Y, Negeri Z, He C, et al. Patient Health Questionnaire-9 scores do not accurately estimate depression prevalence: individual participant data meta-analysis. J Clin Epidemiol 2020:115-128 [FREE Full text] [doi: 10.1016/j.jclinepi.2020.02.002] [Medline: 32105798]
76. Tahir WB, Khalid S, Almutairi S, Abohashrh M, Memon SA, Khan J. Depression detection in social media: a comprehensive review of machine learning and deep learning techniques. IEEE Access 2025;13:12789-12818. [doi: 10.1109/access.2025.3530862]
77. Nanggala K, Elwirehardja GN, Pardamean B. Systematic literature review of transformer model implementations in detecting depression. 2023 Presented at: Proceedings of the 6th International Conference of Computer and Informatics Engineering (IC2IE) 2023; Sep 14-15, 2023; Lombok, Indonesia. [doi: 10.1109/ic2ie60547.2023.10331448]
78. Phiri D, Makowa F, Amelia VL, Phiri YVA, Dlamini LP, Chung M. Text-based depression prediction on social media using machine learning: systematic review and meta-analysis. J Med Internet Res 2025;27:e59002 [FREE Full text] [doi: 10.2196/59002] [Medline: 40215481]

XSL•FO
RenderX

79. Kelley SW, Mhaonaigh CN, Burke L, Whelan R, Gillan CM. Machine learning of language use on Twitter reveals weak and non-specific predictions. NPJ Digit Med 2022;5(1):35 [FREE Full text] [doi: 10.1038/s41746-022-00576-y] [Medline: 35338248]
80. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev 2021;10(1):89 [FREE Full text] [doi: 10.1186/s13643-021-01626-4] [Medline: 33781348]
81. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
82. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378 [FREE Full text] [doi: 10.1136/bmj.q902] [Medline: 38636956]
83. Lam B, Chrysafi P, Chiasakul T, Khosla H, Karagkouni D, McNichol M, et al. Machine learning natural language processing for identifying venous thromboembolism: systematic review and meta-analysis. Blood advances 2024;8(12):2991-3000 [FREE Full text] [doi: 10.1182/bloodadvances.2023012200] [Medline: 38522096]
84. Casey A, Davidson E, Poon M, Dong H, Duma D, Grivas A, et al. A systematic review of natural language processing applied to radiology reports. BMC Med Inform Decis Mak 2021;21(1):179 [FREE Full text] [doi: 10.1186/s12911-021-01533-7] [Medline: 34082729]
85. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. Journal of Clinical Epidemiology 2003;56(11):1129-1135. [doi: 10.1016/s0895-4356(03)00177-x]
86. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science. A practical primer for t-tests and ANOVAs 2013;4:863 [FREE Full text] [doi: 10.3389/fpsyg.2013.00863] [Medline: 24324449]
87. Cohen J. Statistical power analysis for the behavioral sciences (Rev. ed). Hillsdale, NJ. US: Lawrence Erlbaum Associates, Inc; 1977.
88. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21(11):1539-1558. [doi: 10.1002/sim.1186] [Medline: 12111919]
89. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: I is not an absolute measure of heterogeneity. Res Synth Methods 2017;8(1):5-18. [doi: 10.1002/jrsm.1230] [Medline: 28058794]
90. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023:101014 [FREE Full text] [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
91. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014;14:25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
92. Bown MJ, Sutton AJ. Quality control in systematic reviews and meta-analyses. Eur J Vasc Endovasc Surg 2010;40(5):669-677 [FREE Full text] [doi: 10.1016/j.ejvs.2010.07.011] [Medline: 20732826]
93. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 2004;56(2):455-463 [FREE Full text] [doi: 10.1111/j.0006-341x.2000.00455.x] [Medline: 10877304]
94. Sterne JA, Egger M, Smith GD. Systematic reviews in health care: investigating and dealing with publication and other biases in meta-analysis. BMJ 2001;323(7304):101-115 [FREE Full text] [doi: 10.1136/bmj.323.7304.101] [Medline: 11451790]
95. Duval S. The trim and fill method. In: Publication Bias in Meta-Analysis. Chichester, UK: John Wiley & Sons, Ltd; 2005:127-144.
96. Guyatt G, Oxman A, Akl E, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64(4):383-394 [FREE Full text] [doi: 10.1016/j.jclinepi.2010.04.026] [Medline: 21195583]
97. Geraci J, Wilansky P, de Luca V, Roy A, Kennedy JL, Strauss J. Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. Evid Based Ment Health 2017;20(3):83-87 [FREE Full text] [doi: 10.1136/eb-2017-102688] [Medline: 28739578]
98. Tlachac ML, Rundensteiner E. Screening for depression with retrospectively harvested private versus public text. IEEE J Biomed Health Inform 2020;24(11):3326-3332. [doi: 10.1109/JBHI.2020.2983035] [Medline: 32224470]
99. Shin D, Kim K, Lee S, Lee C, Bae YS, Cho WI, et al. Detection of depression and suicide risk based on text from clinical interviews using machine learning: possibility of a new objective diagnostic marker. Frontiers in Psychiatry 2022;13:801301. [doi: 10.3389/fpsyt.2022.801301] [Medline: 35686182]
100. Munthuli A, Pooprasert P, Klangpornkun N, Phienphanich P, Onsuwan C, Jaisin K, et al. Classification and analysis of text transcription from Thai depression assessment tasks among patients with depression. PLoS One 2023;18(3):e0283095 [FREE Full text] [doi: 10.1371/journal.pone.0283095] [Medline: 36996118]

XSL•FO
RenderX

101. Tlachac M, Shrestha A, Shah M, Litterer B, Rundensteiner EA. Automated construction of lexicons to improve depression screening with text messages. IEEE J Biomed Health Inform 2023;27(6):2751-2759. [doi: 10.1109/JBHI.2022.3203345] [Medline: 36044503]
102. Oh J, Lee T, Chung ES, Kim H, Cho K, Kim H, et al. Development of depression detection algorithm using text scripts of routine psychiatric interview. Front Psychiatry 2023;14:1256571 [FREE Full text] [doi: 10.3389/fpsyt.2023.1256571] [Medline: 38239906]
103. Xu C, Chen Y, Tao Y, Xie W, Liu X, Lin Y. Deep learning-based detection of depression by fusing auditory, visual and textual clues. Journal of Affective Disorders 2025:119860 [FREE Full text] [doi: 10.1016/j.jad.2025.119860]
104. Senn SJ. Overstating the evidence: double counting in meta-analysis and related problems. BMC Med Res Methodol 2009;9(1):10 [FREE Full text] [doi: 10.1186/1471-2288-9-10] [Medline: 19216779]
105. Zhang Y, He Y, Rong L, Ding Y. A hybrid model for depression detection with transformer and bi-directional long short-term memory. 2022 Presented at: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 6-8, 2022; Las Vegas, NV, USA. [doi: 10.1109/bibm55620.2022.9995184]
106. Kaynak EB, Dibeklio?lu H. Systematic Analysis of Speech Transcription Modeling for Reliable Assessment of Depression Severity. Sakarya University Journal of Computer and Information Sciences. 2024 April 2024;7(1):77-91. [doi: 10.35377/saucis...1381522]
107. Yadav U, Sharma A. A novel automated depression detection technique using text transcript. Int J Imaging Syst Tech 2022;33(1):108-122 [FREE Full text] [doi: 10.1002/ima.22793]
108. Avasthi S, Chauhan R, Acharjya D. Topic modeling techniques for text mining over a large-scale scientific and biomedical text corpus. IJACI 2022;13(1):1-18. [doi: 10.4018/ijaci.293137]
109. Du X, Sun Y. Linguistic features and psychological states: a machine-learning based approach. Front Psychol 2022;13:955850 [FREE Full text] [doi: 10.3389/fpsyg.2022.955850] [Medline: 35936260]
110. Yang K, Kim J, Chun M, Ahn MS, Chon E, Park J, et al. Factors to improve distress and fatigue in cancer survivorship; further understanding through text analysis of interviews by machine learning. BMC Cancer 2021;21(1):741 [FREE Full text] [doi: 10.1186/s12885-021-08438-8] [Medline: 34176470]
111. Ghosh CC, McVicar D, Davidson G, Shannon C, Armour C. What can we learn about the psychiatric diagnostic categories by analysing patients' lived experiences with machine-learning? BMC Psychiatry 2022;22(1):427 [FREE Full text] [doi: 10.1186/s12888-022-03984-2] [Medline: 35751077]
112. Tang SX, Kriz R, Cho S, Park SJ, Harowitz J, Gur RE, et al. Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. NPJ Schizophr 2021;7(1):25 [FREE Full text] [doi: 10.1038/s41537-021-00154-3] [Medline: 33990615]
113. Teleki M, Dong X, Caverlee J. Quantifying the impact of disfluency on spoken content summarization. 2024 Presented at: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024); 2024 May; Torino, Italia.
114. Shang J, Xue K, Yang C, Shi H, Pan L, Zeng Y. Prediction models for health care workers' exposure to type II workplace violence: a systematic review and meta-analysis. J Nurs Care Qual 2026;41(1):29-36. [doi: 10.1097/NCQ.0000000000000897] [Medline: 40658938]
115. Zhang M, Nieuwenhuys A, Zhang Y. Posture prediction models in digital human modeling for ergonomic design: a systematic review. Med Eng Phys 2025;143:104391 [FREE Full text] [doi: 10.1016/j.medengphy.2025.104391] [Medline: 40835359]
116. Gu Z, Zhu Q. MENTALER: toward professional mental health support with LLMs via multi-role collaboration. 2024 Presented at: IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 3-6, 2024; Lisbon, Portugal p. 3-6. [doi: 10.1109/bibm62325.2024.10822643]
117. Otero-González I, Pacheco-Lorenzo M, Fernández-Iglesias MJ, Anido-Rifón LE. Conversational agents for depression screening: a systematic review. Int J Med Inform 2024;181:105272 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105272] [Medline: 37979500]

Abbreviations

ADE: automatic depression estimation
BDI-II: Beck Depression Inventory-II
BERT: Bidirectional Encoder Representations from Transformers
CES-D: Center for Epidemiological Studies Depression Scale
DAIC-WOZ: Distress Analysis Interview Corpus – Wizard-of-Oz
DASS-21: 21-item Depression, Anxiety, and Stress Scale
ES: effect size
GRADE: Grading of Recommendations Assessment, Development and Evaluation
HKSJ: Hartung-Knapp-Sidik-Jonkman
ML: machine learning
NLP: natural language processing

XSL•FO
RenderX

OR: odds ratio
PHQ-8: Patient Health Questionnaire-8
PHQ-9: Patient Health Questionnaire-9
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
SDS: Self-Rating Depression Scale
TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

Edited by S Brini; submitted 20.Aug.2025; peer-reviewed by M Ibrahim, U Pasupulety; comments to author 28.Nov.2025; accepted 10.Jan.2026; published 11.Feb.2026.

Please cite as:
Zhang S, Zhang C, Zhang J
Text-Based Depression Estimation Using Machine Learning With Standard Labels: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e82686
URL: https://www.jmir.org/2026/1/e82686
doi:10.2196/82686
PMID:

©Shengming Zhang, Chaohai Zhang, Jiaxin Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

AI and Wearables for Early Detection of Cognitive Impairment and Dementia: Systematic Review

Ander Cejudo[1,2], MSc; Markel Arrojo[1], MSc; Cristina Martín[1,2,3], PhD; Aitor Almeida[2], PhD

[1]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Donostia-San Sebastián, Spain

[2]Faculty of Engineering, University of Deusto, Bilbao, Spain

[3]Biodonostia Health Research Institute, e-Health Group, San Sebastián, Spain

Corresponding Author:
Ander Cejudo, MSc
Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
Mikeletegi 57
Donostia-San Sebastián, 20009
Spain
Phone: 34 943 309 230
Email: acejudo@vicomtech.org

Abstract

Background: Traditional cognitive screening relies on episodic clinical assessments and may miss early changes preceding cognitive impairment and dementia. Wearable and mobile health technologies enable continuous monitoring of sleep, physical activity, and circadian rhythms, generating digital biomarkers that may support scalable early detection and prevention. However, current evidence remains fragmented across devices, analytic approaches, and cognitive outcomes.

Objective: This study synthesizes and critically evaluates recent evidence on wearable devices for early detection and prevention of cognitive impairment and dementia, focusing on device categories, cognitive outcomes, analytic approaches, and prevention relevance.

Methods: We searched PubMed, Scopus, ACM Digital Library, and SpringerLink for peer-reviewed studies published between January 2020 and December 1, 2025. Eligible studies included human participants with a mean age ≥50 years, continuous wearable-derived data collected for ≥24 hours, and validated cognitive outcomes; reviews, protocols, smartphone-only studies, and pharmacological interventions were excluded. Two reviewers independently screened studies, extracted data, and assessed risk of bias using the Appraisal Tool for Cross-Sectional Studies, Newcastle-Ottawa Scale, Cochrane Risk of Bias tool, and Quality Assessment of Diagnostic Accuracy Studies-2. Owing to substantial heterogeneity in devices, outcomes, and analytic methods, quantitative meta-analysis was not feasible; a structured narrative synthesis was conducted in accordance with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidance. This study was not prospectively registered.

Results: We included 49 studies, with sample sizes ranging from 14 to 91,948 participants (>200,000 total) and a median sample size of 145. Most used research-grade actigraphy (43/49, 87.8%), while fewer used commercial wearables (7/49, 14.3%). Cognitive outcomes most frequently relied on global screening instruments, including the Mini-Mental State Examination (18/49, 36.7%), followed by *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*)–based clinical diagnoses (7/49, 14.3%) and the Montreal Cognitive Assessment (7/49, 14.3%). Analytic approaches were predominantly statistical (36/49, 73.5%), with fewer studies applying machine learning (7/49, 14.3%) or deep learning methods (6/49, 12.2%). Statistical analyses linked disrupted sleep, circadian rhythm fragmentation, and irregular activity patterns to worse cognitive outcomes, with modest-to-moderate effect sizes. Machine learning and deep learning approaches reported classification performance with area under the curve values between approximately 0.70 and 0.95. Approximately one-quarter of the studies (13/49, 26.5%) addressed early detection or prevention through longitudinal risk estimation or predictive modeling. Key limitations included small sample sizes, short monitoring durations, and limited external validation.

Conclusions: Wearable-derived behavioral markers show promise for early risk stratification. This review advances the field by shifting from descriptive associations toward a digital phenotyping framework evaluating artificial intelligence–driven prediction in the preclinical window. Unlike prior reviews focused on established dementia, it differentiates direct predictive evidence from indirect correlational findings and critically assesses methodological maturity. Continuous, passive monitoring may enable scalable detection of subtle behavioral changes, supporting earlier and more personalized risk reduction strategies.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e86262) doi:10.2196/86262

KEYWORDS

actigraphy; artificial intelligence; circadian rhythms; cognitive impairment; dementia; machine learning; mHealth; mobile health; prevention; sleep; wearable devices

Introduction

Background

Cognitive decline and dementia represent a major public health concern, with significant global and regional prevalence. In 2021, approximately 57 million people were affected by dementia worldwide, with more than 60% residing in low- and middle-income countries [1]. As the global population ages, dementia is expected to impact 78 million individuals by 2030 and 139 million by 2050 [2]. This increasing prevalence is attributed to both the rising number of older adults and the fact that dementia disproportionately affects the older population. Furthermore, the severity of the problem is anticipated to rise more sharply in regions with significant demographic changes, particularly low- and middle-income countries, where health care resources are often limited [3].

Dementia severely affects the quality of life, as it is a leading cause of disability and dependency in older individuals [4]. In addition to the physical and cognitive impairments experienced by individuals, dementia imposes a heavy emotional and economic burden on families and caregivers. Caregivers, often family members, are responsible for a substantial amount of care, with an average of 5 hours of daily care and supervision per individual affected. The global cost of dementia is estimated to exceed US $1.3 trillion, with informal caregiving accounting for approximately half of this expenditure. Health care and social systems face immense strain in responding to the needs of individuals with dementia, highlighting the urgent need for preventive measures and early intervention [2].

There are several modifiable risk factors for cognitive decline and dementia, including physical inactivity, unhealthy diets, smoking, and cardiovascular conditions such as hypertension, diabetes, and obesity [5]. Early identification of cognitive decline is crucial for enabling timely interventions. Traditional screening methods, such as the Mini-Mental State Examination (MMSE) and Montreal Cognitive Assessment (MoCA) [6], have limitations due to their episodic nature and reliance on periodic visits to health care providers. These methods are often influenced by situational factors [7,8] and may not detect early signs of decline or track changes in cognitive function over time [9,10]. In addition, performance on these instruments can be affected by educational level, language, and cultural background, and they provide only a cross-sectional snapshot of cognitive status [11]. Therefore, there is a growing need for continuous and more sensitive methods of monitoring cognitive health [12]. At the same time, the use of artificial intelligence (AI)–driven wearable technologies for continuous cognitive monitoring raises ethical and data privacy considerations, as such systems may involve the collection of sensitive behavioral and health-related data, underscoring the importance of appropriate data governance and responsible implementation [13].

Wearable devices present significant potential for the continuous and passive monitoring of key physiological variables, including activity levels, sleep patterns, heart rate, and gait. These devices can detect subtle, longitudinal changes in behavior that may occur before clinical symptoms emerge [14], allowing for the early identification of cognitive decline. AI enhances this process by analyzing data from wearables to identify patterns that indicate potential risks. AI algorithms can process large volumes of data, detecting even small shifts in physiological and behavioral patterns, which can inform personalized interventions. By leveraging AI for predictive analytics, it is possible to improve early detection capabilities, facilitating timely interventions and supporting more effective management of cognitive health [15]. This combination of wearables and AI offers a promising approach [16] to enhancing dementia prevention and care.

Objective

Previous reviews have mainly examined associations between wearable-derived measures and dementia outcomes [16], often focusing on physiological or behavioral differences in individuals already diagnosed with the disease. Other reviews have emphasized specific domains, such as gait and mobility assessment using wearables [17] or the relationship between physical activity and cognitive decline [18]. In contrast, this review focuses on how wearable-derived data are analyzed, with particular attention to advanced analytic approaches, including machine learning and deep learning, used to support early detection and prevention across the cognitive aging spectrum, from subjective cognitive decline to clinical dementia. By synthesizing evidence within a digital phenotyping framework, this review highlights how longitudinal, multimodal wearable data combined with contemporary analytic methods can move beyond descriptive associations toward individualized risk characterization and prevention-oriented applications.

Thus, the main objective of this systematic review is to synthesize and critically evaluate the current evidence base and methodological maturity of wearable devices for the early detection and prevention of cognitive impairment and dementia. We structure our synthesis around 4 specific objectives:

- To evaluate the usage trends and application contexts of wearable technologies, specifically comparing the deployment of research-grade actigraphy vs consumer-grade devices.
- To synthesize the strength of statistical associations between wearable-derived digital biomarkers (eg, sleep and circadian rhythms) and standard clinical cognitive assessments.
- To critically assess the performance and robustness of analytic approaches, comparing the outcomes of conventional statistical modeling, machine learning, and deep learning techniques.

XSL•FO
RenderX

- To assess how early detection is addressed in wearable-based studies by differentiating between studies that directly implement early detection and those that frame their findings as potentially relevant to early detection.

By addressing these specific objectives, the review aims to clarify the current state of knowledge and identify priorities for advancing wearable-based approaches toward clinical and public health application.

Methods

Overview

The reporting of this systematic review was guided by the standards of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement [19] (Multimedia Appendix 1). The literature search strategy was developed and reported in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) [20] to enhance transparency and reproducibility of the search process (Multimedia Appendix 2). The aim was to evaluate the association between wearable device–based sleep and activity measures and the onset or progression of cognitive impairment, including mild cognitive impairment (MCI), Alzheimer disease, and cognitive decline. The review process was carried out in multiple stages: identification of relevant articles, screening for eligibility, data extraction, and assessment of risk of bias. Studies that meet predefined inclusion criteria were included, and their findings were synthesized to provide insights into the role of wearable devices in monitoring cognitive health. No review protocol was prepared or registered before conducting this study.

Information Sources and Search Strategy

A comprehensive search strategy was used across multiple electronic databases, including PubMed, Scopus, ACM Digital Library, and SpringerLink, to identify relevant studies published after 2020. The search was conducted using a combination of MeSH (Medical Subject Headings) terms and keywords related to cognitive impairment, sleep and activity measures, wearable devices, and their associations with cognitive outcomes. Databases were searched individually using their native interfaces. Duplicate records were removed, and studies published before 2020 were discarded to align the scope of this systematic review with contemporary research questions, analytic frameworks, and the use of wearable devices representative of current practice in digital phenotyping and early detection and prevention of cognitive decline. The search strategy was designed to capture a broad range of studies, including randomized controlled trials, cohort studies, and observational studies, to provide a comprehensive overview of the existing evidence.

The search strategy integrates key concepts relevant to studies on the early detection of cognitive impairment and dementia using wearable devices that monitor activity and sleep. Table 1 lists the search terms, organized into categories capturing cognitive impairment, sleep and activity measurements, wearable devices, and their application in early detection. Within each group, similar terms were combined using the OR operator, and the 4 groups were combined using the AND operator. Conceptually, the search followed the structure: (HC_1 OR HC_2 OR ...) AND ($Data_1$ OR $Data_2$ OR ...) AND ($Wearable_1$ OR $Wearable_2$ OR ...) AND ($EarlyDetection_1$ OR $EarlyDetection_2$ OR ...). Full search queries are included in Multimedia Appendix 3.

Alzheimer disease is specifically included under dementia, as it accounts for 60%-70% of all dementia cases [1]. Data capture must include information on sleep, activity, or physiological signals. Wearable devices may be commercial (eg, smartwatches) or noncommercial (ad hoc or research-grade devices) and are typically worn on the body for extended periods. This review focuses on studies advancing early detection and prevention, excluding research that examines features of the condition without the potential to identify cognitive decline before clinical diagnosis or progression to a more advanced stage.

To consider an article in this study, at least 1 search term of each group must be present in either the title or the abstract. This ensures that the manuscript has a clear focus on continuous data collection on sleep, activity, or physiological signals and links these measurements to cognitive impairment outcomes. The search was limited to studies published after 2020 and in English to ensure accessibility to full-text articles.

No additional search methods were used, and no study registries were searched. No targeted website searching or manual browsing of conference proceedings was conducted. Study authors and experts were not contacted for additional data. No published search filters were used. Search strategies were developed specifically for this review and were not adapted from prior reviews. No search updates were performed after the initial search, and the search strategy was not peer reviewed. Reference lists of included studies were screened to identify additional relevant records.

XSL•FO
RenderX

Table 1. Groups of search terms and representative keywords used to identify studies in the systematic review, organized by health condition, data type, wearable technology, and early detection.

Group	Search terms
Health condition	Cognitive impairment, mild cognitive impairment, dementia, cognitive decline, Alzheimer, neurocognitive disorder, and memory impairment
Data	Sleep, sleep duration, sleep quality, circadian rhythm, rest-activity rhythm, physical activity, actigraphy, accelerometry, heart rate, heart rate variability, HRV, respiratory rate, body temperature, respiration, skin temperature, step count, steps, and distance
Wearable	Wearable, wearable device, wearable devices, wearable sensor, wearable sensors, body-worn sensor, body-worn sensors, fitness tracker, smartwatch, smart band, wristband, wristbands, accelerometers, consumer-grade wearable, actigraph, actigraphy, ambulatory monitor, and digital biomarker
Early detection	Risk factor, onset, association, predictor, correlation, early detection, screening, and prevention

Eligibility Criteria

The inclusion and exclusion criteria that guided the selection of studies for this systematic review are presented in Textbox 1. The criteria are designed to ensure that the review captures only the most relevant, recent, and methodologically sound research.

For inclusion, studies must be published between 2020 and 2025 to reflect the most recent developments in the field. Only publications identified by the predefined search queries were considered. Eligible studies must present either new statistical outcomes or AI-based results that provide original contributions. Because the review focuses on prevention and early detection, manuscripts must directly address these topics. Furthermore, included studies were required to involve human participants with a mean age of 50 years or older, ensuring relevance to an aging population. Another important condition was the continuous collection of data from wearable devices for at least 24 hours, which allows for robust and objective measurements. Only peer-reviewed journal articles were considered, discarding gray literature and preprints.

Exclusion criteria further refine the selection. Publications not written in English or not available in full text, for example, when behind an inaccessible paywall, were excluded. Review articles, study protocols, and books were not considered to maintain a focus on original research. Retracted publications were excluded. In addition, studies were not eligible if they did not apply validated measures of cognitive impairment or dementia, focused on pharmacological interventions, relied exclusively on self-reported data or smartphone-based sensing, or included only healthy participants without relevant clinical characteristics. No minimum sample-size threshold was applied during study selection. As a result, included studies span a wide range of sample sizes, including small-sample investigations (eg, n<30).

Textbox 1. Inclusion and exclusion criteria applied in the systematic review.

Inclusion criteria:

- Published between 2020 and 2025
- Identified by the search queries
- Articles presenting new statistical outcomes or artificial intelligence results
- Manuscripts closely related to prevention and early detection
- Human participants must be involved in the study with average age of 50 years or older
- Continuous data captured from wearable devices for 24 hours or more

Exclusion criteria:

- Publication not in English
- Publication behind a paywall that cannot be retrieved
- Review papers, protocols, and books
- Retracted publications
- Validated measures of cognitive impairment and dementia not included
- Focus on drug or pharmacological interventions
- Only used self-reported data or smartphone-based sensing
- Only included healthy participants

XSL•FO
RenderX

Risk of Bias Assessment

The risk of bias in included studies was assessed using the Appraisal Tool for Cross-Sectional Studies, the Newcastle-Ottawa Scale for cohort studies, the Cochrane Risk of Bias tool for randomized controlled trials, and the Quality Assessment of Diagnostic Accuracy Studies-2 for diagnostic studies [21]. These instruments evaluate domains such as selection, performance, detection, attrition, and reporting bias, as well as other potential sources of bias. Each study was independently evaluated by 2 coauthors (AC and MA), with disagreements resolved through discussion or consultation with a third coauthor (CM). The overall risk of bias was classified as low, moderate, or high, and these assessments were considered when interpreting the results of the review.

Study Selection and Data Collection Process

Search results from each database were exported as BibTeX files and managed in Zotero (Corporation for Digital Scholarship), then converted to comma-separated values for automated preprocessing using a custom Python script. This process harmonized metadata across databases, removed duplicates, applied publication year filters, and verified the presence of predefined search terms. Automated preprocessing was conducted by the first author and independently verified by the second author. The following data items were extracted from each eligible study: bibliographic details (author and year), study design, participant characteristics (sample size, age, and diagnosis), wearable device specifications (type, brand, and sensor modality), monitoring duration, cognitive assessment tools, analytic methods (statistical vs AI), and primary outcomes (statistical associations or predictive performance metrics). Study selection followed a 3-stage screening process, with independent title, abstract, and full-text reviews performed by the first and second authors. Discrepancies were resolved through consensus or adjudication by a third author. Eligible studies were synthesized using a structured narrative approach aligned with predefined review objectives, grouping studies by device type, cognitive outcomes, analytic methods, and prevention-oriented contributions.

Analytic approaches were classified according to the overall modeling paradigm rather than individual algorithms. Studies were categorized as AI-based when wearable-derived data were analyzed within predictive modeling frameworks emphasizing automated pattern learning, such as machine learning or deep learning methods, including regression-based models implemented within machine learning pipelines (eg, automated feature selection or cross-validation). In contrast, studies relying on predefined statistical models for inferential or association analyses were classified as statistical model–based approaches. In this review, “direct” evidence was defined for studies demonstrating individual-level early detection capability, including both cross-sectional screening or discriminative analyses and longitudinal or interventional designs enabling risk estimation or prevention. Cross-sectional studies were therefore considered direct when they demonstrated the ability to distinguish preclinical or prodromal disease states at the time of assessment.

Synthesis Methods

Quantitative synthesis was planned only for outcomes estimating a common construct using comparable analytical objectives and metrics. When outcomes cannot be mapped to a shared estimand for early detection, quantitative pooling was not conducted. In such cases, results were synthesized using a structured narrative approach aligned with the predefined review objectives. This approach grouped studies according to analytic strategy, cognitive assessment, and wearable measurement characteristics, without applying formal qualitative synthesis methodologies.

Results

Study Selection

The database search identified 7175 records (SpringerLink, n=3247; Scopus, n=937; ACM Digital Library, n=2471; and PubMed, n=520). After removal of 705 duplicates, 458 records with missing information, and 422 reviews, a total of 5590 records remained for screening. Of these, 1493 were excluded based on publication year (<2020, n=1493), title (n=3527), or abstract (n=452). The remaining 118 reports were sought for full-text retrieval, but 15 could not be obtained. A total of 103 full-text articles were assessed for eligibility, and 54 were excluded for not meeting the inclusion criteria. Finally, a total of 49 studies were included in this systematic review (Figure 1). Figure 2 shows the number of included studies by publication year. At the full-text screening stage, most studies were excluded because they did not include continuous wearable-based data collected over at least 24 hours, did not use wearable-derived data as part of the analysis, or did not report results relevant to the early detection of cognitive impairment or dementia. Table 2 shows the relevant information extracted from the included studies, and additional details are provided in Multimedia Appendix 4.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the study selection process.

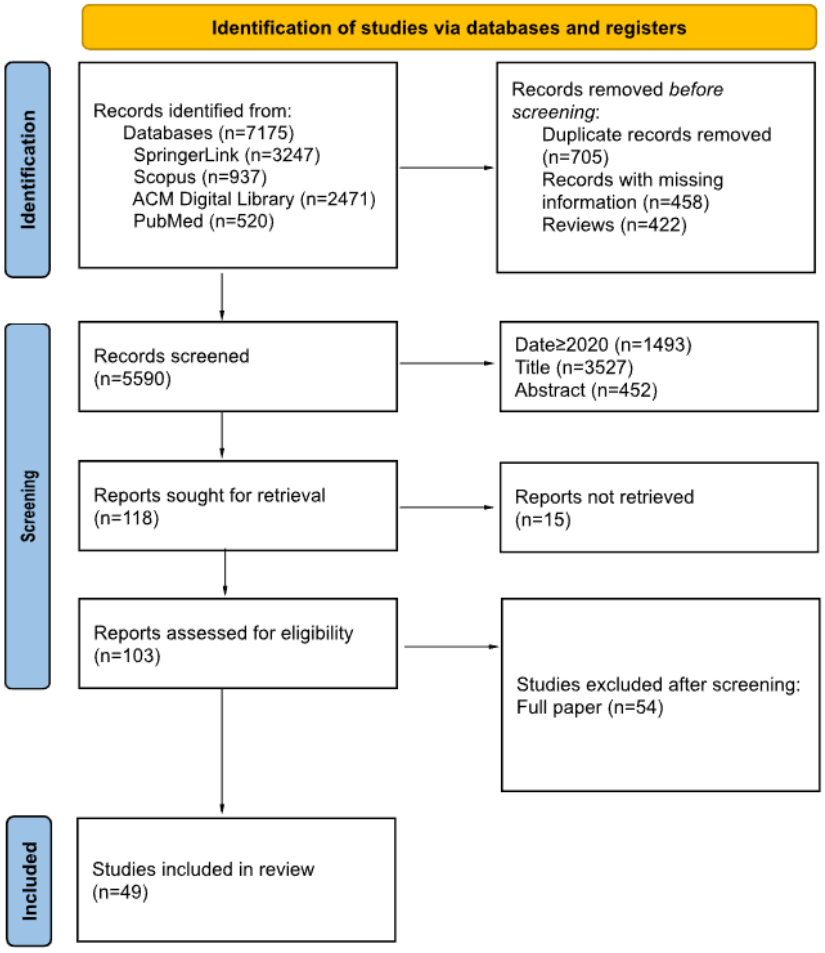

Figure 2. Temporal distribution of included studies by publication year (2020-2025). The bar chart and overlaid trend line illustrate the progression of research interest in the field. Data show a substantial increase in publication volume beginning in 2021, with research activity remaining consistently elevated through the 2024-2025 period. Data for 2025 are incomplete by 1 month.

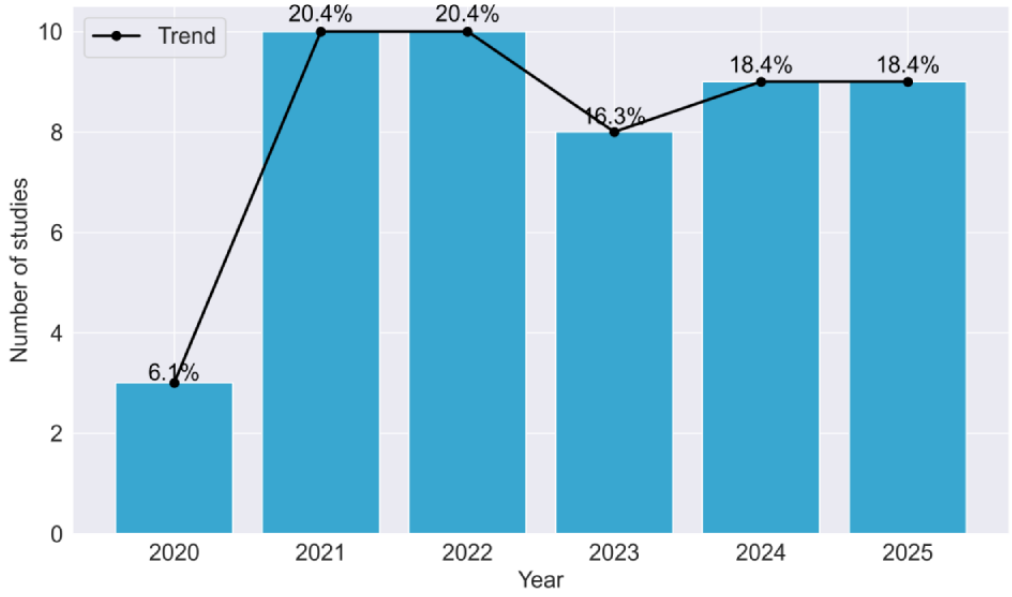

XSL•FO
RenderX

Table 2. Summary of included cross-sectional studies investigating wearable-derived measures in relation to dementia outcomes. The table reports study source, cross-sectional study design, participant groups and mean ages, and assessed risk of bias.

Source	Objective	Groups, n	Age (years), mean (SD)	Risk of bias
Lentzen et al [22]	Early detection	HC[a]=69; DEM[b]=160	69.2 (7.7)	Low
Jones et al [23]	Early detection	NCD[c]=21; MCI[d]=19	67.2 (8.4)	Low
Satomi et al [24]	Early detection	HC=25; MCI=63; DEM=21	79.3 (6.3)	Low
Sauers et al [25]	Risk assessment	HC=196; MCI=47	74	Low
Baril et al [26]	Risk assessment	HC=159; MCI=44	68.25	Low
Seo et al [27]	Early detection	HC=1519; MCI=353; DEM=87	70.1 (6.1)	Low
Huang et al [28]	Preventive intervention	HC=225; MCI=133; DEM=36	71.3(7.8); 72.9 (8.0)	Low
Palmer et al [29]	Progression monitoring	SCI[e]=29; MCI=79	68.8 (8.9)	Low
Lim [30]	Early detection	HC=14; DEM=4	≥65	Low
Boa Sorte Silva et al [31]	Early detection	MCI=113	73.1 (5.7)	Low
Kim et al [32]	Progression monitoring	HC=41; NCD=54	Median 60	Low
Basta et al [33]	General insight	MCI=40; DEM=49	78.3 (6.5)	Low
Roh et al [34]	Progression monitoring	MCI=70; DEM=30	Median 73	Low
Liu et al [35]	Early detection	HC=31; MCI=68	66.5	Low
Espinosa et al [36]	Early detection	SCI=44; MCI=99	67.4 (7.9)	Medium
Manning et al [37]	Risk assessment	HC=7; MCI=2; DEM=5	65.8	Medium
Ghosal et al [38]	Early detection	HC=54; DEM=38	73.4 (7.1)	Medium
Park et al [39]	Early detection	MCI/DEM=177; HC=14,482	74.1 (7); 39 (2.3)	Medium
Palmer et al [40]	Early detection	SCI=23; MCI=44	65.8 (7.9)	Medium
Wei et al [41]	Early detection	HC=20; DEM=8	Mean 59.2	Medium
Corbi and Burgos [42]	Early detection	HC=37; DEM=74	80.5	Medium

[a]HC: healthy controls.

[b]DEM: dementia.

[c]NCD: neurocognitive disorder.

[d]MCI: mild cognitive impairment.

[e]SCI: subjective cognitive impairment.

Characteristics of Included Studies

A total of 49 studies were included in the review (summarized in Tables 3-8). Sample sizes ranged from 14 to 91,948 participants, with a median of 145. A total of 5 large cohorts heavily influenced the mean values (ranging between 47,371 and 91,948 participants). When stratified by analytic approach, studies using traditional statistical methods (36/49, 73.5%) had a median of 100 participants (mean 2597, SD 13,574), while those applying machine learning (7/49, 14.3%) had a median of 99 participants (mean 114.9, SD 76.4). Studies using deep learning (6/49, 12.2%) had a median of 108 participants (mean 397.6, SD 465.4), largely driven by population-scale datasets. Most studies were rated as low risk (34/49, 69.4%), nearly one-third (13/49, 26.5%) were classified as medium risk, and 2/49 (4.1%) as high risk. Participant populations were predominantly composed of individuals with MCI (29/49, 59.2%), dementia (35/49, 71.4%), or healthy controls (36/49, 73.5%). A smaller subset included participants with subjective cognitive decline (4/49, 8.2%) or other at-risk groups (2/49, 4.1%). Regarding study design, half of the studies were cross-sectional (21/49, 42.9%), followed by longitudinal cohorts (24/49, 49%), with fewer randomized controlled trials or diagnostic studies. Most studies (28/49, 57.1%) focused on early detection, while others assessed risk factors (8/49, 19.5%), monitored disease progression (7/49, 14.3%), or preventive interventions (5/49, 10.2%). Risk of bias (complete tables are provided in Multimedia Appendix 5) was generally low (35/49, 71.4%), though higher risk ratings were associated with small sample sizes and lack of external validation.

XSL•FO
RenderX

Table 3. Summary of included cohort studies investigating wearable-derived measures in relation to dementia outcomes. The table reports study source, study objective, cohort study design, participant groups and mean ages, and assessed risk of bias.

Source	Objective	Groups, n	Age (years), mean (SD)	Risk of bias
Baril et al [43]	Progression monitoring	HC[a]=159; MCI[b]=44	68.2 (5.4)	Low
Sun et al [44]	Early detection	HC=807; DEM[c]=270	80.9 (7.3)	Low
Winer et al [45]	Early detection	HC=82,377; DEM=452	62.0 (7.8)	Low
Haghayegh et al [46]	Early detection	HC=90,962; MCI or DEM=555	62.4 (7.8)	Low
Basta et al [47]	Progression monitoring	HC=29; MCI=49; DEM=32	80.3 (6.6)	Low
Milton et al [48]	Risk assessment	HC=476; MCI=164; DEM=93	82.5 (2.9)	Low
Skourti et al [49]	Progression monitoring	HC=146; MCI=231; DEM=128	72.8 (6.7)	Low
Gao et al [50]	Early detection	HC=409; MCI=61; DEM=81	58 (8)	Low
Jeon et al [51]	Early detection	HC=100; MCI=29	69.3 (7.7)	Low
Yi Lee et al [52]	Preventive intervention	HC=123; MCI=51	75.6 (6.9)	Low
Lysen et al [53]	General insight	HC=1262; DEM=60	66.1 (7.6)	Low
Agudelo et al [54]	Early detection	DEM=1035	55.2 (2.5)	Low
Ning et al [55]	Risk assessment	HC=91,212; DEM=736	Median 63	Low
Ning et al [56]	Risk assessment	HC=89,619; DEM=710	Median 64.9	Low
Zhao et al [57]	Risk assessment	HC=87,857; DEM=735	61.9 (7.9)	Low
Lu et al [58]	Early detection	HC=303; MCI=130; DEM=277	81.1 (5.2)	Low
Chan et al [59]	Early detection	HC=46,984; DEM=387	67.0 (4.0)	Low
Shi et al [60]	Early detection	HC=346; DEM=344	70.1 (6.9)	Low
Xiao et al [61]	Progression monitoring	HC=570; MCI=120; DEM=73	84.1	Low
Hoepel et al [62]	Preventive intervention	HC=1849; DEM=50	71.3 (9.26)	Low
Cho et al [63]	Early detection	DEM=222	80.4 (7.4)	Medium
Plotogea et al [64]	Early detection, preventive intervention	HC=25; SCI[d]=7; MCI=17; DEM=25	58.9 (9.8)	Medium
Targa et al [65]	Early detection	DEM=100	Median 76	Medium
Cho et al [66]	Risk assessment	DEM=145	81.2 (6)	Medium

[a]HC: healthy controls.

[b]MCI: mild cognitive impairment.

[c]DEM: dementia.

[d]SCI: subjective cognitive impairment.

Table 4. Summary of included studies with other study designs investigating wearable-derived measures in relation to dementia outcomes. The table reports study source, study objective (early detection or progression monitoring), study design (diagnostic study, randomized controlled trial, or quasi-experimental study), participant groups and ages, and assessed risk of bias.

Source	Objective	Study design	Groups, n	Age (years)	Risk of bias
Rykov et al [67]	Early detection	Quasi-experimental study	MCI[a]=17	60.3 (SD 4.5)	Medium
David et al [68]	Progression monitoring	Randomized controlled trial	DEM[b]=38	70 (SD 7)	Medium
Khosroazad et al [69]	Early detection	Diagnostic study	HC=45; MCI[c]=50	73.6	High
Hossain et al [70]	Early detection	Diagnostic study	HC=22; MCI=6; DEM=5	57.5	High

[a]HC: healthy controls.

[b]DEM: dementia.

[c]MCI: mild cognitive impairment.

XSL•FO
RenderX

Table 5. Wearable devices, recording parameters, and analytic methods used across the included studies. The table lists device type, monitoring duration, product and model, analytic approach (statistical, machine learning, deep learning, or clustering), primary task, and cognitive assessments applied.

Source	Wearable category	Wearable	Product (company name)	Duration (days), n	Model	Cognitive test
Espinosa et al [36]	Research	Actigraph	Actiwatch Spectrum (Philips)	13.7	Statistical	MMSE[a]
Lentzen et al [22]	Commercial research	Smartwatch	Fitbit Charge 3 (Google); AX3 (Axivity); Physilog (Gait Up)	56	Machine learning	Amsterdam IADL[b]
Manning et al [37]	Research	Actigraph	Actiwatch (Philips)	14	Statistical	MoCA[c]
Jones et al [23]	Commercial	Smartwatch	Apple Watch 8 (Apple)	6.9	Statistical	HVLT[d]
Satomi et al [24]	Research	Actigraph	ActTrust (Condor Instruments)	7	Statistical	CDR[e]
Baril et al [43]	Research	Actigraph	Actiwatch (Philips)	6.5	Statistical	RBANS[f]
Sauers et al [25]	Research	Actigraph	Sleep Profiler (Advanced Brain Monitoring); Actiwatch2 (Philips); Alice PDx (Philips)	6	Statistical	CDR
Sun et al [44]	Research	Actigraph	Actical (Philips)	14	Deep learning	NINCDS-ADRDA[g]
Winer et al [45]	Research	Accelerometer	AX3 (Axivity)	7	Statistical	TMT[h]
[67]	Multimodal research	Wrist-worn device	E4 wristband (Empatica)	10	Machine learning	NTB[i]
Haghayegh et al [46]	Research	Accelerometer	AX3 (Axivity)	6	Statistical	*ICD-10*[j]
Basta et al [47]	Ad hoc	Actigraph	Not specified	7	Statistical	*ICD-10*
Baril et al [26]	Research	Actigraph	Actiwatch (Philips)	6.5	Statistical	MMSE
Milton et al [48]	Multimodal research	Actigraph	SleepWatch-O (Ambulatory Monitoring)	3	Statistical	MMSE
Cho et al [63]	Research	Actigraph	ActiGraph wGT3X-BT (ActiGraph)	14	Machine learning	MMSE
Seo et al [27]	Research	Accelerometer	HW-100 (Kao Corporation)	28	Statistical	MMSE
Huang et al [28]	Research	Actigraph	GENEActiv Original (Activinsights Company)	6	Statistical	MMSE
Khosroazad et al [69]	Research	Actigraph	Actiwatch (Philips)	7	Deep learning	HVLT
Skourti et al [49]	Research	Actigraph	Actigraph GT3XP (ActiGraph)	3	Statistical	RAVLT[k]
David et al [68]	Commercial	Smartwatch	Fitbit Charge 2 (Google)	13	Statistical	MoCA
Gao et al [50]	Research	Accelerometer	AX3 (Axivity)	7	Statistical	*ICD-10*
Ghosal et al [38]	Research	Accelerometer	Actigraph GT3x+ (ActiGraph)	7	Machine learning	CDR
Palmer et al [29]	Research	Actigraph	Actiwatch Spectrum (Philips)	10.5	Deep learning	MMSE
Jeon et al [51]	Research	Actigraph	Actiwatch 2 (Philips)	5.5	Statistical	MMSE
Lim et al [30]	Ad hoc	Smartwatch	Not specified	7	Deep learning	MMSE
Plotogea et al [64]	Research	Actigraph	Actiwatch Spectrum Pro (Philips)	7	Statistical	PHES[l]
Boa Sorte Silva et al [31]	Research	Actigraph	MotionWatch8 (CamNtech)	7	Statistical	MoCA
Park et al [39]	Commercial	Accelerometer	Fitmeter (FitNLife); ActiGraph AM-7164 (ActiGraph)	24.6 7	Deep learning	MMSE; CDR; GDS[m]
Palmer et al [40]	Research	Actigraph	Actiwatch Spectrum (Philips)	7	Statistical	MMSE
Kim et al [32]	Research	Actigraph	Spectrum Plus (Philips); Actiwatch Spectrum (Philips)	10.5	Statistical	NIH[n]

XSL•FO
RenderX

Source	Wearable category	Wearable	Product (company name)	Duration (days), n	Model	Cognitive test
Targa et al [65]	Research	Actigraph	Actiwatch 2 (Philips)	14	Statistical	MMSE
Wei et al [41]	Commercial	Smartwatch	Mi Band 2 (Xiaomi)	14	Statistical	MMSE
Hossain et al [70]	Ad hoc	Smartwatch	Not specified	182.5	Machine learning	MMSE
Cho et al [66]	Ad hoc	Actigraph	wGT3X-BT (ActiGraph)	11.5	Statistical	MMSE
Corbi and Burgos [42]	Research	Accelerometer	AX3 (Axivity)	4	Machine learning	GDS
Basta et al [33]	Research	Actigraph	Actigraph GT3XP (ActiGraph)	3	Statistical	MMSE
Yi Lee et al [52]	Research	Actigraph	GENEActiv Original (ActivInsights)	7	Statistical	MoCA
Roh et al [34]	Commercial	Accelerometer	Fitmeter (FitNLife)	4	Statistical	SNSB[o]
Lysen et al [53]	Research	Actigraph	ActiWatch AW4 (CamNtech)	6	Statistical	MMSE
Agudelo et al [54]	Research	Actigraph	Actiwatch Spectrum (Philips)	7	Statistical	B-SEVLT-Sum[p]
Ning et al [55]	Research	Accelerometer	AX3 (Axivity)	7	Statistical	*ICD-10*
Ning et al [56]	Research	Accelerometer	AX3 (Axivity)	7	Statistical	*ICD-10*
Zhao et al [57]	Research	Accelerometer	AX3 (Axivity)	7	Statistical	*ICD-10*
Lu et al [58]	Research	Actigraph	wGT3X-BT (Ametris)	7	Statistical	MoCA
Chan et al [59]	Research	Smartwatch	Not specified	7	Statistical	*ICD-10*
Shi et al [60]	Research	Actigraph	wGT3X-BT (Ametris); Actiwatch Spectrum (Philips)	7; 3	Deep learning	MoCA
Liu et al [35]	Commercial	Accelerometer	W180 (Shenzhen Fitfaith)	14	Machine learning	MoCA
Xiao et al [61]	Research	Actigraph	Actigraph GT3XP (Ametris)	7	Statistical	TMT
Hoepel et al [62]	Research	Actigraph	GENEActiv Original (ActivInsights)	4	Statistical	MMSE

[a]MMSE: Mini-Mental State Examination.

[b]Amsterdam IADL: Amsterdam Instrumental Activities of Daily Living.

[c]MoCA: Montreal Cognitive Assessment.

[d]HVLT: Hopkins Verbal Learning Test-Revised.

[e]CDR: Clinical Dementia Rating.

[f]RBANS: Repeatable Battery for the Assessment of Neuropsychological Status.

[g]NINCDS-ADRDA: National Institute of Neurologic, Communicative Disorders and Stroke and Alzheimer's Disease and Related Disorders Association.

[h]TMT: Trail-Making Test.

[i]NTB: Neuropsychological Test Battery.

[j]ICD-10: International Statistical Classification of Diseases, Tenth Revision.

[k]RAVLT: Rey Auditory Verbal Learning Test.

[l]PHES: Psychometric Hepatic Encephalopathy Score.

[m]GDS: Global Deterioration Scale.

[n]NIH: National Institutes of Health.

[o]SNSB: Seoul Neuropsychological Screening Battery.

[p]B-SEVLT-Sum: Brief Spanish-English Verbal Learning Test.

XSL•FO
RenderX

Table 6. Summary of machine learning used across studies, including model type, main performance metrics, analytical task (classification, regression, or clustering), and the corresponding feature selection or explainable artificial intelligence (XAI) techniques applied. Only the best-performing metric for each model is reported.

Source	Model	Metrics (best results)	Task	Feature selection/XAI
Lentzen et al [22]	LR[a]; DT[b]; RF[c]; XGBoost[d]	AUC[e]=0.73	Binary classification	SHAP[f]
Rykov et al [67]	LR; RF; XGBoost	R^2=0.690; ρ=0.700; MAE[g]=0.460	Regression	Correlation-based feature selection
Cho et al [63]	LR; RF; GBM[h]; SVM[i]	AUC=0.929; accuracy=0.935; precision=0.800; sensitivity=0.956; F_1-score (0.819)	Binary classification	Permutation feature importance
Ghosal et al [38]	GLM[j]; SOFR[k]; SOTDR[l]; SOTDR-L[m]	AUC=0.811; R^2=0.333	Binary classification; regression	LASSO[n]/GEL[o] penalties + functional coefficients
Hossain et al [70]	GBM; SVM; LR + LASSO; CTGAN[p]	Accuracy=0.948	Multiclass classification; regression	Correlation-based feature selection; wrapper feature selection
Corbi and Burgos [42]	Expectation-maximization clustering	Accuracy=0.910	Unsupervised clustering	Feature relevance via variable testing
Liu et al [35]	GBDT[q]; XGBoost[r]	Accuracy=0.757; recall=0.952; AUC=0.628	Binary classification	Permutation importance

[a]LR: logistic regression.
[b]DT: decision tree.
[c]RF: random forest.
[d]XGBoost: Extreme Gradient Boosting.
[e]AUC: area under the curve.
[f]SHAP: Shapley additive explanations.
[g]MAE: mean absolute error.
[h]GBM: gradient boosting machine.
[i]SVM: support vector machine.
[j]GLM: generalized linear model.
[k]SOFR: scalar-on-function regression.
[l]SOTDR: scalar on time-by-distribution regression.
[m]SOTDR-L: SOTDR via time varying L moments.
[n]LASSO: least absolute shrinkage and selection operator.
[o]GEL: group exponential LASSO.
[p]CTGAN: conditional tabular generative adversarial network.
[q]GBDT: gradient boosting decision tree.
[r]XGBoost: Extreme Gradient Boosting.

XSL•FO
RenderX

Table 7. Summary of deep learning used across studies, including model type, main performance metrics, analytical task (classification, regression, or survival analysis), and the corresponding feature selection or explainable artificial intelligence (XAI) techniques applied. Only the best-performing metric for each model is reported.

Source	Model	Metrics (best results)	Task	Feature selection/XAI
Sun et al [44]	CNN[a]; ElasticNet; RSF[b]	C-index=0.840; AUC[c]=0.880	Survival analysis (time to Alzheimer disease onset)	Gradient and Gini feature importance; hazard ratio interpretability
Khosroazad et al [69]	Neural network	AUC=0.880; sensitivity=0.870; specificity=0.890	Binary classification	Intrinsic via time-latency
Palmer et al [29]	MS-GAN[d]; Bayesian LR[e]	Dice=0.730; OR[f]=1.830	Segmentation; regression	Bayesian coefficients (odds ratio and CI)
Lim et al [30]	Neural network + PCA[g]	AUC=0.990	Binary classification	Correlation-based interpretability
Park et al [39]	1D convolutional autoencoder + LR	R^2=0.979	Regression	Backward feature elimination
Shi et al [60]	CDPred[h]; XGBoost[i]	Hip: accuracy=0.84 and AUC=0.86; Wrist: accuracy=0.69 and AUC=0.73	Binary classification	Nonzero predictor relative importance ranking

[a]CNN: convolutional neural network.

[b]RSF: random survival forest.

[c]AUC: area under the curve.

[d]MS-GAN: multispecies generative adversarial network.

[e]LR: logistic regression.

[f]OR: odds ratio.

[g]PCA: principal component analysis.

[h]CDPred: cognitive decline predictor.

[i]XGBoost: Extreme Gradient Boosting.

XSL•FO
RenderX

Table 8. Summary of statistical approaches applied across studies, including model type, reporting metrics, covariates or adjustment factors, and main associations between wearable-derived variables and cognitive impairment or dementia indicators. Only the most relevant and statistically significant results for each study are presented. For wearable-derived metrics, arrows indicate the direction of association: an upward arrow (↑) denotes that higher values of the variable are associated with greater cognitive impairment, whereas a downward arrow (↓) indicates the opposite.

Source	Model	Metrics	Score
Espinosa et al [36]	General linear models	CWP[a]	IV[b] (CWP<.001, ↑); IS[c] (CWP<.001, ↓)
Manning et al [37]	Pearson correlation analysis	Pearson CC[d], *P* value	Activity counts (CC=–0.829; *P*=.041, ↓)
Jones et al [23]	Linear regression models	β coefficient, 95% CI	IV (β≈.40, q=.022, ↑); sleep onset (β=–.28, 95% CI –0.55 to –0.02, ↑); IS (β=–.27, 95% CI –0.54 to 0.00, ↓)
Satomi et al [24]	Multinomial logistic regression	RR[e], 95% CI, *P* value	IV (RR 3.21, 95% CI 1.40-7.34, *P*=.006, ↑); IS (RR 0.44, 95% CI 0.21-0.93, *P*=.03, ↓); M10 (RR 0.40, 95% CI 0.18-0.89, *P*=.02, ↓)
Baril et al [43]	Linear regression models	Standardized β coefficient, *P* value	Sleep duration (β=.384, *P*=.001, ↑)
Sauers et al [25]	Linear regression models	Estimate (β), *P* value	Sleep efficiency (β=–6.026, *P*<.001, ↓); sleep latency (β=11.302, *P*<.001, ↑); number of awakenings (β=6.585, *P*=.001, ↑)
Winer et al [45]	Cox proportional hazards models	HR[g], 95% CI, *P* value	IS (HR 1.25, 95% CI 1.050-1.480; *P*=.007, ↑); amplitude (HR 0.79, 95% CI 0.650-0.960, *P*=.02, ↓); M10 (HR 0.75, 95% CI 0.610-0.940, *P*=.01, ↓); MESOR[h] (HR 0.78, 95% CI 0.590-0.998, *P*=.048, ↓)
Haghayegh et al [46]	Cox proportional hazard models	HR, 95% CI	Amplitude (HR 1.32, 95% CI 1.17-1.49, ↑); M10 (HR 1.28, 95% CI 1.14-1.44, ↑); L5 (HR 1.15, 95% CI 1.10-1.21, ↑); IV (HR 1.14, 95% CI 1.05-1.24, ↑); rest-activity rhythm (HR 1.23, 95% CI 1.16-1.29, ↑)
Basta et al [47]	ANCOVA	*P* value	Sleep duration (night TST[j], *P*<.001, ↑); TiB[k] (night TiB, *P*<.001, ↑)
Baril et al [26]	Linear regression models	*P* value	Sleep duration (*P*<.05, ↑); activity counts (*P*<.05, ↑); circadian rhythm (*P*<.05, ↑)
Milton et al [48]	Multinomial logistic regression	OR[l], 95% CI	Wake after sleep onset (OR 2.26, 95% CI 1.12-4.55, ↑); sleep efficiency (OR 2.15, 95% CI 1.03-4.47, ↓)
Seo et al [27]	Two-way ANCOVA	*P* value	Movement/acceleration (*P*=.03, ↓)
Huang et al [28]	Unconditional multivariable logistic regression	AOR[m], 95% CI	MESOR (AOR 1.99, 95% CI 1.04-3.81, ↑)
Skourti et al [49]	Path models (analysis of moment structures)	Standardized β coefficient, *P* value	Sleep efficiency (direct β=.266, *P*=.001, ↓); wake after sleep onset (direct β=–.211, *P*=.001, ↑); TiB (24-hour TiB, indirect β=–.079, *P*<.001, ↑)
David et al [68]	Spearman rank correlation	Partial correlation coefficient (ρ), *P* value	Moderate-to-vigorous physical activity (ρ=0.558, *P*=.02, ↓)
Gao et al [50]	Cox proportional hazards	HR, 95% CI	Amplitude (HR 1.94, 95% CI 1.53-2.46, *P*<.001, ↑); IV (HR 1.49, 95% CI 1.18-1.88, *P*<.001, ↑); sleep duration (HR 1.28, 95% CI 1.06-1.55, *P*=.01, ↑)
Jeon et al [51]	Multivariate linear regressions	β coefficient, *P* value	Acrophase (β=–.256, *P*=.004, ↑)
Plotogea et al [64]	Multivariate logistic regression	OR, 95% CI, *P* value	Sleep efficiency (OR 0.803, 95% CI 0.711-0.904, *P*=.001, ↓); sleep latency (OR 1.212, 95% CI 1.063-1.383, *P*=.004, ↑)
Boa Sorte Silva et al [31]	Linear regression models	Unstandardized β (β), *P* value	Fragmentation index (β=.004, *P*=.046, ↑)
Palmer et al [40]	Fixel-wise linear regression; Bayesian multiple linear regression	β coefficient, *P* value	L5 (β=.29, *P*<.050, ↑)
Kim et al [32]	Multivariable linear models	Partial rank correlation (ρ), *P* value	IV (ρ=–0.44, *P*<.001, ↑); M10 (ρ=0.45, *P*<.001, ↓); IS (ρ=0.40, *P*=.009, ↓)
Targa et al [65]	Linear regression models	Effect size, *P* value	IV (effect size=–0.715, *P*=.013, ↑)
Wei et al [41]	Descriptive statistics; 1-tailed *t* test	Mean (SD), *P* value	Amplitude (0.93, SD 0.59, *P*=.030, ↓); IS (0.32, SD 0.19, *P*=.02, ↓); acrophase (44, SD 145, *P*<.001, ↓)
Cho et al [66]	Generalized linear mixed model	OR, 95% CI, *P* value	Sleep duration (OR 0.9, 95% CI 0.8-1.0, *P*<.001, ↓); activity counts (OR 0.02, 95% CI 0.0-0.5, *P*=.02, ↓)

XSL•FO
RenderX

Source	Model	Metrics	Score
Basta et al [33]	ANCOVA	Mean (SD), *P* value	Sleep duration (night TST=7.7-hour vs 7.2-hour, *P*=.011, ↑); sleep duration (24-hour TST=8.5-hour vs 7.8-hour, *P*=.012, ↑)
Yi Lee et al [52]	Multivariate logistic regression; multinomial logistic regression	AOR, 95% CI	Percent rhythm (AOR 0.26, 95% CI 0.08-0.79, ↓)
Roh et al [34]	Multiple linear regression	Estimate (β), SE, *P* value	MESOR (β=1.17, SE=0.37, *P*<.001, ↓); L5 (β=3.77, SE=1.22, *P*<.001, ↓)
Lysen et al [53]	Cox proportional hazards	HR, 95% CI	Sleep latency (HR 1.44, 95% CI 1.13-1.83, ↑); TiB (HR 1.40, 95% CI 1.04-1.88, ↑); sleep efficiency (HR 0.72, 95% CI 0.55-0.93, ↓); wake after sleep onset (HR 1.38, 95% CI 1.10-1.74, ↑)
Agudelo et al [54]	Survey linear regression models	β coefficient, *P* value	Sleep latency (β=–.003, *P*<.001, ↑); sleep duration (β=–.070, *P*<.05, ↑)
Ning et al [55]	Cox proportional hazards	HR, 95% CI	Moderate-to-vigorous physical activity (HR 0.69, CI 0.54-0.87, *P*<.001, ↓)
Ning et al [56]	Cox proportional hazards	HR, 95% CI	Moderate-to-vigorous physical activity (HR 0.60, 95% CI 0.40-0.90, ↓)
Zhao et al [57]	Cox proportional hazards	HR, 95% CI	Sleep duration (HR 0.801, 95% CI 0.717-0.893, ↓)
Lu et al [58]	Logistic regression models ANCOVA	OR, 95% CI	Relative amplitude (OR 1.68, 95% CI 1.12-2.50, ↑)
Chan et al [59]	Cox proportional hazards	HR, 95% CI	Bedtime (HR 1.52, 95% CI 1.22-1.85, ↑)
Xiao et al [61]	Cox proportional hazards	HR, 95% CI	MESOR (HR 2.45, 95% CI 1.52-3.94, ↓)
Hoepel et al [62]	Cox proportional hazards	HR, 95% CI	Sedentary behavior (HR 0.36, 95% CI 0.24-0.55, ↓)

[a]CWP: clusterwise *P* value.
[b]IV: intradaily variability.
[c]IS: interdaily stability.
[d]CC: correlation constant.
[e]RR: relative risk.
[f]M10: most active 10-hour.
[g]HR: hazard ratio.
[h]MESOR: midline estimated statistic of rhythm.
[i]L5: least active 5-hour.
[j]TST: total sleep time.
[k]TiB: time in bed.
[l]OR: odds ratio.
[m]AOR: adjusted odds ratio.

Feasibility of Quantitative Synthesis

The feasibility of quantitative meta-analysis for early detection depends on whether the included studies estimate a common outcome construct or estimand. Quantitative synthesis requires studies to measure the same underlying construct. In the current literature, early detection is not operationalized as a single measurable outcome but is addressed through multiple, nonequivalent analytical objectives and metrics. The evidence base is characterized by the following dimensions:

Reported statistical metric: of the 49 included studies, 13 (26.5%) report performance metrics from AI-based approaches, such as accuracy or area under the curve. The remaining 36 studies report heterogeneous statistical effect measures, including hazard ratios for time-to-event analyses (n=16), odds ratios (ORs) for binary risk (n=8), and beta coefficients or correlation measures for continuous associations (n=12). These metrics reflect distinct inferential targets and are not interchangeable.

Device and sensor characteristics: most studies rely on research-grade actigraphy with access to raw accelerometry data (43/49, 87.8%), whereas a smaller subset uses consumer-grade wearables integrating additional sensors and proprietary processing (7/49, 14.3%). Differences in device type, sensor modality, and preprocessing lead to wearable-derived measures that do not map onto a common estimand.

Study design: study designs were divided between cross-sectional analyses assessing concurrent associations (21/49, 42.8%) and longitudinal cohort studies estimating prospective risk or disease onset (24/49, 49%). These designs address different research questions and operate over distinct temporal frameworks.

XSL•FO
RenderX

Clinical endpoint: cognitive outcomes included both categorical clinical diagnoses and continuous screening scores. The MMSE was the most frequently used outcome (18/49, 36.7%), followed by *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*)–based diagnoses and the MoCA (each 7/49, 14.3%). These endpoints differ in scale, sensitivity, and clinical interpretation.

Because early detection is not defined as a single measurable construct across studies, no common estimand can be specified for quantitative pooling. A structured narrative synthesis is therefore adopted to examine how analytical approaches, cognitive assessments, and wearable measurement strategies are used to demonstrate early detection capability across the cognitive continuum.

Wearable Categories

Wearable devices were categorized into 4 groups according to their primary intended use. Research devices were by far the most common, reported in 87.8% (43/49) of studies, and included systems such as ActiGraph, Actiwatch, and accelerometer loggers (eg, Axivity AX3). These devices were purpose-built and validated for assessing activity, sleep, and circadian rhythms in research and clinical contexts. Commercial everyday wearables, such as Fitbit, Apple Watch, and Xiaomi Mi Band, were used in 14.3% (7/49) of studies; these multipurpose consumer devices are affordable and widely accessible but often provide limited access to raw data. Ad hoc prototypes, developed specifically for individual studies, accounted for 8.2% (4/49) of studies. Finally, multimodal research devices were used in 4.1% (2/49) of studies, typically integrating physiological sensing beyond accelerometry, such as heart rate, electrodermal activity, or peripheral arterial tonometry. Counts are nonexclusive because some studies used multiple device types in the same cohort [19].

Across all categories, wear time was typically about 1 week (median, 7 days), with durations ranging from 3-4 days to several months (up to 182.5 days in 1 longitudinal cohort). Most of the studies (29/49, 59.2%) relied on actigraphy devices to capture rest–activity rhythms and sleep, followed by accelerometer devices (12/49, 25.5%) and smartwatches (7/49, 14.3%).

Cognitive Outcomes

The most frequently used cognitive measures were global screening tests, with their distribution shown in Figure 3. The MMSE was the most frequently used measure, applied in 36.7% (18/49) of studies. Clinical diagnoses based on *ICD-10* dementia or MCI codes were reported in 14.3% (7/49) of studies. The MoCA appeared in 14.3% (7/49) of studies and Clinical Dementia Rating was used in 8.2% (7/49) of studies. The Hopkins Verbal Learning Test and the Global Deterioration Scale were each used in 4.1% (2/49) of studies. A wide range of other neuropsychological and functional instruments, including the Amsterdam Instrumental Activities of Daily Living scale, Repeatable Battery for the Assessment of Neuropsychological Status, National Institute of Neurologic, Communicative Disorders and Stroke and Alzheimer's Disease and Related Disorders Association criteria, Trail Making Test, Neuropsychological Test Battery, Rey Auditory Verbal Learning Test, Psychometric Hepatic Encephalopathy Score, National Institutes of Health toolbox, Seoul Neuropsychological Screening Battery, and Brief Spanish-English Verbal Learning Test-Sum, were each applied in 2% (1/49) of studies. The overall distribution of cognitive measures is shown in Figure 3.

One study applied AI-based approaches to directly predict cognitive test scores [67], while others relied on predefined cut-off points for classification [23,36,41,46,55,57]. Some studies used both direct score estimation and classification approaches [43,54]. A separate line of work focused on predicting survival outcomes, which were not directly comparable to cognitive test scores [43]. Finally, 2 studies addressed related regression tasks by estimating white matter characteristics in the brain from wearable-derived data [44,49].

Across the 49 studies, the most frequently reported wearable-derived variables (complete list is provided in Multimedia Appendix 6) used to estimate cognitive assessments (Figure 4) were predominantly sleep-related. Sleep duration was the most frequently reported measure (22/49, 44.9%), followed by wake after sleep onset (16/49, 32.7%), sleep efficiency (16/49, 32.7%), and circadian rhythm metrics (16/49, 32.7%). Activity-related variables were reported less often overall, including movement/acceleration (12/49, 24.5%), activity counts (9/49, 18.4%), step count (7/49, 14.3%), moderate-to-vigorous physical activity (6/49, 12.2%), active minutes (5/49, 10.2%), and sedentary behavior (5/49, 10.2%). Physiological variables were least commonly reported, with skin temperature appearing in a small subset of studies (5/49, 10.2%).

XSL•FO
RenderX

Figure 3. Frequency of cognitive assessment instruments across included studies. The bar chart illustrates the distribution of specific cognitive tests and diagnostic criteria used. Frequencies and percentages represent the prevalence of each tool within the total sample of studies. Amsterdam IADL: Amsterdam Instrumental Activities of Daily Living; B-SEVLT-Sum: Brief Spanish-English Verbal Learning Test; CDR: Clinical Dementia Rating; GDS: Global Deterioration Scale; HVLT: Hopkins Verbal Learning Test-Revised; ICD-10: International Classification of Diseases, Tenth Revision; MMSE: Mini-Mental State Examination; MoCA: Montreal Cognitive Assessment; NIH: National Institutes of Health; NINCDS-ADRDA: National Institute of Neurologic, Communicative Disorders and Stroke and Alzheimer's Disease and Related Disorders Association; NTB: Neuropsychological Test Battery; PHES: Psychometric Hepatic Encephalopathy Score; RAVLT: Rey Auditory Verbal Learning Test; RBANS: Repeatable Battery for the Assessment of Neuropsychological Status; SNSB: Seoul Neuropsychological Screening Battery; TMT: Trail-Making Test.

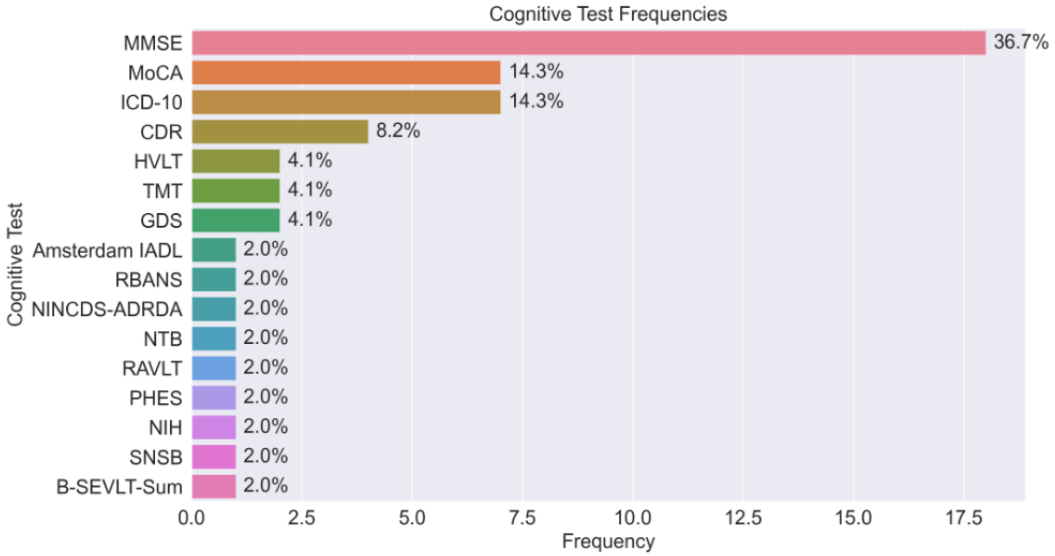

Figure 4. Top 20 most frequently used wearable-derived variables across studies. Variables are grouped by domain: sleep (blue), activity (green), and physiological (red). Frequencies represent the number of studies reporting each variable after standardization and filtering for wearable compatibility.

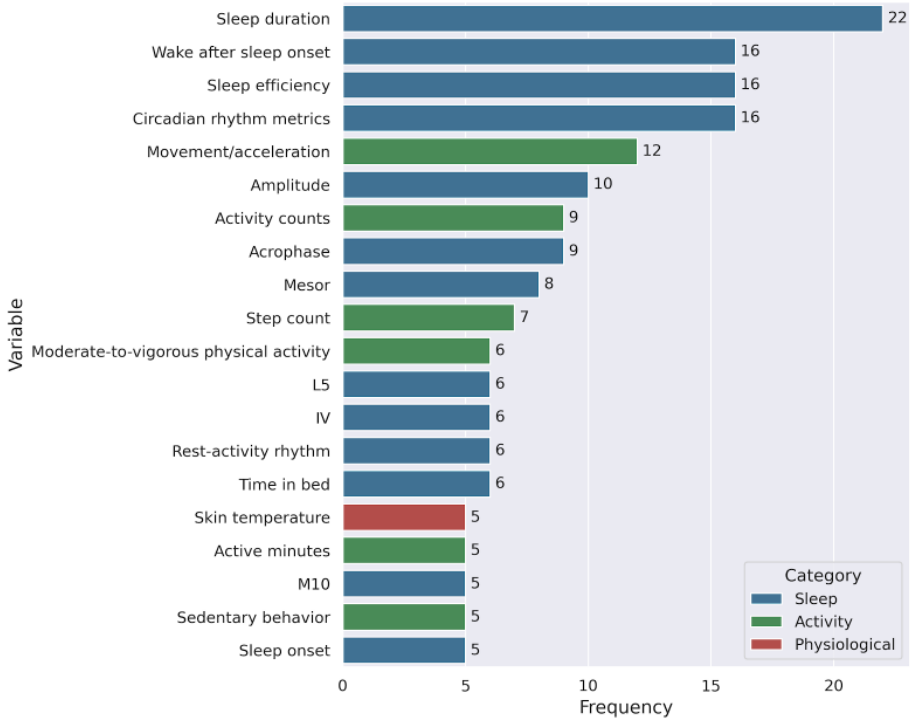

Wearable Data Analysis Approaches

Traditional statistical methods were the most frequently applied, reported in 73.5% (36/49) of studies (Table 3). These included group comparisons, correlation analyses, and regression models evaluating associations between wearable-derived activity or sleep metrics and cognitive outcomes. Commonly reported features were total daily activity counts, sleep duration and efficiency, circadian rhythm indices (intradaily variability, interdaily stability, and relative amplitude), and timing markers such as acrophase or sleep midpoint. The list of wearable-derived features used across these studies, with a description per feature, is included in Multimedia Appendix 6.

XSL•FO
RenderX

Most studies relying on conventional statistical approaches reported significant associations between wearable-derived variables and cognitive outcomes, although the magnitude of reported effects varied with sample size. In smaller studies (n<100), standardized regression coefficients and odds ratios tended to be larger, typically ranging from β≈.35-.55 or OR 2.0-3.1 [24,29,43]. In larger cohort studies (n≥100), effect estimates were generally lower, around β≈.10-.25 or OR 1.3-1.8 [45,46,49,54].

According to established benchmarks for social and behavioral sciences, standardized coefficients are typically categorized as small (β≈.20), medium (β≈.50), or large (β≈.80) [71]. While the estimates from larger cohorts in this review fall within the small-to-moderate range, they must be interpreted within the specific context of cognitive research. Given that cognitive performance is a multifactorial outcome influenced by a broad array of genetic, environmental, and demographic variables, the observation of stable effect sizes between .10 and .25 is highly meaningful. Such values represent robust contributions to the variance that, despite their conservative magnitude, suggest that wearable-derived metrics are reliable markers of cognitive health at a population level. While most associations reached statistical significance (P<.05), only a limited number of studies reported fully adjusted models or external validation [72].

Machine learning methods were applied in 14.3% (7/49) of studies, most often for classification tasks such as distinguishing MCI or dementia from healthy controls. Machine learning models included logistic regression, decision trees, support vector machines, random forests, and gradient-boosting approaches such as Extreme Gradient Boosting (XGBoost). Deep learning approaches were reported in 12.2% (6/49) of studies and were typically applied in larger datasets, where models captured temporal patterns of activity or sleep. Deep learning architectures included neural networks, convolutional neural networks, generative adversarial network–based approaches, and autoencoders. In addition, a small number of studies (2/49, 4.1%) applied clustering techniques to identify subgroups with distinct rest–activity profiles.

Across machine learning and deep learning studies, reported performance metrics varied widely depending on task, cohort size, and study design. Classification performance ranged from moderate discrimination in large, pragmatic settings (eg, area under the receiver operating characteristic curve [AUROC]≈0.73; n=229; dementia=160 and healthy controls=69 [23]) to higher values in smaller or more constrained cohorts (eg, AUROC≈0.93; n=222; affective symptoms=126 and not affective symptoms=96 [63]). Deep learning models reported metrics such as a C-index of approximately 0.84 (n=1077; dementia=270 and healthy controls=807) for long-horizon prediction of Alzheimer disease onset [44] and R^2 values approaching 0.98 (n=14,659; dementia=177 and healthy controls=14,482) for regression tasks estimating brain structural characteristics from wearable-derived data [40]. Several high-performing models were trained on small or imbalanced datasets, including studies with limited case numbers or highly unequal case–control ratios [28,31,40,70].

In addition, machine learning approaches applied to wearable-derived features achieved high discriminative performance across multiple tasks. Reported results included classification accuracies up to 93.5% [63] (n=222; sleep and nighttime behaviors=81 and not sleep and nighttime behaviors=141) for identifying high-risk individuals, accuracies of approximately 94%-95% (n=33; dementia=5, MCI=6, and healthy controls=22) across multiple cognitive impairment levels [70], area-under-the-curve values ranging from approximately 0.73 (n=229; dementia=160 and healthy controls=69) [23] to 0.99 (n=18; dementia=4 and healthy controls=14) [31] for early or prodromal Alzheimer disease detection, sensitivities and specificities exceeding 85% (n=95; MCI=50 and healthy controls=45) [69] for distinguishing MCI from normal cognition, and C-index values up to 0.84 (n=1077; dementia=207 and healthy controls=807) [44] for long-horizon Alzheimer disease onset prediction.

Interpretability approaches were reported in a limited subset of machine learning and deep learning studies. Explicit explainability techniques were applied in 46.1% (6/13) of AI-based studies, including Shapley additive explanations values or permutation feature importance in 2 machine learning studies [23,49], functional coefficients in 1 statistical–machine learning hybrid approach [31], and Bayesian coefficient estimation in 1 deep learning framework [32]. One additional study reported feature relevance through statistical testing in a clustering-based approach [39]. The remaining AI-based studies relied on feature-selection heuristics or dimensionality reduction methods, such as correlation-based selection, wrapper methods, or principal component analysis, without reporting interpretable model outputs [33,35,67,70].

Prevention-Oriented Findings

The included studies were divided into 2 categories according to the type of contribution reported. A total of 22.4% (11/49) of studies explicitly provided quantitative results relevant to the early detection or prevention of cognitive decline (direct evidence), defined as studies demonstrating or validating predictive or preventive applications. The remaining 38 (77.6%) studies contributed indirect evidence, identifying associations between wearable-derived features and cognitive outcomes that may inform, but do not yet constitute, preventive or predictive applications.

Among studies providing direct evidence (n=11), longitudinal or interventional designs were commonly used. Specifically, 45.5% (5/11) of studies used longitudinal follow-up or randomized designs, while 54.5% (6/11) of studies relied on cross-sectional analyses. Direct-evidence studies included randomized intervention trials reporting outcomes such as slower progression to dementia, improved gait speed, or increased adherence to preventive programs, as well as longitudinal cohort studies showing that disrupted rest–activity rhythms, sleep fragmentation, and physical activity patterns predicted incident dementia or MCI over follow-up periods extending up to 8-15 years. Several studies further demonstrated that wearable-derived motor activity features could forecast clinical Alzheimer disease onset, with reported C-index values ranging from approximately 0.80 to 0.84.

XSL•FO
RenderX

In contrast, studies contributing indirect evidence (n=38) more frequently adopted cross-sectional designs. Of these, 52.6% (20/38) of studies were cross-sectional, whereas 47.4% (18/38) of studies used longitudinal designs. These studies primarily focused on identifying associations between wearable-derived sleep, circadian rhythm, activity, or physiological measures and cognitive outcomes, including global cognitive scores, neuropsychological performance, biomarkers, or clinical diagnoses, rather than on prediction or intervention.

Across all included studies, longitudinal designs accounted for 44.9% (22/49) of studies, while 55.1% (27/49) of studies used cross-sectional designs. This distribution reflects differing study aims, with direct evidence studies more often incorporating longitudinal follow-up or intervention components, and indirect evidence studies emphasizing associative analyses.

Overall, wearable technologies contributed to prevention-oriented research both directly, by demonstrating predictive and early detection capabilities in longitudinal and interventional settings, and indirectly, by identifying behavioral and physiological markers associated with future cognitive decline. This dual contribution reflects the role of wearables as both measurement tools for preventive interventions and sources of early behavioral risk markers in dementia research.

Discussion

Principal Findings

This systematic review synthesized 49 studies published since 2020 on wearable devices for the early detection and prevention of cognitive impairment and dementia. The findings are presented in 4 dimensions corresponding to the subobjectives described earlier.

Wearable Devices and Measurement Contexts

Research-grade actigraphy and accelerometry devices were used in most of the included studies (43/49, 87.8%), reflecting both a strong methodological emphasis on validated access to raw accelerometry data and the relative maturity of actigraphy-based approaches in this field [73,74]. Evidence from these studies was derived predominantly from observational designs, including cross-sectional and longitudinal cohorts, and was frequently assessed as having low risk of bias, particularly among cohort studies with large samples [48,53,60]. Across actigraphy-based studies, sample sizes were heterogeneous but predominantly large, with nearly four-fifths enrolling 100 participants or more, and large samples were strongly associated with low risk of bias. While research-grade devices enable robust investigation of behavioral markers associated with cognitive impairment, their limited scalability and accessibility may constrain broader clinical and public health deployment [75].

In contrast, consumer-grade wearable devices were comparatively underrepresented despite their widespread adoption and potential for large-scale, longitudinal monitoring. Evidence from consumer-wearable studies was derived predominantly from cross-sectional designs and exhibited greater heterogeneity in sample size and methodological quality, with smaller studies more frequently assessed as having moderate risk of bias, reflecting restricted access to raw data and reliance on proprietary signal processing [41,68,76]. However, consumer-wearable evidence also included a small number of large cross-sectional studies assessed as having low risk of bias, indicating that methodological rigor is driven more by study scale and design than by device class alone [22,34,35]. Taken together, these findings suggest that different wearable device classes capture complementary aspects of behavior and physiology rather than interchangeable measurements, supporting the potential value of multimodal integration or purpose-driven device selection aligned with specific research or clinical objectives [76].

Wearable-Derived Features and Cognitive Outcomes

Across included studies, cognition was frequently assessed using brief global screening instruments or standardized clinical diagnostic classifications, reflecting a pragmatic choice aligned with clinical practice and the feasibility requirements of observational research. Evidence derived from commonly used cognitive measures was drawn primarily from cross-sectional and cohort designs and was frequently assessed as having low risk of bias, particularly in larger cohort studies [46,50,55,56]. While global cognitive measures facilitate standardized comparisons across heterogeneous cohorts, their limited sensitivity to subtle or domain-specific changes constrains interpretation of associations with wearable-derived features, especially in preclinical populations [77].

Wearable devices were not used to measure cognition directly but to capture continuous behavioral and physiological signals that were statistically associated with cognitive outcomes. Sleep-related and circadian rhythm measures were examined most frequently and were reported across both cross-sectional and longitudinal studies, predominantly in cohort designs with large samples, and were most often assessed as having low risk of bias when derived from validated actigraphy devices and predefined metrics [33,47,59,62]. In contrast, physiological signals beyond accelerometry were examined in only a small subset of studies, which used heterogeneous, primarily noncohort designs, often involved small to medium sample sizes, and showed higher proportions of moderate-to-high–risk-of-bias assessments [35,41] limiting the strength of inferences that can currently be drawn from these measures and highlighting priorities for future methodological development.

Analytical Approaches and Methodological Considerations

Conventional statistical methods were the predominant analytical approach, used in 73.5% (36/49) of studies, reflecting their interpretability and familiarity in clinical and epidemiological research. These methods were applied mainly in cross-sectional and longitudinal observational designs and were frequently assessed as having low risk of bias, particularly in large cohort studies using multivariable or survival models [45,46,50,55-62]. Across these studies, associations between wearable-derived features and cognitive outcomes were generally modest and sensitive to covariate adjustment, especially in large and heterogeneous samples, underscoring

XSL•FO
RenderX

the importance of adequate sample size, rigorous model specification, and replication for reliable inference.

Machine learning and deep learning approaches were applied in a smaller subset of studies (13/49, 26.5%) and were used primarily for classification or prediction tasks. These studies were frequently exploratory, relied on small or imbalanced samples, and were more often assessed as having moderate to high risk of bias, largely due to limited external validation and optimistic performance estimates [67,69,70]. Across AI-based studies, higher predictive performance was more commonly reported in cross-sectional or nonlongitudinal designs using constrained datasets, whereas applications in larger or longitudinal cohorts generally yielded more modest performance estimates but were more often assessed as having low or moderate risk of bias, reflecting improved robustness and generalizability [35,44]. Overall, this pattern suggests a trade-off between performance and methodological robustness, in which performance gains observed in small or exploratory samples may reflect overfitting or cohort-specific structure rather than generalizable predictive signal.

Implications for Early Detection and Prevention

Evidence supporting early detection or prevention was derived from a limited subset of studies providing direct prevention evidence, which more often used cohort-based or predictive modeling designs, enrolled larger samples, and were predominantly assessed as having low risk of bias. Representative examples include large longitudinal or predictive studies demonstrating that wearable-derived behavioral measures, particularly sleep–wake organization and circadian regularity, can precede clinical cognitive impairment by several years [39,46,56]. These findings support the potential role of wearables in early risk stratification rather than post hoc characterization.

In contrast, most of the included studies contributed indirect prevention evidence and relied primarily on cross-sectional or association-focused designs. Although these studies were frequently assessed as having low to moderate risk of bias and consistently reported associations between wearable-derived markers and cognitive status, their typically limited temporal resolution and lack of predictive validation constrained their ability to establish preventive relevance [24,27,33,40]. As a result, their contribution to prevention remains inferential rather than actionable.

Overall, the current evidence supports wearable devices as tools for monitoring and early risk identification rather than as stand-alone preventive interventions [78]. Across both direct and indirect evidence, disrupted sleep–wake patterns and circadian irregularity consistently emerged as markers of elevated cognitive risk [33,39,55]. Translating these associations into effective prevention strategies will require confirmation in larger, longitudinal, and externally validated studies [78] that integrate wearable-based risk stratification with clinical decision-making and behavioral intervention frameworks.

Limitations

Several limitations should be considered when interpreting these findings. The included studies were heterogeneous in design, target populations, device types, and outcome definitions. Devices ranged from consumer-grade wearables to research actigraphy and multimodal physiological instruments, and the metrics derived from them were inconsistently defined. Cognitive outcomes also varied: global screening tools such as the MMSE and MoCA were most frequent, but other studies used different neuropsychological tests, functional measures, or clinical diagnoses. This heterogeneity precluded formal meta-analysis and necessitated a structured narrative synthesis. We did not formally assess the certainty of evidence (eg, using Grading of Recommendations Assessment, Development, and Evaluation), given the methodological heterogeneity of included studies and the predominance of observational designs.

The evidence base is also constrained by small sample sizes and short monitoring periods. In total, 8.2% (4/49) of studies enrolled 30 participants or fewer, while an additional 24.5% (12/49) included between 30 and 100 participants, indicating that more than one-third of the evidence relies on small cohorts. Furthermore, 67.3% (33/49) of studies monitored participants for only 1 week or less, limiting the ability to capture long-term variability in activity, sleep, or circadian rhythms. Only 18.4% (9/49) of studies systematically assessed real-world feasibility, including adherence, comfort, or long-term usability. Moreover, external validation was performed in just 6.1% (3/49) of studies, restricting the generalizability of reported associations or predictive models beyond the original cohorts. Thus, a key limitation of the current evidence is the risk of overinterpretation. Most studies were cross-sectional or small cohorts, and only few have undergone external validation [29,39,63]. In several studies, high accuracy or AUROC values were reported in the context of small or imbalanced datasets, conditions that increase the risk of overfitting and optimism bias. As a result, these performance estimates should be interpreted cautiously in the absence of external validation or evaluation in independent cohorts. While models often report high accuracy, these findings may not generalize. Preventive claims remain speculative until tested in large prospective cohorts or intervention trials. In addition, research-grade devices typically provide access to raw, high-resolution data and validated measurement pipelines, whereas consumer devices often rely on proprietary algorithms and offer limited transparency, which may affect comparability across studies.

The risk of bias was present to varying degrees. Although most studies were rated as low risk (34/49, 69.4%), nearly one-third (13/49, 26.5%) were classified as medium risk, and 4.1% (2/49) as high risk. The high-risk studies exemplify common problems such as selective sampling, inadequate reference standards, and small or restricted populations, all of which likely led to overestimation of diagnostic performance and limited the generalizability of their findings. Publication bias is also possible, as studies reporting positive or novel results are more likely to be published. Together, these issues highlight that the field remains fragmented, with methodological and reporting inconsistencies that restrict the strength and reproducibility of current evidence.

Limitations of the review process itself must also be acknowledged. The restriction to English-language publications and peer-reviewed journals may have introduced language or

XSL•FO
RenderX

publication bias, potentially excluding relevant gray literature or local studies. Additionally, while we searched for 4 major interdisciplinary databases, the exclusion of specialized clinical indices (eg, PsycINFO or Embase) could have resulted in the omission of some relevant records.

Implications for Future Research and Practice

This review highlights several priorities for future research and practice. A key challenge is bridging the gap between research-grade actigraphy and consumer wearables. While actigraphy remains the most validated tool for assessing sleep, activity, and circadian rhythms, limited use of consumer devices constrains scalability [79]. Evidence suggests that some multisensor consumer wearables can achieve accuracy comparable to research-grade actigraphs under controlled conditions [80,81], but direct comparisons across diverse populations and real-world settings remain needed. Where sufficient accuracy and reliability are demonstrated, consumer wearables could support large-scale, longitudinal monitoring for research and preventive care [82].

Another priority is the expanded use of advanced analytics and digital phenotyping. Longitudinal wearable data enable modeling of dynamic behavioral and physiological changes, supporting early and personalized prevention. However, machine learning and deep learning methods remain underused, particularly sequence-based models such as recurrent neural networks, temporal convolutional models, and transformer-based approaches (eg, Informer and TimesFM) that are well suited for time-series health data [83]. In addition, large language model–based frameworks (eg, HealthLLM) offer opportunities to integrate wearable-derived time series with clinical and contextual information [84]. Progress in this area will depend on access to larger, well-annotated datasets and multicenter collaboration, as well as the incorporation of explainable AI methods to ensure interpretability and clinical trust [85].

Finally, more prevention-oriented research is needed. Only 2 randomized controlled trials tested wearable-assisted interventions, and a small subset of observational studies examined prevention-relevant behaviors. Yet prevention is central to dementia care given the absence of disease-modifying treatments. Wearables could play a dual role, both by detecting early signs of decline and by monitoring modifiable lifestyle factors such as sleep, circadian rhythms, and physical activity [86,87]. Positioning wearables within preventive frameworks could therefore be among the most clinically meaningful directions for future research.

From a translational perspective, ethical and regulatory considerations will be critical for the clinical adoption of AI-based dementia risk prediction. Issues such as data privacy, transparency of algorithms, potential psychological impact of early risk labeling, and compliance with regulatory frameworks must be carefully addressed. At present, research-grade wearable devices and analytically transparent models appear more suitable for clinical research settings, whereas many consumer-grade devices and advanced AI approaches remain primarily experimental. Further validation, standardization, and regulatory oversight will be necessary before these methods can be routinely integrated into clinical practice.

Overall, this review advances the field by framing the transition from descriptive statistics to a predictive digital phenotyping framework. Distinct from prior reviews limited to established dementia, it synthesizes evidence specifically for the preclinical window, distinguishing findings with direct predictive utility from those offering only indirect correlational insights. It contributes a critical assessment of methodological maturity, identifying the heavy reliance on research-grade devices and the lack of external validation as primary barriers to implementation. Ultimately, these findings suggest that shifting to continuous, passive monitoring offers a scalable method to detect subtle behavioral deviations, creating opportunities for earlier, personalized risk reduction strategies [88].

Acknowledgments

The authors thank the Vicomtech Foundation for supporting this research. The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing, reformatting, and quality assessment. The GAI tool used was: ChatGPT-5.1 (OpenAI). Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by: “collective responsibility.”

Funding

This research was funded by the Vicomtech Foundation. The funder had no role in the study design, data collection, analysis, interpretation, or writing of the manuscript.

Data Availability

No new primary data were generated in this systematic review. Data extracted from the included studies and the curated list of screened articles are available from the corresponding author upon reasonable request. The Python code used to support automated screening and study management is not publicly available, as it was developed for internal use; however, it can be shared upon reasonable request for purposes of methodological transparency.

XSL•FO
RenderX

Authors' Contributions

Conceptualization was carried out by AC. Methodology was developed by AC and MA, and software was prepared by AC. Validation was conducted by CM and AA. Data curation was performed by AC and MA. The original draft of the manuscript was written by AC and MA, and review and editing were undertaken by AC, MA, CM, and AA. Supervision was provided by CM and AA, while project administration was managed by CM. Funding acquisition was secured by CM. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA 2020 checklist.
[DOCX File , 40 KB - jmir_v28i1e86262_app1.docx]

Multimedia Appendix 2
PRISMA-S checklist.
[DOCX File , 17 KB - jmir_v28i1e86262_app2.docx]

Multimedia Appendix 3
Exact search queries used across all databases to identify records included in this review.
[PDF File (Adobe PDF File), 53 KB - jmir_v28i1e86262_app3.pdf]

Multimedia Appendix 4
Extended study characteristics and principal findings of all the studies included in this systematic review.
[XLSX File (Microsoft Excel File), 16 KB - jmir_v28i1e86262_app4.xlsx]

Multimedia Appendix 5
Risk of bias assessments of included studies using validated tools appropriate to study design.
[PDF File (Adobe PDF File), 276 KB - jmir_v28i1e86262_app5.pdf]

Multimedia Appendix 6
List and definitions of sleep, circadian rhythm, physical activity, and physiological variables extracted from wearable devices across included studies.
[PDF File (Adobe PDF File), 111 KB - jmir_v28i1e86262_app6.pdf]

References

1. Dementia. World Health Organization. 2025. URL: https://www.who.int [accessed 2025-08-30]
2. Chowdhary N, Barbui C, Anstey KJ, Kivipelto M, Barbera M, Peters R. Reducing the risk of cognitive decline and dementia: WHO recommendations. Front Neurol 2021;12:765584 [FREE Full text] [doi: 10.3389/fneur.2021.765584] [Medline: 35082745]
3. Risk Reduction of Cognitive Decline and Dementia: WHO guidelines. Geneva: World Health Organization; 2019:96.
4. Chong TWH, Macpherson H, Schaumberg MA, Brown BM, Naismith SL, Steiner GZ. Dementia prevention: the time to act is now. Med J Aust 2021;214(7):302-304. [doi: 10.5694/mja2.50972] [Medline: 33715184]
5. Jones A, Ali MU, Kenny M, Mayhew A, Mokashi V, He H. Potentially modifiable risk factors for dementia and mild cognitive impairment: an umbrella review and meta-analysis. Dement Geriatr Cogn Disord 2024;53(2):91-106 [FREE Full text] [doi: 10.1159/000536643] [Medline: 38346414]
6. Hardcastle C, Taylor B, Price C. Global Deterioration Scale. In: Encyclopedia of Gerontology and Population Aging. Cham: Springer; 2022:2198-2201.
7. Baron-Shahaf D, Shahaf G. Markers of too little effort or too much alertness during neuropsychological assessment: demonstration with perioperative changes. Brain Behav 2024;14(8):e3649 [FREE Full text] [doi: 10.1002/brb3.3649] [Medline: 39169455]
8. Holding BC, Ingre M, Petrovic P, Sundelin T, Axelsson J. Quantifying cognitive impairment after sleep deprivation at different times of day: a proof of concept using ultra-short smartphone-based tests. Front Behav Neurosci 2021;15:666146. [doi: 10.3389/fnbeh.2021.666146] [Medline: 33927603]
9. Salis F, Costaggiu D, Mandas A. Mini-mental state examination: optimal cut-off levels for mild and severe cognitive impairment. Geriatrics (Basel) 2023;8(1):12 [FREE Full text] [doi: 10.3390/geriatrics8010012] [Medline: 36648917]

XSL•FO
RenderX

10. Arevalo-Rodriguez I, Smailagic N, Roqué-Figuls M, Ciapponi A, Sanchez-Perez E, Giannakou A. Mini-mental state examination (MMSE) for the early detection of dementia in people with mild cognitive impairment (MCI). Cochrane Database Syst Rev 2021;7(7):CD010783. [doi: 10.1002/14651858.CD010783.pub3] [Medline: 34313331]
11. Chithiramohan T, Santhosh S, Threlfall G, Hull L, Mukaetova-Ladinska EB, Subramaniam H. Culture-fair cognitive screening tools for assessment of cognitive impairment: a systematic review. J Alzheimers Dis Rep 2024;8(1):289-306. [doi: 10.3233/adr-230194]
12. Sakal C, Li T, Li J, Li X. Predicting poor performance on cognitive tests among older adults using wearable device data and machine learning: a feasibility study. NPJ Aging 2024;10(1):56 [FREE Full text] [doi: 10.1038/s41514-024-00177-x] [Medline: 39587119]
13. Soumma SB, Mamun A, Ghasemzadeh H. AI-powered wearable sensors for health monitoring and clinical decision making. Curr Opin Biomed Eng 2025;36:100628. [doi: 10.1016/j.cobme.2025.100628]
14. Keshet A, Reicher L, Bar N, Segal E. Wearable and digital devices to monitor and treat metabolic diseases. Nature metabolism 2023;5(4):563-571. [doi: 10.1038/s42255-023-00778-y]
15. Zaman Q, others. The role of artificial intelligence in early disease detection: transforming diagnostics and treatment. Multidiscip J Healthc (MJH) 2024:1-54 [FREE Full text]
16. Cote AC, Phelps RJ, Kabiri NS, Bhangu JS, Thomas. Evaluation of wearable technology in dementia: a systematic review and meta-analysis. Front Med (Lausanne) 2021;7:501104 [FREE Full text] [doi: 10.3389/fmed.2020.501104] [Medline: 33505979]
17. Weizman Y, Tirosh O, Beh J, Fuss FK, Pedell S. Gait assessment using wearable sensor-based devices in people living with dementia: a systematic review. Int J Environ Res Public Health 2021;18(23):12735. [doi: 10.3390/ijerph182312735] [Medline: 34886459]
18. Iso-Markku P, Aaltonen S, Kujala UM, Halme H, Phipps D, Knittle K. Physical activity and cognitive decline among older adults: a systematic review and meta-analysis. JAMA Netw Open 2024;7(2):e2354285. [doi: 10.1001/jamanetworkopen.2023.54285] [Medline: 38300618]
19. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
20. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
21. Ma LL, Wang YY, Yang ZH, Huang D, Weng H, Zeng X. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? Mil Med Res 2020;7(1):7. [doi: 10.1186/s40779-020-00238-8] [Medline: 32111253]
22. Lentzen M, Vairavan S, Muurling M, Alepopoulos V, Atreya A, Boada M. RADAR-AD: assessment of multiple remote monitoring technologies for early detection of Alzheimer's disease. Alzheimers Res Ther 2025;17(1):29. [doi: 10.1186/s13195-025-01675-0] [Medline: 39865315]
23. Jones CD, Wasilko R, Zhang G, Stone KL, Gujral S, Rodakowski J. Detecting sleep/wake rhythm disruption related to cognition in older adults with and without mild cognitive impairment using the myRhythmWatch platform: feasibility and correlation study. JMIR Aging 2025;8:e67294-e67294. [doi: 10.2196/67294] [Medline: 40194310]
24. Satomi E, Apolinário D, Magaldi RM, Busse AL, Vieira Gomes GC, Ribeiro E. Beyond sleep: rest and activity rhythm as a marker of preclinical and mild dementia in older adults with less education. Neurobiol Sleep Circadian Rhythms 2025;18:100110 [FREE Full text] [doi: 10.1016/j.nbscr.2024.100110] [Medline: 39834590]
25. Sauers SC, Toedebusch CD, Richardson R, Spira AP, Morris JC, Holtzman DM. Midpoint of sleep is associated with sleep quality in older adults with and without symptomatic Alzheimer's disease. Sleep Adv 2024;5(1):zpae023. [doi: 10.1093/sleepadvances/zpae023] [Medline: 38711547]
26. Baril AA, Picard C, Labonté A, Sanchez E, Duclos C, Mohammediyan B. Day-to-day sleep variability with Alzheimer's biomarkers in at-risk elderly. Alzheimers Dement Diagn Assess Dis Monit 2024;16(1):e12521 [FREE Full text] [doi: 10.1002/dad2.12521] [Medline: 38371359]
27. Seo K, Takayanagi N, Sudo M, Yamashiro Y, Chiba I, Makino K. Association between daily gait speed patterns and cognitive impairment in community-dwelling older adults. Sci Rep 2023;13(1):2783. [doi: 10.1038/s41598-023-29805-4] [Medline: 36797381]
28. Huang B, Liao G, Lee PMY, Chan CK, Tai L, Tsang CYJ. Association of circadian rhythm with mild cognitive impairment among male pneumoconiosis workers in Hong Kong: a cross-sectional study. Sci Rep 2023;13(1):1650. [doi: 10.1038/s41598-023-28832-5] [Medline: 36717700]
29. Palmer JR, Wang C, Kong D, Cespedes M, Pye J, Hickie IB. Rest-activity rhythms and tract specific white matter lesions in older adults at risk for cognitive decline. Mol Psychiatry 2022;27(8):3410-3416 [FREE Full text] [doi: 10.1038/s41380-022-01641-4] [Medline: 35764707]
30. Lim J. A smart healthcare-based system for classification of dementia using deep learning. Digit Health 2022;8:20552076221131667 [FREE Full text] [doi: 10.1177/20552076221131667] [Medline: 36312848]

XSL•FO
RenderX

31. Boa Sorte Silva NC, Falck RS, Chan PCY, Tai D, Backhouse D, Stein R. The association of sleep and cortical thickness in mild cognitive impairment. Experimental Gerontology 2022;167:111923. [doi: 10.1016/j.exger.2022.111923]
32. Kim M, Liotta EM, Maas MB, Braun RI, Garcia-Canga B, Ganger DR. Rest-activity rhythm disturbance in liver cirrhosis and association with cognitive impairment. Sleep 2021;44(6):zsaa288. [doi: 10.1093/sleep/zsaa288] [Medline: 33367862]
33. Basta M, Zaganas I, Simos P, Koutentaki E, Dimovasili C, Mathioudakis L. Apolipoprotein E ε4 allele is associated with long sleep duration among elderly with cognitive impairment. J Alzheimers Dis 2021;79(2):763-771. [doi: 10.3233/jad-200958]
34. Roh HW, Choi J, Kim NR, Choe YS, Choi JW, Cho SM. Associations of rest-activity patterns with amyloid burden, medial temporal lobe atrophy, and cognitive impairment. EBioMedicine 2020;58:102881. [doi: 10.1016/j.ebiom.2020.102881] [Medline: 32736306]
35. Liu Z, Zhang L, Wu J, Zheng Z, Gao J, Lin Y. Machine learning-based classification of circadian rhythm characteristics for mild cognitive impairment in the elderly. Front Public Health 2022;10:1036886. [doi: 10.3389/fpubh.2022.1036886] [Medline: 36388285]
36. Espinosa N, Hoyos CM, McKinnon AC, Almgren H, Duffy SL, Naismith SL. Rest-activity rhythm fragmentation and synchronization are linked with reduced cortical thickness in older adults "at risk" for dementia. Sleep 2025;48(5):zsaf017. [doi: 10.1093/sleep/zsaf017] [Medline: 40052961]
37. Manning CA, Youngkin A, Quigg M. Sleepiness, sleep time, and depression of caregivers are linked with sleep and behaviors of their paired partners with dementia. J Alzheimers Dis Rep 2025;9:25424823241300981 [FREE Full text] [doi: 10.1177/25424823241300981] [Medline: 40034516]
38. Ghosal R, Varma VR, Volfson D, Urbanek J, Hausdorff JM, Watts A. Scalar on time-by-distribution regression and its application for modelling associations between daily-living physical activity and cognitive functions in Alzheimer's Disease. Sci Rep 2022;12(1):11558. [doi: 10.1038/s41598-022-15528-5] [Medline: 35798763]
39. Park B, Choi BJ, Lee H, Jang JH, Roh HW, Kim EY. Modeling brain volume using deep learning-based physical activity features in patients with dementia. Front Neuroinform 2022;16:795171 [FREE Full text] [doi: 10.3389/fninf.2022.795171] [Medline: 35356447]
40. Palmer JR, Duffy SL, Meares S, Pye J, Calamante F, Cespedes M. Rest-activity functioning is related to white matter microarchitecture and modifiable risk factors in older adults at-risk for dementia. Sleep 2021;44(7):zsab007. [doi: 10.1093/sleep/zsab007] [Medline: 33428761]
41. Wei J, Boger J. Sleep detection for younger adults, healthy older adults, and older adults living with dementia using wrist temperature and actigraphy: prototype testing and case study analysis. JMIR Mhealth Uhealth 2021;9(6):e26462 [FREE Full text] [doi: 10.2196/26462] [Medline: 34061038]
42. Corbi A, Burgos D. Connection between sleeping patterns and cognitive deterioration in women with alzheimer's disease. Sleep Breath 2022;26(1):361-371. [doi: 10.1007/s11325-021-02327-x] [Medline: 33792886]
43. Baril A, Picard C, Labonté A, Sanchez E, Duclos C, Mohammediyan B. Longer sleep duration and neuroinflammation in at-risk elderly with a parental history of Alzheimer's disease. Sleep 2024;47(6):zsae081 [FREE Full text] [doi: 10.1093/sleep/zsae081] [Medline: 38526098]
44. Sun H, Li P, Gao L, Yang J, Yu L, Buchman AS. Altered motor activity patterns within 10-minute timescale predict incident clinical Alzheimer's disease. J Alzheimers Dis 2024;98(1):209-220 [FREE Full text] [doi: 10.3233/JAD-230928] [Medline: 38393904]
45. Winer JR, Lok R, Weed L, He Z, Poston KL, Mormino EC. Impaired 24-h activity patterns are associated with an increased risk of Alzheimer's disease, Parkinson's disease, and cognitive decline. Alzheimers Res Ther 2024;16(1):35 [FREE Full text] [doi: 10.1186/s13195-024-01411-0] [Medline: 38355598]
46. Haghayegh S, Gao C, Sugg E, Zheng X, Yang HW, Saxena R. Association of rest-activity rhythm and risk of developing dementia or mild cognitive impairment in the middle-aged and older population: prospective cohort study. JMIR Public Health Surveill 2024;10:e55211 [FREE Full text] [doi: 10.2196/55211] [Medline: 38713911]
47. Basta M, Bouloukaki I, Skourti E, Zampetakis A, Alexopoulou C, Ganiaris A. Long objective sleep duration is a marker of cognitive impairment in older adults: findings from the Cretan Aging Cohort. J Alzheimers Dis Rep 2024;8(1):927-934 [FREE Full text] [doi: 10.3233/ADR-230203] [Medline: 38910938]
48. Milton S, Cavaillès C, Ancoli-Israel S, Stone KL, Yaffe K, Leng Y. Five-year changes in 24-hour sleep-wake activity and dementia risk in oldest old women. Neurology 2025;104(8):e213403. [doi: 10.1212/WNL.0000000000213403] [Medline: 40106755]
49. Skourti E, Simos P, Zampetakis A, Koutentaki E, Zaganas I, Alexopoulou C. Long-term associations between objective sleep quality and quantity and verbal memory performance in normal cognition and mild cognitive impairment. Front Neurosci 2023;17:1265016. [doi: 10.3389/fnins.2023.1265016] [Medline: 37928739]
50. Gao L, Li P, Gaykova N, Zheng X, Gao C, Lane JM. Circadian rest-activity rhythms, delirium risk, and progression to dementia. Ann Neurol 2023;93(6):1145-1157 [FREE Full text] [doi: 10.1002/ana.26617] [Medline: 36808743]
51. Jeon SY, Byun MS, Yi D, Jung G, Lee J, Kim YK. Circadian rest-activity rhythm and longitudinal brain changes underlying late-life cognitive decline. Psychiatry Clin Neurosci 2023;77(4):205-212 [FREE Full text] [doi: 10.1111/pcn.13521] [Medline: 36527292]

XSL•FO
RenderX

52. Yi Lee PM, Ling Kwok BH, Ting Ma JY, Tse LA. A population-based prospective study on rest-activity rhythm and mild cognitive impairment among Hong Kong healthy community-dwelling older adults. Neurobiol Sleep Circadian Rhythms 2021;10:100065 [FREE Full text] [doi: 10.1016/j.nbscr.2021.100065] [Medline: 33997474]
53. Lysen TS, Luik AI, Ikram MK, Tiemeier H, Ikram MA. Actigraphy-estimated sleep and 24-hour activity rhythms and the risk of dementia. Alzheimers Dement 2020;16(9):1259-1267 [FREE Full text] [doi: 10.1002/alz.12122] [Medline: 32558256]
54. Agudelo C, Tarraf W, Wu B, Wallace DM, Patel SR, Redline S. Actigraphic sleep patterns and cognitive decline in the hispanic community health study/study of latinos. Alzheimers Dement 2021;17(6):959-968. [doi: 10.1002/alz.12250] [Medline: 33350583]
55. Ning Y, Chen M, Yang H, Jia J. Accelerometer-derived "weekend warrior" physical activity, sedentary behavior, and risk of dementia. Alzheimers Res Ther 2025;17(1):67 [FREE Full text] [doi: 10.1186/s13195-024-01657-8] [Medline: 40121491]
56. Ning Y, Chen M, Yang H, Jia J. Accelerometer-measured physical activity timing with incident dementia. Alzheimers Dement 2025;21(2):e14452. [doi: 10.1002/alz.14452] [Medline: 39868637]
57. Zhao B, Zhou S, Chang J, Li A, Geng C, Wei T. Association between weekend recovery sleep and risk of incident dementia: a prospective cohort study in the UK Biobank. J Neurol 2025;272(9):612. [doi: 10.1007/s00415-025-13363-y] [Medline: 40911069]
58. Lu Z, Leung JCS, Feng H, Zhang J, Wing YK, Kwok TCY. Circadian rest-activity rhythms and cognitive decline and impairment in older Chinese adults: a multicohort study with prospective follow-up. Arch Gerontol Geriatr 2023;116:105215. [doi: 10.1016/j.archger.2023.105215] [Medline: 39491073]
59. Chan LLY, Espinoza Cerda MT, Brodie MA, Lord SR, Taylor ME. Daily-life walking speed, running duration and bedtime from wrist-worn sensors predict incident dementia: a watch walk - UK biobank study. Int Psychogeriatr 2025;37(3):100031 [FREE Full text] [doi: 10.1016/j.inpsyc.2024.100031] [Medline: 39827009]
60. Shi C, Babiker N, Urbanek JK, Grossman RL, Huisingh-Scheetz M, Rzhetsky A. Free-living wrist and hip accelerometry forecast cognitive decline among older adults without dementia over 1- or 5-years in two distinct observational cohorts. NPJ Aging 2022;8(1):7. [doi: 10.1038/s41514-022-00087-w] [Medline: 35927250]
61. Xiao Q, Shadyab AH, Rapp SR, Stone KL, Yaffe K, Sampson JN. Rest-activity rhythms and cognitive impairment and dementia in older women: results from the women's health initiative. J Am Geriatr Soc 2022;70(10):2925-2937 [FREE Full text] [doi: 10.1111/jgs.17926] [Medline: 35708069]
62. Hoepel SJW, Ikram MK, Ikram MA, Voortman T, Luik AI. Substitutions of sleep, physical activity, and sedentary behaviour and the risk of dementia and stroke. J Sleep Res 2026;35(1):e70166. [doi: 10.1111/jsr.70166] [Medline: 40793962]
63. Cho E, Kim S, Heo S, Shin J, Hwang S, Kwon E. Machine learning-based predictive models for the occurrence of behavioral and psychological symptoms of dementia: model development and validation. Sci Rep 2023;13(1):8073 [FREE Full text] [doi: 10.1038/s41598-023-35194-5] [Medline: 37202454]
64. Plotogea OM, Diaconu CC, Gheorghe G, Stan-Ilie M, Badea M, Prelipcean CC. The prevalence and association of cognitive impairment with sleep disturbances in patients with chronic liver disease. Brain Sci 2022;12(4):444. [doi: 10.3390/brainsci12040444] [Medline: 35447976]
65. Targa ADS, Benítez ID, Dakterzada F, Fontenele-Araujo J, Minguez O, Zetterberg H. The circadian rest-activity pattern predicts cognitive decline among mild-moderate alzheimer's disease patients. Alzheimers Res Ther 2021;13(1):161 [FREE Full text] [doi: 10.1186/s13195-021-00903-7] [Medline: 34563258]
66. Cho E, Kim S, Hwang S, Kwon E, Heo S, Lee JH, et al. Factors associated with behavioral and psychological symptoms of dementia: prospective observational study using actigraphy. J Med Internet Res 2021;23(10):e29001. [doi: 10.2196/29001] [Medline: 34714244]
67. Rykov YG, Patterson MD, Gangwar BA, Jabar SB, Leonardo J, Ng KP. Predicting cognitive scores from wearable-based digital physiological features using machine learning: data from a clinical trial in mild cognitive impairment. BMC Med 2024;22(1):36 [FREE Full text] [doi: 10.1186/s12916-024-03252-y] [Medline: 38273340]
68. David S, Hohenfeld C, Haberl L, Pahl J, Costa AS, Kilders A. Physical activity monitoring in Alzheimer's disease during sport interventions: a multi-methodological perspective. Front Neurol 2023;14:1195694 [FREE Full text] [doi: 10.3389/fneur.2023.1195694] [Medline: 37808485]
69. Khosroazad S, Gilbert CF, Aronis JB, Daigle KM, Esfahani M, Almaghasilah A. Sleep movements and respiratory coupling as a biobehavioral metric for early Alzheimer's disease in independently dwelling adults. BMC Geriatr 2023;23(1):252 [FREE Full text] [doi: 10.1186/s12877-023-03983-2] [Medline: 37106470]
70. Hossain M, Uddin M, Thapa K, Al Zubaer MA, Islam M, Lee J. Detecting cognitive impairment status using keystroke patterns and physical activity data among the older adults: a machine learning approach. J Healthc Eng 2021;2021:1302989. [doi: 10.1155/2021/1302989] [Medline: 34966518]
71. Kallogjeri D, Spitznagel EL, Piccirillo JF. Importance of defining and interpreting a clinically meaningful difference in clinical research. JAMA Otolaryngol Head Neck Surg 2020;146(2):101-102. [doi: 10.1001/jamaoto.2019.3744] [Medline: 31804662]
72. Rykov YG, Patterson MD, Gangwar BA, Jabar SB, Leonardo J, Ng KP. Predicting cognitive scores from wearable-based digital physiological features using machine learning: data from a clinical trial in mild cognitive impairment. BMC Med 2024;22(1):36 [FREE Full text] [doi: 10.1186/s12916-024-03252-y] [Medline: 38273340]

XSL•FO
RenderX

73. Berger A, Wielgus KK, Young-McCaughan S, Fischer P, Farr L, Lee KA. Methodological challenges when using actigraphy in research. J Pain Symptom Manage 2008;36(2):191-199 [FREE Full text] [doi: 10.1016/j.jpainsymman.2007.10.008] [Medline: 18400460]
74. Patterson MR, Nunes AAS, Gerstel D, Pilkar R, Guthrie T, Neishabouri A. 40 years of actigraphy in sleep medicine and current state of the art algorithms. NPJ Digit Med 2023;6(1):51 [FREE Full text] [doi: 10.1038/s41746-023-00802-1] [Medline: 36964203]
75. Morris AC, Seker A, Telesia L, Wickersham A, Ching BC, Roy R. Adherence to actigraphic devices in elementary school-aged children: systematic review and meta-analysis. J Med Internet Res 2025;27:e79718 [FREE Full text] [doi: 10.2196/79718] [Medline: 41183377]
76. Dobson R, Stowell M, Warren J, Tane T, Ni L, Gu Y. Use of consumer wearables in health research: issues and considerations. J Med Internet Res 2023;25:e52444 [FREE Full text] [doi: 10.2196/52444] [Medline: 37988147]
77. D'Ignazio G, Carlucci L, Sergi MR, Palumbo R, Dattilo L, Terrei M. Is the MMSE enough for MCI? A narrative review of the usefulness of the MMSE. Front Psychol 2025;16:1727738 [FREE Full text] [doi: 10.3389/fpsyg.2025.1727738] [Medline: 41356018]
78. Gu B, Kim HS, Kim H, Yoo J. Advancements in wearable sensor technologies for health monitoring in terms of clinical applications, rehabilitation, and disease risk assessment: systematic review. JMIR Mhealth Uhealth 2026;14:e76084. [doi: 10.2196/76084] [Medline: 41511829]
79. Kainec KA, Caccavaro J, Barnes M, Hoff C, Berlin A, Spencer RMC. Evaluating accuracy in five commercial sleep-tracking devices compared to research-grade actigraphy and polysomnography. Sensors (Basel) 2024;24(2):635 [FREE Full text] [doi: 10.3390/s24020635] [Medline: 38276327]
80. Smith MT, McCrae CS, Cheung J, Martin JL, Harrod CG, Heald JL, et al. Use of actigraphy for the evaluation of sleep disorders and circadian rhythm sleep-wake disorders: an American academy of sleep medicine systematic review, meta-analysis, and GRADE assessment. J Clin Sleep Med 2018;14(7):1209-1230. [doi: 10.5664/jcsm.7228] [Medline: 29991438]
81. Kahawage P, Jumabhoy R, Hamill K, de Zambotti M, Drummond SPA. Validity, potential clinical utility, and comparison of consumer and research-grade activity trackers in Insomnia Disorder I: in-lab validation against polysomnography. J Sleep Res 2020;29(1):e12931. [doi: 10.1111/jsr.12931] [Medline: 31626361]
82. Huhn S, Axt M, Gunga HC, Maggioni MA, Munga S, Obor D. The impact of wearable technologies in health research: scoping review. JMIR Mhealth Uhealth 2022;10(1):e34384 [FREE Full text] [doi: 10.2196/34384] [Medline: 35076409]
83. Ahmed S, Nielsen IE, Tripathi A, Siddiqui S, Ramachandran RP, Rasool G. Transformers in time-series analysis: a tutorial. Circuits Syst Signal Process 2023;42(12):7433-7466. [doi: 10.1007/s00034-023-02454-8]
84. Kim Y, Xu X, McDuff D, Breazeal C, Park HW. Healthllm: large language models for health prediction via wearable sensor data. ArXiv Preprint posted online on January 12, 2024. [doi: 10.48550/arXiv.2401.06866]
85. Saraswat D, Bhattacharya P, Verma A, Prasad VK, Tanwar S, Sharma G. Explainable AI for healthcare 5.0: opportunities and challenges. IEEE Access 2022;10:84486-84517. [doi: 10.1109/access.2022.3197671]
86. Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. Lancet 2020;396(10248):413-446 [FREE Full text] [doi: 10.1016/S0140-6736(20)30367-6] [Medline: 32738937]
87. Kivipelto M, Mangialasche F, Ngandu T. Lifestyle interventions to prevent cognitive impairment, dementia and Alzheimer disease. Nat Rev Neurol 2018;14(11):653-666. [doi: 10.1038/s41582-018-0070-3] [Medline: 30291317]
88. Bufano P, Laurino M, Said S, Tognetti A, Menicucci D. Digital phenotyping for monitoring mental disorders: systematic review. J Med Internet Res 2023;25:e46778 [FREE Full text] [doi: 10.2196/46778] [Medline: 38090800]

Abbreviations

AI: artificial intelligence
ICD-10: International Statistical Classification of Diseases, Tenth Revision
MCI: mild cognitive impairment
MeSH: Medical Subject Headings
MMSE: Mini-Mental State Examination
MoCA: Montreal Cognitive Assessment
OR: odds ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
XGBoost: Extreme Gradient Boosting

XSL•FO
RenderX

Edited by S Brini; submitted 22.Oct.2025; peer-reviewed by B Howell, S Ajayi, S Ye; comments to author 27.Nov.2025; revised version received 23.Jan.2026; accepted 26.Jan.2026; published 23.Feb.2026.

Please cite as:
Cejudo A, Arrojo M, Martín C, Almeida A
AI and Wearables for Early Detection of Cognitive Impairment and Dementia: Systematic Review
J Med Internet Res 2026;28:e86262
URL: https://www.jmir.org/2026/1/e86262
doi:10.2196/86262
PMID:

©Ander Cejudo, Markel Arrojo, Cristina Martín, Aitor Almeida. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Design Practices for Data Dashboards in Health Care: Scoping Review

Heike Vornhagen[1*], MA, PhD; Stephen Barrett[2*], BSc, MSc; Ciara Carroll[2], BCh, BAO, MB; Lydia Kavochi Iladiva[3], MBBCHIR, MPH, MBA; Gregory Martin[2], MBBCHIR, MBA, MPH; Declan McKeown[2], MB, BCh, BAO, MPH; Jennifer Martin[2], MB, BCh, BAO, MPH

[1]Insight Research Ireland Centre for Data Analytics, Data Science Institute, Ollscoil na Gaillimhe – University of Galway, Galway, Ireland

[2]Health Service Executive, Dublin, Ireland

[3]Department of Epidemiology & Public Health, University College Cork, Cork, Ireland

[*]these authors contributed equally

Corresponding Author:
Heike Vornhagen, MA, PhD
Insight Research Ireland Centre for Data Analytics
Data Science Institute
Ollscoil na Gaillimhe – University of Galway
IDA Business Park
Lower Dangan
Galway, H91 AEX4
Ireland
Phone: 353 91 495053
Email: heike.vornhagen@universityofgalway.ie

Abstract

Background: Health care dashboards have the potential to enhance understanding, decision-making, and communication. However, their design, implementation, and evaluation are often hindered by the absence of standardized guidelines. This scoping review synthesizes international evidence to identify common practices for health care dashboard design, providing a foundation for application in the Irish context.

Objective: This study aimed to identify existing guidelines and common practices for health care dashboard design to inform future development and implementation within the Irish health care system.

Methods: A scoping review using an evidence summary approach was conducted. PubMed, Embase, Scopus, and IEEE Xplore (2014-2024) were searched. Practices were extracted and analyzed using reflexive thematic analysis and then grouped into 4 main pillars: approach (engagement of end users and stakeholders), content (data quality, effective insights, and presentation), behavior (usability and accessibility), and adoption (sustainability).

Results: From 1644 initially identified studies, 18 (1.1%) met the inclusion criteria. Most were hospital focused (13/18, 72.2%), with few community- or public-facing dashboards. Only 4 of 18 (22.2%) studies described structured guidelines; most implementations (14/18, 77.8%) were ad hoc. Common practices included user involvement, actionable metrics, data quality, usability, and workflow integration. Divergences were observed: hospitals prioritized clinical indicators, public dashboards emphasized transparency, and community dashboards were underrepresented. Conflicting findings included debate over interactivity vs static simplicity.

Conclusions: Dashboard design remains fragmented, with limited guidance for structured design or implementation. The 4 pillars provide a practical synthesis of best practices, highlighting pathways for evidence-informed, user-centered design. These pillars will inform future consensus building and co-design of health care dashboards in Ireland and can serve as a foundation for broader application in primary care, community, and public health settings.

(J Med Internet Res 2026;28:e77361) doi:10.2196/77361

KEYWORDS

data visualization; dashboard design; health care analytics; evidence-based design; health informatics

XSL•FO
RenderX

Introduction

This review explores how dashboards are developed within the health care sector and which guidelines, if any, steer such development. Drawing insights from international research, this review ultimately aims to identify best practices and demonstrate how visual tools can enhance understanding, decision-making, and communication.

Dashboards are electronic displays that usually combine multiple visualizations to enable users to simultaneously access and interact with information on related topics [1]. Ideally, dashboards should support rapid understanding and reduce cognitive load, a key consideration in high-stress health care environments where time-sensitive decisions are paramount [2]. As visual dashboards can significantly improve decision-making [3], they should enable health care professionals to efficiently and effectively identify care challenges and progress. Health organizations use dashboards for different purposes, such as management and oversight, as well as for monitoring and improving direct patient care. Dashboards can also help improve teamwork and communication among people with different roles, expertise, and schedules [1].

Data visualization refers to the process of converting complex datasets into visual formats, such as charts and graphs. By facilitating clearer and more immediate insights, effective data visualizations have the potential to significantly improve information sharing among health care professionals, leading to enhanced collaborative decision-making [4]. This is supported by Lurie and Mason [5], who highlight that appropriately designed data visualizations not only improve comprehension but also enhance retention of critical information. Similarly, international studies have found that well-designed data visualization tools can reduce medical errors and support more accurate clinical decisions, particularly when used to communicate complex patient data [6].

Several frameworks for data visualization and dashboard design exist outside of health care and provide useful starting points. The nested model by Munzner [7] describes levels of design decision-making, from domain problems to visual encoding, emphasizing alignment between user tasks and visualization choices. The dashboard design principles by Few [8] offer practical guidance on clarity, simplicity, and information density, primarily targeted at business contexts. Human-computer interaction (HCI) standards, such as ISO 9241-210 [9], similarly emphasize usability, accessibility, and iterative user-centered design. These frameworks are valuable, but their application to health care contexts is limited. They often neglect the complexities of health care data, such as interoperability, sensitive patient information, and the unique cognitive demands of clinical environments.

Previous reviews of health care dashboards have advanced the field but remain constrained in several ways. For example, a scoping review by Helminski et al [10] found that few included studies used usability testing or formal theoretical frameworks for design or evaluation. The systematic review of hospital dashboards by Rabiei and Almasi [11] identified both functional and nonfunctional requirements but reported that many studies lacked empirical evaluation and had limited detail on implementation or integration with existing systems. As a result, existing evidence has not yet been consolidated into a structured, theoretically grounded framework that can guide the design of health care dashboards across settings.

To address this gap, we undertook an international scoping review of health care dashboard literature. Our aim was not only to map and catalog existing practices but also to organize them into a cohesive framework that bridges theory and practice. Building on previous work, our contribution is to consolidate existing evidence into 4 core pillars of dashboard design (approach, content, behavior, and adoption), which provide a foundation for more consistent, evidence-informed dashboard development in health care while remaining adaptable to diverse contexts.

Methods

Study Design

This study was conducted as a scoping review, following the framework proposed by Arksey and O'Malley [12] and enhanced by Levac et al [13]. Scoping reviews are particularly suited to mapping the existing literature in areas where guidelines are not well established, identifying key concepts, and synthesizing recurring practices. Our objective was to explore the international literature on health care dashboard design and identify common practices. As such, we did not perform formal quality appraisals, which is consistent with scoping review methodology. We report this review in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist [14] (Multimedia Appendix 1).

Needs Assessment

The initial proposal emerged from discussions between health professionals working on health service improvement initiatives and a data visualization expert. On the basis of a need to develop regional population profiles, a lack of clarity on effective data presentation was identified as an area that needed to be addressed.

Identify the Research Question

The objective of this scoping review was to explore current international health care literature to identify existing guidelines or, in the absence of such guidelines, to identify common practices for health care dashboard design. The following research questions guided the search strategy: (1) What dashboards exist within (national) health systems that support decision-making? (2) Are there good practice guidelines for the creation of such dashboards? and (3) If not, are the design practices sufficiently described to enable the creation of dashboard design guidelines?

Proposal Development and Approval

The questions and the overall time plan were discussed with the project team. Concurrently, discussions with other stakeholders involved in the development of health dashboards

XSL•FO
RenderX

allowed the team to gather feedback regarding the potential impact of developing an evidence summary.

Systematic Literature Search

We conducted an electronic search using PubMed, Embase, Scopus, and IEEE Xplore. The search strategy combined terms for dashboards and data visualization (eg, dashboard, data visualization, and visual analytics) with terms for health care and health systems (eg, health care and health infrastructure). Boolean operators and truncation were used to capture variations. Controlled vocabulary (eg, Medical Subject Headings [MeSH] terms such as data visualization and decision support systems) was included where applicable. These keywords and associated terms were based on the research questions and manual searches for papers about the topic. Searches were limited to English-language articles published between January 2014 and April 2024. The complete reproducible search strings for each database are provided in Multimedia Appendix 2.

Screen and Select Relevant Literature

On completion of the electronic searches, the results were imported into Microsoft Excel. After removal of duplicates, titles and abstracts were screened by 2 independent reviewers. Conflicts or disagreements about selections were resolved by a third independent reviewer.

Eligibility criteria were defined to ensure that the review captured evidence most relevant to contemporary health care dashboard design. We limited the search to publications from 2014 onward, reflecting significant advances in data visualization technologies, health care information systems, and the use of dashboards in health service decision-making over the past decade. Only English-language publications were included due to resource constraints for translation. We included peer-reviewed research papers, systematic reviews, and dissertations or theses, as these sources often provide detailed methodological and design insights. In contrast, conference abstracts, commentaries, and opinion papers were excluded because they generally provide insufficient methodological detail or practical guidance. The specific inclusion and exclusion criteria are summarized in Table 1.

Table 1. Eligibility criteria used for the screening of studies.

	Inclusion criteria	Exclusion criteria
Population	Data visualization and data dashboards	Clinical visualizations and clinical visual aids
Context	All health care settings and any country or health system	None
Types of evidence sources	Research papers, systematic reviews, and dissertations or theses (including empirical findings and author-reported practices)	Conference abstracts and commentaries or opinion papers
Languages	English	Other than English
Time	2014 to 2024	Publications before 2014

Consistent with scoping review methodology and PRISMA-ScR guidance, no formal risk of bias or quality appraisal of included studies was undertaken, as our aim was to map the characteristics and range of existing evidence rather than to assess study validity.

After the initial screening of titles and abstracts, the selected studies were reviewed in full by 2 authors, and any disagreements were resolved by a third reviewer.

A data charting table summarizing key characteristics of the included studies is provided in Multimedia Appendix 3. The table also indicates the extent to which each study described guidelines and specifies which sections of the paper were used for reflexive thematic analysis (RTA) [15].

Identify Pillars From Included Studies

After the full-text screening was complete, the guidelines and practices described were extracted into an Excel sheet. Using RTA, the first author (HV) coded each described practice with a descriptive keyword (eg, *user engagement* or *data quality*) and iteratively grouped related codes into broader conceptual clusters. These clusters were reviewed and discussed by all team members (HV, SB, and CC), leading to the identification of four overarching themes (pillars):

1. Approach—the *methods* and *mindset* that shape the design process
2. Content—the *type of information* that should be provided and displayed
3. Behavior—the *manner of interaction* with the information (access, user-friendliness, and navigation)
4. Adoption—the *practices* that support adoption and ongoing engagement

Multimedia Appendix 4 lists the second-level keywords that informed each pillar. The analysis was conducted in line with the interpretive and reflexive orientation of RTA, emphasizing depth of understanding over coder agreement or reproducibility.

Produce Reports

This paper presents the first report arising from this search. Future papers will detail how the identified pillars and associated practices have been contextualized in the Irish context and describe the overall impact of this initiative.

Ongoing Follow-Up and Dialogue

The pillars and associated practices will inform a series of workshops with users and stakeholders to communicate and contextualize the findings.

XSL•FO
RenderX

Results

After removing duplicates, a total of 1644 potentially relevant reports were identified. Two reviewers (HV and SB) screened the abstracts of the selected papers, while a third reviewer (CC) resolved disagreements between the 2 reviewers. A total of 1471 papers were excluded, while 173 were selected for full-text review. A citation search of the selected papers led to a further 39 records, which were screened as outlined earlier. This added another 13 papers to the 173 papers to be included in the full-text review. After assessing against the eligibility criteria, a total of 18 reports were included in the synthesis of evidence. The entire selection process is represented in a PRISMA-ScR flow diagram (Figure 1; full study characteristics and extracted data are provided in Multimedia Appendix 3) [16].

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension) flow diagram showing selection of relevant papers [16].

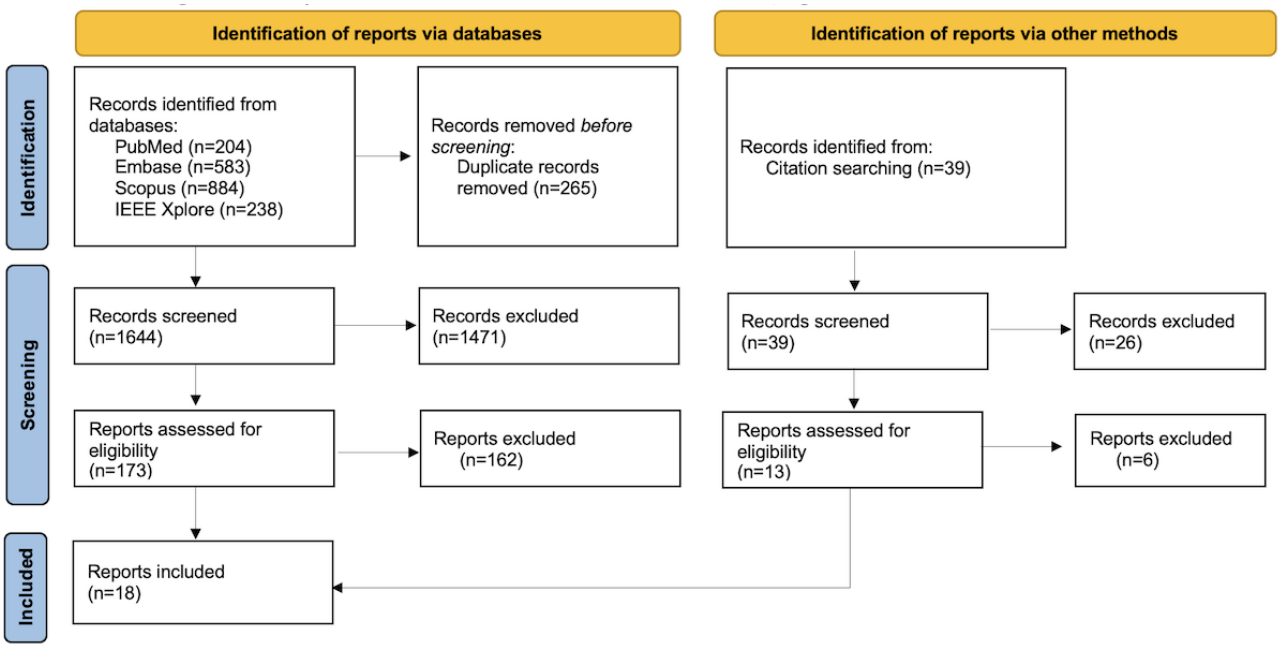

A total of 18 studies were included, with the majority focused on hospital settings (n=13, 72.2%). Few studies addressed national or public dashboards, and dashboards implemented at the community level were notably underrepresented, indicating a gap in evidence for nonhospital contexts. Two papers were review articles. Three papers focused on the development of national dashboards or the integration of a mobile app with a dashboard. Studies based in UK hospitals were the focus in 4 papers, while studies in the United States and the Netherlands were the setting for 2 papers each. The remainder included 1 study each in Australia, Iran, Portugal, Somalia, South Africa, South Korea, and Switzerland. Three papers did not specify the country in which the study was conducted.

In total, 4 papers [17-20] explicitly described comprehensive guidelines for dashboard development from start to finish, highlighting structured and replicable approaches, while an additional 9 papers [11,21-28] presented practices addressing specific aspects, requirements, or phases of dashboard design. The remaining 5 papers [29-33] either outlined specific design goals or described how dashboards were developed in practice but did not offer development practices that could inform broader or generalizable approaches to dashboard design. Overall, most implementations appeared ad hoc rather than following a structured framework, which may limit reproducibility and generalizability. Notably, half of the studies (9/18, 50%) focused primarily on evaluating dashboards after their creation, rather than explicitly detailing robust design principles during development.

Common practices identified across studies included user involvement in design and evaluation [17,28]; use of relevant and actionable metrics [25]; attention to data quality, including data accuracy and timeliness [30]; usability and interactivity [33]; and integration into existing workflows to avoid overburdening staff [20].

To illustrate, van de Baan et al [17] in the Netherlands developed quality and safety dashboards through a 5-stage participatory process with 13 health professionals across 2 care pathways. Their structured, clinician-led approach demonstrated how iterative co-design can enhance ownership and ensure dashboards address real-world needs (pillar: approach). Randell et al [25], drawing on National Health Service clinical audit data, highlighted challenges in selecting indicators and ensuring their accuracy, translating these into clear requirements for dashboards that can support decision-making “from ward to board” (pillar: content). Loorak et al [33] introduced TimeSpan, a dashboard for stroke care pathways, where interactivity enabled clinicians to generate new insights and explore patterns beyond static reporting (pillar: behavior). Finally, Weggelaar-Jansen et al [20] showed that sustainable dashboard adoption required embedding dashboards into governance routines and review meetings, not just technical development (pillar: adoption).

XSL•FO
RenderX

Contrasts and conflicts across settings were notable: hospital dashboards prioritized clinical indicators and immediate operational utility, whereas public dashboards emphasized transparency, accessibility, and stakeholder accountability. Community dashboards were minimally represented, suggesting that practical guidance for these contexts is lacking. Furthermore, divergences in design philosophy emerged: some studies emphasized highly interactive dashboards with rich visualizations, while others favored static simplicity to reduce cognitive load. Similarly, governance and sustainability considerations varied: certain implementations benefited from strong institutional buy-in, whereas others struggled to maintain dashboards over time. These differences highlight potential tensions between usability, engagement, and sustainability and underscore that no single approach fits all contexts.

Methodologically, studies differed in both their analytic rigor and their use of visualization techniques. Hospital-based dashboards typically adopted iterative or co-design methods, enabling clinical validation and rapid refinement of metrics [17,18], whereas national or public-facing dashboards tended to rely on top-down technical builds using aggregated data [26]. These methodological differences shaped the visualization choices observed: co-designed hospital dashboards favored targeted, high-resolution displays with limited indicator sets, while public health dashboards prioritized breadth, using simplified or static visuals to maximize accessibility. Studies using participatory design generally reported stronger usability and contextual fit, while purely technical implementations often struggled with user engagement or indicator relevance.

Context also influenced implementation and sustainability. Hospital dashboards were typically embedded within established governance structures, allowing integration into performance review meetings and continuous data validation [20]. In contrast, public dashboards faced challenges of data standardization, cross-agency coordination, and long-term ownership, which limited ongoing updates. Methodological variation, therefore, reflected contextual constraints: where clinical workflows supported routine data collection, interactive and iterative designs thrived; where data pipelines were diffuse, design flexibility was restricted. These contrasts illustrate that methodological rigor and stakeholder proximity are decisive factors in determining dashboard longevity and real-world impact.

More than half of the papers (11/18, 61.1%) identified evaluation, appropriate visual encoding [7,34], and user involvement as key to developing a dashboard. Other key practices included modularity (the ability to add or delete components) and data quality (including timeliness and accuracy). Interactivity and the need to avoid adding to staff's existing workload were mentioned by 6 papers (33.3%).

The secondary analysis (RTA) resulted in 4 key themes (pillars) being defined: "approach," "content," "behavior," and "adoption." Most practices supported the "approach" pillar (33 unique mentions) (Table 2), while practices for "behavior" were comparatively few (15 unique mentions; Figure 2).

Figure 2. Key terms grouped by pillar: approach, content, behaviour and adoption.

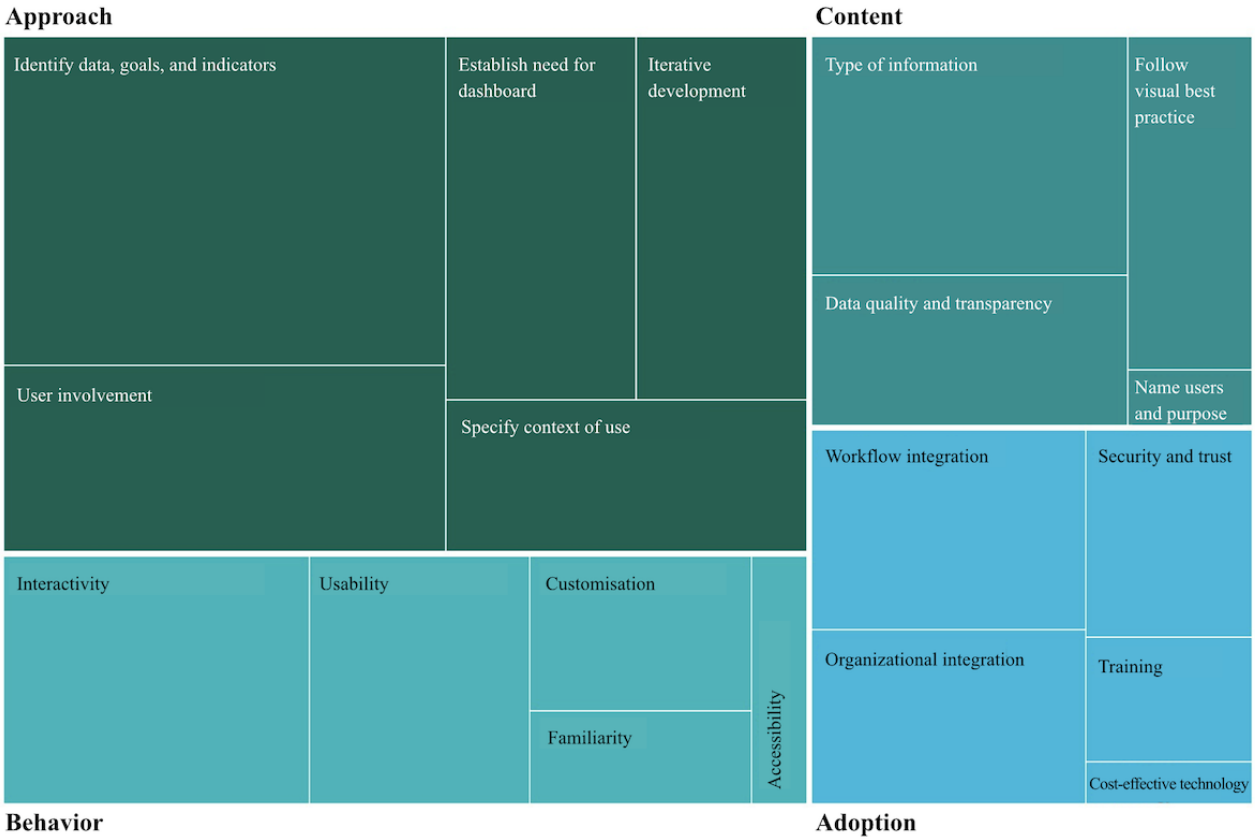

XSL•FO
RenderX

Table 2. Illustrative contributions of included studies to each pillar.

Pillar	Illustrative contributions (examples)
Approach	Participatory and iterative design approaches were reported in several studies [17,18,28], emphasizing co-design and ownership.
Content	Indicator selection and data quality were central concerns, as highlighted by Randell et al [25] on National Health Service clinical audits, Christen et al [32] on modular EMS dashboards, and Ivanković et al [26] on transparency in global COVID-19 dashboards.
Behavior	Usability and interactivity were highlighted in dashboards such as TimeSpan for stroke care [33], population health frameworks [16], and productivity dashboards [31].
Adoption	Sustainable uptake depended on workflow integration and governance structures, as described by Weggelaar-Jansen et al [20], Ahn et al [30], and Kamadjeu and Gathenji [27].

This finding underscores that while dashboard design is frequently addressed, critical attention to adoption, behavior change, and context-specific implementation is limited, suggesting gaps in translating design principles into practical impact.

Discussion

Overview

This review provides an insight into current practices in health care dashboard design. Our findings reveal a significant gap in the use of established guidelines or frameworks during the dashboard design process. While there is broad agreement on some core practices such as evaluation and user involvement, other areas such as data security or staff training are rarely mentioned. These gaps may arise from the reactive and ad hoc nature of dashboard implementation in response to immediate operational pressures, rather than from a lack of awareness of best practices. Moreover, our synthesis highlights how design priorities differ across health care contexts, reflecting the specific needs and constraints of each setting. Understanding these practices provides a foundation for mapping observed patterns to established visualization and HCI frameworks, which underpins the subsequent synthesis of the 4 pillars: approach, content, behavior, and adoption.

Anchoring the 4 Pillars in Established Theory

Theoretical Alignment With Visualization

Each of the 4 pillars aligns with recognized frameworks in visualization and HCI. Approach reflects principles of user-centered and participatory design, emphasizing stakeholder involvement and iterative cocreation (ISO 9241-210 [9]). Content is consistent with the nested model by Munzner [7] and the semiology by Bertin [34], underscoring the importance of selecting actionable metrics and applying effective visual encoding. Behavior draws on HCI principles of usability [35] and the concept of affordances by Norman [36], highlighting interactivity, customization, and accessibility. Finally, adoption resonates with sociotechnical theories such as the technology acceptance model [37] and normalization process theory [38]. By mapping inductively derived themes to these theories, the 4 pillars provide a bridge between conceptual models and applied health care practice, ensuring the framework is both empirically grounded and theoretically anchored.

These theoretical frameworks were not used to generate the pillars but were applied retrospectively as an interpretive lens to examine their coherence and alignment with established concepts in visualization, usability, and sociotechnical theory. This approach ensures that the 4 pillars remain inductively grounded while being conceptually validated against recognized theoretical models.

Pillar 1: Approach

This pillar includes the methods and mindsets that shape the dashboard design process. It focuses strongly on involving and engaging end users and stakeholders in defining needs, iterative cocreation, and a review of existing systems to ensure that the best solution is put forward for development.

Pillar 2: Content

This pillar describes the type of information a dashboard should provide and how it should be displayed. It includes the need for relevant and actionable metrics and an examination of data accuracy and quality. Content should be presented following visual best practices (eg, using percentages instead of raw numbers, eliminating visual noise, and highlighting key signals) to ensure clarity and support decision-making.

Pillar 3: Behavior

Behavior includes practices related to how users interact with the dashboard. This includes interactivity to support engagement and exploration, as well as usability and accessibility considerations to accommodate diverse levels of data literacy.

Pillar 4: Adoption

This pillar focuses on practices that support adoption both by immediate users and the wider organization, thereby ensuring longevity and sustainability. Practices such as the ability to customize the dashboard, workflow integration to minimize additional staff workloads, broader integration into the organizational culture through stakeholder engagement, and continuous improvements help to ensure that the dashboard becomes part of daily operations. Attention to cost-effective tools, data security, privacy, and continuous evaluation further supports sustained adoption.

Principal Findings

A key insight from our review is that 50% (9/18) of studies focused on evaluating dashboards after their creation rather than incorporating robust design principles during development. Evaluation efforts often highlighted issues with usability, data integration, and user engagement, underscoring the importance of proactive design strategies to address these challenges upfront. Ad hoc development and inconsistent application of design principles were common, contributing to persistent gaps

XSL•FO
RenderX

in effective dashboard implementation. Adoption-related factors, such as governance, training, and sustainability, were rarely addressed, despite their known importance for long-term use.

Our analysis also demonstrates that context shapes dashboard priorities: hospitals often emphasize operational and clinical metrics to support immediate decision-making, while publicly accessible dashboards tend to prioritize transparency and stakeholder communication. Community and population health dashboards were scarce, which may explain the limited guidance on integrating diverse data sources or engaging broader stakeholder groups. Conflicting design priorities were also noted: some studies emphasized highly interactive dashboards to enhance engagement, while others favored simpler, static dashboards to reduce cognitive load or resource requirements. Technological, governance, and sustainability challenges were frequently overlooked, particularly in nonhospital or community settings.

Furthermore, while concepts such as user involvement and iterative refinement were mentioned, they were not consistently applied, leaving gaps in achieving effective, user-centered dashboards.

Taken together, the 4 pillars summarize a practical set of considerations for dashboard design and adoption. They provide a lens for assessing both technical and sociotechnical aspects, from user-centered approaches and data quality to usability, governance, and sustainability. For practice, this highlights embedding dashboards into workflows with training and stakeholder engagement. For policy, investment is required in the sociotechnical infrastructure to ensure dashboards are effective and sustainable beyond immediate returns.

Strengths and Limitations

This review contributes to the understanding of best practices in health care dashboard design by synthesizing existing literature and identifying key themes. The integration of theory strengthens the framework and allows critical evaluation of divergences and gaps. Limitations include scope and selection bias, lack of empirical validation, and the overrepresentation of hospital settings. Future research should test the 4 pillars in underrepresented contexts, such as primary care, community health, and public dashboards, and evaluate impacts on usability, adoption, and decision-making.

Implications for Ireland

Although international in scope, these findings provide an evidence base for the next phase of work in the Irish health care system. The 4 pillars will serve as the starting point for a consensus building process with clinicians, public health professionals, service planners, policymakers, IT specialists, and patient or community representatives. This process will contextualize international practices to Irish data infrastructures, governance, and regulatory frameworks, ensuring that future dashboards are both evidence informed and locally relevant.

Future Work

This review represents an initial step toward establishing structured, evidence-based guidelines for health care dashboard design, synthesizing international best practices into the 4 pillars: approach, content, behavior, and adoption. Building on the findings of this international review, the next phase of this work focuses on consensus building and co-design within the Irish health care system. Specifically, the 4 pillars will provide the starting point for structured discussions with Irish stakeholders, ensuring that international evidence is explicitly adapted to local infrastructure, governance, and regulatory requirements. By engaging a diverse range of stakeholders—including public health professionals; service planners; clinicians; IT specialists; policymakers; and public, patient, and community partners—we will refine the practices under each of the 4pillars into actionable, context-specific guidelines. For example, the practice of ensuring data quality (pillar: content) will be mapped to current data sources within Ireland, their interoperability, and Ireland-specific data protection principles (eg, General Data Protection Regulation [39]), illustrating how international guidance informs practical national implementation.

Through this iterative process, the resulting guidelines will be globally informed; locally applicable; and co-designed to foster equity, accessibility, and usability.

Several areas also warrant further investigation to strengthen the evidence base supporting the 4 pillars. First, empirical evaluation of the 4 pillars across diverse health care settings, including primary care, community health, and public-facing dashboards, could test their impact on usability, adoption, and decision-making. Second, the 4 pillars could be translated into practical tools, such as checklists, design templates, or evaluation criteria, for use by health care organizations and developers. Third, comparative studies with other sectors (eg, business intelligence and education) may provide insights into transferable practices and highlight areas requiring health care–specific adaptation.

Conclusions

This review is an essential first step in understanding the global landscape of health care dashboards and provides a foundation for contextualization in Ireland. As outlined earlier, we have identified a significant gap, namely, a lack of standardized principles that could guide the design of dashboards in the health care sector. Arising from this review, we have defined 4 main pillars with associated key practices. These 4 pillars, derived from the international literature, will guide the upcoming consensus building process in Ireland and ensure that international evidence informs locally relevant guidelines. They will serve as the foundation for creating a comprehensive framework for health care data visualization in Ireland.

Ultimately, this phased approach from international review to consensus building and framework development ensures a solid foundation of evidence and co-design for transforming the design of data dashboards in the Irish health care system. By enabling dashboards that are not only functional but also user-friendly and impactful, this work supports improved decision-making, resource allocation, and health outcomes for all stakeholders.

XSL•FO
RenderX

Acknowledgments

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the Generative AI Delegation Taxonomy (GAIDeT) [40], the following task was delegated to GenAI tools under full human supervision: publication support. The GAI tool used was GPT-5-mini. The tool was used solely to assist with responding to reviewer comments, including rewriting text for clarity and consistency, providing clarification, and formulating responses to reviewers. GenAI was not used for ideation, study design, data analysis, interpretation of results, or drafting of the original manuscript. All content generated with AI assistance was reviewed, edited, and approved by the authors. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This publication has emanated from research jointly funded by Taighde Éireann—Research Ireland under grant number 12/RC/2289_P2—Insight Research Ireland Centre for Data Analytics, and supported by a Science Foundation Ireland Public Fellowship Award (project: HSE2 Dashboard Design for Health Service Executive Quality Improvement Reports). For the purpose of open access, the authors have applied a CC BY public copyright license to any author-accepted manuscript version arising from this submission.

Authors' Contributions

HV, SB, and JM developed the initial research objectives. HV, SB, and CC reviewed and selected relevant papers. HV and SB carried out the reflexive thematic analysis of the methodology sections of all papers and, together with CC and LKI, synthesized the themes into pillars. HV and SB drafted the initial manuscript. CC, LKI, GM, DM, and JM reviewed the manuscript and provided suggestions for improvement.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR checklist.
[DOC File , 119 KB - jmir_v28i1e77361_app1.doc]

Multimedia Appendix 2
Electronic search report.
[DOCX File , 23 KB - jmir_v28i1e77361_app2.docx]

Multimedia Appendix 3
Characteristics of reports included.
[DOCX File , 37 KB - jmir_v28i1e77361_app3.docx]

Multimedia Appendix 4
Overview of coding for reflexive thematic analysis.
[DOCX File , 33 KB - jmir_v28i1e77361_app4.docx]

Multimedia Appendix 5
Reports mapped to pillars.
[DOCX File , 26 KB - jmir_v28i1e77361_app5.docx]

References

1. Sarikaya A, Correll M, Bartram L, Tory M, Fisher D. What do we talk about when we talk about dashboards? IEEE Trans Vis Comput Graph 2018 Aug 21:682-692. [doi: 10.1109/TVCG.2018.2864903] [Medline: 30136958]
2. Sweller J. Cognitive load during problem solving: effects on learning. Cogn Sci 2010 Feb 11;12(2):257-285 [FREE Full text] [doi: 10.1207/s15516709cog1202_4]
3. Bateman S, Mandryk RL, Gutwin C, Genest A, McDine D, Brooks C. Useful junk? The effects of visual embellishment on comprehension and memorability of charts. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2010 Presented at: CHI '10; April 10-15, 2010; Atlanta, GA URL: https://doi.org/10.1145/1753326.1753716 [doi: 10.1145/1753326.1753716]

XSL•FO
RenderX

4. Wanderer JP, Nelson SE, Ehrenfeld JM, Monahan S, Park S. Clinical data visualization: the current state and future needs. J Med Syst 2016 Dec;40(12):275. [doi: 10.1007/s10916-016-0643-x] [Medline: 27787779]
5. Lurie NH, Mason CH. Visual representation: implications for decision making. Journal of Marketing 2007 Jan 01;71(1):160-177 [FREE Full text] [doi: 10.1509/jmkg.71.1.160]
6. Abudiyab NA, Alanazi AT. Visualization techniques in healthcare applications: a narrative review. Cureus 2022 Nov;14(11):e31355 [FREE Full text] [doi: 10.7759/cureus.31355] [Medline: 36514654]
7. Munzner T. Visualization Analysis and Design. Boca Raton, FL: CRC Press; 2014.
8. Few S. Information Dashboard Design: The Effective Visual Communication of Data. Sebastopol, CA: O'Reilly Media; 2006.
9. ISO 9241-210:2019 Ergonomics of human-system interaction part 210: human-centred design for interactive systems. International Organization for Standardization. 2019. URL: https://www.iso.org/standard/77520.html [accessed 2025-10-20]
10. Helminski D, Sussman JB, Pfeiffer PN, Kokaly AN, Ranusch A, Renji AD, et al. Development, implementation, and evaluation methods for dashboards in health care: scoping review. JMIR Med Inform 2024 Dec 10;12:e59828 [FREE Full text] [doi: 10.2196/59828] [Medline: 39656991]
11. Rabiei R, Almasi S. Requirements and challenges of hospital dashboards: a systematic literature review. BMC Med Inform Decis Mak 2022 Nov 08;22(1):287 [FREE Full text] [doi: 10.1186/s12911-022-02037-8] [Medline: 36348339]
12. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International Journal of Social Research Methodology 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
13. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
14. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
15. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Jun 13;11(4):589-597 [FREE Full text] [doi: 10.1080/2159676x.2019.1628806]
16. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev 2021 Mar 29;10(1):89 [FREE Full text] [doi: 10.1186/s13643-021-01626-4] [Medline: 33781348]
17. van de Baan FC, Lambregts S, Bergman E, Most J, Westra D. Involving health professionals in the development of quality and safety dashboards: qualitative study. J Med Internet Res 2023 Jun 12;25:e42649 [FREE Full text] [doi: 10.2196/42649] [Medline: 37307058]
18. Elshehaly M, Randell R, Brehmer M, McVey L, Alvarado N, Gale CP, et al. QualDash: adaptable generation of visualisation dashboards for healthcare quality improvement. IEEE Trans Vis Comput Graph 2021 Feb;27(2):689-699 [FREE Full text] [doi: 10.1109/TVCG.2020.3030424] [Medline: 33048727]
19. Concannon D, Herbst K, Manley E. Developing a data dashboard framework for population health surveillance: widening access to clinical trial findings. JMIR Form Res 2019 Apr 04;3(2):e11342 [FREE Full text] [doi: 10.2196/11342] [Medline: 30946016]
20. Weggelaar-Jansen AM, Broekharst DS, de Bruijne M. Developing a hospital-wide quality and safety dashboard: a qualitative research study. BMJ Qual Saf 2018 Dec;27(12):1000-1007 [FREE Full text] [doi: 10.1136/bmjqs-2018-007784] [Medline: 29950323]
21. Lee K, Jung SY, Hwang H, Yoo S, Baek HY, Baek RM, et al. A novel concept for integrating and delivering health information using a comprehensive digital dashboard: an analysis of healthcare professionals' intention to adopt a new system and the trend of its real usage. Int J Med Inform 2017 Jan;97:98-108. [doi: 10.1016/j.ijmedinf.2016.10.001] [Medline: 27919400]
22. Alvarado N, McVey L, Elshehaly M, Greenhalgh J, Dowding D, Ruddle R, et al. Analysis of a web-based dashboard to support the use of national audit data in quality improvement: realist evaluation. J Med Internet Res 2021 Nov 23;23(11):e28854 [FREE Full text] [doi: 10.2196/28854] [Medline: 34817384]
23. Bunting Jr RF, Siegal D. Developing risk management dashboards using risk and quality measures: a visual best practices approach. J Healthc Risk Manag 2017 Oct;37(2):8-28. [doi: 10.1002/jhrm.21287] [Medline: 28960593]
24. Fazaeli S, Khodaveisi T, Vakilzadeh AK, Yousefi M, Ariafar A, Shokoohizadeh M, et al. Development, implementation, and user evaluation of COVID-19 dashboard in a third-level hospital in Iran. Appl Clin Inform 2021 Oct;12(5):1091-1100 [FREE Full text] [doi: 10.1055/s-0041-1740188] [Medline: 34879405]
25. Randell R, Alvarado N, McVey L, Ruddle RA, Doherty P, Gale C, et al. Requirements for a quality dashboard: lessons from National Clinical Audits. AMIA Annu Symp Proc 2019;2019:735-744 [FREE Full text] [Medline: 32308869]
26. Ivanković D, Barbazza E, Bos V, Brito Fernandes Ó, Jamieson Gilmore K, Jansen T, et al. Features constituting actionable COVID-19 dashboards: descriptive assessment and expert appraisal of 158 public web-based COVID-19 dashboards. J Med Internet Res 2021 Feb 24;23(2):e25682 [FREE Full text] [doi: 10.2196/25682] [Medline: 33577467]

XSL•FO
RenderX

27. Kamadjeu R, Gathenji C. Designing and implementing an electronic dashboard for disease outbreaks response - case study of the 2013-2014 Somalia Polio outbreak response dashboard. Pan Afr Med J 2017;27(Suppl 3):22 [FREE Full text] [doi: 10.11604/pamj.supp.2017.27.3.11062] [Medline: 29296157]
28. Khanbhai M, Symons J, Flott K, Harrison-White S, Spofforth J, Klaber R, et al. Enriching the value of patient experience feedback: web-based dashboard development using co-design and heuristic evaluation. JMIR Hum Factors 2022 Feb 03;9(1):e27887 [FREE Full text] [doi: 10.2196/27887] [Medline: 35113022]
29. Ratwani RM, Fong A. 'Connecting the dots': leveraging visual analytics to make sense of patient safety event reports. J Am Med Inform Assoc 2015 Mar;22(2):312-317. [doi: 10.1136/amiajnl-2014-002963] [Medline: 25336592]
30. Ahn E, Liu N, Parekh T, Patel R, Baldacchino T, Mullavey T, et al. A mobile app and dashboard for early detection of infectious disease outbreaks: development study. JMIR Public Health Surveill 2021 Mar 09;7(3):e14837 [FREE Full text] [doi: 10.2196/14837] [Medline: 33687334]
31. Pestana M, Pereira R, Moro S. Improving health care management in hospitals through a productivity dashboard. J Med Syst 2020 Mar 12;44(4):87. [doi: 10.1007/s10916-020-01546-1] [Medline: 32166499]
32. Christen OM, Mösching Y, Müller P, Denecke K, Nüssli S. Dashboard visualization of information for emergency medical services. Stud Health Technol Inform 2020 Nov 23;275:27-31. [doi: 10.3233/SHTI200688] [Medline: 33227734]
33. Loorak MH, Perin C, Kamal N, Hill M, Carpendale S. TimeSpan: using visualization to explore temporal multi-dimensional data of stroke patients. IEEE Trans Visual Comput Graphics 2016 Jan;22(1):409-418. [doi: 10.1109/tvcg.2015.2467325]
34. Bertin J. Sémiologie Graphique: Les Diagrammes, Les Réseaux, Les Cartes. Berlin, Germany: Mouton Publishers; 1973.
35. Nielsen J. Enhancing the explanatory power of usability heuristics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 1994 Presented at: CHI '94; April 24-28, 1994; Boston, MA URL: https://doi.org/10.1145/191666.191729 [doi: 10.1145/191666.191729]
36. Norman D. Affordances and design. jnd.org. 2008. URL: https://jnd.org/affordances-and-design/ [accessed 2025-08-25]
37. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
38. May C, Finch T. Implementing, embedding, and integrating practices: an outline of normalization process theory. Sociology 2009 Jun 15;43(3):535-554 [FREE Full text] [doi: 10.1177/0038038509103208]
39. Data Protection Commission. URL: https://www.dataprotection.ie/en/organisations [accessed 2025-01-15]
40. GAIDeT declaration generator. GitHub. URL: https://panbibliotekar.github.io/gaidet-declaration/index.html [accessed 2025-11-13]

Abbreviations

MeSH: Medical Subject Headings
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RTA: reflexive thematic analysis

Edited by J Sarvestan; submitted 16.May.2025; peer-reviewed by S Guan, G Adewumi, O Nguyen; comments to author 20.Aug.2025; accepted 18.Nov.2025; published 25.Feb.2026.

Please cite as:
Vornhagen H, Barrett S, Carroll C, Iladiva LK, Martin G, McKeown D, Martin J
Design Practices for Data Dashboards in Health Care: Scoping Review
J Med Internet Res 2026;28:e77361
URL: https://www.jmir.org/2026/1/e77361
doi:10.2196/77361
PMID:

©Heike Vornhagen, Stephen Barrett, Ciara Carroll, Lydia Kavochi Iladiva, Gregory Martin, Declan McKeown, Jennifer Martin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Effects of Digital Health Interventions for Family Members in Intensive Care Units: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Chuchu Zhang[1], BSN; Weijing Sui[1], MSN; Weilin Jiang[1], BSN; Yibing Shen[1], BSN; Zhenzhen Huang[1], BSN; Ran Yan[1], BSN; Jia Yi[1], BSN; Jiayu Zhang[1], MSN; Chaoyi Zhang[1], BSN; Yiyu Zhuang[1], MSN; Xiaoyan Gong[1], MSN

Department Nursing, Zhejiang University School of Medicine Sir Run Run Shaw Hospital, Hangzhou, Zhejiang, China

Corresponding Author:
Xiaoyan Gong, MSN
Department Nursing
Zhejiang University School of Medicine Sir Run Run Shaw Hospital
No.3 Qingchun East Road, Shangcheng District
Hangzhou, Zhejiang, 310016
China
Phone: 86 0571 86006616
Email: 3203071@zju.edu.cn

Abstract

Background: Family members of intensive care units (ICUs) often experience psychological difficulties like anxiety and depression, which affect family functioning and their ability to support patient recovery. Digital health interventions (DHIs) offer accessible and timely support; yet, evidence regarding their effectiveness for ICU families demonstrates inconsistent.

Objective: This systematic review and meta-analysis aimed to explore the impact of DHIs on psychological outcomes, quality of life (QoL), and quality of communication (QOC) among ICU family members.

Methods: A total of 9 databases (CNKI [China Knowledge Resource Integrated Database], Wanfang, VIP [Weipu Database], SinoMed, Cochrane Library, PubMed, Embase, CINAHL, and Web of Science) were searched for randomized controlled trials (RCTs) published up to October 11, 2025. Eligible studies evaluated DHIs targeting adult ICU patients' family members and reported psychological outcomes, QoL, or QOC; interventions used solely for monitoring or tracking were excluded. Furthermore, 2 reviewers independently screened studies, extracted data, and assessed the risk of bias using the Cochrane Risk of Bias Tool. Random-effects meta-analyses were performed using the Hartung-Knapp-Sidik-Jonkman method. Prediction intervals (PIs) were calculated to reflect between-study heterogeneity. Subgroup analyses were performed by patients' primary diagnosis, relationship with patients, and type of digital health technology. Evidence quality was assessed using the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach.

Results: In total, 17 RCTs involving 1864 participants were included. Interrater agreement was high across study selection and assessment processes (κ=0.55-1.00). Most studies were judged to have some concerns about the risk of bias. Meta-analyses showed no statistically significant pooled effects of DHIs on anxiety (standardized mean difference [SMD] –0.34, 95% CI –0.68 to 0.00; 95% PI –1.46 to 0.79; *P*=.05), depression (SMD –0.26, 95% CI –0.52 to 0.01; 95% PI –1.11 to 0.60; *P*=.06), posttraumatic stress disorder (SMD –0.21, 95% CI –0.49 to 0.06; 95% PI –1.03 to 0.60; *P*=.11), QoL (SMD 0.09, 95% CI –0.10 to 0.28; 95% PI –0.41 to 0.60; *P*=.36), or QOC (SMD 0.14, 95% CI –0.03 to 0.31; 95% PI –0.26 to 0.55; *P*=.10). Wide PIs indicated substantial variability in intervention effects across settings. The certainty of evidence ranged from very low to moderate.

Conclusions: This review emphasizes significant uncertainty in estimated effects and provides the first thorough synthesis of RCT evidence on DHIs for ICU family members. The inherently high heterogeneity and low certainty of the underlying evidence limit the conclusions of this review, despite its methodological rigor and use of PIs. Compared with previous reviews, this review concentrates on the ICU family members and includes more recent research. Future studies should focus on high-risk subgroups, use mixed methods designs, and develop theory-informed, personalized, and interactive DHIs. DHIs may enable proactive, customized communication techniques for families of ICU patients, even when there are few formal clinical recommendations.

Trial Registration: PROSPERO CRD420251044704; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251044704

(J Med Internet Res 2026;28:e83294) doi:10.2196/83294

XSL•FO
RenderX

KEYWORDS

digital health interventions; intensive care units; family members; psychology; systematic review; meta-analysis; randomized controlled trials

Introduction

Patients admitted to the intensive care units (ICUs) frequently experience varying degrees of organ or system failure, which can be life-threatening [1]. Their families may endure enormous burdens and suffer from psychological disorders like anxiety, depression, and posttraumatic stress disorder (PTSD). The Society of Critical Care Medicine's 2024 Clinical Practice Guidelines on Family-Centered Care for Adult ICUs underscore the need for delivering patient-centered and family-centered care in the ICU to uphold clinical excellence [2]. Nonetheless, various psychological challenges for family members, including anxiety, depression, acute stress disorder, PTSD, and complicated grief, are induced by the uncertainty of disease trajectories, the complexity of medical decisions, and the urgent demand for time. The term "postintensive care syndrome family" (PICS-F) encompasses these mental health disorders [3,4]. In addition to undermining family functioning and integrity, these problems may hinder the family's ability to facilitate the patient's recovery in the ICU [5,6].

A retrospective study [6] showed that the prevalence of psychological disorders related to PICS-F among family members of ICU patients within 6 months post admission was 15.1%, with spouses, as primary caregivers, exhibiting a heightened risk of psychological disorders. Furthermore, 1 research indicated that 10% to 69% of family members of ICU patients experienced moderate to severe psychological distress, including symptoms of anxiety, depression, and PTSD, potentially persisting for up to a year [7,8]. Family members endure not only emotional distress, such as anxiety and depression, but also are influenced by several biopsychosocial factors, including sleep deprivation, diminished quality of life (QoL) and mental health, and insufficient social support. These factors make it difficult for them to provide adequate family support to facilitate the patient's recovery [5]. Common contributors to PICS-F among family members include female sex, relatively younger age, younger patient age, lower educational level, unmarried status, inadequate communication skills of ICU medical personnel, and the degree of family participation in decision-making [3,9]. Several strategies, such as ICU diaries [10], information support, and peer support, have been implemented to mitigate PICS-F but often focus on the patient's hospitalization problems [11,12], and due to the family dynamics, geography, family resources, and family values, the effect of related intervention on PTSD, anxiety, and depression among family members is unclear [13]. Traditional educational tools are rarely implemented in the ICU environment due to high service costs, limited resources, and a lack of personalization [14,15]. With the rapid development of internet technology, mobile health care solutions have become popular, providing innovative psychological empowerment interventions [16].

Digital health technology (DHT) was initially introduced by the World Health Organization in the Global Strategy for Digital Health (2020-2024) and is defined as a comprehensive technical practice that facilitates clinical care, health-related education for patients and professionals, and public health through the extensive usage of electronic information and communication technologies [17,18]. Compared with conventional interventions, digital health interventions (DHIs) amalgamate computing platforms, connectivity, software, networked devices, and sensors—including smartphone apps, wearable devices, web-based communication platforms, videoconferencing, chatbots, artificial intelligence, and virtual reality (VR) technologies—leveraging technological advancements to transcend temporal and spatial limitations [19,20]. DHIs provide family members with greater accessibility, affordability, and convenience in receiving timely support and information, which is critical in the ICU context. In addition, DHIs help reduce health care spending [21,22]. DHIs have been shown to improve the psychological outcomes and QoL of caregivers, such as depressive symptoms, perceived stress, anxiety, and stress in other high-stress situations, such as caregivers of patients with cancer [23], dementia [24,25], and mental illness [15]. This study highlights the potential value of its application to family members of ICU patients. However, other studies have suggested that DHIs have little effect on alleviating the anxiety and depression of these family members [26]. Therefore, there are ambiguities in the existing research results on whether DHIs can effectively alleviate the psychological distress symptoms of family members of ICU patients [26].

After searching relevant databases, no meta-analysis of these intervention outcomes has been found. Randomized controlled trials (RCTs) continue to serve as the gold standard for delivering the most rigorous scientific findings [27]. Therefore, a systematic review and meta-analysis of RCTs focusing on the effects of DHIs on ICU family members is essential. This meta-analysis will assess the effects of DHIs on alleviating psychological symptoms, improving QoL, and enhancing communication quality among ICU family members, thereby providing critical evidence for the formulation of a comprehensive and effective psychological support program.

Methods

Research Design

This systematic review and meta-analysis were planned, conducted, and reported following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) [28] (Multimedia Appendix 1). A systematic literature search, screening, study selection, data extraction, and risk-of-bias assessment were conducted according to the Cochrane methodology [29].

Search Strategy

The search strategy was reported in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-analyses literature search extension) checklist

XSL•FO
RenderX

(Multimedia Appendix 2). In total, 2 investigators (Chuchu Z and WJ) independently systematically searched 9 databases, including CNKI [China Knowledge Resource Integrated Database], Wanfang, VIP (Weipu Database), SinoMed, Cochrane Library, PubMed, Embase, CINAHL, and Web of Science, from inception to October 11, 2025. The search terms included "digital health," "intensive care units," "family," and "randomized controlled trial," which were pilot-tested and validated by independent researchers. No published search filters for study design were used. To identify additional studies, we manually searched the reference lists of included articles and relevant websites, including PROSPERO clinical trial registries. Before the final data analysis, the database searches were updated. Gray literature, such as conference abstracts and academic dissertations, was excluded. The whole search strategy is available in Multimedia Appendix 3.

Eligibility Criteria

Studies were selected based on the following inclusion criteria: (1) population: adult family members (≥18 y) of adult ICU patients (a family member could be a spouse, partner, friend, or relative [blood-related or not]); (2) interventions: all DHIs, which may include one or more methods, such as telephone, video, websites, or mobile apps, wearable devices, and others; (3) control: a comparison group that received either standard care, routine care or a blank control; (4) outcomes: the primary outcomes associated with the psychological health of family members included anxiety, depression, PTSD, and other related outcomes like QoL; and (5) study type: RCTs.

Studies were excluded if they met one of the following criteria: (1) articles were excluded if the DHI was used for monitoring or tracking purposes only; (2) duplicate publications, or (3) full text was not available.

Study Selection

The search results were imported into NoteExpress (Aegean Software) for review. After removing duplicates, 2 investigators (Chuchu Z and RY) independently screened the titles and abstracts, then reviewed the full texts of the remaining records. Any disagreements were resolved by discussion with the third investigator (WS). The κ coefficients were used to specify the consistency of independent work. The degree of consistency was identified as slight, fair, moderate, substantial, and almost perfect when the κ coefficients were calculated as 0.00-0.20, 0.21-0.40, 0.41-0.60, 0.61-0.80, and 0.81-1.00, respectively [30].

Data Extraction

Two authors (Chuchu Z and JY) independently extracted data from each of the included studies using a structured data extraction form, including (1) study: author, year, country, and design; (2) participants: patients' diagnosis, sample size, female, age, relationship with patients, occupation, marital status, education level, and financial burden; (3) control; (4) experimental: intervention measurement, intervention delivery (categorized as mobile app, internet-based, and others [eg, telephone calls, VR, and emails]), intervention duration, and data collection time points; (5) outcomes and measurements; and (6) results. If necessary, the corresponding authors were contacted by email and asked for the missing information. A third reviewer (XG) verified the extracted data.

Study Risk of Bias Assessment

The Cochrane's risk-of-bias tool for randomized trials, version 2 (RoB 2), was used by 2 independent reviewers (JZ and ZH) to assess [31,32]. The RoB 2 tool consists of 5 domains: randomization process, deviations from intended intervention, missing outcome data, measurement of the outcome, and selection of the reported result. The risk of bias was assessed as low, some concerns, or high in the RoB 2 based on responses to signal questions in each domain. The κ coefficient was calculated for each item to identify the degree of consistency.

Quality of Evidence

The Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach was used to assess the certainty of the body of evidence [33]. According to the approach, the evidence quality of RCTs was initially graded as high and downgraded to moderate, low, or very low when any limitations were identified in risk of bias, inconsistency, directness, imprecision, or publication bias. However, the evidence could be upgraded by a significant effect and dose-response gradient. Furthermore, 2 authors (Chaoyi Z and YS) independently assessed the quality of evidence using the GRADE handbook, and the interrater agreement was evaluated using the κ coefficient. Disagreements were resolved through discussion, or help was sought from a third author (YZ).

Statistical Analysis

For continuous outcomes, the Hartung-Knapp-Sidik-Jonkman random-effects method [34] was used to obtain standardized mean differences (SMDs) with 95% CIs when studies used different psychometric scales. In studies with follow-up lasting more than 1 month, data from the shortest follow-up period were used. A meta-analysis was undertaken if more than 2 studies reported outcomes in the same format; otherwise, the studies' findings were described narratively [35]. We applied random-effects models because of suspected heterogeneity across studies to provide a relatively conservative estimate of the combined results [36]. Moreover, the prediction intervals (PI) were reported to indicate heterogeneity in the same metric as SMD and to reflect the variation in actual treatment effects expected in future studies, which aids clinical decision-making [37]. A leave-one-out sensitivity analysis was performed to assess the robustness of the synthesized results. Subgroup analysis was conducted to explore the potential sources of heterogeneity. These analyses were based on patient primary diagnosis (medical vs surgical), relationship with patients (spouse vs children), and type of DHT (education-based vs communication-based vs mental health–related). Finally, if sufficient studies were available for the outcome (n>10), a funnel plot was generated to assess potential small-study effects [38]. The Egger test was performed to further quantify funnel plot asymmetry, with a P value <.05 considered statistically significant [39]. A 2-tailed P<.05 was considered statistically significant. All analyses and visualizations were conducted using the meta package in R software (version 4.5.1; R Core Team).

XSL•FO
RenderX

Ethical Considerations

This study and paper used published data, so no ethics approval was needed.

Results

Study Selection

The PRISMA flow diagram in Figure 1 summarizes the decision pathway for final study inclusion. In total, 2 authors (Chuchu Z and WJ) independently conducted the study selection process, and the κ coefficients for title or abstract screening and full-text reading were 0.82 (P<.001) and 0.84 (P<.001), respectively, indicating almost perfect consistency in these steps. We identified 2624 records through database searches and 2 records through citation searching. After title and abstract screening, we excluded 1944 records and identified 55 records for full-text evaluation. After the full-text screening, we excluded 38 records that did not meet the inclusion criteria and included 17 studies [14,16,26,40-53].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) flow diagram. CNKI: China Knowledge Resource Integrated Database; VIP: Weipu Database.

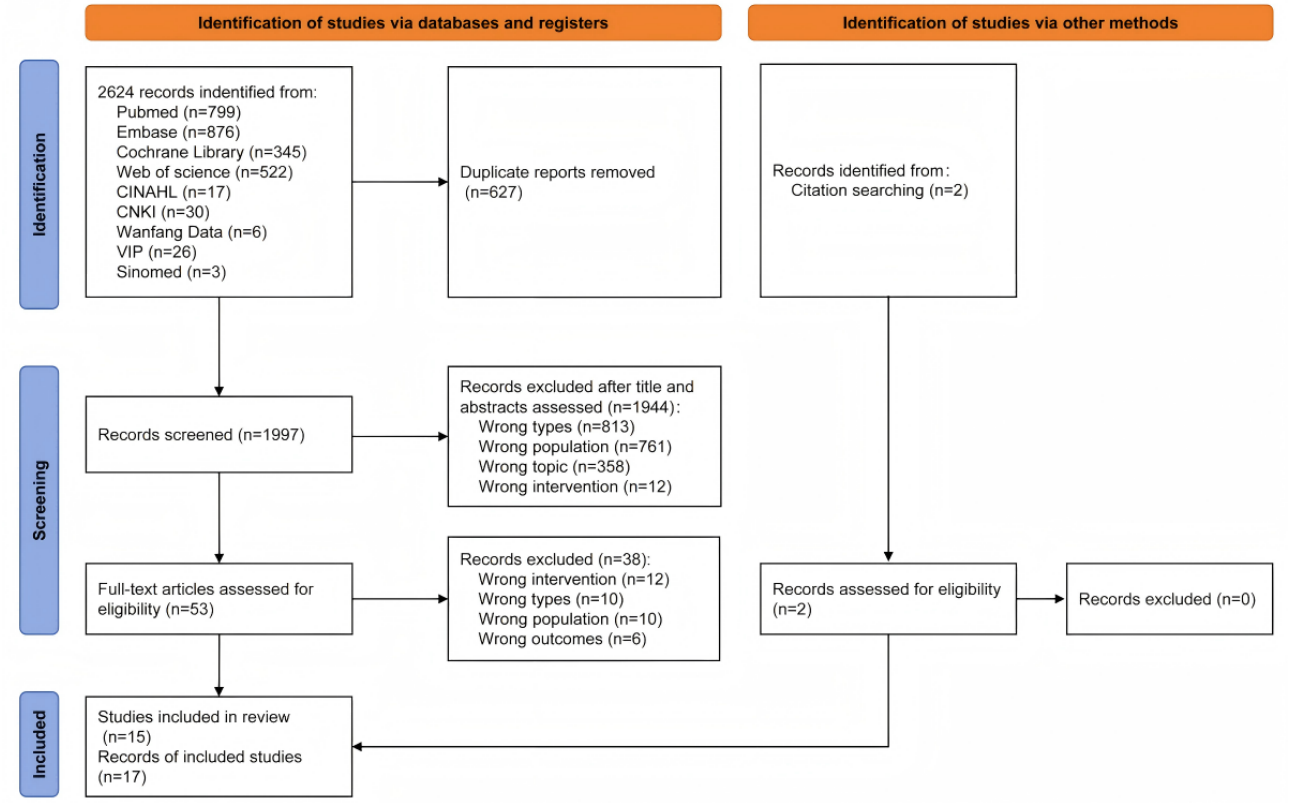

Study Characteristics

Overview

The characteristics of all the included studies are summarized in Table 1. Among the 17 RCTs, 2 were cluster RCTs [41,43], and 4 were pilot RCTs [26,45,47,53]. Studies were conducted in the United States (n=8) [26,41,42,45,47,51-53], China (n=3) [16,40,50], the Netherlands (n=1) [43], Norway (n=1) [48], Germany (n=1) [44], Austria (n=1) [14], Iran (n=1) [46], and Korea (n=1) [49], which were published from 2017 to 2025.

XSL•FO
RenderX

Table 1. Characteristics of the studies and participants in the included studies.

Study (year, country)	Design	Patient's diagnosis	Sample size, n (I/C[a]), % of female	Age (y), mean (SD)	Relationship with patients, n (%)	Occupation, n (%)	Marital status, n (%)	Education level, n (%)	Financial burden, n (%)
Carlozzi et al [51] (2025, United States)	RCT[b]	• Traumatic brain injury 100%	257 (129/128), 79.1	52.25 (14.48)	• Partner 104 (40.9) • Parent 83 (32.7) • Child 36 (14.2) • Sibling 20 (7.9) • Other 6 (2.4)	• Employed 157 (61.8) • Retired 58 (22.8) • Unemployed 15 (5.9) • Disabled 8 (3.1) • Other 16 (6.3)	—[c]	—	—
Chiang et al [40] (2017, China)	RCT	—	74 (39/35), 62.2	—	• Children 32 (43.2) • Parent 21 (28.4) • Spouse 9 (12.2) • Sibling 5 (6.8) • Other 7 (9.5)	—	• Married 45 (60.8) • Single 22 (29.7) • Divorced 7 (9.5)	• Below university 46 (62.2) • University or above 26 (35.1)	• High 32 (43.2) • Low 31 (41.9)
Cox et al [42] (2019, United States)	RCT	• Medical 50.9% • Neurological 22% • Surgical 13.4% • Trauma 11.9% • Cardiology 1.8%	275 (137/138), 73.1	51.20 (12.60)	• Spouse 128 (46.5) • Children 65 (23.6) • Parent 54 (19.6) • Sibling 22 (8) • Other 6 (2.2)	• Employed 184 (66.9) • Retired 54 (19.6) • Disabled 18 (6.5) • Unemployed 17 (6.2) • Other 2 (0.7)	• Married 210 (76.4) • Single 28 (10.2) • Divorced 37 (13.4)	• Below university 157 (57.1) • University or above 118 (42.9)	• Low 130 (42.3) • No 80 (29.1) • Moderate 45 (16.4) • High 20 (7.3)
Cox et al [41] (2024, United States)	Cluster RCT	• Medical 51.6% • Neurological 22% • Shock 14% • Trauma 13%	111 (55/56), 83	51.00 (15.00)	• Spouse 50 (45) • Parent 23 (20.7) • Children 22 (19.8) • Sibling 11 (9.9) • Other 5 (4.5)	• Employed 81 (73) • Retired 19 (17.1) • Disabled 7 (6.3) • Unemployed 4 (3.6)	• Married 67 (60.4) • Single 21 (18.9) • Divorced 23 (20.7)	• Below university 20 (18) • University or above 91 (82)	• No 45 (40.5) • Low 44 (39.6) • Moderate 13 (11.7) • High 9 (8.1)
Cox et al [52] (2025, United States)	RCT	• Medical 37.1% • Neurological 27.8% • Cardiac 24.5% • Surgical 8% • Community medical-surgical 2.7%	151 (76/75), 72.9	57.40 (12.90)	• Spouse 80 (60.6) • Child 36 (27.3) • Parent 12 (9.1) • Sibling 2 (1.5) • Other 2 (1.5)	• Employed 82 (62.1) • Retired 36 (27.3) • Disabled 11 (8.3) • Unemployed 3 (2.3)	• Married 107 (81.1) • Single 13 (9.9) • Divorced 12 (9.1)	—	—
Yuan et al [50] (2023, China)	RCT	• Surgical 100%	98 (50/48)	—	• Children 66 (67.3) • Spouse 24 (24.5)	—	—	—	—

XSL•FO
RenderX

Study (year, country)	Design	Patient's diagnosis	Sample size, n (I/C[a]), % of female	Age (y), mean (SD)	Relationship with patients, n (%)	Occupation, n (%)	Marital status, n (%)	Education level, n (%)	Financial burden, n (%)
Drop et al [43] (2025, Netherlands)	Cluster RCT	• Medical 60.9% • Surgical 31.7% • Trauma 7.5%	189 (100/89), 53.4	47.35 (12.36)	• Children 71 (37.6) • Spouse 62 (32.8) • Sibling 11 (5.8) • Parent 9 (4.8) • Other 7 (3.7)	—	—	• Below university 142 (75.1) • University or above 18 (9.5)	—
Gawlytta et al [44] (2022, Germany)	RCT	• Severe sepsis 100%	25 (12/13), 64	54.00 (11.01)	• Spouse 25 (100)	—	• Married 21 (84)	• Below university 15 (60) • University or above 10 (40)	—
Hoffmann et al [14] (2023, Austria)	RCT	• Nonsurgical 69.7% • Surgical 30.3%	89 (46/43), 66.3	47.00 (13.00)	• Spouse 36 (40.4) • Children 31 (34.8) • Sibling 9 (10.1) • Parent 8 (9) • Other 5 (5.6)	• Employed 67 (75.3) • Unemployed 22 (24.7)	—	• Below university 55 (61.8) • University or above 34 (38.2)	—
Petrinec et al [45] (2023, United States)	Pilot RCT	• Infectious disease 31.7% • Neurological 13.3% • Cardiovascular 11.7% • Gastrointestinal 8.3% • Trauma 8.3%	60 (30/30), 61.7	46.39 (13.04)	• Children 23 (38.3) • Spouse 16 (26.7) • Parent 8 (13.3) • Sibling 7 (11.7) • Other 6 (10)	—	—	—	—
Shariati et al [46] (2021, Iran)	RCT	• Patient with COVID-19	67 (33/34), 46.3	38.68 (11.18)	• Children 37 (55.2) • Spouse 21 (31.3) • Parent 4 (6) • Sibling 1 (1.5) • Other 4 (6)	• Employed 62 (92.5) • Retired 2 (3) • Student 3 (4.5)	• Married 41 (61.2) • Single 26 (38.8)	• Below university 4 (6) • University or above 63 (94)	—
Shin et al [26] (2025, United States)	Pilot RCT	• Respiratory 42.9% • Cardiovascular 21.4% • Sepsis 17.9% • Surgical 17.9%	28 (14/14), 85.7	50.93 (16.05)	• Spouse 19 (67.9) • Parent 4 (14.3) • Sibling 3 (10.7) • Children 2 (7.1)	• Employed 12 (42.9) • Unemployed 6 (21.4) • Retired 8 (28.6) • Disabled 2 (7.1)	—	• Below university 14 (50) • University or above 14 (50)	—

XSL•FO
RenderX

Study (year, country)	Design	Patient's diagnosis	Sample size, n (I/C[a]), % of female	Age (y), mean (SD)	Relationship with patients, n (%)	Occupation, n (%)	Marital status, n (%)	Education level, n (%)	Financial burden, n (%)
Suen et al [47] (2021, United States)	Pilot RCT	• Medical 100%	48 (23/25), 62.5	56.37 (12.95)	• Spouse 16 (33.3) • Children 16 (33.3) • Sibling 4 (8.3) • Parent 4 (8.3) • Other 8 (16.7)	—	—	• Below university 24 (50) • University or above 24 (50)	—
Vranceanu et al [53] (2020, United States)	Pilot RCT	• Neurological 100%	58 (29/29), 67.2	52.25 (14.48)	• Spouse 46 (79.3) • Parent 7 (12.1) • Sibling 2 (3.4) • Other 3 (5.2)	• Employed 41 (70.7) • Retired 10 (17.2) • Other 7 (12.1)	—	• Below university 29 (50) • University or above 29 (50)	—
Watland et al [48] (2025, Norway)	RCT	• Respiratory 37.7% • Cardiovascular 20.4% • Accident 18.4% • Neurological 13.3% • Sepsis 12.7%	196 (101/95), 64.8	46.80 (13.51)	• Children 79 (40.3) • Spouse 45 (23) • Parent 32 (16.3) • Sibling 26 (13.3) • Other 14 (7.1)	—	—	• Below university 73 (37.2) • University or above 123 (62.8)	—
Woo et al [49] (2024, Korea)	RCT	• Surgery 71.1% • Shock 18.4% • Multiple traumas 7.9% • Respiratory 2.6%	38 (18/20), 68.4	51.60 (11.80)	• Spouse 20 (52.6) • Children 8 (21.1) • Parent 5 (13.2) • Sibling 4 (10.5) • Other 1 (2.6)	—	—	—	—
Xiong et al [16] (2025, China)	RCT	• Patients who underwent HVR[d]	100 (50/50), 61	—	• Children 48 (48) • Spouse 42 (42) • Parent 8 (8) • Sibling 1 (1) • Other 1 (1)	• Employed 92 (92) • Retired 8 (8)	• Married 93 (93) • Not married 7 (7)	• Below university 75 (75) • University or above 25 (25)	• Moderate 53 (53) • Low 41 (41) • High 6 (6)

[a]I/C: intervention/control.

[b]RCT: randomized controlled trial.

[c]Not applicable.

[d]HVR: heart valve replacement.

Participants

The sample sizes ranged from 25 to 275, yielding a total of 1864 participants (intervention: 942, 50.5% and control: 922, 49.5%). Regarding patient types, only 1 study [40] did not report; 11 studies [14,26,41-48,52] focused on medical patients; and 5 studies [16,49-51,53] focused on surgical patients. Furthermore, 7 studies [41,42,44,49,50,52,53] recruited not only family

XSL•FO
RenderX

members but also ICU patients or ICU health care staff, whereas the remaining 10 studies recruited only family members. In addition, 14 studies reported the mean ages of family members, ranging from 38.68 (SD 11.18) years to 57.40 (SD 12.90) years. Except for the study by Yuan et al [50], all studies reported that 46.3% to 85.7% of participants were female. Concerning the relationship to patients, 7 studies [16,40,43,45,46,48,50] reported that the main participants were children, and 10 studies reported that they were spouses. Moreover, 9 studies [14,16,26,41,42,46,51-53] reported participants' occupation, indicating most of them were employed; 10 studies [14,26,43,45,47-51,53] did not report marital status; the other 7 did, and the majority of participants were married. Regarding educational level, 5 studies [45,49-52] did not report it; 8 reported that most participants had nontertiary education; and 4 reported that most participants had university education or higher. Financial burden was not reported in 13 studies; 2 studies [41,42] reported no or low burden, and 2 studies [16,40] reported moderate or high burden (Table 1).

Intervention

In the included studies, various types of DHT were used. In total, 9 studies [16,26,40,41,45,49-53] used mobile apps, 5 studies [14,42,44,46,47] used internet-based technology, and 3 used other technologies (1 was a video tool [49], 1 was a digital assessment tool [48], and 1 was VR technology [43]). Furthermore, 10 studies [41-46,49,51-53] reported the specific duration of the intervention, and the remaining studies indicated that family members could use digital intervention tools during the ICU or within months after discharge. Regarding data collection timing, in addition to pre- and postintervention assessments, 6 studies [14,16,26,43,45,51] collected data at 1 month post intervention; 1 study [45] at 2 months post enrollment; 10 studies [14,26,41-43,47,48,51-53] at 3 months post enrollment or discharge; 4 studies [26,42,43,51] at 6 months post discharge; and 2 studies [14,51] at 12 months post admission. No adverse outcomes or events were reported across studies (Table 2).

XSL•FO
RenderX

Table 2. Characteristics of the outcomes and measurements in the included studies.

Study (year, country)	Control	Experimental				Outcomes and measurements	Results
		Intervention measurement	Intervention delivery	Intervention duration	Data collection time points		
Carlozzi et al [51] (2025, United States)	Self-monitoring only	Self-monitoring plus personalized self-care from the CareQOL app	Mobile app	10-day run-in period followed by a 6-month home monitoring period	baseline; 1-12 months	• Caregiver appraisal: CAS[a] • Adaptability: MPAI-4[b] • PTSD[c]: PCL-5[d] • Caregiver-specific anxiety: PROMIS[e] • Sleep-related impairment: PROMIS • Sepression: PROMIS • Fatigue: PROMIS • Perceived stress: PROMIS • Self-efficacy general: PROMIS • Overall health: PROMIS	• Although we did not observe improvements in HRQOL[f], physical activity, or sleep, we found that across all measures, approximately one-third of participants showed clinically meaningful improvements, one-third remained the same, and one-third worsened. • Care partners who reported engaging in the intervention were more likely to show improvement than those who did not. • There was preliminary support for factors that being male, caring for a person with posttraumatic stress symptoms, living in the same household as the person with TBI[g], being a spousal care partner, working, and being diagnosed with COVID-19 during the study were associated with increased risk for adverse outcomes.
Chiang et al [40] (2017, China)	EF-R[h]	EF-T[i]	Interactive mobile technology	During ICU[j]	Baselines; the posttest one day after EF-T or EF-R	• Depression: C-DASS[k] • Anxiety: C-DASS • Stress: C-DASS • Satisfaction: CPCS[l]	• The EF-R group had a slightly lower stress score than the EF-T group at posttest, but the difference was statistically insignificant. • There was also no significant improvement in anxiety after EF-T. • The EF-T group had a significantly better reduction of depression score than the EF-R group. • Information need satisfaction was not significantly different between intervention and control groups.

XSL•FO
RenderX

Study (year, country)	Control	Experimental				Outcomes and measurements	Results
		Intervention measurement	Intervention delivery	Intervention duration	Data collection time points		
Cox et al [42] (2019, United States)	Usual care practices followed by a family meeting	A web-based decision aid	Web-based	2 days	In person at enrollment but before randomization; on study day 3; via telephone at 3 months and 6 months	• Clinician surrogate concordance: CSCS[m] • Comprehension of diagnosis, treatment, and prognosis: MCS[n] • Clinician communication satisfaction: QOC[o] • Anxiety: HADS[p] • Depression: HADS • PTSD: PTSS[q] • Decision-making uncertainty: DCS[r] • Patient-centeredness of care: PPPC[s]	• Concordance improvement did not differ between intervention and control groups. • Intervention surrogates had a greater reduction in decisional conflict than control surrogates. • Intervention and control primary surrogates did not differ in medical comprehension or communication. • Surrogates' total scores on anxiety, depression, and PTSD also did not differ at 3 and 6 months.
Cox et al [41] (2024, United States)	Usual ICU care without protocolized family meetings	ICUconnect uses text messages and emails	Mobile app	7-10 days	Baselines; study days 1, 3, and 7; 3 months	• Needs at the End-of-Life Screening Tool: NEST[t] • Goal concordance of care • Quality of communication: QOC • Patient-centeredness of decision-making, eliciting concerns, discrimination: IPC[u] • Depression: PHQ-9[v] • Anxiety: GAD-7[w] • PTSD: PTSS	• There was greater improvement in NEST scores among intervention recipients between baseline and both day 3 and day 7. • There were no treatment group differences at 3 months in psychological distress symptoms. • Change from baseline to study day 3 was not statistically significantly different between intervention and control groups in IPC-Elicited Concerns, and IPC-Decision Making subscales, as well as QOC scores. • Although intervention family members reported a numerically greater improvement in goal-concordant care compared with control family members, this finding was not statistically significant. • White family members experienced a greater reduction in NEST scores compared with Black family members at day 3 and day 7.
		Intervention measurement	Intervention delivery	Intervention duration	Data collection time points		

XSL•FO
RenderX

Study (year, country)	Control	Experimental				Outcomes and measurements	Results
		Intervention measurement	Intervention delivery	Intervention duration	Data collection time points		
Cox et al [52] (2025, United States)	Usual ICU care	PC planner	Mobile app-based platform	7 days	Baseline; day1; day3; day 7; and 3 month	• Needs at the End-of-Life: NEST • Goal concordant of care: GCC[x] • Quality of communication: QOC • Patient-centeredness: PPPC • Depression: PHQ-9 • Anxiety: GAD-7 • PTSD: PTSS	• Treatment group differences in estimated mean NEST scores were similar at 3 days and at 7 days. • Goal concordance, communication quality, and psychological distress symptoms did not differ. • 29 (38.2%) intervention participants had palliative care consultations, compared with only 3 (4%) among controls (P<.001).
Yuan et al [50] (2023, China)	Routine care	Video visitation	WeChat video call	During the ICU, every afternoon	Admission to the ICU, and discharge from the ICU	• Anxiety: SAS[y] • Satisfaction: questionnaire in the hospital	• There were no statistically significant differences between the groups in patients' anxiety and depression scores, and no statistically significant differences in delirium incidence between the groups. • There were no significant differences in changes in family members' anxiety scores. • A statistically significant difference in satisfaction was found between the two groups of patients, and the result of family members' satisfaction was also statistically significant.
Drop et al [43] (2025, Netherlands)	Standard care	Standard care plus ICU-VR[z]	VR	14 min	Immediately after enrollment at ICU; admission at ICU; discharge at 1, 3, and 6 months post-ICU discharge	• PTSD: IES-R[aa] • Anxiety: HADS • Depression: HADS • Quality of life: SF-36[ab] • Relatives' understanding: CQI[ac]-relatives ICU • Perceived stress and appreciation of ICU-VR: family satisfaction with ICU care tools	• Relatives who received ICU-VR did not experience a decrease in PTSD, anxiety, or depression. • There was no significant difference between the median mental quality of life, physical quality of life, or understanding of ICU care. • Patients in the intervention group highly endorsed ICU-VR, favoring it over traditional informational brochures, and the majority stated it improved their understanding of ICU treatment.
Gawlytta et al [44] (2022, Germany)	5-week waiting period	iCBT[ad]: 10 writing assignments	Internet-based	Two 50-minute sessions per week over 5 weeks	Start of intervention or waiting; end of intervention or waiting; end of intervention in WL[ae] control group		

XSL•FO
RenderX

Study (year, country)	Control	Experimental				Outcomes and measurements	Results
		Intervention measurement	Intervention delivery	Intervention duration	Data collection time points		
						• PTSD: PCL-5 • Depression: BSI-18[af] • Anxiety: BSI-18 • Somatization: BSI-18 • Relationship satisfaction: RAS[ag] • Health-related quality of life: EQ-5D-5L (five-level EuroQol five-dimensional questionnaire)	• There was no evidence of a difference in PCL-5 pre-post change between iCBT and WL. • iCBT led to a larger RAS change than waiting, with effects in favor of waiting. • For all other secondary efficacy outcomes, we did not observe evidence for an association between score changes and iCBT in the primary analysis set.
Hoffmann et al [14] (2023, Austria)	Control website	Intervention website	Online	Not reported	Within 2 days of the patient's admission to the ICU; 30 days after T0; 90, and 365 days after admission	• PTSD: IES • Anxiety: HADS • Depression: HADS • Frequency and times of website use	• An informative website did not result in reduced PTSD and depression symptoms.
Petrinec et al [45] (2023, United States)	Usual care	A mental health app (Sanvello)	Smartphone app	Each module contains 6 to 9 lessons (15 min per lesson)	Upon enrollment (within 5 days of ICU admission); 30 and 60 days after enrollment	• Anxiety: HADS • Depression: HADS • PTSD: PCL-5 • Quality of life: SF-12[ah] • self-efficacy: MHSES[ai] • Use of the app • Satisfaction with the app • Selection of the app	• Anxiety and depression symptom severity decreased significantly over time in the intervention group but not in the control group. • Family members logged in to the app a mean of 11.4 times and spent a mean of 50.16 minutes using the app. • Participants' mean rating of their satisfaction with the mental health app was 4.19 out of 5.
Shariati et al [46] (2021, Iran)	All common interventions except web-based communication	Web-based communication, audio or video (WhatsApp)	Web-based	4 consecutive days for 10 to 15 minutes per day	Before intervention; within 24 hours after the intervention in both groups	• Stress: PSS[aj]	• After the intervention, the mean PSS-14 in the intervention group was significantly lower than that of the control group.
Shin et al [26] (2025, United States)	Standard Android tablet loaded with MyChart	The tablet also contained the MyChart Bedside app and selected games as standard in the hospital	Tablet computer app	During ICU	Baseline; extubation or ICU discharge; 1, 3, and 6 months follow-up	• Anxiety: HADS • Depression: HADS • PTSD: IES-R • Communication difficulty: FCS[ak]	

XSL•FO
RenderX

Study (year, country)	Control	Experimental				Outcomes and measurements	Results
		Intervention measurement	Intervention delivery	Intervention duration	Data collection time points		
							• No statistically significant difference was found between groups in changes in family psychological outcomes. The VidaTalk was associated with a slight to medium improvement in anxiety symptoms at 1 month. • The VidaTalk group had lower PTSD-related symptoms than the AC[al] group with a medium effect size at 1 month and a medium-to-large effect size at 3 months. • Compared with the perceived communication difficulty, although not statistically significant, the mean change in FCS score was larger for the intervention group than for AC, with a small effect size.
Suen et al [47] (2021, United States)	iPad without the tool	iPad preloaded with the Family Support Tool	Interactive website	During ICU	Baseline; 48 h of study enrollment; 5-9 days of study enrollment; 3 months	• Tool usability and acceptability • Perceived effectiveness: internally generated 11-item questionnaire • Quality of communication: QOC	• Surrogates reported that the tool was highly usable, acceptable, and effective. • Surrogates who used the tool reported higher overall quality of communication and higher quality in shared decision-making, but the difference did not reach statistical significance.
Vranceanu et al [53] (2020, United States)	Mimic the dose and duration of RT[am] without teaching RT skills	RT is an active intervention. The first 2 sessions were taught at the bedside, and the next 4 were taught after discharge	Web-based app on a smartphone	30 minutes per session	Baseline; 6 weeks; and 12 weeks	• Feasibility of recruitment and intervention delivery, credibility, and satisfaction • Depression and anxiety: HADS • PTSD: PCL-C • Mindfulness: CAMS-R[an] • Coping: MOCS-A[ao] • Dyadic interactions: DRS[ap]	• Feasibility, adherence, fidelity, satisfaction, credibility, and expectancy exceeded benchmarks set a priori. • Participation in RT was associated with statistically and clinically significant improvement between baseline and postintervention in symptoms of depression, anxiety, and PTSD. • Improvements sustained through the 12-week follow-up visit. • RT-dependent improvement in dyadic interpersonal interactions for survivors.
Watland et al [48] (2025, Norway)	Standard care	The Caregiver Pathway 4-step model	Digital assessment tool	From admission to 3 months after discharge	Baseline; 3 months	• PTSD: IES-R • Anxiety: HADS • Depression: HADS • Quality of life: SF-12 • Self-efficacy: GSE[aq] • Hope: Herth Hope Index	• The Caregiver Pathway is associated with reduced symptoms of PICS-F[ar] in terms of reduced symptoms of PTSD, anxiety, and depression for caregivers of surviving patients.

XSL•FO
RenderX

Study (year, country)	Control	Experimental				Outcomes and measurements	Results
		Intervention measurement	Intervention delivery	Intervention duration	Data collection time points		
Woo et al [49] (2024, Korea)	Daily phone calls	Daily virtual visits	Zoom app	Began on the first day after ICU admission until ICU discharge, lasting for a maximum of 7 days (each limited to 15 min)	At study enrollment after ICU admission; at the end of the study	• Satisfaction concerning ICU care and decision-making processes: FS-ICU 24[as] • Anxiety: HADS • Depression: HADS	• The FS-ICU 24 survey score was significantly higher in the virtual visitation group. • The reduction in HADS-A and HADS-D scores between the 2 time points was significantly larger in the virtual visitation group.
Xiong et al [16] (2025, China)	Standard care	Wab-WPPEP	WeChat applet	Patient's admission to the ICU until 1 month after discharge	Preoperative baseline; before transfer from the ICU; before discharge from the hospital; 1 month after discharge	• Anxiety: HADS • Depression: HADS • PTSD: IES-R • Quality of life: SF-36	• The intervention group exhibited greater improvements in anxiety, depression, and quality of life compared with the control group. • PTSD scores were also significantly lower in the intervention group.

[a]CAS: caregiver appraisal scale.
[b]MPAI-4: Mayo-Portland Adaptability Inventory-Fourth Edition.
[c]PTSD: posttraumatic stress disorder.
[d]PCL-5: PTSD (posttraumatic stress disorder) Checklist for DSM-5 (Diagnostic and Statistical Manual of Mental Disorders, [Fifth Edition]).
[e]PROMIS: Patient-Reported Outcomes Measurement Information System.
[f]HRQOL: health-related quality of life.
[g]TBI: traumatic brain injury.
[h]EF-R: education of families by routine.
[i]EF-T: education of families by interactive tabs.
[j]ICU: intensive care unit.
[k]C-DASS: Depression Anxiety Stress Scale-Chinese.
[l]CPCS: Society of Critical Care Medicine's Family Needs Assessment Questionnaire's Communication and Physical Comfort Scale.
[m]CSCS: clinician-surrogate concordance scale.
[n]MCS: medical comprehension scale.
[o]QOC: quality of communication questionnaire.
[p]HADS: Hospital Anxiety And Depression Scale.
[q]PTSS: posttraumatic stress symptom inventory.
[r]DCS: decision conflict scale.
[s]PPPC: Patient-Perceived Patient-Centeredness scale.
[t]NEST: Needs at the End-of-Life Screening Tool.
[u]IPC: interpersonal processes of care scales.
[v]PHQ-9: patient health questionnaire-9 scale.
[w]GAD-7: generalized anxiety disorder 7 scale.
[x]GCC: goal concordant of care.
[y]SAS: Self-Rating Anxiety Scale.
[z]VR: virtual reality.
[aa]IES-R: Impact of Event Scale-revised.
[ab]SF-36: 36-Item Short-Form Health Survey.
[ac]CQI: Consumer Quality Index.
[ad]iCBT: internet-based cognitive behavioral writing therapy.
[ae]WL: wait list.
[af]BSI-18: Brief Symptom Inventory-18.

XSL•FO
RenderX

[ag]RAS: Relationship Satisfaction Scale.
[ah]SF-12: 12-Item Short-Form General Health Survey.
[ai]MHSES: Mental Health Self-Efficacy Scale.
[aj]PSS: Perceived Stress Scale.
[ak]FCS: Family Communication Scale.
[al]AC: attention control.
[am]RT: Recovering Together.
[an]CAMS-R: Cognitive and Affective Mindfulness Scale-Revised.
[ao]MOCS-A: Measure of Current Status-Part A.
[ap]DRS: Dyadic Relationship Scale.
[aq]GSE: General Self-Efficacy Scale.
[ar]PICS-F: Post-Intensive Care Syndrome Family.
[as]FS-ICU: Family Satisfaction-Intensive Care Unit Questionnaire.

Control

Control groups varied in different studies; 12 studies [16,40-43,45-50,52] adapted usual or routine care, including regular education on the ICU environment, basic intensive care processes, the use of monitoring devices, ICU rules, spontaneous conversations with the ICU nurses and meetings with the physicians, effective communication with critically ill patients, and basic principles of patient care. In addition, 4 studies [14,26,51,53] used a standard app or website that provided only basic information, and 1 study [44] was a blank control (Table 2).

Outcomes and Assessment Tools

Table 2 summarizes the outcomes and measurement tools of each study. All the assessment scales used in the outcome measurement were validated. In total, 5 scales were used to assess the anxiety and depression levels of ICU patients' families. Of the total, 9 of the included studies [14,16,26,42,43,45,48,49,53] adopted the Hospital Anxiety and Depression Scale to assess anxiety and depression. Moreover, 2 studies [41,52] assessed anxiety using the Generalized Anxiety Disorder 7 scale and depression using the Patient Health Questionnaire-9 Scale. Another study [40] assessed anxiety and depression using the Depression Anxiety Stress Scale-Chinese. One study [51] used the Patient-Reported Outcomes Measurement Information System questionnaire. One study [50] used the Self-Rating Anxiety Scale to assess anxiety. One study [44] used the Brief Symptom Inventory-18. The measurement tools used for PTSD varied among studies. In total, 3 studies [41,42,52] measured the Post-Traumatic Stress Symptom Inventory, 3 studies [44,45,53] used the PTSD Checklist for *DSM-5* (Diagnostic and Statistical Manual of Mental Disorders [Fifth Edition]), and 5 studies [14,16,26,43,48] used the Impact of Event Scale-Revised. In addition, 5 studies measured QoL with 3 scales. 1 study [44] measured the Health Questionnaire of the EQ-5D-5L; 2 studies [45,48] measured the 12-Item Short-Form Health Survey, and 2 studies [16,43] measured the 36-Item Short-Form Health Survey.

Risk of Bias

Two authors (JZ and ZH) used the Cochrane Risk of Bias Assessment Tool version 2 independently to assess the methodological quality of the included studies, and the weighted κ coefficients of domains ranged between 0.55 (P=.002) and 1.00 (P<.001), indicating substantial to almost perfect interrater agreement of the independent work in this part. The results are presented in Figure 2 [14,16,26,40-53]. Specifically, all included studies were reported as randomized, but selective bias was evident because 3 studies [45,48,53] did not report allocation concealment. Due to the nature of DHIs, blinding of participants and investigators was not possible; only 1 study [14] used a double-blind design for the intervention because the structure was the same across both websites. However, almost all studies reported using intention-to-treat analysis to estimate the effect of intervention assignment, thereby reducing bias. A total of 5 studies [40,42,45,50,53] reported deviations from the intended intervention arising from the trial context, and are therefore rated as high-bias. In addition, 3 studies [26,43,52] reported incomplete data, which increased the risk of attrition bias. Furthermore, 12 studies were blinded to outcome assessors, and the other 5 studies [26,44,48,49,53] addressed bias in outcome measurement. Moreover, 2 studies [14,45] provided insufficient information to permit judgment and might have led to overestimated effects and introduced selection bias into the reported results. Finally, 4 studies were assessed as "low-risk bias," 10 as "some concerns," and 3 as "high-risk bias." However, in accordance with the consensus agreement of all the authors, the studies with a high risk of overall bias were ultimately reserved.

XSL•FO
RenderX

Figure 2. The results of the risk of bias assessment of included studies. (A) Risk of bias summary, (B) risk of bias graph [14,16,26,40-53].

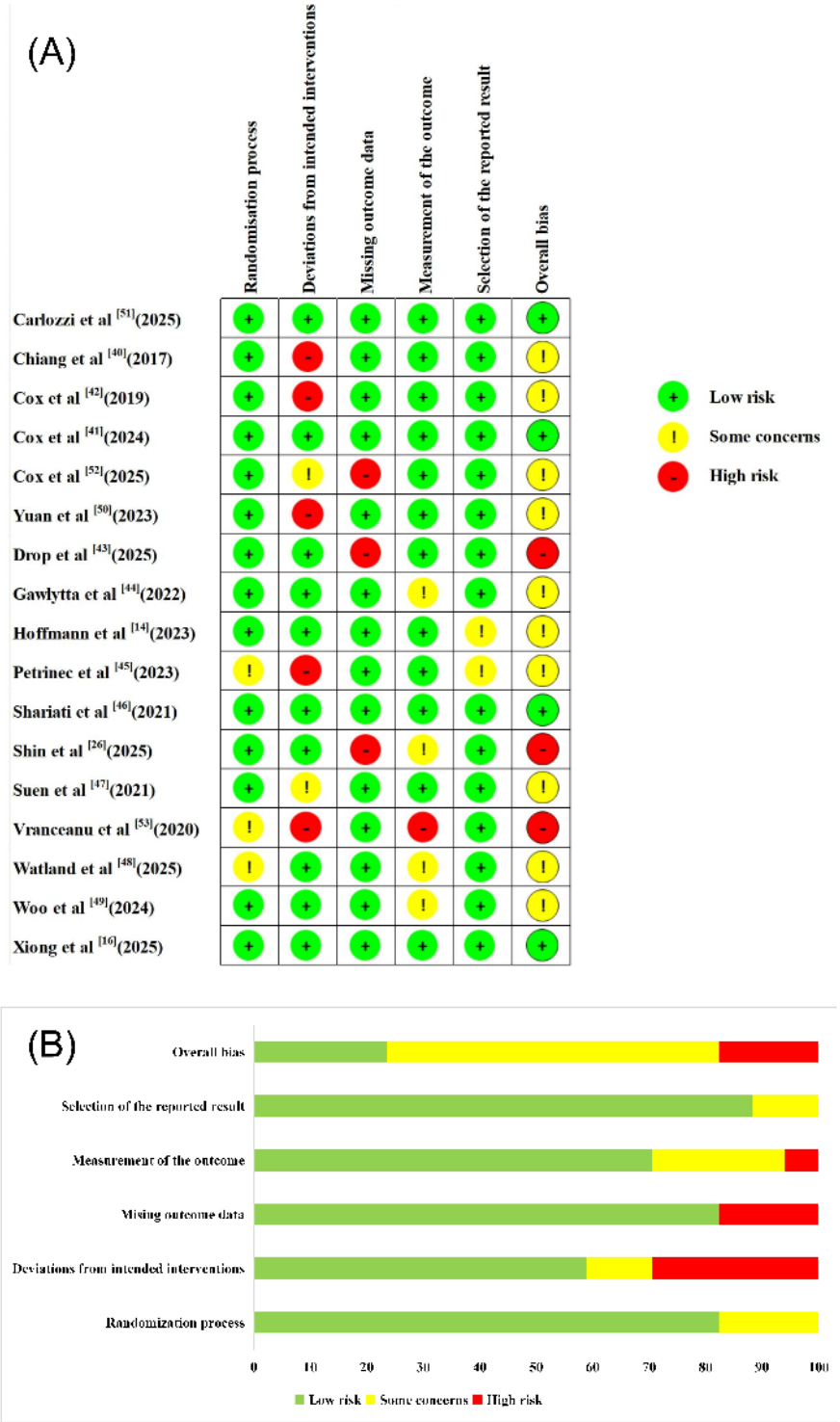

Quality of Evidence

The GRADE approach [33] was used to assess the overall certainty of the evidence. The κ coefficient in this part was 0.88 (P<.001), indicating almost perfect interrater consistency. Among the 7 outcomes analyzed, the certainty of evidence was downgraded to moderate for QoL and quality of communication (QOC). The certainty for anxiety, depression, PTSD, and self-efficacy was downgraded to low. Stress was downgraded to very low certainty of evidence (Table 3).

XSL•FO
RenderX

Table 3. GRADE[a] evidence profile for digital health intervention compared with standard care for intensive care unit family members.

Outcomes	Follow-up		Certainty of the evidence[b] (GRADE)	Relative effect (95% CI)[c]	Anticipated absolute effects	
	Participants[d], n	Studies, n			Risk with standard care	Risk difference with digital health intervention
Anxiety	1209	12 RCTs[e]	• • Low[f,g,h]	—[i]	—	SMD[j] 0.34 SD lower (0.68 lower to 0)
Depression	1181	12 RCTs	• • Low[f,g,h]	—	—	SMD 0.26 SD lower (0.52 lower to 0.01 higher)
PTSD[k]	1077	11 RCTs	• • Low[f,g,h]	—	—	SMD 0.21 SD lower (0.49 lower to 0.06 higher)
Stress	141	2 RCTs	• • Very low[g,h,l,m]	—	—	SMD 1.17 lower (3.53 lower to 1.19 higher)
Self-efficacy	204	2 RCTs	• • Lowf,h,n,o	—	—	SMD 0.01 lower (0.28 lower to 0.27 higher)
QoL[p]	424	5 RCTs	• • Moderate[f,h,n]	—	—	SMD 0.09 SD higher (0.17 lower to 0.35 higher)
Quality of communication	555	4 RCTs	• • Moderate[f,h,n]	—	—	SMD 0.15 SD higher (0.06 lower to 0.35 higher)

[a]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.

[b]High certainty (we are very confident that the true effect lies close to that of the estimate of the effect); moderate certainty (we are moderately confident in the effect estimate: the true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different); low certainty (our confidence in the effect estimate is limited: the true effect may be substantially different from the estimate of the effect); and very low certainty (we have very little confidence in the effect estimate: the true effect is likely to be substantially different from the estimate of effect).

[c]The risk in the intervention group (and its 95% CI) is based on the assumed risk in the comparison group and the relative effect of the intervention (and its 95% CI).

[d]Patient or population: intensive care unit family members; setting: intensive care unit; intervention: digital health intervention; and comparison: standard care.

[e]RCT: randomized controlled trial.

[f]Most of the included studies were assessed as some concerns or high-risk bias.

[g]I^2≥50%.

[h]Direct participants, interventions and outcomes.

[i]Not applicable.

[j]SMD: standardized mean difference.

[k]PTSD: posttraumatic stress disorder.

[l]Most of the included studies were assessed as low-risk or some concerns bias.

[m]Sample size is small and CI was wide.

[n]I^2<50%.

[o]CI was wide.

[p]QoL: quality of life.

Effect of DHIs

Psychological Outcomes

In total, 12 studies assessed the impact of DHIs on anxiety among ICU family members. The pooled results showed that DHIs did not lead to a statistically significant reduction in anxiety levels (SMD –0.34, 95% CI –0.68 to 0.00; *P*=.05). The heterogeneity among studies was substantial (I^2=80.19%; τ=0.49; τ^2=0.24), and the certainty of the evidence was rated as low. The 95% PI of the effect of DHIs on anxiety was –1.46 to 0.79, which indicates that we can be 95% confident that the true effect size from a future individual study would fall within this range (Figure 3 [16,26,40-43,45,48-50,52,53]). Similarly, 12 studies investigated the impact of DHIs on depression among ICU family members. The results revealed no significant effect (SMD –0.26, 95% CI –0.52 to 0.01; *P*=.06), with moderate heterogeneity (I^2=67.80%; τ=0.37; τ^2=0.13). The certainty of this evidence was also rated as low. A wide 95% PI (–1.11 to 0.60) suggests that future studies could show effects ranging

XSL•FO
RenderX

from a large reduction to a modest increase in depression levels (Figure 4 [14,16,26,40-43,45,48,49,52,53]). In total, 11 studies evaluated the effects of DHIs on PTSD among ICU family members. The meta-analysis demonstrated no significant difference in PTSD scores between groups (SMD –0.21, 95% CI –0.49 to 0.06; *P*=.11), with substantial heterogeneity (I^2=69.98%; τ=0.34; τ^2=0.12). The certainty of this evidence was also rated as low. The 95% PI (–1.03 to 0.60) indicates that the actual effect of DHIs on PTSD could vary widely, indicating potential for both beneficial and harmful effects (Figure 5 [14,16,26,41-45,48,52,53]). Furthermore, regarding stress and self-efficacy, which were evaluated in only 2 studies each, we discussed them qualitatively. 2 studies reported different results regarding stress: 1 study [40] found no effect of DHI on stress. In contrast, another [46] reported a significant effect of DHI in reducing stress among ICU family members. Neither study [45,48] on self-efficacy found a significant benefit from DHIs. The certainty of the 2 pieces of evidence was rated as very low and low.

Figure 3. Forest plot of the effectiveness of digital health technology on anxiety. SMD: standardized mean difference [16,26,40-43,45,48-50,52,53].

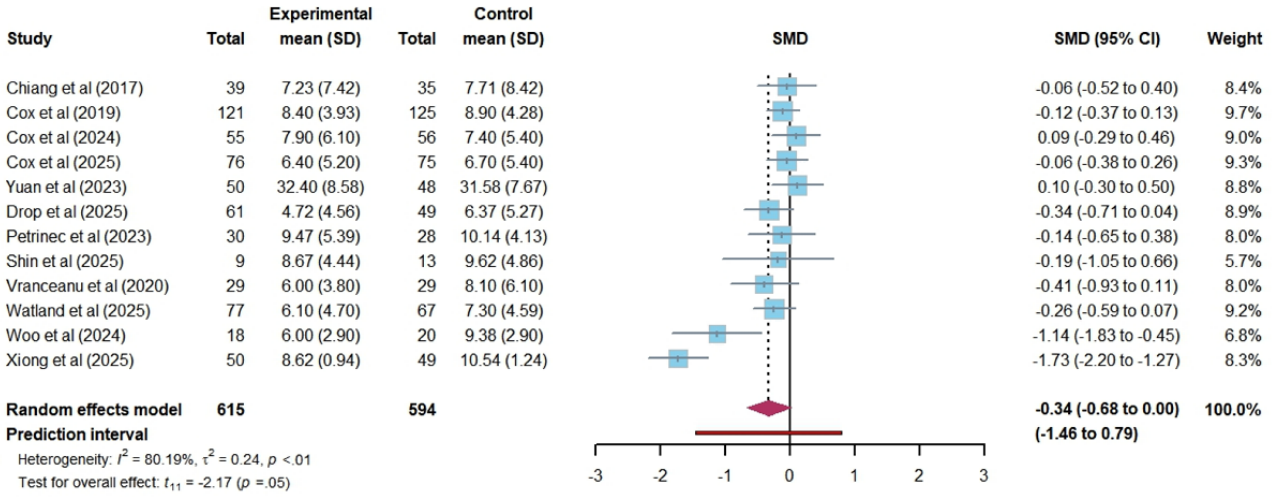

Figure 4. Forest plot of the effectiveness of digital health technology on depression. SMD: standardized mean difference [14,16,26,40-43,45,48,49,52,53].

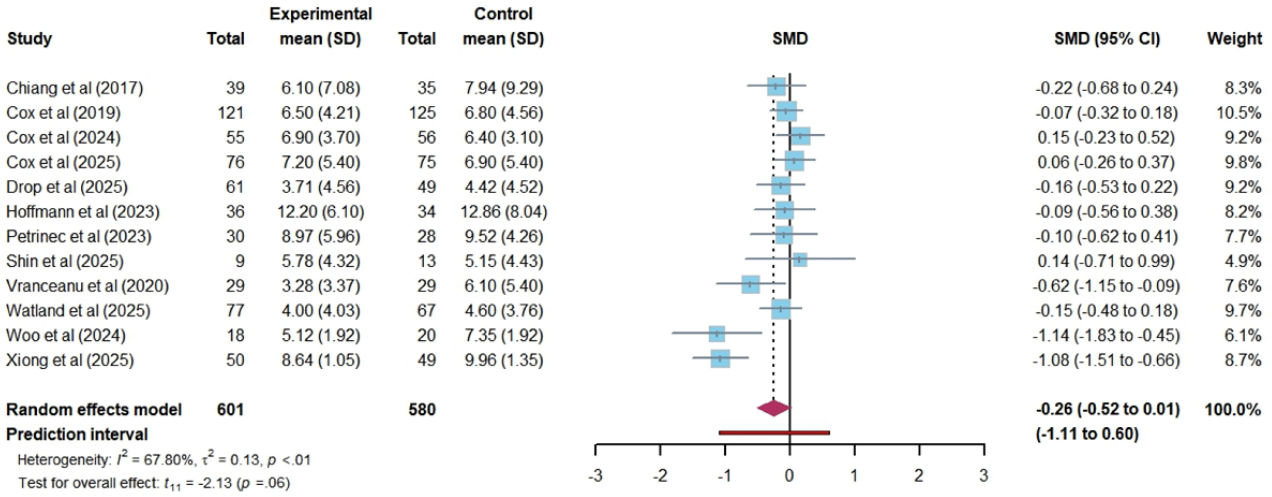

XSL•FO
RenderX

Figure 5. Forest plot of the effectiveness of digital health technology on post-traumatic stress disorder. SMD: standardized mean difference [14,16,26,41-45,48,52,53].

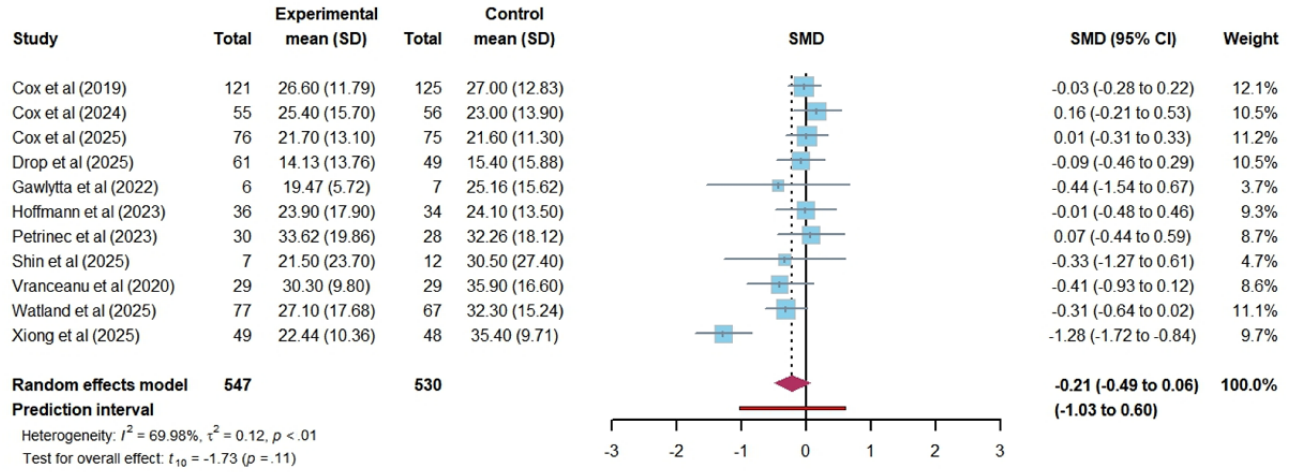

QoL

In total, 5 studies assessed the impact of DHIs on QoL among ICU family members. A random-effect meta-analysis indicated no statistically significant improvement in QoL (SMD 0.09, 95% CI –0.10 to 0.28; *P*=.36). Despite the absence of heterogeneity (I^2=0.00%; τ=0.14; τ^2=0.02), the actual effect of DHIs on QoL in the future falls within the broader PI range (–0.41 to 0.60), suggesting that future outcome could vary significantly across different populations and settings (Figure 6 [16,43-45,48]). The certainty of this evidence was rated as moderate.

Figure 6. Forest plot of the effectiveness of digital health technology on quality of life. SMD: standardized mean difference [16,43-45,48].

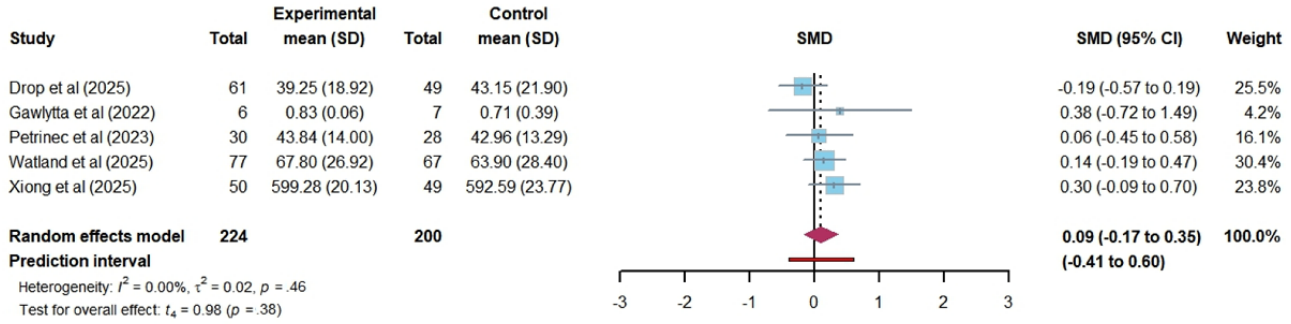

QOC

In total, 4 studies evaluated the effect of DHIs on the QOC among ICU family members. The pooled analysis showed no statistically significant difference between groups (SMD 0.14; 95% CI –0.03 to 0.31; *P*=.10; Figure 7 [41,42,47,52]). Despite the absence of heterogeneity (I^2=0.00%; τ=0.08; τ^2=0.01), the actual effect of DHIs on QoL in the future falls within the broader PI range (–0.26 to 0.55), suggesting that future outcomes could vary significantly across different populations and settings. The certainty of this evidence was rated as moderate.

XSL•FO
RenderX

Figure 7. Forest plot of the effectiveness of digital health technology on the quality of communication. SMD: standardized mean difference [41,42,47,52].

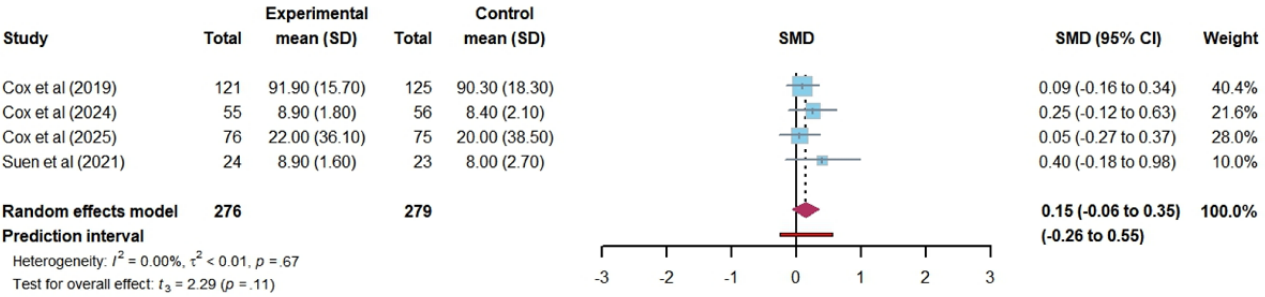

Subgroup Analysis

We conducted subgroup analyses for the three outcome indicators of anxiety, depression, and PTSD based on potential covariates, including patients' primary diagnosis, relationship with patients, and type of DHT (Table S1 and Figures S1-S3 in Multimedia Appendix 4). Regarding anxiety, the results of subgroup analyses showed that there was no significant difference (P>.05) in the patients' primary diagnosis (χ^2_2=2.60; P=.27), relationship with patients (χ^2_1=0.23; P=.63), and type of DHT (χ^2_2=1.43; P=.49), suggesting that these factors may not be related to the change of intervention effect. Within the subgroup, family members of medical patients showed a slight improvement trend (SMD –0.14, 95% CI –0.26 to –0.01; P=.04). From the perspective of heterogeneity sources, the patients' primary diagnosis and the type of DHT may be potential sources of heterogeneity. Regarding depression, subgroup analysis showed significant differences between subgroups with different patients' primary diagnosis (χ^2_2=29.17; P<.01). Among the surgical patients' family members (3 studies), depression scores in the intervention group were significantly reduced (SMD –0.94, 95% CI –1.64 to –0.24, t_2=–5.80; P=.03). In studies focusing on medical patients, no significant effect of the intervention on family members depressive symptoms was observed (SMD –0.04, 95% CI –0.14 to 0.05, t_7=–1.11; P=.30). There was no statistical significance (P>.05) in the subgroup differences in relationship with patients (χ^2_1=0.38; P=.54) and type of DHT (χ^2_2=3.11; P=.21). Among the subgroups of type of DHT, the effect size of the education-based DHT group (4 studies) had a more minor statistical significance (SMD –0.11, 95% CI –0.21 to –0.01, t_3=–3.47; P=.04). Patients' primary diagnosis and type of DHT may be significant sources of heterogeneity. Regarding PTSD, the differences between subgroups in the patients' primary diagnosis (χ^2_1=3.29; P=.07), relationship with patients (χ^2_1=1.30; P=.25), and type of DHT (χ^2_2=2.75; P=.25) did not reach statistical significance (P>.05). However, looking at the sources of heterogeneity, these factors may be the primary drivers. Due to the limited number of studies in some subgroups, these results should be interpreted with caution.

Sensitivity Analysis

For anxiety, it had no impact on the statistical significance of the results after removing the study conducted by Xiong et al [16] (SMD –0.18, 95% CI –0.38 to 0.01; P=.06; Figure S4 in Multimedia Appendix 4), which decreased heterogeneity to 27.36%. Regarding depression, removing the study by Xiong et al [16] (SMD –0.17, 95% CI –0.38 to 0.05; P=.11; Figure S5 in Multimedia Appendix 4) decreased heterogeneity to 36.09%, but it did not alter the statistical significance of the results. For PTSD, sensitivity analyses demonstrated that removing the study by Xiong et al [16] (SMD –0.08, 95% CI –0.21 to 0.05; Figure S6 in Multimedia Appendix 4) resolved heterogeneity, although it did not alter the statistical significance of the results. For QoL, statistical significance was observed only after removing the study by Drop et al [43] (SMD 0.19, 95% CI 0.01 to –0.37; P=.04; Figure S7 in Multimedia Appendix 4). Regarding QOC, the effect of DHIs remained stable across studies, indicating robust findings.

Bias Assessment

Visual inspection of the funnel plots for anxiety, depression, and PTSD showed approximate symmetry (Figures S1-S3 in Multimedia Appendix 5). Egger tests did not provide statistically significant evidence of small-study effects for anxiety (P=.24), depression (P=.20), and PTSD (P=.44; Table S1 in Multimedia Appendix 5). However, we note that small-study effects can arise from various sources beyond publication bias, such as heterogeneity or methodological differences. For other outcomes, there were insufficient numbers of systematic reviews (n<10) to create funnel plots.

Discussion

Principal Findings

In total, 17 studies were included, revealing that compared with control groups, DHIs may not confer a significant benefit in reducing anxiety, depression, PTSD, or stress among ICU family members. Moreover, they also did not notably improve

XSL•FO
RenderX

self-efficacy, QoL, or communication quality. Given the low certainty of the evidence for the proposed association, the substantial study heterogeneity, and the overall risk of bias in the included studies, the findings should be interpreted with caution. Subgroup analyses indicated that patients' primary diagnosis may influence the intervention's effect, with the benefits being more pronounced for families of surgical patients. Patients' primary diagnosis, relationship with patients, and type of DHT may be a significant source of heterogeneity. Given the PI value, adverse effects of the intervention cannot be ruled out in future studies, and more RCTs are needed to provide robust results.

The observed effect of DHIs on anxiety aligns with previous research concerning improvements in anxiety among caregivers of patients with cancer [23]. The ICU environment is characterized by high stress, restricted access, asymmetric information, and uncertain prognosis, which can significantly increase the psychological burden of family members [54,55]. DHI compensates for the absence of in-person support by using mobile apps, network platforms, and virtual visits [56]. It provides timely updates on diseases, health education, and emotional support for family members, thereby reducing uncertainty and equipping them with knowledge and skills that, in turn, alleviate anxiety symptoms [57]. However, the clinical significance of this finding is tempered by its statistical marginality (P=.05), a low certainty of the evidence, and risk of bias in some studies, which significantly reduces the credibility of this potential effect. It is critical to underscore that substantial heterogeneity (I^2=80.19%) yields a wide 95% PI (–1.46 to 0.79), suggesting that the distribution of effects on anxiety is unpredictable. This finding should only be considered a preliminary exploratory conclusion and requires further validation through additional research. No significant effects were observed on depression, PTSD, stress, self-efficacy, QoL, and QOC. It is important to note that the certainty of the evidence for these outcomes is low, with an associated risk of bias, so the reasons for the lack of significant observed effects should be considered from multiple perspectives. From a clinical view, the onset of PTSD is often delayed and complex, typically manifesting months to a year post discharge, making it challenging for short-term interventions and follow-ups to capture long-term effects [14,44]. Additionally, stress, self-efficacy, QOC, and QoL are inherently multidimensional indicators influenced by confounding factors, including economic conditions and social support, beyond psychological state [45,58,59]. From a methodological perspective, existing studies have limitations, including risk of bias, low certainty of the evidence, and imprecision—particularly given that only 2 studies were included for stress and self-efficacy. Current evidence is insufficient to identify the potential benefits or harms of DHIs for these outcomes, and subsequent research is very likely to alter these estimates.

The observed significant heterogeneity across outcomes prompted extensive subgroup analyses. While these analyses pointed to several potential moderators—such as patients' primary diagnosis and type of DHT for depression—their interpretability is limited. For example, the notable reduction in depression among surgical patients (SMD –0.94, 95% CI –1.64 to –0.24), derived from only 3 studies, may reflect a type I error or chance variation rather than an actual differential effect. Similarly, although traditional factors such as family members' sex and relationship with patients did not reach statistical significance in our analyses, existing literature consistently indicates that female caregivers, especially spouses, often endure greater caregiving stress and more persistent psychological distress [14,60,61]. The overall heterogeneity remained largely unexplained by subgrouping, further underscoring the robustness of these comparisons. Thus, these analyses should not be viewed as confirming specific moderators but rather as highlighting priority variables for future, adequately powered studies.

The limited effectiveness of DHIs may also be attributed to the design of the interventions. Most existing interventions are information-centric, providing unidirectional education rather than incorporating structured psychotherapy, skills training, or facilitating two-way communication [42,62]. Such a format appears insufficient for translating into actual skill enhancement, family interaction, or overall well-being improvement [43]. Furthermore, some studies reported positive subjective feedback from families despite nonsignificant quantitative outcomes, indicating that current measurement tools may lack sensitivity to nuanced experiential gains [40,43,44].

Moving forward, future research should use mixed methods designs to capture subjective experience and identify needs of high-risk subgroups; tailor interventions according to patients' diagnosis, cultural context, and characteristics of caregiver; develop more interactive and personalized DHIs formats, such as artificial intelligence–based scenario simulations, two-way communication platforms, and longitudinal support programs; extend follow-up duration and use multimodal outcome measures that combine patients and family members reported experiences with clinically relevant indicators; and explore hybrid online-offline interventions to support continuity of care across different phases of the ICU trajectory.

Strengths and Limitations

This systematic review has several strengths. First, to our knowledge, this is the first meta-analysis to synthesize and appraise the evidence about the effects of DHIs on psychological and related outcomes among family members in the ICU. Second, we strictly followed PRISMA guidelines to strengthen transparency and clarity of reporting. Third, we searched across 9 databases from inception to October 11, 2025, reflecting the most recent advancements in the field of DHI. Moreover, we used the Cochrane RoB 2 and GRADE methodology for quality assessment. Finally, subgroup analyses, sensitivity analyses, and bias assessment were conducted to validate the stability of the results and explore potential sources of heterogeneity.

However, there were several limitations to this review. First, the considerable heterogeneity among studies, as evidenced by the wide PIs for our primary outcomes, indicates uncertainty about the intervention effect. Discrepancies in participants' characteristics and intervention protocols could lead to inconsistencies and bias in the study findings. Second, the interventions included in the review ranged widely, from simple information websites and SMS text messages to VR- and

XSL•FO
RenderX

app-based psychotherapy, and intervention duration varied. Pooling across these fundamentally different modalities may dilute or obscure which components work. Third, while blinding interventionists poses difficulties in DHT care settings, the absence of blinding for study participants and interventionists in most of the included studies introduced risks of bias, resulting in a methodological error. Fourth, the search approach has certain limitations as it exclusively incorporated papers published in English or Chinese, rendering findings from articles in other languages unknown and constraining the global generalizability of the results. Fifth, while no significant small-study effects were detected, the ability of these tests may be constrained by the limited number of included studies. Hence, we cannot conclusively dismiss potential publication bias. Sixth, the majority of the selected trials were conducted in developed countries. This bias may result from the restricted access to technology and the low practicality of technology-based interventions in low-income countries. Seventh, the results regarding the effect of DHIs on psychological outcomes, QoL, and communication quality are inconclusive because of the limited number of studies and the wide PIs. Negligible effects are likely from poor statistical power, not a genuine absence of an intervention effect.

Implications of Nursing Practice

Given the current limitations and low certainty of the evidence, our findings have limited direct implications for clinical practice but provide valuable insights for future research. ICU nurses, as the medical personnel with the most frequent interactions with patients and their families, are ideally positioned to lead DHIs [23]. They can facilitate online meetings involving physicians, patients, and families to dismantle information barriers, guided by shared decision-making theory. More importantly, nurses can leverage digital platforms to enable personalized communication. This nurse-driven, tailored digital communication strategy not only provides adequate informational support to mitigate their negative emotions but also significantly improves communication quality and overall QoL [63]. The subgroup analyses suggest that DHIs' effectiveness may not be universal but can be enhanced by tailoring interventions to specific family or patient characteristics. In the future, using technologies such as videoconferencing, virtual simulations, and advanced language models, we can deliver more comprehensive and personalized psychological support interventions led by nurses via the internet and mobile devices. Simultaneously, it is necessary to advocate large-sample, multicenter, high-quality, and reproducible RCTs to provide more robust data to support DHI's findings on the psychological effects and QoL of ICU family members in China.

Conclusions

In conclusion, this study provides the first comprehensive synthesis of RCT evidence on DHIs for ICU family members, revealing significant uncertainty in the estimated effects. Despite its methodological rigor and use of PIs, the conclusions of this review are constrained by the intrinsically heterogeneous and low certainty of the existing evidence. Unlike previous reviews, this work integrates more recent studies and focuses explicitly on the family members of ICU patients. Future research should prioritize the development of theory-informed, personalized, and interactive DHIs. Incorporating mixed methods designs and targeting high-risk subgroups would be particularly valuable. Although direct clinical recommendations are limited, DHIs hold potential for enabling more proactive and individualized communication strategies for families of ICU patients.

Acknowledgments

We extend our gratitude to the librarians for their invaluable assistance and expertise in conducting the literature search for this study.

We essentially wrote the paper ourselves. Since the authors are not native speakers, we used ChatGPT (OpenAI) to brainstorm verbs and adjectives to express our opinions better.

Funding

This work was supported by the Zhejiang Provincial Natural Science Foundation of China under Grant LTGY24H090006, and the General Program of the Zhejiang Provincial Health Commission under Grant 2025KY904. The funding sources had no role in the design and conduct of the study, the collection, management, analysis, and interpretation of the data, the preparation, review, or approval of the manuscript, or the decision to submit the manuscript for publication. The contents of this paper are solely the responsibility of the authors and do not necessarily represent those of the Zhejiang Provincial Department of Science and Technology or the Zhejiang Provincial Health Commission.

Data Availability

All data used in this study are available in the article and supplementary material. Further inquiries could be directed to the first author.

Authors' Contributions

Chuchu Z: writing - original draft, validation, software, methodology, formal analysis, data curation, reviewing and editing
WS: writing - review and editing, validation, software

XSL•FO
RenderX

WJ: writing - review & editing, software, methodology, data curation
YS: writing - conceptualization, software, methodology, data curation
ZH: writing - formal analysis, data curation
RY: writing - methodology
JY: writing - methodology
JZ: writing - methodology
Chaoyi Z: writing - data curation
YZ: writing - conceptualization, funding acquisition
XG: writing - review & editing, supervision, funding acquisition

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA 2020 checklist.
[DOC File , 124 KB - jmir_v28i1e83294_app1.doc]

Multimedia Appendix 2
PRISMA-S checklist.
[DOCX File , 17 KB - jmir_v28i1e83294_app2.docx]

Multimedia Appendix 3
Search strategy.
[DOCX File , 35 KB - jmir_v28i1e83294_app3.docx]

Multimedia Appendix 4
Subgroup analysis by potential covariates.
[DOCX File , 3412 KB - jmir_v28i1e83294_app4.docx]

Multimedia Appendix 5
Publication bias.
[DOCX File , 33 KB - jmir_v28i1e83294_app5.docx]

References

1. Jiang W, Ni Q, Zhang C, Dong Y, Yi J, Yan R, et al. Factors influencing quality of life after intensive care: A systematic review and meta-analysis. Crit Care Med 2025;53(9):e1803-e1814. [doi: 10.1097/CCM.0000000000006770] [Medline: 40662879]
2. Hwang D, Oczkowski S, Lewis K, Birriel B, Downar J, Farrier C, et al. Society of critical care medicine guidelines on family-centered care for adult ICUs: 2024. Crit Care Med 2025;53(2):e465-e482. [doi: 10.1097/CCM.0000000000006549] [Medline: 39982184]
3. Shirasaki K, Hifumi T, Nakanishi N, Nosaka N, Miyamoto K, Komachi M, et al. Postintensive care syndrome family: A comprehensive review. Acute Med Surg 2024;11(1):e939 [FREE Full text] [doi: 10.1002/ams2.939] [Medline: 38476451]
4. Sui W, Gong X, Qiao X, Zhang L, Cheng J, Dong J, et al. Family members' perceptions of surrogate decision-making in the intensive care unit: A systematic review. Int J Nurs Stud 2023;137:104391. [doi: 10.1016/j.ijnurstu.2022.104391] [Medline: 36442321]
5. Lin Z, Huang Q, Ya R, Huang H, Bai Y, Lin W. Survey on sleep, mental health, and quality of life among family members of critically ill patients in rehabilitation hospitalization. Chin J Rehabil Med 2025;40(7):1071-1075.
6. Shirasaki K, Ohbe H, Hifumi T, Okajima M, Otani N, Yasunaga H. Association between patients' admission to the ICU and psychological disorders in their families: a retrospective matched-pair cohort study. Crit Care Med 2025;53(10):e1963-e1972. [doi: 10.1097/CCM.0000000000006793] [Medline: 40736355]
7. Abdul Halain A, Tang LY, Chong MC, Ibrahim NA, Abdullah KL. Psychological distress among the family members of intensive care unit (ICU) patients: A scoping review. J Clin Nurs 2022;31(5-6):497-507. [doi: 10.1111/jocn.15962] [Medline: 34254377]
8. de Souza JMB, Miozzo AP, da Rosa Minho Dos Santos R, Mocellin D, Rech GS, Trott G, et al. Long-term effects of flexible visitation in the intensive care unit on family members' mental health: 12-month results from a randomized clinical trial. Intensive Care Med 2024;50(10):1614-1621. [doi: 10.1007/s00134-024-07577-3] [Medline: 39172240]

XSL•FO
RenderX

9. Davidson JE, Jones C, Bienvenu OJ. Family response to critical illness: postintensive care syndrome-family. Crit Care Med 2012;40(2):618-624. [doi: 10.1097/CCM.0b013e318236ebf9] [Medline: 22080636]
10. Huang W, Gao Y, Zhou L, Xiao X, Xu H, Lu L, et al. Effects of ICU diaries on psychological disorders and sleep quality in critically ill patients and their family members: A systematic review and meta-analysis. Sleep Med 2024;122:84-91. [doi: 10.1016/j.sleep.2024.08.002] [Medline: 39137664]
11. Nielsen AH, Angel S. How diaries written for critically ill influence the relatives: a systematic review of the literature. Nurs Crit Care 2016;21(2):88-96. [doi: 10.1111/nicc.12158] [Medline: 26487425]
12. McPeake J, Hirshberg E, Christie L, Drumright K, Haines K, Hough C, et al. Models of peer support to remediate post-intensive care syndrome: a report developed by the society of critical care medicine thrive international peer support collaborative. Crit Care Med 2019;47(1):e21-e27 [FREE Full text] [doi: 10.1097/CCM.0000000000003497] [Medline: 30422863]
13. Bohart S, Møller AM, Andreasen AS, Waldau T, Lamprecht C, Thomsen T. Effect of patient and family centred care interventions for adult intensive care unit patients and their families: a systematic review and meta-analysis. Intensive Crit Care Nurs 2022;69:103156. [doi: 10.1016/j.iccn.2021.103156] [Medline: 34753631]
14. Hoffmann M, Jeitziner M, Riedl R, Mueller G, Peer A, Bachlechner A, et al. Effects of an online information tool on post-traumatic stress disorder in relatives of intensive care unit patients: a multicenter double-blind, randomized, placebo-controlled trial (ICU-Families-Study). Intensive Care Med 2023;49(11):1317-1326. [doi: 10.1007/s00134-023-07215-4] [Medline: 37870597]
15. Münchenberg PS, Yessimova D, Panteli D, Kurth T. Digital health interventions for informal family caregivers of people with first-episode psychosis: systematic review on user experience and effectiveness. JMIR Ment Health 2024;11:e63743 [FREE Full text] [doi: 10.2196/63743] [Medline: 39607998]
16. Xiong J, Wu T, Cai H, He J, Wang T, Lin X, et al. Effects of wab-WPPEP on post-intensive care syndrome-family in family caregivers of heart valve replacement patients: A randomized controlled trial. Intensive Crit Care Nurs 2025;86:103817 [FREE Full text] [doi: 10.1016/j.iccn.2024.103817] [Medline: 39426206]
17. NA. Global Strategy on Digital Health 2020-2025. Geneva: World Health Organization; 2021.
18. Zhao Q, Chen X, Zhang X, Wang S. Research progress and hotspot analysis in the digital health field. China Digit Med 2022;17(4):37-44.
19. Dong M, Huang Y, Zhang X, Wu L, Ghazal L, Cao F. Effects of psychological intervention based on digital health technology on psychological outcomes in cancer survivors: systematic review and meta-analysis. Psychooncology 2025;34(5):e70166. [doi: 10.1002/pon.70166] [Medline: 40325624]
20. Guo G, Lan M, Zeng F, Liang J, Guo L, Cai P, et al. Effects of remote interventions with digital health technologies in lung transplant patients:a meta-analysis. Chin J Nurs 2024;59(11):1389-1396.
21. Shaffer KM, Turner KL, Siwik C, Gonzalez BD, Upasani R, Glazer JV, et al. Digital health and telehealth in cancer care: a scoping review of reviews. Lancet Digit Health 2023;5(5):e316-e327 [FREE Full text] [doi: 10.1016/S2589-7500(23)00049-3] [Medline: 37100545]
22. Lau CKY, Saad A, Camara B, Rahman D, Bolea-Alamanac B. Acceptability of digital mental health interventions for depression and anxiety: systematic review. J Med Internet Res 2024;26:e52609 [FREE Full text] [doi: 10.2196/52609] [Medline: 39466300]
23. Zhang Y, Flannery M, Zhang Z, Underhill-Blazey M, Bobry M, Leblanc N, et al. Digital health psychosocial intervention in adult patients with cancer and their families: systematic review and meta-analysis. JMIR Cancer 2024;10:e46116 [FREE Full text] [doi: 10.2196/46116] [Medline: 38315546]
24. Shin Y, Kim SK, Kim Y, Go Y. Effects of app-based mobile interventions for dementia family caregivers: a systematic review and meta-analysis. Dement Geriatr Cogn Disord 2022;51(3):203-213. [doi: 10.1159/000524780] [Medline: 35609526]
25. Leng M, Zhao Y, Xiao H, Li C, Wang Z. Internet-based supportive interventions for family caregivers of people with dementia: systematic review and meta-analysis. J Med Internet Res 2020;22(9):e19468 [FREE Full text] [doi: 10.2196/19468] [Medline: 32902388]
26. Shin JW, Tan A, Tate J, Balas M, Dabelko-Schoeny H, Happ MB. Preliminary efficacy of the vidatalk communication application on family psychological symptoms in the intensive care unit: A pilot study. Heart Lung 2025;70:14-22. [doi: 10.1016/j.hrtlng.2024.11.001] [Medline: 39546958]
27. Hariton E, Locascio JJ. Randomised controlled trials - the gold standard for effectiveness research: Study design: randomised controlled trials. BJOG 2018;125(13):1716 [FREE Full text] [doi: 10.1111/1471-0528.15199] [Medline: 29916205]
28. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
29. Julian H, James T, Jacqueline C, Miranda C, Tianjing L, Matthew P, et al. Cochrane Handbook for Systematic Reviews of Interventions (Current Version). 2024. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current [accessed 2025-07-27]
30. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159-174. [Medline: 843571]

XSL•FO
RenderX

31. RoB 2: A revised cochrane risk-of-bias tool for randomized trials. Cochrane bias. 2019. URL: https://methods.cochrane.org/bias/resources/rob-2-revised-cochrane-risk-bias-tool-randomized-trials [accessed 2025-07-27]
32. Sterne JAC, Savovi J, Page MJ, Elbers RG. RoB 2: A revised tool for assessing risk of bias in randomised trials. BMJ. 2019. URL: https://www.bmj.com/content/366/bmj.l4898.long [accessed 2025-07-27]
33. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64(4):383-394. [doi: 10.1016/j.jclinepi.2010.04.026] [Medline: 21195583]
34. IntHout J, Ioannidis JPA, Borm GF. The hartung-knapp-sidik-jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard dersimonian-laird method. BMC Med Res Methodol 2014;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
35. Ng AYM, Takemura N, Xu X, Smith R, Kwok JY, Cheung DST, et al. The effects of advance care planning intervention on nursing home residents: A systematic review and meta-analysis of randomised controlled trials. Int J Nurs Stud 2022;132:104276. [doi: 10.1016/j.ijnurstu.2022.104276] [Medline: 35667145]
36. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods 2010;1(2):97-111. [doi: 10.1002/jrsm.12] [Medline: 26061376]
37. Borenstein M. How to understand and report heterogeneity in a meta-analysis: The difference between I-squared and prediction intervals. Integr Med Res 2023;12(4):101014 [FREE Full text] [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
38. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
39. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997;315(7109):629-634 [FREE Full text] [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
40. Chiang VCL, Lee RLP, Ho FM, Leung CK, Tang YP, Wong WS, et al. Fulfilling the psychological and information need of the family members of critically ill patients using interactive mobile technology: A randomised controlled trial. Intensive Crit Care Nurs 2017;41:77-83. [doi: 10.1016/j.iccn.2017.03.006] [Medline: 28438476]
41. Cox CE, Ashana DC, Riley IL, Olsen MK, Casarett D, Haines KL, et al. Mobile application-based communication facilitation platform for family members of critically ill patients: a randomized clinical trial. JAMA Netw Open 2024;7(1):e2349666 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.49666] [Medline: 38175648]
42. Cox CE, White DB, Hough CL, Jones DM, Kahn JM, Olsen MK, et al. Effects of a personalized web-based decision aid for surrogate decision makers of patients with prolonged mechanical ventilation: a randomized clinical trial. Ann Intern Med 2019;170(5):285-297 [FREE Full text] [doi: 10.7326/M18-2335] [Medline: 30690645]
43. Drop DLQ, Vlake JH, Wils E, Van Bommel J, Jung C, Hilling DE, et al. Effect of an intensive care unit virtual reality intervention on relatives´ mental health distress: a multicenter, randomized controlled trial. Crit Care 2025;29(1):62 [FREE Full text] [doi: 10.1186/s13054-025-05281-2] [Medline: 39910674]
44. Gawlytta R, Kesselmeier M, Scherag A, Niemeyer H, Böttche M, Knaevelsrud C, et al. Internet-based cognitive-behavioural writing therapy for reducing post-traumatic stress after severe sepsis in patients and their spouses (REPAIR): results of a randomised-controlled trial. BMJ Open 2022;12(3):e050305 [FREE Full text] [doi: 10.1136/bmjopen-2021-050305] [Medline: 35264337]
45. Petrinec AB, Wilk C, Hughes JW, Zullo MD, George RL. Self-care mental health app intervention for post-intensive care syndrome-family: A randomized pilot study. Am J Crit Care 2023;32(6):440-448. [doi: 10.4037/ajcc2023800] [Medline: 37907376]
46. Shariati E, Dadgari A, Talebi SS, Mahmoodi Shan GR, Ebrahimi H. The effect of the web-based communication between a nurse and a family member on the perceived stress of the family member of patients with suspected or confirmed covid-19: a parallel randomized clinical trial. Clin Nurs Res 2021;30(7):1098-1106. [doi: 10.1177/10547738211017688] [Medline: 34044625]
47. Suen AO, Butler RA, Arnold RM, Myers B, Witteman HO, Cox CE, et al. A pilot randomized trial of an interactive web-based tool to support surrogate decision makers in the intensive care unit. Ann Am Thorac Soc 2021;18(7):1191-1201 [FREE Full text] [doi: 10.1513/AnnalsATS.202006-585OC] [Medline: 33326348]
48. Watland S, Solberg Nes L, Ekeberg Ø, Rostrup M, Hanson E, Ekstedt M, et al. The caregiver pathway intervention can contribute to reduced post-intensive care syndrome among family caregivers of icu survivors: a randomized controlled trial. Crit Care Med 2025;53(3):e555-e566. [doi: 10.1097/CCM.0000000000006546] [Medline: 39718436]
49. Woo H, Oh S, Lim L, Im H, Lee H, Ha E, et al. Efficacy of virtual visitation in ICU during COVID-19 pandemic: The ICU visits randomized controlled trial. Crit Care Med 2024;52(12):e593-e603. [doi: 10.1097/CCM.0000000000006429] [Medline: 39283178]
50. Yuan C, Xiao Y, Wang F, Wang Y, Wang Y, Lin F. The effect of video visitation on intensive care unit patients and family members outcomes during the COVID-19 pandemic: A randomised controlled trial. Intensive Crit Care Nurs 2023;76:103394 [FREE Full text] [doi: 10.1016/j.iccn.2023.103394] [Medline: 36731263]

XSL•FO
RenderX

51. Carlozzi NE, Troost JP, Sen S, Choi SW, Wu Z, Miner JA, et al. Improving outcomes for care partners of individuals with traumatic brain injury: results for a mhealth randomized control trial of the careqol app. Arch Phys Med Rehabil 2025;106(4):548-561. [doi: 10.1016/j.apmr.2024.12.022] [Medline: 39798892]
52. Cox CE, Ashana DC, Dempsey K, Olsen MK, Parish A, Casarett D, et al. Mobile app-facilitated collaborative palliative care intervention for critically ill older adults: a randomized clinical trial. JAMA Intern Med 2025;185(2):173-183. [doi: 10.1001/jamainternmed.2024.6838] [Medline: 39680398]
53. Vranceanu A, Bannon S, Mace R, Lester E, Meyers E, Gates M, et al. Feasibility and efficacy of a resiliency intervention for the prevention of chronic emotional distress among survivor-caregiver dyads admitted to the neuroscience intensive care unit: a randomized clinical trial. JAMA Netw Open 2020;3(10):e2020807 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.20807] [Medline: 33052404]
54. Wen J, Hu Z, Yang S, Huang Y, Zhang L, Zhong C, et al. A qualitative study on experience of post-intensive care syndrome-family symptoms. Med Philos 2024;45(19):41-46.
55. Kean S, Donaghy E, Bancroft A, Clegg G, Rodgers S. Theorising survivorship after intensive care: A systematic review of patient and family experiences. J Clin Nurs 2021;30(17-18):2584-2610. [doi: 10.1111/jocn.15766] [Medline: 33829568]
56. Aghaie B, Anoosheh M, Foroughan M, Mohammadi E, Kazemnejad A. A whirlpool of stress in families of intensive care unit patients: a qualitative multicenter study. Crit Care Nurse 2021;41(3):55-64. [doi: 10.4037/ccn2021322] [Medline: 34061191]
57. Anders S, Albert R, Miller A, Weinger MB, Doig AK, Behrens M, et al. Evaluation of an integrated graphical display to promote acute change detection in ICU patients. Int J Med Inform 2012;81(12):842-851 [FREE Full text] [doi: 10.1016/j.ijmedinf.2012.04.004] [Medline: 22534099]
58. Garcia ACM, Camargos Junior JB, Sarto KK, Silva Marcelo CAD, Paiva EMDC, Nogueira DA, et al. Quality of life, self-compassion and mindfulness in cancer patients undergoing chemotherapy: A cross-sectional study. Eur J Oncol Nurs 2021;51:101924. [doi: 10.1016/j.ejon.2021.101924] [Medline: 33610930]
59. Zhang S, Li J, Zhang Y, Hu X. Effects of advance care planning for patients with advanced cancer: A meta-analysis of randomized controlled studies. Int J Nurs Stud 2025;168:105096. [doi: 10.1016/j.ijnurstu.2025.105096] [Medline: 40378812]
60. Liu S, Yang X, Xie W, Huang S, Xiao W, Deng Y, et al. Identifying family experiences and perspectives on involvement in care of adult ICU patients under family-centered care: A systematic review of qualitative studies. Intensive Crit Care Nurs 2025;90:104059. [doi: 10.1016/j.iccn.2025.104059] [Medline: 40300243]
61. Putowski Z, Rachfalska N, Majewska K, Megger K, Krzych K. Identification of risk factors for post-intensive care syndrome in family members (PICS-F) among adult patients: a systematic review. Anaesthesiol Intensive Ther 2023;55(3):168-178 [FREE Full text] [doi: 10.5114/ait.2023.130831] [Medline: 37728444]
62. Reifarth E, Böll B, Köpke S, Altenrath L, Garcia Borrega J, Kochanek M, et al. Multicomponent communication intervention to support family members of the critically ill: a controlled pre-post study. Crit Care Med 2025;53(7):e1437-e1448. [doi: 10.1097/CCM.0000000000006714] [Medline: 40387470]
63. Wolff JL, Aufill J, Echavarria D, Heughan J, Lee KT, Connolly RM, et al. Sharing in care: engaging care partners in the care and communication of breast cancer patients. Breast Cancer Res Treat 2019;177(1):127-136 [FREE Full text] [doi: 10.1007/s10549-019-05306-9] [Medline: 31165374]

Abbreviations

CNKI: China Knowledge Resource Integrated Database
DSM-5: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition)
ICU: intensive care unit
DHI: digital health intervention
DHT: digital health technology
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
PI: prediction interval
PICS-F: Post-intensive Care Syndrome-Family
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-analyses literature search extension
PTSD: posttraumatic stress disorder
QOC: quality of communication
QoL: quality of life
RCT: randomized controlled trial
ROB 2: Cochrane's risk-of-bias tool for randomized trials, version 2
SMD: standardized mean difference
VIP: Weipu Databse
VR: virtual reality

XSL•FO
RenderX

Edited by S Brini; submitted 31.Aug.2025; peer-reviewed by H Maheshwari, I Orimolade; comments to author 07.Oct.2025; accepted 22.Dec.2025; published 25.Feb.2026.

Please cite as:
Zhang C, Sui W, Jiang W, Shen Y, Huang Z, Yan R, Yi J, Zhang J, Zhang C, Zhuang Y, Gong X
The Effects of Digital Health Interventions for Family Members in Intensive Care Units: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e83294
URL: https://www.jmir.org/2026/1/e83294
doi:10.2196/83294
PMID:

©Chuchu Zhang, Weijing Sui, Weilin Jiang, Yibing Shen, Zhenzhen Huang, Ran Yan, Jia Yi, Jiayu Zhang, Chaoyi Zhang, Yiyu Zhuang, Xiaoyan Gong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

AI-Enabled Wearables for Motor Function Assessment and Rehabilitation in Parkinson Disease: Scoping Review

Shengting Li[1,2*], MSc; Siqi Chen[1,2*], MSc; Xiaosong Yu[1,2*], MSc; Huixiang Shang[1,2], MSc; Tingting Tu[1,2], MSc; Mingtao Quan[1,2], MSc

[1]School of Nursing, Zunyi Medical University, Zunyi, China

[2]Affiliated Hospital of Zunyi Medical College, Zunyi, China

[*]these authors contributed equally

Corresponding Author:
Mingtao Quan, MSc
School of Nursing
Zunyi Medical University
No. 201 Dalian Road, Huichuan District
Zunyi, 563000
China
Phone: 86 18185066688
Email: 865768818@qq.com

Abstract

Background: Artificial intelligence (AI)–enabled wearable devices are rapidly emerging in rehabilitation and motor function assessment for patients with Parkinson disease (PD). However, evidence remains fragmented, integration into nursing practice is limited, and comprehensive synthesis is lacking.

Objective: This study aimed to summarize studies on AI-enabled wearable devices for PD rehabilitation and motor function assessment, describing device types, monitored indicators, algorithms, and application characteristics, and identifying research gaps and barriers to clinical translation.

Methods: Guided by the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) framework, 9 databases (China National Knowledge Infrastructure, Wanfang Data, SinoMed, Cochrane Library, PubMed, Web of Science, CINAHL, Scopus, and Embase) were searched from inception to December 2025. Eligible studies were published in English or Chinese from January 1, 2020, onward and enrolled people with PD using noninvasive, body-worn AI-enabled wearable devices for rehabilitation, assessment, or monitoring. Dissertations and full conference papers were included, whereas preprints and conference abstracts were excluded. Methodological quality was appraised using the Mixed Methods Appraisal Tool, 2018 tool. Results were synthesized narratively and mapped to characterize devices, sensing modalities, algorithms, and evaluation methods.

Results: A total of 66 studies involving approximately 3579 participants were included. Wearable devices mainly comprised multisensor modules, smart insoles, and wrist-worn devices, with accelerometers being the most frequently used sensors. Data collection was predominantly passive, and most studies were conducted in laboratory or clinical settings using single- or short-term sessions. Internal validation approaches, particularly leave-one-out and k-fold cross-validation, were common, whereas external validation was rare, and reporting of calibration and clinical decision thresholds was limited. Sensitivity and accuracy were the most frequently reported performance metrics, highlighting substantial heterogeneity in analytical methods and outcome reporting.

Conclusions: This scoping review systematically synthesized evidence on AI-enabled wearable devices for motor function assessment and rehabilitation in PD, complemented by an evidence map and guided by a rehabilitation- and nursing-oriented perspective, and identified key translational gaps between proof-of-concept studies and real-world rehabilitation workflows. Compared with previous reviews that primarily focused on monitoring functions or device performance, this review places greater emphasis on rehabilitation applications and nurse-led translation into practice, and proposes a conceptual “challenges and opportunities” framework to inform the design, evaluation, and reporting of devices and algorithms, while further highlighting key considerations for workflow integration and the implementation of decision-support systems. These findings have practical relevance for advancing continuity of rehabilitation across clinical, home, and community settings, and may help guide nurses

XSL•FO
RenderX

in delivering continuous monitoring, personalized follow-up, and timely intervention, thereby improving the efficiency and accessibility of rehabilitation management.

(J Med Internet Res 2026;28:e85596) doi:10.2196/85596

KEYWORDS

artificial intelligence; Parkinson disease; rehabilitation; scoping review; wearable devices

Introduction

Background

Parkinson disease (PD) is a common neurodegenerative disorder characterized by tremor, bradykinesia, rigidity, and postural or gait disturbances [1,2]. Its high prevalence and disability rates profoundly impair patients' quality of life and impose substantial burdens on families and health care systems. Since 1990, the number of patients with PD has more than doubled—from approximately 2.8 million to over 6.2 million—and is projected to exceed 12 million by 2040 [3]. In China, accelerated population aging has led to a rising prevalence, with an estimated 1.37% among adults aged 60 years and older and more than 3.62 million patients nationwide [4,5]. Rehabilitation remains essential for improving motor function and slowing functional decline [6-8]. However, conventional face-to-face rehabilitation is time- and resource-intensive, often inaccessible in remote regions, and frequently associated with poor adherence [9,10].

In recent years, the rapid development and application of wearable devices have provided significant advantages for assessment and intervention in neurological and psychiatric disorders, such as stroke and depression [11,12]. Wearable devices encompass a wide range of intelligent electronic devices, including smartwatches, wristbands, insoles, exoskeletons, multisensor wearable devices, and electronic textiles [13-16]. These devices support continuous, real-time, and multidimensional monitoring during daily activities or rehabilitation training, capturing key motor indicators such as gait, tremor, joint mobility, postural control, motor coordination, and overall physical activity levels [17-19].

Against this backdrop, the rapid emergence of artificial intelligence (AI)–enabled wearable devices has created new opportunities for the rehabilitation of PD. In this review, AI-enabled wearable devices refer to noninvasive wearable technologies that incorporate machine learning (ML) or deep learning (DL) methods for data-driven analysis beyond rule-based approaches. Unlike traditional approaches that rely on manual observation or intermittent scale-based assessments, AI-enabled wearable devices enable the capture of large volumes of fine-grained data. Supported by ML and DL techniques, these data can be efficiently processed and intelligently interpreted within a data-driven rehabilitation framework, highlighting the importance of AI model interpretability for clinical and nursing decision-making [20-22]. Moreover, AI-enabled wearable devices facilitate continuous remote monitoring and long-term follow-up, making it possible to track disease progression and changes in motor function in real time within home-based rehabilitation settings. These data can then be leveraged by clinicians and rehabilitation teams to design more precise training programs [23,24]. Such an approach enhances the continuity and personalized rehabilitation, offers a feasible pathway for remote and home-based rehabilitation, and holds promising potential for optimizing health care resource allocation and improving patient adherence.

Research Questions and Objectives

Although numerous studies and reviews have explored the use of wearable devices in PD rehabilitation, several limitations remain, including (1) most studies involve small sample sizes and short follow-up periods, providing insufficient longitudinal evidence to evaluate long-term rehabilitation outcomes [17,25,26]; (2) current reviews mainly focus on device performance or symptom monitoring, with limited attention to AI data-processing workflows, algorithmic applications, and their potential value for clinical and nursing practice [18,19]; (3) studies differ substantially in device types, monitoring indicators, data-processing strategies, and AI algorithm selection. The absence of a unifying framework to synthesize these variations hinders a comprehensive understanding of the research landscape [27-29]; and (4) evidence in nursing practice remains limited; despite nurses' central role in rehabilitation, research on AI-enabled tools for nursing decision-making, health education, and adherence support is still scarce [30,31]. Collectively, these issues result in fragmented and unsystematic evidence, limiting its ability to guide clinical rehabilitation nursing practice and theoretical development.

To address these needs, we conducted a scoping review and developed an evidence map to systematically present the current applications of AI-enabled wearable devices in PD rehabilitation and motor function assessment. The scoping review summarized the breadth, depth, and nature of existing evidence [32], while the evidence map visualizes research distribution, patterns, and emerging trends across domains [33]. By integrating these 2 approaches, this study aimed to map and visualize the current evidence on AI-enabled wearable devices for PD rehabilitation and motor function assessment. The objectives were to summarize device types, monitoring indicators, AI algorithms, and application characteristics, and to identify research gaps and future directions to support nursing practice and further digital-health implementation in PD rehabilitation.

Methods

Protocol and Registration

The review followed the methodological guidance outlined in the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [34], and the completed checklist is provided in Multimedia Appendix 1. No protocol was registered for this review.

XSL•FO
RenderX

Eligibility Criteria

The inclusion criteria were defined using the population, concept, and context framework. Eligible studies enrolled participants with a confirmed diagnosis of PD and used noninvasive, body-worn wearable devices integrated with AI-based analytical methods. The focus of included studies was on rehabilitation, motor function assessment, or monitoring. Gray literature such as dissertations and full conference papers was included, whereas preprints and conference abstracts were excluded to ensure methodological consistency and to prioritize evidence that had undergone formal peer review. Only studies published in English or Chinese on or after January 1, 2020, were eligible for inclusion. Detailed inclusion and exclusion criteria are presented in Multimedia Appendix 2.

Information Sources

A comprehensive search was conducted on August 12, 2025, across 9 databases: CNKI (China National Knowledge Infrastructure), Wanfang Data, SinoMed, Cochrane Library, PubMed, Web of Science, CINAHL, Scopus, and Embase. The literature coverage spanned from database inception to the search date. No websites or other nondatabase information sources were searched. Reference lists and forward citations of the included studies were additionally screened to identify further potentially relevant records. To ensure the currency and completeness of the evidence, an updated search using the same strategy was conducted on December 10, 2025.

Search Strategy

The reporting of the search process adhered to the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) [35]. The search strategy was developed by a researcher trained in systematic literature retrieval and combined MeSH (Medical Subject Headings) terms with free-text keywords across four core concepts: (1) wearable or body-worn mobile devices; (2) rehabilitation, assessment, or monitoring; (3) PD; and (4) AI or ML. No limits or restrictions, including language, publication date, or study design, were applied at the search level. All eligibility criteria were applied manually during title/abstract screening and full-text review. The search strategy did not undergo peer review. Complete search strategies for each database and information source are provided in Multimedia Appendix 3.

Selection of Sources of Evidence

The selection process comprised two stages: (1) title/abstract screening and (2) full-text review. In the first stage, 1 reviewer imported all records into EndNote X9 (version 12062; Clarivate), removed duplicates, and excluded studies outside the eligible date range. Two reviewers then independently screened titles and abstracts to determine provisional inclusion. In the second stage, the same 2 reviewers independently assessed the full texts to establish the final set of included studies. Disagreements were resolved through discussion; unresolved cases were adjudicated by a third reviewer. Cohen κ was calculated to measure the interrater agreement [36]. Agreement was high for title and abstract screening (κ=0.86), whereas it was moderate for full-text review (κ=0.57).

Data Charting Process

Data charting was conducted independently by 1 reviewer using a standardized extraction form developed a priori. The form included key study characteristics, wearable device details, AI methods, outcomes, and findings relevant to the review questions. Following Joanna Briggs Institute guidance, charting was iterative, and the form was refined as needed. A second reviewer verified all data, with disagreements resolved by a third reviewer. The form was pilot-tested on 5 studies before full implementation.

Data Items

The data extraction form was developed by the research team based on the objectives of this review and included three core modules: (1) study details, (2) wearable device details, and (3) AI details. When the relevant information was not reported, it was recorded as not available. Each module contained specific data variables, as detailed in Multimedia Appendix 4.

Critical Appraisal of Individual Sources of Evidence

The methodological quality of all included studies was evaluated using the Mixed Methods Appraisal Tool, 2018 version [37]. Two reviewers conducted the assessments independently and resolved discrepancies through discussion; cases in which consensus could not be reached were adjudicated by a third reviewer. The appraisal was used descriptively to identify methodological strengths and limitations of the included studies and was not used as a criterion for study inclusion. The detailed appraisal results are provided in Multimedia Appendix 5 [38-103].

Synthesis of Results

A narrative synthesis approach was used. Extracted data were summarized thematically and presented in descriptive text, tables, and an evidence map to illustrate research trends, methodological characteristics, and application domains of AI-enabled wearable devices in PD.

Ethical Considerations

This scoping review used only published, peer-reviewed literature and therefore did not require approval from an ethics committee or institutional review board. Findings will be submitted to an open access, peer-reviewed journal and presented at relevant medical and engineering conferences.

Results

Study Selection

A total of 1467 records were initially retrieved. Of these, 475 (32.38%) were removed using EndNote X9 due to duplication, publication date, or language incompatibility. The remaining 992 (67.62%) records underwent title and abstract screening, during which 877 (88.41%) were excluded. Among the 115 records reviewed in full text, 47 (40.87%) were unrelated to AI, 7 (6.09%) were unavailable in full text, and 2 (1.74%) lacked extractable data. An additional 7 relevant studies were identified through reference tracing. In total, 66 studies were included in this review [38-103]. The detailed selection process is illustrated

XSL•FO
RenderX

in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram (Figure 1).

Figure 1. Flowchart of the study selection process.

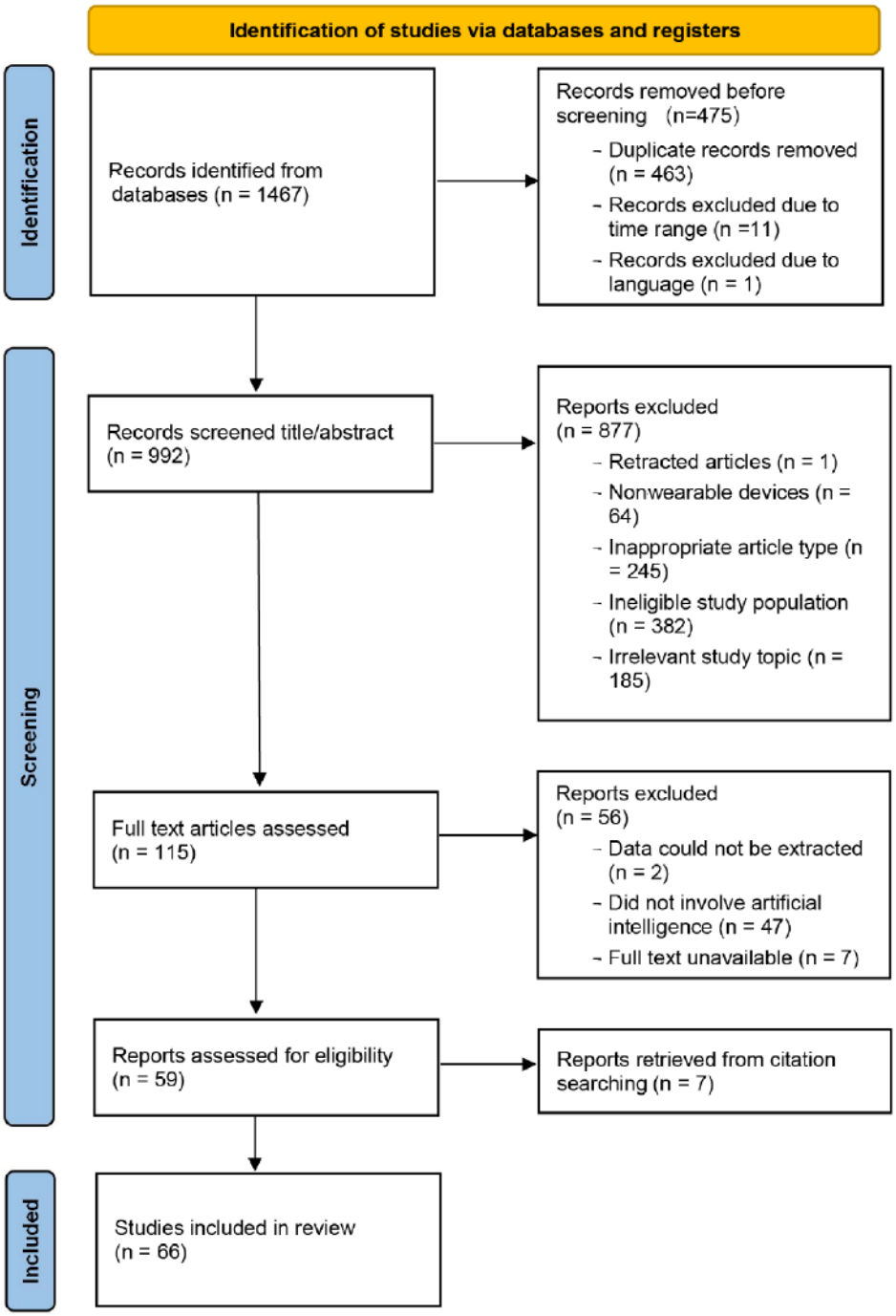

Characteristics of Included Studies

All included studies were published between 2020 and 2025, and the publication volume remained relatively stable over this 6-year period. The highest number of studies was published in 2025 (16/66, 24.24%; Table 1). The included studies originated from multiple countries and regions (Figure 2), with Switzerland contributing the largest proportion (27/66, 40.91%). With respect to publication type, the vast majority were journal articles (61/66, 92.42%).

XSL•FO
RenderX

Table 1. Characteristics of the included studies using artificial intelligence–enabled wearable devices for Parkinson disease (PD; N=66).

Features	Values	References
Year of publication, n (%)		
2020	8 (12.12)	[38,47,50,53,61,63,66,73]
2021	11 (16.67)	[44,49,55,56,59,62,72,82,86,94,96]
2022	9 (13.64)	[39,46,51,58,64,68,80,81,83]
2023	11 (16.67)	[43,57,60,71,74,76,77,84,85,89,91]
2024	11 (16.67)	[48,54,65,67,69,70,78,87,90,93,95]
2025	16 (24.24)	[40-42,45,52,75,79,88,92,97-103]
Type of publication, n (%)		
Journal article	61 (92.42)	[38,40-45,47-69,71-84,86-91,93-103]
Conference paper	4 (6.06)	[39,46,70,92]
Thesis	1 (1.52)	[85]
Country/region of publication, n (%)		
Switzerland	27 (40.91)	[41-44,47,49,55,57,58,62,65-68,71,75,79,80,88,90,91,96,97,99,101-103]
United States	19 (28.79)	[39,46,48,51,54,59,61,64,69,70,73,74,76,78,81,86,89,92,94]
United Kingdom	11 (16.67)	[50,52,53,56,60,63,72,87,93,95,98]
Netherlands	5 (7.58)	[38,40,45,82,100]
China	3 (4.55)	[83-85]
Germany	1 (1.52)	[77]
Number of participants		
Mean (SD; range)	55.06 (132.78; 2-1057)	[38-99,101-103]
1-30, n (%)	41 (62.12)	[38-42,44,46-49,52,55-64,67,69-75,82,85,86,90,91,93-97,99,103]
31-60, n (%)	8 (12.12)	[43,68,76-79,81,92]
61-100, n (%)	9 (13.64)	[45,50,51,53,65,66,83,89,101]
101-500, n (%)	6 (9.09)	[54,84,87,88,98,102]
>500, n (%)	1 (1.52)	[80]
Not reported	1 (1.52)	[100]
Age distribution of participants[a] (years)		
Mean[b] (SD; range)	64.59 (10.03; 25.00-76.10)	[38-47,49-53,56-61,63-76,80,81,83-85,89,91-93,95,96,98,99,101-103]
<60, n (%)	3 (4.55)	[39,40,55]
60-69, n (%)	39 (59.09)	[42-45,47,49,51,53,56,63,65-67,69-77,79-81,83,85,86,88,89,91-93,95,96,98,99,102,103]
70-79, n (%)	14 (21.21)	[38,41,46,50,52,57-61,64,68,84,101]
≥80, n (%)	2 (3.03)	[54,97]
Not reported	8 (12.12)	[48,62,78,82,87,90,94,100]
Sex distribution[c]		
Female (%), mean (SD; range)	35.88 (18.08; 0–100)	[38-41,45-48,50-68,70-73,75,79-81,83-92,94-96,98,99,101,102]
Inclusion details n (%)		
PD only	47 (71.21)	[38,40,41,43,46-49,51,52,54-72,76,81,82,84-88,90,93-99,102,103]
Mixed	18 (27.27)	[39,42,44,45,50,53,73-75,77-80,83,89,91,92,101]
Not reported	1 (1.52)	[100]

XSL•FO
RenderX

Features	Values	References
Application objective[d], n (%)		
Motor function assessment	54 (81.82)	[38,39,41-47,49,50,52-55,57-80,82-89,91,92,95,99-102]
Disease progression/symptom monitoring	26 (39.39)	[39,46,51-53,55,56,65,68,69,71,74,76,79,87,89,90,93,94,96-100,102,103]
Efficacy evaluation	11 (16.67)	[46,54,56,61,70,71,77,78,81,90,103]
Rehabilitation training	8 (12.12)	[40,48,54,55,60,67,77,96]
Others	1 (1.52)	[88]

[a]Reported as mean (SD), median (IQR), or range, depending on the study.

[b]Age was summarized from studies reporting mean (SD); studies with only median or without age were excluded. For mixed populations, only PD patient data were used.

[c]Based only on studies that reported sex data.

[d]Application objectives were classified based on the primary functional aim of the wearable system. Motor function assessment refers to quantitative evaluation of motor performance at specific time points; disease progression/symptom monitoring refers to continuous or longitudinal tracking of disease states; efficacy evaluation refers to the assessment of responses to interventions or treatments; and rehabilitation training refers to systems explicitly designed to support training or feedback to improve motor function. The number of studies does not add up, as several studies have used >1 application objective.

Figure 2. Geographical distribution of the included studies using artificial intelligence–enabled wearable devices for Parkinson disease.

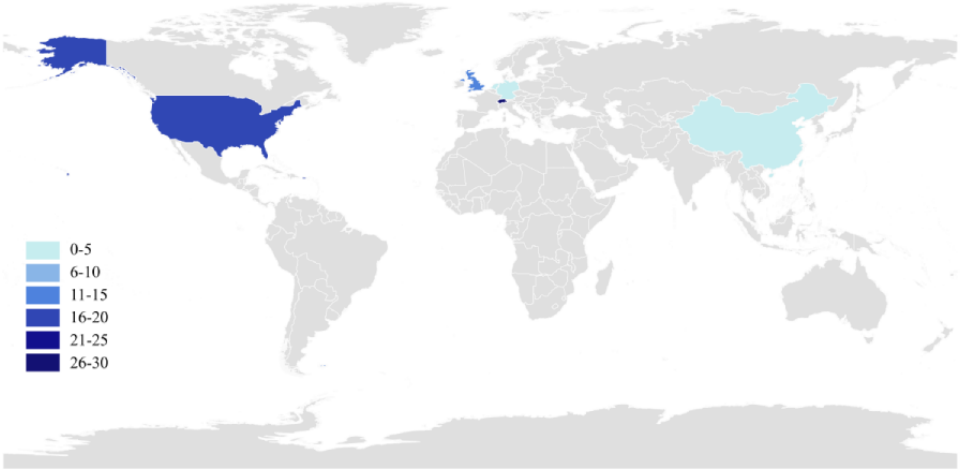

Regarding study populations, sample sizes ranged from 2 to 1057 participants, with a mean sample size of 55.06 (SD 132.78), yielding a total of approximately 3579 participants. Overall, most studies involved relatively small samples with substantial heterogeneity, and the majority included only patients with PD (47/66, 71.21%). In terms of the application objectives of wearable devices, most studies focused on motor function assessment (54/66, 81.82%), whereas rehabilitation-oriented applications were relatively scarce, being reported in only 8 (12.12%) studies. Overall, this distribution suggests that intervention-focused research remains limited. Detailed characteristics of the included studies are presented in Table 1 and Multimedia Appendix 6 [38-103]. For categories in which studies could contribute to more than 1 classification, percentages may exceed 100%; detailed coding rules are provided in the table footnotes.

Characteristics of Wearable Devices

A total of 5 major types of wearable devices were identified across the included studies. Sensor module was the predominant device type, accounting for 84.85% (56/66) of studies, followed by smart insoles (7/66, 10.61%) and smartwatches (5/66, 7.58%; Table 2). Regarding device origin, commercially available devices were more common (41/66, 62.12%) than noncommercial devices (21/66, 31.82%); nevertheless, a substantial proportion of studies still relied on customized or research-grade devices. A total of 14 distinct wearing locations were reported (Figure 3). The most common placements were the shank (25/66, 37.88%) and the wrist (23/66, 34.85%), followed by the foot (19/66, 28.79%) and the ankle (16/66, 24.24%). Many studies used multiple wearing locations within a single study, reflecting the diversity of measurement strategies adopted.

XSL•FO
RenderX

Table 2. Features of artificial intelligence–enabled WDsa used in the included Parkinson disease studies (N=66).

Features	Values, n (%)	References
Type of WD[a,b]		
Sensor module	56 (84.85)	[38-45,47,48,50-75,78-82,84,86-89,91-98,102,103]
Smart insole	7 (10.61)	[39,46,55,57,59,100,101]
Smartwatch	5 (7.58)	[49,63,85,97,99]
Smart wristband	3 (4.55)	[83,85,90]
Stimulator	2 (3.03)	[48,103]
Others (adhesive electrodes, ankle band, or smart glove)	3 (4.55)	[42,76,77]
Status of WD[a,b]		
Commercial	41 (62.12)	[38,42,44,46,48,49,51-59,63-67,70-72,74,75,80,81,84-89,91-93,97-99,101,102]
Noncommercial	21 (31.82)	[39-41,43,45,47,48,50,60,67-69,73,76-78,83,84,90,100,103]
Not reported	8 (12.12)	[61,62,79,82,85,94-96]
Company of WD[a,b]		
APDM Inc	6 (9.09)	[53,62,64,66,86,87]
Shimmer Sensing	5 (7.58)	[38,55,59,85,93]
Activinsights Ltd	4 (6.06)	[43,44,85,97]
MC10 Inc	3 (4.55)	[53,63,75]
Moticon ReGo AG	3 (4.55)	[46,57,101]
Noraxon USA Inc	2 (3.03)	[42,74]
PD Neurotechnology Ltd	2 (3.03)	[57,71]
Great Lakes NeuroTechnologies Inc	2 (3.03)	[56,70]
STMicroelectronics	2 (3.03)	[68,69]
Tekscan Inc	2 (3.03)	[55,59]
Self-developed	13 (19.70)	[39,51,60,61,67,73,76,77,83,84,89,100,103]
Others	16 (24.24)	[43,44,48,50,54,58,63,65,88-92,96,99,102]
Not reported	16 (24.24)	[40,41,45,47,49,52,72,78-82,94,95,97,98]
Placement[b]		
Shank	25 (37.88)	[40,42-44,48,52,55,59,60,67,69,71,74,75,80-82,84,86,88,91,93-95,102]
Wrist	23 (34.85)	[38,44,49,53,56-58,63,69-71,79,83,85,87-92,97,99,102]
Foot	19 (28.79)	[39,40,46,48,55,57,59,62,72-74,87,88,91,96,100-103]
Ankle	16 (24.24)	[38,39,41,45,51,52,61,64,66,70,76,83,93,98,99,103]
Waist	13 (19.70)	[44,47,52,54,57,65,69,71,86,88,89,97,102]
Trunk	12 (18.18)	[45,66,74,75,78,81,84,87,88,92,98,102]
Arm	9 (13.64)	[45,74,75,78,81,84,85,97,98]
Thigh	7 (10.61)	[55,59,68,74,88,95,102]
Hip	6 (9.09)	[62,74,78,82,93,98]
Hand	3 (4.55)	[63,77,91]
Others (Neck, knees, back, or head)	4 (6.06)	[50,51,74,95]
Compatibility with OS[c,d]		
Local logger	16 (24.24)	[38,40,43,44,46,50,54,57,58,60,61,69,73,87,88,93]
Android	6 (9.09)	[64,65,76,85,90,97]

XSL•FO
RenderX

Features	Values, n (%)	References
iOS	5 (7.58)	[48,51,63,65,90]
Windows	3 (4.55)	[68,83,84]
Not applicable	7 (10.61)	[39,52,67,71,94,95,103]
Not reported	31 (46.97)	[41,42,45,47,49,53,55,56,59,62,66,70,72,74,75,77-82,86,89,91,92,96,98-102]
Gateway[e]		
PC[f]	10 (15.15)	[52,68,73,74,76,77,80,86,90,102]
Smartphone	6 (9.09)	[39,48,64,76,85,90]
Tablet	3 (4.55)	[51,80,92]
IoT[g] Gateway	2 (3.03)	[82,103]
Not reported	48 (72.73)	[38,40-47,49,50,53-63,65-67,69-72,75,78,79,81,83,84,87-89,91,93-101]
Host[h]		
PC[f]	35 (53.03)	[38,41,43,44,52,54-59,61,66,68,69,73-77,80,82-84,86-89,91,93,94,96,98,102,103]
Server	6 (9.09)	[39,48,58,71,90,97]
Smartphone	5 (7.58)	[48,64,65,76,85]
On-device	4 (6.06)	[40,60,65,67]
Tablet	3 (4.55)	[51,80,92]
Not reported	18 (27.27)	[42,45-47,49,50,53,62,63,70,72,78,79,81,95,99-101]
Mode of data transfer[i]		
Bluetooth	15 (22.73)	[39,40,48,51,52,64,65,68,76,80,84,86,88,89,102]
Internet	8 (12.12)	[39,48,58,71,83,85,90,103]
Removable media	6 (9.09)	[38,40,43,48,61,69]
Wired	5 (7.58)	[40,44,73,77,103]
Not reported	39 (59.09)	[41,42,45-47,49,50,53-57,59,60,62,63,66,67,70,72,74,75,78,79,81,82,84,87,91-101]
Sensors in the wearables[j]		
Accelerometer	61 (92.42)	[38-54,57-59,61-99,102,103]
Gyroscope	50 (75.76)	[38-43,45-49,51,52,54-57,59-64,66-74,76,78,80,81,83,84,86-89,92-94,96,98,99,102,103]
Magnetometer	11 (16.67)	[45,52,53,66,71,76,78,83,86,87,94]
Pressure sensor	9 (13.64)	[39,40,46,55,57,59,77,100,101]
sEMG[k] sensor	3 (4.55)	[40,42,78]
Flex sensor	2 (3.03)	[40,77]
Measured biosignals[l]		
Acceleration	61 (92.42)	[38-54,57-59,61-99,102,103]
Angular velocity	50 (75.76)	[38-43,45-49,51,52,54-57,59-64,66-74,76,78,80,81,83,84,86-89,92-94,96,98,99,102,103]
Magnetic field signals	11 (16.67)	[45,52,53,66,71,76,78,83,86,87,94]
Pressure/mechanical signals	9 (13.64)	[39,40,46,55,57,59,77,100,101]
EMG[m] signals	3 (4.55)	[40,42,78]
Bending/Flex sensing	2 (3.03)	[40,77]
Sensing type		
Passive	52 (78.79)	[38,40,41,44-60,62,65,71-73,75-79,82-103]
Active	14 (21.21)	[39,42,43,61,63,64,66-70,74,80,81]
Application scenario		

XSL•FO
RenderX

Features	Values, n (%)	References
Clinical	31 (46.97)	[38,48,51,54,55,57,59-61,63,66-68,74,76-78,80,81,83,84,86,88,89,92-96,98,99]
Home	12 (18.18)	[44,47,58,65,71,72,79,82,97,100-102]
Clinical and Home	23 (34.85)	[39-43,45,46,49,50,52,53,56,62,64,69,70,73,75,85,87,90,91,103]
Duration of monitoring/intervention		
Single session	32 (48.48)	[38,40-44,46,48-50,52,55,64,67-69,73,74,76,77,79,83-85,89-91,96,99-102]
Multiple sessions	20 (30.30)	[47,51,53,54,56,57,60,61,63,66,70,75,80,92-95,97,98,103]
Long-term monitoring	9 (13.64)	[39,45,58,62,65,71,72,81,86]
Not reported	5 (7.58)	[59,78,82,87,88]

[a]WD: wearable device.

[b]The number of studies does not add up, as several studies have used >1 wearable device.

[c]OS: operating system.

[d]The number of studies does not add up, as several studies have used >1 wearable device, and many wearable devices are compatible with >1 operating system.

[e]The number of studies does not add up, as several studies used >1 wearable device, and many wearable devices used >1 gateway.

[f]PC: personal computer.

[g]IoT: Internet of Things.

[h]The number of studies does not add up, as several studies used >1 wearable device, and many wearable devices used >1 host.

[i]The number of studies does not add up, as several studies used >1 wearable device, and many wearable devices used >1 mode of data transfer.

[j]The number of studies does not add up, as several studies used >1 wearable device, and most wearable devices have >1 sensor.

[k]sEMG: surface electromyography.

[l]The number of studies does not add up, as several studies used >1 wearable device, and most wearable devices assess >1 biosignal.

[m]EMG: electromyography.

Figure 3. Placement of wearable sensors in studies using artificial intelligence–enabled wearable devices for Parkinson disease.

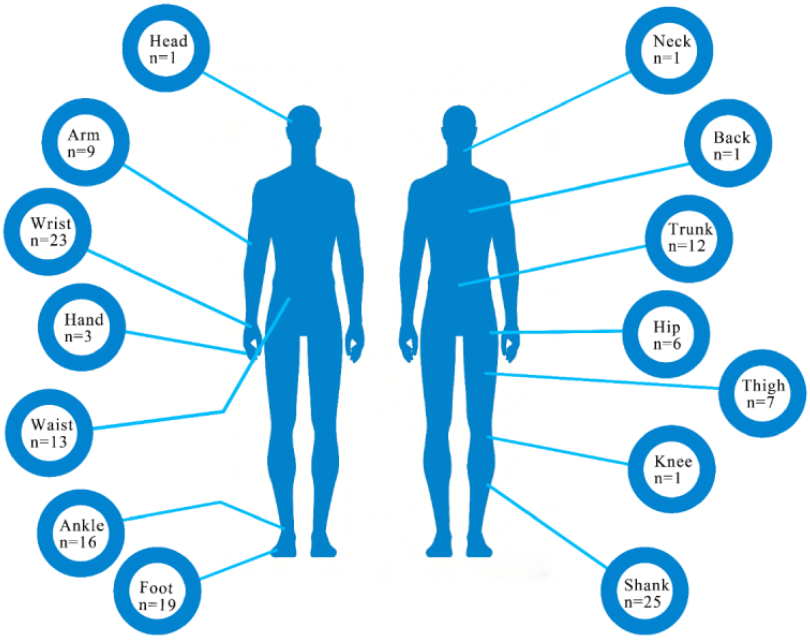

With regard to sensor configuration, devices integrating multiple sensors predominated, accounting for 75.76% (50/66) of the included studies. The collected biosignals were predominantly inertial signals. Accelerometer data were acquired in the vast majority of studies (61/66, 92.42%), followed by gyroscope data (50/66, 75.76%), whereas magnetic field, pressure, and electromyography signals were used less frequently (Table 2).

With respect to application settings, wearable devices were more frequently deployed in clinical environments (31/66, 46.97%), with an additional 34.85% (23/66) of studies spanning both clinical and home environments, whereas studies conducted exclusively in home settings were relatively uncommon (12/66, 18.18%). Regarding the duration of monitoring or intervention, most studies used either single-session assessments (32/66, 48.48%) or repeated short-term testing designs (20/66, 30.30%),

XSL•FO
RenderX

while studies involving long-term continuous monitoring were comparatively limited (9/66, 13.64%). Overall, wearable device research in PD is dominated by sensor module devices, with devices most commonly worn on the lower limbs and wrist. Data collection is primarily passive, and studies largely focus on short-term applications conducted in clinical settings. Additional technical characteristics of the wearable devices are provided in Table 2 and Multimedia Appendix 7 [38-103].

AI Algorithm Characteristics of Wearable Devices

The AI applications in the included studies were categorized into four purposes: (1) monitoring or assessment (35/66, 53.03%), (2) state recognition or functional screening (17/66, 25.76%), (3) prediction (8/66, 12.12%), and (4) rehabilitation and feedback (6/66, 9.09%; Table 3). Among the studies, 62.12% (41/66) used ML algorithms only, 28.79% (19/66) used DL algorithms only, and 9.09% (6/66) combined ML and DL approaches. These studies used algorithms to address classification problems (59/66, 89.39%), regression problems (15/66, 22.73%), and clustering problems (3/66, 4.55%).

XSL•FO
RenderX

Table 3. Features of artificial intelligence (AI) algorithms used in the included Parkinson disease wearable-device studies (N=66).

Features	Values, n (%)	References
AI category		
ML[a]	41 (62.12)	[38,39,43,45,48,50,53-55,57-63,65,66,68,71-77,80-84,86-88,92,96,98,99,101-103]
DL[b]	19 (28.79)	[41,42,44,46,49,51,52,56,64,67,69,70,78,79,89-91,94,100]
ML and DL	6 (9.09)	[40,47,85,93,95,97]
Task type[c]		
Classification	59 (89.39)	[39-55,57-60,62-67,69,71,72,74-79,81-103]
Regression	15 (22.73)	[38,43-45,53,56,61,68-71,77,80,98,102]
Clustering	3 (4.55)	[73,77,98]
AI algorithm[d]		
Support vector machine	28 (42.42)	[38,40,43,45,47,50,53,57,61,65,66,68,74,76,78,81-85,87,88,92,93,96,101-103]
Convolutional neural network	23 (34.85)	[40-42,44,46,47,49,51,52,54,56,67,69,70,78,85,89-91,93,95,97,100]
Random forest	19 (28.79)	[39,45,47,53,57,60,63,79,80,82,83,86,87,92,95,97-99,101]
Logistic regression	15 (22.73)	[39,43,53,61,62,65,71,78-80,87,88,92,98,101]
Decision tree	14 (21.21)	[45,47-50,55,57,59,61,62,76,82,86,88]
Long short-term memory	11 (16.67)	[42,47,52,56,58,69,70,85,93,94,100]
Multilayer perceptron	8 (12.12)	[40,41,44,77,78,88,92,100]
k-nearest neighbors	7 (10.61)	[50,76,83,88,92,93,101]
Naive Bayes	5 (7.58)	[43,50,76,88,101]
Extreme gradient boosting	4 (6.06)	[49,87,93,102]
Gradient boosting	3 (4.55)	[39,57,86]
Self-attention mechanism	3 (4.55)	[42,69,85]
Hidden Markov model	3 (4.55)	[58,72,73]
k-means clustering	3 (4.55)	[75,77,98]
Elastic net	3 (4.55)	[80,87,98]
AdaBoost	3 (4.55)	[76,92,96]
Radial basis function	2 (3.03)	[78,79]
Others	12 (18.18)	[43,45,47,64,67,78-80,82,86,97,102]
Aim of AI algorithm		
Monitoring or assessment	35 (53.03)	[38,39,42-47,50,53,56-59,61,63,65,70-73,79,81,83-85,88,90-93,97-99,102]
State recognition or functional screening	17 (25.76)	[41,49,51,54,62,66,74-78,82,89,94,100,101,103]
Prediction	8 (12.12)	[52,55,64,68,69,80,87,95]
Rehabilitation and feedback	6 (9.09)	[40,48,60,67,86,96]
Validation approach[e]		
Leave-one-out cross-validation	37 (56.06)	[39,41,42,45-47,49,50,53-55,57-63,65-68,70,73,74,76,78,79,81,83,86,88,90,91,93,102,103]
k-fold cross-validation	27 (40.91)	[38-41,47,49-52,54,57,69,72,76,77,84,85,87-92,99,101,103]
Hold-out validation	18 (27.27)	[40,44,48,49,56,61,64,66,76,78,80,82,88,92,94,100-102]
External validation	4 (6.06)	[43,60,69,71]
Not reported	5 (7.58)	[75,95-98]
Performance measures[f]		

XSL•FO
RenderX

Features	Values, n (%)	References
Sensitivity	43 (65.15)	[39-42,46,47,49-55,57-60,62,64-67,69,72,73,75-79,82-85,87,90,94-97,99,101,103]
Accuracy	41 (62.12)	[39-43,45,46,48,50-55,57,58,62,65,67,69,74-77,82-87,89-92,95-98,100-102]
Specificity	27 (40.91)	[46,47,49,50,52-55,59,60,62,64-67,69,76,78,79,82,83,94-96,99,101,103]
F_1-score	27 (40.91)	[39-41,49,51-55,57,67,73,75-78,82,84,85,87,92,93,95,99,101-103]
Area under the curve	25 (37.88)	[39-41,47,52,55,57,58,62-65,69,74,78,79,82,86-88,92,94,96,98,101]
Precision	24 (36.36)	[39-42,46,51,53,54,58,59,62,65-67,75-78,84,85,87,92,95,97]
Correlation coefficient	14 (21.21)	[38,44,53,56,58,61,68,70,71,79-81,88,102]
Receiver operating characteristic	7 (10.61)	[60,63,66,74,89,94,98]
Mean absolute error	7 (10.61)	[43,44,56,68,70,98,102]
Root-mean-square error	6 (9.09)	[38,53,61,68,80,102]
Intraclass correlation coefficient	4 (6.06)	[38,54,77,93]
Cohen κ	3 (3.03)	[77,95,102]
Equal error rate	2 (3.03)	[47,52]
G-mean	2 (3.03)	[51,52]
Negative predictive value	2 (3.03)	[52,65]
Coefficient of determination	2 (3.03)	[70,98]
Others	9 (13.64)	[38,52,54,59,66,67,82,88,99]

[a]ML: machine learning.

[b]DL: deep learning.

[c]The number of studies does not add up, as many studies have used >1 task type.

[d]The number of studies does not add up, as many studies have used >1 AI algorithm.

[e]The number of studies does not add up, as many studies have used >1 validation approach.

[f]The number of studies does not add up, as most studies used >1 performance measure.

A total of 30 distinct AI algorithms were identified across the included studies. The most frequently used were support vector machines (SVMs; 28/66, 42.42%), followed by convolutional neural networks (CNNs; 23/66, 34.85%), random forests (19/66, 28.79%), logistic regression (15/66, 22.73%), decision trees (14/66, 21.21%), long short-term memory (LSTM) networks (11/66, 16.67%), and multilayer perceptrons (8/66, 12.12%; Table 3).

As shown in Figure 4, SVMs and CNNs were predominantly applied in the “monitoring or assessment” and “state recognition or functional screening” categories. Random forests, decision trees, and logistic regression models were distributed across application purposes but were mainly used for monitoring- or assessment-related tasks. Temporal and probabilistic models, such as LSTM networks and hidden Markov models, were more often observed in state recognition or functional screening tasks, although their overall adoption remained limited. In contrast, algorithmic applications in the “prediction” and “rehabilitation or feedback” categories were comparatively sparse, indicating that prospective and closed-loop applications remain at an early stage of development.

XSL•FO
RenderX

Figure 4. Distribution of artificial intelligence (AI) algorithms and application objectives in the included Parkinson disease wearable-device studies.

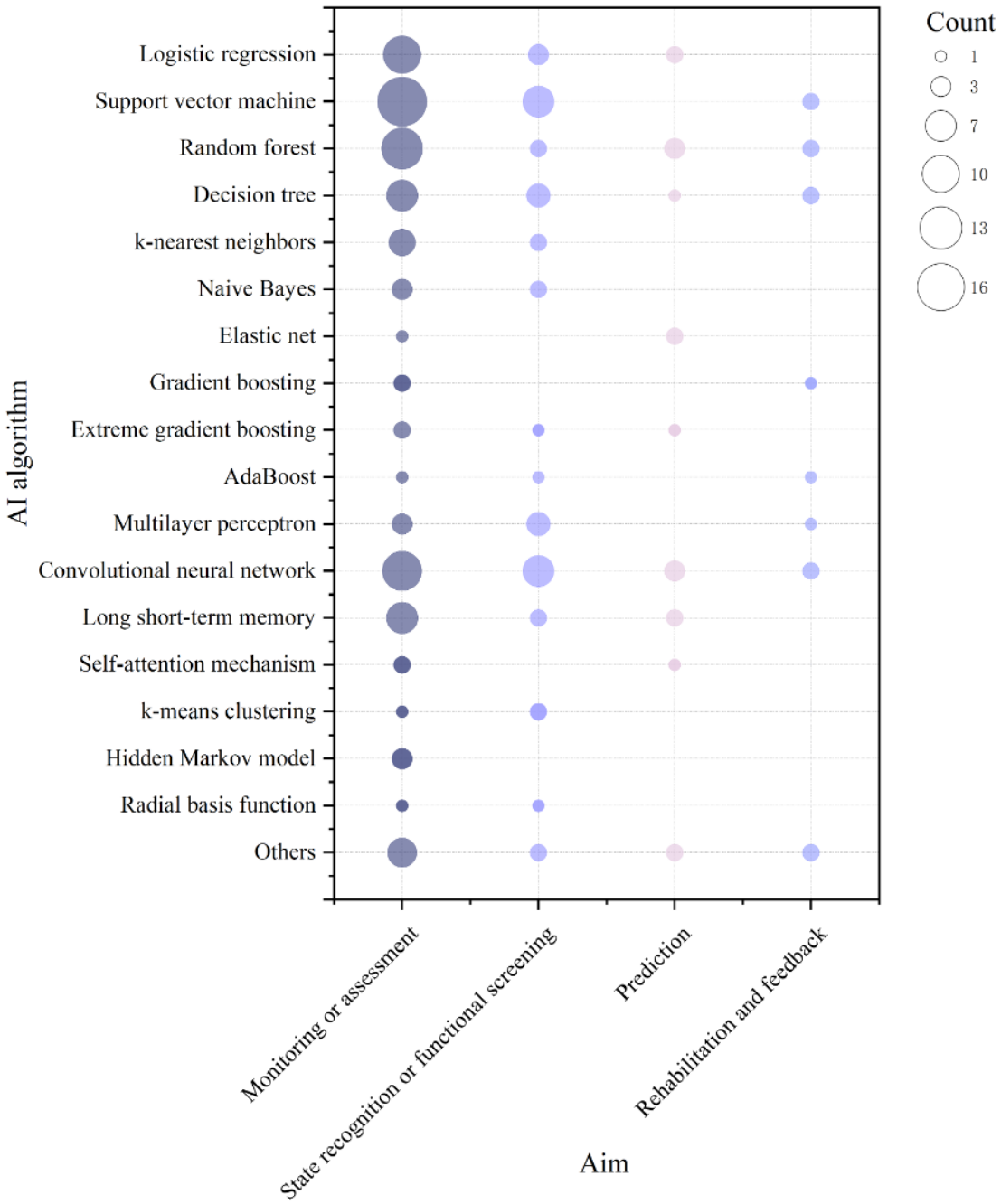

A total of 4 different model validation approaches were used across the included studies, with approximately 31.82% (21/66) using more than 1 method (Table 3). Leave-one-out cross-validation was the most frequently used approach (37/66, 56.06%), followed by k-fold cross-validation (27/66, 40.91%), hold-out validation (18/66, 27.27%), and external validation (4/66, 6.06%). A total of 25 performance metrics were used to evaluate model performance. The most commonly reported metrics were sensitivity (43/66, 65.15%), accuracy (41/66, 62.12%), specificity (27/66, 40.91%), F_1-score (27/66, 40.91%), area under the curve (25/66, 37.88%), precision (24/66, 36.36%), and correlation coefficient (14/66, 21.21%). Detailed characteristics of the AI algorithms applied in each study are provided in Multimedia Appendix 8 [38-103].

Challenges and Opportunities of Wearable Devices

Based on a synthesis of the included evidence, we developed a conceptually integrated "challenges and opportunities" framework grounded in the evidence base. This framework categorizes the challenges and opportunities of AI-enabled wearable devices in clinical applications into six dimensions: (1) evidence and data quality; (2) technical limitations; (3) usability, adherence, and equity; (4) economic and policy barriers; (5) privacy, security, and data governance; and (6)

XSL•FO
RenderX

clinical translation and workflow integration. The framework is intended to inform future research and practice rather than to represent an empirically validated implementation pathway, as illustrated in Figure 5 and detailed in Multimedia Appendix 9.

Figure 5. From monitoring to closed-loop rehabilitation: key opportunity domains for artificial intelligence (AI)–enabled wearable devices in Parkinson disease.

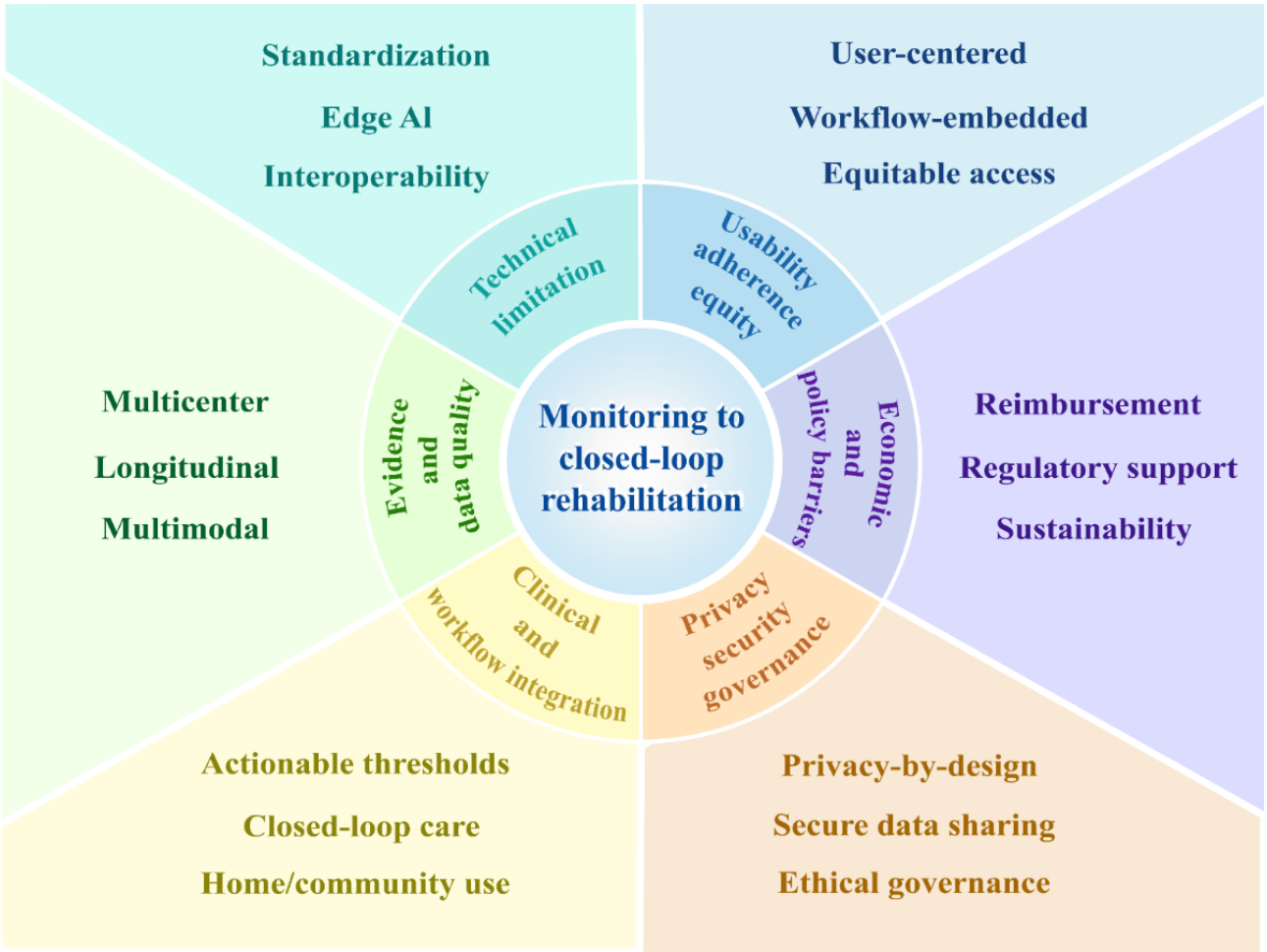

Discussion

Principal Findings

Overview

We synthesized the evidence on AI-enabled wearable devices for PD rehabilitation and motor function assessment, focusing on device types, monitoring indicators, algorithmic approaches, and application characteristics. Overall, the available evidence remains predominantly monitoring- and assessment-oriented, with relatively limited rehabilitation interventions and workflow-integrated applications. Key translational gaps persist, including limited evidence from real-world home/community settings, scarce external validation, underreporting of model calibration and clinical use, and underrepresentation of diverse populations.

Most prior reviews were published between 2020 and 2023 [104-106], and may not fully capture more recent developments, including the wider adoption of remote monitoring devices [107,108], rapid advances in commercial multisensor systems [109], and the evolution of AI models [110]. In contrast to earlier reviews that primarily focused on diagnosis or device technical performance, the present review adopts a rehabilitation- and nursing-oriented perspective and foregrounds evidence relevant to real-world implementation and nurse-led translation. Building on these findings, we propose a conceptual closed-loop framework linking monitoring, assessment, intervention, feedback, and reevaluation, which can be tested and refined in future research.

Study Characteristics

This review covered the 2020-2025 publication window and found that the volume of studies remained consistently high in recent years, suggesting that the field has entered a relatively stable phase of development rather than a period of rapid expansion, consistent with previous reviews [106]. Nevertheless, notable limitations in methodological rigor and sampling design persist: (1) sample sizes varied substantially, with mean values often influenced by outliers [48,80]; (2) sampling was typically based on single-center convenience cohorts [41,64,96]; and (3) subgroup analyses by sex, age, or special populations (such as early-onset or comorbid PD) were largely absent [41,83,84,96]. Many studies enrolled only people with PD, with relatively few including healthy control groups. Outcome reporting tended to prioritize discrimination metrics (eg, accuracy), while model calibration, uncertainty quantification, and clinically actionable thresholding or decision-curve analysis were rarely reported

XSL•FO
RenderX

[57,63,64,72,94-96]. In addition, research has been concentrated in Europe and North America (Table 1), with comparatively limited evidence from low- and middle-income countries and insufficient attention to home/community settings, thereby constraining generalizability and real-world applicability. Overall, persistent methodological limitations—including small, single-center samples, predominantly internal validation, and limited evidence from home/community settings—suggest that growth in study quantity has outpaced advances in translational value.

Application Characteristics and Signal Focus

In terms of application characteristics, AI-enabled wearable devices in PD show a marked imbalance between assessment-oriented and rehabilitation-oriented work, consistent with prior systematic reviews describing the field as "detection-heavy but prediction- and intervention-light" [104-106]. Although several representative studies have demonstrated the feasibility of continuous sensor-based quantification of tremor and bradykinesia [53], objective quantification of medication response [38], and sensor-based early detection of freezing of gait [54,55], most remain at a proof-of-concept stage, with limited evidence for real-world, workflow-integrated, or actionable use in routine care [104,105]. This disparity likely reflects inherent differences in methodological complexity. Assessment-focused studies typically rely on passive data capture and brief, structured tasks that are easier to implement in controlled settings, whereas rehabilitation studies require longer follow-up, personalized feedback, adherence management, and integration with clinical workflows—substantially raising both technical and methodological demands [13,111]. Moreover, most studies adopt motion capture–oriented wearable devices centered on inertial and/or pressure signals, with sensors preferentially placed on the lower limbs or the wrist. This design choice reduces data-collection burden and improves task reproducibility, thereby reinforcing assessment paradigms built around controlled tasks; however, it also structurally steers research toward motor outcomes (eg, gait, freezing of gait, tremor, and bradykinesia), making sleep, mood, and cognition more difficult to capture continuously and comprehensively [38,46,53]. Meanwhile, many applications still rely on short, structured assessments, which further constrain the continuous capture of nonmotor domains in real-world contexts [16,19]. Taken together, the evidence continues to prioritize motor-domain tasks, with insufficient coverage of nonmotor symptoms and their continuous assessment in real-world contexts. As a result, current applications fall short of meeting the comprehensive needs of long-term chronic disease management.

AI and Algorithmic Characteristics

In our review, SVMs and CNNs are the most frequently used approaches, while decision trees and logistic regression are commonly used as interpretable baseline models. Currently, the field reflects a pattern of "traditional ML dominance, parallel advances in convolutional DL, and early-stage development of sequential models and emerging paradigms" [38,40,50,57,68,74], which aligns with the ML/DL distribution observed in our results. This distribution may be attributable to differences in how algorithms adapt to data scale and structure. SVMs tend to provide more stable performance in small-sample, feature-engineered wearable data settings, whereas CNNs are better suited to end-to-end feature learning from high-frequency inertial signals, thereby reducing reliance on manual feature engineering [112,113]. Sequential approaches (eg, LSTM networks and a small number of HMMs) have been explored for gait-phase and state-recognition tasks, but remain relatively uncommon. This suggests that temporal dependencies and state-transition dynamics are still underutilized [36,52,72,73]. Consistent with these constraints, validation remains largely internal (eg, cross-validation), while independent external validation is rare, limiting confidence in generalizability [105]. More complex architectures (eg, self-attention mechanisms and extreme gradient boosting) also appear underused, potentially reflecting feasibility constraints for on-device or real-world deployment [39,75,80,92-96,98].

In summary, AI-enabled wearable devices for PD rehabilitation and assessment remain characterized by a strong emphasis on monitoring, with comparatively weak integration into closed-loop systems. From the perspective of the *JMIR mHealth and uHealth* taxonomy, most devices included in this review are better characterized as mobile health tools, primarily focused on individual-level monitoring and assessment [114]. Although these functions provide valuable objective data, there is still insufficient evidence that they have been systematically embedded into routine nurse-led rehabilitation workflows or broader health care delivery systems as mature eHealth solutions. Building on these findings, future research should move beyond stand-alone monitoring applications toward workflow-integrated eHealth systems with decision-support functions that can be tested in nurse-led rehabilitation pathways.

Practical Recommendations

Implications for Clinical and Nursing Practice

In both clinical and home-based rehabilitation contexts, nurses increasingly require continuous and objective patient information to enhance the precision of assessment and monitoring, and strengthen patient engagement in the rehabilitation process [115]. AI-enabled wearable devices represent a promising approach to addressing this need. Within clinical rehabilitation settings, continuous monitoring of gait, tremor, sleep, and activity levels can augment routine nursing surveillance, facilitate earlier detection of functional deterioration, and support timely intervention or prioritization—moving beyond the inherently fragmented nature of episodic in-clinic observations [25,115]. In remote or home-based rehabilitation, the continuous and objective data generated by wearables can be used to evaluate therapeutic response and adherence, while personalized feedback, symptom prompts, medication-taking support, and nurse-led follow-up via telerehabilitation platforms promote patient self-management [26,115]. Specifically, when algorithmic detection identifies abnormal changes in patients' daily activity levels or tremor severity that exceed an individualized threshold, an alert can be triggered on the nursing interface. This may prompt nurses to rapidly assess symptom changes, medication adherence, and

XSL•FO
RenderX

potential adverse effects and, as appropriate, arrange earlier review, reinforce medication and safety guidance, or adjust follow-up plans [97,115]. Likewise, when long-term monitoring trend reports indicate sustained deterioration in motor function–related indicators or marked fluctuations in motor symptoms, these data can provide objective evidence for the rehabilitation team to deliver targeted health education, adjust the frequency of follow-up, and, when needed, coordinate involvement of rehabilitation therapists or a multidisciplinary team [56,89,99]. These functions align closely with core nursing responsibilities in patient education, symptom monitoring, and long-term chronic disease management, and may help reinforce the full nursing care cycle of assessment, diagnosis, planning, implementation, and evaluation [116].

Priorities for Future Research and Translation

To advance the field, future research may consider five priority areas. (1) Strengthening the evidence base through multicenter, long-term real-world studies with preregistration, appropriate sample size estimation, and attrition management, alongside reporting of external validation, model calibration, uncertainty, and clinical use [117-120]. (2) Enhancing model feasibility through multimodal signal fusion and data-efficient learning strategies to support practical real-world deployment [104]. (3) Exploring conceptual closed-loop rehabilitation frameworks that integrate assessment and intervention, while examining the feasibility of actionable thresholds and their potential integration into routine clinical workflows [120]. (4) Improving device comfort and usability to balance patient adherence with caregiver and clinician workload, address the digital divide, and enhance accessibility [119,120]. (5) Strengthening privacy and data security frameworks, alongside improving insurance reimbursement mechanisms and regulatory support, to promote responsible cross-institutional data sharing [121,122], including explicitly addressing equity, responsible data use, and related ethical considerations in real-world nursing and rehabilitation contexts.

Limitations

This study has several limitations. First, no prospective protocol was registered; however, this is an acceptable practice for scoping reviews, and we clarify this point to enhance methodological transparency. Second, although a systematic search and screening process was conducted, the scope of databases, search strategies, and inclusion criteria may have led to the omission of eligible studies, posing a risk of incompleteness or potential bias. In addition, most included studies had small sample sizes, single-center designs, short follow-up durations, and limited external validation, alongside substantial heterogeneity across study populations, task paradigms, signal sources, algorithmic approaches, and outcome measures. This heterogeneity may limit cross-study comparability and the generalizability of findings. Given the predominantly exploratory and proof-of-concept nature of the included studies, the closed-loop rehabilitation framework and related recommendations proposed in this review should be interpreted as conceptual and practice-informing rather than empirically validated implementation pathways. Finally, the search was restricted to English and Chinese peer-reviewed publications, with limited inclusion of gray literature or other languages, which may introduce time-lag, linguistic, and publication biases. In particular, studies reporting null or negative findings, early-stage failures, or limited model performance may be less likely to be published and therefore underrepresented, potentially leading to an overestimation of the maturity, effectiveness, and translational readiness of AI-enabled wearable technologies.

Conclusions

This study systematically summarized recent evidence on AI-enabled wearable devices for motor function assessment and rehabilitation in people with PD. By combining a scoping review with an evidence map and adopting a rehabilitation-oriented perspective, we synthesized evidence across multiple dimensions, including application aims, sensing modalities, AI methods, and validation practices, thereby identifying key translational gaps between proof-of-concept studies and real-world rehabilitation workflows. Compared with previous reviews that primarily focused on monitoring functions or device performance, this study places greater emphasis on rehabilitation applications and nurse-led translation into practice, and further clarifies the key actionable gaps that limit real-world implementation. On this basis, a conceptually integrated “challenges and opportunities” framework is proposed to inform the design, evaluation, and reporting of devices and algorithms, and to highlight considerations for developing workflow-integrated, decision-support wearable systems. From a real-world perspective, these findings may support continuity of rehabilitation across clinical, home, and community settings by enabling nurses to deliver continuous monitoring, personalized follow-up, and timely intervention, ultimately improving the efficiency and accessibility of PD rehabilitation management.

Acknowledgments

Generative artificial intelligence tools (ChatGPT [OpenAI] and DeepSeek [Hangzhou DeepSeek Artificial Intelligence Co, Ltd]) were used solely under full human supervision to assist with proofreading/editing and translation, including terminology conversion, textual clarification, and translation of reviewer comments. These tools were not used for study design, data processing, analysis, or the generation of scientific content. All AI-assisted outputs were reviewed and revised by the authors, who take full responsibility for the accuracy and integrity of the manuscript.

XSL•FO
RenderX

Data Availability

This scoping review did not generate any primary data. All data extracted and synthesized in this review were obtained from previously published studies, which are publicly available. No additional datasets were created or analyzed.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR checklist.
[DOCX File , 71 KB - jmir_v28i1e85596_app1.docx]

Multimedia Appendix 2
Eligibility criteria.
[DOCX File , 19 KB - jmir_v28i1e85596_app2.docx]

Multimedia Appendix 3
Search strategy.
[DOCX File , 30 KB - jmir_v28i1e85596_app3.docx]

Multimedia Appendix 4
Extracted data form.
[DOCX File , 19 KB - jmir_v28i1e85596_app4.docx]

Multimedia Appendix 5
Quality assessment of the included studies.
[DOCX File , 72 KB - jmir_v28i1e85596_app5.docx]

Multimedia Appendix 6
Detailed characteristics of included studies using artificial intelligence–enabled wearable devices in Parkinson disease research.
[DOCX File , 50 KB - jmir_v28i1e85596_app6.docx]

Multimedia Appendix 7
Detailed characteristics of artificial intelligence–enabled wearable devices used in the included Parkinson disease studies.
[DOCX File , 59 KB - jmir_v28i1e85596_app7.docx]

Multimedia Appendix 8
Detailed characteristics of artificial intelligence algorithms used in the included wearable device studies for Parkinson disease.
[DOCX File , 50 KB - jmir_v28i1e85596_app8.docx]

Multimedia Appendix 9
Challenges and opportunities of artificial intelligence–enabled wearable devices in Parkinson disease research.
[DOCX File , 20 KB - jmir_v28i1e85596_app9.docx]

References

1. Kouli A, Torsney KM, Kuan WL. Parkinson's disease: etiology, neuropathology, and pathogenesis. In: Stoker TB, Greenland JC, editors. Parkinson's Disease: Pathogenesis and Clinical Aspects Internet. Brisbane (AU): Codon Publications; 2018.
2. Abusrair AH, Elsekaily W, Bohlega S. Tremor in Parkinson's disease: from pathophysiology to advanced therapies. Tremor Other Hyperkinet Mov (NY) 2022;12:29 [FREE Full text] [doi: 10.5334/tohm.712] [Medline: 36211804]
3. Dorsey ER, Sherer T, Okun MS, Bloem BR. The emerging evidence of the Parkinson pandemic. J Parkinsons Dis 2018;8(s1):S3-S8 [FREE Full text] [doi: 10.3233/JPD-181474] [Medline: 30584159]
4. Qi S, Yin P, Wang L, Qu M, Kan GL, Zhang H, et al. Prevalence of Parkinson's disease: a community-based study in China. Mov Disord 2021;36(12):2940-2944. [doi: 10.1002/mds.28762] [Medline: 34390510]
5. Zhang ST, Jia XF, Zhang JG. Research progress on the association between dietary nutrition, lifestyle, and Parkinson's disease [in Chinese]. J Environ Occup Med 2025;40(2):149-155 [FREE Full text]
6. Garg D, Dhamija RK. Rehabilitation in Parkinson's disease: current status and future directions. Ann Mov Disord 2020;3(2):79-85. [doi: 10.4103/aomd.aomd_1_20]

XSL•FO
RenderX

7. Yoon SY. Update on Parkinson's disease rehabilitation. Brain Neurorehabil 2022;15(2):e15 [FREE Full text] [doi: 10.12786/bn.2022.15.e15] [Medline: 36743207]
8. Zhang L, Liu Y, Han W. Advance of rehabilitation for Parkinson's disease [in Chinese]. Chin J Contemp Neurol Neurosurg 2017;17(5):328-332 [FREE Full text]
9. Pastana Ramos LF, Vilacorta-Pereira TDCS, Duarte JDS, Yamada ES, Santos-Lobato BL. Feasibility and effectiveness of a remote individual rehabilitation program for people with Parkinson's disease living in the Brazilian Amazon: a randomized clinical trial. Front Neurol 2023;14:1244661 [FREE Full text] [doi: 10.3389/fneur.2023.1244661] [Medline: 37693755]
10. Sevcenko K, Lindgren I. The effects of virtual reality training in stroke and Parkinson's disease rehabilitation: a systematic review and a perspective on usability. Eur Rev Aging Phys Act 2022;19(1):4 [FREE Full text] [doi: 10.1186/s11556-022-00283-3] [Medline: 35078401]
11. Fedor S, Lewis R, Pedrelli P, Mischoulon D, Curtiss J, Picard RW. Wearable technology in clinical practice for depressive disorder. N Engl J Med 2023;389(26):2457-2466. [doi: 10.1056/NEJMra2215898] [Medline: 38157501]
12. Maceira-Elvira P, Popa T, Schmid AC, Hummel FC. Wearable technology in stroke rehabilitation: towards improved diagnosis and treatment of upper-limb motor impairment. J Neuroeng Rehabil 2019;16(1):142 [FREE Full text] [doi: 10.1186/s12984-019-0612-y] [Medline: 31744553]
13. Porciuncula F, Roto AV, Kumar D, Davis I, Roy S, Walsh CJ, et al. Wearable movement sensors for rehabilitation: a focused review of technological and clinical advances. PM R 2018;10(9 Suppl 2):S220-S232 [FREE Full text] [doi: 10.1016/j.pmrj.2018.06.013] [Medline: 30269807]
14. Rodríguez-Fernández A, Lobo-Prat J, Font-Llagunes JM. Systematic review on wearable lower-limb exoskeletons for gait training in neuromuscular impairments. J Neuroeng Rehabil 2021;18(1):22 [FREE Full text] [doi: 10.1186/s12984-021-00815-5] [Medline: 33526065]
15. Meena JS, Choi SB, Jung S, Kim J. Electronic textiles: new age of wearable technology for healthcare and fitness solutions. Mater Today Bio 2023;19:100565 [FREE Full text] [doi: 10.1016/j.mtbio.2023.100565] [Medline: 36816602]
16. Huang X, Xue Y, Ren S, Wang F. Sensor-based wearable systems for monitoring human motion and posture: a review. Sensors (Basel) 2023;23(22):9047 [FREE Full text] [doi: 10.3390/s23229047] [Medline: 38005436]
17. Adams JL, Dinesh K, Snyder CW, Xiong M, Tarolli CG, Sharma S, et al. A real-world study of wearable sensors in Parkinson's disease. NPJ Parkinsons Dis 2021;7(1):106 [FREE Full text] [doi: 10.1038/s41531-021-00248-w] [Medline: 34845224]
18. Lu R, Xu Y, Li X, Fan Y, Zeng W, Tan Y, et al. Evaluation of wearable sensor devices in Parkinson's disease: a review of current status and future prospects. Parkinsons Dis 2020;2020:4693019 [FREE Full text] [doi: 10.1155/2020/4693019] [Medline: 33029343]
19. Bougea A. Application of wearable sensors in Parkinson’s disease: state of the art. JSAN 2025;14(2):23. [doi: 10.3390/jsan14020023]
20. Sotirakis C, Su Z, Brzezicki MA, Conway N, Tarassenko L, FitzGerald JJ, et al. Identification of motor progression in Parkinson's disease using wearable sensors and machine learning. NPJ Parkinsons Dis 2023;9(1):142 [FREE Full text] [doi: 10.1038/s41531-023-00581-2] [Medline: 37805655]
21. Abujrida H, Agu E, Pahlavan K. Machine learning-based motor assessment of Parkinson's disease using postural sway, gait and lifestyle features on crowdsourced smartphone data. Biomed Phys Eng Express 2020;6(3):035005. [doi: 10.1088/2057-1976/ab39a8] [Medline: 33438650]
22. Benyoucef Y, Melliti I, Harmouch J, Asadi B, Del Mastro A, Lapuente-Hernández D, et al. Wearable sensors and artificial intelligence for the diagnosis of Parkinson's disease. J Clin Med 2025;14(12):4207 [FREE Full text] [doi: 10.3390/jcm14124207] [Medline: 40565952]
23. Farabolini G, Baldini N, Pagano A, Andrenelli E, Pepa L, Morone G, et al. Continuous movement monitoring at home through wearable devices: a systematic review. Sensors (Basel) 2025;25(16):4889 [FREE Full text] [doi: 10.3390/s25164889] [Medline: 40871752]
24. Islam MS, Rahman W, Abdelkader A, Lee S, Yang PT, Purks JL, et al. Using AI to measure Parkinson's disease severity at home. NPJ Digit Med 2023;6(1):156 [FREE Full text] [doi: 10.1038/s41746-023-00905-9] [Medline: 37608206]
25. Moreau C, Rouaud T, Grabli D, Benatru I, Remy P, Marques A, et al. Overview on wearable sensors for the management of Parkinson's disease. NPJ Parkinsons Dis 2023;9(1):153 [FREE Full text] [doi: 10.1038/s41531-023-00585-y] [Medline: 37919332]
26. Del Din S, Kirk C, Yarnall AJ, Rochester L, Hausdorff JM. Body-worn sensors for remote monitoring of Parkinson's disease motor symptoms: vision, state of the art, and challenges ahead. J Parkinsons Dis 2021;11(s1):S35-S47 [FREE Full text] [doi: 10.3233/JPD-202471] [Medline: 33523020]
27. Li H, Zecca M, Huang J. Evaluating the utility of wearable sensors for the early diagnosis of Parkinson disease: systematic review. J Med Internet Res 2025;27:e69422 [FREE Full text] [doi: 10.2196/69422] [Medline: 40690754]
28. Gutowski T, Stodulska O, Ćwiklińska A, Gutowska K, Kopeć K, Betka M, et al. Machine learning-based assessment of Parkinson's disease symptoms using wearable and smartphone sensors. Sensors (Basel) 2025;25(16):4924 [FREE Full text] [doi: 10.3390/s25164924] [Medline: 40871788]

XSL•FO
RenderX

29. Wu X, Ma L, Wei P, Shan Y, Chan P, Wang K, et al. Wearable sensor devices can automatically identify the ON-OFF status of patients with Parkinson's disease through an interpretable machine learning model. Front Neurol 2024;15:1387477 [FREE Full text] [doi: 10.3389/fneur.2024.1387477] [Medline: 38751881]
30. Tischendorf T, Heitmann-Möller A, Ruppert S, Marchwacka M, Schaffrin S, Schaal T, et al. Sustainable integration of digitalisation in nursing education-an international scoping review. Front Health Serv 2024;4:1344021 [FREE Full text] [doi: 10.3389/frhs.2024.1344021] [Medline: 38665930]
31. AlDossary MBD, Al-Enezi MAH, Aldosary MM. The new frontier in care: a systematic review of nursing implications of remote patient monitoring (RPM) for managing chronic diseases. Saudi J Med Public Health 2024;1(2):582-589 [FREE Full text] [doi: 10.64483/jmph-167]
32. Tricco AC, Antony J, Zarin W, Strifler L, Ghassemi M, Ivory J, et al. A scoping review of rapid review methods. BMC Med 2015;13:224 [FREE Full text] [doi: 10.1186/s12916-015-0465-6] [Medline: 26377409]
33. Miake-Lye IM, Hempel S, Shanman R, Shekelle PG. What is an evidence map? A systematic review of published evidence maps and their definitions, methods, and products. Syst Rev 2016;5:28 [FREE Full text] [doi: 10.1186/s13643-016-0204-x] [Medline: 26864942]
34. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
35. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
36. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions. Chichester (UK): John Wiley & Sons; 2008.
37. Hong QN, Gonzalez-Reyes A, Pluye P. Improving the usefulness of a tool for appraising the quality of qualitative, quantitative and mixed methods studies, the Mixed Methods Appraisal Tool (MMAT). J Eval Clin Pract 2018;24(3):459-467. [doi: 10.1111/jep.12884] [Medline: 29464873]
38. Aghanavesi S, Westin J, Bergquist F, Nyholm D, Askmark H, Aquilonius SM, et al. A multiple motion sensors index for motor state quantification in Parkinson's disease. Comput Methods Programs Biomed 2020;189:105309. [doi: 10.1016/j.cmpb.2019.105309] [Medline: 31982667]
39. Li Y, Bai Q, Yang X, Zhou X, Sun Y, Yao Z. An abnormal gait monitoring system for patients with Parkinson's disease based on wearable devices. Piscataway, NJ: IEEE; 2022 Presented at: 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI); 2022 November 05-07; Beijing, China p. -15. [doi: 10.1109/cisp-bmei56279.2022.9980005]
40. Ghayvat H, Awais M, Geddam R, Quasim MT, Khowaja SA, Dev K. AiCareGaitRehabilitation: multi-modalities sensor data fusion for AI-IoT enabled realtime electrical stimulation device for pre-fog and post-fog to person with Parkinson’s disease. Information Fusion 2025;122:103155. [doi: 10.1016/j.inffus.2025.103155]
41. Delgado-Terán JD, Hilbrants K, Mahmutović D, Silva de Lima AL, Wezel RJAV, Heida T. Ankle sensor-based detection of freezing of gait in Parkinson's disease in semi-free living environments. Sensors (Basel) 2025;25(6):1895 [FREE Full text] [doi: 10.3390/s25061895] [Medline: 40293010]
42. Liu X, Zhang X, Li J, Pan W, Sun Y, Lin S, et al. Automated UPDRS gait scoring using wearable sensor fusion and deep learning. Bioengineering (Basel) 2025;12(7):686 [FREE Full text] [doi: 10.3390/bioengineering12070686] [Medline: 40722378]
43. Han Y, Liu X, Zhang N, Zhang X, Zhang B, Wang S, et al. Automatic assessments of Parkinsonian gait with wearable sensors for human assistive systems. Sensors (Basel) 2023;23(4):2104 [FREE Full text] [doi: 10.3390/s23042104] [Medline: 36850705]
44. Peraza LR, Kinnunen KM, McNaney R, Craddock IJ, Whone AL, Morgan C, et al. An automatic gait analysis pipeline for wearable sensors: a pilot study in Parkinson's disease. Sensors (Basel) 2021;21(24):8286 [FREE Full text] [doi: 10.3390/s21248286] [Medline: 34960379]
45. Sánchez-Fernández LP, Sánchez-Pérez LA, Martínez-Hernández JM. Computer model for gait assessments in Parkinson's patients using a fuzzy inference model and inertial sensors. Artif Intell Med 2025;160:103059. [doi: 10.1016/j.artmed.2024.103059] [Medline: 39732050]
46. Channa A, Popescu N, Faisal M. Parkinson's disease gait evaluation leveraging wearable insoles and a deep learning approach. : IEEE; 2022 Presented at: 8th International Conference on Control, Decision and Information Technologies (CoDIT); 2022 May 17-20; Istanbul, Turkey p. 543-549. [doi: 10.1109/codit55151.2022.9804064]
47. Sigcha L, Costa N, Pavón I, Costa S, Arezes P, López JM, et al. Deep learning approaches for detecting freezing of gait in Parkinson's disease patients through on-body acceleration sensors. Sensors (Basel) 2020;20(7):1895 [FREE Full text] [doi: 10.3390/s20071895] [Medline: 32235373]
48. Dvorani A, Wiesener C, Salchow-Hommen C, Jochner M, Spieker L, Skrobot M, et al. On-demand gait-synchronous electrical cueing in Parkinson's disease using machine learning and edge computing: a pilot study. IEEE Open J Eng Med Biol 2024;5:306-315 [FREE Full text] [doi: 10.1109/OJEMB.2024.3390562] [Medline: 38766539]

XSL•FO
RenderX

49. Bikias T, Iakovakis D, Hadjidimitriou S, Charisis V, Hadjileontiadis LJ. DeepFoG: an IMU-based detection of freezing of gait episodes in Parkinson's disease patients via deep learning. Front Robot AI 2021;8:537384 [FREE Full text] [doi: 10.3389/frobt.2021.537384] [Medline: 34113654]
50. Aich S, Pradhan PM, Chakraborty S, Kim HC, Kim HT, Lee HG, et al. Design of a machine learning-assisted wearable accelerometer-based automated system for studying the effect of dopaminergic medicine on gait characteristics of Parkinson's patients. J Healthc Eng 2020;2020:1823268 [FREE Full text] [doi: 10.1155/2020/1823268] [Medline: 32148741]
51. Shi B, Tay A, Au WL, Tan DML, Chia NSY, Yen S. Detection of freezing of gait using convolutional neural networks and data from lower limb motion sensors. IEEE Trans Biomed Eng 2022;69(7):2256-2267. [doi: 10.1109/TBME.2022.3140258] [Medline: 34986092]
52. Ezhilarasi J, Senthil Kumar T. Develop a novel, faster mask region-based convolutional neural network model with leave-one-subject-out to predict freezing of gait abnormalities of Parkinson's disease. Neural Comput Applic 2025;37(7):5441-5457. [doi: 10.1007/s00521-024-10832-9]
53. Mahadevan N, Demanuele C, Zhang H, Volfson D, Ho B, Erb MK, et al. Development of digital biomarkers for resting tremor and bradykinesia using a wrist-worn wearable device. NPJ Digit Med 2020;3:5 [FREE Full text] [doi: 10.1038/s41746-019-0217-7] [Medline: 31970290]
54. Chan LLY, Yang S, Aswani M, Kark L, Henderson E, Lord SR, et al. Development, validation, and limits of freezing of gait detection using a single waist-worn device. IEEE Trans Biomed Eng 2024;71(10):3024-3031. [doi: 10.1109/TBME.2024.3407059] [Medline: 38814761]
55. Pardoel S, Shalin G, Nantel J, Lemaire ED, Kofman J. Early detection of freezing of gait during walking using inertial measurement unit and plantar pressure distribution data. Sensors (Basel) 2021;21(6):2246 [FREE Full text] [doi: 10.3390/s21062246] [Medline: 33806984]
56. Hssayeni MD, Jimenez-Shahed J, Burack MA, Ghoraani B. Ensemble deep model for continuous estimation of unified Parkinson's disease rating scale III. Biomed Eng Online 2021;20(1):32 [FREE Full text] [doi: 10.1186/s12938-021-00872-w] [Medline: 33789666]
57. Tsakanikas V, Ntanis A, Rigas G, Androutsos C, Boucharas D, Tachos N, et al. Evaluating gait impairment in Parkinson's disease from instrumented insole and IMU sensor data. Sensors (Basel) 2023;23(8):3902 [FREE Full text] [doi: 10.3390/s23083902] [Medline: 37112243]
58. Brand YE, Schwartz D, Gazit E, Buchman AS, Gilad-Bachrach R, Hausdorff JM. Gait detection from a wrist-worn sensor using machine learning methods: daily living study in older adults and people with Parkinson's disease. Sensors (Basel) 2022;22(18):7094 [FREE Full text] [doi: 10.3390/s22187094] [Medline: 36146441]
59. Pardoel S, Shalin G, Lemaire ED, Kofman J, Nantel J. Grouping successive freezing of gait episodes has neutral to detrimental effect on freeze detection and prediction in Parkinson's disease. PLoS One 2021;16(10):e0258544 [FREE Full text] [doi: 10.1371/journal.pone.0258544] [Medline: 34637473]
60. Li D, Hallack A, Gwilym S, Li D, Hu MT, Cantley J. Investigating gait-responsive somatosensory cueing from a wearable device to improve walking in Parkinson's disease. Biomed Eng Online 2023;22(1):108 [FREE Full text] [doi: 10.1186/s12938-023-01167-y] [Medline: 37974260]
61. Aghanavesi S, Bergquist F, Nyholm D, Senek M, Memedi M. Motion sensor-based assessment of Parkinson's disease motor symptoms during leg agility tests: results from levodopa challenge. IEEE J Biomed Health Inform 2020;24(1):111-119. [doi: 10.1109/JBHI.2019.2898332] [Medline: 30763248]
62. Atrsaei A, Hansen C, Elshehabi M, Solbrig S, Berg D, Liepelt-Scarfone I, et al. Effect of fear of falling on mobility measured during lab and daily activity assessments in Parkinson's disease. Front Aging Neurosci 2021;13:722830 [FREE Full text] [doi: 10.3389/fnagi.2021.722830] [Medline: 34916920]
63. Shawen N, O'Brien MK, Venkatesan S, Lonini L, Simuni T, Hamilton JL, et al. Role of data measurement characteristics in the accurate detection of Parkinson's disease symptoms using wearable sensors. J Neuroeng Rehabil 2020;17(1):52 [FREE Full text] [doi: 10.1186/s12984-020-00684-4] [Medline: 32312287]
64. Naghavi N, Wade E. Towards real-time prediction of freezing of gait in patients with Parkinson's disease: a novel deep one-class classifier. IEEE J Biomed Health Inform 2022;26(4):1726-1736. [doi: 10.1109/JBHI.2021.3103071] [Medline: 34375292]
65. Zampogna A, Borzì L, Rinaldi D, Artusi CA, Imbalzano G, Patera M, et al. Unveiling the unpredictable in Parkinson's disease: sensor-based monitoring of dyskinesias and freezing of gait in daily life. Bioengineering (Basel) 2024;11(5):440 [FREE Full text] [doi: 10.3390/bioengineering11050440] [Medline: 38790307]
66. Reches T, Dagan M, Herman T, Gazit E, Gouskova NA, Giladi N, et al. Using wearable sensors and machine learning to automatically detect freezing of gait during a FOG-provoking test. Sensors (Basel) 2020;20(16):4474 [FREE Full text] [doi: 10.3390/s20164474] [Medline: 32785163]
67. Slemenšek J, Geršak J, Bratina B, van Midden VM, Pirtošek Z, Šafarič R. Wearable online freezing of gait detection and cueing system. Bioengineering (Basel) 2024;11(10):1048 [FREE Full text] [doi: 10.3390/bioengineering11101048] [Medline: 39451423]
68. Borzì L, Mazzetta I, Zampogna A, Suppa A, Irrera F, Olmo G. Predicting axial impairment in parkinson's disease through a single inertial sensor. Sensors (Basel) 2022;22(2):412 [FREE Full text] [doi: 10.3390/s22020412] [Medline: 35062375]

XSL•FO
RenderX

69. Wang W, Lin J, Le X, Li Y, Liu T, Pan L, et al. Addressing multiple challenges in early gait freezing prediction for Parkinson's disease: a practical deep learning approach. IEEE J Biomed Health Inform 2025;29(9):6251-6262. [doi: 10.1109/JBHI.2024.3522664] [Medline: 40030782]
70. Shuqair M, Jimenez-Shahed J, Ghoraani B. Advancing Parkinson's disease management through multi-shared-task self-supervised signal processing. : IEEE; 2024 Presented at: 58th Asilomar Conference on Signals, Systems, and Computers; 2024 October 27-30; Pacific Grove, CA, USA p. 957-961. [doi: 10.1109/ieeeconf60004.2024.10942713]
71. Kanellos FS, Tsamis KI, Rigas G, Simos YV, Katsenos AP, Kartsakalis G, et al. Clinical evaluation in Parkinson's disease: is the golden standard shiny enough? Sensors (Basel) 2023;23(8):3807 [FREE Full text] [doi: 10.3390/s23083807] [Medline: 37112154]
72. Roth N, Küderle A, Ullrich M, Gladow T, Marxreiter F, Klucken J, et al. Hidden Markov model based stride segmentation on unsupervised free-living gait data in Parkinson's disease patients. J Neuroeng Rehabil 2021;18(1):93 [FREE Full text] [doi: 10.1186/s12984-021-00883-7] [Medline: 34082762]
73. Perez-Ibarra JC, Siqueira AAG, Krebs HI. Identification of gait events in healthy subjects and with Parkinson's disease using inertial sensors: an adaptive unsupervised learning approach. IEEE Trans Neural Syst Rehabil Eng 2020;28(12):2933-2943. [doi: 10.1109/TNSRE.2020.3039999] [Medline: 33237863]
74. Meng L, Pang J, Yang Y, Chen L, Xu R, Ming D. Inertial-based gait metrics during turning improve the detection of early-stage Parkinson's disease patients. IEEE Trans Neural Syst Rehabil Eng 2023;31:1472-1482. [doi: 10.1109/TNSRE.2023.3237903] [Medline: 37021993]
75. Dattola S, Ielo A, Quartarone A, De Cola MC. Integrating wearable sensor signal processing with unsupervised learning methods for tremor classification in Parkinson's disease. Bioengineering (Basel) 2025;12(1):37 [FREE Full text] [doi: 10.3390/bioengineering12010037] [Medline: 39851311]
76. Zhao Y, Liu Y, Lu W, Li J, Shan P, Lian C, et al. Intelligent IoT anklets for monitoring the assessment of Parkinson’s diseases. IEEE Sensors J 2023;23(24):31523-31536. [doi: 10.1109/jsen.2023.3331277]
77. Li Y, Yin J, Liu S, Xue B, Shokoohi C, Ge G, et al. Learning hand kinematics for Parkinson's disease assessment using a multimodal sensor glove. Adv Sci (Weinh) 2023;10(20):e2206982 [FREE Full text] [doi: 10.1002/advs.202206982] [Medline: 37150855]
78. Yan F, Gong J, Zhang Q, He H. Learning motion primitives for the quantification and diagnosis of mobility deficits. IEEE Trans Biomed Eng 2024;71(12):3339-3349. [doi: 10.1109/TBME.2024.3404357] [Medline: 38776195]
79. Evers LJW, Raykov YP, Heskes TM, Krijthe JH, Bloem BR, Little MA. Passive monitoring of Parkinson tremor in daily life: a prototypical network approach. Sensors (Basel) 2025;25(2):366 [FREE Full text] [doi: 10.3390/s25020366] [Medline: 39860736]
80. Greene BR, Premoli I, McManus K, McGrath D, Caulfield B. Predicting fall counts using wearable sensors: a novel digital biomarker for Parkinson's disease. Sensors (Basel) 2021;22(1):54 [FREE Full text] [doi: 10.3390/s22010054] [Medline: 35009599]
81. Ricci M, Lazzaro GD, Errico V, Pisani A, Giannini F, Saggio G. The impact of wearable electronics in assessing the effectiveness of levodopa treatment in Parkinson's disease. IEEE J Biomed Health Inform 2022;26(7):2920-2928. [doi: 10.1109/JBHI.2022.3160103] [Medline: 35316198]
82. Ghosh N, Banerjee I. IoT-based freezing of gait detection using grey relational analysis. Internet Things 2021;13:100068. [doi: 10.1016/j.iot.2019.100068]
83. Lin ZR. Quantitative assessment of Parkinson's motor symptoms based on machine learning. Sci Technol Innov Appl 2022;12(34):50-54. [doi: 10.19981/j.CN23-1581/G3.2022.34.013]
84. Sun Y, Rong Z, Wang F. Automatic quantitative study of motor function in patients with Parkinson's disease based on wearable inertial sensors [in Chinese]. China Med Equip 2023;38(10):27-32 [FREE Full text]
85. Li YM. Quantitative Assessment of Parkinson's Motor Symptoms Based on Deep Learning [dissertation]. Lanzhou, China: Lanzhou Jiaotong University; 2023.
86. Donisi L, Cesarelli G, Balbi P, Provitera V, Lanzillo B, Coccia A, et al. Positive impact of short-term gait rehabilitation in Parkinson patients: a combined approach based on statistics and machine learning. Math Biosci Eng 2021;18(5):6995-7009 [FREE Full text] [doi: 10.3934/mbe.2021348] [Medline: 34517568]
87. Sotirakis C, Brzezicki MA, Patel S, Conway N, FitzGerald JJ, Antoniades CA. Predicting future fallers in Parkinson's disease using kinematic data over a period of 5 years. NPJ Digit Med 2024;7(1):345 [FREE Full text] [doi: 10.1038/s41746-024-01311-5] [Medline: 39638907]
88. Hong G, Mao F, Zhang M, Zhang F, Wang X, Ren K, et al. Modeling and validation of wearable sensor-based gait parameters in Parkinson's disease patients with cognitive impairment. Front Aging Neurosci 2025;17:1590224 [FREE Full text] [doi: 10.3389/fnagi.2025.1590224] [Medline: 40787670]
89. Chen M, Sun Z, Xin T, Chen Y, Su F. An interpretable deep learning optimized wearable daily detection system for Parkinson's disease. IEEE Trans Neural Syst Rehabil Eng 2023;31:3937-3946. [doi: 10.1109/TNSRE.2023.3314100] [Medline: 37695969]

XSL•FO
RenderX

90. Channa A, Ruggeri G, Ifrim RC, Mammone N, Iera A, Popescu N. Cloud-connected bracelet for continuous monitoring of Parkinson's disease patients: integrating advanced wearable technologies and machine learning. Electronics 2024;13(6):1002. [doi: 10.3390/electronics13061002]
91. Singh M, Prakash P, Kaur R, Sowers R, Brašić JR, Hernandez ME. A deep learning approach for automatic and objective grading of the motor impairment severity in Parkinson's disease for use in tele-assessments. Sensors (Basel) 2023;23(21):9004 [FREE Full text] [doi: 10.3390/s23219004] [Medline: 37960703]
92. Weikert T, Li Y, Paez-Granados D, Easthope CA. Automated prediction of item-level ARAT scores from wearable sensors. : IEEE; 2025 Presented at: Proceedings of the 2025 IEEE International Conference on Rehabilitation Robotics (ICORR); 2025 May 12–16; Chicago, IL, USA p. 1239-1244. [doi: 10.1109/icorr66766.2025.11063162]
93. Yang PK, Filtjens B, Ginis P, Goris M, Nieuwboer A, Gilat M, et al. Freezing of gait assessment with inertial measurement units and deep learning: effect of tasks, medication states, and stops. J Neuroeng Rehabil 2024;21(1):24 [FREE Full text] [doi: 10.1186/s12984-024-01320-1] [Medline: 38350964]
94. Esfahani AH, Dyka Z, Ortmann S, Langendorfer P. Impact of data preparation in freezing of gait detection using feature-less recurrent neural network. IEEE Access 2021;9:138120-138131. [doi: 10.1109/access.2021.3117543]
95. Sun H, Ye Q, Xia Y. Predicting freezing of gait in patients with Parkinson's disease by combination of manually-selected and deep learning features. Biomed Signal Process Control 2024;88:105639. [doi: 10.1016/j.bspc.2023.105639]
96. Dvorani A, Waldheim V, Jochner MCE, Salchow-Hömmen C, Meyer-Ohle J, Kühn AA, et al. Real-time detection of freezing motions in Parkinson's patients for adaptive gait phase synchronous cueing. Front Neurol 2021;12:720516 [FREE Full text] [doi: 10.3389/fneur.2021.720516] [Medline: 34938252]
97. Khan U, Riaz Q, Hussain M, Zeeshan M, Krüger B. Towards effective Parkinson's monitoring: movement disorder detection and symptom identification using wearable inertial sensors. Algorithms 2025;18(4):203. [doi: 10.3390/a18040203]
98. Park H, Youm C, Cheon SM, Kim B, Choi H, Hwang J, et al. Using machine learning to identify Parkinson's disease severity subtypes with multimodal data. J Neuroeng Rehabil 2025;22(1):126 [FREE Full text] [doi: 10.1186/s12984-025-01648-2] [Medline: 40457336]
99. Goubault E, Martin C, Duval C, Daneault JF, Boissy P, Lebel K. Enhanced detection and segmentation of sit phases in patients with Parkinson's disease using a single smartwatch and random forest algorithms. Sensors (Basel) 2025;25(19):6104 [FREE Full text] [doi: 10.3390/s25196104] [Medline: 41094927]
100. Liu T, Li Z, Ji X, Xu M, Wang L, Chi Y. Friction nanogenerators based on asymmetrically adherent conductive hydrogels for multifunctional sensing and Parkinsonian gait diagnostics. Chem Eng J 2025;524:169026. [doi: 10.1016/j.cej.2025.169026]
101. Nanayakkara T, Herath HMKKMB, Malekroodi HS, Madusanka N, Yi M, Lee BI. Multi-domain CoP feature analysis of functional mobility for Parkinson's disease detection using wearable pressure insoles. Sensors (Basel) 2025;25(18):5859 [FREE Full text] [doi: 10.3390/s25185859] [Medline: 41013102]
102. Ma L, Lin S, Jin J, Wang Z, Wang X, Chen Z, et al. Objective assessment of gait and posture symptoms in Parkinson's disease using wearable sensors and machine learning. Front Aging Neurosci 2025;17:1618764 [FREE Full text] [doi: 10.3389/fnagi.2025.1618764] [Medline: 40860216]
103. Lopes T, Reis Carneiro M, Morgadinho A, Reis Carneiro D, Tavakoli M. ParCuR-a novel AI-enabled gait cueing wearable for patients with Parkinson's disease. Sensors (Basel) 2025;25(22):7077 [FREE Full text] [doi: 10.3390/s25227077] [Medline: 41305286]
104. Channa A, Popescu N, Ciobanu V. Wearable solutions for patients with Parkinson's disease and neurocognitive disorder: a systematic review. Sensors (Basel) 2020;20(9):2713 [FREE Full text] [doi: 10.3390/s20092713] [Medline: 32397516]
105. Mughal H, Javed AR, Rizwan M, Almadhor AS, Kryvinska N. Parkinson's disease management via wearable sensors: a systematic review. IEEE Access 2022;10:35219-35237. [doi: 10.1109/access.2022.3162844]
106. Sigcha L, Borzì L, Amato F, Rechichi I, Ramos-Romero C, Cárdenas A, et al. Deep learning and wearable sensors for the diagnosis and monitoring of Parkinson's disease: a systematic review. Expert Syst Appl 2023;229:120541. [doi: 10.1016/j.eswa.2023.120541]
107. Joo JH, Lieu N, Tang Y, Browne DS, Agusala B, Liao JM. Trends in utilization of remote monitoring in the United States. Health Aff Sch 2025;3(6):qxaf115. [doi: 10.1093/haschl/qxaf115] [Medline: 40575760]
108. Tan SY, Sumner J, Wang Y, Wenjun Yip A. A systematic review of the impacts of remote patient monitoring (RPM) interventions on safety, adherence, quality-of-life and cost-related outcomes. NPJ Digit Med 2024;7(1):192 [FREE Full text] [doi: 10.1038/s41746-024-01182-w] [Medline: 39025937]
109. Hsueh WT, Yu CX, Cheng HC, Chen MY, Wang HM, Fu LM. A comprehensive review of wearable devices for non-invasive biosensing. Trends Anal Chem 2025;193:118425. [doi: 10.1016/j.trac.2025.118425]
110. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature 2023;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]
111. Smuck M, Odonkor CA, Wilt JK, Schmidt N, Swiernik MA. The emerging clinical role of wearables: factors for successful implementation in healthcare. NPJ Digit Med 2021;4(1):45 [FREE Full text] [doi: 10.1038/s41746-021-00418-3] [Medline: 33692479]
112. Xu H, Liu J, Hu H, Zhang Y. Wearable sensor-based human activity recognition method with multi-features extracted from Hilbert-Huang transform. Sensors (Basel) 2016;16(12):2048 [FREE Full text] [doi: 10.3390/s16122048] [Medline: 27918414]

XSL•FO
RenderX

113. Ordóñez FJ, Roggen D. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. Sensors (Basel) 2016;16(1):115 [FREE Full text] [doi: 10.3390/s16010115] [Medline: 26797612]
114. Fiordelli M, Diviani N, Schulz PJ. Mapping mHealth research: a decade of evolution. J Med Internet Res 2013;15(5):e95 [FREE Full text] [doi: 10.2196/jmir.2430] [Medline: 23697600]
115. Evers LJW, Peeters JM, Bloem BR, Meinders MJ. Need for personalized monitoring of Parkinson's disease: the perspectives of patients and specialized healthcare providers. Front Neurol 2023;14:1150634 [FREE Full text] [doi: 10.3389/fneur.2023.1150634] [Medline: 37213910]
116. Toney-Butler TJ, Thayer JM. Nursing Process Internet. Treasure Island (FL): StatPearls Publishing; 2023.
117. Kwon SH, Park JK, Koh YH. A systematic review and meta-analysis on the effect of virtual reality-based rehabilitation for people with Parkinson's disease. J Neuroeng Rehabil 2023;20(1):94 [FREE Full text] [doi: 10.1186/s12984-023-01219-3] [Medline: 37475014]
118. Chuang CS, Chen YW, Zeng BY, Hung CM, Tu YK, Tai YC, et al. Effects of modern technology (exergame and virtual reality) assisted rehabilitation vs conventional rehabilitation in patients with Parkinson's disease: a network meta-analysis of randomised controlled trials. Physiotherapy 2022;117:35-42. [doi: 10.1016/j.physio.2022.07.001] [Medline: 36242929]
119. Yau CE, Ho ECK, Ong NY, Loh CJK, Mai AS, Tan EK. Innovative technology-based interventions in Parkinson's disease: a systematic review and meta-analysis. Ann Clin Transl Neurol 2024;11(10):2548-2562 [FREE Full text] [doi: 10.1002/acn3.52160] [Medline: 39236299]
120. Bowman T, Gervasoni E, Arienti C, Lazzarini SG, Negrini S, Crea S, et al. Wearable devices for biofeedback rehabilitation: a systematic review and meta-analysis to design application rules and estimate the effectiveness on balance and gait outcomes in neurological diseases. Sensors (Basel) 2021;21(10):3444 [FREE Full text] [doi: 10.3390/s21103444] [Medline: 34063355]
121. Giannakopoulou KM, Roussaki I, Demestichas K. Internet of things technologies and machine learning methods for Parkinson's disease diagnosis, monitoring and management: A systematic review. Sensors (Basel) 2022;22(5):1799 [FREE Full text] [doi: 10.3390/s22051799] [Medline: 35270944]
122. Mathias R, McCulloch P, Chalkidou A, Gilbert S. Digital health technologies need regulation and reimbursement that enable flexible interactions and groupings. NPJ Digit Med 2024;7(1):148 [FREE Full text] [doi: 10.1038/s41746-024-01147-z] [Medline: 38890404]

Abbreviations

AI: artificial intelligence
CNKI: China National Knowledge Infrastructure
CNN: convolutional neural network
DL: deep learning
LSTM: long short-term memory
MeSH: Medical Subject Headings
ML: machine learning
PD: Parkinson disease
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
SVM: support vector machine

Edited by S Brini; submitted 10.Oct.2025; peer-reviewed by O Oyetunji, S Jiang, S Mohanadas, S Ajayi; comments to author 01.Dec.2025; accepted 04.Feb.2026; published 26.Feb.2026.

Please cite as:
Li S, Chen S, Yu X, Shang H, Tu T, Quan M
AI-Enabled Wearables for Motor Function Assessment and Rehabilitation in Parkinson Disease: Scoping Review
J Med Internet Res 2026;28:e85596
URL: https://www.jmir.org/2026/1/e85596
doi:10.2196/85596
PMID:

©Shengting Li, Siqi Chen, Xiaosong Yu, Huixiang Shang, Tingting Tu, Mingtao Quan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution,

XSL•FO
RenderX

and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effectiveness of Telerehabilitation Interventions for Self-Management of Tinnitus: Update of a Systematic Review

Sara Demoen[1,2,3,4], PhD; Elise Van Kerchove[1], MSc; Annick Timmermans[1], Prof Dr; Vincent Van Rompaey[2,3], Prof Dr Med; Sarah Michiels[1,3*], Prof Dr; Annick Gilles[2,3,5*], Prof Dr

[1]REVAL Rehabilitation Research Center, Faculty of Rehabilitation Sciences, Hasselt University, Hasselt, Belgium

[2]Department of Translational Neurosciences, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Flanders, Belgium

[3]Department of Otorhinolaryngology and Head and Neck Surgery, Faculty of Medicine and Health Sciences, Antwerp University hospital, Antwerp, Belgium

[4]Rehabilitation Research Group (RERE), Faculty of Health Sciences, Vrije Universiteit Brussel, Brussels, Brussels Capital, Belgium

[5]Department of Education, Humanities and Social Work, Faculty of Human and Social Welfare, University College Ghent, Ghent, Flanders, Belgium

[*]these authors contributed equally

Corresponding Author:
Sarah Michiels, Prof Dr
REVAL Rehabilitation Research Center
Faculty of Rehabilitation Sciences
Hasselt University
Martelarenlaan 42, 3500
Hasselt
Belgium
Phone: 32 011269307
Email: sarah.michiels@uhasselt.be

Abstract

Background: Approximately 14% of the adult population has tinnitus, and current treatments are often costly and time-consuming. Telerehabilitation might reduce treatment costs without compromising effectiveness.

Objective: Telerehabilitation is a quickly evolving research topic. Therefore, this systematic review update aims to give an overview of the research concerning the effectiveness of telerehabilitation interventions for self-management of tinnitus published between 2022 and 2025.

Methods: This systematic review adheres to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020) guidelines. PubMed, ScienceDirect, Scopus, Web of Science, and Cochrane Library were consulted for eligible studies concerning a study intervention of any possible form of self-management or telerehabilitation for adult patients with subjective tinnitus as a primary complaint. The risk of bias (RoB) and certainty of all included studies were assessed respectively by the Cochrane RoB2-tool and GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) framework.

Results: In total, 24 papers were included, of which 6 studied multiple telerehabilitation forms. Internet-based cognitive behavioral therapy with guidance by a psychologist or audiologist was examined in 5 studies (n=619), self-help manuals in 1 study (n=10), technological self-help devices in 3 studies (n=286), smartphone apps in 13 studies (n=23,788), and other internet-based interventions in 5 studies (n=442). These rehabilitation categories were proven to be effective in decreasing tinnitus severity and relieving tinnitus distress as measured by tinnitus questionnaires.

Conclusions: The strength of this review is the gathering of recent studies on the very evolving topic of telerehabilitation for tinnitus. An important limitation of all included studies is that they raised some to great concerns of RoB. As a result, it is necessary to acknowledge that the overall certainty of the evidence ranged from low to moderate certainty. In addition, some crucial confounding parameters, such as the presence of hearing loss, hyperacusis, anxiety, depression, or sleeping problems, were not taken into consideration by all studies. This review gives an indication of the use of different telerehabilitation and self-management interventions for real-world clinical use, stating not only their possibilities but also their limitations. Overall, telerehabilitation was found to be effective in reducing tinnitus severity and distress. It forms a possible tool to improve the self-management capacities of the patient and the accessibility of tinnitus care as a replacement or an addition to in-person care. Nevertheless, barriers such as a lack of time, engagement, motivation, and openness of the patient, causing high dropout, should be taken into

XSL•FO
RenderX

consideration. This review accentuated the shift from internet-based cognitive behavioral therapy to the growing interest in the use of smartphone apps, increasing the accessibility of the treatments even more.

Trial Registration: PROSPERO CRD 42021285450; https://www.crd.york.ac.uk/PROSPERO/view/CRD42021285450

(J Med Internet Res 2026;28:e83529) doi:10.2196/83529

KEYWORDS

tinnitus; telerehabilitation; self-management; mobile phone; smartphone application

Introduction

Background

Tinnitus is defined as the perception of a constant or intermittent sound without the presence of an external auditory source [1,2]. It is a highly prevalent symptom affecting approximately 14% of the adult population [3]. The daily functioning and health-related quality of life of tinnitus patients can directly or indirectly be disrupted by the distress they experience due to the tinnitus sound [4]. Therefore, the most commonly used and evidence-based treatments are tinnitus retraining therapy (TRT) and cognitive behavioral therapy (CBT) [5,6]. The high costs and time consumption of these treatments, combined with the repeated need for care of these patients, resulted in a quest to find a more cost-effective alternative [7-9]. As an answer, telerehabilitation is on the rise to promote self-management in tinnitus patients and reduce their dependency on in-person care. Telerehabilitation, meaning delivery of care from a distance, became a necessity rather than an opportunity during the COVID-19 pandemic [10-12].

The previous systematic review published in 2023 on the topic of telerehabilitation for chronic subjective tinnitus patients indicated that low-contact treatment provided from a distance through several telerehabilitation forms could be used as a substitution for or an addition to in-person clinical care for tinnitus [13]. It can potentially reduce the costs of treatment without sacrificing the level of effectiveness. Telerehabilitation treatment forms are easy to implement in clinical practice and form accessible treatment options [13]. Overall, telerehabilitation in the form of internet-based cognitive behavioral therapy (iCBT) with or without guidance, self-help manuals, self-help devices, smartphone apps, and other internet-based interventions was found to be effective in reducing tinnitus severity and distress [13]. Nevertheless, barriers such as lack of time, engagement, motivation, and openness of the patient need to be considered, as they can cause high dropout rates. The previous review also concluded that future research should further explore which factors are most likely to cause the lack of compliance and how clinicians can counteract these [13].

Due to the advances in technology made every day, telerehabilitation will keep evolving and, therefore, remain a trending topic to follow up on. The application of digital technology in tinnitus therapy is rising stronger than ever. For example, a variety of smartphone apps for people with tinnitus are available, promising relief from the condition. Yet, it is difficult to estimate their potential [14,15]. After the COVID-19 pandemic, the potential benefits of telerehabilitation forms became undeniable, and recently, these new tinnitus treatment forms were even implemented in the guidelines for tinnitus care. Since then, telehealth has been a highly dynamic research field, and several new studies on telerehabilitation for tinnitus have been published in the last couple of years [16-18]. The high number of patients with tinnitus, their repeated need for care, long waiting lists for regular care, and the high costs related to gold-standard treatments are some of the reasons why the need to explore telerehabilitation options is so high. It might form the key to making treatments more cost-effective and more accessible without losing on a level of effectiveness. It can even benefit the effectiveness when it is used as a tool to improve tailoring of the treatment approach to the patient's needs and as a tool to increase adherence and self-management skills of the patient, reducing the need for repeated care.

Objectives

In the past decades, various low-contact and self-management telerehabilitation treatments were developed for patients with tinnitus to search for an alternative or additional tool for the existing treatment options that are time-consuming and costly. Due to the rise in scientific research and the high number of new publications in recent years on this topic, this systematic review aimed to give an updated overview of the research concerning the effectiveness of telerehabilitation interventions for self-management of tinnitus published between 2022 and 2025.

Methods

Study Design

This systematic review is an update of an earlier systematic review published in February 2023 [13]. It adheres to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [19]. The predefined protocol registered in PROSPERO from the earlier review, with registration number CRD42021285450, was repeated identically.

Eligibility Criteria

The eligibility criteria can be found in Textbox 1. These criteria remained the same for the updated review. However, a publication date restriction was added as an exclusion criterion. Papers published before October 18, 2021, were not taken into consideration because they were already included in the previous review. The search ended on February 10, 2025.

XSL•FO
RenderX

Textbox 1. Eligibility criteria.

Inclusion criteria

- Adult patients (>18 y).
- Subjective tinnitus as a primary complaint.
- A study intervention comprising any possible form of self-management or telerehabilitation.
- No publication date restrictions.

Exclusion criteria

- No tinnitus.
- Objective tinnitus.
- Tinnitus as a secondary complaint.
- Full text not available in Dutch, French, German, Greek, or English.

Information Sources

On February 10, 2025, a total of 5 databases were consulted: PubMed, ScienceDirect, Scopus, Web of Science, and Cochrane Library.

Search Strategy

The search queries were composed according to the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) guidelines [20] (checklist provided in Multimedia Appendix 1). The queries comprise free terms and MeSH (Medical Subject Headings) terms for tinnitus (population) and self-management, telerehabilitation, smartphone application (intervention), and their synonyms. All terms and synonyms were controlled for value to the search query. If a term did not result in an additional number of studies found, it was eliminated to keep the search query as concise as possible while maintaining the highest level of efficiency. The search queries were not adapted from the previous review and were rerun identically, because the addition of terms did not result in more relevant eligible studies. The complete search queries can be consulted in Multimedia Appendix 2. All searches were concluded on February 10, 2025. Besides the date restriction that papers had to be from 2022 or later, no additional search filters or restrictions were added to the search strategy. Rayyan was used to eliminate duplicates from the searches.

Selection Process

Double screening of the gathered papers was performed by 2 reviewers (Sara Demoen and EVK) independently and blindly. A first screening on the abstract and title was performed, followed by a screening on the full text. The Rayyan web platform was used to screen all papers efficiently. The final decision on the inclusion of the papers was obtained after a consensus meeting, discussing potential conflicts or papers that were labelled as maybe by at least one of the reviewers.

Data Collection

Both reviewers collected the data independently, and the data extraction tables were composed during a consensus meeting with the reviewers and the supervisors. In the earlier review, 6 subcategories were made: iCBT with guidance (feedback, monitoring, and support by a psychologist or audiologist), iCBT without guidance, self-help devices, self-help manuals, smartphone apps, and other internet-based interventions. For all of these categories, new studies were included during this updated search, except for the subcategory iCBT without guidance. For each subcategory, a data-extraction table was composed, comprising the collected study information according to the applicable items of the subdivisions: PICO (population, intervention, comparison, and outcome).

Data Items

The study and population characteristics that were listed are the author, the year of publication, the study design, the sample size (n), the gender distribution (%), the mean age and the SD (y), the type of tinnitus, the mean duration of tinnitus, and the presence of hearing loss.

Intervention is the first aspect summarized in the table of evidence. This section discusses the following items: (+ if applicable comparison: control intervention, ie, if there is a comparison with a control group [CG], these items are also discussed for the control intervention and not solely the experimental intervention): the used intervention form and specifics, such as guidance, duration, and follow-up. For the next section, Outcome, outcome measures, and dropout were presented, and finally, under the Results section, data results, GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) score, and the conclusions were listed. Each of the 5 subcategories of telerehabilitation included in this review had a table of evidence giving an overview of the information on the intervention, outcome, results, and conclusions of the included study.

Risk of Bias Assessment

The risk of bias (RoB) assessment was performed using the Cochrane RoB2 tool, a tool for assessing RoB in randomized controlled trials (RCTs) [21]. However, it can also be used for other study designs because the domains are, in general, also applicable [21]. The study design of the included studies is diverse, and for this reason, following a separate RoB assessment for each study design would not give a clear overview for comparison [21]. Alternative tools such as ROBINS-I (Risk of Bias in Non-Randomized Studies of

XSL•FO
RenderX

Interventions) were considered, but using a different instrument in the current update would introduce heterogeneity in assessment methods. We therefore applied the Cochrane RoB tool across all designs to ensure methodological consistency with the previous version of this review, and because this tool provides a more conservative, strict assessment of bias for nonrandomized studies rather than an overly lenient one. RoBvis, an RoB visualization tool by Cochrane, was used to summarize the results in a graphic manner [22]. The RoB screening was performed blinded by 2 reviewers (Sara Demoen and EVK), and conflicts were resolved after a consensus meeting.

Data Synthesis

The heterogeneity in reported interventions and outcome measures precluded meta-analysis. A meta-analysis was not performed in the present review update for several methodological reasons. First, this work serves as an update of a previously published systematic review, which also did not pool results due to substantial heterogeneity. Maintaining consistency with the earlier review ensures comparability across iterations. Second, the included studies used a wide range of outcome measures that were operationalized and reported in noncomparable ways, limiting the feasibility of quantitative synthesis. Finally, the statistical reporting across studies varied considerably, with insufficiently detailed or inconsistent data formats that prevented reliable extraction of effect sizes. In combination, these factors precluded a valid and meaningful meta-analysis, and therefore, it was decided to summarize the collected data in a systematic review. This was not different in the current update.

Certainty Assessment

The GRADE framework is a tool to assess the level of certainty by examining RoB, imprecision, inconsistency, indirectness, and publication bias of the included studies [23]. The GRADE assessment was completed according to the GRADE handbook guidelines [23].

Results

Study Selection

The initial database searches on February 10, 2025, resulted in 1192 potentially relevant results (Figure 1 [19]). After removal of the duplicates and of the papers published before October 2021, a total of 496 papers remained for primary screening on title and abstract. During the primary screening, 441 papers were excluded; the reasons for exclusion are specified in Figure 1. In total, 55 papers were sought for retrieval for the screening of full text. Subsequently, the screening of the full text was completed for 46 papers because the full text of 9 papers could not be retrieved. The reason for exclusion is again specified in Figure 1. A total of 24 papers were found to be eligible [24-47]. These eligible papers were divided into the same subgroups used in the previous systematic review [13]. The subgroups were based on the form of telerehabilitation as an intervention: iCBT with guidance, iCBT without guidance, self-help devices, self-help manuals, smartphone apps, and other internet-based interventions.

XSL•FO
RenderX

Figure 1. Flow diagram for study inclusion (PRISMA 2020). Studies can belong to multiple of the listed study objective categories. iCBT: internet-based cognitive behavioral therapy; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses. S

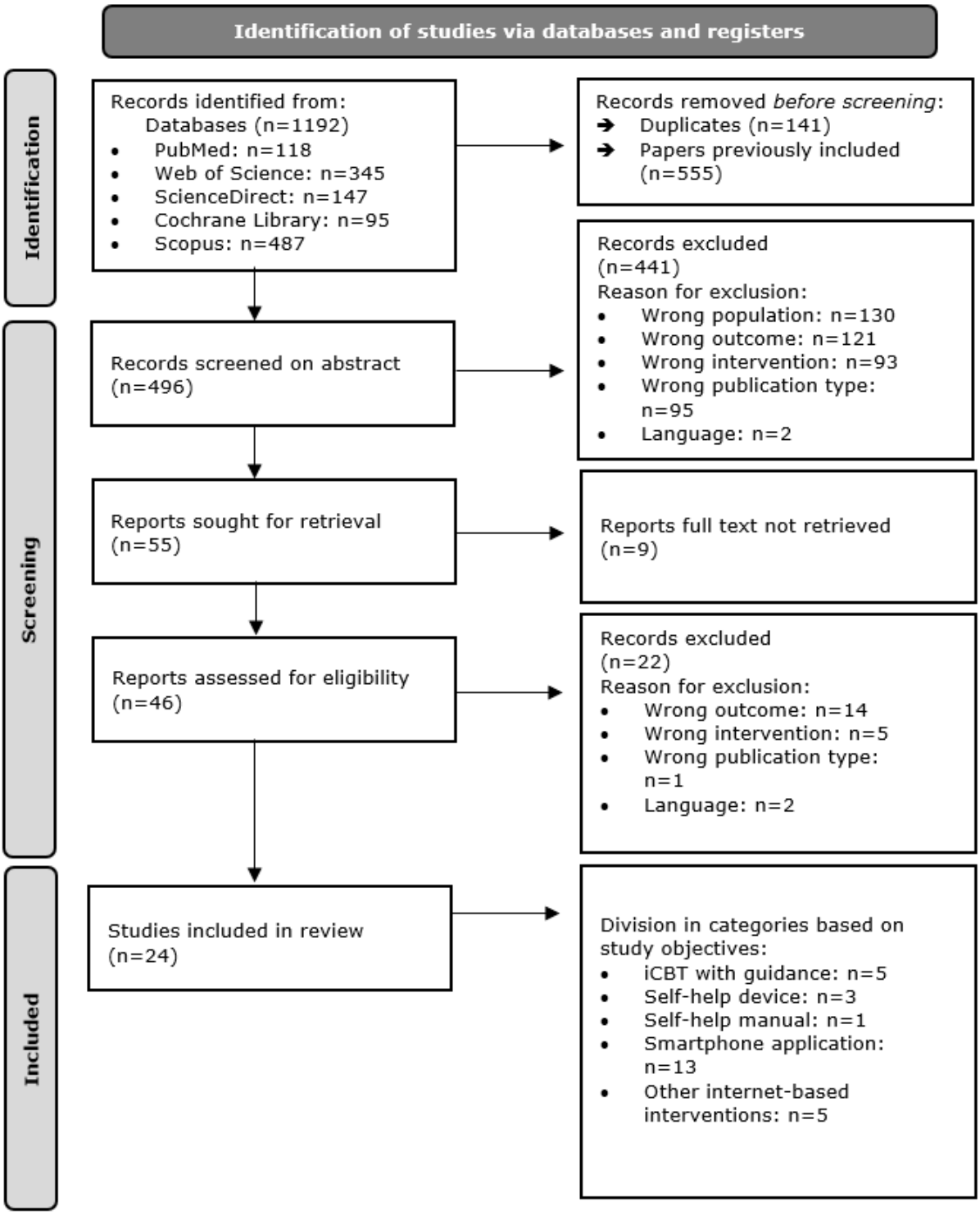

About RoB

Controlled Trials

In total, 11 studies were controlled trials [24,25,30-32,35,37,41,44,45,47], of which 6 studies were RCTs [25,30,31,35,41,47]. Five RCTs had a low RoB arising from the randomization process, except for Altissimi et al [25], who showed a higher RoB due to the lack of information given concerning the randomization method (D1) [25,30,31,35,41,47]. All controlled trials showed some reasons for concern of RoB due to deviations from the intended interventions (D2), except for Goshtasbi et al [30]. RoB, due to missing outcome data (D3), was scored as "high concerns" and "some concerns" for studies with respectively moderate to high dropout rates or no clearly specified plan on how to handle these missing data. All controlled studies were rated "low concerns for RoB" in the

XSL•FO
RenderX

measurement of outcome (D4), except Ravi et al [37], a study with a sample size of only 2 patients and only 1 questionnaire as an outcome measure. Most studies had a full, clear, correct, and specified plan for statistical analysis, meaning the concern for RoB in the selection of the reported result (D5) was low. However, Alonso-Valerdi et al [24] and Ravi et al [37] had a nondetailed protocol, missing information on the exact premised data analysis plan. Overall, studies were rated "some concerns" to "high concerns" of RoB. See Figure 2 [24,25,30-32,35,37,41,44,45,47] for the RoB assessment results of the controlled trials. Additionally, a summarization of RoB results can be found in Multimedia Appendix 3.

Figure 2. Risk of bias in the controlled trials scoring per domain [25,26,31,32,33,36,38,42,45,46,48].

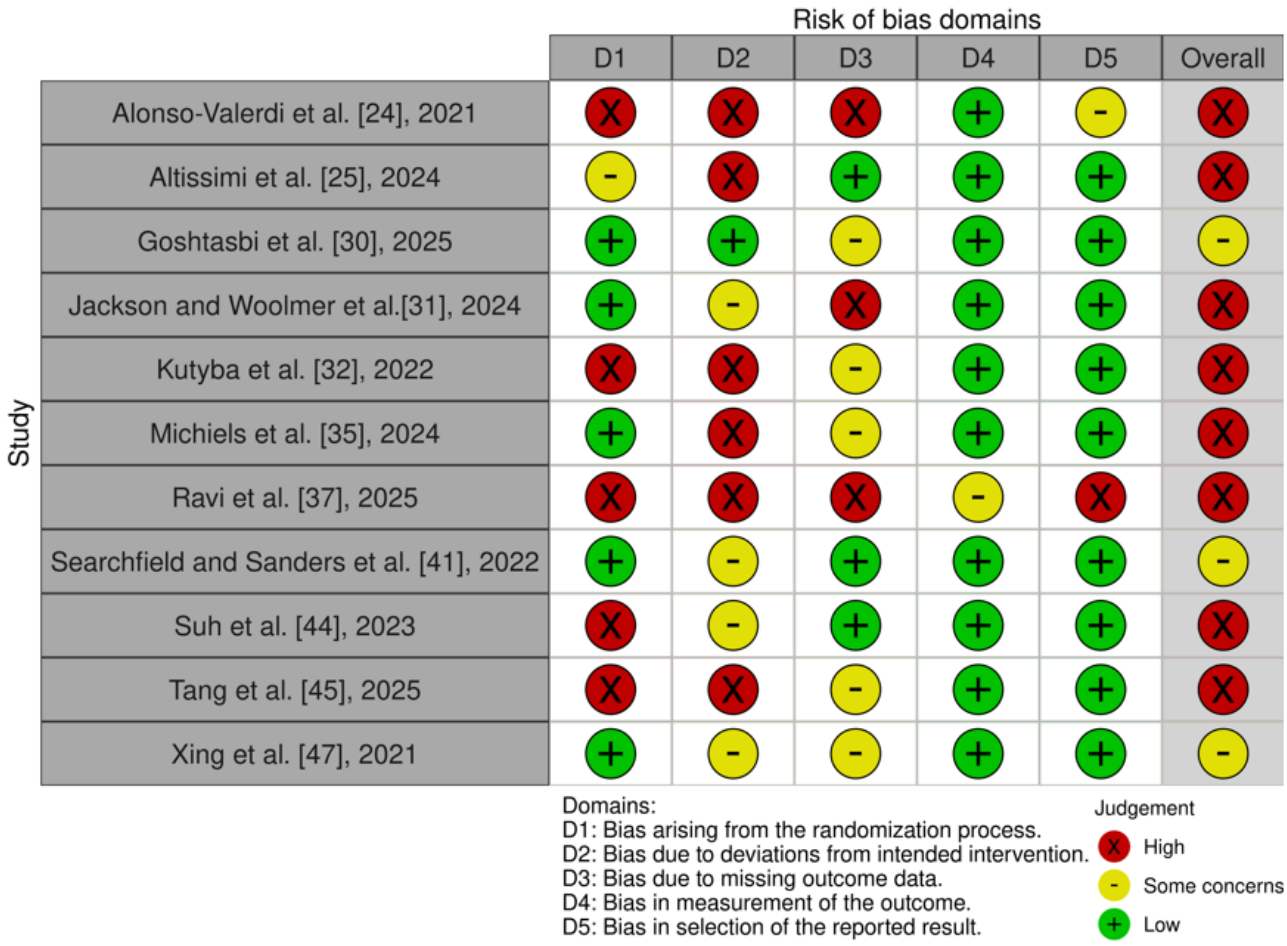

Noncontrolled Trials

A total of 13 studies were noncontrolled trials [26-29,33,34,38-40,42,43,46]. They all show some concerns about RoB due to a lack of randomization, blinding, and CGs (D1 and D2). RoB due to missing outcome data (D3) was again scored high, and there were some concerns for studies with moderate to high dropout rates. The majority of the studies specified their statistical analysis beforehand, resulting in low RoB in the selection of the reported results (D5). See Figure 3 [26-29,32,34,36,38-40,42,43,46] for the RoB assessment results of the noncontrolled trials.

XSL•FO
RenderX

Figure 3. Risk of Bias in the noncontrolled trials per domain [27,28,29,30,33,35,37,39,40,41,43,44,47].

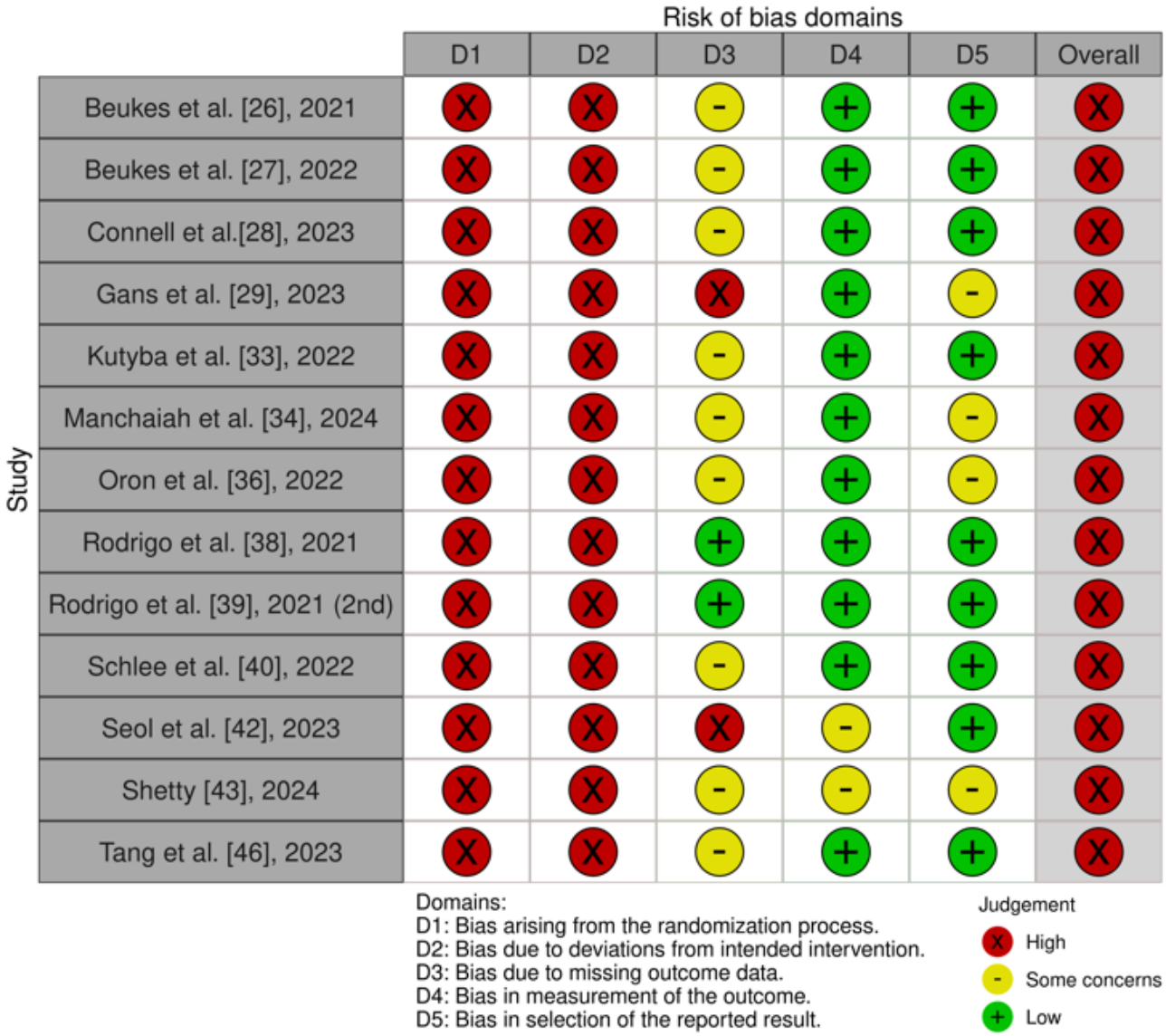

Certainty

In general, the included studies had a low to moderate certainty level according to the GRADE framework [23].

iCBT with Guidance

Study and Patient Characteristics

A total of 3 studies looked into the use of an iCBT program with guidance provided by an audiologist in the form of feedback, monitoring, and support through an online messaging system within the platform [26,27,34]. All 3 studies were cohort studies with an adult study population with chronic tinnitus. Further, 2 studies additionally specified that tinnitus symptoms had to be present for more than 3 months, and it had to be at least a mild severity of tinnitus. Beukes et al [26] stated that patients were only eligible if they scored at least 25 of 100 on the Tinnitus Functional Index (TFI). The mean tinnitus duration ranged from 11.64 to 13.91 years between the studies. Beukes et al [26] was the only study controlling for the presence of hearing loss, while Beukes et al [27] was the only one specifying the tinnitus characteristics, namely the tinnitus location. The total sample is 391 patients receiving iCBT with guidance. For further specifications, see Table S1 in Multimedia Appendix 4. Two studies examined an iCBT program with minimal guidance [38,39]. However, it needs to be noted that both these published studies were from the same author and concerned the same sample data. Both studies used data from a cohort of 228 patients with long-standing chronic tinnitus with a duration of 17.68 (SD 19.42). For further specifications, see Table S1 in Multimedia Appendix 4.

Results of Individual Studies

All 3 studies used a similar iCBT program consisting of an interactive e-learning program with 22 modules based on the original CBT self-help program developed by Andersson and Kaldo [48] (2004) [26,27,34]. The interventions were guided by an audiologist who monitored progress, provided feedback on worksheets that were completed, outlined the content of new modules, and answered questions. The guidance by the audiologist was offered through a 2-way encrypted online

XSL•FO
RenderX

messaging system. The primary tinnitus outcome of all studies was tinnitus severity or distress measured by the TFI. Various secondary outcome measures, using self-report questionnaires, were appended. All studies used data from previous trials. See Table 1 for further details. All studies concluded that the iCBT with guidance provided by an audiologist trained in iCBT principles was effective in the treatment of chronic tinnitus. Beukes et al [26] demonstrated that iCBT has long-term positive effects, and Manchaiah et al [34] added that it can help individuals with tinnitus to think more positively by changing unhelpful patterns. Both studies from Rodrigo et al [38,39] used the same data that was gathered from multiple trials and looked into the use of an iCBT program with minimal guidance through an encrypted messaging system solely for support. This iCBT program was again based on the original CBT self-help program developed by Andersson and Kaldo [48] (2004). However, it was adapted to a set of only 21 modules, of which 5 were optional. These 2 studies also found a decrease in TFI after completing the iCBT program and additionally investigated predictors of iCBT outcomes. For more information, see Table 1.

XSL•FO
RenderX

Table 1. Table of evidence for the studies concerning iCBT[a] with guidance.

Study	Intervention			Outcome			Results				
Author and year	Intervention	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[b]
Beukes et al [26], 2022	Interactive e-learning iCBT consisting of 22 modules and guidance by an audiologist	8 weeks	FU1[c]: postintervention assessment; FU2[d]: 2-months postintervention; FU3[e]:1-year postintervention	TFI[f]	THS[g].	Effect sizes, LMM[h], and the RCI[i].	Significant BL[j]-FU3 TFI mean improvement of 24.3 (SD 22.5; T0: 54.23, SD 20.7; T3: 30.15, SD 24.51). There was a clinically significant change for 56 (42%) of the participants using the reliable change criterion of 26.98 at T3.	Statistically significant BL-FU3 difference with a large effect size (Cohen *d*=1.06, CI 0.80 to 1.32).	68/200 (34%) did not complete FU1, 83/200 (43%) FU2, and 68/200 (34%) FU3.	The study provides evidence of long-term efficacy for an audiologist-guided iCBT tinnitus intervention.	⊕⊕⊕◯
Beukes et al [27], 2022	Interactive e-learning iCBT consisting of 22 modules and guidance by an audiologist	8 weeks	FU[k]=2 months	TFI	GAD-7[l], PHQ-9[m], ISI[n], EQ-5D-5L, THS, TCQ[o], tinnitus and hearing survey.	Pretest and posttest effect size for all primary and secondary outcomes; 2-sided *P* values using α=.05.	A significant, large effect size TFI change postintervention (mean 1.60, 0.91-2.23), which was maintained at FU.	The RCI indicated a prepost score difference of 19.51 on the TFI would be a clinically significant change.	4/27 did not complete FU1 and 9/27 did not complete FU2.	This pilot study has indicated the feasibility of iCBT for tinnitus in the United States.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention			Outcome			Results				
Author and year	Intervention	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[b]
Manchaiah et al [34], 2024	iCBT with guidance by an audiologist	8 weeks	FU: posttreatment	TFI	GAD-7, PHQ-9, ISI, EQ-5D-5L.	Mixed method approach, paired sample *t* test as the data met the assumption of normality using the Shapiro-Wilk test (*P*>.05).	More participants reported positive experiences following the intervention in the overall sample (preintervention mean 0.78, postintervention mean 1.63, *P*<.001).	53/164 (32.3%) provided at least 1 positive experience in both pre and postintervention (group 1), 19/164 (9.1%) provided positive experiences only during preintervention, 49/164 (29.9%) provided positive experiences only during postintervention, and 28.7 participants did not provide any positive experience in both pre- and postintervention.	NR[p].	iCBT can help individuals with tinnitus to think more positively by changing unhelpful thought patterns.	⊕⊕⊕◯
Rodrigo et al [38], 2022	Interactive e-learning iCBT consisting of 21 modules and minimal guidance by an audiologist	8 weeks (2-3 modules a week)	FU1=postintervention assessment; FU2=2-months postintervention	TFI	ISI, PHQ[q], HHIA-S[r], HQ[s], CFQ[t], SWLS[u].	Paired *t* test.	The mean BL tinnitus severity and tinnitus severity following the iCBT intervention were 57.93 (SD 19.17) and 34.22 (SD 22.76), respectively, indicating a significant decrease (*P*<.001).	65% had a successful iCBT outcome, meaning they had a 13-point or higher reduction in TFI after the intervention.	NR.	Some predictors can be noted for a better outcome of iCBT, such as, for example, level of education, BL tinnitus, workload, and disability allowance.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention			Outcome			Results				
Author and year	Intervention	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[b]
Rodrigo et al [39], 2021	iCBT	8 weeks (2-3 modules a week)	FU: postintervention	TFI	ISI, PHQ, HHIA-S, HQ, CFQ, SWLS.	Paired sample *t* tests.	Undertaking iCBT significantly reduced tinnitus severity scores (t_{227}=16.37; P<.001) from a mean BL severity score of 57.93 (SD 19.17) to a mean post-iCBT severity score of 34.22 (SD 22.78), as measured by the TFI.	A clinically significant 13-point change in TFI scores was achieved by 150 of the 228 participants after the intervention.	NR.	Decision tree models, especially the CART[v] and gradient boosting models, appeared to be promising in predicting iCBT outcomes.	⊕⊕⊕◯

[a]iCBT: internet-based cognitive behavioral therapy.

[b]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.

[c]FU1: first follow-up moment.

[d]FU2: second follow-up moment.

[e]FU3: third follow-up moment.

[f]TFI: tinnitus functional index.

[g]THS: tinnitus handicap scale.

[h]LMM: linear mixed effects models.

[i]RCI: Reliable Change Index.

[j]BL: baseline.

[k]FU: follow-up moment.

[l]GAD-7: Generalized Anxiety Disorder-7.

[m]PHQ-9: Patient Health Questionnaire-9.

[n]ISI: Insomnia Severity Index.

[o]TCQ: Tinnitus Cognitions Questionnaire.

[p]NR: not reported.

[q]PHQ: Physical Health Questionnaire.

[r]HHIA-S: Hearing Handicap Inventory for Adults-Short Form.

[s]HQ: Hyperacusis Questionnaire.

[t]CFQ: Cognitive Failures Questionnaire.

[u]SWLS: Satisfaction With Life Scale.

[v]CART: Communication Access Realtime Translation.

XSL•FO
RenderX

Self-Help Devices

Study and Patient Characteristics

Three controlled trials, of which 2 RCTs, studied the use of self-help devices for the treatment of chronic tinnitus [24,25,41]. The studies had a total sample size of 286 patients. The specifications of all 3 studies can be consulted in Table S2 in Multimedia Appendix 4.

Results of Individual Studies

All 3 studies looked into the use of a self-help device delivering a form of sound therapy. For Altissimi et al [25], this was the use of a sound generator for TRT treatment, and this was the CG. The intervention group received TRT treatment using a smartphone app. Searchfield and Sanders [41] also used a smartphone app in combination with the self-help devices for sound therapy, and the CG did as well use an app, but without a self-help device. All studies concluded that the use of a self-help device was effective in reducing tinnitus distress according to the TFI or Tinnitus Handicap Inventory [24,25,41]. Further details can be found in Table 2.

XSL•FO
RenderX

Table 2. Table of evidence for the studies concerning self-help devices.

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Alonso-Valerdi et al [24], 2021	(1) TRT[b]; (2) TEAE[c]; (3) ADT[d]; (4) BBT[e]; (5) music	No treatment.	2 months (60 min daily)	FU1[f]: after 1 week; FU2[g]: after 5 weeks; FU3[h]: after 8 weeks	THI[i]	HADS[j], audiometry, tinnitus characteristics analysis, and BL[k] neurophysiological analysis	Identification of ± or 0 effects *t* test	All negative differences between BL and FU[l] were statistically significant (*P*=.009)	29.4% of the TRT group reported diminution of their tinnitus perception, followed by those in TEAE (18.8%), BBT or ADT (14.3%), and music (7.7%) groups.	Dropout: music 6/16, BBT[m] 4/18, TRT 3/18, TEAE 5/18, ADT 5/18, and control 3/14	TRT most effective sound-based therapy to reduce tinnitus perception for individuals without anxiety. BBT, ADT, and TEAE produced very similar effects.	⊕⊕⊕◯
Altissimi et al [25], 2024	TRT with mobile app	TRT with a traditional sound generator.	6 months	FU1: after 3 months of treatment; FU2: after 6 months of treatment	THI	TRT interview: with questions on tinnitus, hearing loss, hyperacusis, classifying gravity, the annoyance, and the effect on the QoL[n] of tinnitus, using a VAS[o].	Descriptive data by counts of a mixed-effect model Tukey test	A significant difference in the THI scores was found in category 1 patients between those treated with a sound generator vs app (–14.529, *P*<.001)	Improvement was found in category 0 patients, both treated with the app and sound generator; conversely, in category 1, significant improvement was seen only for the group using a sound generator.	No dropout	TRT can be delivered to patients with mild tinnitus (class 0) through mobile apps with results comparable to traditional sound generators.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Searchfield and Sanders [41], 2022	The digital therapeutic consisted of an app for iPhone or Android smartphone, Bluetooth bone conduction headphones, a neck pillow speaker, and a cloud-based clinician dashboard to enable messaging and app personalization	The control app was a popular self-help passive sound therapy app called WN[p] Lite.	12 weeks	FU1: mid-trial=6 weeks; FU2: postintervention	TFI[q]	COSIT[r], SUS[s], MAUQ[t]	ANOVA with 2 groups and 4 repeated measures, intent-to-treat and completer (per protocol) analyses	Mean changes TFI for USL[u] at 6 (16.36, SD 17.96 points) and 12 weeks (17.83, SD 19.87 points) were clinically meaningful (>13 reduction), the mean changes in WN group were not (6 wk 10.77, SD 18.53 points; 12 wk 10.12, SD 21.36 points)	A statistically higher proportion of USL participants achieved meaningful TFI change at 6 weeks (55%) and 12 weeks (65%) than the WN group at 6 weeks (33%) and 12 weeks (43%).	4/35 USL and 6/36 WN	The USL group demonstrated a higher proportion of responders than the WN group. The usability of the USL therapeutic was similar to the established WN app. The digital polytherapeutic demonstrated significant benefit for tinnitus reduction, supporting further development.	⊕⊕⊕◯

[a]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.

[b]TRT: tinnitus retraining therapy.

[c]TEAE: therapy for an enriched acoustic environment.

[d]ADT: auditory discrimination therapy.

[e]BBT: broad band therapy.

[f]FU1: first follow-up moment.

[g]FU2: second follow-up moment.

[h]FU3: third follow-up moment.

[i]THI: Tinnitus Handicap Inventory.

[j]HADS: Hospital Anxiety and Depression Scale.

[k]BL: baseline.

[l]FU: follow-up moment.

[m]BBT: Broad band therapy.

[n]QoL: quality of life.

[o]VAS: visual analog scale.

[p]WN: white noise.

[q]TFI: Tinnitus Functional Index.

XSL•FO
RenderX

[r]COSIT: Client-Oriented Scale of Improvement in Tinnitus.

[s]SUS: System Usability Scale.

[t]MAUQ: mHealth App Usability Questionnaire.

[u]USL: UpSilent app.

Self-Help Manual

Study and Patient Characteristics

Only 1 cohort with a sample size of 10 adult patients with tinnitus and with bilateral mild sloping sensorineural hearing loss who wore a hearing aid used a self-help manual as an intervention [43]. Further details can be found in Table S3 in Multimedia Appendix 4.

Results of Individual Studies

Shetty [43] examined the use of a self-management manual for tinnitus with content such as (1) hearing aids; (2) listening to songs or soothing sounds; (3) tips for food intake; (4) exercise, yoga, and meditation; (5) locking negative thinking; (6) distraction techniques; (7) progressive muscle relaxation; (8) slow breathing; and (9) attention control and imaginary techniques. The self-help manual was prepared using the health literacy recommendations by adapting the United States Department of Health and Human Services. It was concluded that the self-management manual effectively reduces tinnitus perception and associated problems [43]. See Table 3 for further information.

Table 3. Table of evidence for the studies concerning the self-help manual.

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Shetty et al [43], 2024	Self-management manual for tinnitus	N/A[b]	30 days	FU[c]: postintervention	THI[d], TFI[e]	N/A	Independent samples *t* test.	The results revealed a significant reduction in THI after 30 days of manual use (t_9=1.99, *P*=.07). A similar result was found in TFI, in which functional impairment from tinnitus was significantly reduced between visits one and two (t_9=6.13, *P*=.001)	The means of THI and TFI scores in visit 1 were 31 (SD 5.16) and 58 (SD 5.92), respectively. Furthermore, at visit 2, the THI and TFI scores were 29 (SD 4.59) and 45 (SD 7.05), respectively.	No dropout	Patients who have used the manual reported relaxing and coping with tinnitus. It is concluded that the self-management manual effectively reduces tinnitus perception and associated problems.	⊕⊕⊕◯

[a]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.

[b]N/A: not applicable.

[c]FU: follow-up moment.

[d]THI: Tinnitus Handicap Inventory.

[e]TFI: Tinnitus Functional Index.

XSL•FO
RenderX

Smartphone Apps

Study and Patient Characteristics

A total of 13 articles, with a combined sample size of 23,637 participants, discussed the use of a smartphone app as part of the treatment of tinnitus patients [25,30,32,33,35-37,40-42,44-46]. Clarification on the study and patient characteristics of the studies examining the use of a smartphone app is noted in Table S4 in Multimedia Appendix 4.

Results of Individual Studies

The intervention used a smartphone app in 13 studies [25,30,32,33,35-37,40-42,44-46]. First, in 8 studies, this smartphone app was used to deliver a form of sound therapy to the patients. These studies showed that sound therapy through a smartphone app can be beneficial in reducing tinnitus distress. In 2 of these studies, Goshtasbi et al [30] and Searchfield and Sanders [41], the smartphone app combined both sound therapy and CBT or psychoeducation, respectively. Second, 5 studies look into TRT, CBT, or psychoeducational treatment delivered through a smartphone app. TRT with a mobile app appeared to be as effective as traditional TRT with or without a sound generator, and daily training using a CBT-based app targeting maladaptive beliefs may decrease tinnitus intrusiveness and handicap. Psycho-education through a smartphone app gave promising results in reducing tinnitus distress, but did not result in changes in tinnitus loudness. Finally, 1 study investigated the use of a smartphone app in the treatment of a specific subpopulation of tinnitus patients, namely patients with somatic tinnitus [35]. This pilot study provides evidence that an app-based physiotherapy intervention with a cervical spine exercise program could be an effective option for treating somatic tinnitus [35]. See Table 4 for an overview of the detailed information for all studies concerning treatments using smartphone apps.

XSL•FO
RenderX

Table 4. Table of evidence for the studies concerning smartphone apps.

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Altissimi et al [25], 2024	TRT[b] with mobile app	TRT with a traditional sound generator	6 months	FU1[c]: after 3 months of treatment; FU2[d]: after 6 months of treatment	THI[e]	TRT interview: with questions on tinnitus, hearing loss, hyperacusis, classifying gravity, the annoyance, and the effect on the QoL[f] of tinnitus, using a VAS[g].	Descriptive data by counts, a mixed-effect model Tukey test.	A significant difference in the THI scores was found in category 1 patients between those treated with a sound generator vs app (–14.529, *P*<.001), while no significance was found in category 0 patients (–1.186, *P*=.78).	Improvement was found in category 0 patients, both treated with the app and sound generator; conversely, in category 1, significant improvement was seen only for the group using a sound generator.	No dropout.	TRT can be delivered to patients with mild tinnitus (class 0) through mobile apps with results comparable to traditional sound generators.	⊕⊕⊕◯
Goshtasbi et al [30], 2025	Smartphone-based CBT[h] and customized sound therapy.	Waitlisted	8 weeks (1 weekly CBT module and 2 h daily sound therapy)	Posttreatment	TFI[i]	PHQ-9[j], GAD-7[k], PSQI[l], PSS[m].	Independent sample *t* test and chi-square tests of independence.	Subjects in the IG[n] showed greater improvements on their TFI (16.7, SD 14.9 versus 1.9, SD 10.8 (*P*<.001) and on 7 of 8 subscales (*P*<.001-.002).	18 (38.3%) IG patients showed a TFI improvement of 20 or more points posttreatment.	15 patients dropped out between week 2 and week 8 and were not included in the IG.	Smartphone-based CBT, providing CBT and customized sound therapy, was effective in reducing symptom severity and improving anxiety, sleep, and mood.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Kutyba et al [32], 2022	Free ReSound Tinnitus Relief app	N/A[o]	6 months (≥30 min/day)	FU1: after 3 months of treatment FU2: after 6 months of treatment	THI, TFI	User survey.	Repeated measures ANOVA with Bonferroni correction.	THI: $F_{2,102}=624$; $P<.001$; $e2=0.550$, TFI: $F_{2,102}=289$; $P<.001$; $e2=0.362$.	In the study, the general severity decreased after the first 3 months (THI: –6.5 in mean score in comparison to BL[p]; TFI: –6.3 in mean score) and again in the following 3 months of using the application (THI: –19.4; TFI: –14.3).	NR.	The sound therapy through the use of a smartphone app appeared to be beneficial in reducing subjective tinnitus distress.	⊕⊕⊕◯
Kutyba et al [33], 2022	Free ReSound Tinnitus Relief app	Participants not willing to participate in the intervention group	6 months (≥30 min/day)	FU1: after 3 months of treatment FU2: after 6 months of treatment	THI	VAS, user survey.	Independent (unpaired) *t* test or a chi-square test, and a mixed-design ANOVA with Bonferroni adjustment for multiple comparisons.	After 3 months of using the app, the THI global score significantly decreased ($P<.001$) in the study group, decreasing again at 6 months ($P<.001$).	A clinically important change (>20) in the THI was reported by 39% (17/44) of the study group and in none of the controls.	Dropout rate: 53% (50/94) in the study group and 55% (29/53) in the control group.	Use of an app that generates background sounds appears to be an effective way of reducing tinnitus severity.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Michiels et al [35], 2024	App-based cervical spine exercise program	Waiting list	9 weeks	FU: post-treatment	THI miniTQ[q]	N/A.	Independent sample *t* test and chi-square tests of independence, and mixed model ANOVA.	Reduction in THI and miniTQ for the IG, not for CG[r]. Strong significant decrease in THI (from mean 51.33, SD 23.77 to mean 38.96, SD 21.66) and miniTQ (from mean 13.27, SD 5.81 to mean 9.14, SD 5.27).	Effect sizes for THI were d=1.71 (95% CI 0.85 to 2.56) and for miniTQ d=1.02 (95% CI 0.24 to 1.80).	CG: 3 dropouts; IG: 4 dropouts.	This pilot study provides evidence that an app-based physiotherapy intervention could be an effective option for treating somatic tinnitus.	⊕⊕⊕◯
Oron et al [36], 2022	Brief cognitive behavioral training for tinnitus relief using a mobile GGTI[s] app	N/A	16 days (3 min/day)	FU1: at level 24 half way through the app; FU2: level 48 final level of the app	H-THI[t]	VAS, user survey.	Repeated measures ANOVA of completers.	A significant linear trend was present, $F_{1,13}$=14.5, P=.002, d=2.12. Oron et al [36] indicated 5 was a steady decrease in the H-THI score of 28.57, 22.14, and 22.0 for BL, FU1, and FU2, respectively.	50% of completers have shown reliable change (indicated by their RCI[u]).	12/26 (46%).	Several minutes a day of training using a CBT-based app targeting maladaptive beliefs may decrease tinnitus intrusiveness and handicap.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Ravi et al [37], 2025	Mobile-based TRT	Conventional face-to-face TRT	3 months (2×/week)	FU1: after 1 month of treatment; FU2: after 2 months of treatment; FU3: after 3 months of treatment	THI	N/A.	N/A.	Patient A: BL=40, FU1=40; FU2=36; FU3=24. Patient B: BL=36, FU1=24; FU2=16; FU3=14.	N/A.	N/A.	Mobile-based TRT might be a good option for patients who cannot attend in-person sessions.	⊕⊕◯◯
Schlee et al [40], 2022	Educational smartphone app	N/A	4 months	FU[v]: postintervention	THI, TS[w] NRS[x]	N/A.	Wilcoxon Signed Rank Test.	Statistical significance between the pre- and postintervention phase (W=345, *P*=.02).	Improvements on the tinnitus severity NRS reached an effect size of 0.408, while the improvements on the THI[y] were much smaller with an effect size of 0.168.	26 of 52 (50%) participants dropped out.	Educational training on a smartphone app without the guidance of a medical person can result in small to medium improvements in tinnitus distress measures, but not for tinnitus loudness.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Searchfield et al [41], 2022	The digital therapeutic consisted of an app for iPhone or Android smartphones, Bluetooth bone conduction headphones, a neck pillow speaker, and a cloud-based clinician dashboard to enable messaging and app personalization	The control app was a popular self-help passive sound therapy app called WN[z] Lite.	12 weeks	FU1: mid-trial=6 weeks FU2: postintervention	TFI	COSIT[aa], SUS[ab], MAUQ[ac]	ANOVA with 2 groups and 4 repeated measures intent-to-treat and completer (per protocol) analyses	Mean changes in TFI for the USL[ad] group at 6 (16.36, SD 17.96 points) and 12 weeks (17.83, SD 19.87 points) were clinically meaningful (>13 points reduction), the mean change in WN scores were not clinically meaningful (6 weeks 10.77, SD 18.53 points; 12 weeks 10.12, SD 21.36 points).	A statistically higher proportion of USL participants achieved meaningful TFI change at 6 (55%) weeks and 12 (65%) weeks than the WN group at 6 (33%) weeks and 12 (43%) weeks.	4/35 USL and 6/36 WN.	The USL group demonstrated a higher proportion of responders than the WN group. The usability of the USL therapeutic was similar to the established WN app. The digital polytherapeutic demonstrated significant benefit for tinnitus reduction supporting further development.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Seol et al [42], 2023	Sound therapy smartphone app	N/A	6 months	FU1: mid-trial=3 months; FU2: postinterevention	THI	BDI[ae], SSL[af], VAS stress assessment, hearing tests.	1-way RM[ag] ANOVA as it passed the normality test (Shapiro-Wilk).	The THI scores, on the other hand, significantly decreased over the 6-month period (P=.02); when the Bonferroni correction method was applied for all P values, statistical significance was observed for THI scores between visits 1 and 3 (P=.006).	After 6 months of listening to the sound stimuli, participants' perceived tinnitus handicap severity significantly decreased.	NR[ah].	The findings of this study demonstrate the potential benefit of the tinnitus application for tinnitus improvement.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Suh et al [44], 2023	Smart device-based TRT with smart pad app	Conventional TRT	3 months (3 sessions 30-60 min) with 1 month interval	FU1=postintervention assessment; FU2=2-months postintervention	THI	VAS, STAI[ai], BDI, PSQI, pure tone audiometry.	*t* tests and chi-square tests.	The greatest THI improvement at 3 months was –23.3 (95% CI –33.41 to –13.4), and at 2 months –16.8 (95% CI –30.8 to –2.8), and the reductions were significant, *P*<.001.	There was no difference between IG and CG (*P*<.001).	NS[aj].	TRT based on smart devices can be an effective alternative to conventional in-person TRT.	⊕⊕⊕◯
Tang et al [46], 2022	FTRS[ak] mobile app	Cohort	2 h/d	FU every 3 months	THI	HADS[al], AIS[am], VAS, FTQ[an], TCS[ao].	Descriptive statistics.	N/A.	N/A.	Only 22.9% of patients adhered to the application (29,864 patients were excluded because of incomplete records of personal and tinnitus-related information).	FTRS greatly reduced medical costs in comparison to traditional in-person care and enabled patients with tinnitus to arrange their own treatment times.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome measure	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Tang et al [45], 2025	DCRS[ap] through a smartphone app	Unmodified music	3 months (2 h/d)	FU1: 1 month of treatment; FU2: 2 months of treatment; FU3: 3 months of treatment	THI	HADS, AIS, FTQ, TCS.	Linear mixed models and a multiclass logistic model with a stepwise function.	DCRS-treated patients after 3 months of treatment had significant tinnitus relief compared to those in the CG group (P<.001).	At 3 months, 92.5% of patients undergoing DCRS reported tinnitus relief or disappearance.	NS.	DCRS showed to be a promising noninvasive therapy for chronic tinnitus that can be delivered through a mobile app.	⊕⊕⊕◯

[a]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.
[b]TRT: tinnitus retraining therapy.
[c]FU1: first follow-up moment.
[d]FU2: second follow-up moment.
[e]THI: Tinnitus Handicap Inventory.
[f]QoL: quality of life.
[g]VAS: visual analog scale.
[h]CBT: cognitive behavioral therapy.
[i]TFI: Tinnitus Functional Index.
[j]PHQ-9: Patient Health Questionnaire-9.
[k]GAD-7: Generalized Anxiety Disorder-7.
[l]PSQI: Pittsburgh Sleep Quality Index.
[m]PSS: Perceived Stress Scale.
[n]IG: intervention group.
[o]N/A: not applicable.
[p]BL: baseline.
[q]miniTQ: Mini Tinnitus Questionnaire.
[r]CG: control group.
[s]GGTI: GG Tinnitus.
[t]H-THI: Hearing-Tinnitus Handicap Inventory.
[u]RCI: Reliable Change Index.
[v]FU: follow-up moment.
[w]TS: Tinnitus Scale.
[x]NRS: Numeric Rating Scale.
[y]THI: Tinnitus Handicap Inventory.
[z]WN: white noise.
[aa]COSIT: Client-Oriented Scale of Improvement in Tinnitus.
[ab]SUS: System Usability Scale.
[ac]MAUQ: mHealth App Usability Questionnaire.
[ad]USL: UpSilent app.
[ae]BDI: Beck Depression Index.
[af]SSL: Subjective Stress Level.
[ag]RM: repeated measure.
[ah]NR: not reported.

XSL•FO
RenderX

[ai]STAI: State Trait Anxiety Inventory.
[aj]NS: not specified.
[ak]FTRS: Fudan Tinnitus Relieving System.
[al]HADS: Hospital Anxiety and Depression Scale.
[am]AIS: Acceptance of Illness Scale.
[an]FTQ: Fear of Tinnitus Questionnaire.
[ao]TCS: Tinnitus Catastrophizing Scale.
[ap]DCRS: digital customized relieving sound.

Other Internet-Based Interventions

Study and Patient Characteristics

A total of 5 studies looked into the use of other internet-based interventions for chronic tinnitus, ranging from sound therapy, mindfulness, meditation, and intensive auditory cognitive training [24,28,29,31,47]. See Table S5 in Multimedia Appendix 4 for more information.

Results of Individual Studies

In total, 5 studies examined other internet-based interventions, not CBT-related [24,28,29,31,47]. All 5 studies looked into a different treatment, some more related to sound therapy, others focusing more on mindfulness or cognitive training. All studies reported promising results concerning the effectiveness of the treatment, except for Xing et al [47], who concluded that auditory intensive cognitive training was not associated with changes in self-reported tinnitus. Further details can be found in Table 5.

XSL•FO
RenderX

Table 5. Table of evidence for the studies concerning other internet-based interventions.

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Alonso-Valerdi et al. [24], 2021	(1) TRT[b], (2) TEAE[c], (3) ADT[d], (4) BBT[e], (5) music	No treatment	2 months (60 min daily)	FU1[f]: after 1 week; FU2[g]: after 5 weeks; FU3[h]: after 8 weeks	THI[i]	HADS[j], audiometry, tinnitus characteristics analysis, and BL[k] neurophysiological analysis.	Identification of ± or 0 effects *t* test.	All negative differences between BL and FU[l] were statistically significant (*P*=.009).	29.4% of tinnitus sufferers in the TRT group reported diminution of their tinnitus perception, followed by those in TEAE (18.8%), BBT or ADT (14.3%), and music (7.7%) groups.	Dropout: music 6/16; BBT[m] 4/18; TRT 3/18; TEAE 5/18; ADT 5/18; control 3/14.	TRT most effective sound-based therapy to reduce tinnitus perception for individuals without anxiety. BBT, ADT, and TEAE produced very similar effects.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Connell et al [28], 2023	Customized acoustic therapy delivered through a web-based platform	N/A[n]	6 weeks (2 h daily)	Every week of treatment	FiveQ tinnitus symptoms severity scores	N/A.	Wilcoxon signed-rank tests, Mann-Whitney U-tests, linear mixed models, Tukey method *P* value adjustment for multiple variables, and a Kenward-Roger degrees of freedom estimation.	The mean difference between pre- and posttreatment of –9.3 (SD 15.5) was statistically significant (*P*<.001).	At treatment end point, 39/58 (67.2%) patients demonstrated an overall improvement in FiveQ, 4/58 (6.9%) had no change, and 15/58 (25.9%) declined.	51% (57/115) dropout.	The findings of this study support a customized acoustic therapy program as an efficacious tinnitus treatment modality that can improve symptom severity. Patients with low-frequency tinnitus and mild or above symptom severity at baseline demonstrated a greater treatment response.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Gans et al [29], 2023	i-MBT-SR[o] focusing on applying mindfulness to tinnitus	N/A	8 weeks (2 h lesson/wk and 30 min daily instructor-led meditation practice)	FU1: posttreatment; FU2: 6 months posttreatment	TFI[p]	PSS[q] TIQ[r].	Repeated-measures 2 (group) × 3 (time) repeated-measures (using Wilks' Lambda because of heterogeneity of variance) MANOVA[s], ANOVA with the Greenhouse-Geisser method.	Post hoc contrast comparisons with Bonferroni adjustments were used to examine the differences between the levels of each main effect. Contrasts revealed that the greatest gains were made in TFI and PSS between time 1 and either time 2 or time 3 time periods.	31 (72%) patients experienced clinically meaningful improvements of 13 or more on the TFI.	534/577 (94%) did not complete FU1, and 555/577 (98%) FU2.	The iMBT-SR course appears to be a viable and effective treatment modality for the tinnitus population.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Jackson and Woolmer [31], 2024	Online body scan meditation intervention	Waiting list	8 weeks (3-4 times a week)	FU1: posttreatment; FU2: 6 months posttreatment	TFI	TCQ[t] MAAS[u].	1-way ANOVA design.	A significant main effect of condition on self-reported tinnitus distress measured by the TFI ($F_{1,100}$=18.318; P<.001; 2P=.15) was present.	The mean TFI change over 8 weeks of body scan meditation was –9.52 in comparison to –0.58 for patients on the waiting list. 29.63% of the IG[v] experienced a clinically significant change on the TFI of more than 14/100.	15/48 lost to follow-up in the control group and 9/54 in IG ≥LOCF[w].	These results evidence that those in the body scan meditation group experienced a significant reduction in tinnitus distress, an increase in mindful thought, and a decrease in negative tinnitus cognitions.	⊕⊕⊕◯

XSL•FO
RenderX

Study	Intervention				Outcome			Results				
Author and year	Intervention	Control group	Treatment period	Follow-up	Primary tinnitus outcome	Additional outcome measures	Statistical analysis method	Data results	Statistical analysis results	Dropout	Conclusion	GRADE[a]
Xing et al [47], 2021	Brain HQ[x] auditory intensive training	Brain HQ nonauditory intensive training	8 weeks (5 d/wk, 20 min/d)	FU1: posttreatment; FU2: 1 month posttreatment	TFI	TGBS[y], CFQ[z], BAI[aa], PHQ-9[ab].	Spearman correlation mixed effects model with an unrestricted variance-covariance matrix.	Marginal mean differences for IG and CG[ac] were at 1 month (0.24, 95% CI –11.20 to 10.7), at 2 months 2.17 (95% CI –8.50 to 13.3), and at 3 months 3.36 (95% CI –7.91 to 14.6).	At 3 months, 13% of the CG and 19% of the IG experienced a clinically relevant reduction in the TFI of 13 or more.	>20%	Auditory intensive cognitive training was not associated with changes in self-reported tinnitus.	⊕⊕⊕◯

[a]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.
[b]TRT: tinnitus retraining therapy.
[c]TEAE: therapy for an enriched acoustic environment.
[d]ADT: auditory discrimination therapy.
[e]BBT: Broad Band Therapy.
[f]FU1: first follow-up moment.
[g]FU2: second follow-up moment.
[h]FU3: third follow-up moment.
[i]THI: Tinnitus Handicap Inventory.
[j]HADS: Hospital Anxiety and Depression Scale.
[k]BL: baseline.
[l]FU: follow-up moment.
[m]BBT: broad band therapy.
[n]N/A: not applicable.
[o]iMBTSR: internet-based mindfulness-based tinnitus stress reduction.
[p]TFI: Tinnitus Functional Index.
[q]PSS: Perceived Stress Scale.
[r]TIQ: Tinnitus Impact Questionnaire.
[s]MANOVA: multivariate ANOVA.
[t]TCQ: Tinnitus Cognitions Questionnaire.
[u]MAAS: Mindful Attention Awareness Scale.
[v]IG: intervention group.
[w]LOCF: Last Observation Carried Forward.
[x]HQ: Hyperacusis Questionnaire.
[y]TGBS: Tinnitus Global Bothersome Scale.
[z]CFQ: Cognitive Failures Questionnaire.
[aa]BAI: Beck Anxiety Inventory.
[ab]PHQ-9: Patient Health Questionnaire-9.
[ac]CG: control group.

XSL•FO
RenderX

Discussion

Principal Findings

The previous systematic review published in 2023 on the topic of telerehabilitation for patients with chronic subjective tinnitus indicated that low-contact treatment provided from a distance through several telerehabilitation forms could be used as a substitution for or an addition to in-person clinical care for tinnitus [13]. It can potentially reduce the costs of treatment without sacrificing the level of effectiveness. Due to the advances in technology made every day, telerehabilitation is a quickly evolving research topic, and several new studies on telerehabilitation for tinnitus have been published in the last couple of years. Therefore, the aim of this systematic review was to give an updated overview of the research concerning the effectiveness of telerehabilitation interventions for self-management of tinnitus published between 2022 and 2025. The recent studies included in this review confirm the earlier findings that telerehabilitation shows promising results in effectively reducing tinnitus distress in a less costly and more accessible way. However, the recent findings also demonstrate some changes within the field of telerehabilitation for tinnitus.

In the last decade, a lot of research has been published concerning the development and effectiveness of iCBT programs. These studies formed the vast majority of the preceding review. In the last 3 years, a shift in telerehabilitation research has been noticeable. The interest in iCBT has abated, and the use of smartphone apps for telerehabilitation for tinnitus is a growing topic. Not only in the field of tinnitus, but in medicine and rehabilitation in general, the use of smartphone apps to offer treatment from a distance is on the rise and forms the newest trending topic in the field of medicine [49,50]. The implementation of smartphone apps offers several advantages over conventional internet-based interventions that typically rely on desktop or browser-based platforms. First, it ensures greater accessibility and reach, increasing the likelihood of consistent patient engagement and adherence. They also enable more flexible and context-sensitive delivery of rehabilitation tasks. Push notifications, reminders, and just-in-time prompts can support adherence and promote habitual use, increasing spontaneous use. Smartphone apps also tend to perform better in terms of user friendliness and patient satisfaction [51-53]. It is not just the use of apps, but also the content of the researched apps that has shifted. While earlier research investigated the use of apps specifically for sound therapy for tinnitus patients, the newer research also shows other treatments in the form of an app [54,55]. The majority of the researched apps still gravitate more toward the use of sound therapy, but also, CBT, TRT, and psycho-education are made more accessible in this way. It is, however, noticeable that while research concerning sound therapy in app form has scaled up its sample size in the last couple of years to strengthen its power, this is not yet the case for the other treatments in app form. CBT, TRT, and psycho-education without sound therapy were still investigated in smaller sample sizes and pilot studies. Additionally, 1 study looked into the use of an app with a cervical exercise program for patients with somatic tinnitus. These patients experienced neck complaints in addition to their tinnitus, and these complaints can influence the tinnitus sound and loudness due to somatosensory interference. As a tailored care approach detecting present influencing factors is recommended, it is good that telerehabilitation is expanding and not only focusing on the psychosocial aspect, but also on specific other potential factors that might be of influence, such as neck complaints.

Another trend noticeable from the research of the last couple of years is that internet-based programs are also expanding beyond iCBT with or without guidance. This review included studies looking into several forms of sound therapy, customized acoustic therapy, mindfulness, meditation, and even cognitive training through a web-based platform. These treatments all demonstrated promising results for reducing tinnitus distress. Except for the cognitive training program with acoustic cognitive exercises, which was not reported with self-reported changes in tinnitus experiences. However, given the promising results concerning the link between cognition and tinnitus and the potential use of cognitive training for tinnitus, future studies on cognitive training telerehabilitation forms for tinnitus remain relevant [56-58]. Sound therapy, meditation, and mindfulness have been widely researched in a nontelerehabilitation context for tinnitus, making it a great avenue to explore how these treatment forms can be further improved by increasing accessibility.

Important notes from the preceding review were the high dropout rates from iCBT programs and the lack of reporting dropout in studies concerning the use of smartphone apps [13]. The same note can still be made in the current review. For iCBT, 2 of 5 studies reported dropout, which was around 34% at postintervention follow-up. This is in line with previous similar studies concerning iCBT in this population. According to Beukes et al [26], patients withdraw from treatment due to time constraints or lack of engagement. In accordance, the preceding review stated that earlier research indicated that barriers such as lack of time, engagement, motivation, and openness of the patient cause high dropout and should be taken into consideration. The previous review also concluded that future research should further explore which factors are most likely to cause the lack of compliance and how clinicians can counteract these. Therefore, it is crucial that studies analyze and report where the dropout is coming from, as Beukes et al [26] did extensively in their discussion. For the smartphone apps, a change could be noted, namely, the dropout rates were reported in 8 of 13 studies and ranged from no dropout to 53%. In other self-management intervention forms discussed, the rate ranged up to 94%, but on average was between 34% and 63%. The reasons for dropout noted were related to practical obstacles or treatment preferences from the patients. Oron et al [36] offered that monitoring programs might enhance the retention of treatment. In other treatment forms, the dropout rate and reasons for dropout were often not listed. As stated in the preceding review, to enhance the effectiveness of telerehabilitation, the limited current understanding of predictors of dropout should be tackled. Additionally, potential ways to improve adherence, such as therapist guidance levels, motivational feedback, automated reminders, etc, should be explored further [59].

XSL•FO
RenderX

Limitations

The included studies of this systematic review presented similar limitations as earlier research concerning this topic summarized in the previous review. All included studies scored some to great concerns of RoB. As a result, it needs to be acknowledged that the overall certainty of evidence ranged from low to moderate certainty. In addition, some crucial confounding parameters, such as the presence of hearing loss, hyperacusis, anxiety, depression, or sleeping problems, were not taken into consideration by all studies. However, it needs to be noted that concerning the presence of hearing loss, improvements were made on this level. A total of 15 of 24 studies reported how large the proportion of participants with hearing loss was or how many patients used hearing aids. Yet, how they defined the presence of hearing loss was often not clarified. An audiometric examination of tinnitus patients is therefore of value when researching patients with tinnitus because hearing loss is an important risk factor for developing tinnitus [1]. Due to the additional distress of hearing loss, tinnitus patients with hearing loss might present with worse tinnitus distress levels in comparison to tinnitus patients without hearing loss, according to earlier research [60].

While hearing loss is an important confounding factor to keep in mind, other factors also need to be considered: gender, age, mental state, hyperacusis, and neck and jaw complaints [61-71]. These factors might play a role in how patients perceive their tinnitus and how they react to treatment forms [61-71]. Additionally, the link between tinnitus and comorbidities such as anxiety, depression, stress, sleeping problems, hyperacusis, etc, are already widely researched [66,72-74]. The more serious these comorbidities are, the worse the prognosis of the tinnitus [73]. All the factors mentioned above might have an impact on the effectiveness of a treatment and should, therefore, be considered in future research. Additionally, for this reason, there is an increasing need for personalization of the treatment approach. Currently, often 1 factor is taken into consideration; however, a multidisciplinary and tailored approach is recommended. Therefore, the generalizability of these results is low, tinnitus mechanisms can greatly differ individually, and the limited tailoring options of these self-management interventions cause the evidence concerning the efficacy to remain rather limited. Due to a lack of high-quality and homogeneous studies concerning this topic, no confident conclusions can be made concerning the generalizability of self-management interventions for patients with tinnitus. Telerehabilitation has the potential to make treatment tailored to the needs of the patients more accessible; however, future research should invest more into using this potential. Future studies should focus on delivering high-quality study designs with homogenous outcomes to build the evidence of the strengths of telerehabilitation, which are now considered potential benefits.

Conclusions

This updated systematic review concerning the effectiveness of telerehabilitation for patients with chronic subjective tinnitus confirms the conclusion stated in the preceding review. The results of this systematic review indicate that telerehabilitation in the form of iCBT with guidance, self-help manuals, self-help devices, smartphone apps, and other internet-based interventions (except for cognitive training) are effective in relieving tinnitus distress. Telerehabilitation might form an alternative or additional tool to the recommended in-person care that patients currently receive. This review accentuated the shift from iCBT to the growing interest in the use of smartphone apps, increasing the accessibility of the treatments even more. However, the most significant barrier to the success of telerehabilitation remains the lack of compliance with the treatment. The great expansion of research within the field of telerehabilitation for tinnitus treatment in recent years indicates that it remains an evolving topic to keep track of.

Acknowledgments

The authors gratefully acknowledge Fonds Wetenschappelijk Onderzoek for its financial support. No generative artificial intelligence tools (including large language models such as ChatGPT, OpenAI) were used in the writing, editing, data analysis, interpretation of results, or preparation of tables and figures for this paper. All content was produced solely by the authors.

Funding

This study was supported by the Research Foundation – Flanders (Fonds Wetenschappelijk Onderzoek). The funding agency had no role in the design of this study; in the collection, analysis, and interpretation of data; or in writing the paper.

Data Availability

All data underlying the findings of this systematic review are derived from the studies included in the review and are publicly available through the original publications, which are fully referenced in this paper. No additional primary data were generated or collected for this work. Extracted datasets and materials used for the review process are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

PRISMA checklist for systematic review.
[PDF File (Adobe PDF File), 48 KB - jmir_v28i1e83529_app1.pdf]

Multimedia Appendix 2
Search strategy updated.
[DOC File , 33 KB - jmir_v28i1e83529_app2.doc]

Multimedia Appendix 3
Risk of bias summarization of newly included papers.
[DOCX File , 91 KB - jmir_v28i1e83529_app3.docx]

Multimedia Appendix 4
Tables of study and patient characteristics of newly included papers.
[DOCX File , 29 KB - jmir_v28i1e83529_app4.docx]

References

1. Baguley D, McFerran D, Hall D. Tinnitus. Lancet 2013;382(9904):1600-1607. [doi: 10.1016/s0140-6736(13)60142-7]
2. De Ridder D, Schlee W, Vanneste S, Londero A, Weisz N, Kleinjung T, et al. Tinnitus and tinnitus disorder: theoretical and operational definitions (an international multidisciplinary proposal). Prog Brain Res 2021;260:1-25. [doi: 10.1016/bs.pbr.2020.12.002] [Medline: 33637213]
3. Jarach CM, Lugo A, Scala M, van den Brandt PA, Cederroth CR, Odone A, et al. Global prevalence and incidence of tinnitus: a systematic review and meta-analysis. JAMA Neurol 2022;79(9):888-900. [doi: 10.1001/jamaneurol.2022.2189] [Medline: 35939312]
4. Demoen S, Cardon E, Jacquemin L, Timmermans A, Van Rompaey V, Gilles A, et al. Health-related quality of life in subjective, chronic tinnitus patients: a scoping review. J Assoc Res Otolaryngol 2024;25(2):103-129. [doi: 10.1007/s10162-024-00926-5] [Medline: 38253898]
5. Soni A, Dubey A. Chronic primary tinnitus: a management dilemma. Audiol Res 2020;10(2):55-66. [doi: 10.3390/audiolres10020010] [Medline: 33255533]
6. Han BI, Lee HW, Ryu S, Kim J. Tinnitus update. J Clin Neurol 2021;17(1):1-10. [doi: 10.3988/jcn.2021.17.1.1] [Medline: 33480192]
7. Maes IH, Cima RFF, Vlaeyen JW, Anteunis LJC, Joore MA. Tinnitus: a cost study. Ear Hear 2013;34(4):508-514. [doi: 10.1097/AUD.0b013e31827d113a] [Medline: 23411656]
8. Trochidis I, Lugo A, Borroni E, Cederroth CR, Cima R, Kikidis D, et al. Systematic review on healthcare and societal costs of tinnitus. Int J Environ Res Public Health 2021;18(13):6881. [doi: 10.3390/ijerph18136881] [Medline: 34206904]
9. Simões JP, Neff PKA, Langguth B, Schlee W, Schecklmann M. The progression of chronic tinnitus over the years. Sci Rep 2021;11(1):4162. [doi: 10.1038/s41598-021-83068-5] [Medline: 33602995]
10. Bokolo AJ. Exploring the adoption of telemedicine and virtual software for care of outpatients during and after COVID-19 pandemic. Ir J Med Sci 2021;190(1):1-10. [doi: 10.1007/s11845-020-02299-z] [Medline: 32642981]
11. Scaglione T, Kuzbyt B. Tinnitus management: the utilization of a hybrid telehealth and in-person delivery model. Semin Hear 2021;42(2):115-122. [doi: 10.1055/s-0041-1731692] [Medline: 34381295]
12. Beukes EW, Lourenco MPCG, Biot L, Andersson G, Kaldo V, Manchaiah V, et al. Suggestions for shaping tinnitus service provision in Western Europe: lessons from the COVID-19 pandemic. Int J Clin Pract 2021;75(7):e14196. [doi: 10.1111/ijcp.14196] [Medline: 33837639]
13. Demoen S, Chalimourdas A, Timmermans A, Van Rompaey V, Vanderveken OM, Jacquemin L, et al. Effectiveness of telerehabilitation interventions for self-management of tinnitus: systematic review. J Med Internet Res 2023;25:e39076. [doi: 10.2196/39076] [Medline: 36757768]
14. Searchfield GD, Sanders PJ, Doborjeh Z, Doborjeh M, Boldu R, Sun K, et al. A state-of-art review of digital technologies for the next generation of tinnitus therapeutics. Front Digit Health 2021;3:724370 [FREE Full text] [doi: 10.3389/fdgth.2021.724370] [Medline: 34713191]
15. Langguth B, Kleinjung T, Schlee W, Vanneste S, De Ridder D. Tinnitus guidelines and their evidence base. J Clin Med 2023;12(9):3087 [FREE Full text] [doi: 10.3390/jcm12093087] [Medline: 37176527]
16. Thomas EE, Haydon HM, Mehrotra A, Caffery LJ, Snoswell CL, Banbury A, et al. Building on the momentum: sustaining telehealth beyond COVID-19. J Telemed Telecare 2020;28(4):301-308. [doi: 10.1177/1357633x20960638]
17. Saharkhiz M, Rao T, Parker-Lue S, Borelli S, Johnson K, Cataife G. Telehealth expansion and medicare beneficiaries' care quality and access. JAMA Netw Open 2024;7(5):e2411006 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.11006] [Medline: 38739388]

XSL•FO
RenderX

18. Morelli S, Daniele C, D'Avenio G, Grigioni M, Giansanti D. Optimizing telehealth: leveraging key performance indicators for enhanced telehealth and digital healthcare outcomes (telemechron study). Healthcare (Basel) 2024;12(13):1319 [FREE Full text] [doi: 10.3390/healthcare12131319] [Medline: 38998854]
19. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
20. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
21. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
22. McGuinness LA, Higgins JPT. Risk-of-bias visualization (robvis): an R package and Shiny web app for visualizing risk-of-bias assessments. Res Synth Methods 2021;12(1):55-61. [doi: 10.1002/jrsm.1411] [Medline: 32336025]
23. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336(7650):924-926. [doi: 10.1136/bmj.39489.470347.ad]
24. Alonso-Valerdi LM, González-Salazar JI, Ibarra-Zarate DI. Neuropsychological monitoring of current acoustic therapies as alternative treatment of chronic tinnitus. Am J Otolaryngol 2021;42(6):103109. [doi: 10.1016/j.amjoto.2021.103109] [Medline: 34175772]
25. Altissimi G, Musacchio A, Pace A, Greco A, Magliulo G, Iagnocco G, et al. Sound therapy in patients with tinnitus: traditional sound generators vs. mobile apps. Eur Rev Med Pharmacol Sci 2024;28(11):3781-3786. [doi: 10.26355/eurrev_202406_36384] [Medline: 38884513]
26. Beukes EW, Andersson G, Fagelson M, Manchaiah V. Audiologist-supported internet-based cognitive behavioral therapy for tinnitus in the United States: a pilot trial. Am J Audiol 2021;30(3):717-729. [doi: 10.1044/2021_aja-20-00222]
27. Beukes EW, Andersson G, Manchaiah V. Long-term efficacy of audiologist-guided internet-based cognitive behaviour therapy for tinnitus in the United States: a repeated-measures design. Internet Interv 2022;30:100583. [doi: 10.1016/j.invent.2022.100583] [Medline: 36353148]
28. Connell JT, Bassiouni A, Harrison E, Laden S, O'Brien S, Sahota R, et al. Customised acoustic therapy delivered through a web-based platform-an innovative approach to tinnitus treatment. Clin Otolaryngol 2023;48(2):226-234. [doi: 10.1111/coa.14027] [Medline: 36550768]
29. Gans JJ, Holst J, Holmes C, Hudock D. Healing from home: examination of an online mindfulness-based tinnitus stress reduction course during the 2020 COVID pandemic. Am J Audiol 2023;32(1):160-169. [doi: 10.1044/2022_aja-22-00063]
30. Goshtasbi K, Tawk K, Khosravi P, Abouzari M, Djalilian HR. Smartphone-based cognitive behavioral therapy and customized sound therapy for tinnitus: a randomized controlled trial. Ann Otol Rhinol Laryngol 2025;134(2):125-133. [doi: 10.1177/00034894241297594] [Medline: 39506291]
31. Jackson JG, Woolmer CD. An 8-week online body scan meditation intervention for tinnitus: accessibility, adherence, and rates of clinically meaningful success. Mindfulness 2024;15(5):1176-1187. [doi: 10.1007/s12671-024-02357-y]
32. Kutyba J, Gos E, Jędrzejczak WW, Raj-Koziak D, Karpiesz L, Niedziałek I, et al. Effectiveness of tinnitus therapy using a mobile application. Eur Arch Otorhinolaryngol 2022;279(3):1257-1267 [FREE Full text] [doi: 10.1007/s00405-021-06767-9] [Medline: 33783597]
33. Kutyba JJ, Jędrzejczak WW, Gos E, Raj-Koziak D, Skarzynski PH. Chronic tinnitus and the positive effects of sound treatment via a smartphone app: mixed-design study. JMIR mHealth uHealth 2022;10(4):e33543. [doi: 10.2196/33543] [Medline: 35451975]
34. Manchaiah V, Beukes EW, Andersson G, Bateman E, Swanepoel DW, Uhler K, Vinay. Individuals with tinnitus report more positive experiences following internet-based cognitive behavioral therapy. Clin Pract 2024;14(4):1615-1624 [FREE Full text] [doi: 10.3390/clinpract14040130] [Medline: 39194934]
35. Michiels S, Wölflick S, Simões J, Schlee W. Exploring app-based physiotherapy for somatic tinnitus: results from a pilot study. J Clin Med 2024;13(23):7203 [FREE Full text] [doi: 10.3390/jcm13237203] [Medline: 39685670]
36. Oron Y, Ben David BM, Doron G. Brief cognitive-behavioral training for tinnitus relief using a mobile application: a pilot open trial. Health Inf J 2022;28(1):14604582221083483 [FREE Full text] [doi: 10.1177/14604582221083483] [Medline: 35349777]
37. Ravi P, Samayan K, Sivagnanapandian D. Tinnitus retraining therapy: in person versus mobile application based - two case scenarios. Indian J Otolaryngol Head Neck Surg 2025;77(2):1142-1146. [doi: 10.1007/s12070-024-05300-8] [Medline: 40070760]
38. Rodrigo H, Beukes EW, Andersson G, Manchaiah V. Internet-based cognitive-behavioural therapy for tinnitus: secondary analysis to examine predictors of outcomes. BMJ Open 2021;11(8):e049384 [FREE Full text] [doi: 10.1136/bmjopen-2021-049384] [Medline: 34417217]
39. Rodrigo H, Beukes EW, Andersson G, Manchaiah V. Exploratory data mining techniques (decision tree models) for examining the impact of internet-based cognitive behavioral therapy for tinnitus: machine learning approach. J Med Internet Res 2021;23(11):e28999 [FREE Full text] [doi: 10.2196/28999] [Medline: 34726612]

XSL•FO
RenderX

40. Schlee W, Neff P, Simoes J, Langguth B, Schoisswohl S, Steinberger H, et al. Smartphone-guided educational counseling and self-help for chronic tinnitus. J Clin Med 2022;11(7):1825 [FREE Full text] [doi: 10.3390/jcm11071825] [Medline: 35407432]
41. Searchfield GD, Sanders PJ. A randomized single-blind controlled trial of a prototype digital polytherapeutic for tinnitus. Front Neurol 2022;13:958730. [doi: 10.3389/fneur.2022.958730] [Medline: 35989940]
42. Seol HY, Jo M, Moon IJ. The effectiveness of a smartphone application for tinnitus relief. Healthcare (Basel) 2023;11(17):2368. [doi: 10.3390/healthcare11172368] [Medline: 37685402]
43. Shetty HN. Development and validation of self-management manual for acute tinnitus. Indian J Otolaryngol Head Neck Surg 2024;76(3):2654-2659. [doi: 10.1007/s12070-024-04504-2] [Medline: 38883543]
44. Suh MW, Park MK, Kim Y, Kim YH. The treatment outcome of smart device-based tinnitus retraining therapy: prospective cohort study. JMIR mHealth uHealth 2023;11:e38986 [FREE Full text] [doi: 10.2196/38986] [Medline: 36633890]
45. Tang D, Peng Y, Gu D, Wu Y, Li H. Digital frequency customized relieving sound for chronic subjective tinnitus management: prospective controlled study. J Med Internet Res 2025;27:e60150 [FREE Full text] [doi: 10.2196/60150] [Medline: 39823588]
46. Tang D, Wang H, Gu D, Ye L, Sun S, Li H. The fudan tinnitus relieving system application for tinnitus management. BMC Med Inform Decis Mak 2023;23(1):76. [doi: 10.1186/s12911-023-02164-w] [Medline: 37085904]
47. Xing M, Kallogjeri D, Piccirillo JF. Investigating the impact of cognitive training for individuals with bothersome tinnitus: a randomized controlled trial. Otolaryngol Head Neck Surg 2021;165(6):854-861. [doi: 10.1177/0194599821994742] [Medline: 33650921]
48. Andersson G, Kaldo V. Internet-based cognitive behavioral therapy for tinnitus. J Clin Psychol 2004;60(2):171-178. [doi: 10.1002/jclp.10243] [Medline: 14724924]
49. Nussbaum R, Kelly C, Quinby E, Mac A, Parmanto B, Dicianno BE. Systematic review of mobile health applications in rehabilitation. Arch Phys Med Rehabil 2019;100(1):115-127. [doi: 10.1016/j.apmr.2018.07.439] [Medline: 30171827]
50. Moral-Munoz JA, Zhang W, Cobo MJ, Herrera-Viedma E, Kaber DB. Smartphone-based systems for physical rehabilitation applications: a systematic review. Assist Technol 2021;33(4):223-236. [doi: 10.1080/10400435.2019.1611676] [Medline: 31112461]
51. Lee C, Ahn J, Lee BC. A systematic review of the long-term effects of using smartphone- and tablet-based rehabilitation technology for balance and gait training and exercise programs. Bioengineering (Basel) 2023;10(10):1142. [doi: 10.3390/bioengineering10101142] [Medline: 37892872]
52. Arensman R, Kloek C, Pisters M, Koppenaal T, Ostelo R, Veenhof C. Patient perspectives on using a smartphone app to support home-based exercise during physical therapy treatment: qualitative study. JMIR Hum Factors 2022;9(3):e35316 [FREE Full text] [doi: 10.2196/35316] [Medline: 36098993]
53. Simmich J, Ross MH, Russell T. Real-time video telerehabilitation shows comparable satisfaction and similar or better attendance and adherence compared with in-person physiotherapy: a systematic review. J Physiother 2024;70(3):181-192 [FREE Full text] [doi: 10.1016/j.jphys.2024.06.001] [Medline: 38879432]
54. Mehdi M, Dode A, Pryss R, Schlee W, Reichert M, Hauck FJ. Contemporary review of smartphone apps for tinnitus management and treatment. Brain Sci 2020;10(11):867 [FREE Full text] [doi: 10.3390/brainsci10110867] [Medline: 33212928]
55. Mehdi M, Stach M, Riha C, Neff P, Dode A, Pryss R, et al. Smartphone and mobile health apps for tinnitus: systematic identification, analysis, and assessment. JMIR mHealth uHealth 2020;8(8):e21767 [FREE Full text] [doi: 10.2196/21767] [Medline: 32808939]
56. Kallogjeri D, Piccirillo JF, Spitznagel E, Hale S, Nicklaus JE, Hardin FM, et al. Cognitive training for adults with bothersome tinnitus: a randomized clinical trial. JAMA Otolaryngol Head Neck Surg 2017;143(5):443-451. [doi: 10.1001/jamaoto.2016.3779] [Medline: 28114646]
57. Tegg-Quinn S, Bennett RJ, Eikelboom RH, Baguley DM. The impact of tinnitus upon cognition in adults: a systematic review. Int J Audiol 2016;55(10):533-540. [doi: 10.1080/14992027.2016.1185168] [Medline: 27240696]
58. Henry JL, Wilson PH. An evaluation of two types of cognitive intervention in the management of chronic tinnitus. Scand J Behav Ther 1998;27(4):156-166. [doi: 10.1080/02845719808408510]
59. Simon N, Robertson L, Lewis C, Roberts NP, Bethell A, Dawson S, et al. Internet-based cognitive and behavioural therapies for post-traumatic stress disorder (PTSD) in adults. Cochrane Database Syst Rev 2021;5(5):CD011710 [FREE Full text] [doi: 10.1002/14651858.CD011710.pub3] [Medline: 34015141]
60. Mahafza N, Zhao F, El Refaie A, Chen F. A comparison of the severity of tinnitus in patients with and without hearing loss using the tinnitus functional index (TFI). Int J Audiol 2021;60(3):220-226. [doi: 10.1080/14992027.2020.1804081] [Medline: 32787604]
61. Seydel C, Haupt H, Olze H, Szczepek AJ, Mazurek B. Gender and chronic tinnitus: differences in tinnitus-related distress depend on age and duration of tinnitus. Ear Hear 2013;34(5):661-672. [doi: 10.1097/AUD.0b013e31828149f2] [Medline: 23439056]
62. Han TS, Jeong J, Park S, Kim JJ. Gender differences affecting psychiatric distress and tinnitus severity. Clin Psychopharmacol Neurosci 2019;17(1):113-120. [doi: 10.9758/cpn.2019.17.1.113] [Medline: 30690946]

XSL•FO
RenderX

63. Hiller W, Goebel G. Factors influencing tinnitus loudness and annoyance. Arch Otolaryngol Head Neck Surg 2006;132(12):1323-1330. [doi: 10.1001/archotol.132.12.1323] [Medline: 17178943]
64. Lugo A, Trpchevska N, Liu X, Biswas R, Magnusson C, Gallus S, et al. Sex-specific association of tinnitus with suicide attempts. JAMA Otolaryngol Head Neck Surg 2019;145(7):685-687. [doi: 10.1001/jamaoto.2019.0566] [Medline: 31046059]
65. Van der Wal A, Luyten T, Cardon E, Jacquemin L, Vanderveken OM, Topsakal V, et al. Sex differences in the response to different tinnitus treatment. Front Neurosci 2020;14:422. [doi: 10.3389/fnins.2020.00422] [Medline: 32477049]
66. Boecking B, Biehl R, Brueggemann P, Mazurek B. Health-related quality of life, depressive symptoms, anxiety, and somatization symptoms in male and female patients with chronic tinnitus. J Clin Med 2021;10(13):2798 [FREE Full text] [doi: 10.3390/jcm10132798] [Medline: 34202097]
67. Al-Swiahb J, Park SN. Characterization of tinnitus in different age groups: a retrospective review. Noise Health 2016;18(83):214-219 [FREE Full text] [doi: 10.4103/1463-1741.189240] [Medline: 27569409]
68. Shargorodsky J, Curhan SG, Curhan GC, Eavey R. Change in prevalence of hearing loss in US adolescents. JAMA 2010;304(7):772-778. [doi: 10.1001/jama.2010.1124] [Medline: 20716740]
69. Michiels S, Cardon E, Gilles A, Goedhart H, Vesala M, Schlee W. Somatosensory tinnitus diagnosis: diagnostic value of existing criteria. Ear Hear 2022;43(1):143-149. [doi: 10.1097/AUD.0000000000001105] [Medline: 34261856]
70. Michiels S, Sanchez TG, Oron Y, Gilles A, Haider HF, Erlandsson S, et al. Diagnostic criteria for somatosensory tinnitus: a Delphi process and face-to-face meeting to establish consensus. Trends Hear 2018;22:2331216518796403. [doi: 10.1177/2331216518796403] [Medline: 30213235]
71. Michiels S, Naessens S, Van de Heyning P, Braem M, Visscher CM, Gilles A, et al. The effect of physical therapy treatment in patients with subjective tinnitus: a systematic review. Front Neurosci 2016;10:545 [FREE Full text] [doi: 10.3389/fnins.2016.00545] [Medline: 27965530]
72. Li YL, Hsu Y, Lin C, Wu J. Sleep disturbance and psychological distress in adult patients with tinnitus. J Formosan Med Assoc 2022;121(5):995-1002. [doi: 10.1016/j.jfma.2021.07.022] [Medline: 34366185]
73. Hébert S. Psychological comorbidities of tinnitus. In: Searchfield GD, Zhang J, editors. The Behavioral Neuroscience of Tinnitus. Cham: Springer International Publishing; 2021:349-359.
74. Jacquemin L, Cardon E, Michiels S, Luyten T, Van der Wal A, De Hertogh W, et al. Hyperacusis: demographic, audiological, and clinical characteristics of patients at the ENT department. Eur Arch Oto-Rhino-Laryngol 2022;279(10):4899-4907. [doi: 10.1007/s00405-022-07336-4] [Medline: 35298688]

Abbreviations

CBT: cognitive behavioral therapy
CG: control group
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
iCBT: internet-based cognitive behavioral therapy
MeSH: Medical Subject Headings
PICO: population, intervention, comparison, and outcome
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
RCT: randomized controlled trial
RoB: risk of bias
ROBINS-I: Risk of Bias in Non-Randomized Studies of Interventions
TFI: Tinnitus Functional Index
TRT: tinnitus retraining therapy

Edited by S Brini; submitted 05.Sep.2025; peer-reviewed by A Chalimourdas, AA Danesh; comments to author 06.Nov.2025; revised version received 07.Jan.2026; accepted 07.Jan.2026; published 27.Feb.2026.

Please cite as:
Demoen S, Van Kerchove E, Timmermans A, Van Rompaey V, Michiels S, Gilles A
Effectiveness of Telerehabilitation Interventions for Self-Management of Tinnitus: Update of a Systematic Review
J Med Internet Res 2026;28:e83529
URL: https://www.jmir.org/2026/1/e83529
doi:10.2196/83529
PMID:

XSL•FO
RenderX

©Sara Demoen, Elise Van Kerchove, Annick Timmermans, Vincent Van Rompaey, Sarah Michiels, Annick Gilles. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Digital Health Interventions to Promote Physical Activity Among Adolescents: Systematic Review

Rui Shi Fan[1,2*], BEd; Jia Jun Jiang[1], BEd; Qing Yuan Zhou[1], BEd; Xin Yue Zhang[1,2*], MEd; Zhou Hang Wu[1,2], MEd; Liu Ji[1,2], PhD

[1]College of Physical Education and Health, East China Normal University, ShangHai, Shanghai, China

[2]Key Laboratory of Adolescent Health Assessment and Exercise Intervention of Ministry of Education, East China Normal University, ShangHai, China

[*]these authors contributed equally

Corresponding Author:
Liu Ji, PhD
College of Physical Education and Health
East China Normal University
500 Dongchuan Road
ShangHai, 200241
China
Phone: 86 13901817299
Email: lji@tyxx.ecnu.edu.cn

Abstract

Background: Insufficient physical activity among adolescents is a major global public health concern. Digital health interventions (DHIs) have gained increasing attention as a promising approach to promoting physical activity in adolescents. However, existing systematic reviews predominantly focus on single-intervention formats or specific study designs, while reviews that integrate multiple DHIs and diverse study designs remain scarce.

Objective: This systematic review aims to synthesize evidence from diverse DHIs and multiple study designs to assess their effectiveness in promoting physical activity among adolescents.

Methods: The review protocol was registered in PROSPERO (International Prospective Register of Systematic Reviews; CRD420251117923). This systematic review searched literature published between January 1, 2014, and June 30, 2025, across Web of Science, PubMed, EBSCO, Scopus, Embase, the Cochrane Library, ProQuest, and Google Scholar. The final search was completed on August 3, 2025. Using the PICOS (population, intervention, comparator, outcomes, and study design) framework, the review included adolescents aged 10-19 years and focused on evidence-based research promoting physical activity through DHIs. The review was limited to peer-reviewed English-language literature and excluded studies solely focused on measurement tools, those not evaluating intervention effectiveness, or those not involving adolescents. Two reviewers independently screened studies and extracted data. Research quality was assessed using the Joanna Briggs Institute tool. Findings were synthesized through narrative synthesis and qualitative content analysis.

Results: A total of 24 studies were included, involving approximately 12,183 adolescents. Study designs comprised 10 randomized controlled trials, 4 quasi-experimental studies, 3 quantitative research studies, 3 cross-sectional studies, and 4 mixed methods studies. Overall, 7 (29%) studies were of high quality, 16 (67%) were of moderate quality, and 1 (4%) was of low quality. Study populations included general adolescents as well as subgroups with specific health risks: insufficient physical activity (1/24, 4%), obesity or overweight (4/24, 17%), attention-deficit/hyperactivity disorder (1/24, 4%), cancer survivors (1/24, 4%), and at-risk youth (1/24, 4%). DHIs were categorized into 3 types: single-driver interventions (14/24, 58%), multimodal integrated interventions (7/24, 29%), and interaction-enhanced interventions (3/24, 13%). Most studies reported positive outcomes, including direct effectiveness (15/24, 63%), indirect effectiveness (8/24, 33%), and unclear effectiveness (1/24, 4%).

Conclusions: This systematic review synthesizes evidence from diverse research designs and multiple types of DHIs, offering a more comprehensive perspective than previous reviews focused on single designs or technological formats. The results indicate that DHIs generally enhance adolescent physical activity levels, although their effectiveness varies considerably across intervention types and study designs. The review fills key research gaps and highlights the critical role of intervention adaptability and implementation context. It also addresses practical concerns, including adolescents with special health conditions, digital health inequalities, and technology dependency. Despite limitations related to methodological quality and insufficient follow-up, this

XSL•FO
RenderX

review provides important evidence to inform practical application, policy development, and the equitable promotion of DHIs to enhance adolescent physical activity. Against the backdrop of rising global adolescent physical inactivity and widening health disparities, it also outlines directions for future high-quality research.

Trial Registration: PROSPERO CRD420251117923; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251117923

(J Med Internet Res 2026;28:e82395) doi:10.2196/82395

KEYWORDS

adolescents; physical activity; digital health interventions; systematic review; health promotion; behavior change; youth health promotion

Introduction

Background

Adolescence represents a pivotal developmental period for the establishment of long-term healthy lifestyle behaviors. Engaging in moderate physical activity during this stage contributes to physical development, improves cardiorespiratory fitness, supports mental health regulation, and shapes future health trajectories, including the prevention and management of chronic diseases. However, the prevalence of insufficient physical activity among adolescents has increased markedly in recent years and is now widely recognized as a significant global public health concern affecting youth populations [1].

Against this backdrop, digital health interventions (DHIs), an emerging approach to health promotion, have been increasingly implemented among adolescent populations [2]. These interventions typically utilize digital technologies—such as mobile health (mHealth) apps, wearable devices, and online fitness platforms—to facilitate behavioral change among adolescents. They offer advantages such as high accessibility and real-time feedback, rendering them particularly suitable for this demographic group [3,4]. For example, within the domain of gamified mobile apps, fitness platforms such as “Nike Training Club” enhance daily exercise participation through mechanisms including instructional video tutorials and task-based check-ins. Wearable devices, such as Fitbit and Garmin, deliver real-time feedback on physical activity indicators, including step count, heart rate, and energy expenditure, thereby supporting users in enhancing self-regulation of physical activity [5]. Overall, DHIs exhibit growing diversity and adaptability in promoting physical activity among adolescents, while simultaneously presenting novel research opportunities and methodological challenges within the field.

As DHIs evolve, understanding the underlying behavioral mechanisms is key to optimizing intervention design and improving effectiveness. The Behavior Change Wheel (BCW) explains how DHIs influence behavior change and provides a foundation for developing targeted strategies. Research suggests that the COM-B (capability, opportunity, motivation—behavior) system is a robust framework for designing DHIs. The COM-B system elucidates the mechanisms underlying behavior change [6], explaining how different technologies and strategies operate across various contexts. Proposed by Michie et al [6], this model is central to the BCW and has been widely applied in health behavior research. It explains how multimechanism strategies work together to promote health behavior change. The system proposes that behavior requires 3 elements: (1) capability, (2) opportunity, and (3) motivation. These elements interact to form a behavioral system; for example, enhanced capability can increase motivation, favorable opportunities can facilitate the use of capability, and higher motivation can drive further pursuit and development.

In recent years, DHIs have undergone continuous innovation in content strategy. The content of such interventions has evolved from single-dimensional information delivery to multidimensional models that integrate behavioral incentives, motivational guidance, and social support. Similarly, intervention modalities have transitioned from single-medium approaches to integrated, multicomponent systems that combine mobile apps, wearable technologies, and social networking platforms. Despite these advancements, research on DHIs aimed at promoting physical activity among adolescents remains substantially underdeveloped. At the level of primary research, existing studies targeting physical activity promotion among children and adolescents demonstrate considerable diversity in intervention modalities; however, a standardized classification framework remains absent. Moreover, substantial variation is observed across studies with respect to intervention tools, implementation frequency, and intervention duration, contributing to pronounced heterogeneity in reported outcomes [7]. Simultaneously, methodological weaknesses persist in some studies—such as the absence of control groups, small sample sizes, short intervention periods, or a lack of long-term follow-up—thereby undermining the interpretability and external validity of the findings [8].

Existing reviews primarily focus on specific DHIs, with few systematic reviews encompassing multiple intervention types and diverse study designs. Previous reviews, such as that by Wang et al [9], focused on mHealth app interventions using randomized controlled trials (RCTs) and reported benefits for total physical activity, sedentary behavior, BMI, agility, and muscle strength in children and adolescents. He et al [10] examined smartphone-based interventions using RCTs, showing increases in total physical activity and step counts among children and adolescents, with app-based interventions proving more effective. Sequí-Domínguez et al [11] examined eHealth interventions in RCTs and found that they reduced sedentary time in children and adolescents.

The scope of these previous systematic reviews was relatively narrow, focusing solely on specific DHIs or intervention types. The included studies were limited to a single study design,

XSL•FO
RenderX

resulting in incomplete interpretation of the evidence. Consequently, there remains a lack of comprehensive integration and comparison across multiple DHIs, as well as a paucity of reviews and further analyses encompassing diverse study designs.

Objectives

Against this backdrop, this systematic review aims to conduct a focused analysis of the existing literature targeting adolescent populations. It seeks to identify and examine the capabilities of different research designs and various types of DHIs in promoting physical activity among adolescents, thereby addressing current evidence gaps. The review is intended to provide theoretical support and an empirical foundation for optimizing subsequent intervention design and research.

Methods

Review Design and Protocol Registration

This systematic review adopts a systematic methodology to comprehensively examine the current landscape, intervention types, and implementation effectiveness of DHIs aimed at promoting physical activity among adolescents in recent years. In addition, it evaluates and synthesizes the methodological quality and practical outcomes reported in the existing literature. The review strictly adheres to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 reporting guidelines [12], encompassing key steps including literature retrieval, screening, data extraction, and quality appraisal. To ensure scientific rigor and interstudy comparability, a structured risk-of-bias assessment was performed using the Joanna Briggs Institute (JBI) Critical Appraisal Tool [13]. The review protocol was prospectively registered with the international systematic review registry PROSPERO (International Prospective Register of Systematic Reviews; registration number CRD420251117923), and the study was conducted and reported in strict accordance with the preestablished protocol. The study protocol can be accessed via the corresponding registration record on the PROSPERO platform.

Eligibility Criteria

Inclusion Criteria

The systematic review utilized the PICOS (population, intervention, comparator, outcomes, and study design) framework to define the following inclusion criteria.

- Population: Studies must include adolescent participants aged 10-19 years, with a primary focus on physical activity behaviors during this developmental stage.
- Intervention: Studies must examine DHIs aimed at promoting physical activity among adolescents. Eligible intervention formats include, but are not limited to, mHealth, wearable technologies, eHealth platforms, fitness tracking devices, virtual fitness programs, and gamified health apps.
- Comparison: Acceptable control conditions may include no intervention, traditional physical activity programs, or other nondigital health promotion approaches.
- Outcomes: Included studies must report outcomes related to adolescent physical activity behaviors, including but not limited to physical activity levels, exercise frequency, intensity of participation, sports involvement, and engagement in fitness-related activities.
- Study design: Eligible study designs include RCTs, quasi-experimental studies (QESs), quantitative research (QR), cross-sectional studies (CSSs), and mixed methods studies (MMSs).
- All included studies must be peer-reviewed publications written in English and published no later than June 30, 2025.

Exclusion Criteria

- Systematic reviews that exclusively examined physical activity assessment tools or indicators without evaluating intervention effectiveness.
- Systematic reviews involving participants who were not adolescents or whose data did not allow clear differentiation of adolescent subgroups.
- Systematic reviews classified as dissertations (master's or doctoral), review articles, conference papers, or other non–peer-reviewed publications.
- Systematic reviews for which the full text was not accessible.
- Systematic reviews with unclear research designs or markedly low methodological quality.

Information Sources

This systematic review retrieved relevant studies from multiple databases and information sources. Primary databases included PubMed (National Library of Medicine), Web of Science (Clarivate Analytics), EBSCO (EBSCOhost), Scopus (Elsevier), Embase (Elsevier), the Cochrane Library (Wiley Online Library), and ProQuest (ProQuest Platform), with Google Scholar serving as a supplementary search source. All databases were searched independently on their respective platforms, with no simultaneous searches conducted on a single platform.

To identify potentially overlooked studies, the research team conducted manual supplementary searches. This process involved reviewing the reference lists of included studies and examining references from relevant systematic reviews and meta-analyses. No additional studies or data were obtained from authors, experts, or other individuals. The review comprised 2 rounds of searching: the initial search was completed on June 10, 2025, for primary study identification, and a second, updated search was completed on August 3, 2025, to capture newly published studies. Both searches employed consistent search strategies.

Search Strategy

The literature search strategy was constructed in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Search Extension) guidelines to ensure transparency and reproducibility [14]. The search strategy was organized around 3 core concept categories. First, physical activity behaviors: "physical activity," "exercise," "physical fitness," "sports participation," and "fitness activity." Second, DHIs: "digital health," "mobile health," "wearable device," "electronic health," and "fitness tracking device." Third, study

XSL•FO
RenderX

participants: “adolescent” and “teenager.” Search terms within the same concept category were combined using the Boolean operator “OR,” while different concepts were linked with “AND.” Search fields and expressions were adjusted according to database-specific syntax and subject heading systems (eg, Medical Subject Headings [MeSH]).

The search time frame spanned from January 1, 2014, to June 30, 2025, and was limited to English-language original empirical research. This systematic review did not utilize published standardized search filters; the retrieval strategy was specifically designed for this study and was not adapted or reused from previous systematic reviews or meta-analyses. The complete search terms and database-specific parameters are detailed in Multimedia Appendix 1. Beyond these methods, this systematic review did not search study registries, conduct purposeful browsing of conference proceedings, government websites, or other online resources; employ additional complementary search methods; or undergo a dedicated peer review of the search strategy.

Selection Process

All search results were first imported into the EndNote reference management software (Clarivate Plc) for preliminary deduplication. Subsequently, the Rayyan online reference management platform was used for literature screening. The screening process was conducted independently by 2 researchers (RSF and JJJ) and comprised 2 stages: title and abstract screening, followed by full-text screening. Disagreements during screening were resolved through discussion and negotiation between the 2 researchers. If consensus could not be reached, a third researcher (XYZ) intervened to make the final decision, ensuring the objectivity and consistency of the screening process.

Data Collection Process

After study inclusion, data were collected using preestablished standardized data extraction forms. Two researchers (RSF and JJJ) independently extracted data on study characteristics, including basic study information, methodological features, forms of DHIs, and intervention outcomes. When critical information was missing or unclear, the researchers contacted the corresponding author to obtain supplementary data. Discrepancies identified during data extraction were resolved through discussion and consensus, and unresolved disagreements were referred to a third researcher (XYZ) for deliberation and final adjudication.

Study Outcomes (Data Items)

Data extraction was performed independently by 2 researchers using predefined standardized data extraction forms. Extracted data included primary and secondary outcomes. Primary outcomes comprised basic study information (eg, authors, study population characteristics, sample size), study design and methodological features, and specific forms of DHIs along with their effects on adolescent physical activity–related outcomes, summarized in conjunction with the reported outcome measures and overall intervention results.

Secondary outcomes included supplementary information supporting outcome classification and interpretation, such as geographic region, intervention duration, and follow-up or sample attrition rates. Disagreements during the process were resolved by a third researcher. Complete data extraction forms and the classification framework are presented in Tables 1 and 2.

XSL•FO
RenderX

Table 1. Overview of included studies (N=24).

Study	Age, mean (SD); grade; gender; condition; population	Region	Study design	Intervention duration	Digital health interventions	Effects of digital health interventions
Soltero et al [15]	• 14.9 (0.91); 53% female; obese/overweight	Spain	Quantitative research	12 weeks	Wearable device (Fitbit Watch) + SMS text messaging	High engagement in self-monitoring behavior and perceived increase in activity
Goodyear et al [16]	• 13-14; 53% female; general adolescents	The United Kingdom	Quantitative research	8 weeks	Wearable device (Fitbit) + virtual platform (with activity goal setting)	Encouraged youth to engage in more physical exercise
Van Dyck et al [17]	• 12-14; 49% female; general adolescents	Belgium	Quantitative research	1 week	SMS text messaging/chatbot + social media (Facebook) + SMS	Facebook and SMS text messaging were considered promising methods for physical activity interventions among vocational school adolescents
Dinç et al [18]	• 14.27 (0.44); 56% female; general adolescents	Turkey	Cross-sectional study	N/A[a]	mHealth[b]: smartphone app	Effectively improves physical activity levels and health awareness
Mojica et al [19]	• 11-14; 100% female; general adolescents	The United States	Cross-sectional study	N/A	Social media: cell phones, computers, game consoles, and internet	Increased physical activity and more frequent participation in daily physical education classes after the intervention
Ng et al [20]	• 11-15; 42% female; general adolescents	Finland	Cross-sectional study	N/A	Wearable device: physical activity tracker (smartwatches and heart rate monitors)	Positive correlation between physical activity behavior and use of wearables/apps; further research needed to verify relationship
Mendoza et al [21]	• • Intervention: 16.9 (1.5); 41.4% female; cancer survivors • Control: 16.3 (SD 1.5); 40% female; cancer survivors	The United States	Randomized controlled trial	10 weeks	mHealth: Fitbit + Facebook	Demonstrates good feasibility and acceptability
Larsen et al [22]	• 14.7 (2.1); insufficient physical activity	The United States	Randomized controlled trial	12 weeks	Virtual platform: intervention website	Shows good feasibility and acceptability, with a significant increase in self-reported physical activity levels
Thompson et al [23]	• 14-17; 51.88% female; general adolescents (group: control, pedometer only, pedometer + goal prompts, and pedometer + goal prompts + self-determination theory texts)	The United States	Randomized controlled trial	12 weeks	mHealth: smartphone (SMS text messaging) + pedometer	Moderate increase in average daily steps and moderate-to-vigorous physical activity
Guthrie et al [24]	• 12-14; 55% female; general adolescents	The United States	Randomized controlled trial	6 weeks	Virtual platform: online intervention system (Zamzee)	Positive impact on moderate-to-vigorous physical activity levels
Direito et al [25]	• 14-17; general adolescents	New Zealand	Randomized controlled trial	8 weeks	mHealth: smartphone (immersive application)	Demonstrated feasibility
Chen et al [26]	• 14.9 (1.7); 42% female; obesity or overweight	The United States	Randomized controlled trial	6 months	mHealth: + wearable device + virtual platform: SMS text messaging, Fitbit Flex, and online educational program	Increased number of days per week engaged in physical activity

XSL•FO
RenderX

Study	Age, mean (SD); grade; gender; condition; population	Region	Study design	Intervention duration	Digital health interventions	Effects of digital health interventions
Caillaud et al [27]	• Intervention: 10.9 (0.7); 51% female; general adolescents • Control: 10.4 (0.5); 62% female; general adolescents	Australia	Randomized controlled trial	5 weeks	Virtual platform: app (iEngage Program)	Improved physical activity goals, academic performance, and moderate-to-vigorous physical activity levels
Egilsson et al [28]	• 15.6 (0.26); 41% female; general adolescents	Iceland	Randomized controlled trial	6 weeks	mHealth: smartphone (Mobile Health Program)	Feasible and usable
Staiano et al [29]	• 11.2 (0.8); 46% female; obesity or overweight	The United States	Randomized controlled trial	24 weeks	Game: Squad Intervention (gaming console)	Improved physical activity levels
Stasinaki et al [30]	• PathMate2: 12.6 (range 11.4-16.9); 38.9% female; obesity or overweight • Control: 13.7 (range 10.9-16.8); 46.2% female; obesity or overweight	Switzerland	Randomized controlled trial	5.5 months	mHealth: smartphone (PathMate2)	Significant and sustained improvements in physical capacity and body composition
Ortega and Cushing [31]	• 13-18; 45% female; general adolescents	The United States	Mixed methods study	27 days	SMS text messaging/chatbot: text bot (temporally augmented goal setting)	Increased participation, with observable changes in physical activity before and after the intervention
Willinger et al [32]	• 12.6 (1.7);27% female; general adolescents	Germany	Mixed methods study	4 months	mHealth: smartphone (the KIJANI app)	Helps oneself and others to be more active in daily life
Schoenfelder et al [33]	• 15.5 (1.4); 54% female; with attention-deficit/hyperactivity disorder	The United States	Mixed methods study	4 weeks	mHealth + Fitbit Flex + Facebook	Average weekly step count increased; intervention shows promise for promoting physical activity in adolescents with attention-deficit/hyperactivity disorder
Koorts et al [34]	• 13.7 (0.4); 49%female; general adolescents	Australia	Mixed methods study	12 weeks	mHealth + wearable device (wrist-worn Fitbit)	Perceived short-term positive effect on exercise motivation
Glaser et al [35]	• Ninth to eleventh grade; 28% female; at-risk youth	Israel	Quasi-experimental study	8 months	Virtual platform: Friends Online Intervention Program—Zoom/video chat with online physical activities and dialogue	The program effectively increased physical activity among adolescents and reduced risky behaviors
Garde et al [36]	• 8-13; general adolescents	Canada	Quasi-experimental study	2 weeks	mHealth: smartphone game (Mobile Kids Monster Manor)	After the intervention, more physical activity was observed; the game's role in promoting activity needs further validation
Mateo-Orcajada et al [37]	• 13.96 (1.21); 48% female; general adolescents	Spain	Quasi-experimental study	10 weeks	mHealth: mobile app	Postintervention improvements in physical activity level, body composition, and physical fitness quality
Cushing et al [38]	• 13-18; 75% female; general adolescents	The United States	Quasi-experimental study	3 weeks	SMS/chatbot: tailored SMS text messaging intervention	Increased physical activity and reduced sedentary behavior

[a]N/A: not applicable.

[b]mHealth: mobile health.

XSL•FO
RenderX

Table 2. Effectiveness of DHIs[a] to promote PA[b] in adolescents: a classification and ranking by intervention type.

Study	Measurement	Key findings	Result	Effect
Single-driver intervention				
Dinç et al [18]	Questionnaire	• PA-themed (40.5%) and healthy nutrition–themed (33.3%) apps were the most frequently used. Adolescents who preferred PA apps had significantly higher healthy lifestyle belief scores (mean 61.82, SD 0.70) than those who preferred Ministry of Health apps (mean 43.54, SD 2.91; P<.001). Similarly, users of self-monitoring health apps scored higher (mean 68.20, SD 0.28) than those using sleep-monitoring apps (mean 55.35, SD 1.24; P<.001).	Intrinsic motivation increased	I[c]
Mojica et al [19]	App data	• Girls who owned a mobile phone were more likely than those who did not to engage in PA for more than 5 days in the past week (odds ratio 5.5, 95% CI 2.1-14) and to participate in daily physical education classes (odds ratio 2.6, 95% CI 1.1-5.9).	Enhanced PA	D[d]
Mendoza et al [21]	Measurable PA + questionnaire	• Exploratory analyses found no significant differences in MVPAe (4.4 vs 5.0 minutes/day; P=.92) or sedentary time (–4.5 vs 1.0 minutes/day; P=.73) between the intervention and control groups. Mean differences were reported for these outcomes, with P values indicating no significant effects. Some modest differences were observed for select subscales of quality of life and motivation for PA.	Intervention feasible/acceptable	I
Larsen et al [22]	Questionnaire + semistructured interview + accelerometer measurement	• According to physical activity recall measurements, participants' MVPA time increased to an average of 79.4 (SD 46.8) minutes at follow-up, representing a per capita increase of 58.8 (SD 46.3) minutes. This difference was statistically significant (P<.001). Overall, PA levels showed a significant improvement from baseline to 12 weeks (P<.001).	Enhanced PA	D
Thompson et al [23]	Measurable PA + questionnaire	• Program satisfaction: The mean score was 17.47/20, with no significant differences between groups (P=.068). The feasibility criterion—a program satisfaction score of ≥15 out of 20 points—was met. • Internal consistency: All measures exceeded the feasibility criterion ($\alpha \geq .70$), with the lowest α being .74 for social desirability. • PA trends: The all-intervention group showed the greatest increases in steps (+317.8 steps/day) and MVPA (+1.73 minutes/day) compared with other groups. However, no significant group-by-time interactions were observed (P>.05).	Enhanced PA	D
Guthrie et al [24]	Zamzee data	• Participants receiving the Zamzee intervention exhibited average MVPA levels 54% higher than the passive control group (P<.0001) and 68% higher than the active control group using commercially available active video games (P<.0001).	Enhanced PA compared with the control group	D
Direito et al [25]	Self-report + accelerometer measurement	• N/Af	The intervention will further validate its effectiveness.	I
Caillaud et al [27]	Accelerometer measurement + Physical Activity Questionnaire for Children + fitness test + iEngage app recording	• Participants increased their daily step count by 30% (+2647 steps/day, P<.001), and the proportion of days exceeding 11,000 steps rose to 48%.	Enhanced PA	D
Egilsson et al [28]	Scale + app data	• Self-reported daily PA increased by 19.61% in the intervention group but decreased by 26.21% in the control group. Self-efficacy levels increased by 8.23% in the intervention arm, compared with a 3.03% decrease in the control group.	Enhanced PA	D

XSL•FO
RenderX

Study	Measurement	Key findings	Result	Effect
Ortega and Cushing [31]	Accelerometer + questionnaire	• The temporally augmented goal setting intervention was generally feasible, acceptable, and technically and functionally reliable. Adolescents demonstrated adequate levels of engagement. Changes in MVPA from pre- to postintervention were small, approximately a 2-minute increase.	Enhanced PA	D
Willinger et al [32]	Semistructured interviews + questionnaires	• Most participants (n=16) believed that using KIJANI would increase their PA levels.	Enhanced PA	D
Garde et al [36]	PA monitor	• During the intervention week, both groups showed significant increases in PA levels (1191 and 796 steps/day, respectively; P=.01). • In the game group, children with higher BMI z scores exhibited greater PA levels, with an average increase of 964 steps/day for each unit increase in BMI z score (P=.03; 95% CI 98-1829).	Further validation is required.	U[g]
Mateo-Orcajada et al [37]	Questionnaire + physical fitness test + anthropometric measurements	• Compared with the control group, adolescents in the experimental group demonstrated higher levels of PA, along with improvements in body composition and physical fitness following the intervention.	Enhanced PA compared with the control group and improvement in health indicators	D
Cushing et al [38]	Questionnaire + accelerometer	• Compared with the attention control group, the intervention group spent an average of 20.84 minutes more per day in MVPA (β=20.84, SE 8.19).	Enhanced PA compared with the control group	D
Multimodal integrated intervention				

XSL•FO
RenderX

Study	Measurement	Key findings	Result	Effect
Soltero et al [15]	Interview	• Young adults reported increases in both PA and self-efficacy for PA.	Intervention feasible and acceptable	I
Goodyear et al [16]	Interview	• Daily goals of 10,000 steps and calorie burn motivate young people to engage in more PA.	Exercise goals encourage PA	I
Van Dyck et al [17]	Questionnaire + interview	• Participants indicated that SMS text messages provide a simple way to receive information about PA, while Facebook group pages are ideal for sharing information with others.	The intervention shows promise as a method for increasing PA	I
Ng et al [20]	Scale	• Compared with nonusers, app users had higher odds of engaging in daily MVPA (males: ORh 1.27, 95% CI 1.04-1.55; females: OR 1.49, 95% CI 1.20-1.85) and of being members of sports clubs (males: OR 1.37, 95% CI 1.15-1.62; females: OR 1.25, 95% CI 1.07-1.50). These associations were even stronger among wearable device users, with MVPA ORs up to 2.25. Male wearable users also had higher odds of engaging in active travel (OR 1.39, 95% CI 1.04-1.86).	Compared with the control group, the intervention showed a positive correlation with PA.	I
Chen et al [26]	Anthropometric measurements + questionnaire + Fitbit Flex data + interview	• Compared with participants in the control group, those in the mobile phone–based intervention showed significant improvements in BMI (z=–4.37, P=.001), diastolic blood pressure (z=–3.23, P=.001), PA days/week (z=2.58, P=.01), TV and computer time (z=–3.34, P=.001), servings of fruits and vegetables per day (z=2.74, P=.006), servings of soda and sweetened drinks (z=–3.19, P=.001), PA self-efficacy (z=2.75, P=.006), and dietary self-efficacy (z=5.05, P=.001). Medium to large effect sizes were observed for these outcomes.	Enhanced PA and increased self-efficacy.	D
Schoenfelder et al [33]	Quantitative measures + Fitbit Flex data + questionnaires + interview	• Participants significantly increased their average weekly step count and showed improvements in attention-deficit/hyperactivity disorder inattention symptoms.	Enhanced PA	D
Koorts et al [34]	Questionnaire + interview	• Adolescents perceived the intervention content as easy to understand (100/120, 83.3%) and the Fitbit as easy to use (112/120, 93.3%). Half of the adolescents found the SMS text messages useful (61/120, 50.8%), while 47.5% (57/120) liked the weekly challenges, and 38.3% (46/120) liked the Facebook videos. • Teenagers reported that the Fitbit Flex increased their awareness of PA and their likelihood of being more active in the short term.	Enhanced PA compared with the control group.	I
Interaction-enhanced intervention				

XSL•FO
RenderX

Study	Measurement	Key findings	Result	Effect
Staiano et al [29]	Anthropometry + full-body scanner + accelerometer + questionnaire	• The intervention group showed a significant reduction in BMI z score, with a marginal difference in the intent-to-treat analysis (mean −0.06, SE 0.03 for the intervention group vs mean 0.02, SE 0.03 for the control group, P=.065). Compared with the control group, the intervention group also showed improvements in systolic blood pressure, diastolic blood pressure, total cholesterol, low-density lipoprotein-cholesterol, and MVPA (all P<.05).	Enhanced PA and improvement in health indicators	D
Stasinaki et al [30]	Physical fitness test (Dordel-Koch test) + questionnaire + application data	• Both groups showed significant improvements in muscle mass, strength, and agility at T2, whereas only the PathMate2 group demonstrated a significant reduction in body fat at both T1 and T2. The average daily PathMate2 app usage rate was 71.5%.	Enhanced PA and improvement in health indicators	D
Glaser et al [35]	Questionnaire	• Compared with the control group, the intervention group experienced a 57% increase in PA.	Enhanced PA	D

[a]DHI: digital health intervention.

[b]PA: physical activity.

[c]DHIs are indirectly effective in promoting PA among adolescents (indirectly effective).

[d]DHIs are directly effective in promoting PA among adolescents (directly effective).

[e]MVPA: moderate-to-vigorous physical activity.

[f]N/A: not applicable.

[g]U: The effectiveness of DHIs in promoting PA among adolescents remains unclear (unclear).

[h]OR: odds ratio.

Study Risk-of-Bias Assessment

To ensure methodological rigor and assess internal validity, structured risk-of-bias assessments were conducted using JBI critical appraisal tools, each aligned with the corresponding study design. This included assessment checklists applicable to CSSs, RCTs, QESs, QR, and MMSs [13].

Risk of bias was assessed independently by 2 researchers using a 3-color coding system: green for “met,” yellow for “partially met,” and red for “not met.” Based on these assessments, each study’s methodological quality was categorized as high, moderate, or low. Disagreements were resolved through discussion and consensus, and if consensus was not reached, a third researcher adjudicated. Studies with reliability concerns—such as bias, unclear criteria, incomplete reporting, inappropriate methods, or insufficient justification—were flagged for further discussion and potential exclusion. No studies were excluded solely based on quality scores. Risk of bias was incorporated into the narrative synthesis to contextualize the strengths and limitations of the evidence. Detailed assessment results and scoring criteria are presented in Figures 1-4.

XSL•FO
RenderX

Figure 1. Quality assessment tool: cross-sectional studies. JBI: Joanna Briggs Institute.

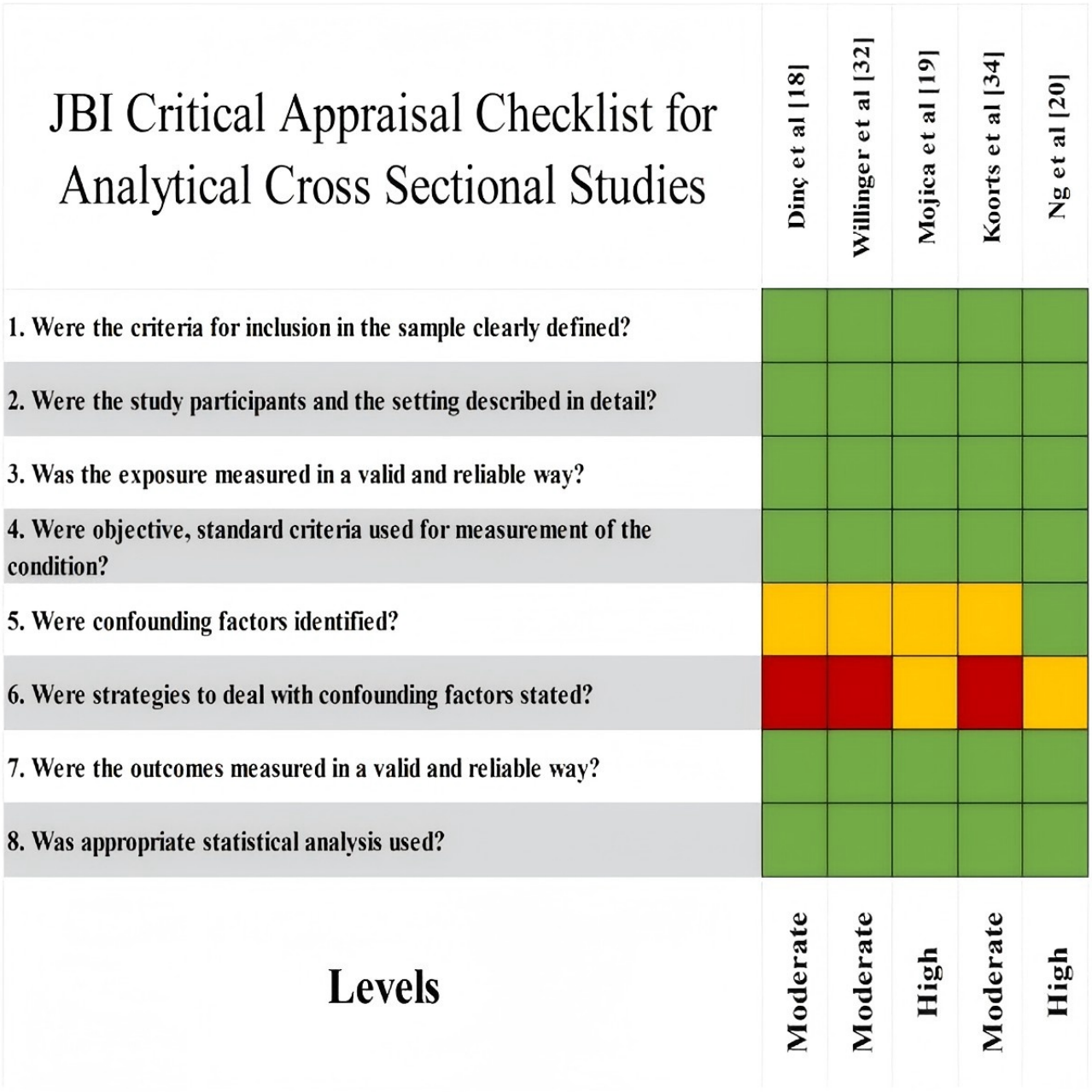

Effect Measures

Effect measures reported in the studies were diverse, encompassing both quantitative metrics and qualitative outcomes. Quantitative studies primarily presented intervention effects through comparisons of mean values or change values (eg, physical activity duration, step count, and their changes before and after intervention or between intervention and control groups), odds ratios, and regression model estimates, supplemented by statistical indicators such as CIs and *P* values to support conclusions. Some studies reported only the direction of change or statistical significance of outcomes without providing uniformly convertible effect sizes.

Additionally, some studies reported intervention effects through qualitative descriptions (eg, user experience, satisfaction, or perceived changes). Given the inconsistencies in effect measure formats and statistical reporting methods, this review did not uniformly convert or pool effect sizes. Instead, it focused on comprehensively describing the directional trends and overall performance of intervention effects.

Synthesis Methods

Given the substantial differences among included studies in research design, intervention modalities, outcome measurement, and statistical analysis methods, and considering that many studies did not report effect size parameters suitable for uniform conversion or pooling, this systematic review did not conduct a quantitative meta-analysis to avoid producing statistically unstable or potentially misleading pooled results. Prerequisites for meta-analysis include comparability or reasonable convertibility of outcome measurements and effect size definitions across studies. However, the included studies employed both self-reported and objectively measured outcomes,

XSL•FO
RenderX

with inconsistent measurement time points, making these conditions unattainable.

The research team therefore employed a narrative synthesis and qualitative content analysis approach to integrate the findings. During synthesis, a classification framework based on core forms of DHIs and their effect characteristics was developed to determine study applicability across synthesis units and to facilitate group comparisons. This classification and synthesis were performed independently by 2 researchers, with disagreements resolved through discussion and, when necessary, adjudicated by a third researcher. Findings were primarily presented using structured tables and typological narratives (Tables 1 and 2), with interpretations of outcome variations grounded in intervention formats, study designs, and implementation characteristics. As no statistical synthesis was conducted, this review did not perform formal heterogeneity testing or sensitivity analyses.

Reporting Bias Assessment

This systematic review employed a narrative synthesis and did not perform statistical tests for publication bias. Given the absence of a quantitative meta-analysis and the substantial heterogeneity in study design and outcome reporting across included studies, methods such as funnel plots were considered inapplicable. During evidence synthesis and result interpretation, the research team conducted a qualitative assessment of potential reporting bias. By comparing the consistency between study objectives, methods, and reported outcomes, and by incorporating study registration information (where available) and author explanations, the team cautiously discussed the potential impact of missing results on study conclusions.

Certainty Assessment

This systematic review conducted a qualitative synthesis of the overall credibility of the evidence, considering the design types of included studies, the results of the JBI methodological quality assessment, and the consistency of study conclusions across different studies. Given that no quantitative meta-analysis was performed and the included studies exhibited substantial heterogeneity in design types and outcome measures, quantitative evidence grading tools such as GRADE (Grading of Recommendations, Assessment, Development and Evaluation) were not applied. The assessment of evidence certainty was conducted independently by 2 researchers (RSF and JJJ). The reliability and limitations of the evidence were interpreted in the “Results” and “Discussion” sections in conjunction with specific study findings.

Ethical Approval

Ethics approval was not required for this systematic review.

Results

Study Selection

A total of 2127 English-language articles were initially retrieved from 7 major databases, including Web of Science (n=601), PubMed (n=518), EBSCO (n=50), Scopus (n=305), Embase (n=164), Cochrane Library (n=128), ProQuest (n=78), and Google Scholar (n=283). After initial deduplication using EndNote, 1288 duplicate records were removed. Subsequently, an additional 164 duplicates were removed using the Rayyan online literature management platform, resulting in 675 records retained for title and abstract screening. Following this initial screening, 213 articles remained for full-text assessment. Of these, 189 articles were excluded based on predefined exclusion criteria, including study population, intervention characteristics, and research design. The specific reasons for exclusion were as follows: studies not focused on adolescents (n=36), absence of DHIs (n=14), insufficient clarity in intervention or outcome measures (n=21), conference abstracts (n=8), studies focused solely on instrument validation (n=49), review articles (n=35), and studies with unrelated topics (n=26). Ultimately, 24 English-language studies met the inclusion criteria and were included in the final qualitative synthesis of this systematic review (see Figure 5 and Multimedia Appendix 2).

XSL•FO
RenderX

Figure 5. PRISMA Flow Diagram of the Study Selection Process.

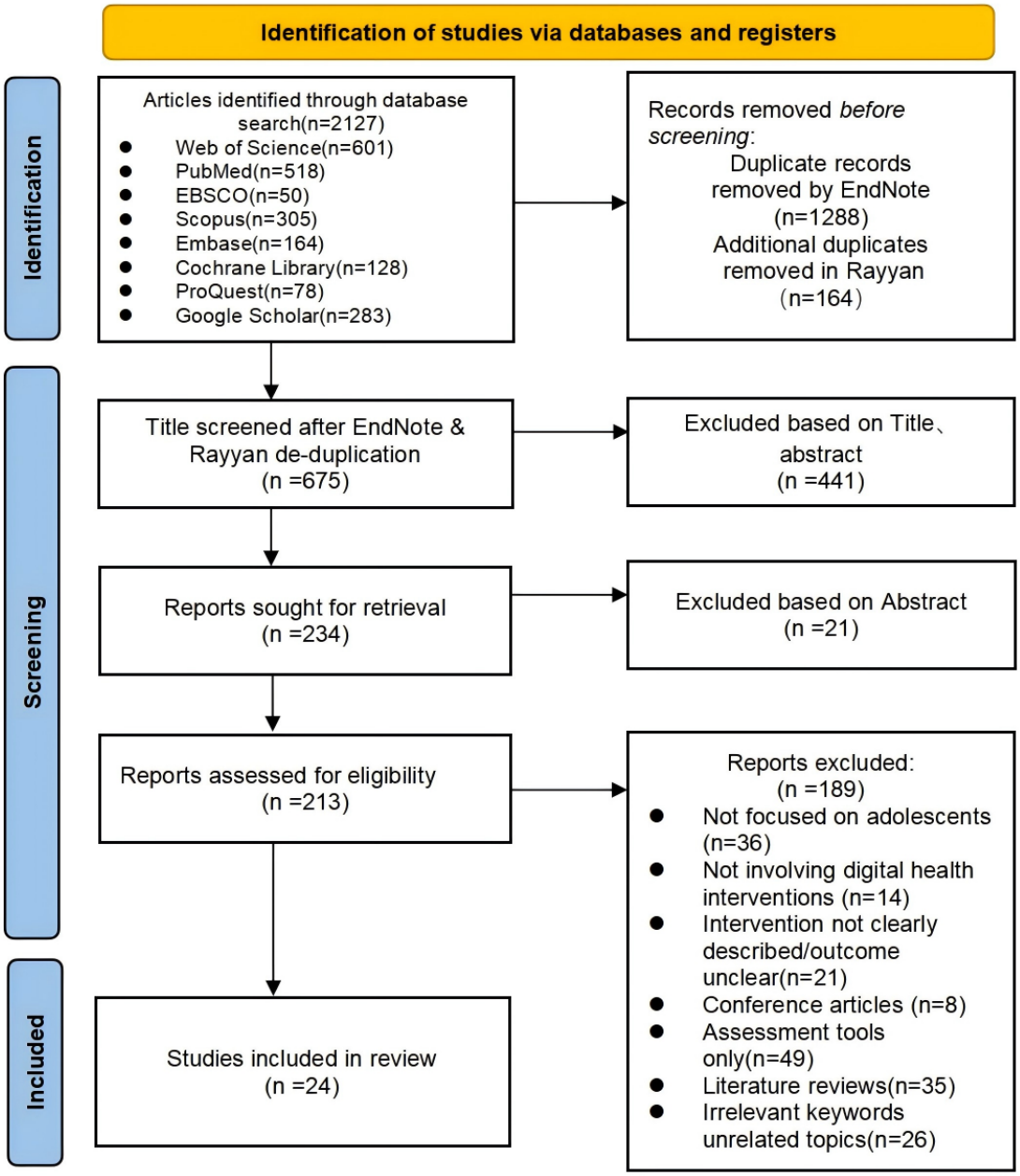

Study Characteristics

Overview

This systematic review included 24 peer-reviewed articles published between 2014 and 2025, spanning multiple countries and regions, thereby reflecting the global uptake of DHIs among adolescents. The included studies employed a range of methodological designs, including RCTs, MMS, QR, and other empirical approaches. All interventions targeted adolescents aged 10-19 years, including the general population, individuals with insufficient physical activity, and those with specific health conditions. Sample sizes ranged from 10 to 9940 participants, with the majority of studies involving fewer than 500 individuals. Intervention duration ranged from 2 weeks to 8 months. The included studies employed diverse intervention formats, utilizing either single-modality or multicomponent digital technology approaches. Detailed study characteristics are summarized in Table 1.

Participant Characteristics

This systematic review encompassed approximately 12,183 adolescent participants. Sample sizes varied widely, ranging from 10 to 9,940 participants. The overall mean sample size was approximately 507, whereas the median was substantially lower at 46. Most studies employed small-to-medium sample sizes, with a few large-scale studies exerting a disproportionate influence on the overall mean. Among studies that reported age data, participants' age ranged from 8 to 19 years, with the highest concentration in the 13-15 age group, representing over one-third of the total sample. The study populations included both general adolescent cohorts and specific subgroups characterized by distinct health risks or behavioral profiles. Of

XSL•FO
RenderX

the total 24 studies, 8 (33%) explicitly targeted specific subgroups, including 4 (17%) on adolescents with obesity or overweight, 1 (4%) on adolescents with insufficient physical activity, 1 (4%) on individuals with attention-deficit/hyperactivity disorder, 1 (4%) on cancer survivors, and 1 (4%) on at-risk adolescents. Additional details are provided in Multimedia Appendix 3.

Geographic Distribution

The 24 included studies were conducted across multiple countries and regions. North America accounted for the highest concentration, contributing 11 (46%) studies of the total, significantly exceeding other regions. Europe followed with 8 (33%) studies, while the Middle East and Oceania contributed 2 (8%) and 3 (13%) studies, respectively. At the country level, the United States emerged as the leading research site, accounting for 10 (42%) studies—the highest among all countries. Spain and Australia each contributed 2 studies. Other countries—namely, Switzerland, Turkey, Germany, Israel, Belgium, New Zealand, Iceland, Canada, the United Kingdom, and Finland—each contributed 1 study. Regional variations were also observed in the selection and application of intervention technologies. North American countries (eg, the United States and Canada) generally favored the combination of wearable devices and mobile apps, providing integrated data tracking and personalized feedback. By contrast, several European countries primarily integrated traditional intervention models with digital platforms, including web-based systems and educational software. Although studies from Middle Eastern and Oceanic countries featured smaller sample sizes, their intervention formats were diverse, incorporating tools such as SMS text messaging and interactive platforms (see Multimedia Appendix 3).

Intervention Duration and Frequency

Considerable variation was observed in the duration of interventions. The shortest intervention lasted 1 week (excluding 3 CSSs), while the longest extended to 8 months. The most commonly adopted duration was 12 weeks (3 months), reported in 4 (17%) studies, followed by durations of 6, 8, and 10 weeks, each reported in 2 (8%) studies. Other durations included 1 week, 2 weeks, and 8 months, each reported in a single study (4%). Overall, approximately one-third of the studies fell within the 6- to 12-week range. Based on duration, short-term interventions (≤4 weeks) were reported in 5 (21%) studies; medium-term interventions (5-12 weeks) were the most prevalent, appearing in 11 (46%) studies; and long-term interventions (>12 weeks) were implemented in 5 (21%) studies. Long-term interventions were predominantly associated with high-quality RCTs, whereas short-term interventions were distributed across a variety of study designs. In summary, the majority of studies employed medium-term interventions, typically lasting 2-3 months.

Risk of Bias in Studies

Among the 24 included studies, research designs comprised 10 (42%) RCTs, 4 (17%) QES, 3 (13%) QR, 3 (13%) CSS, and 4 (17%) MMS. Quality appraisal results indicated that 7 (29%) studies were rated as high quality, 16 (67%) as moderate quality, and 1 (4%) as low quality. Further analysis revealed that most RCTs achieved high or moderate quality ratings, while QES, CSS, and MMS studies generally met moderate or higher quality standards. Notably, QR studies demonstrated outstanding performance, with 2 projects achieving high quality (see Figure 2).

XSL•FO
RenderX

Figure 2. Quality assessment tool: qualitative research. JBI: Joanna Briggs Institute.

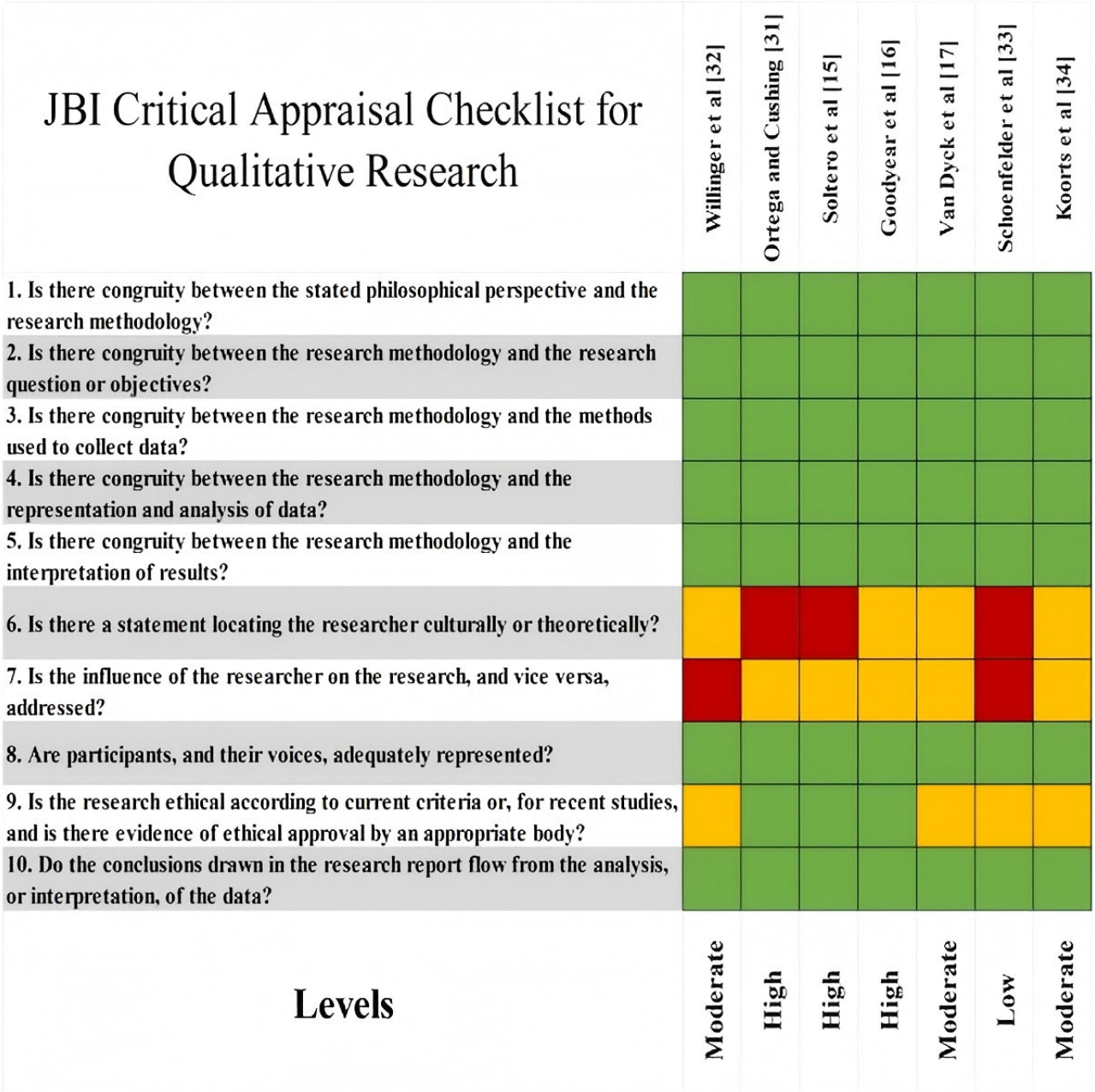

XSL•FO
RenderX

Figure 3. Quality assessment tool: randomized controlled trials. JBI: Joanna Briggs Institute.

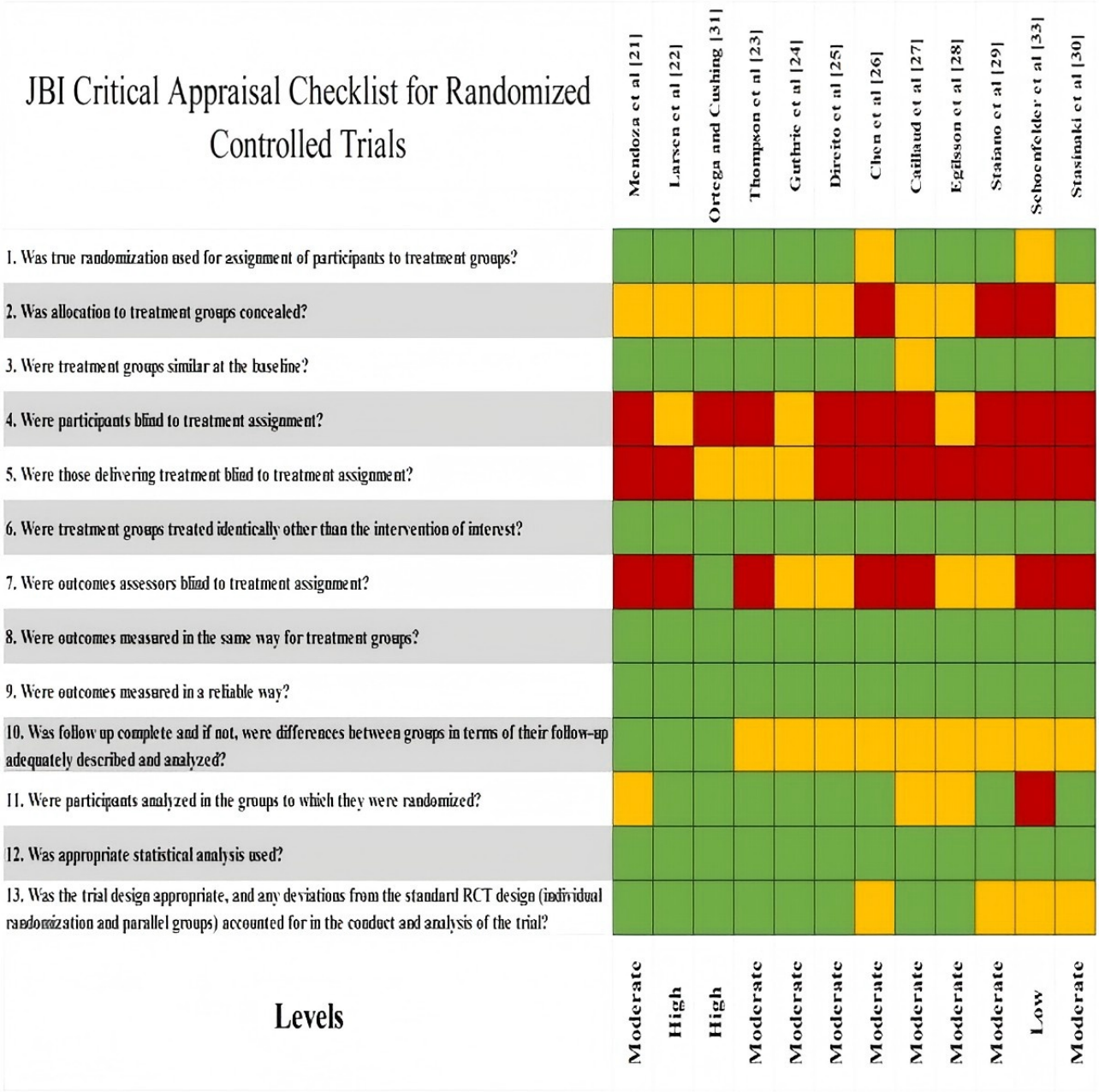

Classification and Outcomes of DHIs

The classification and corresponding outcomes of DHIs are presented in Table 2. The types of interventions employed across the included studies exhibited considerable diversity.

These interventions were broadly categorized into 3 types: single-driver interventions, multimodal integrated interventions, and interaction-enhanced interventions. Based on the combined findings from Tables 1 and 2, single-driver interventions were the most commonly reported, appearing in 14 (58%) studies. Interventions in this category typically relied on a single technological medium, most commonly smartphones as the primary device. mHealth interventions were the most prevalent, followed by virtual platform–based interventions and robotic systems delivering personalized SMS text messages. Multimodal integrated interventions were identified in 7 (29%) studies. These interventions combined 2 or more digital components, including mHealth apps, wearable devices, and web-based platforms. Interaction-enhanced interventions were relatively limited, reported in only 3 (13%) studies. Characterized by active adolescent engagement—such as promoting physical activity through video chats—these interventions primarily involved gamified strategies, social media platforms, and SMS text message–based messaging systems.

Overall, all 3 categories of DHIs demonstrated positive outcomes. As shown in Table 2, the majority of studies were classified as “D” (directly effective; 15/24, 63%), while smaller proportions were categorized as “I” (indirectly effective; 8/24, 33%) or “U” (unclear; 1/24, 4%).

Among the “D” interventions, most studies reported statistically significant improvements in physical activity outcomes [19,22-24,26-33,35,37,38], including increased step counts, longer durations of moderate-to-vigorous physical activity, and

XSL•FO
RenderX

greater participation in sports. Several studies also reported improvements in physical health indicators related to physical activity [15-18,20,21,25,34], including reductions in BMI, improved body composition and physical fitness, and decreased body fat percentage. Studies classified as "I" did not report direct increases in physical activity levels but demonstrated positive effects on psychological mechanisms, such as enhanced exercise motivation and self-efficacy. These effects included strengthened beliefs in healthy lifestyles, the establishment of exercise goals that promote physical activity, and increased awareness of the benefits of physical activity [36]. Studies categorized as "U" lacked conclusive evidence, making it difficult to determine the effectiveness of the DHIs under investigation.

Results of Syntheses

A comprehensive analysis of the included studies was conducted based on the form of DHIs and their effectiveness characteristics. The synthesis identified significant variation in research designs (eg, QR, CSS, RCT, QES, MMS), sample sizes (10-9940 participants), outcome measurement methods, and methodological quality. Overall study quality was predominantly moderate to high, with most interventions lasting short to medium durations. As a result of substantial heterogeneity in study design, outcome definitions, and statistical analysis methods, as well as the absence of standardized effect size parameters for pooling across studies, a quantitative meta-analysis was not performed. Instead, results were synthesized using narrative synthesis and qualitative content analysis based on intervention type. The findings are presented in structured tables and categorized descriptions (Tables 1 and 2) to compare the effectiveness of different intervention formats and study designs in promoting physical activity among adolescents.

Reporting Biases and Certainty of Evidence

As this systematic review did not perform a quantitative meta-analysis and due to considerable variation in study designs and outcome reporting methods, no formal testing for publication bias was conducted using funnel plots or statistical methods. Instead, a qualitative assessment of potential reporting bias was undertaken during evidence synthesis and interpretation. Evidence certainty was not formally assessed using a quantitative tool; rather, the credibility of the evidence was judged by integrating study design type, methodological quality assessment results, and the consistency of study conclusions. Overall, the included studies were of moderate-to-high quality (Figures 1-4). However, owing to study heterogeneity and methodological limitations, the conclusions should be interpreted within the context of specific research settings.

XSL•FO
RenderX

Figure 4. Quality assessment tool: quasi-experimental studies. JBI: Joanna Briggs Institute.

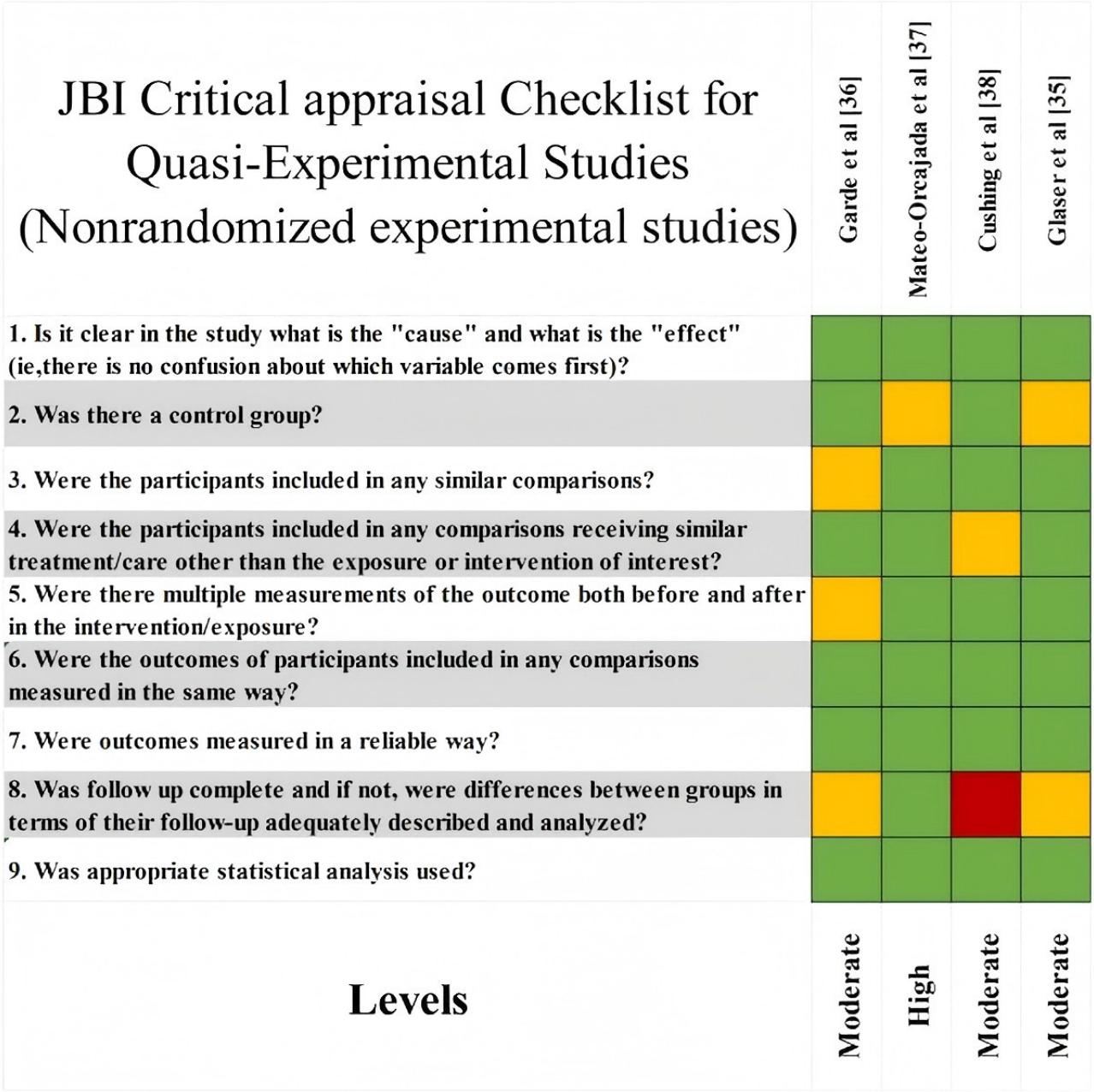

Discussion

Principal Findings

This systematic review synthesizes evidence from diverse DHIs and study designs to evaluate their role in promoting physical activity among adolescents and to explore their adaptability across different research designs and settings. Results from the 24 included studies indicate that DHIs can enhance adolescents' physical activity through direct or indirect mechanisms. Substantial heterogeneity was observed among the studies in terms of research design types, resulting in inconsistent evidence strength and limited generalizability; this heterogeneity represents a key finding of the review. A systematic analysis of intervention formats further revealed that different types of DHIs (single-driver, multimodal integrated, and interaction-enhanced) exhibit distinct effect profiles and are applicable in different contexts. In addition, factors such as target population characteristics and intervention duration significantly influence the interpretation and application of outcomes. Accordingly, this review emphasizes the need for cautious interpretation of research findings, with careful consideration of the scope of inference supported by the study designs and implementation contexts.

This review included 3 QR and 3 CSS. As shown in Tables 1 and 2, the evidence derived from these study types differs. QR evidence reflects the meaning or experience of behaviors or events rather than quantitative measurements [39]. For example, Soltero et al [15] used qualitative methods to assess the feasibility and acceptability of DHIs, identifying potential pathways for promoting physical activity among adolescents with obesity. CSSs measure exposure and outcomes at a single time point, revealing correlations between DHI use, individual characteristics, and physical activity levels [40]. For example,

XSL•FO
RenderX

Ng et al [20] examined the association between ownership and use of personal activity trackers and physical activity behaviors. Consequently, this evidence is better suited to suggesting potential mechanisms and real-world usage scenarios rather than to drawing strong causal conclusions.

This review observes the indirect effectiveness of DHIs on adolescent physical activity within QR and CSS evidence. Although such interventions may not immediately lead to significant behavioral changes, these study types more readily capture positive psychological shifts, such as enhanced motivation, increased behavioral intent, strengthened self-efficacy, improved health cognition, or greater interest in participation [41]. Compared with RCTs, QR emphasizes contextual experiences and subjective perceptions, making it more effective in identifying positive psychological shifts, such as heightened motivation. This provides crucial insights into the indirect mechanisms of intervention. CSSs excel at revealing structural relationships between psychological or behavioral characteristics and DHI use at a single time point, offering key clues for understanding these indirect mechanisms.

According to self-determination theory [42], the persistence of adolescent behaviors depends not only on external reminders or incentives but also on the fulfillment of fundamental psychological needs, such as autonomy, competence, and relatedness. Among adolescents in the early stages of behavior change or with unstable motivation, such psychological improvements are particularly important [43] and may gradually translate into behavioral changes through positive experiences, goal identification, and social support [44]. This mechanism of psychological-behavioral conversion helps explain the delayed behavioral improvements observed in some follow-up studies; however, further longitudinal research is needed to validate these findings.

While this evidence provides valuable insights into underlying mechanisms and the overall methodological quality is high, risks of bias remain. Some QR studies did not adequately disclose the researcher's stance [15-17], and some CSSs did not sufficiently control for confounding factors [18,20]. These issues may compromise the robustness of the findings and therefore warrant cautious interpretation.

QR studies and CSSs have distinct characteristics. The former illuminates indirect effects and underlying mechanisms of DHIs, whereas the latter reveals associations among variables but cannot establish causality. Future research should prioritize high-quality study designs to validate the efficacy and long-term effects of DHIs.

This systematic review included 10 RCTs, which provided evidence with varying focuses. RCTs are considered the gold standard for evaluating intervention effectiveness, as their rigorous design allows for systematic assessment of intervention outcomes over time [45]. The RCTs included in this review involved both general adolescent populations and adolescents with specific health conditions, primarily aiming to validate the feasibility or efficacy of DHIs for these groups [21].

Collectively, these studies indicate that DHIs have potential for promoting physical activity. For populations with specific health conditions, effectiveness often depends more on personalized and sustained support than on technological features [46-48]. Although subgroups share certain commonalities, they also differ in their intervention needs. Commonalities include challenges in behavior change, strong demands for psychological and emotional support, and the need for long-term follow-up and sustained support [10,32,49-53]. Differences are reflected in personalized preferences. Adolescents with attention-deficit/hyperactivity disorder may require more structured and highly interactive tools, whereas cancer survivors may prioritize privacy and self-paced control. Adolescents with obesity or overweight issues may benefit from integrated solutions that combine physical activity, dietary adjustments, and psychological support [54-56]. These differences arise from variations in behavioral mechanisms, cognitive characteristics, and motivational responses across populations. Even with robust technology, failure to align interventions with users' cognition, motivation, and behavior can hinder effective engagement and behavior change. Future research should prioritize methodologically rigorous RCTs and tailored designs that match population characteristics with intervention protocols to enhance evidence interpretability and real-world applicability.

Of the follow-up studies included in this systematic review, 3 were RCTs [22,23,26]. The findings suggest that the presence or absence of follow-up may introduce bias in assessing DHI outcomes. Follow-up is crucial for evaluating the sustainability and long-term impact of interventions, as extended follow-up periods help identify both enduring effects and potential side effects [57]. For example, Larsen et al [22] reported that a 12-week follow-up showed the intervention promoted diverse physical activities. Similarly, Thompson et al [23] found that personalized SMS text message interventions helped adolescents establish long-term exercise habits, with follow-up feedback further optimizing intervention effectiveness.

In this systematic review, studies with follow-up periods allow assessment of the sustainability of intervention effects, whereas studies without follow-up can evaluate only short-term outcomes. Evidence suggests that a 12-week intervention can yield favorable results for adolescents [23], and such feedback helps refine intervention strategies. RCTs provide relatively strong evidence for this review; however, the limited number of studies primarily supports short-term effects and does not adequately validate intervention sustainability or stability. Consequently, robust evidence for the long-term sustainability of DHIs in real-world settings, such as schools and communities, remains insufficient, which may hinder practical implementation and policy support.

The included RCTs indicate that a 3-month (12-week) intervention duration is the most common. This cycle facilitates gradual behavior establishment and consolidation while minimizing negative effects such as fatigue, a finding widely recognized [23]. Research suggests that shorter intervention durations are insufficient to produce significant changes, whereas excessively long durations may increase complexity and reduce adherence [58]. However, no studies specifically examined the impact of varying intervention durations on adolescent populations.

XSL•FO
RenderX

The impact of sample attrition on outcomes cannot be overlooked. Three of the included studies reporting attrition were RCTs [23,28,30]. Our analysis found that attrition typically occurs in studies requiring sustained participation and long-term engagement, making it a critical factor influencing outcomes. Studies with high attrition rates often face challenges such as insufficient intervention personalization, limited feedback, and inadequate technical support, all of which can weaken participants' motivation and commitment.

From a behavioral science perspective, based on the COM-B framework, if interventions fail to sustainably enhance individuals' sense of competence, adolescents are prone to burnout and disengagement. Egilsson et al [28] confirmed this, suggesting that high attrition rates during 6-week follow-ups may be related to insufficient intervention personalization, monotonous formats, or technical barriers. Conversely, other studies found that lower attrition rates are closely associated with adolescent cocreation of intervention content [23]. Thus, enhancing intervention personalization, diversity, and interactivity can reduce attrition, thereby improving research stability and the generalizability of findings.

As shown in Table 2, among the studies demonstrating direct effectiveness, 8 were RCTs. RCTs minimize confounding bias through randomization, which helps clarify intervention effects. Analysis indicates that these interventions promote short-term behavioral changes (eg, increased daily step counts) through goal setting, task execution, and feedback design [59]. This effect depends on clear goals and immediate feedback, enabling adolescents to master behavioral cognition through tracking or outcome comparisons [60]. Such interventions drive strong short-term behavioral change, particularly among participants with initial motivation. The core mechanism enhances individual competence and execution capacity, fosters a sense of accomplishment, and lays the foundation for long-term behavioral strategies.

In summary, RCT designs provided key evidence for this review, although some biases remain. Many studies, due to the nature of DHIs, lacked blinding and faced challenges in follow-up and assessment, which may have affected the reliability of their findings. Future research should employ high-quality RCTs with robust follow-up designs to assess long-term effects and address sample attrition. Given that over half of existing studies used RCTs to evaluate direct effects, future studies should select study designs that align with specific objectives and research contexts.

This systematic review included 4 MMS and 4 QES. MMS combines quantitative and qualitative approaches, enabling simultaneous evaluation of effects and underlying mechanisms, whereas QES assesses impacts but is prone to selection bias due to the absence of randomization or control groups [61]. Given the limited number of MMS and QES, the evidence is scattered. These study types are suitable for identifying feasible strategies and potential mechanisms in real-world contexts, but should not be overgeneralized into universal conclusions. For example, one study from each design type explored the effectiveness and feasibility of DHIs for promoting physical activity in adolescents with specific health conditions. As shown in Table 2, the majority of studies within both research designs demonstrated the direct effectiveness of DHIs, supporting the applicability of these approaches. However, the overall quality was moderate, likely due to small sample sizes and selection bias. Given the limited number of studies, the evidence lacks representativeness and should be interpreted with caution.

Beyond the evidence provided by different study designs, we identified additional relevant findings. Compared with previous systematic reviews [9-11], this review included multiple types of DHIs and categorized them thematically. Results indicate that most studies employed single-driver interventions, while a minority utilized multimodal integrated or interaction-enhanced interventions. Each intervention type exhibits distinct characteristics, tailored to different target populations and intervention phases (see Multimedia Appendix 3 and Table 2).

First, single-driver interventions, such as smartphone apps, promote behavior change through goal setting and reminders and are particularly effective during the initiation phase, leading to significant increases in daily step counts [10,62]. However, their effects are limited by the lack of continuous feedback, restricting their applicability to short-term behavior promotion [63]. Second, multimodal integrated interventions combine wearable devices with mHealth technologies. These interventions are well-suited for the long-term management of high-risk populations, consistently increasing activity frequency and reducing sedentary behavior [64,65]. Nevertheless, their complexity and user burden may hinder acceptance and adherence [66], highlighting the need for simplified workflows to enhance participation.

Third, interaction-enhanced interventions leverage gamification and social interaction to stimulate adolescents' emotional and social motivation, thereby increasing participation enthusiasm and self-efficacy [67]. Their strength lies in enhancing behavioral appeal and fostering sustained engagement, making them potent drivers of short-term behavioral change. The core mechanism involves strengthening individuals' sense of competence and execution capacity, enabling them to experience accomplishment and receive positive feedback upon goal attainment, thereby laying the foundation for long-term behavioral strategies. However, effectiveness may vary among individuals with weaker social motivation or limited digital access [32,49]. Current research predominantly focuses on single-driver interventions due to their operational simplicity. Nonetheless, distinct characteristics and limitations emerge across intervention types when applied to diverse adolescent groups and developmental stages. Consequently, the selection of intervention type can significantly influence outcomes, particularly for specific populations and developmental phases.

On a broader scale, this systematic review highlights 2 critical real-world challenges in DHIs. First, these interventions reflect regional disparities and health inequalities [68]. North America, particularly the United States, leads in research and technological applications, emphasizing personalized feedback and data tracking [69,70]. In Europe, although fewer studies are conducted, there is a focus on educational and universal approaches tailored to diverse sociocultural contexts [71].

XSL•FO
RenderX

Meanwhile, the Middle East and Oceania explore a variety of technological pathways within the constraints of limited resources [72].

However, these regional disparities not only reflect differences in technological adoption but also highlight digital health inequalities across domains. Digital health disparities are becoming increasingly prominent worldwide, particularly in low- and middle-income countries [73], where significant gaps persist in the accessibility and effectiveness of DHIs. These disparities are largely influenced by factors such as infrastructure development and digital literacy [74]. Therefore, in promoting global DHIs, greater focus must be placed on addressing these regional inequalities. Measures should be taken to allocate resources appropriately and enhance technical support to ensure the inclusivity and equity of technology.

Second, the technological dependency of DHIs is both a crucial factor influencing outcomes and a determinant of long-term adherence. When interventions rely on digital platforms, apps, or wearable devices, participants' sustained engagement may be disrupted by technical issues, thereby affecting intervention stability and the assessment of outcomes [75-77].

Mechanistically, the effectiveness of system design in technological interventions largely determines participants' long-term adherence, as supported by previous studies. For example, Jakob et al [78] found that sustained use of mHealth apps is influenced by design factors such as technical stability and personalized push notifications. Similarly, Kelders et al [79] demonstrated that digital interventions incorporating persuasive design significantly enhance user adherence. Therefore, addressing the negative impacts of technological dependency represents a critical future development pathway for DHIs.

Limitations

This systematic review has several key limitations. First, the included studies exhibited considerable heterogeneity in design type, intervention format, feedback mechanisms, duration, and behavioral assessment methods. This heterogeneity limited evidence consistency and hindered the integration and generalizability of the findings. Second, many original studies had brief evaluation periods and lacked medium- to long-term follow-up data, complicating the assessment of intervention sustainability and long-term impact. Additionally, some studies had methodological limitations and insufficient reporting transparency. Although most studies were of moderate to high quality, these issues compromised the reliability and generalizability of the evidence. Finally, despite conducting multidatabase systematic searches, relevant literature may have been omitted due to limitations in search coverage and potential publication bias.

Therefore, future research can be enhanced in several areas. First, the study design should be strengthened to elevate overall evidence quality. Conducting more high-quality RCTs would improve the scientific rigor and feasibility of intervention studies. Standardization of intervention content and behavioral indicators should be promoted, establishing a unified assessment system to enhance comparability and integration of results across studies. Simultaneously, methodological design and reporting transparency should be improved to provide a more reliable evidence base for systematic reviews.

Second, future research should integrate discussions of intervention mechanisms and effects into study methodologies. The subjective experiences and underlying mechanisms of DHIs hold significant research value. Combining subjective experience data with objective behavioral data can generate more comprehensive and representative evidence. A deeper understanding of the transformation processes underlying these mechanisms will provide a foundation for developing personalized intervention models. Finally, continuous research should address practical issues, such as strengthening follow-up. Study designs should incorporate more follow-up components to systematically document postintervention behavioral maintenance and motivational changes. Concurrently, attention should be given to populations with specific health conditions, digital health inequalities, and technology dependency challenges, to enhance the interpretive power and practical applicability of research in real-world contexts.

Conclusions

Against the backdrop of increasing physical inactivity and sedentary behavior among adolescents globally, this systematic review synthesizes evidence from 24 studies encompassing diverse research designs and multiple types of DHIs. Compared with previous systematic reviews that focused exclusively on specific study designs or single technological formats, this review provides a more holistic and comparative perspective on emerging evidence. The integrated findings indicate that DHIs generally contribute to enhancing adolescents' physical activity levels. However, their effectiveness varies considerably across intervention types, research designs, and application contexts. This finding addresses gaps in existing research, suggesting that both the adaptability of interventions and the implementation setting play crucial roles in shaping outcomes. Distinct strengths were observed across different research designs: RCTs provided relatively reliable evidence for direct and long-term intervention effects, while QR and CSS contributed to understanding underlying mechanisms and associations. QES and some MMS offered relatively limited representativeness due to scattered evidence. This finding supplements and refines the existing understanding of the evidence structure for DHIs. Further analysis indicates that the suitability of different DHIs varies across adolescent populations and practice settings, with the choice of intervention type significantly influencing outcomes. Concurrent attention to adolescents with specific health conditions, digital health inequities, and technology dependency challenges provides new perspectives for equitable implementation and sustainable scaling of DHIs in real-world settings. Despite limitations in the existing evidence—including methodological bias, inconsistent study quality, and insufficient follow-up data—the findings of this review remain valuable for understanding the role of DHIs in promoting adolescent physical activity. They provide useful insights for practical application, global health promotion, health equity, and innovation in cross-cultural intervention strategies. Future research should integrate the strengths of diverse study designs to conduct high-quality,

XSL•FO
RenderX

long-term follow-up studies. Such studies should comprehensively consider target population characteristics, intervention duration, and regional variations to enhance the effectiveness and scalability of DHIs in real-world settings.

Acknowledgments

We are also grateful to all participants in this systematic review for their time, effort, and dedication. The authors declare that no generative artificial intelligence was used in the writing of this manuscript.

Funding

The authors received no financial support for the research, authorship, or publication of this article.

Data Availability

All data generated or analyzed during this systematic review are included in this published article and its multimedia appendices. Additional information is available from the corresponding author (LJ) on reasonable request.

Authors' Contributions

Conceptualization: RSF (lead), JJJ (equal), LJ (supporting)
Data curation: QYZ (lead), XYZ (supporting)
Formal analysis: RSF
Funding acquisition: LJ
Investigation: RSF (lead), JJJ (supporting)
Methodology: RSF (lead), LJ (equal), QYZ (supporting)
Project administration: RSF (lead), XYZ (supporting)
Resources: RSF (lead), JJJ (supporting), ZHW (supporting)
Supervision: LJ (lead), XYZ (supporting)
Validation: RSF (lead), QYZ (supporting), XYZ (supporting)
Visualization: RSF (lead), JJJ (supporting), QYZ (supporting)
Writing – original draft: RSF (lead), JJJ (supporting), QYZ (supporting)
Writing – review and editing: RSF (lead), JJJ (supporting), XYZ (supporting), LJ (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 28 KB - jmir_v28i1e82395_app1.docx]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 28 KB - jmir_v28i1e82395_app2.docx]

Multimedia Appendix 3
Additional analysis.
[DOCX File , 516 KB - jmir_v28i1e82395_app3.docx]

References

1. Chaput J, Willumsen J, Bull F, Chou R, Ekelund U, Firth J, et al. 2020 WHO guidelines on physical activity and sedentary behaviour for children and adolescents aged 5-17 years: summary of the evidence. Int J Behav Nutr Phys Act 2020 Nov 26;17(1):141 [FREE Full text] [doi: 10.1186/s12966-020-01037-z] [Medline: 33239009]
2. Rose T, Barker M, Maria Jacob C, Morrison L, Lawrence W, Strömmer S, et al. A systematic review of digital interventions for improving the diet and physical activity behaviors of adolescents. J Adolesc Health 2017 Dec;61(6):669-677 [FREE Full text] [doi: 10.1016/j.jadohealth.2017.05.024] [Medline: 28822682]
3. Sampasa-Kanyinga H, Colman I, Goldfield GS, Janssen I, Wang J, Podinic I, et al. Combinations of physical activity, sedentary time, and sleep duration and their associations with depressive symptoms and other mental health problems in

XSL•FO
RenderX

children and adolescents: a systematic review. Int J Behav Nutr Phys Act 2020 Jun 05;17(1):72 [FREE Full text] [doi: 10.1186/s12966-020-00976-x] [Medline: 32503638]
4. Düking P, Tafler M, Wallmann-Sperlich B, Sperlich B, Kleih S. Behavior change techniques in wrist-worn wearables to promote physical activity: content analysis. JMIR Mhealth Uhealth 2020 Nov 19;8(11):e20820 [FREE Full text] [doi: 10.2196/20820] [Medline: 33211023]
5. Au WW, Recchia F, Fong DY, Wong SHS, Chan DKC, Capio CM, et al. Effect of wearable activity trackers on physical activity in children and adolescents: a systematic review and meta-analysis. Lancet Digit Health 2024 Sep;6(9):e625-e639 [FREE Full text] [doi: 10.1016/S2589-7500(24)00139-0] [Medline: 39112110]
6. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
7. Collin C, Eyraud C, Martin P, Michel M, Le Roux E, Alberti C. A scoping review of outcome selection and accuracy of conclusions in complex digital health interventions for young people (2017-2023): methodological proposals for population health intervention research. BMC Med 2025 Jul 02;23(1):400 [FREE Full text] [doi: 10.1186/s12916-025-04245-1] [Medline: 40604855]
8. Partridge SR, Knight A, Todd A, McGill B, Wardak S, Alston L, et al. Addressing disparities: a systematic review of digital health equity for adolescent obesity prevention and management interventions. Obes Rev 2024 Dec;25(12):e13821. [doi: 10.1111/obr.13821] [Medline: 39256958]
9. Wang J, Zhu Z, Shuling Z, Fan J, Jin Y, Gao Z, et al. Effectiveness of mHealth app-based interventions for increasing physical activity and improving physical fitness in children and adolescents: systematic review and meta-analysis. JMIR Mhealth Uhealth 2024 Apr 30;12:e51478 [FREE Full text] [doi: 10.2196/51478] [Medline: 38687568]
10. He Z, Wu H, Yu F, Fu J, Sun S, Huang T, et al. Effects of smartphone-based interventions on physical activity in children and adolescents: systematic review and meta-analysis. JMIR Mhealth Uhealth 2021 Feb 01;9(2):e22601 [FREE Full text] [doi: 10.2196/22601] [Medline: 33522980]
11. Sequí-Domínguez I, Cavero-Redondo I, Álvarez-Bueno C, López-Gil JF, Martínez-Vizcaíno V, Pascual-Morena C. Effectiveness of eHealth interventions promoting physical activity in children and adolescents: systematic review and meta-analysis. J Med Internet Res 2024 Feb 21;26:e41649 [FREE Full text] [doi: 10.2196/41649] [Medline: 38381490]
12. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Rev Esp Cardiol (Engl Ed) 2021 Sep;74(9):790-799 [FREE Full text] [doi: 10.1016/j.rec.2021.07.010] [Medline: 34446261]
13. Barker T, Stone J, Sears K, Klugar M, Tufanaru C, Leonardi-Bee J, et al. The revised JBI critical appraisal tool for the assessment of risk of bias for randomized controlled trials. JBI Evid Synth 2023 Mar 01;21(3):494-506. [doi: 10.11124/JBIES-22-00430] [Medline: 36727247]
14. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
15. Soltero EG, Musaad SM, O'Connor TM, Thompson D, Norris K, Beech BM. Feasibility of Fit24, a digital diabetes prevention program for Hispanic adolescents: qualitative evaluation study. JMIR Form Res 2024 May 17;8:e54595 [FREE Full text] [doi: 10.2196/54595] [Medline: 38758584]
16. Goodyear VA, Kerner C, Quennerstedt M. Young people's uses of wearable healthy lifestyle technologies; surveillance, self-surveillance and resistance. Sport, Education and Society 2017 Sep 22;24(3):212-225. [doi: 10.1080/13573322.2017.1375907]
17. Van Dyck D, D'Haese S, Plaete J, De Bourdeaudhuij I, Deforche B, Cardon G. Opinions towards physical activity interventions using Facebook or text messaging: focus group interviews with vocational school-aged adolescents. Health Soc Care Community 2019 May;27(3):654-664. [doi: 10.1111/hsc.12679] [Medline: 30398291]
18. Dinç F, Kurt A, Yıldız D. The use of mobile health applications in the development of a healthy lifestyle of adolescents: a cross-sectional study. J Pediatr Nurs 2025;83:1-6. [doi: 10.1016/j.pedn.2025.04.012] [Medline: 40273677]
19. Mojica CM, Parra-Medina D, Yin Z, Akopian D, Esparza LA. Assessing media access and use among Latina adolescents to inform development of a physical activity promotion intervention incorporating text messaging. Health Promot Pract 2014 Jul 19;15(4):548-555 [FREE Full text] [doi: 10.1177/1524839913514441] [Medline: 24357863]
20. Ng K, Kokko S, Tammelin T, Kallio J, Belton S, O'Brien W, et al. Clusters of Adolescent Physical Activity Tracker Patterns and Their Associations With Physical Activity Behaviors in Finland and Ireland: Cross-Sectional Study. J Med Internet Res 2020 Sep 01;22(9):e18509-e18555 [FREE Full text] [doi: 10.2196/18509] [Medline: 32667894]
21. Mendoza JA, Baker KS, Moreno MA, Whitlock K, Abbey-Lambertz M, Waite A, et al. A Fitbit and Facebook mHealth intervention for promoting physical activity among adolescent and young adult childhood cancer survivors: a pilot study. Pediatr Blood Cancer 2017 Dec;64(12):e26660. [doi: 10.1002/pbc.26660] [Medline: 28618158]
22. Larsen B, Benitez T, Cano M, Dunsiger SS, Marcus BH, Mendoza-Vasconez A, et al. Web-based physical activity intervention for Latina adolescents: feasibility, acceptability, and potential efficacy of the Niñas Saludables study. J Med Internet Res 2018 May 09;20(5):e170 [FREE Full text] [doi: 10.2196/jmir.9206] [Medline: 29743151]

XSL•FO
RenderX

23. Thompson D, Cantu D, Ramirez B, Cullen KW, Baranowski T, Mendoza J, et al. Texting to increase adolescent physical activity: feasibility assessment. Am J Health Behav 2016 Jul;40(4):472-483 [FREE Full text] [doi: 10.5993/AJHB.40.4.9] [Medline: 27338994]
24. Guthrie N, Bradlyn A, Thompson SK, Yen S, Haritatos J, Dillon F, et al. Development of an accelerometer-linked online intervention system to promote physical activity in adolescents. PLoS One 2015;10(5):e0128639 [FREE Full text] [doi: 10.1371/journal.pone.0128639] [Medline: 26010359]
25. Direito A, Jiang Y, Whittaker R, Maddison R. Smartphone apps to improve fitness and increase physical activity among young people: protocol of the Apps for IMproving FITness (AIMFIT) randomized controlled trial. BMC Public Health 2015 Jul 11;15:635 [FREE Full text] [doi: 10.1186/s12889-015-1968-y] [Medline: 26159834]
26. Chen J, Guedes CM, Cooper BA, Lung AE. Short-term efficacy of an innovative mobile phone technology-based intervention for weight management for overweight and obese adolescents: pilot study. Interact J Med Res 2017 Aug 02;6(2):e12 [FREE Full text] [doi: 10.2196/ijmr.7860] [Medline: 28768612]
27. Caillaud C, Ledger S, Diaz C, Clerc G, Galy O, Yacef K. iEngage: a digital health education program designed to enhance physical activity in young adolescents. PLoS One 2022;17(10):e0274644 [FREE Full text] [doi: 10.1371/journal.pone.0274644] [Medline: 36197890]
28. Egilsson E, Bjarnason R, Njardvik U. Usage and weekly attrition in a smartphone-based health behavior intervention for adolescents: pilot randomized controlled trial. JMIR Form Res 2021 Feb 17;5(2):e21432 [FREE Full text] [doi: 10.2196/21432] [Medline: 33481750]
29. Staiano AE, Beyl RA, Guan W, Hendrick CA, Hsia DS, Newton RL. Home-based exergaming among children with overweight and obesity: a randomized clinical trial. Pediatr Obes 2018 Nov;13(11):724-733 [FREE Full text] [doi: 10.1111/ijpo.12438] [Medline: 30027607]
30. Stasinaki A, Büchter D, Shih CI, Heldt K, Güsewell S, Brogle B, et al. Effects of a novel mobile health intervention compared to a multi-component behaviour changing program on body mass index, physical capacities and stress parameters in adolescents with obesity: a randomized controlled trial. BMC Pediatr 2021 Jul 09;21(1):308 [FREE Full text] [doi: 10.1186/s12887-021-02781-2] [Medline: 34243738]
31. Ortega A, Cushing CC. Design of a temporally augmented text messaging bot to improve adolescents' physical activity and engagement: proof-of-concept study. JMIR Form Res 2024 Oct 10;8:e60171 [FREE Full text] [doi: 10.2196/60171] [Medline: 39388222]
32. Willinger L, Schweizer F, Böhm B, Scheller DA, Jonas S, Oberhoffer-Fritz R, et al. Evaluation of the gamified application KIJANI to promote physical activity in children and adolescents: a multimethod study. Digit Health 2024;10:20552076241271861 [FREE Full text] [doi: 10.1177/20552076241271861] [Medline: 39161345]
33. Schoenfelder E, Moreno M, Wilner M, Whitlock KB, Mendoza JA. Piloting a mobile health intervention to increase physical activity for adolescents with ADHD. Prev Med Rep 2017 Jun;6:210-213 [FREE Full text] [doi: 10.1016/j.pmedr.2017.03.003] [Medline: 28373931]
34. Koorts H, Salmon J, Timperio A, Ball K, Macfarlane S, Lai SK, et al. Translatability of a wearable technology intervention to increase adolescent physical activity: mixed methods implementation evaluation. J Med Internet Res 2020 Aug 07;22(8):e13573 [FREE Full text] [doi: 10.2196/13573] [Medline: 32763872]
35. Glaser M, Green G, Barak S, Bord S, Levi S, Jakobovich R, et al. The effects of the Friendship Online Intervention Program on physical activity, substance abuse, psychosomatic symptoms, and well-being among at-risk youth. J Adolesc 2024 Feb;96(2):251-265. [doi: 10.1002/jad.12272] [Medline: 37985148]
36. Garde A, Umedaly A, Abulnaga SM, Robertson L, Junker A, Chanoine JP, et al. Assessment of a mobile game ("MobileKids Monster Manor") to promote physical activity among children. Games Health J 2015 Apr;4(2):149-158. [doi: 10.1089/g4h.2014.0095] [Medline: 26181809]
37. Mateo-Orcajada A, Abenza-Cano L, Albaladejo-Saura MD, Vaquero-Cristóbal R. Mandatory after-school use of step tracker apps improves physical activity, body composition and fitness of adolescents. Educ Inf Technol (Dordr) 2023 Jan 24:1-32 [FREE Full text] [doi: 10.1007/s10639-023-11584-0] [Medline: 36714445]
38. Cushing C, Bejarano C, Ortega A, Sayre N, Fedele D, Smyth J. Adaptive mHealth intervention for adolescent physical activity promotion. J Pediatr Psychol 2021 Jun 03;46(5):536-546. [doi: 10.1093/jpepsy/jsaa125] [Medline: 33484137]
39. Moffatt S, White M, Mackintosh J, Howel D. Using quantitative and qualitative data in health services research - what happens when mixed method findings conflict? [ISRCTN61522618]. BMC Health Serv Res 2006 Mar 08;6:28 [FREE Full text] [doi: 10.1186/1472-6963-6-28] [Medline: 16524479]
40. Wang X, Cheng Z. Cross-sectional studies: strengths, weaknesses, and recommendations. Chest 2020 Jul;158(1S):S65-S71. [doi: 10.1016/j.chest.2020.03.012] [Medline: 32658654]
41. Domhardt M, Schröder A, Geirhos A, Steubl L, Baumeister H. Efficacy of digital health interventions in youth with chronic medical conditions: a meta-analysis. Internet Interv 2021 Apr;24:100373 [FREE Full text] [doi: 10.1016/j.invent.2021.100373] [Medline: 33732626]
42. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000 Jan;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]

XSL•FO
RenderX

43. Zhang G, Feng W, Zhao L, Zhao X, Li T. The association between physical activity, self-efficacy, stress self-management and mental health among adolescents. Sci Rep 2024 Mar 06;14(1):5488 [FREE Full text] [doi: 10.1038/s41598-024-56149-4] [Medline: 38448518]
44. Teixeira PJ, Silva MN, Mata J, Palmeira AL, Markland D. Motivation, self-determination, and long-term weight control. Int J Behav Nutr Phys Act 2012 Mar 02;9:22 [FREE Full text] [doi: 10.1186/1479-5868-9-22] [Medline: 22385818]
45. Collin C, Eyraud C, Martin P, Michel M, Le Roux E, Alberti C. A scoping review of outcome selection and accuracy of conclusions in complex digital health interventions for young people (2017-2023): methodological proposals for population health intervention research. BMC Med 2025 Jul 02;23(1):400 [FREE Full text] [doi: 10.1186/s12916-025-04245-1] [Medline: 40604855]
46. Stavric V, Kayes NM, Rashid U, Saywell NL. The effectiveness of self-guided digital interventions to improve physical activity and exercise outcomes for people with chronic conditions: a systematic review and meta-analysis. Front Rehabil Sci 2022;3:925620 [FREE Full text] [doi: 10.3389/fresc.2022.925620] [Medline: 36188933]
47. Michalsen H, Wangberg SC, Hartvigsen G, Jaccheri L, Muzny M, Henriksen A, et al. Physical activity with tailored mHealth support for individuals with intellectual disabilities: protocol for a randomized controlled trial. JMIR Res Protoc 2020 Jun 29;9(6):e19213 [FREE Full text] [doi: 10.2196/19213] [Medline: 32437328]
48. Ma J, Floegel T, Li L, Leese J, De Vera MA, Beauchamp M, et al. Tailored physical activity behavior change interventions: challenges and opportunities. Transl Behav Med 2021 Dec 14;11(12):2174-2181 [FREE Full text] [doi: 10.1093/tbm/ibab106] [Medline: 34424344]
49. Shameli A, Althoff T, Saberi A, Leskovec J. How gamification affects physical activity: large-scale analysis of walking challenges in a mobile application. Proc Int World Wide Web Conf 2017 Apr;2017:455-463 [FREE Full text] [doi: 10.1145/3041021.3054172] [Medline: 28990011]
50. Bandura A. Self-Efficacy: The Exercise of Control. New York, NY: W. H. Freeman & Co; 1997.
51. Fløtlo TN, Tørris C, Riiser K, Almendingen K, Chew HSJ, Fosså A, et al. Digital health interventions to treat overweight and obesity in children and adolescents: an umbrella review. Obes Rev 2025 Jun;26(6):e13905. [doi: 10.1111/obr.13905] [Medline: 39972996]
52. van Baak MA, Pramono A, Battista F, Beaulieu K, Blundell JE, Busetto L, et al. Effect of different types of regular exercise on physical fitness in adults with overweight or obesity: systematic review and meta-analyses. Obes Rev 2021 Jul;22 Suppl 4(Suppl 4):e13239 [FREE Full text] [doi: 10.1111/obr.13239] [Medline: 33939229]
53. Fismen A, Galler M, Klepp K, Chatelan A, Residori C, Ojala K, et al. Weight status and mental well-being among adolescents: the mediating role of self-perceived body weight. a cross-national survey. J Adolesc Health 2022 Aug;71(2):187-195. [doi: 10.1016/j.jadohealth.2022.02.010] [Medline: 35428562]
54. Kollins SH, DeLoss DJ, Cañadas E, Lutz J, Findling RL, Keefe RSE, et al. A novel digital intervention for actively reducing severity of paediatric ADHD (STARS-ADHD): a randomised controlled trial. Lancet Digit Health 2020 Apr;2(4):e168-e178 [FREE Full text] [doi: 10.1016/S2589-7500(20)30017-0] [Medline: 33334505]
55. Wang W, Cheng J, Song W, Shen Y. The effectiveness of wearable devices as physical activity interventions for preventing and treating obesity in children and adolescents: systematic review and meta-analysis. JMIR Mhealth Uhealth 2022 Apr 08;10(4):e32435 [FREE Full text] [doi: 10.2196/32435] [Medline: 35394447]
56. Stamatis CA, Farlow DN, Mercaldi C, Suh M, Maple A, Savarese A, et al. Two single arm trials of AKL-T01, a digital therapeutic for adolescents and adults with ADHD. Npj Ment Health Res 2024 Jun 19;3(1):30 [FREE Full text] [doi: 10.1038/s44184-024-00075-w] [Medline: 38898133]
57. Cuzick J. The importance of long-term follow up of participants in clinical trials. Br J Cancer 2023 Feb;128(3):432-438 [FREE Full text] [doi: 10.1038/s41416-022-02038-4] [Medline: 36456713]
58. Kahwati L, Viswanathan M, Golin CE, Kane H, Lewis M, Jacobs S. Identifying configurations of behavior change techniques in effective medication adherence interventions: a qualitative comparative analysis. Syst Rev 2016 May 04;5:83 [FREE Full text] [doi: 10.1186/s13643-016-0255-z] [Medline: 27209092]
59. Bi S, Yuan J, Wang Y, Zhang W, Zhang L, Zhang Y, et al. Effectiveness of digital health interventions in promoting physical activity among college students: systematic review and meta-analysis. J Med Internet Res 2024 Nov 20;26:e51714 [FREE Full text] [doi: 10.2196/51714] [Medline: 39566049]
60. Kassim PSJ, Muhammad NA, Rahman NFA, Sidik SM, Essau CA, Shah SA. Digital behaviour change interventions to promote physical activity in overweight and obese adolescents: a systematic review protocol. Syst Rev 2022 Sep 05;11(1):188 [FREE Full text] [doi: 10.1186/s13643-022-02060-w] [Medline: 36064617]
61. Barker T, Habibi N, Aromataris E, Stone J, Leonardi-Bee J, Sears K, et al. The revised JBI critical appraisal tool for the assessment of risk of bias for quasi-experimental studies. JBI Evid Synth 2024 Mar 01;22(3):378-388. [doi: 10.11124/JBIES-23-00268] [Medline: 38287725]
62. Gómez-Cuesta N, Mateo-Orcajada A, Meroño L, Abenza-Cano L, Vaquero-Cristóbal R. A mobile app-based intervention improves anthropometry, body composition and fitness, regardless of previous active-inactive status: a randomized controlled trial. Front Public Health 2024;12:1380621 [FREE Full text] [doi: 10.3389/fpubh.2024.1380621] [Medline: 39193194]

XSL•FO
RenderX

63. Silva AG, Simões P, Queirós A, P Rocha N, Rodrigues M. Effectiveness of mobile applications running on smartphones to promote physical activity: a systematic review with meta-analysis. Int J Environ Res Public Health 2020 Mar 27;17(7):2251 [FREE Full text] [doi: 10.3390/ijerph17072251] [Medline: 32230769]
64. Howard-Wilson S, Ching J, Gentile S, Ho M, Garcia A, Ayturk D, et al. Efficacy of a multimodal digital behavior change intervention on lifestyle behavior, cardiometabolic biomarkers, and medical expenditure: protocol for a randomized controlled trial. JMIR Res Protoc 2024 Oct 30;13:e50378 [FREE Full text] [doi: 10.2196/50378] [Medline: 39475852]
65. Böhm B, Karwiese SD, Böhm H, Oberhoffer R. Effects of mobile health including wearable activity trackers to increase physical activity outcomes among healthy children and adolescents: systematic review. JMIR Mhealth Uhealth 2019 Apr 30;7(4):e8298 [FREE Full text] [doi: 10.2196/mhealth.8298] [Medline: 31038460]
66. Singh B, Ahmed M, Staiano AE, Gough C, Petersen J, Vandelanotte C, et al. A systematic umbrella review and meta-meta-analysis of eHealth and mHealth interventions for improving lifestyle behaviours. NPJ Digit Med 2024 Jul 05;7(1):179 [FREE Full text] [doi: 10.1038/s41746-024-01172-y] [Medline: 38969775]
67. Bas-Sarmiento P, Julián-López C, Fernández-Gutiérrez M, Poza-Méndez M, Marín-Paz A. Gamified eHealth interventions for health promotion and disease prevention in children and adolescents: a scoping review. Humanit Soc Sci Commun 2025 Mar 20;12(1):397. [doi: 10.1057/s41599-025-04670-w]
68. Badr J, Motulsky A, Denis J. Digital health technologies and inequalities: a scoping review of potential impacts and policy recommendations. Health Policy 2024 Aug;146:105122 [FREE Full text] [doi: 10.1016/j.healthpol.2024.105122] [Medline: 38986333]
69. Ghadi YY, Shah SFA, Waheed W, Mazhar T, Ahmad W, Saeed MM, et al. Integration of wearable technology and artificial intelligence in digital health for remote patient care. J Cloud Comp 2025 Jul 21;14(1):39. [doi: 10.1186/s13677-025-00759-4]
70. Grand VR, Inc. U.S. Digital Health Market Size, Share. Grand View Research. 2024. URL: https://www.grandviewresearch.com/industry-analysis/us-digital-health-market-report [accessed 2025-11-10]
71. Majcherek D, Hegerty SW, Kowalski AM, Lewandowska MS, Dikova D. Opportunities for healthcare digitalization in Europe: comparative analysis of inequalities in access to medical services. Health Policy 2024 Jan;139:104950 [FREE Full text] [doi: 10.1016/j.healthpol.2023.104950] [Medline: 38061175]
72. El-Jardali F, Bou-Karroum L, Jabbour M, Bou-Karroum K, Aoun A, Salameh S, et al. Digital health in fragile states in the Middle East and North Africa (MENA) region: a scoping review of the literature. PLoS One 2023;18(4):e0285226 [FREE Full text] [doi: 10.1371/journal.pone.0285226] [Medline: 37115778]
73. Sylla B, Ismaila O, Diallo G. 25 years of digital health toward universal health coverage in low- and middle-income countries: rapid systematic review. J Med Internet Res 2025 May 29;27:e59042 [FREE Full text] [doi: 10.2196/59042] [Medline: 40440696]
74. ew GSMA report shows mobile internet connectivity continues to grow globally but barriers for 3.45 billion unconnected people remain. GSMA. 2024 Oct 23. URL: https://www.gsma.com/newsroom/press-release/new-gsma-report-shows-mobile-internet-connectivity-continues-to-grow-globally-but-barriers-for-3-45-billion-unconnected-people-remain/ [accessed 2025-11-10]
75. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
76. Pratap A, Neto EC, Snyder P, Stepnowsky C, Elhadad N, Grant D, et al. Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. NPJ Digit Med 2020;3:21 [FREE Full text] [doi: 10.1038/s41746-020-0224-8] [Medline: 32128451]
77. Pathiravasan CH, Zhang Y, Wang X, Trinquart L, Benjamin EJ, Borrelli B, et al. Factors associated with long-term use of digital devices in the electronic Framingham Heart Study. NPJ Digit Med 2022 Dec 27;5(1):195 [FREE Full text] [doi: 10.1038/s41746-022-00735-1] [Medline: 36572707]
78. Jakob R, Harperink S, Rudolf AM, Fleisch E, Haug S, Mair JL, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371 [FREE Full text] [doi: 10.2196/35371] [Medline: 35612886]
79. Kelders SM, Kok RN, Ossebaard HC, Van Gemert-Pijnen JE. Persuasive system design does matter: a systematic review of adherence to web-based interventions. J Med Internet Res 2012 Nov 14;14(6):e152 [FREE Full text] [doi: 10.2196/jmir.2104] [Medline: 23151820]

Abbreviations

BCW: Behavior Change Wheel
COM-B: capability, opportunity, motivation—behavior
CSS: cross-sectional study
DHI: digital health intervention
GRADE: Grading of Recommendations, Assessment, Development and Evaluation
JBI: Joanna Briggs Institute

XSL•FO
RenderX

MeSH: Medical Subject Headings
mHealth: mobile health
MMS: mixed methods study
PICOS: population, intervention, comparator, outcomes, and study design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Search Extension
PROSPERO: International Prospective Register of Systematic Reviews
QES: quasi-experimental study
QR: quantitative research
RCT: randomized controlled trial

Edited by S Brini; submitted 14.Aug.2025; peer-reviewed by C Ahmadu, O Akhadelor; comments to author 08.Sep.2025; accepted 02.Jan.2026; published 27.Feb.2026.

Please cite as:
Fan RS, Jiang JJ, Zhou QY, Zhang XY, Wu ZH, Ji L
Digital Health Interventions to Promote Physical Activity Among Adolescents: Systematic Review
J Med Internet Res 2026;28:e82395
URL: https://www.jmir.org/2026/1/e82395
doi:10.2196/82395
PMID:

©Rui Shi Fan, Jia Jun Jiang, Qing Yuan Zhou, Xin Yue Zhang, Zhou Hang Wu, Liu Ji. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Digital Technologies That Support Meaningful Connections in Care Homes: Scoping Review

Deborah Muldrew[1], PhD; Rosemary Bradley[2], MSc; Kelly Conway[1], MSc

[1]Institute of Nursing and Health Research, University of Ulster, Belfast, United Kingdom

[2]Institute of Nursing and Health Research, University of Ulster, Londonderry, United Kingdom

Corresponding Author:
Deborah Muldrew, PhD
Institute of Nursing and Health Research
University of Ulster
Ulster University Belfast Campus
2-24 York Street
Belfast, BT15 1AP
United Kingdom
Phone: 44 2895365301
Email: d.muldrew@ulster.ac.uk

Abstract

Background: Meaningful connections in which people feel valued, seen, and heard are essential for social health and well-being. However, individual, systemic, and structural barriers exist within care homes that exacerbate risks for social isolation in this population. Leveraging digital technology to promote meaningful connections has the potential to affect positive health outcomes; however, the evidence base within the care home environment is not clear.

Objective: This scoping review aimed to identify how digital technologies are used in long-term care settings to facilitate meaningful connections.

Methods: Following Joanna Briggs Institute and PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) reporting guidelines, 6 online databases (CINAHL, MEDLINE, PsycINFO, Scopus, IEEE Explore, and ACM) were searched on December 12, 2025. Studies were included if they were published in English; included care home residents, relatives, or staff; and directly discussed the application of a digital technology that focused on building meaningful connections. No date limiters were applied. Gray literature was searched through Google Scholar and conference proceedings, and forward and backward citation searching was also undertaken. Papers were managed in Covidence and reviewed against inclusion criteria by 3 reviewers (DM, RB, and KC). Results were presented using tabular, graphical, and narrative summaries to describe how the findings relate to the review objectives.

Results: In total, 72 studies were included. More than half of the studies were published since 2022 (38/72, 53%). Australia (15/72, 21%) and the Netherlands (12/72, 17%) had the highest frequency of publications. Technologies including robotics, virtual reality, mobile or tablet apps, digital devices, and online programs have been used in care homes. Factors underpinning the development of meaningful connections include getting to know the person, increases autonomy and choice, source of enjoyment and fun, facilitates communication, and builds community. Engagement, well-being or satisfaction, emotional response, quality of life, purpose and meaning, social closeness, loneliness, depression and anxiety, and psychosocial capacity were the main indicators of meaningful connections. Limitations include English-language restriction, potential missed studies on social connection, exclusion of passive technologies, and heterogeneous outcome measures.

Conclusions: This review is the first to explore how digital technologies are applied in care homes to facilitate meaningful connections and identify how digital technology is a catalyst for meaningful human connection rather than a replacement. The application of generative artificial intelligence technology is currently a gap in the evidence base and needs to be considered in conjunction with key stakeholders to ensure that future developments meet the social needs of the entire community. Future research needs to consider how new evaluation metrics can be developed that combine digital engagement data with validated measures of meaningful connection to assess the person-centered impact of generative artificial intelligence technologies in care homes.

(J Med Internet Res 2026;28:e88181) doi:10.2196/88181

XSL•FO
RenderX

KEYWORDS

scoping review; digital technology; meaningful connection; social connection; care homes; nursing homes; older people; aged

Introduction

Social connection describes how individuals relate to each other, encompassing both observable aspects such as engagement with social networks and the experiences of loneliness or social connectedness [1]. People who are socially connected feel an overall sense of belonging to a group or community. The World Health Organization (WHO) recognizes social connection as a global public health priority. Unfortunately, societal trends over a number of decades indicate that, as a population, we have become less socially connected, and a high percentage of the population is lonely [2]. Current global estimates suggest that 1 in 4 older adults experience social isolation [3]. This concern is likely to become more prevalent, given the speed of population aging, which is significantly faster than in previous decades. Current population predictions by the WHO estimate that by 2050, the number of individuals aged 60 years and older will double to 2.1 billion globally, and the number of individuals aged 80 years and older will triple to 426 million [4]. An aging population, coupled with societal shifts in how care is delivered, is predicted to place a substantial rise in the demand on long-term care facilities [5]. Long-term care consists of a range of medical or nursing care, personal care (help with activities of daily living), assistance services (help with instrumental activities of daily living), and support to maintain social relationships and activities [6]. The term “care homes” encompasses 2 types of long-term care settings: residential care homes (for people requiring personal care) and nursing homes (for people requiring personal and nursing care) [7]. Individual, systemic, and structural barriers exist within care homes that exacerbate the risks for social isolation in this population [8].

While the goal toward social connection is well documented, meaningful connections are the first step toward creating an enriching life and can have a significant impact on health and well-being [9]. Meaningful connections are characterized by deeper, more intimate relationships that make the person feel seen, heard, and valued. For the purpose of this research, we have defined meaningful connections as “interpersonal interactions, whether one-off or repeated, that foster feelings of being seen or heard, empathy, or enhanced social belonging; and/or lead to indicators such as reduced loneliness, increased engagement, or improvements in psychosocial outcomes.” An avenue to such meaningful connection is meaningful activity. National Institute for Health and Care Excellence guidelines recommend that people should be offered opportunities to participate in meaningful activity that promotes health and mental well-being [10]. Meaningful activities are defined as “all activities or occupations that are significant or meaningful for the person and reflect someone’s current and past interests, routines, habits, and roles and are adjusted to someone’s abilities” [11]. Engagement in personally meaningful activities leads to improved psychological well-being [12], and research into people with dementia living in care homes reported improvements in behavioral and psychological symptoms through the provision of meaningful activities [13]. However, when the focus becomes the activity itself and not the meaningful connection formed between the people involved, this approach could lead to task-oriented approaches to care, currently being challenged within health care [14]. Furthermore, activities viewed as something separate from everyday routine, and require additional staffing to coordinate in a busy environment where resources and staffing are already overstretched, are less likely to be meaningful or person-centered [15]. Instead, considerations should be given to the meaningful connections that can form between people at any time, not just during planned or purposeful meaningful activity.

According to Kemp et al [16], meaningful connections are essential for well-being and may serve as effective nonpharmacological approaches to mitigate adverse behavioral and psychological manifestations such as anxiety, depression, apathy, agitation, and aggression. Gonnord et al [17] also demonstrated that engagement could positively affect a number of aspects of well-being, such as social connection, personal autonomy, physical health, and mental stimulation. Social connectedness and participation in meaningful activities are important sources of joy, which have been shown in the literature as important to the health and social care of older people [18]. In the nursing home context, meaningful social relationships with peers and families foster a sense of belonging, purpose, achievement, continuity, and significance, contributing to their psychosocial well-being [19]. Within nursing homes, social connectedness may be an important modifiable risk factor for adverse health outcomes [20]. However, barriers to quality social participation exist at an individual level, for example, personal motivations and preexisting social networks, and at a community level, including accessibility and neighborhood cohesion [21]. Although these findings come from research into older adults more widely, their relevance to older adults living in care homes is clear. As highlighted by Nygaard et al [22], there are additional obstacles and challenges to achieving person-centered meaningful connections in care homes including a regimented structure to care home life and a lack of time for staff to spend quality time with individual residents, as well as a lack of connectedness within meaningful social relationships.

Leveraging digital technology to promote social connectedness has the potential to affect positive health outcomes [23]. The introduction of digital technologies was accelerated into care homes as a result of global restrictions during the COVID-19 pandemic, despite reported systemic and structural barriers [24]. Technology-supported interventions can improve old-age social well-being among older adults living in social isolation [25]; however, the evidence base is typically condition-specific (eg, dementia) or focused on treating a clinical outcome (eg, depression), rather than exploring the role of technology to facilitate meaningful connections among residents, relatives, and staff. Furthermore, much of the evidence based around digital technology considers technology solely for passive monitoring purposes such as digital movement sensors or blood pressure monitors. Therefore, a review of the evidence is required to fill the current evidence gap on how digital

XSL•FO
RenderX

technologies can be used to form meaningful connections in care homes. Given the overall limited focus within the evidence base on application of digital technology to support meaningful connections among the wider care home population, formulating a precise question for a systematic review was not possible. Instead, a scoping review supported the mapping of all digital technologies applied to the care home setting with the aim of supporting meaningful connections [26]. A preliminary search of CINAHL, MEDLINE, and *JBI Evidence Synthesis* was conducted, and no current or underway systematic reviews or scoping reviews on the topic were identified.

The aim of this scoping review is to identify how digital technologies are used in long-term care settings to facilitate meaningful connections. The objectives related to this aim are (1) to identify the scope of digital technologies being applied within the care home setting to build meaningful connections, (2) to understand the factors through which digital technologies are able to support meaningful connections, (3) to identify which outcomes are measured to evidence a change in meaningful connections, and (4) to identify gaps in the evidence base to inform future research, policy, and practice.

Methods

Design

A scoping review methodology following guidance from the Joanna Briggs Institute (JBI) and PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews; Multimedia Appendix 1) [27] reporting guidelines was selected as the most appropriate evidence synthesis approach, based on the guidance from Munn et al [26]. The aim was to scope and map the existing literature on digital technologies used in long-term care homes to support meaningful connections, including technologies used, factors showing how the digital technologies promote meaningful connections, and outcome measures of such connections. The systematic approach outlined by the JBI [28] was chosen for this review to provide a high degree of rigor and transparency, to ensure that the review can be replicated easily, and to increase its value to health care decision-makers. Adherence to the items reported in the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) checklist was used to strengthen the transparency of reporting the search [29]. Ethics approval was not required because scoping reviews synthesize existing publicly available literature without involving human participants or identifiable data.

Protocol

A protocol for the scoping review was developed and subject to peer review within the first author's (DM) higher education institution. However, this protocol was not registered externally. Following the review of the protocol, changes to the final review included the addition of "aged care" as a concept, the addition of 2 databases (ACM Digital Library and IEEE Xplore), and the expansion of the gray literature searches to include a range of relevant conference proceedings from 2025.

Eligibility Criteria

As the scoping review question guides and directs the development of the inclusion criteria, the JBI recommends the use of the population, concept, and context mnemonic in ensuring the eligibility of the primary research question [28]. The research questions were developed through consultation with the research team and key stakeholders focusing on the overall aim of the scoping review. Textbox 1 presents the identified population, concept, and context and the primary research question.

Textbox 1. Population, concept, and context framework for defining the inclusion criteria of studies.

Population

Participants include care home residents, relatives, and staff. Care homes include residential and nursing homes. Studies were excluded if they were undertaken in nonresidential (day care) services, supported housing, facilities for people with learning disabilities or enduring mental health conditions, or other community care settings such as in the person's own home.

Concept

Meaningful connections, activities, or engagement as a goal of the digital technology. Studies were excluded if they focused solely on clinical health or other outcomes without data related specifically to meaningful connections.

Context

Digital technologies used actively with or by people in care homes. Studies were excluded if the technology was solely for passive monitoring purposes, such as digital movement sensors or blood pressure monitors, or did not prioritize the creation of meaningful connections through digital means.

Research question

How do digital technologies facilitate meaningful connections in care homes?

Studies published in English were included due to practical limitations (translator not available to the research team). No date limitations were applied to the searches to allow for an exploration of how digital technologies have changed over time in type and frequency of application in care homes. No published search filters were used. This scoping review included all primary research, including qualitative, quantitative, and mixed methods approaches. Opinion papers and editorials were also considered for inclusion in this scoping review; however, they were required to report some primary data.

Information Sources

CINAHL Ultimate (EBSCOhost), MEDLINE (Ovid), Scopus, ACM Digital Library, IEEE XPlore, and PsycINFO (Ovid) were independently searched on December 12, 2025. Gray literature

XSL•FO
RenderX

was searched using Google Scholar (first 100 hits) and websites including the WHO and National Institute for Health and Care Excellence. Published conference proceedings from 2025 were hand-searched to identify any new publications that may not be published and accessible through traditional database searches. Conferences included the CHI Conference on Human Factors in Computing Systems, Good IT Proceedings of the 2025 International Conference on Information Technology for Social Good, COMPASS '25: Proceedings of the 2025 ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies, C&C '25: Proceedings of the 2025 Conference on Creativity and Cognition, MuC '25: Proceedings of the Mensch und Computer 2025, WWW '25: Companion Proceedings of the ACM on Web Conference 2025, and VRST '25: Proceedings of the 2025 31st ACM Symposium on Virtual Reality Software and Technology.

Search Strategy

The search strategy aimed to locate both published and unpublished studies. A 3-step search strategy was used in this review. First, an initial limited search of MEDLINE Ovid and CINAHL Ultimate was undertaken to identify papers on the topic in conjunction with a university librarian and subject specialist. This search was based on the subject expertise of the team and the database searching skills of the librarian. The text words contained in the titles and abstracts of relevant papers and the index terms used to describe the papers were used to develop a full search strategy (Multimedia Appendix 2).

The search strategy, including all identified keywords and index terms, was adapted for each included database and information source. A broad range of keywords and MeSH (Medical Subject Headings) headings were used to reduce the likelihood of missing relevant literature, and then, the inclusion and exclusion criteria were applied strictly to ensure that the papers included were relevant to the objectives. Forward and backward citation searching on the included sources of evidence was undertaken to identify additional studies. No authors were contacted, as all papers identified through the search strategy were available. Email alerts were set up for each database to provide weekly updates of new literature.

Study Selection

Following the search, all identified citations were collated and uploaded into Covidence, and duplicates were removed. Duplication was managed within Covidence by the in-built duplication feature and thorough review from the authors. Studies were sorted alphabetically by author to allow for identification of additional duplicates missed by Covidence. Following a pilot test, titles and abstracts were screened by multiple independent reviewers (DM, KC, and RB) for assessment against the inclusion criteria for the review. Potentially relevant sources were retrieved in full and stored in Covidence. The full text of selected citations was assessed in detail against the inclusion criteria by multiple independent reviewers (DM, KC, and RB). Reasons for exclusion of sources of evidence at full text that did not meet the inclusion criteria were recorded and reported in the scoping review. Any disagreements that arose between the reviewers at each stage of the selection process were resolved through discussion.

Data Charting

Data extraction was undertaken to develop a logical and descriptive summary of the results, aligned to the objectives of the review. Data were extracted from papers included in the scoping review by multiple independent reviewers (DM, KC, and RB) using a data extraction tool developed by the research team. It was based on the JBI data extraction template to meet the objectives of the review. Each researcher (DM, KC, and RB) piloted the data extraction tool on 5 papers before meeting to discuss the utility of the tool. Originally, all findings were collected in one column; however, this was leading to confusion, and key data being overlooked. The data extraction tool was updated to have 3 separate results columns: one for the type of digital technology discussed, one for the factors that create meaningful connections, and one for the outcome measures identified. Any disagreements that arose between the reviewers were resolved through discussion. Once data extraction was completed for the included studies, a second reviewer (RB) repeated the data extraction process for 10% (7/72) of the included studies to ensure that all relevant information was extracted accurately. See Multimedia Appendix 3 for the completed data extraction table.

Data Items

The data extracted included specific details about the author, year of publication, country of origin, population and sample size, methods, type of technology used, factors supporting meaningful connection, and outcome measures.

Quality Appraisal

Quality appraisal was not undertaken, aligned to methodological norms for scoping reviews. However, a descriptive assessment of each study design, based on the levels of evidence identified by Melnyk and Fineout-Overholt [30], was included to support the development of more robust policy and practice recommendations. The rating scale ranged from 1=strong (systematic review or meta-analysis of randomized controlled trials) to 7=weak (opinion pieces and narrative literature review). In total, 65 studies were categorized as level 6, representing evidence from a single qualitative or descriptive study.

Synthesis of Results

An overview of included studies was used to summarize bibliographic information. A narrative summary was developed to describe how the findings relate to the core objectives of the review: the digital technologies used, the factors through which meaningful connections are established, and the outcome measures reported. Three independent researchers (DM, KC, and RB) read and reread the included studies and extracted data relevant to the core objectives. The extracted data were then synthesized to ensure that a comprehensive account of relevant information was presented in the final data extraction table. A narrative summary accompanying the data presented was developed by the first author (DM) and reviewed and edited by the second and third authors (RB and KC).

XSL•FO
RenderX

Results

Selection of Sources of Evidence

In total, 12,742 papers were identified through the database searches, citation searching, and gray literature. Following the removal of duplicates, 12,219 papers were screened against the inclusion and exclusion criteria. The results of the search are shown in the PRISMA flow diagram (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

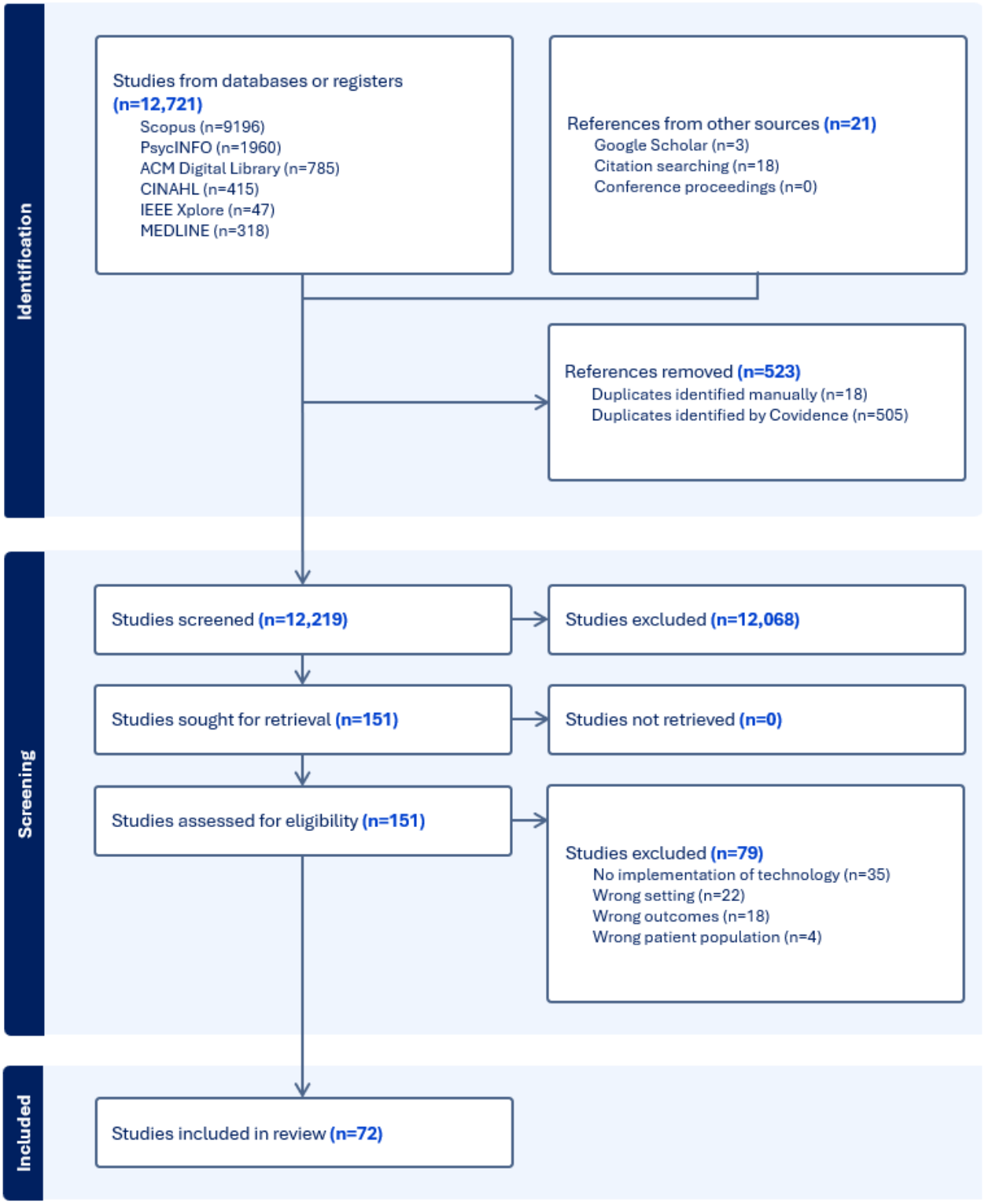

In total, 151 full-text studies were assessed for eligibility, and 79 were excluded at this point. Reasons for exclusion included no digital technologies being implemented in the paper (35/79, 44%), papers focused on clinical outcomes instead of meaningful connection (18/79, 23%), research was undertaken in the wrong setting (22/79, 28%), or papers included a range of settings or populations and did not have a separate analysis available for the focus of the review (4/79, 5%). In total, 72 studies were

XSL•FO
RenderX

identified for inclusion in the final review (Multimedia Appendix 3).

Characteristics of Included Studies

The first study that reported on digital technologies to support meaningful connections in care homes was published in 2002; however, more than half of the studies were published since 2022 (38/72, 53%; Table 1). While older studies focused more on film screenings and videophones, virtual reality (VR) and digital apps were the most popular digital technologies in 2024-2025, showing a change over time in the types of digital technologies explored.

Table 1. Year of publication.

Year	Values, n (%)
2002	1 (1)
2003	1 (1)
2009	1 (1)
2010	1 (1)
2013	1 (1)
2014	1 (1)
2016	2 (3)
2017	1 (1)
2018	3 (4)
2019	4 (6)
2020	9 (13)
2021	10 (14)
2022	11 (15)
2023	7 (10)
2024	14 (19)
2025	5 (7)

In total, 15 studies were undertaken in Australia, 12 studies in the Netherlands, 8 studies in Canada, 5 studies in Norway and the United Kingdom, respectively, 4 studies in the United States, 3 studies across multiple countries, 3 studies in Germany, Italy, and Taiwan, respectively, 2 studies in Denmark, Finland, and Sweden, respectively, and 1 study in Egypt, Ireland, Japan, Jordan, and Switzerland, respectively. Study settings were identified as care homes (17/72, 24%), residential homes (17/72, 24%), nursing homes (13/72, 18%), long-term care settings (16/72, 22%), and other care facilities including “aged care residences” (9/72, 13%). Participants included care home residents (47 studies, 1432 residents), professional caregivers (30 studies, 811 people), and friends and family (18 studies, 407 people) primarily. Some studies included other key stakeholders alongside residents, families, and staff, including high school students, expert older adults, and volunteers. In total, 17 studies specifically included patients with dementia in their sample, including a total of 269 residents with a dementia diagnosis.

Types of Digital Technologies

Types of digital technologies used in care homes to support meaningful connections were grouped into 5 main categories: mobile and tablet apps, VR and immersive technology, robotics, digital devices and ambient technology, and online or computer-assisted programs (Table 2).

XSL•FO
RenderX

Table 2. Summary of intervention types and functions.

Intervention and function	Examples
Mobile and tablet apps (n=18)	
Communication	• Ticket To Talk app • iPad communication app (unnamed) • Video calling (n=4; eg, FaceTime) • Teleconferencing (n=2)
Reminiscence	• MemoryKeeper app • StoryTiling app
Social activity	• Connecting Today app (remote visiting) • A Better Visit app (shared activity) • Cognitive stimulation therapy app (n=2) • Virtual visits to places of personal significance (n=2) • iBeacons and audio server with iOS applications (virtual museum visit; n=2)
Devices and ambient technology (n=21)	
Wearables and sensory stimulation	• MEMENTO wearable reminiscence technology • Hepatic devices during video calls (warm hand, vibrating music strap, weighted collar, warm hug robot, Hey bracelet) • Interactive sound cushion “VITA” (n=3) • Alternative controller musical instrument
Creative engagement	• VENSTER interactive art (n=2) • ABBY interactive touchscreen • AmbientEcho interactive media system • Vodafone smart photo frame
Social activity	• Computer kiosks • Digital television for reminiscence • Interactive table • Motitech bikes • KOMPa “one button computer connecting generations” (n=4) • SAMb companions
Smart home and voice-controlled assistants	• Google Nest Hub Max • Google Home Assistant
Online and computer-assisted programs (n=8)	
Reminiscence	• Videotape (Remembering When) • Film screenings • Computer Interactive Reminiscence and Communication Aid
Social activity	• SitDance (interactive touchscreen with exercise program) • Virtual creative aging program (museum and music) • ArtOnTheBrain (museum)
Communication	• Online social networking • Internet-based videoconferencing
Virtual reality (VR) and immersive technology (n=10)	
Social activity	• VR treadmill • VR Take A Cycle With Me • VR cycling with pedal exercisers and GoPro video footage • VR Vacation HomeImmersive • 3D VR with hands-on horticultural therapy • VR companion Kiera
Reminiscence	• 3D virtual worlds (reminiscence room, virtual tours, gardening) • VR reminiscence with VR photo albums or videos • SENSE-Garden (n=2)
Robotics (n=14)	

XSL•FO
RenderX

Intervention and function	Examples
Social activity	• LiveNature (sheep) robotic social sheep (n=3) • Robotic pets (cat and dog; n=2) • Paro (baby harp seal; n=2) • Zora humanoid • Adam humanoid • Matilda • Pepper semihumanoid • Rebecca social robot • Zoomorphic cleaning robot • Robot-facilitated dance therapy

[a]KOMP: a communication tool developed by Norwegian company No Isolation, designed to bridge the gap between older adults and their digitally experienced families. It consists of a screen with one single button that turns it on and off and an app that is connected to the screen.

[b]SAM: an interactive companion that plays sounds and displays colors in response to touch-based inputs and can mimic or react to the input of a paired device to stimulate social engagement between care home residents.

Mobile apps were discussed in 18 studies [31-48]. The primary aim of these interventions was to increase communication (eg, Ticket to Talk, video calling, and FaceTime), support reminiscence (eg, MemoryKeeper and StoryTiling), increase access to shared activities (eg, A Better Visit, Virtual Museum visiting app paired with iBeacons and audio servers, cognitive stimulation therapy, and virtual visits), and deliver cognitive stimulation therapy. Digital devices and ambient-assisted living technologies were reported in 21 studies [49-69]. New devices included wearables and sensory stimulation such as MEMENTO wearable reminiscence technology and hepatic devices to wear during video calls, creative engagement devices such as VENSTER interactive artwork and AmbientECHO interactive media system, social activities such as Motitech bikes and interactive tables, and smart home and voice-controlled assistants such as Google Nest Hub Max. Online and computer-assisted programs were reported in 8 studies [70-77]. These programs supported reminiscence such as videotapes, film screenings, and computer-interactive reminiscence-specific programs, social activity such as interactive exercise videos (SitDance), creative aging programs (museum visits and music), and communication such as online social networking. VR and immersive experiences were reported in 10 studies [78-87], including social activities such as VR treadmills, VR cycling (Take a Cycle with Me), VR Vacation Home, a VR companion (Kiera), an immersive personalized sensory experience (SENSE-Garden), and reminiscence activities such as 3D virtual words with a reminiscence room and VR reminiscence through virtual photo albums and videos. The application of robotics in care homes was reported in 14 studies [88-101], including robotic animals (sheep, cats, dogs, and seals), humanoids and semihumanoids (Adam, Matilda, Pepper, Rececca, and Zora), a zoomorphic cleaning robot with a social role, and robot-facilitated dance therapy. One study reported on the application of multiple digital technologies including digital music therapy, mobile apps, video games, videoconference, social robots, VR, and other technologies [102].

Outcomes to Measure Meaningful Connection

In total, 38 studies primarily used qualitative analysis techniques such as thematic analysis to assess outcomes, and another 12 studies combined qualitative analysis techniques with quantitative analysis. Observation was also a commonly used technique to observe changes in meaningful connections such as observing verbal and nonverbal responses, enjoyment, ease of responding to stimuli, and eye contact [35,72,88,99]. Engagement, well-being or satisfaction, emotional response, quality of life, purpose and meaning, social closeness, loneliness, depression and anxiety, and psychosocial capacity were the main themes identified from the scales used to indicate whether meaningful connections were present (Table 3). Studies also considered technology acceptance and clinical outcomes such as cognitive function, activities of daily living, and neuropsychiatric symptoms.

XSL•FO
RenderX

Table 3. Summary of outcome measurement instruments.

Instrument	Studies	Focus	Commentary on the tool from the authors
Engagement			
Engagement of a Person With Dementia Scale	[94]	Evaluates user engagement through activity participation (affective, visual, verbal, behavioral, and social).	Multiple raters can be a strength; however, when staff are unfamiliar with using the tool, and different staff complete the tool due to shift changes, this may lead to inconsistencies in the results.
Observational Measurement of Engagement	[88,94,95]	Direct observation measure of activity engagement for persons with dementia (attention and attitude).	Rated based on direct observations on-site by 2 different raters.
Music in Dementia Assessment Scale (adapted)	[70]	Trained raters observe and quantify behaviors (interest, response, initiation, involvement, and enjoyment).	No comment.
Engagement in Meaningful Activities Survey	[70]	Measures the impressions of the meaningfulness of daily activities.	Some of the items focus on domains typically impacted by aging and may influence responses.
Video Coding Incorporating Observed Emotion	[51]	Overview of emotional responses through 6 dimensions of engagement (emotional, verbal, visual, behavioral, collective engagement, and agitation).	A second coder is required to enhance the reliability and validity of the analysis. Automatic interpretation of affective facial expression and qualitative measures of connectedness and social inclusion should be included.
Ethographic and Laban-Inspired Coding System of Engagement	[88]	Measures engagement of a person with dementia through observational behaviors (head, torso, and hands or arms).	Instructions are needed to ensure consistency in the measurement procedures (eg, reduce personal conversations irrelevant to the study).
Visual Focus of Attention	[103]	OpenVINO used to autonomously track 2D positions of users' eyes, ears, and nose.	No comment.
Well-being and satisfaction			
Warwick-Edinburgh Mental Well-Being Scale	[59]	Unidimensional self-report instrument (14 items) assessing well-being administered either in person or via telephone.	Previous studies demonstrated the feasibility of the tool's use with vulnerable populations.
World Health Organization-Five Well-Being Index	[81]	Self-report measure of mental well-being.	Previous studies demonstrated validity in screening for depressive mood.
Satisfaction With Life Scale	[75]	Self-report 5-item scale assessing global satisfaction with life.	No comment.
Ryff's Psychological Well-Being Scale	[85]	Six dimensions—autonomy, environmental mastery, personal development, supportive relationships, purpose in life, and acceptance of oneself.	No comment.
Emotional response			
Observed Emotional Rating Scale	[94]	Measures the extent of emotional expressions (pleasure, anger, anxiety, sadness, and general alertness).	Rated based on direct observations on-site by 2 different raters. Evidence supports interrater reliability for each state and concurrent and discriminant validity for use by trained raters.
Four items measuring enjoyment from the Game Experience Questionnaire	[40]	Self-report 4 items measuring enjoyment in the experience.	No comment.
People Environment Apathy Scale—Apathy subscale	[94]	Four-item Likert scale assessing apathy-related behaviors through facial expression, eye contact, physical engagement, purposeful activity, verbal tone, and verbal expression.	Valuable tool to assess the extent of intrinsic motivation to behave despite being positively or negatively engaged.
Quality of life			

XSL•FO
RenderX

Instrument	Studies	Focus	Commentary on the tool from the authors
EuroQOL[a] EQ-5D-5L	[80]	A 2-page instrument with a descriptive system and visual analog scale, comprising 5 dimensions (mobility, self-care, usual activities, pain or discomfort, and anxiety or depression).	Valid measure of health-related quality of life in people living in residential care.
Multicultural Quality of Life Index	[70]	Captures subjective quality of life over 10 dimensions such as physical well-being and spiritual fulfillment.	Two domains did not apply to long-term care residents, so were removed.
Quality of Life in Late Stage Dementia scale	[70]	Rates observed resident behaviors on 11 items such as crying or appearing uncomfortable over the last week.	No comment.
DEMQOL[b] Proxy instrument	[38]	A 31-item scale completed by proxy to measure quality of life—social relationships items selected (getting in touch with people, having enough company, and being able to help other people).	Evidence supports the DEMQOL Proxy's internal consistency and test-retest reliability when used with people living with mild to moderate and severe dementia, and its content, convergent, and discriminate validity.
Quality of Life—Alzheimer's Disease	[67]	A 13-item questionnaire used to provide a self-reported indication of quality of life.	Multiple assessors may have introduced sources of variability.
Purpose and meaning			
Chinese Health Questionnaire	[86]	A 12-item scale measuring both participant health and mattering (the degree of being valued by significant others), meaning of life, loneliness, and depression.	Valid tool for use in older adults.
Purpose in Life Survey	[86]	Nine items (Likert scales).	Cronbach α coefficient was 0.87 in this study.
Social closeness			
Patient Reported Outcomes Measurement Information System Social Isolation Short Form 4a	[70]	Self-report 4-item scale captures social relationships.	Some of the items focus on domains typically impacted by institutional residence and may influence responses.
Inclusion of Other in Self Scale	[39,40]	Single-item graphical measure that shows 2 circles for self and others at various levels of distance until they substantially overlap. Participants indicate which one best represents the perceived interaction with the onsite companions.	No comment.
Quality of the Caregiving Relationship Questionnaire	[67]	Fourteen items with two subscales, warmth and absence of conflict, measuring quality of relationship between participant and relative.	Ceiling effect noted.
Social Support Behavior Scale	[77]	Three subscales—social support network, quantity of social support, and satisfaction with social support.	No comment.
Lubben Social Networking Scale	[63]	Self-report measure of the size of social networks with friends and family.	Recommended measure for assessing social networks among older adults.
Loneliness			
UCLA[c] Loneliness Scale	[77,80,86,96]	Self-report 3-item scale measuring loneliness.	Previous studies have shown this to be a reliable and valid measure of loneliness.
Agitation			

XSL•FO
RenderX

Instrument	Studies	Focus	Commentary on the tool from the authors
Cohen-Mansfield's Agitation Scale	[86]	Eight problematic behaviors were coded for frequency of occurrence on a scale of 0=did not occur to 3=occurred almost constantly. These included leaving one's seat, cursing or verbal aggression, requests for attention, asking repetitive questions or sentences, screaming, shouting or making strange noises, complaining, engaging in physical aggression to self or others, and exhibiting general restlessness.	Interrater reliability was established at >90% agreement among 3 independent coders. Mood prior to the study was observed to have an impact on agitation observed during the study, potentially skewing the results. Baseline agitation levels should be collected as a baseline to allow for comparison.
Depression and anxiety			
Center for Epidemiologic Studies Depression Scale	[75]	Self-report 8-item scale measuring depressive symptoms in the last week.	No comment.
Cornell Scale of Depression in Dementia	[63]	Clinician rated 19-item scale indexing mood and related signs, behavior disturbance, cyclic function, ideational disturbance, and physical signs of depression.	No comment.
Geriatric Depression Scale	[63,77,96]	Thirty items with a yes or no response.	The validity of the scale decreases when older adults have cognitive impairment.
Geriatric Depression Scale—Residential	[67]	A 12-item depression scale relevant to nursing and residential home populations.	No comment.
Chinese Geriatric Depression Scale	[86]	A 15-item scale measuring feelings of depression over the prior week.	Cronbach α coefficient was 0.90 in this study.
Geriatric Anxiety Inventory	[63]	A 29-item scale with yes or no responses relating to anxiety over the prior week.	Strong utility in detecting anxiety disorders.
Psychosocial capacity			
Mini-ICF[d] Rating for Impairment in psychological activities and capacities	[81]	Assesses psychosocial capacity in nursing homes through (1) adherence to regulations, (2) planning and structuring of tasks, (3) flexibility, (4) competence and knowledge application, (5) capacity to make decisions and judgments, (6) proactivity and spontaneous activities, (7) endurance, (8) self-assertiveness, (9) contact with others, (10) group integration, (11) intimate relationships, (12) self-care, and (13) mobility.	The tool is recommended in social medicine guidelines. Proven to be manageable, reliable, and valid in clinical practice.

[a]EuroQOL: EuroQOL consists of a Research Foundation and Group Association, which provides tools for measuring and valuing health that aim to improve decisions about health and health care globally.

[b]DEMQOL: Dementia Quality of Life is a patient-reported outcome measure, which is designed to enable the assessment of health-related quality of life of people with dementia.

[c]UCLA: University of California, Los Angeles.

[d]ICF: International Classification of Functioning, Disability and Health.

Engagement was measured using the Engagement of a Person With Dementia Scale [94], Observational Measurement of Engagement [94], Music in Dementia Assessment Scale [70], Engagement in Meaningful Activities Survey [70], Video Coding Incorporating Observed Emotion protocol [51], the Ethographic and Laban-Inspired Coding System of Engagement [88], and Visual Focus of Attention [103]. Well-being and satisfaction were measured through the Warwick-Edinburgh Mental Well-Being Scale [59], World Health Organization-Five Well-Being Index [81], the Satisfaction With Life Scale [75], and Ryff's Psychological Well-Being Scale [85]. Emotional response was assessed through the Observed Emotional Rating Scale [38,88], 4 items measuring enjoyment from the Game Experience Questionnaire [39,40], and the People Environment Apathy Scale [94]. Quality of life was measured using the EuroQOL EQ-5D-5 L [80], Multicultural Quality of Life Index [70], Quality of Life in Late Stage Dementia scale [70], the Dementia Quality of Life Proxy instrument [38], and the Quality of Life—Alzheimer's Disease scale [67]. Purpose and meaning were measured by the Chinese Health Questionnaire [86] and the Purpose in Life Survey [86]. Social closeness was measured using the Patient Reported Outcomes Measurement Information System Social Isolation Short Form 4a [70], the Inclusion of Other in Self Scale [39,40], the Quality of the Caregiving Relationship Questionnaire [67], the Social Support Behavior Scale [77], and the Lubben Social Networking Scale [63]. Loneliness was measured using the University of California, Los Angeles 3-Item Loneliness Scale [80]. Agitation was assessed through the Cohen-Mansfield's Agitation Scale [51,72]. Depression was measured using the Center for Epidemiologic

XSL•FO
RenderX

Studies Depression Scale [75], the Cornell Scale of Depression in Dementia [63], the Geriatric Depression Scale [63,77,96], the Geriatric Depression Scale—Residential [67], and the Chinese Geriatric Depression Scale [86]. Anxiety was measured using the Geriatric Anxiety Inventory [63]. Psychosocial capacity was measured using the Mini-ICF Rating for Limitations of Activities and Participation in Psychological Disorders [81], which includes subscales assessing group integration and intimate relationships.

Instruments included a mixture of self-report scales and observation of behavior. Observational tools were more commonly seen when residents also had a diagnosis of dementia. While some authors did not justify their choice of instruments, others provided justification through the provision of evidence on their validity and reliability in similar contexts. Challenges to the use of the tools were also recorded, including potential inconsistencies between different raters [35,67], the impact of the long-term care setting on the accuracy of the results [70], and the conceptual gaps around interpreting affective facial expressions, connectedness, and social inclusion [51].

Factors That Support Meaningful Connection

Overview

A narrative summary was developed to explore how the included studies reported the factors underpinning meaningful connections during the use of digital technology. Five key themes were identified: getting to know the person, increases autonomy and choice, source of enjoyment and fun, facilitates communication, and builds community. Table 4 identifies the studies associated with each theme.

Table 4. Summary of papers associated with factors that support meaningful connections.

Theme and subtheme	Papers
Getting to know the person	
Reminiscence or life story work	[33-36,41,47,49,53,55,57,67,70,71,75,77,78,81-84,87,88,93,94]
Seeing the patient as a person	[36-38,43,46,55,79,84,86,90]
Increases autonomy and choice	
Autonomy and control	[34,52,56,64,70]
Personal choice valued	[34,48,58,74,79,81,84,104]
Source of enjoyment and fun	
Technology was fun	[39-41,52,53,55,56,59,64,65,70,73,74,81,88,89,91,92,104]
Space for creativity	[34,56,70,73,79]
Facilitates communication	
Communicating with a facilitator or HCP[a]	[31,37,46,50,52-54,56-58,63,68,81,88,90,93,96]
Communicating with family	[32,33,36,38,41,43-47,49,50,52,60-62,67,75-78,93]
Builds community	
Shared social activities within the care home	[39,40,51,52,56,63,64,66,71,74,76,78,81,87,89,93,99-103]
Intergenerational and wider community connection	[39,40,42,49,50,71,74,75,80,104]

[a]HCP: health care professional.

Getting to Know the Person

Reminiscence and life story work were identified as a key mechanism through which meaningful connections were formed by providing an opportunity to get to know the person as an individual without focusing on their medical condition. Reminiscence was enabled through a range of activities from virtual cycling by reminiscing on experiences of cycling earlier in life [83] to VR photo albums and videos [85]. Technology supported less pressured nonverbal interactions, allowing personal preferences to emerge, helping practitioners to get to know individuals more effectively [54], and offering multiple channels through which people could express their identity [52]. Shifting from patient to person was seen not only in reminiscence-specific technologies such as MEMENTO, a spontaneous reminiscence therapy smartwatch [58], but also in the use of robotics such as the social sheep [88,94,95] and in the development of personalized multimedia spaces through life story work such as SENSE-Garden [79]. Residents enjoyed when activities led to an increased feeling of value and a sense of being treated seriously, for example, when combining 3D VR with hands-on horticultural therapy [86]. Getting to know the individual resident led to a change in the care culture, for example, gaining deeper insights into residents' life stories to facilitate SENSE-Garden fostered a more profound personal connection and shift in focus to a person-centered approach.

Increases Autonomy and Choice

Digital technologies provided an opportunity for residents to feel in control, empowered, and independent while having their autonomy and personal choice respected, creating spaces for meaningful connections to form. Many of the technologies, for example, virtual talks [34], SENSE-Garden [79], VR Vacation Home [81], and Virtual Creative Ageing Program [70], all promoted a sense of control and autonomy over the activity, which was reported as a key element that promoted meaningful

XSL•FO
RenderX

connections. Computer kiosks provided an opportunity for personal development while being in control, which helped to increase residents' feelings of self-worth and esteem as well as connecting socially [64]. Lifestyle enrichment was also noted in the study by Webber et al [48], which used technology to facilitate virtual visits to places of personal significance. Residents could also meaningfully connect with their families without needing staff as gatekeepers [50]. Promoting ability instead of focusing on limitations created positive experiences for residents and staff.

Source of Enjoyment and Fun

To build meaningful connections, this theme identified the importance of digital technologies as a source of enjoyment and fun for the users. Digital technology was described as playful [53,55,59,89], pleasurable [65], fun [39,40], and enjoyable [52,88,103]. The digital technology created a space to express creativity [70], engage imagination [56], socialize [43], and break routine [34], which relieved stress and boredom for residents. Robotics in particular was identified as a source of shared joy, such as Pepper, the social semihumanoid robot [91], and Paro, the baby harp seal [92].

Facilitates Communication

Digital technologies were often implemented with facilitation from a health care professional or study-specific facilitator. The process of engaging with the facilitator was identified as a key opportunity to connect meaningfully, for example, during the delivery of digital cognitive stimulation therapy [31] or the introduction of the interactive sound cushion VITA [53]. Spending more time with staff while engaging with technology was reported positively by residents [68]. Technology supported meaningful communication with the facilitator, enabling trust and reducing misunderstandings [37]. Robotics such as pet cats and dogs provided something positive to talk about between the residents and health care professionals, opening up a channel of communication [96]. In addition to increasing communication with a facilitator, some technologies increased communication with families, for example, teleconferencing [44,45] and video calling relatives [33,41,47,77]. Apps such as Memory Keeper acted as a conversation starter [36], and new devices, such as a digital television for reminiscence [67], VENSTER interactive artwork, encouraged conversation about a present activity [57]. KOMP supported the maintenance of social bonds through family interaction [60,61], and a digital photo frame improved social connectiveness with family [62].

Builds Community

Opportunities were created to engage in a social activity, which supported residents to find others with shared interests and develop a sense of community. Meaningful social relationships and small group interactions were enabled through the use of online social networking in the study by Ysseldyk et al [75] and the use of a Computer Interactive Reminiscence and Communication Aid in the study by Samuelsson et al [76]. The introduction of new devices such as an interactive table [51], PARO the baby harp seal [100], virtual visits to museums [39,40], technology-assisted music making [66], and shared activities such as SitDance [74] encouraged interaction between residents, staff, and relatives through a shared social activity, reducing loneliness. VR Take a Cycle With Me supported intergenerational rapport and cohesiveness, which fed into an increased sense of community [80], and adaptive Motitech bikes provided a purpose and challenge while building a sense of community with other residents in an inclusive way [63]. Technology improved participation for people living with dementia, with robotic cats and dogs promoting an inclusive community for all abilities in care [90], zoomorphic robots initiating conversations [99], robot-led activities improving engagement [101], and robot-facilitated dance sessions improving group valence [103].

Discussion

What This Study Adds

This scoping review explored the findings from 72 papers that identified the active digital technologies used in care homes to help build meaningful connections, as well as the factors that underpin the formation of meaningful connections and the range of outcome measures used to assess their impact. Based on our search of CINAHL, MEDLINE, and *JBI Evidence Synthesis* up to December 12, 2025, we did not identify any previous scoping review that looks at meaningful connections across the care home, without focusing on a specific condition or clinical outcome. Five main types of digital technology were identified, including mobile apps, VR and immersive technology, robotics, digital devices and ambient technology, and online or computer-assisted programs. A narrative summary described the factors through which meaningful connections were developed, and these included getting to know the person, increases sense of autonomy and choice, source of enjoyment and fun, facilitates communication, and builds community. While most studies relied on some form of qualitative analysis or observation to identify changes in meaningful connections, a wide variety of standardized instruments were also used, which looked for changes in engagement, well-being or satisfaction, emotional response, quality of life, purpose and meaning, social closeness, loneliness, depression and anxiety, and psychosocial capacity.

Comparison of Findings With Existing Literature

The findings from this scoping review concluded that meaningful connections were facilitated through the creation of opportunities for the residents to be known and heard as a person, and the development of group and shared spaces where people could share an enjoyable experience and an increased sense of community. Effective digital technologies amplified human interactions instead of replacing them, enabling increased interaction with family, friends, and community outside of the care home, shared activities and hobbies with other residents in the care home, and opportunities for staff and facilitators to get to know the resident on a deeper level. These findings align with a community-based implementation project of social robots with older adults, which found that social robots facilitated person-person interactions and inspired participants to share experiences with their family and friends after the engagement session [105]. Fostering and maintaining meaningful connections is at the core of social health, which is defined as an adequate

XSL•FO
RenderX

quantity and quality of relationships in a particular context to meet an individual's need for meaningful human connection [3]. These findings align with wider research into the impact of digital technology on patient experience, which reported improvements in patient experience and improved quality of care [106].

While digital technology plays a key role in addressing social health by acting as a catalyst for forming meaningful human connections, caution is also due when considering the potentially harmful effects of overreliance on artificial intelligence companions that could replace social human interaction [107]. Warm technology, which aims to enhance human potential and social connectedness [108], has been developed primarily within the field of dementia care, but could be further developed and tested for application across the wider care home population. Furthermore, focus should be given to ensuring that both staff and residents are familiar with and comfortable using such technologies [109]. However, it is also important to remember that while many of the technologies included a social element, not all meaningful connections were inherently social. Some technologies such as the MEMENTO smartwatch [58] encouraged companionship through spontaneous meaningful conversations and enhanced staff access to residents' stories through reminiscence. SENSE-Garden created space for interactions between staff and residents, during which they engaged with life stories, focused on abilities more than limitations, and supported a culture shift from patient to person [79]. This highlights the fundamental importance of technology as a facilitator of individual meaningful connection, as well as a facilitator of social connection and community. While some studies support the integration of technology-driven solutions into caregiving procedures to promote social health among nursing home residents with dementia [102], it is also important to consider the potential practical, clinical, and ethical limitations of the introduction of such technologies. From a practical perspective, research has shown that the use of technology for social connection in aged care requires careful facilitation from staff, better resourcing and infrastructural support, collaborations with volunteers, and more attention to individual needs [110]. Clinically guided, emotional warmth, complex social interactions, and empathy are essential in the care of older adults [111]. This may be reduced or removed if the care sector becomes overreliant on technology to meet social health needs. Ethically, the use of social robots should be guided by the same ethical guidelines followed in all therapeutic interventions [112]; however, it was noted in a recent qualitative analysis of research into robotics in health care of older adults that limited attention is given to ethical and legal aspects of care robots [113].

The application of such active digital technologies to social health sits in contrast to the use of passive technologies, which improve safety and efficiency but may inadvertently reinforce a surveillance-oriented model and add little to the social well-being of residents in care homes [114]. A scoping review of digital technologies supporting aged care services more broadly found 7 key roles of digital technology, including health monitoring and assessment, remote health care services, assistive technology to support treatment, self-care management, social technology to facilitate interaction, clinical decision support, and aged care quality measurement [115]; however, only social robots to provide better communication were discussed. When looking at policies focused on digital health, the core messages around technology use in health care move toward clinical applications, such as diagnosis, decision-making, virtual care, remote monitoring, remote data capture, and clinical trials within the Global Strategy on Digital Health 2020-2025 [116]. The focus is once again on more passive clinical monitoring than the application of active digital technology to social well-being. This trend continues in new research on generative artificial intelligence (GenAI), which, despite the capacity to address social health concerns, is currently focused on increasing task efficiency such as scheduling and care planning [117]. Alternatively, active technologies have a greater potential to promote engagement and connection [118,119]. While their effectiveness is often dependent on family or professional caregivers to help set up and use the technology to ensure meaningful use rather than superficial engagement [120], the involvement of caregivers is, in itself, a mechanism through which more meaningful connections can be achieved.

The scoping review showed how the types of digital technologies used for meaningful connection have changed over time and have exponentially increased in the last 3 years. Smart Home technologies have been shown to significantly enhance safety and monitoring capabilities for adults living independently [121]; however, this overlooks the social health of older adults. A current area for future exploration is the application of GenAI in care homes to promote meaningful connections. GenAI, while originally appearing in physical form through the chatbot Eliza in the 1960s, has recently boomed in popularity in the public sphere through chatbots such as ChatGPT, and more recently again through its application to health care [122]. Currently, the primary application in clinical practice is to support health care, planning, and information searching [117]. For example, one study successfully applied GenAI to reduce emergency incidents and help daily activities run more smoothly within assisted living environments [123]. Another study explored how GenAI models can improve decision accuracy among emergency nurses; however, they noted the lack of contextual sensitivity in emergency care [124]. Within nursing practice, GenAI is being embedded in patient care, supporting workflow integration, and providing operational support [125]. In gerontological nursing education, GenAI has been integrated with art therapy to support the development of empathy in students [125]. In community care, artificial intelligence–powered reflection tools have been shown to improve communication between older adults and caregivers [126], to support older adults with health information tasks [127,128], and to provide empathetic responses to online queries related to Alzheimer disease and related dementias [129].

However, GenAI with the goal of building meaningful connections has received little attention, specifically in the long-term care setting, which has additional barriers to meaningful connections. The HARMONEE (Harnessing Artificial Intelligence Resources for Mental Well-Being for Older Adults and Nurturing Empathy in Education) project is a novel example of how GenAI can be used to support

XSL•FO
RenderX

reminiscence with older adults in long-term care to improve their mood [119]; however, its dependence on a research team to facilitate generates issues with the sustainability. Technology must be designed so it can be used by the resident or care team independently. Additionally, to ensure that technology is fit for purpose and suited to the unique setting of the care home, care home residents must be part of decision-making when it comes to co-designing and implementing new technologies in order to defy techno-solutionist scripting of dehumanizing images of fragility, dependency, and technical ineptitude embedded in product designs [130]. Research has indicated an openness of older adults to new technologies for physical assistive functions [131]; however, further exploration is needed to understand their attitudes toward technologies aimed at supporting meaningful connections.

This review identified that the measurement of meaningful connection was inconsistent across studies. This may be attributable to the varying operational definitions of meaningful connections, activities, and engagement across these studies. Qualitative analysis and observation were most frequently used to assess concepts associated with meaningful connections. Studies that used scales and instruments considered a wide variety of measures including engagement, well-being or satisfaction, emotional response, quality of life, purpose and meaning, social closeness, loneliness, depression and anxiety, and psychosocial capacity. This highlights a broader challenge around defining and measuring meaningful connections in the literature. It could be argued that meaningful connections in themselves are not measured specifically, but instead outcome measures focused on indicators of meaningful connections (eg, well-being, loneliness, and quality of life). When exploring social connection, Holt-Lunstad [2] discusses similar terminological challenges; for example, loneliness is often used to measure social connection; however, it is defined as a distressing experience resulting from perceived inadequate meaningful connections; therefore, it does not fully capture the breadth of social connection. Similarly with meaningful connections, a clear consensus has not been reached on a definition or approach to measurement. Measurement of social connection is being progressed, for example, through the development of the Connection During Conversations Scale [132]; however, with the focus being on social connection during social interactions, it is possible that it overlooks some of the meaningful connections occurring outside social spaces.

Strengths and Limitations

By adopting a scoping review approach, a comprehensive overview of the broad field of digital technology within care homes could be provided, helping to explore the scope of existing knowledge and identify gaps in the evidence base to highlight for future research. The scoping review was conducted with strict adherence to the JBI methodology and included multiple authors at each stage to minimize the risk of individual bias. However, limitations are also noted. While the language of meaningful connections was left as broad as possible within the search strategy, it is possible that papers with relevant information around social connection or person-centered care delivery were missed. Given the time and resource restraints within the project, the decision was made to limit the search to the English language. However, language bias may result in missing important cultural contexts in non-English–speaking settings, limiting the relevance of findings in some geographical locations. The main focus of this review was on active digital technologies, excluding passive monitoring technologies. However, these technologies contribute to care home life through their influence on care practices, staff time and availability, and safety perceptions that directly impact opportunities for connection. It is possible that passive monitoring technologies could increase staff availability and create a safer care environment, which would lead to increased opportunities for meaningful connections. Through their omission from the review, only a partial understanding of the digital landscape within care homes has been explored. Finally, the heterogeneity of outcome measures reported across the studies limits the comparability of studies and restricts the synthesis of these findings. Clear conclusions regarding the measurement of meaningful connections could not be drawn; however, this did highlight a gap for future research.

Recommendations and Real-World Implications

At present, technologies used within aged care seem to not be meeting end user needs [133]. While technology design for older adults is often focused for family members or caregivers to allow passive monitoring, giving control back to older adults nurtures their sense of autonomy [134]. Future technologies must also be designed with social health in mind, considering both the engagement of the individual and the community. This review has highlighted how technology can facilitate both individual meaningful connections as well as social connections and a sense of community. However, to ensure technology is fit for purpose and suited to the unique setting of the care home, care home residents must be part of decision-making when it comes to co-designing and implementing new technologies in order to defy the dehumanizing images of fragility, dependency, and technical ineptitude embedded in product designs [130]. Furthermore, future research needs to consider the cultural environment of the care home when embedding new active digital technologies with the goal of building meaningful connections, as research has shown that how the cultural environment has the most significant influence on nursing home residents' leisure and social occupational choice [135].

Future research needs to consider the development and psychometric testing of a standardized tool to measure meaningful connection more broadly, as well as in relation to the digital technologies implemented. Consideration needs to be given to both individual and social connection indicators, as well as emotional response, engagement, satisfaction, loneliness, anxiety, depression, and quality of life. There is also a gap specifically around meaningful connections within care homes. The challenge raised with some of the instruments used to collect outcome measures was the uniqueness of the care home setting. Challenges in fostering meaningful engagement in care homes include a lack of understanding of residents' values and interests, staff needing to manage multiple requirements simultaneously, and staff not receiving support to involve residents [136]. Some of the items in the scales were not reflective of, or relevant to, long-term care settings, leading to the need for modifications or the acknowledgment that this may

XSL•FO
RenderX

have impacted on the findings of the study. Future research needs to consider the development of a bespoke instrument that considers meaningful connections within the context of the care home environment.

Research in this space is rapidly expanding, aligned with advances in digital technology availability. However, few studies that focused on meaningful connections considered an implementation design. The reliance on qualitative and observational methods may link to the challenges in measurement or the infancy of the technology; however, future research needs to consider the development of implementation studies, considering broader factors such as cost-effectiveness and organizational readiness. The TIDieR (Template for Intervention Description and Replication) checklist [137] may structure the reporting of pre- and postinterventions from other settings, allowing them to be replicated in the care home context.

Conclusions

This scoping review has presented a detailed exploration of how digital technologies have been used within care homes to facilitate the development of meaningful connections between residents, families, staff, and communities. While the wider literature indicates that many digital technologies overlook social health and well-being in care homes, this review highlights how a range of technologies including mobile apps, VR and immersive technology, robotics, digital devices and ambient technology, and online or computer-assisted programs can be applied to facilitate meaningful connections within care homes. Digital technology is a catalyst for meaningful human connection, which is a key element of social health, rather than a replacement. Further research is needed to explore how digital technologies can support individual meaningful connections with residents as well as enhance the sense of community through shared social activities and spaces. The application of GenAI to the development of meaningful connections and social health in care homes is a current gap in the evidence base. Moving forward, the development of GenAI-based technologies needs to be undertaken in conjunction with care home residents, staff, and communities to ensure that they meet the social needs of the entire community. An ongoing challenge is found in the measurement of the effectiveness of digital technologies to enhance meaningful connections. Future research needs to consider how new evaluation metrics can be developed that combine digital engagement data with validated measures of social and meaningful connection to assess the person-centered impact of digital technologies in care homes.

Acknowledgments

The authors thank the participants of the My Home Life Northern Ireland Leadership Support Programme who helped identify the gap that led to this scoping review and the librarian Kelly Coogan who supported the team during database searching. This manuscript has been prepared without assistance from generative artificial intelligence tools.

Funding

No funding was received to undertake this review. The paper processing fee was paid through the Ulster-JMIR Open Access deal.

Data Availability

All data generated and analyzed during this study are included in this published paper and Multimedia Appendix 3.

Authors' Contributions

Conceptualization: DM, RB
Data curation: DM, RB, KC
Formal analysis: DM (lead), RB (supporting), KC (supporting)
Methodology: DM
Project administration: DM
Supervision: DM
Validation: DM
Visualization: DM, RB
Writing—original draft: DM (lead), RB (supporting)
Writing—review and editing: DM (lead), RB (supporting), KC (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA ScR checklist.
[PDF File (Adobe PDF File), 172 KB - jmir_v28i1e88181_app1.pdf]

XSL•FO
RenderX

Multimedia Appendix 2
Search strings per database from search on December 12, 2025.
[DOCX File , 18 KB - jmir_v28i1e88181_app2.docx]

Multimedia Appendix 3
Data summary table.
[XLSX File (Microsoft Excel File), 35 KB - jmir_v28i1e88181_app3.xlsx]

References

1. Bethell J, Aelick K, Babineau J, Bretzlaff M, Edwards C, Gibson JL, et al. Social connection in long-term care homes: a scoping review of published research on the mental health impacts and potential strategies during COVID-19. J Am Med Dir Assoc 2021;22(2):228-237.e25 [FREE Full text] [doi: 10.1016/j.jamda.2020.11.025] [Medline: 33347846]
2. Holt-Lunstad J. Social connection as a critical factor for mental and physical health: evidence, trends, challenges, and future implications. World Psychiatry 2024 Oct;23(3):312-332 [FREE Full text] [doi: 10.1002/wps.21224] [Medline: 39279411]
3. From loneliness to social connection: charting a path to healthier societies. World Health Organization. 2025. URL: https://www.who.int/publications/i/item/978240112360 [accessed 2025-09-18]
4. Ageing and health. World Health Organization. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health [accessed 2025-09-19]
5. Ageing and long-term care. Organisation for Economic Co-operation and Development. 2024. URL: https://www.oecd.org/en/topics/sub-issues/ageing-and-long-term-care.html [accessed 2025-09-19]
6. Llena-Nozal A, Barszczewski J, Rauet-Tejeda J. How do countries compare in their design of long-term care provision? A typology of long-term care systems. OECD Health Working Papers. 2025 Jun 26. URL: https://www.oecd.org/en/publications/how-do-countries-compare-in-their-design-of-long-term-care-provision_44f5453a-en.html [accessed 2026-02-11]
7. Ettelt S, Williams L, Damant J, Perkins M, Wittenberg R. What kind of home is your care home? A typology of personalised care provided in residential and nursing homes. Ageing Soc 2020;42(5):993-1013. [doi: 10.1017/s0144686x20001142]
8. Boamah SA, Weldrick R, Lee TJ, Taylor N. Social isolation among older adults in long-term care: a scoping review. J Aging Health 2021;33(7-8):618-632 [FREE Full text] [doi: 10.1177/08982643211004174] [Medline: 33779366]
9. Swinton J, Vilapakkam Nagarajan S, Atee M, Morris T, Poulos CJ. Being present and meaningful engagement for aged care residents living with dementia: a mixed-methods evaluation of Australian care worker's experiences. Gerontologist 2024;64(4):gnad131 [FREE Full text] [doi: 10.1093/geront/gnad131] [Medline: 37777877]
10. QS50: Mental wellbeing of older people in care homes: quality standards. National Institute for Health and Care Excellence. 2013. URL: https://www.nice.org.uk/guidance/qs50/chapter/quality-statement-1-participation-in-meaningful-activity [accessed 2025-09-19]
11. Groenendaal M, Loor A, Trouw M, Achterberg WP, Caljouw MA. Perspectives of healthcare professionals on meaningful activities for persons with dementia in transition from home to a nursing home: an explorative study. Healthcare (Basel) 2019;7(3):98 [FREE Full text] [doi: 10.3390/healthcare7030098] [Medline: 31430855]
12. Hooker SA, Masters KS, Vagnini KM, Rush CL. Engaging in personally meaningful activities is associated with meaning salience and psychological well-being. J Posit Psychol 2019;15(6):821-831. [doi: 10.1080/17439760.2019.1651895]
13. Travers C, Brooks D, Hines S, O'Reilly M, McMaster M, He W, et al. Effectiveness of meaningful occupation interventions for people living with dementia in residential aged care: a systematic review. JBI Database System Rev Implement Rep 2016;14(12):163-225. [doi: 10.11124/JBISRIR-2016-003230] [Medline: 28009676]
14. Framework on integrated, people-centred health services. World Health Organization. 2016. URL: https://apps.who.int/gb/ebwha/pdf_files/WHA69/A69_39-en.pdf?ua=1%26ua=1 [accessed 2025-09-19]
15. Smith N, Towers AM, Palmer S, Beecham J, Welch E. Being occupied: supporting ‘meaningful activity’ in care homes for older people in England. Ageing Soc 2017;38(11):2218-2240. [doi: 10.1017/s0144686x17000678]
16. Kemp CL, Bender AA, Morgan JC, Burgess EO, Epps FR, Hill AM, et al. Understanding capacity and optimizing meaningful engagement among persons living with dementia. Dementia (London) 2023;22(4):854-874 [FREE Full text] [doi: 10.1177/14713012231162713] [Medline: 36913646]
17. Gonnord T, Clarys D, Boucard G, Esnard C. Positive impact of social relationships fostered by physical and/or cognitive group activity on older people's quality of life: PRISMA systematic review. Front Psychol 2023;14:1166072 [FREE Full text] [doi: 10.3389/fpsyg.2023.1166072] [Medline: 37767210]
18. Goodwin VA, Hall AJ, Bethel A, Taylor E, Tatnell L, Turner M, et al. Understanding joy amongst older people: a scoping review. Arch Gerontol Geriatr 2023;115:105118 [FREE Full text] [doi: 10.1016/j.archger.2023.105118] [Medline: 37516061]
19. Kang B, Scales K, McConnell ES, Song Y, Lepore M, Corazzini K. Nursing home residents' perspectives on their social relationships. J Clin Nurs 2020;29(7-8):1162-1174 [FREE Full text] [doi: 10.1111/jocn.15174] [Medline: 31889360]
20. Lim E, Nielsen N, Lapane L, Barooah A, Xu S, Qu S, et al. Health effects of social connectedness in older adults living in congregate long-term care settings: a systematic review of quantitative and qualitative evidence. Int J Older People Nurs 2023;18(6):e12577. [doi: 10.1111/opn.12577] [Medline: 37803996]

XSL•FO
RenderX

21. Townsend BG, Chen JTH, Wuthrich VM. Barriers and facilitators to social participation in older adults: a systematic literature review. Clin Gerontol 2021;44(4):359-380. [doi: 10.1080/07317115.2020.1863890] [Medline: 33393443]
22. Nygaard A, Halvorsrud L, Grov EK, Bergland A. 'What matters to you?'—a qualitative study on the views of nursing home residents with dementia regarding the health care they receive. J Clin Nurs 2022;31(1-2):262-274. [doi: 10.1111/jocn.15904] [Medline: 34096110]
23. Wright PJ, Raynor PA, Bowers D, Combs EM, Corbett CF, Hardy H, et al. Leveraging digital technology for social connectedness among adults with chronic conditions: a systematic review. Digit Health 2023;9:20552076231204746 [FREE Full text] [doi: 10.1177/20552076231204746] [Medline: 37799504]
24. Chu CH, Yee A, Stamatopoulos V. Poor and lost connections: essential family caregivers' experiences using technology with family living in long-term care homes during COVID-19. J Appl Gerontol 2022;41(6):1547-1556 [FREE Full text] [doi: 10.1177/07334648221081850] [Medline: 35416076]
25. Ibarra F, Baez M, Cernuzzi L, Casati F. A systematic review on technology-supported interventions to improve old-age social wellbeing: loneliness, social isolation, and connectedness. J Healthc Eng 2020;2020:2036842 [FREE Full text] [doi: 10.1155/2020/2036842] [Medline: 32765823]
26. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018;18(1):143 [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
27. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
28. Peters M, Godfrey C, McInerney P, Munn Z, Trico A, Khalil H. Chapter 11: Scoping reviews. In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis. Adelaide: JBI; 2020.
29. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
30. Melnyk BM, Fineout-Overholt E. Evidence-Based Practice in Nursing & Healthcare: A Guide to Best Practice. Philadelphia, PA: Wolters Kluwer Health; 2015.
31. Abdalrahim A, ALBashtawy M, Alkhawaldeh A, Ayed A. Examining the feasibility and acceptability of digital cognitive stimulation therapy for dementia care in Jordan: a qualitative study. SAGE Open Nurs 2024;10:23779608241272599. [doi: 10.1177/23779608241272599] [Medline: 39314644]
32. Barbosa Neves B, Franz R, Judges R, Beermann C, Baecker R. Can digital technology enhance social connectedness among older adults? A feasibility study. J Appl Gerontol 2019;38(1):49-72. [doi: 10.1177/0733464817741369] [Medline: 29166818]
33. Chapman H, Bethell J, Dewan N, Liougas MP, Livingston G, McGilton KS, et al. Social connection in long-term care homes: a qualitative study of barriers and facilitators. BMC Geriatr 2024;24(1):857 [FREE Full text] [doi: 10.1186/s12877-024-05454-8] [Medline: 39438797]
34. Gleeson E, Early E. Bringing the outside in through facilitated communication technology in long term care: how do facilitated, virtual communication interventions influence peer interaction, social engagement and participant well-being? Irish Association for Applied Linguistics. 2021. URL: https://hdl.handle.net/10468/11698 [accessed 2025-09-19]
35. Hird N, Osaki T, Ghosh S, Palaniappan SK, Maeda K. Enabling personalization for digital cognitive stimulation to support communication with people with dementia: pilot intervention study as a prelude to AI development. JMIR Form Res 2024;8:e51732 [FREE Full text] [doi: 10.2196/51732] [Medline: 38227357]
36. McAllister M, Dayton J, Oprescu F, Katsikitis M, Jones CM. Memory Keeper: a prototype digital application to improve engagement with people with dementia in long-term care (innovative practice). Dementia (London) 2020;19(4):1287-1298. [doi: 10.1177/1471301217737872] [Medline: 29096546]
37. Munk S, Toohey R, Remtilla A, D'Cunha NM, Gibson D, Isbel S, et al. "I forgot she used to make chocolate cake": digital storytelling supporting person-focussed dementia care: a qualitative thematic analysis. Dementia (London) 2025;24(4):666-696 [FREE Full text] [doi: 10.1177/14713012251317761] [Medline: 39876485]
38. O'Rourke HM, Swindle J, Chacinski D, Dal Pizzol FLF, Lee H, Hoben M, et al. Connecting Today: feasibility and acceptability of a remote visiting program for people living with dementia in long-term care homes. Dementia (London) 2023;22(7):1321-1347 [FREE Full text] [doi: 10.1177/14713012231176858] [Medline: 37341515]
39. Pisoni G. Mediating distance: new interfaces and interaction design techniques to follow and take part in remote museum visits. J Syst Inform Technol 2020;22(4):329-350. [doi: 10.1108/jsit-03-2020-0038]
40. Pisoni G, Daniel F, Casati F, Callaway C, Stock O. Interactive remote museum visits for older adults: an evaluation of feelings of presence, social closeness, engagement, and enjoyment in an social visit. 2019 Presented at: IEEE International Symposium on Multimedia (ISM); December 9-11, 2019; San Diego, CA, United States p. 99-102 URL: https://ieeexplore.ieee.org/document/8959064 [doi: 10.1109/ISM46123.2019.00023]
41. Sävenstedt S, Brulin C, Sandman PO. Family members' narrated experiences of communicating via video-phone with patients with dementia staying at a nursing home. J Telemed Telecare 2003;9(4):216-220. [doi: 10.1258/135763303322225544] [Medline: 12952692]

XSL•FO
RenderX

42. Welsh D, Morrissey K, Foley S, McNaney R, Salis C, McCarthy J, et al. Ticket to Talk: supporting conversation between young people and people with dementia through digital media. 2018 Presented at: CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems; April 21-26, 2018; Montreal, QC, Canada p. 1-14. [doi: 10.1145/3173574.3173949]
43. Muñoz D, Pedell S, Sterling L. Evaluating engagement in technology-supported social interaction by people living with dementia in residential care. ACM Trans Comput-Hum Interact 2022;29(5):1-31. [doi: 10.1145/3514497]
44. Hardy MS, Fanaki C, Savoie C, Dallaire C, Wilchesky M, Gallani MC, et al. Acceptability of videoconferencing to preserve the contact between cognitively impaired long-term care residents and their family caregivers: a mixed-methods study. Geriatr Nurs 2022;48:65-73 [FREE Full text] [doi: 10.1016/j.gerinurse.2022.09.006] [Medline: 36155311]
45. Hardy MS, Fanaki C, Savoie C, Wilchesky M, Gagnon MP, Laberge M, et al. Long-term care staffs' experience in facilitating the use of videoconferencing by cognitively impaired long-term care residents during the COVID-19 pandemic: a mixed-methods study. BMC Health Serv Res 2024;24(1):646 [FREE Full text] [doi: 10.1186/s12913-024-11095-9] [Medline: 38769512]
46. Munoz D, Favilla S, Pedell S, Murphy A, Beh J, Petrovich T. Evaluating an app to promote a beter visit through shared activities for people living with dementia and their families. 2021 Presented at: CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; May 8-13, 2021; Yokohama, Japan p. 1-13. [doi: 10.1145/3411764.3445764]
47. Thach KS, Lederman R, Waycott J. Key considerations for the design of technology for enrichment in residential aged care: an ethnographic study. 2023 Presented at: CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems; April 23-28, 2023; Hamburg, Germany p. 1-16. [doi: 10.1145/3544548.3581176]
48. Webber S, Baker S, Waycott J. Virtual visits: reminiscence in residential aged care with digital mapping technologies. Australas J Ageing 2021;40(3):293-300. [doi: 10.1111/ajag.12902] [Medline: 33538405]
49. Akhtar SA. Socially connected while apart: the use of technology to increase social connection between nursing home residents and their relatives. Front Public Health 2024;12:1296524 [FREE Full text] [doi: 10.3389/fpubh.2024.1296524] [Medline: 38333742]
50. Badawy A, Solberg M, Obstfelder AU, Alnes RE. Improvised use of a digital tool for social interaction in a Norwegian care facility during the COVID-19 pandemic: an exploratory study. BMC Health Serv Res 2022;22(1):136 [FREE Full text] [doi: 10.1186/s12913-022-07526-0] [Medline: 35105344]
51. Feng Y, van RR, Yu S, Hu J, Rauterberg M, Barakova E. Using observational engagement assessment method VC-IOE for evaluating an interactive table designed for seniors with dementia. 2017 Presented at: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); June 26-27, 2017; Hong Kong, China p. 26-37. [doi: 10.1007/978-3-319-67964-8_3]
52. Grigorovich A, Marcotte AA, Colobong R, Szabo M, MacNeill C, Blais D, et al. Using voice-activated technologies to enhance well-being of older adults in long-term care homes. Innov Aging 2024;8(12):igae102 [FREE Full text] [doi: 10.1093/geroni/igae102] [Medline: 39664605]
53. Houben M, Brankaert R, Dhaeze E, Kenning G, Bongers I, Eggen B. Enriching everyday lived experiences in dementia care. 2022 Presented at: TEI '22: Proceedings of the Sixteenth International Conference on Tangible, Embedded, and Embodied Interaction; February 13-16, 2022; Daejeon, Republic of Korea p. 1-13. [doi: 10.1145/3490149.3501326]
54. Houben M, Brankaert R, Bakker S, Kenning G, Bongers I, Eggen B. The role of everyday sounds in advanced dementia care. 2020 Presented at: Conference on Human Factors in Computing Systems, CHI 2020; April 25-30, 2020; Honolulu p. 14. [doi: 10.1145/3313831.3376577]
55. Houben M, Brankaert R, Kenning G, Eggen B, Bongers I. The perspectives of professional caregivers on implementing audio-based technology in residential dementia care. Int J Environ Res Public Health 2020;17(17):6333 [FREE Full text] [doi: 10.3390/ijerph17176333] [Medline: 32878116]
56. Jamin G, Luyten T, Delsing R, Braun S. The process of co-creating the interface for VENSTER, an interactive artwork for nursing home residents with dementia. Disabil Rehabil Assist Technol 2018;13(8):809-818. [doi: 10.1080/17483107.2017.1385102] [Medline: 29037109]
57. Luyten T, Braun S, Jamin G, van Hooren S, de Witte L. How nursing home residents with dementia respond to the interactive art installation 'VENSTER': a pilot study. Disabil Rehabil Assist Technol 2018;13(1):87-94 [FREE Full text] [doi: 10.1080/17483107.2017.1290701] [Medline: 28287047]
58. Paay J, Kjeldskov J, Aaen I, Bank M. User-centred iterative design of a smartwatch system supporting spontaneous reminiscence therapy for people living with dementia. Health Informatics J 2022;28(2):14604582221106002 [FREE Full text] [doi: 10.1177/14604582221106002] [Medline: 35658692]
59. van Wingerden E, Vacaru SV, Holstege L, Sterkenburg PS. Hey Google! Intelligent personal assistants and well-being in the context of disability during COVID-19. J Intellect Disabil Res 2023;67(10):973-985. [doi: 10.1111/jir.13064] [Medline: 37317950]
60. Akhtar SA. Together through technology: next of kin's use of technology to maintain social connection with nursing home residents. Ageing Soc 2025;45(11):2339-2355. [doi: 10.1017/s0144686x25000054]

XSL•FO
RenderX

61. Badawy A, Solberg M, Obstfelder AU, Alnes RE. Together, at a distance: experiences with a novel technology for social contact among older people and their relatives in Norway during the COVID-19 pandemic. BMC Geriatr 2023;23(1):218 [FREE Full text] [doi: 10.1186/s12877-023-03869-3] [Medline: 37024822]
62. Biemans M, Van Dijk B. Food for talk: photo frames to support social connectedness for elderly people in a nursing home. 2009 Presented at: ECCE '09: European Conference on Cognitive Ergonomics: Designing beyond the Product --- Understanding Activity and User Experience in Ubiquitous Environments; September 30-October 2, 2009; Helsinki, Finland p. 147-153 URL: https://dl.acm.org/doi/10.5555/1690508.1690528 [doi: 10.5555/1690508.1690528]
63. Brookman R, Hulm Z, Hearn L, Siette J, Mathew N, Deodhar S, et al. Evaluation of an exercise program incorporating an international cycling competition: a multimodal intervention model for physical, psychological, and social wellbeing in residential aged care. BMC Geriatr 2024;24(1):435 [FREE Full text] [doi: 10.1186/s12877-024-05033-x] [Medline: 38755554]
64. Hasan H, Linger H. Enhancing the wellbeing of the elderly: social use of digital technologies in aged care. Educ Gerontol 2016;42(11):749-757. [doi: 10.1080/03601277.2016.1205425]
65. Lignos N, McCloskey R, Donovan C, Ellis K, Herrington M, Kanik M. Use of an ambient activity technology for long-term care residents with dementia. J Gerontol Nurs 2022;48(1):35-41. [doi: 10.3928/00989134-20211206-05] [Medline: 34978492]
66. MacRitchie J, Breaden M, Taylor JR, Milne AJ. Exploring older adult needs and preferences for technology-assisted group music-making. A qualitative analysis of data collected during the participatory user-centred design process. Disabil Rehabil Assist Technol 2024;19(5):1935-1944 [FREE Full text] [doi: 10.1080/17483107.2022.2077461] [Medline: 35658719]
67. Subramaniam P, Woods B. Digital life storybooks for people with dementia living in care homes: an evaluation. Clin Interv Aging 2016;11:1263-1276 [FREE Full text] [doi: 10.2147/CIA.S111097] [Medline: 27698556]
68. Thoolen M, Brankaert R, Lu Y. AmbientEcho: exploring interactive media experiences in the context of residential dementia care. 2020 Presented at: DIS '20: Proceedings of the 2020 ACM Designing Interactive Systems Conference; July 6-10, 2020; Eindhoven, Netherlands p. 1495-1508. [doi: 10.1145/3357236.3395432]
69. Vasseur D, Ipakchian Askari S, Suijkerbuijk S, Nap HH, IJsselsteijn W. Sensory, affective, and social experiences with haptic devices in intramural care practice. Nurs Rep 2024;14(1):230-253 [FREE Full text] [doi: 10.3390/nursrep14010019] [Medline: 38391064]
70. Belgrave M, Palmer K, Luger Motyka TM. Creative aging in virtual spaces: using museum content and music therapy to explore cultural diversity. Front Med (Lausanne) 2023;10:1273000 [FREE Full text] [doi: 10.3389/fmed.2023.1273000] [Medline: 38098846]
71. Breckenridge JP, Kroll T, Wylie G, Salzberg A. Remembering the past, enhancing the present and sharing the future: a qualitative study of the impact of film screenings in care home communities. Ageing Soc 2020;41(11):2615-2640. [doi: 10.1017/s0144686x20000501]
72. Caserta MS, Lund DA. Video Respite® in an Alzheimer's care center. Act Adapt Aging 2008;27(1):13-28. [doi: 10.1300/j016v27n01_02]
73. Howard EP, Retalic T, Rogan J, Murphy K, Swaminathan S, Altschuler A. ArtontheBrain: results of a pilot project among long-term care residents. Res Gerontol Nurs 2021;14(5):235-243. [doi: 10.3928/19404921-20210825-03] [Medline: 34542348]
74. Juul A, Wilding R, Baldassar L. The best day of the week: new technology enhancing quality of life in a care home. Int J Environ Res Public Health 2019;16(6):1000 [FREE Full text] [doi: 10.3390/ijerph16061000] [Medline: 30893945]
75. Ysseldyk R, Morton TA, Haslam C, Haslam SA, Boger J, Giau E, et al. You've Got E-Mail: a pilot study examining the feasibility and impact of a group-based technology-training intervention among older adults living in residential care. Can J Aging 2024;43(1):45-56. [doi: 10.1017/S0714980823000375] [Medline: 37501571]
76. Samuelsson C, Ferm U, Ekström A. "It's Our Gang"—promoting social inclusion for people with dementia by using digital communication support in a group activity. Clin Gerontol 2021;44(4):418-429 [FREE Full text] [doi: 10.1080/07317115.2020.1795037] [Medline: 32791946]
77. Tsai HH, Tsai YF, Wang HH, Chang YC, Chu HH. Videoconference program enhances social support, loneliness, and depressive status of elderly nursing home residents. Aging Ment Health 2010;14(8):947-954. [doi: 10.1080/13607863.2010.501057] [Medline: 21069600]
78. Bradwell HL, Cooper L, Edwards KJ, Baxter R, Tomaz SA, Ritchie J, et al. Staff perceptions towards virtual reality-motivated treadmill exercise for care home residents: a qualitative feedback study with key stakeholders and follow-up interview with technology developer. BMJ Open 2023;13(11):e073307 [FREE Full text] [doi: 10.1136/bmjopen-2023-073307] [Medline: 37996232]
79. Goodall G, Taraldsen K, Granbo R, Serrano JA. Towards personalized dementia care through meaningful activities supported by technology: a multisite qualitative study with care professionals. BMC Geriatr 2021;21(1):468 [FREE Full text] [doi: 10.1186/s12877-021-02408-2] [Medline: 34418973]
80. Isbel S, Frost J, Fearon A, Kellett J, D'Cunha NM. Take a Cycle with Me: an intergenerational group activity with older adults and adolescents. J Intergener Relatsh 2023;23(1):1-11. [doi: 10.1080/15350770.2023.2282499]

XSL•FO
RenderX

81. Li Y, Wilke C, Shiyanov I, Muschalla B. Impact of virtual reality-based group activities on activity level and well-being among older adults in nursing homes: longitudinal exploratory study. JMIR Serious Games 2024;12:e50796 [FREE Full text] [doi: 10.2196/50796] [Medline: 38551635]
82. Sheehy L, Bouchard S, Kakkar A, El Hakim R, Lhoest J, Frank A. Development and initial testing of an artificial intelligence-based virtual reality companion for people living with dementia in long-term care. J Clin Med 2024;13(18):5574 [FREE Full text] [doi: 10.3390/jcm13185574] [Medline: 39337060]
83. D'Cunha NM, Isbel ST, Frost J, Fearon A, McKune AJ, Naumovski N, et al. Effects of a virtual group cycling experience on people living with dementia: a mixed method pilot study. Dementia (London) 2021;20(5):1518-1535. [doi: 10.1177/1471301220951328] [Medline: 32820955]
84. Goodall G, André L, Taraldsen K, Serrano JA. Supporting identity and relationships amongst people with dementia through the use of technology: a qualitative interview study. Int J Qual Stud Health Well-being 2021;16(1):1920349 [FREE Full text] [doi: 10.1080/17482631.2021.1920349] [Medline: 33955324]
85. Khirallah Abd El Fatah N, Abdelwahab Khedr M, Alshammari M, Mabrouk Abdelaziz Elgarhy S. Effect of immersive virtual reality reminiscence versus traditional reminiscence therapy on cognitive function and psychological well-being among older adults in assisted living facilities: a randomized controlled trial. Geriatr Nurs 2024;55:191-203. [doi: 10.1016/j.gerinurse.2023.11.010] [Medline: 38007908]
86. Lin TY, Huang CM, Hsu HP, Liao JY, Cheng VYW, Wang SW, et al. Effects of a combination of three-dimensional virtual reality and hands-on horticultural therapy on institutionalized older adults' physical and mental health: quasi-experimental design. J Med Internet Res 2020;22(11):e19002 [FREE Full text] [doi: 10.2196/19002] [Medline: 33135666]
87. Siriaraya P, Ang CS. Recreating living experiences from past memories through virtual worlds for people with dementia. 2014 Presented at: CHI '14: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; April 26-May 1, 2014; Toronto, Ontario, Canada p. 3977-3986. [doi: 10.1145/2556288.2557035]
88. Feng Y, Perugia G, Yu S, Barakova EI, Hu J, Rauterberg GWM. Context-enhanced human-robot interaction: exploring the role of system interactivity and multimodal stimuli on the engagement of people with dementia. Int J Soc Robot 2021;14(3):807-826. [doi: 10.1007/s12369-021-00823-4]
89. Melkas H, Hennala L, Pekkarinen S, Kyrki V. Impacts of robot implementation on care personnel and clients in elderly-care institutions. Int J Med Inform 2020;134:104041 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.104041] [Medline: 31838291]
90. Paluch R, Müller C. "That's Something for Children": an ethnographic study of attitudes and practices of care attendants and nursing home residents towards robotic pets. 2022 Presented at: Proceedings of the ACM on Human-Computer Interaction, Volume 6, Issue GROUP; January 14, 2022; New York, NY, United States p. 1-35. [doi: 10.1145/3492850]
91. Blindheim K, Solberg M, Hameed IA, Alnes RE. Promoting activity in long-term care facilities with the social robot Pepper: a pilot study. Inform Health Soc Care 2023;48(2):181-195 [FREE Full text] [doi: 10.1080/17538157.2022.2086465] [Medline: 35702818]
92. Chen SC, Davis BH, Kuo CY, Maclagan M, Chien CO, Lin MF. Can the Paro be my Buddy? Meaningful experiences from the perspectives of older adults. Geriatr Nurs 2022;43:130-137. [doi: 10.1016/j.gerinurse.2021.11.011] [Medline: 34883391]
93. De Carolis B, Macchiarulo N, La Forgia A, Melone G, Nardelli F. Designing and evaluating Rebecca: a telepresence robot in the shoes of a social assistant for cognitive stimulation therapy. 2025 Presented at: CHItaly '25: Proceedings of the 16th Biannual Conference of the Italian SIGCHI Chapter; October 6-10, 2022; Salerno, Italy p. 1-8. [doi: 10.1145/3750069.3750119]
94. Feng Y, Barakova EI, Yu S, Hu J, Rauterberg GWM. Effects of the level of interactivity of a social robot and the response of the augmented reality display in contextual interactions of people with dementia. Sensors (Basel) 2020;20(13):3771 [FREE Full text] [doi: 10.3390/s20133771] [Medline: 32635640]
95. Feng Y, Yu S, Van DMD, Barakova E, Hu J, Rauterberg M. LiveNature: ambient display and social robot-facilitated multi-sensory engagement for people with dementia. 2019 Presented at: DIS '19: Proceedings of the 2019 on Designing Interactive Systems Conference; June 23-28, 2019; San Diego, CA, United States p. 1321-1333. [doi: 10.1145/3322276.3322331]
96. Fogelson DM, Rutledge C, Zimbro KS. The impact of robotic companion pets on depression and loneliness for older adults with dementia during the COVID-19 pandemic. J Holist Nurs 2022;40(4):397-409 [FREE Full text] [doi: 10.1177/08980101211064605] [Medline: 34889146]
97. Khosla R, Chu MT. Embodying care in matilda: an affective communication robot for emotional wellbeing of older people in Australian residential care facilities. ACM Trans Manag Inf Syst 2013;4(4):1-33. [doi: 10.1145/2544104]
98. Chang CY, Wang PL, Li CJ, Hwang GJ. From empathy to quality long-term care: a generative AI-based art therapy approach based on the self-directed learning model. Interact Learn Environ 2024;33(5):3333-3353. [doi: 10.1080/10494820.2024.2443072]
99. Marchetti E, Grimme S, Hornecker E, Kollakidou A, Graf P. Pet-robot or appliance? Care home residents with dementia respond to a zoomorphic floor washing robot. 2022 Presented at: CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems; April 29-May 5, 2022; New Orleans, LA, United States p. 1-21. [doi: 10.1145/3491102.3517463]

XSL•FO
RenderX

100. Wangmo T, Duong V, Felber NA, Tian YJA, Mihailov E. No playing around with robots? Ambivalent attitudes toward the use of Paro in elder care. Nurs Inq 2024;31(3):e12645. [doi: 10.1111/nin.12645] [Medline: 38812242]
101. Wu D, Pu L, Jo J, Hexel R, Moyle W. Deploying robot-led activities for people with dementia at aged care facilities: a feasibility study. J Am Med Dir Assoc 2024;25(7):105028. [doi: 10.1016/j.jamda.2024.105028] [Medline: 38772526]
102. Hoel V, Seibert K, Domhoff D, Preuß B, Heinze F, Rothgang H, et al. Social health among German nursing home residents with dementia during the COVID-19 pandemic, and the role of technology to promote social participation. Int J Environ Res Public Health 2022;19(4):1956 [FREE Full text] [doi: 10.3390/ijerph19041956] [Medline: 35206143]
103. Li Y, Liang N, Nejat G. How emotional dances with social robots influence group HRI for older adults. ACM Trans Hum Robot Interact 2025;15(2):1-33. [doi: 10.1145/3773894]
104. Waycott J, Zhao W, Kelly RM, Robertson E. Technology-mediated enrichment in aged care: survey and interview study. JMIR Aging 2022;5(2):e31162 [FREE Full text] [doi: 10.2196/31162] [Medline: 34975014]
105. Ito H, Banh H, Wong KLY, Wong L, Allen S, Hung L. A critical reflection of an intergenerational, student-led team bringing social robots and research to older adults in the community. Gerontol Geriatr Educ 2025;46(4):587-602. [doi: 10.1080/02701960.2025.2507413] [Medline: 40380793]
106. Dodson P, Haase AM, Jeffreys M, Hales C. Capturing patient experiences of care with digital technology to improve service delivery and quality of care: a scoping review. Digit Health 2024;10:20552076241282900 [FREE Full text] [doi: 10.1177/20552076241282900] [Medline: 39449855]
107. Jecker NS, Sparrow R, Lederman Z, Ho A. Digital humans to combat loneliness and social isolation: ethics concerns and policy recommendations. Hastings Cent Rep 2024;54(1):7-12. [doi: 10.1002/hast.1562] [Medline: 38390675]
108. Brankaert R. Inclusive designs in healthcare: warm technology for people with dementia. In: Applied Design Research. Boca Raton, FL: CRC Press; 2022:137-146.
109. Hicks B, Karim A, Jones E, Burgin M, Cutler C, Tang W, et al. Care home practitioners' perceptions of the barriers and facilitators for using off-the-shelf gaming technology with people with dementia. Dementia (London) 2022;21(5):1532-1555 [FREE Full text] [doi: 10.1177/14713012221085229] [Medline: 35427198]
110. Zhao W, Kelly RM, Waycott J. Technology deployment for social connection in residential aged care: care and technology providers' experiences during the COVID-19 pandemic. 2023 Presented at: CHI '23: CHI Conference on Human Factors in Computing Systems; April 23-28, 2023; Hamburg, Germany. [doi: 10.1145/3544549.3585662]
111. Mihailov E, Wangmo T. Does humanness matter? An ethical evaluation of sharing care work with social robots. Sci Eng Ethics 2025;31(4):20. [doi: 10.1007/s11948-025-00547-y] [Medline: 40721552]
112. Ware C, Rigaud AS, Blavette L, Damnée S, Dacunha S, Lenoir H, et al. Development of an ethical framework for the use of social robots in the care of individuals with major neurocognitive disorders: a qualitative study. BMC Geriatr 2025;25(1):260 [FREE Full text] [doi: 10.1186/s12877-025-05891-z] [Medline: 40264012]
113. Khaksar W, Lindblom DS, Bygrave LA, Torresen J. Robotics in elderly healthcare: a qualitative analysis of 20 recent European research projects. ACM Trans Hum Robot Interact 2025;14(2):1-38. [doi: 10.1145/3711936]
114. Hall A, Wilson CB, Stanmore E, Todd C. Implementing monitoring technologies in care homes for people with dementia: a qualitative exploration using normalization process theory. Int J Nurs Stud 2017;72:60-70 [FREE Full text] [doi: 10.1016/j.ijnurstu.2017.04.008] [Medline: 28494333]
115. Theopilus Y, Davis H, Pedell S. Digital technology for supporting aged care services: a scoping review. 2025 Presented at: CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems; April 26-May 1, 2025; Yokohama, Japan p. 1-12. [doi: 10.1145/3706598.3713498]
116. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240020924 [accessed 2025-09-18]
117. Tischendorf T, Hinsche L, Hasseler M, Schaal T. GenAI in nursing and clinical practice: a rapid review of applications and challenges. J Public Health 2025:1-11. [doi: 10.1007/s10389-025-02523-z]
118. Pardini S, Calcagno R, Genovese A, Salvadori E, Ibarra OM. Exploring virtual reality-based reminiscence therapy on cognitive and emotional well-being in people with cognitive impairments: a scoping review. Brain Sci 2025;15(5):500 [FREE Full text] [doi: 10.3390/brainsci15050500] [Medline: 40426671]
119. Reed JM, Dodson T, Petrinec A, Hughes J, Miller RD. The HARMONEE project: using genAI images for reminiscence with older adults in long-term care. Geriatr Nurs 2025;65:103448 [FREE Full text] [doi: 10.1016/j.gerinurse.2025.103448] [Medline: 40618723]
120. Bradley L, Shanker S, Murphy J, Fenge LA, Heward M. Effectiveness of digital technologies to engage and support the wellbeing of people with dementia and family carers at home and in care homes: a scoping review. Dementia (London) 2023;22(6):1292-1313 [FREE Full text] [doi: 10.1177/14713012231178445] [Medline: 37235791]
121. Vrančić A, Zadravec H, Orehovački T. The role of smart homes in providing care for older adults: a systematic literature review from 2010 to 2023. Smart Cities 2024;7(4):1502-1550. [doi: 10.3390/smartcities7040062]
122. Abbasian M, Azimi I, Rahmani AM, Jain R. Conversational health agents: a personalized LLM-powered agent framework. JAMIA Open 2025;8(4):ooaf067 [FREE Full text] [doi: 10.1093/jamiaopen/ooaf067]

XSL•FO
RenderX

123. Khan MA, Din IU, Khan NA, Hassan S, Almogren A. Adaptive generative AI for elderly-assisted living environments: a proactive approach. 2024 Presented at: 7th International Conference on Internet Applications, Protocols, and Services (NETAPPS); November 6-7, 2024; Kuala Lumpur, Malaysia. [doi: 10.1109/netapps63333.2024.10823634]
124. Levin C, Zaboli A, Turcato G, Saban M. Nursing judgment in the age of generative artificial intelligence: a cross-national study on clinical decision-making performance among emergency nurses. Int J Nurs Stud 2025;172:105216 [FREE Full text] [doi: 10.1016/j.ijnurstu.2025.105216] [Medline: 40975905]
125. Bautista JR, Westendorf R, Goth O. Using a socio-technical strategy to identify the use and implications of generative artificial intelligence tools on nursing education and practice. J Adv Nurs 2025. [doi: 10.1111/jan.70401] [Medline: 41288087]
126. Johnson J, Jamshed H, Zuppke R, Leggett A, Provost EM, Brewer RN. “I Felt Listened to”: evaluating an AI-powered reflection tool for care partners. ACM Trans Access Comput 2025;18(3):1-35. [doi: 10.1145/3731562]
127. Karimi P, Martin-Hammond A. Finding the right balance: user control and automation in AI tools for supporting older adults' health information tasks. 2025 Presented at: CHI EA '25: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems; April 26-May 1, 2025; Yokohama, Japan p. 1-10. [doi: 10.1145/3706599.3719773]
128. Karimi P, Martin-Hammond A. Designing intelligent voice assistants for older adults' collaborative care: exploring supportive and non-supportive interactions. Proc ACM Hum-Comput Interact 2025;9(7):1-35. [doi: 10.1145/3757465]
129. Saha K, Jain Y, Liu C, Kaliappan S, Karkar R. AI vs. humans for online support: comparing the language of responses from LLMs and online communities of alzheimer's disease. 2025 Presented at: ACM Transactions on Computing for Healthcare; March 11, 2025; New York, NY, United States. [doi: 10.1145/3709366]
130. Katz S. Aging together-with: The growing older of humans, non-humans and more-than-humans. A commentary. J Aging Stud 2024;71:101280 [FREE Full text] [doi: 10.1016/j.jaging.2024.101280] [Medline: 39608904]
131. Vandemeulebroucke T, Dzi K, Gastmans C. Older adults' experiences with and perceptions of the use of socially assistive robots in aged care: a systematic review of quantitative evidence. Arch Gerontol Geriatr 2021;95:104399. [doi: 10.1016/j.archger.2021.104399] [Medline: 33813208]
132. Okabe-Miyamoto K, Walsh LC, Ozer DJ, Lyubomirsky S. Measuring the experience of social connection within specific social interactions: the Connection During Conversations Scale (CDCS). PLoS One 2024;19(1):e0286408 [FREE Full text] [doi: 10.1371/journal.pone.0286408] [Medline: 38236933]
133. Moyle W, Pu L, Murfield J, Sung B, Sriram D, Liddle J, et al. Consumer and provider perspectives on technologies used within aged care: an Australian qualitative needs assessment survey. J Appl Gerontol 2022;41(12):2557-2565 [FREE Full text] [doi: 10.1177/07334648221120082] [Medline: 35948942]
134. Bhowmick P, Stolterman BE. Exploring tangible user interface design for social connection among older adults: a preliminary review. 2023 Presented at: CHI EA '23: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems; April 23-28, 2023; Hamburg, Germany p. 1-9. [doi: 10.1145/3544549.3585722]
135. Keane J, Ryan C, Usher R. Leisure and social occupational choice within nursing home facilities in Ireland: residents perspectives. Scand J Occup Ther 2024;31(1):2259222 [FREE Full text] [doi: 10.1080/11038128.2023.2259222] [Medline: 37918009]
136. Yuan S, Coghlan S, Lederman R, Waycott J. Meaningful engagement, ethical care, and design opportunities: an ethnographic study on social activities in long-term care. 2025 Presented at: CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems; April 26-May 1, 2025; Yokohama, Japan p. 1-15. [doi: 10.1145/3706598.3713755]
137. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014;348:g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]

Abbreviations

GenAI: generative artificial intelligence
HARMONEE: Harnessing Artificial Intelligence Resources for Mental Well-Being for Older Adults and Nurturing Empathy in Education
JBI: Joanna Briggs Institute
MeSH: Medical Subject Headings
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
TIDieR: Template for Intervention Description and Replication
VR: virtual reality
WHO: World Health Organization

XSL•FO
RenderX

Edited by S Brini; submitted 26.Nov.2025; peer-reviewed by Z Su, W Ghulam Hussain; comments to author 09.Dec.2025; accepted 26.Jan.2026; published 03.Mar.2026.

Please cite as:
Muldrew D, Bradley R, Conway K
Digital Technologies That Support Meaningful Connections in Care Homes: Scoping Review
J Med Internet Res 2026;28:e88181
URL: https://www.jmir.org/2026/1/e88181
doi:10.2196/88181
PMID:

©Deborah Muldrew, Rosemary Bradley, Kelly Conway. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Association Between Digital Biomarkers of Health and Anxiety: Systematic Review and Meta-Analysis

Yolanda Lau[1], MSc; Natalia Chemas[2], MSc; Heema Ajeet Gokani[1], MSc; Rachel Morrell[1], MSc; Harisd Phannarus[1,3], MD; Claudia Cooper[4], MD; Zuzana Walker[1,5], MD; Harriet Demnitz-King[1,4], PhD; Natalie L Marchant[1], PhD

[1]Division of Psychiatry, University College London, London, United Kingdom

[2]Centre for Preventive Neurology, Wolfson Institute of Population Health, Queen Mary University of London, London, United Kingdom

[3]Department of Preventive and Social Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand

[4]Centre for Psychiatry and Mental Health, Wolfson Institute of Population Health, Queen Mary University of London, London, United Kingdom

[5]Essex Partnership University NHS Foundation Trust, Essex, United Kingdom

Corresponding Author:
Natalie L Marchant, PhD
Division of Psychiatry
University College London
Division of Psychiatry, 4th Floor Maple House
149 Tottenham Court Road
London, W1T 5AU
United Kingdom
Phone: 44 20 3108 7961
Email: n.marchant@ucl.ac.uk

Abstract

Background: Digital biomarkers are gaining interest as proxy markers for mental health, as they enable passive and continuous data collection. However, the association between digital biomarkers of health and anxiety, both generalized anxiety disorder and anxiety symptoms, remains unknown.

Objective: This systematic review and meta-analysis examined the association between digital biomarkers of health obtained from wrist-worn wearables and anxiety in adults.

Methods: Systematic literature searches were conducted across 6 databases, including unpublished gray literature. The final search was done on September 21, 2025. Cross-sectional or longitudinal studies investigating the association between digital biomarkers from wrist-worn wearables and anxiety were eligible. Studies using inferential statistics or machine learning methods were both eligible. Studies were excluded if participants received diagnoses of neurodegenerative disorders or physical health conditions. Two risk-of-bias tools were used: the National Heart, Lung, and Blood Institute assessment tool for inferential statistical studies, and the modified version of the Quality Assessment of Diagnostic Accuracy Studies-2 for machine learning studies. Whenever possible, effect sizes were combined across studies, for each digital biomarker of health separately, using random-effects meta-analyses. Sensitivity analyses were performed to assess whether results differed according to anxiety type (state or trait) and age group. Otherwise, studies were synthesized narratively.

Results: A total of 44 studies from 42 articles were eligible. Among these, 36 studies used inferential statistical approaches for analysis (21 reporting sleep characteristics, 8 reporting physical activity, 2 reporting heart rate variability, and 5 reporting more than 1 type), and 8 studies used machine learning approaches. Sample size ranged from 17 to 170,320. Meta-analyses on 4 sleep metrics found no associations: sleep efficiency (Fisher z=–0.07, 95% CI –0.14 to 0.002; P=.06; PI –0.19 to 0.05), wake after sleep onset (Fisher z=0.13, 95% CI –0.04 to 0.30; P=.11; PI –0.15 to 0.41), total sleep time (Fisher z=0.009, 95% CI –0.01 to 0.03; P=.28; PI –0.02 to 0.03), and sleep onset latency (Fisher z=0.04, 95% CI –0.07 to 0.15; P=.08; PI –0.19 to 0.27). Qualitative syntheses revealed that lower physical activity levels and higher heart rate were associated with greater anxiety symptoms. Machine learning studies using wrist-worn wearable data alone showed varied performance, with predictive performance improving when wearable data were combined with other data sources.

Conclusions: This is the first review to synthesize evidence from inferential statistical (mostly fair quality) and machine learning studies examining association between wearable-derived digital biomarkers and anxiety. Meta-analyses found no associations between sleep metrics and anxiety. Although based on limited studies, lower physical activity levels and elevated heart rate were

XSL•FO
RenderX

associated with greater anxiety symptoms. Digital biomarkers may be more useful when integrated with other data sources (eg, self-report and clinical data) rather than used as stand-alone screening tools.

Trial Registration: PROSPERO CRD42023409995; https://www.crd.york.ac.uk/PROSPERO/view/CRD42023409995

(J Med Internet Res 2026;28:e73812) doi:10.2196/73812

KEYWORDS

anxiety; digital; heart rate; machine learning; meta-analysis; physical activity; sleep; systematic review; wearable

Introduction

Anxiety is the most prevalent mental health condition globally, impacting 301 million people in 2019 [1]. Symptoms can include worry, restlessness, fatigue, difficulty concentrating, and sleep disturbances [2]. Among older adults, anxiety disorders and anxiety symptoms are common, with 1.2% to 28% experiencing anxiety disorders (generalized anxiety disorder [GAD] being the most common type) and 15% to 56% experiencing anxiety symptoms in community and clinical populations [3]. In older adults, both GAD and anxiety symptoms have been associated with increased risk of dementia [4-8]. Given the increasing global prevalence of dementia [9] and the lack of effective disease-modifying treatments, it is crucial to prioritize the identification and treatment of modifiable risk factors associated with dementia to reduce dementia cases. Therefore, early identification of anxiety in both the general population and older adults could have significant clinical implications.

Predominant methods to assess anxiety include the use of self-report questionnaires, clinical diagnoses, or a combination of both. These methods have several limitations. Self-report questionnaires measuring trait anxiety are retrospective and rely on an individual's ability, insight, and willingness to respond accurately. Evidence suggests that self-report questionnaires can also be affected by recall bias and social desirability, potentially leading to misreporting of anxiety symptoms [10]. Clinical diagnoses require clinical expertise and are, therefore, time-consuming and resource-heavy. In addition, in older adults, anxiety is often underdiagnosed due to overlapping symptoms with physical health conditions [11]. Both methods only provide a snapshot of an individual's symptomology unless assessed repeatedly, which can increase demands on both individuals and clinicians.

Leveraging digital biomarkers, defined as objective, quantifiable, physiological, and behavioral measures collected using digital devices [12], as proxy markers of anxiety has the potential to overcome the limitations outlined above. Digital biomarkers collect continuous objective data on various aspects of behavior and physiology.

While digital biomarkers can be measured using different technologies (eg, contactless sensors or smartphones), this review focuses on wrist-worn wearable devices, which are widely used by the general public and in existing research [13,14]. Wrist-worn wearables can measure physical activity (eg, different physical intensities), sleep (eg, sleep efficiency [SE]), and heart rate (eg, heart rate variability [HRV]). These variables have been linked to anxiety; for example, lower physical activity levels and more sleep disturbances have been associated with more anxiety symptoms and GAD [15,16]. There is also emerging evidence supporting the effectiveness of wearable-derived metrics in detecting mental health symptoms [17,18]. Studies also found that digital biomarkers collected from other devices can detect mental health symptoms. For instance, a study found that metrics from smartphones (including heart rate metrics) were able to predict changes in anxiety symptoms [19]. Similarly, studies using the Oura Ring also reported associations between sleep metrics and anxiety [20].

Two recent meta-analyses of machine learning studies examined the association between metrics collected from wearable devices and depression and anxiety, both of which are modifiable psychosocial risk factors of dementia [17,18]. They found that the models accurately classified individuals with and without depression in 89% of cases [17] and correctly identified individuals with and without anxiety in 82% of cases [18]. Similar wearable-derived metrics were used as predictors in anxiety and depression models (eg, activity, sleep, and heart rate). These meta-analyses only focused on machine learning studies; existing research also used inferential statistical methods.

Inferential statistical methods and machine learning models offer distinct advantages. Inferential statistical models provide interpretable results, allowing for hypothesis testing and quantifying associations between digital biomarkers of health and anxiety. In contrast, machine learning models often prioritize predictive accuracy and excel at capturing complex, nonlinear relationships that may not be easily captured through inferential statistical models. Together, these 2 approaches offer complementary insights.

Given the growing interest in using wearable-derived metrics in mental health and dementia research, this review aims to contribute to these research fields by systematically reviewing existing studies that have investigated the association between metrics collected from wrist-worn wearables and anxiety using inferential and machine learning statistical approaches.

Methods

Design

This review was conducted in line with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) recommendation [21] and registered with PROSPERO (International Prospective Register of Systematic Reviews) in March 2023 (CRD42023409995). The PRISMA 2020 checklist [22] (Multimedia Appendix 1), the PRISMA Abstract checklist [22] (Multimedia Appendix 2), and the PRISMA-S (Preferred

XSL•FO
RenderX

Reporting Items for Systematic Reviews and Meta-Analyses–Search Extension) checklist [23] (Multimedia Appendix 3) were used.

Search Strategy

A total of 5 databases (Medline, Embase, PsycINFO, Web of Science, and CINAHL) were systematically searched. Further, unpublished gray literature searches were conducted with ProQuest Dissertations and Theses Global. Following published recommendations [24], we also checked the first 300 papers on Google Scholar to supplement the gray literature search. Finally, to supplement the electronic search, references of eligible papers were reviewed to identify any additional eligible studies.

The search strategy combined 2 search term strings with "AND." Each string reflected a concept related to the research question: (1) digital biomarkers and (2) psychosocial risk factors of dementia (Multimedia Appendix 4). The search strategy was informed by previous reviews in related research areas [17], but all search terms and strategies were developed specifically for this review and adapted for each database. There were no restrictions imposed on the date of publication. This review did not restrict the database searches to human studies and studies published in English, and no filters were used. The initial searches were conducted on April 27, 2023; a subsequent search was done on September 24, 2024, and a final search focusing only on anxiety was done on September 21, 2025. While the PROSPERO registration included multiple psychosocial risk factors, this systematic review focused specifically on anxiety, due to the large number of articles that made it past full-text screening. As all eligible anxiety studies in earlier searches were identified via anxiety-specific terms, a final search restricted to anxiety-related search terms was conducted in 2025.

Study Selection

The web platform Covidence (Veritas Health Innovation, 2020) was used for deduplication and to coordinate multiuser title-abstract and full-text screening. Two reviewers independently screened titles and abstracts of all the identified studies (YL screened 100% of the articles, and NC, HAG, RM, and HD-K collectively screened 100%), followed by full-text screening. Any disagreements were resolved by a third reviewer (NM or HD-K), with HD-K only resolving conflicts for articles she had not screened at the title-abstract screening stage. Interrater agreement was assessed using the Cohen κ coefficient.

In accordance with the PROSPERO registration, our search strategies encompassed a range of psychosocial risk factors for dementia. However, given the large number of eligible articles that made it through to the full-text screening stage, we decided to categorize these articles into different psychosocial risk factors before the data extraction stage.

This systematic review focused on anxiety, both clinical anxiety and anxiety symptoms. The PECOS (Population, Exposure, Control, Outcome, and Study design) framework was used to structure the study selection. Studies were selected if they (1) involved adults with a mean age of >18 years (Population); (2) reported data from research- or consumer-grade wrist-worn devices (Exposure); (3) did not require a control group, although studies including participants without anxiety were considered as comparators when applicable (Control); (4) assessed GAD or anxiety symptoms using standardized assessments (eg, self-report questionnaires or clinical diagnoses [Outcome]); and (5) used cross-sectional or longitudinal designs (Study design). Studies published in an English-language, peer-reviewed journal or thesis/dissertation, and studies that examined the relationship between anxiety and digital biomarkers via inferential statistics or machine learning methods were also eligible. Studies were excluded if they (1) focused on a sample that received diagnoses of neurodegenerative disorders or physical health conditions (however, studies were eligible if participants without these conditions were analyzed separately), and (2) primarily focused on participants with a psychiatric comorbidity, neurological disorder, or stroke (>50% of the sample).

Data Extraction

A standardized form was created to extract the following data from eligible inferential statistical studies: (1) authors and year of publication; (2) study characteristics, including type of study (ie, cross-sectional or longitudinal); (3) demographic information; (4) characteristics of digital biomarkers, including digital biomarker type and devices used for data collection; (5) anxiety measurement, and (6) data required for meta-analysis (eg, mean difference and correlation coefficients). For studies using machine learning approaches, another standardized form was created to extract the following details: (1) authors and year of publication, (2) demographic information, (3) ground truth assessment, (4) characteristics of digital biomarkers (5) machine learning task, (6) predictor features, (7) algorithms, (8) validation method, and (9) model performance.

Where necessary, authors were contacted to request additional information.

Quality Assessments

Inferential Statistical Studies

The National Heart, Lung, and Blood Institute assessment tool was used [25]. This quality assessment tool comprises 14 criteria, which are used to evaluate the validity and reliability of studies. For cross-sectional studies, 4 criteria that are relevant to longitudinal study designs are not applicable and were therefore omitted [26]. Scores were averaged, and each study was categorized as having a "Good" (≥80), "Fair" (50%-80%), or "Poor" (50%) rating, as specified in the guidelines.

Machine Learning Studies

Currently, there is no validated quality assessment tool for machine learning studies. We used a modified version of the validated Quality Assessment of Diagnostic Accuracy Studies 2 [27], which evaluates the risk of bias of studies across 4 domains (participants, index test [artificial intelligence algorithms], reference standard [ground truth], and analysis), and evaluates applicability concerns across 3 domains (excluding the analysis domain) [17].

Two independent reviewers assessed the quality of each eligible study. Any disagreements were resolved by a third reviewer.

XSL•FO
RenderX

Synthesis and Analysis

Effect sizes were combined across studies, separately for each type of digital biomarker of health, using random-effects meta-analyses. When effect sizes were unavailable, other available data were used for calculations (eg, *t* test statistic). Random-effects meta-analyses were used as they take into account between-study heterogeneity [28]. For meta-analysis based on a limited number of studies (<20 studies [29]), we applied the Hartung-Knapp-Sidik-Jonkman adjustments when calculating CIs and *P* values, which are recommended for small meta-analyses [30]. Prediction intervals for random-effects meta-analyses were also reported [31].

To ensure only 1 effect size from each included study was used in the primary meta-analysis, studies were selected based on an a priori determined hierarchy. Specifically, we prioritized (1) the largest sample size, (2) unadjusted over adjusted estimates, and (3) trait (chronic/ongoing) over state (present moment) anxiety, when more than 1 study used the same cohort. Trait anxiety refers to the general tendency to respond with anxiety across various situations, whereas state anxiety is a temporary response reflecting the present moment [32]. Since trait anxiety reflects a chronic condition with long-term implications and offers a more stable measure of general anxiety levels, it was the focus of our analysis.

Sensitivity analyses were conducted to examine whether results differed according to anxiety type (trait and state anxiety) and age group (young, middle, and older).

For each meta-analysis, heterogeneity was assessed using τ^2 (between-study variance) and I^2 (proportion of total variation in study estimates that is due to heterogeneity [33]). When there were 10 or more studies included in a meta-analysis, funnel plots would be used to assess small-study effects [34].

All analyses were conducted using the *metafor* package in R software (version 4.2.2; R Foundation for Statistical Computing).

Qualitative synthesis was conducted for studies using inferential statistical approaches that were not included in the meta-analysis. These studies were synthesized based on the type of digital biomarkers of health examined (ie, sleep metrics and activity metrics). On the other hand, studies using machine learning approaches were not meta-analyzed and were synthesized qualitatively, ordered by performance metrics (ie, accuracy, sensitivity, precision, and F_1-score).

Results

Study Selection

After the removal of duplicates, title/abstract screening, and full-text review, 365 articles were eligible (Figure 1). Of these, 47 articles focused on anxiety. Interrater reliability was high, with Cohen κ of 0.73 (substantial agreement) and 0.83 (almost perfect agreement) at the title/abstract screening and full-text review stages, respectively.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram illustrating the systematic process.

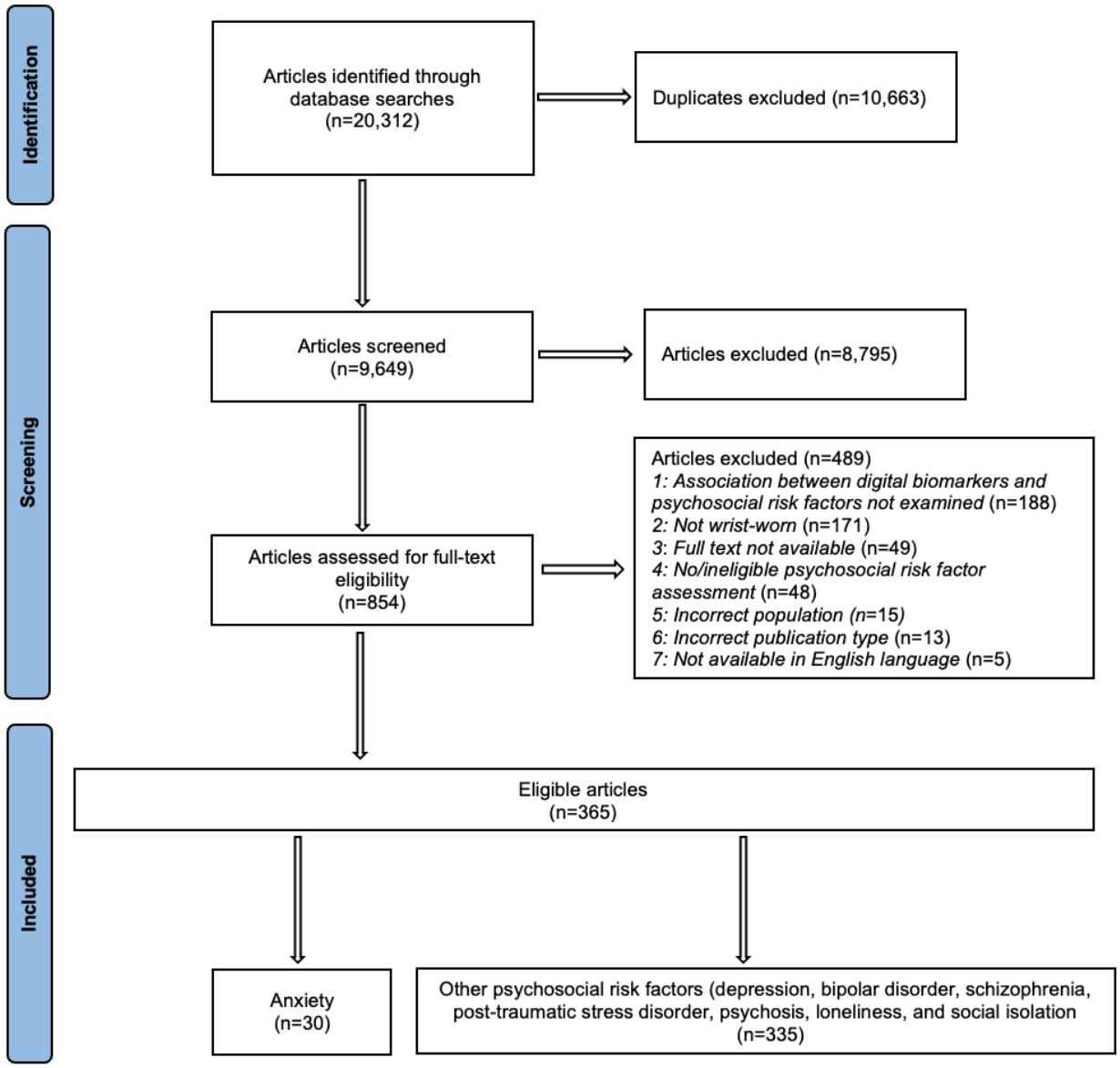

Study and Participant Characteristics

Overview

When an article reported outcomes from different cohorts or separately analyzed different populations (eg, male and female), each result was treated as a different study derived from the same article. This resulted in 44 studies from 42 articles. Of these, 36 (81.8%) studies used inferential statistical approaches for analysis, while 8 (18.2%) studies used machine learning approaches.

Studies Using Inferential Statistical Approaches

Characteristics of studies that used inferential statistical approaches are presented in Table 1. There was a total of 36 studies from 35 articles (cross-sectional: k=21, 58.3%; longitudinal: k=11, 30.6%; both cross-sectional and longitudinal: k=4, 11.1%). The study sizes ranged from 14 to 170,320 participants (median 189, IQR 79-2276). Most of the studies assessed trait anxiety or clinically diagnosed anxiety (k=19, 52.8%), 10 (27.8%) assessed state anxiety, and 7 (19.4%) assessed both state and trait anxiety. The majority of studies were conducted in young adults (k=18, 50%), 8 (22.2%) in middle-aged adults, 8 (22.2%) in older adults, and 2 (5.6%) did not report mean age data.

XSL•FO
RenderX

Table 1. Characteristics of samples eligible for systematic review and/or meta-analysis of studies using inferential statistical approaches. Most studies were cross-sectional (k=21), with sample sizes ranging from 14 to 170,320 (median 189, IQR 79-2276); most assessed trait or clinically diagnosed anxiety (k=21) and were conducted in young adults (k=18).

Study (location)	Study design (follow-up time)	Participants, N	Population type	Age (years), mean (SD) Sex (female), % Education, mean (SD) or categories % Ethnicity (White), %	Anxiety measure (type)	Wearable device (research/consumer-grade)	Digital biomarker type (subtype)
Ahn et al, 2024 (United States) [35]	Cross-sectional	33	Family caregivers	Age: 61.27 (11.2) Female: 84.8% Education: NR[a] White: 90.9%	GAD-7[b] (trait)	ActiGraph GT9X Link (research)	Sleep (SE[c])
Bullis, 2016 (United States) [36][d]	Cross-sectional	30	Clinical samples and healthy controls	Age: 28 (9.0) Female: 70% Education: high school graduate 6.7%; partial college 16.7%; college graduate 46.7%; graduate 30% White: 73.3%	Clinical diagnosis (*DSM-5*[e]—trait)	Actiwatch (research)	Sleep (SOL[f], TST[g], SE, and WASO[h])
Bullis, 2016 (United States) [36][c]	Longitudinal (14 days)	30	Clinical samples and healthy controls	Age: 28 (9.0) Female: 70% Education: high school graduate 6.7%; partial college 16.7%; college graduate 46.7%; graduate 30% White: 73.3%	Clinical diagnosis (*DSM-5*—trait) and daily symptom ratings (Ecological Momentary Assessment—state)	Actiwatch (research)	Sleep (SOL, TST, SE, and WASO)
Casanova et al, 2023 (United Kingdom) [37]	Cross-sectional	95,776	General population	Age: NR Female: NR Education: NR White: NR	GAD-7 (trait)	Axivity AX3 (research)	Activity (physical activity and SB[i])
Cox et al, 2018 (United States) [38]	Cross-sectional	138	Undergraduate students and community adults	Age: 22.5 (9.2) Female: 75.5% Education: NR White: 57.6%	Momentary anxiety (state)	ActiGraph wGT3X-BT (research)	Sleep (TST)
Cox et al, 2018 (United States) [38]	Longitudinal (9 days)	138	Undergraduate students and community adults	Age: 22.5 (9.2) Female: 75.5% Education: NR White: 57.6%	Momentary anxiety (state)	ActiGraph wGT3X-BT (research)	Sleep (TST)
D'Aurizio et al, 2023 (Italy) [39][c]	Cross-sectional	60	Prisoner and healthy control	Age: 36 (8.46) Female: 0% Education: 5-8 years White: NR	STAI[j] (trait and state)	Actiwatch Spectrum Plus (research)	Sleep (SOL, SE, TST, and WASO)
Dillon et al, 2018 (Ireland) [40]	Cross-sectional	397	Patients (population-representative random sample, from a primary care center)	Age: 59.6 (5.5) Female: 53.9% Education: primary 26.7%, secondary/tertiary 73.3% White: NR	HADS-A[k] (state)	GENEActiv accelerometer (research)	Activity (SB, LPA[l], and MVPA[m])
Doane et al, 2015 (United States) [41][c]	Cross-sectional	82	Adolescent transition to college	Age: 18.1 (0.4) Female: 76% Education: NR White: 54%	DASS-A[n] (state)	Actiwatch Score (research)	Sleep (sleep start time variability, SOL, SE, and TST)

XSL•FO
RenderX

Study (location)	Study design (follow-up time)	Participants, N	Population type	Age (years), mean (SD) Sex (female), % Education, mean (SD) or categories % Ethnicity (White), %	Anxiety measure (type)	Wearable device (research/consumer-grade)	Digital biomarker type (subtype)
Doane et al, 2015 (United States) [41][c]	Longitudinal (1 year)	Follow-up 1: 76 Follow-up 2: 71	Adolescent transition to college	NR	DASS-A (state)	Actiwatch Score (research)	Sleep (sleep start time variability, SOL, SE, and TST)
El-Sheikh et al, 2013 (United States; female) [42][c]	Cross-sectional	135	Cohabiting couples	Age: 36.5 (5.9) Female: 100% Education: NR White: NR	BAI[o] (trait)	Octagonal Basic Motionloggers (research)	Sleep (TST, SE, and SOL)
El-Sheikh et al, 2013 (United States; male) [42][c]	Cross-sectional	135	Cohabiting couples	Age: 39.4 (7.3) Female: 0% Education: NR White: NR	BAI (trait)	Octagonal Basic Motionloggers (research)	Sleep (TST, SE, and SOL)
Facer-Childs et al, 2022 (Australia) [43]	Cross-sectional	68	Elite football league athletes	Age: 23.3 (3.4) Female: 0% Education: NR White: NR	GAD-7 (trait) and DASS-A (state)	GENEActiv actigraphy devices (research)	Sleep (snooze time, SE, TST, and TIB[p])
Feng et al, 2021 (United States) [44]	Cross-sectional	14	Ophthalmology residents	Age: 30.2 (2.5) Female: 57.1% Education: NR White: NR	DASS-A (state)	Fitbit Alta HR (consumer)	Sleep (lower average sleep on call) and activity (average daily exercise)
Fuller-Rowell et al, 2024 (United States) [45][c]	Cross-sectional	874	Small towns and semirural communities	Age: 41.0 (7.0) Female: 67% Education: NR White: 67%	BAI (trait)	Octagonal Basic Motionloggers (research)	Sleep (TST, night-to-night variability, SE, and SOL)
Hofman et al, 2022 (Netherlands) [46]	Cross-sectional	1943	General population	Age: 70.9 (9.3) Female: 51.6% Education: primary 6%, lower 36.7%, further/intermediate 30.6%, higher 25.4% White: NR	HADS-A (state)	GENEActiv (research)	Sleep (TST) and activity (SB, LPA, and MVPA)
Huang et al, 2022 (China) [47]	Cross-sectional	321	College students	Age: 19.7 (1.2) Female: 55.8% Education: NR White: NR	SAS[q] (state)	Axivity AX3 (research)	Activity (SB-total, weekend, and weekdays)
Lee et al, 2021 (United States, Australia, South Korea) [48][c]	Cross-sectional	17	Elite esports athletes	Age: 20 (3.5) Female: 0% Education: NR White: NR	STAI (state)	Readiband V5 (research)	Sleep (TST, SOL, WASO, SOT[r], and WT[s])

XSL•FO
RenderX

Study (location)	Study design (follow-up time)	Participants, N	Population type	Age (years), mean (SD) Sex (female), % Education, mean (SD) or categories % Ethnicity (White), %	Anxiety measure (type)	Wearable device (research/consumer-grade)	Digital biomarker type (subtype)
Montanari et al, 2025 (France) [49]	Cross-sectional	211	Older adults	Age: 60-70 years: 105 (50%); 71-80 years: 89 (42%); >80 years: 17 (8.1%) Female: 36% Education: lower education 31%, secondary education 32%, higher education (37%) White: NR	STAI (trait and state)	Actigraph GT3X+ (research)	Sleep (TST, WASO, and SE)
Nguyen et al, 2023 (Australia) [50][c]	Cross-sectional	101	Newly recruited paramedics	Age: 26.0 (5.5) Female: 52% Education: NR White: NR	GAD-7 (trait)	Actiwatch Spectrum or Spectrum PRO (research)	Sleep (TST, WASO, SOL, and SE)
Nguyen et al, 2023 (Australia) [50][c]	Longitudinal (6 months)	101	Newly recruited paramedics	Age: 26.0 (5.5) Female: 52% Education: NR White: NR	GAD-7 (trait)	Actiwatch Spectrum or Spectrum PRO (research)	Sleep (TST, WASO, SOL, and SE)
Spira et al, 2008 (United States) [51][c]	Cross-sectional	59	Adults with primary insomnia	Age: 63.8 (6.9) Female: 74.6% Education: mean 16.3 (SD 2.3) years White: 78%	STAI (trait and state)	Actiwatch-L (research)	Sleep (WASO, SE, and SOL)
Spira et al, 2009 (United States) [52][c]	Cross-sectional	3040	Community-dwelling older women	Age: 83.6 (3.8) Female: 100% Education: mean 12.9 (SD 2.6) White: 89.3%	Goldberg Anxiety Scale (trait)	SleepWatch-O (research)	Sleep (TST, SE, time awake after sleep onset, SOL, and napping time)
Stremler et al, 2017 (Canada) [53]	Cross-sectional	118	Parents of critically ill hospitalized children	Age: 34.1 (7.2) Female: 62.7% Education: <high school 5.1%, high school 22%, university/college 53.4%, postgraduate 17.8%, unknown 1.7% White: 69.4%	STAI (state)	Octagonal Basic Motionlogger actigraphs (research)	Sleep (TST, number of nocturnal awakenings, and individual sleep variability)
Swanson et al, 2023 (United States) [54][c]	Cross-sectional	1197	Postmenopausal women	Age: 65.5 (2.6) Female: 100% Education: less than high school 5.2%, high school 15.3%, some college 30.7%, college 22.6%, post-college 26.2% White: 45.7%	GAD-7 (trait)	Actiwatch-2 (research)	Sleep (sleep irregularity and sleep midpoint outside 2-4 AM)
Uchida and Kurosawa, 2025 (Japan) [55]	Cross-sectional	50	University students	Age: 20.58 (1.34) Female: 64% Education: NR White: NR	Japanese version of STAI (trait and state)	Fitbit Charge 5 (consumer)	Sleep (TST, bedtime, and wake time)
Windmill et al, 2024 [56]	Cross-sectional	83	General population	Age: 24 (5.1) Female: 65.7% Education: NR White: NR	PANAS[t] (state and trait), STAI (state and trait), BAI (state)	GENEActiv, ActivInsights (research)	Sleep (SE and TST)

XSL•FO
RenderX

Study (location)	Study design (follow-up time)	Participants, N	Population type	Age (years), mean (SD) Sex (female), % Education, mean (SD) or categories % Ethnicity (White), %	Anxiety measure (type)	Wearable device (research/consumer-grade)	Digital biomarker type (subtype)
Walters et al, 2020 (Australia) [57]	Cross-sectional	110	Without sleep disorders	Age: 32.3 (12.4) Female: 50.9% Education: NR White: NR	BAI (trait)	Respironics Actiwatch Spectrum Pro (research)	Sleep (SE)
Zhang et al, 2025 (United States) [58]	Cross-sectional	10,129	General population	Age: median 48.0 (IQR 36.0-58.0) Female: 69.3% Education: NR White: 91.4%	GAD-7 (trait)	Fitbit (consumer)	Sleep (eg, TST, wake time, sleep start, percentage of long sleep days, and percentage of short sleep days) Physical activity (eg, steps, sedentary duration, light activity duration, and total active duration) Heart rate metrics (eg, average heart rate and SD)
Zheng et al, 2024 (United States) [59]	Longitudinal (4.5 years)	6785	General population	Age: median 50.2 (IQR 35.7-61.5) Female: 70.8% Education: NR White: 83.6%	Electronic health records	Fitbit (consumer)	Sleep (TST, sleep stages, and sleep regularity)
Bai et al, 2024 (United States) [60]	Longitudinal (16 weeks)	167	University students	Age: NR Female: 79% Education: NR White: 92.2%	Daily survey measuring anxiety	Apple Watch (consumer)	Physical activity (MVPA and steps)
Jo et al, 2024 (Korea) [61]	Longitudinal (4 weeks)	47	Young adults	Age: 28.7 (5.9) Female: 48.9% Education: NR White: NR	GAD-7 (trait)	Samsung Galaxy Active 2 (consumer)	HRV[u] (time-domain measures: RMSSD[v], SDNN[w], SDSD[x], and PNN50[y]; frequency-domain measures: LF[z], HF[aa], and LF/HF[ab])
Kandola et al, 2021 (United Kingdom) [62]	Longitudinal (2 years)	60,235	General population	Age: 55.9 (7.7) Female: 56% Education: college/university degree 48%; a-level/as levels/equivalent 14%; GCSEs[ac]/O-levels/equivalent 19%; CSEs[ad]/equivalent 3.4%; NVQ[ae]/HND[af]/equivalent 4.9%; other professional qualifications 4.9%; none (6.1%) White: 98%	GAD-7 (trait)	Axivity AX3 (research)	Activity (SB, LPA, and MVPA)

XSL•FO
RenderX

Study (location)	Study design (follow-up time)	Participants, N	Population type	Age (years), mean (SD) Sex (female), % Education, mean (SD) or categories % Ethnicity (White), %	Anxiety measure (type)	Wearable device (research/consumer-grade)	Digital biomarker type (subtype)
Liu et al, 2025 (China) [63]	Longitudinal (7-9 years)	84,570	General population	Age: 56.2 (7.8) Female: 55.5% Education: college/university degree 34.8%, other levels of education (64%), unknown (1.2%) White: 96.8%	Patient Health Questionnaire, *ICD-10*[ag] codes (trait)	Axivity AX3 (research)	Physical activity (MVPA)
McDuff et al, 2025 (United States) [64]	Longitudinal (4 weeks)	237	General population	Age: 45.1 (11.8) Female: 67.5% Education: NR White: 87.7%	GAD-7 (trait)	Fitbit Sense 2TM (consumer)	Skin conductance level, heart rate, skin temperature, and step count
Okawara et al, 2024 (Japan) [65]	Longitudinal (60 days)	279	Office workers	Age: 42.0 (10.0) Female: 16.1% Education: NR White: NR	STAI trait subscale (trait)	Apple Watch (consumer)	HRV (HF, LF, LF/HF ratio, and LnccL/H[ah])
Presby et al, 2025 (Switzerland, United States, Australia) [66]	Longitudinal (13 months)	170,320	General population	Age: 37.39 (10.36) Female: 33% Education: NR White: 33.4%	Generalized Anxiety Disorder 2-item (trait)	WHOOP strap version 3.0 and 4.0 (consumer)	Sleep (TST, SE, SD of TST, sleep consistency, sleep start, and wake start) Heart rate (HRV and resting heart rate) Physical activity (total activity)
Straus et al, 2023 (United States) [67]	Longitudinal (8 weeks)	2021	Trauma survivors from the emergency department	Age: 35.8 (13.0) Female: 62.2% Education: <high school 12%, high school graduate 25.6%, some college 41.3%, college graduate 20.7% White: 34.3%	Smartphone-based questionnaires (state—measured 6 times over the 8-week period)	Verily Life Sciences (research)	Sleep (number of transitions between sleep and wake)
Wang et al, 2025 (China) [68]	Longitudinal (12.6 years)	91,800	General population	Age: 61.8 (7.9) Female: 56.1% Education: college above 43.7%, college below 56.3% White: 97%	*ICD-10* (trait)	Axivity AX3 (research)	Physical activity (LPA, moderate intensity, vigorous intensity, and MVPA)

XSL•FO
RenderX

Study (location)	Study design (follow-up time)	Participants, N	Population type	Age (years), mean (SD) Sex (female), % Education, mean (SD) or categories % Ethnicity (White), %	Anxiety measure (type)	Wearable device (research/consumer-grade)	Digital biomarker type (subtype)
Yu et al, 2025 (Japan) [69]	Longitudinal (2 years)	71,556	Healthy adults aged 40-69 years	Age: 62.11 (7.83) Female: 54.4% Education: college or higher 44.4% White: 97.3%	GAD-7 (trait)	Axivity AX3 (research)	Physical activity (MVPA, LPA, and SB)

[a]NR: not reported.

[b]GAD-7: Generalized Anxiety Disorder-7.

[c]SE: sleep efficiency.

[d]Studies included in meta-analyses.

eDSM-5: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition).

[f]SOL: sleep onset latency.

[g]TST: total sleep time.

[h]WASO: wake after sleep onset.

[i]SB: sedentary behavior.

[j]STAI: State-Trait Anxiety Inventory.

[k]HADS-A: Hospital Anxiety and Depression Scale-Anxiety section.

[l]LPA: light-intensity physical activity.

[m]MVPA: moderate to vigorous physical activity.

[n]DASS-A: Depression, Anxiety, and Stress Scale-Anxiety Subscale.

[o]BAI: Beck Anxiety Inventory.

[p]TIB: time in bed.

[q]SAS: Self-Rating Anxiety Scale.

[r]SOT: sleep onset time.

[s]WT: wake time.

[t]PANAS: Positive and Negative Affect Schedule.

[u]HRV: heart rate variability.

[v]RMSSD: root mean square of successive differences between RR intervals (intervals between successive R peaks) interval differences.

[w]SDNN: SD of NN intervals (intervals between successive normal heartbeats).

[x]SDSD: SD of RR interval intervals.

[y]PNN50: percentage of successive RR intervals that differ by more than 50 ms.

[z]LF: absolute power of the low-frequency band.

[aa]HF: absolute power of the high-frequency band.

[ab]LF/HF: ratio of LF to HF power.

[ac]GCSE: General Certificate of Secondary Education.

[ad]CSES: Certificate of Secondary Education.

[ae]NVQ: National Vocational Qualification.

[af]HND: Higher National Diploma.

agICD-10: International Statistical Classification of Diseases, Tenth Revision.

[ah]LnccL/H: log-transformed coefficient of component variance of the low-frequency component to high-frequency component power ratio.

Most studies assessed only sleep (k=21, 58.3%), 8 (22.2%) studies assessed only physical activity, 2 (5.6%) studies assessed only HRV, and 5 (13.9%) studies assessed more than one type. The majority of sleep studies assessed the following 4 sleep metrics: total sleep time (TST; ie, total time spent in bed), wake after sleep onset (WASO; ie, time spent awake after initially falling asleep), SE (ie, ratio of TST to the total time spent in bed), and sleep onset latency (SOL; ie, time it takes to fall asleep after going to bed). Regarding physical activity, sedentary behavior, light physical activity, and moderate to vigorous physical activity (MVPA) were most commonly assessed.

Studies Using Machine Learning Approaches

Characteristics of studies using machine learning approaches are presented in Table 2. A total of 8 studies from 7 articles were included. The study sizes ranged from 40 to 727 participants (median 326, IQR 66-410). All studies used algorithms trained with a labeled dataset to select predictors for predicting anxiety (ie, supervised learning). Included studies used a variety of digital biomarkers of health as predictors in their models, for example, HRV, sleep, and activity.

XSL•FO
RenderX

Table 2. Characteristics of samples for studies using machine learning approaches. All studies used supervised learning, with sample sizes ranging from 40 to 727 (median 326, IQR 66-410).

Study	Population type	Participants, N	Age (years), mean (SD) Sex (female), %	Ground truth assessment, anxiety measure (type)	Wearable device (consumer or research-grade)	Machine learning task	Predictors	Algorithms	Validation method
Coutts et al, 2020 (United Kingdom) Trial 1 [70]	University students	68	Age: 21 (NR[a]) Female: 64%	DASS-A[b] (state) and STAI[c] (trait and state)	Biobeam band (research)	Binary classification (supervised)	Heart rate variability	Deep neural networks (LSTM[d])	Train-test split: training (80%), test (10%), validation (10%)
Coutts et al, 2020 (United Kingdom) Trial 2 [70]	University students	584	Age: NR Female: NR	DASS-A (state) and STAI (trait and state)	Biobeam band (research)	Binary classification (supervised)	Heart rate variability	Deep neural networks (LSTM)	Train-test split: training (80%), test (10%), validation (10%)
Fukuda et al, 2020 (Japan) [71]	Office workers	60	Age: NR Female: NR	DAMS[e] (state)	Fitbit Charge 3 (Consumer)	Binary classification (supervised)	Sleep actigraphy data (13 features, eg, total sleep time and number of wake)	RF[f]	Leave-one-person-out cross-validation
He et al (2025) [72]	University students	40	Age: 21.7 (3.11) Female: 52.5%	STAI Y6 (state)	E4 Empatica Hexoskin smart shirt	Binary classification (unsupervised)	Wrist-worn wearable: skin temprature and EDA[g] Smart shirt: ECG[h]	SVM[i], RF, KNN[j], naïve Bayes	Leave-one-participant-out
Lee et al, 2022 (Korea) [73]	Older adults with mild cognitive impairment	352	Age: 72.48 (5.9) Female: 73%	Clinical diagnosis (Trait)	Fitbit Alta HR2 (Consumer)	Binary classification (supervised)	Model 1: 24-hour activity rhythms and sleep pattern[k] Model 2: Model 1+ minimal K-GAI[l] (5 items)	Logistic regression, SVM, RF, GBM[m]	10-fold stratified cross-validation
Lee et al, 2024 (Korea) [74]	Older adults with mild cognitive impairment	352	Age: 72.5 (5.9) Female: 73%	Clinical diagnosis (Trait)	Fitbit Alta HR2 (Consumer)	Binary classification (supervised)	Model 1: activity, sleep Model 2: activity + minimal K-GAI	Convolutional neural network, LSTM, residual network	Cross-validation
Saylam et al, 2023 (United States) [75]	Office workers	727	Age: NR Female: NR	Single item[n] (state)	Garmin smartwatch (consumer)	Regression	Activity, stress, sleep, heart rate	RF, XGBoost[o], LSTM	80% training, 20% testing
Saylam et al, 2024 (United States) [76]	College students	Baseline: ~700 Follow-up: 300	Age: NR Female: NR	BAI[p]	Fitbit (consumer)	Regression	Activity and sleep	RF, XGBoost, LSTM	Training: 5 semesters Validation: 2 semesters Testing: 2 semesters

XSL•FO
RenderX

[a]NR: not reported.

[b]DASS-A: Depression, Anxiety, and Stress Scale-Anxiety Subscale.

[c]STAI: State Trait Anxiety Inventory

[d]LSTM: Long Short-Term Memory networks.

[e]DAMS: Depression and Anxiety Mood Scale.

[f]RF: random forest.

[g]EDA: electrodermal activity.

[h]ECG: electrocardiogram.

[i]SVM: Support Vector Machine.

[j]KNN: k-nearest neighbors.

[k]Model 1: 24-hour activity rhythms (interdaily stability [the stability of an activity rhythm of a daily pattern], intra-daily variability [the variability of the activity rhythms throughout the day], dominant rest phase onset [the start time of the 5-hour period with the least activity within 24 hours], and sleep patterns (total sleep time, sleep onset latency, wake after sleep onset, and sleep quality).

[l]K-GAI: Korean version of Geriatric Anxiety Inventory.

[m]GBM: gradient boosting machine.

[n]Single-item question "Please select the response that shows how anxious you feel at the moment. Response scale: 1 (not at all anxious), 2 (a little anxious), 3 (moderately anxious), 4 (very anxious), and 5 (extremely anxious)."

[o]XGBoost: extreme gradient boosting.

[p]BAI: Beck Anxiety Inventory.

Quality Assessment

Studies Using Inferential Statistical Approaches

For studies that conducted both cross-sectional and longitudinal analyses, quality assessment was based on the longitudinal studies. Of the 36 studies, 12 (33.3%) received a "Good" quality rating; the remainder received a "Fair" rating (k=24, 66.7%; Multimedia Appendix 5).

Studies Using Machine Learning Approaches

The quality assessment tool for studies using machine learning approaches assesses the risk of bias across 4 domains: participants, index test, reference standard, and analysis. A total of 5 studies were rated as having a high risk of bias in the participants' domain. For the index test, ground truth, and analysis, all studies were rated as either low risk or unclear. There were either unclear or low concerns regarding the applicability of the studies to our review question across the participants, index test, and reference standard domains (Multimedia Appendix 5 [70-76]).

Qualitative Synthesis of Inferential Statistical Studies

Sleep

A total of 8 studies investigated the association between sleep metrics not included in the meta-analyses and anxiety. Sleep start variability [41], sleep onset time [48], wake time [48,55,58], bedtime [55,58], number of nocturnal awakenings [53], sleep variability [53], and sleep timing [54] were not associated with anxiety symptoms. Lower average sleep among doctors who were on call was negatively associated with anxiety symptoms [44]. Other studies found that greater sleep irregularity [54], greater sleep duration variability [45,58], higher sleep start variability [58], greater wake time variability [58], later wake time [58], and a higher percentage of long (>10 hours) and short (<6 hours) sleep days [58] were associated with more anxiety symptoms.

A total of 7 studies reported the longitudinal association between sleep and anxiety [36,38,41,50,59,66,67]. While 1 study found no association between baseline sleep metrics (ie, SE, WASO, SOL, and TST) and anxiety at 6 months follow-up in paramedics [50], the majority of studies did observe an association over varying periods of time. Lower TST compared to an individual's personal TST average was associated with higher anxiety levels on the next day [38]. Anxious mood also predicted next day's poorer sleep (ie, lower SE, longer WASO, and less TST) in individuals with clinical anxiety, but not in healthy controls [30]. Worsening sleep continuity (measured by the number of transitions between sleep and wake) was associated with worsening anxiety over an 8-week period [67], lower SE and longer SOL were associated with more anxiety symptoms at 1 year follow-up [41], and sleep patterns (increased sleep irregularity, decreased deep and rapid eye movement (REM) sleep percentage, and increased light sleep percentage) captured through continuous monitoring over a median period of 4.5 (IQR 2.5-6.5) years were associated with increased odds of GAD [59]. Additionally, they identified a nonlinear, J-shaped relationship between average daily sleep duration and GAD—compared with the median average daily sleep of 6.8 hours, individuals with an average daily sleep duration of 5 hours or 10 hours had increased odds of GAD. One study found that more consistent wake and sleep times were associated with lower anxiety symptoms [66].

Activity

A total of 6 studies investigated the association between sedentary behavior and anxiety [37,40,46,47,58,62]. While 1 study found no association between sedentary behavior and anxiety symptoms [47], the remainder did observe associations. Individuals with moderate to severe anxiety symptoms in primary care settings spent more time in sedentary activity per day than participants with no significant anxiety [40]. This finding was supported by a large study in the UK general population [58]. A study using isotemporal substitution models (which assess the effect of hypothetically replacing 1 activity with another while keeping the total time constant) found similar

XSL•FO
RenderX

findings; they found that replacing MVPA with sedentary behavior was associated with higher odds of having clinically relevant anxiety symptoms [46]. Another study found that while sedentary time was not associated with current GAD and anxiety symptoms [37], it was associated with lower odds of lifetime GAD, and in males when stratified by sex [37]. The association between sedentary behavior and anxiety symptoms was also supported in a longitudinal study, with more baseline sedentary behavior associated with more anxiety symptoms at follow-up [62].

A total of 5 studies examined the association between light activity, MVPA, and anxiety [40,46,58,60,62] and reported mixed findings. Regarding light physical activity, 1 study found that lower light activity duration was associated with more anxiety symptoms. Consistent with this finding, 1 study found that individuals with moderate to severe anxiety symptoms spent less time in light physical activity than those without significant levels of anxiety [40]. Using isotemporal substitution models, 1 study found that replacing sedentary behavior with light physical activity was associated with a decrease in anxiety symptoms [40], while another study found no association [46], and a longitudinal study (UK Biobank) found that replacing sedentary behavior with light physical activity was associated with lower anxiety symptoms at follow-up [62]. Regarding MVPA, 1 study found that replacing sedentary behavior with MVPA was associated with lower odds of clinically relevant anxiety symptoms [46], but 1 study found no association [40]. One longitudinal study found that replacing baseline sedentary behavior with MVPA was associated with lower anxiety symptoms at follow-up [62], and another found that more anxiety symptoms (measured using a daily survey) were associated with lower MVPA [60].

A total of 3 studies examined associations between physical activity and anxiety using data from the UK Biobank [63,68,69]. One study reported an L-shaped association, where higher levels of moderate-intensity physical activity and vigorous-intensity physical activity were associated with lower incident anxiety [68]. Another study also identified a nonlinear relationship, with higher levels of MVPA and total physical activity associated with reduced anxiety risk, whereas light physical activity and sedentary behavior were not associated with anxiety [69]. The third study found that, compared with inactive individuals, those who were regularly active during the week or active only during weekends had lower risks of anxiety [63].

A total of 3 studies examined the associations between step count and anxiety. One study found that lower daily step count and lower SD of daily step count were associated with more anxiety symptoms [58]. A longitudinal study found that higher step count was associated with lower anxiety symptoms [60]; another longitudinal study found no association [64].

Of the 5 studies investigating other activity metrics (average daily exercise duration, total sedentary time on weekdays and weekends, and overall physical activity) [37,44,47,58,66], only 2 did not find associations with anxiety symptoms [44,47]. More physical activity was associated with lower odds of current and lifetime GAD [37] and lower anxiety symptoms [37,66].

Heart Rate

A total of 5 studies investigated the association between heart rate metrics and anxiety symptoms [58,61,64-66]. Four studies investigated the association between HRV metrics and anxiety [61,64-66]. HRV metrics can be examined in either the time-domain (measuring the amount of variability in intervals between heartbeats over time) or the frequency domain (measuring how HRV is distributed across different frequency bands). Within the frequency domain, the high-frequency component has been linked with parasympathetic nervous system (PNS) activation, and the low-frequency component has been linked with sympathetic nervous system (SNS) activation. Both central tendency and variability of metrics can be assessed in the HRV time and frequency domains.

One study focusing on the frequency domain found that, compared to a low trait anxiety group, the high trait anxiety group showed greater variability in HRV frequency metrics associated with PNS activation and the consistency of this activation, SNS activation, and the balance of PNS and SNS activation (eg, low-frequency/high-frequency HRV) in addition to daily changes of this balance [65]. Additionally, a higher mean and greater variance in daily changes in the balance of PNS and SNS activation were associated with increased odds of having trait anxiety [65]. In contrast, a separate study observed that higher mean low frequency (SNS activation) was associated with lower anxiety scores, and that mean high-frequency HRV (PNS activation) and the low frequency/high frequency ratio were not associated with anxiety symptoms [61]. This study also examined the association between HRV time-domain metrics and anxiety at 2 different time points, and found that higher mean deviation from the average HRV and short-term variations in heart rate were associated with lower anxiety scores [61]. One study found no association between HRV root mean square of successive differences and anxiety, a time-domain HRV metric that measures beat-to-beat changes in heart rate [64]. A study examined HRV but did not specify whether it was calculated in the time or frequency domain and found that higher HRV was associated with lower anxiety symptoms [66].

A total of 2 studies examined associations between average heart rate and anxiety, all reporting consistent findings. One study found that higher average heart rate and lower SD of heart rate were associated with more anxiety symptoms [58]. Another study found that individuals with anxiety had elevated heart rate compared with those without [64]. One study examined resting heart rate and found that higher levels were associated with more anxiety symptoms [66].

Other Metrics

One study found that individuals who were anxious had an elevated skin conductance response and skin temperature compared with those without anxiety [64].

Qualitative Synthesis of Machine Learning Studies

Of the 8 studies included in this review, some studies used multiple algorithms within the same study, each with multiple performance metrics. Accuracy measures correct classifications (ie, presence/absence of anxiety). Sensitivity is the ability to

XSL•FO
RenderX

detect true positives (ie, correct identification of individuals with anxiety), while precision is the correct classification among predicted positives (ie, how many individuals classified as having anxiety actually have anxiety). F_1-score balances precision and sensitivity, and mean absolute error (MAE) measures the average prediction error, with lower MAE signifying better performance.

A total of 3 studies reported 13 accuracy estimates, ranging from 56.3% to 94.6% [70,71,73]. Two studies reported 9 sensitivity estimates (range 56.3%-94.6%), 9 precision estimates (range 56.3%-94.8%), and 9 F_1-score estimates (range 55.9%-94.6%) [71,73]. Overall, 56 MAE estimates were reported from 2 studies [75,76], ranging from 0.05 to 0.63.

One article using 2 cohorts found better accuracy in predicting anxiety using frequency-domain HRV and nighttime HRV (range 63.9%-69.4%) data than time-domain HRV and daytime (range 63.2%-63.4%) data [70]. They also found that nighttime HRV data (maximum accuracy: 69.4%) performed better than daytime data (maximum accuracy: 63.2%) [70].

One study with older adults who had mild cognitive impairment found that combining activity (24-hour activity rhythm), sleep, and anxiety questionnaire items improved model performance across accuracy (91.0%-94.6% vs 56.3%-59.9%), precision (91.4%-94.8% vs 56.4%-61.0%), recall (91.0%-94.6% vs 56.3%-59.9%), and F_1-scores (90.9%-94.6% vs 55.9%-59.0%) compared to models that excluded anxiety questionnaire data [73]. Similarly, another study using the same sample found that combining activity (step count) and sleep stages achieved a predictive success rate for participants without anxiety (98%) but had a low predictive success rate identifying those with anxiety (18%) [74]. Models including anxiety questionnaire items improved the predictive success rate for identifying participants with anxiety (82%) and maintained a good predictive success rate at identifying those without anxiety (95%). These studies did not report feature importance in their models.

A study with university students used digital biomarkers of health—HRV, electrodermal activity, and skin temperature—collected from wrist-worn wearables and smart shirts to predict low versus high anxiety levels. The best classifier (support vector machine) achieved an overall accuracy of 90.9%, with a precision of 0.90-0.92, recall of 0.90-0.92, and F_1-score of 0.91 [72].

Two studies predicted anxiety in office workers [71,75]. One study, using 13 sleep metrics, achieved an F_1-score of 75.6% and identified that the top 5 important predictors were REM sleep time ratio, REM sleep minutes, light sleep time ratio, deep sleep time ratio, and total wake time ratio [71]. Another study, using data from wearable devices (activity, stress, sleep, and heart rate data) and smartphones, found that the best non–time-based model (ie, without incorporating temporal data) predicted anxiety with an MAE of 0.6316. Time-based models showed improved performance, with an MAE of 0.3729 when using data from the previous 3 days to predict the next day's anxiety [75]. They found that the top 20 predictors included measures of sleep (bedtime, wake time, and sleep duration), stress (low stress duration), and activity (highly active duration).

Finally, 1 study comparing 4 modeling approaches—conventional, multitask learning approach (model trained to solve multiple tasks simultaneously), time-based approach, and time-based multitask approach—found that time-based models using data from the previous 15 days to predict next day's anxiety achieved the lowest MAE (0.0047), followed by the time-based multitask model (MAE 0.0083), the conventional model (MAE 0.1277), and the multitask model (MAE 0.1347) [76]. While activity, sleep, and heart rate were considered, the top 20 predictors did not include any of these metrics.

Quantitative Synthesis of Cross-Sectional Studies

Random-effects meta-analyses were performed for studies that provided combinable effect sizes. All studies were cross-sectional and assessed associations between sleep metrics and anxiety symptoms.

No association between SOL and anxiety symptoms was observed (k=9, N=3643; Fisher z=0.04, 95% CI –0.07 to 0.15; P=.08; Figure 2 [36,39,41,42,48,50-52] and Multimedia Appendix 6), with evidence of moderate to substantial heterogeneity between studies (I^2=49.77%). The prediction interval ranged from –0.19 to 0.27. Around half of the included studies were rated as fair quality (k=5, 55.6%). The results remained unchanged in sensitivity analyses based on anxiety type (trait and state) and young adults (Multimedia Appendix 6). One study was not eligible for meta-analysis but found that higher SOL was associated with more anxiety symptoms [45].

XSL•FO
RenderX

Figure 2. Forest plot of the associations between sleep onset latency and anxiety symptoms. Effect sizes are Fisher z with corresponding 95% CIs. No association between sleep onset latency and anxiety symptoms was observed (k=9, N=3643; Fisher z=0.04, 95% CI –0.07 to 0.15; *P*=.08). ANX: Goldberg Anxiety Scale; BAI: Beck Anxiety Inventory; DASS: Depression, Anxiety, and Stress Scale-Anxiety Subscale; GAD-7: Generalized Anxiety Disorder-7; HKSJ: Hartung-Knapp-Sidik-Jonkman; STAI: State-Trait Anxiety Inventory [36,39,41,42,48,50,51,52].

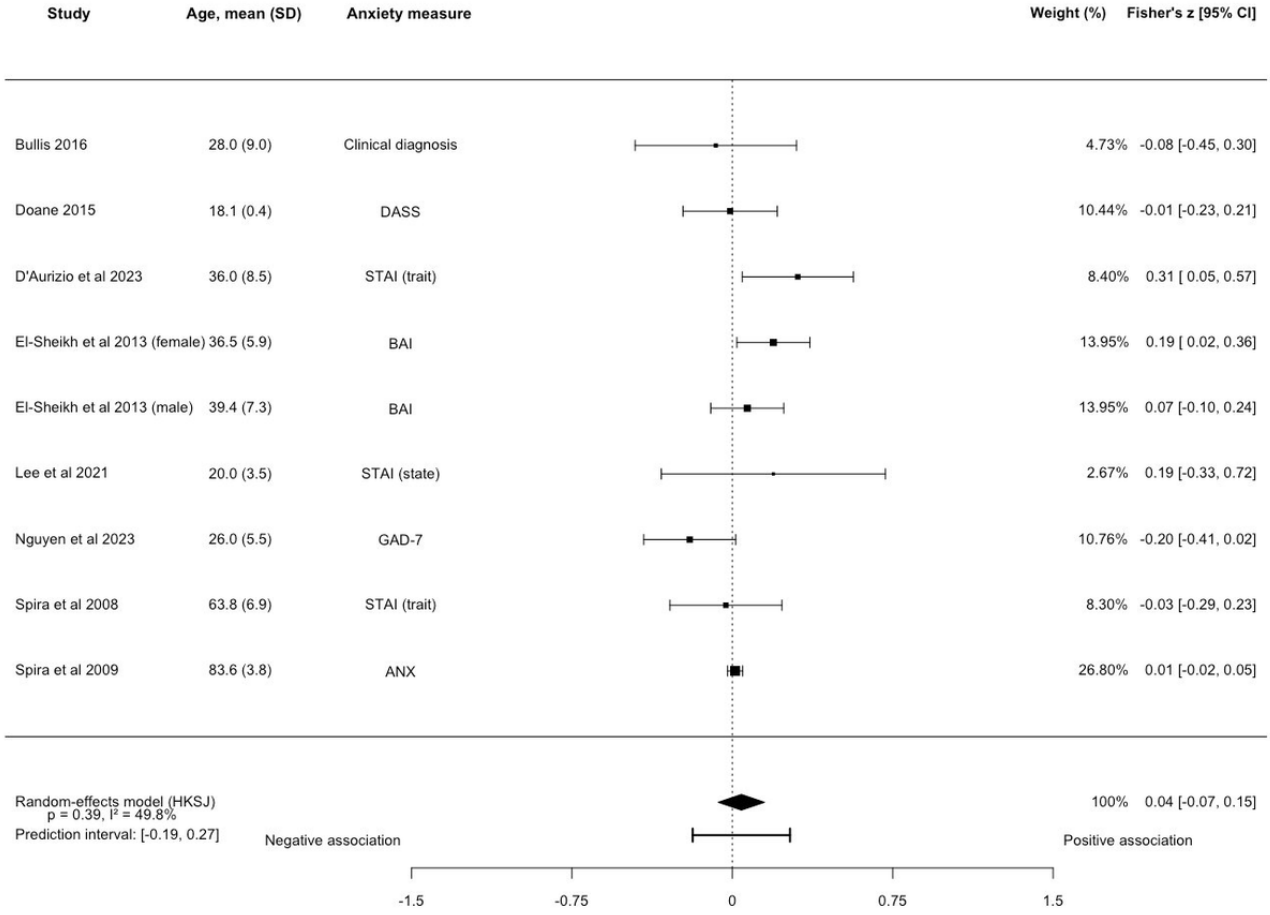

No association between TST and anxiety symptoms was observed (k=12, N=14,721; Fisher *z*=0.009, 95% CI –0.01 to 0.03; *P*=.28; Figure 3 [36,39,41,42,45,48,50,52,55,56,58] and Multimedia Appendix 6), with very low heterogeneity between studies (I^2=3.89%). The prediction interval ranged from –0.02 to 0.03. The majority of the studies were rated as fair quality (k=9, 75%). The results remained unchanged in sensitivity analyses on anxiety type and young adults (Multimedia Appendix 6). Five studies were not eligible for quantitative synthesis. Findings from 4 studies aligned with those from the meta-analysis [38,46,53,66]. One study found a positive association between poor sleep—as measured by a composite including TST and other sleep metrics—and anxiety symptoms [43]. One study reported no association [49].

XSL•FO
RenderX

Figure 3. Forest plot of the association between total sleep time and anxiety symptoms. Effect sizes are Fisher z with corresponding 95% CIs. No association between total sleep time and anxiety symptoms was observed (k=12, N=14,721; Fisher z=0.009, 95% CI –0.01 to 0.03; *P*=.28). ANX: Goldberg Anxiety Scale; BAI: Beck Anxiety Inventory; DASS: Depression, Anxiety, and Stress Scale-Anxiety Subscale; GAD-7: Generalized Anxiety Disorder-7; HKSJ: Hartung-Knapp-Sidik-Jonkman; STAI: State-Trait Anxiety Inventory [36,39,41,42,45,48,50,52,55,56,58].

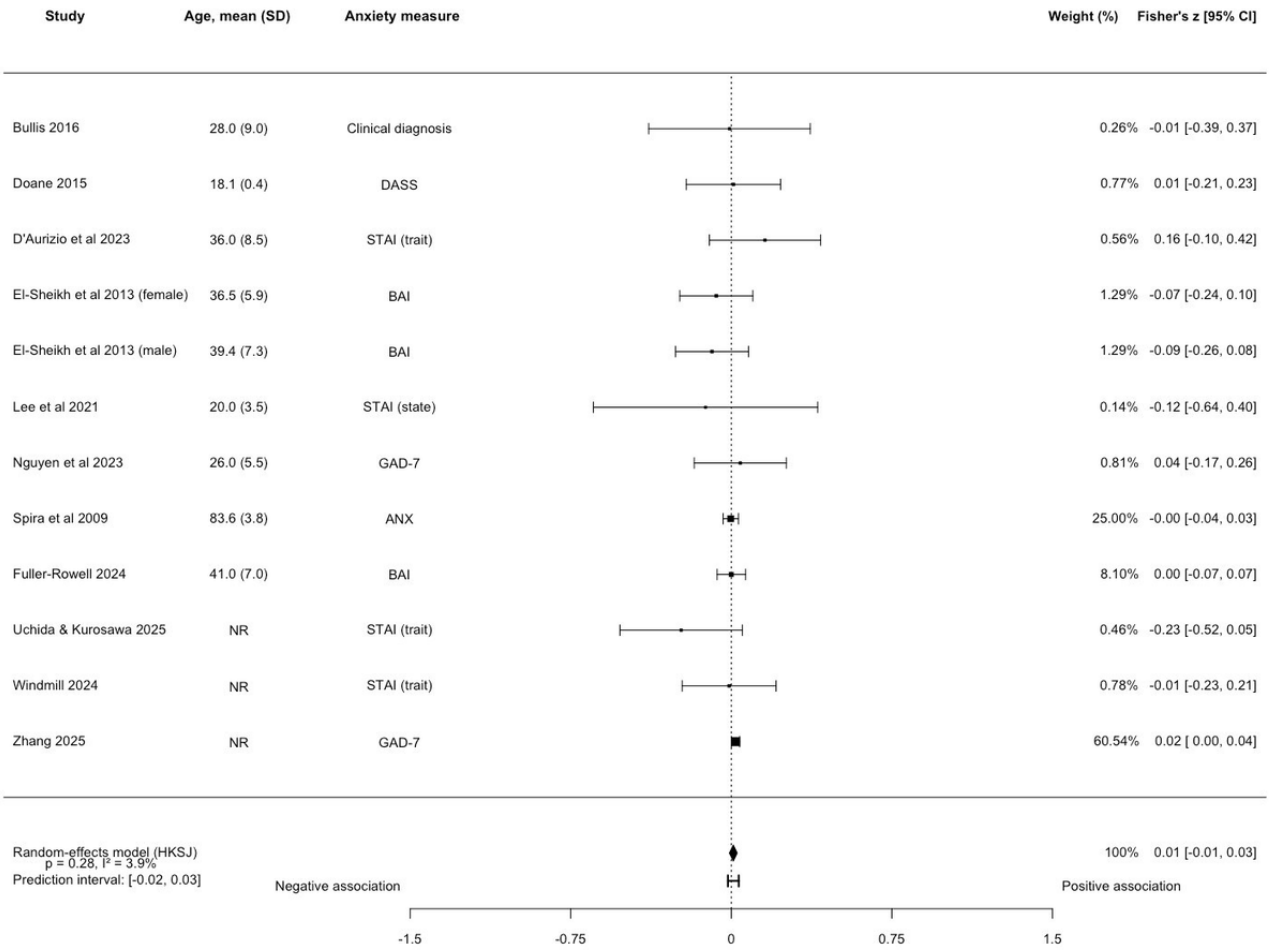

No association between SE and anxiety symptoms was observed (k=9, N=3710; Fisher *z*=–0.07, 95% CI –0.14 to 0.002; *P*=.06; Figure 4 [36,39,41,42,50-52,56] and Multimedia Appendix 6), with low heterogeneity between studies (I^2=20.97%). The prediction interval ranged from –0.19 to 0.05. The majority of studies were rated as fair quality (k=7, 77.8%). These results remained unchanged in sensitivity analyses on anxiety type and young adults (Multimedia Appendix 6). Of the 5 studies ineligible for quantitative synthesis, 3 found no association between SE and anxiety symptoms [45,49,57], 1 found a positive association [66], while 2 studies reported negative associations [43,77].

XSL•FO
RenderX

Figure 4. Forest plot of the association between sleep efficiency and anxiety symptoms. Effect sizes are Fisher z with corresponding 95% CIs. No association between sleep efficiency and anxiety symptoms (k=9, N=3710; Fisher z=-0.07, 95% CI –0.14 to 0.002; *P*=.06). BAI: Beck Anxiety Inventory; DASS: Depression, Anxiety, and Stress Scale-Anxiety Subscale; GAD-7: Generalized Anxiety Disorder-7; HKSJ: Hartung-Knapp-Sidik-Jonkman; STAI: State-Trait Anxiety Inventory [36,39,41,42,50,51,52,56].

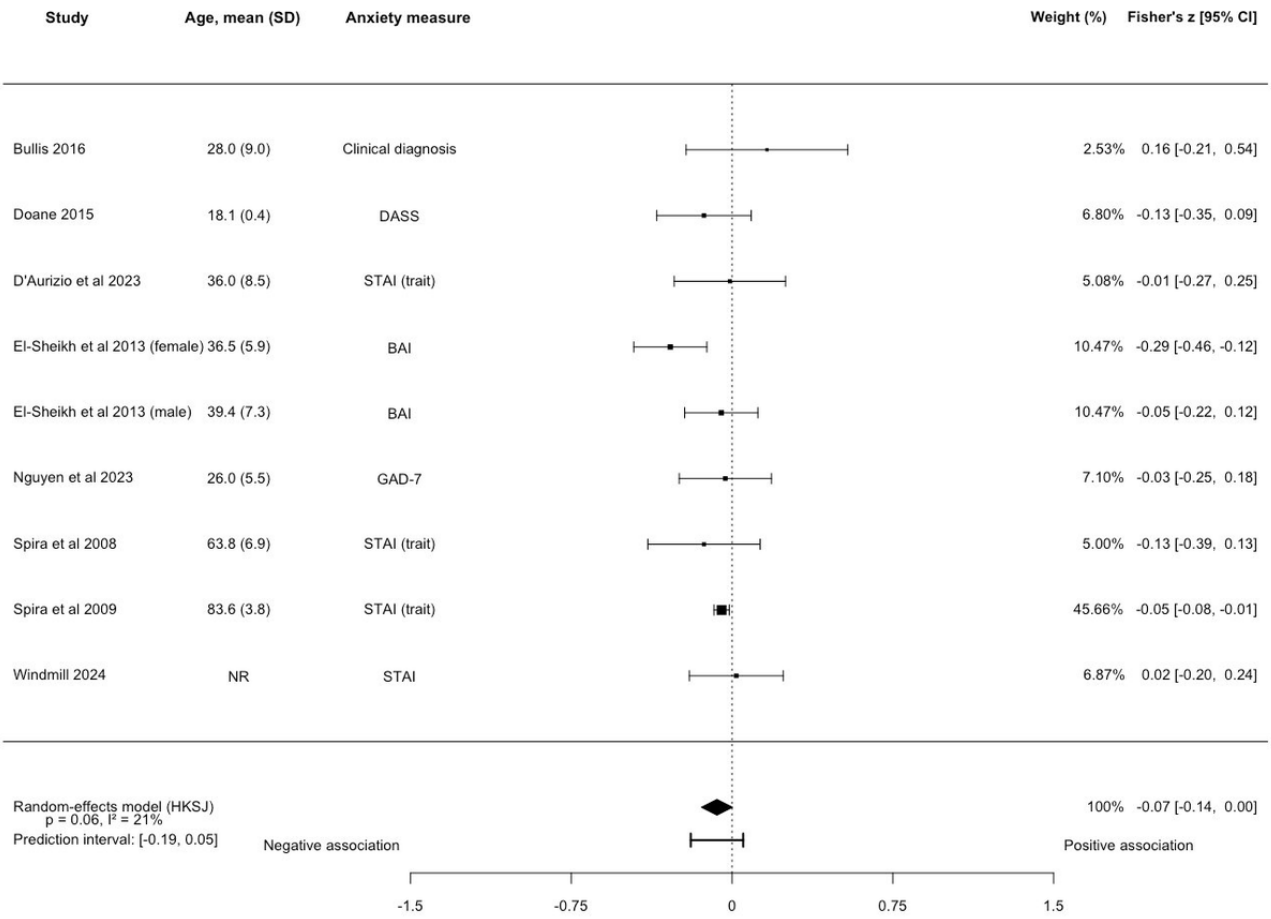

No association between WASO and anxiety symptoms was observed (k=6, N=3291; Fisher *z*=0.13, 95% CI –0.04 to 0.30; *P*=.11; Figure 5 [36,39,48,50-52] and Multimedia Appendix 6), with low to moderate heterogeneity between studies (I^2=40.62%). The prediction interval ranged from –0.15 to 0.41. Most studies were rated as fair quality (k=4, 66.7%). These results remained unchanged in sensitivity analyses on anxiety type and young adults (Multimedia Appendix 6). One study that was not eligible for quantitative synthesis also found no association [49].

XSL•FO
RenderX

Figure 5. Forest plot of the association between wake after sleep onset and anxiety symptoms. Effect sizes are Fisher z with corresponding 95% CIs. No association between wake after sleep onset and more anxiety symptoms (k=6, N=3291; Fisher z=0.13, 95% CI –0.04 to 0.30; *P*=.11). ANX: Goldberg Anxiety Scale; GAD-7: Generalized Anxiety Disorder-7; HKSJ: Hartung-Knapp-Sidik-Jonkman; STAI: State-Trait Anxiety Inventory [36,39,48,50,51,52].

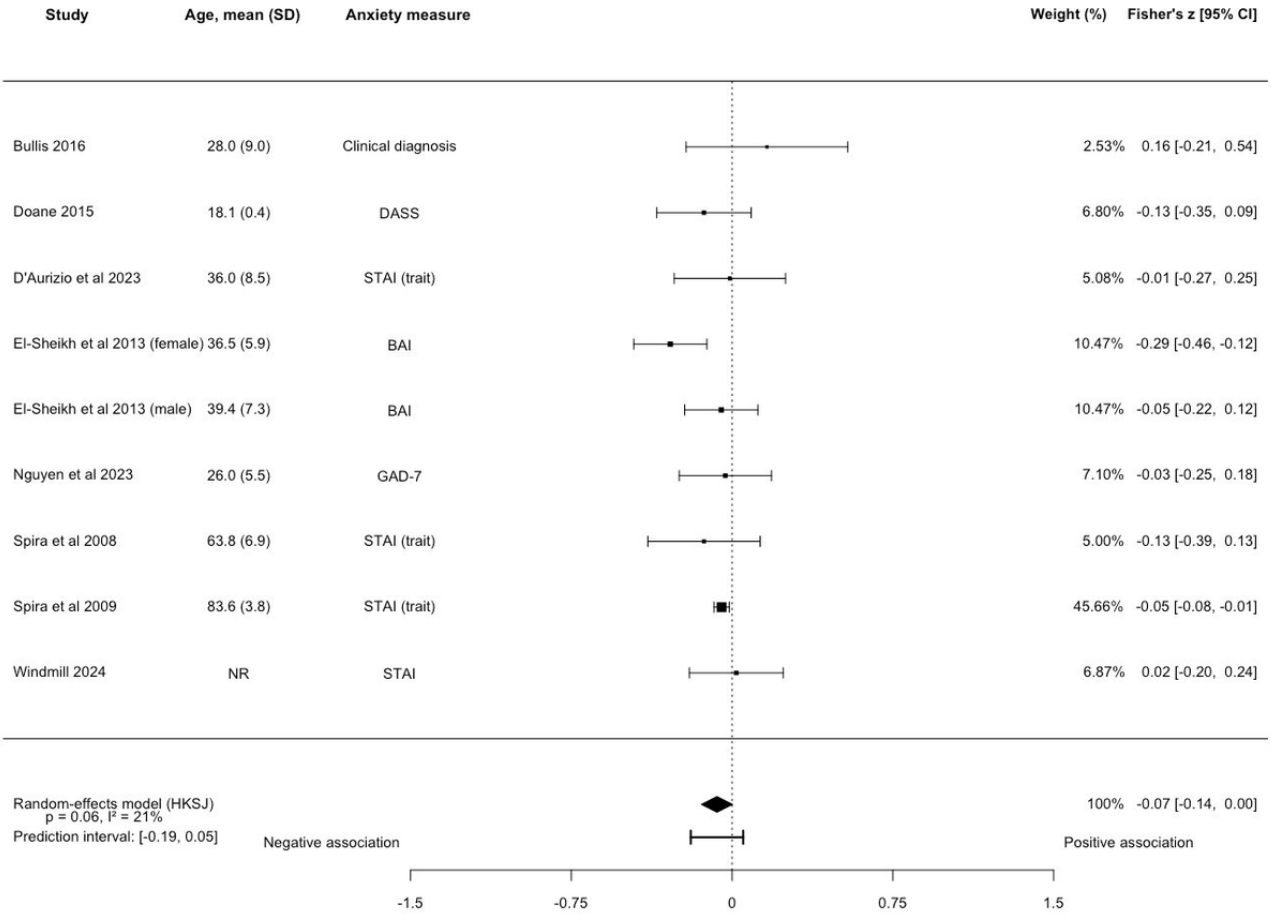

The Hartung-Knapp-Sidik-Jonkman–adjusted CIs reflect uncertainty around the average associations across studies, whereas the prediction intervals estimate the range of effects that may be observed in future studies conducted in different populations or settings. Across the sleep metrics examined, prediction intervals were generally similar to, but slightly wider than, the CIs, suggesting that effects observed in future studies are likely to be close to the average pooled estimates.

Discussion

Overview

This review aimed to advance scientific understanding of the association between digital biomarkers of health collected from wrist-worn devices and anxiety. Specific interest centered on determining which digital biomarkers of health have been the focus of existing research, and which demonstrate the strongest evidence of an association with anxiety. A total of 36 studies were included in this systematic review. Meta-analyses were conducted for 4 sleep metrics (SE, WASO, TST, and SOL), and none of these metrics were associated with anxiety. It was not possible to conduct meta-analyses for physical activity and heart rate metrics; however, qualitative synthesis suggests that lower physical activity levels and elevated heart rate were associated with higher levels of anxiety.

Sleep

Our meta-analytic findings showed that SE, WASO, TST, and SOL were not statistically associated with anxiety. However, the association for SE was close to the significance threshold, suggesting a possible relationship. Heterogeneity was low for TST and SE, indicating that studies included in these meta-analyses found relatively consistent findings. In contrast, heterogeneity was moderate for WASO and SOL, suggesting that results varied across studies, potentially due to differences in populations, devices used to measure sleep metrics, or anxiety measurement. Subgroup analyses were not conducted as this was not prespecified in the protocol, and the number of studies contributing to each meta-analysis was small [78]. Finally, these findings should also be interpreted in the context of study quality. Most studies were rated as fair quality, reflecting some methodological limitations such as the lack of sample size calculation and no adjustment for key potential confounding variables.

The Pittsburgh Sleep Quality Index, commonly used to assess subjective sleep, comprises 7 components (sleep quality, sleep latency, sleep duration, SE, sleep disturbance, use of sleep medication, and daytime dysfunction) [79]. Existing research using the Pittsburgh Sleep Quality Index generally reports that shorter sleep duration, poor sleep quality, lower SE, and longer SOL are associated with anxiety disorders in both young and older adults [80-83]. In addition, worse subjective sleep quality

XSL•FO
RenderX

has also been associated with more nonclinical levels of anxiety symptoms [84,85].

Research examining associations between anxiety and sleep metrics derived from polysomnography, the gold standard measurement of objective sleep, has been limited. One review identified 6 studies investigating polysomnographic sleep metrics, and found decreased SE and TST, and increased WASO in individuals with GAD compared with healthy controls [86]. Another review examined sleep metrics derived from electroencephalography and found that individuals with GAD had decreased TST, and increased WASO and SOL compared with healthy controls; however, no differences in SE were observed [16].

The lack of statistically significant associations observed in our meta-analyses contrasts with some findings from existing literature and may reflect differences in devices and methodological approaches. For instance, only 2 studies included in our meta-analysis asked participants to press an event marker to indicate when they began to fall asleep [36,87]. Without these event markers, estimating SOL may be less accurate [88], as different sleep behaviors can influence wrist movements. The majority of the studies (12/13, 92.3%) in our meta-analyses assessed anxiety symptoms, whereas the studies using polysomnography/electroencephalography focused on individuals with clinical anxiety. The lack of associations in our meta-analyses could also be attributed to differences in sample characteristics. Associations between sleep metrics and anxiety may emerge in clinical anxiety, but not in subthreshold anxiety symptoms.

Physical Activity

Our review suggests that higher levels of physical activity are associated with lower anxiety symptoms. These findings appear to be largely consistent across different aspects of physical activity, including step count, total physical activity, and physical activity intensities. These findings align with existing reviews showing the protective effects of physical activity on reducing anxiety symptoms [89,90].

Our review found that sedentary behavior was associated with greater anxiety symptoms. The observed positive association between sedentary behavior and anxiety symptoms aligns with existing findings [91,92]. A meta-analysis primarily based on studies that used self-report questionnaires found a small positive association between sedentary behavior and anxiety [92]. Similar to our review, their study populations were largely composed of healthy individuals rather than those with anxiety disorders. It may be that stronger associations emerge with more severe anxiety. For example, 1 study included in our review observed a significant association only when assessing clinical vs nonclinical anxiety symptoms [46]. No association was observed when analyzing anxiety as a continuous variable (with a mean anxiety score below the clinical threshold) [46]. This potential threshold effect is further supported by another study included in our review, which found that individuals with more severe anxiety symptoms engaged in more sedentary behavior than those with no significant anxiety symptoms [40].

Regarding physical intensities, the majority of studies examining light physical activity (k=3, 75%) and MVPA (k=3, 75%) found that spending more time in these intensities was associated with lower anxiety symptoms. Similarly, a meta-analysis of physical activity interventions found that these interventions reduced anxiety symptoms [93]. Although the number of studies within each subgroup is limited, subgroup analyses of exercise intensity found that both moderate- and high-intensity exercise reduced anxiety symptoms [93]. Taken together, our review's findings on physical activity intensities align with existing research showing that higher-intensity physical activity is effective in reducing anxiety symptoms.

HRV

Our review found inconclusive evidence on the association between HRV and anxiety, as only 4 studies included in this review examined HRV. An existing meta-analysis on HRV found that individuals with GAD had reduced high-frequency and time-domain HRV [94], which means they may have less flexibility in adapting to stress. This partially aligns with the findings from 2 studies included in this review: 1 found that higher HRV (domain not specified) was associated with anxiety symptoms [66], while another found that higher mean time-domain HRV was associated with lower anxiety symptoms [61]. This study also found that greater high-frequency HRV variability was associated with more anxiety symptoms [61]. The discrepancy in findings may be due to the focus of the existing meta-analysis on clinical anxiety, whereas the 2 studies in our review examined anxiety symptoms.

Elevated resting or average heart rate reflects chronic SNS activation (SNS regulates the body's fight-or-flight response) [95]. Compared to HRV, far fewer studies have examined average heart rate or resting heart rate in relation to anxiety. Existing evidence suggests that a higher resting heart rate is linked to greater anxiety symptoms and anxiety disorders [96,97]. Consistent with this literature, our review found that higher average heart rate and resting heart rate were associated with more anxiety symptoms.

Studies Using Machine Learning Approaches

Machine learning approaches emphasize building models that generalize to new data. These approaches offer advantages over inferential statistical methods [98]. For example, machine learning approaches can capture both linear and nonlinear relationships between variables without strong assumptions about variable relationships. However, machine learning models can be challenging to interpret due to their "black box" nature, where decision-making processes are not transparent [99].

In this review, machine learning models predicting anxiety had accuracies ranging from 56.3% to 90.9%. The accuracy reported here is somewhat comparable to the 2 reviews on machine learning models using any wearable devices to predict depression (70%-89%) [17] and anxiety (82%) [18], which used a range of predictors (eg, heart rate, activity, sleep, audio, and skin temperature). The majority of studies included in this review did not include demographic predictors, which might further enhance model accuracy. Indeed, some studies in the depression and anxiety machine learning reviews used

XSL•FO
RenderX

demographic data as predictors, further supporting the potential use of these predictors to improve the prediction of anxiety. The model with the highest performance (accuracy of 90.9%) combined digital biomarkers of health from more than 1 type of digital device (wrist-worn wearable and smart shirts) [72]. This aligns with existing research showing that models integrating digital biomarkers of health from multiple devices achieved better predictive performance than those relying on a single device [20]. For studies using only digital biomarkers of health from wrist-worn wearables in this review, the models with the highest accuracy included anxiety questionnaires as a predictor to predict clinically diagnosed anxiety. This likely inflates model performance due to circularity. When excluding models that incorporated anxiety questionnaires as inputs, the studies using only digital biomarkers of health from wrist-worn wearables reported accuracies ranging from 56.3% to 69.4%.

Wrist-Worn Wearable Devices

Digital biomarkers from wrist-worn wearables offer several benefits over traditional assessments of anxiety, as they allow for continuous data collection over long periods without the need for repeated administration. In addition, these largely affordable and minimally intrusive devices can be used in natural settings, providing valuable insights into sleep and activity habits. While both research-grade and consumer-grade devices are less precise than polysomnography in measuring sleep [100-102], this may be due to their indirect measurement (eg, based on movement). Despite this limitation, they are generally considered valid tools for sleep assessment [100,103]. Moreover, polysomnography is neither cost-effective nor practical for large-scale real-world studies. Given that the majority of the studies in our review used research-grade devices, a key remaining question is whether these findings generalize to consumer-grade wrist-worn devices, which are more widely accessible to the general population.

There are also benefits to using self-report data. Self-report questionnaires can capture subjective experiences, such as perceived levels of exercise or sleep quality, which wearables cannot measure. Further, they have the potential to record underlying reasons for poor sleep, for example, bad dreams or pain. Therefore, while wrist-worn wearables offer advantages, incorporating self-report measures can provide a more comprehensive understanding of both objective and subjective experience.

Strengths, Limitations, and Future Directions

To the best of our knowledge, this is the first review to examine the association between digital biomarkers of health from wrist-worn wearables and anxiety. We used double screening for all studies to ensure a rigorous selection process and to minimize bias. We also included gray literature to reduce potential publication bias and provide a more comprehensive review of existing studies.

This review also has several limitations. First, most studies focused on young adults. Part of the impetus for this review was to explore whether metrics from wrist-worn wearables could eventually be used to screen for dementia risk, as indicated by changes in anxiety symptoms. This is because anxiety is a potentially modifiable factor associated with dementia risk. However, the number of studies conducted in older adults is scarce. Future studies involving older adults would enrich our understanding of the association between digital biomarkers of health and anxiety in the context of dementia. Second, despite previous literature showing associations between physical activity and anxiety [104,105], few studies in our review focused on activity-related metrics. This precluded a meta-analytic synthesis of these findings. Wrist-worn devices generally provide more objective measurements of activity compared to self-report, which can be influenced by different biases [106], highlighting the need for further research. Third, although our search strategy and eligibility criteria allowed inclusion of any metrics measured by wrist-worn wearables, the majority of studies focused on sleep metrics. The search strategy was not peer-reviewed, and although it is possible that consulting with an information scientist might have led to the inclusion of additional relevant studies, this gap highlights an opportunity for future research to explore a wider range of metrics that could be associated with anxiety. Fourth, while machine learning has potential in predicting anxiety, the quality of these studies needs improvement. For example, future studies need to carefully consider the predictors used. None of the included machine learning studies considered demographics as potential predictors, despite existing research indicating sex differences in anxiety [107]. In addition, future studies should increase the interpretability of machine learning models by routinely reporting Shapley additive explanations values or feature importance scores. Additionally, most of the studies focused on a single metric (eg, sleep), despite the potential for the models to include multiple predictors. More research using machine learning approaches is needed, especially to compare whether this approach is superior to traditional approaches that use inferential statistics in predicting anxiety.

Wrist-worn wearable devices offer a scalable and noninvasive method to passively and continuously monitor anxiety symptoms. Metrics collected from these devices have the potential to serve as a screening or detection tool for anxiety, allowing individuals at risk of developing clinical anxiety to be identified before consulting a clinician. If these metrics and models have high predictive accuracy, they could support clinicians in identifying at-risk individuals and provide timely assessments and interventions. Before such an approach can be implemented in real-world clinical practice, more research is essential. In order to develop a clinically meaningful prediction model, interdisciplinary collaboration is crucial (eg, clinicians or machine learning experts). This will ensure the models are accurate, interpretable, clinically meaningful, and integrated into the health care system rather than remaining a research-focused tool.

Conclusions

This review is the first to comprehensively synthesize evidence on the association between digital biomarkers of health derived from wrist-worn wearable devices and anxiety, integrating findings from both inferential statistical and machine learning studies. By systematically synthesizing the literature, this review identified gaps in evidence for future research.

XSL•FO
RenderX

Our meta-analyses found no associations between SOL, TST, SE, and WASO and anxiety. Although the number of studies was limited, there was evidence that lower physical activity levels and elevated heart rate were associated with more anxiety symptoms. Existing studies on light physical activity, MVPA, and HRV present inconsistent or inconclusive findings, highlighting the need for more research. Further, a comprehensive review of the literature yielded few machine learning studies in this area, and their quality was variable. While machine learning is increasingly used to analyze wearable data, its application for predicting anxiety remains in the early stages.

Machine learning and digital biomarkers could become powerful tools for early detection of anxiety, symptom monitoring and management, and prevention, particularly when used as part of broader assessment frameworks. Continuous data collection could support long-term monitoring of behavioral and physiological patterns relevant to anxiety, and may complement self-report and clinical assessments rather than being used as stand-alone screening tools. Given that anxiety is a potentially modifiable psychological factor associated with dementia risk, wearable-derived digital biomarkers may also offer future opportunities to support early identification and monitoring of dementia risk. However, further high-quality research is required before such approaches can be meaningfully translated into clinical or public health practice.

Funding

YL is a PhD student funded by the Economic & Social Research Council's London (UBEL) Doctoral Training Partnership (ES/P000592/1), embedded within the APPLE-TREE programme, which was funded by an Economic and Social Research Council/National Institute for Health Research programme grant (ES/S010408/1). The views expressed in this publication are those of the authors and not necessarily those of the National Institute for Health Research or the Department of Health and Social Care.

Data Availability

The dataset generated and analyzed during this study is available from the corresponding author on reasonable request.

Authors' Contributions

Protocol development: YL, HD-K, NM
Database search: YL
Screening, data extraction, and risk of bias assessment: YL, NC, HAG, RM
Analysis and interpretation of data: YL, HD-K, NLM
Drafting of manuscript: YL
Critical revision of the manuscript for important intellectual content: YL, NC, HAG, RM, CC, ZW, HD-K, NLM
Study supervision: ZW, HD-K, NLM

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA 2020 Checklist.
[DOCX File , 271 KB - jmir_v28i1e73812_app1.docx]

Multimedia Appendix 2
PRISMA 2020 for abstracts checklist.
[DOCX File , 264 KB - jmir_v28i1e73812_app2.docx]

Multimedia Appendix 3
PRISMA-S Checklist.
[DOCX File , 17 KB - jmir_v28i1e73812_app3.docx]

Multimedia Appendix 4
Overview of search terms.
[DOCX File , 18 KB - jmir_v28i1e73812_app4.docx]

Multimedia Appendix 5
Risk of bias studies.

XSL•FO
RenderX

[DOCX File , 23 KB - jmir_v28i1e73812_app5.docx]

Multimedia Appendix 6
Meta-analytic results for sleep metrics.
[DOCX File , 23 KB - jmir_v28i1e73812_app6.docx]

References

1. Global burden of disease Collaborative Network. Global Burden of Disease Study 2019. Seattle, United States: Institute for Health Metrics and Evaluation (IHME); 2020.
2. Diagnostic and Statistical Manual of Mental Disorders: DSM-5TM, 5th ed. Arlington, VA, US: American Psychiatric Publishing, Inc; 2013:xliv-x947.
3. Bryant C, Jackson H, Ames D. The prevalence of anxiety in older adults: methodological issues and a review of the literature. J Affect Disord 2008;109(3):233-250. [doi: 10.1016/j.jad.2007.11.008] [Medline: 18155775]
4. Gulpers B, Ramakers I, Hamel R, Köhler S, Voshaar RO, Verhey F. Anxiety as a predictor for cognitive decline and dementia: a systematic review and meta-analysis. Am J Geriatr Psychiatry 2016;24(10):823-842. [doi: 10.1016/j.jagp.2016.05.015] [Medline: 27591161]
5. Ford E, Greenslade N, Paudyal P, Bremner S, Smith HE, Banerjee S, et al. Predicting dementia from primary care records: a systematic review and meta-analysis. PLoS One 2018;13(3):e0194735 [FREE Full text] [doi: 10.1371/journal.pone.0194735] [Medline: 29596471]
6. Santabárbara J, Villagrasa B, López-Antón R, Olaya B, Bueno-Notivol J, de la Cámara C, et al. Clinically relevant anxiety and risk of Alzheimer's disease in an elderly community sample: 4.5 years of follow-up. J Affect Disord 2019;250:16-20. [doi: 10.1016/j.jad.2019.02.050] [Medline: 30825716]
7. Santabárbara J, Lipnicki DM, Olaya B, Villagrasa B, Bueno-Notivol J, Nuez L, et al. Does anxiety increase the risk of all-cause dementia? An updated meta-analysis of prospective cohort studies. J Clin Med 2020;9(6):1791 [FREE Full text] [doi: 10.3390/jcm9061791] [Medline: 32526871]
8. Kuring JK, Mathias JL, Ward L. Risk of dementia in persons who have previously experienced clinically-significant depression, anxiety, or PTSD: a systematic review and meta-analysis. J Affect Disord 2020;274:247-261. [doi: 10.1016/j.jad.2020.05.020] [Medline: 32469813]
9. GBD 2019 Dementia Forecasting Collaborators. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. Lancet Public Health 2022;7(2):e105-e125 [FREE Full text] [doi: 10.1016/S2468-2667(21)00249-8] [Medline: 34998485]
10. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc 2016;9:211-217 [FREE Full text] [doi: 10.2147/JMDH.S104807] [Medline: 27217764]
11. Balsamo M, Cataldi F, Carlucci L, Fairfield B. Assessment of anxiety in older adults: a review of self-report measures. Clin Interv Aging 2018;13:573-593 [FREE Full text] [doi: 10.2147/CIA.S114100] [Medline: 29670342]
12. Babrak LM, Menetski J, Rebhan M, Nisato G, Zinggeler M, Brasier N, et al. Traditional and digital biomarkers: two worlds apart? Digit Biomark 2019;3(2):92-102 [FREE Full text] [doi: 10.1159/000502000] [Medline: 32095769]
13. Nelson BW, Low CA, Jacobson N, Areán P, Torous J, Allen NB. Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research. NPJ Digit Med 2020;3(1):90 [FREE Full text] [doi: 10.1038/s41746-020-0297-4] [Medline: 32613085]
14. Ahmed A, Aziz S, Alzubaidi M, Schneider J, Irshaidat S, Abu Serhan H, et al. Comput Methods Programs Biomed Update 2023;3:100095 [FREE Full text] [doi: 10.1016/j.cmpbup.2023.100095] [Medline: 36743720]
15. McDowell CP, Dishman RK, Gordon BR, Herring MP. Physical activity and anxiety: a systematic review and meta-analysis of prospective cohort studies. Am J Prev Med 2019;57(4):545-556. [doi: 10.1016/j.amepre.2019.05.012] [Medline: 31542132]
16. Cox RC, Olatunji BO. A systematic review of sleep disturbance in anxiety and related disorders. J Anxiety Disord 2016;37:104-129. [doi: 10.1016/j.janxdis.2015.12.001] [Medline: 26745517]
17. Abd-Alrazaq A, AlSaad R, Shuweihdi F, Ahmed A, Aziz S, Sheikh J. Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression. NPJ Digit Med 2023;6(1):84 [FREE Full text] [doi: 10.1038/s41746-023-00828-5] [Medline: 37147384]
18. Abd-Alrazaq A, AlSaad R, Harfouche M, Aziz S, Ahmed A, Damseh R, et al. Wearable artificial intelligence for detecting anxiety: systematic review and meta-analysis. J Med Internet Res 2023;25(1):e48754 [FREE Full text] [doi: 10.2196/48754] [Medline: 37938883]
19. Jacobson NC, Bhattacharya S. Digital biomarkers of anxiety disorder symptom changes: personalized deep learning models using smartphone sensors accurately predict anxiety symptoms from ecological momentary assessments. Behav Res Ther 2022;149:104013 [FREE Full text] [doi: 10.1016/j.brat.2021.104013] [Medline: 35030442]
20. Moshe I, Terhorst Y, Opoku Asare K, Sander LB, Ferreira D, Baumeister H, et al. Predicting symptoms of depression and anxiety using smartphone and wearable data. Front Psychiatry 2021;12:625247 [FREE Full text] [doi: 10.3389/fpsyt.2021.625247] [Medline: 33584388]

XSL•FO
RenderX

21. Moher D, Liberati A, Tetzlaff J, Altman DG, Group TP. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
22. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
23. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
24. Haddaway NR, Collins AM, Coughlin D, Kirk S. The role of Google Scholar in evidence reviews and its applicability to grey literature searching. PLoS One 2015;10(9):e0138237 [FREE Full text] [doi: 10.1371/journal.pone.0138237] [Medline: 26379270]
25. Study quality assessment tools. NHLBI, NIH. URL: https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools [accessed 2023-11-18]
26. Berner K, Morris L, Baumeister J, Louw Q. Objective impairments of gait and balance in adults living with HIV-1 infection: a systematic review and meta-analysis of observational studies. BMC Musculoskelet Disord 2017;18(1):325 [FREE Full text] [doi: 10.1186/s12891-017-1682-2] [Medline: 28764704]
27. Jayakumar S, Sounderajah V, Normahani P, Harling L, Markar SR, Ashrafian H, et al. Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study. NPJ Digit Med 2022;5(1):11 [FREE Full text] [doi: 10.1038/s41746-021-00544-y] [Medline: 35087178]
28. A basic introduction to fixed-effect and random-effects models for meta-analysis - Borenstein - 2010 - research synthesis methods -. Wiley Online Library. URL: https://onlinelibrary.wiley.com/doi/10.1002/jrsm.12 [accessed 2024-01-05]
29. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014;14:25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
30. Korevaar E, Turner SL, Forbes AB, Karahalios A, Taljaard M, McKenzie JE. Evaluation of statistical methods used to meta-analyse results from interrupted time series studies: a simulation study. Res Synth Methods 2023;14(6):882-902 [FREE Full text] [doi: 10.1002/jrsm.1669] [Medline: 37731166]
31. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
32. Endler NS, Kocovski NL. State and trait anxiety revisited. J Anxiety Disord 2001;15(3):231-245. [doi: 10.1016/s0887-6185(01)00060-3] [Medline: 11442141]
33. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21(11):1539-1558. [doi: 10.1002/sim.1186] [Medline: 12111919]
34. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
35. Ahn J, Kim J, Park Y, Kim R, Choi H. Nature-based virtual reality relaxation to improve mental health and sleep in undergraduate students: a randomized controlled trial. Digit Health 2025;11:20552076251365140 [FREE Full text] [doi: 10.1177/20552076251365140] [Medline: 40799837]
36. Bullis J. Temporal patterns of sleep disturbance, anxiety, and depressed mood in generalized anxiety disorder. OpenBU. 2016. URL: https://open.bu.edu/handle/2144/19590 [accessed 2023-08-15]
37. Casanova F, O'Loughlin J, Karageorgiou V, Beaumont RN, Bowden J, Wood AR, et al. Effects of physical activity and sedentary time on depression, anxiety and well-being: a bidirectional Mendelian randomisation study. BMC Med 2023;21(1):501 [FREE Full text] [doi: 10.1186/s12916-023-03211-z] [Medline: 38110912]
38. Cox RC, Sterba SK, Cole DA, Upender RP, Olatunji BO. Time of day effects on the relationship between daily sleep and anxiety: an ecological momentary assessment approach. Behav Res Ther 2018;111:44-51 [FREE Full text] [doi: 10.1016/j.brat.2018.09.008] [Medline: 30300779]
39. D'Aurizio G, Tosti B, Tempesta D, Avvantaggiato L, Splendiani A, Sacco S, et al. Reduced sleep amount and increased sleep latency in prisoners: a pilot study in an Italian jail. Brain Sci 2023;13(1):132 [FREE Full text] [doi: 10.3390/brainsci13010132] [Medline: 36672113]
40. Dillon CB, McMahon E, O'Regan G, Perry IJ. Associations between physical behaviour patterns and levels of depressive symptoms, anxiety and well-being in middle-aged adults: a cross-sectional study using isotemporal substitution models. BMJ Open 2018;8(1):e018978 [FREE Full text] [doi: 10.1136/bmjopen-2017-018978] [Medline: 29358436]
41. Doane LD, Gress-Smith JL, Breitenstein RS. Multi-method assessments of sleep over the transition to college and the associations with depression and anxiety symptoms. J Youth Adolesc 2015;44(2):389-404. [doi: 10.1007/s10964-014-0150-7] [Medline: 25034248]
42. El-Sheikh M, Kelly R, Rauer A. Quick to berate, slow to sleep: interpartner psychological conflict, mental health, and sleep. Health Psychol 2013;32(10):1057-1066 [FREE Full text] [doi: 10.1037/a0031786] [Medline: 23544995]

XSL•FO
RenderX

43. Facer-Childs ER, Mascaro L, Hoffman D, Mansfield D, Drummond SPA, Rajaratnam SMW. Sleep as a major determinant for mental health outcomes in elite Australian Football League (AFL) athletes. Med Sci Sports Exerc 2022;54(4):665-672. [doi: 10.1249/MSS.0000000000002825] [Medline: 34772903]
44. Feng S, Yi JS, Deitz G, Ding L, Van Gelder RN, Menda S. Relationships between sleep, activity, and burnout in ophthalmology residents. J Surg Educ 2021;78(3):1035-1040. [doi: 10.1016/j.jsurg.2020.09.003] [Medline: 32967802]
45. Fuller-Rowell TE, Zeringue MM, Saini EK, Yip T, El-Sheikh M. Do sleep problems exacerbate the mental health consequences of discrimination among adults? Psychosom Med 2024;86(4):324-333. [doi: 10.1097/psy.0000000000001305]
46. Hofman A, Voortman T, Ikram MA, Luik AI. Substitutions of physical activity, sedentary behaviour and sleep: associations with mental health in middle-aged and elderly persons. J Epidemiol Community Health 2022;76(2):175-181 [FREE Full text] [doi: 10.1136/jech-2020-215883] [Medline: 34301796]
47. Huang T, Zheng K, Li S, Yang Y, Kong L, Zhao Y. Screen-based sedentary behaviors but not total sedentary time are associated with anxiety among college students. Front Public Health 2022;10:994612 [FREE Full text] [doi: 10.3389/fpubh.2022.994612] [Medline: 36339232]
48. Lee S, Bonnar D, Roane B, Gradisar M, Dunican IC, Lastella M, et al. Sleep characteristics and mood of professional esports athletes: a multi-national study. Int J Environ Res Public Health 2021;18(2):664 [FREE Full text] [doi: 10.3390/ijerph18020664] [Medline: 33466788]
49. Montanari A, Fancello G, Sueur C, Kestens Y, van Lenthe FJ, Chaix B. A sensor-based study on the environmental determinants of sleep in older adults. Environ Res 2025;274:120874. [doi: 10.1016/j.envres.2025.120874] [Medline: 39855412]
50. Nguyen E, Meadley B, Harris R, Rajaratnam SMW, Williams B, Smith K, et al. Sleep and mental health in recruit paramedics: a 6-month longitudinal study. Sleep 2023;46(8):zsad050 [FREE Full text] [doi: 10.1093/sleep/zsad050] [Medline: 36861384]
51. Spira AP, Friedman L, Aulakh JS, Lee T, Sheikh JI, Yesavage JA. Subclinical anxiety symptoms, sleep, and daytime dysfunction in older adults with primary insomnia. J Geriatr Psychiatry Neurol 2008;21(2):149-153. [doi: 10.1177/0891988707317120] [Medline: 18474724]
52. Spira AP, Stone K, Beaudreau SA, Ancoli-Israel S, Yaffe K. Anxiety symptoms and objectively measured sleep quality in older women. Am J Geriatr Psychiatry 2009;17(2):136-143 [FREE Full text] [doi: 10.1097/JGP.0b013e3181871345] [Medline: 19155746]
53. Stremler R, Haddad S, Pullenayegum E, Parshuram C. Psychological outcomes in parents of critically ill hospitalized children. J Pediatr Nurs 2017;34:36-43. [doi: 10.1016/j.pedn.2017.01.012] [Medline: 28274664]
54. Swanson LM, Hood MM, Hall MH, Avis NE, Joffe H, Colvin A, et al. Sleep timing, sleep regularity, and psychological health in early late life women: findings from the Study of Women's Health Across the Nation (SWAN). Sleep Health 2023;9(2):203-210 [FREE Full text] [doi: 10.1016/j.sleh.2022.11.001] [Medline: 36509657]
55. Uchida Y, Kanamori M, Fukushima S, Takemura K. Interdependent culture and older adults' well-being: health and psychological happiness in Japanese communities. Curr Opin Psychol 2024;55:101729 [FREE Full text] [doi: 10.1016/j.copsyc.2023.101729] [Medline: 38096782]
56. Windmill H, Dyer A, Smith AD, Roser ME, Dhanda A, Hall SD. Characterising the inconsistency between perceived and actual sleep and its impact upon cognition and mood. Sci Rep 2024;14(1):29342 [FREE Full text] [doi: 10.1038/s41598-024-80683-w] [Medline: 39592657]
57. Walters EM, Phillips AJ, Hamill K, Norton PJ, Drummond SP. Anxiety predicts dyadic sleep characteristics in couples experiencing insomnia but not in couples without sleep disorders. J Affect Disord 2020;273:122-130. [doi: 10.1016/j.jad.2020.04.031] [Medline: 32421592]
58. Zhang Y, Stewart C, Ranjan Y, Conde P, Sankesara H, Rashid Z, et al. Large-scale digital phenotyping: identifying depression and anxiety indicators in a general UK population with over 10,000 participants. J Affect Disord 2025;375:412-422 [FREE Full text] [doi: 10.1016/j.jad.2025.01.124] [Medline: 39892753]
59. Zheng NS, Annis J, Master H, Han L, Gleichauf K, Ching JH, et al. Sleep patterns and risk of chronic disease as measured by long-term monitoring with commercial wearable devices in the All of Us Research Program. Nat Med 2024;30(9):2648-2656. [doi: 10.1038/s41591-024-03155-8] [Medline: 39030265]
60. Bai Y, Burns R, Ma C, Curl A, Hudziak J, Copeland WE. Tracking well-being: a comprehensive analysis of physical activity and mental health in college students across COVID-19 phases using ecological momentary assessment. Scand J Med Sci Sports 2024;34(10):e14738. [doi: 10.1111/sms.14738] [Medline: 39410823]
61. Jo YT, Lee SW, Park S, Lee J. Association between heart rate variability metrics from a smartwatch and self-reported depression and anxiety symptoms: a four-week longitudinal study. Front Psychiatry 2024;15:1371946 [FREE Full text] [doi: 10.3389/fpsyt.2024.1371946] [Medline: 38881544]
62. Kandola AA, Del Pozo Cruz B, Osborn DPJ, Stubbs B, Choi KW, Hayes JF. Impact of replacing sedentary behaviour with other movement behaviours on depression and anxiety symptoms: a prospective cohort study in the UK Biobank. BMC Med 2021;19(1):133 [FREE Full text] [doi: 10.1186/s12916-021-02007-3] [Medline: 34134689]
63. Liu Y, Huang C, Wei K, Lin R, Yao Y, Qin G, et al. Accelerometer-derived "weekend warrior" physical activity, genetic susceptibility, and risk of depression and anxiety: the UK Biobank study. J Affect Disord 2025;380:55-62. [doi: 10.1016/j.jad.2025.03.110] [Medline: 40122250]

XSL•FO
RenderX

64. McDuff D, Galatzer-Levy I, Thomson S, Barakat A, Heneghan C, Abdel-Ghaffar S, et al. Evidence of differences in diurnal electrodermal, temperature and heart rate patterns by mental health status in free-living data. BMJ Ment Health 2025;28(1):e301307 [FREE Full text] [doi: 10.1136/bmjment-2024-301307] [Medline: 40819881]
65. Okawara H, Shiraishi Y, Sato K, Nakamura M, Katsumata Y. Visually assessing work performance using a smartwatch via day-to-day fluctuations in heart rate variability. Digit Health 2024;10:20552076241239240 [FREE Full text] [doi: 10.1177/20552076241239240] [Medline: 38533307]
66. Presby D, Jasinski S, Capodilupo E, Holmes KE, von Hippel W, Grosicki GJ, et al. Inter- and intrapersonal associations between physiology and mental health: a longitudinal study using wearables and mental health surveys. J Med Internet Res 2025;27:e64955 [FREE Full text] [doi: 10.2196/64955] [Medline: 40700646]
67. Straus LD, An X, Ji Y, McLean SA, Neylan TC. AURORA Study Group. Utility of wrist-wearable data for assessing pain, sleep, and anxiety outcomes after traumatic stress exposure. JAMA Psychiatry 2023;80(3):220-229 [FREE Full text] [doi: 10.1001/jamapsychiatry.2022.4533] [Medline: 36630119]
68. Wang Z, Cao Z, Min J, Duan T, Xu C. Associations between device-measured and self-reported physical activity and common mental disorders: findings from a large-scale prospective cohort study. BMJ Evid Based Med 2025;30(1):45-54. [doi: 10.1136/bmjebm-2024-112933] [Medline: 38997150]
69. Yu K, Yang Q, Wang J, Zeng B. Accelerometer-derived physical activity, sedentary behavior, and the risk of depression and anxiety in middle-aged and older adults: a prospective cohort study of 71,556 UK Biobank participants. Am J Geriatr Psychiatry 2025;33(5):512-523. [doi: 10.1016/j.jagp.2024.10.015] [Medline: 39532554]
70. Coutts LV, Plans D, Brown AW, Collomosse J. Deep learning with wearable based heart rate variability for prediction of mental and general health. J Biomed Inform 2020;112:103610 [FREE Full text] [doi: 10.1016/j.jbi.2020.103610] [Medline: 33137470]
71. Fukuda S, Matsuda Y, Tani Y, Arakawa Y, Yasumoto K. Predicting depression and anxiety mood by wrist-worn sleep sensor. 2020 Presented at: 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops); 2020 March 23-27; Austin, TX, USA p. 1-6. [doi: 10.1109/percomworkshops48775.2020.9156176]
72. He M, Cerna J, Mathew R, Zhao J, Zhao J, Espina E, et al. Objective anxiety level classification using unsupervised learning and multimodal physiological signals. Smart Health 2025;36:100572. [doi: 10.1016/j.smhl.2025.100572]
73. Lee TR, Kim GH, Choi MT. Identification of geriatric depression and anxiety using activity tracking data and minimal geriatric assessment scales. Appl Sci 2022;12(5):2488. [doi: 10.3390/app12052488]
74. Lee TR, Kim GH, Choi MT. Geriatric depression and anxiety screening via deep learning using activity tracking and sleep data. Int J Geriatr Psychiatry 2024;39(2):e6071. [doi: 10.1002/gps.6071] [Medline: 38372966]
75. Saylam B, İncel Ö. Quantifying digital biomarkers for well-being: stress, anxiety, positive and negative affect via wearable devices and their time-based predictions. Sensors (Basel) 2023;23(21) [FREE Full text] [doi: 10.3390/s23218987] [Medline: 37960685]
76. Saylam B, İncel Ö. Multitask learning for mental health: depression, anxiety, stress (DAS) using wearables. Diagnostics (Basel) 2024;14(5):501 [FREE Full text] [doi: 10.3390/diagnostics14050501] [Medline: 38472973]
77. Ahn S, Lobo JM, Davis EM, Howie-Esquivel J, Chung ML, Logan JG. Characterization of sleep efficiency transitions in family caregivers. J Behav Med 2024;47(2):308-319. [doi: 10.1007/s10865-023-00461-3] [Medline: 38017251]
78. Spineli LM, Pandis N. Problems and pitfalls in subgroup analysis and meta-regression. Am J Orthod Dentofacial Orthop 2020;158(6):901-904. [doi: 10.1016/j.ajodo.2020.09.001] [Medline: 33250104]
79. Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 1989;28(2):193-213. [Medline: 2748771]
80. van den Berg JF, Luijendijk HJ, Tulen JHM, Hofman A, Neven AK, Tiemeier H. Sleep in depression and anxiety disorders: a population-based study of elderly persons. J Clin Psychiatry 2009;70(8):1105-1113. [doi: 10.4088/JCP.08m04448] [Medline: 19607762]
81. van Mill JG, Hoogendijk WJG, Vogelzangs N, van Dyck R, Penninx BWJH. Insomnia and sleep duration in a large cohort of patients with major depressive disorder and anxiety disorders. J Clin Psychiatry 2010;71(3):239-246. [doi: 10.4088/JCP.09m05218gry] [Medline: 20331928]
82. Potvin O, Lorrain D, Belleville G, Grenier S, Préville M. Subjective sleep characteristics associated with anxiety and depression in older adults: a population-based study. Int J Geriatr Psychiatry 2014;29(12):1262-1270. [doi: 10.1002/gps.4106] [Medline: 24733621]
83. Ramsawh HJ, Stein MB, Belik SL, Jacobi F, Sareen J. Relationship of anxiety disorders, sleep quality, and functional impairment in a community sample. J Psychiatr Res 2009;43(10):926-933. [doi: 10.1016/j.jpsychires.2009.01.009] [Medline: 19269650]
84. Zhu Y, Meng R, Jiang C, Yang N, Huang M, Wang X, et al. Sleep quality and subjective well-being in healthcare students: examining the role of anxiety and depression. Front Public Health 2023;11:1281571 [FREE Full text] [doi: 10.3389/fpubh.2023.1281571] [Medline: 38213643]
85. Gould CE, Spira AP, Liou-Johnson V, Cassidy-Eagle E, Kawai M, Mashal N, et al. Association of anxiety symptom clusters with sleep quality and daytime sleepiness. J Gerontol B Psychol Sci Soc Sci 2018;73(3):413-420 [FREE Full text] [doi: 10.1093/geronb/gbx020] [Medline: 28379498]

XSL•FO
RenderX

86. Monti JM, Monti D. Sleep disturbance in generalized anxiety disorder and its treatment. Sleep Med Rev 2000;4(3):263-276. [doi: 10.1053/smrv.1999.0096] [Medline: 12531169]
87. Doane LD, Thurston EC. Associations among sleep, daily experiences, and loneliness in adolescence: evidence of moderating and bidirectional pathways. J Adolesc 2014;37(2):145-154. [doi: 10.1016/j.adolescence.2013.11.009] [Medline: 24439620]
88. Fekedulegn D, Andrew ME, Shi M, Violanti JM, Knox S, Innes KE. Actigraphy-based assessment of sleep parameters. Ann Work Expo Health 2020;64(4):350-367 [FREE Full text] [doi: 10.1093/annweh/wxaa007] [Medline: 32053169]
89. Wanjau MN, Möller H, Haigh F, Milat A, Hayek R, Lucas P, et al. Physical activity and depression and anxiety disorders: a systematic review of reviews and assessment of causality. AJPM Focus 2023;2(2):100074 [FREE Full text] [doi: 10.1016/j.focus.2023.100074] [Medline: 37790636]
90. Singh B, Olds T, Curtis R, Dumuid D, Virgara R, Watson A, et al. Effectiveness of physical activity interventions for improving depression, anxiety and distress: an overview of systematic reviews. Br J Sports Med 2023;57(18):1203-1209 [FREE Full text] [doi: 10.1136/bjsports-2022-106195] [Medline: 36796860]
91. Teychenne M, Costigan SA, Parker K. The association between sedentary behaviour and risk of anxiety: a systematic review. BMC Public Health 2015;15(1):513 [FREE Full text] [doi: 10.1186/s12889-015-1843-x] [Medline: 26088005]
92. Allen MS, Walter EE, Swann C. Sedentary behaviour and risk of anxiety: a systematic review and meta-analysis. J Affect Disord 2019;242:5-13. [doi: 10.1016/j.jad.2018.08.081] [Medline: 30170238]
93. Lin Y, Gao W. The effects of physical exercise on anxiety symptoms of college students: a meta-analysis. Front Psychol 2023;14:1136900 [FREE Full text] [doi: 10.3389/fpsyg.2023.1136900] [Medline: 37063553]
94. Chalmers JA, Quintana DS, Abbott MJA, Kemp AH. Anxiety disorders are associated with reduced heart rate variability: a meta-analysis. Front Psychiatry 2014;5:80 [FREE Full text] [doi: 10.3389/fpsyt.2014.00080] [Medline: 25071612]
95. Kishi T. Heart rate is the clinical indicator of sympathetic activation and prognostic value of cardiovascular risks in patients with hypertension. Hypertension 2020;76(2):323-324. [doi: 10.1161/hypertensionaha.120.14898]
96. Kemp AH, Brunoni AR, Santos IS, Nunes MA, Dantas EM, Carvalho de Figueiredo R, et al. Effects of depression, anxiety, comorbidity, and antidepressants on resting-state heart rate and its variability: an ELSA-Brasil cohort baseline study. Am J Psychiatry 2014;171(12):1328-1334. [doi: 10.1176/appi.ajp.2014.13121605] [Medline: 25158141]
97. Latvala A, Kuja-Halkola R, Rück C, D'Onofrio BM, Jernberg T, Almqvist C, et al. Association of resting heart rate and blood pressure in late adolescence with subsequent mental disorders: a longitudinal population study of more than 1 Million men in Sweden. JAMA Psychiatry 2016;73(12):1268-1275. [doi: 10.1001/jamapsychiatry.2016.2717] [Medline: 27784035]
98. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nat Methods 2018;15(4):233-234 [FREE Full text] [doi: 10.1038/nmeth.4642] [Medline: 30100822]
99. Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, et al. Interpreting black-box models: a review on explainable artificial intelligence. Cogn Comput 2024;16(1):45-74. [doi: 10.1007/s12559-023-10179-8]
100. Lehrer HM, Yao Z, Krafty RT, Evans MA, Buysse DJ, Kravitz HM, et al. Comparing polysomnography, actigraphy, and sleep diary in the home environment: the Study of Women's Health Across the Nation (SWAN) sleep study. Sleep Adv 2022;3(1):zpac001 [FREE Full text] [doi: 10.1093/sleepadvances/zpac001] [Medline: 35296109]
101. Blackwell T, Yaffe K, Ancoli-Israel S, Schneider JL, Cauley JA, Hillier TA, Study of Osteoporotic Fractures Group. Poor sleep is associated with impaired cognitive function in older women: the study of osteoporotic fractures. J Gerontol A Biol Sci Med Sci 2006;61(4):405-410. [doi: 10.1093/gerona/61.4.405] [Medline: 16611709]
102. Chinoy ED, Cuellar JA, Huwa KE, Jameson JT, Watson CH, Bessman SC, et al. Performance of seven consumer sleep-tracking devices compared with polysomnography. Sleep 2021;44(5):zsaa291 [FREE Full text] [doi: 10.1093/sleep/zsaa291] [Medline: 33378539]
103. Marino M, Li Y, Rueschman MN, Winkelman JW, Ellenbogen JM, Solet JM, et al. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. Sleep 2013;36(11):1747-1755 [FREE Full text] [doi: 10.5665/sleep.3142] [Medline: 24179309]
104. De Mello MT, Lemos VDA, Antunes HKM, Bittencourt L, Santos-Silva R, Tufik S. Relationship between physical activity and depression and anxiety symptoms: a population study. J Affect Disord 2013;149(1-3):241-246. [doi: 10.1016/j.jad.2013.01.035] [Medline: 23489405]
105. Azevedo Da Silva M, Singh-Manoux A, Brunner EJ, Kaffashian S, Shipley MJ, Kivimäki M, et al. Bidirectional association between physical activity and symptoms of anxiety and depression: the Whitehall II study. Eur J Epidemiol 2012;27(7):537-546 [FREE Full text] [doi: 10.1007/s10654-012-9692-8] [Medline: 22623145]
106. Prince SA, Adamo KB, Hamel ME, Hardt J, Connor Gorber S, Tremblay M. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. Int J Behav Nutr Phys Act 2008;5:56 [FREE Full text] [doi: 10.1186/1479-5868-5-56] [Medline: 18990237]
107. McLean CP, Asnaani A, Litz BT, Hofmann SG. Gender differences in anxiety disorders: prevalence, course of illness, comorbidity and burden of illness. J Psychiatr Res 2011;45(8):1027-1035 [FREE Full text] [doi: 10.1016/j.jpsychires.2011.03.006] [Medline: 21439576]

XSL•FO
RenderX

Abbreviations

GAD: generalized anxiety disorder
HRV: heart rate variability
MAE: mean absolute error
MVPA: moderate to vigorous physical activity
PECOS: Population, Exposure, Control, Outcome, and Study design
PNS: parasympathetic nervous system
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search Extension
PROSPERO: International Prospective Register of Systematic Reviews
REM: rapid eye movement
SE: sleep efficiency
SNS: sympathetic nervous system
SOL: sleep onset latency
TST: total sleep time
WASO: wake after sleep onset

Edited by S Brini; submitted 12.Mar.2025; peer-reviewed by EDA Bacil, R Chigevenga; comments to author 12.Sep.2025; revised version received 06.Feb.2026; accepted 09.Feb.2026; published 09.Mar.2026.

Please cite as:
Lau Y, Chemas N, Ajeet Gokani H, Morrell R, Phannarus H, Cooper C, Walker Z, Demnitz-King H, Marchant NL
Association Between Digital Biomarkers of Health and Anxiety: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e73812
URL: https://www.jmir.org/2026/1/e73812
doi:10.2196/73812
PMID:

©Yolanda Lau, Natalia Chemas, Heema Ajeet Gokani, Rachel Morrell, Harisd Phannarus, Claudia Cooper, Zuzana Walker, Harriet Demnitz-King, Natalie L Marchant. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effects of Digital Health Interventions on Functional and Psychological Outcomes in Older Patients With Hip Fractures: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Wei Fan[1], MD; Qi Zhang[2*], PhD; Qunfeng Lu[2*], PhD

[1]School of Medicine, Tianjin Tianshi College, Tianjin, China

[2]Department of Nursing, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Xuhui District, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Qunfeng Lu, PhD
Department of Nursing
Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine
No.600 Yishan Road
Xuhui District, Shanghai, 200233
China
Phone: 86 18917128006
Email: luluroom2004@163.com

Abstract

Background: Hip fractures in older adults increasingly challenge public health, making traditional rehabilitation very challenging. Digital health interventions (DHIs) have emerged as a promising solution for postoperative rehabilitation. However, evidence on DHIs' effects on functional and psychological outcomes remains insufficient.

Objective: This systematic review aimed to comprehensively examine the effects of DHIs on functional and psychological outcomes in older adults with hip fractures.

Methods: Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, we searched 9 databases (PubMed, Embase, CENTRAL, APA PsycINFO, Web of Science, PEDro, CNKI, WANFANG, and SinoMed) from inception to November 13, 2025. Included studies enrolled adults aged 60 years and older with hip fractures, delivered DHIs, assessed functional and psychological outcomes, set usual care or no intervention as the control, and had a randomized controlled trial design. Studies were excluded if they enrolled nonhospitalized patients in the emergency department, patients discharged to nonhome settings, or had inaccessible full text or insufficient data. Study quality was evaluated using the Cochrane Risk of Bias tool 2.0 (Cochrane Collaboration), and evidence certainty was assessed using GRADE (Grading of Recommendations, Assessment, Development and Evaluation). The literature screening, data extraction, and quality assessment were independently conducted by 2 researchers, and any disputes were resolved by the third researcher. We performed analysis using R version 4.0.3 (R Foundation for Statistical Computing) with a random-effects model.

Results: Of 17,723 studies screened, 13 met the inclusion criteria. DHIs, compared to the control, significantly improved hip function (standardized mean difference [SMD] 0.80, 95% CI 0.33-1.26; 95% prediction interval [PI] –0.24 to 1.83; *P*=.007) and functional independence (SMD 1.23, 95% CI 0.34-2.11; 95% PI –0.98 to 3.34; *P*=.02). Despite favorable pooled effects, a wide 95% PI spanning positive or negative values signals substantial heterogeneity. No significant difference was observed in balance function, risk of falling, and quality of life. Only a single available study reported a 70% adherence rate in the DHIs group. Subgroup analyses stratified by intervention duration revealed no significant intersubgroup differences for hip function (χ_1^2=0.1; *P*=.75) or functional independence (χ_1^2=2.93; *P*=.09). For hip function, the point estimate favored the 3 months subgroup (SMD 0.89, 95% CI 0.36-1.41; I^2=7%; *P*=.41) over the <3 months subgroup. Conversely, for functional independence, the point estimate favored shorter intervention duration (SMD 0.67, 95% CI 0.12-1.23; I^2=0%; *P*=.72).

XSL•FO
RenderX

Conclusions: This review incorporates the latest randomized controlled trials and comprehensively assesses functional and psychological outcomes of DHIs in older patients with hip fractures, distinct from prior studies focusing solely on functional outcomes. While the 95% CI supports the potential of DHIs to improve hip function and functional independence, the wide 95% PI indicating substantial real-world response variability, which calls for cautious interpretation, informs the design of targeted DHI-based rehabilitation regimens, warranting further research into optimal techniques and dosages in clinical practice.

Trial Registration: PROSPERO CRD42024626186; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024626186

(J Med Internet Res 2026;28:e79563) doi:10.2196/79563

KEYWORDS

digital health; older patients; hip fractures; systematic review; meta-analysis

Introduction

Hip fractures, particularly prevalent in older adults, present a significant and escalating public health challenge, characterized by well-documented morbidity, mortality, and health care burden [1-3]. By 2050, the annual global incidence of hip fractures is expected to reach 4.5 million [4], exerting ongoing pressure on existing health care and welfare systems [2,5]. With the general increase in human life expectancy, the systemic medical challenges posed by this condition are expected to intensify further. As per existing literature, the 30-day mortality rate after hip fractures can reach up to 10%, escalating further to 25% within the first year [6], while most surviving patients experience a significant decline in quality of life (QoL) [7]. Effective postsurgical rehabilitation is essential to mitigating the adverse outcomes of older adults with hip fractures, such as functional decline [8,9], psychological distress [10,11], and reduced QoL [9,12]. The majority of rehabilitation must now take place outside of hospitals due to shorter hospital stays brought on by a lack of health care resources and the need to increase hospital bed turnover rates [13].

Traditional rehabilitation programs face significant challenges in older adults, including limited access to health care services, low intervention adherence to prescribed rehabilitation programs, and insufficient psychological support. Adherence to rehabilitation programs remains a major issue, with many patients failing to complete recommended programs, thereby compromising functional recovery and increasing the risk of complications [14]. In addition, older adults may face specific barriers, such as limited mobility, geographic constraints, and resource shortages, which further reduce the feasibility and success of traditional rehabilitation methods [15-19]. Digital health interventions (DHIs) have emerged as a promising solution. DHIs, defined as the application of digital technologies in health care, leverage both conventional and innovative forms of information and communication technologies, such as mobile apps, telemedicine, wearable devices, and online platforms, to enhance health care delivery and patient engagement [20].

By bridging gaps in distance and accessibility, DHIs offer innovative solutions to overcome traditional health care barriers, particularly for populations with limited mobility, such as older patients with hip fractures [21]. Moreover, as digital technologies continue to advance, DHIs are evolving in both scope and sophistication, capable of providing more personalized, scalable, and effective approaches for the care of most patients whose rehabilitation training needs to be carried out outside the hospital [20,22]. DHIs have been used in the rehabilitation of older patients with hip fractures, and their potential effects have also been preliminarily demonstrated [13,21,23-26]. A study carried out by Gilboa et al [27] investigating a home-based tele-rehabilitation program, where occupational therapists remotely guided patients through personalized goal-setting and daily activity planning, demonstrated significant improvements in activities of daily living, health-related QoL, and mobility, while also reducing caregiver burden. In contrast to studies supporting DHIs, some researchers have argued that there are no or only minor effects detected in functional and psychological outcomes [28,29].

While the above findings highlight the potential impact of DHIs on functional recovery and psychological well-being in older adults with hip fractures, the evidence remains limited. Additionally, most studies focus primarily on functional recovery without adequately addressing psychological well-being and intervention adherence. Considering that the rehabilitation of older patients with hip fractures is an overall concept, it not only involves the recovery of function but also the restoration of psychological outcomes [30]. At the same time, studying the influence of intervention on intervention adherence is essential for the development of successful interventions in the rehabilitation of older patients with hip fractures [31]. The current research gaps underscore the need for a more comprehensive evaluation of DHIs in the context of the rehabilitation of hip fractures. With ongoing advancements in DHIs, a plethora of original studies focused on functional and psychological outcomes in older adults with hip fractures have been recently published. This makes it possible to review DHIs' applications and guide future design and implementation of DHIs technology in clinical practice.

Thus, our systematic review aimed to comprehensively examine the effects of DHIs on functional and psychological outcomes, as well as intervention adherence in older adults with hip fractures. Compared with previous systematic reviews, this systematic review incorporates more recent research, includes a greater number of studies addressing psychological outcomes, and assesses the quality of the evidence [21].

Methods

Overview

This systematic review and meta-analysis was reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement

XSL•FO
RenderX

(Multimedia Appendix 1) and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) to strengthen transparency (Multimedia Appendix 2) [32,33]. The review protocol was registered on the PROSPERO International Prospective Register of Systematic Reviews (CRD42024626186).

Search Strategy

This systematic review was systematically searched in 9 scientific databases, including PubMed, Embase, CENTRAL, APA PsycINFO, Web of Science, PEDro, CNKI, WANFANG, and SinoMed. The literature search was updated by re-executing the predefined search string in the same 9 databases on November 13, 2025, and newly retrieved records were screened following the identical inclusion and exclusion criteria. Search strategies were originally designed by our research team and underwent a peer review process conducted by an independent medical doctor with extensive evidence-based medicine expertise, who verified the comprehensiveness of search terms, database selection, and eligibility criteria prior to the initiation of retrieval. No substantive modifications to the strategy were required following this review. No published search filters were used in our review. No search strategies from previously published literature reviews were adapted or reused for any part of our literature retrieval process.

Reference lists of eligible studies and cited studies (identified via Google Scholar) were manually screened to retrieve additional relevant literature. The ClinicalTrials.gov (United States Clinical Trials Registry) and the Chinese Clinical Trial Registry were searched for eligible studies.

Search engines and citation chasers were also screened to capture all relevant studies. For studies where the full text could not be obtained, corresponding authors were contacted to attempt to acquire the full text. Considering the availability of academic resources and the language proficiency of the research team, the included studies were limited to those published in English and Chinese with no date restriction. The full search strategies are available in Multimedia Appendix 3.

Eligibility Criteria

We included the studies that related to (1) older adults (including patients aged 60 years or older) with hip fractures defined as any type of hip fractures, such as femoral neck, intertrochanteric, subtrochanteric fractures, and no specified type of hip fractures; (2) the intervention consisted of DHIs according to classifications of the World Health Organization (WHO) and must be provided or supported through DHIs for patients rather than trials targeting health care providers or other stakeholders [34]; (3) the primary outcome in this systematic review including functional outcomes (eg, hip function, walking capacity, and balance ability) and secondary outcomes including psychological outcomes (eg, depression and QoL) and intervention adherence; (5) the comparator was usual care, control group or no intervention; and (6) the study design limited to randomized controlled trials (RCTs). The following studies were excluded: (1) patients enrolled in the emergency department but not admitted to the hospital, (2) patients discharged from the hospital to a setting other than home, (3) studies involving DHIs that did not directly intervene with patients (including database management), and (4) studies with inaccessible full text or insufficient data for evaluating the characteristics of DHIs or quantitative synthesis.

Study Selection and Data Extraction

After automatic and manual removal of duplicate records using the reference management software EndNote 21 (Clarivate Analytics), 2 review authors (WF and QZ) independently screened titles and abstracts and then full texts against predefined eligibility criteria. A senior author (QL) resolved conflicts.

Two authors (WF and QZ) independently but in parallel collected relevant data in a standardized data collection sheet, including but not limited to characteristics of the included studies, such as the authors, publication year, country, and study design, as well as participant characteristics, including sample size, sex ratio, average age, and the details of the intervention and control groups.

For studies with multiple intervention arms, data from the relevant arms were combined to create a single pairwise comparison, as recommended in the Cochrane Handbook, to avoid double-counting participants in the control group. For studies reporting multiple time points for the same outcome, data at the end of the intervention were extracted for meta-analysis. Any discrepancies between groups were resolved through discussion and consensus, and when necessary, a senior reviewer (QL) was consulted to adjudicate unresolved differences.

Risk-of-Bias and Grading of Recommendations Assessment, Development, and Evaluation

Risk of bias in each study was assessed independently by 2 authors (WF and QZ). The RCTs were evaluated using the Cochrane Risk of Bias tool 2.0 (Cochrane Collaboration), which assesses random sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, selective reporting, and other bias [35]. The reviewers made a judgment of “yes” (low bias), “no” (high bias), or “unclear” (lack of relevant information or uncertainty of bias) for each item. Included studies were also subjected to scoring for certainty of evidence using GRADE (Grading of Recommendations, Assessment, Development and Evaluation) criteria independently by 2 authors (WF and QZ). Any discrepancies between groups were resolved through discussion and consensus, and when necessary, a third reviewer (QL) was consulted to adjudicate unresolved differences.

Meta-Analysis Methods

R (version 4.0.3; R Foundation for Statistical Computing) was used to conduct the meta-analyses. The data analysis was based on postintervention means and SDs, with conversion performed using established formulas when only change values were reported in the original studies [36]. The effect measures were estimated through the standardized mean difference (SMD) based on Hedges *g*, along with the 95% CI. To assess statistical heterogeneity, τ^2 was calculated. The Cochran Q test and I^2 were also derived, which together quantify the extent and significance

XSL•FO
RenderX

of the observed heterogeneity. Considering the expected significant heterogeneity among studies, we adopted a random-effects model to combine the effect sizes. The estimation of the CI was performed by the HKSJ (Hartung-Knapp-Sidik-Jonkman) method. For outcomes with enough studies (n3), we calculated 95% prediction intervals (PIs) to estimate the range of true effects in future settings. Forest plots are used to present the results of the merger.

Subgroup analyses were conducted to explore sources of heterogeneity, stratified by (1) type of DHIs (mobile health [mHealth]: telephone calls, text messages, and apps; eHealth: internet based, websites, and emails; or digital device: wearables and Bluetooth-connected devices) [37]; and (2) the duration of intervention (<3 months vs ≥3 months). The duration threshold was chosen based on common clinical rehabilitation milestones. If a prespecified subgroup analysis is not feasible due to a small number (<3 studies) of included studies, we will use narrative synthesis to summarize the included studies.

In addition, sensitivity analyses were conducted by sequentially excluding one study at a time to assess the reliability and stability of the main outcome (hip function). A funnel plot and Egger test were used to assess the small study effect. All tests were 2-tailed with a significance level of α=.05.

Results

Search Findings

The initial database search yielded 17,708 records, and an additional 15 records were identified through other sources, resulting in a total of 17,723 records. After removing 6105 duplicates, 11,618 records were excluded by title and abstract screening. Subsequently, the remaining 43 studies were screened for full text. Finally, a subset of 13 studies [29,38-49] was included. The PRISMA flow diagram is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Flowchart of the study selection process.

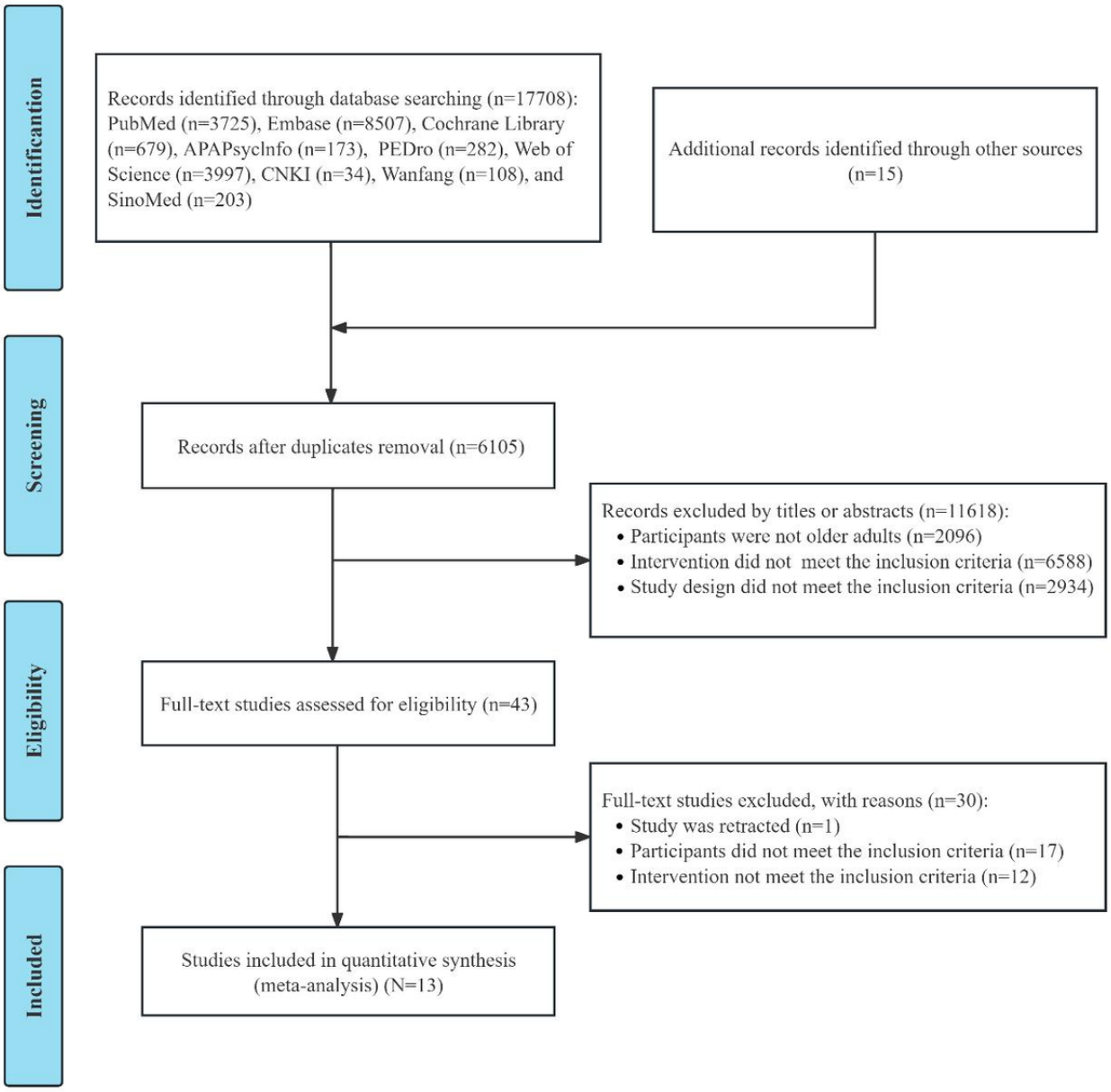

Characteristics of Included Studies

This systematic review included 13 studies [29,38-49] with 737 participants in total and published between 2015 and 2025 (Table 1). Eight studies were implemented in China [29,38-44], and one each in Israel [45], Sweden [46], Spain [47], Thailand [48], and Canada [49]. The age of participants ranged from 65.7 (SD 7.8) to 86.0 (SD 16.0) years, and the percentage of males ranged from 27.2% (38/138) to 80.6% (25/31). The intervention lasted between 3 weeks and 4 months, and the duration of follow-up was from 3 weeks to 9 months.

XSL•FO
RenderX

Table 1. Characteristics of included studies.

Study, year, country, and study design	Sample size (IG[a]:CG[b]); sex (M[c]:F[d]); age (IG:CG)	Intervention type and provider	Contents of the intervention group	Control group	Outcome and outcome measurement tools
Kalron et al [45], 2018, Israel: RCT[e] (FS[f]), Blind (participants, investigators)	32 (15:17); 20:12; 65.7(7.8):67.3(9.5)	mHealth[g] (app): physiotherapist	• Video lesson (includes health education, rehabilitation training standards, and key points of training movements) and progressive rehabilitation training. • Frequency: 5 times per week. • Duration: 4 months.	Regular manage	• Balance function and risk of falling: TUG[h]
Li et al [38], 2022, China, Hong Kong: RCT, Blind (outcome assessors)	31 (15:16); 25:6; 76.5 (8.6):82.1 (9.7)	eHealth (internet-based): physiotherapist	• Telerehabilitation was delivered through the Caspar Health e-system (CASPAR Health) to initiate falls prevention information, coaching to remain active, problem-solving skills, mobility goal setting, and advice to help participants maintain and increase their prescribed home exercises. • Frequency: personalized. • Duration: 8 weeks.	Usual care	• Functional independence: BI[i] • Balance function and risk of falling: TUG
Howell et al [46], 2023, Sweden: RCT (FS), Blind (participants)	21 (10:11); 15:6; 86.0 (16.0):83.0 (13.0)	Digital device (wearables): physiotherapist	• Participant wore a wearable device to measure body position and movements; the therapist receives feedback and can ensure that the rehabilitation program is tailored to the needs of the individual. • Frequency: no mention. • Duration: 4 weeks.	Standard rehabilitation	• Functional independence: BI • Balance function and risk of falling: FBG[j] • QoL[k]: EQ-5D
Langford et al [49], 2015, Canada: RCT, Blind (investigators, outcome assessors)	30 (15:15); 19:11; 83.0 (8.0):82.0 (10.0)	mHealth (app): physician	• Exercises relating to movement, strength of the lower limbs, and balance performance were included by using a mobile app. • Frequency: personalized. • Duration: 3 weeks.	Usual care	• QoL: EQ-5D
Cheng et al [29], 2022, China, Hong Kong: RCT, Open label	39 (19:20); 20:12; 75.8 (7.2):79.0 (8.8)	mHealth (app): physiotherapist	• Exercises with embedded videos in the app, receive a home-based rehabilitation program, and caregiver education. • Frequency: once per day. • Duration: 2 months.	Standard rehabilitation	• Functional independence: MFAC[l]
Prieto-Moreno et al [47], 2024, Spain and Belgium: RCT, Blind (outcome assessors)	105 (51:54); 30:75; 79.6 (7.1):80.1 (7.7)	eHealth: (internet-based); no mention	• Health knowledge, education, and telerehabilitation guidance, including physical exercise and an occupational therapy session. • Frequency: twice a week. • Duration: 12 weeks.	Routine rehabilitation	• Functional independence: FIM[m] • QoL: EQ-5D

XSL•FO
RenderX

Study, year, country, and study design	Sample size (IG[a]:CG[b]); sex (M[c]:F[d]); age (IG:CG)	Intervention type and provider	Contents of the intervention group	Control group	Outcome and outcome measurement tools
Li [44], 2024, China: RCT, No mention	80 (38:42); 47:33; 68.5 (5.6):69.4 (5.0)	mHealth (app): physician	• Remote rehabilitation guidance to the patients and their family caregivers via WeChat (Tencent) video, answered the patients' questions, and guided functional exercises. • Frequency: once each in the second, fourth, and eighth weeks. • Duration: 12 weeks.	Routine health education	• Hip Function: HHS[n] • QoL: SF-12[o]
Zhang [42], 2024, China: RCT, Blind (outcome assessors)	61 (30:31); 36:25; 69.0 (4.5):70.0 (7.4)	mHealth (app): physiotherapist	• The experimental group provided exercise guidance and training feedback to the subjects through a mobile app (Joymotion software, Shanghai Fudong Medical Management Co, Ltd), and conducted real-time 2-way audio and video interactive consultations with physical therapists. • Frequency: 5 times a week. • Duration: 4 weeks.	Routine rehabilitation	• Hip function: HOOS[p] • QoL: SF-12
Gao et al [39], 2021 China: RCT (No mention)	80 (40:40); 30:50; 77.8(5.9):78.1(5.6)	mHealth (app): physician	• Doctors transmitted text, pictures, voice, and video to the group to guide and urge patients to perform rehabilitation exercises through WeChat. For special patients, doctors would provide personalized and targeted guidance. • Frequency: no mention. • Duration: 2 months.	Routine rehabilitation	• Hip function: HHS
Zhang (1) [40], 2021, China: RCT, Blind (participants, outcome assessors)	51 (27:24); 18:33; 77.0 (7.9):75.2 (7.7)	eHealth (internet-based): physiotherapist	• The experimental group adopted the "Family-Oriented Postoperative Rehabilitation Management System for Elderly Hip Fractures" developed by Tianjin Hospital of Tianjin, which was used for remote home rehabilitation. This system is based on technologies such as Internet Plus, the Internet of Things, cloud computing, mobile Internet, and big data. It adopts the approach of "integrated application of technologies and rapid combination of modules" to establish a family rehabilitation treatment platform. • Frequency: no mention. • Duration: 3 months.	Usual care	• Hip function: HHS • Functional independence: FIM • Balance function and risk of falling: TUG • QoL: SF-36[q]
Zhang (2) [41], 2021, China: RCT (No mention)	36 (18:18); 15:21; 72.4 (8.4):68.8 (12.5)	mHealth (app): rehabilitation team		Usual care	• Hip Function: HHS • QoL: SF-36

XSL•FO
RenderX

Study, year, country, and study design	Sample size (IG[a]:CG[b]); sex (M[c]:F[d]); age (IG:CG)	Intervention type and provider	Contents of the intervention group	Control group	Outcome and outcome measurement tools
			• Participants received remote rehabilitation based on network media support, mainly using 2 social software programs, WeChat and TikTok (ByteDance). First, establish a WeChat group composed of rehabilitation physicians, surgeons, physical therapists, head nurses, patients, and their families. Make full use of their spare time to provide real-time rehabilitation guidance to patients and answer their questions and their families' questions. • Frequency: no mention. • Duration: 10 weeks.		
Dajpratham et al [48], 2025, Thailand: RCT, Blind (investigators)	33 (16:17); 6:27; 75.1 (9.5):78.3 (7.7)	mHealth (app): physiotherapist	• Real-time videoconference exercises with a physiotherapist via the LINE app. • Frequency: 3 times per week, with each session lasting approximately 45 minutes. • Duration: 12 weeks.	Usual care	• Balance function and risk of falling: SPPB[r] • Intervention adherence: percentage format
Shui [43], 2025, China: RCT (No mention)	138 (69:69); 38:100; 76.0 (5.7):75.2 (5.6)	mHealth (app): rehabilitation team	• Family remote rehabilitation program training: the attending physician, rehabilitation physician, and responsible nurse established an online communication group through the WeChat platform to strengthen patient management and guide the training content. • Frequency: 3 times per month. • Duration: 3 months.	Usual care	• Hip function: HHS • Functional independence: FIM • QoL: EQ-5D

[a]IG: intervention group.

[b]CG: control group.

[c]M: male.

[d]F: female.

[e]RCT: randomized controlled trial.

[f]FS: feasibility study.

[g]mHealth: mobile health.

[h]TUG: Timed Up and Go test.

[i]BI: Barthel Index.

[j]FBG: Functional Balance for Geriatric Patients.

[k]QoL: quality of life.

[l]MFAC: Modified Functional Ambulatory Category.

[m]FIM: Functional Independence Measure.

[n]HHS: Harris Hip Score.

[o]SF-12: Short Form–12 Health Survey.

[p]HOOS: Hip Disability and Osteoarthritis Outcome Score.

[q]SF-36: Medical Outcomes Study 36-Item Short Form Health Survey.

[r]SPPB: Short Physical Performance Battery.

XSL•FO
RenderX

In the 13 included studies [29,38-49], the type of DHIs could be divided into 3 categories, which included mHealth (n=9 [29,39,41-45,48,49]), eHealth (n=3 [38,40,47]), and digital device (n=1 [46]). The involvement of health care providers in DHIs varied across studies: a rehabilitation team was involved in 2 studies [41,43], physiotherapists in 7 studies [29,38,40,42,45,46,48], physicians in 3 studies [39,44,49], and 1 study [47] did not specify the intervention provider.

Of the 13 included trials [29,38-49], outcomes of our interest included hip function (6 studies [39-44]), functional independence (6 studies [29,38,40,43,46,47]), balance function and risk of falling (5 studies [38,40,45,46,48]), QoL (8 studies [40-44,46,47,49]), and intervention adherence (1 study [48]). Although the instruments used for outcome evaluation varied across studies, all were valid scales, and the process of data collection was carried out. Six studies [39-44] that reported hip function were assessed using the Harris Hip Score (HHS) [39-41,43,44] and Hip Disability and Osteoarthritis Outcome Score (HOOS) [42]. Six trials [29,38,40,43,46,47] evaluated functional independence using the Functional Independence Measure (FIM) [40,43,47], the Modified Functional Ambulatory Category (MFAC) [29], and the Barthel index [38,46]. Instruments used to evaluate balance function and risk of falling included the Timed Up and Go test (TUG) [38,40,45], Short Physical Performance Battery (SPPB) [48], and Functional Balance for Geriatric Patients (FBG) [46]. Eight studies [40-44,46,47,49] that reported QoL were assessed using the Medical Outcomes Study (MOS) 36-Item Short Form Health Survey (SF-36) [40,41], the EQ-5D scale [43,46,47,49], and the Short Form-12 Health Survey (SF-12) [42,44]. Intervention adherence was described using the percentage format in a single study [48].

Risk of Bias

Overall, 5 studies [40,42,45,47,48] had low risk of bias, and 5 studies [38,39,41,43,44] had some concerns. Another 3 studies [29,46,49] had a high risk of bias, 1 study [29] had a high risk in terms of the randomization process, 1 study [46] had a high risk in terms of the measurement of the outcome, and 1 study [49] had a high risk in terms of the selection of the reported result. Detailed results are presented in Figure 2.

Figure 2. Risk of bias.

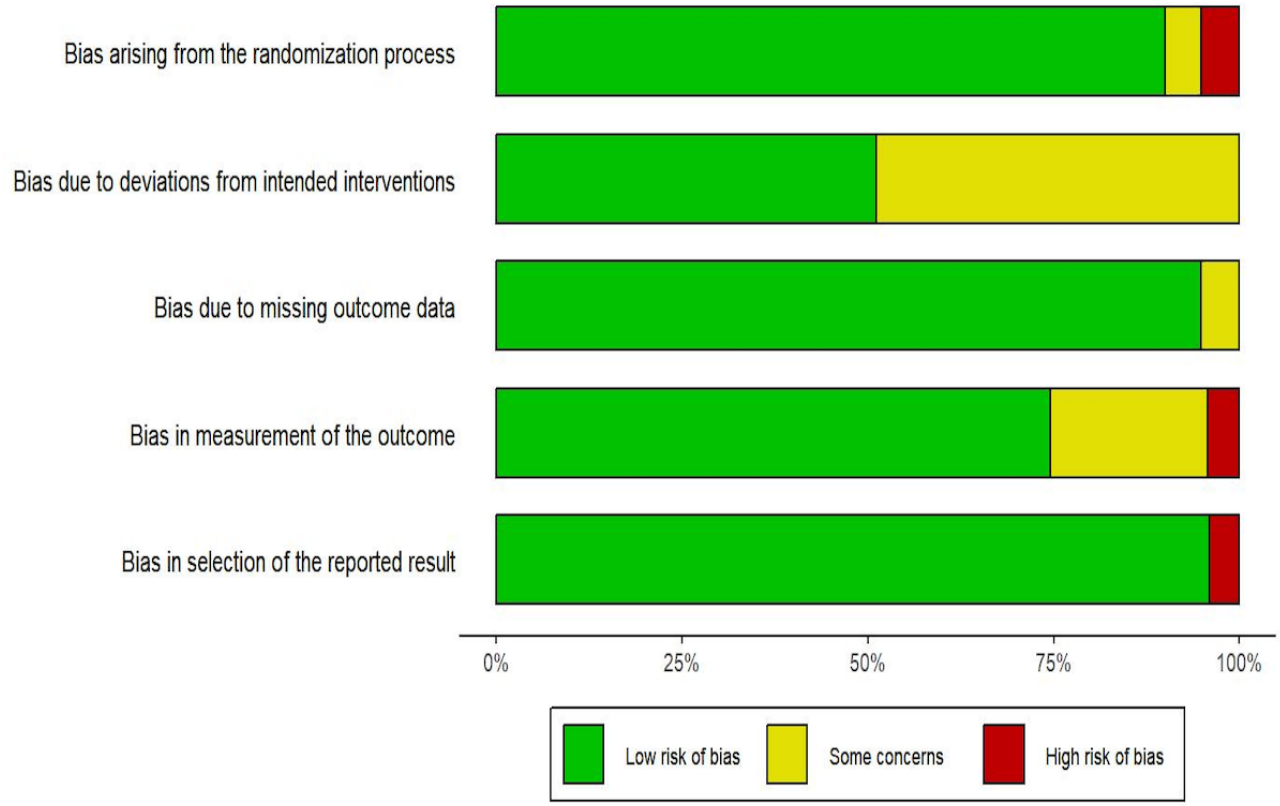

Certainty of Evidence

The GRADE assessment scores ranged from low to moderate (Table 2). Evidence was downgraded if (1) the risk of bias was apparent; (2) inconsistency was demonstrated, with I^2>75%; (3) there was indirectness in participants or comparators; (4) there was evidence of imprecision; or (5) there was publication bias. An overall GRADE rating was agreed upon (WF, QZ, and QL) for each outcome at 4 levels, including very low, low, moderate, and high.

XSL•FO
RenderX

Table 2. GRADE (Grading of Recommendations, Assessment, Development and Evaluation) summary of findings.

Outcomes	Number of participants (studies)	Quality of evidence (GRADE[a])	Relative effect (SMD[b], 95% CI)	Anticipated absolute effect	
				Control group, mean (SD)	DHIs[c] group, mean (95% CI)
Hip function	435 (6 RCTs[d])	(⊠) Moderate[e,f] (due to imprecision)	0.80 (0.33-1.26)	HHS[g]: 56.71 (14.71)	68.48 (61.56-75.24)
Functional independence	369 (6 RCTs)	(⊠) Low[e,f] (due to inconsistency, imprecision)	1.23 (0.34-2.11)	FIM[h]: 82.67 (5.69)	89.67 (84.60-94.68)
Balance function and risk of fall	163 (5 RCTs)	(⊠) Moderate[f] (due to imprecision)	–0.09 (–0.99 to 0.81)	TUG[i]: 34.16 (15.07)	32.80 (19.24-46.37)
Quality of life	501 (8 RCTs)	(⊠) Low[e,f] (due to inconsistency, imprecision)	0.65 (–0.21 to 1.50)	SF-36[j]: 56.71 (14.71)	66.27 (53.62-78.78)
Intervention adherence	33 (1 RCT)	—[k]	—	—	—

[a]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.

[b]SMD: standard mean difference.

[c]DHI: digital health intervention.

[d]RCT: randomized controlled trial.

[e]Inconsistency.

[f]Imprecision.

[g]HHS: Harris Hip Score.

[h]FIM: Functional Independence Measure.

[i]TUG: Timed Up and Go.

[j]SF-36: MOS 36-Item Short Form Health Survey.

[k]Not available.

Findings of Meta-Analysis

Effects of DHIs on Hip Function

As shown in Figure 3, 6 trials [39-44] with 435 patients were eligible for pooling for the analysis of hip function outcome. The pooled results from the random-effects model indicated that DHIs significantly improved hip function compared to the control group, with an SMD of 0.80 (95% CI 0.33-1.26; *P*=.007). The 95% PI (–0.24 to 1.83) further quantified the range of potential treatment effects across individual patients and clinical settings. Although pooled results suggest that DHIs yield a favorable average effect, some patients or subgroups might experience negligible or even marginal negative responses to the intervention, which aligns with the substantial heterogeneity observed (χ_5^2=15.1; *P*=.01; τ^2=0.13; I^2=67%).

XSL•FO
RenderX

Figure 3. Forest plot for the effect of digital health interventions on hip function [39-44]. DHI: digital health intervention; SMD: standardized mean difference.

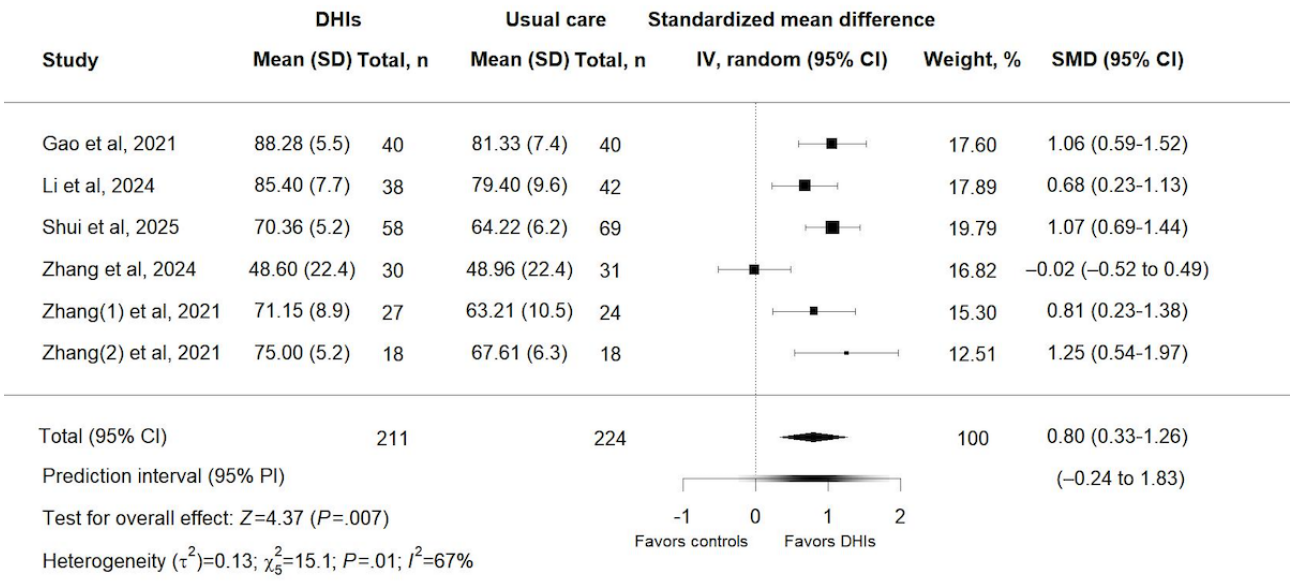

Effects of DHIs on Functional Independence

The effects of DHIs on functional independence were measured in 6 trials [29,38,40,43,46,47] involving 369 participants (Figure 4). Based on a random-effects model, pooled results showed a statistically significant beneficial effect of DHIs on enhancing functional independence in older patients with hip fractures, with an SMD of 1.23 (95% CI 0.34-2.11; *P*=.02). In stark contrast, the 95% PI (–0.98 to 3.34) reflects the wide spectrum of individual responses to DHIs, indicating that while DHIs yield a favorable mean effect, some patients may experience no improvement or even a slight decline in functional independence, whereas others could achieve marked benefits. This pronounced variability aligns with the extremely high heterogeneity (χ_5^2=39.7; *P*=.02; τ^2=0.62; I^2=87%) observed in the meta-analysis, highlighting the critical role of patient-specific and trial-related factors in modulating the effects of DHIs.

XSL•FO
RenderX

Figure 4. Forest plot for the effect of digital health interventions on functional independence [29,38,40,43,46,47]. DHI: digital health intervention; SMD: standardized mean difference.

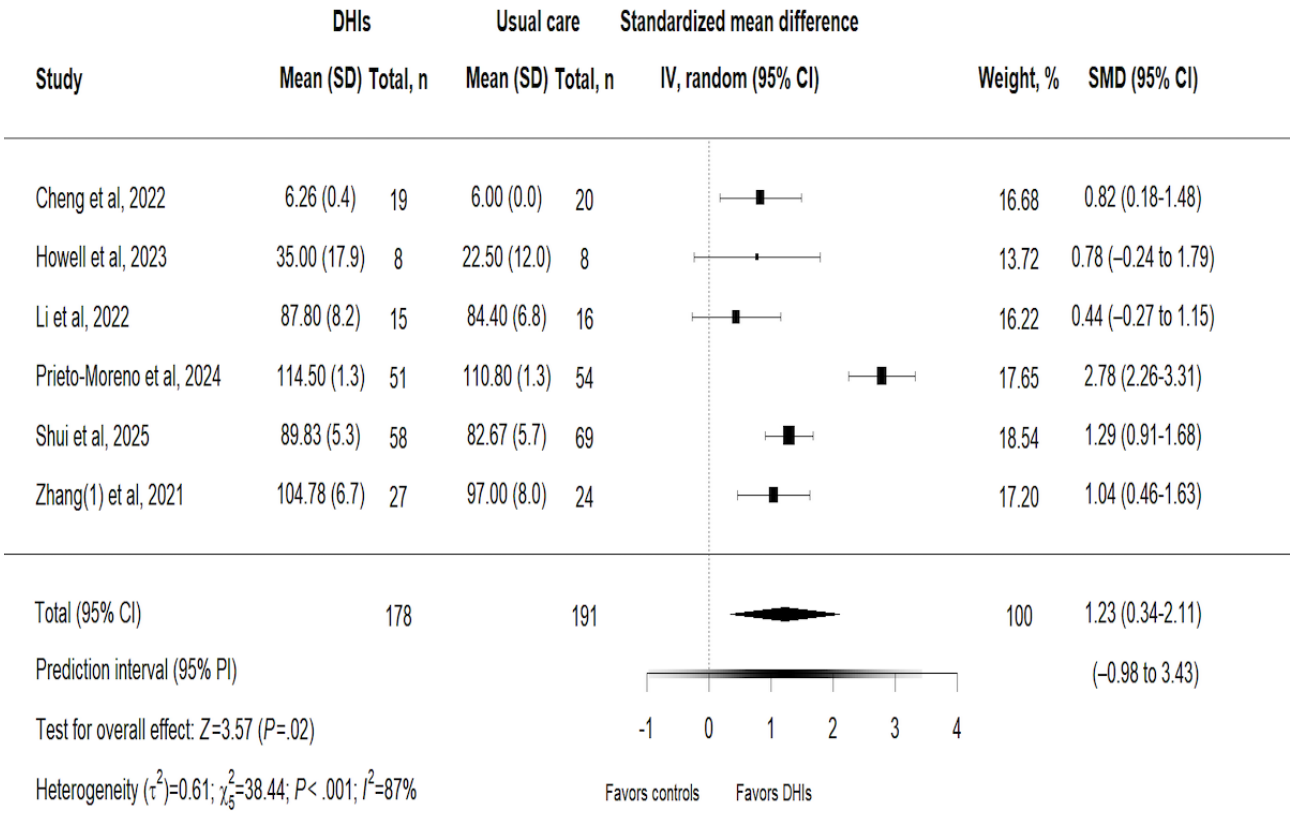

Effects of DHIs on Balance Function and Risk of Falling

In terms of balance function and risk of falling, it was measured in 5 trials [38,40,45,46,48] involving 163 participants. Based on a random-effects model, the results showed that compared to usual care, DHIs had no significant effects on improving balance function and reducing risk of falling (SMD –0.09, 95% CI –0.99 to 0.81; *P*=.80). The 95% PI (–2.07 to 1.89) was further estimated to delineate the variability of individual treatment responses and indicated that while the mean effect is negligible, some patients might experience meaningful improvements in balance function with DHIs, whereas others could exhibit worse outcomes. This pronounced variability in individual responses corresponds well with the high heterogeneity (χ_4^2=16.7; *P*=.002;τ^2=0.40; I^2=75%; Figure 5).

XSL•FO
RenderX

Figure 5. Forest plot for the effect of digital health interventions on balance function and risk of falling [38,40,45,46,48]. DHI: digital health intervention; SMD: standardized mean difference.

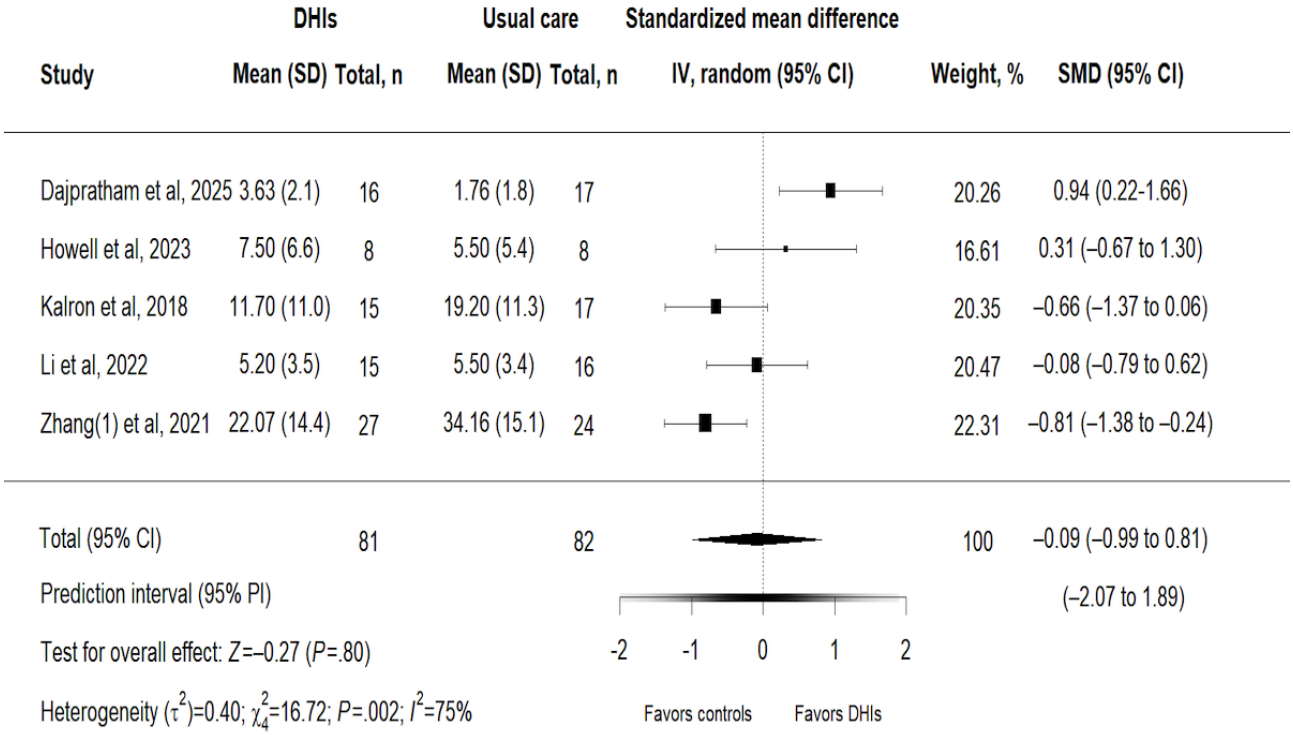

Effects of DHIs on Quality of Life

The effect of DHIs on QoL was measured in 8 trials [40-44,46,47,49] involving 501 participants. This meta-analysis adopted a random-effects model for data pooling. The overall results indicated that there was no statistically significant difference in QoL between groups, with an SMD of 0.65 (95% CI –0.21 to 1.50; *P*=.12). The 95% PI (–1.83 to 3.13) was further calculated, and substantial statistical heterogeneity was observed among the included trials (χ_7^2=89.5; *P*<.001; τ^2=0.97; I^2=93%; Figure 6).

XSL•FO
RenderX

Figure 6. Forest plot for the effect of digital health interventions on quality of life [46,49,44,47,43,42,41,40]. DHI: digital health intervention; SMD: standardized mean difference.

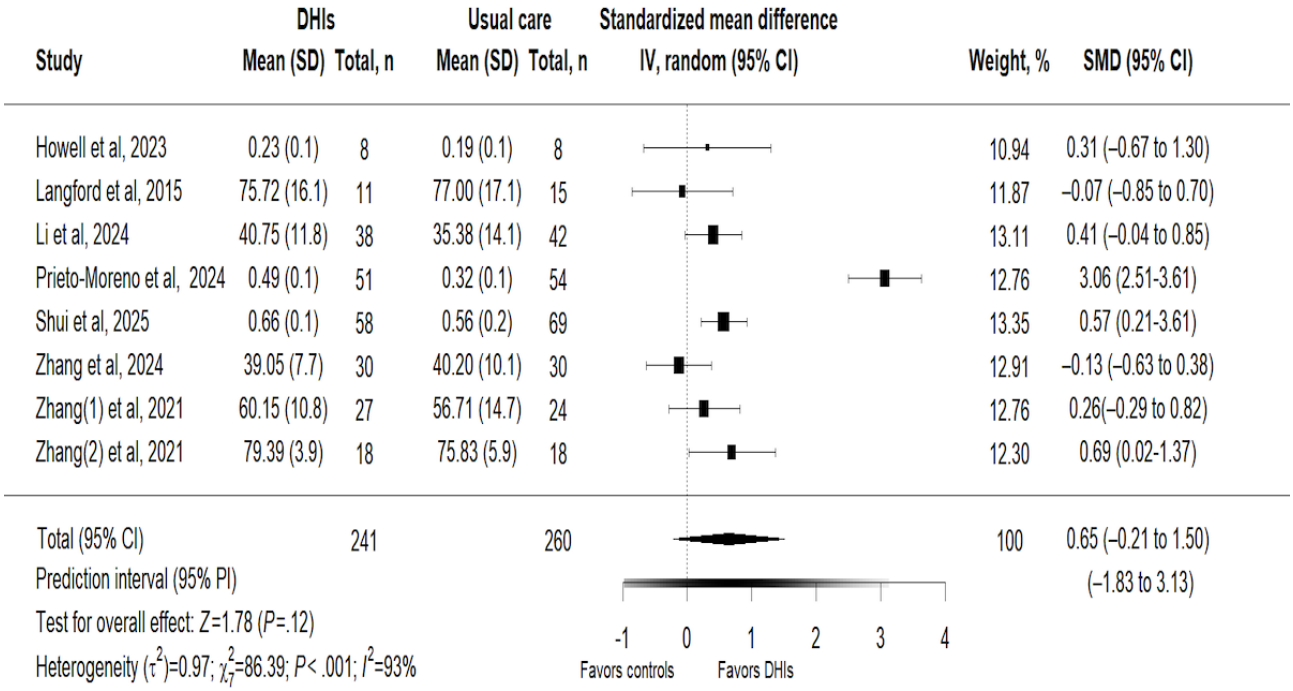

Effects of DHIs on Intervention Adherence

Due to the limited reporting of intervention adherence data across studies, a meta-analysis was not possible; only a single available study reported a 70% intervention adherence rate in the intervention group [48].

Subgroup Analyses

Subgroup Analysis by Type of Intervention

A subgroup analysis based on the type of DHIs was planned. However, the number of studies in each predefined intervention category (mHealth; n=9 [29,39,41-45,48,49], eHealth; n=3 [38,40,47], and digital device; n=1 [46]) fell below the threshold of 3 for all outcomes, and the planned subgroup analysis based on the type of DHIs was therefore forfeited due to the inability to provide meaningful quantitative estimates.

Subgroup Analysis by Intervention Duration

Subgroup analysis was conducted based on the intervention duration (<3 months vs ≥3 months) to explore the effects of DHIs on hip function. The subgroup analysis for hip function revealed no statistically significant difference between subgroups (χ_1^2=0.1; *P*=.75), with interventions lasting ≥3 months showing a greater effect (SMD 0.89, 95% CI 0.36-1.41; I^2=7%; *P*=.41) compared to those lasting <3 months (Figure 7). The test for subgroup differences in functional independence suggested a marginally significant variation in effects (χ_1^2=2.93; *P*=.09), indicating that intervention duration may be a potential contributor to the overall heterogeneity of the meta-analysis, albeit not reaching formal statistical significance, while the point estimate still favored shorter intervention duration (SMD 0.67, 95% CI 0.12-1.23; I^2=0%; *P*=.72; Figure 8).

XSL•FO
RenderX

Figure 7. Subgroup analysis on hip function [39,42,41,44,43,40]. DHI: digital health intervention; RE: Random effect; SMD: standardized mean difference.

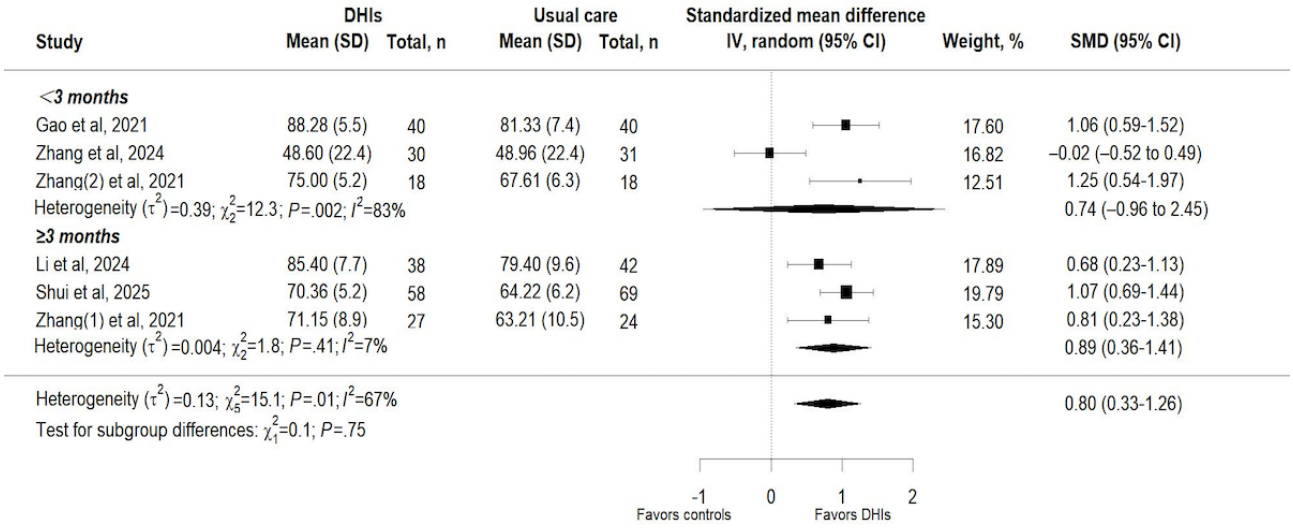

Figure 8. Subgroup analysis on functional independence [29,38,40,43,46,47]. DHI: digital health intervention; SMD: standardized mean difference.

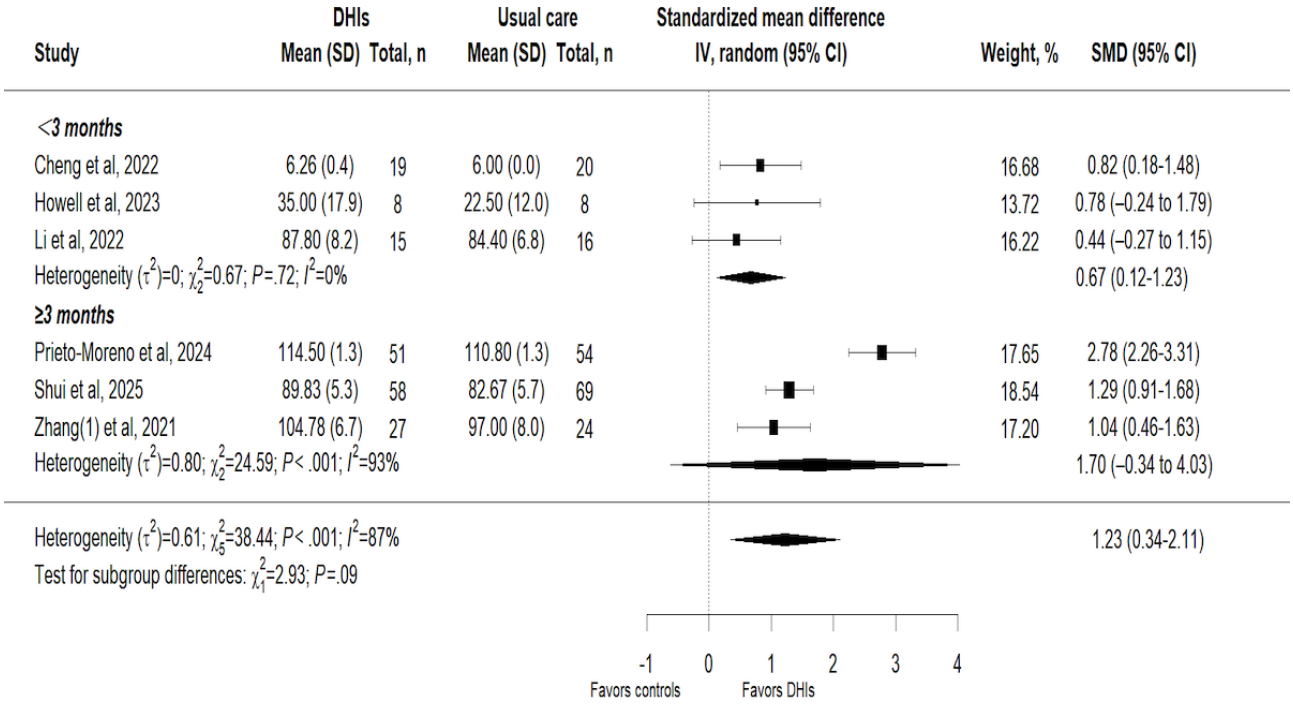

Small-Study Effects and Sensitivity Analysis

The funnel plot for hip function underlying the meta-analyses was symmetrical, and the Egger test was conducted to assess potential small-study effects, and no statistically significant evidence of such effects was identified in this meta-analysis (*z*-score=0.84; *P*=.40; Multimedia Appendix 4). Sensitivity analysis, performed by sequentially removing each study, confirmed that no single study unduly influenced the overall pooled estimate for hip function, demonstrating the robustness of this finding (Multimedia Appendix 4).

XSL•FO
RenderX

Discussion

Principal Findings

This meta-analysis confirms that DHIs confer statistically reliable average benefits for hip function and functional independence in older adults with hip fractures, whereas their effects on balance and QoL remain inconclusive. A wide 95% PI, however, indicates substantial real-world variability in individual DHI responses, restricting broad clinical deployment. Subgroup analyses also reflect this inconsistency—with intervention duration providing partial explanatory power—but the limited number of included studies necessitates further research to elucidate the underlying sources of such heterogeneity. Additionally, intervention adherence was reported in only 1 study [48], precluding quantitative assessment of DHI-related effects on this outcome. These results underscore that the potential of DHIs to promote functional and psychological recovery in this population requires further validation.

Low-to-moderate-certainty evidence, as determined by the GRADE assessment, suggests that DHIs are superior to control conditions in improving hip function and functional independence among older adults with hip fractures. This finding aligns with previous research supporting the role of telerehabilitation interventions in enhancing physical function following joint surgery and has suggested that this is related to the ability of DHIs to provide effective communication and feedback, as well as an educational component [21,50,51]. Health education during the postoperative rehabilitation stage is crucial for patients' recovery. Postoperative patients need to receive numerous recommendations, including early ambulation, exercise and rehabilitation programs, preventing postoperative complications, modifying home hazards, and taking nutrition supplements [21,52]. DHIs mostly adopt mobile apps as the mode of delivery, which can provide real-time or asynchronous communication and feedback. It can break down the boundaries of time and space, providing an effective communication platform between physiotherapists and patients, which is conducive to accurately transmitting rehabilitation knowledge and ultimately achieving positive functional recovery effects [53,54].

In addition, most DHIs systems incorporate exercise training, allowing physiotherapists to provide exercise guidance through voice, video, and other means, while receiving feedback from patients [55]. This helps patients correct improper movements promptly and ensures training quality [56]. Studies have shown that appropriate exercise after orthopedic surgery can counteract muscle weakness and functional disorders that may be caused by limited limb usage and also improve the range of joint motion, which is conducive to the recovery of the hip joint and physical function [21,57]. Currently, DHIs predominantly use combined exercises, such as resistance or strength training paired with balance training, to maximize functional improvement. Resistance training and strength training can not only enhance muscle strength but also lead to the upregulation of bone turnover and subsequent bone deposition, which has a positive regulatory effect on bone health [58-60]. Balance training is expected to improve balance function and prevent falls.

However, it is critical to emphasize that the effects of DHIs exhibited substantial real-world variability. For hip function, the 95% PI spanned from marginally negative to large positive treatment responses, which quantifies the true heterogeneity across populations and settings far more pragmatically than I^2 alone. For functional independence, the 95% PI covered an even broader range of individual patient responses, consistent with extremely high between-study heterogeneity. This pattern indicates inconsistent DHIs' self-efficacy across studies and patient subgroups in real-world settings. In addition, variations in study design (eg, differences in individual characteristics, DHIs delivery modes, and control group interventions) may have introduced bias, potentially confounding the pooled effect estimates. The low-to-moderate certainty of evidence reflects limitations, including imprecision in effect estimates and inconsistency across trials. Collectively, these factors mandate cautious clinical translation of the findings.

Moderate-certainty evidence suggested that the effects of DHIs on balance function and fall risk were uncertain, which contrasts with previous studies. Previous studies have reported the positive effects of DHIs on improving balance ability in populations, such as older adults and patients with neurological diseases [61-63]. A study investigating the mechanism of balance improvement in older adults after balance training found that balance training can increase the duration of cocontraction of the soleus or tibialis anterior, which can enhance joint stability and thereby improve the balance control ability of older adults [64]. However, the effects of balanced training also seem to have a dose-response relationship [65]. Earlier studies showed that a high practice intensity and an increased session frequency led to significant retention over the longer term [66,67]. However, increasing the intensity, frequency, and duration of such protocols could also have disadvantages, such as activity-induced fatigue and reduced participation [68]. It has also been found that the plasticity of the human brain provides physiological preparation for the restoration of balance function, but this still requires long-term persistence [69]. Therefore, the fact that no significant improvement in balance function was observed in this systematic review might be due to the lack of a suitable balance training program for older patients with hip fractures, or the duration of intervention was not sufficient to restore balance function.

It should also be noted that the wide 95% PI for balance function and fall risk reflects pronounced variability in individual patient responses, aligning with the high between-study heterogeneity observed. Additionally, many trials lacked standardized assessment tools for these outcomes and failed to implement outcome assessor blinding, and these methodological limitations likely introduced bias, confounded pooled effect precision, and contributed to the observed heterogeneity. The evidence for effects of DHIs on balance function and fall risk is rated as moderate certainty, owing to imprecise effect sizes. Collectively, these factors indicate that the null average effect should not be interpreted as definitive proof of DHIs' ineffectiveness, and instead, the broad PI underscores the potential for

XSL•FO
RenderX

subgroup-specific benefits or harms that aggregate analyses cannot capture.

Contrary to our expectations, low-certainty evidence showed no effect of DHIs on QoL. Molina-Garcia et al [70] indicated that telerehabilitation interventions can significantly improve QoL in patients with musculoskeletal disorders. This difference may be attributed to the fact that Molina-Garcia et al [70] focused on a broader range of musculoskeletal diseases, while our meta-analysis specifically targeted patients with hip fractures—a population typically characterized by advanced age, multiple diseases, and slow recovery [71]. Additionally, the intensity and duration of DHIs conducted in existing studies may not be sufficient to overcome the profound physical and psychological challenges associated with this injury and restore the QoL of older patients to the level before the fracture [12,72]. A cohort study also showed that social support and pain while walking can predict QoL among older adults 3 months after discharge from rehabilitation following a surgical hip fracture repair [73]. It indicates that the failure to effectively address pain may indirectly lead to a lack of statistically significant improvement in QoL outcomes. Therefore, in order to improve the psychological outcomes of older patients with hip fractures, it is best to integrate pain management and social support measures into DHIs.

The wide 95% PI also reflects pronounced variability in QoL in individual patient responses to DHIs, which aligns closely with the extremely high between-study heterogeneity observed. Included trials lacked a uniform QoL instrument and failed to account for confounding factors, such as baseline mental health status and social support levels, introducing methodological bias that may have inflated heterogeneity and compromised the precision of pooled estimates. Consequently, the evidence for effects of DHIs on QoL is rated as low certainty, constrained by imprecise effect sizes and inconsistent reporting of long-term QoL trajectories. Collectively, these findings indicate that the null average effect should not be equated with a lack of DHIs' efficacy, and instead, the broad PI highlights the potential for substantial subgroup-specific benefits or null effects that aggregate analyses cannot fully delineate.

Although a formal meta-analysis on intervention adherence was not feasible due to the lack of quantifiable data, the role of adherence in improving prognosis cannot be overlooked. In the included studies, Dajpratham et al [48] reported that teleresistance exercise showed an adherence rate of 70%, higher than that of the control group [48]. A recently published meta-analysis also indicates that telerehabilitation is associated with higher levels of medical adherence compared to conventional care [74]. All these reflect the potential of DHIs in improving intervention adherence. Theoretically, the convenience, reminders, and engaging nature of DHIs are posited to enhance adherence. However, future research must prioritize the standardized measurement and reporting of adherence metrics (eg, login frequency and exercise completion rates) to quantitatively validate this claim and understand its relationship with clinical outcomes.

Subgroup Analyses

Given the high heterogeneity, we performed subgroup analyses based on the type and duration of the intervention. Although only 1 study [46] used digital device-based DHIs (wearable devices), which prevented subgroup analysis by the type of intervention, the potential of digital device-based DHIs, especially wearable devices, to provide objective and continuous data on physical activity and movement patterns is immense [75,76]. It is reported that wearable devices have been used to collect body posture and movement data, providing real-time feedback to therapists to ensure rehabilitation programs are precisely tailored to individual needs [46]. It is evident that their role in enabling precise dosing of rehabilitation and detecting early signs of functional decline is a promising frontier [76]. The emerging field of wearable technology warrants dedicated investigation in the future, with more trials to isolate its specific contribution to the effectiveness of DHIs for hip fracture recovery [75].

Subgroup analyses stratified by intervention duration were performed to explore sources of heterogeneity in DHIs' effects on hip function, with findings interpreted in light of heterogeneity and risk of bias. The test for subgroup differences revealed no statistically significant effects, indicating that intervention duration was not a contributor to the overall heterogeneity observed in the meta-analysis. The 3 months subgroup yielded a significant pooled treatment effect, which suggests that prolonged intervention duration may mitigate outcome variability by enabling cumulative rehabilitation gains and sustained behavior modification. Additionally, more uniform long-term intervention protocols across included trials may have reduced methodological heterogeneity. This result may also be related to the dose-response relationship of rehabilitation [77,78]. A longer intervention period allows for a greater cumulative "dose" of therapy, which is necessary to counteract the severe physical deconditioning post fracture.

Interestingly, the subgroup analysis on functional independence based on intervention duration yielded inconsistent findings, with interventions lasting <3 months appearing more effective than those lasting 3 months. Notably, however, each of the 3 individual studies [3,40,43] in the 3 months subgroup demonstrated a statistically significant improvement in functional independence for the DHIs group relative to controls. When pooled, these results became statistically insignificant—a discrepancy attributable to marked between-study heterogeneity within this subgroup. These 3 studies used the same measurement tools, and the degree of improvement evaluated by Prieto Moreno et al [47] was much higher than that of the other 2 studies [40,43], which might be related to their higher frequency, intervention, and usage rate. Prieto Moreno et al [47] had an intervention frequency of twice a week and reported a relatively high app usage rate. The other 2 studies, however, had no reports on usage rate or adherence rate and had a lower intervention frequency.

Collectively, these findings underscore the complexity of optimizing DHI protocols for older adults with hip fractures. While prolonged intervention duration conferred consistent benefits for hip function, its effects on functional independence

XSL•FO
RenderX

appear contingent on ensuring sufficient intervention frequency and monitoring adherence. Future research should therefore prioritize standardized reporting of intervention dose and adherence metrics, alongside rigorous protocol design that specifies both duration and frequency of DHIs delivery, to clarify and replicate these subgroup-specific effects.

Comparison with Prior Work

With the increasing popularity of DHIs, scholars are paying more and more attention to their effects in the field of bone and joint rehabilitation. Several similar systematic reviews and meta-analyses have been published. While Molina-Garcia et al [70] provided a broad overview of DHIs across musculoskeletal disorders (including hip fractures), our study focused specifically on older patients with hip fractures, which made the evidence we provided direct and specifically applicable to clinicians and rehabilitation experts managing older patients with hip fractures, enabling them to make more precise decisions regarding DHIs.

Two published systematic reviews and meta-analyses have focused on the older population with hip fractures, as we have. One of the studies was published by Pliannuom et al [21] in 2024, focusing only on the impact of DHIs on functional outcomes, while we simultaneously evaluated functional and psychological outcomes and intervention adherence [21]. Furthermore, their paper was limited by the number of studies (only 6 RCTs [39-44] were included), and they were unable to conduct subgroup analyses to further clarify which subgroups could benefit the most from DHIs. Compared with another meta-analysis by Xiao et al [74], our population mainly focuses on older patients with hip fractures rather than all patients after hip fractures and replacement surgeries. Furthermore, compared with Xiao et al [74], our analysis included more of the latest published studies and conducted a GRADE assessment on the included studies to ensure the quality of the studies and the reliability of the results of the meta-analysis [47,48]. Our study and the study by Xiao et al [74] both performed subgroup analyses of hip function based on intervention duration, and both found that a longer intervention period was associated with a greater improvement effect on hip function [74]. However, in the study by Xiao et al [74], the 6-month subgroups showed extremely high heterogeneity (I^2=95%), indicating instability. In our study, the heterogeneity of the ≥3 months subgroup was low (I^2=7%), which indicates that our findings are more robust and reliable.

However, our review indicates that only low-to-moderate certainty evidence supports the effects of DHIs on our outcomes of interest. This suggests that while there may be some positive impact, the evidence is not strong enough to draw definitive conclusions. The variability in study quality, sample sizes, and methodologies may contribute to the uncertainty, highlighting the need for more robust and well-conducted trials to strengthen the evidence base. Therefore, the results should be interpreted carefully, as future research might yield divergent findings.

Strengths and Limitations

This meta-analysis followed the guidelines for performing rigorous systematic reviews. We also proposed a rigorous screening and search strategy to identify the most comprehensive literature. Sensitivity analyses revealed robust findings. Consequently, the results of our review support a widespread belief. However, there are also limitations in this systematic review. First, high statistical heterogeneity was observed in some results, which could not be fully explained by the subgroup analysis we prespecified. This may stem from the diversity of intervention components (such as types, dose, and intervention providers), control groups, and patient characteristics. Second, the limited number of studies on certain outcomes has restricted the statistical power and accuracy of estimates. Third, some studies had high risk or some concerns, particularly in terms of randomization and selection of reported results. This may have influenced the overall findings. Fourth, only one study [48] included in our review reported intervention adherence, which limits the quantitative evaluation for adherence of DHIs. Finally, as most of the research (61.5%) was conducted in China, the universality of the research results may be limited, and their applicability to other health care systems and cultures needs to be verified.

Further Research and Clinical Implications

For further research, the sources of heterogeneity should be identified based on patient characteristics (eg, prefracture functional level), intervention dose (eg, frequency and duration), as our meta-analysis observed significant heterogeneity. The high heterogeneity observed across studies also underscores the need for standardized protocols of DHIs. Future research should also aim to identify the most effective components of DHIs and develop standardized intervention frameworks to reduce variability and improve the generalizability of findings. Second, as the existing RCTs lack assessment of psychological outcomes and intervention adherence, future RCTs should consider including these outcomes, which are particularly important for a more comprehensive evaluation of the effect of DHIs. Third, due to the limited number of studies using digital devices for DHIs, subgroup analysis based on intervention types was not achieved in this systematic review. Future research needs to explore the effects of digital devices to further clarify the impact of intervention types on improving the functional and psychological outcomes of patients with hip fractures. Fourth, future research should further explore the dose-response of exercise rehabilitation on the recovery of physical function after hip fractures to achieve the best intervention effect. Clinically, our research results indicate that DHIs have potential value in improving hip function and functional independence in older patients with hip fractures. Clinicians should consider incorporating DHIs into clinical practice and using them as routine care after hip fractures in older adults. This is of great significance for optimizing the allocation of medical resources and expanding the scope of rehabilitation services.

Conclusions

This review incorporates the latest RCTs and comprehensively assesses both functional and psychological outcomes of DHIs in older patients with hip fractures, distinct from prior studies focusing solely on functional outcomes. While 95% CI supports the potential of DHIs to improve hip function and functional independence, the wide 95% PI indicating substantial real-world response variability calls for cautious interpretation, informing

XSL•FO
RenderX

the design of targeted DHI-based rehabilitation regimens, and warrants further research into optimal techniques and dosages in clinical practice.

Acknowledgments

No artificial intelligence tools were used in any aspect of this manuscript, from study implementation to manuscript preparation. All content reflects the authors' original work. We also sincerely acknowledge Dr Shang Wenru from the Hospital Management Center of Shanghai Sixth People's Hospital for peer-reviewing our search strategy by verifying search terms, database selection, and eligibility criteria prior to retrieval.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated or analyzed during this systematic review are available from the corresponding author on reasonable request.

Authors' Contributions

QZ and QFL were responsible for conceptualization, methodology, supervision, validation, and writing–original draft. QFL handled project administration and resources. WF and QZ completed data curation, formal analysis, software, and visualization. WF, QZ, and QL contributed to writing–review & editing. QZ and QFL are considered co-corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISIMA 2020 checklist.
[DOCX File , 28 KB - jmir_v28i1e79563_app1.docx]

Multimedia Appendix 2
PRISMA-S checklist.
[DOCX File , 16 KB - jmir_v28i1e79563_app2.docx]

Multimedia Appendix 3
Search strategy.
[DOCX File , 19 KB - jmir_v28i1e79563_app3.docx]

Multimedia Appendix 4
Sensitivity analysis.
[DOCX File , 88 KB - jmir_v28i1e79563_app4.docx]

References

1. Neuman MD, Silber JH, Magaziner JS, Passarella MA, Mehta S, Werner RM. Survival and functional outcomes after hip fracture among nursing home residents. JAMA Intern Med 2014;174(8):1273-1280 [FREE Full text] [doi: 10.1001/jamainternmed.2014.2362] [Medline: 25055155]
2. Baji P, Patel R, Judge A, Johansen A, Griffin J, Chesser T, REDUCE Study Group. Organisational factors associated with hospital costs and patient mortality in the 365 days following hip fracture in England and Wales (REDUCE): a record-linkage cohort study. Lancet Healthy Longev 2023;4(8):e386-e398 [FREE Full text] [doi: 10.1016/S2666-7568(23)00086-7] [Medline: 37442154]
3. Sing C, Lin TC, Bartholomew S, Bell JS, Bennett C, Beyene K, et al. Global epidemiology of hip fractures: secular trends in incidence rate, post-fracture treatment, and all-cause mortality. J Bone Miner Res 2023;38(8):1064-1075 [FREE Full text] [doi: 10.1002/jbmr.4821] [Medline: 37118993]
4. Gullberg B, Johnell O, Kanis J. World-wide projections for hip fracture. Osteoporos Int 1997;7(5):407-413. [doi: 10.1007/pl00004148] [Medline: 9425497]

XSL•FO
RenderX

5. Feng J, Zhang C, Li B, Zhan S, Wang S, Song C. Global burden of hip fracture: The Global Burden of Disease Study. Osteoporos Int 2024;35(1):41-52. [doi: 10.1007/s00198-023-06907-3] [Medline: 37704919]
6. Nordström P, Gustafson Y, Michaëlsson K, Nordström A. Length of hospital stay after hip fracture and short term risk of death after discharge: a total cohort study in Sweden. BMJ 2015;350:h696 [FREE Full text] [doi: 10.1136/bmj.h696] [Medline: 25700551]
7. Branca Vergano L, Monesi M, Vicenti G, Bizzoca D, Solarino G, Moretti B. Posterior approaches in malleolar fracture: when, why and how. J Biol Regul Homeost Agents 2020;34(3 Suppl. 2):89-95. [Medline: 32856446]
8. Araiza-Nava B, Méndez-Sánchez L, Clark P, Peralta-Pedrero ML, Javaid MK, Calo M, et al. Short- and long-term prognostic factors associated with functional recovery in elderly patients with hip fracture: a systematic review. Osteoporos Int 2022;33(7):1429-1444. [doi: 10.1007/s00198-022-06346-6] [Medline: 35247062]
9. Yang T, Kuo Y, Huang S, Chen Y. Minimal short-term decline in functional performance and quality of life predicts better long-term outcomes for both in older Taiwanese adults after hip fracture surgery: a prospective study. J Orthop Surg Res 2023;18(1):791 [FREE Full text] [doi: 10.1186/s13018-023-04278-3] [Medline: 37872535]
10. Nightingale S, Holmes J, Mason J, House A. Psychiatric illness and mortality after hip fracture. Lancet 2001;357(9264):1264-1265 [FREE Full text] [doi: 10.1016/S0140-6736(00)04421-4] [Medline: 11418154]
11. Milton-Cole R, Ayis S, Lambe K, O'Connell MDL, Sackley C, Sheehan KJ. Prognostic factors of depression and depressive symptoms after hip fracture surgery: systematic review. BMC Geriatr 2021;21(1):537 [FREE Full text] [doi: 10.1186/s12877-021-02514-1] [Medline: 34627160]
12. Alexiou K, Roushias A, Varitimidis S, Malizos K. Quality of life and psychological consequences in elderly patients after a hip fracture: a review. Clin Interv Aging 2018;13:143-150 [FREE Full text] [doi: 10.2147/CIA.S150067] [Medline: 29416322]
13. Wang J, Tong Y, Jiang Y, Zhu H, Gao H, Wei R, et al. The effectiveness of extended care based on Internet and home care platform for orthopaedics after hip replacement surgery in China. J Clin Nurs 2018;27(21-22):4077-4088. [doi: 10.1111/jocn.14545] [Medline: 29851157]
14. Rosas Hernández AM, Alarcón T, Menéndez-Colino R, Martín Maestre I, González-Montalvo JI. Factors affecting exercise program adherence in patients with acute hip fracture and impact on one-year survival. Braz J Phys Ther 2020;24(6):479-487 [FREE Full text] [doi: 10.1016/j.bjpt.2019.07.008] [Medline: 31378633]
15. Taylor N, Rimayanti MU, Peiris CL, Snowdon DA, Harding KE, Semciw AI, et al. Hip fracture has profound psychosocial impacts: a systematic review of qualitative studies. Age Ageing 2024;53(9):afae194. [doi: 10.1093/ageing/afae194] [Medline: 39238124]
16. Crotty M. Rehabilitation interventions for improving physical and psychosocial functioning after hip fracture in older people. Cochrane Database Syst Rev 2010:20091644. [doi: 10.1002/14651858.cd007624]
17. Handoll H, Cameron ID, Mak J, Panagoda CE, Finnegan TP. Multidisciplinary rehabilitation for older people with hip fractures. Cochrane Database Syst Rev 2021;11(11):Cd007125 [FREE Full text] [doi: 10.1002/14651858.CD007125.pub3] [Medline: 34766330]
18. Blackburn J, Yeowell G. Patients' perceptions of rehabilitation in the community following hip fracture surgery. a qualitative thematic synthesis. Physiotherapy 2020;108:63-75. [doi: 10.1016/j.physio.2020.02.001] [Medline: 32711229]
19. Lewis S, Macey R, Lewis J, Stokes J, Gill J, Cook JA, et al. Surgical interventions for treating extracapsular hip fractures in older adults: a network meta-analysis. Cochrane Database Syst Rev 2022;2(2):CD013405 [FREE Full text] [doi: 10.1002/14651858.CD013405.pub2] [Medline: 35142366]
20. Recommendations on digital interventions for health system strengthening. World Health Organization.: World Health Organization; 2019. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2026-01-11]
21. Pliannuom S, Pinyopornpanish K, Buawangpong N, Wiwatkunupakarn N, Mallinson PAC, Jiraporncharoen W, et al. Characteristics and effects of home-based digital health interventions on functional outcomes in older patients with hip fractures after surgery: systematic review and meta-analysis. J Med Internet Res 2024;26:e49482. [doi: 10.2196/49482]
22. Backman C, Papp S, Harley A, Skidmore B, Green M, Shah S, et al. Patient-clinician digital health interventions for the hip fracture population: a scoping review. BMC Health Serv Res 2023;23(1):1052 [FREE Full text] [doi: 10.1186/s12913-023-09784-y] [Medline: 37784118]
23. Zhang J, Yang M, Ge Y, Ivers R, Webster R, Tian M. The role of digital health for post-surgery care of older patients with hip fracture: a scoping review. Int J Med Inform 2022;160:104709 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104709] [Medline: 35144101]
24. Chan JKY, Klainin-Yobas P, Chi Y, Gan JKE, Chow G, Wu XV. The effectiveness of e-interventions on fall, neuromuscular functions and quality of life in community-dwelling older adults: a systematic review and meta-analysis. Int J Nurs Stud 2021;113:103784. [doi: 10.1016/j.ijnurstu.2020.103784] [Medline: 33120138]
25. Morita PP, Sahu KS, Oetomo A. Health monitoring using smart home technologies: scoping review. JMIR Mhealth Uhealth 2023;11:e37347. [doi: 10.2196/37347]
26. Oyibo K, Wang K, Morita PP. Using smart home technologies to promote physical activity among the general and aging populations: scoping review. J Med Internet Res 2023;25:e41942 [FREE Full text] [doi: 10.2196/41942] [Medline: 37171839]

XSL•FO
RenderX

27. Gilboa Y, Maeir T, Karni S, Eisenberg ME, Liebergall M, Schwartz I, et al. Effectiveness of a tele-rehabilitation intervention to improve performance and reduce morbidity for people post hip fracture - study protocol for a randomized controlled trial. BMC Geriatr 2019;19(1):135 [FREE Full text] [doi: 10.1186/s12877-019-1141-z] [Medline: 31109289]
28. Pfeiffer K, Kampe K, Klenk J, Rapp K, Kohler M, Albrecht D, et al. Effects of an intervention to reduce fear of falling and increase physical activity during hip and pelvic fracture rehabilitation. Age Ageing 2020;49(5):771-778. [doi: 10.1093/ageing/afaa050] [Medline: 32832985]
29. Cheng KC, Lau KMK, Cheng AS, Lau TSK, Lau FOT, Lau MCH, et al. Use of mobile app to enhance functional outcomes and adherence of home-based rehabilitation program for elderly with hip fracture: a randomized controlled trial. Hong Kong Physiother J 2022;42(2):99-110 [FREE Full text] [doi: 10.1142/S101370252250010X] [Medline: 37560168]
30. Câmara SMA, Falvey JR, Orwig D, Gruber-Baldini AL, Auais M, Feng Z, et al. Associations between living alone, social interactions, and physical performance differ by sex: results from the Baltimore Hip Studies. J Am Geriatr Soc 2023;71(9):2788-2797. [doi: 10.1111/jgs.18403] [Medline: 37171145]
31. Shams F, Tai AM, Kim J, Boyd M, Meyer M, Kazemi A, et al. Adherence to e-health interventions for substance use and the factors influencing it: systematic Review, meta-analysis, and meta-regression. Digit Health 2023;9:20552076231203876 [FREE Full text] [doi: 10.1177/20552076231203876] [Medline: 37780062]
32. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
33. Rethlefsen M, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
34. Classification of digital health interventions v1.0.: World Health Organization; 2018. URL: https://www.who.int/publications/i/item/WHO-RHR-18.06 [accessed 2026-01-11]
35. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The cochrane collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928 [FREE Full text] [doi: 10.1136/bmj.d5928] [Medline: 22008217]
36. Xu YP, Yu XY, Fu WJ, Tao Y, Wu JY, He Q, et al. Introduction to the methods of missing standard deviation conversion and standardized mean difference estimation in Meta-Analysis. Chinese journal of evidence-based cardiovascular medicine, 2017 Feb 1;8(12):1412-1415.
37. Zangger G, Bricca A, Liaghat B, Juhl CB, Mortensen SR, Andersen RM, et al. Benefits and harms of digital health interventions promoting physical activity in people with chronic conditions: systematic review and meta-analysis. J Med Internet Res 2023;25:e46439 [FREE Full text] [doi: 10.2196/46439] [Medline: 37410534]
38. Li CT, Hung GK, Fong KN, Gonzalez PC, Wah S, Tsang HW. Effects of home-based occupational therapy telerehabilitation via smartphone for outpatients after hip fracture surgery: a feasibility randomised controlled study. J Telemed Telecare 2022;28(4):239-247. [doi: 10.1177/1357633X20932434] [Medline: 32594832]
39. Gao S, Wang Y, Zhang G, Zhang W. Rehabilitation guidance for hip fracture patients during the COVID-19 pandemic using chat software: a new model. J Back Musculoskelet Rehabil 2021;34(3):337-342. [doi: 10.3233/BMR-200324] [Medline: 33896813]
40. Zhang Y. Research on the Application of Internet + Rehabilitation Management System in Family Rehabilitation after Hip Fracture Surgery in the Elderly [dissertation]. Tianjin, China: Tianjin University of Sport; 2021.
41. Zhang GZ, Ding L, Lin QJ, Gu LD, Wang X. Evaluation of the remote rehabilitation effect supported by network media for patients after total hip arthroplasty. Chinese Journal of Bone and Joint Injury 2021 Oct;36(10):1053-1055.
42. Zhang X, Lu YT, Zhang C, Li H, Lu SD, Zhang ZA. Comparison of the therapeutic effects of remote home rehabilitation and inpatient rehabilitation after total hip replacement. Journal of Tongji University(Medical Science) 2024 Feb;45(1):57-65.
43. Shui XM, Xu Q, Sun YT, Xu MP, Kong HX. Clinical study on the impact of family remote rehabilitation training programs on postoperative rehabilitation outcomes of hip fractures in the elderly. Chinese Journal of Bone and Joint 2025 Jul 30;14(7):640-644.
44. Li ZW, Fan JJ, Wang TR, Yuan Z. Evaluation of the effect of remote rehabilitation on elderly patients and their family caregivers after intertrochanteric fracture surgery. Journal of Air Force Medical University 2024;45(11):1262-1266.
45. Kalron A, Tawil H, Peleg-Shani S, Vatine JJ. Effect of telerehabilitation on mobility in people after hip surgery: a pilot feasibility study. Int J Rehabil Res 2018;41(3):244-250. [doi: 10.1097/MRR.0000000000000296] [Medline: 29794545]
46. Howell DF, Malmgren Fänge A, Rogmark C, Ekvall Hansson E. Rehabilitation outcomes following hip fracture of home-based exercise interventions using a wearable device-A randomized controlled pilot and feasibility study. Int J Environ Res Public Health 2023;20(4):3107 [FREE Full text] [doi: 10.3390/ijerph20043107] [Medline: 36833801]
47. Prieto-Moreno R, Mora-Traverso M, Estévez-López F, Molina-Garcia P, Ortiz-Piña M, Salazar-Graván S, et al. Effects of the activeHip+ mHealth intervention on the recovery of older adults with hip fracture and their family caregivers: a multicentre open-label randomised controlled trial. EClinicalMedicine 2024;73:102677 [FREE Full text] [doi: 10.1016/j.eclinm.2024.102677] [Medline: 38911836]

XSL•FO
RenderX

48. Dajpratham P, Komas J, Yamthed R, Chanthon P, Kovintaset K, Claikhem T, et al. Feasibility and efficacy of real-time teleresistance exercise programs for physical function in elderly patients after hip fracture surgery: a randomized controlled trial. BMC Geriatr 2025;25(1):647 [FREE Full text] [doi: 10.1186/s12877-025-06230-y] [Medline: 40836222]
49. Langford D, Fleig L, Brown KC, Cho NJ, Frost M, Ledoyen M, et al. Back to the future - feasibility of recruitment and retention to patient education and telephone follow-up after hip fracture: a pilot randomized controlled trial. Patient Prefer Adherence 2015;9:1343-1351 [FREE Full text] [doi: 10.2147/PPA.S86922] [Medline: 26491262]
50. Suso-Martí L. Effectiveness of telerehabilitation in physical therapist practice: an umbrella and mapping review with meta-meta-analysis. Phys Ther 2021;101(5):pzab075. [doi: 10.1093/ptj/pzab075]
51. Pastora-Bernal JM, Martín-Valero R, Barón-López FJ, Estebanez-Pérez MJ. Evidence of benefit of telerehabitation after orthopedic surgery: systematic review. J Med Internet Res 2017;19(4):e142 [FREE Full text] [doi: 10.2196/jmir.6836] [Medline: 28455277]
52. Freitag MH, Magaziner J. Post-operative considerations in hip fracture management. Curr Rheumatol Rep 2006;8(1):55-62. [doi: 10.1007/s11926-006-0027-1] [Medline: 16515767]
53. Wu W, Zhang Y, Zhang Y, Peng B, Xu W. Clinical effectiveness of home-based telerehabilitation program for geriatric hip fracture following total hip replacement. Orthop Surg 2023;15(2):423-431 [FREE Full text] [doi: 10.1111/os.13521] [Medline: 36448261]
54. Vincent C, Wegier P, Chien V, Kurahashi AM, Ginsburg S, Molla Ghanbari H, et al. Qualitative evaluation of a Novel Educational tool to communicate individualized hip fracture prognostic information to patients and surrogates: my hip fracture (My-HF). Geriatr Orthop Surg Rehabil 2021;12:21514593211050513 [FREE Full text] [doi: 10.1177/21514593211050513] [Medline: 34707914]
55. Pak SS, Janela D, Freitas N, Costa F, Moulder R, Molinos M, et al. Comparing digital to conventional physical therapy for chronic shoulder pain: randomized controlled trial. J Med Internet Res 2023;25:e49236 [FREE Full text] [doi: 10.2196/49236] [Medline: 37490337]
56. Tore NG, Oskay D, Haznedaroglu S. The quality of physiotherapy and rehabilitation program and the effect of telerehabilitation on patients with knee osteoarthritis. Clin Rheumatol 2023;42(3):903-915 [FREE Full text] [doi: 10.1007/s10067-022-06417-3] [Medline: 36279075]
57. Brown MC, Lin CQ, Jin C, Rohde M, Rocos B, Belding J, et al. Effect of a digital health exercise program on the intention for spinal surgery in adult spinal deformity: exploratory cross-sectional survey. JMIR Form Res 2025;9:e66889 [FREE Full text] [doi: 10.2196/66889] [Medline: 40300159]
58. Salehian F, Mahmoudzadeh-Sagheb Z, Yoosefinejad AK, Zakerabasali S. A home-based tele-rehabilitation exercise system for patients after knee replacement surgery. BMC Musculoskelet Disord 2024;25(1):605 [FREE Full text] [doi: 10.1186/s12891-024-07731-4] [Medline: 39085850]
59. Chang X, Xu S, Zhang H. Regulation of bone health through physical exercise: mechanisms and types. Front Endocrinol (Lausanne) 2022;13:1029475 [FREE Full text] [doi: 10.3389/fendo.2022.1029475] [Medline: 36568096]
60. Troy K. Exercise Early and Often: Effects of Physical Activity and Exercise on Women's Bone Health. Int J Environ Res Public Health 2018;15(5).
61. Park C, Lee B. A Systematic review of the effects of interactive telerehabilitation with remote monitoring and guidance on balance and gait performance in older adults and individuals with neurological conditions. Bioengineering (Basel) 2024;11(5):460 [FREE Full text] [doi: 10.3390/bioengineering11050460] [Medline: 38790328]
62. He Z, Wu H, Zhao G, Zhang Y, Li C, Xing Y, et al. The effectiveness of digital technology-based Otago Exercise Program on balance ability, muscle strength and fall efficacy in the elderly: a systematic review and meta-analysis. BMC Public Health 2025;25(1):71 [FREE Full text] [doi: 10.1186/s12889-024-21251-9] [Medline: 39773392]
63. Gaspar AGM, Lapão LV. eHealth for addressing balance disorders in the elderly: systematic review. J Med Internet Res 2021;23(4):e22215 [FREE Full text] [doi: 10.2196/22215] [Medline: 33908890]
64. Alizadehsaravi L, Koster RA, Muijres W, Maas H, Bruijn SM, van Dieën JH. The underlying mechanisms of improved balance after one and ten sessions of balance training in older adults. Hum Mov Sci 2022;81:102910 [FREE Full text] [doi: 10.1016/j.humov.2021.102910] [Medline: 34864610]
65. McCrum C, Bhatt TS, Gerards MHG, Karamanidis K, Rogers MW, Lord SR, et al. Perturbation-based balance training: principles, mechanisms and implementation in clinical practice. Front Sports Act Living 2022;4:1015394 [FREE Full text] [doi: 10.3389/fspor.2022.1015394] [Medline: 36275443]
66. Pai Y, Yang F, Bhatt T, Wang E. Learning from laboratory-induced falling: long-term motor retention among older adults. Age (Dordr) 2014;36(3):9640 [FREE Full text] [doi: 10.1007/s11357-014-9640-5] [Medline: 24668268]
67. Bhatt T, Yang F, Pai Y. Learning to resist gait-slip falls: long-term retention in community-dwelling older adults. Arch Phys Med Rehabil 2012;93(4):557-564. [doi: 10.1016/j.apmr.2011.10.027]
68. McCrum C, Vaes AW, Delbressine JM, Koopman M, Liu W, Willems P, et al. A pilot study on the feasibility and effectiveness of treadmill-based perturbations for assessing and improving walking stability in chronic obstructive pulmonary disease. Clin Biomech (Bristol) 2022;91:105538 [FREE Full text] [doi: 10.1016/j.clinbiomech.2021.105538] [Medline: 34823220]

XSL•FO
RenderX

69. Liu H, Yin H, Yi Y, Liu C, Li C. Effects of different rehabilitation training on balance function in stroke patients: a systematic review and network meta-analysis. Arch Med Sci 2023;19(6):1671-1683 [FREE Full text] [doi: 10.5114/aoms/167385] [Medline: 38058731]
70. Molina-Garcia P, Mora-Traverso M, Prieto-Moreno R, Díaz-Vásquez A, Antony B, Ariza-Vega P. Effectiveness and cost-effectiveness of telerehabilitation for musculoskeletal disorders: a systematic review and meta-analysis. Ann Phys Rehabil Med 2024;67(1):101791 [FREE Full text] [doi: 10.1016/j.rehab.2023.101791] [Medline: 38128150]
71. Peeters CM, Visser E, Van de Ree CL, Gosens T, Den Oudsten BL, De Vries J. Quality of life after hip fracture in the elderly: a systematic literature review. Injury 2016;47(7):1369-1382. [doi: 10.1016/j.injury.2016.04.018]
72. Soukkio PK, Suikkanen SA, Sintonen H, Kukkonen-Harjula KT, Kautiainen H, Kääriä SM, et al. Health-related quality of life after hip fracture: effects of a 12-month home-based exercise intervention-secondary analyses of an RCT. Qual Life Res 2024;33(2):541-550. [doi: 10.1007/s11136-023-03545-4] [Medline: 37932555]
73. Gilboa Y, Maeir T, Weber A, Maeir A, Rotenberg S. Predictors of community reintegration and quality of life after hip fracture among community-dwelling older adults. Int J Rehabil Res 2019;42(3):234-239. [doi: 10.1097/MRR.0000000000000355] [Medline: 31394559]
74. Xiao H, Zeng W, Lu L, Yuan J, Yan Z, Wang J. Effectiveness of telerehabilitation in postoperative outcomes in patients on hip fracture surgery: a meta-analysis of randomized controlled trials. BMC Sports Sci Med Rehabil 2025;17(1):130 [FREE Full text] [doi: 10.1186/s13102-025-01169-4] [Medline: 40420240]
75. Marmor MT, Grimm B, Hanflik AM, Richter PH, Sivananthan S, Yarboro SR, et al. Use of wearable technology to measure activity in orthopaedic trauma patients: a systematic review. Indian J Orthop 2022;56(7):1112-1122 [FREE Full text] [doi: 10.1007/s43465-022-00629-0] [Medline: 35813536]
76. Choi S, Lee J, Bae KJ, Kim DK, Lee YH, Lee Y. Clinical effect of rehabilitation after distal radius fracture surgery using a wearable device: a comparative prospective cohort study. Hand Surg Rehabil 2024;43(6):101779. [doi: 10.1016/j.hansur.2024.101779] [Medline: 39349296]
77. Nuzzo JL, Pinto MD, Kirk BJC, Nosaka K. Resistance exercise minimal dose strategies for increasing muscle strength in the general population: an overview. Sports Med 2024;54(5):1139-1162 [FREE Full text] [doi: 10.1007/s40279-024-02009-0] [Medline: 38509414]
78. Spiering B, Mujika I, Sharp MA, Foulis SA. Maintaining physical performance: The minimal dose of exercise needed to preserve endurance and strength over time. J Strength Cond Res 2021;35(5):1449-1458. [doi: 10.1519/JSC.0000000000003964] [Medline: 33629972]

Abbreviations

DHI: digital health intervention
FBG: Functional Balance for Geriatric Patients
FIM: Functional Independence Measure
GRADE: Grading of Recommendations, Assessment, Development and Evaluation
HHS: Harris Hip Score
HKSJ: Hartung-Knapp-Sidik-Jonkman
HOOS: Hip Disability and Osteoarthritis Outcome Score
MFAC: Modified Functional Ambulatory Category
mHealth: mobile health
MOS: Medical Outcomes Study
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
QoL: quality of life
RCT: randomized controlled trial
SF-12: Short Form-12 Health Survey
SF-36: MOS 36-Item Short Form Health Survey
SMD: standardized mean difference
SPPB: Short Physical Performance Battery
TUG: Timed Up and Go Test
WHO: World Health Organization

XSL•FO
RenderX

Edited by S Brini; submitted 01.Jul.2025; peer-reviewed by F Ye, J Zhang; comments to author 30.Oct.2025; accepted 05.Jan.2026; published 12.Mar.2026.

Please cite as:
Fan W, Zhang Q, Lu Q
Effects of Digital Health Interventions on Functional and Psychological Outcomes in Older Patients With Hip Fractures: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e79563
URL: https://www.jmir.org/2026/1/e79563
doi:10.2196/79563
PMID:

©Wei Fan, Qi Zhang, Qunfeng Lu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Use of Health and Welfare Technology in Palliative Care: State-of-the-Art Review

Viktoria Zander[1*]; Maja Holm[2,3*]; Monir Mazaheri[2,4*]; Christine Gustafsson[1,2*]; Sara Landerdahl Stridsberg[5*]; Ragnhild Hedman[2,6*]

[1]School of Health, Care and Social Welfare, Mälardalen University, Eskilstuna, Sweden

[2]Department of Nursing Science, Sophiahemmet University College, Stockholm, Stockholm, Sweden

[3]Department of Health Care Sciences, Palliative Research Centre, Marie Cederschiöld University, Stockholm, Sweden

[4]Department of Neurobiology Care Sciences and Society, Nursing Division, Karolinska Institutet, Stockholm, Sweden

[5]University Library, Mälardalen University, Västerås, Sweden

[6]Stockholm Gerontology Research Center, Stockholm, Sweden

[*]all authors contributed equally

Corresponding Author:
Viktoria Zander
School of Health, Care and Social Welfare
Mälardalen University
SE-Box 325
Eskilstuna, 631 05
Sweden
Phone: 46 0736620515
Email: viktoria.zander@mdu.se

Abstract

Background: As more individuals live longer with complex conditions, the need for effective palliative care (PC) grows. It has been stated that access to PC should be integrated early and delivered in a timely manner to patients with life-threatening illnesses. Health and welfare technologies (HWTs) offer tools to enhance care delivery, particularly in home and rural settings. Although there is a profound lack of evidence regarding the impact when used in PC, it is necessary to critically assess the current state of knowledge regarding impacts and consequences of technologies, ensuring that their integration considers broader implications for patients, caregivers, and health care systems in PC.

Objective: This review explores health and welfare technology used in PC, aiming to inform practice and improve care quality.

Methods: This state-of-the-art review included empirical studies describing the use of HWT in PC for adult patients. We used a thematic synthesis approach to compare studies and provide a synthesis of the key points.

Results: Based on the inclusion criteria, 94 studies were included. PC is both a clinical specialty and an overall approach to care that focuses on improving quality of life and relieving suffering for patients and families facing serious illness, based on needs and not prognosis. HWT shows potential to increase access and continuity of care, for symptom management to support patients to remain at home and prevent frequent emergency visits. It can have the potential to build and remain relationships between patients, their families, and the health care team, as well as for interprofessional collaboration and support. However, there are challenges to overcome that might affect the quality of care when using technology.

Conclusions: HWT shows potential as a complement to usual PC. Our findings point toward the importance of caution in choosing when to use HWT in PC, and for which patients.

(J Med Internet Res 2026;28:e79637) doi:10.2196/79637

KEYWORDS

digital technology; health and welfare technology; palliative care; state-of-the-art review; systematic review

XSL•FO
RenderX

Introduction

As more individuals live longer with complex conditions, the need for effective palliative care (PC) grows. It has been stated that access to PC should be integrated early and delivered in a timely manner to patients with life-threatening illnesses. However, research has demonstrated that health care systems fail in this regard. Home-based PC is gaining attention, reflecting patient preferences to remain at home and potential cost savings from reduced hospitalizations. Health and welfare technologies (HWTs) offer tools to enhance care delivery, particularly in home and rural settings. However, questions about their feasibility and acceptability remain.

Today, an increasing number of individuals are living to older age with complex conditions and a great need for PC, including symptom relief and family support [1,2]. However, reports show that less than 15 % of patients requiring PC receive it in a timely and accurate manner [3]. PC is a holistic approach to the care of people with life-limiting conditions and near the end of life. It aims to acknowledge and attend to all aspects of the patients' and family members' needs when they are facing severe illness, including physiological, psychological, existential, and social issues. Symptom management, support of family members, and interprofessional teamwork are emphasized as crucial elements [3]. In recent decades, PC has evolved from primarily being offered to patients with incurable cancer, to currently being recommended to all patients with chronic and life-limiting illness [4]. Nonspecialized PC can be delivered by all care professionals, whereas specialized PC is provided in hospices, hospital PC units, and by specialized PC teams in home care [3]. Further, it has been highlighted that PC should be delivered early and integrated with specific illness treatment [5]. However, it has been demonstrated that access to PC varies and that patients with cancer still receive it more often than patients with other diagnoses [6,7].

Lately, there has been an increased focus on home-based care, especially since many patients wish to be cared for and die in their homes [8,9]. Home-based PC services offer many benefits, such as a sense of normalcy, choice, and comfort [10]. The prospect of dying at home is regarded as a more comfortable and dignified experience than dying in a hospital [10]. However, challenges exist in providing an optimal service. Unmet needs, uncoordinated care, and insufficient communication with health care professionals [11,12], as well as the demanding collaboration between specialists and home care professionals, make this challenging [13]. Moreover, home-based PC relies on the contribution of family caregivers [14], who often find themselves in a situation of managing multiple responsibilities and often forget their own needs to attend to those of the patient [10]. This informal care work often goes unnoticed and unaddressed [10,15]. From a societal point of view, home care has been associated with lower costs, as repeated hospital admissions are a major driver of expenditure in PC [16]. Although, the evidence for this is uncertain [17].

The COVID-19 pandemic has accelerated the demand for technologies in health care [18]. HWT comprises technology-based interventions that aim to maintain or promote health, well-being, quality of life, and/or increase efficiency in the operational delivery of welfare, social, and health care services, while improving working conditions for staff [19]. There are high expectations for HWT as solutions to challenges such as aging populations and limited resources [20-22]. HWT is expected to enhance the delivery and accessibility of PC, particularly in home settings [23] and in rural areas [24]. For instance, videoconferencing systems enable the remote delivery of multispecialty care, and artificial intelligence–driven wearable and nonwearable technologies facilitate remote assessments of patients' conditions in their home environments, leading to more comprehensive clinical evaluations and empowering patients to monitor their own health [25]. Additionally, digital care plans can streamline home care delivery, potentially reducing avoidable hospitalizations [26].

Previous research has sought to gather evidence on the use of HWT in PC with a focus on telehealth (describing the provision of health care remotely by means of a variety of telecommunication tools and video consultations) [27-32], telemedicine (the use of remote technology and telecommunications) [33,34], eHealth (the use of information and communication technology for health care provision) [35], or with a focus on video consultations [36]. There is also specific focus on populations or settings, such as older adults [23], the professional perspective [30], or the patient perspective [29,35], home-based PC [23,32,37], or PC in rural areas [24,27,31,33]. However, there is a lack of literature reviews that have included the overall variety of HWT for all patients in PC and focused on impacts on patients as well as informal and formal caregivers.

Overall, HWT is frequently met with expectations, often based on biases, such as optimistic assumptions about what it could achieve and the belief that a technology will be as good as or better than a human, regardless of the task [38]. While this optimism emphasizes the potential for HWT, it often overlooks the complexities and challenges of implementing technology in sensitive areas such as PC [39]. There is also a profound lack of evidence of benefits and harms, as well as of impacts, when HWT is used in PC [40]. Other than telehealth, videoconferencing, or after-hours telephone support, there is little evidence for HWT used in PC [37]. Moreover, it is still unclear whether PC delivered remotely or with support of HWT is equivalent to more resource-intensive methods of in-person care [41]. Therefore, it is necessary to critically assess the current state of knowledge regarding the use of technologies for patients, caregivers, and health care systems in PC. This review explores HWT used in PC, aiming to inform practice and improve care quality.

This systematic review responds to the following research questions: (1) Which HWT is used in PC? (2) What impact does the use of HWT in PC have on patients and informal and formal caregivers? (3) What knowledge gaps and research needs are identified related to the use of HWT in PC? (The last research question is reported elsewhere).

XSL•FO
RenderX

Methods

Study Design

We conducted a state-of-the-art review. This type of review is appropriate to shape a comprehensive understanding of the current state of knowledge in a specific area [42]. The review followed PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [43]. The PRISMA-ScR checklist for this review is presented in Multimedia Appendix 1. A protocol was prospectively registered in the Open Science Framework on October 10, 2024 [44].

Eligibility Criteria

Eligibility criteria were formulated in dialogue within the research group, based on the aim of the review. In line with Radbruch et al [4], PC was defined as holistic care of patients with severe illness, especially near the end of life, regardless of diagnosis and care place. Studies involving patients receiving curative treatment, or with the possibility of receiving curative treatment such as, kidney transplantation, were excluded.

Studies focusing on technology-based interventions for safety, activity, participation, and independence and/or increased efficiency in PC, and working conditions for professional caregivers were included. Further inclusion criteria were empirical studies (qualitative, quantitative, mixed methods, and case studies) describing the use of HWT in PC of adult patients (aged ≥18 years). The search was limited to studies published after 2012.

Studies that used technologies only for data collection were excluded, as were studies not reporting technology use in the intended setting. Study protocols, reviews, and studies describing the development of technologies without involvement of patients, informal caregivers, or health care professionals were also excluded, as were studies in languages other than English and studies that could not be obtained in full text.

Information Sources

The following electronic databases were searched: PubMed, APA PsycINFO, Cochrane Library, CINAHL Plus, Scopus, and Web of Science Core Collection.

Searches were conducted on October 27, 2022. Additional update searches were completed on November 23, 2023.

Search Strategy

Searches were conducted by an academic librarian (SLS). The search terms were organized to match the review's goals, covering concepts such as PC, digital health and welfare technology, and home monitoring. Free-text words were searched in article titles and abstracts, along with database-specific subject headings, like MeSH (Medical Subject Headings) terms, when relevant. The search targeted articles published after 2012, excluding topics related to neonatal and pediatric care using the Boolean operator NOT to avoid content on children's PC. Textbox 1 contains an example of search terms. The complete search strategies for all databases are available in Multimedia Appendix 2.

Textbox 1. Search strategy PubMed.

PubMed search strategy:

- "Palliative Care"[MeSH Terms] OR "Palliative Medicine"[MeSH Terms] OR "Hospice and Palliative Care Nursing"[MeSH Terms] OR "Terminally ill"[MeSH Terms] OR "Terminal Care"[MeSH Major Topic] OR "Hospice Care"[MeSH Terms]
- "palliati*"[Title/Abstract] OR "hospice care"[Title/Abstract] OR "hospice nursing"[Title/Abstract] OR "terminal care"[Title/Abstract] OR "supportive care"[Title/Abstract] OR "terminal stage"[Title/Abstract] OR "terminal disease"[Title/Abstract] OR "terminally ill"[Title/Abstract] OR "end stage"[Title/Abstract] OR "end of life"[Title/Abstract] OR "hospice program*"[Title/Abstract] OR "advanced illness"[Title/Abstract]
- "eHealth"[Title/Abstract] OR "e-health"[Title/Abstract] OR "telemedicine"[Title/Abstract] OR "telehealth"[Title/Abstract] OR "mhealth"[Title/Abstract] OR "m-health"[Title/Abstract] OR "mobile health"[Title/Abstract] OR "e-homecare"[Title/Abstract] OR "digital health"[Title/Abstract]
- "Telemedicine"[MeSH Terms]
- "welfare technolog*"[Title/Abstract] OR "ambient assisted living*"[Title/Abstract] OR "ambient intelligence*"[Title/Abstract]
- "Ambient Intelligence"[MeSH Terms]
- "home monitoring"[Title/Abstract] OR "distance monitoring"[Title/Abstract] OR "distance care"[Title/Abstract] OR "distance nursing"[Title/Abstract] OR "distance medicine"[Title/Abstract] OR "environmental control*"[Title/Abstract] OR "remote sensing"[Title/Abstract] OR "Distance Counseling"[Title/Abstract] OR "Internet-Based Intervention"[Title/Abstract] OR "ambulatory monitoring"[Title/Abstract] OR "remote consultation"[Title/Abstract] OR "telecommunication*"[Title/Abstract] OR "E-Counseling"[Title/Abstract] OR "e therapy"[Title/Abstract] OR "distance spanning"[Title/Abstract] OR "health informatics"[Title/Abstract] OR "health information technology"[Title/Abstract] OR "medical information science"[Title/Abstract]
- "Remote Sensing Technology"[MeSH Terms] OR "Internet-Based Intervention"[MeSH Terms] OR "monitoring, ambulatory"[MeSH Terms] OR "Remote Consultation"[MeSH Terms] OR "Telecommunications"[MeSH Terms]

Selection of Sources of Evidence

After the search, references were uploaded to Covidence, a web-based collaboration software for literature reviews [45]. Following the automatic removal of duplicates by Covidence and the manual removal of duplicates by the reviewers, a total of 3662 studies were examined for eligibility. Five of the authors (CG, VZ, MH, MM, and RH) conducted the abstract and full

XSL•FO
RenderX

text screening. To start with, 40 abstracts were screened by 2 or 3 of the authors in various constellations, and their decisions to include or exclude were compared and discussed to ensure consistency. Subsequently, the abstract screening was equally divided among the authors. After the title and abstract screening, the remaining 601 articles were read in full and assessed for eligibility. A subset of 10 articles was initially read by all 5 authors separately and then compared. The remaining studies were divided among the authors and discussed when needed. All articles that matched the inclusion criteria were included in the review. In accordance with established practice for state-of-the-art reviews, no formal quality assessment of the included studies was undertaken [42].

The screening process is described in a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. HWT: health and welfare technology; PC: palliative care.

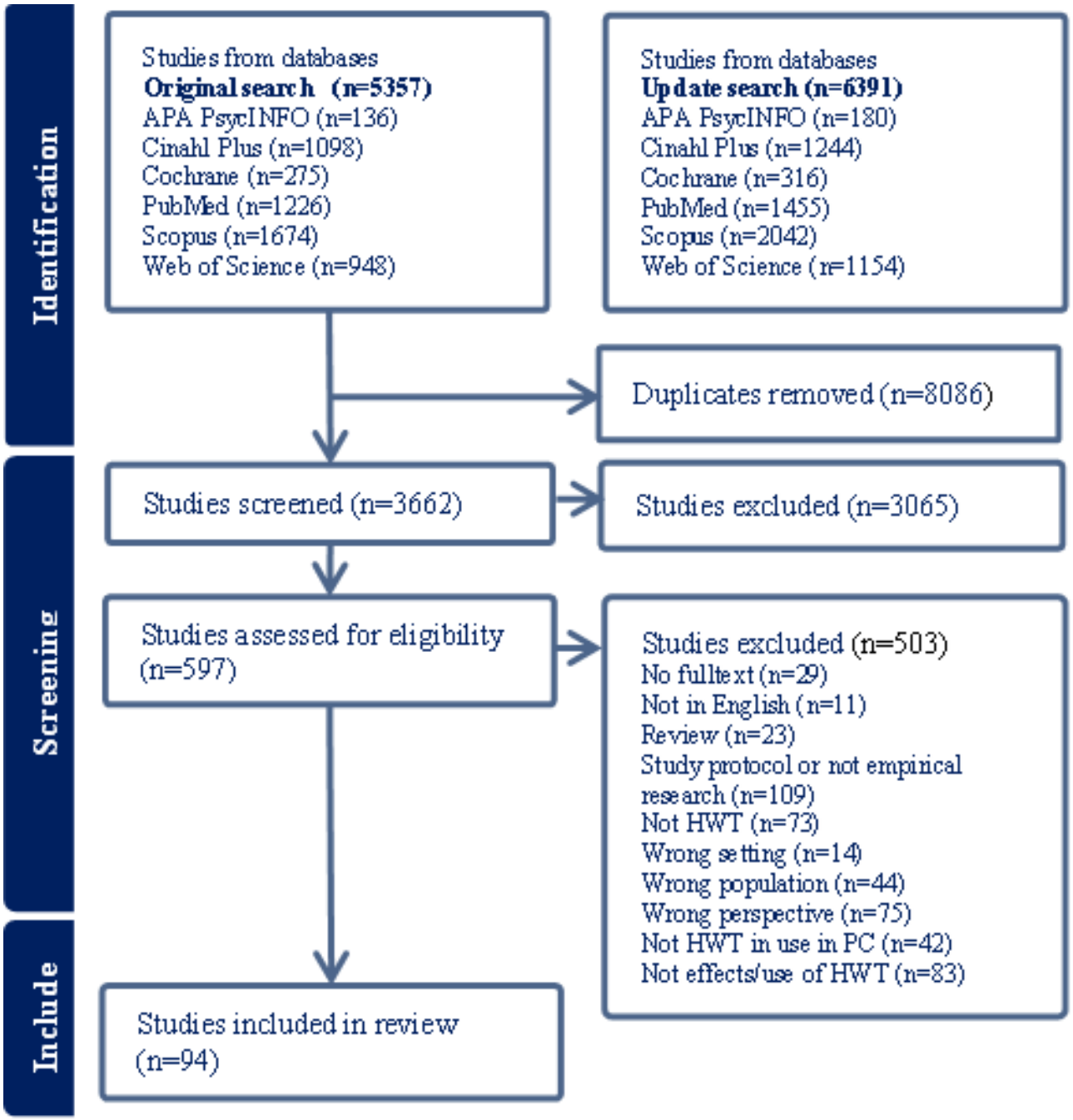

Data Charting Process

After reviewing eligible studies, a data extraction template was generated to guide the data extraction process, including citation, study location, study aim and design, sample characteristics, intended target group, PC description, characteristic of technology, impacts of technology use in PC, knowledge gaps and research needs, and conclusions (knowledge gaps and research needs have been reported elsewhere).

XSL•FO
RenderX

Synthesis of Results

Study characteristics were entered into an Excel (Microsoft Corp) spreadsheet. A thematic synthesis approach, led by VZ, supported by RH and MH, and discussed within the author group, was used to compare the studies and provide a synthesis of the key points. The process of the synthesis was guided by the first 2 steps described by Thomas and Harden [46]. These steps were (1) coding of the text—identification of themes across the included studies. Each study's findings were identified and put into a metric. The findings were then coded line by line. Each sentence was read inductively to identify underpinning themes or concepts, which were labeled with a code; (2) developing descriptive themes—the descriptive codes were compared and organized into descriptive themes. Descriptive themes across the different categories of technology were compared and organized into one overall synthesis of the use of HWT in PC.

Results

Description of Included Studies

The selection process resulted in the inclusion of 94 articles describing the use of HWT in PC [47-140]. The included studies were published between 2012 and 2023. They were conducted in North America (Canada and United States; n=44), South America (Brazil and Chile; n=3), Australia (n=14), Africa (Kenya, sub-Saharan Africa, and Tanzania; n=4), Asia (China, India, Indonesia, and Taiwan; n=5), and Europe (Austria, Belgium, Finland, Georgia, Germany, Netherlands, Norway, Italy, Sweden, and United Kingdom; n=26). Some were cross-national. The review included studies using quantitative (n=50), qualitative (n=19), as well as mixed methods (n=25). The sample size of the studies ranged from 1 to 3178. The most common target group was patients, who were involved in 64 studies, followed by informal caregivers (involved in 31 studies) and health care professionals (involved in 21 studies). Several studies involved more than one target group. The most common diagnoses were cancer (n=38), followed by a few articles focusing on a variety of illnesses such as heart failure (n=2), chronic obstructive pulmonary disease (COPD, n=1), cirrhosis (n=1), dementia (n=1), and cognitive impairments in amyotrophic lateral sclerosis (n=1). Besides diagnoses, articles focused on older adults suffering from chronic illnesses (n=3), terminal illness and the end-of-life phase (n=12), and palliative care regardless of diagnosis (n=24). The characteristics of the studies included are presented in Multimedia Appendix 3.

Health and Welfare Technology in Palliative Care

The HWT in the included studies were categorized based on function (Table 1). In some of the studies, HWT from more than 1 category were included, such as video consultations based on symptom monitoring.

Table 1. Categorization of health and welfare technology (HWT) in included studies.

Category of HWT	Examples of type of HWT
Symptom monitoring (n=25)	• Systems to alert clinicians of changes in patient's symptoms • Symptom reporting systems and applications • Systems for wireless data transfer • Wearable monitoring devices
Telehealth consultations and conferences (n=51)	• Videoconferencing • Teleconsultations • Telehealth-delivered care
Sharing of patient information (n=2)	• Electronic care coordination systems • Needs rounds via telehealth
Remote therapy and treatment interventions (n=7)	• VR[a] headset • Therapeutic interventions in mobile apps • Web-based therapeutic interventions • Therapy delivered via email
Education and support (n=27)	• Mobile apps and web-based education platforms for patients, caregivers, or health care personnel • Teleconferencing technology to support and train health care personnel remotely • Virtual webinar sessions • Facebook support groups

[a]VR: virtual reality.

XSL•FO
RenderX

The Use of Health and Welfare Technology in Palliative Care

Overview

The included studies described the use of HWT in PC on an individual level (including patients, informal caregivers, and health care professionals) and on an organizational level. The synthesis resulted in 7 themes, which are presented in the following subsections (Table 2 provides an overview).

Table 2. Descriptive themes and codes.

Descriptive themes	Codes
Symptom control and disease progression (n=26)	• Impact on disease progression • Impact on physical and mental symptoms • Management of symptoms • Disease-related stress
Quality of life and death (n=10)	• Patients' and caregivers' quality of life • Support in the end-of-life phase and death
Competence and palliative literacy (n=9)	• Knowledge and understanding • Self-efficacy • Insight into norms and values
Palliative care provision (n=44)	• Care satisfaction • Medication use and compliance with therapies • Prevention of excessive care • Care coordination • Accelerating advanced care planning • Access to care
Multidimensional care relationships (n=18)	• Partnership between patients and the health care team • Companionship with and among significant others • Interprofessional collaboration
Facilitating a comprehensive support system (n=33)	• Professionals' understanding • Patient empowerment • Support for the caring role
Organizational outcome (n=8)	• Cost-effectiveness • Time and travel • Job satisfaction and distress

Symptom Control and Disease Progression

This theme focuses on the potential of HWT to support and prevent disease progression, physical and mental symptoms, and to relieve disease-related stress [47-72]. Types of HWT involved in symptom control and disease progression were technology for symptom monitoring [47,50,52-55,60,64,65,69,71], telehealth consultations and conferences [47,48,52,55-57,60,61,63,64,66-68,70], remote therapy and treatment interventions [58,59,62,72], and technology for education and support [49,51,56].

Among the included studies, the potential of HWT to support and prevent disease progression was mainly reported and discussed by 2 studies. Kazankov et al [47] described how a digital health solution with monitoring of heart rate, blood pressure, weight, body water, and cognitive function, for individuals with cirrhosis could facilitate timely intervention to prevent disease progression. Jiang et al [48] described how a model with a communication platform resulted in less functional decline. A total of 3 studies explored the effects of HWT on mortality, none of which could show any significant effects [49-51].

The impact of HWT on physical and mental symptoms was described by several studies. Positive impact on physical symptoms related to cancer illness and treatment was described by Tumeh et al [52], Mooney et al [53], Besse et al [54], and Cornetta et al [55]. Ngoma et al [50] reported no differences in overall symptoms when using remote symptom assessment and care coordination, although there were effects on symptom severity. Mark et al [56] reported a positive impact on dyspnea with an online teaching intervention for persons with COPD. Rafter [57] showed that an eHealth system allowed hospice professionals to deliver more effective care, with improvement in pressure ulcers for end-of-life patients. Impact on mental symptoms, such as anxiety and depression, was reported in 7 of the included studies. Positive effects were reported in cancer care studies using symptom monitoring [52], a virtual reality headset [58], a mindfulness intervention [59], telemedicine [60], videoconferencing [61], and web-based psychotherapy for informal caregivers [62]. On the other hand, one study using teleconsultations showed no differences in depression scores and increased anxiety scores [63]. Eight of the included studies showed the potential of HWT to contribute support in symptom management [64-71].

XSL•FO
RenderX

HWT was also used to relieve disease-related stress. For example, studies showed that an online support system in the care of patients with cancer has the potential to reduce symptom distress [49], mindfulness via video could reduce cancer-related stress [59], telehealth music therapy had positive affective effects [72], and teleconsultations resulted in lower distress symptom scores [68]. Maguire et al [69] showed that using a remote symptom monitoring system provided reassurance about symptom experience and the feeling of being listened to. However, Hoek et al [63], on the other hand, reported that adding weekly teleconsultations to usual PC led to a higher total distress score among home-dwelling patients with advanced cancer.

Quality of Life and Death

This theme describes patients' and caregivers' quality of life [52,56,60,69,73,74] and support in end-of-life phase and death [75-78]. It involves HWT for symptom monitoring [52,60,69], telehealth consultations and conferences [52,56,60,74,75,77,78], sharing of patient information [76], remote therapy and treatment interventions [73], and education and support [56,74].

Five of the included studies showed some effects on patients' and informal caregivers' quality of life with the use of HWT for symptom monitoring [52,60], training delivered over Skype for patients with COPD [56], a supportive care mobile app intervention [74], and a cognitive-behavioral therapy mobile app for anxiety [73]. Maguire et al [69] found that a system with daily symptom reports provided reassurance, but it did not lead to changes in quality of life.

Outcomes regarding support in the end-of-life phase and death were reported in 4 of the included studies. The potential to support dying at home or in another community setting was reported using an electronic PC coordination system [76], home care delivery with point-of-care technology and remotely located health care professionals [78], and a PC after-hours telephone number [75]. Johnston et al [77] reported the potential of Skype calls with dying persons and family members to bring closure and reconciliation, inclusion in the dying process, and healthy grieving.

Competence and Palliative Literacy

Competence and palliative literacy refer to knowledge and understanding, self-efficacy, and insight into norms and values [74,75,79-85]. This theme includes HWT for symptom monitoring [80,81], telehealth consultations and conferences [74,83], remote therapy and treatment interventions [85], and education and support [74,75,79,82-84].

Different HWT interventions have been shown to increase knowledge and understanding of health, diagnosis, symptoms, and concerns among patients [74,79], informal caregivers [80,81], and health care professionals [82,83]. An e-learning intervention for nurses to increase knowledge of PC and attitudes toward dying patients and death showed positive effects [82]. Teleconferencing technology to support and train hospice nurses remotely was perceived to improve knowledge as well as self-efficacy in caring of patients in PC [83]. The intervention had given them access to education that otherwise would have been difficult to obtain due to geography. Self-efficacy to manage their own condition was reported by patients with advanced cancer using a digital support app [75]. Digital information and education were also shown useful for addressing insights into one's own thoughts and values related to advanced dementia among patients [84] and related to caring for a person with advanced cancer among informal caregivers, including accepting experiences of negative thoughts and feelings and being more aware of personal values [85].

Palliative Care Provision

PC provision includes subthemes such as care satisfaction [50,87], medication use and compliance with therapies [72,73,81,82,87,88], prevention of excessive care [48,51,64,65,68,69,76,86,89-98], care coordination [75,83,94,102-110], accelerating advanced care planning [51,84,87,93,94,97,99-101], and access to care [68,74,96,106,108,111-117]. This was mainly related to use of technology for telehealth consultations and conferences [48,64,68,74,75,83,84,87,88,91-96,98,100,102-104,106,108,109,111-116], but also technology for symptom monitoring [50,64,65,69,81,89,96,98,106,107,110,117], education and support [51,74,82-84,86,90,97,99,101,117], sharing of patient information [76,100], and remote therapy and treatment interventions [72,105].

The use of HWT was shown to be promising in the provision of care, including care satisfaction [50,87] and medication use and compliance with therapies [73,82,87,88]. Zeiser et al [72] reported that music therapy telehealth services increased compliance with other therapies. Other studies reported decreased polypharmacy and potentially decreased adverse drug events [88], but no effects on analgesic adherence [82] and no differences in anticipatory medication prescribing [87].

Some studies reported positive impacts on care coordination [102], care provision [83], increased efficiency [103-108], increased care resource use [75], and positive outcomes related to quality of care [94,106]. For example, Groothuizen et al [109] reported that virtual team meetings resulted in increased flexibility, reduced travel time, and easier real-time access to patient information for health care professionals. Health care professionals in the study by Oelschlägel et al [110], on the other hand, found that organizational challenges made it difficult to obtain and share the information necessary to provide seamless and optimal service to patients.

Prevention of excessive care is a core purpose of technology in PC and relates to the use of hospice care, emergency care, and hospitalization. Among the included studies, no effects were shown on hospice enrollment or length of stay [51], nor on time until entry into hospice [98]. Emergency department visits and admission could be avoided or decreased in several studies [65,69,76,86,89,90,92]. The HWT used in these studies varied. Technology to support patients (symptom monitoring and means to communicate needs to health care professionals) showed positive effects [65,69,89,92], as did technology used to support health care professionals in their work [76,86,90]. Manz et al [51] showed that opt-out text messages to prompt serious illness conversations decreased end-of-life systemic therapy relative to controls, but there was no effect on hospice enrollment or length of stay, inpatient death, or end-of-life intensive care unit

XSL•FO
RenderX

use. Nor did a teleconsultation service with a triage system change the number of emergency visits [91]. HWT to support patients in communication with health care professionals [48,91,94,96] showed potential to decrease the need for hospital referrals. Unnecessary hospital transfers could also be prevented by improved collaboration among physicians using telemedicine [95]. Moreover, an advance care planning video program intervention increased documented "Do Not Hospitalize" orders among nursing home residents with advanced illness, but did not significantly reduce hospitalizations [97].

Technology was also used to accelerate advance care planning, with promising results in 7 studies [51,84,94,97,99-101], but no results in 2 studies [87,93].

Other positive effects reported were increased access to care [69,96,106,108,111-114] and increased continuity of care [113,115,116]. HWT was shown to enable health care professionals to respond quickly to patients' care needs [96,106], allowing them to reach more patients [68,111], especially during epidemics or when out of town. For patients and informal caregivers, the technology increased the sense of security related to the ability to contact the clinic when needed [96,117].

Multidimensional Care Relationships

Multidimensional care relationships concern the ability to use HWT to form partnership between patients and health care teams, companionship with significant others, and interprofessional collaboration [65,69,77,96,104,110,114,116,118-127]. HWT used in the studies related to this theme included technology for symptom monitoring [65,69,96,110,119,120], telehealth consultations and conferences [77,96,104,110,114,116,118,119,122,127], and education and support [121,123-126].

Whether remote care can affect the partnership between patients and health care teams varies between studies and within studies. Rosa et al [118] described positive as well as negative impacts on the quality of relationships with patients, families, and between health care professionals. For health care professionals, increased availability to patients and their caregivers was positive, but there was a perceived loss of nonverbal cues, which made communication more difficult. Some health care professionals also experienced difficulties being supportive remotely during difficult times. It also put a strain on relationships among coworkers. Other studies reported no negative effects on relationships [114], enhanced connectivity with the care team [70,96], and contributions of unique insight into the daily lives of patients [116]. Over time, care delivered with support of HWT, such as telehealth consultations, can result in trustful relationships [116]. The introduction of technologies has the potential to alter the dynamic of relationships between patients, families, and community PC clinicians, serving as a means to complement in-person care [119]. Among the included studies, there were also reports of positive impacts on patient–caregiver communication regarding symptom management [66] and perceptions that it was easier to discuss psychological and care needs remotely [120].

Moreover, HWT can be used to facilitate companionship with and among significant others and informal caregivers [67,78,104,121-125]. The use of technology, such as video calls, was shown to be able to connect patients and family members at the end of life [78]. Another area of use is online support groups for informal caregivers, with the aim of forming companionship between peers [121].

Other studies have shown the ability of HWT to increase professional collaboration [126,127], offering an understanding of each other's professions [126].

Facilitating a Comprehensive Support System

This theme focuses on the ability to use HWT to support professionals' understanding, patient empowerment, and support for the caring role among informal caregivers. Telehealth consultation and conference technology [56,57,67,91,96,102,103,110-112,114,117,118,127,129,133] and technology for education and support [56,86,125,126,128,130-135,137-139] were the most common HWT, followed by technology for symptom monitoring [96,110,117,138,139] and remote therapy and treatment interventions [62].

Among the included studies, HWT was used to support professionals' understanding. Oelschlägel et al [110] described how remote home care helped municipal health care professionals shift their perspective toward patients' priorities. In the study by Rosa et al [118], remote care used as an alternative to in-person care during the COVID-19 pandemic led to changes in perceived self-efficacy among health care professionals in managing their job responsibilities. HWT, such as videoconferences, can also be used by senior professionals to support and mentor the junior workforce [86].

According to the included studies, HWT can play a role in empowering patients, for example by increasing involvement and enabling patients to take an active part in their care [96,117], feel in control of treatment [117], and manage everyday life [110]. HWT was useful in preparing patients before care visits [128], providing support [91,102,112,129], facilitating comfort, safety, and independence for patients [103,114], and increasing patient satisfaction [68].

Technologies were also used to support informal caregivers. Examples of technologies addressing their needs included interventions for mental and emotional support [63,130,131]. These were shown to have potential for preparing individuals for the caring role [86] and addressing unmet needs [63]. Digital technology interventions were used to support informal caregivers in managing everyday challenges and stress [57,132-135]. For example, an education intervention delivered through a smartphone app was effective in improving family readiness and quality of life among family members [135]. However, other studies were not able to show this potential. According to Dionne-Odom et al [130] a telehealth intervention to educate and support informal caregivers of patients with heart failure did not provide any significant effects on quality of life, mood, or caregiver burden.

Other technologies, which primarily focus on patient's needs, may also be used to support the caregiver role, with increased caregiver involvement [136], increased connection between caregivers and patients [125], reduced caregiver burden

XSL•FO
RenderX

[64,137], and effects such as less negative mood and emotional distress [138], as well as reduced loneliness [134]. Access to hospice personnel provided by HWT might be perceived as comforting [68]. Among the included studies, one study was not able to show any significant effects on caregiver burden, self-efficacy, or quality of life with the use of an eHealth self-management application for caregivers of patients with incurable cancer [139].

Organizational Outcomes

Organizational outcomes reported were related to cost-effectiveness [61,111], time and travel [96,108,109,114,140], and job satisfaction and distress [62,68,111,118]. The most common HWT related to the theme was telehealth consultations and conferences [61,62,68,96,108,109,111,114,118,140], although one study also involved HWT for symptom monitoring [96].

Two studies reported results related to cost-effectiveness. Compared with traditional in-person service, telehealth services resulted in equivalent costs but greater efficiency by allowing PC to reach more patients [111]. Video consultations with the PC team for rural patients were found feasible and resulted in travel and cost savings for patients [62]. The potential to minimize travel is clear, as reported by several studies. According to the case report by Morgan et al [96], telehealth-supported care was an effective adjunct to routine clinical care, Groothuizen [109] reported that virtual PC team meetings reduced travel time. Others reported that video-based consultations reduced the burden and expense of travel for patients, families, and consultants [108,114], and patient-reported satisfaction with telehealth in oncology was mainly attributed to advantages in travel and time savings [140]. Using telehealth services and eHealth systems increased job satisfaction due to the patient-centered nature of the care service, increased peer support, and increased professional development [111]; having 24/7 access to hospice triage personnel [68]; and effective care and good patient experiences [58]. On the other hand, a multidisciplinary PC team delivering telecare for hospitalized patients with cancer during the COVID-19 pandemic expressed distress related to competing loyalties (such as institutional obligations, ethical obligations to patients, resentment, and distrust of leadership) and feelings of disempowerment (due to guilt in providing subpar support, decisional regret, and loss of identity as a provider) [118].

Discussion

Principal Results

According to the findings, HWT has the potential to facilitate interventions to support symptom management. Different HWT interventions have shown usefulness to increase knowledge and understanding of health, diagnosis, symptoms, and concerns among patients, informal caregivers, and health care professionals. There were a variety among the reports regarding impact on quality of life among patients and caregivers; although, HWT can sometimes be useful to support dying persons and family members to bring closure and reconciliation. Related to PC provision, the use of HWT has shown promising results on care satisfaction and on medication use and compliance with therapies. Technology has also been used to accelerate advance care planning, with promising results and positive effects on the use of emergency and hospital care. One large benefit is the increased access to care using HWT, which means comfort and security for patients and families. HWT can be useful to include and involve significant others. On the other hand, research has shown both positive and negative impacts on the quality of partnership between patients, families, and health care teams when using HWT, suggesting situations when physical meetings are preferred and others when online meetings are suitable to build and maintain good quality care. From an organizational perspective, HWT has the potential to save time and costs due to decreased travelling by using digital means to consultations. The use of digital tools can increase job satisfaction but also contribute to perceived job strains and job distress.

Comparison With Prior Work

PC is a holistic approach to care for people with life-limiting conditions and near end of life. It aims to acknowledge and attend to all aspects of the patients' and family members' needs, including physiological, psychological, existential, and social issues [3]. Symptom management is identified as one of the critical areas in PC [141]. According to the current review, HWT can have a positive impact on physical and mental symptoms and support symptom management. Patients in PC suffer from advanced and chronic conditions, often with a negative progression. There was some evidence for the potential of using remote monitoring and care support to prevent disease progression or functional decline, at least for some time. Not very surprisingly, there were no studies presenting any significant effects on mortality.

For symptom management, timely and close contact with the health care team is essential [141]. It has long been recognized that access to and communication with the PC team are vitally important for patients as well as informal caregivers [142,143]. Based on the findings, HWT provides new ways of access and continuity of care, with improved opportunities to follow patients' needs without travelling. The support of patients to remain and continue to receive care at home [31,32] is especially important in cases when clinical conditions or geographic location prevent patients from accessing conventional care [33].

Although there is a concern that the partnership between patients, informal caregivers, and the PC team might be suffering. According to included studies, digital means of communication could be perceived as difficult due to loss of nonverbal cues and limited proximity to the patient when not being in the same room. Meeting online instead of in person can affect the possibilities for the health care professional to be supportive and comforting during difficult situations. Other reviews, on the other hand, have concluded that technology has the potential to not only facilitate communication through the inherent flexibility provided by technology [35] but also to build and enhance relationships [30,31]. These contradictory findings point toward the importance of caution in choosing when to use HWT in PC and for what patients.

Informal caregivers have an important role and take great responsibility for everyday symptom and medicine management

XSL•FO
RenderX

[143]. As having insight into the patient's well-being and needs, they are an important link between patient and the care team, relaying concerns, managing medication, and generally advocating for the patient when they are unable or unwilling to do so [143]. However, they also carry a heavy burden, while they must manage their own health and grief. It is, therefore, vital that they are adequately supported. Among the included studies, there was HWT addressing the caring role, aimed at supporting informal caregivers. Moreover, HWT focusing on the patients' needs were also shown to have an impact on the situation for informal caregivers. PC should take a holistic view of the patient and the informal caregiver, the concerns of both being intertwined and interdependent [143]. Based on the findings, there is research supporting the potential for HWT to complement the ordinary PC to increase support for the caregiving role.

Besides symptom management, access to the health care team, and support of informal caregivers, interprofessional teamwork is emphasized as a crucial element in PC [3]. From a professional perspective, the use of HWT means both possibilities and challenges. According to the included studies, the use of HWT may increase the possibility for professionals to reach the patients and their families when needed, for example, to follow progression and adjust treatment. On the other hand, according to some of the studies, this means yet another responsibility, which may contribute to increased job strain among health care professionals. Previous research suggests that HWT can improve the coordination of care and help build and enhance personal and professional relationships [30]. Other potential effects are time saving and reduction of no-show rates [27].

Limitations and Future Directions

The methodology of this study involved a comprehensive, state-of-the-art review of the existing literature on HWT in PC. With the purpose of enhancing reliability and contributing to the generalizability of findings, the search was conducted in a variety of electronic databases—APA, PsycINFO, Cochrane Library, CINAHL Plus, PubMed, Scopus, and Web of Science Core Collection. A supplementary search ensured up-to-date coverage and inclusion of recent research developments in the rapidly evolving field of HWT. In the search, several concepts were used to describe technology-based interventions in health care, often used interchangeably [144]. HWT is a common term in Nordic countries, but not as commonly used elsewhere. To ensure broad inclusion of relevant studies, broader terms like mHealth, telemedicine, and ambient intelligence were used, although specific technologies such as alarms or monitoring were avoided [145]. This approach may have limited our results to more general articles, excluding those focused on specific technologies.

Notably, among the included studies, there was a predominance of studies focusing on patients with cancer, often neglecting other palliative conditions. Furthermore, the limited diversity in participant demographics—with studies often focused on specific age groups, genders, or cultural contexts—highlights a gap in the representativeness of the findings. Future research should aim to include a wider range of diagnoses and settings to ensure that the findings are applicable to the broader PC population [36]. Moreover, there is a need for more rigorous research related to patient outcomes and evidence regarding effectiveness [29], health system outcomes (eg, usage and costs) [146], and best practices for quality remote PC [27]. New literature reviews should be conducted in a few years to update the state of the art.

Conclusions

PC is both a clinical specialty and an overall approach to care that focuses on improving quality of life and relieving suffering for patients and families facing serious illnesses, based on need and not prognosis. HWT shows potential as a complement to usual PC to increase access and continuity of care, for symptom management, to support patients to remain at home and prevent frequent emergency visits. In some situations, it has the potential to build and maintain relationships between patients, their families, and the health care team, as well as serve as a means for increased interprofessional collaboration and support. However, there are challenges to overcome that might affect the quality of care using HWT. It is unclear whether PC delivered remotely or with support of HWT is equivalent to the more resource-intensive in-person care. Our findings point toward the importance of caution in choosing when to use HWT in PC and for which patients.

Acknowledgments

The authors wish to express their gratitude to Sophiahemmet University and Mälardalen University for their support and provision of resources, which were instrumental in conducting this state-of-the-art review.

Funding

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

XSL•FO
RenderX

Authors' Contributions

SLS completed database searching. CG, MH, MM, RH, and VZ completed article screening and data extraction. MH, RH, and VZ were responsible for data curation and analysis and data interpretation. MH, RH, and VZ wrote the original draft. CG, MM, and SLS provided advice. All authors contributed to reviewing and editing the final manuscript and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR checklist.
[DOCX File , 85 KB - jmir_v28i1e79637_app1.docx]

Multimedia Appendix 2
Documentation database searches.
[DOCX File , 42 KB - jmir_v28i1e79637_app2.docx]

Multimedia Appendix 3
Characteristics of included studies.
[DOCX File , 61 KB - jmir_v28i1e79637_app3.docx]

References

1. Centeno C, Garralda E, Carrasco JM, den Herder-van der Eerden M, Aldridge M, Stevenson D. The palliative care challenge: analysis of barriers and opportunities to integrate palliative care in Europe in the view of national associations. J Palliat Med 2017 Nov;20(11):1195-1204. [doi: 10.1089/jpm.2017.0039] [Medline: 28509657]
2. Knaul FM, Farmer PE, Krakauer EL, De Lima L, Bhadelia A, Jiang Kwete X. Alleviating the access abyss in palliative care and pain relief-an imperative of universal health coverage: the Lancet Commission report. Lancet 2018;391(10128):1391-1454. [doi: 10.1016/S0140-6736(17)32513-8] [Medline: 29032993]
3. Palliative care. World Health Organization. 2018. URL: https://www.who.int/news-room/fact-sheets/detail/palliative-care [accessed 2026-01-16]
4. Radbruch L, De Lima L, Knaul F, Wenk R, Ali Z, Bhatnaghar S. Redefining palliative care—a new consensus-based definition. J Pain Symptom Manage 2020;60(4):754-764 [FREE Full text] [doi: 10.1016/j.jpainsymman.2020.04.027] [Medline: 32387576]
5. Kaasa S, Loge JH, Aapro M, Albreht T, Anderson R, Bruera E. Integration of oncology and palliative care: a Lancet Oncology Commission. Lancet Oncol 2018;19(11):e588-e653 [FREE Full text] [doi: 10.1016/S1470-2045(18)30415-7] [Medline: 30344075]
6. Hansen MB, Rojas-Concha L, Petersen MA, Adsersen M, Groenvold M. Differences in palliative care needs between cancer patients and non-cancer patients at the start of specialized palliative care: a nationwide register-based study. Palliat Med 2024;38(9):1021-1032. [doi: 10.1177/02692163241269705] [Medline: 39194071]
7. Donath C, Ostgathe C, Heckel M. Dementia patients in palliative care according to data from the German National Hospice and Palliative Care Register (2009-2021). BMC Palliat Care 2024;23(1):185 [FREE Full text] [doi: 10.1186/s12904-024-01509-0] [Medline: 39054518]
8. Arnold E, Finucane AM, Oxenham D. Preferred place of death for patients referred to a specialist palliative care service. BMJ Support Palliat Care 2015;5(3):294-296 [FREE Full text] [doi: 10.1136/bmjspcare-2012-000338] [Medline: 24644165]
9. Ali M, Capel M, Jones G, Gazi T. The importance of identifying preferred place of death. BMJ Support Palliat Care 2019;9(1):84-91. [doi: 10.1136/bmjspcare-2015-000878] [Medline: 26408428]
10. Hudson P. Home-based support for palliative care families: challenges and recommendations. Med J Aust 2003;179(S6):S35-S37. [doi: 10.5694/j.1326-5377.2003.tb05576.x] [Medline: 12964935]
11. Mack JW, Jacobson J, Frank D, Cronin AM, Horvath K, Allen V. Evaluation of patient and family outpatient complaints as a strategy to prioritize efforts to improve cancer care delivery. Jt Comm J Qual Patient Saf 2017;43(10):498-507. [doi: 10.1016/j.jcjq.2017.04.008] [Medline: 28942774]
12. Ventura AD, Burney S, Brooker J, Fletcher J, Ricciardelli L. Home-based palliative care: a systematic literature review of the self-reported unmet needs of patients and carers. Palliat Med 2014;28(5):391-402. [doi: 10.1177/0269216313511141] [Medline: 24292156]
13. Lundereng ED, Dihle A, Steindal SA. Nurses' experiences and perspectives on collaborative discharge planning when patients receiving palliative care for cancer are discharged home from hospitals. J Clin Nurs 2020;29(17-18):3382-3391. [doi: 10.1111/jocn.15371] [Medline: 32533726]

XSL•FO
RenderX

14. Aranda SK, Hayman-White K. Home caregivers of the person with advanced cancer: an Australian perspective. Cancer Nurs 2001;24(4):300-307. [doi: 10.1097/00002820-200108000-00011] [Medline: 11502039]
15. Sharpe L, Butow P, Smith C, McConnell D, Clarke S. The relationship between available support, unmet needs and caregiver burden in patients with advanced cancer and their carers. Psychooncology 2005;14(2):102-114. [doi: 10.1002/pon.825] [Medline: 15386783]
16. Nze Ossima A, Szfetel D, Denoyel B, Beloucif O, Texereau J, Champion L. End-of life medical spending and care pathways in the last 12 months of life: a comprehensive analysis of the national claims database in France. Medicine (Baltimore) 2023;102(31):e34555 [FREE Full text] [doi: 10.1097/MD.0000000000034555] [Medline: 37543784]
17. Shepperd S, Gonçalves-Bradley DC, Straus S, Wee B. Hospital at home: home-based end-of-life care. Cochrane Database Syst Rev 2021;3(3):CD009231 [FREE Full text] [doi: 10.1002/14651858.CD009231.pub3] [Medline: 33721912]
18. Budd J, Miller BS, Manning EM, Lampos V, Zhuang M, Edelstein M. Digital technologies in the public-health response to COVID-19. Nat Med 2020;26(8):1183-1192. [doi: 10.1038/s41591-020-1011-4] [Medline: 32770165]
19. Andersson SW, Richardson MX, Cozza M, Lindén M, Redekop K. Addressing evidence in health and welfare technology interventions from different perspectives. Health Policy Technol 2021;10(2):100519. [doi: 10.1016/j.hlpt.2021.100519]
20. Danaher J. Techno-optimism: an analysis, an evaluation and a modest defence. Philos. Technol 2022;35(2):54 [FREE Full text] [doi: 10.1007/s13347-022-00550-2]
21. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
22. Kuoppamäki S. The application and deployment of welfare technology in Swedish municipal care: a qualitative study of procurement practices among municipal actors. BMC Health Serv Res 2021;21(1):918 [FREE Full text] [doi: 10.1186/s12913-021-06944-w] [Medline: 34488740]
23. Basile I, Consolo L, Colombo S, Rusconi D, Rampichini F, Lusignani M. Technology to support older adults in home palliative care: a scoping review. Am J Hosp Palliat Care 2024;41(6):673-690 [FREE Full text] [doi: 10.1177/10499091231189502] [Medline: 37473720]
24. Marshall C, Virdun C, Phillips JL. Evidence-based models of rural palliative care: a systematic review. Palliat Med 2023;37(8):1129-1143. [doi: 10.1177/02692163231183994] [Medline: 37537971]
25. Luis-Martínez R, Monje MHG, Antonini A, Sánchez-Ferro Á, Mestre TA. Technology-enabled care: integrating multidisciplinary care in Parkinson's disease through digital technology. Front Neurol 2020;11:575975 [FREE Full text] [doi: 10.3389/fneur.2020.575975] [Medline: 33250846]
26. Windle N, Alam A, Patel H, Street JM, Lathwood M, Farrington T. A retrospective cohort study evaluating the association between implementation of a digital care plan and hospitalization rates for home care residents in the United Kingdom. Home Health Care Manage Pract 2022;35(2):123-128 [FREE Full text] [doi: 10.1177/10848223221135560]
27. Gordon B, Mason B, Smith SLH. Leveraging telehealth for delivery of palliative care to remote communities: a rapid review. J Palliat Care 2022;37(2):213-225 [FREE Full text] [doi: 10.1177/08258597211001184] [Medline: 33730904]
28. Hancock S, Preston N, Jones H, Gadoud A. Telehealth in palliative care is being described but not evaluated: a systematic review. BMC Palliat Care 2019;18(1):114 [FREE Full text] [doi: 10.1186/s12904-019-0495-5] [Medline: 31835998]
29. Hutchinson R, Chiu EJ, Belin SC, Klein-Fedyshin M, Impagliazzo CR, Costanza L. How is telehealth used to increase access to specialty palliative care? A systematic review. J Pain Symptom Manage 2025;69(4):e303-e314. [doi: 10.1016/j.jpainsymman.2024.12.017] [Medline: 39755286]
30. Lundereng ED, Nes AAG, Holmen H, Winger A, Thygesen H, Jøranson N. Health care professionals' experiences and perspectives on using telehealth for home-based palliative care: scoping review. J Med Internet Res 2023;25:e43429 [FREE Full text] [doi: 10.2196/43429] [Medline: 36989024]
31. Namasivayam P, Bui DT, Low C, Barnett T, Bridgman H, Marsh P. The use of telehealth in the provision of after-hours palliative care services in rural and remote Australia: a scoping review. PLoS One 2022;17(9):e0274861 [FREE Full text] [doi: 10.1371/journal.pone.0274861] [Medline: 36156089]
32. Steindal SA, Nes AAG, Godskesen TE, Holmen H, Winger A, Österlind J. Advantages and challenges of using telehealth for home-based palliative care: systematic mixed studies review. J Med Internet Res 2023;25:e43684 [FREE Full text] [doi: 10.2196/43684] [Medline: 36912876]
33. Sánchez-Cárdenas MA, Iriarte-Aristizábal MF, León-Delgado MX, Rodríguez-Campos LF, Correa-Morales JE, Cañón-Piñeros A. Rural palliative care telemedicine for advanced cancer patients: a systematic review. Am J Hosp Palliat Care 2023;40(8):936-944. [doi: 10.1177/10499091221130329] [Medline: 36331174]
34. Tang M, Reddy A. Telemedicine and its past, present, and future roles in providing palliative care to advanced cancer patients. Cancers (Basel) 2022;14(8):1884 [FREE Full text] [doi: 10.3390/cancers14081884] [Medline: 35454791]
35. Widberg C, Wiklund B, Klarare A. Patients' experiences of eHealth in palliative care: an integrative review. BMC Palliat Care 2020;19(1):158 [FREE Full text] [doi: 10.1186/s12904-020-00667-1] [Medline: 33054746]
36. Jess M, Timm H, Dieperink KB. Video consultations in palliative care: a systematic integrative review. Palliat Med 2019;33(8):942-958. [doi: 10.1177/0269216319854938] [Medline: 31187688]

XSL•FO
RenderX

37. Disalvo D, Agar M, Caplan G, Murtagh FE, Luckett T, Heneka N. Virtual models of care for people with palliative care needs living in their own home: a systematic meta-review and narrative synthesis. Palliat Med 2021;35(8):1385-1406. [doi: 10.1177/02692163211024451] [Medline: 34169759]
38. Greenhalgh T. Five biases of new technologies. Br J Gen Pract 2013;63(613):425 [FREE Full text] [doi: 10.3399/bjgp13X670741] [Medline: 23972179]
39. Mormina M. Science, technology and innovation as social goods for development: rethinking research capacity building from Sen's capabilities approach. Sci Eng Ethics 2019;25(3):671-692 [FREE Full text] [doi: 10.1007/s11948-018-0037-1] [Medline: 29497970]
40. Hofmann B. Progress bias versus status quo bias in the ethics of emerging science and technology. Bioethics 2020;34(3):252-263. [doi: 10.1111/bioe.12622] [Medline: 31617222]
41. Finucane AM, O'Donnell H, Lugton J, Gibson-Watt T, Swenson C, Pagliari C. Digital health interventions in palliative care: a systematic meta-review. NPJ Digit Med 2021;4(1):64 [FREE Full text] [doi: 10.1038/s41746-021-00430-7] [Medline: 33824407]
42. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Info Libr J 2009;26(2):91-108 [FREE Full text] [doi: 10.1111/j.1471-1842.2009.00848.x] [Medline: 19490148]
43. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
44. Gustafsson C, Zander V, Stridsberg S, Holm M, Mazaheri M, Hedman R. Health and welfare technology in palliative care. A state-of-the-art review. Open Science Framework. 2022. URL: https://osf.io/yr85w [accessed 2026-01-23]
45. Covidence systematic review software. Veritas Health Innovation. URL: https://www.covidence.org/ [accessed 2026-01-16]
46. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008;8:45 [FREE Full text] [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
47. Kazankov K, Novelli S, Chatterjee DA, Phillips A, Balaji A, Raja M. Evaluation of CirrhoCare—a digital health solution for home management of individuals with cirrhosis. J Hepatol 2023;78(1):123-132. [doi: 10.1016/j.jhep.2022.08.034] [Medline: 36087864]
48. Jiang B, Bills M, Poon P. Integrated telehealth-assisted home-based specialist palliative care in rural Australia: a feasibility study. J Telemed Telecare 2023;29(1):50-57. [doi: 10.1177/1357633X20966466] [Medline: 33079611]
49. Gustafson D, DuBenske LL, Namkoong K, Hawkins R, Chih MY, Atwood AK. An eHealth system supporting palliative care for patients with non-small cell lung cancer: a randomized trial. Cancer 2013;119(9):1744-1751 [FREE Full text] [doi: 10.1002/cncr.27939] [Medline: 23355273]
50. Ngoma M, Mushi B, Morse RS, Ngoma T, Mahuna H, Lambden K. mPalliative Care Link: examination of a mobile solution to palliative care coordination among Tanzanian patients with cancer. JCO Glob Oncol 2021;7:1306-1315 [FREE Full text] [doi: 10.1200/GO.21.00122] [Medline: 34406856]
51. Manz CR, Zhang Y, Chen K, Long Q, Small DS, Evans CN. Long-term effect of machine learning–triggered behavioral nudges on serious illness conversations and end-of-life outcomes among patients with cancer. JAMA Oncol 2023;9(3):414-418 [FREE Full text] [doi: 10.1001/jamaoncol.2022.6303] [Medline: 36633868]
52. Tumeh IBRG, Bergerot CD, Lee D, Philip EJ, Freitas-Júnior R. mHealth program for patients with advanced cancer receiving treatment in a public health hospital in Brazil. Psychooncology 2023;32(1):125-132. [doi: 10.1002/pon.6059] [Medline: 36284459]
53. Mooney K, Whisenant MS, Wilson CM, Coombs LA, Lloyd J, Alekhina N. Technology-assisted mHealth caregiver support to manage cancer patient symptoms: a randomized controlled trial. J Pain Symptom Manage 2023;66(1):33-43. [doi: 10.1016/j.jpainsymman.2023.02.320] [Medline: 36889453]
54. Besse KTC, Faber-te Boveldt ND, Janssen GHP, Vernooij-Dassen M, Vissers KCP, Engels Y. Pain assessment with short message service and interactive voice response in outpatients with cancer and pain: a feasibility study. Pain Pract 2016;16(3):320-326. [doi: 10.1111/papr.12278] [Medline: 25581306]
55. Cornetta K, Nyariki S, Manji I, Kiplimo I, Korir M, Muinga E. Telehospice for cancer patients discharged from a tertiary care hospital in Western Kenya. J Pain Symptom Manage 2023;65(5):378-387 [FREE Full text] [doi: 10.1016/j.jpainsymman.2023.01.027] [Medline: 36773732]
56. Mark DD, Ikehara C, Matsuura C, Hara K, Li D. Validating the impact of teaching pursed-lips breathing with Skype. J Hosp Palliat Nurs 2013;15(8):424-432. [doi: 10.1097/njh.0000000000000015]
57. Rafter L. Employing e-health in the palliative care setting to manage pressure ulcers. Wounds UK 2016;12(1):114-123. [doi: 10.1093/oso/9780198508434.003.0028]
58. Moscato S, Sichi V, Giannelli A, Palumbo P, Ostan R, Varani S. Virtual reality in home palliative care: brief report on the effect on cancer-related symptomatology. Front Psychol 2021;12:709154 [FREE Full text] [doi: 10.3389/fpsyg.2021.709154] [Medline: 34630217]
59. Milbury K, Li Y, Durrani S, Liao Z, Tsao AS, Carmack C. A mindfulness-based intervention as a supportive care strategy for patients with metastatic non-small cell lung cancer and their spouses: results of a three-arm pilot randomized controlled trial. Oncologist 2020;25(11):e1794-e1802 [FREE Full text] [doi: 10.1634/theoncologist.2020-0125] [Medline: 32621630]

XSL•FO
RenderX

60. Nemecek R, Huber P, Schur S, Masel EK, Baumann L, Hoeller C. Telemedically augmented palliative care: empowerment for patients with advanced cancer and their family caregivers. Wien Klin Wochenschr 2019;131(23-24):620-626. [doi: 10.1007/s00508-019-01562-3] [Medline: 31667588]
61. Watanabe SM, Fairchild A, Pituskin E, Borgersen P, Hanson J, Fassbender K. Improving access to specialist multidisciplinary palliative care consultation for rural cancer patients by videoconferencing: report of a pilot project. Support Care Cancer 2013;21(4):1201-1217. [doi: 10.1007/s00520-012-1649-7] [Medline: 23161339]
62. Applebaum AJ, Buda KL, Schofield E, Farberov M, Teitelbaum ND, Evans K. Exploring the cancer caregiver's journey through web-based meaning-centered psychotherapy. Psychooncology 2018;27(3):847-856 [FREE Full text] [doi: 10.1002/pon.4583] [Medline: 29136682]
63. Hoek PD, Schers HJ, Bronkhorst EM, Vissers KCP, Hasselaar JGJ. The effect of weekly specialist palliative care teleconsultations in patients with advanced cancer -a randomized clinical trial. BMC Med 2017;15(1):119 [FREE Full text] [doi: 10.1186/s12916-017-0866-9] [Medline: 28625164]
64. Ho YX, Morse RS, Lambden K, Mushi BP, Ngoma M, Mahuna H. How a digital case management platform affects community-based palliative care of sub-Saharan African cancer patients: clinician-users' perspectives. Appl Clin Inform 2022;13(5):1092-1099 [FREE Full text] [doi: 10.1055/s-0042-1758223] [Medline: 36384234]
65. LeBaron V, Alam R, Bennett R, Blackhall L, Gordon K, Hayes J. Deploying the Behavioral and Environmental Sensing and Intervention for Cancer Smart Health System to support patients and family caregivers in managing pain: feasibility and acceptability study. JMIR Cancer 2022;8(3):e36879 [FREE Full text] [doi: 10.2196/36879] [Medline: 35943791]
66. Graven LJ, Abbott L, Schluck G. The coping in heart failure (COPE-HF) partnership intervention for heart failure symptoms: implications for palliative care. Prog Palliat Care 2022;31(3):169-178. [doi: 10.1080/09699260.2022.2124144]
67. Cameron P. Caregiver comfort with telehospice. Home Healthc Now 2021;39(1):32-38. [doi: 10.1097/NHH.0000000000000928] [Medline: 33417360]
68. Hennemann-Krause L, Lopes AJ, Araújo JA, Petersen EM, Nunes RA. The assessment of telemedicine to support outpatient palliative care in advanced cancer. Palliat Support Care 2015;13(4):1025-1030. [doi: 10.1017/S147895151400100X] [Medline: 25159308]
69. Maguire R, Connaghan J, Arber A, Klepacz N, Blyth KG, McPhelim J. Advanced Symptom Management System for patients with malignant pleural mesothelioma (ASyMSmeso): mixed methods study. J Med Internet Res 2020;22(11):e19180 [FREE Full text] [doi: 10.2196/19180] [Medline: 33180025]
70. Palma A, Rojas V, Ihl F, Ávila C, Plaza-Parrochia F, Estuardo N. Implementation of a palliative hospital-centered spiritual and psychological telehealth system during COVID-19 pandemic. J Pain Symptom Manage 2021;62(5):1015-1019 [FREE Full text] [doi: 10.1016/j.jpainsymman.2021.04.016] [Medline: 33957254]
71. Gustafson DH, DuBenske LL, Atwood AK, Chih M, Johnson RA, McTavish F. Reducing symptom distress in patients with advanced cancer using an e-alert system for caregivers: pooled analysis of two randomized clinical trials. J Med Internet Res 2017;19(11):e354 [FREE Full text] [doi: 10.2196/jmir.7466] [Medline: 29138131]
72. Zeiser SR. Quality improvement evaluation of a Veterans' palliative care telehealth music therapy program. Music Ther Perspect 2023;41(1):54-62. [doi: 10.1093/mtp/miac017]
73. Greer J, Jacobs J, Pensak N, MacDonald JJ, Fuh CX, Perez GK. Randomized trial of a tailored cognitive-behavioral therapy mobile application for anxiety in patients with incurable cancer. Oncologist 2019;24(8):1111-1120. [doi: 10.1634/theoncologist.2018-0536] [Medline: 30683710]
74. Merz A, Mohamed A, Corbett C, Herring K, Hildenbrand J, Locke SC. A single-site pilot feasibility randomized trial of a supportive care mobile application intervention for patients with advanced cancer and caregivers. Support Care Cancer 2022;30(10):7853-7861. [doi: 10.1007/s00520-022-07224-x] [Medline: 35718794]
75. Baird-Bower D, Roach J, Andrews M, Onslow F, Curnin E. Help is just a phone call away: after-hours support for palliative care patients wishing to die at home. Int J Palliat Nurs 2016;22(6):286-291. [doi: 10.12968/ijpn.2016.22.6.286] [Medline: 27349847]
76. Finucane AM, Davydaitis D, Horseman Z, Carduff E, Baughan P, Tapsfield J. Electronic care coordination systems for people with advanced progressive illness: a mixed-methods evaluation in Scottish primary care. Br J Gen Pract 2020;70(690):e20-e28 [FREE Full text] [doi: 10.3399/bjgp19X707117] [Medline: 31848198]
77. Johnston J, Hillier LM, Purdon M, Pears D, Robson M. Opportunities for improving the quality of life of dying persons: use of technology to connect people at the end of life. J Palliat Care 2018;28(3):178-180. [doi: 10.1177/082585971202800309]
78. Royackers AR, Regan S, Donelle L. The eShift model of care: informal caregivers' experience of a new model of home-based palliative care. Prog Palliat Care 2016;24(2):84-92. [doi: 10.1179/1743291x15y.0000000006]
79. LoCastro M, Sanapala C, Mendler JH, Norton S, Bernacki R, Carroll T. Adaptation of Serious Illness Care Program to be delivered via telehealth for older patients with hematologic malignancy. Blood Adv 2023;7(9):1871-1884 [FREE Full text] [doi: 10.1182/bloodadvances.2022008996] [Medline: 36521100]
80. Harding R, Carrasco JM, Serrano-Pons J, Lemaire J, Namisango E, Luyirika E. Design and evaluation of a novel mobile phone application to improve palliative home-care in resource-limited settings. J Pain Symptom Manage 2021;62(1):1-9 [FREE Full text] [doi: 10.1016/j.jpainsymman.2020.09.045] [Medline: 33246073]

XSL•FO
RenderX

81. Wilkie DJ, Yao Y, Ezenwa MO, Suarez ML, Dyal BW, Gill A. A stepped-wedge randomized controlled trial: effects of eHealth interventions for pain control among adults with cancer in hospice. J Pain Symptom Manage 2020;59(3):626-636 [FREE Full text] [doi: 10.1016/j.jpainsymman.2019.10.028] [Medline: 31711969]
82. Hao Y, Zhan L, Huang M, Cui X, Zhou Y, Xu E. Nurses' knowledge and attitudes towards palliative care and death: a learning intervention. BMC Palliat Care 2021;20(1):50 [FREE Full text] [doi: 10.1186/s12904-021-00738-x] [Medline: 33765995]
83. White C, McIlfatrick S, Dunwoody L, Watson M. Supporting and improving community health services-a prospective evaluation of ECHO technology in community palliative care nursing teams. BMJ Support Palliat Care 2019;9(2):202-208. [doi: 10.1136/bmjspcare-2015-000935] [Medline: 26628535]
84. Chiu Wu CH, Perng S, Shi C, Lai H. Advance care planning and advance directives: a multimedia education program in community-dwelling older adults. J Appl Gerontol 2020;39(8):811-819. [doi: 10.1177/0733464819831596] [Medline: 30795711]
85. Köhle N, Drossaert CHC, Jaran J, Schreurs KMG, Verdonck-de Leeuw IM, Bohlmeijer ET. User-experiences with a web-based self-help intervention for partners of cancer patients based on acceptance and commitment therapy and self-compassion: a qualitative study. BMC Public Health 2017;17(1):225 [FREE Full text] [doi: 10.1186/s12889-017-4121-2] [Medline: 28245794]
86. Ariss SMB, Taylor P, Fitzsimmons D, Kyerematen S, Mawson S. Mobile technology and delegated work in specialist community services: the EnComPaSS Integration project. BMJ Support Palliat Care 2021:e1144-e1151 [FREE Full text] [doi: 10.1136/bmjspcare-2020-002288] [Medline: 34916238]
87. Samara JL, Liu W, Kroon W, Harvie B, Hingeley R, Johnston N. Telehealth palliative care needs rounds during a pandemic. The Journal for Nurse Practitioners 2021;17(3):335-338. [doi: 10.1016/j.nurpra.2020.12.009]
88. Shirley LD, DiScala S, Brooks A, Vartan CM, Silverman M. Pilot of a pharmacist-integrated interprofessional team to optimize prescribing in a telemedicine palliative care clinic. J Am Coll Clin Pharm 2021;4(9):1093-1099. [doi: 10.1002/jac5.1478]
89. Bhargava R, Keating B, Isenberg SR, Subramaniam S, Wegier P, Chasen M. RELIEF: a digital health tool for the remote self-reporting of symptoms in patients with cancer to address palliative care needs and minimize emergency department visits. Curr Oncol 2021;28(6):4273-4280 [FREE Full text] [doi: 10.3390/curroncol28060363] [Medline: 34898539]
90. Marco D, McMillan-Drendel E, Philip JAM, Williamson T, Le B. Establishment of the first Australian public and health-professional palliative care advice service: exploring caller needs and gaps in care. Aust Health Rev 2023;47(5):569-573. [doi: 10.1071/AH23108] [Medline: 37516935]
91. Atreya S, Kumar G, Samal J, Bhattacharya M, Banerjee S, Mallick P. Patients'/caregivers' perspectives on telemedicine service for advanced cancer patients during the COVID-19 pandemic: an exploratory survey. Indian J Palliat Care 2020;26(Suppl 1):S40-S44 [FREE Full text] [doi: 10.4103/IJPC.IJPC_145_20] [Medline: 33088085]
92. Teter CM, Whitted K, Vardaman SA. Reducing hospital visit rates in hospice patients using telemedicine. Comput Inform Nurs 2021;39(11):682-688. [doi: 10.1097/CIN.0000000000000769] [Medline: 34145206]
93. Bange EM, Li Y, Kumar P, Doucette A, Gabriel P, Parikh R. The association between telemedicine, advance care planning, and unplanned hospitalizations among high-risk patients with cancer. Cancer 2024;130(4):636-644 [FREE Full text] [doi: 10.1002/cncr.35116] [Medline: 37987207]
94. Baxter KE, Kochar S, Williams C, Blackman C, Himmelvo J. Development of a palliative telehealth pilot to meet the needs of the nursing home population. J Hosp Palliat Nurs 2021;23(5):478-483. [doi: 10.1097/NJH.0000000000000784] [Medline: 34225341]
95. Bückmann A, Bernhardt F, Eveslage M, Storck M, Thölking G, Buss H. Telemedical consultations in palliative care: benefits through knowledge exchange and intercollegiate collaboration—findings from the German oVID project. Cancers (Basel) 2023;15(9):2512 [FREE Full text] [doi: 10.3390/cancers15092512] [Medline: 37173978]
96. Morgan DD, Swetenham K, To THM, Currow DC, Tieman JJ. Telemonitoring via self-report and video review in community palliative care: a case report. Healthcare (Basel) 2017;5(3):51 [FREE Full text] [doi: 10.3390/healthcare5030051] [Medline: 28858221]
97. McCreedy EM, Yang X, Mitchell SL, Gutman R, Teno J, Loomer L. Effect of advance care planning video on do-not-hospitalize orders for nursing home residents with advanced illness. BMC Geriatr 2022;22(1):298 [FREE Full text] [doi: 10.1186/s12877-022-02970-3] [Medline: 35392827]
98. Takahashi PY, Hanson GJ, Thorsteinsdottir B, Van Houten HK, Shah ND, Naessens JM. The impact of telemonitoring upon hospice referral in the community: a randomized controlled trial. Clin Interv Aging 2012;7:445-451 [FREE Full text] [doi: 10.2147/CIA.S36461] [Medline: 23152674]
99. Hayek S, Nieva R, Corrigan F, Zhou A, Mudaliar U, Mays D. End-of-life care planning: improving documentation of advance directives in the outpatient clinic using electronic medical records. J Palliat Med 2014;17(12):1348-1352. [doi: 10.1089/jpm.2013.0684] [Medline: 24988497]
100. Rainsford S, Hall Dykgraaf S, Phillips C. Effectiveness of telehealth palliative care Needs Rounds in rural residential aged care during the COVID-19 pandemic: a hybrid effectiveness-implementation study. Aust J Rural Health 2022;30(1):108-114 [FREE Full text] [doi: 10.1111/ajr.12789] [Medline: 34757687]

XSL•FO
RenderX

101. Taubert M, Norris J, Edwards S, Snow V, Finlay IG. Talk CPR - a technology project to improve communication in do not attempt cardiopulmonary resuscitation decisions in palliative illness. BMC Palliat Care 2018;17(1):118 [FREE Full text] [doi: 10.1186/s12904-018-0370-9] [Medline: 30340632]
102. Beesley VL, Turner J, Chan RJ, Yates P, Gordon LG, Burge M. Supporting patients and carers affected by pancreatic cancer: a feasibility study of a counselling intervention. Eur J Oncol Nurs 2020;46:101729. [doi: 10.1016/j.ejon.2020.101729] [Medline: 32339910]
103. Alcalde Castro MJ, Zaig S, Nissim R, O'Connor B, Lau J, Mak E. Telehealth outpatient palliative care in the COVID-19 pandemic: patient experience qualitative study. BMJ Support Palliat Care 2024;14:e1512-e1520. [doi: 10.1136/spcare-2023-004189] [Medline: 36828625]
104. Bandini JI, Scherling A, Farmer C, Fratkin M, Mistler B, Ast K. Experiences with telehealth for outpatient palliative care: findings from a mixed-methods study of patients and providers across the United States. J Palliat Med 2022;25(7):1079-1087. [doi: 10.1089/jpm.2021.0545] [Medline: 35506994]
105. Bentley B, O'Connor M, Williams A, Breen LJ. Dignity therapy online: piloting an online psychosocial intervention for people with terminal illness. Digit Health 2020;6:2055207620958527 [FREE Full text] [doi: 10.1177/2055207620958527] [Medline: 33014409]
106. Bonsignore L, Bloom N, Steinhauser K, Nichols R, Allen T, Twaddle M. Evaluating the feasibility and acceptability of a telehealth program in a rural palliative care population: TapCloud for palliative care. J Pain Symptom Manage 2018;56(1):7-14 [FREE Full text] [doi: 10.1016/j.jpainsymman.2018.03.013] [Medline: 29551433]
107. Lind LK, Karlsson D. Electronic patient-reported symptom assessment in palliative end-of-life home care. Health Syst 2017;2(3):171-180. [doi: 10.1057/hs.2013.4]
108. Read Paul L, Salmon C, Sinnarajah A, Spice R. Web-based videoconferencing for rural palliative care consultation with elderly patients at home. Support Care Cancer 2019;27(9):3321-3330. [doi: 10.1007/s00520-018-4580-8] [Medline: 30613908]
109. Groothuizen JE, Aroyewun E, Zasada M, Harris J, Hewish M, Taylor C. Virtually the same? Examining the impact of the COVID-19 related shift to virtual lung cancer multidisciplinary team meetings in the UK National Health Service: a mixed methods study. BMJ Open 2023;13(6):e065494 [FREE Full text] [doi: 10.1136/bmjopen-2022-065494] [Medline: 37328174]
110. Oelschlägel L, Christensen VL, Moen A, Heggdal K, Österlind J, Dihle A. Patients' experiences with a welfare technology application for remote home care: a longitudinal study. J Clin Nurs 2023;32(17-18):6545-6558. [doi: 10.1111/jocn.16592] [Medline: 36437490]
111. Haydon HM, Snoswell CL, Thomas EE, Broadbent A, Caffery LJ, Brydon J. Enhancing a community palliative care service with telehealth leads to efficiency gains and improves job satisfaction. J Telemed Telecare 2021;27(10):625-630. [doi: 10.1177/1357633X211048952] [Medline: 34726990]
112. Hutchinson RN, Anderson EC, Ruben MA, Manning N, John L, Daruvala A. A formative mixed-methods study of emotional responsiveness in telepalliative care. J Palliat Med 2022;25(8):1258-1267 [FREE Full text] [doi: 10.1089/jpm.2021.0589] [Medline: 35417249]
113. Pasanen L, Le Gautier R, Wong A, Wawryk O, Collins A, Schwetlik S. Telehealth in outpatient delivery of palliative care: a qualitative study of patient and physician views. Palliat Support Care 2023;21(6):980-987. [doi: 10.1017/S1478951522000670] [Medline: 35818898]
114. Tasneem S, Kim A, Bagheri A, Lebret J. Telemedicine video visits for patients receiving palliative care: a qualitative study. Am J Hosp Palliat Care 2019;36(9):789-794. [doi: 10.1177/1049909119846843] [Medline: 31064195]
115. Soliman AA, Akgün KM, Coffee J, Kapo J, Morrison LJ, Hopkinson E. Quality of telehealth-delivered inpatient palliative care during the early COVID-19 pandemic. J Pain Symptom Manage 2023;65(1):6-15 [FREE Full text] [doi: 10.1016/j.jpainsymman.2022.09.014] [Medline: 36206949]
116. van Gurp J, van Selm M, Vissers K, van Leeuwen E, Hasselaar J. How outpatient palliative care teleconsultation facilitates empathic patient-professional relationships: a qualitative study. PLoS One 2015;10(4):e0124387 [FREE Full text] [doi: 10.1371/journal.pone.0124387] [Medline: 25902263]
117. Viitala A, Åstedt-Kurki P, Lehto JT, Palonen M. Online follow-up with a mobile device improves incurable cancer patients' coping - a qualitative study. Eur J Oncol Nurs 2021;55:102047 [FREE Full text] [doi: 10.1016/j.ejon.2021.102047] [Medline: 34688053]
118. Rosa WE, Lynch KA, Hadler RA, Mahoney C, Parker PA. "Doing palliative care with my hands tied behind my back": telepalliative care delivery for oncology inpatients during a COVID-19 surge. Transl Behav Med 2022;12(7):816-824 [FREE Full text] [doi: 10.1093/tbm/ibac044] [Medline: 35716077]
119. Collier A, Morgan DD, Swetenham K, To TH, Currow DC, Tieman JJ. Implementation of a pilot telehealth programme in community palliative care: a qualitative study of clinicians' perspectives. Palliat Med 2016;30(4):409-417. [doi: 10.1177/0269216315600113] [Medline: 26290500]
120. Misplon S, Marneffe W, Himpe U, Hellings J, Demedts I. Evaluation of the implementation of value-based healthcare with a weekly digital follow-up of lung cancer patients in clinical practice. Eur J Cancer Care (Engl) 2022;31(6):e13653. [doi: 10.1111/ecc.13653] [Medline: 35819130]

XSL•FO
RenderX

121. Benson JJ, Oliver DP, Washington KT, Rolbiecki AJ, Lombardo CB, Garza JE. Online social support groups for informal caregivers of hospice patients with cancer. Eur J Oncol Nurs 2020;44:101698 [FREE Full text] [doi: 10.1016/j.ejon.2019.101698] [Medline: 31816508]
122. Elma A, Cook D, Howard M, Takaoka A, Hoad N, Swinton M. Use of video technology in end-of-life care for hospitalized patients during the COVID-19 pandemic. Am J Crit Care 2022;31(3):240-248. [doi: 10.4037/ajcc2022722] [Medline: 35118491]
123. Halton-Hernandez E, Abrams R, Cooney G, Ali K. Exploring audio recording in terminally ill patients receiving hospice care: “Stories for Life Charity” model. J Hosp Palliat Nurs 2023;25(5):271-276. [doi: 10.1097/NJH.0000000000000958] [Medline: 37347958]
124. Kuntz JG, Kavalieratos D, Esper GJ, Ogbu N, Mitchell J, Ellis CM. Feasibility and acceptability of inpatient palliative care e-family meetings during COVID-19 pandemic. J Pain Symptom Manage 2020;60(3):e28-e32 [FREE Full text] [doi: 10.1016/j.jpainsymman.2020.06.001] [Medline: 32505643]
125. Guo Q, Cann B, McClement S, Thompson G, Chochinov HM. Keep in Touch (KIT): feasibility of using internet-based communication and information technology in palliative care. BMC Palliat Care 2017;16(1):29 [FREE Full text] [doi: 10.1186/s12904-017-0203-2] [Medline: 28477630]
126. Potter G, Pesut B, Hooper BP, Erbacker L. Team-based education in a palliative approach for rural nurses and unlicensed care providers. J Contin Educ Nurs 2015;46(6):279-288. [doi: 10.3928/00220124-20150518-04] [Medline: 26057165]
127. van Gurp J, van Selm M, van Leeuwen E, Vissers K, Hasselaar J. Teleconsultation for integrated palliative care at home: a qualitative study. Palliat Med 2016;30(3):257-269. [doi: 10.1177/0269216315598068] [Medline: 26269323]
128. Kamal AH, Wolf S, Nicolla JM, Friedman F, Xuan M, Bennett AV. Usability of PCforMe in patients with advanced cancer referred to outpatient palliative care: results of a randomized, active-controlled pilot trial. J Pain Symptom Manage 2019;58(3):382-389 [FREE Full text] [doi: 10.1016/j.jpainsymman.2019.05.007] [Medline: 31163259]
129. Middleton-Green L, Gadoud A, Norris B, Sargeant A, Nair S, Wilson L. ‘A friend in the corner’: supporting people at home in the last year of life via telephone and video consultation—an evaluation. BMJ Support Palliat Care 2019;9(4):e26. [doi: 10.1136/bmjspcare-2015-001016] [Medline: 26850118]
130. Dionne-Odom JN, Ejem DB, Wells R, Azuero A, Stockdill ML, Keebler K. Effects of a telehealth early palliative care intervention for family caregivers of persons with advanced heart failure: The ENABLE CHF-PC randomized clinical trial. JAMA Netw Open 2020;3(4):e202583 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.2583] [Medline: 32282044]
131. Parker D, Hudson P, Tieman J, Thomas K, Saward D, Ivynian S. Evaluation of an online toolkit for carers of people with a life-limiting illness at the end-of-life: health professionals' perspectives. Aust J Prim Health 2021;27(6):473-478 [FREE Full text] [doi: 10.1071/PY21019] [Medline: 34802508]
132. Cruz-Oliver DM, Abshire Saylor M, Nelson KE, Milner GE, Blinka MD, Durkin N. Hospice family caregiver perceptions of benefits and challenges of a telenovela educational intervention. J Palliat Med 2022;25(6):945-951 [FREE Full text] [doi: 10.1089/jpm.2021.0628] [Medline: 35446674]
133. Koufacos NS, Gottesman EM, Dorisca E, Howe JL. Supporting caregivers of veterans with dementia. J Soc Work End Life Palliat Care 2023;19(1):12-22. [doi: 10.1080/15524256.2023.2184597] [Medline: 36856029]
134. Olesen LK, la Cour K, With H, Mahoney AF, Handberg C. A cross-sectional evaluation of acceptability of an online palliative rehabilitation program for family caregivers of people with amyotrophic lateral sclerosis and cognitive and behavioral impairments. BMC Health Serv Res 2022;22(1):697 [FREE Full text] [doi: 10.1186/s12913-022-07986-4] [Medline: 35610609]
135. Rochmawati EP, Putranto D. Mobile application-based education to improve family caregivers’ readiness: feasibility study. J Nurse Pract 2022;18(10):1097-1101. [doi: 10.1016/j.nurpra.2022.08.023]
136. Schuler T, King C, Matsveru T, Back M, Clark K, Chin D. Wearable-triggered ecological momentary assessments are feasible in people with advanced cancer and their family caregivers: feasibility study from an outpatient palliative care clinic at a cancer center. J Palliat Med 2023;26(7):980-985. [doi: 10.1089/jpm.2022.0535] [Medline: 37134212]
137. DuBenske LL, Gustafson DH, Namkoong K, Hawkins RP, Atwood AK, Brown RL. CHESS improves cancer caregivers' burden and mood: results of an eHealth RCT. Health Psychol 2014;33(10):1261-1272 [FREE Full text] [doi: 10.1037/a0034216] [Medline: 24245838]
138. Chih M, DuBenske LL, Hawkins RP, Brown RL, Dinauer SK, Cleary JF. Communicating advanced cancer patients’ symptoms via the internet: a pooled analysis of two randomized trials examining caregiver preparedness, physical burden, and negative mood. Palliat Med 2013;27(6):533-543 [FREE Full text] [doi: 10.1177/0269216312457213] [Medline: 22988042]
139. Schuit AS, Rienks MM, Hooghiemstra N, Jansen F, Lissenberg-Witte BI, Cuijpers P. Reach and efficacy of the eHealth application Oncokompas, facilitating partners of incurably ill cancer patients to self-manage their caregiver needs: a randomized controlled trial. Support Care Cancer 2022;30(12):10191-10201 [FREE Full text] [doi: 10.1007/s00520-022-07441-4] [Medline: 36367563]
140. Hamilton E, Van Veldhuizen E, Brown A, Brennan S, Sabesan S. Telehealth in radiation oncology at the Townsville Cancer Centre: service evaluation and patient satisfaction. Clin Transl Radiat Oncol 2019;15:20-25 [FREE Full text] [doi: 10.1016/j.ctro.2018.11.005] [Medline: 30582017]

XSL•FO
RenderX

141. Wilkie DJ, Ezenwa MO. Pain and symptom management in palliative care and at end of life. Nurs Outlook 2012;60(6):357-364 [FREE Full text] [doi: 10.1016/j.outlook.2012.08.002] [Medline: 22985972]
142. Murray CD, McDonald C, Atkin H. The communication experiences of patients with palliative care needs: a systematic review and meta-synthesis of qualitative findings. Palliat Support Care 2015;13(2):369-383. [doi: 10.1017/S1478951514000455] [Medline: 24784479]
143. Chapman EJ, Paley CA, Pini S, Ziegler LE. Correction: updating a conceptual model of effective symptom management in palliative care to include patient and carer perspective: a qualitative study. BMC Palliat Care 2024;23(1):237 [FREE Full text] [doi: 10.1186/s12904-024-01571-8] [Medline: 39367358]
144. Otto L, Harst L, Schlieter H, Wollschlaeger B, Richter P. Towards a unified understanding of ehealth and related terms: proposal of a consolidated terminological basis. 2018 Presented at: Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) – HEALTHINF; January 18-20, 2018; Funchal. [doi: 10.5220/0006651005330539]
145. Borg J, Gustafsson C, Landerdahl Stridsberg S, Zander V. Implementation of welfare technology: a state-of-the-art review of knowledge gaps and research needs. Disabil Rehabil Assist Technol 2023;18(2):227-239 [FREE Full text] [doi: 10.1080/17483107.2022.2120104] [Medline: 36103349]
146. Kavalieratos D, Gelfman LP, Tycon LE, Riegel B, Bekelman DB, Ikejiani DZ. Palliative care in heart failure: rationale, evidence, and future priorities. J Am Coll Cardiol 2017;70(15):1919-1930 [FREE Full text] [doi: 10.1016/j.jacc.2017.08.036] [Medline: 28982506]

Abbreviations

COPD: chronic obstructive pulmonary disease
HWT: health and welfare technology
MeSH: Medical Subject Headings
PC: palliative care
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by K Liew; submitted 25.Jun.2025; peer-reviewed by WJ She, SHX Ma; comments to author 19.Sep.2025; revised version received 18.Dec.2025; accepted 10.Jan.2026; published 12.Mar.2026.

Please cite as:
Zander V, Holm M, Mazaheri M, Gustafsson C, Landerdahl Stridsberg S, Hedman R
Use of Health and Welfare Technology in Palliative Care: State-of-the-Art Review
J Med Internet Res 2026;28:e79637
URL: https://www.jmir.org/2026/1/e79637
doi:10.2196/79637
PMID:

©Viktoria Zander, Maja Holm, Monir Mazaheri, Christine Gustafsson, Sara Landerdahl Stridsberg, Ragnhild Hedman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Effects of Digital Health Interventions on Motor Symptoms, Nonmotor Symptoms, and Quality of Life in Patients With Parkinson Disease: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Ruwen Liu[1,2*], MS; Sirui Zhang[1,2*], MS; Yi Xiao[1,2], MD; Yangfan Cheng[1,2], MD; Huifang Shang[1,2], MD

[1]Department of Neurology, Laboratory of Neurodegenerative Disorders, National Clinical Research Center for Geriatric, West China Hospital of Sichuan University, Chengdu, China

[2]Rare Disease Center, West China Hospital of Sichuan University, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Huifang Shang, MD
Department of Neurology, Laboratory of Neurodegenerative Disorders, National Clinical Research Center for Geriatric
West China Hospital of Sichuan University
37 Guoxue Alley
Chengdu, 610041
China
Phone: 1 028 85422114
Fax: 1 028 85582944
Email: shanghuifang@scu.edu.cn

Abstract

Background: Parkinson disease (PD) is a progressive neurodegenerative disorder with increasing global prevalence, necessitating innovative management. Digital health interventions (DHIs) offer potential advantages for PD care; yet, a comprehensive systematic review and synthesis across all DHI types and core outcomes is still lacking.

Objective: This review aimed to assess the effectiveness of DHIs for improving motor symptoms, nonmotor symptoms, and quality of life in patients with PD and to summarize the reach, uptake, and feasibility.

Methods: We searched PubMed, Ovid Embase, Web of Science, CINAHL, Cochrane Central Register of Controlled Trials, and APA PsycINFO up to November 2025. Pooled standardized mean differences (SMDs) were calculated using random-effects models. We calculated 95% prediction intervals (PIs) to estimate the true effects. The revised Cochrane Risk of Bias 2 tool was used to assess risk of bias. Heterogeneity was assessed using I^2, τ^2, and 95% PI. Subgroup analyses, meta-regression, and sensitivity analyses were conducted to address heterogeneity and potential bias. The quality of evidence was assessed using GRADE (Grading of Recommendations Assessment, Development, and Evaluation).

Results: The review included 112 randomized controlled trials involving 5594 participants. Significant postintervention improvements were identified in motor symptoms (SMD=–0.39, 95% CI –0.60 to –0.18, 95% PI –1.75 to 0.99; I^2=80.3%) and overall nonmotor symptoms (SMD=–0.26, 95% CI –0.49 to –0.03, 95% PI –0.56 to 0.03; I^2=13.8%), including cognitive function (SMD=0.47, 95% CI 0.22 to 0.72, 95% PI –0.41 to 1.35; I^2=63.5%) and psychiatric symptoms (SMD=–0.42, 95% CI –0.74 to –0.09, 95% PI –1.82 to –0.99; I^2=85.4%); however, there was no significant enhancement in quality of life (SMD=–0.19, 95% CI –0.47 to 0.09, 95% PI –1.50 to 1.12; I^2=81.2%). The certainty of evidence was very low for quality of life, motor, and psychiatric symptoms and low for cognitive function and overall nonmotor symptoms. Improvements in motor symptoms and cognitive function remained stable at follow-up. Meta-regression analysis indicated that age, percentage of female participants, and supervision mode were possible sources of heterogeneity. Overall, 94 studies reported reach (median 37.5%), 38 reported fidelity (95.7%), and 105 reported dropout rates (9.1%).

Conclusions: In contrast to previous reviews focused on single technologies or outcomes, this review provided the first comprehensive synthesis across all DHI types on multiple outcomes and indicated their potential as nonpharmacological interventions for PD management. However, current evidence is of low to very low certainty, and wide 95% PIs, together with

XSL•FO
RenderX

high risk of bias and substantial heterogeneity, indicate considerable uncertainty regarding the true effect in future implementations. Therefore, findings should be interpreted with caution. These findings provide integrated evidence to guide the design and prioritization of future research. The results have important real-world implications, supporting cautious implementation while underscoring the need for more robust trials, particularly in resource-limited settings.

Trial Registration: PROSPERO CRD42023492123; https://www.crd.york.ac.uk/PROSPERO/view/CRD42023492123

(J Med Internet Res 2026;28:e79935) doi:10.2196/79935

KEYWORDS

digital; eHealth; meta-analysis; motor symptoms; non-motor symptoms; Parkinson disease; Parkinson's disease; quality of life

Introduction

Parkinson disease (PD) is the most common serious movement disorder, characterized clinically by bradykinesia, rest tremor, rigidity, and postural and gait abnormalities [1,2]. As the global population ages, the global prevalence of PD is projected to exceed 12 million by 2040 [3,4], representing a significant public health challenge. In addition to motor symptoms, nearly all patients experience complex nonmotor manifestations, including hyposmia, autonomic dysfunction, psychiatric symptoms, cognitive decline, and sleep disorders [1,5,6], which also impair the quality of life of patients with PD [7]. PD management mainly relies on dopamine replacement therapy and incorporates a range of medical and surgical treatments for complex symptoms [5].

However, PD management still faces many obstacles that must be overcome. It remains largely limited to symptomatic treatment and is challenged by long-term motor complications as well as many nonmotor symptoms, which increase the complexity of medication regimens. Crucially, there are still no clinically established disease-modifying therapies to slow or halt neurodegeneration [2,8]. Consequently, long-term, individualized pharmacological optimization and multidisciplinary care are essential for people with PD; yet, delivering sustained, high-quality care in routine outpatient and community settings remains challenging. This challenge may underscore the potential value of digital health approaches.

In addition, nonpharmacological interventions such as physical and cognitive training have shown promise for improving both motor and nonmotor symptoms of PD [5,9], but access to these interventions can be challenging for patients. Patients often need to visit medical institutions in person to receive rehabilitation training, which requires sufficient time and financial resources. Furthermore, the disease may make it difficult or cause reluctance for patients to go out. Moreover, in some economically underdeveloped regions, few medical institutions offer such rehabilitation training. In addition, patient adherence also poses a significant challenge. The combination of high prevalence, multifaceted symptomatology, and absence of disease-modifying treatments makes PD management particularly challenging; therefore, strategies that improve symptoms and quality of life are essential to reduce the burden on patients and health care systems.

Digital technology is now revolutionizing PD management, creating new possibilities across the entire spectrum of patient care, from disease identification and diagnosis [10,11] to treatment and prognosis [12,13]. It offers opportunities to enhance and extend nonpharmacological interventions [14,15]. Emerging evidence highlights the potential of digital health interventions (DHIs) in the long-term management of PD [16]. DHIs, defined as the "discrete functionality of digital technology that is applied to achieve health objectives" [17-19], may be especially valuable for long-term PD management.

A growing number of randomized controlled trials (RCTs) have demonstrated that various types of DHIs, including computer-based cognitive training [20] and telerehabilitation [21], may benefit patients with PD. These digital interventions use diverse digital tools, spanning from wearable sensors [22] to smartphone apps [23], to improve both motor and nonmotor outcomes. DHIs offer several potential advantages (1) they may reduce the frequency of in-person visits and thus relieve pressure on medical services and regional resource imbalances in medical resources through information and communication technology [24], (2) they may augment and enhance conventional rehabilitation and pharmacotherapy to improve engagement and outcomes [25], and (3) they may address patients' individualized demands in real time by leveraging adaptive systems that analyze real-time health data [26].

Current DHIs for PD management include robot-assisted physical therapy [27], exergaming [28], videoconferencing [29], smartphone apps [30], wearable sensors [31], and others. Although these studies demonstrate the considerable potential of DHIs in PD care, no study has summarized current evidence systematically and comprehensively in this field, as previous reviews primarily focused on single digital intervention types (eg, virtual reality) [32] or specific outcomes (eg, motor symptoms) [33]. A comprehensive synthesis across DHI modalities and health-related outcomes in PD is still lacking.

To fill this gap, we performed a large-scale, updated systematic review and meta-analysis to evaluate the effectiveness of diverse DHIs across multiple health outcomes in patients with PD. We additionally summarized the reported reach, fidelity, and feasibility of these interventions. Our findings aim to inform clinicians in selecting management strategies for PD and to guide priorities for future research.

Methods

Protocol and Registration

This systematic review and meta-analysis were preregistered on PROSPERO (CRD42023492123) and were conducted according to the PRISMA (Preferred Reporting Items for

XSL•FO
RenderX

Systematic Reviews and Meta-Analyses) guidelines for transparent and comprehensive reporting of methodology and results [34]. The PRISMA 2020 checklist and the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) checklist are provided in Multimedia Appendices 1 and 2, respectively. Any deviations from the preregistered protocol are outlined and explained in Multimedia Appendix 3.

Search Strategy and Study Selection

Searches were first performed in PubMed (via NCBI), Ovid Embase (via Ovid), Web of Science (via Clarivate), CINAHL (via EBSCOhost), Cochrane Central Register of Controlled Trials (via Wiley), and APA PsycINFO (via EBSCOhost) from inception to November 21, 2023. Updated database searches were performed in the above databases via the same platforms from November 21, 2023, to January 5, 2025. Final database searches were conducted on November 22, 2025, for articles published in 2025, and email alerts were maintained to capture newly published studies until final data analysis. No study registries, websites, or other non–peer-reviewed online resources were systematically searched to ensure the reliability of the included data. The reference lists of included articles and other relevant systematic reviews were examined. Authors of potentially eligible studies were contacted via email to inquire about additional data or unpublished results.

Three groups of search terms were used: terms related to PD, terms related to DHIs (eg, digital, technology, telemedicine, internet, robotics, virtual reality, and computers), and terms related to RCTs. Complete search strategies and search limits are detailed in Multimedia Appendix 4. Search results were imported into EndNote 20 (Clarivate), and duplicates were removed using the software's automatic deduplication feature, followed by manual review. The titles and abstracts of the involved studies were first screened, and then the full texts of potentially eligible studies were reviewed according to the inclusion criteria.

The inclusion criteria were developed based on the PICOS (Population, Intervention, Comparison, Outcomes, and Study Type) framework. First, for population, we included studies involving patients with a confirmed diagnosis of PD. Second, for intervention, interventions provided through the computer, smartphone, tablet, virtual reality, wearable sensor, robot, or any other digital technology for physical rehabilitation, care aid, or cognitive training were defined as DHIs in this review. DHIs were eligible regardless of setting (eg, used in medical institution vs at home), mode of interaction and supervision, intervention content, or intervention duration. Although digital technology also contributes to the application of deep brain stimulation, transcranial direct current stimulation, and repetitive transcranial magnetic stimulation, these interventions with specific therapeutic aims were not in the scope of this review. DHIs solely used for monitoring parameters related to the disease symptoms were not included. Third, for comparison, comparators were categorized into passive control and active control. Passive control refers to waitlist control or no intervention. Active control included traditional, nontechnological caregiving methods, such as conventional physical rehabilitation, standard care aid, or drug treatments. Studies comparing the effectiveness of different modes of DHIs were excluded. Fourth, for outcome, we considered any outcomes related to motor symptoms, nonmotor symptoms, and quality of life in patients with PD as eligible to provide a comprehensive summary. Studies that only reported the reach, fidelity, or feasibility of the DHIs in PD were also included. Fifth, for study type, only RCT studies were included in this review. Studies with fewer than 5 participants in each group were excluded from the quantitative meta-analysis to ensure the robustness of the results.

The entire process of study selection was performed by 2 researchers independently, and disagreements were resolved with the involvement of a third researcher. Interrater reliability was assessed using Cohen κ [35], which was 0.831 for full-text review, indicating a high agreement between the 2 researchers.

Data Extraction

One reviewer extracted data using a comprehensive data extraction form: publication and author details (title, year, first author's name, and country), study details (study design, study setting, sample size, dropout, and follow-up duration), participants' characteristics (age, sex, and disease severity), intervention information (type of intervention, purpose of the intervention, intervention duration, and supervision mode), comparison information (type of comparator), and outcomes (scales used for assessing outcomes, reach, fidelity, and feasibility of interventions). The extracted data were checked by a second reviewer. Discrepancies in data extraction between the 2 researchers were resolved through discussion with a third reviewer.

Types of DHIs

DHIs were classified into three categories: (1) digital databases, (2) online classes, and (3) technology-based rehabilitation devices, based on the interaction modalities influenced by the technical implementation and the level of supervision inherent to the intervention, following the classification framework recommended by a previous review [36].

Digital databases are platforms that store health- or disease-related information in the form of videos or textual materials and offer self-paced access to predesigned resources, characterized by asynchronous interaction and a minimal level of supervision. Online classes involve health management and telerehabilitation sessions delivered in real time through telephone or video conferencing, offering synchronous interaction and direct supervision by professionals remotely and allowing for immediate feedback and guidance. Technology-based rehabilitation devices encompass a broad range of technologies, including exergaming, virtual reality, robotics, and mobile apps, which support physical rehabilitation or cognitive training. These devices involve close interactions with users, who respond to guidance and instructions provided by the DHIs, which in turn provide cues and feedback based on the users' actions. The level of supervision in these interventions varies depending on the technical features of the devices (eg, sensor integration) and the study design (eg, use under professional supervision).

XSL•FO
RenderX

Additionally, DHIs were categorized into physical rehabilitation, cognitive training, and care aid according to the intervention purpose.

Outcome Measures

The primary outcomes of the review were changes in symptom severity, including motor symptoms, cognitive function, psychiatric symptoms, overall nonmotor symptoms, and quality of life from baseline to postintervention and last follow-up. The scales used for assessing the severity of each symptom and quality of life depended on the original study. When one study utilized multiple assessment scales for one type of syndrome (eg, Parkinson's Disease Questionnaire-8 and Parkinson's Disease Questionnaire-39 for quality of life), the most frequently used one among all included studies was selected for analysis to reduce potential heterogeneity between studies. The assessment scales for each outcome used in each study are summarized in Multimedia Appendix 5.

The secondary outcomes included percentage reach, fidelity, dropout rate, and feasibility. According to the Medical Research Council process evaluation framework and Proctor implementation outcomes, we used the more common term reach to replace penetration, which is defined as the integration of a practice within a service setting and its subsystems [37,38]; therefore, we calculated the proportion of people who come into contact with the intervention. Fidelity was defined as the degree to which an intervention was implemented as prescribed in the original protocol or as intended by the program developers [38], and we extracted the fulfillment of the intervention. Feasibility was defined as the extent to which a new treatment or innovation can be successfully used or carried out within a given agency or setting [38] and can be measured by retention, acceptability, adherence, satisfaction, preference, safety, and cost-effectiveness [39]. Feasibility was summarized according to the definition of each study, and the aforementioned items were extracted even if the study did not discuss feasibility.

Quality Assessment

The revised Cochrane Risk of Bias 2 tool [40] was used to assess study quality across the following domains: randomization process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported result. Studies were categorized as low risk, some concerns, or high risk. We used the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) criteria [41] to assess the certainty of the overall evidence, which incorporates 5 key considerations: study limitations, inconsistency of effects, indirectness, imprecision, and publication bias. Any deviation across these 5 domains resulted in downgrading of the quality of evidence. The overall certainty of evidence was classified as high, moderate, low, or very low. All quality assessments were independently conducted by 2 reviewers, and any discrepancies were resolved through discussion.

Meta-Analysis Methods

Pairwise meta-analyses were performed to evaluate the effect of DHIs on the primary outcomes when more than 5 effect estimates were available in the analysis. The means and SDs of within-group differences from baseline to postintervention and follow-up assessments were used. If means and SDs were not reported in a study, the corresponding author was contacted via email to obtain the data. Otherwise, means and SDs were calculated based on available data using recommended formulas [42,43].

Standardized mean differences (SMDs; Hedges g) at postintervention and the last follow-up assessment were used as the effect measure, allowing comparison of data from different assessment scales. In studies with more than 2 intervention groups, the intervention group demonstrating the most favorable effect was selected for the main analysis as it represented the most well-designed DHIs. In studies featuring multiple control groups, the most active control group was chosen to adopt a more conservative analytical approach.

Considering that different assessment scales may reflect symptom severity in opposite directions, data from the original studies for the same outcome were standardized based on the most commonly used scale. For instance, while higher scores on the Unified Parkinson's Disease Rating Scale Part 3 indicate more severe motor symptoms, higher scores on the 6-Minute Walk Test suggest milder symptoms. Therefore, the results of the 6-Minute Walk Test were inverted (multiplied by –1) to align with the Unified Parkinson's Disease Rating Scale Part 3 for the analysis.

Due to the presence of potential heterogeneity, the random-effects model and the DerSimonian-Laird estimator were applied to calculate pooled estimates across studies. The Hartung-Knapp-Sidik-Jonkman method was applied to increase the robustness of the results [44]. Forest plots were generated to visually display the effect size and 95% CIs for individual studies and the overall pooled estimate. To present the expected range of true effects in similar studies, the 95% prediction intervals (PIs) were also calculated and displayed in the forest plot [45]. When the number of studies was fewer than 10, the method proposed by Nagashima et al [46] was applied to adjust the 95% PI. The 95% CI reflects the precision of the estimated overall average effect, whereas the 95% PI estimates the distribution of effects within which the true effect size is expected to fall in future similar studies. These 2 intervals complement each other to provide comprehensive information for result interpretation.

Subgroup analyses were conducted based on the intervention type (technology-based rehabilitation devices, online classes, and digital databases), considering the different inherent characteristics of various DHIs. Pairwise comparisons between subgroups were performed using random-effects meta-analysis with Hartung-Knapp adjustment and were limited to subgroups with 3 or more studies. Funnel plots and Egger regression intercept were used to assess small-study effects, which may indicate the potential publication bias. A sensitivity analysis was conducted to evaluate whether the overall results were statistically and significantly affected by 1 individual study using the leave-one-out analysis method. In addition, another sensitivity analysis restricting the synthesis to studies with a low risk of bias was performed for outcomes for which at least 10 such studies were available. All meta-analyses were

XSL•FO
RenderX

conducted using R (version 4.2.3; R Foundation for Statistical Computing).

Heterogeneity and Meta-Regression

The presence of heterogeneity among studies was assessed using Cochran Q test and the I^2 statistic, whereas 95% PI, τ, and τ^2 were calculated to evaluate the magnitude of heterogeneity [45]. By estimating the expected range of true effects for future studies, the 95% PI directly evaluates the variability of the intervention effect across diverse settings [47]. τ represents the estimated SD of the between-study effects, and τ^2 denotes the variance, which quantifies the spread of the true study effects around the mean [48].

Because substantial heterogeneity was anticipated to result from various types of DHIs in the analysis, univariate meta-regression analyses were performed using the mixed-effects model with the DerSimonian-Laird estimator for between-study variance and the Hartung-Knapp-Sidik-Jonkman method for SE adjustment to identify potential moderators that could account for heterogeneity. These potential moderators included individual-level variables (mean age, sex, and disease severity) and study-level variables (publication years, country income level, intervention type, intervention purpose, intervention setting, type of control, intervention duration, and supervision mode). In addition, bubble plots were provided as visual representation of the meta-regressions.

Results

Study Search and Selection

The search initially yielded 30,918 related study records from 6 databases in November 2023, among which 12,909 duplicate records were first removed during screening. The titles and abstracts of the remaining 18,009 records were screened, and the full texts of 284 studies were further assessed. A total of 90 studies from 91 articles evaluating the effect of DHI in patients with PD were included.

A supplementary search was conducted on January 5, 2025, and 4179 new publications were identified using the same keywords and databases as the initial search. After removing duplicates, a total of 3019 unique records were screened based on title and abstract, and the full texts of 66 potentially eligible studies were then reviewed. A total of 11 studies meeting the inclusion criteria were included.

The final search was performed on November 22, 2025, to ensure the inclusion of the most recent evidence and identified 3980 new publications. Following deduplication, a total of 2481 unique records were screened by title and abstract. Subsequently, a total of 44 studies underwent full-text review, resulting in the inclusion of 11 studies that met the eligibility criteria. During the full-text screening process, no studies were identified that appeared to meet the inclusion criteria but were ultimately excluded; all exclusions were straightforward decisions based on eligibility criteria.

Ultimately, a total of 112 unique studies from 115 articles [27,30,31,49-158] were eligible for the systematic review, and 97 unique studies from 98 articles were included in the quantitative meta-analysis. Of these 112 studies, 73 evaluated motor symptoms, 8 evaluated overall nonmotor symptoms, 26 evaluated cognitive function, 30 evaluated psychiatric symptoms, and 41 evaluated quality of life. The PRISMA flow diagram of the study search and selection is presented in Figure 1.

Figure 1. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for study selection.

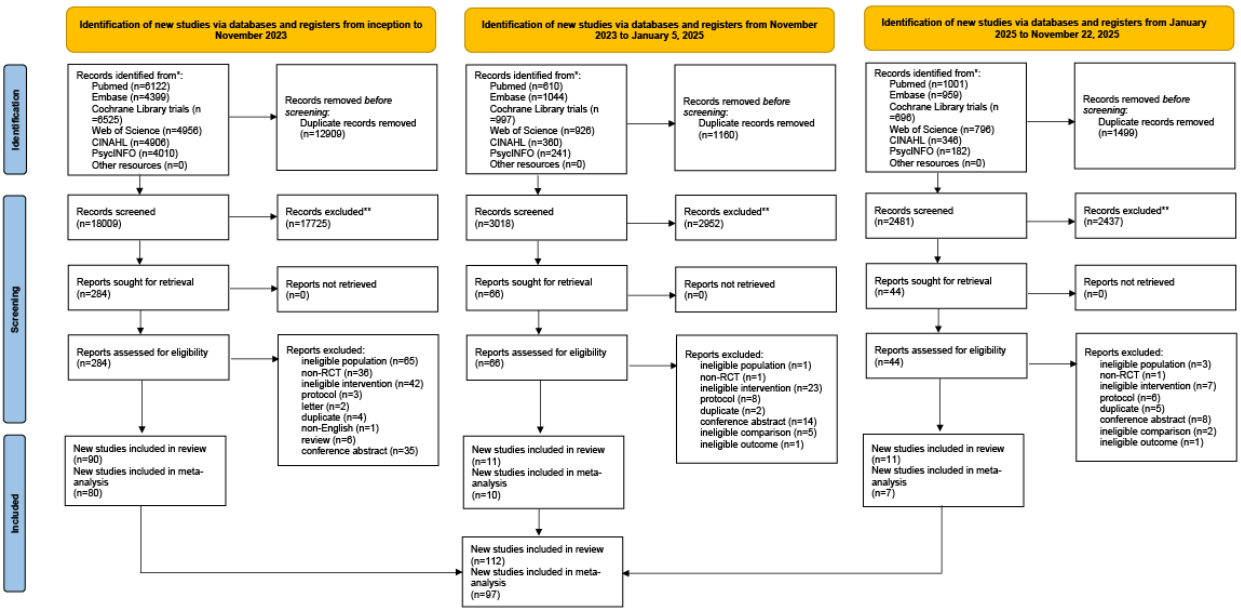

XSL•FO
RenderX

Study Characteristics

A total of 112 included studies comprising 5594 participants were identified, among which 84 studies (4658 participants) were performed in high-income countries, whereas only 28 studies (936 participants) were performed in low- and middle-income countries (LMICs). According to the predetermined classification criteria, the DHIs used in 91 studies were categorized as technology-based rehabilitation devices, 17 as online classes, and 8 as digital databases. The intervention purposes of 74 studies were physical rehabilitation, 28 were cognitive training, and 17 were care aid.

Active comparators (eg, traditional and nontechnological caregiving methods) were used in 91 trials, whereas others used passive comparators (eg, waitlist control or no intervention). The study setting of 51 studies was at home, and that of 55 studies was in medical institutions. Participants in the study by Halpern et al [88] first received intervention in the clinic for 9 sessions and then at home.

Regarding delivery mode, most interventions were delivered through computers and virtual reality devices; other modes included wearable sensors, robots, tablets, smartphones, and others. Most interventions lasted for at least 4 weeks, whereas 4 did not [51,54,92,115]. Among the 39 (34.8%) studies that reported follow-up assessments, the follow-up duration ranged from 4 to 24 weeks in 38 studies, except for the study by Yang et al [143], which implemented a shorter 2-week follow-up period.

A total of 97 of 112 studies were included in the meta-analysis, with the remaining 15 retained in the systematic review because of unavailable outcome data. A comprehensive summary of the characteristics of the included studies is presented in Multimedia Appendix 6.

Quality of Studies

The methodological limitations assessed by the Cochrane Risk of Bias 2 tool revealed that ≥20% of the included studies exhibited some concerns or high risk in the following domains (Figure 2): randomization process (postintervention: 53/180, 29% and follow-up: 11/55, 20%), deviations from intended interventions (postintervention: 95/180, 53% and follow-up: 19/55, 34%), missing outcome data (postintervention: 46/180, 26% and follow-up: 13/55, 24%), and selection of reported results (postintervention: 96/180, 53% and follow-up: 33/55, 60%). The main sources of bias were inadequate reporting of randomization methods (eg, lack of explicit allocation concealment) and failure to blind participants or caregivers in trials. Regarding motor symptoms, 12 studies were assessed as having a low risk of bias, 43 studies as having some concerns, and 18 studies as having a high risk of bias. Regarding cognitive function, 6 studies were assessed as having a low risk of bias, 16 studies as having some concerns, and 4 studies as having a high risk of bias. Regarding psychiatric symptoms, 4 studies were assessed as having a low risk of bias, 15 studies as having some concerns, and 10 studies as having a high risk of bias. Regarding overall nonmotor symptoms, 5 studies were assessed as having some concerns and 3 as having a high risk of bias. Regarding quality of life, 5 studies were assessed as having a low risk of bias, 22 studies as having some concerns, and 14 studies as having a high risk of bias (Multimedia Appendix 7).

XSL•FO
RenderX

Figure 2. Risk of bias across effect estimates at postintervention and follow-up assessments based on the Cochrane Risk of Bias 2 tool for randomized trials.

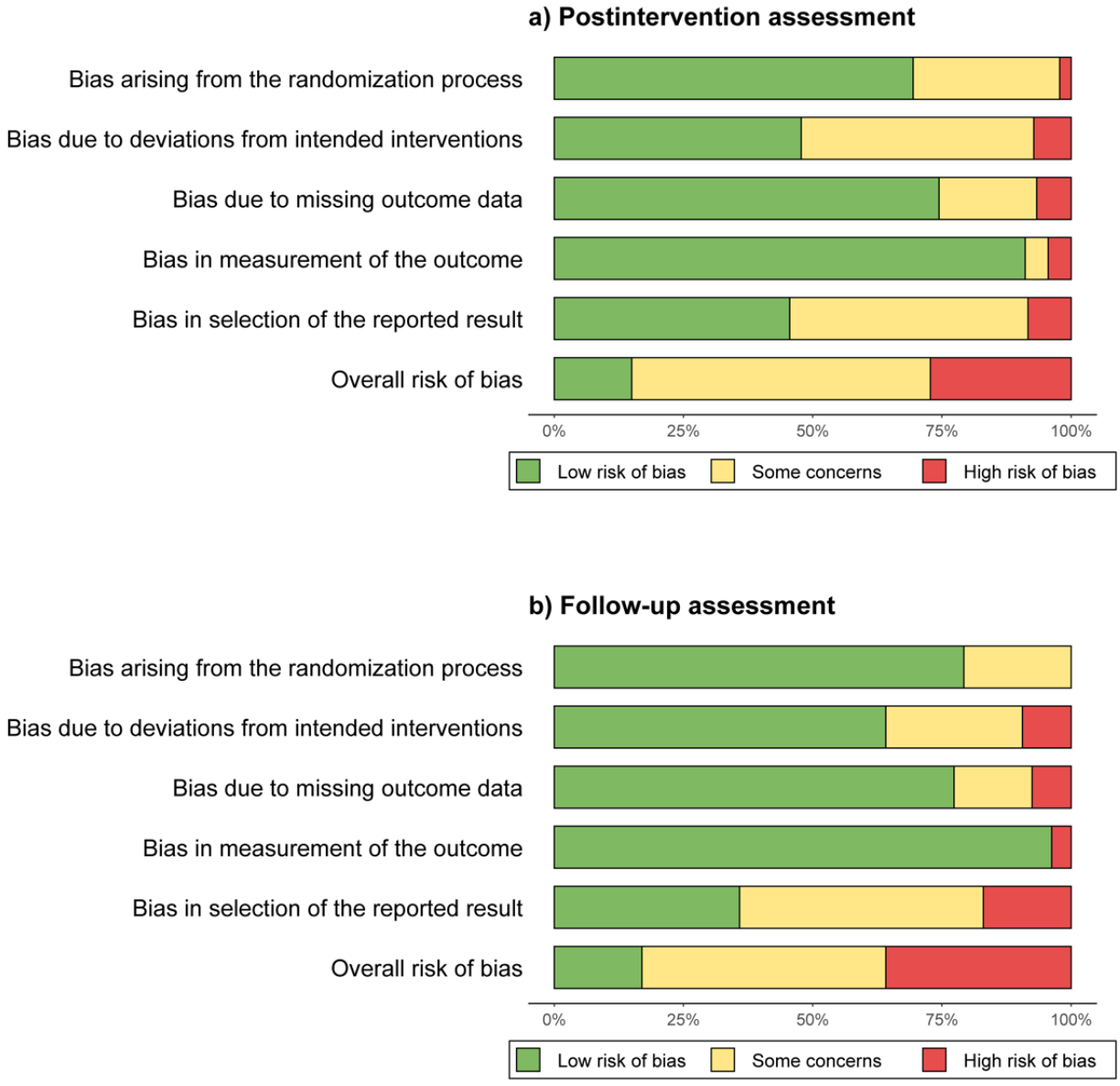

Effect of DHIs on PD Across Different Symptoms

Motor Symptoms

DHIs significantly improve motor symptoms (SMD=–0.39, 95% CI –0.60 to –0.18), while the 95% PI (–1.75 to 0.99) indicates that the true effect is uncertain and could range from a substantial benefit to being inferior to the control condition in future implementations (Figure 3 [27,31,49-53,56-59,61,62,65,68,69,71,72,74-79,82,84-87,89-97,99,100,103,106,110-113,115-117,119,120,122-124,126-132,135,136,141-146,151-153,156]). Although the intervention group had fewer symptoms on average (SMD=–0.39), the PI suggests considerable variability in future outcomes. DHIs may improve symptoms (SMD=–1.75) in some populations, while in other populations DHIs may be less effective than control groups (SMD=0.99). Also, DHIs may not affect other populations since the PI contains the 0 value. However, only 1 of 73 included studies suggested that DHIs were inferior to standard training, while both led to improvements in motor symptoms [89]. Subgroup analyses revealed no statistically significant differences in pairwise comparisons among the 3 intervention types (Multimedia Appendix 8). Only technology-based rehabilitation devices significantly improve motor symptoms (SMD=–0.36, 95% CI –0.59 to –0.13, 95% PI –1.80 to 1.07). Neither online classes nor digital databases exhibited favorable effects, with SMD=–0.19 (95% CI –0.87 to 0.48) for online classes, and SMD=–0.76 (95% CI –1.85 to 0.33) for digital databases. However, the subgroup analysis revealed no statistically significant difference in effects across intervention types (P=.39). Significant heterogeneity was found (I^2=80.3%, P<.001), and the heterogeneity was substantial (τ=0.680 and

XSL•FO
RenderX

τ^2=0.4623). The intervention types did not completely explain the significant heterogeneity, with I^2=81.3% for technology-based rehabilitation devices (P<.001) and I^2=83% for digital databases (P<.001). In the follow-up analysis, DHIs also exhibited significant improvement in the motor symptoms of patients with PD (Multimedia Appendix 9). The leave-one-out analysis obtained a consistent result. The sensitivity analysis of 12 low-risk-of-bias studies showed a significant but modest benefit of DHIs on motor symptoms (SMD=–0.22, 95% CI –0.40 to –0.04), with negligible heterogeneity (I^2=0.6%; τ=0.025; τ^2=0.0006; 95% PI –0.43 to –0.01). The 95% PI indicated a high certainty that the intervention will yield at least a small beneficial effect in future similar study settings. GRADE ratings indicated that the certainty of the evidence for the effect of DHIs on motor symptoms of patients with PD was very low owing to the high risk of bias of included studies, high heterogeneity, and potential publication bias (Figure 4).

Figure 3. Forest plot of the meta-analysis of digital health interventions' efficacy on postintervention motor symptoms [27,31,49-53,56-59,61,62,65,68,69,71,72,74-79,82,84-87,89-97,99,100,103,106,110-113,115-117,119,120,122-124,126-132,135,136,141-146,151-153,156]. Standardized mean differences (SMDs) with 95% CIs were calculated using a random-effects model. Negative SMD values favor the experimental intervention. Subgroup analyses were performed by intervention type (technology-based rehabilitation devices, online classes, digital databases). The square represents the individual study estimate. The rhombus shape represents the pooled estimates of the lowest accuracy in all studies. *P* values are shown as p.

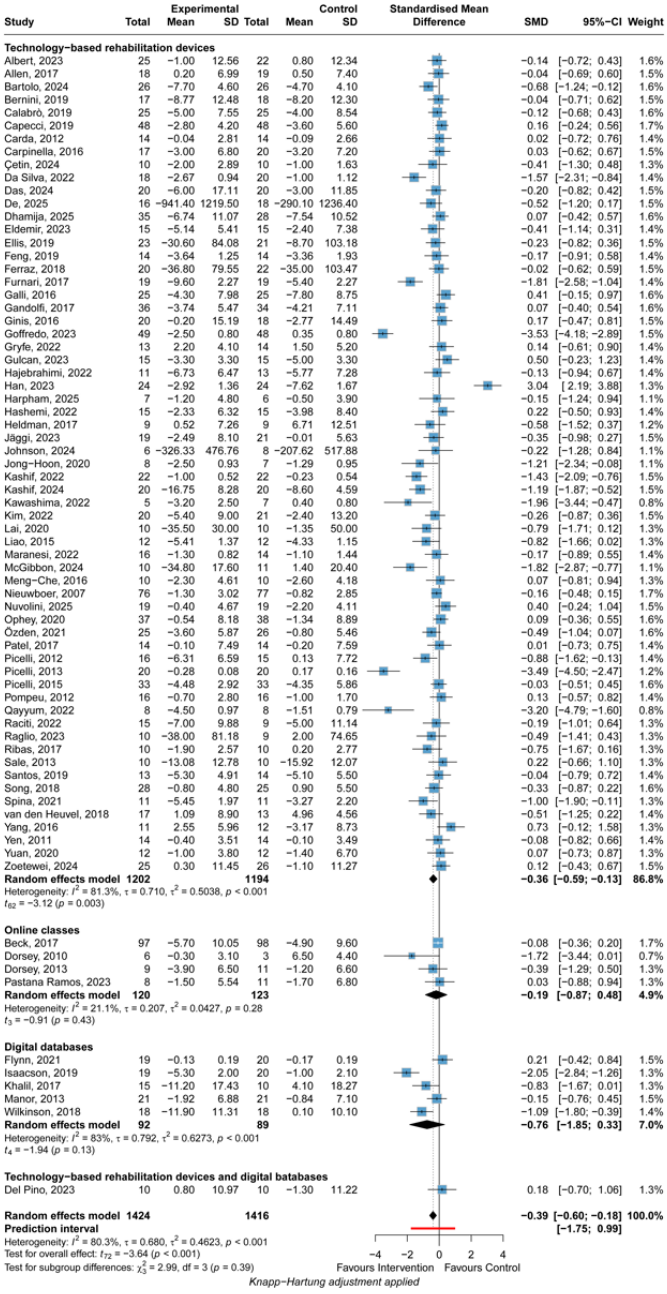

XSL•FO
RenderX

Figure 4. GRADE (Grading of Recommendations Assessment, Development, and Evaluation) ratings at postintervention assessments.

Digital health intervention compared to convetional intervention/usual care/waitlist group for people with Parkinson's disease
Bibliography:

Certainty assessment							Summary of findings				
							Study event rates (%)			Anticipated absolute effects	
Participants (studies) Follow-up	Risk of bias	Inconsistency	Indirectness	Imprecision	Publication bias	Overall certainty of evidence	With convetional intervention/usual care/waitlist group	With digital health intervention	Relative effect (95% CI)	Risk with convetional intervention/usual care/waitlist group	Risk difference with digital health intervention
Motor symptoms											
2840 (73 RCTs)	serious[a]	very serious[b]	not serious	not serious	publication bias strongly suspected[c]	⨁◯◯◯ Very low[a,b,c]	1416	1424	-	-	SMD **0.39 SD lower** (0.6 lower to 0.18 lower)
Psychiatric symptoms											
1741 (30 RCTs)	serious[d]	very serious[e]	not serious	not serious	none	⨁◯◯◯ Very low[d,e]	871	870	-	-	SMD **0.42 SD lower** (0.74 lower to 0.09 lower)
Cognitive function											
1148 (26 RCTs)	not serious	serious[f]	not serious	not serious	publication bias strongly suspected[g]	⨁⨁◯◯ Low[f,g]	568	580	-	-	SMD **0.47 SD higher** (0.22 higher to 0.72 higher)
Overall Non-motor symptoms											
572 (8 RCTs)	serious[h]	not serious	not serious	serious[i]	none	⨁⨁◯◯ Low[h,i]	294	278	-	-	SMD **0.26 SD lower** (0.49 lower to 0.03 higher)
Quality of life											
1929 (41 RCTs)	serious[j]	very serious[k]	not serious	serious[l]	none	⨁◯◯◯ Very low[j,k,l]	970	959	-	-	SMD **0.19 SD lower** (0.47 lower to 0.09 higher)

CI: confidence interval; **SMD:** standardised mean difference

Explanations

a. 18 of 73 studys are rated at overall high risk with some limitations for multiple criteria.
b. I^2 = 80.3%, τ^2 = 0.462, P<0.001, 95% prediction interval [-1.75, 0.99]
c. Egger's test indicated potential small-study effect. (P = 0.013)
d. 9 of 30 studys are rated at overall high risk with some limitations for multiple criteria.
e. I^2 = 85.4%, τ^2 = 0.451 P<0.001, 95% prediction interval [-1.82, 0.99]
f. I^2 = 63.5%, τ^2 = 0.172, P<0.001, 95% prediction interval [-0.41, 1.35]
g. Egger's test indicated potential small-study effect. (P = 0.002)
h. 3 of 8 studys are rated at overall high risk, with only one study being rated at overall low risk.
i. Small sample size.
j. 14 of 41 studys are rated at overall high risk with some limitations for multiple criteria.
k. I2 = 81.2%, τ^2 = 0.405, p <0.001, 95% prediction interval [-1.50; 1.12]
l. 95% Confidence Interval crosses null effect

Overall Nonmotor Symptoms

DHIs significantly improved overall nonmotor symptoms (SMD=–0.26, 95% CI –0.49 to –0.03, 95% PI –0.56 to 0.03; I^2=0; Figure 5 [30,52,65,91,100,120,134,152]). No significant heterogeneity was observed for the analysis (I^2=13.8%; P=.32). The leave-one-out analysis did not obtain a consistent result after omitting the studies by Dhamija et al [152], and So et al [134], suggesting the instability of this finding. GRADE ratings indicated that the certainty of the evidence for the effect of DHIs on overall nonmotor symptoms of patients with PD was low owing to the high risk of bias of included studies and serious imprecision (Figure 4).

XSL•FO
RenderX

Figure 5. Forest plot of the meta-analysis of digital health interventions' efficacy on postintervention overall nonmotor symptoms [30,52,65,91,100,120,134,152]. Standardized mean differences (SMDs) with 95% CIs were calculated using a random-effects model. Negative SMD values favor the experimental intervention. Subgroup analyses were performed by intervention type (technology-based rehabilitation devices, online classes, digital databases). The square represents the individual study estimate. The rhombus shape represents the pooled estimates of the lowest accuracy in all studies. *P* values are shown as p.

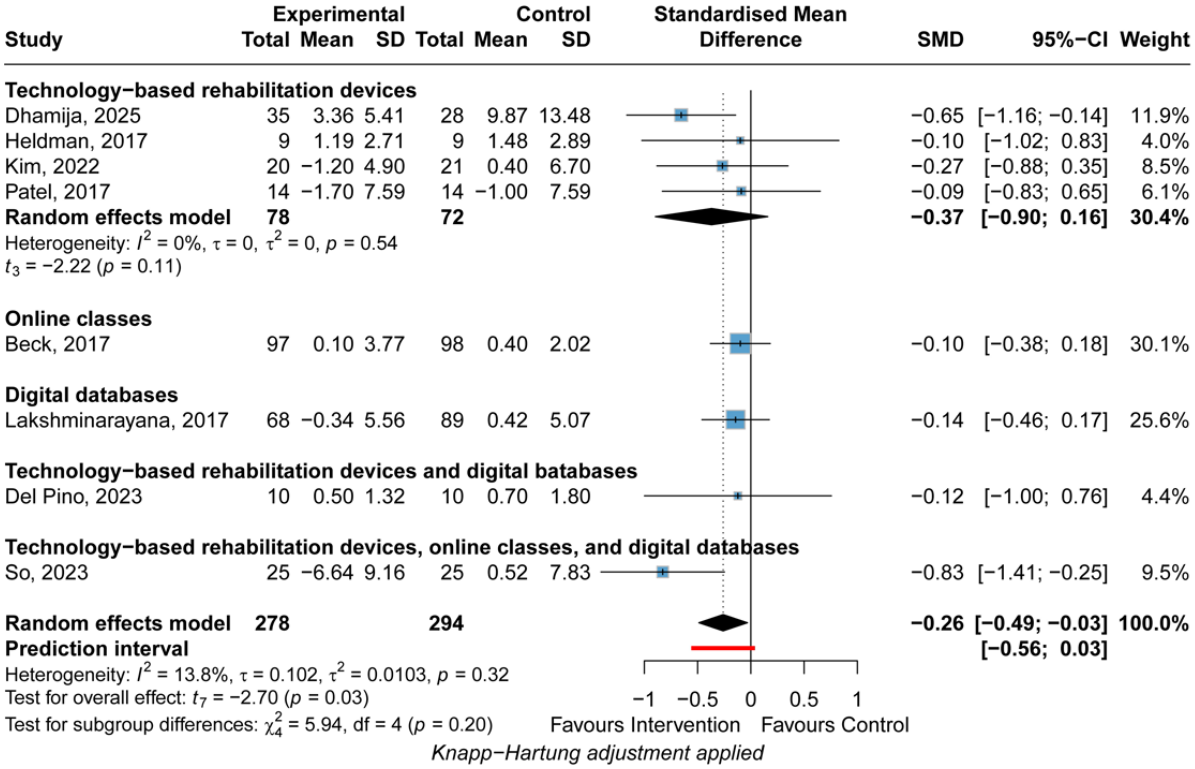

Psychiatric Symptoms

DHIs significantly improved psychiatric symptoms (SMD=–0.42, 95% CI –0.74 to –0.09), while the 95% PI (–1.82 to 0.99) indicates that the true effect is uncertain and could range from a substantial benefit to being inferior to the control condition in future implementations (Figure 6 [30,52,53,55,57,63,66-68,73,75,77,82,85,87,102,107,109,111,117,118,120,125,137,139,141,142,145,157,158]). Although the intervention group had fewer symptoms on average (SMD=–0.42), the PI suggests considerable variability in future outcomes. DHIs may improve symptoms (SMD=–1.82) in some populations, while in other populations DHIs may be less effective than control groups (SMD=0.99). Also, DHIs may not affect other populations since the PI contains the 0 value. However, only 1 of 30 included studies suggested that DHIs were significantly harmful to psychiatric symptoms. Subgroup analyses revealed no statistically significant differences in pairwise comparisons among the intervention types (Multimedia Appendix 8). No types of DHIs exhibited significantly favorable effects in the subgroup analysis, with SMD=–0.49 (95% CI –0.98 to 0.01) for technology-based rehabilitation devices, SMD=–0.05 (95% CI –1.89 to 1.79) for digital databases, and SMD=–0.42 (95% CI –0.98 to 0.13) for online classes. Significant heterogeneity was found (I^2=85.4%; P<.001), and the heterogeneity was substantial (τ=0.671; τ^2=0.4506; 95% PI –1.82 to 0.99). The intervention types could not completely explain the significant heterogeneity, with I^2=88.1% for technology-based rehabilitation devices (P<.001) and I^2=80% for online classes (P<.001). In the follow-up analysis, DHIs did not exhibit significant improvement in the psychiatric symptoms of patients with PD (Multimedia Appendix 9). The leave-one-out analysis obtained a consistent result. GRADE ratings indicated that the certainty of the evidence for the effect of DHIs on psychiatric symptoms of patients with PD was very low owing to the high risk of bias of included studies and high heterogeneity (Figure 4).

XSL•FO
RenderX

Figure 6. Forest plot of the meta-analysis of digital health interventions' efficacy on postintervention psychiatric symptoms [30,52,53,55,57,63,66-68,73,75,77,82,85,87,102,107,109,111,117,118,120,125,137,139,141,142,145,157,158]. Standardized mean differences (SMDs) with 95% CIs were calculated using a random-effects model. Negative SMD values favor the experimental intervention. Subgroup analyses were performed by intervention type (technology-based rehabilitation devices, online classes, digital databases). The square represents the individual study estimate. The rhombus shape represents the pooled estimates of the lowest accuracy in all studies. *P* values are shown as p.

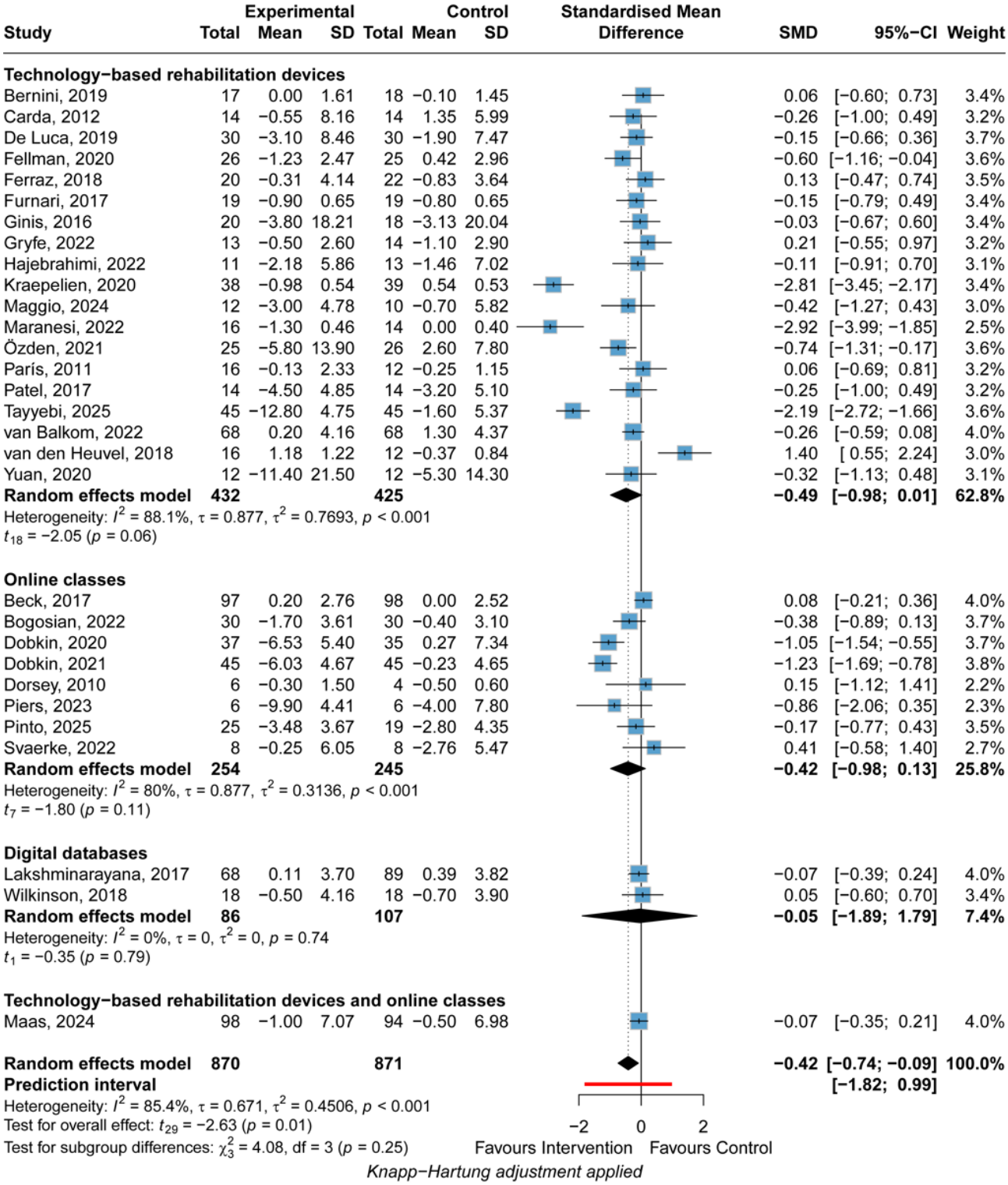

Cognitive Function

DHIs significantly improve cognitive function (SMD=0.47, 95% CI 0.22 to 0.72), while the 95% PI (–0.41 to 1.35) indicates that the true effect is uncertain and could range from a substantial benefit to being inferior to the control condition in future implementations (Figure 7 [50,52-54,59,62,65,68,73,80,82,85,87,92,108,109,112,116,118,126,135,137,139,140,155,156]). Although the intervention group had fewer symptoms on average (SMD=0.47), the PI suggests considerable variability in future outcomes. DHIs may improve symptoms (SMD=1.35) in some populations, while in other populations DHIs may be less effective than control groups (SMD=–0.41). Also, DHIs may

XSL•FO
RenderX

not affect other populations since the PI contains the 0 value. However, no included studies suggested that DHIs were significantly harmful to cognitive function. Technology-based rehabilitation devices demonstrated significantly greater improvement in cognitive function compared with online classes (SMD=0.39, 95% CI 0.01 to 0.77; *P*=.049; Multimedia Appendix 8). Technology-based rehabilitation devices exhibited significant improvement in cognitive functions of patients with PD (SMD=0.53, 95% CI 0.24 to 0.82), but not online classes (SMD=0.14, 95% CI –0.44 to 0.72). Significant heterogeneity was found (I^2=63.5%; *P*<.001), and the heterogeneity was substantial (τ=0.415 and τ^2=0.1723). The intervention types could not explain the significant heterogeneity, with I^2=68.2% for technology-based rehabilitation devices (*P*<.001). In the follow-up assessment, DHIs also exhibited significant improvement in the cognitive function of patients with PD (Multimedia Appendix 9). The leave-one-out analysis obtained a consistent result. GRADE ratings indicated that the certainty of the evidence for the effect of DHIs on the cognitive function of patients with PD was low owing to high heterogeneity and potential publication bias (Figure 4).

Figure 7. Forest plot of the meta-analysis of digital health interventions' efficacy on postintervention cognitive function [50,52-54,59,62,65,68,73,80,82,85,87,92,108,109,112,116,118,126,135,137,139,140,155,156]. Standardized mean differences (SMDs) with 95% CIs were calculated using a random-effects model. Negative SMD values favor the experimental intervention. Subgroup analyses were performed by intervention type (technology-based rehabilitation devices, online classes, digital databases). The square represents the individual study estimate. The rhombus shape represents the pooled estimates of the lowest accuracy in all studies. *P* values are shown as p.

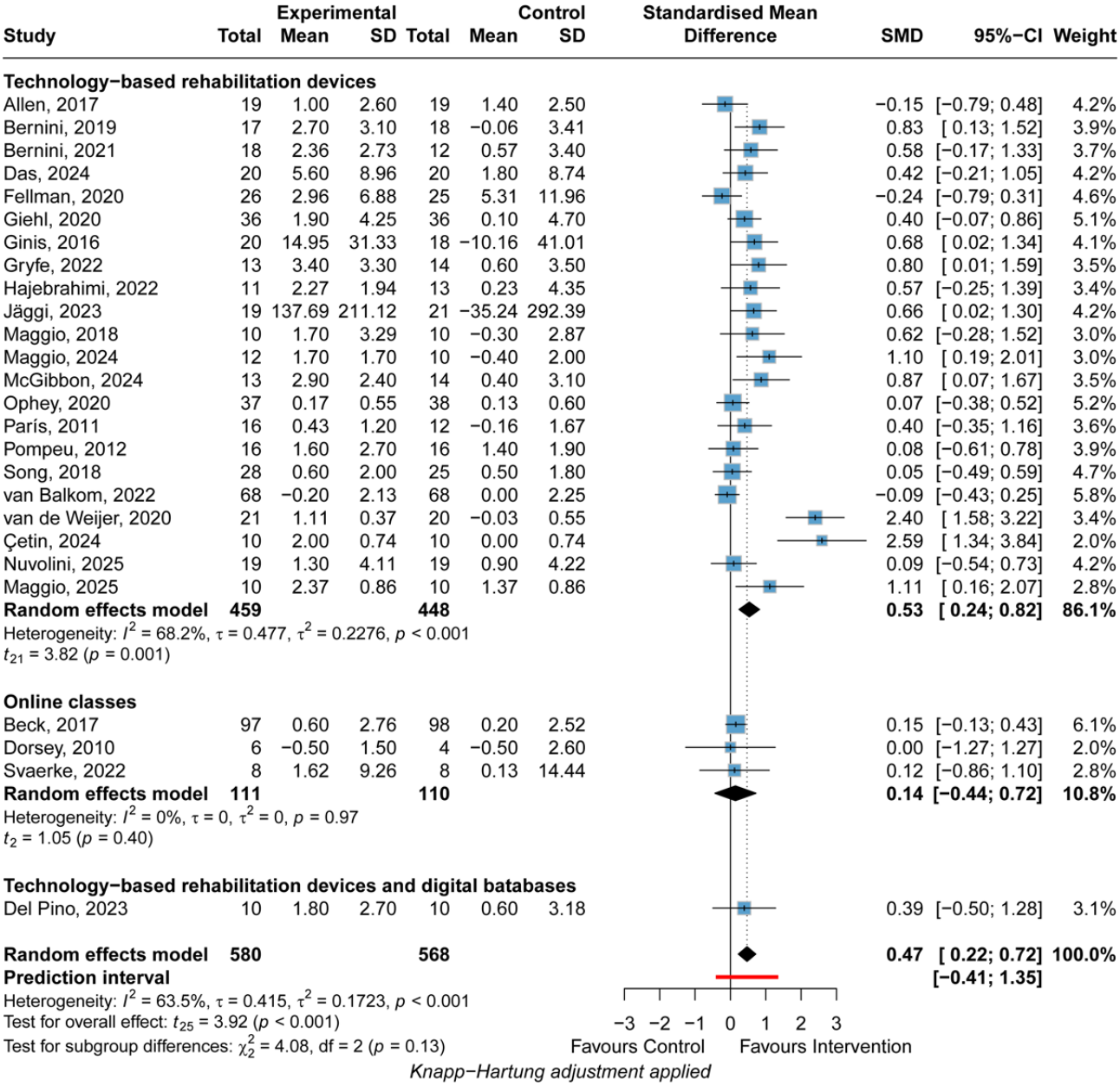

XSL•FO
RenderX

Quality of Life

DHIs did not improve quality of life in patients with PD (SMD=–0.19, 95% CI –0.47 to 0.09), and the 95% PI (–1.50 to 1.12) also indicates the uncertain true effect (Figure 8 [27,30,31,50,52,58,59,61,62,65,68,69,71,72,75,79,83,85,87,91,93,97,102,107,111,115,118-121,129,132,134,136-138,141,143,152,155,157]). Subgroup analyses revealed no statistically significant differences in pairwise comparisons among the intervention types (Multimedia Appendix 8). The analysis of each intervention type obtained a consistent result, with SMD=–0.13 (95% CI –0.52 to 0.25) for technology-based rehabilitation devices, SMD=–0.24 (95% CI –0.77 to 0.29) for online classes, and SMD=–0.19 (95% CI –9.86 to 9.49) for digital databases. Significant heterogeneity was found (I^2=81.2%; P<.001), and the heterogeneity was substantial (τ=0.634; τ^2=0.4053; 95% PI –1.50 to 1.12). The intervention types could not completely explain the heterogeneity, with I^2=84.1% for technology-based rehabilitation devices (P<.001) and I^2=53.6% for online classes (P=.04). The follow-up analysis indicated a consistent result (Multimedia Appendix 9). The leave-one-out analysis obtained a consistent result. GRADE ratings indicated that the certainty of the evidence for the effect of DHIs on quality of life of patients with PD was very low owing to the high risk of bias of included studies, high heterogeneity, and serious imprecision (Figure 4).

XSL•FO
RenderX

Figure 8. Forest plot of the meta-analysis of digital health interventions' efficacy on postintervention quality of life [27,30,31,50,52,58,59,61,62,65,68,69,71,72,75,79,83,85,87,91,93,97,102,107,111,115,118-121,129,132,134,136-138,141,143,152,155,157]. Standardized mean differences (SMDs) with 95% CIs were calculated using a random-effects model. Negative SMD values favor the experimental intervention. Subgroup analyses were performed by intervention type (technology-based rehabilitation devices, online classes, digital databases). The square represents the individual study estimate. The rhombus shape represents the pooled estimates of the lowest accuracy in all studies. *P* values are shown as p.

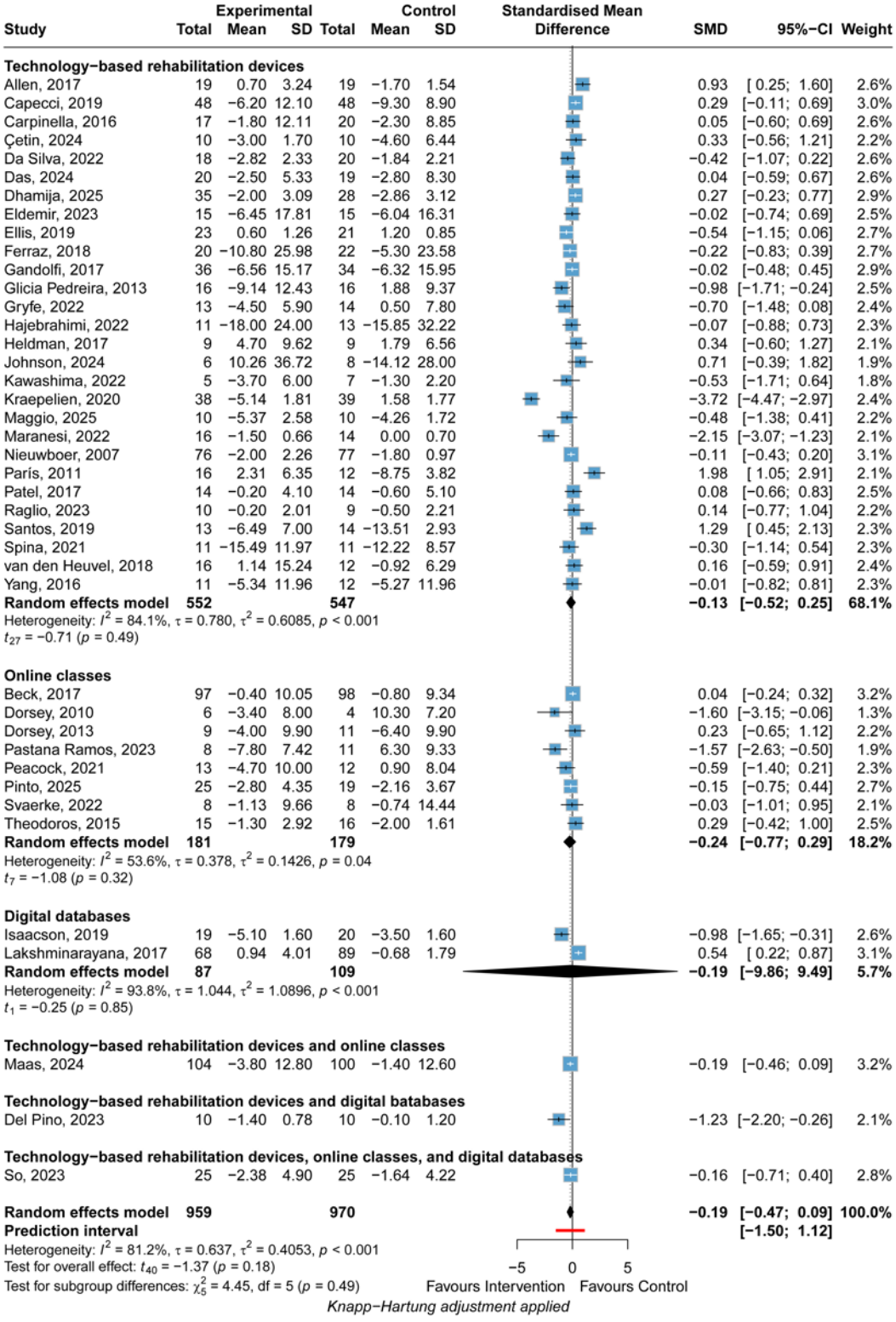

Publication Bias

Small-study effect was assessed via visual inspection of a funnel plot and Egger test (Multimedia Appendix 10). The Egger test indicated a significant small-study effect in the analysis of the effect of DHIs on motor (intercept=–2.07, 95% CI –3.69 to –0.45; *P*=.02) and cognitive function (intercept=2.54, 95% CI –1.12 to 3.97; *P*=.002) of patients with PD. Further, trim-and-fill analysis (Multimedia Appendix 10) was performed, and we observed that the effects were no longer significant for both motor symptoms (SMD=–0.04, 95% CI –0.29 to 0.21) and cognitive function (SMD=0.16, 95% CI –0.14 to 0.47).

XSL•FO
RenderX

Meta-Regressions

To explore potential sources of significant heterogeneity, we performed univariate meta-regression analyses examining 3 individual-level and 8 study-level variables (Multimedia Appendix 11), and we also present bubble plots to visualize the relationship involving the continuous variables (Multimedia Appendix 12). We conducted multivariable meta-regressions using the variables mentioned above as predictors of between-study variance when 2 or more variables accounted for R^2. In the analysis of cognitive functions, 2 of these study variables accounted for R^2: mean age of participants (model Q statistic [QM]=0.08; P=.78; R^2=0.49%) and intervention setting (QM=1.16; P=.29; R^2=8.88%), with a total R^2=1.38% (P=.53). There are 4 variables associated with the improvement of psychiatric symptoms with a total R^2=9.83% (P=.35): the percentage of female participants (QM=2.73; P=.11; R^2=7.42%), publication year (QM=3.39; P=.08; R^2=4.93%), country income level (QM=0.38; P=.54; R^2=2.64%), and intervention purpose (QM=1.44; P=.25; R^2=11.19%). As for motor symptoms, mean age of participants (QM=4.22; P=.04; R^2=2.96%) and supervision mode (QM=1.50; P=.22; R^2=3.96%) may explain part of the heterogeneity, with a total R^2=9.42% (P=.04). Supervision mode (QM=1.41; P=.25; R^2=10.38%) and mean age (QM=0.90; P=.35; R^2=.64%) moderated the effect of DHIs on quality of life with a total R^2 of 14.56% (P=.14). These meta-regression results should be interpreted as exploratory. The low R^2 values indicate that the factors account for only a small proportion of the total heterogeneity. It reminds us that the variability in intervention effects is likely due to a complex combination of clinical and methodological factors not fully captured here. Nevertheless, these findings generate valuable hypotheses for future research.

Reach, Fidelity, and Feasibility of DHIs

Most studies reported their eligible population and participant enrollment (Multimedia Appendix 13). A total of 94 studies reported on intervention reach with a median reach of 37.5% (range 4.5%-94.7%), and all included studies reported the randomly assigned populations, with a median of 50% (range 30.6%-100%) randomly assigned to the intervention groups. A total of 38 studies reported intervention fidelity, revealing a high degree of fidelity across a diverse range, of which 34 reported percentage fidelity with a median of 95.7% (range 26.7%-100%), and another 4 studies provided relevant information regarding fidelity. Dropout rates were reported in 105 studies, with a median of 9.1% (range 0%-61.1%).

A total of 32 studies reported that the interventions were feasible to deliver with high satisfaction, adoption, preference for interventions, high interest, high retention rates, convenience, high acceptance, reliability, safety, comfort, flexibility, high adherence, cost-effectiveness, user-friendliness, high recruitment, and high fidelity. Only 5 of these reports were from LMICs. A total of 35 studies did not mention feasibility directly, but they reported a high rate of usability, no serious adverse events, favorable satisfaction ratings, willingness to recommend to others, low cost, high participant interactivity, and low dropout rate. The remaining 44 studies did not report any data on feasibility.

Discussion

In contrast to previous reviews focusing on single DHI modalities or specific symptoms, this is the first systematic review and meta-analysis to comprehensively evaluate the effectiveness of all existing DHIs across motor symptoms, nonmotor symptoms, and quality of life in patients with PD, while providing a concurrent analysis of implementation feasibility. In this review, we systematically evaluated the effectiveness and implementation of currently reported DHIs in patients with PD, collecting data from 97 RCTs, including 4404 participants from 24 countries. Favorable postintervention effects of DHIs were found over control groups on motor symptoms and specific nonmotor domains (cognitive function, psychiatric symptoms, and overall burden), with the improvement of motor symptoms and cognitive function stable at follow-up assessments. However, substantial and unexplained heterogeneity was detected, and studies with a moderate to high risk of bias accounted for a high proportion; hence, the GRADE approach rated the overall certainty of evidence as low to very low. The wide 95% PIs indicate substantial heterogeneity, suggesting that the true effect remains uncertain. Future implementations might find DHIs produce strong benefits in some populations, yet be ineffective or even worse than control in others, with the results spanning from no effect to a clear positive or negative outcome. In addition, a high proportion of the included studies were assessed as having moderate to high risk of bias, which calls for more rigorously designed studies in the future. Therefore, clinicians should interpret these findings with caution. Additionally, we summarized the reach, fidelity, and feasibility of DHIs, which is valuable for the design of future trials as well as clinical decision-making regarding the application of DHIs. This review brings to the field an updated and integrative evidence that can inform clinical decision-making and guide the design of future DHIs. In real-world implications, our findings highlight that while DHIs hold promise for scalable and accessible PD management, their clinical application should be cautious, personalized, and supported by further high-quality evidence, especially in underserved regions.

Previously reported meta-analyses of the use of DHIs in PD mostly focus on the effects of virtual reality technology [159-167], and others focused on telemedicine [168-170], computerized cognitive training [20], robot-assisted training [112,171-174], mobile app [109,175], and wearable technology [51,176,177]. A recent umbrella review including 8 meta-analyses showed that virtual reality training significantly improved motor performance, mainly including balance ability and stride length in patients with PD [178]. A meta-analysis found that robot-assisted gait training was helpful for the improvement of motor function and balance function [179], while another showed that the efficacy was very uncertain [180]. All of the above meta-analyses were based on a specific type of technology, and their outcomes focused primarily on improvements in motor symptoms. A recent network

XSL•FO
RenderX

meta-analysis compared 4 technology-based interventions (internet-based, proprioceptive, robot, and virtual reality) and found that the virtual reality–based intervention fares the best in terms of improving motor symptoms and quality of life [181]. Notably, no study in the field of PD has included all RCTs that met the definition of DHIs and conducted a systematic synthesis. With technological advancements leading to a rapid expansion of DHI applications in PD, a comprehensive analysis of existing DHI modalities in PD is essential to establish an evidence-based framework for guiding future development.

Our review included all types of DHIs, systematically categorizing them by type, purpose, and implementation characteristics while focusing on the 5 key health outcomes in PD. Regarding motor symptoms and cognitive function, our findings are consistent with most previous meta-analyses that DHIs are effective compared with control groups. In addition, we identified significant small-study effects, a critical issue that has not been reported in prior studies. This may reflect publication bias, whereby smaller studies with null results remain unpublished; genuine heterogeneity between smaller and larger trials is also one of the important sources of small-study effects. We used the trim-and-fill method to estimate and adjust for the number and outcomes of missing studies, and the postintervention effects on both motor symptoms and cognitive function became statistically insignificant. This implies that small-study effects might have an impact on these outcomes, and the wide 95% PIs emphasize the uncertainty of the benefits of DHIs. These results highlight the need for further high-quality studies with larger sample sizes and rigorous designs to provide more definitive evidence for updated meta-analyses. For psychiatric symptoms, although the result was statistically significant, it must be interpreted with caution due to the very low certainty of evidence according to the GRADE framework. This low rating may be driven by several factors: the wide 95% PIs, substantial heterogeneity, and a high risk of bias across the included studies.

As for overall nonmotor symptoms, although the results were statistically significant, they should be considered with caution due to the small sample size. Notably, So et al [134] and Dhamija et al [152] were the only 2 trials among 8 to report significant improvements in nonmotor symptoms, and omitting either of them would result in a null effect of the result. It is potentially attributable to its multimodal design (face-to-face education, telecounseling, and smartphone or wearable tools) and early-stage PD cohort. Hence, we did not use the results as a point of reference. In terms of quality of life, unlike previous studies, we did not find a significant improvement. It may be because the majority of scales measuring quality of life are self-reported, such as the most widely used Parkinson's Disease Questionnaire-39. It lacks high sensitivity for detecting mild to moderate improvements in quality of life, particularly for patients with PD with milder symptoms (Hoehn and Yahr stage ≤3) [182]. In addition, most studies tend to consider quality of life as a secondary outcome, which may affect the design of DHIs. Moreover, complex or long-term interventions may impose an additional burden on patients. Standardized quality of life instruments evaluate a multidimensional construct including emotional well-being, social support, and communication abilities, but the most widely discussed applications of DHIs in PD were motor or cognitive training. Therefore, improvements in isolated domains (eg, mobility or cognitive function) may not substantially enhance overall quality of life. Future DHIs should adopt a user-centered design approach to improve accessibility, reduce treatment-related burdens, and optimize patient experience.

Of the included studies, the majority reported reach or fidelity, and some researchers further reported the feasibility of the intervention. Nearly half of the researchers mentioned that the intervention was feasible and provided quantitative data and qualitative interpretation. This implies that researchers are consciously valuing the feasibility of their interventions. However, the existing problem is the lack of a widely accepted framework for evaluating them, which makes comparing the feasibility of different DHIs across studies challenging. In this review, when analyzing feasibility, we extracted data on reach, fidelity, and dropout rates and summarized the authors' definitions and explanations related to the feasibility of DHIs as mentioned in the study. For those not directly mentioning feasibility, we also extracted information that may be relevant, including satisfaction, adverse effects, user-friendliness, and so on. Some studies reported high safety with few adverse events [61,85,114,119,133,149], as DHIs that provide real-time surveillance of patients can detect unexpected conditions promptly. Furthermore, physical therapy that incorporates digital technology can help prevent falls and freezing of gait [22] during the training process, such as through the use of robot-assisted gait training [179,183]. The reasons for patients dropping out may include the complexity of the intervention system, and face-to-face guidance before DHIs may increase adherence [184]. Future studies should take into account the aforementioned factors to design more feasible intervention protocols.

The included studies exhibited substantial heterogeneity. The primary source of heterogeneity may be that all DHIs were comprehensively summarized rather than a specific type. The types of DHIs range from the simplest intervention designs of only watching relevant videos to more complex designs involving specifically developed robotic devices and even using multiple digital tools simultaneously. Therefore, we conducted subgroup analyses according to the intervention type, yet substantial heterogeneity persisted within subgroups. Although only the technology-based rehabilitation devices subgroup of the 3 subgroups showed a statistically significant pooled effect on motor symptoms, the test for subgroup differences was not significant; therefore, these findings should not be interpreted as evidence that this subgroup is superior to the other 2 subgroups. Moreover, while the original RCTs may demonstrate that a specifically designed DHI is superior to the control group, they do not imply that individuals using different DHIs across trials were randomized. Therefore, such cross-trial comparisons are invalid for establishing causal relationships between different types of DHIs. Consequently, observed subgroup effects may reflect differences in study populations, intervention intensity, comparators, durations, or outcome instruments across studies, rather than genuine differences between the DHIs themselves. Meta-regression revealed that the percentage of female

XSL•FO
RenderX

participants, mean age, intervention setting, publication year, country income level, intervention purpose, and supervision mode might explain a modest portion of the I^2 heterogeneity. While we cannot fully account for all heterogeneity, this finding suggests that the effectiveness of DHIs may vary depending on individual patient characteristics and supervision mode. Therefore, comprehensive consideration of these factors is essential when selecting DHIs for patients. Moreover, the 95% PIs for every primary outcome were wide and consistently crossed the null value. This finding shows substantial heterogeneity and indicates that the effect of DHIs in future settings could range from being less effective than the control condition to being clearly beneficial due to this clinical and methodological diversity. It reminds us that the same intervention may not work equally well for every patient. Hence, clinicians should consider patient responses individually, and future research could aim to identify which types of DHIs are most suitable for specific patient groups.

Clinicians should interpret our findings and their applicability to local settings with caution, as 75% of participants were recruited from high-income countries. The effectiveness and feasibility of these DHIs may differ substantially in LMICs due to variations in resources, infrastructure, and cultural context. Future research in LMICs is critically needed to provide evidence that is generalizable to these settings. As a majority of the world's population has access to mobile technologies [185], the implementation of digital technology in LMICs could potentially help to close the treatment gap [186,187].

Strengths of this review include that only RCTs were included, which are of high evidence quality levels; the review was conducted strictly by the methodology outlined in the Cochrane guidelines; a large number of studies were included, enrolling a large population from 24 countries; the review encompassed a broad range of outcomes, including motor symptoms, overall nonmotor symptoms, cognitive functions, psychiatric symptoms, and quality of life; the Hartung-Knapp-Sidik-Jonkman method was applied to obtain more conservative results; and the 95% PI was provided to estimate the potential true effect.

This review also has limitations. First, this review comprehensively included various types of DHIs for analysis, leading to high heterogeneity in the analysis. We explored heterogeneity through prespecified subgroup analyses and meta-regression and identified 3 variables accounting for the heterogeneity significantly, although to a small degree. Second, we included a large number of studies, which resulted in a longer period from literature search to manuscript preparation. Therefore, secondary and third screenings were conducted to ensure that as many recent studies as possible were comprehensively included. In addition, we observed possible publication bias in the analyses; more rigorous studies are needed to confirm these findings.

In summary, this systematic review provided the first comprehensive synthesis of evidence across all DHI types for PD, which distinguishes it from previous reviews that focused on single technologies or specific outcomes. DHIs were associated with improvement in motor symptoms, overall nonmotor symptoms, cognitive function, and psychiatric symptoms, but not with quality of life, compared with the non-DHI group. This study supports the potential of DHIs as a promising nonpharmacological intervention for PD. It brings to the field an integrated evidence base that can guide the design of future research and inform the careful implementation of DHIs in clinical practice. However, in real-world implications, the conclusions should be interpreted with caution given the wide 95% PIs, relatively high risk of bias, small-study effects, and the very-low GRADE ratings. Therefore, more well-designed DHIs and high-quality research are urgently needed, especially in underresourced regions.

Acknowledgments

We would like to thank the authors of the primary studies for providing the data used in this meta-analysis.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Funding

This research was funded by the Sichuan Science and Technology Program (grant number 2022ZDZX0023) and Chengdu Science and Technology Project Key Research and Development Support Program (grant number 2024-YF08-00038-GX).

Authors' Contributions

RL and SZ contributed to conceptualization, methodology, investigation, data curation, formal analysis, and writing of the original draft. YX contributed to investigation, data curation, formal analysis, validation, and writing—review and editing. YC and HS contributed to supervision, project administration, and writing—review and editing. All authors approved the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA 2020 checklist.

XSL•FO
RenderX

[DOCX File , 32 KB - jmir_v28i1e79935_app1.docx]

Multimedia Appendix 2
PRISMA-S checklist.
[DOCX File , 17 KB - jmir_v28i1e79935_app2.docx]

Multimedia Appendix 3
Deviations from the original protocol in the final review.
[DOCX File , 21 KB - jmir_v28i1e79935_app3.docx]

Multimedia Appendix 4
Search strategies per database.
[DOCX File , 24 KB - jmir_v28i1e79935_app4.docx]

Multimedia Appendix 5
Outcome measurement tools utilized in individual studies.
[DOCX File , 38 KB - jmir_v28i1e79935_app5.docx]

Multimedia Appendix 6
Characteristics of included studies.
[DOCX File , 68 KB - jmir_v28i1e79935_app6.docx]

Multimedia Appendix 7
Risk of bias for included effect estimates based on the Cochrane risk-of-bias tool for randomized trials (RoB2).
[DOCX File , 44 KB - jmir_v28i1e79935_app7.docx]

Multimedia Appendix 8
Pairwise Comparisons of Intervention Type Subgroups.
[DOCX File , 21 KB - jmir_v28i1e79935_app8.docx]

Multimedia Appendix 9
Meta-analysis of digital health interventions' efficacy on follow-up motor symptoms, psychiatric symptoms, cognitive function, and quality of life.
[DOCX File , 448 KB - jmir_v28i1e79935_app9.docx]

Multimedia Appendix 10
Publication bias analyses – Funnel plots and trim and fill analyses at post-intervention assessments on motor symptoms, psychiatric symptoms, cognitive function, overall non-motor symptoms and quality of life.
[DOCX File , 176 KB - jmir_v28i1e79935_app10.docx]

Multimedia Appendix 11
Meta-regression analyses of the effect on motor symptoms, quality of life, cognitive function and psychiatric symptoms.
[DOCX File , 28 KB - jmir_v28i1e79935_app11.docx]

Multimedia Appendix 12
Bubble plots for the continuous variables in meta-regression.
[DOCX File , 365 KB - jmir_v28i1e79935_app12.docx]

Multimedia Appendix 13
Reach, uptake, dropout rate, and feasibility of included interventions.
[DOCX File , 36 KB - jmir_v28i1e79935_app13.docx]

References

1. Bloem BR, Okun MS, Klein C. Parkinson's disease. Lancet 2021;397(10291):2284-2303. [doi: 10.1016/S0140-6736(21)00218-X] [Medline: 33848468]

XSL•FO
RenderX

2. Tanner CM, Ostrem JL. Parkinson's disease. N Engl J Med 2024;391(5):442-452. [doi: 10.1056/NEJMra2401857] [Medline: 39083773]
3. Dorsey ER, Sherer T, Okun MS, Bloem BR. The emerging evidence of the parkinson pandemic. J Parkinsons Dis 2018;8(s1):S3-S8 [FREE Full text] [doi: 10.3233/JPD-181474] [Medline: 30584159]
4. GBD 2016 Parkinson's Disease Collaborators. Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol 2018;17(11):939-953 [FREE Full text] [doi: 10.1016/S1474-4422(18)30295-3] [Medline: 30287051]
5. Foltynie T, Bruno V, Fox S, Kühn AA, Lindop F, Lees AJ. Medical, surgical, and physical treatments for Parkinson's disease. Lancet 2024;403(10423):305-324. [doi: 10.1016/S0140-6736(23)01429-0] [Medline: 38245250]
6. Leite Silva ABR, Gonçalves de Oliveira RW, Diógenes GP, de Castro Aguiar MF, Sallem CC, Lima MPP. Premotor, nonmotor and motor symptoms of Parkinson's disease: a new clinical state of the art. Ageing Res Rev 2023;84:101834. [doi: 10.1016/j.arr.2022.101834] [Medline: 36581178]
7. Li S, Jia C, Li T, Le W. Hot topics in recent Parkinson's disease research: where we are and where we should go. Neurosci Bull 2021;37(12):1735-1744 [FREE Full text] [doi: 10.1007/s12264-021-00749-x] [Medline: 34313916]
8. The Lancet. What next in Parkinson's disease? Lancet 2024;403(10423):219. [doi: 10.1016/S0140-6736(24)00094-1] [Medline: 38245237]
9. Langeskov-Christensen M, Franzén E, Grøndahl Hvid L, Dalgas U. Exercise as medicine in Parkinson's disease. J Neurol Neurosurg Psychiatry 2024;95(11):1077-1088. [doi: 10.1136/jnnp-2023-332974] [Medline: 38418216]
10. Schalkamp AK, Peall KJ, Harrison NA, Sandor C. Wearable movement-tracking data identify Parkinson's disease years before clinical diagnosis. Nat Med 2023;29(8):2048-2056. [doi: 10.1038/s41591-023-02440-2] [Medline: 37400639]
11. Rábano-Suárez P, Del Campo N, Benatru I, Moreau C, Desjardins C, Sánchez-Ferro Á. Digital outcomes as biomarkers of disease progression in early Parkinson's disease: a systematic review. Mov Disord 2025;40(2):184-203. [doi: 10.1002/mds.30056] [Medline: 39613480]
12. Hyeon B, Shin J, Lee J, Kim W, Kwon J, Lee H. Integrating artificial intelligence and optogenetics for Parkinson's disease diagnosis and therapeutics in male mice. Nat Commun 2025;16(1):7797 [FREE Full text] [doi: 10.1038/s41467-025-63025-w] [Medline: 40841722]
13. Qi W, Shen S, Dong C, Zhao M, Zang S, Zhu X. Digital biomarkers for Parkinson disease: bibliometric analysis and a scoping review of deep learning for freezing of gait. J Med Internet Res 2025;27:e71560 [FREE Full text] [doi: 10.2196/71560] [Medline: 40392578]
14. Park Y, Kim EJ, Park S, Lee M. Digital health intervention effect on older adults with chronic diseases living alone: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2025;27:e63168 [FREE Full text] [doi: 10.2196/63168] [Medline: 40163849]
15. Bai X, Zhang H, Jiao Y, Yuan C, Ma Y, Han L. Digital health interventions for chronic wound management: a systematic review and meta-analysis. J Med Internet Res 2024;26:e47904 [FREE Full text] [doi: 10.2196/47904] [Medline: 39012684]
16. Ellis TD, Earhart GM. Digital therapeutics in Parkinson's disease: practical applications and future potential. J Parkinsons Dis 2021;11(s1):S95-S101 [FREE Full text] [doi: 10.3233/JPD-202407] [Medline: 33646177]
17. Xiong S, Lu H, Peoples N, Duman EK, Najarro A, Ni Z. Digital health interventions for non-communicable disease management in primary health care in low-and middle-income countries. NPJ Digit Med 2023;6(1):12 [FREE Full text] [doi: 10.1038/s41746-023-00764-4] [Medline: 36725977]
18. Boima V, Doku A, Agyekum F, Tuglo LS, Agyemang C. Effectiveness of digital health interventions on blood pressure control, lifestyle behaviours and adherence to medication in patients with hypertension in low-income and middle-income countries: a systematic review and meta-analysis of randomised controlled trials. EClinicalMedicine 2024;69:102432 [FREE Full text] [doi: 10.1016/j.eclinm.2024.102432] [Medline: 38333367]
19. Zangger G, Bricca A, Liaghat B, Juhl CB, Mortensen SR, Andersen RM. Benefits and harms of digital health interventions promoting physical activity in people with chronic conditions: systematic review and meta-analysis. J Med Internet Res 2023;25:e46439 [FREE Full text] [doi: 10.2196/46439] [Medline: 37410534]
20. Gavelin HM, Domellöf ME, Leung I, Neely AS, Launder NH, Nategh L. Computerized cognitive training in Parkinson's disease: a systematic review and meta-analysis. Ageing Res Rev 2022;80:101671. [doi: 10.1016/j.arr.2022.101671] [Medline: 35714854]
21. Darbandsari P, Pescatello LS, Piscitelli D, Smith JM, Ugolini A, Colón-Semenza C. Effect of telerehabilitation in Parkinson disease: a systematic review and meta-analysis. Phys Ther 2025;105(11):376-383. [doi: 10.1093/ptj/pzaf121] [Medline: 41065729]
22. Kim J, Porciuncula F, Yang HD, Wendel N, Baker T, Chin A. Soft robotic apparel to avert freezing of gait in Parkinson's disease. Nat Med 2024;30(1):177-185. [doi: 10.1038/s41591-023-02731-8] [Medline: 38182783]
23. Schootemeijer S, de Vries NM, Darweesh SKL, Ascherio A, Schwarzschild MA, Macklin EA. Promoting physical activity in people with Parkinson's disease through a smartphone app: a pilot study. J Neurol Phys Ther 2025;49(2):74-81. [doi: 10.1097/NPT.0000000000000507] [Medline: 39763021]

XSL•FO
RenderX

24. Hinman RS, Campbell PK, Kimp AJ, Russell T, Foster NE, Kasza J. Telerehabilitation consultations with a physiotherapist for chronic knee pain versus in-person consultations in Australia: the PEAK non-inferiority randomised controlled trial. Lancet 2024;403(10433):1267-1278. [doi: 10.1016/S0140-6736(23)02630-2] [Medline: 38461844]
25. Rawson TM, Zhu N, Galiwango R, Cocker D, Islam MS, Myall A. Using digital health technologies to optimise antimicrobial use globally. Lancet Digit Health 2024;6(12):e914-e925. [doi: 10.1016/S2589-7500(24)00198-5] [Medline: 39547912]
26. Knitza J, Gupta L, Hügle T. Rheumatology in the digital health era: status quo and quo vadis? Nat Rev Rheumatol 2024;20(12):747-759. [doi: 10.1038/s41584-024-01177-7] [Medline: 39482466]
27. Capecci M, Pournajaf S, Galafate D, Sale P, Le Pera D, Goffredo M. Clinical effects of robot-assisted gait training and treadmill training for Parkinson's disease. A randomized controlled trial. Ann Phys Rehabil Med 2019;62(5):303-312 [FREE Full text] [doi: 10.1016/j.rehab.2019.06.016] [Medline: 31377382]
28. Hajebrahimi F, Velioglu HA, Bayraktaroglu Z, Helvaci Yilmaz N, Hanoglu L. Clinical evaluation and resting state fMRI analysis of virtual reality based training in Parkinson's disease through a randomized controlled trial. Sci Rep 2022;12(1):8024 [FREE Full text] [doi: 10.1038/s41598-022-12061-3] [Medline: 35577874]
29. Dorsey ER, Venkataraman V, Grana MJ, Bull MT, George BP, Boyd CM. Randomized controlled clinical trial of "virtual house calls" for Parkinson disease. JAMA Neurol 2013;70(5):565-570 [FREE Full text] [doi: 10.1001/jamaneurol.2013.123] [Medline: 23479138]
30. Lakshminarayana R, Wang DL, Burn D, Chaudhuri KR, Galtrey C, Guzman NV. Erratum: author correction: using a smartphone-based self-management platform to support medication adherence and clinical consultation in Parkinson's disease. NPJ Parkinsons Dis 2017;3(1):32 [FREE Full text] [doi: 10.1038/s41531-017-0034-0] [Medline: 29152558]
31. Isaacson SH, Boroojerdi B, Waln O, McGraw M, Kreitzman DL, Klos K. Effect of using a wearable device on clinical decision-making and motor symptoms in patients with Parkinson's disease starting transdermal rotigotine patch: a pilot study. Parkinsonism Relat Disord 2019;64:132-137. [doi: 10.1016/j.parkreldis.2019.01.025] [Medline: 30948242]
32. Sarasso E, Gardoni A, Tettamanti A, Agosta F, Filippi M, Corbetta D. Virtual reality balance training to improve balance and mobility in Parkinson's disease: a systematic review and meta-analysis. J Neurol 2022;269(4):1873-1888. [doi: 10.1007/s00415-021-10857-3] [Medline: 34713324]
33. Elena P, Demetris S, Christina M, Marios P. Differences between exergaming rehabilitation and conventional physiotherapy on quality of life in Parkinson's disease: a systematic review and meta-analysis. Front Neurol 2021;12:683385 [FREE Full text] [doi: 10.3389/fneur.2021.683385] [Medline: 34434157]
34. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
35. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
36. Herold F, Theobald P, Gronwald T, Kaushal N, Zou L, de Bruin ED. Alexa, let's train now! - a systematic review and classification approach to digital and home-based physical training interventions aiming to support healthy cognitive aging. J Sport Health Sci 2024;13(1):30-46 [FREE Full text] [doi: 10.1016/j.jshs.2023.01.004] [Medline: 36736727]
37. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W. Process evaluation of complex interventions: medical research council guidance. BMJ 2015;350:h1258 [FREE Full text] [doi: 10.1136/bmj.h1258] [Medline: 25791983]
38. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011;38(2):65-76 [FREE Full text] [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
39. Moschonis G, Siopis G, Jung J, Eweka E, Willems R, Kwasnicka D. Effectiveness, reach, uptake, and feasibility of digital health interventions for adults with type 2 diabetes: a systematic review and meta-analysis of randomised controlled trials. Lancet Digit Health 2023;5(3):e125-e143 [FREE Full text] [doi: 10.1016/S2589-7500(22)00233-3] [Medline: 36828606]
40. Chapter 8: assessing risk of bias in a randomized trial. Cochrane Handbook for Systematic Reviews of Interventions. 2019. URL: http:\\www.training.cochrane.org/handbook [accessed 2019-10-01]
41. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336(7650):924-926 [FREE Full text] [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
42. Luo D, Wan X, Liu J, Tong T. Optimally estimating the sample mean from the sample size, median, mid-range, and/or mid-quartile range. Stat Methods Med Res 2018;27(6):1785-1805. [doi: 10.1177/0962280216669183] [Medline: 27683581]
43. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC Med Res Methodol 2014;14:135 [FREE Full text] [doi: 10.1186/1471-2288-14-135] [Medline: 25524443]
44. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014;14:25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
45. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023;12(4):101014 [FREE Full text] [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]

XSL•FO
RenderX

46. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
47. IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open 2016;6(7):e010247 [FREE Full text] [doi: 10.1136/bmjopen-2015-010247] [Medline: 27406637]
48. Ruppar T. Meta-analysis: how to quantify and explain heterogeneity? Eur J Cardiovasc Nurs 2020;19(7):646-652. [doi: 10.1177/1474515120944014] [Medline: 32757621]
49. Alberts JL, Kaya RD, Penko AL, Streicher M, Zimmerman EM, Davidson S. A randomized clinical trial to evaluate a digital therapeutic to enhance gait function in individuals with Parkinson's disease. Neurorehabil Neural Repair 2023;37(9):603-616. [doi: 10.1177/15459683231184190] [Medline: 37465959]
50. Allen NE, Song J, Paul SS, Smith S, O'Duffy J, Schmidt M. An interactive videogame for arm and hand exercise in people with Parkinson's disease: a randomized controlled trial. Parkinsonism Relat Disord 2017;41:66-72. [doi: 10.1016/j.parkreldis.2017.05.011] [Medline: 28528804]
51. Bartolo M, Castelli A, Calabrese M, Buttacchio G, Zucchella C, Tamburin S. A wearable system for visual cueing gait rehabilitation in Parkinson's disease: a randomized non-inferiority trial. Eur J Phys Rehabil Med 2024;60(2):245-256 [FREE Full text] [doi: 10.23736/S1973-9087.24.08381-3] [Medline: 38483335]
52. Beck CA, Beran DB, Biglan KM, Boyd CM, Dorsey ER, Schmidt PN. National randomized controlled trial of virtual house calls for Parkinson disease. Neurology 2017;89(11):1152-1161 [FREE Full text] [doi: 10.1212/WNL.0000000000004357] [Medline: 28814455]
53. Bernini S, Alloni A, Panzarasa S, Picascia M, Quaglini S, Tassorelli C. A computer-based cognitive training in mild cognitive impairment in Parkinson’s disease. Neuro Rehabilitation 2019;44(4):555-567. [doi: 10.3233/nre-192714]
54. Bernini S, Panzarasa S, Barbieri M, Sinforiani E, Quaglini S, Tassorelli C. A double-blind randomized controlled trial of the efficacy of cognitive training delivered using two different methods in mild cognitive impairment in Parkinson's disease: preliminary report of benefits associated with the use of a computerized tool. Aging Clin Exp Res 2021;33(6):1567-1575. [doi: 10.1007/s40520-020-01665-2] [Medline: 32895890]
55. Bogosian A, Hurt CS, Hindle JV, McCracken LM, Vasconcelos E Sa DA, Axell S. Acceptability and feasibility of a mindfulness intervention delivered via videoconferencing for people with Parkinson's. J Geriatr Psychiatry Neurol 2022;35(1):155-167 [FREE Full text] [doi: 10.1177/0891988720988901] [Medline: 33504245]
56. Calabrò RS, Naro A, Filoni S, Pullia M, Billeri L, Tomasello P. Walking to your right music: a randomized controlled trial on the novel use of treadmill plus music in Parkinson's disease. J Neuroeng Rehabil 2019;16(1):68 [FREE Full text] [doi: 10.1186/s12984-019-0533-9] [Medline: 31174570]
57. Carda S, Invernizzi M, Baricich A, Comi C, Croquelois A, Cisari C. Robotic gait training is not superior to conventional treadmill training in parkinson disease: a single-blind randomized controlled trial. Neurorehabil Neural Repair 2012;26(9):1027-1034. [doi: 10.1177/1545968312446753] [Medline: 22623206]
58. Carpinella I, Cattaneo D, Bonora G, Bowman T, Martina L, Montesano A. Wearable sensor-based biofeedback training for balance and gait in Parkinson disease: a pilot randomized controlled trial. Arch Phys Med Rehabil 2017;98(4):622-630.e3 [FREE Full text] [doi: 10.1016/j.apmr.2016.11.003] [Medline: 27965005]
59. Çetin B, Kılınç M, Çakmaklı GY. The effects of exergames on upper extremity performance, trunk mobility, gait, balance, and cognition in Parkinson's disease: a randomized controlled study. Acta Neurol Belg 2024;124(3):853-863. [doi: 10.1007/s13760-023-02451-3] [Medline: 38182919]
60. Constantinescu G, Theodoros D, Russell T, Ward E, Wilson S, Wootton R. Treating disordered speech and voice in Parkinson's disease online: a randomized controlled non-inferiority trial. Int J Lang Commun Disord 2011;46(1):1-16. [doi: 10.3109/13682822.2010.484848] [Medline: 21281410]
61. Da Silva KG, Nuvolini RA, Bacha JMR, De Freitas TB, Doná F, Torriani-Pasin C. Comparison of the effects of an exergame-based program with conventional physiotherapy protocol based on core areas of the European guideline on postural control, functional mobility, and quality of life in patients with Parkinson's disease: randomized clinical trial. Games Health J 2023;12(3):228-241. [doi: 10.1089/g4h.2022.0039] [Medline: 36206023]
62. Das J, Barry G, Walker R, Vitorio R, Celik Y, McDonald C. The feasibility of a visuo-cognitive training intervention using a mobile application and exercise with stroboscopic glasses in Parkinson's: findings from a pilot randomised controlled trial. PLOS Digit Health 2024;3(12):e0000696. [doi: 10.1371/journal.pdig.0000696] [Medline: 39693278]
63. De Luca R, Latella D, Maggio MG, Di Lorenzo G, Maresca G, Sciarrone F. Computer assisted cognitive rehabilitation improves visuospatial and executive functions in Parkinson's disease: preliminary results. Neuro Rehabilitation 2019;45(2):285-290. [doi: 10.3233/NRE-192789] [Medline: 31498141]
64. de Melo GEL, Kleiner AFR, Lopes JBP, Dumont AJL, Lazzari RD, Galli M. Effect of virtual reality training on walking distance and physical fitness in individuals with Parkinson's disease. Neuro Rehabilitation 2018;42(4):473-480. [doi: 10.3233/NRE-172355] [Medline: 29660956]
65. Del Pino R, de Echevarría AO, Díez-Cirarda M, Ustarroz-Aguirre I, Caprino M, Liu J. Virtual coach and telerehabilitation for Parkinson's disease patients: vCare system. J Public Health (Berl.) 2023;33(7):1583-1596. [doi: 10.1007/s10389-023-02082-1]

XSL•FO
RenderX

66. Dobkin RD, Mann SL, Gara MA, Interian A, Rodriguez KM, Menza M. Telephone-based cognitive behavioral therapy for depression in Parkinson disease. Neurology 2020;94(16):e1764-e1773. [doi: 10.1212/wnl.0000000000009292]
67. Dobkin RD, Mann SL, Weintraub D, Rodriguez KM, Miller RB, St Hill L. Innovating Parkinson's care: a randomized controlled trial of telemedicine depression treatment. Mov Disord 2021;36(11):2549-2558. [doi: 10.1002/mds.28548] [Medline: 33710659]
68. Dorsey ER, Deuel LM, Voss TS, Finnigan K, George BP, Eason S. Increasing access to specialty care: a pilot, randomized controlled trial of telemedicine for Parkinson's disease. Mov Disord 2010;25(11):1652-1659. [doi: 10.1002/mds.23145] [Medline: 20533449]
69. Dorsey ER, Venkataraman V, Grana MJ, Bull MT, George BP, Boyd CM. Randomized controlled clinical trial of "virtual house calls" for Parkinson disease. JAMA Neurol 2013;70(5):565-570 [FREE Full text] [doi: 10.1001/jamaneurol.2013.123] [Medline: 23479138]
70. Edwards JD, Hauser RA, O'Connor ML, Valdés EG, Zesiewicz TA, Uc EY. Randomized trial of cognitive speed of processing training in Parkinson disease. Neurology 2013;81(15):1284-1290 [FREE Full text] [doi: 10.1212/WNL.0b013e3182a823ba] [Medline: 24014503]
71. Eldemir S, Guclu-Gunduz A, Eldemir K, Saygili F, Yilmaz R, Akbostancı MC. The effect of task-oriented circuit training-based telerehabilitation on upper extremity motor functions in patients with Parkinson's disease: a randomized controlled trial. Parkinsonism Relat Disord 2023;109:105334. [doi: 10.1016/j.parkreldis.2023.105334] [Medline: 36917914]
72. Ellis TD, Cavanaugh JT, DeAngelis T, Hendron K, Thomas CA, Saint-Hilaire M. Comparative effectiveness of mHealth-supported exercise compared with exercise alone for people with Parkinson disease: randomized controlled pilot study. Phys Ther 2019;99(2):203-216. [doi: 10.1093/ptj/pzy131] [Medline: 30715489]
73. Fellman D, Salmi J, Ritakallio L, Ellfolk U, Rinne JO, Laine M. Training working memory updating in Parkinson's disease: a randomised controlled trial. Neuropsychol Rehabil 2020;30(4):673-708. [doi: 10.1080/09602011.2018.1489860] [Medline: 29968519]
74. Feng H, Li C, Liu J, Wang L, Ma J, Li G. Virtual reality rehabilitation versus conventional physical therapy for improving balance and gait in Parkinson's disease patients: a randomized controlled trial. Med Sci Monit 2019;25:4186-4192 [FREE Full text] [doi: 10.12659/MSM.916455] [Medline: 31165721]
75. Ferraz DD, Trippo KV, Duarte GP, Neto MG, Bernardes Santos KO, Filho JO. The effects of functional training, bicycle exercise, and exergaming on walking capacity of elderly patients with Parkinson disease: a pilot randomized controlled single-blinded trial. Arch Phys Med Rehabil 2018;99(5):826-833. [doi: 10.1016/j.apmr.2017.12.014] [Medline: 29337023]
76. Flynn A, Preston E, Dennis S, Canning CG, Allen NE. Home-based exercise monitored with telehealth is feasible and acceptable compared to centre-based exercise in Parkinson's disease: a randomised pilot study. Clin Rehabil 2021;35(5):728-739. [doi: 10.1177/0269215520976265] [Medline: 33272025]
77. Furnari A, Calabrò RS, De Cola MC, Bartolo M, Castelli A, Mapelli A. Robotic-assisted gait training in Parkinson's disease: a three-month follow-up randomized clinical trial. Int J Neurosci 2017;127(11):996-1004. [doi: 10.1080/00207454.2017.1288623] [Medline: 28132574]
78. Galli M, Cimolin V, De Pandis MF, Le Pera D, Sova I, Albertini G. Robot-assisted gait training versus treadmill training in patients with Parkinson's disease: a kinematic evaluation with gait profile score. Funct Neurol 2016;31(3):163-170 [FREE Full text] [doi: 10.11138/fneur/2016.31.3.163] [Medline: 27678210]
79. Gandolfi M, Geroin C, Dimitrova E, Boldrini P, Waldner A, Bonadiman S. Virtual reality telerehabilitation for postural instability in Parkinson's disease: a multicenter, single-blind, randomized, controlled trial. Biomed Res Int 2017;2017:7962826 [FREE Full text] [doi: 10.1155/2017/7962826] [Medline: 29333454]
80. Giehl K, Ophey A, Hammes J, Rehberg S, Lichtenstein T, Reker P. Working memory training increases neural efficiency in Parkinson's disease: a randomized controlled trial. Brain Commun 2020;2(2):fcaa115 [FREE Full text] [doi: 10.1093/braincomms/fcaa115] [Medline: 32954349]
81. Giehl K, Ophey A, Reker P, Rehberg S, Hammes J, Barbe MT. Effects of home-based working memory training on visuo-spatial working memory in Parkinson's disease: a randomized controlled trial. J Cent Nerv Syst Dis 2020;12:1179573519899469 [FREE Full text] [doi: 10.1177/1179573519899469] [Medline: 32002011]
82. Ginis P, Nieuwboer A, Dorfman M, Ferrari A, Gazit E, Canning CG. Feasibility and effects of home-based smartphone-delivered automated feedback training for gait in people with Parkinson's disease: a pilot randomized controlled trial. Parkinsonism Relat Disord 2016;22:28-34 [FREE Full text] [doi: 10.1016/j.parkreldis.2015.11.004] [Medline: 26777408]
83. Pedreira G, Prazeres A, Cruz D, Gomes I, Monteiro L, Melo A. Virtual games and quality of life in Parkinson's disease: a randomised controlled trial. Advances Parkinson's Disease 2013;02(04):97-101. [doi: 10.4236/apd.2013.24018]
84. Goffredo M, Baglio F, De Icco R, Proietti S, Maggioni G, Turolla A. Efficacy of non-immersive virtual reality-based telerehabilitation on postural stability in Parkinson's disease: a multicenter randomized controlled trial. Eur J Phys Rehabil Med 2023;59(6):689-696 [FREE Full text] [doi: 10.23736/S1973-9087.23.07954-6] [Medline: 37847247]
85. Gryfe P, Sexton A, McGibbon CA. Using gait robotics to improve symptoms of Parkinson's disease: an open-label, pilot randomized controlled trial. Eur J Phys Rehabil Med 2022;58(5):723-737 [FREE Full text] [doi: 10.23736/S1973-9087.22.07549-9] [Medline: 35708047]

XSL•FO
RenderX

86. Gulcan K, Guclu-Gunduz A, Yasar E, Ar U, Sucullu Karadag Y, Saygili F. The effects of augmented and virtual reality gait training on balance and gait in patients with Parkinson's disease. Acta Neurol Belg 2023;123(5):1917-1925 [FREE Full text] [doi: 10.1007/s13760-022-02147-0] [Medline: 36443623]
87. Hajebrahimi F, Velioglu HA, Bayraktaroglu Z, Helvaci Yilmaz N, Hanoglu L. Clinical evaluation and resting state fMRI analysis of virtual reality based training in Parkinson's disease through a randomized controlled trial. Sci Rep 2022;12(1):8024 [FREE Full text] [doi: 10.1038/s41598-022-12061-3] [Medline: 35577874]
88. Halpern AE, Ramig LO, Matos CEC, Petska-Cable JA, Spielman JL, Pogoda JM. Innovative technology for the assisted delivery of intensive voice treatment (LSVT®LOUD) for Parkinson disease. Am J Speech Lang Pathol 2012;21(4):354-367. [doi: 10.1044/1058-0360(2012/11-0125)] [Medline: 23071195]
89. Han T, Liu Q, Hu Y, Wang Y, Xue K. Effect of Pro-kin visual feedback balance training on balance function of individuals with early Parkinson's disease: a randomized controlled pilot trial. Afr Health Sci 2023;23(2):582-588 [FREE Full text] [doi: 10.4314/ahs.v23i2.67] [Medline: 38223579]
90. Hashemi Y, Taghizadeh G, Azad A, Behzadipour S. The effects of supervised and non-supervised upper limb virtual reality exercises on upper limb sensory-motor functions in patients with idiopathic Parkinson's disease. Hum Mov Sci 2022;85:102977. [doi: 10.1016/j.humov.2022.102977] [Medline: 35932518]
91. Heldman DA, Harris DA, Felong T, Andrzejewski KL, Dorsey ER, Giuffrida JP. Telehealth management of Parkinson's disease using wearable sensors: an exploratory study. Digit Biomark 2017;1(1):43-51 [FREE Full text] [doi: 10.1159/000475801] [Medline: 29725667]
92. Jäggi S, Wachter A, Adcock M, de Bruin ED, Möller JC, Marks D. Feasibility and effects of cognitive-motor exergames on fall risk factors in typical and atypical Parkinson's inpatients: a randomized controlled pilot study. Eur J Med Res 2023;28(1):30 [FREE Full text] [doi: 10.1186/s40001-022-00963-x] [Medline: 36647177]
93. Johnson JK, Longhurst JK, Gevertzman M, Jefferson C, Linder SM, Bethoux F. The use of telerehabilitation to improve movement-related outcomes and quality of life for individuals with Parkinson disease: pilot randomized controlled trial. JMIR Form Res 2024;8:e54599 [FREE Full text] [doi: 10.2196/54599] [Medline: 39083792]
94. Jong-Hoon M, Jin-Hwa J, Hwi-Young C. Effects of balance training using a wii fit balance board on balance, gait and activities of daily living in patients with Parkinson disease: a pilot, randomized controlled trial. Medico-Legal 2020;20(1):1799-1803. [doi: 10.37506/v20/i1/2020/mlu/194564] [Medline: 162448856]
95. Kashif M, Ahmad A, Bandpei MAM, Syed HA, Raza A, Sana V. A randomized controlled trial of motor imagery combined with virtual reality techniques in patients with Parkinson's disease. J Pers Med 2022;12(3):450 [FREE Full text] [doi: 10.3390/jpm12030450] [Medline: 35330450]
96. Kashif M, Albalwi AA, Zulfiqar A, Bashir K, Alharbi AA, Zaidi S. Effects of virtual reality versus motor imagery versus routine physical therapy in patients with parkinson's disease: a randomized controlled trial. BMC Geriatr 2024;24(1):229 [FREE Full text] [doi: 10.1186/s12877-024-04845-1] [Medline: 38443801]
97. Kawashima N, Hasegawa K, Iijima M, Nagami K, Makimura T, Kumon A. Efficacy of wearable device gait training on Parkinson's disease: a randomized controlled open-label pilot study. Intern Med 2022;61(17):2573-2580 [FREE Full text] [doi: 10.2169/internalmedicine.8949-21] [Medline: 35135928]
98. Kegelmeyer DA, Minarsch R, Kostyk SK, Kline D, Smith R, Kloos AD. Use of a robotic walking device for home and community mobility in Parkinson disease: a randomized controlled trial. J Neurol Phys Ther 2024;48(2):102-111. [doi: 10.1097/NPT.0000000000000467] [Medline: 38441461]
99. Khalil H, Busse M, Quinn L, Nazzal M, Batyha W, Alkhazaleh S. A pilot study of a minimally supervised home exercise and walking program for people with Parkinson's disease in Jordan. Neurodegener Dis Manag 2017;7(1):73-84. [doi: 10.2217/nmt-2016-0041] [Medline: 28097926]
100. Kim H, Kim E, Yun SJ, Kang M, Shin HI, Oh B. Robot-assisted gait training with auditory and visual cues in Parkinson's disease: a randomized controlled trial. Ann Phys Rehabil Med 2022;65(3):101620 [FREE Full text] [doi: 10.1016/j.rehab.2021.101620] [Medline: 34896605]
101. Kluger BM, Katz M, Galifianakis NB, Pantilat SZ, Hauser JM, Khan R. Patient and family outcomes of community neurologist palliative education and telehealth support in Parkinson disease. JAMA Neurol 2024;81(1):39-49. [doi: 10.1001/jamaneurol.2023.4260] [Medline: 37955923]
102. Kraepelien M, Schibbye R, Månsson K, Sundström C, Riggare S, Andersson G. Individually tailored internet-based cognitive-behavioral therapy for daily functioning in patients with Parkinson’s disease: a randomized controlled trial. J Parkinsons Dis 2020;10(2):653-664. [doi: 10.3233/jpd-191894]
103. Lai B, Bond K, Kim Y, Barstow B, Jovanov E, Bickel CS. Exploring the uptake and implementation of tele-monitored home-exercise programmes in adults with Parkinson's disease: a mixed-methods pilot study. J Telemed Telecare 2020;26(1-2):53-63. [doi: 10.1177/1357633X18794315] [Medline: 30134777]
104. Lau J, Regis C, Burke C, Kaleda MJ, McKenna R, Muratori LM. Immersive technology for cognitive-motor training in Parkinson's disease. Front Hum Neurosci 2022;16(1):863930-863936 [FREE Full text] [doi: 10.3389/fnhum.2022.863930] [Medline: 35615742]

XSL•FO
RenderX

105. Li KP, Zhang ZQ, Zhou ZL, Su JQ, Wu XH, Shi BH. Effect of music-based movement therapy on the freezing of gait in patients with Parkinson's disease: a randomized controlled trial. Front Aging Neurosci 2022;14:924784 [FREE Full text] [doi: 10.3389/fnagi.2022.924784] [Medline: 36337701]
106. Liao YY, Yang YR, Wu YR, Wang RY. Virtual reality-based wii fit training in improving muscle strength, sensory integration ability, and walking abilities in patients with Parkinson's disease: a randomized control trial. Int J Gerontol 2015;9(4):190-195. [doi: 10.1016/j.ijge.2014.06.007]
107. Maas JJL, de Vries NM, IntHout J, Bloem BR, Kalf JG. Effectiveness of remotely delivered speech therapy in persons with Parkinson's disease - a randomised controlled trial. E Clinical Medicine 2024;76:102823 [FREE Full text] [doi: 10.1016/j.eclinm.2024.102823] [Medline: 39309726]
108. Maggio MG, De Cola MC, Latella D, Maresca G, Finocchiaro C, La Rosa G. What about the role of virtual reality in Parkinson disease's cognitive rehabilitation? Preliminary findings from a randomized clinical trial. J Geriatr Psychiatry Neurol 2018;31(6):312-318. [doi: 10.1177/0891988718807973] [Medline: 30360679]
109. Maggio MG, Luca A, Cicero CE, Calabrò RS, Drago F, Zappia M. Effectiveness of telerehabilitation plus virtual reality (Tele-RV) in cognitive e social functioning: a randomized clinical study on Parkinson's disease. Parkinsonism Relat Disord 2024;119:105970. [doi: 10.1016/j.parkreldis.2023.105970] [Medline: 38142630]
110. Manor Y, Mootanah R, Freud D, Giladi N, Cohen JT. Video-assisted swallowing therapy for patients with Parkinson's disease. Parkinsonism Relat Disord 2013;19(2):207-211. [doi: 10.1016/j.parkreldis.2012.10.004] [Medline: 23131836]
111. Maranesi E, Casoni E, Baldoni R, Barboni I, Rinaldi N, Tramontana B. The effect of non-immersive virtual reality exergames versus traditional physiotherapy in Parkinson's disease older patients: preliminary results from a randomized-controlled trial. Int J Environ Res Public Health 2022;19(22):14818 [FREE Full text] [doi: 10.3390/ijerph192214818] [Medline: 36429537]
112. McGibbon CA, Sexton A, Gryfe P. Exercising with a robotic exoskeleton can improve memory and gait in people with Parkinson's disease by facilitating progressive exercise intensity. Sci Rep 2024;14(1):4417 [FREE Full text] [doi: 10.1038/s41598-024-54200-y] [Medline: 38388571]
113. Shih MC, Wang RY, Cheng SJ, Yang YR. Effects of a balance-based exergaming intervention using the Kinect sensor on posture stability in individuals with Parkinson's disease: a single-blinded randomized controlled trial. J Neuroeng Rehabil 2016;13(1):78 [FREE Full text] [doi: 10.1186/s12984-016-0185-y] [Medline: 27568011]
114. Mirelman A, Rochester L, Maidan I, Del Din S, Alcock L, Nieuwhof F. Addition of a non-immersive virtual reality component to treadmill training to reduce fall risk in older adults (V-TIME): a randomised controlled trial. Lancet 2016;388(10050):1170-1182. [doi: 10.1016/S0140-6736(16)31325-3] [Medline: 27524393]
115. Nieuwboer A, Kwakkel G, Rochester L, Jones D, van Wegen E, Willems AM. Cueing training in the home improves gait-related mobility in Parkinson's disease: the RESCUE trial. J Neurol Neurosurg Psychiatry 2007;78(2):134-140 [FREE Full text] [doi: 10.1136/jnnp.200X.097923] [Medline: 17229744]
116. Ophey A, Giehl K, Rehberg S, Eggers C, Reker P, van Eimeren T, et al. Effects of working memory training in patients with Parkinson's disease without cognitive impairment: a randomized controlled trial. Parkinsonism Relat Disord 2020;72:13-22. [doi: 10.1016/j.parkreldis.2020.02.002] [Medline: 32078917]
117. Özden F. Letter to the editor: "comparison of virtual reality rehabilitation and conventional rehabilitation in Parkinson's disease: a randomised controlled trial". Physiotherapy 2021;110:87. [doi: 10.1016/j.physio.2020.04.001] [Medline: 32736882]
118. París AP, Saleta HG, de la Cruz Crespo Maraver M, Silvestre E, Freixa MG, Torrellas CP. Blind randomized controlled study of the efficacy of cognitive training in Parkinson's disease. Mov Disord 2011;26(7):1251-1258. [doi: 10.1002/mds.23688] [Medline: 21442659]
119. Pastana Ramos LF, Vilacorta-Pereira TDCS, Duarte JDS, Yamada ES, Santos-Lobato BL. Feasibility and effectiveness of a remote individual rehabilitation program for people with Parkinson's disease living in the Brazilian Amazon: a randomized clinical trial. Front Neurol 2023;14:1244661 [FREE Full text] [doi: 10.3389/fneur.2023.1244661] [Medline: 37693755]
120. Patel S, Ojo O, Genc G, Oravivattanakul S, Huo Y, Rasameesoraj T. A computerized cognitive behavioral therapy randomized, controlled, pilot trial for insomnia in Parkinson disease (ACCORD-PD). J Clin Mov Disord 2017;4(1):16 [FREE Full text] [doi: 10.1186/s40734-017-0062-2] [Medline: 28852567]
121. Peacock D, Yoneda J, Thomson V, Wile D. Tailoring the use of wearable systems and telehealth for Parkinson's disease. Parkinsonism Relat Disord 2021;89:111-112. [doi: 10.1016/j.parkreldis.2021.07.004] [Medline: 34274619]
122. Picelli A, Melotti C, Origano F, Waldner A, Gimigliano R, Smania N. Does robotic gait training improve balance in Parkinson's disease? A randomized controlled trial. Parkinsonism Relat Disord 2012;18(8):990-993. [doi: 10.1016/j.parkreldis.2012.05.010] [Medline: 22673035]
123. Picelli A, Melotti C, Origano F, Neri R, Waldner A, Smania N. Robot-assisted gait training versus equal intensity treadmill training in patients with mild to moderate Parkinson's disease: a randomized controlled trial. Parkinsonism Relat Disord 2013;19(6):605-610. [doi: 10.1016/j.parkreldis.2013.02.010] [Medline: 23490463]
124. Picelli A, Melotti C, Origano F, Neri R, Verzè E, Gandolfi M. Robot-assisted gait training is not superior to balance training for improving postural instability in patients with mild to moderate Parkinson's disease: a single-blind randomized controlled trial. Clin Rehabil 2015;29(4):339-347 [FREE Full text] [doi: 10.1177/0269215514544041] [Medline: 25082957]

XSL•FO
RenderX

125. Piers RJ, Farchione TJ, Wong BN, Rosellini AJ, Cronin-Golomb A. Telehealth transdiagnostic cognitive behavioral therapy for depression in Parkinson's disease: a pilot randomized controlled trial. Mov Disord Clin Pract 2023;10(1):79-85 [FREE Full text] [doi: 10.1002/mdc3.13587] [Medline: 36704072]
126. Pompeu JE, Mendes FADS, Silva KGD, Lobo AM, Oliveira TDP, Zomignani AP. Effect of Nintendo Wii™-based motor and cognitive training on activities of daily living in patients with Parkinson's disease: a randomised clinical trial. Physiotherapy 2012;98(3):196-204. [doi: 10.1016/j.physio.2012.06.004] [Medline: 22898575]
127. Qayyum S, Hashmi Z, Waqas S, Tariq M, Mughal MW, Tanvir M. Effects of exer–gaming on balance and gait in Parkinson's patients. Pak J Med Health Sci 2022;16(12):213-214. [doi: 10.53350/pjmhs20221612213]
128. Raciti L, Pignolo L, Perini V, Pullia M, Porcari B, Latella D. Improving upper extremity bradykinesia in Parkinson's disease: a randomized clinical trial on the use of gravity-supporting exoskeletons. J Clin Med 2022;11(9):2543 [FREE Full text] [doi: 10.3390/jcm11092543] [Medline: 35566669]
129. Raglio A, De Maria B, Parati M, Giglietti A, Premoli S, Salvaderi S. Movement sonification techniques to improve balance in Parkinson's disease: a pilot randomized controlled trial. Brain Sci 2023;13(11):1586 [FREE Full text] [doi: 10.3390/brainsci13111586] [Medline: 38002546]
130. Ribas CG, Alves da Silva L, Corrêa MR, Teive HG, Valderramas S. Effectiveness of exergaming in improving functional balance, fatigue and quality of life in Parkinson's disease: a pilot randomized controlled trial. Parkinsonism Relat Disord 2017;38(3):13-18. [doi: 10.1016/j.parkreldis.2017.02.006] [Medline: 28190675]
131. Sale P, De Pandis MF, Le Pera D, Sova I, Cimolin V, Ancillao A. Robot-assisted walking training for individuals with Parkinson's disease: a pilot randomized controlled trial. BMC Neurol 2013;13(3):50-57 [FREE Full text] [doi: 10.1186/1471-2377-13-50] [Medline: 23706025]
132. Santos P, Machado T, Santos L, Ribeiro N, Melo A. Efficacy of the nintendo wii combination with conventional exercises in the rehabilitation of individuals with Parkinson's disease: a randomized clinical trial. NeuroRehabilitation 2019;45(2):255-263. [doi: 10.3233/NRE-192771] [Medline: 31498138]
133. Sekimoto S, Oyama G, Hatano T, Sasaki F, Nakamura R, Jo T. A randomized crossover pilot study of telemedicine delivered via iPads in Parkinson's disease. Parkinsons Dis 2019;2019:9403295 [FREE Full text] [doi: 10.1155/2019/9403295] [Medline: 30723541]
134. So HY, Kim SR, Kim S, Park YS, Jo S, Park KW. Effect of home-based self-management intervention for community-dwelling patients with early Parkinson's disease: a feasibility study. J Community Health Nurs 2023;40(2):133-146. [doi: 10.1080/07370016.2022.2133566] [Medline: 36920114]
135. Song J, Paul SS, Caetano MJD, Smith S, Dibble LE, Love R. Home-based step training using videogame technology in people with Parkinson's disease: a single-blinded randomised controlled trial. Clin Rehabil 2018;32(3):299-311. [doi: 10.1177/0269215517721593] [Medline: 28745063]
136. Spina S, Facciorusso S, Cinone N, Armiento R, Picelli A, Avvantaggiato C. Effectiveness of robotic balance training on postural instability in patients with mild Parkinson's disease: a pilot, single blind, randomized controlled trial. J Rehabil Med 2021;53(2):jrm00154 [FREE Full text] [doi: 10.2340/16501977-2793] [Medline: 33585943]
137. Svaerke K, Faerk AK, Riis A, Stiegnitz von Ehrenfels SEM, Mogensen J, Lokkegaard A. Effects of computer-based cognitive rehabilitation on attention, executive functions, and quality of life in patients with Parkinson's disease: a randomized, controlled, single-blinded pilot study. Dement Geriatr Cogn Disord 2021;50(6):519-528. [doi: 10.1159/000520591] [Medline: 34986487]
138. Theodoros DG, Hill AJ, Russell TG. Clinical and quality of life outcomes of speech treatment for Parkinson's disease delivered to the home via telerehabilitation: a noninferiority randomized controlled trial. Am J Speech Lang Pathol 2016;25(2):214-232. [doi: 10.1044/2015_ajslp-15-0005]
139. van Balkom TD, Berendse HW, van der Werf YD, Twisk JWR, Peeters CFW, Hoogendoorn AW. Effect of eight-week online cognitive training in Parkinson's disease: a double-blind, randomized, controlled trial. Parkinsonism Relat Disord 2022;96(1):80-87 [FREE Full text] [doi: 10.1016/j.parkreldis.2022.02.018] [Medline: 35248830]
140. van de Weijer SCF, Duits AA, Bloem BR, de Vries NM, Kessels RP, Köhler S. Feasibility of a cognitive training game in Parkinson's disease: the randomized Parkin'Play study. Eur Neurol 2020;83(4):426-432 [FREE Full text] [doi: 10.1159/000509685] [Medline: 32756067]
141. van den Heuvel MRC, Kwakkel G, Beek PJ, Berendse HW, Daffertshofer A, van Wegen EEH. Effects of augmented visual feedback during balance training in Parkinson's disease: a pilot randomized clinical trial. Parkinsonism Relat Disord 2014;20(12):1352-1358 [FREE Full text] [doi: 10.1016/j.parkreldis.2014.09.022] [Medline: 25283070]
142. Wilkinson JR, Spindler M, Wood SM, Marcus SC, Weintraub D, Morley JF. High patient satisfaction with telehealth in Parkinson disease: a randomized controlled study. Neurol Clin Pract 2016;6(3):241-251 [FREE Full text] [doi: 10.1212/CPJ.0000000000000252] [Medline: 27347441]
143. Yang WC, Wang HK, Wu RM, Lo CS, Lin KH. Home-based virtual reality balance training and conventional balance training in Parkinson's disease: a randomized controlled trial. J Formos Med Assoc 2016;115(9):734-743 [FREE Full text] [doi: 10.1016/j.jfma.2015.07.012] [Medline: 26279172]

XSL•FO
RenderX

144. Yen CY, Lin KH, Hu MH, Wu RM, Lu TW, Lin CH. Effects of virtual reality-augmented balance training on sensory organization and attentional demand for postural control in people with Parkinson disease: a randomized controlled trial. Phys Ther 2011;91(6):862-874. [doi: 10.2522/ptj.20100050] [Medline: 21474638]
145. Yuan RY, Chen SC, Peng CW, Lin YN, Chang YT, Lai CH. Effects of interactive video-game-based exercise on balance in older adults with mild-to-moderate Parkinson's disease. J Neuroeng Rehabil 2020;17(1):91 [FREE Full text] [doi: 10.1186/s12984-020-00725-y] [Medline: 32660512]
146. Zoetewei D, Herman T, Ginis P, Palmerini L, Brozgol M, Thumm PC. On-demand cueing for freezing of gait in Parkinson's disease: a randomized controlled trial. Mov Disord 2024;39(5):876-886. [doi: 10.1002/mds.29762] [Medline: 38486430]
147. Tagliente S, Minafra B, Aresta S, Santacesaria P, Buccoliero A, Palmirotta C. Effectiveness of a home-based computerized cognitive training in Parkinson's disease: a pilot randomized cross-over study. Front Psychol 2024;15:1531688 [FREE Full text] [doi: 10.3389/fpsyg.2024.1531688] [Medline: 39850970]
148. Asadiof F, Hasanpour GZ, Naeim M, Mirzai F, Mohammadi Y. Effectiveness of virtual reality-based cognitive behavioral group therapy on reducing anxiety and depression in Parkinson's patients: a randomized controlled trial. Ann Med Surg (Lond) 2025;87(3):1318-1323. [doi: 10.1097/ms9.0000000000003035]
149. Cubo E, Rohani M, Eissazade N, Garcia-Bustillo Á, Ramírez-Sanz JM, Garrido-Labrador JL. Cost-utility analysis of a coadjutant telemedicine intervention for fall prevention in Parkinson's disease. Eur J Neurol 2025;32(1):e16561. [doi: 10.1111/ene.16561] [Medline: 39838709]
150. Das J, Barry G, Vitorio R, Walker R, McDonald C, Morris R. "It was testing me, but not distressing me." Stroboscopic visual training in Parkinson's disease: a qualitative study. BMC Digit Health 2025;3(1):39. [doi: 10.1186/s44247-025-00182-4]
151. De Vleeschhauwer J, Nackaerts E, D'Cruz N, Zhang Y, Janssens L, Vandenberghe W. The effects of task-specific home-based touchscreen training in people with Parkinson's disease: a pilot randomized controlled trial. J Neurol 2025;272(5):328. [doi: 10.1007/s00415-025-13065-5] [Medline: 40204991]
152. Dhamija RK, Saluja A, Garg D, Chauhan S, Majumdar R, Bhardwaj SB. Teleneurorehabilitation and motor and nonmotor symptoms and quality of life in Parkinson disease: the TELEPARK randomized clinical trial. JAMA Neurol 2025;82(4):376-383. [doi: 10.1001/jamaneurol.2024.5387] [Medline: 39992667]
153. Harpham C, Gunn H, Marsden J, Bescos-Garcia R, Murgatroyd C, Connolly L. Home-based high-intensity interval training for people with Parkinson's: a randomized, controlled, feasibility trial. Health Sci Rep 2025;8(7):e71024 [FREE Full text] [doi: 10.1002/hsr2.71024] [Medline: 40661734]
154. Kratz E, Scheffer J, Volc D, Seebacher B. Comparing face-to-face and online LSVT®LOUD speech training using LSVT®Coach in patients with Parkinson's disease: a pilot randomised controlled trial. BMC Neurol 2025;25(1):183 [FREE Full text] [doi: 10.1186/s12883-025-04161-0] [Medline: 40287683]
155. Maggio MG, Rizzo A, Benenati A, Giambò Mauro F, Cannavò A, De Pasquale P. Efficacy and feasibility of virtual reality-based cognitive tele-rehabilitation in Parkinson's disease: a pilot randomized controlled trial on patients and caregivers. Digit Health 2025;11:20552076251376534 [FREE Full text] [doi: 10.1177/20552076251376534] [Medline: 41018508]
156. Nuvolini RA, Silva KGD, Freitas TBD, Doná F, Torriani-Pasin C, Pompeu JE. Exergame-based program and conventional physiotherapy based on core areas of the European Guideline similarly improve gait and cognition in people with Parkinson's disease: randomized clinical trial. Games Health J 2025;14(5):358-368. [doi: 10.1089/g4h.2024.0116] [Medline: 40160128]
157. Pinto C, Norton S, Cubi-Molla P, Hurt C, Chowdhury S, Dooley N. Evaluating a digital mental health intervention for people with Parkinson's (PACT): acceptability and feasibility randomised controlled trial. Aging Ment Health 2025;29(9):1694-1702 [FREE Full text] [doi: 10.1080/13607863.2025.2478508] [Medline: 40135828]
158. Tayyebi G, Asadiof F, Hashempour B, Lotfi M, Taheri M, Naeim M. Efficacy of virtual reality-based cognitive behavioral group therapy in enhancing emotional well-being and quality of life in Parkinson's disease: a randomized controlled trial. Clin Park Relat Disord 2025;12:100316 [FREE Full text] [doi: 10.1016/j.prdoa.2025.100316] [Medline: 40161512]
159. Chen Y, Gao Q, He CQ, Bian R. Effect of virtual reality on balance in individuals with Parkinson disease: a systematic review and meta-analysis of randomized controlled trials. Phys Ther 2020;100(6):933-945. [doi: 10.1093/ptj/pzaa042] [Medline: 32157307]
160. Lina C, Guoen C, Huidan W, Yingqing W, Ying C, Xiaochun C. The effect of virtual reality on the ability to perform activities of daily living, balance during gait, and motor function in Parkinson disease patients: a systematic review and meta-analysis. Am J Phys Med Rehabil 2020;99(10):917-924. [doi: 10.1097/PHM.0000000000001447] [Medline: 32304383]
161. Wang B, Shen M, Wang YX, He ZW, Chi SQ, Yang ZH. Effect of virtual reality on balance and gait ability in patients with Parkinson's disease: a systematic review and meta-analysis. Clin Rehabil 2019;33(7):1130-1138. [doi: 10.1177/0269215519843174] [Medline: 31016994]
162. Triegaardt J, Han TS, Sada C, Sharma S, Sharma P. The role of virtual reality on outcomes in rehabilitation of Parkinson's disease: meta-analysis and systematic review in 1031 participants. Neurol Sci 2020;41(3):529-536 [FREE Full text] [doi: 10.1007/s10072-019-04144-3] [Medline: 31808000]
163. Chuang CS, Chen YW, Zeng BY, Hung CM, Tu YK, Tai YC. Effects of modern technology (exergame and virtual reality) assisted rehabilitation vs conventional rehabilitation in patients with Parkinson's disease: a network meta-analysis of randomised controlled trials. Physiotherapy 2022;117:35-42. [doi: 10.1016/j.physio.2022.07.001] [Medline: 36242929]

XSL•FO
RenderX

164. Parra AG, Gonzalez-Medina G, Perez-Cabezas V, Casuso-Holgado MJ, Vinolo-Gil MJ, García-Muñoz C. Dropout rate of participants in randomized clinical trials that use virtual reality to train balance and gait in Parkinson's disease. A systematic review with meta-analysis and meta-regression. J Med Syst 2023;47(1):46. [doi: 10.1007/s10916-023-01930-7] [Medline: 37010723]
165. Kwon SH, Park JK, Koh YH. A systematic review and meta-analysis on the effect of virtual reality-based rehabilitation for people with Parkinson's disease. J Neuroeng Rehabil 2023;20(1):94 [FREE Full text] [doi: 10.1186/s12984-023-01219-3] [Medline: 37475014]
166. Sarasso E, Gardoni A, Tettamanti A, Agosta F, Filippi M, Corbetta D. Virtual reality balance training to improve balance and mobility in Parkinson's disease: a systematic review and meta-analysis. J Neurol 2022;269(4):1873-1888. [doi: 10.1007/s00415-021-10857-3] [Medline: 34713324]
167. Cortés-Pérez I, Zagalaz-Anula N, Montoro-Cárdenas D, Lomas-Vega R, Obrero-Gaitán E, Osuna-Pérez MC. Leap motion controller video game-based therapy for upper extremity motor recovery in patients with central nervous system diseases. A systematic review with meta-analysis. Sensors (Basel) 2021;21(6):2065 [FREE Full text] [doi: 10.3390/s21062065] [Medline: 33804247]
168. León-Salas B, González-Hernández Y, Infante-Ventura D, de Armas-Castellano A, García-García J, García-Hernández M. Telemedicine for neurological diseases: a systematic review and meta-analysis. Eur J Neurol 2023;30(1):241-254. [doi: 10.1111/ene.15599] [Medline: 36256522]
169. Chen YY, Guan BS, Li ZK, Yang QH, Xu TJ, Li HB, et al. Application of telehealth intervention in Parkinson's disease: a systematic review and meta-analysis. J Telemed Telecare 2018;26(1-2):3-13. [doi: 10.1177/1357633x18792805]
170. Alnajjar AZ, Abouelmagd ME, Krayim A, AbdelMeseh M, Bushara N, Nabil Y. Efficacy of cognitive behavioral therapy for anxiety and depression in Parkinson's disease patients: an updated systematic review and meta-analysis. Neurol Sci 2024;45(11):5277-5290. [doi: 10.1007/s10072-024-07659-6] [Medline: 38958793]
171. Tao Y, Luo J, Tian J, Peng S, Wang H, Cao J. The role of robot-assisted training on rehabilitation outcomes in Parkinson's disease: a systematic review and meta-analysis. Disabil Rehabil 2024;46(18):4049-4067. [doi: 10.1080/09638288.2023.2266178] [Medline: 37818694]
172. Jiang X, Zhou J, Chen Q, Xu Q, Wang S, Yuan L. Effect of robot-assisted gait training on motor dysfunction in Parkinson's patients: a systematic review and meta-analysis. J Back Musculoskelet Rehabil 2024;37(2):253-268. [doi: 10.3233/BMR-220395] [Medline: 37955075]
173. Alwardat M, Etoom M, Al Dajah S, Schirinzi T, Di Lazzaro G, Sinibaldi Salimei P. Effectiveness of robot-assisted gait training on motor impairments in people with Parkinson's disease: a systematic review and meta-analysis. Int J Rehabil Res 2018;41(4):287-296. [doi: 10.1097/MRR.0000000000000312] [Medline: 30119060]
174. Kegelmeyer DA, Minarsch R, Kostyk SK, Kline D, Smith R, Kloos AD. Use of a robotic walking device for home and community mobility in Parkinson disease: a randomized controlled trial. J Neurol Phys Ther 2024;48(2):102-111. [doi: 10.1097/NPT.0000000000000467] [Medline: 38441461]
175. Özden F. The effect of mobile application-based rehabilitation in patients with Parkinson's disease: a systematic review and meta-analysis. Clin Neurol Neurosurg 2023;225:107579. [doi: 10.1016/j.clineuro.2022.107579] [Medline: 36603336]
176. Barrachina-Fernández M, Maitín AM, Sánchez-Ávila C, Romero JP. Wearable technology to detect motor fluctuations in Parkinson's disease patients: current state and challenges. Sensors (Basel) 2021;21(12):4188 [FREE Full text] [doi: 10.3390/s21124188] [Medline: 34207198]
177. Zhang T, Meng DT, Lyu DY, Fang BY. The efficacy of wearable cueing devices on gait and motor function in Parkinson disease: a systematic review and meta-analysis of randomized controlled trials. Arch Phys Med Rehabil 2024;105(2):369-380. [doi: 10.1016/j.apmr.2023.07.007] [Medline: 37532166]
178. Yu J, Wu J, Lu J, Wei X, Zheng K, Liu B. Efficacy of virtual reality training on motor performance, activity of daily living, and quality of life in patients with Parkinson's disease: an umbrella review comprising meta-analyses of randomized controlled trials. J Neuroeng Rehabil 2023;20(1):133 [FREE Full text] [doi: 10.1186/s12984-023-01256-y] [Medline: 37777748]
179. Xue X, Yang X, Deng Z. Efficacy of rehabilitation robot-assisted gait training on lower extremity dyskinesia in patients with Parkinson's disease: a systematic review and meta-analysis. Ageing Res Rev 2023;85:101837. [doi: 10.1016/j.arr.2022.101837] [Medline: 36634871]
180. Lazzarini SG, Mosconi B, Cordani C, Arienti C, Cecchi F. Effectiveness of robot-assisted training in adults with Parkinson's disease: a systematic review and meta-analysis. J Neurol 2024;272(1):22. [doi: 10.1007/s00415-024-12798-z] [Medline: 39666104]
181. Yau CE, Ho ECK, Ong NY, Loh CJK, Mai AS, Tan E. Innovative technology-based interventions in Parkinson's disease: a systematic review and meta-analysis. Ann Clin Transl Neurol 2024;11(10):2548-2562 [FREE Full text] [doi: 10.1002/acn3.52160] [Medline: 39236299]
182. Hagell P, Nygren C. The 39 item Parkinson's disease questionnaire (PDQ-39) revisited: implications for evidence based medicine. J Neurol Neurosurg Psychiatry 2007;78(11):1191-1198 [FREE Full text] [doi: 10.1136/jnnp.2006.111161] [Medline: 17442762]

XSL•FO
RenderX

183. Xu H, Merryweather A, Foreman KB, Zhao J, Hunt M. Dual-task interference during gait on irregular terrain in people with Parkinson's disease. Gait Posture 2018;63:17-22. [doi: 10.1016/j.gaitpost.2018.04.027] [Medline: 29702370]
184. Kwok JYY, Lee JJ, Choi EPH, Chau PH, Auyeung M. Stay mindfully active during the coronavirus pandemic: a feasibility study of mHealth-delivered mindfulness yoga program for people with Parkinson's disease. BMC Complement Med Ther 2022;22(1):37 [FREE Full text] [doi: 10.1186/s12906-022-03519-y] [Medline: 35130894]
185. Li H, Lewis C, Chi H, Singleton G, Williams N. Mobile health applications for mental illnesses: an Asian context. Asian J Psychiatr 2020;54:102209 [FREE Full text] [doi: 10.1016/j.ajp.2020.102209] [Medline: 32623190]
186. Araya R, Menezes PR, Claro HG, Brandt LR, Daley KL, Quayle J. Effect of a digital intervention on depressive symptoms in patients with comorbid hypertension or diabetes in Brazil and Peru: two randomized clinical trials. JAMA 2021;325(18):1852-1862 [FREE Full text] [doi: 10.1001/jama.2021.4348] [Medline: 33974019]
187. Ouedraogo I, Some BMJ, Oyibo K, Benedikter R, Diallo G. Using serious mobile games to improve health literacy in rural Sub-Saharan Africa: a literature review. Front Public Health 2022;10:768252 [FREE Full text] [doi: 10.3389/fpubh.2022.768252] [Medline: 36466518]

Abbreviations

DHI: digital health intervention
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
LMIC: low- and middle-income country
PD: Parkinson disease
PI: prediction interval
PICOS: Population, Intervention, Comparison, Outcomes, and Study Type
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
QM: model Q statistic
RCT: randomized controlled trial
SMD: standardized mean difference

Edited by S Brini; submitted 01.Jul.2025; peer-reviewed by I Adefolaju, X Liang; comments to author 11.Sep.2025; accepted 07.Jan.2026; published 12.Mar.2026.

Please cite as:
Liu R, Zhang S, Xiao Y, Cheng Y, Shang H
The Effects of Digital Health Interventions on Motor Symptoms, Nonmotor Symptoms, and Quality of Life in Patients With Parkinson Disease: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e79935
URL: https://www.jmir.org/2026/1/e79935
doi:10.2196/79935
PMID:

©Ruwen Liu, Sirui Zhang, Yi Xiao, Yangfan Cheng, Huifang Shang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Effects of Wearable Devices on Parkinson Disease: Systematic Review and Meta-Analysis of Randomized Controlled Trials Within the International Classification of Functioning, Disability, and Health Framework

Jiarong Wu[1], MSc; Wanli Zang[1], PhD; Mingqing Fang[2], MM; Ningkun Xiao[3], PhD; Xianzuo Zhang[4], PhD; Su Wang[5], PhD; Qiuxia Zhang[1], PhD

[1]School of Physical Education, Soochow University, Suzhou, Jiangsu, China

[2]Xiangya School of Medicine, Central South University, Changsha, Hunan, China

[3]Laboratory for Brain and Neurocognitive Development, Department of Psychology, Institution of Humanities, Ural Federal University, Yekaterinburg, Russian Federation

[4]Department of Orthopedics, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, China

[5]College of Sports Science and Health, University of Harbin Sport, Harbin, Heilongjiang, China

Corresponding Author:
Qiuxia Zhang, PhD
School of Physical Education
Soochow University
No. 50 East Ring Road
Suzhou, Jiangsu, 215021
China
Phone: 86 13915542576
Email: qxzhang@suda.edu.cn

Abstract

Background: Parkinson disease (PD) impairs gait, balance, and quality of life, and wearable devices have been proposed to support rehabilitation, but evidence for their clinical efficacy remains uncertain.

Objective: This study aimed to evaluate, within the International Classification of Functioning, Disability, and Health (ICF) framework, the effects of wearable-device interventions on gait performance, balance, and health-related quality of life in people with PD by conducting a systematic review and meta-analysis of randomized controlled trials (RCTs).

Methods: We searched PubMed, Web of Science, Cochrane Library, Embase, and ClinicalTrials.gov from inception to November 18, 2025, for RCTs in people with PD comparing wearable-device interventions with control conditions. We used Hartung-Knapp random-effects models to pool mean differences (MDs) or standardized mean differences (SMDs) and reported 95% prediction intervals when ≥3 studies were pooled. Risk of bias was assessed using the Cochrane Risk of Bias (RoB) tool, and certainty of evidence was rated using Grading of Recommendations Assessment, Development, and Evaluation (GRADE).

Results: Nine RCTs involving 260 participants were included. Wearable devices produced a small improvement in stride length (MD 0.10 meter, 95% CI 0.03-0.17), but there was no clear benefit for the 10-Meter Walk Test time (MD 0.04 second, 95% CI –0.06 to 0.15). Double support time showed no reduction (MD –1.59% gait cycle, 95% CI –3.79 to 0.61). Freezing of gait (Freezing of Gait Questionnaire [FOG-Q] and New Freezing of Gait Questionnaire [NFOG-Q]) did not significantly improve (SMD –0.24, 95% CI –0.72 to 0.24). Motor severity (Unified Parkinson Disease Rating Scale Part III [UPDRS III]) showed a small, nonsignificant trend favoring wearable devices (MD –2.16 points, 95% CI –4.39 to 0.07). For balance, pooled results from the Berg Balance Scale (BBS), Mini Balance Evaluation Systems Test (Mini-BESTest), and Performance-Oriented Mobility Assessment Balance Subscale (POMA balance) suggested a borderline effect (SMD 0.48, 95% CI –0.02 to 0.98). Wearable devices did not meaningfully improve Parkinson Disease Questionnaire (PDQ) scores (SMD –0.28, 95% CI –0.74 to 0.17), EQ-5D utility (MD 0.10, 95% CI –0.24 to 0.44), or Falls Efficacy Scale–International (FES-I) scores (MD –0.04, 95% CI –1.10 to 1.02). Prediction intervals frequently crossed the null, suggesting effects may vary by setting and population.

XSL•FO
RenderX

Conclusions: Wearable device interventions for Parkinson disease produced a small improvement on average in stride length, with no consistent benefits for other gait outcomes, balance, or patient-centered outcomes. By integrating ICF mapping with Hartung-Knapp meta-analysis, prediction intervals, and GRADE, and avoiding pooling of conceptually distinct gait measures used in prior reviews, this review clarifies where evidence is most consistent, supports using wearables as adjuncts to rehabilitation, and underscores the need for larger, longer RCTs with standardized outcomes to determine who benefits and how to implement them.

Trial Registration: PROSPERO CRD42024585686; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024585686

(J Med Internet Res 2026;28:e85914) doi:10.2196/85914

KEYWORDS

gait performance; balance function; wearable devices; meta-analysis; International Classification of Functioning, Disability and Health (ICF)

Introduction

Parkinson disease (PD) is the second most common neurodegenerative disorder, characterized by the loss of dopaminergic neurons in the substantia nigra and the presence of Lewy bodies containing α-synuclein [1]. Globally, the estimated number of patients with PD has increased from over 6 million in 1990 to more than 6 million in 2016 [2], and it is expected to double, reaching over 12 million by 2040 [3]. PD poses significant health challenges to patients and places considerable burdens on health care systems and the economy, primarily due to productivity loss and rising health care costs. As the disease progresses, motor symptoms such as bradykinesia, muscle rigidity, resting tremor, and postural instability become more pronounced [4-7]. These changes lead to secondary issues, including diminished walking ability (eg, reduced speed, stride length, and step frequency), increased fall risk, and severe limitations in community participation, which significantly impair patients' quality of life [8-10]. Therefore, restoring walking and balance functions is a crucial aspect of PD rehabilitation, contributing to improved quality of life and independence for patients. While many studies have shown that levodopa can increase stride length and walking speed, treatment options become limited as the disease progresses. Often, after 3-5 years of levodopa use, the duration of its effect decreases, resulting in a worsening of symptoms before the next dose [11,12]. Moreover, long-term use of dopaminergic medications may lead to reduced efficacy and adverse effects such as dyskinesia [13], underscoring the need for alternative and promising PD treatments, such as wearable devices [14-16].

Wearable cueing and sensor-based devices provide a lightweight and portable means of delivering external cues or monitoring motor symptoms in everyday environments, without requiring the continuous presence of health care professionals [17-20]. In PD, these technologies have been used to provide auditory, visual, or somatosensory cues during gait, to quantify tremor and bradykinesia, and to track gait abnormalities, balance, activity levels, and sleep patterns in real time [21-23]. Among these technologies, inertial measurement units (IMUs)—typically combining triaxial accelerometers and gyroscopes into body-worn sensors—have become the most widely used platform for quantitative mobility assessment in both in-clinic and home-based settings [24]. Recent clinical and experimental studies suggest that such devices can acutely improve selected gait parameters and support more individualized adjustment of pharmacological or stimulation therapies [25]. The growing number of clinical and research applications has also prompted international initiatives, such as Movement Disorder Society task forces and national guidance (eg, the 2023 National Institute for Health and Care Excellence recommendations on device-based monitoring in PD), which highlight both the potential and the current uncertainties surrounding wearable technologies in routine care [26]. In addition, qualitative research indicates that long-term use is often limited by issues, such as physical discomfort, interface complexity, privacy concerns, and fluctuating motivation, which may reduce adherence and attenuate real-world effectiveness [27]. Overall, current evidence points to wearable devices as a promising but still incompletely understood adjunct in the management and rehabilitation of people with PD.

Despite this rapid technological development, the specific benefits of wearable devices in improving gait, balance, and quality of life for people with PD have not been fully confirmed. Many studies use observational or small pilot designs, involve heterogeneous device types and training regimens, and focus primarily on laboratory-based gait measures, with relatively less attention to balance, fear of falling, and broader patient-reported outcomes [28,29]. Existing reviews and meta-analyses have often combined wearable and nonwearable cueing interventions or have not applied contemporary random-effects methods, prediction intervals, or Grading of Recommendations Assessment, Development, and Evaluation (GRADE) to formally appraise the certainty of evidence [30]. Moreover, sensor-derived mobility metrics can be influenced by psychological, cognitive, environmental, and technical factors and may differ substantially between supervised laboratory assessments and unsupervised home recordings [30,31], further complicating interpretation of existing data [26,31]. As a result, the overall impact of wearable-device interventions on gait performance, balance, and health-related quality of life in PD remains uncertain.

Therefore, the aim of this study was to conduct a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)-aligned systematic review and meta-analysis of randomized controlled trials (RCTs) to evaluate the effects of wearable-device interventions on gait performance, balance, and health-related quality of life in people with PD, interpreted within the International Classification of Functioning, Disability,

XSL•FO
RenderX

and Health (ICF) framework. We specifically sought to (1) quantify the effects of wearable devices on key gait and motor outcomes (eg, stride length, short-distance gait speed, double support time (DST), freezing of gait, and Unified Parkinson Disease Rating Scale Part III (UPDRS III)), (2) examine their impact on balance and patient-reported outcomes such as quality of life and fear of falling, and (3) appraise the overall certainty of evidence using the GRADE approach. We hypothesized that wearable devices would yield modest improvements in specific gait parameters, whereas evidence for consistent benefits on balance and quality of life would be more limited.

Methods

Overview

This study is based on the ICF framework [32], analyzing the intervention effects of wearable devices on gait, balance, and quality of life in patients with PD. The Population, Intervention, Comparison, and Outcome (PICO) structure is detailed in Table 1. This study has been registered in PROSPERO (International Prospective Register of Systematic Reviews) [33], with registration number CRD42024585686. The review protocol is available in the PROSPERO record.

Table 1. Population, Intervention, Comparison, and Outcome (PICO) framework of wearable device interventions on gait, balance, and quality of life in patients with Parkinson disease under the International Classification of Functioning, Disability, and Health (ICF) framework.

Population and intervention	Comparison	Outcome
Patients with Parkinson disease		
• Intervention settings	Comparison between the wearable device group and the control group	• Gait performance (10MWT[a], Stride length, DST[b], FOG-Q[c], UPDRS III[d])
Hoehn and Yahr stages 1-3		
• Hospital	—[e]	• d450 Walking
• Home and community	—	• d450 Walking
• Laboratory	—	• b730 Muscle power functions
• Intervention personnel	—	• Balance function (BBSf, Mini-BESTestg)
• Physiotherapist	—	• b235 Vestibular functions
• Clinical researchers	—	• d410 Changing basic body position
• Professional instructors	—	• Quality of life (PDQh, EQ-5D-utility, FES-Ii)
• Interventions		
• Use of wearable devices	—	• b130 Energy and drive functions
• Conventional rehabilitation	—	• d920 Recreation and leisure
• Intervention protocol	—	• d570 Looking after one's health
• Type	—	• d760 Family relationships
• Frequency	—	—
• Duration	—	—

[a]10MWT: 10-Meter Walk Test.

[b]DST: double support time.

[c]FOG-Q: Freezing of Gait Questionnaire.

[d]UPDRS III: Unified Parkinson's Disease Rating Scale Part III.

[e]Not applicable.

[f]BBS: Berg Balance Scale.

[g]Mini-BESTest: Mini Balance Evaluation Systems Test.

[h]PDQ: Parkinson Disease Questionnaire.

[i]FES-I: Falls Efficacy Scale–International.

XSL•FO
RenderX

Information Sources

We conducted a comprehensive literature search in PubMed (National Library of Medicine), Embase (Elsevier, Embase.com), the Cochrane Library (Wiley), and Web of Science Core Collection (Clarivate) from inception to November 18, 2025. We also searched ClinicalTrials.gov on November 18, 2025, to identify ongoing or unpublished RCTs evaluating wearable-device interventions in PD.

Databases were searched separately via their native interfaces (ie, no multidatabase searching on a single platform). No search updates were performed after the final search date, and no email alerts were set. No additional online or print sources (eg, journal tables of contents, conference proceedings, or websites) were purposefully searched or browsed.

We screened the reference lists of relevant systematic reviews and all included trials (backward citation searching) to identify additional eligible studies. Reference list screening was performed during full-text assessment and was completed on November 18, 2025. We did not perform forward citation tracking (citing-reference searches) or set up citation alerts. When necessary, we contacted corresponding authors of included studies to request clarification or missing outcome data. No additional information sources or search methods (eg, gray literature databases, preprint servers, or Google Scholar) were used beyond those described above.

Search Strategy

The search strategy combined controlled vocabulary terms (eg, MeSH [Medical Subject Headings] and Emtree) and free-text keywords related to PD, wearable or sensor-based devices, and RCTs. We did not use published search filters (eg, validated RCT filters); trial design was captured using a combination of controlled vocabulary and free-text terms. Search strategies were developed de novo and were not adapted or reused from previous literature reviews.

Searches were limited to English-language publications involving human adults; no restrictions were placed on publication status. These restrictions were applied to align with the eligibility criteria (adult PD populations and RCTs evaluating intervention efficacy) and for feasibility of screening and data extraction.

The search strategy and its reporting followed PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension). The full search strings for each database and information source (including all search terms, Boolean operators, limits, and the date last searched) are provided in S1 in Multimedia Appendix 1. The search strategy was not formally peer-reviewed.

Eligibility Criteria

Inclusion Criteria

Studies were eligible if they met all of the following criteria:

(1) Study design: RCTs evaluating wearable-device interventions in people with PD.

(2) Participants: adults (≥18 years) with a clinical diagnosis of PD based on internationally accepted criteria. Participants were required to be clinically stable; where pharmacotherapy was used (eg, levodopa or other anti-Parkinsonian medications), medication regimens were stable during the trial period (ie, no major medication changes). No restrictions were applied with respect to sex.

(3) Interventions and comparators: the experimental group received a wearable-device intervention (eg, wearable cueing or feedback or sensor-based device, assisted rehabilitation), either as a standalone intervention or as an adjunct to usual care (eg, conventional training and/or pharmacotherapy). Control conditions included usual care without an active wearable-device component and could include conventional training and/or pharmacotherapy, with or without sham or placebo wearable devices.

(4) Outcomes: studies were eligible if they reported at least one prespecified outcome in any of the following domains: gait and mobility, motor severity, balance, and health-related quality of life outcomes. Detailed outcome definitions and measurement instruments are provided in Section 2.6 (Data items—outcomes).

Exclusion Criteria

Studies were excluded if they:

(1) Were not RCTs (eg, nonrandomized studies, observational designs, case series, and qualitative studies);

(2) Enrolled participants younger than 18 years or did not clearly involve a PD population;

(3) Did not evaluate an eligible wearable-device intervention and/or lacked an appropriate control condition;

(4) Did not report any outcomes of interest and/or did not provide sufficient data for extraction (eg, missing summary statistics), and the necessary data could not be obtained from the report or study authors; or

(5) Had unavailable full texts.

For duplicate publications or multiple reports from the same trial, the most complete or most recent report was retained, and companion reports were used to supplement missing details where applicable.

Selection Process

Two reviewers (JW and WZ) independently screened all retrieved records. In the first stage, titles and abstracts were assessed for relevance and study design. In the second stage, full texts were evaluated against the predefined inclusion and exclusion criteria. Any discrepancies were resolved by discussion; if consensus could not be reached, a third reviewer adjudicated. The study selection process was documented using a PRISMA 2020 flow diagram. The completed PRISMA checklist is provided in Multimedia Appendix 2. No automation tools were used for study selection. As only English-language reports were eligible, no translation was required for screening.

XSL•FO
RenderX

Data Collection Process

Data were extracted independently by 2 reviewers (MF and NX) using a piloted, standardized extraction form. Discrepancies in data extraction were resolved by discussion, with adjudication by a third reviewer when necessary. For each prespecified outcome, we extracted data at baseline and at the end of the intervention (postintervention values and/or change scores, as reported), prioritizing the end-of-intervention assessment for the primary synthesis. Risk-of-bias information for each included study was extracted in parallel to support risk-of-bias assessments and GRADE judgments. When required data were missing or unclear, attempts were made to contact study authors; otherwise, available data were analyzed as reported. No automation tools were used for data extraction. As only English-language reports were eligible, no translation was required for data collection.

Data Items (Outcomes)

Outcome data were extracted for the prespecified outcome domains and measurement instruments, with scale directions checked to ensure consistent interpretation. Outcomes of interest included:

1. Gait and mobility: 10-Meter Walk Test (10MWT) time, stride length, DST, and freezing of gait assessed by the Freezing of Gait Questionnaire (FOG-Q) or New Freezing of Gait Questionnaire (NFOG-Q).
2. Motor severity: UPDRS III.
3. Balance: Berg Balance Scale (BBS), Mini Balance Evaluation Systems Test (Mini-BESTest), and Performance-Oriented Mobility Assessment—Balance subscale (POMA balance).
4. Health-related quality of life outcomes: Parkinson Disease Questionnaire (PDQ-39 or PDQ-8), EQ-5D utility index, and Falls Efficacy Scale-International (FES-I).

For each outcome, we extracted the assessment time points and baseline and end-of-intervention data (postintervention values and/or change scores as reported), including the mean and SD or SE and sample size per group.

Data Items (Other Variables)

The following additional data items were extracted:

1. Study characteristics: first author, publication year, country, study design (restricted to RCTs), recruitment source, sample size per arm, follow-up duration, trial registration, ethics approval, funding sources, and declared conflicts of interest.
2. Participant baseline characteristics: age, sex distribution, disease duration, PD severity or stage, medication status (ON or OFF or stable medication), and baseline comparability between groups.
3. Intervention and comparator details: type and modality of wearable device (eg, auditory, visual or vibrotactile cueing, feedback features, stimulation system, and sensor placement), training context (clinic, home, or laboratory), intervention dose (frequency, session duration, or total intervention period), cointerventions, adherence, compliance, dropouts, and adverse events.

Study Risk of Bias Assessment

Risk of bias was assessed at the study level for included RCTs using the Cochrane Risk of Bias tool (RoB 1), following the Cochrane Handbook (version 5.1.0). The following domains were evaluated, including random sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, selective reporting, and other sources of bias. Each domain was judged as low, high, or unclear risk of bias. An overall risk-of-bias judgment was derived as follows: studies were rated low overall if all domains were low risk, high overall if one or more domains were high risk, and unclear overall otherwise. Two reviewers (JW and WZ) independently performed the assessments, and disagreements were resolved through discussion, with adjudication by a third reviewer (QZ) when necessary. Where required, study authors were contacted to clarify information relevant to risk-of-bias judgments. No automation tools were used for risk-of-bias assessment.

Effect Measures

All prespecified outcomes were continuous measures. For outcomes assessed using the same instrument and reported on the same scale across studies, we synthesized effects as mean differences (MDs) with 95% CIs (eg, 10MWT time [seconds], stride length [meters], double support time [% gait cycle], UPDRS III [points], EQ-5D utility, FES-I [points]). When the same construct was assessed using different instruments, we used standardized mean differences (SMDs; Hedges *g*) with 95% CIs (eg, pooling balance outcomes across BBS, Mini-BESTest, and POMA balance; pooling quality-of-life outcomes across PDQ-39 and PDQ-8). Effect directions were checked before synthesis to ensure consistent interpretation across scales (eg, lower scores indicating improvement for PDQ, FOG-Q and NFOG-Q, UPDRS III, and 10MWT time; higher scores indicating improvement for balance scales).

For each outcome, we prioritized end-of-intervention assessments. We preferentially extracted post-intervention values; when only change-from-baseline data were reported, these were used as reported. Where change scores were not reported but could be derived from available pre- and postintervention summary statistics, they were calculated using standard formulas; otherwise, postintervention values were used. For all meta-analyses, effect estimates were presented with 95% CIs and, when 3 or more studies were available, 95% prediction intervals. We did not prespecify universal thresholds for “small, moderate, or large” effects; interpretation emphasized clinical relevance on the original scale (for MDs) and certainty of evidence (GRADE).

Synthesis Methods

Eligibility for Each Synthesis

For each prespecified outcome, studies were eligible for quantitative synthesis if they were randomized controlled trials and reported the relevant outcome with extractable data at baseline and at the end of the intervention for at least one wearable-device versus control comparison. When fewer than 2 studies were available for a given outcome, results were summarized narratively without pooling.

XSL•FO
RenderX

Preparing for Synthesis

Prior to synthesis, outcome definitions, units, and scale directions were checked to ensure consistent interpretation of effect estimates (eg, higher scores consistently indicated better or worse status as appropriate). Where required summary statistics were not directly reported, we attempted to derive them from available information (eg, SEs or 95% CIs) using standard formulas; when this was not possible, study authors were contacted. If key statistics remained unavailable after these attempts, the outcome was not pooled.

Methods Used to Tabulate and Visually Display Results

Study characteristics and extracted outcome data were summarized in tables. Meta-analysis results were presented using forest plots. Risk-of-bias judgments were summarized both graphically (eg, traffic-light plots) and in tabular form.

Methods Used to Synthesize Results

Studies were grouped for synthesis by outcome domain and measurement instrument. We conducted separate meta-analyses for (1) gait performance (10MWT, stride length, double support time, FOG-Q, and NFOG-Q), (2) motor severity (UPDRS III), (3) balance (BBS, Mini-BESTest, and POMA balance), and (4) health-related quality of life outcomes (PDQ-39, PDQ-8, EQ-5D utility, and FES-I). When multiple time points were reported, we prioritized the end-of-intervention assessment for the primary synthesis.

Meta-analyses were conducted in R (meta package; R Core Team). For continuous outcomes, pooled effects were expressed as MD when studies used the same measurement scale or SMD (Hedges *g*) when different instruments were used to assess the same construct, both with 95% CIs. Given expected clinical and methodological variability across interventions, all meta-analyses used a random-effects model [34]. Between-study variance (τ^2) was estimated using restricted maximum likelihood (REML) [35], and the Hartung-Knapp-Sidik-Jonkman method was applied to obtain more robust CIs [36]. Statistical heterogeneity was quantified using I^2 and τ/τ^2 statistics. Where at least three studies were available, 95% prediction intervals were calculated using the bootstrap approach proposed by Nagashima et al [37], (method.predict="NNF" in the meta package; requires the pimeta package) to improve performance in meta-analyses with few studies [38].

Methods Used to Explore Possible Causes of Heterogeneity

Formal investigations of heterogeneity (eg, subgroup analyses, meta-regression, or statistical tests for interaction) were not undertaken because most syntheses included few studies and statistical heterogeneity was generally low, making effect-modifier analyses unreliable.

Sensitivity Analyses

To examine robustness to measurement heterogeneity, prespecified sensitivity analyses were conducted when feasible (typically when 3 or more studies were available) by restricting analyses to studies using the same instrument for a given construct (eg, pooling FOG-Q only when most studies used FOG-Q rather than NFOG-Q; pooling BBS only when multiple balance instruments were used). Sensitivity analysis results were reported alongside the primary analyses.

Reporting Bias Assessment

Because the number of included studies contributing to each synthesis was small, formal assessment of reporting bias due to missing results (eg, funnel plot asymmetry and Egger regression test for small-study effects) was planned only when 10 or more studies were available for a given outcome. When fewer than 10 studies were available, formal statistical and graphical assessments were not undertaken because of limited power and interpretability. In such cases, the potential impact of missing results was considered qualitatively when judging certainty of evidence within the GRADE framework (publication bias domain).

Certainty Assessment

The certainty (confidence) in the body of evidence for each prespecified outcome was assessed using the GRADE approach, considering five domains: risk of bias, inconsistency, imprecision, indirectness, and publication bias [39]. Evidence certainty was categorized as high, moderate, low, or very low. Two reviewers (JW and WZ) independently performed the GRADE assessments, and disagreements were resolved by discussion, with adjudication by a third reviewer (QZ) when necessary.

Results

Study Selection

A total of 2300 relevant articles were identified. After the initial screening, 866 duplicate articles were excluded. Based on the titles and abstracts, a further 1394 articles were excluded. Following a full-text review according to the inclusion and exclusion criteria, 31 additional articles were excluded. Ultimately, 9 studies were included in the analysis [25,40-47]. The literature screening process is shown in Figure 1. Full-text reports that appeared potentially eligible but were excluded, together with reasons for exclusion, are listed in S2 in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram of study selection.

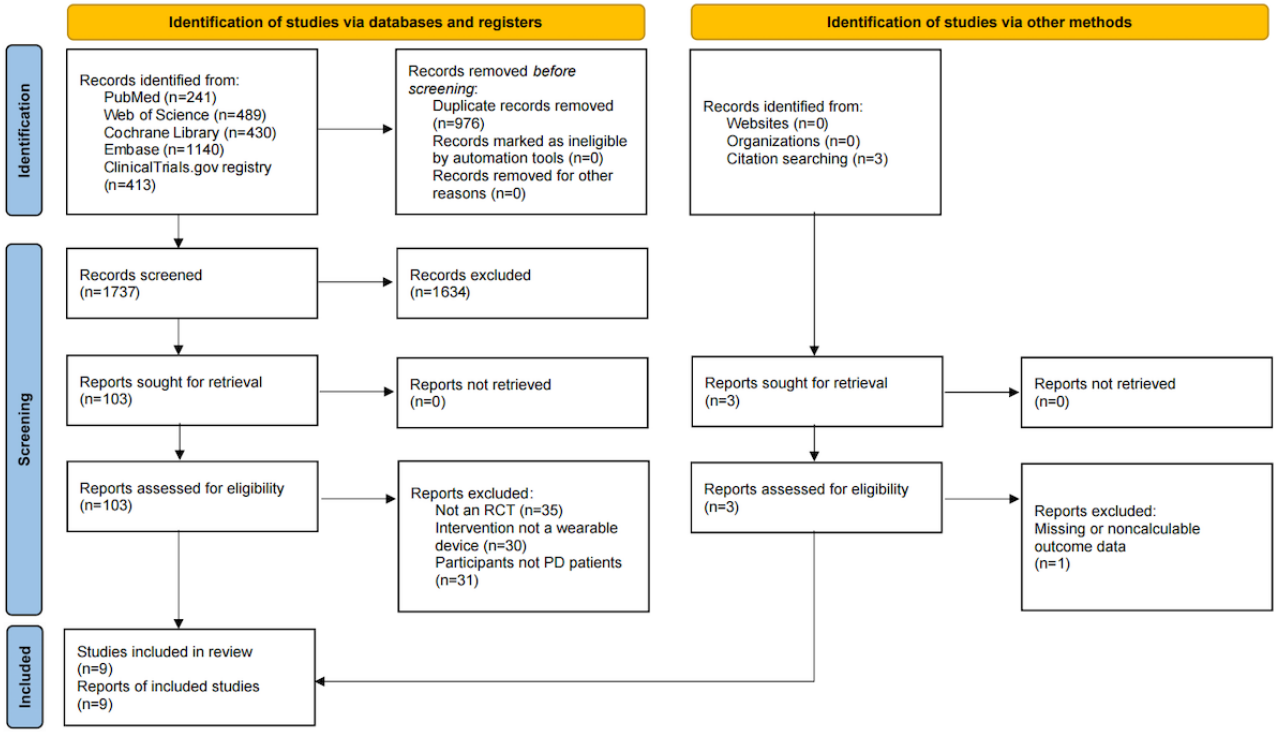

Risk of Bias in Studies

Figures 2A and 2B summarize the risk-of-bias assessment [25,40-47]. Overall, methodological quality was mixed. Most trials reported adequate random sequence generation, whereas details of allocation concealment were frequently insufficient or not reported, leading to many domains being judged as having an unclear risk of bias. Because blinding of participants and personnel was largely not feasible given the nature of wearable-device interventions, several studies were rated at high risk of performance bias. In contrast, blinding of outcome assessors was implemented in approximately half of the trials. Incomplete outcome data and selective reporting were generally judged to be at low risk of bias, with most studies reporting attrition and prespecified outcomes adequately. Other potential sources of bias were usually rated as unclear.

XSL•FO
RenderX

Figure 2. Quality assessment of included studies. (A) Risk of bias summary. (B) Risk of bias graph: percentage summary across all included studies for each type of bias [25,40-46,48].

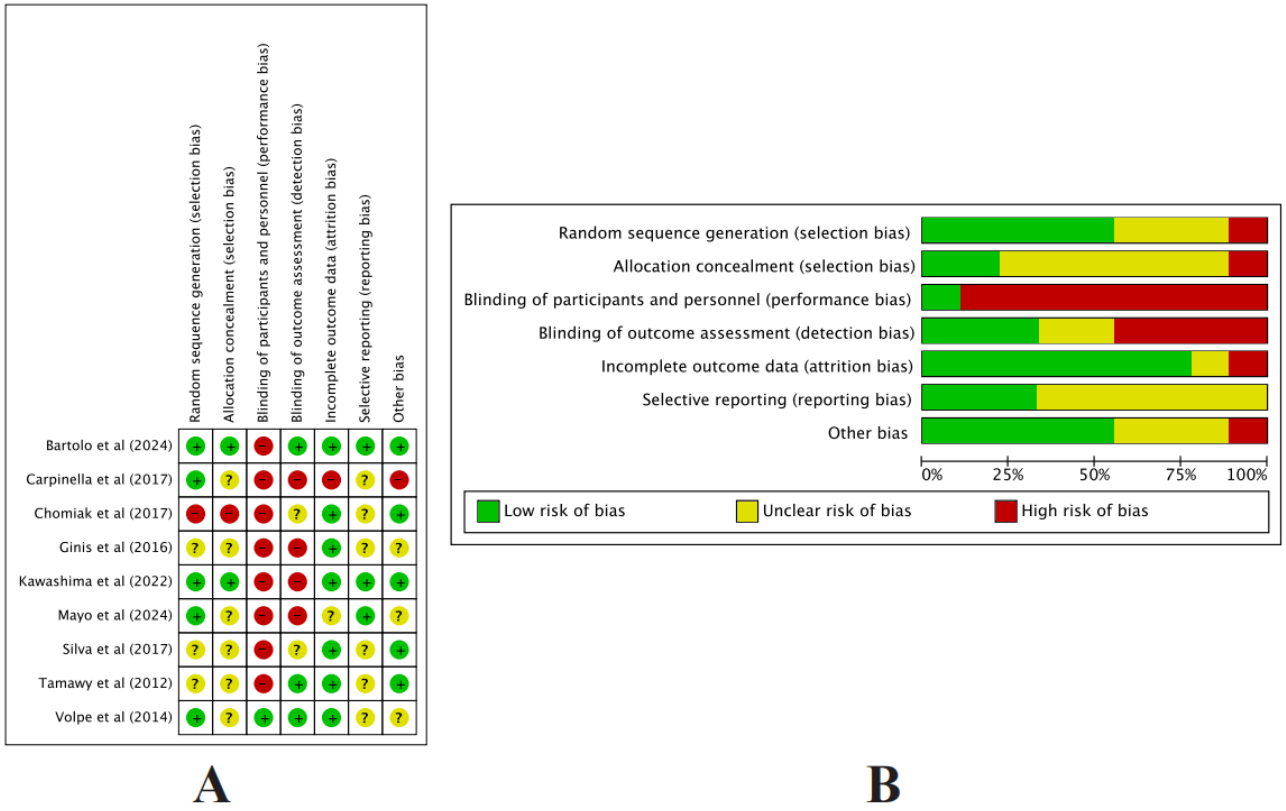

Study Characteristics

A total of 9 studies were included [25,40-47], all of which were published in English, involving a total of 260 PD patients. In most studies, the average age of patients was over 60 years, and the average disease duration exceeded 3 years. Two studies [42,45] conducted 10MWT analysis, 4 studies [40,41,43,46] analyzed stride length, 3 studies [41,43,46] analyzed DST, 4 studies [41,42,44,45] analyzed FOG-Q, 5 studies [41,42,45-47] analyzed UPDRS III, 5 studies [41,42,45-47] analyzed balance function, 3 studies [42,45,47] analyzed PDQ-39, 2 studies [25,45] analyzed EQ-5D utility, and 2 studies [41,44] analyzed FES-I. In most of these studies, patients maintained stable medication regimens before and during the treatment period. Detailed information on the patients and wearable device variables used in these studies is shown in Table 2.

XSL•FO
RenderX

Table 2. Characteristics of included studies

Included studies	T[a]/C[b] (n)	Age in years (T/C), mean (SD)	Hoehn and Yahr stage (T/C)	Disease duration (T/C) years	Wearable prompt device	Feedback characteristics	Stimulation system	Intervention method, (T/C)	Training environment	Intervention duration	Outcome measures
El-Tamawy et al (2012) [40]	15/15	61.4 (7.28)/63.2 (5.6)	2.8 (0.5)/2.6 (0.4)	4.0 (0.9)/3.8 (0.9)	Vibration device placed in the shoes	Proprioception	Open-loop system	Proprioceptive neuromuscular facilitation technique and vibration stimulation + physical therapy program / conventional physical therapy	Laboratory	8 weeks, 3 times/week, 45 minutes/session	Stride length
Volpe et al (2014) [47]	20/20	66.5 / 69.	3.0 / 3.0	6.0 / 6.5	Equistasi nanotech device (Equistasi Srl) placed on C7 and both soleus tendons	Proprioception	Open-loop system	Proprioceptive stimulation (Equistasi) + balance training / balance training only	Laboratory setting (hospital)	8 weeks, 5 times/week, 60 minutes/session	UPDRS III (Unified Parkinson Disease Rating Scale Part III); DST (double support time); BBS (Berg Balance Scale); PDQ (Parkinson Disease Questionnaire)

XSL•FO
RenderX

Included studies	T[a]/C[b] (n)	Age in years (T/C), mean (SD)	Hoehn and Yahr stage (T/C)	Disease duration (T/C) years	Wearable prompt device	Feedback characteristics	Stimulation system	Intervention method, (T/C)	Training environment	Intervention duration	Outcome measures
Ginis et al (2016) [41]	20/18	67.3 (8.13)/ 66.11 (8.07)	2.3 (0.44)/2.2 (0.39)	10.65 (5.39)/11.67 (7.63)	CuPiD system	Auditory feedback	Open-loop system	Real-time gait feedback training provided by the CuPiD system / personalized gait advice	Home	6 weeks, 3 times/week, 30 minutes/session	Stride length; UPDRS III (Unified Parkinson Disease Rating Scale Part III); FOG-Q (Freezing of Gait Questionnaire); DST (double support time); FES-I (Falls Efficacy Scale - International); Mini-BESTest (Mini Balance Evaluation Systems Test)
Carpinella et al (2017) [42]	17/20	73 (7.1)/ 75.6 (8.2)	2.7 (0.7)/2.9 (0.5)	7.5 (3.2)/ 10.3 (5.7)	Gamepad system	Visual and auditory feedback	Open-loop system	Gait and balance training using the Gamepad system / personalized physical therapy	Hospital	20 sessions, 3 times/week, 45 minutes/session	10MWT (10-Meter Walk Test); UPDRS III (Unified Parkinson Disease Rating Scale Part III); FOG-Q (Freezing of Gait Questionnaire); BBS (Berg Balance Scale); PDQ (Parkinson Disease Questionnaire)

XSL•FO
RenderX

Included studies	T[a]/C[b] (n)	Age in years (T/C), mean (SD)	Hoehn and Yahr stage (T/C)	Disease duration (T/C) years	Wearable prompt device	Feedback characteristics	Stimulation system	Intervention method, (T/C)	Training environment	Intervention duration	Outcome measures
Lirani-Silva et al (2017) [43]	10/9	70.4 (6.87)/72 (6.2)	2 (0.5)/ 1.9 (0.4)	N/A[c]	Textured insoles	Proprioceptive feedback	Open-loop system	Using textured insoles / using conventional insoles	Home	1 week	Stride length; DST (double support time)
Chomiak et al (2017) [44]	5/6	70.8 (5.6)/ 69 (5.7)	2.5 (0.50/2.7 (0.41)	15.4 (5.4)/ 11.2 (5)	Ambulosono platform	Auditory feedback	Closed-loop system	Feedback through music playback / Feedback through CBC podcasts	Home	4 weeks, 3 times/week, 10-20 minutes/session	FOG-Q (Freezing of Gait Questionnaire); FES-I (Falls Efficacy Scale - International)
Kawashima et al (2022) [45]	5/7	76.6 (5.3)/ 75.4 (5.7)	2.4 (0.55)/2.4 (0.79)	11.2 (5.8)/ 12.4 (4.6)	Step Management Assist exoskeleton	Proprioceptive feedback	Open-loop + closed-loop system	Gait training using the SMA / conventional gait training	Home-based	3 months, 10 sessions, 30 minutes per session	10MWT (10-Meter Walk Test); stride length; UPDRS III (Unified Parkinson Disease Rating Scale Part III); FOG-Q (Freezing of Gait Questionnaire); BBS (Berg Balance Scale); PDQ (Parkinson Disease Questionnaire); EQ-5D-utility

XSL•FO
RenderX

Included studies	T[a]/C[b] (n)	Age in years (T/C), mean (SD)	Hoehn and Yahr stage (T/C)	Disease duration (T/C) years	Wearable prompt device	Feedback characteristics	Stimulation system	Intervention method, (T/C)	Training environment	Intervention duration	Outcome measures
Bartolo et al (2024) [46]	26/26	73.0 (7.3)/ 70.3 (11)	2 (0.6)/2.1 (0.7)	9.4 (3.1)/ 9.8 (3.9)	Q-Walk system	Visual feedback	Closed-loop system	Gait training using the Q-Walk system / conventional gait training	Clinical rehabilitation	2 weeks, 5 sessions per week, 90 minutes per session	10MWT (10-Meter Walk Test); UPDRS III (Unified Parkinson Disease Rating Scale Part III); DST (double support time); POMA Balance (Performance-Oriented Mobility Assessment Balance Subscale)
Mayo et al (2024) [25]	14/7	70.2 (8.5)/ 70.7 (8.8)	2～3	N/A	Heel2Toe sensor	Auditory feedback	Closed-loop system	Training with Heel2Toe sensor / training according to the manual	Home-based	3 months, twice daily, 5 minutes per session	PDQ (Parkinson Disease Questionnaire); EQ-5D-utility

[a]T: trials (experimental group).

[b]C: controls (control group).

[c]N/A: not applicable.

Results of Syntheses

Gait Performance

10MWT

Two studies [42,45] evaluated the 10MWT outcome. Between-study heterogeneity was negligible (τ^2 0, τ=0; I^2=0%; Q test *P*=.92). The pooled analysis showed no clear evidence of a difference between the wearable-device and control groups in 10MWT (MD 0.04, 95% CI –0.06 to 0.15; *P*=.12; Figure 3A). Prediction intervals were not calculated because fewer than 3 studies were available.

XSL•FO
RenderX

Figure 3. Meta-analyses of gait performance and motor severity outcomes for wearable-device interventions versus control in Parkinson disease. (A) 10-Meter Walk Test (10MWT) time (seconds). (B) Stride length (meters). (C) Double support time (DST, % gait cycle). (D) Freezing of gait, pooled across the Freezing of Gait Questionnaire (FOG-Q) and the New Freezing of Gait Questionnaire (NFOG-Q), expressed as standardized mean differences (SMDs). (E) Sensitivity analysis for freezing of gait restricted to studies using the FOG-Q only, expressed as mean differences (points). (F) Unified Parkinson Disease Rating Scale Part III (UPDRS III) score (points) [40-46,48].

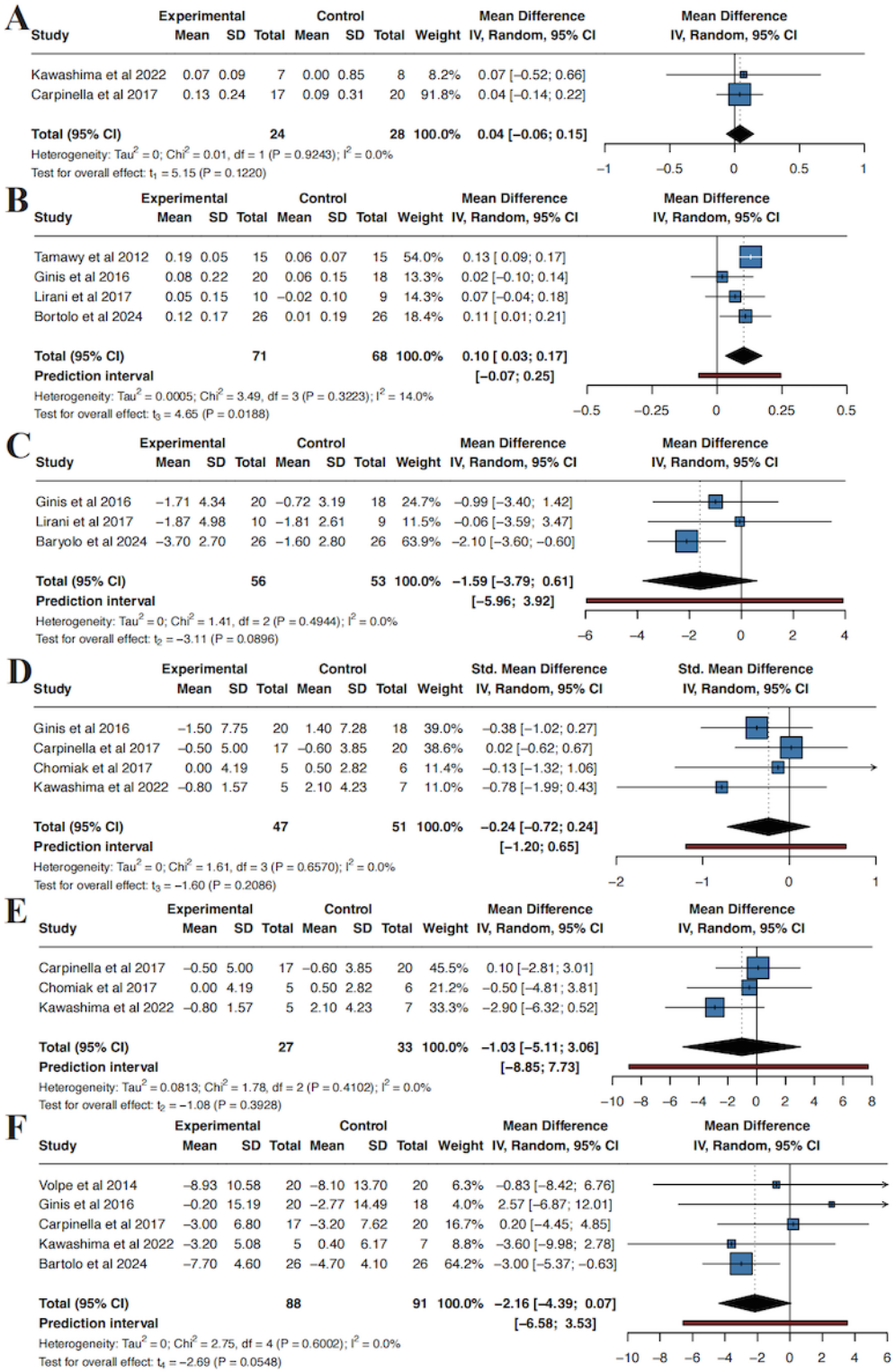

Stride Length

Four studies [40,41,43,46] evaluated stride length (wearable-device group n=71; control group n=68). Between-study heterogeneity was low (τ^2=0.0005; τ=0.02; I^2=14.0%; Q test *P*=.32). Based on our prespecified random-effects model (REML with Hartung–Knapp adjustment), wearable-device interventions were associated with a small but statistically significant improvement on average in stride length compared with controls (MD 0.10 m, 95% CI 0.03-0.17; *P*=.02). The 95% prediction interval was –0.07 to 0.25 m (Figure 3B).

XSL•FO
RenderX

DST

Three studies [41,43,46] evaluated DST (wearable-device group n=56; control group n=53). Between-study heterogeneity was negligible (τ^2=0; τ=0; I^2=0.0%; Q test P=.49). Based on our prespecified random-effects model (REML with Hartung–Knapp adjustment), the pooled analysis showed no clear evidence of a difference in DST between wearable-device interventions and controls (MD –1.59% gait cycle, 95% CI –3.79 to 0.61; P=.09). The 95% prediction interval ranged from –5.96 to 3.92% gait cycle (Figure 3C).

FOG-Q

Four studies [41,42,44,45] evaluated freezing of gait using either the FOG-Q or NFOG-Q (wearable-device group n=47; control group n=51). Between-study heterogeneity was negligible (τ^2=0; τ=0; I^2=0.0%; Q test P=.66). Using our prespecified random-effects model (REML with Hartung–Knapp adjustment), the pooled analysis showed no clear evidence of a difference in freezing of gait between wearable-device interventions and controls (SMD –0.24, 95% CI –0.72 to 0.24; P=.21). The 95% prediction interval was –1.20 to 0.65 (Figure 3D). In sensitivity analyses restricted to studies using the same instrument (FOG-Q only; 3 studies [42,44,45], wearable-device group n=27; control group n=33), heterogeneity remained low (τ^2=0.0813; τ=0.29; I^2=0.0%; Q test P=.41). The pooled effect again showed no statistically significant between-group difference (MD –1.03 points, 95% CI –5.11 to 3.06; P=.39). The 95% prediction interval ranged from –8.85 to 7.73 points (Figure 3E).

UPDRS III

Five studies [41,42,45-47] evaluated UPDRS III (wearable-device group n=88; control group n=91). Between-study heterogeneity was negligible (τ^2=0; τ=0; I^2=0.0%; Q test P=.60). Using the prespecified random-effects model (REML with Hartung–Knapp adjustment), the pooled analysis found no statistically significant difference in UPDRS III between wearable-device interventions and controls (MD –2.16 points, 95% CI –4.39 to 0.07; P=.06). The 95% prediction interval was –6.85 to 3.53 points (Figure 3F).

Balance Function

Three studies [42,45,47] reported results using the BBS, one study [41] reported using the Mini-BESTest, and one study [46] reported using the POMA balance (wearable-device group n=88; control group n=91). Because different scales were used to measure the same construct, results were pooled using SMD. Between-study heterogeneity was low (τ^2=0.0510; I^2=25.5%; Q test P=.25). Using a random-effects model (REML with Hartung–Knapp adjustment), the pooled analysis showed no statistically significant difference in balance between wearable-device interventions and controls (SMD 0.48, 95% CI –0.02 to 0.98; P=.06). The 95% prediction interval ranged from –0.57 to 1.57 (Figure 4A).

XSL•FO
RenderX

Figure 4. Meta-analyses of balance, health-related quality of life, and fear-of-falling outcomes for wearable-device interventions versus control in Parkinson disease. (A) Balance performance expressed as standardized mean differences (SMDs), pooling the Berg Balance Scale (BBS), Mini Balance Evaluation Systems Test (Mini-BESTest), and the balance subscale of the Performance-Oriented Mobility Assessment (POMA balance). (B) Sensitivity analysis for balance restricted to studies using the BBS only, expressed as mean differences (MDs, points). (C) Health-related quality of life assessed with Parkinson Disease Questionnaire (PDQ) instruments, pooling PDQ-39 and PDQ-8 as SMDs. (D) Sensitivity analysis for PDQ restricted to PDQ-39 only, expressed as MDs (points). (E) EQ-5D utility index scores. (F) Fear of falling assessed with the Falls Efficacy Scale–International (FES-I, points) [25,41,42,44-46,48].

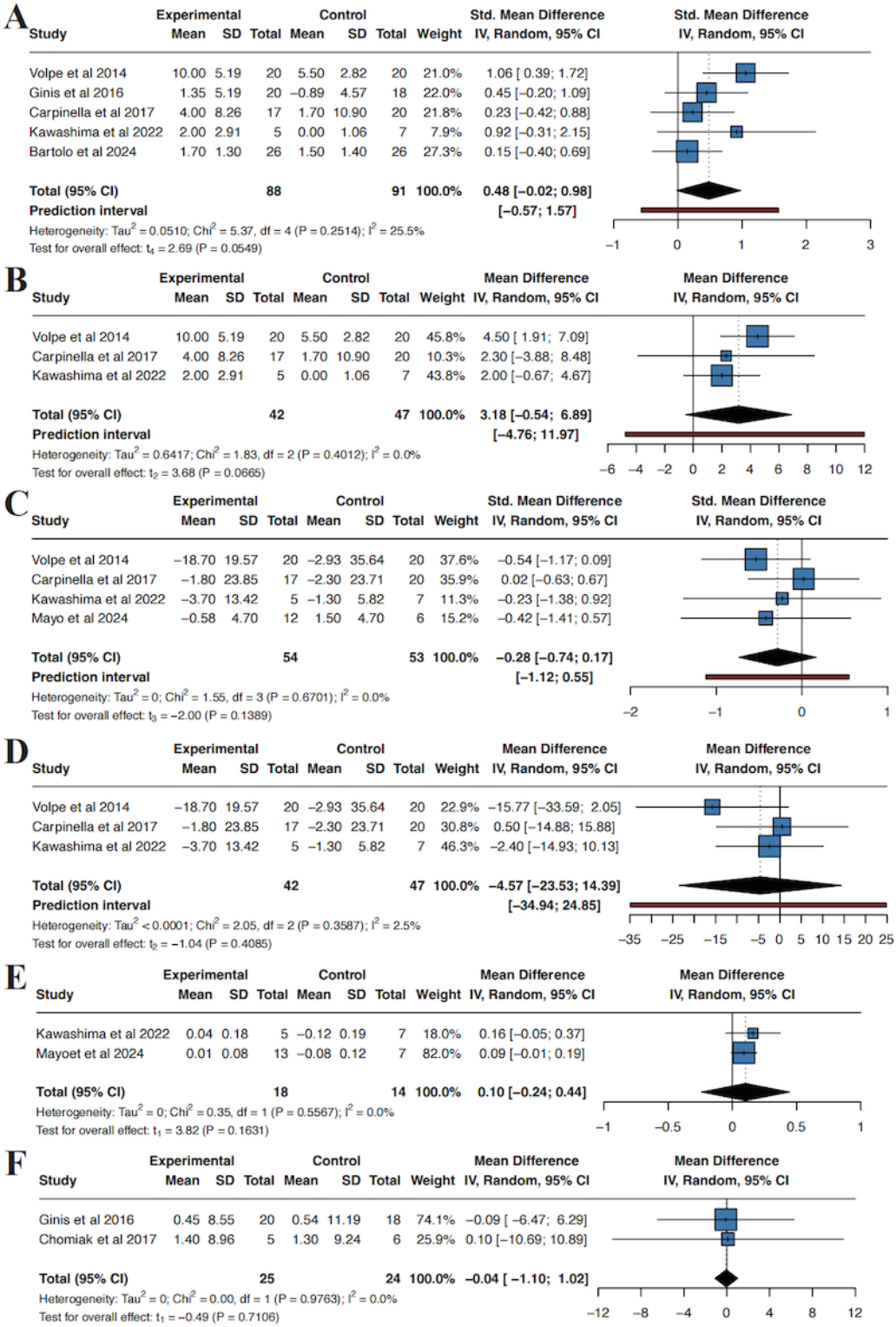

To examine the robustness of this SMD result to measurement heterogeneity, we performed a sensitivity analysis restricted to the three studies using the same balance instrument (BBS; wearable-device group n=42; control group n=47). Heterogeneity remained negligible (τ^2=0.6417; I^2=0%; Q test *P*=.40). The pooled effect was not statistically significant (MD 3.18 BBS points, 95% CI –0.54 to 6.89; *P*=.07). The 95%

XSL•FO
RenderX

prediction interval ranged from –4.76 to 11.97 BBS points (Figure 4B).

Quality of Life

PDQ

Four studies [25,42,45,47] assessed quality of life using PDQ instruments. Three studies used the PDQ-39 [42,45,47], whereas 1 study [25] used the PDQ-8 (wearable-device group n=54; control group n=53). Therefore, the primary meta-analysis pooled these outcomes using SMD (Hedges *g*). Between-study heterogeneity was negligible (τ^2=0; τ=0; I^2=0%; Q test *P*=.67). Based on our prespecified random-effects model (REML with Hartung–Knapp adjustment), the pooled analysis showed no statistically significant difference in PDQ scores between wearable-device interventions and controls (SMD –0.28, 95% CI –0.74 to 0.17; *P*=.14). The 95% prediction interval ranged from –1.12 to 0.55 (Figure 4C). To examine the robustness of the pooled effect to measurement heterogeneity, we repeated the analysis after excluding the PDQ-8 study and pooling only PDQ-39 outcomes using MD (wearable-device group n=42; control group n=47). Heterogeneity remained low (τ^2<0.0001; I^2=2.5%; Q test *P*=.36). The pooled result showed no statistically significant difference between groups (MD –4.57, 95% CI –23.53 to 14.39; *P*=.41). The 95% prediction interval was –34.94 to 24.85 (Figure 4D).

EQ-5D Utility

Two studies [25,45] evaluated EQ-5D utility (wearable-device group n=18; control group n=14). Between-study heterogeneity was negligible (τ^2=0; τ=0; I^2=0.0%; Q test *P*=.56). Using the prespecified random-effects model (REML with Hartung–Knapp adjustment), the pooled analysis showed no statistically significant difference in EQ-5D utility between wearable-device interventions and controls (MD 0.10, 95% CI –0.24 to 0.44; *P*=.16; Figure 4E). Prediction intervals were not calculated because fewer than three studies were available.

FES-I

Two studies [41,44] evaluated fear of falling using the FES-I (wearable-device group n=25; control group n=24). Between-study heterogeneity was negligible (τ^2=0; τ=0; I^2=0.0%; Q test *P*=.98). Using the prespecified random-effects model (REML with Hartung–Knapp adjustment), the pooled analysis showed no statistically significant difference in FES-I scores between wearable-device interventions and controls (MD –0.04 points, 95% CI –1.10 to 1.02; *P*=.71; Figure 4F). Prediction intervals were not calculated because fewer than three studies were available.

Certainty of Evidence

According to GRADE, the overall certainty of evidence ranged from low to very low for all outcomes. For stride length, the certainty was rated as low, reflecting consistent direction of effect across four randomized trials and reasonably precise estimates, but downgrading for very serious risk of bias related to unclear allocation concealment, lack of blinding of outcome assessors, and incomplete outcome data. For all other prespecified outcomes—including 10MWT, DST, freezing of gait, UPDRS III, balance scales, PDQ, EQ-5D utility, and FES-I—the certainty of evidence was judged very low. These ratings were primarily driven by very serious risk of bias and very serious imprecision, with small sample sizes per outcome and wide confidence and prediction intervals that encompassed both no effect and potentially important benefit or harm. Inconsistency and indirectness were generally not considered serious (Table 3). Because the number of included studies was small, we could not formally evaluate funnel plot asymmetry; therefore, the possibility of small-study effects could not be ruled out [48].

XSL•FO
RenderX

Table 3. Certainty of evidence assessment of included studies.

Certainty assessment							Number of patients		Effect			
Number of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Intervention	Comparison	Relative (95% CI)	Absolute (95% CI)	Certainty	Importance
10-Meter Walk Test												
2	Randomized trials	Very serious[a]	Not serious	Not serious	Very serious[b]	None	24	28	—[c]	MD[d] 0.04s (95% CI –0.06 to 0.15)	Very low[a,b]	Important
Stride length												
4	Randomized trials	Very serious[a]	Not serious	Not serious	Not serious	None	71	68	—	MD 0.10 m, (95% CI 0.03-0.17)	Low[a]	Important
Double support time												
3	Randomized trials	very serious[a]	Not serious	Not serious	Very serious[b]	None	56	53	—	MD 1.59 % gait cycle, (95% CI –3.79 to 0.61)	Very low[a,b]	Important
Freezing of Gait Questionnaire												
4	Randomized trials	Very serious[a]	Not serious	Not serious	Very serious[b]	None	47	51	—	SMD[e] –0.24, (95% CI –0.72 to 0.24)	Very low[a,b]	Important
Unified Parkinson Disease Rating Scale, Part III												
5	Randomized trials	Very serious[a]	Not serious	Not serious	Serious[b]	None	88	91	—	MD –2.16 points, (95% CI –4.39 to 0.07)	Very low[a,b]	Important
Balance function												
5	Randomized trials	Very serious[a]	Not serious	Not serious	Serious[b]	None	88	91	—	SMD 0.48, (95% CI –0.02 to 0.98)	Very low[a,b]	Important
Parkinson Disease Questionnaire												
4	Randomized trials	Very serious[a]	Not serious	Not serious	Serious[b]	None	55	53	—	SMD –0.28, (95% CI –0.74 to 0.17)	Very low[a,b]	Important
EQ-5D questionnaire												
2	Randomized trials	Very serious[a]	Not serious	Not serious	Very serious[b]	None	18	14	—	MD 0.10, (95% CI –0.24 to 0.44)	Very low[a,b]	Important
Falls Efficacy Scale–International												

XSL•FO
RenderX

Certainty assessment							Number of patients		Effect			
Number of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Intervention	Comparison	Relative (95% CI)	Absolute (95% CI)	Certainty	Importance
2	Randomized trials	Very serious[a]	Not serious	Not serious	Very serious[b]	None	25	24	—	MD –0.04, (95% CI –1.10 to 1.02)	Very low[a,b]	Important

[a]Risk of bias: downgraded 1 level when most contributing trials had unclear or high risk of bias in one or more key domains (eg, allocation concealment, blinding of outcome assessment, incomplete outcome data); downgraded 2 levels when the pooled evidence was dominated by trials at high risk of bias across multiple key domains.

[b]Imprecision: downgraded 1 level when the total information size was small and the 95% CI included the null effect; downgraded 2 levels when the total information size was very small and the 95% CI was wide, spanning both clinically important benefit and no effect (or harm).

[c]not applicable.

[d]MD: mean difference.

[e]SMD: standardized mean difference.

Discussion

Interpretation

This systematic review and meta-analysis synthesized RCTs evaluating the effects of wearable-device interventions on gait, balance, and quality of life in people with PD, interpreting the findings within the ICF framework [49]. We found a small but statistically significant improvement on average in stride length; however, most other outcomes—including short-distance gait speed, double support time, freezing of gait, balance scales, health-related quality of life, and fear of falling—did not show clear between-group differences. These results suggest that current wearable technologies may have only modest effects on specific gait parameters at the level of body functions and activities, with insufficient evidence for consistent improvement in balance or overall quality of life. Importantly, prediction intervals for several outcomes crossed the null, indicating that effects may vary across populations and implementation contexts and may be minimal or absent in some settings.

Regarding gait and motor outcomes, wearable devices produced a small improvement on average in stride length with low statistical heterogeneity, although the prediction interval suggests that effects may be minimal in some settings, consistent with the idea that cueing and feedback can help increase step amplitude in PD. In contrast, we did not observe clear between-group differences in 10-meter walk performance, double support time, freezing-of-gait scores, or UPDRS III, and the corresponding prediction intervals generally spanned both small benefits and no effect. Overall, these results point to at most modest and uncertain benefits of wearable-device interventions on gait and motor severity. Our findings on short-distance gait performance differ from those of a previous meta-analysis by Zhang et al [30], which reported a statistically significant improvement in gait speed with wearable cueing devices. One likely explanation is that Zhang et al [30] pooled several conceptually different outcomes (10MWT, 6-minute walk test, and treadmill-based speed measures) under a single “gait velocity” construct, whereas we analyzed the 10-Meter Walk Test separately and did not combine it with endurance or treadmill outcomes [30]. In addition, we used a more conservative random-effects approach (REML with Hartung–Knapp adjustment and prediction intervals [36]), which yields wider uncertainty intervals and may partly explain why the apparent speed benefit did not remain statistically significant or robust across settings in our analyses. The small but statistically significant improvement in stride length, in the absence of clear effects on 10MWT or freezing measures, may reflect that step amplitude is more directly modulated by external cueing and feedback than more complex phenomena such as gait initiation, freezing episodes, or sustained walking speed; moreover, many participants were in Hoehn and Yahr stages 1-3 with relatively preserved gait, which may limit the observable incremental benefit of wearable devices on standard clinical tests [50].

For balance-related outcomes, this meta-analysis is, to our knowledge, the first to synthesize randomized evidence on the effects of wearable-device interventions on balance function in people with PD. We pooled BBS, Mini-BESTest, and POMA balance subscores using SMDs, as they reflect a common construct of postural control within the ICF domain of body functions and activities. The combined analysis suggested a borderline improvement in balance favoring wearable devices, but the confidence and prediction intervals were wide and included no effect, indicating substantial uncertainty. A key contextual factor is that, among the included trials, only one study explicitly designed the wearable intervention as a balance-focused training program, whereas the others primarily targeted gait initiation or step regulation; thus, the cueing content and training priorities may not have been optimal for producing measurable changes on balance scales. In contrast, a meta-analysis in a nonspecific older-adult population by Mao et al [51] reported clearer balance gains from sensor-based interventions, suggesting that differences in underlying pathology, intervention content, and training dose between PD cohorts and general older adults may be important.

Quality-of-life and fear-of-falling outcomes map more closely to the ICF domains of participation and contextual or personal factors [49]. Across trials, wearable-device interventions did

XSL•FO
RenderX

not result in statistically significant improvements in PDQ scores, EQ-5D utility, or FES-I scores, and, where calculated, prediction intervals were wide or included no effect. This pattern suggests that short-term, device-focused interventions may not be sufficient to translate modest improvements in gait or motor scores into perceived gains in overall quality of life or reduced fear of falling. Nonmotor symptoms, psychosocial and environmental factors, and real-world participation constraints were rarely targeted explicitly in the included trials, which may further limit the impact of wearable devices on these broader patient-centered outcomes. From a clinical perspective, these findings underscore that wearable devices are unlikely to replace comprehensive multidisciplinary rehabilitation; rather, they may serve as adjunct tools within a broader program that also addresses cognition, mood, balance confidence, and environmental adaptation [52].

When the findings were interpreted in light of heterogeneity, risk of bias, and GRADE assessments, the overall certainty of evidence for most outcomes was low to very low. Although we restricted inclusion to RCTs, several studies had unclear or high risk of bias in domains such as allocation concealment, blinding of outcome assessment, and incomplete outcome data. Many meta-analyses included only 2-5 trials, leading to wide confidence and, where applicable, prediction intervals that encompassed both no effect and potentially clinically relevant benefit or harm. This imprecision, together with some inconsistency in results and indirectness arising from heterogeneous devices and protocols, resulted in downgrading of GRADE ratings for multiple outcomes [39]. Consequently, our conclusions should be interpreted with caution and considered hypothesis-generating rather than definitive.

Limitations of Evidence

This study has several limitations that should be considered when interpreting the results. First, the number of eligible trials and the total sample size for each outcome were modest, limiting statistical power and precision, especially for balance, quality-of-life, and fear-of-falling measures. Second, there was substantial diversity in the technical characteristics and application methods of the wearable devices (eg, cueing modality, feedback algorithms, and sensor placement) and in training protocols (setting, frequency, and duration), which may have diluted or masked device-specific effects. Third, most participants were in Hoehn and Yahr stages 1-3, with relatively preserved motor function and independence, restricting the generalizability of our findings to individuals with more advanced disease. Fourth, intervention periods and follow-up durations were relatively short, so the long-term sustainability of any observed benefits remains unknown. Finally, the number of studies per outcome was too small to formally evaluate funnel plot asymmetry; therefore, small-study effects cannot be ruled out.

Limitations of Review Processes

Several limitations of our review processes should be acknowledged. First, we restricted eligibility to English-language reports, which may have introduced language bias and could have led to missed trials published in other languages. Second, our search did not include gray literature sources (eg, preprint servers, theses, conference proceedings, or general search engines), and we did not perform forward citation tracking; consequently, relevant unpublished or difficult-to-index studies may not have been identified. Third, because all outcomes were informed by fewer than 10 studies, formal assessments of small-study effects (eg, funnel plots and Egger test) were not feasible; therefore, reporting bias due to missing results cannot be excluded. Finally, although we contacted study authors when key information was unclear or missing, incomplete responses may have limited data availability for some syntheses, potentially contributing to imprecision.

Implications

This review provides an updated and methodologically rigorous synthesis of RCT evidence on wearable devices for gait, balance, and quality of life in Parkinson disease. Innovatively, we integrated the ICF framework with contemporary random-effects meta-analysis (Hartung–Knapp adjustment and prediction intervals) and GRADE to contextualize both the pooled average effects and the expected range of effects across real-world settings. Compared with prior reviews, we avoided pooling conceptually different gait outcomes into a single construct, enabling a more implementation-relevant interpretation of heterogeneity and generalizability. This work contributes to the field by clarifying where evidence is most consistent (eg, stride length) and where effects remain uncertain across settings, particularly for balance and patient-centered outcomes. In real-world practice, wearable technologies are best viewed as adjuncts rather than stand-alone rehabilitation strategies and may need to be integrated with task-specific gait and balance training, behavioral strategies, and environmental modifications to translate modest gait gains into meaningful improvements in confidence and quality of life. Future trials should enroll larger and more diverse populations, extend intervention and follow-up periods, standardize ICF-aligned outcomes, and transparently report device characteristics and implementation strategies to clarify for whom, under what conditions, and how wearables can be optimally integrated into routine care.

Conclusion

This systematic review and meta-analysis of randomized controlled trials suggests that wearable-device interventions may provide modest benefits on average for specific gait parameters in PD, particularly stride length, but do not consistently improve short-distance gait speed, double support time, freezing of gait, balance performance, health-related quality of life, or fear of falling. Within the ICF framework, current evidence indicates that wearable devices may support selected body functions and activities, yet there is insufficient and uncertain evidence that they reliably enhance broader participation or overall quality of life.

When interpreted alongside heterogeneity estimates, prediction intervals, risk-of-bias assessments, and GRADE ratings, the certainty of evidence for most outcomes is low to very low. Many meta-analyses were based on a small number of heterogeneous trials with modest sample sizes and relatively short intervention and follow-up durations, leading to wide ranges of plausible effects across settings, including minimal or no benefit. Accordingly, our findings should be viewed as

XSL•FO
RenderX

cautious and hypothesis-generating rather than definitive and do not currently justify the use of wearable devices as a stand-alone rehabilitation strategy in PD.

From a clinical perspective, wearable technologies are best considered as adjuncts to comprehensive multidisciplinary rehabilitation. Combining wearable cueing and feedback with task-specific gait and balance training, education, and psychosocial and environmental interventions may be required to translate small gains in gait parameters into meaningful improvements in balance confidence, participation, and quality of life. Future trials should recruit larger and more diverse patient populations, including those with more advanced disease; standardize outcome measures across ICF domains; extend intervention and follow-up periods; and transparently report device characteristics and implementation strategies to clarify which patients benefit most and how wearable devices can be optimally integrated into routine care.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The analytic code used for the meta-analyses and figure generation and the extracted data underlying the analyses are available from the corresponding author upon reasonable request. The data collection forms and additional materials are available upon request.

Acknowledgments

Generative artificial intelligence (AI; OpenAI; ChatGPT) was used solely to support language editing (eg, grammar checking and rewording for clarity and consistency). It was not used to generate study data, perform analyses, interpret results, or draw conclusions. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the manuscript.

Authors' Contributions

JW, WZ, MF, NX, SW, and QZ contributed to the conceptualization and methodology of the study. JW, WZ, MF, NX, SW, and QZ conducted the investigation. JW, WZ, MF, NX, and QZ performed the formal analysis. WZ, SW, and QZ provided supervision. All authors reviewed and approved the final manuscript.

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Multimedia Appendix 1
Full electronic search strategies for all databases and registers, excluded full-text articles with reasons, and PRISMA-S checklists.
[DOC File , 117 KB - jmir_v28i1e85914_app1.doc]

Multimedia Appendix 2
The PRISMA 2020 statement.
[DOCX File , 22 KB - jmir_v28i1e85914_app2.docx]

References

1. Fu Y, Zhou L, Li H, Hsiao JT, Li B, Tanglay O, et al. Adaptive structural changes in the motor cortex and white matter in Parkinson's disease. Acta Neuropathol 2022;144(5):861-879 [FREE Full text] [doi: 10.1007/s00401-022-02488-3] [Medline: 36053316]
2. GBD 2016 Parkinson's Disease Collaborators. Global, regional, and national burden of Parkinson's disease, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol 2018;17(11):939-953 [FREE Full text] [doi: 10.1016/S1474-4422(18)30295-3] [Medline: 30287051]
3. Dorsey ER, Sherer T, Okun MS, Bloem BR. The emerging evidence of the Parkinson pandemic. J Parkinsons Dis 2018;8(s1):S3-S8 [FREE Full text] [doi: 10.3233/JPD-181474] [Medline: 30584159]
4. Simon DK, Tanner CM, Brundin P. Parkinson disease epidemiology, pathology, genetics, and pathophysiology. Clin Geriatr Med 2020;36(1):1-12 [FREE Full text] [doi: 10.1016/j.cger.2019.08.002] [Medline: 31733690]
5. Foltynie T, Bruno V, Fox S, Kühn AA, Lindop F, Lees AJ. Medical, surgical, and physical treatments for Parkinson's disease. Lancet 2024;403(10423):305-324. [doi: 10.1016/S0140-6736(23)01429-0] [Medline: 38245250]
6. Morris HR, Spillantini MG, Sue CM, Williams-Gray CH. The pathogenesis of Parkinson's disease. Lancet 2024;403(10423):293-304. [doi: 10.1016/S0140-6736(23)01478-2] [Medline: 38245249]
7. Lei J, Tang L, You H. Pathological pain: non-motor manifestations in Parkinson disease and its treatment. Neurosci Biobehav Rev 2024;161:105646. [doi: 10.1016/j.neubiorev.2024.105646] [Medline: 38569983]

XSL•FO
RenderX

8. Pang MY. Physiotherapy management of Parkinson's disease. J Physiother 2021;67(3):163-176 [FREE Full text] [doi: 10.1016/j.jphys.2021.06.004] [Medline: 34154949]
9. Jost ST, Aloui S, Evans J, Ashkan K, Sauerbier A, Rizos A, et al. Neurostimulation for advanced parkinson disease and quality of life at 5 years: a nonrandomized controlled trial. JAMA network open 2024;7:e2352177. [doi: 10.1101/2023.06.07.23291019]
10. Tosin MH, Goetz CG, Stebbins GT. Patient with Parkinson disease and care partner perceptions of key domains affecting health-related quality of life: systematic review. Neurology 2024;102(3):e208028 [FREE Full text] [doi: 10.1212/WNL.0000000000208028] [Medline: 38215353]
11. Poewe W, Espay AJ. Long duration response in Parkinson's disease: levodopa revisited. Brain 2020;143(8):2332-2335. [doi: 10.1093/brain/awaa226] [Medline: 32844192]
12. Ramesh V, Bilal E. Detecting motor symptom fluctuations in Parkinson's disease with generative adversarial networks. NPJ Digit Med 2022;5(1):138 [FREE Full text] [doi: 10.1038/s41746-022-00674-x] [Medline: 36085350]
13. Boelens Keun JT, Arnoldussen IA, Vriend C, van de Rest O. Dietary approaches to improve efficacy and control side effects of levodopa therapy in Parkinson's disease: a systematic review. Adv Nutr 2021;12(6):2265-2287 [FREE Full text] [doi: 10.1093/advances/nmab060] [Medline: 34113965]
14. Bowman T, Pergolini A, Carrozza MC, Lencioni T, Marzegan A, Meloni M, et al. Wearable biofeedback device to assess gait features and improve gait pattern in people with Parkinson's disease: a case series. J Neuroeng Rehabil 2024;21(1):110 [FREE Full text] [doi: 10.1186/s12984-024-01403-z] [Medline: 38926876]
15. Sapienza S, Tsurkalenko O, Giraitis M, Mejia AC, Zelimkhanov G, Schwaninger I, et al. Assessing the clinical utility of inertial sensors for home monitoring in Parkinson's disease: a comprehensive review. NPJ Parkinsons Dis 2024;10(1):161 [FREE Full text] [doi: 10.1038/s41531-024-00755-6] [Medline: 39164257]
16. Kim J, Porciuncula F, Yang HD, Wendel N, Baker T, Chin A, et al. Soft robotic apparel to avert freezing of gait in Parkinson's disease. Nat Med 2024;30(1):177-185. [doi: 10.1038/s41591-023-02731-8] [Medline: 38182783]
17. Babu M, Lautman Z, Lin X, Sobota MH, Snyder MP. Wearable devices: implications for precision medicine and the future of health care. Annu Rev Med 2024;75(1):401-415 [FREE Full text] [doi: 10.1146/annurev-med-052422-020437] [Medline: 37983384]
18. Olsen RJ, Hasan SS, Woo JJ, Nawabi DH, Ramkumar PN. The fundamentals and applications of wearable sensor devices in sports medicine: a scoping review. Arthroscopy 2025;41(2):473-492. [doi: 10.1016/j.arthro.2024.01.042] [Medline: 38331364]
19. Ginsburg GS, Picard RW, Friend SH. Key issues as wearable digital health technologies enter clinical care. N Engl J Med 2024;390(12):1118-1127. [doi: 10.1056/NEJMra2307160] [Medline: 38507754]
20. Longhini J, Marzaro C, Bargeri S, Palese A, Dell'Isola A, Turolla A, et al. Wearable devices to improve physical activity and reduce sedentary behaviour: an umbrella review. Sports Med Open 2024;10(1):9 [FREE Full text] [doi: 10.1186/s40798-024-00678-9] [Medline: 38219269]
21. Ma Y, Zhang Y, Li R, Cheng W, Wu F. The experience and perception of wearable devices in Parkinson's disease patients: a systematic review and meta-synthesis of qualitative studies. J Neurol 2025;272(5):350. [doi: 10.1007/s00415-025-13085-1] [Medline: 40252116]
22. Caroppo A, Manni A, Rescio G, Carluccio AM, Siciliano PA, Leone A. Movement disorders and smart wrist devices: a comprehensive study. Sensors (Basel) 2025;25(1):266 [FREE Full text] [doi: 10.3390/s25010266] [Medline: 39797057]
23. Chatzaki C, Skaramagkas V, Tachos N, Christodoulakis G, Maniadi E, Kefalopoulou Z, et al. The smart-insole dataset: gait analysis using wearable sensors with a focus on elderly and Parkinson's patients. Sensors (Basel) 2021;21(8):2821 [FREE Full text] [doi: 10.3390/s21082821] [Medline: 33923809]
24. Espay AJ, Hausdorff JM, Sánchez-Ferro Á, Klucken J, Merola A, Bonato P, Movement Disorder Society Task Force on Technology. A roadmap for implementation of patient-centered digital outcome measures in Parkinson's disease obtained using mobile health technologies. Mov Disord 2019;34(5):657-663 [FREE Full text] [doi: 10.1002/mds.27671] [Medline: 30901495]
25. Mayo NE, Mate KKV, Fellows LK, Morais JA, Sharp M, Lafontaine A, et al. Real-time auditory feedback for improving gait and walking in people with Parkinson's disease: a pilot and feasibility trial. Pilot Feasibility Stud 2024;10(1):115 [FREE Full text] [doi: 10.1186/s40814-024-01542-z] [Medline: 39192343]
26. Moreau C, Rouaud T, Grabli D, Benatru I, Remy P, Marques A, et al. Overview on wearable sensors for the management of Parkinson's disease. NPJ Parkinsons Dis 2023;9(1):153 [FREE Full text] [doi: 10.1038/s41531-023-00585-y] [Medline: 37919332]
27. Warmerdam E, Hausdorff JM, Atrsaei A, Zhou Y, Mirelman A, Aminian K, et al. Long-term unsupervised mobility assessment in movement disorders. Lancet Neurol 2020;19(5):462-470. [doi: 10.1016/S1474-4422(19)30397-7] [Medline: 32059811]
28. Maddumage Dona SP, Chand PB, Haden C, Kalyani N, Lehn A, Sullivan K, et al. Effectiveness of haptic cues on gait in people with Parkinson's disease - a systematic review and meta-analysis. Gait Posture 2025;122:358-374 [FREE Full text] [doi: 10.1016/j.gaitpost.2025.08.069] [Medline: 40850056]

XSL•FO
RenderX

29. Rovini E, Maremmani C, Cavallo F. How wearable sensors can support Parkinson's disease diagnosis and treatment: a systematic review. Front Neurosci 2017;11:555 [FREE Full text] [doi: 10.3389/fnins.2017.00555] [Medline: 29056899]
30. Zhang T, Meng D, Lyu D, Fang B. The efficacy of wearable cueing devices on gait and motor function in Parkinson disease: a systematic review and meta-analysis of randomized controlled trials. Arch Phys Med Rehabil 2024;105(2):369-380. [doi: 10.1016/j.apmr.2023.07.007] [Medline: 37532166]
31. Sigcha L, Borzì L, Amato F, Rechichi I, Ramos-Romero C, Cárdenas A, et al. Deep learning and wearable sensors for the diagnosis and monitoring of Parkinson's disease: a systematic review. Expert Systems with Applications 2023;229:120541. [doi: 10.1016/j.eswa.2023.120541]
32. Jaiswal A, Paramasivam A, Budhiraja S, Santhakumaran P, Gravel C, Martin J, et al. The international classification of functioning, disability and health (ICF) core sets for deafblindness, part II of the systematic review: linking data to the ICF categories. Eur J Phys Rehabil Med 2024;60(5):893-902 [FREE Full text] [doi: 10.23736/S1973-9087.24.07984-X] [Medline: 39235255]
33. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. bmj 2021;372:n71-n71 [FREE Full text] [doi: 10.31222/osf.io/jb4dx] [Medline: 33782057]
34. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods 2010;1(2):97-111. [doi: 10.1002/jrsm.12] [Medline: 26061376]
35. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023;12(4):101014 [FREE Full text] [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
36. IntHout J, Ioannidis JP, Borm GF. The hartung-knapp-sidik-jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014;14:25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
37. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
38. IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open 2016;6(7):e010247 [FREE Full text] [doi: 10.1136/bmjopen-2015-010247] [Medline: 27406637]
39. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336(7650):924-926 [FREE Full text] [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
40. El-Tamawy M, Darwish M, Khallaf M. Effects of augmented proprioceptive cues on the parameters of gait of individuals with Parkinson's disease. Ann Indian Acad Neurol 2012;15(4):267-272 [FREE Full text] [doi: 10.4103/0972-2327.104334] [Medline: 23349591]
41. Ginis P, Nieuwboer A, Dorfman M, Ferrari A, Gazit E, Canning CG, et al. Feasibility and effects of home-based smartphone-delivered automated feedback training for gait in people with Parkinson's disease: a pilot randomized controlled trial. Parkinsonism Relat Disord 2016;22:28-34 [FREE Full text] [doi: 10.1016/j.parkreldis.2015.11.004] [Medline: 26777408]
42. Carpinella I, Cattaneo D, Bonora G, Bowman T, Martina L, Montesano A, et al. Wearable sensor-based biofeedback training for balance and gait in Parkinson disease: a pilot randomized controlled trial. Arch Phys Med Rehabil 2017;98(4):622-630.e3 [FREE Full text] [doi: 10.1016/j.apmr.2016.11.003] [Medline: 27965005]
43. Lirani-Silva E, Vitório R, Barbieri FA, Orcioli-Silva D, Simieli L, Gobbi LTB. Continuous use of textured insole improve plantar sensation and stride length of people with Parkinson's disease: a pilot study. Gait Posture 2017;58:495-497. [doi: 10.1016/j.gaitpost.2017.09.017] [Medline: 28946113]
44. Chomiak T, Watts A, Meyer N, Pereira FV, Hu B. A training approach to improve stepping automaticity while dual-tasking in Parkinson's disease: a prospective pilot study. Medicine (Baltimore) 2017;96(5):e5934 [FREE Full text] [doi: 10.1097/MD.0000000000005934] [Medline: 28151878]
45. Kawashima N, Hasegawa K, Iijima M, Nagami K, Makimura T, Kumon A, et al. Efficacy of wearable device gait training on Parkinson's disease: a randomized controlled open-label pilot study. Intern Med 2022;61(17):2573-2580 [FREE Full text] [doi: 10.2169/internalmedicine.8949-21] [Medline: 35135928]
46. Bartolo M, Castelli A, Calabrese M, Buttacchio G, Zucchella C, Tamburin S, et al. A wearable system for visual cueing gait rehabilitation in Parkinson's disease: a randomized non-inferiority trial. Eur J Phys Rehabil Med 2024;60(2):245-256. [doi: 10.23736/S1973-9087.24.08381-3] [Medline: 38483335]
47. Volpe D, Giantin MG, Fasano A. A wearable proprioceptive stabilizer (Equistasi®) for rehabilitation of postural instability in Parkinson's disease: a phase II randomized double-blind, double-dummy, controlled study. PLoS One 2014;9(11):e112065 [FREE Full text] [doi: 10.1371/journal.pone.0112065] [Medline: 25401967]
48. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]

XSL•FO
RenderX

49. World Health Organization. International classification of functioning, disability and health: ICF. In: International Classification of Functioning, Disability and Health. Geneva: World Health Organization; 2001.
50. Hoehn MM, Yahr MD. Parkinsonism: onset, progression and mortality. Neurology 1967;17(5):427-442. [doi: 10.1212/wnl.17.5.427] [Medline: 6067254]
51. Mao Q, Zhang J, Yu L, Zhao Y, Luximon Y, Wang H. Effectiveness of sensor-based interventions in improving gait and balance performance in older adults: systematic review and meta-analysis of randomized controlled trials. J Neuroeng Rehabil 2024;21(1):85. [doi: 10.1186/s12984-024-01375-0] [Medline: 38807117]
52. Nielsen C, Siersma V, Ghaziani E, Beyer N, Magnusson SP, Couppé C. Health-related quality of life and physical function in individuals with Parkinson's disease after a multidisciplinary rehabilitation regimen-a prospective cohort feasibility study. Int J Environ Res Public Health 2020;17(20):7668 [FREE Full text] [doi: 10.3390/ijerph17207668] [Medline: 33096677]

Abbreviations

10MWT: 10-Meter Walk Test
BBS: Berg Balance Scale
DST: double support time
FES-I: Falls Efficacy Scale-International
FOG-Q: Freezing of Gait Questionnaire
GRADE: Grading of Recommendations Assessment, Development and Evaluation
ICF: International Classification of Functioning, Disability and Health
IMU: inertial measurement unit
MD: mean difference
MeSH: Medical Subject Headings
Mini-BESTest: Mini Balance Evaluation Systems Test
NFOG-Q: New Freezing of Gait Questionnaire
PD: Parkinson disease
PDQ: Parkinson Disease Questionnaire
PDQ-8: 8-item Parkinson Disease Questionnaire
PDQ-39: 39-item Parkinson Disease Questionnaire
PICO: Population, Intervention, Comparison, and Outcome
POMA balance: Performance-Oriented Mobility Assessment balance subscore
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
PROSPERO: International Prospective Register of Systematic Reviews
RCT: randomized controlled trial
REML: restricted maximum likelihood
RoB: Risk of Bias
SMD: standardized mean difference
UPDRS III: Unified Parkinson Disease Rating Scale Part III

Edited by S Brini; submitted 15.Oct.2025; peer-reviewed by O Enaworu, CY Su; comments to author 14.Nov.2025; accepted 28.Dec.2025; published 18.Mar.2026.

Please cite as:
Wu J, Zang W, Fang M, Xiao N, Zhang X, Wang S, Zhang Q
Effects of Wearable Devices on Parkinson Disease: Systematic Review and Meta-Analysis of Randomized Controlled Trials Within the International Classification of Functioning, Disability, and Health Framework
J Med Internet Res 2026;28:e85914
URL: https://www.jmir.org/2026/1/e85914
doi:10.2196/85914
PMID:

©Jiarong Wu, Wanli Zang, Mingqing Fang, Ningkun Xiao, Xianzuo Zhang, Su Wang, Qiuxia Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet

XSL•FO
RenderX

Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Determinants and Health Outcomes of Digital Health Literacy in Patients With Cardiovascular Disease: Systematic Review and Meta-Analysis

Eun-Jung Kim[1], RN, MSN; Da-Young Kim[1], RN, MSN; Ha-Jin Kim[2], RN, BSN; Youn-Jung Son[3], RN, CCAPN, PhD

[1]Graduate School of Nursing, Chung-Ang University, Seoul, Republic of Korea

[2]Department of Nursing, Asan Medical Center, Seoul, Republic of Korea

[3]Red Cross College of Nursing, Chung-Ang University, Seoul, Republic of Korea

Corresponding Author:
Youn-Jung Son, RN, CCAPN, PhD
Red Cross College of Nursing
Chung-Ang University
Building 103
84 Heuksukro, Dongjak-gu
Seoul, 06974
Republic of Korea
Phone: 82 028205198
Email: yjson@cau.ac.kr

Abstract

Background: With expansion of technology-enabled care, digital health literacy (DHL) has become integral to effective cardiovascular disease (CVD) management. However, quantitative evidence regarding determinants and health outcomes of DHL in CVD remains limited and heterogeneous, necessitating comprehensive evidence synthesis.

Objective: This study aimed to (1) estimate DHL levels, (2) synthesize DHL-associated factors, and (3) examine DHL-related health outcomes in CVD.

Methods: A systematic review and meta-analysis of DHL in adults with CVD was conducted per PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines. PubMed, Embase, Cochrane CENTRAL, CINAHL, Scopus, Web of Science, and Google Scholar were searched for peer-reviewed studies published between 2006 and January 31, 2026. Quantitative studies enrolling adults with CVD, which reported a measure of DHL were included. Studies focusing exclusively on primary cerebrovascular disease and non–peer-reviewed articles were excluded. Risk of bias (ROB) was assessed using the Appraisal Tool for Cross-Sectional Studies tool, the Newcastle-Ottawa Scale, the Revised Cochrane Risk-of-Bias Tool for Randomized Trials, and the Risk of Bias in Nonrandomized Studies of Interventions. Certainty of evidence was evaluated using the Grading of Recommendations Assessment, Development, and Evaluation approach. Pooled mean eHealth Literacy Scale (eHEALS) scores were synthesized using a random-effects meta-analysis. Heterogeneity was quantified using the I^2 statistic and 95% prediction intervals.

Results: Twenty studies involving 8581 adults with CVD were included. The overall pooled mean eHEALS score was 24.26 (95% CI 21.19-27.32), with substantial heterogeneity (I^2=98.4%; τ^2=15.55; τ=3.94) and a wide 95% prediction interval (14.66-33.85). Lower DHL was consistently associated with older age, lower educational attainment, female sex, limited social support, and less experience with digital technologies. Higher DHL was associated with more favorable health-related outcomes, including health behaviors, better quality of life, and greater use and acceptance of digital health technologies. Subgroup analyses showed no statistically significant differences in DHL by region, disease type, or age group. The certainty of evidence was rated as low to very low, and substantial heterogeneity persisted across analyses.

Conclusions: Our findings underscore DHL as a foundational capability for digitally supported self-management in CVD care and reveal disparities associated with age and socioeconomic factors. By integrating evidence on DHL levels, associated factors, and DHL-related health outcomes in CVD populations, this review provides a more comprehensive, clinically relevant understanding of DHL beyond studies relying on a single instrument (eg, eHEALS) or examining isolated domains. DHL appears to be a

XSL•FO
RenderX

context-dependent competency shaped by broader structural and social determinants. From a clinical and health system perspective, digital health interventions should be accompanied by structured digital inclusion strategies, including routine assessment of DHL and care delivery to patients' digital capacities. Further longitudinal and interventional studies are warranted to clarify the causal pathways linking DHL to health outcomes in adults with CVD and to incorporate provider- and system-level perspectives beyond individual-level assessments.

Trial Registration: PROSPERO International Prospective Register of Systematic Reviews CRD420251068000; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251068000

(J Med Internet Res 2026;28:e89102) doi:10.2196/89102

KEYWORDS

cardiovascular disease; digital health; health literacy; systematic review; meta-analysis

Introduction

Cardiovascular disease (CVD) encompasses disorders of the heart and vasculature, including angina, myocardial infarction, stroke, heart failure, arrhythmia, and valvular heart disease [1]. CVD is typically chronic and requires lifelong management. Effective patient self-management, which includes medication adherence, physical activity, and dietary management, is essential for optimal outcomes [2]. Self-management is an active process in which patients understand their conditions, manage symptoms, adhere to prescribed medications, and modify lifestyle habits. These actions are critical for improving health-related quality of life [3]. Digital health technologies have increasingly been applied to support self-management by enabling real-time monitoring, personalized feedback, and behavioral adjustment [4]. Despite variability in effectiveness across modalities and implementation contexts, the adoption of these technologies continues to expand [5,6]. Digital health technologies facilitate the delivery of personalized health information and enable real-time monitoring and feedback through mobile platforms, wearable devices, and remote patient monitoring systems [7,8]. By leveraging digital infrastructures, these technologies facilitate large-scale dissemination of health information across diverse settings, including medically underserved and resource-limited environments, enhancing their relevance to population and public health contexts [7].

However, substantial differences exist in access to digital technologies, digital capabilities, and trust in digital health systems [7,8]. Older adults and socioeconomically disadvantaged populations are more frequently exposed to environments characterized by limited digital health literacy (DHL), restricted access, and reduced trust, thereby increasing the risk of digital exclusion [8]. In a US study of patients with heart failure, more than half of the participants lacked access to mobile health technologies, and limited access was strongly associated with older age, rural residence, and socioeconomic factors [9]. These findings underscore the presence of digital disparities in both access to and effective use of digital health technologies among patients with CVD [9]. Furthermore, the uptake and utilization of digital health technologies are influenced by social determinants of health, including educational attainment, income, race and ethnicity, and sex. DHL has been identified as a key component of digital health equity [10]. Consequently, effective use of digital health technologies is closely associated with DHL, which may determine whether disparities in access translate into successful digitally supported self-management [11].

With the migration of health information delivery to online and mobile platforms, DHL, an extension of traditional health literacy, has emerged as a central construct [12,13]. DHL refers to the ability to access, understand, evaluate, and apply health information obtained in digital environments to support effective health behaviors and improve or maintain health-related quality of life [14,15]. Higher DHL enables individuals to accurately interpret and use digital health information, promoting more effective self-management [16]. Given the complexity and chronicity of CVD, patients require continuous self-management, for which DHL is a critical resource [17].

Studies assessing DHL in CVD have been conducted across diverse settings and have used heterogeneous instruments, most commonly the eHealth Literacy Scale (eHEALS), the Digital Health Literacy Instrument (DHLI), and the eHealth Literacy Questionnaire (eHLQ). These instruments differ in item content and measurement domains depending on their intended purpose and target population and are used to quantify DHL levels [18-20]. In addition, DHL varies across populations according to age, educational attainment, socioeconomic status, and employment status. Older adults and individuals with lower education or income typically report lower DHL, which is associated with reduced use of digital health information and less effective self-management [16,17].

Supporting effective self-management in CVD requires assessment of individual DHL levels, identification of their determinants, and evaluation of associated health outcomes [21]. Despite the widespread availability of mobile apps, remote monitoring tools, and online resources, individuals with low DHL may be unable to use these resources effectively. This limitation increases exposure to inaccurate information and contributes to distrust or avoidance of digital health resources [22]. Emerging evidence suggests that DHL is associated with health behaviors and outcomes in CVD populations; however, quantitative data on DHL levels, determinants, and associations with clinical and patient-reported outcomes remain limited and heterogeneous [21,23].

This systematic review and meta-analysis aimed to assess DHL levels, synthesize factors associated with DHL, and examine DHL-related health outcomes among patients with CVD.

XSL•FO
RenderX

Methods

Study Design, Protocol, and Registration

This systematic review and meta-analysis synthesized factors associated with DHL and its health outcomes among patients with CVD. The protocol was registered with PROSPERO (CRD420251068000). Reporting followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 expanded checklist, the PRISMA 2020 (checklist provided in Multimedia Appendix 1) for abstracts checklist [24], and PRISMA-S (PRISMA literature search extension) [25].

Two deviations from the registered protocol were made. First, although the PROSPERO record specified inclusion of English-language publications only, no language restrictions were applied during the final search to minimize potential language bias. Second, studies that focused exclusively on stroke or transient ischemic attack were excluded during full-text screening to enhance clinical homogeneity.

Eligibility Criteria

Eligibility criteria were based on the participants, intervention or interest, comparisons, outcomes, and study design (PICOS) framework, as follows: adults aged ≥18 years diagnosed with CVD (participants); studies measuring DHL among patients with CVD (intervention or interest); no comparison group was required, as this review did not focus on comparative studies (comparison); determinants of DHL and its health outcomes (outcomes); and quantitative studies and mixed methods studies containing quantitative data were included (study design). Purely qualitative studies and those reporting only DHL scores without examining associated factors or health outcomes were excluded.

For this review, CVD was operationalized as cardiac disease, including coronary artery disease, myocardial infarction, heart failure, arrhythmias, and valvular disease. Primary cerebrovascular disease, such as stroke and transient ischemic attack, was excluded to minimize confounding of DHL assessment by postevent language or cognitive impairment [26]. To capture DHL comprehensively, related constructs including eHealth literacy, electronic health literacy, and mobile health literacy were also considered [14]. The literature search included studies published from 2006 onward, the year in which the concept of eHealth literacy was first introduced [14,27]. No language restrictions were applied. Exclusion criteria were (1) studies not measuring DHL; (2) studies involving children or adolescents; and (3) non–peer-reviewed articles, unpublished works, research protocols, or conference abstracts (Table 1).

Table 1. Eligibility criteria based on the PICOS[a] framework.

Item	Inclusion criteria	Exclusion criteria
Participants	Studies involving adult patients (≥18 years) with CVD[b]	Studies involving children or adolescents (<18 years) or targeting non-CVD populations
Intervention or interest	Studies measuring DHL[c], including eHealth literacy, electronic health literacy, or mobile health literacy	Studies not measuring DHL or reporting only clinician-reported outcomes
Comparison	Not applicable (no comparator required)	Not applicable
Outcomes	Studies reporting DHL levels, associated factors, or health outcomes related to DHL	Studies not reporting DHL outcomes or related factors
Publication period	Published between January 2006 and January 2026	Published before 2006 or after January 2026
Study designs	Original quantitative studies (cross-sectional, cohort, quasi-experimental, randomized controlled trials, or mixed methods with quantitative outcomes)	Reviews, qualitative-only studies, protocols, conference abstracts, dissertations, editorials, or commentaries

[a]PICOS: participants, intervention or interest, comparison, outcomes, and study design.

[b]CVD: cardiovascular disease.

[c]DHL: digital health literacy.

Information Sources

A systematic literature search on DHL in patients with CVD was conducted in PubMed (via the National Library of Medicine), Embase (via Elsevier), Cochrane Central Register of Controlled Trials (CENTRAL; via EBSCOhost), CINAHL Plus with Full Text (via EBSCOhost), Scopus (via Elsevier), Web of Science (via Clarivate), and Google Scholar. No additional study registries or online resources were searched. Studies published from 2006 through June 30, 2025, were included. To ensure inclusion of the most recent literature, an updated search was conducted on January 31, 2026, covering the period up to that date. Reference lists of all included studies were manually screened for additional relevant publications; however, no further eligible studies were identified. We contacted the corresponding authors of studies with missing or incomplete data to request additional information; however, no additional data were provided.

Search Strategy

A combination of MeSH (Medical Subject Headings) terms, free-text keywords, and Boolean operators was used to ensure comprehensive retrieval. The search strategy was independently developed for this study by synthesizing terms from preliminary literature reviews and MeSH databases, without using pre-established search filters or existing search strings. Search terms were developed based on DHL and CVD. In Google Scholar, where Boolean operators and advanced filters are limited, the main keywords (“digital health literacy or eHealth literacy” AND “cardiovascular disease”) were applied. The

XSL•FO
RenderX

search strategy was developed in consultation with a librarian and reviewed by the research team prior to implementation. Detailed information on the search process is provided in Multimedia Appendix 2.

Study Selection

All search records were imported into EndNote 21, and duplicate records were removed electronically. After removing duplicates, the remaining records were exported to Microsoft Excel for screening. During the initial screening of titles and abstracts, the reviewers manually identified and removed any additional duplicates not detected by the software. Two reviewers (EJK and YJS), both experienced in conducting systematic reviews and meta-analyses, independently assessed the records against the inclusion and exclusion criteria and evaluated full texts for eligibility. Discrepancies were resolved through discussion until consensus was reached, and eligible studies were included in the review.

Data Extraction

Data were extracted on study descriptors (first author, publication year, and country), participant characteristics (sample size, age, and sex), study measures (DHL measurement tools, scale structure and domains, and DHL level), and outcome indicators (factors associated with DHL and related health outcomes). Adjusted estimates for factors associated with DHL and health outcomes were extracted whenever available. When both unadjusted and adjusted results were reported, the final adjusted estimates from multivariable analysis were included. When effect estimates could be derived from reported values, reviewers systematically calculated and extracted them.

Risk of Bias Assessment

Risk of bias (ROB) and methodological quality were independently assessed by two reviewers (EJK and YJS). The Appraisal Tool for Cross-Sectional Studies (AXIS) was used for cross-sectional studies, the Newcastle-Ottawa Scale (NOS) for cohort or case-control studies, the Revised Cochrane Risk-of-Bias Tool for Randomized Trials (RoB 2), and the Risk of Bias in Nonrandomized Studies of Interventions (ROBINS-I) for nonrandomized interventional studies [28-31]. Disagreements were resolved through consensus.

AXIS comprises 20 items covering study design, sampling, analysis, reporting, and discussion, with each item rated as yes, no, or do not know [28]. NOS evaluates selection, comparability, and outcome, awarding stars within each domain; total scores range from 0 to 9, with higher values indicating better quality [29]. RoB 2 appraises 5 domains: the randomization process, deviations from intended interventions, missing outcome data, measurement of outcomes, and selection of the reported result. Each domain is rated as low risk, some concerns, or high risk. Overall risk is classified as low if all domains are low risk, as some concerns if at least one domain raises concerns but none are high risk, and as high if any domain is high risk or multiple domains raise concerns [30]. ROBINS-I assesses 7 domains: confounding, participant selection, intervention classification, deviations from intended interventions, missing data, outcome measurement, and selection of the reported result. Each domain is graded as low, moderate, serious, or critical ROB, with an overall judgement derived accordingly [31].

Data Analysis

When at least two studies reported mean eHEALS scores, the means were meta-analyzed using an inverse-variance–weighted random-effects model [32,33]. To avoid double-counting in these analyses, multiple publications from the same cohort were checked, and country-specific studies with clearly defined samples were prioritized to ensure the independence of observations. For assessing determinants and health outcomes of DHL, effect estimates were extracted where reported. Owing to substantial heterogeneity in outcome definitions and analytical approaches, these findings were synthesized narratively.

Random-effects meta-analyses were conducted using the Hartung-Knapp-Sidik-Jonkman method to obtain more robust CIs. This approach accounts for uncertainty in estimating between-study variance and is particularly appropriate when heterogeneity is high or the number of studies is limited [34]. Heterogeneity was evaluated using Cochran Q and quantified with Higgins I^2; values of 25%, 50%, and 75% were interpreted as low, moderate, and high heterogeneity, respectively [35]. Although I^2 was reported as a descriptive measure of heterogeneity, it was interpreted with caution because it does not reflect the magnitude of variability in true effects across populations or settings.

Therefore, where applicable, 95% prediction intervals were calculated to quantify between-study heterogeneity in a clinically interpretable manner and to facilitate the real-world interpretation of the pooled estimates. Prediction intervals were derived from the random-effects model, incorporating the estimated between-study variance [36].

Meta-regression was prespecified to explore potential sources of heterogeneity; however, it was not conducted owing to the limited number of included studies [37]. Subgroup analyses were performed to examine variability in mean scores according to three predefined criteria: (1) study region (Asia vs Europe), (2) disease type (heart failure only vs mixed CVD), and (3) mean age (<65 years vs ≥65 years). A leave-one-out sensitivity analysis was performed to assess robustness by examining the contribution of each study to the pooled estimate [38]. All analyses were conducted in R (version 4.4.2; R Foundation for Statistical Computing) using the meta and metafor packages [39-41]. Small-study effects were not assessed in meta-analyses including fewer than 10 studies, as funnel plots and related statistical tests, such as Egger regression, have limited power in analyses with a small number of studies [42,43].

Certainty of the Evidence

The certainty of evidence for each key outcome was assessed using the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach. Assessment domains included ROB, inconsistency, indirectness, imprecision, and publication bias. The GRADE approach was applied to associations between DHL and health outcomes. Based on this assessment, the certainty of evidence for each outcome was classified as high, moderate, low, or very low. Two reviewers

XSL•FO
RenderX

(EJK and DYK) independently assessed the certainty of evidence for each outcome, and any disagreements were resolved through consensus [44].

Ethics Approval

The institutional review board of Chung-Ang University approved the study protocol (1041078-20250603-HR-172).

Results

Study Selection

The initial database search, conducted through June 30, 2025, yielded 17,001 records. Screening and eligibility assessment were performed according to prespecified criteria, and the study selection process is summarized in the PRISMA 2020 flow diagram (Figure 1). After exporting records to EndNote 21 and removing duplicates, 13,341 records remained for title and abstract screening. Of these, 1578 records were excluded owing to incompatible study designs (eg, qualitative studies, protocols, and letters), 4943 records were excluded because participants were not adults with CVD, 6615 records were excluded because DHL was not assessed, and 129 records were removed as remaining duplicates. Full texts of 76 studies were assessed for eligibility, and 58 were excluded owing to ineligible study design (eg, conference abstracts and protocols), ineligible population, absence of DHL assessment, or lack of DHL-related factors or health outcomes. Eighteen studies were initially selected from the primary search. To ensure inclusion of the most recent literature, an updated search was conducted in January 2026, yielding 40 additional records. Following the same screening and eligibility process, 38 records were excluded, and 2 additional studies were included. Consequently, a total of 20 studies met the inclusion criteria and were included in this review (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study selection. This figure illustrates the process of identification, screening, eligibility assessment, and inclusion of studies in the systematic review in accordance with the PRISMA 2020 guidelines. Records were identified through database searching up to June 30, 2025, with an updated search conducted in January 2026. DHL: digital health literacy.

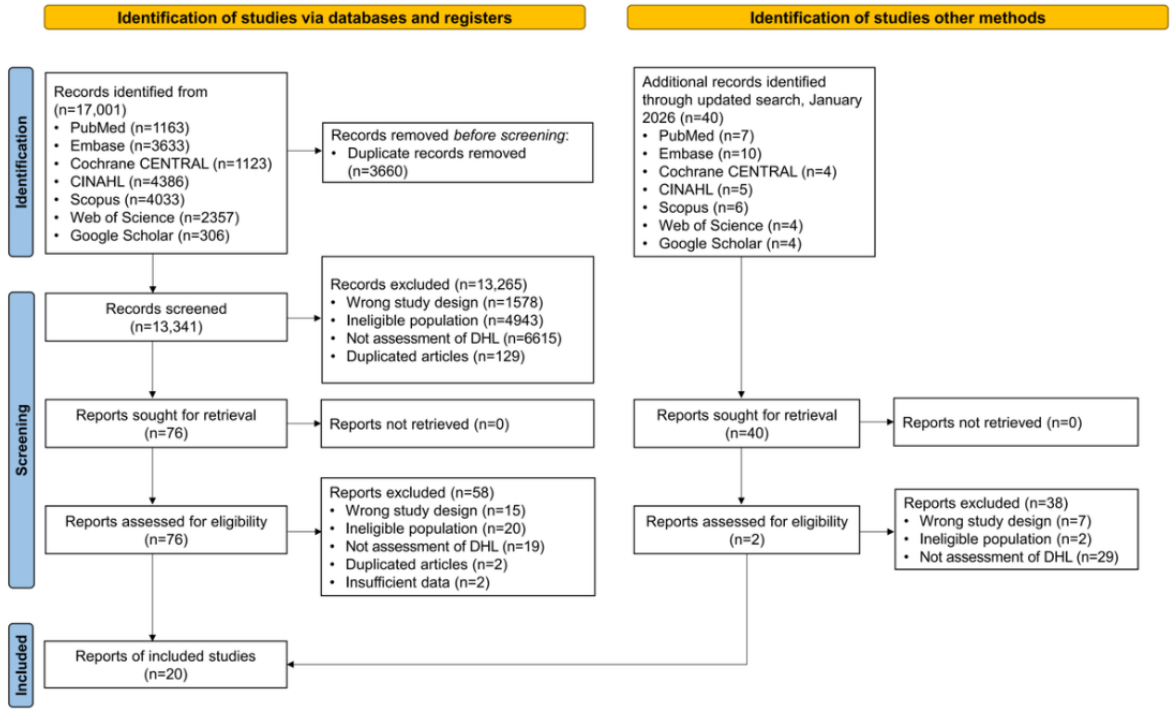

Study Characteristics

The 20 studies included in this review were published from 2018 onward, with 14 published after 2023. Most studies used a cross-sectional design (n=11, 55%), followed by 3 quasi-experimental studies and 2 studies each with randomized controlled trial, longitudinal observational, and prospective observational designs. In 5 experimental and cohort studies [45-49], DHL was assessed at multiple time points. The studies were conducted across several countries, predominantly in Europe and Asia.

Heart failure was the most common condition among the study populations (n=10, 50%). Participants included patients who had undergone coronary artery bypass grafting, those with ischemic heart disease, those who had valve surgery or percutaneous coronary intervention, and those with atrial fibrillation, cardiac arrhythmias with a cardiac implantable electronic device, atherosclerosis, or coronary artery disease. The reported mean age of participants ranged from 57.6 to 79.9 years. Most studies recruited adults aged ≥18 years, although some enrolled specific age groups, such as those aged ≥40 years [49], ≥45 years [50], or ≥60 years [51,52]. The total sample size across all studies was 8581 participants, with proportions of males ranging from 49.5% to 83% and females ranging from 17% to 50.5% (Table 2).

XSL•FO
RenderX

Table 2. Characteristics of included studies (N=20).

Authors (year)	Study design	Country	Health status	Sample size	Age (years)	Sex (female), n (%)
Melholt et al (2018) [45]	Quasi-experimental	Denmark	CABG[a], HF[b], IHD[c], VS[d]	49	• Mean 60.64 (SD 10.75)	9 (18.4)
Chuang et al (2019) [53]	Cross-sectional	Taiwan	HF	141	• Mean 65.2 (SD 11.9)	63 (44.7)
Lin et al (2020) [51]	Longitudinal	Iran	HF	468	• Mean 69.3 (SD 7.3)	230 (49.2)
Rodríguez Parrado et al (2022) [46]	Quasi-experimental	Colombia	HF	28	• Mean 67.3 (SD 12.9)	7 (25.0)
Spindler et al (2022) [47]	Randomized controlled trial	Denmark	HF	137	• Intervention group: mean 61.73 (SD 10.75) • Control group: mean 61.36 (SD 11.46)	32 (23.4)
Yun et al (2022) [54]	Randomized controlled trial	Spain	HF	178	• Median 74	73 (41.0)
Bakhshayesh et al (2023) [55]	Cross-sectional	Iran	HF	200	• ≤50 years: n=45 (22.5%) • >50 years: n=155 (77.5%)	70 (35.0)
Bäuerle et al (2023) [56]	Cross-sectional	Germany	CHF[e], IHD	290	• Mean 57.59 (SD 13.33)	149 (51.4)
Brørs et al (2023) [48]	Prospective	Denmark and Norway	PCI[f]	2924	• Mean 66 (SD 11)	582 (19.9)
Ramstad et al (2023) [57]	Prospective	Norway	PCI	1970	• Mean 66 (SD 11)	427 (21.7)
Rush et al (2023) [58]	Cross-sectional	Canada	AF[g]	195	• Mean 65.36 (SD 10.32)	73 (37.4)
Son et al (2023) [49]	Quasi-experimental	Korea	HF	100	• Mean 58.78 (SD 8.83)	17 (17.0)
Vitolo et al (2023) [59]	Cross-sectional	Italy	Arrhythmia, CIED[h] recipient	300	• Median 75 (IQR 66-84)	109 (36.3)
Mohajeri et al (2024) [60]	Cross-sectional	Germany	Atherosclerosis	310	• Mean 61.72 (SD 11.44)	130 (41.9)
van Schalkwijk et al (2024) [61]	Longitudinal	Netherlands	Arrhythmia, CHD[i], HF	305	• Mean 64.07 (SD 9.98)	112 (36.7)
Astuti et al (2025) [62]	Cross-sectional	Indonesia	HF	119	• Mean 60.18 (SD 13.13)	58 (48.7)
Cuppen et al (2025) [63]	Cross-sectional	Netherlands	HF	61	• Mean 79.9 (SD 9.5)	24 (39.3)
Zhao et al (2025) [52]	Cross-sectional	China	CHD	396	• Mean 69.78 (SD 6.36)	200 (50.5)
Dibek et al (2025) [64]	Cross-sectional	Turkey	HF	250	• ≤60 years: n=76 (30.4%) • >60 years: n=174 (69.6%)	113 (45.2)
Sun et al (2025) [50]	Cross-sectional	China	CHD	594	• 45-59 years: n=203 (34.2%) • 60-74 years: n=264 (44.4%) • ≥75 years: n=127 (21.4%)	235 (42.6)

[a]CABG: coronary artery bypass grafting.

[b]HF: heart failure.

[c]IHD: ischemic heart disease.

[d]VS: valve surgery.

[e]CHF: congestive heart failure.

[f]PCI: percutaneous coronary intervention.

XSL•FO
RenderX

[g]AF: atrial fibrillation.

[h]CIED: cardiac implantable electronic device.

[i]CHD: coronary heart disease.

The most frequently used instrument for assessing DHL was the eHEALS [18] (n=11), followed by the DHLI [19] (n=2) and the eHLQ [20] (n=1). The Digital Health Readiness Questionnaire (DHRQ) [65] was used in one study (n=1). Because newer DHL instruments, such as the eHLQ and DHRQ, incorporate attitudinal components (eg, motivation and attitudes) into the construct of DHL, some studies additionally assessed digital-related attitudes, such as internet anxiety and digital confidence [56,60]. One study used a self-developed instrument [46] (Table 3; Multimedia Appendix 3 [36-48,51,53-56]). In studies with incomplete or unreported essential data [55,56,60], the corresponding authors were contacted to request additional information; however, no responses were received.

XSL•FO
RenderX

Table 3. Digital health literacy measurement tools and levels of the included studies.

Authors (year)	DHL[a] measurement	DHL level
Melholt et al (2018) [45]	eHEALS[b]	• Mean 28.80 (SD 5.96)[c] (baseline)
Chuang et al (2019) [53]	eHEALS	• Mean 26.2 (SD 5.7)
Lin et al (2020) [51]	eHEALS	• Mean 28.16 (SD 5.46)
Rodríguez Parrado et al (2022) [46]	Researcher-developed questionnaire	• Mean 2.33 (SD 1.25)[c] (baseline)
Spindler et al (2022) [47]	eHLQ[d]	• Intervention group: mean 2.98 (SD 3.22) • Control group: mean 2.71 (SD 3.17) (after 6 months)[e]
Yun et al (2022) [54]	Researcher-developed questionnaire	• 72% lower ICT[f] skills (n=128) • 28% middle or higher ICT skills (n=50)
Bakhshayesh et al (2023) [55]	eHEALS	• Mean 18.09 (SD 9.08)
Bäuerle et al (2023) [56]	Internet anxiety, digital confidence, prior experiences with mobile health interventions	• Not reported
Brørs et al (2023) [48]	eHEALS	• Mean 27.27 (SD 6.28) (baseline)
Ramstad et al (2023) [57]	eHEALS	• Mean 25.71 (SD 6.22) (baseline)
Rush et al (2023) [58]	30-item composite tool	• Computer self-efficacy: mean 5.38 (SD 1.27) • Health technology self-efficacy: mean 5.63 (SD 1.01) • Attitude toward health technology: mean 5.46 (SD 0.87)
Son et al (2023) [49]	eHEALS	• Intervention group: mean 24.32 (SD 9.55) • Control group: mean 25.58 (SD 9.79) (baseline)
Vitolo et al (2023) [59]	DHLI[g]	• Mean 48.58 (SD 24.16)
Mohajeri et al (2024) [60]	Internet anxiety, digital confidence, digital overload	• Internet anxiety: mean 1.58 (SD 0.80) • Digital confidence: mean 3.87 (SD 0.98) • Digital overload: mean 1.94 (SD 1.01)
van Schalkwijk et al (2024) [61]	DHLI	• Mean 2.97 (SD 0.60)
Astuti et al (2025) [62]	eHEALS	• Mean 26.35 (SD 7.40)
Cuppen et al (2025) [63]	DHRQ[h]	• Mean 38.4 (SD 17.7)
Zhao et al (2025) [52]	eHEALS	• Mean 18.18 (SD 10.11)
Dibek et al (2025) [64]	eHEALS	• Mean 21.9 (SD 9.7)
Sun et al (2025) [50]	eHEALS	• Median 21 (IQR 13-29)

[a]DHL: digital health literacy.
[b]eHEALS: eHealth Literacy Scale.
[c]Scores calculated by authors.
[d]eHLQ: eHealth Literacy Questionnaire.
[e]Baseline data were not reported in this study.
[f]ICT: information and communications technology.
[g]DHLI: Digital Health Literacy Instrument.
[h]DHRQ: Digital Health Readiness Questionnaire.

XSL•FO
RenderX

Quality Appraisal

The included studies comprised 11 cross-sectional, 2 longitudinal observational, 2 prospective observational, 2 randomized controlled trials, and 3 quasi-experimental studies. ROB and methodological quality were appraised using tools appropriate for each study design, and a summary of these assessments is presented in Table 4.

Table 4. Summary of quality assessment of the included studies[a].

Appraisal tool and study design	Overall quality assessment
AXIS[b]	
Cross-sectional	
Chuang et al (2019) [53]	18/20
Bakhshayesh et al (2023) [55]	17/20
Bäuerle et al (2023) [56]	17/20
Rush et al (2023) [58]	18/20
Vitolo et al (2023) [59]	16/20
Mohajeri et al (2024) [60]	18/20
Astuti et al (2025) [62]	17/20
Cuppen et al (2025) [63]	16/20
Zhao et al (2025) [52]	18/20
Dibek et al (2025) [64]	17/20
Sun et al (2025) [50]	17/20
NOS[c]	
Longitudinal/prospective	
Lin et al (2020) [51]	7/9
Brørs et al (2023) [48]	6/9
Ramstad et al (2023) [57]	5/9
van Schalkwijk et al (2024) [61]	5/9
RoB 2[d]	
Randomized controlled trial	
Spindler et al (2022) [47]	High risk
Yun et al (2022) [54]	Some concerns
ROBINS-I[e]	
Quasi-experimental	
Melholt et al (2018) [45]	Moderate
Rodríguez Parrado et al (2022) [46]	Moderate
Son et al (2023) [49]	Moderate

[a]Methodological quality was assessed using validated appraisal tools appropriate for each study design. Overall quality ratings or scores are reported as provided by each appraisal framework.

[b]AXIS: Appraisal Tool for Cross-Sectional Studies.

[c]NOS: Newcastle-Ottawa Quality Assessment Scale.

[d]RoB 2: Revised Cochrane Risk-of-Bias Tool for Randomized Trials.

[e]ROBINS-I: Risk of Bias in Nonrandomized Studies of Interventions.

The 11 cross-sectional studies were evaluated using the AXIS tool, with quality scores ranging from 16 to 18 out of a maximum of 20 points. However, most studies did not report information on nonresponders or provide response rates. In addition, characteristics of nonresponders or comparisons with responders were not reported (Multimedia Appendix 4 [50,52,53,55,56,58-60,62-64]). The 4 cohort studies (longitudinal, prospective observational studies) were assessed using the NOS, with scores ranging from 5 to 7 out of 9. According to the Agency for Healthcare Research and Quality (AHRQ) thresholds, 2 studies were rated as good quality and 2 as poor quality. The studies rated as poor quality relied on

XSL•FO
RenderX

self-reported outcomes and lacked clear reporting of follow-up rates or attrition (Multimedia Appendix 5 [48,51,57,61]).

The 2 randomized controlled trials were appraised using RoB 2 and were rated as having "some concerns" and "high ROB," respectively. In both studies, the randomization process and blinding procedures were not clearly reported, and insufficient information was available to support a low risk in these domains. Furthermore, neither study provided adequate details regarding missing outcome data (Multimedia Appendix 6 [47,54]). The 3 quasi-experimental studies were evaluated using ROBINS-I and were all judged to have a moderate ROB. Concerns primarily arose in the measurement of the outcomes domain, where outcome assessors may have been aware of participants' intervention status (Multimedia Appendix 7 [45,46,49]).

Factors Associated With DHL

Across the 20 included studies, factors associated with DHL were categorized into 5 domains: demographic (age, sex, education, and residence), clinical (cognitive function, laboratory findings, and disease severity), physical health (functional status and frailty), psychosocial (social support and caregiving needs), and digital technology use (prior experience) (Table 5; Multimedia Appendix 8 [45-49,51-63]). Lower educational attainment (n=6) and older age (n=5) were most frequently associated with lower DHL. Other factors included female sex (n=2), disease severity (n=2), lower or higher social support (n=2), and limited prior experience with digital technologies (n=2).

Table 5. Factors associated with digital health literacy of the included studies.

Category/subcategory	Significant factors[a]	References
Demographic variables		
Age	Older age	[49,50,54,55,64]
Sex	Female	[54,58]
Education	Lower educational attainment	[45,50,54-56,64]
Marital status	Married status	[64]
Economic status	Lower economic status	[64]
Employment status	Unemployment	[64]
Residence	Rural residence	[56]
Clinical variables		
Cognitive function	Lower cognitive function	[54]
Laboratory findings	Lower BMI, DBP[b], Hb[c], NT-proBNP[d]	[54]
Diagnosis duration	>6 years	[64]
Disease severity	Higher NYHA[e] (or HF stage) functional class	[54,64]
Prior cardiac intervention	Absence of PCI or CABG	[50]
Physical health		
Functional status	Greater functional dependence	[54]
Frailty	Higher frailty status	[59]
Smoking status	Smokers	[64]
Psychosocial variables		
Health behaviors	Lower health behaviors	[64]
Social support	Higher social support	[54]
Social support	Lower social support	[52]
Caregiving needs	Greater caregiver needs	[54]
Digital technology use		
Prior experience with digital technology	Low use of digital tools (computer, internet, national health portal)	[45,57]

[a]All listed factors were negatively associated with digital health literacy.

[b]DBP: diastolic blood pressure.

[c]Hb: hemoglobin.

[d]NT-proBNP: N-terminal pro-B-type natriuretic peptide.

[e]NYHA: New York Heart Association classification of heart failure functional status.

XSL•FO
RenderX

Association Between DHL and Health Outcomes

Health outcomes associated with DHL among adults with CVD were categorized as cognitive outcomes (disease knowledge and empowerment), health behaviors (smoking, self-care behaviors, physical activity, and medication adherence), clinical outcomes (cardiac events), health-related outcomes (depression, anxiety, and quality of life), and digital health-related outcomes (use of digital health technology and acceptance of digital health; Table 6).

Across the included studies, higher DHL was most frequently associated with digital health–related outcomes, particularly greater use and higher acceptance of digital health technologies (n=8). Other reported associations included improved self-care behaviors (n=3) and enhanced quality of life (n=2).

Most studies reported their results using regression analysis, while others used correlation analysis, logistic regression, or Cox regression for survival analyses. Effect sizes were reported in various forms depending on study design, including standardized beta coefficient (β), unstandardized coefficient (*B*), odds ratio, correlation coefficient (*r*), and hazard ratio (Multimedia Appendix 8).

XSL•FO
RenderX

Table 6. Association between digital health literacy and health outcomes.

Category, subcategory, and observed health outcomes	References	Certainty[a]
Cognitive outcomes		
Disease knowledge		
Greater HF[b] knowledge	[53]	⊕OOO
Empowerment		
Increased empowerment	[46]	⊕OOO
Health behaviors		
Smoking		
Lower likelihood of smoking	[48]	⊕OOO
Self-care behaviors		
Higher self-care behaviors	[53,55,62]	⊕OOO
Physical activity		
Higher physical activity	[48]	⊕OOO
Medication adherence		
Improved medication adherence	[51]	⊕OOO
Clinical outcomes		
Cardiac events		
Lower risk of cardiac events	[51]	⊕OOO
Symptom status		
Lower symptom burden	[64]	⊕OOO
Health-related outcomes		
Depression		
Reduced depressive symptoms	[48]	⊕OOO
Anxiety		
Reduced anxiety symptoms	[48]	⊕OOO
Quality of life		
Improved quality of life	[51,55]	⊕OOO
Digital health-related outcomes		
Use of digital health technology		
Greater use of the internet, mobile apps, and electronic health applications	[47,57,61]	⊕⊕OO[c] ⊕OOO[d]
Higher technology self-efficacy and motivation; lower technophobia	[52]	⊕OOO
Acceptance of digital health		
Increased acceptance of mobile health; greater willingness to participate in telemonitoring	[56,60,63]	⊕OOO
Higher telehealth satisfaction and more positive attitudes toward health care technology; lower telehealth satisfaction with higher computer self-efficacy	[58]	⊕OOO

[a]The certainty of evidence for each outcome was evaluated using the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) approach: ⊕⊕⊕⊕ indicates high certainty, ⊕⊕⊕O moderate certainty, ⊕⊕OO low certainty, and ⊕OOO very low certainty.

[b]HF: heart failure.

[c]Certainty of evidence based on randomized controlled trial evidence.

[d]Certainty of evidence based on observational studies.

XSL•FO
RenderX

Meta-Analysis Results

Nine studies reported mean eHEALS scores, which were pooled using a random-effects model. The overall pooled mean eHEALS score was 24.26 (95% CI 21.19-27.32; $P<.001$; $I^2=98.4\%$; $\tau^2=15.55$; $\tau=3.94$). The 95% prediction interval ranged from 14.66 to 33.85, indicating substantial variability in true eHEALS scores across populations and settings. Subgroup analyses were conducted by region (n=9), disease type (n=9), and mean age (n=7). Studies conducted in Europe reported slightly higher eHEALS scores than those conducted in Asia. Studies with a mean age <65 years exhibited higher scores than those with a mean age ≥65 years. Studies that included only patients with heart failure demonstrated eHEALS scores nearly identical to those reported in studies of mixed CVD populations. No statistically significant subgroup differences were observed in analyses stratified by region ($P=.50$), disease type ($P=.98$), or age ($P=.38$; Figure 2). Leave-one-out sensitivity analysis indicated that the pooled eHEALS score ranged from 23.70 to 25.02, with a maximum variation of 1.32 points across iterations.

XSL•FO
RenderX

Figure 2. Forest plot of pooled mean eHEALS scores across included studies: (A) overall pooled mean eHEALS score, (B) subgroup meta-analysis by study region (Asia vs Europe), (C) subgroup meta-analysis by disease type (heart failure only vs mixed CVD conditions, including percutaneous coronary intervention or coronary heart disease), and (D) subgroup meta-analysis by mean age group (<65 years vs ≥65 years). This figure summarizes results from 9 studies published between 2018 and 2025, including 5 cross-sectional studies, 2 quasi-experimental studies, 1 longitudinal study, and 1 prospective cohort study. Study populations consisted of adult patients with heart failure or other cardiovascular diseases, recruited across Asia (Taiwan, Iran, South Korea, Indonesia, and China) and Europe (Denmark, Norway, and Turkey). Sample sizes ranged from 49 to 1970 participants per study, with a total of 3693 participants included in the overall meta-analysis [45-47,51,54,56,60,62,63]. eHEALS: eHealth Literacy Scale; HKSJ: Hartung-Knapp-Sidik-Jonkman method; MRAW: mean raw; PCI/CHD: percutaneous coronary intervention/coronary heart disease.

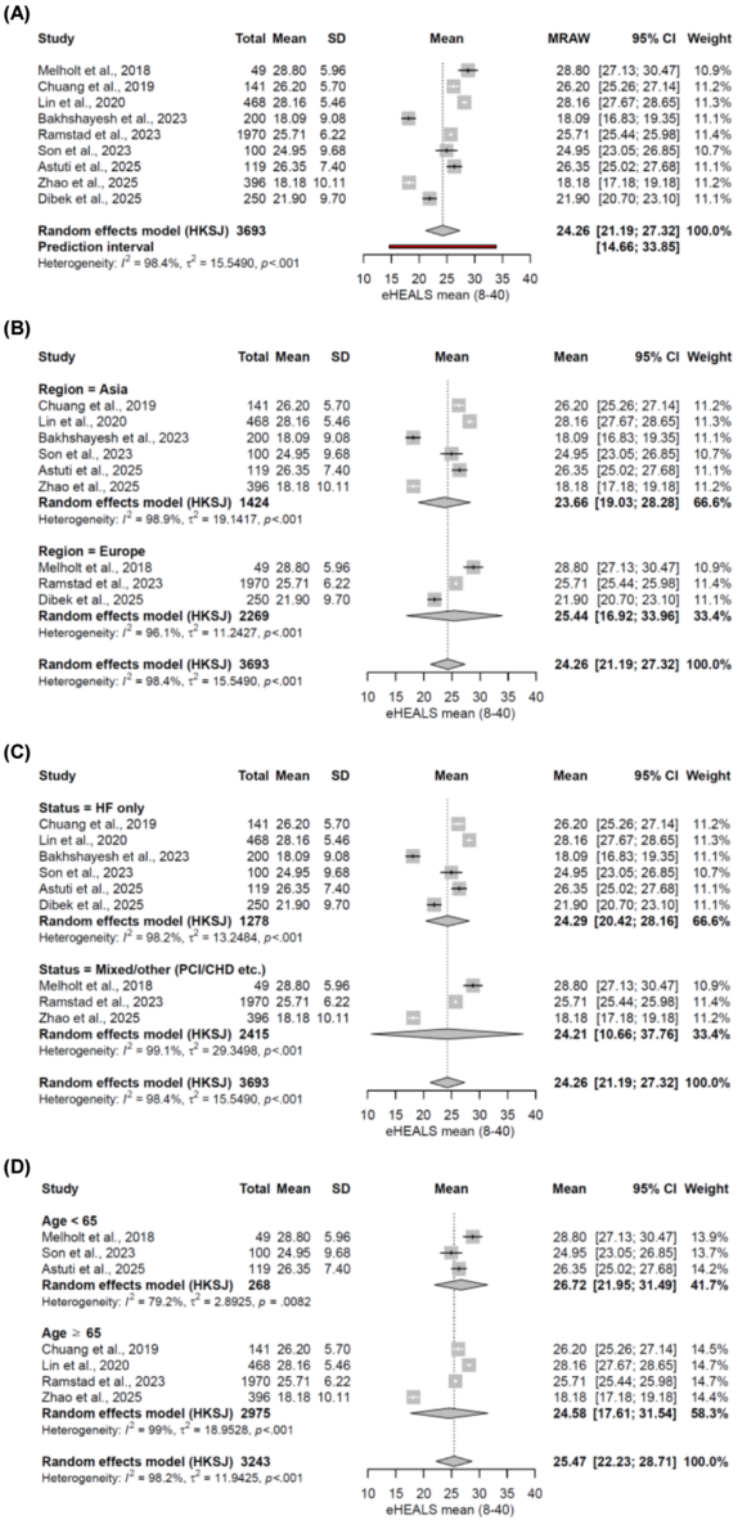

XSL•FO
RenderX

Assessment of Reporting Biases

Assessment of reporting biases was not conducted because fewer than 10 studies were included in the meta-analysis, as described in the Methods section.

Certainty of Evidence

The certainty of evidence for each outcome, as assessed using the GRADE framework, was rated as low or very low (Table 6). The evidence was initially low, reflecting its primary reliance on observational studies. Certainty was further downgraded for imprecision, attributable to the limited number of studies and wide CIs. Detailed ratings and the rationale for downgrades are provided in Multimedia Appendix 9.

Discussion

Principal Findings

This review synthesized evidence on DHL levels, factors associated with DHL, and related health outcomes among adults with CVD, a population for whom sustained self-management is critical. Overall, the findings indicate that DHL levels among adults with CVD are relatively low and characterized by substantial between-study heterogeneity. The wide prediction interval further reflects this heterogeneity, indicating that DHL levels may vary considerably across clinical contexts even when the pooled mean estimate is similar [36]. Sociodemographic characteristics, particularly older age, lower educational attainment, and female sex, were commonly associated with lower DHL. Higher DHL was associated with more favorable health outcomes, including improved self-care behaviors, better quality of life, and greater acceptance of digital health technologies. Collectively, these findings provide an empirical overview of the role of DHL in supporting self-management in patients with CVD.

DHL Levels Across Studies

DHL is increasingly recognized as a key determinant of both public and individual health [66]. Using the commonly applied cutoff of 26 on the eHEALS to indicate limited DHL [67], the pooled eHEALS score identified in this review was below this threshold, suggesting generally low DHL levels among adults with CVD. Compared with findings from a previous systematic review of individuals with chronic diseases, including diabetes mellitus and hypertension [37], DHL levels observed in the present review were comparatively lower. According to the American Heart Association, the prevalence of CVD increases substantially with advancing age, with the highest burden observed among older adults [68]. Accordingly, the relatively low DHL levels identified in this review should be interpreted in the context of the older age distribution of CVD populations. This finding is consistent with a previous meta-analysis by Kanejima et al [26], which reported a high prevalence of limited health literacy among patients with CVD. Effective management of CVD requires adherence to medication regimens, dietary modifications, and timely symptom management and response [12]. These tasks increasingly depend on accessing, interpreting, and applying health information delivered through digital platforms [69]. The cognitive and technological demands associated with CVD management may contribute to challenges in engaging with digital health information and may partially explain the lower observed DHL levels observed in this population [70].

Subgroup analyses were limited to variables consistently reported across studies, including region, disease type, and mean age. No statistically significant subgroup differences were identified, which may reflect limited statistical power due to the small number of included studies [71]. Nevertheless, directional trends warrant consideration. Studies conducted in Europe tended to report higher DHL levels than those conducted in Asia, potentially reflecting differences in digital health infrastructure, access to health technologies, and system-level support [72]. In addition, higher DHL levels observed among participants younger than 65 years suggest that age-related disparities in DHL persist within CVD. The CIs around the pooled estimate reflect uncertainty in the average DHL level across studies and should be interpreted with caution, given the substantial between-study heterogeneity. Moreover, the wide prediction interval indicates considerable variability across settings [36]. This variability suggests that DHL proficiency is not uniformly low among adults with CVD but is likely influenced by contextual and system-level factors [73].

Determinants and Health Outcomes of DHL

The factors identified in this review suggest that DHL among adults with CVD is shaped by both individual characteristics and broader social and digital contexts. Notably, the included studies predominantly focused on individual-level digital experiences, competencies, and perceptions, rather than structural or system-level digital environments. In addition, older age and lower educational attainment were consistently associated with lower DHL in adults with CVD. According to the 2026 American Heart Association Heart Disease and Stroke Statistics, CVD and heart failure disproportionately affect older adults [68]. Furthermore, large-scale cohort studies consistently demonstrate that lower educational attainment is associated with an increased risk of CVD [74]. Collectively, these sociodemographic characteristics provide critical context for interpreting DHL levels in patients with CVD. In this review, additional factors frequently associated with lower DHL included female sex, rural residence, poorer self-rated health, lower social support, and limited prior experience with digital technologies. These findings align with a recent meta-analysis of 17 studies in mixed populations, which identified age, ethnicity, income, employment status, education, and health status as key correlates of eHealth literacy [75].

However, rather than framing these demographic characteristics solely as individual deficits, they should be interpreted through the lens of Digital Determinants of Health [76]. According to the American Heart Association, Digital Determinants of Health—which encompass DHL, digital infrastructure, access, and digital inequity—play a pivotal role in shaping cardiovascular outcomes [77]. Most studies included in this review treated DHL as an individual-level attribute, emphasizing personal skills and behaviors. This approach reflects a broader tendency in the literature to frame DHL as an individual responsibility, with limited consideration of health system structures, provider practices, or service-level contexts that may

XSL•FO
RenderX

influence digital engagement [76]. From this perspective, disparities in DHL may not be fully explained by individual-level characteristics alone, highlighting the potential relevance of structural and provider-level factors [76,77].

Health outcomes associated with DHL clustered into 5 domains: cognitive, health behavior, clinical, health-related, and digital health–related outcomes. Across these domains, higher DHL was most consistently associated with improvements in cognitive understanding and health behaviors, including greater disease knowledge, increased empowerment, and more effective self-care practices such as medication adherence and physical activity. These findings suggest that DHL may facilitate self-management by enhancing information processing and supporting the translation of health information into everyday behaviors [14]. Digital health–related outcomes were most consistently associated with DHL in this review. Higher DHL was associated with greater acceptance and use of digital health technologies, stronger technology self-efficacy, and lower technophobia. These digital indicators appear closely aligned with the functional attributes of DHL, conceptualized as the ability to seek, appraise, and apply health information in digital contexts [14,27]. Therefore, DHL functions as a foundational competency that underpins a range of outcomes related to the use of digital health tools, supporting sustained self-management [22].

In this review, evidence linking DHL to clinical outcomes was comparatively limited. While some included studies reported associations between higher DHL and lower risks of cardiac events [51] or improved symptom status [64], these findings were based on a small number of studies and heterogeneous outcome measures. Further longitudinal research is required to elucidate the pathways linking DHL to hard clinical outcomes, including studies that integrate patient-reported perspectives on providers and system-level digital support.

The overall certainty of these findings is limited by study design characteristics and ROB across the included studies. Most synthesized studies were observational and therefore inherently susceptible to confounding. In addition, quality assessment revealed methodological limitations, particularly incomplete reporting of participant flow, nonresponders, and attrition. Given that DHL is influenced by sociodemographic factors such as age, educational attainment, and socioeconomic status [20], insufficient information on sample selection and follow-up may introduce selection bias and affect the interpretation of DHL estimates.

As digital health environments continue to evolve rapidly, transparent reporting of participant characteristics and study procedures is increasingly critical for contextualizing findings [78]. Improving reporting of sample representativeness and participant flow would enhance the interpretability of future studies and contribute to a more robust evidence base on DHL among adults with CVD. Considering the observed heterogeneity, ROB, and limitations identified in the GRADE framework, the overall certainty of evidence regarding associations between DHL and health outcomes was rated as low. Consequently, these findings should be interpreted as indicative of broad trends rather than a precise estimate applicable to all individual patients.

Implications for Clinical Practice and Health Policy

These findings have key implications for clinical practice. DHL and digital access should be considered central determinants of health equity that influence patients' capacity for effective self-management, rather than merely reflecting individual-level limitations [73]. The emphasis on individual-level factors in the existing literature underscores the need to also consider the health care system and provider-level influences on patients' digital engagement [76]. Accordingly, health care systems and providers may play a critical role in supporting equitable self-care by addressing digital barriers [77].

In clinical settings, assessing patients' digital access and competencies as part of routine assessment—alongside other social and contextual factors—may help identify individuals at risk of digital exclusion [73]. These assessments should not be treated solely as screening tools, but as a foundation for adapting service delivery and provider practices to accommodate varying levels of digital readiness [76]. For patients with low DHL or limited access, the mere provision of digital technologies may be insufficient. System-level strategies—such as supported-use models integrated into routine care and service designs that enhance accessibility—may be required, alongside patient education programs or involvement of family caregivers where appropriate [69,73,77]. Adopting a digital inclusion–oriented approach may help ensure that digital health technologies act as facilitators rather than barriers to care, thereby supporting more effective and sustained self-management among adults with CVD [69,77].

Strengths and Limitations

This review has several strengths. First, to our knowledge, it is the first systematic review to comprehensively synthesize health outcomes associated with DHL in patients with CVD using a multidimensional framework encompassing 5 distinct domains. Second, by incorporating findings from multidimensional instruments such as the DHLI [19], eHLQ [20], and DHRQ [65], this study aligns with the evolving conceptual definition of DHL [14]. This approach extends beyond traditional eHealth literacy and captures both functional skills and attitudinal components (eg, motivation and digital confidence) required in contemporary digital health environments.

Several limitations warrant consideration. First, most included studies were cross-sectional and showed substantial variability in analytic methods and confounder adjustment, limiting comparability. Second, evidence was limited for psychosocial determinants as well as objective clinical endpoints and biological markers. Many biomarkers and clinical events were evaluated in only a small number of studies, often from single-center samples with modest sizes, which constrains generalizability. Third, reporting of nonresponders and participant flow was often incomplete. This limited transparency may compromise assessments of sample representativeness and increase the risk of selection bias, particularly given known sociodemographic gradients in DHL [21]. Finally, substantial between-study heterogeneity was observed across study

XSL•FO
RenderX

populations, settings, measurement instruments, and analytic approaches. Accordingly, pooled estimates and prediction intervals should be interpreted with caution.

Conclusions

To our knowledge, this review is one of the first to integrate evidence on DHL levels, associated factors, and DHL-related health outcomes among adults with CVD. Overall, DHL varied widely and showed consistent gradients by age and socioeconomic disadvantage, with lower DHL generally associated with poorer engagement with digital health resources and less favorable self-management–related outcomes. Although the pooled mean eHEALS score indicates modest DHL on average, the wide prediction interval indicates considerable variability across populations and clinical settings. By synthesizing DHL levels, determinants, and health outcomes in CVD populations, this review offers an integrated evidence base that extends beyond single-instrument assessments (eg, eHEALS) and supports interpretation of DHL in relation to structural and health system contexts.

Our findings highlight the need for routine assessment of DHL and targeted support for patients with limited DHL, as well as the importance of designing and delivering digital tools that are accessible and easy to use in routine CVD care. Future studies using longitudinal and interventional designs are needed to clarify causal pathways linking DHL to self-management and health outcomes, and to examine how provider- and health system–level factors shape and support DHL in routine care.

Acknowledgments

We acknowledge the contributions of individuals who supported this work but did not meet the criteria for authorship. Generative AI tools were not used in the preparation of this manuscript.

Funding

This work was supported by a grant from the National Research Foundation of Korea (NRF), funded by the Korean government (MSIT) (grant RS-2025-00562535). The funder had no role in the study design, data collection, analysis, interpretation, or manuscript preparation.

Data Availability

The data underlying the findings of this study are available from the corresponding authors upon reasonable request.

Authors' Contributions

Conceptualization: EJK, YJS
Data curation: EJK, DYK
Formal analysis: EJK, DYK, HJK
Methodology: YJS
Project administration: YJS
Software: EJK, DYK, HJK
Supervision: YJS
Validation: EJK, DYK, HJK, YJS
Visualization: EJK, DYK, HJK
Writing – original draft: EJK, DYK, HJK, YJS
Writing – review & editing: EJK, DYK, YJS

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA 2020 reporting checklists (PRISMA 2020, PRISMA Abstract, and PRISMA-S).
[DOCX File , 44 KB - jmir_v28i1e89102_app1.docx]

Multimedia Appendix 2
Literature search strategy.
[DOCX File , 30 KB - jmir_v28i1e89102_app2.docx]

Multimedia Appendix 3
Summary of digital health literacy measurement tools included in this review.
[DOCX File , 33 KB - jmir_v28i1e89102_app3.docx]

XSL•FO
RenderX

Multimedia Appendix 4
Quality assessment for cross-sectional studies using the AXIS (Appraisal Tool for Cross-Sectional Studies) tool.
[DOCX File , 30 KB - jmir_v28i1e89102_app4.docx]

Multimedia Appendix 5
Newcastle–Ottawa Quality Assessment of longitudinal observational and prospective observational studies.
[DOCX File , 25 KB - jmir_v28i1e89102_app5.docx]

Multimedia Appendix 6
Risk-of-bias tool for randomized controlled trials.
[DOCX File , 26 KB - jmir_v28i1e89102_app6.docx]

Multimedia Appendix 7
Risk of bias in nonrandomized studies of interventions.
[DOCX File , 28 KB - jmir_v28i1e89102_app7.docx]

Multimedia Appendix 8
Study-level digital health literacy levels, associated factors, and health outcomes.
[DOCX File , 34 KB - jmir_v28i1e89102_app8.docx]

Multimedia Appendix 9
GRADE (Grading of Recommendations Assessment, Development, and Evaluation) assessment.
[DOCX File , 27 KB - jmir_v28i1e89102_app9.docx]

References

1. What is cardiovascular disease? American Heart Association. 2024 Jan 10. URL: https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease [accessed 2026-03-09]
2. Cruz-Ramos NA, Alor-Hernández G, Colombo-Mendoza LO, Sánchez-Cervantes JL, Rodríguez-Mazahua L, Guarneros-Nolasco LR. mHealth apps for self-management of cardiovascular diseases: a scoping review. Healthcare (Basel) 2022;10(2):322 [FREE Full text] [doi: 10.3390/healthcare10020322] [Medline: 35206936]
3. Huang Y, Li S, Lu X, Chen W, Zhang Y. The effect of self-management on patients with chronic diseases: a systematic review and meta-analysis. Healthcare (Basel) 2024;12(21):2151 [FREE Full text] [doi: 10.3390/healthcare12212151] [Medline: 39517362]
4. Yuen E, Winter N, Savira F, Huggins CE, Nguyen L, Cooper P, et al. Digital health literacy and its association with sociodemographic characteristics, health resource use, and health outcomes: rapid review. Interact J Med Res 2024;13:e46888 [FREE Full text] [doi: 10.2196/46888] [Medline: 39059006]
5. Graham-Brown MPM, Smith AC, Greenwood SA. Digital health interventions in chronic kidney disease: levelling the playing field? Clin Kidney J 2023;16(5):763-767 [FREE Full text] [doi: 10.1093/ckj/sfac259] [Medline: 37151422]
6. Endalamaw A, Zewdie A, Wolka E, Assefa Y. A scoping review of digital health technologies in multimorbidity management: mechanisms, outcomes, challenges, and strategies. BMC Health Serv Res 2025;25(1):382 [FREE Full text] [doi: 10.1186/s12913-025-12548-5] [Medline: 40089752]
7. Hu S, Song D, Wan S, Zhang S, Luo C, Li N, et al. Digital health: current applications, challenges, and future directions for enhancing healthcare quality and safety. Front Public Health 2025;13:1646802. [doi: 10.3389/fpubh.2025.1646802] [Medline: 41089861]
8. Livieri G, Mangina E, Protopapadakis ED, Panayiotou AG. The gaps and challenges in digital health technology use as perceived by patients: a scoping review and narrative meta-synthesis. Front Digit Health 2025;7:1474956. [doi: 10.3389/fdgth.2025.1474956] [Medline: 40212901]
9. Diamond JE, Kaltenbach LA, Granger BB, Fonarow GC, Al-Khalidi HR, Albert NM, et al. Access to mobile health interventions among patients hospitalized with heart failure: insights into the digital divide from the CONNECT-HF mHealth substudy. Circ Heart Fail 2024;17(2):e011140. [doi: 10.1161/CIRCHEARTFAILURE.123.011140] [Medline: 38205653]
10. Kim DS, Eltahir AA, Ngo S, Rodriguez F. Bridging the gap: how accounting for social determinants of health can improve digital health equity in cardiovascular medicine. Curr Atheroscler Rep 2024;27(1):9. [doi: 10.1007/s11883-024-01249-9] [Medline: 39576395]
11. Wamala Andersson S, Gonzalez MP. Digital health literacy – a key factor in realizing the value of digital transformation in healthcare. Front Digit Health 2025;7:1461342. [doi: 10.3389/fdgth.2025.1461342] [Medline: 40538571]

XSL•FO
RenderX

12. Magnani JW, Mujahid MS, Aronow HD, Cené CW, Dickson VV, Havranek E, et al. Health literacy and cardiovascular disease: fundamental relevance to primary and secondary prevention: a scientific statement from the American Heart Association. Circulation 2018;138(2):e48-e74 [FREE Full text] [doi: 10.1161/CIR.0000000000000579] [Medline: 29866648]
13. Ndayishimiye C, Lopes H, Middleton J. A systematic scoping review of digital health technologies during COVID-19: a new normal in primary health care delivery. Health Technol (Berl) 2023;13(2):273-284 [FREE Full text] [doi: 10.1007/s12553-023-00725-7] [Medline: 36628261]
14. Ban S, Kim Y, Seomun G. Digital health literacy: a concept analysis. Digit Health 2024;10:20552076241287894 [FREE Full text] [doi: 10.1177/20552076241287894] [Medline: 39381807]
15. Griebel L, Enwald H, Gilstad H, Pohl A, Moreland J, Sedlmayr M. eHealth literacy research—quo vadis? Inform Health Soc Care 2018;43(4):427-442. [doi: 10.1080/17538157.2017.1364247] [Medline: 29045164]
16. Shao YJ, Duan XC, Xu XJ, Guo HY, Zhang ZY, Zhao S, et al. Latent profile and determinants of self-management behaviors among older adult patients with chronic diseases: a cross-sectional study. Front Public Health 2025;13:1506545. [doi: 10.3389/fpubh.2025.1506545] [Medline: 39975786]
17. Ullrich G, Bäuerle A, Vogt H, Mahabadi AA, Paldán K, Messiha D, et al. Digital health literacy and attitudes toward eHealth technologies among patients with cardiovascular disease and their implications for secondary prevention: survey study. JMIR Form Res 2025;9:e63057 [FREE Full text] [doi: 10.2196/63057] [Medline: 40106277]
18. Norman CD, Skinner HA. eHEALS: the eHealth literacy scale. J Med Internet Res 2006;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
19. van der Vaart R, Drossaert C. Development of the digital health literacy instrument: measuring a broad spectrum of Health 1.0 and Health 2.0 skills. J Med Internet Res 2017;19(1):e27 [FREE Full text] [doi: 10.2196/jmir.6709] [Medline: 28119275]
20. Kayser L, Karnoe A, Furstrand D, Batterham R, Christensen KB, Elsworth G, et al. A multidimensional tool based on the ehealth literacy framework: development and initial validity testing of the eHealth literacy questionnaire (eHLQ). J Med Internet Res 2018;20(2):e36 [FREE Full text] [doi: 10.2196/jmir.8371] [Medline: 29434011]
21. Estrela M, Semedo G, Roque F, Ferreira PL, Herdeiro MT. Sociodemographic determinants of digital health literacy: a systematic review and meta-analysis. Int J Med Inform 2023;177:105124 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105124] [Medline: 37329766]
22. Qi Y, Mohamad E, Azlan AA, Zhang C, Ma Y, Wu A. Digital health solutions for cardiovascular disease prevention: systematic review. J Med Internet Res 2025;27:e64981 [FREE Full text] [doi: 10.2196/64981] [Medline: 39847411]
23. Verweel L, Newman A, Michaelchuk W, Packham T, Goldstein R, Brooks D. The effect of digital interventions on related health literacy and skills for individuals living with chronic diseases: a systematic review and meta-analysis. Int J Med Inform 2023;177:105114. [doi: 10.1016/j.ijmedinf.2023.105114] [Medline: 37329765]
24. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
25. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
26. Kanejima Y, Shimogai T, Kitamura M, Ishihara K, Izawa KP. Impact of health literacy in patients with cardiovascular diseases: a systematic review and meta-analysis. Patient Educ Couns 2022;105(7):1793-1800. [doi: 10.1016/j.pec.2021.11.021] [Medline: 34862114]
27. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
28. Downes MJ, Brennan ML, Williams HC, Dean RS. Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). BMJ Open 2016;6(12):e011458 [FREE Full text] [doi: 10.1136/bmjopen-2016-011458] [Medline: 27932337]
29. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. The Ottawa Hospital Research Institute. URL: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp [accessed 2026-03-09]
30. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
31. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919 [FREE Full text] [doi: 10.1136/bmj.i4919] [Medline: 27733354]
32. Barili F, Parolari A, Kappetein PA, Freemantle N. Statistical primer: heterogeneity, random- or fixed-effects model analyses? Interact Cardiovasc Thorac Surg 2018;27(3):317-321 [FREE Full text] [doi: 10.1093/icvts/ivy163] [Medline: 29868857]
33. Shim SR, Kim S. Intervention meta-analysis: application and practice using R software. Epidemiol Health 2019;41:e2019008 [FREE Full text] [doi: 10.4178/epih.e2019008] [Medline: 30999738]
34. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014;14:25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]

XSL•FO
RenderX

35. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327(7414):557-560 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
36. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023;12(4):101014 [FREE Full text] [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
37. Zaghloul H, Fanous K, Ahmed L, Arabi M, Varghese S, Omar S, et al. Digital health literacy in patients with common chronic diseases: systematic review and meta-analysis. J Med Internet Res 2025;27:e56231 [FREE Full text] [doi: 10.2196/56231] [Medline: 40854231]
38. Liu X, Valdez D, Parker MA, Mai A, Walsh-Buhi ER. Quality of cancer-related information on new media (2014-2023): systematic review and meta-analysis. J Med Internet Res 2025;27:e73185 [FREE Full text] [doi: 10.2196/73185] [Medline: 41061257]
39. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2024. URL: https://www.R-project.org/ [accessed 2026-03-09]
40. Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. Evid Based Ment Health 2019;22(4):153-160 [FREE Full text] [doi: 10.1136/ebmental-2019-300117] [Medline: 31563865]
41. Viechtbauer W. Conducting meta-analyses in R with the meta for package. J Stat Softw 2010;36(3):1-48. [doi: 10.18637/jss.v036.i03]
42. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ. Cochrane Handbook for Systematic Reviews of Interventions, Version 6.5. The Cochrane Collaboration. 2024 Aug. URL: https://www.cochrane.org/handbook [accessed 2026-03-09]
43. Schünemann H, Brożek J, Guyatt G, Oxman A. GRADE Handbook for Grading the Quality of Evidence and the Strength of Recommendations Using the GRADE Approach. Cochrane. 2013 Oct. URL: https://gdt.gradepro.org/app/handbook/handbook.html [accessed 2026-03-09]
44. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336(7650):924-926 [FREE Full text] [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
45. Melholt C, Joensson K, Spindler H, Hansen J, Andreasen JJ, Nielsen G, et al. Cardiac patients' experiences with a telerehabilitation web portal: implications for eHealth literacy. Patient Educ Couns 2018;101(5):854-861. [doi: 10.1016/j.pec.2017.12.017] [Medline: 29305064]
46. Rodríguez Parrado IY, Achury Saldaña DM. Digital health literacy in patients with heart failure in times of pandemic. Comput Inform Nurs 2022;40(11):754-762 [FREE Full text] [doi: 10.1097/CIN.0000000000000883] [Medline: 35234702]
47. Spindler H, Dyrvig AK, Schacksen CS, Anthonimuthu D, Frost L, Gade JD, et al. Increased motivation for and use of digital services in heart failure patients participating in a telerehabilitation program: a randomized controlled trial. Mhealth 2022;8:25 [FREE Full text] [doi: 10.21037/mhealth-21-56] [Medline: 35928510]
48. Brørs G, Dalen H, Allore H, Deaton C, Fridlund B, Norman CD, et al. The association of electronic health literacy with behavioural and psychological coronary artery disease risk factors in patients after percutaneous coronary intervention: a 12-month follow-up study. Eur Heart J Digit Health 2023;4(2):125-135 [FREE Full text] [doi: 10.1093/ehjdh/ztad010] [Medline: 36974270]
49. Son Y, Kim S, Lee W, Won H, Cho JH, Hong J, et al. The effects of a 24-week interactive text message-based mobile health intervention for enhancing self-care behaviours of patients with heart failure: a quasi-experimental study. Nurs Open 2023;10(9):6309-6319 [FREE Full text] [doi: 10.1002/nop2.1878] [Medline: 37313589]
50. Sun S, Guo R, Wang Y, Jiao L, Zhao C, Qiao S, et al. Correlation analysis of electronic health literacy, compliance behavior, and quality of life in middle-aged and older patients with coronary heart disease: a cross-sectional study. Front Public Health 2025;13:1680950. [doi: 10.3389/fpubh.2025.1680950] [Medline: 41446526]
51. Lin C, Ganji M, Griffiths MD, Bravell ME, Broström A, Pakpour AH. Mediated effects of insomnia, psychological distress and medication adherence in the association of eHealth literacy and cardiac events among Iranian older patients with heart failure: a longitudinal study. Eur J Cardiovasc Nurs 2020;19(2):155-164. [doi: 10.1177/1474515119873648] [Medline: 31516036]
52. Zhao J, Hu D, Du H, Wang H, Tu X, Wang A. Social support and technophobia in older patients with coronary heart disease: the mediating roles of eHealth literacy and healthcare technology self-efficacy. PLOS One 2025;20(6):e0325202 [FREE Full text] [doi: 10.1371/journal.pone.0325202] [Medline: 40465587]
53. Chuang HW, Kao CW, Lin WS, Chang YC. Factors affecting self-care maintenance and management in patients with heart failure: testing a path model. J Cardiovasc Nurs 2019;34(4):297-305. [doi: 10.1097/JCN.0000000000000575] [Medline: 31033857]
54. Yun S, Enjuanes C, Calero-Molina E, Hidalgo E, José-Bazán N, Ruiz M, et al. Usefulness of telemedicine-based heart failure monitoring according to 'eHealth literacy' domains: insights from the iCOR randomized controlled trial. Eur J Intern Med 2022;101:56-67. [doi: 10.1016/j.ejim.2022.04.008] [Medline: 35483994]

XSL•FO
RenderX

55. Bakhshayesh M, Hojatansari M, Rabie Siahkali S. Investigating the relationship of electronic health literacy and self-care behaviors with life quality in patients with heart failure. Prevent Care Nurs Midwifery J 2023;13(4):30-40. [doi: 10.61186/pcnm.13.4.30]
56. Bäuerle A, Mallien C, Rassaf T, Jahre L, Rammos C, Skoda E, et al. Determining the acceptance of digital cardiac rehabilitation and its influencing factors among patients affected by cardiac diseases. J Cardiovasc Dev Dis 2023;10(4):174 [FREE Full text] [doi: 10.3390/jcdd10040174] [Medline: 37103053]
57. Ramstad KJ, Brørs G, Pettersen TR, Deaton C, Palm P, Rotevatn S, et al. eHealth technology use and eHealth literacy after percutaneous coronary intervention. Eur J Cardiovasc Nurs 2023;22(5):472-481. [doi: 10.1093/eurjcn/zvac087] [Medline: 36190843]
58. Rush KL, Burton L, Seaton CL, Loewen P, O'Connor BP, Moroz L, et al. Telehealth satisfaction in patients receiving virtual atrial fibrillation care: quantitative exploratory study. JMIR Hum Factors 2023 Sep 14;10:e50232. [doi: 10.2196/50232] [Medline: 37707881]
59. Vitolo M, Ziveri V, Gozzi G, Busi C, Imberti JF, Bonini N, et al. DIGItal health literacy after COVID-19 outbreak among frail and non-frail cardiology patients: the DIGI-COVID study. J Pers Med 2022;13(1):99 [FREE Full text] [doi: 10.3390/jpm13010099] [Medline: 36675760]
60. Mohajeri D, Jahre LM, Bäuerle A, Schieffers T, Messiha D, Rammos C, et al. Acceptance, needs, and demands for nutritional mhealth support in patients with cardiovascular disease. Nutrients 2024;16(23):4155 [FREE Full text] [doi: 10.3390/nu16234155] [Medline: 39683549]
61. van Schalkwijk D, Lodder P, Everaert J, Widdershoven J, Habibović M. Latent profiles of telehealth care satisfaction during the COVID-19 pandemic among patients with cardiac conditions in an outpatient setting. Cardiovasc Digit Health J 2024;5(2):85-95 [FREE Full text] [doi: 10.1016/j.cvdhj.2023.11.022] [Medline: 38765625]
62. Astuti RP, Nurdina G, Herdiman H. Factors influencing self-care among patients with heart failure in Indonesia. Med Surg J Nurs Res 2025;4(1):11-22. [doi: 10.70331/jpkmb.v4i1.47]
63. Cuppen S, van Leunen M, Henken T, Goevaerts M, Scherrenberg M, Falter M, et al. Association of technology-related skills and self-efficacy with willingness to participate in heart failure telemonitoring: cross-sectional observational study. JMIR Form Res 2025;9:e68992 [FREE Full text] [doi: 10.2196/68992] [Medline: 40505140]
64. Dibek EN, Eren MG, Sert H, Kocayigit I. Relationship of health behaviors and symptom status mediated by digital health literacy in persons with heart failure. Nurs Res 2025;74(5):349-356. [doi: 10.1097/NNR.0000000000000828] [Medline: 40307233]
65. Scherrenberg M, Falter M, Kaihara T, Xu L, van Leunen M, Kemps H, et al. Development and internal validation of the digital health readiness questionnaire: prospective single-center survey study. J Med Internet Res 2023;25:e41615 [FREE Full text] [doi: 10.2196/41615] [Medline: 36897627]
66. Barbati C, Maranesi E, Giammarchi C, Lenge M, Bonciani M, Barbi E, et al. Effectiveness of eHealth literacy interventions: a systematic review and meta-analysis of experimental studies. BMC Public Health 2025;25(1):288 [FREE Full text] [doi: 10.1186/s12889-025-21354-x] [Medline: 39849354]
67. Xie L, Zhang S, Xin M, Zhu M, Lu W, Mo PK. Electronic health literacy and health-related outcomes among older adults: a systematic review. Prev Med 2022;157:106997. [doi: 10.1016/j.ypmed.2022.106997] [Medline: 35189203]
68. Palaniappan LP, Allen NB, Almarzooq ZI, Anderson CAM, Arora P, Avery CL, et al. 2026 heart disease and stroke statistics: a report of US and global data from the American Heart Association. Circulation 2026;153(9):e275-e906 [FREE Full text] [doi: 10.1161/CIR.0000000000001412] [Medline: 41562125]
69. Avoke D, Elshafeey A, Weinstein R, Kim CH, Martin SS. Digital health in diabetes and cardiovascular disease. Endocr Res 2024;49(3):124-136. [doi: 10.1080/07435800.2024.2341146] [Medline: 38605594]
70. Shao Y, Hou X, Peng Q, Hu M, Li C. A meta-synthesis of qualitative studies on cardiovascular disease patients' experiences using digital health tools. Front Public Health 2025;13:1709562. [doi: 10.3389/fpubh.2025.1709562] [Medline: 41346736]
71. Zeymer U, Freund A, Hochadel M, Ostadal P, Belohlavek J, Rokyta R, et al. Venoarterial extracorporeal membrane oxygenation in patients with infarct-related cardiogenic shock: an individual patient data meta-analysis of randomised trials. Lancet 2023;402(10410):1338-1346. [doi: 10.1016/S0140-6736(23)01607-0] [Medline: 37643628]
72. Okpechi IG, Muneer S, Ye F, Zaidi D, Ghimire A, Tinwala MM, et al. Global eHealth capacity: secondary analysis of WHO data on eHealth and implications for kidney care delivery in low-resource settings. BMJ Open 2022;12(3):e055658 [FREE Full text] [doi: 10.1136/bmjopen-2021-055658] [Medline: 35321893]
73. Sieck CJ, Sheon A, Ancker JS, Castek J, Callahan B, Siefer A. Digital inclusion as a social determinant of health. NPJ Digit Med 2021;4(1):52. [doi: 10.1038/s41746-021-00413-8] [Medline: 33731887]
74. Tao J, Zhao X, Li B, Sun H, Hu Y, Chen S, et al. Associations of educational attainment and traditional risk factor control with cardiovascular disease. Am J Prev Cardiol 2025;23:101031 [FREE Full text] [doi: 10.1016/j.ajpc.2025.101031] [Medline: 40585338]
75. Hua Z, Yuqing S, Qianwen L, Hong C. Factors influencing eHealth literacy worldwide: systematic review and meta-analysis. J Med Internet Res 2025;27:e50313 [FREE Full text] [doi: 10.2196/50313] [Medline: 40063939]
76. Chidambaram S, Jain B, Jain U, Mwavu R, Baru R, Thomas B, et al. An introduction to digital determinants of health. PLOS Digit Health 2024 Jan;3(1):e0000346 [FREE Full text] [doi: 10.1371/journal.pdig.0000346] [Medline: 38175828]

XSL•FO
RenderX

77. Powell-Wiley TM, Brewer LC, Burke LE, Hernandez R, Landsbaugh Kaar J, Kepper M, et al. Role of technology in promoting heart healthy behavior change to increase equity in optimal cardiovascular health: a scientific statement from the American Heart Association. Circulation 2025;151(18):e972-e985 [FREE Full text] [doi: 10.1161/CIR.0000000000001314] [Medline: 40190270]
78. Perrin Franck C, Babington-Ashaye A, Dietrich D, Bediang G, Veltsos P, Gupta PP, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023;25:e46694 Erratum in: J Med Internet Res. 2023;25:e49027. doi: 10.2196/49027 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]

Abbreviations

AHRQ: Agency for Healthcare Research and Quality
AXIS: Appraisal Tool for Cross-Sectional Studies
CVD: cardiovascular disease
DHL: digital health literacy
DHLI: Digital Health Literacy Instrument
DHRQ: Digital Health Readiness Questionnaire
eHEALS: eHealth Literacy Scale
eHLQ: eHealth Literacy Questionnaire
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
MeSH: Medical Subject Headings
NOS: Newcastle-Ottawa Scale
PICOS: participants, intervention or interest, comparisons, outcomes, and study design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
ROB: risk of bias
RoB 2: Revised Cochrane Risk-of-Bias Tool for Randomized Trials
ROBINS-I: Risk of Bias in Nonrandomized Studies of Interventions

Edited by S Brini; submitted 08.Dec.2025; peer-reviewed by D Nutbeam, M Harris; comments to author 27.Jan.2026; revised version received 22.Feb.2026; accepted 24.Feb.2026; published 24.Mar.2026.

Please cite as:
Kim EJ, Kim DY, Kim HJ, Son YJ
Determinants and Health Outcomes of Digital Health Literacy in Patients With Cardiovascular Disease: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e89102
URL: https://www.jmir.org/2026/1/e89102
doi:10.2196/89102
PMID:41874540

©Eun-Jung Kim, Da-Young Kim, Ha-Jin Kim, Youn-Jung Son. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Robot-Assisted Therapy for Upper Limb Rehabilitation After Stroke: Umbrella Review

Sijia Liu[1], MD; Xianghong Zhang[2], MSc; Lijun Zhou[2], MBBS; Jiashan Zhang[2], MBBS; Jialin Liu[1], MD; Chengqi He[1], MD

[1]West China Hospital of Sichuan University, Chengdu, China

[2]West China School of Clinical Medicine, Sichuan University, Chengdu, China

Corresponding Author:
Chengqi He, MD
West China Hospital of Sichuan University
No.37 Guoxuexiang, Wuhou District
Chengdu, 610041
China
Phone: 86 13880982177
Fax: 1 388 098 2177
Email: hxkfhcq2015@126.com

Abstract

Background: Stroke is a leading cause of long-term upper limb disability, severely impacting patients' independence and quality of life. Robot-assisted therapy (RAT) has emerged as a promising, high-intensity rehabilitation alternative. However, conclusions from existing systematic reviews on its efficacy are inconsistent and often lack a holistic framework, limiting their use for guiding personalized clinical decisions.

Objective: This study aims to systematically synthesize recent evidence on RAT for upper limb rehabilitation after stroke. Guided by the International Classification of Functioning, Disability and Health framework, it moves beyond singular outcomes to provide a multidimensional evaluation across body function, activity, and participation levels. The review aims to provide stratified guidance for clinical decision-making based on patient- and intervention-specific characteristics, thereby supporting evidence-based practice and informing future research.

Methods: This study included systematic reviews and meta-analyses published from January 1, 2019, to December 26, 2025, comparing RAT with conventional therapy for upper limb rehabilitation after stroke. Overall, 6 databases, including PubMed, Web of Science, and Embase, were searched. Two reviewers (XZ and LZ) independently performed study selection, data extraction, and quality assessment using the AMSTAR 2 tool. The synthesis integrated outcome measures and subgroup analyses derived from the included studies.

Results: This umbrella review included 21 meta-analyses encompassing 535 randomized controlled trials and 27,598 patients across acute, subacute, and chronic stroke stages. According to AMSTAR 2, 17 reviews were high quality, 3 moderate, and 1 critically low. The synthesis demonstrated that RAT was superior in improving upper limb motor function, but no statistically significant advantages were observed in activities of daily living compared to conventional therapy. Subgroup analyses revealed that treatment effects were influenced by stroke stage, upper limb motor impairment level, and robot type.

Conclusions: RAT is an effective intervention for improving upper limb motor function after stroke. However, its benefits are primarily observed at the level of body function, with limited evidence for long-term maintenance. The current evidence is constrained by significant outcome heterogeneity and methodological limitations inherent to umbrella reviews. Future research should validate these findings in broader clinical practice, focus on translating functional gains into sustained improvements in daily activities and participation, and include cost-effectiveness evaluations.

Trial Registration: PROSPERO CRD42024497183; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024497183

(J Med Internet Res 2026;28:e79363) doi:10.2196/79363

KEYWORDS

robot-assisted rehabilitation; stroke; upper limb rehabilitation; umbrella review; robot therapy effectiveness

XSL•FO
RenderX

Introduction

As one of the diseases with the highest disability rate worldwide, stroke severely affects upper limb motor function, significantly impacting patients' ability to perform activities of daily living (ADL) and their quality of life [1]. Upper limb motor impairment is highly prevalent post stroke, affecting nearly 80% of patients in the acute phase, with persistent deficits observed in 50%-60% of survivors at 6 months [2,3]. Consequently, early and effective motor intervention is crucial to improve upper limb function and enhance performance in daily activities. Although traditional physical therapy can effectively improve upper limb function in survivors with stroke, it is resource-intensive, costly, and often relies on potentially limited specialized facilities [4,5].

In recent years, robot-assisted therapy (RAT) has emerged as a major focus of research in upper limb rehabilitation, owing to its advantages, including quantifiability, high reproducibility, and task-oriented exercise [6,7]. Evidence from clinical studies suggests that RAT can enhance motor outcomes and promote neuroplasticity [8]. For instance, robotic-exoskeleton training has been shown to improve motor performance and cortical excitability [9]. Furthermore, task-oriented upper limb training with robotic assistance can modify sensorimotor cortex neuroplasticity and support motor control and learning [8]. Numerous systematic reviews have concluded that RAT can improve upper limb motor function and functional activity in survivors with stroke, irrespective of the specific device used [10-12].

However, the efficacy of RAT remains controversial [13]. For example, one study concluded that robotic rehabilitation did not yield significant advantages over conventional therapy in measures such as Fugl-Meyer Assessment-Upper Extremity (FMA-UE), Action Research Arm Test, and Motor Activity Log scores [14], while other studies have reported positive effects [15,16]. This heterogeneity may stem from factors such as variability in study populations (eg, stroke type) [17], differences in intervention protocols (eg, robot type) [18], and inconsistent assessment tools [19-21]. More critically, existing systematic reviews predominantly focus on single outcome measures (eg, motor function) and lack a comprehensive analysis of multidimensional recovery within the International Classification of Functioning, Disability and Health (ICF) framework [22], which distinguishes between body function, activity, and participation levels. This limitation prevents the provision of stratified guidance for clinical decision-making, which is tailored to patients' specific recovery goals and functional levels.

An umbrella review represents an advanced form of systematic review, distinguished by its capacity to synthesize findings from multiple existing systematic reviews or meta-analyses, thereby offering a more holistic and integrated perspective on a given research topic [23]. This methodology is particularly suited to mapping contradictions, identifying consistent patterns across reviews, and assessing the overall strength of evidence in a fragmented field. Building upon this methodology, the present study aims to systematically integrate the available systematic review evidence concerning RAT for upper limb rehabilitation following stroke. Through comprehensive analysis, we seek to elucidate not only the overall therapeutic effects but also identify key moderating factors that influence outcomes. Therefore, under the guidance of the ICF framework, this umbrella review aims to (1) synthesize and appraise evidence from existing systematic reviews on robot-assisted upper limb rehabilitation after stroke; (2) examine the consistency and strength of evidence across ICF domains; and (3) provide multidimensional, stratified guidance for clinical practice and future research directions.

Methods

Overview

The design, conduct, and reporting of this umbrella review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Multimedia Appendix 1) and its PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) checklist (Multimedia Appendix 2) [24,25]. Additionally, we adhered to the PRIOR (Preferred Reporting Items for Overviews of Reviews) checklist (Multimedia Appendix 3) to ensure comprehensive reporting [26].

Eligibility Criteria

Given the complexity of upper limb motor function and its critical role in daily life, upper limb rehabilitation robots have reached a relatively mature stage of research, development, and application. These systems enable high-precision, high-intensity, and personalized training, which is designed to promote functional recovery, quality of life, and self-care ability. Accordingly, this review specifically investigated the effects of robotic interventions on upper limb recovery post stroke. We excluded reviews that focused only on general stroke outcomes or those evaluating robotic interventions for lower limb rehabilitation. Furthermore, included studies were required to demonstrate clinical relevance through rigorous methodological quality and provide complete datasets to support their conclusions. To account for the rapid technological evolution in this field, we limited inclusion to meta-analyses published within the past 7 years.

Eligible studies involved adult patients with stroke undergoing upper limb rehabilitation and compared interventions incorporating RAT (as a stand-alone or adjunct treatment) against conventional therapy. The outcomes focused on clinical efficacy, encompassing upper limb motor function [27], muscle strength [28], spasticity [29], and ADL [30]. Meta-analyses lacking sufficient data and non-English publications were excluded.

Search Strategy

We performed a systematic literature search on December 26, 2025, using the Sichuan University Library Discovery System. The following databases were searched: PubMed, Web of Science, Embase, the Cochrane Library, IEEE Xplore, and Scopus. The search was restricted to peer-reviewed journal articles published in English between January 1, 2019, and the search date. The complete search syntax is detailed in Table S1

XSL•FO
RenderX

in Multimedia Appendix 4. The search strategy was developed de novo through team discussion to ensure conceptual adequacy. It was not based on any previously published strategy and was not updated after its initial execution. No supplementary search techniques were used, such as citation tracking, manual searches of grey literature, or contacting authors for unpublished data. The study protocol was prospectively registered on PROSPERO (CRD42024497183), and no additional trial registries were consulted. All retrieved records were imported into EndNote 20 for management. Duplicates were removed using the software's automated function, supplemented by manual verification, to compile a unique set of citations for subsequent screening.

Selection Process

Two reviewers (XZ and LZ) independently appraised the methodological quality of each included meta-analysis using the AMSTAR 2 tool [31]. Any disagreements were resolved through discussion with a third reviewer (SL) to reach consensus. The overall quality of each study was rated as high, moderate, low, or critically low.

Data Items

Two reviewers (XZ and LZ) independently performed data extraction using Microsoft Excel 2019. For the qualitative evidence synthesis, key information was extracted from each meta-analysis. This included all outcome comparisons between RAT and conventional therapy, as well as the moderators examined in prespecified subgroup analyses. The extracted moderators comprised intervention intensity, stroke stage (subacute or chronic), robot type (unilateral vs bilateral), training design (exoskeleton vs end-effector), and baseline upper limb motor impairment level.

Assessment of Overlap

To assess the degree of primary study overlap across the included systematic reviews, we constructed a citation matrix in tabular form and used the corrected covered area (CCA) index to quantify the overall extent of overlap, calculated using the following formula: $CCA = (N – r) \times (r \times c) – r$

where represents the total number of included publications (including duplicates), *r* denotes the number of index publications (the count of unique primary studies or the number of rows in the matrix), and is the number of reviews (or the number of columns in the matrix).

The CCA value was interpreted using commonly applied thresholds derived from Pieper et al [32] work: <5% (slight), 5%-10% (moderate), 11%-15% (high), >15% (very high). For meta-analyses with incomplete information, the following handling approach was adopted: (1) Maximized matching: using the available author and year information, cross-referencing and matching were performed against the references in the other 20 meta-analyses. (2) Conservative estimation: for studies that could not be uniquely matched, they were treated as independent studies in the calculation of the overall CCA to avoid underestimating overlap.

Synthesis Methods

Given the methodological limitations of statistically pooling effect sizes from different meta-analyses, this study used a narrative synthesis approach to summarize and evaluate the evidence. The results for each outcome measure were categorized based on the direction of the effect and its statistical significance: if RAT was significantly superior to conventional therapy, it was classified as "Significant;" if there was no significant difference or if conventional therapy was superior, it was classified as "Non-significant." The determination of statistical significance was based on whether the 95% CIs reported in the original meta-analyses included the null value.

In the synthesis and interpretation of the evidence, we fully incorporated considerations of heterogeneity and risk of bias. For heterogeneity, in addition to examining statistical indicators such as I^2, we focused on analyzing its methodological sources, including population characteristics, intervention protocols, and measurement tools. When heterogeneity was high, interpretations were made cautiously, relying on subgroup results. For the risk of bias, we considered both the methodological quality of the included meta-analyses (assessed using AMSTAR 2) and potential biases in the original studies. These factors were integrated to grade and critically discuss the evidence during result interpretation.

Protocol Amendments

During the study screening phase, the publication date range specified in the registration protocol (CRD42024497183) was updated from December 30, 2013-December 30, 2023, to January 1, 2019-December 26, 2025, to obtain the most recent evidence. All other methodological aspects followed the original protocol.

Results

Overview

From an initial pool of 1307 records identified, 21 meta-analyses [10,16,19-22,33-47] were included following deduplication and multiple rounds of screening. The study selection flow is detailed in Figure 1. The 21 meta-analyses collectively cited primary studies 591 times. After deduplication, 310 independent primary studies were obtained. Among these, multiple studies were repeatedly included in different systematic reviews. All studies that were cited 2 or more times (≥2) and their citation frequencies are summarized in Table S2 in Multimedia Appendix 5. The CCA was 4.532%, indicating a slight degree of overlap among the primary studies in this research.

XSL•FO
RenderX

Figure 1. Flowchart of study selection process in the umbrella review of robot-assisted therapy for upper limb rehabilitation poststroke.

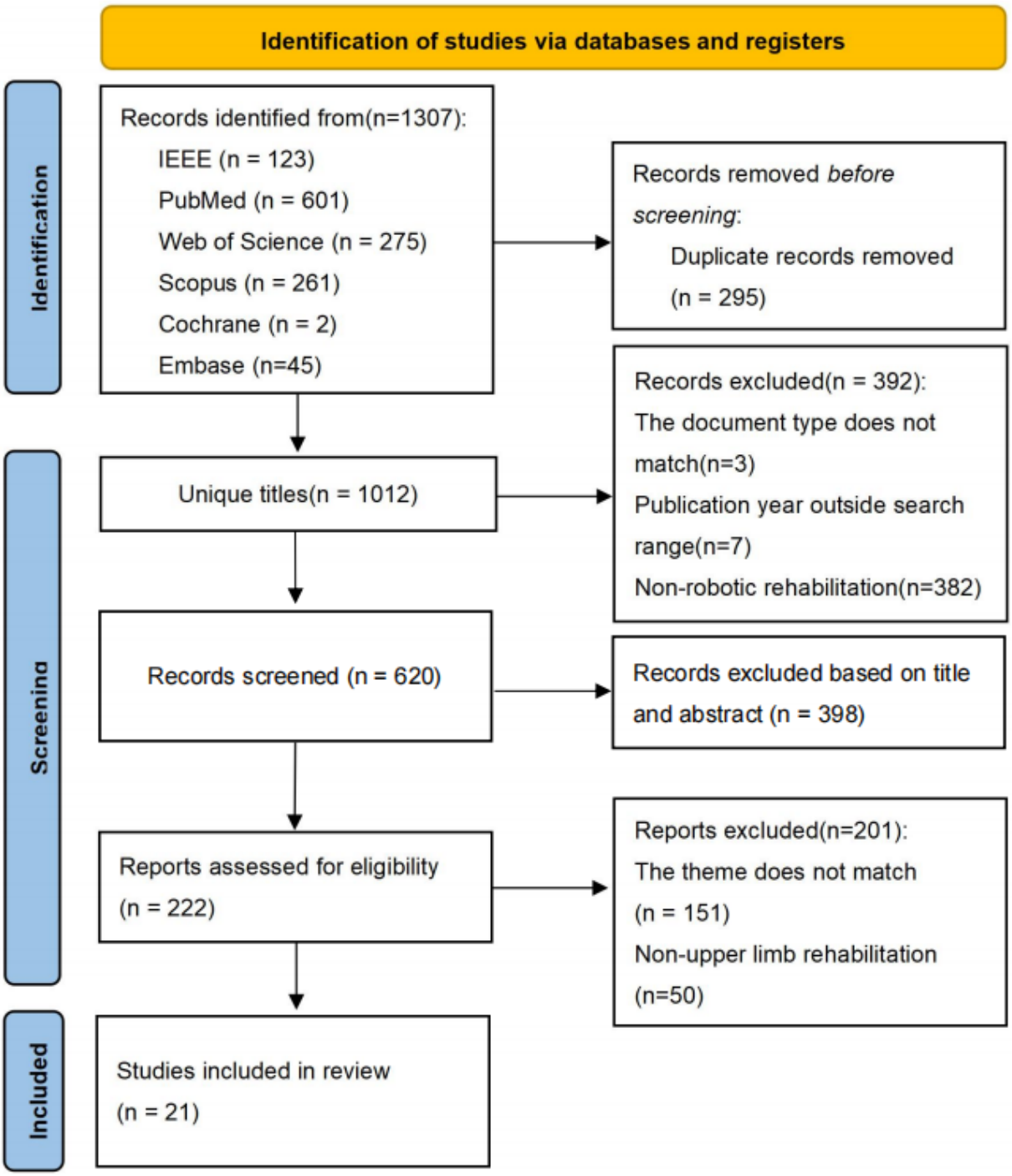

Overview of the Study Included

This review ultimately included 21 meta-analyses [10,16,19-22,33-47], encompassing a total of 535 randomized controlled trials and involving 27,598 patients with upper limb dysfunction after stroke. The study participants included patients in the acute, subacute, and chronic phases of stroke. The types of robots included are end-effector, exoskeleton, soft robotic gloves, etc. Table 1 provides a detailed overview of all the included meta-analyses.

XSL•FO
RenderX

Table 1. Characteristics of the meta-analyses included in the umbrella review on robot-assisted therapy for upper limb rehabilitation post stroke.

Authors	Number of papers	Sample size	Clinical status	Follow-up	Experimental group	Control group	Robotic type	Measurement instrument	Outcome
Tseng et al [33] (2024)	9	295 participants (*EG*[a], *n*=150; *CG*[b], *n*=145)	Survivors with stroke	—[c]	Portable rehabilitation robot. The average duration was 87.78 min. The average number of training sessions per week is 4.56. The average training period was 4.22 weeks	Conventional therapy or no therapy	End-effector-based, exoskeleton-based, and orthosis	FMA[d]	Portable robots prove to be effective (FMA: SMD[e]=0.696, 95%=0.099 to .293, *P*<.05).
Wu et al [34] (2021)	41	1916	Patients with stroke	—	Unilateral RAT[f] or bilateral RAT	Dose-matched conventional rehabilitation	End effector robots, exoskeleton robots	FMA	RAT is an effective intervention for improving upper extremity motor impairment in patients with stroke.
Marotta et al [35] (2021)	26	892	Acute-subacute (<8 weeks) people, chronic people	—	RAT	Conventional therapy	Robotic exoskeleton	FMA	RAT is significantly useful for people with stroke.
Ko et al [36] (2023)	8	309	Patients with post-stroke hemiparesis who had received or were scheduled to receive rehabilitation	—	Rehabilitation programs involving soft robotic gloves or other similar devices	Physical therapy and occupational therapy	Soft robotic gloves or other similar devices	FMA, MBI[g], MAS[h], FIM[i], WMFT[j]	Soft robotic gloves can promote the functional abilities of the upper extremities.
Bazan et al [37] (2022)	9	142, intervention group 73, and control group 69	Adults with an objective diagnosis of unilateral spatial neglect after stroke	—	Robotic limb activation in patients with unilateral spatial neglect after stroke	Conventional rehabilitation for unilateral spatial neglect after stroke	Exoskeleton or end-effector	MVPT-3[k], LBT[l], SCT[m], AT[n], CBS[o]	Limb activation through robotic therapy can improve midline perception. However, there was no impact on tasks assessing visual scanning, functionality, or activities of daily living.

XSL•FO
RenderX

Authors	Number of papers	Sample size	Clinical status	Follow-up	Experimental group	Control group	Robotic type	Measurement instrument	Outcome
Johansen et al [21] (2023)	18	1295, with an age range from 20 to 95 years of age	Persons with stroke	—	Commercialized RAT	Traditional occupational and physiotherapy	End-effector robots, exoskeleton	MAS, Jamar, BI[p]	A statistically significantly higher treatment effect in the robotic-assisted exercise group (P=<0.0001) compared to the traditional treatment group, with a total effect size of 0.44 (CI 0.22 to –0.65).
Carrillo et al [19] (2023)	14	1141, the age range of 20 to 85 years, with a median age of 57 years	Participants (age > 18 years) with stroke resulting in functional deficits in their upper extremities.	—	RAT	Conventional therapy	End-effector robots, exoskeleton	FMA	It is unclear whether robot-assisted therapy accelerates upper extremity recovery poststroke when used in conjunction with conventional therapy.
Huo et al [38] (2023)	13	330 (EG, n=175; CG, n=155)	Patients who have been suffering stroke, subacute	—	EMG[q]-based robot	Conventional rehabilitation	EMG-based robot	FMA, MAS, ADL[r]	The outcomes postintervention were significantly improved in the EMG-based robot group.
Chen et al [22] (2020)	35	2241	Patients diagnosed with stroke and having upper limb motor dysfunction	—	RAT. Time per session ranged from 30 minutes to 5 hours. The duration of the intervention ranged from 2 weeks to 12 weeks.	Conventional occupational therapy, physical therapy, task-specific training, ADL training, and constraint-induced movement therapy	Exoskeleton and end-effector	FMA, ARAT[s], WMFT, BBT[t], 9-HPT[u], CAHAI[v], AMAT[w]	RAT was slightly superior in motor impairment recovery.

XSL•FO
RenderX

Authors	Number of papers	Sample size	Clinical status	Follow-up	Experimental group	Control group	Robotic type	Measurement instrument	Outcome
Chien et al [10] (2020)	11	493, the participants were aged 18–65 years	6 months after the onset of stroke	—	The number of sessions for RT ranged from 9 to 40, and each session lasted for 30–120 mins. Participants received RAT for 5 days per week for 2–12 weeks	Usual care	End-effector robots, exoskeleton	FMA, FIM, BI, ARAT, WMFT	RAT produced benefits similar, but not significantly superior, to those from usual care for improving functioning and disability in patients diagnosed with stroke within 6 months.
Iaco et al [39] (2024)	86	4240	Patients with upper limb limitations post-stroke	—	Upper limb-robot therapy. Intervention ranged from 2 weeks to 12 weeks, with a mean upper limb-robot therapy time of 53.55 minutes per session and 4.56	Any other rehabilitative intervention (usual care or specific interventions), placebo, or no treatment	End-effector, exoskeleton	FMA	Small significant improvements in upper limb-muscle synergism, muscle power, motor performance, and basic ADLs.
Moggio et al [40] (2022)	5	149(78 males and 71 females)	Patients with finger-hand motor impairment stroke	—	Exoskeleton and end-effector RAT	Traditional or conventional rehabilitation therapy	End-effector (Amadeo) and exoskeleton	FMA, Quick-DASH[x]	MI[y] showed a significant improvement (P<.05) in the robotic intervention group.
Yang et al [16] (2023)	14	1275	Patients aged 18-80 years with a stroke diagnosis	—	RAT	Any comparative therapy, as well as treatment, as usual or no treatment	End-effector	FMA, MBI, MAS, FIM, WMFT	RAT can significantly enhance the upper limb motor function and activities of daily life in patients with stroke undergoing upper limb rehabilitation.

XSL•FO
RenderX

Authors	Number of papers	Sample size	Clinical status	Follow-up	Experimental group	Control group	Robotic type	Measurement instrument	Outcome
Zhang et al [20] (2022)	46	2533 participants with a mean age ranging from 46.20 to 75.5 years	The stroke patients were over 18 years old	≥ three months	The arm robot used in the intervention group included the Mirror Image Movement Enabler, UL-EXO7, Amadeo Robotic System, InMotion ARM 2.0 Robot, Aremo Spring, Bi-Manu-Track, Myomo e100, Neuro-Rehabilitation Robot, electromyography-driven robot, REJOYCE robot, Pneu-WREX, ReoGo system, and Gloreha robot. Patients received RAT 4 sessions per week for 6 weeks.	Non-robotic therapy	End-effector, exoskeleton, hybrid robot	FMA, MBI, FIM	RAT has significant immediate benefits for motor control and functional activity of the hemiparetic upper limb in patients after stroke.
Zhao et al [41] (2022)	22	758	Participants of either gender over 18 years of age after any duration of stroke	—	Robot-assisted distal training. The training lasted for ≤ 4 weeks with 20 sessions conducted 5 times per week for ≤ 60min each.	Therapist-assisted training or passive range of motion exercise	Exoskeleton robot, end-effector robot, and self-developed devices	FMA, 9-HPT, BBT, MAS, MRC^z, MI	The overall effect of robot-assisted distal training on the motor function of the wrists and hands was a significant improvement.
Bardsworth et al [42] (2025)	54	2744	Patients with stroke	Ranging from 2 weeks to 6 months	Guide patients to perform high-intensity, repetitive upper limb task training using mechanical-assisted devices.	Conventional rehabilitation	End-effector, Exoskeleton	FIM, BI, ADL, MAL	Robotic rehabilitation had a small, statistically significant positive effect on upper limb capacity compared with conventional rehabilitation (SMD 0.14, 95% CI 0.02-0.26).
Hwang et al [43] (2024)	31	708	Patients with acute stroke	—	RAT as the assistive technology	Conventional treatments	End effectors such as the Armeo Spring, sensor-based devices, InMotion2, ReHapticKnob, and Reo therapy systems.	FMA, ARAT, BBT, WMFT, FIM, MI	Upper-limb robots did not demonstrate significant superiority over conventional treatments in improving the function of upper limbs.

XSL•FO
RenderX

Authors	Number of papers	Sample size	Clinical status	Follow-up	Experimental group	Control group	Robotic type	Measurement instrument	Outcome
Jin et al [44] (2025)	15	574	Patients with stroke	—	The intervention included robot-assisted task-oriented training, without restrictions on the types of robots, training durations, intensities, or frequencies.	Conventional therapies	End-effector, Exoskeleton	FMA-UE[aa], MBI	Robot-assisted task-oriented training significantly enhances the rehabilitation of upper limb function and the recovery of daily living skills in patients with stroke.
Su et al [45] (2024)	18	573	Patients with stroke	—	Upper limb robot rehabilitation training	Conventional rehabilitation	End-effector, Exoskeleton	FMA-UE, WMD[ab], ARAT, MBI, MAS	Upper limb robot–assisted training is superior to conventional training in terms of improving upper limb motor impairment, ability to perform daily living activities, and muscle tone recovery.
Verola et al [46] (2025)	85	3452	Patients with stroke	—	RAT	Conventional physiotherapy	Exoskeleton, End-effector, Hand and Finger Rehabilitation Robots, Bilateral or Mirror-training Robots, Hybrid or Integrated Robotic Systems	FMA, BI, SIS[ac], FIM, BBT, ARAT, WMFT, MAS	RAT produces some significant improvements for the upper limb, but these differences are not clinically relevant when compared to other therapies.

XSL•FO
RenderX

Authors	Number of papers	Sample size	Clinical status	Follow-up	Experimental group	Control group	Robotic type	Measurement instrument	Outcome
Wang et al [47] (2025)	31	1538	Patients with stroke	—	RAT+conventional rehabilitation therapy	Conventional rehabilitation therapy	End-effector robot-assisted therapy, exoskeleton robot-assisted therapy	FIM, FMA-UE, MAS, MBI	RAT combined with routine rehabilitation therapy can effectively improve the upper limb motor function and activities of daily life of patients with stroke.

[a]EG: experimental group.
[b]CG: control group.
[c]NA: not applicable.
[d]FMA: Fugl-Meyer Assessment.
[e]SMD: standardized mean difference.
[f]RAT: robot-assisted therapy.
[g]MBI: Modified Barthel Index.
[h]MAS: Modified Ashworth Scale.
[i]FIM: Functional Independence Measure.
[j]WMFT: Wolf Motor Function Test,
[k]MVPT-3: motor-free visual perception test 3rd.
[l]LBT: line bisection test.
[m]SCT: star cancellation test.
[n]AT: Albert's test.
[o]CBS: Catherine Bergego Scale.
[p]BI: Barthel Index.
[q]EMG: Electromyography.
[r]ADL: activity of daily living.
[s]ARAT: Action Research Arm Test.
[t]BBT: Box and Blocks Test.
[u]9-HPT: Nine Hole Peg Test.
[v]CAHAI: Chedoke Arm and Hand Activity Inventory.
[w]AMAT: Arm Motor Ability Test.
[x]QuickDASH: quick version of disabilities of the arm, shoulder, and hand.
[y]MI: motricity index.
[z]MRC: Medical Research Council Scale.
[aa]FMA-UE: Fugl-Meyer Assessment-upper extremity.
[ab]WMD: weighted mean differences.
[ac]SIS: Stroke Impact Scale.

Methodological Quality of Included Reviews

The methodological quality of the included reviews was assessed using the AMSTAR 2 tool. We synthesized the reported findings from these reviews rather than conducting a new meta-analysis of primary data. According to the AMSTAR 2 appraisal, 17 of the 21 included meta-analyses were of high quality [10,16,20-22,34,35,37,39-47], 3 were of moderate quality [19,33,36], and one was of critically low quality [38] (Table 2). Downgrading was primarily due to factors such as unmet justification for significant statistical heterogeneity [36] or inadequate consideration of potential biases from the primary randomized controlled trials [38]. Despite these limitations, the evidence base is reliable, as the key conclusions are underpinned by a preponderance of medium- to high-quality studies (95.2%), indicating robust primary findings [40].

XSL•FO
RenderX

Table 2. Methodological quality assessment of included meta-analyses using the AMSTAR 2 tool.

Studies	Item 1	Item 2*	Item 3	Item 4*	Item 5	Item 6	Item 7*	Item 8	Item 9*	Item 10	Item 11*	Item 12	Item 13*	Item 14	Item 15*	Item 16	AMSTAR 2 overall rating
Tseng et al [33] (2024)	Yes	Yes	Yes	Partial Yes	No	No	Partial Yes	Partial Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Moderate
Wu et al [34] (2021)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Yes	Partial Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Marotta et al [35] (2021)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Partial Yes	Partial Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Ko et al [36] (2023)	Yes	Partial Yes	Yes	Partial Yes	Yes	Yes	Yes	Partial Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Moderate
Bazan [37] (2022)	Yes	Partial Yes	Yes	Partial Yes	Yes	Yes	Yes	Partial Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Johansen [21] (2023)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Partial Yes	Partial Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Carrillo [19] (2023)	Yes	Partial Yes	Yes	Partial Yes	Yes	Yes	Partial Yes	Partial Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Moderate
Huo [38] (2023)	Yes	Yes	Yes	Partial Yes	Yes	Yes	No	Partial Yes	Yes	No	Yes	Yes	Yes	No	No	Yes	Critically low
Chen et al [22] (2020)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Partial Yes	Partial Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Chien et al [10] (2020)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Partial Yes	Partial Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Iaco et al [39] (2024)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Partial Yes	Partial Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Moggio et al [40] (2022)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Partial Yes	Partial Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Yang et al [16] (2023)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Partial Yes	Partial Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	High
Zhang et al [20] (2022)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Partial Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Zhao et al [41] (2022)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Partial Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High

XSL•FO
RenderX

Studies	Item 1	Item 2*	Item 3	Item 4*	Item 5	Item 6	Item 7*	Item 8	Item 9*	Item 10	Item 11*	Item 12	Item 13*	Item 14	Item 15*	Item 16	AMSTAR 2 overall rating
Boardsworth et al [42] (2025)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Hwang et al [43] (2024)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	High
Jin et al [44] (2025)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Su et al [45] (2024)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Partial Yes	Yes	No	Yes	Yes	High
Verola et al [46] (2025)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	High
Wang et al [45] (2025)	Yes	Yes	Yes	Partial Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Partial Yes	Yes	No	Yes	Yes	High

*Indicates that the corresponding item is critical.

Primary Outcome: Upper Limb Motor Function

The Fugl-Meyer Assessment (FMA) is the “gold standard” for assessing upper-extremity sensorimotor function, due to its ability to assess aspects such as movement within synergies, mixing synergies, reflexes, wrist, hand, grip, coordination, and speed movements, thus providing a large amount of information that is very useful for understanding the sensorimotor capacity of the affected upper limb after a stroke [48]. The FMA was adopted as the primary outcome measure in 18 out of the 21 included meta-analyses (85.7%) to evaluate the restoration of upper limb motor function.

However, the specific clinical questions explored by various studies using FMA differ significantly, covering multiple dimensions such as the immediacy and sustainability of therapeutic effects [36], responses in populations at different stroke stages [16,21], and impacts on other functional domains [35]. By synthesizing these studies, the consistent use of FMA provides highly comparable and reliable evidence for a comprehensive, multifaceted evaluation of the efficacy of robot-assisted therapy.

Immediate Effects on Upper Limb Function

This study synthesizes evidence on the immediate effects of RAT from multiple meta-analyses. As shown in Table 3, various studies used standardized mean difference (SMD), mean difference (MD), weighted mean differences (WMD), and Hedges gas effect size metrics. Eight independent effect size estimates from different meta-analyses demonstrated statistically significant positive effects of RAT [20,33,34,36,42,44,45,47], supporting its immediate benefits. However, one study reported a non-significant result [43], and variations were observed in the magnitude of effects. The meta-analysis demonstrated moderate heterogeneity (I^2=35%), with variations among the included studies in terms of population characteristics, device types, and treatment protocols. Moreover, the study’s own assessment indicated potential risks of bias in areas such as intervention implementation and completeness of outcome data.

XSL•FO
RenderX

Table 3. Summary of meta-analyses on the immediate effects of robot-assisted therapy compared to conventional therapy on upper limb function in patients with stroke.

Studies	Effect size	Value	95% CI	Statistical significance
Tseng et al [33] (2024)	SMD[a]	0.696	0.09-1.293	Significant
Zhang [20] (2022)	SMD	0.2	0.08-0.32	Significant
Ko et al [36] (2023)	MD[b]	6.52	3.65-9.39	Significant
Wu et al [34] (2021)	Hedges *g*	0.25	0.11-0.38	Significant
Boardsworth et al [42] (2025)	SMD	0.14	0.02-0.26	Significant
Hwang et al [43] (2024)	MD	4.99	–0.07 to 10.05	Non-significant
Jin et al [44] (2025)	SMD	1.01	0.57-1.45	Significant
Su et al [45] (2024)	WMD	5.27	3.36-7.17	Significant
Wang et al [47] (2025)	MD[c]	5.92	3.52-8.32	Significant

[a]SMD: standardized mean difference.

[b]MD: mean difference.

[c]WMD: weighted mean differences.

Long-Term Maintenance of Upper Limb Function

Regarding the long-term maintenance effects of RAT, this review confirms that RAT demonstrates sustained short-term efficacy in improving upper limb function [20,36,42,45], as shown in Table 4. Meanwhile, our comprehensive analysis reveals a significant time-gradient effect: while significant functional improvements were observed immediately after the intervention and during short-term follow-up (≤12 weeks) [36,45], these advantages generally diminished or even disappeared during long-term follow-up (>12 weeks) [20,42]. This time-dependent pattern of efficacy decay indicates notable limitations in current robotic rehabilitation strategies for maintaining long-term therapeutic effects.

Table 4. Effects of robot-assisted therapy on the long-term maintenance of upper limb function in stroke survivors, stratified by follow-up duration.

Studies	Follow-up	Effect size	Value	95% CI	Statistical significance
Ko et al [36] (2023)	≤12 weeks	MD[a]	7.79	5.03-10.55	Significant
Zhang et al [20] (2022)	>12weeks	SMD[b]	–0.07	–0.21 to 0.07	Non-significant
Boardsworth et al [42] (2025)	NA	SMD	0.05	–0.13 to 0.24	Non-significant
Su et al [45] (2024)	＞4weeks	WMD[c]	6.63	3.46-9.8	Significant
Su et al [45] (2024)	≤4 weeks	WMD	4.49	2.11-6.88	Significant

[a]MD: mean difference.

[b]SMD: standardized mean difference.

[c]WMD: weighted mean differences.

Secondary Outcomes

Spasticity

The impact of RAT on spasticity represents another critical dimension for evaluation. The Motor Assessment Scale (MAS) serves as a core, standardized tool for assessing upper limb spasticity following stroke [49]. Although 4 meta-analyses have investigated this issue [16,21,38,41], and multiple studies have reported that robotic training can significantly improve spasticity, as shown in Table 5, their conclusions have been challenged by key research. For instance, some studies contained directional contradictions or reporting flaws, making the results difficult to interpret reasonably [41].

XSL•FO
RenderX

Table 5. Meta-analysis results of the effects of robot-assisted therapy on upper limb spasticity (measured by Modified Ashworth Scale) in patients after stroke.

Studies	Effect size	Value	95% CI	*P* value	Statistical significance
Johansen et al [21] (2023)	MD[a]	–0.24	–1.33 to 0.22	<.00001	Significant
Huo et al [38] (2023)	MD	–0.42	–0.82 to –0.03	.03	Non-significant
Zhao et al [41] (2022)	MD	0.18	–0.32 to –0.04	.01	Significant
Yang et al [16] 2023	SMD[b]	–1.49	–2.85 to –0.12	.05	Non-significant

[a]MD: mean difference.

[b]SMD: standardized mean difference.

Current evidence demonstrates marked heterogeneity (MD –0.42, 95% CI –0.82 to –0.03, I^2=83%, *P*=.03) [38], primarily due to differences among the included studies in terms of population characteristics, intervention protocols, and assessment methods. This heterogeneity makes it difficult to directly compare or synthesize findings across studies. Therefore, the use of the RAT was not more significant than conventional treatment in improving spasticity [16].

Activity of Daily Living

Six meta-analyses evaluated the effect of robot-assisted therapy on the ADL of patients with stroke [20-22,41,42,46]. The evidence regarding its overall immediate effects remains inconsistent, as shown in Table 6. Although some studies reported significant positive effects at the end of the intervention period, heterogeneity among the studies limits the robustness of the conclusions [20,41].

Table 6. Effect sizes for the impact of robot-assisted therapy on the activity of daily living in patients with stroke.

Studies	SMD[a]	95% CI	*P* value	Statistical significance
Johansen et al [21] (2023)	0.11	–0.04 to 0.25	.17	Non-significant
Chen et al [22] (2023)	0.0049	–0.055 to 0.17	.15	Non-significant
Zhang et al [20] (2022)	0.32	0.16-0.47	<.0001	Significant
Zhao et al [41] (2022)	0.7	–0.29 to 1.11	<.001	Significant
Boardsworth et al [42] (2025)	0.04	–0.05 to 0.13	.86	Non-significant
Verola et al [46] 2025	0.29	0.15-0.43	＜.01	Significant

[a]SMD: standardized mean difference.

Long-term follow-up data further indicate that the initial therapeutic benefits may not be sustainable. For instance, a meta-analysis by Zhang et al [20] demonstrated that the effect size diminished and became statistically non-significant during the follow-up period (SMD=0.09, 95% CI –0.06 to 0.23, I^2=38%), suggesting that the improvement in ADL from RAT may diminish over time. Subgroup analysis provided key insights for understanding the heterogeneity in treatment effects, indicating that the degree of ADL improvement is moderated by patients' baseline upper limb motor impairment levels and robotic training modalities. Among these, the level of active engagement during training emerged as a significant influencing factor [20].

Subgroup Analysis Results

Device Types and Patient Upper Limb Motor Impairment Levels

Based on current evidence, the differential efficacy of RAT for patients with mild-to-moderate versus severe upper limb motor impairment demonstrates a complex interaction pattern, as shown in Table 7. Device type may well be the key moderating factor explaining this heterogeneity. Wu et al [34] demonstrated that end-effector robots exhibited clear advantages (Hedges *g*=0.22, 95% CI 0.09-0.36, I^2=35.4%). This finding provides a crucial explanation for their overall subgroup results. However, the study by Su et al [45] found that exoskeletons (WMD 6.90, 95% CI 4.33-9.47) demonstrated a superior effect on upper limb motor impairment compared to end-effector devices (WMD 3.28, 95% CI 0.44-6.12).

XSL•FO
RenderX

Table 7. Subgroup analysis of the efficacy of robot-assisted therapy on upper limb function in patients with stroke, stratified by baseline motor impairment level.

Subgroups and studies	Effect size	Value	95% CI	Statistical significance
Mild-to-moderate				
Wu et al [34] (2021)	Hedges *g*	0.19	–0.01 to 0.4	Non-significant
Zhang et al [20] (2022)	SMD[a]	0.26	0.09-0.42	Significant
Severe				
Wu et al [34] (2021)	Hedges *g*	0.27	0.08-0.46	Significant
Zhang et al [20] (2022)	SMD	0.14	–0.01 to 0.3	Non-significant

[a]SMD: standardized mean difference.

The meta-analysis by Zhang et al [20], which included end-effector, exoskeleton, and hybrid robotic devices, yielded results more supportive of the therapy's benefits for patients with mild-to-moderate upper limb motor impairment. The discrepancy between the findings of these 2 studies strongly suggests a specific matching relationship between robot types and patients' upper limb motor impairment levels: end-effector robots may hold unique value for patients with severe impairment, while more complex devices like hybrid systems may provide additional benefits for those with mild-to-moderate upper limb motor impairment [20].

Stroke Stage

Based on current meta-analytical evidence, the efficacy of RAT for upper limb rehabilitation following stroke demonstrates a distinct phase-dependent characteristic [50]. As shown in Table 8, within the population of patients with chronic stroke, the evidence demonstrates a high degree of consistency. Six studies [16,21,34-36,38] reported statistically significant functional improvements, supporting the effectiveness of this therapy during this phase.

Table 8. Subgroup analysis comparing the efficacy of robot-assisted therapy on upper limb function in patients with subacute stroke.

Subacute subgroup and studies	Effect size	Value	95% CI	Statistical significance
Johansen et al [21] (2023)	SMD[a]	0.5	0.2-0.8	Significant
Huo et al [38] (2023)	SMD	1.18	0.29-2.07	Significant
Marotta et al [35] (2021)	SMD	4.09	1.31-6.87	Significant
Yang et al [16] (2023)	SMD	–0.16	–0.56 to 0.24	Non-significant

[a]SMD: standardized mean difference.

In contrast, the evidence for patients in the subacute phase demonstrates marked heterogeneity. Among four relevant studies [16,21,35,38], three reported significant benefits [21,35,38], while one showed no significant effects [36], indicating that a clear consensus regarding its efficacy at this stage has yet to be established. These divergent outcomes may primarily stem from variations in intervention protocols, types of robotic devices, training dosage, and program design.

Secondly, the ambiguity in study population definitions may introduce bias. The criteria for defining the "subacute phase" across studies could lead to differences in the actual recovery stages and potential of included patients [51]. Finally, methodological heterogeneity cannot be overlooked. Variations in inclusion criteria, statistical methods, and bias risk assessment among different meta-analyses could all affect the final synthesis of results. These factors collectively highlight the need for caution when interpreting the evidence, and future research should aim to clarify these complex relationships through more detailed subgroup analyses and standardized definition criteria.

Discussion

Principal Findings

Based on the systematic synthesis of existing meta-analytical evidence conducted in this umbrella review, we are able to examine the value and position of RAT in poststroke upper limb rehabilitation from a broader perspective. Importantly, our analysis is guided by the ICF framework, which provides a structured approach to evaluating outcomes across different dimensions of recovery. The main findings, clinical implications, and future research directions of this study are outlined as follows.

Efficacy Across ICF Levels: From Body Function to Activity

This review confirms that RAT demonstrates certain efficacy in improving upper limb motor function at the ICF body functions level, as measured by the FMA [36]. However, our synthesis reveals a crucial gradient in treatment effects across ICF levels. First, the therapeutic effects show significant stage-dependency, with the most robust and consistent evidence supporting its effectiveness for patients in the chronic phase

XSL•FO
RenderX

[16,34,35], while its efficacy for subacute patients remains inconsistent due to variations in intervention protocols and patient characteristics [16]. Second, a prominent finding is the limited translation of benefits to the ICF activity level. Although it effectively improves upper limb function and reduces spasticity [34,41], its effects on higher-level functions such as hand dexterity, ADL, and social participation are limited and inconsistent [22,41].

Specifically, when assessing manual dexterity, studies using the Nine-Hole Peg Test demonstrated that the robot-assisted group completed tasks significantly faster than the conventional therapy group [22,38], whereas studies using the Box and Block Test found no significant advantage [41]. More importantly, although RAT performed no worse than conventional therapist-guided training in improving both ADL and enhancing social participation, neither domain showed statistically significant superiority for the robotic approach [22,41]. This dissociation between body functions and activity-level outcomes represents a key challenge in robotic rehabilitation.

Moderators of Treatment Response

Within the ICF framework, we identified several critical moderators that explain the substantial heterogeneity in treatment effects. Intervention intensity is a critical factor influencing therapeutic efficacy. Zhang et al [20] found that when the total treatment duration exceeded 15 hours, the RAT group showed significant improvements in both motor control and functional activity. However, Yang et al [16] reported that excessively long treatment cycles may lead to diminishing returns. Notably, Carrillo et al [19] evaluated the impact of different training intensities on FMA scores and demonstrated that higher-intensity interventions were most beneficial for functional recovery. Furthermore, multiple dimensions of training intensity (session duration, frequency, total cycle) need to be considered synergistically [45,47]. Johansen et al [21] found that significant therapeutic effects were achieved when the total intervention dose reached 1375.33 minutes, while Ko et al [36] subgroup analysis regarding total training duration indicated no significant improvements, regardless of whether the training period exceeded or fell short of 2 weeks. This finding suggests that we need to consider interactions among dose parameters [36].

Device Characteristics and Training Modalities

Technical characteristics significantly moderate therapeutic efficacy. Wu et al [34] found that end-effector robots demonstrated superior effectiveness in improving upper limb motor impairment compared to conventional rehabilitation therapy (Hedges g=0.22; 95% CI 0.09-0.36, I^2=35.4%), potentially attributable to their multijoint coordination training mechanism [52]. Moggio et al [40] further compared device types and revealed that exoskeleton devices showed advantages in enhancing overall hand function and reducing disability, while the end-effector device demonstrated better performance in specific motor control. The selection of training modes is equally crucial for achieving functional gains. Wu et al [34] confirmed that unilateral robot training demonstrated clear efficacy, whereas the advantages of bilateral training and combined unilateral-bilateral training regimens remain unsubstantiated [34,38]. Yang et al [16] emphasized that training modes requiring active patient participation are essential for achieving improvements in ADL.

Integration of Secondary Outcomes

When comprehensively evaluating the effects of RAT on poststroke upper limb function, changes in muscle strength, as one of the core indicators at the ICF body functions level, hold significant clinical reference value. This umbrella review synthesizing relevant meta-analyses found that robotic therapy demonstrates a positive trend in improving muscle strength, though the strength of evidence remains limited. For instance, the analysis by Zhao et al [41] indicated that the upper limb strength in the RAT group was significantly greater than in the control group (SMD=0.42, 95% CI 0.07-0.78, I^2=4%), a finding consistent with the results reported by Johansen et al [21] (SMD=0.43, 95% CI 0.16-0.71, I^2=46%). However, evidence in this area shows inconsistency, as some studies found no significant between-group differences (MD=0.51, 95% CI –0.06 to 1.09, I^2=0%) [41]. This heterogeneity may stem from differences in assessment tools, insufficient specificity of training protocols for strength enhancement, and variations in patients' baseline levels.

Clinical Translation and Health Economic Perspectives

From a health economic perspective, the cost-effectiveness of rehabilitation robotics still warrants careful evaluation. Currently, the price range for such assistive robotic products varies widely, ranging from US $9000 to US $100,000 [33,53]. Robot-assisted rehabilitation requires not only a high initial investment but also involves relatively high system maintenance and operational costs [54,55]. Notably, the ICF framework helps explain the current cost-effectiveness challenges: the clinical benefits are largely confined to the body functions level, with limited impact on the activity level and virtually no high-quality evidence regarding participation-level outcomes. Furthermore, the implementation of robotic rehabilitation is influenced by health care system disparities and cultural contexts [56]. The high costs may pose significant barriers in resource-limited settings, while cultural differences in technology acceptance can affect patient engagement and treatment adherence [57]. Concurrently, the clinical benefits derived from robotic rehabilitation by both patients and health care institutions remain relatively limited, resulting in a cost-effectiveness ratio that has not yet reached an ideal level [58].

Furthermore, patients' attitudes toward RAT present a mixed picture. Multiple studies indicate that participants perceive robotic therapy as more precise and less invasive, and hold higher expectations regarding treatment outcomes [59,60]. Such positive perceptions may enhance clinical acceptance of the technology. However, a considerable proportion of patients express concerns about the risk of device malfunctions, with some even mistakenly believing that robotic therapy has higher error rates than conventional methods [60-62]. This crisis of trust could significantly impact treatment adherence and ultimately compromise therapeutic outcomes [63].

Currently, evidence regarding the ability of rehabilitation robots to reduce overall health care costs remains insufficient, and their

XSL•FO
RenderX

potential advantages in controlling medical expenditures require further validation [64]. Considering multiple factors, including equipment costs, patient acceptance, health care system differences, and cultural contexts, the pathway for clinical integration of robot-assisted rehabilitation still demands careful planning. Future research should not only focus on more rigorous economic evaluations but also address how to enhance patient trust through technical transparency and improved clinician-patient communication, while developing implementation strategies adapted to diverse health care systems and cultural environments, thereby ultimately improving the cost-benefit profile.

Limitations and Future Directions

It should be noted that this umbrella review has several limitations. First, although our synthesis approach accounted for heterogeneity and bias, the inability to reanalyze individual participant data from the primary studies limited a deeper exploration of heterogeneity sources. Second, some included meta-analyses contained methodological flaws or had missing outcome data, factors that could affect the accuracy of the conclusions. Additionally, as a secondary synthesis of existing evidence, this review cannot circumvent potential biases inherent in the evidence base, including possible publication bias (where studies with positive results are more likely to be published) and the fact that some primary studies may have received commercial funding or involved conflicts of interest. Such factors could compromise the objectivity of the evidence. Third, the comprehensive cross-checking and deduplication of the included systematic reviews and meta-analyses, which themselves contain extensive references to indexed primary studies, required immense and highly time-consuming effort [65]. This posed a significant methodological challenge in the implementation of this study.

Building on the ICF framework, we recommend several key directions for future research. Future studies should prioritize the investigation of participation level outcomes to fully understand the technology's impact on patients' lives. Furthermore, research should explicitly target the transfer of gains from body functions to activities and participation levels. Future umbrella reviews could enhance the reliability of their findings by expanding the literature search scope, strengthening the assessment of bias risks in primary studies, and raising the quality thresholds for included studies. In future research, artificial intelligence technologies could be leveraged to further optimize the identification and processing of literature overlap, thereby enhancing both efficiency and accuracy.

These limitations suggest that the conclusions of this review should be interpreted with due caution. However, despite these limitations, our application of the ICF framework provides a robust structure for understanding and advancing the field of robot-assisted upper limb rehabilitation.

Conclusions

This umbrella review demonstrates that RAT serves as an effective intervention for poststroke upper limb rehabilitation, with the most robust and consistent evidence supporting its efficacy in improving motor function, particularly demonstrating clear benefits for patients in the chronic phase. However, its therapeutic advantages are primarily concentrated in short-term functional improvement, while long-term efficacy maintenance remains insufficient, constituting a core bottleneck in current clinical application. The therapeutic effect is significantly modulated by robot type, training mode, and intervention intensity. Future research should focus on developing stratified and individualized treatment protocols based on patient injury characteristics, rehabilitation stage, and device features, and conduct cross–health care system health economic evaluations to advance the clinical translation of precision rehabilitation technologies.

Data Availability

All data generated or analyzed during this study are included in its supplementary information files.

Funding

This study was funded by grants from the National Natural Science Foundation of China (No.82102680) and the Sichuan Provincial Department of Science and Technology (No.2024YFFK0143). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Authors' Contributions

Conceptualization: LS (lead) and ZX (equal)
Data curation: ZX
Formal analysis: ZX (lead), ZL (supporting), and ZJ (supporting)
Funding acquisition: LS
Investigation: LS
Methodology: LS and ZX
Project administration: LS (lead), LJ (equal), and HC (equal)
Supervision: HC (lead) and LJ (equal)
Validation: HC
Visualization: ZX
Writing – original draft: LS (lead) and ZX (equal)

XSL•FO
RenderX

Writing – review and editing:LS (lead) and ZX (equal)

Conflicts of Interest

The authors declare that they have no competing interests.

Multimedia Appendix 1
PRISMA Checklist.
[PDF File (Adobe PDF File), 97 KB - jmir_v28i1e79363_app1.pdf]

Multimedia Appendix 2
PRISMA-S Checklist.
[PDF File (Adobe PDF File), 65 KB - jmir_v28i1e79363_app2.pdf]

Multimedia Appendix 3
PRIOR Checklist.
[PDF File (Adobe PDF File), 260 KB - jmir_v28i1e79363_app3.pdf]

Multimedia Appendix 4
Search processes for different databases.
[PDF File (Adobe PDF File), 73 KB - jmir_v28i1e79363_app4.pdf]

Multimedia Appendix 5
Studies cited two or more times and their citation frequencies.
[PDF File (Adobe PDF File), 96 KB - jmir_v28i1e79363_app5.pdf]

References

1. Kuriakose D, Xiao Z. Pathophysiology and treatment of stroke: present status and future perspectives. Int J Mol Sci 2020;21(20):7609 [FREE Full text] [doi: 10.3390/ijms21207609] [Medline: 33076218]
2. Langhorne P, Coupar F, Pollock A. Motor recovery after stroke: a systematic review. Lancet Neurol 2009;8(8):741-754. [doi: 10.1016/S1474-4422(09)70150-4] [Medline: 19608100]
3. Wen X, Li L, Li X, Zha H, Liu Z, Peng Y, et al. Therapeutic role of additional mirror therapy on the recovery of upper extremity motor function after stroke: a single-blind, randomized controlled trial. Neural Plast 2022;2022:8966920 [FREE Full text] [doi: 10.1155/2022/8966920] [Medline: 36624743]
4. Hu J, Zhao L. Rehab mirror: the application of virtual mirror therapy in post-stroke upper limb rehabilitation. : IEEE; 2024 Presented at: Proceedings of the 17th International Symposium on Computational Intelligence and Design (ISCID); December 14-15, 2024; Hangzhou, China p. 222-225. [doi: 10.1109/iscid63852.2024.00057]
5. Shen YC, Kim AS, Hsia RY. Treatments and patient outcomes following stroke center expansion. JAMA Netw Open 2024;7(11):e2444683 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.44683] [Medline: 39535793]
6. Morone G, Cocchi I, Paolucci S, Iosa M. Robot-assisted therapy for arm recovery for stroke patients: state of the art and clinical implication. Expert Rev Med Devices 2020;17(3):223-233. [doi: 10.1080/17434440.2020.1733408] [Medline: 32107946]
7. Zhang Y, Zhang Y, Zheng B, Chen S, Yu H, Dai L, et al. The effects of combining anodal transcranial direct current stimulation with robot-assisted gait training on lower limb motor function and the motor cortex regulation of stroke patients. J Neuroeng Rehabil 2025;22(1):230 [FREE Full text] [doi: 10.1186/s12984-025-01731-8] [Medline: 41163056]
8. Xie H, Li X, Huang W, Yin J, Luo C, Li Z, et al. Effects of robot-assisted task-oriented upper limb motor training on neuroplasticity in stroke patients with different degrees of motor dysfunction: a neuroimaging motor evaluation index. Front Neurosci 2022;16:957972 [FREE Full text] [doi: 10.3389/fnins.2022.957972] [Medline: 36188465]
9. Singh N, Saini M, Kumar N, Srivastava MVP, Mehndiratta A. Evidence of neuroplasticity with robotic hand exoskeleton for post-stroke rehabilitation: a randomized controlled trial. J Neuroeng Rehabil 2021;18(1):76 [FREE Full text] [doi: 10.1186/s12984-021-00867-7] [Medline: 33957937]
10. Chien WT, Chong YY, Tse MK, Chien CW, Cheng HY. Robot-assisted therapy for upper-limb rehabilitation in subacute stroke patients: a systematic review and meta-analysis. Brain Behav 2020;10(8):e01742 [FREE Full text] [doi: 10.1002/brb3.1742] [Medline: 32592282]
11. Mehrholz J, Pohl M, Platz T, Kugler J, Elsner B. Electromechanical and robot-assisted arm training for improving activities of daily living, arm function, and arm muscle strength after stroke. Cochrane Database Syst Rev 2018;9(9):CD006876 [FREE Full text] [doi: 10.1002/14651858.CD006876.pub5] [Medline: 30175845]

XSL•FO
RenderX

12. Mehrholz J, Pollock A, Pohl M, Kugler J, Elsner B. Systematic review with network meta-analysis of randomized controlled trials of robotic-assisted arm training for improving activities of daily living and upper limb function after stroke. J Neuroeng Rehabil 2020;17(1):83 [FREE Full text] [doi: 10.1186/s12984-020-00715-0] [Medline: 32605587]
13. Park JM, Park HJ, Yoon SY, Kim YW, Shin JI, Lee SC. Effects of robot-assisted therapy for upper limb rehabilitation after stroke: an umbrella review of systematic reviews. Stroke 2025;56(5):1243-1252. [doi: 10.1161/STROKEAHA.124.048183] [Medline: 40115991]
14. Coskunsu DK, Akcay S, Ogul OE, Akyol D, Ozturk N, Zileli F, et al. Effects of robotic rehabilitation on recovery of hand functions in acute stroke: a preliminary randomized controlled study. Acta Neurol Scand 2022;146(5):499-511. [doi: 10.1111/ane.13672] [Medline: 35855628]
15. Bertani R, Melegari C, De Cola MC, Bramanti A, Bramanti P, Calabrò RS. Effects of robot-assisted upper limb rehabilitation in stroke patients: a systematic review with meta-analysis. Neurol Sci 2017;38(9):1561-1569. [doi: 10.1007/s10072-017-2995-5] [Medline: 28540536]
16. Yang X, Shi X, Xue X, Deng Z. Efficacy of robot-assisted training on rehabilitation of upper limb function in patients with stroke: a systematic review and meta-analysis. Arch Phys Med Rehabil 2023;104(9):1498-1513. [doi: 10.1016/j.apmr.2023.02.004] [Medline: 36868494]
17. Qu H, Zeng F, Tang Y, Shi B, Wang Z, Chen X, et al. The clinical effects of brain-computer interface with robot on upper-limb function for post-stroke rehabilitation: a meta-analysis and systematic review. Disabil Rehabil Assist Technol 2024;19(1):30-41. [doi: 10.1080/17483107.2022.2060354] [Medline: 35450498]
18. Carrión-Téllez V, Pastor-Zaplana J, Compañ-Gabucio LM, Peral-Gómez P, García-Aracil NM. Use of robotic devices in upper limb rehabilitation for adult stroke patients: protocol for an umbrella review. Health Sci Rep 2026;9(1):e71684 [FREE Full text] [doi: 10.1002/hsr2.71684] [Medline: 41467237]
19. Carrillo C, Tilley D, Horn K, Gonzalez M, Coffman C, Hilton C, et al. Effectiveness of robotics in stroke rehabilitation to accelerate upper extremity function: systematic review. Occup Ther Int 2023;2023:7991765 [FREE Full text] [doi: 10.1155/2023/7991765] [Medline: 37927581]
20. Zhang L, Jia G, Ma J, Wang S, Cheng L. Short and long-term effects of robot-assisted therapy on upper limb motor function and activity of daily living in patients post-stroke: a meta-analysis of randomized controlled trials. J Neuroeng Rehabil 2022;19(1):76 [FREE Full text] [doi: 10.1186/s12984-022-01058-8] [Medline: 35864524]
21. Johansen T, Sørensen L, Kolskår KK, Strøm V, Wouda MF. Effectiveness of robot-assisted arm exercise on arm and hand function in stroke survivors - a systematic review and meta-analysis. J Rehabil Assist Technol Eng 2023;10:20556683231183639 [FREE Full text] [doi: 10.1177/20556683231183639] [Medline: 37426037]
22. Chen Z, Wang C, Fan W, Gu M, Yasin G, Xiao S, et al. Robot-assisted arm training versus therapist-mediated training after stroke: a systematic review and meta-analysis. J Healthc Eng 2020;2020:8810867 [FREE Full text] [doi: 10.1155/2020/8810867] [Medline: 33194159]
23. Jain R, Meena ML, Rana KB. Risk factors of musculoskeletal symptoms among mobile device users during work from home. Int J Occup Saf Ergon 2022 Dec;28(4):2262-2268. [doi: 10.1080/10803548.2021.1979318] [Medline: 34514964]
24. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
25. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
26. Gates M, Gates A, Pieper D, Fernandes RM, Tricco AC, Moher D, et al. Reporting guideline for overviews of reviews of healthcare interventions: development of the PRIOR statement. BMJ 2022;378:e070849 [FREE Full text] [doi: 10.1136/bmj-2022-070849] [Medline: 35944924]
27. Chen P, Liu T, Tse MMY, Lai CKY, Tsoh J, Ng SSM. The predictive role of hand section of fugl-meyer assessment and motor activity log in action research arm test in people with stroke. Front Neurol 2022;13:926130 [FREE Full text] [doi: 10.3389/fneur.2022.926130] [Medline: 35873769]
28. Naqvi U, Margetis K, Sherman A. Muscle Strength Grading. Treasure Island (FL): StatPearls Publishing URL: https://www.ncbi.nlm.nih.gov/books/NBK436008/ [accessed 2025-11-03]
29. Freire B, Bochehin do Valle M, Lanferdini FJ, Foschi CVS, Abou L, Pietta-Dias C. Cut-off score of the Modified Ashworth Scale corresponding to walking ability and functional mobility in individuals with chronic stroke. Disabil Rehabil 2023;45(5):866-870. [doi: 10.1080/09638288.2022.2037753] [Medline: 35171052]
30. Hong I, Woo HS, Shim S, Li C, Yoonjeong L, Velozo CA. Equating activities of daily living outcome measures: the functional independence measure and the Korean version of modified barthel index. Disabil Rehabil 2018;40(2):217-224. [doi: 10.1080/09638288.2016.1247468] [Medline: 27930889]
31. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017;358:j4008 [FREE Full text] [doi: 10.1136/bmj.j4008] [Medline: 28935701]

XSL•FO
RenderX

32. Pieper D, Antoine S, Mathes T, Neugebauer EAM, Eikermann M. Systematic review finds overlapping reviews were not mentioned in every other overview. J Clin Epidemiol 2014;67(4):368-375. [doi: 10.1016/j.jclinepi.2013.11.007] [Medline: 24581293]
33. Tseng KC, Wang L, Hsieh C, Wong AM. Portable robots for upper-limb rehabilitation after stroke: a systematic review and meta-analysis. Ann Med 2024;56(1):2337735 [FREE Full text] [doi: 10.1080/07853890.2024.2337735] [Medline: 38640459]
34. Wu J, Cheng H, Zhang J, Yang S, Cai S. Robot-assisted therapy for upper extremity motor impairment after stroke: a systematic review and meta-analysis. Phys Ther 2021;101(4):pzab010. [doi: 10.1093/ptj/pzab010] [Medline: 33454787]
35. Marotta N, Demeco A, Moggio L, Ammendolia A. The adjunct of transcranial direct current stimulation to robot-assisted therapy in upper limb post-stroke treatment. J Med Eng Technol 2021;45(6):494-501. [doi: 10.1080/03091902.2021.1922527] [Medline: 34038313]
36. Ko MJ, Chuang YC, Ou-Yang LJ, Cheng YY, Tsai YL, Lee YC. The application of soft robotic gloves in stroke patients: a systematic review and meta-analysis of randomized controlled trials. Brain Sci 2023;13(6):900 [FREE Full text] [doi: 10.3390/brainsci13060900] [Medline: 37371378]
37. Bazan R, Fonseca BHDS, Miranda JMDA, Nunes HRDC, Bazan SGZ, Luvizutto GJ. Effect of robot-assisted training on unilateral spatial neglect after stroke: systematic review and meta-analysis of randomized controlled trials. Neurorehabil Neural Repair 2022;36(8):545-556. [doi: 10.1177/15459683221110894] [Medline: 35880666]
38. Huo Y, Wang X, Zhao W, Hu H, Li L. Effects of EMG-based robot for upper extremity rehabilitation on post-stroke patients: a systematic review and meta-analysis. Front Physiol 2023;14:1172958 [FREE Full text] [doi: 10.3389/fphys.2023.1172958] [Medline: 37256069]
39. De Iaco L, Veerbeek JM, Ket JCF, Kwakkel G. Upper limb robots for recovery of motor arm function in patients with stroke: a systematic review and meta-analysis. Neurology 2024;103(2):e209495. [doi: 10.1212/WNL.0000000000209495] [Medline: 38870442]
40. Moggio L, de Sire A, Marotta N, Demeco A, Ammendolia A. Exoskeleton end-effector robot-assisted therapy for finger-hand motor recovery in stroke survivors: systematic review and meta-analysis. Top Stroke Rehabil 2022;29(8):539-550. [doi: 10.1080/10749357.2021.1967657] [Medline: 34420498]
41. Zhao M, Wang G, Wang A, Cheng LJ, Lau Y. Robot-assisted distal training improves upper limb dexterity and function after stroke: a systematic review and meta-regression. Neurol Sci 2022;43(3):1641-1657. [doi: 10.1007/s10072-022-05913-3] [Medline: 35089447]
42. Boardsworth K, Rashid U, Olsen S, Rodriguez-Ramirez E, Browne W, Alder G, et al. Upper limb robotic rehabilitation following stroke: a systematic review and meta-analysis investigating efficacy and the influence of device features and program parameters. J Neuroeng Rehabil 2025;22(1):164 [FREE Full text] [doi: 10.1186/s12984-025-01662-4] [Medline: 40671078]
43. Hwang S, Min KC, Song CS. Assistive technology on upper extremity function for stroke patients: a systematic review with meta-analysis. J Hand Ther 2024;37(4):507-519. [doi: 10.1016/j.jht.2023.12.014] [Medline: 38796397]
44. Jin C, Chen Y, Ma Y. Effectiveness of robot-assisted task-oriented training intervention for upper limb and daily living skills in stroke patients: a meta-analysis. PLoS One 2025;20(1):e0316633 [FREE Full text] [doi: 10.1371/journal.pone.0316633] [Medline: 39752454]
45. Su T, Wang M, Chen Z, Feng L. Effect of upper robot-assisted training on upper limb motor, daily life activities, and muscular tone in patients with stroke: a systematic review and meta-analysis. Brain Behav 2024;14(11):e70117 [FREE Full text] [doi: 10.1002/brb3.70117] [Medline: 39482838]
46. Verola S, Ugolini A, Pellicciari L, Di Bari M, Paci M. Clinical relevance of the effects of robotic rehabilitation for upper limb recovery after stroke in randomized studies: a systematic review with meta-analysis. Arch Physiother 2025;15:118-130 [FREE Full text] [doi: 10.33393/aop.2025.3209] [Medline: 40416935]
47. Wang H, Wu X, Li Y, Yu S. Efficacy of robot-assisted training on upper limb motor function after stroke: a systematic review and network meta-analysis. Arch Rehabil Res Clin Transl 2025;7(1):100387 [FREE Full text] [doi: 10.1016/j.arrct.2024.100387] [Medline: 40463769]
48. de Blas-Zamorano P, Montagut-Martínez P, Pérez-Cruzado D, Merchan-Baeza J. Fugl-meyer assessment for upper extremity in stroke: a pyschometric systematic review. J Hand Ther 2025;S0894-1130(25):00057-00052 [FREE Full text] [doi: 10.1016/j.jht.2025.04.004] [Medline: 40533328]
49. Ahmad Puzi A, Sidek SN, Mat Rosly H, Daud N, Md Yusof H. Modified Ashworth Scale (MAS) model based on clinical data measurement towards quantitative evaluation of upper limb spasticity. IOP Conf Ser: Mater Sci Eng 2017;260:012024. [doi: 10.1088/1757-899x/260/1/012024]
50. Wade E, Winstein CJ. Virtual reality and robotics for stroke rehabilitation: where do we go from here? Top Stroke Rehabil 2011;18(6):685-700. [doi: 10.1310/tsr1806-685] [Medline: 22436307]
51. Jung HY. Rehabilitation in subacute and chronic stage after stroke. In: Stroke revisited: Diagnosis and Treatment of Ischemic Stroke. Cham: Springer; 2017:351-360.

XSL•FO
RenderX

52. Qiu X, Rahimzamani A, Wang L, Ren B, Mao Q, Durham T, et al. Inferring causal gene regulatory networks from coupled single-cell expression dynamics using scribe. Cell Syst 2020;10(3):265-274.e11 [FREE Full text] [doi: 10.1016/j.cels.2020.02.003] [Medline: 32135093]
53. Qian Z, Bi Z. Recent development of rehabilitation robots. Advances in Mechanical Engineering 2014;7(2):563062. [doi: 10.1155/2014/563062]
54. Zanatta F, Giardini A, Pierobon A, D'Addario M, Steca P. A systematic review on the usability of robotic and virtual reality devices in neuromotor rehabilitation: patients' and healthcare professionals' perspective. BMC Health Serv Res 2022;22(1):523 [FREE Full text] [doi: 10.1186/s12913-022-07821-w] [Medline: 35443710]
55. Lai TJ, Heggie R, Kamaruzaman HF, Bouttell J, Boyd K. Economic evaluations of robotic-assisted surgery: methods, challenges and opportunities. Appl Health Econ Health Policy 2025;23(1):35-49. [doi: 10.1007/s40258-024-00920-1] [Medline: 39333303]
56. Ali A. Ethical and regulatory considerations in rehabilitation robotics: ensuring safe and equitable care. In: Design and Control of Rehabilitation Robots. Studies in Systems, Decision and Control. Cham: Springer; 2025:399-422.
57. Demofonti A, Carpino G, Zollo L, Johnson MJ. Affordable robotics for upper limb stroke rehabilitation in developing countries: a systematic review. IEEE Trans Med Robot Bionics 2021;3(1):11-20. [doi: 10.1109/tmrb.2021.3054462]
58. Tang Y, Dou B. Cost-effectiveness analysis of robotic surgery in healthcare for older individuals: a systematic review based on randomized controlled trials. Front Public Health 2025;13:1614654 [FREE Full text] [doi: 10.3389/fpubh.2025.1614654] [Medline: 40880918]
59. Chang J, Wu C, Hinton Z, Ryan S, Jiranek W, Bolognesi M, et al. Patient perceptions and interest in robotic-assisted total joint arthroplasty. Arthroplast Today 2024;26:101342 [FREE Full text] [doi: 10.1016/j.artd.2024.101342] [Medline: 38481560]
60. Pagani NR, Moverman MA, Puzzitiello RN, Menendez ME, Barnes CL, Kavolus JJ. Online crowdsourcing to explore public perceptions of robotic-assisted orthopedic surgery. J Arthroplasty 2021;36(6):1887-1894.e3. [doi: 10.1016/j.arth.2021.02.027] [Medline: 33741241]
61. Wu Q, Pei H, Ran X, Chen X, Jiang L, Wei A, et al. Qualitative study on the information needs of patients undergoing da vinci robotic surgery. Clin Nurs Res 2023;32(2):433-440. [doi: 10.1177/10547738221103337] [Medline: 35699515]
62. Muaddi H, Zhao X, Leonardelli GJ, de Mestral C, Nathens A, Stukel TA, et al. Fear of innovation: public's perception of robotic surgery. Surg Endosc 2022;36(8):6076-6083. [doi: 10.1007/s00464-022-09011-5] [Medline: 35059838]
63. Jauniaux B, Anand A, Abbas R, Harji DP. From expectations to experiences: a systematic review of patient and public perspectives on robotic surgery. J Robot Surg 2025;19(1):484. [doi: 10.1007/s11701-025-02649-y] [Medline: 40810856]
64. Cano-de-la-Cuerda R, Blázquez-Fernández A, Marcos-Antón S, Sánchez-Herrera-Baeza P, Fernández-González P, Collado-Vázquez S, et al. Economic cost of rehabilitation with robotic and virtual reality systems in people with neurological disorders: a systematic review. J Clin Med 2024;13(6):1531 [FREE Full text] [doi: 10.3390/jcm13061531] [Medline: 38541755]
65. Ballard M, Montgomery P. Risk of bias in overviews of reviews: a scoping review of methodological guidance and four-item checklist. Res Synth Methods 2017;8(1):92-108. [doi: 10.1002/jrsm.1229] [Medline: 28074553]

Abbreviations

ADL: activity of daily living
CCA: corrected covered area
FMA: Fugl-Meyer Assessment
FMA-UE: Fugl-Meyer Assessment-Upper Extremity
ICF: International Classification of Functioning, Disability and Health
MAS: Modified Ashworth Scale
MD: mean difference
PRIOR: Preferred Reporting Items for Overviews of Reviews
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
RAT: robot-assisted therapy
SMD: standardized mean difference
WMD: weighted mean difference

XSL•FO
RenderX

Edited by S Brini; submitted 20.Jun.2025; peer-reviewed by N Döring, H Wang; comments to author 17.Oct.2025; accepted 02.Feb.2026; published 25.Mar.2026.

Please cite as:
Liu S, Zhang X, Zhou L, Zhang J, Liu J, He C
Robot-Assisted Therapy for Upper Limb Rehabilitation After Stroke: Umbrella Review
J Med Internet Res 2026;28:e79363
URL: https://www.jmir.org/2026/1/e79363
doi:10.2196/79363
PMID:41879816

©Sijia Liu, Xianghong Zhang, Lijun Zhou, Jiashan Zhang, Jialin Liu, Chengqi He. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Technology-Based Interventions for Prevention of Type 2 Diabetes Following Gestational Diabetes: Systematic Review and Meta-Analysis

Claire Eades[1], BSc, MSc, PhD; Anh Nguyen-Hoang[1], BSc, MSc, PhD; Louise Hoyle[1], BNurs, MSc, PhD; Dawn Cameron[2], BSN, MSc, PhD; Josie MM Evans[3], MA, MPH, PhD

[1]Health Sciences, Faculty of Health Sciences and Sport, University of Stirling, Stirling, Stirlingshire, United Kingdom

[2]School of Health Sciences and Life Sciences, University of the West of Scotland, Lanarkshire, South Lanarkshire, United Kingdom

[3]Public Health Scotland, Edinburgh, Lothian, United Kingdom

Corresponding Author:
Claire Eades, BSc, MSc, PhD
Health Sciences
Faculty of Health Sciences and Sport
University of Stirling
Pathfoot Building
Stirling, Stirlingshire, FK9 4LA
United Kingdom
Phone: 44 01786466340
Email: c.e.eades@stir.ac.uk

Abstract

Background: Previous gestational diabetes incurs an 8-fold risk of developing type 2 diabetes, but lifestyle change can prevent or delay progression. Technology-based interventions may help overcome challenges women face in making postpartum lifestyle changes.

Objective: This study aimed to assess whether technology-based diabetes prevention interventions improve outcomes related to the onset of type 2 diabetes among women with a previous diagnosis of gestational diabetes.

Methods: Cochrane Central Register of Controlled Trials, CINAHL, Embase, PsycINFO, and Midwives Information and Resource Service were searched to October 2025 using subject headings and free-text terms. Titles and abstracts were independently screened by 2 authors, as were retrieved full-text articles. Studies were eligible if they examined technology-based diabetes prevention interventions delivered between gestational diabetes diagnosis and any time post partum, assessing anthropometric outcomes, glycemic control, health behavior, or psychological outcomes. Risk of bias was assessed by 1 reviewer using the National Institute for Clinical Excellence checklist, and certainty of evidence was assessed by 2 reviewers using the Grading of Recommendations Assessment, Development, and Evaluation. Data were summarized narratively, and results were pooled, where possible, using a random effects model.

Results: This review identified 15 studies, including 1257 participants. Pooled analysis of 7 studies showed significantly greater weight loss among those receiving technology-based interventions (mean difference –1.01, SE 0.35, 95% CI –1.86 to –0.16 kg; *P*=.03). Interventions delivered using technology only showed increased weight loss (mean difference –1.13, 95% CI –3.12 to 0.86 kg) as did those with a longer follow-up (mean difference –1.58, 95% CI –3.93 to 0.76 kg) compared with combined technology and telemedicine approaches (mean difference –0.89, 95% CI –2.51 to 0.73 kg) and studies with shorter follow-up (mean difference –0.7, 95% CI –1.21 to –0.18 kg), but these differences were not significant (mode of delivery: χ^2_1=0.08; *P*=.78; follow-up: χ^2_1=1.06; *P*=.30). Meta-analysis showed no significant differences in BMI (mean difference –0.22, SE 0.1, 95% CI –0.4 to –0.01 kg/m^2; *P*=.27; n=2 studies), fasting glucose (mean difference –0.03, SE 0.16, 95% CI –0.49 to 0.49 mmol/L; *P*=.99; n=4 studies), 2-hour glucose (mean difference 0.12, SE 0.19, 95% CI –0.47 to 0.72 mmol/L; *P*=.56; n=4 studies), hemoglobin A_{1c}(mean difference –0.01%, SE 0.02%, 95% CI –0.24% to 0.23%; *P*=.74; n=2 studies), or homeostasis model assessment of insulin resistance (mean difference 0.07, SE 0.02, 95% CI –0.16 to 0.31; *P*=.16; n=2). Certainty of evidence for all pooled outcomes was very low.

XSL•FO
RenderX

Conclusions: Technology-based interventions may help support women in reducing their risk of type 2 diabetes following gestational diabetes mellitus, but substantial heterogeneity, significant risk of bias, and very low certainty in the evidence mean that the findings should be interpreted cautiously. Trials with larger samples and longer follow-up are required to draw firm conclusions.

Trial Registration: PROSPERO CRD42024324019; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024324019

(J Med Internet Res 2026;28:e78841) doi:10.2196/78841

KEYWORDS

diabetes prevention; gestational diabetes; health behavior change; mHealth; type 2 diabetes

Introduction

Diabetes is a growing global public health concern and a priority for preventive efforts [1]. The number of adults globally with diabetes in 2017 was estimated to be 451 million and is expected to rise to 693 million by 2045, with type 2 diabetes accounting for over 90% of cases [2]. Women who have had gestational diabetes mellitus (GDM) are at an 8-fold increased risk of developing type 2 diabetes [3]. GDM affects around 5% of pregnant women in Europe and is defined as glucose intolerance that is first diagnosed in pregnancy [4]. Although normal glucose regulation usually returns shortly after delivery, rates of type 2 diabetes after a diagnosis of GDM have been reported to be as high as 70% [5], with the greatest risk thought to be in the first 5 years following a pregnancy with GDM [6]. However, women who make positive lifestyle changes can prevent or delay progression to type 2 diabetes [7].

Challenges that women face in making lifestyle changes in the postpartum period include competing demands from work and home life, and a tendency to underestimate their personal risk and the severity of type 2 diabetes [8-10]. Technology-based interventions offer a possible solution to help to overcome some of these challenges by offering behavior change education and support remotely via messages, smartphone apps, or websites. A recent systematic review of technology-driven diabetes prevention interventions found that around two-thirds of included interventions were effective in achieving shorter-term weight loss and one-third were effective in achieving longer-term weight loss, but the review excluded studies focusing on participants with a previous diagnosis of GDM [11].

Although technology-driven interventions have been shown to be effective in supporting the management of GDM [12], to date, no reviews have been identified that focus on interventions primarily delivered using technology for diabetes prevention in women who have previously had GDM. A systematic review and meta-analysis carried out by Halligan et al [13] investigated the effectiveness of interventions with any technology or telemedicine component for women with previous GDM and found clinically relevant but non–statistically significant improvements in weight and BMI, but currently the effectiveness of primarily technology-based interventions is not known. The review by Halligan et al [13] included interventions with any technology component, and so for many of the included articles, the technology-based intervention was an adjunct to a telemedicine or face-to-face intervention. Telemedicine interventions involve real-time communication with a health care provider via telephone or video call, and although they have some practical benefits over a face-to-face appointment, telemedicine still relies on both parties arranging a convenient time for synchronous contact with a health care provider and so cannot offer the flexible access to advice and help that digital interventions can offer. Interventions delivered primarily using technology may offer the most benefits in terms of flexibility for women in the postpartum period and have the potential to be more cost-effective than telemedicine.

The review by Halligan et al [13] did not assess the behavior change techniques (BCTs) present in the interventions reviewed. BCTs are the active ingredients of an intervention that are designed to promote behavior change [14], and a taxonomy has been developed providing a list of standardized labels and definitions of BCTs. Using the shared language and descriptions provided by the BCT taxonomy allows the content of interventions to be unpicked and described in a consistent and comparable way, which may help to provide insight into the effective components of technology-based interventions. The objective of this review is to assess whether interventions focusing on diabetes prevention (not self-management of GDM), that are primarily technology-based, are effective in improving outcomes related to the onset of type 2 diabetes among women with a previous diagnosis of GDM and to identify and describe the BCTs present in interventions.

Methods

Overview

The systematic review and meta-analysis were conducted and are reported according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses; Multimedia Appendix 1) guidelines [15], PRISMA abstract checklist (Multimedia Appendix 2) and the complementary PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension; Multimedia Appendix 3) for reporting literature searches within systematic reviews [16]. The protocol for the systematic review was registered with the International Prospective Register of Systematic Reviews, and no amendments were made to the protocol [17].

Eligibility Criteria

The inclusion and exclusion criteria used to assess eligibility in the review are presented in Textbox 1.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Adults aged 18 years or older who have received a previous diagnosis of gestational diabetes.
- Studies that examine the use of a technology-based intervention (eg, text, smartphone app, email, video, and website) in the antenatal or postpartum period with the stated aim of preventing or reducing the risk of type 2 diabetes.
- Interventions using both technology-based and face-to-face delivery, where the technology-based delivery is the primary means of intervention delivery and the technology-based intervention could feasibly be delivered as a stand-alone intervention.
- The outcomes of interest are anthropometric (weight, BMI, percent body fat, and waist or hip circumference), glycemic control (hemoglobin A_{1c}, fasting plasma glucose, insulin resistance, prediabetes incidence/prevalence, and type 2 diabetes incidence), behavioral outcomes (diet, physical activity, or sedentary behavior), and psychological outcomes (beliefs about type 2 diabetes, self-efficacy for health behavior change, and behavioral intentions).
- Experimental and nonexperimental study designs, including observational studies, randomized trials, and nonrandomized trials.
- Studies published in English.

Exclusion criteria

- Studies of women who have received a diagnosis of type 1 or type 2 diabetes.
- Studies describing interventions where the technology-based component was only used to supplement a face-to-face intervention or to deliver real-time support with a human coach.
- Studies focusing on the self-management of gestational diabetes.
- Studies reporting only feasibility and acceptability related outcomes.
- Studies using a qualitative design only.

Information Sources and Search Strategy

The databases MEDLINE, Central Register of Controlled Trials, CINAHL, Embase, PsycINFO, and Midwives Information and Resource Service were searched separately via the EBSCO platform in March 2024 using a combination of database subject headings and free-text words (Multimedia Appendix 4). No dedicated peer review of the search strategy was conducted. The search was rerun in October 2025 to check for papers published since the initial search in March 2024. No restrictions were applied for language, type, or date of publication, and no published search filters were used. No online resources, print sources, or citing references of included papers were searched. The reference lists of included papers were checked to identify any other potentially relevant papers, but experts in the field were not contacted due to the time-consuming nature of this process. Duplicates were removed in RefWorks (ProQuest LLC).

Selection Process

The titles and abstracts of all articles were screened by 1 author (CE), and independent screening was split between 3 other authors, with DC screening one-quarter, LH screening one-quarter, and AN-H screening the other half. The full texts of papers were retrieved for studies that were considered relevant, but also for those that contained insufficient information to allow judgment of relevance. These were checked against the inclusion criteria by CE and independently by AN-H, LH, and DC. Where there were disagreements between reviewers about the inclusion of a paper, a consensus was reached through discussion among all authors.

Data Collection Process and Data Items

A data extraction form was developed for this review and piloted on 3 papers, and then refined. Information was extracted by 1 reviewer (CE) under the following headings: authors, study location, type of intervention, mode of delivery, theory or model base for intervention, aim of the study, study design, inclusion criteria, size of the population, baseline characteristics, type of comparator, tailoring/individual patient–based, duration of intervention and measurement, outcome measures, measurement tools, sample size intervention group, sample size control group, dropout, findings summary, baseline outcomes intervention, baseline outcomes control, follow-up outcomes intervention, and follow-up outcomes control. Data were extracted for all available follow-up time points and outcomes available within each study. We attempted to contact authors where data were missing or not reported in the correct format for meta-analysis. If the data were not available or we received no response, studies were excluded from the meta-analysis. A second author (AN-H) independently checked all outcome data, and any errors were discussed and checked against the published paper by both CE and AN-H. The BCT taxonomy (version 1) [14] was used by 1 reviewer (AN-H) to code the BCTs present in the descriptions of included interventions. A second reviewer (CE) independently coded a random selection of 20% of interventions, and any disagreements were resolved by discussion.

Risk of Bias Assessment

Internal and external validity of included studies was assessed by 1 reviewer (CE) using the National Institute for Clinical Excellence [18] checklist for quantitative intervention studies of any design. This checklist contains 27 items addressing 4 specific domains (population, method of allocation to intervention, outcomes, and analyses) and gives an appraisal of internal and external validity using a "++," "+," or "–" rating scale depending on the extent to which criteria were satisfied,

XSL•FO
RenderX

with a "++" indicating lowest risk of bias and "–" highest risk of bias.

Data Analysis

All data were grouped by outcome for synthesis and summarized narratively. Meta-analysis was planned for any outcome of interest recorded at both baseline and follow-up by 2 or more studies. Consistent with the Cochrane Collaboration guidance, only studies with a randomized controlled trial design were included in the meta-analysis [19]. After extraction, the following outcomes were synthesized using meta-analysis: weight, BMI, fasting plasma glucose, 2-hour glucose, hemoglobin A_{1c} (HbA_{1c}), and homeostasis model assessment of insulin resistance (HOMA-IR). Results were pooled in SPSS (version 28; IBM Corp) using a random effects model to analyze the between-group differences in each outcome at follow-up. Forest plots were used to display individual study results and the syntheses. When there were follow-up data available at multiple time points, data from the last follow-up time point were used in the analysis. Where there were multiple publications reporting on the same sample, the data were only included in the analyses once to avoid double-counting of participants. When SD values were not reported, these were calculated from 95% CIs when possible. All pooled outcomes were continuous and were summarized using unstandardized mean difference with Sidik-Jonkman estimator and Knapp-Hartung SE adjustment as recommended for meta-analyses of a small number of studies [20]. Two-sided *P* values and 95% CIs were calculated for each outcome. Heterogeneity was quantified using between-study variance τ^2. The I^2 statistic is reported but interpreted with caution [21]. Prediction intervals were not calculated as the number of studies in the analyses fell below the recommended minimum number [22]. Analysis of publication bias was not conducted due to the small number of studies eligible for inclusion. Subgroup analyses were planned for the mode of intervention delivery (entirely digital or mixed) and length of follow-up (less than 6 months versus 6 months or more).

Certainty Assessment

Certainty of evidence was evaluated using the Grading of Recommendations Assessment, Development, and Evaluation (GRADE; [23]) criteria for the main outcomes (weight, fasting glucose, 2-hour glucose, HbA_{1c}, or HOMA-IR). GRADE classifies evidence into 4 levels: high, moderate, low, and very low based on assessment of risk of bias, inconsistency, indirectness, imprecision, and publication bias. With evidence rated as high quality, we can have very high certainty that the true effect is close to the estimated effect, whereas in evidence rated as very low, we can have almost no certainty in the estimates reported. Two reviewers (CE and LH) independently rated studies included in the meta-analysis using GRADEpro, and any disagreements were resolved by discussion. The table summarizing the results of the assessment can be found in Multimedia Appendix 5.

Results

Study Selection

Figure 1 shows a PRISMA flow diagram of studies identified by the search. A total of 4704 papers were returned by the search, of which 4582 were excluded at the title and abstract screening stage. We reviewed the full text of 122 papers, and 15 studies [24-38] reported in 17 papers [39,40] were deemed eligible for inclusion in the systematic review. At the full-text screening stage, the following exclusions were made: 26 nondigital interventions, 23 protocols, 16 interventions where the technology-based component was a supplement to face-to-face, 12 interventions for the management or prevention of gestational diabetes, 11 papers that did not report relevant outcomes, 5 duplicates, 4 review papers, 2 papers not published in English, 2 papers that included women with additional cardiometabolic risk factors in the sample, and 1 conference abstract. A total of 4 conference abstracts were reviewed that did not contain enough information to assess eligibility. We searched the 4704 abstracts identified by our review by author name and intervention name, as outlined in the conference publications, to identify any further publications relating to these interventions. For 3 of the 4 conference abstracts, further publications with more information were identified that allowed assessment of eligibility. There were no related papers found for the fourth conference abstract, and no email addresses could be found for the authors through web searches, so the study was excluded from the review [41].

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

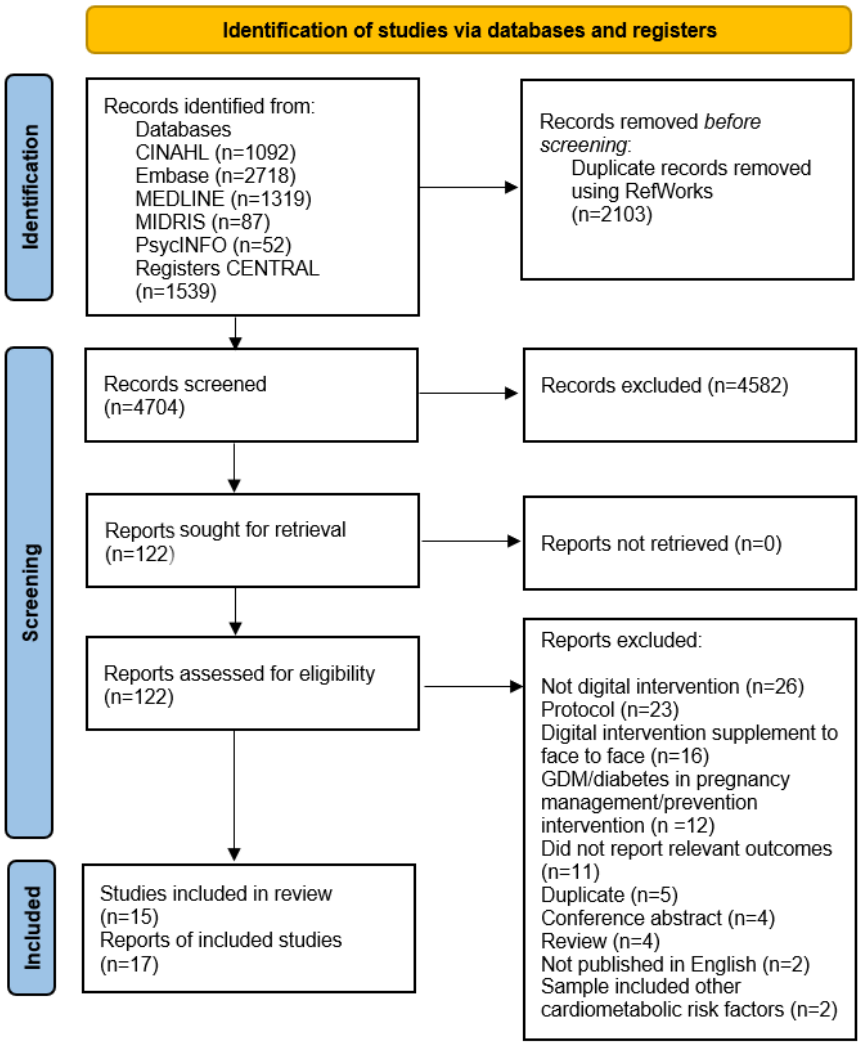

Study Characteristics

The characteristics of the 15 included studies are presented in Table 1. Studies were published between 2012 and 2024, with almost two-thirds published between 2020 and 2024 (n=10, 66%). They were conducted in the United States (n=6) [28,31,32,34,36,37], Australia (n=4) [24,25,35,38], Asia (n=4) [26,28-30], and Europe (n=1) [33]. The mean age of participants among studies that reported age (n=13) [24-26,29-38] ranged from 31.5 to 40.4 years, with most means falling between 31 and 34 years (n=9, 69.2%). Study design was primarily randomized controlled trial (n=11) [24,25,27,29,30,32-36,38], with 2 studies adopting a before-and-after single-arm design [31,37] and 2 a nonrandomized controlled trial design [26,28].

XSL•FO
RenderX

Table 1. Summary of the characteristics of included studies.

Author (year), location	Enrolment setting	Design	Definition GDM[a] used	IG[b]/CG[c] sample size, n	Attrition IG/CG, %	Outcomes assessed	Follow-up duration	Risk of Bias	Age in IG (years), mean (SD)	Age in CG (years), mean (SD)	BMI in IG (years), mean (SD)	BMI in CG (years), mean (SD)
Cheung (2019), Sydney, Australia [24]	1 tertiary care hospital	RCT[d]	FPG[e] ≥ 5.5 mmol/L and/or a 2hPG[f] ≥ 8 mmol/L on 75 g OGTT[g]	40/20	28/35	Weight, PA[h], diet	8.5 months	++/+	34 (4)	34 (4)	NR[i]	NR
Cheung (2024), Sydney, Australia [25]	3 tertiary care hospitals	RCT	FPG ≥ 5.5 mmol/L and/or a 2hPG ≥ 8 mmol/L on 75 g OGTT (at 2 hospitals) or one of the following: FPG ≥ 7.0 mmol/L; 2hPG ≥ 11.1 mmol/L on 75 g OGTT; random plasma glucose ≥ 11.1 mmol/L (at 1 hospital)	88/89	7/12	Weight, PA, diet	6 months	++/+	32.2 (4.6)	32.2 (4.6)	Prepregnancy: 29.4 (6.9)	Prepregnancy: 28.4 (5.9)
Ghaderi (2019), Tehran, Iran [26]	1 treatment center	Quasi-experimental with control group	NR	45/45	2/4	Risk perceptions for T2D[j]	NR	++/+	33.66 (3.93)	32.55 (4.44)	Prepregnancy: 26.44 (1.63)	Prepregnancy: 27.09 (2.13)
Kim (2012), Michigan, United States [27]	University health system (large nonprofit managed care plan with several private practices)	RCT	NR	28/21	7/23	Weight, WC[k], FPG, 2hPG, PA, self-efficacy for diet and PA, risk perceptions type 2 diabetes	3 months	++/++	NR	NR	29.8 (6.8)	29.8 (6.8)

XSL•FO
RenderX

Author (year), location	Enrolment setting	Design	Definition GDM[a] used	IG[b]/CG[c] sample size, n	Attrition IG/CG, %	Outcomes assessed	Follow-up duration	Risk of Bias	Age in IG (years), mean (SD)	Age in CG (years), mean (SD)	BMI in IG (years), mean (SD)	BMI in CG (years), mean (SD)
Kim (2021), South Korea [28]	3 hospitals	Quasi-experimental with control group	NR	64/64	8/3	Weight, body fat, FPG, HbA_{1c}[l], diet, health-promoting lifestyle profile, T2D knowledge	3 months	+/+	Over 35 45.6%	Over 35 50%	NR	NR
Liew (2023), Singapore [29]	NR	RCT	NR	31/30	10/6	Weight, BMI, FPG, 2hPG	2.5 months	++/+	36.5 (3.7)	35.4 (3.7)	24.0 (4.4)	22.7 (4.0)
Lim (2021), Singapore [30]	1 hospital	RCT	FPG ≥ 5.1 mmol/L and/or 1hPG[m] ≥ 10.0 and/or 2hPG ≥ 8.5 mmol/L on 75 g OGTT	101/99	5/6	Weight, WC, diet, FPG, 2hPG, health directed behavior	4 months	++/+	32.6 (4.5)	32.4 (4.2)	Median 26.9 (IQR 23.0–29.5)	Median 25.8 (IQR 22.8–28.6)
Nicholson (2016), United States [31]	2 university-based prenatal clinics and 1 community-based practice	Single-arm pre-post study	NR	16/none	30	Weight, HbA_{1c}, self-efficacy for PA	NR	+/++	31.5 (4.7)	NR	Prepregnancy: 29.4 (11.2)	NR
Nicklas (2014 [25], 2019 [39], 2020 [40]), Boston, United States	1 hospital	RCT	Meets ≥ 2 following 100 g OGTT: FPG ≥ 5.3 mmol/L, 1hPG ≥ 10.0 mmol/L, 2hPG ≥ 8.6 mmol/L, 3hPG[n] ≥ 7.8 mmol/L. Or medical record documented clinician diagnosis	36/39	17/13	Weight, BMI, WC, FPG, 2hPG, HbA_{1c}, HOMA-IR[o], IGT[p] and T2D prevalence, diet, PA	12 months	++/+	33.6 (4.8)	33.3 (5.8)	31.2 (5.8)	31.6 (5.5)

XSL•FO
RenderX

Author (year), location	Enrolment setting	Design	Definition GDM[a] used	IG[b]/CG[c] sample size, n	Attrition IG/CG, %	Outcomes assessed	Follow-up duration	Risk of Bias	Age in IG (years), mean (SD)	Age in CG (years), mean (SD)	BMI in IG (years), mean (SD)	BMI in CG (years), mean (SD)
Potzel (2022), Germany [33]	University medical centre in Munich, Germany; Diabetes center in Dusseldorf; Institute for diabetes research in Tubingen	RCT	FPG ≥ 5.1 mmol/L and/or a 2hPG ≥ 8.5 mmol/L on 75 g OGTT	33/33	18/15	Weight, body fat, IGT prevalence, PA, diet, change in health habits	6 months	++/+	37.0 (4.4)	35.7 (3.1)	Median 26 (IQR 22.6–28.9)	Median 27.3 (IQR 23.9–30.3)
Reutrakul (2022), Chicago, United States [34]	NR	RCT	NR	9/6	0/17	Sleep duration, PA, FPG, 2hPG, area under the curve for glucose, HOMA-IR	1.5 months	+/++	42.0 (2.9)	38.7 (6.0)	32.7 (5.3)	32.6 (4.6)
Rollo (2020), New South Wales, Australia [35]	NR	3-arm RCT	NR	28 (15 low personalization [HP] and 13 low personalization [LP])/14	29/21	Weight, BMI, WC, body fat, HbA_{1c}, diet quality, PA	6 months	++/++	HP[q]: 34 (4.5); LP[r]: 32.8 (3.6)	33.6 (3.8)	HP: 33.9 (3.6), LP: 33.9 (3.6)	31.1 (4.8)
Saxon (2023), Boston and Colorado, United States [36]	NR	RCT	Meets ≥ 2 following 100 g OGTT: FPG ≥ 5.3 mmol/L, 1hPG ≥ 10.0 mmol/L, 2hPG ≥ 8.6 mmol/L, 3hPG ≥ 7.8 mmol/L	91/90	48/49	Diet and PA self-efficacy, readiness to change	24 months	-/-	32.5 (4.9)	32.2 (5.5)	31.4 (5.9)	31.3 (5.8)
Seely (2020), Boston, United States [37]	Single federally qualified community health center	Single-arm pre-post study	NR	21/none	14	Weight, self-efficacy for PA and healthy eating	1.5 months	-/++	Whole sample: 33 (6.9)	NR	NR	NR

XSL•FO
RenderX

Author (year), location	Enrolment setting	Design	Definition GDM[a] used	IG[b]/CG[c] sample size, n	Attrition IG/CG, %	Outcomes assessed	Follow-up duration	Risk of Bias	Age in IG (years), mean (SD)	Age in CG (years), mean (SD)	BMI in IG (years), mean (SD)	BMI in CG (years), mean (SD)
Taylor (2022), New South Wales, Australia [38]	NR	3-arm RCT	NR	48 (HP 25, medium personalization [MP] 23)/28	16/4	Weight, diet quality, PA, self-efficacy weight management	3 months	+/++	HP: 36.0 (5.4); MP[s]: 34.6 (5.3)	37.3 (5.3)	30.7 (4.3)	30.0 (4.8)

[a]GDM: gestational diabetes mellitus.

[b]IG: intervention group.

[c]CG: control group.

[d]RCT: randomized controlled trial.

[e]FPG: fasting plasma glucose.

[f]2hPG: 2-hour plasma glucose.

[g]OGTT: oral glucose tolerance test.

[h]PA: physical activity.

[i]NR: not reported.

[j]T2D: type 2 diabetes.

[k]WC: waist circumference.

[l]HbA1c: hemoglobin A_{1c}.

[m]1hPG: 1-hour plasma glucose.

[n]3hPG: 3-hour plasma glucose.

[o]HOMA-IR: homeostatic model assessment of insulin resistance.

[p]IGT: impaired glucose tolerance.

[q]HP: high personalization.

[r]LP: low personalization.

[s]MP: medium personalization.

Intervention Characteristics

Table 2 summarizes the characteristics of the interventions reported in the 15 included studies. Among the included studies, 2 [36,38] were full trials of interventions piloted in other studies included in the review [32,35]. Most interventions targeted diet and physical activity (n=12) [24,25,27-33,35,37,38]. Among the studies that did not target diet and physical activity, 1 targeted risk perceptions for type 2 diabetes [26], 1 addressed sleep [34], and 1 targeted self-efficacy for diet and physical activity [36]. Of the 12 studies targeting diet and physical activity behaviors, 7 also targeted additional behaviors or factors in the intervention, including breastfeeding [25,32], risk perceptions [27,31], self-efficacy [27], parenting [25], infant health/first aid [25,28], stress [28,29], sleep [29,33], and emotions [33]. A total of 4 studies stated that they were informed by 1 or more theories [25,31,33,37], with social cognitive theory the most frequently used (n=3). Among trials with a control arm (n=11), control participants were provided with information about gestational diabetes or diabetes prevention in 5 studies [24,28,32,33,36], received routine care in 4 studies [25-27,30], received the same app/website/devices as the intervention group but without messaging or coaching 2 studies [29,38], were on a waitlist for the intervention in 1 study [35], and received weekly emails with general health education and weekly brief telephone contact from a coach in 1 study [34].

XSL•FO
RenderX

Table 2. Summary of intervention characteristics.

Author (year)	Intervention name; description	Intervention duration	Time period intervention started	Intervention target	Mode of delivery	Control	Theoretical basis	Any tailoring
Cheung (2019) [24]	Smart Mums with Smart Phones; Lifestyle advice, education and support delivered via text messages and counseling, including feedback from activity monitor	8.5 months	Birth of baby	Diet, PA[a]	Text messaging (3 per week), provision of activity monitor, 2 × 30 min diet counseling sessions (1 remote and 1 face-to-face)	Leaflet	NR[b]	Text messages customized according to activity monitor data, breastfeeding status and ethnicity
Cheung (2024) [25]	Smart Mums with Smart Phones 2; Healthy lifestyle promotion delivered via text messages including feedback from activity monitor	6 months	Birth of baby	Diet, PA, breastfeeding, infant health, parenting	Text messaging (4 per week), activity monitor	Routine care	TRA[c], SCT[d], HBM[e]	Text messages customized to breastfeeding status and activity monitor data
Ghaderi (2019) [26]	NR; Education on risk of type 2 diabetes	NR	NR	Risk perceptions	Smartphone app	Routine care	NR	NR
Kim (2012) [27]	NR; Structured, web-based pedometer program giving personalized goals and education regarding lifestyle modification	3 months	Within 3 years of GDM diagnosis	Diet, PA, risk perceptions, self-efficacy	Website, text messaging, internet forum	Routine care	NR	Customized step count goals
Kim (2021) [28]	NR; Virtual reality program where various exercises could be viewed and performed and diet recorded	3 months	Birth of baby	Diet, PA, stress, infant first aid	Smartphone, virtual reality application and headset	Written educational materials on GDM[f]	NR	Adequacy of diet and calories were presented based upon BMI
Liew (2023) [29]	NR; Self-monitoring intervention using health tracker. Data from tracker reviewed by health coaches and personalized recommendations delivered by coaches	1 month	Within 10 years of GDM diagnosis	Diet, PA, sleep, stress	Smartphone app, continuous glucose monitor, health tracking ring, 2 sessions of chatroom messaging with health coach	Same app & devices as IG[g] but no messaging	NR	Customized recommendations for diet, PA, sleep and stress
Lim (2021) [30]	SPAROW; App offering short videos on lifestyle, diet and activity tracking, personalized goals and live chat function	4 months	Birth of baby	Diet, PA	Smartphone app, live chat with health care professionals (5 to 10 min a day for first week then 2 to 3 times a week)	Routine care	NR	Customized weight and activity goals
Nicholson (2016) [31]	GooDMomS; Web-based lessons on healthy lifestyle, weight tracking and feedback on weight	7.5 months	From diagnosis of GDM	Diet, PA, risk perceptions	Website, 4 live webinars, message board, text messages/emails	NR	SCT	Customized feedback about weight, PA and diet

XSL•FO
RenderX

Author (year)	Intervention name; description	Intervention duration	Time period intervention started	Intervention target	Mode of delivery	Control	Theoretical basis	Any tailoring
Nicklas (2014) [32]; (2019) [39]; (2020) [40]	Balance after Baby; Lifestyle modification adapted from DPP[h] delivered via website and coach. Goals, weight and physical activity recorded on website	12 months	6 weeks post partum	Diet, PA, breastfeeding	Website, email and weekly telephone coaching for 12 weeks then every second week for 12 weeks and monthly thereafter	Leaflet	NR	NR
Potzel (2022) [33]	TRIANGLE; Self-paced DPP-based intervention with individualized challenges, chat coaching and behavior change technique library	6 months	3–18 months following GDM diagnosis	Diet, PA, sleep, emotions	Smartphone app, chat-based coaching	Leaflet	Kok's taxonomy of behavior change methods	Customized and interactive challenges
Retrakul (2022) [34]	Sleep-Extend; Sleep tracker and app, didactic content and telephone coaching aimed at improving sleep	1.5 months	NR	Sleep	Smartphone app, wearable sleep tracker, email, weekly telephone coaching (5 to 20 min each)	Weekly health education emails (eg, vaccination for adults, cancer screening) & weekly brief telephone contact from the coach	NR	During coaching feedback given on sleep data, progress, and barriers explored and goals set
Rollo (2020) [35]	Body Balance Beyond; Website-based content on health lifestyles, goal setting, feedback on diet and activity from coaches, and text messaging support	6 months	Within 2 years of GDM diagnosis	Diet, PA	Website, text messaging, 6 coaching video calls (20–30 min each) with dietician and exercise physiologist	Waitlist control	NR	Coaching sessions reviewed diet and PA, explored barriers to diet and PA, refined goals and personalized strategies for achieving goals
Saxon (2020) [36]	Balance after Baby; See Nicklas 2014 description	12 months	6 weeks post partum	Self-efficacy and readiness to change for PA and diet	Website and weekly telephone, email and text messaging coaching for 12 weeks then every second week for 12 weeks and monthly thereafter	Access to websites with information about diabetes prevention	NR	NR
Seely (2020) [37]	Hola Bebé, Adiós Diabetes; Audio visual modules on healthy eating and physical activity delivered via app	2 months	Within 5 years of GDM diagnosis	Diet, PA	Smartphone app	NR	SCT	NR

XSL•FO
RenderX

Author (year)	Intervention name; description	Intervention duration	Time period intervention started	Intervention target	Mode of delivery	Control	Theoretical basis	Any tailoring
Taylor (2022) [38]	Body Balance Beyond; See Rollo 2020 description	3 months	Within 5 years of GDM diagnosis	Diet, PA	Website, text messaging, 5 coaching video calls (20–30 min each) with a dietician or exercise physiologist	Access to intervention website only	NR	Coaching sessions reviewed diet and PA, explored barriers to diet and PA, set goals and strategies for achieving goals

[a]PA: physical activity.

[b]NR: not reported.

[c]TRA: theory of reasoned action.

[d]SCT: social cognitive theory.

[e]HBM: health belief model.

[f]GDM: gestational diabetes mellitus.

[g]IG: intervention group.

[h]DPP: diabetes prevention program.

The duration of interventions ranged from 1 to 12 months. In most studies, the intervention was carried out in the time period between diagnosis of GDM and 18 months post partum (8/13, 60% of studies reporting when the intervention was delivered) [24,25,28,30-33,36]. A total of 4 studies offered the intervention to women between a maximum of 2 and 5 years following diagnosis of GDM [27,35,37,38], and 1 study offered the intervention up to 10 years following diagnosis [29]. The technology-based modes of delivery used included smartphone apps [26,28-30,33,34,37], SMS text messaging [24,25,27,31,35,36,38], emails [31,32,34,36], websites [27,31,32,35,36,38], online message boards [27,31], virtual reality headset [28], activity/sleep trackers [24,25,34], webinars [31], and video calls [35,38], with most studies used multiple technology-based methods of delivery (n=11). A total of 7 studies used only technology-based modes of intervention delivery [25-30] and 8 used technology-based delivery plus synchronous contact with a health coach/interventionist through live chats [33], telephone calls [32,34,36], video calls [35,38], a combination of face-to-face and remote methods [24], or via webinars [31].

Risk of Bias and Certainty of Evidence

A summary of the quality assessments can be found in Table 1. A total of 9 studies achieved a high-quality rating for internal validity, indicating that they minimized bias across multiple criteria [24-27,29,30,32,33,35]. Additionally, 6 studies received a high-quality rating for external validity, indicating that they minimized risk of bias in the generalizability of the study population across multiple criteria [27,31,34,35,37,38]. The GRADE assessment indicated that the certainty of evidence was very low for all pooled outcomes (weight, fasting plasma glucose, 2-hour plasma glucose, HbA_{1c}, and HOMA-IR; Table 3).

XSL•FO
RenderX

Table 3. Grading of Recommendations Assessment, Development, and Evaluation assessment.

Outcome and certainty assessment							Patients, n		Effect		Certainty	Importance
Studies, n	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Technology-based interventions	Nontechnology-based interventions	Relative (95% CI)	Absolute (95% CI)		
Weight (kg; follow-up: range 10-52 weeks)												
7	Randomized trials	Serious[a,b]	Not serious	Serious[c]	Serious[d]	None	181	159	NR[e]	MD[f] 1.01 SD lower (1.86-0.16 lower)	very low[a,b,c,d]	Critical
BMI (kg/m^2; follow-up: range 10-13 weeks)												
2	Randomized trials	Serious[a,b]	Not serious	Serious[c,g]	Serious[d]	None	47	51	NR	MD 0.22 SD lower (1.48-1.03 higher)	very low[a,b,c,d,g]	Critical
Fasting plasma glucose (follow-up: range 10-52 weeks)												
4	Randomized trials	Serious[a,b]	Not serious	Serious[c]	Serious[d]	None	96	91	NR	MD 0 SD (0.5-0.49 higher)	very low[a,b,c,d]	Critical
2-hour plasma glucose (follow-up: range 10-52 weeks)												
4	Randomized trials	Serious[a,b]	Not serious	Serious[c]	Serious[d]	None	96	91	NR	MD 0.12 SD higher (0.47-0.72 higher)	very low[a,b,c,d]	Critical
Hemoglobin A_{1c} (follow-up: range 26-52 weeks)												
2	Randomized trials	Serious[a,b]	Not serious	Serious[c]	Serious[d]	None	54	53	NR	MD 0.01 SD lower (0.24-0.23 higher)	very low[a,b,c,d]	Critical
HOMA-IR[h] (follow-up: range 6-52 weeks)												
2	Randomized trials	Serious[a,b]	Not serious	Serious[c]	Serious[d]	None	48	44	NR	MD 0.07 SD higher (0.16-0.31 higher)	very low[a,b,c,d]	Critical

[a]The nature of the technological interventions in this review (websites and smartphone apps) makes blinding difficult to achieve. A few studies managed

XSL•FO
RenderX

to blind researchers and outcome assessors, but most studies did not blind participants or researchers or did not report in detail.

[b]Some studies did not adequately describe allocation concealment.

[c]Most studies are small, single region, pilot studies using surrogate/short-term outcomes which limits the directness of the evidence and applicability to the broader population of women with previous gestational diabetes.

[d]All studies have very small sample sizes of <400 and wide CIs.

[e]NR: not reported.

[f]MD: mean difference.

[e]Variation in interventions is present.

[h]HOMA-IR: homeostatic model assessment of insulin resistance.

Anthropometric Outcomes

Weight was reported as an outcome in 12 studies [24,25,27-33,35,37,38], waist circumference in 4 [27,30,32,35], body composition in 3 [28,33,35], and BMI in 2 studies [27,29]. A total of 5 studies could not be included in the meta-analysis for weight: 2 were single-arm studies [31,37], 1 reported the proportion of women who met weight goals [25], 1 did not report SDs, and they could not be calculated [30], and 1 was a nonrandomized controlled trial [28]. Of the 5 studies that could not be pooled, only 1 [28] reported findings that favored the intervention in relation to weight loss.

Pooled analysis showed significantly greater weight loss in the intervention group compared with the control (mean difference –1.01, SE 0.35, 95% CI –1.86 to –0.16 kg; *P*=.03; n=7; GRADE very low; Figure 2 [24,27,29,32,33,35,38]). τ^2=0.59; τ=0.77; and I^2=69% indicated higher heterogeneity. The effect did not remain significant across different modes of delivery and lengths of follow-up in subgroup analyses, although there was a larger reduction in weight noted among studies using technology-based delivery only (mean difference –1.13, 95% CI –3.12 to 0.86 kg for technology-based only, mean difference –0.89, 95% CI –2.51 to 0.73 kg for mixed delivery; Figure 2) and with a longer follow-up duration (mean difference –0.70, 95% CI –1.21 to –0.18 kg for follow-up <6 months; mean difference –1.58, 95% CI –3.93 to 0.76 kg for 6-12 months; Figure 3 [24,27,29,32,33,35,38]) but between-group differences were not significant for mode of delivery (χ^2_1=0.08; *P*=.78) or follow-up duration (χ^2_1=1.06; *P*=.30).

Figure 2. Random effects meta-analysis of the mean difference in weight (kg) comparing the intervention group and the control group according to the mode of delivery (digital only vs mixed) [24,27,29,32,33,35,38].

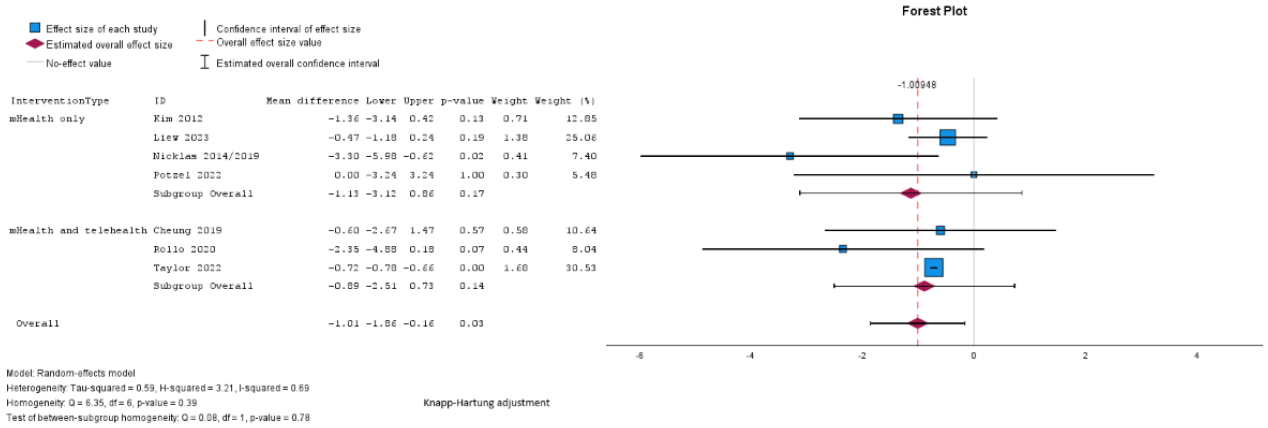

Figure 3. Random effects meta-analysis of the mean difference in weight (kg) comparing the intervention group and the control group according to the length of follow-up [24,27,29,32,33,35,38].

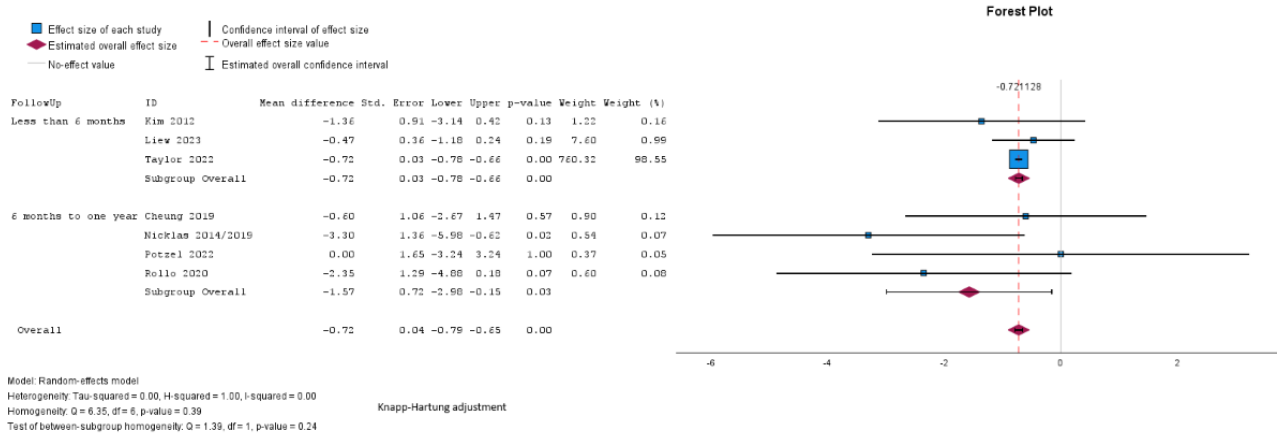

XSL•FO
RenderX

Pooled analysis showed a reduction in BMI in the intervention group compared to the control group, but the difference was not statistically significant (mean difference –0.22, SE 0.1, 95% CI –0.4 to –0.01 kg/m^2; *P*=.27; n=2 [27,29]; GRADE very low; Figure 4 [27,29]). The I^2=41%; τ=0.1; and τ^2=0.01 indicated lower heterogeneity. Subgroup analyses were not possible for BMI because of the small number of studies assessing these outcomes and the uneven distribution of studies across the different subgroups.

Figure 4. Random effects meta-analysis of the mean difference in BMI comparing the intervention group and the control group [27,29].

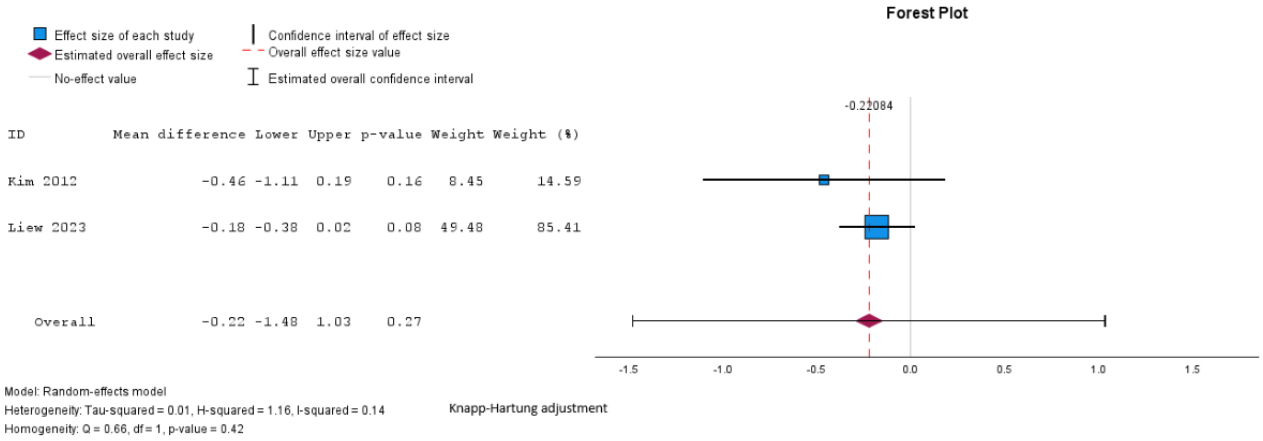

None of the 4 studies assessing waist circumference reported significant differences between the intervention and control groups [27,30,32,35]. Of the 3 studies measuring body composition [28,33,35], only 1 found a significant reduction in body fat percentage in the intervention group compared to the control group [28]. There were too few studies reporting data on waist circumference and body composition at both baseline and follow-up to allow for meta-analysis of these outcomes.

Glycemic Outcomes

Fasting glucose levels were reported in 5 studies [27-29,32,34], 2-hour glucose following an oral glucose tolerance test in 5 [27,29,30,34,39], HbA_{1c} in 4 [28,31,35,39], and HOMA-IR in 2 studies [34,39]. Only 1 study found a significant difference in glycemic outcomes, reporting lower fasting glucose and HbA_{1c} in the intervention group compared to the control group [28]. One study assessed the prevalence of impaired glucose tolerance [33] and another reported both impaired glucose tolerance and type 2 diabetes prevalence [32], with neither finding significant differences in prevalence among women receiving the intervention compared to the control group. Three studies reporting glycemic outcomes could not be pooled: SDs were not reported and could not be calculated for fasting and 2-hour glucose in 1 study [30], 1 study assessing fasting glucose and HbA_{1c} was a nonrandomized controlled trial [28], and another reporting HbA_{1c} was a single-arm design [31]. Subgroup analyses were not possible for glycemic outcomes because of the small number of studies assessing these outcomes and the uneven distribution of studies across the different subgroups of mode of delivery and length of follow-up.

Pooled analysis showed no significant difference between groups in fasting glucose (mean difference –0.03, SE 0.16, 95% CI –0.49 to 0.49 mmol/L; *P*=.99; n=4; GRADE very low; Figure 5 [27,29,32,34]), 2-hour glucose (mean difference 0.12, SE 0.19, 95% CI –0.47 to 0.72 mmol/L; *P*=.56; n=4; GRADE very low; Figure 6 [27,29,32,34]), HbA_{1c} (mean difference –0.01%, SE 0.02%, 95% CI –0.24% to 0.23%; *P*=.74; n=2 studies; GRADE very low; Figure 7 [32,35]), or HOMA-IR (mean difference 0.07, SE 0.02, 95% CI –0.16 to 0.31; *P*=.16; n=2; GRADE very low; Figure 8 [32,34]). The τ=0.24; τ^2=0.06; and I^2=55% indicate higher heterogeneity among the studies assessing fasting plasma glucose, and 2-hour glucose (I^2=42%; τ=0.58; τ^2=0.34), and lower heterogeneity in the studies assessing HbA_{1c} (I^2=31%; τ=0.02; τ^2=0.0004) and HOMA-IR (I^2=0%; τ=0.001; τ^2=0.000002).

XSL•FO
RenderX

Figure 5. Random effects meta-analysis of the mean difference in fasting glucose comparing the intervention group and the control group [27,29,32,34].

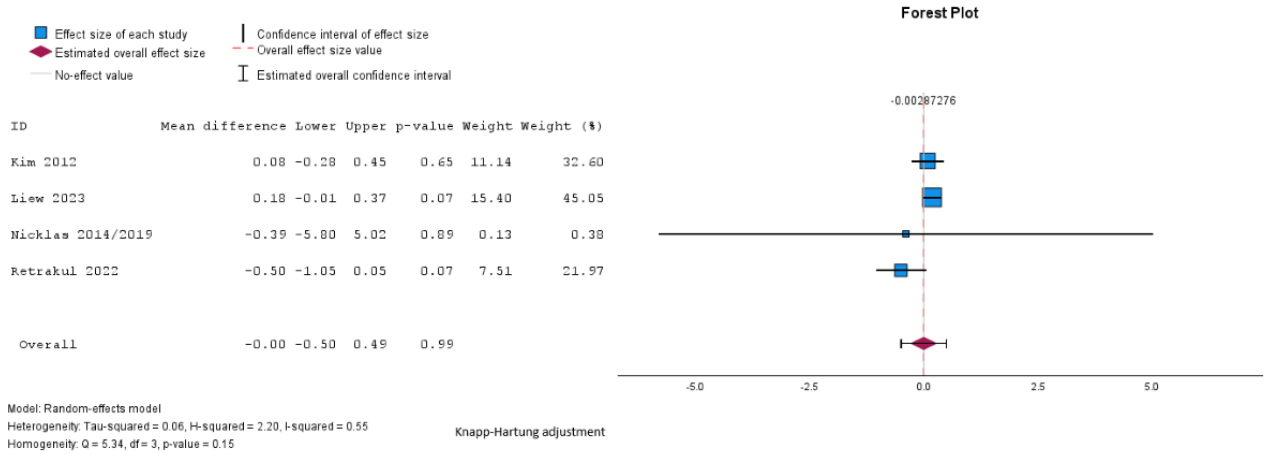

Figure 6. Random effects meta-analysis of the mean difference in 2-hour glucose comparing the intervention group and the control group [27,29,32,34].

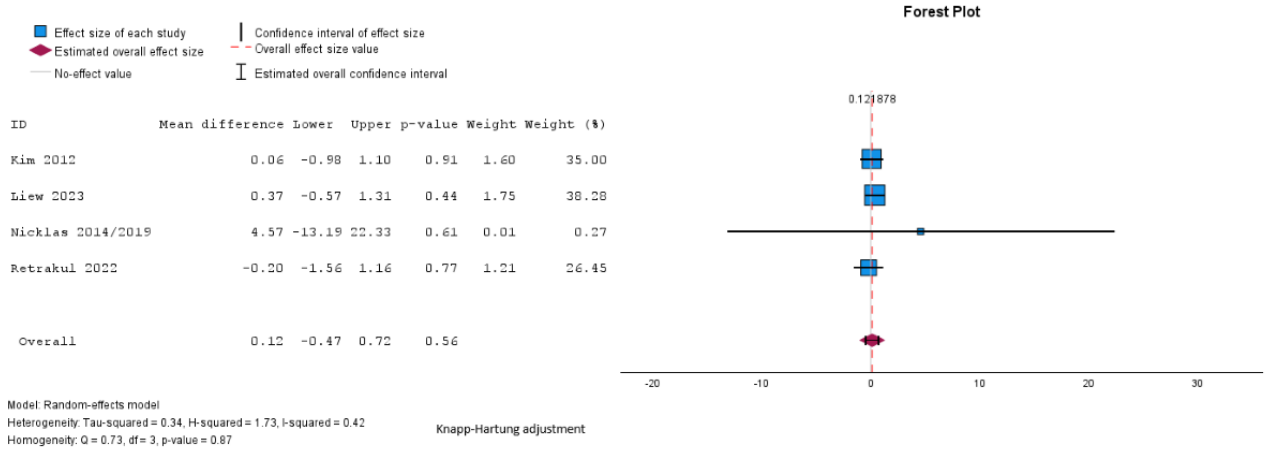

Figure 7. Random effects meta-analysis of the mean difference in hemoglobin A1c comparing the intervention group and the control group [32,35].

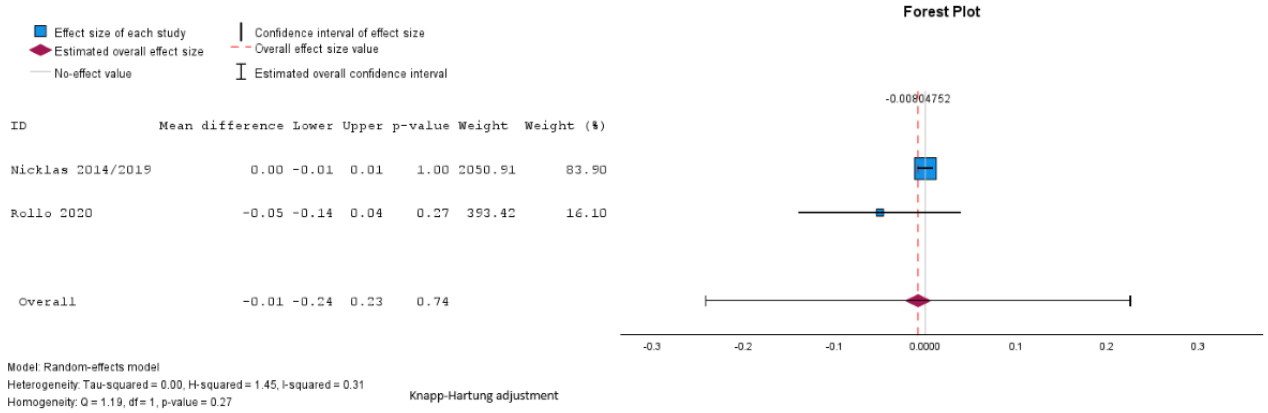

XSL•FO
RenderX

Figure 8. Random effects meta-analysis of the mean difference in homeostasis model assessment of insulin resistance comparing the intervention group and the control group [32,34].

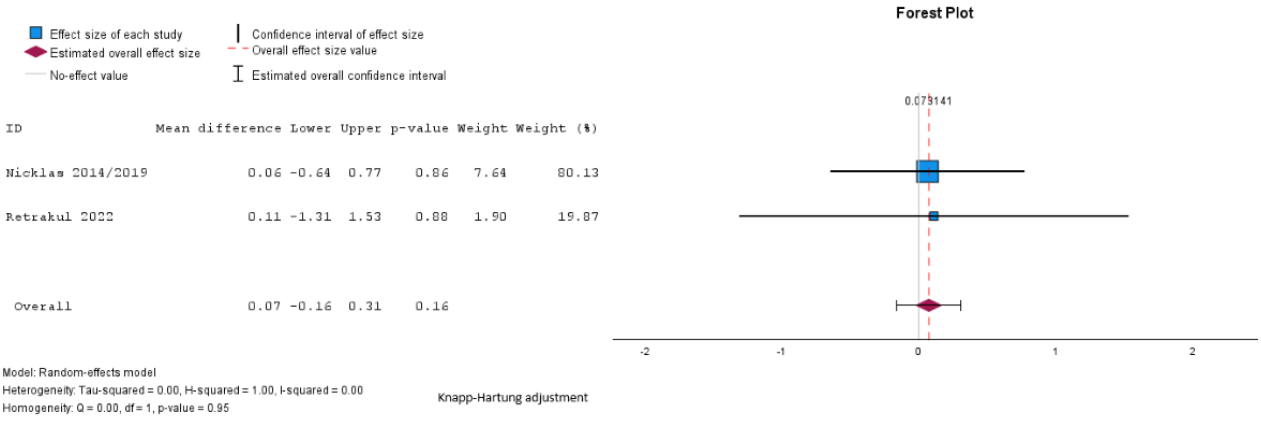

Behavioral Outcomes

A total of 8 studies reported diet outcomes [24,25,28,30,33,35,38,42], and 8 reported physical activity outcomes [24,25,27,32-35,38]. Only 1 study, testing an intervention aimed at improving sleep quality, reported any significant improvements in physical activity among women in the intervention group in comparison to the control group [34]. Two studies reported a significantly reduced calorie intake among those receiving the intervention compared to the control group [30,32], and 1 reported significantly more positive dietary habits among intervention participants [28]. Two studies reported on general measures of health-promoting lifestyles [28,30] and both found that the intervention group had significantly higher scores compared to the control group, indicating more healthy behaviors overall. Similarly, significantly more women in the intervention group of the study by Potzel et al [33] reported making changes to their health-related habits compared to the control group. It was not possible to synthesize behavioral outcomes using meta-analysis due to the variability in how outcomes were measured and reported.

Psychological Outcomes

A total of 4 studies measured self-efficacy for physical activity [27,31,36,37], 3 self-efficacy for diet [27,36,37], and 1 study measured self-efficacy for weight management [38]. Two studies found significant differences in self-efficacy for physical activity [36,37], but in 1, the differences were accounted for by the fact that the control group's self-efficacy worsened, whereas the intervention group remained stable [36]. One study found significant differences in self-efficacy for weight management [38], and 1 found significant differences in self-efficacy for diet [37]. Risk perceptions for type 2 diabetes were reported in 2 studies [26,27], with 1 finding significantly more changes in risk perceptions among women in the intervention group compared to the control group [26]. It was not possible to synthesize psychological outcomes using meta-analysis due to the variability in how these outcomes were measured and reported.

BCTs

A total of 34 different BCTs were coded across the 15 included studies out of the 93 BCTs described in the BCT taxonomy. The number of BCTs in each study ranged from 6 to 17, with a mean of 11.3 (SD 2.82) per study. A 2-tailed independent samples *t* test found no significant difference in the mean number of BCTs in effective interventions (mean 10.5, SD 1.5) and noneffective interventions (mean 12.3, SD 3.7; $t_{7.7}$=–1.19; *P*=.27). Table 4 summarizes the BCTs coded in all studies, and in effective vs noneffective interventions. The most commonly used BCTs were "instruction on how to perform the behavior," which was included in all studies, "credible source" (14/15, 93% of studies), "self-monitoring of behavior" (13/15, 87%), "feedback on behavior" (11/15, 73%), "adding objects to the environment" (10/15, 66%), "prompts and cues" (10/15, 66%), "goal setting (behavior)" (10/15, 66%), "action planning" (10/15, 66%), "problem-solving" (9/15, 60%), and "information about health consequences" (8/15, 53%).

XSL•FO
RenderX

Table 4. Summary of behavior change techniques in effective and noneffective interventions.

Behavior change technique	All studies (n=15), n (%)	Effective any outcome (n=8), n (%)	Not effective for any outcome (n=7), n (%)
Goals and planning			
Goal setting (behavior)	10 (66)	4 (50)	6 (86)
Problem-solving	9 (60)	4 (50)	5 (71)
Goal setting (outcome)	2 (13)	2 (25)	0 (0)
Action planning	10 (66)	5 (63)	5 (71)
Review behavior goal(s)	6 (40)	2 (25)	4 (57)
Discrepancy between current behavior and goal	1 (7)	0 (0)	1 (14)
Review outcome goals	1 (7)	1 (13)	0 (0)
Feedback and monitoring			
Feedback on behavior	11 (73)	5 (63)	6 (86)
Self-monitoring of behavior	13 (87)	7 (88)	6 (86)
Self-monitoring of outcomes of behavior	4 (27)	3 (38)	1 (14)
Feedback on outcomes of behavior	1 (7)	0 (0)	1 (14)
Social support			
Social support (unspecified)	9 (60)	3 (38)	6 (89)
Social support (practical)	2 (13)	2 (25)	0 (0)
Social support (emotional)	1 (7)	1 (13)	0 (0)
Shaping knowledge			
Instruction on how to perform the behavior	15 (100)	8 (100)	7 (100)
Natural consequences			
Information about health consequences	8 (53)	4 (50)	4 (57)
Salience of consequences	3 (20)	3 (38)	0 (0)
Information about social and environmental consequences	4 (27)	2 (25)	2 (29)
Information about emotional consequences	1 (7)	1 (13)	0 (0)
Comparison on behavior			
Demonstration of the behavior	7 (47)	6 (75)	1 (14)
Social comparison	1 (7)	0 (0)	1 (14)
Associations			
Prompts/cues	10 (66)	6 (75)	4 (57)
Repetition and substitution			
Behavioral practice/rehearsal	2 (13)	2 (25)	0 (0)
Habit formation	1 (7)	1 (13)	0 (0)
Graded tasks	2 (13)	0 (0)	2 (29)
Comparison on outcomes			
Credible source	14 (93)	8 (100)	6 (86)
Pros and cons	1 (7)	0 (0)	1 (14)
Reward and threat			
Material reward (behavior)	1 (7)	0 (0)	1 (14)
Nonspecific reward	1 (7)	0 (0)	1 (14)
Social reward	5 (33)	2 (25)	3 (43)
Regulation			

XSL•FO
RenderX

Behavior change technique	All studies (n=15), n (%)	Effective any outcome (n=8), n (%)	Not effective for any outcome (n=7), n (%)
Reduce negative emotions	1 (7)	0 (0)	1 (14)
Antecedents			
Restructuring the physical environment	1 (7)	0 (0)	1 (14)
Adding objects to the environment	10 (66)	7 (88)	3 (43)
Identity			
Framing/reframing	4 (27)	1 (13)	3 (43)

There were 6 BCTs identified as being present in 75% (6/8) or more of effective interventions: "self-monitoring of behavior," "instruction on how to perform the behavior," "demonstration of the behavior," "prompts/cues, credible source," and "adding objects to the environment." However, these BCTs were all present in very similar proportions of the studies that were not effective, except for "demonstration of the behavior," which was found only in 14% (1/7) of noneffective studies compared to 75% (6/8) of effective ones, and "adding objects to the environment," which was present in 43% (3/7) of noneffective interventions compared to 88% (7/8) of effective interventions.

Differences between effective and noneffective interventions were noted for 3 other BCTs: "goal setting (behavior)," "social support (unspecified)," "salience of consequences." The "goal setting (behavior)" BCT was used more frequently in noneffective interventions (6/7, 86%) than in effective interventions (4/8, 50%). Effective interventions less frequently used the BCT "social support (unspecified)" (3/8, 38%) compared to noneffective interventions (6/7, 86%), but more frequently used practical (2/8, 25%) or emotional social support (1/8, 13%), which were not used in any noneffective interventions. The "salience of consequences" BCT was used in 38% (3/8) of effective interventions and not used in any noneffective interventions. However, all of the comparisons between effective and noneffective interventions must be interpreted with caution due to the small number of studies in each group.

Discussion

Principal Findings

This systematic review identified 15 studies [24-38] reported in 17 papers [39,40], investigating interventions aiming to prevent type 2 diabetes among women with a history of gestational diabetes using primarily technology-based delivery. Weight loss was the most frequently assessed outcome [24,25,27-33,35,37,38], but other outcomes related to diabetes risk assessed included waist circumference [27,30,32,35], body composition [28,33,35], BMI [27,29], glycemic control [27-32,34,35,39], health behaviors [24,25,27,28,30,32-35,38], self-efficacy for behavior change [27,31,36-38], and risk perceptions of type 2 diabetes [26,27]. Only 1 study assessed type 2 diabetes prevalence and found no significant differences between the intervention and control group [32]. Pooled analysis of 7 studies [24,27,29,32,33,35,38] revealed a significantly greater weight loss among those receiving technology-based interventions. Interventions delivered using technology [27,29,32,33] and with a longer follow-up [25,32,33,35] resulted in increased weight loss compared to those using combined technology and telemedicine approaches [24,35,38] and studies with a shorter follow-up [27,29,38]. Meta-analysis showed no significant differences in BMI [27,29], fasting glucose [27,29,32,34], 2-hour glucose [27,29,32,34], HbA_{1c} [32,35], or HOMA-IR [32,34] between intervention and control groups. There was heterogeneity in the study populations, dropout rates, target, timing, and duration of the intervention, and modes of delivery. Comparable heterogeneity has been reported in other systematic reviews of technology-based or telemedicine interventions for diabetes prevention [11,13].

Comparison With Other Literature

Although most studies included in this review did not individually report significant effects of technology-based interventions on weight loss, pooled analysis revealed a significantly greater weight loss among those receiving technology-based interventions with a mean difference of –1.01 (95% CI –1.86 to –0.16) kg. Although this difference is relatively small, 1 study found that weight loss of 1 kg in 1079 participants of a diabetes prevention intervention reduced the risk of type 2 diabetes by 16% over 3 years [43]. A systematic review of primarily face-to-face lifestyle interventions for women with a history of gestational diabetes reported a mean difference of –1.06 (95% CI −1.68 to −0.44) kg in pooled analysis of 5 trials, suggesting that technology-based interventions may be close to face-to-face interventions in effectiveness for weight loss in this population [44], but with possible benefits in terms of the resources required for delivery. Research has supported the cost-effectiveness of digital health interventions for supporting behavioral change in other populations [45,46], and so it is possible that technology-based interventions could offer a cost-effective means of supporting lifestyle change among women following a diagnosis of GDM. However, none of the studies in this review included any economic analysis, and this should therefore be explored in future trials.

Subgroup analyses carried out in this meta-analysis showed that interventions delivered using technology-based methods only resulted in more weight loss compared to those interventions using a combination of technology-based and telemedicine approaches, but these differences were not statistically significant. Studies with a follow-up of 6 months or more also resulted in increased weight loss than those with a shorter follow-up, which is consistent with a previous systematic review by Goveia et al [47] that included both technology-based and

XSL•FO
RenderX

face-to-face lifestyle interventions for diabetes prevention in women with GDM. Intervention length and follow-up length in studies in this review were largely similar, and so these findings suggest that technology-based interventions for this population should have a duration of 6 months or more to maximize effectiveness. The need for help over a longer period is supported by women's descriptions of the challenges of managing lifestyle change alongside the multiple roles they fulfill in the postpartum period [10].

The findings in relation to glycemic outcomes were less positive than those for weight. Only 1 of the 5 studies assessing glycemic control reported significant differences favoring the intervention group, and pooled analyses reported no significant differences in fasting glucose, 2-hour glucose, HbA_{1c}, or HOMA-IR. The review by Goveia et al [47] also found no significant effect of lifestyle intervention on fasting and plasma glucose but did find a positive effect for HbA_{1c}. Goveia et al [47] found a reduction in the incidence of type 2 diabetes, which was larger in interventions that started within 6 months of women giving birth. Only 1 study in this review assessed type 2 diabetes and found no significant differences in prevalence between the intervention and control group, but the intervention started up to 3 years postdelivery in this study.

It was not possible to pool studies on physical activity and diet in this review, but individually, most studies did not show significant effects of the intervention on either outcome. In contrast, half of the studies assessing self-efficacy for physical activity and diet, or risk perceptions, reported significant differences that favored the intervention group. This suggests that technology-based interventions included in this review do appear to be capable of changing the antecedents of behavior, even if studies do not demonstrate significant changes in lifestyle behavior. However, it is difficult to draw conclusions from individual studies in this review due to the limitations of the studies in relation to sample size, as discussed in the Limitations section below.

There was no significant difference in the number of BCTs in effective studies compared to noneffective studies in this review. This is in contrast to the findings of a systematic review of technology-driven type 2 diabetes prevention interventions in high-risk groups, excluding women with GDM, by Van Rhoon et al [11], which found that effective interventions had more BCTs on average. In our review, there was an average of 11.3 (SD 2.82) BCTs across all interventions compared to an average of 9 in the interventions included in Van Rhoon et al's review [11]. It has been argued that it is important to include fewer, more effective BCTs in technology-based interventions where attrition rates can be high, and too many BCTs might make the intervention more time-consuming and less engaging [48]. Research on BCTs in technology-based interventions aiming to promote physical activity and reduce sedentary behavior in adults suggests that the specific combination of BCTs may also be important, with certain combinations being more effective than others [48]. Given that there was no difference in the number of BCTs between effective and noneffective interventions in this review, and the challenges and competing demands that women report when attempting lifestyle change in the postpartum period [10], future research should arguably focus on what BCTs are likely to be effective in women with GDM, rather than the quantity of BCTs.

There was little difference in the BCTs contained in effective interventions and noneffective interventions in this review except for "demonstration of behavior," "adding objects to the environment," and specific types of "social support" (eg, "emotional or practical") being more common in effective interventions, whereas less effective interventions were more likely to contain "unspecified social support." Furthermore, descriptions of interventions were often lacking in detail, and these differences in social support may represent differences in reporting rather than true differences in intervention approaches. Clearer specification of interventions is needed in trials to build a cumulative evidence base of effective interventions that can be replicated and refined.

The most commonly included BCTs across all studies in this review were: "instruction on how to perform the behavior," "credible source," "self-monitoring of behavior," "feedback on behavior," "adding objects to the environment," "prompts and cues," "goal setting (behavior)," "problem-solving," "action planning," and "information about health consequences." The systematic review of technology-driven type 2 diabetes prevention interventions in high-risk groups, excluding women with GDM, by Van Rhoon et al [11] found that 90% of effective interventions included the BCTs "social support (unspecified)," "goal setting (outcome/behavior)," "feedback on behavior," and "self-monitoring of outcomes" [11]. Of these, "social support (unspecified)," "goal setting (outcomes)," and "self-monitoring (outcomes)" were not among the most used BCTs in this review, and future research should consider the inclusion of these BCTs and assess their transferability from other high-risk groups to women with previous GDM.

Limitations

Limitations of the review include that non-English papers were excluded, gray literature was not identified, and data were not independently extracted but instead checked by a second reviewer. Prediction intervals could not be calculated in the meta-analysis because of the small number of eligible studies. By reporting only CIs and not prediction intervals, we cannot draw conclusions about the possible variation in the intervention effect across different populations and settings [22]. Key limitations of the studies included in the review were the small sample sizes and lack of power calculations [49]. Sample sizes in all included studies were very small, with all studies containing around 100 participants or fewer in each arm. Two-thirds of studies included in the review either did not report carrying out a power calculation or reported being underpowered. Given that meta-analysis showed significant reductions in weight favoring the interventions when most individual studies did not, it is possible that interventions did result in behavioral changes, but studies were not adequately powered to detect these differences [49]. The certainty of the evidence for all pooled outcomes in this review was rated as very low.

Interventions were often poorly described, making it difficult to code and identify potentially effective ingredients, and most

XSL•FO
RenderX

lacked an explicit theoretical basis to explain and understand how they might work. The explicit use of theory in interventions may increase the likelihood that an intervention will be effective by ensuring that the causal determinants of the behavior are understood and addressed in the intervention [50]. There were not enough studies using theory in this review to suggest a particular model or framework that should be used, but instead, we recommend that, to maximize the potential effectiveness of technology-based interventions in this group and facilitate replicability, future research should clearly describe the content of interventions and consider the theoretical basis for the intervention. There was also a variety in the mode of delivery of digital interventions, and future research should explore whether providing participants with additional devices (eg, fitness trackers) offers any significant benefit over them using their own devices (eg, smartphones) that outweighs the costs of these devices.

Conclusions

This review is the first to summarize interventions with primarily technology-based delivery to prevent type 2 diabetes in women following a diagnosis of GDM. Previously published reviews included interventions with any digital component, and in many of the studies reviewed, the digital aspect was an adjunct to in-person or telemedicine delivery, meaning that conclusions could not be clearly drawn about the benefit of primarily digital interventions [13]. This review is also the first to code the BCTs in interventions with the aim of preventing type 2 diabetes in this population, and so can suggest techniques that should be explored in future interventions. The findings of the review suggest that there may be potential for technology-based interventions to support women in reducing their risk of type 2 diabetes following GDM. However, substantial heterogeneity, significant risk of bias, and very low certainty in the evidence according to the GRADE assessment mean that the findings must be interpreted cautiously. Further, high-quality trials with larger samples, longer follow-up, and clearer reporting of interventions are required to allow firm conclusions to be drawn.

Acknowledgments

The authors thank the Faculty of Health Sciences and Sport, University of Stirling, for supporting the publication of this manuscript. No generative artificial intelligence was used at any stage of this meta-analysis or writing of the manuscript.

Data Availability

The datasets generated during this study are available in the additional supporting files for this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA checklist.
[PDF File (Adobe PDF File), 190 KB - jmir_v28i1e78841_app1.pdf]

Multimedia Appendix 2
PRISMA abstract.
[PDF File (Adobe PDF File), 143 KB - jmir_v28i1e78841_app2.pdf]

Multimedia Appendix 3
PRISMA-S checklist.
[PDF File (Adobe PDF File), 137 KB - jmir_v28i1e78841_app3.pdf]

Multimedia Appendix 4
Full search strategy.
[DOCX File , 26 KB - jmir_v28i1e78841_app4.docx]

Multimedia Appendix 5
Data file.
[XLSX File (Microsoft Excel File), 9 KB - jmir_v28i1e78841_app5.xlsx]

References

XSL•FO
RenderX

1. Cho NH, Shaw JE, Karuranga S, Huang Y, da Rocha Fernandes JD, Ohlrogge AW, et al. IDF diabetes atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res Clin Pract 2018;138:271-281. [doi: 10.1016/j.diabres.2018.02.023] [Medline: 29496507]
2. Xu G, Liu B, Sun Y, Du Y, Snetselaar LG, Hu FB, et al. Prevalence of diagnosed type 1 and type 2 diabetes among US adults in 2016 and 2017: population based study. BMJ 2018;362:k1497 [FREE Full text] [doi: 10.1136/bmj.k1497] [Medline: 30181166]
3. Dennison RA, Chen ES, Green ME, Legard C, Kotecha D, Farmer G, et al. The absolute and relative risk of type 2 diabetes after gestational diabetes: a systematic review and meta-analysis of 129 studies. Diabetes Res Clin Pract 2021;171:108625 [FREE Full text] [doi: 10.1016/j.diabres.2020.108625] [Medline: 33333204]
4. Eades CE, Cameron DM, Evans JM. Prevalence of gestational diabetes mellitus in Europe: a meta-analysis. Diabetes Res Clin Pract 2017;129:173-181. [doi: 10.1016/j.diabres.2017.03.030] [Medline: 28531829]
5. Ferrara A. Increasing prevalence of gestational diabetes mellitus: a public health perspective. Diabetes Care 2007;30 Suppl 2:S141-S146. [doi: 10.2337/dc07-s206] [Medline: 17596462]
6. Kim C, Newton KM, Knopp RH. Gestational diabetes and the incidence of type 2 diabetes: a systematic review. Diabetes Care 2002;25(10):1862-1868. [doi: 10.2337/diacare.25.10.1862] [Medline: 12351492]
7. Ratner RE, Christophi CA, Metzger BE, Dabelea D, Bennett PH, Pi-Sunyer X, Diabetes Prevention Program Research Group. Prevention of diabetes in women with a history of gestational diabetes: effects of metformin and lifestyle interventions. J Clin Endocrinol Metab 2008;93(12):4774-4779 [FREE Full text] [doi: 10.1210/jc.2008-0772] [Medline: 18826999]
8. Lie MLS, Hayes L, Lewis-Barned NJ, May C, White M, Bell R. Preventing type 2 diabetes after gestational diabetes: women's experiences and implications for diabetes prevention interventions. Diabet Med 2013;30(8):986-993. [doi: 10.1111/dme.12206] [Medline: 23534548]
9. Eades CE, France EF, Evans JMM. Postnatal experiences, knowledge and perceptions of women with gestational diabetes. Diabet Med 2018;35(4):519-529. [doi: 10.1111/dme.13580] [Medline: 29338094]
10. Parsons J, Ismail K, Amiel S, Forbes A. Perceptions among women with gestational diabetes. Qual Health Res 2014;24(4):575-585. [doi: 10.1177/1049732314524636] [Medline: 24682021]
11. Van Rhoon L, Byrne M, Morrissey E, Murphy J, McSharry J. A systematic review of the behaviour change techniques and digital features in technology-driven type 2 diabetes prevention interventions. Digit Health 2020;6:2055207620914427 [FREE Full text] [doi: 10.1177/2055207620914427] [Medline: 32269830]
12. Leblalta B, Kebaili H, Sim R, Lee SWH. Digital health interventions for gestational diabetes mellitus: a systematic review and meta-analysis of randomised controlled trials. PLOS Digit Health 2022;1(2):e0000015 [FREE Full text] [doi: 10.1371/journal.pdig.0000015] [Medline: 36812531]
13. Halligan J, Whelan ME, Roberts N, Farmer AJ. Reducing weight and BMI following gestational diabetes: a systematic review and meta-analysis of digital and telemedicine interventions. BMJ Open Diabetes Res Care 2021;9(1):e002077 [FREE Full text] [doi: 10.1136/bmjdrc-2020-002077] [Medline: 33853849]
14. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
15. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021;372:n160 [FREE Full text] [doi: 10.1136/bmj.n160] [Medline: 33781993]
16. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
17. PROSPERO. URL: https://www.crd.york.ac.uk/prospero/ [accessed 2026-02-26]
18. National Institute for Health and Care Excellence. Process and methods guides. In: Methods for the Development of NICE Public Health Guidance [Internet]. London: National Institute for Health and Care Excellence (NICE); 2020.
19. Cumpston M, Li T, Page MJ, Chandler J, Welch VA, Higgins JP, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane handbook for systematic reviews of interventions. Cochrane Database Syst Rev 2019;10(10):ED000142 [FREE Full text] [doi: 10.1002/14651858.ED000142] [Medline: 31643080]
20. IntHout J, Ioannidis JPA, Borm GF. The hartung-knapp-sidik-jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard derSimonian-laird method. BMC Med Res Methodol 2014;14:25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
21. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. Res Synth Methods 2017;8(1):5-18. [doi: 10.1002/jrsm.1230] [Medline: 28058794]
22. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023;12(4):101014 [FREE Full text] [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]

XSL•FO
RenderX

23. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64(4):383-394. [doi: 10.1016/j.jclinepi.2010.04.026] [Medline: 21195583]
24. Cheung NW, Blumenthal C, Smith BJ, Hogan R, Thiagalingam A, Redfern J, et al. A pilot randomised controlled trial of a text messaging intervention with customisation using linked data from wireless wearable activity monitors to improve risk factors following gestational diabetes. Nutrients 2019;11(3):590 [FREE Full text] [doi: 10.3390/nu11030590] [Medline: 30862052]
25. Cheung NW, Simmons D, Marschner S, Thiagalingam A, Pasupathy D, Smith BJ, et al. Randomised controlled trial of a customised text messaging and activity monitor program for lifestyle improvement after gestational diabetes. Nutrients 2024;16(6):820 [FREE Full text] [doi: 10.3390/nu16060820] [Medline: 38542731]
26. Ghaderi M, Farahani MA, Hajiha N, Ghaffari F, Haghani H. The role of smartphone-based education on the risk perception of type 2 diabetes in women with gestational diabetes. Health Technol 2019;9(5):829-837. [doi: 10.1007/s12553-019-00342-3]
27. Kim C, Draska M, Hess ML, Wilson EJ, Richardson CR. A web-based pedometer programme in women with a recent history of gestational diabetes. Diabet Med 2012;29(2):278-283 [FREE Full text] [doi: 10.1111/j.1464-5491.2011.03415.x] [Medline: 21838764]
28. Kim SH, Kim HJ, Shin G. Self-management mobile virtual reality program for women with gestational diabetes. Int J Environ Res Public Health 2021;18(4):1539 [FREE Full text] [doi: 10.3390/ijerph18041539] [Medline: 33562853]
29. Liew SJ, Soon CS, Chooi YC, Tint MT, Eriksson JG. A holistic approach to preventing type 2 diabetes in Asian women with a history of gestational diabetes mellitus: a feasibility study and pilot randomized controlled trial. Front Clin Diabetes Healthc 2023;4:1251411 [FREE Full text] [doi: 10.3389/fcdhc.2023.1251411] [Medline: 37841647]
30. Lim K, Chan SY, Lim SL, Tai BC, Tsai C, Wong SR, et al. A smartphone app to restore optimal weight (SPAROW) in women with recent gestational diabetes mellitus: randomized controlled trial. JMIR Mhealth Uhealth 2021;9(3):e22147 [FREE Full text] [doi: 10.2196/22147] [Medline: 33724204]
31. Nicholson WK, Beckham AJ, Hatley K, Diamond M, Johnson L, Green SL, et al. The gestational diabetes management system (GooDMomS): development, feasibility and lessons learned from a patient-informed, web-based pregnancy and postpartum lifestyle intervention. BMC Pregnancy Childbirth 2016;16(1):277 [FREE Full text] [doi: 10.1186/s12884-016-1064-z] [Medline: 27654119]
32. Nicklas JM, Zera CA, England LJ, Rosner BA, Horton E, Levkoff SE, et al. A web-based lifestyle intervention for women with recent gestational diabetes mellitus: a randomized controlled trial. Obstet Gynecol 2014;124(3):563-570 [FREE Full text] [doi: 10.1097/AOG.0000000000000420] [Medline: 25162257]
33. Potzel AL, Gar C, Banning F, Sacco V, Fritsche A, Fritsche L, et al. A novel smartphone app to change risk behaviors of women after gestational diabetes: a randomized controlled trial. PLoS One 2022;17(4):e0267258 [FREE Full text] [doi: 10.1371/journal.pone.0267258] [Medline: 35476681]
34. Reutrakul S, Martyn-Nemeth P, Quinn L, Rydzon B, Priyadarshini M, Danielson KK, et al. Effects of sleep-extend on glucose metabolism in women with a history of gestational diabetes: a pilot randomized trial. Pilot Feasibility Stud 2022;8(1):119 [FREE Full text] [doi: 10.1186/s40814-022-01076-2] [Medline: 35659776]
35. Rollo ME, Baldwin JN, Hutchesson M, Aguiar EJ, Wynne K, Young A, et al. The feasibility and preliminary efficacy of an eHealth lifestyle program in women with recent gestational diabetes mellitus: a pilot study. Int J Environ Res Public Health 2020;17(19):7115 [FREE Full text] [doi: 10.3390/ijerph17197115] [Medline: 32998401]
36. Saxon CE, Seely EW, Bertin KB, Suresh K, Skurnik G, Roche AT, et al. Self-efficacy and readiness to change among women with recent gestational diabetes engaging in a web-based lifestyle intervention: the balance after baby intervention trial. Am J Lifestyle Med 2026;20(3):385-395. [doi: 10.1177/15598276231155147] [Medline: 41640988]
37. Seely EW, Weitzman PF, Cortes D, Romero Vicente S, Levkoff SE. Development and feasibility of an app to decrease risk factors for type 2 diabetes in Hispanic women with recent gestational diabetes (hola bebé, adiós diabetes): pilot pre-post study. JMIR Form Res 2020;4(12):e19677 [FREE Full text] [doi: 10.2196/19677] [Medline: 33382039]
38. Taylor R, Rollo ME, Baldwin JN, Hutchesson M, Aguiar EJ, Wynne K, et al. Evaluation of a type 2 diabetes risk reduction online program for women with recent gestational diabetes: a randomised trial. Int J Behav Nutr Phys Act 2022;19(1):35 [FREE Full text] [doi: 10.1186/s12966-022-01275-3] [Medline: 35346229]
39. Nicklas JM, Rosner BA, Zera CA, Seely EW. Association between changes in postpartum weight and waist circumference and changes in cardiometabolic risk factors among women with recent gestational diabetes. Prev Chronic Dis 2019;16:E47 [FREE Full text] [doi: 10.5888/pcd16.180308] [Medline: 31002638]
40. Nicklas JM, Skurnik G, Roche T, Schultz C, Suresh K, Seely EW. 191-OR: a web-based lifestyle intervention to reduce postpartum weight retention in women with recent gestational diabetes: the balance after baby intervention trial. Diabetes 2020;69(Supplement_1):191-OR. [doi: 10.2337/db20-191-or]
41. Beeharry D, Swarbrick H, Ahmed H, Nair S, Pond J, Joseph F. Clinical care and other categories posters: education and self-management: P310 multimedia approach to diabetes structured education: the effects of “Keeping Healthy After Gestational Diabetes (GDM)”, a multimedia based education program on patient knowledge, self-efficacy and acceptability. Diabetic Medicine 2017;34(S1):107-130 [FREE Full text] [doi: 10.1111/dme.24_13304]

XSL•FO
RenderX

42. Equality Act 2010. legislation.gov.uk. 2010. URL: https://www.legislation.gov.uk/ukpga/2010/15/contents [accessed 2026-02-26]
43. Hamman RF, Wing RR, Edelstein SL, Lachin JM, Bray GA, Delahanty L, et al. Effect of weight loss with lifestyle intervention on risk of diabetes. Diabetes Care 2006;29(9):2102-2107 [FREE Full text] [doi: 10.2337/dc06-0560] [Medline: 16936160]
44. Gilinsky AS, Kirk AF, Hughes AR, Lindsay RS. Lifestyle interventions for type 2 diabetes prevention in women with prior gestational diabetes: a systematic review and meta-analysis of behavioural, anthropometric and metabolic outcomes. Prev Med Rep 2015;2:448-461 [FREE Full text] [doi: 10.1016/j.pmedr.2015.05.009] [Medline: 26844102]
45. Kyaw TL, Ng N, Theocharaki M, Wennberg P, Sahlen KG. Cost-effectiveness of digital tools for behavior change interventions among people with chronic diseases: systematic review. Interact J Med Res 2023;12:e42396 [FREE Full text] [doi: 10.2196/42396] [Medline: 36795470]
46. Gentili A, Failla G, Melnyk A, Puleo V, Tanna GLD, Ricciardi W, et al. The cost-effectiveness of digital health interventions: a systematic review of the literature. Front Public Health 2022;10:787135 [FREE Full text] [doi: 10.3389/fpubh.2022.787135] [Medline: 36033812]
47. Goveia P, Cañon-Montañez W, Santos DDP, Lopes GW, Ma RCW, Duncan BB, et al. Lifestyle intervention for the prevention of diabetes in women with previous gestational diabetes mellitus: a systematic review and meta-analysis. Front Endocrinol (Lausanne) 2018;9:583 [FREE Full text] [doi: 10.3389/fendo.2018.00583] [Medline: 30344509]
48. Schroé H, Van Dyck D, De Paepe A, Poppe L, Loh WW, Verloigne M, et al. Which behaviour change techniques are effective to promote physical activity and reduce sedentary behaviour in adults: a factorial randomized trial of an e- and m-health intervention. Int J Behav Nutr Phys Act 2020;17(1):127 [FREE Full text] [doi: 10.1186/s12966-020-01001-x] [Medline: 33028335]
49. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. Emerg Med J 2003;20(5):453-458 [FREE Full text] [doi: 10.1136/emj.20.5.453] [Medline: 12954688]
50. Michie S, Prestwich A. Are interventions theory-based? Development of a theory coding scheme. Health Psychol 2010;29(1):1-8 [FREE Full text] [doi: 10.1037/a0016939] [Medline: 20063930]

Abbreviations

BCT: behavior change technique
GDM: gestational diabetes mellitus
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
HbA1c: hemoglobin A1c
HOMA-IR: homeostasis model assessment of insulin resistance
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension

Edited by S Brini; submitted 12.Jun.2025; peer-reviewed by V Mohan, M Lof; comments to author 08.Sep.2025; accepted 13.Feb.2026; published 01.Apr.2026.

Please cite as:
Eades C, Nguyen-Hoang A, Hoyle L, Cameron D, Evans JMM
Technology-Based Interventions for Prevention of Type 2 Diabetes Following Gestational Diabetes: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e78841
URL: https://www.jmir.org/2026/1/e78841
doi:10.2196/78841
PMID:

©Claire Eades, Anh Nguyen-Hoang, Louise Hoyle, Dawn Cameron, Josie MM Evans. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 01.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Predictive Value of Machine Learning for Poststroke Mortality Risk: Systematic Review and Meta-Analysis

Yujie Chen[1*]; Zhujing Ou[1*]; Yiting Deng[1]; Anling Luo[1]; Xuzi Li[1]; Yujia Yang[1]; Qin Yang[1]; Xintong Wu[1*]; Muke Zhou[1*]

Department of Neurology, West China Hospital, Sichuan University, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Muke Zhou
Department of Neurology, West China Hospital, Sichuan University
37 Guo Xue Xiang
Chengdu,
China
Phone: 86 18980601686
Email: muke.zhou@scu.edu.cn

Abstract

Background: People with stroke face a high mortality risk, and an accurate prediction model is essential to the guidance of clinical decision-making in this population. Recently, with growing attention paid to machine learning (ML) in stroke care, some researchers have investigated the effectiveness of ML in predicting the mortality risk in stroke. However, systematic evidence is still lacking for its effectiveness.

Objective: This systematic review aims to evaluate the value of ML in predicting the stroke mortality risk. The findings are expected to offer an evidence-based basis for developing and assessing clinical risk prediction tools.

Methods: A search was made in Cochrane Library, PubMed, Embase, and Web of Science up to June 23, 2025, and studies that reported a complete performance of ML in predicting stroke mortality were included. Studies with only risk factors analyzed were excluded. The risk of bias of the included studies was assessed using PROBAST (Prediction model Risk of Bias Assessment Tool). Pooled risk ratios with 95% CIs and prediction intervals (PIs) were derived using the Hartung-Knapp-Sidik-Jonkman method under a random-effects model. Subgroup analyses were also conducted by model type, stroke type, patient source, and treatment background. Moreover, a metaregression was conducted on the C-index for out-of-hospital mortality at different time points to explore the influence of time factors on the model's predictive performance.

Results: Sixty-eight studies were included (23 predicting in-hospital mortality and 45 predicting out-of-hospital mortality), describing the development of 75 prediction models and 43 external validations. The follow-up period was 1 month to 15 years. For predicting in-hospital mortality, the external validation set had a pooled C-index of 0.727 (95% CI 0.677-0.781, 95% PI 0.521-1.000), with sensitivity and specificity of 0.64 (95% CI 0.57-0.70) and 0.74 (95% CI 0.70-0.77), respectively. For predicting out-of-hospital mortality, the pooled C-index was 0.847 (95% CI 0.808-0.887, 95% PI 0.750-0.956) in the external validation set, with sensitivity and specificity of 0.71 (95% CI 0.55-0.82) and 0.76 (95% CI 0.74-0.78), respectively. Comparatively, the overall pooled C-indexes were 0.788 (95% CI 0.766-0.810, 95% PI 0.621-0.999) and 0.812 (95% CI 0.798-0.826, 95% PI 0.693-0.952), respectively. The metaregression revealed a gradual decline in the predictive performance of the overall model and logistic regression model alone, whereas a random forest model maintained sustained performance. Age, National Institutes of Health Stroke Scale score, and stroke-related complications were the most frequently used variables for modeling.

Conclusions: This is the first meta-analysis to demonstrate that ML-based prediction of stroke mortality is feasible. The performance of ML supports its role as an auxiliary tool for identifying high-risk populations, thereby optimizing clinical monitoring and resource allocation. However, due to substantial heterogeneity and a relatively high risk of bias in available studies, caution is warranted in real-world application. The effectiveness of ML may vary across settings, and external validation is recommended before broader implementation.

Trial Registration: PROSPERO CRD420251086321; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251086321

(J Med Internet Res 2026;28:e83821) doi:10.2196/83821

XSL•FO
RenderX

KEYWORDS

stroke; artificial intelligence; mortality; prognosis; logistic regression; prediction model

Introduction

Rationale

Stroke, an acute focal neurological deficit with cerebrovascular causes, is primarily classified into ischemic stroke and hemorrhagic stroke (intracerebral hemorrhage and subarachnoid hemorrhage) according to its pathogenesis. Acute ischemic stroke (AIS) is the most frequent type, accounting for 60%-80% of the total [1]. As a leading cause of mortality and long-term disability globally, stroke greatly reduces patients' quality of life and brings a heavy socioeconomic burden [2,3]. Notably, stroke mortality is projected to rise by 50% from 2020 to 2050 [4]. Therefore, stroke has become a serious threat to human health.

The functional prognosis of some patients with ischemic stroke (especially cases within the time window) can benefit from aggressive treatments (eg, intravenous and mechanical thrombolysis). However, a large number of patients are still at risk of adverse outcomes, which deserves adequate attention. The global stroke mortality is high, with approximately 7.5 million deaths in 2021 [5]. In countries with high population density (eg, China), people with AIS have an in-hospital mortality or discharge against medical advice of about 6% [6], a 3-month mortality of 1.5%-3.2%, and a 1-year mortality of 3.4%-6% [7,8]. The 90-day mortality of AIS shows no significant difference between patients undergoing endovascular thrombectomy and those not undergoing endovascular thrombectomy [9]. Moreover, people with intracerebral hemorrhage have an in-hospital mortality or discharge against medical advice of approximately 21.8%, and a 1-year mortality of 17.9% [10]. Notably, unlike highly lethal acute-phase hemorrhagic stroke, ischemic stroke presents a unique pattern of "low acute-phase mortality and high medium- to long-term risks" [11,12], and its mortality is mostly attributed to complications, recurrence, and long-term dysfunction. As a result, time-dynamic risk prediction tools are urgently needed for patients with stroke. Developing stage-specific prediction protocols for high-risk stroke is of great value in early intervention, medical counseling, and resource optimization.

Although numerous studies have sought to predict poststroke functional outcome [13,14], relevant models apply primarily to those patients with rehabilitation potential. However, critically ill intensive care unit patients are faced with a unique and urgent requirement during clinical decision-making, that is, to accurately assess mortality risk to guide physician-patient communication, treatment strategies, and allocation of health care resources [15]. Nowadays, validated prediction tools are still lacking for the stroke mortality risk, and clinical assessment relies on traditional tools (eg, Acute Physiology and Chronic Health Evaluation [APACHE] II score). However, such a single-mode assessment is static and subjective, ignoring the inherent complexity of conditions [16]. Some studies have also been conducted on specific biomarkers. For example, the triglyceride-glucose index (an important biochemical indicator for insulin resistance) can predict AIS mortality [17], and adiponectin is independently associated with post-AIS 5-year mortality [18]. However, early prediction of poststroke death by these prediction tools is still lacking in clinical practice. In addition, the Delphi method can also be applied to the prediction of mortality risk, but the resulting bias is a serious clinical challenge. Therefore, developing early prediction tools for the mortality risk is urgently needed.

Objectives

As big data and artificial intelligence (AI) emerge, modeling by the growing volume of patient-related data has become a hot spot, which enables individualized outcome prediction [19]. For example, Bonkhoff and Grefkes [20] argued that machine learning (ML) plus clinical, electrophysiological, and neuroimaging data exhibit a high potential for poststroke functional recovery; Akay et al [21] found that AI achieves individualized outcome prediction and assists in treatment decision-making by analyzing multimodal data. The advantage of ML lies in its capability to integrate multisource heterogeneous data from traditional clinical practice, such as medical histories and medical images, and transform them into efficient predictive systems [22]. Unlike traditional methods reliant on rigorous statistical hypotheses, ML presets no data distributions and possesses the capability to process high-dimensional, large-scale data. Therefore, ML demonstrates higher accuracy and efficacy in disease classification and prognosis evaluation [23,24]. Given that patients with stroke often present with complex conditions and multiple complications, and predictions rely heavily on subjective neurological examinations nowadays, it is significant to integrate multimodal clinical data by ML for enhancing the predictive power for stroke mortality [25]. Under this context, some studies have also further investigated ML for predicting short- [26,27] or long-term [28,29] poststroke mortality based on individual data. However, the predictive performance of these models lacks systematic evidence due to the diversity of ML methods and variables, which is a challenge against the subsequent development of simpler clinical prediction tools. No studies have yet quantitatively assessed the effects of these ML models, so the clinical usefulness of the available predictive systems in routine practice remains unclear. Therefore, we conducted this meta-analysis with the following objectives: (1) to measure the predictive performance of ML for stroke mortality across scenarios and summarize the application status of different ML models in stroke mortality prediction, and (2) to evaluate the methodological quality of available studies and determine whether ML can serve as a novel tool for predicting in-hospital and out-of-hospital stroke mortality. The findings are expected to offer an evidence-based basis for facilitating AI development in this field and further developing simple and clinically useful ML-based prediction tools.

XSL•FO
RenderX

Methods

Study Registration

This meta-analysis followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 Guidelines [30] in Multimedia Appendix 1. The protocol was prospectively registered with PROSPERO (CRD420251086321). To incorporate constructive refinements suggested during the peer-review process, several deviations from the original PROSPERO protocol were implemented to enhance the study's rigor. These updates primarily include (1) restructuring the inclusion or exclusion criteria based on the PICOS framework; (2) expanding the analytical scope, including the addition of subgroup and sensitivity (SE) analyses; and (3) refining the statistical methodology by adopting the Hartung-Knapp-Sidik-Jonkman (HKSJ) approach. Accordingly, the protocol has been updated to version 2.0 on the PROSPERO platform to ensure full alignment with the final manuscript.

Eligibility Criteria

The eligibility criteria are as follows (Textbox 1):

Textbox 1. Eligibility criteria for study selection.

Inclusion criteria

- Populations: patients with stroke, including acute ischemic stroke and unspecified stroke types.
- Model construction: studies using artificial intelligence techniques such as machine learning for automated prediction.
- Outcomes: a complete prediction model for mortality risk.
- Study type: case-control, cohort, nested case-control, or case-cohort studies.
- Language: studies reported in English.

Exclusion criteria

- Populations: only patients with hemorrhagic stroke were included.
- Model construction: only risk factors were analyzed, without constructing a mortality prediction model; validation studies solely assessing established scales.
- Outcomes: any of the following evaluation metrics were lacking: receiver operating characteristic curve, C-index, sensitivity, specificity, accuracy, recall, precision, confusion matrix, F_1-score, or calibration curves.
- Study type: conference abstracts published without peer review, meta-analysis/review/guidelines/letter/case report/protocol.
- Publication year: earlier than 2005.
- Language: non-English language original studies.

Information Sources and Search Strategy

Literature search was carried out in adherence to PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) guidelines [31]. Cochrane Library (CENTRAL), PubMed (NCBI), Embase (Embase.com), and Web of Science (Clarivate) were thoroughly searched as of June 21, 2025. In addition, manual searches were conducted on references cited in systematic review reports on the same or similar topics. Developed based on the database requirements, the search strategy combined both medical subject headings ("Stroke," "Machine learning," "Deep learning," and "Mortality") and free-text keywords, as well as Boolean operators. To ensure a comprehensive retrieval, free-text keywords and relevant subordinate terms were identified via PubMed and Embase in addition to primary MeSH or Emtree terms and were further supplemented based on our clinical expertise. Details of the search strategy are available in Multimedia Appendix 2. Duplicate records were removed with EndNote 20 software (Clarivate Analytics). Search parameters were set to include only peer-reviewed papers, with no restrictions on geographical location or publication year, and no published search filters were used. No peer review of the search strategy was conducted.

Selection Process

All identified studies were imported into EndNote 20 software. All identified studies were imported into EndNote 20 software. After duplicate publications were removed, we used EndNote's filtering functionality to exclude records based on publication type (eg, reviews, case reports, letters, protocols, animal studies, and meeting abstracts), language (non-English), and publication date (pre-2005). Subsequently, 2 independent investigators (YJC and ZJO) screened the titles and abstracts of the remaining records based on predefined criteria regarding study population (stroke), intervention (machine learning), and outcomes (mortality). Thereafter, a rigorous full-text review was performed to exclude studies lacking model evaluation parameters or those that did not involve the development of original models. The screen was followed by cross-checking. In case of any disagreements, a third investigator (MKZ) was consulted for resolution.

Data Collection Process and Data Items

A spreadsheet was created to extract data, including title, first author, type of study, year of publication, patient source, dataset source, stroke type, treatment background (whether to undergo acute-phase vascular recanalization), follow-up time, total number of death cases in the training set, overfitting methods, number of cases in the validation set, handling methods for

XSL•FO
RenderX

missing value, variable selection, model types, and modeling variables. During data extraction, studies with missing outcome data were excluded directly from the meta-analysis. Three independent investigators (YJC, YTD, and ALL) implemented the data extraction, followed by cross-checking. In case of any disagreements, a third investigator (XTW) was consulted for resolution.

Risk of Bias Assessment

The included studies were assessed for risk of bias (RoB) using the PROBAST (Prediction model Risk of Bias Assessment Tool) [32] by structured questions across 4 domains: participants (2 questions), predictors (3 questions), outcomes (6 questions), and analysis (9 questions). Each question was answered as "Yes," "Probably yes," "No," "Probably no," or "No information." Each domain was rated as high RoB if 1 or more questions were answered as "No" or "Probably no," as low RoB if all answers were "Yes" or "Probably yes," and as unclear RoB if answered "No information." The overall risk was deemed low only when all domains were rated as low RoB, whereas any high RoB was considered overall high risk. In the absence of high-risk domains, the overall risk was unclear when 1 or more domains had unclear RoB. Two investigators (YJC and YJY) independently performed RoB assessment and cross-checking. Discrepancies were settled by third-party adjudication.

Synthesis Methods

The C-index, a measure for the overall predictive accuracy of ML models, underwent a meta-analysis. If the C-index was reported without its 95% CI or standard error, the SE was estimated using the method described by Debray et al [33]. Similarly, SE and specificity (SP) were synthesized by a bivariate mixed-effects model. Meta-analyses of these metrics typically rely on diagnostic 4-fold tables, but most primary studies included did not report such tables. In this case, we sought to reconstruct the diagnostic 4-fold table by leveraging other available performance metrics, such as SE, SP, precision, accuracy, and the total number of events (deaths). Some studies were based on the same database, but significant heterogeneity was present across these studies in the data extraction period, variable selection, model construction, and parameter adjustment rule. Therefore, all eligible studies were included, and given this potential heterogeneity, a random-effects model was adopted for meta-analysis [34]. Multiple entries from the same study represented distinct prediction models or validation methods, which were treated as independent units of evidence [35]. We performed random-effects meta-analyses using the HKSJ method to estimate pooled effect sizes with 95% CIs. The HKSJ method was chosen as it provides more robust inference, particularly when the number of studies is small. In addition, we calculated 95% prediction intervals (PIs) to present the expected range of true effects in similar studies, across clinical settings, or in future studies, and to better express the extent of heterogeneity [36,37]. The width of PIs was directly influenced by the number of included studies and the extent of heterogeneity [38]. Model performance was evaluated separately in training and validation sets. Furthermore, the performance of models for in-hospital and out-of-hospital mortality was separately synthesized.

To explore potential sources of heterogeneity, subgroup analyses were conducted, stratified by stroke type (AIS vs not specific), patient source (public database vs local hospital), treatment background (reperfusion therapy vs not specific), and model type. Furthermore, to evaluate the generalizability of the models and minimize optimistic bias associated with internal validation [39], we conducted a subgroup analysis in the validation set by validation method (internal validation vs external validation). When reporting the results, we prioritized the performance of the external validation subgroup based on the primary importance of external validation results for assessing clinical generalization [35]. The overall pooled performance (combining internal and external validation) was subsequently reported as a reference during development setting. Given the limited number of external validation studies, subgroup analyses were based on the overall performance. We used a random-effects model to calculate pooled estimates within each subgroup and a fixed-effects (plural) model to compare effect sizes across subgroups. A random-effects metaregression was conducted to investigate the association between the C-index of model performance and the follow-up time after discharge in studies reporting out-of-hospital mortality. The small-study effect in the validation sets of included studies and different models (≥10 studies) in the validation sets was qualitatively and quantitatively assessed with funnel plots and Egger regression tests, respectively. Funnel plot asymmetry was visualized to detect the small-study effect.

SE analyses were performed in 2 stages. In the first stage, to assess the potential impact of patient cohort overlap from using the same public databases (Medical Information Mart for Intensive Care [MIMIC-III/IV], eICU Collaborative Research Database [eICU-CRD], and Sentinel Stroke National Audit Programme [SSNAP]) on the pooled results, only 1 study with a large sample size, high methodological quality (based on PROBAST), and complete analysis (using external validation) was retained per public database. All studies from single-center or independent regional registries were also retained. Using independent datasets filtered in the 2-stage ways, a meta-analysis of the C-index for in-hospital and out-of-hospital mortality prediction models was conducted.

To mitigate potential bias arising from studies reporting multiple models, we established a priority rule for the second stage: selecting only the best-performing model from each study. Specifically, for studies with multiple models, we prioritized the optimal model among those with external validation; if no external validation was performed, the model with the highest internal performance was selected. In this way, we could obtain the evidence of the model with the greatest potential for clinical generalization or optimal performance from each study of independent sets, thereby assessing the robustness of the main findings under this stricter criterion. Additionally, following this criterion (the principle of selecting the study with the largest sample size among overlapping validation sets and subsequently selecting the optimal model within each independent validation set), a separate meta-analysis of the C-index was conducted in the external validation set. It should be noted that the SE

XSL•FO
RenderX

analyses were performed post hoc, and thus they were not prespecified in the preregistered study protocol or initial data analysis plan. Statistical analyses and visualizations were performed using Stata (version 15.0; StataCorp, LLC) and R (version 4.5.1; R Foundation for Statistical Computing), using metafor, forestplot, and ggplot2 packages.

Protocol Deviations and Updates

To incorporate constructive refinements from the peer-review process, several deviations from the original PROSPERO protocol were implemented to enhance the study's rigor. These updates primarily include (1) restructuring the inclusion or exclusion criteria according to the PICOS framework, (2) refining the statistical methodology by adopting the HKSJ method and reporting PIs, and (3) expanding the scope of analysis by adding SE and subgroup analyses (eg, internal vs external validation). Accordingly, the study protocol has been updated to version 2.0 on the PROSPERO platform to ensure full alignment with the final paper.

Results

Study Selection

We initially retrieved 21,177 records, of which 6491 duplicate publications were eliminated. After title and abstract review, 14,544 studies were further excluded. Then the full text of the remainder was examined, of which 38 were excluded due to no modeling, 31 due to no definite outcome metrics, and 5 due to a focus only on validation of previous scores. Finally, 68 studies were included [26-29,40-103] (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

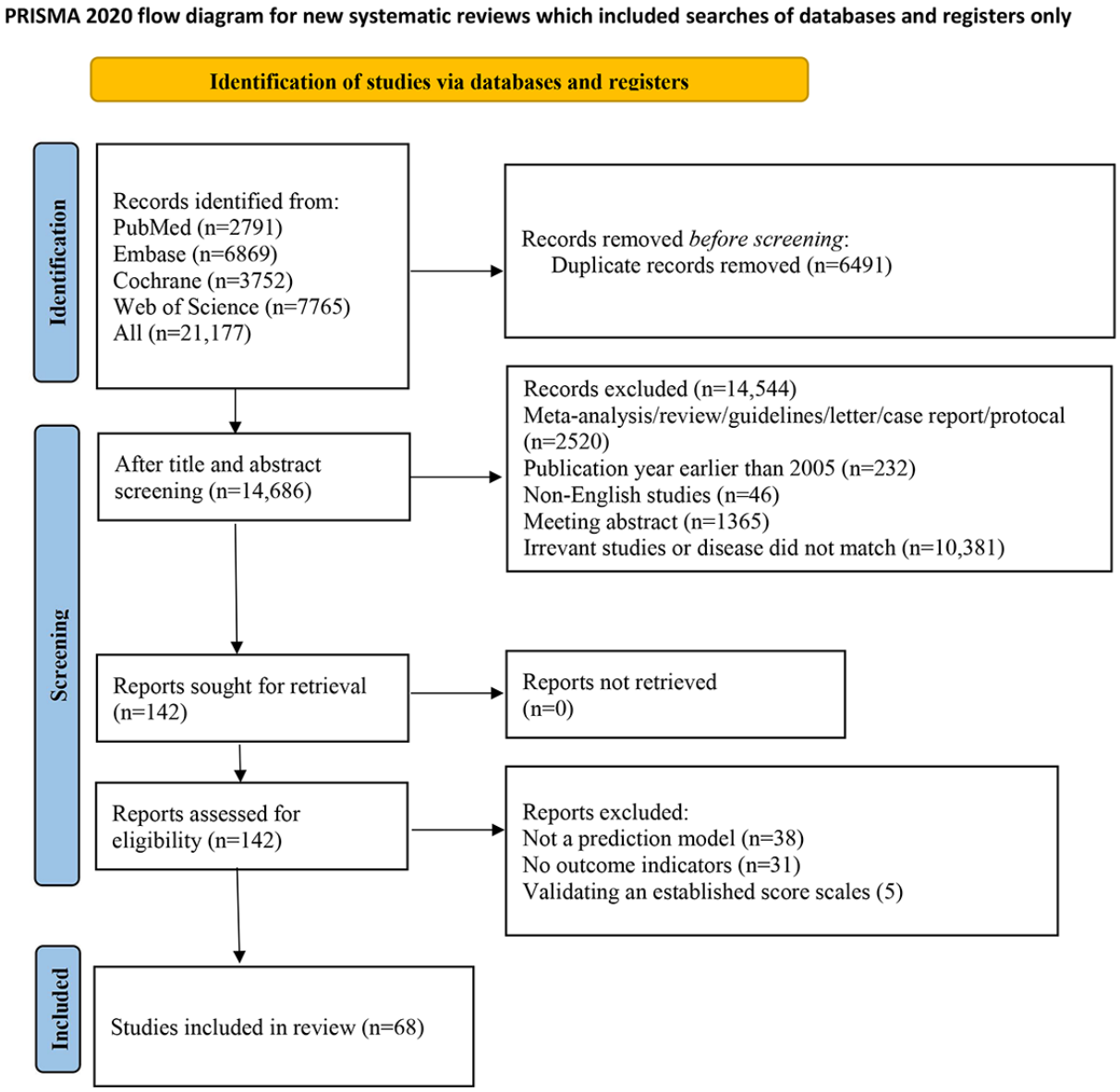

XSL•FO
RenderX

Study Characteristics

We included 68 studies published between 2005 and 2025, mainly from 2007 to 2025. They came from 19 countries and involved 1,759,472 participants, of whom 225,598 (12.8%) died. Twenty-three studies predicted the in-hospital mortality risk, and 45 studies predicted the out-of-hospital mortality risk. Three studies reported both in-hospital and out-of-hospital mortality [41,67,100]. Out-of-hospital mortality was followed up for 1-12 months, as well as 2 years [67], 3 years [46,72], 5 years [82], 10 years [45,75], and 15 years [40] in a few studies. Public databases, including MIMIC, eICU-CRD, the Minimum Basic Data Set, and SSNAP, were used in 28 studies, and there were 26 single-center studies and 14 multicenter studies. Forty-two studies focused on AIS, and the remaining 26 did not specify the stroke type. Seven studies predicted the mortality risk following acute-phase vascular recanalization. Logistic regression (LR) was the most frequently used model (51/68, 75%), followed by extreme gradient boosting (XGBoost) (21/68, 31%), random forest (RF) (20/68, 29%), support vector machine (11/68, 16%), and deep learning (DL) (2/68, 3%). For modeling variables, radiomics features were adopted in only 2 studies, and clinical features were adopted in the remainder. The generation method for the validation set was clearly described in 64 studies. Thirteen studies performed independent external validation, and 11 studies adopted both internal and external validation. For internal validation, random sampling was used in 44 studies, Bootstrap in 9 studies, k-fold cross-validation in 6 studies, holdout validation in 2 studies, and temporal validation in 1 study (Table S1 in Multimedia Appendix 2).

RoB in Studies

For participant selection, PROBAST considers retrospective cohort or case-control studies that do not use public databases with high RoB. Since 8 case-control studies did not use publicly available databases, they were rated as high RoB. All included studies developed reasonable eligibility criteria, so no RoB was introduced in this dimension.

The predictors had the same definition and assessment among participants in all studies, and they were all valid, so all studies were rated as low RoB. The outcome event was death, and the possibility of assessment with knowledge of the outcome could not be ruled out for case-control studies, so we rated 8 studies as high RoB.

The outcome in this meta-analysis was death or nondeath, and it was reasonably categorized and defined. The definition of the outcome eliminated the influence of predictors and was consistent among participants, and the outcome was assessed objectively with or without knowledge of the predictor. Therefore, all studies were rated as low RoB in this dimension.

For statistical analysis, 11 studies were assessed as high RoB because of inappropriate sample sizes; 11 as high RoB because subjects with missing data were directly excluded, without appropriate handling, and 3 as unclear RoB due to no description of the handling method for missing data. The RoB in 13 studies was high since only univariate analysis was conducted. The RoB in 10 studies was high for failing to account for data complexity, including high dimensionality, multicollinearity, or unbalanced data handling. Three studies were rated as high RoB for inappropriate internal validation. All studies were rated as low RoB since they all appropriately handled and incorporated continuous or categorical independent variables and assessed model performance.

To sum up, the overall RoB was low in 33 studies and high in 35 studies (Table S2 in Multimedia Appendix 2 [26-29,40-103] and Multimedia Appendix 3).

Results of Synthesis

In-hospital Mortality

Twenty-five ML models for predicting in-hospital mortality were extracted and included 1,345,411 patients, of whom 159,409 (11.7%) died. The pooled C-index was 0.815 (95% CI 0.783-0.848, 95% PI 0.666-0.997) (Table 1 and Figure 2). The C-index was 0.807 (95% CI 0.770-0.846, 95% PI 0.669-1.000) in studies on AIS, 0.811 (95% CI 0.771-0.853, 95% PI 0.652-1.000) in studies based on public databases, 0.825 (95% CI 0.759-0.897, 95% PI 0.657-1.000) in studies based on hospital databases, and 0.774 (95% CI 0.745-0.803, 95% PI 0.739-0.810) in studies explicitly reporting acute-phase vascular recanalization. LR, RF, and XGBoost models were dominated, and the LR model had a C-index of 0.803 (95% CI 0.742-0.868, 95% PI 0.638-1.000).

In addition, diagnostic metrics could be calculated in 24 studies in the training set, with a pooled SE of 0.81 (95% CI 0.70-0.88) and SP of 0.82 (95% CI 0.76-0.87). The SE and SP of the LR model were 0.71 (95% CI 0.61-0.79) and 0.78 (95% CI 0.69-0.85) (Table 2 and Figures S1 and S2 in Multimedia Appendix 2).

XSL•FO
RenderX

Table 1. Results of meta-analysis of C-index of machine learning for predicting in-hospital mortality in stroke.

Subgroup analysis	Training set						Validation set					
	n	Sample size	C-index (95% CI)	95% PI[a]	Tau	I^2 (%)	n	Sample size	C-index (95% CI)	95% PI	Tau	I^2 (%)
Model												
ANN[b]	1	14,8891	0.651 (0.646-0.656)	N/A[c]	N/A	N/A	8	6837	0.794 (0.696-0.905)	0.544-1.000	0.151	98.40
DT[d]	—[e]	—	—	—	—	—	5	5233	0.720 (0.648-0.800)	0.581-0.892	0.068	73.00
EL[f]	—	—	—	—	—	—	2	845	0.864 (0.373-1.000)	N/A	N/A	N/A
GBM[g]	2	2633	0.842 (0.400-1.000)	N/A	N/A	N/A	2	3813	0.813 (0.248-1.000)	N/A	N/A	N/A
KNN[h]	—	—	—	—	—	—	9	10,555	0.756 (0.707-0.805)	0.590-0.967	0.101	92.80
LASSO[i]	1	77,653	0.879 (0.871-0.886)	N/A	N/A	N/A	1	9141	0.884 (0.863-0.905)	N/A	N/A	N/A
LightGBM[j]	—	—	—	—	—	—	4	2223	0.828 (0.685-1.000)	0.549-1.000	0.114	92.80
LR[k]	8	17,2088	0.803 (0.742-0.868)	0.638-1.000	0.091	99.10	26	419,933	0.802 (0.762-0.844)	0.631-1.000	0.114	99.80
NB[l]	2	1300	0.838 (0.720-0.976)	N/A	N/A	N/A	5	2464	0.621 (0.339-1.139)	0.146-1.000	0.475	95.40
RF[m]	5	15,8453	0.789 (0.727-0.856)	0.647-0.963	0.065	98.30	14	13,721	0.811 (0.773-0.851)	0.682-0.963	0.077	93.20
SVM[n]	1	2031	0.924 (0.909-0.939)	N/A	N/A	N/A	7	7056	0.763 (0.715-0.814)	0.643-0.906	0.064	90.50
XGBoost[o]	5	12,128	0.844 (0.738-0.965)	0.609-1.000	0.107	99.30	16	269,990	0.803 (0.761-0.847)	0.650-0.991	0.096	97.60
Stroke type												
AIS[p]	20	56,6742	0.807 (0.770-0.846)	0.669-1.000	0.099	99.80	65	108,059	0.788 (0.766-0.810)	0.642-0.965	0.101	99.60
Not specific	5	8435	0.848 (0.769-0.936)	0.669-1.000	0.078	97.00	34	643,752	0.784 (0.733-0.840)	0.551-1.000	0.171	97.20
Data resource												
Public database	18	561,389	0.811 (0.771-0.853)	0.652-1.000	0.101	99.80	87	723,433	0.775 (0.753-0.799)	0.613-0.980	0.117	99.50
Local hospital	7	13,788	0.825 (0.759-0.897)	0.657-1.000	0.087	93.50	12	28,378	0.879 (0.840-0.921)	0.749-1.000	0.070	95.70
Treatment background												
EVT[q]/IVT[r]	4	572,769	0.774 (0.745-0.803)	0.739-0.810	0.000	0.00	—	—	—	—	—	—
Not specific	21	2408	0.823 (0.786-0.862)	0.663-1.000	0.101	99.80	—	—	—	—	—	—
Validation method												
Internal validation	—	—	—	—	—	—	67	480,774	0.815 (0.796-0.835)	0.681-0.976	0.089	99.60
External validation	—	—	—	—	—	—	32	271 037	0.727 (0.677-0.781)	0.521-1.000	0.170	96.20
Overall	25	575,177	0.815 (0.783-0.848)	0.666-0.997	0.096	99.70	99	751,811	0.788 (0.766-0.810)	0.621-0.999	0.119	99.50

[a]PI: prediction interval.

XSL•FO
RenderX

[b]ANN: artificial neural network.

[c]N/A: not applicable.

[d]DT: decision tree.

[e]Not available.

[f]EL: ensemble learning.

[g]GBM: gradient boosting machine.

[h]KNN: k-nearest neighbor.

[i]LASSO: least absolute shrinkage and selection operator.

[j]LightGBM: light gradient boosting machine.

[k]LR: logistic regression.

[l]NB: Naive Bayes.

[m]RF: random forest.

[n]SVM: support vector machine.

[o]XGBoost: extreme gradient boosting.

[p]AIS: acute ischemic stroke.

[q]EVT: endovascular thrombectomy.

[r]IVT: intravenous thrombolysis.

Figure 2. Forest plot for meta-analysis of C-index of predictive models for in-hospital mortality in stroke in the training set. AIS: acute ischemic stroke; ANN: artificial neural network; EVT: endovascular thrombectomy; GBM: gradient boosting machine; IVT: intravenous thrombolysis; LASSO: least absolute shrinkage and selection operator; LR: logistic regression; NB: Naive Bayes; PI: prediction interval; RF: random forest; SVM: support vector machine; XGBoost: extreme gradient boosting.

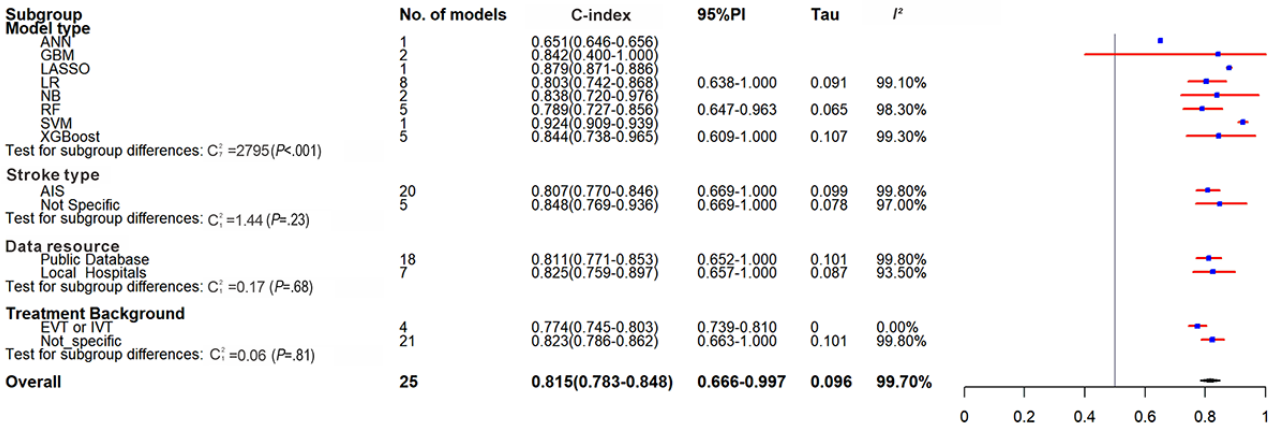

XSL•FO
RenderX

Table 2. Results of meta-analysis of sensitivity and specificity of machine learning for predicting in-hospital mortality in stroke.

Subgroup analysis	Training set				Validation set			
	n	Sample size	SE[a] (95% CI)	SP[b] (95% CI)	n	Sample size	SE (95% CI)	SP (95% CI)
Model								
ANN[c]	2	14,9231	0.92 (0.86-0.98)	0.81 (0.64-0.97)	9	7177	0.74 (0.58-0.85)	0.79 (0.72-0.85)
DT[d]	—[e]	—	—	—	5	5233	0.65 (0.44-0.81)	0.78 (0.52-0.92)
EL[f]	—	—	—	—	2	845	0.88 (0.82-0.94)	0.89 (0.86-0.91)
GBM[g]	2	2633	0.88 (0.72-0.79)	0.79 (0.75-0.83)	2	3813	0.58 (0.33-0.83)	0.78 (0.58-0.98)
KNN[h]	—	—	—	—	9	10,555	0.52 (0.34-0.69)	0.83 (0.70-0.91)
LASSO[i]	1	77,653	0.81	0.81	1	9141	0.79	0.84
LightGBM[j]	—	—	—	—	4	2223	0.73 (0.58-0.83)	0.82 (0.71-0.90)
LR[k]	7	166,373	0.71 (0.61-0.79)	0.78 (0.69-0.85)	19	189,436	0.72 (0.66-0.78)	0.78 (0.70-0.84)
NB[l]	2	1300	0.82 (0.70-0.93)	0.74 (0.69-0.77)	5	2464	0.64 (0.36-0.85)	0.70 (0.61-0.77)
RF[m]	4	152,738	0.66 (0.62-0.70)	0.78 (0.68,0.86)	13	12,292	0.74 (0.63-0.82)	0.81 (0.69-0.89)
SVM[n]	1	2031	0.90	0.85	7	7056	0.61 (0.51-0.70)	0.77 (0.71-0.82)
XGBoost[o]	5	6753	0.78 (0.62-0.88)	0.90 (0.77-0.96)	13	138,201	0.69 (0.62-0.75)	0.73 (0.68-0.78)
Stroke type								
AIS[p]	17	549,597	0.73 (0.66-0.79)	0.82 (0.75-0.88)	58	102,373	0.70 (0.66-0.73)	0.77 (0.74-0.80)
Not specific	7	9115	0.92 (0.67-0.98)	0.83 (0.71-0.90)	31	286,063	0.67 (0.57-0.75)	0.82 (0.74-0.88)
Data resource								
Public database	15	544,244	0.74 (0.67-0.80)	0.82 (0.75-0.88)	76	359,666	0.70 (0.66-0.73)	0.75 (0.72-0.77)
Local hospital	9	14,468	0.89 (0.67-0.97)	0.82 (0.72-0.89)	13	28,770	0.63 (0.43-0.80)	0.93 (0.87-0.97)
Treatment background								
EVT[q]/IVT[r]	4	2408	0.68 (0.64-0.72)	0.76 (0.69-0.82)	N/A[s]	N/A	N/A	N/A
Not specific	20	556,304	0.83 (0.71-0.91)	0.83 (0.77-0.88)	N/A	N/A	N/A	N/A
Validation method								
Internal validation	—	—	—	—	56	111,5921	0.71 (0.66-0.76)	0.81 (0.77-0.85)
External validation	—	—	—	—	33	272,515	0.64 (0.57-0.70)	0.74 (0.70-0.77)
Overall	24	558,712	0.81 (0.70-0.88)	0.82 (0.76-0.87)	89	388,436	0.69 (0.65-0.73)	0.79 (0.76-0.82)

[a]SE: sensitivity.
[b]SP: specificity.
[c]ANN: artificial neural network.
[d]DT: decision tree
[e]Not available.
[f]EL: ensemble learning.
[g]GBM: gradient boosting machine.
[h]KNN: k-nearest neighbor.
[i]LASSO: least absolute shrinkage and selection operator.
[j]LightGBM: light gradient boosting machine.
[k]LR: logistic regression.
[l]NB: Naive Bayes.
[m]RF: random forest.
[n]SVM: support vector machine.
[o]XGBoost: extreme gradient boosting.

XSL•FO
RenderX

[p]AIS: acute ischemic stroke.

[q]EVT: endovascular thrombectomy.

[r]IVT: intravenous thrombolysis.

[s]N/A: not applicable.

Based on externally validated models, the analysis indicated a moderate level of generalizability, with a pooled C-index of 0.727 (95% CI 0.677-0.781, 95% PI 0.521-1.000), with an SE of 0.64 (95% CI 0.57-0.70) and an SP of 0.74 (95% CI 0.70-0.77). These performance estimates were lower than the overall pooled results which yielded a C-index of 0.788 (95% CI 0.766-0.810, 95% PI 0.621-0.999), an overall SE of 0.69 (95% CI 0.65-0.73), and an SP of 0.79 (95% CI 0.76-0.82). The models using hospital databases outperformed those using public databases; the former had a C-index of 0.879 (95% CI 0.840-0.921, 95% PI 0.621-0.999), with SP and SE of 0.93 (95% CI 0.87-0.97) and 0.63 (95% CI 0.43-0.80), and the latter had a C-index of 0.775 (95% CI 0.753-0.799, 95% PI 0.613-0.980), with SP and SE of 0.75 (95% CI 0.72-0.77) and 0.70 (95% CI 0.66-0.73). The studies on AIS had a C-index of 0.788 (95% CI 0.766-0.810, 95% PI 0.642-0.965), with SP and SE of 0.70 (95% CI 0.66-0.73) and 0.77 (95% CI 0.74-0.80), and the LR model had a C-index of 0.802 (95% CI 0.762-0.844, 95% PI 0.631-1.000), with SE and SP of 0.72 (95% CI 0.66-0.78) and 0.78 (95% CI 0.70-0.84), consistent with the overall results. The internal validation subgroup aligned with the overall trend but yielded more optimistic estimates, with a pooled C-index of 0.815 (95% CI 0.796-0.835, 95% PI 0.681-0.976), an SE of 0.71 (95% CI 0.66-0.76), and an SP of 0.81 (95% CI 0.77-0.85) (Tables 1 and 2, Figure 3, and Figures S3 and S4 in Multimedia Appendix 2).

Figure 3. Forest plot for meta-analysis of C-index of predictive models for inhospital mortality in stroke in the validation set. AIS: acute ischemic stroke; ANN: artificial neural network; DT: decision tree; EL: ensemble learning; GBM: gradient boosting machine; KNN: k-nearest neighbor; LASSO: least absolute shrinkage and selection operator; LightGBM: light gradient boosting machine; LR: logistic regression; NA: Not applicable; NA:Not applicable; NB: Naive Bayes; PI: prediction interval; RF: random forest; SVM: support vector machine; XGBoost: extreme gradient boosting.

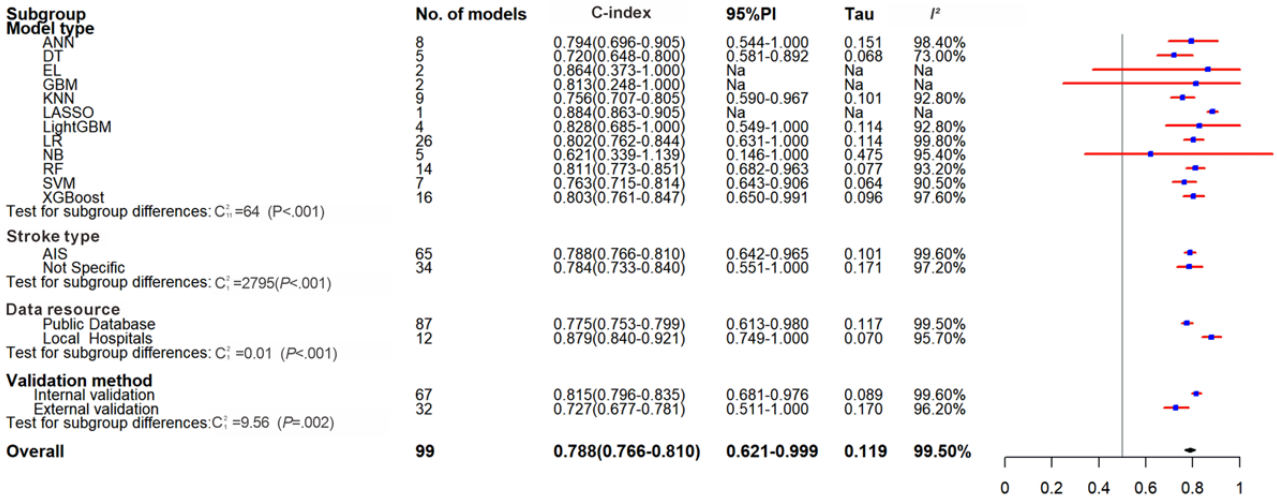

Out-of-Hospital Mortality

Fifty ML models for predicting out-of-hospital mortality were extracted and included 414,061 patients, of whom 66,189 (16%) died. The pooled C-index was 0.823 (95% CI 0.798-0.848, 95% PI 0.672-1.000) (Table 3 and Figure 4), and the diagnostic SE and SP were 0.66 (95% CI 0.60-0.72) and 0.88 (95% CI 0.83-0.91) (Figures S5 and S6 in Multimedia Appendix 2).

XSL•FO
RenderX

Table 3. Results of meta-analysis of C-index of machine learning for predicting out-of-hospital mortality in stroke.

Subgroup analysis	Training set						Validation set					
	n	Sample size	C-index (95% CI)	95% PI[a]	Tau	I^2 (%)	n	Sample size	C-index (95% CI)	95% PI	Tau	I^2 (%)
Model												
ANN[b]	3	4330	0.813 (0.768-0.860)	0.779-0.847	0.001	46.80	4	2615	0.822 (0.653-1.000)	0.498-1.000	0.140	95.3
DT[c]	1	474	0.793 (0.697-0.889)	N/A[d]	N/A	N/A	3	1906	0.69 5 (0.585-0.825)	0.556-0.868	0.025	0.0
EL[e]	—[f]	—	—	—	—	—	3	17,053	0.857 (0.708-1.000)	0.588-1.000	0.075	97.4
GBM[g]	3	3333	0.863 (0.660-1.000)	0.515-1.000	0.103	92.20	5	2076	0.788 (0.717-0.866)	0.652-0.953	0.058	57.6
KNN[h]	2	2859	0.669 (0.0356-1.000)	N/A	N/A	N/A	4	5205	0.737 (0.637-0.852)	0.578-0.939	0.065	80.1
LASSO[i]	—	—	—	—	—	—	—	—	—	—	—	—
LightGBM[j]	1	1388	0.840 (0.790-0.890)	N/A	N/A	N/A	2	2116	0.926 (0.665-1.000)	N/A	N/A	N/A
LR[k]	24	151,284	0.822 (0.795-0.850)	0.698-0.968	0.077	98.10	27	149,308	0.817 (0.793-0.841)	0.708-0.942	0.068	96.8
NB[l]	3	3333	0.832 (0.628-1.000)	0.490-1.000	0.105	86.60	6	6687	0.809 (0.741-0.883)	0.651-1.000	0.076	86.3
RF[m]	5	9502	0.835 (0.699-0.996)	0.549-1.000	0.137	99.50	16	17,729	0.833 (0.801-0.865)	0.724-0.957	0.063	83.1
RSF[n]	—	—	—	—	—	—	2	383	0.789 (0.630-0.986)	N/A	N/A	N/A
SVM[o]	2	2859	0.762 (0.249-1.000)	N/A	N/A	N/A	7	7225	0.789 (0.721-0.863)	0.620-1.000	0.091	92.1
XGBoost[p]	3	7660	0.857 (0.685-1.000)	0.552-1.000	0.088	97.70	10	11,327	0.800 (0.747-0.856)	0.643-0.994	0.091	96.8
DL[q]	1	1546	0.955 (0.925-0.984)	N/A	N/A	N/A	2	1161	0.873 (0.799-0.955)	N/A	N/A	N/A
COX[r]	1	425	0.822 (0.804-0.839)	N/A	N/A	N/A	5	4991	0.792 (0.720-0.871)	0.632-0.993	0.074	92.2
CatBoost[s]	1	1471	0.895 (0.878-0.912)	N/A	N/A	N/A	5	3902	0.818 (0.715-0.937)	0.599-1.000	0.102	94.1
AdaBoost[t]	—	—	—	—	—	—	1	1968	0.850 (0.811-0.889)	N/A	N/A	N/A
Stroke type												
AIS[u]	43	176,770	0.827 (0.801-0.855)	0.678-1.000	0.097	98.90	66	163,990	0.796 (0.778-0.813)	0.677-0.935	0.080	91.4
Not specific	7	13,694	0.796 (0.719-0.882)	0.597-1.000	0.110	99.20	36	71,662	0.841 (0.819-0.862)	0.733-0.964	0.066	97.8
Data resource												
Public database	24	171,811	0.826 (0.798-0.854)	0.696-0.979	0.081	99.50	30	161,344	0.789 (0.763-0.815)	0.661-0.941	0.085	99.4
Local hospital	26	18,653	0.818 (0.776-0.862)	0.633-1.000	0.015	91.70	72	74,308	0.824 (0.808-0.841)	0.712-0.955	0.073	87.1
Treatment background												
EVT[v]/IVT[w]	3	920	0.853 (0.796-0.914)	0.783-0.929	0.008	0.00	7	1338	0.817 (0.760-0.879)	0.681-0.981	0.068	77.5

XSL•FO
RenderX

Subgroup analysis	Training set						Validation set					
	n	Sample size	C-index (95% CI)	95% PI[a]	Tau	I^2 (%)	n	Sample size	C-index (95% CI)	95% PI	Tau	I^2 (%)
Not specific	47	189,544	0.821 (0.795-0.848)	0.666-1.000	0.103	99.10	95	234,314	0.811 (0.797-0.826)	0.691-0.953	0.081	98.4
Validation method												
Internal validation	—	—	—	—	—	—	92	230,828	0.808 (0.793-0.823)	0.687-0.950	0.081	98.5
External validation	—	—	—	—	—	—	10	4824	0.847 (0.808-0.887)	0.750-0.956	0.050	71.3
Overall	50	190,464	0.823 (0.798-0.848)	0.672-1.000	0.099	99.10	102	235,652	0.812 (0.798-0.826)	0.693-0.952	0.080	98.3

[a]PI: prediction interval.
[b]ANN: artificial neural network.
[c]DT: decision tree.
[d]Not applicable.
[e]EL: ensemble learning.
[f]Not available.
[g]GBM: gradient boosting machine.
[h]KNN: k-nearest neighbors.
[i]LASSO: least absolute shrinkage and selection operator.
[j]LightGBM: light gradient boosting machine.
[k]LR: logistic regression.
[l]NB: Naive Bayes.
[m]RF: random forest.
[n]RSF: random survival forest.
[o]SVM: support vector machine.
[p]XGBoost: extreme gradient boosting.
[q]DL: deep learning.
[r]COX: Cox proportional hazards.
[s]CatBoost: categorical boosting.
[t]AdaBoost: adaptive boosting.
[u]AIS: acute ischemic stroke.
[v]EVT: endovascular thrombectomy.
[w]IVT: intravenous thrombolysis.

XSL•FO
RenderX

Figure 4. Forest plot for meta-analysis of C-index of predictive models for out-of-hospital mortality in stroke in the training set. AIS: acute ischemic stroke; ANN: artificial neural network; CatBoost: categorical boosting; COX: Cox proportional hazards; DL: deep learning; DT: decision tree; EVT: endovascular thrombectomy; GBM: gradient boosting machine; IVT: intravenous thrombolysis; KNN: k-nearest neighbor; LightGBM: light gradient boosting machine; LR: logistic regression; NA: Not applicable; NB: Naive Bayes; PI: prediction interval; RF: random forest; SVM: support vector machine; XGBoost: extreme gradient boosting.

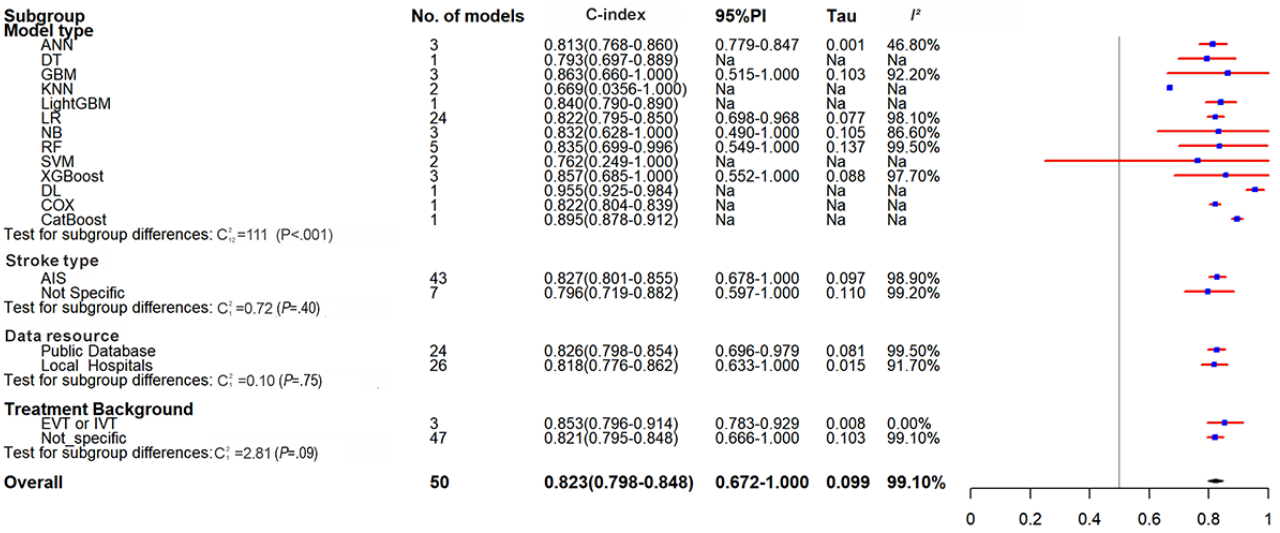

The pooled C-index was 0.827 (95% CI 0.801-0.855, 95% PI 0.678-1.000) in studies on AIS, 0.826 (95% CI 0.798-0.854, 95% PI 0.696-0.979) in studies based on public databases, 0.818 (95% CI 0.776-0.862, 95% PI 0.633-1.000) in studies based on hospital databases, and 0.853 (95% CI 0.796-0.914, 95% PI 0.783-0.929) in studies reporting acute-phase vascular recanalization. LR remained the most used model (24/50, 48%), with a C-index of 0.824 (95% CI 0.798-0.850) (Table 4 and Figure 4).

For out-of-hospital mortality, the pooled performance in external validation yielded a C-index of 0.847 (95% CI 0.808-0.887, 95% PI 0.750-0.956), with an SE of 0.71 (95% CI 0.55-0.82) and an SP of 0.76 (95% CI 0.74-0.78). The overall pooled analysis yielded a C-index of 0.812 (95% CI 0.798-0.826, 95% PI 0.693-0.952), with an SE of 0.68 (95% CI 0.63-0.72) and an SP of 0.82 (95% CI 0.78-0.85). Studies on AIS had a pooled C-index of 0.796 (95% CI 0.778-0.813, 95% PI 0.677-0.935), with SE and SP of 0.65 (95% CI 0.59-0.71) and 0.81 (95% CI 0.75-0.85). Studies based on public databases had a pooled C-index of 0.789 (95% CI 0.763-0.815, 95% PI 0.661-0.941), with SE and SP of 0.69 (95% CI 0.62-0.76) and 0.77 (95% CI 0.74-0.80). Studies based on hospital databases had a pooled C-index of 0.824 (95% CI 0.808-0.841, 95% PI 0.712-0.955), with SE and SP of 0.68 (95% CI 0.62-0.74) and 0.84 (95% CI 0.79-0.88). Studies reporting acute-phase vascular recanalization had a pooled C-index of 0.817 (95% CI 0.760-0.879, 95% PI 0.681-0.981), with SE and SP of 0.77 (95% CI 0.69-0.84) and 0.79 (95% CI 0.72-0.85). Internal validation subgroup yielded a C-index of 0.808 (95% CI 0.793-0.823, 95% PI 0.687-0.950), with SE and SP of 0.67 (95% CI 0.62-0.72) and 0.82 (95% CI 0.78-0.85), respectively. The subgroup results were all consistent with the overall trend. In the validation process for out-of-hospital mortality prediction models, those constructed with data sourced from private hospitals show superior overall performance when contrasted with models that rely on data from public datasets.

LR also demonstrated better performance, with a C-index, SE, and SP of 0.817 (95% CI 0.793-0.841, 95% PI 0.708-0.942), 0.74 (95% CI 0.69-0.79), and 0.79 (95% CI 0.73-0.84), respectively. DL models for out-of-hospital mortality were created in 2 studies, and they exhibited good discriminatory power (C-index: 0.873, 95% CI 0.799-0.955) but unsatisfactory diagnostic performance (SE 0.48, SP 0.76). (Tables 3 and 4, Figure 5, and Figures S7 and S8 in Multimedia Appendix 2).

XSL•FO
RenderX

Table 4. Results of meta-analysis of sensitivity and specificity of machine learning for predicting out-of-hospital mortality in stroke.

Subgroup analysis	Training set				Validation set			
	n	Sample size	SE[a] (95% CI)	SP[b] (95% CI)	n	Sample size	SE (95% CI)	SP (95% CI)
Model								
ANN[c]	3	4330	0.58 (0.55-0.61)	0.80 (0.73-0.87)	5	2860	0.61 (0.38-0.80)	0.85 (0.74-0.92)
DT[d]	—[e]	—	—		4	2217	0.62 (0.24-0.89)	0.72 (0.45-0.89)
EL[f]	—	—	—	—	3	17,053	0.76 (0.71-0.81)	0.80 (0.78-0.81)
GBM[g]	2	2859	0.55 (0.39-0.71)	0.84 (0.75-0.93)	5	2076	0.56 (0.38-0.72)	0.80 (0.52-0.94)
KNN[h]	2	2859	0.45 (0.45-0.46)	0.92 (0.92-0.93)	4	5205	0.53 (0.33-0.72)	0.80 (0.69-0.88)
AdaBoost[ii]	—	—	—	—	1	1968	0.80	0.81
LightGBM[j]	1	1388	0.53	0.83	2	2116	0.77 (0.67-0.86)	0.90 (0.82-0.97)
LR[k]	14	44,677	0.76 (0.69-0.81)	0.80 (0.76-0.83)	21	29,892	0.74 (0.69-0.79)	0.79 (0.73-0.84)
NB[l]	2	2859	0.51 (0.50-0.52)	0.87 (0.82-0.92)	6	6687	0.68 (0.55-0.79)	0.81 (0.67-0.89)
RF[m]	3	3313	0.68 (0.52-0.83)	0.81 (0.68-0.93)	11	13,978	0.64 (0.46-0.79)	0.88 (0.72-0.95)
RSF[n]	—	—	—	—	2	383	0.45 (0.21-0.69)	0.59 (0.42-0.76)
SVM[o]	2	2859	0.49 (0.41-0.57)	0.81 (0.75-0.86)	7	7225	0.67 (0.53-0.78)	0.87 (0.73-0.94)
XGBoost[p]	1	1471	0.33	0.98	8	9179	0.70 (0.48-0.86)	0.84 (0.70-0.91)
DL[q]	2	3821	0.92 (0.83-1.00)	0.99 (0.99-1)	2	1161	0.47 (0.47-0.48)	0.76
COX[r]	1	425	0.62	0.87	3	1101	0.70 (0.61-0.78)	0.74 (0.70-0.78)
CatBoost[s]	1	1471	0.68	0.90	3	2741	0.71 (0.52-0.89)	0.78 (0.77-0.79)
Stroke type								
AIS[t]	29	62,174	0.64 (0.56-0.70)	0.89 (0.84-0.93)	49	33,411	0.65 (0.59-0.71)	0.81 (0.75-0.85)
Not specific	5	10,158	0.78 (0.74-0.80)	0.75 (0.74-0.76)	38	72,431	0.70 (0.64-0.76)	0.83 (0.78-0.87)
Data resource								
Public database	17	55,453	0.64 (0.57-0.71)	0.89 (0.84-0.93)	30	54,088	0.69 (0.62-0.76)	0.77 (0.74-0.80)
Local hospital	17	16,879	0.69 (0.58-0.78)	0.87 (0.77-0.93)	57	51,754	0.68 (0.62-0.74)	0.84 (0.79-0.88)
Treatment background								
EVT[u]/IVT[v]	—	—	—	—	5	1044	0.77 (0.69-0.84)	0.79 (0.72-0.85)
Not specific	—	—	—	—	82	104,798	0.67 (0.62-0.72)	0.82 (0.78-0.85)
Validation method								
Internal validation	—	—	—	—	82	103,116	0.67 (0.62-0.72)	0.82 (0.78-0.85)
External validation	—	—	—	—	5	2726	0.71 (0.55-0.82)	0.76 (0.74-0.78)
Overall	34	72,332	0.66 (0.60-0.72)	0.88 (0.83-0.91)	87	105,842	0.68 (0.63- 0.72)	0.82 (0.78-0.85)

[a]SE: sensitivity.
[b]SP: specificity.
[c]ANN: artificial neural network.
[d]DT: decision tree
[e]Not available.
[f]EL: ensemble learning.
[g]GBM: gradient boosting machine.
[h]KNN: k-nearest neighbor.
[i]AdaBoost: adaptive boosting.

XSL•FO
RenderX

[j]LightGBM: light gradient boosting machine.

[k]LR: logistic regression.

[l]NB: Naive Bayes.

[m]RF: random forest.

[n]RSF: random survival forest.

[o]SVM: support vector machine.

[p]XGBoost: extreme gradient boosting.

[q]DL: deep learning.

[r]COX: Cox proportional hazards.

[s]CatBoost: categorical boosting.

[t]AIS: acute ischemic stroke.

[u]EVT: endovascular thrombectomy.

[v]IVT: intravenous thrombolysis.

Figure 5. Forest plot for meta-analysis of C-index of predictive models for out-of-hospital mortality in stroke in the validation set. AdaBoost: adaptive boosting; AIS: acute ischemic stroke; ANN: artificial neural network;CatBoost: categorical boosting; COX: Cox proportional hazards; DL: deep learning; DT: decision tree; EL: ensemble learning; EVT: endovascular thrombectomy; GBM: gradient boosting machine; IVT: intravenous thrombolysis; KNN: k-nearest neighbor; LightGBM: light gradient boosting machine; LR: logistic regression; NA:Not applicable; NB: Naive Bayes; PI: prediction interval; RF: random forest; RSF: random survival forest; SVM: support vector machine; XGBoost: extreme gradient boosting.

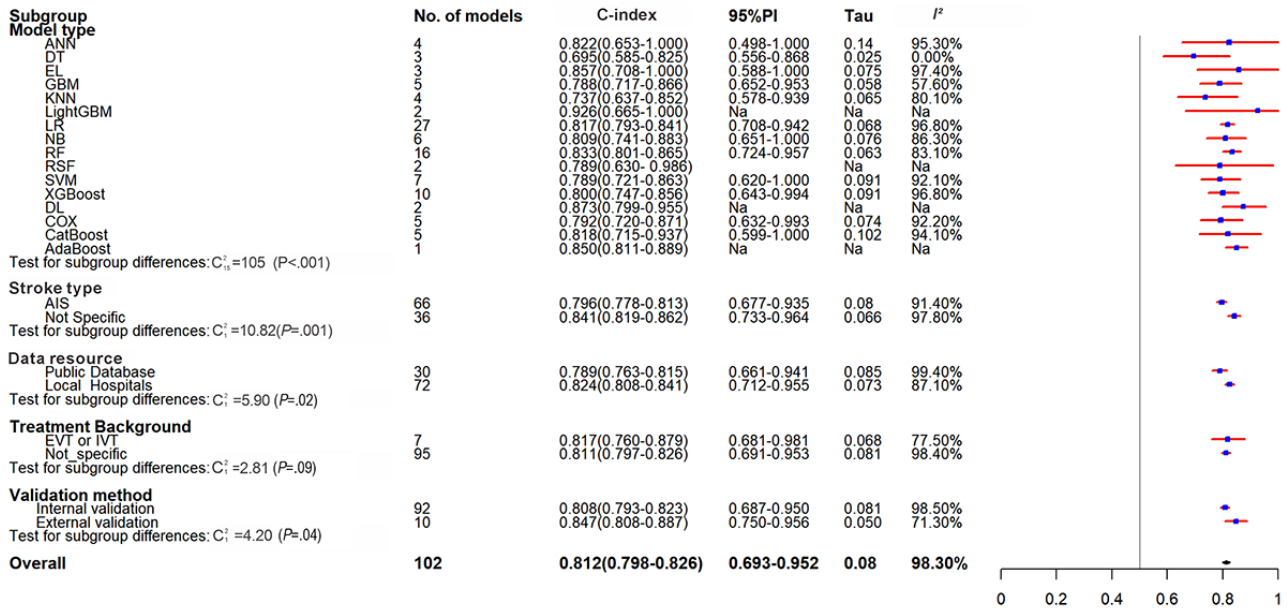

Surprisingly, only a few studies adopted survival analysis modeling (only 1 COX [Cox Proportional Hazards] model in the training set, and only 5 COX models and 2 random survival forest models in the validation set). Furthermore, metaregression revealed a gradual decline in the predictive performance of the LR models and a sustained performance of the RF and XGBoost models (Figure 6A and B and Figures S9-S11 in Multimedia Appendix 2).

XSL•FO
RenderX

Figure 6. Metaregression bubble plot of follow-up time in the predictive models for (A) training set and (B) validation set.

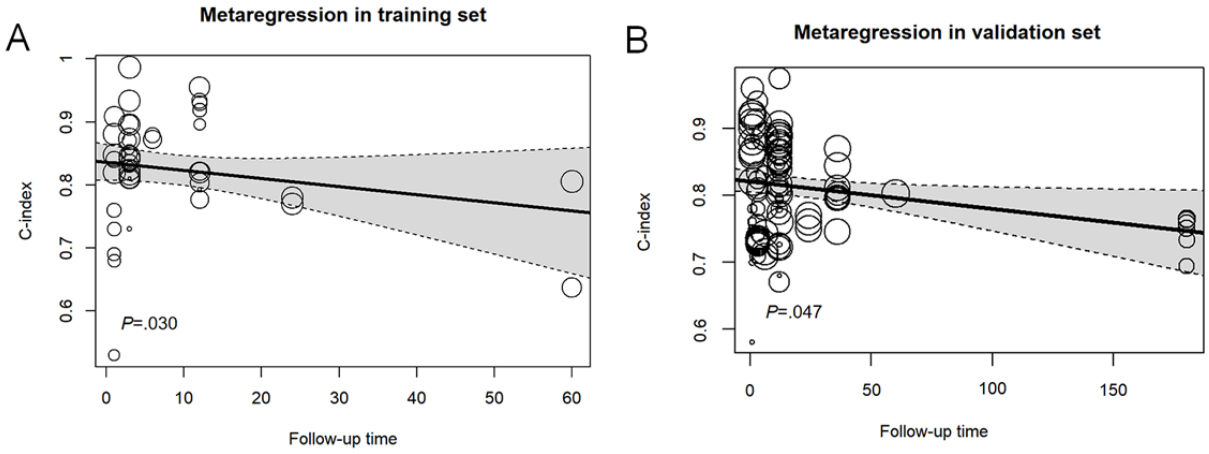

Modeling Variables

Of the 155 clinical variables included, high-frequency variables (>5 occurrences) are shown in Figure 7, with age (n=36), National Institutes of Health Stroke Scale (NIHSS) score (n=16), and hypertension (n=16) as the top 3. These variables were distributed across comorbidities (smoking, previous stroke history, cancer, and in-hospital infections), acute severity of illness (NIHSS, Glasgow Coma Scale, and premorbid modified Rankin Scale [mRS] scores), and admission characteristics (time of admission). Few studies involved biomarkers such as platelet-to-neutrophil ratio (PNR) (n=1) and neutrophil-percentage-to-albumin ratio (NPAR) (n=1).

Figure 7. Machine learning models established by predictors for poststroke mortality. GCS: Glasgow Coma Scale; mRS: modified Rankin Scale; NIHSS: National Institutes of Health Stroke Scale.

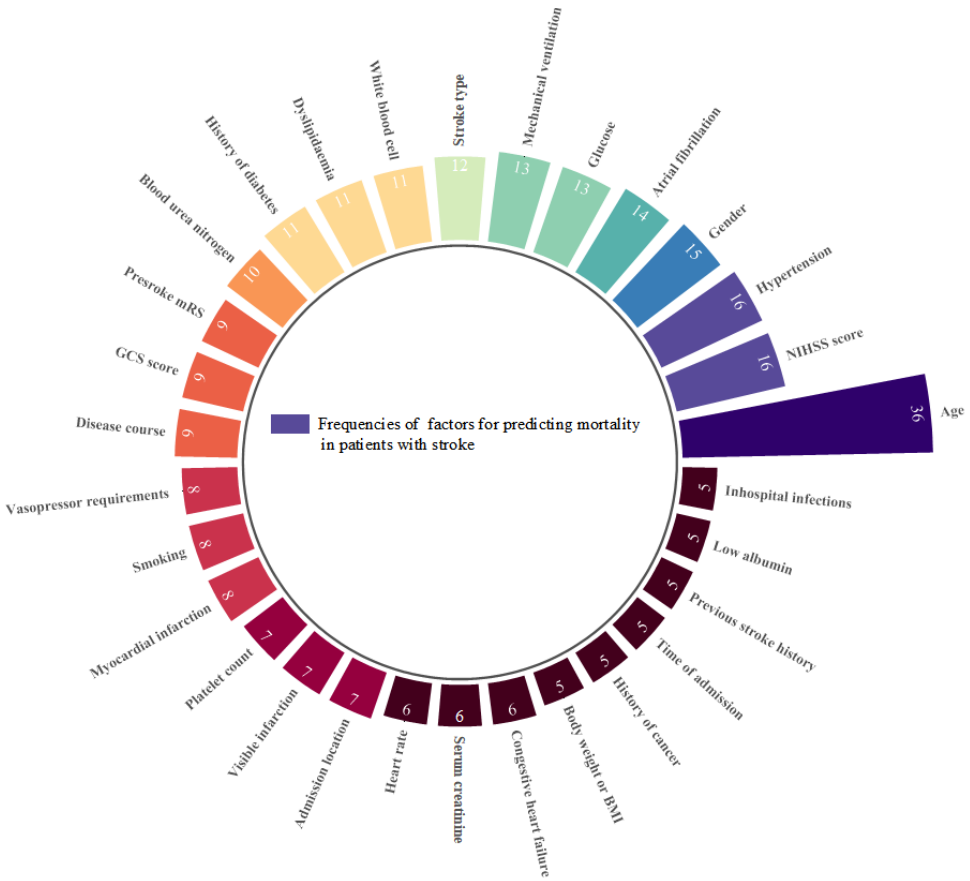

XSL•FO
RenderX

Small-Study Effect Assessment

As revealed by the funnel plot asymmetry (with fewer studies skewing toward lower performance values) and significant results of Egger tests (P<.001), significant small study effects were present in the included studies on out-of-hospital mortality (Figure S13 in Multimedia Appendix 2). Meanwhile, no small study effects were present in the included studies on in-hospital mortality and LR and RF models (Figures S12-S15 in Multimedia Appendix 2).

SE Analysis

SE analyses were performed after controlling for cohort overlap, yielding pooled C-indexes based on selected studies: For in-hospital mortality, Ouyang et al [50] (MIMIC-III), Cummins et al [48] (MIMIC-IV), Huang et al [58] (eICU-CRD), and Wang et al [43] (SSNAP) were selected; for out-of-hospital mortality, Li et al [81] (MIMIC-III) and Zhu et al [96] (MIMIC-IV) were selected as representatives. Thus, the pooled C-indexes for in-hospital and out-of-hospital mortality prediction models were 0.801 (95% CI 0.777-0.828, 95% PI 0.682-0.992) and 0.825 (95% CI 0.811-0.840, 95% PI 0.711-0.957), respectively. Similar to the primary analysis, these results indicated that the main findings of this meta-analysis were robust to cohort overlap. From further single original studies, 16 and 32 optimal models for predicting in-hospital and out-of-hospital mortality were retained, respectively, with pooled C-indexes of 0.852 (95% CI 0.811-0.895, 95% PI 0.699-1.000) and 0.842 (95% CI 0.817-0.868, 95% PI 0.720-0.986), respectively (Figures S16 [27,43,45,48,50,51,58,67,68,70,76,83,99,101-103] and S17 [29, 40, 42, 44, 46, 49, 53-56, 59, 60, 65, 67, 69, 71, 72, 74, 75, 77-82, 86, 89, 91, 93, 95, 96, 100] in Multimedia Appendix 2). Besides, 5 and 6 external validation studies were included for in-hospital and out-of-hospital mortality, respectively. They yielded pooled C-indexes of 0.785 (95% CI 0.698-0.884, 95% PI 0.594-1.000) and 0.853 (95% CI 0.781-0.931, 95% PI 0.708-1.000), respectively (Figures S18 [50,51,53,63,99] and S19 [29,54,56,71,74,79] in Multimedia Appendix 2).

Discussion

Main Findings

To our knowledge, this is the first meta-analysis to evaluate ML as a promising tool for predicting stroke mortality. Different from previous reviews, we conducted subgroup analyses across multiple dimensions, including the model type, patient source, and time. Sixty-eight studies were included in this study, involving numerous ML models for predicting poststroke mortality. Despite methodological concerns of these studies, the most rigorous evidence on generalizability derived from the externally validated subset. The results demonstrated favorable discriminatory power, with pooled C-indices of approximately 0.73 for in-hospital and 0.85 for out-of-hospital mortality, alongside maintaining balanced SE and SP for both in-hospital and out-of-hospital mortality.

The predictive performance of LR models for long-term out-of-hospital mortality might gradually decline, while RF and XGBoost achieved a robust and sustained performance. Subgroup analyses suggested that the models for AIS and the presence or absence of acute-phase vascular recanalization had no great difference from the overall model. Accounting for approximately 50% of the total, LR had slightly inferior performance to other ML methods. Besides, the PI for the C-index for in-hospital or out-of-hospital mortality did not cover the reference line, suggesting that the LR model consistently outperforms random prediction across clinical settings and possesses translational potential. Notably, the PIs were broader in some subgroups due to the limited number of included studies, so further validation is needed in the future. To sum up, ML holds significant value in predicting both in-hospital and out-of-hospital stroke mortality. These findings offer an evidence basis for developing precise, dynamic clinical risk prediction tools in the future.

Comparison With Previous Reviews

Previous reviews have also investigated ML for early judgment of stroke prognosis. Yang et al [104] conducted a meta-analysis of AI models for predicting ischemic stroke outcomes by binary prediction of 90-day mRS, with 7 studies included. They found a pooled area under the curve (AUC) up to 0.872 of the fixed-effects model, confirming the value of AI in predicting a good prognosis of stroke (mRS ≤2). However, they ignored the assessment of mortality (mRS=6). Schwartz et al [105] and Wang et al [106] included 25 and 7 original studies, respectively, and argued that ML achieves better predictive effects on short- and long-term stroke mortality. However, they performed only a qualitative assessment, thus quantitative evidence was lacking. In addition, available studies on ML for predicting stroke mortality do not classify prediction models and consider the influence of time factors on the model accuracy, generating concerns about heterogeneity in population and methodology. Therefore, we expanded the model performance beyond previous systematic reviews by quantitatively synthesizing evidence from external validations to evaluate model generalizability, in addition to assessing the goodness-of-fit in training sets.

Modeling Variables

Modeling variables (clinical features, radiomics features, genetic information, and microbiomics) also constituted a challenge to this study. Due to the critical conditions of severe stroke, some examinations such as magnetic resonance imaging are difficult and consume vast human and material resources. Only 2 studies used radiomics [28,54], which relied primarily on interpretable clinical features. We synthesized interpretable clinical features and confirmed that age, NIHSS score, and hypertension were the most used modeling variables, consistent with previous reviews [21,105]. Meanwhile, some studies also focused on specific laboratory indicators. NPAR combines inflammation with nutrient metabolic status, which is more reflective of the body's acute stress response and overall reserve status [107]. Therefore, NPAR possesses greater predictive power. Chunjuan et al [62] created a nomogram using NPAR and the systematic inflammatory response index for predicting stroke mortality, with an AUC of 0.637. Chung et al [108] proposed that NPAR in combination with simple clinical indicators demonstrates excellent predictive performance for severe illness, with an AUC (0.929) comparable with the Sequential Organ Failure Assessment score, which enables better clinical judgment and

XSL•FO
RenderX

interventions. Geng et al [93] adopted PNR to reflect thrombotic and inflammatory status, in which AUC increases from 0.736 (PNR alone) to 0.888 (PNR incorporating age and NIHSS scores). Therefore, the association of changes in some variables with the mortality risk can be observed in the future.

For intensive care unit patients, simple biomarkers and clinical parameters can be combined with laboratory tests to predict the disease severity. In addition, radiomics can be further introduced to reflect individual neuropathologic features and predict prognosis [109]. Besides, models from individual hospitals generally yielded higher C-indexes. The possible reason is as follows: electronic health records at medical centers often provide more detailed, abundant, and clinically relevant predictor variables than public databases covering broad populations, enabling models to capture more complex prediction patterns [110,111].

Model Selection

Besides modeling variables, the model types (well-interpretable LR, COX, competing risk, and DT) should also be taken into account. These models can quantify the association of variables with risk and thus assist in determining the specific scoring tools, but their accuracy is unsatisfactory. In addition, traditional ML displays high accuracy [112], but its interpretability is challenging. Although ML models interpret "black-box models" using various Shapley additive explanations (SHAP) diagrams at the local and global levels in some studies, the SHAP value under its computing paradigm provides the mean contribution of features in the global scope rather than local logical explanations for specific predictions. Due to this essential attribute, SHAP is difficult to meet the requirements for causality and fidelity in high-reliability scenarios such as clinical diagnosis. Consequently, SHAP values often fail to accurately capture the actual impact of features and may generate misleading feature importance rankings, which restricts their credibility and applicability in high-risk decision-making [113,114]. In the subgroup analysis, LR was dominant in the included studies, with favorable SE and C-index, suggesting that routine interpretable models can early predict stroke mortality. Because of its theoretical simplicity and ease of implementation, LR is widely favored in clinical prediction models [115]. In particular, datasets may be relatively simple and exhibit linear trends in the distribution when data predicting stroke mortality are primarily from clinical history and testing. Therefore, the performance of LR is often comparable with that of complex ML models [112,116]. Meanwhile, LR can quantify the association between various influencing factors and mortality risk and further create a nomogram to better visualize results and thus develop convenient online prediction tools [117]. Thus, despite growing attention to ML, LR remains dominant in data analysis.

Besides predictive accuracy, model interpretability must also be considered in current clinical applications. Generally, highly interpretable models, such as LR, COX, and DT, exhibit limitations in predictive accuracy. In contrast, less interpretable models, such as RF, support vector machine, neural networks, and DL, possess superior accuracy but low interpretability due to their "black-box" property [118,119]. Consequently, less interpretable models face challenges in clinical practice, including limited use by health care providers and difficulty in patient management [120]. Therefore, highly interpretable models should be prioritized. If their accuracy fails to meet clinical requirements, model interpretation techniques (eg, Grad-CAM and LIME) should be introduced to enhance understanding and trust in complex model outputs while maintaining predictive performance [121]. Such interpretable methods are crucial for advancing AI integration and application in clinical settings [122,123]. Besides, DL has high efficiency for image processing and excellent time effect capture ability when processing time series data [124,125], so it has the potential for judging stroke site and time. However, DL is less applied to the early prediction of poststroke mortality, possibly attributed to the difficulty in its interpretability and high requirements for techniques [126]. In the future, DL using multimodal data may be more accurate in prognostic prediction of stroke.

Our meta-analysis revealed a gradual decline in the predictive performance of the overall model and LR model for long-term mortality, consistent with previous reviews [127,128]. The reason may be related to variations in clinical background, alterations in body status, and introduction of time-varying confounders. Notably, RF achieved a robust performance in long-term mortality prediction, with high accuracy, suggesting that RF may be an effective ML model for predicting stroke mortality. Given rehabilitation progresses postdischarge, condition and complications of patients with stroke often present dynamic evolution, so the reliability of AI and ML models relying solely on information at a single time point during hospitalization is restricted in prediction and assessment. Therefore, real-time collection and analysis of patient data are crucial for accurately identifying individual phenotypic changes [129]. Establishing a collaborative real-time data network has become a key foundation for long-term, dynamic clinical decision-making [19]. Moreover, for predicting out-of-hospital mortality, survival analysis modeling was adopted in very few studies. The clinical significance greatly varied between survival analysis modeling and nonsurvival analysis modeling, while the 2 had a minor difference in short-term follow-up; dynamic mortality risk prediction was often a focus in clinical practice, which could be better achieved by survival analysis modeling, with a statistical basis for decision-making [130]. In contrast, nonsurvival analysis modeling was difficult to achieve dynamic mortality risk prediction. Although LR dominates in the analysis currently, dynamic prediction models should be more considered for long-term mortality in the future to develop predictive strategies in real time.

Prospects and Outlook

The potential of ML for predicting stroke mortality was confirmed in this study, but several technical challenges remain. The population is unbalanced data and the dataset balance is a serious challenge [131]. This meta-analysis found a mortality of about 13% from public databases and 11% from hospital databases. However, the impact of data imbalance was considered in few studies. We synthesized SE and SP of models of unbalanced data and found that despite good performance in predicting negative events (ie, survival), these models were less

XSL•FO
RenderX

accurate in predicting positive events (ie, mortality) and were difficult to produce clinical benefits [132]. Therefore, the impact of data imbalance should be taken into account in future modeling.

Another challenge arises from the limited availability and divergent performance patterns of external validation evidence. For in-hospital mortality, the performance in external validation was more conservative than the overall performance and performance in internal validation, probably due to optimistic bias in high-dimensional prognostic settings and simple internal validations [133,134], However, the external validation set remained clinically meaningful and represented an appropriate basis for assessing model generalizability. Conversely, regarding out-of-hospital mortality, the external validation set had more satisfactory C-index than internal validation. This phenomenon was observed in all the 4 included studies [29,54,71,79], which might be related to the fact that patients had more severe conditions, higher mortality, and greater case-mix variation in external validation, thus enhancing the model's discriminatory power [29,135]. However, due to the limited number and potential representativeness concerns [136], the generalizability of this finding requires cautious interpretation.

This meta-analysis revealed significant small study effects in out-of-hospital mortality, with small sample studies tending to report lower C-index values. This may also be attributed to limited standardized data, introducing data noise that attenuated the model's apparent performance [137]. Furthermore, the optimal performance still generally fell below the average of large sample studies. Small sample studies also possessed a narrower case-mix (ie, more homogeneous population), and established evidence suggests that the C-index is naturally lower in homogeneous populations [39]. The lower pooled C-index observed in small sample studies likely reflected clinical heterogeneity or implementation quality, rather than publication bias or outcome reporting bias [138]. Consequently, the inclusion of these studies probably yielded a conservative underestimate of the true performance of robust models developed with sufficient resources. Also, our findings of subgroup analyses should be interpreted with caution. Since algorithm types were not randomly assigned across studies, these differences were observed and might be confounded by other study-level factors rather than the algorithms themselves. Therefore, these results represented associations rather than definitive causal evidence [139].

Furthermore, the translation of these models into clinical practice requires rigorous validation. The vast majority of included studies did not distinguish independent test sets and lacked valid external validation. As a result, the system suitability of the AI-based clinical decision support system was limited in clinical workflows [21]. Therefore, future research should conduct external validation of algorithms by program standardization in larger prospective cohorts and compare multiple existing models by external validation to determine the most useful one [140]. Moreover, the models can be compared with relevant clinicians' judgment, such as providing clinicians (including intensivists and neurologists) with routine clinical information. Additional metrics can also be assessed, such as family acceptance of AI results, resource utilization and planning, and adjustments to subsequent medical plans [141]. Additionally, these algorithms still need improvement to identify mortality in all at-risk populations (low-, medium-, and high-risk populations), and they should be trained by a large number of outlier samples or oversampling techniques, such as random or boundary oversampling or adaptive synthetic sampling. These approaches will facilitate the critical deployment of AI in stroke mortality prediction.

Advantages and Limitations

Advantages

This meta-analysis first quantitatively reviewed the value of ML in predicting in-hospital and out-of-hospital mortality in stroke. The findings are expected to offer an accurate framework for poststroke mortality risk assessment and clinical decision-making.

Limitations

Some limitations are worth noting. First, this meta-analysis exhibited great heterogeneity, which can be attributed to the variations in algorithms and study populations. Few studies reporting acute-phase vascular recanalization were included, and most of the studies did not classify the treatment background in detail, which may affect the overall performance assessment. Also, data mostly came from available cohort studies or conventional registries, and some were single-center studies, introducing some bias and restricting the accuracy. Therefore, prospective multicenter studies with detailed subgroup analyses (including treatment backgrounds) are needed. Second, outcome metrics were missing in a small number of studies, so reasonable estimation was implemented, but the model's predictive accuracy estimates were still affected. Third, internal validation was used in the majority of the studies and external validation in only 20% of the studies, weakening the model's extrapolation capability. Fourth, multiple studies included were from a small number of large public databases, potentially leading to cohort overlap. Although the limited impact of potential cohort overlap on the pooled effect was confirmed by the SE analysis, we recognized that this method could not fully eliminate residual overfitting bias, given that such bias may arise from systematic similarities in patient populations and clinical protocols across these databases [142]. This concealed overlap could still inflate the summary effect estimates and complicate the interpretation of heterogeneity. Therefore, the extrapolation of our conclusions to more heterogeneous clinical settings warrants caution.

Conclusions

ML demonstrates promising performance in stroke mortality prediction, and some ML models are more valuable in predicting long-term mortality. However, more efforts are needed before application in clinical practice. For example, we should focus on population SP, search for more specific biomarkers, introduce more multimodal data, optimize model balance, and increase external validations. Furthermore, ML models should be used to develop prediction tools for stroke mortality risk applicable in clinical settings to advance their translation into clinical practice. In this way, data-based doctor-patient shared decision-making can be facilitated, and more efficient resource

XSL•FO
RenderX

allocation can be achieved in health care institutions, ultimately ameliorating patient outcomes.

Acknowledgments

The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript.

Funding

This study was funded by the Regional Innovation Cooperation Project of the Science and Technology Department of Sichuan Province (2024YFHZ0142), Sichuan Science and Technology Program (no. 2023YFQ0109), and Health Commission of Sichuan Province Medical Science and Technology Program (2023-109 and 24LCYJPT08), as well as the National Key R&D Program of China (no. 2022YFC2503805). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

All data generated or analyzed during this study are included in this published paper and its supplementary information files.

Authors' Contributions

All authors contributed to the study conception and design. YC and ZO participated in writing – original draft preparation, and writing – review and editing. YD contributed to conceptualization. YC and YY contributed to methodology. YC, ZO, YD, AL, XL, YY, and QY conducted the formal analysis and investigation. MZ and XW participated in supervision. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 97 KB - jmir_v28i1e83821_app1.pdf]

Multimedia Appendix 2
Details of the search strategy.
[DOCX File , 3381 KB - jmir_v28i1e83821_app2.docx]

Multimedia Appendix 3
Results of Prediction model Risk of Bias Assessment Tool assessment of risk of bias in included studies.
[PNG File , 57 KB - jmir_v28i1e83821_app3.png]

References

1. Hilkens NA, Casolla B, Leung TW, de Leeuw FE. Stroke. Lancet 2024;403(10446):2820-2836. [doi: 10.1016/S0140-6736(24)00642-1] [Medline: 38759664]
2. Feigin VL, Brainin M, Norrving B, Martins SO, Pandian J, Lindsay P, et al. World stroke organization: Global stroke fact sheet 2025. Int J Stroke 2025;20(2):132-144 [FREE Full text] [doi: 10.1177/17474930241308142] [Medline: 39635884]
3. Pendlebury ST, Rothwell PM. Prevalence, incidence, and factors associated with pre-stroke and post-stroke dementia: a systematic review and meta-analysis. Lancet Neurol 2009;8(11):1006-1018. [doi: 10.1016/S1474-4422(09)70236-4] [Medline: 19782001]
4. Feigin VL, Owolabi MO. Pragmatic solutions to reduce the global burden of stroke: a world stroke organization-lancet neurology commission. Lancet Neurol 2023;22(12):1160-1206. [doi: 10.1016/S1474-4422(23)00277-6] [Medline: 37827183]
5. Global, regional, and national burden of stroke and its risk factors, 1990-2021. A systematic analysis for the global burden of disease study 2021. Lancet Neurol 2024;23(10):973-1003 [FREE Full text] [doi: 10.1016/S1474-4422(24)00369-7] [Medline: 39304265]
6. Gu HQ, Yang X, Wang CJ, Zhao XQ, Wang YL, Liu LP, et al. Clinical characteristics, management, and in-hospital outcomes in patients with stroke or transient ischemic attack in china. JAMA Netw Open 2021;4(8):e2120745 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.20745] [Medline: 34387677]

XSL•FO
RenderX

7. Wang M, Wang CJ, Gu HQ, Meng X, Jiang Y, Yang X, et al. Sex differences in short-term and long-term outcomes among patients with acute ischemic stroke in china. Stroke 2022;53(7):2268-2275 [FREE Full text] [doi: 10.1161/STROKEAHA.121.037121] [Medline: 35130717]
8. Tu WJ, Chao BH, Ma L, Yan F, Cao L, Qiu H, et al. Case-fatality, disability and recurrence rates after first-ever stroke: a study from bigdata observatory platform for stroke of China. Brain Res Bull 2021;175:130-135. [doi: 10.1016/j.brainresbull.2021.07.020] [Medline: 34329730]
9. Goyal M, Menon BK, van Zwam WH, Dippel DWJ, Mitchell PJ, Demchuk AM, HERMES Collaborators. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. Lancet 2016;387(10029):1723-1731. [doi: 10.1016/S0140-6736(16)00163-X] [Medline: 26898852]
10. Tu WJ, Wang LD. China stroke surveillance report 2021. Mil Med Res 2023;10(1):33 [FREE Full text] [doi: 10.1186/s40779-023-00463-x] [Medline: 37468952]
11. van Asch CJ, Luitse MJ, Rinkel GJ, van der Tweel I, Algra A, Klijn CJ. Incidence, case fatality, and functional outcome of intracerebral haemorrhage over time, according to age, sex, and ethnic origin: a systematic review and meta-analysis. Lancet Neurol 2010;9(2):167-176. [doi: 10.1016/S1474-4422(09)70340-0] [Medline: 20056489]
12. Wu S, Wu B, Liu M, Chen Z, Wang W, Anderson CS, et al. Stroke in China: advances and challenges in epidemiology, prevention, and management. Lancet Neurol 2019;18(4):394-405. [doi: 10.1016/S1474-4422(18)30500-3] [Medline: 30878104]
13. Montellano FA, Ungethüm K, Ramiro L, Nacu A, Hellwig S, Fluri F, et al. Role of blood-based biomarkers in ischemic stroke prognosis. Stroke 2021;52(2):543-551. [doi: 10.1161/strokeaha.120.029232]
14. Kremers F, Venema E, Duvekot M, Yo L, Bokkers R, Lycklama À Nijeholt G, et al. Outcome prediction models for endovascular treatment of ischemic stroke: Systematic review and external validation. Stroke 2022;53(3):825-836 [FREE Full text] [doi: 10.1161/STROKEAHA.120.033445] [Medline: 34732070]
15. Xie P, Hu Y, Li J, Ma Y, Xiao J. Unlocking the potential of real-time ICU mortality prediction: redefining risk assessment with continuous data recovery. NPJ Digit Med 2025;8(1):733 [FREE Full text] [doi: 10.1038/s41746-025-02114-y] [Medline: 41315868]
16. Solomon OM, Gu Y, Wei Y. Risk factors for enteral nutrition feeding intolerance in stroke patients: a scoping review. Nurs Crit Care 2025;30(3):e13183. [doi: 10.1111/nicc.13183] [Medline: 39428862]
17. Yang Y, Huang X, Wang Y, Leng L, Xu J, Feng L, et al. The impact of triglyceride-glucose index on ischemic stroke: a systematic review and meta-analysis. Cardiovasc Diabetol 2023;22(1):2 [FREE Full text] [doi: 10.1186/s12933-022-01732-0] [Medline: 36609319]
18. Efstathiou SP, Tsioulos DI, Tsiakou AG, Gratsias YE, Pefanis AV, Mountokalakis TD. Plasma adiponectin levels and five-year survival after first-ever ischemic stroke. Stroke 2005;36(9):1915-1919. [doi: 10.1161/01.STR.0000177874.29849.f0] [Medline: 16109902]
19. Cecconi M, Greco M, Shickel B, Vincent JL, Bihorac A. Artificial intelligence in acute medicine: a call to action. Crit Care 2024;28(1):258. [doi: 10.1186/s13054-024-05034-7] [Medline: 39075468]
20. Bonkhoff AK, Grefkes C. Precision medicine in stroke: towards personalized outcome predictions using artificial intelligence. Brain 2022;145(2):457-475 [FREE Full text] [doi: 10.1093/brain/awab439] [Medline: 34918041]
21. Akay EMZ, Hilbert A, Carlisle BG, Madai VI, Mutke MA, Frey D. Artificial intelligence for clinical decision support in acute ischemic stroke: a systematic review. Stroke 2023;54(6):1505-1516 [FREE Full text] [doi: 10.1161/STROKEAHA.122.041442] [Medline: 37216446]
22. Zhang J, Wu J, Zhou XS, Shi F, Shen D. Recent advancements in artificial intelligence for breast cancer: image augmentation, segmentation, diagnosis, and prognosis approaches. Semin Cancer Biol 2023;96:11-25. [doi: 10.1016/j.semcancer.2023.09.001] [Medline: 37704183]
23. Borchert RJ, Azevedo T, Badhwar A, Bernal J, Betts M, Bruffaerts R, et al. Artificial intelligence for diagnostic and prognostic neuroimaging in dementia: a systematic review. Alzheimers Dement 2023;19(12):5885-5904. [doi: 10.1002/alz.13412] [Medline: 37563912]
24. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
25. Radley NM, Soh I, Saad AM, Wickramarachchi M, Dawson A, Hin JNC, et al. Risk of bias assessment of post-stroke mortality machine learning predictive models: systematic review. J Stroke Cerebrovasc Dis 2025;34(6):108291 [FREE Full text] [doi: 10.1016/j.jstrokecerebrovasdis.2025.108291] [Medline: 40089217]
26. Liu W, Ma W, Bai N, Li C, Liu K, Yang J, et al. Identification of key predictors of hospital mortality in critically ill patients with embolic stroke using machine learning. Biosci Rep 2022;42(9):BSR20220995 [FREE Full text] [doi: 10.1042/BSR20220995] [Medline: 35993194]
27. Kurtz P, Peres IT, Soares M, Salluh JIF, Bozza FA. Hospital length of stay and 30-day mortality prediction in stroke: a machine learning analysis of 17,000 ICU admissions in Brazil. Neurocrit Care 2022;37(Suppl 2):313-321. [doi: 10.1007/s12028-022-01486-3] [Medline: 35381967]

XSL•FO
RenderX

28. Kniep HC, Elsayed S, Nawabi J, Broocks G, Meyer L, Bechstein M, et al. Imaging-based outcome prediction in posterior circulation stroke. J Neurol 2022;269(7):3800-3809 [FREE Full text] [doi: 10.1007/s00415-022-11010-4] [Medline: 35257203]
29. Bray BD, Campbell J, Cloud GC, Hoffman A, James M, Tyrrell PJ, et al. Derivation and external validation of a case mix model for the standardized reporting of 30-day stroke mortality rates. Stroke 2014;45(11):3374-3380. [doi: 10.1161/STROKEAHA.114.006451] [Medline: 25293667]
30. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
31. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
32. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019;170(1):W1-W33 [FREE Full text] [doi: 10.7326/M18-1377] [Medline: 30596876]
33. Debray TP, Damen JA, Riley RD, Snell K, Reitsma JB, Hooft L, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Stat Methods Med Res 2019;28(9):2768-2786 [FREE Full text] [doi: 10.1177/0962280218785504] [Medline: 30032705]
34. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods 2010;1(2):97-111 [FREE Full text] [doi: 10.1002/jrsm.12] [Medline: 26061376]
35. Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ 2017;356:i6460 [FREE Full text] [doi: 10.1136/bmj.i6460] [Medline: 28057641]
36. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
37. Röver C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. BMC Med Res Methodol 2015;15:99 [FREE Full text] [doi: 10.1186/s12874-015-0091-1] [Medline: 26573817]
38. Partlett C, Riley RD. Random effects meta-analysis: coverage performance of 95% confidence and prediction intervals following REML estimation. Stat Med 2017;36(2):301-317 [FREE Full text] [doi: 10.1002/sim.7140] [Medline: 27714841]
39. Damen JAA, Moons KGM, van Smeden M, Hooft L. How to conduct a systematic review and meta-analysis of prognostic model studies. Clin Microbiol Infect 2023;29(4):434-440 [FREE Full text] [doi: 10.1016/j.cmi.2022.07.019] [Medline: 35934199]
40. Norouzi S, Ahmadi S, Alinia S, Farzipoor F, Shahsavari A, Hajizadeh E, et al. Machine learning predictive models for survival in patients with brain stroke. Health Promot Perspect 2025;15(1):63-72. [doi: 10.34172/hpp.025.43635] [Medline: 40453683]
41. Zhou Z, Chen B, Mei Z, Chen W, Cao W, Xu E, et al. Machine learning-based scoring model for predicting mortality in ICU-admitted ischemic stroke patients with moderate to severe consciousness disorders. Front Neurol 2025;16:1534961 [FREE Full text] [doi: 10.3389/fneur.2025.1534961] [Medline: 40170899]
42. Xu L, Li C, Zhang J, Guan C, Zhao L, Shen X, et al. Personalized prediction of mortality in patients with acute ischemic stroke using explainable artificial intelligence. Eur J Med Res 2024;29(1):341 [FREE Full text] [doi: 10.1186/s40001-024-01940-2] [Medline: 38902792]
43. Wang W, Rudd AG, Wang Y, Curcin V, Wolfe CD, Peek N, et al. Risk prediction of 30-day mortality after stroke using machine learning: a nationwide registry-based cohort study. BMC Neurol 2022;22(1):195 [FREE Full text] [doi: 10.1186/s12883-022-02722-1] [Medline: 35624434]
44. Wang K, Gu L, Liu W, Xu C, Yin C, Liu H, et al. The predictors of death within 1 year in acute ischemic stroke patients based on machine learning. Front Neurol 2023;14:1092534 [FREE Full text] [doi: 10.3389/fneur.2023.1092534] [Medline: 36908612]
45. Someeh N, Mirfeizi M, Asghari-Jafarabadi M, Alinia S, Farzipoor F, Shamshirgaran SM. Predicting mortality in brain stroke patients using neural networks: outcomes analysis in a longitudinal study. Sci Rep 2023;13(1):18530 [FREE Full text] [doi: 10.1038/s41598-023-45877-8] [Medline: 37898678]
46. Scrutinio D, Ricciardi C, Donisi L, Losavio E, Battista P, Guida P, et al. Machine learning to predict mortality after rehabilitation among patients with severe stroke. Sci Rep 2020;10(1):20127 [FREE Full text] [doi: 10.1038/s41598-020-77243-3] [Medline: 33208913]
47. Petrović I, Broggi S, Killer-Oberpfalzer M, Pfaff JAR, Griessenauer CJ, Milosavljević I, et al. Predictors of in-hospital mortality after thrombectomy in anterior circulation large vessel occlusion: a retrospective, machine learning study. Diagnostics (Basel) 2024;14(14):1531 [FREE Full text] [doi: 10.3390/diagnostics14141531] [Medline: 39061668]
48. Cummins JA, Gerber BS, Fukunaga MI, Henninger N, Kiefe CI, Liu F. In-hospital mortality prediction among intensive care unit patients with acute ischemic stroke: a machine learning approach. Health Data Sci 2025;5:0179 [FREE Full text] [doi: 10.34133/hds.0179] [Medline: 40099281]

XSL•FO
RenderX

49. Park E, Chang HJ, Nam HS. A bayesian network model for predicting post-stroke outcomes with available risk factors. Front Neurol 2018;9:699 [FREE Full text] [doi: 10.3389/fneur.2018.00699] [Medline: 30245663]
50. Ouyang Y, Cheng M, He B, Zhang F, Ouyang W, Zhao J, et al. Interpretable machine learning models for predicting in-hospital death in patients in the intensive care unit with cerebral infarction. Comput Methods Programs Biomed 2023;231:107431. [doi: 10.1016/j.cmpb.2023.107431] [Medline: 36827826]
51. Luo X, Li B, Zhu R, Tai Y, Wang Z, He Q, et al. Development and validation of an interpretable machine learning model for predicting in-hospital mortality for ischemic stroke patients in ICU. Int J Med Inform 2025;198:105874. [doi: 10.1016/j.ijmedinf.2025.105874] [Medline: 40073651]
52. Mai TD, Nguyen DT, Tran CC, Duong HQ, Nguyen HN, Dang DP, et al. Factors associated with 90-day mortality in Vietnamese stroke patients: prospective findings compared with explainable machine learning, multicenter study. PLoS One 2024;19(9):e0310522 [FREE Full text] [doi: 10.1371/journal.pone.0310522] [Medline: 39302916]
53. Lehmann ALCF, Alfieri DF, de Araújo MCM, Trevisani ER, Nagao MR, Pesente FS, et al. Immune-inflammatory, coagulation, adhesion, and imaging biomarkers combined in machine learning models improve the prediction of death 1 year after ischemic stroke. Clin Exp Med 2022;22(1):111-123. [doi: 10.1007/s10238-021-00732-w] [Medline: 34120242]
54. Kim C, Kwon JM, Lee J, Jo H, Gwon D, Jang JH, et al. Deep learning model integrating radiologic and clinical data to predict mortality after ischemic stroke. Heliyon 2024;10(10):e31000 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e31000] [Medline: 38826743]
55. Huang R, Liu J, Wan TK, Siriwanna D, Woo YMP, Vodencarevic A, et al. Stroke mortality prediction based on ensemble learning and the combination of structured and textual data. Comput Biol Med 2023;155:106176. [doi: 10.1016/j.compbiomed.2022.106176] [Medline: 36805232]
56. Huang J, Liu X, Jin W. Clinical decision support systems for 3-month mortality in elderly patients admitted to ICU with ischemic stroke using interpretable machine learning. Digit Health 2024;10:20552076241280126 [FREE Full text] [doi: 10.1177/20552076241280126] [Medline: 39314817]
57. Huang J, Jin W, Duan X, Liu X, Shu T, Fu L, et al. Twenty-eight-day in-hospital mortality prediction for elderly patients with ischemic stroke in the intensive care unit: interpretable machine learning models. Front Public Health 2022;10:1086339 [FREE Full text] [doi: 10.3389/fpubh.2022.1086339] [Medline: 36711330]
58. Huang J, Chen H, Deng J, Liu X, Shu T, Yin C, et al. Interpretable machine learning for predicting 28-day all-cause in-hospital mortality for hypertensive ischemic or hemorrhagic stroke patients in the ICU: a multi-center retrospective cohort study with internal and external cross-validation. Front Neurol 2023;14:1185447 [FREE Full text] [doi: 10.3389/fneur.2023.1185447] [Medline: 37614971]
59. Hoffman H, Wood J, Cote JR, Jalal MS, Otite FO, Masoud HE, et al. Development and internal validation of machine learning models to predict mortality and disability after mechanical thrombectomy for acute anterior circulation large vessel occlusion. World Neurosurg 2024;182:e137-e154. [doi: 10.1016/j.wneu.2023.11.060] [Medline: 38000670]
60. Fernandez-Lozano C, Hervella P, Mato-Abad V, Rodríguez-Yáñez M, Suárez-Garaboa S, López-Dequidt I, et al. Random forest-based prediction of stroke outcome. Sci Rep 2021;11(1):10071 [FREE Full text] [doi: 10.1038/s41598-021-89434-7] [Medline: 33980906]
61. Fang J, Song B, Li L, Tong L, Jiang M, Yan J. RGX ensemble model for advanced prediction of mortality outcomes in stroke patients. BME Front 2024;5:0077 [FREE Full text] [doi: 10.34133/bmef.0077] [Medline: 39600589]
62. Chunjuan Z, Yulong W, Xicheng Z, Xiaodong M. Machine learning consensus clustering for inflammatory subtype analysis in stroke and its impact on mortality risk: a study based on NHANES (1999-2018). Front Neurol 2025;16:1562247 [FREE Full text] [doi: 10.3389/fneur.2025.1562247] [Medline: 40276469]
63. Chen Y, Yang Z, Liu Y, Li Y, Zhong Z, McDowell G, et al. Exploring the prognostic impact of triglyceride-glucose index in critically ill patients with first-ever stroke: insights from traditional methods and machine learning-based mortality prediction. Cardiovasc Diabetol 2024;23(1):443 [FREE Full text] [doi: 10.1186/s12933-024-02538-y] [Medline: 39695656]
64. Chen H, Yang F, Duan Y, Yang L, Li J. A novel higher performance nomogram based on explainable machine learning for predicting mortality risk in stroke patients within 30 days based on clinical features on the first day ICU admission. BMC Med Inform Decis Mak 2024;24(1):161 [FREE Full text] [doi: 10.1186/s12911-024-02547-7] [Medline: 38849903]
65. Che Nawi CMNH, Mohd Hairon S, Wan Yahya WNN, Wan Zaidi WA, Musa KI. Machine learning models for predicting stroke mortality in malaysia: an application and comparative analysis. Cureus 2023;15(12):e50426 [FREE Full text] [doi: 10.7759/cureus.50426] [Medline: 38222138]
66. Abujaber A, Yaseen S, Imam Y, Nashwan A, Akhtar N. Machine learning-based prediction of one-year mortality in ischemic stroke patients. Oxf Open Neurosci 2024;3:kvae011. [doi: 10.1093/oons/kvae011] [Medline: 39569400]
67. Abedi V, Avula V, Razavi SM, Bavishi S, Chaudhary D, Shahjouei S, et al. Predicting short and long-term mortality after acute ischemic stroke using EHR. J Neurol Sci 2021;427:117560 [FREE Full text] [doi: 10.1016/j.jns.2021.117560] [Medline: 34218182]
68. Zhou L, Wu Y, Wang J, Wu H, Tan Y, Chen X, et al. Development of a predictive nomogram for intra-hospital mortality in acute ischemic stroke patients using LASSO regression. Clin Interv Aging 2024;19:1423-1436 [FREE Full text] [doi: 10.2147/CIA.S471885] [Medline: 39139210]

XSL•FO
RenderX

69. Zhang X, Yuan K, Wang H, Gong P, Jiang T, Xie Y, et al. Nomogram to predict mortality of endovascular thrombectomy for ischemic stroke despite successful recanalization. J Am Heart Assoc 2020;9(3):e014899 [FREE Full text] [doi: 10.1161/JAHA.119.014899] [Medline: 31973604]
70. Zhang RX, Zhang WW, Luo YT, Liu GW. An mNUTRIC-based nomogram for predicting the in-hospital death risk in patients with acute stroke. Eur J Clin Nutr 2022;76(10):1464-1469. [doi: 10.1038/s41430-022-01127-0] [Medline: 35411025]
71. Zhang J, Zhuang S, Zhang Y, An L, Li D, Liu Z, et al. Ultrasonic optic nerve sheath diameter as a new predictor for the mortality of patients with large hemispheric infarction. Sci Rep 2025;15(1):460 [FREE Full text] [doi: 10.1038/s41598-024-84720-6] [Medline: 39748001]
72. Zeng M, Smith L, Bird A, Trinh VQ, Bacchi S, Harvey J, et al. Predictions for functional outcome and mortality in acute ischaemic stroke following successful endovascular thrombectomy. BMJ Neurol Open 2024;6(1):e000707 [FREE Full text] [doi: 10.1136/bmjno-2024-000707] [Medline: 38932996]
73. Çelik G, Baykan Ö, Kara Y, Tireli H. Predicting 10-day mortality in patients with strokes using neural networks and multivariate statistical methods. J Stroke Cerebrovasc Dis 2014;23(6):1506-1512. [doi: 10.1016/j.jstrokecerebrovasdis.2013.12.018] [Medline: 24674954]
74. Wen R, Wang M, Bian W, Zhu H, Xiao Y, He Q, et al. Nomogram to predict 6-month mortality in acute ischemic stroke patients treated with endovascular treatment. Front Neurol 2023;14:1330959 [FREE Full text] [doi: 10.3389/fneur.2023.1330959] [Medline: 38249750]
75. Szlachetka WA, Pana TA, Mamas MA, Bettencourt-Silva JH, Metcalf AK, Potter JF, et al. Predicting 10-year stroke mortality: development and validation of a nomogram. Acta Neurol Belg 2022;122(3):685-693 [FREE Full text] [doi: 10.1007/s13760-021-01752-9] [Medline: 34406610]
76. Sun W, Li G, Liu Z, Miao J, Yang Z, Zhou Q, et al. A nomogram for predicting the in-hospital mortality after large hemispheric infarction. BMC Neurol 2019;19(1):347 [FREE Full text] [doi: 10.1186/s12883-019-1571-4] [Medline: 31884967]
77. Solberg OG, Dahl M, Mowinckel P, Stavem K. Derivation and validation of a simple risk score for predicting 1-year mortality in stroke. J Neurol 2007;254(10):1376-1383. [doi: 10.1007/s00415-007-0555-2] [Medline: 17934885]
78. Sha L, Xu T, Ge X, Shi L, Zhang J, Guo H. Predictors of death within 6 months of stroke onset: a model with barthel index, platelet/lymphocyte ratio and serum albumin. Nurs Open 2021;8(3):1380-1392 [FREE Full text] [doi: 10.1002/nop2.754] [Medline: 33378600]
79. Ni H, Zhao L, Ma X, Lv P, Ding Y, Liu Z, et al. Technical risk stratification nomogram model for 90-day mortality prediction in patients with acute basilar artery occlusion undergoing endovascular thrombectomy: a multicenter cohort study. J Am Heart Assoc 2024;13(6):e032107 [FREE Full text] [doi: 10.1161/JAHA.123.032107] [Medline: 38471827]
80. Liljehult J, Christensen T, Christensen KB. Early prediction of one-year mortality in ischemic and haemorrhagic stroke. J Stroke Cerebrovasc Dis 2020;29(4):104667. [doi: 10.1016/j.jstrokecerebrovasdis.2020.104667] [Medline: 32044222]
81. Li XD, Li MM. A novel nomogram to predict mortality in patients with stroke: a survival analysis based on the MIMIC-III clinical database. BMC Med Inform Decis Mak 2022;22(1):92 [FREE Full text] [doi: 10.1186/s12911-022-01836-3] [Medline: 35387672]
82. Kim TJ, Lee JS, Oh MS, Kim JW, Park SH, Yu KH, et al. Stroke measures analysis of pRognostic testing-mortality nomogram predicts long-term mortality after ischemic stroke. Int J Stroke 2025;20(1):75-84. [doi: 10.1177/17474930241278808] [Medline: 39150095]
83. Lee J, Morishima T, Kunisawa S, Sasaki N, Otsubo T, Ikai H, et al. Derivation and validation of in-hospital mortality prediction models in ischaemic stroke patients using administrative data. Cerebrovasc Dis 2013;35(1):73-80. [doi: 10.1159/000346090] [Medline: 23429000]
84. Jin G, Hu W, Zeng L, Ma B, Zhou M. Prediction of long-term mortality in patients with ischemic stroke based on clinical characteristics on the first day of ICU admission: an easy-to-use nomogram. Front Neurol 2023;14:1148185 [FREE Full text] [doi: 10.3389/fneur.2023.1148185] [Medline: 37122313]
85. Jin G, Hu W, Zeng L, Diao M, Chen H, Chen J, et al. Development and verification of a nomogram for predicting short-term mortality in elderly ischemic stroke populations. Sci Rep 2023;13(1):12580 [FREE Full text] [doi: 10.1038/s41598-023-39781-4] [Medline: 37537270]
86. Gao H, Sun X, Li W, Gao Q, Zhang J, Zhang Y, et al. Development and validation of a risk score to predict 30-day mortality in patients with atrial fibrillation-related stroke: GPS-GF score. Neurol Res 2018;40(7):532-540. [doi: 10.1080/01616412.2018.1451431] [Medline: 29544401]
87. Fang L, Zhou M, Mao F, Diao M, Hu W, Jin G. Development and validation of a nomogram for predicting 28-day mortality in patients with ischemic stroke. PLoS One 2024;19(4):e0302227 [FREE Full text] [doi: 10.1371/journal.pone.0302227] [Medline: 38656987]
88. Cappellari M, Zivelonghi C, Turcato G, Forlivesi S, Micheletti N, Tomelleri G, et al. A nomogram to predict the probability of mortality after first-ever acute manifestations of cerebral small vessel disease. J Neurol Sci 2018;385:92-95. [doi: 10.1016/j.jns.2017.12.020] [Medline: 29406922]

XSL•FO
RenderX

89. Schwartz J, Wang Y, Qin L, Schwamm LH, Fonarow GC, Cormier N, et al. Incorporating stroke severity into hospital measures of 30-day mortality after ischemic stroke hospitalization. Stroke 2017;48(11):3101-3107. [doi: 10.1161/STROKEAHA.117.017960] [Medline: 28954922]
90. Reitzle L, Rohmann JL, Kurth T, Audebert HJ, Piccininni M. External validation of risk prediction models for post-stroke mortality in Berlin. BMJ Open 2025;15(6):e089320 [FREE Full text] [doi: 10.1136/bmjopen-2024-089320] [Medline: 40480675]
91. Hwangbo L, Kang YJ, Kwon H, Lee JI, Cho H, Ko J, et al. Stacking ensemble learning model to predict 6-month mortality in ischemic stroke patients. Sci Rep 2022;12(1):17389 [FREE Full text] [doi: 10.1038/s41598-022-22323-9] [Medline: 36253488]
92. Huang Y, Douiri A, Fahey M. A dynamic model for predicting survival up to 1 year after ischemic stroke. J Stroke Cerebrovasc Dis 2020;29(10):105133 [FREE Full text] [doi: 10.1016/j.jstrokecerebrovasdis.2020.105133] [Medline: 32912566]
93. Geng L, Zuo L, Li H, Wang Y, Zhang Q, Ran J, et al. Association of platelet-to-neutrophil ratios with 1-year outcome and mortality in patients with acute ischemic stroke. Neurosci Lett 2023;798:137016. [doi: 10.1016/j.neulet.2022.137016] [Medline: 36529389]
94. Zhang S, Wang J, Pei L, Liu K, Gao Y, Fang H, et al. Interpretability analysis of one-year mortality prediction for stroke patients based on deep neural network. IEEE J Biomed Health Inform 2022;26(4):1903-1910. [doi: 10.1109/JBHI.2021.3123657] [Medline: 34714758]
95. Abujaber AA, Albalkhi I, Imam Y, Nashwan A, Akhtar N, Alkhawaldeh IM. Machine learning-based prognostication of mortality in stroke patients. Heliyon 2024;10(7):e28869 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e28869] [Medline: 38601648]
96. Zhu E, Chen Z, Ai P, Wang J, Zhu M, Xu Z, et al. Analyzing and predicting the risk of death in stroke patients using machine learning. Front Neurol 2023;14:1096153 [FREE Full text] [doi: 10.3389/fneur.2023.1096153] [Medline: 36816575]
97. Chang Y, Zhang L, Li Y, Wang D, Fang Q, Tang X. Derivation and validation of a new visceral adiposity index for predicting short-term mortality of patients with acute ischemic stroke in a chinese population. Brain Sci 2023;13(2):297 [FREE Full text] [doi: 10.3390/brainsci13020297] [Medline: 36831840]
98. Yan X, Xia P, Tong H, Lan C, Wang Q, Zhou Y, et al. Development and validation of a dynamic nomogram for predicting 3-month mortality in acute ischemic stroke patients with atrial fibrillation. Risk Manag Healthc Policy 2024;17:145-158 [FREE Full text] [doi: 10.2147/RMHP.S442353] [Medline: 38250220]
99. Wang W, Otieno JA, Eriksson M, Wolfe CD, Curcin V, Bray BD. Developing and externally validating a machine learning risk prediction model for 30-day mortality after stroke using national stroke registers in the UK and Sweden. BMJ Open 2023;13(11):e069811 [FREE Full text] [doi: 10.1136/bmjopen-2022-069811] [Medline: 37968001]
100. Easton JF, Stephens CR, Angelova M. Risk factors and prediction of very short term versus short/intermediate term post-stroke mortality: a data mining approach. Comput Biol Med 2014;54:199-210. [doi: 10.1016/j.compbiomed.2014.09.003] [Medline: 25303114]
101. Lea-Pereira MC, Amaya-Pascasio L, Martínez-Sánchez P, Rodríguez Salvador MDM, Galván-Espinosa J, Téllez-Ramírez L, et al. Predictive model and mortality risk score during admission for ischaemic stroke with conservative treatment. Int J Environ Res Public Health 2022;19(6):3182 [FREE Full text] [doi: 10.3390/ijerph19063182] [Medline: 35328867]
102. Süt N, Celik Y. Prediction of mortality in stroke patients using multilayer perceptron neural networks. Turk J Med Sci 2011:886-893. [doi: 10.3906/sag-1105-20]
103. Gattringer T, Posekany A, Niederkorn K, Knoflach M, Poltrum B, Mutzenbach S, et al. Predicting early mortality of acute ischemic stroke. Stroke 2019;50(2):349-356. [doi: 10.1161/strokeaha.118.022863]
104. Yang Y, Tang L, Deng Y, Li X, Luo A, Zhang Z, et al. The predictive performance of artificial intelligence on the outcome of stroke: a systematic review and meta-analysis. Front Neurosci 2023;17:1256592 [FREE Full text] [doi: 10.3389/fnins.2023.1256592] [Medline: 37746141]
105. Schwartz L, Anteby R, Klang E, Soffer S. Stroke mortality prediction using machine learning: systematic review. J Neurol Sci 2023;444:120529. [doi: 10.1016/j.jns.2022.120529] [Medline: 36580703]
106. Wang W, Kiik M, Peek N, Curcin V, Marshall IJ, Rudd AG, et al. A systematic review of machine learning models for predicting outcomes of stroke with structured data. PLoS One 2020;15(6):e0234722 [FREE Full text] [doi: 10.1371/journal.pone.0234722] [Medline: 32530947]
107. Yu Y, Zhong Z, Yang W, Yu J, Li J, Guo X, et al. Neutrophil percentage-to-albumin ratio and risk of mortality in patients on peritoneal dialysis. J Inflamm Res 2023;16:6271-6281 [FREE Full text] [doi: 10.2147/JIR.S437256] [Medline: 38146321]
108. Chung J, Ahn J, Ryu JA. Beyond SOFA and APACHE II, novel risk stratification models using readily available biomarkers in critical care. Diagnostics (Basel) 2025;15(9):1122 [FREE Full text] [doi: 10.3390/diagnostics15091122] [Medline: 40361939]
109. Lin J, Frontera JA. Decompressive hemicraniectomy for large hemispheric strokes. Stroke 2021;52(4):1500-1510 [FREE Full text] [doi: 10.1161/STROKEAHA.120.032359] [Medline: 33719518]
110. Maddox TM, Rumsfeld JS, Payne PRO. Questions for artificial intelligence in health care. JAMA 2019;321(1):31-32. [doi: 10.1001/jama.2018.18932] [Medline: 30535130]

XSL•FO
RenderX

111. Lewis AE, Weiskopf N, Abrams ZB, Foraker R, Lai AM, Payne PRO, et al. Electronic health record data quality assessment and tools: a systematic review. J Am Med Inform Assoc 2023;30(10):1730-1740 [FREE Full text] [doi: 10.1093/jamia/ocad120] [Medline: 37390812]
112. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004] [Medline: 30763612]
113. Oka S, Yamazaki T, Takefuji Y. Addressing bias in machine learning feature importance for food quality assessment. Food Chem 2025;494:146171 [FREE Full text] [doi: 10.1016/j.foodchem.2025.146171] [Medline: 40902562]
114. Allgaier J, Mulansky L, Draelos RL, Pryss R. How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. Artif Intell Med 2023;143:102616 [FREE Full text] [doi: 10.1016/j.artmed.2023.102616] [Medline: 37673561]
115. Zabor EC, Reddy CA, Tendulkar RD, Patil S. Logistic regression in clinical studies. Int J Radiat Oncol Biol Phys 2022;112(2):271-277. [doi: 10.1016/j.ijrobp.2021.08.007] [Medline: 34416341]
116. Yang J, Zeng S, Cui S, Zheng J, Wang H. Predictive modeling of acute respiratory distress syndrome using machine learning: systematic review and meta-analysis. J Med Internet Res 2025;27:e66615 [FREE Full text] [doi: 10.2196/66615] [Medline: 40359510]
117. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform 2008;77(2):81-97 [FREE Full text] [doi: 10.1016/j.ijmedinf.2006.11.006] [Medline: 17188928]
118. La Cava WG, Lee PC, Ajmal I, Ding X, Solanki P, Cohen JB, et al. A flexible symbolic regression method for constructing interpretable clinical prediction models. NPJ Digit Med 2023;6(1):107 [FREE Full text] [doi: 10.1038/s41746-023-00833-8] [Medline: 37277550]
119. Petch J, Di S, Nelson W. Opening the black box: the promise and limitations of explainable machine learning in cardiology. Can J Cardiol 2022;38(2):204-213 [FREE Full text] [doi: 10.1016/j.cjca.2021.09.004] [Medline: 34534619]
120. Bergomi L, Nicora G, Orlowska MA, Podrecca C, Bellazzi R, Fregosi C, et al. Which explanations do clinicians prefer? A comparative evaluation of XAI understandability and actionability in predicting the need for hospitalization. BMC Med Inform Decis Mak 2025;25(1):269 [FREE Full text] [doi: 10.1186/s12911-025-03045-0] [Medline: 40671048]
121. Nazim S, Alam MM, Rizvi SS, Mustapha JC, Hussain SS, Suud MM. Advancing malware imagery classification with explainable deep learning: a state-of-the-art approach using SHAP, LIME and grad-cam. PLoS One 2025;20(5):e0318542 [FREE Full text] [doi: 10.1371/journal.pone.0318542] [Medline: 40435333]
122. Kuo RYL, Harrison C, Curran TA, Jones B, Freethy A, Cussons D, et al. Artificial intelligence in fracture detection: a systematic review and meta-analysis. Radiology 2022;304(1):50-62 [FREE Full text] [doi: 10.1148/radiol.211785] [Medline: 35348381]
123. Khokhar PB, Gravino C, Palomba F. Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review. Artif Intell Med 2025;164:103132 [FREE Full text] [doi: 10.1016/j.artmed.2025.103132] [Medline: 40258308]
124. Ruan X, Fu S, Storlie CB, Mathis KL, Larson DW, Liu H. Real-time risk prediction of colorectal surgery-related post-surgical complications using GRU-D model. J Biomed Inform 2022;135:104202 [FREE Full text] [doi: 10.1016/j.jbi.2022.104202] [Medline: 36162805]
125. Bednarski BP, Singh AD, Zhang W, Jones WM, Naeim A, Ramezani R. Temporal convolutional networks and data rebalancing for clinical length of stay and mortality prediction. Sci Rep 2022;12(1):21247 [FREE Full text] [doi: 10.1038/s41598-022-25472-z] [Medline: 36481828]
126. Shurrab S, Guerra-Manzanares A, Magid A, Piechowski-Jozwiak B, Atashzar SF, Shamout FE. Multimodal machine learning for stroke prognosis and diagnosis: a systematic review. IEEE J Biomed Health Inform 2024;28(11):6958-6973. [doi: 10.1109/JBHI.2024.3448238] [Medline: 39172620]
127. Armoundas AA, Narayan SM, Arnett DK, Spector-Bagdady K, Bennett DA, Celi LA, et al. Use of artificial intelligence in improving outcomes in heart disease: a scientific statement from the American Heart Association. Circulation 2024;149(14):e1028-e1050 [FREE Full text] [doi: 10.1161/CIR.0000000000001201] [Medline: 38415358]
128. Mansournia MA, Etminan M, Danaei G, Kaufman JS, Collins G. Handling time varying confounding in observational research. BMJ 2017;359:j4587. [doi: 10.1136/bmj.j4587] [Medline: 29038130]
129. Braun RG, Edwardson MA, Aldridge CM, Cramer SC, Cruchaga C, Engelter ST, et al. Blood biospecimen recommendations for research on stroke outcomes and recovery. Stroke 2025;56(10):e279-e290. [doi: 10.1161/STROKEAHA.125.050793] [Medline: 40982604]
130. Fisher LD, Lin DY. Time-dependent covariates in the cox proportional-hazards regression model. Annu Rev Public Health 1999;20:145-157. [doi: 10.1146/annurev.publhealth.20.1.145] [Medline: 10352854]
131. Chen X, Wang L, W YYYEDCL. Mitigating data imbalance and noise: a divergence-based approach with enhanced sample selection. Neurocomputing 2024;605:1-10. [doi: 10.1016/j.neucom.2024.128269]
132. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020;368:m689 [FREE Full text] [doi: 10.1136/bmj.m689] [Medline: 32213531]

XSL•FO
RenderX

133. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol 2003;56(5):441-447. [doi: 10.1016/s0895-4356(03)00047-7] [Medline: 12812818]
134. Dubray-Vautrin A, Gravrand V, Marret G, Lamy C, Klijanienko J, Vacher S, et al. Internal validation strategy for high dimensional prognosis model: a simulation study and application to transcriptomic in head and neck tumors. Comput Struct Biotechnol J 2025;27:3792-3802 [FREE Full text] [doi: 10.1016/j.csbj.2025.08.035] [Medline: 40989918]
135. Teale EA, Forster A, Munyombwe T, Young JB. A systematic review of case-mix adjustment models for stroke. Clin Rehabil 2012;26(9):771-786. [doi: 10.1177/0269215511433068] [Medline: 22257504]
136. Peters U, Sherling HR, Chin-Yee B. Hasty generalizations and generics in medical research: a systematic review. PLoS One 2024;19(7):e0306749 [FREE Full text] [doi: 10.1371/journal.pone.0306749] [Medline: 38968284]
137. An D, Lim M, Lee S. Challenges for data quality in the clinical data life cycle: systematic review. J Med Internet Res 2025;27:e60709 [FREE Full text] [doi: 10.2196/60709] [Medline: 40266662]
138. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
139. Borenstein M, Higgins JPT. Meta-analysis and subgroups. Prev Sci 2013;14(2):134-143. [doi: 10.1007/s11121-013-0377-7] [Medline: 23479191]
140. Labarère J, Renaud B, Fine MJ. How to derive and validate clinical prediction models for use in intensive care medicine. Intensive Care Med 2014;40(4):513-527. [doi: 10.1007/s00134-014-3227-6] [Medline: 24570265]
141. Lulin Z, Yiranbon E, Asante Antwi H. Complementarity of clinician judgment and evidence based models in medical decision making: antecedents, prospects, and challenges. Biomed Res Int 2016;2016:1425693 [FREE Full text] [doi: 10.1155/2016/1425693] [Medline: 27642588]
142. Jeong S, Shivakumar M, Jung S, Won H, Nho K, Huang H, et al. Addressing overfitting bias due to sample overlap in polygenic risk scoring. Alzheimers Dement 2025;21(4):e70109. [doi: 10.1002/alz.70109] [Medline: 40189831]

Abbreviations

AI: artificial intelligence
AIS: acute ischemic stroke
AUC: area under the curve
DL: deep learning
eICU-CRD: eICU Collaborative Research Database
HKSJ: Hartung-Knapp-Sidik-Jonkman
LR: logistic regression
MIMIC: Medical Information Mart for Intensive Care
ML: machine learning
mRS: modified Rankin Scale
NIHSS: National Institutes of Health Stroke Scale
NPAR: neutrophil-percentage-to-albumin ratio
PI: prediction interval
PNR: platelet-to-neutrophil ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
PROBAST: Prediction model Risk of Bias Assessment Tool
RF: random forest
RoB: risk of bias
SE: sensitivity
SHAP: Shapley additive explanations
SP: specificity
SSNAP: Sentinel Stroke National Audit Programme
XGBoost: extreme gradient boosting

XSL•FO
RenderX

Edited by S Brini; submitted 09.Sep.2025; peer-reviewed by GI Han, JMI Arockiasamy; comments to author 31.Oct.2025; accepted 19.Feb.2026; published 02.Apr.2026.

Please cite as:
Chen Y, Ou Z, Deng Y, Luo A, Li X, Yang Y, Yang Q, Wu X, Zhou M
Predictive Value of Machine Learning for Poststroke Mortality Risk: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e83821
URL: https://www.jmir.org/2026/1/e83821
doi:10.2196/83821
PMID:

©Yujie Chen, Zhujing Ou, Yiting Deng, Anling Luo, Xuzi Li, Yujia Yang, Qin Yang, Xintong Wu, Muke Zhou. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Artificial Intelligence, Connected Care, and Enabling Digital Health Technologies in Rare Diseases With a Focus on Lysosomal Storage Disorders: Scoping Review

Alberta MC Spreafico[1,2,3], MSc, MBA, PhD; Luca Neri[4,5], MD; Kim Angel[6]; Cinzia Maria Bellettato[7], PhD; Maurizio Scarpa[7], Prof Dr Med

[1]Health Innovation, EVERSANA INTOUCH, Salerno, Italy

[2]Government, Health and Not for Profit Division, SDA Bocconi School of Management, Milano, Italy

[3]Innovation for Global Health Institute, Milan, Italy

[4]WINFOCUS, Milan, Italy

[5]Iuliu Ha ieganu University of Medicine and Pharmacy, Cluj-Napoca, Romania

[6]International MPS Network, Vancouver, BC, Canada

[7]MetabERN, Regional Coordinating Center for Rare Diseases, Udine University Hospital, Udine, Italy

Corresponding Author:
Alberta MC Spreafico, MSc, MBA, PhD
Health Innovation
EVERSANA INTOUCH
Palazzo Innovazione Piazza Abate Conforti
Salerno, 84121
Italy
Phone: 39 3386337713
Email: alberta.spreafico@eversana.com

Abstract

Background: Rare diseases affect more than 300 million people globally, and only about 5% have approved therapies. Lysosomal storage disorders (LSDs) exemplify the diagnostic and long-term care complexity typical of rare diseases, and digital health technologies (DHTs), especially artificial intelligence (AI) and connected care (CC), are emerging tools to support LSD management.

Objective: We aimed to map and synthesize peer-reviewed and gray literature from the past decade on DHTs relevant for LSD care, with a primary analytic focus on AI-enabled and CC solutions and a contextual mapping of other enabling DHTs. Evidence distribution was charted by population, care-journey phase, and outcome domains to identify gaps, methodological limitations, and timely priorities relevant for research, clinical practice implementation, and policies.

Methods: We conducted a scoping review guided by a population, concept, context framework and operationalized through a Population, Intervention, Comparison, and Outcome (PICO)-informed data-charting structure to map study characteristics and reported outcomes, without causal or effectiveness assumptions and without risk-of-bias assessment. We searched PubMed, Google Scholar, and ClinicalTrials.gov for studies published between October 2015 and September 2024, complemented by AI-assisted discovery tools for citation extension. Reproducibility logs (search strings, run dates, filters, and stepwise counts) were maintained. Of 1751 records retrieved, 245 were included. Evidence was charted by LSD population, intervention class (AI, CC, and other enabling DHTs), outcome domains (patient, health care, and societal), and phase of the care journey.

Results: Among 245 included records, 92.2% (226/245) were peer-reviewed, and 7.8% (19/245) were gray literature; no completed and published randomized controlled trials or LSD-specific systematic reviews were identified, with evidence dominated by small, single-center observational studies. Overall, 40 peer-reviewed records reported AI-driven DHTs, 89 reported CC DHTs, and 144 reported other enabling DHTs (some multilabeled). Evidence was concentrated mostly in Gaucher and Fabry diseases. Nearly half of the mapped literature focused on screening and diagnosis, with fewer records addressing treatment intensification, rehabilitation, and end-of-life care. Outcomes were predominantly health care delivery performance measures, with fewer patient and societal outcomes. AI applications mainly supported diagnostic decision support, phenotyping, monitoring, tracking, and risk stratification; CC commonly involved telemedicine, remote monitoring, and patient-engagement platforms; enabling DHTs included interoperable data systems, registries, and digital infrastructures.

XSL•FO
RenderX

Conclusions: The evidence base is appreciable for a niche field and reflects growing interest in AI and CC for LSD care, but heterogeneity and methodological limitations preclude inferences on effectiveness or routine implementation. This evidence map highlights relatively stronger areas and gaps, providing a structured foundation to inform timely expert consensus-building and research prioritization. Key priorities include interoperable data infrastructures and data availability, prospective multicenter evaluations, transparent reporting of algorithms and workflows, and implementation-relevant outcomes to support safe, equitable, and scalable adoption aligned with evolving European Union and global rare-disease priorities.

(J Med Internet Res 2026;28:e73612) doi:10.2196/73612

KEYWORDS

digital health; artificial intelligence; connected care; telemedicine; rare diseases; lysosomal storage disorders; remote monitoring; registries

Introduction

Rationale

Lysosomal storage disorders (LSDs) are a subset of rare diseases (RDs), resulting from inherited deficiencies in lysosomal enzymes or transporters. They affect approximately 1 in 7000 to 8000 newborns and encompass over 70 disorders. They exemplify rare, chronic, and multisystem diseases that require coordinated, longitudinal, multispecialty care [1-4]. The 2021 United Nations resolution on “Addressing the Challenges of Persons Living with a Rare Disease and their Families” [5], together with several policy statements from the European Union (EU) and the World Health Organization, recognizes RDs as a global policy priority and calls for coordinated action and strengthened care networks to advance equitable Universal Health Coverage by 2030 [3].

Across RDs, the diffusion of digital health technologies (DHTs) has accelerated in both research and care delivery, reflecting the need to enhance early and accurate diagnosis, reduce geographic barriers, improve inclusion for dispersed populations, and support longitudinal follow-up [6,7]. DHTs have also been increasingly adopted in RD clinical trials to enable remote assessments and decentralized research models [8]. In parallel, regulators and health systems are integrating digital health into health policies and health technology assessment (HTA) frameworks, with growing emphasis on clinical validation, transparency, and safety. These cross-cutting developments are especially relevant to LSDs, where chronic multisystem trajectories require coordinated multidisciplinary care, timely specialist input, and monitoring across the life course.

However, despite this rapid diffusion and growing policy attention, the landscape of DHT applications in LSD care remains poorly characterized. While the burden and care complexity of RDs, such as LSDs, are well established, the evolving contributions of artificial intelligence (AI) and connected care (CC), as distinct yet complementary classes of DHTs with the potential to enhance LSD management, have not been systematically mapped [3,5,9]. Over the past decade, AI applications (eg, pattern recognition for early diagnosis, predictive analytics, and data-driven decision support systems [DSS]) and CC solutions (eg, televisits, remote monitoring, and digital medical devices [DMD]) have been increasingly reported in LSD contexts. However, evidence is dispersed across heterogeneous study designs, gray literature, and emerging evaluations, which limits informed decision-making for clinicians, patient organizations, and policymakers. Prior reviews survey broader rare-disease contexts or single DHT classes, but none provide an LSD-specific evidence map specifying the studied interventions, populations, care-journey stages, and outcome domains—information fundamental to responsible research, clinical adoption, and policy development.

Objectives

Considering this evolving context, we conducted, to our knowledge, the first scoping review conducted in accordance with PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [10] to describe and map the published and gray literature from the past decade on DHTs relevant to LSD care, with a primary analytic focus on AI-enabled and CC applications and a contextual mapping of other enabling DHTs.

Our aims were to (1) clarify scope and applications across AI, CC, and other enabling DHTs; (2) chart the distribution of evidence across 4 axes (populations, intervention types, care-journey phases, and outcome domains) using an explicit charting framework (Textbox 1 and Figures 1-2); and (3) identify gaps, methodological limitations, and research priorities relevant to clinical implementation and policy development [3,5,9].

This review is explicitly descriptive in nature and does not seek to determine effectiveness or infer causal impact.

XSL•FO
RenderX

Textbox 1. Intervention-based categorization framework to map, cluster, and assess literature for AI-driven, connected care, and other digital health technologies.

Artificial intelligence–driven digital health technologies

- Artificial intelligence (AI) algorithms
  - Machine learning (ML) models, deep learning (DL) and neural networks (NN), natural language processing (NLP), and large language models (LLMs)
- AI-driven digital health technology (DHT) clinical applications
  - Risk stratification, facial recognition, clinical imaging analysis, predictive analytics, decision support, intelligent genomic counselling, virtual agents, and assistants
- AI-driven DHTs research and development applications
  - Drug development, enhanced or simulated clinical trials, disease and phenotypic pattern recognition, and disease epidemiological and progression modeling

Connected care DHTs

- Telemedicine
  - Health care provider (HCP)-patient televisits (including tele-genetics), HCP-HCP teleconsultations, collaborative health care networks
- Digital medical devices
  - Remote patient monitoring (RPM) devices, electronic patient-reported outcomes (ePRO), Mobile Health (mHealth) apps, patient care support programs, digital therapeutics (DTx)
- Data integration systems
  - Electronic health records (EHRs), health information systems (HIS), and national and international registries
- Patient engagement platforms
  - Online patient portals, networks, and peer-to-peer communities

Other DHTs

- Clinical DHTs
  - Precision medicine platforms, digital microfluidics (DMF) platforms, high-performance liquid chromatography (HPLC)–tandem mass spectrometry (MS/MS), bioinformatics tools, enzyme replacement therapy (ERT) optimization tools, virtual reality (VR) & augmented reality (AR), biomedical imaging, decision support systems (DSS)
- Data integration systems enablers
  - Blockchain for health data security
- Research and development DHTs
  - Clinical trial simulation tools, computational modeling
  - Multipurpose omics and molecular analysis DHTs
  - Genomics (including next-generation sequencing), pharmacogenomics, metabolomics, and proteomics applications and platforms.

XSL•FO
RenderX

Figure 1. Outcome-based framework to map, cluster, and assess literature for patient, health care, and societal impact. HCP: health care provider.

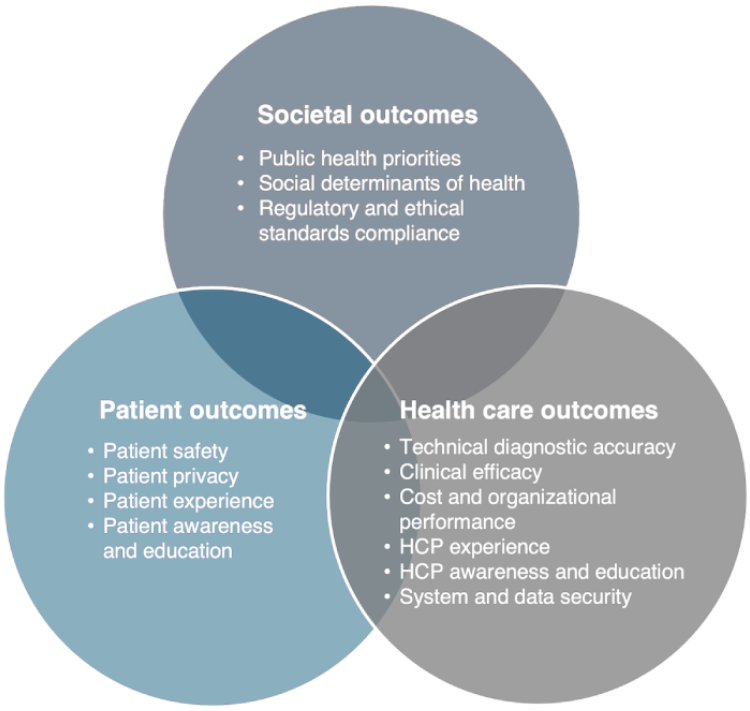

Figure 2. Timed framework to map, cluster, and assess literature for LSD natural history and care journey phases.

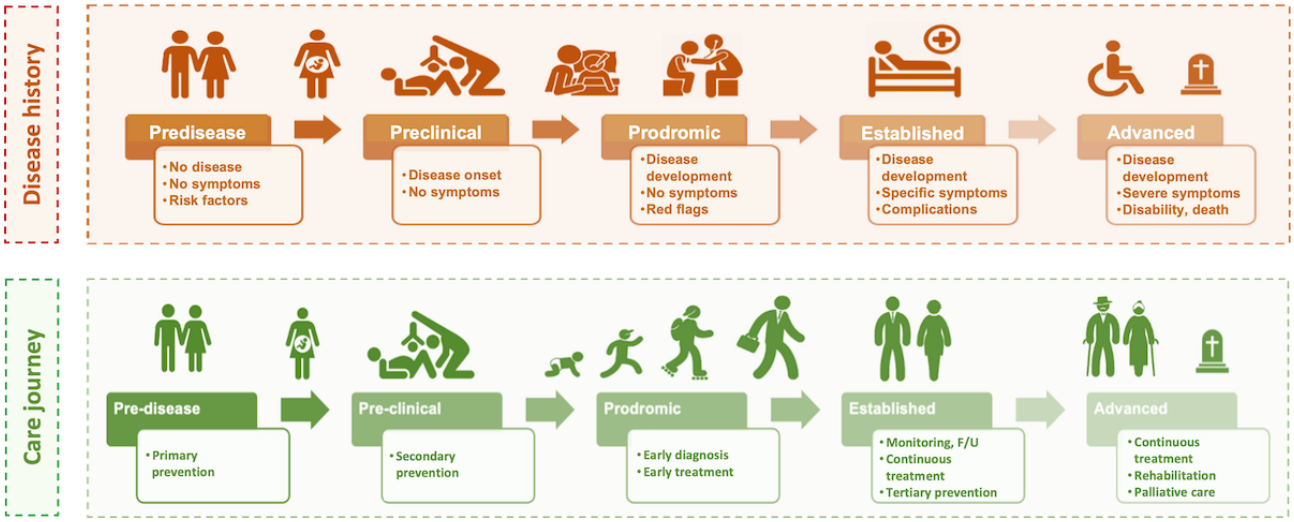

Methods

Overview

The dynamic and rapidly advancing field of DHTs and the variability in the quality of existing evidence render traditional systematic reviews and graded recommendations unsuitable, while making scoping reviews particularly applicable and informative. We used a multidimensional, PICO (population, intervention, comparison, and outcome)-informed, charting framework to classify and map evidence (Textbox 1 and Figures 1-2), while framing the overarching review question using a population–concept–context (PCC) logic. In this study, "PICO-informed" refers to a structured approach to categorize populations, intervention, and outcome domains for mapping purposes, not to a causal comparison framework—consistently with the aim of scoping reviews to map the breadth, characteristics, and gaps of evidence rather than to estimate comparative effectiveness.

The work adheres to the 5-stage scoping review framework advanced by Arksey and O'Malley [11] and follows the PRISMA-ScR guidelines [10,12]. Specifically, reporting is aligned with the 22-item PRISMA-ScR checklist (Multimedia Appendix 1). A prospective protocol was not registered; we acknowledge this as a limitation for transparency and therefore provide full search strings, dates, and flows to support reproducibility.

Identifying the Research Question

Our review question was as follows: In adult and pediatric populations with LSDs (including at-risk/suspected cases), their caregivers, and/or health care professionals (Population), what types of DHTs, particularly AI-enabled and connected care solutions, along with other enabling digital technologies

XSL•FO
RenderX

(Concept), have been reported over the last decade, and how is the evidence distributed across the LSD care-journey phases and across patient, health care delivery, and societal impact domains (Context/Outcomes)?

Comparators were not required for inclusion; when present, they were recorded descriptively. Outcomes were extracted as reported and mapped across predefined domains (Figure 1) to support descriptive synthesis and identification of evidence gaps, rather than causal inference.

Identifying Relevant Studies

We searched PubMed (including MEDLINE) and Google Scholar as core sources, complemented by queries on ClinicalTrials.gov and AI-assisted discovery tools, to broaden recall beyond records explicitly indexed in the core databases. Filters were applied to include studies published in the last decade, regardless of language, and to capture relevant gray literature, including non-peer-reviewed preprint studies, trial registries, theses, reports, expert interviews, book chapters, websites, and institutional or patient association portals. Overall, the multisource search retrieved 1751 records after prescreening. Full database-by-database reproducibility logs (strings, run-dates, filters, and stepwise counts) are provided in Tables S1-S7 in Multimedia Appendix 2.

On September 30, 2024, the baseline PubMed query returned 218 records (full stepwise string and fixed publication date filter reported in #8 in Multimedia Appendix 2). To increase sensitivity while preserving specificity for LSDs, we expanded the DHT concept using 55 keywords, consolidated from >80 DHT intervention terms identified during preliminary mapping, yielding an additional 190 records (total PubMed yield 408; see #5 and #10 in Multimedia Appendix 2). The term “rare disease” was not included, as it captured an excessive number of papers (>1900) with poor specificity for LSDs.

Google Scholar was searched on the same date using the predefined query and year filter (2015-2024), yielding approximately 4070 results (including duplicates and gray literature). To run specific reproducible screening under platform ranking variability, we applied a prespecified stopping rule by exporting and screening only the first 1000 results sorted by relevance (see #24 and #25 in Multimedia Appendix 2).

A similar research query was run on the randomized controlled trial (RCT) registry ClinicalTrials.gov (full string and fixed publication-date filter reported in #32 in Multimedia Appendix 2). We preliminarily identified 14 ongoing or completed, but unpublished, RCTs.

Consensus, SciSpace, and Connected Papers were used as supplementary AI-driven search tools (query expansion and citation chasing) on September 30, 2024, retrieving respectively 50, 163, and 102 records, including duplicates across the outputs (see #34, #35, and #47 in Multimedia Appendix 2). These tools were used only to identify candidate records; they did not autonomously determine eligibility, extract data, or draw conclusions. For reproducibility, we logged for each tool the run date and reference period, inputs (keywords, filters, or seed articles), and the number of candidate records retrieved. The author team reviewed all retrieved outputs and underwent the same deduplication and screening workflow as all other records.

An additional 14 records were finally added from citations, authors’ suggestions, and general web searches.

Study Selection

Records were de-duplicated using Paperpile (non-AI**,** reference manager software) and then double-checked manually. Title and abstract screening was performed by one reviewer (LN) after calibration with a second reviewer (AMCS) to refine eligibility criteria and resolve ambiguities. To mitigate single-screener selection error, AMCS independently screened (1) all records marked as uncertain and (2) a random sample of excluded records; discrepancies were resolved by discussion and, when needed, adjudication by a third reviewer (MS). LN and AMCS assessed full-text eligibility independently, with MS adjudicating disagreements. All researchers reached consensus on the final selection of studies included for data synthesis.

Given the emerging nature of this research field, the eligibility criteria were kept broad to include all study types published over the past 10 years in any language. Gray literature was also taken into consideration if particularly relevant. However, only records that directly addressed our scoping review question, explicitly involving both LSDs and at least one DHT mapped in our taxonomy (Textbox 1), were selected for analysis.

Inclusion criteria were (1) populations with a confirmed LSD diagnosis or at-risk LSD status (including screening, suspected, or undifferentiated LSD cases), and/or their caregivers or health care professionals; (2) digital interventions classified as AI, CC, or other DHTs according to the taxonomy in Textbox 1; and (3) outcomes that could be mapped to at least one of the predefined patient, health care delivery, or societal impact domains.

We excluded records for which (1) full text, or a minimum set of extractable data, could not be retrieved; or records that were (2) not LSD-specific and/or did not involve a DHT as defined in Textbox 1; where cited in Discussion, it served contextual purposes only. Non–peer-reviewed items were retained only when a stable, traceable source was accessible, and core data could be verified.

Data Charting and Synthesis

To convey source heterogeneity, we labeled each included record by peer-reviewed and gray source type, and summarized counts and distribution by evidence tier (Table 1). During synthesis, peer-reviewed items were always summarized separately from gray literature. Source validation and interpretive weighting followed a simple hierarchy: peer-reviewed original studies and reviews formed the primary basis for interpretation; gray sources were used descriptively only to contextualize emerging initiatives and evidence gaps. No clinical or system-level statement in this review rests solely on non–peer-reviewed sources.

XSL•FO
RenderX

Table 1. Literature source types of included records (distribution): corpus split into peer-reviewed and gray literature with subcategories per tier.

Codes and source type		Unique papers, n	% Source tier[a], % (n/N)	% Total papers[b], % (n/N)	Role in synthesis
L.1 Peer-reviewed		226	100 (226/226)	92.2 (226/245)	—[c]
L.1.1	RCTs[d], systematic reviews (SR), meta-analyses	0	0	0	Primary
L.1.2	Narrative, scoping reviews	44	19.5 (44/226)	18 (44/245)	Secondary
L.1.3	Original research (non-RCT)	176	77.9 (176/226)	71.8 (176/245)	Secondary
L.1.4	Case reports, case series	6	2.7 (6/226)	2.4 (6/245)	Context only
L.1.5	Editorials, letters, commentaries, perspectives	0	0	0	Context only
L.2 Gray literature		19	100 (19/19)	7.8 (19/245)	—
L.2.1	Registered, ongoing RCTs	7	36.8 (7/19)	2.9 (7/245)	Context only
L.2.2	Preprints	2	10.5 (2/19)	0.8 (2/245)	Context only
L.2.3	Conference abstracts, proceedings	1	5.3 (1/19)	0.4 (1/245)	Context only
L.2.4	Theses, dissertations	4	21.1 (4/19)	1.6 (4/245)	Context only
L.2.5	Books, book chapters	2	10.5 (2/19)	0.8 (2/245)	Context only
L.2.6	Websites, web posts	1	5.3 (1/19)	0.4 (1/245)	Context only
L.2.7	Reports, white papers, interviews	1	5.3 (1/19)	0.4 (1/245)	Context only
L.2.8	Patents	1	5.3 (7/19)	0.4 (1/245)	Context only
L.1+L.2 All included records		245	—	100 (245/245)	—

[a]% of Total Papers uses 245 as the denominator.

[b]% of Source Tier uses 226 for peer-reviewed (182 research articles + 44 reviews) and 19 for gray literature (7 registered RCTs + 12 miscellaneous sources).

[c]Not available.

[d]RCT: randomized controlled trial.

In line with scoping review guidance, and due to the heterogeneity of study designs, objectives, and sources of evidence, no formal assessment of the risk of bias or methodological quality was performed. Instead, we charted design, sample size, setting, and other quality-relevant features and used these descriptively to temper interpretation in the Results and Discussion sections.

Evidence was mapped and synthesized along the 4 dimensions derived from our multidimensional review question: Population, Intervention class (Textbox 1), Outcome domains (patient, health care delivery, and societal-level outcomes; Figure 1); and Care-journey phase (Figure 2). This population–intervention–outcome–care-phase analytic framework accommodated heterogeneous designs; it was descriptive and label-based, relying on counts and cross-tabulation rather than formal meta-analysis or effect-size estimation.

Population in scope was represented by adults and children diagnosed (or at-risk of being affected) with LSDs, with a particular focus on Gaucher disease (GD), Fabry disease (FD), mucopolysaccharidosis (I, II, IIIA, IIIB, IVA, VI, VII), Pompe disease (PD), Niemann-Pick disease, Krabbe disease, metachromatic leukodystrophy, GM1 and GM2 Gangliosidosis (GM1G and GM2G), Batten disease, and aspartylglucosaminuria, among more than 70 classified entities. Where possible, insights were charted and reported by age group and LSD subtype to reflect differences in clinical presentation and care needs.

We clustered DHT interventions into 3 domains that were applied consistently during study selection, charting, and synthesis: AI-driven DHTs, CC DHTs, and other DHTs (Textbox 1). AI-driven DHTs are adaptive and, to varying degrees, autonomous systems capable of learning, reasoning, and generating outputs from input data (eg, machine learning [ML] and deep learning [DL] models, natural language processing [NLP and large language models [LLMs]). In clinical contexts, these tools typically support early diagnosis and phenotyping, risk prediction, and decision support. CC DHTs include telemedicine; DMDs such as devices monitoring clinical parameters and electronic patient-reported outcomes (ePROs), mobile health (mHealth) apps, and digital therapeutics (DTx); interoperable data-integration systems (electronic health record [EHR], health information systems [HIS], registries, and emerging health-data spaces); and patient-engagement platforms. The category of other DHTs comprises clinical and research and development (R&D) tools that are neither AI-based nor connected-care systems (eg, digital microfluidics, omics and bioinformatics platforms, non-AI imaging tools, and trial-simulation software).

XSL•FO
RenderX

The impact of DHTs was then assessed across outcomes grouped into 3 domains (Figure 1): patient-related outcomes (patient safety, privacy, experience, awareness and education), health care delivery outcomes (technical diagnostic accuracy, clinical efficacy, cost and organizational performance, system and data security, health care provider [HCP] experience, awareness and education), and societal outcomes (public health and social determinants of health, regulatory and ethical compliance). Each included record was coded to one or more (≥1) outcome domains; when a study reported multiple domains, it was multilabeled. As a result, totals in outcome tables may exceed the number of unique records. The Discussion interpreted cross-domain patterns (rather than effectiveness) in line with scoping aims.

We also generalized the natural history of LSDs and summarized how DHTs were positioned across key care-journey phases**:** primary prevention, secondary prevention (screening) and diagnosis, treatment, tertiary prevention and rehabilitation, monitoring and follow-up, and end-of-life care (Figure 2; Multimedia Appendix 3). Each included record was mapped to ≥1 phase depending on where in the journey the intervention (AI, CC, or other DHTs) was evaluated or deployed; multiphase studies were multicoded. The visual representation does not imply causation; it provides placement along the pathway to support descriptive synthesis in the Results and higher-order interpretation in the Discussion.

To interpret plausible causal pathways and specify testable hypotheses for LSD care, we also drew on evidence from adjacent non-LSD clinical domains where AI (eg, diagnostic triage and risk prediction) and CC (eg, televisit, teleconsultation, remote follow-up and monitoring adoption, and adherence) have been more extensively studied. These analogies were cited only as contextual signals (not pooled with LSD-specific data) and used to motivate prospective evaluations (eg, multisite designs, external validation, and standardized outcomes) in LSD settings.

Along the charting process, risks, challenges, and biases associated with DHTs, evidence limitations, and future research needs were identified and reported in the Results and Discussion sections. Key examples are also summarized in Multimedia Appendix 4.

Results

Overview

After the identification of 1751 records, we applied a rigorous selection process adapted to the nature of this multidimensional scoping review and mapping framework (PRISMA-ScR flow in Figure 3). The screening first led to 1225 items after removing duplicates and ineligible entries. The following screening process on titles and abstracts led to the exclusion of 437 records that did not meet relevance criteria coherent with the review question and eligibility criteria. A detailed assessment was then conducted on 788 records. Thorough eligibility assessment ensured each document was relevant to the specific focus, by embracing both LSDs management and DHTs integration, resulting in 245 records being included for final consideration: 182 research articles, 44 reviews, 19 “grey” literature records (including 7 ongoing or unpublished registered RCTs and other sources such as preprints, thesis, book chapters, periodicals, reports, and web posts). Finally, within the peer-reviewed literature, 40 items focused on AI-driven DHTs, 89 on CC DHTs, and 144 on other enabling DHTs, with some reporting more than one category. Nearly all records were published in English, only a few in Italian, German, and Russian.

In addition to LSD-focused records, the searches surfaced many articles on digital health applications in other rare and non-rare conditions. A subset of 492 journal articles and 51 pieces of gray literature was consulted only for contextual illustration. However, these contextual references were not systematically reviewed and did not influence study selection, data extraction, or synthesis.

Data were summarized in a PRISMA-ScR flow diagram (Figure 3), reporting identification, screening, eligibility, and inclusion counts for the corpus. Numbers correspond to records after deduplication and reflect all intervention classes (AI, CC, other DHTs). Reasons for exclusion at full text are summarized in the right-hand box and detailed in the Methods section. Transparency is provided via full search strings, logs, and the PRISMA-ScR checklist (Multimedia Appendix 1).

Expanded study-level evidence charting tables, showing distribution across literature sources, and comparing population characteristics, DHT categories, and care-journey mapping, are provided in Table 1 and Multimedia Appendix 5.

XSL•FO
RenderX

Figure 3. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) flow diagram adapted for a multidimensional scoping review and evidence-mapping framework [9]. AI: artificial intelligence; DHT: digital health technology; LSD: lysosomal storage disorder; RCT: randomized controlled trial.

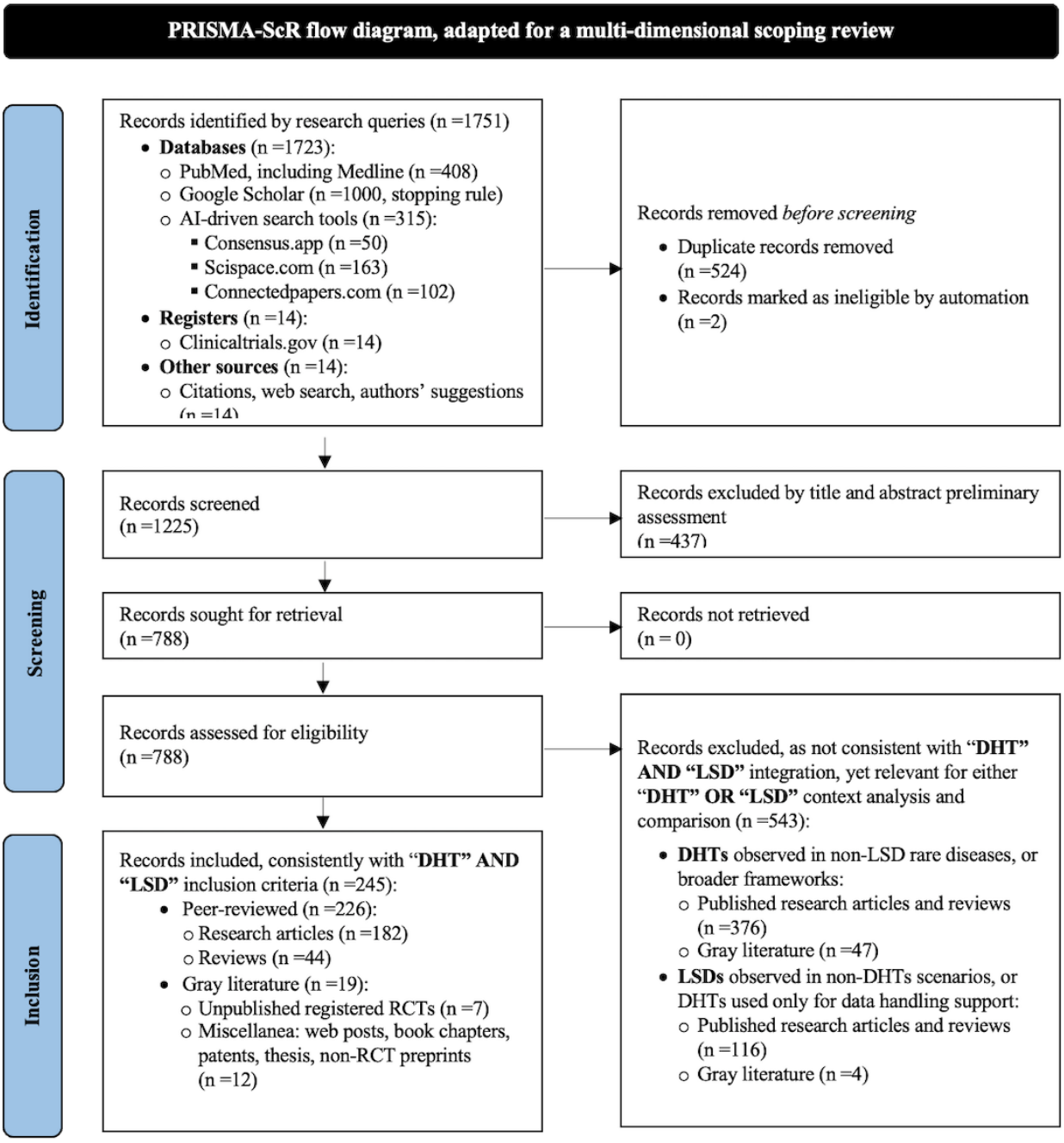

A synthesis at-a-glance of literature source distribution (Table 1) shows that, across the 245 included records, more than 90% (226/245) are peer-reviewed journal items and 7.8% (19/245) are gray literature (registered or ongoing trials, theses, reports, web-based documents, and other non–peer-reviewed sources). No completed RCTs, systematic reviews, or meta-analyses with LSDs-specific digital interventions were identified. Instead, the corpus is dominated by original observational, nonrandomized research and narrative or scoping reviews, with a small tail of case reports and contextual gray sources. Consistent with our scoping aim, peer-reviewed studies form the core basis for interpretation, whereas gray sources are used descriptively to map context and evidence gaps.

Disease-level comparative labeling (Table S1 in Multimedia Appendix 5) shows that evidence is unevenly distributed across LSDs. GD and FD together account for nearly 60% (130/226) of peer-reviewed papers, with additional clusters in mucopolysaccharidosis and PD, and only sparse coverage for other LSDs. To avoid overgeneralization, our synthesis is weighted by this coverage: where an LSD has multiple, convergent peer-reviewed studies, we treat findings as relatively more robust; where counts are very low, we label insights as signal-only and refrain from extrapolating cross-disease claims. This approach preserves disease-specific detail while explicitly acknowledging both sounder evidence and gaps.

XSL•FO
RenderX

Standardized intervention labeling (Table S2 in Multimedia Appendix 5) yielded 380 labels across AI, CC, and other DHTs. Although AI-specific labels represent only about 14% (52/380) of all intervention labels, they appear in approximately 18% (40/226) of peer-reviewed papers, largely in diagnostic and clinical-decision support applications. CC labels account for 32% of all labels (122/380), with DMDs and telemedicine applications most frequently represented. The remaining half of labels fall under other, mostly enabling DHTs (eg, omics platforms, bioinformatics pipelines, and non-AI imaging tools), which provide critical infrastructure. Since multicategory records were multicoded, labeled-paper totals may exceed; we therefore interpret these distributions qualitatively as relative signal strength rather than prevalence estimates.

Outcome-domain labeling (Table S3 in Multimedia Appendix 5) yielded 430 outcome labels across 245 records. Roughly two-thirds of labels (279/430, 64.9%) fall within health care delivery outcomes, led by technical diagnostic accuracy (157/430, 36.5%), with cost and organizational performance plus system and data security together contributing a further 75/430 (17.4%) labels. Patient-level outcomes account for just over one quarter of labels (117/430, 27.2%), dominated by patient experience (45/430, 10.5%) and privacy (39/430, 9.1%), whereas patient safety, awareness, and education remain relatively infrequent (33/430, 7.7% combined). Societal outcomes are least represented, with only 34 of 430 labels (7.9%), 10 of 430 (2.3%) for public health and social determinants of health, and 24 of 430 (5.6%) for regulatory and ethical compliance. Technical diagnostic accuracy alone accounts for more than one-third of all outcome labels and about half of peer-reviewed outcome reports, underscoring a strong focus on accuracy, detection, and classification—for now, more than on clinical efficacy and long-term patient or societal impact. This distribution reflects an evidence landscape characteristic of emerging innovations, in which establishing technical reliability and safety benchmarks precedes the generation of robust data on clinical effectiveness, patient-centered outcomes, and broader societal impact.

Care-journey mapping (Table S4 in Multimedia Appendix 5) shows that nearly half of all phase labels (222/472, 47%) concentrate in secondary prevention (screening) and diagnosis, followed by monitoring and follow-up (116/472, 24.6%) and treatment (95/472, 20.1%), while preconception primary prevention accounts for only 11/472 (2.3%) labels. Rehabilitation, tertiary prevention, and end-of-life care together contribute 28/472 (5.9%) labels, indicating much slimmer digital activity in later or more complex phases of the care journey. Across phases, other DHTs dominate diagnostic and treatment phases (eg, 120/222 labels, 54.1%, in secondary prevention and diagnosis), whereas CC labels are relatively more prominent in monitoring and follow-up (47/116, 40.5%), and AI labels remain numerically modest in downstream phases. These patterns, aligned with the timed framework in Figure 2, highlight a key digital footprint and exploration of applications to enhance early diagnosis and ongoing monitoring, with comparatively fewer digital interventions evaluated in primary prevention and in later phases of the care journey. The subsequent Results subsections unpack these phase-specific signals descriptively, while the Discussion interprets their implications for evaluation and implementation.

Population-Specific Insights by Disease

Coverage Patterns Across LSDs

As shown in Table S1 in Multimedia Appendix 5, disease coverage is highly asymmetric: FD and GD account for >50% of peer-reviewed records, whereas most other LSDs have limited or no DHT literature. We therefore synthesize findings by LSD and avoid cross-disease inference when evidence is sparse.

Although age was not charted as a separate axis, several records describe distinct digital use-cases in pediatric vs adult cohorts. In pediatric care, tools emphasize early detection (eg, neonatal blood spot screening and facial recognition), adherence and progression monitoring, and child- and family-centered education (including virtual reality [VR] and gamified approaches), alongside concerns about limited evidence, access inequities, and privacy [13-20]. In adults, mHealth apps, portals, and telemedicine emphasize self-management and access, while analytics support personalized planning and detection of late-onset manifestations [21-23]. Data scarcity, heterogeneity, and privacy governance recur across age groups [13-23].

Fabry Disease

In FD, AI use-cases cluster around early detection, phenotyping and organ involvement: facial recognition and image analytics (eg, magnetic resonance imaging [MRI], urinary sediment imaging, electrocardiography, echo, and MRI) to detect brain, renal, and cardiac manifestations, and support prognosis, plus EHR-based ML to flag high-risk patients for screening [24-33]. Variant-level prediction of pharmacologic responsiveness from structural and proteomic features is also reported [34-37].

CC and enabling tools support longitudinal management through telemedicine, DMDs, ePROs, registries, and EHRs for symptom tracking, adherence support, and remote follow-up, including app-based support programs integrating wearable or patient-entered data [22,38-43]. Enabling diagnostics include non-AI EHR risk models, digital microfluidics (DMF) + tandem mass spectrometry (MS/MS) newborn screening, and next-generation sequencing (NGS)–based confirmation and precision workflows [14,44-47]. Reported risks include overreliance on digital outputs and discordance with patient experience and clinical status [38,48,49].

Gaucher Disease

In GD, AI is most often applied to imaging-based phenotyping (eg, MRI-derived bone complication prediction; automated spleen-volume estimation) and to EHR and registry analytics supporting risk stratification and screening [8,22,44,50-59]. CC components include televisits, teleconsultations (highlighted during COVID-19) and, in some records, wearables, mHealth, and ePROs for intervisit monitoring of gait, sleep, pain, and stress [6,7,22,60,61]. Enabling diagnostics include genomic-analysis workflows and DMF-based assays supporting mutation detection and newborn-screening pathways [62-64]. Reported limitations include small datasets, bias risk, and limited evidence on long-term impact [56].

XSL•FO
RenderX

Mucopolysaccharidosis

For mucopolysaccharidosis, records describe ML-supported differential diagnosis and prediction of treatment response, complemented by CC (telemedicine and tele-genetics) for remote consultation, adherence monitoring, and education [22,60,65-67]. Enabling tools include genomic pipelines for mutation detection and DMF platforms for diagnostic assays [47,68]. Reported challenges include data integration, regulatory compliance, and evolving practices that require ongoing updating of digital tools [67,69].

Pompe Disease

For PD, AI-enabled tools are reported for EHR-based case finding and progression prediction, and for variant, enzymatic analytics within multiomics research pipelines [35,70-73]. CC applications include telemedicine-supported physiotherapy supervision and cognitive training, plus mHealth trackers and ePROs to monitor neuromuscular function and correlate patient-reported metrics with clinical outcomes, particularly in late-onset PD [60,74-79]. Genomic workflows support mutation identification and personalization [47,80]. Reported implementation barriers include technical and regulatory constraints, costs, fragmented care pathways, and variable user acceptance [74,75,81].

Other LSDs

Beyond FD, GD, mucopolysaccharidosis, and PD, evidence is sparse and typically disease-specific.

In Niemann-Pick disease, AI-driven tools and metabolomic profiling enhance diagnostic accuracy, and DMF platforms streamline diagnostic assays [68,82,83].

Computational tools enhance newborn screening in Krabbe disease by improving specificity and reducing costs, effectively differentiating true-positive from false-positive cases [84].

In MLD, AI algorithms aid in quantifying demyelination load in MRI scans [85,86] and simulate disease mechanisms to predict drug efficacy before clinical trials [87].

Telehealth services improve health care access and satisfaction for patients with Batten disease, enhancing care delivery and potential long-term benefits [20].

MRI scoring and quantitative spectroscopy can be used as surrogates of clinical disease progression in GM1G, as they correlate with ambulation and expressive language gradual impairment [88]. Wearable devices and ePROs may monitor patients with GM2G remotely, assessing health aspects like fatigue, stress, and self-esteem [89].

ML analysis of thalamus MRI images may diagnose aspartylglucosaminuria in patients with neurocognitive symptoms like progressive learning disabilities, with low error rates and promising diagnostic potential [90].

Intervention-Specific Insights by AI-Driven DHTs

Overview

As summarized in Table S2 in Multimedia Appendix 5, intervention subtypes by record, AI-driven DHTs constitute one of the 3 main classes of interventions mapped in this review, with studies clustering predominantly around diagnostic decision support, image and signal analysis, risk prediction, and data-driven patient stratification. The subsections below provide illustrative examples of how AI is being applied along LSD care pathways, highlighting typical use-cases rather than attempting a formal performance ranking.

AI for Risk Prediction and Patient Stratification

ML models identify patients at risk of developing specific diseases, predict the risk of complications, and stratify patients based on disease severity. This is generally achieved by analyzing patient demographics, clinical features, and genetic information [30,56,59,71].

AI for Clinical Diagnostics and Therapeutics

ML models identify potential LSD drug candidates by screening compound libraries and matching data against protein or genomic patient-specific data. They optimize treatment strategies, such as enzyme replacement therapy (ERT), and predict the responsiveness of individual variants to specific treatments [8,28,35,37,52,56]. AI-powered DSS shows use in metabolic conditions where overlapping symptoms often cause diagnostic delays. Intelligent systems may achieve high accuracy (87.4%) in differential diagnosis of LSDs by generating a list of diagnostic hypotheses based on clinical data [16,91,92].

DL algorithms, as a subset of ML that uses artificial neural networks with multiple layers to analyze complex patterns, are widely used for analyzing medical images (eg, brain or cardiac MRI and computed tomography, echocardiography, electrocardiogram, tissue or blood smears, facial features) to detect clinical biomarkers, structural anomalies, and disease-specific LSD patterns. For example, AI applications improve early detection of spleen and liver enlargement [54] or osteonecrosis in GD [51-53], or brain [26,90] and cardiac [28-32,34,93-95] involvement, mostly in FD and PD. AI-driven facial recognition software, used in diagnosing GD and FD, detects distinctive facial features commonly presented in large patient cohorts with early diagnosis potential [24,96,97].

NLP and LLMs extract critical information from unstructured data sources, such as clinical notes, interviews, research papers, insurance claims, and other textual sources. These insights can be integrated into AI-powered symptom assessment or DSS, informing diagnostic and treatment support for LSDs, particularly where clinical expertise is limited or symptoms are challenging to diagnose. NLP enables computers to understand and process human language, allowing them to extract meaningful insights from textual data, identifying patterns and correlations. LLMs have demonstrated remarkable capabilities in understanding and generating human-like text, summarizing patient narratives, extracting key medical concepts from clinical notes or visits, and generating comprehensive reports [33,98,99].

AI for R&D

There are already significant applications of ML and DL models in research on LSDs, including disease modeling, clinical trial simulations, drug discovery, and data analysis. ML tools identify disease-relevant compounds, predict their potential as inhibitors or chaperones, and link them to disease targets, advancing R&D

XSL•FO
RenderX

efforts [37,72,73,100,101]. DL and facial recognition software aid in the investigation and analysis of facial phenotypes in patients with LSDs, contributing to the understanding of these RDs [21,102]. Researchers use AI and simulations to study lysosomal enzymes and their role in storage diseases. These tools may speed up drug development, especially for gene therapies and targeted treatments. AI may optimize preclinical trials and simulate disease progression and treatment response, predict how diseases may evolve, explore new treatments, and simulate genetic mutations and their effects, aiding in establishing standardization and personalization workflows.

Intervention-Specific Insights by CC DHTs

Overview

As summarized in Table S2 in Multimedia Appendix 5, CC solutions are most often used to extend specialist input across distance, support long-term follow-up, and coordinate multidisciplinary care around complex, chronic trajectories. The narrative examples below illustrate how teleconsultation, home-based and hybrid models of care, and registry- or platform-based coordination are being deployed in practice, while acknowledging that formal comparative, cost-effectiveness, and equity analyses remain limited.

Telemedicine

Televisits among providers and patients (HCP-patient) enhance LSD care delivery. Evidence highlights the key role televisits played during the COVID-19 pandemic when in-person consultations were limited [6,15,17,39-41,60,74,103-112].

Platforms supporting virtual visits, genetic counseling (tele-genetics), and real-time monitoring have proven invaluable for maintaining continuity of care, particularly for patients in remote or underserved areas. For instance, telemedicine ensured that patients receiving ERT for conditions like GD and FD continued to receive necessary follow-ups, reducing “missed appointments” rates. Tele-genetics has specifically improved access to genetic testing and counseling, enhancing screening, early diagnosis, and treatment planning [20,43,66,67,75,113].

Tele-consultations among HCPs and collaborative health care networks enable multidisciplinary teams, including geneticists, pediatricians, cardiologists, neurologists, nephrologists, and primary care providers, to discuss complex cases. This collaborative approach has been crucial for LSDs like GD and FD, where care often requires input from various specialists, and nonspecialized professionals may need support from expert centers to enhance early diagnosis or continuous care. European Reference Networks already foster multidisciplinary care for RDs and, specifically, for LSDs through cross-border collaboration and could be enhanced by telemedicine and digital health applications [105,114-116].

Digital Medical Devices (DMDs)

Certified medical device software, such as remote patient monitoring devices, including medical-grade wearables and mHealth apps supported by proof of clinical efficacy, allows for reiterative tracking of real-world inter-visit biomarkers, like heart rate variability, respiratory function, mobility (including gait and fatigue patterns), providing early warnings for disease progression and symptom exacerbations. Examples include monitoring cardiovascular health for FD and activity trackers to help measure physical function as a surrogate for disease progression in GD, FD, PD, and G2MG [7,43,61,76-78,89].

Monitoring of ePROs has also demonstrated efficacy in collecting long-term data from patients with RDs, providing HCPs with detailed insights into daily symptom variations. For instance, patients with FD use ePROs to report neuropathic pain symptoms, enabling clinicians to timely adjust pain management protocols [43,49,61,78,79,89,117-121].

The adoption of digital patient care support programs, often delivered through mHealth applications, correlates with better adherence to management and treatment plans. They empower patients to take an active role in managing their care, leading to improved outcomes. Studies have shown that mHealth apps enhance patient experience, medication adherence, and help track daily symptoms, which are crucial for the long-term care of patients with LSD [7,41,43,61,76-78].

Safe and effective DTx [122] may enhance care through digitally-delivering treatments for common LSD symptoms and comorbidities, in particular, psychological symptoms, such as depression and sleeping disorders [123]. Digitally delivered cognitive behavioral-based interventions are already recommended in recent depression and insomnia treatment guidelines [124,125] and in Germany, as of December 2025, 29 DTx for mental health are already reimbursed and can be prescribed, 3 of which for sleeping disorders [126].

Patient Engagement Tools

Patient portals and online patient networks and communities have fostered greater patient engagement, particularly for RDs like LSDs. In managing long-term conditions like FD, patients who regularly access portals to review their health records and appointments are more likely to adhere to their treatments [8,57,117]. While it is important to ensure correct and up-to-date information, social platforms have created spaces for patients with LSDs to share experiences, discuss treatment side effects, and find emotional support, which in turn improves overall adherence to therapy.

Health Care Data Integration Systems and Disease Registries

Shared EHRs, HIS, and patient registries enhance RDs’ care and research, and enhance multidisciplinary collaboration across care teams and disease networks. In LSD management, EHRs allow geneticists, nephrologists, and other specialists to collaborate on treatment plans, streamlining the care process and improving decision-making. Research highlights that EHRs reduce the likelihood of miscommunication, which is critical for patients requiring long-term monitoring and multiple specialists [44,55,58,70,71].

Integrated international registries, often reported among the sources of published studies, have contributed significantly to understanding disease progression and treatment response in LSDs. These interfaces facilitated data sharing across health care institutions and HCPs, improving research into RDs by

XSL•FO
RenderX

offering large-scale epidemiological data on disease prevalence and long-term outcomes [2,22,38,59,118,127-134].

Intervention-Specific Insights by Other DHTs

Beyond AI and CC, other enabling DHTs underpin much of LSD diagnostics and research and often supply the data and infrastructure that later feed AI or CC applications (Table S2 in Multimedia Appendix 5). Key examples include: high-throughput biochemical diagnostics (DMF and advanced MS/high-performance liquid chromatography-MS/MS workflows), genomics and pharmacogenomics pipelines, bioinformatics and multiomics platforms, digitized imaging and 3D modeling, VR and augmented reality education, selected cognitive and rehabilitation use-cases, non-AI DSS, secure data architectures (including blockchain) for registry- and trial-enabling data sharing, and computational and trial-simulation models for progression and study-design optimization [10,13,16,42,46,47,62,64,68,72,73,80,83-87,92,100,101,135-156].

Outcome-Specific Insights by Domain

Overview

Table S3 in Multimedia Appendix 5 summarizes how outcomes were labeled across the 245 records, grouped into patient, health care delivery, and societal domains. Across all intervention classes, 65% of labels fall within health care delivery outcomes, 27% within patient outcomes, and fewer than 10% within societal outcomes. The following subsections provide a detailed, outcome-based synthesis of the results, grouped across the 3 predefined key domain areas.

Patient Outcome Domain

At the patient level, labeled outcomes mainly address safety, privacy, experience awareness, and education, and are less frequently represented than health care delivery outcomes. AI-enabled predictive analytics, telemedicine, and remote-monitoring DMDs are typically assessed for their ability to anticipate deterioration, flag early warning signs, and reduce avoidable unscheduled visits or admissions in chronic LSD manifestations (eg, cardiomyopathy and skeletal complications) [41,55,61]. Patient experience is usually captured through satisfaction, convenience, perceived access to expertise, usability of telehealth and mHealth tools, while privacy, awareness, and education outcomes appear in studies on consent, information provision, and trust in digital platforms [43,49,60,61,63,78,79,103,108,118,120,157,158].

Observational studies and pilots report that AI-driven risk-prediction tools, telemonitoring platforms, and wearable DMDs are used to identify early signals of deterioration or treatment-related complications, with outcomes reported as fewer emergency contacts, earlier detection of complications, or improved adherence to safety-monitoring protocols [7,41,55,61]. However, safety signals are predominantly process-oriented (alerts generated; actions triggered) rather than hard clinical endpoints, and sample sizes remain small. Reported risks include alert fatigue, overreliance on algorithm outputs, and missed events due to connectivity or data-quality failures [20,41,61].

Privacy outcomes are mostly described through technical and governance safeguards (encryption, access control, and occasionally blockchain-based architectures for registry and genomic data) and through documentation of GDPR-aligned policies, consent capture, and cross-border data-flow governance [63,103,108,152,153,157]. Direct assessment of privacy perceptions is less common, but some reports note that transparency on data use and clear governance arrangements support sustained trust in AI and CC deployments [21,60,69,159,160].

Telemedicine visits, home-based ERT programs, mHealth apps, wearables, VR tools, and online communities are frequently evaluated via patient experience and satisfaction endpoints, with commonly reported benefits including reduced travel burden to expert centers, perceived continuity of specialist input, and more flexible scheduling and follow-up [43,60,61,76-79,118]. Remote monitoring and digital support programs are also linked to perceived gains in self-management and shared decision-making, while digital literacy, connectivity, and device access can limit uptake and drive preference for hybrid models combining in-person care with digital touchpoints [20,60,69,107].

Digital support programs, disease-specific portals, and AI-enhanced educational tools (eg, symptom-assessment apps and interactive e-clinics) deliver tailored information on disease mechanisms, treatment options, and self-management; outcomes typically include self-reported improvements in knowledge, confidence, and engagement with care plans [20,33,99,107,116,117,158]. Online communities and social media groups may add peer-to-peer support, but raise concerns about variable information quality and the need for expert moderation [69,117].

Health Care Delivery Outcome Domain

Health care delivery outcomes account for nearly two-thirds of all outcome labels, dominated by technical diagnostic accuracy, clinical efficacy, system and data security, cost and organizational performance, and health care provider experience and education. AI, CC, and other DHTs are most evaluated on how they improve detection and classification, streamline workflows, support multidisciplinary coordination, and integrate data across settings [46,51-54,68,70,80,97,102,128,134,161].

Security outcomes are reported mainly for registries, EHR-linked algorithms, telemedicine platforms, and blockchain-based prototypes [63,70,103,108,128,152,153]. Evaluations focus on adherence to security standards (encryption, authentication, and audit trails), resilience to cyber-risks, and governance arrangements for access control and data sharing. Some registry and blockchain initiatives explicitly frame security and privacy as prerequisites for cross-border research collaboration and patient trust, but empirical testing of security robustness is rarely detailed beyond high-level descriptions [69,128,152,153].

Several reports and value-based care initiatives describe gains in organizational performance from telemedicine, home-based ERT, digital registries, and automated data flows, including reduced travel and visit burden, fewer missed appointments,

XSL•FO
RenderX

and perceived efficiencies in multidisciplinary coordination [20,23,50,76,116,128,134,161]. Cost-related outcomes are reported as descriptive estimates of resource savings (eg, avoided hospital visits and reduced inpatient days) or as qualitative assessments by clinicians and managers. Up-front investment needs, maintenance costs, and training requirements are also highlighted, particularly in under-resourced settings [23,69,159,161].

HCP experience outcomes concern the usability and perceived value of teleconsultation networks, interoperable HER and registry platforms, and DSS (AI-enabled or rule-based) [16,99,116,139,146,149,150]. Reported benefits include faster access to expert opinions, reduced professional isolation, and more structured approaches to complex diagnostic workups. At the same time, several studies note workflow disruption, documentation burden, and uncertainty around medicolegal responsibility as barriers to sustained adoption, especially when digital tools are introduced without sufficient co-design and training [116,149,162,163].

Digital education platforms, simulation tools, and decision-support interfaces are used to disseminate guidelines, RD curricula, and case-based learning for LSDs and other RDs [116,139,146,149,158]. Outcomes commonly include self-reported changes in knowledge, confidence, or diagnostic accuracy in simulated scenarios. Evidence on long-term behavior change or patient-level impact is limited, but these initiatives are frequently cited as enablers for broader DHT uptake and for strengthening expertise in non-specialist settings.

Societal Outcome Domain

Societal outcomes are least represented and cluster into public-health and social-determinants impacts, and regulatory and ethical compliance, appearing mainly in policy reports, registry initiatives, and mixed-methods studies on access, equity, and governance [1,60,69,153,160]. Public health endpoints are described primarily for telemedicine, networked registries, and outreach programs (eg, geographic reach, referral patterns, and continuity of care during disruptions), with digital divides in connectivity, device access, and digital literacy as recurrent constraints [1,6,60,69,105,109,130,153]. Regulatory and ethical outcomes focus on alignment with evolving frameworks for AI, medical devices, HTA, and data protection (including EU AI Act and national DTx pathways), emphasizing transparency, consent, privacy, and postmarket oversight while noting limited empirical evaluation of regulatory impact [9,21,38,122,126,128,152,153,160,162-164].

Outcome-Specific Insights by Care Journey Phase

Overview

Table S4 in Multimedia Appendix 5 and Figure 2, integrated by Multimedia Appendix 3, summarize how AI, CC and other DHTs applications and relative research are distributed along the LSD care journey; the paragraphs below provide a descriptive narrative of how these technologies are deployed across primary prevention, secondary prevention and diagnosis, treatment, monitoring and follow-up, rehabilitation, tertiary prevention, and end-of-life care, without inferring comparative effectiveness.

Primary Prevention

Primary prevention evidence is sparse and centers on preconception genetic risk identification. Genomic screening pipelines (including NGS, pharmacogenomics, and related digital platforms) and, in some reports, AI-enabled analytics support risk assessment and counseling, including via tele-genetics [14,45-47,62-64,66,68,80,84,165,166]. Ethical considerations include privacy and genetic discrimination risks.

Secondary Prevention and Diagnosis

Across the included studies, AI is used mainly for diagnostic support and phenotyping (eg, facial analysis; imaging analytics across MRI, CT, ultrasound, and smears; EHR-based case finding and severity prediction) [16,24,26,31,33,70,91,92,97-99,102,167]. A smaller set of papers reports NLP and LLM approaches to extract clinically relevant signals from unstructured text to support diagnostic reasoning [33,98,99]. CC solutions complement this pathway through televisits or teleconsultation and EHR-linked data integration, including ePRO and mHealth streams that support longitudinal assessment and diagnostic refinement, particularly for geographically dispersed families [6,20,40,43,60,61,66,76,89].

In secondary prevention and diagnosis, digital tools cluster around high-throughput screening, confirmatory pipelines, and decision support. DMF and related biochemical screening technologies underpin prenatal and neonatal screening and can accelerate early identification of disorders such as FD and GD [14,45]. Genomic workflows (NGS and other omics pipelines) support mutation detection or confirmation, and precision profiling, supplying structured data streams to downstream analytics [46,47,80,168].

Treatment

Treatment-phase applications emphasize personalization and longitudinal adjustment. AI-enabled analytics leverage genomic, proteomic, and metabolomic inputs to support therapy tailoring and to model treatment responsiveness, with reports spanning compound discovery and variant-level prediction [8,28,35,36,72]. CC components (televisits, mHealth, and teleconsultation) support timely initiation and adjustment of therapies such as ERT or substrate reduction, facilitate follow-up and multidisciplinary coordination, and perform continuous-care models integrating ePROs and biometric data for responsive management in conditions such as PD and FD [22,39-41,60,74,76,169].

Monitoring and Follow-Up

For long-term follow-up, CC and monitoring DHTs enhance assessment from episodic visits to inter-visit real-world data. Remote-monitoring DMDs, wearables, ePROs, and digital patient support programs are used to track multidimensional manifestations (eg, cardiac and respiratory function, mobility and gait, sleep, pain, and mood), enabling timely treatment adjustments and adherence support [7,40,43,61,76-78]. Telemedicine and teleconsultation platforms extend specialist oversight and multidisciplinary coordination, while EHR-linked integration (and, in some reports, AI-enabled analytics) consolidates wearable and clinical data into more holistic care

XSL•FO
RenderX

plans, particularly for chronic complications such as cardiomyopathy in FD and skeletal disease in GD [22,29,33,51].

Rehabilitation

Evidence in rehabilitation is limited, but suggests that mHealth apps, patient support programs, tele-rehabilitation platforms, and selected DTx may support at-home therapy and monitoring, sometimes integrating psychological support [61,76-78,118,123,148]. The literature emphasizes the need for clinical validation and implementation integration before these approaches can be scaled [69].

Tertiary Prevention

Tertiary-prevention records mainly describe remote monitoring tools (wearables and apps), and, in some cases, AI-enabled DSS to adjust care plans using real-world markers to prevent complications [71,76-78,89,128]. Blockchain is discussed to strengthen privacy and trust for long-term data sharing; equity, accessibility, and preservation of human contact are recurrent implementation considerations [38,69,103].

End-of-Life Care

End-of-life care evidence is limited, but suggests that telemedicine and remote monitoring can support palliative coordination by facilitating family and team communication and timely symptom-responsive interventions, with emphasis on preserving comfort, dignity, and ethical governance [6,20,43,61,66,109,110,170].

Discussion

Principal Findings

This scoping review shows a rich research landscape on digital health applications for LSDs care, although the overall quality and maturity of the evidence remain limited. Most studied digital interventions cluster upstream along the care journey, focusing on enhancing diagnosis and early monitoring, with comparatively fewer applications in treatment intensification, rehabilitation, tertiary prevention, and end-of-life care. DHTs are used to shorten the road to diagnosis and to stabilize follow-up, and less to support care interventions, where disability, caregiver burden, and health system management tend to peak. In the context of the LSDs' natural history and care journey phases (Figures 2 and 4), this pattern indicates that current evidence is more consistent on shortening diagnostic delays and optimizing surveillance, with a potential positive impact on years lived with disability and years of life lost.

Figure 4. Towards a digitally enhanced care journey for lysosomal storage diseases. HCP: health care provider; LSD: lysosomal storage disorder.

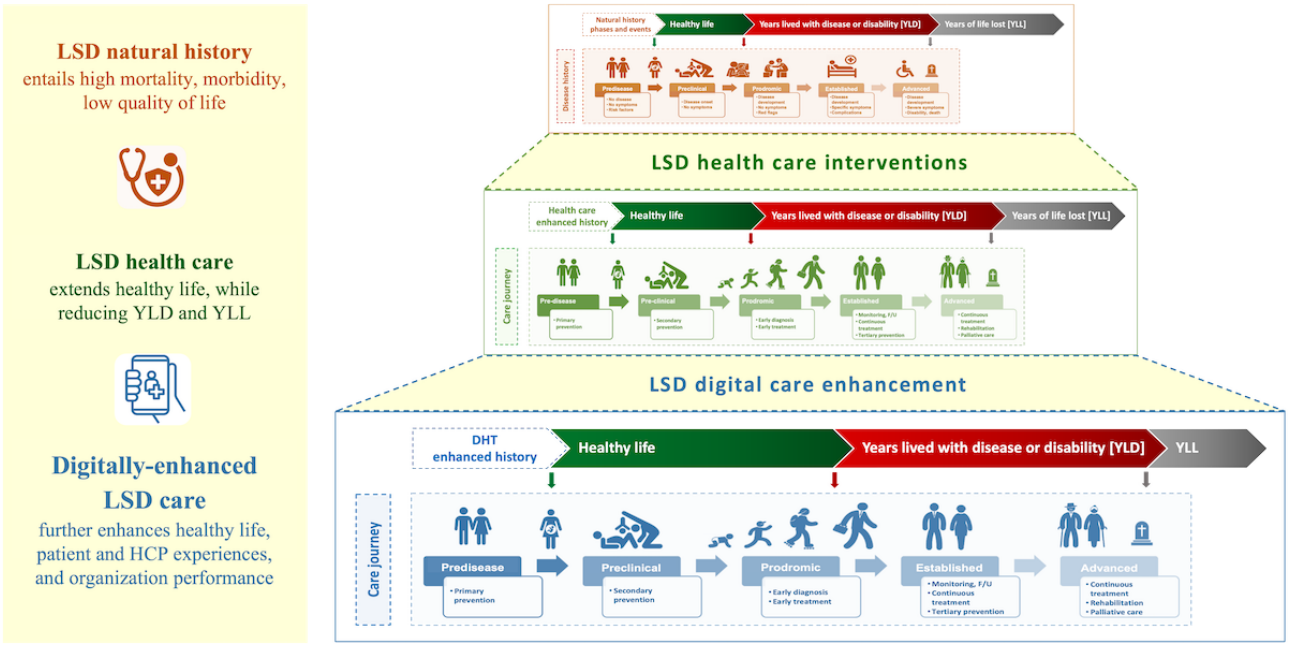

Within this landscape, CC tools anchor real-world delivery (telemedicine, remote patient monitoring, ePRO, and registries), AI is used mainly for diagnostic decision support and risk stratification, and enabling informatics (omics, bioinformatics, non-AI imaging analyses) underpin discovery pipelines. Coverage is uneven across diseases: evidence is concentrated in FD and GD, with only small clusters for other LSDs and little or no activity for many of the 70+ additional entities (Table S2 in Multimedia Appendix 5). To avoid overgeneralization, we therefore synthesized findings stratified by LSD, labeled very low-count areas as signal-only or evidence gaps, and prioritized them in the Future Research section.

Across the evidence base, most outcomes relate to health care delivery performance, particularly technical diagnostic accuracy and workflow coordination. Technical accuracy alone represents over one-third of all outcome labels and nearly half of peer-reviewed reports, indicating that DHT evaluations in LSDs remain centered on proximal process measures (how well tools detect, classify, or transmit information) rather than on longer-term clinical, patient-centered, or societal outcomes. This pattern is typical of an emerging innovation field: early studies prioritize technical reliability and feasibility, while robust assessments of clinical benefit, quality of life, HCP and patient experience, cost-effectiveness, equity, or broader public-health impact are still limited.

XSL•FO
RenderX

These patterns must be methodologically interpreted also within the constraints of the underlying evidence profile. Although title and abstract screening was primarily conducted by one reviewer (AMCS), we used calibration and second-reviewer (LN) audits; residual selection error remains possible. This review was explicitly descriptive and aimed to map what has been studied rather than to infer effectiveness. Across included records, signals for AI (eg, diagnostic support, early detection, and risk stratification) and CC (eg, televisits, teleconsultations, remote follow-up, and monitoring) are emerging, but the predominance of small, single-center, observational designs and the presence of gray literature limit generalizability and constrain inference. Consistent with the scoping-review methodology, we did not conduct a formal risk-of-bias assessment. Instead, we charted study-level features and caveats descriptively and used them to temper interpretation. We accordingly avoided causal language and treated findings as hypothesis-generating and as implementation cues to inform future, more robust evaluations (eg, multisite studies using transparent datasets, predefined comparators, and standardized outcome measures). Our interpretations prioritized peer-reviewed evidence, while policy and strategy documents were used as contextual background only; we did not advance causal or policy claims that were not supported by study design and reporting quality. Additional constraints include language and time-window limits and potential publication and selection biases typical of rapidly evolving digital scenarios.

Taken together, these outcome- and phase-level patterns delineate where evidence is accumulating versus where it is still largely hypothetical. They highlight a need for future evaluations that (1) move beyond accuracy and process metrics toward robust clinical efficacy measures, as well as patient-centered, system-level, and societal endpoints; (2) include later care-journey phases such as rehabilitation, transition to adulthood, and palliative care; and (3) explicitly measure equity, safety, and cost implications. The subsequent Discussion subsections on challenges, counterevidence, and future research interpret these gaps as priorities for staged, theory-informed evaluation rather than as a mandate for immediate large-scale implementation.

Challenges, Risks, and Biases

The evidence base also includes null or inconclusive findings, such as CC applications that did not improve adherence or use due to alert fatigue, workflow misalignment, or connectivity limitations; engagement platforms showing low sustained uptake; and AI models whose performance declined outside the derivation site or failed to outperform simpler baselines in small, heterogeneous cohorts [21,22,38,103,108].

Overall, Figures 2 and 4 suggest a functional split: AI is most often applied in early phases (screening, diagnosis, phenotyping), while CC supports continuity (follow-up, coordination, long-term monitoring). Given the largely descriptive evidence, deployment should be coupled with prospective, multisite, and equity-aware evaluations and transparent reporting [21,38,69,70].

AI-related risks reflect rare-disease data constraints: small single-center cohorts, incomplete or heterogeneous EHR data, missingness, and limited external validation drive overfitting and bias risk [21,38,69,70]. Proposed mitigation strategies include cross-registry pooling under findable, accessible, interoperable, and reusable principles, transfer learning, data augmentation or synthetic data, and routine auditing for bias and performance drift, aligned with emerging governance expectations (eg, EU AI Act) [103,164,171-173]. For CC deployments, the dominant concerns are interoperability with HIS and EHR, and equity (digital literacy, device, and connectivity access), such that CC can either bridge or widen the digital divide depending on implementation design [20-22,38,40,103,108,134,162,163].

Across included records, recurrent barriers to DHT scale-up in LSDs include infrastructure constraints, limited alignment across regulatory, HTA, and reimbursement requirements, and workforce skill gaps [38,128]. Interoperability, provider training, and robust privacy and security governance are repeatedly positioned as prerequisites, alongside patient trust, and clear accountability for data use [14,21,69,102].

Future Research Priorities

Given rapid innovation and evolving governance, multistakeholder consensus processes (eg, consensus conferences informed by PICO-structured evidence syntheses) can help translate emerging evidence into fit-for-purpose recommendations for development, adoption, and evaluation of AI and CC in LSD care and in RDs more broadly [35,36].

Most included evidence is nonrandomized, single-center, with small samples and limited follow-up, constraining causal inference and generalizability [37,71,72,80]. Priorities, therefore, include standardized outcome sets, interoperable data capture, and multisite designs (including pragmatic and, where feasible, randomized evaluations) that can assess longer-term effectiveness, safety, equity, and implementation outcomes [28,29,35,72,79,80]. In rare, geographically dispersed LSD populations, DHTs may need to act both as interventions and as trial enablers (supporting decentralized recruitment, continuous outcome capture, and longitudinal follow-up) [28,79].

Conclusions

This PRISMA-ScR–conformant review provides, to the authors' knowledge, the first comprehensive mapping of DHT applications in LSDs care, spanning disease populations, intervention classes, outcome domains, and care-journey phases. Across 245 records, research activity clustered around a subset of LSDs, particularly FD and GD, and was concentrated in earlier care-journey phases such as screening, diagnosis, and early monitoring, with markedly fewer applications reported in treatment intensification, rehabilitation, transition, or end-of-life care. Care-delivery outcomes (including diagnostic accuracy, workflow coordination, and access to specialist expertise) dominated reporting, whereas patient-centered and societal outcomes remained comparatively scarce. Figure 4 synthesizes how these digitally enhanced activities align with LSDs' natural history and conventional care pathways [24,102].

However, current evidence is insufficient to infer effectiveness or justify routine implementation. Across AI, CC, and other

XSL•FO
RenderX

DHTs, most studies were small, single-center, observational, and exploratory, with only a modest contribution from gray literature. Consistent with the aims of a scoping review, these findings should be interpreted as descriptive and hypothesis-generating. Our mapping is intended to characterize patterns and identify gaps, not to support causal claims or comparative-effectiveness judgments.

Considering these findings, near-term priorities include stronger evidence generation (through RCTs, high-quality real-world evidence studies, and pragmatic, theory-informed implementation trials of AI and CC technologies), using transparent datasets and standardized outcome sets aligned with established evidence standards. Given that the current evidence base is appreciable in size, but heterogeneous and largely observational, expert consensus processes are needed to translate emerging signals into appropriate, practice-relevant guidance and to define minimum requirements for safe and meaningful clinical use. Policy integration and cross-border collaboration are not conclusions drawn from the current data but represent necessary enabling conditions for high-quality data availability in RDs, harmonized governance, interoperability, and future evidence generation. The 4-axis mapping framework and care-journey perspective presented here may help guide the design of these future evaluations and consensus efforts for LSDs and, where appropriate, other RDs.

Acknowledgments

Consistently with *Journal of Medical Internet Research* Publications' guidance on generative artificial intelligence (GenAI), we disclose that GenAI tools, such as Consensus, SciSpace, and Connected Papers, were used to support operational refinement and expansion of the literature search strategy (eg, brainstorming and harmonizing candidate digital health technologies keywords later manually verified and incorporated by the authors) and for language-level editing and quality-control checks (grammar, clarity, translation, terminology harmonization, and internal consistency checks, including verification of internal cross-references to figures, tables, and multimedia appendices). These GenAI tools were not used for study selection decisions, data extraction, data analysis, evidence synthesis, interpretation, or formulation of conclusions. All suggested edits were reviewed, validated, and approved by the author team, who remain fully accountable for the manuscript content. The search prompts used are available in Multimedia Appendix 2 (AI-assisted search reproducibility logs).

Funding

The study was funded by Chiesi Global Rare Diseases. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

AMCS, LN, and MS conceptualized the article. AMCS and LN conducted the detailed literature review and analysis. AMCS provided expertise on digital health innovations and health policies, LN guided the methodology, and MS provided clinical expertise on lysosomal storage disorders and rare diseases. CB complemented the clinical perspective, and KA provided expertise from an international patient association network perspective. AMCS drafted the initial manuscript, which was integrated by LN, KA, and MS. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The 22-item PRISMA-ScR checklist for scoping reviews.
[DOCX File , 2397 KB - jmir_v28i1e73612_app1.docx]

Multimedia Appendix 2
Reproducibility logs of the scoping review search queries.
[DOCX File , 6639 KB - jmir_v28i1e73612_app2.docx]

Multimedia Appendix 3
Examples of AI-driven and Connected Care (CC) Digital Health Technology (DHT) applications to enhance Lysosomal Storage Disease care pathways across all stages.
[DOCX File , 3540 KB - jmir_v28i1e73612_app3.docx]

Multimedia Appendix 4
Examples of challenges for systemic integration of AI-driven and Connected Care (CC) Digital Health Technologies (DHTs).
[DOCX File , 4175 KB - jmir_v28i1e73612_app4.docx]

XSL•FO
RenderX

Multimedia Appendix 5
Comparative and distribution evidence charting tables across LSD population, digital health interventions, patient-healthcare-societal outcomes, and care journey phases.
[DOCX File , 4518 KB - jmir_v28i1e73612_app5.docx]

References

1. Adachi T, El-Hattab AW, Jain R, Nogales Crespo KA, Quirland Lazo CI, Scarpa M, et al. Enhancing equitable access to rare disease diagnosis and treatment around the world: a review of evidence, policies, and challenges. Int J Environ Res Public Health 2023;20(6):4732 [FREE Full text] [doi: 10.3390/ijerph20064732] [Medline: 36981643]
2. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet 2020;28(2):165-173 [FREE Full text] [doi: 10.1038/s41431-019-0508-0] [Medline: 31527858]
3. Rare diseases. European Commission. 2024. URL: https://health.ec.europa.eu/rare-diseases-and-european-reference-networks/rare-diseases_en [accessed 2026-01-01]
4. Tanpaiboon P. Practical management of lysosomal storage disorders (LSDs). Transl Sci Rare Dis 2019;4(3-4):133-157. [doi: 10.3233/trd-190047]
5. Addressing the challenges of persons living with a rare disease and their families : revised draft resolution / Brazil, Central African Republic, Côte d'Ivoire, Cyprus, Ecuador, Equatorial Guinea, France, Italy, Peru, Portugal, Qatar, South Africa, Spain, Ukraine and Vanuatu. United Nations Digital Library. 2021. URL: https://digitallibrary.un.org/record/3947524 [accessed 2026-01-01]
6. Donald A, Cizer H, Finnegan N, Collin-Histed T, Hughes DA, Davies EH. Measuring disease activity and patient experience remotely using wearable technology and a mobile phone app: outcomes from a pilot study in Gaucher disease. Orphanet J Rare Dis 2019;14(1):212 [FREE Full text] [doi: 10.1186/s13023-019-1182-6] [Medline: 31488169]
7. Hurvitz N, Dinur T, Revel-Vilk S, Agus S, Berg M, Zimran A, et al. A feasibility open-labeled clinical trial using a second-generation artificial-intelligence-based therapeutic regimen in patients with Gaucher disease treated with enzyme replacement therapy. J Clin Med 2024;13(11):3325 [FREE Full text] [doi: 10.3390/jcm13113325] [Medline: 38893036]
8. Spreafico AMC, Tarricone R, Stern AD. Comprehensive policies for scaling systemic and equitable integration of digital health technologies. NPJ Digit Med 2025;8(1):404 [FREE Full text] [doi: 10.1038/s41746-025-01820-x] [Medline: 40615530]
9. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). European Union EUR-Lex. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj [accessed 2026-01-01]
10. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473. [doi: 10.7326/m18-0850] [Medline: 30178033]
11. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005;8(1):19-32. [doi: 10.1080/1364557032000119616]
12. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
13. Millington DS, Sista R, Eckhardt A, Rouse J, Bali D, Goldberg R, et al. Digital microfluidics: a future technology in the newborn screening laboratory? Semin Perinatol 2010;34(2):163-169 [FREE Full text] [doi: 10.1053/j.semperi.2009.12.008] [Medline: 20207266]
14. Millington DS, Bali DS. Current state of the art of newborn screening for lysosomal storage disorders. Int J Neonatal Screen 2018;4(3):24 [FREE Full text] [doi: 10.3390/ijns4030024] [Medline: 33072946]
15. Shur N, Atabaki SM, Kisling MS, Tabarani A, Williams C, Fraser JL, et al. Rapid deployment of a telemedicine care model for genetics and metabolism during COVID-19. Am J Med Genet A 2021;185(1):68-72 [FREE Full text] [doi: 10.1002/ajmg.a.61911] [Medline: 33051968]
16. Kobrinskii BA, Blagosklonov NA, Demikova NS, Nikolaeva EA, Kotalevskaya YY, Melikyan LP, et al. A computer system for differential diagnosis of lysosomal storage diseases based on artificial intelligence methods [Article in Russian]. Bull Sib Med 2022;21(2):67-73. [doi: 10.20538/1682-0363-2022-2-67-73]
17. Aydemir S, Ocak S, Saygılı S, Hopurcuoğlu D, Haşlak F, Kıykım E, et al. Telemedicine applications in a tertiary pediatric hospital in Turkey during COVID-19 pandemic. Telemed J E Health 2021;27(10):1180-1187. [doi: 10.1089/tmj.2020.0381] [Medline: 33301373]
18. Bouwman MG, Teunissen QGA, Wijburg FA, Linthorst GE. 'Doctor Google' ending the diagnostic odyssey in lysosomal storage disorders: parents using internet search engines as an efficient diagnostic strategy in rare diseases. Arch Dis Child 2010;95(8):642-644. [doi: 10.1136/adc.2009.171827] [Medline: 20418338]
19. Scherr JF, Albright C, de Los Reyes E. Utilizing telehealth to create a clinical model of care for patients with batten disease and other rare diseases. Ther Adv Rare Dis 2021;2:26330040211038564 [FREE Full text] [doi: 10.1177/26330040211038564] [Medline: 37181116]

XSL•FO
RenderX

20. Uribe-Carretero E, Rey V, Fuentes JM, Tamargo-Gómez I. Lysosomal dysfunction: connecting the dots in the landscape of human diseases. Biology (Basel) 2024;13(1):34 [FREE Full text] [doi: 10.3390/biology13010034] [Medline: 38248465]
21. Elstein D, Giugliani R, Muenzer J, Schenk J, Schwartz IVD, Anagnostopoulou C. Impact of the COVID-19 pandemic on the standard of care for patients with lysosomal storage diseases: a survey of healthcare professionals in the Fabry, Gaucher, and Hunter Outcome Survey registries. Mol Genet Metab Rep 2021;28:100788 [FREE Full text] [doi: 10.1016/j.ymgmr.2021.100788] [Medline: 34367919]
22. Nuovo Monitor 47: Telemedicina e Intelligenza Artificiale a Supporto Dell'assistenza Territoriale. Vol II. Agenas [Web page in Italian]. AGENAS - Agenzia Nazionale per i Servizi Sanitari Regionali. 2022. URL: https://www.agenas.gov.it/monitor-47 [accessed 2026-01-01]
23. Nishith N, Siddiqui SH, R Raja SK, Agrawal N, Phadke S, Sharma S. Utility of morphologic assessment of bone marrow biopsy in diagnosis of lysosomal storage disorders. Indian J Pathol Microbiol 2023;66(1):91-95 [FREE Full text] [doi: 10.4103/ijpm.ijpm_725_21] [Medline: 36656217]
24. D'Angelantonio E, Tocci T, Sansone L, Mencattini A, Russo M, Pallotti A. Fabry disease automatic recognition through image processing. : IEEE; 2022 Presented at: Proceedings of the 2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA); June 22-24, 2022; Messina, Italy p. 1-3. [doi: 10.1109/memea54994.2022.9856583]
25. Montella A, Tranfa M, Scaravilli A, Barkhof F, Brunetti A, Cole J, et al. Assessing brain involvement in Fabry disease with deep learning and the brain-age paradigm. Hum Brain Mapp 2024;45(5):e26599 [FREE Full text] [doi: 10.1002/hbm.26599] [Medline: 38520360]
26. Uryu H, Migita O, Ozawa M, Kamijo C, Aoto S, Okamura K, et al. Automated urinary sediment detection for Fabry disease using deep-learning algorithms. Mol Genet Metab Rep 2022;33:100921 [FREE Full text] [doi: 10.1016/j.ymgmr.2022.100921] [Medline: 36186840]
27. Moura ARS. Detecting important electrocardiogram characteristics for the diagnosis of Fabry disease via statistics and machine learning techniques. ProQuest. 2022. URL: https://search.proquest.com/openview/372a62f0566d6dd4e2056eff49218b4b/1?pq-origsite=gscholar&cbl=2026366&diss=y [accessed 2026-01-01]
28. Jefferies J, Aguiar P, Biondetti G, Warnock D, Kallish S, Nelson M, et al. Estimation of arrhythmia risk in patients with Fabry disease using a machine learning model. J Heart Lung Transplant 2023;42(4):S331. [doi: 10.1016/j.healun.2023.02.765]
29. Vijapurapu R. Multi parametric cardiac magnetic resonance imaging and arrhythmic risk stratification in Fabry disease. University of Birmingham - Institute of Cardiovascular Sciences. 2023. URL: http://etheses.bham.ac.uk/id/eprint/13276/9/Vijapurapu2023PhD_Redacted.pdf [accessed 2026-01-01]
30. Akiki E, Abbasi M, El-Am E, Geske JB, Bos M, Ackerman MJ, et al. Artificial intelligence electrocardiography for the evaluation of cardiac involvement in Fabry disease. J Am Coll Cardiol 2024;83(13):2581. [doi: 10.1016/s0735-1097(24)04571-6]
31. Vijapurapu R, Nordin S, Baig S, Liu B, Rosmini S, Augusto J, et al. Global longitudinal strain, myocardial storage and hypertrophy in Fabry disease. Heart 2019;105(6):470-476. [doi: 10.1136/heartjnl-2018-313699] [Medline: 30282640]
32. Hwang IC, Chun EJ, Kim PK, Kim M, Park J, Choi HM, et al. Artificial intelligence-derived extracellular volume fraction for diagnosis and prognostication in patients with light-chain cardiac amyloidosis. medRxiv 2024 Mar 22. [doi: 10.1101/2024.03.20.24304642]
33. Beck M, Ramaswami U, Hernberg-Ståhl E, Hughes DA, Kampmann C, Mehta AB, et al. Twenty years of the Fabry outcome survey (FOS): insights, achievements, and lessons learned from a global patient registry. Orphanet J Rare Dis 2022;17(1):238 [FREE Full text] [doi: 10.1186/s13023-022-02392-9] [Medline: 35725623]
34. Woodard J, Zheng W, Zhang Y. Protein structural features predict responsiveness to pharmacological chaperone treatment for three lysosomal storage disorders. PLoS Comput Biol 2021;17(9):e1009370 [FREE Full text] [doi: 10.1371/journal.pcbi.1009370] [Medline: 34529671]
35. Tebani A, Mauhin W, Abily-Donval L, Lesueur C, Berger MG, Nadjar Y, et al. A proteomics-based analysis reveals predictive biological patterns in Fabry disease. J Clin Med 2020;9(5):1325 [FREE Full text] [doi: 10.3390/jcm9051325] [Medline: 32370284]
36. Tebani A, Barbey F, Dormond O, Ducatez F, Marret S, Nowak A, et al. Deep next-generation proteomics and network analysis reveal systemic and tissue-specific patterns in Fabry disease. Transl Res 2023;258:47-59. [doi: 10.1016/j.trsl.2023.02.006] [Medline: 36863609]
37. Michalski AA, Lis K, Stankiewicz J, Kloska SM, Sycz A, Dudziński M, et al. Supporting the diagnosis of fabry disease using a natural language processing-based approach. J Clin Med 2023;12(10):3599 [FREE Full text] [doi: 10.3390/jcm12103599] [Medline: 37240705]
38. Laney D, Germain D, Oliveira J, Burlina AP, Cabrera GH, Hong GR, et al. Fabry disease and COVID-19: international expert recommendations for management based on real-world experience. Clin Kidney J 2020;13(6):913-925 [FREE Full text] [doi: 10.1093/ckj/sfaa227] [Medline: 33391734]
39. Riccio E, Pieroni M, Limoneglli G, Pisani A. Impact of COVID-19 pandemic on patients with fabry disease: an Italian experience. Mol Genet Metab 2020;131(1-2):124-125 [FREE Full text] [doi: 10.1016/j.ymgme.2020.07.008] [Medline: 32741663]

XSL•FO
RenderX

40. Nowicki M, Bazan-Socha S, Kłopotowski M, Błażejewska-Hyżorek B, Kusztal M, Pawlaczyk K, et al. Considerations for home-based treatment of fabry disease in Poland during the COVID-19 pandemic and beyond. Int J Environ Res Public Health 2021;18(16):8242 [FREE Full text] [doi: 10.3390/ijerph18168242] [Medline: 34443990]
41. Study based on electronic health records to identify patients at high-risk of Fabry disease (HOPE Fabry). ClinicalTrials.gov. 2024. URL: https://clinicaltrials.gov/study/NCT05671770?cond=Lysosomal%20Storage%20Diseases&intr=electronic&rank=1 [accessed 2026-01-01]
42. D'Amore S, Mckie M, Fahey A, Bleloch D, Grillo G, Hughes M, et al. Fabry app: the value of a portable technology in recording day-to-day patient monitored information in patients with Fabry disease. Orphanet J Rare Dis 2024;19(1):13 [FREE Full text] [doi: 10.1186/s13023-023-02999-6] [Medline: 38212814]
43. Gragnaniello V, Burlina AP, Commone A, Gueraldi D, Puma A, Porcù E, et al. Newborn screening for fabry disease: current status of knowledge. Int J Neonatal Screen 2023;9(2):31 [FREE Full text] [doi: 10.3390/ijns9020031] [Medline: 37367212]
44. Implementation and evaluation of a rare disease algorithm to identify persons at risk of Gaucher disease using data from electronic health records (EHRs) in the United States (project searchlight). ClinicalTrials.gov. 2024. URL: https://clinicaltrials.gov/study/NCT05908656?cond=Lysosomal%20Storage%20Diseases&intr=electronic&rank=2 [accessed 2026-01-01]
45. Zanetti A, D'Avanzo F, Bertoldi L, Zampieri G, Feltrin E, De Pascale F, et al. Setup and validation of a targeted next-generation sequencing approach for the diagnosis of lysosomal storage disorders. J Mol Diagn 2020;22(4):488-502 [FREE Full text] [doi: 10.1016/j.jmoldx.2020.01.010] [Medline: 32036093]
46. Málaga DR, Brusius-Facchin AC, Siebert M, Pasqualim G, Saraiva-Pereira ML, Souza CFD, et al. Sensitivity, advantages, limitations, and clinical utility of targeted next-generation sequencing panels for the diagnosis of selected lysosomal storage disorders. Genet Mol Biol 2019;42(1 suppl 1):197-206 [FREE Full text] [doi: 10.1590/1678-4685-GMB-2018-0092] [Medline: 30985853]
47. Biegstraaten M, Arngrímsson R, Barbey F, Boks L, Cecchi F, Deegan PB, et al. Recommendations for initiation and cessation of enzyme replacement therapy in patients with Fabry disease: The European fabry working group consensus document. Orphanet J Rare Dis 2015;10:36 [FREE Full text] [doi: 10.1186/s13023-015-0253-6] [Medline: 25885911]
48. Berry L, Walter J, Johnson J, Alton J, Powers J, Llòria X, et al. Patient-reported experience with Fabry disease and its management in the real-world setting: Results from a double-blind, cross-sectional survey of 280 respondents. Orphanet J Rare Dis 2024;19(1):153 [FREE Full text] [doi: 10.1186/s13023-024-03090-4] [Medline: 38605390]
49. Kadali S, Naushad SM, Radha Rama Devi A, Bodiga VL. Biochemical, machine learning and molecular approaches for the differential diagnosis of mucopolysaccharidoses. Mol Cell Biochem 2019;458(1-2):27-37. [doi: 10.1007/s11010-019-03527-6] [Medline: 30903511]
50. Sharma GB, Robertson DD, Laney DA, Gambello MJ, Terk M. Machine learning based analytics of micro-MRI trabecular bone microarchitecture and texture in type 1 Gaucher disease. J Biomech 2016;49(9):1961-1968. [doi: 10.1016/j.jbiomech.2016.04.010] [Medline: 27109052]
51. Valero-Tena E, Roca-Espiau M, Verdú-Díaz J, Diaz-Manera J, Andrade-Campos M, Giraldo P. Advantages of digital technology in the assessment of bone marrow involvement in Gaucher's disease. Front Med (Lausanne) 2023;10:1098472 [FREE Full text] [doi: 10.3389/fmed.2023.1098472] [Medline: 37250646]
52. Yu B, Whitmarsh T, Riede P, McDonald S, Kaggie JD, Cox TM, et al. Deep learning-based quantification of osteonecrosis using magnetic resonance images in Gaucher disease. Bone 2024;186:117142 [FREE Full text] [doi: 10.1016/j.bone.2024.117142] [Medline: 38834102]
53. Azuri I, Wattad A, Peri-Hanania K, Kashti T, Rosen R, Caspi Y, et al. A deep-learning approach to spleen volume estimation in patients with Gaucher disease. J Clin Med 2023;12(16):5361 [FREE Full text] [doi: 10.3390/jcm12165361] [Medline: 37629403]
54. Wilson A, Chiorean A, Aguiar M, Sekulic D, Pavlick P, Shah N, et al. Development of a rare disease algorithm to identify persons at risk of Gaucher disease using electronic health records in the United States. Orphanet J Rare Dis 2023;18(1):280 [FREE Full text] [doi: 10.1186/s13023-023-02868-2] [Medline: 37689674]
55. Tenenbaum A, Revel-Vilk S, Gazit S, Roimi M, Gill A, Gilboa D, et al. A machine learning model for early diagnosis of type 1 Gaucher disease using real-life data. J Clin Epidemiol 2024;175:111517. [doi: 10.1016/j.jclinepi.2024.111517] [Medline: 39245415]
56. Revel-Vilk S. Digital health platform customized for patients with Gaucher disease. ClinicalTrials.gov. 2024. URL: https://clinicaltrials.gov/study/NCT04997772?cond=Lysosomal%20Storage%20Diseases&viewType=Card&intr=digital&rank=1 [accessed 2026-01-01]
57. Lobel CJ, Laney DA, Yang J, Jacob D, Rickheim A, Ogg CZ, et al. FDrisk: development of a validated risk assessment tool for fabry disease utilizing electronic health record data. J Rare Dis (Berlin) 2024;3(1):2 [FREE Full text] [doi: 10.1007/s44162-023-00026-7] [Medline: 38187171]
58. Andrade-Campos MM, de Frutos LL, Cebolla JJ, Serrano-Gonzalo I, Medrano-Engay B, Roca-Espiau M, et al. Identification of risk features for complication in Gaucher's disease patients: a machine learning analysis of the Spanish registry of Gaucher disease. Orphanet J Rare Dis 2020;15(1):256 [FREE Full text] [doi: 10.1186/s13023-020-01520-7] [Medline: 32962737]

XSL•FO
RenderX

59. Karabul N, Rose F, Cipra G. Home treatment for lysosomal diseases during COVID-19: German experience. Mol Genet Metab 2021;132(2):S54. [doi: 10.1016/j.ymgme.2020.12.119]
60. Elmonem MA, Belanger-Quintana A, Bordugo A, Boruah R, Cortès-Saladelafont E, Endrakanti M, et al. The impact of COVID-19 pandemic on the diagnosis and management of inborn errors of metabolism: a global perspective. Mol Genet Metab 2020;131(3):285-288 [FREE Full text] [doi: 10.1016/j.ymgme.2020.09.004] [Medline: 33004274]
61. La Cognata V, Guarnaccia M, Morello G, Ruggieri M, Polizzi A, Cavallaro S. Design and validation of a custom NGS panel targeting a set of lysosomal storage diseases candidate for NBS applications. Int J Mol Sci 2021;22(18):10064 [FREE Full text] [doi: 10.3390/ijms221810064] [Medline: 34576242]
62. Ryan E, Jong T, Sidransky E. Newborn screening in Gaucher disease: a bright and complicated future. OBM Genet 2022;06(03):1-21. [doi: 10.21926/obm.genet.2203165]
63. Hirachan R, Horman A, Burke D, Heales S. Evaluation, in a highly specialised enzyme laboratory, of a digital microfluidics platform for rapid assessment of lysosomal enzyme activity in dried blood spots. JIMD Rep 2024;65(2):124-131 [FREE Full text] [doi: 10.1002/jmd2.12413] [Medline: 38444576]
64. D'Souza A, Ryan E, Sidransky E. Facial features of lysosomal storage disorders. Expert Rev Endocrinol Metab 2022;17(6):467-474 [FREE Full text] [doi: 10.1080/17446651.2022.2144229] [Medline: 36384353]
65. Bajaj S, Thawani S, Iyer G, Setty PN, Kore M, Jadhav V. Telegenetics aids the diagnosis of hunter syndrome caused due to a novel IDS variant in rural India, during COVID-19 pandemic. Int J Contemp Pediatr 2022;9(9):851. [doi: 10.18203/2349-3291.ijcp20222126]
66. İnci A, Okur İ, Tümer L, Biberoğlu G, Öktem M, Ezgü F. Clinical and event-based outcomes of patients with mucopolysaccharidosis VI receiving enzyme replacement therapy in Turkey: a case series. Orphanet J Rare Dis 2021;16(1):438 [FREE Full text] [doi: 10.1186/s13023-021-02060-4] [Medline: 34666789]
67. Millington D, Norton S, Singh R, Sista R, Srinivasan V, Pamula V. Digital microfluidics comes of age: High-throughput screening to bedside diagnostic testing for genetic disorders in newborns. Expert Rev Mol Diagn 2018;18(8):701-712 [FREE Full text] [doi: 10.1080/14737159.2018.1495076] [Medline: 30004274]
68. Mistry PK, Kishnani P, Wanner C, Dong D, Bender J, Batista JL, et al. Rare lysosomal disease registries: lessons learned over three decades of real-world evidence. Orphanet J Rare Dis 2022;17(1):362 [FREE Full text] [doi: 10.1186/s13023-022-02517-0] [Medline: 36244992]
69. Lin S, Nateqi J, Weingartner-Ortner R, Gruarin S, Marling H, Pilgram V, et al. An artificial intelligence-based approach for identifying rare disease patients using retrospective electronic health records applied for Pompe disease. Front Neurol 2023;14:1108222 [FREE Full text] [doi: 10.3389/fneur.2023.1108222] [Medline: 37153672]
70. Rustamov J, Rustamov Z, Mohamad MS, Zaki N, Al Tenaiji A, Al Harbi M, et al. An expert rule-based approach for identifying infantile-onset pompe disease patients using retrospective electronic health records. Sci Rep 2024;14(1):21523 [FREE Full text] [doi: 10.1038/s41598-024-72259-5] [Medline: 39277702]
71. Gómez-Cebrián N, Gras-Colomer E, Poveda Andrés JL, Pineda-Lucena A, Puchades-Carrasco L. Omics-based approaches for the characterization of Pompe disease metabolic phenotypes. Biology (Basel) 2023;12(9):1159 [FREE Full text] [doi: 10.3390/biology12091159] [Medline: 37759559]
72. Esmail S, Danter WR. DeepNEU: artificially induced stem cell (aiPSC) and differentiated skeletal muscle cell (aiSkMC) simulations of infantile onset POMPE disease (IOPD) for potential biomarker identification and drug discovery. Front Cell Dev Biol 2019;7:325 [FREE Full text] [doi: 10.3389/fcell.2019.00325] [Medline: 31867331]
73. Brunetti-Pierri N, Fecarotta S, Staiano A, Strisciuglio P, Parenti G. Ensuring continuity of care for children with inherited metabolic diseases at the time of COVID-19: the experience of a metabolic unit in Italy. Genet Med 2020;22(7):1178-1180 [FREE Full text] [doi: 10.1038/s41436-020-0831-4] [Medline: 32409735]
74. Priya K. ClinicalTrials.gov. 2021. URL: https://clinicaltrials.gov/study/NCT02950298?cond=Lysosomal%20Storage%20Diseases&intr=telemedicine&rank=1 [accessed 2026-01-01]
75. Ricci G, Baldanzi S, Seidita F, Proietti C, Carlini F, Peviani S, et al. A mobile app for patients with Pompe disease and its possible clinical applications. Neuromuscul Disord 2018;28(6):471-475. [doi: 10.1016/j.nmd.2018.03.005] [Medline: 29655528]
76. Claramunt M, Idelssonhn S, James M, Corti M, Anton V, Byrne B, et al. P.84 Gait analysis of patients with Pompe disease using a portable system. Neuromuscular Disorders 2022;32:S76. [doi: 10.1016/j.nmd.2022.07.143]
77. Hamed A, Curran C, Gwaltney C, DasMahapatra P. Mobility assessment using wearable technology in patients with late-onset Pompe disease. NPJ Digit Med 2019;2:70 [FREE Full text] [doi: 10.1038/s41746-019-0143-8] [Medline: 31341956]
78. Harfouche M, Kishnani PS, Krusinska E, Gault J, Sitaraman S, Sowinski A, et al. Use of the patient-reported outcomes measurement information system (PROMIS®) to assess late-onset Pompe disease severity. J Patient Rep Outcomes 2020;4(1):83 [FREE Full text] [doi: 10.1186/s41687-020-00245-2] [Medline: 33034771]
79. Sniderman King L, Pan Y, Nallamilli BRR, Hegde M, Jagannathan L, Ramachander V, et al. Pompe disease ascertained through the lantern project, 2018-2021: next-generation sequencing and enzymatic testing to overcome obstacles to diagnosis. Mol Genet Metab 2023;139(1):107565 [FREE Full text] [doi: 10.1016/j.ymgme.2023.107565] [Medline: 37087815]

XSL•FO
RenderX

80. Pruniski B, Lisi E, Ali N. Newborn screening for Pompe disease: impact on families. J Inherit Metab Dis 2018;41(6):1189-1203. [doi: 10.1007/s10545-018-0159-2] [Medline: 29594646]
81. Carmona CJ, German-Morales M, Elizondo D, Ruiz-Rodado V, Grootveld M. Urinary metabolic distinction of Niemann-Pick class 1 disease through the use of subgroup discovery. Metabolites 2023;13(10):1079 [FREE Full text] [doi: 10.3390/metabo13101079] [Medline: 37887404]
82. Moreno-Barea FJ, Franco L, Elizondo D, Grootveld M. Data augmentation techniques to improve metabolomic analysis in Niemann-Pick type C disease. : Springer International Publishing; 2022 Presented at: Proceedings of the Computational Science – ICCS 2022 22nd International Conference; June 21–23, 2022; London, UK p. 78-91. [doi: 10.1007/978-3-031-08757-8_8]
83. Minter Baerg MM, Stoway SD, Hart J, Mott L, Peck DS, Nett SL, et al. Precision newborn screening for lysosomal disorders. Genet Med 2018;20(8):847-854 [FREE Full text] [doi: 10.1038/gim.2017.194] [Medline: 29120458]
84. Tillema J, Derks MG, Pouwels PJW, de Graaf P, van Rappard DF, Barkhof F, et al. Volumetric MRI data correlate to disease severity in metachromatic leukodystrophy. Ann Clin Transl Neurol 2015;2(9):932-940 [FREE Full text] [doi: 10.1002/acn3.232] [Medline: 26401514]
85. Clas P, Groeschel S, Wilke M. A semi-automatic algorithm for determining the demyelination load in metachromatic leukodystrophy. Acad Radiol 2012;19(1):26-34. [doi: 10.1016/j.acra.2011.09.008] [Medline: 22142678]
86. Esmail S, Danter WR. Artificially induced pluripotent stem cell-derived whole-brain organoid for modelling the pathophysiology of metachromatic leukodystrophy and drug repurposing. Biomedicines 2021;9(4):440 [FREE Full text] [doi: 10.3390/biomedicines9040440] [Medline: 33923989]
87. Regier DS, Kwon HJ, Johnston J, Golas G, Yang S, Wiggs E, et al. MRI/MRS as a surrogate marker for clinical progression in GM1 gangliosidosis. Am J Med Genet A 2016;170(3):634-644. [doi: 10.1002/ajmg.a.37468] [Medline: 26646981]
88. Davies EH, Johnston J, Toro C, Tifft CJ. A feasibility study of mHealth and wearable technology in late onset GM2 gangliosidosis (Tay-Sachs and Sandhoff Disease). Orphanet J Rare Dis 2020;15(1):199 [FREE Full text] [doi: 10.1186/s13023-020-01473-x] [Medline: 32746863]
89. Ruohola A, Salli E, Roine T, Tokola A, Laine M, Tikkanen R, et al. Detection of aspartylglucosaminuria patients from magnetic resonance images by a machine-learning-based approach. Brain Sci 2022;12(11):1522 [FREE Full text] [doi: 10.3390/brainsci12111522] [Medline: 36358448]
90. Kobrinskii BA, Blagosklonov NA, Gribova VC, Shalfeeva EA. Expert system for the diagnosis of orphan diseases. : Springer International Publishing; 2023 Presented at: Proceedings of the Sixth International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’22); October 31 – November 6, 2022; Istanbul, Turkey p. 251-260. [doi: 10.1007/978-3-031-19620-1_24]
91. Blagosklonov N, Kobrinskii B. Model of integral evaluation of expert knowledge for the diagnosis of lysosomal storage diseases. CEUR Workshop Proceedings 2020;2648(22):1-20 [FREE Full text]
92. Kamel MA, Abbas MT, Kanaan CN, Awad KA, Baba Ali N, Scalia IG, et al. How artificial intelligence can enhance the diagnosis of cardiac amyloidosis: a review of recent advances and challenges. J Cardiovasc Dev Dis 2024;11(4):118 [FREE Full text] [doi: 10.3390/jcdd11040118] [Medline: 38667736]
93. Martini N, Aimo A, Barison A, Della Latta D, Vergaro G, Aquaro GD, et al. Deep learning to diagnose cardiac amyloidosis from cardiovascular magnetic resonance. J Cardiovasc Magn Reson 2020;22(1):84 [FREE Full text] [doi: 10.1186/s12968-020-00690-4] [Medline: 33287829]
94. Aimo A, Martini N, Barison A. 70 Deep learning to diagnose cardiac amyloidosis from cardiac magnetic resonance findings. Eur Heart J Suppl 2020;22(Supplement_N):N116-N130. [doi: 10.1093/eurheartj/suaa205]
95. Hsieh TC, Lesmann H, Krawitz PM. Facilitating the molecular diagnosis of rare genetic disorders through facial phenotypic scores. Curr Protoc 2023;3(10):e906. [doi: 10.1002/cpz1.906] [Medline: 37812136]
96. Reiter AMV, Pantel JT, Danyel M, Horn D, Ott C, Mensah MA. Validation of 3 computer-aided facial phenotyping tools (DeepGestalt, GestaltMatcher, and D-Score): comparative diagnostic accuracy study. J Med Internet Res 2024;26:e42904 [FREE Full text] [doi: 10.2196/42904] [Medline: 38477981]
97. Liévin V, Hansen JM, Lund A, Elstein D, Matthiesen ME, Elomaa K, et al. FindZebra online search delving into rare disease case reports using natural language processing. PLOS Digit Health 2023;2(6):e0000269 [FREE Full text] [doi: 10.1371/journal.pdig.0000269] [Medline: 37384616]
98. Pankow A, Meißner N, Fouquette L, Türk E, Gilbert S, Tübler T, et al. POS1551-HPR medical expert knowledge meets AI: how expert knowledge can improve symptom assessment apps - a new approach in rare diseases. Ann Rheum Dis 2022;81:1120. [doi: 10.1136/annrheumdis-2022-eular.3734]
99. Gómez J, Artigas L, Valls R, Gervas-Arruga J. An in silico approach to identify early damage biomarker candidates in metachromatic leukodystrophy. Mol Genet Metab Rep 2023;35:100974 [FREE Full text] [doi: 10.1016/j.ymgmr.2023.100974] [Medline: 37275681]
100. Pade LR, Stepler KE, Portero EP, DeLaney K, Nemes P. Biological mass spectrometry enables spatiotemporal 'omics: from tissues to cells to organelles. Mass Spectrom Rev 2024;43(1):106-138 [FREE Full text] [doi: 10.1002/mas.21824] [Medline: 36647247]

XSL•FO
RenderX

101. Pantel JT, Hajjir N, Danyel M, Elsner J, Abad-Perez AT, Hansen P, et al. Efficiency of computer-aided facial phenotyping (DeepGestalt) in individuals with and without a genetic syndrome: Diagnostic accuracy study. J Med Internet Res 2020;22(10):e19263 [FREE Full text] [doi: 10.2196/19263] [Medline: 33090109]
102. Akgun A, Gokcay G, Mungan NO, Sivri HS, Tezer H, Zeybek CA, et al. Expert-opinion-based guidance for the care of children with lysosomal storage diseases during the COVID-19 pandemic: an experience-based Turkey perspective. Front Public Health 2023;11:1092895 [FREE Full text] [doi: 10.3389/fpubh.2023.1092895] [Medline: 36794069]
103. Koracin V, Loeber JG, Mlinaric M, Battelino T, Bonham JR, Groselj U, et al. Global impact of COVID-19 on newborn screening programmes. BMJ Glob Health 2022;7(3):e007780 [FREE Full text] [doi: 10.1136/bmjgh-2021-007780] [Medline: 35236661]
104. Lampe C, Dionisi-Vici C, Bellettato CM, Paneghetti L, van Lingen C, Bond S, et al. The impact of COVID-19 on rare metabolic patients and healthcare providers: results from two MetabERN surveys. Orphanet J Rare Dis 2020;15(1):341 [FREE Full text] [doi: 10.1186/s13023-020-01619-x] [Medline: 33272301]
105. Mlinaric M, Bonham JR, Kožich V, Kölker S, Majek O, Battelino T, et al. Newborn screening in a pandemic-lessons learned. Int J Neonatal Screen 2023;9(2):21 [FREE Full text] [doi: 10.3390/ijns9020021] [Medline: 37092515]
106. Özalp Akın E, Eminoğlu FT, Doğulu N, Koç Yekeduz M, Öncül U, Akpınar F, et al. Unmet needs of children with inherited metabolic disorders in the COVID-19 pandemic. Turk Arch Pediatr 2022;57(3):335-341 [FREE Full text] [doi: 10.5152/TurkArchPediatr.2022.21367] [Medline: 35781238]
107. Ramaswami U, D'Amore S, Finnegan N, Hughes D, Kazemi M, Lysosomal Disorders Team, Royal Free London NHS Foundation Trust. Impact of SARS-CoV-2 (COVID-19) pandemic on patients with lysosomal storage disorders and restoration of services: experience from a specialist centre. Intern Med J 2021;51(10):1580-1593 [FREE Full text] [doi: 10.1111/imj.15473] [Medline: 34487419]
108. Schwartz IVD, Randon DN, Monsores N, Moura de Souza CF, Horovitz DDG, Wilke MVMB, et al. SARS-CoV-2 pandemic in the Brazilian community of rare diseases: a patient reported survey. Am J Med Genet C Semin Med Genet 2021;187(3):301-311 [FREE Full text] [doi: 10.1002/ajmg.c.31883] [Medline: 33474836]
109. Tobór-Świętek E, Sykut-Cegielska J, Bik-Multanowski M, Walczak M, Rokicki D, Kałużny Ł, et al. COVID-19 pandemic and patients with rare inherited metabolic disorders and rare autoinflammatory diseases-organizational challenges from the point of view of healthcare providers. J Clin Med 2021;10(21):4862 [FREE Full text] [doi: 10.3390/jcm10214862] [Medline: 34768381]
110. Zubarioglu T, Hopurcuoglu D, Uygur E, Ahmadzada S, Oge-Enver E, Isat E, et al. The impact of telemedicine for monitoring and treatment of phenylketonuria patients on metabolic outcome during coronavirus disease-19 outbreak. Telemed J E Health 2022;28(2):258-265. [doi: 10.1089/tmj.2020.0569] [Medline: 33719565]
111. Paneghetti L, Bellettato CM, Sechi A, Stepien KM, Scarpa M. One year of COVID-19: infection rates and symptoms in patients with inherited metabolic diseases followed by MetabERN. Orphanet J Rare Dis 2022;17(1):109 [FREE Full text] [doi: 10.1186/s13023-022-02247-3] [Medline: 35246188]
112. Sun A. Lysosomal storage disease overview. Ann Transl Med 2018;6(24):476 [FREE Full text] [doi: 10.21037/atm.2018.11.39] [Medline: 30740407]
113. Tumiene B, Graessner H, Mathijssen IM, Pereira AM, Schaefer F, Scarpa M, et al. European reference networks: challenges and opportunities. J Community Genet 2021;12(2):217-229 [FREE Full text] [doi: 10.1007/s12687-021-00521-8] [Medline: 33733400]
114. Heard JM, Bellettato C, van Lingen C, Scarpa M, MetabERN collaboration group. Research activity and capability in the European reference network metabERN. Orphanet J Rare Dis 2019;14(1):119 [FREE Full text] [doi: 10.1186/s13023-019-1091-8] [Medline: 31142374]
115. Talarico R, Aguilera S, Alexander T, Amoura Z, Andersen J, Arnaud L, et al. The added value of a European reference network on rare and complex connective tissue and musculoskeletal diseases: Insights after the first 5 years of the ERN ReCONNET. Clin Exp Rheumatol 2022;40 Suppl 134(5):3-11 [FREE Full text] [doi: 10.55563/clinexprheumatol/d2qz38] [Medline: 35349419]
116. Rossi A, Hoogeveen IJ, Lubout CMA, de Boer F, Fokkert-Wilts MJ, Rodenburg IL, et al. A generic emergency protocol for patients with inborn errors of metabolism causing fasting intolerance: a retrospective, single-center study and the generation of www.emergencyprotocol.net. J Inherit Metab Dis 2021;44(5):1124-1135 [FREE Full text] [doi: 10.1002/jimd.12386] [Medline: 33844307]
117. Collin-Histed T, Stoodley M, Beusterien K, Elstein D, Jaffe DH, Revel-Vilk S, et al. A global neuronopathic gaucher disease registry (GARDIAN): a patient-led initiative. Orphanet J Rare Dis 2023;18(1):195 [FREE Full text] [doi: 10.1186/s13023-023-02828-w] [Medline: 37480076]
118. Boychuk NA, Mulrooney NS, Kelly NR, Goldenberg AJ, Silver EJ, Wasserstein MP. Parental depression and anxiety associated with newborn bloodspot screening for rare and variable-onset disorders. Int J Neonatal Screen 2022;8(4):59 [FREE Full text] [doi: 10.3390/ijns8040059] [Medline: 36412585]
119. Morand O, Johnson J, Walter J, Atkinson L, Kline G, Frey A, et al. Symptoms and quality of life in patients with Fabry disease: results from an international patient survey. Adv Ther 2019;36(10):2866-2880 [FREE Full text] [doi: 10.1007/s12325-019-01061-x] [Medline: 31435831]

XSL•FO
RenderX

120. Hughes DA, Aguiar P, Deegan PB, Ezgu F, Frustaci A, Lidove O, et al. Early indicators of disease progression in Fabry disease that may indicate the need for disease-specific treatment initiation: findings from the opinion-based PREDICT-FD modified Delphi consensus initiative. BMJ Open 2020;10(10):e035182 [FREE Full text] [doi: 10.1136/bmjopen-2019-035182] [Medline: 33039984]
121. ISO/TR 11147:2023(en), health informatics — personalized digital health — digital therapeutics health software systems. International Organization for Standardization (ISO). 2023. URL: https://www.iso.org/obp/ui/#iso:std:iso:tr:11147:ed-1:v1:en [accessed 2026-01-01]
122. Staretz-Chacham O, Choi JH, Wakabayashi K, Lopez G, Sidransky E. Psychiatric and behavioral manifestations of lysosomal storage disorders. Am J Med Genet B Neuropsychiatr Genet 2010;153B(7):1253-1265. [doi: 10.1002/ajmg.b.31097] [Medline: 20872765]
123. Riemann D, Baglioni C, Bassetti C, Bjorvatn B, Dolenc Groselj L, Ellis JG, et al. European guideline for the diagnosis and treatment of insomnia. J Sleep Res 2017;26(6):675-700. [doi: 10.1111/jsr.12594] [Medline: 28875581]
124. Depression in adults: treatment and management. National Institute for Health and Care Excellence (NICE). 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK583074/ [accessed 2026-01-01]
125. DiGA-Verzeichnis – Digitale Gesundheitsanwendungen (DiGA) [Web page in German]. Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM). 2024. URL: https://diga.bfarm.de/de [accessed 2026-01-01]
126. Muenzer J, Jones SA, Tylki-Szymańska A, Harmatz P, Mendelsohn NJ, Guffon N, et al. Ten years of the hunter outcome survey (HOS): insights, achievements, and lessons learned from a global patient registry. Orphanet J Rare Dis 2017;12(1):82 [FREE Full text] [doi: 10.1186/s13023-017-0635-z] [Medline: 28464912]
127. Klein TL, Bender J, Bolton S, Collin-Histed T, Daher A, De Baere L, et al. A rare partnership: patient community and industry collaboration to shape the impact of real-world evidence on the rare disease ecosystem. Orphanet J Rare Dis 2024;19(1):262 [FREE Full text] [doi: 10.1186/s13023-024-03262-2] [Medline: 38987844]
128. Clarke LA, Giugliani R, Guffon N, Jones SA, Keenan HA, Munoz-Rojas MV, et al. Genotype-phenotype relationships in mucopolysaccharidosis type I (MPS I): insights from the international MPS I registry. Clin Genet 2019;96(4):281-289 [FREE Full text] [doi: 10.1111/cge.13583] [Medline: 31194252]
129. Choudhury MC, Chaube P. Integrating rare disease management in public health programs in India: exploring the potential of national health mission. Orphanet J Rare Dis 2022;17(1):43 [FREE Full text] [doi: 10.1186/s13023-022-02194-z] [Medline: 35144652]
130. Malavera A, Cadilhac DA, Thijs V, Lim JY, Grabsch B, Breen S, et al. Screening for fabry disease in young strokes in the Australian stroke clinical registry (AuSCR). Front Neurol 2020;11:596420 [FREE Full text] [doi: 10.3389/fneur.2020.596420] [Medline: 33324335]
131. Opladen T, Gleich F, Kozich V, Scarpa M, Martinelli D, Schaefer F, et al. U-IMD: the first unified European registry for inherited metabolic diseases. Orphanet J Rare Dis 2021;16(1):95 [FREE Full text] [doi: 10.1186/s13023-021-01726-3] [Medline: 33602304]
132. Wanner C, Ortiz A, Wilcox WR, Hopkin RJ, Johnson J, Ponce E, et al. Global reach of over 20 years of experience in the patient-centered Fabry Registry: advancement of Fabry disease expertise and dissemination of real-world evidence to the Fabry community. Mol Genet Metab 2023;139(3):107603 [FREE Full text] [doi: 10.1016/j.ymgme.2023.107603] [Medline: 37236007]
133. Schoenmakers DH, van den Berg S, Timmers L, Adang LA, Bäumer T, Bosch A, et al. Framework for multistakeholder patient registries in the field of rare diseases: focus on neurogenetic diseases. Neurology 2024;103(6):e209743. [doi: 10.1212/WNL.0000000000209743] [Medline: 39173102]
134. Gucciardi A, Legnini E, Di Gangi IM, Corbetta C, Tomanin R, Scarpa M, et al. A column-switching HPLC-MS/MS method for mucopolysaccharidosis type I analysis in a multiplex assay for the simultaneous newborn screening of six lysosomal storage disorders. Biomed Chromatogr 2014;28(8):1131-1139. [doi: 10.1002/bmc.3133] [Medline: 24449175]
135. La Cognata V, Cavallaro S. Detection of structural variants by NGS: revealing missing alleles in lysosomal storage diseases. Biomedicines 2022;10(8):1836 [FREE Full text] [doi: 10.3390/biomedicines10081836] [Medline: 36009380]
136. Ducatez F, Mauhin W, Boullier A, Pilon C, Pereira T, Aubert R, et al. Parsing fabry disease metabolic plasticity using metabolomics. J Pers Med 2021;11(9):898 [FREE Full text] [doi: 10.3390/jpm11090898] [Medline: 34575675]
137. Gojobori T, Ikeo K, Katayama Y, Kawabata T, Kinjo AR, Kinoshita K, et al. VaProS: a database-integration approach for protein/genome information retrieval. J Struct Funct Genomics 2016;17(4):69-81 [FREE Full text] [doi: 10.1007/s10969-016-9211-3] [Medline: 28012137]
138. Heinken A, El Kouche S, Guéant-Rodriguez RM, Guéant JL. Towards personalized genome-scale modeling of inborn errors of metabolism for systems medicine applications. Metabolism 2024;150:155738 [FREE Full text] [doi: 10.1016/j.metabol.2023.155738] [Medline: 37981189]
139. Pinto E Vairo F, Rojas Málaga D, Kubaski F, Fischinger Moura de Souza C, de Oliveira Poswar F, Baldo G, et al. Precision medicine for lysosomal disorders. Biomolecules 2020;10(8):1110 [FREE Full text] [doi: 10.3390/biom10081110] [Medline: 32722587]
140. Parenti G, Medina DL, Ballabio A. The rapidly evolving view of lysosomal storage diseases. EMBO Mol Med 2021;13(2):e12836 [FREE Full text] [doi: 10.15252/emmm.202012836] [Medline: 33459519]

XSL•FO
RenderX

141. Underhill HR, Golden-Grant K, Garrett LT, Uhrich S, Zielinski BA, Scott CR. Detecting the effects of Fabry disease in the adult human brain with diffusion tensor imaging and fast bound-pool fraction imaging. J Magn Reson Imaging 2015;42(6):1611-1622 [FREE Full text] [doi: 10.1002/jmri.24952] [Medline: 26018987]
142. Sodi A, Lenzetti C, Bacherini D, Finocchio L, Verdina T, Borg I, et al. Quantitative analysis of conjunctival and retinal vessels in Fabry disease. J Ophthalmol 2019;2019:4696429 [FREE Full text] [doi: 10.1155/2019/4696429] [Medline: 31093369]
143. Korlimarla A, Spiridigliozzi GA, Crisp K, Herbert M, Chen S, Malinzak M, et al. Novel approaches to quantify CNS involvement in children with Pompe disease. Neurology 2020;95(6):e718-e732 [FREE Full text] [doi: 10.1212/WNL.0000000000009979] [Medline: 32518148]
144. Kung S, Walters M, Claes P, Goldblatt J, Le Souef P, Baynam G. A dysmorphometric analysis to investigate facial phenotypic signatures as a foundation for non-invasive monitoring of lysosomal storage disorders. JIMD Rep 2013;8:31-39 [FREE Full text] [doi: 10.1007/8904_2012_152] [Medline: 23430517]
145. Mayhew D, Palmer K, Wilson I, Watson S, Stepien KM, Jenkins P, et al. Airway and anaesthetic management of adult patients with mucopolysaccharidoses undergoing cardiac surgery. J Clin Med 2024;13(5):1366 [FREE Full text] [doi: 10.3390/jcm13051366] [Medline: 38592237]
146. Cox-Brinkman J, Vedder A, Hollak C, Richfield L, Mehta A, Orteu K, et al. Three-dimensional face shape in Fabry disease. Eur J Hum Genet 2007;15(5):535-542. [doi: 10.1038/sj.ejhg.5201798] [Medline: 17327876]
147. Ulaşlı AM, Türkmen U, Toktaş H, Solak O. The complementary role of the kinect virtual reality game training in a patient with metachromatic leukodystrophy. PM R 2014;6(6):564-567. [doi: 10.1016/j.pmrj.2013.11.010] [Medline: 24412636]
148. Ronicke S, Hirsch MC, Türk E, Larionov K, Tientcheu D, Wagner AD. Can a decision support system accelerate rare disease diagnosis? Evaluating the potential impact of Ada DX in a retrospective study. Orphanet J Rare Dis 2019;14(1):69 [FREE Full text] [doi: 10.1186/s13023-019-1040-6] [Medline: 30898118]
149. Alves R, Piñol M, Vilaplana J, Teixidó I, Cruz J, Comas J, et al. Computer-assisted initial diagnosis of rare diseases. PeerJ 2016;4:e2211 [FREE Full text] [doi: 10.7717/peerj.2211] [Medline: 27547534]
150. Kobrinskii BA, Demikova NS, Blagosklonov NA. Knowledge engineering in construction of expert systems on hereditary diseases. In: Communications in Computer and Information Science. Cham: Springer International Publishing; 2018:35-45.
151. Piccininni M, Rohmann JL, Logroscino G, Kurth T. Blockchain-based innovations for population-based registries for rare neurodegenerative diseases. Front. Blockchain 2020;3. [doi: 10.3389/fbloc.2020.00020]
152. Vervoort D, Guetter CR, Peters AW. Blockchain, health disparities and global health. BMJ Innov 2021;7(2):506-514. [doi: 10.1136/bmjinnov-2021-000667]
153. Ghasemi M, Rahgozar M, Kavousi K. Complex disease genes identification using a heterogeneous network embedding approach. IEEE/ACM Trans Comput Biol Bioinform 2023;20(2):875-882. [doi: 10.1109/TCBB.2022.3175598] [Medline: 35594221]
154. Thirumal Kumar D, Eldous HG, Mahgoub ZA, George Priya Doss C, Zayed H. Computational modelling approaches as a potential platform to understand the molecular genetics association between Parkinson's and Gaucher diseases. Metab Brain Dis 2018;33(6):1835-1847. [doi: 10.1007/s11011-018-0286-3] [Medline: 29978341]
155. Warnock DG. 7th update on Fabry disease: biomarkers, progression and treatment opportunities in 2022. Nephron 2022;146 Suppl 1:1-51 [FREE Full text] [doi: 10.1159/000524933] [Medline: 35613555]
156. Abdallah S, Sharifa M, I Kh Almadhoun MK, Khawar MM, Shaikh U, Balabel KM, et al. The impact of artificial intelligence on optimizing diagnosis and treatment plans for rare genetic disorders. Cureus 2023;15(10):e46860 [FREE Full text] [doi: 10.7759/cureus.46860] [Medline: 37954711]
157. Al-Jasmi F, Moldovan L, Clarke JT. Hunter disease eClinic: interactive, computer-assisted, problem-based approach to independent learning about a rare genetic disease. BMC Med Educ 2010;10:72 [FREE Full text] [doi: 10.1186/1472-6920-10-72] [Medline: 20973983]
158. Di Brino E, Basile M, Rumi F, Fortunato A, Antonini D, Annicchiarico G, et al. Value for Patient - results of a value based healthcare workshop for the management of lysosomal storage disease in 3 Italian regions [Article in Italian]. GIHTAD 2022 Dec;15(5):1-7 [FREE Full text]
159. Big data and artificial intelligence for achieving universal health coverage: an international consultation on ethics. World Health Organization. URL: https://www.who.int/publications/i/item/WHO-HMM-IER-REK-2018-2 [accessed 2026-01-01]
160. Rao N, Kanago D, Morris M, Narayan V, Varshney K, Gn S, et al. Telegenetics: the experience of an Indian center (Centre for Human Genetics) during the COVID-19 pandemic. J Genet Couns 2021;30(5):1224-1232 [FREE Full text] [doi: 10.1002/jgc4.1517] [Medline: 34596296]
161. Almyranti M, Sutherland E, Ash N, Eiszele S. Artificial intelligence and the health workforce: perspectives from medical associations on AI in health. Organisation for Economic Co-Operation and Development (OECD). URL: https://www.oecd.org/en/publications/artificial-intelligence-and-the-health-workforce_9a31d8af-en.html [accessed 2024-11-20]
162. Tarricone R, Petracca F, Weller HM. "Towards harmonizing assessment and reimbursement of digital medical devices in the EU through mutual learning". NPJ Digit Med 2024;7(1):268 [FREE Full text] [doi: 10.1038/s41746-024-01263-w] [Medline: 39354125]

XSL•FO
RenderX

163. Spreafico A, Coder M, Boccardo V. DTx policy pathways: the evolving scenario in Europe. DTx Policy Pathways. 2023. URL: https://dtxpolicylandscape.org/ [accessed 2026-01-01]
164. Yoldaş Çelik M, Canda E, Yazıcı H, Erdem F, Kalkan Uçar S, Çöker M. Impact of the COVID-19 pandemic on inherited metabolic diseases: evaluation of enzyme replacement treatment adherence with telemedicine. J Pediatr Res Galenos Yayinevi 2022;9(4):391-396. [doi: 10.4274/jpr.galenos.2022.04206]
165. Zierhut HA, MacFarlane IM, Ahmed Z, Davies J. Genetic counselors' experiences and interest in telegenetics and remote counseling. J Genet Couns 2018;27(2):329-338. [doi: 10.1007/s10897-017-0200-x] [Medline: 29362948]
166. Hu H, Tjaden A, Knapp S, Antolin AA, Müller S. A machine learning and live-cell imaging tool kit uncovers small molecules induced phospholipidosis. Cell Chem Biol 2023;30(12):1634-1651.e6 [FREE Full text] [doi: 10.1016/j.chembiol.2023.09.003] [Medline: 37797617]
167. La Cognata V, Cavallaro S. A comprehensive, targeted NGS approach to assessing molecular diagnosis of lysosomal storage diseases. Genes (Basel) 2021;12(11):1750 [FREE Full text] [doi: 10.3390/genes12111750] [Medline: 34828358]
168. Fiumara A, Lanzafame G, Sapuppo A, Arena A, Cirnigliaro L, Barone R. Positive impact of home ERT for mucopolysaccharidoses and pompe disease: the lesson learnt from the COVID-19 pandemic. Healthcare (Basel) 2023;11(8):1176 [FREE Full text] [doi: 10.3390/healthcare11081176] [Medline: 37108010]
169. Finucane AM, O'Donnell H, Lugton J, Gibson-Watt T, Swenson C, Pagliari C. Digital health interventions in palliative care: a systematic meta-review. NPJ Digit Med 2021;4(1):64 [FREE Full text] [doi: 10.1038/s41746-021-00430-7] [Medline: 33824407]
170. Welter D, Juty N, Rocca-Serra P, Xu F, Henderson D, Gu W, et al. FAIR in action - a flexible framework to guide FAIRification. Sci Data 2023;10(1):291 [FREE Full text] [doi: 10.1038/s41597-023-02167-2] [Medline: 37208349]
171. Decherchi S, Pedrini E, Mordenti M, Cavalli A, Sangiorgi L. Opportunities and challenges for machine learning in rare diseases. Front Med (Lausanne) 2021;8:747612 [FREE Full text] [doi: 10.3389/fmed.2021.747612] [Medline: 34676229]
172. AI brings hope for patients with lysosomal storage disease. Today's Clinical Lab. 2023. URL: https://www.clinicallab.com/ai-brings-hope-for-patients-with-lysosomal-storage-disease-27351 [accessed 2026-01-01]
173. The act texts. EU Artificial Intelligence Act. URL: https://artificialintelligenceact.eu/the-act/ [accessed 2026-01-01]

Abbreviations

AI: artificial intelligence
CC: connected care
DHT: digital health technology
DL: deep learning
DMD: digital medical device
DMF: digital microfluidics
DSS: decision support system
DTx: digital therapeutics
EHR: electronic health record
ePRO: electronic patient-reported outcome
ERT: enzyme replacement therapy
EU: European Union
FD: Fabry disease
GD: Gaucher disease
GM1G: GM1 Gangliosidosis
GM2G: GM2 Gangliosidosis
HCP: health care provider
HIS: health information systems
HTA: health technology assessment
LLM: large language model
LSD: lysosomal storage disorder
mHealth: mobile health
ML: machine learning
MRI: magnetic resonance imaging
MS/MS: tandem mass spectrometry
NGS: next-generation sequencing
NLP: natural language processing
PCC: population, concept, context
PD: Pompe disease
PICO: Population, Intervention, Comparison, and Outcome

XSL•FO
RenderX

PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RCT: randomized controlled trial
RD: rare disease
VR: virtual reality

Edited by J Sarvestan; submitted 22.Mar.2025; peer-reviewed by B Wolfe-Piggott, E Ofoka; comments to author 18.Aug.2025; accepted 02.Feb.2026; published 02.Apr.2026.

Please cite as:
Spreafico AMC, Neri L, Angel K, Bellettato CM, Scarpa M
Artificial Intelligence, Connected Care, and Enabling Digital Health Technologies in Rare Diseases With a Focus on Lysosomal Storage Disorders: Scoping Review
J Med Internet Res 2026;28:e73612
URL: https://www.jmir.org/2026/1/e73612
doi:10.2196/73612
PMID:41661224

©Alberta MC Spreafico, Luca Neri, Kim Angel, Cinzia Maria Bellettato, Maurizio Scarpa. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Translating Telehealth Communication Research Into Patient-Centered, Implementable Practice

Rachel Pittmann[1*], PhD, CCC-SLP; Paula D Koppel[2*], NBC-HWC, AHN-BC, GNP-BC, RNC, PhD; David Barrett[3], RN, PhD

[1]Center for Interprofessional Education and Practice, MGH Institute of Health Professions, 36 1st Avenue, Charlestown, MA, United States

[2]School of Nursing, Duke University, Durham, NC, United States

[3]Department of Health Sciences, University of York, Heslington, York, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Rachel Pittmann, PhD, CCC-SLP
Center for Interprofessional Education and Practice, MGH Institute of Health Professions, 36 1st Avenue, Charlestown, MA, United States

Related Article:

https://www.jmir.org/2026/1/e79399

Abstract

Understanding both patient and clinician perspectives on communication challenges in virtual primary care consultations is important to ensure safe and effective care. This commentary reviews the important work of Alboksmaty and colleagues, highlighting their contributions and noting some limitations. To further support translation of clinical practice–based recommendations from this original research, this commentary relates the study findings to patient-centered communication frameworks and evidence-based virtual communication strategies. This commentary also extends the helpful mitigation strategies offered in the original research, demonstrating how they apply to the Expert Recommendations for Implementing Change (ERIC) implementation framework to facilitate organization-based change. The need for additional research differentiating phone-based and videoconferencing-based primary care consultation communication strategies is highlighted.

(J Med Internet Res 2026;28:e93690) doi:10.2196/93690

KEYWORDS

telemedicine; telehealth; patient-centered communication; videoconferencing; phone-based virtual care

Introduction

We congratulate Alboksmaty and colleagues [1] on their interesting study analyzing the communication challenges that occur during virtual consultations and thank them for generating mitigation strategies to address these challenges. Using 2 frameworks, the Shannon Weaver communication model (SWCM) and the capability, opportunity, motivation, and behavior (COM-B) model, to support their analyses, the researchers analyzed the data they collected during focus groups with UK-based general practitioners and patients to understand the communication barriers and gaps in virtual consultations. The SWCM and COM-B provided an interesting perspective when applied to telehealth communication. After reviewing this study closely, we offer some guidance for viewing and implementing the findings in practice.

Disaggregate Phone-Based and Videoconferencing-Based Virtual Consultation Results

In Alboksmaty and colleagues' study [1], phone-based and videoconferencing-based consultations were considered together under the umbrella heading of "virtual" consultations. Our view would be that treating virtual consultations as a homogeneous entity and combining these two modalities into one analysis limits the ability to fully interpret the communication challenges. For example, the first theme in the findings highlights the patients' and clinicians' challenges in interpreting emotional and clinical cues, resulting in feelings of disconnection related primarily to phone-based virtual care. Similarly, the second theme of compromised clarity when discussing complex or sensitive issues was also more relevant to phone-based virtual care, as was the third theme, relating to challenges in trust and confidence centered around the absence of nonverbal interactions. We propose that separately analyzing phone-based

XSL•FO
RenderX

and videoconferencing-based virtual care could provide a clearer understanding of the unique strengths and challenges of each modality. Such an approach might have led to different interpretations of the fifth and sixth themes—aligning virtual modalities with clinical needs and balancing contextual expectations with patient preferences. Disaggregating the data into phone-based and videoconferencing-based modalities could better support shared decision-making between clinicians and patients by helping to determine which option is most appropriate for a given clinical scenario. This approach aligns with recent calls for researchers to directly compare the use of phones and videoconferencing in primary care to inform a structured process for modality selection [2].

Use Patient-Centered Communication Models and Principles in Telehealth Research

We would have found it helpful to understand the rationale for the researchers' selection of the SWCM to support their analysis. The SWCM was originally a model developed in the mathematical theory of communications and, though it has been used in health care research and education, we do not believe it enables a holistic analysis of the complex nature of human interactions between clinicians and patients. Using the Roter interaction analysis system, often leveraged to measure verbal and nonverbal communication patterns in health care, including telehealth [3], or the patient-centered communication (PCC) framework, associated with quality in health care [4], may have supported more meaningful interpretation of the data.

We offer an expansion of Alboksmaty et al's [1] findings by aligning them with PCC principles whereby clinicians empathically engage with patients' needs and values and share in decision-making. PCC, as described by Moser et al [5], is organized around 7 core functions, which include ensuring adequate time and attention for patients to ask questions and address their emotional needs (functions 1, 2, and 6), sharing decision-making (function 3), enabling patients to understand how to take care of their health (function 4), providing clinical explanations that are understandable (function 5), and supporting patients in managing feelings of uncertainty (function 7). PCC has been shown to enhance patient trust and medication adherence as well as overall patient satisfaction [6].

In Table 1, we align Alboksmaty et al's [1] study findings with the PCC core functions and evidence-based, patient-centered virtual communication tools. As can be seen in the table, training and skill development for virtual consultations relate to 4 of the 6 study themes, reflecting the importance of building capacity (ie, knowledge and skill) for clinicians, patients, and organizations. Clinicians should consider that phone-based and videoconferencing-based consultations require distinct communication adaptations, and that specific needs and preferences of patients may also require accommodation. We encourage clinicians to consider their role in the development of the patient-clinician relationship and reference evidence-based telehealth PCC strategies [7-9] in the table to ensure the virtual environment supports patients' preferences and needs.

XSL•FO
RenderX

Table . Study findings in relation to patient-centered communication (PCC) challenges, mitigation strategies, and evidence-based virtual communication tools.

Communication challenges and associated mitigation strategies identified by Alboksmaty et al [1]	PCC functions at risk if mitigation strategies are not used[a]	Evidence-based, patient-centered virtual communication tools [7-9]
• Theme 1 challenges: missed cues and misinterpretations (sender-encoder) • Theme 1 strategies: training and skill development for virtual consultations (capability)	• Question (function 1) • Decisions (function 3) • Understood (function 4) • Explain (function 5)	• Create a safe, private, and comfortable environment for the visit in which the patient and clinician can adequately see and hear one another. • Maintain attentiveness and presence. • Personalize communication based on patient's information needs (literacy, preferred learning style, desired level of detail, sensory challenges). • Use summarizing statements and open-ended questions. • Avoid medical jargon. • Check for understanding (teach-back and follow-up questions). • Ensure adequate time for discussion and questions. • Demonstrate empathy using verbal and, whenever possible, nonverbal means.
• Theme 2 challenge: compromises related to lack of clarity in virtual communication (message) • Theme 2 strategy: addressing barriers to effective communication (opportunity)	• Question (function 1) • Decisions (function 3) • Understood (function 4) • Explain (function 5)	• Maintain attentiveness and presence. • Personalize communication based on patient's information needs (literacy, preferred learning style, desired level of detail, sensory challenges). • Use summarizing statements and open-ended questions. • Avoid medical jargon. • Check for understanding (teach-back and follow-up questions). • Ensure adequate time for discussion and questions. • Demonstrate and clearly state what is being asked of patients during assessment.
• Theme 3 challenges: technical issues, poor connectivity, and accessibility issues (channel) • Theme 3 strategies: training and skill development for virtual consultations (capability); triage and risk mitigation (opportunity); addressing barriers to effective communication (opportunity); training and resources to support patient preparedness (opportunity and motivation)	• Understood (function 4) • Time (function 6)	• Have a working knowledge of telehealth technologies and problem-solving skills. • Demonstrate a willingness to adapt to changing situations with ease and inform patients of a backup plan for technology failures. • Demonstrate respect for the patient challenges that are associated with virtual care. • Ensure patients know how to proceed when technology issues arise.
• Theme 4 challenge: lack of trust in discussing complex needs (receiver-decoder feedback) • Theme 4 strategies: patient and clinician preferences in consultation modalities (motivation); training and skill development for virtual consultations (capability)	• Question (function 1) • Attention (function 2) • Decision (function 3) • Understood (function 4) • Explain (function 5) • Time (function 6) • Feelings (function 7)	• Develop a joint agenda and an agreement on roles and expectations. • Maintain attentiveness and presence. • Take time to ensure a safe and comfortable environment (ensure privacy, create an interpersonal connection with the patient). • Demonstrate empathy using verbal and, whenever possible, nonverbal means.

XSL•FO
RenderX

Communication challenges and associated mitigation strategies identified by Alboksmaty et al [1]	PCC functions at risk if mitigation strategies are not used[a]	Evidence-based, patient-centered virtual communication tools [7-9]
• Theme 5 challenge: feeling less empathy and emotional engagement from clinician (receiver-decoder feedback) • Theme 5 strategies: training and skill development for virtual consultations (capability)	• Attention (function 2) • Understood (function 4) • Time (function 6) • Feelings (function 7)	• Maintain attentiveness and presence. • Intentionally create an interpersonal connection at the beginning of each visit. • Ensure the technology and the virtual space facilitate interpersonal connection (ie, eye contact, minimal distractions, adequate lighting). • Magnify use of active listening techniques (verbal and nonverbal). • Heighten self-awareness to ensure behaviors reflect empathy and care. • Demonstrate empathy using verbal and, whenever possible, nonverbal means.
• Theme 6 challenge: social norms, cultural factors, and age (context)	• Attention (function 2) • Decision (function 3) • Understood (function 4) • Explain (function 5) • Time (function 6) • Feelings (function 7)	• Demonstrate respect for the patient challenges associated with virtual care. • Demonstrate willingness to adapt to changing situations with ease and assist patients with technology. • Personalize communication approach as noted above to accommodate patient needs and/or preferences.
• Theme 7 challenges: lack of patient autonomy and inclusivity leading to decreased engagement and willingness to participate in virtual care	• Question (function 1) • Attention (function 2) • Decision (function 3) • Understood (function 4) • Explain (function 5) • Time (function 6) • Feelings (function 7)	• Demonstrate respect for the patient challenges associated with virtual care. • Demonstrate a willingness to adapt to changing situations with ease and assist patients with technology. • Personalize communication approach as noted above to accommodate patient needs and/or preferences.

[a]Full description of the PCC core functions [5]: (1) question—adequate time for patient to ask all health-related questions; (2) attention—adequate attention to patients' feelings and emotions; (3) decision—shared decision-making; (4) understood—ensuring patient understands how to manage their health; (5) explain—ensuring patient understands what clinician explains; (6) time— ensuring there is adequate time spent with the patient; (7) feelings—adequate time to manage patients' feelings of uncertainty about their health.

Use Implementation Frameworks to Facilitate Transfer of Research Findings to Clinical Practice

To further facilitate the translation of Alboksmaty et al's [1] findings and the associated evidence-based, patient-centered virtual communication tools into clinical practice, we propose using strategies from the Expert Recommendations for Implementing Change (ERIC) implementation framework [10]. Given the challenges, mitigation strategies, and evidence-based tools shown above, we propose using several ERIC strategies at both the individual clinician level and the organizational level for implementation of mitigation strategies. Specifically, creating a community of telehealth practitioners composed of champion clinicians, who will develop educational materials and outreach initiatives and who will model and simulate change prior to the implementation phase, supports the effective implementation of strategies into practice. Additionally, inviting patients as partners in the community of practice enables their voices to be central to the patient-centered change implementation.

Alboksmaty et al [1] analyzed telehealth communication challenges identified by clinicians and patients using unique frameworks. This commentary found it was feasible to apply their findings to the PCC and ERIC models, providing additional support to students and clinicians adapting their communication to telehealth. Future research focused on better differentiating communication strategies for phone-based and videoconferencing-based consultations is necessary to maximize the quality of virtual care.

Funding

The authors declare that no financial support was received for this work.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Alboksmaty A, Lunova T, Darzi A, Neves AL. Communication challenges and mitigation strategies in primary care virtual consultations: qualitative study. J Med Internet Res 2026 Jan 20;28(1):e79399. [doi: 10.2196/79399] [Medline: 41557949]
2. Byambasuren O, Greenwood H, Bakhit M, et al. Comparison of telephone and video telehealth consultations: systematic review. J Med Internet Res 2023 Nov 17;25(1):e49942. [doi: 10.2196/49942] [Medline: 37976100]
3. Agha Z, Roter DL, Schapira RM. An evaluation of patient-physician communication style during telemedicine consultations. J Med Internet Res 2009 Sep 30;11(3):e36. [doi: 10.2196/jmir.1193] [Medline: 19793720]
4. Institute of Medicine (US) Committee on Quality of Health Care in America. Crossing the Quality Chasm: A New Health System for the 21st Century: National Academies Press (US); 2001. URL: http://www.ncbi.nlm.nih.gov/books/NBK222274 [accessed 2026-03-10]
5. Moser RP, Trivedi N, Murray A, Jensen RE, Willis G, Blake KD. Patient-centered communication (PCC) scale: psychometric analysis and validation of a health survey measure. PLOS ONE 2022;17(12):e0279725. [doi: 10.1371/journal.pone.0279725] [Medline: 36584146]
6. Street RL, Makoul G, Arora NK, Epstein RM. How does communication heal? Pathways linking clinician-patient communication to health outcomes. Patient Educ Couns 2009 Mar;74(3):295-301. [doi: 10.1016/j.pec.2008.11.015] [Medline: 19150199]
7. Pittmann R, Danaher-Garcia N, Adair White BA, Thompson A. Development and validation of the Telehealth Etiquette Competency Checklist: a Delphi study. J Telemed Telecare 2025 Oct;31(9):1308-1316. [doi: 10.1177/1357633X241279494] [Medline: 39311041]
8. Koppel PD, De Gagne JC, Webb M, et al. Guidelines for rapport-building in telehealth videoconferencing: interprofessional e-Delphi study. JMIR Med Educ 2025 Aug 7;11:e76260. [doi: 10.2196/76260] [Medline: 40773683]
9. Gustin TS, Kott K, Rutledge C. Telehealth etiquette training: a guideline for preparing interprofessional teams for successful encounters. Nurse Educ 2020;45(2):88-92. [doi: 10.1097/NNE.0000000000000680] [Medline: 31022072]
10. Powell BJ, Waltz TJ, Chinman MJ, et al. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. Implement Sci 2015 Feb 12;10:21. [doi: 10.1186/s13012-015-0209-1] [Medline: 25889199]

Abbreviations

COM-B: capability, opportunity, motivation, and behavior
ERIC: Expert Recommendations for Implementing Change
PCC: patient-centered communication
SWCM: Shannon Weaver communication model

Edited by T Leung; submitted 17.Feb.2026; this is a non–peer-reviewed article;accepted 25.Feb.2026; published 19.Mar.2026.

Please cite as:
Pittmann R, Koppel PD, Barrett D
Translating Telehealth Communication Research Into Patient-Centered, Implementable Practice
J Med Internet Res 2026;28:e93690
URL: https://www.jmir.org/2026/1/e93690
doi:10.2196/93690

© Rachel Pittmann, Paula D Koppel, David Barrett. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Association Between Telemedicine Adoption and Physician Job Satisfaction: Cross-Sectional Study

Xiaohui Zhai[1], PhD; Zhongliang Zhou[2], PhD; Zhichao Wang[2], PhD; Guanping Liu[2], PhD; Peter C Coyte[3], PhD

[1]School of Public Health, Health Science Center, Xi'an Jiaotong University, Xi'an, Shaanxi, China

[2]School of Public Policy and Administration, Xi'an Jiaotong University, No. 28 Xianning West Road, Xi'an, Shaanxi, China

[3]Institute of Health Policy, Management & Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Zhongliang Zhou, PhD
School of Public Policy and Administration, Xi'an Jiaotong University, No. 28 Xianning West Road, Xi'an, Shaanxi, China

Abstract

Background: Telemedicine has expanded rapidly in recent years, with particularly pronounced growth following the COVID-19 pandemic. By improving access to care and offering greater flexibility in service delivery, it has become an important component of health care. Although the benefits of telemedicine for patients are well documented, its effects on physician job satisfaction remain insufficiently understood. Given the importance of job satisfaction for workforce stability, physician well-being, and quality of care, further examination of how telemedicine affects physician job satisfaction is warranted.

Objective: This study aims to examine the association between telemedicine adoption and physician job satisfaction and to assess whether the physician-patient relationship mediates this association.

Methods: A cross-sectional survey was conducted among health care professionals in Xi'an, China. Data were collected between November 7 and December 8, 2023, via an online questionnaire administered using the REDCap (Research Electronic Data Capture; Vanderbilt University) platform. A total of 12,052 physicians were included in the analysis. Physician job satisfaction was measured using a validated 6-point Likert scale. Telemedicine adoption was assessed through self-report. A partial proportional odds model was used to examine the association between telemedicine adoption and job satisfaction, adjusting for a comprehensive set of potential confounders. Additionally, the Karlson-Holm-Breen (KHB) decomposition method was used to explore the mediating role of physician-patient relationship quality in this association.

Results: Among 12,052 surveyed physicians, 1642 (13.62%) reported adopting telemedicine, whereas 10,410 (86.38%) did not. After adjusting for demographic characteristics, work-related factors, psychological factors, and physician-patient relationship, telemedicine adoption was significantly associated with higher job satisfaction (odds ratio [OR] 1.17, 95% CI 1.05 - 1.30). Findings were robust across multiple sensitivity analyses. Subgroup analyses indicated that the association did not vary across physician subgroups, and no significant interaction effects were observed. Mediation analysis revealed a total effect of telemedicine on job satisfaction of 0.33 (95% CI 0.17 - 0.50), with an indirect effect of 0.10 (95% CI 0.07 - 0.13) through improved physician-patient relationships, accounting for 30.30% of the total effect.

Conclusions: These findings suggest that telemedicine adoption is positively associated with physician job satisfaction, partially mediated by the physician-patient relationship. Policies should promote telemedicine adoption while prioritizing platform designs that support effective physician-patient interactions to enhance provider well-being and care outcomes.

(J Med Internet Res 2026;28:e82285) doi:10.2196/82285

KEYWORDS

telemedicine; job satisfaction; technology adoption; cross-sectional study; physician patient relationship

Introduction

Telemedicine refers to the provision of health care services over a distance using information and communication technologies, enabling remote diagnosis, consultation, treatment, and patient monitoring without the need for physical co-location [1]. Telemedicine has demonstrated significant potential to address longstanding challenges in health care systems. It effectively mitigates barriers such as limited access to care in rural and underserved populations, inefficiencies caused by overcrowded facilities and prolonged patient wait times, uneven distribution of health care resources, and escalating costs associated with traditional in-person care [2]. Globally, the use of telemedicine has increased significantly, especially following the COVID-19 pandemic [3]. According to a World Health Organization (WHO) survey, by 2022, over 70% of member states in the WHO European Region had implemented some form of telemedicine [4].

XSL•FO
RenderX

Telemedicine has become increasingly integrated into diverse clinical contexts, including primary care [5], chronic disease management [6], mental health services [7,8], and specialist consultations [9]. For physicians, this development has significantly altered traditional modes of clinical work. Remote consultations alter the pace and structure of daily workflows, reduce face-to-face interactions, and often require adaptation to new technologies [10]. Clinical decision-making is increasingly supported by digital tools [11], and communication with both patients and colleagues often takes place via electronic platforms.

These shifts have important implications for physicians' work experience [12], particularly in relation to job satisfaction. Job satisfaction is widely recognized as a critical component of physician well-being and a key determinant of health care system performance. Higher levels of physician satisfaction are consistently associated with improved clinical performance [13,14], lower turnover intentions [15,16], reduced burnout [17], and better patient care [18]. Nevertheless, empirical evidence on the association between telemedicine and physicians' job satisfaction remains mixed. In Australia, a large national survey of over 7000 physicians demonstrated that those using digital health technologies were approximately 14 percentage points more likely to report higher job satisfaction versus nonusers [19]. Similarly, a US study found that the use of specific telemedicine modalities, particularly video consultations and electronic health record (EHR) integrated platforms, was positively associated with higher physician job satisfaction [20]. However, evidence from a large ambulatory cohort indicated potential drawbacks, as higher telemedicine intensity was associated with increased after-hours EHR workload, which may negatively affect job satisfaction and work-life balance [21]. These divergent findings highlight the need for further investigation into the relationship between telemedicine adoption and physicians' overall job satisfaction.

Although prior studies have explored the effects of telemedicine on physician job satisfaction, limited attention has been paid to whether this relationship varies across different physician subgroups. Given the substantial variation in physicians' demographic and professional characteristics, it is plausible that telemedicine may not affect all physicians in the same way [22]. Examining these potential subgroup differences can improve our understanding of how telemedicine influences physician satisfaction and help inform more targeted and effective implementation strategies.

Moreover, few studies have examined the mechanisms underlying the relationship between telemedicine adoption and physician job satisfaction. One potential mediator is the physician-patient relationship. Telemedicine transforms how doctors and patients communicate by enabling clinical interactions to occur on virtual platforms. The adoption of telemedicine can modify key aspects of clinical communication, enabling more flexible information exchange, reduced time burden, and more efficient clarification and follow-up of clinical issues [23]. Such improvements can strengthen the quality of physician-patient interactions and foster a more positive relationship. In turn, this improved relationship is plausibly linked to physicians' job satisfaction, according to Relational Job Design Theory [24], which posits that more frequent, prolonged, and meaningful interactions with beneficiaries strengthen employees' perceptions of their positive impact and affective commitment, thereby increasing their motivation to make a prosocial difference. For physicians, stronger prosocial motivation is likely to enhance their engagement in patient care and reinforce their sense of competence and social worth, both of which are strongly linked to higher job satisfaction [25,26]. Taken together, these insights suggest that the physician-patient relationship may serve as a key pathway linking telemedicine adoption to physicians' job satisfaction.

Given the increasing integration of telemedicine into routine clinical practice, understanding its impact on physician well-being has become increasingly important. This study aims to evaluate the impact of telemedicine adoption on physicians' job satisfaction and to examine whether this association varies across different physician subgroups. Moreover, the study explores the underlying mechanisms of this relationship by assessing the mediating role of the physician-patient relationship. The findings are expected to provide practical guidance for improving telemedicine implementation and enhancing physician well-being.

Methods

Study Design and Participants

This study used data from a large-scale cross-sectional survey conducted among health care professionals in Xi'an, China. A stratified cluster sampling design based on probability proportional to size was used. In total, 47 hospitals and 59 primary health care facilities were selected. All employees within the selected institutions were invited to participate in the survey. Data were collected between November 7 and December 8, 2023, using an online survey administered through REDCap (Research Electronic Data Capture; version 15.2.1; Vanderbilt University). The questionnaire encompassed multiple domains, including demographic characteristics, health status, work-related experiences, attitudes toward internet-based health care, and perceptions of workplace changes.

For the purposes of this study, only physicians were included in the analysis, as they are the primary providers of telemedicine-based clinical consultations. Physicians were identified based on professional licensure status, including those registered as licensed physicians or assistant physicians in either Western or Traditional Chinese medicine. A total of 20,396 physicians responded to the survey. After excluding incomplete and ineligible responses, 12,052 valid cases remained in the final analytic sample. A detailed overview of the data cleaning and selection process is presented in Figure 1. The reporting of this cross-sectional study adheres to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist (Checklist 1).

XSL•FO
RenderX

Figure 1. Flowchart of the sample selection process.

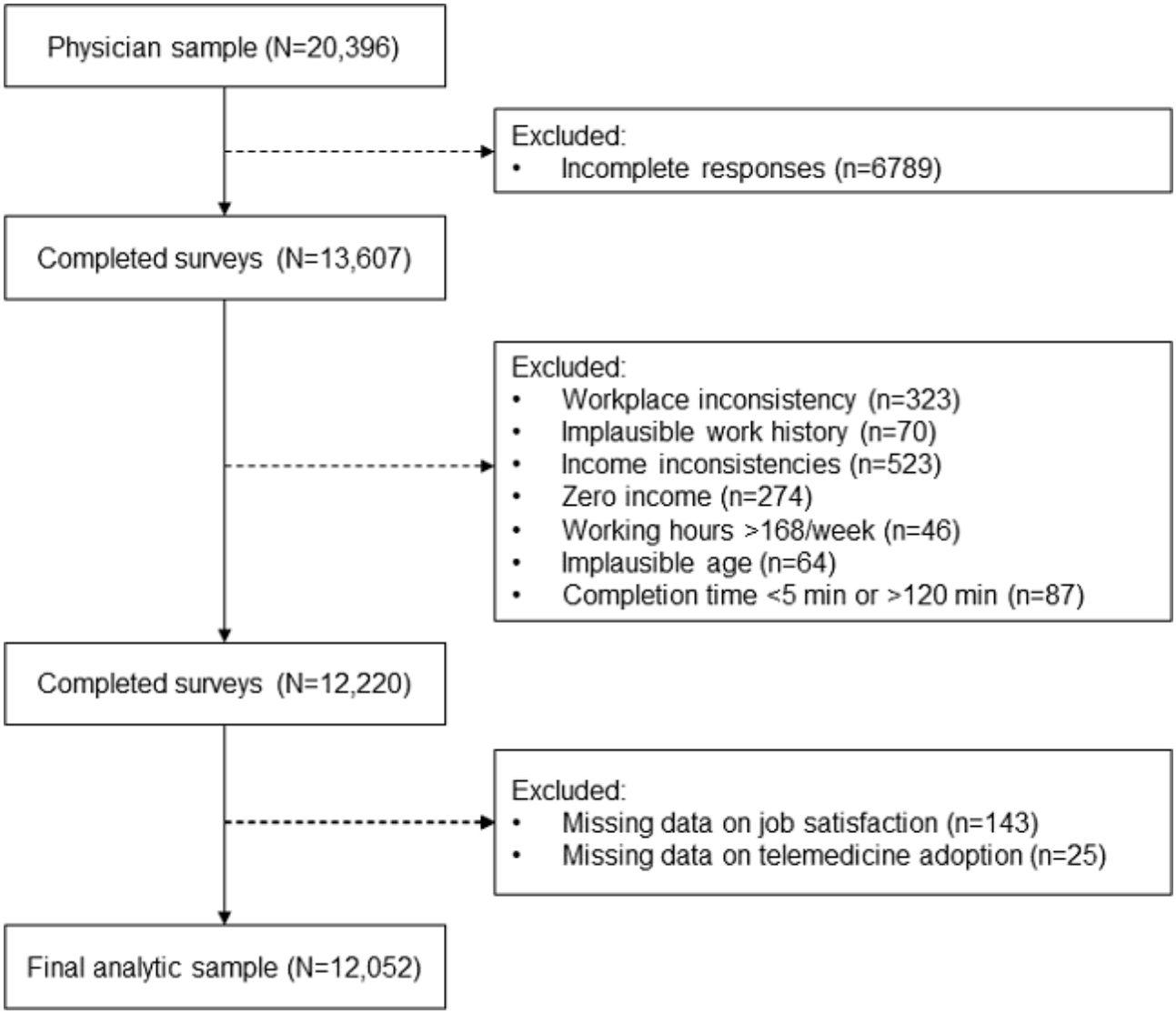

Measures

Job Satisfaction

Job satisfaction was measured using a single-item statement, "Overall, I am very satisfied with my current job." Respondents rated their agreement using a 6-point Likert-type scale with the following response options: (1) strongly disagree, (2) disagree, (3) somewhat disagree, (4) somewhat agree, (5) agree, and (6) strongly agree. The use of a single-item measure to capture overall job satisfaction has been validated in prior studies, demonstrating adequate levels of reliability and convergent validity [27,28]. The decision to adopt a 6-point scale, rather than a traditional 5-point or 7-point format, was based on evidence suggesting that even-point Likert scales can reduce central tendency bias by eliminating the neutral option and may provide greater discriminatory power across response categories [29,30].

Telemedicine Adoption

The primary independent variable in this study was physicians' adoption of telemedicine services, which reflects the integration of digital technologies into clinical practice. This variable was measured by asking physicians whether they had provided medical consultations or treatments via internet-based platforms. Responses were binary, with physicians answering "yes" if they had delivered services online and "no" if they had not.

Physician-Patient Relationship

To measure the physician-patient relationship from the physician's viewpoint, a 5-item physician-patient relationship scale was developed based on core relational dimensions identified in prior research and validated instruments such as the Patient-Doctor Relationship Questionnaire (PDRQ-9) [31]. The scale focused on 5 key relational components, including trust, respect, recognition, interpersonal satisfaction, and overall relational quality. Specifically, physicians were asked to rate their perceptions on the following aspects: (1) To what extent do you feel patients respect you? (2) To what extent do you feel patients trust the services you provide? (3) How would you rate the physician-patient relationship? (4) To what extent do you feel patients recognize the value of your work? (5) In most cases, how satisfied are patients with the care you provide? All items were rated on a 5-point Likert scale ranging from 1 ("very poor") to 5 ("very good"). The average score of the 5 items was calculated to represent the overall level of physician-patient

XSL•FO
RenderX

relationship, with higher scores indicating more positive relationships.

To assess the measurement validity of the 5-item physician-patient relationship scale, we conducted an exploratory factor analysis, which yielded a Kaiser-Meyer-Olkin (KMO) value of 0.82 and a significant Bartlett test of sphericity (χ^2_{10}=19,569.89; P<.001). All 5 items loaded strongly on the factor (loadings ranged from 0.59 to 0.74). The internal consistency of the scale was good, with a Cronbach α of 0.81.

Control Variables

We included three sets of control variables in the analysis: demographic, work-related characteristics, and psychological factors. Physician demographics comprised sex (male or female), age (categorized as ≤30, 31‐44, 45‐59, and ≥60 years), marital status (married or unmarried), and educational attainment (categorized as associate degree or below, bachelor's degree, master's degree, or doctoral degree).

Work-related characteristics included professional title, hospital type, employment status, medical specialty, managerial position, years in practice, years of service at the current institution, monthly income, and weekly working hours. In China, physicians hold professional titles that indicate their career stage. Professional titles were used as a proxy for career seniority and categorized into 4 levels according to the Chinese medical title system, including no title, junior, mid-level, and senior. As hierarchical levels are defined and routinely applied within the public hospital system in China, and such information was not collected in our data for private hospitals, hospitals were categorized into 4 types, including tertiary, secondary, primary, and private hospitals. Employment status was measured by asking whether physicians held a permanent (tenured) position at their institution. Responses were coded as 1 for permanent employee and 0 for contract employee. Medical specialty was classified according to the physician's department and included the following categories: clinical, medical technology, general practitioner, and others. The managerial position was coded as a binary variable indicating whether the physician held any administrative or leadership roles within their department, with 1 representing "yes" and 0 representing "no." Years in practice were measured by asking physicians how many years they have been practicing medicine since completing their training. Years of service at the current institution were assessed by asking physicians how long they had been employed at their current workplace.

Psychological factors were measured using indicators of work stress, depression, and anxiety. Work stress was assessed using "Overall, I feel a high level of work-related stress," with responses on a 6-point Likert scale ranging from strongly disagree to strongly agree. For analysis, responses were grouped into 3 categories: low stress (combining strongly disagree and disagree), moderate stress (combining somewhat disagree and somewhat agree), and high stress (combining agree and strongly agree). Depression and anxiety were assessed by self-reported clinical diagnosis, with participants indicating whether they have ever been diagnosed with depression or anxiety disorders.

Statistical Analysis

All statistical analyses were performed using STATA (version 17 MP; StataCorp LLC) and R (version 4.5.0; R Foundation for Statistical Computing). Statistical significance was set at a 2-sided P<.05.

Descriptive statistics were first calculated to summarize the characteristics of physicians. Given the skewed distribution of continuous variables, medians and IQRs were reported. Categorical variables were summarized using frequencies and percentages. To compare differences between physicians who provided telemedicine services and those who did not, the Mann-Whitney U test was applied for continuous variables, and the chi-square test was used for categorical variables.

Given the ordered nature of the job satisfaction outcome, we initially fitted an ordered logistic regression model to examine the association between telemedicine adoption and physicians' job satisfaction. The proportional odds assumption was evaluated using the Brant test [32]. The overall test indicated that the assumption was violated (χ^2_{116}=357.04; P<.001), although telemedicine adoption met the proportionality requirement (χ^2_4=4.16; P=.39). The detailed Brant test results are presented in Table S1 in Multimedia Appendix 1. To address these violations and avoid biased estimates, we used a partial proportional odds model, which relaxes the proportional odds constraint for variables that do not meet the assumption while maintaining the constraint for variables that do [33]. The final model included telemedicine adoption as the primary independent variable and adjusted for all control variables, including physicians' demographic characteristics, work-related factors, psychological factors, and the physician-patient relationship. Subgroup analyses were performed to examine potential variations in the relationship between telemedicine adoption and job satisfaction across all control variables. Continuous variables (eg, years in practice and weekly working hours) were dichotomized at their median values to ensure balanced group sizes for subgroup comparisons.

We used the Karlson-Holm-Breen (KHB) method to examine whether the physician-patient relationship mediated the association between telemedicine adoption and job satisfaction. The KHB approach allows for the decomposition of total effects into direct and indirect components in non-linear probability models [34]. This method effectively corrects for the rescaling problem that arises when comparing coefficients across nested nonlinear models, thereby providing unbiased estimates of mediation effects [35].

We conducted additional analyses to verify the robustness of our main finding. First, we re-estimated the main models using ordinary least squares regression as an alternative specification. Second, we used an alternative measure of job satisfaction based on a 5-point Likert item from the questionnaire, "Overall, I am very satisfied with my current job," with responses ranging from strongly disagree to strongly agree, to test the consistency of results across different satisfaction measures. Third, we used physicians' weekly online patient volume as a continuous measure of telemedicine adoption to assess the robustness of our findings.

XSL•FO
RenderX

Ethical Considerations

This study was approved by the Biomedical Ethics Committee of Xi'an Jiaotong University (XJTUAE2646). Electronic informed consent was obtained from all participants. All responses were anonymous, and no personal identifying information was collected. No compensation was provided to participants.

Results

Descriptive Statistics

The basic characteristics of physicians are displayed in Table 1. Among the 12,052 physicians, 1642 (13.62%) reported adopting telemedicine services, while 10,410 (86.38%) did not use telemedicine. The sample comprised 7351 out of 12,052 (61.08%) female physicians. Over half of the participants (6318/12,052; 52.42%) were aged 31-44 years. Most physicians were married. Regarding educational attainment, 5380 out of 12,052 (44.64%) held a bachelor's degree and 4512 out of 12,052 (37.44%) held a master's degree. Physicians with mid-level professional titles accounted for 39.17% (4712/12,052) of the sample, followed by those with a junior title (3850/12,052; 32.00%). More than half of the physicians (6765/12,052; 56.20%) were employed at tertiary hospitals. About 65.73% (7282/12,052) of the physicians held permanent positions. Clinicians accounted for 53.71% (6465/12,052) of the sample. The proportion of physicians holding managerial positions was 24.89% (3000/12,052). The median years of practice and years of service at the current institution were 12 and 8, respectively. The median monthly income was 8000 yuan (US $1159.34). The median number of weekly working hours was 50. Physicians who perceived work stress as "high" and "moderate" accounted for 48.14% (5800/12,052) and 42.98% (5179/12,052), respectively. The prevalence of depression and anxiety among participants was 4.32% (521/12,052) and 10.33% (1245/12,052), respectively.

XSL•FO
RenderX

Table . Characteristics of physicians stratified by telemedicine adoption.

Variables[a]	Total (N=12,052)	Telemedicine adoption		*P* value
		Yes (n=1642)	No (n=10,410)	
Sex, n (%)				<.001
Female	7351 (61.08)	797 (48.60)	6554 (63.05)	
Male	4684 (38.92)	843 (51.40)	3841 (36.95)	
Age (years), n (%)				<.001
≤30	2409 (19.99)	113 (6.88)	2296 (22.06)	
31-44	6318 (52.42)	858 (52.25)	5460 (52.45)	
45-59	3077 (25.53)	612 (37.27)	2465 (23.68)	
≥60	248 (2.06)	59 (3.59)	189 (1.82)	
Marital status, n (%)				<.001
Married	9862 (82.00)	1481 (90.30)	8381 (80.69)	
Unmarried[b]	2165 (18.00)	159 (9.70)	2006 (19.31)	
Education, n (%)				<.001
Associate degree or below	1449 (12.02)	137 (8.34)	1312 (12.60)	
Bachelor’s degree	5380 (44.64)	658 (40.07)	4722 (45.36)	
Master’s degree	4512 (37.44)	615 (37.45)	3897 (37.44)	
Doctoral degree	711 (5.90)	232 (14.13)	479 (4.60)	
Professional titles, n (%)				<.001
No title	266 (2.21)	13 (0.79)	253 (2.43)	
Junior	3850 (32.00)	220 (13.41)	3630 (34.93)	
Mid-level	4712 (39.17)	564 (34.39)	4148 (39.92)	
Senior	3203 (26.62)	843 (51.40)	2360 (22.71)	
Hospital type, n (%)				<.001
Tertiary hospitals	6765 (56.20)	1063 (64.78)	5702 (54.84)	
Secondary hospitals	2242 (18.62)	203 (12.37)	2039 (19.61)	
Primary hospitals	2233 (18.55)	163 (9.93)	2070 (19.91)	
Private hospitals	798 (6.63)	212 (12.92)	586 (5.64)	
Employment status, n (%)				<.001
Permanent employee	7282 (65.73)	1067 (75.46)	6215 (64.30)	
Contract employee	3797 (34.27)	347 (24.54)	3450 (35.70)	
Medical specialty, n (%)				<.001
Clinical	6465 (53.71)	1079 (65.75)	5386 (51.81)	
Medical technology	376 (3.12)	28 (1.71)	348 (3.35)	
General practitioner	1224 (10.17)	117 (7.13)	1107 (10.65)	
Others	3971 (32.99)	417 (25.41)	3554 (34.19)	
Managerial position, n (%)				<.001
No	9052 (75.11)	1033 (62.91)	8019 (77.03)	
Yes	3000 (24.89)	609 (37.09)	2391 (22.97)	
Years in practice, median (IQR)	12 (6 - 21)	17 (10 - 26)	11 (5 - 20)	<.001
Years of service at the current institution, median (IQR)	8 (3-15)	12 (6 - 20)	7 (3-14)	<.001

XSL•FO
RenderX

Variables[a]	Total (N=12,052)	Telemedicine adoption		*P* value
		Yes (n=1642)	No (n=10,410)	
Monthly income[c], median (IQR)	8000 (5000 - 10,000)	10,000 (7000 - 15,000)	7500 (5000 - 10,000)	<.001
Weekly working hours, median (IQR)	50 (40 - 60)	50 (45 - 60)	50 (40 - 60)	<.001
Work stress, n (%)				<.001
Low	1070 (8.88)	132 (8.04)	938 (9.01)	
Moderate	5179 (42.98)	633 (38.57)	4546 (43.68)	
High	5800 (48.14)	876 (53.38)	4924 (47.31)	
Depression, n (%)				.009
No	11,531 (95.68)	1551 (94.46)	9980 (95.87)	
Yes	521 (4.32)	91 (5.54)	430 (4.13)	
Anxiety, n (%)				<.001
No	10,807 (89.67)	1423 (86.66)	9384 (90.14)	
Yes	1245 (10.33)	219 (13.34)	1026 (9.86)	
Physician-patient relationship, median (IQR)	3.8 (3.4 - 4)	3.8 (3.4 - 4.2)	3.8 (3.4 - 4)	<.001

[a]Counts are based on valid (nonmissing) responses for each variable.

[b]Individuals who are single (never married), divorced, or widowed.

[c]Income is measured in Chinese Yuan (RMB), in 2023, the average exchange rate was approximately US $1=7.08 RMB.

Significant differences were observed between physicians who adopted telemedicine and those who did not across multiple characteristics. Compared to those who did not adopt telemedicine, physicians who adopted telemedicine were more likely to be male, older, married, hold a doctoral degree and senior professional titles, work in tertiary hospitals, be permanent employees, work in clinical departments, possess managerial roles, have longer years of practice and tenure at their current institution, and report higher monthly income (all *P*<.001). Additionally, telemedicine adopters reported higher work stress and were more likely to experience symptoms of depression and anxiety (all *P*<.01).

Association Between Telemedicine Adoption and Physician Job Satisfaction

We used a partial proportional odds model to estimate the association between telemedicine adoption and job satisfaction among physicians. The telemedicine adoption met the proportional odds assumption, indicating a consistent effect across cumulative logits. As shown in Table 2, telemedicine adoption was positively associated with higher job satisfaction. Specifically, physicians who adopted telemedicine had 17% higher odds of reporting a higher satisfaction level compared with those who did not (odds ratio [OR] 1.17, 95% CI 1.05 - 1.30).

XSL•FO
RenderX

Table . Association between telemedicine adoption and physician job satisfaction.

Variables	Physician job satisfaction, OR[a] (95% CI)				
	Very dissatisfied	Dissatisfied	Somewhat dissatisfied	Somewhat satisfied	Satisfied
Telemedicine adoption (reference=no)					
Yes	1.17[b] (1.05-1.30)	1.17[b] (1.05-1.30)	1.17[b] (1.05-1.30)	1.17[b] (1.05-1.30)	1.17[b] (1.05-1.30)
Sex (reference=female)					
Male	0.70[b] (0.55-0.90)	0.87 (0.74-1.02)	1.13[c] (1.02-1.25)	1.13[b] (1.03-1.23)	1.24[d] (1.11-1.39)
Age (years), (reference ≤30)					
31-44	0.97 (0.85-1.10)	0.97 (0.85-1.10)	0.97 (0.85-1.10)	0.97 (0.85-1.10)	0.97 (0.85-1.10)
45-59	1.13 (0.78-1.63)	1.08 (0.82-1.42)	0.99 (0.79-1.25)	1.10 (0.88-1.36)	1.26 (1.00-1.59)
≥60	1.45 (0.98-2.14)	1.45 (0.98-2.14)	1.45 (0.98-2.14)	1.45 (0.98-2.14)	1.45 (0.98-2.14)
Marital status (reference=unmarried)					
Married	1.78[d] (1.33-2.37)	1.40[b] (1.15-1.71)	1.25[b] (1.09-1.43)	1.08 (0.96-1.22)	1.05 (0.90-1.23)
Education (reference=associate degree or below)					
Bachelor's degree	0.80[b] (0.68-0.93)	0.80[b] (0.68-0.93)	0.80[b] (0.68-0.93)	0.80[b] (0.68-0.93)	0.80[b] (0.68-0.93)
Master's degree	0.73[b] (0.60-0.88)	0.73[b] (0.60-0.88)	0.73[b] (0.60-0.88)	0.73[b] (0.60-0.88)	0.73[b] (0.60-0.88)
Doctoral degree	0.71[b] (0.56-0.90)	0.71[b] (0.56-0.90)	0.71[b] (0.56-0.90)	0.71[b] (0.56-0.90)	0.71[b] (0.56-0.90)
Professional titles (reference=no title)					
Junior	0.80 (0.62-1.03)	0.80 (0.62-1.03)	0.80 (0.62-1.03)	0.80 (0.62-1.03)	0.80 (0.62-1.03)
Mid-level	0.58[d] (0.45-0.76)	0.58[d] (0.45-0.76)	0.58[d] (0.45-0.76)	0.58[d] (0.45-0.76)	0.58[d] (0.45-0.76)
Senior	0.60[d] (0.45-0.79)	0.60[d] (0.45-0.79)	0.60[d] (0.45-0.79)	0.60[d] (0.45-0.79)	0.60[d] (0.45-0.79)
Hospital type (reference =tertiary hospitals)					
Secondary hospitals	0.83[b] (0.75-0.93)	0.83[b] (0.75-0.93)	0.83[b] (0.75-0.93)	0.83[b] (0.75-0.93)	0.83[b] (0.75-0.93)
Primary hospitals	0.71[d] (0.62-0.81)	0.71[d] (0.62-0.81)	0.71[d] (0.62-0.81)	0.71[d] (0.62-0.81)	0.71[d] (0.62-0.81)
Private hospitals	4.67e-11[d] (4.36e-11 to 5.00e-11)	23.41 (0.01-590922.3)	0.48 (0.19-1.20)	1.69 (0.79-3.64)	2.30[c] (1.14-4.65)
Employment status (reference=contract employee)					
Permanent employee	0.86[d] (0.79-0.93)	0.86[d] (0.79-0.93)	0.86[d] (0.79-0.93)	0.86[d] (0.79-0.93)	0.86[d] (0.79-0.93)
Medical specialty (reference=clinical)					
Medical technology	1.25[c] (1.02-1.53)	1.25[c] (1.02-1.53)	1.25[c] (1.02-1.53)	1.25[c] (1.02-1.53)	1.25[c] (1.02-1.53)
General practitioner	1.02 (0.89-1.16)	1.02 (0.89-1.16)	1.02 (0.89-1.16)	1.02 (0.89-1.16)	1.02 (0.89-1.16)
Others	1.05 (0.97-1.13)	1.05 (0.97-1.13)	1.05 (0.97-1.13)	1.05 (0.97-1.13)	1.05 (0.97-1.13)
Managerial position (reference=no)					
Yes	1.44[d] (1.32-1.58)	1.44[d] (1.32-1.58)	1.44[d] (1.32-1.58)	1.44[d] (1.32-1.58)	1.44[d] (1.32-1.58)

XSL•FO
RenderX

Variables	Physician job satisfaction, OR[a] (95% CI)				
	Very dissatisfied	Dissatisfied	Somewhat dissatisfied	Somewhat satisfied	Satisfied
Years in practice	1.00 (0.99-1.01)	1.00 (0.99-1.01)	1.00 (0.99-1.01)	1.00 (0.99-1.01)	1.00 (0.99-1.01)
Years of service at the current institution	0.99[c] (0.99-1.00)	0.99[c] (0.99-1.00)	0.99[c] (0.99-1.00)	0.99[c] (0.99-1.00)	0.99[c] (0.99-1.00)
ln(Income)[e]	1.03 (0.99-1.08)	1.03 (0.99-1.08)	1.03 (0.99-1.08)	1.03 (0.99-1.08)	1.03 (0.99-1.08)
Weekly working hours	1.00[d] (0.99-1.00)	1.00[d] (0.99-1.00)	1.00[d] (0.99-1.00)	1.00[d] (0.991.00)	1.00[d] (0.99-1.00)
Work stress (reference=low)					
Moderate	4.76[d] (2.74-8.26)	2.44[d] (1.78-3.33)	0.73[b] (0.58-0.91)	0.24[d] (0.20-0.29)	0.21[d] (0.18-0.25)
High	1.04 (0.67-1.64)	1.05 (0.79-1.40)	0.57[d] (0.46-0.72)	0.38[d] (0.32-0.45)	0.33[d] (0.28-0.38)
Depression (reference=no)					
Yes	0.49[d] (0.34-0.71)	0.49[d] (0.37-0.65)	0.52[d] (0.41-0.65)	0.73[b] (0.58-0.92)	1.14 (0.80-1.63)
Anxiety (reference=no)					
Yes	0.42[d] (0.31-0.58)	0.62[d] (0.49-0.77)	0.79[b] (0.67-0.92)	0.66[d] (0.57-0.77)	0.72[b] (0.57-0.92)
Physician-patient relationship	3.05[d] (2.47-3.76)	3.03[d] (2.61-3.51)	2.79[d] (2.53-3.08)	3.41[d] (3.13-3.71)	4.54[d] (4.06-5.08)

[a]OR: odds ratio.

[b]$P<.01$.

[c]$P<.05$.

[d]$P<.001$.

[e]ln(Income) refers to the natural logarithm of monthly income.

In addition to telemedicine adoption, several covariates were significantly associated with job satisfaction, and many exhibited clear nonproportional effects across the satisfaction thresholds. For example, in the lowest satisfaction category, males are less likely to report higher satisfaction levels than females (OR 0.70, 95% CI 0.55 - 0.90), whereas at higher thresholds the direction of the association reversed, with males more likely to report higher satisfaction (eg, OR 1.24, 95% CI 1.11 - 1.39). Married physicians were more likely to report higher satisfaction than unmarried physicians at the lower thresholds, but the association diminished and was not statistically significant at the higher satisfaction levels. Physicians with moderate stress were more likely to report higher satisfaction at the lower thresholds but were less likely to report higher levels of satisfaction at the higher thresholds. High stress was associated with lower satisfaction at the higher thresholds. Physicians with depressive symptoms were less likely to report higher satisfaction at the lower and middle thresholds, with the association weakening at the highest satisfaction level. Physicians with anxiety were less likely to report higher satisfaction across all thresholds. Better physician-patient relationships were strongly and consistently associated with higher satisfaction across all thresholds, with stronger effects observed at the highest satisfaction level.

For covariates meeting the proportional-odds assumption, the estimated effects were consistent across all satisfaction thresholds. Physicians with higher education and those holding mid-level or senior professional titles exhibited significantly lower job satisfaction. Employment in secondary and primary hospitals was associated with reduced satisfaction relative to tertiary hospitals. Permanent employees reported lower levels of job satisfaction than contract employees (OR 0.86, 95% CI 0.79 - 0.93). Physicians in medical technology specialties reported higher satisfaction than those in clinical specialties (OR 1.25, 95% CI 1.02 - 1.53). Holding a managerial position showed a positive association with job satisfaction (OR 1.44, 95% CI 1.32 - 1.58). Longer tenure was associated with slightly higher odds of reporting lower satisfaction (OR 0.99, 95% CI 0.99 - 1.00). Longer weekly working hours were negatively associated with job satisfaction (OR 1.00, 95% CI 0.99 - 1.00).

Subgroup Analyses

Subgroup analyses were conducted to examine whether the association between telemedicine adoption and physician job satisfaction differed by physician characteristics. The association was generally consistent across groups, with no significant interaction effects detected (all P for interaction >.05), as illustrated in Figure 2. Notably, the association was statistically significant among female physicians (OR 1.20, 95% CI 1.03 - 1.40), those aged 31 - 44 years (OR 1.22, 95% CI 1.05 - 1.42), married physicians (OR 1.19, 95% CI 1.06 - 1.33), those with a master's degree (OR 1.21, 95% CI 1.02 - 1.44), permanent employees (OR 1.17, 95% CI 1.03 - 1.33), and physicians in the "others" specialty group (OR 1.29, 95% CI 1.05 - 1.59). In addition, significant associations were found among physicians with managerial positions (OR 1.25, 95% CI 1.03 - 1.50), those with fewer than 12 years of practice (OR

XSL•FO
RenderX

1.27, 95% CI 1.05 - 1.55), those with less than 8 years of service at the current institution (OR 1.25, 95% CI 1.04 - 1.51), those with monthly income below 8000 RMB (US $1159.34; OR 1.30, 95% CI 1.06 - 1.60), and those working fewer than 50 hours per week (OR 1.32, 95% CI 1.11 - 1.57). Finally, telemedicine adoption was positively associated with job satisfaction among physicians reporting moderate work stress (OR 1.27, 95% CI 1.07 - 1.51), those without depression (OR 1.15, 95% CI 1.02 - 1.28), and those without anxiety symptoms (OR 1.14, 95% CI 1.02 - 1.29).

Figure 2. Forest plot of subgroup analyses on the association between telemedicine adoption and physician job satisfaction. Income is measured in Chinese Yuan (RMB). In 2023, the average exchange rate was approximately US $1=7.08 RMB. OR: odds ratio.

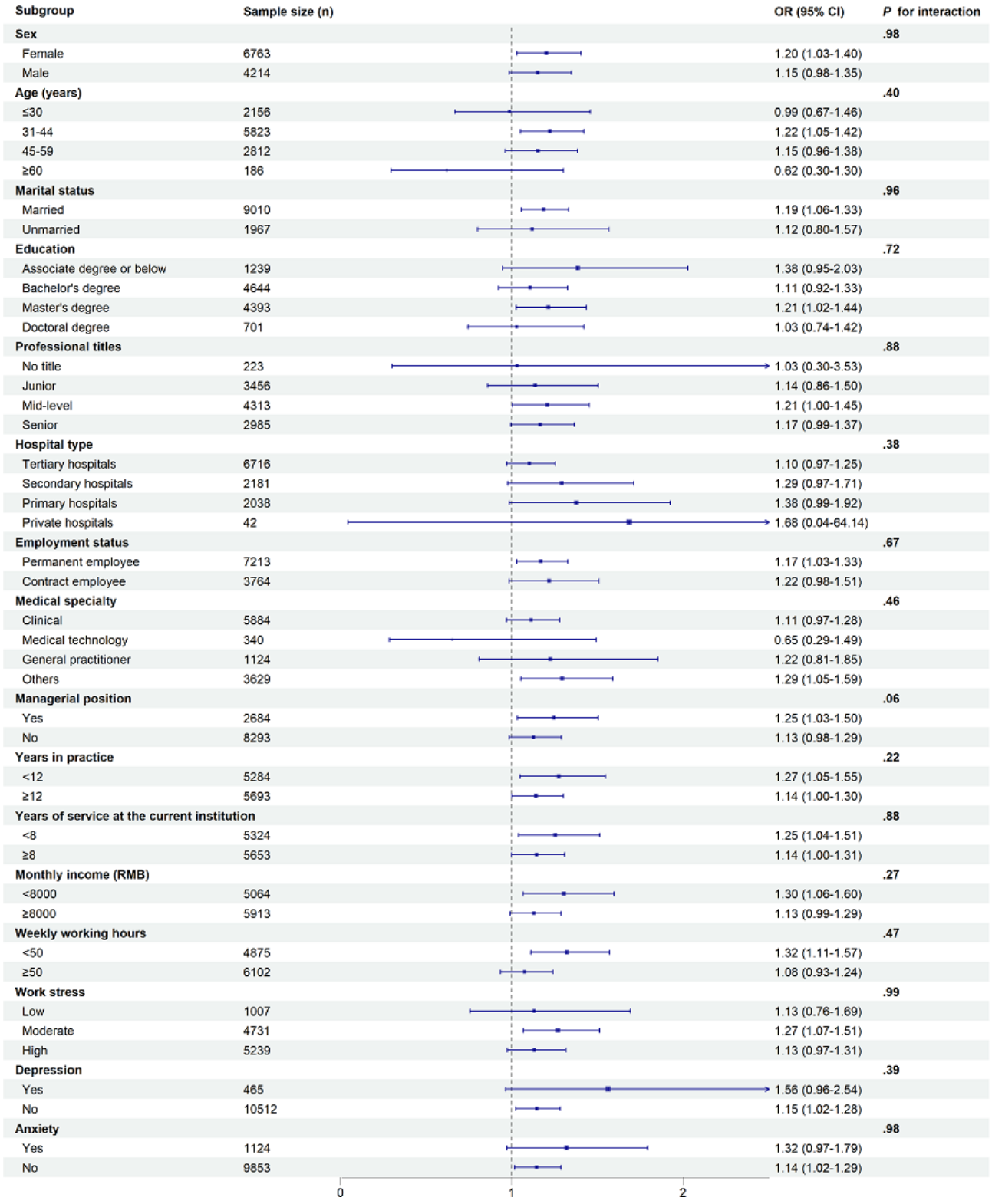

XSL•FO
RenderX

Mediation Analyses

Table 3 presents the results of the KHB decomposition analysis. The results indicated a significant partial mediation effect. The total effect of telemedicine adoption on job satisfaction was 0.33 (95% CI 0.17-0.50). After accounting for the physician-patient relationship, the direct effect decreased to 0.23 (95% CI 0.07-0.39), while the indirect effect through the physician-patient relationship was 0.10 (95% CI 0.07-0.13). This indirect pathway accounted for 30.30% of the total effect, suggesting that the physician-patient relationship plays an important mediating role between telemedicine adoption and job satisfaction.

Table . Mediation effect of physician-patient relationship on telemedicine and job satisfaction.

Effect	β (95% CI)	*P* value	Proportion of mediating effect
Total effect	.33 (0.17-0.50)	<.001	—[a]
Direct effect	.23 (0.07-0.39)	.006	69.70%
Indirect effect	.10 (0.07-0.13)	<.001	30.30%

[a]Not available.

Sensitivity Analysis

The sensitivity analyses supported the robustness of our main findings. First, we re-estimated the models using ordinary least squares regression as a robustness check, and the results remained consistent with the original models in terms of direction and statistical significance (Table S2 in Multimedia Appendix 1). Second, when job satisfaction was measured using an alternative 5-point Likert item, the association between telemedicine adoption and job satisfaction remained robust (Table S3 in Multimedia Appendix 1). Third, we replaced the binary indicator of telemedicine adoption with a continuous measure reflecting the number of patients a physician consulted online per week. The results remained robust, further supporting the validity of our main findings (Table S4 in Multimedia Appendix 1).

Discussion

Principal Findings

In this cross-sectional study of physicians, telemedicine adoption was significantly associated with higher levels of job satisfaction. Physicians who adopted telemedicine reported higher levels of job satisfaction compared with those who did not, and this association remained robust across multiple sensitivity analyses. Mediation analysis further suggested that the physician-patient relationship partially explained this association. These findings contribute to a growing body of literature examining how digital health technologies affect physicians' work experiences and underscore the importance of preserving strong interpersonal relationships in virtual care settings.

Comparison With Prior Work

The telemedicine adoption rate in our study was 13.62%, which appears relatively low compared to telemedicine adoption rates reported in other countries, where physician uptake during the COVID-19 pandemic ranged from approximately 30% to 87% [36-38]. This difference may be partly attributable to variations in the measurement of telemedicine adoption. In our study, adoption was defined specifically as whether physicians had provided medical consultations or treatments via internet-based platforms, whereas some international studies use broader definitions of telemedicine that include not only direct patient consultations but also activities such as exchanging referrals with other providers, ordering and reviewing diagnostic tests, or other forms of remote care, thereby yielding higher reported rates. Additionally, regional differences and health care system characteristics in Xi'an and broader China may play a role.

Consistent with previous research, our study found that telemedicine adoption is positively associated with physician job satisfaction [19,20]. This association may reflect several potential advantages of telehealth services that directly influence physicians' work experiences. Previous studies have found that telemedicine enhances scheduling flexibility and work-life balance among physicians [19], reduces administrative burdens [39], and increases practice productivity [40], all of which are closely linked to lower stress and higher job satisfaction. Additionally, telemedicine may support more efficient patient follow-up and continuity of care, particularly in primary care [41] and chronic disease management settings [42,43], thereby improving the clinical workflow. Taken together, these features may contribute to a more sustainable and fulfilling clinical practice.

Subgroup analyses demonstrated that the positive association between telemedicine adoption and physician job satisfaction was generally consistent across various demographic and professional groups, with no significant interaction effects observed. These findings imply that telemedicine adoption is broadly beneficial for physician job satisfaction, regardless of demographic or work-related characteristics. Nonetheless, the significant associations within certain subgroups indicate that the effect may be more detectable among physicians sharing specific characteristics.

The mediation analysis revealed a significant partial mediating effect, indicating that telemedicine adoption is associated with a more positive physician-patient relationship, which in turn contributes to higher job satisfaction. Specifically, physicians who used telemedicine tended to report higher physician-patient relationship scores. Empirical studies similarly show that telemedicine enables more frequent, timely, and continuous contact, which strengthens physicians' sense of connection and familiarity with their patients, thereby fostering a more trusting and collaborative physician-patient relationship [44,45]. Such positive relational experiences have been associated with higher

XSL•FO
RenderX

job satisfaction [46], partly because strong physician-patient relationships enhance physicians' sense of professional efficacy, reduce workplace stress, and reinforce the meaningfulness of their clinical work. Overall, these findings suggest that improvements in physician-patient relationships represent an important pathway through which telemedicine may enhance physicians' job satisfaction.

Building on these findings, several policy implications can be drawn. Health care systems should facilitate telemedicine adoption through targeted training and resource allocation tailored to physicians' diverse needs. The design of telemedicine platforms should explicitly incorporate features aimed at improving the quality of the physician-patient relationship. Additionally, workload management strategies and support mechanisms are necessary to mitigate potential burnout associated with telemedicine use.

Limitations

Several limitations of this study warrant careful consideration. First, the cross-sectional design inherently restricts the ability to draw causal inferences regarding the relationship between telemedicine use and physicians' job satisfaction. Longitudinal studies are needed to establish temporal dynamics and causal pathways. Second, job satisfaction was assessed using a single-item measure. While this approach has been validated in previous research and is suitable for large-scale surveys with space constraints, multi-item instruments could offer a more detailed understanding of satisfaction dimensions. Third, the survey measured telemedicine use based on whether physicians reported using telemedicine for diagnosis or treatment, without distinguishing specific modalities such as video consultations, asynchronous messaging, or phone visits. Future studies with more detailed information on specific telemedicine modalities could clarify how different forms of telemedicine influence physicians' job satisfaction. Finally, although a stratified probability proportional to size sampling approach was used to enhance representativeness, the sample is limited to physicians within Xi'an, a specific urban context in China. This geographic concentration may limit the external validity of findings, as organizational cultures, telemedicine infrastructure, and health care policies vary substantially across regions and countries.

Conclusions

In this cross-sectional study, telemedicine adoption was positively associated with physician job satisfaction. This association was partially mediated by the physician-patient relationship, suggesting that enhanced communication and connection through telemedicine contribute to higher satisfaction. To optimize these benefits, physicians need to use telemedicine appropriately to balance its benefits and challenges. Meanwhile, health care institutions should provide targeted support, including communication training and stress management, to promote effective and sustainable telemedicine practice.

Acknowledgments

The authors would like to thank all participating physicians for their time and valuable contributions to this study. We are also grateful to the administrative staff and coordinators at the participating hospitals and primary health care facilities in Xi'an for their support during data collection.

Funding

This study was funded by leading talents project in philosophy and social sciences of the National Social Science Foundation of China (2022LJRC02).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: XZ
Formal analysis: XZ
Funding acquisition: ZZ
Methodology: XZ
Supervision: ZZ, PCC
Visualization: XZ
Writing – original draft: XZ
Writing – review & editing: XZ, ZZ, ZW, GL, PCC

Conflicts of Interest

None declared.

Multimedia Appendix 1
Results of Brant test and robustness checks.
[DOCX File, 43 KB - jmir_v28i1e82285_app1.docx]

XSL•FO
RenderX

Checklist 1
The STROBE checklist.
[DOCX File, 33 KB - jmir_v28i1e82285_app2.docx]

References

1. World Health Organization. Telemedicine: Opportunities and Developments in Member States: Report on the Second Global Survey on eHealth: World Health Organization; 2010:93 URL: https://iris.who.int/handle/10665/44497 [accessed 2026-03-19]
2. Bashshur RL, Shannon GW, Smith BR, et al. The empirical foundations of telemedicine interventions for chronic disease management. Telemed J E Health 2014 Sep;20(9):769-800. [doi: 10.1089/tmj.2014.9981] [Medline: 24968105]
3. Suran M. Increased use of medicare telehealth during the pandemic. JAMA 2022 Jan 25;327(4):313. [doi: 10.1001/jama.2021.23332]
4. Gullslett MK, Ronchi E, Lundberg L, et al. Telehealth development in the WHO European region: results from a quantitative survey and insights from Norway. Int J Med Inform 2024 Nov;191:105558. [doi: 10.1016/j.ijmedinf.2024.105558] [Medline: 39084085]
5. Orozco-Beltran D, Sánchez-Molla M, Sanchez JJ, Mira JJ, ValCrònic Research Group. Telemedicine in primary care for patients with chronic conditions: the ValCrònic quasi-experimental study. J Med Internet Res 2017 Dec 15;19(12):e400. [doi: 10.2196/jmir.7677] [Medline: 29246881]
6. Lewinski AA, Walsh C, Rushton S, et al. Telehealth for the longitudinal management of chronic conditions: systematic review. J Med Internet Res 2022 Aug 26;24(8):e37100. [doi: 10.2196/37100] [Medline: 36018711]
7. Fortney JC, Bauer AM, Cerimele JM, et al. Comparison of teleintegrated care and telereferral care for treating complex psychiatric disorders in primary care: a pragmatic randomized comparative effectiveness trial. JAMA Psychiatry 2021 Nov 1;78(11):1189-1199. [doi: 10.1001/jamapsychiatry.2021.2318] [Medline: 34431972]
8. Wu M, Li C, Hu T, et al. Effectiveness of telecare interventions on depression symptoms among older adults: systematic review and meta-analysis. JMIR Mhealth Uhealth 2024 Jan 17;12:e50787. [doi: 10.2196/50787] [Medline: 38231546]
9. Di Cerbo A, Morales-Medina JC, Palmieri B, Iannitti T. Narrative review of telemedicine consultation in medical practice. Patient Prefer Adherence 2015;9:65-75. [doi: 10.2147/PPA.S61617] [Medline: 25609928]
10. Greenhalgh T, Rosen R, Shaw SE, et al. Planning and evaluating remote consultation services: a new conceptual framework incorporating complexity and practical ethics. Front Digit Health 2021;3:726095. [doi: 10.3389/fdgth.2021.726095] [Medline: 34713199]
11. Shah KM, Shah MM, Shah IP, Patel HH. The role of technology in enhancing clinical decision-making for health professiona. Front Health Inform 2024 Dec 26:920-929. [doi: 10.53555/fhi.v13i7.1577]
12. Kaihlanen AM, Laukka E, Nadav J, et al. The effects of digitalisation on health and social care work: a qualitative descriptive study of the perceptions of professionals and managers. BMC Health Serv Res 2023 Jun 30;23(1):714. [doi: 10.1186/s12913-023-09730-y] [Medline: 37386423]
13. Scheepers RA. Physicians' professional performance: an occupational health psychology perspective. Perspect Med Educ 2017 Dec;6(6):425-428. [doi: 10.1007/s40037-017-0382-9] [Medline: 29064072]
14. Linn LS, Brook RH, Clark VA, Davies AR, Fink A, Kosecoff J. Physician and patient satisfaction as factors related to the organization of internal medicine group practices. Med Care 1985 Oct;23(10):1171-1178. [doi: 10.1097/00005650-198510000-00006] [Medline: 4058071]
15. Buchbinder SB, Wilson M, Melick CF, Powe NR. Primary care physician job satisfaction and turnover. Am J Manag Care 2001 Jul;7(7):701-713. [Medline: 11464428]
16. Hall CB, Brazil K, Wakefield D, Lerer T, Tennen H. Organizational culture, job satisfaction, and clinician turnover in primary care. J Prim Care Community Health 2010 Apr 1;1(1):29-36. [doi: 10.1177/2150131909360990] [Medline: 23804066]
17. Hodkinson A, Zhou A, Johnson J, et al. Associations of physician burnout with career engagement and quality of patient care: systematic review and meta-analysis. BMJ 2022 Sep 14;378:e070442. [doi: 10.1136/bmj-2022-070442] [Medline: 36104064]
18. Williams ES, Skinner AC. Outcomes of physician job satisfaction: a narrative review, implications, and directions for future research. Health Care Manage Rev 2003;28(2):119-139. [doi: 10.1097/00004010-200304000-00004] [Medline: 12744449]
19. Zaresani A, Scott A. Does digital health technology improve physicians' job satisfaction and work-life balance? A cross-sectional national survey and regression analysis using an instrumental variable. BMJ Open 2020 Dec 12;10(12):e041690. [doi: 10.1136/bmjopen-2020-041690] [Medline: 33310807]
20. Sengupta A, Sarkar S, Bhattacherjee A. The relationship between telemedicine tools and physician satisfaction, quality of care, and patient visits during the COVID-19 pandemic. Int J Med Inform 2024 Oct;190:105541. [doi: 10.1016/j.ijmedinf.2024.105541] [Medline: 38996654]
21. Lawrence K, Nov O, Mann D, Mandal S, Iturrate E, Wiesenfeld B. The impact of telemedicine on physicians' after-hours electronic health record "Work Outside Work" during the COVID-19 pandemic: retrospective cohort study. JMIR Med Inform 2022 Jul 28;10(7):e34826. [doi: 10.2196/34826] [Medline: 35749661]

XSL•FO
RenderX

22. Leigh JP, Kravitz RL, Schembri M, Samuels SJ, Mobley S. Physician career satisfaction across specialties. Arch Intern Med 2002 Jul 22;162(14):1577-1584. [doi: 10.1001/archinte.162.14.1577] [Medline: 12123400]
23. Haleem A, Javaid M, Singh RP, Suman R. Telemedicine for healthcare: capabilities, features, barriers, and applications. Sens Int 2021;2:100117. [doi: 10.1016/j.sintl.2021.100117] [Medline: 34806053]
24. Grant AM. Relational job design and the motivation to make a prosocial difference. AMR 2007 Apr;32(2):393-417. [doi: 10.5465/amr.2007.24351328]
25. Gkliati A, Saiti A. Work engagement and job satisfaction in the medical sector: two aspects of psychological well-being among medical staff. Int J Health Sci 2022 Jul 25;6(S6):7753-7761. [doi: 10.53730/ijhs.v6nS6.11070]
26. Bailey C, Yeoman R, Madden A, Thompson M, Kerridge G. A review of the empirical literature on meaningful work: progress and research agenda. Hum Resour Dev Rev 2019 Mar;18(1):83-113. [doi: 10.1177/1534484318804653]
27. Wanous JP, Reichers AE, Hudy MJ. Overall job satisfaction: how good are single-item measures? J Appl Psychol 1997 Apr;82(2):247-252. [doi: 10.1037/0021-9010.82.2.247] [Medline: 9109282]
28. Dolbier CL, Webster JA, McCalister KT, Mallon MW, Steinhardt MA. Reliability and validity of a single-item measure of job satisfaction. Am J Health Promot 2005;19(3):194-198. [doi: 10.4278/0890-1171-19.3.194] [Medline: 15693347]
29. Quality of psychology test between Likert scale 5 and 6 points. J Soc Sci 2010 Mar 1;6(3):399-403. [doi: 10.3844/jssp.2010.399.403]
30. Leung SO. A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-Point Likert Scales. J Soc Serv Res 2011 Jul;37(4):412-421. [doi: 10.1080/01488376.2011.580697]
31. Van der Feltz-Cornelis CM, Van Oppen P, Van Marwijk HWJ, De Beurs E, Van Dyck R. A patient-doctor relationship questionnaire (PDRQ-9) in primary care: development and psychometric evaluation. Gen Hosp Psychiatry 2004;26(2):115-120. [doi: 10.1016/j.genhosppsych.2003.08.010] [Medline: 15038928]
32. Long JS, Freese J. Regression Models for Categorical Dependent Variables Using Stata, 3rd edition: College Station, Texas: Stata Press; 2014.
33. Williams R. Generalized ordered logit/partial proportional odds models for ordinal dependent variables. Stata J 2006 Feb;6(1):58-82. [doi: 10.1177/1536867X0600600104]
34. Breen R, Karlson KB, Holm A. Total, direct, and indirect effects in logit and probit models. Sociol Methods Res 2013 May;42(2):164-191. [doi: 10.1177/0049124113494572]
35. Karlson KB, Holm A, Breen R. Comparing regression coefficients between same-sample nested models using logit and probit. Sociol Methodol 2012 Aug;42(1):286-313. [doi: 10.1177/0081175012444861]
36. Majority of UK physicians see continued rise in telemedicine usage, reveals globaldata. GlobalData. 2024. URL: https://www.globaldata.com/media/pharma/majority-uk-physicians-see-continued-rise-telemedicine-usage-reveals-globaldata/ [accessed 2025-08-11]
37. Scott A, Bai T, Zhang Y. Association between telehealth use and general practitioner characteristics during COVID-19: findings from a nationally representative survey of Australian doctors. BMJ Open 2021 Mar 24;11(3):e046857. [doi: 10.1136/bmjopen-2020-046857] [Medline: 33762248]
38. Pylypchuk Y, Barker W. Use of telemedicine among office-based physicians. In: ASTP Health IT Data Brief: Office of the Assistant Secretary for Technology Policy; 2021. [Medline: 39475515]
39. Hopkins L, Pedwell G. The COVID PIVOT - re-orienting child and youth mental health care in the light of pandemic restrictions. Psychiatr Q 2021 Sep;92(3):1259-1270. [doi: 10.1007/s11126-021-09909-8] [Medline: 33754240]
40. Stefos T, Carey K, Shen ML, Poe S, Oh DH, Moran E. The effect of telehealth services on provider productivity. Med Care 2021 May 1;59(5):456-460. [doi: 10.1097/MLR.0000000000001529] [Medline: 33821831]
41. Carrillo de Albornoz S, Sia KL, Harris A. The effectiveness of teleconsultations in primary care: systematic review. Fam Pract 2022 Jan 19;39(1):168-182. [doi: 10.1093/fampra/cmab077] [Medline: 34278421]
42. Mabeza RMS, Maynard K, Tarn DM. Influence of synchronous primary care telemedicine versus in-person visits on diabetes, hypertension, and hyperlipidemia outcomes: a systematic review. BMC Prim Care 2022 Mar 21;23(1):52. [doi: 10.1186/s12875-022-01662-6] [Medline: 35313804]
43. Guarino M, Cossiga V, Fiorentino A, Pontillo G, Morisco F. Use of telemedicine for chronic liver disease at a single care center during the COVID-19 pandemic: prospective observational study. J Med Internet Res 2020 Sep 21;22(9):e20874. [doi: 10.2196/20874] [Medline: 32896833]
44. Sandberg J, Trief PM, Izquierdo R, et al. A qualitative study of the experiences and satisfaction of direct telemedicine providers in diabetes case management. Telemed J E Health 2009 Oct;15(8):742-750. [doi: 10.1089/tmj.2009.0027] [Medline: 19780691]
45. Maria ARJ, Serra H, Castro MG, Heleno B. Interaction at the primary-secondary care interface: patients' and physicians' perceptions of teleconsultations. Digit Health 2022;8:20552076221133698. [doi: 10.1177/20552076221133698] [Medline: 36465985]
46. Li T, Guan L, Zhang R, et al. Roles of doctor-patient relationship perception and job satisfaction in the impact of workplace violence on medical professionals' turnover intentions in the early phase of COVID-19: a cross-sectional study in China. BMJ Open 2023 Sep 4;13(9):e074744. [doi: 10.1136/bmjopen-2023-074744] [Medline: 37666559]

XSL•FO
RenderX

Abbreviations

EHR: electronic health record
KHB: Karlson-Holm-Breen
KMO: Kaiser-Meyer-Olkin
OR: odds ratio
PDRQ-9: Patient-Doctor Relationship Questionnaire-9 item
REDCap: Research Electronic Data Capture
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
WHO: World Health Organization

Edited by A Stone; submitted 12.Aug.2025; peer-reviewed by GY Han, R Guo; revised version received 20.Jan.2026; accepted 20.Jan.2026; published 31.Mar.2026.

Please cite as:
Zhai X, Zhou Z, Wang Z, Liu G, Coyte PC
Association Between Telemedicine Adoption and Physician Job Satisfaction: Cross-Sectional Study
J Med Internet Res 2026;28:e82285
URL: https://www.jmir.org/2026/1/e82285
doi:10.2196/82285

© Xiaohui Zhai, Zhongliang Zhou, Zhichao Wang, Guanping Liu, Peter C Coyte. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Integrating Generative AI Into Patient-Centered Clinical Decision Support: Viewpoint on Research and Practice Considerations

Prashila Dullabh[1], MD; Courtney Zott[1], MPH; Nicole Gauthreaux[1], MPH; Caroline Peterson[1], MPH; Abigail Aronoff[1], MPH; Kistein Monkhouse[2], MPA; Dean F Sittig[3], PhD

[1]Health Sciences Department, NORC at the University of Chicago, 1828 L Street NW, 9th Floor, Washington, DC, United States

[2]Patient Orator Inc, New York, NY, United States

[3]Informatics Review LLC, Lake Oswego, OR, United States

Corresponding Author:
Prashila Dullabh, MD
Health Sciences Department, NORC at the University of Chicago, 1828 L Street NW, 9th Floor, Washington, DC, United States

Abstract

There is growing interest in understanding how generative artificial intelligence (GenAI) can support patients and caregivers in making informed health care decisions, known as patient-centered clinical decision support (PC CDS). In this viewpoint, we present example applications for GenAI-supported PC CDS for patients, caregivers, clinicians, and patient-clinician interactions and examine the opportunities, challenges, and potential solutions associated with these applications. We conducted a targeted document review of our work in the Agency for Healthcare Research and Quality's Clinical Decision Support Innovation Collaborative focusing on GenAI-enabled PC CDS, supplemented by snowball sampling and targeted searches to identify additional applications. Findings were refined and validated through solicited feedback from a 20-member multidisciplinary expert panel. Through our work, we highlight six critical needs that must be addressed to fully realize GenAI's potential in PC CDS: (1) engage and ensure representation of patients and caregivers in design and development; (2) build the science of effective PC CDS implementation to support patient engagement; (3) develop risk-based policies for when to use GenAI; (4) establish independent testing and vetting criteria; (5) periodically reassess to identify and address algorithmic drift and verify performance; and (6) establish policies to promote transparency and patient consent in the use of GenAI. Understanding the applications and their potential implications for health care quality is essential to further the beneficial, ethical, and safe development of GenAI-supported PC CDS.

(J Med Internet Res 2026;28:e81628) doi:10.2196/81628

KEYWORDS

generative artificial intelligence; clinical decision support systems; patient-centered care; patient involvement; artificial intelligence; AI; quality of health care

Introduction

The clinical landscape is experiencing rapid growth in artificial intelligence (AI), especially generative AI (GenAI) systems that learn the underlying structure of their training data to create new content based on those learned patterns, such as new text, images, and data prompts [1]. This technology has sparked interest in its potential to integrate new data and analytics into patient-centered clinical decision support (PC CDS). PC CDS encompasses decision-making tools that significantly incorporate patient-centered factors across 4 key dimensions: knowledge (evidence based on comparative effectiveness research and research findings on patient-centered outcomes), informed by data that are directly generated from patients (ie, patient-generated health data [PGHD], patient-reported outcomes [PROs], and/or preferences), delivery (engaging patients and caregivers across settings through patient portals, apps, and other digital tools), and use (facilitating shared decision-making [SDM], a process where patients, caregivers, and care teams share and discuss health information and patients' values and preferences to reach mutually acceptable health-related decisions) [2]. Studies have suggested that GenAI has the potential to impact health care delivery, facilitate more personalized care, influence patient outcomes, and reshape the clinician-patient and clinician-caregiver relationship [3-5]. At the same time, researchers have identified barriers to realizing GenAI's potential, including limited GenAI implementation frameworks that consider the full sociotechnical environment [6], variability in requirements for reporting and validating GenAI [7], and mixed perceptions about the utility and safety of the technology among clinicians and patients [8]. Additionally, while many articles have described GenAI's short-term impact on clinicians and health care systems, such as relieving clinician workload and improving operational efficiency, to our knowledge, relatively few focus on the decision-making tools that support patient-centered care or on tools where the patient is the user [9-11]. Because these tools

XSL•FO
RenderX

directly engage patients in understanding health information, managing their conditions, and participating in care decisions, they raise distinct expectations around health literacy, such as supporting patients' ability to obtain, process, and understand health information [12]; self-management [13], such as strengthening patients' problem-solving, decision-making, and action-planning skills; and SDM, such as enhancing patients' knowledge, awareness, and ability to cope and engage in collaborative decisions with their care teams [1,14]. There is a need to understand the various ways and intricacies of leveraging GenAI to engage patients and caregivers in their care and how these tools can enhance patient-clinician communications and support the delivery of evidence-based care.

This viewpoint summarizes the novel applications of GenAI in PC CDS through illustrative use cases that can provide a better understanding of its potential among patients, caregivers, clinicians, and PC CDS developers. On the basis of these illustrative use cases, we identified common themes related to the benefits and challenges of GenAI-supported PC CDS tools on health care quality and delivery, with the purpose of acknowledging their advantages while mitigating their potential risks. Drawing from the challenges, we discuss ways to advance trustworthy GenAI-based PC CDS to chart a path forward for research and practice.

Targeted Review Approach

We used 3 methods to identify the use cases.

First, we conducted an in-depth review of 4 reports from the Agency for Healthcare Research and Quality's Clinical Decision Support Innovation Collaborative that addressed AI applications, including a landscape assessment on using AI to scale PC CDS [15], a report on patient and caregiver perspectives on GenAI in PC CDS [16], and 2 assessments of AI-supported PC CDS tools [17,18].

Second, we used a snowball sampling approach to identify additional relevant literature from the reference lists and conducted targeted searches in the PubMed and Google Scholar databases. In total, we identified 53 peer-reviewed sources. From these sources, we abstracted the type of AI technology used, the specific use cases, and the intended end users of these systems.

Third, we solicited feedback from a 20-member Steering Committee including patients, clinicians, informaticians, researchers, electronic health record (EHR) developers, payers, and policymakers. After the research team (PD, CZ, NG, and AA) completed the targeted review analysis and developed a draft use case exhibit, the lead author presented the preliminary findings to the Steering Committee and invited feedback on additional GenAI-supported PC CDS use cases to consider, the categorization of use cases, the validation of the identified benefits and considerations, and recommendations regarding critical needs for the effective use of GenAI-supported PC CDS. Meeting minutes documented key points of agreement, which were subsequently incorporated by the research team.

Use Cases for GenAI-Supported PC CDS

Figure 1 depicts use cases for GenAI-supported PC CDS that operate across categories based on their primary users and functions. Before reaching end users, GenAI can streamline the technical development of PC CDS for developers and implementers by optimizing knowledge artifacts and reducing development time and cost [15,19]. For end users, patient- and caregiver-focused tools emphasize health and well-being, chronic condition self-management, and educational resources. Tools supporting patient-clinician interactions facilitate communication through conversational agents, symptom monitoring systems, and treatment personalization for patients, which are discussed through SDM with their clinician. Clinician-focused applications assist with information management, particularly for analyzing large volumes of PGHD, while also providing diagnostic support and risk prediction. The potential impact and clinical significance of these use cases will vary across specialties, care settings, and patient populations, and the figure is not intended to imply a hierarchy of importance among them.

XSL•FO
RenderX

Figure 1. Use cases of generative artificial intelligence in patient-centered clinical decision support (PC CDS) technology. PGHD: patient-generated health data.

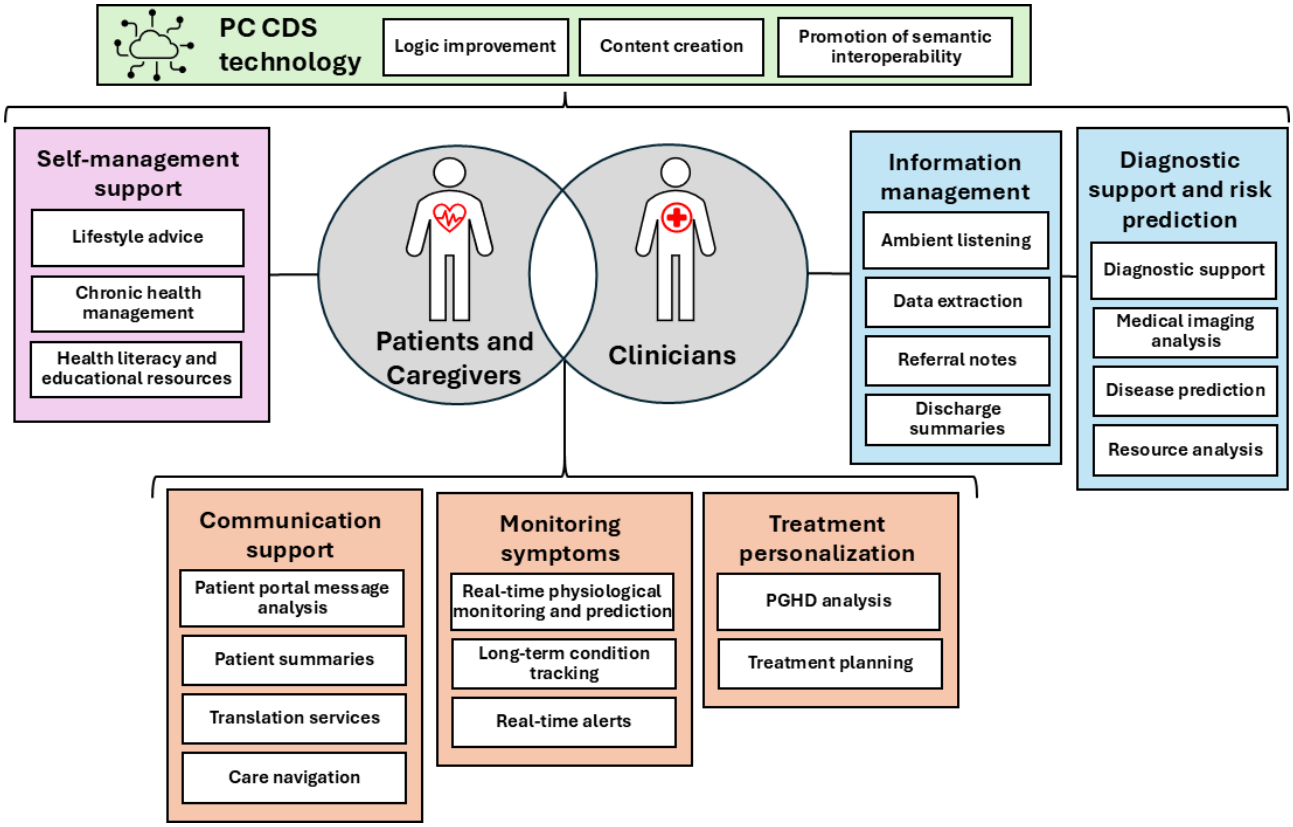

Salient Benefits and Considerations on the Use of GenAI in PC CDS

Overview

The use cases for GenAI-supported PC CDS illuminate several crosscutting themes that highlight the benefits and challenges of implementing these systems in health care settings. The themes do not apply uniformly across all use cases but rather vary in relevance depending on factors such as the clinical context, the decisions being made, and the patient-clinician interactions involved. Some themes also naturally intersect. For example, challenges such as hallucinations and misinformation cut across multiple applications and mediate a range of GenAI's benefits. In the following discussion, the examples specify the applicable use cases and raise challenges where they are highly relevant, illustrating how these themes can manifest across different applications.

Personalization and Precision

GenAI can support the delivery of personalized care that is responsive to a patient's unique health profile. GenAI is equipped to rapidly process extensive amounts of patient-centered health information from a range of sources, such as patients' EHRs, genomic data, medical devices, wearable technologies, and questionnaires assessing patients' preferences and goals [20]. It can then leverage complex algorithms to generate personalized treatment advice and recommendations, facilitating SDM [21-25]. For example, an AI-powered remote monitoring system for diabetes can collect and analyze patients' blood glucose levels, exercise levels, and eating habits to generate personalized meal plans and exercise routines for patients [26]. One study developed an AI-assisted Internet of Things wearable smartwatch prototype for older people to proactively detect and manage frailty through the collection of their physical activity data [27].

However, one of the key limitations of using GenAI to provide individualized health recommendations is its inconsistent reasoning capabilities and clinical judgment, particularly when it must synthesize information across long conversations, complex clinical scenarios, or high-volume data sources such as wearables. A systematic review found several issues with large language models (LLMs) such as fragile reasoning performance, diagnostic inaccuracies, and overly cautious clinical judgments across a range of use cases including treatment personalization, monitoring systems, communication support, and diagnostic support and risk prediction [28]. Others have raised concerns about personalized AI's potential to limit exposure to critical information, fostering "echo chambers" that reinforce users' specific beliefs [29].

Patient Engagement and Empowerment

Emerging use cases for GenAI center on its potential to empower patients through self-management and communication support. GenAI can make health information more understandable and equip patients with the information needed to take a more active role in their care. For example, health systems can use GenAI-powered chatbots or digital health assistants as patient-facing self-management resources and/or as communication support tools for clinicians to provide health information, answer questions, help assess symptoms, share

XSL•FO
RenderX

medication reminders, and facilitate appointment scheduling [30,31].

GenAI has the potential to support SDM between patients and clinicians by providing data or educational resources [32]. One study explored perspectives on the use of GenAI in SDM for knee replacement surgery, finding that patients viewed GenAI as another source of information that could enhance their understanding of their risk profile, empowering them to make decisions based on their values and preferences [10]. Despite its potential, patients with less experience using technology or navigating health information from electronic sources (ie, limited digital health literacy) may find using these tools challenging, leading to a sense of disempowerment [16]. With GenAI in particular, emerging evidence suggests that patients may mistake the confident tone of model outputs for factual accuracy, increasing the likelihood that they follow guidance that is incomplete or inaccurate [33-35].

Quality of Care and Safety

GenAI shows promise in improving patient safety and quality of care. It can facilitate real-time physiological monitoring and prediction of patients' conditions and send recommendations to patients or alerts to clinicians to support care coordination [36]. An ongoing study of a GenAI-supported PC CDS monitoring system for patients with asthma uses voice biomarker technology to recognize variations in the patient's recorded voice and sends immediate alerts to the patient for treatment intervention. The digital tool calculates a Respiratory Symptoms Risk Score and allows for remote care coordination, as necessary [37]. GenAI can also enhance diagnostic accuracy and risk prediction based on patient health histories [38], unstructured clinical notes [39,40], PGHD [41], or medical imaging analysis [42]. In some cases, this can inform allocation of resources and treatments based on individual- and system-level data. Mercy Healthcare System integrated the Chen Chemotherapy Model into their EHR to predict hospitalization risk from chemotherapy side effects in adult patients without leukemia. The texting platform prompts patients to report and rate their symptoms, then sends this PRO data to clinicians for review [43]. This allows clinicians to proactively manage symptoms before hospitalization becomes necessary. Furthermore, GenAI can summarize disparate information for clinicians to allow them more time for direct patient interactions, which can improve care quality [44]. With ambient listening technology, GenAI-supported tools convert recorded patient-clinician conversations into structured visit summary notes. These tools reduce the time clinicians spend on documentation during visits, enabling them to focus more attention on direct patient communication and interaction [45].

At the same time, overreliance on GenAI can pose risks to patient safety. One study found that GenAI models exhibited larger cognitive biases for medical decisions when compared with practicing clinicians, illustrating how LLMs could lead to diagnostic or treatment errors in complex medical cases [46]. Another study found that more than half of a sample of primary care physicians reviewing AI-generated patient portal messages containing errors did not identify and remedy all the errors and 35% to 45% of physicians submitted erroneous messages entirely unedited, indicating the need for additional guardrails beyond physician oversight [47]. Additionally, liability questions arise when GenAI is involved in decision-making, especially for "black box" algorithms where the internal decision-making process is hidden, meaning users can only see the input data they provide and the resulting output without understanding the reasoning process for medical predictions or decisions [48]. This makes it more difficult to determine who is accountable if there are adverse outcomes or whether the health system should take responsibility if they do not properly implement, maintain, or train users on the GenAI-supported system [5,49].

Access to Care

GenAI-supported PC CDS has the potential to improve access to medical care. It can support translation service use cases by using speech recognition and generating speech output in any language, thus overcoming language barriers between patients and their clinicians [50]. It can also be used for self-management support use cases in settings where there is a shortage of health care professionals, such as in rural areas or when more immediate access to clinical expertise is needed (regardless of geographic setting) to assist patients with nonurgent medical conditions. For example, Spänig et al [51] demonstrated that an autonomous medical AI interface designed to identify patients at risk for type 2 diabetes mellitus could function as an initial point of guidance in settings with limited clinician availability by offering preliminary risk assessment and directing patients toward appropriate follow-up care.

Despite this potential, GenAI systems may produce uneven health outcomes if not carefully implemented. Models trained on limited or skewed datasets could lead to less accurate self-management recommendations or physiological monitoring predictions for different patient populations [52]. Biased GenAI outputs can result from a range of common data quality issues in training models, such as limited demographic information [53,54], inaccurate or missing data [55], and unrepresentative data that do not include adequate information on specific patient subpopulations [56,57]. While bias can be mitigated by preprocessing data used to train AI models with retrieval-augmented generation (coupling models with an external and reliable database or knowledge base) [58], selecting models that prioritize transparency, and postprocessing the model output to correct for bias, these approaches require developer assessments of what constitutes fairness [59].

Beyond its potential for bias and error, a wide range of factors limit GenAI's integration into patients' lives, which can influence access to care when a GenAI-supported PC CDS tool serves as the first point of guidance in patient-facing use cases (eg, chronic health management and treatment planning). There may be usability challenges if digital and health literacy levels, community-specific factors, and local context are not considered during the design of GenAI-supported tools [60,61]. Additionally, chief medical information officers and other executives are essential since GenAI-supported tools require technological and financial resources to deploy and train care teams on their effective use. This creates the potential to exacerbate the digital divide between patients in health care organizations with robust health information technology support

XSL•FO
RenderX

systems versus patients in lower-resourced health care organizations, leaving some patients without access to GenAI-supported self-management, communication, or monitoring tools that can help them recognize when care is needed and/or remain engaged between visits.

Development of Clinical Decision Support Artifacts and Scaling

In addition to the clinician and patient and caregiver use cases, GenAI has the potential to improve the technology driving PC CDS. It can write code, map variables, create value sets [15], and generate realistic synthetic patient data that reduce the time needed for training and validation cycles [62]. GenAI can also suggest improvements to PC CDS logic [19] by analyzing alert overrides that contribute to clinician fatigue and personalizing alert criteria [63]. Furthermore, GenAI could solve a critical challenge in health care interoperability: the use of multiple incompatible information systems to record health data [64]. By using LLMs to quickly transform heterogeneous data into structured, standardized formats, it can facilitate the sharing of patient information across health systems. Despite this potential, developing interoperable GenAI systems is complex due to the quality of available patient data, the lack of implementation frameworks and standards for GenAI applications, and the cost of computational resources [65].

Areas to Advance GenAI-Supported PC CDS

Overview

The successful integration of GenAI in PC CDS across the use cases requires health system leadership, researchers, developers, informaticians, and policymakers to address at least the 6 areas discussed in subsequent sections to ensure these tools deliver meaningful, safe, and unbiased benefits. As with the benefits and challenges, these areas of opportunity are interconnected in nature and vary in relevance based on the clinical context, the decisions being made, and the patient-clinician interactions involved in the use case. Several of these areas align with the National Academy of Medicine's AI Code of Conduct framework for the development and application of responsible AI in health and medicine [66]. In addition, since PC CDS does not occur in isolation, its success depends on effective public health strategies that promote community health and well-being.

Engaging and Ensuring Representation of Patients and/or Caregivers in Design and Development

To capitalize on GenAI's ability to support the PC CDS use cases illustrated in this manuscript, engaging patients and caregivers during the design and development process is essential [67]. As reflected in the National Academy of Medicine's AI Code of Conduct, by involving end users in usability testing and other design activities—particularly those with varying language preferences, cultural norms, and digital and health literacy—developers can ensure GenAI-supported tools are tailored to account for the unique preferences, contexts, communication styles, and needs of individual patients [68,69]. Additionally, patient input can help ensure personalized treatment recommendations provided by GenAI are actionable for different populations [70] and accommodate the complexity of patients with multimorbidities. The degree to which patients and caregivers influence the design of GenAI-supported PC CDS tools will vary by use case, with greater influence typically possible in patient-facing applications and more limited involvement in clinician-facing functions such as diagnostic support or risk prediction.

Building the Science of Effective PC CDS Implementation to Support Patient Engagement

Advances in methods to promote the adoption of evidence-based practices are needed to address the barriers to patient engagement posed by patient-facing GenAI-supported PC CDS in particular. Research is needed on the multilevel factors that (1) influence the use of communication support tools such as chatbots and digital health assistants, such as inaccurately processing patient inputs or lack of personal connection [70]; (2) support their seamless integration into clinical workflows and patient lifeflows so that the information provided by the PC CDS does not undermine clinician communication; and (3) support GenAI's ability to facilitate SDM through visualizations that contain the appropriate level of detail [71,72]. A randomized clinical trial for a GenAI-enabled clinical decision aid used some of the aforementioned strategies through the incorporation of PRO measures, patient education, and tailoring to patient preferences, resulting in improvements in patients' decisional quality, SDM, and functional outcomes [73]. Although the research on effective PC CDS implementation is nascent, this evidence supports the importance of further exploring multilevel factors influencing patient engagement to improve uptake of GenAI-supported PC CDS.

Developing Risk-Based Policies for Deciding When GenAI Use Is Appropriate and What Level of Clinician Involvement Is Required

GenAI's ability to integrate into such a wide range of PC CDS use cases highlights the importance of developing clear policies for its use. Risk-based guardrails should define when GenAI can operate autonomously in supporting patients' knowledge and data capture (eg, by translating clinician-authored recommendations into plain language, generating visit summaries, or helping patients record symptoms and goals for minor issues) and where clinician review and sign-off are required before outputs are used to guide care (eg, for patients with multimorbidities who may face conflicting guidelines) [74]. Human-in-the-loop represents a promising approach for risk-based guardrails. Although its current use has largely focused on clinician-facing use cases such as verifying the accuracy and reliability of AI-generated discharge summaries or personalized recommendations, it also has potential to support patients with low digital health literacy in patient-facing use cases such as self-management support, communication support, and monitoring systems, where information is delivered digitally and patients may require additional support in interpretation. As seen in SDM research, patients with low digital health literacy benefit from clinician facilitation to interpret and apply complex health information, suggesting similar human-in-the-loop support may be beneficial for patient-facing GenAI-supported PC CDS tools [75]. While organizations bear

XSL•FO
RenderX

primary responsibility for ensuring the safe and appropriate use of the GenAI tools they deploy, patients' own GenAI literacy can also play a role in promoting appropriate use. Critical health AI literacy, the ability to critically evaluate GenAI outputs, recognizing how health systems and structural factors may shape them, and using this awareness to make informed decisions, is an emerging concept that can complement organizational safeguards by helping patients engage more safely and effectively with GenAI PC CDS [76]. Overall, well-defined policies will support postdeployment monitoring and help mitigate risks such as overreliance on GenAI, unclear liability, and inappropriate delegation of clinical judgment to technology.

Establishing Independent Testing and Vetting Criteria to Ensure Safety and Accuracy

To ensure the quality and safety of GenAI-supported PC CDS, developers must rigorously evaluate and independently test processes before deployment [74]. Unlike static decision support systems, GenAI-supported tools require initial validation not only for accuracy but also for fairness, interpretability, and robustness against biases [46]. Suggested approaches include conducting centralized testing through third parties, as well as requiring local oversight by adapting models such as the Clinical Laboratory Improvement Amendments for AI [77]. Establishing standardized oversight for GenAI helps build trust in its recommendations and ensure its deployment improves, not compromises, care quality.

Periodically Reassessing to Ensure No Algorithmic Drift and Verify Performance

GenAI algorithms and the data they operate on evolve, necessitating reassessment by the organization implementing GenAI-supported PC CDS to identify and address algorithmic drift, where shifts in data patterns lead to degraded performance or unintended outputs [74]. Because GenAI-supported PC CDS is integrated into real clinical workflows, monitoring approaches and metrics need to remain feasible and usable by clinical and organizational leaders, not just data scientists and researchers [78]. Several early prototypes, such as recent work developing human-centered AI monitoring systems for health care, demonstrate the utility of evaluating models along 4 core dimensions: performance, process, outcomes, and fairness [78]. For GenAI-supported PC CDS, outcomes and fairness can be assessed using similar principles, that is, evaluating whether recommendations lead to desired clinical and safety outcomes overall and across patient subgroups. Performance and process evaluation are more complex for GenAI because outputs are nondeterministic and language-based, and validated measures and methods are still emerging. Monitoring must therefore address behaviors such as hallucinations [79], reasoning quality [80], sycophantic behavior [81], and input robustness or feature drift [82]. While some elements require human expert review, automated and semiautomated tools are emerging. Examples include LLM-as-judge for evaluating correctness and guideline adherence or assessing fidelity to expected answers or text-based summaries of clinical information [83]; and statistical monitors that track token-level patterns or vocabulary shifts indicating changes in model behavior [82]. Organizations should establish clear triggers for reassessment—such as major guideline updates, new evidence affecting care recommendations, demographic changes in the patient population, the passage of time, or early signals of performance deterioration—and establish a reassessment frequency based on the clinical risk and operational impact of the GenAI-supported application [84,85].

Establishing Policies to Promote Transparency and Patient Consent in the Use of GenAI

Patients need clear explanations of when and how GenAI is used in PC CDS tools that influence their care, along with the technology's limitations [86]. The fair, appropriate, valid, effective, and safe principles [87] offer a blueprint for providing this transparency through specific, standardized technical and performance information. Furthermore, GenAI-supported PC CDS tools may require access to sensitive health information (eg, ambient listening tools and monitoring systems) with which patients may have varying levels of familiarity and comfort. Patients should have the right to opt out of the use of GenAI in their care and have full knowledge of who will access their data and how it will be used if they choose to participate. The development of patient-centered transparency and consent guidelines represents a critical area for advancing GenAI-supported PC CDS, including establishing which information is most relevant for patients to understand, determining how it should be effectively communicated, and considering when dynamic consent models may be warranted given the adaptive nature of GenAI systems [88].

Conclusions

The integration of GenAI into PC CDS holds promise for transforming health care delivery. GenAI can enhance the personalization and precision of care; empower patients through improved engagement and SDM; and support clinicians, patients, and caregivers in making more informed decisions. However, the successful implementation of GenAI in PC CDS requires consideration of several critical factors. These include addressing potential biases in GenAI algorithms, ensuring the transparency and explainability of GenAI-driven recommendations, and fostering trust among patients and clinicians. Additionally, it is essential to engage patients and caregivers in the design and development of GenAI-supported tools, develop clear policies for GenAI use, and establish rigorous testing and validation processes to ensure safety and accuracy. By addressing these challenges, the full potential of GenAI can begin to be harnessed to improve patient outcomes and advance the field of PC CDS.

Acknowledgments

The authors would like to acknowledge Edwin Lomotan, MD, and James Swiger, MBE, for their critical review and feedback on the manuscript. In addition, the authors acknowledge the following members of the Clinical Decision Support Innovation

XSL•FO
RenderX

Collaborative (CDSiC) Steering Committee and CDSiC Innovation Center Planning Committee: Dr Joel Andress; Dr James Cimino; Deborah Collyar; Angela Dobes; Dr Jordan Everson; Dr Sonja Fulmer; Dr Robert Greenes; Dr Tonya Hongsermeier; Dr Kensaku Kawamoto; Dr Gilad Kuperman; Dr Brian Levy; Dr Dave Little; Dr David Lobach; Dr J Marc Overhage; Tiffany Peterson; Dr Gerasimos Petratos; Dr Jonathan Teich; Dr Lipika Samal; Wesley Sargent; Patrick Schoen; Dr Richard Schreiber; Dr Scott Weingarten; Michael Wittie; and Dr Haipeng (Mark) Zhang.

Funding

This work is based on research conducted by NORC at the University of Chicago under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, Maryland (contract number 75Q80120D00018/75Q80121F32003). The views expressed in this paper are those of the authors and do not necessarily represent the official positions of the AHRQ or the US Department of Health and Human Services.

Data Availability

No datasets were generated or analyzed during this study.

Authors' Contributions

PD conceived of the presented idea and supervised the project. PD, CZ, NG, CP, and AA wrote the manuscript with input from all authors. DFS and KM provided critical feedback and helped shape the figure and manuscript.

Conflicts of Interest

None declared.

References

1. Feuerriegel S, Hartmann J, Janiesch C, Zschech P. Generative AI. Bus Inf Syst Eng 2024 Feb;66:111-126. [doi: 10.1007/s12599-023-00834-7]
2. Dullabh P, Sandberg SF, Heaney-Huls K, et al. Challenges and opportunities for advancing patient-centered clinical decision support: findings from a horizon scan. J Am Med Inform Assoc 2022 Jun 14;29(7):1233-1243. [doi: 10.1093/jamia/ocac059] [Medline: 35534996]
3. Ouanes K, Farhah N. Effectiveness of artificial intelligence (AI) in clinical decision support systems and care delivery. J Med Syst 2024 Aug 12;48(1):74. [doi: 10.1007/s10916-024-02098-4] [Medline: 39133332]
4. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
5. Braun M, Hummel P, Beck S, Dabrock P. Primer on an ethics of AI-based decision support systems in the clinic. J Med Ethics 2020 Apr 3;47(12):e3. [doi: 10.1136/medethics-2019-105860] [Medline: 32245804]
6. McCradden MD, Joshi S, Anderson JA, London AJ. A normative framework for artificial intelligence as a sociotechnical system in healthcare. Patterns (N Y) 2023 Nov 10;4(11):100864. [doi: 10.1016/j.patter.2023.100864] [Medline: 38035190]
7. Liaw ST, Liyanage H, Kuziemsky C, et al. Ethical use of electronic health record data and artificial intelligence: recommendations of the Primary Care Informatics Working Group of the International Medical Informatics Association. Yearb Med Inform 2020 Aug;29(1):51-57. [doi: 10.1055/s-0040-1701980] [Medline: 32303098]
8. Anjara SG, Janik A, Dunford-Stenger A, et al. Examining explainable clinical decision support systems with think aloud protocols. PLoS One 2023 Sep 14;18(9):e0291443. [doi: 10.1371/journal.pone.0291443] [Medline: 37708135]
9. Funer F, Schneider D, Heyen NB, et al. Impacts of clinical decision support systems on the relationship, communication, and shared decision-making between health care professionals and patients: multistakeholder interview study. J Med Internet Res 2024 Aug 23;26:e55717. [doi: 10.2196/55717] [Medline: 39178023]
10. Gould DJ, Dowsey MM, Glanville-Hearst M, et al. Patients' views on AI for risk prediction in shared decision-making for knee replacement surgery: qualitative interview study. J Med Internet Res 2023 Sep 18;25:e43632. [doi: 10.2196/43632] [Medline: 37721797]
11. Bjerring JC, Busch J. Artificial intelligence and patient-centered decision-making. Philos Technol 2021 Jun;34:349-371. [doi: 10.1007/s13347-019-00391-6]
12. Sørensen K, Van den Broucke S, Fullam J, et al. Health literacy and public health: a systematic review and integration of definitions and models. BMC Public Health 2012 Jan 25;12:80. [doi: 10.1186/1471-2458-12-80] [Medline: 22276600]
13. Lorig KR, Holman H. Self-management education: history, definition, outcomes, and mechanisms. Ann Behav Med 2003 Aug;26(1):1-7. [doi: 10.1207/S15324796ABM2601_01] [Medline: 12867348]
14. Elwyn G, Durand MA, Song J, et al. A three-talk model for shared decision making: multistage consultation process. BMJ 2017 Nov 6;359:j4891. [doi: 10.1136/bmj.j4891] [Medline: 29109079]
15. Kawamoto K, Ryan S, Heaney-Huls K, et al. Implementation, Adoption, and Scaling Workgroup: Landscape Assessment on the Use of Artificial Intelligence to Scale PC CDS: Agency for Healthcare Research and Quality; 2024.

XSL•FO
RenderX

16. Desai P, Dobes A, Shah A, et al. Trust and Patient-Centeredness Workgroup: Patient and Caregiver Perspectives on Generative Artificial Intelligence in Patient-Centered Clinical Decision Support [Internet]: Agency for Healthcare Research and Quality (US); 2024.
17. Zott C, Sittig DF, Gauthreaux N, et al. PAIGE Chatbot For Patient-Clinician Communication: Usability and Utility Assessment: Agency for Healthcare Research and Quality (US); 2024.
18. Gauthreaux N, Zott C, Boxwala A, Dullabh PM, Sittig DF. Quartz App to Support Medication Adherence: Usability and Feasibility Assessment [Internet]: Agency for Healthcare Research and Quality (US); 2024.
19. Liu S, Wright AP, Patterson BL, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023 Jun 20;30(7):1237-1245. [doi: 10.1093/jamia/ocad072] [Medline: 37087108]
20. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. In: Artificial Intelligence in Healthcare: Academic Press; 2020:25-60. [doi: 10.1016/B978-0-12-818438-7.00002-2]
21. Love-Koh J, Peel A, Rejon-Parrilla JC, et al. The future of precision medicine: potential impacts for health technology assessment. Pharmacoeconomics 2018 Dec;36(12):1439-1451. [doi: 10.1007/s40273-018-0686-6] [Medline: 30003435]
22. Horiuchi D, Tatekawa H, Shimono T, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. Neuroradiology 2024 Jan;66(1):73-79. [doi: 10.1007/s00234-023-03252-4] [Medline: 37994939]
23. Fathima M, Moulana M. Revolutionizing breast cancer care: AI-enhanced diagnosis and patient history. Comput Methods Biomech Biomed Engin 2025 Apr;28(5):642-654. [doi: 10.1080/10255842.2023.2300681] [Medline: 38178694]
24. Ramkumar PN, Haeberle HS, Ramanathan D, et al. Remote patient monitoring using mobile health for total knee arthroplasty: validation of a wearable and machine learning-based surveillance platform. J Arthroplasty 2019 Oct;34(10):2253-2259. [doi: 10.1016/j.arth.2019.05.021] [Medline: 31128890]
25. Fozoonmayeh D, Le HV, Wittfoth E, et al. A scalable smartwatch-based medication intake detection system using distributed machine learning. J Med Syst 2020 Feb 28;44(4):76. [doi: 10.1007/s10916-019-1518-8] [Medline: 32112271]
26. Khalifa M, Albadawy M, Iqbal U. Advancing clinical decision support: the role of artificial intelligence across six domains. Comput Methods Programs Biomed Update 2024;5:100142 Retracted inComput Methods Programs Biomed Update. 2025;8:100204. [doi: 10.1016/j.cmpbup.2025.100204]
27. Ciubotaru BI, Sasu GV, Goga N, et al. Prototype results of an Internet of Things system using wearables and artificial intelligence for the detection of frailty in elderly people. Appl Sci 2023;13(15):8702. [doi: 10.3390/app13158702]
28. Souza GD, Melo G, Schneider D. Tailoring treatment in the age of AI: a systematic review of large language models in personalized healthcare. Informatics 2025;12(4):113. [doi: 10.3390/informatics12040113]
29. Kostick-Quenet KM. A caution against customized AI in healthcare. NPJ Digit Med 2025 Jan 7;8(1):13. [doi: 10.1038/s41746-024-01415-y] [Medline: 39774331]
30. Thorat V, Rao P, Joshi N, Talreja P, Shetty AR. Role of artificial intelligence (AI) in patient education and communication in dentistry. Cureus 2024 May 7;16(5):e59799. [doi: 10.7759/cureus.59799] [Medline: 38846249]
31. Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, Temsah MH. Artificial intelligence (AI) chatbots in medicine: a supplement, not a substitute. Cureus 2023 Jun 25;15(6):e40922. [doi: 10.7759/cureus.40922] [Medline: 37496532]
32. Abbasgholizadeh Rahimi S, Cwintal M, Huang Y, et al. Application of artificial intelligence in shared decision making: scoping review. JMIR Med Inform 2022 Aug 9;10(8):e36199. [doi: 10.2196/36199] [Medline: 35943793]
33. Hart R. Chatbots can trigger a mental health crisis. What to know about "AI psychosis". TIME. 2025 Aug 5. URL: https://time.com/7307589/ai-psychosis-chatgpt-mental-health/ [accessed 2025-12-10]
34. Dober C. Using generative AI for therapy might feel like a lifeline—but there's danger in seeking certainty in a chatbot. The Guardian. 2025 Aug 3. URL: https://www.theguardian.com/commentisfree/2025/aug/03/generative-ai-chatbot-therapy-dangers-risks [accessed 2025-12-10]
35. Draelos RL, Afreen S, Blasko B, et al. Large language models provide unsafe answers to patient-posed medical questions. arXiv. Preprint posted online on Jul 25, 2025. [doi: 10.48550/arXiv.2507.18905]
36. A novel patient-facing mobile platform to collect and implement patient-reported outcomes and voice biomarkers in underserved adult patients with asthma. Agency for Healthcare Research and Quality. URL: https://digital.ahrq.gov/ahrq-funded-projects/novel-patient-facing-mobile-platform-collect-and-implement-patient-reported [accessed 2025-02-11]
37. ASTHMAXcel voice mobile application to improve chronic disease management and patient outcomes.: Agency for Healthcare Research and Quality URL: https://web.archive.org/web/20251110170833/https://digital.ahrq.gov/program-overview/research-stories/asthmaxcel-voice-mobile-application-improve-chronic-disease [accessed 2025-12-03]
38. Feng T. Applications of artificial intelligence to diagnosis of neurodegenerative diseases. Stud Health Technol Inform 2023 Nov 23;308:648-655. [doi: 10.3233/SHTI230896] [Medline: 38007795]
39. Jee J, Fong C, Pichotta K, et al. Automated real-world data integration improves cancer outcome prediction. Nature 2024 Dec;636(8043):728-736. [doi: 10.1038/s41586-024-08167-5] [Medline: 39506116]
40. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. NPJ Digit Med 2022 Dec 26;5(1):194. [doi: 10.1038/s41746-022-00742-2] [Medline: 36572766]

XSL•FO
RenderX

41. Ye J, Woods D, Jordan N, Starren J. The role of artificial intelligence for the application of integrating electronic health records and patient-generated data in clinical decision support. AMIA Jt Summits Transl Sci Proc 2024 May 31;2024:459-467. [Medline: 38827061]
42. Armato SG, Drukker K, Hadjiiski L. AI in medical imaging grand challenges: translation from competition to research benefit and patient care. Br J Radiol 2023 Oct;96(1150):20221152. [doi: 10.1259/bjr.20221152] [Medline: 37698542]
43. DeFreitas M. Can AI help monitor chemotherapy side effects? HealthLeaders. 2024 Jan 8. URL: https://www.healthleadersmedia.com/technology/can-ai-help-monitor-chemotherapy-side-effects [accessed 2025-12-09]
44. Maleki Varnosfaderani S, Forouzanfar M. The role of AI in hospitals and clinics: transforming healthcare in the 21st century. Bioengineering (Basel) 2024 Mar 29;11(4):337. [doi: 10.3390/bioengineering11040337] [Medline: 38671759]
45. Eastwood B. Ambient listening in healthcare: dictation, documentation and AI. HealthTech. 2024 Aug 7. URL: https://healthtechmagazine.net/article/2024/08/ambient-listening-inhealthcare-perfcon [accessed 2025-04-08]
46. Wang J, Redelmeier DA. Cognitive biases and artificial intelligence. NEJM AI 2024 Nov 27;1(12). [doi: 10.1056/AIcs2400639]
47. Biro JM, Handley JL, Malcolm McCurry J, et al. Opportunities and risks of artificial intelligence in patient portal messaging in primary care. NPJ Digit Med 2025 Apr 24;8(1):222. [doi: 10.1038/s41746-025-01586-2] [Medline: 40275104]
48. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics 2021 Mar 18:medethics-2020-106820. [doi: 10.1136/medethics-2020-106820] [Medline: 33737318]
49. Maliha G, Gerke S, Cohen IG, Parikh RB. Artificial intelligence and liability in medicine: balancing safety and innovation. Milbank Q 2021 Sep;99(3):629-647. [doi: 10.1111/1468-0009.12504] [Medline: 33822422]
50. Kim H, Jin HM, Jung YB, You SC. Patient-friendly discharge summaries in Korea based on ChatGPT: software development and validation. J Korean Med Sci 2024 Apr 29;39(16):e148. [doi: 10.3346/jkms.2024.39.e148] [Medline: 38685890]
51. Spänig S, Emberger-Klein A, Sowa JP, Canbay A, Menrad K, Heider D. The virtual doctor: an interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes. Artif Intell Med 2019 Sep;100:101706. [doi: 10.1016/j.artmed.2019.101706] [Medline: 31607340]
52. Bullerman B. Joint commission on accreditation of healthcare organizations. United States of America. Implicit Bias in Healthcare: A Quick Safety Review. 2026 Mar 23. URL: https://coilink.org/20.500.12592/gbpfpg [accessed 2024-05-01]
53. Akay EM, Hilbert A, Carlisle BG, Madai VI, Mutke MA, Frey D. Artificial intelligence for clinical decision support in acute ischemic stroke: a systematic review. Stroke 2023 Jun;54(6):1505-1516. [doi: 10.1161/STROKEAHA.122.041442] [Medline: 37216446]
54. Jeong HK, Park C, Henao R, Kheterpal M. Deep learning in dermatology: a systematic review of current approaches, outcomes, and limitations. JID Innov 2022;3(1):100150. [doi: 10.1016/j.xjidi.2022.100150] [Medline: 36655135]
55. Balla Y, Tirunagari S, Windridge D. Pediatrics in artificial intelligence era: a systematic review on challenges, opportunities, and explainability. Indian Pediatr 2023 Jul 15;60(7):561-569. [Medline: 37424120]
56. Clement J, Maldonado AQ. Augmenting the transplant team with artificial intelligence: toward meaningful AI use in solid organ transplant. Front Immunol 2021;12:694222. [doi: 10.3389/fimmu.2021.694222] [Medline: 34177958]
57. Buchlak QD, Esmaili N, Leveque JC, et al. Machine learning applications to clinical decision support in neurosurgery: an artificial intelligence augmented systematic review. Neurosurg Rev 2020 Oct;43(5):1235-1253. [doi: 10.1007/s10143-019-01163-8] [Medline: 31422572]
58. Shi W, Zhuang Y, Zhu Y, Iwinski H, Wattenbarger M, Wang MD. Retrieval-augmented large language models for adolescent idiopathic scoliosis patients in shared decision-making. Presented at: ACM-BCB 2023: 14th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics; Sep 3-6, 2023. [doi: 10.1145/3584371.3612956]
59. Ferrara E. Fairness and bias in artificial intelligence: a brief survey of sources, impacts, and mitigation strategies. Sci 2024;6(1):3. [doi: 10.3390/sci6010003]
60. Aung YYM, Wong DCS, Ting DSW. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. Br Med Bull 2021 Sep 10;139(1):4-15. [doi: 10.1093/bmb/ldab016] [Medline: 34405854]
61. Nair M, Svedberg P, Larsson I, Nygren JM. A comprehensive overview of barriers and strategies for AI implementation in healthcare: mixed-method design. PLoS One 2024;19(8):e0305949. [doi: 10.1371/journal.pone.0305949] [Medline: 39121051]
62. Reddy S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. Implement Sci 2024 Mar 15;19(1):27. [doi: 10.1186/s13012-024-01357-9] [Medline: 38491544]
63. Liu S, McCoy AB, Peterson JF, et al. Leveraging explainable artificial intelligence to optimize clinical decision support. J Am Med Inform Assoc 2024 Apr 3;31(4):968-974. [doi: 10.1093/jamia/ocae019] [Medline: 38383050]
64. Iroju O, Soriyan A, Gambo I, Olaleke J. Interoperability in healthcare: benefits, challenges and resolutions. Int J Innov Appl Stud 2013;3(1):262-270 [FREE Full text]
65. Yim D, Khuntia J, Parameswaran V, Meyers A. Preliminary evidence of the use of generative AI in health care clinical services: systematic narrative review. JMIR Med Inform 2024 Mar 20;12:e52073. [doi: 10.2196/52073] [Medline: 38506918]
66. An Artificial Intelligence Code of Conduct for Health and Medicine: Essential Guidance for Aligned Action: National Academy of Medicine; 2025.

XSL•FO
RenderX

67. Dullabh P, Dungan R, Raj M, et al. Trust & Patient-Centeredness Workgroup: Methods for Involving End-Users in PC CDS Co-Design [Internet]: Agency for Healthcare Research and Quality (US); 2023.
68. Desai PJ, Zott C, Gauthreaux N, et al. Trust & Patient-Centeredness Workgroup: An Introductory Handbook for Patient Engagement Throughout the Patient-Centered Clinical Decision Support Lifecycle [Internet]: Agency for Healthcare Research and Quality (US); 2023.
69. Pham Q, Gamble A, Hearn J, Cafazzo JA. The need for ethnoracial equity in artificial intelligence for diabetes management: review and recommendations. J Med Internet Res 2021 Feb 10;23(2):e22320. [doi: 10.2196/22320] [Medline: 33565982]
70. Nadarzynski T, Knights N, Husbands D, et al. Achieving health equity through conversational AI: a roadmap for design and implementation of inclusive chatbots in healthcare. PLOS Digit Health 2024 May 2;3(5):e0000492. [doi: 10.1371/journal.pdig.0000492] [Medline: 38696359]
71. Milne-Ives M, de Cock C, Lim E, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020 Oct 22;22(10):e20346. [doi: 10.2196/20346] [Medline: 33090118]
72. Robinson R, Liday C, Lee S, et al. Artificial intelligence in health care-understanding patient information needs and designing comprehensible transparency: qualitative study. JMIR AI 2023;2:e46487. [doi: 10.2196/46487] [Medline: 38333424]
73. Jayakumar P, Moore MG, Furlough KA, et al. Comparison of an artificial intelligence-enabled patient decision aid vs educational material on decision quality, shared decision-making, patient experience, and functional outcomes in adults with knee osteoarthritis: a randomized clinical trial. JAMA Netw Open 2021 Feb 1;4(2):e2037107. [doi: 10.1001/jamanetworkopen.2020.37107] [Medline: 33599773]
74. Labkoff S, Oladimeji B, Kannry J, et al. Toward a responsible future: recommendations for AI-enabled clinical decision support. J Am Med Inform Assoc 2024 Nov 1;31(11):2730-2739. [doi: 10.1093/jamia/ocae209] [Medline: 39325508]
75. Gartner B, Leysen D, Mcgowan H, et al. Digitally supported shared decision-making for exercise prescription in the secondary prevention of cardiovascular disease. Eur J Prev Cardiol 2025 May 19;32(Supplement_1). [doi: 10.1093/eurjpc/zwaf236.362]
76. Campos H, Salmi L. Critical AI health literacy as liberation technology: a new skill for patient empowerment. NAM Perspect 2025;12. [doi: 10.31478/202512a]
77. Jackson BR, Sendak MP, Solomonides A, Balu S, Sittig DF. Regulation of artificial intelligence in healthcare: Clinical Laboratory Improvement Amendments (CLIA) as a model. J Am Med Inform Assoc 2025 Feb 1;32(2):404-407. [doi: 10.1093/jamia/ocae296] [Medline: 39657218]
78. Salwei ME, Davis SE, Reale C, et al. Human-centered design of an artificial intelligence monitoring system: the Vanderbilt Algorithmovigilance Monitoring and Operations System. JAMIA Open 2025 Oct;8(5):ooaf136. [doi: 10.1093/jamiaopen/ooaf136] [Medline: 41180894]
79. Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med 2025 May 13;8(1):274. [doi: 10.1038/s41746-025-01670-7] [Medline: 40360677]
80. Qiu P, Wu C, Liu S, et al. Quantifying the reasoning abilities of LLMs on clinical cases. Nat Commun 2025 Nov 6;16(1):9799. [doi: 10.1038/s41467-025-64769-1] [Medline: 41198657]
81. Chen S, Gao M, Sasse K, et al. When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. NPJ Digit Med 2025 Oct 17;8(1):605. [doi: 10.1038/s41746-025-02008-z] [Medline: 41107408]
82. Kellogg KC, Ye B, Hu Y, Savova GK, Wallace B, Bitterman DS. Large language models require a new form of oversight: capability-based monitoring. arXiv. Preprint posted online on Nov 5, 2025. [doi: 10.48550/arXiv.2511.03106]
83. Croxford E, Gao Y, First E, et al. Evaluating clinical AI summaries with large language models as judges. NPJ Digit Med 2025 Nov 5;8(1):640. [doi: 10.1038/s41746-025-02005-2] [Medline: 41193667]
84. Davis SE, Walsh CG, Matheny ME. Open questions and research gaps for monitoring and updating AI-enabled tools in clinical settings. Front Digit Health 2022;4:958284. [doi: 10.3389/fdgth.2022.958284] [Medline: 36120717]
85. Davis SE, Embí PJ, Matheny ME. Sustainable deployment of clinical prediction tools-a 360° approach to model maintenance. J Am Med Inform Assoc 2024 Apr 19;31(5):1195-1198. [doi: 10.1093/jamia/ocae036] [Medline: 38422379]
86. Sittig DF, Singh H. Recommendations to ensure safety of AI in real-world clinical care. JAMA 2025 Feb 11;333(6):457-458. [doi: 10.1001/jama.2024.24598] [Medline: 39602298]
87. Health data, technology, and interoperability: certification program updates, algorithm transparency, and information sharing. Office of the National Coordinator for Health Information Technology (ONC), Department of Health and Human Services (HHS). 2023. URL: https://healthit.gov/wp-content/uploads/2023/12/hti-1-final-rule.pdf [accessed 2025-03-17]
88. Dullabh P, Dhopeshwarkar R, Leaphart D, et al. Trustworthy Artificial Intelligence (TAI) for Patient-Centered Outcomes Research (PCOR): Office of the Assistant Secretary for Planning and Evaluation (ASPE); 2023.

Abbreviations

AI: artificial intelligence
EHR: electronic health record
GenAI: generative artificial intelligence
LLM: large language model

XSL•FO
RenderX

CDS: patient-centered clinical decision support
PGHD: patient-generated health data
PRO: patient-reported outcome
SDM: shared decision-making

Edited by A Mavragani; submitted 01.Aug.2025; peer-reviewed by D Hynes, H Lv, K Kier, N Newton; revised version received 03.Feb.2026; accepted 27.Feb.2026; published 01.Apr.2026.

Please cite as:
Dullabh P, Zott C, Gauthreaux N, Peterson C, Aronoff A, Monkhouse K, Sittig DF
Integrating Generative AI Into Patient-Centered Clinical Decision Support: Viewpoint on Research and Practice Considerations
J Med Internet Res 2026;28:e81628
URL: https://www.jmir.org/2026/1/e81628
doi:10.2196/81628

© Prashila Dullabh, Courtney Zott, Nicole Gauthreaux, Caroline Peterson, Abigail Aronoff, Kistein Monkhouse, Dean F Sittig. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Deep Research Agents: Major Breakthrough or Incremental Progress for Medical AI?

Matthew Yu Heng Wong[1], BA, MB, BChir; Ariel Yuhan Ong[2,3], MBChB, FRCOphth; David A Merle[2,3], MD, FEBO; Pearse A Keane[2,3], MD, FRCOphth

[1]School of Clinical Medicine, University of Cambridge, Addenbrooke's Hospital, Hills Rd, Cambridge, United Kingdom
[2]NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust, London, United Kingdom
[3]Institute of Ophthalmology, University College London, London, United Kingdom

Corresponding Author:
Matthew Yu Heng Wong, BA, MB, BChir
School of Clinical Medicine, University of Cambridge, Addenbrooke's Hospital, Hills Rd, Cambridge, United Kingdom

Abstract

Deep research agents are autonomous large language model–based systems capable of iterative web search, retrieval, and synthesis. They are increasingly positioned as the next major leap in medical artificial intelligence. In this viewpoint, we argue that while these agents mark progress in information access and workflow automation, they represent an incremental evolution rather than a paradigm shift. We review current applications of deep research agents in biomedical scenarios, including literature review generation, clinical evidence synthesis, guideline comparison, and patient education. Across these early use cases, the tools demonstrate the ability to rapidly gather and structure up-to-date information, often producing outputs that appear comprehensive and well-referenced. However, these strengths coexist with unresolved and clinically significant limitations. Citation fidelity remains inconsistent across models, with subtle misinterpretations or unreliable references still common. Their retrieval processes and evidence-ranking mechanisms remain opaque, raising concerns about reproducibility and hidden biases. Moreover, overreliance on artificial intelligence–generated syntheses risks eroding clinicians' critical appraisal skills and may introduce automation bias at a time when medicine increasingly requires deeper scrutiny of information sources. Safety constraints are also less predictable within multistep research pipelines, increasing the risk of harmful or inappropriate outputs. Finally, current evidence is largely limited to proof-of-concept evaluations, with little evidence from real-life clinical deployment. We contend that deep research agents should be embraced as assistive research tools rather than pseudoexperts. Their value lies in accelerating information gathering, not replacing rigorous human judgment. Realizing their potential will require transparent retrieval architectures, robust benchmarking, and explicit educational integration to preserve clinicians' evaluative reasoning. Used judiciously, these systems could enrich medical research and practice; used uncritically, they risk amplifying errors at scale. We contend that deep research agents should be embraced as assistive research tools rather than pseudoexperts. Their value lies in accelerating information gathering, not replacing rigorous human judgment. Realizing their potential will require transparent retrieval architectures, robust benchmarking, and explicit educational integration to preserve clinicians' evaluative reasoning. Used judiciously, these systems could enrich medical research and practice; used uncritically, they risk amplifying errors at scale.

(J Med Internet Res 2026;28:e88195) doi:10.2196/88195

KEYWORDS

artificial intelligence; AI; medical research; scientific writing; large language models; LLMs

Introduction

Large language models (LLMs) have now become extensively tested and used for a range of medical tasks and capabilities [1,2]. However, current LLMs are limited by issues such as hallucinations, opaque reasoning processes, and reliance on prior static training data. Retrieval-augmented generation (RAG) aims to mitigate some of these limitations by allowing LLMs to draw from external information sources (eg, specialized medical textbooks and journals) [3]. In late 2024, another advancement in the development of LLMs was introduced in the form of "deep research agents"—autonomous research assistants that combine LLMs with real-time internet search and citation capabilities [4]. These agents leverage the capabilities of RAG by replacing this external data source with up-to-date, live internet access [5].

Beyond simple responses to standard chat-based prompting, these deep research agents can autonomously perform web searches, retrieve, cross-reference, and cite literature, positioning themselves as potential replacements for a myriad of human-led tasks. OpenAI's Deep Research (launched in February 2025) was designed to rapidly access and analyze vast datasets of literature, synthesize their findings, and then generate tailored written outputs [6]. These tools promise to bridge the gap

XSL•FO
RenderX

between an ever-growing medical knowledge base and the limited capacity clinicians have to digest it. With minimal user input, they can autonomously generate comprehensive, well-structured, and referenced reviews of prespecified topics. In this viewpoint, we use "deep research agents" (lowercase) to denote the general class of agentic, multistep, research-oriented LLM systems. The capitalized term "Deep Research" is reserved exclusively for specific commercial implementations (eg, OpenAI's Deep Research) and is used only where product-level distinctions are necessary.

Deep research agents are a relatively recent development, and their application in biomedical contexts remains only partially investigated. In this study, we review how deep research agents have been explored in early use cases in medicine, ranging from literature reviews and medical education to clinical queries, highlighting their anticipated potential. We also identify areas where these agents are limited and require further study and careful validation as usage and adoption grow. Overall, we argue that while these agents mark progress in information access and workflow automation, they represent an incremental evolution in medical artificial intelligence (AI) rather than a paradigm shift.

Deep Research Agents Versus Conventional LLM Chatbots for Medicine

Deep research agents represent a new evolution of conventional chatbot-style LLMs, which integrate RAG workflows and more complex attention mechanisms. Unlike standard LLM chatbots, which typically rely on the model's internal weights and a single input context window, deep research systems typically embed a multistep reasoning pipeline: they first query external knowledge sources, retrieve relevant documents, and then feed those passages into the generator, which uses cross-attention to integrate information across multiple sources before producing a response (Figure 1). Importantly, unlike single-shot retrieval in traditional RAG pipelines, web exploration and source retrieval are iterative and agent-driven, enabling deeper coverage of sparse or scattered evidence [7].

Figure 1. Deep research agents autonomously guide the entire workflow: starting from user input, optionally clarifying intent and planning the search, iteratively using tools for offline data retrieval (databases, vector stores), online search (APIs, browsers), and higher-level functions such as analytics or multimodal content creation (text, image), before assembling the findings into a structured report [8]. API: application programming interface; LLM: large language model.

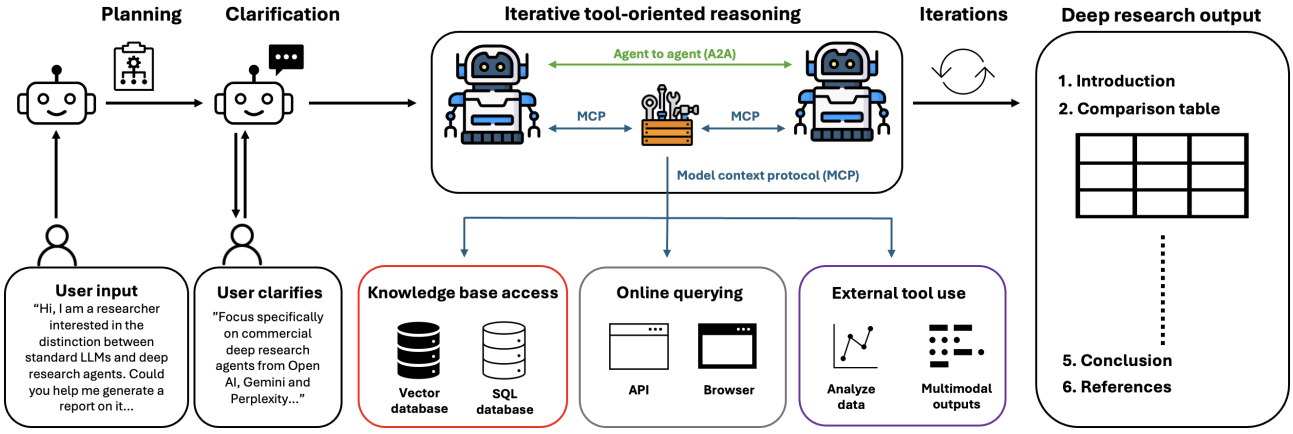

Deep research agents extend RAG by combining retrieval with real-time tool use and iterative reasoning, allowing the system not only to fetch relevant documents but also to plan, refine, and synthesize information as part of a single workflow. Dynamically engaging with external tools and managing multistage research tasks in real time ostensibly allows these agents to provide greater autonomy, context-awareness, and accuracy in their outputs (Table 1) [8].

XSL•FO
RenderX

Table . Architectural and functional comparison of deep research agents vs standard retrieval-augmented generation chatbots.

Dimension	Deep research agents (eg, OpenAI Deep Research, Gemini Deep Research, Perplexity Pro)	Standard RAG[a] or web-enabled Chatbots (single-shot retrieval + generation; optional browsing)
Core behavior	Multistep autonomous pipeline: *plan* → *search* → *retrieve* → *read* → *synthesize* → *cite*. Persistent reasoning and external grounding.	Nonagentic workflow: answer generation supported by one-shot retrieval or ad hoc browsing; limited planning and iteration.
External information integration	Integrated web and file retrieval; dynamic multisource querying with internal memory.	Single-pass retrieval (*top-k* docs); weaker systematic coverage and cross-source reconciliation.
Processing time or run length	Long-running sessions (≈5 - 30 min typical) processing many documents and verifying citations.	Short latency (seconds to a few minutes) typically fewer tool calls and shallower document review.
Citation handling or attribution	Structured outputs with traceable references and in-line links; still requires human verification for claim-source alignment.	Can provide citations when configured or prompted. May show weaker claim-source alignment.
RAG	Full multihop RAG with iterative query refinement, retrieval, cross-checking, and synthesis.	Single-hop RAG; limited multihop behavior unless manually orchestrated.
Control and reasoning architecture	Agentic controller layer atop the LLM[b] for planning, tool selection, and multisource integration.	Fixed LLM + retriever. Reasoning and evidence integration occur implicitly within a single LLM pass.
Output style	Structured research reports with citations and sectioning (eg, “Findings,” “Limitations”), can be tailored with prompt engineering.	Conversational prose or short summaries without specific prompting.
Limitations or risks	Improved grounding but still susceptible to retrieval errors, misweighting of evidence, opaque rankings of sources, and hallucinations.	Similar risks (retrieval noise, cherry-picked snippets, hallucinations) and weaker systematic coverage. Higher risk of hallucinated references.
Potential use cases	Systematic reviews and evidence synthesis, where provenance and depth matter, and slow generation is acceptable.	Point queries, quick updates, short summaries with citations. For tasks where speed matters.

[a]RAG: retrieval-augmented generation.

[b]LLM: large language model.

Compared with their early 2025 iterations, current deep research agents have shifted from text-only, single-model research bots into multimodal, tool-integrated, multimodel research systems. This evolution includes enhanced perceptual capabilities, allowing agents to interpret not only text but also visual elements such as figures, images, and complex documents, alongside expanded access to public sources (eg, PDFs from the web) through external connectors. In parallel, successive model upgrades have incorporated advanced self-reflection mechanisms, increases in scalability, and longer context windows. For example, newer iterations of Gemini’s Deep Research employ “Flash Thinking” to support more explicit planning and iterative self-critique during synthesis [9].

From a clinical perspective, this architecture could one day transform how evidence is synthesized and applied in medicine and health care. By autonomously retrieving and cross-referencing peer-reviewed studies, clinical guidelines, and trial reports, deep research agents could assist in rapid reviews**,** clinical decision support, and precision medicine literature surveillance (Table 2). For instance, a clinician could enter a query to “Identify emerging 2024 - 2025 biomarkers predictive of immunotherapy response across cancer types” and receive a sourced, citation-backed synthesis reflecting the latest evidence—something standard LLM chatbots, limited to static parametric memory, cannot reliably provide. This output strongly hints at the future of AI-assisted knowledge retrieval. Moreover, the use of cross-attention between retrieved documents allows the model to contextualize and reconcile conflicting studies**,** potentially supporting meta-analytic or guideline-development workflows [8]. We briefly tested this workflow and evaluated the resulting retrieval and output of the model (Figures 2 and 3).

XSL•FO
RenderX

Table . Illustrative deep research queries showcasing the breadth of use cases for clinicians and academics.

Category	Example query	Potential benefits
Clinical updates	"What emerging biomarkers beyond programmed death-ligand 1 predict response to immunotherapy, based on recent clinical studies?"	Provides clinicians with current knowledge without manually searching journals and reviewing multiple articles.
Guideline comparison	"How do European Society for Medical Oncology and American Society of Clinical Oncology guidelines address immunotherapy response in lung cancer?"	Rapid evidence synthesis across guideline bodies.
Basic science research	"What biological mechanisms underlie emerging biomarkers predictive of immunotherapy response?"	Elucidates underlying molecular mechanisms and pathways.
Translational research	"Which emerging biomarkers are sufficiently validated and assay-ready to inform immunotherapy patient selection?"	Highlights cutting-edge translational trends.
Patient communication	"Explain in simple terms to a patient how their biomarkers may help predict whether immunotherapy will work."	Generates layperson-friendly patient information.
Academic writing	"Draft an introduction for a review summarizing emerging biomarkers predictive of immunotherapy response."	Saves time on first-draft writing.
Whimsical or exploratory	"What will clinicians recognize as the moment when immunotherapy response first became meaningfully predictable?"	Engages creativity and unlocks novel perspectives.

Figure 2. An example of how a deep research agent (OpenAI's Deep Research) engages in an iterative dialog with the user to clarify intent before conducting literature searches. By confirming details such as cancer type and biomarker scope, the agent supports a more focused and structured synthesis—a workflow that extends beyond typical single-turn large language model interactions.

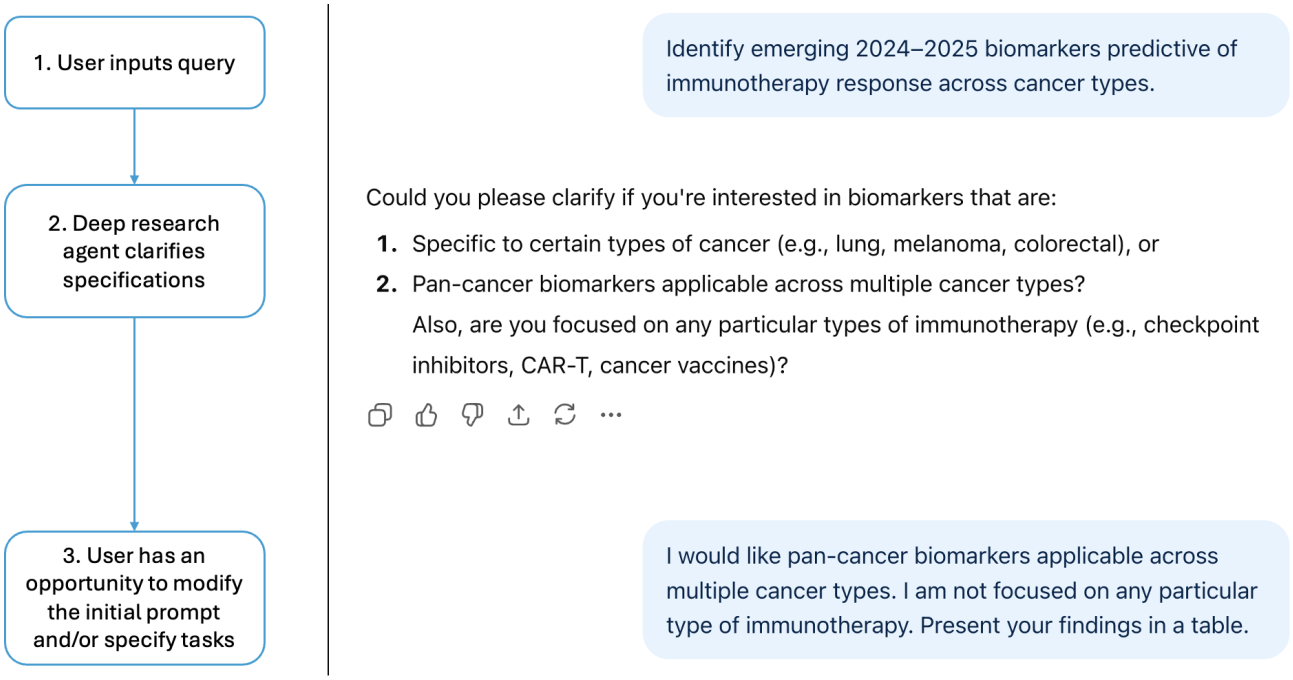

XSL•FO
RenderX

Figure 3. Example of a structured evidence table generated by a deep research agent in response to the earlier query, "identify emerging 2024 - 2025 biomarkers predictive of immunotherapy response across cancer types." The agent autonomously retrieved, synthesized, and formatted recent literature into a reference-supported summary in less than 8 minutes. In this instance, all Deep Research–generated content and citations were independently checked against the original publications and were found to accurately reflect the source literature, with no hallucinated references. This figure, therefore, illustrates a successful deep research workflow. However, polished outputs of this kind should not be assumed to be uniform across tasks or systems. The superscript annotations highlight illustrative areas where known risks may arise: †potential inconsistencies in citation fidelity (eg, incorrect or semantically mismatched source attribution) and ‡possible overstatement or imprecise reporting of quantitative effects. These annotations are hypothetical and do not imply errors in the example shown but are included to emphasize the need for verification despite apparent surface reliability. Citations in this figure are shown in Multimedia Appendix 1. AUC: area under the curve; CTRAM: class I–restricted T-cell–associated molecule; ICI: immune checkpoint inhibitor; MSI: microsatellite instability; OS: overall survival; PD-L1: programmed death-ligand 1; PFS: progression-free survival; TMB: tumor mutational burden.

Biomarker	Type	Immune Mechanism/Pathway	Cancer Types Studied	Key Findings	Source
CRTAM expression	Transcriptomic (T cell/NK cell marker)	Activation/differentiation of cytotoxic and regulatory T cells (CRTAM is a transmembrane protein on T/NK cells)	Pan-cancer (diverse solid tumors; highest expression in melanoma and small intestine)	High tumor CRTAM transcript levels were linked to significantly longer PFS and OS on ICI therapy. In a 514-patient pan-cancer cohort, "high" CRTAM expression (top ≥25%) predicted better response to checkpoint blockade (e.g. HR≈0.60 for PFS, HR≈0.39 for OS)‡ even after adjusting for PD-L1, MSI, and TMB[A]†. This association was not seen in patients who did not receive ICI, indicating CRTAM's predictive value.	[A]
Third Dimensional Stromal (TDS) score	Transcriptomic (stromal gene signature)	Stromal-rich tumor ecosystem (distinct from immune-infiltrated "hot" tumors)	Pan-cancer (multiple solid tumors including lung, urinary, skin, kidney, etc.)	A novel 66-gene "TDS" signature (enriched in stromal cells) was identified. **High TDS** predicted *poor* ICI outcomes: meta-analysis showed OS HR≈1.56 and PFS HR≈1.65 (p<0.0001)‡ for high-TDS tumors[A]†. High-TDS tumors lost the usual benefit of ICI versus chemotherapy, underscoring TDS as a negative predictive biomarker.	[A]
Immunotherapy-Responsive Ecotype Signature (IRE.Sig)	Transcriptomic (single-cell-derived)	Pro-immunogenic tumor "ecosystem" (cell-state composition)	Pan-cancer (validated across multiple cancer types and ICI-treated cohorts)	IRE.Sig is a gene expression signature reflecting a favorable immune "ecotype." In analysis of 34 scRNA-seq datasets and bulk pan-cancer cohorts (11,001 patients, 32 cancer types), IRE.Sig robustly predicted response to ICI. Machine-learning models using IRE.Sig achieved AUC ~0.72 in validation cohorts[B]†, indicating it generalizes across cancers.	[B]

Importantly, although multihop retrieval and iterative self-correction introduce greater autonomy and coverage, the underlying architecture of these agents remains conceptually continuous with established RAG techniques. Their principal innovation lies in the orchestration layer that coordinates planning, retrieval, and tool use, rather than in a fundamentally new reasoning paradigm (eg, no symbolic reasoning or formal logic introduced). These agents may therefore represent a stepwise functional improvement in autonomy but are only an incremental evolution in architecture.

As such, deep research agents inherit many of the same fundamental limitations as prior RAG-based systems. Current implementations remain limited by their dependence on open-access data, lack of transparent ranking mechanisms, and inability to independently verify information. Consequently, these systems continue to operate as unpredictable black boxes, generating probabilistic responses with limited reasoning transparency. As these constraints are progressively addressed, deep research tools may increasingly contribute to the early stages of systematic reviews and evidence-based clinical support.

Deep Research Agents for Literature Reviews and Knowledge Synthesis

One potential use of deep research agents is drafting medical literature reviews or summaries of evidence. Ong et al [5] evaluated several such tools (Gemini 1.5 Pro Deep Research, Perplexity's Deep Research, and OpenAI's Deep Research) by tasking them with writing a literature review on "oculomics" (ocular biomarkers of systemic health). They found that the OpenAI agent produced the most compelling draft—a logically structured, readable, and informative narrative that nonexpert readers could hardly distinguish from a human-written review. In another study, Jesudason et al [10] reported that OpenAI's Deep Research could generate a multipage referenced text on complex medical topics in minutes, including controversial clinical questions such as antithrombotic therapy for carotid dissection. Meanwhile, Wang et al's [11] DeepEvidence allows deep research to be performed across various biomedical knowledge graphs, achieving substantial performance gains in evidence synthesis across drug discovery and clinical trial development. Such comprehensive reports, complete with citations, suggest deep research agents could save researchers substantial time on initial literature reviews.

Another study benchmarked multiple agents for systematic review title and abstract screening in radiology. They found traditional LLMs showed variable sensitivity but high specificity in classifying abstracts, while deep research agents (OpenAI and Google's Gemini) could autonomously retrieve literature with high precision but moderate recall [12]. Another recent evaluation by Loke et al [13] benchmarked deep research agents (OpenAI, Perplexity, and Gemini) and autonomous agentic systems like Manus AI against a published systematic review and meta-analysis on myopia risk. Although this work shares

XSL•FO
RenderX

partial authorship with the present manuscript, it suggests systems such as OpenAI and Perplexity Deep Research can identify studies meeting inclusion criteria and perform generally accurate qualitative data extraction, indicating utility for certain stages of the review process.

However, these studies also reveal important caveats. Citation accuracy arises as a common concern. Ong et al [5] reported multiple instances of hallucinations or erroneous citations, including references to nonexistent articles and incorrect attribution of fictitious articles to real researchers. In independent tests, Keplinger et al [14] tasked 5 deep research agent models with generating a dermatology literature review. They found OpenAI's Deep Research performed best in reference accuracy—approximately 95% of its citations were identifiable, with approximately 70% completely correct. In contrast, Google Gemini 2.5 Pro Deep Research and Perplexity AI Deep Research frequently produced fabricated citations: roughly 47% to 50% of references from those models had fake authors or titles. Another study similarly found OpenAI's Deep Research and xAI's Grok 3 achieved citation precision exceeding 90%, surpassing Gemini and Perplexity but still clearly lagging behind human reviewers [15]. Together, these findings show that while deep research agents can rapidly synthesize literature, citation reliability remains inconsistent across models and still falls short of human standards.

Citation–claim alignment is another notable problem across the board. Keplinger et al [14] found that over 50% of OpenAI's Deep Research's cited statements contained at least one subtle inaccuracy or misrepresentation of the source. The AI often misinterpreted study results, cited out of context, or hallucinated details not in the cited papers. These citation errors are difficult to detect without thorough cross-checking by an expert, raising concerns that an AI-generated review could unwittingly propagate misinformation. Ong et al [5] similarly observed factual errors in the AI-generated oculomics review that appeared innocuous to nonexperts—for example, claims about nonexistent trials or incorrect statements about regulatory approvals. Meanwhile, Loke et al [13] found that none of the evaluated deep research agents could conduct a genuine meta-analysis or generate accurate, publication-quality figures and tables. For instance, frequent omissions and hallucinations were observed when agents were asked to construct summary tables of included studies.

Besides accuracy, bias represents a key concern. Deep research agents often rely on RAG-modeled retrieval to select what they consider the "best" or most relevant evidence, using vector similarity search, ranking algorithms, and other techniques. However, the opacity of what the model deems to be "optimal" or "most relevant" presents a real challenge to the generation of verifiable, reproducible knowledge in clinical and translational research [16]. Given that English-language publications from Western nations dominate the internet, deep research tools may be more likely to draw from these, which risks reinforcing global inequalities in biomedical research. Without conscious efforts to diversify their input sources, these agents risk perpetuating existing publication biases and overlooking unconventional or minority perspectives in research [17].

In summary, deep research agents can dramatically speed up information gathering or evidence summarizing in medicine, producing text that is well organized and richly referenced. However, the consensus among all studies is that human oversight remains crucial for now.

Deep Research Agents for Patient Education and Clinical Q&A

Another emerging medical application of deep research agents is in patient education and answering clinical questions. By integrating up-to-date information from the web, these agents can potentially serve as advanced, real-time medical chatbots. Gültekin et al [18] compared OpenAI's Deep Research with DeepSeek R1 (a contemporary model with a similar "deep think" web-search capability) for answering common patient questions about anterior cruciate ligament surgery. Notably, OpenAI's Deep Research delivered more comprehensive answers, covering questions in greater depth (completeness score 4.0 vs 3.2 for DeepSeek). DeepSeek's answers, however, were clearer and more accessible to laypeople, written at a significantly lower reading level (8.9 vs 14.2). The authors noted that excessive detail may overwhelm patients, while oversimplification risks omitting critical information. This suggests integrating the strengths of both models (the thoroughness of ChatGPT and the plain-language clarity of DeepSeek) could be ideal for patient education materials. Nevertheless, readability need not be an intrinsic model limitation: deep research systems can ultimately be prompted to adjust language complexity, allowing clinicians to tailor explanations to patient literacy levels.

A similar multimodel evaluation by Yetkin et al [19] queried 11 AI chatbots—including GPT-4 Deep Research, Perplexity's Deep Research mode, and others—with common patient questions about sarcoidosis. The models that integrated RAG (ChatGPT-4.0 Deep Research, Perplexity, and xAI's Grok Deep Search) achieved the highest scores on validated metrics such as DISCERN (measuring the quality of health publications) and Patient Education Materials Assessment Tool (PEMAT; patient education and understandability). These deep research–enabled models significantly outperformed standard chatbots that lacked web access. However, the improved comprehensiveness came again at the expense of readability. The top-performing models delivered information at a very high reading level—Flesch-Kincaid grade more than 16 (college-level difficulty). This reiterates the importance of thoughtful prompting and clinician oversight to align outputs with patient health literacy.

Taken together, these studies show that deep research agents can dramatically accelerate the generation of high-quality patient education content, but their output may not yet be patient-accessible without clinical mediation. In practice, a clinician could use deep research agents to quickly gather accurate medical information for a patient, then refine the output into more digestible terms. Moreover, patients have become increasingly AI literate and can adopt simple prompting strategies (eg, "rewrite this explanation for an eighth-grade reading level" or "explain this using analogies and avoid medical jargon") to request clearer, lay-friendly reformulations of

XSL•FO
RenderX

AI-generated content. However, while prompting can improve readability, it cannot guarantee accuracy. Patients need to be made aware of the potential pitfalls of AI-generated medical information, including the importance of corroborating outputs against trusted sources (eg, National Health Service patient information resources). As time progresses, deep research agents could gradually shift the clinician's role from information provider to information steward: verifying accuracy, simplifying complexity, and ensuring alignment with a patient's health literacy level and emotional needs.

Emerging Medical-Specific Deep Research Agents and Benchmarks

As deep research agents have matured, evaluation has shifted from general retrieval toward complex, multistep reasoning in medical contexts. Medically tailored benchmarks have been introduced alongside the development of domain-specific deep research agents.

For example, MedBrowseComp was introduced in 2025 as the first benchmark for autonomous, multihop biomedical research [20]. It featured over 1000 physician-curated questions that simulate realistic clinical research tasks. Importantly, MedBrowseComp evaluated agents not solely on output accuracy but on their ability to locate evidence across multiple authoritative sources and reconcile fragmented biomedical information. The benchmark revealed that leading proprietary deep research agents, including Open AI's o3, Gemini 2.5 Pro, and Perplexity, could answer only around a quarter of these multistep questions correctly.

Meanwhile, domain-specialized agents such as IQVIA's Med-R1 (an 8B parameter medical reasoning model with an agentic framework) and the open-source MedResearcher-R1 (32B) emerged. These incorporated dense medical knowledge and custom retrieval tools, enabling them to match and even outperform much larger general models on certain biomedical tasks. On MedBrowseComp, MedResearcher-R1 (32B) achieved a new state-of-the-art 27.5% accuracy, surpassing OpenAI's o3 Deep Research (25.5%) and Google's Gemini 2.5 Pro (~25%) [21]. Likewise, IQVIA's Med-R1 Deep Research Agent achieved comparable results to OpenAI's o3 and Google's Gemini 2.5 Pro despite using only an 8B parameter model [22]. These medical-specific systems demonstrate more advanced multihop clinical reasoning and synthesis across specialized databases (eg, PubMed, clinical trial registries, and Food and Drug Administration data). By leveraging domain-specific retrieval engines and expert knowledge, they may be able to synthesize complex chains of biomedical facts that generalist agents may miss. In addition, these specialized deep research agents attain their results with relatively compact model sizes (8B for IQVIA or 32B parameters for MedResearcher-R1). This efficiency suggests promising potential for deployment in computationally constrained health care settings, where smaller, domain-focused AI systems may be integrated.

Current limitations underscore that adapting these techniques to medical contexts is still in its early stages. Tasks requiring beyond 4 to 5 reasoning hops remain largely unsolved, as even state-of-the-art agents saw a sharp drop in accuracy on 5-hop questions. Their performance profile on these benchmarks reinforces that such systems should not yet be considered reliable clinical assistants, but rather exploratory tools requiring substantial human oversight. Moreover, the scope of the MedBrowseComp benchmark is itself constrained: it uses automated answer judging (without full clinical expert review) and focuses on hematology and oncology datasets alone, omitting certain subdomains such as advanced genomics or multimodal cases.

Benefits and Limitations of Current Deep Research Agents

Current studies illustrate an emerging picture of the potential roles of deep research agents in biomedical settings. In summary, these agents are being explored as advanced research assistants and information mediators: to draft literature reviews, summarize evidence on controversies, assist in writing academic papers, educate patients with up-to-date information, and even support clinical decision queries. The common thread is their ability to autonomously comb through and synthesize vast text corpora to produce a structured, referenced answer in a fraction of the time a human might take.

This has already led some to herald a new era of AI-augmented medical research. Editorials have described "AI's deep research revolution" and the "next evolutionary leap" [23], noting that current iterations of LLMs can speed up the summarization of literature and even facilitate comparing study results—such that "for text-based research, such as systematic reviews or meta-analyzes, AI may replace them efficiently" [15]. Scientific publishing may be entering a "post-LLM" era, where AI agents equipped with deep reasoning and retrieval capabilities are reshaping how literature is synthesized and reviewed.

On the other hand, virtually all medical evaluations of these tools so far have included strong warnings (Figure 4). A recurring concern is the integrity of citations and content. While the latest models have largely solved the problem of blatantly fabricated references that plagued earlier LLMs, subtler issues persist—slight errors in bibliographic details or correct references attached to unsupported claims. These are not always obvious, especially to nonexperts or when an AI output looks polished. If clinicians or researchers take AI-generated text at face value, there is a risk of false information seeping into publications or clinical guidance.

XSL•FO
RenderX

Figure 4. Limitations for deep research agents include system-level reliability and safety, human and societal risks, and the lack of real-world clinical validation. AI: artificial intelligence.

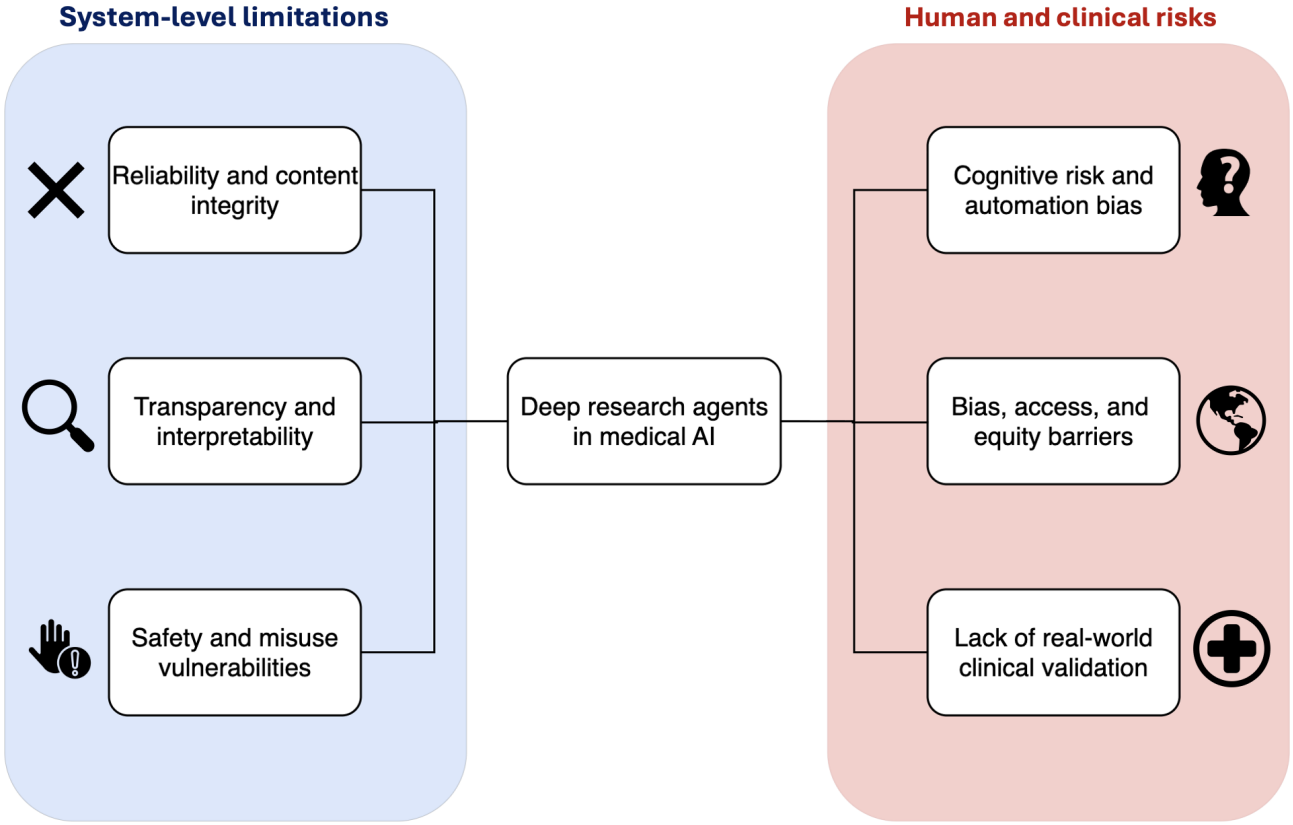

Another concern is the potential for overreliance. Jesudason et al [10] urge that medical trainees should avoid leaning too heavily on AI-prepared syntheses, as they may lose the ability to critically appraise evidence themselves. Cognitive off-loading, where learners delegate lower-level mental tasks to external systems to free up working memory, can support higher-order reasoning and enhance learning if used judiciously [24]. By contrast, automation bias—an overreliance on automated systems for important tasks—arises when users defer to system outputs even when independent judgment is required. In medical training, critical appraisal of studies, evidence weighting, data interpretation, and clinical reasoning represent non-negotiable competencies that should not be off-loaded. Excessive off-loading of these tasks risks fostering not only automation bias (overreliance on automated systems) but also deskilling (loss of previously acquired competencies), never-skilling (failure to develop essential expertise), and mis-skilling (reinforcement of incorrect reasoning due to AI errors or bias) [25]. Over time, excessive reliance on predigested syntheses may erode clinicians' capacity for deep critical thinking, analogous to early concerns about the erosion of manual skills with surgical robotics. In the same vein, Keplinger et al [14] expressed concerns that outsourcing the initial draft of a review to AI could distort authors' own mental map of the field and undermine their engagement with the literature, ultimately harming scientific creativity.

Another key but often overlooked aspect is safety. Recent work has raised broader concerns that increasing autonomy and multistep planning in AI systems introduces qualitatively new safety challenges, particularly around oversight and loss of control [26]. In addition, it has been suggested that deep research agents may circumvent traditional safety filters under certain conditions. Unlike standalone LLMs, which are engineered to refuse unsafe prompts, deep research agents' multistep reasoning pipelines may potentially weaken those refusal mechanisms. Chen et al [27] note that deep research agents may process a broader range of query types than standard chat interfaces because the research pipeline is treated as an information-gathering task rather than a direct response to a prompt. This could inadvertently allow malicious or high-risk instructions—such as instructions related to biological weapons development, hacking, or privacy breaches—to be pursued under the guise of research. This does not mean safety filters are obsolete or permanently bypassed in research-oriented systems. Rather, they highlight a current tension between autonomous information-gathering workflows and content moderation mechanisms that were originally designed for short "chatbot" conversational use. As vendors continue to iterate on agentic architectures and develop guardrails specific to these agents, such safety vulnerabilities may be mitigated.

Importantly, ethical and practical limitations exist. Deep research agents are constrained to open-access information sources, as they cannot directly retrieve paywalled journal content, which could bias them toward certain sources. Their outputs also reflect the data they find—if the online content is skewed or outdated, the AI's output will inevitably be biased as well. Moreover, access to deep research agents themselves is likely to be uneven. Owing to their computational intensity, such systems are expected to be deployed through subscription-based or premium access models. This introduces an economic barrier that may limit adoption in resource-constrained settings, potentially reinforcing disparities in global medical research. As a result,

XSL•FO
RenderX

agentic research workflows may not only reflect prevailing publication biases but also exacerbate inequities in who is able to perform large-scale, synthesis-driven research. Over time, wider licensing or open-source alternatives may mitigate these effects, but economic access remains an important consideration in the near term. Moreover, the development of smaller, lightweight models with deep research capabilities (eg, IQVIA's Med-R1 8B model) may enable more computationally affordable evidence synthesis in resource-limited settings.

Transparency is another issue: it is not always clear how AI algorithms prioritize which sources to use or what reasoning led to a given conclusion. This opacity can make it hard for users to fully trust the results without verification, especially in medicine, where decisions require provenance and auditability. This is in contrast to evidence retrieval platforms such as OpenEvidence, which demonstrate a more transparent paradigm. OpenEvidence pulls directly from curated biomedical databases (eg, PubMed/Medline, clinical trial registries) and exposes every citation used, allowing users to trace each claim back to the original paper or paragraph [28,29]. In the future, deep research systems could adopt a similar architecture—combining autonomous reasoning with constrained, citation-verifiable retrieval pipelines.

Finally, there remains a notable lack of studies that rigorously evaluate the real-world capabilities of deep research agents within medicine. The existing literature largely consists of perspective pieces and proof-of-concept studies, with few empirical studies providing systematic benchmarking of these agents against established evidence synthesis standards or real-world clinical deployment. This gap underscores the need for methodologically robust research to determine whether deep research agents can reliably replicate or augment real-world medical workflows [30]. Recent efforts have begun to recognize this benchmarking gap. Du et al [31] introduced *DeepResearchBench*, the first comprehensive, multidomain benchmark specifically designed for deep research agents. Other initiatives, such as Rigorous Bench, further highlight how current evaluation schemes for LLMs fail to capture the multistep reasoning, citation integrity, and evidence synthesis capabilities characteristic of deep research systems [32]. Together, these studies emphasize that robust, clinically tailored benchmarks are urgently needed to measure not only the generative fluency of these agents but also their fidelity to source material, transparency of reasoning, and reproducibility of evidence synthesis. Moving forward, such benchmarks should also incorporate clinically relevant outcomes, including time saved, workload reduction, and decision support quality.

A Practical Verification and Transparency Framework for Clinical Use

To support the judicious use of deep research agents in clinical and research settings, we propose a structured evidence verification and transparency framework operating at both the system level and the end-user level.

At the system level, developers and vendors should adopt standards for logging retrieval queries, source rankings, and cross-attention or contribution weights linking generated statements to retrieved passages. While these signals do not provide full interpretability of model reasoning, they could enable partial auditing of multistep workflows by exposing which documents were retrieved and which sources most strongly influenced each synthesized claim [33]. Importantly, these logs should be made requestable and interpretable for clinicians and researchers, recognizing that domain experts (rather than system developers) are best positioned to interpret biomedical relevance and evidence quality. As these signals are not currently exposed by most proprietary systems, they should be regarded as forward-looking policy and design recommendations for model developers rather than tools currently available for clinical end-users.

At the user level, clinicians can follow a three-stage workflow when assessing agent outputs: (1) checking that citations originate from authoritative clinical repositories such as PubMed, Food and Drug Administration, guideline bodies, etc; (2) conducting selective audits of the most clinically actionable statements against original papers or guidelines, ensuring these claims are backed by multiple independent sources; and (3) assessing semantic coherence and task alignment to confirm the output remains focused on the specified clinical scope and does not exhibit internal contradictions or overconfidence in weak evidence. By combining architectural transparency with verification at the clinician level, this framework may structure responsible evaluation and improve confidence in the adoption of deep research agents for biomedical evidence synthesis.

Ultimately, trade-offs between the speed advantages of deep research agents and the time required for rigorous manual verification are likely, particularly in high-stakes clinical use. In practice, the net time savings of these systems may arise less from end-to-end automation than from accelerating early-stage information gathering and narrowing the search space for subsequent human verification.

Conclusion

The evolution of general-purpose LLMs to more specialized research-oriented agents, including deep research agents (eg, Open AI's Deep Research, Google's Gemini Deep Research, and Perplexity's Deep Research), marks significant progress in how medical knowledge can be accessed and synthesized. However, while they may represent a stepwise functional improvement in workflow automation, they remain an incremental extension of existing reasoning architectures and inherit many of their limitations. Current evaluations highlight the need for careful, evidence-based integration into medical practice. Their outputs continue to require human scrutiny to ensure fidelity, as they can miscite, misrepresent, or overgeneralize evidence in ways that are not immediately obvious. In medicine, these tools should complement, but not replace, rigorous human-driven research and review processes.

From a practical perspective, the idea of a "deep research" revolution may already be here, and if used judiciously, it can empower health care professionals—speeding up literature analysis and freeing up human experts to focus on critical thinking, creative insights, and patient-centered

XSL•FO
RenderX

decision-making. We believe the most sustainable path forward lies in collaboration with these agents, grounded in transparency and clinical accountability.

Acknowledgments

MYHW is supported by a pre-employment grant from the Ministry of Health Holdings, Singapore. AYO is supported by a National Institute for Health Research (NIHR)—Moorfields Eye Charity (MEC) Doctoral Fellowship (NIHR303691). DAM is supported by a EURETINA Retinal Medicine Clinical Research Grant. PAK is supported by a UK Research & Innovation Future Leaders Fellowship (MR/T019050/1) and The Rubin Foundation Charitable Trust. The views expressed in this publication are those of the authors and not necessarily those of the National Health Service, the NIHR, or the Department of Health and Social Care.

Funding

The authors declared no financial support was received for this work.

Disclaimer

ChatGPT 4o and o1 (OpenAI) were used to conduct the experiments discussed in this article. Generative artificial intelligence was not used in the writing of this Viewpoint.

Data Availability

The data that support the findings of this study are included in this published article and its supplementary information file.

Authors' Contributions

Conceptualization: MYHW, PAK
Methodology: MYHW
Data curation: MYHW, AYO, DAM
Supervision: PAK
Visualization: MYHW
Writing – original draft: MYHW
Writing – review & editing: MYHW, AYO, DAM, PAK

Conflicts of Interest

PAK is a cofounder of Cascader Ltd and has acted as a consultant for Retina Consultants of America, Roche, Boehringer Ingelheim, and Bitfount and is an equity owner in Big Picture Medical. He has received speaker fees from Zeiss, Thea, Apellis, and Roche. He has received travel support from Bayer and Roche. He has attended advisory boards for Topcon, Bayer, Boehringer Ingelheim, and Roche. The remaining authors do not have any conflicts of interest to declare.

Multimedia Appendix 1
Raw chat logs with Open AI's Deep Research.
[DOCX File, 483 KB - jmir_v28i1e88195_app1.docx]

References

1. Yu E, Chu X, Zhang W, et al. Large language models in medicine: applications, challenges, and future directions. Int J Med Sci 2025;22(11):2792-2801. [doi: 10.7150/ijms.111780] [Medline: 40520893]
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
3. Brown A, Roman M, Devereux B. A systematic literature review of retrieval-augmented generation: techniques, metrics, and challenges. Big Data Cogn Comput 2025;9(12):320. [doi: 10.3390/bdcc9120320]
4. Jones N. OpenAI's "Deep Research" tool: is it useful for scientists? Nature New Biol 2025 Feb 6. [doi: 10.1038/d41586-025-00377-9] [Medline: 39915598]
5. Ong AY, Merle DA, Wagner SK, Keane PA. Exploring the dilemma of AI use in medical research and knowledge synthesis: a perspective on deep research tools. J Med Internet Res 2025 Jul 15;27:e75666. [doi: 10.2196/75666] [Medline: 40663724]
6. Introducing deep research. OpenAI. URL: https://openai.com/index/introducing-deep-research/ [accessed 2026-01-17]
7. Zhang W, Li X, Zhang Y, et al. Deep research: a survey of autonomous research agents. arXiv. Preprint posted online on Aug 18, 2025. [doi: 10.48550/arXiv.2508.12752]
8. Huang Y, Chen Y, Zhang H, et al. Deep research agents: a systematic examination and roadmap. arXiv. Preprint posted online on Jun 22, 2025. [doi: 10.48550/arXiv.2506.18096]
9. Gemini deep research. Gemini. URL: https://gemini.google/overview/deep-research/ [accessed 2026-01-17]

XSL•FO
RenderX

10. Jesudason D, Gao C, Seth I, Chan WO, Bacchi S. OpenAI's 'Deep Research' for the generation of comprehensive referenced medical text: uses and cautions. ANZ J Surg 2025 Nov;95(11):2250-2252. [doi: 10.1111/ans.70161] [Medline: 40325519]
11. Wang Z, Chen Z, Yang Z, et al. DeepEvidence: empowering biomedical discovery with deep knowledge graph research. arXiv. Preprint posted online on Dec 23, 2025. [doi: 10.48550/arXiv.2601.11560]
12. Dogra S, Arabshahi S, Wei J, et al. Evaluating large language models for radiology systematic review title and abstract screening. Acad Radiol 2025 Dec;32(12):7023-7034. [doi: 10.1016/j.acra.2025.08.014] [Medline: 40849232]
13. Loke WT, Srinivasan S, Yang GD, et al. Can “Deep Research” agents and general AI agentic systems autonomously perform systematic review and meta-analysis? Eye (Lond) 2026 Jan;40(1):147-149. [doi: 10.1038/s41433-025-04138-w] [Medline: 41315732]
14. Keplinger LE, Frashure LK, Duran SA, Hu G. Assessment of Deep Research for dermatology literature reviews: deep concern over the hype. J Eur Acad Dermatol Venereol 2025 Sep 4:40904191. [doi: 10.1111/jdv.70035] [Medline: 40904191]
15. Thurzo A, Varga I. Revisiting the role of review articles in the age of AI-agents: integrating AI-reasoning and AI-synthesis reshaping the future of scientific publishing. Bratisl Med J 2025 Apr;126(4):381-393. [doi: 10.1007/s44411-025-00106-8]
16. Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: a survey. arXiv. Preprint posted online on Dec 18, 2024. [doi: 10.48550/arXiv.2312.10997]
17. Yang R, Wong MYH, Li H, et al. Retrieval-augmented generation in medicine: a scoping review of technical implementations, clinical applications, and ethical considerations. arXiv. Preprint posted online on Nov 8, 2025. [doi: 10.48550/arXiv.2511.05901]
18. Gültekin O, Inoue J, Yilmaz B, et al. Evaluating DeepResearch and DeepThink in anterior cruciate ligament surgery patient education: ChatGPT-4o excels in comprehensiveness, DeepSeek R1 leads in clarity and readability of orthopaedic information. Knee Surg Sports Traumatol Arthrosc 2025 Aug;33(8):3025-3031. [doi: 10.1002/ksa.12711] [Medline: 40450565]
19. Yetkin NA, Baran B, Rabahoğlu B, Tutar N, Gülmez İ. Evaluating the reliability and quality of sarcoidosis-related information provided by AI chatbots. Healthcare (Basel) 2025 Jun 5;13(11):1344. [doi: 10.3390/healthcare13111344] [Medline: 40508957]
20. Chen S, Moreira P, Xiao Y, et al. MedBrowseComp: benchmarking medical deep research and computer use. 2025 Presented at: 39th Conference on Neural Information Processing Systems (NeurIPS 2025); Dec 6-7, 2025 URL: https://openreview.net/pdf?id=pSDvq2ZgPk [accessed 2026-03-11]
21. Yu A, Yao L, Liu J, et al. MedResearcher-R1: expert-level medical deep researcher via a knowledge-informed trajectory synthesis framework. . Preprint posted online on Aug 20, 2025. [doi: 10.48550/arXiv.2508.14880]
22. Introducing IQVIA med-R1 deep research agent. IQVIA. URL: https://www.iqvia.com/blogs/2025/12/introducing-iqvia-med-r1-deep-research-agent [accessed 2026-01-18]
23. Wang YM, Chen TJ. AI’s deep research revolution: transforming biomedical literature analysis. J Chin Med Assoc 2025;88(6):415-416. [doi: 10.1097/JCMA.0000000000001231] [Medline: 40128176]
24. Chirayath G, Premamalini K, Joseph J. Cognitive offloading or cognitive overload? How AI alters the mental architecture of coping. Front Psychol 2025;16:1699320. [doi: 10.3389/fpsyg.2025.1699320] [Medline: 41356052]
25. Abdulnour REE, Gin B, Boscardin CK. Educational strategies for clinical supervision of artificial intelligence use. N Engl J Med 2025 Aug 21;393(8):786-797. [doi: 10.1056/NEJMra2503232] [Medline: 40834302]
26. Bengio Y, Cohen M, Fornasiere D, et al. Superintelligent agents pose catastrophic risks: can scientist AI offer a safer path? Superintell Robot Saf Align 2025;2(5). [doi: 10.70777/si.v2i5.15569]
27. Chen S, Li Z, Han Z, et al. Deep research brings deeper harm. arXiv. Preprint posted online on Oct 13, 2025. [doi: 10.48550/arXiv.2510.11851]
28. Hurt RT, Stephenson CR, Gilman EA, et al. The use of an artificial intelligence platform OpenEvidence to augment clinical decision-making for primary care physicians. J Prim Care Community Health 2025;16:21501319251332215. [doi: 10.1177/21501319251332215] [Medline: 40238861]
29. Jagarapu J, Babata K, Chamarthi S, Hoyt R. The accuracy and repeatability of OpenEvidence on complex medical subspecialty scenarios: a pilot study. medRxiv. Preprint posted online on Dec 4, 2025. [doi: 10.64898/2025.11.29.25341091]
30. Moons P, Dou B, Desmedt CP. Google’s AI co-scientist and OpenAI’s deep research: new partners in health research? Eur J Cardiovasc Nurs 2025 Jul 21;24(5):800-807. [doi: 10.1093/eurjcn/zvaf066] [Medline: 40678974]
31. Du M, Xu B, Zhu C, Wang X, Mao Z. DeepResearch bench: a comprehensive benchmark for deep research agents. arXiv. Preprint posted online on Jun 13, 2025. [doi: 10.48550/arXiv.2506.11763]
32. Yao Y, Wang Y, Zhang Y, et al. A rigorous benchmark with multidimensional evaluation for deep research agents: from answers to reports. arXiv. Preprint posted online on Oct 2, 2025. [doi: 10.48550/arXiv.2510.02190]
33. Zhu Y. Applications of attention mechanisms in explainable machine learning. Acad J Comput Inf Sci 2025;8(10). [doi: 10.25236/AJCIS.2025.081008]

Abbreviations

AI: artificial intelligence
LLM: large language model

XSL•FO
RenderX

PEMAT: Patient Education Materials Assessment Tool
RAG: retrieval-augmented generation

Edited by A Coristine; submitted 21.Nov.2025; peer-reviewed by KM Sajjadul Islam, K Chen, S Saini; revised version received 25.Feb.2026; accepted 25.Feb.2026; published 26.Mar.2026.

Please cite as:
Wong MYH, Ong AY, Merle DA, Keane PA
Deep Research Agents: Major Breakthrough or Incremental Progress for Medical AI?
J Med Internet Res 2026;28:e88195
URL: https://www.jmir.org/2026/1/e88195
doi:10.2196/88195

© Matthew Yu Heng Wong, Ariel Yuhan Ong, David A Merle, Pearse A Keane. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

From Knowledge Graphs to Digital Twins: Perspectives on Modeling Patient Outcomes for Health Care Quality Assessment

Anna-Katharina Nitschke[1*], MSc; Juan G Diaz Ochoa[2*], Dr rer nat; Simone Neumaier[3], Dr med; Markus Knott[4], MSc

[1]Department of Physics, Heidelberg University, Heidelberg, Germany

[2]PerMediQ, Pelargusstraße 2, Stuttgart, Germany

[3]Institut für klinische und experimentelle Transfusionsmedizin, Tübingen, Germany

[4]Klinikum Stuttgart, Stuttgart Cancer Center, Stuttgart, Germany

[*]these authors contributed equally

Corresponding Author:
Juan G Diaz Ochoa, Dr rer nat
PerMediQ, Pelargusstraße 2, Stuttgart, Germany

Abstract

Medical applications of mathematical modeling, including machine learning models, knowledge graphs, and health digital twins, primarily involve the prediction of patient outcomes. This expert perspective examines how mathematical modeling can contribute to health care quality management. Definitions of procedures, patient outcomes, and quality metrics are provided with a quantitative focus. The emphasis is subsequently placed on 3 categories of patient-centered quality of care, namely, patient safety, procedure accuracy, and procedure efficacy, for which a conceptual and mathematical description is provided. Different levels of modeling tasks essential for managing patient-centered quality of care are identified. This article facilitates a deeper understanding of the topic by assigning relevant publications to these 3 quality categories. Focus is placed on the applicability of graph-based methods, including knowledge graphs and health digital twins, to improve quality management in health care. We have presented a clinical scenario and provided information on methodological limitations, future research directions, and practical implications.

(J Med Internet Res 2026;28:e81946) doi:10.2196/81946

KEYWORDS

quality management; medicine; patient outcomes; machine learning; knowledge graphs; digital twins; human digital shadows

Introduction

In recent years, considerable efforts have been made to improve patient outcomes by standardizing medical care across health care systems for patients [1]. This has been achieved in particular through the creation and application of evidence-based medical guidelines [2]. Here, standardization (ie, the design of medical procedures via guidelines aimed at the systematic integration of scientific evidence) plays a relevant role. This is achieved by selecting the most appropriate health decisions required during the patient's journey. Standardizing patient care is expected to improve cost and workflow efficiency, as well as resource allocation [3].

Many forms of medical interventions (eg, the selection of a therapy) can be seen as part of well-defined processes (ie, a series of actions taken to achieve an outcome) [4]. This concept intersects with the industry, where processes refer to a set of interconnected tasks that convert input into specific outputs. Therefore, modern medicine can be seen as a set of planned processes recommended by guidelines. This overlap has led to industrialization in medicine, with both positive (evidence-based treatments) and negative aspects (eg, dehumanization of medicine and focus on economic aspects) [3].

Furthermore, process-oriented health care focuses on process optimization and economic optimization instead of patient care (for a review of the industrialization of health care and its criticisms, see the chapter by Da Silva [5]). On the other hand, a patient-centered and process-oriented perspective can lead to positive aspects, such as breaking silos in clinical data, optimizing outcomes (ie, patient recovery after therapy), minimizing medical errors and expenses for diagnosis and treatment, and maximizing clinical outcomes [6]. In such cases, the balance between standardization and customization is a paramount problem [3].

Records containing medical interventions generated when a patient receives medical treatment (eg, electronic health records [EHRs]) are essentially defined as processes with different steps, including diagnosis, medical intervention (such as medication, surgery, etc), and tracking of a patient's condition to prevent further health problems [7]. Thus, it is assumed that a patient's process can be modeled in a similar way to the dynamics of a productive industrial process. Through the logical integration of different steps, both the well-being of the patient and the economic performance of the institution (eg, hospital) can be optimized.

XSL•FO
RenderX

Mathematical modeling can significantly improve the way medical processes are implemented and meet quality standards. However, medical data are often not used beyond direct patient care. The reasons for this are manifold and include a lack of digitalization, inadequate structured health data, stringent data protection laws, insufficient technical infrastructure, and limited personnel or financial capacities. Health care providers often use medical jargon and classification schemes that are country-specific and even center-specific (an exception is the use of the *International Classification of Diseases* [*ICD*] for disease classification), and information contained in the form of unstructured data is often not machine-readable.

In addition, health care practitioners (HCPs) are still not sufficiently trained in machine learning (ML) technologies. Moreover, IT specialists rarely understand clinical processes in health care facilities. In these dynamic times, it is crucial to bring different disciplines together if ML applications are to provide real benefits in everyday clinical practice [8]. Assuming that notions of causality can be identified in medicine and used to identify mechanisms and derive predictive models [9,10], ML applications can provide information about patient outcomes, as well as unexpected or undesired events, and support medical decision-making.

This article presents a novel perspective on modeling patient outcomes to improve health care quality assessment. The applicability of mathematical models (focusing on graph-based methods and health digital twins [HDTs]) is considered within a quantitative concept of health care quality management, considering 3 categories of patient-centered quality of care (PCQC), namely, patient safety, procedure accuracy, and procedure efficacy, thereby providing a consistent taxonomy. To date, this relevant topic has been explored only to a limited extent. This work aims to contribute to improving health and well-being (according to the UN sustainability goals, SDG3) [11]. We discuss the potential, limitations, and practical implications of the application of these technologies for modeling PCQC.

Definition of Patient Outcomes for Quantitative Quality Assessment

Overview

In medicine, it is generally difficult to define patient outcomes quantitatively (related to measurable information: exact metrics and observables captured by sensors) and qualitatively (related to nonmeasurable descriptions: categories) because patients are heterogeneous and complex organisms, and an overall assessment of their health status therefore remains a challenge.

Our goal is to identify these variables of interest and place them in a broader understanding of quality assessment, paving the way for mathematical modeling in this field. To this end, we aim to introduce a description of patient outcomes as a fundamental concept for evaluating quality indicators (QIs). These indicators are relevant for decision-making because they provide an estimate for which option holds the greatest desirability or value [12], that is, improvement of the quality of care (eg, improvement of patient well-being, avoidance of patient death, optimal use of resources, etc).

Patient Outcomes

From a medical perspective, a patient outcome is defined in terms of what is meaningful and valuable to the individual patient [13]. However, this definition is very general and difficult to relate to a quantifiable metric. There is still a conceptual problem with the general definition of results in the medical context, such as patient-relevant outcomes (morbidity, mortality, and quality of life) and surrogate or biomarker outcomes [14]. These outcomes can differ significantly depending on the underlying disease. Liu et al [15] summarized the definition of patient outcomes related to clinical practice "as any change [within the patient's health status] that results from health care [for which each profession] has developed outcome measures that focus on the standards, activities, and impact of its discipline."

Thus, patient outcomes are manifold and range from PCQC [14] to institutional performance [16] and to patient-reported outcomes [17], which can be assessed from different data sources (for a comprehensive overview, see information provided by Busse et al [18]).

In this perspective article, we focus only on the description of PCQC (Table 1), for which a comprehensive review was published by Kersting et al [14]. We aim to provide a reference for the general concept, which is why we generally refer to medical procedures as follows: A *medical procedure* is the overarching phrasing for any type of observation, measurement, diagnosis, or treatment performed on the patient.

However, not all patient outcomes can be measured objectively, and only a limited number can be translated into clear, quantifiable metrics. In other words, a mathematical model can reflect only a fraction of the clinical reality.

XSL•FO
RenderX

Table . Subdivision of patient outcomes into 3 areas (patient-centered quality of care, institutional performance, and patient-reported outcomes).

Variable[a]	Patient-centered quality of care	Institutional performance	Patient-reported outcomes
Definition	Patient-centered QIs[b] in health care [19] measured or captured by HCPs[c] [14].	Process execution and resource consumption on an institutional statistical evaluation level [16].	Patient's subjective evaluation of health experience. Report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else [17].
Exemplary categories	Medical evaluation of: • Patient safety • Procedure accuracy • Procedure efficacy	Financial evaluation of: • Patient waiting time • Economic efficiency of medical procedures • Productivity	Questionnaires evaluating: • Symptoms • Quality of life • Health care experience

[a]For each area, a definition and a further subdivision into categories are given to enable a quantitative description of the different areas.

[b]QI: quality indicator.

[c]HCP: health care practitioner.

In the context of clinical trials, great efforts have been made to establish quantifiable patient outcomes. Clinical trials use predefined endpoints to measure patient outcomes. Typical endpoints are patient survival at a given time, incidence of clinical events (such as stroke), clinical performance measures, or patient-reported outcomes [20].

Considering the impact that EHRs have on the quality of care, we provide a definition of patient outcomes on the basis of changes registered in the health record and defined as events [21] (assuming that the data relevant for the definition of patient outcomes are machine-readable). We refer to *patient outcomes* as events registered in the health records. Such events can include changes in a patient's health state or changes in medical procedures.

These outcomes enable health care providers and organizations to measure the impact of their services on patient well-being, identify key areas for improvement, and enhance the overall quality of care.

Medical Guidelines

Medical guidelines are systematically developed statements that reflect the current state of medical knowledge. Medical experts prepare these guidelines in cooperation with professional associations. These evidence-based recommendations are based on clinical studies published in scientific literature and/or other existing evidence informed by expert experience and consensus.

The extracted information is then assessed and published by organizations, such as the World Health Organization, for international guidance or by country-specific professional societies (in Germany, the *Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V.* [22]). Guidelines are not a static corpus and are continuously adapted and further developed on the basis of current clinical studies and scientific findings. They intend to support the decision-making of HCPs in order to enable appropriate care for certain health problems. Medical guidelines consist of complex workflows in mostly narrative form and can also contain multiple, equivalent treatment alternatives for the same disease unit. In general, comparing performed procedures with guidelines is challenging, especially considering that guideline recommendations should not be understood as rigid limits. In medical guidelines, a distinction is made between several grades of recommendations ("can" or "should") on the basis of available evidence (see, for example, the AWMF register for prostate carcinoma [in German] [23]). Additionally, medical guidelines are country-specific and may differ across national states and regions.

Information on QIs

To assess in part how guidelines are applied within an institution and to improve the quality of care, several QIs have been developed by different institutions. These QIs serve as internal quality management tools for medical institutions, enable benchmarking with other institutions, and aim to improve the quality of care for patients [24]. In Germany and Switzerland, routine billing data from hospitals are used for this purpose, if a health care facility decides to participate voluntarily in such initiatives [25]. In Germany, QIs for cancer treatment are defined annually by the German Cancer Society (*Deutsche Krebsgesellschaft* [DKG]). Hospitals that meet the criteria defined by the DKG are eligible to receive DKG certification. The outcomes of cancer patients treated in these DKG-certified centers are superior in terms of patient survival [26]. However, these QIs are assessed retrospectively on an institutional basis and cover aspects of the application of medical guidelines.

Quality Metrics for PCQC

Mathematical modeling is used to represent a real-world phenomenon by capturing and analyzing the most significant relationships between different observables through formalized relationships between mathematical parameters. For the quantitative description of patient outcomes, they need to be described by numerically or categorically measurable parameters. These parameters capture various relevant aspects of the overall performance of patient-centered care quality. These numerical and categorical parameters are referred to as quality metrics (QMs) because they enable an evaluation of patient care. There is a hierarchical ordering that assigns each QI to several QMs, which might be defined in a disease- or context-specific way, to place them into the perspective of

XSL•FO
RenderX

mathematical modeling. We refer to *QM* as a quantitative measure of QIs that relates to adherence to medical guidelines (procedural accuracy); QM also refers to patient-centered care (patient safety and effectiveness of the procedure).

Information on PCQC

Overview

We present an overview of the broad concept of PCQC. According to the classification provided in Table 1, we define in detail the 3 basic categories used for the further discussion of health care quality assessment: patient safety, procedure accuracy, and procedure efficacy.

For PCQC, there is a strong reference to medical guidelines and appropriate clinical practice. Owing to the increasing complexity of medical knowledge reflected within medical guidelines, it is becoming increasingly difficult for health care professionals to keep track of changes. The application of mathematical modeling is intended to help health care professionals comply with the guidelines (ie, a sequence of recommended procedures and support for medical decision-making). Medical practitioners select their decisions on the basis of the probability of initiating a change in the patient's health state (assuming that physicians are either using heuristics or making rational decisions on the basis of available information [27]). Use cases for the application of mathematical modeling for PCQC are presented in Multimedia Appendix 1.

Mathematical modeling can be included for predicting QMs to enhance health care quality. In this context, the patient's journey can be understood as a complex chain of procedures that can be logically interconnected. Therefore, it is imperative to look at the whole patient journey as a process in which each procedure is just a point in time (t_{i-1}, t_{i+1}). As illustrated in Figure 1, the general way in which health care quality is assessed can be described as follows: collecting data from a patient (element 1), which a clinician uses to make a medical decision regarding the patient's further procedure (element 2) and whose impact is measured after the procedure is carried out by looking at the patient outcome (element 3), which allows assessment of the quality of the health care provided (element 4). One can distinguish between two methods of implementation: (1) use the prediction in a prospective way (dashed arrow; meaning that before the medical decision is made, the clinician receives feedback from the model and can adjust the decision, leading to potentially different outcomes and improved health care quality) and (2) use a retrospective inclusion (dotted arrow), allowing an evaluation of the observed patient outcome and an assessment of the quality of care.

Another distinguishing feature of mathematical modeling is whether the model works for individual patients or for patient cohorts. Hence, for modeling the QM, we can consider the following representation levels:

1. Patient-cohort predictions for quality management of patient populations: It involves statistical modeling of the patient population for controlling domestic and global medical practices. At this level, mathematical models describe an "average and black box virtual patient" that tests or corroborates published and domestic guidelines.
2. Patient-centered predictions: It involves modeling of specific targets and the inclusion of personal information in gray (partial knowledge of model details) or white-box (full knowledge of model details) models to model "virtual individual patients" and evaluate published or domestic guidelines at an individual level. This implies that models should reflect and predict individual deviations from the expected results of the main patient cohort (see, for example, information provided by Peng et al [28]).

The patient journey can be represented as a chain of observed patient outcomes (for example, the main disease and corresponding codiseases) and possible medical decisions that lead again to distinct patient outcomes over time that are patient journey dependent. Medical guidelines specify possible medical decisions following patient outcomes. The quality of the health care received during the patient journey can be assessed by comparing different elements within this decision-tree structure. Three categories of PCQC can be distinguished: patient safety, procedure accuracy, and procedure efficacy. In the following sections, the 3 different categories of modeled patient outcomes are defined by describing their context to the patient journey in a qualitative and quantitative way, as well as identifying suitable QMs.

XSL•FO
RenderX

Figure 1. Visualization of key parameters for health care quality management and mathematical model integration throughout the patient's journey over time. The predicted quality metric of the mathematical model can be implemented in 2 ways: prospective (dashed line) or retrospective (dotted line). It includes the following elements: (1) data collection from the patient, (2) medical decision made by a clinician, (3) patient outcome after the procedure, (4) quality assessment, (5) quality metric prediction from the mathematical model, and (6) data storage. This can analogously be performed at the population level (patient-cohort predictions) instead of at the patient level (patient-centered predictions).

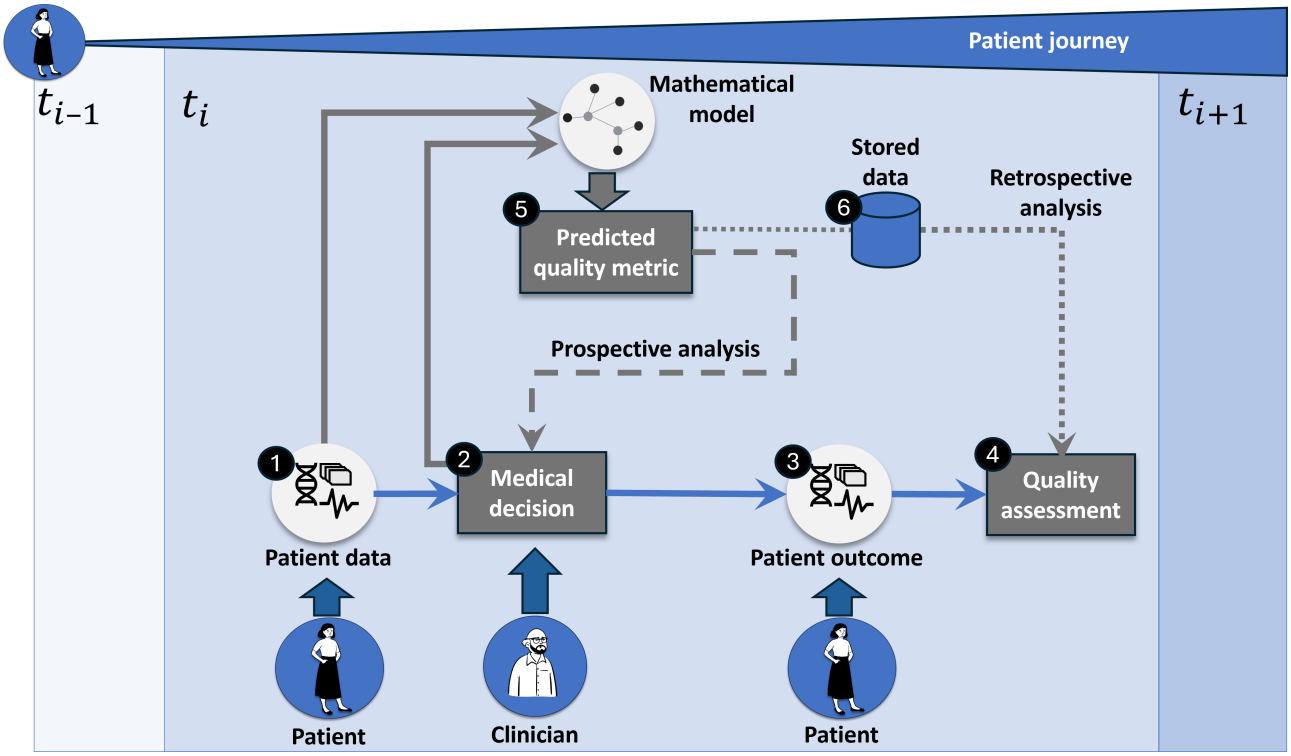

Patient Safety

Providing an exact definition of patient safety is challenging, as this term remains under debate in the literature [29]. Patient safety (related to the taxonomy introduced in Table 1) can be measured by patient safety indicators, which refer mostly to the avoidance of accidental and unwanted events in health care. Patient safety is partially measured within health care facilities via predefined inpatient QIs. Patient safety indicators and inpatient QIs were developed by the Agency for Healthcare Research and Quality (AHRQ [30]), and together, they provide a comprehensive picture of QIs for patient safety. These QIs are mostly determined from the billing data of the hospitals and the medical documentation data contained therein (*ICD* codes for diagnoses and International Classification of Health Interventions [ICHI] codes [31,32] for medical procedures, although each country can have its own classification schema). Possible examples of QIs that reflect patient safety can be extracted from the AHRQ [30], such as complication-related indicators, procedure-related errors, medical-related errors, and hospital system indicators.

There are a variety of possible patient outcomes, some of which are desirable and some of which are adverse (an adverse outcome is evaluated depending on the specific clinical context). Assuming that physicians are rational decision makers, the medical decision is an evaluation of the risk of previous and future outcomes (according to the classic *Theory of Rational Decision Making*; see, for example, information provided previously [33]). A mathematical model for patient safety can provide support through checking (retrospective) or predicting (prospective) how likely the outcome of a medical decision will be an element of the set of adverse outcomes.

QMs can be related to the respective QIs for the quantitative representation of patient safety. Examples can be defined as follows: (1) Mortality rate and failure to rescue, which have emerged as important QMs [34]; (2) Emergency readmission to the hospital after a procedure among outpatients, which can happen, for instance, if prophylaxis has been forgotten or if safety guidelines have not been properly followed; (3) Number of reoperations or complications shortly after the operation, which might indicate a need for investigation within the facility; and (4) Admission to intensive care units (ICUs) and length of stay (LOS) in the ICU among inpatients (different from the intermediate medical care unit).

Table 2 depicts the relationships of the example QMs to the respective QIs, the patient status, the data type, and the QM type. The "Combination event with a time period" column indicates that the observed metric is provided in conjunction with a time parameter.

XSL•FO
RenderX

Table . QIs[a] and QMs[b], data type, correlation involving patient status, and combination event with a time period for patient safety.

Patient safety - QM	QI	QM kind	Data type	Correlation involving patient status	Combination event with a time period
Failure to rescue	Patient safety	Disease specific	Binary	ICHI[c] with mortality	Yes (eg, within 30 days)
Hospital readmission	Hospital readmission	Context specific	Binary	—[d]	Yes (eg, within 30 days)
Number of reoperations or complications	Patient safety	Context specific	Numeric	—	Yes (eg, within 90 days or 1 year)
Admission to ICU[e] and ICU-LOS[f]	Patient safety	Context specific	Binary	ICHI with critical care	Yes (eg, within 48 - 72 hours)

[a]QI: quality indicator.

[b]QM: quality metric.

[c]ICHI: International Classification of Health Interventions.

[d]Not applicable.

[e]ICU: intensive care unit.

[f]ICU-LOS: length of stay in the intensive care unit.

Procedure Accuracy

The purpose of medical guidelines is to reduce errors and disparities in medical care while supporting best practices and responsibilities in medicine on the basis of the current state of scientific knowledge. Following established medical guidelines increases the probability of achieving better clinical outcomes [35-37]. Therefore, an appropriate measure of PCQC (related to the taxonomy introduced in Table 1) is the adherence of HCPs to medical guidelines, including the correct and accurate application of a medical procedure, which we refer to as procedure accuracy. To date, there is no structured or automated assessment of whether recommended medical guidelines have been followed by HCPs for an individual patient.

For the inclusion of mathematical modeling for health care quality assessment, it is necessary to identify metrics that can be defined precisely (if possible, with binary or clear multilevel indicators). In most cases, these metrics cannot be defined universally and require disease-specific development. Thus, with respect to procedure accuracy, mathematical modeling can provide useful insights as follows:

- Anomaly detection and mathematical models can help discover common patterns in the relationship between diagnoses and medical procedures [38] or in the retrospective analysis of clinical narratives via natural language processing [39]. Such pattern recognition can be used to discover for which patient the pattern deviates from the guidelines and if such anomalies are an indication of a patient-centered procedure [40].
- Local guidelines (which are not merely deviations from national standards but adaptive tools shaped by real-world experience) can be created by hospitals or health care institutions. These include local adjustments to general guidelines when differences are identified, and the efficacy of alternative methods has been demonstrated [41].
- Patient-centered interpretation guidelines can be created, and the integration of additional data, such as physiological or genetic data, into models can provide recommendations for action that consider specific patient variations. Hence, models can enable the implementation of patient-centered quality management.

Clear results include checking whether the recommendations of a guideline for performing a standardized procedure or a set of procedures have been followed in an individual patient. This information can be retrieved from EHRs, laboratory information management systems, or even reimbursement data. In addition to compliance with a guideline, medical errors can also be reduced. Examples of QMs for procedure accuracy are as follows (Table 3):

- Detection of patient outliers: The average amount of hospitalization required by a medical procedure (LOS in hospital compared with the average) and the appearance of adverse events (number of infections, etc, compared with the average) can be assessed.
- Expected medical procedure: On the basis of the *ICD* diagnosis, specific ICHI procedures can be predicted, and deviations in the ICHI code from the prediction can be detected.
- Revision operation: The number of necessary additional surgical interventions after surgery (eg, due to complications of the initial surgical intervention) can be assessed.

While these metrics are easy to extract at the site level, they risk missing relevant contextual factors. It is desirable to integrate additional qualitative or clinical data points for a more holistic assessment. However, in clinical practice, simpler modeling techniques are used.

XSL•FO
RenderX

Table . QIs[a] and QMs[b], data type, correlation involving patient status, and combination event with a time period for procedure accuracy.

Procedure accuracy - QM	QI	QM kind	Data type[c]	Correlation involving patient status	Combination event with a time period
Detection of patient outliers	Accuracy of procedure application	Context specific	Binary	ICHI[d] (patient) with ICHI (population)	No
Number of deviations of procedures with respect to guidelines (Δ[e] ICHI)	Accuracy in the application of medical guidelines	Context specific	Multilabel	ICHI (patient) with ICHI (population)	No
Revision of operations	Patient-specific management	Disease specific	Binary	—[f]	Yes (eg, within 90 days or 1 year)

[a]QI: quality indicator.

[b]QM: quality metric.

[c]Binary data correspond to 2 possible values (yes/no or 0/1).

[d]ICHI: International Classification of Health Interventions.

[e]The symbol Δ refers to a difference with respect to a reference value.

[f]Not applicable.

Procedure Efficacy

A general definition of efficacy has been provided by Lynch [42] ("efficacy is the capacity [of a medical intervention] to produce an effect"). For quality assessment, procedure efficacy (related to the taxonomy introduced in Table 1) refers to the optimal selection of procedures aiming at achieving the best possible patient outcome.

Retrospective models could compare the medical decisions chosen by HCPs with other possible choices, which may lead to better outcomes. Hence, procedure efficacy can also be considered for monitoring a decision (made under imperfect information) by observing the patient's response or for making a prognosis and a subsequent decision under imperfect information. Owing to its nature, procedure efficacy requires the implementation of mechanisms and causal relationships where target/trial emulation is needed. This type of modeling contrasts with conventional predictive risk modeling tasks (which are applied, for instance, to evaluate patient safety and procedure accuracy).

The following are relevant QMs for procedure efficacy (Table 4):

- Biomarker levels: Predictions and projections of biomarker levels and their relationships with accepted levels to assess and predict a patient's condition (acceptable [ie, considering parameters below a threshold] or deteriorating [ie, parameters above a threshold]) can lead to specific medical decisions. For example, a stable blood glucose range over time indicates the efficacy of diabetes treatment.
- Tumor-free survival: Time (in months or years) from complete tumor resection until tumor recurrence.
- Symptom risk score: This score quantifies, for example, the risk of an improvement or deterioration in a specific medical condition.
- Overall survival: Indicates the time (in months or years) from diagnosis until death from any cause.
- Pathologic complete remission: The proportion of patients showing no residual viable tumor cells on histopathological assessment after the completion of tumor therapy.

XSL•FO
RenderX

Table . QIs[a] and QMs[b], data type, correlation involving patient status, and combination event with a time period for procedure efficacy.

Procedure efficacy - QM	QI	QM kind	Data type[c]	Correlation involving patient status	Combination event with a time period
Biomarker levels	Surrogate for disease control	Disease specific	Time series	Biomarker with ICHI[d]	—[e]
Tumor-free survival	Tumor control	Disease/context specific	Binary	Biomarker/time	Yes (eg, after 1, 3, or 5 years)
Recurrence-free survival	Tumor control	Disease/context specific	Binary	Biomarker	Yes (eg, 12, 24, or 36 months)
Symptom risk scores	Satisfactory patient condition	Disease specific	Binary	—	Yes (eg, directly or within 6 - 12 weeks after treatment)
Overall survival	Tumor control	Context specific	Binary	Biomarker	Yes (eg, after 1, 3, or 5 years)
Pathologic complete remission	Tumor control	Disease specific	Binary	Biomarker	Yes (eg, directly or within 6 - 12 weeks after treatment)

[a]QI: quality indicator.

[b]QM: quality metric.

[c]Binary data correspond to 2 possible values (yes/no or 0/1).

[d]ICHI: International Classification of Health Interventions.

[e]Not applicable.

In legal or patient-communication contexts, terms such as “survivability” are typically used, which refer to an inherent capacity or probability of survival but are not considered valid clinical endpoints. In this study, we refer to well-defined outcome measures such as tumor-free survival.

Mathematical Modeling of PCQC

Overview

A mathematical representation of the patient’s condition to predict the outcome is essential for a precise evaluation of quality management. In summary, patient safety (Table 2) and procedure efficacy (Table 4) are related to clinical endpoints, whereas procedure accuracy (Table 3) is closely related to precision in the implementation of clinical guidelines and in reference to medical evidence. To avoid conceptual conflation in QMs, procedure accuracy is always defined with respect to documents (scientific literature or guidelines) containing medical or scientific evidence. Importantly, some events, such as patient outliers (Table 3), are estimated inside a time horizon and are mapped into binary states. Despite their convenience for mathematical modeling, there are relevant trade-offs, such as information loss or threshold choices, which can be problematic for the final validity of the models.

In the context of the 3 quality management categories, mathematical models need to perform distinct modeling tasks. First, the information coming from the patient needs to be stored and ordered in a data storage system (graph-based structures, such as knowledge graphs [KGs], can be used to encode patient data and information in a systematic manner). On the basis of the available data, single mathematical models (we focus on ML models) are trained to accomplish tasks that can either be classification tasks of the patient state or prediction tasks of the evolution of a parameter describing the patient response. The model predictions can be either prospective or retrospective.

We focus on KGs to integrate information from single events within the patient journey. By analyzing patterns from these KGs extracted via methods, such as graph neural networks (GNNs), the PCQC can be evaluated. For some QMs, it is necessary to consider the medical process in its specific context within the patient’s journey, which can be understood as a chain of events. This is an important aspect when assessing procedure efficacy, as it strongly depends on the previous and subsequent procedures. We present the concept of HDTs as a model integration framework, enabling a patient path representation.

The growing interest and relevance of mathematical modeling of patient outcomes are indicated in Table 5, which includes the number of published articles from the last 7 years (see Multimedia Appendix 2 for the methodology used to extract these data). However, the number of publications including digital twins (DTs) or GNNs is only a fraction of the total number of publications addressing the application of ML to predict patient outcomes (for a comprehensive review, see the article by Kline et al [43]).

XSL•FO
RenderX

Table . Publications in the last 7 years.

Variable[a]	Year, n						
	2019	2020	2021	2022	2023	2024	2025
Machine learning	899	1517	2289	2662	3192	4701	7146
Graph neural network	13	7	10	16	20	39	55
Digital twin	2	5	7	7	20	63	161

[a]Queries considering "patient outcomes" and "machine learning," "graph neural network," or "digital twin" (see Multimedia Appendix 2 for the method used to construct the table; the content was actualized on January 5, 2026).

Graph-Based Methods for PCQC

With EHRs, extended research integrating different patient data for health care quality assessment is now possible. Regarding this development, Si et al [44] published a comprehensive survey about the use of current advances in patient representation based on data stored in EHRs.

In recent years, the application of linked knowledge in graphs has become relevant, in part because it allows holistic individualized representation, either by considering individual characteristics in an entire population [38] or by providing a patient-centered representation of the disease and the corresponding outcomes. Data representation as a graph is a method to leverage transparency in model construction. Graph models are also a way to map unstructured data, such as clinical narratives, into machine-readable data via large language models (LLMs). Additionally, graph models, including GNNs, have been considered to represent a way to close the gap between ML and symbolic reasoning [45], introducing the desired interpretability for ML methods.

The sample health care KG in Figure 2 can serve to represent the concept of graph models in health care: the health status of a patient can be modeled as the interlinking of various factors, such as the disease's attributes (including type, symptoms, and anatomy), the patient's genetics (including pathways and biological processes), the kind of drug compounds, and the characterization of side effects. Hence, graph models can function as holistic representations of a patient's health status through the integration of different data sources into a single model.

XSL•FO
RenderX

Figure 2. A simple health care knowledge graph inspired by the article of Abu-Salih et al [46]. The abstract graph shows how a disease is connected to its attributes (such as type, symptoms, and anatomy) and to compounds (and attributes such as side effects and pharmacological classes), and how it is associated with genes (and attributes such as pathways and biological processes). Interactions with the disease, which we refer to as medical procedures (such as diagnosis, medication, and measurement), can be added to the graph and can lead to the development of the disease over time (t_i–t_{i+1}).

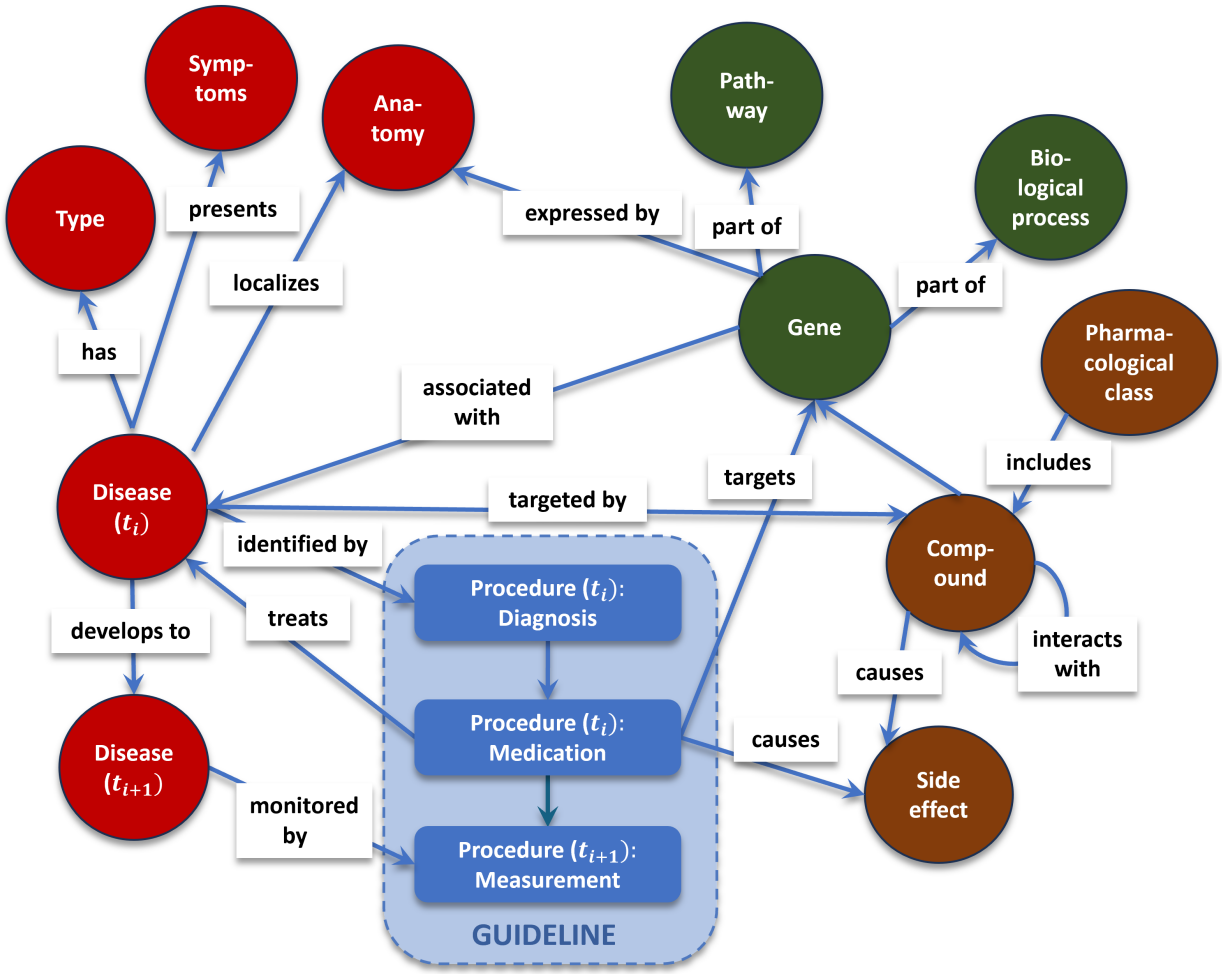

In graph models, representation learning can be applied to convert multivariate time series and static features into nodes of the graph [47], for example, when multiple steps, such as various procedures in different time steps or disease stages, are converted into nodes in the entire graph.

Hence, starting from the disease stage at time point t_i, procedures, such as diagnosis, can be performed, which introduces a new node into the KG. As clinical work is ideally performed within the frame of the guidelines, suitable medication will be proposed and applied. Any action taken through the performance of a medical procedure can lead to an evolution of the patient's disease state, which, at a time point, can be monitored by the next procedure, again inducing new nodes in the KG. Relating medical procedures to specific patient characteristics (such as the genetic profile of the patient, side effects, etc) enables a holistic view, allowing guidelines to be tailored to the individual patient and to differ from institutional guidelines. Graphs are not limited to representing a single event and can represent a whole process (ie, the patient's journey). This is useful for DT implementation, which will be discussed in the following section.

In addition to patient-centered representations, graph models for disease-centered representations can also be constructed, as shown in Figure 3. These can be created from electronic patient records because hospital records contain rich information about the interrelationship of diagnoses (diagnoses and corresponding codiagnoses), the interrelationship of medical procedures (one medical intervention leads to another), and the interrelationship of diagnoses and procedures (due to the causal relationship between the diagnoses and the associated procedures). If a malignant neoplasm of the lungs (C34) is diagnosed, it is already known from earlier observations that it is often accompanied by chronic obstructive pulmonary disease (J44) and secondary malignant neoplasms of the brain (C79.3). Against the background of these relationships, 4 possible methods can be identified: computed tomography, bronchoscopy, magnetic resonance imaging, and positron emission tomography.

In Multimedia Appendix 3, we present an overview of the literature concerning the application of graph models for patient safety, procedure accuracy, and procedure efficacy.

XSL•FO
RenderX

Figure 3. Exemplary knowledge graph related to disease diagnoses and medical procedures. In this example, we present the relationship involving a malignant neoplasm of the lung (C34). We also present typical codiagnoses of this disease on the left of the image (J44 and C79.3). On the right of the image, we present corresponding medical procedures that integrate the main diagnosis and codiagnoses (such as computed tomography, bronchoscopy, magnetic resonance imaging, and positron emission tomography). This example has been extracted from guidelines for lung cancer in oncology [48]. *ICD*: *International Classification of Diseases*; OPS: Operationen- und Prozedurenschlüssel.

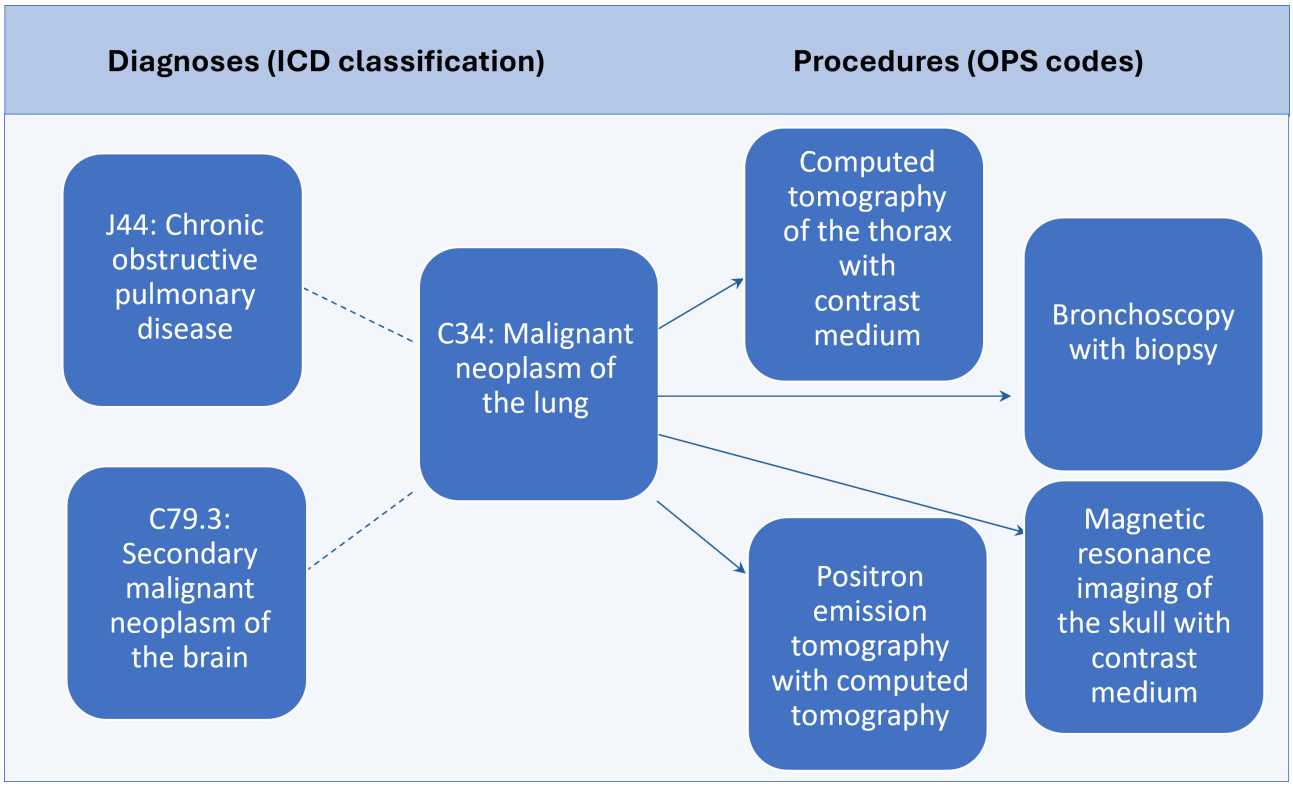

Limitations of Graph-Based Methods for PCQC

As the graph grows in size, querying and updating can become increasingly resource-intensive, potentially impacting the model's performance. In addition, algorithms are often based on the message-passing paradigm, which consists of node-by-node iteration (where each node aggregates information from neighboring nodes), which has significant limitations for graphs consisting of highly connected nodes [49]. For this reason, graphs with explicit heterogeneous edge distributions pose a problem and are difficult to evaluate because of the excessive compression of exponentially growing information in fixed-size vectors [50]. Recently, methods based on a different approach than the node-centered approach, which is based on differential geometry, have shown promising results for overcoming graph bottlenecks [51].

KGs often have incomplete data. If the graph does not have all the relevant entities and relationships, the model may lack important information. Incorrect or outdated information can lead to erroneous insights and conclusions.

HDTs and Health Digital Shadows for Modeling Procedure Efficacy

To model procedure efficacy, it is necessary to analyze the system's (ie, patient's) response to an external perturbation (eg, administration of a drug). It is crucial to gain deep insights into a patient's condition to assess the efficacy of procedures. Here, we want to present how these modeling tasks can be accomplished by HDTs [52] and by health digital shadows (HDSs), specifically when using graph-based methods.

For modeling procedure efficacy, the targets are not only binary values indicating patient outcomes or medical procedures but also biomarkers recorded as time series. For example, it is currently not possible to predict in advance whether a patient will face an allergic reaction or intolerance to medication. However, the ability to predict such events is relevant, as these reactions can be life-threatening and lead to treatment delays or morbidity.

Another problem is that physicians lack the time to review all available medical records of a patient. In the case of complex diseases, such as cancer, with various genetic mutations that could change during the procedure, prior or regular modeling is needed. Procedure adaptations or medication adjustments can be made in cases of side effects or low procedure efficacy.

For example, genetic polymorphisms in drug metabolism are already known for some substances, which must be determined before the substance is administered (eg, dihydropyrimidine dehydrogenase testing before 5-fluorouracil administration). Possible drug interactions must also be considered if several diseases are present. Many patients take several medications, which can sometimes lead to unforeseeable interactions. For patients, these interactions are increasingly difficult to oversee and evaluate, and individual factors also play a role in tolerability. The combination of different target values, along with different time spans, makes the problem definition much more complex than for patient safety and procedure accuracy.

XSL•FO
RenderX

Hence, an algorithmic architecture capable of considering a wide range of patient information and potential interactions would help model patient responses to medical procedures, predict patient outcomes, and test different counterfactual scenarios (ie, estimating a patient's outcome on the basis of a treatment regime that is distinct from the one that was actually followed and the observed outcome) [53,54]. Therefore, we model the integration of different procedures within a process. Technically speaking, this task represents the need for the integration of various models along a process (ie, patient's journey).

For this modeling task, the concept of a DT, which originated from engineering sciences [55], can be adapted into a clinical context. The aim of a DT is to generate a digital representation of an existing object within a process or time evolution in a digital space. Its concept is designed to combine time-evolving information from real objects, such as data generated from the Internet of Things, and mathematical modeling in the digital space [56]. Therefore, this digital representation allows the adaptation or optimization of processes in a digital space before transformations or modifications of real processes are made. For example, by integrating different and disparate information sources, it is possible to create different models not only for industrial production but also for the solution of critical and strategic problems, such as pollution, by generating universal avatars [57]. In summary, the design of a DT has the following main characteristics [56]:

1. Enablement of bidirectional communication between the model and real assets.
2. Integration of different information sources, such as genetic information, comorbidities, medication history, medical history, and family history.
3. Integration of different models, such as mechanistic models (eg, physiologically based pharmacokinetic modeling [PBPK] models [58]) or inductive models (statistical models such as structured nested mean models or ML models), at one time point or for one procedure [53].
4. Coupling of individual procedure models within a process, for instance, in a patient journey.
5. Simulation of procedures over time and process.
6. Model improvement through learning from the digital cohort (sum of all DTs).
7. Compliance with context-specific requirements arising from the field of application (eg, clinical health service).

In the industry, objects can be fully digitalized (with a bidirectional flow between the object and the digital object), which is limited only by the current evidence and understanding of the underlying complexity of the physical or biochemical processes or materials that should be mimicked. However, in the medical context, patients are not only too biologically complex but also cannot be objectivized in a straightforward way [59]. For this reason, we prefer to refer to HDSs, even when the literature currently mainly uses the term HDTs [60].

Furthermore, the connection between KGs and HDSs represents a logical step in the way different and disparate information sources are connected to create descriptive and predictive models. Relevant patient information as well as medical guidelines (in the form of KGs) can be integrated into a full graph containing patient data (eg, genetic or socioeconomic data) and additional relevant information (eg, drug structure and drug interactions).

In its role as an information and model integrating method, an HDS is well-suited for the following [61]:

- Generic representations: The predicted value of a quantity of interest is within the expected range of the global population. The representations are on an abstract and qualitative scale.
- Population-specific representations: The predicted value of a quantity of interest is evaluated with respect to a patient cohort. Such representations are more precise, enabling quantitative predictions.
- Specific representations: The predicted value of a quantity of interest has high accuracy in relation to the expected value for a single patient. These models are used for personalized predictions.

The HDS concept is interesting for prospective analysis, that is, in the simulation and prediction of a patient's health status and response to procedures. For prognostic purposes, it includes the ability to integrate all available patient information. Doctors treating patients can verify in advance the planned treatment regimen, such as medication or radiotherapy, on the patient's digital representation.

The HDS model runs through all the interactions between the procedure and the patient. In the case of pharmacological treatment, medication interactions are simulated (eg, by integrating ML models for drug-drug interactions [62]). Then, an individual recommendation for the substance and the individual dosage and medication can be obtained by applying appropriate mathematical models (ie, ML). Additionally, if the treatment response is not sufficient, the HDS could propose additional procedures or further treatment strategies. Current databases for potential drug targets could also be considered.

Figure 4 shows how all these steps are included in a closed information loop. Bidirectional communication between the real world and the model is enabled. Therefore, the general model obtains patient-specific information at time point t_i (element 1). This enables the human HDS to make a prognosis (element 2), which can prospectively be used by the clinician to adapt the medical decision (element 3). Alternatively, model prediction can be used retrospectively to evaluate medical decisions by examining patient outcomes at time points t_{i+1} (element 4). As time evolves, the patient's journey produces increasing patient-specific information, leading to better possible adaptation of the general model to a personalized human HDS. As the HDS cohort grows, model readjustments for the general model are performed after certain time periods (element 5).

XSL•FO
RenderX

Figure 4. Representation of a health digital shadow (HDS) as an information system. The HDS integrates patient data into a generally derived mathematical model (1) and allows model prognosis (2). Thus, the model supports the work of medical personnel as additional information for medical decision-making (3), which leads to a patient outcome at the next time point. In this way, bidirectional communication between the HDS and the real world is enabled. The model is also able to simulate the patient's journey, although the model's prognosis can be used to directly estimate the patient's outcome (4). The information gained by the individual human HDS can be used to readjust the general model constructed by the digital cohort (5).

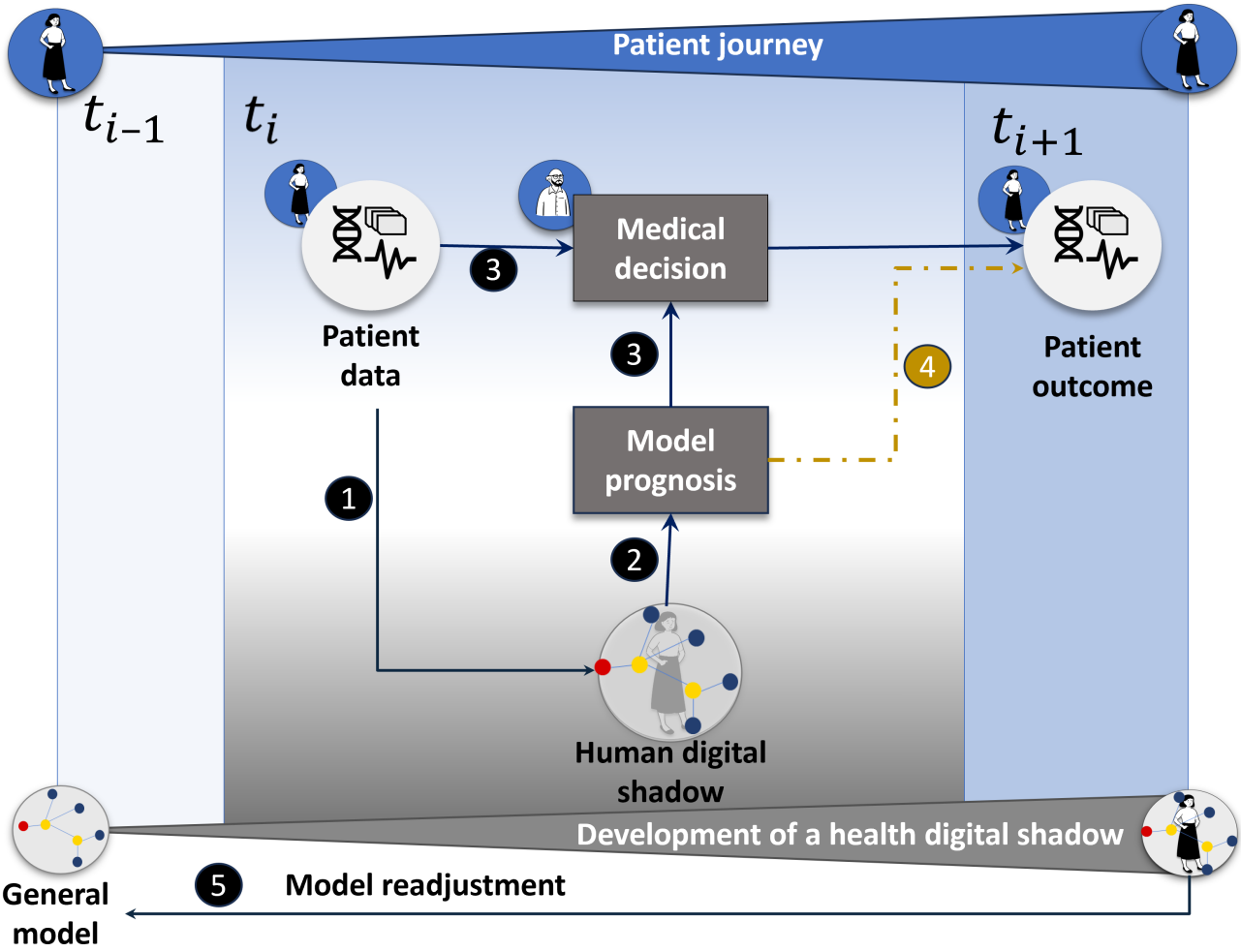

Thus, HDSs could increase the safety of drug therapy and its efficacy by providing personalized recommendations based on all available information [63]. For example, a similar concept, which is based on a whole-body representation via PBPK models and the integration of multiagent systems representing hepatocyte metabolism, can be used for pharmacology, particularly for the representation of the avatar's response to a medication [64,65]. This example demonstrates that through the analysis of the avatar's response, it is possible to plan procedures and even consider individual patient characteristics (eg, personalized metabolite levels that depend on the patient's genetics) that lead to personalized patient care.

This example implies that procedures can be tested on the DHS to forecast procedure success; therefore, the recommendation level of a procedure can be assessed before being applied to a real patient. Such an application is equivalent to the so-called dynamic HDT, where forecasts based on time series are performed [65].

Consequently, HDSs can be useful for modeling procedure efficacy, as in the following examples [66]:

- Human HDSs for inpatients:
  - We analyze which combination of medical procedures is most appropriate for patients depending on their socioeconomic context, environment, genetics, etc.
  - We analyze whether medical decisions will have an effective and desired outcome.
- Human HDSs for outpatients:
  - Help physicians recognize dangerous changes in a patient's health state through the analysis of patient data that can be delivered in real time (wearables or machines, for instance, in nephrology).

HDSs are increasingly being used for the estimation of the quality of patient care [55]. To this end, the prediction of patient outcomes is relevant, for example, in a personalized patient assessment to decide whether a procedure with high risk (death, side effects, patient burden, etc) is needed. This prevents the deterioration of a patient's quality of life. Figure 5 shows an exemplary representation of a human HDS that integrates different data sources (such as omics, EHR, and documentation), which are measured in different time periods along the patient journey, into different models to predict patient outcomes.

XSL•FO
RenderX

Figure 5. Schematic representation of a health digital shadow (HDS). An HDS is not a single model but a collection of different models to perform different predictions depending on the system's parameters of interest. Such predictions can include the analysis and prediction of time series at different time points of the patient's journey. In this image, we present a scheme of a patient's journey: after making the first medical decision on the basis of clinical guidelines (1), different information sources are integrated to predict quality metrics in different time periods depending on the patient's evolution (2). For example, the pharmacological response can be estimated via a whole-body model (eg, the physiologically based pharmacokinetic model [67]). Patient outcomes can be subsequently estimated via, for example, cluster analysis (3) [68]. EHR: electronic health record; HDT: health digital twin.

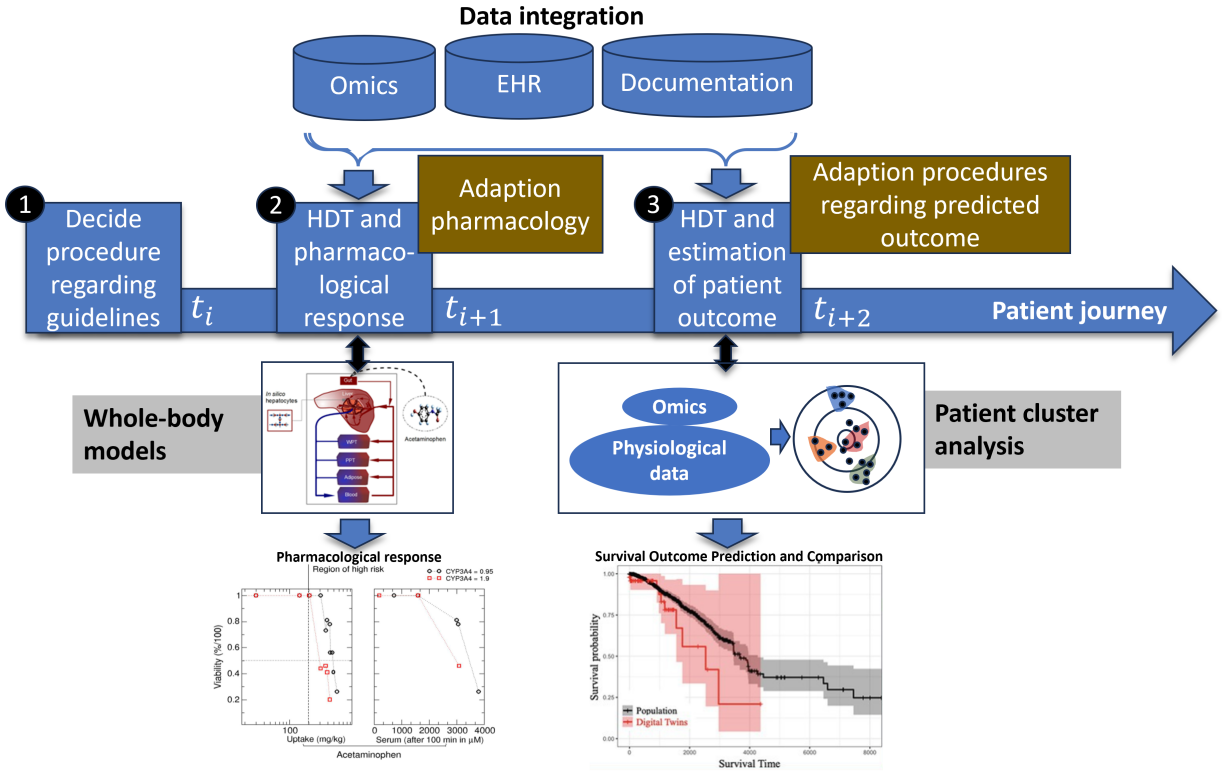

As shown in Figure 5, after choosing the best possible medical procedure according to clinical guidelines, different information sources are integrated to predict QMs in different time periods depending on the patient's evolution. For example, a whole-body (PBPK) model can be used to integrate patient-specific data within the HDT to predict a patient's pharmacological response. Using pharmacological response prediction and patient data, another model can estimate patient outcomes (eg, through cluster analysis) for survival time prediction.

Chang et al presented a remarkable example of possible DHS realization through the creation of a pipeline for predicting patient outcomes, particularly survival heterogeneity, for patients with breast cancer [68,69]. In this implementation, 3 different data sources were integrated, where low-dimensional embeddings of clinical and molecular features enriched by an external annotation database were the main data sources. In this algorithm, a dimensional reduction method, Uniform Manifold Approximation and Projection, is implemented. This is a general-purpose manifold learning and dimension reduction algorithm to reduce the dimension of the data space. A dynamic HDS requires a pipeline that is accordingly adapted (ie, new data and additional models must be integrated and appropriately adapted according to other outcomes) [68].

Nitschke et al [68] presented a general and unspecialized design that is applicable for HDSs and is predictive, modular, evolving, informed, interpretable, and explainable. This design combines KGs and ensemble learning to represent the patient's entire clinical journey and assist clinicians in their decision-making. Broad clinical application was shown by presenting 2 explicit cases: prostate cancer biopsy and glioma treatment decisions [68].

Limitations of HDSs for Modeling Procedure Efficacy

HDSs often oversimplify real systems. These approximations can lead to discrepancies between the model and the actual system. The accuracy of the HDS depends on the quality and granularity of the data used. Since more than one model is coupled to generate a full HDS, different models trained on different data and granularities can lead to unbalanced representations. Finally, poor or sparse data can lead to unreliable models [70]. Notably, all these problems are exacerbated by issues related to the aggregation and integration of different types of data that occur in health care, along with the strict privacy policies that must be followed [71], a problem that may be alleviated with the introduction of EHRs [72].

Creating and maintaining an HDS can be computationally intensive, requiring significant resources for simulation and real-time data processing. Integrating an HDS with other

XSL•FO
RenderX

systems, sensors, and data sources can be complex and may require sophisticated interfaces and protocols. Ensuring that the HDS reflects real-time changes in the physical system can be challenging, especially in systems with high dynamics [73].

An interesting option is the relation and validation of HDSs using in vitro methods, such as organ-on-a-chip. This option opens the possibility to compensate for potential defects in model predictions and identify their potential problems by validating and comparing results with advanced in vitro patient representations [74]. This, however, is still a challenge in the field of oncology and requires more research to obtain a correct representation of patient outcomes.

Example of the Implementation of KGs and HDSs

The clinical scenario in Textbox 1 illustrates the integration of KGs and HDSs and demonstrates how our proposed framework evaluates all 3 QM categories while simultaneously assessing treatment safety in patients with comorbidities, ensuring guideline accuracy while allowing for personalized adaptation, and optimizing efficacy through predictive modeling of multiple therapeutic pathways for patients with non–small cell lung cancer. This case highlights how KG/HDS systems add value beyond static guidelines by dynamically integrating genomic data, resistance patterns, and real-world evidence to support complex clinical decision-making and establish prospective monitoring protocols.

Textbox 1. Clinical scenario.

A 64-year-old never-smoker with advanced lung adenocarcinoma undergoes comprehensive genomic profiling. Data: electronic health record showing diagnosis code C34.1 with brain metastases present; next-generation sequencing results showing EGFR L858R primary, T790M resistance mutation, and concurrent TP53; PD-L1 tumor proportion score showing a result of 2%; and baseline liver function test showing an alanine aminotransferase level of 45 U/L. Knowledge graph identifies treatment hierarchy according to National Comprehensive Cancer Network guidelines and maps resistance patterns to subsequent therapy options. Health digital shadow predicts the intracranial response rate with osimertinib (efficacy quality metric [QM]=0.70), hepatotoxicity probability (safety QM=0.18), and time to next progression (median 11.2 months). Accuracy assessment flags missing baseline brain magnetic resonance imaging according to the guidelines (accuracy QM=0.85). The system recommends osimertinib 80 mg daily with a central nervous system–specific imaging schedule, hepatic monitoring every 2 weeks initially, and liquid biopsy at 3 months for emerging resistance detection. The model also preidentifies next-line options based on likely resistance mechanisms, enabling proactive treatment planning.

Issues and Future Research Directions

Our analysis revealed that patient safety, the accuracy of the procedure, and the effectiveness of the procedure are paramount. Using the present taxonomy, we can define the types of challenges and the type of ML implementation to be performed. However, this taxonomy is far from perfect. For example, distinguishing between a diagnostic procedure and a therapeutic procedure may require a more specialized and precise taxonomy. In addition, the metrics required to evaluate patient outcomes and estimate the QI may not always be perfectly calibrated [75], which can limit the applicability of such methods. In general, patient outcomes are still highly debated [76]. Despite this, we believe that it is important to keep any classification and taxonomy as simple as possible to facilitate the implementation of mathematical models.

We concluded that holistic quality management analysis (especially regarding procedure efficacy) requires the integration of several data sources. This is because there is no single decision (in decision support) or event (assignment of a diagnosis), but whole processes impact the patient's health status. Quality management and assessment can be very challenging for model implementation since it requires the integration of a constant information flow, which can continuously change within the process. As a possible solution to address these issues, we have introduced the concept of the human HDS.

Guidelines serve as guidance to define and implement processes necessary for the management of patients' health status. They are the result of clinical studies and are intended to help doctors make decisions. However, they contain complex workflows that should be evaluated individually. While a deviation from a guideline may be a potential problem in disease management for one patient because a step in the workflow has not been taken into account, which is a reduction in the quality of the hospital (ie, a decrease in procedural accuracy), for other patients, this deviation can be seen as a necessary adjustment needed for more personalized health care (ie, increasing the effectiveness of the procedure). In this context, especially in the evaluation of personalized practices, ML methods can influence the way guidelines are applied. Furthermore, we are cautious about full model–assisted personalization as well. Some critics have reported that data integration and target modeling have the potential to move away from real personalized patient centers and more human medicine [77]. In this case, individual modeling, which is embedded in a full modeling lifecycle, as shown in Multimedia Appendix 4, is just a way to quantitatively evaluate guidelines and their potential effects on the expected quality management of the disease at the patient level.

In this framework, graph representation is appealing, given that graph models are essentially an integration of data and information. Furthermore, when time is considered, it is possible to generate a digital representation of the patient (ie, an HDS). Furthermore, this concept has considerable potential for both inpatients and outpatients. This assumes that data generated by portable devices can track the patient's condition at home. Furthermore, this concept can be coupled with digitally stored patient journeys (ie, the digital cohort) [78].

Currently, an HDS essentially involves the integration of different models within a process. We believe that in the future, the application of KGs will contribute to the rapid development of HDSs. For example, recent work has suggested that the combination of KGs and LNNs could be the key to the implementation of productive avatars in medicine [79].

Novel decentralized data storage concepts may help train and define accurate models with sufficient data volumes. For

XSL•FO
RenderX

example, by integrating different and disparate information sources and real-world data built on the FAIR (findability, accessibility, interoperability, and reusability) principle [80], some of which are interconnected through the semantic web (according to the Worldwide Web 3.0 standards, for example, semantic web, such as Solid [81]), it is possible to create models not only for industrial production but also for the solution of critical and strategic problems by generating universal avatars [57]. Several recent initiatives are developing such ideas to implement DTs on the basis of semantic web principles (see, for instance, the EDITH-CSA project [61]). In such initiatives, it is remarkable to find not only the definition and exploration of models but also the definition of data standards required for data interoperability, most of which are based on Fast Healthcare Interoperability Resource (FHIR) standards.

HDSs face additional challenges. In the event of poor data safety standards, criminals could gain access to medical records and HDSs for blackmail or ransom demands. Companies, such as health insurance or disability insurance, might also demand the patient's HDS prognosis before enrollment. Patients could be denied insurance, insurance rates could be increased, or existing insurance conditions could be adapted regularly on the basis of HDSs. This could ultimately lead to the selection of people who will not receive insurance at all or who will receive only reduced coverage. Institutions and employers could also demand an HDS model before hiring people in order to select healthy people.

Another relevant issue is data contamination by ML models. For example, if a machine catches early signs of sepsis and doctors treat it, this creates a "contaminated association" in the data. Furthermore, an intervention triggered by one model can quietly disrupt another, even if they are focused on entirely different outcomes [82]. This can be particularly critical for HDTs. A potential solution is to preserve clean data from datasets where predictive models are not used.

The level of mandatory treatment recommendations based on HDSs must be discussed and assessed. Patients' wishes are crucial. What happens if the physician or patient deviates from an HDS's recommendations? In this case, the insurance company may not cover the alternative treatments since the HDS does not predict them to be most effective. This question is related to the overall reliability of HDSs. In retrospect, the accuracy of such recommendations can be assessed (eg, by checking procedure efficacy and tolerability).

In summary, maintaining a healthy balance between the sense of euphoria and the concern for validity in the implementation of mathematical models is imperative. Graph models require clean labels, accurate graph structures, and sufficient initial node features; however, real-world graph data often have noise and sparse labels, whereas different datasets have distinct feature constructions [83]. Furthermore, DHSs often oversimplify the systems they represent, are challenged by in vivo conditions, and are difficult to maintain.

Recently, foundation large x models (LxMs) trained on large databases (eg, LLMs) have accelerated the implementation of most mathematical models while reducing development costs. However, these models are expensive and difficult to generalize (even LLMs have limited applicability). In such cases, a balance between LxMs (for instance, LLMs) and tailored models (for instance, BERT [Bidirectional Encoder Representations from Transformers]) can be the most suitable, cost-effective, and safe solution to all these challenges.

Implications of the Study for Practice

The following steps and checks should help minimize some of the limitations of KGs applied to PCQC:

- Define a realistic system size that can be represented by KGs and maintain a critical view of the overscaling and size of KGs.
- Define aggregation functions to derive partial graph folding and compensate for incomplete node categories [51].
- Keep the graph updated with new knowledge and ensure consistency across different sources [84].
- Define appropriate data structures, such as Resource Description Frameworks, to make KGs interoperable.

Similarly, the following checklist can guide the correct implementation of HDSs applied to PCQC:

- Ensure data harmonization and minimization/access control of different data sources and digital records.
- Implement human-in-the-loop overriding for the correct interpretation of HDSs and avoidance of problems such as alert fatigue.
- Use drift detection/recertification and calibration standards (eg, using organ-on-a-chip) for safety in QM implementation on HDSs.
- Implement staged rollouts (eg, silent trials) before HDS activation.

Finally, the following are the key findings and conclusions derived throughout this article:

- Patient outcomes are not abstract quality terms since they can be formalized and linked to measurable clinical processes.
- Medical guidelines are essential but often underutilized in quality monitoring.
- Real-world clinical decisions are complex and require personalization.
- KGs and GNNs help bridge fragmented clinical data.
- HDSs offer a vision for prospective, patient-specific quality evaluation.
- Quality management should shift from static checklists to dynamic, data-driven guidance.

Conclusion

The integration of KGs and DTs into health care holds transformative potential for improving clinical outcomes and the quality of care. However, this optimistic view is feasible only when the technologies are implemented in an ethical way, avoiding blind customer overconfidence in the delivered results. Furthermore, implementation requires accounting for the perspectives of stakeholders (as well as their own cultural environment and background): What is considered a desirable outcome from a patient-centered perspective (eg, symptom relief

XSL•FO
RenderX

and preserved quality of life) may diverge from institutional or systemic goals such as shorter hospital stay or cost efficiency.

As such, mathematical models for health care quality must clearly define whether the primary focus lies in maximizing individual patient well-being, the operational performance of the health care system, or economic sustainability. These dimensions are not mutually exclusive, but the prioritization among them fundamentally shapes the modeling framework and interpretation of safety-related events. From this perspective, we also provide concrete checklists and recommendations for the correct deployment of models, for example, for the generation of KGs or the deployment of HDSs.

The combination of patient outcomes, procedures, and guidelines is not only interesting from a technical point of view but also extremely relevant for improving the quality of hospitals and health centers. The integration of guidelines implies the application of ML in different ways, from alerting and predicting critical conditions to the application of similar methods used in fraud detection to recognize potential deviations from guidelines.

The topic of quality management in medicine and the implementation of more detailed mathematical definitions are expected to become as relevant as diagnosis and image recognition are today, and we expect an increasing interest in further ML applications for quality management in the next few years.

Acknowledgments

AKN would like to thank Matthias Weidemüller for insightful and inspiring discussions, as well as for the invaluable support provided throughout this work. Additionally, AKN would like to thank Fabian Egersdörfer and Carlos Brandl for proofreading the text and providing recommendations. JGDO would like to thank Faizan E Mustafa for inspiring discussions and providing support in the research about graph knowledge, and Elena Ramirez for contributing to the assessment of the relationship of medical decisions in the framework of the theory of rational decision-making. Moreover, JGDO would like to thank Felix Weil for providing constant support and MaryLena Bleile for pointing out the relationship between health digital shadows and testing different counterfactual scenarios. We are grateful to the editor and reviewers who provided comments and helped us improve this article. No identifiable patient-level data were accessed or analyzed in this conceptual viewpoint.

All the ideas, texts, and figures were generated by the authors. No artificial intelligence tool was used to generate substantive content, and only language polishing was supported by the use of such a tool. According to the STM taxonomy [85], the artificial intelligence tool used for the preparation of this manuscript belongs to category 1: language refinement was carried out via Rubriq to detect spelling and typographical errors.

Funding

This project was supported by the Ministry for Economics, Labor and Tourism from Baden-Württemberg, Germany via grant agreement number BW1_1456 (AI4MedCode).This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2181/1 - 390900948 (the Heidelberg STRUCTURES Excellence Cluster). This project was also supported by broad funding to the Klinikum Stuttgart, Stuttgart Cancer Center – Tumorzentrum Eva Mayr-Stihl, from the Eva Mayr-Stihl Foundation, Waiblingen, Germany.

Authors' Contributions

The initial idea was proposed and drafted by AKN and JGDO. JGDO contributed to the overview of the knowledge graphs. SN, AKN, and JGDO contributed to the digital twin overview. SN and MK contributed to the medical background, examples, and critical evaluation of the various methods described in this article. All the authors contributed to the final version of this article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Insight into the different aspects of the assessment of patient-centered quality of care through possible mathematical modeling applications.
[DOCX File, 21 KB - jmir_v28i1e81946_app1.docx]

Multimedia Appendix 2
Analysis of the available literature on PubMed.
[DOCX File, 17 KB - jmir_v28i1e81946_app2.docx]

Multimedia Appendix 3
Literature overview of the application of graph models for patient safety, procedure accuracy, and procedure efficacy.

XSL•FO
RenderX

[DOCX File, 31 KB - jmir_v28i1e81946_app3.docx]

Multimedia Appendix 4
Lifecycle of the data and models in relation to clinical guidelines.
[DOCX File, 530 KB - jmir_v28i1e81946_app4.docx]

References

1. McLachlan S, Kyrimi E, Dube K, Hitman G, Simmonds J, Fenton N. Towards standardisation of evidence-based clinical care process specifications. Health Informatics J 2020 Dec;26(4):2512-2537. [doi: 10.1177/1460458220906069] [Medline: 32186428]
2. Timmermans S, Mauck A. The promises and pitfalls of evidence-based medicine. Health Aff (Millwood) 2005;24(1):18-28. [doi: 10.1377/hlthaff.24.1.18] [Medline: 15647212]
3. Sinsky CA, Bavafa H, Roberts RG, Beasley JW. Standardization vs customization: finding the right balance. Ann Fam Med 2021;19(2):171-177. [doi: 10.1370/afm.2654] [Medline: 33685879]
4. Alexopoulos K, Chryssolouris G. Process. In: Chatti S, Laperrière L, Reinhart G, Tolio T, editors. CIRP Encyclopedia of Production Engineering: Springer; 2019:1349-1352. [doi: 10.1007/978-3-662-53120-4_6567]
5. Da Silva N. The industrialization of healthcare and its critiques. In: Diaz-Bone R, Larquier G, editors. Handbook of Economics and Sociology of Conventions: Springer; 2023:1-25. [doi: 10.1007/978-3-030-52130-1_45-2]
6. Dodwad SS. Quality management in healthcare. Indian J Public Health 2013;57(3):138-143. [doi: 10.4103/0019-557X.119814] [Medline: 24125927]
7. Rojas E, Munoz-Gama J, Sepúlveda M, Capurro D. Process mining in healthcare: a literature review. J Biomed Inform 2016 Jun;61:224-236. [doi: 10.1016/j.jbi.2016.04.007] [Medline: 27109932]
8. May M. Eight ways machine learning is assisting medicine. Nat Med 2021 Jan;27(1):2-3. [doi: 10.1038/s41591-020-01197-2] [Medline: 33442003]
9. Bach JF. Causality in medicine. C R Biol 2019;342(3-4):55-57. [doi: 10.1016/j.crvi.2019.03.001] [Medline: 30981720]
10. Diaz Ochoa JG. A unified method for assessing the observability of dynamic complex systems. Comput Biol Med 2023 Jun;160:107012. [doi: 10.1016/j.compbiomed.2023.107012] [Medline: 37187137]
11. Ensure healthy lives and promote well-being for all at all ages. United Nations. URL: https://sdgs.un.org/goals/goal3 [accessed 2025-08-01]
12. Steele K, Stefánsson HO. Decision theory. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy (Winter 2020 Edition): Metaphysics Research Lab, Stanford University; 2020. URL: https://plato.stanford.edu/archives/win2020/entries/decision-theory/ [accessed 2024-05-12]
13. Epstein RM, Street RL. The values and value of patient-centered care. Ann Fam Med 2011;9(2):100-103. [doi: 10.1370/afm.1239] [Medline: 21403134]
14. Kersting C, Kneer M, Barzel A. Patient-relevant outcomes: what are we talking about? A scoping review to improve conceptual clarity. BMC Health Serv Res 2020 Jun 29;20(1):596. [doi: 10.1186/s12913-020-05442-9] [Medline: 32600321]
15. Liu Y, Avant KC, Aungsuroch Y, Zhang XY, Jiang P. Patient outcomes in the field of nursing: a concept analysis. Int J Nurs Sci 2014 Mar;1(1):69-74. [doi: 10.1016/j.ijnss.2014.02.006]
16. Hadian SA, Rezayatmand R, Shaarbafchizadeh N, Ketabi S, Pourghaderi AR. Hospital performance evaluation indicators: a scoping review. BMC Health Serv Res 2024 May 1;24(1):561. [doi: 10.1186/s12913-024-10940-1] [Medline: 38693562]
17. Wang XS, Gottumukkala V. Patient-reported outcomes: is this the missing link in patient-centered perioperative care? Best Pract Res Clin Anaesthesiol 2021 Dec;35(4):565-573. [doi: 10.1016/j.bpa.2020.10.006] [Medline: 34801218]
18. Busse R, Klazinga N, Panteli D. In: Quentin W, editor. Improving Healthcare Quality in Europe: Characteristics, Effectiveness and Implementation of Different Strategies: European Observatory on Health Systems and Policies; 2019.
19. Dovetail Editorial Team. Patient outcomes: overview, measures & ways to improve. Dovetail. 2023. URL: https://dovetail.com/patient-experience/what-are-patient-outcomes/ [accessed 2025-08-06]
20. McLeod C, Norman R, Litton E, Saville BR, Webb S, Snelling TL. Choosing primary endpoints for clinical trials of health care interventions. Contemp Clin Trials Commun 2019 Dec;16:100486. [doi: 10.1016/j.conctc.2019.100486] [Medline: 31799474]
21. Yanamadala S, Morrison D, Curtin C, McDonald K, Hernandez-Boussard T. Electronic health records and quality of care. Medicine (Baltimore) 2016;95(19):e3332. [doi: 10.1097/MD.0000000000003332] [Medline: 27175631]
22. Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e V [Website in German]. URL: https://www.awmf.org/ [accessed 2025-08-06]
23. S3-Leitlinie Prostatakarzinom [Article in German]. AWMF online. URL: https://register.awmf.org/de/leitlinien/detail/043-022OL [accessed 2025-08-06]
24. Langer T, Follmann M. Das Leitlinienprogramm Onkologie (OL): Nukleus einer evidenzbasierten, patientenorientieren, interdisziplinären Onkologie? [Article in German]. Z Evid Fortbild Qual Gesundhwes 2015;109(6):437-444. [doi: 10.1016/j.zefq.2015.09.007] [Medline: 26474648]

XSL•FO
RenderX

25. Schick J. IQM Positionspapier Ergebnisqualität [Article in German]. Initiative Qualitätsmedizin e V. 2025. URL: https://www.initiative-qualitaetsmedizin.de/positionspapier [accessed 2025-08-06]
26. Schmitt J, Klinkhammer-Schalke M, Bierbaum V, et al. Initial cancer treatment in certified versus non-certified hospitals. Dtsch Arztebl Int 2023 Sep 29;120(39):647-654. [doi: 10.3238/arztebl.m2023.0169] [Medline: 37583089]
27. Gigerenzer G, Gray JAM. Bessere Ärzte, Bessere Patienten, Bessere Medizin Aufbruch in Ein Transparentes Gesundheitswesen: Mit Einem Vorwort von Günther Jonitz Strüngmann [Book in German]: Medizinisch Wissenschaftliche Verlagsgesellschaft; 2013.
28. Peng J, Jury EC, Dönnes P, Ciurtin C. Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: applications and challenges. Front Pharmacol 2021;12:720694. [doi: 10.3389/fphar.2021.720694] [Medline: 34658859]
29. Emanuel L, Berwick D, Conway J, et al. What exactly is patient safety? J Med Regul 2009 Mar;95(1):13-24. [doi: 10.30770/2572-1852-95.1.13]
30. Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/ [accessed 2025-08-06]
31. Nimptsch U, Mansky T, Busse R. G-IQI – German Inpatient Quality Indicators Version 54: Universitätsverlag der TU Berlin; 2022.
32. Zaiss A, Dauben HP. ICHI-International Classification of Health Interventions: a balancing act between the demands of statistics and reimbursement [Article in German]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2018 Jul;61(7):778-786. [doi: 10.1007/s00103-018-2747-6] [Medline: 29808283]
33. Decision-making as reasoning. In: Cognitive Economics: Springer; 2008:49-67. [doi: 10.1007/978-3-540-71347-0_4]
34. Lafonte M, Cai J, Lissauer ME. Failure to rescue in the surgical patient: a review. Curr Opin Crit Care 2019 Dec;25(6):706-711. [doi: 10.1097/MCC.0000000000000667] [Medline: 31567517]
35. Greenes RA. Clinical Decision Support: The Road to Broad Adoption, 2nd edition: Academic Press; 2014.
36. Barth JH, Misra S, Aakre KM, et al. Why are clinical practice guidelines not followed? Clin Chem Lab Med 2016 Jul 1;54(7):1133-1139. [doi: 10.1515/cclm-2015-0871] [Medline: 26650076]
37. Börm P. Leitlinienbasierter Clinical Decision Support – Anforderungen an evidenzbasierte Entscheidungsunterstützungssysteme [Article in German]. OP-Journal 2021 Apr;37(1):28-35. [doi: 10.1055/a-1284-3193]
38. Diaz Ochoa JG, Mustafa FE. Graph neural network modelling as a potentially effective method for predicting and analyzing procedures based on patients' diagnoses. Artif Intell Med 2022 Sep;131:102359. [doi: 10.1016/j.artmed.2022.102359] [Medline: 36100347]
39. Becker M, Kasper S, Böckmann B, Jöckel KH, Virchow I. Natural language processing of German clinical colorectal cancer notes for guideline-based treatment evaluation. Int J Med Inform 2019 Jul;127:141-146. [doi: 10.1016/j.ijmedinf.2019.04.022] [Medline: 31128826]
40. Alharbe N, Rakrouki MA, Aljohani A. A healthcare quality assessment model based on outlier detection algorithm. Processes (Basel) 2022 Jun;10(6):1199. [doi: 10.3390/pr10061199]
41. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
42. Lynch SS. Evaluating drug efficacy and safety. MSD Manuals. URL: https://www.msdmanuals.com/professional/clinical-pharmacology/concepts-in-pharmacotherapy/drug-efficacy-and-safety [accessed 2024-08-08]
43. Kline A, Wang H, Li Y, et al. Multimodal machine learning in precision health: a scoping review. NPJ Digit Med 2022 Nov 7;5(1):171. [doi: 10.1038/s41746-022-00712-8] [Medline: 36344814]
44. Si Y, Du J, Li Z, et al. Deep representation learning of patient data from electronic health records (EHR): a systematic review. J Biomed Inform 2021 Mar;115:103671. [doi: 10.1016/j.jbi.2020.103671] [Medline: 33387683]
45. Liu M, Zhang F, Huang P, Niu S, Ma F, Zhang J. Learning the satisfiability of pseudo-boolean problem with graph neural networks. In: Simonis H, editor. Principles and Practice of Constraint Programming CP 2020 Lecture Notes in Computer Science, Vol 12333: Springer; 2020:885-898. [doi: 10.1007/978-3-030-58475-7_51]
46. Abu-Salih B, Al-Qurishi M, Alweshah M, Al-Smadi M, Alfayez R, Saadeh H. Healthcare knowledge graph construction: a systematic review of the state-of-the-art, open issues, and opportunities. J Big Data 2023;10(1):81. [doi: 10.1186/s40537-023-00774-9] [Medline: 37274445]
47. Zheng Z, Guo C, Chen J, Li J. Graph neural network-based representation learning for medical time series. In: Iliadis L, Papaleonidas A, Angelov P, Jayne C, editors. Artificial Neural Networks and Machine Learning – ICANN 2023: Springer Nature; 2023:194-205. [doi: 10.1007/978-3-031-44223-0_16]
48. S3-Leitlinie Prävention, Diagnostik, Therapie und Nachsorge des Lungenkarzinoms [Report in German]. : Leitlinienprogramm Onkologie; 2024 URL: https://www.leitlinienprogramm-onkologie.de/fileadmin/user_upload/Downloads/Leitlinien/Lungenkarzinom/Version_3/LL_Lungenkarzinom_Langversion_3.0.pdf [accessed 2025-08-06]
49. Sato R. A survey on the expressive power of graph neural networks. arXiv. Preprint posted online on Mar 9, 2020. [doi: 10.48550/arXiv.2003.04078]
50. Alon U, Yahav E. On the bottleneck of graph neural networks and its practical implications. Open Review. 2021. URL: https://openreview.net/forum?id=i80OPhOCVH2 [accessed 2022-09-07]

XSL•FO
RenderX

51. Topping J, Di Giovanni F, Chamberlain BP, Dong X, Bronstein MM. Understanding over-squashing and bottlenecks on graphs via curvature. Open Review. 2022. URL: https://openreview.net/forum?id=7UmjRGzp-A [accessed 2025-12-01]
52. Kulkarni C, Quraishi A, Raparthi M, et al. Hybrid disease prediction approach leveraging digital twin and metaverse technologies for health consumer. BMC Med Inform Decis Mak 2024 Apr 5;24(1):92. [doi: 10.1186/s12911-024-02495-2] [Medline: 38575951]
53. Chakraborty B, Moodie EEM. Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine: Springer; 2013. [doi: 10.1007/978-1-4614-7428-9]
54. Chakraborty B, Moodie EEM. Estimation of optimal DTRs by directly modeling regimes. In: Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine: Springer; 2013:79-100. [doi: 10.1007/978-1-4614-7428-9_5]
55. Vallée A. Digital twin for healthcare systems. Front Digit Health 2023;5:1253050. [doi: 10.3389/fdgth.2023.1253050] [Medline: 37744683]
56. Richter J, Lange F, Scheper T, Solle D, Beutel S. Digitale Zwillinge in der Bioprozesstechnik – Chancen und Möglichkeiten [Article in German]. Chemie Ingenieur Technik 2023 Apr;95(4):498-510. [doi: 10.1002/cite.202200166]
57. Akroyd J, Mosbach S, Bhave A, Kraft M. Universal digital twin - a dynamic knowledge graph. Data-Centric Engineering 2021;2:e14. [doi: 10.1017/dce.2021.10]
58. Peters SA. Physiologically‐Based Pharmacokinetic (PBPK) Modeling and Simulations: Principles, Methods, and Applications in the Pharmaceutical Industry: John Wiley & Sons, Inc; 2012. [doi: 10.1002/9781118140291]
59. John A, Alhajj R, Rokne J. A systematic review of AI as a digital twin for prostate cancer care. Comput Methods Programs Biomed 2025 Aug;268:108804. [doi: 10.1016/j.cmpb.2025.108804] [Medline: 40347618]
60. Katsoulakis E, Wang Q, Wu H, et al. Digital twins for health: a scoping review. NPJ Digit Med 2024 Mar 22;7(1):77. [doi: 10.1038/s41746-024-01073-0] [Medline: 38519626]
61. Building the European virtual human twin. : Edith; 2023 URL: https://www.edith-csa.eu/wp-content/uploads/2023/04/EDITH_D3.1_vision_roadmap_outline_final.pdf [accessed 2025-08-06]
62. Mei S, Zhang K. A machine learning framework for predicting drug-drug interactions. Sci Rep 2021 Sep 2;11(1):17619. [doi: 10.1038/s41598-021-97193-8] [Medline: 34475500]
63. Diasio RB, Offer SM. Testing for dihydropyrimidine dehydrogenase deficiency to individualize 5-fluorouracil therapy. Cancers (Basel) 2022 Jun 30;14(13):3207. [doi: 10.3390/cancers14133207] [Medline: 35804978]
64. Diaz Ochoa JG, Bucher J, Niklas J, Mauch K. Multi-scale modeling for individualized spatiotemporal prediction of drug effects. Toxicol Lett 2013 Aug;221:S83-S84. [doi: 10.1016/j.toxlet.2013.05.093]
65. Mulder ST, Omidvari AH, Rueten-Budde AJ, et al. Dynamic digital twin: diagnosis, treatment, prediction, and prevention of disease during the life course. J Med Internet Res 2022 Sep 14;24(9):e35675. [doi: 10.2196/35675] [Medline: 36103220]
66. Sun T, He X, Li Z. Digital twin in healthcare: recent updates and challenges. Digit Health 2023;9:20552076221149651. [doi: 10.1177/20552076221149651] [Medline: 36636729]
67. Diaz Ochoa JG, Bucher J, Péry ARR, Zaldivar Comenges JM, Niklas J, Mauch K. A multi-scale modeling framework for individualized, spatiotemporal prediction of drug effects and toxicological risk. Front Pharmacol 2012;3:204. [doi: 10.3389/fphar.2012.00204] [Medline: 23346056]
68. Nitschke AK, Brandl C, Egersdörfer F, Görtz M, Hohenfellner M, Weidemüller M. Design for a digital twin in clinical patient care. NPJ Health Syst 2025;3(1):11. [doi: 10.1038/s44401-025-00060-1]
69. Chang HC, Gitau AM, Kothapalli S, Welch DR, Sardiu ME, McCoy MD. Understanding the need for digital twins' data in patient advocacy and forecasting oncology. Front Artif Intell 2023;6:1260361. [doi: 10.3389/frai.2023.1260361] [Medline: 38028666]
70. Margaria T, Ryan S. Data and data management in the context of digital twins. In: Crespi N, Drobot AT, Minerva R, editors. The Digital Twin: Springer International Publishing; 2023:253-278. [doi: 10.1007/978-3-031-21343-4_10]
71. Awrahman BJ, Aziz Fatah C, Hamaamin MY. A review of the role and challenges of big data in healthcare informatics and analytics. Comput Intell Neurosci 2022;2022:5317760. [doi: 10.1155/2022/5317760] [Medline: 36210978]
72. Venkatesh KP, Raza MM, Kvedar JC. Health digital twins as tools for precision medicine: considerations for computation, implementation, and regulation. NPJ Digit Med 2022 Sep 22;5(1):150. [doi: 10.1038/s41746-022-00694-7] [Medline: 36138125]
73. Tao F, Zhang H, Zhang C. Advancements and challenges of digital twins in industry. Nat Comput Sci 2024 Mar;4(3):169-177. [doi: 10.1038/s43588-024-00603-w] [Medline: 38532139]
74. Zhai X, Chen M, Esfahani SS, et al. Heterogeneous system-on-chip-based Lattice-Boltzmann visual simulation system. IEEE Systems Journal 2020;14(2):1592-1601. [doi: 10.1109/JSYST.2019.2952459]
75. Endo H, Uchino S, Hashimoto S, et al. Development and validation of the predictive risk of death model for adult patients admitted to intensive care units in Japan: an approach to improve the accuracy of healthcare quality measures. J Intensive Care 2021 Feb 15;9(1):18. [doi: 10.1186/s40560-021-00533-z] [Medline: 33588956]
76. Reiss J, Ankeny RA. Philosophy of medicine. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy (Spring 2022 Edition): Metaphysics Research Lab, Stanford University; 2022. URL: https://plato.stanford.edu/archives/spr2022/entries/medicine/ [accessed 2023-08-02]

XSL•FO
RenderX

77. Abettan C, Welie JVM. The impact of twenty-first century personalized medicine versus twenty-first century medicine's impact on personalization. Philos Ethics Humanit Med 2020 Nov 18;15(1):11. [doi: 10.1186/s13010-020-00095-2] [Medline: 33203434]
78. Böhler D, Friebe M. (Digital) patient journey and empowerment: digital twin. In: Friebe M, editor. Novel Innovation Design for the Future of Health: Springer; 2022:169-178. [doi: 10.1007/978-3-031-08191-0_17]
79. Nye L. Digital twins for patient care via knowledge graphs and closed-form continuous-time liquid neural networks. arXiv. Preprint posted online on Jul 8, 2023. [doi: 10.48550/arXiv.2307.04772]
80. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018. [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
81. About Solid. Solid. URL: https://solidproject.org/about [accessed 2025-08-06]
82. Vaid A. Medical AI can transform medicine - but only if we carefully track the data it touches. Nature New Biol 2025 Jun;642(8069):864-866. [doi: 10.1038/d41586-025-01946-8] [Medline: 40550881]
83. Huang C, Ren X, Tang J, Yin D, Chawla N. Large language models for graphs: progresses and directions. Presented at: ACM Web Conference 2024; May 13-17, 2024. [doi: 10.1145/3589335.3641251]
84. Lu Z, Liu Y, Wen G, Zhou B, Zhang W, Zhang J. Noise-resistant graph neural networks with manifold consistency and label consistency. Expert Syst Appl 2024 Jul;245:123120. [doi: 10.1016/j.eswa.2023.123120]
85. Do you allow the use of ChatGPT or other generative language models and how should this be reported? JMIR Publications. 2025. URL: https://support.jmir.org/hc/en-us/articles/13387268671771-Do-you-allow-the-use-of-ChatGPT-or-other-generative-language-models-and-how-should-this-be-reported [accessed 2026-01-05]

Abbreviations

AHRQ: Agency for Healthcare Research and Quality
BERT: Bidirectional Encoder Representations from Transformers
DKG: Deutsche Krebsgesellschaft (German Cancer Society)
DT: digital twin
EHR: electronic health record
GNN: graph neural network
HCP: health care practitioner
HDS: health digital shadow
HDT: health digital twin
***ICD*:** *International Classification of Diseases*
ICHI: International Classification of Health Interventions
ICU: intensive care unit
KG: knowledge graph
LLM: large language model
LOS: length of stay
LxM: large x model
ML: machine learning
PBPK: physiologically based pharmacokinetic modeling
PCQC: patient-centered quality of care
QI: quality indicator
QM: quality metric

Edited by J Sarvestan; submitted 06.Aug.2025; peer-reviewed by A Shivanna, S Narayan, UK Chilakalapalli; revised version received 15.Jan.2026; accepted 16.Jan.2026; published 31.Mar.2026.

Please cite as:
Nitschke AK, Diaz Ochoa JG, Neumaier S, Knott M
From Knowledge Graphs to Digital Twins: Perspectives on Modeling Patient Outcomes for Health Care Quality Assessment
J Med Internet Res 2026;28:e81946
URL: https://www.jmir.org/2026/1/e81946
doi:10.2196/81946

© Anna-Katharina Nitschke, Juan G Diaz Ochoa, Simone Neumaier, Markus Knott. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.Mar.2026. This is an open-access article distributed under the terms of the Creative

XSL•FO
RenderX

Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Extrinsic Trust as a Contractual Framework for Accountable AI in Health Care: Viewpoint

Anthony Kelly[1,2], PhD

[1]Department of Electronic and Computer Engineering, University of Limerick, Castletroy, Limerick, Ireland

[2]Health Research Institute, University of Limerick, Limerick, Ireland

Corresponding Author:
Anthony Kelly, PhD
Department of Electronic and Computer Engineering, University of Limerick, Castletroy, Limerick, Ireland

Abstract

Artificial intelligence (AI) promises efficiency and equity in health care. However, adoption remains fragmented due to weak foundations of trust. This Viewpoint highlights the gap between intrinsic trust, based on interpretability, and extrinsic trust, based on functional validation. We propose a contractual framework between AI systems and users defined by 3 promises: reliability, scope and equity, and shift and uncertainty. Illustrated through a vignette, we show how health systems can operationalize these promises through structured evidence and governance, translating trustworthy AI into accountable clinical deployment.

(J Med Internet Res 2026;28:e83903) doi:10.2196/83903

KEYWORDS

artificial intelligence; AI; explainable artificial intelligence; explainable AI; XAI; trust; machine learning; ML; mental health; decision support system; intrinsic trust

Introduction

Artificial intelligence (AI) has the potential to transform clinical care, with applications spanning triage, diagnosis, risk prediction, and resource allocation [1]. However, despite this technical promise, real-world adoption remains fragmented. Trust is a central barrier to adoption as clinicians and decision-makers remain hesitant to rely on AI in safety-critical environments, where errors carry significant consequences for patient outcomes, liability, and public confidence [2].

Adoption requires structured, credible evidence. A recent scoping review identified impediments including limited external validation data, transparency and equity concerns, workflow integration difficulties, and unclear accountability and highlighted facilitators such as robust validation, governance, and postdeployment monitoring [3].

Conceptually, the gap between AI engagement and adoption corresponds to a distinction between intrinsic and extrinsic trust. Intrinsic trust is the sense that clinicians gain when AI outputs are interpretable and aligned with clinical reasoning; extrinsic trust (or functionality trust) [4] rests on empirical evidence that the model performs reliably across diverse populations and under real-world conditions [5]. Intrinsic trust is necessary to encourage engagement and build confidence in the system, but extrinsic trust ultimately influences deployment decisions [5,6].

Establishing extrinsic trust is fundamentally a data and infrastructure challenge. We argue that health systems require structured approaches to generate, verify, and act on validation evidence under real-world conditions. Accordingly, we introduce a contractual framing of extrinsic trust supported by a minimum evidence package and a clinical vignette. Rather than proposing a new reporting artifact, the framework complements existing tools such as model cards, algorithmic audits, and assurance cases by treating trust as a small set of explicit, testable promises. Each promise is linked to defined evidence requirements, verification roles, and breach conditions that trigger governance action across development, deployment, and postdeployment monitoring.

Intrinsic and Extrinsic Trust in Clinical AI

Clinicians engage with AI systems in stages, beginning with how outputs align with their professional reasoning [7]. This early stage of intrinsic trust develops when predictions are interpretable and presented in ways that resonate with clinical heuristics [8-10].

Intrinsic trust provides a foundation for engagement, but decisions about deployment depend on further evidence. Clinicians require independent validation that the model performs reliably across different patient groups and that it can indicate when deference to human judgment is warranted in the face of predictive uncertainty [11-15]. Interpretable AI models are central to the sensemaking involved in intrinsic trust, but decisions about adoption require evidence of performance under real-world conditions [16]. To express this issue more succinctly, explainability fosters engagement, but validation governs deployment [5,12,14]. Intrinsic trust lowers the barrier to initial use, while extrinsic trust provides the assurance needed

XSL•FO
RenderX

for safe integration into workflows and health system decision-making.

Extrinsic trust validates that the model can generalize from the training examples in the dataset to previously unseen examples. As such, it must be either invariant to shift in the statistical distribution of the unseen data or convey increasing uncertainty commensurate with the shift. Since supervised machine learning models are known to be sensitive to out-of-distribution data shift, appropriate validation involves verifying that the model's sensitivity to out-of-distribution data shift is reliably conveyed through an appropriate uncertainty metric. Shifts within the distribution may occur in subgroups of the data, for example, in demographic or clinical subtypes. Appropriate validation for such in-distribution shifts involves verifying performance invariance. AI model calibration is concerned with how well a model's predicted probabilities align with actual outcomes and is therefore the foundation of reliability [17]. Calibration provides a foundation for trust by enabling users to gauge whether the AI "knows when it knows." Without external validation, especially under real-world conditions where data distributions may shift, clinicians may withhold trust or misplace it, leading to underuse or overreliance [12,14,18].

Consequently, the following issues need to be addressed through validation:

- The model should be reliable for in-distribution unseen data examples, demonstrating that predicted probabilities align with actual outcomes.
- Reliability should be maintained for in-distribution subgroups.
- Uncertainty metrics should convey higher uncertainty for out-of-distribution data, demonstrating that the model "knows when it knows."

Together, these questions define a pragmatic minimum for extrinsic trust rather than a complete taxonomy. They reflect 3 recurrent failure modes in clinical AI deployment: miscalibrated confidence, uneven subgroup performance, and silent degradation under distributional shift. Reliability and uncertainty are complementary. Reliability characterizes confidence alignment within scope, while uncertainty governs safe behavior as that alignment weakens. Scope and equity are treated jointly because subgroup shifts represent clinically salient in-distribution variation that population-level metrics can obscure.

Although these requirements are most relevant for discriminative AI models that predict classifications, there is emerging evidence that generative AI text models (large language models) can be queried about their reliability and uncertainty [19], perhaps allowing for future application of this framework.

Core Data Requirements for Extrinsic Trust

Similarly to trust in people, trust in AI has been framed as a contract [11,20,21]. Following this notion of contractual trust, users can rely on an AI system when explicit promises are stated and kept. Therefore, users must know what the AI is being trusted with and possess a means of evaluating whether the contract is adhered to [11]. This adherence depends on the context, and the means of determining adherence are dependent on the nature of the contract [8,11]. Drawing on these validation questions, we proposed 3 evidence clauses that should be adhered to for safe deployment (Figure 1) and detail the required metrics in Textbox 1.

XSL•FO
RenderX

Figure 1. Concentric framework of extrinsic trust. The 3 contractual promises of trust—reliability, scope and equity, and shift and uncertainty—are represented as radial sectors. Each promise operates across 3 levels: model-level evidence (inner ring), user-level trust and workflow fit (middle ring), and system-level governance and accountability (outer ring). At the core lies extrinsic trust, integrating these dimensions into the basis for responsible clinical artificial intelligence adoption. ECE: expected calibration error; SCE: static calibration error.

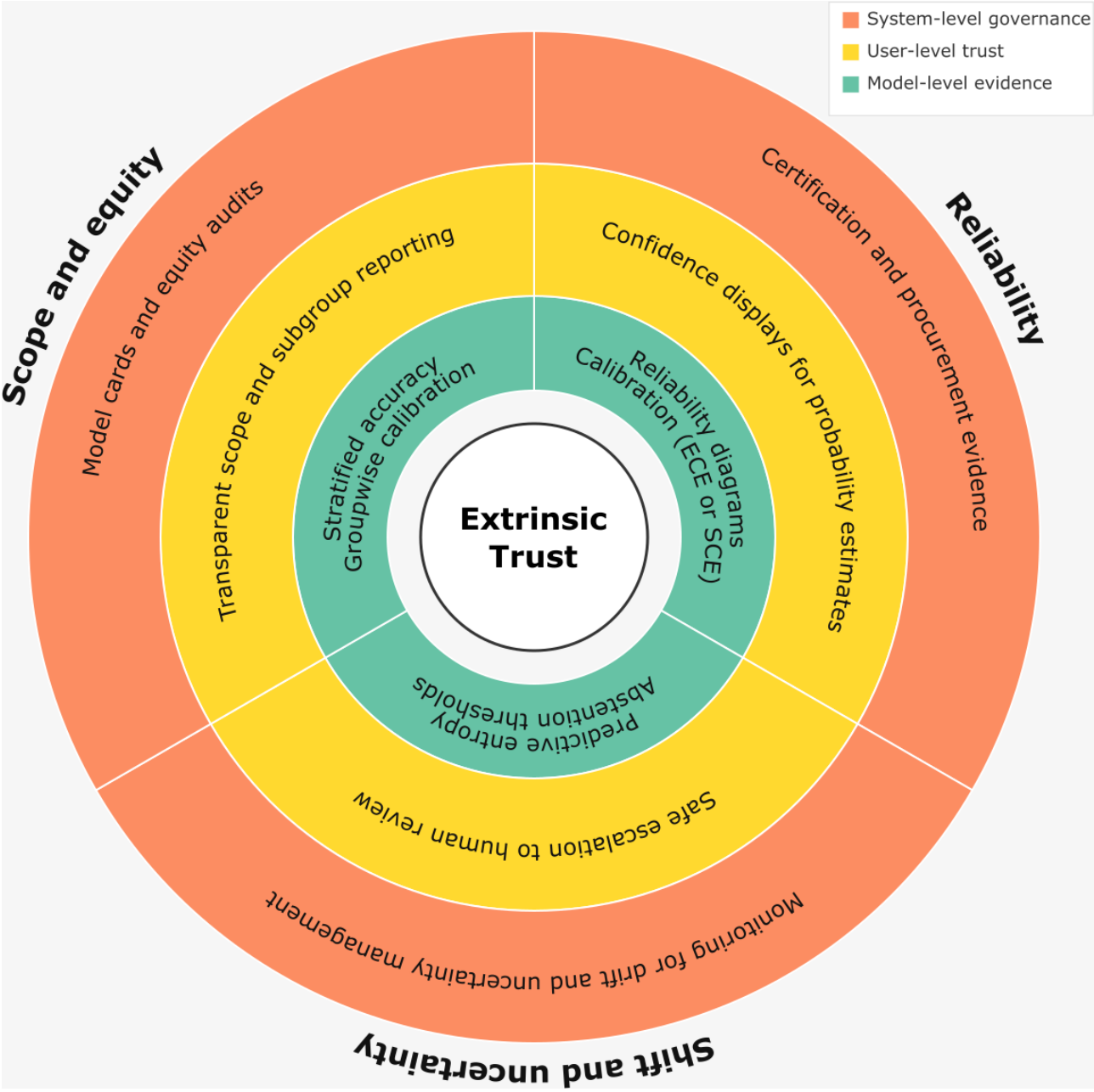

XSL•FO
RenderX

Textbox 1. Metrics for operationalizing the 3 promises of extrinsic trust in artificial intelligence systems.

Reliability promise

- Population-level performance metrics (eg, sensitivity, area under the receiver operating characteristic curve, and F_1-score)
- Expected calibration error or static calibration error: scalar measures of how closely predicted probabilities align with observed outcomes
- Reliability diagrams: visual summaries that clinicians can interpret to judge whether model confidence reflects real-world accuracy

Scope and equity promise

- Stratified performance metrics (eg, sensitivity, area under the receiver operating characteristic curve, and F_1-score): reported across clinically and demographically relevant subgroups to detect disparities
- Groupwise calibration plots: showing whether confidence is systematically over or underestimated for particular populations

Shift and uncertainty promise

- Negative log-likelihood: evaluates the quality of probabilistic predictions under noise or perturbation
- Predictive entropy or abstention rate: label-agnostic measures of uncertainty, useful for detecting out-of-distribution cases and supporting safe deferral to human review

The term "contractual" refers to a professional and governance compact rather than a legal indemnity framework and does not define liability. It makes explicit the performance promises that an AI system claims to uphold, the evidence required to support them, and the governance actions that follow if they are breached, thereby informing procurement and oversight.

In this compact, developers or vendors make explicit performance claims, health systems commission and govern verification, and clinicians rely on outputs under defined workflow conditions. A promise is meaningful only if keeping or breaching it is observable through agreed artifacts and escalation rules.

Recent work on trust in clinical AI emphasizes that trust is relational and bidirectional, emerging from interactions among models, clinicians, workflows, and institutional safeguards rather than from model properties alone. From this perspective, AI systems depend on human inputs, data quality, and organizational practices to remain safe and effective over time [22]. The contractual promises proposed in this paper are intended to complement this view by specifying the minimum evidence and governance commitments required to sustain trust within such sociotechnical systems, particularly through algorithmic audits and postdeployment monitoring.

Reliability Promise

The first requirement is a reliability promise: an AI system should provide probability estimates that clinicians can trust in the intended use context. High-confidence predictions should typically be correct, and uncertainty should be visible when the model is unsure.

Reliability depends on calibration: a well-calibrated model produces probability estimates that align with observed outcomes, giving clinicians a clear sense of when the system's confidence can be relied upon [23,24]. This alignment is crucial in safety-critical settings, where misplaced confidence can lead to harmful decisions. Tools such as reliability diagrams and scalar metrics such as the expected calibration error or its multiclass analogue, the static calibration error (SCE), provide concise ways of demonstrating this alignment by showing a good fit of the reliability curve to the diagonal and low values for the metrics (perfectly calibrated models have an SCE or expected calibration error of 0) [17,25,26]. Monitoring changes in these metrics or the reliability curve fit is useful as acceptable threshold levels are context dependent. Postcalibration scaling can further improve poorly calibrated models [26,27].

For health systems, requiring calibration evidence ensures that AI models offer dependable guidance that supports safe decision-making [12].

Scope and Equity Promise

The second requirement is a scope and equity promise: AI systems should clearly state who and what they are designed for and then provide evidence that they perform consistently across relevant clinical and demographic subgroups. Declaring scope makes explicit the intended populations, settings, and workflow assumptions, while equity validation ensures that no patient group is systematically disadvantaged.

Validation of the scope and equity promise involves stratified evaluation. Standard performance measures such as accuracy, sensitivity, area under the receiver operating characteristic curve, or F_1-score [28] should be reported not just at the population level but separately for subgroups defined by age, sex, ethnicity, or clinical subtype [29]. Calibration should also be examined within each stratum to detect systematic over- or underestimation. These analyses highlight where models are consistent and where subgroup-specific recalibration or data augmentation may be required. It should be acknowledged that small sample sizes may result from intersectional subgroups that may limit the ability to perform subgroup analysis in some cases.

From a regulatory and governance perspective, AI model cards offer a practical tool for this promise. They function as structured "labels," detailing intended use cases, target populations, known limitations, subgroup performance, and fairness considerations [30]. This transparency allows clinicians,

XSL•FO
RenderX

procurement teams, and regulators to assess whether systems are safe and appropriate for deployment.

Equity is fundamental because health populations are heterogeneous. A model that performs well overall may still underperform in underrepresented groups, amplifying disparities in care [12-14]. Mandating scope declarations and subgroup validation enables health systems to recognize both strengths and limits, supporting targeted safeguards where needed.

Shift and Uncertainty Promise

The third requirement is a shift and uncertainty promise: AI systems must provide safeguards when confronted with inputs that differ from the data on which they were trained. In evolving clinical environments, populations, referral patterns, and data quality change over time. Without mechanisms to detect and manage such shifts, models risk silent failure in precisely the contexts where reliability is most critical [12,31].

This promise requires evidence that the system can signal uncertainty, degrade gracefully, and defer to human review when appropriate. Metrics such as negative log-likelihood or predictive entropy capture probabilistic quality and uncertainty in unfamiliar cases [26,32,33]. In practice, confidence should decline, and error signals should increase in a predictable manner when inputs deviate or contain errors.

By enabling abstention or escalation under uncertainty, models help protect patients from misleading outputs, preserve equity across heterogeneous populations, and reduce risks of overreliance [34]. Embedding uncertainty clauses into validation frameworks ensures that decision-makers have the information to guarantee that AI remains trustworthy under real-world conditions.

Minimum Evidence Package for Extrinsic Trust

To formalize the contractual framework, we outline a minimum evidence package (Table 1) that specifies what constitutes proof of each promise, who is responsible for generating and verifying that proof, and what breach conditions require mitigation or suspension of deployment. In many cases, there is no accepted threshold for many of the metrics that constitute proof, although statistical threshold limits may be calculated on held-out calibration sets. For example, a material breach may be defined as a statistically significant upward shift in calibration error relative to the validated deployment baseline as determined by locally agreed monitoring procedures. Therefore, the minimum evidence package provides the foundation for a contract between the AI system and its users: a shared understanding that the model's behavior is recorded and that deviations can trigger governance actions to mitigate emerging risks before they manifest in clinical practice.

XSL•FO
RenderX

Table . Minimum evidence package for extrinsic trust. Baselines refer to the model's validated performance at deployment or the last approved update, with material deviation determined relative to locally defined governance thresholds.

Promise	Artifacts	Producer or verifier	Breach condition and action
All	• Model card documenting intended population, features, and workflow assumptions	• Producer: vendor or developer • Verifier: health system governance and clinical safety officer • Trigger: model commissioning and model update	• Breach: model card not satisfactory • Action: remedial documentation or clarification required before deployment or update
Reliability	• Population-level performance metrics (eg, sensitivity, AUC[a], and F_1-score) • Calibration metrics: ECE[b] or SCE[c] • Reliability diagrams	• Producer: vendor or developer • Verifier: health system governance and clinical safety officer • Trigger: annually; model update or change in patient profiles	• Breach: performance or calibration metrics deviate materially from the validated deployment baseline • Action: (1) recalibration or retraining, (2) investigation of the root cause, or (3) possible suspension
Scope and equity	• Stratified performance metrics (eg, subgroup: sensitivity, AUC, and F_1-score) • Groupwise calibration plots	• Producer: vendor or developer • Verifier: health system governance and clinical safety officer • Trigger: annually; model update or change in patient profiles	• Breach: subgroup performance or calibration deviates materially from the validated deployment baseline • Action: (1) subgroup analysis, (2) recalibration, (3) model card update to exclude subgroups, or (4) possible suspension
Shift and uncertainty	• Predictive entropy or abstention rate • NLL[d]	• Producer: vendor or developer • Verifier: health system governance and clinical safety officer • Trigger: annually; model update or change in patient profiles	• Breach: uncertainty metrics (eg, entropy or NLL) deviate materially from the validated deployment baseline • Action: (1) retraining of model with new data, (2) reverification of model, (3) model card update, or (4) possible suspension

[a]AUC: area under the receiver operating characteristic curve.

[b]ECE: expected calibration error.

[c]SCE: static calibration error.

[d]NLL: negative log-likelihood.

For clinicians and health system decision-makers, the contractual framing also provides a practical cognitive benefit. By organizing technical, regulatory, and statistical requirements into a small number of clear promises, it offers a concise mental model for what trustworthy AI deployment entails, helping cut through the complexity of standards and guidance.

The 3 promises align closely with emerging international regulatory requirements for the analysis and management of AI risks. For example, the European Union AI Act [35] requires transparency (Article 13), accuracy and robustness (Article 15), and risk monitoring (Articles 9 and 72). Although this applies primarily to AI medical devices [36], it nevertheless also implies best practice for lower-risk AI. Table 2 maps each promise to obligations in the European Union AI Act [35], Food and Drug Administration predetermined change control plan [37], and International Organization for Standardization and International Electrotechnical Commission 42001 standard [38], demonstrating that the proposed contractual framework provides a practical implementation framework consistent with regulatory trajectories in high-risk clinical AI.

XSL•FO
RenderX

Table . Mapping the 3 promises to emerging regulatory requirements.

Promise	EU[a] AI[b] Act (Regulation [EU] 2024/1689)	FDA[c] PCCP[d]	ISO[e] and IEC[f] 42001:2023 standard (AI management system)
Reliability promise	• Article 15: high-risk AI must achieve appropriate accuracy, robustness, and cybersecurity, and AI providers must declare accuracy metrics and levels and test or validate them • Article 9: requires an ongoing risk management system • Article 11+annex IV: technical documentation must include information such as performance characteristics and metrics and testing or validation evidence (as specified in annex IV)	The PCCP comprises (1) description of modifications, (2) modification protocol, and (3) impact assessment. The modification protocol should specify verification and validation activities, including performance evaluation methods, performance metrics, statistical tests, and predefined acceptance criteria, and describe how the manufacturer will determine that a modification is acceptable before implementation.	Clause 8.1 (operational planning and control) requires planned and controlled AI operations. Clause 9.1 (monitoring, measurement, analysis, and evaluation) requires organizations to determine what needs to be monitored and measured and how and when to evaluate AI system performance. Clause 10 (improvement) requires correction and continuous improvement when performance deviates.
Scope and equity promise	• Article 10: training, validation, and testing data must be sufficiently relevant and representative for the intended purpose, and AI providers must examine and mitigate bias risks • Article 13: transparency and instructions enabling appropriate use, including information needed for deployers to interpret outputs and use the system properly	The evidence and testing data supporting the PCCP should be representative of intended use populations (eg, demographic factors where relevant). The PCCP and its modifications must keep the device within the authorized intended use and indications; a change to what was described or authorized (eg, new population not covered) may require a new marketing submission.	Clauses 4.1 - 4.3 require definition of organizational context, interested parties, and the scope of the AI management system. Clause 6.1 (actions to address risks and opportunities) requires identification and treatment of AI risks, including risks to individuals and groups. Clause 6.1.4 (AI system impact assessment) requires assessment of impacts on individuals and groups affected by AI systems.
Shift and uncertainty promise	• Article 72: AI providers must operate a postmarket monitoring system to collect, document, and analyze performance and risks during real-world use • Article 15: robustness expectations (accuracy, robustness, and cybersecurity) support managing performance degradation in foreseeable conditions	The PCCP's modification protocol should define how the manufacturer will detect the need for change and evaluate modifications; examples include identifying triggers, such as when drift in data is observed. The protocol should specify how performance will be assessed after the change (metrics, tests, or acceptance criteria) and state that modifications should not be implemented if acceptance criteria are not met.	Clause 9.1 requires ongoing monitoring and evaluation of AI systems. Clause 9.2 (internal audit) requires periodic audits to detect nonconformity. Clause 10.2 (nonconformity and corrective action) requires organizations to respond to deviations, investigate causes, and implement corrective actions supporting structured responses to model drift and performance degradation.

[a]EU: European Union.

[b]AI: artificial intelligence.

[c]FDA: Food and Drug Administration.

[d]PCCP: predetermined change control plan.

[e]ISO: International Organization for Standardization.

[f]IEC: International Electrotechnical Commission.

Case Study Vignette: Operationalizing the 3 Promises Using the Probabilistic Integrated Scoring Model

To illustrate how extrinsic trust can be operationalized, we turn to the Probabilistic Integrated Scoring Model (PrISM) [39], a prototype AI classification system developed for initial treatment screening in routine mental health care. PrISM is a transparent, interpretable neural network that integrates the structured logic of psychometric instruments with probabilistic estimation.

The PrISM study involved a secondary analysis of questionnaire and demographic data from 1068 patients treated in Internetpsykiatrien, Denmark's national internet-based mental health service. Further details related to this vignette can be found in Multimedia Appendix 1.

Regarding the reliability promise, reliability diagrams show how closely the model's predicted probabilities align with observed outcomes. Figure 2 shows that PrISM confidence estimates remain near the ideal diagonal, with an SCE of 4.10%. For a clinician reviewing a referral, this provides assurance that, when the model appears confident, it is generally correct,

XSL•FO
RenderX

supporting decisions about when to trust AI guidance and when to seek further input.

Figure 2. Illustrative example of a reliability diagram for a predictive multiclass model (Probabilistic Integrated Scoring Model). The static calibration error (SCE) of 4.10% and close adherence to the diagonal indicate that the predicted probabilities closely match observed outcomes.

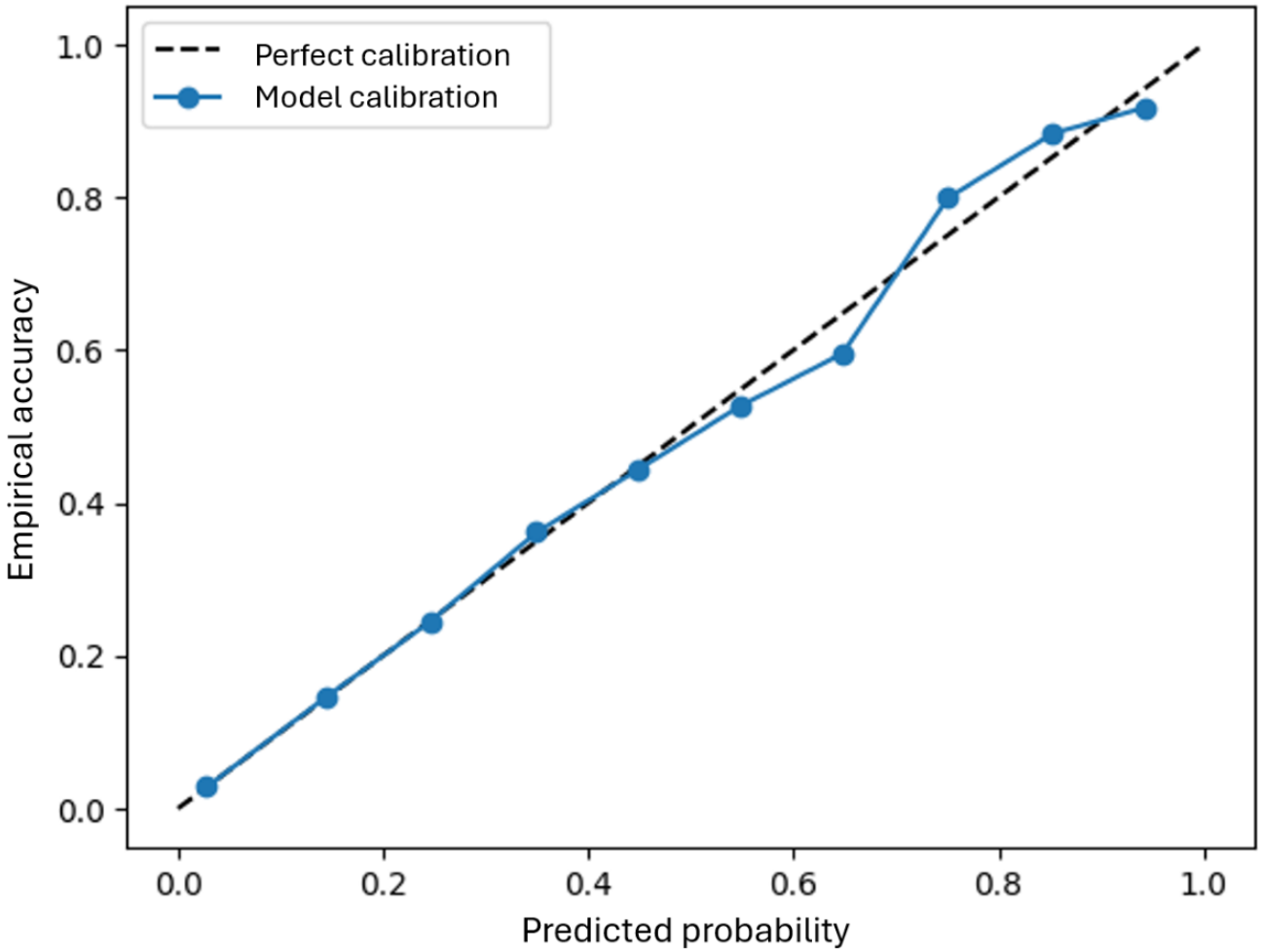

Regarding the scope and equity promise, PrISM is explicitly designed and framed as a tool for screening referrals in a digital clinic for mild to moderate depression and anxiety, clearly stating who and what it is designed for. Validation includes stratified calibration and discrimination (eg, by sex or presenting problem), with deployment contingent on subgroup performance remaining within a predefined deviation from the population baseline. This analysis highlights where performance is consistent and where disparities might emerge, enabling clinicians and service managers to anticipate limitations and plan mitigations rather than encountering them unexpectedly in practice.

Regarding the shift and uncertainty promise, in deployment, AI systems encounter cases that differ from their training data. PrISM addresses this by signaling uncertainty. When data inputs contain errors or deviate from their familiar distribution, the uncertainty metric (negative log-likelihood) rises, and the system may abstain from issuing a recommendation, routing the case for human review (Figure 3). For example, cases exceeding the 95th percentile of validation set uncertainty may trigger abstention and clinician review, with abstention rates monitored longitudinally as a safety signal. For clinicians, this safeguard functions as a clear indicator that the model recognizes its limits, ensuring that safe escalation is possible.

XSL•FO
RenderX

Figure 3. Illustrative example of deterioration of uncertainty (Probabilistic Integrated Scoring Model). As data noise increases (x-axis), the negative log-likelihood (NLL) uncertainty metric increases, whereas the accuracy performance metric stays relatively constant. This illustrates that the model's performance may be robust to data errors, but the model signals its increasing lack of confidence appropriately.

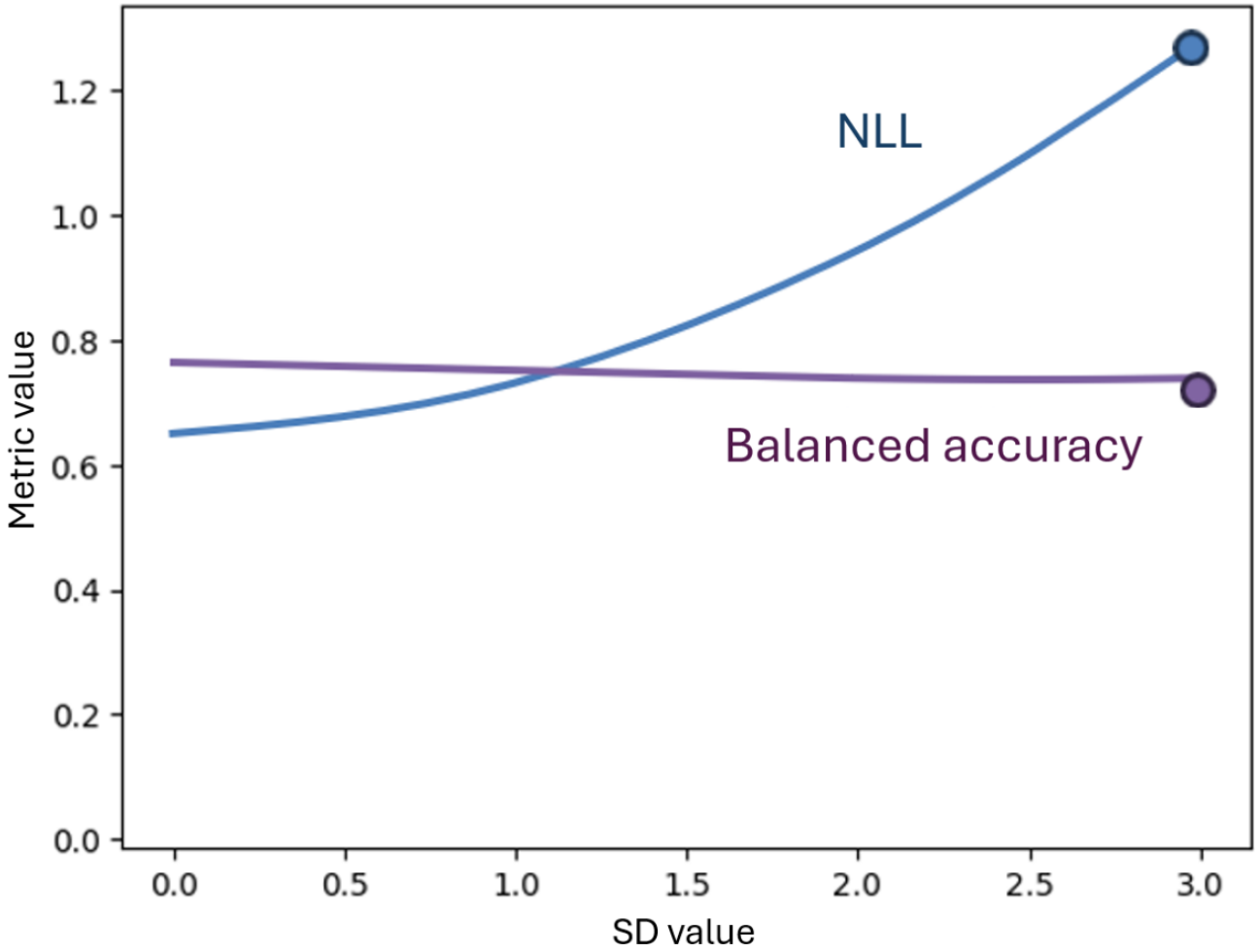

This vignette demonstrates that extrinsic trust can be embedded through design choices that make reliability, scope, and uncertainty measurable. The PrISM vignette shows how trust promises can be translated into clinical evidence, allowing health systems to evaluate, monitor, and govern AI as accountable decision support tools.

Health System Implications and Future Directions

Embedding extrinsic trust requires health systems to translate the 3 contractual promises into routine governance and clinical practice. In practical terms, this means requiring calibration, subgroup performance, and uncertainty evidence at procurement; defining verification roles within clinical safety structures; and specifying escalation actions when promises are breached.

Evidence of trust must be accessible and actionable at the point of care. Simple signals such as confidence or abstention indicators can reduce cognitive burden for frontline clinicians, while dashboards can surface detailed calibration, equity, and drift metrics for governance teams. When systems defer under uncertainty, clear escalation pathways are needed to avoid automation bias or clinician deskilling and preserve human oversight.

Looking ahead, this contractual framing aligns with regulatory shifts toward continuous assurance rather than one-time approval. By treating reliability, scope and equity, and shift and uncertainty as enforceable commitments, health systems gain a lightweight but robust mechanism to govern clinical AI across its life cycle as real-world conditions evolve.

Conclusions

Extrinsic trust is the decisive factor that determines whether clinical AI remains experimental or becomes a sustainable component of care. While intrinsic trust, supported by interpretability, enables clinician engagement, adoption in safety-critical settings depends on structured evidence of performance under real-world conditions.

By framing extrinsic trust as a contractual relationship defined by 3 promises of reliability, scope and equity, and shift and uncertainty, this Viewpoint provides a practical lens for evaluating, governing, and sustaining clinical AI systems. The PrISM vignette demonstrates that these promises can be operationalized through calibration, subgroup validation, and uncertainty-aware escalation. Together, this approach shifts trust from an abstract aspiration to a testable, accountable basis for responsible clinical AI deployment.

XSL•FO
RenderX

Acknowledgments

The author would like to express gratitude to the researchers and staff of Internetpsykiatrien, Denmark's national internet-based mental health service, whose conversations and interactions helped inspire this viewpoint. Generative artificial intelligence was used to improve clarity and language in some author-written paragraphs.

Funding

The author is funded by Innovation Fund Denmark.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to ethical and legal constraints surrounding patient data but can be made available from the corresponding author on reasonable request.

Authors' Contributions

AK conceived the viewpoint, developed the models and results, and wrote the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details of the Probabilistic Integrated Scoring Model vignette.
[DOCX File, 40 KB - jmir_v28i1e83903_app1.DOCX]

References

1. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
2. Afroogh S, Akbari A, Malone E, Kargar M, Alambeigi H. Trust in AI: progress, challenges, and future directions. Humanit Soc Sci Commun 2024;11(1):1568. [doi: 10.1057/s41599-024-04044-8]
3. Hassan M, Kushniruk A, Borycki E. Barriers to and facilitators of artificial intelligence adoption in health care: scoping review. JMIR Hum Factors 2024 Aug 29;11:e48633. [doi: 10.2196/48633] [Medline: 39207831]
4. Choung H, David P, Ross A. Trust and ethics in AI. AI Soc 2023 Apr;38(2):733-745. [doi: 10.1007/s00146-022-01473-4]
5. Shin D. The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. Int J Hum Comput Stud 2021 Feb;146:102551. [doi: 10.1016/j.ijhcs.2020.102551]
6. Choung H, David P, Ross A. Trust in AI and its role in the acceptance of AI technologies. Int J Hum Comput Interact 2023 May 28;39(9):1727-1739. [doi: 10.1080/10447318.2022.2050543]
7. Lipton ZC. The mythos of model interpretability. Commun ACM 2018 Sep 26;61(10):36-43. [doi: 10.1145/3233231]
8. Kelly A, Bhardwaj N, Holmberg Sainte-Marie TT, et al. Investigating how clinicians form trust in an AI-based mental health model: qualitative case study. JMIR Hum Factors 2025 Dec 19;12:e79658. [doi: 10.2196/79658] [Medline: 41417472]
9. Ali S, Abuhmed T, El-Sappagh S, et al. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. Inf Fusion 2023 Nov;99:101805. [doi: 10.1016/j.inffus.2023.101805]
10. Joyce DW, Kormilitzin A, Smith KA, Cipriani A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. NPJ Digit Med 2023 Jan 18;6(1):6. [doi: 10.1038/s41746-023-00751-9] [Medline: 36653524]
11. Jacovi A, Marasović A, Miller T, Goldberg Y. Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI. Presented at: FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; Mar 3-10, 2021. [doi: 10.1145/3442188.3445923]
12. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digit Med 2021 Jan 5;4(1):4. [doi: 10.1038/s41746-020-00367-3] [Medline: 33402680]
13. Malinin A, Gales M. Predictive uncertainty estimation via prior networks. Presented at: NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems; Dec 3-8, 2018. [doi: 10.5555/3327757.3327808]
14. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. Nat Mach Intell 2019;1(1):20-23. [doi: 10.1038/s42256-018-0004-1]
15. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci Rep 2016 May 17;6:26094. [doi: 10.1038/srep26094] [Medline: 27185194]
16. Gawlikowski J, Tassi CRN, Ali M, et al. A survey of uncertainty in deep neural networks. Artif Intell Rev 2023 Oct;56(S1):1513-1589. [doi: 10.1007/s10462-023-10562-9]

XSL•FO
RenderX

17. Riley RD, Archer L, Snell KIE, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ 2024 Jan 15;384:e074820. [doi: 10.1136/bmj-2023-074820] [Medline: 38224968]
18. Zhu F, Zhang XY, Cheng Z, Liu CL. Revisiting confidence estimation: towards reliable failure prediction. IEEE Trans Pattern Anal Mach Intell 2024 May;46(5):3370-3387. [doi: 10.1109/TPAMI.2023.3342285] [Medline: 38090830]
19. Wei J, Karina N, Chung HW, et al. Measuring short-form factuality in large language models. arXiv. Preprint posted online on Nov 7, 2024. [doi: 10.48550/arXiv.2411.04368]
20. Hawley K. Trust, distrust and commitment. Nous 2014 Mar;48(1):1-20. [doi: 10.1111/nous.12000]
21. Tallant J, Donati D. Trust: from the philosophical to the commercial. Philos Manag 2020 Mar;19(1):3-19. [doi: 10.1007/s40926-019-00107-y]
22. Sagona M, Dai T, Macis M, Darden M. Trust in AI-assisted health systems and AI's trust in humans. NPJ Health Syst 2025;2(1):10. [doi: 10.1038/s44401-025-00016-5]
23. Nixon J, Dusenberry MW, Jerfel G, Zhang L, Tran D. Measuring calibration in deep learning. Presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019); Jun 15-20, 2019 URL: https://openaccess.thecvf.com/content_CVPRW_2019/html/Uncertainty_and_Robustness_in_Deep_Visual_Learning/Nixon_Measuring_Calibration_in_Deep_Learning_CVPRW_2019_paper.html [accessed 2026-02-26]
24. Ovadia Y, Fertig E, Ren J, et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Presented at: NIPS'19: 33rd International Conference on Neural Information Processing Systems; Dec 8-14, 2019. [doi: 10.5555/3454287.3455541]
25. Naeini MP, Cooper GF, Hauskrecht M. Obtaining well calibrated probabilities using Bayesian binning. Proc AAAI Conf Artif Intell 2015 Jan;2015:2901-2907 [FREE Full text] [Medline: 25927013]
26. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Presented at: 34th International Conference on Machine Learning (ICML 2017); Aug 6-11, 2017 URL: https://proceedings.mlr.press/v70/guo17a.html [accessed 2026-02-26]
27. Laves MH, Ihler S, Kortmann KP, Ortmaier T. Well-calibrated model uncertainty with temperature scaling for dropout variational inference. Presented at: 4th Workshop on Bayesian Deep Learning (NeurIPS 2019); Dec 13-14, 2019 URL: https://www.bayesiandeeplearning.org/2019/papers/77.pdf [accessed 2026-02-26]
28. Hicks SA, Strümke I, Thambawita V, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep 2022 Apr 8;12(1):5979. [doi: 10.1038/s41598-022-09954-8] [Medline: 35395867]
29. James G, Witten D, Hastie T, Tibshirani R, Taylor J. An Introduction to Statistical Learning: With Applications in Python: Springer; 2023. URL: https://link.springer.com/book/10.1007/978-3-031-38747-0 [doi: 10.1007/978-3-031-38747-0]
30. Gilbert S, Adler R, Holoyad T, Weicken E. Could transparent model cards with layered accessible information drive trust and safety in health AI? NPJ Digit Med 2025 Feb 25;8(1):124. [doi: 10.1038/s41746-025-01482-9] [Medline: 40000736]
31. Ku WL, Min H. Evaluating machine learning stability in predicting depression and anxiety amidst subjective response errors. Healthcare (Basel) 2024 Mar 10;12(6):625. [doi: 10.3390/healthcare12060625] [Medline: 38540589]
32. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); Dec 4-9, 2017. [doi: 10.5555/3295222.3295387]
33. Rasmussen CE, Quiñonero-Candela J, Sinz F, Bousquet O, Schölkopf B. Evaluating predictive uncertainty challenge. In: Quiñonero-Candela J, Dagan I, Magnini B, d'Alché-Buc F, editors. Machine Learning Challenges Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment: Springer; 2006:1-27. [doi: 10.1007/11736790_1]
34. He W, Jiang Z, Xiao T, Xu Z, Li Y. A survey on uncertainty quantification methods for deep learning. ACM Comput Surv 2023;37(4):111 [FREE Full text]
35. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). European Union. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng [accessed 2026-02-27]
36. van Kolfschooten H, van Oirschot J. The EU Artificial Intelligence Act (2024): implications for healthcare. Health Policy 2024 Nov;149:105152. [doi: 10.1016/j.healthpol.2024.105152] [Medline: 39244818]
37. Predetermined change control plans for medical devices. U.S. Food and Drug Administration. 2024. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/predetermined-change-control-plans-medical-devices [accessed 2026-02-27]
38. ISO/IEC 42001:2023 - information technology — artificial intelligence — management system. International Organization for Standardization. 2023. URL: https://www.iso.org/standard/42001 [accessed 2026-02-27]
39. Kelly A, Jensen EK, Grua EM, Mathiasen K, Van de Ven P. An interpretable model with probabilistic integrated scoring for mental health treatment prediction: design study. JMIR Med Inform 2025 Mar 26;13:e64617. [doi: 10.2196/64617] [Medline: 40138679]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
PrISM: Probabilistic Integrated Scoring Model
SCE: static calibration error

Edited by A Coristine; submitted 10.Sep.2025; peer-reviewed by S Mohanadas, T Dai; revised version received 02.Feb.2026; accepted 04.Feb.2026; published 05.Mar.2026.

Please cite as:
Kelly A
Extrinsic Trust as a Contractual Framework for Accountable AI in Health Care: Viewpoint
J Med Internet Res 2026;28:e83903
URL: https://www.jmir.org/2026/1/e83903
doi:10.2196/83903

© Anthony Kelly. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Twins for Just-in-Time Adaptive Interventions (JITAIs): Framework for Optimizing and Continually Improving JITAIs

Asim H Gazi[1], PhD; Daiqi Gao[1], PhD; Susobhan Ghosh[1], BS; Ziping Xu[1], PhD; Anna L Trella[1], BS; Predrag Klasnja[2], PhD; Susan A Murphy[1], PhD

[1]School of Engineering and Applied Sciences, Harvard University, 29 Oxford Street, Cambridge, MA, United States

[2]School of Information, University of Michigan–Ann Arbor, Ann Arbor, MI, United States

Corresponding Author:
Asim H Gazi, PhD
School of Engineering and Applied Sciences, Harvard University, 29 Oxford Street, Cambridge, MA, United States

Abstract

In the context of digital health, just-in-time adaptive interventions (JITAIs) are nascent precision medicine systems that can extend personalized health care support to everyday life. A challenge in designing JITAIs is that personalized support often involves sophisticated decision-making algorithms. These decision-making algorithms can require numerous nontrivial design decisions that must be made between successive JITAI deployments (eg, hyperparameter selection for an artificial intelligence algorithm). Making design decisions between deployments—rather than during deployment—ensures intervention fidelity and enhances the ability to replicate results. Yet, each deployment can be costly, precluding the use of A/B testing for every design decision. How should design decisions be made strategically between JITAI deployments? This paper introduces "digital twins for just-in-time adaptive interventions (JITAI-Twins)" to address this question. JITAI-Twins are "digital twins of a subpopulation" (term used in the 2023 National Academies workshop proceedings on digital twins). JITAI-Twins are used to virtually simulate the potential outcomes of a JITAI's design decisions for an upcoming deployment. Based on simulation results, design decisions are made for the deployed JITAI. To continually improve the JITAI, data collected during deployment are used to update the JITAI-Twin—and this bidirectional feedback between deployments and simulation environments continues. JITAI-Twins are thus "fit-for-purpose" (term used in the National Academies 2024 consensus report on digital twins) instantiations of the digital twin concept. In this paper, we elucidate the specifics and design process of JITAI-Twins, with examples of prior use in clinical settings. JITAI-Twins highlight continuity over the course of a JITAI's optimization and continual improvement, emphasizing the need for bidirectional feedback between versions of a simulation environment and a JITAI's deployments.

(J Med Internet Res 2026;28:e72830) doi:10.2196/72830

KEYWORDS

digital twin; just-in-time adaptive intervention; mobile health; digital health; reinforcement learning; artificial intelligence; machine learning; precision medicine; simulation; deployment; JITAI; mHealth; AI; RL

Introduction

Motivation

Just-in-time adaptive interventions (JITAIs) are promising intervention systems that have the potential to impact mobile health (mHealth) [1,2]. By leveraging wireless technologies such as wearable sensors to infer when an individual is at risk for negative change and/or receptive to positive change [3], JITAIs can proactively provide support in everyday life that is adapted to an individual's health state or environmental context via "decision rules" [4]. Decision rules specify how intervention delivery should be adapted depending on "tailoring variables" such as the weather outside or an individual's heart rate.

While adaptive intervention delivery can improve outcomes [5], it is nontrivial to design effective decision rules, and their design has increasingly relied on sophisticated artificial intelligence (AI) and control engineering methods. These methods optimize decision rules to better match the target individual or sample of individuals, rather than individuals from whom previous data were collected [6,7]. To use AI and control engineering methods, nontrivial design decisions must often be made, including which tailoring variables to use, which AI or control engineering method to use, or what hyperparameters should be selected for an AI algorithm [8].

Two competing forces make it challenging to make such design decisions for a JITAI. First, deployments can be costly and are critical to a JITAI's continued development, especially in clinical settings [9,10]. It is imperative that a JITAI performs as optimally as possible when deployed. Second, JITAIs must remain autonomous during each deployment [7]. It is unscalable to require manual modifications to a JITAI's decision-making algorithms during deployment, and unplanned modifications can compromise intervention fidelity and replicability [11,12].

XSL•FO
RenderX

The desire for a JITAI to perform optimally must be balanced with the need for a JITAI to operate autonomously—motivating the development of methods to evaluate and make informed design decisions before a JITAI is deployed.

Digital Twins for Just-in-Time Adaptive Interventions

In this paper, we introduce the framework "digital twin for just-in-time adaptive intervention (JITAI-Twin)" and elucidate how the JITAI-Twin framework can be used to strategically make design decisions prior to a JITAI's deployments (Figure 1). As discussed in a recent Consensus Report from the National Academies [13], digital twins are characterized by two components: (1) the "modeling and simulation to create a virtual representation of a physical counterpart" and (2) a "bidirectional interaction between the virtual and the physical." This bidirectional interaction refers to the use of the virtual twin to inform decisions made for the physical twin, and the use of data from the physical twin to inform the virtual twin. JITAI-Twins include both virtual representation and bidirectional feedback components. JITAI-Twins take data from a JITAI's target subpopulation (physical twin) and construct a set of virtual approximations to the subpopulation's behavior (virtual twin). These virtual approximations are computational models of the subpopulation's behavior when the JITAI is deployed (also known as "simulation environments"). Computational models are fit using prior deployment data and then used to simulate the subpopulation's behavior. Simulations help evaluate multiple sets of potential design decisions (ie, candidate decision-making algorithms), to evaluate which candidate decision-making algorithms are likely to perform optimally. Based on simulation results, the JITAI design team can decide which decision-making algorithm to deploy on the JITAI for a sample of the subpopulation. This deployment produces additional data from the JITAI's target subpopulation, and these data are again used to update and fit the computational models to simulate upcoming deployments and evaluate design decisions. This bidirectional feedback cycle continues for the lifespan of the JITAI-Twin.

Figure 1. Digital twin for just-in-time adaptive intervention framework. Just-in-time adaptive interventions are designed for a subpopulation of individuals—not just a single individual, as highlighted in green to emphasize a key difference from most digital twins in health. This subpopulation serves as the physical twin. When a just-in-time adaptive intervention is deployed, data are collected from (a sample of) the physical twin. These data are used to construct simulation models that serve as the virtual twin. These models are used to simulate the performance of candidate decision-making algorithms for an upcoming deployment. The best-performing algorithm is then deployed within the just-in-time adaptive intervention. Data collected during deployment are then used to update the simulation models—and the modeling, simulation, decision-making, and data collection cycle continues—with bidirectional feedback taking place between just-in-time adaptive intervention deployments, as highlighted in green. Note the existence of data-deployment mismatches highlighted in pink, where data-deployment mismatches could involve differences in the prior data's sample from the upcoming deployment's sample or an inclusion of new variables. JITAI-Twin: digital twin for just-in-time adaptive intervention.

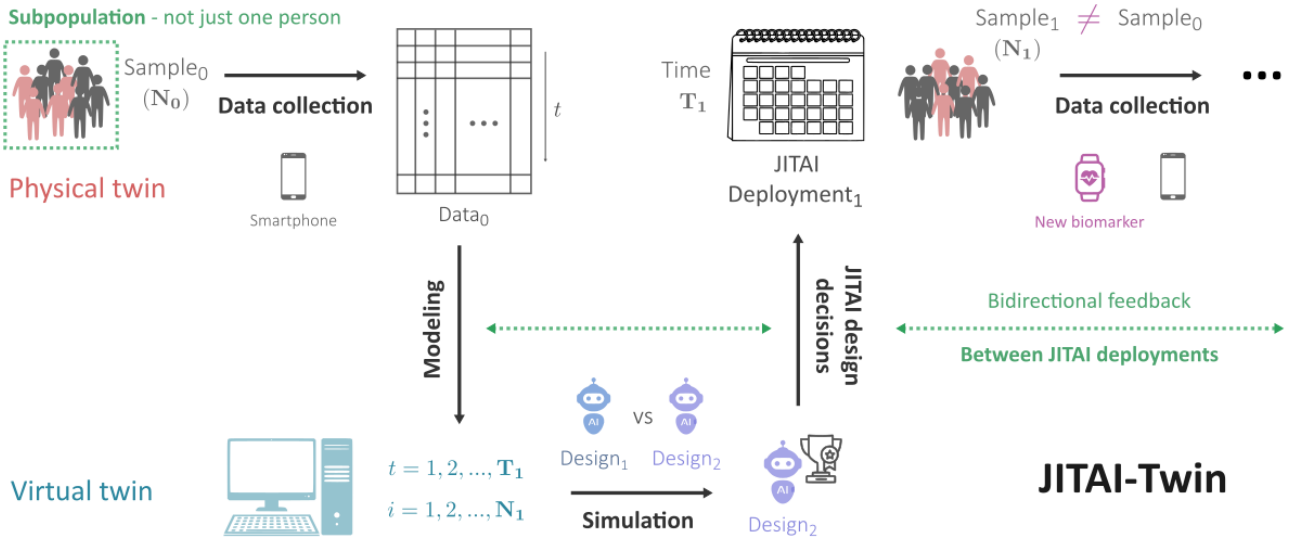

JITAI-Twins are "fit for [the] purpose" of modeling a deployment subpopulation's behavior over the course of a JITAI's continual improvement [13-15], and this presents unique challenges that differ from challenges discussed in previous literature on digital twins for health [16-23] (refer to Section S1 of Multimedia Appendix 1 for a comprehensive overview of related work). A specific deployment's sample rarely shares participants with a previous or follow-up deployment's samples (Figure 1). Thus, models of the subpopulation are not fit or updated to replicate the behavior of specific individuals. Instead, the models should approximate the behavior and heterogeneity of the entire subpopulation, while only sampling from that subpopulation. In contrast, most digital twin frameworks in health focus on modeling a single individual to make model-informed decisions (eg, model predictive control for an artificial pancreas' insulin delivery), where the person-specific model is updated with data specifically from the target individual [20]. Exceptions to this include "digital twins of a subpopulation" for rare disease treatment, discussed in the context of in silico clinical trials to better understand the potency of a drug [24]. Although only briefly mentioned in workshop proceedings of the National Academies without specific guidance on design or implementation examples [24], digital twins of a subpopulation for rare disease treatment are the most relevant example of digital twins. In fact, JITAI-Twins are another instantiation of the digital twin of a subpopulation concept. Other challenges that commonly arise for JITAI-Twins include "data-deployment mismatches." Data-deployment

XSL•FO
RenderX

mismatches are mismatches between the data available to construct the virtual twin and the upcoming real-world deployment characteristics that need to be simulated. Bidirectional feedback for JITAI-Twins takes place at the timescale of deployments, rather than any fixed unit of time (Figure 1). Due to the varying time lengths of deployments and the months or years that could pass between subsequent JITAI deployments, mismatches between previously collected data and future deployments can arise as society, science, and digital health technology evolve.

This Paper's Contributions

Compared to existing literature, the JITAI-Twin framework is the first instantiation of the digital twin of a subpopulation concept with associated guidance on design and implementation examples. Digital twins of a subpopulation are important to making informed design decisions and maintaining intervention fidelity when deploying a JITAI for a group of individuals. Relative to the literature on simulation testbed design for reinforcement learning (RL) algorithms for JITAIs [7], the JITAI-Twin framework highlights the need for bidirectional feedback. Feedback from virtual to physical involves design decisions that are informed via JITAI-Twin simulations. Feedback from physical to virtual helps improve a JITAI-Twin and ensure it matches the ever-changing deployment subpopulation for a JITAI. JITAIs are not designed and deployed as one-time efforts. JITAIs require continual improvement (eg, redesign and deployment) as scientific progress is made, technology evolves, and human behavior changes. Similarly, modeling, simulation, decision-making, deployment, and data collection should be viewed as an ongoing cycle (Figure 1).

This paper is organized as follows. We first provide a motivating example for JITAI-Twins based on a JITAI in the physical activity space that has required continual improvement over multiple deployments [5,25,26]. We then provide guidance on JITAI-Twin design and elucidate the step-by-step process of designing and updating JITAI-Twins. We also enumerate potential solutions to data-deployment mismatches that often complicate a JITAI-Twin's design and updates. Penultimately, we discuss the iterative evaluation process of JITAI-Twins under the digital twin framework of verification, validation, and uncertainty quantification (VVUQ). Finally, we conclude with guidance on selecting a JITAI's decision-making algorithm based on JITAI-Twin simulation results and discussion. We thereby make the following contributions:

1. Formalize the JITAI-Twin framework to instantiate digital twins of a subpopulation for the optimization and continual improvement of JITAIs.
2. Provide guiding principles for the design and updates of a JITAI-Twin.
3. Elucidate the design and iterative evaluation of JITAI-Twins, with examples of imputation and sensitivity analyses to address data-deployment mismatches.

Motivating Example of a JITAI's Continual Improvement: HeartSteps

HeartSteps is a JITAI that has been designed, optimized, and continually improved over the past decade to promote physical activity for sedentary adults [5,27]. In this section, we outline the development journey of HeartSteps, highlighting specific aspects of HeartSteps' development that benefited from the use of a JITAI-Twin. We also highlight the challenges faced in designing the JITAI-Twin due to real-world considerations with HeartSteps.

The initial version of HeartSteps ("HeartSteps V1") was aimed at increasing physical activity levels for sedentary adults by encouraging bouts of physical activity during daily life [5]. HeartSteps V1 consisted of a Jawbone Up wrist-worn activity tracker and a smartphone app that provided self-monitoring and "push" intervention components. The push interventions included 2 types of contextually tailored activity suggestions—intended to help individuals disrupt their sedentary behavior and take short (~10 minutes) walks throughout the day, respectively—and planning prompts, intended to help participants plan a longer activity bout for the following day.

To quantify the effects of the activity suggestions on subsequent physical activity and optimize future development of a JITAI to intelligently deliver activity suggestions targeting walking bouts (ie, walking suggestions), a 6-week microrandomized trial (MRT) was conducted. MRTs are designed to quantify how potential intervention components affect near-term outcomes of interest (eg, step counts), how these effects are moderated by context (eg, weather outside), and how these effects change over time [14]. The data from this initial MRT were then used to optimize the provision of both types of activity suggestions in the revised version of HeartSteps ("HeartSteps V2").

To optimize the provision of walking suggestions for HeartSteps V2, the team developed an RL algorithm that personalized the provision of walking suggestions. This personalization was aimed at maximizing the impact of walking suggestions on the intended proximal outcome: step count in the subsequent 30 minutes [26]. To develop an appropriate RL algorithm and select its associated hyperparameters, the team designed simulation environments that we frame here as the first iteration of a JITAI-Twin for HeartSteps. The JITAI-Twin helped evaluate candidate RL algorithms and strategically select the appropriate RL algorithm to deploy [26].

A challenge in designing the JITAI-Twin for HeartSteps was the existence of data-deployment mismatches. First, HeartSteps V2 focused on a different population—individuals with type 1 hypertension living in a different geographical location (Seattle for V2 vs Michigan for V1)—and involved changes to the data collection methodology compared to HeartSteps V2. Second, the Jawbone sensor was no longer available, so HeartSteps V2 used Fitbit Versa smartwatches to track and collect physical activity data. Third, HeartSteps V2 included daily and weekly collection of ecological momentary assessment data that were different from the data collected when HeartSteps V1 was

XSL•FO
RenderX

deployed, in part due to the team's growing understanding of the psychosocial variables that mediate and moderate physical activity patterns. For example, an app engagement variable was included in HeartSteps V2 to help the RL algorithm personalize the provision of walking suggestions, but HeartSteps V1 did not include measurements of app engagement. After addressing these data-deployment mismatches (discussed in the section "Designing and Updating JITAI-Twins"), the JITAI-Twin simulation results informed the RL algorithm that was then deployed and evaluated for HeartSteps V2 [26].

To address limitations of the prior RL algorithm and incorporate the latest findings from the science of behavior change literature, the team is now working on an improved RL algorithm for deployment (this latest iteration of the JITAI is termed "HeartSteps V3" in this paper) [25,28]. The RL algorithm for HeartSteps V2 is optimized for increased step counts [26], but recent evidence demonstrates that long-term commitment to behavior change, such as the maintenance of physical activity, is more closely determined by affective attitudes toward the desired behavior, rather than short-term bouts of the desired behavior [29-31]. To incorporate this insight into HeartSteps V3, the upcoming deployment will use affective attitudes toward physical activity as outcome variables, and the walking suggestion component was redesigned to pair suggestions for walking with affectively positive stimuli to both increase receptivity to the suggestions and impact affective attitudes. The JITAI-Twin for HeartSteps is being updated accordingly based on the data collected during the deployment of HeartSteps V2. The JITAI-Twin is again being used to evaluate and select between candidate RL algorithms for HeartSteps V3 [25], and is again faced with data-deployment mismatches, as we describe in the section "Designing and Updating JITAI-Twins."

Guiding Principles to Follow for JITAI-Twin Design

When designing and updating a JITAI-Twin, several modeling and simulation decisions must be made. To inform these decisions, we propose the following guiding principles (Table 1) based on lessons learned from prior JITAI-Twin design and related work [7,26,32,33].

Table . Guiding principles for JITAI-Twin[a] design.

Guiding principles	Brief description
Documentation	Document the purpose and plans for the design of a JITAI-Twin, adjustments made during design and updates, and simulation outcomes prior to each deployment
Team science	Include domain scientist(s), data scientist(s), and a deployment lead to ensure the necessary interdisciplinary expertise exists on the JITAI-Twin design team
Model generalizability	Use data modeling strategies that maximize generalizability from prior data to upcoming deployments
The devil is in the details	Judiciously investigate each aspect of a JITAI-Twin's design, especially aspects that are critical to simulation performance
Refrain from data removal	To the extent possible, refrain from removing data that is representative of real-world deployments (eg, disengaged participants) when designing and updating the virtual twin

[a]JITAI-Twin: digital twin for just-in-time adaptive intervention.

Documentation

A JITAI-Twin should be fit for purpose [13], and this purpose should be clearly documented. The design and update plans should be detailed, assumptions made in modeling the physical twin should be documented, and the objectives for each of the simulations conducted should be prespecified. During the JITAI-Twin design, adjustments to the plan will be necessary—these design decisions should be documented for computational reproducibility and continuity [34]. In general, a well-designed and documented JITAI-Twin has the potential to outlive its initial designers and provide a quality return on investment. Since JITAI-Twin simulations are conducted offline between JITAI deployments, computational cost is less likely to be a limiting factor compared to the time and effort in designing and maintaining a JITAI-Twin (ie, labor costs). Thus, like any computational resource, poor documentation can lead to wasted time and effort during updates to a JITAI-Twin, which can incur significantly more costs compared to spending time initially and at each update step on documentation.

Team Science

Just as an interdisciplinary team is recommended when designing and deploying a JITAI, an interdisciplinary team is recommended when designing and leveraging a JITAI-Twin for decision-making [35]. The JITAI-Twin design team should, at a minimum, consist of the deployment lead, a domain scientist, and a computational scientist. The deployment lead is the researcher in charge of executing the JITAI deployment. The deployment lead's insight is necessary to faithfully model practical aspects of a deployment and the target subpopulation (eg, expected disengagement from the JITAI). The domain scientist is the expert in physiology, psychology, etc, relevant to the health application (eg, suicidologist for a suicide prevention study). This domain scientist's insight is necessary to design simulation models that reasonably approximate psychophysical changes and outcomes. Finally, the data scientist

XSL•FO
RenderX

(eg, computer scientist, engineer, or statistician) is needed to mathematically abstract psychophysical characteristics into virtual constructs (eg, simulation models). Ideally, this data scientist has interdisciplinary training to improve cross-disciplinary communication and abstraction.

Model Generalizability

"All models are wrong," but "useful" models have explanatory power, capturing relationships that generalize from past data to future deployments [36]. We highlight 2 interpretations of explanatory power relevant to JITAI-Twin design [37,38]. First, explanatory power reflects a model's ability to quantitatively capture known qualitative relationships, a characteristic often attributed to mechanistic models. This is especially important in the context of counterfactual outcomes, where a simulation model will need to generate plausible outcomes for alternative intervention options that were not selected in prior data. A recommended approach is to assess the limitations of prior data and ensure aspects of the model informed by domain knowledge account for blind spots in the data. For example, if prior data do not account for a portion of the JITAI's target subpopulation, construct variants of the computational models that account for potential differences in behavior informed by domain knowledge [39]. Second, explanatory power refers to a model's ability to generate data that align with real-world observations, a strength of data-driven machine learning models. JITAI-Twin models should balance the strengths of mechanistic and data-driven modeling to enhance generalizability [40]. Generalizability refers to how well computational models can explain or reconstruct data that were not used to train the models and has been extensively discussed in the literature [41,42]. A wealth of literature exists on the challenges of model selection for generalizability and approaches to consider [43,44]. In general, approaches to enhance generalizability ensure that a model's complexity is commensurate with data availability. This is related to the bias-variance trade-off that has been shown empirically to be of value to model generalizability [45]. The bias-variance trade-off generally indicates that generalization error will follow a U-like curve, where large errors with too few parameters indicate too much bias and too little variance (ie, too little flexibility to capture the underlying patterns), and large errors with too many parameters indicate too little bias and too much variance (ie, too much flexibility causing the model to overfit). Recent work on the surprising performance of overparameterized machine learning models adds a new layer to this guidance, showing that the "double descent phenomenon" can be observed with overparameterized models (eg, deep neural networks) [46]. The complexity of model selection makes team science even more critical, as data scientists may have helpful intuition from experience that contradicts general guidance for a specific application setting.

Note that a JITAI-Twin's simulation models should not only capture the underlying relationships of interest in the data (eg, relationships between intervention options and outcomes of interest), but also the unstructured variability within the data (ie, the "noise" characteristics). When candidate decision-making algorithms are evaluated, simulations will need to be conducted with stochastically generated noise to evaluate how well decision-making algorithms perform in settings that are not accounted for by the learned aspects of the models. Generalizable models are critical because a JITAI-Twin's simulations inform real-world decisions for settings that will likely differ from prior data.

The Devil Is in the Details

"Inattention to detail is the hallmark of mediocrity" for JITAI-Twin design [47]. JITAIs and real-world deployments are sophisticated, and because JITAI-Twins are used to make design decisions for real-world JITAI deployments, details cannot be overlooked or the risk of failure can become significant (eg, the Mars Orbiter that crashed partly due to a mismatch between metric and imperial units [48]). JITAI-Twin design must match the level of sophistication by attending to details such as the assumptions and limitations associated with a modeling approach (eg, assuming linearity vs modeling nonlinearities that may require more data). Of course, "perfect is the enemy of good," so a JITAI-Twin design team must judiciously worry about aspects of a JITAI-Twin's design that are critical to modeling the physical twin accurately (ie, "worrying selectively") [36]. Team science helps guide the team in worrying selectively. Protocol details that are critical to model accurately can be highlighted by deployment leads. Domain scientists can prioritize which psychophysical relationships are most important to model. Data scientists can clearly articulate the limitations and the assumptions of modeling approaches they use. Due to the level of detail required in designing a JITAI-Twin, its design should be viewed as a research project itself. Adequate time must be budgeted for collaborative prioritization, modeling, simulation, evaluation, and iteration. Iteration will help data scientists learn what aspects of the target subpopulation and deployment are most important to model accurately for informative simulations. Iteration will also help deployment leads and domain scientists learn what correspondence is reasonable between the physical twin and virtual twin and maintain that standard.

Refrain From Data Removal

Data removal should be avoided to the extent possible when designing a JITAI-Twin. JITAI-Twins approximate the true behavior of a target subpopulation using a JITAI. A fundamental truth of JITAI deployments is the harsh reality of user disengagement, missing data, etc. When disengaged users or users with insufficient data are removed or ignored, the JITAI-Twin becomes biased toward the subsample of prior participants or users that remained engaged [49,50]. However, future JITAI deployments will not only include engaged users, but will also include users who disengage. For a JITAI that seeks to personalize intervention delivery, disengaged users may in fact be the most important to consider [32]. Removing incomplete data would overlook users who may benefit the most from a well-designed decision-making algorithm. That said, if data are incorrect or are unrepresentative of the physical twin due to measurement or recording discrepancies, these data may need to be removed. Hence, the suggestion is not to avoid data removal entirely. Rather, the prudent JITAI-Twin designer should judiciously refrain from removing data representative of upcoming deployments.

XSL•FO
RenderX

Designing and Updating JITAI-Twins

Designing and updating a JITAI-Twin involves 5 overarching steps, as visually summarized in Figure 2:

Figure 2. Overarching steps to designing (and updating) digital twins for just-in-time adaptive intervention. The 5 steps are illustrated from left to right and detailed in the main text. Verification and validation are iterative and continue until the base simulation testbed meets the desired performance. The construction of testbed variants is a form of uncertainty quantification that, along with other uncertainty quantification strategies, ensures just-in-time adaptive intervention design decisions are made in an uncertainty-informed manner. JITAI: just-in-time adaptive intervention.

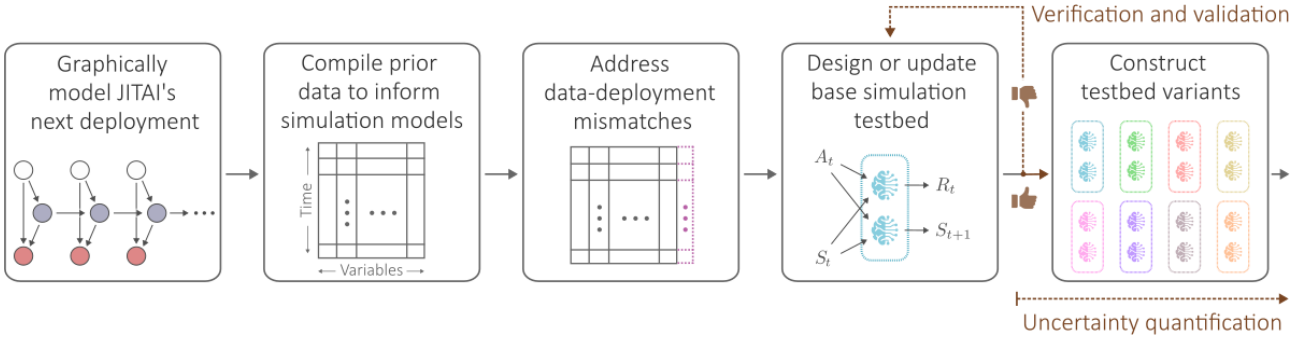

1. The first step is to graphically model the upcoming deployment's characteristics (eg, how often intervention can be delivered), intervention targets (eg, intervention will try to enhance positive affect toward physical activity), and target subpopulation (eg, sedentary adults). A recommended approach is to construct a causal directed acyclic graph (DAG) for the upcoming deployment to qualitatively represent expected user behavior in response to interventions, moderated by context variables [51,52].
2. The second step is to compile prior data that mirror the data to be collected for the upcoming deployment to the greatest extent possible. These data might be compiled from a previously conducted MRT, as was the case for HeartSteps V1 data [5,26]. Note that data from an MRT are not required, as will be discussed in the context of data-deployment mismatches.
3. The third step is to address data-deployment mismatches. For example, if the upcoming deployment involves an intervention component that was not present in prior deployments, the effect size for the intervention component must be imputed. The imputed effect size captures what domain scientists anticipate regarding the strength of the intervention component's effects on outcomes, relative to non–intervention-related changes in outcomes (ie, the "signal-to-noise" ratio). The imputed effect size enables simulations of hypothetical responses to delivery of the intervention [53].
4. The fourth step is to construct a base "simulation testbed" informed by the graphical model, using the data compiled and imputation decisions made. A simulation testbed is defined as a collection of computational models used to simulate the performance of a JITAI's decision-making algorithms. The base simulation testbed will include a set of simulation models (eg, outcome and state transition models) and parameters to vary. Some parameters will correspond to values imputed or specified to address data-deployment mismatches. Since JITAI-Twins are digital twins of a subpopulation and not of a single individual, the simulation models must capture heterogeneity between individuals in the subpopulation. The base testbed will need to be iteratively evaluated and redesigned until satisfactory evaluation results, as illustrated in Figure 2 and discussed in the next section, "Iterative Evaluation: VVUQ."
5. The fifth step is to create variants of the base simulation testbed. These variants are used to evaluate how robust candidate algorithms are to potentially incorrect modeling assumptions. For example, if an intervention's effect size had to be imputed (in step 3) or the intervention's effects may differ from prior data, one can form testbed variants by varying the intervention's effect size to understand how candidate algorithms perform across potential scenarios. This is a form of uncertainty quantification (Figure 2), as discussed in the next section, "Iterative Evaluation: VVUQ." The variants and the base simulation testbed together form the virtual twin of the JITAI-Twin.

To make explanations more concrete, the remainder of this paper will focus on JITAI-Twins that simulate the performance of candidate RL algorithms (eg, the HeartSteps JITAI-Twin). The steps we describe remain applicable to simulating the performance of other decision-making algorithms for JITAIs, such as control engineering methods, decision trees, expert systems, etc [54]. RL algorithms are AI algorithms that iteratively optimize the selection of actions based on previously observed rewards and states [55]. In the context of JITAIs, RL algorithms are decision-making algorithms that can select between intervention options (actions) to optimize proximal outcomes (rewards) based on an individual's tailoring variables (states). Example actions could include whether a physical activity suggestion is delivered or which suggestion is delivered. An example reward variable could include an individual's step count as a surrogate for moderate-to-vigorous physical activity. Example state variables could include an individual's environmental context, such as the weather or information about the individual, such as their current location. State-of-the-art JITAIs use RL algorithms to personalize intervention delivery, progressively determining which intervention option is likely to optimize outcomes based on the inferred state of an individual [26]. For a primer on RL and simulation modeling, refer to Section S2 of Multimedia Appendix 1.

XSL•FO
RenderX

Graphically Model the Subpopulation and JITAI Deployment Characteristics

A graphical model helps ground subsequent JITAI-Twin design and update steps. With a graphical model, the research team can identify variables of interest and compile data accordingly for training models. The team can also identify which variables or relationships are important for which no data constitutes data impoverishment. When designing the base simulation testbed, the computational models used can then be structured according to the relationships specified in the graphical model. The graphical model can also help inform which variants to construct to investigate uncertainties in the graphical model.

For HeartSteps V3, the deployment characteristics and subpopulation behavior are currently being modeled by the causal DAG shown in Figure 3, adapted from Gao et al [25]. Note that this DAG is still under refinement but has undergone multiple iterations to collaboratively develop. A DAG visually describes causal relationships between the specified variables using directed arrows that are acyclic [51]. Importantly, arrows between variables that are omitted from the DAG reflect effects that are believed to be weak or nonexistent. Constructing a graphical model of this form requires the team to first identify the variables of importance to the upcoming trial. At a minimum, this consists of context variables (C in Figure 3), intervention options (A in Figure 3), and outcome variables (R and E in Figure 3). Identifying these variables translates to specifying states, actions, and rewards for the RL algorithm simulation testbed. The conceptual model should also specify mediators (M and N in Figure 3) that may affect the relationships between context, interventions, and outcomes. Since JITAI deployments involve decisions that are made over multiple time points (ie, sequential decision-making), it is important that the conceptual model captures time dependencies. For example, Figure 3 also illustrates dependencies across consecutive weeks and across consecutive days within a week. Assumed independence in time is also captured with the absence of arrows. Although independence between variables may not perfectly mirror the physical twin, it is important to selectively worry about the relationships most consequential to the upcoming deployment. A recommended approach to do this is to rank order the relationships between variables and decide on a cutoff point below which all relationships will be assumed inconsequential. Variants of the simulation testbed can be used to test the robustness of candidate decision-making algorithms to relationships that fall slightly below the cutoff point [25], as discussed in the later section “Construct Variants of the Base Simulation Testbed.”

Note that the variables included in the graphical model map to the subpopulation and deployment characteristics and not to the JITAI directly. This distinction is critical between JITAI-Twin models and a JITAI’s models used for decision-making. The JITAI may not measure or have access to all the variables that are important for simulations. The variables important for simulations are the variables needed to satisfactorily approximate how the target subpopulation will respond to the JITAI for the upcoming deployment. The graphical model needs to encompass all variables that are crucial for the JITAI-Twin’s simulation models.

XSL•FO
RenderX

Figure 3. Example causal directed acyclic graph used for the HeartSteps digital twin for just-in-time adaptive intervention. The directed acyclic graph shown is being iteratively refined to graphically model HeartSteps V3. Each arrow indicates a causal relationship from the variable at the start of the arrow to the variable at the end of the arrow. The subscript w represents week (ie, w-1 refers to previous week), so the subscript w,1 refers to the first decision time within the week, while subscript w,K refers to the Kth decision time within the week. The context, C, represents an individual's recent step count, GPS location, etc The proximal outcome, M, is the step count after an intervention is or is not delivered. Intervention, A, represents the decision to send a physical activity suggestion or not. Proximal outcome, N, represents whether a user clicks on an intervention message or opens the app. Intermediate outcome, E, represents a higher-level "app engagement" variable. The outcome variable, R, represents affective attitude toward physical activity.

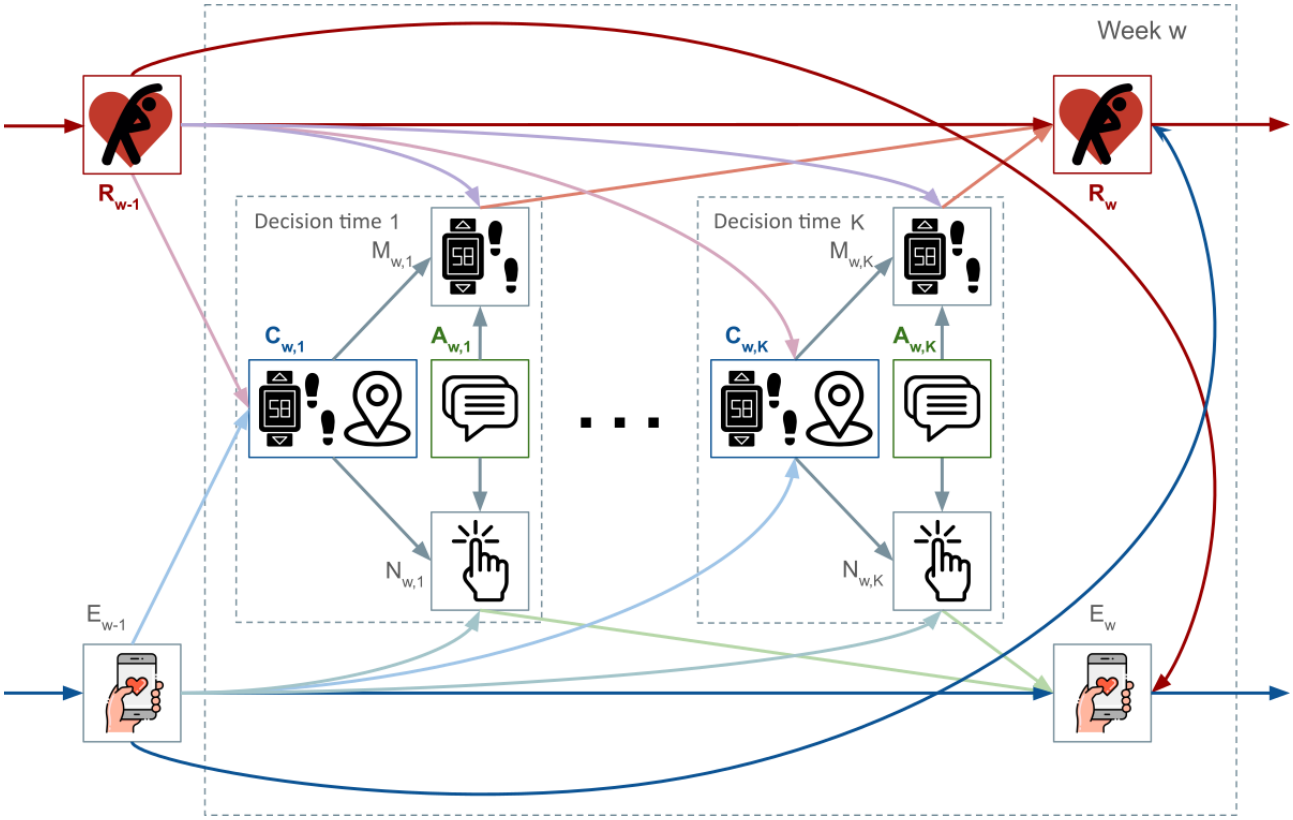

Compile Prior Data to Inform Simulation Models

Prior data are necessary to train and otherwise inform the computational models used for a JITAI-Twin. If prior deployment data are not available (ie, the JITAI-Twin is being designed for the first time, rather than being updated), the study team should identify prior data most relevant to the upcoming deployment's simulations (similar variables, timescales of measurement, etc). The graphical model discussed in the previous section ("Graphically Model the Subpopulation and JITAI Deployment Characteristics") can help guide data compilation. Relevant prior data should help inform relationships between variables in the graphical model.

For HeartSteps, the initial JITAI-Twin was designed using data from the HeartSteps MRT (ie, HeartSteps V1). MRT data are ideal for JITAI-Twin creation because MRT data contain dynamic measurements of outcomes and context relevant to the intervention components being investigated—and the provision of interventions is microrandomized. The microrandomization of intervention provision provides data for use in quantifying the relationships between intervention options and outcomes, and how the relationships are moderated by context variables. The outcomes and context variables map to rewards and state variables that a simulation testbed can generate in response to an RL algorithm's simulated actions (ie, intervention provision).

When compiling prior data, it is important to avoid data leakage [56]. The study team should structure prior data in a way that mirrors how data will be used to inform algorithms for deployment. For example, some RL algorithms, such as Thompson Sampling algorithms used by HeartSteps, enable the use of informative Bayesian priors or other hyperparameters [57]. If the plan is to use a single dataset for both simulation and to inform the Bayesian priors, then the data should be split into disjoint datasets for modeling and Bayesian prior construction or hyperparameter selection, respectively [58].

Address Data-Deployment Mismatches

The data compiled (see previous section, "Compile Prior Data to Inform Simulation Models") are likely missing variables of interest from the graphical model (see previous section, "Graphically Model the Subpopulation and JITAI Deployment Characteristics") or are deficient in some way when it comes to informing relationships of interest for the simulation models. We term these deficiencies "data-deployment mismatches." Data-deployment mismatches are a reality of JITAI-Twin design and updates. During the continual improvement process, research teams often want to incorporate the latest innovations in sensing and intervention technology [59]. The science of behavior change, psychophysiology, and other relevant domain sciences is also progressing rapidly [28,60-62]. This progress and rapidly evolving technology often add additional variables

XSL•FO
RenderX

or dependencies to consider for the upcoming deployment that are absent from previous deployment data. This absence of variables in prior data necessary to simulate upcoming deployment decisions is termed "data impoverishment." Other forms of data-deployment mismatches include domain shift and sampling bias. Deployments may take place years apart, where characteristics of the subpopulation or their behavior may change. The time of the year or the specific sample investigated for each deployment may differ from previous deployments.

To characterize the data impoverishment faced for JITAI-Twin's design or updates, identify the missing variables and/or relationships in the graphical model that cannot be fit using the data compiled. This data impoverishment can be overcome through judicious imputation decisions and the variation of unknown model parameters. These variations are a form of sensitivity analysis, where the uncertainty in the model parameters is propagated through to the simulation results (eg, if parameter x falls below value a, then decision-making algorithm 1 performs better than algorithm 2). Team science is critical to plausibly overcoming data impoverishment. Domain scientists and trial leads help ensure imputation and sensitivity analyses are conducted realistically.

Identifying domain shift and sampling bias requires careful comparison of the prior data, the overall subpopulation, and the upcoming deployment's characteristics. For many deployments, it is not possible to fully anticipate the upcoming sample's characteristics. However, it is important to recognize any oversampling or undersampling of a particular participant group in prior data. Sampling bias may also occur with respect to time. Prior data collected in the winter for a physical activity study will differ from what is expected if deploying a physical activity JITAI in the summer. When sampling bias is less obvious to the eye, methods to detect sampling bias include statistical tests to compare sample demographics and other characteristics to mean values or proportions known to be representative of the subpopulation. Domain shifts can still take place even if sampling was done in a representative manner. Over time, the characteristics of the subpopulation can change (eg, expectations of technology can change rapidly, as with the recent proliferation of generative AI tools), and how technology or a JITAI is deployed can also change. The detection of significant domain shifts (ie, "distribution shifts" or "dataset shifts") is an active area of research, with methods that involve combining dimensionality reduction (of features such as baseline survey responses) and statistical testing performing the best in practice [63]. To overcome these data-deployment mismatches, domain knowledge can be used to adjust parameters of the simulation models according to expected differences between the domain or sample present in prior data and the upcoming deployment. Parameters of the simulation model that capture variability in the subpopulation or data collection environment can also be enlarged. Performing simulations with enlarged variances can help account for sources of variability in the upcoming deployment that may not have been observed in prior data due to sampling bias or domain shift.

For HeartSteps V3, outcome variables are impoverished [25]. Progress in the science of behavior change indicates that optimizing for short-term bouts of a desired activity is less indicative of sustained behavior change compared to optimizing affective attitudes toward the desired activity [28], which applies to the maintenance of physical activity [64-67]. Affective attitudes toward physical activity were not collected in prior data. To address this data impoverishment, affective attitudes toward physical activity are now being generated when updating the JITAI-Twin to simulate the upcoming deployment. Specifically, a linear state-space model is fit using system identification to infer latent (ie, not measured) changes in affective attitudes toward physical activity from observed (ie, measured) bouts of unprompted and prompted physical activity [25]. The relationship between affective attitudes toward physical activity and bouts of unprompted physical activity was captured in the DAG (Figure 3). For more examples of how data impoverishment has been addressed in the past for state and action variables, see Section S3 of Multimedia Appendix 1.

Design the Base Simulation Testbed

Designing the base simulation testbed is an iterative process. Computational models are designed to generate (1) outcomes and (2) state transitions each time a candidate decision-making algorithm takes an action. These models must account for heterogeneity between individuals in the subpopulation. The computational models are then iteratively evaluated to ensure plausible outputs are generated. A detailed discussion of iterative JITAI-Twin evaluation is provided in the next section, "Iterative Evaluation: VVUQ." Note that the model fitting and evaluation procedure for simulation models for JITAI-Twins differs from the procedures used for prediction and statistical inference models typically discussed in the machine learning and statistics literature. A simulation model for a JITAI-Twin is used to evaluate candidate decision-making algorithms, so the simulation model should not only be able to capture the underlying relationships between decisions, context variables, and outcomes of interest—but the simulation model should also be able to model noise characteristics properly. Simulations will need to be conducted with stochastically generated noise to evaluate how well candidate decision-making algorithms perform in settings that are not accounted for by the learned aspects of the models.

Outcome Modeling

Outcome generation requires a model that maps the state of the simulation environment and the action an RL algorithm takes to some set of outcomes. These outcomes can then be used to form rewards for an RL algorithm. Note the subtle distinction between outcomes and rewards. Rewards are used by an RL algorithm for learning and can differ from one RL algorithm to the next. Outcomes are generated by the JITAI-Twin regardless of which RL algorithm is used. The RL algorithms may use those outcomes to form their own reward signals.

Outcomes can be modeled probabilistically, $y(s, a)=P(Y_t=y|S_{tw}=s, A_t=a)$. The superscript W emphasizes that this is the "world state" (ie, the all-encompassing state of the simulation environment, rather than the limited-in-knowledge state of specific RL algorithms to be evaluated). See the RL primer in the supplement for further explanation (Section S2 of Multimedia Appendix 1). The

XSL•FO
RenderX

outcome model will need to be trained on prior data, after addressing data-deployment mismatches (section "Address Data-Deployment Mismatches"). The model used for y (s, a) can be of several forms [68]. See the work on HeartSteps V2 and HeartSteps V3 for examples [25,26]. The guiding principles presented in this paper and the availability of data should inform the model class and the model fitting process used [69,70].

State Transition Modeling

A state transition model is of the form W(s′, s, a)=P(St+1w=s′|Stw=s, At=a). Again, the superscript W emphasizes that this is the "world state." See Section S2 of Multimedia Appendix 1 for further explanation. This model quantifies the probability of reaching state S′, given that the current state is S and the RL algorithm for the JITAI takes action a. The simulation environment's state transition model dictates how the simulation environment's state changes over time in response to actions taken by the simulated JITAI, regardless of which simulated JITAI is used. Our ongoing work for HeartSteps V3 learns a state transition model to continually improve the JITAI-Twin for HeartSteps [25]. Note that by concatenating the next state, St+1w and the outcome, Y_t, into a single vector, a single probabilistic model conditioned on S_t and A_t can be used to produce a vector of next state and outcome variables. This approach is used by Gao et al [25] and captured by the DAG in Figure 3, where next states are not differentiated from outcomes.

A simpler approach taken for HeartSteps V2 was to view state generation from a model-based data augmentation perspective [26]. Models are used only for the outcome function and to address state impoverishments. The prior data's state sequences are then used as is to generate state transitions. This preserves the dependence across time in state variables but makes the simplifying assumption that the actions of the candidate decision-making algorithms have minimal effects on the sequences of states observed in prior data (eg, states such as weather and location). This decision is based on the principle that digital therapeutics tend to have little immediate effect on autonomous state dynamics relative to residual noise (ie, implicit assumption of a simpler RL algorithm called a "contextual bandit").

Modeling Heterogeneity Between Individuals in a Subpopulation

A JITAI-Twin's outcome and state transition models must account for heterogeneity between individuals [71]. Several approaches exist for modeling heterogeneity between individuals in a subpopulation. One approach that has been used for the HeartSteps JITAI-Twin is the construction of several user-specific models to represent a virtual cohort of users [25,26]. By fitting a user-specific model to each user present in prior data, the variability across model parameters can be used to quantify the heterogeneity across individuals. The user-specific modeling approach is useful when variability lacks a clear distributional structure. However, the downside of the approach is that it assumes prior data are representative of the subpopulation. An alternative approach is to assume that the heterogeneity across individuals in the subpopulation exhibits some structure (eg, normal distribution) or can be stratified in some way (eg, categorizing users by some phenotype). Stratified modeling or mixed-effects modeling (eg, with random effects that are normally distributed) can be used to form a single model or a smaller subset of models to simulate the responses of a virtual cohort of users. An inherent trade-off that exists is that imposing structure may lead to underfitting, but assuming no structure risks overfitting to prior data. More flexible data generation approaches (eg, via artificial neural networks) that use proxy features to account for user identities or characteristics could be used to generate a virtual cohort [72]. The downside of more flexible data generation methods is that they typically require extensive data to produce realistic outputs.

Additional mHealth Considerations

A few additional aspects of mHealth interventions are important to model. Engagement with an intervention tends to decay over time [5,73]. This can translate to reduced effect sizes of the intervention (captured by the outcome model) over the course of a simulated trial, as well as participant dropout. As long as data on participants who dropped out were not removed from the prior trial's data, the prior trial's data can be used to inform the expected dropout rate over the course of a simulated trial. User burden in response to the intervention should also be considered. If an intervention is delivered too often, habituation may reduce the effect size of the intervention over time (captured by the state transition model's dynamics). Engagement may also diminish over time.

Construct Variants of the Base Simulation Testbed

Variants of the JITAI-Twin's base simulation testbed can be used to evaluate the robustness of candidate algorithms to potentially incorrect assumptions made during graphical modeling or decisions made to overcome data-deployment mismatches. The use of testbed variants is critical because the base simulation testbed on its own will fail to replicate characteristics of the upcoming deployment. However, the base testbed along with a well-designed set of variants can be used to cast a broader net over the possible scenarios the JITAI may encounter in the upcoming deployment.

The variants constructed should all satisfy "matching conditions." Matching conditions are defined here as marginal statistics or other outputs of the simulation testbeds that must be met to ensure scientifically plausible simulations. For example, if the study team decides that the degree of between-subject variability observed must be greater than some value, then plausible variants must produce between-subject variability above that value. Calibration can be used to ensure matching conditions are satisfied.

For HeartSteps V3, a simplified DAG is used to inform the RL algorithm (simplified version of the DAG shown in Figure 3; see Gao et al for details) [25]. Certain variants of the simulation testbed intentionally cause mismatches between the DAG used to inform the RL algorithm and the DAG used to generate the JITAI-Twin simulation testbed's rewards. The RL algorithm selected for deployment should be robust to these mismatches. Imputed effect size trajectories were also varied to construct additional variants. These variants are used to test the

XSL•FO
RenderX

performance of candidate RL algorithms for the entire range of plausible effect sizes (a matching condition) that could be observed during the upcoming deployment.

Iterative Evaluation: VVUQ

The evaluation process for JITAI-Twins is iterative and aligns with the digital twin evaluation framework of VVUQ [13,74], as shown in Figure 4. Verification ensures that the virtual constructs (eg, simulation models) created computationally (eg, using floating point approximations) mirror the desired mathematical descriptions. Validation ensures that the digital twin adequately represents the physical twin for the purposes of its design. Uncertainty quantification encompasses sensitivity analyses, the quantification of confidence in simulation results, and the quantification of uncertainty in data collected when designing and updating a JITAI-Twin. Sensitivity analyses determine how simulation outcomes are affected by varying learned parameters and/or modeled variables. Confidence measures are produced by performing numerous stochastic simulations and by using statistical methods to quantify confidence and credible intervals (ie, uncertainty quantification) for simulation outcomes. Data uncertainty can arise due to corrupted data, among other sources of uncertainty. Although difficult, characterizing sources of uncertainty (eg, heterogeneity in the subpopulation or stochasticity in simulation) and quantifying uncertainty accordingly are critical tasks.

Figure 4. Verification, validation, and uncertainty quantification for digital twins for just-in-time adaptive interventions. Verification, validation, and uncertainty quantification is an iterative process. For each iteration of the digital twin for just-in-time adaptive intervention, multiple repetitions of verification and validation are often necessary before satisfactory simulation performance is achieved. Uncertainties must then be quantified in simulations to inform decision-making for a just-in-time adaptive intervention's deployment. Data from a just-in-time adaptive intervention's deployment will contain sources of error that also need to be quantified when updating the digital twin for just-in-time adaptive intervention's computational models. The verification, validation, and uncertainty quantification cycle continues. JITAI: just-in-time adaptive intervention; JITAI-Twin: digital twin for just-in-time adaptive intervention; VVUQ: Verification, validation, and uncertainty quantification.

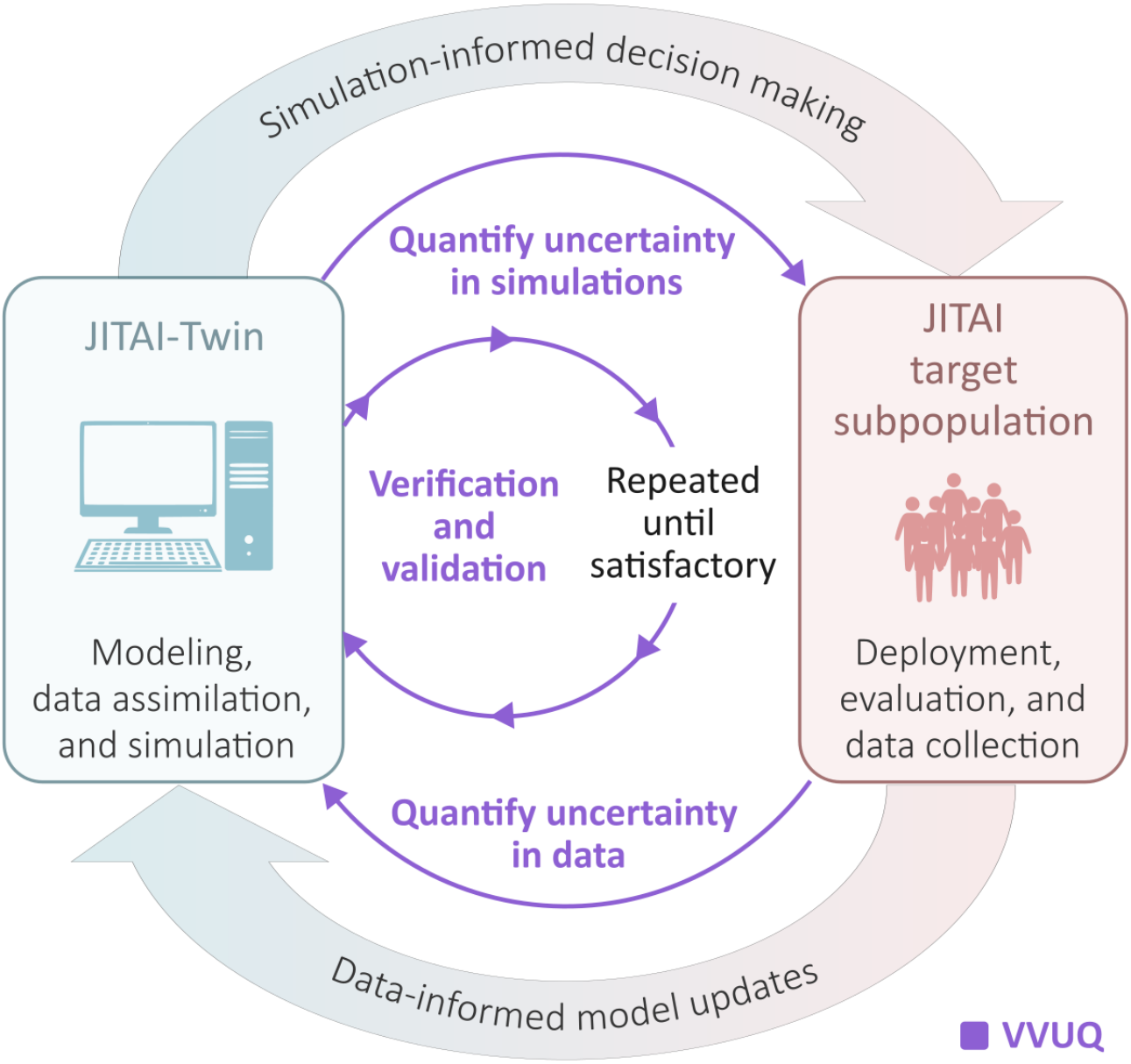

XSL•FO
RenderX

Verification

Verification is necessary for any computational model or piece of code. Standard practices to ensure proper code execution, negligible floating-point approximation errors, and appropriate data generation should all be first-pass evaluation criteria. One verification technique relevant to generative modeling is replication analysis. One can ensure replicable simulations by keeping track of the seeds used for random number generation and rerunning for verification. Other ways to improve verification include publicly sharing the code, procedural information, and data (if possible). Other researchers and practitioners can then participate in verification and improvements to the procedures and code.

Validation

Data from the upcoming deployment are necessary to truly validate a JITAI-Twin's simulations. However, once data are collected for the upcoming deployment, it is too late to update the JITAI-Twin according to validation results. Thus, alternative strategies for predeployment validation are necessary to ascertain JITAI-Twin quality. We separate JITAI-Twin validation strategies into 2 categories: mechanistic and empirical.

Mechanistic

Mechanistic validations are used to determine whether the generated simulation data are reasonable in the context of domain science and deployment experience. Specific examples include simulated engagement with a JITAI or participant dropout. Decaying engagement with an intervention and participant dropout rates should be modeled in a manner that aligns with deployment expectations. Other examples include simulated dynamics for state variables, outcomes, etc A concrete example from HeartSteps is to evaluate whether generated step count outcomes should decrease over time when the simulated virtual participant experiences an increase in positive affect toward physical activity (should generally not be the case in this example).

Empirical

Empirical validation requires detailed, data-driven evaluation of the JITAI-Twin's generative models and their outputs (see the supplementary material of Gao et al for an example from HeartSteps) [25]. Empirical evaluation strategies can be separated into 3 categories. The first category involves goodness-of-fit assessments on prior data used to fit the simulation models. Goodness-of-fit measures describe how well the trained model can generate outputs (in the absence of noise or residual components) similar to the prior data outputs when given prior data inputs (eg, coefficient of determination for regression). Coefficients of the model can also be assessed to ensure that the weights are of the appropriate scale, sign, and magnitude according to domain science.

The second category of empirical validation involves out-of-sample prediction (eg, using held-out data). For models fit to several participants' data simultaneously, a subset of participants' data can be held out to evaluate agreement between simulated and real data when the models are used to predict the held-out outcomes and state transitions given the held-out states and actions. When simulations are conducted without noise, generated data should align with the held-out data, on average (ie, the expected values, in a statistical sense, should align). When simulations are conducted with noise (as is done over multiple iterations when evaluating decision-making algorithms), the out-of-sample data should align with the distribution of data stochastically generated by the models. Importantly, the aim of out-of-sample prediction is not to replicate the held-out data during data generation. The aim of out-of-sample prediction is to instead evaluate how well marginal statistics and other diagnostic metrics agree with held-out data. In cases where a set of participant-specific models are trained to act as virtual participant models, participant-specific data can be held out and used for validation.

The third category of empirical validation diagnoses whether the models generate data that satisfy diagnostic checks. This is related to model calibration [75,76]. These diagnostic checks evaluate whether statistics computed across multiple simulated trials (eg, effect sizes and variability across virtual participants [7,25]) agree with the statistics computed from physical trials. It is recommended to inspect several statistics when validating a JITAI-Twin's generated data.

Uncertainty Quantification

Uncertainty quantification ensures that decisions informed by a JITAI-Twin are made with an understanding of the uncertainty in JITAI-Twin modeling and simulation results [77]. In this context, uncertainty quantification refers to computing measures of variability and alternative outcomes to accompany a JITAI-Twin's simulation results, such as 95% CIs for a candidate decision-making algorithm's simulated performance or simulation performances using a different testbed variant. There are several sources of uncertainty that, to the best of the JITAI-Twin design team's capacity, need to be considered, quantified, and propagated to quantify uncertainty for a JITAI-Twin's simulation results. We outline several important sources of uncertainty for JITAI-Twins in this subsection.

One source of uncertainty is stochasticity in simulations. Stochasticity is critical because simulations often involve several probabilistic outcomes that may differ from one simulation to the next (eg, the outcome of a coin flip). Stochasticity manifests either through probabilistic modeling used for the outcome and state transition models or probabilistic decision-making via a candidate decision-making algorithm. To account for this stochasticity, several seeds should be used for random number generation to investigate the variability in results across multiple simulation runs. CIs can be constructed accordingly, and algorithm evaluation decisions can then be made with a notion of statistical significance [32]. Note that for implementation purposes, the simulations corresponding to each random seed can be parallelized to potentially save time.

An additional source of uncertainty is the uncertainty in trained parameters and imputed variables. Parameters of a simulation model that are trained using prior data may not be accurate in representing upcoming deployment characteristics due to limitations in the quality and quantity of prior data. Variables that are imputed using domain knowledge to address data impoverishment may behave differently from what is observed on average in domain sciences. Sensitivity analyses (eg, via

XSL•FO
RenderX

variants of a simulation testbed) can be a useful way to investigate how simulation results vary as a function of changes in imputed variables or changes in learned parameters [78].

A final source of uncertainty we discuss here is uncertainty in the data compiled from prior deployments and studies, as shown in Figure 4. Accounting for uncertainty in collected data is important because the data compiled will significantly influence what the JITAI-Twin's simulation models capture during training. Uncertainty in compiled data can exist due to inference and measurement errors. Inference errors exist when a computational model is used to map measured data to variables of interest, such as context or outcome variables (eg, an algorithm that computes step counts will produce estimates with error). Measurement errors exist due to the inherent uncertainty in measurements, such as surveys of psychological constructs or sensor-based estimates of quantities that can be corrupted by external sources of noise. Distinguishing uncertainty in measurements from uncertainty in model predictions due to limited knowledge has been discussed extensively in the statistics and machine learning literature [79]. These sources of uncertainty can be simulated and propagated through to the simulation results by introducing noise terms with variances that can be scaled in magnitude to help quantify the sensitivity of simulation results to these errors [80].

VVUQ Is Continual

The VVUQ process is continual for a JITAI-Twin. Once a JITAI-Twin is used to inform decision-making in practice, a JITAI-Twin should be reevaluated following the completed deployment. This reevaluation will help inform the JITAI-Twin prior to its use in decision-making for the subsequent deployment—and the cycle continues, as depicted in Figure 4.

Evaluating Candidate Algorithms According to JITAI-Twin Simulation Results

Once a JITAI-Twin has been evaluated satisfactorily, the JITAI-Twin's simulation results can be used to decide which decision-making algorithm to deploy as part of the upcoming JITAI. Note that these simulations should be conducted with noise variables that match the characteristics of the noise observed in previously collected data (eg, informed by the simulation models' residuals when fit to prior data). This helps evaluate how well candidate decision-making algorithms will perform in real-world settings that are not completely accounted for by the learned aspects of the simulation models. For each candidate decision-making algorithm, the JITAI-Twin will produce a set of simulation results. These results will correspond to multiple simulated deployments for each variant of the base simulation testbed, as illustrated in Figure 5 (for a real example of simulation results, see Trella et al [7]). The goal at this stage is to identify the specific algorithm to be used for deployment based on the results shown across all simulations.

XSL•FO
RenderX

Figure 5. Illustrative example of a digital twin for just-in-time adaptive intervention's simulation results. For each simulation testbed variant of a digital twin for just-in-time adaptive intervention, candidate digital twin for just-in-time adaptive intervention decision-making algorithms will be evaluated according to metrics that the team deems useful for deciding which algorithm to deploy onboard the real digital twin for just-in-time adaptive intervention. Performance metrics may not always agree with respect to which algorithm to select. Trade-offs must often be made based on the performance priorities for the digital twin for just-in-time adaptive intervention to be deployed. JITAI: digital twin for just-in-time adaptive intervention; JITAI-Twin: digital twin for just-in-time adaptive intervention.

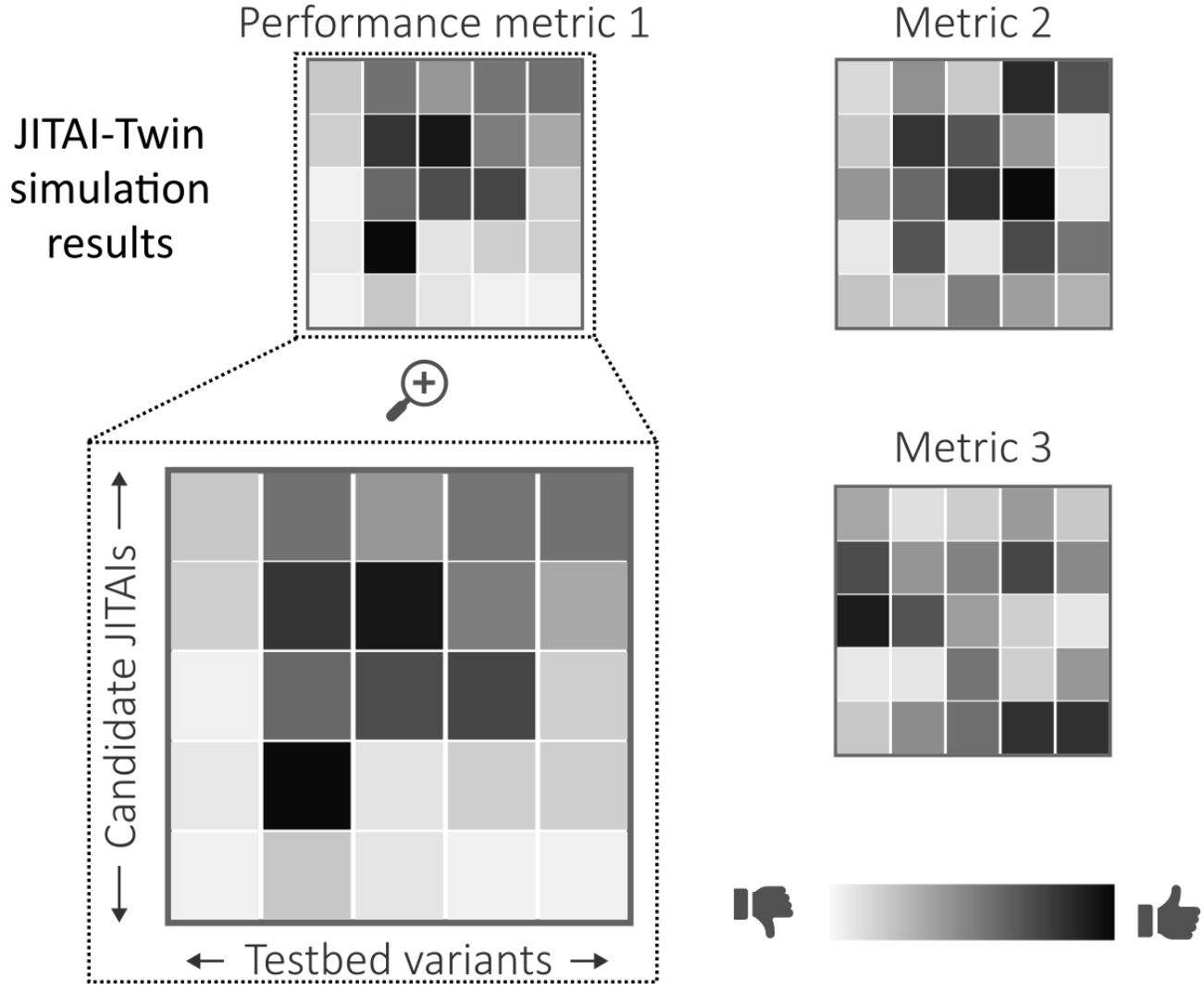

The metrics used to evaluate candidate algorithms and decide which to deploy will depend on the research team's priorities. If the primary goal of an intervention is to elicit a positive impact on average across the entire deployment sample, then a measure of central tendency across all virtual participants is likely sufficient. An example measure would be the mean across all virtual participants of the average cumulative reward across all simulated deployments (ie, $1N\sum n=1N1S\sum s=1S\sum t=1TRt$, s, n, where s indexes random seed, *n* indexes virtual participant, and *t* indexes time within each simulated deployment). However, if one goal of an intervention is to ensure that individuals with disorders that are more treatment resistant experience some benefit, then a lower percentile might be taken into consideration (eg, 25th percentile of $1S\sum s=1S\sum t=1TRt,s,n$). Other metrics could be used to investigate considerations related to time. For example, if an intervention trial may be terminated early if outcomes are not met [81], the algorithms may be evaluated for an initial subset of time (ie, instead of $1N\sum n=1N1S\sum s=1S\sum t=1TRt,s,n$, one could use $1N\sum n=1N1S\sum s=1S\sum t=1T`Rt,s,n$ for $T'<T$). If early participant disengagement is a primary concern, several T` values can be used to compare the speed of learning across the candidate RL algorithms. Table 2 summarizes the 3 example metrics described. Note that if multiple metrics are used to compare algorithms, design trade-offs will be necessary, as illustrated in Figure 5.

XSL•FO
RenderX

Table . Example metrics to evaluate candidate algorithms[a].

Example metric	Use case	Formula
Mean cumulative reward	Choose the algorithm that elicits the greatest positive impact on average across the entire deployment sample	$1N\sum n=1N1S\sum s=1S\sum t=1TRt,s,n$
Lower percentile of the cumulative reward (eg, 25th percentile)	Choose the algorithm that improves outcomes the most for those that are least benefited (eg, treatment resistant)	Lower percentile of $1S\sum s=1S\sum t=1TRt,s,n$
Mean, early terminated reward	Choose the algorithm that most improves the mean positive impact within a shorter time period than the full deployment	$1N\sum n=1N1S\sum s=1S\sum t=1T`Rt,s,n$ for $T'<T$

[a]$s\in\{1,2,...,S\}$ indexes random seed, $n\in\{1,2,...,N\}$ indexes virtual participant, and $t\in\{1,2,...,T\}$ indexes time within each simulated deployment.

For HeartSteps V2, the average total reward was used as the primary metric [26]. For HeartSteps V3, the average and 95% CIs for the cumulative reward are also reported [25]. CIs provide an indication of variability across virtual participants.

Discussion

Implications for Research and Practice

The JITAI-Twin framework we elucidate and the guiding principles we provide are informed by prior work in designing RL algorithms for JITAIs. Aside from the HeartSteps JITAI described in Guiding Principles to Follow for JITAI-Twin Design [25,26], 2 RL-empowered JITAIs that have been deployed in clinical trials were used to inform this paper. For the "Oralytics" JITAI designed to improve oral health in disadvantaged populations [82], the first iteration of a JITAI-Twin was designed and used to inform design decisions (hyperparameters of the RL algorithm's reward [53]) for the JITAI's clinical trial that started in fall 2023 and concluded in summer 2024 [83]. A unique challenge faced when designing the Oralytics JITAI—that was simulated as part of the JITAI-Twin—was delayed access to oral self-care behavior data from participants due to the use of a commercial smartbrush. Issues with data accessibility are common with commercial products and should be simulated by a JITAI-Twin. For the "MiWaves" JITAI designed to reduce cannabis use in emerging adults, the first iteration of a JITAI-Twin was used to select the reBandit algorithm over baseline approaches [32]. MiWaves was then deployed in a preregistered feasibility and acceptability study which was completed in fall 2024 [84]. A challenge faced with MiWaves was that the reward for the RL algorithm was a measure of engagement with the JITAI, but engagement with a JITAI is latent and difficult to quantify. A surrogate reward was thus defined and used for JITAI-Twin simulations to evaluate candidate algorithms. In general, the JITAI-Twin framework can be extended to many other domains, so long as the guiding principles of team science and attention to detail are followed to address domain-specific challenges.

The more costly it is to evaluate design decisions in the real world, the more valuable a JITAI-Twin is for optimizing and continually improving a JITAI. A/B testing is a tried-and-true approach to comparing design choices in practice [85], and is the basis for MRTs and other optimization experiments [14]. A significant downside to A/B testing and optimization experiments is that factorial designs can quickly become untenable when the cost of collecting data is nonnegligible and several nonobvious design decisions need to be made prior to the deployment of a JITAI. Especially for JITAIs that leverage RL and control engineering approaches, several design decisions must often be made just for model selection [8,43]. On the other hand, some design decisions may be difficult to model using prior data and simulate via a JITAI-Twin. An example would be whether a new intervention component that differs substantially from previous components should be added to a JITAI (eg, noninvasive vagus nerve stimulation in tandem with mindfulness interventions for stress management [86]). Optimization experiments are best reserved for these difficult-to-simulate design decisions, especially if the experimental cost of evaluating decisions is non-negligible. Thus, the recommended path forward is to use JITAI-Twins and optimization experiments in tandem to optimize and continually improve JITAIs.

Compared to person-specific digital twins, there are practical advantages to using digital twins of a subpopulation when optimizing and continually improving a JITAI. A person-specific digital twin becomes obsolete as soon as the specific person is no longer being considered for upcoming deployments of a JITAI. Digital twins of a subpopulation remain useful so long as the JITAI is being continually improved for the target subpopulation. Practically speaking, digital twins of a subpopulation also reduce the need to collect large amounts of data from specific individuals. Although person-specific digital twins enhance personalization and circumvent the need to model subpopulation heterogeneity, prior work shows that months of exploratory data collection can be required to satisfactorily construct person-specific models for a JITAI [87]. In behavioral settings where adherence remains an issue and disengagement often spells the end of an individual's participation, collecting large amounts of data from a single individual may be a challenge. Future work should consider balancing the 2 approaches to better balance learning within a deployment for improved personalization (person-specific) and learning between deployments to continually improve a JITAI system (subpopulation-level).

Other opportunities for future research on JITAI-Twins include the use of large language models to improve the abstraction of domain science (often described using natural language) into computational constructs, the use of large language models to

XSL•FO
RenderX

simulate participant interactions with a JITAI [88], and the use of foundation models to leverage large sums of previously collected data that may be openly available but rife with data-deployment mismatches. See Section S4 of the Multimedia Appendix 1 for further discussion of future work.

Limitations

Our discussion of JITAI-Twins is limited in a few respects. The underlying concepts behind JITAI-Twins may find broader applicability beyond JITAI deployments. Other adaptive digital interventions may also benefit from digital twins. However, this paper is restricted to JITAI applications to ensure guidance is informed by practical experience. The example JITAI-Twins in this paper only include those that the coauthors of this paper have designed. Although others have designed JITAIs empowered by RL algorithms [89,90], the use of JITAI-Twins is not yet standardized and the publication of simulation environments and results is not yet commonplace, limiting our ability to include other simulations and JITAIs as examples. One of the goals of this paper is to bring JITAI-Twins to the forefront of discussion to overcome this gap. Future work is also necessary to better elucidate practical considerations, such as the computational and labor costs associated with designing and maintaining a JITAI-Twin and potential comparisons with the costs of conducting optimization experiments for the same design decisions. Other practical considerations include scalability. Although the framework as described is not dependent on the amount of data available or the scale of a JITAI's deployment, future work is necessary to evaluate the generalizability of the JITAI-Twin framework to larger-scale deployments. Guidance is needed on adapting and improving RL algorithms based on JITAI-Twin simulation results. Although the previous section of this paper provides initial guidance on RL algorithm selection, we leave further discussion to future work, as the scope of this paper focuses on JITAI-Twins. Future work is also needed to empirically validate the benefit of each of the proposed guiding principles and design steps for JITAI-Twins to potentially improve or provide additional evidence for aspects of the framework.

Conclusions

In this paper, we present the JITAI-Twin framework as an instantiation of "digital twins of a subpopulation" for JITAI optimization and continual improvement [24]. We detail how JITAI-Twin simulation results can be used to make difficult design decisions, such as which decision-making algorithm to deploy as part of a JITAI. This serves as information flow from the virtual twin to the physical twin. We also show how data collected from a JITAI's deployment can be used to update the JITAI-Twin prior to the subsequent deployment, demonstrating information flow from the physical twin to the virtual twin. JITAI-Twins are thus "fit for [the] purpose" of simulating the behavior of a JITAI deployment's subpopulation [13]. The impact that JITAI-Twins can have would be greatly enhanced through collaborative efforts to share computational resources for designing and evaluating JITAI-Twins (eg, codebases); curate data that could be used to initialize, train, or update JITAI-Twins; and establish benchmarks to help practitioners better understand when a JITAI-Twin is ready to inform decisions or to help researchers evaluate new methods for JITAI-Twin modeling or simulation. Toolboxes that enhance the accessibility of JITAI-Twin design (eg, as a possible addition to the JustIn platform [91]) would also facilitate more widespread adoption of JITAI-Twin methods to improve the optimization and continual improvement of JITAIs.

Acknowledgments

In preparing this article, generative artificial intelligence was used for tasks that fall within classification 1 of the taxonomy published by the International Association of Scientific, Technical & Medical Publishers: "refinement, correction, editing, or formatting the manuscript to improve the clarity of language."

Funding

The work of AHG was supported by Schmidt Science Fellows, in partnership with the Rhodes Trust. This research was funded by National Institutes of Health grants K99EB037411, P50DA054039, P41EB028242, R01HL125440, 2R01HL125440, U01CA229445, R01LM013107, UH3DE028723, and P30AG073107-03 GY3 Pilots.

Data Availability

This paper is accompanied by Multimedia Appendix 1. Section S1 provides a broader overview of related literature, Section S2 provides a primer on RL algorithms and simulation models in the context of JITAIs, Section S3 provides additional examples of how data impoverishment was addressed in prior work, and Section S4 describes open questions and future directions in JITAI-Twin design. Example simulation testbeds used as single deployment iterations of JITAI-Twins (ie, these simulation testbeds were used to inform a single deployment – future work is necessary to continually adapt these testbeds and form true JITAI-Twins) are available online [92,93].

Authors' Contributions

Conceptualization: AHG, SAM
Writing – original draft: AHG
Visualization: AHG, DG

XSL•FO
RenderX

Funding Acquisition: AHG, PK, SAM
Writing – review & editing: DG, SG, ZX, AT, PK, SAM
Supervision: SAM

Conflicts of Interest

SAM holds concurrent appointments at Harvard University and as an Amazon Scholar. This paper describes work performed at Harvard University and is not associated with Amazon.

Multimedia Appendix 1
Includes a more extensive review of related literature, a primer on reinforcement learning algorithms and simulation models, additional examples of data impoverishment faced during digital twin for just-in-time adaptive intervention design and updates, and additional discussion on open questions and future directions in digital twin for just-in-time adaptive intervention design.
[DOCX File, 310 KB - jmir_v28i1e72830_app1.docx]

References

1. Spruijt-Metz D, Nilsen W. Dynamic models of behavior for just-in-time adaptive interventions. IEEE Pervasive Comput 2014;13(3):13-17. [doi: 10.1109/MPRV.2014.46]
2. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462. [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
3. Nahum-Shani I, Hekler EB, Spruijt-Metz D. Building health behavior models to guide the development of just-in-time adaptive interventions: a pragmatic framework. Health Psychol 2015 Dec;34S:1209-1219. [doi: 10.1037/hea0000306] [Medline: 26651462]
4. Spruijt-Metz D, Wen CKF, O'Reilly G, et al. Innovations in the use of interactive technology to support weight management. Curr Obes Rep 2015 Dec;4(4):510-519. [doi: 10.1007/s13679-015-0183-6] [Medline: 26364308]
5. Klasnja P, Smith S, Seewald NJ, et al. Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of HeartSteps. Ann Behav Med 2019 May 3;53(6):573-582. [doi: 10.1093/abm/kay067] [Medline: 30192907]
6. Hekler EB, Rivera DE, Martin CA, et al. Tutorial for using control systems engineering to optimize adaptive mobile health interventions. J Med Internet Res 2018 Jun 28;20(6):e214. [doi: 10.2196/jmir.8622] [Medline: 29954725]
7. Trella AL, Zhang KW, Nahum-Shani I, Shetty V, Doshi-Velez F, Murphy SA. Designing reinforcement learning algorithms for digital interventions: pre-implementation guidelines. Algorithms 2022 Aug;15(8):255. [doi: 10.3390/a15080255] [Medline: 36713810]
8. Probst P, Boulesteix AL, Bischl B. Tunability: importance of hyperparameters of machine learning algorithms. J Mach Learn Res 2019;18:1-32. [doi: 10.5555/3322706.3361994]
9. Russell LB, Norton LA, Pagnotti D, et al. Using clinical trial data to estimate the costs of behavioral interventions for potential adopters: a guide for trialists. Med Decis Making 2021 Jan;41(1):9-20. [doi: 10.1177/0272989X20973160] [Medline: 33218296]
10. Monashefsky A, Alon D, Baranowski T, et al. How much did it cost to develop and implement an eHealth intervention for a minority children population that overlapped with the COVID-19 pandemic? Contemp Clin Trials 2023 Feb;125:107044. [doi: 10.1016/j.cct.2022.107044] [Medline: 36473682]
11. Mohr DC, Schueller SM, Riley WT, et al. Trials of intervention principles: evaluation methods for evolving behavioral intervention technologies. J Med Internet Res 2015 Jul 8;17(7):e166. [doi: 10.2196/jmir.4391] [Medline: 26155878]
12. An M, Dusing SC, Harbourne RT, Sheridan SM, START-Play Consortium. What really works in intervention? Using fidelity measures to support optimal outcomes. Phys Ther 2020 May 18;100(5):757-765. [doi: 10.1093/ptj/pzaa006] [Medline: 31944249]
13. National Academies of Sciences Engineering and Medicine. Foundational Research Gaps and Future Directions for Digital Twins: The National Academies Press; 2024:1-202. [doi: 10.17226/26894]
14. Qian T, Walton AE, Collins LM, et al. The microrandomized trial for developing digital interventions: experimental design and data analysis considerations. Psychol Methods 2022 Oct;27(5):874-894. [doi: 10.1037/met0000283] [Medline: 35025583]
15. Collins LM, Murphy SA, Strecher V. The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions. Am J Prev Med 2007 May;32(5 Suppl):S112-S118. [doi: 10.1016/j.amepre.2007.01.022] [Medline: 17466815]
16. Laubenbacher R, Mehrad B, Shmulevich I, Trayanova N. Digital twins in medicine. Nat Comput Sci 2024 Mar;4(3):184-191. [doi: 10.1038/s43588-024-00607-6] [Medline: 38532133]
17. Kamel Boulos MN, Zhang P. Digital twins: from personalised medicine to precision public health. J Pers Med 2021 Jul 29;11(8):745. [doi: 10.3390/jpm11080745] [Medline: 34442389]

XSL•FO
RenderX

18. Björnsson B, Borrebaeck C, Elander N, et al. Digital twins to personalize medicine. Genome Med 2020 Dec;12(1):1-4. [doi: 10.1186/s13073-019-0701-3]
19. Laubenbacher R, Sluka JP, Glazier JA. Using digital twins in viral infection. Science 2021 Mar 12;371(6534):1105-1106. [doi: 10.1126/science.abf3370] [Medline: 33707255]
20. Katsoulakis E, Wang Q, Wu H, et al. Digital twins for health: a scoping review. NPJ Digit Med 2024 Mar 22;7(1):77. [doi: 10.1038/s41746-024-01073-0] [Medline: 38519626]
21. Zhang K, Zhou HY, Baptista-Hon DT, et al. Concepts and applications of digital twins in healthcare and medicine. Patterns (N Y) 2024 Aug 9;5(8):101028. [doi: 10.1016/j.patter.2024.101028] [Medline: 39233690]
22. Elkefi S, Asan O. Digital twins for managing health care systems: rapid literature review. J Med Internet Res 2022 Aug 16;24(8):e37641. [doi: 10.2196/37641] [Medline: 35972776]
23. Armeni P, Polat I, De Rossi LM, Diaferia L, Meregalli S, Gatti A. Digital twins in healthcare: is it the beginning of a new era of evidence-based medicine? A critical review. J Pers Med 2022 Jul 30;12(8):8. [doi: 10.3390/jpm12081255] [Medline: 36013204]
24. National Academies of Sciences Engineering and Medicine. Opportunities and challenges for digital twins in biomedical research—proceedings of a workshop in brief 2023. : The National Academied Press; 2023. [doi: 10.17226/26922]
25. Gao D, Lai HY, Klasnja P, Murphy SA. Harnessing causality in reinforcement learning with bagged decision times. . Preprint posted online on Oct, 2024 URL: https://arxiv.org/abs/2410.14659v1 [accessed 2024-11-04]
26. Liao P, Greenewald K, Klasnja P, Murphy S. Personalized HeartSteps: a reinforcement learning algorithm for optimizing physical activity. Proc ACM Interact Mob Wearable Ubiquitous Technol 2020 Mar;4(1):1. [doi: 10.1145/3381007] [Medline: 34527853]
27. Spruijt-Metz D, Marlin BM, Pavel M, et al. Advancing behavioral intervention and theory development for mobile health: the HeartSteps II protocol. Int J Environ Res Public Health 2022 Feb 17;19(4):2267. [doi: 10.3390/ijerph19042267] [Medline: 35206455]
28. Science of Behavior Change (SOBC). National Institutes of Health. URL: https://commonfund.nih.gov/behaviorchange [accessed 2024-11-04]
29. Simonton KL. Testing a model of personal attributes and emotions regarding physical activity and sedentary behaviour. Int J Sport Exerc Psychol 2021 Sep 3;19(5):848-865. [doi: 10.1080/1612197X.2020.1739112]
30. Mullen SP, Olson EA, Phillips SM, et al. Measuring enjoyment of physical activity in older adults: invariance of the physical activity enjoyment scale (paces) across groups and time. Int J Behav Nutr Phys Act 2011 Sep 27;8(1):103. [doi: 10.1186/1479-5868-8-103] [Medline: 21951520]
31. Mohiyeddini C, Pauli R, Bauer S. The role of emotion in bridging the intention–behaviour gap: the case of sports participation. Psychol Sport Exerc 2009 Feb;10(2):226-234. [doi: 10.1016/j.psychsport.2008.08.005]
32. Ghosh S, Guo Y, Hung PY, et al. ReBandit: random effects based online RL algorithm for reducing cannabis use. . Preprint posted online on Feb, 2024 URL: http://arxiv.org/abs/2402.17739 [accessed 2024-08-13]
33. Holford NHG, Hale M, Ko HC, et al. Simulation in drug development: good practices. : CDDS; 1999 URL: https://holford.fmhs.auckland.ac.nz/docs/simulation-in-drug-development-good-practices.pdf [accessed 2025-02-23]
34. Ghosh S, Guo Y, Hung PY, et al. Reproducible workflow for online AI in digital health. Preprint posted online on Sep, 2025 URL: https://www.arxiv.org/pdf/2509.13499 [accessed 2025-09-29]
35. Hall KL, Vogel AL, Huang GC, et al. The science of team science: a review of the empirical evidence and research gaps on collaboration in science. Am Psychol 2018;73(4):532-548. [doi: 10.1037/amp0000319] [Medline: 29792466]
36. Box GEP. Science and statistics. J Am Stat Assoc 1976 Dec;71(356):791-799. [doi: 10.1080/01621459.1976.10480949]
37. Baker RE, Peña JM, Jayamohan J, Jérusalem A. Mechanistic models versus machine learning, a fight worth fighting for the biological community? Biol Lett 2018 May;14(5):5. [doi: 10.1098/rsbl.2017.0660] [Medline: 29769297]
38. Biecek P, Burzykowski T. Explore, explain, and examine predictive models. In: Explanatory Model Analysis: Chapman and Hall/CRC; 2021. [doi: 10.1201/9780429027192]
39. Shivaswamy P, Chandrashekar A. Bias-variance decomposition for ranking. 2021 Mar 8 Presented at: WSDM 2021 - Proceedings of the 14th ACM International Conference on Web Search and Data Mining; Virtual Event Israel p. 472-480. [doi: 10.1145/3437963.3441772]
40. Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. Nat Rev Phys 2021;3(6):422-440. [doi: 10.1038/s42254-021-00314-5]
41. Wagg D, Burr C, Shepherd J, Xuereb Conti Z, Enzer M, Niederer S. The philosophical foundations of digital twinning. . Preprint posted online on Jan, 2024. [doi: 10.31224/3500]
42. Geron A. Hands-on Machine Learning with Scikit-Learn, Keras & Tensorflow, 2nd edition: O'Reilly Media; 2019.
43. Forster MR. Key concepts in model selection: performance and generalizability. J Math Psychol 2000 Mar;44(1):205-231. [doi: 10.1006/jmps.1999.1284] [Medline: 10733865]
44. Kernbach JM, Staartjes VE. Foundations of machine learning-based clinical prediction modeling: Part II—generalization and overfitting. Acta Neurochir 2022:15-21. [doi: 10.1007/978-3-030-85292-4_3]

XSL•FO
RenderX

45. Kohavi R, Wolpert D. Bias plus variance decomposition for zero-one loss functions. 1996 Presented at: Proceedings of the Thirteenth International Conference on Machine Learning; Jul 3-6, 1996 URL: https://dl.acm.org/doi/10.5555/3091696.3091730 [accessed 2026-02-18]
46. Dar Y, Muthukumar V, Baraniuk RG. A farewell to the bias-variance tradeoff? An overview of the theory of overparameterized machine learning. . Preprint posted online on Sep, 2021 URL: https://arxiv.org/pdf/2109.02355 [accessed 2026-02-09]
47. Benk J, Beyersdorf F. Commentary: Inattention to detail is the hallmark of mediocrity. J Thorac Cardiovasc Surg 2022 Dec;164(6):1888-1889. [doi: 10.1016/j.jtcvs.2021.09.026] [Medline: 34602255]
48. Why the mars probe went off course - IEEE spectrum. IEEE Spectr. URL: https://spectrum.ieee.org/why-the-mars-probe-went-off-course [accessed 2025-09-29]
49. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in machine learning. J Big Data 2021;8(1):140. [doi: 10.1186/s40537-021-00516-9] [Medline: 34722113]
50. Singh J, Sato M, Ohkuma T. On missingness features in machine learning models for critical care: observational study. JMIR Med Inform 2021 Dec 8;9(12):e25022. [doi: 10.2196/25022] [Medline: 34889756]
51. Digitale JC, Martin JN, Glymour MM. Tutorial on directed acyclic graphs. J Clin Epidemiol 2022 Feb;142:264-267. [doi: 10.1016/j.jclinepi.2021.08.001] [Medline: 34371103]
52. Lipsky AM, Greenland S. Causal directed acyclic graphs. JAMA 2022 Mar 15;327(11):1083-1084. [doi: 10.1001/jama.2022.1816] [Medline: 35226050]
53. Trella AL, Zhang KW, Nahum-Shani I, Shetty V, Doshi-Velez F, Murphy SA. Reward design for an online reinforcement learning algorithm supporting oral self-care. Proc Innov Appl Artif Intell Conf 2023 Jun;37(13):15724-15730. [doi: 10.1609/aaai.v37i13.26866]
54. Turban E. Review of expert systems technology. IEEE Trans Eng Manage 1988 May;35(2):71-81. [doi: 10.1109/17.6007]
55. Sutton R, Barto A. Reinforcement Learning: An Introduction: MIT Press; 2018. URL: http://incompleteideas.net/book/the-book-2nd.html [accessed 2026-02-18]
56. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y) 2023 Sep 8;4(9):100804. [doi: 10.1016/j.patter.2023.100804] [Medline: 37720327]
57. Daniel JR, Benjamin V, Abbas K, Ian O, Zheng W. A tutorial on Thompson sampling. Found Trends Mach Learn 2018 Jul 12;11(1):1-99. [doi: 10.1561/2200000070]
58. van de Schoot R, Depaoli S, King R, et al. Bayesian statistics and modelling. Nat Rev Methods Primers 2021;1(1):1-26. [doi: 10.1038/s43586-020-00001-2]
59. Cuff A. The evolution of digital health and its continuing challenges. BMC Digit Health 2023;1(1):3. [doi: 10.1186/s44247-022-00004-x] [Medline: 38014367]
60. Sumner JA, Carey RN, Michie S, Johnston M, Edmondson D, Davidson KW. Using rigorous methods to advance behaviour change science. Nat Hum Behav 2018 Nov;2(11):797-799. [doi: 10.1038/s41562-018-0471-8] [Medline: 30931398]
61. Berntson GG, Cacioppo JT, Sarter M. Psychophysiology. Biol Psychiatry 2002 Jul:123-138. [doi: 10.1002/0470854871]
62. Bach DR, Castegnetti G, Korn CW, Gerster S, Melinscak F, Moser T. Psychophysiological modeling: current state and future directions. Psychophysiology 2018 Nov;55(11):e13214. [doi: 10.1111/psyp.13209] [Medline: 30175471]
63. Rabanser S, Unnemann S. Failing loudly: an empirical study of methods for detecting dataset shift. Presented at: Proceedings of the 33rd International Conference on Neural Information Processing Systems; Dec 8-14, 2019. [doi: 10.5555/3454287.3454412]
64. Van Cappellen P, Rice EL, Catalino LI, Fredrickson BL. Positive affective processes underlie positive health behaviour change. Psychol Health 2018 Jan;33(1):77-97. [doi: 10.1080/08870446.2017.1320798] [Medline: 28498722]
65. Rhodes RE, Kates A. Can the affective response to exercise predict future motives and physical activity behavior? A systematic review of published evidence. Ann Behav Med 2015 Oct;49(5):715-731. [doi: 10.1007/s12160-015-9704-5] [Medline: 25921307]
66. Brand R, Ekkekakis P. Affective–reflective theory of physical inactivity and exercise. Ger J Exerc Sport Res 2018 Mar;48(1):48-58. [doi: 10.1007/s12662-017-0477-9]
67. Williams DM, Dunsiger S, Ciccolo JT, Lewis BA, Albrecht AE, Marcus BH. Acute affective response to a moderate-intensity exercise stimulus predicts physical activity participation 6 and 12 months later. Psychol Sport Exerc 2008 May;9(3):231-245. [doi: 10.1016/j.psychsport.2007.04.002] [Medline: 18496608]
68. Sankaran K, Holmes SP. Generative models: an interdisciplinary perspective. Annu Rev Stat Appl 2023 Mar 10;10(1):325-352 [FREE Full text] [doi: 10.1146/annurev-statistics-033121-110134]
69. Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. Neural Comput 1992 Jan;4(1):1-58. [doi: 10.1162/neco.1992.4.1.1]
70. Roelof R, Fridovich-Keil S, Miller J, et al. A meta-analysis of overfitting in machine learning. 2019 Presented at: Proceedings of the 33rd International Conference on Neural Information Processing Systems; Dec 8-14, 2019. [doi: 10.5555/3454287.3455110]
71. Bryan CJ, Tipton E, Yeager DS. Behavioural science is unlikely to change the world without a heterogeneity revolution. Nat Hum Behav 2021 Aug;5(8):980-989. [doi: 10.1038/s41562-021-01143-3] [Medline: 34294901]

XSL•FO
RenderX

72. Nikbakht M, Gazi AH, Zia J, et al. Synthetic seismocardiogram generation using a transformer-based neural network. J Am Med Inform Assoc 2023 Jun 20;30(7):1266-1273. [doi: 10.1093/jamia/ocad067] [Medline: 37053380]
73. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11. [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
74. Sel K, Hawkins-Daarud A, Chaudhuri A, et al. Survey and perspective on verification, validation, and uncertainty quantification of digital twins for precision medicine. NPJ Digit Med 2025 Jan 17;8(1):40. [doi: 10.1038/s41746-025-01447-y] [Medline: 39825103]
75. Bingham D, Butler T, Estep D. Inverse problems for physics-based process models. Annu Rev Stat Appl 2024 Apr 22;11(1):461-482. [doi: 10.1146/annurev-statistics-031017-100108]
76. Song H, Song M, Liu X. Online autonomous calibration of digital twins using machine learning with application to nuclear power plants. Appl Energy 2022 Nov;326:119995. [doi: 10.1016/j.apenergy.2022.119995]
77. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. Nat Mach Intell 2019;1(1):20-23. [doi: 10.1038/s42256-018-0004-1]
78. Saltelli A, Aleksankina K, Becker W, et al. Why so many published sensitivity analyses are false: a systematic review of sensitivity analysis practices. Environ Model Softw 2019 Apr;114:29-39. [doi: 10.1016/j.envsoft.2019.01.012]
79. Gal Y, Koumoutsakos P, Lanusse F, Louppe G, Papadimitriou C. Bayesian uncertainty quantification for machine-learned models in physics. Nat Rev Phys 2022;4(9):573-577. [doi: 10.1038/s42254-022-00498-4]
80. Lin DJ, Gazi AH, Kimball J, Nikbakht M, Inan OT. Real-time seismocardiogram feature extraction using adaptive Gaussian mixture models. IEEE J Biomed Health Inform 2023 Aug;27(8):3889-3899. [doi: 10.1109/JBHI.2023.3273989] [Medline: 37155395]
81. Ellenberg SS, Shaw PA. Early termination of clinical trials for futility - considerations for a data and safety monitoring board. NEJM Evid 2022 Jul;1(7):EVIDctw2100020. [doi: 10.1056/EVIDctw2100020] [Medline: 38319261]
82. Trella AL, Zhang KW, Carpenter SM, et al. Oralytics reinforcement learning algorithm. . Preprint posted online on Jun, 2024 URL: https://arxiv.org/abs/2406.13127v1 [accessed 2024-08-13]
83. Trella AL, Zhang KW, Jajal H, et al. A deployed online reinforcement learning algorithm in an oral health clinical trial. . Preprint posted online on Sep, 2024 URL: https://arxiv.org/abs/2409.02069v1 [accessed 2024-09-13]
84. Ghosh S, Guo Y, Hung PY, et al. MiWaves reinforcement learning algorithm. Aug. Preprint posted online on 2024 URL: https://arxiv.org/abs/2408.15076v1 [accessed 2024-09-11]
85. Siroker D, Koomen P. A/B Testing: The Most Powerful Way to Turn Clicks into Customers: John Wiley & Sons; 2015. URL: https://books.google.com/books/about/A_B_Testing.html?id=9xUWCgAAQBAJ [accessed 2025-09-29]
86. Gazi AH, An S, Sanchez-Perez JA, et al. Modeling latent dynamics of the autonomic nervous system in response to trauma recall and non-invasive vagus nerve stimulation. IEEE Trans Biomed Eng 2025;73(1):220-233. [doi: 10.1109/TBME.2025.3580051]
87. Park J, Kim M, El Mistiri M, et al. Advancing understanding of just-in-time states for supporting physical activity (Project JustWalk JITAI): protocol for a system ID study of just-in-time adaptive interventions. JMIR Res Protoc 2023 Sep 26;12:e52161. [doi: 10.2196/52161] [Medline: 37751237]
88. Karine K, Marlin BM. Enhancing adaptive behavioral interventions with LLM inference from participant-described states. Jul. Preprint posted online on 2025 URL: https://arxiv.org/pdf/2507.03871 [accessed 2025-09-30]
89. Figueroa CA, Aguilera A, Chakraborty B, et al. Adaptive learning algorithms to optimize mobile applications for behavioral health: guidelines for design decisions. J Am Med Inform Assoc 2021 Jun 12;28(6):1225-1234. [doi: 10.1093/jamia/ocab001] [Medline: 33657217]
90. Gönül S, Namlı T, Coşar A, Toroslu İ. A reinforcement learning based algorithm for personalization of digital, just-in-time, adaptive interventions. Artif Intell Med 2021 May;115:102062. [doi: 10.1016/j.artmed.2021.102062] [Medline: 34001322]
91. Introducing JustIn. d3center. URL: https://d3c.isr.umich.edu/justin/ [accessed 2025-09-29]
92. ReBandit. GitHub. URL: https://github.com/susobhang70/reBandit [accessed 2026-02-09]
93. StatisticalReinforcementLearningLab. GitHub. URL: https://github.com/StatisticalReinforcementLearningLab/oralytics_algorithm_design [accessed 2026-02-09]

Abbreviations

AI: artificial intelligence
DAG: directed acyclic graph
JITAI: just-in-time adaptive intervention
JITAI-Twin: digital twin for just-in-time adaptive intervention
mHealth: mobile health
MRT: microrandomized trial
RL: reinforcement learning
VVUQ: verification, validation, and uncertainty quantification

XSL•FO
RenderX

Edited by J Sarvestan; submitted 09.May.2025; peer-reviewed by KH Lin, VG Prabhu; revised version received 19.Dec.2025; accepted 20.Dec.2025; published 06.Mar.2026.

Please cite as:
Gazi AH, Gao D, Ghosh S, Xu Z, Trella AL, Klasnja P, Murphy SA
Digital Twins for Just-in-Time Adaptive Interventions (JITAIs): Framework for Optimizing and Continually Improving JITAIs
J Med Internet Res 2026;28:e72830
URL: https://www.jmir.org/2026/1/e72830
doi:10.2196/72830

© Asim H Gazi, Daiqi Gao, Susobhan Ghosh, Ziping Xu, Anna L Trella, Predrag Klasnja, Susan A Murphy. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI Agents and Epidemic Intelligence on Respiratory Infectious Diseases: Toward a Conceptual Framework Integrating Decision Support

Liuyang Yang[1,2,3*], PhD; Liyu Shan[1*], MD; Xiaolin Cao[4*], PhD; Jinzhao Cui[2], MPH; Michael Tong[5], PhD; Yan Niu[6], PhD; Ting Zhang[2], PhD

[1]The Third Affiliated Hospital of Kunming Medical University, Yunnan Cancer Hospital, Peking University Cancer Hospital Yunnan, Kunming, Yunnan, China

[2]School of Population Medicine and Public Health, Chinese Academy of Medical Sciences & Peking Union Medical College, No. 9 Dongdan Santiao, Beijing, Beijing, China

[3]School of Data Science, Fudan University, Shanghai, Shanghai, China

[4]Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, China

[5]National Centre for Epidemiology and Population Health, The Australian National University, Canberra, Australian Capital Territory 2601, Australia

[6]Public Health Emergency Center, Chinese Center for Disease Control and Prevention, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Ting Zhang, PhD
School of Population Medicine and Public Health, Chinese Academy of Medical Sciences & Peking Union Medical College, No. 9 Dongdan Santiao, Beijing, Beijing, China

Abstract

Traditional epidemic intelligence relies heavily on human epidemiologists for data interpretation and reporting, which makes it resource intensive, slow to respond, and vulnerable to variability in professional expertise. To overcome these limitations, we propose an expanded conceptual epidemic intelligence quadripartite framework that extends the classical trinity of (1) surveillance, (2) risk evaluation, and (3) early warning with a fourth pillar, (4) decision support and intervention optimization through AI agents. Acting as 24/7 digital epidemiologists, multiagent systems can integrate heterogeneous signals from multisource surveillance systems, conduct contextual risk evaluation and adaptive forecasting, generate tailored early warnings, and provide actionable recommendations for targeted control—closing the loop between detection and response. Embedding interpretability and mandatory human-in-the-loop oversight enhances trust and accountability. Nonetheless, real-world deployment requires addressing context-specific challenges of data quality, interoperability, robustness, governance, circular reporting, and equity. If designed with transparency, inclusiveness, and resilience, AI agents have the potential to transform epidemic intelligence into a continuously adaptive and globally connected system.

(J Med Internet Res 2026;28:e86936) doi:10.2196/86936

KEYWORDS

artificial intelligence agents; respiratory infectious diseases; surveillance; risk evaluation; early warning; decision support

Introduction

Respiratory diseases, such as seasonal influenza, respiratory syncytial virus (RSV), and emerging zoonotic respiratory pathogens, remain among the most significant causes of global morbidity and mortality. Seasonal influenza infects up to 1 billion people annually, causing an estimated 290,000 to 650,000 deaths each year [1], while RSV is responsible for approximately 33 million cases and 100,000 child deaths worldwide each year [2]. The rapid transmissibility, unpredictable evolution, and potential for interspecies spillover events make these pathogens a persistent global health threat. Recent outbreaks, such as the COVID-19 pandemic, underscore the urgent need for advanced epidemic intelligence. In particular, capturing the critical "golden window"— a short but decisive period during which timely detection and intervention can accelerate responses such as vaccine development and deployment—is essential to limit transmission, prevent missed opportunities for effective containment, and reduce the broader societal and economic impacts of respiratory infectious diseases.

Traditional respiratory disease surveillance systems, relying on clinician reporting, laboratory confirmation, and manual data collation, are often constrained by delayed signal detection and fragmented workflows. While event-based surveillance platforms, such as ProMED-mail and HealthMap, have integrated nontraditional signals, the transition from detection

XSL•FO
RenderX

to actionable response remains suboptimal. Emerging artificial intelligence (AI)–driven frameworks, such as BlueDot [1] and the World Health Organization (WHO) Mosaic surveillance framework [2,3], demonstrate that fusing multisource signals with real-time analytics can substantially reduce analytical latency. However, technical acceleration alone cannot bridge the gap to intervention; the "golden window" is frequently missed due to governance bottlenecks, including bureaucratic friction and varying political will [4]. Consequently, there is an evolving need for orchestrated multiagent systems (MAS) that move beyond passive dashboards to provide active, evidence-based decision support, helping to navigate both data complexity and institutional inertia [5].

The "epidemic intelligence trinity," a framework recently consolidated by the authors, comprises three functional stages: (1) surveillance and detection, (2) risk evaluation, and (3) early warning [6]. Surveillance aggregates heterogeneous outbreak indicators, risk evaluation contextualizes them to estimate potential spread and impact, and early warning translates these insights into alerts and intervention triggers. However, in current practice, these stages often remain linear and siloed. Information flows from surveillance to warning but often lacks the feedback loop required to optimize interventions immediately.

This Viewpoint presents a conceptual framework that advances beyond existing models by introducing "decision support" as a fourth, explicit pillar and using AI agents to drive the workflow. Unlike the WHO Mosaic framework, which primarily focuses on the diversity of data sources ("tiles"), or the traditional trinity, which emphasizes situational awareness, our quadripartite framework prioritizes autonomy and action. It proposes moving from passive dashboards to active MAS that not only display data but autonomously reason over it to suggest specific interventions, closing the loop between intelligence and response. Throughout this paper, we illustrate this potential through a recurring case scenario of a hypothetical zoonotic spillover to demonstrate tangible impacts.

AI agents—autonomous computational entities capable of perceiving their environment, reasoning, and taking actions—offer a pathway to unify this process. Moving beyond static rule-based algorithms, modern AI agents use large language models (LLMs) to perform chain-of-thought reasoning and tool orchestration. This viewpoint examines how AI agents can enhance respiratory epidemic intelligence within this expanded framework (Figure 1).

Figure 1. Artificial intelligence (AI) agents–driven integrated closed-loop framework for respiratory infectious disease surveillance, risk evaluation, early warning, and decision support. The framework illustrates an autonomous cycle for respiratory infectious disease management consisting of four interactive stages: (1) surveillance and detection—autonomous collection, integration, and preprocessing of multisource data; (2) risk evaluation—automated prediction and estimation of epidemic risk using machine learning and deep learning; (3) early warning—generation of timely alerts based on risk analysis; and (4) decision support—provision of targeted prevention and control recommendations through a decision database. At the core, the AI agent serves as the "brain," performing reasoning and orchestration to manage data flow and workflow across these stages. The framework supports a hybrid interaction model: it primarily operates in a proactive mode by pushing integrated "surveillance-to-action" reports to decision-makers, while simultaneously supporting reactive queries from expert supervisors through an iterative question and answer process.

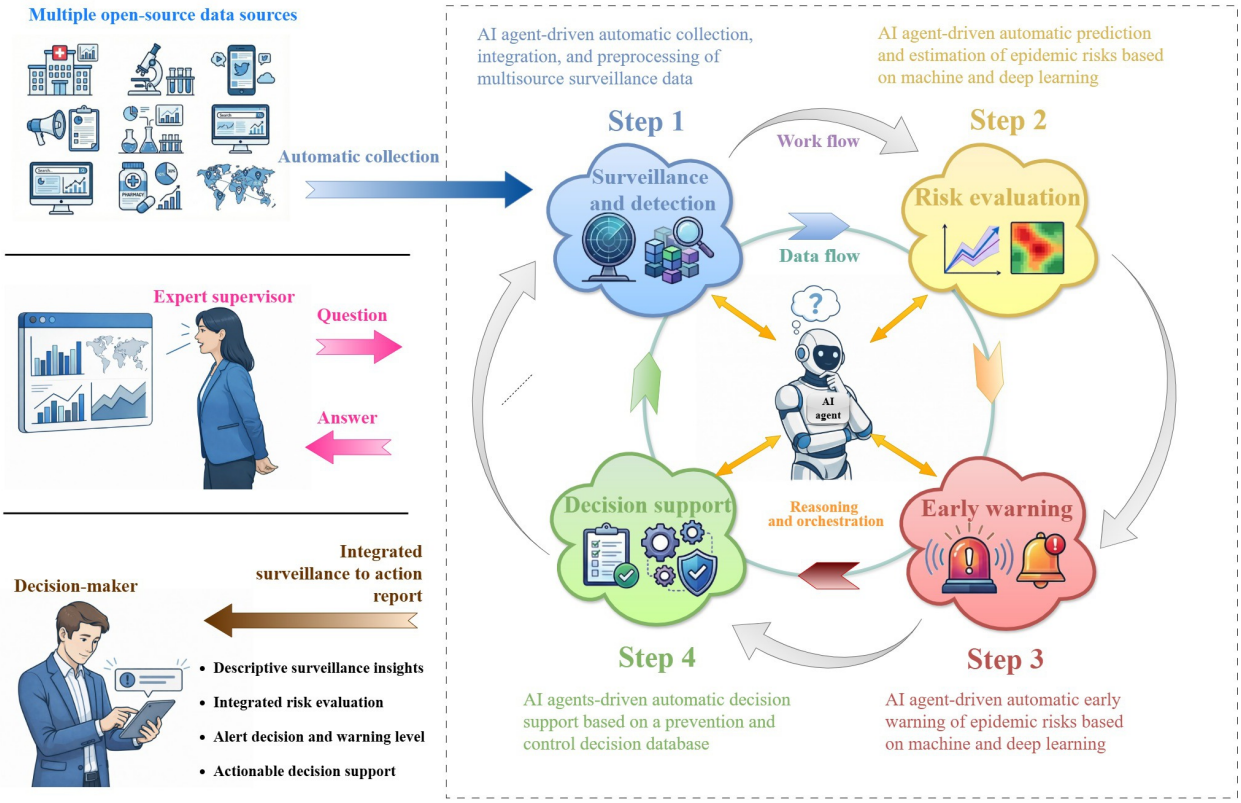

XSL•FO
RenderX

Defining AI Agents in Epidemiology

To understand the potential of this framework, it is necessary to define what constitutes an "AI agent" in this context. Unlike traditional deep learning models that passively process input to produce a prediction (eg, forecasting case counts), an AI agent is an autonomous system capable of three core functions.

Second, reasoning—using LLMs as a cognitive engine, the agent decomposes complex epidemiological problems into steps (chain-of-thought), interprets context, and plans tasks. For example, it can decide that a spike in "fever" search terms requires validation against pharmacy sales data.

Third, action—the agent executes tasks using external tools (application programming interface; API), such as running a transmission simulation model, drafting an alert email to health authorities, or querying a clinical guideline database.

In a MAS, specialized agents collaborate, handing off tasks to one another. This shift allows epidemic intelligence to move from a human-dependent "pull" system to an AI-driven "push" system, where the software proactively identifies threats and proposes solutions.

Surveillance and Detection

Surveillance and detection constitute the first pillar of epidemic intelligence, aiming to identify early signals of outbreaks or abnormal disease activities. Traditional respiratory surveillance predominantly relies on single-source data, such as hospital reports of influenza-like illness (ILI), laboratory confirmation of pathogens, or clinician-reported unusual cases. While effective for confirmed cases, this passive and delayed approach may fail to detect emerging threats, particularly novel or geographically dispersed events.

In contrast, multisource "mosaic" surveillance integrates heterogeneous data streams to create a more sensitive and timely view of epidemic activity. The WHO has advocated a mosaic framework for respiratory virus surveillance, emphasizing that combining multiple "tiles" of data produces a more complete outbreak picture [2].

Given the complexity of diverse data modalities, an MAS approach is superior to a single monolithic agent. Specialized agents can autonomously ingest data from formal channels, such as EHRs, laboratory test results, pharmacy sales, and sentinel clinic reports, and from informal sources, including news media, ProMED-mail bulletins, social media trends, search index [7], participatory symptom reports, animal health bulletins, and environmental or climate datasets. Unconventional indicators, such as increased sales of cough or cold medication, spikes in fever-related online searches, and elevated SARS-CoV-2 RNA levels in wastewater, have been shown to precede clinical surges by days or weeks [8]. AI agents can continuously integrate and analyze such signals, functioning as tireless "digital epidemiologists."

Traditional ILI surveillance has relied heavily on passive reporting, where clinicians manually record and submit case counts after aggregating data from their observations. This approach is inherently subjective, prone to underreporting, and often delayed. Increasingly, health systems are shifting to automated extraction of ILI indicators from EHRs, leveraging structured fields such as chief concerns, physical examination findings, and diagnostic codes. These systems can automatically flag and count encounters matching ILI case definitions in near real time.

Natural language processing further enhances surveillance by enabling AI agents to mine structured and unstructured data for outbreak clues. In addition to EHRs, agents can process informal text sources such as local news and social media [9], extracting disease-relevant facts, classifying signals by syndrome, and distinguishing credible reports from rumors. For instance, an agent might detect an increase in local language posts mentioning "pneumonia," cross-reference this with wastewater testing data showing a coronavirus RNA spike, and confirm concurrent increases in web searches for "fever" and "cough." The system can produce composite alerts with higher confidence and shorter detection lag than manual reporting or single-source monitoring by fusing such diverse signals. A critical advantage of AI agents lies in their ability to filter out noise while retaining sensitivity to subtle signals. In contrast, traditional systems often raise alert thresholds to avoid false alarms, which may result in missed events.

Moreover, AI agents can flag low-intensity anomalies and internally validate them against other data streams or reference knowledge bases. For instance, an agent detecting a sudden rise in emergency visits for respiratory illness could immediately check for similar patterns in nearby regions or seek plausible explanations, such as environmental events. The system would escalate an alert to human experts only when multiple indicators converge, thereby reducing false positives while preserving early warning capacity.

By automating and democratizing multisource data collection and analysis, AI agents extend surveillance to regions with limited public health infrastructure. This capability shortens detection lead times and reduces "blind spots" in low-resource settings, contributing to more equitable and timely global epidemic intelligence.

Case Scenario

An AI agent designed for respiratory threats might continuously scan unstructured data. In a hypothetical scenario of a novel zoonotic spillover, the "surveillance agent" detects a statistically significant cluster of social media posts mentioning "severe cough" and "poultry die-off" in a specific province. Simultaneously, it cross-references this with real-time wastewater data showing an unidentified viral RNA spike in the same locality. Unlike a passive dashboard, the agent autonomously links these disparate signals and flags them for immediate processing. The practical superiority of these orchestrated systems in reducing analytical latency is further evidenced by recent empirical studies, which show significant improvements in lead time and sensitivity compared to traditional methods (Table 1 provides a detailed comparison).

XSL•FO
RenderX

Table . Comparison of traditional surveillance systems vs artificial intelligence (AI)–driven systems.

Study	Conceptual approach and data integration	Performance advantage (timeliness and accuracy)	Role in decision support and public health utility
Du et al [5]	LLM[a]-driven multimodal fusion: uses LLMs to synthesize heterogeneous data streams (news, reports, and clinical data) dynamically rather than static statistical modeling.	Real-time integration: outperforms baseline models in complex dynamics by processing unstructured data in real time.	Operational readiness and resource mobilization: provides rapid situational awareness to justify early resource allocation before clinical confirmation is available.
Idahor et al [10]	Paradigm shift (reactive to proactive): a theoretical framework contrasting rigid, siloed traditional reporting with fluid, big data–driven AI ecosystems.	Lag reduction: traditional systems lag by weeks; AI enables near–real-time monitoring by bypassing bureaucratic reporting delays.	Strategic planning and gap analysis: supports high-level policy shifts from reactive containment to proactive mitigation; identifies "blind spots" in existing infrastructure.
Xie et al [9]	Digital signal detection (social sensing): leverages LLMs to interpret social media sentiment and symptom reporting (Twitter [subsequently rebranded X, X Corp]) as a proxy for early epidemic signals.	Early warning (approximately 1 week lead): achieves approximately 7.63 days lead time over official epidemiological data; strong temporal correlation.	Preclinical alerting and risk communication: triggers public health messaging and targeted testing before hospitals see a surge, allowing for "flattening the curve" interventions.
Zhang et al [11]	Hybrid surveillance (environmental+clinical): integrates wastewater viral load monitoring with traditional clinical case reporting to form a dual verification system.	Leading indicator (0-7 days): wastewater signals precede clinical reports; combined systems show higher robustness than single-source methods.	Noninvasive community risk assessment: enables decision-makers to assess community transmission levels without mass individual testing; guides localized lockdowns or interventions.
Zhang et al [12]	Algorithmic anomaly detection: compares machine learning (supervised or unsupervised) against static statistical thresholds (ADTM)[b] for outbreak detection.	Sensitivity enhancement: warnings issued approximately 5 days earlier. Unsupervised ML[c] achieved 100% sensitivity vs 71% for traditional thresholds.	Automated triage and sensitivity tuning: reduces "missed alarms" for critical threats; allows officials to adjust sensitivity thresholds based on available response capacity.

[a]LLM: large language model.

[b]ADTM: algorithmic detection and threshold monitoring.

[c]ML: machine learning.

Risk Evaluation

Risk evaluation is the second core function of epidemic intelligence in respiratory infectious disease, focusing on interpreting detected signals to determine an outbreak's likelihood, severity, and potential consequences. While detection identifies unusual patterns, risk evaluation contextualizes them, distinguishing events requiring urgent intervention from those of limited concern [3].

Traditional risk assessments rely on expert review of epidemiological, clinical, and contextual information. However, manual processes can be slow, particularly when integrating heterogeneous datasets such as pathogen genomic sequences, case demographics, population immunity, mobility data, and health care system capacity. AI agents can accelerate this process by automating the collation, integration, and preliminary analysis of such datasets in near real time.

Modern AI agents can apply predictive models to estimate outbreak growth rates, reproduction numbers, and geographical spread, incorporating uncertainties from multiple data sources [13]. They can also use machine learning to identify populations or regions at higher risk based on vaccination coverage, comorbidities, demographic profiles, and environmental factors. For example, an agent might assess whether a novel influenza strain detected in wastewater coincides with low immunity levels in nearby populations, indicating a heightened potential for community spread.

Importantly, risk evaluation by AI agents is not limited to epidemiological modeling. AI agents can incorporate operational factors, such as testing capacity, supply chain status for antivirals, or intensive care unit (ICU) bed availability, to provide decision-makers with a comprehensive risk profile. However, it is important to acknowledge that obtaining real-time operational data is challenging due to the fragmented nature of health information systems and inventory databases. These assessments can be updated dynamically as new information becomes available, ensuring that public health authorities receive timely, data-driven insights to guide intervention priorities.

Case Scenario

On receiving the signal from the surveillance agent, the "risk evaluation agent" automatically initiates a probabilistic assessment. It retrieves current population immunity data for influenza subtypes in the affected region, factors in local hospital capacity (ICU bed availability), and runs a transmission simulation. It calculates that the effective reproduction number (Rt) could exceed 2.5 within 7 days if left unchecked, classifying the event as "high risk."

XSL•FO
RenderX

Early Warning

Early warning aims to alert relevant stakeholders quickly enough for interventions to prevent widespread transmission [14,15]. Traditional systems often delay alerts until events are verified through official channels, which can reduce timeliness. AI agents can bridge this gap by issuing graded alerts based on probability thresholds, while continuing to refine assessments as new data become available.

An AI agent's early warning process typically follows three steps: (1) aggregating multisource indicators into a composite risk score; (2) comparing the score against predefined alert thresholds; and (3) communicating the alert in structured, actionable formats. By using anomaly detection and probabilistic forecasting models, agents can provide both short-term outbreak forecasts and medium-term epidemic scenarios [16].

Unlike static rule-based systems, AI agents can adapt their alerting logic over time by incorporating feedback from both human experts and real-world outcomes, thereby creating a continuously improving feedback loop. In practice, this adaptation relies on human-in-the-loop review, where epidemiologists confirm or dismiss alerts, as well as retrospective validation once ground truth data become available. Alerts can be stratified by urgency; for example, an "advisory" for low-probability but high-impact threats and an "action alert" when confidence is high and immediate measures are recommended.

Dissemination channels can also be diversified: AI agents can automatically generate reports for public health dashboards, push notifications for field epidemiologists, and plain language summaries for community stakeholders. This ensures that early warnings reach the right audience in time to act, a critical factor in respiratory pathogen control where delays of even a few days can lead to exponential case growth.

Case Scenario

The "early warning agent" takes the "high risk" classification and drafts a stratified alert. It generates a technical "action alert" for the local Centers for Disease Control, detailing the viral anomalies and projected trajectory. Simultaneously, it prepares a plain language "advisory" draft for community health workers, emphasizing vigilance for patients who have had contact with poultry. These drafts are held for human verification but are prepared instantly, saving hours of manual drafting.

Decision Support

Decision support extends beyond warning, guiding the choice and timing of interventions to minimize health and socioeconomic impacts. Respiratory epidemic control includes recommendations for targeted vaccination, nonpharmaceutical interventions, antiviral distribution, and communication strategies.

AI agents can synthesize epidemiological forecasts, resource availability, and intervention effectiveness data to recommend optimal response strategies [4]. For example, given an emerging RSV outbreak in a region with limited ICU capacity, an agent might recommend targeted nonpharmaceutical interventions for high-risk populations, rapid deployment of prophylactic treatments, and public messaging to reduce exposure risk.

Given the severe safety risks associated with LLM hallucination (fabrication) in medical contexts, decision support modules must implement strict guardrails. Recommendations must be grounded in verified clinical protocols and subjected to retrieval-augmented generation techniques to ensure factual accuracy. Decision support also benefits from scenario simulation. AI agents can run counterfactual models to estimate their potential impact by comparing projected outcomes with and without specific interventions. These simulations can incorporate real-time updates, allowing response plans to adapt dynamically as the situation evolves.

Transparency is essential: decision recommendations must be explainable, clearly indicating underlying assumptions, data sources, and model limitations. This ensures that human decision-makers can assess the validity and acceptability of AI-driven guidance before it is implemented. In this role, AI agents act not as autonomous decision-makers but as analytical partners—augmenting human expertise with rapid, data-rich, and evidence-based insights that can shorten reaction times and improve intervention precision.

Case Scenario

The "decision support agent" queries a database of evidence-based interventions. On the basis of the projected Rt and local resource constraints, it proposes a ranked list of interventions: (1) immediate targeted closure of live poultry markets in the identified district, (2) reallocation of antiviral stockpiles from the neighboring province, and (3) activation of a fever-screening protocol at local transport hubs. These recommendations are presented to human decision-makers with citations to relevant clinical protocols.

Challenges and Practical Implementation Guidance

While promising, real-world deployment demands addressing specific challenges through concrete practical measures.

Data Quality and Interoperability

AI agents depend on high-quality, diverse inputs. Informal sources may be noisy or biased, while data gaps in low-income regions risk reinforcing inequities, necessitating careful curation and cross-verification to prevent the system from optimizing primarily for data-rich regions. A critical risk is "circular reporting," where agents amplify their own signals; therefore, rigorous source provenance tracking is essential to distinguish between organic signals and AI-generated content. To address data fragmentation, such as inconsistent formats in wastewater surveillance reports or a lack of standardized electronic medical record coding, implementation must adhere to open standards, such as HL7 Fast Healthcare Interoperability Resources and WHO protocols, to ensure seamless interoperability [12].

XSL•FO
RenderX

Furthermore, to overcome data sovereignty barriers, privacy-preserving technologies such as federated learning can enable global model training without transferring sensitive local data.

Model Robustness and Validation

Complex models risk false positives or false negatives, potentially triggering inappropriate responses or missing critical threats. Rigorous validation is necessary, measuring lead time, sensitivity, specificity, and forecast accuracy against benchmarks. However, optimizing these metrics requires a delicate balance; an overly sensitive system risks generating "alert fatigue," where a high volume of low-confidence warnings desensitizes human analysts. If operators are inundated with false alarms, they may eventually ignore genuine threats, rendering the surveillance system ineffective. Validating on historical data using internet-connected LLMs carries a risk of "look-ahead bias," as these models may have "seen" the outcomes during training. Therefore, "prospective shadow testing" is recommended as a more rigorous alternative, where the agent runs in parallel to existing systems without releasing active alerts, allowing for safe evaluation of performance. Transparent, adjustable decision thresholds and iterative human-AI feedback loops can enhance reliability. Initial deployment should involve close human oversight, with gradual autonomy as confidence in model performance grows.

Explainability and Trust

High-stakes decisions require explainable outputs. Explainable AI approaches can improve transparency by highlighting relevant data signals and causal rules. LLM agents can generate narrative rationales, but these must reflect actual reasoning to avoid conveying misleading confidence. Trust in AI agents will grow as they demonstrate value under expert supervision; however, high-profile errors could set progress back.

Political and Bureaucratic Friction

Even the most precise AI-driven warnings may encounter friction as they move upward through institutional hierarchies. Decisions such as introducing lockdowns or prioritizing vaccine distribution are not solely forecasting problems; they are high-stakes governance and optimization problems influenced by economic concerns, political incentives, and competing narratives. AI agents must therefore be viewed not as a panacea for delay but as tools to provide clearer evidence-based support under radical uncertainty, potentially reducing, although not eliminating, the friction within political decision processes.

Ethical and Legal Considerations

Privacy, consent, and accountability are critical concerns. AI agents must respect anonymity and avoid individual identification without consent. Legal frameworks for automated public health decisions remain underdeveloped, raising questions of liability in the event of errors. Human-in-the-loop governance must be framed not merely as a method for enhancing trust but as a mandatory legal and safety requirement before any "decision support" outputs are acted upon. Equity demands ensuring benefits for low-resource settings, and WHO-led global AI intelligence services could help provide access for all countries [17]. Furthermore, the economic sustainability of 24/7 digital epidemiologists must be considered, as continuous inference on large models is computationally expensive [18].

Cybersecurity and Misuse

Adversarial attacks, such as false data injection, could disrupt or manipulate surveillance outputs. Systems must authenticate data and resist manipulation. Risks of misuse, including political exploitation for restrictive measures, necessitate algorithmic transparency, independent audits, and adherence to human rights norms.

Workforce, Infrastructure, and Human-AI Interaction

Effective integration requires both skilled personnel and adequate infrastructure. Human epidemiologists should work alongside AI agents (digital epidemiologists) to validate outputs and enrich datasets. However, we must strictly mitigate the risks of "automation bias" and overreliance. If public health officials passively accept algorithmic outputs without scrutiny, it can lead to dangerous "misprioritization," where resources are diverted based on high-confidence but erroneous model predictions. Furthermore, long-term dependence on automated agents risks a "loss of expertise," where the workforce's ability to independently interpret raw epidemiological signals and nuances degrades. Therefore, workflows must be designed to keep humans cognitively engaged—using AI as a "triage tool" rather than a replacement—and regular "manual mode" training simulations should be mandated to maintain core analytical competencies. Low-resource settings may rely on cloud-based services to host AI capabilities. Regular evaluations, including simulated outbreak drills, are necessary to test readiness and refine performance.

Strategic Road Map

To transition this framework from concept to reality, we propose four priority actions: (1) standardization—develop global data dictionaries and API standards for respiratory surveillance to facilitate cross-border AI agent collaboration; (2) validation—establish standard "epidemic benchmark datasets" to rigorously test AI agent performance against historical ground truths; (3) pilot programs—deploy shadow-mode AI agents in sentinel surveillance sites to evaluate real-world utility and workflow integration; and (4) governance frameworks—draft international guidelines defining liability and accountability for AI-supported public health decisions.

Conclusions

The convergence of advanced AI and the urgent need for improved epidemic intelligence offers a transformative opportunity. By extending the traditional trinity into a proactive loop with decision support, AI agents embody the "universal

XSL•FO
RenderX

epidemic intelligence model" [7]. Through autonomous workflows, ranging from detecting weak signals in wastewater to recommending targeted market closures, agents can contextualize respiratory outbreaks days or weeks sooner than traditional methods. Realizing this vision demands investment, transparency, and equity, alongside the practical safeguards and road map outlined earlier. If implemented responsibly, these "force multipliers" can network globally to create a distributed immune system, unifying fragmented data into actionable knowledge—a capability essential in an interconnected world where pathogens ignore borders. The integration of AI agents into epidemic intelligence represents a transformative opportunity to enhance global health security. By addressing the outlined challenges and implementing the proposed road map, we can create a continuously adaptive and globally connected system. We call upon policymakers, researchers, and public health professionals to collaborate in realizing this vision, ensuring a more resilient future against respiratory infectious diseases.

Acknowledgments

The authors thank colleagues and collaborators who provided valuable discussions and conceptual input during the development of this manuscript. The authors also acknowledge the support of institutional research platforms that facilitated literature retrieval and manuscript preparation. No additional individuals or organizations contributed in a manner that meets the criteria for authorship. The authors declare that no generative artificial intelligence tools were used in the writing, data analysis, figure preparation, or editing of this manuscript. All content was produced, reviewed, and approved solely by the authors, who take full responsibility for the accuracy, originality, and integrity of the work.

Funding

This research was funded by the National Key Research and Development Program of China (grant 2023YFC2308701) and Yunnan Fundamental Research Projects (grant 202501AU070024).

Authors' Contributions

LY: conceptualization, framework design, writing—original draft, funding acquisition
LS: conceptualization, visualization, writing—review and editing
XC: conceptualization, literature review, writing—review and editing
JC: literature review, writing—review and editing
MT: conceptualization, critical revision of the manuscript
YN: writing—review and editing
TZ: supervision, project administration, funding acquisition, critical revision of the manuscript
All authors: intellectual development of the manuscript and final approval

Conflicts of Interest

None declared.

References

1. Transforming public health surveillance. BlueDot. URL: https://bluedot.global/transforming-public-health-surveillance/ [accessed 2026-02-22]
2. Mott JA, Bergeri I, Lewis HC, Mounts AW, Briand SC. Facing the future of respiratory virus surveillance: "the mosaic surveillance framework". Influenza Other Respir Viruses 2023 Mar 21;17(3):e13122. [doi: 10.1111/irv.13122] [Medline: 36970570]
3. Morgan OW, Abdelmalik P, Perez-Gutierrez E, et al. How better pandemic and epidemic intelligence will prepare the world for future threats. Nat Med 2022 Aug;28(8):1526-1528. [doi: 10.1038/s41591-022-01900-5] [Medline: 35764683]
4. Koren O, Weidmann NB. Infectious disease outbreaks drive political mistrust. Proc Natl Acad Sci U S A 2025 Jul 22;122(29):e2506093122. [doi: 10.1073/pnas.2506093122] [Medline: 40674429]
5. Du H, Zhao Y, Zhao J, et al. Advancing real-time infectious disease forecasting using large language models. Nat Comput Sci 2025 Jun;5(6):467-480. [doi: 10.1038/s43588-025-00798-6] [Medline: 40481184]
6. Zhang T, Feng L, Lai S, Peng Z, Lan Y, Yang W. Epidemic intelligence trinity: detection, risk assessment, and early warning. Chin Med J (Engl) 2024 Jun 5;137(11):1313-1315. [doi: 10.1097/CM9.0000000000002856] [Medline: 37758650]
7. Yang L, Zhang T, Han X, et al. Influenza epidemic trend surveillance and prediction based on search engine data: deep learning model study. J Med Internet Res 2023 Oct 17;25:e45085. [doi: 10.2196/45085] [Medline: 37847532]
8. Wu F, Zhang J, Xiao A, et al. SARS-CoV-2 titers in wastewater are higher than expected from clinically confirmed cases. mSystems 2020 Jul 21;5(4):e00614-20. [doi: 10.1128/mSystems.00614-20] [Medline: 32694130]

XSL•FO
RenderX

9. Xie J, Zhang Z, Zeng S, et al. Leveraging large language models for infectious disease surveillance-using a web service for monitoring COVID-19 patterns from self-reporting tweets: content analysis. J Med Internet Res 2025 Feb 20;27:e63190. [doi: 10.2196/63190] [Medline: 39977859]
10. Idahor CO, Esomu EJO, Ogbonna N, et al. Infectious disease surveillance in the era of big data and AI: opportunities and pitfalls. Cureus 2025 Oct;17(10):e93929. [doi: 10.7759/cureus.93929] [Medline: 41200593]
11. Zhang X, Lei Z, Wang Q, et al. Comparative performance of wastewater, clinical, and digital surveillance indicators for COVID-19 monitoring in routine practice: retrospective observational study. J Med Internet Res 2025 Nov 6;27:e70232. [doi: 10.2196/70232]
12. Zhang T, Yang L, Fan Z, et al. Comparison between threshold method and artificial intelligence approaches for early warning of respiratory infectious diseases - Weifang City, Shandong Province, China, 2020-2023. China CDC Wkly 2024 Jun 28;6(26):635-641. [doi: 10.46234/ccdcw2024.119] [Medline: 38966311]
13. Reich NG, Brooks LC, Fox SJ, et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. Proc Natl Acad Sci U S A 2019 Feb 19;116(8):3146-3154. [doi: 10.1073/pnas.1812594116] [Medline: 30647115]
14. Kostkova P. A roadmap to integrated digital public health surveillance: the vision and the challenges. Presented at: Proceedings of the 22nd International Conference on World Wide Web; May 13-17, 2017 URL: https://dl.acm.org/doi/10.1145/2487788.2488024 [accessed 2026-02-22]
15. Allam Z. The rise of machine intelligence in the COVID-19 pandemic and its impact on health policy. In: Allam Z, editor. Surveying the Covid-19 Pandemic and Its Implications: Urban Health, Data Technology and Political Economy: Elsevier; 2020:89-96. [doi: 10.1016/B978-0-12-824313-8.00006-1]
16. Smolinski MS, Crawley AW, Olsen JM, Jayaraman T, Libel M. Participatory disease surveillance: engaging communities directly in reporting, monitoring, and responding to health threats. JMIR Public Health Surveill 2017 Oct 11;3(4):e62. [doi: 10.2196/publichealth.7540] [Medline: 29021131]
17. Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models: World Health Organization; 2025.
18. Nagarajan R, Kondo M, Salas F, et al. Economics and equity of large language models: health care perspective. J Med Internet Res 2024 Nov 14;26:e64226. [doi: 10.2196/64226] [Medline: 39541580]

Abbreviations

AI: artificial intelligence
API: application programming interface
EHR: electronic health record
ICU: intensive care unit
ILI: influenza-like illness
LLM: large language model
RSV: respiratory syncytial virus
WHO: World Health Organization

Edited by A Coristine; submitted 01.Nov.2025; peer-reviewed by K Adegoke, M Chakit, Z Liu; revised version received 27.Jan.2026; accepted 29.Jan.2026; published 09.Mar.2026.

Please cite as:
Yang L, Shan L, Cao X, Cui J, Tong M, Niu Y, Zhang T
AI Agents and Epidemic Intelligence on Respiratory Infectious Diseases: Toward a Conceptual Framework Integrating Decision Support
J Med Internet Res 2026;28:e86936
URL: https://www.jmir.org/2026/1/e86936
doi:10.2196/86936

© Liuyang Yang, Liyu Shan, Xiaolin Cao, Jinzhao Cui, Michael Tong, Yan Niu, Ting Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Scalable Agile Framework for Execution in AI for Medical AI Ethics Policy Design in Small- and Medium-Sized Enterprises

Ion Nemteanu[1], BS, MS; Adir Mancebo Jr[2], BA, MA, PhD; Leslie Joe[2], MSDS; Ryan Lopez[2], BS; Patricia Lopez[2], ME, MASC; Warren Woodrich Pettine[3,4], MD

[1]Nemsee LLC, San Diego, CA, United States

[2]Data Science Alliance, San Diego, CA, United States

[3]Department of Psychiatry, University of Utah, 501 Chipeta Way, Salt Lake City, UT, United States

[4]Mountain Biometrics, Inc., Salt Lake City, UT, United States

Corresponding Author:
Warren Woodrich Pettine, MD
Department of Psychiatry, University of Utah, 501 Chipeta Way, Salt Lake City, UT, United States

Abstract

Artificial intelligence (AI) is transforming patient care, but it also raises ethical questions, such as bias and transparency. While a range of well-established frameworks exist to guide responsible AI practice, most were designed for academic or regulatory settings and can be hard to operationalize within fast-moving, resource-limited small and medium-sized enterprises (SMEs). We report on the collaborative design of the SAFE-AI (Scalable Agile Framework for Execution in AI), an approach that embeds ethical safeguards, including fairness, transparency, responsibility metrics, and continuous monitoring, directly into standard Agile development cycles. In keeping with established Agile principles, SAFE-AI provides "just enough structure" to integrate ethical oversight into existing workflows without prescribing extensive new governance layers. Similar to other Agile frameworks, such as Scrum, which is described as a "lightweight framework" designed to help teams solve complex problems through iterative learning and minimal process overhead, SAFE-AI aims to remain practical for organizations that may not have dedicated ethics or compliance staff. Rather than simplifying technical methods, SAFE-AI simplifies when and how ethical review is triggered and documented, making responsible AI practices feasible even in environments with limited ethics, governance, or compliance resources. SAFE-AI assumes the presence of qualified data scientists and engineers, and it does not replace the need for statistical or technical expertise but instead provides a lightweight structure for coordinating and documenting work that those experts already perform. We followed a design-science, practice-oriented approach over 20 weeks. After a discovery workshop, a cross-functional team was assembled that included SME employees, ethics researchers, and academic partners. The SME's role was limited to informing design constraints and feasibility considerations during the co-design phase. No operational pilot or production deployment was conducted as part of this study. To reduce the risk of internal design bias and improve generalizability, we also consulted external stakeholders through structured feedback sessions, including clinicians, health care domain experts, and regulatory specialists. Their feedback was incorporated into each prototype-feedback cycle, ensuring that priorities reflected not only the SME's immediate context but also broader clinical and regulatory perspectives. The co-design process produced a 4-phase SAFE-AI life cycle: discovery, assessment, development, and monitoring. SAFE-AI's phase-specific checklists meld acceptance, fairness, and transparency metrics into each Agile sprint. A novel scenario-based probability analogy mapping method was added to translate model risk and uncertainty into plain-language narratives for nontechnical stakeholders, forming the framework's core "responsibility metrics" layer. SAFE-AI is presented as a proposed framework showing that meaningful ethical safeguards can be embedded easily within common workflows used by SMEs that already use basic Agile or iterative development practices. Its checklist-driven phases and automatic review triggers provide a defensible way to track fairness, transparency, and responsibility throughout the model lifecycle.

(J Med Internet Res 2026;28:e80028) doi:10.2196/80028

KEYWORDS

digital health; eHealth; artificial intelligence; machine learning; medical ethics; small business; agile methodology; software development life cycle; risk assessment; fairness metrics; transparency; trustworthy AI

XSL•FO
RenderX

Introduction

Background

Artificial intelligence (AI) systems deployed in health care, whether directly in clinical decision-making or through supporting technologies, carry significant ethical responsibilities. Their outputs can influence diagnoses, treatment options, operational workflows, and ultimately patient well-being. For organizations developing or integrating these tools, including small and medium-sized enterprises (SMEs) that support patient care, ethical considerations, such as fairness, transparency, and accountability, are essential from the outset.

While significant academic work has explored the ethics of deploying AI in medical decision support systems, these insights are often inaccessible to SMEs [1-6]. The complexity of AI ethics, including concepts like fairness, transparency, and accountability, presents a significant barrier, as does the prohibitive cost of dedicated ethics teams. Furthermore, many existing toolkits lack the practical guidance or real-world case studies necessary for SMEs [7,8]. Additionally, embedding ethics into AI development can be transformative for SMEs. Investments in ethical governance contribute to the value of intangible assets, such as intellectual property, loss aversion, brand reputation, and organizational goodwill. In health care markets, where regulatory trust and patient safety are paramount, demonstrable ethical alignment can also serve as a differentiator that attracts ethically aware customers, clinical partners, and investors. As a result, operationalizing AI ethics is not merely a compliance strategy but a value-creating asset for emerging technology companies. Proactive approaches tend to emphasize ways in which investments drive important intangibles that are reflected in a differentiated reputation and robust corporate culture. Mature firms link these intangible outcomes directly to investments in AI ethics and governance, using criteria, such as recognition of industry leadership and support of environmental, social, and governance efforts, as well as the long-term ability to manage risks [9].

In April 2024, HHS OCR (Department of Health and Human Services - Office for Civil Rights) clarified that decision-support tools, including AI, must not produce discriminatory outcomes under §1557 of the Affordable Care Act [10]. Outside software as a medical device (SaMD) pathways, comprehensive federal requirements remain unsettled. The US Food and Drug Administration (FDA) oversight and emerging guidance coexist with evolving proposals [11-13]. In this environment, organizations must anticipate stricter expectations while operating with practical, defensible governance. SAFE-AI (Scalable Agile Framework for Execution in AI) offers a pragmatic path for SMEs to meet today's expectations and prepare for tomorrow's.

To facilitate SMEs in addressing this challenge, we introduce a simplified framework in alignment with established Agile principles for responsibly integrating AI in medical settings. SAFE-AI aims to provide "just enough structure" to guide ethical oversight without imposing extensive new governance requirements, consistent with industry definitions of lightweight frameworks, such as Scrum, which emphasizes minimal process overhead and iterative adaptation [14,15,16]. Using this framework will help ensure SMEs develop medical AI solutions that meet the standard of large hospital systems and prioritize patient care.

Emphasizing AI responsibility, SAFE-AI addresses fairness, transparency, and ethical deployment in a manner proportional to organizational maturity and available expertise. A well-defined process facilitates collaboration between health care practitioners and data scientists, ensuring that ethical considerations, fairness, and equity are systematically assessed and documented. This approach ensures that each phase of AI development aligns with these critical values, leading to more effective and responsible AI tools for health care.

Study Design and Project Context

This work followed a design-science, practice-oriented methodology to develop SAFE-AI, a scalable internal governance framework intended for broad adoption by SMEs building medical-AI products. The project was a collaboration among Medical Timeseries Networks (MTN; an SME), the Data Science Alliance (DSA), Nemsee LLC, and the University of Utah. MTN served as the primary design partner, while DSA, Nemsee, and academic partners contributed ethics, regulatory, and methodological expertise. MTN's role in this study was limited to articulating operational constraints, workflow realities, and feasibility considerations during framework development; no operational deployment or live pilot was conducted as part of this study.

The purpose of the consortium was to embed ethics, regulatory, and methodological expertise into the SAFE-AI framework itself, rather than to provide ongoing expert participation in day-to-day development workflows.

Informal Evidence Gathering

Between December 2024 and April 2025, the design team conducted an ad hoc search of peer-reviewed literature, regulatory guidance (eg, FDA, SaMD, Good Machine-Learning Practice [17], HHS §1557 rule [15], and EU (European Union; AI Act drafts) [18], and industry white papers. Searches were performed opportunistically in PubMed, Scopus, IEEE Xplore, and gray-literature repositories as questions arose during development. All relevant sources were captured in a shared Zotero library and tagged by topic. Key findings were reviewed in weekly design committee meetings.

Framework Outputs

The co-design process yielded:

1. A 4-phase life-cycle diagram (discovery, assessment, development, and monitoring)
2. Stage-specific checklists covering acceptance, fairness, and transparency metrics
3. Practical implementation guidance (eg, role assignments and artifact templates) for each phase.

XSL•FO
RenderX

Consensus and Documentation

All design decisions were made through real-time discussion until unanimous verbal agreement was reached. No formal voting or Delphi procedures were used. Version history, meeting minutes, and rationale for changes were maintained in a locked Google Drive folder that serves as the project's audit trail.

Future Work: Planned Pilot Implementation and Evaluation

The following pilot implementation and evaluation activities are planned for future work and were not conducted as part of the present design study. SAFE-AI will be applied to 2 upcoming MTN initiatives, including a workflow-orchestration product and a data-engineering pipeline. An evaluation plan was prespecified to capture:

1. Process metrics: time (hours) required to complete each SAFE-AI phase.
2. Governance metrics: count and category of fairness metrics defined per project.
3. Usability: postimplementation survey of participating staff (5-item Likert scale; 1=strongly disagree, 5=strongly agree).

Descriptive statistics will summarize these data after the first pilot quarter.

Ethical Considerations

The project involved policy development only; no human participants or identifiable health data were used. Institutional Review Board (IRB) oversight was therefore not required.

SAFE-AI Ethical Evaluation Framework

We present the SAFE-AI, which is a structured process designed to embed ethical evaluation into the technical development life cycle of AI products (summary in Figure 1, full process in Figure 2). A concise checklist distinguishing required and recommended steps is provided in Checklist 1 to support implementation consistency. SAFE-AI enters the workflow after business requirements or system objectives have been defined and prioritized, and tasks are ready for execution by the development team. It complements Agile and Scrum methodologies by introducing structured ethical oversight at critical stages without disrupting project momentum.

Figure 1. SAFE-AI (Scalable Agile Framework for Execution in AI) summary workflow, highlighting each core phase and feedback loop from prioritization to deployment and monitoring. AI: artificial intelligence.

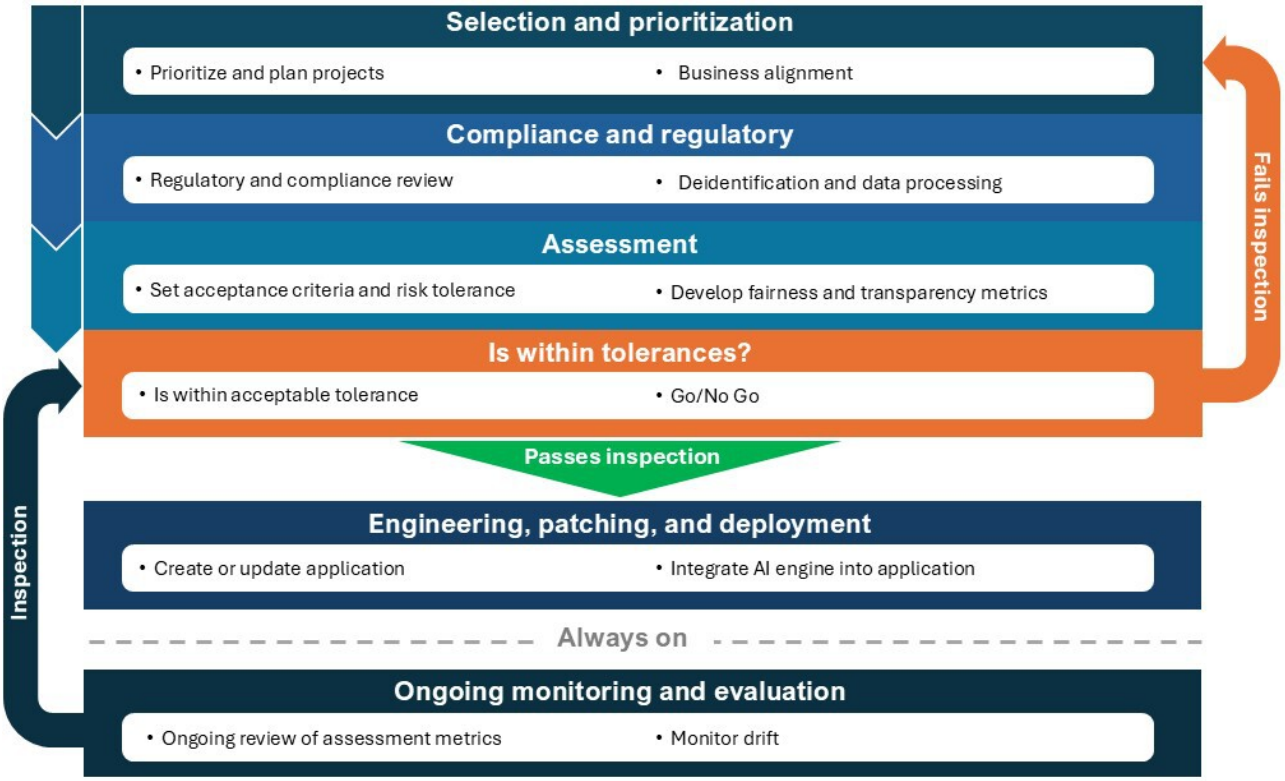

XSL•FO
RenderX

Figure 2. Detailed ethical evaluation process SAFE-AI (Scalable Agile Framework for Execution in AI) - conducted continuously throughout product development as illustrated above. The process begins with a discussion of priorities and alignment and proceeds through model building and inference implementation, emphasizing appropriate tolerance levels at each iteration. AI: artificial intelligence; SDLC: software development life cycle.

XSL•FO
RenderX

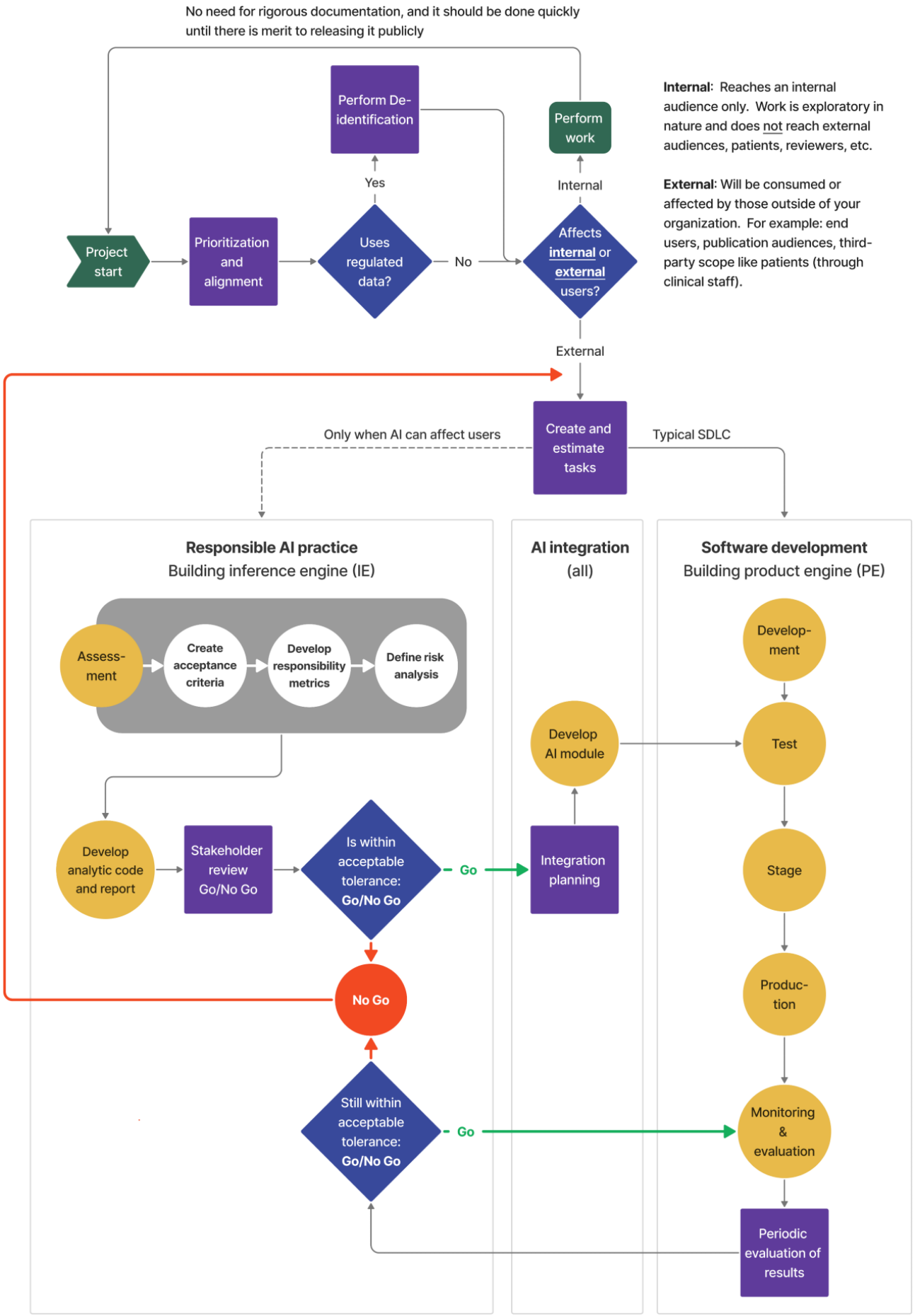

XSL•FO
RenderX

The SAFE-AI framework is not a one-time gate. It is designed to be re-entered at any point in the development lifecycle whenever changes occur that could materially affect the behavior, fairness, or transparency of the AI system. This includes retraining with new data, hyperparameter adjustments, updates to feature selection, or integrating external datasets.

Decisions to reactivate safe-ai should be made jointly by the product owner and technical lead, in consultation with the cross-functional team during natural Agile "sprint review" mechanisms when all the stakeholders are already present. This ensures that ethical checkpoints are triggered not only by technical changes (eg, retraining, parameter tuning) but also by product or stakeholder concerns (eg, shifts in clinical use cases). Embedding this responsibility into sprint reviews and release planning creates a consistent governance mechanism without requiring a separate oversight body.

Practitioners are encouraged to revisit SAFE-AI whenever there are any modifications to their models, regardless of perceived simplicity. Even seemingly minor technical tweaks may cause unintended shifts in model performance or equity. Further recommendations are outlined in the section titled "Quick Note on Model Updates and Considerations," where we suggest re-evaluating the model against all the previously established metrics (refer to Part 2 [Assessment]) to ensure continued ethical integrity.

In this paper, the reference to a "data science team" is used as a generalized term to encompass data scientists and the other skilled roles and internal stakeholders that contribute to the project, including product and program managers, data engineers, analysts, statisticians, software engineers, and development operations personnel. However, when referring specifically to data scientists, we mean those roles that are directly responsible for researching and building the AI algorithms for use in the business and in their products.

SAFE-AI: Part 1 (Discovery)

Determine if the Project Will Use Regulated Data

First, we will determine if the planned work will use or access regulated data. Organizations must comply with all relevant regional, state, federal, and local laws before any access and manipulation of data begins. Protected health information and personally identifiable information are among the many regulated data types. Existing acts, such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) set specific compliance requirements.

Best practices include anonymizing or aggregating data to protect individual privacy. Additionally, informed consent is required for legally effective participation in research and must be obtained in accordance with the principle of respect for persons [19]. This principle ensures that individuals are given the opportunity to choose whether they wish to participate in research studies.

A comprehensive regulatory review should be conducted before any AI development begins. If specific handling techniques require documentation, they should be securely stored and referenced throughout development. When designing systems that impact external audiences, regulatory requirements should be listed as acceptance criteria and revisited at every development stage. If the same data are used repeatedly, the regulatory review can cover repeated use and should have a documented process (if necessary) to ensure compliance for future access and proper use. This will ensure that resources do not have to review the data every time it is accessed.

Determine if Scope of Work Impacts Internal or External Users

iDetermine External Effect

During typical prioritization meetings, it should be discussed whether the work will affect external stakeholders like patients, clinicians, or other health care providers. This is an important distinction since there are many tasks that should be done quickly and match the pace of the organization and do not require the extra burden of documentation. Realistically, not every task requires burdensome processes and documentation. This framework aims to only target the tasks that have a direct effect on the health care outcomes as a direct result of the product being developed. The first step in this process is a quick discussion during the planning meeting that determines the upcoming scope of work for the development team and clarifies if outputs can affect external stakeholders. Clarifications are detailed below.

Internal Stakeholders

Project work that includes exploratory research, feasibility studies, or internal management tools that do not directly impact patient outcomes can proceed without additional scrutiny. The process involves moving forward with the task as normal, without the need for additional scrutiny. Upon completion, the process is restarted from the beginning.

External Stakeholders

Work that affects those outside of your organization. For example, patients, clinicians, or other health care providers; end users, and publication audiences. In this case, the outcome of these tasks can or will have a direct impact on an external audience. Internal projects may also carry external implications. For instance, human resource initiatives that use protected employee data involve internal stakeholders, but mishandling that data can create risks that extend beyond the organization. The goal is to recognize when the project scope introduces potential external risk, even if it originates internally. In this case, the process involves moving on to the next step.

Create and Estimate Tasks

This phase begins with breaking down high-level project goals into smaller, manageable user stories or tasks, as seen in Agile Software Development Processes. These tasks often span data ingestion, cleaning, feature engineering, model development, evaluation, and integration. During planning, the data science team estimates each task's duration based on the scope of work for iterative delivery while leaving room for learning, model iteration, and integration into production workflows.

To ensure that ethical and social considerations are not deferred until the end of development, or left out entirely, this framework incorporates structured ethical reasoning into the task creation

XSL•FO
RenderX

process from the start. The foundation for this integration draws from the RESOLVEDD process (Review, Estimate, Solutions, Outcomes, Likely impact, Values, Evaluate, Decide, and Defend), a structured ethical decision-making framework originally introduced by Pfeiffer and Forsberg [20,21]. RESOLVEDD guides practitioners through steps, such as reviewing the facts, estimating the ethical conflict, surveying alternatives, opening options to stakeholders, listing risks, valuing impacts, evaluating options, deciding, and defending the decision. While originally applied to workplace ethics, the logic of RESOLVEDD has broad applicability in responsible AI contexts.

Building on this foundation, Vakkuri et al. [22] introduced ECCOLA, a card-based toolkit tailored to Agile AI and autonomous systems development. ECCOLA does not follow an acronym, but its sprint-by-sprint format maps well to the RESOLVEDD process. It consists of a deck of 21 cards aligned to 8 core AI ethics themes (eg, transparency, accountability, and bias), each designed to provoke discussion, surface risks, and frame decisions during active development. The cards can be used in planning meetings to ensure that ethical issues are not only acknowledged but also actively scoped into the project as discrete tasks.

SAFE-AI adapts this card logic but moves one step further: rather than selecting from a predefined deck, teams create their own custom cards directly from the objectives and risks identified in the project. Each card captures the problem, the related ethical concern, and the concrete analysis tasks that will address it. These cards are then written into the backlog as user stories or research spikes, estimated during sprint planning, and assigned clear ownership. By embedding them directly into the backlog, ethical concerns are tracked with the same visibility and accountability as technical work. For example, a card generated from the objective “Minimize Gender Bias” might be translated into a user story with acceptance criteria tied to distributional checks, fairness metrics, and subgroup error analyses.

This approach converts abstract ethical principles into concrete, time-boxed actions that can be tracked and revisited throughout the sprint cycle. It ensures that responsibility is not just a post hoc evaluation, but a proactive part of scoping and estimating work. By embedding the card principle in this way, SAFE-AI enables collaborative assignment of each card and its tasks, supports ongoing discussion between stakeholders, and provides a shared method for everyone to follow. A simple template for these cards anchors the ethical problem with other “User Stories,” with the analysis tasks listed below (Table 1).

Table . SAFE-AI[c] ethical card template for planning and backlog integration (eg, Minimize Gender Bias[a,b]).

What to do	Output
Training data balance. Quantify overall and conditional distributions by gender (and salient covariates)	• Distribution tables and joint density plots
Split integrity. Verify gender balance across training, validation, and testing (or rolling production batches).	• Per-split distribution summaries
Performance by subgroup: Report declared metrics by gender (eg, AUROC[d], F_1-score, PPV[e], or NPV[f] calibration) and compute gaps and confidence intervals.	• Per-subgroup metric table • Gap estimates • Calibration plots
Downstream effect at acceptance testing: Compare consequential outcomes by gender (eg, referral or redirect rates) under final acceptance criteria.	• Outcome distributions with statistical tests and effect sizes

[a]Do all genders have similar performance and minimized bias?

[b]Problem: The model may yield systematically different performance and error profiles across genders, a protected characteristic in health care decision support.

[c]SAFE-AI: Scalable Agile Framework for Execution in AI.

[d]AU: area under the receiver operating characteristic.

[e]PPV: positive predictive value.

[f]NPV: negative predictive value.

These tasks are customized per project and sprint. They should be time-boxed and assigned within the sprint backlog, ensuring fairness concerns are scoped into the same process as technical work.

Integrating Into Agile Software Development Life Cycle

In practice, these SAFE-AI cards should be treated in the same way as user stories or research spikes within Agile planning. During sprint planning, teams can pull the relevant cards into the backlog and frame them as discrete work items with acceptance criteria, effort estimates, and clear ownership. This ensures that ethical questions are not abstract side discussions but concrete backlog items tracked alongside feature development. For example, a card highlighting fairness concerns may become a “bias audit story,” or a transparency card may generate a “model interpretability spike.” By embedding these cards directly into sprint discussions and backlog management, responsibility metrics are scoped, prioritized, and delivered with the same visibility as any other technical deliverable without introducing new governance layers or slowing iterative delivery.

Importantly, integrating SAFE-AI into Agile requires only minimal additional ceremony. Teams typically spend about 50 minutes during preplanning to surface ethical risks and create the initial cards and approximately 10 minutes during

XSL•FO
RenderX

sprint-planning ceremonies to refine and convert those cards into user stories or spikes (Figure 3). These brief additions occur within meetings that already exist, meaning the procedural overhead is minimal. The more substantial time investment (generally up to 6 hours within the sprint) is not overhead created by SAFE-AI itself but rather the time spent actually performing the ethical tasks, such as fairness checks, interpretability analyses, drift reviews, or documentation improvements. In many teams, these activities either occurred informally, inconsistently, or were never undertaken at all. SAFE-AI simply makes this work visible, trackable, and accountable by assigning it a discrete story with an owner and acceptance criteria.

Figure 3. SAFE-AI (Scalable Agile Framework for Execution in AI) card workflow integrated into the Agile software development life cycle, including an estimate of the additional time required to incorporate SAFE-AI activities into standard Agile ceremonies. In preplanning, project objectives are defined, and ethical risks are identified, leading to the creation of custom cards that specify the problem, analysis tasks, and outputs. During sprint planning, these cards are converted into backlog items, such as user stories or research spikes, estimated, assigned, and given acceptance criteria. In the sprint phase, the ethical tasks are executed and validated alongside technical stories, ensuring responsibility metrics are tracked within the same Agile workflow with about 1-7 hours of additional time per sprint depending on the needs.

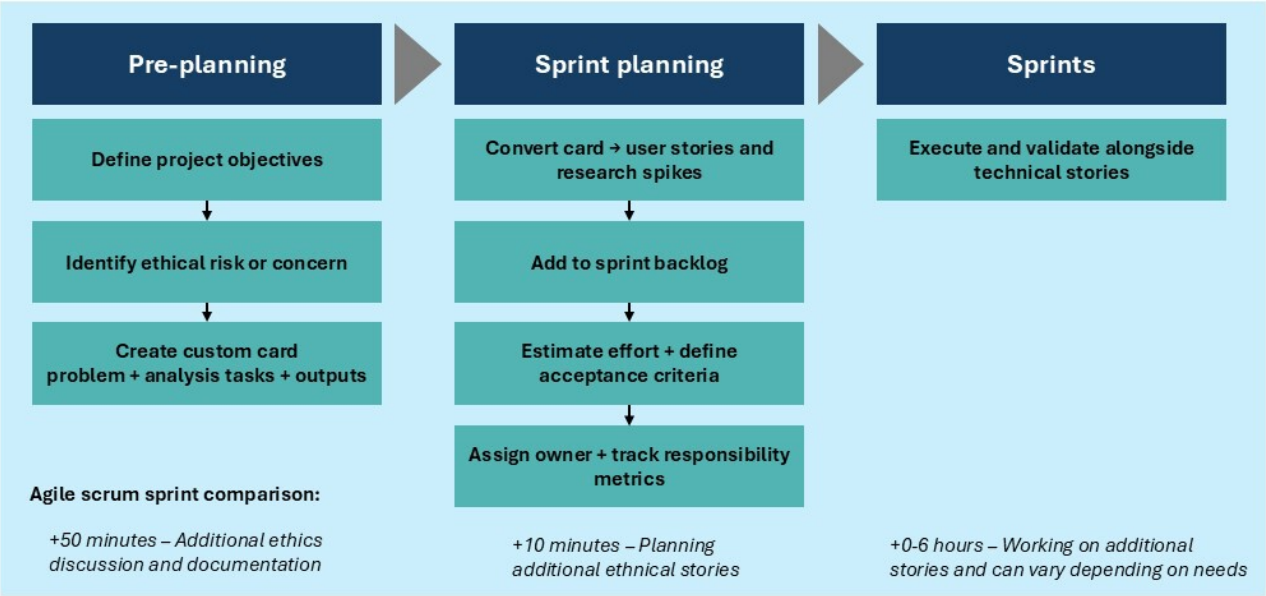

SAFE-AI cards progress from the identification of ethical risks in preplanning to conversion into backlog items during sprint planning and finally to execution and validation alongside technical stories (Figure 3). This demonstrates that SAFE-AI adds only a small amount of process time while ensuring that ethical responsibilities are now explicitly planned, resourced, and delivered with the same rigor as technical features.

SAFE-AI: Part 2 (Assessment)

Overview

This section occurs during the inference engine (IE) phase and outlines key steps the data science team, both technical and product stakeholders, should complete before progressing to development and integration. These steps ensure that AI IEs operate within defined boundaries of uncertainty and acceptable risk, aligning with ethical standards and strategic business goals.

It is important that the following steps be incorporated alongside traditional software testing as a routine part of development practice, with consistent attention to their execution. While these steps may initially appear to add time or complexity—requiring deeper evaluation, more test cycles, and added stakeholder input—when planned appropriately, these activities do not disrupt schedules. They are instead embedded naturally as part of the process and become part of the schedule.

The SAFE-AI framework is designed to be iterative and collaborative, ensuring that ethical assessments are embedded incrementally and transparently throughout the development lifecycle. Crucially, tasks are not imposed in isolation, as they are co-developed and agreed upon by all stakeholders, including data scientists, engineers, product leads, and domain experts, which keeps timelines aligned and priorities clear.

Ethical testing is treated with the same rigor and visibility as functional testing, making it easier to forecast and manage. While often overlooked due to perceived resource constraints or competitive pressure, these steps are essential to building systems that are robust, compliant, and trustworthy. The framework's goal is to define the minimum necessary ethical safeguards required to ensure responsible AI while preserving agility and long-term strategic value.

There are 3 tasks that are especially relevant during the assessment phase, each of which is detailed below

Creating Acceptance Criteria

Acceptance criteria establish the conditions a system must meet before release. Borrowed from traditional software development life cycle (SDLC), they provide clear, testable expectations. In AI systems, this still begins with baseline functionality "Does the model work as intended?" For example, a clinical diagnostic tool might require ≥95% accuracy overall before being considered ready for integration into workflows. Such criteria

XSL•FO
RenderX

help establish shared expectations and support regulatory alignment before AI models are integrated into clinical workflows. Atlassian describes acceptance criteria as "clear, concise, and testable statements that focus on providing positive customer results" [23,24].

In AI systems, especially those used in high-stakes domains like health care, defining acceptance criteria is not just a technical step—it is a normative one. The choice of what counts as "acceptable" performance can have real-world implications for safety, equity, and trust. It is therefore essential that these baseline criteria be co-developed with key stakeholders, including clinicians, patients, domain experts, and end users. AI engineers may assume they understand what matters most to users, but this assumption often overlooks critical perspectives—particularly around risk tolerance, fairness, and real-world utility. Engaging stakeholders in defining success conditions helps ensure that the system reflects the values and priorities of those it is intended to serve.

Acceptance criteria are only the first step. Once minimum functionality is defined, additional layers of responsibility metrics are needed to answer two further questions: "Does it work fairly, and is it understandable to those making directions on behalf of the company and those it serves?" Fairness and transparency extend acceptance criteria into the ethical domain, ensuring that the system not only functions but also functions responsibly.

Develop Responsibility Metrics

The Importance of Responsibility Metrics

Once baseline acceptance criteria confirm that a model works, the next step is to evaluate how it will work in practice. Responsibility metrics extend beyond functionality to address whether the system is fair, trustworthy, and aligned with stakeholder values. These metrics help organizations detect bias, communicate uncertainty, and ensure that model outputs can be understood and acted upon responsibly. Together, they provide the ethical guardrails that acceptance criteria alone cannot supply.

Fairness Metrics

Bias and equitable performance across groups within AI systems will be defined and measured. As BSA (Business Software Alliance [25]) states, fairness is subjective and requires selecting the most appropriate metrics for the system being developed. These metrics should be documented and reviewed collaboratively by subject matter experts and data scientists [25]. In practice, fairness metrics are used to detect and mitigate disparate impacts, ensuring that one group or class (eg, based on race, gender, age, or geography) is not disproportionately harmed or excluded compared to others. This may involve evaluating performance across subgroups, measuring disparities in false positives or negatives, or assessing equal access to beneficial outcomes. The goal is not only to identify statistical imbalance but also to actively choose metrics that reflect the ethical priorities and risk tolerances of the system's intended users and affected populations [26-30]

Transparency Metrics

Metrics will be defined that focus on ensuring that AI model predictions are interpretable, explainable, and actionable by end users. One technique used in this framework is a scenario-based probability analogy mapping method (SPAMM)—a structured approach to communicating AI predictions by framing them as outcome narratives grounded in real-world context and relevant to the business goals. This method, which we refer to as the SPAMM, enhances transparency by mapping probabilities of outcomes related to fairness metrics and other acceptance and risk criteria to familiar analogies that illustrate the likelihood of different scenarios, the conditions under which errors occur, and their potential consequences. Rather than focusing solely on model internals, SPAMM presents predictions in narrative form - conveying not only what the model predicts, but why, how alternative outcomes compare, and which inputs most influence the result (Figure 4). This approach builds stakeholder trust by making uncertainty and risk both visible and relatable.

Figure 4. Example of scenario-based probability analogy mapping method compared to technical metrics. AI: artificial intelligence; SHAP: Shapley Additive Explanations; SPAMM: scenario-based probability analogy mapping method.

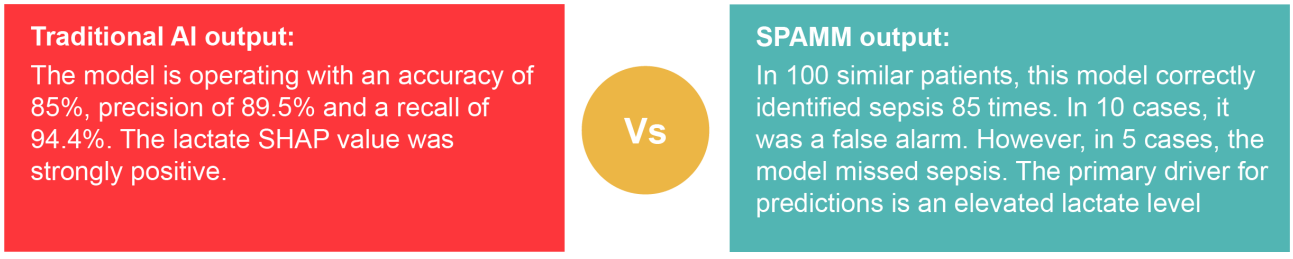

This approach is particularly important in high-stakes domains like health care, where decision-makers, such as clinicians, patients, or administrators must weigh the risks of false positives and false negatives. By highlighting the probability of misclassification and mapping those scenarios to familiar contexts, stakeholders can assess how often and under what circumstances the model may be wrong - and what the real-world consequences of those errors might be. For example, if a model is correct 85% of the time but consistently fails on a particular patient subgroup, this insight can guide mitigation strategies, prompt workflow adjustments, or justify additional human oversight. These dimensions of model behavior—error distribution, uncertainty, and impact—are often absent in standard AI evaluation pipelines.

Traditional interpretability tools like SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), partial dependence plots (PDP), and accumulated local effects (ALE) are valuable for understanding model

XSL•FO
RenderX

behavior. SHAP, which is based on cooperative game theory, assigns each feature an importance value representing its contribution to a particular prediction—similar to calculating how much each player contributed to a team's win. LIME works by perturbing the input data near a given instance and fitting a simple model to explain the local decision boundary. PDP and ALE methods help assess how the average model output changes when a single feature is varied, which supports global interpretability, especially when exploring relationships between inputs and outcomes.

These methods help data scientists and auditors understand what the model is doing, but they may not fully address the kinds of interpretive needs that arise when communicating overall impact with clinical or business stakeholders [31-35]. Technical performance metrics that evaluate true and positive rates, such as root mean square error, precision, and/or recall, can be difficult to translate to nontechnical audiences. As Ribeiro et al [36] noted, trust in machine learning (ML) models depends on the user's ability to anticipate and understand model behavior in practical terms— not just accuracy, but reliability and risk.

As a result, we apply SPAMM, which adds a critical layer to the transparency toolkit by reframing predictive outputs in terms that mirror how people naturally reason under uncertainty. At its core, the method walks users through a model's output by anchoring it to the specific input conditions that led to that prediction and by showing how often those conditions have historically produced similar outcomes.

Presenting outputs in this context helps stakeholders understand not just what the model predicted, but why it arrived at that decision, how likely alternative outcomes were, and how much confidence they can place in it. While this can become complex in models with high-dimensional input spaces or large numbers of interacting variables, the point is not to replicate the entire model logic, but to provide a real-world testing lens that frames predictions through the lens of practical "what-if" scenarios. This helps surface critical edge cases, increase interpretability, and build trust, especially in settings where human oversight and risk-based decisions are essential [37-39].

Crucially, this framing also enables teams to assign tangible costs to errors by explicitly asking "What can happen if this prediction is wrong?" and "What is the risk impact to the organization or patient?" For example, a false negative in a sepsis prediction model may delay critical treatment, leading to clinical harm, regulatory exposure, and reputational damage. A false positive in an insurance fraud detector may result in a wrongful denial of benefits, eroding user trust and increasing legal liability. By mapping these risks to predicted outcomes and associated variables, organizations can better understand which errors matter most, estimate their financial or operational consequences, and prioritize safeguards accordingly. Integrating these cost-of-error considerations into the transparency process transforms interpretability from a purely technical exercise into a tool for risk mitigation and strategic alignment [40,41].

Ultimately, this approach supports a more robust, stakeholder-centered view of transparency. It enables users to understand reliability, anticipate error, and make informed decisions based on modeled outcomes—an essential capability for ethical and accountable AI deployment in health care and other sensitive domains (Table 2).

Table . Illustrative acceptance criteria and responsibility metrics across common project types, including enterprise automation, consumer-facing artificial intelligence, and clinical decision support.

Project type	Acceptance criteria	Responsibility metrics
Enterprise automation	• API[a] response time ≤200 ms • Workflow accuracy ≥98% across standard inputs	• Fairness: Ensure no subgroup of users is excluded from automation • Transparency: Apply SPAMM[b] to communicate error risks, automation actions, and results, and maintain audit logs
Consumer-facing AI	• Recommendation relevance ≥35% via click-through, ≥85 satisfaction • Subgroup accuracy within ±5%, ≥10% Interaction with the next action	• Fairness: Balanced recommendation exposure across demographic groups • Transparency: Apply SPAMM to explain recommendations, actions, and results, and expose key features
Clinical decision support	• Diagnostic accuracy ≥95% overall • Subgroup disparity ≤5% across age, race, and gender	• Fairness: Subgroup error analysis, including false positives and negatives • Transparency: Apply SPAMM to frame predictions and risks in a clinical context

[a]API (application programming interface):

[b]SPAMM (scenario-based probability analogy mapping method):

This framework invites further development in areas such as visualization, user-centered metrics, and storytelling techniques for AI transparency. While this paper does not formalize the full methodology, it highlights the importance of designing explanation strategies that reflect how real-world users evaluate uncertainty, reliability, and relevance—not just technical performance.

In addition to defining fairness and transparency requirements, teams should specify the statistical methods that will be used later in monitoring to verify whether these metrics hold in

XSL•FO
RenderX

production. For example, subgroup outcome disparities can be tracked using chi-square or Fisher exact tests, distribution shifts can be detected with Kolmogorov–Smirnov or Jensen–Shannon divergence, and overall population stability can be assessed with Population Stability Index (PSI). By embedding these statistical guardrails directly into the acceptance criteria, SAFE-AI ensures that monitoring processes in Part 4 are both preplanned and directly linked to the responsibility metrics established during development.

Define Risk Analysis

Risk analysis should be conducted collaboratively, including consultation with relevant stakeholders like product teams, data scientists, and legal experts. There are many possible approaches, and the following are a few examples of structured risk analysis processes that already exist in the field of AI and health care.

1. The NIST (National Institute of Standards and Technology) AI Risk Management Framework provides structured guidelines for evaluating risks in AI systems, emphasizing trustworthiness and resilience through iterative assessments [42,43].
2. McKinsey's "Derisking AI by Design" approach highlights the necessity of integrating risk management throughout the AI lifecycle, focusing on proactive monitoring and governance [44].
3. Stanford's AI health care risk assessment framework outlines liability risks in medical AI applications, suggesting adaptive monitoring strategies to address unforeseen ethical and operational challenges [45].
4. The US Food and Drug Administration's (FDA) Total Product Life Cycle (TPLC) framework ensures continuous oversight from AI model development through postmarket surveillance. The FDA writes how generative artificial intelligence (GenAI) presents potential risks that may require varying levels of risk controls for different applications, as is true of other technologies. For example, hallucinations, particularly those that may appear to be authentic outputs to users, may present a significant challenge in certain health care applications where highly accurate, truthful information is critical. They state how current evaluation approaches, such as those used to evaluate computer-assisted triage, detection, and diagnostic devices, may still be applicable for GenAI-enabled devices, albeit with additional supporting evidence. However, it may be challenging to determine the evidence that may be needed for certain GenAI implementations [46,47].

All of the above approaches include hazard identification, risk assessment, and mitigation strategies tailored to AI's dynamic nature in health care. By incorporating structured risk evaluation, teams can proactively address potential harms and regulatory challenges. Key considerations include:

1. Misuse potential: Could the product be used incorrectly, leading to harm?
2. Error costs: What are the consequences of false predictions?
3. Societal implications: Are there broader ethical concerns?

Legal experts should be engaged to ensure compliance with evolving regulatory landscapes and provide specific tests and metrics that can be used to create risk boundaries for AI models that can help pass regulatory reviews as well as build trust with the public.

Quick Note on Model Updates and Considerations

Once an IE has been approved and integrated into a production environment, it is common practice to revisit the model for improvements through techniques, such as retraining with new data, hyperparameter tuning, or feature modifications. While these changes may be perceived as incremental or routine, they carry the potential to significantly alter the algorithm's behavior, calibration, and fairness properties. Even minor updates can reintroduce latent biases, shift subgroup performance, or affect the interpretability and reliability of outputs.

From an ethical standpoint, any modification to the IE, including changes to a single model or updates to the system of models, should be treated as a new model iteration when following the SAFE-AI framework, regardless of whether the foundational architecture remains unchanged. Although it is not necessary to recreate the entire ethical evaluation pipeline, it is considered best practice to reapply the original assessment metrics established during initial development (outlined in Part 2 [Assessment]). This allows teams to validate that the updated model continues to meet ethical expectations for fairness, transparency, and performance consistency.

In high-stakes domains, such as health care, the tension between deployment speed and responsible oversight is a persistent challenge. Rapid iteration is often essential for competitiveness; however, this urgency must be weighed against the risk of unintended harm. Ethical oversight should not be perceived as a barrier but rather as a safeguard that ensures continuity of ethical alignment across the model's lifecycle.

Therefore, organizations are strongly encouraged to adopt a policy of proactive reassessment following any material change to an AI system. By incorporating these evaluations into their standard operating procedures, teams can preserve ethical integrity, mitigate risks, and maintain regulatory and stakeholder confidence.

SAFE-AI: Part 3 (Development and Integration)

Transition to Development

After defining metrics and performing the necessary reviews, the next step is to move from validated concepts into functional systems. This phase encompasses 3 key stages already defined in Part 2: the IE, the product engine (PE), and AI integration.

1. IE: As described earlier, the IE represents the culmination of algorithm development, transforming exploratory research into validated, production-ready code. Activities include hypothesis testing, feature engineering, model training, and validation, until the algorithm meets defined SAFE-AI thresholds for accuracy, risk, and subgroup reliability.
2. PE: In parallel, product engineers build the functional application layer. Here, validated models are embedded into the product architecture so they can import data,

XSL•FO
RenderX

process it through the IE, and output results for downstream systems and users. Inputs, outputs, and thresholds are explicitly defined and jointly agreed upon. This work ensures the product is both technically robust and usable for stakeholders.

3. AI integration: Integration is the convergence point where the IE and PE are combined into a single, deployable system. At this stage, focus shifts from feasibility to operational viability, ensuring compatibility with data pipelines, embedding monitoring systems, and preparing the organization for deployment (training, regulatory compliance, user support, and related functions).

Together, these stages operationalize the SAFE-AI framework by moving from validated research into accountable, production-ready systems that can be deployed and monitored in real-world settings.

Develop Analytic Code and Report

As part of the IE phase, data scientists transform exploratory research into validated inference code. This includes feature engineering, data exploration, model training and tuning, and refining the algorithmic process until it meets the established acceptance criteria and responsibility metrics. The key output of this phase is a report, which serves as the deliverable for review in the assessment and go or no-go decision.

This report functions as a consensus-building document: it (1) demonstrates how the model performs against the minimum functional standards defined in the acceptance criteria; (2) presents the results of fairness and transparency evaluations under the responsibility metrics, including subgroup performance, parity checks, error distributions, and interpretability narratives, such as SPAMM; (3) assesses overall risks and costs with implications for both users and the organization; (4) establishes a baseline of monitoring metrics and visualizations for long-term oversight; and (5) ultimately facilitates an informed go or no-go decision among stakeholders.

Stakeholder Review Go or No-Go

After completing the above, the decision-making stakeholders should come together as part of an existing review or showcase discussion to review the report, analyze the results, and come to a consensus on whether or not the work is “good enough” to proceed to deployment. If the group determines that further refinement is needed, then the team updates the relevant metrics, documents the discussion, and begins another iteration of the assessment process within the same Agile cycle. Importantly, this review is not a new governance requirement being added to Agile. Most teams already make informal go or no-go decisions in sprint reviews, demos, or ad hoc leadership discussions. SAFE-AI simply formalizes and documents these joint decisions, so the rationale is transparent and traceable. It is important to include organizational leadership with the authority to make the executive decision in this regard, including people with ownership of balancing the risks of the algorithm against the needs of the organization.

AI developers and AI deployers should maintain a governance framework that is backed by sufficient executive oversight. In addition to developing and approving the substance of the governance framework’s policies, senior management should play an active role in overseeing the company’s AI product development lifecycle. For high-risk systems that may negatively impact people in consequential ways, company leadership should be accountable for making “go or no-go” decisions [25]. SAFE-AI does not add a separate approval gate, but rather aligns the accountability expected in high-risk AI with the existing cadence of Agile decision points, making explicit what is often handled implicitly.

Integration Planning

Now that the inferential process has been completed and all stakeholders are aligned on the algorithm, the data scientist can work with the software engineers to lock down the deployment constraints. In some cases, this can be done before the “go or no-go” decision if inputs and outputs will not change significantly. It is important to have these completed before proceeding, although in practice, many teams already discuss these topics informally earlier in the development cycle. SAFE-AI does not introduce new tasks here, but simply formalizes conversations that often happen ad hoc, providing clearer documentation and coordination. It is good practice to perform integration planning as early as possible. However, performing this too early can create conflicts when the inputs and outputs change during the research phases. It’s best to perform this planning phase when there is good certainty that the inputs and outputs are well defined and not likely to change. By treating these steps as structured, documented alignment points rather than informal hallway conversations, teams avoid misunderstandings and reduce rework without adding meaningful overhead. For example:

1. Input or output contract: The data engineering process will need to know what data to create to feed into the algorithm and how often, while also asking for the algorithm outputs in return. This contract can be version controlled to say, “The system will always give you this data in this format, and your IE will always give certain data in a certain format in return.” This makes updating the algorithm in the future easier, which will be discussed later.
2. Update process: This should be a plan for which version of the production will integrate the first model. It is typical that an algorithm’s code may change more often than the production version and should have a process to deploy updates to the algorithm quickly without waiting for a full product release.

Develop an AI Module

Now it is time to integrate the models as a new feature within the home application’s development environment that is now aligned with a planned production release. At this point, the work follows typical SDLC processes where the algorithms are engineered into the overall product architecture and involves performing all the other typical processes involved with a new feature release, such as marketing, educating stakeholders, etc

Integration

Once the code is ready, developers move it into production, following the standard SDLC. It is essential to recognize that ML models and the broader application often have different

XSL•FO
RenderX

deployment schedules, requiring a well-orchestrated Continuous Integration and Continuous Delivery process as illustrated in Textbox 1.

Textbox 1. Defining distinct task categories that are relevant to the SAFE-AI (Scalable Agile Framework for Execution in AI) process.

Developing product engine (PE): Traditional software development where the product architecture is built. These tasks are deterministic and repeatable, such as constructing databases, application programming interfaces, or user interfaces. Outcomes are binary (eg, schema exists or not; field matches or does not), and successful runs always yield identical results. The PE provides the stable scaffolding into which artificial intelligence (AI) components can later be embedded.

Developing an inference engine (IE): Unlike deterministic product code, IEs rely on models that incorporate randomness, sampling, or continuous learning, meaning the same code can yield different outputs across runs. This makes validation more complex, where predictions must be evaluated across multiple runs, data splits, and subgroups to ensure they fall within acceptable thresholds rather than simple pass or fail testing. As such, running the same code twice can produce different results. IE development spans from exploratory research (hypotheses, tuning cycles, and modeling notebooks) to validated inference code ready for integration.

Organizations often manage two parallel codebases during this phase: (1) research code is maintained by data scientists and (2) validated prediction engine code is used within the product engine. Only when the team confirms that the engine meets uncertainty and risk criteria does it become eligible for integration.

AI Integration (AII): The convergence point where prediction engine code is embedded into the product codebase. AI transforms research into an operational component that accepts input, runs the algorithm, and returns predictions through the product interface.

By the time the code reaches this phase, the model has been fully validated and is no longer being evaluated for feasibility—just functionality. The code is versioned independently and integrated in coordination with product releases to ensure seamless deployment.

To illustrate these differences, Figure 5 demonstrates how these phases operate in practice.

As shown, the IE and PE evolve on separate but parallel development tracks, each with its own branching and deployment cadence. For example, Live IE version 1.1 may remain in production while a newer Dev IE v1.2 is being validated, just as Dev PE v2.2 advances independently of earlier releases. The AII phase serves as the coordination point where validated AI models are embedded into product-ready systems, and inputs and outputs are aligned and locked. This moment (marked by the star icon in Figure 5) represents the culmination of both tracks, where a stable IE is paired with its corresponding product version for release. Once deployed, the system enters a new live state (eg, Live PE v2.2 with Live IE v1.2), ensuring that only approved, tested combinations reach end users. This disciplined integration workflow supports transparency, traceability, and accountability across the AI development lifecycle.

It is important to note that changes to the IE don't always involve modifications to input or output formats. The structure of the data pipeline may remain consistent even as the underlying model evolves through algorithm updates, parameter tuning, or other enhancements. The purpose of the diagram is simply to illustrate that, at some point, the IE and PE must converge during the AI Integration phase. However, since IE updates can occur independently of PE deployments, it's critical that any such changes, regardless of how minimal they may appear, still undergo the SAFE-AI evaluation process. This ensures that even behind-the-scenes model improvements are held to the same standards of validation, risk assessment, and accountability.

Figure 5. Example of product engine and inference engine version control with parallel development cycles. AI: artificial intelligence; PE: product engine.

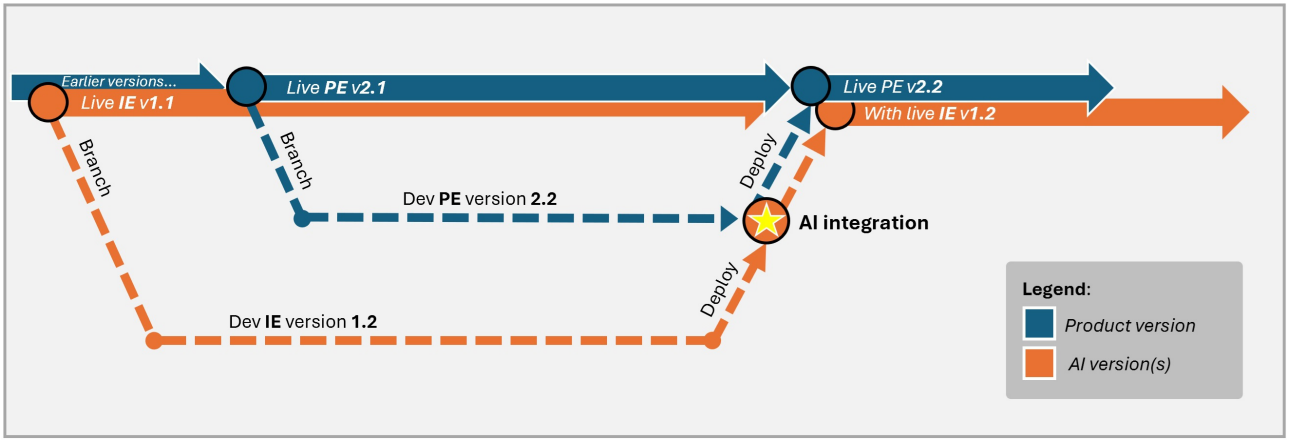

For example, ML models may need to be retrained and updated more frequently than full product releases. This raises important considerations, such as how to introduce new parameters or model updates without disrupting the existing environment.

Ensuring a smooth deployment process requires clear coordination between the data science and software development teams. These discussions should focus on versioning strategies, model refresh cycles, and seamless integration workflows to maintain system stability while enabling continuous improvements (Textbox 2).

XSL•FO
RenderX

Textbox 2. Example illustration of SPAMM (Scenario-based Probability Analogy Mapping Method) in practice

To illustrate SPAMM, the project is assumed to be a system that helps to triage emergency room visits for nonurgent patients only (not to include resuscitation or emergent patients) based on incoming patient data and demographics to try and reduce patient wait times and complications caused by long waits and overcrowding. The goal is to translate your inference engine's preferred error metrics, such as recall, accuracy, root mean square error, and other technical metrics, into the real actions and events that will happen as a result of the predictions.

This is a demonstration example only. It illustrates how the SPAMM method can be applied, but it is not a real clinical analysis.

Step 1 (complete assessment steps): Before building scenarios, all relevant metrics and goals found in the acceptance criteria, responsibility metrics, and risk analysis steps in SAFE-AI (Scalable Agile Framework for Execution in AI) should be defined and documented. That will ensure the organization has a shared baseline of "what matters," how to measure it, and the cost of getting it wrong.

Step 2 (translate acceptance criteria into scenarios): For a relevant metric, these should be associated with the simulated outcomes from the inference engine that is translated into how those predictions relate to the real-world actions and consequences of the model.

Overall Scenarios

- The Emergency Department receives about 152 "nonurgent" patients a day, with 4 out of 5 of those patients experiencing no change in how they would normally be treated.
- If all of the prescribed actions are implemented:
- About 15% (23/152) of these patients would end up waiting longer than they normally would, since the system would reprioritize other cases ahead of them.
- About 5% (8/152) of these patients would be classified as higher urgency than they normally would, meaning they would be escalated for faster treatment than under standard triage.
- Of the 15% (23/152) of patients who would experience longer wait times, only about 1 in 1000 would experience an adverse effect due to the delay (approximately 9 cases per year).
- For the remaining ~23 patients per day who waited longer but did not experience harm, their delayed care effectively freed capacity for more urgent patients. This reallocation reduced other ER-related adverse events by an estimated 5% (around 25 cases per year).
- **Demographic Scenarios** • Age: For every 100 older patients, about 8 are predicted incorrectly. These patients have a 50% higher chance of harm and accounted for nearly all of the adverse events caused by longer wait times. • Race: For every 100 Black patients, about 6 are predicted incorrectly, with about a 5% higher chance of harm. • Gender: There are no meaningful differences in error rates by gender.

Conclusion

- Most patients (121/152, 80%) are triaged the same way they would be under the current system.
- Roughly 8 patients per day are upgraded, meaning they receive care faster than before.
- About 24 patients per day are downgraded, which creates some additional risk, but only around 1 in 1000 of these cases leads to actual harm (about 9 cases per year).
- By comparison, the reallocation of resources reduces other ER-related harms by an estimated 25 cases per year, suggesting an overall safety benefit.
- The most significant concern is that the small number of adverse events disproportionately affects older patients, and Black patients show slightly higher error rates.
- These fairness issues, together with the rare but high-impact missed urgent cases, represent the biggest risks to the hospital's reputation and liability. Smaller risks, such as false alarms or mild misclassifications, carry lower per-case costs but can still influence efficiency.

Overall Demonstration Conclusion: This example illustrates how a triage support system could reduce wait times and improve patient flow while also exposing the disproportionate costs of rare but severe errors and subgroup disparities. From a business perspective, the primary risks are tied to patient harm, liability exposure, and reputational impact. The exercise is not a real analysis but a demonstration of how SPAMM can be applied to create scenario-based narratives that connect acceptance criteria, responsibility metrics, and risk analysis into a single story for decision-makers.

SAFE-AI: Part IV- Monitoring and Evaluation

System monitoring and evaluation at regular intervals can help ensure the prevention of potential harm creation or perpetuation. While initial model results may meet responsibility metrics during development, changes in data patterns or underlying system characteristics could cause deviations in production performance.

According to Data Science Alliance [48] and Guan et al [49], 5 important types of concerns exist when conducting monitoring:

1. Concept drift: Also known as model drift, this occurs when the patterns or relationships estimated by the model shift from their state at deployment. These shifts can manifest instantaneously, gradually, cyclically, or temporarily, potentially compromising the model's effectiveness. This can occur abruptly (eg, new treatments altering outcome rates) or gradually ("AI aging"), requiring continuous recalibration [49].
2. Data drift: This phenomenon emerges when the distribution of production data significantly differs from the training data. While the negative impact of data drift on model

XSL•FO
RenderX

performance typically develops slowly, its effects compound over time, leading to increasingly unreliable results [48].

3. Data integrity: Production environments may introduce unexpected variables, incompatible data types, or inconsistent measurements that differ from the model's training environment. These integrity issues can severely impact the model's accuracy and reliability, requiring immediate attention and correction [48].
4. Bias drift: Even a model developed with careful consideration for fairness and bias mitigation can become unfair in production due to bias drift. As societies evolve, existing biases may diminish while new ones emerge, requiring continuous monitoring and adjustment to maintain ethical performance [48].
5. Calibration drift: Highlighted by Guan et al [49] as a distinct concern in health care: a model may maintain discrimination but lose calibration, producing systematically over- or underestimated risk scores that undermine trust and safety [49].

The categories of drift outlined by Data Science Alliance [48] can be directly linked to statistical methods described by Guan et al [49], which provide concrete means of distinguishing meaningful changes from random variation. By pairing each drift type with corresponding identification strategies, it becomes possible to operationalize conceptual risks into measurable safeguards. Table 3 shows how SAFE-AI connects Data Science Alliance's [48] framework of monitoring concerns with Guan et al [49] recommended statistical approaches.

Table . Statistical tests and drift identification strategies (adapted from Guan et al [49]).

Drift type and formula or test	When to use
Concept drift	
Cumulative Sum and Shewhart or SPC[a] control charts	• For monitoring performance metrics, such as AUC[b] or calibration slope, over time to detect gradual "AI[c] aging" or abrupt performance changes that are effective for real-time detection of sustained deviations.
Data drift	
Maximum mean discrepancy	• For high-dimensional feature spaces (eg, latent embeddings from imaging or EHR[d] data) • It detects subtle covariate shifts.
Wasserstein Distance (Earth Mover's Distance)	• For comparing distributions of latent features, like in medical imaging or pathology data • It is sensitive to spatial domain shifts
Kolmogorov–Smirnov test	• For univariate continuous variables (eg, lab values and vital signs) • It outputs a test statistic and a *P* value to assess whether the distributions differ significantly.
Jensen–Shannon divergence	• For comparing feature or prediction distributions across temporal windows, be robust to temporal shifts.
Bias drift	
Chi-square test	• For categorical variables (eg, sex, race, and hospital unit) • Applied in metadata to detect subgroup distribution shifts
Fisher exact test	• For small sample sizes where chi-square assumptions fail (eg, minority subgroup outcome analysis).
Calibration drift	
Calibration slope or intercept and dynamic calibration curves	• For tracking whether predicted probabilities remain aligned with observed outcome frequencies.
Data integrity	
Rule-based validation and schema checks	• For detecting anomalies, such as new, unexpected, or missing variable encodings and inconsistent measurement units

[a]SPC: statistical process control.

[b]AUC: area under the curve.

[c]AI: artificial intelligence.

[d]EHR: electronic health record.

XSL•FO
RenderX

To distinguish these forms of drift from random variation, Guan et al [49] emphasize the importance of embedding formal statistical testing within monitoring pipelines. Data drift can be identified using statistical distance measures, such as Maximum mean discrepancy or Wasserstein distance for high-dimensional features, Kolmogorov-Smirnov tests for continuous variables, and Jensen-Shannon divergence for monitoring temporal distributional shifts. Subgroup outcome disparities should be routinely assessed with categorical association tests, such as chi-square or Fisher exact test, ensuring fairness is maintained across demographic groups. Calibration drift requires continuous evaluation of calibration slope and intercept, or dynamic calibration curves, to confirm that predicted probabilities remain aligned with observed outcomes. For concept drift and temporal performance degradation, including the gradual loss of discrimination, sequential statistical process control methods, such as cumulative summation and Shewhart control charts, are effective for detecting "AI aging" in deployed systems (Table 3). In this way, the SAFE-AI framework operationalizes Guan et al [49] recommendations by mapping conceptual drift categories to concrete identification strategies relevant for health care contexts.

Operationalizing these methods requires automated pipelines that compute drift, calibration, and fairness metrics at scheduled intervals, with alerts triggered when thresholds are exceeded. However, detection alone is insufficient [49,50]. The need for multidisciplinary review is stressed, where data scientists, clinicians, and ethics experts interpret signals and determine whether retraining, recalibration, or workflow changes are warranted. Equally important is the structured documentation of drift events, calibration measures, and corrective actions, which provides transparency, supports regulatory alignment, and builds trust. Embedding such statistical guardrails within governance processes reduces false alarms and enables timely interventions before performance degradation compromises patient care.

Below are the recommended steps to implement an effective monitoring structure:

1. Automate monitoring and alerting: Deploy automated pipelines that calculate drift and fairness metrics at scheduled intervals. Integrate with alert systems (eg, dashboards, email, and Slack notifications) to flag deviations as soon as they occur. For example, trigger an alert if PSI >0.25 or subgroup accuracy difference >10%.
2. Scheduled reviews: Implement periodic review sessions (eg, monthly or quarterly) involving stakeholders (data scientists, clinicians, and ethics experts) to analyze monitoring reports and determine if intervention, such as model retraining or recalibration, is needed.
3. Document changes: Maintain a structured log (eg, an SQL table or Pandas DataFrame) capturing timestamps, versions, drift statistics, fairness scores, and actions taken.

Table 4 illustrates how SAFE-AI's automated monitoring and documentation process operates in practice. The table represents a simplified monthly log of model performance and fairness metrics tracked through an automated pipeline. In this example, Model 1.2.1 initially performs within acceptable bounds, showing low data drift (PSI <0.10) and stable fairness scores.

Table . Example of monitoring statistics table.

Date	Model_ID	Data_Drift (PSI[a])	Accuracy	F_1-score	Fairness_Parity	Bias_Flag	Notes
January 3, 2025	1.2.1	0.07	0.86	0.91	0.94	Pass	Stable
January 4, 2025	1.2.1	0.08	0.81	0.92	0.93	Pass	Stable
January 5, 2025	1.2.1	0.11	0.79	0.89	0.90	Warning	Stable
January 6, 2025	1.2.1	0.15	0.74	0.85	0.88	Alert	Notified developer
January 7, 2025	1.2.1	0.19	0.68	0.80	0.84	Fail	Failed
January 8, 2025	1.3.0	0.05	0.88	0.92	0.95	Pass	Updated

[a]PSI (population stability index):

By May 2025, drift begins to rise and accuracy drops slightly, triggering a warning flag. In June, the system automatically issues an alert notification to developers as drift and subgroup disparity increase. By July, both performance and fairness fall below defined SAFE-AI thresholds, producing a fail status that prompts model retraining.

A month later, version 1.3.0 was deployed. The updated model restores accuracy, stability, and fairness parity to acceptable levels, confirming successful remediation.

This example demonstrates how SAFE-AI's automated monitoring, alerting, and documentation pipeline detects early signs of degradation, escalates issues for review, and records corrective actions, creating a transparent, auditable trail of model health over time.

1. User feedback integration: Establish channels for end-user (clinician and patient) feedback regarding both model performance and explainability. Use structured surveys and focus groups to gather qualitative data that can inform further improvements.
2. Iterative improvement: Based on audit findings and feedback, implement retraining or fine-tuning cycles that reincorporate recent data. Validate updated models against the established responsibility metrics before redeployment.

XSL•FO
RenderX

Effective system monitoring and evaluation represent crucial components of responsible AI deployment. Potential issues can be identified and addressed early on with the right tools in place. This proactive approach, combined with structured feedback mechanisms and continuous improvement processes, helps ensure that AI systems remain both technically sound and ethically aligned with their intended purposes.

Much of the monitoring and evaluation effort falls within the "0 - 6 hours" of additional sprint time associated with SAFE-AI integration (Figure 3). Importantly, this represents the structured version of work that most teams already perform informally, such as checking whether performance has shifted, manually reviewing error cases, or responding to unexpected model behavior in production, rather than a new category of tasks introduced by SAFE-AI. This framework formalizes existing monitoring responsibilities, provides statistical methods to replace ad hoc practices, and ensures that time spent diagnosing drift, fairness issues, or degraded calibration is tracked as intentional sprint work with owners and acceptance criteria. In this way, the monitoring step reflects not additional governance layers, but a clearer, more accountable articulation of work that must occur to sustain responsible AI performance over time.

Discussion

Balancing Ethical Oversight With Business Objectives

The SAFE-AI framework distinguishes itself by aligning ethical rigor with business priorities and development speed. Whereas existing frameworks, such as ECCOLA [51] or academic or regulatory models, are often perceived as prescriptive or cumbersome, SAFE-AI integrates oversight into Agile workflows without slowing delivery. Ethical review is framed not as external compliance but as a strategic safeguard against regulatory, reputational, and financial risks.

SAFE-AI frames ethical assessment not as an overhead cost, but as a strategic risk mitigation process that helps ensure model integrity, maintain user trust, and prevent potential regulatory, reputational, or financial harm. SAFE-AI is purposefully designed to complement the natural rhythms of modern software and AI product development by integrating ethical review seamlessly into existing Agile and Scrum workflows.

One of the primary distinctions of SAFE-AI is its emphasis on efficiency without compromising ethical standards. Rather than exhaustive checklists or governance layers, SAFE-AI establishes the minimum necessary safeguards to preserve responsibility while remaining agile. This proportional approach makes it especially practical for SMEs with limited resources.

A key aspect of this balance is recognizing that the time SAFE-AI introduces into the development cycle, such as the additional 6 hours of work during a sprint, is not a totally new burden but a structured articulation of tasks that development teams typically perform informally. Most teams already spend time investigating unexpected model behavior, resolving data inconsistencies, responding to stakeholder concerns, or informally reviewing fairness or performance anomalies. SAFE-AI does not create totally new categories of work, but it simply moves these ad hoc activities into the sprint backlog, where they can be estimated, tracked, owned, and completed with the same discipline as technical tasks to ensure they are completed. By making previously invisible labor explicit rather than adding new governance layers, SAFE-AI preserves development speed while ensuring that ethical safeguards are delivered consistently and transparently.

A central innovation is the SPAMM. Unlike SHAP, LIME, or other interpretability tools that emphasize technical explanations, SPAMM translates model behavior into real-world narratives and cost-based analogies. This enables nontechnical stakeholders to grasp the implications of model errors, facilitating informed trade-offs between ethical risk and business priorities.

SAFE-AI also emphasizes iterative oversight. AI systems evolve through retraining, tuning, and updates. SAFE-AI requires previously defined acceptance and responsibility metrics to be revisited after each substantive change. This ensures updates do not reintroduce bias or compromise fairness while avoiding unnecessary rework.

Finally, SAFE-AI frames ethical responsibility as a shared duty across cross-functional teams, embedding accountability within everyday workflows rather than delegating it to specialized ethics committees.

The Critical Role of Risk Interpretability and Metric Development

A core principle of ethical AI development in health care is ensuring that the inherent risks of automated decision-making are both interpretable and measurable. Modeling risk interpretability is not a secondary concern but a foundational necessity for responsible AI deployment. Without clear and agreed-upon mechanisms for understanding how an algorithm's predictions translate into real-world outcomes, health care providers and affected individuals are left to navigate an opaque and potentially harmful system.

The introduction of the SAFE-AI framework addresses this challenge directly by developing probability analogy methods to bridge the gap between complex model behavior and the practical, human-driven ethical decision-making processes within health care settings. Unlike traditional explainability tools, which often focus exclusively on mathematical metrics or technical explanations, while emphasizing narrative-driven, probabilistic scenarios that map model behavior to understandable analogies. This approach acknowledges the reality that nontechnical stakeholders, including clinicians, patients, and regulatory bodies, must ultimately be able to evaluate and engage with the outputs of AI systems in order to trust them.

The significance of this method is especially relevant during the assessment phase of SAFE-AI. Setting responsibility metrics is not merely a technical exercise but an ethical obligation. Without these metrics, it is impossible to contextualize and evaluate how well the algorithm is performing against real-world risks. The development of responsibility metrics, including fairness and transparency metrics, must occur in tandem with risk interpretation strategies to ensure that metrics are not abstract or arbitrary but rather aligned with the actual impact on the user population.

XSL•FO
RenderX

Additionally, one of the most overlooked sources of ethical risk in AI systems is introduced not during the initial development but during subsequent algorithmic updates, such as model retraining or hyperparameter tuning. Even seemingly minor updates can alter the balance of an algorithm in subtle ways that inadvertently introduce ethical imbalances or exacerbate existing disparities. As described in earlier sections, SAFE-AI mandates that any update to an existing inference engine—whether through additional training data, tuning, or rebalancing—should be treated as a new project cycle if there is any risk that new training data can create predictions that fall out of bounds. Although previously established assessment metrics can and should be reused, they must be explicitly re-evaluated to verify that any modifications have not unintentionally undermined fairness, transparency, or safety.

This approach balances the need for iterative improvement with the ethical imperative to avoid harm. Too often, pressure to deploy quickly or incrementally improve system performance results in bypassing critical ethical assessments; however, many algorithms use random samples as part of their retraining process, and a quick iteration can inadvertently create vastly different outcomes. SAFE-AI stresses that, even in fast-paced environments, evaluating updated models against established metrics is essential. The cost of skipping this step is not merely technical debt but ethical liability, with real consequences for patient safety and equity.

Finally, it is important to emphasize that the development of these metrics and risk interpretability mechanisms cannot be left solely to technical teams. The most effective way to safeguard against ethical harm is through multidisciplinary collaboration. Risk interpretation is inherently contextual, requiring input from health care professionals, regulatory experts, ethicists, and product stakeholders who understand the social and operational realities in which the AI system will function. This collaborative approach ensures that metrics are not only technically valid but also ethically relevant and meaningful to the people most impacted by the model's behavior.

By prioritizing transparency, risk interpretability, and ongoing evaluation using the approaches outlined in the "Develop Responsibility Metrics" section, the framework supports the development of AI systems that are not only effective but also fair, accountable, and ethically sound.

Validating the Framework in Real-World Health Care AI Settings

To assess SAFE-AI's real-world relevance and impact, future work should focus on comprehensive field-testing across diverse health care AI applications. The empirical analysis would involve implementing the SAFE-AI framework with SMEs developing various medical AI systems, including diagnostic support tools, patient outcome assessment, and wound trajectory prediction. Using a pre-post intervention design, these studies would evaluate the framework's impact on development practices, fairness outcomes, and organizational dynamics in real-world health care AI deployments.

Within health care settings, the field studies would specifically examine how the SAFE-AI framework performs across various clinical applications, where ethical considerations regarding patient care and outcomes are particularly critical. The research would investigate how effectively SAFE-AI addresses issues, such as training on biased datasets, transparency requirements, preventing provider overreliance on AI systems across different clinical contexts, and how to ensure its seamless integration into existing processes. By studying implementation across diverse medical specialties, it is possible to assess SAFE-AI's adaptability to varying clinical workflows, data types, and ethical priorities.

The proposed implementation process would include a baseline assessment to capture preframework metrics, structured implementation of the responsibility protocols, and postimplementation evaluation at multiple intervals. This systematic approach would allow for the identification of potential adoption challenges, including resource constraints, workflow disruptions, and organizational resistance. Field testing would evaluate various implementation strategies, such as gradual integration, iterative refinement, and leadership engagement models, to determine which approaches most effectively support framework adoption while maintaining rigorous ethical standards.

Building on these future validation efforts, the framework presented here offers a practical approach to embedding ethical considerations throughout health care AI development while maintaining efficient workflows. By transforming abstract principles into concrete processes with measurable outcomes, this framework helps bridge the gap between technical implementation and ethical governance. As health care AI systems become increasingly integrated into clinical decision-making, such structured approaches to responsible development will be essential for building trustworthy systems that genuinely advance patient care without introducing new forms of bias or harm.

Comparison With Existing Frameworks

Table 5 situates SAFE-AI alongside ECCOLA and CARE-AI (Collaborative Assessment for Responsible and Ethical AI Implementation), two of the most prominent ethical AI frameworks. While ECCOLA provides broad governance prompts and CARE-AI emphasizes health care assurance cases, both approaches are resource-intensive and often external to fast-paced product development. In contrast, SAFE-AI was co-designed with practitioners to embed ethical safeguards directly into Agile or Scrum sprints, collaboratively customize responsibility metrics to the project at hand, and lower adoption barriers for SMEs. While the development of SAFE-AI benefited from an expert consortium, this involvement intentionally front-loaded ethical and regulatory complexity into the framework design so that downstream SME users can apply its artifacts without requiring comparable ethics or academic partnerships.

XSL•FO
RenderX

Table . Example of SPAMM[a] compared to technical metrics.

Criteria	ECCOLA	CARE-AI[b]	SAFE-AI[c]
Origin or context	Academic or consortium-led, broad AI[d] governance guidelines	Health care–oriented ethical assurance	SME[e]-practical, co-designed with practitioners
Workflow integration	Provides ethical prompts or checklists, but external to development sprints	Process guidance tied to assurance cases; heavy documentation	Fully embedded into Agile or Scrum sprints with sprint-based checkpoints
Fairness coverage	High-level bias prompts, no quantitative thresholds	Addresses equity in health care safety cases	Defines fairness metrics with subgroup thresholds and SPAMM narratives
Transparency coverage	Suggests explainability but lacks concrete methods	Assurance-case documentation for regulators	Combines methods like SHAP[f] or LIME[g] with SPAMM for all domains (scenario-based narrative analogies)
Monitoring	Minimal; assumes postdeployment audits	Focus on compliance reporting and auditability	Continuous monitoring with statistical drift detection triggers
Adoption burden	High (lengthy prompts or checklists and steep learning curve)	High (documentation intensive; requires assurance specialists)	Low to moderate (checklists and metrics integrated into existing development tools)
Scalability to SMEs	Limited; designed for larger institutions	Limited; high compliance load	Explicitly designed for SMEs; pragmatic, “minimum necessary” safeguards

[a]SPAMM: scenario-based probability analogy mapping method.

[b]CARE-AI: (Collaborative Assessment for Responsible and Ethical AI Implementation).

[c]SAFE-AI: Scalable Agile Framework for Execution in AI.

[d]AI: artificial intelligence.

[e]SME: small and medium-sized enterprise.

[f]SHAP: Shapley Additive Explanations.

[g]LIME: Local Interpretable Model-Agnostic Explanations.

A key differentiator is SAFE-AI’s explicit use of fairness metrics and SPAMM narratives to operationalize equity and transparency across domains, whereas ECCOLA and CARE-AI stop at higher-level guidance. Similarly, SAFE-AI integrates continuous monitoring with statistical drift detection, allowing ongoing oversight without the compliance burden of full assurance-case documentation.

Taken together, these differences highlight SAFE-AI’s pragmatic contribution, as it delivers a “minimum necessary” governance structure that is rigorous enough to support fairness, transparency, and monitoring, yet lean enough for SMEs to adopt without specialized compliance staff.

Conclusion

The SAFE-AI ethical development framework proposed in this paper offers a practical, business-aligned approach for integrating responsible AI practices into the everyday workflows of product and data science teams. By mirroring established Agile and Scrum methodologies, SAFE-AI avoids treating ethics as a parallel or external process, instead embedding ethical evaluation directly into the structure of software and AI development. This alignment ensures that ethical considerations are addressed without compromising delivery timelines or development agility—a critical factor for SMEs operating under resource constraints.

Central to SAFE-AI is the principle that AI development introduces uncertainty and risk that cannot be adequately managed through traditional software engineering approaches alone. Accordingly, SAFE-AI emphasizes the early and ongoing use of measurable ethical metrics—acceptance criteria, fairness metrics, and transparency metrics—to evaluate model behavior. These metrics serve as practical tools for managing the probabilistic nature of AI systems, enabling teams to assess whether their models meet not only performance expectations but also ethical and social obligations.

Importantly, the SAFE-AI framework extends beyond initial deployment by incorporating a monitoring and evaluation phase that supports continuous ethical oversight. It recognizes that AI systems are dynamic and must be re-evaluated whenever retraining, tuning, or environmental shifts occur. This ensures that ethical integrity is not treated as a static checkpoint but rather as an evolving requirement that must be actively maintained throughout the product lifecycle.

A key innovation introduced by SAFE-AI is the scenario-based probability analogy mapping process as an interpretability technique designed to enhance trust among nontechnical stakeholders. Translating probabilistic model behavior into familiar, narrative-based analogies with associated risks and costs helps explain how and why AI systems reach certain predictions—especially in high-stakes health care settings where

XSL•FO
RenderX

transparency is critical. Unlike traditional explainability tools that focus on model internals, SAFE-AI centers on real-world interpretability, offering an accessible way to communicate model behavior, associated risks, and uncertainty in decision-making. As such, it not only supports transparency but also serves as a bridge between ethical modeling and user-centered design.

Taken together, the contributions of SAFE-AI establish a scalable pathway for responsible AI adoption in health care and beyond where modeling excellence, software best practices, and ethical integrity are not at odds but rather deeply interwoven into a cohesive and actionable development process.

Acknowledgments

The authors gratefully acknowledge Professor David Danks, University of California, San Diego, for insightful comments that improved the clarity and rigor of the manuscript. The author prepared portions of this manuscript with the assistance of ChatGPT (GPT-5, OpenAI, 2025) to support phrasing refinement. The author reviewed, verified, and revised all AI-assisted text for accuracy and originality, ensuring that all factual content, references, and interpretations are the author's own. No generative AI was used for data analysis, study design, or interpretation of results.

Funding

This study was partially supported by the National Institute on Aging's Artificial Intelligence and Technology Collaboratories for Aging Research Grant P30AG073104 (WWP), the US Army xTech AI Grand Challenge (WWP), the Huntsman Mental Health Foundation (WWP); as well as Mountain Biometrics, Inc. (WWP), Nemsee, LLC (IN), and the Data Science Alliance (AM, LJ, RL, and PL). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the US Army or the US Government.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

Authors' Contributions

Conceptualization: IN, AM, LJ, RL, WWP
Investigation: IN, AM, LJ, RL, WWP
Methodology: IN, AM, LJ, RL, WWP
Project administration: IN, WWP, PL
Supervision: IN, WWP, PL
Validation: IN, AM, RL, WWP
Visualization: IN, LJ, WWP
Writing – original draft: IN
Writing – review & editing: IN, AM, WWP

Conflicts of Interest

WWP is a shareholder in Mountain Biometrics Inc., and IN is a shareholder in Nemsee, LLC. All other authors declare no competing interests.

Checklist 1
SAFE-AI (Scalable Agile Framework for Execution in AI) phase summary and implementation checklist.
[XLSX File, 10 KB - jmir_v28i1e80028_app1.xlsx]

References

1. What is Scrum? Scrum.org. URL: https://www.scrum.org/resources/what-scrum-module [accessed 2025-12-02]
2. The 2020 Scrum Guide. Scrum Guides. URL: https://scrumguides.org/scrum-guide.html [accessed 2025-12-02]
3. Ali SJ, Christin A, Smart A, Katila R. Walking the walk of AI ethics: organizational challenges and the individualization of risk among ethics entrepreneurs. Presented at: FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency; Jun 12-15, 2023; Chicago IL USA p. 217-226 URL: https://dl.acm.org/doi/proceedings/10.1145/3593013 [doi: 10.1145/3593013.3593990]
4. Crockett K, Colyer E, Gerber L, Latham A. Building trustworthy AI solutions: a case for practical solutions for small businesses. IEEE Trans Artif Intell 2023 Aug;4(4):778-791. [doi: 10.1109/TAI.2021.3137091]
5. Hajianhosseinabadi H, Lindau N. Startup interaction with AI - A multiple case study on startups' ethical navigation when using Generative AI tools.: University of Gothenburg; 2024. URL: https://gupea.ub.gu.se/handle/2077/82378 [accessed 2025-02-26]

XSL•FO
RenderX

6. Kostick-Quenet KM, Cohen IG, Gerke S, et al. Mitigating racial bias in machine learning. J Law Med Ethics 2022;50(1):92-100. [doi: 10.1017/jme.2022.13] [Medline: 35243993]
7. Ning Y, Liu X, Collins GS, et al. An ethics assessment tool for artificial intelligence implementation in healthcare: CARE-AI. Nat Med 2024 Nov;30(11):3038-3039. [doi: 10.1038/s41591-024-03310-1]
8. Berman G, Goyal N, Madaio M. A scoping study of evaluation practices for responsible AI tools: steps towards effectiveness evaluations. Presented at: CHI '24: CHI Conference on Human Factors in Computing Systems; May 11-6, 2024; Honolulu, HI p. 1-24 URL: https://dl.acm.org/doi/proceedings/10.1145/3613904 [doi: 10.1145/3613904.3642398]
9. Winecoff AA, Watkins EA. Artificial concepts of artificial intelligence: institutional compliance and resistance in AI startups. AIES '22: AAAI/ACM Conference on AI, Ethics, and Society :788-799. [doi: 10.1145/3514094.3534138]
10. Fjeld J, Achten N, Hilligoss H, Nagy A, Srikumar M. Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI. SSRN Journal 2020. [doi: 10.2139/ssrn.3518482]
11. Clinical decision support software frequently asked questions (FAQs). US Food and Drug Administration. 2025. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/clinical-decision-support-software-frequently-asked-questions-faqs [accessed 2025-01-23]
12. Clinical decision support software guidance for industry and food and drug administration staff. US Food and Drug Administration. 2022. URL: https://www.fda.gov/media/109618/download [accessed 2025-01-23]
13. Rep. arrington JC [R-T-19]. H.R.1 - 119th congress (2025-2026): one big beautiful bill act. Congress.gov. 2025. URL: https://www.congress.gov/bill/119th-congress/house-bill/1/text [accessed 2025-06-30]
14. Domin H, Rossi F, Goehring B, Ganapini M, Berente N, Bevilacqua M. On the ROI of AI ethics and governance investments: from loss aversion to value generation. California Management Review. 2024 Jul 29. URL: https://cmr.berkeley.edu/2024/07/on-the-roi-of-ai-ethics-and-governance-investments-from-loss-aversion-to-value-generation/ [accessed 2025-09-19]
15. Nondiscrimination in health programs and activities. Centers for Medicare & Medicaid Services. 2024. URL: https://www.federalregister.gov/documents/2024/05/06/2024-08711/nondiscrimination-in-health-programs-and-activities [accessed 2025-01-23]
16. Bevilacqua M, Berente N, Domin H, Goehring B, Rossi F. The return on investment in AI ethics: a holistic framework. arXiv. Preprint posted online on Nov 7, 2023. [doi: 10.48550/arXiv.2309.13057]
17. Good machine learning practice for medical device development: guiding principles. US Food and Drug Administration.: FDA; 2023. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles [accessed 2025-01-23]
18. Artificial intelligence and machine learning in software as a medical device. US Food and Drug Administration. 2025. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device [accessed 2025-01-23]
19. Informed consent faqs. US Department of Health and Human Services. URL: https://www.hhs.gov/ohrp/regulations-and-policy/guidance/faq/informed-consent/index.html [accessed 2025-02-05]
20. Pfeiffer RS, Forsberg RP. Ethics on the Job: Cases and Strategies, 4th edition: Wadsworth; 2013.
21. Pfeiffer RS. Teaching ethical decision-making: the case study method and the RESOLVEDD strategy. Teach Philos 1992 May 1;15(2):175-184. [doi: 10.5840/teachphil199215221]
22. Vakkuri V, Kemell KK, Jantunen M, Halme E, Abrahamsson P. ECCOLA — A method for implementing ethically aligned AI systems. J Syst Softw 2021 Dec;182:111067. [doi: 10.1016/j.jss.2021.111067]
23. Acceptance criteria: examples and best practices. Atlassian. URL: https://www.atlassian.com/work-management/project-management/acceptance-criteria [accessed 2025-01-23]
24. Everything you need to know plus examples. Scrum Alliance. URL: https://resources.scrumalliance.org/Article/need-know-acceptance-criteria [accessed 2025-01-23]
25. Confronting bias: BSA's framework to build trust in AI. : BSA; 2021 URL: https://www.bsa.org/files/reports/2021bsaaibias.pdf
26. Ratwani RM, Sutton K, Galarraga JE. Addressing AI algorithmic bias in health care. JAMA 2024 Oct 1;332(13):1051-1052. [doi: 10.1001/jama.2024.13486] [Medline: 39230911]
27. Assessment. Fairlearn. URL: https://fairlearn.org/v0.12/user_guide/assessment/index.html [accessed 2025-01-23]
28. Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. Science 2021 Jul 16;373(6552):284-286. [doi: 10.1126/science.abg1834]
29. Abujaber AA, Nashwan AJ. Ethical framework for artificial intelligence in healthcare research: a path to integrity. World J Methodol 2024 Sep 20;14(3):94071. [doi: 10.5662/wjm.v14.i3.94071] [Medline: 39310239]
30. Weerts H, Dudík M, Edgar R, Jalali A, Lutz R, Madaio M. Fairlearn: assessing and improving fairness of AI systems. arXiv. Preprint posted online on Mar 29, 2023. [doi: 10.48550/arXiv.2303.16626]
31. Molnar C. Local Model-Agnostic Methods. In: Interpret Mach Learn 2018. URL: https://christophm.github.io/interpretable-ml-book/agnostic.html [accessed 2025-02-05]
32. Watson D. Conceptual challenges for interpretable machine learning. SSRN Journal 2020 Jan:65. [doi: 10.2139/ssrn.3668444]

XSL•FO
RenderX

33. Kim S. Explainable AI (XAI) methods part 3 - accumulated local effects (ALE). Towards Data Science. 2022. URL: https://towardsdatascience.com/explainable-ai-xai-methods-part-3-accumulated-local-effects-ale-cf6ba3387fde/ [accessed 2026-01-25]
34. Mutegeki H, Nahabwe A, Nakatumba-Nabende J, Marvin G. Interpretable machine learning-based triage for decision support in emergency care. 2023 Presented at: 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI); Apr 11-13, 2023; Tirunelveli, India p. 983-990. [doi: 10.1109/ICOEI56765.2023.10125918]
35. The AI/ML wars: “explain” or test black box models? error stat philos. Error Statistics. 2022. URL: https://errorstatistics.com/2022/03/23/the-ai-ml-wars-explain-or-test-black-box-models/ [accessed 2025-02-05]
36. Ribeiro MT, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. arXiv. Preprint posted online on Jun 16, 2016. [doi: 10.48550/arXiv.1606.05386]
37. Aamodt A, Plaza E. Case-based reasoning: foundational issues, methodological variations, and system approaches. Eur J Artif Intell 1994;7(1):39-59. [doi: 10.3233/AIC-1994-7104]
38. Cai CJ, Jongejan J, Holbrook J. The effects of example-based explanations in a machine learning interface. 2019 Mar 17 Presented at: IUI ’19; Mar 17-20, 2019; Marina del Ray California p. 258-262 URL: https://dl.acm.org/doi/proceedings/10.1145/3301275 [doi: 10.1145/3301275.3302289]
39. IONOS Editorial Team. What is case-based reasoning? IONOS. 2025. URL: https://www.ionos.com/digitalguide/websites/web-development/case-based-reasoning/ [accessed 2025-02-05]
40. Costa M. Social return on investment. In: Idowu SO, Capaldi N, Zu L, Gupta AD, editors. Encyclopedia of Corporate Social Responsibility: Springer; 2013:2238-2244. [doi: 10.1007/978-3-642-28036-8_734]
41. Millar R, Hall K. Social Return on Investment (SROI) and Performance Measurement: the opportunities and barriers for social enterprises in health and social care. Public Manag Rev 2013 Sep;15(6):923-941. [doi: 10.1080/14719037.2012.698857]
42. NIST AI RMF Playbook. NIST AIRC. URL: https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook [accessed 2025-01-23]
43. Blueprint for trustworthy AI implementation guidance and assurance for healthcare version 10. : Coalition for Health AI; 2023 URL: https://assets.ctfassets.net/7s4afyr9pmov/4AXIWGIlcrjWDaW2ueTaRS/f98e5cb2528187635895cce6ba5ec309/Blueprint_for_Trustworthy_AI.pdf
44. Baquero JA, Burkhardt R, Govindarajan A, Wallace T. Derisking AI: risk management in AI development. McKinsey. 2020. URL: https://www.mckinsey.com/capabilities/quantumblack/our-insights/derisking-ai-by-design-how-to-build-risk-management-into-ai-development?utm_source=chatgpt.com [accessed 2025-02-14]
45. Mello MM, Guha N. Understanding liability risk from healthcare AI. : Stanford University: Human-Centered Artificial Intelligence; 2024 URL: https://hai-production.s3.amazonaws.com/files/2024-02/Liability-Risk-Healthcare-AI.pdf
46. Total product life cycle for medical devices. US Food and Drug Administration. 2023. URL: https://www.fda.gov/about-fda/cdrh-transparency/total-product-life-cycle-medical-devices [accessed 2025-02-14]
47. Reddy GP, Pavan Kumar YV, Prakash KP. Hallucinations in large language models (LLMs). 2024 Apr 25 Presented at: 2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream); Apr 25-25, 2024; Vilnius, Lithuania p. 1-6. [doi: 10.1109/eStream61684.2024.10542617]
48. The framework for responsible data science practices. : Data Science Alliance; 2024 URL: https://drive.google.com/file/d/17cdW49APhyv7qEy0EyQV6vv7qkw_EcIQ/view?usp=embed_facebook
49. Guan H, Bates D, Zhou L. Keeping medical AI healthy and trustworthy: a review of detection and correction methods for system degradation. arXiv. Preprint posted online on Dec 2, 2025. [doi: 10.48550/arXiv.2506.17442]
50. Saenz AD, McCoy T, Mantha AB, et al. Establishing responsible use of AI guidelines: a comprehensive case study for healthcare institutions. npj Digit Med 2024 Nov 30;7(1):1-6. [doi: 10.1038/s41746-024-01300-8]
51. Vakkuri V, Kemell KK, Abrahamsson P. Ethically aligned design: an empirical evaluation of the RESOLVEDD-strategy in software and systems development context. 2019 Presented at: 2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA); Aug 28-30, 2019; Kallithea-Chalkidiki, Greece p. 46-50. [doi: 10.1109/SEAA.2019.00015]

Abbreviations

AI: artificial intelligence
ALE: accumulated local effects
BSA: Business Software Alliance
CCPA: California Consumer Privacy Act
DSA: Data Science Alliance
EU: European Union
FDA: US Food and Drug Administration
GDPR: General Data Protection Regulation
GenAI: generative artificial intelligence
HHS OCR: Department of Health and Human Services - Office for Civil Rights
HIPAA: Health Insurance Portability and Accountability Act

XSL•FO
RenderX

IE: inference engine
IRB: Institutional Review Board
LIME: Local Interpretable Model-Agnostic Explanations
ML: machine learning
MTN: Medical Timeseries Networks
NIST: National Institute of Standards and Technology
PDP: partial dependence plots
PE: product engine
PSI: Population Stability Index
SAFE-AI: Scalable Agile Framework for Execution in AI
SaMD: software as a medical device
SDLC: software development life cycle
SHAP: Shapley Additive Explanations
SME: small and medium-sized enterprise
SPAMM: Scenario-based Probability Analogy Mapping Method
TPLC: Total Product Life Cycle

Edited by A Coristine; submitted 03.Jul.2025; peer-reviewed by A Munoz-Zavala, VV Sangaraju, X Liang; accepted 18.Dec.2025; published 25.Feb.2026.

Please cite as:
Nemteanu I, Mancebo Jr A, Joe L, Lopez R, Lopez P, Pettine WW
Scalable Agile Framework for Execution in AI for Medical AI Ethics Policy Design in Small- and Medium-Sized Enterprises
J Med Internet Res 2026;28:e80028
URL: https://www.jmir.org/2026/1/e80028
doi:10.2196/80028

© Ion Nemteanu, Adir Mancebo Jr, Leslie Joe, Ryan Lopez, Patricia Lopez, Warren Woodrich Pettine. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

From Agents to Governance: Essential AI Skills for Clinicians in the Large Language Model Era

Weiping Cao[1,2], MMed; Qing Zhang[1*], MD; Jialin Liu[3,4*], MD; Siru Liu[5], PhD

[1]Department of Cardiology, West China Hospital, Sichuan University, Chengdu, Sichuan, China

[2]Department of Cardiology, The People's Hospital of Leshan, Leshan, Sichuan, China

[3]Information Center, West China Hospital, Sichuan University, Chengdu, China

[4]Department of Medical Informatics, West China Medical School, Sichuan University, Chengdu, China

[5]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

[*]these authors contributed equally

Corresponding Author:
Jialin Liu, MD
Information Center
West China Hospital
Sichuan University
No.37 Guo Xue Xiang Street
Chengdu, 610041
China
Phone: 86 28 85422416
Fax: 86 28 85422607
Email: dljl8@163.com

Abstract

Large language models are rapidly transitioning from pilot schemes to routine clinical practice. This creates an urgent need for clinicians to develop the necessary skills to strike the right balance between seizing opportunities and taking accountability. We propose a 3-tier competency framework to support clinicians' evolution from cautious users to responsible stewards of artificial intelligence (AI). Tier 1 (foundational skills) defines the minimum competencies for safe use, including prompt engineering, human–AI agent interaction, security and privacy awareness, and the clinician-patient interface (transparency and consent). Tier 2 (intermediate skills) emphasizes evaluative expertise, including bias detection and mitigation, interpretation of explainability outputs, and the effective clinical integration of AI-generated workflows. Tier 3 (advanced skills) establishes leadership capabilities, mandating competencies in ethical governance (delineating accountability and liability boundaries), regulatory strategy, and model life cycle management—specifically, the ability to govern algorithmic adaptation and change protocols. Integrating this framework into continuing medical education programs and role-specific job descriptions could enhance clinicians' ability to use AI safely and responsibly. This could standardize deployment and support safer clinical practice, with the potential to improve patient outcomes.

(J Med Internet Res 2026;28:e86550) doi:10.2196/86550

KEYWORDS

large language model; clinician; agent; competency; artificial intelligence; education; continuing medical education

Introduction

The convergence of generative artificial intelligence (AI) and health care is catalyzing a paradigm shift in clinical practice, with significant implications for the future of medicine [1-3]. Large language models (LLMs), exemplified by recent advances, such as GPT-4 and Gemini, demonstrate a transformative capacity to process multimodal data and generate context-aware responses, increasingly positioning them as integral components in frontline clinical decision support [4,5].

Although LLMs have the potential to improve clinical effectiveness, ensuring that their application optimizes patient safety, ethical alignment, and long-term benefits remains a substantial challenge [2,5]. This complexity is compounded by the intersection of regulatory requirements and ethical obligations. Evolving legal frameworks, such as the European Union (EU) AI Act and the US Food and Drug Administration

XSL•FO
RenderX

(FDA) guidance, explicitly mandate human oversight for high-risk AI systems [6,7]. Simultaneously, global ethical standards from the World Health Organization (WHO) and the American Medical Association emphasize the necessity of physician leadership and accountability [2,8,9]. However, a gap remains in translating these high-level mandates into actionable clinical skills. Without the active leadership and input of clinicians, these technologies risk imposing unintended burdens and may fail to achieve their full potential [1].

The imperative for advanced AI governance arises from a fundamental shift from passive information retrieval to autonomous task execution. While conventional LLM paradigms rely on user-initiated prompt response exchanges, clinicians query the model and verify its text outputs, and the system does not autonomously call external tools. By contrast, emerging agentic workflows introduce a perceive-plan-act (and often reflect) loop [10]. In this mode, the system interprets high-level clinical intents (eg, hypertension management); decomposes them into subtasks; and autonomously executes actions via application programming interfaces, such as accessing electronic health record (EHR) data or calculating risk scores [11,12]. This transition reframes supervision; clinicians must move beyond prompt engineering to govern how autonomy is delegated, how actions are constrained, and how escalation pathways are formalized.

To address these regulatory, ethical, and technical demands, we propose a foundational, tiered AI competency framework for clinicians. The framework is structured around progressive tiers: tier 1 (foundational skills), tier 2 (intermediate skills), and tier 3 (advanced skills). We describe the core competencies at each tier, outline the framework's limitations, and propose priority directions for validation to sustain its relevance amid an evolving regulatory landscape.

AI Competency Framework

LLM-enabled care necessitates a transition in the roles of clinicians (physicians, nurses, pharmacists, and allied health professionals)—from interpreting predictive outputs to supervising agentic workflows. Drawing on previous research, a narrative synthesis of evolving digital health competencies, and an analysis of the technical capabilities of LLMs [13-17], we propose a 3-tier, governance-aligned framework that articulates core LLM competencies. As illustrated in Figure 1, the framework progresses from foundational safe use (tier 1) to evaluative proficiency (tier 2) and ultimately to governance and leadership (tier 3).

Figure 1. Essential artificial intelligence (AI) skills for clinicians in the large language model era.

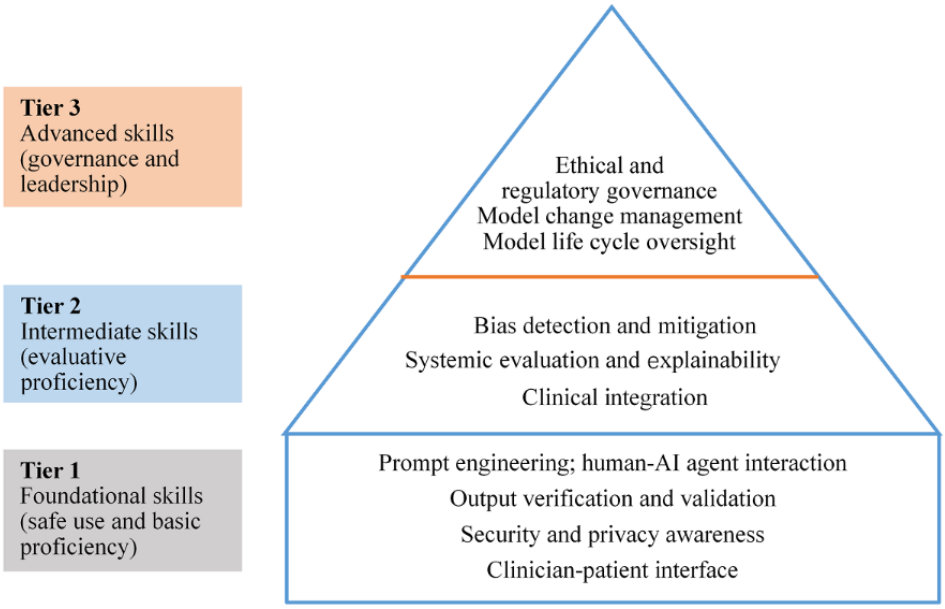

Tier 1: Foundational Skills (Safe Use and Basic Proficiency)

These entry-level competencies prioritize basic interaction with LLMs, enabling clinicians to leverage AI for routine tasks without compromising clinical autonomy. The key elements are described subsequently. First, prompt engineering (task specification for clinician-initiated and hybrid workflows) is used to craft precise, context-aware instructions—with explicit roles, required inputs, constraints, task steps, and output formats (including citation and traceability requirements)—to elicit task-appropriate outputs (eg, structured outlines for differential diagnosis). This competency primarily supports clinician-initiated chat and hybrid workflows, as fully agentic perceive-plan-act execution is typically governed by system-level prompts and policies rather than user-generated prompts. When paired with verification and source grounding, this approach may reduce hallucinations and improve relevance and completeness [18]. Second, human-AI agent interaction (agent supervision) ensures that agents operate within bounded autonomy with explicit roles, goals, and guardrails. Clinicians must maintain awareness of least privilege tool permissions and system constraints (eg, data minimization, time and step limits, and sandboxed execution) [19], with clear termination and escalation criteria. Clinicians also monitor and validate the perceive-plan-act-reflect loop using provenance and citation requirements; protected health information redaction; and audit

XSL•FO
RenderX

logging of prompts, tool calls, and human overrides [20-22]. When confidence or calibration thresholds are not met (eg, coverage targets and abstention or deferral rules), clinicians intervene, interrupt the agent, or revert to manual workflows and document the event for review. Third, output verification and validation involve clinicians critically evaluating individual LLM outputs for accuracy, relevance, and internal consistency. Generated content (eg, summaries, diagnostic considerations, and treatment plans) is cross-referenced against the EHR, structured data, and established clinical evidence to detect hallucinations, omissions, or misstatements. In culturally diverse settings, clinicians must assess outputs for cultural safety and linguistic accuracy. This involves verifying that translated instructions and culturally specific dietary or lifestyle advice are appropriate for the patient's context. Clinicians should also check for Anglocentric bias that could conflict with local norms or the patient's language proficiency. This human-in-the-loop verification is essential for ensuring patient safety in individual clinical encounters [23,24]. The fourth element is security and privacy awareness. To comply with regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation, clinicians must adhere to foundational safeguards centered on data minimization and appropriate tool use. For nonintegrated or open access LLM interfaces, this includes avoiding entry of protected health information and direct identifiers and, when clinically necessary to discuss a case, removing or generalizing nonessential identifiers before input [2,5,25,26]. In contrast, for authorized, integrated enterprise agents operating within a secure EHR environment, manual deidentification is often neither feasible nor necessary; instead, clinicians verify minimum necessary access, confirm the agent is scoped to the correct patient context, and ensure permissions are aligned with the specific clinical task through role-based access control and least privilege settings. Rather than conducting technical audits themselves, clinicians prevent inadvertent privacy breaches by distinguishing approved tools from unapproved ones and escalating permission or access-scope concerns through institutional channels. These baseline competencies are prerequisites for safe AI use in routine clinical workflows [5]. Fifth, clinician-patient interface (transparent communication and shared decision-making) involves incorporating AI-assisted content into the clinical encounter without undermining patient trust or the therapeutic alliance. Clinicians should disclose when AI is used (eg, AI-scribed summaries, patient-portal messages, and patient education materials) to uphold patient autonomy and informed consent [26].

Tier 2: Intermediate Skills (Evaluative Proficiency)

Building on foundational knowledge, these competencies center on critically assessing and integrating LLMs into clinical reasoning workflows while addressing bias and uncertainty in generative AI. First, bias detection and mitigation require clinicians to interpret algorithmic bias audit reports and uncertainty quantification outputs (eg, confidence intervals, prediction intervals, or conformal prediction sets when provided) to assess reliability across patient subgroups. Clinicians initiate and validate remediation actions—such as recommending prompt or workflow adjustments and defining escalation and deferral rules—in coordination with technical teams, ensuring adherence to prespecified fairness metrics and minimum subgroup performance thresholds [27,28]. For example, in tumor grading, clinicians review reported subgroup performance using minimum sample-size thresholds, calibration and coverage, abstention rates, and uncertainty displays (including confidence sets). They assess model rationale and interpret between-group performance gaps. Second, systemic evaluation and explainability involve moving beyond checking individual outputs to assessing the broader reliability, calibration, and failure modes of the AI system. Clinicians should be able to interpret model performance metrics (eg, sensitivity, specificity, error and hallucination rates, and performance in specific subpopulations) and evaluate available explainability outputs (eg, feature importance scores, reason codes, or saliency maps where available) to understand why a model reached a conclusion [29]. This evaluation must include equity audits that assess model performance across distinct subgroups (eg, race, ethnicity, and language) [28]. For instance, a model may demonstrate high overall accuracy but fail disproportionately for languages spoken by minority groups or specific dialects. Clinicians leading the evaluation must identify such disparities and determine if the model is safe for deployment in diverse populations. These skills enable clinicians to judge systemic trustworthiness and identify appropriate clinical use cases and target populations for which the model is calibrated, effective, and equitable [30]. Third, clinical integration requires clinicians to use domain knowledge to refine model outputs (eg, align treatment suggestions with evidence-based guidelines) while monitoring for potential deskilling. Clinicians maintain human-AI collaboration and specify deferral and escalation rules (eg, abstention thresholds and human-review triggers) and document these events for auditability [31,32].

Tier 3: Advanced Skills (Governance and Leadership)

Unlike the foundational skills in tiers 1 and 2, this tier represents a specialized track for clinician-leaders, clinical informaticists, and physician builders assuming governance roles. These competencies focus on the strategic oversight and architectural direction of AI systems. The main competencies involved are described subsequently. First, ethical and regulatory governance involves overseeing the development of institutional policies for LLM use to ensure alignment with ethical principles, data protection laws (eg, General Data Protection Regulation and HIPAA), and international guidance [26]. This requires establishing governance infrastructure—such as AI steering committees and ethics review boards—to specify authorized use cases, roles and responsibilities, liability frameworks, and compliance protocols. Crucially, policies must explicitly delineate accountability boundaries among supervising clinicians, health care institutions, and AI developers and vendors, particularly for autonomous or semiautonomous agentic workflows. In this capacity, clinician leaders do not personally conduct technical audits; instead, they serve as the strategic link between medical staff and technical bodies, ensuring that institutional processes reflect clinical realities and patient safety risks. Second, model change management requires supervision of domain adaptation (eg, task- or specialty-specific tuning) within a multidisciplinary process. In this capacity, clinicians

XSL•FO
RenderX

bridge clinical needs and technical implementation, upholding standards for validity, equity, and safety. This supervision necessitates predefined evaluation plans, comprehensive documentation (eg, model cards), and rigorous external validation (including multicenter, temporal, and geographic shift tests) before production deployment. Leaders must specify minimum performance thresholds and mandate shadow deployment phases to validate safety before full patient exposure [33,34]. Third, model life cycle oversight entails governing AI systems across their full life cycle—from validation through postmarket monitoring, updating, and decommissioning. This includes orchestrating institutional processes for drift detection, performance re-evaluation, and version control [35,36] (Textbox 1). Leaders must navigate complex regulatory mechanisms for iterative improvement, such as the predetermined change control plans (PCCPs) by the FDA [6] and the postmarket surveillance requirements of the EU AI Act [7]. They collaborate with informatics, regulatory, and quality teams to ensure that updates, retraining, or expanded indications are clinically justified, transparently communicated, and supported by robust evidence and incident review protocols [35,37].

Textbox 1. Clinical vignette—governance in action.

Scenario: executing a manual rollback protocol

- A clinical informatics director oversees a deployed discharge-summary agent. During routine postmarket surveillance, the monitoring dashboard signals that the model's summarization accuracy has dropped below the prevalidated threshold of 95% (a metric specified in the Food and Drug Administration–accepted predetermined change control plan). Attributing the decline to data drift caused by a recent update in the hospital's note-template format, the director initiates a rollback protocol—leveraging either institutional version control or a vendor-mediated pathway specified in the service-level agreement. The deployment is rolled back to the previous stable version (version 2.0) while the technical team remediates and revalidates the updated model (version 2.1).
- In deployments where direct rollback is technically unsupported (eg, some software as a service–based integrations), the protocol mandates pausing or disabling the agent and reverting to manual workflows until remediation is complete.

Alignment and Differentiation From Existing Frameworks

This framework is broadly aligned with the American Medical Association's guidance on augmented intelligence, prioritizing physician leadership, transparency, and patient benefit [38,39]. Furthermore, it adheres to competency-based digital education frameworks from the WHO and the Association of American Medical Colleges, both of which prioritize observable behaviors and measurable learning outcomes [40-42]. It also builds on recent competency proposals in AI and digital health that foreground digital health literacy, awareness of data bias, and the ethical use of assistive tools [10,43]. As summarized in Table 1, our contribution lies in extending these earlier frameworks to the agentic LLM era. First, we explicitly differentiate between predictive and informational paradigms and agentic workflows. Accordingly, we move from clinicians interpreting decision support outputs to supervising and governing active, tool-using agents. Second, we introduce model life cycle literacy as an explicit competency domain, encompassing familiarity with mechanisms for ongoing monitoring, updating, and regulatory adaptation. Within this broader, jurisdiction-agnostic concept, PCCPs in the FDA context are presented as one concrete example, alongside emerging requirements under frameworks, such as the EU AI Act. To our knowledge, previous frameworks have not explicitly integrated agent supervision and life cycle–oriented governance into a tiered, clinician-facing competency model.

Table 1. Comparison of the agent to the governance framework and existing digital health competency frameworks.

Feature and dimension	Agent to governance framework	Existing frameworks (eg, World Health Organization, American Medical Association, and Association of American Medical Colleges)
Technological scope	Agentic and autonomous: agentic workflows (perceive-plan-act loops) and tool-using large language models that execute multistep tasks	Predictive and informational: clinical decision support, diagnostic classifiers, and standard information retrieval systems
Clinician's role	Supervisor and governor: human-on-the-loop oversight for task delegation, monitoring agent behavior, and managing bounded autonomy	Interpreter and decision-maker: human-in-the-loop integration, focusing on the critical appraisal of risk scores and diagnostic suggestions
Verification skills	Output verification and logic checking: detection of hallucinations in generative text and verification of agentic tool calls (eg, application programming interface actions)	Statistical and evidence-based appraisal: evaluation of model performance metrics (eg, sensitivity and specificity), data quality, and automation bias
Regulatory and life cycle	Life cycle management: specific literacy in predetermined change control plans, algorithmic drift detection, and postmarket surveillance (eg, European Union Artificial Intelligence Act and Food and Drug Administration)	Foundational ethics and compliance: adherence to core bioethical principles (beneficence and equity), privacy standards (Health Insurance Portability and Accountability Act and General Data Protection Regulation), and informed consent
Target audience and structure	Tiered differentiation: distinguishes between frontline users (tiers 1 and 2) and a specialized leadership track (tier 3) for governance	Universal digital literacy: baseline digital health competencies applicable to the broad health care workforce to ensure safe general use

XSL•FO
RenderX

Operationalizing Competencies for Education and Assessment

Translation of this framework into continuing medical education (CME) curricula requires the specification of observable, assessable behaviors aligned with competency-based medical education principles. Given that the clinical workforce encompasses diverse roles—including physicians, nurses, and allied health professionals—implementation and assessment should be tailored to role-specific scope of practice and role-based EHR access controls. For example, behavioral indicators involving least privilege enforcement or the rejection of agent actions may be operationalized differently depending on the individual's credentialed permissions and administrative privileges. To ensure implementation feasibility and mitigate workforce burden, only tier 1 competencies are intended for the general clinical workforce, whereas tiers 2 and 3 are reserved for smaller groups of superusers and clinician leaders in formal governance roles. To avoid adding entirely new courses, these competencies are designed to be integrated into existing curricula (eg, evidence-based medicine, clinical reasoning, and quality and safety) and CME activities. Institutions are responsible for resourcing and coordinating these training activities, ensuring that individual clinicians are not expected to acquire advanced competencies (tiers 2 and 3) without appropriate organizational support and protected time.

Table 2 links each tier to behavioral indicators written as active, measurable learning outcomes. Indicators span tier 1 (eg, identifying hallucinations) to tier 3 (eg, initiating life cycle protocols) and should be tailored to role-specific responsibilities and decision rights. These indicators provide curriculum developers with a concrete scaffolding to design simulation-based, workplace-based, and microlearning assessments that verify skill acquisition in clinical practice. Ultimately, these anchors facilitate the incorporation of this framework into CME curricula and clinical job descriptions, thereby promoting institutional transparency, accountability, and regulatory alignment [40,43,44].

Table 2. Sample behavioral indicators for continuing medical education assessment and clinical application.

Core competency	Behavioral indicator (observable action)
Tier 1: foundational (frontline user)	
Prompt engineering	Formulates a context-aware prompt that includes explicit role definitions (eg, act as a cardiologist), constraints, and required output formats, without disclosing PHI[a]
Human-AI[b] agent interaction	Identifies and intercepts inappropriate agent requests (eg, social history for refills) and enforces denial or escalation protocols based on least privilege guardrails and predefined termination and handoff criteria
Output verification and validation	Detects and corrects a hallucinated reference or dosage in a large language model–generated draft by cross-referencing with the patient's structured laboratory data and trusted guidelines
Security and privacy awareness	Uses minimum necessary data; deidentifies data for nonintegrated tools; for electronic health record agents, verifies patient context and least privilege access, and escalates PHI or policy risks
Clinician-patient interface	Informs patients when AI is used, explains AI-derived insights in patient-appropriate language (including uncertainties and limitations), and documents consent or refusal when clinically indicated.
Tier 2: intermediate (superuser or champion)	
Bias detection and mitigation	Interprets stratified subgroup performance and uncertainty reports, flags clinically meaningful disparities, triggers mitigation (eg, threshold adjustments or human-review rules), and verifies improvement via updated audit reports
Systemic evaluation	Evaluates a confusion matrix for a diagnostic AI tool to determine if the false-negative rate is acceptable for a specific screening population
Explainability	Interprets available explainability outputs (eg, feature importance, reason codes, or saliency maps where available) to detect spurious cues and document potential failure modes
Clinical integration	Defines where AI outputs enter the workflow; assigns roles, documentation, and escalation steps; and maintains clinician accountability when AI recommendations conflict
Tier 3: advanced (governance leader)	
Ethical and regulatory governance	Drafts and implements an institutional policy that establishes escalation pathways and explicitly delineates accountability and liability boundaries among the supervising clinician, the institution, and the AI developer for autonomous agentic workflows
Model change management	Initiates and justifies model change requests (eg, recalibration, retraining, or expanded indication), defining the clinical rationale, validation plan, and monitoring criteria consistent with the predetermined change control plan
Model life cycle oversight	Oversees monitoring of model performance and drift (eg, calibration, error rates, and data shifts); ensures execution of incident protocols and predefined controls (eg, roll back and human review)

[a]PHI: protected health information.

[b]AI: artificial intelligence.

XSL•FO
RenderX

Limitations and Future Work

We propose a governance-aligned competency framework designed to guide clinicians in the safe and effective use of LLMs in clinical practice. However, several limitations should be acknowledged. First, external validity may differ by specialty, care setting (inpatient vs ambulatory), health-system maturity, and EHR integration capacity. Critically, the institutional infrastructure required for tier 3—specifically, the establishment of AI steering committees—may currently be feasible only in resource-rich academic medical centers. Mandating such governance structures in resource-constrained community hospitals may be impractical. This feasibility gap extends to global health contexts; the framework requires adaptation in low- and middle-income settings where informatics infrastructure, governance capacity, and regulatory regimes differ substantially. Moreover, the objective structured clinical examination blueprint [44] and key performance indicators remain surrogate end points. By themselves, these measures do not guarantee improvements in patient-centered outcomes (eg, adverse events and readmissions). Second, a prospective, multicenter external evaluation is still necessary. Although we specify fairness analyses and minimum subgroup sample size and performance thresholds, real-world coverage across languages, cultural contexts, pediatrics, geriatrics, and rare-disease pathways is likely incomplete. Third, the regulatory environment remains dynamic as harmonized standards under the EU AI Act and FDA change control frameworks (eg, PCCPs) continue to evolve. Accordingly, operationalized procedures and performance thresholds should be periodically reassessed—particularly following material model updates—to sustain regulatory compliance and clinical relevance. Fourth, the automation paradox (skill decay) warrants attention. While AI agents improve efficiency, they may precipitate clinician deskilling over time. Safe fallback protocols (eg, reverting to manual workflows during system failure) are feasible only if clinicians maintain underlying diagnostic and procedural competence. To mitigate this risk, organizations should integrate automation-off scenarios into simulation training and downtime contingency plans to ensure clinicians remain capable of detecting errors and safely resuming control.

Given the rapid evolution of clinical AI, this framework requires ongoing refinement across 5 strategic areas. First, validation studies use expert consensus (eg, Delphi methods) and multicenter educational trials to link tiered competencies to observable clinical behaviors, such as error interception, safe deferral, and workflow efficiency. Second, assessment science strengthens psychometric measurement by refining objective structured clinical examination stations, bias and calibration checklists, and reliability targets (eg, generalizability coefficient [G] ≥0.70) [45,46]. Third, capacity building establishes faculty development programs and reusable educational resources—including deidentified sandboxes, annotated audit log exemplars, and HIPAA-aligned exercise sets—to support cross-specialty implementation. Fourth, advanced equity and uncertainty quantification cultivates practical competence in algorithmic fairness and uncertainty management through routine subgroup audits, coverage targets and reporting (using conformal prediction where appropriate), and bedside remediation playbooks. Fifth, simulation and institutional integration evaluate the feasibility and effectiveness of embedding automation-on and automation-off scenarios and rollback procedures within existing simulation programs and downtime contingency planning (eg, multidisciplinary team discussions). Outcomes include competency attainment, error interception during failures, auditability, and pathways for formal recognition within CME credit structures and institutional job descriptions.

Conclusions

The progressive competency model integrates technical proficiency with ethical governance to provide clinicians with essential AI skills for the LLM era. By embedding these competencies into CME standards and job descriptions—using clear, observable behaviors—institutions can standardize safe and accountable AI use. Preparing for an AI-augmented future requires integrating governance-focused skills into medical training and professional development. This approach positions clinicians as responsible stewards of AI, ensuring adoption remains sustainable and centered on patient care.

Acknowledgments

Following the initial drafting of the manuscript, the authors used ChatGPT (OpenAI) for artificial intelligence–assisted language editing to enhance clarity and style. All suggested edits were subject to critical review, and the authors subsequently made additional substantive revisions. The authors assume complete responsibility for the integrity and accuracy of the final content.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Data Availability

Data sharing is not applicable to this paper as no new datasets were generated or analyzed during this study.

XSL•FO
RenderX

Authors' Contributions

JL, QZ, and SL conceived and designed the study. WC and JL led data curation, with literature analysis by WC, JL, QZ, and SL. JL, WC, QZ, and SL contributed to writing the original draft. All authors critically revised the manuscript, approved the final version, and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

References

1. Angus DC, Khera R, Lieu T, Liu V, Ahmad FS, Anderson B, JAMA Summit on AI. AI, health, and health care today and tomorrow: the JAMA Summit Report on Artificial Intelligence. JAMA 2025 Nov 11;334(18):1650-1664. [doi: 10.1001/jama.2025.18490] [Medline: 41082366]
2. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. World Health Organization. 2025 Mar 25. URL: https://www.who.int/publications/i/item/9789240084759 [accessed 2025-09-10]
3. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
4. Liu S, Huang SS, McCoy AB, Wright AP, Horst S, Wright A. Optimizing order sets with a large language model-powered multiagent system. JAMA Netw Open 2025 Sep 02;8(9):e2533277 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.33277] [Medline: 40986301]
5. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, FUTURE-AI Consortium. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ 2025 Feb 05;388:e081554 [FREE Full text] [doi: 10.1136/bmj-2024-081554] [Medline: 39909534]
6. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions. United States Food and Drug Administration. 2025 Aug. URL: https://tinyurl.com/5dr3xx6j [accessed 2025-09-12]
7. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). European Union. 2024 Jun 13. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng [accessed 2025-09-10]
8. Augmented intelligence in health care. American Medical Association. 2018. URL: https://www.ama-assn.org/system/files/2019-01/augmented-intelligence-policy-report.pdf [accessed 2025-09-16]
9. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. 2021 Jun 28. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2025-09-10]
10. Xi Z, Chen W, Guo X, He W, Ding Y, Hong B, et al. The rise and potential of large language model based agents: a survey. Sci China Inf Sci 2025 Jan 17;68(2):121101. [doi: 10.1007/s11432-024-4222-0]
11. Jiang Y, Black KC, Geng G, Park D, Zou J, Ng AY, et al. MedAgentBench: a virtual EHR environment to benchmark medical LLM agents. NEJM AI 2025 Aug 28;2(9):1-9. [doi: 10.1056/aidbp2500144]
12. Skalidis I, Maurizi N, Salihu A, Fournier S, Cook S, Iglesias JF, et al. Artificial intelligence and advanced digital health for hypertension: evolving tools for precision cardiovascular care. Medicina (Kaunas) 2025 Sep 04;61(9):1597 [FREE Full text] [doi: 10.3390/medicina61091597] [Medline: 41010987]
13. Liu J, Liu F, Fang J, Liu S. The application of Chat Generative Pre-trained Transformer in nursing education. Nurs Outlook 2023 Nov;71(6):102064. [doi: 10.1016/j.outlook.2023.102064] [Medline: 37879261]
14. Schuitmaker L, Drogt J, Benders M, Jongsma K. Physicians' required competencies in AI-assisted clinical settings: a systematic review. Br Med Bull 2025 Jan 16;153(1):ldae025. [doi: 10.1093/bmb/ldae025] [Medline: 39821209]
15. Scott IA, Shaw T, Slade C, Wan TT, Barmanray R, Coorey C, et al. Proposing core competencies for physicians in using artificial intelligence tools in clinical practice. Intern Med J 2025 Aug;55(8):1403-1409. [doi: 10.1111/imj.70112] [Medline: 40576330]
16. Goodman KE, Rodman AM, Morgan DJ. Preparing physicians for the clinical algorithm era. N Engl J Med 2023 Aug 10;389(6):483-487. [doi: 10.1056/NEJMp2304839] [Medline: 37548320]
17. Russell RG, Lovett Novak L, Patel M, Garvey KV, Craig KJ, Jackson GP, et al. Competencies for the use of artificial intelligence-based tools by health care professionals. Acad Med 2023 Mar 01;98(3):348-356. [doi: 10.1097/ACM.0000000000004963] [Medline: 36731054]
18. Liu J, Liu F, Wang C, Liu S. Prompt engineering in clinical practice: tutorial for clinicians. J Med Internet Res 2025 Sep 15;27:e72644 [FREE Full text] [doi: 10.2196/72644] [Medline: 40955776]
19. Naveen K, Sajja WS, Nerella A. Building secure AI agents for autonomous data access in compliance/regulatory-critical environments. Comput Fraud Secur 2024;2024(9):363-373 [FREE Full text] [doi: 10.52710/cfs.746]
20. Huang K. Agentic AI: Theories and Practices. Cham, Switzerland: Springer; 2025.

XSL•FO
RenderX

21. Acharya DB, Kuppan K, Divya B. Agentic AI: autonomous intelligence for complex goals—a comprehensive survey. IEEE Access 2025;13:18912-18936. [doi: 10.1109/access.2025.3532853]
22. Borkowski AA, Ben-Ari A. Multiagent AI systems in health care: envisioning next-generation intelligence. Fed Pract 2025 May;42(5):188-194. [doi: 10.12788/fp.0589] [Medline: 40831649]
23. Lu JG, Song LL, Zhang LD. Cultural tendencies in generative AI. Nat Hum Behav 2025 Nov 20;9(11):2360-2369. [doi: 10.1038/s41562-025-02242-1] [Medline: 40542181]
24. Grazhdanski G, Vasilev V, Vassileva S, Taskov D, Antova I, Koychev I, et al. SynthMedic: utilizing large language models for synthetic discharge summary generation, correction and validation. J Biomed Inform 2025 Oct;170:104906 [FREE Full text] [doi: 10.1016/j.jbi.2025.104906] [Medline: 40962129]
25. Jonnagaddala J, Wong ZS. Privacy preserving strategies for electronic health records in the era of large language models. NPJ Digit Med 2025 Jan 16;8(1):1-3 [FREE Full text] [doi: 10.1038/s41746-025-01429-0] [Medline: 39820020]
26. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023 Aug 11;25:e48009 [FREE Full text] [doi: 10.2196/48009] [Medline: 37566454]
27. Ganta T, Kia A, Parchure P, Wang MH, Besculides M, Mazumdar M, et al. Fairness in predicting cancer mortality across racial subgroups. JAMA Netw Open 2024 Jul 01;7(7):e2421290 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.21290] [Medline: 38985468]
28. Omar M, Sorin V, Agbareia R, Apakama DU, Soroush A, Sakhuja A, et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review. Int J Equity Health 2025 Feb 26;24(1):57 [FREE Full text] [doi: 10.1186/s12939-025-02419-0] [Medline: 40011901]
29. Mesinovic M, Watkinson P, Zhu T. Explainability in the age of large language models for healthcare. Commun Eng 2025 Jul 17;4(1):128 [FREE Full text] [doi: 10.1038/s44172-025-00453-y] [Medline: 40676176]
30. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature 2025 Jun 09;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
31. Berzin TM, Topol EJ. Preserving clinical skills in the age of AI assistance. The Lancet 2025 Oct;406(10513):1719. [doi: 10.1016/s0140-6736(25)02075-6]
32. Kim SH, Wihl J, Schramm S, Berberich C, Rosenkranz E, Schmitzer L, et al. Human-AI collaboration in large language model-assisted brain MRI differential diagnosis: a usability study. Eur Radiol 2025 Sep 07;35(9):5252-5263. [doi: 10.1007/s00330-025-11484-6] [Medline: 40055233]
33. Dorfner FJ, Dada A, Busch F, Makowski MR, Han T, Truhn D, et al. Evaluating the effectiveness of biomedical fine-tuning for large language models on clinical tasks. J Am Med Inform Assoc 2025 Jun 01;32(6):1015-1024. [doi: 10.1093/jamia/ocaf045] [Medline: 40190132]
34. Makhni S, Rico J, Cerrato P, Hill B, Overgaard S, Wu J, et al. A comprehensive approach to responsible AI development and deployment. Mayo Clin Proc Digit Health 2025 Dec;3(4):100294 [FREE Full text] [doi: 10.1016/j.mcpdig.2025.100294] [Medline: 41282583]
35. M S AR, C R N, B R S, Lahza H, Lahza HF. A survey on detecting healthcare concept drift in AI/ML models from a finance perspective. Front Artif Intell 2023 Apr 17;5:955314 [FREE Full text] [doi: 10.3389/frai.2022.955314] [Medline: 37139355]
36. Kore A, Abbasi Bavil E, Subasri V, Abdalla M, Fine B, Dolatabadi E, et al. Empirical data drift detection experiments on real-world medical imaging data. Nat Commun 2024 Feb 29;15(1):1887 [FREE Full text] [doi: 10.1038/s41467-024-46142-w] [Medline: 38424096]
37. Moskalenko V, Kharchenko V. Resilience-aware MLOps for AI-based medical diagnostic system. Front Public Health 2024 Mar 27;12:1342937 [FREE Full text] [doi: 10.3389/fpubh.2024.1342937] [Medline: 38601490]
38. Augmented intelligence development, deployment, and use in health care. American Medical Association. 2024 Nov. URL: https://www.ama-assn.org/system/files/ama-ai-principles.pdf [accessed 2025-09-16]
39. Augmented intelligence in medicine. American Medical Association. URL: https://www.ama-assn.org/practice-management/digital-health/augmented-intelligence-medicine [accessed 2025-09-10]
40. Digital education for building health workforce capacity. World Health Organization. 2020 Apr 14. URL: https://www.who.int/publications/i/item/9789240000476 [accessed 2025-09-10]
41. Global competency and outcomes framework for the essential public health functions. World Health Organization. 2024 Jun 10. URL: https://www.who.int/publications/i/item/9789240091214 [accessed 2025-09-18]
42. Artificial intelligence competencies for medical educators. Association of American Medical Colleges. URL: https://tinyurl.com/44t88c4c [accessed 2025-10-21]
43. Gazquez-Garcia J, Sánchez-Bocanegra CL, Sevillano JL. AI in the health sector: systematic review of key skills for future health professionals. JMIR Med Educ 2025 Feb 05;11:e58161 [FREE Full text] [doi: 10.2196/58161] [Medline: 39912237]
44. Milan FB, Grochowalski JH. A resource efficient and reliable standard setting method for OSCEs: borderline regression method using standardized patients as sole raters in clinical case encounters with medical students. Med Teach 2022 Aug;44(8):878-885. [doi: 10.1080/0142159X.2022.2041586] [Medline: 35234562]
45. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. Med Teach 2012;34(11):960-992. [doi: 10.3109/0142159X.2012.703791] [Medline: 23140303]

XSL•FO
RenderX

46. Park SY, Lee SH, Kim MJ, Ji KH, Ryu JH. Acceptability of the 8-case objective structured clinical examination of medical students in Korea using generalizability theory: a reliability study. J Educ Eval Health Prof 2022 Sep 08;19:26 [FREE Full text] [doi: 10.3352/jeehp.2022.19.26] [Medline: 36071557]

Abbreviations

AI: artificial intelligence
CME: continuing medical education
EHR: electronic health record
EU: European Union
FDA: Food and Drug Administration
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
PCCP: predetermined change control plan
WHO: World Health Organization

Edited by A Coristine; submitted 26.Oct.2025; peer-reviewed by S Palmieri, A Algumaei; comments to author 24.Nov.2025; revised version received 23.Dec.2025; accepted 23.Dec.2025; published 14.Jan.2026.

Please cite as:
Cao W, Zhang Q, Liu J, Liu S
From Agents to Governance: Essential AI Skills for Clinicians in the Large Language Model Era
J Med Internet Res 2026;28:e86550
URL: https://www.jmir.org/2026/1/e86550
doi:10.2196/86550
PMID:

©Weiping Cao, Qing Zhang, Jialin Liu, Siru Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Collaborative and Cooperative Hospital "In-House" Medical Device Development and Implementation in the AI Age: The European Responsible AI Development (EURAID) Framework Compatible With European Values

Anett Schönfelder[1], MBA; Maria Eberlein-Gonska[2], Prof Dr; Manfred Hülsken-Giesler[3], Prof Dr; Florian Jovy-Klein[4], MSc; Jakob Nikolas Kather[1,5,6,7], Prof Dr; Elisabeth Kohoutek[8], RA; Thomas Lennefer[9], Dr rer nat, MSc; Elisabeth Liebert[10], MA; Myriam Lipprandt[11], Prof Dr; Rebecca Mathias[1], MSc; Hannah Sophie Muti[1,12], MD, Dr med; Julius Obergassel[13], MD, MHBA; Thomas Reibel[4], MSc; Ulrike Rösler[14], PhD; Moritz Schneider[15], MSc; Larissa Schlicht[16], MSc; Hannes Schlieter[17], Prof Dr; Malte L Schmieding[18], MD, MSc; Nils Schweingruber[13], MD; Martin Sedlmayr[19], Prof Dr; Reinhard Strametz[20], Prof Dr; Barbara Susec[21], BA; Magdalena Katharina Wekenborg[1], Dr rer nat; Eva Weicken[22], MD; Katharina Weitz[22], Dr rer nat; Anke Diehl[10*], Dr med, MD; Stephen Gilbert[1,17*], Prof Dr

[1]Else Kroener Fresenius Center for Digital Health, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Saxony, Germany

[2]University Hospital Carl Gustav Carus, Dresden, Germany

[3]Department of Nursing Science, Institute of Health Research and Education, University of Osnabrück, Osnabrück, Germany

[4]Institute for Technology and Innovation Management (TIM), RWTH Aachen University, Aachen, Germany

[5]Department of Medicine I, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Saxony, Germany

[6]Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany

[7]Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, United Kingdom

[8]Luther Rechtsanwaltsgesellschaft mbH, Cologne, Germany

[9]Department for Prevention, AOK Federal Association, Berlin, Germany

[10]Department for Digital Transformation, University Medicine Essen, Essen, Germany

[11]Institute of Medical Informatics, University Hospital RWTH Aachen, Aachen, North Rhine-Westphalia, Germany

[12]Department for Visceral, Thoracic and Vascular Surgery, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany

[13]University Medical Center Hamburg - Eppendorf, Hamburg, Germany

[14]Federal Institute for Occupational Safety and Health, Berlin, Germany

[15]Institute for Occupational Safety and Health of the German Social Accident Insurance, Sankt Augustin, Germany

[16]Faculty of Humanities and Social Sciences, Karlsruhe Institute of Technology, Karlsruhe, Germany

[17]Faculty of Business and Economics, TUD Dresden University of Technology, Dresden, Germany

[18]German Federal Ministry of Health, Berlin, Germany

[19]Institute for Medical Informatics and Biometry, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany

[20]Wiesbaden Institute for Healthcare Economics and Patient Safety (WiHelP), Wiesbaden, Germany

[21]ver.di, Berlin, Germany

[22]Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Berlin, Germany

[*]these authors contributed equally

Corresponding Author:
Anett Schönfelder, MBA
Else Kroener Fresenius Center for Digital Health
Faculty of Medicine and University Hospital Carl Gustav Carus
TUD Dresden University of Technology
Fetscherstr. 74
Dresden, Saxony, 01307
Germany

XSL•FO
RenderX

Phone: 49 35145815840
Email: anett.schoenfelder@ukdd.de

Abstract

The last years have seen an acceleration in the development and uptake of artificial intelligence (AI) systems by "early adopter" hospitals, caught between the pressures to "perform" and "transform" in a struggling health care system. This transformation has raised concerns among health care providers as their voices and location-specific workflows have often been overlooked, resulting in technologies that fail to integrate meaningfully into routine care and worsen rather than improve care processes. How can positive AI implementation be carried out in health care, aligned with European values? Based on a perspective that spans all stakeholders, we have created EURAID (European Responsible AI Development), a practical, human-centric framework for AI development and implementation based on agreed goals and values. We illustrate this approach through the co-development of a narrow-purpose "in-house" AI system, designed to help bridge the AI implementation gap in real-world clinical settings. This example is then expanded to address the broader challenges associated with complex, multiagent AI systems. By portraying all key stakeholders across the AI development life cycle and highlighting their roles and contributions within the process, real use cases, and methods for achieving iterative consensus, we offer a unique practical approach for safe and fast progress in hospital digital transformation in the AI age.

(J Med Internet Res 2026;28:e80754) doi:10.2196/80754

KEYWORDS

AI Act; digital transformation; in-house medical device development; agentic AI; AI life cycle; artificial intelligence

The Transformation of Future Medicine Through Artificial Intelligence Technologies

Will the slogans already heard in health care system strikes, such as "Trust Nurses, Not AI" and "AI has got to go!"[1,2], become more common? They reflect growing concerns about the evolving role of health care professionals (HCPs) in a changing health system, which persist despite reports that 20% of National Health Service (NHS) doctors are already using artificial intelligence (AI) daily [3]. Although the importance of digital transformation to enhance the efficiency of care delivery and to provide better models of care suited to modern age [4-6] is well recognized within care systems [7-11], it often cannot be comprehensively addressed, as health care systems worldwide find themselves caught between the need to both "perform" and "transform" in a system facing "firefighting" ongoing challenges [12-17]. The application of AI technologies has the potential to address some of those aspects (Table 1), as it can speed digital transformation and can (at least if applied well and if the associated potential barriers and uncertainties are jointly recognized and resolved) make health care more accessible, effective, and economically sustainable [18]. Examples of the positive impact of good AI implementation are (1) enhancement of clinical practice, particularly in areas such as diagnosis and personalized medicine [11,19,20]; (2) workflow improvements, by supporting administrative tasks such as transcription, patient communication, and patient-related recordkeeping [21,22]; and (3) increased operational efficiency, through the optimization of routine processes, enabling HCPs to work in a more patient-centered way [23], and potentially contribute to cost reductions [24,25]. With the recent introduction of "agentic AI" [26-29] and autonomous AI-enabled systems [30,31], far more systematic complexity can be handled by AI [32].

XSL•FO
RenderX

Table 1. Problems artificial intelligence (AI)–enabled transformation can address, approaches, challenges, and possible unintended consequences.

Current health system problems	Possible digital and AI[a] solutions	Implementation challenges and risks
Administrative workload unrelated to direct patient care [33,34], inefficient workflows, and fragmented communication burden on HCPs[b].	Automation of administrative and routine tasks, and AI-driven workflow optimization, allowing people to focus on patients.	• Different perspectives on which tasks to automate. • Increase in workload in some cases. • Risk of overreliance on AI outcomes with insufficient human oversight. • Automation of the current way of providing care without restructuring and rethinking processes. • Concerns about job security, the transformation of job roles, and medical malpractice.
Stress, duplication (eg, medical history) [35], and discontinuous care resulting from disconnected devices, limited interoperability, and manual coordination.	Adjusting the hospital's IT environment as an AI-sustained platform, characterized by high interoperability in itself and with other providers supporting seamless patient journeys.	• Deficient data quality, data silos, inadequate computational resources, a shortage of specialized expertise, and poor or nonexistent infrastructure between providers. • Concerns about safety and regulation.
Poor information flow and HCP training deficit.	AI-supported knowledge management to build confidence in usage.	• Shortage of HCPs limits time for training. • Various adoption readiness levels among HCPs. • Concerns about trust in technologies.

[a]AI: artificial intelligence.

[b]HCP: health care professional.

However, AI is not a panacea, and initial evaluations of real-world performance in clinical settings are mixed [36-38]. One reason is that AI implementation projects have often underestimated the importance of individual AI medical devices operating as interconnected clinical and technological infrastructures rather than being a collection of isolated, standalone algorithms. AI in health care over the next years needs to be seen as interacting, interdependent, and flexible applications [39], involving both broad- and narrow-purpose tools and models that closely interact with and reshape human workflows, while simultaneously, human workflows, adaptations, and experience reshape the use of AI, particular to the local setting and local approach to health care delivery.

Integration of Interactive AI Systems in Clinical Workflows Requires HCPs at the Core, Not as Observers

This future model needs HCPs at its core, not only as users interacting with AI systems, but as active participants in their co-design, procurement, implementation, monitoring, and evaluation. This idea is rooted in organizational and implementation theories, such as the "socio-technical systems theory" [40], that emphasizes the importance of a holistic perspective to jointly bridge human and technological capabilities, particularly in the context of autonomous technologies [41,42], and the "normalization process theory" [43], which acknowledges users' cognitive participation and collective action as key determinants in implementing, embedding, and integrating complex and new interventions (eg, AI systems) in daily practice [44,45]. "Human-centered AI" can take a cross-theoretical perspective by viewing AI systems not as stand-alone technologies, but as integral components of a broader sociotechnical system. Two perspectives are relevant: humans being able to understand AI and AI being able to understand humans [46]. For example, explainable AI (XAI) methods should not only address the technical transparency of machine learning models but also focus on human understanding [47]. On the other hand, AI systems need to take into account the needs, requirements, and mental models of humans [48] and the context of clinical decisions [49] to create explanations that are supportive in the clinical setting.

Yet, despite the substantial body of research on theoretical foundations, the translation of the underlying principles into everyday implementation of AI systems and clinical reality is lagging behind [50-54], often key aspects are neglected, and many implementation projects fail [55]. Problems often begin during the development of AI systems, which are frequently designed and tested in settings that are far removed from the everyday realities of clinical practice [56], and with HCPs and location-specific workflows often overlooked. The consequences of systems designed without sustained input from HCP and patients [57] are visible as they fail to demonstrate their suitability and worsen rather than improve processes, leading to the perception that the introduction of digital technologies into health care adds to the burden [57,58] (Figure 1), although general relief through well-implemented work aids would be very welcomed. That misalignment has been associated with increased stress among HCPs [59] (including "technostress" [60,61]), disconnected patient care [62,63], and has even resulted in other unintended negative consequences, such as HCPs resisting the use of the technologies [15], using technologies in unanticipated ways [64], or developing workarounds that may endanger patient care [65]. Insufficient digital health literacy and training among HCPs amplify these effects, leaving HCPs unprepared for the demands of interacting with intelligent systems [66]. Other consequences appearing in real-world implementation are model uncertainty [67], "AI hallucinations" or clinically harmful recommendations, bias [14], and context misalignment [68], which risk fragmented care and diminish patients' trust in technology-assisted decisions.

XSL•FO
RenderX

Figure 1. The introduction of artificial intelligence (AI) into clinical workflows is changing everyday clinical care and could, at least theoretically, enhance satisfaction, empower, upskill, and provide a better work environment and better interactions for health care professionals (HCPs) and patients; however, the reality is often much less positive. The upper circle is showing the current situation of health care delivery, which is characterized by a low level of digitalization and an ever-increasing amount of nonpatient-related activities, causing moderate satisfaction and happiness among both HCPs and patients. Care delivery transformation through AI can bring positive effects as shown in the green circle on the right (such as delivering better, more efficient and even more patient-centered care through optimized processes and well-balanced support systems) or, as is frequently the case, negative effects (red circle on the left), causing frustration, disconnection and stress of HCPs and patients because of interoperability issues with AI implementations that were never properly designed with the user needs in mind.

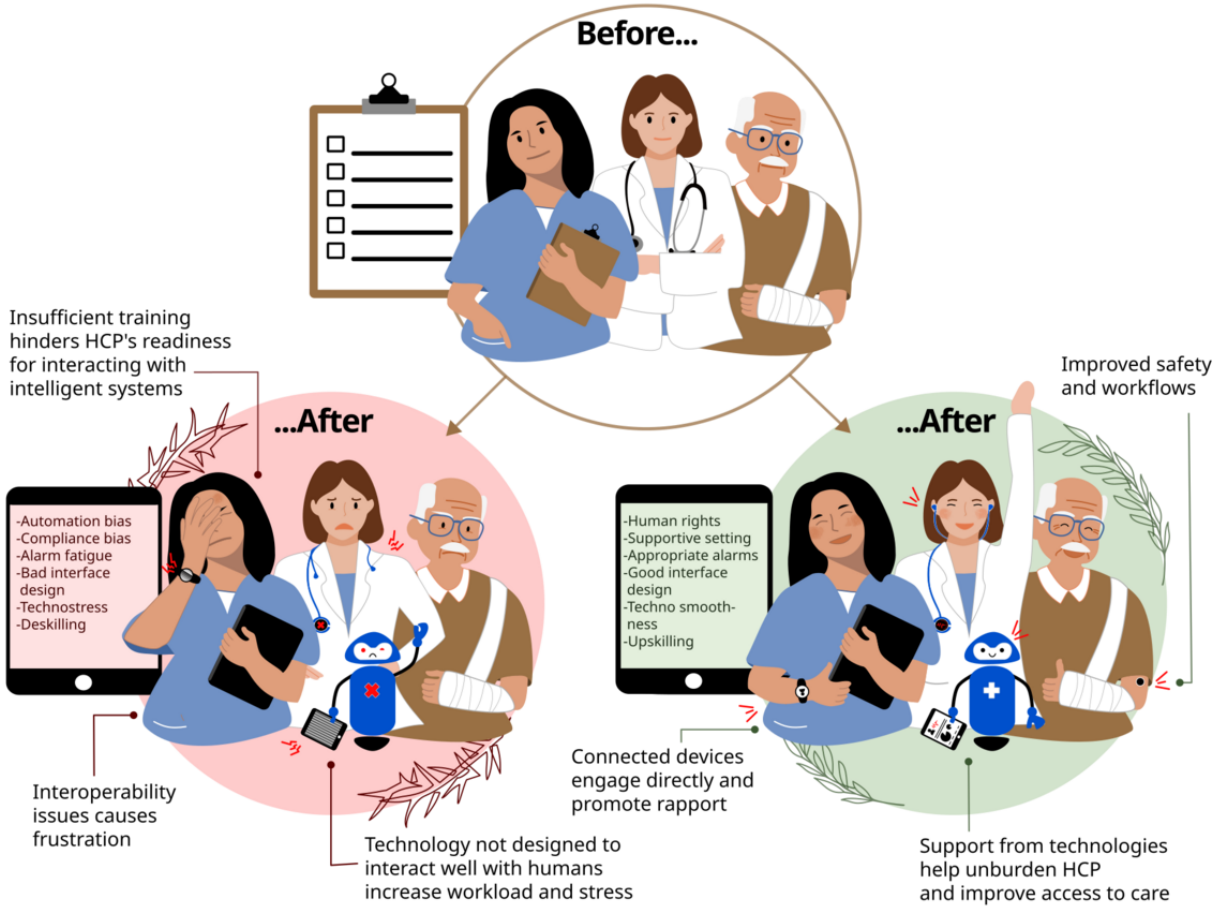

Improving Adoption by Co-Development Across the AI Life Cycle

Overview

The real-world challenges discussed underscore that successful AI development and implementation are less a technical task than a comprehensive change management process [57] that needs active participation, transparent governance, continuous feedback, and development beyond technical metrics, including systematic real-world evaluation of human-AI interaction, and a focus on non-technical design criteria such as usability, workflow fit, trust, and acceptance.

To bridge this gap, we propose EURAID (European Responsible AI Development), a practical framework of human-centric AI development and implementation in hospitals, which is cooperative and collaborative and based on shared goals in accordance with European values according to Article 2 of the Treaty on European Union (TEU; ie, human dignity [69,70], freedom [69,71], democracy [72], equality [69], rule of law [73], and human rights [69,74]) and European laws (Table 2).

XSL•FO
RenderX

Table 2. Regulations in the European Union and its member states that guide AI use in health care (nonexhaustive).

Regulation or law	Scope	Approach
Medical Device Regulation (MDR; 2017/745)	Governs medical devices (including digital systems) used for diagnostic or therapeutic purposes.	• The medical devices' intended purpose defines the associated performance claims, which must be substantiated through clinical evaluation. • GSPRs[a] must be met, including structured risk management (ISO[b] 14971:2019), a certified QMS[c] (ISO 13485:2016), usability engineering (IEC[d] 62366-1:2015+A1:2020), and a planned and documented development process (IEC 62304:2006+A1:2015), depending on the respective product category.
Artificial Intelligence Act (AI Act; 2024/1689)	Governs the development, market entry, and use of AI[e] systems.	• Classifies high-risk AI systems (including AI-enabled medical devices, class IIa+) and GPAI[f] (that can perform a wide range of tasks, not limited to one clear intended purpose) and LLM[g] models, depending on both the function performed and the systems' intended purpose. • Additional transparency obligations apply for certain systems such as emotion recognition, biometric categorization, and interactive or generative AI.
EU Occupational Safety and Health Directive (89/391/EEC 1989) and national laws	Ensures workers' health and safety.	• Systematic risk assessments and preventive measures. • Worker consultation and participation.
Professional regulations (eg, Federal Medical Code for doctors) and labor laws (eg, German Works Constitution Acts)	Defines autonomy and participation rights of HCPs[h].	• Protection of professional independence in decision-making. • Co-determination rights of employee representatives, for example, when adopting systems that monitor behavior and/or performance.

[a]GSPR: general safety and performance requirements.

[b]ISO: International Organization for Standardization.

[c]QMS: quality management system.

[d]IEC: International Electrotechnical Commission.

[e]AI: artificial intelligence.

[f]GPAI: general-purpose artificial intelligence.

[g]LLM: large language model.

[h]HCP: health care professional.

In detail, we describe the appropriate stakeholder circle, the approaches needed for implementing new and highly integrated, localized, and adaptive AI models, and optimal techniques for building consent. While this paper emphasizes that AI systems are increasingly evolving into system-level tools with broad intended purposes, it is nevertheless valuable to explore the development of a narrow-purpose, limited-functionality tool as a simple entry point in the consideration of AI system implementation. This example serves as a foundation for discussing the broader challenges associated with a broad intended purpose and multiagent AI systems. We describe the co-development of an "in-house" AI system [75] that is developed within a health institution to address specific needs [76,77], rather than the implementation of an externally developed "off the shelf" AI system, as this allows more aspects of the collaborative process to be described.

This pragmatic approach was developed in part through in-depth individual consultation and 4 flexible multistakeholder workshops, which are described in more detail in Table 3. By bringing together all the relevant players in the health care ecosystem, we were able to set agreed goals and processes for the development, integration, use, and oversight of health AI. These insights from the workshops informed aspects of the development of the overall framework presented in this viewpoint, alongside the perspective of the authors.

XSL•FO
RenderX

Table 3. Methodological design of the stakeholder workshops. Since workshops are platforms to jointly identify and explore complex domains, and help to gain relevant insights beyond the individual stakeholders' scope of knowledge [78], they offer a valuable basis for a framework that has consensus-building at its core.

Aspect	Approach
Stakeholder definition	• An individual or group who is affected by or can influence the digital transformation in hospitals, particularly with a focus on AIa-enabled systems.
Identification of stakeholders	• Stakeholders were identified using the 7Ps framework [79], which serves as a guide for engaging diverse and relevant interest groups. We modified the categories and definitions of the 7Ps according to our context: • Patients and the public: As this is not a traditional patient-focused study, but rather a practical, expert-driven implementation guide for human-centric digital transformation in a hospital setting, stakeholders were viewed both as domain experts and as potential patients. Additionally, we had feedback from two different international patient representative organizations. • Providers: Individuals who provide care to patients and offer relevant insights from their clinical work were included. The selected clinicians represent various medical fields, including psychology, and are balanced in their seniority and professional position. • Purchasers: Since digital transformation must be financed individually by each hospital, we included stakeholders responsible for the high-level management of digital transformation in large hospitals who manage strategic decisions about cost underwriting based on a specific internal budget. • Payers: In Germany, digital hospital transformation is supported through federal programs. Therefore, we involved stakeholders working at the Federal Ministry of Health and stakeholders who are actively translating those programs into clinical practice. Additionally, we included employees of insurance companies, as insurers play, in general, a critical role in creating patient-centric digital ecosystems and in incentivizing digital health solutions. • Policy makers: Policy makers and supporters of digital transformation in hospitals were included, particularly those who support a human-centric approach while ensuring the rights of HCPs[b] and patients are in place, spanning stakeholders from labor unions to occupational health and safety experts, as well as relevant legal and ethical perspectives. • Product makers: As EURAID[c] highlights the need for a well-balanced stakeholder group developing and implementing AI in health care, the stakeholders representing the "in-house" manufacturers are in their profession AI system developers, psychologists and human-centered AI development professionals, as well as experts in medical device regulation, quality and clinical risk management, medical informatics, and in occupational health and safety at work. • Principal investigators: The researchers included were from a background of clinical AI, medical device regulation, nursing science, medical informatics, digital health, patient safety, psychology, and ethics.
Stakeholder engagement	• Objectives: The goal of stakeholder engagement was to achieve a common agreement on the theme by balancing the differences of individual viewpoints (eg, between calls for greater space for innovation or rather tighter regulation), and developing a framework that all stakeholders agree with. • Methods: Stakeholders were engaged through participatory workshops (three dealt with relevant aspects EURAID should focus on and were initiated by the German Federal Institute for Occupational Safety and Health (BAuA), in 2024 and 2025, with 25, 24, and 17 participants respectively; and one dealt with aspects of HCP integration and current health system problems AI-enabled transformation might solve (Table 1) and was organized by the Else Kröner Fresenius Center for Digital Health in February 2025, with 5 participants). The participating stakeholders spanned the whole 7P categories. Based on this data and a critical review of the literature exploring existing frameworks and gaps, AS and SG developed the concept for the paper and wrote the first draft of EURAID. The stakeholders reviewed the paper, validated its content, and provided further expert insights during a 4-month iterative consensus process.

[a]AI: artificial intelligence.

[b]HCP: health care professional.

[c]EURAID: European Responsible AI Development.

Step 1: Comprehensive and Inclusive Stakeholder Involvement to Build Consensus and Ensure Goal-Oriented Development and Implementation

The selection and active participation of stakeholders and the building of consensus are critical to the success of the AI system development and implementation. The stakeholders involved should be balanced in disciplinarity (clinical, technical, and administrative [80]) and operational responsibilities (professional positions, employee representatives, etc) as well as in age and gender. In Table 4, we highlight the key stakeholders involved, and in particular their role in the implementation process. Each stakeholder is selected for their contribution, ranging from strategic aspects (management board) to safety perspectives (employee representatives, quality management, clinical experts, and users), and data-driven issues (AI system developer, data scientists, and IT and regulatory specialists). In principle, stakeholders in their profession are not mutually exclusive; instead, one could fulfill several roles simultaneously.

XSL•FO
RenderX

Table 4. Key stakeholders and their roles in shaping and guiding AI development and implementation in health care. Each stakeholder is selected for their contribution to the process and expertise.

Stakeholder	Important areas of stakeholder involvement and key aspects they can address
Management Board 	The management board sets an overall vision and strategy, leading change management [57,81], and providing investment [82] in staff, hardware, and supporting infrastructure [17]. They foster an institutional culture that tolerates experimentation (and failure) [80], serve as the institution's most credible communicator (ensuring transparency around risks and benefits), and manage external relationships by forging alliances with industry innovators, researchers, professional associations, and policymakers.
Employee Representatives 	The foremost priority of employee representatives is to defend and improve working conditions, including occupational safety, workload management, and job security. Although large-scale staff redundancies are unlikely consequences of the near-term implementation of AI[a] in hospital health care systems, which are operating against a backdrop of large staff shortage [83,84], anxiety about automation and transformation of job roles is real [85]. Employee representatives ensure that AI is implemented in a way that eases staff workload and safeguards their well-being and autonomy. In the mid- to long-term, they also negotiate fair compensation policies [86] and career development frameworks that reflect changing roles and skills in an increasingly digital workplace.
AI-System Owner and team[b] 	The AI system owner holds primary accountability for the system's performance, safety, and operational impact. They lead the project and ensure alignment with strategic goals and regulatory compliance, while understanding the users' "pain points" both from a clinical and organizational perspective. Their responsibilities include bridging the communication gap between technical and nontechnical language, balancing different perspectives, and developing educational approaches [66] to increase user adoption.
Clinical Experts 	Clinical experts identify clinical relevance and utility, which are interpreted and transcribed into a specific scope (intended purpose that specifies clinical indication and initial target group). They provide crucial input to clinical validation and safety, ensuring the AI system integrates effectively into workflows, as well as initiate, oversee, and conduct clinical trial–based AI studies.
AI System Developer 	To design and develop machine learning algorithms tailored to specific needs, the AI system developer must integrate and harmonize data from different sources [15]. They also validate the AI model and detect and mitigate model bias to ensure the systems are fair, scalable, adaptable, and verifiable in real-world environments [87].
Users (HCP[c] or patient) 	Users with varying levels of digital literacy [57,88] provide real-world, iterative feedback on the system's usability, workflow integration, and perceived value. They often become multipliers for AI adoption, and by their active participation in co-designing educational materials [66], they support evolving digital competence among peers.
Data Scientist 	The data scientist safeguards the quality of the data foundation on which the AI system depends during preparation, collection, and checking of the data, for example, by keeping data collection protocols and detecting data imbalance, bias, or outliers across age, sex, gender, race, or ethnicity to prevent disparities and underperformance before they arise [89].
IT Specialists 	This role provides the essential technical infrastructure and ensures secure, seamless integration with existing systems, like EHR[d] platforms or laboratory systems, requiring technical, syntactic, semantic, and organizational interoperability [15,90]. Beyond integration, they build and maintain structures for data security, access control, and real-time support, and establish data backup and disaster-recovery systems.
Regulatory Specialists 	Regulatory Specialists provide expertise in medical device and AI law, data protection, and human rights. They ensure regulation standards (like the MDR[e] and the AI Regulation) are met throughout the product lifecycle, which is essential to mitigate legal risks and prevent potential breaches.
Notified Body 	The role of the Notified Body is to assess whether medical devices meet European legislation, like MDR. This includes determining the correct classification, evaluating legal compliance, and reviewing technical documentation [75,91]. The Notified Body only has a direct role where a CE[f]-mark is sought for medium or high-risk AI systems.
Quality Management 	Quality management ensures continuous patient safety by monitoring and measuring performance, outcomes, and the integrity of clinical workflows [87]. They establish comprehensive risk management systems (eg, handling device failures or malfunctions) [87] and drive standardization. This role also promotes safe system use by co-designing educational programs [66] for both HCP and patients.

[a]AI: artificial intelligence.

[b]Role of the stakeholders whose input is coordinated through the AI-System Owner.

[c]HCP: health care professional.

XSL•FO
RenderX

[d]EHR: electronic health record.

[e]MDR: Medical Device Regulation (2017/745).

[f]CE: Conformité Européenne.

An interactive environment, with all critical stakeholder groups adequately represented, enables and encourages the integration of stakeholder insights and experiential learnings, while promoting careful consideration of how AI systems are best built to be suited to clinical workflows, as well as where existing workflows may need to be modified to adapt to the AI system. This does not mean that every stakeholder group is involved in every decision and has an equal say in the progress of digitalization. Creating this impression could lead to disillusionment and eroded trust in digitalization, and would probably slow down the whole process. Each stakeholder group is involved in some part of the process, with their precise stages of involvement and roles depending on their potential contribution to the process, and it is essential that each stakeholder is aware of the degree of their involvement.

A crucial success factor alongside the development and implementation is the role of the "product owner," who takes the coordinating lead. As in-house development in health care institutions often does not have a commercial development focus, we use the term "AI-System Owner" to denote the "product owner." Although the title may vary by organization, this role usually combines both the entire lifecycle product ownership responsibilities and the domain expertise in health care and AI. The absence of a single person taking responsibility for the development and performance of the system will generally result in a range of negative consequences, such as poor stakeholder communication, a lack of clear vision, scope, and prioritization, and other issues, as real-world examples [92] have shown. We therefore highlight the AI-System Owner as a central stakeholder leading a team of other stakeholders (Figure 2).

Figure 2. The (ongoing) product development in a dynamic team led by the AI-System Owner. The AI-System Owner fulfils a crucial role as he is leading a core team of relevant stakeholders during the process of development and implementation. In a hospital setting, team members will often fulfill several roles simultaneously. AI: artificial intelligence; HCP: health care professional.

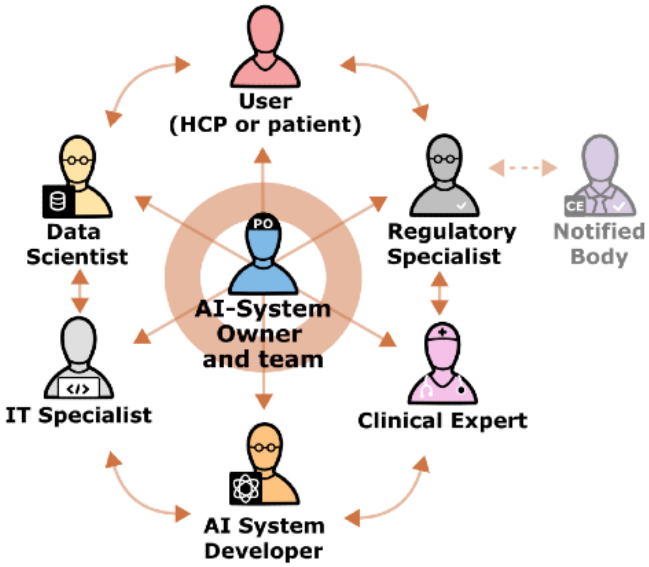

Step 2: Agreement on the Overall Goals and "Device" Purpose

The collaborative and effective implementation of an AI system into clinical workflows starts with a collective agreement on the goals of the implementation, for example, using methods such as SMART (specific, measurable, attainable, relevant, and time-bound), particularly the specific user (generally an HCP or patient) whose needs the system is intended to address. These identified needs are then interpreted and transcribed into a specific scope of the device, known as the "Intended Purpose," which specifies the clinical indication, how the system addresses this clinical indication, and the (initial) target group needs.

Although the regulations for AI-system design and implementation do not formally require the direct involvement of any other health care system actors than the "user" of the AI and its "deployer" (in a broad sense), we argue that the sustainable and beneficial implementation of AI systems needs early and proportional agreement on goals and input from all stakeholders. This includes discussion between the management board, employee representatives, quality management, and the AI-System Owner and their team (Figure 3). Later product development steps require feedback between the AI-System Owner team (including clinical experts and the users of the system), and selected stakeholders (as shown in Figure 3), with management "checkpoints" periodically to ensure that the development of the AI-system is following the initially agreed plan for the AI system. Given the complexity of multistakeholder involvement, it is useful to have a set of rules for working together at the beginning, and to repeatedly build consent along the AI development life cycle, for which we highlight techniques in Figure 3.

XSL•FO
RenderX

Figure 3. Stakeholder interaction and consent-building across the AI life cycle. The figure describes the co-development of an "in-house" AI system, ie, one that is developed within a health care organization to address specific needs. During the development phases, which build on each other and can be repeated, different groups of stakeholders interact to improve the AI system by providing feedback and optimizing the system's adaptation to health care professional workflows. Building consent with a range of different stakeholders with varying levels of experience and backgrounds is not easy. We therefore highlight techniques for building consent at each stage of development to ensure an efficient and safe process that is in line with European values and regulations. AI: artificial intelligence; MDR: Medical Device Regulation; ML: machine learning; PCCP: predetermined change control plan; RWP: real-world performance; SaMD: software as a medical device.

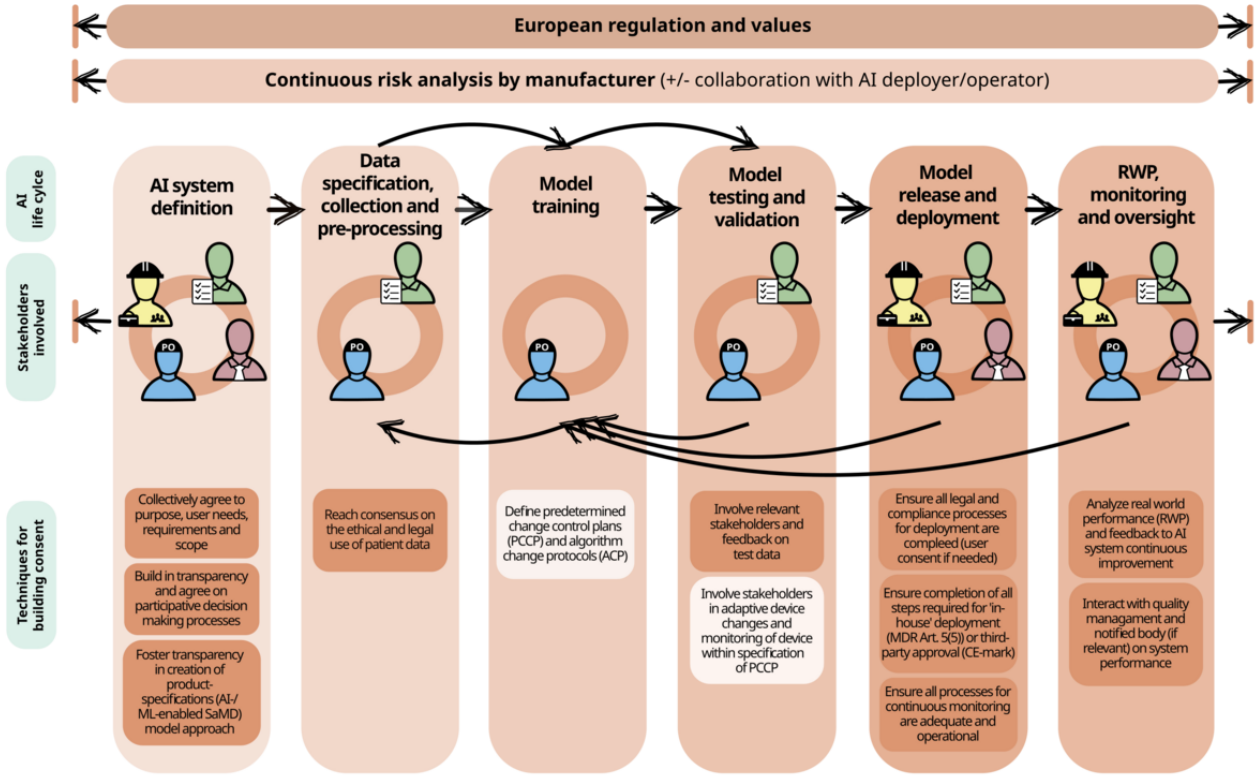

Step 3: AI System Development "In-House"

While generally medical devices must undergo a conformity assessment procedure and must be marked with a CE (Conformité Européenne)-mark before being used, the European Union (EU) exempts certain devices from this general obligation and allows individual health institutions to develop and use "in-house" medical AI systems involved in the diagnosis or therapy of disease without the obligation to conduct a conformity assessment procedure, as long as safety standards and those for quality management are in place. Based on Article 5(5) of the EU Medical Device Regulation (MDR; 2017/745) [75], this exemption applies only for in-house use on a nonindustrial scale and if the needs of the targeted patient groups cannot be met through available and equivalent devices on the market [75,93]. Also covered is the in-house combination or modification of existing systems or devices [93,94]. For example, in Table 5, we have outlined 3 practical examples of AI systems, which have been developed in-house in a German hospital setting, each of them with a unique intended purpose, clinical indication, and target group. We highlight for each the technical approach used as well as the stakeholders included during development and potential prospective trial designs.

XSL•FO
RenderX

Table 5. Practical examples of AI applications developed in-house and their stakeholder integration. The AI use cases presented originate from the SmartHospital.NRW [95] research project, funded by the Federal State of North Rhine-Westphalia, Germany. The project is limited to research and development activities; therefore, the use cases are confined to the development stage. Clinical testing and product commercialization are explicitly beyond the project's scope.

Use case	Automated discharge summary	AI[a]-powered voice assistant for bedside patient support	AI-supported prevention of adverse events
Intended purpose	Automates and optimizes the creation of discharge letters within hospital workflows to reduce clinician workload and improve communication regarding patient care.	Enables patients at the bedside to interact via natural speech, facilitating access to medication schedules, personal calendars, diary management, and support to overcome language barriers through oral translation and simplified language.	Focuses on early and reliable detection of nursing-relevant risks by enhancing existing risk models based on structured nursing assessments and integrating LLMs[b] to analyze clinical progress notes and identify patient-specific risk factors.
Clinical indication	Addresses the challenge of time-intensive medical documentation, particularly discharge summaries following inpatient stays.	Designed for patients requiring accessible communication support, especially those experiencing language barriers, vision impairments, or limited mobility, while promoting autonomy without providing direct medical advice.	Designed to support systematic, early identification of nursing-related risks, including falls, pressure ulcers, and malnutrition, augmenting safety and enabling individualized care planning.
Target group	Primarily, hospital physicians with indirect patient benefits, such as improved continuity of care and efficient information transfer to general practitioners.	Hospitalized patients who require assistance in accessing information and communicating effectively.	Nursing staff responsible for patient care and hospitalized patients are actively involved in care processes.
Technical approach	Uses generative AI language models interfaced with hospital information systems to autonomously extract structured clinical data and generate contextually relevant text suggestions for documentation.	Uses on-premises LLMs within dedicated patient devices; enables localized processing of voice input streams independent of hospital system integration, thereby preserving data sovereignty.	Integrates structured clinical data, unstructured data derived from speech-to-text conversion of nursing assessments, and patient-reported outcomes to facilitate comprehensive risk detection.
Stakeholders included during development	Management Board, AI System Developer, AI-System Owner, IT Specialists, Clinical Experts, and Users.	Management Board, AI System Developer, IT Specialists, Clinical Experts, and Users.	Management Board, AI System Developer, AI-System Owner, IT Specialists, Clinical Experts, and Users.
Experience of development	Developed iteratively as a prototype, validated with real clinical data, while ensuring compliance with regulatory, privacy, and interoperability standards.	Followed an iterative development approach with thorough curation of informational content; faced technical challenges such as limited server access before full deployment of open-source models.	Development prioritized screening instruments to assess signs and symptoms of nursing care, optimization of AI risk detection models, and ensuring data privacy using pseudonymization and anonymization techniques.
Potential prospective trial designs	Cluster-randomized controlled trial at the ward level, comparing standard discharge processes versus AI-assisted summaries. Primary endpoints: clinician documentation time and report quality (as judged by independent review).	Patient-level crossover trial with and without AI voice assistant support. Main outcomes: patient autonomy, effectiveness of information access, and user satisfaction, controlling for intrapatient variability.	Pragmatic controlled trial in clinical wards comparing standard care with and without AI-based risk detection algorithms. Outcomes: incidence of adverse events (falls, pressure ulcers, and malnutrition), timeliness of risk identification, and changes in clinical workflow.

[a]AI: artificial intelligence.

[b]LLM: large language model.

In contrast to commercial deployments, in-house systems offer a distinctive opportunity for embedding participatory ethics, iterative design cycles, and real-world validation and feedback loops directly into the lifecycle of medical AI. This allows the creation of a highly customized solution that fits in location-specific clinical workflows and staff practices, which can be extended to multiple systems within the same platform and institution [76]. Moreover, a key advantage is the use of the hospital's own data; however, this requires a well-developed data infrastructure and processes for obtaining patient consent. Considerations include interoperability and data preparation, such as labeling (although label-free approaches are becoming more common), structuring, and collection (requirements also under the AI Regulation), in order to know which data can be used for a specific solution.

Step 4: AI-System Testing, Validation, and Clinical Evaluation

Health care AI demands rigorous, multidimensional evaluation that must encompass not only technical performance, but also clinical integration, and verify safety, usability, ethical robustness, and regulatory compliance.

XSL•FO
RenderX

Independent assessment of device performance can be generated through statistically sound test plans, which generate information separate from the training data set [96]. Since validation in real-world settings is still a bottleneck [97], prospective, noninterventional silent trials [98,99] (where AI is tested within the clinical pathway in real time without affecting patients) can enhance transparency and facilitate informed deployment decisions. For large language models (LLMs) and, in particular, adaptive AI models that evolve over time, continuous validation frameworks are needed [100]. Recent studies have highlighted, substantial challenges to the reliability and safety of LLMs in health care persist, including hallucinations [101], metacognitive deficiencies [102], vulnerability to bias [103] and data-poisoning [104], and problems in integration in existing workflows [105], making single evaluation dimensions insufficient. Therefore, multidimensional methods could help to operationalize feasibility, score diagnostic accuracy or unsafe recommendations, and detect bias and usability issues. Examples are "QUEST" [106] to score outputs, or agentic-based simulations such as "CRAFT-MD" [107] for clinical workflow evaluation. Alignment with international AI standards (eg, ISO/IEC [International Organization for Standardization/ International Electrotechnical Commission] 42001:2023 [108], FG-AI4H [Focus Group on AI for Health] clinical evaluation framework [109]) further strengthens interoperability and safety.

Beyond objective data and algorithm quality, subjective feedback from users is essential [57,110]. Evaluations should capture how AI systems integrate into existing workflows and routines, their ease of use, and their perceived performance and interface design. Researchers highlighted several approaches for evaluation, such as through integrated feedback systems [110,111] or through organizational internalization by creating an "AI-QI"-unit responsible for quality improvement and assurance [87], interacting as a "glue" between different entities.

Evaluation should follow a risk-tiered approach that links the level of regulatory and ethical scrutiny to the severity of the health decision involved (Figure 4). For instance, AI systems used for administrative optimization or appointment scheduling may require a lower level of risk mitigation, while those supporting diagnostic or therapeutic decisions demand significantly higher safeguards. This tiering can draw on the EU AI Act's risk classes and MDR risk classifications, and should be developed in consensus with relevant stakeholders, including clinical risk management and regulatory specialists.

XSL•FO
RenderX

Figure 4. Risk-based tiering of safeguards. With a proportional approach to regulatory and ethical safeguards aligned with the severity of the health decisions affected by an AI system, this provides a useful link between risk classification (eg, under Medical Device Regulation or the EU AI Act) and the required level of human oversight, transparency, and stakeholder involvement. AI: artificial intelligence; EHR: electronic health record; EU: European Union; LLM: large language model.

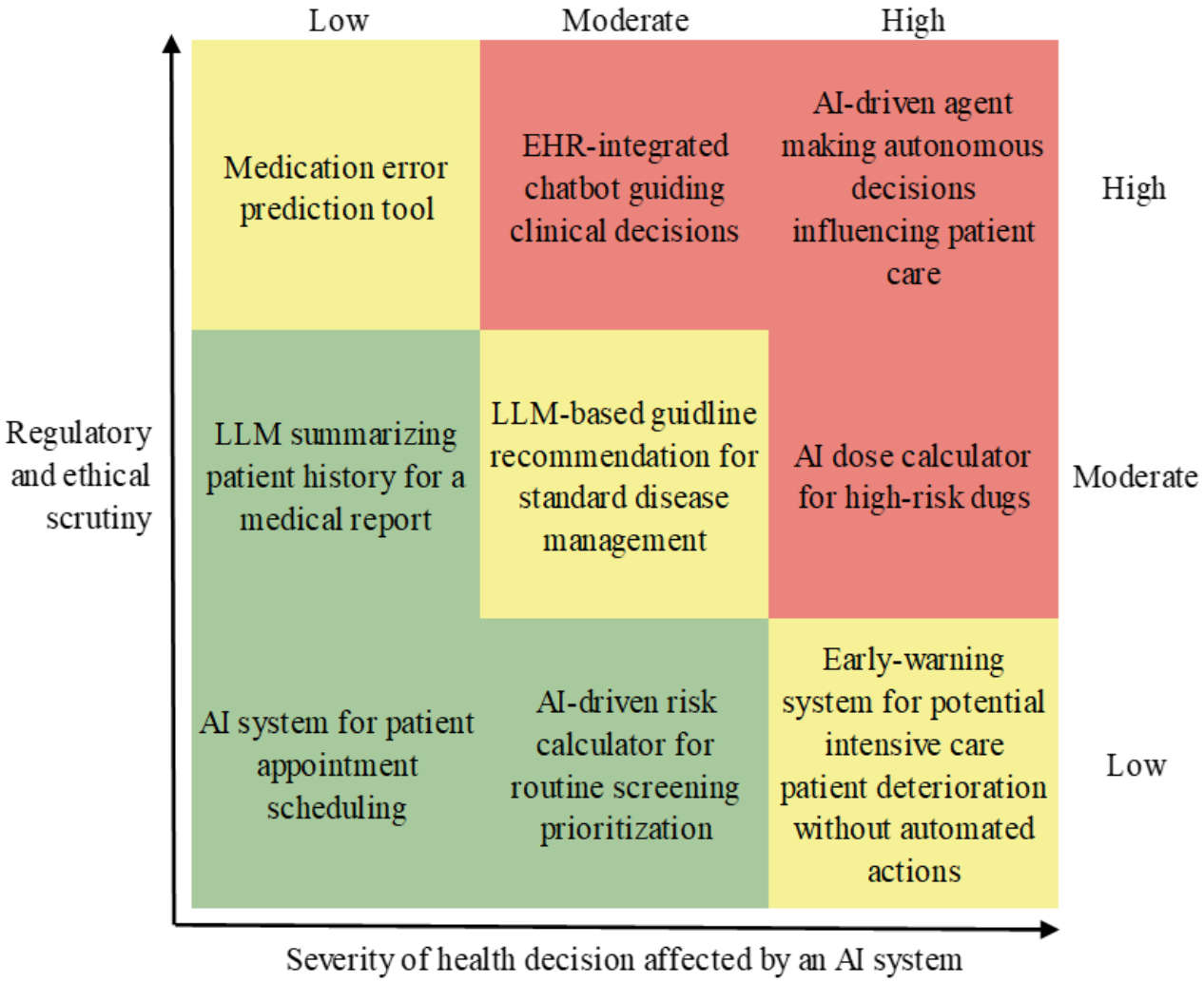

To ensure that the AI system is compatible with European values, ethics-based auditing frameworks like capAI, grounded in the EU AI Act, can guide risk identification in each phase of the AI lifecycle from an ethical point of view [112]. The integration of tools like the self-assessment list for trustworthy AI (ALTAI) [113], developed by the EU High-Level Expert Group on AI, into ethics-based auditing of AI systems can further support responsible usage of AI and foster user trust. Yet, ethical guidelines are just that: guidelines. They rarely or incompletely answer concrete ethical questions regarding the use of an AI system in a specific situation, such as the question of specific moral responsibility if mistakes of AI systems lead to patient harm. This is a highly discussed topic in ethics [114] and becomes even more severe in the context of black-box problems, eventually leading to moral responsibility gaps [115]. Other still unsolved ethical questions occur, for example, regarding data ownership in the context of the principle of beneficence (ie, promoting others' benefit and preventing harm [116,117]) and informed consent [118] or anthropomorphization of AI [119]. Therefore, embedding ethical points of view into the whole life cycle of AI is necessary [120].

Step 5: Development and Deployment of Training Approaches

The successful adoption of AI by hospital employees correlates with continuous development and training [88]. Although training is also a requirement of the EU AI Act [121], it is of note that only 24% of the health care institutions provide AI training programs and workshops [122]. This underscores a gap in education and certification, leaving clinicians without the necessary tools to harness the full potential of AI. However, there are various ways to support confidence in AI technologies among HCPs. For example, (1) by investing in comprehensive training programs that help to gain necessary skills [88] while also extending existing programs with AI literacy, or (2) by developing and provisioning resources and mechanisms to build and strengthen connections among peers and innovators to share their AI-related knowledge and experiences [80]. And more importantly, AI training should be a fixed part of any professional education and competency assessment, as well as included in further training (eg, through integration into Continual Medical Education programs) [123] to build confidence in its use among the next generation of HCP and achieve a symbiotic relationship between humans and AI [124].

XSL•FO
RenderX

In order to build AI literacy among HCP in a safe and controlled environment, training methods such as simulation-based modules [125,126] (ie, practice in realistic settings [125,127]), case-based exercises [128], and interactive workshops [129] can help to explore tools repeatedly without risking patient safety while facilitating experimental learning. Another method of providing HCPs with hands-on experience using AI tools in a controlled environment is to conduct a pilot phase, during which AI is tested by selected clinical users in a narrow area of practice, or shadow deployment, in which AI operates in shadow mode alongside clinicians in real time and is guided by predefined safety and workflow indicators [130]. This will also influence trust and adoption among users and foster psychological safety, since evidence from human-computer interaction research indicates that a positive attitude toward AI is not only a function of system transparency or explainability, but also depends on users' self-efficacy, previous experience, and the perceived fairness and predictability of the system [131]. With regard to content, it is important to define responsibilities within the company regarding who will take ownership of training the users in basic competencies of AI literacy. The AI-System Owner and his or her team would be the best fit, as they combine the entire use case-relevant expertise through different perspectives, ranging from clinical experts to system developers.

Training should foster understanding of AI systems and facilitate interaction and use of AI systems, and is relevant not just for direct users, but for all HCPs who will work alongside care systems influenced by AI (Figure 5) [15,66]. Key competencies are a basic understanding of when and how to use AI, knowledge about the use of the systems' elements, the ability to make informed decisions based on a risk-benefit analysis, the awareness of legal and ethical considerations, and, to adapt to new tools and applications [123,132]. Components of health care AI training that are generic do not need to be developed de novo by the health institution. However, specific training directly related to the AI-system to be deployed will generally be required, and it is often necessary to provide ongoing training which takes account of the learning curve of the HCP in the use of the AI, emergent problems such as automation bias [133] and deskilling [134], and changes and further development of the AI-systems.

Figure 5. The learning curve of the health care professional (HCP) in the use of artificial intelligence (AI) systems in health care. After training in the basic AI principles and their use, as well as health care–specific guidelines for AI integration, on the first day of the system's clinical use, the HCP should be trained in the operational knowledge of the system being deployed. The HCP will then develop their skills through experience in their use.

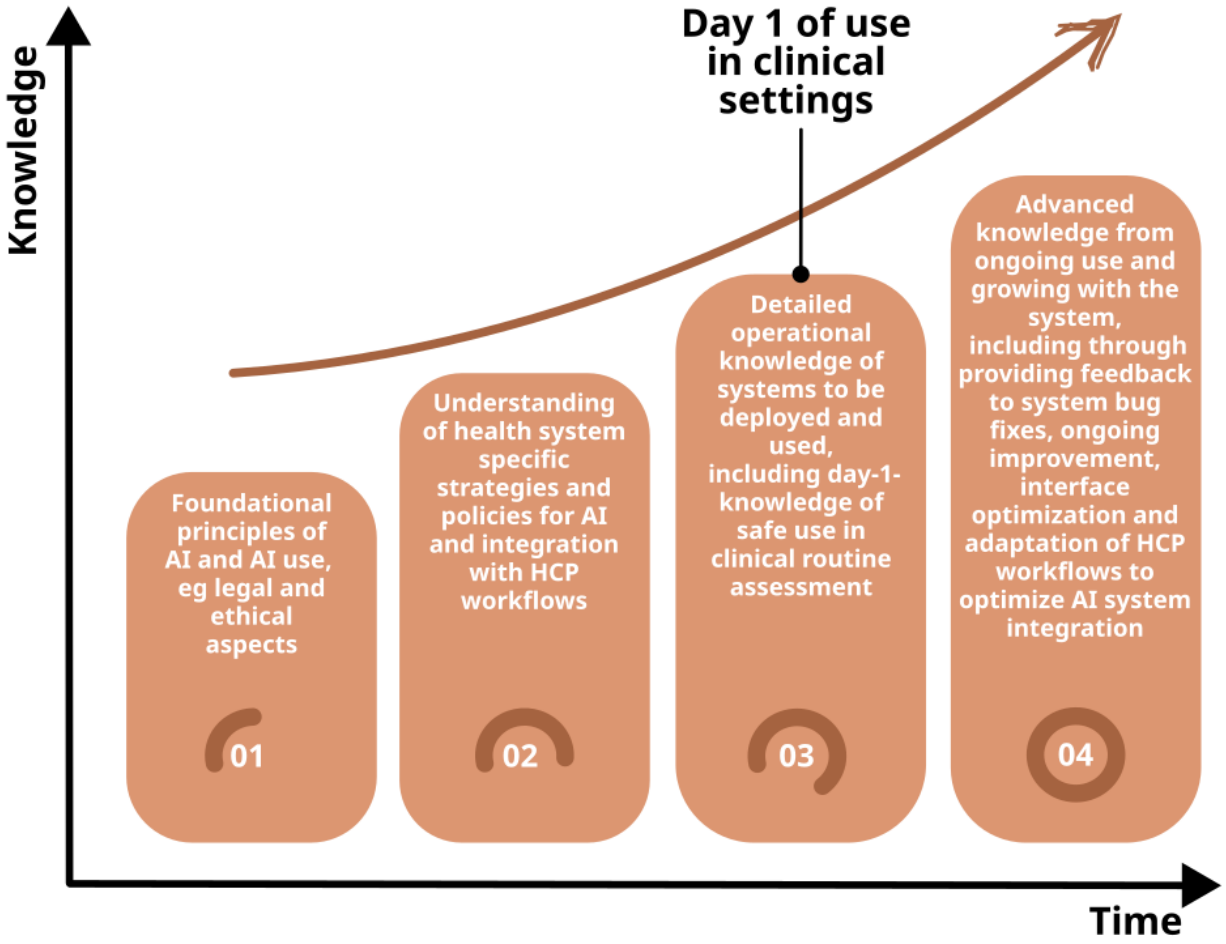

XSL•FO
RenderX

Step 6: AI-System Deployment, Real-World Performance Monitoring, and Later Decommissioning

After model creation and testing, the goal is to place the system in real-world clinical settings to improve patient care and outcomes [135] according to the previously defined overall goals and device purpose. This needs transparency, and compliance with legal and ethical processes (user consent), as well as the completion of all steps required for the exemption to conduct a conformity assessment under in-house deployment ("MDR Article 5(5)") or third-party approval (CE-mark). Therefore, looping in all stakeholders is needed to collaboratively address associated challenges. A key role is played by the management board and AI-System Owner to provide a clear external and internal communication that signals the prioritization of human well-being during the whole process, and users as multipliers to promote trust for broad widespread acceptance and use.

Involving all stakeholders also applies to monitoring and oversight of real-world performance, as it needs constant feedback from different perspectives to improve system performance and data-related processes. The goal of monitoring is to raise an alarm when unintended or special cases occur [87], which emphasizes the importance of finding solutions through collaboration and collective intelligence. The "AI-QI" unit described above could consolidate and strengthen the established stakeholder structure within the company long term. In addition, algorithmic audits can serve as a framework for continuously monitoring AI systems and understanding errors, how and why these adverse events occurred, while anticipating their potential consequences [136]. Real-world performance monitoring must adequately account for model drift (degradation of AI system performance over time) due to changes in external factors such as patient populations, data collection, or medical practice [137].

Running a "legacy system" usually means facing layers of technical debt, which slows down development and complicates maintenance, and leads to several risks, such as the technology becoming less reliable and decreasing in performance, or exposing systems to vulnerabilities such as cyberattacks. However, decommissioning can be an option to abstract and secure data in a newer system [138]. This process needs to be carried out by IT and regulatory specialists, as well as data scientists and quality management, in consultation with users, the management board, and employee representatives, and notified bodies where required.

Special Considerations for Adaptive, "Agentic" and "off-the-Shelf" AI Systems

Some recent AI approaches are developed so that they learn and adapt from data and feedback from the real world, allowing them to change continuously without explicit interventions from the developer [139,140]. Ensuring such systems are safe, effective, and of high quality while being flexible requires a more interactive and participatory approach than traditional systems that follow static and predefined rules. This is especially true when self-learning systems are combined with agentic AI systems that are able to handle multilevel tasks, coordinate tools, centralize human communication, and basically act as health care teammates [26-29]. Autonomous AI systems and LLM-enabled clinical decision systems have already been approved in Europe [30,141,142]. As the approval and use increase, and as these systems continuously encounter new settings and tasks, it is essential to define clear boundaries, controlled environments with clinician oversight [27], ongoing auditing [26], and adequate training capacities for HCPs [27]. As broad models may be applied across multiple hospital departments and clinical contexts (eg, simultaneously used in an emergency department and psychiatry clinic) with dynamic or variable workflow integration, transparent communication, and iterative feedback across stakeholders (as presented in this paper) are also critical to ensure adaptability and to address the more complex ethical, legal, and social implications.

For off-the-shelf AI systems provided by external companies, the interaction between stakeholders should be focused on integration, compliance, and validation to meet operational and regulatory needs. These systems may limit the level of innovation achievable (no bottom-up activism from internal users and developers to continually contribute improvements and features that better meet unique requirements) and may lead to trust issues due to less transparency in the handling of data and underlying algorithms [14], requiring proactive communication and change management. Responsibilities for monitoring and model updating, especially with proprietary algorithms, become more complex and need to be clarified between external collaborators and internal stakeholders [87]. Platforms for delivering off-the-shelf AI systems now allow the co-hosting of in-house developed AI models, alongside the CE-marked models, enabling both approaches to coexist, and making clear the need and possibilities for the co-design, embedding, and co-implementation of commercial and in-house approaches [143].

Discussion

Studies show a persistent gap between research and clinical implementation [144,145], with medical AI adoption still very slow [144,146] and limited to a few use cases [147]. Reasons include the difficulty of aligning diverse stakeholder perspectives within complex health care systems, the rigidity of regulatory frameworks, and the limited consideration of design approaches of work and organizational psychology [148]. As a result, achieving both technological effectiveness, in the sense of medical accuracy and system performance, and user acceptance among HCP and patients is often perceived as conflicting goals.

A balance is therefore needed between ensuring safety and enabling innovation [149]. EURAID finds this "sweet spot," accelerating digital transformation in a human-centric way. Unlike existing frameworks, which focus narrowly on user perspectives [80,150,151], isolated implementation aspects [150,152-155] (such as evaluation, safety, or ethics), serve as decision support tool for choosing the most fitting available AI solution [156], or have a limited clinical scope [157-160], EURAID explicitly maps all key stakeholders across the AI development life cycle, clarifies their roles and key aspects they

XSL•FO
RenderX

can address (Table 4) in co-creating, guiding, and governing "in-house" AI development and deployment. It also details stakeholder roles in real use cases, and methods for achieving iterative consensus at each development stage across disciplines that reflect shared goals in alignment with European values, and strengthening the understanding of training methods, content, and key competencies.

However, EURAID has some limitations. The resources or specialized staff needed for iterative development and testing are more limited in smaller hospitals, necessitating concentrating multiple roles on fewer people, which can lead to a shortage of expertise, but, on the other hand, may also speed up processes. Although our approach can likely better address creative problem-solving, traditional, rigid, and hierarchical structures common in health care may hinder stakeholder selection based on their contributions and expertise rather than their positions and level of seniority. Although "in-house" AI devices may not require CE marking, they are not exempt from regulation and have legal liability implications. Health institutions must comply with a number of obligations that may discourage them from doing it at all, which slows down both innovation and digitalization. A practical solution is to designate key staff for legal or ethical liaison roles or establish a multidisciplinary AI advisory board and data governance council within the institution to ensure compliance and continuity.

Conclusions

EURAID is a pragmatic, solution-oriented framework, compatible with European values and regulations, and ensures that barriers to "in-house" AI development and implementation in hospitals are acknowledged early and resolved through collaborative problem-solving. The underlying principle is that the likely future of medicine, driven by integrated, localized, and adaptive AI technologies, will need all critical stakeholders (which we portray individually in this paper) adequately represented, and their various perspectives embedded in the co-design, procurement, implementation, and oversight of AI systems, ensuring that digital transformation in health care truly benefits the people who will use them every day. Additionally, as AI systems used vary by type and clinical setting, we propose a risk-tiered approach that provides a useful link between risk classification and the required level of human oversight, transparency, and stakeholder involvement.

To translate EURAID into action, hospitals should begin by conducting internal readiness assessments, establishing cross-functional AI governance structures, and defining clear, role-specific responsibilities for ethical, legal, technical, and clinical oversight. Regulators and professional bodies should, in parallel, create structures that connect local innovation with next-generation European legislation, for governance that is as intelligent as the technology built.

Acknowledgments

We acknowledge the use of the ChatGPT language model (GPT-3.5, GPT-4, and GPT-5; OpenAI) for assisting in refining some text of this paper. Responsibility for the final manuscript lies entirely with the authors. The graphical elements in this paper were designed using Inkscape.

Funding

This work was supported by the European Commission under the Horizon Europe Program, as part of the project ASSESS-DHT (101137347) via funding to SG and RM. The views and opinions expressed herein are, however, the authors' responsibility only, and do not necessarily reflect those of the European Union, the United Kingdom, the European Health and Digital Executive Agency (HaDEA), UK Research and Innovation (UKRI), or the National Institute for Health and Care Excellence (NICE); the European Union, United Kingdom, and granting authorities cannot be held responsible for the views, opinions, and information contained herein.

Authors' Contributions

AS and SG developed the concept of the study. AS and SG wrote the first draft of the paper. AS, MEG, MHG, FJK, JNK, EK, TL, EL, ML, RM, HSM, JO, TR, UR, M Schneider, LS, HS, MLS, NS, M Sedlmayr, RS, BS, MKW, EW, KW, AD, and SG contributed to the writing, interpretation of the content, and editing of the paper, revising it critically for important intellectual content. AS, MEG, MHG, FJK, JNK, EK, TL, EL, ML, RM, HSM, JO, TR, UR, M Schneider, LS, HS, MLS, NS, M Sedlmayr, RS, BS, MKW, EW, KW, AD, and SG had final approval of the completed version. AS, MEG, MHG, FJK, JNK, EK, TL, EL, ML, RM, HSM, JO, TR, UR, M Schneider, LS, HS, MLS, NS, M Sedlmayr, RS, BS, MKW, EW, KW, AD, and SG take accountability for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

The authors expertise ranges from medical device regulation (AS, SG, RM, and ML), to high-level management of digital transformation in big hospitals (AD), and includes experts in quality and clinical risk management (RS, MEG, and SG), medical informatics (ML, M Schneider, SG, MLS, and M Sedlmayr) and occupational health and safety at work (UR, LS, and TL), as well as relevant insights from clinical experts and HCP (EW, JNK, HSM, NS, JO, and AD), AI system developers (JNK, NS, and JO) and expertise in psychology and human-centered AI development (MKW, KW, and TL). In addition, we included relevant legal (EK), ethical (EL), and federal policy (MLS) perspectives, health and social accident insurance companies (TL and M Schneider), labor unions (BS), and from academia (MHG, HS, TR, FJK, AS, RM, HSM, SG, JNK, and MKW).

XSL•FO
RenderX

Conflicts of Interest

SG declares a nonfinancial interest as an Advisory Group member of the EY-coordinated "Study on Regulatory Governance and Innovation in the field of Medical Devices" conducted on behalf of the Directorate-General for Health and Food Safety (SANTE) of the European Commission. He declares the following competing financial interests: SG has or has had consulting relationships with Una Health GmbH, Lindus Health Ltd, Flo Ltd, ICURA ApS, Rock Health Inc, Thymia Ltd, FORUM Institut für Management GmbH, High-Tech Gründerfonds Management GmbH, Directorate-General for Research and Innovation of the European Commission, and Ada Health GmbH, and holds share options in Ada Health GmbH. JNK declares consulting services for Bioptimus, France; Panakeia, UK; AstraZeneca, UK; and MultiplexDx, Slovakia. Furthermore, he holds shares in StratifAI, Germany, Synagen, Germany, and Ignition Lab, Germany; has received an institutional research grant from GSK; and has received honoraria from AstraZeneca, Bayer, Daiichi Sankyo, Eisai, Janssen, Merck, MSD, BMS, Roche, Pfizer, and Fresenius. JO has received travel grants from Abbott and research grants from German Heart Foundation (DSHF), German Center for Cardiovascular Research (DZHK), the University of Hamburg (UHH), and the German Federal Ministry of Education and Research (BMBF), and is co-founder and former managing director of IDM GmbH. MLS reports no conflicts of interest. The opinions expressed in this article are his own and do not necessarily reflect the views held by the German Federal Ministry of Health. None declared by the other authors.

References

1. Blum K. California nurses protest 'untested' AI as it proliferates in health care. Association of Health Care Journalists. URL: https://healthjournalism.org/blog/2024/08/california-nurses-protest-untested-ai-as-it-proliferates-in-health-care/ [accessed 2024-08-09]
2. Bruce G. Nurses protest AI at Kaiser Permanente. Becker's Health IT. URL: https://www.beckershospitalreview.com/healthcare-information-technology/nurses-protest-ai-at-kaiser-permanente/ [accessed 2024-04-22]
3. Blease CR, Locher C, Gaab J, Hägglund M, Mandl KD. Generative artificial intelligence in primary care: an online survey of UK general practitioners. BMJ Health Care Inform 2024;31(1):e101102 [FREE Full text] [doi: 10.1136/bmjhci-2024-101102] [Medline: 39288998]
4. Fernandopulle R. We must stop trying to deliver 21st-century care with a 19th-century delivery model. MedGenMed 2005;7(2):50 [FREE Full text] [Medline: 16369428]
5. Kennedy PJ. Our health system is built on an antiquated model of care. The Hill. 2020 Aug 25. URL: https://thehill.com/opinion/healthcare/513615-our-health-system-is-built-on-an-antiquated-model-of-care/ [accessed 2025-04-03]
6. Mele M. Antiquated methods put patients at risk. Beckers's Clinical Leadership. URL: https://www.beckershospitalreview.com/quality/antiquated-methods-put-patients-at-risk/ [accessed 2019-03-14]
7. Mauro M, Noto G, Prenestini A, Sarto F. Digital transformation in healthcare: assessing the role of digital technologies for managerial support processes. Technol Forecast Soc Change 2024;209:123781. [doi: 10.1016/j.techfore.2024.123781]
8. Marques ICP, Ferreira JJM. Digital transformation in the area of health: systematic review of 45 years of evolution. Health Technol 2019;10(3):575-586. [doi: 10.1007/s12553-019-00402-8]
9. Barbieri C, Neri L, Stuard S, Mari F, Martín-Guerrero JD. From electronic health records to clinical management systems: how the digital transformation can support healthcare services. Clin Kidney J 2023;16(11):1878-1884 [FREE Full text] [doi: 10.1093/ckj/sfad168] [Medline: 37915897]
10. Mulukuntla S, Pamulaparthyvenkata S. Digital transformation in healthcare: assessing the impact on patient care and safety. Int J Med Health Sci 2020;6(3) [FREE Full text]
11. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
12. Otero-García L, Mateos JT, Esperato A, Llubes-Arrià L, Regulez-Campo V, Muntaner C, et al. Austerity measures and underfunding of the Spanish health system during the COVID-19 pandemic-perception of healthcare staff in Spain. Int J Environ Res Public Health 2023;20(3):2594 [FREE Full text] [doi: 10.3390/ijerph20032594] [Medline: 36767958]
13. MOSCIARO M, KAIKA M, ENGELEN E. Financializing healthcare and infrastructures of social reproduction: How to bankrupt a hospital and be unprepared for a pandemic. J Soc Policy 2022;53(2):261-279. [doi: 10.1017/s004727942200023x]
14. Dennstädt F, Hastings J, Putora PM, Schmerder M, Cihoric N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. NPJ Digit Med 2025;8(1):143 [FREE Full text] [doi: 10.1038/s41746-025-01476-7] [Medline: 40050366]
15. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, Martinez EZ, Zucoloto ML, Østengaard L, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023;6(1):161 [FREE Full text] [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
16. Rane N, Choudhary S, Rane J. Acceptance of artificial intelligence: key factors, challenges, and implementation strategies. SSRN Electron J 2024:19. [doi: 10.2139/ssrn.4842167]
17. Karpathakis K, Morley J, Floridi L. A justifiable investment in AI for healthcare: aligning ambition with reality. SSRN Electron J 2024. [doi: 10.2139/ssrn.4795198]

XSL•FO
RenderX

18. Artificial intelligence in healthcare. European Commission. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/artificial-intelligence-healthcare_en [accessed 2025-12-20]
19. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards accurate differential diagnosis with large language models. Nature 2025;642(8067):451-457 [FREE Full text] [doi: 10.1038/s41586-025-08869-4] [Medline: 40205049]
20. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature 2025;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
21. Anderson BJ, Zia ul Haq M, Zhu Y, Hornback A, Cowan AD, Mott M, et al. Development and evaluation of a model to manage patient portal messages. NEJM AI 2025;2(3) [FREE Full text] [doi: 10.1056/aioa2400354]
22. Hassan H, Zipursky AR, Rabbani N, You JG, Tse G, Orenstein E, et al. Clinical implementation of artificial intelligence scribes in health care: a systematic review. Appl Clin Inform 2025;16(4):1121-1135 [FREE Full text] [doi: 10.1055/a-2597-2017] [Medline: 40306686]
23. Olson KD, Meeker D, Troup M, Barker TD, Nguyen VH, Manders JB, et al. Use of ambient AI scribes to reduce administrative burden and professional burnout. JAMA Netw Open 2025;8(10):e2534976 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.34976] [Medline: 41037268]
24. Chatzikou M, Latsou D, Apostolidis G, Billis A, Charisis V, Rigas ES, et al. Economic evaluation of artificially intelligent (AI) diagnostic systems: Cost consequence analysis of clinician-friendly interpretable computer-aided diagnosis (ICADX) tested in cardiology, obstetrics, and gastroenterology, from the HosmartAI horizon 2020 project. Healthcare (Basel) 2025;13(14):1661 [FREE Full text] [doi: 10.3390/healthcare13141661] [Medline: 40724686]
25. El Arab RA, Al Moosa OA. Systematic review of cost effectiveness and budget impact of artificial intelligence in healthcare. NPJ Digit Med 2025;8(1):548 [FREE Full text] [doi: 10.1038/s41746-025-01722-y] [Medline: 40858882]
26. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature 2023;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]
27. Zou J, Topol EJ. The rise of agentic AI teammates in medicine. The Lancet 2025;405(10477):457. [doi: 10.1016/s0140-6736(25)00202-8]
28. Moritz M, Topol E, Rajpurkar P. Coordinated AI agents for advancing healthcare. Nat Biomed Eng 2025;9(4):432-438. [doi: 10.1038/s41551-025-01363-2] [Medline: 40169759]
29. Qiu J, Lam K, Li G, Acharya A, Wong TY, Darzi A, et al. LLM-based agentic systems in medicine and healthcare. Nat Mach Intell 2024;6(12):1418-1420. [doi: 10.1038/s42256-024-00944-1]
30. DERM makes medical history as world's first autonomous skin cancer detection system is approved for clinical decisions in Europe. Skin Analytics. URL: https://skin-analytics.com/news/regulatory-certification/derm-class-iii-ce-mark/ [accessed 2025-12-20]
31. Gilbert S, Dai T, Mathias R. Consternation as congress proposal for autonomous prescribing AI coincides with the haphazard cuts at the FDA. NPJ Digit Med 2025;8(1):165 [FREE Full text] [doi: 10.1038/s41746-025-01540-2] [Medline: 40102664]
32. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021;8(2):e188-e194 [FREE Full text] [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
33. Myny D, Van Goubergen D, Gobert M, Vanderwee K, Van Hecke A, Defloor T. Non-direct patient care factors influencing nursing workload: a review of the literature. J Adv Nurs 2011;67(10):2109-2129. [doi: 10.1111/j.1365-2648.2011.05689.x] [Medline: 21722164]
34. Woolhandler S, Himmelstein DU. Administrative work consumes one-sixth of U.S. physicians' working hours and lowers their career satisfaction. Int J Health Serv 2014;44(4):635-642. [doi: 10.2190/hs.44.4.a]
35. Steinkamp J, Kantrowitz JJ, Airan-Javia S. Prevalence and sources of duplicate information in the electronic medical record. JAMA Netw Open 2022;5(9):e2233348 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.33348] [Medline: 36156143]
36. Fritz P, Kleinhans A, Raoufi R, Sediqi A, Schmid N, Schricker S, et al. Evaluation of medical decision support systems (DDX generators) using real medical cases of varying complexity and origin. BMC Med Inform Decis Mak 2022;22(1):254 [FREE Full text] [doi: 10.1186/s12911-022-01988-2] [Medline: 36153527]
37. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023;330(1):78-80 [FREE Full text] [doi: 10.1001/jama.2023.8288] [Medline: 37318797]
38. Ng JJW, Wang E, Zhou X, Zhou KX, Goh CXL, Sim GZN, et al. Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: a systematic review. BMC Med Inform Decis Mak 2025;25(1):236 [FREE Full text] [doi: 10.1186/s12911-025-03061-0] [Medline: 40598136]
39. Mathias R, McCulloch P, Chalkidou A, Gilbert S. Digital health technologies need regulation and reimbursement that enable flexible interactions and groupings. NPJ Digit Med 2024;7(1):148 [FREE Full text] [doi: 10.1038/s41746-024-01147-z] [Medline: 38890404]
40. Appelbaum SH. Socio‐technical systems theory: an intervention strategy for organizational development. Management Decision 1997;35(6):452-463. [doi: 10.1108/00251749710173823]
41. Behymer KJ, Flach JM. From autonomous systems to sociotechnical systems: designing effective collaborations. She Ji J Des Econ Innov 2016;2(2):105-114 [FREE Full text] [doi: 10.1016/j.sheji.2016.09.001]

XSL•FO
RenderX

42. Kudina O, Van de Poel I. A sociotechnical system perspective on AI. Minds Mach 2024;34(3):21. [doi: 10.1007/s11023-024-09680-2]
43. May C, Finch T. Implementing, embedding, and integrating practices: an outline of normalization process theory. Sociology 2009;43(3):535-554. [doi: 10.1177/0038038509103208]
44. Finch TL, Rapley T, Girling M, Mair FS, Murray E, Treweek S, et al. Improving the normalization of complex interventions: measure development based on normalization process theory (NoMAD): study protocol. Implement Sci 2013;8:43 [FREE Full text] [doi: 10.1186/1748-5908-8-43] [Medline: 23578304]
45. Murray E, Treweek S, Pope C, MacFarlane A, Ballini L, Dowrick C, et al. Normalisation process theory: a framework for developing, evaluating and implementing complex interventions. BMC Med 2010;8:63 [FREE Full text] [doi: 10.1186/1741-7015-8-63] [Medline: 20961442]
46. Riedl MO. Human‐centered artificial intelligence and machine learning. Hum Behav & Emerg Tech 2019;1(1):33-36. [doi: 10.1002/hbe2.117]
47. Dawoud K, Samek W, Eisert P, Lapuschkin S, Bosse S. Human-centered evaluation of XAI methods. : IEEE; 2023 Presented at: Proceedings of the 2023 IEEE International Conference on Data Mining Workshops (ICDMW); Dec 4, 2023; Shanghai, China p. 912-921. [doi: 10.1109/icdmw60847.2023.00122]
48. Holzinger A, Kargl M, Kipperer B, Regitnig P, Plass M, Muller H. Personas for artificial intelligence (AI) an open source toolbox. IEEE Access 2022;10:23732-23747. [doi: 10.1109/access.2022.3154776]
49. Combi C, Amico B, Bellazzi R, Holzinger A, Moore JH, Zitnik M, et al. A manifesto on explainability for artificial intelligence in medicine. Artif Intell Med 2022;133:102423 [FREE Full text] [doi: 10.1016/j.artmed.2022.102423] [Medline: 36328669]
50. Woolf SH. The meaning of translational research and why it matters. JAMA 2008;299(2):211-213. [doi: 10.1001/jama.2007.26] [Medline: 18182604]
51. Westerlund A, Sundberg L, Nilsen P. Implementation of implementation science knowledge: the research-practice gap paradox. Worldviews Evid Based Nurs 2019;16(5):332-334 [FREE Full text] [doi: 10.1111/wvn.12403] [Medline: 31603625]
52. Sanderson C, Douglas D, Lu Q, Schleiger E, Whittle J, Lacey J, et al. AI ethics principles in practice: perspectives of designers and developers. IEEE Trans Technol Soc 2023;4(2):171-187. [doi: 10.1109/tts.2023.3257303]
53. Tidjon LN, Khomh F. The different faces of AI ethics across the world: a principle-to-practice gap analysis. IEEE Trans Artif Intell 2023;4(4):820-839. [doi: 10.1109/tai.2022.3225132]
54. Lukkien DRM, Nap HH, Buimer HP, Peine A, Boon WPC, Ket JCF, et al. Toward responsible artificial intelligence in long-term care: a scoping review on practical approaches. Gerontologist 2023;63(1):155-168 [FREE Full text] [doi: 10.1093/geront/gnab180] [Medline: 34871399]
55. Oludapo S, Carroll N, Helfert M. Why do so many digital transformations fail? A bibliometric analysis and future research agenda. J Bus Res 2024;174:114528. [doi: 10.1016/j.jbusres.2024.114528]
56. Wekenborg MK, Gilbert S, Kather JN. Examining human-AI interaction in real-world healthcare beyond the laboratory. NPJ Digit Med 2025;8(1):169 [FREE Full text] [doi: 10.1038/s41746-025-01559-5] [Medline: 40108434]
57. Safi S, Thiessen T, Schmailzl KJ. Acceptance and resistance of new digital technologies in medicine: qualitative study. JMIR Res Protoc 2018;7(12):e11072 [FREE Full text] [doi: 10.2196/11072] [Medline: 30514693]
58. Sujan M, Baber C, Salomon P, Pool R, Chozos N, Aceves-González C. Human factors ergonomics in healthcare AI. Chartered Institute of Ergonomics & Human Factors 2021:45. [doi: 10.13140/RG.2.2.22455.85924]
59. Wosny M, Strasser LM, Hastings J. Experience of health care professionals using digital tools in the hospital: qualitative systematic review. JMIR Hum Factors 2023;10:e50357 [FREE Full text] [doi: 10.2196/50357] [Medline: 37847535]
60. Wekenborg MK, Förster K, Schweden F, Weidemann R, Bechtolsheim FV, Kirschbaum C, et al. Differences in physicians' ratings of work stressors and resources associated with digital transformation: cross-sectional study. J Med Internet Res 2024;26:e49581 [FREE Full text] [doi: 10.2196/49581] [Medline: 38885014]
61. Brod C. Technostress: The Human Cost of the Computer Revolution. Boston, MA: Addison-Wesley; 1984.
62. Alkureishi MA, Choo ZY, Rahman A, Ho K, Benning-Shorb J, Lenti G, et al. Digitally disconnected: qualitative study of patient perspectives on the digital divide and potential solutions. JMIR Hum Factors 2021;8(4):e33364 [FREE Full text] [doi: 10.2196/33364] [Medline: 34705664]
63. Tabche C, Raheem M, Alolaqi A, Rawaf S. Effect of electronic health records on doctor-patient relationship in Arabian gulf countries: a systematic review. Front Digit Health 2023;5:1252227 [FREE Full text] [doi: 10.3389/fdgth.2023.1252227] [Medline: 37877127]
64. Zheng K, Abraham J, Novak LL, Reynolds TL, Gettinger A. A survey of the literature on unintended consequences associated with health information technology: 2014–2015. Yearb Med Inform 2018;25(01):13-29. [doi: 10.15265/iy-2016-036]
65. Holden RJ, Rivera-Rodriguez AJ, Faye H, Scanlon MC, Karsh B. Automation and adaptation: Nurses' problem-solving behavior following the implementation of bar coded medication administration technology. Cogn Technol Work 2013;15(3):283-296 [FREE Full text] [doi: 10.1007/s10111-012-0229-4] [Medline: 24443642]
66. Antecedents of constructive human-AI collaboration: an exploration of human actors' key competencies. In: IFIP Advances in Information and Communication Technology. Cham, Switzerland: Springer International Publishing; 2021:113-124.

XSL•FO
RenderX

67. Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach Learn 2021;110(3):457-506 [FREE Full text] [doi: 10.1007/s10994-021-05946-3]
68. Dung L. Current cases of AI misalignment and their implications for future risks. Synthese 2023;202(5):138 [FREE Full text] [doi: 10.1007/s11229-023-04367-0]
69. Charter of fundamental rights of the European union (2000/C 364/01). European Parliament, the Council and the Commission of the European Union. URL: https://www.europarl.europa.eu/charter/pdf/text_en.pdf [accessed 2025-12-24]
70. Human dignity in the European Union (EU). Values@VET. 2025. URL: https://valuesatvet.si/files/2025/06/Human-dignity-in-the-European-Union.pdf [accessed 2025-12-20]
71. Freedom in the European Union (EU). Values@VET. URL: https://valuesatvet.si/files/2025/06/Freedom-in-the-European-Union.pdf [accessed 2025-12-20]
72. EU mechanism on democracy, the rule of law and fundamental rights: European Parliament resolution of 25 October 2016 with recommendations to the commission on the establishment of an EU mechanism on democracy, the rule of law and fundamental rights (2015/2254(INL)) (2018/C 215/25). European Parliament. 2018. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX%3A52016IP0409&utm [accessed 2025-12-20]
73. Klamert M, Kochenov D. Article 2 TEU. In: The EU Treaties and the Charter of Fundamental Rights: A Commentary. New York: Oxford Academic; 2019:22-30.
74. European Convention on Human Rights, as amended by protocols nos. 11, 14 and 15 supplemented by protocols nos. 1, 4, 6, 7, 12, 13 and 16. European Court of Human Rights. URL: https://www.echr.coe.int/documents/d/echr/convention_ENG [accessed 2025-12-20]
75. Regulation (EU) 2017/745 of 5 April 2017 on medical devices, amending directive 2001/83/EC, regulation (EC) No 178/2002 and regulation (EC) No 1223/2009 and repealing council directives 90/385/EEC and 93/42/EEC. European Parliament and Council of the European Union. URL: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng [accessed 2025-03-25]
76. Gilbert S, Mathias R, Schönfelder A, Wekenborg M, Steinigen-Fuchs J, Dillenseger A, et al. A roadmap for safe, regulation-compliant Living Labs for AI and digital health development. Sci Adv 2025;11(20):eadv7719 [FREE Full text] [doi: 10.1126/sciadv.adv7719] [Medline: 40367163]
77. Calderaro J, Morement H, Penault-Llorca F, Gilbert S, Kather JN. The case for homebrew AI in diagnostic pathology. J Pathol 2025;266(4-5):390-394 [FREE Full text] [doi: 10.1002/path.6438] [Medline: 40613320]
78. Ørngreen R, Levinsen KT. Workshops as a research methodology. Electron J E-Learn 2017;15(1):70-81 [FREE Full text]
79. Concannon TW, Meissner P, Grunbaum JA, McElwee N, Guise J, Santa J, et al. A new taxonomy for stakeholder engagement in patient-centered outcomes research. J Gen Intern Med 2012;27(8):985-991 [FREE Full text] [doi: 10.1007/s11606-012-2037-1] [Medline: 22528615]
80. Understanding healthcare workers confidence in artificial intelligence (AI) (Part 1). NHS Artificial Intelligence (AI) Lab, Health Education England (HEE). 2022. URL: https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/understanding-healthcare-workers-confidence-in-ai [accessed 2025-12-20]
81. Hess T, Matt C, Benlian A, Wiesböck F. Options for formulating a digital transformation strategy. In: Strategic Information Management: Theory and Practice. Oxfordshire, UK: Routledge; 2020:494.
82. Kejriwal M. AI in practice and implementation: issues and costs. In: Artificial Intelligence for Industries of the Future. Cham, Switzerland: Springer International Publishing; 2023:25-45.
83. Džakula A, Relić D. Health workforce shortage - doing the right things or doing things right? Croat Med J 2022;63(2):107-109 [FREE Full text] [doi: 10.3325/cmj.2022.63.107] [Medline: 35505643]
84. Global strategy on human resources for health: workforce 2030. World Health Organization. 2016. URL: https://iris.who.int/handle/10665/250368 [accessed 2025-02-13]
85. Rony MKK, Parvin MR, Wahiduzzaman M, Debnath M, Bala SD, Kayesh I. "I Wonder if my years of training and expertise will be devalued by Machines": Concerns about the replacement of medical professionals by artificial intelligence. SAGE Open Nurs 2024;10:23779608241245220 [FREE Full text] [doi: 10.1177/23779608241245220] [Medline: 38596508]
86. Kochan TA. Artificial intelligence and the future of work: a proactive strategy. AI Mag 2021;42(1):16-24. [doi: 10.1002/j.2371-9621.2021.tb00006.x]
87. Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. NPJ Digit Med 2022;5(1):66 [FREE Full text] [doi: 10.1038/s41746-022-00611-y] [Medline: 35641814]
88. Kumawat E, Datta A, Prentice C, Leung R. Artificial intelligence through the lens of hospitality employees: a systematic review. Int J Hosp Manag 2025;124:103986. [doi: 10.1016/j.ijhm.2024.103986]
89. de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med 2022;5(1):2 [FREE Full text] [doi: 10.1038/s41746-021-00549-7] [Medline: 35013569]
90. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. NPJ Digit Med 2019;2(1):79 [FREE Full text] [doi: 10.1038/s41746-019-0158-1] [Medline: 31453374]

XSL•FO
RenderX

91. Regulation (EU) 2017/746 of 5 April 2017 on in vitro diagnostic medical devices and repealing directive 98/79/EC and commission decision 2010/227/EU. European Parliament and Council of the European Union. URL: https://eur-lex.europa.eu/eli/reg/2017/746/oj/eng [accessed 2025-12-20]
92. Lohr S. What ever happened to IBM's Watson? The New York Times. 2016. URL: https://www.nytimes.com/2021/07/16/technology/what-happened-ibm-watson.html [accessed 2025-12-20]
93. Guidance on the health institution exemption under Article 5(5) of Regulation (EU) 2017/745 and Regulation (EU) 2017/746 (MDCG 2023-1). Medical Device Coordination Group (MDCG). 2023. URL: https://dskb.dk/wp-content/uploads/2021/09/In-house-guidance_stakeholders.pdf [accessed 2025-12-20]
94. Boyle G, Melvin T, Verdaasdonk RM, Van Boxtel RA, Reilly RB. Hospitals as medical device manufacturers: keeping to the medical device regulation (MDR) in the EU. BMJ Innov 2024;10(3):74-80. [doi: 10.1136/bmjinnov-2023-001150]
95. Mit Künstlicher Intelligenz das Krankenhaus von morgen gestalten. SmartHospital.NRW. URL: https://smarthospital.nrw/ [accessed 2025-12-20]
96. Good machine learning practice for medical device development: guiding principles. Medicines and Healthcare products Regulatory Agency (MHRA). URL: https://www.gov.uk/government/publications/good-machine-learning-practice-for-medical-device-development-guiding-principles/good-machine-learning-practice-for-medical-device-development-guiding-principles#guiding-principles [accessed 2021-10-27]
97. Arun S, Grosheva M, Kosenko M, Robertus JL, Blyuss O, Gabe R, et al. Systematic scoping review of external validation studies of AI pathology models for lung cancer diagnosis. NPJ Precis Oncol 2025;9(1):166 [FREE Full text] [doi: 10.1038/s41698-025-00940-7] [Medline: 40483288]
98. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019;25(9):1337-1340. [doi: 10.1038/s41591-019-0548-6] [Medline: 31427808]
99. McCradden MD, London AJ, Gichoya JW, Sendak M, Erdman L, Stedman I, et al. CANAIRI: the collaboration for translational artificial intelligence trials in healthcare. Nat Med 2025;31(1):9-11. [doi: 10.1038/s41591-024-03364-1] [Medline: 39762426]
100. Hellmeier F, Brosien K, Eickhoff C, Meyer A. Beyond one-time validation: a framework for adaptive validation of prognostic and diagnostic AI-based medical devices. ArXiv Preprint posted online on September 7, 2024. [doi: 10.48550/ARXIV.2409.04794]
101. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature 2024;630(8017):625-630 [FREE Full text] [doi: 10.1038/s41586-024-07421-0] [Medline: 38898292]
102. Griot M, Hemptinne C, Vanderdonckt J, Yuksel D. Large language models lack essential metacognition for reliable medical reasoning. Nat Commun 2025;16(1):642 [FREE Full text] [doi: 10.1038/s41467-024-55628-6] [Medline: 39809759]
103. Omar M, Soffer S, Agbareia R, Bragazzi NL, Apakama DU, Horowitz CR, et al. Sociodemographic biases in medical decision making by large language models. Nat Med 2025;31(6):1873-1881. [doi: 10.1038/s41591-025-03626-6] [Medline: 40195448]
104. Alber DA, Yang Z, Alyakin A, Yang E, Rai S, Valliani AA, et al. Medical large language models are vulnerable to data-poisoning attacks. Nat Med 2025;31(2):618-626. [doi: 10.1038/s41591-024-03445-1] [Medline: 39779928]
105. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 2024;30(9):2613-2622. [doi: 10.1038/s41591-024-03097-1] [Medline: 38965432]
106. Tam TYC, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med 2024;7(1):258 [FREE Full text] [doi: 10.1038/s41746-024-01258-7] [Medline: 39333376]
107. Mehandru N, Miao BY, Almaraz ER, Sushil M, Butte AJ, Alaa A. Evaluating large language models as agents in the clinic. NPJ Digit Med 2024;7(1):84 [FREE Full text] [doi: 10.1038/s41746-024-01083-y] [Medline: 38570554]
108. ISO/IEC 42001:2023 Information technology - artificial intelligence - management system. International Organization for Standardization. 2023. URL: https://www.iso.org/standard/81230.html#lifecycle [accessed 2025-12-20]
109. FG-AI4H DEL7.4 - Clinical evaluation of AI for health. International Telecommunication Union. 2023. URL: https://www.itu.int/pub/T-FG-AI4H-2023-3 [accessed 2025-12-20]
110. Welzel C, Cotte F, Wekenborg M, Vasey B, McCulloch P, Gilbert S. Holistic human-serving digitization of health care needs integrated automated system-level assessment tools. J Med Internet Res 2023;25:e50158 [FREE Full text] [doi: 10.2196/50158] [Medline: 38117545]
111. Mathias R, Vasey B, Chalkidou A, Riedemann L, Melvin T, Gilbert S. Safe AI-enabled digital health technologies need built-in open feedback. Nat Med 2025;31(2):370-375. [doi: 10.1038/s41591-024-03397-6] [Medline: 39905271]
112. Floridi L, Holweg M, Taddeo M, Amaya Silva J, Mökander J, Wen Y. capAI - A procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act. SSRN Electron J 2022:91. [doi: 10.2139/ssrn.4064091]
113. Directorate General for Communications Networks, Content and Technology. The assessment list for trustworthy artificial intelligence (ALTAI) for self assessment. European Commission. 2020. URL: https://data.europa.eu/doi/10.2759/002360 [accessed 2025-06-07]

XSL•FO
RenderX

114. Coeckelbergh M. Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci Eng Ethics 2020;26(4):2051-2068 [FREE Full text] [doi: 10.1007/s11948-019-00146-8] [Medline: 31650511]
115. Santoni de Sio F, Mecacci G. Four responsibility gaps with artificial intelligence: why they matter and how to address them. Philos Technol 2021;34(4):1057-1084 [FREE Full text] [doi: 10.1007/s13347-021-00450-x]
116. Beauchamp T. The principle of beneficence in applied ethics. The Stanford Encyclopedia of Philosophy. 2019. URL: https://plato.stanford.edu/archives/spr2019/entries/principle-beneficence/ [accessed 2025-12-20]
117. Varkey B. Principles of clinical ethics and their application to practice. Med Princ Pract 2021;30(1):17-28 [FREE Full text] [doi: 10.1159/000509119] [Medline: 32498071]
118. Porsdam Mann S, Savulescu J, Sahakian BJ. Facilitating the ethical use of health data for the benefit of society: electronic health records, consent and the duty of easy rescue. Philos Trans A Math Phys Eng Sci 2016;374(2083):20160130 [FREE Full text] [doi: 10.1098/rsta.2016.0130] [Medline: 28336803]
119. Placani A. Anthropomorphism in AI: hype and fallacy. AI Ethics 2024;4(3):691-698 [FREE Full text] [doi: 10.1007/s43681-024-00419-4]
120. McLennan S, Fiske A, Tigard D, Müller R, Haddadin S, Buyx A. Embedded ethics: a proposal for integrating ethics into the development of medical AI. BMC Med Ethics 2022 26;23(1):6 [FREE Full text] [doi: 10.1186/s12910-022-00746-3] [Medline: 35081955]
121. Regulation (EU) 2024/1689 of the European Parliament and of the council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). European Parliament and Council of the European Union. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng [accessed 2025-12-20]
122. Early successes, untapped potential, lingering questions: AI adoption in healthcare report 2024. Healthcare Information and Management Systems Society (HIMSS), Medscape. 2024. URL: https://cdn.sanity.io/files/sqo8bpt9/production/68216fa5d161adebceb50b7add5b496138a78cdb.pdf [accessed 2025-12-20]
123. Schubert T, Oosterlinck T, Stevens RD, Maxwell PH, van der Schaar M. AI education for clinicians. EClinicalMedicine 2025;79:102968 [FREE Full text] [doi: 10.1016/j.eclinm.2024.102968] [Medline: 39720600]
124. Zirar A, Ali SI, Islam N. Worker and workplace artificial intelligence (AI) coexistence: emerging themes and research agenda. Technovation 2023;124:102747. [doi: 10.1016/j.technovation.2023.102747]
125. Elendu C, Amaechi DC, Okatta AU, Amaechi EC, Elendu TC, Ezeh CP, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024;103(27):e38813 [FREE Full text] [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
126. So HY, Chen PP, Wong GKC, Chan TTN. Simulation in medical education. J R Coll Physicians Edinb 2019;49(1):52-57. [doi: 10.4997/jrcpe.2019.112]
127. Datta R, Upadhyay K, Jaideep C. Simulation and its role in medical education. Med J Armed Forces India 2012;68(2):167-172 [FREE Full text] [doi: 10.1016/S0377-1237(12)60040-9] [Medline: 24623932]
128. Thistlethwaite JE, Davies D, Ekeocha S, Kidd JM, MacDougall C, Matthews P, et al. The effectiveness of case-based learning in health professional education. A BEME systematic review: BEME Guide No. 23. Medical Teacher 2012;34(6):e421-e444. [doi: 10.3109/0142159x.2012.680939]
129. Mukurunge E, Reid M, Fichardt A, Nel M. Interactive workshops as a learning and teaching method for primary healthcare nurses. Health SA 2021;26:1643 [FREE Full text] [doi: 10.4102/hsag.v26i0.1643] [Medline: 34956654]
130. Daye D, Wiggins WF, Lungren MP, Alkasab T, Kottler N, Allen B, et al. Implementation of clinical artificial intelligence in radiology: who decides and how? Radiology 2022;305(3):555-563 [FREE Full text] [doi: 10.1148/radiol.212151] [Medline: 35916673]
131. Hoff KA, Bashir M. Trust in automation: integrating empirical evidence on factors that influence trust. Hum Factors 2015;57(3):407-434. [doi: 10.1177/0018720814547570] [Medline: 25875432]
132. Jiang T, Sun Z, Fu S, Lv Y. Human-AI interaction research agenda: a user-centered perspective. Data Inf Manag 2024;8(4):100078. [doi: 10.1016/j.dim.2024.100078]
133. Vered M, Livni T, Howe PDL, Miller T, Sonenberg L. The effects of explanations on automation bias. Artificial Intelligence 2023;322:103952. [doi: 10.1016/j.artint.2023.103952]
134. Choudhury A, Chaudhry Z. Large language models and user trust: consequence of self-referential learning loop and the deskilling of health care professionals. J Med Internet Res 2024;26:e56764 [FREE Full text] [doi: 10.2196/56764] [Medline: 38662419]
135. Ng MY, Kapur S, Blizinsky KD, Hernandez-Boussard T. The AI life cycle: a holistic approach to creating ethical AI for health decisions. Nat Med 2022;28(11):2247-2249 [FREE Full text] [doi: 10.1038/s41591-022-01993-y] [Medline: 36163298]
136. Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. Lancet Digit Health 2022;4(5):e384-e397. [doi: 10.1016/s2589-7500(22)00003-6]
137. Faust L, Wilson P, Asai S, Fu S, Liu H, Ruan X, et al. Considerations for quality control monitoring of machine learning models in clinical practice. JMIR Med Inform 2024;12:e50437 [FREE Full text] [doi: 10.2196/50437] [Medline: 38941140]

XSL•FO
RenderX

138. Planning for managing legacy systems and decommissioning digital healthcare technologies. NHS AI and Digital Regulations Service for Health and Social Care. URL: https://www.digitalregulations.innovation.nhs.uk/regulations-and-guidance-for-adopters/all-adopters-guidance/planning-for-managing-legacy-systems-and-decommissioning-digital-healthcare-technologies/ [accessed 2023-11-13]
139. Sharma A, Nayancy, Verma R. The Confluence of Cryptography, Blockchain and Artificial Intelligence. Florida, USA: CRC Press; 2025.
140. MHRA, Brunel University. Project Report: Research into Methodology for Determining Significant Change in the Way That an Adaptive AI Algorithm Medical Device Is Working and How Such Change Should Be Regulated. https://www.gov.uk. URL: https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device [accessed 2025-03-29]
141. The most trusted AI in mental healthcare: scale behavioral health with clinical AI. Limbic. 2025. URL: https://www.limbic.ai/ [accessed 2025-12-20]
142. We provide validated information for healthcare professionals. Prof. Valmed - Validated Medical Information GmbH. URL: https://profvalmed.com/ [accessed 2025-12-20]
143. Frequently asked questions. deepc GmbH. URL: https://www.deepc.ai/learn/faq [accessed 2025-12-20]
144. Study on the deployment of AI in healthcare: final report. Publications Office of the European Union. 2025. URL: https://data.europa.eu/doi/10.2875/2169577 [accessed 2025-08-07]
145. Eskofier BM, Klucken J. Predictive models for health deterioration: understanding disease pathways for personalized medicine. Annu Rev Biomed Eng 2023;25(1):131-156 [FREE Full text] [doi: 10.1146/annurev-bioeng-110220-030247] [Medline: 36854259]
146. Goldfarb A, Taska B, Teodoridis F. Artificial intelligence in health care? Evidence from online job postings. AEA Pap Proc 2020;110:400-404 [FREE Full text] [doi: 10.1257/pandp.20201006]
147. Wu K, Wu E, Theodorou B, Liang W, Mack C, Glass L, et al. Characterizing the clinical adoption of medical AI devices through U.S. insurance claims. NEJM AI 2024;1(1). [doi: 10.1056/aioa2300030]
148. Ulfert AS, Le Blanc P, González-Romá V, Grote G, Langer M. Are we ahead of the trend or just following? The role of work and organizational psychology in shaping emerging technologies at work. Eur J Work Organ Psychol 2024;33(2):120-129. [doi: 10.1080/1359432x.2024.2324934]
149. Gilbert S, Anderson S, Daumer M, Li P, Melvin T, Williams R. Learning from experience and finding the right balance in the governance of artificial intelligence and digital health technologies. J Med Internet Res 2023;25:e43682 [FREE Full text] [doi: 10.2196/43682] [Medline: 37058329]
150. Ganesan S, Somasiri N. Navigating the integration of machine learning in healthcare: challenges, strategies, and ethical considerations. J Comput Cogn Eng 2024. [doi: 10.47852/bonviewJCCE42023600]
151. Developing healthcare workers' confidence in artificial intelligence (AI) (Part 2). NHS Artificial Intelligence (AI) Lab, Health Education England (HEE). 2023. URL: https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/developing-healthcare-workers-confidence-in-ai [accessed 2025-12-20]
152. Reddy S, Rogers W, Makinen VP, Coiera E, Brown P, Wenzel M, et al. Evaluation framework to guide implementation of AI systems into healthcare settings. BMJ Health Care Inform 2021;28(1):e100444 [FREE Full text] [doi: 10.1136/bmjhci-2021-100444] [Medline: 34642177]
153. Moreno-Sánchez PA, Ser JD, Gils MV, Hernesniemi J. A design framework for operationalizing trustworthy artificial intelligence in healthcare: requirements, tradeoffs and challenges for its clinical adoption. Information Fusion 2025;127:103812. [doi: 10.2139/ssrn.5249603]
154. Nair M, Nygren J, Nilsen P, Gama F, Neher M, Larsson I, et al. Critical activities for successful implementation and adoption of AI in healthcare: towards a process framework for healthcare organizations. Front Digit Health 2025;7:1550459 [FREE Full text] [doi: 10.3389/fdgth.2025.1550459] [Medline: 40453810]
155. Nilsen P, Svedberg P, Neher M, Nair M, Larsson I, Petersson L, et al. A framework to guide implementation of AI in health care: protocol for a cocreation research project. JMIR Res Protoc 2023;12:e50216 [FREE Full text] [doi: 10.2196/50216] [Medline: 37938896]
156. Dagan N, Devons-Sberro S, Paz Z, Zoller L, Sommer A, Shaham G, et al. Evaluation of AI solutions in health care organizations — The OPTICA tool. NEJM AI 2024;1(9). [doi: 10.1056/aics2300269]
157. Mittermaier M, Raza M, Kvedar JC. Collaborative strategies for deploying AI-based physician decision support systems: challenges and deployment approaches. NPJ Digit Med 2023;6(1):137 [FREE Full text] [doi: 10.1038/s41746-023-00889-6] [Medline: 37543707]
158. Davahli MR, Karwowski W, Fiok K, Wan T, Parsaei HR. Controlling safety of artificial intelligence-based systems in healthcare. Symmetry 2021;13(1):102 [FREE Full text] [doi: 10.3390/sym13010102]
159. Labkoff S, Oladimeji B, Kannry J, Solomonides A, Leftwich R, Koski E, et al. Toward a responsible future: recommendations for AI-enabled clinical decision support. J Am Med Inform Assoc 2024;31(11):2730-2739. [doi: 10.1093/jamia/ocae209] [Medline: 39325508]
160. Lekadira K, Osuala R, Gallin C. FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. ArXiv Preprint posted online on September 20, 2021. [doi: 10.48550/arXiv.2109.09658]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
CE: Conformité Européenne
EU: European Union
EURAID: European Responsible AI Development
FG-AI4H: Focus Group on AI for Health
HCP: health care professional
ISO/IEC: International Organization for Standardization/ International Electrotechnical Commission
LLM: large language model
MDR: Medical Device Regulation
NHS: National Health Service
SMART: specific, measurable, attainable, relevant, and time-bound
TEU: Treaty on European Union
XAI: explainable AI

Edited by J Sarvestan; submitted 16.Jul.2025; peer-reviewed by I Schlömer, KH Lin; comments to author 22.Aug.2025; revised version received 05.Nov.2025; accepted 06.Nov.2025; published 29.Jan.2026.

Please cite as:
Schönfelder A, Eberlein-Gonska M, Hülsken-Giesler M, Jovy-Klein F, Kather JN, Kohoutek E, Lennefer T, Liebert E, Lipprandt M, Mathias R, Muti HS, Obergassel J, Reibel T, Rösler U, Schneider M, Schlicht L, Schlieter H, Schmieding ML, Schweingruber N, Sedlmayr M, Strametz R, Susec B, Wekenborg MK, Weicken E, Weitz K, Diehl A, Gilbert S
Collaborative and Cooperative Hospital "In-House" Medical Device Development and Implementation in the AI Age: The European Responsible AI Development (EURAID) Framework Compatible With European Values
J Med Internet Res 2026;28:e80754
URL: https://www.jmir.org/2026/1/e80754
doi:10.2196/80754
PMID:

©Anett Schönfelder, Maria Eberlein-Gonska, Manfred Hülsken-Giesler, Florian Jovy-Klein, Jakob Nikolas Kather, Elisabeth Kohoutek, Thomas Lennefer, Elisabeth Liebert, Myriam Lipprandt, Rebecca Mathias, Hannah Sophie Muti, Julius Obergassel, Thomas Reibel, Ulrike Rösler, Moritz Schneider, Larissa Schlicht, Hannes Schlieter, Malte L Schmieding, Nils Schweingruber, Martin Sedlmayr, Reinhard Strametz, Barbara Susec, Magdalena Katharina Wekenborg, Eva Weicken, Katharina Weitz, Anke Diehl, Stephen Gilbert. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

AI Triage in Primary Care: Building Safer and More Equitable Real-World Evidence

Aymn Alamoudi[1,2], MSc; Evangelos Kontopantelis[3,4], PhD; Salwa S Zghebi[1], PhD; Benjamin Brown[1], PhD

[1]Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, England, United Kingdom

[2]Department of Public Health, School of Nursing and Health Sciences, Jazan University, Jazan, Jazan Region, Saudi Arabia

[3]Division of Informatics, Imaging and Data Sciences, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, England, United Kingdom

[4]Division of Family Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

Corresponding Author:
Aymn Alamoudi, MSc
Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, Faculty of Biology, Medicine and Health
University of Manchester
Williamson Building, 5th Floor
Oxford Road
Manchester, England, M13 9PL
United Kingdom
Phone: 44 161 306 6000
Email: aymn.alamoudi@postgrad.manchester.ac.uk

Abstract

Artificial intelligence triage in general practice is developing rapidly within the primary care digital transformation, promising efficiency gains and safety standardization in overwhelmed primary care systems. However, current evidence is drawn from retrospective validations, emergency settings, or vignettes, with scant evaluation of real-world outcomes and almost no equity-stratified safety data, despite known disparities across age, ethnicity, language, and deprivation. From a sociotechnical standpoint, which considers the fit between people, tasks, technology, and organizational context, risks arise not only from algorithmic bias and undertriage but also from human factors, workflow misalignment, governance gaps, and inadequate postdeployment monitoring. We argue that ensuring artificial intelligence triage is safe and equitable requires real-world evaluations in primary care settings, equity-focused performance reporting using theoretically informed frameworks, and rigorous postmarket surveillance. Without these, deployment may widen existing health inequalities rather than moderate them.

(J Med Internet Res 2026;28:e88396) doi:10.2196/88396

KEYWORDS

artificial intelligence; AI; triage; primary care; general practice; patient safety; health equity; real-world evidence; clinical decision support

Introduction

Globally, primary care faces sustained growth in demand, increased patient complexity, and a workforce whose full-time equivalent growth has not matched demand, resulting in persistent access pressures and delays [1,2]. The COVID-19 pandemic accelerated the adoption of remote and digital access, including online consultations and reinforced strategic commitments to “digital front door” models within health systems, such as the National Health Service [3-5]. Online consultation submissions in England rose from approximately 2.7 million in October 2023 to a peak of 8.3 million in October 2025, as seen in Figure 1, highlighting the rapid and sustained growth of digital entry points into general practice (GP) [6]. Evidence also suggests that digital access is not equity neutral. A systematic review of inequalities in remote GP consultations found differential use by sociodemographic characteristics, with internet-based consultations more frequently used by younger, more affluent, and more educated groups, and noted that the impact of these inequalities on clinical outcomes remains uncertain [7].

XSL•FO
RenderX

Figure 1. Growth in monthly online consultation submissions in England. Data source: National Health Service England release [6]. Information from NHS England, licenced under the current version of the Open Government Licence.

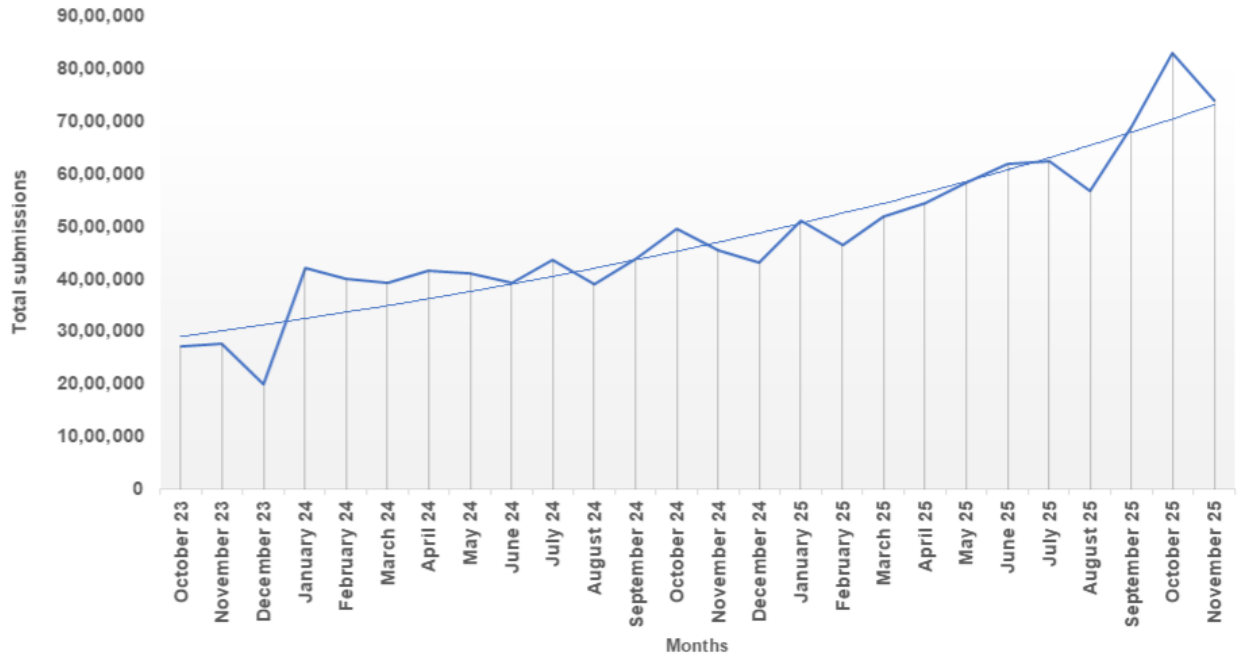

In this context, artificial intelligence (AI)–enabled triage combines structured questions, red-flag pathways, and machine learning (ML) risk stratification with electronic health record (EHR) integration and clinician oversight to route patients more efficiently and potentially improve safety [8-10]. In this viewpoint, we distinguish between 3 related but conceptually distinct system types: symptom checkers, clinical decision support systems (CDSSs), and AI triage. Symptom checkers are patient-facing digital tools that provide health advice or triage recommendations directly to users, often without clinician oversight, and have been widely evaluated in consumer and emergency contexts [8]. CDSSs are clinician-facing tools embedded within clinical workflows or EHRs that support decision-making through alerts, risk scores, or guideline-based recommendations [9,10]. AI-enabled triage refers to digital systems that collect patient-reported information and generate urgency or routing recommendations (eg, self-care, routine review, urgent GP assessment, or emergency referral), with or without clinician oversight [11]. These systems may be embedded within online consultation platforms, patient-facing symptom checkers, or CDSS. Importantly, not all online consultation systems are AI enabled and not all AI-enabled triage systems function as stand-alone symptom checkers.

This viewpoint advances 3 linked arguments. First, we argue that triage in primary care is a safety-critical and equity-sensitive function, such that errors or delays can produce serious harm and unequal outcomes. Second, we show that the current evidence base for AI-supported triage is dominated by emergency department (ED), vignette, and retrospective studies, with little real-world or equity-stratified evaluation in GP. Third, we argue that AI triage operates as a sociotechnical system shaped by human behavior, workflows, and governance, meaning that algorithmic accuracy alone cannot guarantee safety or fairness. This viewpoint aims to outline a practical agenda for evaluating and governing AI-enabled triage in GP that integrates real-world safety outcomes, equity-stratified performance reporting, and sociotechnical implementation and surveillance. The intended audience includes GP clinicians and practice leaders, digital health and AI developers, evaluators and implementation scientists, and policymakers and regulators responsible for deployment and monitoring. Our contribution is to consolidate a practical, real-world evaluation and governance agenda for AI triage in GP that integrates sociotechnical safety (workflow and human factors), equity-stratified performance reporting (including an example of fairness), and postdeployment surveillance.

Triage in Primary Care Is Safety Critical and Equity Sensitive

In health care, triage refers to the systematic process of assessing patient urgency and risk to determine the appropriate level, timing, and pathway of care [12,13]. In primary care, triage does not establish a diagnosis but prioritizes patients for self-care, routine review, urgent GP assessment, or emergency referral based on presenting symptoms, clinical risk, and service capacity [14,15]. This function is safety critical because misclassification can lead to delayed diagnosis, inappropriate self-management, or unnecessary escalation [8-10]. AI-enabled triage promises standardization and auditability but introduces novel patient-safety risks, such as automation bias, algorithmic mistriage, and digital exclusion, particularly for socially disadvantaged groups [8,16,17]. Moreover, model performance (eg, sensitivity, specificity, calibration, and error rates) may vary by age, ethnicity, language, or limited digital access, unless these dimensions are intentionally tested and monitored [18-22].

XSL•FO
RenderX

The absence of equity evaluation may pose a significant risk. If models are calibrated primarily on majority language, younger, or White-majority cohorts, AI triage may systematically de-escalate or deprioritize patients whose symptom descriptions diverge due to cultural or linguistic factors. Coupled with cognitive and automation biases in human users, the most vulnerable groups risk unsafe disposition, such as self-care advice when urgent assessment is indicated.

The "equity blind spot" in AI triage is not merely a technical glitch; it reflects broader systemic oversight. To operationalize safe, equitable AI, embedding framework-informed stratification, for example, PROGRESS-Plus, fairness metrics, and multidimensional performance reporting into every stage of model development, validation, and deployment is needed. Without this, AI triage may reinforce and even amplify existing health care disparities if deployed without adequate safeguards.

What the Evidence Shows and What It Misses

What Current Studies Show

Controlled evaluations of AI triage report high technical performance, with area under the receiver operating characteristic (singular) curve values typically between 0.82 and 0.94 and sensitivities often exceeding 0.75. However, these studies are predominantly retrospective, vignette based, or conducted in EDs and hospital settings. They rarely reflect routine workflows in GP [23-26].

Recently, Abualruz et al [12] reviewed 22 studies on AI-supported triage, most of which were carried out in emergency, acute, or hospital-based settings. Only a small subset examined outpatient or primary-care use. As a result, the current literature tells us little about how AI triage performs in everyday GP or how it affects patient safety in real workflows. Table 1 presents the distribution of published AI-supported triage studies by clinical setting.

Table 1. Distribution of published artificial intelligence–supported triage studies by clinical setting (N=22).

Clinical setting	Study type	Studies, n (%)
Emergency department or hospital	Real patient data	19 (86)
Primary care	Clinical vignettes or qualitative studies	3 (14)
Primary care	Real patient data	0 (0)

Most studies have been conducted in emergency or hospital settings using real patient data (19/22, 86%). In total, 14% (3/22) of the studies relied on clinical vignettes or qualitative interviews, and none (0/22, 0%) evaluated AI-supported triage using real-world patient data in routine GP.

Why This Evidence Is Not Sufficient for Safe and Equitable GP Deployment

Equity reporting is also sparse. Few studies disaggregate performance by age, ethnicity, language, or socioeconomic status [27-30]. Intersectional analyses, for example, age×ethnicity or ethnicity×deprivation, are almost absent. Subgroup calibration, false-negative rates, and false-positive rates (FPRs) are rarely reported.

Study design further limits interpretability. Vignette-based and retrospective analyses do not capture real-world pressures, such as workload variation, free-text symptom input, clinician overrides, or case-mix drift [31]. Prospective designs, such as controlled interrupted time series or cluster-randomized trials, are almost never used in GP settings [1,3]. Without these designs, safety effects cannot be attributed reliably to AI deployment.

Postdeployment monitoring is also underdeveloped (Table 2). Few studies report ongoing calibration checks, subgroup performance dashboards, or systematic incident reporting aligned to the World Health Organization (WHO) International Classification for Patient Safety [32,33]. As a result, health systems lack visibility into how AI triage safety changes over time.

Table 2. What the evidence on artificial intelligence triage shows and what it misses.

Domain	What studies typically show	What is usually missing
Accuracy	High area under the receiver operating characteristic curve (≈0.82-0.94) and reasonable sensitivity in retrospective and vignette-based studies	Performance under real general practitioner workload, with free-text input, comorbidity, and clinician override
Setting	Predominantly emergency departments, acute care, or simulated cases	Routine general practice, community clinics, and longitudinal follow-up
Safety outcomes	Agreement with clinicians or reference standards	Delayed diagnosis, avoidable emergency use, or patient harm
Equity	Rare or absent subgroup reporting	Performance by age, ethnicity, language, deprivation, or intersectional groups
Monitoring	One-off validation at model development	Postdeployment drift, subgroup miscalibration, and incident tracking

XSL•FO
RenderX

Consistent with this, an ED-focused scoping review found limited demographic breakdowns and no multidimensional analyses, leaving equity implications unclear [27]. Within UK GP settings, stratified data on undertriage or performance by deprivation or ethnicity are rare [16,29], and subgroup calibration metrics or true-positive rate (TPR) and FPR reporting are notably absent. A recent international review found variable triage accuracy, poor calibration reporting, and limited deployment-level evaluation, reinforcing that the evidence gap is global [34].

Why AI Triage Is a Sociotechnical System

Overview

AI-enabled triage systems are not isolated algorithms; they operate within complex care delivery systems where human factors, workflows, and trust dynamics profoundly shape safety. A sociotechnical system perspective, exemplified by the systems engineering initiative for patient safety framework, analyzes how people, tasks, tools, and organizational structures interact to influence patient safety.

By contrast, implementation frameworks focus on organizational readiness, technology adoption, and sustainability [35,36]. Together, these complementary approaches emphasize that deploying an accurate ML model alone does not guarantee safe outcomes; safety depends on both workflow integration and organizational adoption.

Human factors and trust are central. Health care professionals and patients must interpret AI-generated recommendations within their cognitive, ethical, and emotional contexts. A recent qualitative work on AI-based triage in Swedish primary care underlines how trust emerges from lived experience, transparency, and perceived reliability. Both patients' and professionals' trust is contingent on real-world usability and clear decision roles and not just model accuracy [31].

Similar issues are emerging in teledentistry and dental triage, where AI-enabled chatbots and triage systems are used to prioritize pain, infection, trauma, or urgent referral pathways [37]. Early work includes prototype "intelligent dental triage systems" and evaluations of AI chatbots for dental queries, but the same core risks apply: safety-critical undertriage, unequal performance for patients with language barriers or limited digital access, and workflow integration challenges in busy dental practices [38]. AI tools in dental assessment and smile analysis, such as Dynasmile, a video-based AI smile analysis platform in orthodontics, illustrate the expanding role of AI beyond workflow triage into diagnostic and aesthetic decision support in oral care [39].

This cross-domain comparison reinforces our central claim: AI triage should be evaluated as a sociotechnical intervention with equity-stratified safety reporting and postdeployment monitoring, regardless of clinical specialty.

Explainable AI to Support Calibrated Trust and Reduce Automation Bias

In safety-critical triage, explanations should aim to support calibrated trust not persuasion. Evidence from clinical decision-support research suggests that well-designed explanations can improve clinician understanding and trust calibration, while poorly designed explanations can increase overreliance and automation bias [40,41].

In practice, explainable AI for GP triage should be workflow integrated and low burden, including the following: (1) a short list of the main drivers for escalation (eg, red-flag symptoms and abnormal risk profile), (2) uncertainty indicators or confidence bands, (3) an "override required" prompt for high-risk edge cases, and (4) safety-netting text that is consistent with the triage rationale. Explanation stability is also important; near-identical inputs should not produce inconsistent rationales, as this undermines trust and may increase unsafe deference [40,42].

These requirements for explanation design reinforce the importance of sociotechnical fit described subsequently.

What Real-World Deployments Show

AI triage does not operate as a stand-alone algorithm. In GP, it is embedded in symptom checkers, online consultations, patient apps, and EHR-based decision-support tools. These systems influence how patients describe symptoms, how clinicians prioritize work, and how care is delivered.

Evidence from multiple countries shows potential efficiency gains. In Iceland, an ML triage model for respiratory symptoms improved previsit risk stratification in community clinics [43]. Similarly, Brazil's primary care referral triage system demonstrated improved appropriateness of specialist referrals [44]. Studies from Sweden and Italy reported improved workflow transparency but persistent concerns about trust, usability, and clinician acceptance [31,33].

However, most deployments remain early stage, small scale, or limited to specific pathways. Systematic reviews of symptom checkers across Europe, the United States, Spain, Canada, and Asia report wide variability in triage accuracy and frequent mismatches between algorithmic and clinician assessments, particularly for complex multimorbidity and non–native-language users [8,14,34,45].

Why Workflow, Trust, and Integration Matter

Qualitative evidence highlights that sociotechnical fit is critical. Steerling et al [31] found that both health care professionals and patients require alignment with clinical judgment, transparency, and oversight before trusting AI-based triage. Similarly, Siira et al [46] identified 3 interacting barriers in Swedish primary care: (1) professional skepticism or resistance and trust, (2) organizational readiness and digital maturity, and (3) technical limitations and poor EHR integration. Successful sites mitigated these by hands-on leadership and staff training, "superuser" networks, and iterative codevelopment with vendors. Even where efficiency gains were perceived, unresolved integration gaps and complex case-mix sustained workload and safety concerns.

Evidence from the UK primary care e-visits (14 practices; 16 staff, 24 patients; 2020-2021) identified 7 concrete AI use cases—workflow routing, directing, prioritization, postsubmission adaptive questioning, writing assistance,

XSL•FO
RenderX

self-help information, and autobooking—and found acceptability hinged on clinical oversight, timely responses, and ongoing evaluation. Perceived upsides were workload relief and faster help, while risks were depersonalization and mistriage, if poorly implemented [47].

What This Means for Safety

These findings show that AI triage safety depends on how tools interact with people, workflows, and organizational routines. This aligns with sociotechnical safety theory, particularly the systems engineering initiative for patient safety framework, which emphasizes the fit between tasks, technology, and organizational context [32,48,49].

AI also offers enhanced structured data capture, natural language processing (NLP)–enabled symptom interpretation, EHR-integrated safety netting, and auditable decision trails [50]. NLP-enabled symptom interpretation can, for example, recognize "feeling pressure in chest" as equivalent to angina, supporting safer triage for patients who do not use standard terminology [26,45]. However, these benefits are offset by risks. Algorithmic mistriage, automation bias, and poor integration can delay escalation or overload urgent pathways [8,13].

Without governance mechanisms, such as version control, audit logs, safety dashboards, and periodic revalidation, performance may degrade over time as populations, language, and risk profiles change [32]. Therefore, sociotechnical alignment is not optional; it appears essential for safe AI-enabled triage.

The Equity Blind Spot

Although AI-enabled triage systems are promoted as equitable tools for managing primary care demand, the current evidence reveals a persistent equity blind spot, driven by underrepresentation, limited fairness measurement, and neglect of equity-focused monitoring [51].

Equity in digital health demands more than equal treatment; it requires fair opportunity to achieve safe, good outcomes, especially when baseline disadvantages exist. Equity reporting should use structured tools. The PROGRESS-Plus framework, developed by the Cochrane-Campbell equity methods group, was designed to systematically identify and report on equity-relevant factors in health research [17]. It extends the original PROGRESS (place of residence, race and ethnicity, occupation, gender, religion, education, socioeconomic status, and social capital) acronym with "plus" dimensions, including age, disability, and language [17]. The framework was created to help researchers illuminate disparities that might otherwise be masked in aggregated data and has since been widely applied in clinical trials, systematic reviews, and digital health evaluations.

Alternative frameworks, such as the health equity impact assessment tool, as seen in Textbox 1, are used prospectively to anticipate equity impacts before interventions are deployed [52]. The SIITHIA (Strengthening the Integration of Intersectionality Theory in Health Inequality Analysis) checklist provides structured criteria for identifying inequities in digital health [53]. More recently, the digital health equity framework extends this approach to digital interventions and multidimensional analyses [54]. In practice, these can be operationalized by setting thresholds (eg, sensitivity gaps ≤5% between groups), with breaches triggering model review and corrective action. Complementing this, algorithmic fairness metrics, such as equal opportunity (equal TPRs), equalized odds (matching both TPR and FPR), and calibration integrity, are critical for measuring subgroup performance and detecting systematic bias. Without these frameworks, inequities may go undetected beneath aggregated performance.

Textbox 1. Frameworks for evaluating safety and equity in artificial intelligence triage.

Safety and sociotechnical performance

- Systems engineering initiative for patient safety describes how people, tasks, tools, and organizational context interact to shape patient safety [48].
- World Health Organization International Classification for Patient Safety provides a standardized taxonomy for reporting and classifying safety incidents [32,49,55].

Implementation and adoption

- Consolidated Framework for Implementation Research assesses organizational readiness and barriers to and facilitators of implementation.
- Nonadoption, abandonment, scale-up, spread, and sustainability [35,36] framework evaluates the complexity and long-term viability of digital health technologies.
- Reach, effectiveness, adoption, implementation, and maintenance framework evaluates long-term reach, effectiveness, adoption, and maintenance [56].
- Human, organization, and technology-fit framework examines alignment between human, organizational, and technical factors [57].

Equity and fairness

- PROGRESS-Plus identifies social stratifiers, such as age, ethnicity, language, deprivation, and disability [17].
- Health Equity Impact Assessment tool evaluates potential equity impacts before deployment [52].
- The SIITHIA (Strengthening the Integration of Intersectionality Theory in Health Inequality Analysis) checklist and the digital health equity framework support intersectional and digital-specific equity analysis [53,54].

XSL•FO
RenderX

The frameworks presented in Textbox 1 allow AI triage to be evaluated across safety, implementation, and equity dimensions. Combining them enables a more comprehensive assessment than any single lens can provide.

Textbox 2 provides an illustrative example of how to report equal opportunity (true positive rate; sensitivity) across intersectional PROGRESS-Plus strata.

Textbox 2. Illustrative example: reporting equal opportunity across intersectional strata.

- Equal opportunity requires similar true positive rates (TPRs; sensitivity) across groups for individuals who truly need urgent care.
- Step 1 includes defining the safety-critical outcome (Y=1), for example, "urgent same-day clinical assessment required" based on a reference standard (eg, clinician adjudication, emergency department attendance within 24 to 48 hours, or diagnosis of a time-critical condition).
- Step 2 involves computing TPR within each subgroup.
- TPR (sensitivity) = true positive (TP) / (TP + false negative [FN])
- Step 3 involves reporting TPR across PROGRESS-Plus strata and intersectional strata, for example, age group×ethnicity or ethnicity×deprivation quintile. An example of reporting is as follows:
- White, least deprived: TP=180; FN=20 → TPR=0.90
- White, most deprived: TP=150; FN=30 → TPR=0.83
- Minority ethnicity, least deprived: TP=70; FN=20 → TPR=0.78
- Minority ethnicity, most deprived: TP=55; FN=25 → TPR=0.69
- Step 3a involves quantifying uncertainty in subgroup estimates. For each subgroup, TPRs should be reported with measures of uncertainty (eg, 95% CIs or standard errors), particularly where subgroup sample sizes are small. This enables assessment of whether observed differences are robust or compatible with random variation.
- Step 4 summarizes the disparity. This approach makes equity risks visible that would be hidden in overall performance metrics.
- Absolute gap (maximum TPR – minimum TPR) = 0.90 – 0.69 = 0.21
- Flag threshold: investigate if the gap is greater than 0.05 or if any subgroup TPR is less than 0.80
- These figures are illustrative. In practice, equity assessments should consider statistical uncertainty (eg, CI overlap and subgroup sample size) alongside point estimates. Empirical data on equity-stratified artificial intelligence triage performance in primary care remain limited.

Several dimensions of bias have been documented or are anticipated in an AI-driven triage system, as mentioned subsequently.

The first dimension is age. Older adults are frequently underrepresented in development datasets, increasing the risk of atypical presentations or those with limited digital literacy [58,59].

The second dimension is language and ethnicity. NLP models are extremely sensitive to linguistic variation, dialects, multilingual input, or limited English proficiency. However, model evaluations rarely account for these, threatening safe triage in diverse populations [60].

Furthermore, broader AI research (outside primary care) shows that racial and ethnic biases in algorithmic systems persist [61]. For instance, algorithms that relied on health care cost as a proxy for illness systematically undertriaged Black patients due to unequal access-driven cost differences [60]. Similarly, AI in imaging frequently underdiagnoses emergent pathology in marginalized groups. Black women have shown significantly higher underdiagnosis rates in medical imaging models.

Telephone or digital triage evaluations suggest that low-income individuals, ethnic minority groups, and displaced patients experience worse outcomes, though quantitative data remain sparse [62].

Single-axis analysis (age or ethnicity) is insufficient. Intersecting vulnerabilities (eg, older adults from minority ethnic backgrounds with language barriers) can compound risk and increase mistriage. However, only a limited number of studies disaggregate safety performance by intersectional subgroups, leaving some of the most disadvantaged populations effectively invisible in assessments. Such analyses often lack power; therefore, findings should be treated as exploratory unless supported by large, multisite cohorts.

What Must Change: a Research and Policy Agenda

In this viewpoint paper, we argue that AI triage stands at a crossroad; it has the potential to improve safety and access in primary care but requires real-world evaluation, equity-focused monitoring, and sociotechnical governance. Figure 2 summarizes the proposed real-world evaluation and governance loop for AI triage in GP. Digital entry points feed into AI-supported triage and clinician workflow, producing safety and service-use outcomes that inform monitoring and governance (including equity dashboards, drift detection, and incident reporting). Governance outputs drive model and workflow updates, enabling continuous improvement.

XSL•FO
RenderX

Figure 2. Conceptual loop for evaluating and governing artificial intelligence (AI)–enabled triage in general practice as a sociotechnical intervention.

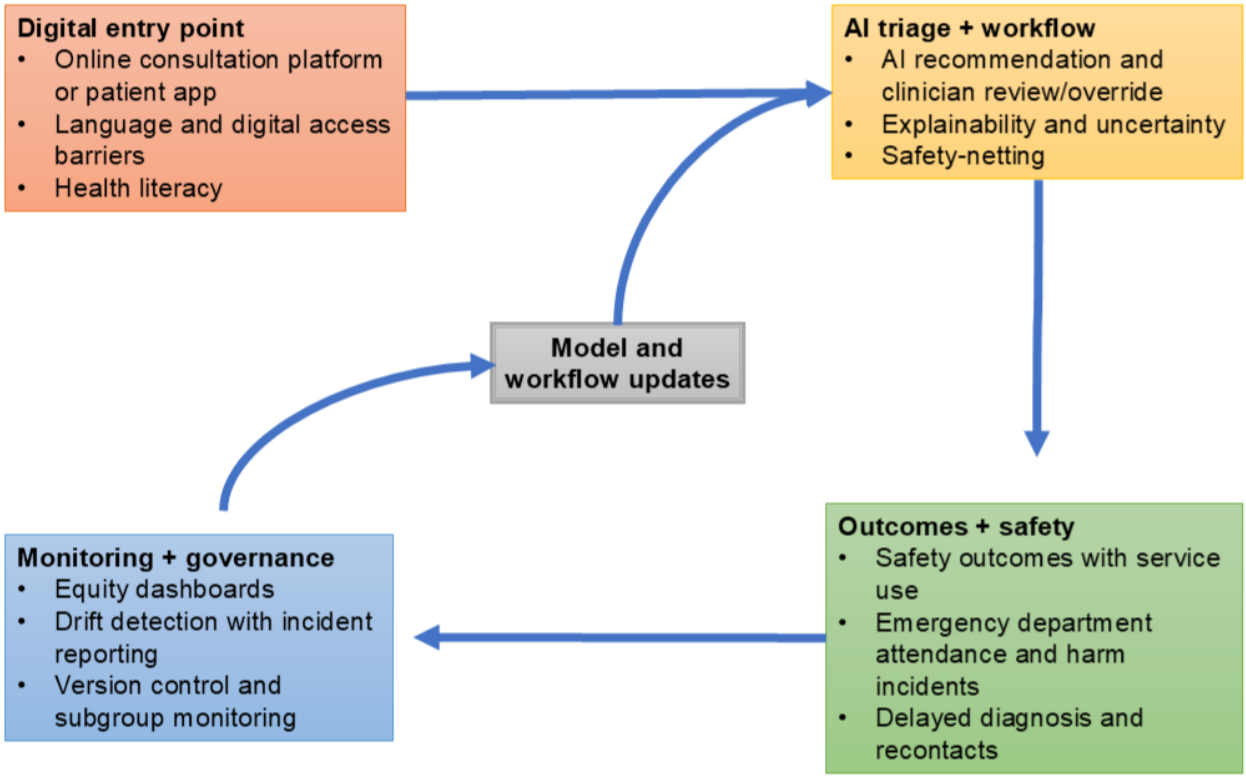

To shift AI triage from a hypothetical promise to an equitable, safe reality in GP and primary care settings, we propose 5 interrelated priorities, as mentioned subsequently.

The first priority is real-world evaluations in primary care (prospective or retrospective). Current evidence is dominated by vignette experiments or ED contexts. Prospective, real-world evaluations—such as controlled interrupted time series or (cluster) randomized controlled trials (RCTs)—that assess patient safety outcomes (eg, delayed diagnoses and avoidable emergency use), workflow effects, and override behaviors in GP practice are urgently needed. Where randomization is infeasible, controlled interrupted time series with matched controls and prespecified safety outcomes can provide strong quasi-experimental evidence.

The second priority is equity-stratified performance reporting. AI triage systems should be evaluated through the equity lenses, such as PROGRESS-Plus and other fairness metrics. This means disaggregating performance (TPR, FPR, and calibration) by age, ethnicity, language, deprivation, and their multidimensional aspects to identify and mitigate differential risks. Without such reporting, disparities will remain hidden.

The third priority is causal evaluation designs. Observational signals are helpful, but causality demands rigorous designs. Interrupted time series around AI deployment and RCTs, where feasible (to attribute safety effects more definitively to AI interventions), and propensity score methods, instrumental variables, or target trial emulation, where RCTs are infeasible, should be conducted.

The fourth priority is postmarket surveillance and governance infrastructure. Effective governance is not optional. Organizations should adopt frameworks such as people, process, technology, and operations to ensure structured oversight across personnel, process, technology, and operations. AI triage requires continuous monitoring, with statistically principled detection of drift, subgroup miscalibration, and emerging hazards [63].

The fifth priority is human-AI collaboration and implementation research. Research should shift from algorithm-centric evaluation to sociotechnical integration. Studies should examine how clinicians interpret, override, or trust AI suggestions; how AI supports (rather than disrupts) workflow; and how organizational culture shapes safe AI adoption. Mixed methods research combining qualitative insights with quantitative safety metrics will be critical.

Prospective evaluations in GP are challenging but feasible. Cluster-randomized trials and interrupted time series require careful handling of contamination between clinicians, cointerventions during rollout, seasonal and demand shocks, and variation in practice digital maturity. Outcome measurement also depends on data linkage (eg, EHR, urgent care, ED attendance, and diagnostic follow-up), and governance processes can slow implementation. Although such designs remain underused for AI-enabled triage, they are well-established in evaluating complex service interventions in primary care and are methodologically appropriate for this context [64,65].

Conclusions

AI triage offers potential for improving primary care efficiency, safety, and consistency, but current evidence leaves critical gaps. Without intersectional, real-world safety evaluations,

XSL•FO
RenderX

implementation is not just uncertain; it may be ethically risky and may inadvertently magnify existing health inequities. Over the coming years, this field must commit to responsible, equity-focused, system-aware evidence generation. That means embedding AI evaluation within the disordered realities of practice, governance mechanisms that ensure fairness and transparency, and human-AI systems that augment care rather than add workload. Operationally, this requires three commitments: (1) prospective, real-world evaluations in GP; (2) equity-stratified performance reporting guided by frameworks, such as PROGRESS-Plus; and (3) rigorous postmarket surveillance with drift and subgroup monitoring and WHO International Classification for Patient Safety–aligned incident reporting. Without these, deployment risks amplify inequities rather than reducing them. With these commitments, AI triage can better deliver on its potential of safer, more equitable primary care.

Acknowledgments

Generative artificial intelligence (ChatGPT, version 5.2; OpenAI) was used to support language editing for clarity and readability. All content was reviewed and revised by the authors, who take full responsibility for the final manuscript.

Funding

This work received no specific funding. EK is partly funded by the National Institute for Health and Care Research HealthTech Research Centre in Emergency and Acute Care (grant NIHR205301) and the Manchester British Heart Foundation Centre for Research Excellence (grant RE/24/130017).

Conflicts of Interest

None declared.

References

1. Sidaway-Lee K, Pereira Gray SD, Khan N, Abraham L, Evans P. GP continuity: the keystone of general practice. InnovAiT 2024 Apr 24;17(7):313-320. [doi: 10.1177/17557380241246742]
2. Hobbs FD, Bankhead C, Mukhtar T, Stevens S, Perera-Salazar R, Holt T, National Institute for Health Research School for Primary Care Research. Clinical workload in UK primary care: a retrospective analysis of 100 million consultations in England, 2007-14. Lancet 2016 Jun 04;387(10035):2323-2330 [FREE Full text] [doi: 10.1016/S0140-6736(16)00620-6] [Medline: 27059888]
3. The NHS long term plan. NHS England. 2019 Jan. URL: https://www.longtermplan.nhs.uk/ [accessed 2025-05-25]
4. Greenhalgh T, Wherton J, Shaw S, Morrison C. Video consultations for COVID-19. BMJ 2020 Mar 12;368:m998. [doi: 10.1136/bmj.m998] [Medline: 32165352]
5. Reidy C, Papoutsi C, Kc S, Gudgin B, Laverty AA, Greaves F, et al. Qualitative evaluation of the implementation and national roll-out of the NHS App in England. BMC Med 2025 Jan 21;23(1):20 [FREE Full text] [doi: 10.1186/s12916-024-03842-w] [Medline: 39838384]
6. Submissions via online consultation systems in general practice. NHS England. URL: https://digital.nhs.uk/data-and-information/publications/statistical/submissions-via-online-consultation-systems-in-general-practice [accessed 2026-01-24]
7. Parker RF, Figures EL, Paddison CA, Matheson JI, Blane DN, Ford JA. Inequalities in general practice remote consultations: a systematic review. BJGP Open 2021 Jun;5(3):040 [FREE Full text] [doi: 10.3399/BJGPO.2021.0040] [Medline: 33712502]
8. Chambers D, Cantrell AJ, Johnson M, Preston L, Baxter SK, Booth A, et al. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. BMJ Open 2019 Aug 01;9(8):e027743 [FREE Full text] [doi: 10.1136/bmjopen-2018-027743] [Medline: 31375610]
9. Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc 2003;10(6):523-530 [FREE Full text] [doi: 10.1197/jamia.M1370] [Medline: 12925543]
10. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 2005 Apr 02;330(7494):765 [FREE Full text] [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]
11. Technologies T. Technologies to address wait times in the emergency department. Health Technologies, National Library of Medicine 2025 Jul:EN0058 [FREE Full text] [Medline: 40857378]
12. Abualruz H, Yasin I, Abu Sabra MA, Abunab HY, Azayzeh R, Zubidi Y, et al. The role of artificial intelligence in enhancing triage decisions in healthcare settings: a systematic review. Appl Nurs Res 2025 Dec;86:152024. [doi: 10.1016/j.apnr.2025.152024] [Medline: 41330654]
13. Peta D, Day A, Lugari WS, Gorman V, Ahayalimudin N, Pajo VM. Triage: a global perspective. J Emerg Nurs 2023 Nov;49(6):814-825. [doi: 10.1016/j.jen.2023.08.004] [Medline: 37925222]

XSL•FO
RenderX

14. Tahernejad A, Sahebi A, Abadi AS, Safari M. Application of artificial intelligence in triage in emergencies and disasters: a systematic review. BMC Public Health 2024 Nov 18;24(1):3203 [FREE Full text] [doi: 10.1186/s12889-024-20447-3] [Medline: 39558305]
15. Singh H, Sittig DF. Advancing the science of measurement of diagnostic errors in healthcare: the Safer Dx framework. BMJ Qual Saf 2015 Feb;24(2):103-110 [FREE Full text] [doi: 10.1136/bmjqs-2014-003675] [Medline: 25589094]
16. Smart C, Newman C, Hartill L, Bunce S, McCormick J. Workload effects of online consultation implementation from a job-characteristics model perspective: a qualitative study. BJGP Open 2022 Nov 21;7(1):BJGPO.2022.0024. [doi: 10.3399/bjgpo.2022.0024]
17. O'Neill J, Tabish H, Welch V, Petticrew M, Pottie K, Clarke M, et al. Applying an equity lens to interventions: using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health. J Clin Epidemiol 2014 Jan;67(1):56-64. [doi: 10.1016/j.jclinepi.2013.08.005] [Medline: 24189091]
18. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan 7;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
19. Barocas S, Hardt M, Narayanan A. Fairness And Machine Learning: Limitations And Opportunities. Cambridge, MA: The MIT Press; 2023.
20. Akter S, Dwivedi YK, Sajib S, Biswas K, Bandara RJ, Michael K. Algorithmic bias in machine learning-based marketing models. J Bus Res 2022 May;144:201-216. [doi: 10.1016/j.jbusres.2022.01.083]
21. Jain A, Brooks JR, Alford CC, Chang CS, Mueller NM, Umscheid CA, et al. Awareness of racial and ethnic bias and potential solutions to address bias with use of health care algorithms. JAMA Health Forum 2023 Jun 02;4(6):e231197 [FREE Full text] [doi: 10.1001/jamahealthforum.2023.1197] [Medline: 37266959]
22. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci U S A 2020 Jun 09;117(23):12592-12594 [FREE Full text] [doi: 10.1073/pnas.1919012117] [Medline: 32457147]
23. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019 Oct;1(6):e271-e297 [FREE Full text] [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
24. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020 Mar 25;368:m689 [FREE Full text] [doi: 10.1136/bmj.m689] [Medline: 32213531]
25. Delshad S, Dontaraju VS, Chengat V. Artificial intelligence-based application provides accurate medical triage advice when compared to consensus decisions of healthcare providers. Cureus 2021 Aug;13(8):e16956 [FREE Full text] [doi: 10.7759/cureus.16956] [Medline: 34405077]
26. Ivanov O, Wolf L, Brecher D, Lewis E, Masek K, Montgomery K, et al. Improving ED emergency severity index acuity assignment using machine learning and clinical natural language processing. J Emerg Nurs 2021 Mar;47(2):265-78.e7 [FREE Full text] [doi: 10.1016/j.jen.2020.11.001] [Medline: 33358394]
27. Tyler S, Olis M, Aust N, Patel L, Simon L, Triantafyllidis C, et al. Use of artificial intelligence in triage in hospital emergency departments: a scoping review. Cureus 2024 May;16(5):e59906 [FREE Full text] [doi: 10.7759/cureus.59906] [Medline: 38854295]
28. Bernardi FA, Alves D, Crepaldi N, Yamada DB, Lima VC, Rijo R. Data quality in health research: integrative literature review. J Med Internet Res 2023 Oct 31;25:e41446 [FREE Full text] [doi: 10.2196/41446] [Medline: 37906223]
29. Wise J. Electronic consultations offer few benefits for GP practices, says study. BMJ 2017 Nov 06:j5141. [doi: 10.1136/bmj.j5141]
30. Karran EL, Cashin AG, Barker T, Boyd MA, Chiarotto A, Dewidar O, et al. Using PROGRESS-plus to identify current approaches to the collection and reporting of equity-relevant data: a scoping review. J Clin Epidemiol 2023 Nov;163:70-78 [FREE Full text] [doi: 10.1016/j.jclinepi.2023.09.017] [Medline: 37802205]
31. Steerling E, Svedberg P, Nilsen P, Siira E, Nygren J. Influences on trust in the use of AI-based triage-an interview study with primary healthcare professionals and patients in Sweden. Front Digit Health 2025 May 20;7:1565080 [FREE Full text] [doi: 10.3389/fdgth.2025.1565080] [Medline: 40463579]
32. Runciman W, Hibbert P, Thomson R, Van Der Schaaf T, Sherman H, Lewalle P. Towards an international classification for patient safety: key concepts and terms. Int J Qual Health Care 2009 Feb;21(1):18-26 [FREE Full text] [doi: 10.1093/intqhc/mzn057] [Medline: 19147597]
33. Mahlknecht A, Engl A, Piccoliori G, Wiedermann CJ. Supporting primary care through symptom checking artificial intelligence: a study of patient and physician attitudes in Italian general practice. BMC Prim Care 2023 Sep 04;24(1):174 [FREE Full text] [doi: 10.1186/s12875-023-02143-0] [Medline: 37661285]
34. Riboli-Sasco E, El-Osta A, Alaa A, Webber I, Karki M, El Asmar ML, et al. Triage and diagnostic accuracy of online symptom checkers: systematic review. J Med Internet Res 2023 Jun 02;25:e43803 [FREE Full text] [doi: 10.2196/43803] [Medline: 37266983]

XSL•FO
RenderX

35. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 07;4:50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
36. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
37. Tuzlalı M, Baki N, Aral K, Aral CA, Bahçe E. Evaluating the performance of AI chatbots in responding to dental implant FAQs: a comparative study. BMC Oral Health 2025 Oct 08;25(1):1548 [FREE Full text] [doi: 10.1186/s12903-025-06863-w] [Medline: 41063105]
38. Kaushik R, Rapaka R. A patient-centered perspectives and future directions in AI-powered teledentistry. Discoveries (Craiova) 2024;12(4):e199. [doi: 10.15190/d.2024.18] [Medline: 40109877]
39. Chen K, Qiu L, Xie X, Bai Y. Dynasmile: Video-based smile analysis software in orthodontics. SoftwareX 2025 Feb;29:102004. [doi: 10.1016/j.softx.2024.102004]
40. Abbas Q, Jeong W, Lee SW. Explainable AI in clinical decision support systems: a meta-analysis of methods, applications, and usability challenges. Healthcare (Basel) 2025 Aug 29;13(17):2154 [FREE Full text] [doi: 10.3390/healthcare13172154] [Medline: 40941506]
41. Abdelwanis M, Alarafati HK, Tammam MM, Simsekler MC. Exploring the risks of automation bias in healthcare artificial intelligence applications: a Bowtie analysis. J Saf Sci Resil 2024 Dec;5(4):460-469. [doi: 10.1016/j.jnlssr.2024.06.001]
42. Salimparsa M, Sedig K, Lizotte DJ, Abdullah SS, Chalabianloo N, Muanda FT. Explainable AI for clinical decision support systems: literature review, key gaps, and research synthesis. Informatics 2025 Oct 28;12(4):119. [doi: 10.3390/informatics12040119]
43. Ellertsson S, Hlynsson HD, Loftsson H, Sigur Sson EL. Triaging patients with artificial intelligence for respiratory symptoms in primary care to improve patient outcomes: a retrospective diagnostic accuracy study. Ann Fam Med 2023;21(3):240-248 [FREE Full text] [doi: 10.1370/afm.2970] [Medline: 37217331]
44. Vergara PO, de Oliveira JC, Mattiello R, Montelongo A, Roman R, Katz N, et al. Accuracy of artificial intelligence for gatekeeping in referrals to specialized care. JAMA Netw Open 2025 Jun 02;8(6):e2513285 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.13285] [Medline: 40459894]
45. Wallace W, Chan C, Chidambaram S, Hanna L, Acharya A, Daniels E, et al. Evaluating the diagnostic and triage performance of digital and online symptom checkers for the presentation of myocardial infarction: a retrospective cross-sectional study. PLOS Digit Health 2024 Aug;3(8):e0000558. [doi: 10.1371/journal.pdig.0000558] [Medline: 39102377]
46. Siira E, Tyskbo D, Nygren J. Healthcare leaders' experiences of implementing artificial intelligence for medical history-taking and triage in Swedish primary care: an interview study. BMC Prim Care 2024 Jul 24;25(1):268. [doi: 10.1186/s12875-024-02516-z] [Medline: 39048973]
47. Moschogianis S, Darley S, Coulson T, Peek N, Cheraghi-Sohi S, Brown B. Seven opportunities for artificial intelligence in primary care electronic visits: qualitative study of staff and patient views. Ann Fam Med 2025 May 27;23(3):214-222 [FREE Full text] [doi: 10.1370/afm.240292] [Medline: 40425478]
48. Holden RJ, Carayon P, Gurses AP, Hoonakker P, Hundt AS, Ozok AA, et al. SEIPS 2.0: a human factors framework for studying and improving the work of healthcare professionals and patients. Ergonomics 2013;56(11):1669-1686 [FREE Full text] [doi: 10.1080/00140139.2013.838643] [Medline: 24088063]
49. Patient safety: making health care safer. World Health Organization (WHO). 2017. URL: https://www.who.int/publications/i/item/WHO-HIS-SDS-2017.11 [accessed 2025-07-14]
50. Lobach D, Sanders GD, Bright TJ, Wong A, Dhurjati R, Bristow E, et al. Enabling health care decisionmaking through clinical decision support and knowledge management. Evid Rep Technol Assess (Full Rep) 2012 Apr(203):1-784. [Medline: 23126650]
51. Darley S, Coulson T, Peek N, Moschogianis S, van der Veer SN, Wong DC, et al. Understanding how the design and implementation of online consultations affect primary care quality: systematic review of evidence with recommendations for designers, providers, and researchers. J Med Internet Res 2022 Oct 24;24(10):e37436 [FREE Full text] [doi: 10.2196/37436] [Medline: 36279172]
52. Olyaeemanesh A, Takian A, Mostafavi H, Mobinizadeh M, Bakhtiari A, Yaftian F, et al. Health Equity Impact Assessment (HEIA) reporting tool: developing a checklist for policymakers. Int J Equity Health 2023 Nov 18;22(1):241 [FREE Full text] [doi: 10.1186/s12939-023-02031-0] [Medline: 37980523]
53. Government of Canada. How to integrate intersectionality theory in quantitative health equity analysis? A rapid review and checklist of promising practices. Public Health Agency of Canada. Ottawa, ON URL: https://www.canada.ca/en/public-health/services/publications/science-research-data/how-integrate-intersectionality-theory-quantitative-health-equity-analysis.html [accessed 2026-01-24]
54. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119 [FREE Full text] [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
55. Carayon P. Sociotechnical systems approach to healthcare quality and patient safety. Work 2012;41 Suppl 1(0 1):3850-3854 [FREE Full text] [doi: 10.3233/WOR-2012-0091-3850] [Medline: 22317309]

XSL•FO
RenderX

56. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
57. Yusof MM, Kuljis J, Papazafeiropoulou A, Stergioulas LK. An evaluation framework for health information systems: human, organization and technology-fit factors (HOT-fit). Int J Med Inform 2008 Jun;77(6):386-398. [doi: 10.1016/j.ijmedinf.2007.08.011] [Medline: 17964851]
58. Chu CH, Donato-Woodger S, Khan SS, Nyrup R, Leslie K, Lyn A, et al. Age-related bias and artificial intelligence: a scoping review. Humanit Soc Sci Commun 2023 Aug 17;10(1):510. [doi: 10.1057/s41599-023-01999-y]
59. Shiwani T, Relton S, Evans R, Kale A, Heaven A, Clegg A, Ageing Data Research Collaborative (Geridata) AI group, et al. New Horizons in artificial intelligence in the healthcare of older people. Age Ageing 2023 Dec 01;52(12):afad219 [FREE Full text] [doi: 10.1093/ageing/afad219] [Medline: 38124256]
60. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453 [FREE Full text] [doi: 10.1126/science.aax2342] [Medline: 31649194]
61. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019 Apr 04;380(14):1347-1358. [doi: 10.1056/nejmra1814259]
62. Williams C, Shang D. Telehealth usage among low-income racial and ethnic minority populations during the COVID-19 pandemic: retrospective observational study. J Med Internet Res 2023 May 12;25:e43604 [FREE Full text] [doi: 10.2196/43604] [Medline: 37171848]
63. Maleki Varnosfaderani S, Forouzanfar M. The role of AI in hospitals and clinics: transforming healthcare in the 21st century. Bioengineering (Basel) 2024 Mar 29;11(4):337 [FREE Full text] [doi: 10.3390/bioengineering11040337] [Medline: 38671759]
64. Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol 2017 Feb 01;46(1):348-355 [FREE Full text] [doi: 10.1093/ije/dyw098] [Medline: 27283160]
65. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M, Medical Research Council Guidance. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 2008 Sep 29;337:a1655 [FREE Full text] [doi: 10.1136/bmj.a1655] [Medline: 18824488]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
ED: emergency department
EHR: electronic health record
FPR: false-positive rate
GP: general practice
ML: machine learning
NLP: natural language processing
RCT: randomized controlled trial
SIITHIA: Strengthening the Integration of Intersectionality Theory in Health Inequality Analysis
TPR: true-positive rate
WHO: World Health Organization

Edited by A Mavragani; submitted 24.Nov.2025; peer-reviewed by KMS Islam, M Kokash, J Grosser, K Chen; comments to author 07.Jan.2026; accepted 28.Jan.2026; published 04.Mar.2026.

Please cite as:
Alamoudi A, Kontopantelis E, Zghebi SS, Brown B
AI Triage in Primary Care: Building Safer and More Equitable Real-World Evidence
J Med Internet Res 2026;28:e88396
URL: https://www.jmir.org/2026/1/e88396
doi:10.2196/88396
PMID:

©Aymn Alamoudi, Evangelos Kontopantelis, Salwa S Zghebi, Benjamin Brown. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN

XSL•FO
RenderX

1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Using an Adult-Designed Wearable for Pediatric Monitoring: Practical Tutorial and Application in School-Aged Children With Obesity

Alessandra Angelucci[1], PhD; Andrea Aliverti[1], PhD; Matteo Vandoni[2], PhD; Daniela Lucini[3,4], MD, PhD; Cristiana Larizza[5], PhD; Vittoria Carnevale Pellino[2], PhD; Gianvincenzo Zuccotti[6,7*], MD; Valeria Calcaterra[7,8*], MD

[1]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan, Italy

[2]Laboratory of Adapted Motor Activity (LAMA), Department of Public Health, Experimental Medicine and Forensic Science, University of Pavia, Pavia, Italy

[3]BIOMETRA Department, University of Milan, Milan, Italy

[4]Exercise Medicine Unit, Istituto di Ricovero e Cura a Carattere Scientifico Istituto Auxologico Italiano, Milan, Italy

[5]Department of Computer, Electrical and Biomedical Engineering, University of Pavia, Pavia, Italy

[6]Department of Biomedical and Clinical Science, University of Milan, Milan, Italy

[7]Pediatric Department, Buzzi Children's Hospital, Milan, Italy

[8]Department of Internal Medicine and Therapeutics, University of Pavia, Pavia, Italy

[*]these authors contributed equally

Corresponding Author:
Andrea Aliverti, PhD
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan, Italy

Abstract

This tutorial presents a step-by-step guide on how to use an adult-oriented wearable (Fitbit) to collect and analyze activity and cardiovascular data in a pediatric population of school-aged children with obesity. After outlining the current landscape of commercial wearables for adults and children, the tutorial illustrates the distinct considerations required for accurate pediatric monitoring, especially for cardiovascular metrics and derived features. The text provides a clinical application, highlighting how data from these devices were gathered and integrated with standard clinical measurements (ie, 1 week of monitoring with the wearable compared with the 6-minute walk test). The tutorial also discusses potential correlations, which should be interpreted as exploratory, given the small sample size (n=16), as well as limitations and future perspectives on using wearables for long-term pediatric monitoring of school-aged children, aiming to inform clinicians, researchers, and other stakeholders about the additional considerations that are needed to use wearables designed for adults to monitor this age group.

(J Med Internet Res 2026;28:e76166) doi:10.2196/76166

KEYWORDS

childhood obesity; wearable devices; activity monitoring; cardiovascular monitoring; wearable data quality

Introduction

Wearable Devices

Wearables are considered key elements of future health care because they enable ecological measurements without interfering with daily life. Feedback provided by consumer devices, especially those equipped with displays, can induce behavioral reactivity, potentially nudging toward active lifestyles [1]. Wearables find many applications in the field of telemonitoring of patients [2], consumer health and well-being, sports performance assessment (particularly in endurance disciplines and predominantly during training rather than competition [3]), environmental monitoring [4], and emergency settings [5]. Examples of wearables include watches (such as smartwatches and activity trackers), eyeglasses [6], visors [7], rings [8], patches [9,10], sensorized garments (such as shirts, gloves, and socks) [11], elastic bands applied to the torso [12,13], earbuds [14], and devices clipped to the shoes [15].

Even if it is often unclear to what extent nonmedical-grade commercial devices are validated, such devices have been proposed in the medical field as self-assessment tools [16,17] and may provide data associated with information of interest, such as activity levels and physiological parameters.

Activity metrics range from the number of steps performed in a fixed period [18], most commonly in a day, to a wide variety of parameters, such as distance, minutes in different heart rate zones (HRZ), and the number of stand hours [19].

Physiological measurements can consider a single parameter, for example, heart rate (HR) or pulse rate (PR) [20], peripheral

XSL•FO
RenderX

blood oxygen saturation [20], respiratory rate [21], skin temperature, and galvanic skin response or electrodermal activity, or could be more complex measures combining data from several sensors, such as in the case of sleep analysis, estimation of energy expenditure, and analysis of physiological parameters with respect to activity [22-25].

Devices for School-Aged Children

A first distinction must be made between wearable systems for neonates, infants, and school-aged children. Certain US Food and Drug Administration–cleared medical-grade patches, such as ANNE One (Sibel Health Inc) [26,27], have been successfully deployed for continuous vital sign monitoring in neonates and infants, particularly in intensive care settings [27]. These devices are designed for short-term clinical use under medical supervision and are optimized for the anatomical and physiological characteristics of newborns (eg, very small body size, high resting HR, thin skin, and limited mobility). However, such solutions are not directly applicable to ambulatory or home-based pediatric monitoring in school-aged children. Children exhibit larger body dimensions, mobility, and behavioral variability, which require wearables that are robust, reusable, and comfortable for daily use outside the hospital environment.

Another distinction should be made between devices for adults and children. Physiological monitoring in the school-aged child population is inherently more complex than in adults due to several factors. First, children have markedly different cardiovascular physiology; for instance, normal HRs in healthy children can reach upward of 180 beats per minute [28]. Second, anatomical scaling poses challenges in terms of device design, as wearables must be smaller and adapt to smaller vessels, thinner tissues, variable subcutaneous fat, and smaller chest circumference. Third, children's patterns of movement and physical activity are more dynamic and unpredictable, which increases motion artifacts, signal noise, and intermittent sensor loss [29]. Failure to monitor cardiovascular metrics continuously in children can delay early detection of hemodynamic compromise, mask subtle progression of dysfunction, and limit timely interventions, with consequences for morbidity and long-term outcomes [30]. The limitations of wearables for pediatric monitoring add up to the already existing limitations of wearable devices, even more so when commercial devices are involved.

In general, the results of studies about the accuracy and precision of commercial smartwatches used by adults are inconsistent. Regarding PR measurement from a Fitbit wrist-worn device, a study [31] found that the accuracy of PR estimation from a photoplethysmographic (PPG) sensor (<10% mean absolute error with respect to a gold-standard electrocardiogram) across 24 hours and during various activities was acceptable. Conversely, other studies [32] reported significant accuracy variations depending on the activity type, and a notable proportion of PR measurements were substantially inaccurate. Focusing on physical activity metrics, a 2015 study [33] found that Fitbit's step counts strongly correlated with those obtained by gold-standard methods. However, they observed low accuracy at slower walking speeds. Additionally, such results are valid in the adult population and cannot be immediately translated to pediatric applications.

It becomes even more complicated when children are involved. In fact, most research on the use of wearables in children is focused on activity tracking [34]. At the market level, popular models of wearables for children are limited in the number of integrated sensors and, therefore, in functionality. Most models only perform the function of activity trackers, such as the Fitbit Ace 3 [35] and the Garmin vívofit jr. 3 [36]. The Fitbit Ace 3, for instance, is only equipped with an accelerometer and not with a PPG sensor, thus making it not usable to assess cardiovascular features. Brands such as Apple and Polar do not offer any consumer product that is specifically designed for children. Polar used to offer the Polar Active [37] device with a business-to-business model for physical activity programs in school-aged children [38].

Nash et al [39] evaluated the accuracy of Apple Watch–derived cardiac features in children, both for the PPG and electrocardiography (ECG). In terms of PPG, they assessed whether PR estimates were accurate during both sinus rhythm and arrhythmias and found that there was a good agreement between the PPG and the gold-standard ECG during sinus rhythm (intraclass correlation coefficient [ICC] 0.98 - 0.99) and poor agreement during arrhythmias (ICC 0.24 - 0.27).

In a prospective cohort of children referred for 24-hour Holter monitoring (n=31, mean age 13.2, SD 3.6 y), a PPG-based device with the European Union Medical Device Regulation approval for use in adults, CardioWatch 287-2B (Corsano Health BV), achieved a mean accuracy of 84.8% (SD 8.7%) of readings; accuracy was significantly higher at lower HR (mean 90.9%, SD 9.3%) than at higher HR (mean 79.0%, SD 10.6%), and Bland–Altman analysis showed a bias of −1.4 bpm with 95% limits of agreement from −18.8 to +16.0 bpm [40].

In addition, ECG has been explored in smartwatches for pediatric applications. Unlike optical PPG-based wearables, the ScanWatch (Withings) uses a one-lead ECG approach. In a study of 100 children aged 5 to 17 years, the ICC for HR was 0.97, 0.86 for the PR interval, and 0.80 for the QT interval, whereas QRS and QTc interval agreement was only moderate [41].

Rationale and Aim of This Tutorial

The market for wearables for pediatric monitoring of school-aged children presents a gap in the offer and availability of monitors of cardiovascular function, as the most popular commercial wearable brands do not offer PPG-equipped (or ECG-equipped) devices for children or do not offer devices for children at all. It is nowadays necessary to use adult-designed wearables if there is a need for cardiovascular monitoring. The rationale for this tutorial focuses on the methodological adaptation required to use adult-grade devices equipped with PPG sensors and other integrated modules for cardiovascular monitoring in school-aged children. Dedicated child-specific devices with comparable sensing capabilities are largely unavailable, and even when they are marketed as child-specific devices, it is often not disclosed whether the changes are only aesthetic and in shape or also from a sensors and algorithms

XSL•FO
RenderX

point of view. This tutorial, therefore, highlights both the opportunities and the technical and physiological limitations of repurposing such devices for pediatric research and clinical applications. Rather than testing clinical hypotheses, the primary aim of this work is to demonstrate methodological feasibility by showing how an adult-designed wearable can be deployed, processed, and interpreted when used in school-aged children and by providing a transparent, reproducible workflow for researchers and clinicians. The clinical example included in the manuscript serves to illustrate the application of these methods, not to evaluate the effectiveness of the wearable or to draw population-level clinical inferences.

Step-by-Step Tutorial

Overview

The workflow is organized into 12 steps that readers can replicate in pediatric studies using adult-designed wearables. Each step describes what to do and how to do it, including defining the clinical context (steps 1 - 4), managing pediatric-specific reference values (step 5), selecting and importing Fitbit files (steps 6 and 7), extracting activity and cardiovascular metrics (steps 8 - 10), assessing data quality (step 11), and conducting statistical analysis (step 12).

Step 1: Framing the Clinical Case

The first step must be taken before collecting any data and consists of framing the clinical case that will be investigated by using adult-designed wearables in a pediatric population. This is a necessary step that guides the future choices of the parameters that will be retained or discarded and the clinical outcomes.

This tutorial includes a practical example based on data collected from children with obesity to demonstrate how wearable-derived metrics can be dealt with and compared with conventional clinical measures when adult-grade devices are used in children. Childhood obesity, defined according to the World Health Organization growth reference as a BMI greater than +2 SDs above the age- and sex-specific mean [42], represents a major global health concern [43]. BMI remains the most widely used indicator for classification, with equivalent percentile-based criteria also adopted by the Centers for Disease Control and Prevention [44]. By including this clinical case, the tutorial aims to illustrate how data extracted from an adult-designed wearable can be contextualized and analyzed in relation to standard clinical outcomes, such as the 6-minute walk test (6MWT) and blood pressure percentiles.

Exercise, as a nonpharmacological approach, can help delay obesity-related conditions, improve cardiovascular health, and regulate inflammation in both children and adults [45]. In addition, exercise promotes improvement in functional capacity, which is also reduced in the pediatric population with obesity [46].

Traditional tests and measures of functional capacity evaluate different aspects of physical activity and exercise tolerance; some are subjective (eg, evaluation scales), while others provide objective or quantitative data. Among the objective tests, the 6MWT [47] is considered a simple field-based approximation of submaximal exercise capacity, whereas cardiopulmonary exercise testing represents the gold standard for assessing maximal aerobic capacity but is expensive and time-consuming and thus reserved for selected cases [48,49]. In particular, the 6MWT evaluates exercise tolerance and consists of making the patient walk at a self-paced speed up and down a 30-m-long corridor for a fixed period of 6 minutes [47], and the total distance traveled in meters (6MWT distance or 6-minute walk test distance) is the primary result of this test.

Intermediate approaches such as exercise electrocardiogram stress testing [50] offer additional physiological insight but require clinical supervision. It is also worth noting that in nonclinical or educational contexts, such as school-based physical activity programs, functional capacity is often assessed using simpler field tests (eg, shuttle run or step tests), which provide practical but less accurate estimates of exercise tolerance than laboratory-based evaluations (eg, 6MWT and cardiopulmonary exercise testing) [51].

Step 2: Ethics Committee Approval

The second step consists of obtaining approval to collect data with adult-designed wearables in a cohort of children chosen based on the specific clinical case, which has already been defined in step 1. Approval from an ethics committee or institutional review board is required, and compliance with local laws on the privacy of minors is up to the investigators. To obtain this approval, it is necessary to outline the cohort and the experimental procedures to be followed. The same dataset is generally used for more than one research aim.

In this case, data were obtained from a cohort of children with obesity studied before and after a 12-week online exercise training program, the clinical outcomes of which were published by Calcaterra et al [52]. The study was approved by the institutional ethics committee (Milano Area 1 protocol number 2020/ST/298, approval date December 2, 2020); the study was performed in line with the principles of the Declaration of Helsinki. Written informed consent for participation and continuous wearable monitoring was obtained from the parents or legal guardians of all children. All data were pseudonymized before analysis, stored on secure institutional servers, and processed in full compliance with the General Data Protection Regulation (EU GDPR 2016/679). Given that the dataset includes health- and activity-related data from children (a vulnerable population), sharing individual-level data is subject to strict legal and ethical safeguards under the GDPR and the Italian Personal Data Protection Code. Owing to these restrictions and because this paper presents a tutorial and is not intended as a clinical validation, individual-level data will not be publicly shared.

Step 3: Experimental Data Collection

The third step consists of collecting experimental data, which come not only from the wearable device but also from different sources.

In this application, 2 device models (Fitbit Charge 2 and Charge 3) were used for wearable data collection; assignment to participants was determined by device availability and was not logged at the individual level.

XSL•FO
RenderX

Personal data used in this study included age (computed in months), sex, standing height, weight, BMI, and BMI *z* score (as explained later). A Harpenden stadiometer (Holtain Ltd) with a fixed vertical backboard and an adjustable headpiece was used to measure standing height. Weight was measured with the patients standing upright at the center of the scale platform (Seca) without shoes and wearing light clothing. They faced the recorder, kept their hands at their sides, and looked straight ahead.

The clinical outcomes evaluated in this study were the 6MWT and its *z* score with respect to the normative data reported by Geiger et al [53] and the percentiles of systolic arterial pressure (SAP) and diastolic arterial pressure (DAP). The latter two were measured in the supine position using an electronic mercury sphygmomanometer (A&D Medical) with an appropriately sized cuff on the right arm after 5 minutes of rest.

Step 4: Aim of the Research Work

As mentioned in step 2, the same dataset is generally used for more than one research aim. For this reason, a specific research project often requires further selection of the data collected that are needed to be retained for that specific aim.

Of the wearable data available in this study, only those regarding the first week of the program were selected and compared with clinical outcomes in the research work, so that the values could be considered as baseline because the effect of the exercise training program would be negligible after 1 week.

Conversely, this tutorial goes deeper into showing the opportunities that can be exploited by using the whole activity and cardiovascular dataset, without necessarily having used them in this specific research work. Additionally, the tutorial explains how to deal with anthropometric data in the case of children. Specifically, this tutorial is primarily intended for researchers, biomedical engineers, and clinicians interested in the scientific use of commercial wearables for pediatric monitoring, specifically for monitoring school-aged children. Its purpose is not to provide consumer guidance but to illustrate, step-by-step, how to extract, preprocess, and interpret wearable-derived physiological and activity data in a research context. By combining practical technical explanations (eg, data file structure and feature computation) with clinical interpretation, the tutorial aims to bridge the gap between data science and pediatric health care applications.

Step 5: *z* Score and Percentile Data in Children

In the case of pediatric patients, it must be highlighted that parameters are generally expressed relative to normative data for a given age and sex. This is mostly expressed in terms of *z* score or percentile. The fifth step consists of adjusting all relevant parameters with respect to normative data so that children of different ages and sexes can still be compared and analyzed together.

Age was used in this study to obtain relevant reference data for the computation of *z* scores and percentiles. When the mean of the sample μ and the SD σ are made available, equation (1) is used to compute the *z* score:

$$(1) z = x - \mu\sigma$$

Here, z denotes the *z* score, and x represents the observed value.

When μ is available, but not σ, z can be computed by knowing the SE of the mean and the numerosity of the sample n (ie, the number of people from which data were collected), equation (2) can be used instead:

$$(2) z = x - \mu SE \cdot n$$

To obtain the BMI *z* score, the BMI was then converted into its *z* scores using World Health Organization reference values [54]. In this scenario, because μ and σ are provided as reference values, equation (1) can be applied.

Resting PR (RPR) *z* scores are determined using national reference data from the 1999 - 2008 US population across all age groups [55]. In this case, μ, SE, and n are available, so equation (2) is used.

The blood pressure percentile was determined for each child following recent guidelines [56-58].

Step 6: Device Characteristics

The previously listed steps can be applied to data from all adult-designed wearables. The tutorial can also be used in the case of other devices collecting similar variables but requires more adaptations. The sixth step consists of knowing specific model characteristics and data formats, which may present slight changes even in the case of subsequent versions of the same device. It is always important to thoroughly consult all available documents and manuals before starting to analyze data.

Although some technical details of the considered devices are publicly available and downloadable from the company's website, a concise summary is included here to guide researchers in identifying which data files and parameters are most relevant for pediatric monitoring studies. The purpose of this section is not to reproduce the device manual but to provide practical instructions on how to locate, interpret, and preprocess key Fitbit outputs for subsequent scientific analysis. Although Fitbit labels its optical measurement as HR in device interfaces and documentation, this value is in fact a PR estimate derived from a peripheral PPG sensor; the use of the term "HR" reflects commercial and user-facing conventions rather than physiological accuracy. For this reason, throughout this tutorial, we refer to this signal as PR when we talk about the physiological variable, while we maintain the original nomenclature when we refer to the file or the Fitbit-computed parameter (eg, in the case of HRZ). We refer to PR over steps as PROS and to resting PR as RPR, as previously defined.

While some parameters—such as PR and step counts—are calculated directly by the Fitbit, other metrics can be derived or estimated from the downloadable data. This section is dedicated to the parameters that can be extracted directly from the Fitbit device.

Throughout this protocol, the Fitbit Charge 2 [59] and Charge 3 [60] were used interchangeably as stated before in the manuscript. These models were selected because they have been validated previously in several peer-reviewed studies assessing the accuracy of PPG-based PR and step count estimation [61,62], ensuring data comparability with existing literature.

XSL•FO
RenderX

Furthermore, they were the devices available to our research group at the time of data collection (in 2021), which allowed maintaining consistency across participants.

Although we used both the Fitbit Charge 2 and Charge 3 devices and did not perform a direct head-to-head comparison, existing independent validation studies show that the Charge 2 provides acceptable PR measurement versus ECG in 24-hour ecological recordings in adult populations (mean bias −3.47 bpm) [63], while the Charge 3 showed a larger bias (mean bias −7 bpm) but in a protocol focused on more dynamic activities when compared to a chest ‐ strap reference [64]. On this basis, one can assume a functional equivalence of the used models for the primary metrics (step count and PR), acknowledging that minor differences in sensor or firmware may exist. Newer Fitbit generations were not used to avoid potential discrepancies arising from firmware or algorithmic updates that could compromise data uniformity and comparability with validated models.

Both devices are equipped with a micro-electromechanical system 3-axis accelerometer to track motion patterns, an altimeter to track altitude changes, and an optical PR tracker based on PPG. As stated previously, in Fitbit documentation and app interfaces, this parameter is referred to as HR for commercial and communication purposes; however, as it is derived from peripheral PPG rather than ECG, it technically represents PR rather than HR [65].

Both devices contain a Bluetooth 4.0 radio transceiver to send data to a mobile app and can store 7 days of minute-by-minute statistics and 30 days of summary totals.

Step 7: Import Data

The seventh step consists of a first selection of the files to be imported or discarded and of importing the files of interest for the analysis.

Different files are available for analysis for each patient, all in .json format. Data structures vary between files: some contain time series of specific parameters with associated time stamps, others contain daily recaps, and still others contain weekly recaps. Synchronization between different files was performed using the time stamps in the files that were downloaded from the Fitbit device.

Before importing the data, the first essential filtering step is required. Any file whose content depends on age-based proprietary algorithms must be excluded because Fitbit profiles must be set to 18 years or older, and several derived metrics (eg, sleep stages, sleep score, caloric expenditure, and V˙O2 max–related fitness scores) are computed using adult-specific thresholds and models. For this reason, files related to sleep analysis, HRZ, calories, or fitness level should be discarded, as their values would be systematically biased in a pediatric cohort. More details regarding discarded metrics are reported in Figure 1.

Providing a fully reproducible data import script is not useful to readers because Fitbit intraday files do not follow a fixed header structure; field names differ across device models, firmware versions, application programming interface generations and types (eg, consumer vs research interface), and, more recently, Google-managed data export formats, requiring manual inspection and adjustment at import. More details on this can be found in the Data Import section of Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. Flowchart illustrating the process from all wearable data available from Fitbit to the list of retained variables and the variables obtained with further signal processing (highlighted in purple). The purple tab is dedicated to the additional processing pipeline, that is, to explain how additional variables can be obtained. HRZ: heart rate zone; PR: pulse rate; PROS: pulse rate over steps; RPR: resting pulse rate.

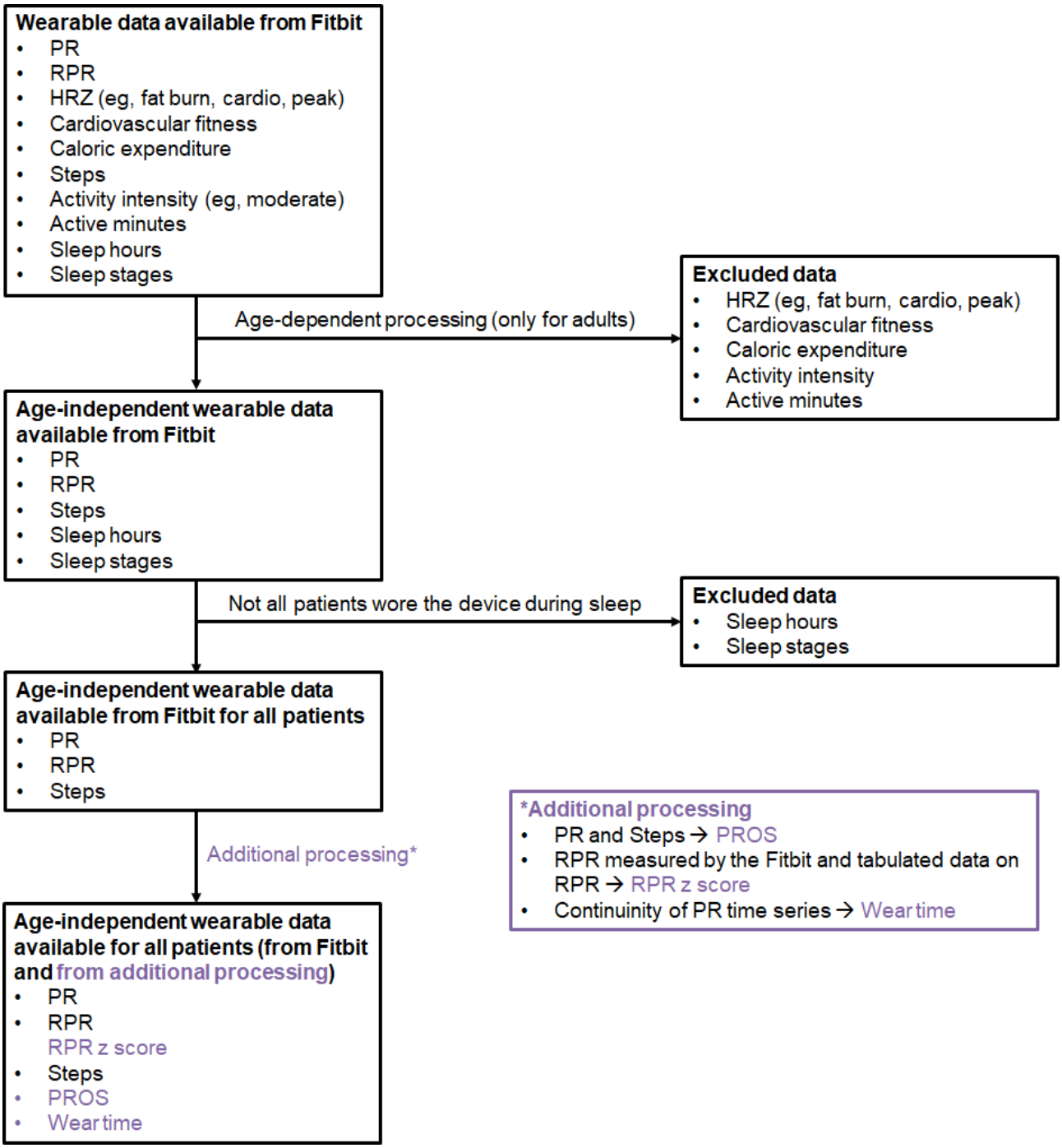

Step 8: Activity Metrics

The eighth step consists of exploring files with parameters related to the activity and selecting age-independent activity-related metrics.

"Activities" File and Derived

The "activities" file contains a daily summary of Fitbit activity data and goals, covering metrics such as steps, floors, active minutes, distance, and calories burned. The main fields include "activities," "goals," and "summary."

Specifically, the field "activities" is an array listing any detailed logged activities. If this array is empty, this indicates that there were no manually logged activities on that day.

The field "goals" describes the user's target goal for the day in terms of "activeMinutes" (target active minutes), "caloriesOut" (target calories to be burned), "distance" (target distance), "floors" (target floors), and "steps" (target steps).

The field "summary" reports the actual daily data, specifically "activeScore" (an internal scoring metric used by Fitbit), "activityCalories" (calories burned during active minutes), "caloriesBMR" (basal metabolic rate—calories burned at rest),

XSL•FO
RenderX

"caloriesOut" (total calories burned, including basal metabolic rate and activity), "distances" (an array of distance values broken down by activity levels), "elevation" (floors climbed in terms of elevation gain; may be 0 if not measured), "fairlyActiveMinutes" (minutes of moderate activity), "floors" (floors climbed), "heartRateZones" (an array showing the time spent in each HRZ—out of range, fat burn, cardio, and peak—along with estimated calories burned in each zone), "lightlyActiveMinutes" (minutes of light activity), "marginalCalories" (extra calories burned while being lightly active), "sedentaryMinutes" (minutes spent in a sedentary state), "steps" (total number of steps taken), and "veryActiveMinutes" (minutes of very active exercise).

All the fields regarding the count of calories ("activeCalories," "caloriesOut," "caloriesBMR," and "marginalCalories"), as well as the "calories" file, contain estimations of caloric consumption that are likely based on age, sex, weight, and height used as input in the Fitbit app. As the Fitbit device is designed for adults, it was only possible to set the age to 18 years or older to have the app working. For this reason, these metrics cannot be considered reliable in the present application, as already mentioned in step 7. The weekly recap files associated with these variables must also be excluded from the analysis.

The same can be said for the information related to the intensity of the activities based on HRZ values, which are still the subject of research in adults and are computed considering the maximum HR for a given age (which was set as 18 y in the application). In addition, in this case, the weekly recaps have been excluded.

For this reason, the only reliable metric is "steps." The number of walked steps is computed using the 3-axis accelerometer embedded in the Fitbit with dedicated processing algorithms. A traditional threshold of 10,000 daily steps is well accepted as indicative of an active lifestyle in adults [66]. To evaluate patient performance, one can consider either the mean daily steps or the maximum daily steps, representing the highest number of steps achieved during the test period. The latter is more representative of the ability to walk a certain distance [67].

The file "range_activities_steps" reports the total number of daily steps for a week. Visualizing this information, extracted either from the daily metrics or from the weekly recap files, can be used to immediately visualize behavioral shifts of patients. Figure 2 shows the distribution of daily steps in the 12 weeks of the intervention in the case of one patient. The higher step counts observed during weeks 7 to 9 likely coincide with the central weeks of August, when many families in Italy are on summer vacation, possibly leading to increased physical activity; however, this interpretation remains speculative.

Figure 2. Weekly steps distribution over the 12-wk online exercise training program (boy, aged 10 y; BMI=23.7 kg/m^2; BMI *z* score=1.74).

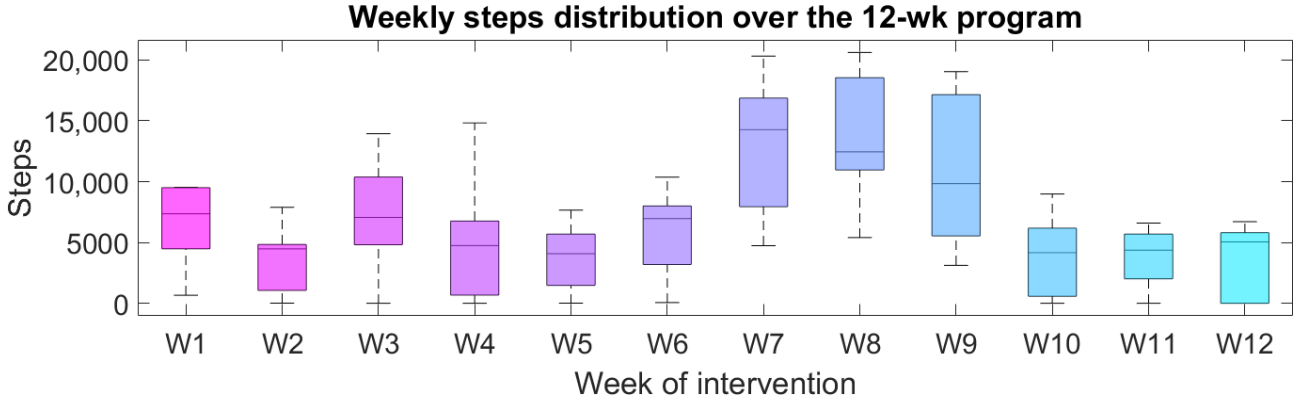

"Steps" File

Unlike the steps derived from the "activities" files family, the daily "steps" file shows the distribution of the steps during each day, in addition to the total number of steps. This allows to see when steps were taken during the day and better track the times of day when there has been the most activity. Examples of this can be seen in Figure 3.

In the considered clinical case example, there were 3 cases when an overall daily step count was provided, but no data on the distribution of the steps. This lack of detailed information on the distribution prevents the computation of other parameters (PROS, explained in a subsequent section).

XSL•FO
RenderX

Figure 3. Example of pulse rate (PR), steps, and pulse rate over steps (PROS) obtained in 1 wk of acquisitions (boy, aged 12 y; BMI=27.1 kg/m^2; BMI *z* score=1.86). Areas with the gray background correspond to night hours (22:00-06:59), while areas with the white background correspond to day hours (07:00-21:59). bpm: beats per minute.

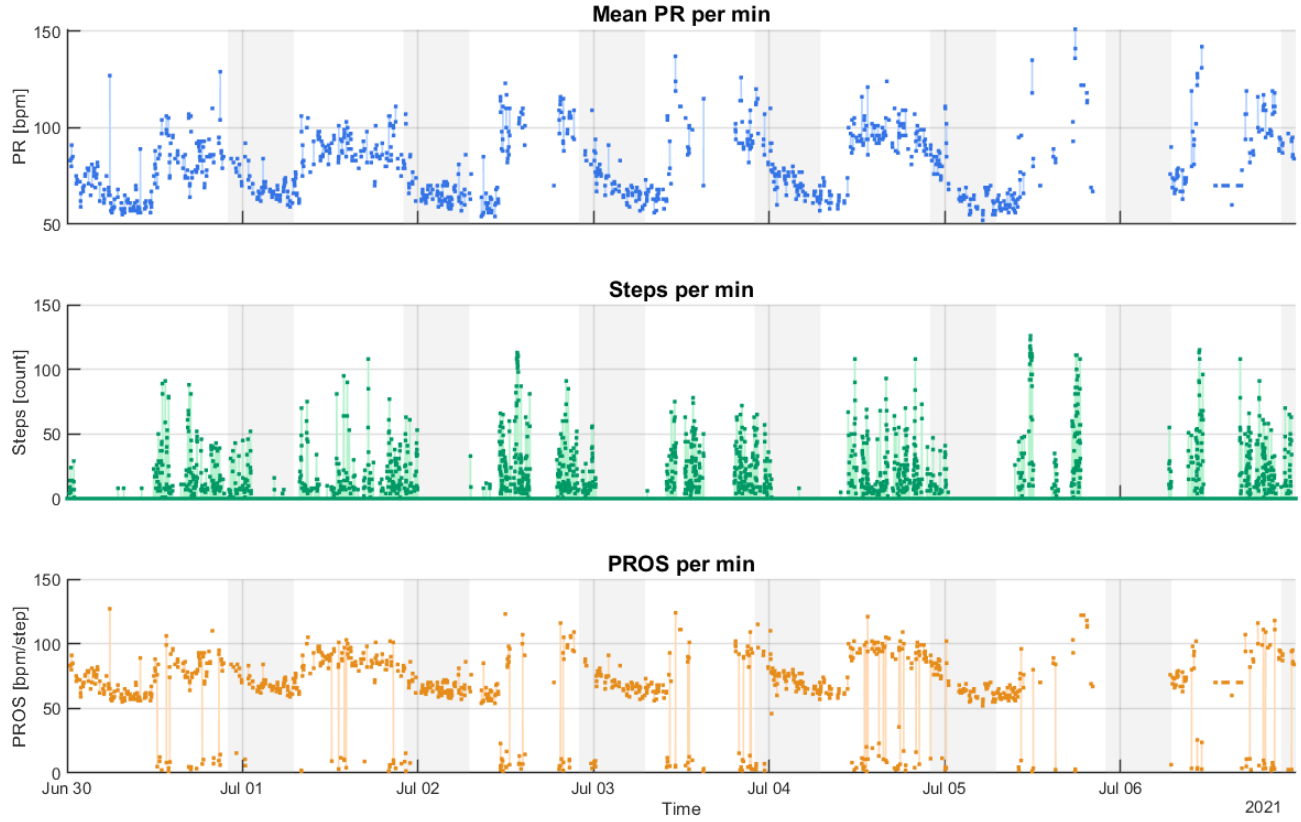

Step 9: Cardiovascular Metrics

The ninth step consists of exploring files with parameters related to the cardiovascular metrics and selecting age-independent metrics. As was mentioned earlier, some data, such as RPR, need to be adjusted for age and sex even if acquired from the wearable devices; in all relevant cases, step 5 should be repeated.

PR is continuously estimated via the embedded PPG sensor, with the raw data processed by an internal proprietary algorithm [67].

"hr" File

The daily "hr" file contains information regarding the already processed PR. There is an array ("activities-heart") containing the summary of the daily information, and then there is the "activities-heart-intraday" array with more details. This array has "dataset" as the primary subfield, which is an array of time-stamped entries; each entry has time and value. Additionally, there are 2 metadata fields, "datasetInterval" and "datasetType," which describe how the intraday data are structured.

PR measurement is affected by the location of the device and movement, as both factors can alter optical coupling, blood perfusion, and motion-induced artifacts, leading to transient inaccuracies or missing data—limitations common to all wrist-based PPG systems [68]. Furthermore, data are available at different sampling rates, with an interval in the order of seconds, one from the other at rest, and an increased sampling rate when exercise is detected. RPR is calculated over a 24-hour period and is known from the literature to reach its lowest value at night [69]. Consequently, nighttime RPR values are directly comparable across patients, whereas daytime values may not reflect the true minimum.

PR variability (PRV) is highly correlated with the HR variability, which quantifies the extent to which the R-R interval or HR changes from one cardiac cycle to the next. This measure can provide a lot of information about how the autonomic nervous system regulates the heart. HR variability has been correlated with cardiac diseases [70], fitness, and functional reserve [71] and is positively affected by exercise [72]. During our study, no PRV values were made available for download, nor could they be derived as no raw signal was given. As we reported in a previous study [67], PR data retrieved from a Fitbit device presented considerable gaps in the recordings, even for continuous recordings. As stated by the manufacturer, the normal sampling period should be between 5 and 15 seconds; however, in the results of that study, lack of data for 15 seconds or more was common. This makes it challenging to compute PRV reliably, so this metric should be analyzed with caution if obtained from this type of device.

Fitbit Charge 2 and 3 also provide information regarding HRZ and consequently the number of minutes spent in each zone. In adults, these data can be used to estimate the physical activity level of the person. However, in the present pediatric application, their reliability cannot be verified because the underlying proprietary algorithms were developed and validated on adult populations, and their exact computation methods are not publicly available.

XSL•FO
RenderX

Step 10: Combining Activity and Cardiovascular Metrics

The tenth step consists of combining activity and cardiovascular metrics. The parameters computed in this work were selected based on the clinical case and research aims, but other parameters can be considered as well.

Combining the metrics obtained from the activity and the cardiovascular domain is possible, and one example is reported in this tutorial. Mishra et al [73] used a Fitbit device to detect SARS-CoV-2 infection before symptom onset. To achieve this, the authors introduced and calculated the HR over steps index to identify potential anomalies in cardiac response relative to physical activity. In this study, this index is referred to as PROS to maintain homogeneity with the rest of the nomenclature.

PROS is computed as the instantaneous ratio between PR and the number of steps recorded within the corresponding time window. As PR is sampled approximately every 5 seconds and steps are aggregated per minute, each minute-level step count was matched to all PR values recorded within that same minute. To avoid division by zero during sedentary periods, 1 step was added to the denominator. The resulting value can therefore be expressed as in equation (3):

$$\text{PROS}(t)=\frac{\text{PR}(t)}{\text{steps}(t)+1} \quad (3)$$

where PR(t) represents PR at time t (beats per minute) and steps(t) is the number of steps detected within that minute. PROS provides a minute-by-minute estimate of the patient's cardiovascular effort and functional capacity. A high PROS value indicates a rapid heartbeat during moderate physical activity, while a low value suggests a restrained heartbeat despite challenging physical activity.

Mishra et al [73] developed a structured algorithm for computing PROS and identifying anomalous values, leveraging a dispersion-based approach that detects deviations from a normal distribution using the elliptic envelope method. The code used to process the data was retrieved from the GitHub repository associated with the original study. The anomalies are detected using a contamination of 0.1, where the contamination means the number of outliers in the dataset. In this manner, the algorithm identifies some outliers in the computed feature, corresponding to approximately 10% of the total. Following the method in the original paper, an iterative approach was used, incrementally increasing contamination values until a stable number of outliers was identified.

A possibility is to evaluate PROS at low quantiles, such as the 1% or the 5% quantile, to account for the possibility that the minimum PROS is an outlier.

As was mentioned before in the section dedicated to steps, it was not possible to compute PROS in 3 patients, as Fitbit provided the overall step count but not the details of the distribution of the steps during the day.

Step 11: Data Quality Assessment

The eleventh step of this tutorial consists of assessing the quality of the collected data, especially in terms of completeness of the dataset. Wearable data collection is unsupervised, and for this reason, patient adherence to the data collection protocol, that is, wear time, must be carefully evaluated in each study. In addition, signal loss is an important parameter to be evaluated, as explained later in the section.

To assess signal completeness and compliance, wear time can be quantified based on PR data continuity. In this dataset, the time intervals between consecutive PR samples were computed for each participant; intervals of 2 minutes or less were considered "worn," while longer gaps indicated device removal. Wear time was then expressed as the total duration of "worn" periods divided by the total duration of the analyzed time window. The 2-minute PR gap threshold was selected as a conservative compromise, as short interruptions of a few seconds are common even when the device is correctly worn. Adult Fitbit studies have proposed different definitions: Dominick et al [74] considered 60 minutes or more of missing signal as nonwear; Orstad et al [75] allowed up to 90 minutes of interruption to consider a day valid (with ≥10 h of recording including the interruption); Claudel et al [76] used minute-level PR gaps to classify nonwear episodes. Given the higher motion and frequent repositioning typical of children, the 2-minute criterion minimizes false nonwear detection while maintaining sensitivity to genuine removal periods.

To quantify how this threshold influenced data retention, it is advised to perform a sensitivity analysis. We performed such calculations using all 16 participants and recalculated wear time with nonwear gap thresholds of 60, 90, 120, 180, and 300 seconds. Weekly wear time changed only minimally across thresholds (60 s: 100.8 h; 90 s: 100.9 h; 120 s: 101.1 h; 180 s: 101.4 h; and 300 s: 101.8 h), with differences never exceeding 1.0 hour per week on average. Similarly, the percentage of days achieving 70% or greater completeness remained constant at 41% for all thresholds. These results indicate that the 120-second rule does not artificially inflate wear time estimates and that more permissive thresholds (eg, 180‐300 s) offer no meaningful advantage in terms of data retention in this cohort. Day and night were defined as 07:00 to 21:59 and 22:00 to 06:59, respectively, to align with typical waking and sleeping schedules in school-aged children. The theoretical acquisition window corresponded to 168 hours per week, split into approximately 105 hours daytime and 63 hours nighttime.

In addition to wear time, a second essential component of data quality is the assessment of signal loss, which reflects gaps in acquisitions even when the device is being worn. Unlike nonwear episodes, signal loss is typically caused by transient motion artifacts, insufficient optical coupling, or abrupt changes in perfusion—phenomena that are especially common in children due to rapid, irregular movements. To quantify signal loss, it is advised to examine the distribution of intersample intervals across the PR time series and identify gaps that exceed the expected sampling pattern (typically 5‐15 s at rest, with denser sampling during activity). These gaps should be flagged as within-wear missing segments, which differ clinically and analytically from nonwear periods. For example, brief clusters of missing samples may be acceptable when computing aggregated daily features (eg, daily steps and mean PR) but can severely affect features that rely on continuous sampling, such as PROS, nocturnal minima, or any attempt to estimate PR

XSL•FO
RenderX

variability. Figure 1 reports a flow diagram of the Fitbit-derived metrics retained or excluded from the analysis, the reasons for exclusion, a list of the metrics that have been computed with additional processing, and a tab dedicated to how additional processing is performed for each variable.

Step 12: Statistical Analysis

The final step is the statistical analysis and is again generalizable like steps 1 to 5. First, traditional statistical metrics such as mean and SD should be computed for each parameter, possibly further divided into subpopulations as done in Table 1, to characterize the population. Then, further steps of statistical analysis must be tailored to the specifics of the study.

In this clinical case, correlation analyses were performed to obtain the Pearson correlation coefficient ρ, the coefficient of determination R^2, and the P value; then a multiple linear regression analysis [77] was conducted to predict the 6 clinical outcomes listed in Textbox 1. Both the raw resting RPR and its age- and sex-adjusted z score were included as predictors in the regression models. Although these variables are mathematically related, they were retained together to explore whether absolute RPR and its standardized deviation from pediatric norms contribute differently to the prediction of functional outcomes. This approach was adopted intentionally to capture both absolute physiological values and relative, age-adjusted variability, acknowledging that some degree of correlation between the two variables may occur.

To mitigate overfitting due to the small sample size, an additional model selection analysis was conducted using leave-one-out cross-validation (LOOCV) combined with least absolute shrinkage and selection operator (LASSO) regularization. This approach automatically shrinks or removes redundant predictors, allowing a more parsimonious set of variables to emerge for each clinical outcome. The models were then refitted using only the selected predictors to verify their explanatory power.

Table . Summary of demographic, cardiovascular, and physical activity metrics of the whole cohort (N=16) and then divided by boys (n_{boys}=11) and girls (n_{girls}=5).

Parameter	Overall (N=16), mean (SD)	Boys (n_{boys}=11), mean (SD)	Girls (n_{girls}=5), mean (SD)
Age (y)	10.2 (1.7)	10.2 (1.3)	10.2 (2.6)
BMI (kg/m²)	28.1 (3.7)	27.1 (3.0)	30.2 (4.5)
BMI z score	2.2 (0.3)	2.1 (0.3)	2.4 (0.4)
RPR[a] (bpm)	79.4 (9.4)	79.3 (8.9)	79.5 (11.3)
RPR z score	−0.06 (0.45)	−0.03 (0.47)	−0.15 (0.42)
SAP[b] (mm Hg)	111.7 (11.1)	109.9 (12.6)	116.5 (2.4)
SAP percentile (%)	76 (27)	72 (30)	88 (8)
DAP[c] (mm Hg)	73.1 (8)	71.5 (8)	77.5 (7)
DAP percentile (%)	82 (16)	79 (17)	91 (10)
Daily steps	7742 (3025)	7838 (3629)	7530 (1148)
Maximum daily steps	12,895 (4964)	13,279 (5873)	12,050 (2210)
6MWT[d] distance (m)	490.5 (55.0)	487.2 (55.6)	517.0 (50.4)
6MWT z score	−2.7 (1.1)	−3.0 (0.9)	−2.0 (1.3)
Total wear time (weekly) (h)	101.1 (44.6)	105.5 (43.8)	91.4 (49.8)
Total wear time (weekly) (%)	60.2 (26.5)	62.8 (26.1)	54.4 (29.6)
Day wear time (weekly) (h)	72.3 (27.3)	73.8 (22.9)	68.8 (38.2)
Day wear time (weekly) (%)	68.8 (26.0)	70.3 (21.8)	65.5 (36.4)
Night wear time (weekly) (h)	28.8 (23.0)	31.6 (23.6)	22.6 (22.8)
Night wear time (weekly) (%)	45.7 (36.5)	50.2 (37.4)	35.9 (36.2)

[a]RPR: resting pulse rate

[b]SAP: systolic arterial pressure

[c]DAP: diastolic arterial pressure

[d]6MWT: 6-minute walk test

XSL•FO
RenderX

Textbox 1. Input data used for the analysis and the clinical outcomes of the presented clinical application.

Input data

- Personal data (anthropometric)
  - Age
  - BMI
  - BMI *z* score
- Wearable data (week 1)
  - 6-minute walk test
  - 6-minute walk test *z* score
  - Systolic arterial pressure
  - Systolic arterial pressure percentile
  - Diastolic arterial pressure
  - Diastolic arterial pressure percentile

Clinical outcomes

- Mean daily steps
- Maximum daily steps
- Resting pulse rate
- Resting pulse rate *z* score
- Pulse rate over steps

Results of the Considered Clinical Case

Population Characteristics

A summary of demographic, cardiovascular, and physical activity metrics for the cohort is presented in Table 1, including data divided by boys and girls. These values are expressed in terms of mean (SD).

Quantitative wear time statistics per participant can be reported in greater detail, as proposed in Table S1 in Multimedia Appendix 1.

Collected Data

An example of time series plots of PR, steps, and PROS obtained in a week of acquisitions is shown in Figure 3.

To further illustrate the temporal pattern of PR recordings and the occurrence of signal loss, Figure 4 shows a representative 15-minute segment extracted from the same week of acquisition plotted in Figure 3. This high-resolution view highlights the irregular sampling of PR data and the short interruptions occurring when the signal was not acquired by the Fitbit device.

XSL•FO
RenderX

Figure 4. Representative 15-min window of PR data from the same participant as in Figure 3 (boy, aged 12 y; BMI=27.1 kg/m^2, BMI *z* score=1.86). Blue dots represent individual PR samples, while light orange shaded areas indicate missing intervals where the time between consecutive samples exceeded 1.5×the expected sampling period. bpm: beats per minute.

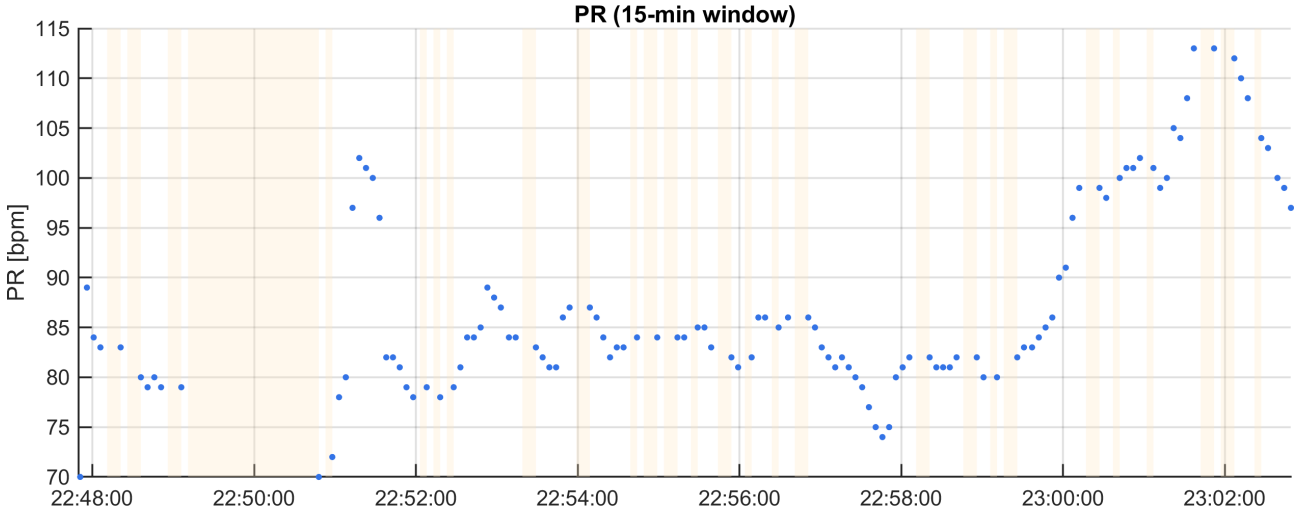

Statistical Analysis

Table 2 reports Pearson correlation coefficients (ρ) to explore the relationships among wearable data and clinical outcomes.

It is noteworthy that RPR *z* score has a significant correlation with 6MWT results (ρ=−0.53; R^2=0.28; *P*=.04), and it is a negative correlation as is expected from physiology. These results directly show the utility of wearable-derived data in clinical outcome assessment.

Two additional statistically significant correlations are observed between maximum weekly steps and the 6MWT (ρ=−0.52; R^2=0.27; *P*=.04) and between maximum weekly steps and the 6MWT *z* score (ρ=−0.60; R^2=0.36; *P*=.01); however, such results are not what one would intuitively expect, nor do they confirm what was obtained in a population of older patients [67]; therefore, it is likely a spurious correlation in both cases, as is discussed in more detail later in the paper.

The mean daily steps showed a moderate negative correlation with the DAP percentile (ρ=−0.42; R^2=0.18; *P*=.12) but did not reach statistical significance.

PROS displayed only weak correlations with functional capacity measures, indicating that this variable alone might have limited predictive value in this context. RPR and RPR *z* score showed a moderate, nonsignificant correlation with DAP (ρ=0.41; R^2=0.17; *P*=.13; and ρ=0.43; R^2=0.19; *P*=.12, respectively), and the RPR *z* score also showed a moderate correlation with DAP percentile (ρ=0.42; R^2=0.17; *P*=.14).

Multiple linear regression models were constructed to examine the associations between wearable data and clinical outcomes. Figure 5 shows the observed and the fitted values of each of the variables, with statistical significance defined as *P*<.05.

with The regression model for fitting 6MWT distance yielded an adjusted R^2 of 0.51, indicating a moderate fit of the model; the regression model fitting the 6MWT *z* score had a stronger fit (adjusted R^2=0.79), and, overall, the best performance among the 6 clinical outcomes considered in the study. In this case, 2 variables, that is, RPR and its *z* score, made a statistically significant contribution to the model (*P*=.03 and *P*=.02, respectively), indicating the strong dependence of the 6MWT *z* score on parameters relating to cardiac activity.

SAP, SAP percentile, DAP, and DAP performed poorly in the case of adjusted R^2, with the value being negative because there are too many variables. Given the small cohort size (n=16), all regression models should be considered underpowered; therefore, their estimates are intended to be illustrative of the methodological workflow and not to provide confirmatory inference. Still, all predictors in the full multivariable models, with coefficient, SE, *P* value, R^2, adjusted R^2, and n, are reported for each outcome in Table S2 of Multimedia Appendix 1.

The LASSO procedure identified age, BMI, maximum daily steps, and RPR *z* score as predictors of the 6MWT distance (adjusted R^2=0.29) and age, BMI *z* score, mean and maximum daily steps, RPR, RPR *z* score, and PROS for the 6MWT *z* score (adjusted R^2=0.75). Details on the predictors selected by LASSO with LOOCV for each outcome are reported in Table S3 of Multimedia Appendix 1. For blood pressure–related outcomes, the regularization excluded most predictors, retaining only maximum daily steps for systolic measures and BMI for diastolic pressure, both with negligible explanatory value (adjusted R^2≈0).

XSL•FO
RenderX

Table . Correlation of wearable data with clinical outcomes.

Variable	Clinical outcomes																	
	SAP[a]			SAP percentile			DAP[b]			DAP percentile			6MWT[c]			6MWT z score		
	ρ	R^2	*P* value	ρ	R^2	*P* value	ρ	R^2	*P* value	ρ	R^2	*P* value	ρ	R^2	*P* value	ρ	R^2	*P* value
Wearable data																		
Mean steps	−0.04	0.00	.90	−0.13	0.02	.64	−0.22	0.05	.44	−0.42	0.18	.12	0.07	0.01	.79	0.05	0.00	.85
Maximum steps	0.12	0.01	.67	0.14	0.02	.63	0.20	0.04	.47	0.16	0.03	.56	−0.52	0.27	*.04*[d]	−0.60	0.36	.01[d]
RPR[e]	0.15	0.03	.57	0.12	0.01	.67	0.41	0.17	.13	0.35	.12	.21	−0.38	0.14	.15	0.03	0.00	.90
RPR z score	0.24	0.06	.40	0.23	0.06	.44	0.43	0.19	.12	0.42	0.17	.14	*−0.53*	*0.28*	*.04*[d]	−0.34	0.12	.21
PROS[f]	−0.14	0.02	.66	−0.17	0.03	.59	0.08	0.01	.81	0.20	0.04	.52	0.20	0.04	.52	0.21	0.04	.49

[a]SAP: systolic arterial pressure.

[b]DAP: diastolic arterial pressure.

[c]6MWT: 6-min walk test.

[d]Statistically significant correlations.

[e]RPR: resting pulse rate.

[f]PROS: pulse rate over steps.

These results confirm that, in this limited cohort, 6MWT and 6MWT *z* score remain the only variables moderately predictable from wearable-derived metrics, while SAP, SAP percentile, DAP, and DAP percentile show minimal dependence on these predictors. The analysis demonstrates that the negative adjusted R^2 values observed in the unregularized multiple regressions shown in Figure 5 are primarily due to sample size limitations rather than methodological flaws.

XSL•FO
RenderX

Figure 5. Observed versus fitted clinical outcomes (6-min walk test [6MWT], 6MWT *z* score, systolic arterial pressure [SAP], SAP percentile, diastolic arterial pressure [DAP], and DAP percentile). The scatter plot illustrates the relationship between observed and fitted distances, with the red dashed line indicating an ideal fit.

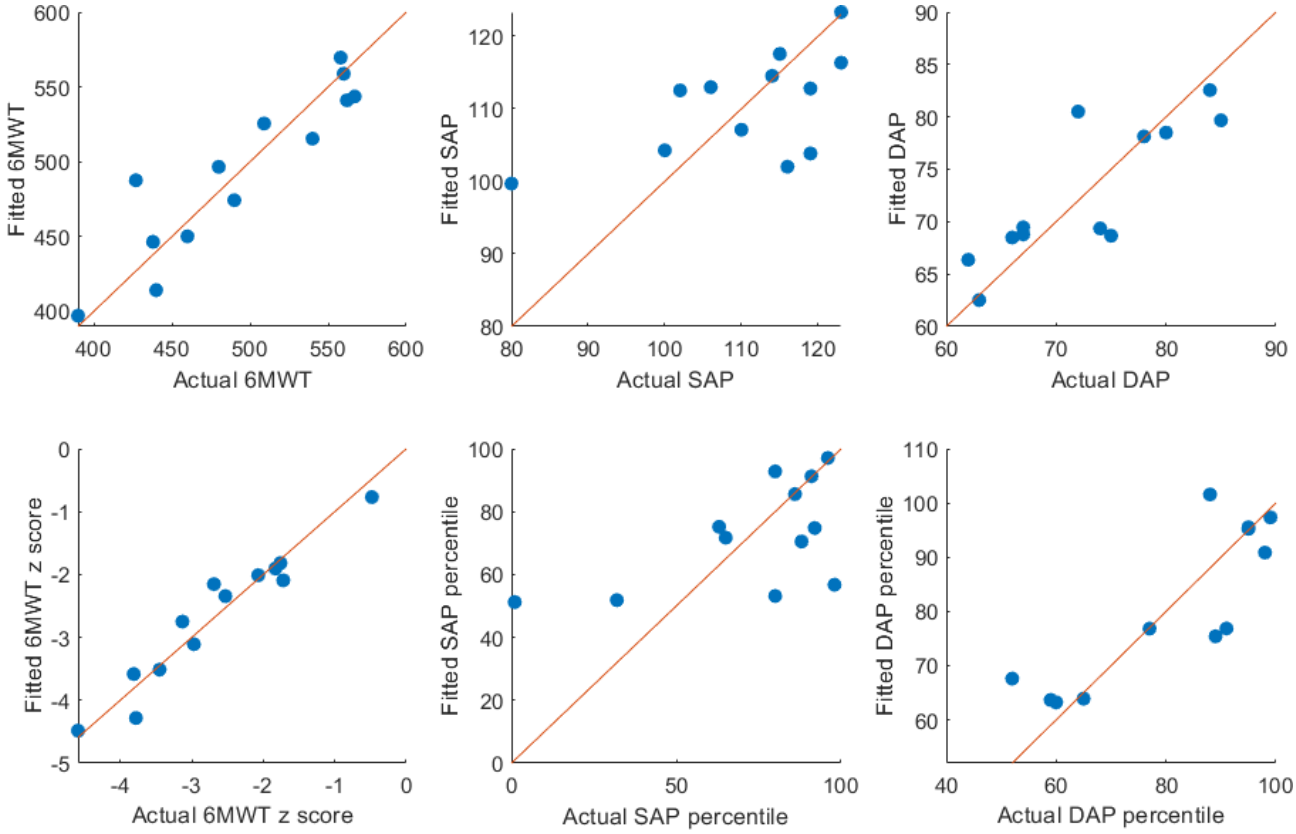

Discussion

This tutorial explains step by step how to process and further engineer activity and cardiovascular data obtained from Fitbit when used in pediatric patients and presents a clinical application in children with obesity.

Study-Specific Findings

The data gathered from Fitbit Charge 2 and Charge 3 in children with obesity demonstrate that consumer-grade wearable devices can provide clinically relevant information about physical activity (eg, mean or maximum daily steps) and cardiovascular function (eg, RPR and PROS) in this population. As shown, these metrics alone do not exhibit strong correlations with standard reference tests (eg, the 6MWT in children or systolic or diastolic blood pressure percentiles). However, by supplementing them with anthropometric indicators and appropriate pediatric reference values, a more robust framework can be built for understanding functional capacity and exercise tolerance. Such results are also expected to improve with larger datasets available. Among the observed statistically significant associations, there were inverse correlations between maximum daily steps and both the 6MWT distance and its *z* score. These associations should be interpreted as exploratory, as the analysis is constrained by the small sample size (n=16), which limits statistical power and the generalizability of the findings. Additionally, the counterintuitive finding of having inverse correlations may have several explanations. First, given the small sample size and multiple comparisons, a spurious correlation cannot be excluded. Second, the 6MWT reflects submaximal exercise tolerance measured under clinical supervision, whereas maximum daily step counts reflect spontaneous physical activity in free-living conditions. Children with lower functional capacity may have been more engaged to increase their daily steps during the monitoring week, potentially due to parental encouragement after clinical evaluation. Measurement variability may also contribute: step count accuracy in children can be affected by shorter stride length, irregular gait, and differences in device placement, leading to over- or under-counting at certain speeds. Additionally, the week chosen for wearable data collection may not coincide with typical activity patterns—for example, periods of school holidays or illness could bias step count data independently of actual fitness.

Overall, our findings indicate that consumer-grade wearable devices, such as the Fitbit Charge 2 and Charge 3, can provide meaningful information on both physical activity and cardiovascular parameters in children with obesity. These results suggest a potential role for such devices as complementary tools in clinical and research settings to monitor behavior and physiological responses outside the laboratory environment, hence the need to provide practical guidelines and indications on possible data processing pipelines.

An additional limitation of our work stems from the fact that the Fitbit app enforces a minimum user age of 18 years or older, which inevitably introduces bias in the estimation of HRZ and caloric output and led us to discard such parameters. As the algorithmic assignment of HRZ is age-based (often through formulas such as “220 − age”), pediatric users are forced to

XSL•FO
RenderX

adopt adult parameterization. Consequently, derived metrics such as time in "cardio" or "peak" zones and estimated energy expenditure are likely to be inaccurate when applied to cohorts of school-aged children. Further complicating matters, children's basal metabolic rates and activity energy expenditure differ systematically from adult models (eg, children exhibit relatively higher mass-specific metabolic rates) [78]. Thus, caloric estimates based on adult-tuned algorithms may misrepresent true energy usage in minors.

Setting all profiles to 18 years or older affects multiple Fitbit-derived metrics beyond calories. Cardiovascular fitness score (a proxy of $V\dot{O}_2$ max) is explicitly age-dependent and sex-dependent [79], so values in minors will be misestimated. Finally, activity classification and intensity inference are trained on adult patterns or rely on adult cutpoints [80]; studies show poor transfer to children and wide variability when applying adult or nonpediatric thresholds, motivating pediatric-specific models [81]. In practical terms, these biases limit the validity of continuous cardiovascular monitoring using standard Fitbit commercial pipelines in children: improvements or deteriorations in cardiovascular load may be masked or exaggerated.

Some issues appear easier to overcome than others; for instance, the lack of PRV data and the presence of inconsistencies in the sampling could be dealt with by using Fitbit smartwatches instead of activity trackers. It is, in fact, possible to design custom apps for Fitbit smartwatches, but not for trackers, by exploiting the software development kit available. Nevertheless, our experience suggests that, when used in a controlled protocol and combined with rigorous data cleaning, even simple trackers can yield reliable step counts and consistent PR data—even in a cohort of children.

A key limitation of this study is the small sample size (n=16), which substantially restricts the statistical power of correlation and regression analyses. With multiple predictors included in the models, the degrees of freedom are limited, increasing the likelihood of overfitting and unstable coefficient estimates. Accordingly, the regression outputs presented here primarily demonstrate the feasibility of integrating wearable-derived features into statistical models, rather than serving as robust or generalizable predictors. Owing to this, the absence of significant associations should be interpreted cautiously, as these results may reflect insufficient power rather than a true lack of relationship between wearable-derived and clinical variables. In addition, the negative adjusted R^2 values observed in several models, together with the presence of multicollinearity among personal and wearable data, indicate that the LASSO and LOOCV results should not be overinterpreted. With such a small sample and correlated predictors, penalized regression can select unstable subsets of variables, and cross-validation may provide overly optimistic estimates of predictive value. Future studies with larger, more heterogeneous samples will be required to confirm these findings, refine the regression models, and enhance generalizability.

While data acquisition from the Fitbit devices was successful in this cohort, this analysis did not assess broader aspects of feasibility, such as long-term adherence, cost, or regulatory considerations, nor are such aspects the focus of this tutorial.

General Considerations

The 12 steps of the tutorial can be adapted to other studies with Fitbit devices with minimal adjustments and with other devices by maintaining some of the steps (steps 1-5 and step 12) and adapting the remaining steps to the specific file format and available data. This pipeline could be easily replicated in other studies using similar devices. In a previous study conducted by our research group [67], we enrolled 31 older patients (mean age 76.1 years) to assess whether parameters that could be obtained from a Fitbit Inspire 2 device were correlated with 6MWT results. We obtained statistically significant correlation in the cases of nonexercise testing cardiorespiratory fitness (ρ=0.68; P<.001), RPR (ρ=−0.39; P=.03), HR over steps at 1% quantile (ρ=−0.39; P=.04), mean number of steps (ρ=0.59; P<.001), logarithmic transformation of the mean number of steps (ρ=0.56; P=.001), and $V\dot{O}_2$ max (ρ=0.58; P=.006).

On the basis of our findings across multiple studies, data from commercial wearable devices could offer a low-cost alternative to in-hospital assessments such as the 6MWT. This approach is particularly advantageous for patients who live far from the hospital or when hospital access is restricted, as was the case during the recent COVID-19 pandemic. Furthermore, the presented technology allows tracking one's progress over time. Results obtained in children demonstrate that even models designed for adults could be useful; however, further considerations should be applied when devices are used on children, particularly for company-provided metrics that are not immediately connected to the raw values of parameters.

Fitbit Charge 2 and Charge 3 are low-cost, basic trackers. More recent models (eg, Fitbit Charge 6) can be purchased with less than €150 (about US $173) per unit. An advantage of such devices is that they are widespread, specifically among young people and thus already present in a large part of the population, constituting a mine of data that is constantly supplied with even more data. Beyond childhood obesity, wrist-worn PR activity workflows have relevance across multiple clinical domains. In congenital heart disease, an American Heart Association science advisory supports wearable devices for continuous home monitoring and outlines clinical, ethical, and integration considerations [82]; pediatric studies also show consumer smartwatches (in this case, the Apple Watch) can capture clinically relevant arrhythmias not detected by standard ambulatory monitors [83]. In respiratory disease, cystic fibrosis cohorts have derived individualized PR thresholds from activity trackers (the so-called HRZ) to quantify moderate-to-vigorous activity [84], while asthma programs using multimodal wearable home monitoring have shown feasibility for assessing compliance outside the clinic [85]. These represent only a few examples of possible clinical applications of wearables in school-aged children.

The main limitations of currently available commercial devices are that they generally do not provide raw data, and their metrics (such as the number of minutes of an activity of a given intensity) are computed based on adult data and a black box, generally artificial intelligence-based [86] model from the point

XSL•FO
RenderX

of view of the end user. This makes it difficult, if not impossible, to understand how the algorithm was designed and on which assumptions it is based, particularly because private companies do not disclose their algorithms. If this lack of transparency could still be acceptable in the case of adult populations, where the assumption holds that substantial amounts of data have been collected in the past by major companies, this is no longer true in the case of pediatric applications.

As noted in the state-of-the-art section, most commercial devices for children are focused on activity metrics, but there are no devices that perform cardiovascular measurements. Both raw data and dedicated studies on continuous monitoring of cardiovascular parameters would enhance our ability to evaluate data quality and accuracy.

Conclusions

This tutorial primarily aims to establish the methodological feasibility of using adult-oriented wearable devices for pediatric monitoring, rather than to test clinical hypotheses or generate clinical conclusions. Our findings show that, when appropriate preprocessing and data quality assessments are applied, low-cost wrist-worn wearables can provide some usable data in school-aged children. The clinical example illustrates how such methodological steps translate into practice, but the interpretation remains exploratory and is not intended as a validation study. Although consumer devices may not offer raw signals or entirely pediatric-calibrated metrics, the overall experience we describe indicates that step counting and PR data can still be leveraged in clinical or research contexts involving children. By applying careful methods to manage and analyze the data, clinicians and investigators can minimize inaccuracies and still benefit from remote monitoring capabilities.

To perform reliable clinical trials using wearables on the pediatric population, activity trackers have demonstrated their performance, while further research is needed to assess the accuracy of cardiovascular parameters. The methodology proposed in this tutorial, however, gives indications on how to process such data with caution and maintaining scientific rigor, which is necessary for applications in clinical trials.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT (Generative Artificial Intelligence Delegation Taxonomy) (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing. The GAI tool used was GPT-5.

Responsibility for the final manuscript lies entirely with the authors.

GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was funded by: (1) the PODiaCar Project 101128946-PODiaCar-EU4H-2022-PJ-3 co-funded by European Union; (2) the National Recovery and resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3-Call for proposal No. 341 of 15 March 2022 of Italian Ministry of University and Research funded by the European Union-NextGenerationEU. Project code PE00000003, Concession Decree No. 1550 of 11 October 2022 adopted by the Italian Ministry of University and Research, CUP D93C22000890001, Project title “ON Foods-Research and innovation network on food and nutrition Sustainability, Safety and Security-Working ON Foods”; (3) the National Plan for National Recovery and Resilience Plan (NRRP) Complementary Investments (PNC, established with the decree-law May 6, 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of the Research Initiatives for Technologies and Innovative Trajectories in the health and care sectors (Directorial Decree n. 931 of 06–06-2022)–AdvaNced Technologies for Hu-man-centrEd Medicine (ANTHEM) under Project PNC0000003. This work reflects only the authors’ views and opinions, neither the Ministry for University and Research nor European Commission can be considered responsible for them. The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data import and supplementary tables.
[DOCX File, 58 KB - jmir_v28i1e76166_app1.docx]

References

1. Laranjo L, Ding D, Heleno B, et al. Do smartphone applications and activity trackers increase physical activity in adults? Systematic review, meta-analysis and metaregression. Br J Sports Med 2021 Apr;55(8):422-432. [doi: 10.1136/bjsports-2020-102892]

XSL•FO
RenderX

2. Angelucci A, Aliverti A. Telemonitoring systems for respiratory patients: technological aspects. Pulmonology 2020 Jul 30;26(4):221-232. [doi: 10.1016/j.pulmoe.2019.11.006]
3. Aliverti A, Evangelisti M, Angelucci A. Wearable tech for long-distance runners. In: Canata GL, Jones H, Krutsch W, Thoreux P, Vascellari A, editors. The Running Athlete: A Comprehensive Overview of Running in Different Sports: Berlin, Heidelberg: Springer Berlin Heidelberg; 2022. [doi: 10.1007/978-3-662-65064-6_10]
4. Bernasconi S, Angelucci A, Aliverti A. A scoping review on wearable devices for environmental monitoring and their application for health and wellness. Sensors (Basel) 2022;22(16):5994. [doi: 10.3390/s22165994]
5. Curone D, Secco EL, Tognetti A, et al. Smart garments for emergency operators: the ProeTEX project. IEEE Trans Inform Technol Biomed 2010;14(3):694-701. [doi: 10.1109/TITB.2010.2045003]
6. Solbiati S, Charkas S, Mozzini F, et al. Comparison of ECG-free algorithms for heart rate computation from head-BCG signals obtained with smart eyewear. Presented at: 2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE); Oct 21-23, 2024; St Albans, United Kingdom p. 213-218. [doi: 10.1109/MetroXRAINE62247.2024.10795866]
7. Floris C, Solbiati S, Landreani F, et al. Feasibility of heart rate and respiratory rate estimation by inertial sensors embedded in a virtual reality headset. Sensors (Basel) 2020;20(24):7168. [doi: 10.3390/s20247168]
8. Kinnunen H, Rantanen A, Kenttä T, Koskimäki H. Feasible assessment of recovery and cardiovascular health: accuracy of nocturnal HR and HRV assessed via ring PPG in comparison to medical grade ECG. Physiol Meas 2020 May 7;41(4):MID. [doi: 10.1088/1361-6579/ab840a] [Medline: 32217820]
9. Chu M, Nguyen T, Pandey V, et al. Respiration rate and volume measurements using wearable strain sensors. NPJ Digital Med 2019;2(1):1-9. [doi: 10.1038/s41746-019-0083-3]
10. Angelucci A, Villamar OWP, Agostoni P, Aliverti A. Design and evaluation of a wearable single-lead ECG for continuous monitoring. Presented at: 2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE); Oct 25-27, 2023. [doi: 10.1109/MetroXRAINE58569.2023.10405813]
11. Angelucci A, Cavicchioli M, Cintorrino IA, et al. Smart textiles and sensorized garments for physiological monitoring: a review of available solutions and techniques. Sensors (Basel) 2021;21(3):814. [doi: 10.3390/s21030814]
12. Angelucci A, Kuller D, Aliverti A. A home telemedicine system for continuous respiratory monitoring. IEEE J Biomed Health Inform ;25(4):1247-1256. [doi: 10.1109/JBHI.2020.3012621]
13. Angelucci A, Aliverti A. Detrended fluctuation analysis of day and night breathing parameters from a wearable respiratory holter. Comput Biol Med 2025 Apr;188:109907. [doi: 10.1016/j.compbiomed.2025.109907]
14. Röddiger T, Wolffram D, Laubenstein D, Budde M, Beigl M. Towards respiration rate monitoring using an in-ear headphone inertial measurement unit. Presented at: UbiComp '19: The 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing; Sep 9, 2019; London United Kingdom p. 48-53. [doi: 10.1145/3345615.3361130]
15. García-Pinillos F, Roche-Seruendo LE, Marcén-Cinca N, Marco-Contreras LA, Latorre-Román PA. Absolute reliability and concurrent validity of the Stryd system for the assessment of running stride kinematics at different velocities. Journal of Strength & Conditioning Research 2021;35(1):78-84. [doi: 10.1519/JSC.0000000000002595]
16. Evenson KR, Goto MM, Furberg RD. Systematic review of the validity and reliability of consumer-wearable activity trackers. Int J Behav Nutr Phys Act 2015 Dec;12(1):1-22. [doi: 10.1186/s12966-015-0314-1]
17. Jachymek M, Jachymek MT, Kiedrowicz RM, Kaźmierczak J, Płońska-Gościniak E, Peregud-Pogorzelska M. Wristbands in home-based rehabilitation-validation of heart rate measurement. Sensors (Basel) 2021 Dec 23;22(1):60. [doi: 10.3390/s22010060] [Medline: 35009603]
18. Angelucci A, Canali S, Aliverti A. Digital technologies for step counting: between promises of reliability and risks of reductionism. Front Digit Health 2023;5:1330189. [doi: 10.3389/fdgth.2023.1330189] [Medline: 38152629]
19. Lyons K, Hei Man AH, Booth D, Rena G. Defining activity thresholds triggering a "stand hour" for Apple Watch Users: cross-sectional study. JMIR Form Res 2024 Jun 10;8(1):e53806. [doi: 10.2196/53806]
20. Angelucci A, Bernasconi S, D'Andrea M, et al. Integration of a body sensor network of wearable devices for cardio-respiratory monitoring. Presented at: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Jul 24-27, 2023; Sydney, Australia p. 1-4. [doi: 10.1109/EMBC40787.2023.10340495]
21. Antonelli A, Guilizzoni D, Angelucci A, et al. Comparison between the Airgo™ Device and a Metabolic Cart during Rest and Exercise. Sensors (Basel) 2020 Jul 15;20(14):3943. [doi: 10.3390/s20143943] [Medline: 32679882]
22. Angelucci A, Camuncoli F, Galli M, Aliverti A. A wearable system for respiratory signal filtering based on activity: a preliminary validation. Presented at: 2022 IEEE International Workshop on Sport, Technology and Research (STAR); Jul 6-8, 2022; Trento - Cavalese, Italy p. 19-23. [doi: 10.1109/STAR53492.2022.9860001]
23. Angelucci A, Camuncoli F, Dotti F, et al. A wireless body sensor network for cardiorespiratory monitoring during cycling. Presented at: 2023 IEEE International Workshop on Sport, Technology and Research (STAR); Sep 14-16, 2023; Cavalese - Trento, Italy p. 1-4. [doi: 10.1109/STAR58331.2023.10302438]
24. Angelucci A, Aliverti A. An IMU-based wearable system for respiratory rate estimation in static and dynamic conditions. Cardiovasc Eng Technol 2023 Jun;14(3):351-363. [doi: 10.1007/s13239-023-00657-3] [Medline: 36849621]
25. Angelucci A, Camuncoli F, Dotti F, et al. Validation of a body sensor network for cardiorespiratory monitoring during dynamic activities. Biocybern Biomed Eng 2024 Oct;44(4):794-803. [doi: 10.1016/j.bbe.2024.09.002]

XSL•FO
RenderX

26. Clinical care - Sibel Health, Inc. SIbEL. URL: https://sibelhealth.com/clinical-care/ [accessed 2025-10-14]
27. Chung HU, Rwei AY, Hourlier-Fargette A, et al. Skin-interfaced biosensors for advanced wireless physiological monitoring in neonatal and pediatric intensive-care units. Nat Med 2020 Mar;26(3):418-429. [doi: 10.1038/s41591-020-0792-9]
28. Tandon A, de Ferranti SD. Wearable biosensors in pediatric cardiovascular disease. Circulation 2019 Jul 30;140(5):350-352. [doi: 10.1161/CIRCULATIONAHA.119.038483]
29. Peng DM, Edelson JB, Gureddygari HR, et al. Capturing physiological data in children with heart failure using wearable digital technology: lessons from pilot project. CJC Pediatric and Congenital Heart Disease 2025 Aug;4(4):226-230. [doi: 10.1016/j.cjcpc.2025.03.007]
30. Drummond CK, Tandon A. Advancing wearable technology for monitoring heart activity in paediatric populations. CJC Pediatric and Congenital Heart Disease 2023 Aug;2(4):196-197. [doi: 10.1016/j.cjcpc.2023.06.003]
31. Nelson BW, Allen NB. Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: intraindividual validation study. JMIR Mhealth Uhealth 2019 Mar 11;7(3):e10828. [doi: 10.2196/10828] [Medline: 30855232]
32. Etiwy M, Akhrass Z, Gillinov L, et al. Accuracy of wearable heart rate monitors in cardiac rehabilitation. Cardiovasc Diagn Ther 2019 Jun;9(3):262-271. [doi: 10.21037/cdt.2019.04.08]
33. Diaz KM, Krupka DJ, Chang MJ, et al. Fitbit®: an accurate and reliable device for wireless physical activity tracking. Int J Cardiol 2015 Apr;185:138-140. [doi: 10.1016/j.ijcard.2015.03.038]
34. Au WW, Recchia F, Fong DY, et al. Effect of wearable activity trackers on physical activity in children and adolescents: a systematic review and meta-analysis. Lancet Digit Health 2024 Sep;6(9):e625-e639. [doi: 10.1016/S2589-7500(24)00139-0] [Medline: 39112110]
35. Fitbit Ace 3: user manual version 10. Google Store for Google Made Devices & Accessories. URL: https://staticcs.fitbit.com/content/assets/help/manuals/manual_ace_3_en_US.pdf [accessed 2025-05-26]
36. VÍVOFIT® JR. 3: owner's manual. Garmin Official Site. URL: https://www8.garmin.com/manuals/webhelp/GUID-BE237499-CE69-4300-B563-57DAA26CF813/EN-US/vivofit_JR3_OM_EN-US.pdf [accessed 2025-05-26]
37. Polar active user manual. Polar Support | Polar Global. URL: https://support.polar.com/e_manuals/Active/Polar_Active_user_manual_English/manual.pdf [accessed 2025-03-26]
38. Kim Y, Lochbaum M. Comparison of polar active watch and waist- and wrist-worn ActiGraph accelerometers for measuring children's physical activity levels during unstructured afterschool programs. IJERPH 2018 ;15(10):2268. [doi: 10.3390/ijerph15102268]
39. Nash D, Shah MJ, Shehab O, et al. "But for the blind spot": accuracy and diagnostic performance of smart watch cardiac features in pediatric patients. Heart Rhythm 2024 May;21(5):581-589. [doi: 10.1016/j.hrthm.2024.01.021]
40. Hardon HJ, Van Kerkhof YN, Bartelds B, Kammeraad JAE, Van Deutekom AW. The validation and accuracy of wearable heart rate trackers in children with heart disease: prospective cohort study. JMIR Form Res 2025 Sep 30;9(1):e70835-e70835. [doi: 10.2196/70835]
41. Ernstsson J, Svensson B, Liuba P, Weismann CG. Validation of smartwatch electrocardiogram intervals in children compared to standard 12 lead electrocardiograms. Eur J Pediatr 2024 Sep;183(9):3915-3923. [doi: 10.1007/s00431-024-05648-7] [Medline: 38918230]
42. Obesity and overweight. World Health Organization. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2025-03-23]
43. Kumar S, Kelly AS. Review of childhood obesity: from epidemiology, etiology, and comorbidities to clinical assessment and treatment. Mayo Clin Proc 2017 Feb;92(2):251-265. [doi: 10.1016/j.mayocp.2016.09.017] [Medline: 28065514]
44. Ogden CL. Defining overweight in children using growth charts. Md Med 2004;5(3):19-21. [Medline: 15495854]
45. Juonala M, Magnussen CG, Berenson GS, et al. Childhood adiposity, adult adiposity, and cardiovascular risk factors. N Engl J Med 2011 Nov 17;365(20):1876-1885. [doi: 10.1056/NEJMoa1010112]
46. Özgen İ, Çakır E, Torun E, Güleş A, Hepokur MN, Cesur Y. Relationship between functional exercise capacity and lung functions in obese chidren. Jcrpe 2015 Sep 5;7(3):217-221. [doi: 10.4274/jcrpe.1990]
47. Enright PL. The six-minute walk test. Respir Care 2003 Aug;48(8):783-785. [Medline: 12890299]
48. Wijeysundera DN, Pearse RM, Shulman MA, et al. Assessment of functional capacity before major non-cardiac surgery: an international, prospective cohort study. The Lancet 2018 Jun;391(10140):2631-2640. [doi: 10.1016/S0140-6736(18)31131-0]
49. Contini M, Angelucci A, Aliverti A, et al. Comparison between PtCO2 and PaCO2 and derived parameters in heart failure patients during exercise: a preliminary study. Sensors (Basel) 2021 Oct 1;21(19):6666. [doi: 10.3390/s21196666]
50. Ward KM, White DA, McMahon M, Griffith G. Pediatric Exercise Testing: Springer; 2024:993-1040. [doi: 10.1007/978-3-031-07563-6_38]
51. Cruz-León C, Expósito-Carrillo P, Sánchez-Parente S, et al. Feasibility and safety of field-based physical fitness tests: a systematic review. Sports Med Open 2025 Jan 24;11(1):8. [doi: 10.1186/s40798-024-00799-1] [Medline: 39853552]
52. Calcaterra V, Bernardelli G, Malacarne M, et al. Effects of endurance exercise intensities on autonomic and metabolic controls in children with obesity: a feasibility study employing online exercise training. Nutrients 2023 Feb 20;15(4):1054. [doi: 10.3390/nu15041054] [Medline: 36839416]

XSL•FO
RenderX

53. Geiger R, Strasak A, Treml B, et al. Six-minute walk test in children and adolescents. J Pediatr 2007 Apr;150(4):395-399. [doi: 10.1016/j.jpeds.2006.12.052] [Medline: 17382117]
54. Child growth standards. World Health Organization. URL: https://www.who.int/tools/child-growth-standards [accessed 2024-09-02]
55. Ostchega Y, Porter KS, Hughes J, Dillon CF, Nwankwo T. Resting pulse rate reference data for children, adolescents, and adults: United States, 1999-2008. Natl Health Stat Report 2011 Aug 24(41):1-16. [Medline: 21905522]
56. Rosner B, Cook N, Portman R, Daniels S, Falkner B. Determination of blood pressure percentiles in normal-weight children: some methodological issues. Am J Epidemiol 2008 Mar 15;167(6):653-666. [doi: 10.1093/aje/kwm348] [Medline: 18230679]
57. Flynn JT, Kaelber DC, Baker-Smith CM, et al. Clinical practice guideline for screening and management of high blood pressure in children and adolescents. Pediatrics 2017 Sep;140(3):e20171904. [doi: 10.1542/peds.2017-1904] [Medline: 28827377]
58. Age-based pediatric blood pressure reference charts. Baylor College of Medicine. URL: https://www.bcm.edu/bodycomplab/BPappZjs/BPvAgeAPPz.html [accessed 2024-09-02]
59. Fitbit charge 2: user manual version 13. Google Store for Google Made Devices & Accessories. URL: https://staticcs.fitbit.com/content/assets/help/manuals/manual_charge_2_en_US.pdf [accessed 2024-10-28]
60. Fitbit charge 3: user manual version 35. Google Store for Google Made Devices & Accessories. URL: https://staticcs.fitbit.com/content/assets/help/manuals/manual_charge_3_en_US.pdf [accessed 2024-10-28]
61. Evenson KR, Goto MM, Furberg RD. Systematic review of the validity and reliability of consumer-wearable activity trackers. Int J Behav Nutr Phys Act 2015 Dec;12(1). [doi: 10.1186/s12966-015-0314-1]
62. Shcherbina A, Mattsson C, Waggott D, et al. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. JPM 2017 Jun 1;7(2):3. [doi: 10.3390/jpm7020003]
63. Nelson BW, Allen NB. Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: intraindividual validation study. JMIR Mhealth Uhealth 2019 Mar 11;7(3):e10828. [doi: 10.2196/10828]
64. Muggeridge DJ, Hickson K, Davies AV, et al. Measurement of heart rate using the Polar OH1 and Fitbit Charge 3 wearable devices in healthy adults during light, moderate, vigorous, and sprint-based exercise: validation study. JMIR Mhealth Uhealth 2021 Mar 25;9(3):e25313. [doi: 10.2196/25313] [Medline: 33764310]
65. Wong JS, Lu WA, Wu KT, Liu M, Chen GY, Kuo CD. A comparative study of pulse rate variability and heart rate variability in healthy subjects. J Clin Monit Comput 2012 Apr;26(2):107-114. [doi: 10.1007/s10877-012-9340-6]
66. Le-Masurier GC, Sidman CL, Corbin CB. Accumulating 10,000 steps: does this meet current physical activity guidelines? Res Q Exerc Sport 2003 Dec;74(4):389-394. [doi: 10.1080/02701367.2003.10609109]
67. Angelucci A, Greco M, Canali S, et al. Fitbit data to assess functional capacity in patients before elective surgery: pilot prospective observational study. J Med Internet Res 2023;25:e42815. [doi: 10.2196/42815]
68. Charlton PH, Marozas V, Mejía-Mejía E, Kyriacou PA, Mant J. Determinants of photoplethysmography signal quality at the wrist. PLOS Digit Health 2025 Jun 1;4(6):e0000585. [doi: 10.1371/journal.pdig.0000585]
69. Nanchen D. Resting heart rate: what is normal? Heart 2018 Jul;104(13):1048-1049. [doi: 10.1136/heartjnl-2017-312731]
70. Monmeneu JV, Chorro FJ, Bodí V, et al. Relationships between heart rate variability, functional capacity, and left ventricular function following myocardial infarction: an evaluation after one week and six months. Clin Cardiol 2001 Apr;24(4):313-320. [doi: 10.1002/clc.4960240411] [Medline: 11303700]
71. Lewis MJ, Kingsley M, Short AL, Simpson K. Rate of reduction of heart rate variability during exercise as an index of physical work capacity. Scandinavian Med Sci Sports 2007 Dec;17(6):696-702. [doi: 10.1111/j.1600-0838.2006.00616.x]
72. Ståhle A. Aerobic group training improves exercise capacity and heart rate variability in elderly patients with a recent coronary event. A randomized controlled study. Eur Heart J 1999 Nov;20(22):1638-1646. [doi: 10.1053/euhj.1999.1715]
73. Mishra T, Wang M, Metwally AA, et al. Pre-symptomatic detection of COVID-19 from smartwatch data. Nat Biomed Eng 2020 Dec;4(12):1208-1220. [doi: 10.1038/s41551-020-00640-6] [Medline: 33208926]
74. Dominick GM, Winfree KN, Pohlig RT, Papas MA. Physical activity assessment between consumer- and research-grade accelerometers: a comparative study in free-living conditions. JMIR Mhealth Uhealth 2016 Sep 1;4(3):e110. [doi: 10.2196/mhealth.6281]
75. Orstad SL, Gerchow L, Patel NR, et al. Defining valid activity monitor data: a multimethod analysis of weight-loss intervention participants' barriers to wear and first 100 days of physical activity. Informatics (MDPI) 2021 Jun 1;8(2):39. [doi: 10.3390/informatics8020039]
76. Claudel SE, Tamura K, Troendle J, et al. Comparing methods to identify wear-time intervals for physical activity with the Fitbit Charge 2. J Aging Phys Act 2020 Dec 16;29(3):529-535. [doi: 10.1123/japa.2020-0059]
77. Jobson JD. Multiple Linear Regression: Springer; 1991:219-398. [doi: 10.1007/978-1-4612-0955-3_4]
78. Zhou F, Yin X, Hu R, et al. Energy expenditure estimation in children, adolescents and adults by using a respiratory magnetometer plethysmography system and a deep learning model. Nutrients 2022 Oct 1;14(19):4190. [doi: 10.3390/nu14194190]
79. Klepin K, Wing D, Higgins M, Nichols J, Godino JG. Validity of Cardiorespiratory Fitness Measured with Fitbit Compared to V˙O2max. Medicine & Science in Sports & Exercise 2019;51(11):2251-2256. [doi: 10.1249/MSS.0000000000002041]

XSL•FO
RenderX

80. Hosseini A, Fazeli S, Vliet EV, et al. Children activity recognition: challenges and strategies. Presented at: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); Jul 18-21, 2018; Honolulu, HI p. 4331-4334. [doi: 10.1109/EMBC.2018.8513320]
81. Howe CA, Clevenger KA, Leslie RE, Ragan MA. Comparison of accelerometer-based cut-points for children's physical activity: counts vs. steps. Children (Basel) 2018 Aug 3;5(8):105. [doi: 10.3390/children5080105] [Medline: 30081457]
82. Tandon A, Avari Silva JN, Bhatt AB, et al. Advancing wearable biosensors for congenital heart disease: patient and clinician perspectives: a science advisory from the American Heart Association. Circulation 2024 May 7;149(19):e1134-e1142. [doi: 10.1161/CIR.0000000000001225] [Medline: 38545775]
83. Zahedivash A, Chubb H, Giacone H, et al. Utility of smart watches for identifying arrhythmias in children. Commun Med (Lond) 2023 Dec 13;3(1):167. [doi: 10.1038/s43856-023-00392-9] [Medline: 38092993]
84. Tanriver G, Stanojevic S, Filipow N, et al. Using heart rate data from wrist worn activity trackers to define thresholds for moderate to vigorous physical activity in children and young people with cystic fibrosis. J Cyst Fibros 2025 Mar;24(2):412-417. [doi: 10.1016/j.jcf.2024.10.014]
85. Kouis P, Michanikou A, Anagnostopoulou P, et al. Use of wearable sensors to assess compliance of asthmatic children in response to lockdown measures for the COVID-19 epidemic. Sci Rep 2021 Dec 1;11(1). [doi: 10.1038/s41598-021-85358-4]
86. Angelucci A, Li Z, Stoimenova N, Canali S. The paradox of the artificial intelligence system development process: the use case of corporate wellness programs using smart wearables. AI Soc 2022 Sep 26:1-11. [doi: 10.1007/s00146-022-01562-4] [Medline: 36185063]

Abbreviations

6MWT: 6-minute walk test
DAP: diastolic arterial pressure
ECG: electrocardiography
GAI: generative artificial intelligence
GAIDeT: Generative Artificial Intelligence Delegation Taxonomy
HR: heart rate
HRZ: heart rate zone
ICC: intraclass correlation coefficient
LASSO: least absolute shrinkage and selection operator
LOOCV: Leave-one-out cross-validation
PPG: photoplethysmography
PR: pulse rate
PROS: pulse rate over steps
RPR: resting pulse rate
SAP: systolic arterial pressure

Edited by A Mavragani; submitted 17.Apr.2025; peer-reviewed by A Martinko, J Rogers, M Mansoor; revised version received 31.Dec.2025; accepted 20.Jan.2026; published 20.Mar.2026.

Please cite as:
Angelucci A, Aliverti A, Vandoni M, Lucini D, Larizza C, Carnevale Pellino V, Zuccotti G, Calcaterra V
Using an Adult-Designed Wearable for Pediatric Monitoring: Practical Tutorial and Application in School-Aged Children With Obesity
J Med Internet Res 2026;28:e76166
URL: https://www.jmir.org/2026/1/e76166
doi:10.2196/76166

© Alessandra Angelucci, Andrea Aliverti, Matteo Vandoni, Daniela Lucini, Cristiana Larizza, Vittoria Carnevale Pellino, Gianvincenzo Zuccotti, Valeria Calcaterra. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

SMARTCLOTH Prototype for Dietary Management in Patients With Diabetes Mellitus: Tutorial on Human-Centered Design Methodology for Health Care Hardware Development

Jose M Palomares[1,2,3], PhD; Rafael Molina-Luque[1,4], RN, MSc, PhD; Fernando León-García[1,2,3], PhD; Irene Casares-Rodríguez[1], MSc; María García-Rodríguez[1,5], MSc, DN, PhD; María Pilar Villena Esponera[1,6], PhD; Guillermo Molina-Recio[1,4], RN, MSc, PhD

[1]Lifestyles, Innovation and Health Associated Group, Maimonides Institute for Biomedical Research (IMIBIC), Avd. Menéndez Pidal, N/N, Córdoba, Spain

[2]Advanced Informatics Research Group (GIIA), University of Cordoba (UCO), Córdoba, Spain

[3]Department of Electronic and Computer Engineering, University of Cordoba (UCO), Córdoba, Spain

[4]Department of Nursing, Pharmacology, and Physiotherapy, University of Cordoba (UCO), Córdoba, Spain

[5]Faculty of Health Sciences, Atlantic-Mediterranean Technological University (UTAMED), Málaga, Spain

[6]Faculty of Arts and Social Sciences, International University of La Rioja (UNIR), Logroño, Spain

Corresponding Author:
Rafael Molina-Luque, RN, MSc, PhD
Lifestyles, Innovation and Health Associated Group, Maimonides Institute for Biomedical Research (IMIBIC), Avd. Menéndez Pidal, N/N, Córdoba, Spain

Abstract

Background: Developing user-centered digital health hardware requires systematic design methods applicable across clinical contexts. As diabetes mellitus continues to rise globally and contributes to morbidity, mortality, and costs, effective nutritional management remains essential—yet adherence is often poor. Digital health interventions grounded in human-centered design may enhance adherence by better aligning solutions with patients' real needs.

Objective: This tutorial aims to provide replicable guidance on applying the design thinking approach to health care hardware development, illustrated through the design, development, and preliminary usability evaluation of SMARTCLOTH (GA-16: Lifestyles, Innovation, and Health), a smart tablecloth prototype intended to facilitate dietary management and support adherence to nutritional recommendations among individuals with diabetes.

Methods: We demonstrate a systematic design thinking approach adaptable to other hardware contexts, using the Double Diamond model. In mapping, we performed a structured preassessment to define project scope and feasible functionalities. To characterize end user needs, we conducted 6 in-depth interviews with health care professionals and applied persona, empathy map, and customer journey map tools. In exploring, 5 focus groups (patients and diabetes educators) identified barriers, facilitators, and desired functionalities for dietary self-management. In building, we created low- and high-fidelity wireframes and interactive web prototypes using Phaser 3 (HTML5/JS) to simulate a kitchen workspace for meal assembly. In testing, 7 patients with different diabetes profiles participated in 3 iterative usability sessions. Using think-aloud, video analysis, and structured tasks, we documented completion times, errors, and the level of required assistance, enabling refinements. Development progressed through 15 internal versions and 3 user-tested prototypes with real-time adjustments when feasible.

Results: Interviews and focus groups yielded three user profiles guiding design: (1) adolescents with type 1 diabetes navigating social and dietary challenges, (2) working-age adults with type 2 diabetes who were motivated but inconsistent, and (3) older adults with type 2 diabetes showing low adherence due to entrenched habits. Iterative usability testing indicated that the system was intuitive, with improvements in layout, labeling, and navigation. Quantitative metrics showed refinement, with simple tasks being completed in under 1 minute in later iterations, while complex meal simulations took longer. Error rates and required guidance decreased as prototypes evolved. Qualitative feedback highlighted clarity, motivational value, and educational potential, while older participants requested larger text and simplified controls. Despite usability gains, motivational barriers persisted among low-adherence older adults.

Conclusions: This tutorial demonstrates that systematic human-centered design can yield feasible and well-accepted digital health hardware. SMARTCLOTH emerged as a promising tool for dietary management in diabetes, though effectiveness and clinical outcomes were not evaluated. The methodology can be adapted by teams developing hardware for chronic disease management.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e75744) doi:10.2196/75744

KEYWORDS

type 1 diabetes mellitus; type 2 diabetes mellitus; dietary adherence; user-centered design: usability; digital health

Introduction

Developing effective digital health hardware—such as wearable devices, smart medical equipment, or assistive technologies—presents unique methodological challenges that distinguish it from software-based interventions. Hardware development requires balancing user needs with physical constraints, manufacturing feasibility, and cost considerations, while ensuring usability across diverse populations. Human-centered design (HCD) and design thinking (DT) methodologies offer structured approaches to navigate these challenges, yet practical, step-by-step guidance on applying these methods to health care hardware development remains limited in the literature. This tutorial addresses this gap by providing detailed methodological guidance, illustrated through a real-world development case.

Diabetes mellitus (DM) is a chronic disease in which glucose metabolism is altered due to increased resistance to insulin action or low insulin production [1]. The prevalence of this pathology has increased significantly over the last few decades and will continue to do so in the coming decades. Data show that in 2024, an estimated 589 million adults were living with DM, a figure that could reach 853 million by 2050. More than 90% of these cases are type 2 diabetes mellitus (T2DM), making this disease a global public health challenge [2,3].

In addition to the significance of prevalence figures, it is worth noting that DM places a high demand on health care personnel due to the multiple comorbidities it can cause, including nephropathies, vasculopathies, and neuropathies, among others [4], with disease management playing a crucial role in this context. Diabetes management is a dynamic and long-term process that begins at diagnosis, when initial education and treatment are introduced, and continues with daily self-management tasks such as diet, medication adherence, and glucose monitoring. Periodic follow-up is also essential to prevent complications and adapt treatment as the disease progresses. In this context, diabetes self-management education and support should be provided at diagnosis, during critical transitions, and on an ongoing basis throughout the care process [5]. Diabetes self-management education and support contributes to both daily management and the acquisition of self-care skills, which are crucial for long-term adherence. Among lifestyle behaviors, nutrition is particularly influential [6,7], yet adherence to healthy dietary patterns remains inadequate in many patients [8,9]. Our research group focuses on the stage of learning and consolidating dietary self-management skills and on supporting sustained adherence through innovative tools adapted to patients' contexts that improve their clinical situation.

Among dietary strategies, carbohydrate counting stands out as the cornerstone of nutritional management in type 1 diabetes mellitus (T1DM), as it enables the adjustment of insulin doses [10]. However, its complexity often hinders adherence. The rapid growth of digital technologies has created new opportunities in this field. Mobile health apps have been shown to support lifestyle changes and chronic disease management [11-13]. In diabetes, both generalist apps (eg, MyFitnessPal [Mike Lee] and CalorieMama [Azumio Inc]) and T1DM-specific apps (eg, CarbAndMove, SocialDiabetes [SocialDiabetes SL], and Carb Manager [Wombat Apps LLC]) are widely available. Nevertheless, they share essential limitations, including low usability and adherence [14,15], limited involvement of health care professionals in their development [16], technical problems, and poor integration with glucose-monitoring devices [17,18]. This fragmentation often forces patients to use multiple apps simultaneously, complicating management and increasing the risk of abandonment.

Recent reviews also highlight that evidence of their effectiveness remains inconsistent, and many tools fail to adequately meet patients' real needs [19,20]. Successful solutions must therefore actively involve end users in the design process and incorporate features such as educational content, monitoring tools, goal setting, gamification, reminders, and social interaction options [21-23]. Positioning these initiatives within health behavior change frameworks [23-25] provides a comprehensive foundation for understanding how capability, opportunity, and motivation interact to enable sustained dietary change. In this regard, technological tools should not only support day-to-day management but also strengthen patients' ability to learn, consolidate, and adhere to effective self-care strategies over time [26].

This complex clinical challenge—requiring sustained user engagement, multidisciplinary collaboration, and hardware-software integration—provides an ideal context for demonstrating the systematic application of HCD methodology. The following sections present a replicable approach that other teams can adapt to their own hardware development challenges in chronic disease management.

In this scenario, HCD emerges as an innovation method particularly suited to the development of solutions in the highly ambiguous field of health care. It focuses on creating high-value solutions through a people-centered approach that leverages diversity of thought, creativity, and ethnographic insights, which are refined through iterative testing [27,28]. The iterative nature of HCD reduces risks in terms of time and budget, mitigates cognitive biases among researchers, empowers individuals and teams, and enhances both creative confidence and solution quality [29-33]. Despite these advantages, HCD remains relatively new and is inconsistently applied in health care innovation [34]. This fact is particularly relevant for addressing health care disparities, including the digital divide, which can be mitigated through collaborative solutions developed in partnership with patients and stakeholders [35,36].

HCD also complements classical scientific approaches by improving the translation of evidence into practice and

XSL•FO
RenderX

identifying opportunities for intervention in target populations [33,37]. Although it is often used interchangeably with DT [27,30,34,38], important distinctions exist between the two. While both are iterative and user-centered, HCD applies more rigorous methods aimed at usability and user satisfaction in digital health products. In contrast, DT has a broader scope that emphasizes innovation across diverse contexts [37-39]. In short, HCD can be considered a comprehensive framework, while DT operates as a flexible strategy for problem-solving within that framework [40].

At the same time, designing any logical or physical system is inherently complex, shaped by both enabling and constraining factors. The classical system design cycle outlines 5 main stages, namely objectives, research, requirements, design, and evaluation, yet traditionally involves users only in the final evaluation phase [41]. In practice, the process is iterative, with later phases informing adjustments to earlier ones. However, excluding users from early stages often results in products that lack adequate user experience, making people feel uncomfortable when using them with tools that were not developed according to their real needs and requirements.

For this reason, some methodologies have evolved to involve end users and stakeholders throughout the design cycle [29-31,34,38,39]. HCD has proven particularly effective for developing systems and interfaces tailored to users' needs while aligning with institutional objectives [32,33,37,38]. Because users are deeply engaged throughout the process, the resulting interfaces provide a better fit and more satisfying usability experience [27,28]. User input can be collected through both direct methods (eg, questionnaires and interviews) and indirect approaches (eg, observation of prototype interactions and session recordings), complemented by quantitative measures, such as task completion time or error frequency [33].

This tutorial demonstrates the practical application of these principles throughout a complete hardware development cycle.

In this context, a novel initiative called SMARTCLOTH has been developed. SMARTCLOTH is a 3-year project funded by the Spanish 2021 Health Research Projects of the Strategic Action in Health 2017 - 2020 call. It aims to develop and preliminarily test the usability of a "smart tablecloth" designed to enhance dietary management and adherence in individuals with diabetes, supporting both the portion diet system (standard in type 1 diabetes) and the plate method (recommended in type 2 diabetes). This tutorial provides step-by-step guidance on applying HCD methodology and iterative prototyping to health care hardware development, illustrated through the SMARTCLOTH project. The approach emphasizes usability testing strategies that ensure accessibility for patients or caregivers of any age and level of technological literacy, principles that can be adapted to other digital health hardware development contexts.

Methods

Overview

This tutorial provides step-by-step guidance on applying HCD methodology and iterative prototyping to health care hardware development, illustrated through the SMARTCLOTH project. The approach emphasizes usability testing strategies that ensure accessibility for patients or caregivers of any age and level of technological literacy, principles that can be adapted to other digital health hardware development contexts. For each phase, we describe (1) the purpose and key objectives of the phase, (2) specific tools and techniques used, (3) practical implementation details, and (4) decision-making processes when challenges arose.

We present the phases sequentially for clarity but acknowledge that in practice, iteration between phases is common and often necessary. Timelines and resource allocations are provided as reference points but will vary based on project scope and available resources.

The research team applied the DT method, following the Double Diamond model proposed by the UK Design Council [42] and adapted according to Gasca and Zaragozá [43,44]. This Double Diamond structure (Multimedia Appendix 1) identifies 4 general stages (mapping, exploring, building, and testing). In turn, the first three are subdivided into another 6. In their proposal, these authors offer specific tools, which are not mutually exclusive, to complete each stage and substage. All the tools used in the development of SMARTCLOTH are detailed in Table 1.

XSL•FO
RenderX

Table . Stages, substages, and tools in the development of SMARTCLOTH.

Stage and substage	Description	Tools
Mapping	It is used to delimit the context of the work (what is known, what is unknown, and what needs to be known) and the scope and objectives of the project.	—[a]
Mapping the team	Map the internal context of the interdisciplinary research team.	• Strengths, Weaknesses, Opportunities, and Threats (SWOT) • In/Out
Mapping customers	This tool helps know the profiles of patients who would benefit from using SMARTCLOTH (external context).	• In-depth interviews • Person • Empathy map • Customer journey map
Exploring	It focuses on using qualitative techniques to understand and delimit the problem. It also gathers information on what is currently being done to address it.	—
Research	It is essential to understand the problem and its related factors in depth.	• Focus group (patients) • Focus group (experts)
Synthesis	The specification of the problem in the design challenge, in addition to synthesizing the research into a challenge, orients the team and focuses on devising solutions on something concrete.	• Design challenge
Building	It consists of embodying ideas in a prototype to define and transmit an idea quickly.	—
Devise	Through divergent thinking, the devise substage allows a broad spectrum of possible solutions to the design challenge to be explored before converging on the most effective and feasible proposals.	• Brainstorming • Cocreation sessions
Prototyping	It consists of materializing ideas. It is the process of quickly defining and transmitting an idea and materializing it into something physical that can be tested.	• Wireframes and web prototyping • 3D printing and physical prototyping
Testing	It is used to get feedback on each idea presented to end users. This stage allows us to learn what works and what doesn't and to test functionalities without the need for the final product.	• User test on the web • User test on a physical prototype • Usability test

[a]Not applicable.

Phase 1: Mapping (January-February 2022)

Scope Definition and User Profiling

The key aspects of this phase include:

1. Purpose: Systematic problem assessment, team alignment, and user characterization
2. Adaptable duration: 1 - 3 months depending on project scope
3. Key outputs: User profiles, project scope definition, and initial requirements
4. Application to SMARTCLOTH: According to the Double Diamond DT, the initial stage is the mapping, in which the problem is assessed, and the available resources are determined. This phase started in January 2022 and finished by the end of February 2022.

Team

In the team mapping phase, conducted in February 2022, we used the SWOT (Strengths, Weaknesses, Opportunities, and Threats) and IN/OUT tools to identify and categorize the team members' strengths, weaknesses, opportunities, and threats and clearly define the elements within or outside the project's scope. Both health care professionals and engineers participated in these sessions.

User and Market

Subsequently, for patient mapping, 6 in-depth interviews were conducted with health care professionals who regularly deal with patients with diabetes (4 nurses and 2 nutritionists), selected by nonrandom convenience sampling between February and March 2022. We used user-centered tools to synthesize the information collected, including "persona," "empathy map" and "customer journey map." The persona tool allows us to create

XSL•FO
RenderX

detailed profiles of fictitious users based on actual data. Each "persona" represents a specific profile of patients who might use SMARTLOTH, detailing their demographics, behaviors, needs, and goals [45]. The "Empathy Map" was used to capture and organize information about what users think, feel, say, and do about diabetes dietary management. This tool helped us understand users' emotions and perspectives and identify their main frustrations and desires [46]. Finally, the "Customer Journey Map" was used to map the complete journey of people with diabetes through the health care system, from the first contact to the end of their interaction. It highlighted the touchpoints, emotions, and potential barriers users face throughout their experience [47]. By identifying opportunities for improvement at each stage of the journey, we were able to define the most effective and successful functionalities of SMARTCLOTH, as well as the point in education of patients with diabetes where it could be used. Ultimately, these tools allowed us to develop detailed patient profiles, deeply understand their needs, emotions, and experiences, and comprehensively visualize their interaction with health services over time, thus facilitating the identification of critical points and opportunities for improvement that SMARTCLOTH could represent.

Data Analysis for Mapping

A phenomenological approach was used to analyze the data from the in-depth interviews with health care professionals specialized in diabetes education. The interviews were transcribed and meticulously reviewed. Subsequently, initial coding of the transcripts was undertaken, identifying emerging themes related to perceptions and experiences of dietary management and technology use. The codes were grouped into broad themes, such as experiences with technology, challenges in diabetes education, and recommendations for technology improvements, which are essential for identifying common patterns and developing detailed user profiles.

User-centered design tools, such as "persona," "empathy mapping," and "consumer journey mapping," were developed using the information extracted from the transcripts and themes. Different user profiles were identified based on familiar patterns, and their demographic, psychographic, and target characteristics were detailed. The phenomenological approach was the most appropriate for this analysis, as it allowed capturing the subjective and emotional experiences of health care professionals and their patients, providing a comprehensive and richly detailed view that facilitated the identification of pain points and opportunities for improvement in the use of diet management hardware in patients with diabetes.

Phase 2: Exploring (April-June 2022)

Research, Feasibility, and Challenge Formulation

The main aspects of this phase are as follows:

1. Purpose: Deep investigation of user needs and technical feasibility
2. Adaptable duration: 2 - 4 months depending on complexity
3. Key outputs: Design challenge formulation, technical framework selection, and detailed requirements
4. Application to SMARTCLOTH: The research stage was carried out after establishing the initial set of objectives. We noticed that at least 2 research areas, clinical and technical, had to be considered.

Clinical Research

During the substage of the research, conducted between April and June 2022, we completed 5 focus groups to gain an in-depth understanding of the needs, experiences, barriers, and facilitators of different patient profiles regarding adherence to the recommended dietary pattern in diabetes. These included 3 focus groups with patients from the 3 profiles previously identified in the mapping stage and 2 focus groups with nurses responsible for diabetes education and management in both primary and specialized care.

Technical Research

We studied different software and hardware design methodologies. Hardware design requires determining the components, integrating all of them, placing each element in the device, building the device, interconnecting everything, testing it, and checking the conformance with the objectives and the experience of use of the product by different users. This task is rather long and time-consuming. Therefore, most companies do not build the final product until all the elements have been tested and approved. The users do all the testing, and in this sense, as the HCD requires large interactions with the users, it is highly suitable for combining both.

For this reason, we decide to include a quick and easy mechanism for creating a simulated device to ask users about the product without building it physically. This tool would reduce the costs of the overall design and development process, and users would provide accurate usage information for the designer. Therefore, the physical product will finally be built with the most accepted design from the users.

We analyzed several options for prototyping, including (1) creating a simulator from scratch using Python (Python Software Foundation) or other similar programming language, (2) building a mobile or tablet app, or (3) making a dynamic web page. The first option is the most powerful, as the programmer can design any behavior, but it requires much effort to develop the simulator. Moreover, adapting the simulator for every platform where it will be executed would be required. The second option has been widely used because tablets and smartphones are ubiquitous. Therefore, the app is developed once and used on lots of devices. The last option was creating a dynamic response web page. This third is the most versatile option, as any device connected to the internet with a web browser is suitable. Simulating the behavior of the SMARTCLOTH device could be like playing a game. In fact, many games are developed for HTML5-compliant web browsers, executing the code on the device and not on the server. Regarding HTML5 development, some of the most relevant game-creation frameworks are Phaser 3, Matter.js, Kiwi.js, Quintus, and CreateJS. All of them are similar, with slightly different variations and compatibilities with various web browsers and platforms. Phaser 3 is probably the most versatile HTML5 game development framework, as

XSL•FO
RenderX

it has fast rendering and multitouch input capabilities on mobile and desktop browsers.

Synthesis

We use the design challenge tool in the synthesis substage to analyze and organize the information gathered. This tool is used to frame the core design problem and helps define the scope and objectives of the project, establishing a common starting point for the team. Typically, the challenge is formulated as an open-ended question that encourages creativity and the exploration of multiple possible solutions [48,49]. While using this tool, we focus on identifying and articulating the primary needs and desires of the users, as well as the constraints and opportunities of the project context. The formulation of the design challenge was agreed upon in a collaborative and interdisciplinary process between the health care professionals and the engineering team, trying to synthesize all the data and insights collected in the previous stages and substages.

Data Analysis for Mapping

Similarly, for the exploration phase, all focus group sessions with different profiles, which were audio- and video-recorded, were transcribed verbatim by a specialized company. Ricoeur's method of hermeneutic analysis[50] a 4-stage inductive approach, was used with the support of NVivo (Lumivero) software to analyze qualitative data. First, 2 researchers reviewed the focus group data and transcripts multiple times to gain a general understanding and compile a list of main ideas. Then, the data were coded, and nodes with predefined categories were identified, establishing 13 analytical subcategories. In the third stage, new nodes were generated based on the affinity of meanings, creating 4 codes and 18 subcodes to achieve greater abstraction and discriminate the characteristics of the same idea. In the fourth stage, a greater understanding of the initial categories was achieved through the "hermeneutic arc," considering both the initial experience and the rereading of the corpus of data. This analysis enabled us to identify which elements or characteristics act as barriers or facilitators from the perspectives of both patients and diabetes education nurses.

Phase 3: Building (Iterative Timeline)

From Low-Fidelity Wireframes to Interactive Web Simulation

The relevant aspects of this phase include:

1. Purpose: Translation of requirements into tangible prototypes
2. Adaptable approach: Can use various prototyping tools based on resources
3. Key outputs: Low-fidelity wireframes, high-fidelity prototypes, and interactive simulations
4. Application to SMARTCLOTH: The team built a functional system for users to test at this stage.

Devise

The devise (ideation) substage was based on the implementation of brainstorming techniques and cocreation sessions and focused on the definition and integration in a single device of the functionalities that SMARTCLOTH should present to respond to the design challenge. During the brainstorming, the interdisciplinary team proposed creative and varied solutions, regardless of their initial feasibility, to maximize creativity and divergent thinking. This process allowed for exploring multiple approaches before selecting the most feasible ideas for further development [51]. The cocreation sessions also involved the interdisciplinary group to harness the diversity of thinking and experience, enrich the design process, and ensure that solutions were comprehensive and well-founded, facilitating a holistic and user-centered approach [52]. Ultimately, both sessions were characterized by interdisciplinarity, seeking solutions from the perspective of health care workers and patients but accepting the technological and time constraints for the completion of the project.

Prototyping

Subsequently, the prototyping substage started with the creation of low- and high-fidelity wireframes and then moved on to web prototypes. The low-fidelity wireframes were initially used to outline the basic structure and layout of the elements in our interfaces without focusing on design details. These simple, quick sketches allowed us to agilely explore different configurations and functionalities.

Finally, we developed high-fidelity wireframes, incorporating finer details, including typography, colors, and graphics, providing a more realistic representation of the final product [53,54]. Once the high-fidelity wireframes were approved, we created interactive web prototypes, which allowed us to simulate the end-user experience and evaluate the functionality and fluidity of the interface in a realistic digital environment, screen and button locations, icons, and general usability ideas. This strategy ensured the design was visually appealing, intuitive, and functional for users [22]. Phaser 3 framework enables drawing 2D objects, adding event-handling hook functions, and managing different functionalities using JavaScript. Several assumptions and simplifications had to be made. For example, we need to simulate the part of preparing the meal by humans. In real life, this process would take place at the kitchen table, by cutting portions of food or using a plate or bowl to serve each item. However, these actions fall outside the scope of the SMARTCLOTH physical device. The design will only deal with the device and its internal functioning. However, to simulate the behavior of the SMARTCLOTH virtual device, the web was split into 2 areas. On the right part, we add an area named "YOUR KITCHEN," in which a dish and several different foods could be dragged to place the dish above the balance or add or remove pieces of food to the dish or from the dish, or even directly above the weight. As a simplification, 2 buttons in the "YOUR KITCHEN" area were added to add 25 g of a given previously activated food or to remove 25 g from the piece of food. The piece of food grows bigger or smaller according to the number of 25 g clicks the user has pressed. If the piece of food goes to 0 g, it disappears from "YOUR KITCHEN." We decided that the "YOUR KITCHEN" area should be placed on the left part of the screen. On the right side of the screen, the SMARTCLOTH device was simulated.

The SMARTCLOTH web prototype was designed and developed using the Phaser 3 JavaScript framework. The initial

XSL•FO
RenderX

prototype, shown in Figure 1, was created using simple forms (circles and rectangles). Members of the project team tested this prototype internally to check the conceptual functionality of the elements.

Figure 1. SMARTCLOTH initial web prototype – not shown to users, internal testing only.

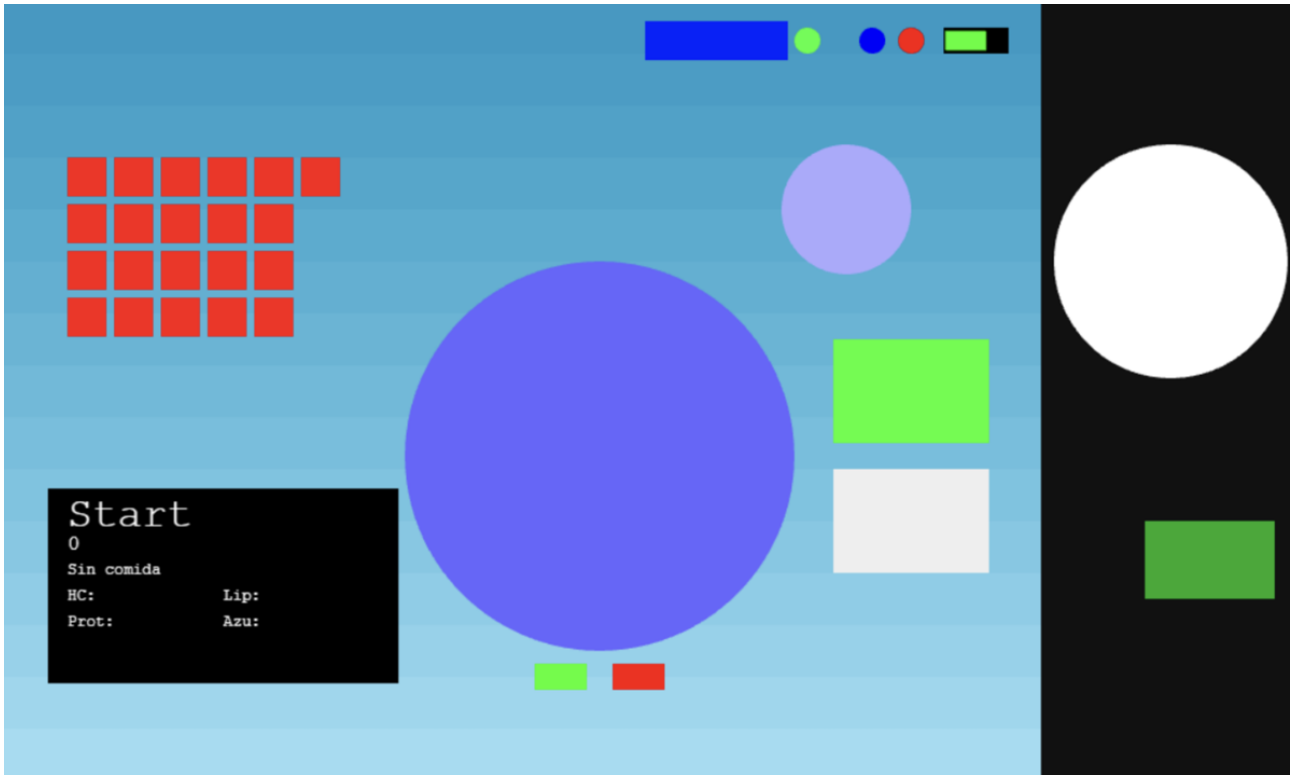

Phase 4: Testing (October-December 2022)

Iterative Usability Evaluation and Refinement

The main elements of this phase include:

1. Purpose: Iterative usability evaluation and refinement
2. Adaptable cycles: 2 - 5 testing iterations recommended
3. Key outputs: Validated design, usability metrics, user feedback, and refined prototype
4. Application to SMARTCLOTH: SMARTCLOTH web prototypes designed in the previous phase were used for the testing stage. Five testing sessions were conducted between October and December 2022 (Multimedia Appendix 2), involving 7 patients with diabetes representing the 3 previously identified profiles. Each prototype was tested by users and supervised by project team members. During these sessions, participants performed a series of specific tasks, detailed in Table 2, to evaluate the functionality and usability of the prototype. Each test was recorded with a video camera so that their behavior could be checked in the future. The users provided some annotated feedback, which was used to make all the changes for the following prototype evolution.

XSL•FO
RenderX

Table . Tasks applied in the SMARTCLOTH user tests.

Test	Elements involved	Difficulty
Weigh one portion of grilled salmon (300 g)	• Dish • Sliced salmon	Easy
Weigh a 150 g apple	• Apple	Easy
Weigh a plate of 150 g of pasta, 50 g of tomato, and 50 g of canned tuna.	• Dish • Pasta (boiled)[a] • Tomato, natural • Canned tuna (raw)	Moderate
Prepare a chickpea salad and a plate of 150 g of pasta, 50 g of tomato, and 50 g of canned tuna.	• Chickpeas (boiled) • Tomato, natural • Canned tuna (raw)	Moderate
Simulate two meals on the same day: • Breakfast: 150 g whole milk + 100 g whole wheat bread + 50 g butter • Lunch: 200 grams of boiled zucchini + 50 grams of oil + 150 g of breast + 1 boiled egg + 150 grams of banana	• Glass • Plate • Whole milk • Butter • Wholemeal bread • Dish[b] • Courgette (boiled) • Olive oil • Breast • Hard-boiled egg • Banana	Hard

[a]SMARTCLOTH can differentiate between raw and boiled foods.

[b]When having 2 meals, 2 dishes are needed.

The think-aloud technique was used to collect qualitative data on the user experience during the development of the tests. This technique involves participants verbalizing their thoughts, emotions, and actions as they interact with the prototype, providing insight into their cognitive process and identifying usability issues [55,56]. The choice of the think-aloud technique, widely recognized in usability research, was based on its ability to reveal critical aspects of user behavior that would not be evident through other evaluation methods [57]. The data collected was analyzed to identify common patterns, recurring problems, and SMARTCLOTH usability, providing the basis for subsequent design iterations and ensuring that the final product was intuitive and functional for end users.

We designed a set of questions for the users after each test. Each user was asked about how many kilocalories that meal had, how many rations of carbohydrates that meal had, and some other similar questions. We took note of the errors they committed and the overall interaction with the simulator. Finally, we asked them to express their feelings and suggestions about using that simulator (Multimedia Appendix 3).

For each recorded test, we wrote down several direct data points extracted from the videos, including number of attempts for each test, total time to complete the test, best time try, average time of the attempts, number of errors, number of explanations and positive reinforcements, and stops in the attempts (direct instructions and indirect indications). Some other indirect data were extracted from the videos of the answers after the test, including responses completed without help, with help, and unanswered or incorrectly answered. In the following, we describe every term extracted from the videos:

1. Attempt: Time interval in which the user works with the prototype to accomplish the proposed task.
2. Explanation: Time interval in which the user receives a more or less detailed explanation of the task, operation of the device, or some other information related to the experiment.
3. Error: Detection of malfunctioning of the prototype. Depending on its severity, it can generate a restart of the attempt.
4. Positive reinforcement: Affirmative expression confirming to the user that he or she is proceeding correctly.
5. Direct help: Explicit indication of the next step to follow.
6. Indirect help: Implicit indication, usually in the form of a question, to give the user clues on how to continue.

In each test, we showed the SMARTCLOTH web prototype on a large (50 inches) screen connected to a computer with an internet connection. We explained every button and element to users, describing their use and behavior. After that, one by one, each user was asked to manage different types of meals with increasing complexity. Each user entered the room where the test was carried out at a given time without interaction with the rest of the users so that no bias could arise when exchanging information among them.

Data Analysis for Testing

A qualitative approach based on content analysis was used to analyze the data obtained through the think-aloud methodology

XSL•FO
RenderX

and the annotations made during testing. The sessions were also videotaped, transcribed, and coded. User comments and behaviors were analyzed and grouped into categories related to usability, understanding of the hardware, difficulties encountered, and suggestions for improvement.

Positionality Statement

The research team comprised health care professionals (nurses and nutritionists) and engineers (computer and electronic engineering) affiliated with academic and biomedical research institutions in Spain. Our backgrounds combine expertise in diabetes education, nutrition therapy, digital health innovation, and HCD methodologies. Several members have extensive clinical experience supporting individuals with diabetes and are committed to enhancing patient education and dietary adherence. Others contribute technical expertise in prototyping, usability, and interactive system design.

We acknowledge that our values—particularly a belief in the potential of technology to enhance lifestyle management and our commitment to patient-centered care—have influenced the research process. These perspectives informed our emphasis on usability, patient engagement, and contextual adaptation of SMARTCLOTH. At the same time, we recognize that our orientation toward health care innovation could predispose us to highlight the positive aspects of digital interventions.

To address these influences, we used reflexive practices, triangulated perspectives within an interdisciplinary team, and integrated continuous feedback from patients and health care professionals throughout the design and testing process. This collaborative and iterative approach aimed to ensure that the findings reflect not only our assumptions but also the lived experiences and challenges of people with diabetes.

Ethical Considerations

The project complied with current Spanish regulations on bioethical research, personal data protection, and bioethics, following Law 3/Dec 5, 2018. In addition, the fundamental principles of the Declaration of Helsinki (1964), the Council of Europe Convention on Human Rights and Biomedicine (1997), and the United Nations Educational, Scientific and Cultural Organization (UNESCO) Universal Declaration on the Human Genome and Human Rights (1997) were respected. The research was approved by the Ethics and Research Committee of Cordoba (Act n° 273, ref. 3754). Since video and audio recordings were made of the participants, they were also asked to sign a consent for transferring image rights. In the case of minors, these consents were signed by their parents.

This study was conducted in accordance with ethical standards on privacy and confidentiality. No identifying information of participants is included in the manuscript or supplementary materials, and all potentially identifiable data were omitted unless scientifically essential. All participants signed informed consent and were able to withdraw it at any time if they considered it appropriate; in the case of minors, informed consent forms were signed by their parents. Since video and audio recordings were made of the participants, they were also asked to sign a consent form for transferring image rights. No compensation was provided to participants.

Results

Methodological Insights and Case Example Outcomes

This section presents the outcomes of applying the Double Diamond methodology to SMARTCLOTH development, organized to highlight both the specific results obtained in our case example and the generalizable methodological insights that emerged from the process. We emphasize decision points, adaptations, and lessons learned that can inform other hardware development projects.

Mapping

Team Mapping

The team considered that the main threats identified included the recent pandemic (and its long-term consequences), the difficulty in changing dietary habits, and the increasing number of similar products on the market. Significant weaknesses included limited time availability and lack of human resources, the disparity of interests among the researchers, and the fact that they were involved in several projects (multitasking). However, the project had important strengths, such as skills in web technologies, support from various institutions, and a strong motivation for health education. Notable opportunities included (1) direct access to patients and diabetes care professionals, (2) a positive view of technology in the wake of the COVID-19 pandemic, (3) the availability of funding, and (4) the opportunity to apply to different national calls for hiring professionals to reinforce the team, all of which supported the feasibility and potential success of the project.

In addition, it was crucial to determine which technological challenges could be addressed by the team. Among the functionalities of SMARTCLOTH that the team felt they were capable of developing were food weighing, barcode reading, design and physical printing of prototypes, design of buttons for user interactions, estimation of calories and macronutrients ingested at each meal with the ability to make longitudinal records of intake, and a website where each user could submit and consult this information, among others. However, it was decided that other functionalities (shopping list, voice recognition, and offering feedback on the diet quality) exceeded the time and economic capabilities of the equipment. A graphical summary of the results of these dynamics is available in Multimedia Appendix 4.

User Profiles Mapping

In the in-depth interviews conducted to identify user profiles, 6 professionals (4 diabetes nurse educators and 2 nutritionists) with varying experience in diabetes management but all over 3 years participated. The analysis of these interviews helped us identify potential users of SMARTCLOTH, identifying 3 profiles.

The first profile (given the name Kevin) is an adolescent with T1DM. Kevin manages his insulin well and understands his condition, but he faces challenges related to social life and eating out. He represents teenagers who think changing their diet is complicated and do not want to be the “weirdo” among their friends. They perceive their health as out of control and

XSL•FO
RenderX

sometimes feel frustrated because they cannot eat like everyone else. These users seek to enjoy and have fun while eating but face occasional family conflicts and struggle to adjust their eating habits. In their journey as users, they face (together with their family) initial doubts and fears, mistrust and loneliness, but show interest and confidence in seeking new tools to manage their condition.

The second profile, Julia, represents mature adults (between 50 and 65 years old) with T2DM. This profile corresponds to active workers, which makes it difficult for them to maintain a regular diet due to their schedules. They are motivated to improve their dietary knowledge and are concerned about their health but find it difficult to plan and prepare healthy meals. Overall, they maintain acceptable control of their disease. However, they perceive diet as a significant sacrifice and feel uncertainty and decreasing motivation when faced with limited consultation time and the confusion of scattered and, in many cases, contrary information they can find on other media, such as the internet. They often commit dietary transgressions linked to family gatherings, parties, and work shifts. In their journey as users, they experience an internal struggle to follow the diet and improve their eating habits while seeking to maintain an active social life and enjoy food with family and friends.

The third profile (Paco) corresponds to older people (aged more than 65 years) with T2DM and low adherence to dietary treatment. Generally speaking, they have a long history of the disease and face insecurities and a lack of interest in following the diet. In addition, they have no objection to adhering to pharmacological treatment, although they say that insulin injections frighten them a little. They prefer traditional food and do not want to give up what they enjoy. This group of users thinks it is difficult to change their eating habits at their age, and they feel distrustful and uncertain about the effectiveness of the diet. In the case of men, it is common for them to delegate responsibility for their diet to their wives (which can lead to family conflicts), and in general, they have not changed their lifestyle and do not follow the recommended diet. In their journey as users, this resistance to change stands out; they argue with their partners about meals and feel overwhelmed by the amount of information about diet despite understanding the importance of taking care of themselves. SMARTCLOTH could be a solution for this group, provided they are motivated. However, the fact that they are not motivated and have poor disease control makes it unlikely that they will ultimately be users of this device. These profiles helped design SMARTCLOTH to meet the specific needs of each user group, providing solutions tailored to their particular contexts and challenges.

Multimedia Appendix 5 (in Spanish) summarizes the tools used to define these profiles (Persona, Empathy Map, and Consumer Journey Map).

Exploring

Once these 3 profiles were identified, 5 focus groups were conducted, 1 for each profile, and 2 other groups that included primary and specialized care professionals responsible for diabetes education. The characteristics of the participants are detailed in Table 3 for the patients and Table 4 for the nurses.

XSL•FO
RenderX

Table . Descriptive variables of the patients participating in the focus groups.

Code	Age (years)	Sex	Marital status	Area	Employment status	Educational level	Socioeconomic status	Diagnosis DM[a] (year)
Patients with T2DM[b]								
T2DMP1	71	F[c]	MR[d]	U[e]	HW[f]	P[g]	ID[h]	2018
T2DMP2	65	F	MR	U	RT[i]	P	IE2[j]	2007
T2DMP3	70	F	MR	U	HW	P	IE1[k]	2004
T2DMP4	77	F	MR	U	RT	P	IE1	2006
T2DMP5	61	F	MR	U	AJS[l]	P	IE1	2009
T2DMP6	68	M[m]	MR	U	RT	P	IE1	—[n]
T2DMP7	67	M	MR	U	RT	P	ID	2012
T2DMP8	53	F	MR	U	A[o]	Sec[p]	IE1	2018
Patients with T1DM[q] and parents								
T1DMP1	16	F	SN[r]	R[s]	ST[t]	Sec	ID	2014
T1DMP2	17	M	SN	U	ST	Sec	ID	2014
T1DMP3	15	M	SN	U	ST	Sec	IA1[u]	2019
T1DMP4	15	M	SN	U	ST	Sec	IC[v]	2018
PART1DMP1[w]	55	M	MR	R	A	P	ID	—
PART1DMP2	50	F	MR	R	HW	P	ID	—
PART1DMP3	45	F	SD[x]	U	A	S	ID	—
PART1DMP4	48	F	MR	U	A	US[y]	IA1	—
PART1DMP5	46	F	MR	R	HW	P	IC	—
PART1DMP6	45	M	MR	R	A	Sec	IC	—

[a]DM: diabetes mellitus.
[b]T2DM: type 2 diabetes mellitus.
[c]F: female.
[d]MR: married.
[e]U: urban.
[f]HW: housewife.
[g]P: primary.
[h]ID: 1313 - 1602€/month (1€=US $1.16).
[i]RT: retired.
[j]IE2: less than 745€/month.
[k]IE1: 745 - 1312€/month.
[l]AJS: active job search.
[m]M: male.
[n]Not available.
[o]A: active
[p]S: secondary.
[q]T1DM: type 1 diabetes mellitus.
[r]SN: single.
[s]R: rural.
[t]ST: student.
[u] IA1: more than 3005€/month.
[v]IC: 1603 - 2145€/month.
[w]FAMXX: A family member or primary caregiver of a patient with type 1 diabetes mellitus.

XSL•FO
RenderX

[x]SD: separated or divorced.

[y]US: university studies.

[z]PAR: parents.

Table . Descriptive variables of the diabetes nurse educators participating in the focus groups.

Code	Age (years)	Sex	Educational level	Socioeconomic status	Years of experience in diabetes education
DNE1[a]	42	F[b]	U[c]	IB[d]	16 (as a primary nurse, although in health centers, they are not exclusively dedicated to this).
DNE 2	55	F	U	IB	17 (as a primary nurse, although in health centers, they are not exclusively dedicated to this). She was diagnosed with diabetes 30 years ago.
DNE 3	58	F	U	IB	32 (as a primary nurse, although in health centers, they are not exclusively dedicated to this).
DNE 4	55	F	U	IB	6 years as a diabetes nurse educator in the endocrinology department.
DNE 5	55	F	U	IB	6 years as a diabetes nurse educator in the endocrinology department.
DNE6	59	F	U	IB	4 years as a diabetes nurse educator in the endocrinology department.

[a]DNE: diabetes nurse educators.

[b]F: female.

[c]U: university studies.

[d]IB: 2146 - 2451€/month (1€=US $1.16).

Patients with T1DM and their caregivers (fathers, mothers, or both) indicated that diabetes management is more difficult in situations outside the daily routine, such as when eating out or at social events. They also mentioned that a lack of knowledge and confidence in portion measurement and carbohydrate counting can be a significant barrier. Other factors highlighted were the influence of the social environment and the need for flexibility in dietary choices. However, motivations for maintaining adherence included the desire for future well-being and effective diabetes management to avoid long-term complications. Possible functionalities we extracted for this group included food scanning and carbohydrate counting, personalized alerts and reminders, a healthy recipe database, real-time monitoring and feedback on glucose levels, adjusting dietary recommendations based on current readings, compatibility with glucose sensors, and educational and motivational functionalities.

Participants belonging to profile 2 (with T2DM and with acceptable metabolic control) highlighted that diabetes management is affected by factors, such as the social environment, the availability of healthy food at home, and the lack of structured support to follow a proper diet. Many mentioned that they find it difficult to follow dietary recommendations due to the temptation to consume unhealthy foods readily available in their environment. In addition, some patients expressed that the lack of clear and personalized information about recommended diets complicates their adherence. On the other hand, motivations for maintaining adherence include avoiding major complications and maintaining a good quality of life. Functionalities for this group included personalized dietary recommendations with personalized meal plans, food scanning and analysis, reminders and alerts (for food intake, medication, and glucose level monitoring), interactive nutrition education, connection to monitoring devices, community and motivational support

XSL•FO
RenderX

(creating a community support platform to share experiences and advice, and receive support), and physical activity recording.

Finally, from the group of patients belonging to profile 3, participants mentioned the difficulty of following a diet due to the constant temptation to consume unhealthy foods available at home, the lack of social and motivational support, and the perception that measuring and weighing food is tedious and difficult to maintain in the long term. Some patients expressed that physical exercise was used to justify consuming nonrecommended foods, believing that physical activity would compensate for dietary excesses. Lack of consistency and additional health problems were also important barriers. On the other hand, the main motivation for following the diet was to improve quality of life and avoid serious health complications. However, there was little motivation to adhere to the recommended dietary pattern. Continuous monitoring of glucose levels, recording of physical activity, alerts and reminders, nutritional and motivational education, food scanning, and carbohydrate counting were also highlighted as functionalities for managing their disease.

According to specialist care nurses for patients with diabetes, patients with T1DM are often newcomers, those who are newly diagnosed, or those with years of evolution who require ongoing education on the management of their condition, including insulin administration and carbohydrate counting. On the other hand, patients with T2DM, especially those poorly controlled (profile 3), tend to be more resistant to changing habits and face additional complications that make adherence to dietary guidelines complex and are the reason for referral from primary care. According to these nurses, the main barrier for patients with T1DM is the lack of adequate initial education, especially at the time of diagnosis, when both patients and their families feel overwhelmed. Adolescents with T1DM (profile 1) sometimes have low adherence to dietary recommendations due to rebelliousness and preference for unhealthy foods. In addition, constant carbohydrate counting and weighing is tedious for many, which can lead to errors in insulin management. However, structured and ongoing diabetes education is a crucial facilitator, as are family support and the use of technological tools, such as mobile apps and continuous glucose monitoring devices, which help to improve adherence. For patients with T2DM, the most prominent barriers include ingrained eating habits and the perception that following a diabetic diet is complicated and tedious. Cultural and social factors also play an important role in dietary adherence. Many patients prioritize medication over diet, which hinders adequate glycemic control. On the other hand, facilitators for these patients include ongoing education and awareness of the importance of diet in diabetes management. Social and family support can motivate patients to follow an appropriate diet, and using personalized meal plan tools is also beneficial.

As a result, the specialized care nurses believe that SMARTCLOTH could include functionalities such as the implementation of a food scanning and analysis system that provides detailed nutritional information that would help patients with carbohydrate counting, personalized meal plans, and detailed menus to avoid the monotony and perceived complication of following a diabetic diet, and automatic alerts and reminders for taking medications, meals, and glucose measurements. In addition, they comment that mobile apps that help to calculate portions and adjust insulin intake according to the meal would be a valuable tool, as well as interactive and ongoing nutrition education, including modules on nutrition and diabetes management, which can improve knowledge and adherence.

Primary care nurses mainly care for patients with T2DM and, to a lesser extent, patients with T1DM, with one of the participating nurses having T1DM, which brought a unique perspective to the session. According to the professionals, patients with T1DM (profile 1) often arrive well educated and controlled from diagnosis, especially if they debut in childhood, but face barriers such as the perception of diet as tedious and lack of adequate initial education. At the same time, facilitators include good diabetes education and family support and engagement. Patients with T2DM face barriers, such as ingrained dietary habits, cultural influences, and a passive attitude toward diabetes, with facilitators including continuing education, social support, and the use of practical tools. For all these reasons, they agree with the other nurses that functionalities, such as food scanning, personalized meal plans, automatic alerts and reminders, carbohydrate counting apps, interactive nutrition education, and a community support platform could improve dietary adherence.

Building

The qualitative assessment of patients' perspectives revealed a set of needs that were systematically translated into essential functionalities, which necessarily had to be incorporated into the SMARTCLOTH device. This process ensured that the design responded directly to real user requirements, providing core features, such as real-time conversion of food weights into macronutrient portions, integrated calorie estimation, food group identification, and barcode or QR code scanning (Table 5). However, other needs could not be satisfied due to technical constraints, challenges encountered during the prototyping phase, or because they exceeded the project's scope and objectives. These unmet needs, also documented in Table 5, included, for example, the provision of insulin dosing guidelines, auditory feedback when pressing keys, and additional nutritional recommendations, such as specific weight-loss or lactose-free diets. All this information was subsequently transferred to a SMARTCLOTH simulation website that supported the first round of user testing.

XSL•FO
RenderX

Table . Identified needs and functionalities developed to meet them[a].

Need identified	Functionality developed to respond to the need
• Difficulty converting food weight in grams into macronutrient portions. • Lengthy calculations, leading to disengagement and discontinuation. • Uncertainty regarding carbohydrate intake when estimating the insulin dose to be administered (T1DM)	The screen provides real-time display of both weight in grams and the corresponding macronutrient portions.
• Need to rely on multiple devices (scales, calculators, plates, depending on the food) to calculate each dish.	SMARTCLOTH integrates all required tools into a single device.
• Lack of information on caloric intake (particularly relevant for some patients with T2DM and obesity)	The screen provides real-time estimation of the caloric content of each dish.
• Lack of knowledge and difficulty identifying foods, the group they belong to, and their main macronutrient	SMARTCLOTH includes a panel of 20 food groups illustrated with images; selecting a group displays a list of foods within it.
• Difficulty estimating macronutrient portions in packaged foods.	SMARTCLOTH incorporates barcode and QR code scanning.
• Lack of overview of daily and weekly caloric and macronutrient intake.	SMARTCLOTH stores consumption data, accessible via a web platform.

[a]Needs that have not been met: (1) guidelines for insulin dosing based on carbohydrate intake; (2) auditory feedback when pressing device keys; (3) provision of healthy recipes or meal plans; (4) additional nutritional recommendations (eg, weight-loss and lactose-free diets); (5) touchscreen interface (requested primarily by younger patients).

The SMARTCLOTH web prototype had 15 internal versions, which provided 3 prototypes to be tested by the users. All the versions were tested internally to check whether the system behaved correctly. After those internal prototypes, users tested each new version and modified it accordingly. Main modifications came from the clinical part as new meals were designed, and thus, the “YOUR KITCHEN” had to be altered to add and remove food to create the meal.

Some other modifications were required because of the users’ feedback about the position or size of the elements (screen, balance, or buttons). These modifications were the most interesting from the design point of view, as they provided firsthand feedback from the users’ usage experience.

Besides, the researchers performed express modification sprints during testing sessions to change some behaviors or correct minor errors that would not affect the rest of the users in the testing. Some of these modifications were done in less than 10 minutes, from detecting the requirement to be modified or added to uploading the new code to the web server. These modifications could be done in real-time as the modified file was on the web server. Therefore, each new prototype version was available just after reloading the web page.

The following version, shown in Figure 2, represented a significant change. All the simple forms were substituted by images representing real buttons, screens, etc, so users could feel the usage experience as real as possible.

Figure 2. SMARTCLOTH’s first web prototype shown to users.



XSL•FO
RenderX

All the users in the test were surprised about the SMARTCLOTH prototype. They all liked it very much and were committed to helping in the test the best they could. We noticed that the users did not "read" the screen. The screen's position in the lower-left corner of the SMARTCLOTH prototype was not the best choice. All the users expected it in the upper left corner. They did not recognize most foods in the buttons representing food groups. The quality or size of those buttons did not allow the users to understand the foods included in every food group well. Some of them suggested adding examples on the screen showing the foods included in each group. Some of them did not like the position of the buttons for adding a new dish or removing a dish from the meal. They preferred all the control buttons related to the meals to be in the same area. The older users asked us to add text to the control buttons. Therefore, we added "TARA" (in Spanish) to the tare button.

The second version of the SMARTCLOTH web prototype, shown in Figure 3, changed the composition. As suggested in the previous test, the screen and the group of food buttons were swapped. Besides, we added new buttons related to cooked or raw foods. The clinical team requested this for the project. We decided to move the add dish and remove dish buttons below the weight, beside the Tare and Save buttons, as indicated in the previous test. Those two new buttons (cooked and raw) were placed on top of the weight. We added text inside the buttons describing the functionality of each of them, as mentioned in the previous test.

Figure 3. The second evolution of the SMARTCLOTH web prototype shown to users.



In the second evolution test, we noticed that the Tare button was useless in how users interacted with the prototype. The system did the tare when a new group of food buttons was pressed. Therefore, we removed the Tare button and the necessity of pressing 2 buttons.

Besides, users asked us to remove them from the final prototype, as neither barcode nor Bluetooth functionalities were included. In the same way, as the power consumption was not simulated in the web prototype, the battery icon was removed in the SMARTCLOTH final prototype. All these changes led us to the final SMARTCLOTH web prototype, shown in Figure 4.

XSL•FO
RenderX

Figure 4. Final testing of SMARTCLOTH HTML5-compliant web prototype.

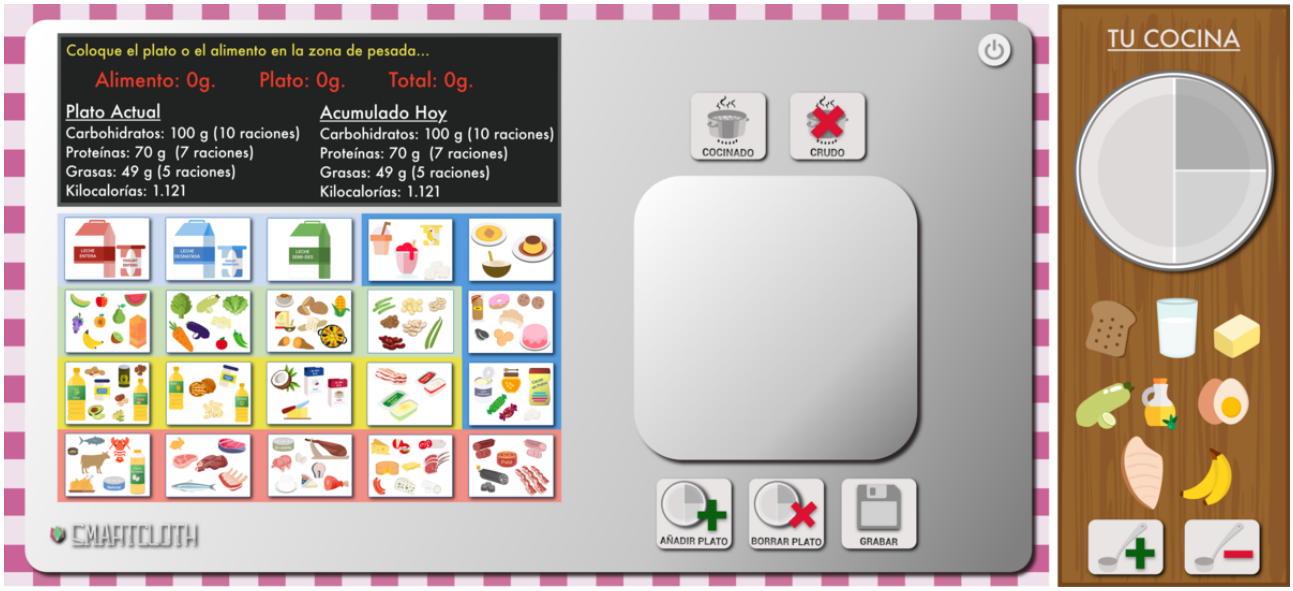

Testing

Three sessions were carried out. Every session was recorded and analyzed afterwards.

Tests 1 and 2

The SMARTCLOTH first web prototype, shown in Figure 2, was used in tests 1 and 2. The first test was a meal of 150 g of apple, and the second was 300 g of salmon. The analysis of the recorded video provided the results shown in Multimedia Appendix 6. Each user was anonymized by the acronym profile (P1/P2/P3) and user (1-5). For instance, P1.3 represents user 3 of profile 1 (T1DM).

Tests 3 and 4

The second version of the SMARTCLOTH web prototype, shown in Figure 3, was used in tests 3 and 4. Test 3 consisted of 150 g of apple, 250 g of yogurt, and 100 g of nuts. Test 4 consisted of 150 g of pasta, 50 g of tomato sauce, and 50 g of tuna. The analysis of the recorded video provided the results shown in Multimedia Appendix 7.

Test 5

The final version of the SMARTCLOTH web prototype, shown in Figure 4, was used in test 5, a complete meal composed of 3 dishes. The analysis of the recorded video provided the results shown in Multimedia Appendix 8.

Figure 5 shows the average time of the last try and the average time of the tries for each test, according to user profiles.

XSL•FO
RenderX

Figure 5. Average completion times for user tests. Bars represent the average time of the tries for each test and user type, while the lines represent the average time of the last try for each test and user type. T1DM: type 1 diabetes mellitus; T2DM: type 2 diabetes mellitus.

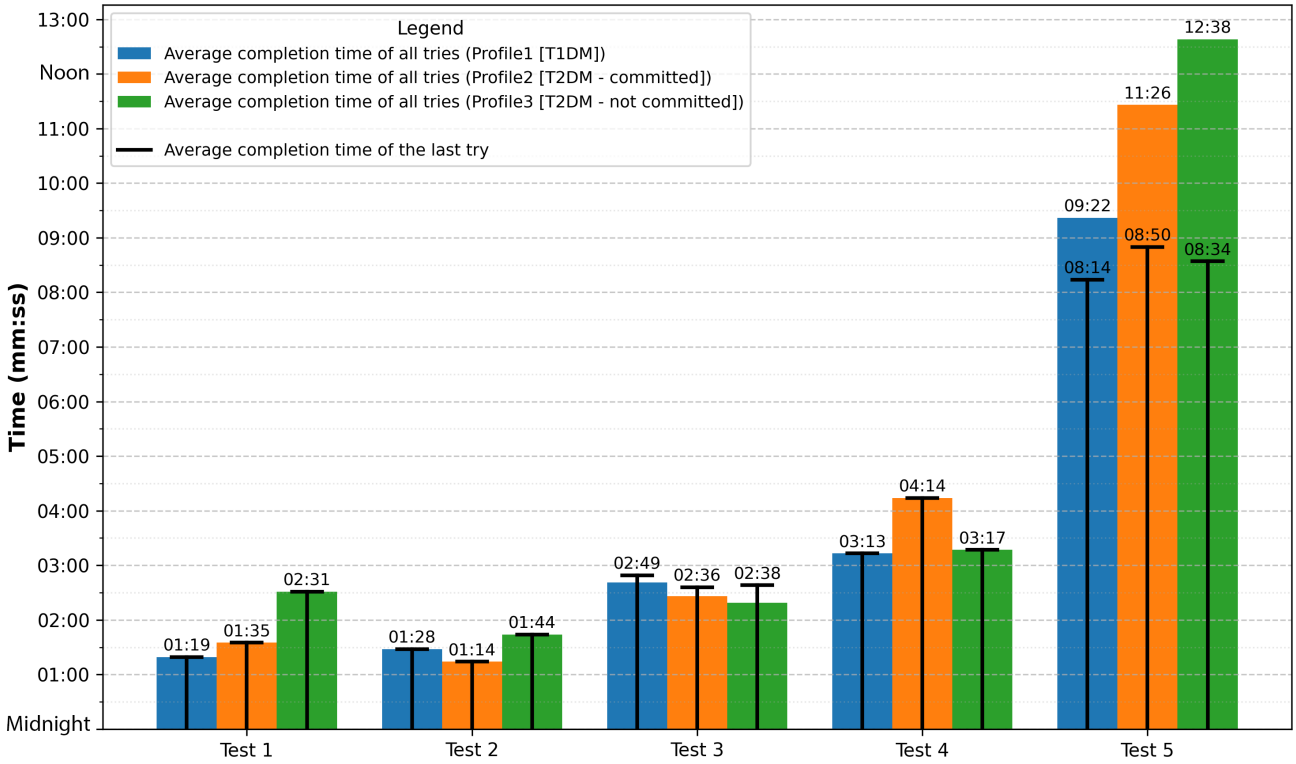

Discussion

Principal Findings

This tutorial has demonstrated the systematic application of HCD and DT methodology to develop digital health hardware, using SMARTCLOTH as an illustrative case example.

This DT methodology is widely used to create technological solutions for chronic health problems. For example, the study by Petersen and Hempler et al [58] used it to develop a mobile app that helps patients with diabetes manage their disease more effectively. The researchers found that the app significantly improved glycemic control and adherence to treatment. Another study used this methodology to design a telemedicine platform for patients with hypertension, resulting in a marked improvement in blood pressure monitoring and patient satisfaction [59]. Like our work, both studies showed optimal results, highlighting the effectiveness of DT in developing technological solutions for managing chronic diseases, where treatment adherence has been shown to be a key element.

Reflections on the HCD Process

Mapping

Team Mapping

Concerning the research team's preassessment, despite the threats and weaknesses identified, this technique offered an interesting approach for researchers to maximize their strengths and take advantage of the opportunities available. The recent pandemic presented significant challenges, coupled with the difficulty in changing dietary habits and competition with similar products on the market. In this sense, an honest and strategic assessment of the project's human, temporal, and economic capacities beforehand allowed the development of all stages of the DT methodology to be adjusted. Despite this, a dearth of studies uses or reports the results of these preassessment strategies. However, the orientation and focus that originated in the team underlines the need for robust methods to identify the potential of interdisciplinary groups and concretize the work in technological projects to improve adherence to lifestyles in chronic diseases. Recent studies have highlighted similar challenges and proposed similar solutions to overcome identified weaknesses in health-related technology projects [60-62].

User Profiles Mapping

Using end-user-centered design tools, such as "persona," "empathy mapping," and "customer journey mapping" has been particularly interesting. Traditionally used in business and the private sector to improve the benefits of business propositions, these tools are applied novelly in the health sector. However, this approach has been evidenced to help increase usability and adherence to use by patients with chronic conditions. For example, a recent study used the persona tool to design a medication-tracking app for patients with cardiovascular disease, significantly improving treatment adherence [63]. Another study implemented the "customer journey map" in developing a telerehabilitation program for patients with chronic obstructive pulmonary disease, finding that the tool helped identify friction points and improve the user experience, thereby increasing patient engagement and satisfaction [64]. Similarly, using the "empathy map" in developing a support platform for patients with diabetes allowed developers to better understand users' emotional and functional needs, leading to a more intuitive and effective design [65]. Like the studies discussed above, our work has also found these tools very useful in understanding and segmenting our users, having identified three profiles (1)

XSL•FO
RenderX

First profile comprised one of adolescents and their caregivers with T1DM, who manage their insulin well but face social and dietary challenges; (2) a second profile, of mature adults with T2DM who, while having difficulty maintaining a diet, try to engage with treatment; and (3) a third profile, corresponding to older people with T2DM and with low adherence to treatment and severe difficulties in modifying lifestyles.

Comparing these profiles with those reported in the literature, we found important similarities and differences. For example, a study on medication adherence in patients with diabetes identified that levels of family support play a crucial role in adherence to treatment. Patients who perceive greater instrumental support from family members, such as help with diet and medical appointments, tend to show better adherence to treatment regimens [66,67]. This finding aligns with our second profile, where family support is a key facilitator. Another study highlighted that barriers, such as lack of knowledge about diabetes and unsupportive family behaviors, such as negative interference with medication adherence, are significantly associated with poorer glycemic control outcomes [68]. This finding is particularly relevant for our third profile of older people with T2DM, who often have difficulty following dietary recommendations due to ingrained dietary habits and lack of structured support [69].

In addition, implementing continuous quality improvement programs in diabetes management has been shown to improve treatment adherence, promote self-care, and increase patients' knowledge about their disease [66,70]. This strategy could benefit all the profiles identified in our study, providing a constant support structure and ongoing education.

Ultimately, these tools have allowed us to identify each group's specific needs and barriers, facilitating the design of the functionalities that SMARTCLOTH should present as the most effective and personalized solutions to improve adherence and dietary management of diabetes. These were the key issues addressed in the next stage.

Exploring

A determining factor in the design of a technological solution is its adaptation to the needs of the end users. In this sense, during the exploring stage, factors that facilitate or hinder adherence to a diet that helps glycemic control in patients with diabetes were detected. At this point, we emphasized those identified barriers or facilitators relevant to the design of the proposed technological solution.

Among the 3 profiles identified, the environment (social events, eating outside the home, family pressure, constant exposure to unhealthy foods...) appears as a common element that gets in the way of following an adequate diet, something described in studies carried out in other countries [71,72]. This situation makes glycemic control difficult, either by eating high-calorie meals, high consumption of proteins and/or fats that alter glucose absorption, or difficulty counting carbohydrates ingested [73,74].

Of the elements mentioned in the previous paragraph, the high availability and exposure to high-calorie foods, unhealthy fats, and free sugars, especially ultraprocessed foods, are particularly important today. Studies show that consumption of these foods is very high [75], so it must be assumed that SMARTCLOTH users will incorporate them into their diet. In this context, a barcode scanner was envisaged to monitor the intake of these products accurately and to increase control and awareness of what is involved in eating them [76].

On the other hand, the difficulty in counting carbohydrates is not only limited to social life but also at specific times (the first months after diagnosis of the disease, adolescence, after years with the disease...), there is insecurity and/or relaxation due to lack of training, in the first case, or fatigue due to the tediousness of the process required to do so, in the second [77,78]. The counting of this macronutrient, especially in patients with T1DM, is fundamental in glycemic control, as it defines the insulin dose [79]. Therefore, the mainstay of SMARTCLOTH is the calculation of grams and servings of the 3 immediate principles and the monitoring of intakes throughout the day. Farooqi et al [80] found that this type of functionality is highly relevant in diabetes management and in increasing adherence to the necessary therapeutic approaches, as also found by other authors [81-83].

Furthermore, the ability to structure an adequate diet is an important limitation for most of the population. During the focus groups, patients with T2DM openly expressed this problem [71,84]. This difficulty is a major problem, as the correct dietary therapeutic approach (quality food, reduction of free sugars, moderate carbohydrate diet, and reduction of caloric intake) is essential to improve metabolic control [85]. In this context, SMARTCLOTH presents itself as an ideal tool for dietary management. In addition to the aforementioned functions, the digital tablecloth has to allow the monitoring of caloric intake at each meal and accumulated caloric intake over a day, which will allow moderation of energy intake and loss of body fat, which has emerged as the main goal set by health care workers for patients with T2DM [86]. However, SMARTCLOTH should be accompanied by dietary recommendations made by qualified personnel tailored to the individual patient's needs.

A lack of culinary knowledge also limits the ability to structure and follow an appropriate diet [87,88]. This shortage means that most meals are not organoleptically comparable to those offered by the food industry and require more preparation time [89,90]. Therefore, taking advantage of the barcode reader, the researchers decided to design a book and a database of standardized healthy recipes that would (1) improve the quality of meals, thanks to a better choice of foods; (2) reduce the time spent in the kitchen, through simple recipes; (3) increase the palatability of meals, through the introduction of very tasty recipes; and (4) reduce the difficulty of controlling meals, because the simple reading of the barcode associated with the recipe will allow the portion to be weighed directly without the need to consider the ingredients separately.

Monitoring blood glucose and its coupling with dietary intake is of particular concern to health care staff caring for this group of patients [91-93]. In this case, SMARTCLOTH will upload all intakes recorded on the device to a database. This information will be uploaded to a website, allowing the patient to enter their blood glucose and graphically observe the correlation between

XSL•FO
RenderX

blood glucose and intake. This functionality can help to establish a pattern or detect errors in the patient's self-care and help to intervene more precisely.

Finally, it should be noted that although SMARTCLOTH is designed and developed according to the needs expressed by patients, motivation will play a key role in its use, as, like any intervention, it will require commitment from users [94-96]. As evidenced throughout this research, it is unlikely that those who do not want to change their habits (profile 3) are likely to do so.

Undoubtedly, SMARTCLOTH will not address all the barriers that were identified, especially some related to the environment. This is because the final prototype will not be of a size that allows it to be used outside the home. However, a solution may be offered by future iterations of SMARTCLOTH, which may be more portable so that it can be used outside the home.

Building

Initially, the prototype was planned to be developed physically. However, adopting the web-based implementation strategy has greatly benefited the project. Web technologies have made it possible to speed up prototyping and make changes at no economic cost.

Three significant evolutions of the prototype (with multiple small internal modifications) have occurred. Agile methodologies guided these evolutions, and mini prototypes were rapidly developed. The development team tested these internally, and if they passed the quality standards, they were integrated into the prototype to be tested by the users.

Two factors limited the set of changes applied to the prototypes. The main limiting factor was the time until the next user testing session. The next factor that recommended limiting the set of changes in each prototype was the need to evaluate the impact on the users while maintaining the functionality, as including many changes could make detailed evaluation too tricky.

Due to this prototyping process's exploratory and discovery nature, it cannot be separated from the testing stage. The experience gained from this project has revealed that the methodology for developing the best prototype cannot be separated into watertight stages. Being user-centered, the process of specification, development, testing, and respecification must be continuous and cannot be disaggregated. Users' own evaluations generate proposals for modifications, which, in turn, require agility in development, with a very rapid process of specifying new requirements. In many cases, this entire process was reduced to 10 minutes. This extreme speed did not allow for a standard process of requirements gathering, analysis, design, and reimplementation. This section followed a software development paradigm [41] that did not precisely conform to the spiral or rapid prototyping paradigm, although it took many elements from both approaches.

Because of this rapidity, many of the usual thorough error-checking processes cannot be applied. This means that the developed prototypes should not be used for unsupervised external deployments but only for controlled testing. In other words, prototypes should be discarded and only be used to extract fundamental information for constructing the final system.

In short, all these evolutions have made it possible to substantially change the functioning and the layout of the different visual and information input elements.

Testing

At this point, it is necessary to differentiate between testing and user evaluation. Testing aims to verify the code. Therefore, tests ensure that the code does not contain errors and that the execution provides consistent results. User evaluation seeks validation of the product, that is, the degree of user acceptance of the system. This user evaluation does not seek the correctness of the results, as this is a later phase. Once the product has been developed per the requirements gathered from the users, it can be used by the research team to achieve the project's objectives.

Focusing on the internal tests, we must distinguish between those carried out before and during the user evaluations. The developments carried out before the evaluations had a short time frame, less than a month, but with the capacity to carry out a sufficient set of tests on the code. These tests detected lexical and syntactic coding faults and even validated results with output verification and acceptable ranges. The testing procedure followed the usual software engineering standards [97]. The code developed before the user evaluations had minor coding errors in specific performances that were difficult to trace. It can be considered good quality code with respect to the specifications previously established for each prototype evolution. However, due to time constraints, code testing was not performed for code modifications made during the evaluations. In this respect, it is essential that the developer involved have extensive experience in programming, in the JavaScript programming language, in Phaser 3, and in the development of web-based systems. The quality of this code varied considerably, although, in general, the code developed could pass the minimum quality requirements.

With regard to user evaluations, focusing on the product development aspect and the acquisition and adjustment of user requirements associated with the system, the procedure had to be more systematic. Follow-up documents were produced and filled in with the subjective evaluation of the user tests. In addition, video recordings were made to allow for a more analytical review of the user tests. In the documents and the video, we did not ask the same questions in all cases, so the users' answers may have been somewhat biased in some cases. However, we think this bias did not affect the sense of the suggested modifications, since the indications were mostly oriented to help users with the assignment of the different foods to the food group buttons. Only when users got stuck in the procedure for too long was full guidance provided until the process was completed. The evaluation of the reasons for the impossibility of completing the process was analyzed. In some cases, the size of the buttons and the lack of knowledge about the assignment of foods to food groups were problematic.

In general, user tests allowed the prototypes to evolve into the final design, which saved the project a lot of time and money.

XSL•FO
RenderX

Although SMARTCLOTH is conceived as a hardware device rather than a mobile app, its integrated intelligent functionalities—such as real-time conversion of food weights into macronutrient portions, barcode or QR scanning, and automated calorie estimation—are expected to reinforce both glycemic control and dietary adherence. From a clinical perspective, improved precision in dietary monitoring is anticipated to contribute to reductions in HbA_{1c}, as demonstrated in studies evaluating technology-based wearable interventions compared with standard care [98]. Beyond glycemic outcomes, SMARTCLOTH could also increase adherence by simplifying complex dietary tasks, reducing the cognitive burden of carbohydrate counting, and making the process more understandable. Usability and clarity of device interfaces have been shown to influence treatment adherence in diabetes care, with user-friendly designs being associated with better long-term engagement [99]. This characteristic is particularly relevant during the early stages after diagnosis, when patients often report errors and insecurity in carbohydrate counting, and an assistive hardware device may function as both an educational and behavioral reinforcement tool. Nevertheless, the translation of these functionalities into measurable clinical outcomes will depend on sustained device engagement, data completeness, and seamless integration into daily routines.

Methodological Reflections: Limitations and Strengths of the HCD Approach

These reflections on the limitations and strengths of our HCD application provide important considerations for other teams adapting this methodology to their own hardware development projects.

This project presents several strengths that enhance the validity and applicability of its findings. First, the use of an HCD methodology ensured that the prototype was developed based on the real needs and experiences of patients and health care professionals, thereby increasing usability and acceptance. Second, the iterative prototyping and testing process enabled continuous refinements, contributing to a functional and user-friendly system. Third, the inclusion of diverse patient profiles (adolescents with T1DM, working-age adults with T2DM, and older adults with low adherence) provided a comprehensive understanding of different user needs and improved the generalizability of the results.

Nevertheless, several limitations should be acknowledged. The user testing involved a relatively small sample, which may restrict representativeness. Evaluations were conducted in controlled settings rather than during everyday use, limiting inferences about long-term adoption. Finally, the current prototype is oriented toward home use, which may limit its applicability in outdoor contexts where dietary management can be especially challenging.

Although the HCD approach enabled us to gather a wide range of ideas and proposals from patients and professionals, not all of them proved feasible—whether due to technological or economic constraints of the project—or guaranteed a significant clinical impact. This fact is an inherent limitation of participatory processes in complex areas, such as diabetes management, particularly when patients express needs that are difficult to address without a highly active role on their part (eg, motivation). Nevertheless, the iterative nature of the design process and the triangulation with health care professionals allowed us to refine these contributions, prioritizing those that were more feasible and had greater potential applicability. In this way, HCD was useful not only for capturing perceptions and needs but also for guiding a process of selection and refinement oriented toward realistic technological solutions aligned with the precise needs of users.

Limitations and Adaptability Considerations

The methodology presented reflects our experience with a specific clinical context (diabetes dietary management) and available resources (an academic research team with engineering and clinical expertise). Teams adopting this approach should consider the following:

1. Resource allocation: Our timeline (8 months from mapping to testing) may need adjustment based on team size and funding.
2. Technical expertise: We chose web-based prototyping (Phaser 3) for accessibility; teams with different capabilities might select alternative frameworks.
3. User recruitment: Access to patient engagement boards or clinical partnerships will influence the feasibility of iterative testing.
4. Cultural and linguistic adaptation: Our Spanish-speaking population required specific interface considerations that will differ in other contexts.

The methodology is broadly applicable to other chronic disease management contexts and health care hardware development scenarios. By sharing our experiences, decision-making processes, and lessons learned, we aim to support other multidisciplinary teams in developing user-centered digital health solutions. Future work should explore adaptations of this approach for resource-constrained settings, integration with established clinical workflows, and methods for scaling from prototype to clinical implementation and evaluation.

Conclusions

This tutorial has provided step-by-step methodological guidance for applying HCD and the Double Diamond model to health care hardware development. Through the SMARTCLOTH case example, we have demonstrated how systematic user engagement, iterative prototyping, and structured usability testing can inform the development of digital health solutions tailored to patient needs. Key methodological contributions of this tutorial include (1) a replicable framework for user needs assessment combining multiple HCD tools (persona, empathy mapping, and customer journey mapping) that can be adapted to different chronic disease contexts; (2) strategies for synthesizing multidisciplinary input from clinical and technical team members through structured cocreation sessions; (3) cost-effective prototyping approaches using web-based frameworks (Phaser3/HTML5) that enable rapid iteration before committing to physical manufacturing; (4) structured usability testing protocols combining quantitative metrics (task completion time and error rates) with qualitative feedback

XSL•FO
RenderX

(think-aloud technique) across diverse user populations; and (5) practical guidance on adapting design requirements for users with varying technological literacy levels, including specific accommodations for older adults and users with limited digital experience.

In our specific application to diabetes dietary management, the resulting SMARTCLOTH prototype demonstrated strong usability and user acceptance across 3 distinct user profiles. The iterative development process, guided by continuous patient and professional input, yielded a tool perceived as intuitive and potentially valuable for dietary self-management support.

The methodology presented in this tutorial is broadly applicable to other chronic disease management contexts and health care hardware development scenarios where user-centered solutions are needed. By sharing our experiences, decision-making processes, and lessons learned—including both successes and limitations—we aim to support other multidisciplinary teams in developing effective digital health hardware.

Future applications of this methodology could explore adaptations for different resource settings, integration with existing clinical workflows and electronic health record systems, methods for scaling from prototype to clinical implementation, and approaches for conducting effectiveness evaluations that measure both usability and clinical outcomes. For SMARTCLOTH specifically, next steps include larger-scale usability testing, evaluation of clinical effectiveness on dietary adherence and glycemic control, and development of portable versions suitable for use beyond the home environment.

Acknowledgments

The authors thank all patients, caregivers, and health care professionals who participated in this study for their time and valuable contributions.

Funding

This research (PI21/01602) has been made possible because of the Strategic Health Action 2021 call, funded by the "Instituto de Salud Carlos III (Carlos III Health Institute)" and cofunded by the European Union.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the sensitive and potentially identifiable nature of these materials (qualitative data, including transcripts of interviews, focus groups, and video recordings) and to protect participant privacy and confidentiality but are available from the corresponding author on reasonable request and subject to appropriate ethical approval and data-sharing agreements.

Authors' Contributions

Conceptualization: GMR (lead), RML (equal)
Data curation: MPVE (focus groups), JMP (prototype testing)
Formal analysis: MGR (focus groups), FLG (prototype testing; lead), ICR (prototype testing; equal)
Funding acquisition: GMR
Investigation: GMR (focus groups, lead), RML (focus groups, equal), MPVE (focus groups, equal), JMP (prototype software design), FLG (prototype hardware design)
Methodology: GMR (lead), JMP (equal), FLG (equal)
Project administration: GMR (lead), RML (equal)
Resources: JMP (lead), FLG (equal), MGR (equal)
Software: JMP (lead), JMP (equal); IGR (supporting)
Supervision: GMR
Validation: GMR (lead), JMP (equal), RML (supporting)
Visualization: GMR (lead), FLG (supporting)
Writing–original draft: JMP, RML, FLG, and GMR (all four equally)
Writing–review & editing: RML and GMR

Conflicts of Interest

None declared.

Multimedia Appendix 1
General phases in the SMARTCLOTH development project. Adapted from Gasca and Zaragoza´, 2021.
[PNG File, 250 KB - jmir_v28i1e75744_app1.png]

Multimedia Appendix 2
Different user tests and patients using SMARTCLOTH.

XSL•FO
RenderX

[PNG File, 908 KB - jmir_v28i1e75744_app2.png]

Multimedia Appendix 3
Grid of actions for user test 2 task using SMARTCLOTH.
[PNG File, 248 KB - jmir_v28i1e75744_app3.png]

Multimedia Appendix 4
Results of In/Out and Strengths, Weaknesses, Opportunities, and Threats.
[PNG File, 1777 KB - jmir_v28i1e75744_app4.png]

Multimedia Appendix 5
Persona, empathy map, and consumer journey map.
[PDF File, 111657 KB - jmir_v28i1e75744_app5.pdf]

Multimedia Appendix 6
Tests 1 and 2 results.
[DOCX File, 17 KB - jmir_v28i1e75744_app6.docx]

Multimedia Appendix 7
Tests 3 and 4 results.
[DOCX File, 17 KB - jmir_v28i1e75744_app7.docx]

Multimedia Appendix 8
Test 5 results.
[DOCX File, 16 KB - jmir_v28i1e75744_app8.docx]

References

1. Brutsaert EF. Diabetes mellitus. MSD Manual. 2023 Oct. URL: https://acortar.link/jyD5sn [accessed 2024-05-29]
2. Diabetes. World Health Organziation. 2023 Apr. URL: https://acortar.link/EGwDd [accessed 2024-05-29]
3. IDF Diabetes Atlas 2025.: International Diabetes Federation URL: https://diabetesatlas.org/resources/idf-diabetes-atlas-2025/ [accessed 2025-09-22]
4. Nowakowska M, Zghebi SS, Ashcroft DM, et al. The comorbidity burden of type 2 diabetes mellitus: patterns, clusters and predictions from a large English primary care cohort. BMC Med 2019 Jul 25;17(1):145. [doi: 10.1186/s12916-019-1373-y] [Medline: 31345214]
5. ElSayed NA, McCoy RG, Aleppo G, et al. Summary of revisions: standards of care in diabetes—2025. Diabetes Care 2025 Jan 1;48(Supplement_1):S6-S13. [doi: 10.2337/dc25-SREV]
6. Lemieux I. Reversing type 2 diabetes: the time for lifestyle medicine has come!. Nutrients 2020 Jul 3;12(7):1974. [doi: 10.3390/nu12071974] [Medline: 32635141]
7. Sami W, Ansari T, Butt NS, Hamid MRA. Effect of diet on type 2 diabetes mellitus: a review. Int J Health Sci (Qassim) 2017;11(2):65-71. [Medline: 28539866]
8. Pérez Unanua MP, Alonso Fernández M, López Simarro F, et al. Adherence to healthy lifestyle behaviours in patients with type 2 diabetes in Spain. SEMERGEN 2021 Apr;47(3):161-169. [doi: 10.1016/j.semerg.2020.08.009] [Medline: 33160855]
9. Chong S, Ding D, Byun R, Comino E, Bauman A, Jalaludin B. Lifestyle changes after a diagnosis of type 2 diabetes. Diabetes Spectr 2017 Feb;30(1):43-50. [doi: 10.2337/ds15-0044] [Medline: 28270714]
10. Gillespie SJ, Kulkarni KD, Daly AE. Using carbohydrate counting in diabetes clinical practice. J Am Diet Assoc 1998 Aug;98(8):897-905. [doi: 10.1016/S0002-8223(98)00206-5] [Medline: 9710660]
11. Hanlon P, Daines L, Campbell C, McKinstry B, Weller D, Pinnock H. Telehealth interventions to support self-management of long-term conditions: a systematic metareview of diabetes, heart failure, asthma, chronic obstructive pulmonary disease, and cancer. J Med Internet Res 2017 May 17;19(5):e172. [doi: 10.2196/jmir.6688] [Medline: 28526671]
12. Moreno-Ligero M, Moral-Munoz JA, Salazar A, Failde I. mHealth intervention for improving pain, quality of life, and functional disability in patients with chronic pain: systematic review. JMIR Mhealth Uhealth 2023 Feb 2;11:e40844. [doi: 10.2196/40844] [Medline: 36729570]
13. Sakane N, Suganuma A, Domichi M, et al. The effect of a mHealth app (KENPO-app) for specific health guidance on weight changes in adults with obesity and hypertension: pilot randomized controlled trial. JMIR Mhealth Uhealth 2023 Apr 12;11:e43236. [doi: 10.2196/43236] [Medline: 37043287]
14. Vlahu-Gjorgievska E, Burazor A, Win KT, Trajkovik V. mHealth apps targeting obesity and overweight in young people. App review and analysis. JMIR Mhealth Uhealth 2023 Jan 19;11:e37716. [doi: 10.2196/37716] [Medline: 36656624]

XSL•FO
RenderX

15. Pulman A, Taylor J, Galvin K, Masding M. Ideas and enhancements related to mobile applications to support type 1 diabetes. JMIR Mhealth Uhealth 2013 Jul 25;1(2):e12. [doi: 10.2196/mhealth.2567] [Medline: 25100684]
16. Lamprinos I, Papadaki C, Schmuhl HH, Demski H, Hildebrand C, Plößnig M. Mobile personal health application for empowering diabetic patients. J Int Soc Telemed eHealth 2014;2(1):3-11 [FREE Full text]
17. Jakob R, Harperink S, Rudolf AM, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371. [doi: 10.2196/35371] [Medline: 35612886]
18. Wang X, Shu W, Du J, et al. Mobile health in the management of type 1 diabetes: a systematic review and meta-analysis. BMC Endocr Disord 2019 Feb 13;19(1):21. [doi: 10.1186/s12902-019-0347-6] [Medline: 30760280]
19. Kebede MM, Zeeb H, Peters M, Heise TL, Pischke CR. Effectiveness of digital interventions for improving glycemic control in persons with poorly controlled type 2 diabetes: a systematic review, meta-analysis, and meta-regression analysis. Diabetes Technol Ther 2018 Nov;20(11):767-782. [doi: 10.1089/dia.2018.0216] [Medline: 30257102]
20. Martínez-Pérez B, de la Torre-Díez I, López-Coronado M. Mobile health applications for the most prevalent conditions by the World Health Organization: review and analysis. J Med Internet Res 2013 Jun 14;15(6):e120. [doi: 10.2196/jmir.2600] [Medline: 23770578]
21. Molina-Recio G, Molina-Luque R, Romero-Saldaña M. The importance of knowing and listening to all those involved in the design and use of nutrition mobile apps. Getting to know the Great GApp. Nutr Hosp 2021 Jun 10;38(3):555-562. [doi: 10.20960/nh.03385] [Medline: 33813833]
22. Saparamadu A, Fernando P, Zeng P, et al. User-centered design process of an mHealth app for health professionals: case study. JMIR Mhealth Uhealth 2021 Mar 26;9(3):e18079. [doi: 10.2196/18079] [Medline: 33769297]
23. Lövestam E, Vivanti A, Steiber A, et al. The International Nutrition Care Process and Terminology Implementation Survey: towards a Global Evaluation Tool to assess individual practitioner implementation in multiple countries and languages. J Acad Nutr Diet 2019 Feb;119(2):242-260. [doi: 10.1016/j.jand.2018.09.004] [Medline: 30552017]
24. Atkins L, Michie S. Designing interventions to change eating behaviours. Proc Nutr Soc 2015 May;74(2):164-170. [doi: 10.1017/S0029665115000075] [Medline: 25998679]
25. Classification of digital interventions, services and applications in health: a shared language to describe the uses of digital technology for health, 2nd ed. World Health Organization. 2023. URL: https://www.who.int/publications/i/item/9789240081949 [accessed 2025-12-08]
26. Daly A, Hovorka R. Technology in the management of type 2 diabetes: present status and future prospects. Diabetes Obes Metab 2021 Aug;23(8):1722-1732. [doi: 10.1111/dom.14418] [Medline: 33950566]
27. Levander XA, VanDerSchaaf H, Barragán VG, et al. The role of human-centered design in healthcare innovation: a digital health equity case study. J Gen Intern Med 2024 Mar;39(4):690-695. [doi: 10.1007/s11606-023-08500-0] [Medline: 37973709]
28. Dopp AR, Parisi KE, Munson SA, Lyon AR. Aligning implementation and user-centered design strategies to enhance the impact of health services: results from a concept mapping study. Implement Sci Commun 2020;1(1):17. [doi: 10.1186/s43058-020-00020-w] [Medline: 32885179]
29. Kaur N, Pluye P. Delineating and operationalizing the definition of patient-oriented research: a modified e-Delphi study. J Patient Cent Res Rev 2019;6(1):7-16. [doi: 10.17294/2330-0698.1655] [Medline: 31414019]
30. Crowe B, Gaulton JS, Minor N, et al. To improve quality, leverage design. BMJ Qual Saf 2022 Jan;31(1):70-74. [doi: 10.1136/bmjqs-2021-013605] [Medline: 34510018]
31. Chen E, Neta G, Roberts MC. Complementary approaches to problem solving in healthcare and public health: implementation science and human-centered design. Transl Behav Med 2021 May 25;11(5):1115-1121. [doi: 10.1093/tbm/ibaa079] [Medline: 32986098]
32. Dopp AR, Parisi KE, Munson SA, Lyon AR. Integrating implementation and user-centred design strategies to enhance the impact of health services: protocol from a concept mapping study. Health Res Policy Sys 2019 Dec;17(1):1-13. [doi: 10.1186/s12961-018-0403-0]
33. Hookway S, Johansson MF, Svensson A, Heiden B. The problem with problems: reframing and cognitive bias in healthcare innovation. Des J 2019 Apr 1;22(sup1):553-574. [doi: 10.1080/14606925.2019.1595438]
34. Altman M, Huang TTK, Breland JY. Design thinking in health care. Prev Chronic Dis 2018 Sep 27;15:E117. [doi: 10.5888/pcd15.180128] [Medline: 30264690]
35. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119. [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
36. Sieck CJ, Sheon A, Ancker JS, Castek J, Callahan B, Siefer A. Digital inclusion as a social determinant of health. NPJ Digit Med 2021 Mar 17;4(1):52. [doi: 10.1038/s41746-021-00413-8] [Medline: 33731887]
37. Larusdottir M, Cajander Å, Roto V. User-centered design approaches and software development processes. In: Vanderdonckt J, Palanque P, Winckler M, editors. Handbook of Human Computer Interaction: Springer; 2023:1-4. [doi: 10.1007/978-3-319-27648-9_104-1]
38. Baker FW III, Moukhliss S. Concretising design thinking: a content analysis of systematic and extended literature reviews on design thinking and human-centred design. Rev Educ 2020 Feb;8(1):305-333. [doi: 10.1002/rev3.3186]

XSL•FO
RenderX

39. Rösch N, Tiberius V, Kraus S. Design thinking for innovation: context factors, process, and outcomes. Eur J Innov Manag 2023 Dec 18;26(7):160-176. [doi: 10.1108/EJIM-03-2022-0164]
40. Pomar P. Differences between design thinking and human centered design. Thinkernautas. 2017 May 27. URL: https://thinkernautas.com/diferencias-design-thinking-human-centered-design [accessed 2025-12-08]
41. Pressman R, Maxim B. Software Engineering: A Practitioner's Approach, 9th edition: McGraw-Hill Education; 2020.
42. The double diamond. Design Council. 2019. URL: https://www.designcouncil.org.uk/our-resources/the-double-diamond/ [accessed 2024-05-27]
43. Gasca J. Designpedia: 80 Herramientas Para Construir Tus Ideas [Designpedia: 80 Tools to Build Your Ideas], 5th edition: LID editorial; 2021.
44. Gasca J, Zaragozá R. El Workbook de Designpedia [Workbook of Designpedia], 1st edition: LID editorial.
45. Pruitt J, Adlin T. The Persona Lifecycle: Keeping People in Mind Throughout Product Design, 2nd edition: Morgan Kaufmann; 2006. [doi: 10.1016/B978-012566251-2/50002-2]
46. Gray D. The empathy map canvas. Gamestorming. 2017. URL: https://gamestorming.com/wp-content/uploads/2017/07/Empathy-Map-Canvas-006.pdf [accessed 2025-12-23]
47. Kalbach J. Mapping Experiences: A Complete Guide to Creating Value through Journeys, Blueprints, and Diagrams, 2nd edition: O'Reilly Media; 2020.
48. Lewrick M, Link P, Leifer L. The Design Thinking Toolbox: A Guide to Mastering the Most Popular and Valuable Innovation Methods: Wiley; 2020.
49. Stickdorn M, Hormess ME, Lawrence A, Schneider J. This Is Service Design Doing: Applying Service Design Thinking in the Real World, 2nd edition: O'Reilly Media; 2020. URL: https://www.thisisservicedesigndoing.com/ [accessed 2026-01-12]
50. Tan H, Wilson A, Olver I. Ricoeur's theory of interpretation: an instrument for data interpretation in hermeneutic phenomenology. Int J Qual Methods 2009 Dec;8(4):1-15. [doi: 10.1177/160940690900800401]
51. Osborn AF. Applied Imagination: Principles and Procedures of Creative Thinking, 4th edition: Creative Education Foundation; 2020.
52. Sanders EBN, Stappers PJ. Convivial Toolbox: Generative Research for the Front End of Design, 2nd edition: BIS Publishers; 2020.
53. McElroy K. Prototyping for Designers: Developing the Best Digital and Physical Products, 2nd edition: O'Reilly Media; 2017.
54. Rutter T. Designing Interfaces: Patterns for Effective Interaction Design, 3rd edition: O'Reilly Media; 2021.
55. Fonteyn ME, Kuipers B, Grobe SJ. A description of think aloud method and protocol analysis. Qual Health Res 1993 Nov;3(4):430-441. [doi: 10.1177/104973239300300403]
56. Ericsson KA, Simon HA. Protocol Analysis: Verbal Reports as Data Revised Ed: MIT Press; 2023.
57. Nielsen J, Molich R. Heuristic evaluation of user interfaces. 1990 Presented at: CHI '90: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; Apr 1-5, 1990; Seattle, WA p. 249-256. [doi: 10.1145/97243.97281]
58. Petersen M, Hempler NF. Development and testing of a mobile application to support diabetes self-management for people with newly diagnosed type 2 diabetes: a design thinking case study. BMC Med Inform Decis Mak 2017 Sep 12;17(1):133. [doi: 10.1186/s12911-017-0525-2] [Medline: 28899375]
59. Griffin AC, Khairat S, Bailey SC, Chung AE. A chatbot for hypertension self-management support: user-centered design, development, and usability testing. JAMIA Open 2023 Jul 4;6(3):ooad073. [doi: 10.1093/jamiaopen/ooad073]
60. Blandford A, Gibbs J, Newhouse N, Perski O, Singh A, Murray E. Seven lessons for interdisciplinary research on interactive digital health interventions. Digit Health 2018 Jan;4:2055207618770325. [doi: 10.1177/2055207618770325]
61. Krause-Jüttler G, Weitz J, Bork U. Interdisciplinary collaborations in digital health research: mixed methods case study. JMIR Hum Factors 2022 May 4;9(2):e36579. [doi: 10.2196/36579] [Medline: 35507400]
62. Fowe IE. Evaluating organizational readiness for change in the implementation of telehealth and mobile health interventions for chronic disease management. AMIA Jt Summits Transl Sci Proc 2021;2021(1):210-219. [Medline: 34457135]
63. Haldane V, Koh JJK, Srivastava A, et al. User preferences and persona design for an mHealth intervention to support adherence to cardiovascular disease medication in Singapore. JMIR Mhealth Uhealth 2019 May 28;7(5):e10465. [doi: 10.2196/10465] [Medline: 31140445]
64. Liu Y, Dickerson T, Early F, Fuld J, Jiang C, Clarkson PJ. Understanding the influences of COPD Patient's capability on the uptake of pulmonary rehabilitation in the UK through an inclusive design approach. Int J Chron Obstruct Pulmon Dis 2021;16(2):1717-1740. [doi: 10.2147/COPD.S305145] [Medline: 34168438]
65. Imhemed SM, Kamaruddin A, Atan R. Using empathy approach to design type 2 diabetic user persona. J Adv Sci Eng Res 2021;7(1):27-42 [FREE Full text]
66. Horta C, Quaresma G, Lucas P. Adherence to therapeutic regimen for people with diabetes through the implementation of continuous quality improvement projects – Scoping Review. New Trends Qual Res 2022;13:1-11. [doi: 10.36367/ntqr.13.2022.e678]
67. Mayberry LS, Osborn CY. Family involvement is helpful and harmful to patients' self-care and glycemic control. Patient Educ Couns 2014 Dec;97(3):418-425. [doi: 10.1016/j.pec.2014.09.011] [Medline: 25282327]

XSL•FO
RenderX

68. Mayberry LS, Osborn CY. Family support, medication adherence, and glycemic control among adults with type 2 diabetes. Diabetes Care 2012 Jun 1;35(6):1239-1245. [doi: 10.2337/dc11-2103]
69. Beverly EA, Ritholz MD, Brooks KM, et al. A qualitative study of perceived responsibility and self-blame in type 2 diabetes: reflections of physicians and patients. J Gen Intern Med 2012 Sep;27(9):1180-1187. [doi: 10.1007/s11606-012-2070-0] [Medline: 22549299]
70. Edupuganti S, Bushman J, Maditz R, Kaminoulu P, Halalau A. A quality improvement project to increase compliance with diabetes measures in an academic outpatient setting. Clin Diabetes Endocrinol 2019 Dec;5(1):15. [doi: 10.1186/s40842-019-0084-9]
71. Siopis G, Colagiuri S, Allman-Farinelli M. People with type 2 diabetes report dietitians, social support, and health literacy facilitate their dietary change. J Nutr Educ Behav 2021 Jan;53(1):43-53. [doi: 10.1016/j.jneb.2020.09.003] [Medline: 33077370]
72. Swarna Nantha Y, Kalasivan A, Ponnusamy Pillai M, et al. The validation of the Malay Yale Food Addiction Scale 2.0: factor structure, item analysis and model fit. Public Health Nutr 2020 Feb;23(3):402-409. [doi: 10.1017/S1368980019002684] [Medline: 31538554]
73. Al-Adwi ME, Al-Haswsa ZM, Alhmmadi KM, et al. Effects of different diets on glycemic control among patients with type 2 diabetes: a literature review. Nutr Health 2023 Jun;29(2):215-221. [doi: 10.1177/02601060221112805] [Medline: 35795964]
74. Paterson M, Bell KJ, O'Connell SM, Smart CE, Shafat A, King B. The role of dietary protein and fat in glycaemic control in type 1 diabetes: implications for intensive diabetes management. Curr Diab Rep 2015 Sep;15(9):61. [doi: 10.1007/s11892-015-0630-5] [Medline: 26202844]
75. Mertens E, Colizzi C, Peñalvo JL. Ultra-processed food consumption in adults across Europe. Eur J Nutr 2022 Apr;61(3):1521-1539. [doi: 10.1007/s00394-021-02733-7] [Medline: 34862518]
76. Sob C, Siegrist M, Hartmann C. The Positive Eating Scale: Associations with eating behavior, food choice, and body mass index. Eat Behav 2023 Jan;48:101706. [doi: 10.1016/j.eatbeh.2023.101706] [Medline: 36773373]
77. Bayram S, Kızıltan G, Akın O. Effect of adherence to carbohydrate counting on metabolic control in children and adolescents with type 1 diabetes mellitus. Ann Pediatr Endocrinol Metab 2020 Sep;25(3):156-162. [doi: 10.6065/apem.1938192.096] [Medline: 32871653]
78. Tascini G, Berioli MG, Cerquiglini L, et al. Carbohydrate counting in children and adolescents with type 1 diabetes. Nutrients 2018 Jan 22;10(1):109. [doi: 10.3390/nu10010109] [Medline: 29361766]
79. Uliana GC, Carvalhal M, Berino TN, et al. Adherence to carbohydrate counting improved diet quality of adults with type 1 diabetes mellitus during social distancing due to COVID-19. Int J Environ Res Public Health 2022 Aug 9;19(16):9776. [doi: 10.3390/ijerph19169776] [Medline: 36011412]
80. Farooqi MH, Abdelmannan DK, Al Buflasa MM, et al. The impact of telemonitoring on improving glycemic and metabolic control in previously lost-to-follow-up patients with type 2 diabetes mellitus: a single-center interventional study in the United Arab Emirates. Int J Clin Pract 2022;2022:6286574. [doi: 10.1155/2022/6286574] [Medline: 35685530]
81. Russell-Minda E, Jutai J, Speechley M, Bradley K, Chudyk A, Petrella R. Health technologies for monitoring and managing diabetes: a systematic review. J Diabetes Sci Technol 2009 Nov 1;3(6):1460-1471. [doi: 10.1177/193229680900300628] [Medline: 20144402]
82. Eberle C, Löhnert M, Stichling S. Effectiveness of disease-specific mhealth apps in patients with diabetes mellitus: scoping review. JMIR Mhealth Uhealth 2021 Feb 15;9(2):e23477. [doi: 10.2196/23477] [Medline: 33587045]
83. Stephen DA, Nordin A, Nilsson J, Persenius M. Using mHealth applications for self-care - an integrative review on perceptions among adults with type 1 diabetes. BMC Endocr Disord 2022 May 25;22(1):138. [doi: 10.1186/s12902-022-01039-x] [Medline: 35614419]
84. Moutou KE, England C, Gutteridge C, Toumpakari Z, McArdle PD, Papadaki A. Exploring dietitians' practice and views of giving advice on dietary patterns to patients with type 2 diabetes mellitus: a qualitative study. J Human Nutrition Diet 2022 Feb;35(1):179-190. [doi: 10.1111/jhn.12939]
85. Kahleova H, Znayenko-Miller T, Smith K, et al. Effect of a dietary intervention on insulin requirements and glycemic control in type 1 diabetes: a 12-week randomized clinical trial. Clin Diabetes 2024;42(3):419-427. [doi: 10.2337/cd23-0086] [Medline: 39015168]
86. Zhang S, Jiang H, Wang L, et al. Longitudinal relationship between body fat percentage and risk of type 2 diabetes in Chinese adults: evidence from the China Health and Nutrition Survey. Front Public Health 2022;10:1032130. [doi: 10.3389/fpubh.2022.1032130]
87. Stotz SA, Ricks KA, Eisenstat SA, Wexler DJ, Berkowitz SA. Opportunities for interventions that address socioeconomic barriers to type 2 diabetes management: patient perspectives. Sci Diabetes Self Manag Care 2021 Apr;47(2):153-163. [doi: 10.1177/01457217219962911] [Medline: 34078177]
88. Moore SE, McEvoy CT, Prior L, et al. Barriers to adopting a Mediterranean diet in Northern European adults at high risk of developing cardiovascular disease. J Human Nutrition Diet 2018 Aug;31(4):451-462. [doi: 10.1111/jhn.12523]
89. Monsivais P, Aggarwal A, Drewnowski A. Time spent on home food preparation and indicators of healthy eating. Am J Prev Med 2014 Dec;47(6):796-802. [doi: 10.1016/j.amepre.2014.07.033] [Medline: 25245799]

XSL•FO
RenderX

90. Vanstone M, Rewegan A, Brundisini F, Giacomini M, Kandasamy S, DeJean D. Diet modification challenges faced by marginalized and nonmarginalized adults with type 2 diabetes: a systematic review and qualitative meta-synthesis. Chronic Illn 2017 Sep;13(3):217-235. [doi: 10.1177/1742395316675024] [Medline: 27884930]
91. Adu MD, Malabu UH, Malau-Aduli AEO, Malau-Aduli BS. Enablers and barriers to effective diabetes self-management: a multi-national investigation. PLoS One 2019;14(6):e0217771. [doi: 10.1371/journal.pone.0217771] [Medline: 31166971]
92. Moström P, Ahlén E, Imberg H, Hansson PO, Lind M. Adherence of self-monitoring of blood glucose in persons with type 1 diabetes in Sweden. BMJ Open Diabetes Res Care 2017;5(1):e000342. [doi: 10.1136/bmjdrc-2016-000342] [Medline: 28611921]
93. Sousa C, Neves JS, Dias CC, Sampaio R. Adherence to glucose monitoring with intermittently scanned continuous glucose monitoring in patients with type 1 diabetes. Endocrine 2023 Mar;79(3):477-483. [doi: 10.1007/s12020-022-03288-1] [Medline: 36574148]
94. Martinez K, Frazer SF, Dempster M, Hamill A, Fleming H, McCorry NK. Psychological factors associated with diabetes self-management among adolescents with Type 1 diabetes: a systematic review. J Health Psychol 2018 Nov;23(13):1749-1765. [doi: 10.1177/1359105316669580] [Medline: 27663288]
95. Koenigsberg MR, Corliss J. Diabetes self-management: facilitating lifestyle change. Am Fam Physician 2017 Sep 15;96(6):362-370. [Medline: 28925635]
96. Vilafranca Cartagena M, Tort-Nasarre G, Rubinat Arnaldo E. Barriers and facilitators for physical activity in adults with type 2 diabetes mellitus: a scoping review. Int J Environ Res Public Health 2021 May 18;18(10):5359. [doi: 10.3390/ijerph18105359] [Medline: 34069859]
97. Ammann P, Offutt J. Introduction to Software Testing, 2nd edition: Cambridge University Press; 2016. [doi: 10.1017/9781316771273]
98. Luo J, Zhang K, Xu Y, Tao Y, Zhang Q. Effectiveness of wearable device-based intervention on glycemic control in patients with type 2 diabetes: a system review and meta-analysis. J Med Syst 2021 Dec 24;46(1):11. [doi: 10.1007/s10916-021-01797-6] [Medline: 34951684]
99. Toletti G, Boaretto A, Di Loreto C, Fornengo R, Gigante A, Perrone G. Enhancing diabetes therapy adherence: a comprehensive study on glucometer usability for type 2 diabetes patients. Front Clin Diabetes Healthc 2024;5:1328181. [doi: 10.3389/fcdhc.2024.1328181] [Medline: 38807703]

Abbreviations

DM: diabetes mellitus
DT: design thinking
HCD: human-centered design
SWOT: Strengths, Weaknesses, Opportunities, and Threats
T1DM: type 1 diabetes mellitus
T2DM: type 2 diabetes mellitus
UNESCO: United Nations Educational, Scientific and Cultural Organization

Edited by N Cahill; submitted 25.Apr.2025; peer-reviewed by C Hueso-Montoro, P Brauer; revised version received 07.Nov.2025; accepted 10.Nov.2025; published 21.Jan.2026.

Please cite as:
Palomares JM, Molina-Luque R, León-García F, Casares-Rodríguez I, García-Rodríguez M, Villena Esponera MP, Molina-Recio G SMARTCLOTH Prototype for Dietary Management in Patients With Diabetes Mellitus: Tutorial on Human-Centered Design Methodology for Health Care Hardware Development
J Med Internet Res 2026;28:e75744
URL: https://www.jmir.org/2026/1/e75744
doi:10.2196/75744

© Jose M Palomares, Rafael Molina-Luque, Fernando León-García, Irene Casares-Rodríguez, María García-Rodríguez, María Pilar Villena Esponera, Guillermo Molina-Recio. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Developing a Trauma-Informed Social Media Campaign to Disseminate Endometriosis-Specific Qualitative Art-Based Research Findings: Tutorial

Kerry Marshall[1,2*], RN, MN; Hargun Dhillon[3*]; A Fuchsia Howard[1,2], RN, PhD; Heather Noga[2], MA; Grace J Yang[3], BSc; William Zhu[3], BSc; Jessica Sutherland[4], BA; Sarah Lett[4]; Anna Leonova[2,3], MSc; Paul J Yong[2,3], MD, PhD; Natasha L Orr[1,2,3], PhD

[1]School of Nursing, University of British Columbia, Gateway Health Building, 5955 University Blvd., Vancouver, BC, Canada

[2]Women's Health Research Institute, BC Women's Hospital & Health Centre, Vancouver, BC, Canada

[3]Department of Obstetrics and Gynecology, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

[4]Patient Research Advisory Board, Endometriosis and Pelvic Pain Lab, University of British Columbia, Vancouver, BC, Canada

[*]these authors contributed equally

Corresponding Author:
A Fuchsia Howard, RN, PhD
School of Nursing, University of British Columbia, Gateway Health Building, 5955 University Blvd., Vancouver, BC, Canada

Abstract

Trauma-informed approaches can promote the creation of systems that prioritize safety and empowerment to improve patient well-being. These approaches are especially important in sexual and reproductive health care, where patients are often asked to disclose sensitive and personal information. This disclosure is particularly relevant in the context of endometriosis, a condition that affects 10% of reproductive-aged women and causes debilitating pelvic pain. Our team led a trauma-informed social media campaign to raise awareness and improve the understanding of endometriosis by sharing research findings from a photovoice study focusing on Asian women's experiences of endometriosis during the COVID-19 pandemic in Canada (*EndoPhoto Study*). In this paper, we describe how we adapted and applied trauma-informed approaches to the development and implementation of the social media campaign. To do this, we followed five adapted trauma-informed principles: (1) support and collaboration, (2) trustworthiness and transparency, (3) safety, (4) empowerment and voice, and (5) cultural and gender sensitivity, and four steps: (1) frame the campaign, (2) create content and manage the campaign, (3) measure campaign impact, and (4) conduct postcampaign reflections. We co-designed this campaign with patient partners having lived experience of endometriosis to facilitate support and collaboration. Additionally, we shared details about the funders of this study to increase trust and transparency, moderated comments and deidentified images to promote participant safety, chose safer platforms to enhance empowerment and voice, avoided stereotypes, and shared authentic experiences of Asian women with endometriosis to support cultural and gender sensitivity. The campaign launched on Instagram and Pinterest in March 2025 to coincide with Endometriosis Awareness Month. The social media campaign received 8,540,528 total impressions over the course of the month and had engagement rates of 6.23% and 1.4% on Instagram and Pinterest, respectively.

(J Med Internet Res 2026;28:e83491) doi:10.2196/83491

KEYWORDS

trauma-informed approach; social media; knowledge translation; endometriosis; information dissemination; content creation

Background and Rationale

Overview

Endometriosis is a chronic inflammatory condition characterized by the presence of endometrial-like tissue outside the uterus [1]. The symptoms may vary, but they often include severe pelvic pain, painful periods, painful sexual intercourse, and infertility [2]. Despite affecting approximately 10% of reproductive-aged women and girls, and an unmeasured number of gender diverse people, endometriosis remains significantly underdiagnosed and misunderstood [1,3]. Although diagnostic delays average 5 years in Canada, some individuals have reported a formal diagnosis taking up to 20 years [3,4]. The invisibility of symptoms, stigma surrounding sexual and menstrual health, and dismissal of women's pain all contribute to misinformation and present barriers to timely diagnosis and treatment [3,5,6]. Ultimately, these aspects all affect the mental and physical health of those with endometriosis. Furthermore, racialized populations may experience additional barriers to endometriosis diagnosis and care [7]. For instance, one study found that East and/or Southeast Asian women were 8 times

XSL•FO
RenderX

more likely than their White counterparts to experience severe disease before being referred to more specialized care [8].

Globally, the COVID-19 pandemic further exacerbated the gaps in endometriosis care as it upended the health care system, causing resource redirection toward patients with COVID-19 and interrupting the continuity of care for patients with chronic conditions like endometriosis [9,10]. In Canada, appointments and surgeries for people with endometriosis were postponed or canceled as hospitals became overwhelmed and health care providers transitioned to virtual environments [11]. Concurrently, mandatory self-isolation measures dramatically altered people's levels of social support, contributing to worsening psychological symptoms such as depression and anxiety [12]. Additionally, the COVID-19 pandemic was marked by a global rise in anti-Asian sentiment, with people of Asian descent reporting increasing episodes of violence and feelings of vulnerability to discrimination [13].

Given these compounding factors, our team conducted a study—the *EndoPhoto Study*—with 22 South, East, and/or Southeast Asian cisgender women with endometriosis in Canada to better understand the experiences of people in these communities during the COVID-19 pandemic. This study used photovoice, an art-based methodology that provides opportunities to use photos to share experiences and emotions related to stigmatized or hidden conditions [14]. Results from the *EndoPhoto Study* are published elsewhere [15,16] and highlight several key themes. These themes include the ways in which the pandemic exacerbated feelings of isolation and created additional challenges in accessing health care for those living with endometriosis. Participants also built resilience during the pandemic by accepting social support from peers, advocating for themselves in health care interactions, and taking empowering actions to self-manage their conditions. More details on the methodology and data analysis of the *EndoPhoto Study* are available in other studies [15,16]. The *EndoPhoto Study* was approved by the University of British Columbia Children's and Women's Research Ethics Board (reference number: H22-02390).

Findings from the *EndoPhoto Study* and our team's previous research highlight the importance of sharing evidence that validates the experiences of people affected by endometriosis, helps people feel they are not alone, fosters hope, and recognizes the strengths of those affected. Our team's pre-existing website showcased *EndoPhoto* results via images and quotes (*EndoPhoto* website [17]). The original website was co-created by researchers, clinicians, and patient partners to disseminate information and resources related to endometriosis. As guided by our Patient Research Advisory Board (PRAB; a group of people with lived experience of endometriosis), we chose to disseminate the *EndoPhoto Study* findings and *EndoPhoto* website to a public audience through a social media campaign. The goal of the campaign was to amplify the stories shared by Asian women regarding their experiences during the COVID-19 pandemic while focusing on disrupting silence related to the medical dismissal, social isolation, and cultural stigma of pelvic pain and endometriosis. As such, we recognized the relevance of using a trauma-informed approach to develop and implement the campaign. See Figure 1 for a project overview and the campaign development process.

XSL•FO
RenderX

Figure 1. Project overview and campaign development process. PRAB: Patient Research Advisory Board.

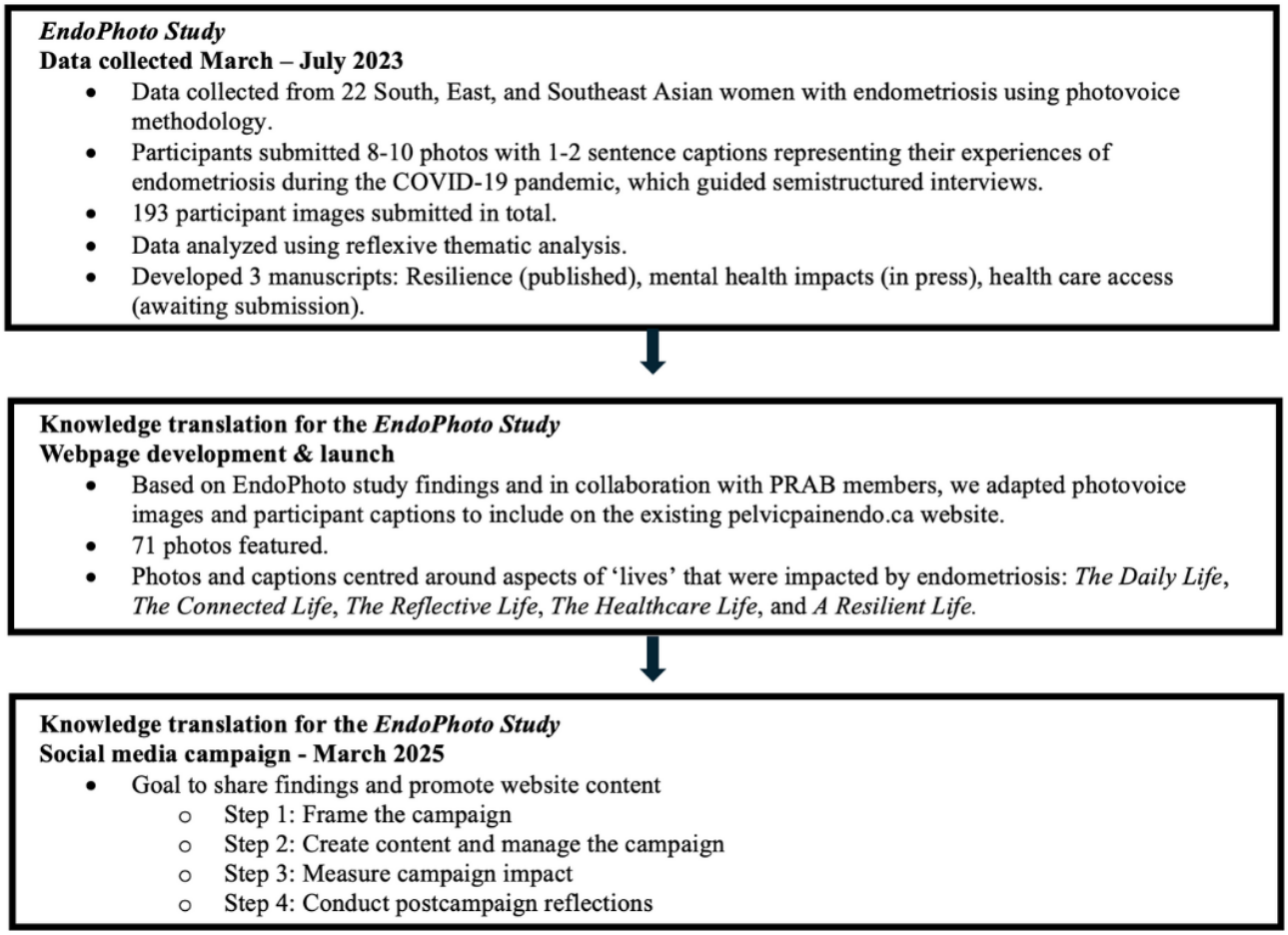

Objective

Our primary objective is to provide information on the steps we took when developing a social media campaign informed by the principles of a trauma-informed approach. Our secondary objective is to share the engagement results of the social media campaign. The target audience includes individuals and teams interested in trauma-informed social media campaigns, particularly those disseminating health-adjacent research findings. We begin by sharing information about our team, followed by information on trauma-informed approaches and social media dissemination. Lastly, we share the four steps that can be taken when developing a trauma-informed social media campaign: (1) frame the campaign, (2) create content and manage the campaign, (3) measure campaign impact, and (4) conduct postcampaign reflections.

Our Team

We are a multigenerational team with diverse genders, sexualities, and ethnicities, and are committed to improving the understanding and awareness of endometriosis through cutting-edge interdisciplinary research and knowledge translation. We recognize the importance of disseminating intentionally curated, evidence-based, and nuanced research findings to the endometriosis community and the public. Our team includes researchers, clinicians, health care trainees, and patient partners who are part of our PRAB. We are affiliated with the Endometriosis and Pelvic Pain Laboratory at the University of British Columbia, Canada.

What Are Trauma-Informed Approaches?

The formal conceptualization of trauma-informed care was first introduced by Harris and Fallot [18] in the context of mental health and substance use treatment systems. However, these principles have long-standing roots in community-based practices, including those within Indigenous traditions [19]. Since its inception, trauma-informed care has been adapted to various disciplines, with the framework of the Substance Abuse and Mental Health Services Administration (SAMHSA) often cited [20]. Trauma-informed approaches often acknowledge that trauma is widespread, and they actively support creating systems that promote physical and psychological safety [20].

SAMHSA defines trauma broadly, encompassing experiences at individual and structural levels that can be considered emotionally harmful [20]. SAMHSA's trauma-informed approach rests on four key assumptions: (1) realizing that trauma is widespread and can deeply affect individuals, communities, and societies; (2) recognizing the signs of trauma; (3) responding to trauma by integrating trauma-informed approaches; and (4) resisting retraumatization [20]. These assumptions are operationalized through six guiding principles: (1) safety; (2) trustworthiness and transparency; (3) peer support; (4) collaboration; (5) empowerment, voice, and choice; and (6) cultural, historical, and gender issues [20].

In health care settings, trauma-informed approaches in the provision of care have been shown to improve negative mental health symptoms and increase patient satisfaction, especially

XSL•FO
RenderX

among populations with histories of trauma or medical dismissal [21]. Trauma-informed approaches are particularly relevant in sexual and reproductive health, where patients are frequently asked to disclose sensitive information and are more likely to have experienced prior health care–related trauma [22]. Trauma-informed approaches are relevant for people with endometriosis as they have described feelings of shame and emotional distress related to their health care encounters where their symptoms have been diminished, normalized, or dismissed [23]. Interactions with health care systems and providers—as well as with broader public discourses that minimize people's experiences of endometriosis—have been further characterized as harmful, disempowering, and socially isolating. These experiences highlight the importance of using a trauma-informed approach that prioritizes safety, empowerment, and collaboration [23].

Social Media, Knowledge Dissemination, and Trauma-Informed Approaches

Social media is broadly defined as a digital space centered around information sharing and human connection [24]. Social media has increasingly become a pervasive aspect of everyday life and a powerful knowledge dissemination tool where various social media platforms, such as Facebook and TikTok, have been used by the health care community for patient education, peer support, and advocacy [25-27]. Social media platforms are participatory and easy to access, allowing the dissemination of information and rapid engagement of large, globally-connected audiences [28]. Content shared on social media platforms can provide unique opportunities to build a community, share experiences, and influence public health discourse [26].

There are also challenges with social media. For example, the nature of instant access to information and a lack of fact-verifying measures can lead to the unchecked and rapid spread of misinformation and disinformation to the public. Moreover, algorithms may incidentally lead to retraumatization and feelings of stigmatization [29-31]. Furthermore, considering people often consume social media content in isolation, it is difficult for content creators to recognize potential traumatization or retraumatization of viewers, highlighting the importance of intentionally creating and sharing content.

As knowledge dissemination of health-adjacent information increasingly moves onto social media platforms, trauma-informed approaches appear particularly relevant and potentially useful when working in these virtual environments. While there is limited guidance for applying trauma-informed approaches in digital spaces [32], the literature is emerging. We drew upon 3 frameworks that highlighted the potential of these approaches to reduce harm when sharing health-adjacent information digitally.

First, Josephs et al [33] emphasized 3 key pillars for digital trauma-informed design specific to sexual and reproductive health: privacy and confidentiality, intuitive and representative designs, and inclusive language. Second, *trauma-informed computing*, introduced by Chen et al [34], presents a framework guiding the adaptation of trauma-informed principles to digital design. This framework recognizes that digital tools can cause or exacerbate trauma and seeks to enable safer technological experiences [34]. Key adaptations of trauma-informed principles for online settings include safety, trustworthiness, peer support and collaboration, empowerment and choice, and cultural sensitivity.

Lastly, Scott et al [30] built upon the framework from Chen et al [34], adding specific aspects and examples to consider when applying trauma-informed approaches to social media engagement. The framework from Scott et al [30] outlines six guiding principles: (1) safety (eg, safe data collection and storage, and relaxing colors); (2) trustworthiness and transparency (eg, transparent about what user data are collected and why); (3) peer support (eg, protection for those sharing their unique stories); (4) collaboration and mutuality (eg, co-design with people having lived experience); (5) empowerment, voice, and choice (eg, no real names); and (6) cultural and historical gender issues (eg, acknowledge algorithmic biases).

Developing Our Trauma-Informed Social Media Campaign

Step 1: Frame the Campaign

Determine Relevant Trauma-Informed Principles

Based on the previously mentioned frameworks, we adapted our campaign to focus on the following principles of a trauma-informed approach: (1) support and collaboration, (2) trustworthiness and transparency, (3) safety, (4) empowerment and voice, and (5) cultural and gender sensitivity. These principles were used throughout the campaign development process. Table 1 provides an overview of how we approached incorporating trauma-informed principles throughout the campaign.

XSL•FO
RenderX

Table . Overview of our approach.

Trauma-informed guiding principle	Our approach and examples	Considerations

Support and collaboration	• Step 1: Frame the campaign
	  • Co-designed with PRAB[a] members and a diverse interdisciplinary team of experts
	  • Incorporated feedback throughout the process
	  • Shared content from lived experiences
	  • Engaged with influencers we had a previous relationship with to promote content
	• Step 4: Conduct postcampaign reflections
	  • Reflected as a team on successes, challenges, and learnings for future social media engagement
		• More time may be needed to include the feedback and ideas of all team members, and thus, there may be a longer timeline to project completion

Trustworthiness and transparency	• Step 1: Frame the campaign
	  • Informed participants of the purpose/content when consenting to the original study
	  • Obtained explicit and ongoing consent related to the use of data
	• Step 2: Create content and manage the campaign
	  • Created content highlighting our research team's positionality statement
	  • Created content highlighting transparency around funding and what this meant
	• Step 3: Measure campaign impact
	  • Intentionally collected metrics that were not overly invasive
	  • Intentionally collected metrics that were inclusive of multiple ways of engaging with content, recognizing that people may engage with content differently
		• Viewers may have personal negative feelings about funders
		• Additional human resources are needed to gather confirmatory consent from participants
		• Opt-out versus opt-in could incidentally include photos that participants did not want to share, but they did not see the email

Safety	• Step 1: Frame the campaign
	  • Engaged with specific platforms (Instagram and Pinterest)
	• Step 2: Create content and manage the campaign
	  • Created content using gentle, muted color palettes in content creation (eg, using light yellow)
	  • Created content in grouped images in collage format and shared select images
	  • Created content with deidentified images by blurring faces or including images where people were masked
	  • Moderated comments
	  • Included content warnings regarding sensitive topics
		• Participants may provide images that violate design principles and gentle colors
		• Additional human resources are required for moderating comments
		• Not including all experiences captured, as many images were not selected
		• Not as wide a reach due to including only select friendly platforms
		• Temporary stories reduce reach
		• Deidentifying images or using the photos without the original captions may alter the intended goal and impact of the image, limiting participant creativity

XSL•FO
RenderX

Trauma-informed guiding principle	Our approach and examples	Considerations

Empowerment and voice	• Step 1: Frame the campaign
	  • Leveraged platforms to amplify voices
	• Step 2: Create content and manage the campaign
	  • Created content that used friendly, nonstigmatizing, and everyday language when captioning photos
	  • Made changes to the content based on algorithm limitations while staying grounded in authentic patient voice and experiences
		• Potential for less reach when only using select platforms, and the reach to certain groups (eg, older people who more often use Facebook) might be reduced; demographics can vary across platforms
		• Participant caption is not always included with a photo, which could change the intended meaning
		• Focus on "positive" aspects due to the algorithm could represent a one-sided or skewed representation of experiences
		• Sharing content that is "trauma-informed" may incidentally portray the information and experiences as being neutral or positive

Cultural and gender sensitivity	• Step 2: Create content and manage the campaign
	  • Created content that ensured sharing only "positive" language and experiences due to the Meta algorithm
	  • Created content that avoided stereotypes and stigmatization
	  • Created content that avoided hypergendered content
		• Excluding images that depict too much pain or have identifying information may reduce the transparency of people's experiences
		• Some people might feel that the nonhypergendered colors and content do not relate to them as much

[a]PRAB: Patient Research Advisory Board.

Identify a Theoretical Approach

Theoretical approaches provide the foundation for a research study, offering structure and guidance when developing objectives, methods, and analysis [35]. Throughout the development of the campaign, including content creation, data interpretation, and dissemination strategies, we were primarily guided by the principles of intersectional feminism and integrated knowledge translation (IKT). An intersectional feminist perspective informed our understanding of how overlapping identities, such as race, gender, and country of origin, shape individuals' health care experiences. IKT is a collaborative process that emphasizes partnership between researchers and knowledge users throughout all stages of research [36]. IKT informed how we identified priorities, designed methods, interpreted data, and shared results [36]. Unlike traditional models that position researchers as the primary producers of knowledge, IKT recognizes the expertise of both researchers and community partners, aiming to minimize power differentials and promote equitable, contextually relevant knowledge creation [37].

Identify Campaign Goals and Messaging

Our main goal for the social media campaign was to share key research findings from the *EndoPhoto Study*, increase awareness of Asian women's experiences living with endometriosis, and direct viewers to our newly developed interactive *EndoPhoto* website [17]. In conducting the campaign, a second goal was to foster a sense of validation, emotional safety, and support for both previous research participants and audiences, and minimize their risk of retraumatization.

The campaign's central message—*your pain is real, you are believed, and you are not alone*—matched the overall messaging of content produced by our team and shared on the website. Additionally, this message was consistently emphasized across all digital platforms. The team deliberately declined to create a new hashtag for the campaign, given the objective of using social media as a mechanism to reach a wide audience of users rather than share isolated content with limited reach. As such, all the posts incorporated the widely recognized community hashtag #ThisIsEndo, supplemented by topic-relevant hashtags.

Determine When and Where to Launch the Campaign

We launched the campaign in March 2025 to coincide with Endometriosis Awareness Month. The Endometriosis and Pelvic Pain Laboratory had a previously established Instagram profile (@pelvicpainendo) with approximately 1000 followers before the campaign and had a Pinterest account (@pelvicpainendo), which was created for this and future campaigns. These platforms aligned with the campaign's visual and trauma-informed goals, offering features such as content warnings and comment moderation. The content shared was similar for both platforms but adapted in format to best use each platform's features. For example, Instagram's reels, stories, and carousels supported a balance of educational content and personal narratives, while Pinterest enabled thematic curation through boards and infographics. Pinterest has a different set of users and is more aligned with artistic communities.

Considering that Pinterest policies restrict the use of paid advertisements for new accounts, a soft launch on the platform began on February 19, 2025. Because of this, we published several posts prior to the full campaign, which allowed us to generate early interest, establish baseline engagement, and be considered an established account.

XSL•FO
RenderX

Step 2: Create Content and Manage the Campaign

Overview

The content that was shared during the social media campaign was initially designed by HD (quote posts), GJY (image posts), and WZ (videos/reels), who drew upon findings from the *EndoPhoto Study* and website content under the guidance of the announcement post creator and content lead (HN). Before final approval, all content was reviewed and discussed with the broader research team during team meetings every 2 weeks and PRAB meetings every 2 months. In total, the campaign featured 41 posts between March 1 and March 31, 2025, across Instagram and Pinterest. In this paper, we chose to share the images of content only from Instagram owing to the easy viewability of the entire post. In the following section, we discuss how we enacted these aspects in content creation and when managing the campaign. It is important to note that although these aspects are presented as being separate ideas, many have overlapping and related actions.

Incorporating Support and Collaboration

We defined support and collaboration as intentional actions taken to meaningfully engage with those having lived experience and others in the community doing similar work. As such, PRAB members defined campaign goals; reviewed and co-developed content; and ensured that materials were safe, empowering, and contextually relevant. The PRAB involvement created an opportunity for safety by foregrounding lived experience and ensuring that social media content reflected their values and perspectives. The campaign was also managed by an external communications agency, where the partner and social media specialist (SL) was someone with lived experience of endometriosis who simultaneously acted as a member of the PRAB.

We also took a collaborative approach to promoting the campaign, leveraging our known and existing networks. This involved approaching familiar social media accounts, including science communicators, nonprofit organizations, news outlets, and independent influencers, and asking them to share and promote our content. The campaign also partnered with advocacy organizations, including The Endometriosis Network Canada, to broaden outreach and ensure alignment with existing efforts in the endometriosis advocacy landscape.

Incorporating Trustworthiness and Transparency

When creating content, we considered trustworthiness and transparency aspects that involved disclosing who is behind the campaign, our goals, the funders, and where the content came from through "announcement" posts (Figure 2). Prior to the campaign, we were transparent with the participants of the *EndoPhoto Study* about how we were using their data.

Figure 2. Example of an announcement-style post on Instagram.

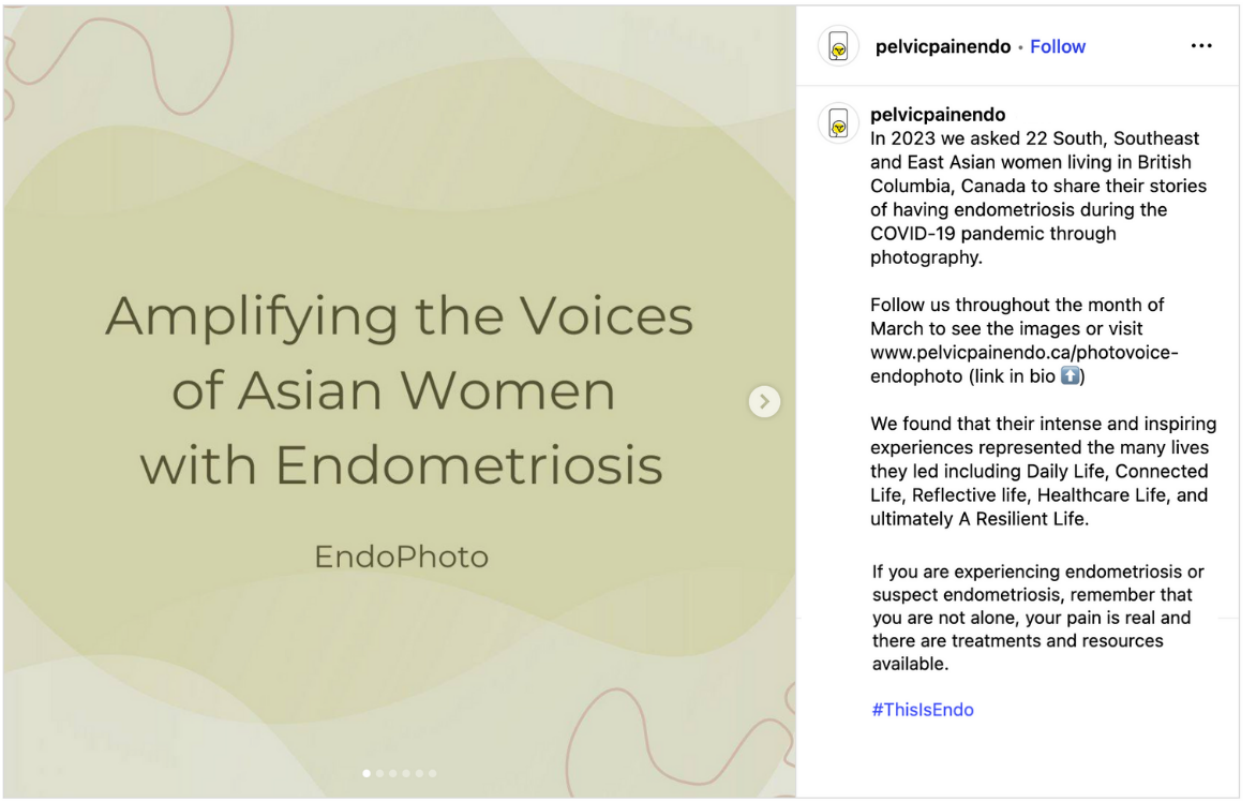

All participants in the *EndoPhoto Study* provided explicit informed consent to have their photos used in a social media campaign. We also provided participants with a lay summary of the results via email that included an "action required" message, showed participants the website [17] where photos had been included in a virtual gallery, and gave participants the ability to withdraw their photos and quotes at any point. We also strongly encouraged *EndoPhoto Study* participants to review their photos to ensure they were comfortable with these being shared. No participants opted out of their photos being shared.

XSL•FO
RenderX

Incorporating Safety

When considering safety, we focused on the principles of privacy and confidentiality from Josephs et al [33], and prioritized safety and preventing retraumatization of participants of the *EndoPhoto Study* whose images we shared. First, when creating content, to maintain the emotional safety of those who participated in the research and shared their photos, we intentionally curated visual content in a way that still honored participants' lived experiences. We did this by choosing images that were not intimately personal or overtly medical in nature, or that did not depict individual people in moments of visible distress. Instead, the campaign showcased strength-based visuals, such as nature scenes, symbolic objects, and comforting moments like participants' pets offering support (Figures 3 and 4).

Figure 3. Example of a collage of nature and social support.

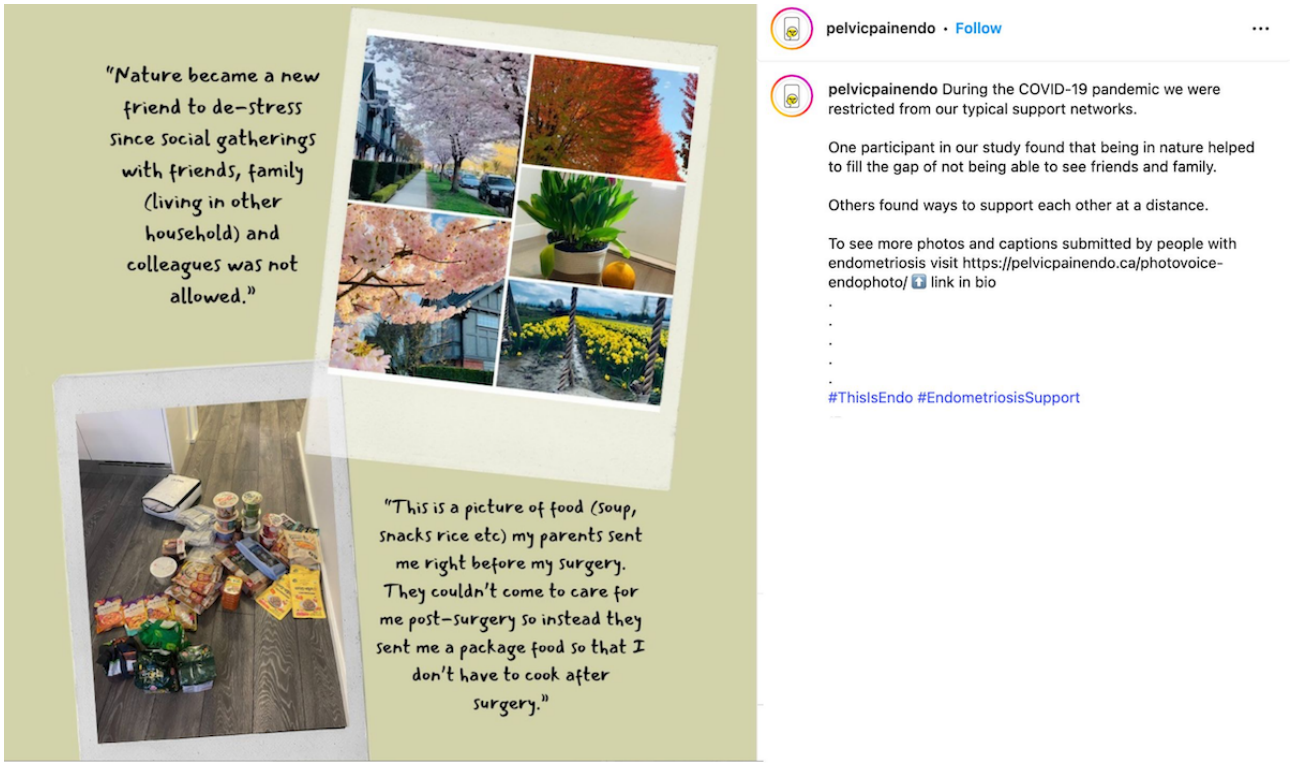

Figure 4. Example of a post about pets providing comfort.

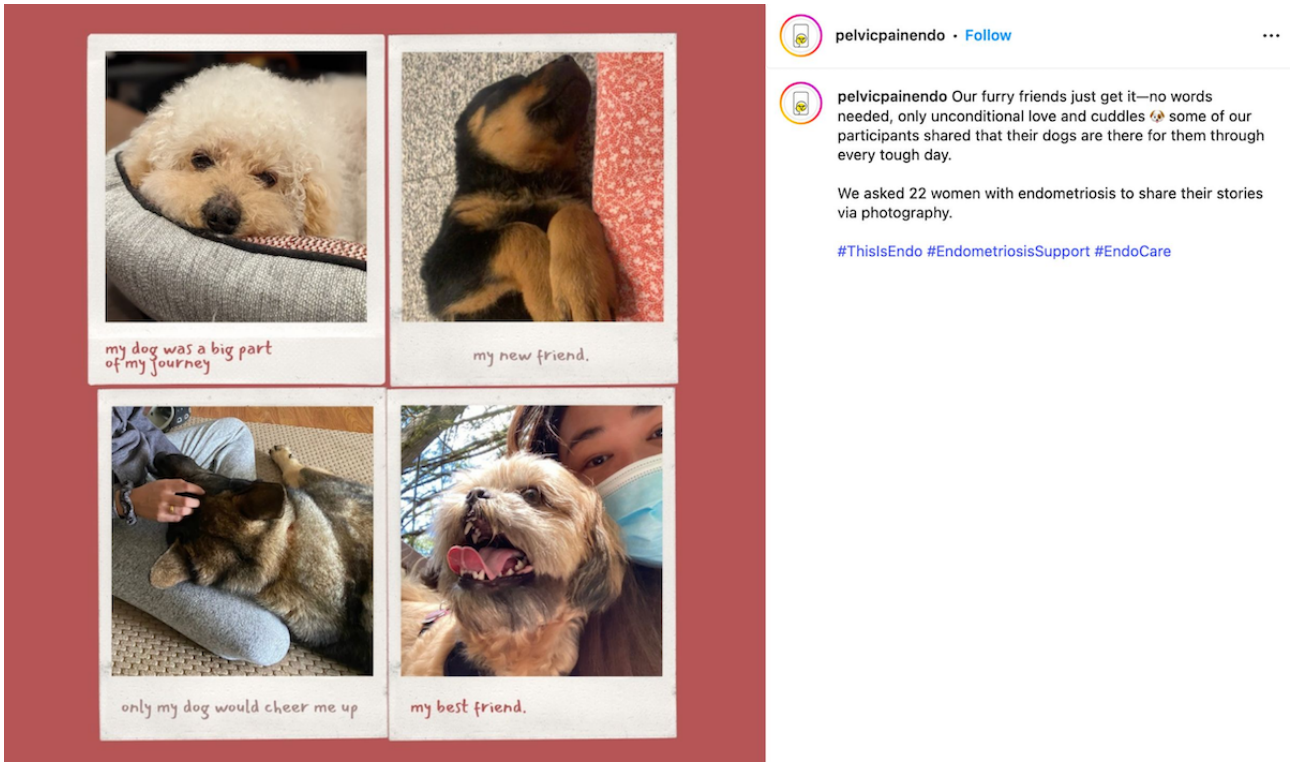

XSL•FO
RenderX

When possible, in content creation, images were paired with participant-authored captions that emphasized resilience, healing, and other personally meaningful themes to balance narrative authenticity with emotional safety. We also intentionally blurred faces to protect identities or selected images that further supported anonymity, such as those in which individuals used a surgical mask (Figures 4 and 5).

Figure 5. Example of identity protection on Instagram.



Second, when creating content, careful consideration was given to how participant photos were shared. We opted to present images as collages or grouped images rather than posting them individually (Figure 6). This approach was chosen to minimize the risk of certain photos receiving disproportionately more "likes" or "shares" than others, which could cause distress among some participants if they noticed their photos were less "liked."

XSL•FO
RenderX

Figure 6. Example of a shared collage of participant photos from an Instagram reel.

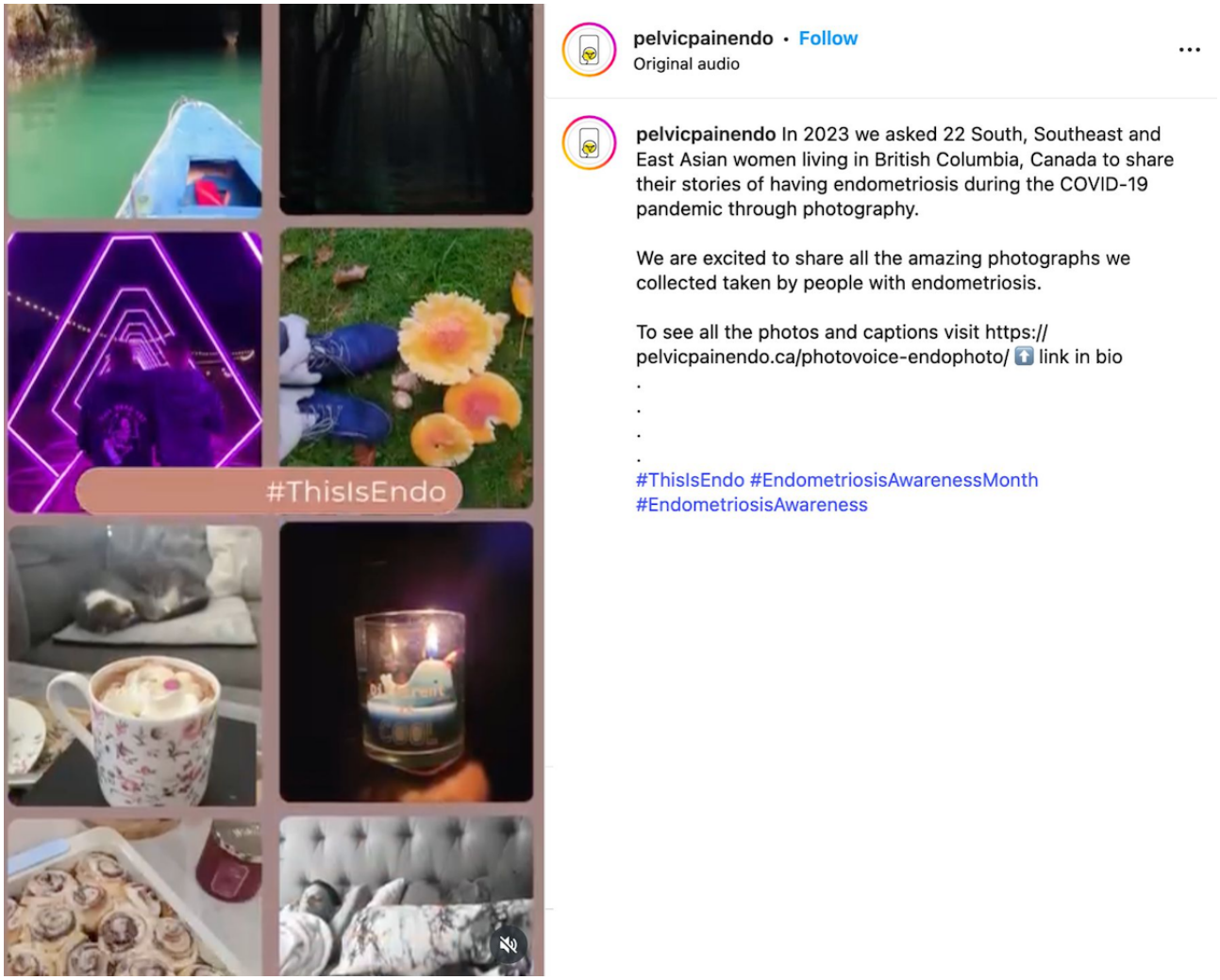

Third, based on recommendations from our social media specialist and patient partners, Instagram and Pinterest were specifically chosen as platforms for this campaign, given that they are well-suited to image-based storytelling and, anecdotally, were considered less volatile during the dates of the campaign, with lower rates of reproductive health online harassment compared with other well-known platforms.

Fourth, elements of safety were further considered through active moderation of comments to identify and remove hate speech, trolling, and unsolicited medical advice; however, we did not find that these were issues in this campaign.

Incorporating Empowerment and Voice

We considered empowerment and voice to focus on ensuring that we were truthfully representing participant experiences of the *EndoPhoto Study*, opting to create content that was more strength-based and showcased resilience and empowerment while also sharing the reality of experiencing endometriosis (Figures 7 and 8). In order to accomplish this, we created content that used nonstigmatizing, everyday language and often incorporated participants' own words in explaining the context of the photos (Figure 9).

XSL•FO
RenderX

Figure 7. Example of quote-based content on strength while living with endometriosis.

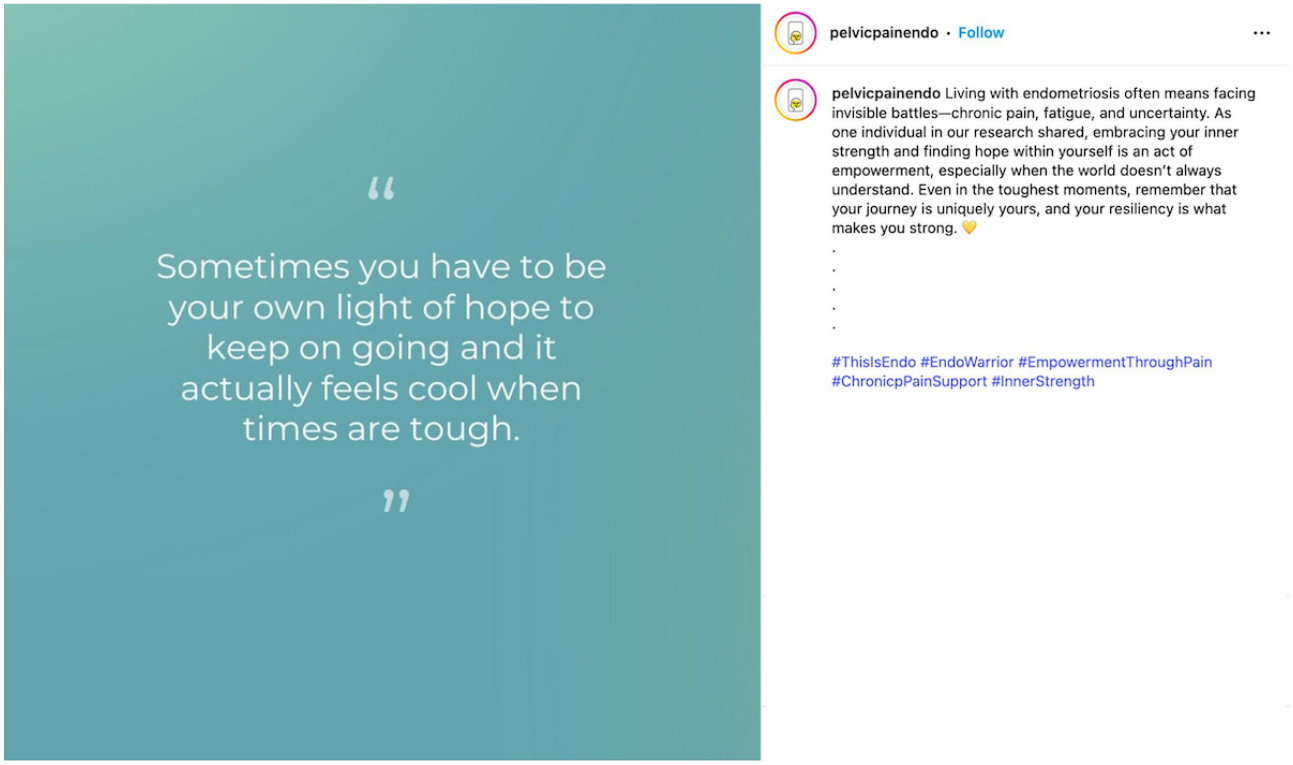

Figure 8. Example of quote-based content on living with endometriosis.

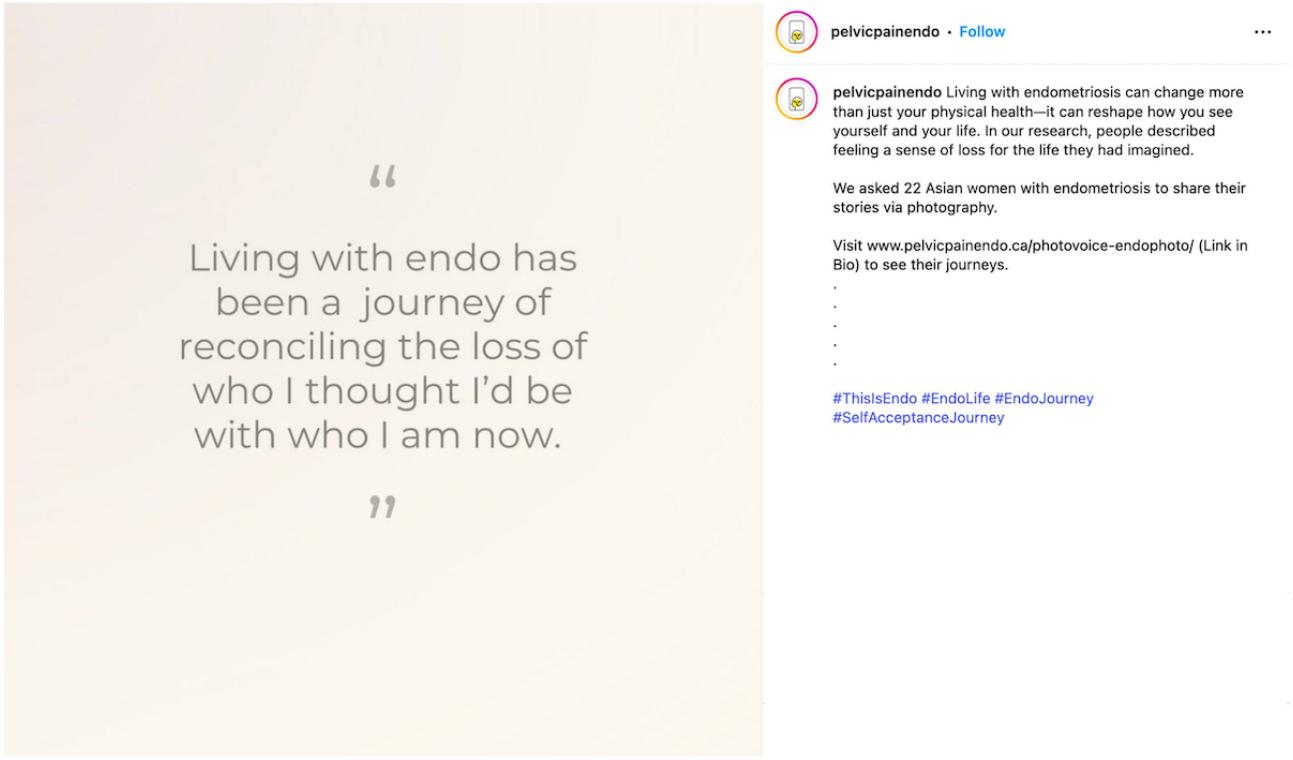

XSL•FO
RenderX

Figure 9. Example of everyday language use in a caption.

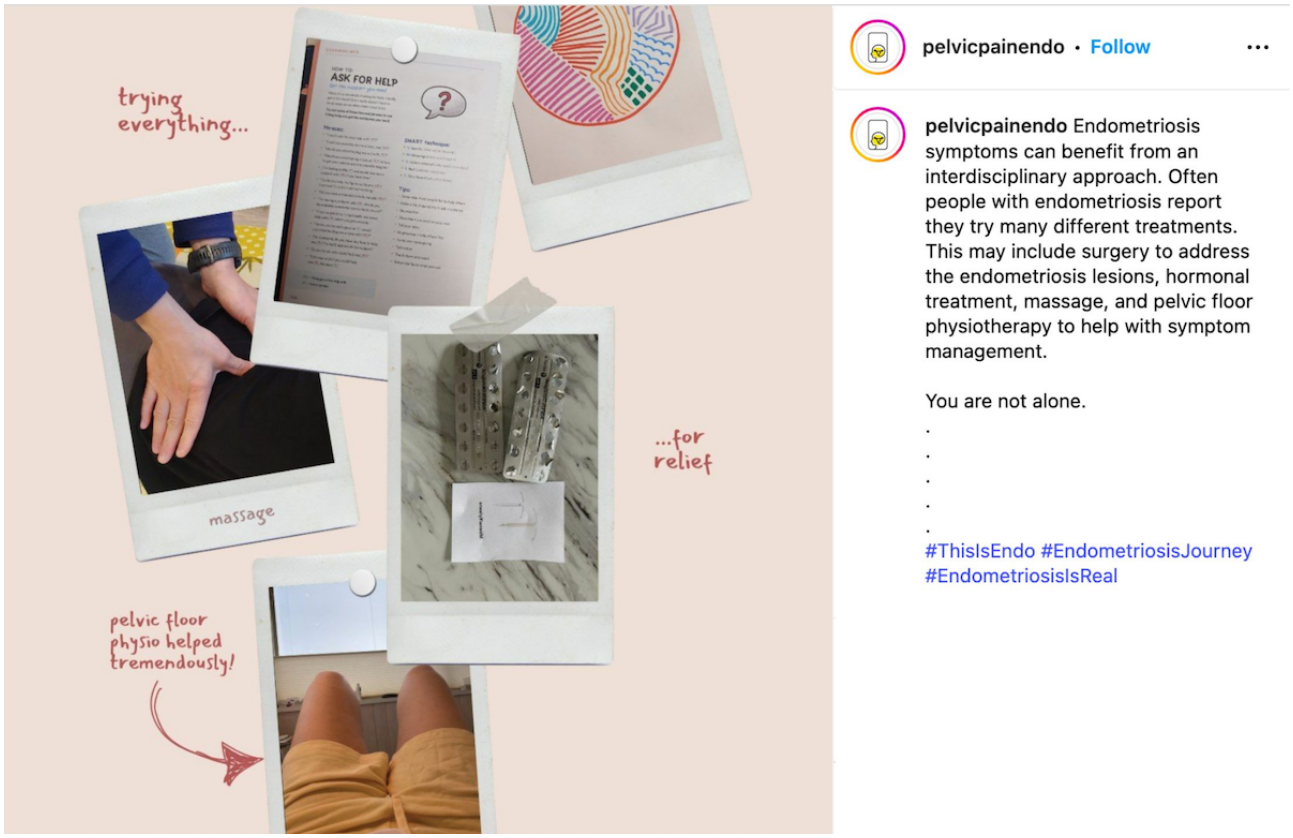

By integrating visual narratives with participant-authored captions, the campaign created opportunities for individuals to reclaim agency in narrating their health care experiences, particularly where medical and workplace systems had previously been invalidating. For example, 1 post focused on how cultural taboos surrounding menstruation can lead to a lack of communication and discussion of pain. This caption drew attention to the compounding effects of stress, isolation, and reduced access to care (Figure 10). Together, we intended these posts to help humanize the lived realities of people with endometriosis while fostering empathy, reducing stigma, and encouraging public dialogue.

XSL•FO
RenderX

Figure 10. Example of content highlighting cultural taboos around menstruation.

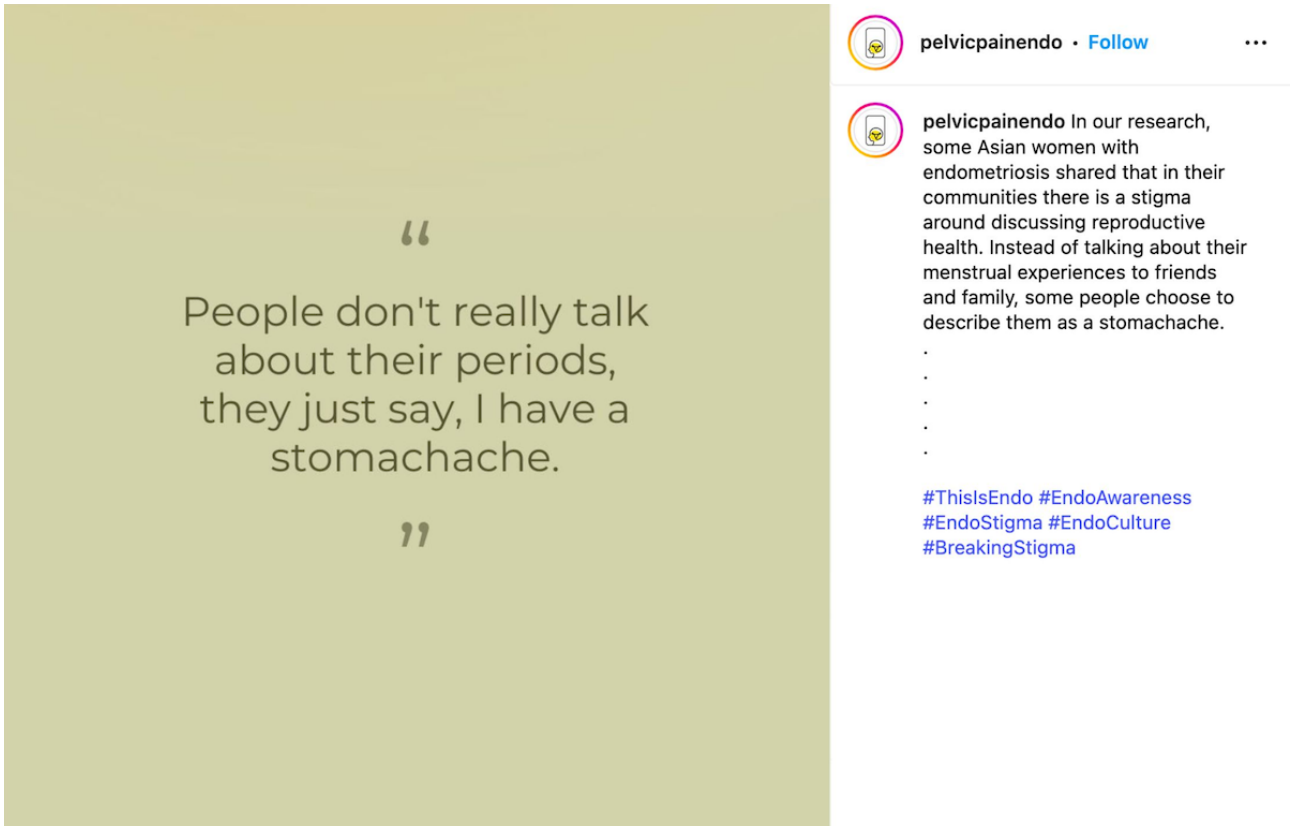

Incorporating Cultural and Gender Sensitivity

Striving for cultural and gender sensitivity, we carefully curated content to avoid sensationalism, incidental stigmatization, stereotypes, clinical or diagnostic language, and potentially distressing imagery. The campaign content aimed to disrupt the silence surrounding pelvic pain and endometriosis, particularly the effects of medical dismissal, social isolation, and cultural stigma. Considering the gendered nature of endometriosis—and although all the participants whose photos we shared identified as cisgender women—we intentionally avoided making the content hyperfeminized or gendered toward women exclusively. We also aimed to avoid perpetuating stereotypes and hyperfeminized content by choosing a color palette that was intentionally calming while not overly gendered (Figure 11).

XSL•FO
RenderX

Figure 11. Example of the avoidance of hypergendered or stigmatizing language or colors.

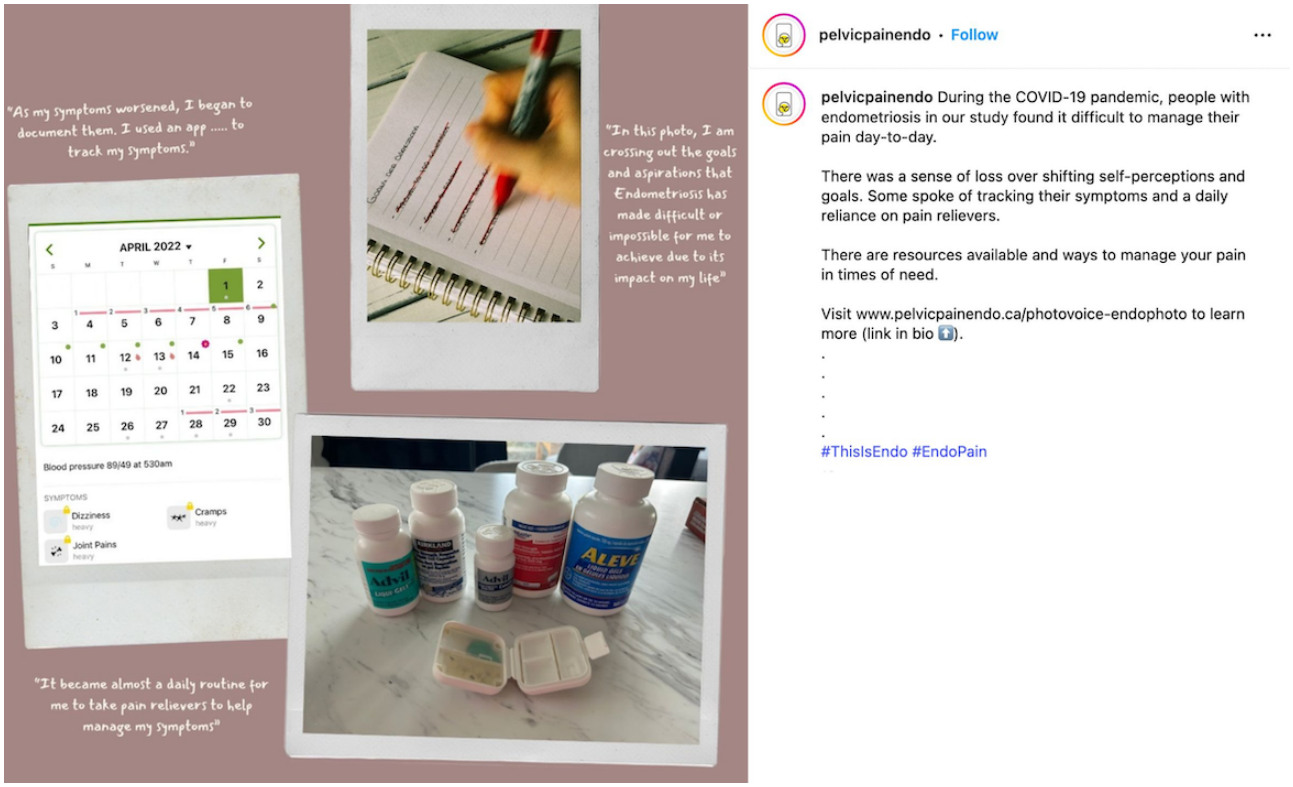

Step 3: Measure Campaign Impact

We used platform-integrated analytics (Instagram Insights and Pinterest Analytics) to monitor primary performance indicators. The metrics included reach, engagement (likes, shares, comments, profile visits, and link clicks), and website page visits (for definitions, see Table 2). We reviewed these metrics throughout the campaign to enable real-time optimization of posting frequency and timing, as well as advertising spend based on platform recommendations and observed audience behavior. Over the 31-day campaign, the website attracted 6326 unique users (for additional engagement measures, see Table 3).

Table . Consolidated definitions of terms used to describe campaign engagement.

Term	Definition
Reach	Reach is defined as the number of unique users who have seen the online content at least once
Engagement	Likes, shares, comments, profile visits, and link clicks
Impressions	The total number of times the content is presented to potential users on a screen
Volume of sessions	Number of visits to the website
Dwell time	The duration of time people view the content
Click-through rate	The percentage of people who click on a link to the website within the content

XSL•FO
RenderX

Table . Engagement results from Instagram and Pinterest.

Platform and metric	Value
Overall	
Total social impressions, n	8,540,528
Instagram	
Engagement[a], %	6.23
Advertisement impressions, n	7,941,457
Advertisement reach, n	3,550,309
Pinterest	
Total impressions, n	581,081
Total engagement, n	5528

[a]With regard to engagement, on Instagram, anything above 6% is considered high engagement, and on Pinterest, anything between 1% and 2% is considered average engagement [38,39].

Instagram generated both the greatest volume of sessions and the longest dwell time. Pinterest's shorter dwell time likely reflected the account's infancy (new profile, first campaign, and algorithmic learning period). The high number of impressions was largely attributed to 3 posts that "went viral" during the campaign, meaning they received over 1 million views each (Figures 12-14). These posts reflected messages of support, resiliency, and healing. Advertisements drew in the largest number of platform users to the page, accounting for most of the sessions and engagement, with arguably the least human resources.

Figure 12. Still image from a reel that "went viral" during the campaign.



XSL•FO
RenderX

Figure 13. A post that “went viral” during the campaign.

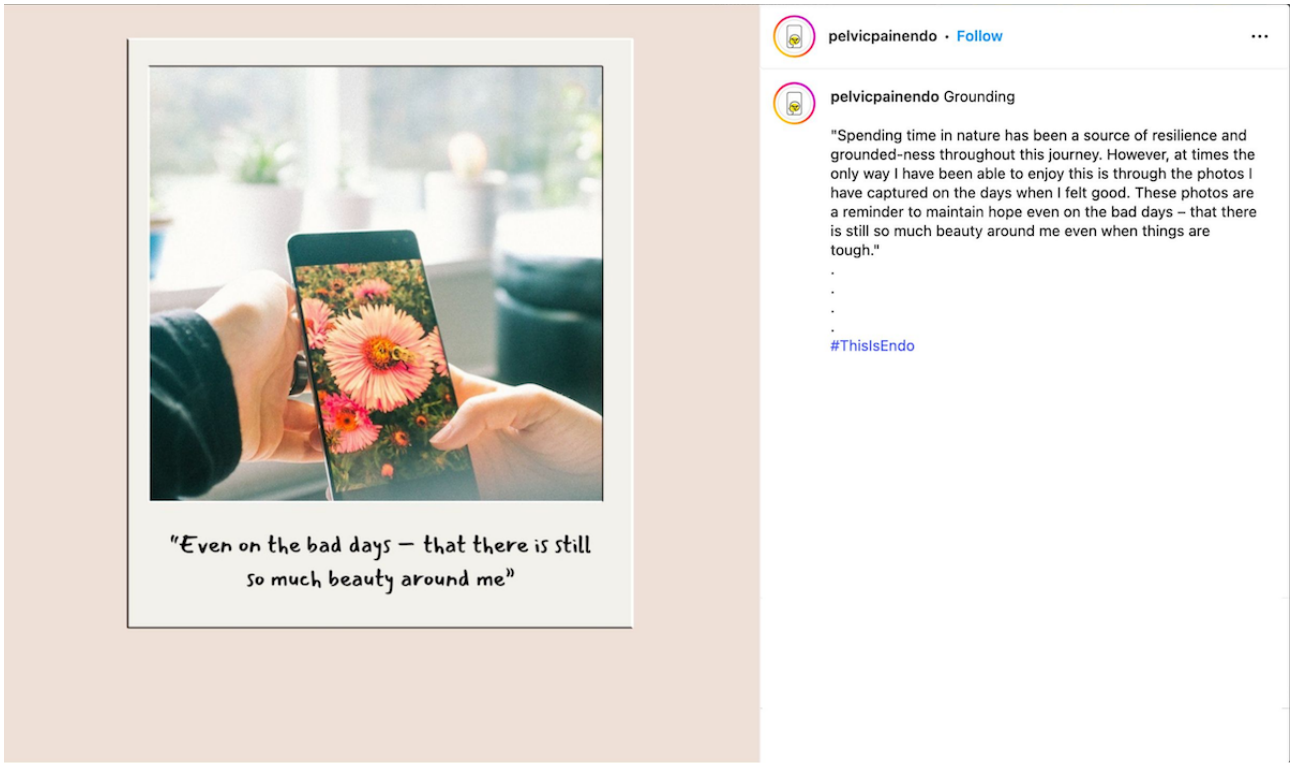

Figure 14. Still image taken from the second reel that “went viral” during the campaign.

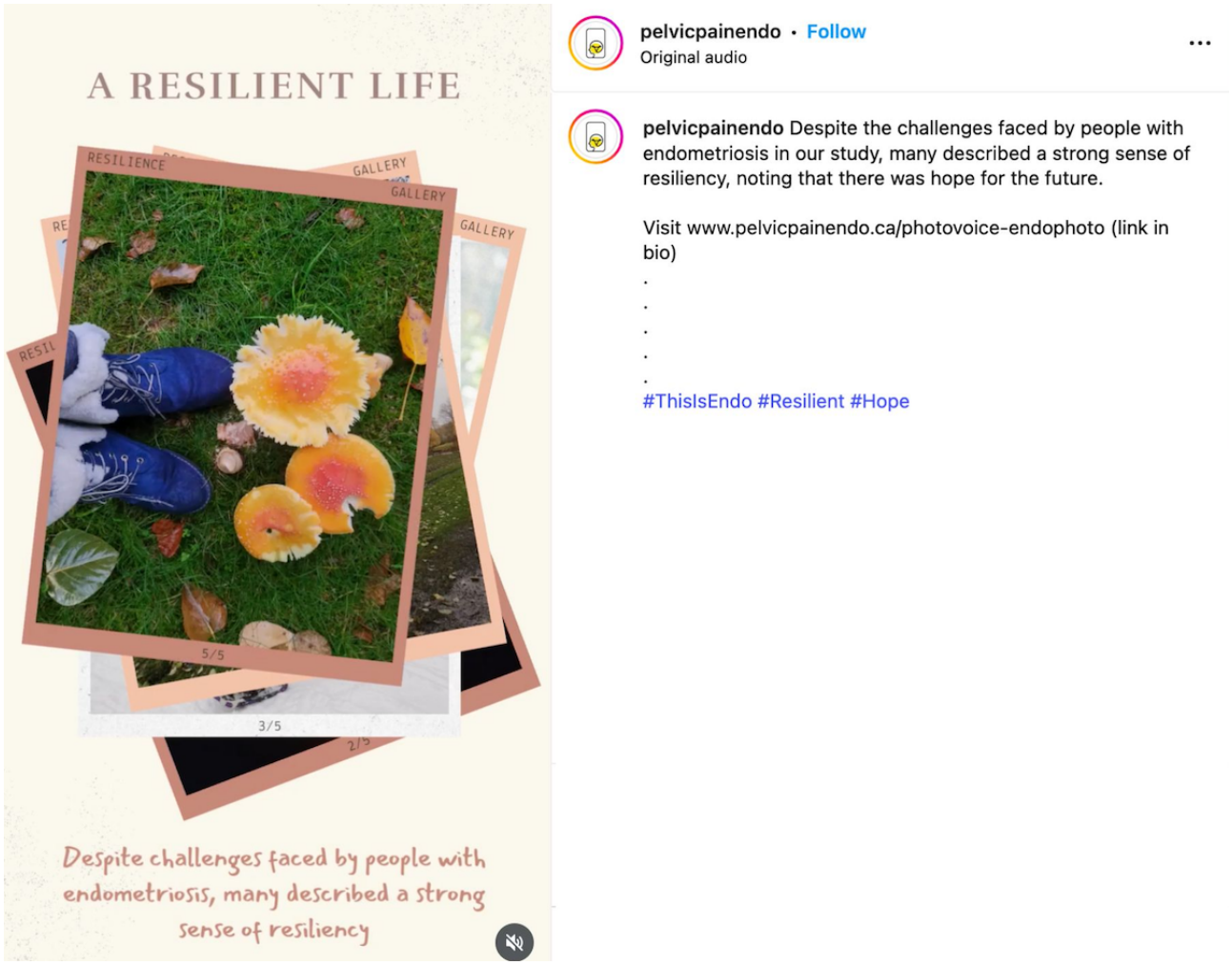

XSL•FO
RenderX

Step 4: Conduct Postcampaign Reflections

Identify What Went Well

To increase the reach of the social media campaign and the number of people seeing the online content at least once, we paid for advertisements on both platforms, with a total budget of CAD $3000 (US $2160). We found that advertisements required few human resources and were a relatively inexpensive way to increase reach. Posts that were advertised resulted in greater reach than posts that were not advertised. For our paid advertisements, a detailed audience profile was developed to guide content creation and advertisement targeting. The intended audience included individuals who either had a confirmed or suspected endometriosis diagnosis, were assigned female at birth, were of reproductive age (inclusive of all gender identities, sexual orientations, and relationship statuses), had moderate levels of health literacy, and understood English. We did not explicitly target Asian women and people living with endometriosis, as race and disease data are not captured for users by social media platforms explicitly. High-frequency search terms related to pelvic pain, endometriosis, and Asian experiences were identified to optimize advertisement discoverability (eg, pain, symptoms, self-care, journey, resilience, and reflection).

One of the campaign's primary strengths was its collaborative, interdisciplinary approach. The involvement of patient partners through the PRAB ensured content authenticity and emotional safety. In addition, working with a social media specialist was invaluable for navigating the landscape of digital media. The campaign also benefited from sufficient funding, allowing strategic investment in high-performing advertisement formats and continuous optimization based on analytics.

Our team's expertise, including content development led by those with lived, clinical, and research experience, added legitimacy, credibility, and multiple perspectives that countered misinformation in digital health spaces. Importantly, the campaign filled a representational void by centering the narratives of South, East, and Southeast Asian women with endometriosis, a demographic historically underrepresented in both research and advocacy.

Identify Challenges

Early in the campaign (March 3, 2025), our Instagram account was flagged by the Meta algorithm and included in a category called "Health and Wellness" owing to a post that was identified as being "negative" (Figure 15). This category was designed by Meta to reduce "negative" advertising and monetization of organizations that use advertising to sell products and services. Although we were not advertising products, our placement in this category limited what content could be promoted. Posts that included any features flagged as "negative" by the algorithm were prohibited from being advertised. Considering that endometriosis experiences often involve challenges, our preplanned content required significant changes in order to meet the algorithm's criteria.

Figure 15. An original post on Instagram that was flagged by the algorithm as being "negative".

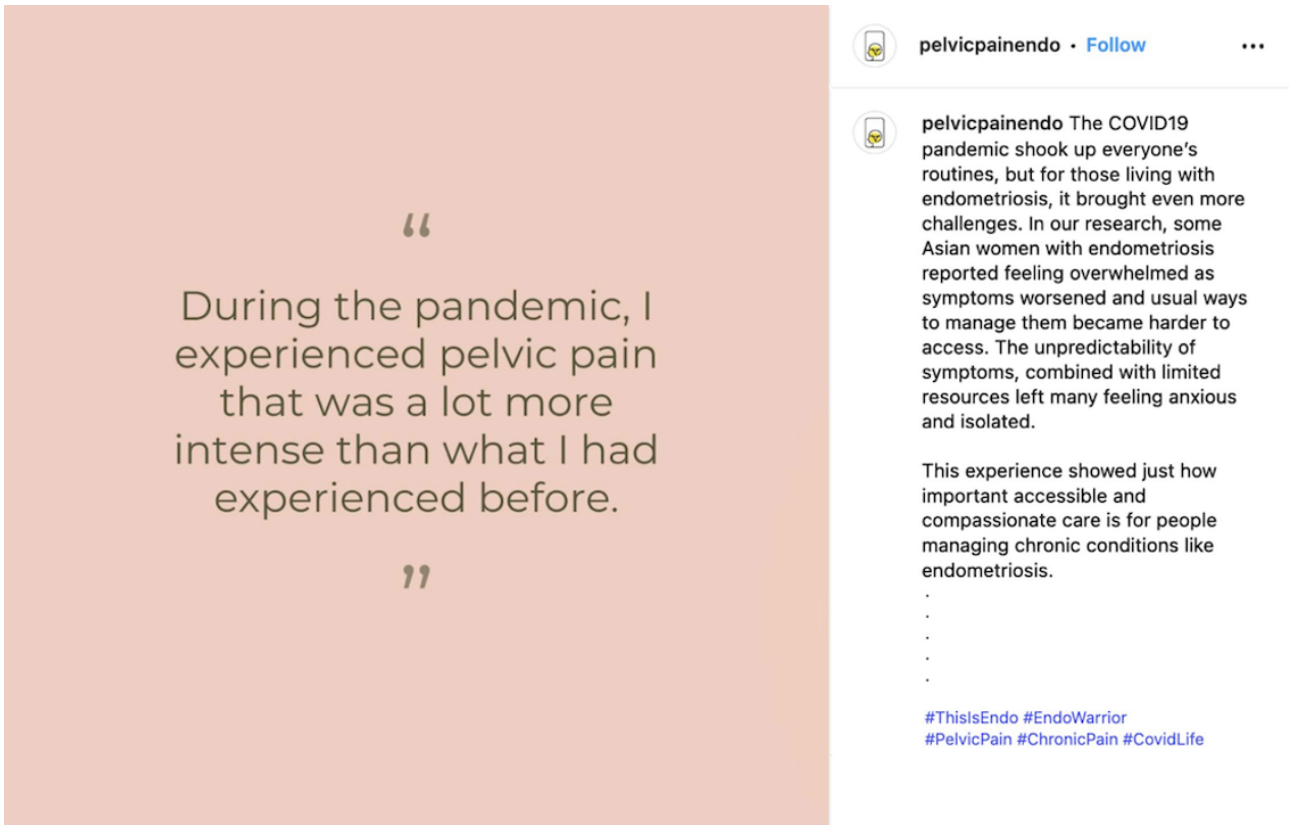

XSL•FO
RenderX

Discussion

Lessons Learned

This manuscript describes the development of a trauma-informed social media campaign designed to disseminate findings from the *EndoPhoto Study*, which explored the experiences of South, East, and Southeast Asian women living with endometriosis in Canada during the COVID-19 pandemic. We intend for this work to serve as a guide for others seeking to share their research findings through social media, with broader applicability for those interested in trauma-informed campaign development. By integrating trauma-informed principles, the campaign not only centered the voices of underrepresented communities but also demonstrated the potential for digital platforms to promote trauma-informed knowledge dissemination.

Upon reviewing our engagement metrics, we found that images with quote-based content (as opposed to image- or quote-only–based content) produced the highest click-through rates on both platforms; announcement posts (eg, study overview) generated the greatest engagement; and Instagram advertisements, particularly images with quotes, outperformed other creative formats (Table 4). Although it is challenging to measure how the content truly impacted viewers, some of these indicators may provide insights into the meaning viewers garnered from the content. For example, the option to "save" content on Pinterest may suggest that some of the posts resonated enough for viewers to want to share or review the content at a later date, perhaps indicating a feeling of added value.

Table . Examples of posts and engagement.

Type of content and example in the manuscript (figure)	Instagram reach	Instagram impressions	Pinterest impressions
Image-based			
Figure 5	484 unique viewers	658 presentations	—[a]
Quote-based			
Figure 8	295	422	259; 96,371 presentations with advertisements
Announcement style			
Figure 2	1396 unique viewers	2361 presentations	—
Reel			
Figure 6	509 unique viewers	840 presentations	—
Advertisement			
Figure 14	1,924,605 unique viewers	2,506,671 presentations	—
Figure 12	1,044,746 unique viewers	1,099,259 presentations	—
Figure 9	1,043,201 unique viewers	1,455,368 presentations	—
Figure 13	1,007,878 unique viewers	1,559,976 presentations	—
Figure 10	—	—	81,523 presentations
Figure 3	14,172 unique viewers	16,963 presentations	23,674 presentations

[a]Not applicable.

At this time, we are unable to say definitively as to why some post types had higher levels of reach and impressions (outside of advertisements). To the best of our knowledge, there is no research literature that shows which specific types of social media content tend to perform better, as it is highly based on the content and target audience. Some social media blogs have suggested that reels, carousel-type posts or those that have multiple images (our announcement posts), and relatable content tend to have higher engagement [40]. Although these content categories are broad, they could indicate that working as a team that has lived knowledge, clinical knowledge, and social media knowledge proved beneficial in understanding the types of content that may be engaging for our audience.

While trauma-informed principles are increasingly recognized in clinical and community settings, their application to digital media, particularly social media, remains largely underexplored. This tutorial highlights how principles, such as collaboration, safety, trustworthiness, voice, and cultural and gender sensitivity, can be applied in online spaces to mitigate harm and increase engagement. Additionally, this project helps to address the critical gap in the representation of racialized individuals, particularly Asian women, in endometriosis advocacy and online discourse.

One unanticipated lesson learned was the suppression of women's sexual and reproductive health information on Meta platforms. A recent report published by the Center for Intimacy Justice [41] highlighted that a bias exists on major Meta platforms, where organizations felt that their content and advertisements related to women's sexual and reproductive health, including fertility and pelvic health, were being censored and overmoderated. Social media algorithms, driven by artificial intelligence, limited content visibility when posts included

XSL•FO
RenderX

"sensitive" health-related terms or were deemed to violate vague platform policies such as Meta's "Personal Health and Appearance" guidelines. This report aligned with our experience, as our content was flagged as not aligning with community standards owing to its "negative" nature and association with health care. This flag necessitated a shift toward resilience-focused and positively framed messaging only, which may have constrained the full scope of participant narratives.

Limitations of the Campaign

Considering that platform selection was intentional, using only Instagram and Pinterest (chosen for their visual nature and perceived safety) may have excluded audiences who primarily engage with platforms like TikTok, X (formerly Twitter), or Facebook. Additionally, the metrics used to gauge success, such as views, impressions, and likes, offer limited insight into true impact. While some posts featuring animals gained viral traction, it is unclear whether the viral nature was due to their relevance to endometriosis or due to the important role that animals can play in people's lives, prompting questions about whether high engagement with content truly reflected increased awareness or understanding of endometriosis specifically.

In addition, because the campaign relied on Meta-owned platforms, our content was influenced by algorithmic restrictions that often suppressed posts containing sexual or reproductive health terms. This may have unintentionally narrowed the range of experiences we were able to highlight. While the campaign focused on the voices of Asian women, our sample does not capture the full diversity of people living with endometriosis, including those with different cultural backgrounds, gender identities, or varying levels of access to digital spaces. Finally, even with trauma-informed strategies in place, sharing health-related stories online carries the risk of re-exposure or secondary trauma for some viewers, especially when content reflects their own experiences.

Key Takeaways

The key takeaways are as follows:

- Trauma-informed principles can be adapted for digital health communication and effectively applied in social media campaigns.
- Posts that featured images with quotes, support networks (including pets), and announcements consistently maximized engagement, suggesting that these formats may be prioritized to ensure engagement.
- Early budget reallocation toward viral creative assets improved cost-efficiency.
- Paid advertisements created opportunities for ensuring wider reach and may be helpful in providing opportunities for content to be viewed by a larger audience.
- Inclusive online storytelling that prioritizes participant voice and emotional safety resonates with audiences and supports effective knowledge translation.
- Algorithmic biases targeted toward women's sexual and reproductive health may necessitate creativity to avoid messages being flagged or framing messaging in "positive" ways. This can ensure a wider reach but highlights gender bias within social media platforms.

Acknowledgments

We would like to thank the participants of the EndoPhoto Study who shared their stories and experiences with us. Moreover, we are grateful to have worked with Mass Velocity Media to develop this social media campaign. Generative artificial intelligence was not used in any form while writing this manuscript.

Funding

This work was funded by a 2024 Michael Smith Health Research BC Reach Award (RA-2024-04266).

Data Availability

Data sharing is not applicable to this article as no data sets were generated or analyzed during this study.

Authors' Contributions

Conceptualization: AFH, HN, NLO
Data curation: HN
Formal analysis: SL
Funding acquisition: KM, AFH, HN, JS, AL, PJY, NLO
Investigation: HN, SL
Methodology: KM, HD, AFH, HN, GJY, WZ, JS, SL, AL, PJY, NLO
Project administration: HN, NLO
Resources: HN
Supervision: KM, AFH, HN, NLO
Visualization: KM, HD, HN, GJY, WZ, AL, NLO
Writing – original draft: KM, HD, GJY, WZ, NLO
Writing – review & editing: KM, HD, AFH, HN, GJY, WZ, JS, SL, AL, PJY, NLO

XSL•FO
RenderX

Conflicts of Interest

There are no major conflicts of interest. However, SL acted as a patient partner and a paid consultant from Mass Velocity Media who contributed to the development and strategies of the social media campaign.

References

1. Zondervan KT, Becker CM, Missmer SA. Endometriosis. N Engl J Med 2020 Mar 26;382(13):1244-1256. [doi: 10.1056/NEJMra1810764] [Medline: 32212520]
2. Johnson NP, Hummelshoj L, Adamson GD, et al. World Endometriosis Society consensus on the classification of endometriosis. Hum Reprod 2017 Feb;32(2):315-324. [doi: 10.1093/humrep/dew293] [Medline: 27920089]
3. Wahl KJ, Yong PJ, Bridge-Cook P, Allaire C, EndoAct C. Endometriosis in Canada: it is time for collaboration to advance patient-oriented, evidence-based policy, care, and research. J Obstet Gynaecol Can 2021 Jan;43(1):88-90. [doi: 10.1016/j.jogc.2020.05.009] [Medline: 32753352]
4. Singh S, Soliman AM, Rahal Y, et al. Prevalence, symptomatic burden, and diagnosis of endometriosis in Canada: cross-sectional survey of 30 000 women. J Obstet Gynaecol Can 2020 Jul;42(7):829-838. [doi: 10.1016/j.jogc.2019.10.038] [Medline: 32001176]
5. Greene R, Stratton P, Cleary SD, Ballweg ML, Sinaii N. Diagnostic experience among 4,334 women reporting surgically diagnosed endometriosis. Fertil Steril 2009 Jan;91(1):32-39. [doi: 10.1016/j.fertnstert.2007.11.020] [Medline: 18367178]
6. Sims OT, Gupta J, Missmer SA, Aninye IO. Stigma and endometriosis: a brief overview and recommendations to improve psychosocial well-being and diagnostic delay. Int J Environ Res Public Health 2021 Aug 3;18(15):8210. [doi: 10.3390/ijerph18158210] [Medline: 34360501]
7. Bougie O, Nwosu I, Warshafsky C. Revisiting the impact of race/ethnicity in endometriosis. Reprod Fertil 2022 Apr 1;3(2):R34-R41. [doi: 10.1530/RAF-21-0106] [Medline: 35514542]
8. Williams C, Long AJ, Noga H, et al. East and South East Asian ethnicity and moderate-to-severe endometriosis. J Minim Invasive Gynecol 2019;26(3):507-515. [doi: 10.1016/j.jmig.2018.06.009] [Medline: 29935381]
9. Kabani Z, Ramos-Nino ME, Ramdass P. Endometriosis and COVID-19: a systematic review and meta-analysis. Int J Mol Sci 2022 Oct 26;23(21):12951. [doi: 10.3390/ijms232112951] [Medline: 36361745]
10. Demetriou L, Cox E, Lunde CE, et al. The global impact of COVID-19 on the care of people with endometriosis. Front Glob Womens Health 2021;2:662732. [doi: 10.3389/fgwh.2021.662732] [Medline: 34816218]
11. Leonardi M, Horne AW, Vincent K, et al. Self-management strategies to consider to combat endometriosis symptoms during the COVID-19 pandemic. Hum Reprod Open 2020;2020(2):hoaa028. [doi: 10.1093/hropen/hoaa028] [Medline: 32509977]
12. Schwab R, Stewen K, Kottmann T, et al. Mental health and social support are key predictors of resilience in German women with endometriosis during the COVID-19 pandemic. J Clin Med 2022 Jun 26;11(13):3684. [doi: 10.3390/jcm11133684] [Medline: 35806968]
13. Leigh JP, Moss SJ, Tiifu F, et al. Lived experiences of Asian Canadians encountering discrimination during the COVID-19 pandemic: a qualitative interview study. CMAJ Open 2022;10(2):E539-E545. [doi: 10.9778/cmajo.20220019] [Medline: 35700997]
14. Han CS, Oliffe JL. Photovoice in mental illness research: a review and recommendations. Health (London) 2016 Mar;20(2):110-126. [doi: 10.1177/1363459314567790] [Medline: 25673051]
15. Marshall K, Howard AF, Marshall N, Noga H, Rojas HE, Leonova A, et al. Impacts of the COVID-19 pandemic on the mental health of asian women with endometriosis in canada: a photovoice study. SAGE Women's Health (forthcoming) 2025.
16. Marshall N, Howard AF, Marshall K, et al. Endometriosis and expressions of self-management and resilience among asian women living in canada during the COVID-19 pandemic: a photovoice study. J Public Health Res 2026 Jan;15(1). [doi: 10.1177/22799036251407192]
17. EndoPhoto. Endometriosis and Pelvic Pain Laboratory. URL: https://pelvicpainendo.ca/photovoice-endophoto [accessed 2025-07-15]
18. Harris M, Fallot RD. Envisioning a trauma-informed service system: a vital paradigm shift. New Dir Ment Health Serv 2001(89):3-22. [doi: 10.1002/yd.23320018903] [Medline: 11291260]
19. Makosis P, Greenwood M. What's new is really old: trauma-informed health practices through an understanding of historic trauma. National Collaborating Centre for Indigenous Health (NCCIH). 2017. URL: https://www.nccih.ca/495/Webinar__What_s_new_is_really_old__Trauma_informed_health_practices_through_an_understanding_of_historic_trauma.nccah?id=205 [accessed 2026-01-14]
20. SAMHSA's concept of trauma and guidance for a trauma-informed approach. : Substance Abuse and Mental Health Services Administration; 2014 URL: https://www.health.ny.gov/health_care/medicaid/program/medicaid_health_homes/docs/samhsa_trauma_concept_paper.pdf [accessed 2026-01-14]

XSL•FO
RenderX

21. Raja S, Hasnain M, Hoersch M, Gove-Yin S, Rajagopalan C. Trauma informed care in medicine: current knowledge and future research directions. Fam Community Health 2015;38(3):216-226. [doi: 10.1097/FCH.0000000000000071] [Medline: 26017000]
22. Caring for patients who have experienced trauma: ACOG Committee opinion, number 825. Obstet Gynecol 2021 Apr 1;137(4):e94-e99. [doi: 10.1097/AOG.0000000000004326] [Medline: 33759830]
23. Parmar G, Howard AF, Noga H, et al. Pelvic pain & endometriosis: the development of a patient-centred e-health resource for those affected by endometriosis-associated dyspareunia. BMC Med Inform Decis Mak 2025 Feb 13;25(1):79. [doi: 10.1186/s12911-025-02907-x] [Medline: 39948529]
24. Burgess J, Marwick A, Poell T. Editors' introduction. In: Burgess J, Marwick A, Poell T, editors. The SAGE Handbook of Social Media: SAGE Publications; 2018:1-10. [doi: 10.4135/9781473984066.n1]
25. Davis JL. Social media. In: Mazzoleni G, editor. The International Encyclopedia of Political Communication: John Wiley & Sons, Inc; 2016. [doi: 10.1002/9781118541555.wbiepc004]
26. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85. [doi: 10.2196/jmir.1933] [Medline: 23615206]
27. Ezeilo CO, Leon N, Jajodia A, Han HR. Use of social media for health advocacy for digital communities: descriptive study. JMIR Form Res 2023 Nov 14;7:e51752. [doi: 10.2196/51752] [Medline: 37962914]
28. Shawky S, Kubacki K, Dietrich T, Weaven S. Using social media to create engagement: a social marketing review. JSOCM 2019 Apr 8;9(2):204-224. [doi: 10.1108/JSOCM-05-2018-0046]
29. Aïmeur E, Amri S, Brassard G. Fake news, disinformation and misinformation in social media: a review. Soc Netw Anal Min 2023;13(1):30. [doi: 10.1007/s13278-023-01028-5] [Medline: 36789378]
30. Scott CF, Marcu G, Anderson RE, Newman MW, Schoenebeck S. Trauma-informed social media: towards solutions for reducing and healing online harm. Presented at: 2023 CHI Conference on Human Factors in Computing Systems; Apr 23-28, 2023. [doi: 10.1145/3544548.3581512]
31. Abdulai AF, Howard AF, Yong PJ, Currie LM. Defining destigmatizing design guidelines for use in sexual health-related digital technologies: a Delphi study. PLOS Digit Health 2023 Jul;2(7):e0000223. [doi: 10.1371/journal.pdig.0000223] [Medline: 37436972]
32. Zheng W, Walquist E, Datey I. "It's not what we were trying to get at, but I think maybe it should be": learning how to do trauma-informed design with a data donation platform for online dating sexual violence. Presented at: 2024 CHI Conference on Human Factors in Computing System; May 11-16, 2024. [doi: 10.1145/3613904.3642045]
33. Josephs JC, Bungay V, Guta A, Gilbert M, Abdulai AF. Trauma-informed technology design in digital sexual health interventions. Stud Health Technol Inform 2024 Jul 24;315:773-774. [doi: 10.3233/SHTI240323] [Medline: 39049423]
34. Chen JX, McDonald A, Zou Y, et al. Trauma-informed computing: towards safer technology experiences for all. Presented at: 2022 CHI Conference on Human Factors in Computing System; Apr 29 to May 5, 2022. [doi: 10.1145/3491102.3517475]
35. Heale R, Noble H. Integration of a theoretical framework into your research study. Evid Based Nurs 2019 Apr;22(2):36-37. [doi: 10.1136/ebnurs-2019-103077] [Medline: 30894364]
36. Kothari A, Wathen CN. A critical second look at integrated knowledge translation. Health Policy 2013 Feb;109(2):187-191. [doi: 10.1016/j.healthpol.2012.11.004] [Medline: 23228520]
37. Crosschild C, Huynh N, De Sousa I, Bawafaa E, Brown H. Where is critical analysis of power and positionality in knowledge translation? Health Res Policy Syst 2021 Jun 11;19(1):92. [doi: 10.1186/s12961-021-00726-w] [Medline: 34116685]
38. What's a good engagement rate on Pinterest? Hudson Design Company. URL: https://www.hudsondesigncompany.com/whats-a-good-engagement-rate-on-pinterest-exploring-the-industry-standards [accessed 2025-07-15]
39. Polishchuk D. What is a good Instagram engagement rate? Promo Republic. 2022. URL: https://promorepublic.com/en/blog/what-is-a-good-instagram-engagement-rate [accessed 2025-07-15]
40. Rose-Collins F. Content types that perform best on Instagram. Ranktracker. 2025. URL: https://www.ranktracker.com/blog/content-types-that-perform-best-on-instagram [accessed 2025-12-12]
41. The digital gag: supression of sexual and reproductive health on Meta, TikTok, Amazon and Google. : Center for Intimacy Justice; 2025 URL: https://www.intimacyjustice.org/ [accessed 2026-01-14]

Abbreviations

IKT: integrated knowledge translation
PRAB: Patient Research Advisory Board
SAMHSA: Substance Abuse and Mental Health Services Administration

XSL•FO
RenderX

Edited by A Mavragani; submitted 04.Sep.2025; peer-reviewed by SS Yilmaz, T Harpel; accepted 23.Dec.2025; published 27.Jan.2026.

Please cite as:
Marshall K, Dhillon H, Howard AF, Noga H, Yang GJ, Zhu W, Sutherland J, Lett S, Leonova A, Yong PJ, Orr NL
Developing a Trauma-Informed Social Media Campaign to Disseminate Endometriosis-Specific Qualitative Art-Based Research Findings: Tutorial
J Med Internet Res 2026;28:e83491
URL: https://www.jmir.org/2026/1/e83491
doi:10.2196/83491

© Kerry Marshall, Hargun Dhillon, A Fuchsia Howard, Heather Noga, Grace J Yang, William Zhu, Jessica Sutherland, Sarah Lett, Anna Leonova, Paul J Yong, Natasha L Orr. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Tutorial

Team-Based Analysis of Large-Scale Qualitative Data: Tutorial Using a Nationwide SMS Text Messaging Poll of Youth

Melissa DeJonckheere[1,2], PhD; Samantha A Chuisano[1], MPH; Marika Waselewski[1], MPH; Kendrin Sonneville[3], SCD, RD; Tammy Chang[1,2], MD, MPH, MS

[1]Department of Family Medicine, University of Michigan, Ann Arbor, MI, United States

[2]Institute for Healthcare Policy and Innovation, University of Michigan, Ann Arbor, MI, United States

[3]School of Public Health, Department of Nutritional Sciences, University of Michigan, Ann Arbor, MI, United States

Corresponding Author:
Melissa DeJonckheere, PhD
Department of Family Medicine
University of Michigan
1018 Fuller Street
Ann Arbor, MI, 48105
United States
Phone: 1 7349361927
Email: mdejonck@umich.edu

Abstract

With the growing use of technology in qualitative data collection and analysis, there is an opportunity to gather rich and varied perspectives to improve health and well-being. However, large-scale qualitative datasets can be difficult to manage using traditional qualitative methods, and there are few examples of the application of large-scale qualitative analysis. In the context of digital health, large qualitative datasets are increasingly made up of short text segments, which need to be analyzed differently from lengthy transcripts from interviews or focus groups. Therefore, this tutorial describes the use of traditional qualitative methods to analyze a large corpus of qualitative text data. We use examples from a nationwide SMS text messaging poll of youth to highlight the opportunities to use this team-based analysis approach, which has been accessible and meaningful to youth researchers and novice qualitative researchers. These large-scale qualitative strategies may benefit novice researchers analyzing large volumes of qualitative data and short text segments, including SMS text messaging, social media posts, medical notes, and open-ended survey questions, among others.

International Registered Report Identifier (IRRID): RR2-10.2196/resprot.8502

(J Med Internet Res 2026;28:e72526) doi:10.2196/72526

KEYWORDS

qualitative; large-scale; survey methods; data analysis; adolescence

Introduction

Large-scale qualitative methods can gather the diverse perspectives of many people to augment our understanding of complex health and social issues. Traditional qualitative methods have focused on depth rather than breadth of analysis, resulting in smaller samples with rich detail. However, large-scale qualitative methods can emphasize both depth and breadth, responding to common critiques of qualitative research related to external validity (ie, generalizability in quantitative research or transferability in qualitative research). Large-scale qualitative approaches are appropriate for some studies and research questions, including those that use large primary or secondary text data sources (eg, social media posts, documents, and medical records) [1], multiple sites for data collection [2], longitudinal studies [3], or mixed-methods studies where understanding experiences across all participants is desirable (eg, clinical trials) [4-6]. Advances in technology have also increased access to large qualitative data sources through data mining techniques, such as social media sites, electronic health records [7,8], and patient portal messages [4]. As a result, mobile health and other digital health researchers have access to vast qualitative datasets of short text data.

Much of the existing literature on large-scale qualitative studies has focused on collaborative research teams that collect data across multiple sites, institutions, states, or nations. Traditional approaches for qualitative analysis leverage a line-by-line review

XSL•FO
RenderX

of each interview transcript or open-ended text response. However, the time and resources required to apply these approaches to a dataset with thousands of qualitative datapoints would be restrictive for nearly all research teams. Little practical guidance exists for researchers on how to approach data management and analysis of these large datasets [6,9]. As a result, some teams may intentionally narrow their analytic focus using purposive sampling techniques [3,5], including a subsample of responses into their analysis, or use rapid approaches to qualitative analysis [10,11]. Furthermore, many teams leverage software to assist with data management, coding, or summarization [12-14]. However, these approaches may introduce the need for additional time and resources for learning and acquiring new tools or new methodologies [13,15].

The growth of artificial intelligence (AI) has shaped some approaches to qualitative data analysis as some researchers seek methods to improve efficiency or increase objectivity [16]. For example, natural language processing has been used to support text analysis of large text datasets through the initial development of a coding structure or validation of themes [4,17,18]. Large language models, like ChatGPT (OpenAI), have been used to explore the use of qualitative analysis through summarization and theme development [19-21]. However, AI-supported qualitative analysis has thus far been limited in its ability to understand nuance and context in qualitative data and develop interpretive themes [18,21]. Furthermore, some qualitative researchers have rejected the use of AI in interpretive and reflexive qualitative approaches, which are intentionally human and subjective processes used to make meaning of human thoughts and experiences [22].

As human interpretation is still integral to qualitative research, practical guidance for researchers seeking to conduct analysis of increasingly large and complex studies in digital health is needed [6,9]. While analysis can be burdensome, human-based qualitative processes can be conducted efficiently on large datasets without sacrificing methodological rigor. Likewise, analysis of large qualitative datasets can be completed without the need for advanced methodologies or acquisition of new software [15]. Therefore, the purpose of this tutorial is to describe an example of team-based qualitative analysis of a large-scale SMS text message poll using traditional qualitative methods and freely available, easy to learn collaboration tools, including cloud-based spreadsheets. While spreadsheets and matrices have been used to organize and analyze transcripts from interviews and focus groups, [23,24] there is limited guidance on strategies for datasets with a large volume of short text responses common in digital health studies. We describe our experience conducting team-based qualitative analysis of open-ended survey responses from a nationwide SMS text messaging poll of youth across the United States [25]. Our approach has been used to efficiently and effectively engage both new and experienced researchers in qualitative research, including youth, students, and trainees.

Methods

Example Study

To illustrate our analysis approach, we use MyVoice, a nationwide poll of individuals aged 14 to 24 years across the United States that is administered via SMS text messaging [25]. The goal of MyVoice is to gather diverse youth perspectives across the United States to inform policies and programs that impact youth. Incorporating youth voices can create more relevant, effective, sustainable, and impactful interventions that ultimately improve the health of young people [26,27]. Increasingly, research has included the perspectives of adolescents, rather than adult proxies, while also considering the unique needs of this age group [28]. Participatory approaches to research have also successfully integrated adolescents into the research team to increase the relevance of research to youth populations while also promoting youth development and capacity [29-31]. However, engaging adolescents in research remains challenging when conducted in traditional settings and when using methods that are not accessible or approachable to young people.

MyVoice is designed to address challenges related to engaging youth in research through familiar and easy-to-use methods, like SMS text messaging to collect personal opinions and stories. Each MyVoice poll consists of 4-5 primarily open-ended questions designed to elicit youth perspectives on complex health and social issues. Participants respond to questions directly via SMS text messaging, rather than following a link to an external platform. The MyVoice research protocol was previously described after enrolling approximately 100 participants [25]. While the study rationale and overall methodology have not changed since 2016, there are currently approximately 800 MyVoice participants in each cohort to maximize understanding of variations in their lived experiences to better inform policy and practice. A brief description of the overall methods is presented below, highlighting considerations for large-scale data collection relevant to this tutorial. For additional details on recruitment, screening, and enrollment of MyVoice participants, see Multimedia Appendix 1.

Weekly Polls

SMS text messaging polls are administered using a cloud-based platform to send and receive SMS messages directly, which includes Textizen and the University of Michigan instance of REDCap (Research Electronic Data Capture; Vanderbilt University) with Twilio integration. REDCap is a secure, web-based software platform designed to support validated data capture, data management, and integration for research studies [32,33]. Questions are sent one at a time, and respondents must send an SMS text message to receive the next question. Participants are given 7 days to complete the poll before the next poll is automatically sent. Participants who have not responded after 2-3 days are sent a reminder to complete the poll. Each week, approximately 2750-3250 open-ended responses are collected. Since 2016, we have collected over 120 SMS text messaging polls and 45,000 unique responses.

XSL•FO
RenderX

Ethical Considerations

The MyVoice study was approved by the Institutional Review Board at the University of Michigan Medical School (HUM00119982). All participants provide online informed consent or assent (for those under the age of 18 years) prior to enrollment in the study. Parental consent is waived for minors. Participants receive US $1 for each poll that they answer (each poll consists of 4-5 questions) and US $5 for completion of a demographic survey upon enrollment. All study data are collected and stored in secure web-based platforms. Data are deidentified before shared with team members and any external collaborators.

Data Analysis of Large-Scale Qualitative Data

Analysis Overview

Below, we outline the efficient and reliable method we have used for team-based coding of large qualitative datasets. The overall approach includes five distinct steps: (1) data cleaning and management, (2) codebook development, (3) coding, (4) reconciliation, and (5) presentation of findings, which are described in detail below. Our approach leverages traditional qualitative approaches (ie, manual human coding) and enables collaboration among teams at various experience levels by using a simple spreadsheet – a freely accessible tool that can be used with minimal to no training. While spreadsheets have been used to structure coding of interview and focus group data, few studies [15,23,24,34] have described the use of spreadsheets for qualitative analysis of short open-ended text responses. Worked examples can be particularly beneficial in highlighting strategies that have been successfully implemented by other researchers to manage qualitative data and make study design decisions [15,35]. In the example provided, we highlight specific considerations for large-scale short text responses from the MyVoice study; however, this method could be easily applied to any study where data are available in a table or a .csv format (ie, social media data, unstructured medical record text, open-ended survey responses, etc).

Step 1: Data Cleaning and Management

SMS text messaging responses are downloaded from the data collection platform as a .csv file. Data are deidentified when downloaded and identified by a participant identification number. Cleaning includes merging responses with participant demographic data, rearranging lines of text that were separated, preliminary identification of junk or skipped responses, and filtering out nonresponses (ie, when a respondent does not answer a question).

The resulting deidentified spreadsheet is organized with the following columns: participant study ID, responses separated by question, and demographic survey responses, each in a separate column. Each row in the file includes the data for a unique participant.

For all MyVoice studies, analyses include line-by-line coding of deidentified participant responses in a spreadsheet. Teams primarily use a cloud-based platform (eg, Google Sheets) to ensure that all members of the research team (including high school students, undergraduate students, and other trainees) have access to the data throughout the analysis process and are able to work synchronously and asynchronously.

Step 2: Codebook Development

Codebooks are used to systematically record and organize qualitative data analysis by providing a list of key ideas in a dataset (ie, codes) and ensuring consistency and transparency in analysis [13]. MyVoice teams typically use a structured codebook to record the emerging ideas in their large-scale text data, including codes, definitions, and examples [13]. The codebook development process includes three overarching and iterative steps: (1) data familiarization through open coding and memoing, (2) identifying shared concepts across team members, and (3) defining and refining codes. Our codebooks are developed iteratively and collaboratively, with multiple team members offering interpretations of the important ideas that are present in the dataset and revising the list of codes throughout the analysis process [36].

Data Familiarization

Within each study team, at least 2 members independently read a portion of participant responses. Team members read approximately 100 responses per question, which in most datasets is approximately 10-20% of the responses. Multiple research team members independently use open coding to list the main concepts in the responses in an analytic memo, a document reflecting the researcher's thinking throughout the analysis process [13,37]. In our team, memos typically include a running list of concepts that are repeated in the data that will inform future codes. Some researchers include personal reflections, emerging patterns, and connections to the literature in their analytic memos [13].

Identifying Shared Concepts

Team members compare analytic memos to create a synthesized list of concepts that were identified in the responses. These concepts form the basis of the codes, although they may be renamed or modified in subsequent steps.

Defining and Refining Codes

Each concept is defined in a structured codebook that includes the code (ie, label for the concept), definition, and example response [13,38]. Some teams also include inclusion and exclusion criteria in their codebook, depending on the type of question and desired overlap between codes [38]. In studies with poll questions that are close-ended or when using a more deductive analysis approach, responses are typically categorized into discrete codes with clear inclusion and exclusion criteria. Teams may also structure their codebook with rules for co-coding responses, which can be helpful when analyzing responses that include both categorical and narrative responses.

Figure 1 displays an excerpt from a structured codebook that was used to analyze a dataset on youth perspectives on telemedicine before and after the start of the COVID-19 pandemic [39]. In total, 5 open-ended questions were asked in October 2019, and a new set of 5 questions was asked in October 2020. A separate coding scheme was created for all 10 questions.

XSL•FO
RenderX

Figure 1. Excerpt from structured codebook organized by poll question (Waselewski et al, 2022 [39]).

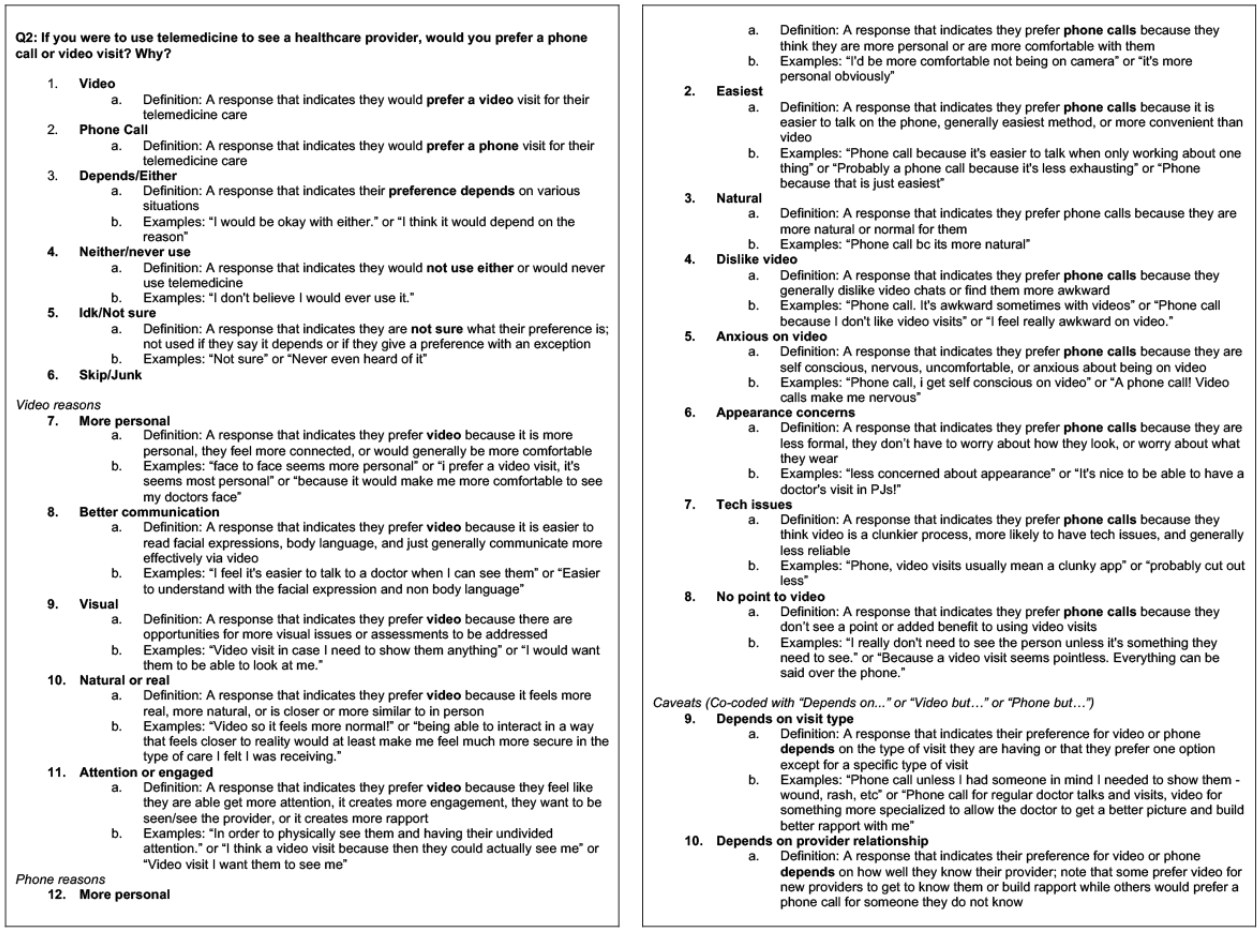

Q2: If you were to use telemedicine to see a healthcare provider, would you prefer a phone call or video visit? Why?

1. **Video**
   a. Definition: A response that indicates they would **prefer a video** visit for their telemedicine care
2. **Phone Call**
   a. Definition: A response that indicates they would **prefer a phone** visit for their telemedicine care
3. **Depends/Either**
   a. Definition: A response that indicates their **preference depends** on various situations
   b. Examples: "I would be okay with either." or "I think it would depend on the reason"
4. **Neither/never use**
   a. Definition: A response that indicates they would **not use either** or would never use telemedicine
   b. Examples: "I don't believe I would ever use it."
5. **Idk/Not sure**
   a. Definition: A response that indicates they are **not sure** what their preference is; not used if they say it depends or if they give a preference with an exception
   b. Examples: "Not sure" or "Never even heard of it"
6. **Skip/Junk**

Video reasons

7. **More personal**
   a. Definition: A response that indicates they prefer **video** because it is more personal, they feel more connected, or would generally be more comfortable
   b. Examples: "face to face seems more personal" or "i prefer a video visit, it's seems most personal" or "because it would make me more comfortable to see my doctors face"
8. **Better communication**
   a. Definition: A response that indicates they prefer **video** because it is easier to read facial expressions, body language, and just generally communicate more effectively via video
   b. Examples: "I feel it's easier to talk to a doctor when I can see them" or "Easier to understand with the facial expression and non body language"
9. **Visual**
   a. Definition: A response that indicates they prefer **video** because there are opportunities for more visual issues or assessments to be addressed
   b. Examples: "Video visit in case I need to show them anything" or "I would want them to be able to look at me."
10. **Natural or real**
    a. Definition: A response that indicates they prefer **video** because it feels more real, more natural, or is closer or more similar to in person
    b. Examples: "Video so it feels more normal!" or "being able to interact in a way that feels closer to reality would at least make me feel much more secure in the type of care I felt I was receiving."
11. **Attention or engaged**
    a. Definition: A response that indicates they prefer **video** because they feel like they are able get more attention, it creates more engagement, they want to be seen/see the provider, or it creates more rapport
    b. Examples: "In order to physically see them and having their undivided attention." or "I think a video visit because then they could actually see me" or "Video visit I want them to see me"

Phone reasons

12. **More personal**
    a. Definition: A response that indicates they prefer **phone calls** because they think they are more personal or are more comfortable with them
    b. Examples: "I'd be more comfortable not being on camera" or "it's more personal obviously"
13. **Easiest**
    a. Definition: A response that indicates they prefer **phone calls** because it is easier to talk on the phone, generally easiest method, or more convenient than video
    b. Examples: "Phone call because it's easier to talk when only working about one thing" or "Probably a phone call because it's less exhausting" or "Phone because that is just easiest"
14. **Natural**
    a. Definition: A response that indicates they prefer phone calls because they are more natural or normal for them
    b. Examples: "Phone call bc its more natural"
15. **Dislike video**
    a. Definition: A response that indicates they prefer **phone calls** because they generally dislike video chats or find them more awkward
    b. Examples: "Phone call. It's awkward sometimes with videos" or "Phone call because I don't like video visits" or "I feel really awkward on video."
16. **Anxious on video**
    a. Definition: A response that indicates they prefer **phone calls** because they are self conscious, nervous, uncomfortable, or anxious about being on video
    b. Examples: "Phone call, i get self conscious on video" or "A phone call! Video calls make me nervous"
17. **Appearance concerns**
    a. Definition: A response that indicates they prefer **phone calls** because they are less formal, they don't have to worry about how they look, or worry about what they wear
    b. Examples: "less concerned about appearance" or "It's nice to be able to have a doctor's visit in PJs!"
18. **Tech issues**
    a. Definition: A response that indicates they prefer **phone calls** because they think video is a clunkier process, more likely to have tech issues, and generally less reliable
    b. Examples: "Phone, video visits usually mean a clunky app" or "probably cut out less"
19. **No point to video**
    a. Definition: A response that indicates they prefer **phone calls** because they don't see a point or added benefit to using video visits
    b. Examples: "I really don't need to see the person unless it's something they need to see." or "Because a video visit seems pointless. Everything can be said over the phone."

Caveats (Co-coded with "Depends on..." or "Video but…" or "Phone but…")

20. **Depends on visit type**
    a. Definition: A response that indicates their preference for video or phone **depends** on the type of visit they are having or that they prefer one option except for a specific type of visit
    b. Examples: "Phone call unless I had someone in mind I needed to show them - wound, rash, etc" or "Phone call for regular doctor talks and visits, video for something more specialized to allow the doctor to get a better picture and build better rapport with me"
21. **Depends on provider relationship**
    a. Definition: A response that indicates their preference for video or phone **depends** on how well they know their provider; note that some prefer video for new providers to get to know them or build rapport while others would prefer a phone call for someone they do not know

For example, in the October 2020 poll, 1 question read, "If you were to use telemedicine to see a health care provider, would you prefer a phone call or video visit? Why?" The research team inductively developed codes for this question that included categorical codes that captured the first part of the question (ie, preference for phone or video visits) and descriptive codes for the subsequent explanation. Categorical codes included "Video," "Phone call," "Depends or either," "Neither or never use," or "I don't know or not sure" to understand participants' preferences for video or phone visits when using telemedicine. Here, there is a clear distinction in the categories, and responses cannot be categorized into multiple codes.

Each categorical code was also accompanied by a list of reasons that were derived from the text responses explaining their preference between phone and video visits (ie, descriptive codes). As seen in Figure 1, respondents who indicated a preference for telemedicine using video often indicated in their responses that video visits were more personal, improved communication, included visuals that could help with assessment, better approximated an in-person visit, and improved engagement with the provider. These reasons were then translated into descriptive codes that could be co-coded alongside "Video" (eg, "more personal," and "better communication"). Respondents who preferred phone visits described them as more personal, easy, and natural, among other factors. These reasons were translated into descriptive codes that could only be co-coded with "Phone call." While this team had rules for co-coding such that the concept of "more personal" had a code for preferring video visits and a separate code for preferring phone calls, an alternative strategy would be to have a single code of "more personal" that could be applied across categorical codes. Decisions are made by the research team to determine the scope of their codebook and direction for analysis. In this example, having separate codes allowed for quick calculations of frequency for each code in the content analysis.

Most MyVoice teams create a separate coding scheme for each question (as in Figure 1). A separate codebook for each question is helpful when conducting a content analysis, when the questions and, therefore, responses vary significantly, or to focus the analysis. It is also possible to create an overarching coding scheme that can be applied across questions in a single dataset. A single, overarching coding scheme would be appropriate when the poll questions are closely tied, and content from multiple responses can be captured with related codes. Alternatively, if analyzing all the data by individual respondent, rather than by question, a single coding scheme would also be optimal. This approach would closely align with most thematic analysis approaches that proceed with line-by-line coding using an individual participant or individual interview transcript, observation, or other textual data as the unit of analysis.

See, for example, an excerpt from a codebook developed by high school student co-researchers [40]. During the MYHealth research training program, high school students developed a

XSL•FO
RenderX

study to understand youth perspectives on the impact of social media on adolescents and young adults. Poll questions included the following: (1) How do you typically use social media (posting, watching videos, scrolling, messaging, commenting, etc.)? (2) What benefits do you get from social media? (3) How do you create a positive experience for yourself while using social media? (4) Who or what do you enjoy seeing or interacting with on social media? Why? (5) Tell us about a time when social media made you feel good about yourself. This team elected to analyze responses across individual participants, rather than by question, to track patterns across all 5 questions. See Figure 2 for an excerpt from the resulting codebook.

Figure 2. Example excerpt from MYHealth social media dataset with single codebook.

Code Name	Code Definition	Examples
SM Breaks	limits the amount of time they spend on social media, takes breaks from social media, or deletes social media apps	"I try to follow creators that are beneficial to my mental health and try to limit my screen time" "i recently decided to do a social media cleanse so as of right now i don't use a ton of social media..." "By being selective in what I allow on my screen"
Tailor content	Makes an effort to follow, view, post positive content or unfollow/block negative content.	"By blocking accounts that make me unhappy and only viewing specific things so that the algorithm suggests related posts" "I block toxic people" "I don't really look at celebrities or stuff that'll make me feel bad about myself, I don't spend a lot of time on social media either. And I don't post content for others to judge good or bad "
Maintain personal circle	This person tries to use SM to connect with, post content to, or share content with ***ONLY*** people they know personally. This code can also be used if the person says they avoid content from celebrities and influencers	"...I only let people I personally know or have known follow me." "I am mostly a lurker and I rarely post anything or interact beyond liking unless I personally know the poster" (in response to Q4) "My friends because it's personal" "With personal social media like Snapchat and instagram I only follow and let people follow me that are close"
Don't worry about other people's opinions	When the person says they try not to worry about what others think about them or avoid comments on their own posts	"i focus on myself, and don't let what others say, both negatively and positively, impact my self respect and self esteem" "Avoiding getting caught up being concerned about likes or comments" "I don't read the negative comments or most comments on anything I post...."
Reduce loneliness	Using social media to feel less isolated or alone or to feel accepted (that it is okay to be themselves)	'I have severe scoliosis and I found a community of other women who struggle with the same issue and it makes me **feel less isolated**" "laughter, sharing with friends, and realizing **I'm not alone** in thoughts/feelings/actions" "I enjoy seeing posts about mental health because **it makes me feel less alone and** about my favorite books/tv shows because it's fun to see content about things I enjoy" "It has helped me a lot to see other trans and queer people. It's hard to feel accepted in a conservative state, but being able to connect with others reminds me **that I am okay, and that I'm not alone**. "
Community	Uses social media to connect with people or connect with a community or people with similar backgrounds or interests	"I enjoy interacting with my friends as well as online users who are in the same community as I am such as shared hobbies" "I love sports content (especially related to my favorite teams and fans I've made friends with) and creative content (photography, design, etc.) because it helps create a sense of community" "Feeling a sense of community and staying informed" Feel apart of a community" ""I enjoy seeing people of color, women, and queer people. I share identities with and they are also widely underrepresented and promoted on social media"
Distraction	Using social media as a way to distract themselves or as a way to pass or fill time. This can be positive distraction or avoidance of negative things	"Scrolling on my phone because **I'm bored**" "It's a nice **distraction** and I use it to keep in contact with friends" "**Distraction** in case I need it." "**Turns my brain off**" "feeling plugged in, humor, a mind numbing way to **pass the time**"
Receives Attention	When the respondent says they have received positive attention or support from SM	"when people like/comment/share their posts/**receiving attention**" "**positive reinforcement**/positivity" "When I was included in an **appreciation post**"
Entertainment	This person uses social media as a way to find amusement, boost their mood, or get inspired. This can also include someone saying they use SM to find funny content or to laugh	"I find **entertainment** from watching videos about dogs, Scrolling through videos and watching inspiration from them give me a mood booster and is entertaining to me." "I enjoy interacting with niche influencers that align with my passions and morals. they **inspire** me to further shoot for my goals & also give me **inspo** for different aspects of my life" "... **inspiration** from other people" "I post reels on my story, text my friends , scroll through posts, send funny or **inspiring** or relatable things to my friends"

Step 3: Coding

In Google Sheets, Microsoft Excel, or another spreadsheet, each code from the finalized codebook is entered into a new, separate column. Within each research team, at least 2 members independently perform line-by-line coding of each participant's response. Cloud-based tools allow team members to collaboratively code data in real-time.

Team members use the codebook developed in the previous step to apply codes to each participant's response. Where a concept is described by the participant, a “1” is entered into the spreadsheet cell. All other cells remain blank (or could be entered as “0”). In most cases, multiple codes may be applied to a participant’s response to reflect the many ideas that are present. Many responses include multiple sentences, phrases, or lists of ideas that are unique concepts.

For example, in response to the question about preference between video or phone visits for telemedicine, one respondent shared:

> *Video visit, for sure. Because the non-verbal cues that come along with being able to see someone are extremely important. And with my primary, I wouldn’t feel satisfied with a phone call because being able to interact in a way that feels closer to reality would at least make me feel more secure in the type of care I was receiving.*

This response included several important concepts, which should each be coded to accurately reflect the data. The research team

XSL•FO
RenderX

identified the following codes: (1) video, (2) better communication, (3) natural or real, and (4) works better (represented in the highlighted row in Figure 3).

Figure 3. Example coding sheet and completed coding.

Study ID	If you were to use telemedicine to see a healthcare provider, would you prefer a phone call or video visit? Why?	Video	Call	Depends /either	More personal	Better com	Visual	Natural /real	Attention /engaged	Dislike phone	Anxious on phone	Tech issues	Works better	More personal	Easiest	Natural	Dislike video
95	Phone call for regular doctor talks and visits, video for something more specialized to allow the doctor to get a better picture and build better rapport with me			1			1		1								
98	Video call because it feels more personal to me	1			1												
108	Video visit. It seems more personal to me	1			1												
111	Video visit, for sure. Because the non-verbal cues that come along with being able to see someone are extremely important. And with my primary, I wouldn't feel satisfied with a phone call because being able to interact in a way that feels closer to reality would at least make me feel much more secure in the type of care I felt I was receiving.	1				1		1					1				
112	No because I want them to see my issues in person																
131	the first time i'd want a video call to see who i'm talking to, but after that i'd be fine to do it over a phone call			1													
141	Video to get the nonverbals	1				1											
145	Video visit so I can also show what I am talking about	1					1										

Step 4: Reconciliation

Depending on the analysis strategy, teams often review and reconcile discrepancies in coding to reach agreement on the meaning of each response and final assignment of codes. Teams will use a spreadsheet formula to quickly identify discrepancies in applied codes across the large dataset. Disagreements are labeled in a new column so they can easily be filtered and reviewed. To resolve disagreements and reach consensus, the 2 independent coders compare their applied codes, discuss the rationale for the codes, and address the discrepancy by either agreeing on the final application of codes or asking a third member of the research team to review. With online tools, reconciliation can take place in real-time to support collaboration and integration of multiple perspectives.

Step 5: Development of Findings (eg, Themes or Descriptions)

Content Analysis

After the coding process is complete, MyVoice teams often summarize codes with frequencies to determine the prevalence of codes and categories (ie, groups of related codes) across participants. Teams will often present code frequencies (or the proportion of responses that are aligned with a code) in their findings. Presentations of frequencies are often grouped by question or overarching main ideas across the dataset, supported by sample responses. For example, Schmid et al [41] examined the impact of COVID-19 on weight and shape control behaviors among youth. The authors conducted a content analysis that involved developing initial themes, constructing a structured codebook with codes and definitions, independently applying codes to responses for each survey question, comparing discrepancies between each pair of coders, and finalizing themes and subthemes that summarized the main ideas in responses to each question. Content analysis of the prompt, “Since the pandemic, have you changed your eating or exercise habits because you were worried about your weight or shape. Tell us about it,” resulted in 2 themes: changes for weight or shape control and changes in overall attitude or concerns. Overall, 4 subthemes that detail the responses common in each theme were presented alongside the frequency of responses. By capturing the frequencies in responses, the authors concluded that one-third of the sample implemented weight or shape control behaviors during the COVID-19 pandemic.

Thematic Analysis

Theme development occurs by looking for relationships across codes, both within and across questions in the dataset. MyVoice teams often develop themes through iterative team discussion where codes are grouped into related categories or concepts, patterns are identified across categories or concepts, and variation in participant experiences is described. For example, a MyVoice study examined youth perspectives on use and access to primary care [42]. Questions included:

1. Where do you go for a health problem you can’t solve on your own (family doctor, ER, urgent care, school clinic, etc)? Why do you go there?
2. Do you have a regular doctor you go to for check-ups or when you feel sick? If you don’t have a regular doctor, why not?
3. Do you think it’s important to have a regular doctor you go to for check-ups or when you feel sick? Why or why not?
4. How could your doctor’s office make your visits easier?
5. Have you had a bad experience with a doctor or clinic visit? What happened?

Using a thematic analysis approach, the research team followed the analysis steps outlined above, including data familiarization, codebook development, line-by-line coding, code reconciliation,

XSL•FO
RenderX

and consolidation of conceptually related categories. To develop themes, the team held many discussions where 6 faculty and program staff reflected on patterns in the data. They considered prevalent categories (eg, convenience) in addition to those that were salient but less prevalent (eg, experiences with physical and emotional harm in health care settings). One prominent theme was "Feeling respected, valued, and heard was essential to youth having positive experiences in a primary care health care setting." In the description of this theme and in the selection of representative quotes in the final manuscript, the authors presented the range of experiences (ie, both positive and negative previous experiences with primary care) among MyVoice participants to fully elaborate on the breadth of responses [42].

Quantitative Analysis

Occasionally, teams will pursue a quantitative or mixed methods approach that incorporates statistical analyses, in addition to the methods described above [43,44]. Quantitative analysis of large-scale qualitative data may be tempting because of the volume of data and the need to efficiently summarize responses. In addition, the use of spreadsheets for coding facilitates frequency counts and other descriptive methods that are less common in many approaches to qualitative research [35,38].

Several considerations should be made when performing quantitative analysis of qualitative data. First, the goal of qualitative research is typically not to generalize to a population. Decisions in the study design, including sampling, data collection, and data analysis, will likely impose limitations for a quantitative analysis. SMS text messaging responses in MyVoice are limited in that follow-up and clarifying questions cannot be asked. Therefore, analysis is restricted to the words participants choose to share. When questions are framed as very broad and open-ended, participants may only share some of their perspectives. Researchers cannot be sure that participants would not have endorsed a theme, just that they did not share it in their written responses, which creates potential missing data. As a result, quantitative analyses should align with the scope of the research question and poll question.

For example, Schuiteman et al [44] investigated self-management of health care among youth in a mixed methods study. The four MyVoice poll questions were:

1. Who schedules your doctor's appointments? Why?
2. Who usually goes with you to your doctor's appointments? Why?
3. Who usually picks up your prescription medicine? Why?
4. Do you wish you were more or less involved in your health care? How?

The primary outcome was whether respondents desired to be more, less, or equally involved in their health care (ie, responses to the fourth question), and "independence scores" were calculated to assess variation in this outcome by summing participants' levels of independence in the first 3 questions (ie, through identification of self or help from others). Differences in the primary outcome were compared by demographic characteristics using chi-square tests. In this example, questions were worded in a way that supported coding of responses into categorical variables. Qualitative analyses explored the "why" and "how" narrative responses provided across questions, and resulting themes were compared between participants with high independence and low independence scores.

Second, careful consideration should also be given to what differences by demographics represent, particularly demographic characteristics that are socially constructed [45]. Analyses should have sufficient justification in the literature or existing theory, particularly in cases where testing for differences by demographic characteristics is desired. For example, one could expect, based on the literature, that nonbinary and transgender youth have more negative experiences with health care providers than cisgender youth [46]. In the MyVoice study of youth perspectives on primary care use and access, the research team conducted exploratory quantitative analyses to compare endorsement of each theme by different demographic groups using chi-square (for gender, race, and income level) and *t* tests (1-tailed; for age). In the study sample, more than one-third (284/789) reported previous bad experiences in a healthcare setting. Among these participants, 47% of nonbinary or transgender participants and 44% of cisgender women reported negative experiences with health care providers compared to 24% of men, which was statistically significant ($P<.001$) [42].

Discussion

Although technology has enabled the data collection and analysis of larger volumes of text data, there has been limited practical guidance on how researchers can manage large databases and rigorously analyze large datasets using qualitative methods [3,5]. In this tutorial, we presented steps and tools we have used for team-based analysis of large-scale text data collected through a nationwide SMS text messaging poll of youth. These steps have been used by teams that include youth and trainees of all levels, supporting the development of research interest, experiences, and skills through meaningful and authentic participation in research. Novice researchers can apply the methods described here using familiar tools, like spreadsheets, to reduce the time and resources spent using other qualitative data software tools. In addition to open-ended survey responses, the data analysis methods described can be applied by other digital and mobile health researchers conducting qualitative studies with substantial samples of short text data.

This tutorial describes the application of a qualitative analysis approach to short text responses, like those in SMS text messaging, open-ended survey questions, social media posts, and more. SMS text messaging polls allow youth to share their perspectives using an easy, familiar, and accessible tool, which may support youth to feel more comfortable engaging in research [47,48]. However, responses can vary in length, detail, and quality. Unlike some other qualitative data collection methods, like semistructured interviews, MyVoice SMS text message polls are limited to a predetermined set of questions. Teams are unable to ask follow-up questions or probe for more detail, which likely limits some of the richness of the data and therefore understanding of participant perspectives and experiences [49]. Indeed, Tran et al [50] found that significantly more responses were needed in open-ended survey responses

XSL•FO
RenderX

to reach the point of data sufficiency than is commonly recommended in qualitative analysis of interviews or focus groups. As a result, studies using short text responses can require larger datasets that can apply the team-based approach detailed in this tutorial. Further research is needed to illustrate analysis methods for large-scale qualitative research with longer text responses, such as interviews or observations.

Our initial study protocol proposed the use of natural language processing to support analysis of SMS text message responses. We used natural language processing to analyze several datasets and to try to streamline research processes for novice researchers [4,18]. However, time and resource limitations (eg, sufficient expertise among team members and updating algorithms for the context of each unique dataset) and concerns about the utility and validity of outputs restricted our regular use of this approach to date. Advancements in AI, including large language models and tools within qualitative data analysis software, will likely improve the efficiency and validity of AI-assisted qualitative analysis over time.

Our team includes youth as active team members in the research process to enhance the validity and relevance of the research [29,51,52]. We have found that line-by-line coding and pattern identification are processes and tools that are familiar to youth and similar to strategies taught in classrooms, enabling youth researchers to quickly engage in meaningful research opportunities. Through the MYHealth research training program, over 150 high school students and recent high school graduates have successfully used the methods in this tutorial to understand youth perspectives on future planning, social media, academic expectations, health education, and health care access [25,53]. In addition, novice researchers and researchers new to qualitative methods have successfully used our team-based approach to collaboratively and efficiently analyze large text datasets on a variety of health and health policy topics.

Despite the recent proliferation of AI tools, manual analysis procedures are still preferred in many qualitative and mixed methods studies to ensure understanding of the nuance and context available through qualitative data and limiting the bias introduced by AI that can influence findings [16,18]. This tutorial has provided an example of using a team-based qualitative analysis approach to analyze a large corpus of data collected through SMS text messaging polls of youth with commonly available tools. This tutorial adds to the limited research describing the step-by-step application of qualitative analysis techniques to large-scale qualitative data, particularly short text responses. Our approach can be used by research teams seeking to analyze short response text data using qualitative methods.

Funding

Establishment of the MyVoice study was supported by the Michigan Institute for Clinical and Health Research, the University of Michigan MCubed program, and the University of Michigan Department of Family Medicine and continues to receive funding through contracts with collaborating teams. The University of Michigan Research Electronic Data Capture (REDCap) instance used to collect data throughout MyVoice is maintained by the Michigan Institute for Clinical & Health Research. Michigan Institute for Clinical & Health Research is supported by the National Center for Advancing Translational Sciences at the National Institutes of Health (grant number UM1TR004404). A portion of MD's time is supported by a career development award funded through the National Institute of Diabetes and Digestive and Kidney Diseases and the National Institutes of Health (grant number K01DK134766). The MYHealth program is funded by a Science Education Partnership Award through the National Institute of General Medical Sciences and the National Institutes of Health (grant number R25GM137361). The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

Data sharing is not applicable to this article as no data sets were generated or analyzed during this study.

Authors' Contributions

Conceptualization: MD (lead), SAC (supporting), MW (supporting), KS (supporting), TC (supporting)
Writing – original draft: MD (lead), SAC (supporting), MW (supporting)
Writing – review & editing: MD (lead), SAC (supporting), MW (supporting), KS (supporting), TC (supporting)

Conflicts of Interest

MD is an Associate Editor for *JMIR mHealth and uHealth*.

Multimedia Appendix 1
Recruitment, Screening and Enrollment Process.
[DOCX File , 22 KB - jmir_v28i1e72526_app1.docx]

References

XSL•FO
RenderX

1. Davidson E, Edwards R, Jamieson L, Weller S. Big data, qualitative style: a breadth-and-depth method for working with large amounts of secondary qualitative data. Qual Quant 2019;53(1):363-376 [FREE Full text] [doi: 10.1007/s11135-018-0757-y] [Medline: 30930496]
2. Nallamothu BK, Guetterman TC, Harrod M, Kellenberg JE, Lehrich JL, Kronick SL, et al. How do resuscitation teams at top-performing hospitals for in-hospital cardiac arrest succeed? A qualitative study. Circulation 2018;138(2):154-163 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.118.033674] [Medline: 29986959]
3. Winskell K, Singleton R, Sabben G. Enabling analysis of big, thick, long, and wide data: Data management for the analysis of a large longitudinal and cross-national narrative data set. Qual Health Res 2018;28(10):1629-1639 [FREE Full text] [doi: 10.1177/1049732318759658] [Medline: 29557295]
4. Chang T, DeJonckheere M, Vydiswaran VGV, Li J, Buis LR, Guetterman TC. Accelerating mixed methods research with natural language processing of big text data. Journal of Mixed Methods Research 2021;15(3):398-412. [doi: 10.1177/15586898211021196]
5. Robins CS, Eisen K. Strategies for the effective use of NVivo in a large-scale study: qualitative analysis and the repeal of don't ask, don't tell. Qualitative Inquiry 2017;23(10):768-778. [doi: 10.1177/1077800417731089]
6. Yang Y, Pankow J, Swan H, Willett J, Mitchell SG, Rudes DS, et al. Preparing for analysis: a practical guide for a critical step for procedural rigor in large-scale multisite qualitative research studies. Qual Quant 2018;52(2):815-828 [FREE Full text] [doi: 10.1007/s11135-017-0490-y] [Medline: 37904883]
7. Vydiswaran VGV, Romero DM, Zhao X, Yu D, Gomez-Lopez I, Lu JX, et al. Uncovering the relationship between food-related discussion on twitter and neighborhood characteristics. J Am Med Inform Assoc 2020;27(2):254-264 [FREE Full text] [doi: 10.1093/jamia/ocz181] [Medline: 31633756]
8. Liu T, Waselewski EA, Dougherty A, Lee JD, Hill NE, Marzolf BA, et al. Use of and medical decision-making in portal messages among patients with type 2 diabetes: mixed methods study. J Med Internet Res 2025;27:e79413 [FREE Full text] [doi: 10.2196/79413] [Medline: 40997326]
9. White DE, Oelke ND, Friesen S. Management of a large qualitative data set: establishing trustworthiness of the data. Int J Qual Methods 2012;11(3):244-258. [doi: 10.1177/160940691201100305]
10. St George SM, Harkness AR, Rodriguez-Diaz CE, Weinstein ER, Pavia V, Hamilton AB. Applying rapid qualitative analysis for health equity: lessons learned using "EARS" with latino communities. Int J Qual Methods 2023;22:10.1177/16094069231164938 [FREE Full text] [doi: 10.1177/16094069231164938] [Medline: 38463016]
11. Watkins DC. Rapid and rigorous qualitative data analysis: the "RADaR" technique for applied research. Int J Qual Methods 2017;16(1):160940691771213. [doi: 10.1177/1609406917712131]
12. Parkinson S, Eatough V, Holmes J, Stapley E, Midgley N. Framework analysis: a worked example of a study exploring young people's experiences of depression. Qual Res Psychol 2015;13(2):109-129. [doi: 10.1080/14780887.2015.1119228]
13. Saldana J. The Coding Manual for Qualitative Researchers. Thousand Oaks, CA: SAGE Publications; 2021.
14. Woods M, Paulus T, Atkins DP, Macklin R. Advancing qualitative research using qualitative data analysis software (QDAS)? Reviewing potential versus practice in published studies using ATLAS.ti and NVivo, 1994–2013. Soc Sci Comput Rev 2016;34(5):597-617. [doi: 10.1177/0894439315596311]
15. Ose SO. Using excel and word to structure qualitative data. J Appl Soc Sci 2016;10(2):147-162. [doi: 10.1177/1936724416664948]
16. Prescott MR, Yeager S, Ham L, Rivera Saldana CD, Serrano V, Narez J, et al. Comparing the efficacy and efficiency of human and generative AI: qualitative thematic analyses. JMIR AI 2024;3:e54482 [FREE Full text] [doi: 10.2196/54482] [Medline: 39094113]
17. Abram MD, Mancini KT, Parker RD. Methods to integrate natural language processing into qualitative research. Int J Qual Methods 2020;19:1609406920984608. [doi: 10.1177/1609406920984608]
18. Guetterman TC, Chang T, DeJonckheere M, Basu T, Scruggs E, Vydiswaran VGV. Augmenting qualitative text analysis with natural language processing: methodological study. J Med Internet Res 2018;20(6):e231 [FREE Full text] [doi: 10.2196/jmir.9702] [Medline: 29959110]
19. Hamilton L, Elliott D, Quick A, Smith S, Choplin V. Exploring the use of AI in qualitative analysis: a comparative study of guaranteed income data. Int J Qual Methods 2023;22:16094069231201504. [doi: 10.1177/16094069231201504]
20. Hitch D. Artificial intelligence augmented qualitative analysis: the way of the future? Qual Health Res 2024;34(7):595-606 [FREE Full text] [doi: 10.1177/10497323231217392] [Medline: 38064244]
21. Morgan DL. Exploring the use of artificial intelligence for qualitative data analysis: the case of ChatGPT. Int J Qual Methods 2023;22:16094069231211248. [doi: 10.1177/16094069231211248]
22. Jowsey T, Braun V, Clarke V, Lupton D, Fine M. We reject the use of generative artificial intelligence for reflexive qualitative research. Qualitative Inquiry 2025. [doi: 10.1177/10778004251401851]
23. Meyer DZ, Avery LM. Excel as a qualitative data analysis tool. Field Methods 2008;21(1):91-112. [doi: 10.1177/1525822x08323985]
24. Ndungo I, Balimuttajjo S, Akugizibwe E. Excel-based qualitative data analysis: methodological guidelines for systematic organization, iterative coding, and thematic visualization. Int J Educ Technol Sci 2025;5(3):271-288. [doi: 10.5281/zenodo.17007908]

XSL•FO
RenderX

25. DeJonckheere M, Nichols LP, Moniz MH, Sonneville KR, Vydiswaran VGV, Zhao X, et al. MyVoice national text message survey of youth aged 14 to 24 years: study protocol. JMIR Res Protoc 2017;6(12):e247 [FREE Full text] [doi: 10.2196/resprot.8502] [Medline: 29229587]
26. Aceves-Martins M, Aleman-Diaz AY, Giralt M, Solà R. Involving young people in health promotion, research and policy-making: practical recommendations. Int J Qual Health Care 2019;31(2):147-153. [doi: 10.1093/intqhc/mzy113] [Medline: 29788085]
27. Patton GC, Sawyer SM, Santelli JS, Ross DA, Afifi R, Allen NB, et al. Our future: a Lancet commission on adolescent health and wellbeing. Lancet 2016;387(10036):2423-2478 [FREE Full text] [doi: 10.1016/S0140-6736(16)00579-1] [Medline: 27174304]
28. Schelbe L, Chanmugam A, Moses T, Saltzburg S, Williams LR, Letendre J. Youth participation in qualitative research: challenges and possibilities. Qualitative Social Work 2014;14(4):504-521. [doi: 10.1177/1473325014556792]
29. Anyon Y, Bender K, Kennedy H, Dechants J. A systematic review of youth participatory action research (YPAR) in the United States: methodologies, youth outcomes, and future directions. Health Educ Behav 2018;45(6):865-878. [doi: 10.1177/1090198118769357] [Medline: 29749267]
30. Ballonoff Suleiman A, Ballard PJ, Hoyt LT, Ozer EJ. Applying a developmental lens to youth-led participatory action research: a critical examination and integration of existing evidence. Youth & Society 2021;53(1):26-53. [doi: 10.1177/0044118x19837871]
31. Kennedy H, DeChants J, Bender K, Anyon Y. More than data collectors: a systematic review of the environmental outcomes of youth inquiry approaches in the United States. Am J Community Psychol 2019;63(1-2):208-226. [doi: 10.1002/ajcp.12321] [Medline: 30843254]
32. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
33. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
34. Bree RT, Gallagher G. Using Microsoft Excel to code and thematically analyse qualitative data: a simple, cost-effective approach. AISHE-J: The All Ireland Journal of Teaching and Learning in Higher Education 2016;8. [doi: 10.62707/aishej.v8i2.281]
35. Rosen RK, Gainey M, Nasrin S, Garbern SC, Lantini R, Elshabassi N, et al. Use of framework matrix and thematic coding methods in qualitative analysis for mHealth: NIRUDAK study data. Int J Qual Methods 2023;22:10.1177/16094069231184123 [FREE Full text] [doi: 10.1177/16094069231184123] [Medline: 38817641]
36. Reyes V, Bogumil E, Welch LE. The living codebook: documenting the process of qualitative data analysis. Sociol Methods Res 2021;53(1):89-120. [doi: 10.1177/0049124120986185]
37. Charmaz K. Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis. Thousand Oaks, CA: SAGE Publications; 2006.
38. MacQueen KM, McLellan E, Kay K, Milstein B. Codebook development for team-based qualitative analysis. CAM Journal 1998;10(2):31-36. [doi: 10.1177/1525822X980100020301]
39. Waselewski ME, Waselewski EA, Wasvary M, Wood G, Pratt K, Chang T, et al. Perspectives on telemedicine from a national study of youth in the United States. Telemed J E Health 2022;28(4):575-582. [doi: 10.1089/tmj.2021.0153] [Medline: 34415810]
40. Chuisano SA, Rafferty J, Allen A, Chang T, Diemer M, Harris K, et al. Increasing representation and diversity in health research: a protocol of the MYHealth research training program for high school students. PLoS One 2023;18(9):e0281790 [FREE Full text] [doi: 10.1371/journal.pone.0281790] [Medline: 37768968]
41. Schmid JC, Rose KL, Hadler NL, Amaro X, Frank A, Wilkie E, et al. Content analysis of the impact of COVID-19 on weight and shape control behaviors and social media content of U.S. adolescents and young adults. Eat Behav 2022;45:101635 [FREE Full text] [doi: 10.1016/j.eatbeh.2022.101635] [Medline: 35567879]
42. DeJonckheere M, Zhao A, Lane J, Plegue MA, Ranalli L, Wagner E, et al. Results of a national text message poll of youth: perspectives on primary care. J Adolesc Health 2020;67(6):814-820. [doi: 10.1016/j.jadohealth.2020.04.017] [Medline: 32620346]
43. DeJonckheere M, Waselewski M, Amaro X, Frank A, Chua KP. Views on COVID-19 and use of face coverings among U.S. youth. J Adolesc Health 2021;68(5):873-881 [FREE Full text] [doi: 10.1016/j.jadohealth.2021.02.015] [Medline: 33896552]
44. Schuiteman S, Chua KP, Plegue MA, Ilyas O, Chang T. Self-management of health care among youth: implications for policies on transitions of care. J Adolesc Health 2020;66(5):616-622 [FREE Full text] [doi: 10.1016/j.jadohealth.2020.01.009] [Medline: 32113903]
45. Lett E, Asabor E, Beltrán S, Cannon AM, Arah OA. Conceptualizing, contextualizing, and operationalizing race in quantitative health sciences research. Ann Fam Med 2022;20(2):157-163 [FREE Full text] [doi: 10.1370/afm.2792] [Medline: 35045967]

XSL•FO
RenderX

46. Goldenberg T, Gamarel KE, Reisner SL, Jadwin-Cakmak L, Harper GW. Gender affirmation as a source of resilience for addressing stigmatizing healthcare experiences of transgender youth of color. Ann Behav Med 2021;55(12):1168-1183 [FREE Full text] [doi: 10.1093/abm/kaab011] [Medline: 33761531]
47. French RS, McCarthy O, Baraitser P, Wellings K, Bailey JV, Free C. Young people's views and experiences of a mobile phone texting intervention to promote safer sex behavior. JMIR Mhealth Uhealth 2016;4(2):e26 [FREE Full text] [doi: 10.2196/mhealth.4302] [Medline: 27083784]
48. Perry RCW, Kayekjian KC, Braun RA, Cantu M, Sheoran B, Chung PJ. Adolescents' perspectives on the use of a text messaging service for preventive sexual health promotion. J Adolesc Health 2012;51(3):220-225. [doi: 10.1016/j.jadohealth.2011.11.012] [Medline: 22921131]
49. Robinson OC. Probing in qualitative research interviews: theory and practice. Qual Res Psychol 2023;20(3):382-397. [doi: 10.1080/14780887.2023.2238625]
50. Tran VT, Porcher R, Falissard B, Ravaud P. Point of data saturation was assessed using resampling methods in a survey with open-ended questions. J Clin Epidemiol 2016;80:88-96. [doi: 10.1016/j.jclinepi.2016.07.014] [Medline: 27492788]
51. Balazs CL, Morello-Frosch R. The three R's: How community based participatory research strengthens the rigor, relevance and reach of science. Environ Justice 2013;6(1):10.1089/env.2012.0017 [FREE Full text] [doi: 10.1089/env.2012.0017] [Medline: 24260590]
52. Malorni A, Lea CH, Richards-Schuster K, Spencer MS. Child Youth Serv Rev 2022;136:106399 [FREE Full text] [doi: 10.1016/j.childyouth.2022.106399]
53. Durden A, Lucien J, Chuisano SA, Palazzolo B, Ahmed Z, Ceaser N, et al. Youth perspectives of school-based health education: findings from a US youth-led nationwide qualitative survey. Health Educ J 2025:10.1177/00178969251376859. [doi: 10.1177/00178969251376859] [Medline: 41393786]

Abbreviations

AI: artificial intelligence
REDCap: Research Electronic Data Capture

Edited by T de Azevedo Cardoso, A Stone; submitted 11.Feb.2025; peer-reviewed by D Levine, J Cornelius; comments to author 01.Aug.2025; revised version received 04.Feb.2026; accepted 10.Feb.2026; published 27.Feb.2026.

Please cite as:
DeJonckheere M, Chuisano SA, Waselewski M, Sonneville K, Chang T
Team-Based Analysis of Large-Scale Qualitative Data: Tutorial Using a Nationwide SMS Text Messaging Poll of Youth
J Med Internet Res 2026;28:e72526
URL: https://www.jmir.org/2026/1/e72526
doi:10.2196/72526
PMID:

©Melissa DeJonckheere, Samantha A Chuisano, Marika Waselewski, Kendrin Sonneville, Tammy Chang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

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

Huanzhuo Mai[1,2,3*], MPH; Jianyuan Liu[1,2,3*], BSc; Liping Cheng[1,2,3*], BSc; Hao Liang[1,2,4], PhD; Ping Cui[2,4], PhD; Qin Yao[5], MSc; Qin Cao[1,2,3], BSc; Heng Zhang[1,2,3], BSc; Wenping Liao[1,2,3], BSc; Shiran He[1,2,3], BSc; Xing Yang[6*], PhD; Ping Cen[7*], PhD; Jiegang Huang[1,2,3*], PhD

[1]School of Public Health, Guangxi Medical University, 22, Shuangyong Road, Nanning, Guangxi, China

[2]Guangxi Key Laboratory of AIDS Prevention and Treatment, Guangxi Medical University, Nanning, Guangxi, China

[3]Key Laboratory of Prevention and Control of Highly Prevalent Diseases (Guangxi Medical University), Education Department of Guangxi Zhuang Autonomous Region, Nanning, Guangxi, China

[4]Life Sciences Institute, Guangxi Medical University, Nanning, Guangxi, China

[5]Nursing College, Guangxi Medical University, Nanning, Guangxi, China

[6]Health Management Research Institute, People's Hospital of Guangxi Zhuang Autonomous Region and Guangxi Academy of Medical Sciences, Nanning, Guangxi, China

[7]Infectious Disease Laboratory, Guangxi AIDS Clinical Treatment Centre (Nanning), The Fourth People's Hospital of Nanning, Nanning, Guangxi, China

[*]these authors contributed equally

Corresponding Author:
Jiegang Huang, PhD
School of Public Health, Guangxi Medical University, 22, Shuangyong Road, Nanning, Guangxi, China

Abstract

Background: Social media intervention may enhance HIV prevention among men who have sex with men, but the effect of this intervention in resource-limited settings remains unclear.

Objective: This randomized controlled trial evaluated whether a social media intervention grounded in the information-motivation-behavioral skills (IMB) model could be beneficial for HIV prevention among men who have sex with men in resource-limited settings.

Methods: Participants were recruited in Nanning, China, between April 2023 and April 2024. Eligible participants were randomly assigned to either the social media intervention group or the routine HIV prevention services control group. Participants in the intervention group received a 3-month social media intervention, which included completing video-based tasks. Baseline surveys were conducted, followed by follow-up surveys every 3 months, for a total of 2 follow-ups. Outcomes included HIV testing uptake, high-risk behavior, AIDS-related knowledge, safe sex self-efficacy, and attitude.

Results: A total of 180 eligible men who have sex with men were enrolled (90 per group). Follow-up rates were 97.8% (88/90) and 95.5% (86/90) for the intervention and control groups, respectively. At the follow-ups, the intervention group demonstrated significantly higher uptake of HIV testing, a lower proportion of participants reporting high-risk sexual behaviors, and higher condom use self-efficacy compared to the control group (all P<.05). After controlling for sociodemographic variables, generalized estimating equations analysis revealed that the intervention group had significantly higher odds of HIV testing (risk ratio [RR] 1.739, 95% CI 1.110 - 2.730), HIV self-testing (RR 2.306, 95% CI 1.593 - 3.340), and consistent condom use (RR 2.457, 95% CI 1.636 - 3.690) than the control group. Cochran-Armitage trend tests within the intervention group revealed that with increasing intervention duration, both HIV testing and HIV self-testing significantly increased, while high-risk sexual behaviors significantly decreased (all P<.05).

Conclusions: The social media intervention guided by the IMB model demonstrated a positive effect on expanding HIV testing coverage, reducing high-risk behavior, enhancing AIDS-related knowledge, and improving safer sex self-efficacy among men who have sex with men in resource-limited settings. These findings provide valuable guidance for future HIV prevention and control efforts targeting this population.

XSL•FO
RenderX

Trial Registration: Chinese Clinical Trial Registry ChiCTR2500095176; https://www.chictr.org.cn/hvshowproject.html?id=267136

(J Med Internet Res 2026;28:e84279) doi:10.2196/84279

KEYWORDS

men who have sex with men; social media intervention; information-motivation-behavioral skills model; HIV testing; high-risk behaviors

Introduction

Men who have sex with men represent a high-risk group for HIV infection. As of 2024, men who have sex with men account for 7.6% of global HIV infections among individuals aged 15‐49 years [1]. In China, the HIV infection rate among men who have sex with men is approximately 5.7%, with rates as high as 10.7% in the southwestern region, and infection rates continue to rise [2,3]. A significant proportion of men who have sex with men remain unaware of their HIV status. By 2020, only 62.2% of men who have sex with men in China had ever undergone HIV testing [4,5], which hinders progress toward achieving the first “95” target of “95-95-95” set by the Joint United Nations Programme on HIV/AIDS (UNAIDS)—aiming for 95% of people living with HIV to know their status [6]. Men who have sex with men often engage in behaviors associated with high HIV risk, including unprotected sex, having multiple sexual partners, and recreational drug use [7]. For example, in one study conducted in China, about 37% of men who have sex with men reported engaging in unprotected anal intercourse in the past 6 months [8]. These behaviors increase the likelihood of sexual transmission and exacerbate the epidemic. Reducing such high-risk sexual behaviors can effectively lower the risk of HIV transmission among the men who have sex with men population [9]. Furthermore, they often face social discrimination and stigma [10], creating significant barriers and difficulties to seeking health services and information. These structural and social factors facilitate the spread of HIV within the population. Therefore, implementing effective strategies to improve HIV testing uptake and reduce high-risk behaviors among men who have sex with men is an urgent priority for current HIV prevention and control efforts.

In recent years, social media–based interventions have garnered prominence and have been increasingly adopted for HIV prevention among men who have sex with men [11-13]. Social media refers to various internet-based communication channels, such as social networking sites, blogs, online social media networks, and other forms of communication technology [14]—essentially, digital technology-based platforms using networks for information dissemination. With the rapid advancement of the internet and social media, mobile devices have become ubiquitous personal items. This gives social media interventions the advantage of offering fragmented learning opportunities, addressing limitations of traditional offline interventions [15]. To date, existing studies of social media–based HIV prevention interventions for men who have sex with men in China have largely focused on economically advanced metropolitan regions, including Beijing [16], Anhui [17], and Guangdong [18], which are characterized by relatively high GDP (gross domestic product), well-developed infrastructure, and strong regional influence. In contrast, there is a scarcity of research exploring the effectiveness of such interventions in resource-limited settings within China. Although Nanning serves as the capital of Guangxi, it remains relatively underresourced compared with China’s more developed cities, facing structural and resource constraints in economic development, infrastructure, and research capacity amid ongoing rapid urbanization [19,20]. Therefore, investigating the impact of social media interventions on improving HIV testing uptake and reducing high-risk behaviors among men who have sex with men in these underserved areas is crucial for enhancing local HIV prevention and control efforts.

To maximize the effectiveness of social media interventions, integrating them with established behavioral intervention theories is essential. Compared to other primary behavioral intervention theories [21,22], the information-motivation-behavioral skills (IMB) model, proposed by Fisher, comprehensively considers the influence of psychosocial factors on behavior. It categorizes factors influencing behavior change into 3 components: information, motivation, and behavioral skills. The model posits and validates that possessing disease-prevention information and motivation to engage in prevention can trigger preventive behaviors through behavioral skills to initiate and sustain action [23]. Currently, interventions based on the IMB model are widely applied in HIV behavioral interventions [24,25]. Furthermore, with the rapid development of new media technologies, integrating the IMB model into social media interventions has emerged as a growing trend in public health. Therefore, exploring the feasibility of combining social media interventions with the IMB model holds significant importance for developing more effective HIV prevention and control strategies.

Previous studies in China have demonstrated that social media–based interventions can improve HIV testing uptake among men who have sex with men, but evaluations have primarily focused on economically developed metropolitan areas [16-18]. Although this approach has been shown to increase HIV testing, evidence regarding its effectiveness in reducing high-risk behaviors remains limited. Moreover, in prior studies, intervention effects have mostly been limited to increasing HIV testing willingness or behavior, while effects on promoting broader key behavioral changes such as condom use and pre-exposure prophylaxis (PrEP) adoption remain unclear; most studies only conducted online interventions and data collection, making it difficult to ensure the quality of online information; interventions often lacked grounding in established behavioral change theories; furthermore, outcomes were often measured only at the end of the intervention period without

XSL•FO
RenderX

follow-up assessments to determine whether the effects were sustained over time [16-18]. Guangxi, a resource-limited province in southern China, has one of the highest HIV burdens nationwide, particularly among men who have sex with men [26,27]. To address these gaps, we conducted a 2-arm, parallel-group randomized controlled trial (RCT) in Guangxi, with additional emphasis on outcomes related to promoting broader key behavioral changes such as condom use and PrEP adoption, incorporated offline data collection methods to ensure quality, and the application of the IMB model to guide the intervention process, in order to evaluate whether a social media–based intervention could both increase HIV testing and reduce high-risk behaviors. Furthermore, a continued follow-up assessment was incorporated after the intervention to evaluate the durability of the intervention effects over time. This study aimed to provide context-specific evidence to evaluate a social media–based intervention's efficacy and sustainability in increasing HIV testing and reducing high-risk behaviors among men who have sex with men, thereby guiding more targeted and scalable HIV prevention strategies for men who have sex with men, supporting optimal allocation of limited health resources in comparable settings, and contributing to the achievement of the 95-95-95 targets.

Methods

Trial Design and Participants

This study used a 2-armed, parallel-group RCT design. Recruitment of participants among men who have sex with men was conducted using a combination of convenience sampling and peer-driven sampling methods through Guangxi Caihong, a gay-led community-based organization (CBO) based in Nanning, China. On-site peer educators posted online recruitment notices and registration questionnaires on social platforms such as WeChat Moments, WeChat Official Accounts, WeChat/QQ groups, and Blued. Individuals who completed online registration underwent an initial screening by staff, and the results were notified via SMS text message. Those who passed the initial screening were required to arrive at the CBO site to proceed with the next steps of the survey. Participants who were selected and completed the on-site survey received a cash compensation of RMB 20 (approximately US $3), regardless of whether they were eventually enrolled in the study. Additionally, men who have sex with men who had registered, participated, or were subsequently enrolled were encouraged to refer friends or partners who met the inclusion criteria to join the study.

The entire social media intervention procedure was conducted online via WeChat, the predominant social media platform in China [28]. Other research procedures, including informed consent, baseline assessment, and follow-up, were carried out offline. The 2-armed study design was (1) social media intervention group and (2) routine HIV prevention (control) services group. All participants completed assessments at 3 time points: baseline (preintervention), 3 months postbaseline (postintervention), and 6 months postbaseline (3 months after postintervention; Figure 1).

Figure 1. Trial design diagram. The study used a 2-armed, parallel-group randomized controlled trial design. Participants were recruited from among men who have sex with men through the Guangxi Caihong organization in Nanning, China. After recruitment and baseline survey, a 3-month intervention period commenced, with follow-up assessments conducted at the 3rd and 6th months (focusing on the 6 assessment categories shown under "follow-up endpoint" in the figure). Both the intervention and control groups participated in HIV prevention activities over 3 months, while the intervention group additionally received a social media–based intervention. PEP: post-exposure prophylaxis; PrEP: pre-exposure prophylaxis.

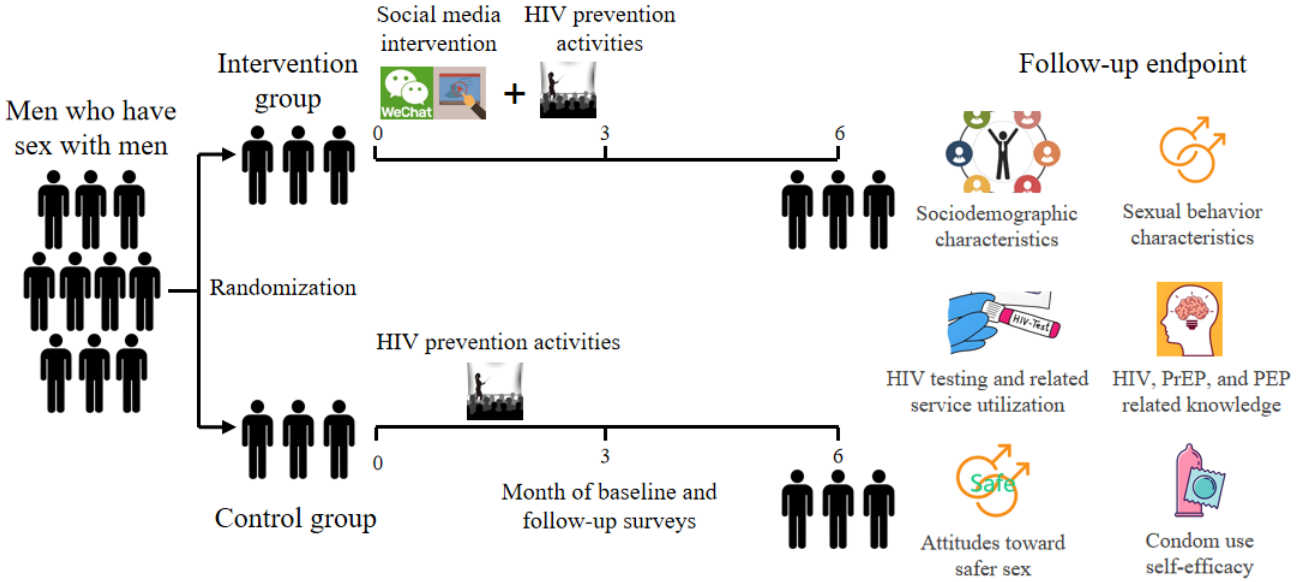

Participants were eligible if they met the following criteria: (1) born biologically male, (2) aged 18‐45 years, (3) current Nanning residents planning to reside here for most of the next year, (4) self-reported male-to-male sex in the past 6 months, (5) proficient in using WeChat and willing to connect with the study account, (6) tested HIV-negative within the past 3 months (with test results provided by participants or verified on-site by CBO), and (7) demonstrated sufficient comprehension to complete study procedures. Individuals who had significant psychological, psychiatric, and intellectual disabilities preventing survey participation, or who were currently participating in other related intervention studies, were excluded.

XSL•FO
RenderX

We followed the standard guidelines for reporting parallel group RCTs of CONSORT (Consolidated Standards of Reporting Trials) 2010 [29]. Further details regarding study design and intervention development are provided in Multimedia Appendix 1.

Sample Size

The sample size was calculated for this on-site RCT based on previous similar studies. The calculation formula is as follows:

$$N=Z\alpha 2\bar{P}(1-\bar{P})+Z\beta P1(1-P1)+P2(1-P2)(P1-P2)2$$

The study has 2 primary outcomes: HIV testing rate and the incidence of high-risk sexual behaviors. Previous studies have shown that digital media interventions significantly increase HIV testing rates or HIV self-testing rates (by 48.7% - 63%) [16,17]. Meanwhile, prior research has reported a relatively small expected difference in the incidence of high-risk sexual behaviors (17% - 35%) [9,30]. According to the RCT sample size formula, calculating based on the incidence of high-risk sexual behaviors yields a larger required sample size. To ensure sufficient study power and reflect the most conservative estimate, this study's sample size was determined based on the expected difference in the incidence of high-risk sexual behaviors between the two men who have sex with men groups.

Previous studies had reported that 25% - 42% of men who have sex with men engaged in high-risk sexual behaviors in the past 6 months [8,30]. Based on this evidence, we assumed a baseline prevalence (ρ) of 35%. Prior interventions had demonstrated reductions of 17% - 35% in such behaviors [9,31]. To ensure a conservative estimate, we assumed a 15% reduction, setting the expected prevalence in the intervention group (P_1) at 20% and in the control group (P_2) at 35%. With a significance level (α) of .05 (Z_α=1.96) and power (1 – β) of 80% (Z_β=0.84), the required sample size per group was calculated as n=78. Accounting for an estimated 10% loss to follow-up, we planned to recruit 90 participants per group, yielding a total of 180 men who have sex with men.

Randomization and Blinding

To ensure the privacy of study participants, each was assigned a unique identification number for all subsequent follow-ups. For randomization, the allocation sequence was generated by an independent researcher, who was not involved in participant recruitment or any subsequent intervention procedures, using the "randomizr" package in R software (R Foundation for Statistical Computing). Participants were then randomly allocated in a 1:1 ratio to either the intervention or control group based on their assigned identification numbers. Participants were blinded to their group assignment and instructed to maintain confidentiality regarding their intervention status. Throughout the recruitment, baseline assessment, and all follow-up stages, the research staff responsible for these tasks remained unaware of the group allocations, thereby ensuring allocation concealment. While research staff administering the interventions were necessarily unblinded, the statisticians conducting data analysis remained blinded to group allocations throughout the study to ensure objective outcome assessment.

Intervention Exposure

All participants added the study team's WeChat work account for information notifications. The control arm received routine HIV prevention services provided by Guangxi Caihong. These services, conducted offline, included HIV/AIDS education and awareness campaigns, regular HIV testing services, and structured peer education sessions, held at least once a month.

Participants in the intervention group received the same services as the control group, with the addition of a social media–based intervention. This intervention was delivered via "Rainbow Home," a WeChat mini-program developed by the research team. The research team assisted participants in the intervention group with guided registration, provided a detailed explanation of the mini-program's features, and conducted weekly check-in tasks (Multimedia Appendix 2). To minimize intervention contamination between groups, researchers provided both verbal and written instructions to all participants in the intervention group before the intervention began, requesting that they refrain from sharing the mini-program or any task-related content with others until the end of the study period. This request did not impose any coercion or restriction on the participants. Furthermore, the WeChat mini-program used in this study was configured with restricted access and anonymous log-in. Access privileges were granted exclusively to the intervention group participants by the researchers via dynamic invitation codes, ensuring that unauthorized individuals could not use the mini-program. Participants in the intervention group were unable to view any information about other members within their group. Additionally, service schedules of the intervention group differed from the control group, while all other elements—including activity content and study personnel—remained consistent across both groups.

The intervention based on the IMB model was structured around 4 core components: information, motivation, behavioral skills, and behavior. Content mainly contained comprehensive HIV/AIDS knowledge, real-world case studies highlighting HIV infection risks and prevention narratives, practical guidance on safer sexual practices and correct use of prevention tools, and provision of free HIV self-testing kits. See Table 1 for detailed intervention content.

XSL•FO
RenderX

Table . The detailed intervention content based on the information-motivation-behavioral skills (IMB) model[a].

Component	Content
Information	Delivered knowledge via animated videos on HIV/AIDS, sexually transmitted infections, correct HIV testing and self-testing, proper use of PrEP[b] and PEP[c], correct condom use, recreational drug side effects, and HIV case studies with treatment.
Motivation	Raised awareness of safer sex and self-protection through sharing HIV infection case studies and provided one-on-one psychological counseling to participants in need.
Behavioral skills	Weekly video-based check-in tasks were sent via mini-program, with staff available for consultation and tailored advice. Free HIV self-test kits with usage guidance were provided, and partner testing was encouraged.
Behavior	Change attitudes toward safer sex, improve self-efficacy in condom use, promote intentions to engage in safer sexual practices, increase awareness of HIV testing, and reduce high-risk sexual behaviors.

[a]Guided by the IMB model, the intervention content was structured around the 4 core components of the model: information, motivation, behavioral skills, and behavior.

[b]PrEP: pre-exposure prophylaxis.

[c]PEP: post-exposure prophylaxis.

The intervention was delivered primarily through video viewing; for example, procedural knowledge (such as PrEP usage guidelines and HIV self-testing procedures) was provided through animated videos. Research staff released 1‐2 video-based tasks per week (a total of 12 tasks) via the mini-program backend. The main content, video duration of each task, and its alignment with the 4 components based on the IMB model are provided in Multimedia Appendix 3. Tasks were pushed every Sunday at 7:30 PM and remained accessible for 2 weeks to allow flexibility in completion. Automated reminder notifications were sent through WeChat, and participants completing all tasks within the 3-month intervention period received financial incentives. In parallel, trained staff provided real-time consultation through WeChat, addressing participant inquiries and encouraging feedback on platform usability. At the 3- and 6-month follow-ups, intervention participants were encouraged to conduct HIV testing on-site, request free HIV self-testing kits by mail or on-site pickup, and invite their sexual partners to undergo testing.

During the follow-up process, for participants suspected of being HIV-positive, the research team members provided them with corresponding psychological support and medical referral services, assisting them in visiting the local Centers for Disease Control or designated hospitals for further confirmatory testing and subsequent follow-up management. Once a participant was confirmed positive, it was considered an endpoint event, and further follow-up was discontinued from the date of diagnosis.

The intervention in this study was conducted in accordance with the research protocol. After the intervention materials were developed, they were reviewed by 3 public health experts and peer educators before being finalized and disseminated. All research staff received standardized training prior to the intervention, and they documented and reported the implementation of interventions in real time. Dedicated staff monitored each participant's intervention engagement and task completion rate via the system backend, and backend monitoring data were exported and evaluated every 2 weeks.

After the conclusion of this study, participants in the control group will be granted the same access to the WeChat mini-program and all intervention content as the intervention group. However, the program still requires further optimization and is not currently intended for public release.

Outcomes, Data Collection, and Management

The primary outcome was the uptake of HIV testing and HIV self-testing (verified through testing at the CBO and photo-submitted self-test results to the staff at the 3-month and 6-month follow-ups after enrollment, respectively), along with self-reported high-risk behaviors among men who have sex with men in the past 3 months (defined as engaging in anal intercourse with a male partner without consistent condom use throughout the act or experiencing condom breakage during intercourse). Secondary outcomes included the uptake of syphilis testing (verified through testing at the CBO), as well as knowledge of AIDS, PrEP, and post-exposure prophylaxis (PEP), self-efficacy for condom use, and safe sex attitudes. Validated scales were used to measure knowledge, self-efficacy, and safe sex attitudes. Definitions of outcomes and methods of outcome assessment were detailed in Multimedia Appendix 4.

All self-reported outcomes were evaluated via follow-up surveys administered at enrollment and at 3 and 6 months postenrollment, conducted offline at the CBO site through one-on-one interviews by research staff for data collection.

All original questionnaires and backend-exported data were independently double-entered by two researchers after quality control checks. Personal information was removed, and the data were stored on an encrypted hard drive accessible only to the principal investigator of the research team. The data used during the analysis phase were anonymized. Both the original paper questionnaires and backup electronic questionnaire files are stored in the research team's encrypted hard drive.

XSL•FO
RenderX

Statistical Analysis

Descriptive statistics were used to summarize the baseline characteristics. Continuous variables are presented as mean (SD) and were compared using *t* tests or ANOVA. Categorical variables are presented as frequency (percentage) and were compared using chi-square tests or Fisher exact test. The Cochran-Armitage trend test was used for repeated categorical measures within the intervention group. To investigate the effect of the intervention, generalized estimating equations (GEE) were used to analyze correlations for repeated measures data (eg, HIV testing and condom use) across baseline and follow-up time points and between intervention and control groups. Bonferroni correction was applied for post hoc pairwise comparisons. All tests were 2-tailed, with a significance level defined as α=.05; *P* values <.05 were considered statistically significant. All data analyses were completed using R Studio (version 4.2.1; Posit Software) or SPSS (version 26.0; IBM Inc).

Ethical Considerations

This RCT was retrospectively registered at the Chinese Clinical Trial Registry (ChiCTR2500095176). Ethical approval was obtained from the Medical Ethics Committee of Nanning Center for Disease Control and Prevention (2023014) and the Medical Ethics Committee of Guangxi Medical University (2024-KY0148).

Following ethics committee approval, the study was submitted to the Chinese Clinical Trial Registry. However, due to the platform's review process and timeline, participant recruitment began under the ethics-approved protocol before the official trial registration number was issued, resulting in a retrospective registration. Throughout the study, all procedures adhered strictly to the original ethics-approved protocol, with no changes to the study content or primary outcomes before or after registration.

Participation in this study was voluntary, and written informed consent was provided by all participants. All analytical data were processed anonymously to ensure participant confidentiality. All research staff committed to not disclosing any participant information. At baseline and each follow-up, participants received a cash compensation of RMB 20 (approximately US $3) upon completing the survey. Additionally, those who completed all tasks during the intervention period were provided an extra cash compensation of RMB 20 (approximately US $3).

Results

Participant Flow and Intervention Adherence

Recruitment took place between April 2023 and August 2023, during which 392 men who have sex with men were initially screened. After applying the inclusion and exclusion criteria, 180 eligible participants were enrolled and randomly allocated to either the social media intervention group (n=90) or the routine HIV prevention control group (n=90). The formal intervention began in September 2023, and all follow-ups were completed by April 2024. All participants were retained at the 3-month follow-up assessment. At the 6-month follow-up, loss to follow-up occurred in 2 participants (2/90, 2.2%) from the intervention group and 4 participants (4/90, 4.5%) from the control group, yielding follow-up rates of 97.8% (88/90) and 95.5% (86/90), and an overall follow-up rate of 96.7% (174/180). The participant recruitment, intervention, and follow-up flowchart is detailed in Figure 2. During the 3-month intervention period, 12 video check-in tasks were delivered to the intervention group. The average duration per video task was 112 seconds. Participants completed a total of 861 check-ins, representing a task completion rate of 79.7% (861/1080). Due to the low proportion of missing data, no data imputation was performed in this study. The analysis of this study was based on complete case analysis, including only participants with complete outcome data (ie, those who completed all follow-ups).

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram. Inclusion criteria: (1) born biologically male, (2) aged 18-45 years, (3) current Nanning residents planning to reside here for most of the next year, (4) self-reported male-to-male sex in the past 6 months, (5) proficient in using WeChat and willing to connect with the study account, (6) tested HIV-negative within the past 3 months, and (7) demonstrated sufficient comprehension to complete study procedures.

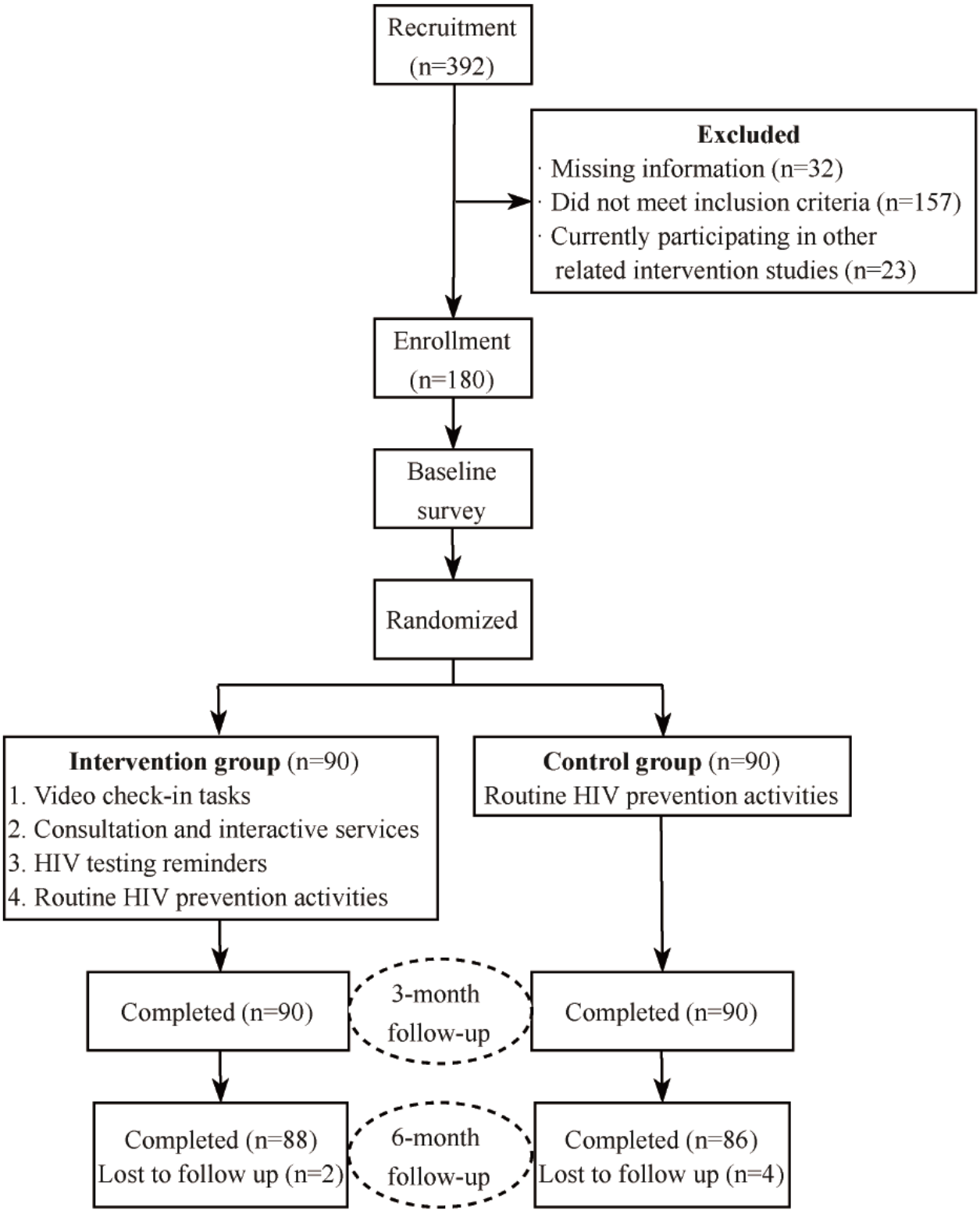

Baseline Characteristics of Participants

The mean age of participants was 28.5 (SD 6.9) years. The majority (143/180, 79.4%) had an educational level of college/university or above. Most participants were employed in government institutions or as office workers (91/180, 50.6%). Regarding sexual behavior characteristics, 68.3% (123/180) identified as homosexual in orientation. More than half (104/180, 57.8%) reported engaging in high-risk sexual behaviors. Of the 180 participants, 106 (58.9%) reported having

XSL•FO
RenderX

performed HIV self-testing in the past 3 months. After randomization, there were no statistically significant differences in demographic or behavior characteristics between the two groups, indicating comparability (Table 2).

XSL•FO
RenderX

Table . Sociodemographic and behavioral characteristics of study participants.

Characteristics	Total (n=180), n (%)	Control group (n=90), n (%)	Intervention group (n=90), n (%)	Chi-square (*df*)	*P* value
Age (year)				0.7 (2)	.70
≤24	51 (28.3)	23 (25.6)	28 (31.1)		
25-34	87 (48.3)	45 (50.0)	42 (46.7)		
≥35	42 (23.3)	22 (24.4)	20 (22.2)		
Household registration				0.4 (2)	.82
Nanning City	84 (46.7)	40 (44.4)	44 (48.9)		
Other cities within the province	80 (44.4)	42 (46.7)	38 (42.2)		
Other cities outside the province	16 (8.9)	8 (8.9)	8 (8.9)		
Ethnicity				2.1 (2)[a]	.34
Han	100 (55.6)	47 (52.2)	53 (58.9)		
Zhuang	75 (41.7)	39 (43.3)	36 (40.0)		
Other	5 (2.8)	4 (4.4)	1 (1.1)		
Education level				2.3 (2)[a]	.34
Junior high school or below	9 (5.0)	6 (6.7)	3 (3.3)		
Vocational school/high school	28 (15.6)	11 (12.2)	17 (18.9)		
College/bachelor's degree or higher	143 (79.4)	73 (81.1)	70 (77.8)		
Occupation				4.1 (4)[a]	.38
Student	35 (19.4)	15 (16.7)	20 (22.2)		
Public institution/staff	91 (50.6)	48 (53.4)	43 (47.8)		
Farmer/worker	11 (6.1)	3 (3.3)	8 (8.9)		
Self-employed	24 (13.3)	12 (13.3)	12 (13.3)		
Unemployed	19 (10.6)	12 (13.3)	7 (7.8)		
Monthly income (CNY)				0.7 (2)	.70
<3000 (US $435)	63 (35.0)	32 (35.6)	31 (34.4)		
3000-5000 (US $435-$725)	67 (37.2)	31 (34.4)	36 (40.0)		
>5000 (US $725)	50 (27.8)	27 (30.0)	23 (25.6)		
Marital status				1.2 (2)	.55
Unmarried	160 (88.8)	78 (86.7)	82 (91.1)		
Married	10 (5.6)	7 (7.8)	3 (3.3)		
Divorced	10 (5.6)	5 (5.5)	5 (5.6)		
Living arrangement				0.6 (2)	.75
Living alone	95 (52.8)	47 (52.2)	48 (53.3)		
Living with family/partner	42 (23.3)	23 (25.6)	19 (21.1)		
Dormitory/shared rental	43 (23.9)	20 (22.2)	23 (25.6)		

XSL•FO
RenderX

Characteristics	Total (n=180), n (%)	Control group (n=90), n (%)	Intervention group (n=90), n (%)	Chi-square (*df*)	*P* value
Sexual orientation				1.1 (2)[a]	.59
Homosexual	123 (68.3)	60 (66.7)	63 (70.0)		
Bisexual	49 (27.2)	27 (30.0)	22 (24.4)		
Uncertain	8 (4.4)	3 (3.3)	5 (5.6)		
Sexual role				0.1 (2)	.96
Insertive role	48 (26.7)	23 (25.6)	25 (27.8)		
Receptive role	65 (36.1)	33 (36.7)	32 (35.6)		
Versatile	67 (37.2)	34 (37.8)	33 (36.7)		
Age at first anal sex (years)				0.8 (1)	.45
≤18	74 (41.1)	34 (37.8)	40 (44.4)		
>18	106 (58.9)	56 (62.2)	50 (55.6)		
Number of regular partners				0.5 (2)	.85
0	50 (27.8)	27 (30.0)	23 (25.6)		
1-2	120 (66.7)	58 (64.4)	62 (68.9)		
≥3	10 (5.6)	5 (5.6)	5 (5.6)		
Number of casual partners				2.0 (2)	.38
0	77 (42.8)	34 (37.8)	43 (47.8)		
1-2	46 (25.6)	24 (26.7)	22 (24.4)		
≥3	57 (31.7)	32 (35.6)	25 (27.8)		
Awareness of partner's HIV status				2.3 (1)	.13
No	78 (43.3)	44 (48.9)	34 (37.8)		
Yes	102 (56.7)	46 (51.1)	56 (62.2)		
Sexting				1.1 (1)	.39
No	46 (25.6)	26 (28.9)	20 (22.2)		
Yes	134 (74.4)	64 (71.1)	70 (77.8)		
Popper/alkyl nitrite inhalant use				0.3 (1)	.62
No	131 (72.8)	64 (71.1)	67 (74.4)		
Yes	49 (27.2)	26 (28.9)	23 (25.5)		
Consistent condom use				3.3 (1)	.10
Yes	76 (42.2)	32 (35.6)	44 (48.9)		
No	104 (57.8)	58 (64.4)	46 (51.1)		
HIV self-testing				0.4 (1)	.65
No	74 (41.1)	35 (38.9)	39 (43.3)		
Yes	106 (58.9)	55 (61.1)	51 (56.7)		
Syphilis test result				1.6 (1)	.31
Negative	163 (90.6)	84 (93.3)	79 (87.8)		
Positive	17 (9.4)	6 (6.7)	11 (12.2)		
HIV testing				1.5 (1)	.15
No	43 (23.9)	18 (20.0)	25 (27.8)		

XSL•FO
RenderX

Characteristics	Total (n=180), n (%)	Control group (n=90), n (%)	Intervention group (n=90), n (%)	Chi-square (*df*)	*P* value
Yes	137 (76.1)	72 (80.0)	65 (72.2)		

[a]Fisher exact test.

Social Media Interventions Drive Increases Testing and Reductions in High-Risk Behavior Among Men Who Have Sex With Men

At the 3-month follow-up, the HIV self-testing rate was significantly higher in the intervention group than in the control group (*P*<.001), and the proportion of participants reporting high-risk sexual behaviors was significantly lower (*P*<.05). The number of casual partners and the rate of popper/alkyl nitrite inhalant use in the intervention group were significantly lower than in the control group (*P*<.05). The intervention group had a significantly higher rate of awareness of their sexual partners' HIV status compared to the control group (*P*<.05; Table 3).

XSL•FO
RenderX

Table . Comparison of sexual behavior characteristics and HIV testing status between the intervention group and the control group among study participants after 6-month follow-up.

Variables and groups	Baseline (intervention: n=90; control: n=90), n (%)	3-month follow-up (intervention: n=90; control: n=90), n (%)	6-month follow-up (intervention: n=88; control: n=86), n (%)
HIV testing			
Intervention	65 (72.2)	74 (82.2)	81 (92.0)[a,b]
Control	72 (80.0)	66 (73.3)	56 (65.1)
HIV self-testing			
Intervention	51 (56.7)	63 (70.0)[a]	63 (71.6)[a,c]
Control	55 (61.1)	31 (34.4)	32 (37.2)
Syphilis test result (positive)			
Intervention	11 (12.2)	3 (3.3)	4 (4.5)
Control	6 (6.7)	13 (14.4)	7 (8.1)
Number of regular partners[d]			
0			
Intervention	23 (25.6)	24 (26.7)	25 (28.4)
Control	27 (30.0)	33 (36.7)	21 (26.7)
1 - 2			
Intervention	62 (68.9)	69 (76.7)	60 (68.2)
Control	58 (64.4)	52 (57.8)	61 (76.7)
≥3			
Intervention	5 (5.6)	5 (5.6)	3 (3.4)
Control	5 (5.6)	5 (5.6)	4 (5.5)
Number of casual partners[e]			
0			
Intervention	43 (47.8)	54 (60.0)	38 (43.2)
Control	34 (37.8)	46 (51.1)	47 (60.0)
1 - 2			
Intervention	22 (24.4)	31 (34.4)	32 (36.4)
Control	24 (26.7)	30 (33.3)	31 (34.4)
≥3			
Intervention	25 (27.8)	11 (12.2)	13 (14.8)
Control	32 (35.6)	14 (15.5)	8 (12.2)
Awareness of partner's HIV status[f]			
Intervention	56 (62.2)	69 (76.7)[d]	62 (74.7)[a]
Control	46 (51.1)	50 (62.5)	40 (49.4)
High-risk sexual behavior[f]			
Intervention	46 (51.1)	33 (36.7)[d]	26 (31.3)[g]
Control	58 (64.4)	45 (55.5)	36 (44.4)
Sexting			
Intervention	70 (77.8)	63 (70.0)	56 (63.6)[c]
Control	64 (71.1)	57 (63.3)	58 (67.4)
Popper/alkyl nitrite inhalant use			

XSL•FO
RenderX

Variables and groups	Baseline (intervention: n=90; control: n=90), n (%)	3-month follow-up (intervention: n=90; control: n=90), n (%)	6-month follow-up (intervention: n=88; control: n=86), n (%)
Intervention	23 (25.5)	16 (17.8)[d]	12 (13.6)[c,e]
Control	26 (28.9)	31 (34.4)	26 (30.2)

[a]*P*<.001 compared to the control group.

[b]*P*<.001 compared to the baseline.

[c]*P*<.05 compared to the baseline.

[d]*P*<.05 compared to the control group.

[e]*P*<.01 compared to the control group.

[f]There are missing values as some study participants had no sexual partners in the last 3 months and were therefore excluded from the analysis.

[g]*P*<.01 compared to the baseline.

At the 6-month follow-up, both HIV testing rate and HIV self-testing rate were significantly higher in the intervention group compared to the control group (*P*<.001). The proportions of participants reporting high-risk sexual behaviors and popper/alkyl nitrite inhalant use were lower in the intervention group than in the control group (*P*<.05). Additionally, participants in the intervention group were significantly more likely to be aware of their sexual partners' HIV status than those in the control group (*P*<.001). Regarding sexually transmitted infections, 16 cases of syphilis were detected at the 3-month follow-up (3 in the intervention group and 13 in the control group), and 11 cases were identified at the 6-month follow-up (4 in the intervention group and 7 in the control group). No participants tested positive for HIV at either follow-up assessment (Table 3).

Cochran-Armitage trend tests within the intervention group revealed significant linear trends over time for all measured variables except number of regular sexual partners, number of casual sexual partners, and awareness of partner HIV status (*P*<.05). These variables included HIV testing, HIV self-testing, high-risk sexual behaviors, sexting behavior, and popper/alkyl nitrite inhalant use. Specifically, the positive effects observed in the intervention group at month 3 (the end of the intervention)—such as the promotion of HIV testing and HIV self-testing, reduction of high-risk sexual behaviors, and reduction of Rush popper use—were maintained and even strengthened at the 6-month follow-up assessment (3 months after the intervention ended; Table 3).

After controlling for sociodemographic variables, repeated measures ANOVA (with time as the within-subject factor and group as the between-subject factor) revealed: The intervention group had significantly higher odds of HIV testing (risk ratio [RR] 1.739, 95% CI 1.110 - 2.730), HIV self-testing (RR 2.306, 95% CI 1.593 - 3.340), and consistent condom use (RR 2.457, 95% CI 1.636 - 3.690) compared to the control group. The intervention group had significantly lower odds of popper/alkyl nitrite inhalant use compared to the control group (RR 0.554, 95% CI 0.365 - 0.841). Compared to the control group, the intervention group showed significantly increased odds of HIV self-testing at the 3-month follow-up (RR 5.651, 95% CI 2.321 - 13.761); also, significantly increased odds of both HIV testing (RR 10.473, 95% CI 3.276 - 33.491) and HIV self-testing (RR 5.651, 95% CI 2.321 - 13.761) were observed at the 6-month follow-up (Table 4).

XSL•FO
RenderX

Table . Analysis of the impact of new media intervention on high-risk sexual behaviors and HIV testing before and after intervention[a].

Variable	β value	RR[b] (95% CI)	*P* value[c]
HIV testing (no=0, yes=1)			
Group (intervention vs control)	0.553	1.739 (1.110 - 2.730)	*.02*
Time1 (3-month follow-up)	0.208	1.187 (0.709 - 1.990)	.51
Time2 (6-month follow-up)	0.172	0.475 (0.732 - 2.070)	.43
Group * Time1	0.893	2.443 (0.876 - 6.808)	.09
Group * Time2	2.349	10.473 (3.276 - 33.491)	*<.001*
HIV self-testing (no=0, yes=1)			
Group (intervention vs control)	0.835	2.306 (1.593 - 3.340)	*<.001*
Time1 (3-month follow-up)	−0.225	0.798 (0.512 - 1.250)	.32
Time2 (6-month follow-up)	−0.151	0.860 (0.551 - 1.340)	.51
Group * Time1	1.732	5.651 (2.321 - 13.761)	*<.001*
Group * Time2	1.686	5.395 (2.221 - 13.117)	*<.001*
Consistent condom use[d] (no=0, yes=1)			
Group (intervention vs control)	0.899	2.457 (1.638 - 3.687)	*<.001*
Time1 (3-month follow-up)	0.699	2.011 (1.252 - 3.233)	*.004*
Time2 (6-month follow-up)	0.873	2.395 (1.475 - 3.885)	*<.001*
Group * Time1	−0.222	0.801 (0.310 - 2.068)	.65
Group * Time2	0.013	1.013 (0.382 - 2.683)	.98
Sexting (no=0, yes=1)			
Group (intervention vs control)	0.242	0.785 (0.530 - 1.163)	.23
Time1 (3-month follow-up)	−0.509	0.636 (0.375 - 0.964)	*.04*
Time2 (6-month follow-up)	−0.564	0.748 (0.351 - 0.923)	*.02*
Group * Time1	−0.125	0.882 (0.340 - 2.290)	.80
Group * Time2	−0.560	0.571 (0.216 - 1.510)	.26
Popper/alkyl nitrite inhalant use (no=0, yes=1)			
Group (intervention vs control)	−0.591	0.554 (0.365 - 0.841)	*.006*
Time1 (3-month follow-up)	−0.031	0.970 (0.595 - 1.579)	.90
Time2 (6-month follow-up)	−0.327	1.060 (0.436 - 1.193)	.21
Group * Time1	−0.808	0.446 (0.167 - 1.190)	.11
Group * Time2	−0.883	0.414 (0.147 - 1.160)	.09

[a]The generalized estimating equations model was adjusted for age, household registration, ethnicity, education level, occupation, and monthly income.

[b]RR: risk ratio.

[c]Italic values represent statistically significant at $P<.05$.

[d]There are missing values as some study participants had no sexual partners within the last 3 months and were therefore excluded from the analysis.

Social Media Intervention Significantly Increases AIDS-Related Knowledge

Knowledge scores for HIV/AIDS, PrEP, and PEP at baseline, 3 months, and 6 months were presented in Table 5 (mean scores shown). Comparison results indicated that compared to the control group, the intervention group showed significantly higher PrEP knowledge scores at both the 3-month and 6-month follow-ups (both $P<.05$). Furthermore, within the intervention group, knowledge scores for HIV/AIDS, PrEP, and PEP were significantly higher at both follow-up time points compared to baseline (all $P<.05$; Table 5).

XSL•FO
RenderX

Table . Comparison of awareness levels of relevant health knowledge between the intervention group and the control group before and after intervention.

Variables and groups	Baseline, mean (SD)	3-month follow-up, mean (SD)	6-month follow-up, mean (SD)
AIDS awareness level			
Intervention	13.10 (3.14)	15.10 (2.28)[a,b]	14.50 (2.02)[a,b]
Control	13.10 (3.14)	13.08 (2.32)	13.09 (2.75)
PrEP[c] awareness level			
Intervention	3.31 (1.56)	5.91 (1.39)[a,b]	6.01 (1.69)[a,b]
Control	3.14 (1.41)	4.87 (2.17)[a]	4.79 (2.38)[a]
PEP[d] awareness level			
Intervention	5.71 (1.83)	6.71 (1.60)[a,b]	6.68 (1.48)[a,b]
Control	5.58 (1.75)	5.48 (2.17)	5.42 (2.36)

[a]*P*<.05 compared to the baseline.

[b]*P*<.05 compared to the control group.

[c]PrEP: pre-exposure prophylaxis.

[d]PEP: post-exposure prophylaxis.

Social Media Intervention Improves Condom Use Self-Efficacy and Modifies Sex Attitudes in Men Who Have Sex With Men

Scores for condom use self-efficacy and safer sex attitudes at baseline, 3 months, and 6 months are presented in Table 6 (mean scores shown). Results demonstrated that compared to the control group, condom use self-efficacy was significantly higher in the intervention group at the 3-month follow-up (*P*<.05). Within the intervention group, condom use self-efficacy scores increased significantly from baseline to both the 3-month and 6-month follow-ups (both *P*<.05), with scores at 6 months being significantly higher than those at 3 months (*P*<.05; Table 6).

Table . Comparison of safe behavior attitudes and self-efficacy in condom use between the intervention group and the control group before and after intervention.

Variables and groups	Baseline, mean (SD)	3-month follow-up, mean (SD)	6-month follow-up, mean (SD)
Condom use self-efficacy			
Intervention	30.18 (6.27)	32.29 (5.38)[a,b]	33.21 (5.06)[a,c]
Control	29.03 (6.24)	30.82 (4.83)[a]	32.89 (5.83)[a,c]
Safe behavior attitudes			
Intervention	45.95 (8.65)	47.74 (7.82)	47.65 (7.20)
Control	45.39 (8.06)	46.73 (7.57)	46.87 (7.38)

[a]*P*<.05 compared to the control group.

[b]*P*<.05 compared to the baseline.

[c]*P*<.05 compared to the 3-month follow-up.

Discussion

Concerning the impact of the social media intervention on promoting HIV testing among men who have sex with men, this study found a higher uptake of HIV testing in the intervention group compared to the control group. This finding aligns with the results of some previous similar studies, such as two WeChat-based interventions conducted in Anhui, China, and Guangdong, China, where the HIV testing rates in the intervention groups were significantly higher than those in the control groups [17,18]. Our research further confirms the effectiveness of WeChat-based social media interventions in promoting HIV testing among the men who have sex with men population, indicating a positive effect of the social media–based strategy in increasing HIV testing coverage. Particularly for young, internet-savvy men who have sex with men, social media interventions could more effectively convey HIV prevention knowledge and enhance their testing awareness and capacity [32].

Regarding the effect of the social media intervention on reducing high-risk behaviors among men who have sex with men, our results demonstrated that the intervention group exhibited significantly lower frequencies of high-risk sexual behaviors and popper/alkyl nitrite inhalant use compared to the control group. This indicated that the social media intervention effectively reduces high-risk sexual behaviors among men who have sex with men, aligning with findings from multiple similar studies [9,31,33]. It demonstrates the broad application potential

XSL•FO
RenderX

of social media interventions in reducing high-risk behaviors and promoting HIV prevention. However, in China, research exploring the impact of WeChat-based social media interventions on reducing high-risk sexual behaviors remains relatively scarce. A WeChat-based intervention study conducted in Anhui Province, China, showed no significant difference in condom use behavior between the intervention and control groups, which may be related to the measures implemented in the control group [17]. This indicates that further and more in-depth research is still needed regarding the effects of WeChat-based social media interventions on reducing high-risk sexual behaviors in China. Furthermore, within the intervention group, after receiving the social media intervention, significant reductions were observed in the frequency of high-risk sexual behaviors, sexting behavior, and popper/alkyl nitrite inhalant use compared to baseline. GEE model analysis confirmed the positive effect of the intervention on reducing high-risk behaviors.

Furthermore, our study found that the positive effects observed in the intervention group at month 3 (the end of the intervention)—such as the promotion of HIV testing and HIV self-testing, reduction of high-risk sexual behaviors, and reduction of Rush popper use compared to the control group—were maintained and even strengthened at the 6-month follow-up assessment (3 months after the intervention ended). This indicates that the social media intervention in this study had sustained effects in promoting risk behavior change among men who have sex with men, continuing to positively influence their health behaviors even after the intervention concluded. This highlights the potential long-term benefits of social media as a health education platform for men who have sex with men, and the important role of social media as a platform for information dissemination and health education.

Our results showed that AIDS-related knowledge levels significantly increased from baseline to postintervention in both groups, with a more pronounced increase in the intervention group. This suggested that, compared to conventional methods, the social media intervention was more effective in enhancing relevant knowledge levels among men who have sex with men. Due to societal discrimination and stigma, especially in resource-limited settings, traditional sexual education and counseling services are often inaccessible or ineffective for men who have sex with men [34,35]. Emerging internet-based social media interventions, popular in men who have sex with men research, use engaging graphics, animations, and short videos to capture attention, foster interactivity and participation, and create a positive atmosphere for interaction. This strengthened men who have sex with men's understanding and retention of health information, further promoting the dissemination and depth of HIV knowledge delivery. Moreover, PrEP has been proven effective in reducing HIV risk [3]. As a novel and effective educational tool, social media interventions played a crucial role in improving AIDS knowledge, particularly awareness and willingness to use PrEP and PEP, among men who have sex with men. Therefore, future interventions should continue to leverage internet platforms and the strengths of social media to provide more personalized and acceptable services for men who have sex with men.

Using GEE models, this study also evaluated the intervention's impact on safer sex attitudes and condom use self-efficacy. For condom use self-efficacy, results showed a significant improvement in the intervention group compared to the control group. Within the intervention group, scores significantly increased at both 3 and 6 months compared to baseline. This indicated the positive role of the new media intervention in promoting condom use self-efficacy among men who have sex with men. Bogale et al [36] found significantly increased condom use intention among rural Ethiopian women following an audio-based HIV intervention. Hirshfield and Hennessy et al [37,38] found that television- and radio-based media interventions can increase the intention to use condoms among African American adolescents. Our study further confirms that media-based interventions can enhance individuals' confidence and capability to use condoms during sex practices. However, no significant differences were found in safer sex attitudes between the groups. This might indicate that men who have sex with men have relatively fixed attitudes toward sexual behavior, shaped by long-term sociocultural influences and personal experiences, which were difficult to change through short-term interventions. Future research could further extend the intervention and follow-up period to confirm its effects on attitudes toward safe sexual behavior.

Our study demonstrates that, compared to routine HIV prevention services alone, a social media–based intervention via WeChat significantly increased HIV testing rates, reduced high-risk sexual behaviors, and enhanced HIV-related knowledge among men who have sex with men. As a leading social media platform in China, WeChat features Mini Programs that require no additional installation, offering low user accessibility barriers [28,39]. This makes it particularly suitable for use in resource-limited settings such as Guangxi. Our research provides novel insights and a reference for integrating social media interventions with routine HIV prevention services to develop diversified strategies in resource-limited regions.

The IMB model categorizes factors influencing behavior change into 3 components: information, motivation, and behavioral skills. It proposes and validates that possessing disease prevention information and having the motivation to engage in prevention can, through behavioral skills, trigger preventive actions to initiate and sustain behavioral change [23]. This study found that interventions based on the IMB model effectively promoted the adoption of HIV preventive behaviors, following the theoretically hypothesized progressive pathway: First, procedural knowledge (such as PrEP usage guidelines and HIV self-testing procedures) provided through animated videos established a solid informational foundation (eg, individuals learned and accumulated knowledge related to HIV testing), while case studies and psychological counseling simultaneously enhanced personal risk perception and awareness of social support, collectively addressing the issues of "knowledge" and "motivation" (eg, individuals recognized the harms of not undergoing HIV testing through case studies, developed a sense of urgency, and strengthened their motivation and support for HIV testing through psychological counseling). Subsequently, skill-building components—such as weekly video check-ins, personalized consultations, and the distribution of free

XSL•FO
RenderX

self-testing kits—facilitated the practical mastery of operational skills and improved self-efficacy by creating a low-risk practice environment (eg, individuals learned and mastered the skills for HIV testing). These efforts successfully translated the accumulated information and motivation into actionable behavioral capabilities. Ultimately, this multicomponent synergistic effect drove significant behavioral improvements (eg, individuals ultimately decided to undergo HIV testing). In this study, this was specifically evidenced by increased HIV testing rates, enhanced self-efficacy in condom use, and reduced frequency of high-risk sexual behaviors. Thus, the study validates the utility of the IMB model in constructing effective social media interventions for men who have sex with men populations.

This study holds significant research and policy implications. First, by adopting an RCT design and integrating social media interventions with an established behavior change theory—the IMB model—we were able to more comprehensively account for factors influencing individual behavior change. Second, conducting the research in Guangxi, China—a relatively resource-limited region with a high HIV burden—expands the demographic and geographical diversity of studies on social media–based HIV prevention interventions among men who have sex with men. Third, by using multiple outcome measures to evaluate the intervention's effectiveness from various dimensions, we demonstrated that our approach effectively promoted HIV testing and reduced high-risk behaviors among men who have sex with men. Fourth, a 3-month postintervention follow-up was conducted, which confirmed the sustained effects of the social media intervention in promoting HIV testing and reducing high-risk behaviors. These findings provide new insights and methodologies for HIV prevention behavioral interventions.

Several limitations should be acknowledged. First, the participants were recruited from one site in China, which may limit the generalizability of the findings due to potential homogeneity within the recruited sample. Second, the survey relied on staff-administered one-on-one interviews, which might introduce discrepancies, especially for sensitive information like sexual behaviors. During follow-up, participants might have been influenced by social desirability bias, potentially underreporting high-risk behaviors and leading to report bias. Social desirability bias can distort the authenticity of data, weaken or even reverse the true relationships between variables. Third, the 6-month duration of this study limited the assessment of long-term effects postintervention. Fourth, although we aimed to reduce the risk of contamination by instructing participants to maintain confidentiality and differentiating service schedules, the use of respondent-driven sampling means that participants in the intervention and control groups may have social connections. Therefore, the possibility of intervention information spreading (contamination) cannot be entirely ruled out theoretically. Fifth, due to the nature of the intervention, complete blinding of participants was challenging, which may have allowed them to deduce their group assignment, potentially introducing expectation bias and reporting bias. Future studies should adopt a combination of strategies—such as implementing "anonymous" designs, using specialized measurement methods like the "randomized response technique," and incorporating social desirability scales for statistical control during data analysis—to systematically identify and mitigate the impact of social desirability bias, while also expanding recruitment to more diverse geographic regions to verify the effectiveness of social media–based strategies in resource-limited settings, extending the postintervention period of follow-up to examine their sustained impact on HIV prevention behaviors, considering designs such as cluster randomization or randomization based on social network communities to more effectively avoid contamination risks, and implementing a structured active control design to maintain participant blinding.

In conclusion, social media intervention based on the IMB model represented a promising strategy for HIV prevention in resource-limited settings, aimed at men who have sex with men. It significantly contributed to expanding HIV testing coverage, reducing high-risk behavior, enhancing AIDS-related knowledge, and improving safer sex self-efficacy within this high-risk population. Moreover, the positive effects became more pronounced with increasing the intervention and follow-up period. It demonstrated considerable potential and promising application prospects in the broader fight against HIV/AIDS. To achieve the 95-95-95 targets, adapting the IMB model–based social media intervention in low-resource settings and other key populations could prove highly beneficial.

Acknowledgments

We are grateful to the Nanning Center for Disease Control and Prevention (CDC), the Guangxi Caihong, and their peer educators for their invaluable support and assistance throughout this study. No generative artificial intelligence tools were used for manuscript preparation.

Funding

This work was supported by funds from the National Natural Science Foundation of China (NSFC, grant 82273694), the Natural Science Foundation of Guangxi Zhuang Autonomous Region (grants 2023GXNSFAA026287 and 2023GXNSFDA026036), and the Open Research Project from the Key Laboratory of Prevention and Control of Highly Prevalent Diseases (KLPCHPD202501). The funding source played a role in the design and conduct of the study and the preparation of the manuscript.

Data Availability

The deidentified data used and/or analyzed during the study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: HM, JL, LC
Data curation: HL, PC, QY, QC, HZ, WL, SH
Methodology: HM, JL, LC
Supervision: XY, PC, JH
Writing – original draft: HM, JL, LC
Writing – review & editing: XY, PC, JH

Conflicts of Interest

None declared.

Editorial Notice

This randomized study was only retrospectively registered, explained by authors due to delays with the registration platform's review process. The editor granted an exception from ICMJE rules mandating prospective registration of randomized trials, because the risk of bias appears low along with the authors' assertion that there were no deviations from the registered protocol once recruitment began. However, readers are advised to carefully assess the validity of any potential explicit or implicit claims related to primary outcomes or effectiveness, as retrospective registration does not prevent authors from changing their outcome measures retrospectively.

Multimedia Appendix 1
Intervention development and pilot study.
[PDF File, 1405 KB - jmir_v28i1e84279_app1.pdf]

Multimedia Appendix 2
Operational process of WeChat mini-program.
[PDF File, 1209 KB - jmir_v28i1e84279_app2.pdf]

Multimedia Appendix 3
Video tasks of the intervention plan based on the information-motivation-behavioral skills model.
[PDF File, 88 KB - jmir_v28i1e84279_app3.pdf]

Multimedia Appendix 4
Definitions of outcomes and methods of outcome assessment.
[PDF File, 155 KB - jmir_v28i1e84279_app4.pdf]

Checklist 1
CONSORT 2010 checklist.
[PDF File, 219 KB - jmir_v28i1e84279_app5.pdf]

References

1. Latest global and regional statistics on the status of the AIDS epidemic. : UNAIDS; 2025 URL: https://www.unaids.org/sites/default/files/2025-07/2025_Global_HIV_Factsheet_en.pdf [accessed 2026-03-05]
2. Dong MJ, Peng B, Liu ZF, et al. The prevalence of HIV among MSM in China: a large-scale systematic analysis. BMC Infect Dis 2019 Nov 27;19(1):1000. [doi: 10.1186/s12879-019-4559-1] [Medline: 31775654]
3. Gao Y, Hu Q, Leuba SI, et al. Medication non-adherence and condomless anal intercourse increased substantially during the COVID-19 pandemic among MSM PrEP users: a retrospective cohort study in four Chinese metropolises. Front Med (Lausanne) 2022;9:738541. [doi: 10.3389/fmed.2022.738541] [Medline: 35573017]
4. He N. Research progress in the epidemiology of HIV/AIDS in China. China CDC Wkly 2021 Nov 26;3(48):1022-1030. [doi: 10.46234/ccdcw2021.249] [Medline: 34888119]
5. Zhou Y, Lu Y, Ni Y, et al. Monetary incentives and peer referral in promoting secondary distribution of HIV self-testing among men who have sex with men in China: a randomized controlled trial. PLOS Med 2022 Feb;19(2):e1003928. [doi: 10.1371/journal.pmed.1003928] [Medline: 35157727]
6. Frescura L, Godfrey-Faussett P, Feizzadeh A A, et al. Achieving the 95 95 95 targets for all: a pathway to ending AIDS. PLOS One 2022;17(8):e0272405. [doi: 10.1371/journal.pone.0272405] [Medline: 35925943]
7. Shoptaw S. HIV prevention for people who use substances: evidence-based strategies. J Food Drug Anal 2013 Dec;21(4):S91-S94. [doi: 10.1016/j.jfda.2013.09.042] [Medline: 25214751]

XSL•FO
RenderX

8. Luo R, Xie Z, Silenzio VMB, Kuang Y, Luo D. Gay app use, sexuality traits, and high-risk sexual behaviors among men who have sex with men in China: mediation analysis. J Med Internet Res 2023 Nov 1;25:e49137. [doi: 10.2196/49137] [Medline: 37910154]
9. Johnson WD, Diaz RM, Flanders WD, et al. Behavioral interventions to reduce risk for sexual transmission of HIV among men who have sex with men. Cochrane Database Syst Rev 2008 Jul 16(3):CD001230. [doi: 10.1002/14651858.CD001230.pub2] [Medline: 18646068]
10. Romijnders K, de Groot L, Vervoort S, et al. The perceived impact of an HIV cure by people living with HIV and key populations vulnerable to HIV in the Netherlands: a qualitative study. J Virus Erad 2022 Mar;8(1):100066. [doi: 10.1016/j.jve.2022.100066] [Medline: 35280938]
11. Cao B, Gupta S, Wang J, et al. Social media interventions to promote HIV testing, linkage, adherence, and retention: systematic review and meta-analysis. J Med Internet Res 2017 Nov 24;19(11):e394. [doi: 10.2196/jmir.7997] [Medline: 29175811]
12. Young SD, Cumberland WG, Nianogo R, Menacho LA, Galea JT, Coates T. The HOPE social media intervention for global HIV prevention in Peru: a cluster randomised controlled trial. Lancet HIV 2015 Jan;2(1):e27-e32. [doi: 10.1016/S2352-3018(14)00006-X] [Medline: 26236767]
13. Pagkas-Bather J, Young LE, Chen YT, Schneider JA. Social network interventions for HIV transmission elimination. Curr HIV/AIDS Rep 2020 Oct;17(5):450-457. [doi: 10.1007/s11904-020-00524-z] [Medline: 32720253]
14. Yujie Z, Al Imran Yasin M, Alsagoff S, Hoon AL. The mediating role of new media engagement in this digital age. Front Public Health 2022;10:879530. [doi: 10.3389/fpubh.2022.879530] [Medline: 35586011]
15. Muessig KE, Bien CH, Wei C, et al. A mixed-methods study on the acceptability of using eHealth for HIV prevention and sexual health care among men who have sex with men in China. J Med Internet Res 2015 Apr 21;17(4):e100. [doi: 10.2196/jmir.3370] [Medline: 25900881]
16. Wang C, Ong JJ, Zhao P, et al. Expanding syphilis test uptake using rapid dual self-testing for syphilis and HIV among men who have sex with men in China: a multiarm randomized controlled trial. PLOS Med 2022 Mar;19(3):e1003930. [doi: 10.1371/journal.pmed.1003930] [Medline: 35235573]
17. Zhu X, Zhang W, Operario D, et al. Effects of a mobile health intervention to promote HIV self-testing with MSM in China: a randomized controlled trial. AIDS Behav 2019 Nov;23(11):3129-3139. [doi: 10.1007/s10461-019-02452-5] [Medline: 30852728]
18. Li JS, Gu YZ, Hou FS, et al. Delivery of WeChat-based HIV result e-reports in social networks for recruitment of high-risk population: baseline data from a cluster randomized controlled trial. J Med Internet Res 2023 Jun 15;25:e46793. [doi: 10.2196/46793] [Medline: 37318850]
19. Zou Y, Deng M, Wang Q, Zhang Q, Rong Y. Evolution characteristics of new urbanization in the provincial capital cities of Western China. Front Environ Sci 2022;10. [doi: 10.3389/fenvs.2022.926444]
20. Ya-Feng Z, Min D, Ya-Jing L, Yao R. Evolution characteristics and policy implications of new urbanization in provincial capital cities in Western China. PLOS One 2020;15(5):e0233555. [doi: 10.1371/journal.pone.0233555] [Medline: 32453772]
21. Prabawanti C, Dijkstra A, Riono P, Hartana G. A survey on HIV-related health-seeking behaviors among transgender individuals in Jakarta, based on the theory of planned behavior. BMC Public Health 2015 Nov 16;15:1138. [doi: 10.1186/s12889-015-2480-0] [Medline: 26573618]
22. Fishbein M. The role of theory in HIV prevention. AIDS Care 2000 Jun;12(3):273-278. [doi: 10.1080/09540120050042918] [Medline: 10928203]
23. Kalichman SC, Picciano JF, Roffman RA. Motivation to reduce HIV risk behaviors in the context of the information, motivation and behavioral skills (IMB) model of HIV prevention. J Health Psychol 2008 Jul;13(5):680-689. [doi: 10.1177/1359105307082456] [Medline: 18519441]
24. Krishnan A, Ranjit YS, Zhou X, Altice FL. Predicting antiretroviral medication adherence among substance-using people with HIV: test and extension of the information-motivation-behavioral skills model. J Behav Med 2025 Jun;48(3):403-413. [doi: 10.1007/s10865-025-00557-y] [Medline: 39924604]
25. Wan Mohamad Darani WNS, Mat Ruzlin AN, Azhar ZI, Chen XW. Development and validation of the information-motivation-behavioural skills model-based human immunodeficiency virus education kit for adolescents in Malaysia. Sci Rep 2024 Mar 22;14(1):6890. [doi: 10.1038/s41598-024-57593-y] [Medline: 38519534]
26. Chen H, Luo L, Pan SW, et al. HIV epidemiology and prevention in Southwestern China: trends from 1996-2017. Curr HIV Res 2019;17(2):85-93. [doi: 10.2174/1570162X17666190703163838] [Medline: 31269884]
27. Pang X, Wei H, Huang J, et al. Patterns and risk of HIV-1 transmission network among men who have sex with men in Guangxi, China. Sci Rep 2021 Jan 12;11(1):513. [doi: 10.1038/s41598-020-79951-2] [Medline: 33436843]
28. Zhang X, Wen D, Liang J, Lei J. How the public uses social media wechat to obtain health information in China: A survey study. BMC Med Inform Decis Mak 2017 Jul 5;17(Suppl 2):66. [doi: 10.1186/s12911-017-0470-0] [Medline: 28699549]
29. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c332. [doi: 10.1136/bmj.c332] [Medline: 20332509]
30. Liu L, Liu JH, Ruan YH, et al. Rush poppers use and associated factors among men who have sex with men in Guangxi, China. Chin J AIDS STD 2024;30(1):23-28. [doi: 10.13419/j.cnki.aids.2024.01.05]

XSL•FO
RenderX

31. Knight R, Karamouzian M, Salway T, Gilbert M, Shoveller J. Online interventions to address HIV and other sexually transmitted and blood-borne infections among young gay, bisexual and other men who have sex with men: a systematic review. J Int AIDS Soc 2017 Nov;20(3). [doi: 10.1002/jia2.25017] [Medline: 29091340]
32. Moreno MA, D'Angelo J. Social media intervention design: applying an affordances framework. J Med Internet Res 2019 Mar 26;21(3):e11014. [doi: 10.2196/11014] [Medline: 30912754]
33. Carpenter KM, Stoner SA, Mikko AN, Dhanak LP, Parsons JT. Efficacy of a web-based intervention to reduce sexual risk in men who have sex with men. AIDS Behav 2010 Jun;14(3):549-557. [doi: 10.1007/s10461-009-9578-2] [Medline: 19499321]
34. Nelson KM, Carey MP. Media literacy is an essential component of HIV prevention for young men who have sex with men. Arch Sex Behav 2016 May;45(4):787-788. [doi: 10.1007/s10508-016-0712-8] [Medline: 26895006]
35. Gyamerah AO, Taylor KD, Atuahene K, et al. Stigma, discrimination, violence, and HIV testing among men who have sex with men in four major cities in Ghana. AIDS Care 2020 Aug;32(8):1036-1044. [doi: 10.1080/09540121.2020.1757020] [Medline: 32362131]
36. Bogale GW, Boer H, Seydel ER. Effects of a theory-based audio HIV/AIDS intervention for illiterate rural females in Amhara, Ethiopia. AIDS Educ Prev 2011 Feb;23(1):25-37. [doi: 10.1521/aeap.2011.23.1.25] [Medline: 21341958]
37. Hirshfield S, Chiasson MA, Joseph H, et al. An online randomized controlled trial evaluating HIV prevention digital media interventions for men who have sex with men. PLOS One 2012;7(10):e46252. [doi: 10.1371/journal.pone.0046252] [Medline: 23071551]
38. Hennessy M, Romer D, Valois RF, et al. Safer sex media messages and adolescent sexual behavior: 3-year follow-up results from project iMPPACS. Am J Public Health 2013 Jan;103(1):134-140. [doi: 10.2105/AJPH.2012.300856] [Medline: 23153149]
39. Cheng A, Ren G, Hong T, Nam K, Koo C. An exploratory analysis of travel-related WeChat mini program usage: an affordance theory perspective. In: Information and Communication Technologies in Tourism 2019: Springer; 2019:333-343. [doi: 10.1007/978-3-030-05940-8_26]

Abbreviations

CBO: community-based organization
CONSORT: Consolidated Standards of Reporting Trials
GDP: gross domestic product
GEE: generalized estimating equation
IMB: information-motivation-behavioral skills
PEP: post-exposure prophylaxis
PrEP: pre-exposure prophylaxis
RCT: randomized controlled trial
UNAIDS: United Nations Programme on HIV/AIDS

Edited by A Mavragani; submitted 17.Sep.2025; peer-reviewed by E Ocean, X Li; accepted 30.Jan.2026; published 07.Apr.2026.

Please cite as:
Mai H, Liu J, Cheng L, Liang H, Cui P, Yao Q, Cao Q, Zhang H, Liao W, He S, Yang X, Cen P, Huang J
Social Media Intervention Based on the Information-Motivation-Behavioral Skills Model Promotes HIV Testing and Reduces High-Risk Behaviors Among Men Who Have Sex With Men in Resource-Limited Settings in China: Randomized Controlled Trial
J Med Internet Res 2026;28:e84279
URL: https://www.jmir.org/2026/1/e84279
doi:10.2196/84279

© Huanzhuo Mai, Jianyuan Liu, Liping Cheng, Hao Liang, Ping Cui, Qin Yao, Qin Cao, Heng Zhang, Wenping Liao, Shiran He, Xing Yang, Ping Cen, Jiegang Huang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

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

Maximilian Scholl[1], MA; Claas Lendt[1,2], MSc; Sebastian Appelbaum[3,4], MSc; Bianca Biallas[1], PhD; Katja Brenk-Franz[5], PhD; Chloé Chermette[6], PhD; Friederike Frank[7,8], MSc; Angeli Gawlik[6], PhD; Lisa Giesen[9], PhD; Martina Heßbrügge[10], MD; Lucas Küppers[11], MD; Larisa Pilic[12,13], MSc; Marcus Redaèlli[9], MD; Lara Schneider[14], BA; Frank Vitinius[14,15], MD; Stefan Wilm[16], MD; Uwe Konerding[3,4], PhD

[1]Institute of Movement Therapy and Movement-oriented Prevention and Rehabilitation, Section I, German Sports University Cologne, Cologne, Germany

[10]Institute of General Practice, University of Duisburg-Essen, Essen, Germany

[11]Institute of Family Medicine and General Practice, University of Bonn, Bonn, Germany

[12]Institute of General Practice, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany

[13]Department of Child and Adolescent Psychiatry, Psychosomatic Medicine and Psychotherapy, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany

[14]Department of Psychosomatics and Psychotherapy, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany

[15]Department of Psychosomatic Medicine, Robert Bosch Hospital, Stuttgart, Germany

[16]Institute of General Practice, Centre for Health and Society, Medical Faculty and University Hospital, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[2]Department of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany

[3]Department of Psychology and Psychotherapy, Witten/Herdecke University, Witten, Germany

[4]Trimberg Research Academy, University of Bamberg, An der Weberei 5, Bamberg, Germany

[5]Institute of Psychosocial Medicine, Psychotherapy and Psychooncology, Jena University Hospital, Jena, Germany

[6]Institute of Psychology, Section I: Health and Social Psychology, German Sport University Cologne, Cologne, Germany

[7]Institute for Digitalization and General Medicine, Medical Faculty, RWTH Aachen University, Aachen, Germany

[8]Center for Rare Diseases Aachen (ZSEA), Medical Faculty, RWTH Aachen University, Aachen, Germany

[9]Institute of Health Economics and Clinical Epidemiology, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany

Corresponding Author:
Uwe Konerding, PhD
Department of Psychology and Psychotherapy, Witten/Herdecke University, Witten, Germany

Abstract

Background: The targeted application and design of digital health interventions (DHIs) require an understanding of usage determinants. Usage includes uptake (initial use) and frequency (extent of use), but it is unclear whether both components are driven by the same determinants.

Objective: This study aimed to examine the determinants of uptake and frequency of use and assess whether they differ.

Methods: The investigated DHI was a web portal provided in an intervention for improving disease-related self-management. This study is a secondary analysis of intervention group data from a parallel-group randomized controlled trial. Eligibility criteria were being an adult and being diagnosed with type 2 diabetes and/or coronary heart disease. Sociodemographic, psychological, and health-related variables were examined as determinants. Determinants were analyzed using simple and multiple regression models. Uptake was analyzed using logistic regression, and frequency was analyzed using negative binomial regression with robust SEs. Frequency was analyzed for those who used the DHI at least once. Except for sociodemographic variables, all other variables were standardized to a range from 0 to 1. For simple regression, inflation of the α error due to multiple testing was controlled via the approach of Benjamini and Hochberg, and for multiple regression, it was controlled via the significance of the complete multiple regression model.

Results: Of 462 intervention group members, 199 (43.1%) used the web portal at least once. After controlling for inflation of the α error, simple regression for uptake yielded significant effects for higher education (B=0.56, 95% CI 0.18-0.95; *P*=.004),

XSL•FO
RenderX

openness (B=1.08, 95% CI 0.33-1.83; *P*=.005), intention regarding physical activity (B=2.28, 95% CI 1.30-3.26; *P*<.001), and intention regarding healthy nutrition (B=2.30, 95% CI 1.30-3.31; *P*<.001). The multiple regression model for uptake was highly significant (*P*<.001), with significant positive associations for intentions regarding physical activity (B=1.86, 95% CI 0.74-2.97; *P*=.001) and healthy nutrition (B=2.22, 95% CI 1.00-3.44; *P*<.001), as well as a significant negative association for patient activation (B=−3.20, 95% CI −4.95 to −1.46; *P*<.001). After controlling for inflation of the α error, simple regression for frequency yielded no statistically significant effect, and the multiple regression model for frequency was not significant (*P*=.07).

Conclusions: This study is innovative in jointly examining determinants of the uptake and frequency of use of the same DHI within a single context and sample. By demonstrating that factors driving uptake do not necessarily increase the frequency of use, it advances existing research. The study contributes to a more differentiated understanding of DHI use and shows that distinct strategies are required to promote adoption versus sustained engagement. Applying this approach to other DHIs and settings may support more targeted and equitable digital health implementation in real-world contexts, thereby optimizing digital health deployment strategies overall.

Trial Registration: German Clinical Trials Registry (DRKS) DRKS00020592; https://drks.de/search/de/trial/DRKS00020592

International Registered Report Identifier (IRRID): RR2-10.1186/s13063-021-05636-4

(J Med Internet Res 2026;28:e80895) doi:10.2196/80895

KEYWORDS

digital health interventions; eHealth; type 2 diabetes; coronary heart disease; predictors of use; chronic disease self-management

Introduction

Chronic noncommunicable diseases, such as type 2 diabetes mellitus (T2DM) and coronary heart disease (CHD), place an enormous burden on patients and health care providers worldwide [1,2]. Improving patient self-management could help to reduce this burden [3,4]. The use of digital health interventions (DHIs), that is, interventions meant to influence health via digital tools, is a promising approach that might serve this purpose. Some examples are patient portals [5], mobile health apps [6], and web-based interventions [7]. DHIs make it possible to provide personalized, location-independent, and cost-effective support to patients [8,9]. Accordingly, DHIs have already been used in rehabilitation [10], medication adherence [11], and chronic disease management [12-15], as well as in the promotion of healthy lifestyles [16,17]. Recent systematic reviews and meta-analyses have indicated that DHIs help to improve hemoglobin A_{1c} levels [13,18,19], CHD status [14], and health-promoting behaviors, such as physical activity and healthy nutrition intake [14,20].

DHIs can only be effective if they are actually used. In some cases, repeated use might be necessary to have an impact [21,22]. Accordingly, purposeful design of interventions requires knowledge regarding the determinants of usage, that is, knowledge regarding conditions under which the DHI in question is used more and knowledge regarding the conditions under which the DHI is used less. Knowledge regarding the conditions under which the DHI is used less can be helpful in two regards: (1) it might show how use could be increased by adapting the DHI itself and/or its mode of presentation, and (2) it might show where usage is limited, that is, when neither the condition can be changed nor the DHI can be adapted. In this case, alternative interventions for the purpose addressed by the DHI in question are needed, and these might very well be nondigital interventions.

Previous research regarding potential determinants of DHI usage primarily focused on sociodemographic factors, especially age, gender, and education. Findings regarding age [23-26] and gender [24-27] were inconsistent across studies, whereas education showed a more consistent positive association with DHI usage [28-30]. Further, previous research regarding the potential determinants of DHI referred to psychological variables, such as the “big five” [29]; loneliness, depression, and anxiety [31]; health literacy [31,32]; patient activation [33,34]; and health-related behavioral intentions [26,35]. Among the “big five,” openness appeared to be positively associated with usage, whereas extraversion appeared to be negatively associated [29]. There were mixed results for mental health factors [31,35] and behavioral intentions [26,35], whereas health literacy appeared to be positively associated with usage [31,32]. Moreover, mixed results were reported for patient activation, with some studies reporting no association [36] and others reporting a positive association [37].

A major problem in research regarding the determinants of DHI usage is the operationalization of usage itself [38]. In previous research, the following two different operationalizations were mainly applied: (1) uptake, that is, the act of using the DHI at least once [26,30,39], and (2) frequency, that is, the amount of use thereafter [23-25]. Uptake should, at best, be investigated in all persons who have used the DHI, and frequency should be investigated only in those who have used it at least once. Both operationalizations constitute different components of usage. These 2 components might be determined by completely different factors. Therefore, in research regarding the determinants of DHI use, both components should be distinguished from each other. This is not only important for the sake of scientific precision but also relevant with regard to deriving implications for the targeted design of the DHI and the corresponding communication and outreach strategies. Insights into the determinants of uptake are important for the design of communication and outreach strategies, whereas insights into the determinants of frequency are relevant for the design of the DHIs themselves.

XSL•FO
RenderX

The best approach for identifying which variables affect uptake and which affect frequency is to investigate both effects for the same DHI in the same context within the same sample. There is hardly any previous research on this approach. The objective of this study is to demonstrate this approach, that is, to investigate the determinants of uptake and the frequency of use for the same DHI in the same context within the same sample.

Methods

Study Design

This study is a secondary analysis of a randomized controlled trial (RCT) involving a complex intervention called the Personalized Self-management Support Program (P-SUP; 2019 - 2024) [40]. The P-SUP consists of the following four modules: (1) peer support groups, which met for weekly group-based exercise sessions and regularly scheduled digital expert education classes; (2) telephone coaching, which is provided to a subgroup of patients with low health literacy and low patient activation [41]; (3) feedback reports, which are provided to patients and contain routine data on medical parameters that can be used as a basis for discussing their state of health with their general practitioners during routine check-ups; and (4) a browser-based web portal, which provides practical and theoretical content on exercise, nutrition, and motivation. The last component is the DHI investigated in this study.

The P-SUP was implemented and evaluated within disease management programs (DMPs) [42] for T2DM and CHD in North Rhine-Westphalia, Germany. DMPs are structured treatment programs that address individuals with chronic conditions and that are added to standard primary care. The trial followed a parallel-group randomized controlled design with a delayed intervention for the control group and was conducted consecutively (ie, participants started sequentially rather than simultaneously). Participants were randomized to the intervention group and control group with an allocation ratio of 6:7. In the first 18 months, the intervention group received the intervention, whereas the control group received regular DMP care. After this interval, the control group was offered the intervention for 9 months. The primary outcome of the original trial was the number of hospital days during the year preceding the end of the intervention. Secondary outcomes covered three domains: (1) self-reported physical health status, (2) self-reported health-related behaviors, and (3) self-reported thoughts and feelings (ie, psychological variables).

Patient and Public Involvement

No formal patient or public involvement took place in the design, conduct, or reporting of the trial.

Changes to the Trial Protocol

In some aspects, the conduction of the trial deviated from the protocol. Not all peer support group meetings could be guided by members of the groups because not enough participants were willing to take on this role. Moreover, some peer support group meetings had to be held online because of the COVID-19 pandemic, and the feedback report only started with a delay due to organizational problems.

Harms

In accordance with the trial protocol, harms and potential adverse events were monitored and documented during peer support group meetings and telephone coaching sessions. No adverse events or harms were reported.

Sample Size

The sample size of the original trial was determined based on the primary outcome (ie, the number of hospital days in the last year of the intervention interval) and was powered to detect small effect sizes (Cohen d=0.2) with a significance level of .05 and a power of 0.80, resulting in a required sample size of 1001 participants, as detailed in the trial protocol [40]. No separate power calculation was conducted for this secondary analysis, which focused on a specific intervention component and on the intervention group alone.

Investigated DHI

The browser-based web portal provided as part of the complete intervention (ie, the DHI investigated in this study) was developed based on the self-determination theory (SDT) [43]. Among other things, the SDT posits that individuals strive to fulfill the following 3 basic psychological needs: autonomy (eg, possession of choice and action scope), competence (eg, personal growth and skill development), and social relatedness (eg, relationship building and integration). The DHI was developed within this conceptual framework, with a particular focus on supporting autonomy and enhancing competence. The development was conducted in three stages: (1) a workshop, (2) content design by core teams, and (3) consolidation of content and posting on the web portal.

The aim of the workshop was to define the content, communication style, and visual design appropriate for the target group. This workshop was guided by an external digital agency based in Cologne. Participants of the workshop included members of the interdisciplinary project team (sports science, health science, psychology, and medicine), a practicing physician, and a potential study participant. Based on the workshop results, content and functions for the web portal were developed. Subsequently, the following four core teams were established (each for a different topic): (1) medicine, (2) physical activity, (3) nutrition, and (4) motivation. These core teams systematically developed the intervention content. In the final step, the developed content was consolidated and successively integrated into the DHI by a specific project team member.

The web portal included the following: (1) evidence-based information regarding T2DM and CHD, physical activity, nutrition, and self-motivation; (2) exercise videos conducted by a sports therapist focusing on strength, endurance, coordination, and flexibility; (3) recipes for healthy food; (4) a motivational 12-week coaching program; and (5) a download area for nutritional help tools (Figure 1). Usability and visual design of the complete web portal were tested formatively during the development process with a cohort comparable to the final study sample. The 12-week coaching program was developed iteratively according to the Integrate, Design, Assess, and Share (IDEAS) framework [44-46]. For further details on the content and design of the web portal, see Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. Screen capture images from the P-SUP (Personalized Self-management Support Program) web portal.

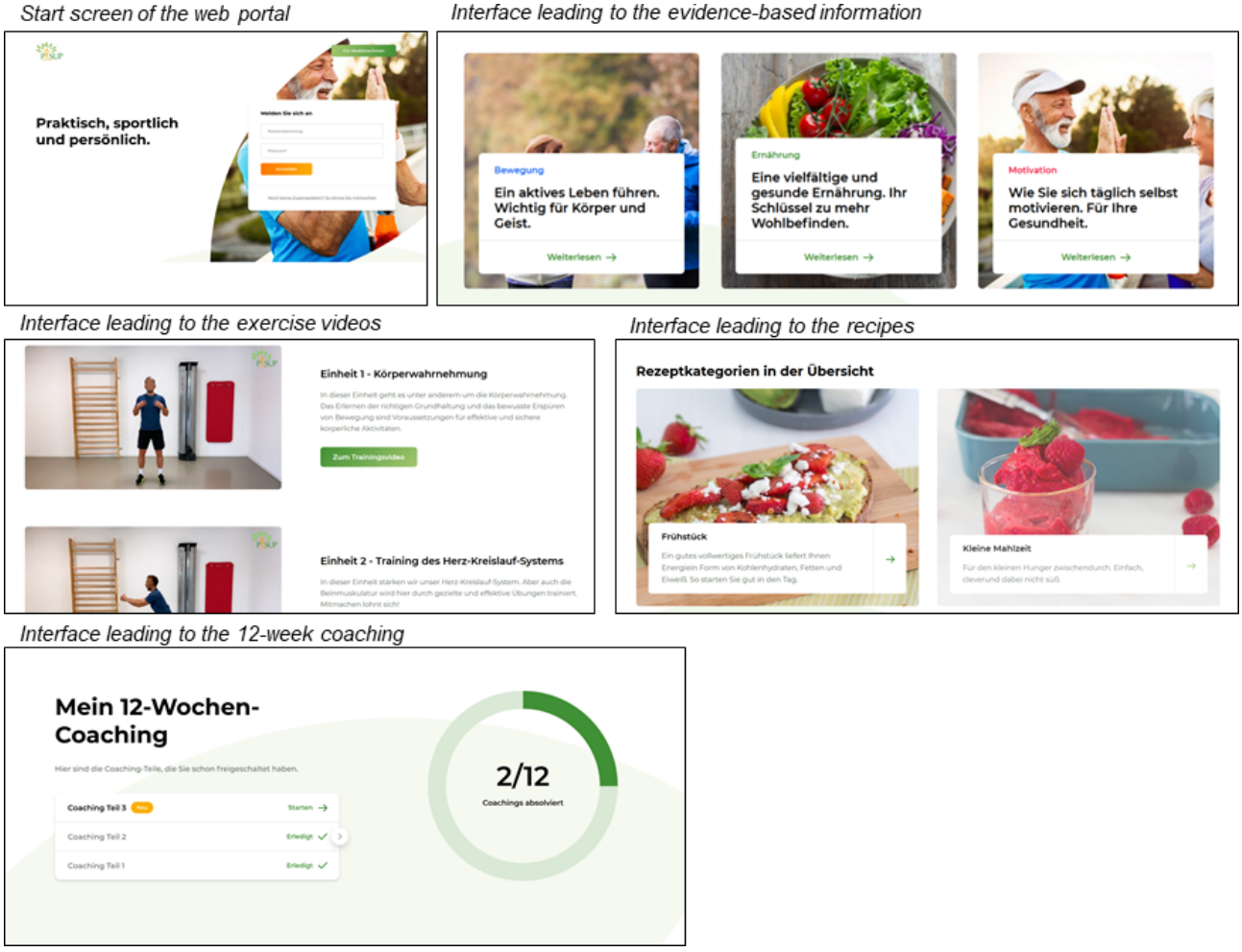

The final web portal was made accessible to all participants of the intervention group by sending them a log-in identification number and a password by mail prior to the start of the study. In addition to the log-in credentials, they received a letter that included an explanation of the web portal contents, data protection information (eg, data storage, server location, and privacy), a step-by-step guide for the registration process, and contact details (email and telephone) for technical support. To maintain participant interest in the application, the web portal was updated with new content at irregular intervals. To accompany these updates and support user engagement, a total of 10 newsletters were distributed via email at regular intervals throughout the intervention period. These newsletters explained the different components of the P-SUP intervention and provided direct links to newly published content.

Study Participants

The study participants of the complete RCT were recruited between February and October 2021 from 6 different regions in North Rhine-Westphalia, Germany, including Aachen, Bonn, Cologne, Düsseldorf, Essen, and Bünde, via general practitioners, newspapers, and online advertisements. General practitioners informed eligible patients about the intervention during routine DMP appointments and enrolled them directly. Individuals who responded to newspaper or online advertisements were enrolled in the study by the study team if they met the inclusion criteria. The inclusion criteria for patients were as follows: (1) being ≥18 years of age, (2) being enrolled in a DMP for T2DM and/or CHD [42], and (3) being a resident of North Rhine-Westphalia. The exclusion criteria were as follows: (1) insufficient German language skills, (2) severe cognitive or physical impairments, and (3) severe comorbidities. Those persons who were assigned to the intervention group and who were finally included in the analyses for examining the effectiveness of the complete intervention were also included in the secondary analysis presented in this article.

Recruiting Sites and Personnel

Recruitment was conducted through general practitioners participating in the DMPs for T2DM and/or CHD. The eligibility for physician recruitment was limited to standard DMP accreditation. No additional eligibility criteria, training, or intervention delivery responsibilities were applicable, and no additional personnel involved in delivering or supporting the DHI were investigated in this secondary analysis.

Data Collection and Measures

DHI Usage

DHI usage was determined based on server data from the web portal, covering the entire 18-month intervention period. Because of the consecutive nature of the study, these periods began in the period from April 2021 to February 2022. The

XSL•FO
RenderX

collected data included information on log-ins and viewed content.

However, for some participants, the documentation of web portal usage did not start with an identified log-in. This indicated technical problems with the documentation. As a result, the log-ins, which were meant to be the basic units of DHI usage, could not be determined via the log-in data. Instead, log-ins were partly reconstructed based on timestamps of viewed content. A new log-in was assumed whenever there was a gap of at least 2 hours between 2 consecutive content views. This threshold was deliberately set longer than the 30 or 60 minutes recommended in previous literature [47,48], because certain types of content on the web portal, such as exercise videos, lasted around 30 minutes. Thus, if a participant logged into the web portal and subsequently watched 1-2 videos, the activity following this should still be attributed to the same session.

Uptake was operationalized as the incidence of at least one log-in, whereas frequency was operationalized as the total number of log-ins minus the first one. The first log-in was subtracted because the frequencies were analyzed to obtain information about processes that occur after the first log-in. As a sensitivity analysis, all analyses referring to frequencies were also performed with log-ins defined using 30- and 60-minute intervals.

All variables determined for operationalizing usage only reflected direct use of the DHI. Indirect use was possible by downloading exercise videos or healthy food recipes and using them offline. This shifted some web portal activities from tracked online sessions to unobserved offline use and thus likely led to an underestimation of usage frequency.

Possible Determinants of DHI Usage

Possible determinants of DHI usage were selected from two sources: (1) register data from the organization unit coordinating the whole project, and (2) data from the baseline survey conducted prior to the start of the intervention. The register data were collected and updated from January 2021 to April 2024. The baseline survey was conducted from February to December 2021. The criterion for selecting variables from these 2 sources was that these variables had already been considered as possible determinants in previous research regarding the determinants of DHI usage.

Sociodemographic Variables

Sociodemographic characteristics included age, gender, and educational attainment. Age and gender were assessed via both the register data and the baseline survey. Educational attainment was only assessed via the baseline survey. In this survey, participants were asked whether they had the qualification required to attend a higher education institution, such as a university of applied sciences or an ordinary university. Those who answered “yes” were classified as having a higher level of educational attainment, while those who answered “no” were classified as having a lower level. Data for age and gender were available for all participants. Missing data for education were multiply imputed by producing 50 different datasets using the multiple imputation procedure provided as standard by SPSS (IBM Corp).

Psychological Variables

Psychological variables were solely assessed via the baseline survey. These variables can be assigned to two different categories: (1) general psychological variables, and (2) health-related psychological variables.

Five general psychological variables included the “big five” (ie, openness, conscientiousness, extraversion, agreeableness, and neuroticism). Three further general psychological variables included loneliness, depression, and anxiety. The “big five” variables were assessed using the Big Five Inventory-10 items (BFI-10) [49], which consists of 10 items, with each dimension addressed by 2 items. Loneliness was assessed using the 6-item short version of the revised University of California Los Angeles (UCLA) Loneliness Scale [50,51]. Depression and anxiety were assessed using the Patient Health Questionnaire-4 items (PHQ-4). The PHQ-4 is a 4-item self-report questionnaire that measures depression using the first 2 items and anxiety using the last 2 items [52].

The health-related psychological variables were health literacy, patient activation, and intentions regarding physical activity or healthy nutrition. Health literacy was assessed using the 5-item Brief Health Literacy Screen (BHLS) [53], and patient activation was evaluated using the Patient Activation Measure-13 (PAM13-D) [54,55]. Each of the 2 intention variables was assessed using 1 item with a 7-category scale. Physical activity was operationalized as engaging in at least 150 minutes of moderate or vigorous physical activity per week (“Do you intend to engage in at least two-and-a-half hours of moderate or vigorous physical activity per week in the coming months?”), and healthy nutrition was operationalized as performing at least five of nine items of healthy nutrition behavior (“Do you intend to follow at least five of the dietary behaviors described above in the coming months?”) listed in a self-constructed questionnaire (Multimedia Appendix 2).

Missing data for psychological variables were multiply imputed by producing 50 different datasets using the multiple imputation procedure provided as standard by SPSS. Except for the PAM13-D, all variables based on more than one item were constructed by summing up the scale values assigned to the different answer categories, with all answer categories coded with equal distances between all adjacent categories. For the PAM13-D, the sums determined in this way were additionally transformed into scale values obtained by Rasch scaling [56] (Multimedia Appendix 3).

All psychological variables were standardized via a linear transformation, with 0 being the lowest and 1 the highest value that could be obtained with the respective measurement instrument. This standardization was performed because all these variables possess at best an interval scale level, and thus, the corresponding measurement values have no absolute meaning but only a meaning relative to 2 reference points. For all variables, the boundaries of the corresponding measurement instrument constitute the best available reference points, and coding these reference points with 0 and 1 makes it possible to compare the magnitudes of the effects of these variables on DHI usage. All variables coded this way were pooled according to the denomination applied for the respective variable (ie, 1 is

XSL•FO
RenderX

the highest level of extraversion, agreeableness, openness, conscientiousness, neuroticism, loneliness, depression, anxiety, patient activation, health literacy, intention regarding physical activity, or intention regarding healthy nutrition).

Statistical Analysis

Patient characteristics were documented based on complete data. This was put into practice separately for nonusers and users, as well as for all investigated persons together. Gender and education were described using absolute and relative frequencies; all other variables were described using means and SDs. To provide a clear overview of overall usage of the web portal, the 18-month intervention period (78 weeks) was divided into 13 intervals of 6 weeks each. This segmentation was chosen because 78 is divisible by 13, allowing for consistent time intervals across the entire period. For each 6-week interval, user activity was described using relative and absolute frequencies. Absolute frequencies included (1) the absolute frequency of log-ins per interval; (2) the absolute frequency of users per interval; and (3) the absolute frequency of first log-ins per interval. Relative frequencies included (1) the relative frequency of log-ins per interval, expressed as a percentage of all log-ins across all intervals; (2) the relative frequency of users per interval, expressed as a percentage of all user occurrences across all intervals; and (3) the relative frequency of first log-ins per interval, expressed as a percentage of all first log-ins across all intervals.

All further analyses were conducted with missing values being multiply imputed. In response to comments raised during the review process, a test for data missing completely at random (MCAR) was performed [57]. However, the results of this test have no implications for the interpretation of the results of the further analyses, because statistical analyses based on multiply imputed data are unbiased as long as the much less restrictive condition of MAR (ie, data missing at random) is fulfilled [58]. However, for this condition, no feasible statistical test presently exists.

The analyses regarding uptake and frequency were conducted with different samples. For uptake, these were all members of the intervention group, while for frequency, these were only those who had used the web portal at least once (ie, users). In both cases, three different analyses were performed: (1) an analysis of the correlations between the variables investigated as possible determinants; (2) simple regression analyses for all possible determinants, with either uptake or frequency as a criterion; and (3) a multiple regression analysis with all possible determinants entered as predictors in a common model, with either uptake or frequency as a criterion. If the usage variable was uptake, logistic regression was applied, while if the usage variable was frequency, negative binomial regression with data-based (robust) SEs was applied. The latter procedure was chosen because negative binomial regression is specifically designed for count variables and because data-based SEs are not based on a specific model, so that inference statistical tests based on these errors cannot be compromised by invalidity of the presupposed model [59].

Several different statistics were computed in connection with the regression analyses. These were regression coefficients, corresponding 95% CIs, and corresponding *P* values for all regression analyses. In logistic regression, Nagelkerke pseudo R^2 was used to assess how well the determinants explain the actual values. In negative binomial regression, the proportion of explained variance was used for this purpose. For negative binomial regression, the dispersion parameter was also computed. This parameter refers to the variance of the actual values that are predicted by the same predicted value. If this variance is equal to or lower than the predicted value, the dispersion parameter is 0. Otherwise, this parameter is larger than 0, with larger dispersion parameters indicating larger variances. To investigate how the different predictor variables interfere with each other when regression coefficients in multiple regression are estimated, variance inflation factors were computed for both multiple regressions. These factors are equal to 1 if there is no interference at all, and they increase with increasing interference. With a variance inflation factor larger than 5, the estimates of the regression coefficients are not very trustworthy.

Inflation of α errors due to multiple testing was controlled differently for simple and multiple regression. For simple regression, the approach of Benjamini and Hochberg was applied [60]. In this approach, a specific comparator is computed for each *P* value. For this purpose, the *P* values for the different tests are arranged in ascending order. The comparator for each *P* value is as follows:

$$\text{comparator}_i = \frac{i}{m}\alpha \quad (1)$$

where *i* is the test number in ascending order, *m* is the number of tests, and α is the significance level applied for an individual test. The largest *i* for which the *P* value is smaller than the comparator is identified, and all tests with an index smaller than or equal to this *i* are considered to be significant. In multiple regression, the final decision regarding the significance was based on the result of the significance test of the complete regression model. If this test yielded a significant result, all significant results of tests regarding the individual regression coefficients were considered as actually significant; otherwise, none of the individual test results were considered as significant.

To compare the regression models for uptake and frequency, the rank correlation between the *P* values was computed.

All data were analyzed using IBM SPSS Statistics version 29.

Ethical Considerations

Ethical approval for the complete RCT was obtained from the Ethics Committee of the Medical Faculty of the University Hospital Cologne in July 2020 (20-1155). All patients provided written informed consent prior to participation in the RCT. This consent also covered the use of pseudonymized data for secondary analyses without requiring additional consent. Patient data were stored within three separate organizational units at the University Hospital Cologne: (1) Trust Center, which held address data linked to study identification numbers; (2) Data Warehouse, which held evaluation data linked to study identification numbers; and (3) Multiple Pseudonym Assignment Unit, which maintained the assignment list linking personal and study identification numbers. These units were strictly separated both digitally and in terms of personnel, ensuring confidentiality

XSL•FO
RenderX

and preventing reidentification. Participants received no financial incentives.

No images in the article or supplementary materials allow for the identification of individual participants. Any visible human faces have been deliberately blurred to ensure anonymity, or the faces originate from commercially available stock images and do not depict study participants. No consent from identifiable individuals was therefore required.

Results

General Information

A total of 1227 patients were enrolled in the RCT, with 566 patients allocated to the intervention group and 661 patients allocated to the control group. Ultimately, 1002 patients were included in the RCT analyses (intervention group: n=462; control group: n=540; Figure 2). The analyses presented here are restricted to the 462 participants allocated to the intervention group with available baseline questionnaire data and register data at the start of the intervention. The flow diagram in Figure 2 follows the structure recommended by CONSORT (Consolidated Standards of Reporting Trials) 2025 [61]; however, only the elements relevant to this secondary analysis are addressed in the flow diagram. The CONSORT checklist is presented in Checklist 1. Follow-up assessments and analyses of primary and secondary outcomes were conducted as part of the trial but are not relevant to the analyses reported here and will be presented in the effectiveness evaluation in a separate publication.

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram. DHI: digital health intervention; DMP: disease management program; RCT: randomized controlled trial.

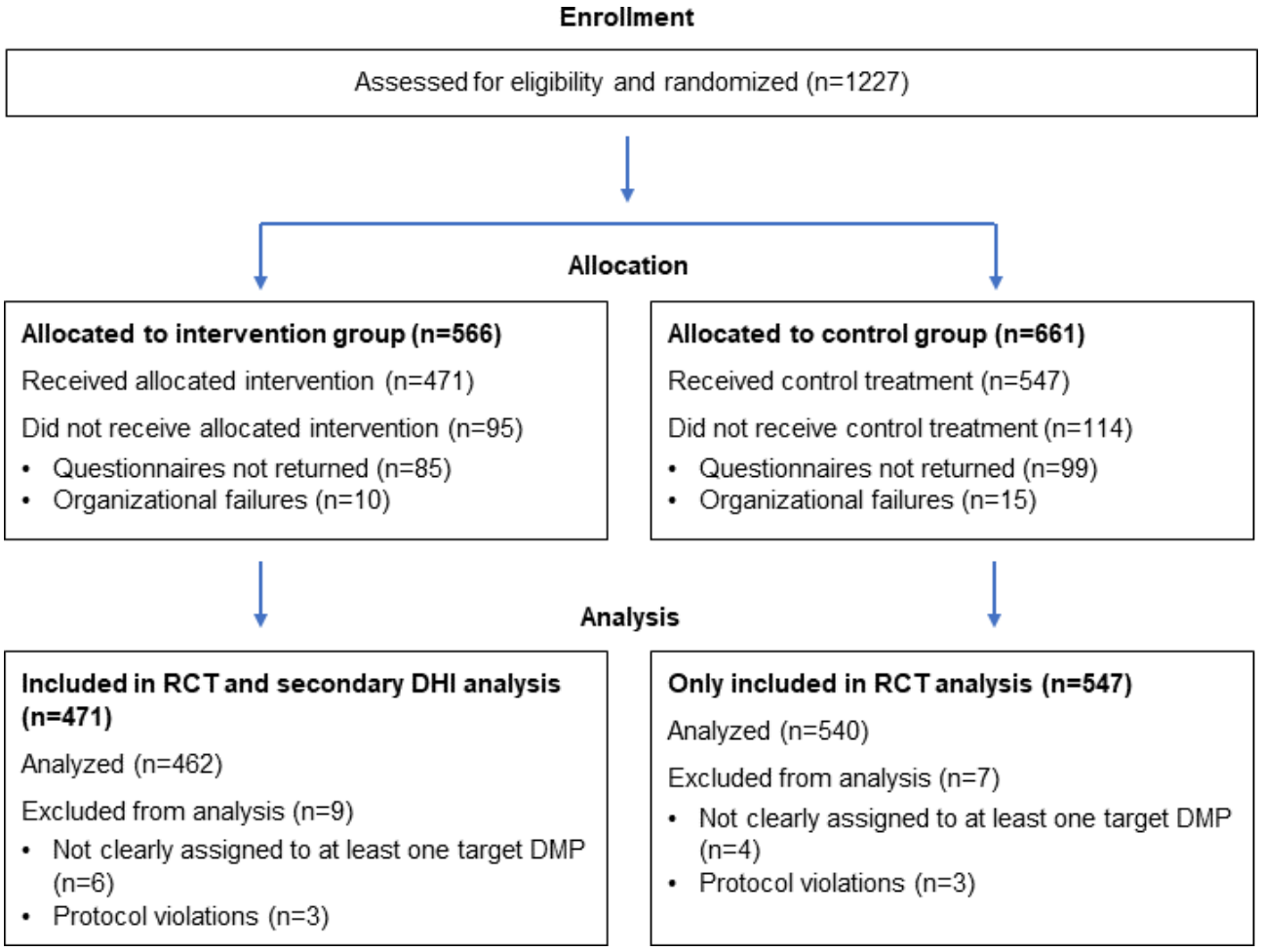

Among the data included in the secondary analysis, an average of 2.2% (range 0.7%-4.8%) of data were missing. MCAR was not fulfilled (P=.008). The mean participant age at the time of the survey was 64.8 (SD 10.2) years. Moreover, 49.8% (230/462) of participants were female, 38.6% (170/462) had a university entrance qualification, and 43.1% (199/462) had at least one log-in to the web portal (Table 1). Participants in the last group were classified as users.

XSL•FO
RenderX

Table . Patient characteristics[a].

Variable	Nonusers		Users		Total	
	n[b]	Statistics	n[b]	Statistics	n[b]	Statistics
Sociodemographic variable						
Age (years), mean (SD)	263	65.4 (11.0)	199	64.0 (9.2)	462	64.8 (10.2)
Female gender[c], n (%)	263	125 (47.5)	199	92 (46.2)	462	230 (49.8)
Higher education, n (%)	248	81 (32.7)	192	89 (46.4)	440	170 (38.6)
Psychological variables[d]						
General psychological variables, mean (SD)						
Extraversion	254	0.60 (0.24)	195	0.60 (0.26)	449	0.60 (0.25)
Agreeableness	258	0.56 (0.21)	195	0.56 (0.19)	453	0.56 (0.20)
Openness	259	0.55 (0.26)	195	0.62 (0.24)	454	0.58 (0.25)
Conscientiousness	254	0.70 (0.21)	195	0.69 (0.20)	449	0.70 (0.21)
Neuroticism	261	0.49 (0.23)	195	0.48 (0.23)	456	0.49 (0.23)
Loneliness	254	0.32 (0.28)	194	0.35 (0.27)	448	0.33 (0.28)
Depression	255	0.24 (0.23)	198	0.24 (0.24)	453	0.24 (0.24)
Anxiety	255	0.23 (0.25)	198	0.22 (0.23)	453	0.23 (0.24)
Health-related psychological variables, mean (SD)						
Patient activation	254	0.67 (0.16)	197	0.65 (0.13)	451	0.66 (0.15)
Health literacy	250	0.70 (0.21)	196	0.73 (0.18)	446	0.71 (0.20)
Intention–physical activity	257	0.79 (0.27)	198	0.89 (0.16)	455	0.84 (0.23)
Intention–healthy nutrition	261	0.78 (0.25)	198	0.88 (0.15)	459	0.82 (0.22)

[a]Patients are members of the intervention group of the randomized controlled trial conducted within the P-SUP (Personalized Self-Management Support Program) project. They were enrolled in disease management programs for type 2 diabetes and/or coronary heart disease and were residing in North Rhine-Westphalia, Germany. Data were collected during the baseline assessment of the P-SUP project.

[b]Number of persons with complete data for the respective variable.

[c]None of the participants reported their gender as diverse.

[d]For better comparability, all psychological variables were scaled to a range from 0 to 1. Higher values indicate a higher expression of the construct.

Among users, 86.4% (172/199) had their first log-in within the first 6 weeks of the complete intervention period. The final first log-in was recorded in week 39. The total number of log-ins was 637 within these first 6 weeks. In the following 6-week intervals, both the number of persons using the web portal and the total number of log-ins decreased almost continuously (Figure 3). With the first log-in included, participants classified as users visited the web portal on average 7.4 (SD 8.5; range 1 - 50) times, if log-ins were determined using an interval of 120 minutes between 2 subsequent activities on the web portal. For intervals of 60 and 30 minutes, the corresponding values were 7.7 (SD 8.8; range 1 - 52) times and 8.2 (SD 9.5; range 1 - 59) times, respectively.

XSL•FO
RenderX

Figure 3. User activity.

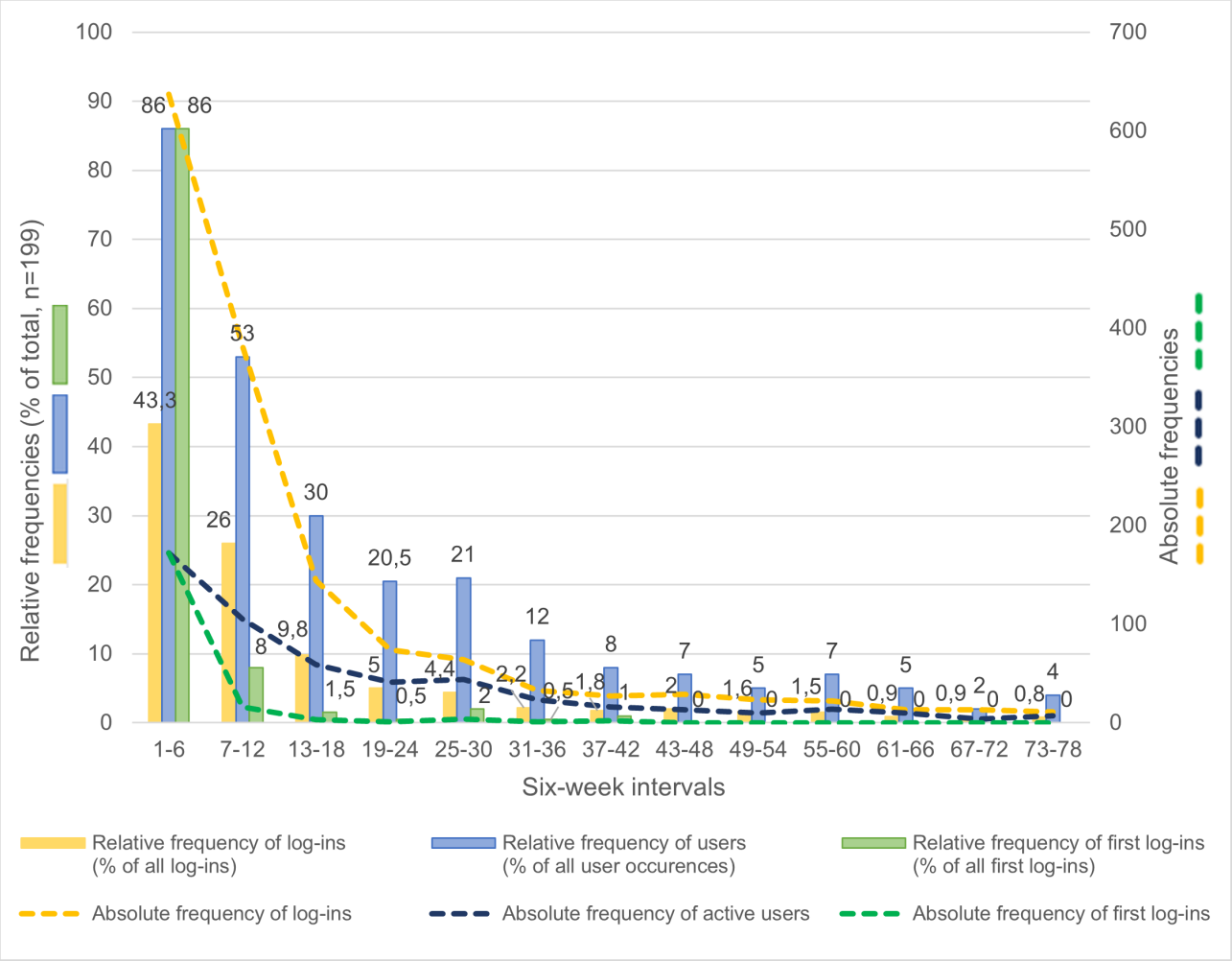

Uptake

Many of the variables taken as possible determinants for uptake correlated moderately with each other. This applied especially to loneliness, depression, and anxiety, as well as to intentions regarding physical activity and healthy nutrition (Multimedia Appendix 4). Nevertheless, the largest inflation factor was smaller than 3 (Table 2). Accordingly, all possible determinants could be included in the multiple regression.

XSL•FO
RenderX

Table . Simple and multiple regression analyses for predictors of uptake[a].

Variable	Simple regression analysis[b]			Multiple regression analysis[b]		
	B[c] (95% CI)	*P* value	NK[d]	B[c] (95% CI)	*P* value	VIF[e]
Sociodemographic variables						
Age in years	−0.01 (−0.03 to 0.00)	.12	0.006	−0.00 (−0.02 to 0.02)	.90	1.14
Female gender	−0.25 (−0.62 to 0.12)	.18	0.005	−0.30 (−0.73 to 0.12)	.16	1.15
Higher education	0.56 (0.18 to 0.95)	.004[f]	0.024	0.29 (−0.15 to 0.73)	.20	1.14
Psychological variables[g]						
General psychological variables						
Extraversion	−0.05 (−0.81 to 0.70)	.89	0.000	−0.08 (−1.00 to 0.84)	.87	1.27
Agreeableness	−0.18 (−1.09 to 0.73)	.70	0.000	0.02 (−1.01 to 1.06)	.97	1.09
Openness	1.08 (0.33 to 1.83)	.005[f]	0.024	0.81 (−0.06 to 1.67)	.07	1.18
Conscientiousness	−0.31 (−1.21 to 0.58)	.49	0.001	−0.03 (−1.11 to 1.04)	.95	1.23
Neuroticism	−0.14 (−0.94 to 0.67)	.74	0.000	−0.05 (−1.06 to 0.96)	.92	1.35
Loneliness	0.34 (−0.33 to 1.01)	.32	0.003	0.54 (−0.43 to 1.51)	.27	1.76
Depression	0.07 (−0.72 to 0.85)	.87	0.000	0.08 (−1.33 to 1.48)	.92	2.64
Anxiety	−0.08 (−0.85 to 0.68)	.83	0.000	−0.38 (−1.83 to 1.07)	.61	2.96
Health-related psychological variables						
Patient activation	−1.17 (−2.48 to 0.14)	.08	0.010	−3.20 (−4.95 to −1.46)	<.001[f]	1.36
Health literacy	1.05 (0.08 to 2.03)	.03[f]	0.014	1.01 (−0.26 to 2.29)	.12	1.43
Intention–physical activity	2.28 (1.30 to 3.26)	<.001[f]	0.071	1.86 (0.74 to 2.97)	.001[f]	1.23
Intention–healthy nutrition	2.30 (1.30 to 3.31)	<.001[f]	0.067	2.22 (1.00 to 3.44)	<.001[f]	1.34

[a]Patients are members of the intervention group of the randomized controlled trial conducted within the P-SUP (Personalized Self-Management Support Program) project. They were enrolled in disease management programs for type 2 diabetes and/or coronary heart disease and were residing in North Rhine-Westphalia, Germany. Data were collected during the baseline assessment of the P-SUP project.

[b]Simple and multiple regression analyses performed on data after multiple imputation (n=462).

[c]B: regression coefficient B.

[d]NK: Nagelkerke pseudo R^2.

[e]VIF: variance inflation factor.

[f]Significant value.

[g]For better comparability, all psychological variables were scaled to a range from 0 to 1. Higher values indicate a higher expression of the construct.

In the simple regression analyses, education, openness, health literacy, intention regarding physical activity, and intention regarding healthy nutrition had statistically significant effects on uptake (education: *P*=.004; openness: *P*=.005; health literacy: *P*=.03; intention-physical activity: *P*<.001; intention-healthy nutrition: *P*<.001; Table 2). In all 5 cases, these effects were positive, with the effects for the 2 intention variables being the largest. After controlling for multiple testing according to Benjamini and Hochberg [60], the effect for health literacy disappeared, but the effects for the other 4 variables with significant effects in the individual tests remained. In the multiple regression analysis, the pattern of statistically significant effects changed. Education (*P*=.20), openness (*P*=.07), and health literacy (*P*=.12) no longer had statistically significant effects. Instead, patient activation had a strong statistically significant negative effect on uptake (*P*<.001), that

XSL•FO
RenderX

is, persons with high patient activation were less likely to use the web portal. The intentions regarding physical activity and healthy nutrition remained statistically significant positive predictors (intention-physical activity: P=.001; intention-healthy nutrition: P<.001). However, the positive effects of these 2 variables were smaller than the negative effect of patient activation. The complete model used in the multiple regression analysis had high statistical significance (P<.001; Table 2). This indicates that the statistically significant effects found for 3 of the 15 variables cannot simply be explained by a higher chance of α errors due to multiple testing. Nagelkerke pseudo R^2 for the complete model was 0.18.

Frequency

If the correlations between the possible determinants are only computed for participants who used the web portal at least once, the pattern of correlations would be very similar to the pattern of correlations for all investigated persons. Many of the variables analyzed were moderately correlated with each other. Again, the highest correlations were among loneliness, depression, and anxiety. However, the correlation between the intentions regarding physical activity and healthy nutrition decreased (Multimedia Appendix 5). Notwithstanding the moderate correlations, the largest inflation factor was smaller than 3, similar to the finding in the analyses for uptake (Table 3). Accordingly, all possible determinants could be included in the multiple regression.

XSL•FO
RenderX

Table . Simple and multiple regression analyses for predictors of frequency[a].

Variable	Simple regression analysis[b]				Multiple regression analysis[b]		
	B[c] (95% CI)	*P* value	DP[d]	PEV[e]	B[c] (95% CI)	*P* value	VIF[f]
Sociodemographic variables							
Age in years	0.03 (0.01 to 0.05)	.005[g]	1.66	0.024	0.02 (0.00 to 0.04)	.03[g]	1.19
Female gender	0.12 (−0.25 to 0.49)	.53	1.72	0.002	0.25 (−0.09 to 0.59)	.15	1.25
Higher education	0.01 (−0.36 to 0.39)	.94	1.73	0.000	−0.02 (−0.36 to 0.33)	.93	1.15
Psychological variables[h]							
General psychological variables							
Extraversion	−0.41 (−1.16 to 0.33)	.28	1.71	0.004	−0.63 (−1.37 to 0.12)	.10	1.43
Agreeableness	−0.42 (−1.35 to 0.51)	.37	1.72	0.004	−0.65 (−1.43 to 0.14)	.11	1.21
Openness	−0.14 (−0.88 to 0.60)	.72	1.72	0.000	0.04 (−0.64 to 0.72)	.91	1.15
Conscientiousness	1.09 (0.25 to 1.93)	.01[g]	1.66	0.039	1.00 (0.16 to 1.84)	.02[g]	1.34
Neuroticism	0.37 (−0.40 to 1.14)	.35	1.72	0.003	0.50 (−0.32 to 1.33)	.23	1.53
Loneliness	−0.62 (−1.26 to 0.01)	.06	1.69	0.020	−0.64 (−1.42 to 0.15)	.11	1.94
Depression	−0.89 (−1.58 to −0.19)	.01[g]	1.68	0.015	−0.74 (−1.64 to 0.17)	.11	2.58
Anxiety	−0.24 (−1.05 to 0.58)	.57	1.72	0.001	0.60 (−0.52 to 1.72)	.29	2.88
Health-related psychological variables							
Patient activation	1.25 (−0.12 to 2.62)	.07	1.69	0.018	0.52 (−0.95 to 1.98)	.49	1.57
Health literacy	0.35 (−0.78 to 1.48)	.55	1.72	0.002	0.09 (−1.08 to 1.25)	.89	1.56
Intention–physical activity	−0.19 (−1.12 to 0.75)	.69	1.72	0.001	−0.63 (−1.49 to 0.22)	.15	1.17
Intention–healthy nutrition	0.49 (−0.82 to 1.79)	.47	1.72	0.002	0.48 (−0.61 to 1.57)	.39	1.19

[a]Patients are members of the intervention group of the randomized controlled trial conducted within the P-SUP (Personalized Self-Management Support Program) project. They were enrolled in disease management programs for type 2 diabetes and/or coronary heart disease and were residing in North Rhine-Westphalia, Germany. Data were collected during the baseline assessment of the P-SUP project.

[b]Simple and multiple regression analyses performed on data after multiple imputation (n=199).

[c]B: regression coefficient B.

[d]DP: dispersion parameter.

[e]PEV: proportion of explained variance.

[f]VIF: variance inflation factor.

[g]Significant value.

[h]For better comparability, all psychological variables were scaled to a range from 0 to 1. Higher values indicate a higher expression of the construct.

The results of the regression analyses were virtually the same for all 3 intervals applied for identifying new log-ins. Therefore, only the results for the 120-minute interval have been reported. These results differed strongly from those for uptake (Table 3).

In the simple regression analyses, only 3 of the 15 variables had statistically significant effects (ie, age: *P*=.005; conscientiousness: *P*=.01; and depression: *P*=.01). The first 2 variables related positively to frequency, whereas there was a

XSL•FO
RenderX

negative effect for depression. However, after controlling for multiple testing according to Benjamini and Hochberg [60], all 3 effects disappeared. In the multiple regression analysis, there were effects for age (P=.03) and conscientiousness (P=.02), whereas the effect for depression was no longer observed (P=.11; Table 3). However, in contrast to the complete multiple regression model for predicting uptake, the complete multiple regression model for predicting frequency was not statistically significant (P=.07). Accordingly, whether age and conscientiousness actually affect frequency is questionable. The dispersion parameter for the complete model was 1.48, and the proportion of explained variance was 0.16.

Comparison Between Uptake and Frequency

The comparison of the multiple regression models for uptake and frequency showed that the rank order of the P values differed markedly between uptake and frequency. The correlation between the rank orders was −0.50, with a statistically significant deviation from 0 (P=.009).

Discussion

Principal Findings

The main finding of this study is that uptake and frequency of use are influenced in different ways by the variables considered as potential determinants. In simple regression analyses, uptake of the web portal was positively associated with higher education, openness, health literacy, and intentions regarding physical activity and healthy nutrition; however, the effect for health literacy did not remain significant after controlling for multiple testing. In the multiple regression model, uptake was primarily explained by strong intentions regarding physical activity and healthy nutrition, alongside a significant negative association with patient activation. In contrast, frequency of use showed a different pattern. In simple regression analyses, frequency was positively associated with age and conscientiousness and negatively associated with depression; however, in the multiple regression model, only age and conscientiousness remained associated. After controlling for inflation of the α error, none of the associations for frequency of use remained statistically significant, and the multiple regression model for frequency was not significant.

The observation that the multivariate regression model has a strong effect for uptake but no statistically significant effect for frequency may be partly attributable to the smaller sample size available for the frequency analyses. As higher values of Nagelkerke pseudo R^2 are more difficult to obtain than higher values of proportions of explained variance for continuous variables [62], the comparison of these 2 statistics suggests that the set of variables investigated in this study has a stronger impact on uptake than on frequency of use. A further difference between the 2 analyses concerns the rank order of the P values for the 2 multiple regressions, which deviated markedly between uptake and frequency. This strongly indicates that the considered variables affect uptake and frequency in very different ways. This demonstrates that both components of usage are distinct and should be investigated and interpreted separately from each other.

Uptake

In this study, intentions had a very strong positive association with uptake. This finding is supported by research on behavior change, which emphasizes the central role of intention in bridging the gap between attitudes and actual behavior [63]. According to the Theory of Planned Behavior [64], intention is viewed as the proximal determinant of behavior. According to this, people are more likely to perform a behavior if they strongly intend to do so. In a previous study that most closely resembles this study, Voncken-Brewster et al [26] examined a web-based self-management intervention for individuals with or at high risk for chronic obstructive pulmonary disease (COPD). In a common multiple regression model, they considered age, gender, education, and intention to change behavior alongside clinical factors such as COPD status, smoking status, physical activity level, and dyspnea status and found no statistically significant effects. In contrast, this study identified a strong positive association between intention and uptake. The difference between the 2 results might be explained by the fact that both studies refer to different kinds of intentions. In the study by Voncken-Brewster et al [26], intention referred to a change in a specific behavior, while in this study, intention referred to the performance of a specific behavior. The latter seems to be a strong motivating factor for uptake, whereas the former does not show the same effect.

In this study, patient activation showed a strong negative association with uptake. This contrasts with previous findings, where either a positive association [37] or no effect was reported [36]. A central assumption is that higher patient activation is associated with greater engagement in health-promoting behaviors [65]. However, DHIs are not necessarily health-promoting, and the DHI investigated in this study may provide little additional value for individuals who already engage in the targeted behaviors. Highly activated individuals may therefore be less inclined to use such interventions. In contrast, less activated individuals in this study may have been more likely to perceive the DHI as beneficial, given that individuals with lower activation levels are more likely to experience poorer clinical outcomes and higher rates of hospitalization [65], and thus have greater potential to benefit from additional self-management support. When initial use does not require immediate behavioral change, such tools may represent an accessible entry point into health-related activation. This suggests that the relationship between patient activation and DHI uptake may depend on perceived relevance and expected benefit. Differences in the added value of the DHIs investigated across studies may therefore explain the divergent findings. Whether this mechanism accounts for the negative association observed in this study remains a question for future research.

In this study, the effect of openness was only seen in the simple analysis and not in the multiple regression analysis. This finding might be unexpected, as individuals with high openness are generally more willing to explore unfamiliar tools and information sources [66]. Moreover, in a previous study by Su et al [29], where a smartphone-based app for improving self-management in people with T2DM was examined, openness had a statistically significant positive effect on uptake. One

XSL•FO
RenderX

possible explanation for this discrepancy is that Su et al [29] did not include an intention variable in their multiple regression analysis. Given the strong and proximal role of intention in predicting behavior, intention may represent a more proximal and stronger determinant of uptake, thereby attenuating the independent effect of more distal personality traits, such as openness, when both variables are considered simultaneously. This interpretation is further supported by evidence suggesting that openness primarily influences behavior by moderating the intention-behavior relationship rather than exerting a strong independent effect. Openness has been shown to enhance the predictive strength of intention for behavior, while its direct association diminishes once intention is accounted for [67].

Taken together, the findings for uptake suggest that individuals who either have low intention to engage in health-promoting behaviors or consider their current levels of physical activity and healthy nutrition to be sufficient may not be reached by such interventions. To increase uptake among individuals with low intention with regard to the targeted behaviors, this intention may first need to be developed. This could be achieved through targeted outreach strategies, motivational inputs that emphasize the relevance and potential benefits of the intervention, or integration into routine care settings such as general practice consultations [68]. In particular, endorsement by trusted health care professionals may help legitimize the intervention and lower initial barriers to engagement [69]. Highly activated individuals might benefit from DHIs that go beyond structured guidance and motivational support by providing tools that support the optimization of routines, monitoring of progress, and maintenance of long-term health-promoting behaviors. In contrast, less activated patients may benefit from low-threshold DHIs that provide consolidated, easily accessible information and basic guidance. Such interventions may reduce initial barriers to engagement by requiring little effort or commitment at the outset and may serve as an accessible starting point for health-related activation.

Frequency

Without controlling for multiple testing, both simple and multiple regression analyses showed statistically significant positive effects of age and conscientiousness on frequency of use. After controlling for multiple testing, none of these effects remained statistically significant in any of the regression analyses. The following considerations regarding age and conscientiousness are therefore exploratory and are intended to contextualize observed tendencies only.

The tendency regarding age found in this study is consistent with findings reported by Goyal et al [24], who examined a smartphone-based app for the assessment and prevention of heart disease. In that study, a core feature of the app was the proposal of health-related activities, referred to as challenges, and the frequency of use was operationalized as the number of challenges completed within the first 30 days of app use. An explanation of this effect or tendency might be that health-related goals tend to increase over a person's life course [70,71] because health becomes a more central concern due to increasing functional limitations and health issues [72]. Furthermore, older individuals may particularly appreciate DHIs that bundle relevant information in an easily accessible and user-friendly platform and allow self-paced engagement [73]. The lower uptake tendency among younger individuals may reflect different usage preferences rather than a lack of interest in health topics. Younger individuals may be more likely to access health-related information through multiple digital channels and may prefer more flexible or interactive formats due to their higher digital health literacy [74]. Further studies should investigate whether different types of DHIs are preferentially used by different age groups, which could inform more age-sensitive design and implementation strategies.

There was no gender effect in this study. This contrasts with findings from previous studies showing that women are generally more likely than men to engage in health-promoting behavior [75-77] and that gender influences DHI use [24,78]. However, the direction of this gender effect was not specified in the study by Goyal et al [24], whereas Rising et al [78] reported higher DHI use among women. One explanation for the absence of a gender effect in this study is that prior studies used substantially larger sample sizes (n=69,952 and n=6789), which increased the likelihood of detecting gender differences. A further reason might be the measurement of usage. Goyal et al [24] assessed usage based on the completion of real-life health challenges, such as engaging in physical activity or preparing healthy meals, while this study focused on direct interaction with the DHI. Rising et al [78] relied on self-reported patterns of digital health use derived from questionnaire data, distinguishing users based on ownership and subjective use of smartphones, apps, and tracking devices rather than on actual interaction with a specific DHI. These differences in how usage is operationalized may lead to divergent findings regarding gender effects.

As no statistically significant effects on frequency of use were observed after controlling for multiple testing, no practical implications were derived from these findings.

Limitations

The study has 5 main limitations regarding the interpretability of the results. First, the selection of potential determinants was limited to variables available in the register and survey data. As the survey did not assess participants' access to digital devices or their level of digital health literacy [79], the results will presumably not cover all possible determinants of usage. Second, the operationalization of the frequency of use was a limitation. The participants had the option to download the exercise videos provided within the DHI. Since neither the number of downloads nor the frequency of offline usage could be tracked, actual usage may have been underestimated, potentially affecting the results. Third, the study was embedded in a complex, multicomponent intervention. The other components may have interacted with the effect of the web portal on DHI usage. Fourth, just as in all other studies addressing the relationship between person-related variables and DHI use, the variables examined might correlate with other variables that are not examined, and these variables, in turn, might be the actual determinants of DHI usage. Fifth, the investigations presented here refer to a very specific DHI and target group. All participants were enrolled in a German DMP

XSL•FO
RenderX

and recruited in a specific region in Germany, where digital infrastructure, care pathways, and culture may differ from those in other regions or countries. Thus, usage patterns and determinants of usage may not be generalizable to other DHIs, other persons, or other settings.

Conclusion

This study differs from previous studies in jointly examining the determinants of uptake and frequency of use for the same DHI within the same context and sample. It contributes to a more differentiated understanding of DHI use by demonstrating that uptake and frequency of use are influenced by distinct patterns of determinants. Uptake was primarily driven by strong behavioral intentions, whereas associations with frequency of use were less robust after adjustment for multiple testing. These findings suggest that initial adoption and sustained engagement represent related but analytically and practically distinct processes. Consequently, strategies aimed at promoting uptake may need to focus on strengthening health-related intentions, while efforts to foster continued use may require different, potentially more context-sensitive approaches. Applying the approach of this study to other DHIs and settings may support more targeted, effective, and equitable digital health development and implementation.

Acknowledgments

We would like to thank August-Wilhelm Bödecker, Cosima Bootz, Suzan Botzenhardt, Iris Dapper, Katja Dehnen, Anja Engels, Christian Funke, Ekaterini Giagkou, Anette Graute, Larisa Pilic, Jasmin Marikar, Mareike Michely, Inés van der Arend, and Birgitta Weltermann for their cooperation in recruiting study participants; Matthias Funken and Michael Frank for coordinating work; and Pete Bereza for his linguistic advice.

The authors declare the use of generative artificial intelligence (AI) in the research and writing process. According to the GAIDeT taxonomy (2025) [80], the following tasks were delegated to generative AI tools under full human supervision: proofreading, editing, and translation. The generative AI tool used was ChatGPT 5.2 (OpenAI). Responsibility for the final manuscript lies entirely with the authors. Generative AI tools are not listed as authors and do not bear responsibility for the final outcomes. The declaration has been submitted by the first author (MS).

Funding

The study was funded by the Innovation Fund of the Federal Joint Committee (funding number: 01NVF18033).

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the publication agreement between the project consortium partners but are available from the corresponding author upon reasonable request.

Authors' Contributions

MS contributed to developing the investigated digital health intervention (DHI), developed essential parts of the conception, wrote essential parts of the manuscript, performed some of the statistical analyses, and did the literature research. UK contributed essentially to developing the conception and to writing the manuscript, advised MS in the statistical analyses, and performed some of the statistical analyses. CL consolidated the different components of the DHI and integrated them into the final DHI. SA advised on the statistical analyses. MR advised on the conception. KB-F advised on the Patient Activation Measure. BB, CC, AG, LG, and LS contributed to developing the DHI. FF, MH, LK, LP, FV, and SW contributed to recruiting the study participants.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Web portal design and content.
[DOC File, 3458 KB - jmir_v28i1e80895_app1.doc]

Multimedia Appendix 2
Nutritional behavior questionnaire
[DOCX File, 15 KB - jmir_v28i1e80895_app2.docx]

Multimedia Appendix 3
Rasch scaling.
[DOC File, 49 KB - jmir_v28i1e80895_app3.doc]

Multimedia Appendix 4

XSL•FO
RenderX

Correlation analysis for uptake.
[DOC File, 112 KB - jmir_v28i1e80895_app4.doc]

Multimedia Appendix 5
Correlation analysis for frequency.
[DOC File, 112 KB - jmir_v28i1e80895_app5.doc]

Checklist 1
CONSORT checklist
[PDF File, 141 KB - jmir_v28i1e80895_app6.pdf]

References

1. Hajat C, Stein E. The global burden of multiple chronic conditions: a narrative review. Prev Med Rep 2018 Dec;12:284-293. [doi: 10.1016/j.pmedr.2018.10.008] [Medline: 30406006]
2. Joseph P, Leong D, McKee M, et al. Reducing the global burden of cardiovascular disease, part 1: the epidemiology and risk factors. Circ Res 2017 Sep 1;121(6):677-694. [doi: 10.1161/CIRCRESAHA.117.308903] [Medline: 28860318]
3. Bekele BB, Negash S, Bogale B, et al. Effect of diabetes self-management education (DSME) on glycated hemoglobin (HbA1c) level among patients with T2DM: systematic review and meta-analysis of randomized controlled trials. Diabetes Metab Syndr 2021;15(1):177-185. [doi: 10.1016/j.dsx.2020.12.030] [Medline: 33360516]
4. Reynolds R, Dennis S, Hasan I, et al. A systematic review of chronic disease management interventions in primary care. BMC Fam Pract 2018 Jan 9;19(1):11. [doi: 10.1186/s12875-017-0692-3] [Medline: 29316889]
5. Scheckel B, Schmidt K, Stock S, Redaèlli M. Patient portals as facilitators of engagement in patients with diabetes and chronic heart disease: scoping review of usage and usability. J Med Internet Res 2023 Aug 25;25:e38447. [doi: 10.2196/38447] [Medline: 37624629]
6. Cheah KJ, Abdul Manaf Z, Fitri Mat Ludin A, Razalli NH, Mohd Mokhtar N, Md Ali SH. Mobile apps for common noncommunicable disease management: systematic search in app stores and evaluation using the mobile app rating scale. JMIR Mhealth Uhealth 2024 Mar 12;12(1):e49055. [doi: 10.2196/49055] [Medline: 38532298]
7. Lavoie A, Dubé V. Web-based interventions to promote healthy lifestyles for older adults: scoping review. Interact J Med Res 2022 Aug 23;11(2):e37315. [doi: 10.2196/37315] [Medline: 35998024]
8. Gentili A, Failla G, Melnyk A, et al. The cost-effectiveness of digital health interventions: a systematic review of the literature. Front Public Health 2022;10:787135. [doi: 10.3389/fpubh.2022.787135] [Medline: 36033812]
9. Kvedar J, Coye MJ, Everett W. Connected health: a review of technologies and strategies to improve patient care with telemedicine and telehealth. Health Aff (Millwood) 2014 Feb;33(2):194-199. [doi: 10.1377/hlthaff.2013.0992] [Medline: 24493760]
10. Rawstorn JC, Gant N, Direito A, Beckmann C, Maddison R. Telehealth exercise-based cardiac rehabilitation: a systematic review and meta-analysis. Heart 2016 Aug 1;102(15):1183-1192. [doi: 10.1136/heartjnl-2015-308966] [Medline: 26936337]
11. Mistry N, Keepanasseril A, Wilczynski NL, et al. Technology-mediated interventions for enhancing medication adherence. J Am Med Inform Assoc 2015 Apr;22(e1):e177-e193. [doi: 10.1093/jamia/ocu047] [Medline: 25726568]
12. Captieux M, Pearce G, Parke HL, et al. Supported self-management for people with type 2 diabetes: a meta-review of quantitative systematic reviews. BMJ Open 2018 Dec 14;8(12):e024262. [doi: 10.1136/bmjopen-2018-024262] [Medline: 30552277]
13. Greenwood DA, Gee PM, Fatkin KJ, Peeples M. A systematic review of reviews evaluating technology-enabled diabetes self-management education and support. J Diabetes Sci Technol 2017 Sep;11(5):1015-1027. [doi: 10.1177/1932296817713506] [Medline: 28560898]
14. Palacios J, Lee GA, Duaso M, et al. Internet-delivered self-management support for improving coronary heart disease and self-management-related outcomes: a systematic review. J Cardiovasc Nurs 2017;32(4):E9-E23. [doi: 10.1097/JCN.0000000000000392] [Medline: 28107251]
15. Su D, Zhou J, Kelley MS, et al. Does telemedicine improve treatment outcomes for diabetes? A meta-analysis of results from 55 randomized controlled trials. Diabetes Res Clin Pract 2016 Jun;116:136-148. [doi: 10.1016/j.diabres.2016.04.019] [Medline: 27321329]
16. Flores Mateo G, Granado-Font E, Ferré-Grau C, Montaña-Carreras X. Mobile phone apps to promote weight loss and increase physical activity: a systematic review and meta-analysis. J Med Internet Res 2015 Nov 10;17(11):e253. [doi: 10.2196/jmir.4836] [Medline: 26554314]
17. Middelweerd A, Mollee JS, van der Wal CN, Brug J, Te Velde SJ. Apps to promote physical activity among adults: a review and content analysis. Int J Behav Nutr Phys Act 2014 Jul 25;11:97. [doi: 10.1186/s12966-014-0097-9] [Medline: 25059981]

XSL•FO
RenderX

18. Hou C, Carter B, Hewitt J, Francisa T, Mayor S. Do mobile phone applications improve glycemic control (HbA1c) in the self-management of diabetes? A systematic review, meta-analysis, and GRADE of 14 randomized trials. Diabetes Care 2016 Nov;39(11):2089-2095. [doi: 10.2337/dc16-0346] [Medline: 27926892]
19. Pal K, Eastwood SV, Michie S, et al. Computer-based diabetes self-management interventions for adults with type 2 diabetes mellitus. Cochrane Database Syst Rev 2013 Mar 28;2013(3):CD008776. [doi: 10.1002/14651858.CD008776.pub2] [Medline: 23543567]
20. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010 Feb 17;12(1):e4. [doi: 10.2196/jmir.1376] [Medline: 20164043]
21. Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N. A systematic review of the impact of adherence on the effectiveness of e-therapies. J Med Internet Res 2011 Aug 5;13(3):e52. [doi: 10.2196/jmir.1772] [Medline: 21821503]
22. Schoeppe S, Alley S, Van Lippevelde W, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2016 Dec 7;13(1):127. [doi: 10.1186/s12966-016-0454-y] [Medline: 27927218]
23. Böhm AK, Jensen ML, Sørensen MR, Stargardt T. Real-world evidence of user engagement with mobile health for diabetes management: longitudinal observational study. JMIR Mhealth Uhealth 2020 Nov 6;8(11):e22212. [doi: 10.2196/22212] [Medline: 32975198]
24. Goyal S, Morita PP, Picton P, Seto E, Zbib A, Cafazzo JA. Uptake of a consumer-focused mHealth application for the assessment and prevention of heart disease: the <30 days study. JMIR Mhealth Uhealth 2016 Mar 24;4(1):e32. [doi: 10.2196/mhealth.4730] [Medline: 27012937]
25. Poduval S, Ahmed S, Marston L, Hamilton F, Murray E. Crossing the digital divide in online self-management support: analysis of usage data from HeLP-Diabetes. JMIR Diabetes 2018 Dec 6;3(4):e10925. [doi: 10.2196/10925] [Medline: 30522988]
26. Voncken-Brewster V, Amoureus M, de Vries H, et al. The impact of participant characteristics on use and satisfaction of a web-based computer-tailored chronic obstructive pulmonary disease self-management intervention: a process evaluation. JMIR Form Res 2017 Aug 21;1(1):e1. [doi: 10.2196//formative.6585] [Medline: 30684399]
27. Signal V, McLeod M, Stanley J, et al. A mobile- and web-based health intervention program for diabetes and prediabetes self-management (BetaMe/Melon): process evaluation following a randomized controlled trial. J Med Internet Res 2020 Dec 1;22(12):e19150. [doi: 10.2196/19150] [Medline: 33258776]
28. Coorey G, Peiris D, Scaria A, et al. An internet-based intervention for cardiovascular disease management integrated with primary care electronic health records: mixed methods evaluation of implementation fidelity and user engagement. J Med Internet Res 2021 Apr 26;23(4):e25333. [doi: 10.2196/25333] [Medline: 33900204]
29. Su J, Dugas M, Guo X, Gao GG. Influence of personality on mHealth use in patients with diabetes: prospective pilot study. JMIR Mhealth Uhealth 2020 Aug 10;8(8):e17709. [doi: 10.2196/17709] [Medline: 32773382]
30. Ruotsalainen I, Valtanen M, Kärsämä R, et al. Influencing factors in digital health intervention uptake: the interplay of education, lifestyle, and digital literacy. In: Särestöniemi M, Keikhosrokiani P, Singh D, Harjula E, Tiulpin A, Jansson M, editors. Digital Health and Wireless Solutions NCDHWS 2024 Communications in Computer and Information Science, Vol 2084: Springer; 2024:322-332. [doi: 10.1007/978-3-031-59091-7_21]
31. Yan M, Sun W, Tan C, Wu Y, Liu Y. Analysis of factors influencing the willingness of Chinese older adults to use mHealth devices: nationwide cross-sectional survey study. J Med Internet Res 2025 Mar 4;27(1):e66804. [doi: 10.2196/66804] [Medline: 40053781]
32. Glasgow RE, Strycker LA, King DK, Toobert DJ. Understanding who benefits at each step in an internet-based diabetes self-management program: application of a recursive partitioning approach. Med Decis Making 2014 Feb;34(2):180-191. [doi: 10.1177/0272989X13498156] [Medline: 23913917]
33. Brenk-Franz K, Leonhardt LJ, Strauß B. Attachment and patient activation as predictors of the interest and use of telemedical health applications -results of an observational study in primary health care. BMC Prim Care 2022 Apr 29;23(1):96. [doi: 10.1186/s12875-022-01711-0] [Medline: 35488203]
34. Dolničar V, Petrovčič A, Škafar M, Laznik J, Prevodnik K, Hvalič-Touzery S. Determinants of the intention to use mHealth in the future: evidence from an intervention study of patients with chronic diseases in Slovenia. Int J Med Inform 2024 Oct;190:105537. [doi: 10.1016/j.ijmedinf.2024.105537] [Medline: 39002206]
35. Ochnik D, Cholewa-Wiktor M, Jakubiak M, Pataj M. eHealth tools use and mental health: a cross-sectional network analysis in a representative sample. Sci Rep 2024 Mar 2;14(1):5173. [doi: 10.1038/s41598-024-55910-z] [Medline: 38431653]
36. Ancker JS, Osorio SN, Cheriff A, Cole CL, Silver M, Kaushal R. Patient activation and use of an electronic patient portal. Inform Health Soc Care 2015;40(3):254-266. [doi: 10.3109/17538157.2014.908200] [Medline: 24786648]
37. Hibbard JH, Greene J. Who are we reaching through the patient portal: engaging the already engaged? Int J Pers Cent Med 2011;1(4):788-793 [FREE Full text]
38. Short CE, DeSmet A, Woods C, et al. Measuring engagement in eHealth and mHealth behavior change interventions: viewpoint of methodologies. J Med Internet Res 2018 Nov 16;20(11):e292. [doi: 10.2196/jmir.9397] [Medline: 30446482]

XSL•FO
RenderX

39. Engelen M, van Gaal B, Vermeulen H, Zuidema R, Bredie S, van Dulmen S. The influencing contexts and potential mechanisms behind the use of web-based self-management support interventions: realistic evaluation. JMIR Hum Factors 2022 Jul 1;9(3):e34925. [doi: 10.2196/34925] [Medline: 35776437]
40. Konerding U, Redaèlli M, Ackermann K, et al. A pragmatic randomised controlled trial referring to a Personalised Self-management SUPport Programme (P-SUP) for persons enrolled in a disease management programme for type 2 diabetes mellitus and/or for coronary heart disease. Trials 2021 Sep 27;22(1):659. [doi: 10.1186/s13063-021-05636-4] [Medline: 34579783]
41. Gawlik A, Nacak Y, Kleinert J, Konerding U, Vitinius F. Theoretical derivation of a telephone-based health coaching intervention for promoting physical activity and healthy nutrition. Int J Environ Res Public Health 2023 Jun 30;20(13):6271. [doi: 10.3390/ijerph20136271] [Medline: 37444117]
42. Busse R. Disease management programs in Germany's statutory health insurance system. Health Aff (Millwood) 2004;23(3):56-67. [doi: 10.1377/hlthaff.23.3.56] [Medline: 15160803]
43. Ryan RM, Deci EL. Overview of self-determination theory: an organismic-dialectical perspective. In: Handbook of Self-Determination Research: University of Rochester Press; 2002:3-33.
44. Mummah SA, Robinson TN, King AC, Gardner CD, Sutton S. IDEAS (integrate, design, assess, and share): a framework and toolkit of strategies for the development of more effective digital interventions to change health behavior. J Med Internet Res 2016 Dec 16;18(12):e317. [doi: 10.2196/jmir.5927] [Medline: 27986647]
45. Lendt C, Gawlik A, Chermette C, Biallas B, Kleinert J. Entwicklung einer webbasierten intervention zur aktivitätsförderung bei patient:innen mit typ-2 diabetes mellitus und koronarer herzkrankheit [Article in German]. Presented at: 31 Rehabilitationswissenschaftliches Kolloquium: Deutscher Kongress für Rehabilitationsforschung Rehabilitation: Neue Wege, neue Chancen; Mar 7-9, 2022 URL: https://fis.dshs-koeln.de/de/publications/entwicklung-einer-webbasierten-intervention-zur-aktivit%C3%A4tsf%C3%B6rderu/ [accessed 2026-03-08]
46. Reck C, Lendt C, Biallas B. User engagement and usability of a digital behavior change intervention to promote physical activity in older patients with type-2 diabetes and coronary heart disease. Presented at: 26th Annual Congress of the European College of Sport Science; Sep 8-10, 2021 URL: https://fis.dshs-koeln.de/en/publications/user-engagement-and-usability-of-a-digital-behavior-change-interv/ [accessed 2026-03-08]
47. Steitz BD, Cronin RM, Davis SE, Yan E, Jackson GP. Long-term patterns of patient portal use for pediatric patients at an academic medical center. Appl Clin Inform 2017 Aug 2;8(3):779-793. [doi: 10.4338/ACI-2017-01-RA-0005] [Medline: 28765865]
48. Halfaker A, Keyes O, Kluver D, et al. User session identification based on strong regularities in inter-activity time. arXiv. Preprint posted online on Nov 11, 2014. [doi: 10.48550/arXiv.1411.2878]
49. Rammstedt B, Kemper CJ, Klein MC, Beierlein C, Kovaleva A. Big Five Inventory (BFI-10). GESIS – Leibniz Institute for the Social Sciences - Zusammenstellung Sozialwissenschaftlicher Items Und Skalen (ZIS). 2014. URL: https://search.gesis.org/instruments_tools/zis76 [accessed 2026-03-08]
50. Russell D, Peplau LA, Cutrona CE. The revised UCLA loneliness scale: concurrent and discriminant validity evidence. J Pers Soc Psychol 1980 Sep;39(3):472-480. [doi: 10.1037//0022-3514.39.3.472] [Medline: 7431205]
51. Wongpakaran N, Wongpakaran T, Pinyopornpanish M, et al. Development and validation of a 6-item revised UCLA loneliness scale (RULS-6) using Rasch analysis. Br J Health Psychol 2020 May;25(2):233-256. [doi: 10.1111/bjhp.12404] [Medline: 31999891]
52. Löwe B, Wahl I, Rose M, et al. A 4-item measure of depression and anxiety: validation and standardization of the Patient Health Questionnaire-4 (PHQ-4) in the general population. J Affect Disord 2010 Apr;122(1-2):86-95. [doi: 10.1016/j.jad.2009.06.019] [Medline: 19616305]
53. Wallston KA, Cawthon C, McNaughton CD, Rothman RL, Osborn CY, Kripalani S. Psychometric properties of the brief health literacy screen in clinical practice. J Gen Intern Med 2014 Jan;29(1):119-126. [doi: 10.1007/s11606-013-2568-0] [Medline: 23918160]
54. Brenk-Franz K, Hibbard JH, Herrmann WJ, et al. Validation of the German version of the patient activation measure 13 (PAM13-D) in an international multicentre study of primary care patients. PLoS ONE 2013;8(9):e74786. [doi: 10.1371/journal.pone.0074786] [Medline: 24098669]
55. Hibbard JH, Mahoney ER, Stockard J, Tusler M. Development and testing of a short form of the patient activation measure. Health Serv Res 2005 Dec;40(6 Pt 1):1918-1930. [doi: 10.1111/j.1475-6773.2005.00438.x] [Medline: 16336556]
56. Brenk-Franz K, Hibbard J, Tiesler F, Gensichen J. Manual Zur Deutschen Version Des PAM 13 [Book in German]: Institut für Allgemeinmedizin, Friedrich-Schiller-Universität Jena; 2014.
57. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988 Dec;83(404):1198-1202. [doi: 10.1080/01621459.1988.10478722]
58. van Buuren S. Flexible Imputation of Missing Data, 2nd edition: Chapman and Hall/CRC; 2018. [doi: 10.1201/9780429492259]
59. Hilbe JM. Modeling Count Data: Cambridge University Press; 2014. [doi: 10.1017/cbo9781139236065]
60. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 1995 Jan 1;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]

XSL•FO
RenderX

61. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025 Apr 14;389:e081123. [doi: 10.1136/bmj-2024-081123] [Medline: 40228833]
62. Smith TJ, McKenna CM. A comparison of logistic regression pseudo R2 indices. Multiple Linear Regression Viewpoints 2013;39(2):17-26 [FREE Full text]
63. Admassu W, Gorems K. Analyzing health service employees' intention to use e-health systems in southwest Ethiopia: using UTAUT-2 model. BMC Health Serv Res 2024 Sep 27;24(1):1136. [doi: 10.1186/s12913-024-11567-y] [Medline: 39334209]
64. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
65. Janamian T, Greco M, Cosgriff D, Baker L, Dawda P. Activating people to partner in health and self-care: use of the Patient Activation Measure. Med J Aust 2022 Jun 6;216 Suppl 10(Suppl 10):S5-S8. [doi: 10.5694/mja2.51535] [Medline: 35665937]
66. Zhu Y, Jiang H, Zhou Z. Information adoption behavior in online healthcare communities from the perspective of personality traits. Front Psychol 2022;13:973522. [doi: 10.3389/fpsyg.2022.973522] [Medline: 36329756]
67. Pérez-Fuentes MDC, Molero Jurado MDM, Gázquez Linares JJ, Oropesa Ruiz NF, Simón Márquez MDM, Saracostti M. Self-expressive creativity in the adolescent digital domain: personality, self-esteem, and emotions. Int J Environ Res Public Health 2019 Nov 15;16(22):4527. [doi: 10.3390/ijerph16224527] [Medline: 31731778]
68. Lau WY, Lian J, Yap M. The effectiveness of interventions used to improve general health check uptake by the older adult population: a systematic review and meta-analysis. PLOS Glob Public Health 2025;5(3):e0004362. [doi: 10.1371/journal.pgph.0004362] [Medline: 40163543]
69. Catapan SDC, Sazon H, Zheng S, et al. A systematic review of consumers' and healthcare professionals' trust in digital healthcare. NPJ Digit Med 2025 Feb 21;8(1):115. [doi: 10.1038/s41746-025-01510-8] [Medline: 39984678]
70. Heckhausen J. Developmental regulation across adulthood: primary and secondary control of age-related challenges. Dev Psychol 1997;33(1):176-187. [doi: 10.1037/0012-1649.33.1.176]
71. Nurmi JE, Pulliainen H, Salmela-Aro K. Age differences in adults' control beliefs related to life goals and concerns. Psychol Aging 1992 Jun;7(2):194-196. [doi: 10.1037//0882-7974.7.2.194] [Medline: 1610507]
72. Baltes PB, Baltes MM. Psychological perspectives on successful aging: the model of selective optimization with compensation. In: Baltes PB, Baltes MM, editors. Successful Aging: Perspectives from the Behavioral Sciences: Cambridge University Press; 1990:1-34. [doi: 10.1017/CBO9780511665684.003]
73. Yin R, Martinengo L, Smith HE, Subramaniam M, Griva K, Tudor Car L. The views and experiences of older adults regarding digital mental health interventions: a systematic review of qualitative studies. Lancet Healthy Longev 2024 Nov;5(11):100638. [doi: 10.1016/j.lanhl.2024.08.007] [Medline: 39491881]
74. Estrela M, Semedo G, Roque F, Ferreira PL, Herdeiro MT. Sociodemographic determinants of digital health literacy: a systematic review and meta-analysis. Int J Med Inform 2023 Sep;177:105124. [doi: 10.1016/j.ijmedinf.2023.105124] [Medline: 37329766]
75. Baker P. Who self‐cares wins: an updated perspective on men and self‐care. Trends Urol & Men's Health 2019 May;10(3):19-22. [doi: 10.1002/tre.694]
76. Höhn A, Gampe J, Lindahl-Jacobsen R, Christensen K, Oksuyzan A. Do men avoid seeking medical advice? A register-based analysis of gender-specific changes in primary healthcare use after first hospitalisation at ages 60+ in Denmark. J Epidemiol Community Health 2020 Jun;74(7):573-579. [doi: 10.1136/jech-2019-213435] [Medline: 32303595]
77. Seidler ZE, Dawes AJ, Rice SM, Oliffe JL, Dhillon HM. The role of masculinity in men's help-seeking for depression: a systematic review. Clin Psychol Rev 2016 Nov;49:106-118. [doi: 10.1016/j.cpr.2016.09.002] [Medline: 27664823]
78. Rising CJ, Jensen RE, Moser RP, Oh A. Characterizing the US population by patterns of mobile health use for health and behavioral tracking: analysis of the National Cancer Institute's Health Information National Trends Survey data. J Med Internet Res 2020 May 14;22(5):e16299. [doi: 10.2196/16299] [Medline: 32406865]
79. Kayser L, Karnoe A, Furstrand D, et al. A multidimensional tool based on the eHealth literacy framework: development and initial validity testing of the eHealth Literacy Questionnaire (eHLQ). J Med Internet Res 2018 Feb 12;20(2):e36. [doi: 10.2196/jmir.8371] [Medline: 29434011]
80. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2026 Apr;33(3):2544331. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]

Abbreviations

BFI-10: Big Five Inventory-10 items
BHLS: Brief Health Literacy Screen
CHD: coronary heart disease
CONSORT: Consolidated Standards of Reporting Trials
COPD: chronic obstructive pulmonary disease
DHI: digital health intervention

XSL•FO
RenderX

DMP: disease management program
IDEAS: Integrate, Design, Assess, and Share
MCAR: missing completely at random
P-SUP: Personalized Self-management Support Program
PAM13-D: Patient Activation Measure-13
PHQ-4: Patient Health Questionnaire-4 items
RCT: randomized controlled trial
SDT: self-determination theory
T2DM: type 2 diabetes mellitus
UCLA: University of California Los Angeles

Edited by S Brini; submitted 18.Jul.2025; peer-reviewed by A Ojaruega, MH Rahman, O Oluwole; accepted 14.Feb.2026; published 25.Mar.2026.

Please cite as:
Scholl M, Lendt C, Appelbaum S, Biallas B, Brenk-Franz K, Chermette C, Frank F, Gawlik A, Giesen L, Heßbrügge M, Küppers L, Pilic L, Redaèlli M, Schneider L, Vitinius F, Wilm S, Konerding U
Determinants of the Uptake and Frequency of Use of a Web Portal Digital Health Intervention in Patients With Type 2 Diabetes and/or Coronary Heart Disease: Secondary Analysis of a Randomized Controlled Trial
J Med Internet Res 2026;28:e80895
URL: https://www.jmir.org/2026/1/e80895
doi:10.2196/80895

© Maximilian Scholl, Claas Lendt, Sebastian Appelbaum, Bianca Biallas, Katja Brenk-Franz, Chloé Chermette, Friederike Frank, Angeli Gawlik, Lisa Giesen, Martina Heßbrügge, Lucas Küppers, Larisa Pilic, Marcus Redaèlli, Lara Schneider, Frank Vitinius, Stefan Wilm, Uwe Konerding. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

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

Shuanghui Zheng[1*], MSc; Shiyu Wang[1*], BSc; Xiaoyu Shui[1*], BSc; Jing Zhao[2], MSN; Ying Lin[2], BSc; Qian Huang[1], BSc; Siqi Li[1], BSc; Yang Chen[1], BSc; Zhijie Zou[1], PhD; Jin Zheng[3], MEng; Xiaoli Chen[1], PhD

[1]School of Nursing, Wuhan University, Building 2, School of Medicine, Wuhan University, 4th Fl, No. 115, Donghu Road, Wuhan, China

[2]Department of Nursing, Maternal and Child Health Hospital of Hubei Province, Wuhan, China

[3]Qianjin 4th Road Campus, Hospital of Stomatology, Wuhan University, Wuhan, China

[*]these authors contributed equally

Corresponding Author:
Zhijie Zou, PhD
School of Nursing, Wuhan University, Building 2, School of Medicine, Wuhan University, 4th Fl, No. 115, Donghu Road, Wuhan, China

Abstract

Background: The identification and management of depression during pregnancy is an important public health issue. Although many existing psychological intervention programs are effective, their implementation is plagued by issues, such as insufficient professional resources and lengthy intervention cycles. Studies have suggested that internet-based problem management plus (IPM+) can effectively address the aforementioned challenges in the management of general depression. However, its application in the pregnant population remains to be verified.

Objective: The objective of our study was to explore the effect of IPM+ on depressive symptoms in pregnant women.

Methods: From April to October 2024, 80 pregnant women were recruited from the Hubei General Hospital and randomly assigned to the intervention group (n=40) and the control group (n=40). The control group received routine care, while the intervention group received a 5-week IPM+ intervention (1.5 hours per week) in addition to routine care. The primary outcome measure is depressive mood, and the secondary outcome measures include anxiety symptoms, perceived stress, and sleep quality. The data were collected using the Edinburgh Postnatal Depression Scale, Pregnancy Anxiety Questionnaire, 4-item Perceived Stress Scale, and Insomnia Severity Index at baseline (T1), immediately after the intervention (T2), and at 3 months (T3).

Results: Compared with the control group, the intervention group showed statistically greater improvements in depressive symptoms (between-group effect size: Hedges *g*=0.74, 95% CI 0.23-1.24 at T2; Hedges *g*=0.76, 95% CI 0.25-1.28 at T3), anxiety symptoms (between-group effect size: Hedges *g*=0.51, 95% CI 0.03-0.98 at T2; Hedges *g*=0.52, 95% CI 0.04-0.99 at T3), and perceived stress (between-group effect size: Hedges *g*=0.71, 95% CI 0.23-1.18 at T2; Hedges *g*=0.62, 95% CI 0.15-1.11 at T3) at both T2 and T3. In terms of sleep quality, the intervention group demonstrated statistically greater improvement compared with the control group only at T3 (between-group effect size: Hedges *g*=0.68, 95% CI 0.19-1.17).

Conclusions: IPM+ effectively alleviates depression and anxiety symptoms, reduces perceived stress, and improves sleep quality in pregnant women with antenatal depression. The intervention effects are sustained for up to 3 months postintervention.

Trial Registration: Chictr.org.cn ChiCTR2400083324; https://www.chictr.org.cn/showprojEN.html?proj=222789

(J Med Internet Res 2026;28:e81998) doi:10.2196/81998

KEYWORDS

antenatal depression; problem management plus; PM+; randomized controlled trial; digital intervention; mental health

Introduction

Antenatal depression (AND) refers to depressive episodes occurring during pregnancy (from conception to before delivery) and is considered a clinical subtype of perinatal depression [1]. The global prevalence of AND is reported to be approximately 20.7%, with significant regional variations: the prevalence in low- and middle-income countries (30.3%) is substantially higher than in high-income countries (18.1%) [2]. The prevalence of prenatal depression among women in China (19.7%) has exceeded the average level in high-income countries [3]. If left untreated, it can result in a dual crisis for maternal and infant health. Mothers may experience obstetric complications, such as hyperemesis gravidarum, preeclampsia, miscarriage, preterm birth, and postpartum depression [4]. In severe cases, suicidal ideation or infanticide may occur [5]. For the offspring, long-term negative effects, including cognitive developmental delays and emotional regulation disorders, are

XSL•FO
RenderX

possible, and there is a risk of intergenerational transmission of mental health problems [6-9].

Psychotherapy has now become the first-line treatment for AND [10], including evidence-based methods, such as cognitive behavioral therapy (CBT), interpersonal psychotherapy (IPT), and mindfulness-based interventions [11]. Existing evidence suggests that these therapies significantly reduce depressive symptoms and concurrently improve anxiety, sleep quality, and social support levels [12-15]. However, traditional psychotherapy is confronted with 3 core dilemmas: one lies in the barrier of human resources, as its implementation highly relies on highly qualified professionals, involving lengthy training periods and exorbitant labor costs; another manifests as the intervention is excessively time-consuming for participants, which is hardly compatible with the fragmented and mobile schedules of perinatal women, directly resulting in a high dropout rate [16]; and a deeper obstacle stems from sociocultural factors, where the prevalence of stigma leads to only 28.5% of patients with postpartum depression actively seeking professional support [17].

Internet-based problem management plus (IPM+) may become a breakthrough strategy to improve the accessibility of psychological service. Problem management plus (PM+), developed by the World Health Organization (WHO) [18], is a low-intensity intervention that integrates CBT and problem-solving therapy. It not only features transdiagnostic applicability, short duration, and implementability by nonprofessionals (such as community nurses and volunteers) but also consists of 4 modules: stress management, problem-solving–oriented action planning, continuous behavioral activation, and interpersonal relationship enhancement [19]. Several randomized controlled trials have demonstrated that PM+ has significant effects on depression, posttraumatic stress disorder, and anxiety in the general population, with effects lasting up to 3 months postintervention [20-23]. At the same time, compared to traditional offline methods, online psychological interventions offer 3 main advantages: overcoming spatial limitations by relying on various digital platforms, allowing perinatal women to receive interventions anywhere; time flexibility, with asynchronous courses (eg, prerecorded videos, artificial intelligence question and answer) that fit into fragmented schedules; and stigma reduction, as the anonymous environment encourages help-seeking behavior [24]. IPM+ not only boasts the advantages of internet interventions, such as geographical reach, time flexibility, and reduction of stigma, but also retains the strengths of PM+ as an integrated multimodule intervention. However, the application of PM+ in pregnant populations remains in the exploratory stage, since only 1 cultural adaptation study conducted in Pakistan worldwide has involved this [23].

To explore the intervention effect of IPM+ on AND, this study developed a randomized controlled trial, which includes a 5-week intervention period and a 3-month follow-up period. From the perspective of empirical research, it aims to answer the scientific question of “whether IPM+ can improve AND.”

Methods

Study Design

This study adopts a single-blind, 1:1 parallel-group randomized controlled trial design. The study was conducted at Hubei General Hospital, a large Grade A tertiary hospital located in Wuhan, China, and strictly followed the CONSORT (Consolidated Standards of Reporting Trials; Checklist 1) guidelines.

Ethical Considerations

The research was approved by the Ethics Committee of the Medical School of Wuhan University (WHU-LFMD-IRB2024006) and registered with the Chinese Clinical Trial Registry (ChiCTR2400083324). All participants provided written informed consent before completing the baseline assessment. To ensure participant privacy and confidentiality, all collected data were deidentified and stored on a secure, password-protected server, with access strictly restricted to the research team. Regarding compensation, no financial or material incentives were provided to the participants for completing the intervention or the assessment sessions.

Participants

Participants were recruited through convenience sampling from the Obstetrics Department of Hubei General Hospital between April 2024 and October 2024. Eligible patients who met the inclusion criteria were invited to participate in the study. The inclusion criteria were as follows: (1) age 18 years and above, (2) gestational age of 28 weeks or less, (3) singleton pregnancy, (4) no plans for elective termination of pregnancy, (5) Edinburgh Postnatal Depression Scale (EPDS) score of 10 or higher (if EPDS score of 13 or higher, further assessment was required to determine suitability for enrollment), and (6) ability to understand the study questionnaire. The exclusion criteria were as follows: (1) presence of suicidal thoughts or behaviors, (2) serious organic diseases or functional disorders, (3) severe mental or cognitive disorders (eg, severe intellectual disability, dementia), (4) substance or alcohol dependence or abuse behaviors within the past 6 months, and (5) previous experience with PM+ treatment or simultaneous participation in other psychological interventions.

Blinding

Non-researchers generated a random number sequence using a random number table, which was then imported into the Excel software and arranged in ascending order. Based on the sorted random number sequence, the first 50% (n=40) of the participants were assigned to the intervention group, and the remaining 50% (n=40) were assigned to the control group. To ensure allocation concealment, the random allocation scheme was sealed in opaque envelopes, numbered sequentially. These envelopes were opened by research assistants, who were not involved in the study evaluation, after the participants completed the baseline data collection to execute the final group assignments. The study adopted a single-blind design, with all participants unaware of their group allocation. Throughout the recruitment, informed consent, and the entire study process, we consistently utilized the unified project title “Perinatal Mental

XSL•FO
RenderX

Health Online Support Project." All participants were informed that the project involved 2 different forms of online guidance programs, both of which were beneficial for pregnancy health. Furthermore, we ensured that both the control group and the intervention group included weekly one-on-one meetings with an online facilitator to maintain consistency. Due to the nature of the intervention, the implementers were not blinded.

Intervention Group

On the basis of routine care, the intervention group received a 5-week IPM+ intervention (1.5 hours per week). The intervention plan was compiled based on the "PM+" manual of the Chinese version of the WHO [25] and tailored to address the specific characteristics of perinatal depression. Key adaptations included pregnancy-specific case examples and stressors, the incorporation of safe and appropriate activities for the antenatal period, and an emphasis on social support derived from peer interactions among pregnant women. The adapted protocol was reviewed and validated by a multidisciplinary panel of experts (including a clinical psychologist, an obstetrician, a PM+ supervisor, and a senior midwife) for cultural relevance, safety, and feasibility. Furthermore, a usability pretest was conducted with 4 pregnant women (who met the inclusion criteria but did not participate in the main trial) to preliminarily assess its acceptability and comprehensibility. The intervention was conducted in an individualized online format, with structured 1.5-hour sessions every week via the WeChat platform. The intervention was delivered by 2 nonspecialist providers, both of whom were nursing graduate students. Prior to study initiation, they completed a standardized 40-hour training program based on the Chinese version of the WHO PM+ manual [25]. Competency was subsequently assessed through role-play–based simulations.

The intervention content adopted a modular design, focusing on the development of 4 core psychological abilities for pregnant and postpartum women: stress regulation, problem-solving ability, behavioral activation, and social support ability. The intervention process followed a gradual approach and was implemented in 5 stages: the first stage focused on building a good therapeutic relationship, helping participants understand their responses to adversity and master basic relaxation techniques; the second stage systematically trained problem-solving skills, using methods such as brainstorming to find feasible solutions; the third stage used behavioral experiments to verify the effectiveness of the solutions and established a regular activity pattern; the fourth stage strengthened the social support network and improved the quality of interpersonal interactions; and the final stage involved an overall review and planning for future applications, ensuring the long-term maintenance of intervention effects. After each intervention, targeted homework assignments were given, and progress was monitored using standardized assessment tools. Detailed content is presented in Multimedia Appendix 1. Detailed problem management steps are presented in Multimedia Appendix 2.

To ensure the standardization and consistency of the intervention, the research team established a comprehensive quality control system, including audio recording analysis of 10% (n=17) of the intervention sessions, biweekly supervision meetings focusing on adherence and competence, and random quality checks. Participants who exhibited severe deterioration in mental health or were identified as having a risk of self-harm or suicide during the intervention period were immediately referred to a psychiatrist for specialized care and were withdrawn from the study.

Control Group

The participants in the control group received a 5-week regimen of standard prenatal care, which was administered by obstetric nursing staff who underwent uniform training through a WeChat platform. The standard care program included weekly systematic health follow-ups, during which the nursing staff utilized standardized communication scripts to assess the physical and mental health status of the pregnant women. Additionally, professional prenatal health guidance was provided, covering basic health education services, such as prenatal examination precautions, key aspects of perinatal mental health maintenance, and newborn care knowledge.

Outcomes

The primary outcome measure was depressive symptoms, assessed using the EPDS [26,27]. This scale consists of 10 items, with a 4-point Likert scale for scoring, where the frequency of symptoms is rated from "never" to "always," corresponding to scores of 0-3. The Chinese version of the EPDS [28] demonstrates good internal reliability (Cronbach α=0.89). The clinical diagnostic criteria are as follows: a total score of less than 9 indicates a normal state, 9 - 12 points suggest possible depression, and a total score of 13 or higher or a score greater than 0 on item 10 (self-harm thoughts) indicates a depressive state.

The following secondary outcome measures were also assessed: pregnancy-related anxiety: Pregnancy Anxiety Questionnaire (PAQ) [29]; perceived stress via the 4-item Perceived Stress Scale (PSS) [30,31]; and sleep quality via the Insomnia Severity Index (ISI) [32].

Additionally, baseline general information was collected, including age, gestational age, last menstrual period, height, current or prepregnancy weight, education level, marital status, occupation, residence, per capita family monthly income, living arrangements with elderly family members, knowledge of pregnancy-related information, and a history of prepregnancy depression. The data were collected at baseline (T1), immediately postintervention (T2), and 3 months postintervention (T3). The 3-month follow-up was selected, as it represents a standard medium-term evaluation period in behavioral intervention trials and aligns with prior WHO PM+ research, thereby facilitating cross-study comparison.

Sample Size

The study employs the sample size estimation method for comparing the means of 2 samples, with depression scores as the calculation indicator, and uses the following formula to calculate the sample size:

$$(1)n1=n2=2\times[(\mu\alpha+\mu\beta)/d]2$$

XSL•FO
RenderX

A 2-tailed test was adopted with α=.05 and β=.10. The anticipated effect size for this study was informed by the meta-analysis conducted by Ansaari et al [33] on online psychological interventions for perinatal depression, from which a Hedges *g* of 0.86 was derived for the primary outcome. Accounting for an estimated 20% loss to follow-up, the calculated sample size was approximately 37 participants per group, resulting in a total required sample of 74 participants. This attrition estimate is based on the typical range (15% - 25%) observed in comparable trials of PM+ [34].

Statistical Analysis

Descriptive statistical analysis was used for baseline data: unordered categorical variables were expressed as percentages, and the chi-square test was used for between-group comparisons. When the expected count was less than 5, the Fisher exact test was applied. Ordinal categorical variables were described as median (IQR), and the Mann-Whitney *U* test was used for between-group comparisons. Normally distributed continuous variables were presented as mean (SD), with the independent samples *t*-test used for between-group comparisons; non-normally distributed continuous variables were described as median (IQR), and the Mann-Whitney *U* test was also used for between-group comparisons.

In this study, intention-to-treat analysis was adopted for primary and secondary outcomes. Based on the data after multiple imputation, linear mixed-effects models were used for analysis. The covariates included in the model were baseline variables with between-group differences (whether depression was present before pregnancy), and the random effect was set as the individual random intercept.

Standardized mean differences at time points T2 and T3 were calculated to represent between-group effect sizes, while standardized mean differences between T2, T3, and T1 within the same group served as within-group effect sizes. Both were quantified using Hedges *g* values and 95% CI. SPSS 26.0 (IBM Corp) was used for data analysis in this study, with the significance level set at $P<.05$.

Results

Overview

A total of 515 participants were screened for eligibility assessment between April 2024 and October 2024, and finally, 80 eligible participants were included in this study, who were randomly assigned to the intervention group (n=40) or the control group (n=40). In the intervention group, 34 (85.0%) participants and 32 (80%) participants completed the assessments at T2 and T3, respectively. In the control group, 32 (80%) participants and 28 (70.0%) participants completed the assessments at T2 and T3, respectively. In accordance with the intention-to-treat principle, all 80 randomized participants (40 per group) were included in the final analyses after missing data were handled using multiple imputation. The flowchart of participant recruitment and grouping is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Flowchart of participants. ITT: intention-to-treat.

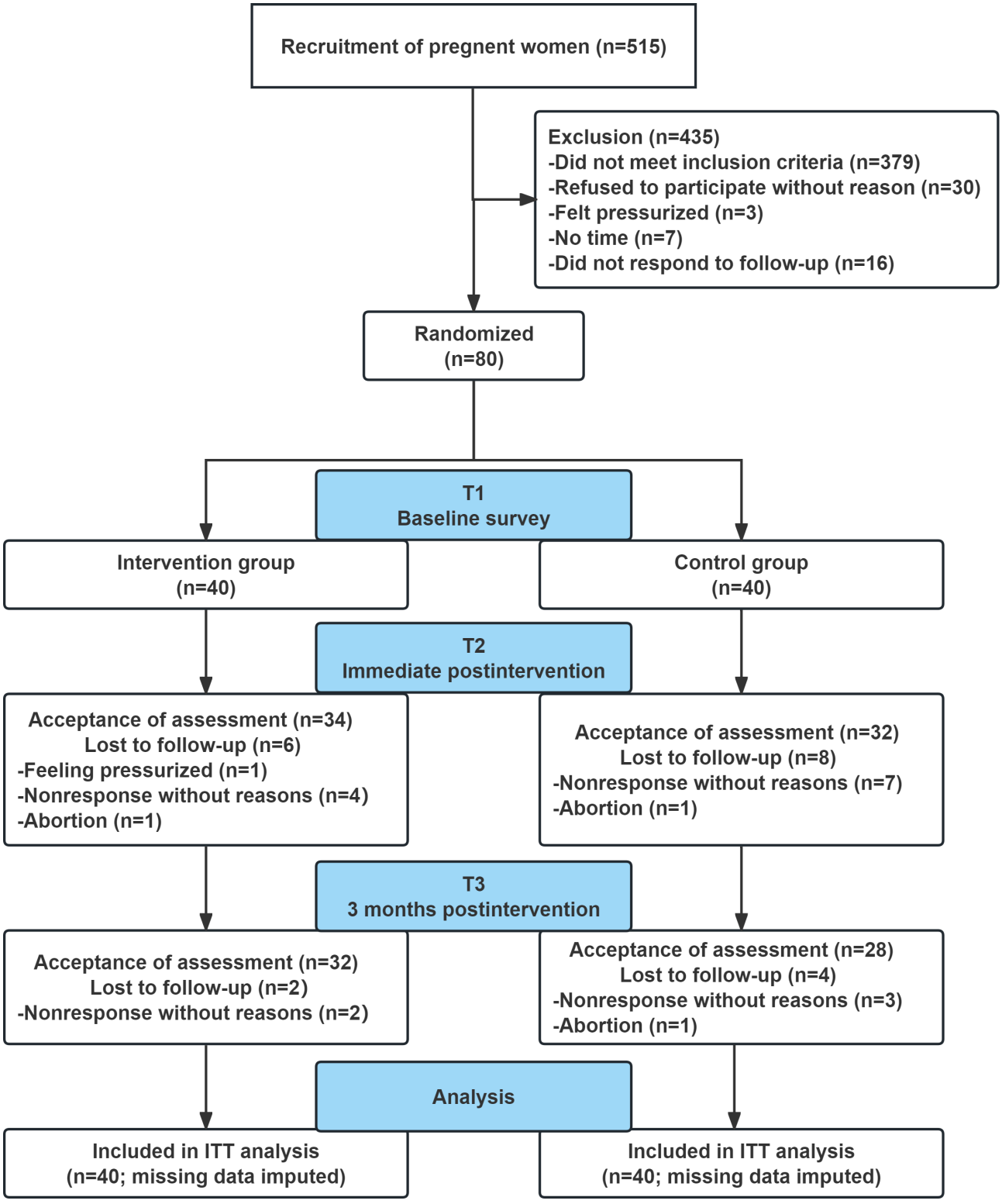

Baseline Assessment

The baseline data of both the intervention group and the control group are presented in Table 1. A comparison of the 2 groups in terms of demographic characteristics revealed a statistically significant difference in the history of depression before pregnancy (P=.005), with the intervention group showing a higher proportion of individuals who had experienced depression prior to pregnancy. Regarding the scores of the EPDS, PAQ, PSS, and ISI, the differences between the 2 groups did not reach statistical significance.

XSL•FO
RenderX

Table . Baseline characteristics of participants from the intervention and control groups (n=80).

General information indicators	Intervention group (n=40)	Control group (n=40)	*P* value
Age (years), mean (SD)	30.63 (4.78)	30.43 (4.29)	.84
Gestational age, median (IQR)	13.5 (12.0-23.7)	14.5 (11.0-22.0)	.43
Number of pregnancies, n (%)			>.99
First pregnancy	24 (60)	24 (60)	
Nonfirst pregnancy	16 (40)	16 (40)	
Number of deliveries, n (%)			.45
Primiparous	31 (77.5)	28 (70.0)	
Multiparous	9 (22.5)	12 (30.0)	
Prepregnancy BMI (kg/m²), median (IQR)	22.0 (20.0-24.0)	21.0 (20.0-25.0)	.78
Baseline BMI (kg/m²), median (IQR)	23.0 (20.0-25.0)	22.5 (12.0-27.0)	.54
Employment status, n (%)			.45
Employed	28 (70)	31 (77.5)	
Other	12 (30)	9 (22.5)	
Place of residence, n (%)			.53
Rural	5 (12.5)	7 (17.5)	
Urban or city	35 (87.5)	33 (82.5)	
Whether living with elderly family members, n (%)			.50
Yes	20 (50)	23 (57.5)	
No	20 (50)	17 (42.5)	
History of depression before pregnancy, n (%)			.005
Yes	8 (20)	0 (0)	
No	32 (80)	40 (100)	
Education level, n (%)			.05
Elementary school or below	0 (0)	0 (0)	
Middle school	1 (2.5)	5 (12.5)	
High school or vocational school	8 (20)	10 (25.0)	
Bachelor’s degree	22 (55.0)	21 (52.5)	
Master’s degree or higher	9 (22.5)	4 (10)	
Per capita monthly family income (yuan [US $]/month), n (%)			.52
3000-5000 [432-720]	7 (17.5)	7 (17.5)	
5001-10,000 [720-1440]	17 (42.5)	19 (47.5)	
>10,000 [>1440]	14 (35)	11 (27.5)	
Knowledge level of pregnancy, n (%)			.75
Not at all knowledgeable	3 (7.5)	3 (7.5)	
Somewhat knowledgeable	35 (87.5)	36 (90)	
Very knowledgeable	2 (5.0)	1 (2.5)	
Depressive symptoms (EPDS[a]), mean (SD)	12.32 (1.62)	11.85 (2.14)	.07

XSL•FO
RenderX

General information indicators	Intervention group (n=40)	Control group (n=40)	*P* value
Anxiety symptoms (PAQ[b]), mean (SD)	25.85 (2.92)	26.00 (4.25)	.89
Stress level (PSS[c]), mean (SD)	7.40 (1.94)	7.97 (1.34)	.47
Sleep quality (ISI[d]), mean (SD)	10.60 (5.65)	11.32 (4.89)	.66

[a]EPDS: Edinburgh Postnatal Depression Scale.

[b]PAQ: Pregnancy Anxiety Questionnaire.

[c]PSS: 4-item Perceived Stress Scale.

[d]ISI: Insomnia Severity Index.

Outcome Measures

The linear mixed model uses the total score of the EPDS, the score of the PAQ, the score of the PSS, and the score of the ISI as dependent variables, with the control group and baseline time point (T1) as reference levels. It includes group (intervention group and control group), time (T1 or T2 or T3), and group×time interaction term as fixed effects, and controls for the covariate of a history of depression before pregnancy.

Depressive Mood

Table 2 shows that the between-group effect ($F_{1,95.747}$=13.564; P<.001), time effect ($F_{2,97.600}$=116.563; P<.001), and group×time interaction effect ($F_{2,97.600}$=24.203; P<.001) of depressive mood were all statistically significant.

Table . Results of fixed effects tests for each dependent variable.[a]

Effect type		*F* test (*df*)	*P* value
EPDS[b]			
	Between-group effect	13.564 (1, 95.747)	<.001
	Time effect	116.563 (2, 97.600)	<.001
	Interaction effect (time × between-group)	24.203 (2, 97.600)	<.001
PAQ[c]			
	Between-group effect	8.388 (1, 69.106)	.005
	Time effect	30.762 (2, 100.407)	<.001
	Interaction effect (time × between-group)	7.122 (2, 100.407)	.001
PSS[d]			
	Between-group effect	20.444 (1, 75.596)	<.001
	Time effect	29.707 (2, 90.567)	<.001
	Interaction effect (time × between-group)	7.255 (2, 90.567)	.001
ISI[e]			
	Between-group effect	13.458 (1, 73.646)	<.001
	Time effect	37.581 (2, 97.521)	<.001
	Interaction effect (time × between-group)	5.566 (2, 97.521)	.005

[a]The control group, T1 (baseline), and a history of depression before pregnancy were used as the reference categories.

[b]EPDS: Edinburgh Postnatal Depression Scale.

[c]PAQ: Pregnancy Anxiety Questionnaire.

[d]PSS: 4-item Perceived Stress Scale.

[e]ISI: Insomnia Severity Index.

The fixed-effects model (Table 3) analysis showed that the baseline effect in the intervention group was not statistically significant (β=.55; P=.24). The time effects were statistically significant at both T2 (β=−2.04; P<.001, 95% CI −2.92 to −1.16) and T3 (β=−2.34; P<.001, 95% CI −3.51 to −1.18) versus T1. We observed significant time×group interactions at T2

XSL•FO
RenderX

(β=−3.26; *P*<.001, 95% CI −4.51 to −2.01) and T3 (β=−3.75; *P*<.001, 95% CI −5.34 to −2.17) versus the control group T1. No history of depression before pregnancy showed an effect coefficient of 0.33 (*P*=.68; 95% CI −1.21 to 1.87), with no statistical significance. We reanalyzed the participants without a history of depression (32 participants in the intervention group and 40 participants in the control group). The results showed that the impact of the intervention on the primary outcomes (EPDS) of T2 (β=−3.40; *P*<.001) and T3 (β=−3.76; *P*<.001) remained very significant.

Table . Fixed effect estimation results of depression.[a]

Parameter	Coefficient (β)	SE	*t* test (*df*)	*P* value	95% CI	
					Lower bound	Upper bound
Intercept	12.00	0.91	13.15 (73.2)	<.001	10.21	13.79
Intervention group	0.55	0.47	1.17 (69.6)	.24	−0.37	1.46
History of depression before pregnancy						
No	0.33	0.79	0.42 (72.2)	.68	−1.21	1.87
Time						
T2	−2.04	0.45	−4.57 (87.8)	<.001	−2.92	−1.16
T3	−2.34	0.59	−3.99 (85.2)	<.001	−3.51	−1.18
Time×group interaction						
T2	−3.26	0.64	−5.12 (87.8)	<.001	−4.51	−2.01
T3	−3.75	0.80	−4.68 (85.2)	<.001	−5.34	−2.17

[a]The intervention group was referenced to the control group. Time points T2 and T3 were referenced to time point T1. The group without a history of depression before pregnancy was referenced to the group with a history of depression before pregnancy.

The trend of mean EPDS total scores over the 3 time points in the intervention group and the control group is shown in Figure 2. The score trajectories of the 2 groups diverged immediately after the intervention (T2). The scores in the intervention group decreased sharply at T2 and remained stable at T3, whereas the scores in the control group showed only a modest decline and remained consistently higher than those in the intervention group.

XSL•FO
RenderX

Figure 2. Line chart of the mean scores of Edinburgh Postnatal Depression Scale (EPDS) total scores in the 2 groups at 3 time points.

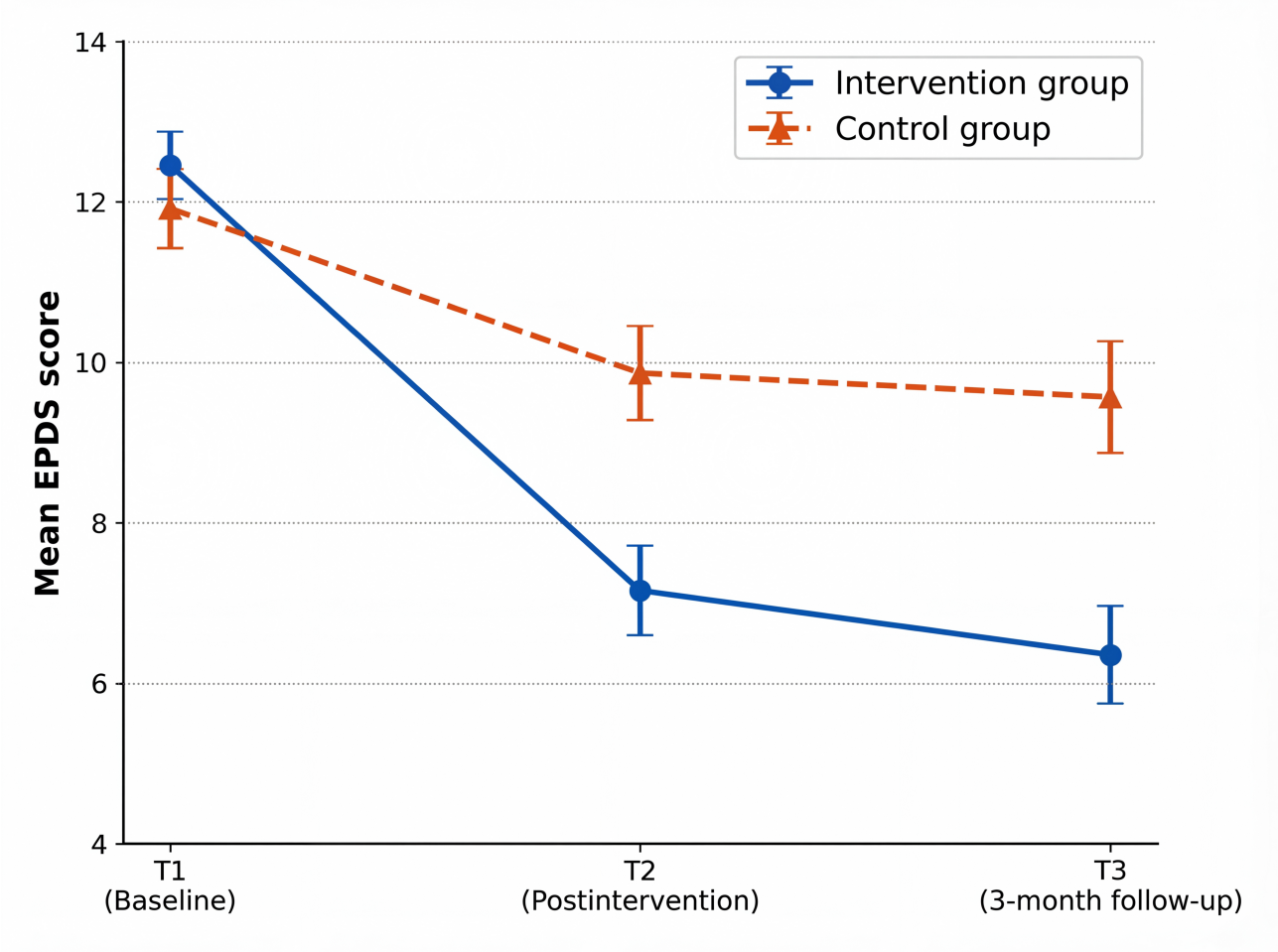

Further intra- and intergroup effect comparisons were conducted. The analysis of intragroup effect sizes (Table 4) showed that the intervention group exhibited significant reductions in depressive scores at T2 (Hedges *g*=1.69, 95% CI 1.18-2.20) and T3 (Hedges *g*=1.75, 95% CI 1.23-2.26), with large effect sizes. In contrast, the reduction in the control group was smaller (T2: Hedges *g*=0.58, 95% CI 0.13-1.02; T3: Hedges *g*=0.59, 95% CI 0.13-1.03). In intergroup comparisons, the intervention group scored significantly lower than the control group in terms of depression scores at T2 (Hedges *g*=0.74, 95% CI 0.23-1.24) and T3 (Hedges *g*=0.76, 95% CI 0.25-1.28; Table 4, Figure 2).

XSL•FO
RenderX

Table . Between-group and within-group effect sizes.[a]

	Intervention group (n=40)			Control group (n=40)			Between-group	
	Mean (SD)	Hedges *g* (95% CI)	*P* value	Mean (SD)	Hedges *g* (95% CI)	*P* value	Hedges *g* (95% CI)	*P* value
EPDS[b]								
T1	12.46 (2.67)	—[c]	—	11.92 (3.11)	—	—	—	—
T2	7.16 (3.51)	1.69 (1.18 to 2.20)	<.001	9.87 (3.71)	0.58 (0.13 to 1.02)	.003	0.74 (0.23 to 1.24)	.001
T3	6.36 (3.85)	1.75 (1.23 to 2.26)	<.001	9.57 (4.40)	0.59 (0.13 to 1.03)	<.001	0.76 (0.25 to 1.28)	<.001
PAQ[d]								
T1	24.72 (5.34)	—	—	24.89 (6.07)	—	—	—	—
T2	19.82 (5.06)	0.91 (0.44 to 1.37)	<.001	22.70 (6.13)	0.35 (0.06 to 0.62)	.008	0.51 (0.03 to 0.98)	.03
T3	20.54 (5.27)	0.77 (0.31 to 1.23)	<.001	23.56 (6.05)	0.21 (−0.03 to 0.45)	.08	0.52 (0.04 to 0.99)	.02
PSS[e]								
T1	7.65 (2.25)	—	—	8.33 (2.61)	—	—	—	—
T2	5.85 (2.51)	0.74 (0.27 to 1.22)	<.001	7.83 (2.96)	0.17 (−0.14 to 0.48)	.20	0.71 (0.23 to 1.18)	.002
T3	5.50 (2.70)	0.82 (0.34 to 1.31)	<.001	7.34 (3.01)	0.34 (−0.07 to 0.75)	.09	0.62 (0.15 to 1.11)	.005
ISI[f]								
T1	9.83 (4.36)	—	—	10.54 (4.91)	—	—	—	—
T2	5.06 (4.71)	1.01 (0.46 to 1.57)	<.001	7.26 (5.44)	0.61 (0.13 to 1.09)	.002	0.42 (−0.08 to 0.94)	.06
T3	4.88 (4.86)	1.03 (0.48 to 1.59)	<.001	8.52 (5.51)	0.37 (−0.03 to 0.79)	.079	0.68 (0.19 to 1.17)	.002

[a]Small, medium, and large effect sizes are defined as 0.2, 0.5, and 0.8, respectively. Lower scores indicate lower levels of depression. Statistical significance is determined when the 95% CI of Hedges *g* does not include "0." T1 is used as the reference.

[b]EPDS: Edinburgh Postnatal Depression Scale.

[c]Not applicable.

[d]PAQ: Pregnancy Anxiety Questionnaire.

[e]PSS: 4-item Perceived Stress Scale.

[f]ISI: Insomnia Severity Index.

Secondary Outcome

Among the secondary outcome indicators, the between-group effect ($F_{1,69.106}$=8.388; *P*=.005), time effect ($F_{2,100.407}$=30.762; *P*<.001), and interaction effect ($F_{2,100.407}$=7.122; *P*=.001) of anxiety symptoms; the between-group effect ($F_{1,75.596}$=20.444; *P*<.001), time effect ($F_{2,90.567}$=29.707; *P*<.001), and interaction effect ($F_{2,90.567}$=7.255; *P*=.001) of perceived stress; and the between-group effect ($F_{1,73.646}$=13.458; *P*<.001), time effect ($F_{2,97.521}$=37.581; *P*<.001), and interaction effect ($F_{2,97.521}$=5.566; *P*=.005) of sleep quality were all statistically significant (Table 2).

The results of fixed-effect estimates (Multimedia Appendix 3) revealed the dynamic change characteristics of secondary outcomes. For anxiety symptoms (Table S3 in Multimedia Appendix 3), the baseline levels were comparable between the 2 groups (β=−0.17, $t_{71.2}$=−0.17; *P*=.87). The time effect was significant at T2 (β=−2.19, $t_{107.5}$=−3.18; *P*<.001) and marginally significant at T3 (β=−1.33, $t_{101.0}$=−1.95; *P*=.05). Additionally, the group×time interaction effects were significant at both T2 (β=−2.72, $t_{107.5}$=−2.78; *P*=.006) and T3 (β=−2.86, $t_{101.0}$=−2.99; *P*=.003), suggesting that the intervention group had more prominent improvements in anxiety symptoms.

Regarding perceived stress (Table S4 in Multimedia Appendix 3), there was no significant difference in baseline levels between the 2 groups (β=−.68, $t_{74.8}$=−1.67; *P*=.096). The time effect was significant at T3 (β=−.98, $t_{88.4}$=−2.73; *P*=.007) but not at T2 (β=−0.49, $t_{100.7}$=−1.52; *P*=.14). The interaction effect showed that the reduction in the PSS in the intervention group during period T2 was significantly greater than that in the control group

XSL•FO
RenderX

(β=−1.30, $t_{100.7}$=−3.17; *P*=.002); the between-group difference persisted during period T3 (β=−1.16, $t_{88.4}$=−2.19; *P*=.03), but the effect intensity was slightly weaker than that at the T2 time point.

The fixed-effect estimates for sleep quality (Table S5 in Multimedia Appendix 3) indicated that the baseline levels were comparable between the 2 groups (β=−.71, $t_{54.2}$=−0.87, *P*=.38). The time effects were significant at both T2 (β=−3.27, $t_{98.3}$=−4.01; *P*<.001) and T3 (β=−2.01, $t_{118.1}$=−2.66; *P*=.008), showing a steady improvement in sleep quality over time. Moreover, the group×time interaction effect was significant only at T3 (β=−2.93, $t_{118.1}$=−2.72; *P*=.007).

Line plots (Figures 3-5) depict the trajectories of anxiety, perceived stress, and sleep quality across the 3 assessment time points. In the intervention group, anxiety decreased markedly after the intervention (T2) and showed a slight rebound at follow-up (T3) while remaining well below baseline levels, whereas the control group exhibited only minor fluctuations over time. Perceived stress in the intervention group declined substantially at T2 and was maintained at follow-up, indicating sustained effects, while stress levels in the control group remained largely stable throughout the study. Differences in sleep quality between the groups were the smallest immediately postintervention (T2); however, a clear divergence emerged at follow-up (T3), with continued improvement observed in the intervention group and a rebound trend in the control group.

Figure 3. Line chart of the mean scores of Pregnancy Anxiety Questionnaire (PAQ) total scores in the 2 groups at 3 time points.

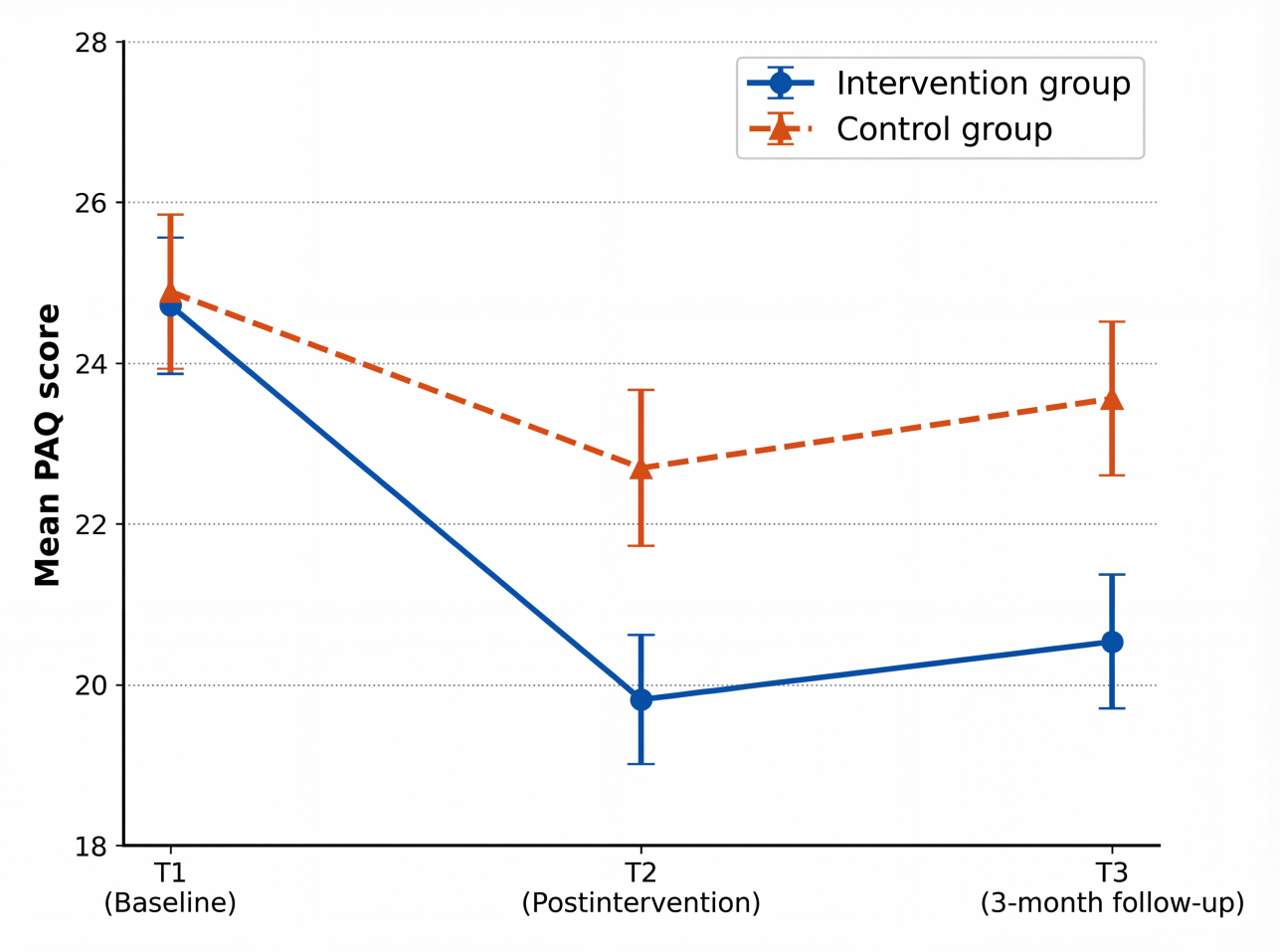

XSL•FO
RenderX

Figure 4. Line chart of the mean scores of the 4-item Perceived Stress Scale (PSS) total scores in the 2 groups at 3 time points.

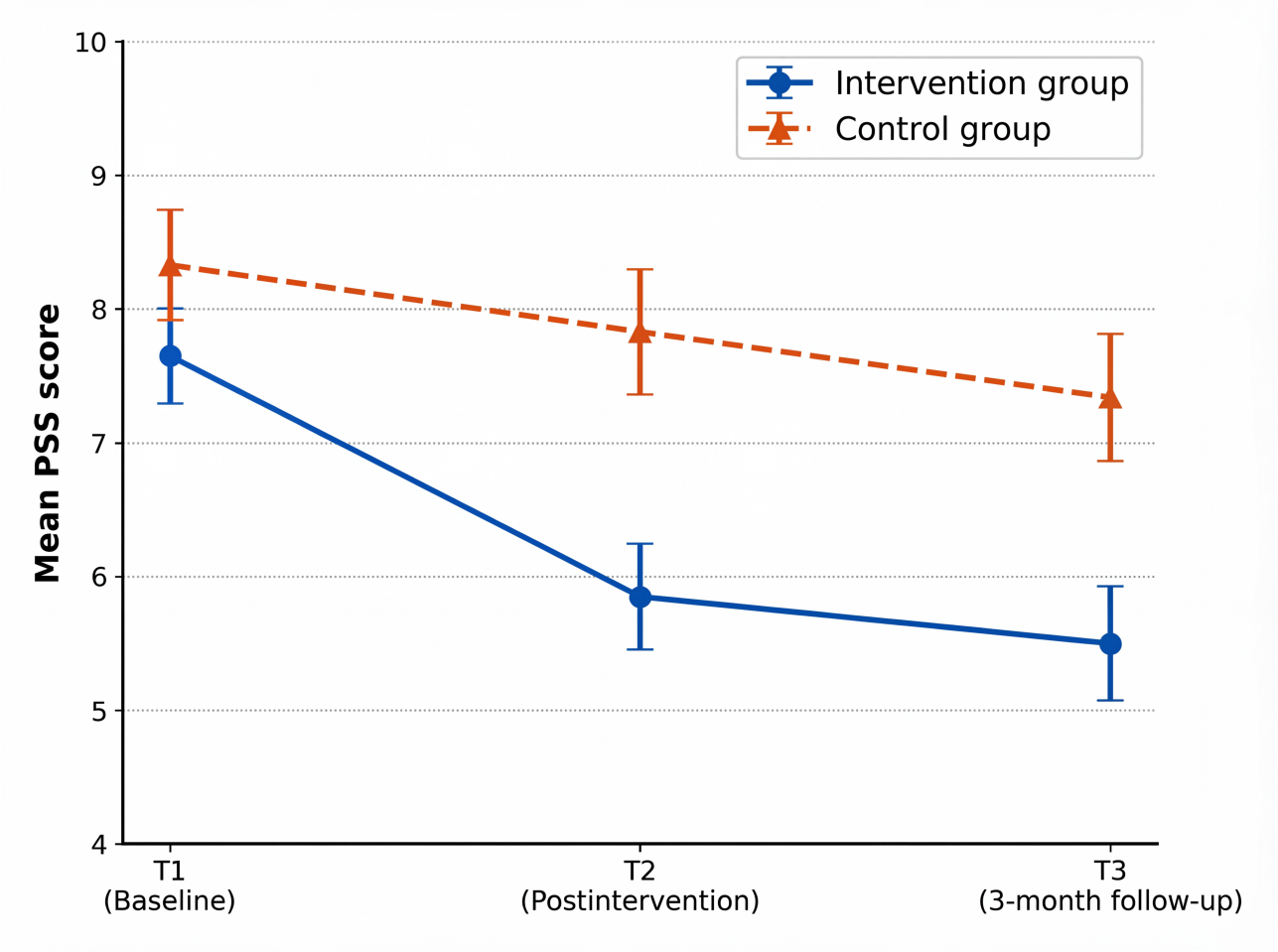

XSL•FO
RenderX

Figure 5. Line chart of the mean scores of Insomnia Severity Index (ISI) total scores in the 2 groups at 3 time points.

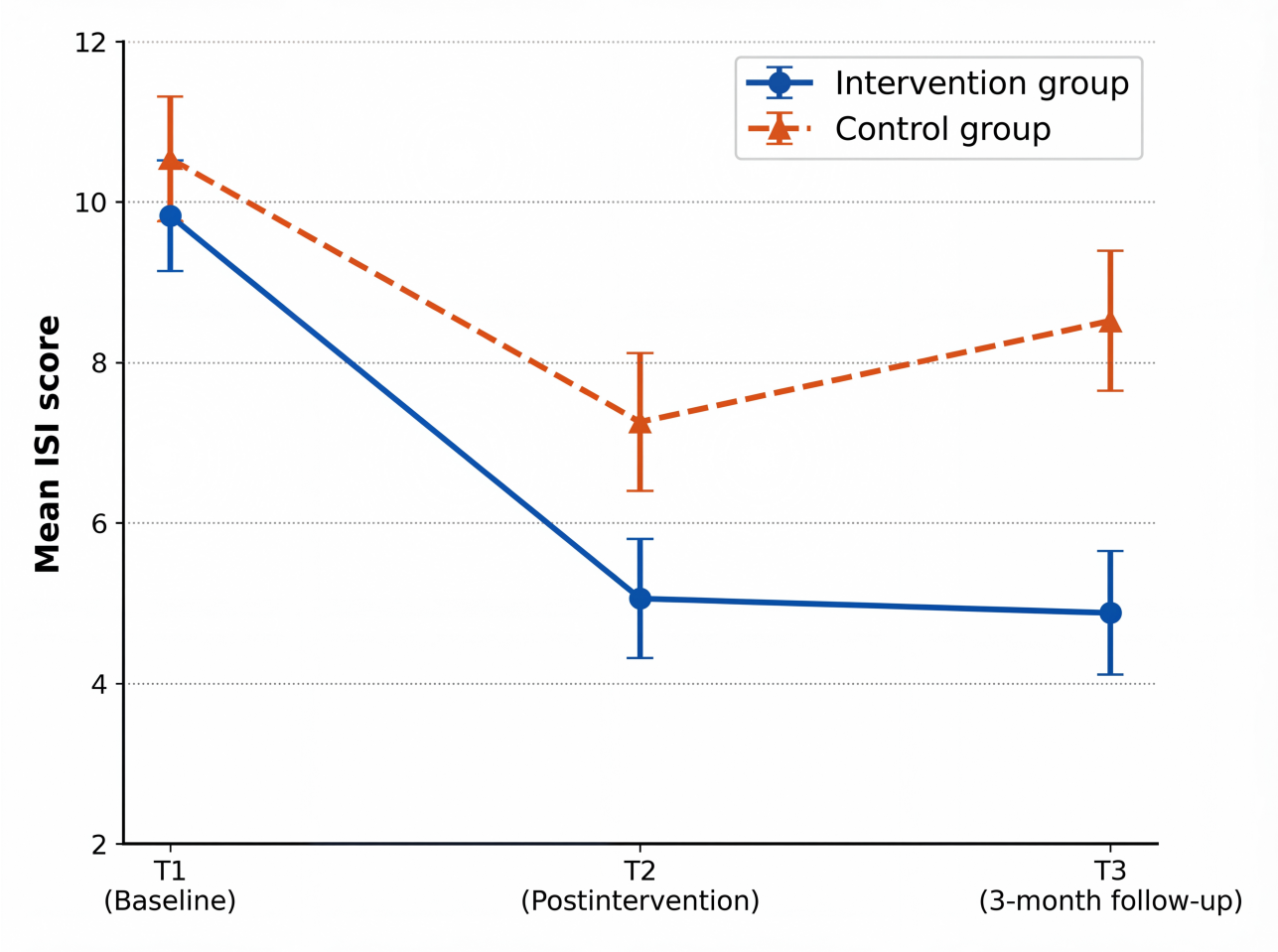

In terms of between-group effect sizes (Table 4), for anxiety symptoms, the scores of the intervention group at T2 (Hedges *g*=0.51, 95% CI 0.03-0.98) and T3 (Hedges *g*=0.52, 95% CI 0.04-0.99) were significantly lower than those of the control group, showing a medium effect. For perceived stress, the scores of the intervention group at T2 (Hedges *g*=0.71, 95% CI 0.23-1.18) and T3 (Hedges *g*=0.62, 95% CI 0.15-1.11) were significantly lower than those of the control group, presenting a medium effect. For sleep quality, the score of the intervention group at T3 (Hedges *g*=0.68, 95% CI 0.19-1.17) was significantly lower than that of the control group, with a medium effect, while the effect size at T2 was 0.42 (95% CI –0.08 to 0.94), which was not statistically significant.

Sensitivity Analysis

To assess the potential impact of missing data on the results, this study conducted a linear mixed-effects model analysis on the original un-imputed data and compared the effect sizes (Hedges *g*), direction, magnitude, confidence intervals (95% CI), and significance levels (*P*-values) before and after multiple imputation. The results showed that the missing data did not alter the significance or direction of the core outcome measures, supporting the robustness of the study's conclusions. The detailed results are presented in Table S6 in Multimedia Appendix 3.

Safety and Adverse Events

No serious adverse events related to the intervention were reported throughout the trial. In the intervention group, 1 participant withdrew at T2 due to feeling stressed by the pace of the program, which was recorded as a nonserious intervention–related adverse event. Another participant withdrew because of a miscarriage; this event was determined to be unrelated to the intervention. No participants required referral to psychiatric services due to acute suicidal risk or severe deterioration in mental health.

Discussion

Principal Findings

To our knowledge, this study represents the first randomized controlled trial in a Chinese cultural context to examine the effectiveness of an IPM+ intervention for antenatal depression. The results demonstrate that IPM+ was associated with significant and sustained improvements in depressive symptoms, anxiety, and perceived stress up to 3 months after the intervention. Notably, improvements in sleep quality exhibited a delayed pattern, with clearer benefits observed at follow-up (T3).

XSL•FO
RenderX

The Intervention Effect of IPM+ on Prenatal Depressive Mood

The results of this study demonstrate that the IPM+ intervention effectively alleviated AND, with effects sustained up to 3 months postintervention, consistent with the findings by Fink et al [35] showing prolonged benefits of telephone-delivered PM+ on maternal psychological distress. The significant efficacy observed in this study may stem from the synergistic intervention of the IPM+ modular design on the multidimensional causes of perinatal depression, including stress management techniques that regulate physiological stress responses, problem-solving strategies that enhance perceived control and self-efficacy, behavioral activation that counters low mood and social withdrawal, and social support facilitation that buffers stress and isolation. By simultaneously targeting the physiological, cognitive, behavioral, and social dimensions of depression, IPM+ may achieve sustained benefits within a relatively brief intervention period of 5 weeks. Notably, despite a higher proportion of participants with a history of prepregnancy depression in the intervention group, analyses indicated that prepregnancy depressive history did not significantly influence improvements in depressive symptoms, suggesting that IPM+ may be broadly applicable across perinatal women with varying levels of depression risk.

Compared with internet-based cognitive behavioral therapy (iCBT), IPM+ appears to be more time-efficient. A meta-analysis by Pan et al [36] indicated that interventions lasting fewer than 9 weeks did not produce significant improvements in postnatal depression (standardized mean difference=−0.36; P=.06), whereas the interventions of 9 weeks or longer demonstrated large and significant effects (standardized mean difference=−0.85; P<.01). In contrast, the IPM+ intervention achieved significant reductions in depressive symptoms within only 5 weeks, thereby reducing participant time burden and potentially enhancing the accessibility of psychological interventions for perinatal women. From a cost perspective, the IPM+ allows for delivery by nonspecialists, offering a scalable and economical alternative to therapist-intensive models, such as iCBT. This is particularly vital for implementation in resource-constrained settings.

Relative to existing internet-based mindfulness-based interventions, IPM+ also demonstrated comparatively larger effect sizes. For example, a randomized controlled trial by Kim et al. [37] reported moderate effects of a mobile-based mindfulness intervention on depressive severity (partial η^2=0.06), whereas this study observed a large effect size for perinatal depression. This disparity likely stems from the divergent mechanisms underlying each intervention. While internet-based mindfulness-based interventions primarily foster the awareness and acceptance of emotional distress, IPM+ is anchored in structured problem-solving—a framework that directly targets the concrete, real-world challenges inherent to the perinatal period, such as role transitions and shifting relationship dynamics. By equipping women to actively resolve these tangible stressors, IPM+ appears mechanistically better aligned with the needs of this population, ultimately driving the larger effect sizes observed in our study.

Furthermore, although IPT is well established as an effective treatment, IPM+ offers notable advantages in scalability. A systematic review by Couch et al [38] showed that IPT relies almost exclusively on specialist providers and typically requires 8‐14 sessions, limiting its large-scale implementation in resource-constrained settings. By contrast, the structured 5-week IPM+ program can be delivered by nonspecialist providers, substantially lowering the threshold for dissemination of depression interventions.

The Intervention Effect of IPM+ on Anxiety, Perceived Stress, and Sleep

This study further demonstrates that the IPM+ intervention exerted significant and sustained positive effects on anxiety and perceived stress among pregnant women, supporting its effectiveness as a transdiagnostic intervention recommended by the WHO. The intervention group showed sharp reductions in anxiety (PAQ) and stress scores immediately after the intervention (T2), with effects maintained at a moderate magnitude at the 3-month follow-up (T3). This pattern of rapid onset and durable improvement is highly consistent with the findings from face-to-face PM+ trials reported by Dawson et al [39] and Jordans et al [40], which showed superiority over usual care, providing strong evidence that the therapeutic effects of PM+ remain robust when extended to digital delivery formats. The marked improvements in anxiety and stress are likely attributable to the synergistic effects of the stress management and problem-solving modules, which may confer advantages over mindfulness-based interventions that primarily emphasize emotional awareness [41] or iCBT approaches that focus on cognitive restructuring [42] when addressing the unique stressors of the perinatal period. The stress management module offers relaxation techniques aimed at regulating heightened physiological arousal, consistent with established principles of relaxation training for anxiety management [11]. Importantly, the problem-solving module guides participants through structured steps to transform diffuse anxiety-provoking worries—such as excessive concerns about fetal health—into concrete, manageable problems with actionable plans. This process may directly target core cognitive maintenance factors of anxiety and stress, including intolerance of uncertainty and avoidance. The sustained reduction in stress observed at follow-up may further reflect the continued application of problem-solving skills in daily life, enhancing perceived control, alongside the facilitative role of the social support module in mobilizing available resources.

Sleep quality exhibited a distinct pattern of delayed improvement in our study, which contrasts with the findings from a CBT for insomnia study on perinatal insomnia by Quin et al [43], where interventions directly targeting sleep-related behaviors typically led to immediate postintervention improvements. The therapeutic pathway of PM+ may involve initial reductions in daytime emotional burden and cognitive rumination through the stress management and problem-solving modules, leading to decreased nocturnal psychophysiological arousal. Building on this foundation, the behavioral activation module may contribute to more regular daily routines and increased daytime activity, thereby supporting circadian rhythm stabilization, while the social support module may alleviate

XSL•FO
RenderX

feelings of loneliness and burden during pregnancy by strengthening social connectedness. Together, these changes form a gradual cycle of "reduced daytime emotional distress → enhanced nighttime calmness → progressive establishment of healthier sleep patterns." This cycle likely requires several weeks to consolidate into measurable sleep improvements, providing a plausible explanation for why sleep-related effects were not fully evident immediately postintervention (T2) but became significant at follow-up (T3). This finding aligns with emerging evidence highlighting emotion regulation as a key mechanism underlying improvements in comorbid insomnia [44].

Limitations and Suggestions

This study has several limitations that should be acknowledged. First, the sample size is relatively small (n=80) and of limited representativeness. Participants were recruited from a single tertiary hospital in Wuhan, where the levels of education and access to health care are generally high. Therefore, the generalizability of the study findings to the maternal population in rural areas or primary health care institutions may be limited. Future studies should adopt multicenter designs with larger, more diverse samples that include women from different geographic regions, health care tiers, and socioeconomic backgrounds to better evaluate the generalizability of IPM+. Second, we did not assess participants' blinding success after the intervention or examine whether group guessing influenced self-reported outcomes. Future studies could systematically evaluate participants' group guesses after the intervention and include this as a covariate in analyses to control for potential expectation effects or reporting bias. Third, symptom outcomes were primarily measured using self-report measures without confirmation via structured clinical interviews based on diagnostic criteria. Combining self-report scales with standardized diagnostic interviews in future work would help mitigate subjective bias and strengthen clinical validity. Finally, the 3-month follow-up period limits conclusions regarding the longer-term impact of IPM+ on postnatal mental health and mother-infant outcomes. Future studies with extended follow-up windows are needed and should include more ecologically meaningful endpoints, such as maternal-infant attachment quality and infant emotional-behavioral development.

This study presents a scalable framework for integrating perinatal mental health support into China's existing health care infrastructure. In the future, the core modules of IPM+ could be integrated into hospital-based official online service platforms as a universal and preventive psychological support tool for all pregnant women, facilitating early identification and timely intervention. In addition, the task-shifting potential of PM+ could be further leveraged through standardized training and supervision of maternal and child health providers in community health centers, enabling the delivery of preliminary yet fidelity-consistent psychological support at the primary care level. Such an approach may help alleviate the imbalance between the high demand for perinatal mental health services and the limited availability of specialist providers.

Conclusion

The findings from this randomized controlled trial indicate that the IPM+ intervention significantly alleviated depressive and anxiety symptoms, reduced perceived stress, and improved sleep quality among pregnant women, with benefits sustained at the 3-month follow-up. By employing structured modules that simultaneously address perinatal-specific stressors and internal emotional distress, IPM+ establishes a targeted and practical support framework. These results support IPM+ as an effective, sustainable, and scalable approach to promoting perinatal mental health. Future implementation should focus on integrating IPM+ into routine antenatal care pathways to enhance accessibility and impact.

Acknowledgments

The authors sincerely appreciate the participants and the support from the medical staff of Renmin Hospital of Wuhan University.

Funding

This study was supported by the National Social Science Foundation of China (Project No. 21BSH073).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

SZ, SW, and XS contributed to conceptualization, methodology, and investigation, with SZ and SW drafting the manuscript and XS verifying results. J Zhao and YL were responsible for data curation. SZ prepared figures and tables. QH, SL, and YC conducted formal analysis and visualization. The study was supervised by ZZ, XC, and J Zheng. Specifically, ZZ as the methodological lead led the intervention adaptation and data analysis; XC as the clinical lead oversaw trial design, ethics, execution, and funding acquisition; and J Zheng contributed to methodology design and provided critical intellectual revision. All authors reviewed and approved the final manuscript. ZZ and XC contributed equally as the corresponding authors of this manuscript. XC can be reached via email at chenxl72@whu.edu.cn.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
The problem management plus (PM+) group intervention program.
[DOCX File, 20 KB - jmir_v28i1e81998_app1.docx]

Multimedia Appendix 2
Problem management steps.
[DOC File, 23 KB - jmir_v28i1e81998_app2.doc]

Multimedia Appendix 3
Fixed effect estimation results of anxiety, perceived stress, and sleep quality; Comparison of analysis results before and after multiple imputation.
[DOCX File, 26 KB - jmir_v28i1e81998_app3.docx]

Checklist 1
CONSORT checklist.
[DOCX File, 38 KB - jmir_v28i1e81998_app4.docx]

References

1. Dagher RK, Bruckheim HE, Colpe LJ, Edwards E, White DB. Perinatal depression: challenges and opportunities. J Womens Health (Larchmt) 2021 Feb;30(2):154-159. [doi: 10.1089/jwh.2020.8862] [Medline: 33156730]
2. Yin X, Sun N, Jiang N, et al. Prevalence and associated factors of antenatal depression: systematic reviews and meta-analyses. Clin Psychol Rev 2021 Feb;83:101932. [doi: 10.1016/j.cpr.2020.101932] [Medline: 33176244]
3. Nisar A, Yin J, Waqas A, et al. Prevalence of perinatal depression and its determinants in mainland China: a systematic review and meta-analysis. J Affect Disord 2020 Dec 1;277:1022-1037. [doi: 10.1016/j.jad.2020.07.046] [Medline: 33065811]
4. Bränn E, Shen Q, Lu D. Perinatal depression and its health impact. BMJ 2024 Jan 9;384:2777. [doi: 10.1136/bmj.p2777] [Medline: 38195147]
5. Yu H, Shen Q, Bränn E, et al. Perinatal depression and risk of suicidal behavior. JAMA Netw Open 2024 Jan 2;7(1):e2350897. [doi: 10.1001/jamanetworkopen.2023.50897] [Medline: 38194232]
6. Slomian J, Honvo G, Emonts P, Reginster JY, Bruyère O. Consequences of maternal postpartum depression: a systematic review of maternal and infant outcomes. Womens Health (Lond Engl) 2019 Jan;15:1745506519844044. [doi: 10.1177/1745506519844044]
7. Ghimire U, Papabathini SS, Kawuki J, Obore N, Musa TH. Depression during pregnancy and the risk of low birth weight, preterm birth and intrauterine growth restriction—an updated meta-analysis. Early Hum Dev 2021 Jan;152:105243. [doi: 10.1016/j.earlhumdev.2020.105243] [Medline: 33190020]
8. Simonovich SD, Nidey NL, Gavin AR, et al. Meta-analysis of antenatal depression and adverse birth outcomes In US populations, 2010-20. Health Aff (Millwood) 2021 Oct;40(10):1560-1565. [doi: 10.1377/hlthaff.2021.00801] [Medline: 34606360]
9. Adeoye IA, Sogbesan A, Esan O. Prevalence, associated factors and perinatal outcomes of antepartum depression in Ibadan Nigeria. BMC Pregnancy Childbirth 2022 Mar 18;22(1):219. [doi: 10.1186/s12884-022-04549-7] [Medline: 35303804]
10. Molenaar NM, Kamperman AM, Boyce P, Bergink V. Guidelines on treatment of perinatal depression with antidepressants: an international review. Aust N Z J Psychiatry 2018 Apr;52(4):320-327. [doi: 10.1177/0004867418762057] [Medline: 29506399]
11. Domínguez-Solís E, Lima-Serrano M, Lima-Rodríguez JS. Non-pharmacological interventions to reduce anxiety in pregnancy, labour and postpartum: a systematic review. Midwifery 2021 Nov;102:103126. [doi: 10.1016/j.midw.2021.103126] [Medline: 34464836]
12. Cuijpers P, Franco P, Ciharova M, et al. Psychological treatment of perinatal depression: a meta-analysis. Psychol Med 2023 Apr;53(6):2596-2608. [doi: 10.1017/S0033291721004529] [Medline: 37310303]
13. Liu H, Yang Y. Effects of a psychological nursing intervention on prevention of anxiety and depression in the postpartum period: a randomized controlled trial. Ann Gen Psychiatry 2021 Jan 4;20(1):2. [doi: 10.1186/s12991-020-00320-4] [Medline: 33397393]
14. Pan WL, Lin LC, Kuo LY, Chiu MJ, Ling PY. Effects of a prenatal mindfulness program on longitudinal changes in stress, anxiety, depression, and mother-infant bonding of women with a tendency to perinatal mood and anxiety disorder: a randomized controlled trial. BMC Pregnancy Childbirth 2023 Jul 31;23(1):547. [doi: 10.1186/s12884-023-05873-2] [Medline: 37525110]
15. Pettman D, O'Mahen H, Blomberg O, Svanberg AS, von Essen L, Woodford J. Effectiveness of cognitive behavioural therapy-based interventions for maternal perinatal depression: a systematic review and meta-analysis. BMC Psychiatry 2023 Mar 29;23(1):208. [doi: 10.1186/s12888-023-04547-9] [Medline: 36991389]

XSL•FO
RenderX

16. Boele FW, Klein M, Verdonck-de Leeuw IM, et al. Internet-based guided self-help for glioma patients with depressive symptoms: a randomized controlled trial. J Neurooncol 2018 Mar;137(1):191-203. [doi: 10.1007/s11060-017-2712-5] [Medline: 29236238]
17. Cacciola E, Psouni E. Insecure attachment and other help-seeking barriers among women depressed postpartum. Int J Environ Res Public Health 2020 May 30;17(11):3887. [doi: 10.3390/ijerph17113887] [Medline: 32486285]
18. Dawson KS, Bryant RA, Harper M, et al. Problem Management Plus (PM+): a WHO transdiagnostic psychological intervention for common mental health problems. World Psychiatry 2015 Oct;14(3):354-357. [doi: 10.1002/wps.20255] [Medline: 26407793]
19. Problem management plus (PM+): individual psychological help for adults impaired by distress in communities exposed to adversity. : World Health Organization; 2016 URL: https://iris.who.int/server/api/core/bitstreams/0d493975-285d-4d0b-86b1-2365cfe45748/content [accessed 2026-01-23]
20. Hamdani SU, Huma ZE, Masood A, et al. Effect of adding a psychological intervention to routine care of common mental disorders in a specialized mental healthcare facility in Pakistan: a randomized controlled trial. Int J Ment Health Syst 2021 Jan 19;15(1):11. [doi: 10.1186/s13033-020-00434-y] [Medline: 33468192]
21. Zhang H, Zhang D, Lin H, Huang L. Problem management plus in the treatment of mental disorders in patients with multiple myeloma. Support Care Cancer 2020 Oct;28(10):4721-4727. [doi: 10.1007/s00520-019-05289-9] [Medline: 31970512]
22. de Graaff AM, Cuijpers P, Elsawy M, et al. The effectiveness of Problem Management Plus at 1-year follow-up for Syrian refugees in a high-income setting. Epidemiol Psychiatr Sci 2024 Oct 25;33:e50. [doi: 10.1017/S2045796024000519] [Medline: 39449610]
23. Coleman SF, Mukasakindi H, Rose AL, et al. Adapting problem management plus for implementation: lessons learned from public sector settings across Rwanda, Peru, Mexico and Malawi. Intervention (Amstelveen) 2021;19(1):58-66. [Medline: 34642580]
24. Danaher BG, Seeley JR, Silver RK, et al. Trial of a patient-directed eHealth program to ameliorate perinatal depression: the MomMoodBooster2 practical effectiveness study. Am J Obstet Gynecol 2023 Apr;228(4):453. [doi: 10.1016/j.ajog.2022.09.027] [Medline: 36174746]
25. Problem management plus (PM+): individual psychological help for adults impaired by distress in communities exposed to adversity [Report in Chinese]. : CITIC Press Group; 2022 URL: https://iris.who.int/server/api/core/bitstreams/54198a49-f09e-4307-ac75-ff2353bf42ff/content [accessed 2026-02-05]
26. Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression: development of the 10-item Edinburgh Postnatal Depression Scale. Br J Psychiatry 1987 Jun;150:782-786. [doi: 10.1192/bjp.150.6.782] [Medline: 3651732]
27. Lee DT, Yip SK, Chiu HF, et al. Detecting postnatal depression in Chinese women: validation of the Chinese version of the Edinburgh Postnatal Depression Scale. Br J Psychiatry 1998 May;172:433-437. [doi: 10.1192/bjp.172.5.433] [Medline: 9747407]
28. Cox J. Thirty years with the Edinburgh Postnatal Depression Scale: voices from the past and recommendations for the future. Br J Psychiatry 2019 Mar;214(3):127-129. [doi: 10.1192/bjp.2018.245] [Medline: 30774059]
29. Xiao L. Development of the pregnancy-specific anxiety questionnaire and research on pregnancy-specific anxiety in early pregnancy [Master's thesis in Chinese]. : Southern Medical University; 2010 URL: https://d.wanfangdata.com.cn/thesis/Ch1UaGVzaXNOZXdTb2xyOVMyMDI2MDExNzA4NTkxNhIHRDEyODgyNBoIcnEyMzd6amg%25253D [accessed 2026-02-05]
30. Leung DY, Lam TH, Chan SS. Three versions of Perceived Stress Scale: validation in a sample of Chinese cardiac patients who smoke. BMC Public Health 2010 Aug 25;10:513. [doi: 10.1186/1471-2458-10-513] [Medline: 20735860]
31. Staneva A, Bogossian F, Pritchard M, Wittkowski A. The effects of maternal depression, anxiety, and perceived stress during pregnancy on preterm birth: a systematic review. Women Birth 2015 Sep;28(3):179-193. [doi: 10.1016/j.wombi.2015.02.003] [Medline: 25765470]
32. Morin CM. Insomnia: Psychological Assessment and Management: Guilford Press; 1996. URL: https://search.worldcat.org/zh-cn/title/27385461 [accessed 2026-02-05]
33. Ansaari N, Rajan SK, Kuruveettissery S. Efficacy of in-person versus digital mental health interventions for postpartum depression: meta-analysis of randomized controlled trials. J Reprod Infant Psychol 2025 Jul;43(4):1067-1087. [doi: 10.1080/02646838.2024.2303470] [Medline: 38221725]
34. Rahman A, Khan MN, Hamdani SU, et al. Effectiveness of a brief group psychological intervention for women in a post-conflict setting in Pakistan: a single-blind, cluster, randomised controlled trial. The Lancet 2019 Apr;393(10182):1733-1744. [doi: 10.1016/S0140-6736(18)32343-2]
35. Fink G, Melero-Dominguez M, Chembe M, et al. Feasibility and acceptability of the Problem Management for Moms programme for improving maternal mental health in Zambia: an open-label trial. Lancet Psychiatry 2024 Dec;11(12):965-974. [doi: 10.1016/S2215-0366(24)00256-6] [Medline: 39515361]
36. Pan J, Luo W, Zhang H, et al. The effects of online cognitive behavioral therapy on postpartum depression: a systematic review and meta-analysis. Healthcare (Basel) 2025 Mar 21;13(7):696. [doi: 10.3390/healthcare13070696] [Medline: 40217993]

XSL•FO
RenderX

37. Kim M, Park JY, Park S, et al. Effectiveness of a mindfulness-based mobile intervention for improving perinatal mental health and reducing depression during pregnancy: randomized controlled trial. J Med Internet Res 2025 Oct 2;27:e75630. [doi: 10.2196/75630] [Medline: 41037777]
38. Couch E, Mai HJ, Kanaan G, et al. Effectiveness of psychological therapies for depression during the perinatal period: a systematic review and meta-analysis. Ann Intern Med 2025 Dec;178(12):1752-1762. [doi: 10.7326/ANNALS-24-03520] [Medline: 41183344]
39. Dawson KS, Schafer A, Anjuri D, et al. Feasibility trial of a scalable psychological intervention for women affected by urban adversity and gender-based violence in Nairobi. BMC Psychiatry 2016 Nov 18;16(1):410. [doi: 10.1186/s12888-016-1117-x] [Medline: 27863515]
40. Jordans MJD, Kohrt BA, Sangraula M, et al. Effectiveness of Group Problem Management Plus, a brief psychological intervention for adults affected by humanitarian disasters in Nepal: a cluster randomized controlled trial. PLoS Med 2021 Jun;18(6):e1003621. [doi: 10.1371/journal.pmed.1003621] [Medline: 34138875]
41. Mao F, Sun Y, Li Y, Cui N, Cao F. Internet-delivered mindfulness-based interventions for mental health outcomes among perinatal women: a systematic review. Asian J Psychiatr 2023 Feb;80:103321. [doi: 10.1016/j.ajp.2022.103321] [Medline: 36423436]
42. Mahoney A, Shiner CT, Grierson AB, et al. Online cognitive behaviour therapy for maternal antenatal and postnatal anxiety and depression in routine care. J Affect Disord 2023 Oct 1;338:121-128. [doi: 10.1016/j.jad.2023.06.008] [Medline: 37295653]
43. Quin N, Tikotzky L, Astbury L, et al. Preventing postpartum insomnia: findings from a three-arm randomized-controlled trial of cognitive behavioral therapy for insomnia, a responsive bassinet, and sleep hygiene. Sleep 2024 Aug 14;47(8):zsae106. [doi: 10.1093/sleep/zsae106] [Medline: 38736364]
44. Meneo D, Samea F, Tahmasian M, Baglioni C. The emotional component of insomnia disorder: a focus on emotion regulation and affect dynamics in relation to sleep quality and insomnia. J Sleep Res 2023 Dec;32(6):e13983. [doi: 10.1111/jsr.13983] [Medline: 37394234]

Abbreviations

AND: antenatal depression
CBT: cognitive behavioral therapy
CONSORT: Consolidated Standards of Reporting Trials
EPDS: Edinburgh Postnatal Depression Scale
iCBT: internet-based cognitive behavioral therapy
IPM+: internet-based problem management plus
IPT: interpersonal psychotherapy
ISI: Insomnia Severity Index
PAQ: Pregnancy Anxiety Questionnaire
PM+: problem management plus
PSS: 4-item Perceived Stress Scale
WHO: World Health Organization

Edited by A Schwartz, M Balcarras; submitted 07.Aug.2025; peer-reviewed by S Alves, X Xiao; accepted 08.Jan.2026; published 27.Mar.2026.

Please cite as:
Zheng S, Wang S, Shui X, Zhao J, Lin Y, Huang Q, Li S, Chen Y, Zou Z, Zheng J, Chen X
Internet-Based Problem Management Plus Intervention for Antenatal Depression: Randomized Controlled Trial
J Med Internet Res 2026;28:e81998
URL: https://www.jmir.org/2026/1/e81998
doi:10.2196/81998

© Shuanghui Zheng, Shiyu Wang, Xiaoyu Shui, Jing Zhao, Ying Lin, Qian Huang, Siqi Li, Yang Chen, Zhijie Zou, Jin Zheng, Xiaoli Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

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

Luisa R Blanco[1], MBA, PhD; Alexandra Klomhaus[2], MS, PhD; Rebecca Dudovitz[3], MSHS, MD; Denise Marquez[4], MPH; Ray Lopez-Chang[5], MUP; Evelyn Aleman[6], MPP; Bas Weerman[5]; Michael Moldoff[5], MA; Keith Norris[2], MD, PhD; Yelba Castellon-Lopez[7], MS, MD

[1]School of Public Policy, Pepperdine University, 24255 Pacific Coast Highway, Malibu, CA, United States

[2]Department of Medicine, David Geffen School of Medicine, Division of General Internal Medicine and Health Services Research, University of California, Los Angeles, Los Angeles, CA, United States

[3]Department of Pediatrics, David Geffen School of Medicine, Division of General Pediatrics, University of California, Los Angeles, Los Angeles, CA, United States

[4]Self-Sufficiency Services, Office of Immigrant and Refugee Affairs, County of San Diego, Health and Human Services Agency, San Diego, CA, United States

[5]Center for Economic and Social Research, USC Dornsife, University of Southern California, Los Angeles, CA, United States

[6]Our Voice: Communities for Quality Education (Nuestra Voz), Los Angeles, CA, United States

[7]Cancer Research Center for Health Equity, Samuel Oschin Comprehensive Cancer Institute, Department of Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA, United States

Corresponding Author:
Luisa R Blanco, MBA, PhD
School of Public Policy, Pepperdine University, 24255 Pacific Coast Highway, Malibu, CA, United States

Abstract

Background: Early in the children's COVID-19 rollout in the United States, racial and ethnic vaccination rate disparities were evident. Based on COVID-19 communication literature and qualitative interviews with Hispanic parents, we developed a mobile phone–delivered digital intervention to address factors associated with low vaccine confidence.

Objective: We conducted a community-based randomized controlled trial of a digital intervention called MiVacunaLA/MyShotLA to increase COVID-19 vaccine uptake among Hispanic children. The fully automated digital intervention was designed in collaboration with community organizations and linguistically and culturally tailored to meet the informational needs of Hispanic caregivers. The intervention focused on families with unvaccinated children 5 to 11 years old but was offered to families with any unvaccinated children 17 years or younger.

Methods: Participants were recruited with community organization partners and trained parent ambassadors via an open online screener. The 4-week intervention consisted of 3 SMS text messages with culturally and linguistically tailored educational information weekly. Intervention materials were delivered digitally through a closed online platform. Study team members were blinded. We used a difference-in-difference model with an intention-to-treat approach. The primary outcome was self-reported COVID-19 vaccine uptake among household children collected via online questionnaires. Secondary outcomes included COVID-19 vaccine knowledge, vaccine trust, and measures of participant engagement. We conducted a sensitivity analysis using the treatment-on-the-treated approach.

Results: In total, 254 participants completed the baseline survey (128 control and 126 intervention). The average participant age was 34 (SD 6.3) years with an average of 1.7 (SD 0.8) minors in the household, and among households, 62.2% (n=158) reported having children aged 5 to 11 years old. Most participants (n=207, 81.5%) reported English as their primary language. We found a statistically significant difference of 13.3% (95% CI 0.3%-26.4%; *P*=.04) points in self-reported vaccine uptake between intervention and control groups among caregivers of Hispanic children aged 5 to 11 years old. We also found a statistically significant point difference of 14.3% (95% CI 0% - 23.7%; *P*=.003) between intervention and control groups in trust of governmental approval processes for the children's COVID-19 vaccine. Most participants reported that the weekly digital videos and educational information were "very" (892/1031, 86.5%) or "extremely" (888/1019, 87.1%) useful.

Conclusions: MiVacunaLA demonstrates that a culturally tailored, community-based, mobile phone–delivered vaccine educational intervention can increase COVID-19 vaccine uptake among Hispanic children and improve caregivers' trust in governmental vaccine processes. MiVacunaLA is innovative in its integration of community-informed design with a fully automated, mobile phone–centric format and builds on prior literature by prospectively evaluating a culturally tailored, SMS text messaging–linked

XSL•FO
RenderX

web curriculum in a community setting. Findings provide evidence that scalable, low-cost, digital strategies can measurably improve trust and uptake in a population facing persistent vaccination gaps. Real-world implications include the portability and adaptability of this approach across diverse communities and settings to support timely, community-engaged vaccination efforts for broader applicability and scalability in public health.

Trial Registration: ClinicalTrials.gov NCT05234372; https://clinicaltrials.gov/study/NCT05234372

(J Med Internet Res 2026;28:e78103) doi:10.2196/78103

KEYWORDS

children digital education; Latino/Hispanic children; health equity; COVID-19 vaccines; mobile-based interventions

Introduction

Background

During the initial rollout of the COVID-19 vaccine for children of ages 5 to 11 years in the United States, children identifying as part of a racial and ethnic minority group were underrepresented among first-dose recipients [1-5]. This was due, in part, to lack of access to accurate information, concerns about vaccine safety and efficacy, exposure to disinformation and misinformation, and lack of trust in governmental vaccine processes [5-8]. Studies show that parental and caregiver acceptance of COVID-19 vaccination for their children is positively associated with beliefs about vaccination benefits [9]. Improving parental and caregiver acceptance is a key public health focus, given the trends in childhood vaccine uptake after the COVID-19 pandemic.

Following the pandemic, vaccination rates among US kindergartners across other childhood vaccinations declined compared to prior years [10,11]. Meanwhile, the United States has seen the resurgence of previously eliminated or highly controlled vaccine-preventable diseases, such as measles-mumps-rubella (MMR) [12,13]. While prepandemic vaccination rate disparities existed among racial and ethnic subgroups [14], these trends were further exacerbated by the pandemic [15]. As such, novel interventions are needed to effectively combat disinformation and misinformation, rebuild trust in vaccination benefits and safety, and engage undervaccinated communities.

In the early stages of the COVID-19 vaccine rollout, leveraging community-engaged strategies aided in promoting vaccine confidence and addressing barriers such as language accessibility, addressing disinformation and misinformation, and building trust in vaccination among the Hispanic population [16,17]. Incorporating community voices and perspectives into vaccine promotion and distribution strategies can have more meaningful impact with community members as opposed to traditional communication methods [18,19]. Digital interventions have also increased uptake of other vaccines, such as human papillomavirus vaccination, suggesting that community-based digital interventions might be particularly effective at increasing vaccine confidence and uptake to address recent declines in children vaccination rates [20,21]. However, few randomized controlled trial (RCT) studies exist on the use of community-based digital interventions to improve COVID-19 vaccination rates among Hispanic children. Testing and identifying such interventions can yield critical public health and primary care approaches and lessons for responding to both current vaccine uptake disparities and future public health emergencies.

Study Objectives

This study describes efforts to increase COVID-19 vaccination rates among Hispanic children of ages 5 to 11 years in the Los Angeles County area through a community-informed culturally and linguistically appropriate RCT. We evaluated the efficacy of a digital intervention, called MiVacunaLA/MyShotLA (MVLA) 2.0, for parents or caregivers of Hispanic children to increase COVID-19 vaccination uptake for children. We also evaluated the intervention's efficacy in increasing knowledge of COVID-19 vaccines as well as trust in the governmental approval process for the vaccine.

Methods

Intervention Content and Approach

MVLA 2.0 builds on MVLA 1.0, which was created in response to the US Food and Drug Administration–approved COVID-19 vaccines for children of ages 12 to 17 years and COVID-19 communication literature [22]. The MVLA 1.0 design and implementation are described in detail elsewhere [22]. Briefly, MVLA 1.0 used a community-partnered approach to develop and distribute COVID-19 vaccine education via SMS text messaging, emails, and brief videos multiple times a week for 4 weeks. MVLA 2.0, the focus of this study, began recruitment between June and August of 2022, with the intervention implemented between August and November 2022.

For MVLA 2.0, we updated and refined the digital materials to focus on vaccination for children aged 5 to 11 years based on parental and community partner feedback. We conducted focus groups with parents and caregivers of children between the ages of 5 and 11 years from MVLA 1.0 to tailor MVLA 2.0 for the 5- to 11-year-old age group. Detailed focus group results are reported in a separate paper [23] and revealed the need to adapt the intervention's educational content to address emerging COVID-19 vaccine concerns specific to children, as well as myths and misinformation [23]. Next, we collaborated with community-based organizations to conduct community sessions with parent advocates from these organizations. Participants suggested several new features to MVLA that included (1) infographics to visually summarize weekly content, (2) voice-overs in English and Spanish of text to address literacy needs, (3) a discussion board where participants could post questions about the vaccine and have a physician respond, and

XSL•FO
RenderX

(4) parent testimonials from those who had vaccinated their children. In addition, community partners identified the need for new content to address questions about emerging COVID strains. These suggestions were all incorporated into the MVLA 2.0 intervention.

For the MVLA 2.0 intervention, participants received 3 SMS text messaging or emails per week (Monday, Wednesday, and Thursday) inviting them to complete a specific activity. On Mondays, participants were invited to watch a relevant informational video for that week. On Wednesdays, participants were invited to visit the platform to read additional relevant educational information. On Thursdays, participants received a summary infographic of the week's material. Figures S1 and S2 in Multimedia Appendix 1 provide examples of the infographics we used for week 3 (English and Spanish), and Table S1 in Multimedia Appendix 1 provides an outline of the intervention material and activities.

To deliver MVLA 2.0, we utilized the digital infrastructure of the Understanding America Study (UAS) platform. We created a participant-only website to provide access to weekly study content. The online platform for MVLA 2.0 included all study information, a discussion board, and downloadable infographics summarizing each week's COVID-19 topic. Participants could select their preferred language (English or Spanish) to engage with the website and all study materials. Table S2 in Multimedia Appendix 1 shows a detailed description of the MVLA 2.0 intervention components (visual summaries, audio content, discussion board, and testimonial videos).

Community Engagement, Recruitment, Timeline, and Other Intervention Logistics

We applied community-partnered research methods throughout the study. Recruitment involved community organization leaders and a group of parent ambassadors—parents who completed MVLA 1.0 and expressed interest in continuing to raise awareness by aiding with recruitment for MVLA 2.0. Recruitment was conducted through public school serving organizations located in East and South Los Angeles, communities with a high proportion of Hispanic families and disproportionately lower COVID-19 vaccination rates in children and youth based on Los Angeles County Department of Public Health Data [24]. We recruited participants using word-of-mouth and community partner social media sites to share the project flyer within their networks. The following organizations supported recruitment efforts by sharing our study flyer: Best Start LA, Coalition for Humane Immigrant Rights of Los Angeles, Communities in School, Families in School, Inner City Struggle, Great Public Schools Now, KIPP Public Charter Schools, Latino Equality Alliance, Mexican American Legal Defense and Education Fund, Mexican American Opportunity Foundation, Nuestra Voz (Our Voice), and Young Men's Christian Association of Metropolitan Los Angeles.

We planned to start recruitment when vaccines for children of ages 5 to 11 were newly released. Therefore, we tailored the intervention to address informational needs of Hispanic parents or caregivers of children in this age group. While the intervention information was relevant across all age groups, recruitment and intervention materials focused on households with children of ages 5 to 11 years as we anticipated the greatest impact among this age group. COVID-19 vaccines for children younger than 5 years were released in June 2022, and therefore, we updated intervention content to reflect up-to-date vaccine approval for the younger group (see Figures S1 and S2 in Multimedia Appendix 1).

We recruited parents and caregivers of Hispanic minors below 18 years of age who had access to a computer or cell phone to receive the intervention (inclusion criteria) via an online access open screener. Recruitment efforts were targeted not only to reach families with unvaccinated children of ages 5 to 11 years but also allowed interested families with at least one child below 18 years to participate in the intervention. While recruitment was focused among Los Angeles County parents, participants were not excluded if they resided in other neighboring areas or were from other racial or ethnic groups (no exclusion criteria).

We ended recruitment 2 weeks prior to starting the intervention to allow time to randomize all participants and enter their information for receiving SMS text messaging and emails into the digital platform. Figure S3 in Multimedia Appendix 1 shows the intervention implementation timeline. Participants first filled out an online screening survey to determine eligibility. Qualifying participants received an SMS text messaging and email inviting them to participate and directing them to the electronic consent. Upon consenting, participants were directed to the baseline survey. We did not establish specific eligibility criteria for sites or for individuals delivering the intervention, as the intervention was delivered digitally.

We used separate block randomization for cohorts, conducted by one UAS staff using Stata software, version 17 (StataCorp LLC). Randomized participants received MVLA 2.0 either at month 1 (intervention) or at month 2 (control). We used a 1:1 allocation ratio for randomization to the intervention and control groups. All participants were invited to complete a baseline and a 1-month follow-up survey. In the first month, the intervention group received educational material for 4 weeks, while the control group received a biweekly message reminding them of the number of days left until the beginning of month 2 when they were scheduled to start MiVacunaLA. After the intervention group completed the 4-week educational intervention, we sent all participants reminders twice a week for 2 weeks to complete the 1-month follow-up survey. The control group received the intervention after all groups completed the follow-up survey.

All intervention materials were delivered digitally. Participants completed the intervention activities in order. If a participant did not complete an intervention activity by the end of the week, they still received content for the following week. However, when participants attempted to access the next activity, they were prompted to first complete the missed activity. Participants had the flexibility to complete all intervention activities by the end of the month, with an additional 1-week grace period before the 1-month follow-up survey became available. Once the survey was available, participants could not complete additional intervention activities.

XSL•FO
RenderX

Ethical Considerations

MiVacunaLA 2.0 was registered and approved by the University of California, Los Angeles Institutional Review Board (21 - 000857). Participants completed an online informed consent form in their preferred language. Consent information was also available to participants as a PDF on the intervention platform for their reference. Participants' contact information was collected in the screening survey. A study principal investigator worked with UAS staff to create identifiers so that all data downloaded from the UAS platform, and subsequently used in data analysis, were deidentified. Thus, no identification of individual participants in any images of the paper or supplementary material is possible. Only 1 principal investigator and 1 UAS staff member had access to participant identifiers, which were password-protected in an encrypted file, as well as access to pretrial randomization information. Study team members were blinded. After randomization, intervention and control participants started the intervention immediately (intervention arm) or in 2 months (waitlist control arm). Participants received a US $40 gift card sent electronically after completing all intervention activities.

Given this was a behavioral intervention and no concomitant care was provided during the intervention, there was minimal risk, and we did not collect data regarding harms. There were no significant changes to the trial after commencement or interim analysis during the data collection phase. We added information about US Food and Drug Administration–approved COVID-19 vaccines for children of ages 6 months to 4 years to the material during the design phase. During consent, participants were informed that they could decline participation at any time and could opt out of any intervention activity they did not feel comfortable completing. In addition to the detailed information provided on the curriculum and digital approach, those interested in accessing all the educational materials and survey documents to replicate this RCT may contact the corresponding author.

Primary Outcomes

The primary outcome was self-reported vaccination status of household children in 3 age groups based on COVID-19 vaccine availability pre- and postintervention. The age groups used were (1) 6 months to 4 years, (2) 5 to 11 years, and (3) 12 to 17 years. The primary outcome was analyzed at the household level with at least 1 child in the specified age groups. The baseline survey asked for vaccination status for all household children. When participants confirmed a specific child's vaccination, only the status of the household's remaining nonvaccinated children was queried in the follow-up survey. We imputed the same vaccination status for the household expressed at baseline and the 1-month follow-up survey. For those with missing data, we assumed that if a household vaccinated 1 child in a specific age group, then the household's children in that same age group were vaccinated at follow-up.

Table S3 in Multimedia Appendix 1 shows the vaccination status of qualifying households with children by age groups at baseline and follow-up for all participants, regardless of loss to follow-up. For those participants who did not complete the follow-up survey, we assumed that the vaccination status remained unchanged from baseline.

Secondary Outcomes

Secondary outcomes included perceived change in COVID-19 vaccine knowledge and trust from baseline and follow-up. For vaccine knowledge, we asked participants to indicate whether the following statement was true or false: "After being fully vaccinated with the COVID-19 vaccine, your chances of being hospitalized or dying from COVID-19 are reduced at least by 90% if you contract the virus." We coded the variable as equal to 1 if participants answered true, and 0 if they answered false or unsure. Second, we asked participants, "How would you rate your knowledge about the COVID-19 vaccine?" This variable was coded equal to 1 if participants answered adequate or superior knowledge, and equal to 0 if the answer was basic, minimal, or no knowledge.

We measured vaccine trust via 2 separate questions: "How much do you trust the governmental approval process to ensure the COVID-19 vaccine is safe for (1) the public and (2) children?" We coded this variable as equal to 1 for those who answered they trust government fully, mostly, or somewhat, and equal to 0 if they answered they do not trust the government for each question. We also measured participant engagement with the intervention regardless of the study arm. This analysis included all participants (intervention and waitlist control groups together) who interacted with the material on the digital platform and completed at least 1 activity. At the end of each week, we collected information on the perceived usefulness of the weekly video and in-platform information and participants' familiarity with the information. We also tracked participants' in-platform click-through rates of a link to the Los Angeles County Department of Public Health appointment portal for COVID-19 vaccination; whether participants downloaded infographics to share; and whether participants utilized audio files for written information and infographics sent midweek.

Statistical Analysis

We evaluated efficacy using an intention-to-treat (ITT) approach. For the ITT analysis, we imputed the vaccination status for those who did not complete the follow-up survey with the same value entered at baseline. This conservative approach assumes no behavioral change among participants who were lost to follow-up. We used the same approach for the secondary outcomes.

We used a treatment-on-the-treated (TOT) approach as a sensitivity analysis. For the TOT analysis, we only included participants who completed the follow-up survey and whose vaccination status was reported at follow-up. We used a difference-in-difference (DID) model estimation for ITT and TOT approaches. Table S4 in Multimedia Appendix 1 shows the description of the DID model specification used for the statistical analysis.

To fit these models, we used a generalized linear model, using maximum likelihood estimation, with a binomial distribution and an identity link function to fit vaccination status as the dependent variable, with main fixed effects for both month (0=baseline and 1=follow-up) and intervention arm (0=control

XSL•FO
RenderX

and 1=intervention), and an interaction term between month and intervention. The interaction term, defined specifically as Month×Intervention, represents the primary variable of interest and reflects the DID in the outcome variable. The models also accounted for repeated measures within the same household and utilized an unstructured correlation structure.

We ran Little's Missing Completely at Random (MCAR) test to evaluate whether our data were missing completely at random. Subsequently, for the TOT analysis, we evaluated whether statistically significant differences exist between those who completed follow-up surveys and those who did not. We then modeled the TOT data using estimated weights based on significantly different characteristics between the 2 groups. This approach was used to help ensure we were not overestimating efficacy based on the characteristics of those who completed follow-up surveys. We estimated a Probit model where we regressed the attrition-related variable on those statistically significant variables between the 2 groups and created weights inverting these probabilities to estimate the intervention's efficacy on COVID-19 vaccine uptake among household children within a specific age group. This RCT was reported using the CONSORT 2025 (Consolidated Standards of Reporting Trials) checklist (Checklist 1) [25].

Results

Demographic Characteristics for Overall, Control, and Intervention Samples

Figure 1 presents the CONSORT (Consolidated Standards of Reporting Trials) flow diagram for the intervention. Based on screening survey completion and study criteria, we invited 512 qualifying participants. Among those, 310 (60.5%) people consented to participate, completed the baseline survey, and were randomized to intervention or control groups. From those who consented and completed the initial baseline survey, we excluded 56 participants who reported no household minors in their baseline responses, which conflicted with their screener responses. Of the remaining 254 parents and caregivers in the baseline sample (126 and 128 participants in intervention and control groups, respectively), we had 216 participants who completed the 1-month follow-up survey (99 and 117 participants in intervention and control groups, respectively) with a retention rate of 85%. We had 17% missing data for the primary outcome or vaccine uptake. Of those, we imputed 37 for a total analytical sample of 247. We had 4% missing data for secondary outcomes or vaccine knowledge and trust. Our total analytic sample for secondary outcomes was 243.

Table 1 provides baseline demographic characteristics for the full sample and for control and intervention groups separately. The sample includes 254 participants with a household minor recruited at baseline. Most participants were Hispanic (n=230, 90.5%) and completed the intervention in English (n=207, 81.5%). Most participants had some college education or more (n=196, 77.2%), were employed (n=202, 79.5%), and over half (n=150, 59.1%) had an annual household income greater than US $50,000. Just over half had children aged 5 to 11 years (n=158, 62.2%), and about one-third had children aged 12 to 17 years (n=92, 36.2%). The average age of parents or caregivers was 33.8 (SD 6.3) years, and the average number of children in the household was 1.7 (SD 0.8). We evaluated whether differences in participant characteristics between study arms were statistically significant using a *t* test.

XSL•FO
RenderX

Figure 1. The CONSORT (Consolidated Standards of Reporting Trials) flow diagram for the intervention.

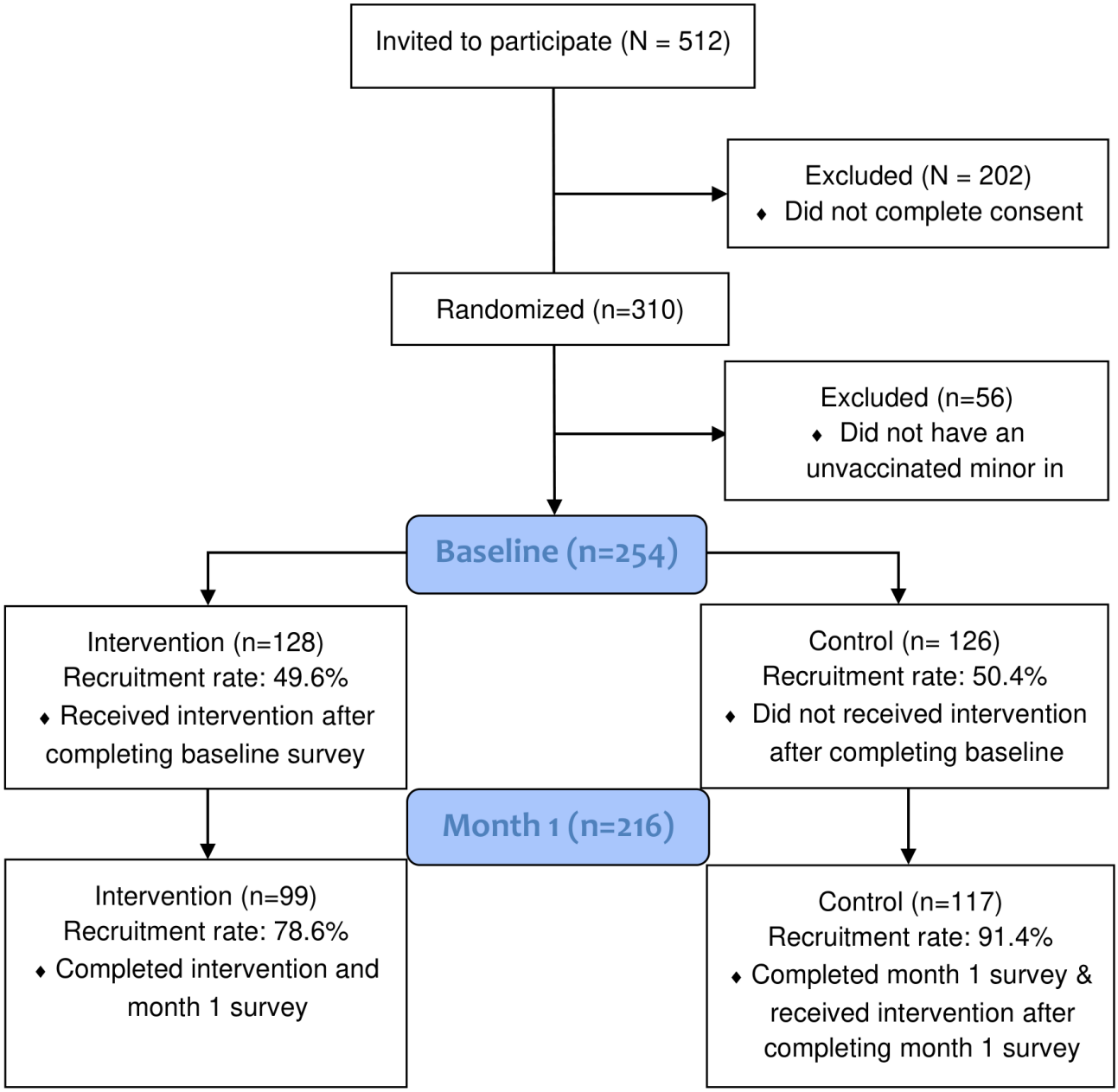

XSL•FO
RenderX

Table . Baseline demographic characteristics by overall, control, and intervention samples.

Characteristic	Overall (N=254)[a]	Control (n=128)	Intervention (n=126)	*P* value[b]
Age of parent, mean (SD)[c]	33.8 (6.3)	33.7 (6.2)	33.9 (6.4)	.68
Number of minors in household, mean (SD)	1.7 (0.8)	1.7 (0.9)	1.6 (0.7)	.82
Language, n (%)				.45
English	207 (81.5)	102 (79.7)	105 (83.3)	
Spanish	47 (18.5)	26 (20.3)	21 (16.7)	
Parent COVID-19 vaccination status, n (%)				.31
Vaccinated	95 (37.4)	51 (39.8)	44 (34.9)	
Not vaccinated	155 (61)	76 (59.4)	79 (62.7)	
Unsure/missing	4 (1.6)	1 (0.8)	3 (2.4)	
Ethnicity, n (%)				.89
Not Hispanic/Spanish origin or missing	24 (9.4)	11 (8.6)	13 (10.3)	
Mexican/Mexican American/Chicano	152 (59.8)	79 (61.7)	73 (57.9)	
Other Hispanic/Spanish origin[d]	78 (30.7)	38 (29.7)	40 (31.8)	
Born in the United States, n (%)				.18
Yes	187 (73.6)	99 (77.3)	88 (69.8)	
No	52 (20.5)	25 (19.5)	27 (21.4)	
Prefer not to respond/did not respond	15 (5.9)	4 (3.1)	11 (8.7)	
Highest education attained, n (%)				.99
Some high school or less	22 (8.7)	11 (8.6)	11 (8.7)	
High school graduate/GED[e]	30 (11.8)	15 (11.7)	15 (11.9)	
Some college or more	196 (77.2)	100 (78.1)	96 (76.2)	
Missing	6 (2.4)	2 (1.6)	4 (3.2)	
Employment status, n (%)				.10
Employed	202 (79.5)	107 (83.6)	95 (75.4)	
Unemployed	8 (3.2)	3 (2.3)	5 (4)	
Other[f]	44 (17.3)	18 (14.1)	26 (20.6)	
Household income, n (%)				.64
<$25,000	51 (20.1)	23 (18)	28 (22.2)	
$25,000-$49,000	46 (18.1)	25 (19.5)	21 (16.7)	
>$50,000	150 (59.1)	77 (60.2)	73 (57.9)	
Missing	7 (2.8)	3 (2.3)	4 (3.2)	
Health insurance status, n (%)				.24
Insured[g]	227 (89.4)	115 (89.8)	112 (88.9)	
Not insured/don’t know	27 (10.6)	13 (10.2)	14 (11.1)	

XSL•FO
RenderX

Characteristic	Overall (N=254)[a]	Control (n=128)	Intervention (n=126)	*P* value[b]
Marital status, n (%)				.40
Currently married	203 (79.9)	105 (82)	98 (77.8)	
Widowed/divorced/separated	20 (7.9)	9 (7)	11 (8.7)	
Never married	20 (7.9)	8 (6.3)	12 (9.5)	
Other[h]	11 (4.3)	6 (4.7)	5 (4)	
Type of household, n (%)				.61
Married with children	207 (81.5)	107 (83.6)	100 (79.4)	
Single/married without children	7 (2.8)	2 (1.6)	5 (4)	
Single with children	16 (6.3)	9 (7)	7 (5.6)	
Other[i]	24 (9.4)	10 (7.8)	14 (11.1)	
Any minors in household under 6 months of age, n (%)				
Yes[j]	8 (3.1)	5 (3.9)	3 (2.4)	.72
Any minors in household from 6 months to 4 years, n (%)				
Yes[j]	105 (41.3)	51 (39.8)	54 (42.9)	.63
Any minors in household from 5 to 11 years, n (%)				
Yes[j]	158 (62.2)	87 (68)	71 (56.4)	.06
Any minors in household from 12 to 17 years, n (%)				
Yes[j]	92 (36.2)	44 (34.4)	48 (38.1)	.54

[a]Excludes n=56 who reported no minors in their household at baseline.

[b]*P* value from χ^2 tests (or Fisher exact tests, when appropriate) for categorical variables and Wilcoxon tests for continuous variables.

[c]Control group: n=125 parents had nonmissing age; intervention group: n=117 had nonmissing age.

[d]Includes Puerto Rican, Cuban, multiple ethnicities, and "other".

[e]GED: general education development.

[f]Includes housekeeper, retired, disabled, temporary employment, student, and "other".

[g]Includes government insurance, insurance through the Veteran Affairs, private insurance, and Medicare.

[h]Includes missing and cohabitation (common law marriage).

[i]Includes don't know, prefer not to respond, missing, and "other".

[j]All other participants indicated "no".

There were no statistically significant differences between intervention and control groups in demographic characteristics. Table S3 in Multimedia Appendix 1 shows more intervention versus control households indicated at least 1 minor aged 5 to 11 years in their household (22.5% vs 19.5%) was vaccinated at baseline, while fewer intervention versus control households reported having at least 1 minor vaccinated at baseline in other age groups (5.6% vs 13.7% for minors 6 mo to 4 y; 27.1% vs 34.1% for minors 12 to 17 years old).

Table S5 in Multimedia Appendix 1 shows demographic characteristics at baseline for those completing the follow-up survey (216/254, 85%) and those who did not (38/254, 15%). Most significant demographic characteristics were associated with follow-up survey completion. Those who completed the follow-up survey were more likely to be unvaccinated, Hispanic, employed, US born, currently married or cohabitating, younger, with a smaller number of children at baseline. They were also more likely to have higher educational attainment and a household income greater than US $50,000. Table S5 also discusses weights considered for TOT estimation related to significant differences between follow-up completers and noncompleters.

DID Model With the ITT Approach

Estimates from the DID model using the ITT approach for the primary outcomes related to vaccination are shown in Table 2 for households with children in each age group individually and all together. Table 2 shows vaccination status estimates for control and intervention groups at baseline and at follow-up in the regression analysis setting, alongside estimates of the difference within and between groups at pre- and postintervention. The control and intervention groups had overall baseline vaccination rates of 25.7% and 15.7%, respectively. At follow-up, the overall vaccination rates were 44.2% and 48.4%, respectively, among these groups. The estimate of interest in Table 2 is the DID estimate shown as "difference."

XSL•FO
RenderX

We found a statistically significant difference between intervention and control groups of 13.3% points (95% CI 0.3%-26.4%; *P*=.04) among children aged 5 to 11 years in change from pre- to postintervention. We did not find a significant DID for other age groups. Given ITT as the conservative estimate of intervention efficacy, we concluded that the intervention had a positive effect on COVID-19 vaccine uptake among the primary target group for this intervention, children aged 5 to 11 years.

Table . Difference-in-difference (DID) regression estimates for primary outcomes with the intention-to-treat (ITT) approach: vaccination status (N=247)[a].

	Baseline, %	1-month follow-up, %	Change Δ (95% CI)	*P* value
6 months-4 years[b]				
Control	13.8	31.4	17.6 (7.0 to 28.2)	.001
Intervention	5.5	27.8	22.3 (11.1 to 33.5)	<.001
Difference	N/A[c]	N/A	4.7 (–10.7 to 20.1)	.55
5 - 11 years[b]				
Control	19.5	34.5	14.9 (7.5 to 22.4)	<.001
Intervention	22.4	50.7	28.3 (17.6 to 38.9)	<.001
Difference	N/A	N/A	13.3 (0.3 to 26.4)	.04
12 - 17 years[b]				
Control	34.1	50	15.9 (5.1 to 26.7)	.004
Intervention	27.1	47.9	20.8 (9.3 to 32.3)	<.001
Difference	N/A	N/A	4.9 (–10.9 to 20.7)	.54
All ages[b]				
Control	26.6	43.6	16.9 (10.3 to 23.6)	<.001
Intervention	21.9	46.3	24.4 (16.8 to 32.1)	<.001
Difference	N/A	N/A	7.5 (–2.6 to 17.7)	.15

[a]Simple (unadjusted) DIDs of vaccination of minors in household. Rates of vaccination are among those participants who indicated having at least 1 minor in the household within that age range and assume that anyone missing a 1-month follow-up had minors who remained unvaccinated. For those who reported a minor was vaccinated at baseline (and thus were not asked the vaccination question at 1 mo), we imputed the "yes" value to the 1-month follow-up, given that the question is not asked again at follow-up for this group. Models are estimated using a generalized linear model and maximum likelihood estimates.

[b]Sample sizes for each age group are as follows: n=105 households for 6 months-4 years; n=158 households for 5-11 years; n=92 households for 12 - 17 years.

[c]N/A: not applicable.

Estimates of the DID model with the ITT approach for secondary outcomes related to vaccine knowledge and trust are shown in Table 3. We found a statistically significant difference between the intervention and control groups of 14.3% points (95% CI 5.0% - 23.7%; *P*=.003) preintervention to postintervention for trust in the governmental approval process to ensure the COVID-19 vaccine is safe for children. We did not find significant effects of the intervention on other secondary outcomes.

XSL•FO
RenderX

Table . Difference-in-difference (DID) regression estimates for secondary outcomes with the intention-to-treat (ITT) approach: vaccine knowledge and trust (N=243)[a].

	Baseline, %	1-month follow-up, %	Change Δ (95% CI)	*P* value
...chances of being hospitalized or dying from COVID-19...[b]				
Control	52.9	65.1	12.2 (3.6 to 20.8)	.005
Intervention	50	65.1	15.0 (5.9 to 24.1)	.001
Difference	N/A[c]	N/A	2.8 (−9.8 to 15.4)	.66
...rate knowledge about COVID-19 vaccine...[d]				
Control	25.2	43.1	17.9 (8.9 to 26.9)	.001
Intervention	30.8	56.7	25.8 (15.9 to 35.8)	<.001
Difference	N/A	N/A	8.0 (–5.5 to 21.4)	.25
...trust governmental approval process... for public...[e]				
Control	81.3	89.4	8.1 (1.9 to 14.3)	.01
Intervention	74.2	90	15.8 (9.3 to 22.4)	<.001
Difference	N/A	N/A	7.7 (–1.3 to 16.7)	.09
...trust governmental approval process... for children...[e]				
Control	78.9	85.4	6.5 (0.7 to 12.4)	.03
Intervention	66.7	87.5	20.8 (13.6 to 28.1)	<.001
Difference	N/A	N/A	14.3 (5.0 to 23.7)	.003

[a]Simple (unadjusted) DID of knowledge and trust. Rates are among those participants who indicated having at least 1 minor in the household and assume that anyone missing a 1-month follow-up did not change their trust and knowledge from baseline. Models are estimated using a generalized linear model and maximum likelihood estimates.

[b]Variable coded equal to 1 if true, equal to 0 if false/unsure.

[c]N/A: not applicable.

[d]Variable coded equal to 1 if adequate/superior knowledge, equal to 0 if no/minimal/basic knowledge.

[e]Variable coded equal to 1 if fully/mostly/somewhat trust, equal to zero if do not trust.

Participant Engagement

Table S6 in Multimedia Appendix 1 shows most participants (228/270, 84% to 228/257, 89%) found the video (provided on Monday) and educational information (provided on Wednesday) very/extremely useful. Materials in weeks 1 and 4 had the highest percentages of participants (236/266 and 220/247, 89%) expressing material was very or extremely useful. When asked about familiarity with that week's provided information, only about half of the participants (108/266, 41% to 132/247, 53%) responded they were very/extremely familiar, with week 1 information showing the lowest percentage of participants (108/266, 41%) responding being very/extremely familiar with information.

Table S7 in Multimedia Appendix 1 summarizes participant engagement data. More than half of the participants (124/265, 47% to 147/252, 58%) clicked the specific link for making a COVID-19 vaccination appointment each week. Among those who completed that week's activity, 89% (219/246) to 90% (227/252) of participants clicked to download the summary infographic to save/share. A large number of participants listened to the audio files from the written information sent on Wednesdays (681/1037, 66%) and the summary infographic sent on Thursdays (562/1037, 54%). Please note that percentages presented in Tables S6 and S7 are based on all participant information collected in our platform, regardless of whether the participant was included in our main analysis of primary and secondary outcomes.

Sensitivity Analysis

A TOT sensitivity analysis with DID model estimation was used. Table S8 in Multimedia Appendix 1 presents estimates from the DID model for the primary vaccination outcome, both for individual age groups and collectively. The analysis revealed an 18.6 percentage point difference (95% CI) in vaccination uptake for children aged 5 to 11 years in the intervention group compared to controls (P=.02), and a 12.1 percentage point increase when considering all age groups (P=.04). Little's MCAR test found evidence that data were not MCAR (P<.01). Weighted estimates from the DID model, which account for statistically significant differences between participants who completed the follow-up survey and those who did not, are provided in Table S9 in Multimedia Appendix 1. These weighted TOT estimates for children aged 5 to 11 years and the overall sample closely align with the unweighted estimates, showing increases of 20.1 percentage points (P=.01) and 12.2 percentage points (P=.04), respectively.

XSL•FO
RenderX

Similarly, TOT analyses of secondary outcomes related to vaccine knowledge and trust (Tables S10 and S11 in Multimedia Appendix 1) are consistent with the ITT results. Specifically, we found a larger increase in trust in the governmental approval process for children post intervention, with TOT unweighted and weighted estimates showing differences of 18.8 and 20.1 percentage points (P=.01), respectively. In addition, these tables indicate a significant difference of 10.8 and 10.6 percentage points (P=.05) in government approval processes for the public for unweighted and weighted TOT analyses, respectively. The detailed estimates using the TOT approach with the DID model are presented in Tables S8 to S11 in Multimedia Appendix 1.

Discussion

The Role of Community-Based Digital Interventions in Addressing Vaccine Hesitancy

In this study, we examined whether a community-based digital intervention could increase COVID-19 vaccination rates and trust in the governmental approval process to ensure vaccine safety in children aged 5 to 11 years. MVLA 2.0 significantly increased COVID-19 vaccine uptake in children aged 5 to 11 years and trust in the COVID-19 vaccine governmental approval processes for children. The MVLA 2.0 digital intervention demonstrated an effect on vaccine uptake comparable to the magnitude of effect observed in another such COVID-19-specific digital intervention [26], which found an adjusted absolute difference of 11.9 (95% CI 4.5 - 19.3) percentage points in their 1-week intervention conducted during emergency room visits. Electronic health record approaches typically used in broader immunization efforts have demonstrated effect sizes between 4.2% (adjusted) and 7% (P=.05) [27,28]. These findings offer new insights into the role of community-informed and culturally tailored digital strategies, such as MVLA 2.0, in increasing trust in vaccines and improving vaccination rates. Furthermore, this digital intervention demonstrated the potential for scaling across diverse communities to address other pediatric vaccination gaps.

MVLA 2.0 intervention impact appears to reflect a combination of delivery, cultural tailoring, and user-centered design. Although we cannot rule out selection or reporting bias as contributors to the observed effects, interviews from MVLA 1.0 participants suggest that several components were likely impactful [23]. These included digitally accessible SMS text messaging delivering a web-based curriculum, with written educational content and videos reinforcing weekly material content over time. We also incorporated culturally tailored, real-time feedback to address parental concerns about COVID-19 vaccinations in children of ages 5 to 11 years and counter misinformation, disinformation, and myths prevalent in the Hispanic community. Finally, we applied user-centered design features such as a discussion board and voice-overs for written content to enhance engagement, accessibility, and adherence.

Although some progress for COVID-19 vaccination was made during the pandemic within the Hispanic community, trends in overall vaccinations remain low. Furthermore, national estimates on intent to get a COVID-19 vaccine among Hispanic groups have dropped from 18.6% to 11.4% between September 2024 and January 2025 [29]. Perhaps more concerning is the postpandemic decline in uptake of other childhood vaccines [2]. In California, where this intervention was conducted, statewide estimates on unvaccinated Hispanics have increased nearly 5% in the 4-month period between the end of September 2024 (15.7%) and the end of January 2025 (19.2%) alone [29]. A January 2025 article received media coverage over its analysis of 2019 - 2022 vaccination data showing an overall decline in MMR and diphtheria, tetanus, and acellular pertussis vaccination rates across 11 states and, more worrisome, that three of those states had MMR statewide rates below herd immunity and broadly vulnerable to an outbreak [30]. Just weeks after these data were published, also widely reported, Texas experienced its largest measles outbreak in 3 decades, with the first US mortality from measles in a decade and the first pediatric mortality in 2 decades [31].

In addition, there are concerns that many of the impediments to COVID vaccination are similarly involved with the more recent downward trends in vaccination elsewhere, and some initial evidence to suggest these concerns may be warranted [32]. Because the MVLA 2.0 intervention was designed to be digital and culturally tailored, this model may have the potential for scaling across diverse communities to address other pediatric vaccination gaps. Future research is needed to explore the use of digital culturally tailored interventions such as MVLA 2.0 more broadly, as a cost-effective strategy to improve vaccination rates and reduce disparities in undervaccination for children.

Both intervention and control groups showed an increase in knowledge postintervention. Although the magnitude of this increase was not significantly different between groups for most outcomes, this study showed a significant intervention effect for increasing trust regarding the governmental approval process of the COVID-19 vaccine for children. A recent systematic review reported that health concerns, vaccine attributes, and mistrust were primary drivers of COVID-19 vaccine hesitancy, whereas trust and confidence, community and social factors, and demographics and identity were key factors associated with vaccine uptake [33]. Altogether, these broader trends in vaccine hesitancy and the associated causes reinforce the need for additional research into broad-based, easily disseminated multifactorial approaches to improve vaccine confidence.

Approaches, such as the mobile phone–centric format used in MVLA 2.0, may help promote vaccine acceptance, especially given the widespread cellphone and smartphone ownership. According to the Pew Research Center, 98% of Americans own some kind of cellphone and 91% some kind of smartphone. The penetration of ownership/type of usage is relatively high regardless of income (95% and 84%, respectively, those making under $30,000 per year) or ethnicity (97%/87% Black and 99%/93% Hispanic) and even age (94%/79% for those above 65 years of age) [34]. Furthermore, considerable support can be found in the literature for adopting digital, mobile-based interventions to change health behaviors, including immunization among adolescents [35], influenza vaccination among pediatric Hispanic patients in low-income urban areas [36], increasing human papillomavirus vaccine uptake among underserved high-risk populations [37], as well as culturally

XSL•FO
RenderX

specific vaccine education to combat misinformation and improve vaccination behaviors [36,38-46]. Prior studies have recognized the ease of implementation as a benefit of mobile phone–delivered text-based interventions [36,38-53].

SMS text messaging interventions were successful in increasing influenza vaccination rates among Hispanic patients in low-income urban areas [36] and in increasing human papillomavirus vaccine uptake among underserved high-risk populations [37], as well as culturally specific vaccine education to combat misinformation [36,38-46]. Prior studies have recognized the ease of implementation as a benefit of mobile phone–delivered text-based interventions [36,38-53].

Study Limitations and Suggested Future Work

This study has several limitations. We used self-reported vaccination status, which is subject to over- or under-reporting. Parents may have over-reported vaccination due to social desirability bias, particularly after participating in a provaccine educational intervention. The lack of objective verification through medical records or vaccination databases reduces confidence in the primary findings. However, the target population for this intervention may be less likely to provide official records due to concerns with immigration and documentation status. Thus, building trust to ensure the collection of official vaccination records for future digital vaccine acceptance interventions is key.

Lower baseline vaccination rates in the intervention group may overestimate efficacy. Due to a 15% attrition rate and systematic differences between completers and noncompleters, the ITT analysis may be compromised, despite attempts to adjust for these differences using weighting in the TOT analysis. The assumption that vaccinating 1 child in an age group translates to all children in that household being vaccinated for those with missing data could lead to inflated effect sizes. In this study, individual children in a household did not have unique IDs; therefore, we were unable to link responses between baseline and follow-up. Thus, future work should aim to analyze vaccination uptake at the individual level. Furthermore, CIs in our results were wide in relation to our primary (vaccination status of children aged 5-11 years) and secondary outcomes (trust in the government process for the approval of COVID-19 vaccines for children). This variability may suggest that estimates of intervention effect lack precision, likely explained by the small sample size. Thus, future studies warrant a larger sample size to provide a more precise estimate of the efficacy of a digital intervention on vaccine uptake and trust.

Although a strength of this study included using a community-based approach to recruit populations typically “under-represented” in research, this also may limit the generalizability of our findings. First, we primarily recruited participants who had already engaged with community organizations and were willing to participate in vaccine research, which could create selection bias. This population could differ systematically from the broader Hispanic community in terms of health engagement, technology access, and baseline vaccine attitudes. Second, most participants completed the intervention in English, so results may not generalize to households who prefer to communicate in Spanish. Interestingly, during the MVLA 1.0 intervention, 79% of participants completed the intervention in Spanish [22]. This shift in language predominance might be explained by differences in the community partner organizations engaged in MVLA 2.0 and the populations they serve. Third, there may have been a bias in participants willing to be screened for eligibility to participate. Future digital intervention studies should work closely with community organizations to incorporate a screening method where the community organization can confirm whether participants are affiliated with their organization to improve recruitment tracking.

Future directions for vaccine interventions should emphasize the development of culturally tailored community-based digital strategies to build trust in vaccines. This includes actively engaging with community leaders and advocates to ensure that interventions are relevant and resonate with the target population. Addressing misinformation could involve using mobile platforms to provide accurate, accessible information and counteract false narratives in the same modality in which they are being disseminated. In addition, improving vaccination behaviors might be achieved by integrating behavioral insights to identify and overcome specific barriers to vaccination within the community. Scaling these strategies to diverse populations is critical to address the challenge of decreasing vaccination rates.

Conclusions

The findings from this RCT evaluating MVLA 2.0 (a culturally tailored, community-based, mobile phone–delivered intervention) demonstrate higher COVID-19 vaccine uptake among Hispanic children and increased parental trust in the governmental vaccine approval process compared with the control group. This study is innovative in combining community-informed content, real-time culturally tailored feedback, and a fully automated, mobile phone–centric delivery model. It adds to existing literature by testing a culturally tailored digital curriculum in a community setting and showing improvements in both uptake and trust. These findings provide rigorous evidence that scalable, low-cost digital strategies can address vaccine confidence and uptake in populations with persistent gaps in vaccine acceptance. Real-world implications include the portability and adaptability of this approach across diverse communities and settings—urban or rural—and its feasibility and potential for broader applicability and scalability to support public health vaccination campaigns.

Acknowledgments

University of California, Los Angeles, and Pepperdine University provided in-kind support. The authors are thankful to Gael Perez for his help supporting the submission of the manuscript.

XSL•FO
RenderX

Funding

Funding for this project came through the VaxUp Challenge, in which the selection process was conducted jointly by the Los Angeles Department of Public Health and Children's Hospital of Los Angeles. The Possibility Labs through the Digital Transformation Fund also provided supplementary funding. This research was also, in part, funded by the National Institutes of Health (NIH) Agreement OT2HL158287, NIH/National Institute on Aging (NIA) Grant P30-AG021684, and NIH/National Center for Advancing Translational Sciences (NCATS) Grant Number UL1TR001881. The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated and analyzed during this study are not publicly available to protect patient confidentiality. Deidentified data, study protocol, and statistical analysis plan may be accessed upon request to the corresponding author and institutional review board approval.

Authors' Contributions

Conceptualization: LRB, YC-L
Methodology: LRB, AK, RL-C, EA, BW, MM, YC-L
Software: LRB, AK, BW, MM, YC-L
Validation: LRB, AK, RD, KN, YC-L
Writing—original draft preparation: LRB, AK, RD, DM, RL-C, EA, BW, MM, KN, YC-L
Writing—review and editing: LRB, AK, RD, DM, RL-C, EA, BW, MM, KN, YC-L
Visualization: LRB, AK, YC-L
Supervision: LRB, YC-L
Project administration: LRB, DM, BW, MM, YC-L
Funding acquisition: LRB, RD, RL-C, EA, KN, YC-L
All authors have read and agreed to the published version of the manuscript.
We certify that no artificial intelligence tools were used in the writing, editing, or preparation of any portion of this manuscript. All content, including text, figures, and tables, was created by the authors without the use of artificial intelligence.

Conflicts of Interest

YCL reports consulting fees for advisory work from Eli Lilly and Company and Exact Sciences for work unrelated to this study. Other authors have no conflicts to declare.

Multimedia Appendix 1
Examples of the culturally tailored infographics used in the MiVacunaLA 2.0 digital intervention.
[PDF File, 579 KB - jmir_v28i1e78103_app1.pdf]

Checklist 1
CONSORT checklist.
[PDF File, 313 KB - jmir_v28i1e78103_app2.pdf]

References

1. Murthy NC, Zell E, Fast HE, et al. Disparities in first dose COVID-19 vaccination coverage among children 5-11 years of age, United States. Emerg Infect Dis 2022 May;28(5):986-989. [doi: 10.3201/eid2805.220166] [Medline: 35226801]
2. COVID-19 vaccination coverage and intent for vaccination, children 6 months through 17 years, United States. Centers for Disease Control and Prevention (CDC). URL: https://www.cdc.gov/covidvaxview/weekly-dashboard/child-coverage-vaccination.html [accessed 2026-03-11]
3. COVID-19 vaccination data - vaccination status by age and race and ethnicity. California Department of Public Health. URL: https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-Vaccine-Data.aspx#Age [accessed 2026-03-04]
4. Kim L, Whitaker M, O'Halloran A, et al. Hospitalization rates and characteristics of children aged <18 years hospitalized with laboratory-confirmed COVID-19 - COVID-NET, 14 states, March 1-July 25, 2020. MMWR Morb Mortal Wkly Rep 2020 Aug 14;69(32):1081-1088. [doi: 10.15585/mmwr.mm6932e3] [Medline: 32790664]
5. Gonzalez SL, Sarik DA, Dean-Olmsted E, Salyakina D. COVID-19 vaccination knowledge, attitudes, and practices within a majority Hispanic/Latino pediatric healthcare system. J Pediatr Nurs 2025;82:e65-e74. [doi: 10.1016/j.pedn.2025.03.020] [Medline: 40246617]
6. Bateman LB, Hall AG, Anderson WA, et al. Exploring COVID-19 vaccine hesitancy among stakeholders in African American and Latinx communities in the Deep South through the lens of the health belief model. Am J Health Promot 2022 Feb;36(2):288-295. [doi: 10.1177/08901171211045038] [Medline: 34719985]

XSL•FO
RenderX

7. Milo RB, Aguayo C, Chico AR, et al. Understanding the barriers to COVID-19 vaccine among Hispanic/Latinx communities. J Prim Care Community Health 2023;14:21501319231174810. [doi: 10.1177/21501319231174810] [Medline: 37306326]
8. Haro-Ramos AY, Sanchez GR, Barreto MA. Health care discrimination and immigration fears: unpacking COVID-19 vaccine hesitancy in Latino adults. Am J Public Health 2024 Jul;114(S6):S505-S509. [doi: 10.2105/AJPH.2024.307668] [Medline: 39083732]
9. Hammershaimb EA, Cole LD, Liang Y, et al. COVID-19 vaccine acceptance among US parents: a nationally representative survey. J Pediatric Infect Dis Soc 2022 Aug 30;11(8):361-370. [doi: 10.1093/jpids/piac049] [Medline: 35748047]
10. Ataç Ö, Peterson LE, Waters TM. The impact of the COVID-19 pandemic on vaccinations in United States primary care practices. PLoS One 2025;20(6):e0325934. [doi: 10.1371/journal.pone.0325934] [Medline: 40493658]
11. Vaccination coverage and exemptions among kindergartners. Centers for Disease Control and Prevention (CDC). 2025. URL: https://www.cdc.gov/schoolvaxview/data/index.html#cdc_data_surveillance_section_2-new-findings-on-vaccination-coverage-and-exemptions-among-kindergartners-2024-2025-school-year [accessed 2025-09-16]
12. Seither R, Yusuf OB, Dramann D, Calhoun K, Mugerwa-Kasujja A, Knighton CL. Coverage with selected vaccines and exemption from school vaccine requirements among children in kindergarten - United States, 2022-23 school year. MMWR Morb Mortal Wkly Rep 2023 Nov 10;72(45):1217-1224. [doi: 10.15585/mmwr.mm7245a2] [Medline: 37943705]
13. Dong E, Saiyed S, Nearchou A, Okura Y, Gardner LM. Trends in county-level MMR vaccination coverage in children in the United States. JAMA 2025 Aug 26;334(8):730-732. [doi: 10.1001/jama.2025.8952] [Medline: 40455620]
14. Hill HA, Yankey D, Elam-Evans LD, Singleton JA, Sterrett N. Vaccination coverage by age 24 months among children born in 2017 and 2018 - National Immunization Survey-Child, United States, 2018-2020. MMWR Morb Mortal Wkly Rep 2021 Oct 15;70(41):1435-1440. [doi: 10.15585/mmwr.mm7041a1] [Medline: 34648486]
15. Cunniff L, Alyanak E, Fix A, et al. The impact of the COVID-19 pandemic on vaccination uptake in the United States and strategies to recover and improve vaccination rates: a review. Hum Vaccin Immunother 2023 Aug 1;19(2):2246502. [doi: 10.1080/21645515.2023.2246502] [Medline: 37671468]
16. Abdul-Mutakabbir JC, Granillo C, Peteet B, et al. Rapid implementation of a community-academic partnership model to promote COVID-19 vaccine equity within racially and ethnically minoritized communities. Vaccines (Basel) 2022 Aug 20;10(8):1364. [doi: 10.3390/vaccines10081364] [Medline: 36016251]
17. Munoz-Lavanderos C, Oluyomi A, Rosales O, Hernandez N, Mensah-Bonsu N, Badr H. Development, implementation, and evaluation of three outreach events to improve COVID-19 vaccine uptake among racial and ethnic minority communities in Houston, Texas, 2022. Public Health Rep 2024;139(1_suppl):71S-80S. [doi: 10.1177/00333549231213848] [Medline: 38140821]
18. Demeke J, McFadden SM, Dada D, et al. Strategies that promote equity in COVID-19 vaccine uptake for undocumented immigrants: a review. J Community Health 2022 Jun;47(3):554-562. [doi: 10.1007/s10900-022-01063-x] [Medline: 35084639]
19. AuYoung M, Rodriguez Espinosa P, Chen WT, et al. Addressing racial/ethnic inequities in vaccine hesitancy and uptake: lessons learned from the California alliance against COVID-19. J Behav Med 2023 Apr;46(1-2):153-166. [doi: 10.1007/s10865-022-00284-8] [Medline: 35066696]
20. Riggle-van Schagen C, Andrade EL, Chandarana S, et al. Assessing changes in COVID-19 vaccine uptake and intentions among the Brigada Digital Latino social media audience: a repeated cross-sectional study. BMC Digit Health 2024;2(1):72. [doi: 10.1186/s44247-024-00131-7]
21. Chandeying N, Thongseiratch T. Systematic review and meta-analysis comparing educational and reminder digital interventions for promoting HPV vaccination uptake. NPJ Digit Med 2023 Aug 29;6(1):162. [doi: 10.1038/s41746-023-00912-w] [Medline: 37644090]
22. Castellon-Lopez YM, Klomhaus AM, Garcia C, et al. MivacunaLA (MyShotLA): a community-partnered mobile phone intervention to improve COVID-19 vaccination behaviors among low-income, Spanish-speaking, and immigrant Latino parents or caregivers. Vaccines (Basel) 2024 May 8;12(5):511. [doi: 10.3390/vaccines12050511] [Medline: 38793762]
23. Panameno M, Blanco LR, Hernandez AM, et al. Using digital technology to build COVID-19 vaccine confidence: a qualitative study among Latinx parents of children aged 5-11 in under-resourced communities across Los Angeles County. Vaccines (Basel) 2023 May 30;11(6):1042. [doi: 10.3390/vaccines11061042] [Medline: 37376431]
24. LA county COVID-19 vaccine dashboard. County of Los Angeles Public Health. 2025. URL: http://publichealth.lacounty.gov/media/coronavirus/vaccine/vaccine-dashboard.htm [accessed 2026-03-04]
25. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025 Apr 14;389:e081123. [doi: 10.1136/bmj-2024-081123] [Medline: 40228833]
26. Rodriguez RM, Nichol G, Eucker SA, et al. Effect of COVID-19 vaccine messaging platforms in emergency departments on vaccine acceptance and uptake: a cluster randomized clinical trial. JAMA Intern Med 2023 Feb 1;183(2):115-123. [doi: 10.1001/jamainternmed.2022.5909] [Medline: 36574256]
27. Fiks AG, Hunter KF, Localio AR, et al. Impact of electronic health record-based alerts on influenza vaccination for children with asthma. Pediatrics 2009 Jul;124(1):159-169. [doi: 10.1542/peds.2008-2823] [Medline: 19564296]

XSL•FO
RenderX

28. Wijesundara JG, Ito Fukunaga M, Ogarek J, et al. Electronic health record portal messages and interactive voice response calls to improve rates of early season influenza vaccination: randomized controlled trial. J Med Internet Res 2020 Sep 25;22(9):e16373. [doi: 10.2196/16373] [Medline: 32975529]
29. Vaccination uptake, intent, and confidence. Centers for Disease Control and Prevention (CDC). 2025. URL: https://www.cdc.gov/respvaxview/dashboards/vaccination-behavioral-social-drivers.html [accessed 2026-03-04]
30. Moghtaderi A, Callaghan T, Luo Q, et al. Evidence on trends in uptake of childhood vaccines and association with COVID-19 vaccination rates. Vaccine 2025 Jan 25;45:126631. [doi: 10.1016/j.vaccine.2024.126631] [Medline: 39700909]
31. Shastri D, Seitz A. A Texas child who was not vaccinated has died of measles, a first for the US in a decade. AP News. 2025. URL: https://apnews.com/article/measles-outbreak-west-texas-death-rfk-41adc66641e4a56ce2b2677480031ab9 [accessed 2025-02-26]
32. Melani D. Child dies of measles: here's a look at the country's re-emerging disease. University of Colorado Anschutz. 2025. URL: https://news.cuanschutz.edu/news-stories/child-dies-of-measles-heres-a-look-at-the-countrys-re-emerging-disease [accessed 2025-03-03]
33. Nwachukwu G, Rihan A, Nwachukwu E, Uduma N, Elliott KS, Tiruneh YM. Understanding COVID-19 vaccine hesitancy in the United States: a systematic review. Vaccines (Basel) 2024 Jul 6;12(7):747. [doi: 10.3390/vaccines12070747] [Medline: 39066385]
34. Mobile fact sheet. Pew Research Center. URL: https://www.pewresearch.org/internet/fact-sheet/mobile [accessed 2025-02-28]
35. Stockwell MS, Kharbanda EO, Martinez RA, Vargas CY, Vawdrey DK, Camargo S. Effect of a text messaging intervention on influenza vaccination in an urban, low-income pediatric and adolescent population: a randomized controlled trial. JAMA 2012 Apr 25;307(16):1702-1708. [doi: 10.1001/jama.2012.502] [Medline: 22535855]
36. Stockwell MS, Hofstetter AM, DuRivage N, et al. Text message reminders for second dose of influenza vaccine: a randomized controlled trial. Pediatrics 2015 Jan;135(1):e83-e91. [doi: 10.1542/peds.2014-2475] [Medline: 25548329]
37. Sandi YDL, Yang LY, Andarini E, Maryam D, Wu LM. Effectiveness of digital education on human papillomavirus knowledge, vaccination intent and completion rates in adolescents and young adults: a meta-analysis. J Adv Nurs 2025 Apr;81(4):2199-2213. [doi: 10.1111/jan.16575] [Medline: 39465620]
38. Marquez C, Kerkhoff AD, Naso J, et al. A multi-component, community-based strategy to facilitate COVID-19 vaccine uptake among Latinx populations: from theory to practice. PLoS One 2021;16(9):e0257111. [doi: 10.1371/journal.pone.0257111] [Medline: 34543291]
39. Balasuriya L, Santilli A, Morone J, et al. COVID-19 vaccine acceptance and access among Black and Latinx communities. JAMA Netw Open 2021 Oct 1;4(10):e2128575. [doi: 10.1001/jamanetworkopen.2021.28575] [Medline: 34643719]
40. Castellon-Lopez YM, Carson SL, Mansfield L, et al. "The System Doesn't Let Us in"-a call for inclusive COVID-19 vaccine outreach rooted in Los Angeles Latinos' experience of pandemic hardships and inequities. Int J Environ Res Public Health 2022 May 10;19(10):5785. [doi: 10.3390/ijerph19105785] [Medline: 35627322]
41. Cialdini RB, Goldstein NJ. Social influence: compliance and conformity. Annu Rev Psychol 2004;55:591-621. [doi: 10.1146/annurev.psych.55.090902.142015] [Medline: 14744228]
42. Garcia J, Vargas N, de la Torre C, Magana Alvarez M, Clark JL. Engaging Latino families about COVID-19 vaccines: a qualitative study conducted in Oregon, USA. Health Educ Behav 2021 Dec;48(6):747-757. [doi: 10.1177/10901981211045937] [Medline: 34596462]
43. Kapteyn A, Angrisani M, Bennett D, Bruin W, Darling J, Gutsche T, et al. Tracking the effect of the COVID-19 pandemic on the lives of American households. Surv Res Methods 2020;14(2):179-186. [doi: 10.18148/srm/2020.v14i2.7737]
44. Lustria MLA, Noar SM, Cortese J, Van Stee SK, Glueckauf RL, Lee J. A meta-analysis of web-delivered tailored health behavior change interventions. J Health Commun 2013;18(9):1039-1069. [doi: 10.1080/10810730.2013.768727] [Medline: 23750972]
45. Webb Hooper M, Nápoles AM, Pérez-Stable EJ. No populations left behind: vaccine hesitancy and equitable diffusion of effective COVID-19 vaccines. J Gen Intern Med 2021 Jul;36(7):2130-2133. [doi: 10.1007/s11606-021-06698-5] [Medline: 33754319]
46. Benefits of getting vaccinated. Centers for Disease Control and Prevention. 2025. URL: https://www.cdc.gov/covid/vaccines/benefits.html [accessed 2026-03-04]
47. Cáceres NA, Shirazipour CH, Herrera E, Figueiredo JC, Salvy SJ. Exploring Latino Promotores/a de Salud (Community Health Workers) knowledge, attitudes, and perceptions of COVID-19 vaccines. SSM Qual Res Health 2022 Dec;2:100033. [doi: 10.1016/j.ssmqr.2021.100033] [Medline: 34904136]
48. Carson SL, Casillas A, Castellon-Lopez Y, et al. COVID-19 vaccine decision-making factors in racial and ethnic minority communities in Los Angeles, California. JAMA Netw Open 2021 Sep 1;4(9):e2127582. [doi: 10.1001/jamanetworkopen.2021.27582] [Medline: 34591103]
49. Gargano LM, Herbert NL, Painter JE, et al. Impact of a physician recommendation and parental immunization attitudes on receipt or intention to receive adolescent vaccines. Hum Vaccin Immunother 2013 Dec;9(12):2627-2633. [doi: 10.4161/hv.25823] [Medline: 23883781]
50. Lee HY, Koopmeiners JS, McHugh J, Raveis VH, Ahluwalia JS. mHealth pilot study: text messaging intervention to promote HPV vaccination. Am J Health Behav 2016 Jan;40(1):67-76. [doi: 10.5993/AJHB.40.1.8] [Medline: 26685815]

XSL•FO
RenderX

51. McAboy K. UCLA-led coalition works with minority communities on COVID-19 and vaccine education. FOX 11 Los Angeles. 2021. URL: https://www.foxla.com/news/ucla-led-coalition-works-with-minority-communities-on-covid-19-and-vaccine-education [accessed 2026-03-04]
52. Silverman-Lloyd LG, Dominguez Cortez J, Godage SK, et al. Immigrant Latino parents demonstrated high interactivity with pediatric primary care text messaging intervention. Mhealth 2020;6:45. [doi: 10.21037/mhealth.2020.01.06] [Medline: 33209916]
53. Szilagyi PG, Thomas K, Shah MD, et al. The role of trust in the likelihood of receiving a COVID-19 vaccine: results from a national survey. Prev Med 2021 Dec;153:106727. [doi: 10.1016/j.ypmed.2021.106727] [Medline: 34280405]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
DID: difference-in-difference
HPV: human papillomavirus
ITT: intention-to-treat
MCAR: Missing Completely at Random
MMR: measles-mumps-rubella
MVLA: MiVacunaLA/MyShotLA
RCT: randomized controlled trial
TOT: treatment-on-the-treated
UAS: Understanding America Survey
UCLA: University of California, Los Angeles

Edited by S Brini; submitted 26.May.2025; peer-reviewed by L Jain, SW Lee; accepted 10.Dec.2025; published 30.Mar.2026.

Please cite as:
Blanco LR, Klomhaus A, Dudovitz R, Marquez D, Lopez-Chang R, Aleman E, Weerman B, Moldoff M, Norris K, Castellon-Lopez Y Digital Intervention (MiVacunaLA 2.0) to Promote COVID-19 Vaccine Acceptance Among Hispanic Children: Community-Based Randomized Controlled Trial
J Med Internet Res 2026;28:e78103
URL: https://www.jmir.org/2026/1/e78103
doi:10.2196/78103

© Luisa R Blanco, Alexandra Klomhaus, Rebecca Dudovitz, Denise Marquez, Ray Lopez-Chang, Evelyn Aleman, Bas Weerman, Michael Moldoff, Keith Norris, Yelba Castellon-Lopez. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

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

Asmaa Housni[1], RD, PhD; Aidan Shulkin[2], MD; Alexandra Katz[3], MSc, MD; Giuliana Giannini[1], RD; Amélie Roy-Fleming[1], RD, CDE, MSc; Meranda Nakhla[4], MSc, MD; Courtney South[1], RD, MSc; Anne-Sophie Brazeau[1], RD, PhD

[1]School of Human Nutrition, McGill University, 21,111 Lakeshore drive, Montreal, QC, Canada

[2]Faculté de médecine, Université de Montréal, Montreal, QC, Canada

[3]Internal Medicine, McGill University, Montreal, QC, Canada

[4]Division of Endocrinology, Montreal Children's Hospital, Montreal, QC, Canada

Corresponding Author:
Anne-Sophie Brazeau, RD, PhD
School of Human Nutrition, McGill University, 21,111 Lakeshore drive, Montreal, QC, Canada

Abstract

Background: Carbohydrate counting (CC) assists people with type 1 diabetes (T1D) adjust mealtime insulin doses; however, it is often burdensome. Mobile apps can simplify this process by automating carbohydrate estimation and insulin calculations, yet no comprehensive solution currently combines photo-based carbohydrate recognition with an integrated bolus calculator.

Objective: This study aimed to identify user-informed design principles from usability testing interviews to optimize a novel app supporting young adults with T1D in CC and insulin dosing.

Methods: We conducted 4 iterative rounds of usability testing interviews, each with 3 to 5 participants, using a think-aloud protocol to evaluate how easily and effectively users interacted with the app and to identify areas for improvement. Interviews were analyzed qualitatively to derive main design principles, and findings from each round informed the refinement of the app prior to subsequent testing.

Results: A total of 18 participants completed the usability testing (median age of 23, IQR 19-24 y and diabetes duration of 9, IQR 6-12 y; n=12, 66.7% young women). Thematic analysis highlighted that a person-centered design that prioritizes the lived experiences of youth with T1D was essential to position the app as a self-management support system, beyond a clinical tool. Personalization was central, including customizable treatment profiles, tailored dashboard metrics, esthetic preferences, and artificial intelligence–driven recommendations based on personal trends. Early usability barriers revealed the need for intuitive navigation, streamlined multistep processes, and clear guidance for data entry and interpretation. Participants valued culturally inclusive content and familiar terminology to enhance accessibility and engagement. Users perceived strong potential for the app to centralize diabetes management tasks, integrate contextual factors (eg, exercise, diet, and timing of insulin) with glucose data, generate sharable reports to facilitate patient-practitioner communication, and strengthen self-efficacy through personalized trend analysis. Concerns about over-reliance on automation underscored the necessity of transparent data verification and user override options to maintain trust in insulin dosing decisions.

Conclusions: Iterative usability testing highlighted the importance of balancing automation with user control, personalization, and contextual understanding of personal trends, as key design principles to enhance engagement and the apps' relevance as a self-management tool. Incorporating these features into a CC and insulin-dosing app could improve self-efficacy in youth living with T1D.

(J Med Internet Res 2026;28:e86024) doi:10.2196/86024

KEYWORDS

type 1 diabetes; carbohydrate counting; mobile apps; youth; usability testing

Introduction

Type 1 diabetes (T1D) is an autoimmune condition resulting in pancreatic β cell destruction [1]. People with T1D need daily administration of insulin to regulate glucose levels and prevent acute events and long-term health risks [2,3]. Intensive insulin therapy through carbohydrate counting (CC) is the recommended treatment for individuals living with T1D due to its association with lowered glycated hemoglobin (HbA_{1c}) levels and the prevention of diabetes-associated complications [4-6]. Moreover, CC offers dietary flexibility by allowing individuals to tailor their mealtime insulin based on the carbohydrate content of

XSL•FO
RenderX

their meals [7]. However, CC adherence can be burdensome and challenging as it requires tracking intake, reading and interpreting nutrition labels, and weighing portions [8,9]. Additionally, calculations are multifactorial and prone to error, due to the complexities of combining insulin sensitivity factor, insulin-to-carbohydrate ratio, and glucose targets [10].

The transition from childhood to adulthood introduces an additional barrier to CC, namely due to social and developmental changes commonly observed in this age group [11-13], and CC tasks transferred from parent to child. A study by Gurmani et al [14] reported that 43.6% of adolescents inaccurately calculated carbohydrates by 10 to 20 g from the actual amount. Incorrectly identified meals either lacked nutrition labels or required fiber subtraction to calculate carbohydrates accurately, underscoring the difficulty young adults face estimating carbohydrates in unpackaged foods [14]. This contributes to the broader challenge of achieving glycemic stability during adolescence and young adulthood, as evidenced by the higher HbA_{1c} levels observed in these age groups compared to others [13]. Given the various benefits of CC, strategies are needed to promote its proper use among youth.

A promising solution lies in mobile health (mHealth), a branch of telemedicine that uses technology such as mobile devices and tablets to improve health care services for both patients and health care providers [15]. In the context of diabetes management, mHealth can automate CC as a novel care approach for individuals living with T1D. In fact, the efficacy of mHealth, that is, the use of smartphone or tablet apps in improving diabetes self-management among adults, is well established [16-18]. Despite the large variety of apps able to offer photo recognition powered by artificial intelligence (AI) to quantify meal composition, most are not designed for diabetes management, few have been tested in a clinical setting, and even fewer limit their search to validated food databases only [19]. Additionally, while there are separate apps available for bolus insulin calculations, few have undergone usability testing or been developed using user-centered design approach, raising concerns about the safety of insulin calculation apps [20]. A needs assessment survey was conducted across Canada to identify the barriers and needs related to CC in people with T1D, including youth. The survey emphasized that existing mobile apps currently used for CC often fail to meet the needs of this population. A novel app incorporating features such as photo recognition, reliable nutrient values, and personalized bolus calculations could help reduce the burden of CC [21].

Scientific literature has demonstrated a need for a comprehensive app tailored for individuals living with T1D that integrates accurate CC and bolus insulin calculations. Thus, we collaborated with Ikigai Development, Inc, a company specializing in app development, to design the Petit Cactus app, which involved users in the development process to promote a user-centered design [22]. The Petit Cactus app was designed to include four main features: (1) the ability to calculate the number of carbohydrates in meals either through AI-powered photo recognition or through manual selection of food items, with data derived from validated Canadian food databases; (2) bolus insulin dose calculation that incorporates multiple variables (glucose levels, insulin-to-carbohydrates ratio, and correction factors), and integrates continuous glucose monitoring (CGM) data to help users visualize how logged events (ie, meals, stress, sick days, physical activities) affect glucose levels; (3) an electronic food journal for reviewing logged meals and event history; and (4) a connecting portal to a health care provider interface.

A primary objective was to identify user-informed design principles that could guide the development of this and future digital tools for supporting self-management in adolescents and young adults with T1D. Building on these principles, a broader objective was to collaboratively optimize the app's usability to ensure it effectively meets the needs of this population.

Methods

Study Design

The study adopts a user-centered design framework, using iterative usability testing with young adults living with T1D [23]. User feedback was systematically collected and integrated into successive stages of app development to inform design refinements and enhance overall usability. We conducted usability testing to evaluate how users interacted with the Petit Cactus app and to identify usability issues, preferences, and unmet needs. Participants completed task-based interactions with the prototype while verbalizing their thoughts using a think-aloud approach, following a semistructured interview. Usability testing with young adults involved 4 iterative rounds, each with a new group of 3 to 5 participants testing an updated version of the Petit Cactus app refined based on feedback from previous rounds. The usability testing interviews were followed by 1 week of independent use and completion of a feedback questionnaire to capture additional insights (Figure 1). Data from observations and questionnaires were analyzed using inductive qualitative thematic analysis to derive user needs and design insights.

XSL•FO
RenderX

Figure 1. Methodology for conducting iterative usability testing. T1D: type 1 diabetes.

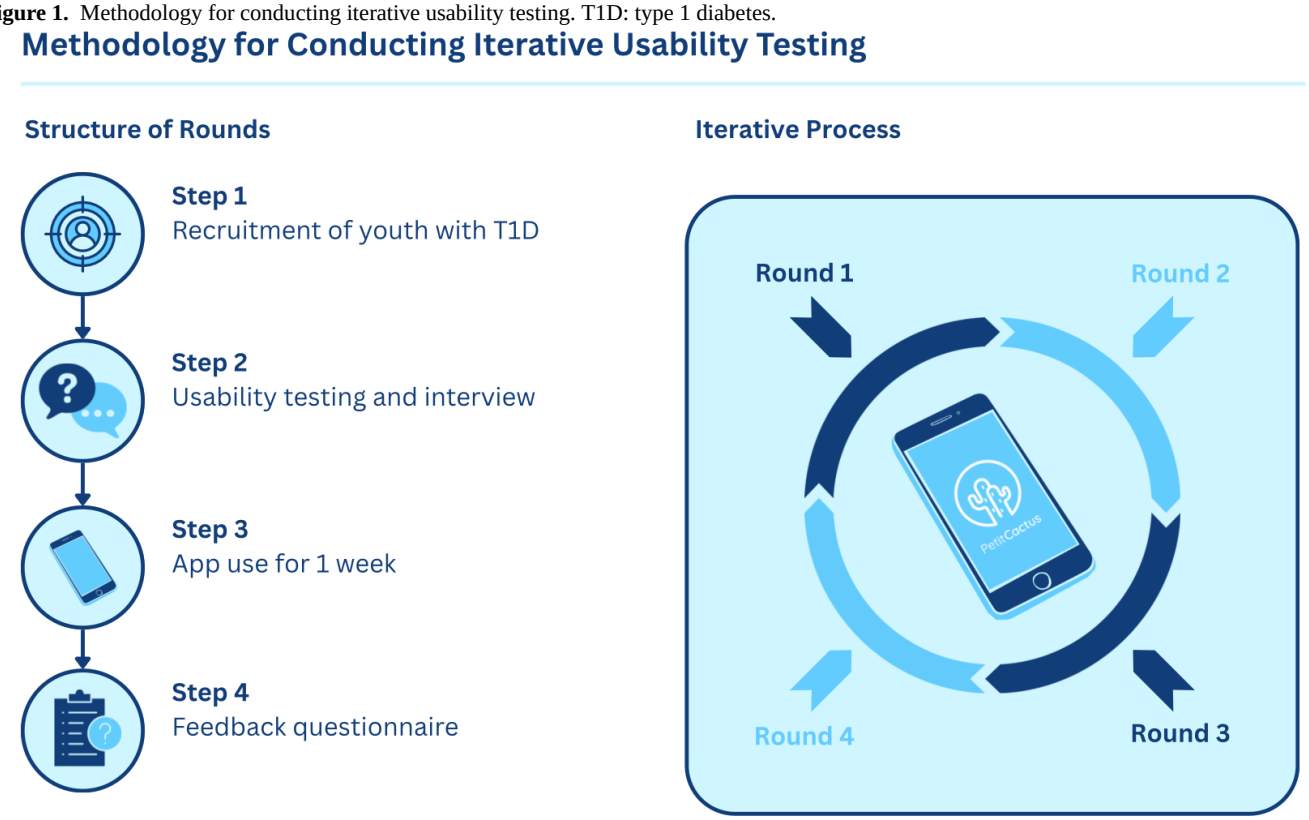

Eligibility Criteria

Young adults aged 14 to 24 years, with a T1D diagnosis of more than 1 year, using intensive insulin therapy (ie, basal-bolus injections or insulin pump therapy), and residing in Canada, were eligible to participate in the usability testing.

Participant Recruitment and Consent

Usability testing typically requires an average of 5 participants, as it was found to be generally sufficient in identifying significant issues for each version [23]. Recruitment was conducted using convenience sampling from BETTER (Behaviors, Therapies, Technologies and Hypoglycemic Risk in Type 1 Diabetes Registry), a volunteer registry of individuals living with T1D in Canada [24], Canadian diabetes organizations, social media, and word-of-mouth. Prior to enrollment in the study, each participant provided an informed consent and completed a demographic and background questionnaire on LimeSurvey (LimeSurvey GmbH) hosted on secure servers.

Ethical Considerations

The study was approved by McGill University Research Ethics Board (#23-04-018). Written informed consent was obtained before participation. Adults aged 18 years or older provided their own consent, and parents or legal guardians provided consent for participants aged 14 to 17 years. All data collected were deidentified to ensure privacy and confidentiality and were hosted on secure McGill servers. Upon completion of the usability testing interview and questionnaire, participants were provided a CAD $50 (US $36.44) gift card of their choice from a selection of retailers as a compensation for their time and participation in the study. The study was conducted in accordance with the principles of the Declaration of Helsinki for research involving humans and/or human data.

Procedures

Following consent, participants were contacted by email to schedule a usability interview. Subsequently, a link to a Microsoft Teams meeting was sent. One or two days preceding the interview, participants received a user guide to the app with access and log-in information. Interviews were 60 to 90 minutes in length and included the participant, the interviewer, and an observer who took notes to complement verbatim transcripts, when necessary. Prior to commencing the interview, participants were reminded of the objective of the usability testing and their role in the co-design process: to provide feedback on their impression of the app and necessary changes required to improve its functionality. Each session consisted of two components: (1) task-based usability testing and (2) a semistructured interview. During the usability testing component, participants were asked to interact with the app in a sequential order, starting with completing their user profile and dashboard, and then later trialing core features (ie, meal entry and the bolus insulin calculator; Multimedia Appendix 1). Using a think-aloud approach, participants verbalized their thoughts while completing predefined tasks. Probing questions from the usability guide were used as needed to clarify navigation challenges, functionality, design, and ease of use. Following task completion, a semistructured interview was conducted to explore participants' broader perceptions of the app, expectations, and experiences using the app. Following the interview, participants used the app for one week and were then requested to complete a final questionnaire including 8 open-ended questions. The questionnaire was developed by the research team to better understand potential barriers to app use,

XSL•FO
RenderX

the usability of core features (ie, the food journal, meal entry and CC, readability of glucose graphs, desired support from health care providers), motivations for trying the app, and suggestions to enhance engagement (Multimedia Appendix 2). The open-ended format was selected to allow participants to elaborate on their experiences and identify unanticipated issues.

Data Collection

The interviews were recorded and transcribed using Microsoft Teams. Three members of the research team (AH, AS, and GG) verified the accuracy of the transcripts and anonymized all identifiable information. The transcripts were then translated from French to English using a forward-backward translation method [25], when necessary by AH and verified by GG and AS, all of whom are fully bilingual.

Data Analysis

The transcripts were open coded independently by 3 researchers (AH, AS, and GG), using MAXQDA (version 24; VERBI GmbH), to develop a preliminary code book [26]. Interviews were conducted in 4 rounds, with thematic saturation reached within each round, defined as the point at which no new ideas emerged and no new codes were identified, and overall saturation achieved by the fourth round. Two researchers then independently coded all transcripts using the finalized codebook. To ensure consistency, 2 transcripts were initially co-coded, and discrepancies were discussed until consensus was reached, leading to refinement of code definitions. The remaining transcripts were coded independently, with regular meetings held to compare coding decisions and resolve any discrepancies through discussion. An inductive thematic analysis was conducted to interpret the data [27]. After transcripts were first coded, codes were then iteratively grouped and refined to identify themes based on patterns and similarities across the dataset.

Results

Participants

Twenty participants were recruited and provided consent. However, 2 participants were excluded due to not meeting the inclusion criteria for age and T1D diagnosis, despite self-reporting eligibility in the initial online survey. A total of 18 eligible participants completed the interview and the questionnaire. About two-thirds of the participants identified as women (n=12, 66.7%), and Caucasian (n=11, 61.1%). The median (IQR) for age was 23 (19-24) years and for diabetes duration was 9 (6-12) years. Almost all participants (n=17, 94.4%) used CGM systems. The majority used an insulin pump to administer insulin (n=14, 77.8%), and 12 (66.7%) used an automated insulin delivery (AID) system. Participants' characteristics are outlined in Table 1.

XSL•FO
RenderX

Table . Participant characteristics.

Characteristics	Value
Demographics (N=18)	
Age (y), median (IQR)	23 (19-24)
Gender (women), n (%)	12 (66.7)
Country of birth (Canada), n (%)	15 (83.3)
Ethnicity[a], n (%)	
Arab	1 (5.6)
Black (African, Afro-American, and Caribbean)	3 (16.7)
Indigenous (First Nation)	1 (5.6)
White or Caucasian	11 (61.1)
Currently studying[b], n (%)	13 (72.2)
Highest level of education acquired, n(%)	
High school level	7 (38.9)
CEGEP[d], vocational or community college	6 (33.3)
University level	5 (27.8)
Medical history and diabetes management (N=18)	
Diabetes duration (y), median (IQR)	9 (6-12)
Had at least 1 diabetes-related consultation in the last year with, n (%)	
Medical specialist (endocrinologist, pediatrician, or internist)	17 (94.4)
Family physician or general practitioner	15 (83.3)
Registered dietitian	12 (66.7)
Registered nurse	12 (66.7)
Use continuous glucose monitoring systems (N=18), n (%)	17 (94.4)
Dexcom G6	10 (55.6)
Freestyle Libre	1 (5.6)
Freestyle Libre 2	1 (5.6)
Guardian Sensor 3	2 (11.1)
Guardian Sensor 4	3 (16.7)
Use insulin pens or syringes exclusively (N=18), n (%)	4 (22.2)
Use an insulin pump (N=18), n (%)	14 (77.8)
Use of pump features in the last week[c] (n=14), n (%)	
Use an integrated bolus wizard or calculator function	14 (100)
Use an automated insulin pump	12 (86)

[a]Two participants do not know their ethnicities.

[b]One participant preferred not to answer.

[c]Proportion among the 14 pump users only.

[d]CEGEP: Collège d'Enseignement Général et Professionnel (general and professional college).

Critical Issues and User Feedback Across Rounds: Thematic Analysis

Overview

Data from the usability testing of the CC app with young adults with T1D was analyzed using thematic analysis through which usability issues and user preferences identified were coded and iteratively grouped based on shared patterns across 4 rounds. This process resulted in 7 themes that represent recurring usability priorities and corresponding design principles. The codes and emerging themes, representing distinct patterns across all rounds of the dataset, are presented in Figure 2. A summary table of all themes, associated codes, and illustrative participant quotes is provided in Multimedia Appendix 3. Evaluations and

XSL•FO
RenderX

suggestions for the app's features were compiled to illustrate how the features evolved across the usability testing rounds (Table 2).

Figure 2. Themes and codes identified through usability testing of a type 1 diabetes (T1D) management app with youth.

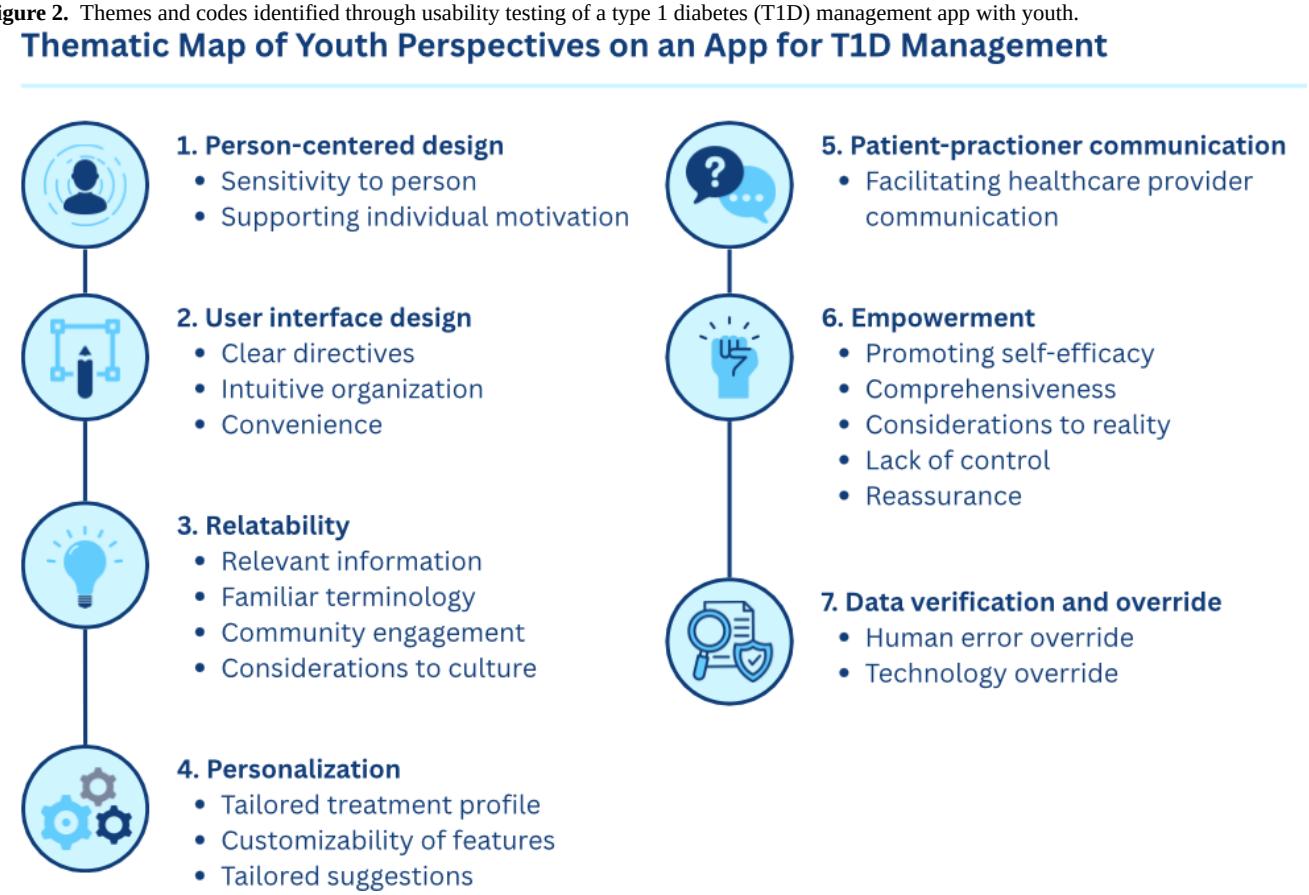

XSL•FO
RenderX

Table . Feature development through usability testing rounds.

Features	Round 1	Round 2	Round 3	Round 4
Photo recognition for carbohydrate suggestions	The photo capture feature for carbohydrate estimations was presented unclearly, leading some participants to mistake it for a tool to simply add pictures of meals.	User feedback was incorporated to improve the presentation of the photo recognition feature, which was clarified by adding prompts to explain its function. Photo uploading remained slow.	Ease of use was enhanced by adding simple options to edit food items incorrectly recognized by AI[a].	The algorithm was optimized for accuracy, with faster photo uploading.
Meal logging	Meal logging was designed with step-by-step navigation instead of showing everything on one page. However, it was overly complex, unintuitive, and too lengthy. The food database also had issues with irrelevant items and inaccurate search results.	The food database was subsequently improved by adding a better search function that presents key food items first, while retaining a large database.	The next iteration addressed this by displaying everything on the same page with streamlined manual entry and dropdown features. The food database was refined to prioritize commonly used items, but estimating quantities remained challenging.	The final iteration enhanced functionality by allowing users to copy or edit previously logged meals, offering more units of measure and options to determine food quantities. It also included options to add notes or assign meal tags (eg, high fat, high protein, pre or post-workout, or injection timing).
Event logging	The feature was not included in this round, but users suggested it to track events other than meals that could impact glucose levels.	Initial concept presented.	The full feature was added, allowing users to track various events beyond meals, such as glucose and insulin (independent of meals), and physical activity.	More events were added, such as the menstrual cycle, medication, and a customizable event, to enable more comprehensive event tracking.
Insulin dosing calculator	Basic model for dosing recommendations.	No changes were made.	The feature now presents an added safety measure. Users are shown their glucose levels and ratios and must confirm entries before calculations.	No changes were made.
Dashboard	The dashboard featured a basic design with limited functionality, presenting statistics on the number of meals, units of insulin, and glucose levels, all compiled over the course of the day.	User feedback was incorporated, resulting in a clearer visual presentation and the addition of adjustable timeframes for compiling statistics.	The final design included a personalized data display with the addition of a graph to show glucose levels alongside logged events (meals, activity, stress, and others).	When the CGM[b] device is connected, glucose levels are displayed in real time, and the cactus avatar dynamically changes to reflect these levels.
Profile settings	Basic settings included name, age, date of birth, gender, target glucose levels, insulin modality, glucose monitoring preferences, and app language.	Settings were expanded to enable users to enter multiple correction factors, input their basal insulin regimen, as well as the option for users to remove information irrelevant to their diabetes management.	Further refinements were made, including the option to add a profile picture or an avatar.	No changes were made.
Connection to a health care provider portal	Not available. Participants recommended a direct method to share data with the health care team.	Initial concept of provider connection presented.	The prototype was tested with basic integration, including an option to link the account with a health care team portal.	No changes were made.
AI chatbot	The AI chatbot was presented unclearly, often mistaken for a help center.	Prompts and examples of queries were added to clarify its use.	In the final iteration, the AI was integrated to provide personalized feedback based on previous discussions.	No changes were made.

[a]AI: artificial intelligence.

[b]CGM: continuous glucose monitoring.

XSL•FO
RenderX

Theme 1: Person-Centered Design

This theme reflects the importance of designing the app with youth living with T1D in mind, prioritizing their perspectives and lived experiences rather than focusing solely on what may be deemed important from a clinician-centered perspective. Participants consistently emphasized the need for the app to acknowledge the person living with T1D, not only the condition it aims to support. This theme evolved across the 4 rounds to reflect a progression and growing interest in emotional well-being and sensitivity to stigma within the app's design.

Initially, participants highlighted the importance of prioritizing personal profile settings (eg, name, gender, and date of birth) before treatment-related parameters during onboarding. This approach was appreciated for making the app feel more personal.

In later rounds, participants further distinguished information relevant to health care providers from that intended for users with T1D. While clinical parameters such as blood glucose targets were considered essential during medical appointments, participants expressed a desire not to have these metrics displayed at all times. Suggestions included the ability to customize dashboard displays, such as hiding A_{1c} levels, which some found stress-inducing. They also recommended avoiding design elements that could unintentionally reinforce negative perceptions (eg, a thumbs down for late insulin administration) and replacing red and green indicators with neutral colors to classify food choices:

> *Especially when you're a young teenager and you see your A_{1c}'s high, it can stress you out. Maybe it could be an option to hide it on your dashboard.* [Woman, 20 years, 13 years living with T1D]

> *Say for example I had a bad few days I wouldn't want to remind myself of it all the time.* [Man, 24 years, 9 years living with T1D]

The final round added features to support individual motivation. For instance, participants suggested integrating challenges and objectives that align with users' preferences for managing their diabetes, focusing not just on what they need to do but on how they want to experience and engage with the process. This also included offering peer support and the option to send encouragement messages to others within the app, if they consent to it.

Theme 2: User Interface Design

During the first round, users identified foundational problems. They reported significant challenges with navigation, struggling to locate key elements such as the photo recognition feature, the dashboard, the settings and treatment profile, and the help menus. There was a lack of clear directives, leaving users uncertain about what needed to be entered and unclear about the descriptions of what was displayed. Prompts, descriptors, and visual cues were suggested to optimize the user's comprehension of entry requirements.

By round 2, navigation improvements reduced these challenges; however, multistep processes remained cumbersome. For instance, meal logging required navigating through multiple pages (first to confirm food items recognized by the AI algorithm, then confirm the detected quantities, and finally to confirm bolus calculations), whereas users suggested a streamlined approach using a list or drop-down menu in the same page. Another important aspect is intuitive organization, ensuring that the order of information display is practical. For example, users instinctively clicked on their profile picture to navigate back to their profile.

In later rounds, critical navigation issues were resolved, and only minor refinements were suggested to improve convenience, such as faster photo upload and quicker analysis of the photos. Participants also highlighted the value of automatic logging of meals in the food journal, when a photo is used to estimate carbohydrate content, along with glucose trends and insulin dosing. They further requested simple data extraction for the food journal and the ability to filter and sort information efficiently. This was further reinforced in the feedback questionnaire (Multimedia Appendix 4).

Theme 3: Relatability

This theme focused on enhancing relatability by providing relevant and accessible information while minimizing complexity. In round 1, participants emphasized the need to streamline the food journal feature by avoiding overly detailed descriptions in the database and instead prioritizing a simple, comprehensive list of commonly used items. The importance of using familiar terminology was highlighted to improve user comprehension and align with user preferences (eg, correction factor vs insulin sensitivity factor, 1 cup vs 250 ml).

In later rounds, fostering community engagement became a key focus. Participants proposed adding an in-app peer messaging feature, and a *Frequently Asked Questions* section to promote peer feedback and provide actionable insights. Additionally, participants stressed the importance of cultural inclusivity by ensuring that the app's content, such as food items in the database, is relevant and accessible to users from diverse backgrounds:

> *I think that the different cultures and cultural foods might be something to take into consideration. [...] We really need to make sure there are many different people and many different cultures represented.* [Woman, 22 years, 11 years living with T1D]

Theme 4: Personalization

In earlier rounds, participants emphasized the importance of tailoring the treatment profile to individual needs, including options for CGM connection, weight tracking, injection reminders, and specifying insulin treatment types (eg, pen or injections vs pump). This was particularly relevant when discussing the dashboard and the types of statistics users wanted displayed, depending on their insulin delivery methods and glucose level monitoring preferences. Beyond diabetes management, customizability of the app itself is important. Participants expressed a desire to personalize its esthetics, such as colors and design, as well as practical features like glucose units, custom tags, data visualization, and the organization of the features to align with their individual routines and preferences.

XSL•FO
RenderX

By later rounds, participants suggested the incorporation of AI technology to provide tailored diabetes management recommendations based on user trends and status:

> *If, say you've had too many [hyperglycemic episodes] and it's dinner time. [...] Maybe the AI could give suggestions based on your statistics for the day.* [Man, 18 years, 12 years living with T1D]

Theme 5: Patient-Practitioner Communication

Participants highlighted the potential for the app to enhance patient-practitioner communication, facilitating more effective diabetes management. Initially, participants reported that these features could empower users to take charge of their management during appointments with their health care team, fostering more productive discussions:

> *I went to see the dietitian, and I had to provide a dietary recall, but trying to remember everything I'd eaten, it's definitely more tedious. But to see that [the food journal] is accessible, to have it on hand all the time, and to be able to enter it as I go along, I find it really practical.* [Woman, 19 years, 10 years living with T1D]

By later rounds, the emphasis shifted to refining and expanding these features. Participants suggested generating downloadable reports that combine dietary data, automatically generated from photos taken with the app, with glucose levels and insulin doses, to facilitate sharing with health care professionals and supporting more informed decisions.

Theme 6: Empowerment

The theme of empowerment focuses on how the app can enhance users' confidence and ability to manage T1D effectively, while addressing concerns about over-reliance on technology. In earlier rounds, participants emphasized how a CC app could empower young adults with T1D by promoting self-efficacy. The app's ability to centralize tools and information into a single platform was highlighted as essential for simplifying diabetes management. Features like long-term tracking and trend analysis allowed users to understand factors influencing glycemia, such as meals, activity, stress, and timing of insulin injections:

> *If I look at my graph later and see that one day I was trending high all day, but then I look back at the food journal and remember what I ate, I'll understand better.* [Woman, 19 years, 10 years since diagnosis]

Additionally, the app reassured users by validating daily behaviors, even those deviating from guidelines, providing a realistic and supportive approach to condition management:

> *I really liked [the feature of inputting insulin before or after a meal]. I sometimes forget to give myself insulin before I eat. I think it's important to see the trend.* [Man, 18 years, 12 years living with T1D]

In later rounds, participants expressed appreciation for these empowering features but raised concerns about the potential trade-off between reliance on technology and their sense of control. Some users noted skepticism regarding the accuracy of the app's recommendations, which they felt could undermine their self-efficacy.

Discussions emphasized the importance of building user trust to address these concerns. Participants suggested that transparent mechanisms to verify the app's outputs and ensure accuracy would help mitigate doubts. This evolution across rounds demonstrates the balance needed between empowering users and maintaining their confidence in the app as a reliable support tool for diabetes management. This is further highlighted in the feedback questionnaire, where participants expressed reluctance to rely on AI-driven photo recognition if the carbohydrate content of meals is already known (Multimedia Appendix 4).

Theme 7: Data Verification and Override

Participants highlighted the importance of data verification and override capabilities within the app to ensure accurate diabetes management. In earlier rounds, the focus was on addressing critical issues, such as implementing mechanisms to override technological errors and the ability to accept or reject automated suggestions made by the app. They also stressed the importance of addressing human error by incorporating prompts that encourage users to verify and correct data entries to enhance reliability.

By later rounds, the emphasis shifted to refining these features, including streamlining the verification process to avoid unnecessary disruptions during app use. Participants suggested simplifying error notifications to ensure they were clear but not intrusive and incorporating an option to review and confirm CGM readings. These minor improvements aimed to strike a balance between maintaining user control and creating a seamless experience.

Discussion

Principal Results

The usability testing revealed 7 major themes highlighting key design principles. These themes focused on young adults' empowerment, enhancing patient-practitioner communication, data verification and override for accuracy in diabetes management, an intuitive user interface design, and the relatability of the information provided. While both the person-centered design and personalization themes reflect participants' desire for individual relevance, they differ in scope and intent. Person-centered design represents a broader guiding principle emphasizing empathy and how the app acknowledges the person behind the condition. In contrast, personalization refers to the practical and technological implementation of this principle ie, how the app's content, settings, and recommendations adapt to the user's specific needs and data.

Comparison With Prior Work

Clinical guidelines recommend adjusting meal-time insulin dose to manage glucose variability due to different meal compositions [28,29]. However, there is no consensus as to which insulin strategies to use. Significant interindividual differences in insulin dose requirements exist, necessitating personalized advice based on postprandial glucose monitoring for several hours afterwards [30]. This is consistent with a major emerging theme: empowering users to gain control over their T1D management by understanding personal trends and factors affecting glycemia. While users appreciated receiving personalized AI-generated

XSL•FO
RenderX

recommendations based on their trends, the sense of empowerment mainly came from accessing personal statistics and tracking various factors.

The interest to learn about personal trends aligned with participants' desire to remain in control and avoid full reliance on the app, which could hinder self-efficacy. Some users were skeptical about the accuracy of AI-generated information, emphasizing the need for safety mechanisms for data verification and the ability to override when necessary. This concern aligns with previous research on the lag between interstitial and capillary glucose levels, which led to data distrust and prompted some CGM users to rely on capillary glucose to clarify high and low readings [31]. Similarly, a recent study examining patient perspectives on AI in health care highlighted varying levels of comfort depending on the role of AI in care, with higher acceptance of apps that play an assistive role (eg, digital scribing or radiology review), and lower comfort when AI systems directly influenced decision-making [32]. Together, these findings highlight the tension between empowerment and over-reliance on technology in contexts where trust in automated systems is still evolving.

This sense of empowerment was further emphasized in the context of patient-practitioner communication. Participants reported that access to their own data enabled them to take a more active role in diabetes management by sharing relevant information and tracking topics they wished to discuss during medical appointments. This form of empowerment is particularly important during adolescence, a time when young people are striving to gain independence and shape their personal and social identities [33]. It is especially relevant given evidence that young adults with T1D often report health care needs and priorities that significantly differ from those perceived by primary caregivers and health care professionals [34]. Consistent with literature on shared medical appointments, which emphasize patient leadership and active participation in care, greater access to and understanding of personal health data has been shown to enhance patients' ability to engage more confidently in discussions with clinicians, supporting more collaborative and patient-led care [35].

Importantly, access to information must be implemented through a person-centered lens to balance what information youth want to access and *how* they want to engage with it. For example, participants did not want glucose data displayed on their dashboard each time they opened the app. It is well documented in the literature that living with and managing T1D is associated with significant psychological burden, with young people at increased risk of psychiatric disorders compared to peers without diabetes [36]. More specifically, data overload and constant exposure to CGM and insulin pump data have been identified as sources of stress [37]. Notably, while CGM use has been associated with improved well-being among young people, only a minority report regularly engaging in detailed data analysis [38]. These findings highlight the importance of selectively presenting information in digital tools intended to support diabetes management. At the individual level, participants voiced different needs: some valued a comprehensive app that consolidates all information necessary for diabetes management and reduces the need to juggle multiple sources, while others felt that repeatedly displaying data such as glucose readings or insulin doses was redundant, given its availability through CGM and insulin pump reports. Filtering information, by suppressing, deleting, or selecting irrelevant information, has been identified as a helpful strategy for mitigating information overload [39]. This aligns with personalization as a key design principle, enabling users to determine which factors to track, how data are visualized, and which trends are emphasized.

Participants also expressed a desire to access consolidated data, with logged meals alongside corresponding glucose values, to better understand personal patterns. The multifactorial nature of glycemic management is well established, as insulin requirements are influenced by factors beyond carbohydrate intake, including physical activity, stress, and illness [40]. This complexity highlights the importance of tools that can help support decision-making in diabetes management. Consistent with this, findings from a needs assessment survey emphasized the importance of tracking factors beyond dietary intake, such as hormonal changes, timing of insulin delivery, and sleep, alongside glucose data [21].

These considerations are particularly important given clinical appointments often prioritize glucose data and glycemic management, which, although essential, may leave limited space to address broader psychosocial, behavioral, and contextual factors that shape day-to-day management, largely due to time constraints [41]. Designing the app around youths' lived experiences, rather than solely around clinical priorities, helps bridge this gap and ensures greater relevance and usability. Incorporating features that reflect real-world practices, such as tracking whether insulin was administered before or after meals despite guidelines recommending injecting before meals [42], acknowledges daily realities and avoids setting unrealistic expectations. Consistent with this, a study on the co-design of a digital mental health platform for young adults highlighted the importance of relatability, with participants emphasizing that the platform should be one they can identify with to support meaningful engagement [43].

Strengths and Limitations

These results should be interpreted in light of the study's limitations. A major limitation is that most participants were users of AID, relying on systems that automate insulin delivery and integrate CGM data. Consequently, feedback may under-represent the perceived value of features aimed at supporting manual bolus calculations, while overrepresenting preferences for features that support retrospective reflection and learning from personal trends rather than real-time decision support. In addition, AID users may represent a subgroup with higher baseline engagement with diabetes technologies, greater digital literacy, and increased comfort with data interpretation. While some app functions may overlap with AID features, the app serves a complementary rather than redundant role. AID systems focus on insulin automation but provide limited insight into the behavioral and contextual factors influencing glucose stability. In contrast, the proposed app supports self-awareness, education, and behavioral reflection by linking glucose trends with meal composition, stress, physical activity, and other lifestyle factors. It also facilitates communication with health

XSL•FO
RenderX

care providers through integrated reports that combine these contextual data. Finally, its relevance extends to individuals on basal-bolus injection therapy, who continue to make daily dosing decisions manually, and to those without access to AID systems due to cost, eligibility, or access limitations. Other limitations include the potential selection bias arising from the nonrandom sample, which may affect generalizability. However, the users represented various ethnicities, genders, education levels, and insulin treatment modalities. Researcher bias was mitigated by using the same interviewer throughout, ensuring consistency in questioning style, tone, and data collection procedures across rounds. In addition to having an observer present during the interviews, subjectivity in data analysis was addressed by having 3 coders reach consensus through discussions. Observer bias was minimized by adhering to a preset questionnaire and having no prior relationships with participants. Due to the potential for authority and response bias, the interviewer mitigated their impact by establishing rapport with participants to create a comfortable and nonjudgmental environment, used neutral language and tone to avoid perceived authority, and encouraged openness by reassuring participants that their honest opinions were valued.

A key strength of the study is that end user needs directly informed the app's development, with participant input integrated iteratively across all design rounds. This co-design approach ensured that the resulting features were grounded in lived experience and reflected the priorities of young adults with T1D. By continuously refining the app based on user feedback, the design process enhanced relevance and usability, laying a strong foundation for future pilot testing and real-world engagement.

Conclusions

Petit Cactus is a mobile app that enables carbohydrate estimations through photo recognition and supports automated insulin dosing within a single, user-informed platform. Through its co-design process, the study identified design principles that extend beyond this specific application, emphasizing person-centered, intuitive, and emotionally sensitive interfaces that align with users' daily routines and language. Participants also underscored the importance of personalization, data accuracy, and effective communication tools to support meaningful interaction with health care providers. These insights could inform the design of digital health tools aimed at supporting diabetes self-management. Usability testing has refined Petit Cactus accordingly, and a pilot trial will next evaluate its impact on engagement and glycemic outcomes among young adults with T1D.

Funding

This work was supported by Mitacs (grant IT31370 ACC). AH received a Mitacs Accelerate scholarship. GG received a Natural Sciences and Engineering Research Council of Canada summer scholarship. AK received a Canadian Graduate Scholarship–Master's Program. CS received a Fonds de Recherche du Québec en Santé–Doctoral scholarship. ASB is a Fonds de Recherche du Québec en Santé research scholar and Mitacs grant recipient. All other authors received no financial support for the research, authorship, and/or publication of this article.

Data Availability

The datasets used and analyzed during the study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: AH, ASB
Formal analysis: AH
Interview facilitation: AH
Interview observation: AS, GG, CS
Methodology: AH
Participation recruitment: AH
Protocol – original draft: AH
Protocol – review & editing: AH, ASB
Project administration: AH
Supervision: ASB
Visualization: AH
Writing – original draft: AH, AS, AK, GG
Writing – review & editing: AH, AS, AK, GG, AR-F, CS, MN, ASB

Multimedia Appendix 1
Interview guide and questions.
[DOCX File, 18 KB - jmir_v28i1e86024_app1.docx]

Multimedia Appendix 2
Feedback questionnaire.

XSL•FO
RenderX

[DOCX File, 18 KB - jmir_v28i1e86024_app2.docx]

Multimedia Appendix 3
Themes and codes description
[DOCX File, 27 KB - jmir_v28i1e86024_app3.docx]

Multimedia Appendix 4
Appreciations and suggestions from the feedback questionnaire after 1 week of app use.
[DOCX File, 19 KB - jmir_v28i1e86024_app4.docx]

References

1. Ilonen J, Lempainen J, Veijola R. The heterogeneous pathogenesis of type 1 diabetes mellitus. Nat Rev Endocrinol 2019 Nov;15(11):635-650. [doi: 10.1038/s41574-019-0254-y] [Medline: 31534209]
2. Diabetes in Canada: Backgrounder. : Diabetes Canada; 2021 URL: https://www.diabetes.ca/DiabetesCanadaWebsite/media/Advocacy-and-Policy/Backgrounder/2021_Backgrounder_Canada_English_FINAL.pdf [accessed 2026-03-13]
3. Diabetes Canada Clinical Practice Guidelines Expert Committee, McGibbon A, Adams L, Ingersoll K, Kader T, Tugwell B. Glycemic Management in Adults With Type 1 Diabetes. Can J Diabetes 2018 Apr;42 Suppl 1:S80-S87. [doi: 10.1016/j.jcjd.2017.10.012] [Medline: 29650115]
4. Diabetes Control and Complications Trial Research Group, Nathan DM, Genuth S, et al. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. N Engl J Med 1993 Sep 30;329(14):977-986. [doi: 10.1056/NEJM199309303291401] [Medline: 8366922]
5. Diabetes Control and Complications Trial Research Group. Effect of intensive diabetes treatment on the development and progression of long-term complications in adolescents with insulin-dependent diabetes mellitus: diabetes control and complications trial. J Pediatr 1994 Aug;125(2):177-188. [doi: 10.1016/S0022-3476(94)70190-3]
6. Gonzalez C, Picón MJ, Tomé M, Pujol I, Fernández-García JC, Chico A. Expert study: Utility of an automated bolus advisor system in patients with type 1 diabetes treated with multiple daily injections of insulin-A crossover study. Diabetes Technol Ther 2016 May;18(5):282-287. [doi: 10.1089/dia.2015.0383] [Medline: 26886163]
7. Tascini G, Berioli MG, Cerquiglini L, et al. Carbohydrate counting in children and adolescents with type 1 diabetes. Nutrients 2018 Jan 22;10(1):109. [doi: 10.3390/nu10010109] [Medline: 29361766]
8. Spiegel G, Bortsov A, Bishop FK, et al. Randomized nutrition education intervention to improve carbohydrate counting in adolescents with type 1 diabetes study: Is more intensive education needed? J Acad Nutr Diet 2012 Nov;112(11):1736-1746. [doi: 10.1016/j.jand.2012.06.001] [Medline: 22975086]
9. Singla R, Bindra J, Singla A, Gupta G, Gupta Y, Aggarwal S. Automation of insulin bolus dose calculation in type 1 diabetes: a feasibility study. Int J Diabetes Dev Ctries 2023 Feb;43(1):66-71. [doi: 10.1007/s13410-022-01054-7]
10. Sussman A, Taylor EJ, Patel M, et al. Performance of a glucose meter with a built-in automated bolus calculator versus manual bolus calculation in insulin-using subjects. J Diabetes Sci Technol 2012 Mar 1;6(2):339-344. [doi: 10.1177/193229681200600218] [Medline: 22538144]
11. Borus JS, Laffel L. Adherence challenges in the management of type 1 diabetes in adolescents: prevention and intervention. Curr Opin Pediatr 2010 Aug;22(4):405-411. [doi: 10.1097/MOP.0b013e32833a46a7] [Medline: 20489639]
12. Shah BR, Lipscombe LL, Booth GL. Glycemic control among people with diabetes in Ontario: a population-based cross-sectional study. Can J Diabetes 2021 Jun;45(4):313-318. [doi: 10.1016/j.jcjd.2020.09.017] [Medline: 33221136]
13. Foster NC, Beck RW, Miller KM, et al. State of type 1 diabetes management and outcomes from the T1D exchange in 2016-2018. Diabetes Technol Ther 2019 Feb;21(2):66-72. [doi: 10.1089/dia.2018.0384] [Medline: 30657336]
14. Gurnani M, Pais V, Cordeiro K, Steele S, Chen S, Hamilton JK. One potato, two potato,… assessing carbohydrate counting accuracy in adolescents with type 1 diabetes. Pediatr Diabetes 2018 Nov;19(7):1302-1308. [doi: 10.1111/pedi.12717] [Medline: 29999219]
15. Klonoff DC. The current status of mHealth for diabetes: Will it be the next big thing? J Diabetes Sci Technol 2013 May 1;7(3):749-758. [doi: 10.1177/193229681300700321] [Medline: 23759409]
16. Liu NF, Brown AS, Folias AE, et al. Stigma in people with type 1 or type 2 diabetes. Clin Diabetes 2017 Jan;35(1):27-34. [doi: 10.2337/cd16-0020] [Medline: 28144043]
17. Whitehead L, Seaton P. The effectiveness of self-management mobile phone and tablet apps in long-term condition management: a systematic review. J Med Internet Res 2016 May 16;18(5):e97. [doi: 10.2196/jmir.4883] [Medline: 27185295]
18. Liang X, Wang Q, Yang X, et al. Effect of mobile phone intervention for diabetes on glycaemic control: a meta-analysis. Diabet Med 2011 Apr;28(4):455-463. [doi: 10.1111/j.1464-5491.2010.03180.x] [Medline: 21392066]
19. Vasiloglou MF, Marcano I, Lizama S, Papathanail I, Spanakis EK, Mougiakakou S. Multimedia data-based mobile applications for dietary assessment. J Diabetes Sci Technol 2023 Jul;17(4):1056-1065. [doi: 10.1177/19322968221085026] [Medline: 35348398]

XSL•FO
RenderX

20. Eiland L, McLarney M, Thangavelu T, Drincic A. App-based insulin calculators: current and future state. Curr Diab Rep 2018 Oct 4;18(11):123. [doi: 10.1007/s11892-018-1097-y] [Medline: 30284645]
21. Housni A, Katz A, Bergeron LJ, et al. Bridging the gap in carbohydrate counting with a mobile app: needs assessment survey. J Med Internet Res 2025 Mar 28;27:e63278. [doi: 10.2196/63278] [Medline: 40153793]
22. Gulliksen J, Göransson B, Boivie I, Blomkvist S, Persson J, Cajander Å. Key principles for user-centred systems design. Behav Inf Technol 2003 Nov;22(6):397-409. [doi: 10.1080/01449290310001624329]
23. Alwashmi MF, Hawboldt J, Davis E, Fetters MD. The iterative convergent design for mobile health usability testing: mixed methods approach. JMIR Mhealth Uhealth 2019 Apr 26;7(4):e11656. [doi: 10.2196/11656] [Medline: 31025951]
24. Brazeau AS, Messier V, Talbo MK, et al. Self-reported severe and nonsevere hypoglycemia in type 1 diabetes: population surveillance through the BETTER Patient Engagement Registry: development and baseline characteristics. Can J Diabetes 2022 Dec;46(8):813-821. [doi: 10.1016/j.jcjd.2022.05.010] [Medline: 35835670]
25. Pan Y, de la Puente M. Census bureau guideline for the translation of data collection instruments and supporting materials: documentation on how the guideline was developed. : Statistical Research Division, US Bureau of the Census; 2005 URL: https://www.census.gov/content/dam/Census/library/working-papers/2005/adrm/rsm2005-06.pdf [accessed 2026-03-13]
26. Zeynali-Tazehkandi M. Analyzing qualitative data with MAXQDA: a book review. Inf Discov Deliv 2022;50 [FREE Full text]
27. Kiger ME, Varpio L. Thematic analysis of qualitative data: AMEE Guide No. 131. Med Teach 2020 Aug;42(8):846-854. [doi: 10.1080/0142159X.2020.1755030] [Medline: 32356468]
28. Pańkowska E, Błazik M, Groele L. Does the fat-protein meal increase postprandial glucose level in type 1 diabetes patients on insulin pump: the conclusion of a randomized study. Diabetes Technol Ther 2012 Jan;14(1):16-22. [doi: 10.1089/dia.2011.0083] [Medline: 22013887]
29. Smart CE, Annan F, Higgins LA, Jelleryd E, Lopez M, Acerini CL. ISPAD Clinical practice consensus guidelines 2018: Nutritional management in children and adolescents with diabetes. Pediatr Diabetes 2018 Oct;27:136-154. [doi: 10.1111/pedi.12738] [Medline: 30062718]
30. Bell KJ, Toschi E, Steil GM, Wolpert HA. Optimized mealtime insulin dosing for fat and protein in type 1 diabetes: application of a model-based approach to derive insulin doses for open-loop diabetes management. Diabetes Care 2016 Sep;39(9):1631-1634. [doi: 10.2337/dc15-2855] [Medline: 27388474]
31. Lawton J, Blackburn M, Allen J, et al. Patients' and caregivers' experiences of using continuous glucose monitoring to support diabetes self-management: qualitative study. BMC Endocr Disord 2018 Feb 20;18(1):12. [doi: 10.1186/s12902-018-0239-1] [Medline: 29458348]
32. Foresman G, Biro J, Tran A, et al. Patient perspectives on artificial intelligence in health care: focus group study for diagnostic communication and tool implementation. J Particip Med 2025 Jul 24;17:e69564. [doi: 10.2196/69564] [Medline: 40705399]
33. King KM, King PJ, Nayar R, Wilkes S. Perceptions of adolescent patients of the "lived experience" of type 1 diabetes. Diabetes Spectr 2017 Feb;30(1):23-35. [doi: 10.2337/ds15-0041] [Medline: 28270712]
34. Chiang Y, Tsay P, Chen C, et al. A Delphi study on the healthcare needs of patients with type 1 diabetes during the transition from adolescence to adulthood: consensus among patients, primary caregivers, and healthcare providers. Int J Environ Res Public Health 2021 Jul 4;18(13):7149. [doi: 10.3390/ijerph18137149] [Medline: 34281086]
35. Wadsworth KH, Archibald TG, Payne AE, Cleary AK, Haney BL, Hoverman AS. Shared medical appointments and patient-centered experience: a mixed-methods systematic review. BMC Fam Pract 2019 Jul 8;20(1):97. [doi: 10.1186/s12875-019-0972-1] [Medline: 31286876]
36. Dybdal D, Tolstrup JS, Sildorf SM, et al. Increasing risk of psychiatric morbidity after childhood onset type 1 diabetes: a population-based cohort study. Diabetologia 2018 Apr;61(4):831-838. [doi: 10.1007/s00125-017-4517-7] [Medline: 29242985]
37. Kruger S, Deacon E, van Rensburg E, Segal D. Adjustment experiences of adolescents living with well-controlled type 1 diabetes using closed-loop technology. Front Clin Diabetes Healthc 2024;5:1445972. [doi: 10.3389/fcdhc.2024.1445972] [Medline: 39483147]
38. Huhn F, Lange K, Jördening M, Ernst G. Real-world use of continuous glucose monitoring systems among adolescents and young adults with type 1 diabetes: reduced burden, but little interest in data analysis. J Diabetes Sci Technol 2023 Jul;17(4):943-950. [doi: 10.1177/19322968221081216] [Medline: 35255729]
39. Kluge A, Antoni CH, Ellwart T. Digitalization as the problem of and the solution to vast amounts of data in future work—challenges for individuals, teams, and organizations. Z Arb Organisationspsychol A&O 2020 Jan;64(1):1-5. [doi: 10.1026/0932-4089/a000317]
40. DiMeglio LA, Evans-Molina C, Oram RA. Type 1 diabetes. Lancet 2018 Jun 16;391(10138):2449-2462. [doi: 10.1016/S0140-6736(18)31320-5] [Medline: 29916386]
41. Irving G, Neves AL, Dambha-Miller H, et al. International variations in primary care physician consultation time: a systematic review of 67 countries. BMJ Open 2017 Nov 8;7(10):e017902. [doi: 10.1136/bmjopen-2017-017902] [Medline: 29118053]

XSL•FO
RenderX

42. Jovanovic L, Giammattei J, Acquistapace M, Bornstein K, Sommermann E, Pettitt DJ. Efficacy comparison between preprandial and postprandial insulin aspart administration with dose adjustment for unpredictable meal size. Clin Ther 2004 Sep;26(9):1492-1497. [doi: 10.1016/j.clinthera.2004.09.001] [Medline: 15531011]
43. Ludlow K, Russell JK, Ryan B, et al. Co-designing a digital mental health platform, “Momentum”, with young people aged 7-17: a qualitative study. Digit Health 2023;9:20552076231216410. [doi: 10.1177/20552076231216410] [Medline: 38033517]

Abbreviations:

AI: artificial intelligence
AID: automated insulin delivery
BETTER: Behaviors, Therapies, Technologies and Hypoglycemic Risk in Type 1 Diabetes Registry
CC: carbohydrate counting
CGM: continuous glucose monitoring
HbA_{1c}: glycated hemoglobin
mHealth: mobile health
T1D: type 1 diabetes

Edited by A Stone; submitted 17.Oct.2025; peer-reviewed by M Ekwueme, O Ajayi, SM Youngblood; revised version received 25.Feb.2026; accepted 25.Feb.2026; published 31.Mar.2026.

Please cite as:
Housni A, Shulkin A, Katz A, Giannini G, Roy-Fleming A, Nakhla M, South C, Brazeau AS
Designing a Carbohydrate Counting App for Young Adults With Type 1 Diabetes: Usability Testing Interview Study
J Med Internet Res 2026;28:e86024
URL: https://www.jmir.org/2026/1/e86024
doi:10.2196/86024

© Asmaa Housni, Aidan Shulkin, Alexandra Katz, Giuliana Giannini, Amélie Roy-Fleming, Meranda Nakhla, Courtney South, Anne-Sophie Brazeau. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of Postdischarge Telephone Calls in Reducing Hospital Utilization: Quasi-Randomized Controlled Trial

Sandie Du[1], MHA; Abigail Anada[1], MM; Maria Montenegro[1], MSc; Maurie Maitland[1], MSN; Sean Chuan[1], MSc; Sonia Jose[1], RN, BSN; Harveer Sihota[1], BSN, MSN; Delia Cooper[1], MSc; Joanna Rivera[1], MPH; Dianne Burditt[1], SEA; Minetaro Naruki-van Velzen[1], MSc; Megan MacPherson[2], PhD

[1]Digital Patient and Provider Experience, Fraser Health, Surrey, BC, Canada

[2]Department of Occupation Science & Occupational Therapy, Faculty of Medicine, University of British Columbia, 13737 96th Avenue Office #249, Surrey, BC, Canada

Corresponding Author:
Megan MacPherson, PhD
Department of Occupation Science & Occupational Therapy, Faculty of Medicine, University of British Columbia, 13737 96th Avenue Office #249, Surrey, BC, Canada

Abstract

Background: Unplanned emergency department (ED) visits and hospital readmissions following discharge contribute to patient distress, increased health care costs, and system inefficiencies. Early postdischarge follow-up can improve care transitions, yet evidence on the effectiveness of telephone-based interventions remains mixed. Telephone calls, a low-barrier form of digital health, may enhance equity and accessibility by reaching patients who face challenges with in-person or higher-technology follow-up.

Objective: This study evaluated the impact of a nurse-led postdischarge telephone intervention delivered by Fraser Health Virtual Care on short-term ED visits and hospital readmissions among recently discharged high-risk patients. Secondary objectives included examining patient experiences with the service and identifying care gaps addressed during follow-up calls.

Methods: A pragmatic quasi-randomized trial was conducted (May 2022-September 2022). Participants were eligible if they were aged 18 years or older and classified as high-risk for readmission using the LACE (Length of stay, Acuity of admission, Comorbidities, and Emergency department use) index (≥10 or <9 and ≥45 y). Participants were allocated to either a postdischarge telephone intervention group or a standard care control group based on daily nurse availability. Intervention participants received a structured nurse-led call 48 hours after discharge assessing understanding of discharge instructions, medication management, follow-up appointments, and home supports. Primary outcomes were ED visits within 7 and 30 days post call; secondary outcomes were hospital readmissions and patient experience. Negative binomial regression models were used to calculate adjusted incident rate ratios (IRRs).

Results: A total of 7091 participants were included (intervention: n=3911, of whom 1752 completed the call; control: n=3180). Postdischarge calls significantly reduced ED visits at both 7 days (adjusted IRR 0.719, 95% CI 0.617 - 0.837; *P*<.001) and 30 days (IRR 0.878, 95% CI 0.783 - 0.983; *P*=.02). No statistically significant reductions were observed in hospital readmissions at either 7 days (IRR 0.809; *P*=.13) or 30 days (IRR 0.942; *P*=.54). Forty percent of completed calls (n=701) identified at least 1 gap in discharge understanding or follow-up care. Most participants found the calls helpful and reported increased confidence in managing their care.

Conclusions: Structured nurse-led postdischarge telephone calls significantly reduced short-term ED utilization but did not impact readmission rates. These findings support the role of telephone-based virtual care as a scalable, low-barrier strategy to improve care transitions and reduce avoidable ED visits. Additional or ongoing interventions may be required to influence hospital readmission outcomes among high-risk patients.

Trial Registration: International Clinical Trials Registry Platform ISRCTN26233632; https://www.isrctn.com/ISRCTN26233632

(J Med Internet Res 2026;28:e80529) doi:10.2196/80529

KEYWORDS

patient readmission; hospitalization; emergency room; continuity of care; digital health

XSL•FO
RenderX

Introduction

Background

High volumes of emergency department (ED) visits and unplanned hospital readmissions following discharge have significant implications for patient outcomes, health care costs, and overall system efficiency. Readmissions within 30 days, which are often considered unplanned and avoidable, are linked to poor patient satisfaction [1], increased mortality risk [2], and substantial financial burden, accounting for over US $2.3 billion annually in Canada [3]. While not all ED visits and readmissions can be prevented, strengthening postdischarge follow-up and care coordination can help reduce these rates and improve patient experience.

The postdischarge period is a particularly vulnerable time for patients as they transition from hospital to home and assume responsibility for their own care. Patients often leave the hospital with new medications, recommended lifestyle changes, or with ongoing care needs, but many may not fully understand their discharge instructions [4]. Limited awareness of their symptoms, difficulty accessing follow-up care, and a lack of support at home can further increase their risk of complications [5]. Without proper guidance and resources, patients are more likely to experience preventable adverse events, leading to ED visits or hospital readmissions [6]. Proactive care coordination during this transition period has the potential to reduce hospital utilization, including ED [7] and hospital readmission rates [8,9]. One promising intervention is postdischarge telephone calls.

Telephone Calls as Digital Health Interventions

While digital health is often associated with apps, wearables, and video consultations, it is important to recognize that telephone-based care is a legitimate and impactful digital health tool, particularly in real-world contexts where access and equity are paramount. Digital health can be broadly defined as the delivery of care at a distance, the digital storage and sharing of health information, and the use of health-related data to improve services and systems. This broad definition allows inclusion of "low-tech" modalities, such as telephone or SMS text messaging–based interventions, alongside more "high-tech" solutions like app-based programs or algorithm-driven self-management tools [10].

Telephones are widely accessible and require no broadband internet, which makes them particularly valuable for reaching populations who may be marginalized by more complex digital platforms (eg, older adults, those with lower incomes, and those with limited English proficiency) [11]. Unlike video visits and other forms of digital care which demand stable internet and technical know-how [12], telephone calls lower the barrier to entry, enabling broader participation in virtual care [13].

From an equity standpoint, telephone-based care helps mitigate the digital divide [14,15]. Research has shown that disparities in mobile phone ownership are significantly less pronounced than those in internet access or smartphone usage [16]. This positions telephone calls as a uniquely inclusive tool in digital health strategies.

Postdischarge Calls

Postdischarge calls aim to bridge information gaps, enhance adherence to care plans, and offer emotional support to patients during their transition from hospital to home [17]. Although postdischarge calls have been implemented across various patient populations [18], the evidence supporting their effectiveness remains mixed. Studies have found that postdischarge calls are feasible [19], can improve patients' experience [20], and may improve follow-up with primary care providers [21]. However, findings on their effectiveness in reducing ED visits and hospital readmissions are inconclusive, partly due to methodological limitations such as small sample size or variability in intervention protocols [22]. Despite these mixed results, the potential for postdischarge calls as part of a comprehensive care intervention continues to be explored, with some studies suggesting their effectiveness in reducing readmissions [23-25].

This study aimed to evaluate the effectiveness of nurse-led postdischarge telephone calls provided by Fraser Health Virtual Care (FHVC) for high-risk patients. Specifically, we examined the impact of a telephone call made 48 hours after hospital discharge on ED visits and hospital readmission within 7 and 30 days following the call. By assessing this intervention, we aim to provide further insight into the role of structured postdischarge follow-up calls in reducing hospital service use and improving patient outcomes.

FHVC Service

Fraser Health is 1 of 5 health authorities in British Columbia, Canada, and provides care to more than 2 million people across 20 diverse communities through 12 acute care hospitals, outpatient and surgery centers, and various community care sites [26].

In response to the COVID-19 pandemic and the growing need for digitally enabled care, Fraser Health launched FHVC in April 2020, a nurse-led virtual health service that delivers clinical assessment and navigation support through telephone, video, and secure web chat. FHVC services include general health advice and information, personalized clinical assessments, and referrals to appropriate Fraser Health programs and services [26].

FHVC is integrated within the health authority's digital infrastructure: clinicians document each encounter directly into the electronic health record (EHR), access real-time clinical data, and coordinate follow-up or referrals electronically. The service also supports medical interpreter access, ensuring linguistic inclusivity and equitable reach across the region.

In March 2021, FHVC expanded to include proactive postdischarge outreach calls to patients within 48 hours of hospital discharge. This initiative was introduced as part of a regional digital quality improvement strategy to strengthen care transitions and reduce avoidable ED visits and readmissions. Calls are initiated automatically from daily EHR-generated discharge lists, and nurses follow a standardized digital protocol addressing symptoms, medications, follow-up appointments, and home supports.

XSL•FO
RenderX

This postdischarge component represents a digitally enabled care coordination model that leverages FHVC's integrated data systems, digital workflows, and virtual communication tools. Unlike usual care, which relies on patients or community providers to initiate follow-up, FHVC's model provides proactive, system-triggered outreach, a scalable approach to virtual care that blends clinical expertise with health-system data infrastructure.

Methods

Study Design

This study was a multicenter, pragmatic quasi-randomized trial conducted across all 12 acute care hospitals in the Fraser Health region (British Columbia, Canada) from May 16, 2022, to September 21, 2022. The trial evaluated the effect of a routine, operationally delivered, postdischarge follow-up phone call on ED visits and hospital readmissions among high-risk patients. The design was pragmatic in nature, embedded within routine care delivery and implemented by the existing FHVC nursing staff, to reflect real-world practice conditions. The trial was retrospectively registered with the ISRCTN registry (registration submission date: September 3, 2025).

The primary objective was to determine the impact of a 1-time follow-up phone call 48 hours after hospital discharge on 30-day ED visits for participants at high risk of readmission. Secondary objectives were to assess 7-day ED visits, 7- and 30-day readmissions, and participant experiences with the service. This trial is reported in accordance with the CONSORT guidelines (Checklist 1).

Participants and Setting

Participants were adults (≥18 y) discharged from any of Fraser Health's 12 acute care hospitals. The inclusion and exclusion criteria reflected routine FHVC operational parameters, designed to identify typical patients most likely to benefit from postdischarge follow-up.

Eligibility was informed by the LACE index [9,27,28], a validated measure of 30-day readmission risk incorporating Length of stay (LOS), Acuity of admission, Comorbidities, and Emergency department use. Participants were eligible if they had a LACE score ≥10, or a score <9, and were aged 45 years or older. These thresholds were derived from internal data showing an elevated risk of unplanned readmission.

Exclusion criteria included patients discharged from psychiatric units, to hospice or community nursing programs (where follow-up is routine), those who received additional postdischarge support, or were readmitted or deceased prior to follow-up.

Intervention

The intervention was a one-time, proactive follow-up call from an FHVC registered nurse approximately 48 hours after discharge. The intervention required no additional infrastructure, using existing FHVC systems, staff, and workflows. Nurses used automated discharge lists generated daily from the EHR to identify eligible participants.

During the calls, nurses followed a semistructured clinical assessment tool covering as follows:

- Understanding of discharge diagnosis and instructions,
- Medication adherence and reconciliation,
- Symptoms or concerns postdischarge, and
- Follow-up appointments and home supports

Calls were tailored to each patient's discharge plan and medical record. If issues were identified, nurses provided clarification, reinforced education, or facilitated referrals to community or primary care services. Interpreter services were available when required. Nurses used an assessment form to track this information as shown in Appendix S2 in Multimedia Appendix 1.

The intervention reflected usual care processes augmented by proactive outreach and was intentionally flexible to accommodate real-world nursing workflow and patient needs.

Comparator (Usual Care)

Participants in the control group received standard care only, which typically includes written discharge instructions and follow-up arranged by hospital units or primary care providers at their discretion. No proactive FHVC calls were made to these participants.

Recruitment and Allocation

Eligible participants were identified through the use of a custom iTracker report that automatically retrieved electronic medical record (EMR) discharge data across Fraser Health hospitals. Each morning, a research assistant applied the eligibility criteria and randomly selected up to 70 participants (using Microsoft Excel's RAND function) for the intervention group, based on daily nurse capacity. The remaining eligible participants were assigned to the control group.

Because the assignment depended on operational capacity, the allocation was quasi-random and not concealed. On days when the total number of eligible discharges was fewer than 70, all available participants were assigned to the intervention group to maintain operational continuity. Conversely, on days with limited nurse availability, more participants were assigned to the control group. This pragmatic approach preserved ecological validity by embedding group assignment within real-world service delivery constraints.

Registered nurses were blinded to the control group list and conducted intervention calls as part of their routine FHVC workload. Participants were unaware of the research component, ensuring naturalistic behavior.

Outcomes

Our primary outcome measure was the number of ED visits within 30 days of receiving the intervention, defined as any visit in which a participant registered in any of the 12 EDs within the Fraser Health region, regardless of whether they were ultimately seen by a clinician.

Secondary outcome measures included the following:

1. ED visits within 7 days of the intervention;

XSL•FO
RenderX

2. Hospital readmissions, at 7 and 30 days postintervention, defined as any admission to a hospital postdischarge; and
3. Participant experience, measured via a survey.

Sample Size Considerations

The sample size calculation was based on a retrospective analysis of EMRs from June 1, 2020, to May 31, 2021. This analysis estimated the average ED visits within 30 days of discharge before and after FHVC's postdischarge calls were implemented (0.73 and 0.59 ED visits, respectively). To detect a statistically significant difference with 80% power and α=.05, a required sample size of 6930 participants (3465 per group) was calculated for 2 independent groups.

Procedures and Data Collection

FHVC nurses received standardized training to conduct postdischarge follow-up calls from a Clinical Nurse Educator, ensuring consistency in care delivery. Calls were conducted using standard FHVC telehealth systems, and data were recorded in a secure Excel file linked to the EHR.

During each call, FHVC nurses recorded participant information, including name, age, sex, phone number, LACE score, and discharging hospital, into a Microsoft Excel file stored on a secure network drive. Additional data included any identified knowledge gaps during the postdischarge call. A research assistant monitored data daily to ensure completeness and accuracy of data collection.

If participants were unreachable, the nurse left a voicemail message with the FHVC phone number, inviting the participant to call back if they needed support. Data on ED visits, hospital readmissions, LOS of initial hospitalization, city of residence, and the day of discharge were extracted from the EMR using unique identifiers.

All participants were invited via a telephone call to complete a patient experience survey 30 days after the intervention until 80 participants were reached for each of the intervention and control groups. Surveys were administered by a research assistant and a patient partner. The survey instruments differed by study arm to reflect the experience received (see Appendix S3 in Multimedia Appendix 1 for survey questions). Participants who received the postdischarge call from FHVC answered questions about the call itself, including its helpfulness in clarifying discharge instructions, medications, follow-up appointments, and confidence in managing care at home. Participants who did not receive a postdischarge call were asked parallel questions about their discharge experience, including understanding of instructions, medications, follow-up appointments, and confidence in self-management.

Statistical Analysis

The analysis was completed by a Fraser Health statistician using SPSS (Statistical Package for the Social Sciences) version 21.0 (IBM Corp). The primary analyses followed an intention-to-treat approach. Differences in demographic and outcome variables between groups were analyzed using chi-square tests for categorical variables and the Welch 2-tailed *t* tests for continuous variables. Mann-Whitney *U* tests were used where data were not normally distributed. Average levels of risk factors were compared between groups using relative risks, 95% CIs, and 2-tailed *P* values. To reduce the risk of type I error due to multiple comparisons, a Bonferroni correction was applied to the *P* values from unadjusted univariate tests conducted on demographic and clinical risk variables. This included both between-group comparisons (intervention vs control) and exploratory within-group comparisons (reached vs not reached in the intervention arm). Adjusted *P* values were reported where appropriate.

Negative binomial regression models were used to examine the relationship between the dependent variables (ie, the number of ED visits and count of readmissions at 7 and 30 d post call) and relevant predictors. This modeling approach was selected to account for the overdispersion in the data due to the high occurrence of zero values in the count data for ED visits and hospital admissions. Independent variables included intervention group, age, sex, LACE score, LOS of initial hospitalization, day of discharge of initial hospitalization, and geographical location of ED of initial hospitalization. Regression models were not corrected for multiple comparisons, as they represented prespecified primary analyses. Unlike univariate tests, regression coefficients are estimated jointly, inherently accounting for covariation among predictors. Therefore, corrections for multiple comparisons were not applied.

Incident rate ratios (IRRs) were reported to describe the association between the independent variables and the relative frequency of count of ED visits and count of readmissions at 7 and 30 days. A sensitivity analysis was conducted in which an outlier participant with 23 ED visits was recoded to the next-highest value observed in the dataset (13 visits). The results remained robust following this adjustment. Due to missing values in the initial LOS variable, 32 observations (9 in the control group and 22 in the intervention group) were excluded from the final regression models.

In addition to the primary and secondary outcomes, the number needed to treat (NNT) was calculated to further assess the potential clinical significance of the intervention [29]. NNT is a measure used to determine how many participants need to be treated with the intervention to prevent one additional adverse outcome, such as an ED visit or hospital readmission. The NNT was calculated for the primary outcome of ED visits and the secondary outcome of hospital readmissions at both 7 and 30 days after the follow-up call.

The formula used for calculating NNT is:

$$NNT = 1/|CER - EER|$$

where CER is the control event rate (ie, the proportion of participants in the control group who experience the adverse outcomes), and EER is the experimental event rate (ie, the proportion of participants in the intervention group who experienced the adverse outcome). The calculated NNT values provide a clinical perspective on the effectiveness of the intervention, helping to assess how many participants need to receive the postdischarge follow-up call to prevent 1 additional adverse outcome.

Descriptive statistics for the participant experience survey were calculated using Microsoft Excel. No inferential tests were

XSL•FO
RenderX

conducted because the survey questions differed between groups. The intervention group was asked whether the postdischarge call improved their understanding, and the control group was asked about their understanding of the initial discharge instructions; therefore, direct between-group comparisons were not possible.

Ethical Considerations

This study was reviewed and approved by the Fraser Health Research Ethics Board (number 2022192). Given that the study evaluated a quality improvement initiative embedded within routine operations and used deidentified secondary administrative data, the requirement for individual informed consent was waived by the ethics committees in accordance with institutional policy. All data were deidentified before analysis. No patient-identifiable information was shared outside the health authority. Data were stored on secure, password-protected servers in compliance with provincial privacy legislation. No financial or material compensation was provided to participants, as all activities represented standard care or evaluation of routine service delivery.

Results

Study Population

A total of 7091 participants were included in the study, with 3911 in the intervention group and 3180 in the control group (Figure 1). The overall sample had a mean age of 68.1 (SD 13.74) years (95% CI 67.8 - 68.4), and 47.2% (n=3349) were female.

Among the participants in the intervention group, 1752 fully completed the postdischarge call, while nurses were unable to reach the remainder directly. Of those who completed the call, 40% (n=701) had at least 1 care gap identified.

Figure 1. Flow diagram of participant inclusion and study design. This flowchart outlines participant selection for the retrospective cohort study evaluating the impact of nurse-led postdischarge telephone calls provided by Fraser Health Virtual Care (FHVC) on emergency department (ED) visits and hospital readmissions. The study included adult patients (≥18 y) discharged from acute care hospitals within Fraser Health, British Columbia, Canada, between January 2022 and December 2023.

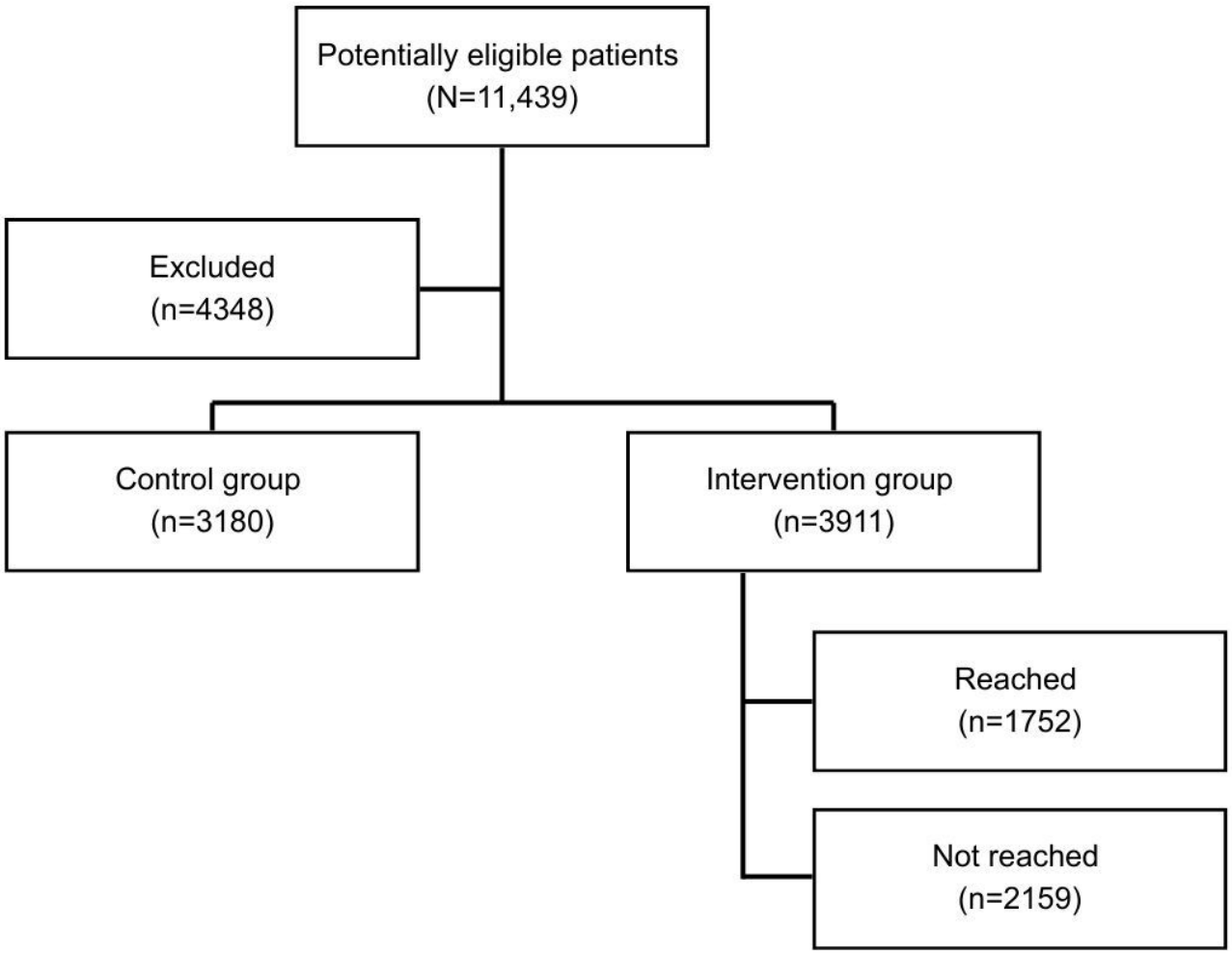

Baseline Characteristics

Baseline characteristics for the intervention and control groups are presented in Table 1. After applying a Bonferroni correction for multiple comparisons (adjusted α=.0042 and adjusted P<.012), the only statistically significant difference was a higher proportion of weekend discharges in the intervention group (33.8% vs 21.5%; adjusted P<.012). No meaningful differences were observed in age, sex, LACE score, LOS, or hospital location. Within the intervention arm, baseline characteristics were similar between participants who completed the call and those who did not (Appendix S4 in Multimedia Appendix 1).

XSL•FO
RenderX

Table . Baseline characteristics of study sample. Characteristics of patients included in the retrospective cohort study assessing Fraser Health Virtual Care (FHVC)'s postdischarge telephone calls.[a]

	Intervention (n=3911)	Control (n=3180)	Difference statistics	Unadjusted *P* value	Bonferroni adjusted *P* value
Age, mean (SD)	67.7 (13.5)	68.6 (14.1)	6001487.5[b]	.01	.132
Sex, n (%)			χ^2=7.4[c]	.006	.072
Female	1790 (45.8)	1559 (49)			
Male	2121 (54.2)	1621 (51)			
LACE[d] score, mean (SD)	8.56 (4.42)	8.29 (4.53)	5943871[b]	.001	.012
LOS[e] days of initial hospitalization, mean (SD)	5.49 (6.94)	6.57 (9.70)	6001849[b]	.05	.6
Discharge timing, n (%)			χ^2=131.2[c]	<.001	<.012
Discharged on weekend	1323 (33.8)	684 (21.5)			
Discharged on weekday	2588 (66.2)	2496 (78.5)			
Hospital location, n (%)			χ^2=3.4[c]	.06	.78
Rural or urban hospital	2278 (58.2)	1783 (56.1)			
Metro hospital	1633 (41.8)	1397 (43.9)			

[a]Data include demographics, clinical risk factors (eg, LACE score and LOS), discharge timing, and hospital location. Comparisons are presented between the intervention group (received nurse-led call within 48 h postdischarge) and matched control group (no call), with difference statistics and *P* values.

[b]Mann-Whitney *U* test.

[c]Chi-square test. All degrees of freedom were 1.

[d]LACE: Length of stay, Acuity of admission, Comorbidities, and Emergency department use

[e]LOS: length of stay.

Primary Outcome: ED Visits

Participants in the intervention group had significantly fewer ED visits at both 7-day and 30-day postdischarge call compared to those in the control group. Within 7 days of the discharge call, 9.33% (365/3911) of intervention participants visited an ED, compared to 11.86% (377/3180) of control participants (Table 2).

The NNT to prevent 1 ED visit within 7 days post call was 40 (95% CI 2774). For ED visits within 30 days, NNT=86; however, the 95% CI crossed zero, indicating nonsignificance.

Table . Trial outcomes with unadjusted 7-day and 30-day emergency department (ED) visits and hospital readmissions following discharge, and counts and proportions of patients with ED use or hospital readmission after discharge among those who received a postdischarge telephone call (intervention) and those who did not (control) in Fraser Health, British Columbia, Canada (2022–2023).

Postdischarge call	Intervention (n=3911)	Control (n=3180)
Within 7 days of call		
ED visits, n	427	472
ED use, n (%)	365 (9.33)	377 (11.9)
Hospital admissions, n	109	115
Hospital use, n (%)	106 (2.7)	109 (3.4)
Within 30 days of call		
ED visits, n	1072	973
ED use, n	759 (19.4)	654 (20.6)
Hospital admissions, n	292	253
Hospital use, n (%)	257 (6.6)	214 (6.7)

XSL•FO
RenderX

Adjusted analyses confirmed that the intervention was associated with a significant reduction in 7-day ED visits (IRR 0.72, 95% CI 0.62 - 0.84; *P*<.001) and a modest reduction at 30 days (IRR 0.88, 95% CI 0.78 - 0.98; *P*=.02). Higher LACE scores were consistently associated with increased ED use. No significant associations were observed for age, sex, LOS, day of discharge, or hospital location (Tables 3 and 4).

Table . Negative binomial regression results for 7-day emergency department (ED) visits following discharge[a].

	IRR[b] (95% CI)	*P* value
Intervention	0.719 (0.617-0.837)	<.001
Age (y)	0.996 (0.990-1.001)	.11
Female (vs male)	0.886 (0.762-1.030)	.12
LACE[c] score	1.051 (1.034-1.068)	<.001
LOS[d] at initial hospitalization	0.992 (0.982-1.002)	.12
Weekday discharge (vs weekend)	0.982 (0.827-1.164)	.84
Metro hospital (vs rural or urban)	0.897 (0.769-1.045)	.16

[a]Adjusted IRR and 95% CIs for factors associated with ED visits within 7 days of hospital discharge among adult patients in Fraser Health, British Columbia (2022–2023). The model includes intervention exposure and demographic and clinical covariates.

[b]IRR: incident rate ratio.

[c]LACE: Length of stay, Acuity of admission, Comorbidities, and Emergency department use.

[d]LOS: length of stay.

Table . Negative binomial regression results for 30-day emergency department (ED) visits following discharge[a].

	IRR[b] (95% CI)	*P* value
Intervention	0.878 (0.783 - 0.983)	.02
Age	0.991 (0.987 - 0.995)	<.001
Female (vs male)	0.941 (0.841 - 1.053)	.29
LACE[c] score	1.082 (1.069 - 1.095)	<.001
LOS[d] at initial hospitalization	1.000 (0.993 - 1.006)	.96
Weekday discharge (vs weekend)	0.957 (0.843 - 1.086)	.50
Metro hospital (vs rural or urban)	0.897 (0.801 - 1.006)	.06

[a]Adjusted IRR for ED visits within 30 days of hospital discharge among adult patients in Fraser Health, British Columbia (2022–2023), comparing those who received a postdischarge Fraser Health Virtual Care (FHVC) nurse call versus matched controls.

[b]IRR: incident rate ratio.

[c]LACE: Length of stay, Acuity of admission, Comorbidities, and Emergency department use.

[d]LOS: length of stay.

Secondary Outcome: Hospital Readmissions

There was no statistically significant reduction in hospital readmissions at 7 or 30 days. Within 7 days, 2.7% of intervention participants were readmitted versus 3.4% of controls (IRR 0.81, 95% CI 0.62 - 1.06; *P*=.13). At 30 days, readmissions were 6.6% versus 6.7% (IRR 0.94, 95% CI 0.78 - 1.14; *P*=.54). LACE score and LOS were significant predictors of readmission, while other covariates were not. NNT estimates to prevent one hospital readmission were 139 (7 days) and 625 (30 days), with 95% CIs crossing zero (Table 2; Appendix S5 in Multimedia Appendix 1).

Patient Experience Survey Results

Between June 14, 2022, and August 29, 2022, a total of 457 participants (control group: n=223; intervention group: n=264). The study design specified a target sample size of 80 participants per group, and recruitment was concluded once this threshold was reached, resulting in 160 participants (n=80 per group), yielding response rates of 36% in the control group and 30% in the intervention group. Among control group respondents, 67 (83.8%) participants were between the ages of 50 and 70 years, 43 (53.8%) were male, and 51 (63.8%) identified as White. Among intervention group respondents, 56 (70%) participants were between the ages of 50 and 70 years, 45 (56.3%) were male, and 57 (71.3%) identified as White.

Participants in the intervention group reported that the postdischarge call improved their understanding of hospital discharge instructions (n=46, 57.5% agreed) and 40% (n=32) stated that the call provided additional information they did not receive at the time of discharge. Most participants found the

XSL•FO
RenderX

postdischarge call helpful in answering their questions (n=33; 41.3% strongly agreed) and 46.3% (n=36) felt more confident managing their own health care after discharge. Nearly all participants viewed the postdischarge call as a valuable service (n=40, 50.6% strongly agreed; n=36, 45.6% agreed). Overall, 71.6% of (n=53) participants in the intervention group visited a primary care provider for follow-up care after discharge.

Among participants in the control group, 47.5% (n=38) strongly agreed that they clearly understood the hospital discharge instructions, and 46.3% (n=37) agreed that they were satisfied with the information provided. Additionally, 51.3% (n=41) felt confident managing their health at home. In total, 60.8% (n=45) of participants in the control group visited a primary care provider for follow-up care after discharge. See Appendix S3 in Multimedia Appendix 1 for full results from patient experience surveys.

Discussion

Overview

This study evaluated a 1-time, nurse-led postdischarge follow-up call delivered through an integrated virtual care program. The intervention was associated with modest improvements in early postdischarge outcomes, with participants in the intervention group experiencing significantly fewer ED visits at both 7 and 30 days. However, hospital readmission rates were not significantly affected.

The observed 28% relative reduction in 7-day ED visits and 12% reduction in 30-day ED visits highlight the clinical relevance of the intervention, supported by the NNT estimates which showed that 39 participants needed to receive the intervention to prevent one 7-day ED visit, and 83 patients to prevent one 30-day ED visit. These are clinically meaningful numbers, especially in the context of high hospital discharge volumes, where even modest improvements can translate into substantial system-wide benefits. While hospital readmission reductions were numerically favorable, they did not reach statistical significance, potentially due to lower event rates and limited power. Alternatively, while a postdischarge phone call can clarify instructions, reduce anxiety, and address basic care gaps, a single call may be insufficient to address the complex medical and social factors that contribute to hospital readmissions [30,31] and patients with elevated LACE scores may require more comprehensive or repeated interventions to meaningfully reduce readmissions.

Patient Experience

Participants receiving postdischarge phone calls reported overall satisfaction with the intervention and found that the calls provided them with additional information, reduced their concerns, and gave them more confidence to manage their own health care at home (Appendix S3 in Multimedia Appendix 1). Interestingly, participants in the control group also reported satisfaction with the discharge information they received, suggesting that unrecognized gaps may remain without structured follow-up. During the postdischarge calls, FHVC nurses asked participants to explain their discharge plan, which often revealed missing or misunderstood information. Nurses followed a standard script which covered essential postdischarge topics. Through this process, 40% of participants in the intervention group were identified as having at least 1 care gap. It is plausible that some participants in the control group had similar gaps, but without prompting, they were not aware of them at the time of the survey, and therefore felt they could manage adequately post discharge. Additionally, 41% of control participants expressed that they would have appreciated FHVC contact information sooner, highlighting the potential value of proactive outreach.

Potential Mechanisms of Action

Follow-up calls likely operate through multiple mechanisms: reinforcing patient education, supporting medication adherence, addressing emotional concerns, and signaling coordinated care across the health system. The intervention may serve both as a direct support tool and as an organizational signal promoting better discharge and follow-up practices. While single-touch calls improved ED utilization, they appear insufficient to address complex factors driving hospital readmissions, underscoring the need for more intensive or repeated interventions for high-risk populations.

Several studies have demonstrated that patient education [32] and medication adherence [33,34] are critical factors in reducing avoidable health care utilization, which may explain the observed reduction in ED visits in our study. Prior research has demonstrated that inadequate understanding of discharge instructions is a major driver of postdischarge complications and unplanned care use [35,36]. By reinforcing these instructions, postdischarge follow-up calls may have improved patient comprehension of their recovery plan, reduced knowledge gaps, and empowered patients to manage symptoms at home, thereby lowering the need for ED visits. In addition, the calls may have alleviated anxiety and promoted confidence in self-management [37] by providing reassurance and tailored guidance, thereby reducing the likelihood of seeking avoidable emergency care.

Medication adherence is another well-documented challenge following hospital discharge, with nonadherence associated with increased ED visits and readmissions [38]. The follow-up calls offered a valuable opportunity to clarify prescriptions, address concerns, and reinforce adherence strategies. This may have helped prevent adverse events related to medication errors. Similar findings have been observed in previous studies where medication reconciliation and adherence support were associated with reductions in acute care use [32]. These results suggest that postdischarge calls extend beyond simple follow-up; they actively address gaps in understanding, medication adherence, and emotional well-being, contributing to their observed effectiveness in reducing ED visits.

Importantly, within the intervention arm, completing the call did not meaningfully change ED or readmission outcomes compared with participants who were not reached, suggesting that the benefits observed may reflect system-level effects of the outreach model rather than call completion alone. For example, the intervention may have functioned as a signal of coordinated care, reinforcing follow-up workflows, prompting timely triage, or improving discharge planning processes across

XSL•FO
RenderX

the health system. Alternatively, exposure misclassification, where a participant assigned to the intervention group did not actually receive a call, may have diluted the measured effect of individual calls, meaning the observed reductions in ED visits reflect the broader impact of the intervention's integration into routine care rather than direct participant-level engagement alone. Recognizing these mechanisms underscores that single-touch outreach can influence outcomes through organizational processes as well as patient-level interactions.

Clinical Implications

Postdischarge calls have been shown to enhance patient engagement with primary care providers, improve medication adherence, and contribute to better overall health outcomes [21]. Given the increasing difficulty in accessing primary care [39,40], these calls may serve as a low-barrier intervention to support care transitions and fill gaps in follow-up care.

Discharge calls can facilitate smoother transitions from hospital to home by proactively identifying and addressing common barriers to recovery. By intervening early, the discharge calls can reduce the risk of preventable complications and improve the overall patient experiences [38] while also relieving pressures on overstretched health care systems [41].

Reducing ED visits has important implications for health care systems, as ED overcrowding remains a critical issue [42]. A reduction in ED visits suggests that patients may be managing their conditions more effectively in outpatient settings, potentially leading to cost savings and improved resource allocation. Future research should explore whether postdischarge calls lead to increased engagement with primary care providers, potentially shifting health care utilization patterns.

Strengths and Limitations

Previous reviews, including 1 Cochrane review [18] and 2 systematic reviews [21,25] have noted that many trials lacked sufficient statistical power, clearly defined sample size calculations, rigorous methodology, or standardized interventions. In contrast, our study addressed these limitations by using a larger sample size, informed by Fraser Health historical data. Additionally, we implemented a structured intervention that included standardized nurse training and call templates, ensuring consistency in care delivery across settings.

A major strength of this trial was its integration into an existing digital health care service. FHVC was already operational before the study, meaning that nurses conducting the calls had prior experience and received training from a clinical nurse educator. This level of standardization may have contributed to the observed effectiveness of the intervention, differentiating our study from others that implemented newly introduced follow-up call programs. Conducted within an existing operational health care service, the study reflected real-world variations in patient care, staffing constraints, and service delivery. As a result, our findings represent an intervention that health care organizations could realistically implement to support patients' transitions from hospital to home.

The broad eligibility criteria further enhance the generalizability of our findings. Although the study focused on high-risk patients, it included patients with a range of health conditions, demographics, and socioeconomic backgrounds. The standardized discharge call protocol was designed to be adaptable across clinical contexts, and a detailed version is provided for reference (see Appendix S2 in Multimedia Appendix 1).

A key limitation is the potential exclusion of patients without reliable phone access. Certain socioeconomic groups may have faced barriers to participating, including a lack of access to phones or data plans. Although FHVC implemented strategies to enhance accessibility, such as using plain language and integrating interpreters to support patients who do not speak English, these barriers remain a challenge for equitable access to virtual care.

Another limitation of this study was the low intervention exposure rate within the intervention group. While the analysis followed an intention-to-treat approach, a substantial proportion of participants did not answer the postdischarge call and therefore were not actually exposed to the intervention. This may have diluted the observed effect and may have underestimated the true impact of the intervention had more participants engaged. Future efforts should focus on improving patient engagement, such as further increasing awareness of the follow-up service at the point of discharge, to enhance reach and effectiveness.

Despite being guided by a sample size calculation, the study faced logistical challenges due to fluctuating discharge volumes and staffing availability. While the goal was to collect data for 58 days, recruitment extended to 84 days to achieve the desired sample size. Nurse capacity constraints also required adjustments to the intervention group sizes, leading to a control group that was smaller than the initially planned sample size (3180 instead of 3465) and an intervention group that was larger (3911 vs 3465). These adjustments reflect the realities of implementing research within an operational care model and underscore the need to balance methodological rigor with service delivery. These deviations from the original allocation plan may impact statistical power and group comparability and should be considered when interpreting results.

Future Directions

Overview

While this study provides strong evidence for the effectiveness of postdischarge follow-up calls, several areas require further exploration to optimize implementation, improve equity, and maximize impact. Future research should focus on leveraging emerging technologies, understanding the mechanisms driving intervention success, addressing disparities in access to virtual care, and evaluating cost-effectiveness. Additionally, qualitative and subgroup analyses may help further refine the intervention's design and delivery to better meet the needs of diverse patient populations.

Leveraging Emerging Technologies for Enhanced Support

Artificial intelligence (AI) has the potential to improve the scalability and efficiency of postdischarge interventions by

XSL•FO
RenderX

augmenting human-led care with data-driven insights. AI-driven decision support systems could help identify patients most likely to benefit from follow-up calls, thereby optimizing resource allocation and ensuring that high-risk individuals receive timely support. By using predictive analytics to prioritize outreach, health care systems could enhance the effectiveness of follow-up programs while also managing staff workload more efficiently.

Additionally, natural language processing technologies could be integrated into call workflows to assist nurses in real-time. These tools could provide structured prompts to ensure comprehensive patient assessments, summarize patient concerns for documentation, or generate automated follow-up messages tailored to each patient's needs. For example, natural language processing could automatically flag patients with high LACE scores who have missed follow-ups, prompting timely outreach and helping prioritize those at greater risk of admission. Future studies should explore how AI can support, rather than replace, clinical decision-making—maintaining the essential human connection that characterizes high-quality transitional care. A hybrid model where AI improves workflow efficiency while nurses remain the primary point of engagement may offer an optimal balance between scalability and personalized care.

Understanding Mechanisms of Action and Patient-Centered Benefits

While this study demonstrated a reduction in ED visits, future research should explore the mechanisms underlying this effect. It remains unclear which components of the intervention, such as patient education, medication adherence support, emotional reassurance, or care coordination, or the simple act of receiving a call, contribute most to the observed improvements. Identifying the most influential elements would support refinement of the intervention to maximize its impact and cost-effectiveness. Involving patients in this refinement process through co-design of the call content or format can further enhance the relevance and effectiveness of the intervention, aligning it more closely with patient needs and preferences.

Beyond clinical outcomes, patient-reported experiences offer valuable insights into the intervention's effectiveness. Future qualitative research should examine how follow-up calls influence patient confidence in managing their recovery, their perceived quality of care, and any remaining unmet needs following discharge. Additionally, understanding the perspectives of nurses and other health care providers delivering follow-up calls can help identify workflow challenges and opportunities for improvement. Research in this area could inform best practices for scaling and sustaining postdischarge calls while also minimizing provider burden and enhancing patient engagement.

Promoting Equity in Postdischarge Support

Ensuring equitable access to follow-up care remains a critical challenge. Patients from lower socioeconomic backgrounds, those with limited English proficiency, and individuals without reliable phone access are at risk of being disproportionately excluded from telephone-based interventions. As a result, those who are most vulnerable to hospital readmission may not receive the support they need. To effectively reach and support diverse populations, follow-up care must also be culturally appropriate and safe. Embedding cultural safety into the design and delivery of interventions can help build trust and improve engagement across communities.

Future research should explore alternative and complementary strategies to reach underserved populations. These may include text-based follow-ups, video consultations, or partnerships with community health workers who can provide in-person support. Additionally, health systems should investigate more seamless integration of interpreter services into virtual care workflows and evaluate the potential of multilingual AI-driven chatbots to help bridge communications barriers. Addressing these disparities is essential to ensure that all patients, regardless of socioeconomic status, digital access, or language, can benefit equitably from postdischarge support.

Economic Analysis of Postdischarge Follow-Up Calls

A critical next step is to conduct a cost-effectiveness analysis to determine whether the observed reduction in ED visits results in net cost savings for the health care system. By evaluating factors such as reduced hospital readmissions, decreased ED use, and the cost of nursing time invested in delivering the intervention, an economic evaluation can guide policy decisions and resource allocation for broader implementation.

In addition to understanding overall cost-effectiveness, future studies should assess differential outcomes across patient subgroups. Certain populations may derive more substantial benefit from follow-up calls, informing a more targeted and efficient use of resources. Conducting subgroup analyses based on factors such as age, comorbidities, hospital discharge diagnosis, and social determinants of health could help refine intervention targeting. For example, patients with chronic conditions such as heart failure or chronic obstructive pulmonary disease may benefit more from medication adherence support, while those with lower health literacy may require additional educational reinforcement. Identifying which subgroups experience the greatest reduction in ED visits or readmissions will enable a more tailored approach, ensuring that resources are directed toward those who need them most.

Conclusions

This study suggests that proactive, digitally enabled outreach may support safer care transitions and reduce early ED utilization, even when delivered as a pragmatic, single-touch intervention within routine clinical operations. The findings highlight the potential for this low-cost, scalable intervention to support safer care transitions, particularly in healthcare systems where timely access to primary care remains a challenge. While the intervention did not significantly reduce hospital readmissions, its impact on ED use underscores its value in improving short-term outcomes and alleviating pressure on acute care services. Future studies should investigate whether certain patient subgroups derive greater benefit and explore ways to improve accessibility for populations with limited phone or digital access.

XSL•FO
RenderX

Acknowledgments

The authors thank the patients and the Fraser Health Virtual Care nurses that participated in this study.

Funding

This work was supported by the Fraser Health Strategic Priorities Grant and the Health Research Foundation of Innovative Medicines Canada Team Grant in Virtual Care. The funders had no role in the study design, data collection, analysis, interpretation, writing of the manuscript, or the decision to submit the paper for publication.

Data Availability

Due to conditions of the informed consent obtained from participants, the institutional and Ministry of Health ethical requirements do not permit us to share participant data from this study.

Authors' Contributions

SD contributed to conceptualization, design, methodology, data curation, writing of the manuscript, and project administration. AA contributed to data curation, project administration, investigation, data acquisition, data analysis, and writing of the manuscript. M Montenegro contributed to methodology, data curation, data analysis, software, and writing of the manuscript. M Maitland contributed to conceptualization, design, methodology, and funding acquisition. SC contributed to software and data analysis. SJ contributed to design, methodology, and project administration. HS contributed to design and methodology. DC contributed to investigation and review and editing of the manuscript. JMR contributed to writing. M MacPherson contributed to writing of the manuscript. DB contributed to the review and editing of the manuscript. MN-vV contributed to conceptualization, design, methodology, and supervision. All authors read and approved the final manuscript.

Conflicts of Interest

Two authors (MN-vV and SJ) are employed by Fraser Health Virtual Care, the program responsible for implementing the intervention examined in this study, and therefore may be perceived as having an interest in positive outcomes. All other authors declare no financial or nonfinancial competing interests.

Multimedia Appendix 1
Appendix S1: PRECIS-2 assessment of the post-discharge follow-up call trial. Appendix S2: Discharge Call Protocol Checklist. Appendix S3: Patient Experience Survey. Appendix S4: Intervention Group Characteristics (those Reached vs Not Reached). Appendix S5: Negative Binomial regression results for Hospital Readmission,
[DOCX File, 85 KB - jmir_v28i1e80529_app1.docx]

Checklist 1
CONSORT checklist.
[PDF File, 177 KB - jmir_v28i1e80529_app2.pdf]

References

1. Howard-Anderson J, Lonowski S, Vangala S, Tseng CH, Busuttil A, Afsar-Manesh N. Readmissions in the era of patient engagement. JAMA Intern Med 2014 Nov;174(11):1870-1872. [doi: 10.1001/jamainternmed.2014.4782] [Medline: 25264930]
2. Staples JA, Thiruchelvam D, Redelmeier DA. Site of hospital readmission and mortality: a population-based retrospective cohort study. CMAJ Open 2014 Apr;2(2):E77-E85. [doi: 10.9778/cmajo.20130053] [Medline: 25077133]
3. Sinden S. Is the path to better outcomes paved in gold? Physician incentives and hospital readmissions. Centre for Advancing Health Outcomes. 2024. URL: https://www.advancinghealth.ubc.ca/research-in-action/is-the-path-to-better-outcomes-paved-in-gold-physician-incentives-and-hospital-readmissions/ [accessed 2026-02-04]
4. Horwitz LI, Moriarty JP, Chen C, et al. Quality of discharge practices and patient understanding at an academic medical center. JAMA Intern Med 2013 Oct 14;173(18):1715-1722. [doi: 10.1001/jamainternmed.2013.9318] [Medline: 23958851]
5. Feigenbaum P, Neuwirth E, Trowbridge L, et al. Factors contributing to all-cause 30-day readmissions: a structured case series across 18 hospitals. Med Care 2012 Jul;50(7):599-605. [doi: 10.1097/MLR.0b013e318249ce72] [Medline: 22354212]
6. Shahid R, Shoker M, Chu LM, Frehlick R, Ward H, Pahwa P. Impact of low health literacy on patients' health outcomes: a multicenter cohort study. BMC Health Serv Res 2022 Sep 12;22(1):1148. [doi: 10.1186/s12913-022-08527-9] [Medline: 36096793]

XSL•FO
RenderX

7. Enard KR, Ganelin DM. Reducing preventable emergency department utilization and costs by using community health workers as patient navigators. J Healthc Manag 2013;58(6):412-427. [doi: 10.1097/00115514-201311000-00007] [Medline: 24400457]
8. Bamforth RJ, Chhibba R, Ferguson TW, et al. Strategies to prevent hospital readmission and death in patients with chronic heart failure, chronic obstructive pulmonary disease, and chronic kidney disease: a systematic review and meta-analysis. PLoS One 2021;16(4):e0249542. [doi: 10.1371/journal.pone.0249542] [Medline: 33886582]
9. Miller DA, Schaper AM. Implementation of a follow-up telephone call process for patients at high risk for readmission. J Nurs Care Qual 2015;30(1):63-70. [doi: 10.1097/NCQ.0000000000000069] [Medline: 24844916]
10. Shaw J, Abejirinde IOO, Agarwal P, Shahid S, Martin D. Digital health and equitable access to care. PLOS Digit Health 2024 Sep;3(9):e0000573. [doi: 10.1371/journal.pdig.0000573] [Medline: 39321176]
11. Tong L, George B, Crotty BH, et al. Telemedicine and health disparities: association between patient characteristics and telemedicine, in-person, telephone and message-based care during the COVID-19 pandemic. IPEM Transl 2022;3:100010. [doi: 10.1016/j.ipemt.2022.100010] [Medline: 36340828]
12. Kaihlanen AM, Virtanen L, Buchert U, et al. Towards digital health equity—a qualitative study of the challenges experienced by vulnerable groups in using digital health services in the COVID-19 era. BMC Health Serv Res 2022 Feb 12;22(1):188. [doi: 10.1186/s12913-022-07584-4] [Medline: 35151302]
13. Rodriguez JA, Betancourt JR, Sequist TD, Ganguli I. Differences in the use of telephone and video telemedicine visits during the COVID-19 pandemic. Am J Manag Care 2021 Jan;27(1):21-26. [doi: 10.37765/ajmc.2021.88573] [Medline: 33471458]
14. Badr J, Motulsky A, Denis JL. Digital health technologies and inequalities: a scoping review of potential impacts and policy recommendations. Health Policy 2024 Aug;146:105122. [doi: 10.1016/j.healthpol.2024.105122] [Medline: 38986333]
15. Thomas-Jacques T, Jamieson T, Shaw J. Telephone, video, equity and access in virtual care. NPJ Digit Med 2021 Nov 18;4(1):159. [doi: 10.1038/s41746-021-00528-y] [Medline: 34795356]
16. García E, Vincent BC, Thakur S, et al. Variables influencing telemedicine utilization via telephone appointments among rural patients. BMC Health Serv Res 2025 Jan 2;25(1):1. [doi: 10.1186/s12913-024-12122-5] [Medline: 39748425]
17. Harrison PL, Hara PA, Pope JE, Young MC, Rula EY. The impact of postdischarge telephonic follow-up on hospital readmissions. Popul Health Manag 2011 Feb;14(1):27-32. [doi: 10.1089/pop.2009.0076] [Medline: 21090991]
18. Mistiaen P, Poot E. Telephone follow-up, initiated by a hospital-based health professional, for postdischarge problems in patients discharged from hospital to home. Cochrane Database Syst Rev 2006 Oct 18;2006(4):CD004510. [doi: 10.1002/14651858.CD004510.pub3] [Medline: 17054207]
19. Hodalova S, Moore S, Dowds J, Murphy N, Martin-Loeches I, Broderick J. Feasibility of telephone follow-up after critical care discharge. Med Sci (Basel) 2020 Mar 14;8(1):16. [doi: 10.3390/medsci8010016] [Medline: 32183263]
20. Dudas V, Bookwalter T, Kerr KM, Pantilat SZ. The impact of follow-up telephone calls to patients after hospitalization. Dis Mon 2002 Apr;48(4):239-248. [doi: 10.1016/s0011-5029(02)90031-3] [Medline: 12021756]
21. Crocker JB, Crocker JT, Greenwald JL. Telephone follow-up as a primary care intervention for postdischarge outcomes improvement: a systematic review. Am J Med 2012 Sep;125(9):915-921. [doi: 10.1016/j.amjmed.2012.01.035] [Medline: 22938927]
22. Woods CE, Jones R, O'Shea E, Grist E, Wiggers J, Usher K. Nurse-led postdischarge telephone follow-up calls: a mixed study systematic review. J Clin Nurs 2019 Oct;28(19-20):3386-3399. [doi: 10.1111/jocn.14951] [Medline: 31162748]
23. Leppin AL, Gionfriddo MR, Kessler M, et al. Preventing 30-day hospital readmissions: a systematic review and meta-analysis of randomized trials. JAMA Intern Med 2014 Jul;174(7):1095-1107. [doi: 10.1001/jamainternmed.2014.1608] [Medline: 24820131]
24. Finlayson K, Chang AM, Courtney MD, et al. Transitional care interventions reduce unplanned hospital readmissions in high-risk older adults. BMC Health Serv Res 2018 Dec 12;18(1):956. [doi: 10.1186/s12913-018-3771-9] [Medline: 30541530]
25. Jayakody A, Bryant J, Carey M, Hobden B, Dodd N, Sanson-Fisher R. Effectiveness of interventions utilising telephone follow up in reducing hospital readmission within 30 days for individuals with chronic disease: a systematic review. BMC Health Serv Res 2016 Aug 18;16(1):403. [doi: 10.1186/s12913-016-1650-9] [Medline: 27538884]
26. Fraser Health Virtual Care. Fraser Health. URL: https://www.fraserhealth.ca/patients-and-visitors/fraser-health-virtual-care [accessed 2026-02-04]
27. van Walraven C, Dhalla IA, Bell C, et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. CMAJ 2010 Apr 6;182(6):551-557. [doi: 10.1503/cmaj.091117] [Medline: 20194559]
28. Gruneir A, Dhalla IA, van Walraven C, et al. Unplanned readmissions after hospital discharge among patients identified as being at high risk for readmission using a validated predictive algorithm. Open Med 2011;5(2):e104-e111. [Medline: 21915234]
29. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. BMJ 1995 Feb 18;310(6977):452-454. [doi: 10.1136/bmj.310.6977.452] [Medline: 7873954]
30. Kaya S, Sain Guven G, Aydan S, Toka O. Predictors of hospital readmissions in internal medicine patients: application of Andersen's model. Int J Health Plann Manage 2019 Jan;34(1):370-383. [doi: 10.1002/hpm.2648] [Medline: 30221793]

XSL•FO
RenderX

31. Kroch E, Duan M, Martin J, Bankowitz RA. Patient factors predictive of hospital readmissions within 30 days. J Healthc Qual 2016;38(2):106-115. [doi: 10.1097/JHQ.0000000000000003] [Medline: 26042742]
32. Wilhelmsen NC, Eriksson T. Medication adherence interventions and outcomes: an overview of systematic reviews. Eur J Hosp Pharm 2019 Jul;26(4):187-192. [doi: 10.1136/ejhpharm-2018-001725] [Medline: 31338165]
33. Allison K. Adherence through education: a call to clinicians to educate all patients on medication use. Ment Health Clin 2012;2(4):83-85. [doi: 10.9740/mhc.n117761]
34. Rosen OZ, Fridman R, Rosen BT, Shane R, Pevnick JM. Medication adherence as a predictor of 30-day hospital readmissions. Patient Prefer Adherence 2017;11:801-810. [doi: 10.2147/PPA.S125672] [Medline: 28461742]
35. Sheikh H, Brezar A, Dzwonek A, Yau L, Calder LA. Patient understanding of discharge instructions in the emergency department: do different patients need different approaches? Int J Emerg Med 2018 Feb 8;11(1):5. [doi: 10.1186/s12245-018-0164-0] [Medline: 29423767]
36. Hastings SN, Barrett A, Weinberger M, et al. Older patients' understanding of emergency department discharge information and its relationship with adverse outcomes. J Patient Saf 2011 Mar;7(1):19-25. [doi: 10.1097/PTS.0b013e31820c7678] [Medline: 21921863]
37. van Loon-van Gaalen M, van Winsen B, van der Linden MC, Gussekloo J, van der Mast RC. The effect of a telephone follow-up call for older patients, discharged home from the emergency department on health-related outcomes: a systematic review of controlled studies. Int J Emerg Med 2021 Feb 18;14(1):13. [doi: 10.1186/s12245-021-00336-x] [Medline: 33602115]
38. Aremu TO, Oluwole OE, Adeyinka KO, Schommer JC. Medication adherence and compliance: recipe for improving patient outcomes. Pharmacy (Basel) 2022 Aug 28;10(5):106. [doi: 10.3390/pharmacy10050106] [Medline: 36136839]
39. Starfield B, Shi L, Macinko J. Contribution of primary care to health systems and health. Milbank Q 2005;83(3):457-502. [doi: 10.1111/j.1468-0009.2005.00409.x] [Medline: 16202000]
40. Duong D, Vogel L. National survey highlights worsening primary care access. CMAJ 2023 Apr 24;195(16):E592-E593. [doi: 10.1503/cmaj.1096049] [Medline: 37094873]
41. Sawesi S, Rashrash M, Phalakornkule K, Carpenter JS, Jones JF. The impact of information technology on patient engagement and health behavior change: a systematic review of the literature. JMIR Med Inform 2016 Jan 21;4(1):e1. [doi: 10.2196/medinform.4514] [Medline: 26795082]
42. Pines JM, Hilton JA, Weber EJ, et al. International perspectives on emergency department crowding. Acad Emerg Med 2011 Dec;18(12):1358-1370. [doi: 10.1111/j.1553-2712.2011.01235.x] [Medline: 22168200]

Abbreviations

AI: artificial intelligence
CER: control event rate
ED: emergency department
EER: experimental event rate
EHR: electronic health record
EMR: electronic medical record
FHVC: Fraser Health Virtual Care
IRR: incident rate ratio
LACE: Length of stay, Acuity of admission, Comorbidities, and Emergency department use
LOS: length of stay
NNT: number needed to treat
SPSS: Statistical Package for the Social Sciences

Edited by A Stone, N Cahill; submitted 11.Jul.2025; peer-reviewed by LS Topaz, NC Favour; revised version received 19.Dec.2025; accepted 24.Dec.2025; published 17.Mar.2026.

Please cite as:
Du S, Anada A, Montenegro M, Maitland M, Chuan S, Jose S, Sihota H, Cooper D, Rivera J, Burditt D, Naruki-van Velzen M, MacPherson M
Effectiveness of Postdischarge Telephone Calls in Reducing Hospital Utilization: Quasi-Randomized Controlled Trial
J Med Internet Res 2026;28:e80529
URL: https://www.jmir.org/2026/1/e80529
doi:10.2196/80529

XSL•FO
RenderX

© Sandie Du, Abigail Anada, Maria Montenegro, Maurie Maitland, Sean Chuan, Sonia Jose, Harveer Sihota, Delia Cooper, Joanna Rivera, Dianne Burditt, Minetaro Naruki-van Velzen, Megan MacPherson. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effects of Self-Compassion and Mindfulness Interventions on Mental Health and Work-Related Outcomes Among Japanese Workers: Randomized Controlled Trial

Takumu Kurosawa*, PhD; Koichiro Adachi, MA; Ryu Takizawa*, MD, PhD

Department of Clinical Psychology, Graduate School of Education, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, Japan

*these authors contributed equally

Corresponding Author:
Ryu Takizawa, MD, PhD
Department of Clinical Psychology, Graduate School of Education, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, Japan

Abstract

Background: Mental health problems among workers are a significant global concern, leading to substantial economic losses, particularly due to presenteeism. While mindfulness and self-compassion practices have shown promise in improving psychological well-being and occupational outcomes, traditional interventions often require a high time commitment. Low-intensity digital interventions remain underexplored, especially in the context of occupational health.

Objective: This study aimed to develop and evaluate the effectiveness of a smartphone-based self-care app that enables independent practice of mindfulness and self-compassion without facilitator support. Specifically, it assessed the effects of self-compassion meditation (SCM) and mindfulness meditation (MM) on mental health and work-related outcomes among nonclinical Japanese workers.

Methods: This open-label, 3-arm randomized controlled trial recruited 300 working adults in Japan, who were randomly assigned to SCM (n=101), MM (n=100), or a waitlist control group (n=99). Participants in the SCM and MM groups engaged in daily guided meditation via a custom-built smartphone app over 4 weeks. Primary outcomes included psychological distress (Kessler Psychological Distress Scale-6) and work performance (World Health Organization Health and Work Performance Questionnaire and the Stanford Presenteeism Scale), while secondary outcomes encompassed cognitive flexibility, self-compassion, perceived stress, work engagement, psychological safety, and self-perceived creativity. Assessments were conducted at baseline, postintervention (4 wk), and 1-month follow-up (8 wk). Recruitment occurred in 2 waves (November 2022-December 2022 and June 2023), with follow-up assessments completed by September 2023.

Results: A total of 300 participants (mean age of 35.44, SD 9.14 y; n=180, 60% female) were randomized. Adherence was high in both intervention groups (mean completed days: SCM 23.30, SD 5.58; MM 22.95, SD 6.25), with an overall dropout rate of 21.4%. No significant group × time interactions were detected for most outcomes, although significant main effects of time were observed for several measures, including work performance and self-compassion, with small effect sizes. Within-group analyses suggested broader improvements in the SCM group, whereas changes in the MM group were more limited. In sensitivity analyses using linear mixed-effects models, a significant group × time interaction was detected for work performance, with a larger pre-post improvement in SCM.

Conclusions: This low-intensity, fully self-guided, smartphone-based meditation program demonstrated good feasibility, as reflected by high adherence and relatively low attrition. Between-group effects were limited; therefore, the findings should be interpreted as preliminary and do not support strong causal claims of intervention superiority. Clinical significance remains unclear in this nonclinical sample and should be evaluated in future studies.

Trial Registration: UMIN-CTR UMIN000049466; https://center6.umin.ac.jp/cgi-open-bin/ctr_e/ctr_view.cgi?recptno=R000056324

International Registered Report Identifier (IRRID): RR2-10.2196/53541

(J Med Internet Res 2026;28:e79991) doi:10.2196/79991

KEYWORDS

self-compassion; mindfulness; mobile apps; digital health; occupational health; mental health; randomized controlled trial; RCT; work performance

XSL•FO
RenderX

Introduction

Mental health problems among workers pose a critical challenge with far-reaching implications for the global economy. According to the World Health Organization (WHO), depression and anxiety disorders account for the loss of approximately 12 billion workdays annually, resulting in an annual economic burden of approximately US $1 trillion [1]. A survey of 10,000 workers in Japan identified mental illnesses, stiff shoulders, and lower back pain as health problems affecting job performance. Among these, presenteeism—employees being present at work but unable to perform at full capacity due to mental illness—causes the most significant economic losses [2]. These findings underscore the need for initiatives that improve workers' mental health and productivity.

Mindfulness and self-compassion are effective approaches for supporting workers' mental health. These approaches are often categorized within the "third wave" of cognitive-behavioral therapies, which emphasize context- and acceptance-based strategies aimed at altering one's relationship with internal experiences rather than primarily modifying cognitive content [3]. Mindfulness is the practice of maintaining nonjudgmental awareness of one's thoughts, emotions, bodily sensations, and surrounding environment in the present moment [4], while self-compassion entails being warm and understanding toward ourselves when we suffer, fail, or feel inadequate, rather than ignoring our pain or flagellating ourselves with self-criticism [5]. Mindfulness practice has been shown to enhance focus at work, self-fulfillment, and sleep quality, while reducing emotional exhaustion, occupational stress, and anxiety [6-8]. It has also been reported to improve job performance, workplace relationships, and creativity [9,10]. Similarly, self-compassion has been found to alleviate stress and burnout and to improve well-being and resilience. However, further research is needed to clarify its specific effects on work-related outcomes [11].

Mindfulness and self-compassion both use meditation as a core intervention strategy and share the goal of fostering nonjudgmental acceptance of one's experiences. However, self-compassion places additional emphasis on responding to difficulty with self-kindness and a sense of common humanity [12]. Furthermore, mindfulness-based interventions may cultivate self-compassion more implicitly than self-compassion–focused programs, which teach these elements explicitly [13]. Elevated levels of self-compassion have been correlated with more adaptive emotion regulation and enhanced well-being [14], and interventions aimed at increasing self-compassion have proven effective in mitigating self-criticism [15]. Within the employee samples, self-compassion has been linked to job-related outcomes, including performance indicators [16], and its related capacities may contribute to the development of psychologically safer environments and promote learning-oriented behaviors within teams [17]. When implementing short-term, low-intensity interventions for busy workers, it is crucial to select the most effective components. Accordingly, directly comparing mindfulness meditation (MM) and self-compassion meditation (SCM) may help clarify which elements are particularly relevant for improving work-related outcomes.

Traditional mindfulness and self-compassion programs typically consist of 8 weeks of in-person sessions with daily practice, which has been identified as a barrier to participation for busy workers [4,5,18]. In contrast, digital interventions using websites or smartphone apps have been shown to be effective in reducing stress and improving well-being, while requiring less time and cost [19,20]. These low-intensity interventions provide a more accessible alternative for users. Digital interventions exhibit varying levels of human support, ranging from fully self-administered programs to guided formats that incorporate contact with a coach or therapist [21]. Although unguided interventions offer high scalability and low cost, they tend to show lower adherence and reduced efficacy compared to guided approaches [21]. In the context of occupational digital mental health interventions, additional support and increased intensity may improve engagement and outcomes, albeit at the expense of scalability and increased implementation demands [22]. Consequently, engagement with mindfulness-based mobile apps varies widely, underscoring the importance of evaluating feasibility and reporting engagement transparently, especially in fully self-guided delivery [23]. However, despite their growing popularity, these low-intensity digital interventions remain underrepresented in randomized controlled trials (RCTs), underscoring the need for further empirical research. Additionally, the COVID-19 pandemic has accelerated the adoption of online psychological support, emphasizing the importance of developing accessible digital tools to assist workers, especially as remote work becomes increasingly prevalent [24]. In response to this need, this study aims to develop a self-care app that facilitates independent practice of mindfulness and self-compassion and to evaluate its effectiveness using mental health and work-related outcomes through an RCT design.

Methods

Study Design

This open-label, 3-arm RCT compared the effects of 2 intervention groups—the SCM and MM groups—with a waitlist control group (WCG). Both intervention groups engaged in guided meditation sessions delivered through a mobile app developed specifically for the study. Further details are provided in the study protocol [25].

Participants

The inclusion criteria were employees who were (1) working more than 20 hours per week, (2) aged between 18 and 54 years, (3) not on a leave of absence, (4) not business owners or students, and (5) not currently diagnosed with a mental disorder and scoring below 13 on the Kessler Psychological Distress Scale-6 [26]. The smartphone app used in this study was only available on iOS, restricting participation to iPhone users. The study involved 300 workers in Japan via open calls from websites. Participants were required to have basic digital literacy to install and use the smartphone app and to complete web-based questionnaires. The participants were not informed of their group allocation.

XSL•FO
RenderX

Ethical Considerations

This study was approved by the ethics committee of the University of Tokyo (22-326). The study was registered with the University Hospital Medical Information Network Clinical Trials Registry (UMIN000049466), and the study protocol was published previously [25]. Electronic informed consent was obtained via a web-based form prior to eligibility screening and randomization, and participants could withdraw at any time without penalty. Survey responses and app log data were managed and analyzed using study-generated participant IDs. Identifying information was stored separately from study data; access to the data was restricted to the research team, and the linkage file was stored offline in a secure location. Participants received a gift card worth up to JP ¥10,000 (approximately US $63; JP ¥158=US $1 as of January 11, 2026), depending on completion of study procedures; compensation was not contingent on adherence to the intervention (eg, the number of meditation days completed).

Procedure

Participants were recruited via open calls during 2 recruitment waves (November 2022-December 2022 and June 2023) through a psychology experiment recruitment website and our laboratory's website, due to insufficient enrollment in the first wave. Study details and contact information were available on website posters. Interested individuals completed a web form with a screening survey (Kessler Psychological Distress Scale-6) and were screened based on the inclusion criteria. The baseline assessment (T0), the posttreatment assessment at 4 weeks (T1), and the follow-up assessment at 8 weeks (T2) were assessed using self-report questionnaires. Questionnaires at each time point were administered independently of the intervention through an external web-based form, rather than within the app, with survey links distributed to participants via email. For the SCM and MM groups, T2 (8 wk) represents the 1-month follow-up after the 4-week intervention period. As assessments were scheduled relative to each participant's intervention timeline, the exact calendar dates varied by participant; however, all follow-up (T2) assessments were completed by September 2023. No significant methodological changes were implemented after trial commencement; any deviations from the protocol regarding outcome measures are described in the "Methods" section.

Randomization and Blinding

After obtaining informed consent and screening participants, those who met the eligibility criteria were randomly assigned to 1 of the 3 groups (1:1:1) via a computerized blocked randomization scheme (block size=15). This randomization was conducted independently by research staff, and participants were immediately assigned to either the intervention (SCM or MM) or WCG. The intervention commenced approximately 1 week after allocation, while the waitlist group received no specific instructions and waited for 4 weeks. Participants received group-specific instructions via automated email. As the participants followed a meditation guide, blinding was not possible, thus making this an open-label RCT. The instructions were sent via automated emails, and although the participants could infer their group, they were not explicitly informed of their assignment.

Interventions

The SCM and MM courses were delivered through a smartphone app developed by the research team. Participants assigned to the SCM and MM groups used the app to engage in guided meditation once daily for 4 weeks. Each course comprised 3 modules that integrated guided self-help meditation and psychoeducation. The guided meditations in weeks 1 and 2 were identical across both courses: a 7-minute breathing meditation in week 1 and a 6-minute short body scan in week 2. In weeks 3 and 4, SCM participants practiced a 12-minute loving-kindness meditation focused on self-kindness and common humanity, while MM participants engaged in a 12-minute breath, sound, and body meditation. The guided meditations were adapted from scripts developed by the University of California, Los Angeles, Mindful Awareness Research Center and modified for a Japanese audience. Psychoeducation covered topics such as stress responses, mindfulness, and self-compassion and was provided as the studies were accessible via the app at any time. Weekly email reminders encouraged participation, though no minimum time commitment was required. Meditation practice was systematically recorded by the app through data logging, and engagement metrics were derived from these logs. The intervention was fully automated and self-directed, with no support from a human instructor or therapist. Participants could view their completed practice days and total practice time within the app.

We did not instruct participants to use the app in a specific location and did not collect data on the settings where sessions were completed (eg, home vs workplace); participants were advised to practice in a quiet, private environment without interruptions. Participants in the WCG received no intervention during the 4-week period; afterward, they were allocated to either the SCM or MM course for an additional 4 weeks of guided meditation. Because WCG participants were offered access to an intervention after the 4-week waiting period for ethical reasons, they did not remain in a controlled condition during weeks 4 to 8 and were therefore not included in the 8-week (T2) follow-up analyses.

Outcome Measures

This study aimed to investigate the efficacy of a brief, self-administered smartphone-based meditation program, focusing on self-compassion and mindfulness, in enhancing workers' mental health and work-related functioning. Therefore, the primary outcomes were defined as (1) psychological distress, which is a direct target of the intervention; and (2) work performance (presenteeism), a pragmatic occupational outcome. As secondary outcomes, we assessed potentially modifiable psychological processes and related indicators (eg, cognitive flexibility, self-compassion, perceived stress, and work engagement) to complement the evaluation of changes in mental health and work-related outcomes. Although daily physiological measures (including heart rate and heart rate variability) were planned in the protocol, they are not reported in this paper because this report focuses on psychological and work-related

XSL•FO
RenderX

outcomes; these physiological outcomes will be reported separately.

Primary Outcomes

The primary outcomes were psychological distress and work performance, assessed at 3 time points: baseline (T0), posttreatment at 4 weeks (T1), and follow-up at 8 weeks (T2). Psychological distress was measured using the Japanese version of the Kessler Psychological Distress Scale-6 [26,27], a 6-item instrument designed to measure nonspecific psychological distress (depressive mood and anxiety) over the preceding 30 days (total score range: 0 - 24; higher scores indicate greater distress), with good internal consistency in this study (Cronbach α at T0=0.85). A cutoff score of 12/13 on the Kessler Psychological Distress Scale-6 has been proposed for screening severe depressive symptoms [28], and thus, an ethical exclusion criterion of Kessler Psychological Distress Scale-6 ≥13 was established. Work performance was evaluated using one presenteeism item from the Japanese version of the World Health Organization Health and Work Performance Questionnaire (WHO-HPQ) [29,30], which assesses overall work performance over the past 28 days (absolute presenteeism score range: 0 - 100; higher scores indicate better performance). Moreover, work inefficiency was assessed using the work impairment score from the Stanford Presenteeism Scale (SPS) [31,32], which captures the extent to which health-related cognitive, emotional, and behavioral difficulties impede task completion and focus (10 items; 1="always" to 5="never"; scored such that higher scores indicate greater impairment). This scale exhibited acceptable internal consistency (Cronbach α=0.72) in this study. While the protocol specified presenteeism assessed using the WHO-HPQ, this report additionally includes the SPS work impairment score to capture health-related work impairment.

Secondary Outcomes

The following self-reported questionnaires were administered as secondary outcome measures at each time point: the Cognitive Flexibility Scale (Cronbach α=0.86) [33,34], which assesses the ability to generate alternative responses and adapt behavior to changing circumstances (12 items; total score range: 12 - 72; higher scores indicate greater cognitive flexibility); the Self-Compassion Scale (SCS; Cronbach α=0.94) [35,36], which assesses compassionate self-responding toward oneself across 6 facets (26 items; total score range: 26 - 130; higher scores indicate greater self-compassion), including its subscales: self-kindness (Cronbach α=0.92), common humanity (Cronbach α=0.80), mindfulness (Cronbach α=0.79), self-judgment (Cronbach α=0.89), isolation (Cronbach α=0.84), and overidentification (Cronbach α=0.81); the Perceived Stress Scale-14 (Cronbach α=0.89) [37-39], which assesses the extent to which situations in the past month are perceived as unpredictable, uncontrollable, and overloaded (total score range: 0 - 56; higher scores indicate greater perceived stress); and the Utrecht Work Engagement Scale-9 (Cronbach α=0.95) [40-42], which assesses vigor, dedication, and absorption (total score range: 0 - 54; higher scores indicate higher engagement), including its subscales: vigor (Cronbach α=0.89), dedication (Cronbach α=0.84), and absorption (Cronbach α=0.88). Additionally, psychological safety was assessed using the Psychological Safety Scale (7 items; Cronbach α=0.80) [43,44], which reflects the extent to which individuals can express themselves at work without fear of negative consequences; items are rated from 1 ("not at all") to 7 ("strongly agree"), yielding total scores of 7 to 49. Creativity (idea generation at work) was assessed using the Self-Perceived Creativity Scale (3 items; Cronbach α=0.91) [45], rated from 1 ("do not agree") to 7 ("agree"), yielding total scores of 3 to 21. For further details on these outcome measures, refer to the study protocol [25].

Other Measures

Demographic variables were assessed using a questionnaire at baseline (T0). The collected variables included age, gender, employment status, marital status, educational background, and prior meditation experience, including mindfulness practices.

Sample Size

In the protocol, the sample size was planned based on detecting a small group-by-time interaction effect (Cohen *f*=0.15) for the primary pre-post comparison across the 3 arms, assuming a correlation of 0.50 between repeated measures, with 80% power at a 2-sided α=.05 (G*Power; Heinrich Heine University Düsseldorf) [25]. This yielded a required sample size of 37 participants per arm. The protocol, therefore, initially planned to randomize 200 participants to account for expected attrition. During the trial conduct, we observed that a proportion of participants did not complete the baseline assessment following randomization; therefore, recruitment was extended and a total of 300 participants were randomized to ensure an adequate evaluable sample size at key assessment time points and to improve the precision of effect estimates. This target sample size was reflected in the trial registration (UMIN000049466). The increase in sample size was an operational decision made without reference to intervention effects, and no interim efficacy analyses were conducted.

Statistical Analyses

Missing Data and Sensitivity Analysis

Analyses were conducted using available cases at each assessment point. Participants with missing data for a specific analysis were excluded from that analysis, resulting in varying sample sizes across time points. As a sensitivity analysis, linear mixed-effects models (LMMs) were fitted using all available data to assess the robustness of the findings in relation to attrition.

Baseline Analyses

Demographic variables and baseline psychological indices were summarized descriptively (mean and SD for continuous variables; n and % for categorical variables) for each group. In accordance with CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidance (Checklist 1), we did not perform significance testing for baseline group comparisons.

XSL•FO
RenderX

Analysis of Preintervention and Postintervention Effects

Changes before and after the intervention or waitlist periods were assessed using 2-tailed paired *t* tests within each intervention group. Effect sizes were calculated as Cohen *d* and interpreted using cutoff values of 0.20, 0.50, and 0.80 for small, medium, and large effects, respectively [38].

Analysis of Effect Size in Group Comparisons

To examine preintervention and postintervention changes and pre-post differences between groups, assessment questionnaire results were analyzed using a 2 (time: pre and post) × 3 (group) repeated-measures ANOVA. Effect sizes for ANOVA were reported as generalized η^2, with cutoff values of 0.01, 0.06, and 0.14 for small, medium, and large effects, respectively [38,46].

Analysis of 1-Month Follow-Up

As the waitlist group transitioned to receive an intervention after T1, all T2 follow-up analyses were restricted to participants initially allocated to SCM or MM. The 1-month follow-up scores were analyzed using a series of one-way within-subjects ANOVAs to compare scores within each intervention group across three time points: (1) preintervention, (2) postintervention, and (3) 1-month follow-up. For measures that showed significant time effects, Bonferroni-adjusted pairwise comparisons were conducted to examine changes within each intervention group.

To assess the maintenance of treatment effects and differences between the intervention groups, a 3 (time: pre, post, and follow-up) × 2 (group: SCM and MM) repeated-measures ANOVA was performed. For measures that showed significant time effects, Bonferroni-adjusted pairwise comparisons were conducted to investigate changes across the intervention groups.

Sensitivity Analyses Using LMMs

Primary inference in this trial was derived from between-group comparisons (group × time). Within-group tests in the main analyses were conducted as exploratory analyses to describe within-group patterns and should not be interpreted as evidence of causal superiority. As a sensitivity analysis, we reanalyzed the primary and secondary outcomes using LMMs to assess the robustness of the findings. These models included fixed effects for group, time, and their interaction, with a random intercept for participants to account for within-subject correlations. Additionally, variables showing notable baseline imbalances were included as covariates in sensitivity analyses to assess the robustness of the findings. The models were fitted using all available data under a missing-at-random assumption. For the pre-post comparison (T0-T1), analyses included all randomized groups (SCM, MM, and WCG), whereas follow-up analyses (T0-T1-T2) were restricted to SCM and MM because the WCG crossed over after T1.

Results

Participant Flow and Sample Characteristics

In total, 300 participants were randomly allocated to SCM (n=101), MM (n=100), and WCG (n=99; Figure 1). Of these 300 randomized participants, 251 completed the baseline assessment (T0; 83.7%), including 86 (85.1%) in SCM, 85 (85%) in MM, and 80 (80.8%) in WCG, and were included in Table 1. At posttreatment (T1), outcome data were available for 234 participants (78%; SCM: n=79, 78.2%; MM: n=79, 79%; WCG: n=76, 76.8%). At follow-up (T2), outcome data were available for 136 (67.7% of the SCM and MM) participants in the intervention arms (SCM: n=70, 69.7%; MM: n=66, 66%). No important harms or unintended effects (including major technical problems or privacy incidents) were reported to the study team during the trial. Among those who completed the baseline assessment (Table 1), the sample consisted of 60% (n=180) female participants, with a high proportion of regular employees (n=57, 67.1%). The mean age was 35.44 (SD 9.14) years. The groups were generally similar at baseline; however, age and prior meditation experience showed some imbalance across groups. Therefore, age (grand-mean centered) and prior meditation experience were included as covariates in the sensitivity analyses using LMMs (Figure 1).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of participant enrollment, allocation, and follow-up assessments. MM: mindfulness meditation; SCM: self-compassion meditation; WCG: waitlist control group.

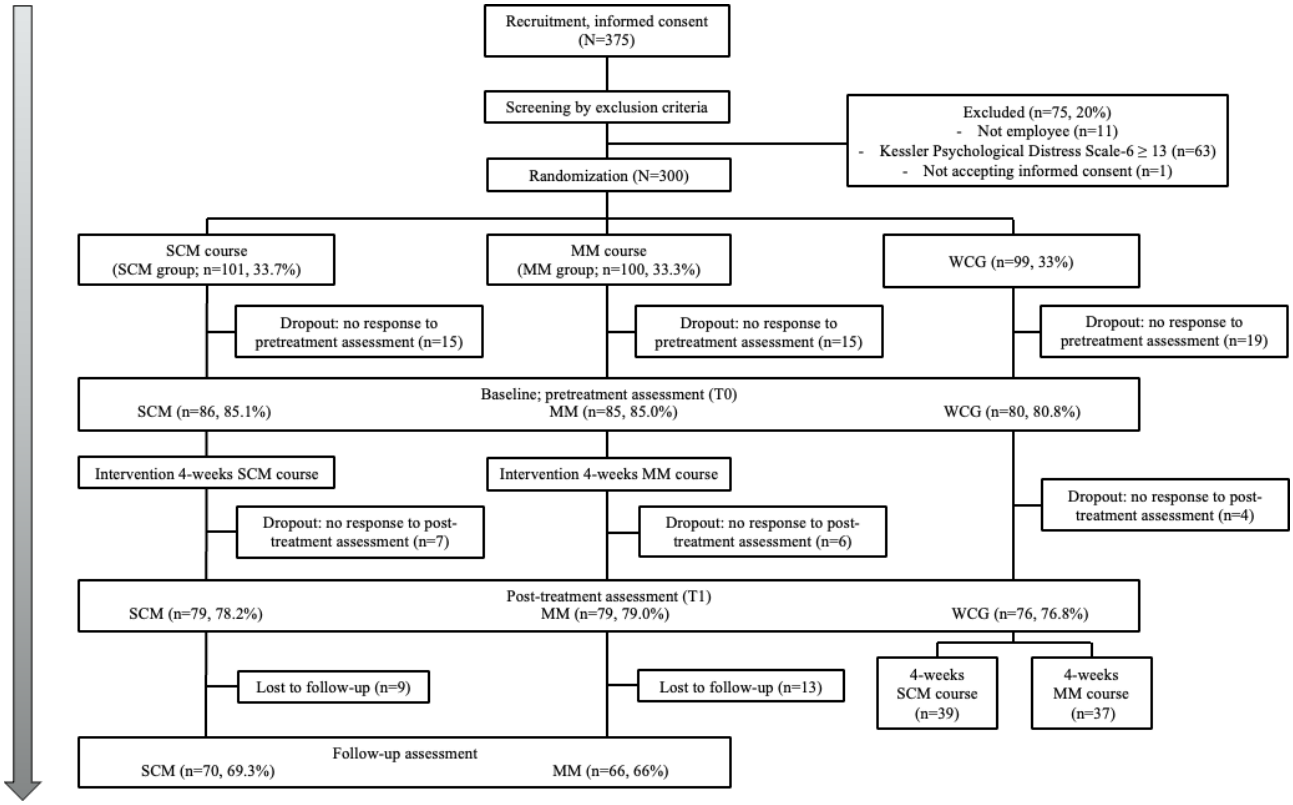

XSL•FO
RenderX

Table . Baseline demographic characteristics of participants by group.

Variable	SCM[a] group (n=86)	MM[b] group (n=85)	Waitlist group (n=80)
Age (y), mean (SD)	33.96 (8.84)	35.02 (9.65)	37.45 (8.65)
Sex, n (%)			
Female	51 (59.3)	47 (55.3)	52 (65)
Male	34 (39.5)	38 (44.7)	28 (35)
Employment status, n (%)			
Regular	59 (68.6)	57 (67.1)	52 (65)
Nonregular	27 (31.4)	27 (31.8)	28 (35)
Marital status, n (%)			
Single	50 (58.1)	52 (61.2)	35 (43.8)
Married	35 (40.7)	32 (37.6)	45 (56.2)
Educational attainment, mean (SD)	15.70 (1.67)	15.88 (1.72)	15.79 (1.63)
Previous meditation experience, n (%)			
Yes	30 (34.9)	18 (21.2)	15 (18.8)
No	56 (65.1)	66 (77.6)	64 (80)
Psychological distress, mean (SD)	7.05 (4.51)	6.23 (4.96)	6.62 (4.92)
Work performance, mean (SD)	62.79 (18.32)	59.88 (18.16)	60.25 (20.68)
Work inefficiency, mean (SD)	26.46 (5.21)	26.89 (6.02)	26.13 (4.92)
Cognitive flexibility, mean (SD)	47.28 (8.43)	46.01 (8.46)	47.00 (9.11)
Self-compassion, mean (SD)	76.39 (17.43)	74.14 (19.46)	79.85 (18.73)
Self-kindness, mean (SD)	15.74 (4.55)	15.02 (4.89)	16.16 (4.81)
Self-judgment, mean (SD)	15.80 (4.82)	15.79 (4.76)	14.83 (5.26)
Common humanity, mean (SD)	10.89 (3.72)	10.85 (3.76)	11.49 (3.80)
Isolation, mean (SD)	10.81 (4.05)	11.74 (4.04)	10.74 (4.44)
Mindfulness, mean (SD)	12.47 (3.07)	11.88 (3.39)	12.81 (3.18)
Overidentification, mean (SD)	13.52 (3.55)	13.75 (3.86)	13.46 (3.72)
Perceived stress, mean (SD)	28.50 (9.58)	28.75 (10.81)	28.50 (9.47)
Work engagement, mean (SD)	28.41 (13.51)	26.14 (12.30)	27.69 (12.20)
Vigor, mean (SD)	8.64 (4.72)	7.79 (4.51)	8.66 (4.52)
Dedication, mean (SD)	10.62 (4.53)	9.98 (4.10)	10.28 (4.14)
Absorption, mean (SD)	9.29 (4.95)	8.38 (4.41)	8.57 (4.47)
Psychological safety, mean (SD)	35.30 (6.65)	34.29 (7.44)	33.60 (7.93)
Creativity, mean (SD)	12.19 (4.37)	11.67 (4.35)	11.36 (4.09)

[a]SCM: self-compassion meditation.

[b]MM: mindfulness meditation.

The rate of participants who dropped out before posttreatment assessment did not differ among groups (χ^2_2=0.66; *P*=.72). On average, the participants practiced for 23.30 (SD 5.58) days in the SCM group and 22.95 (SD 6.25) days in the MM group during the intervention period. The 2-tailed *t* tests showed no significant differences in the number of practice days between the 2 groups (t_{154}=0.38; *P*=.71).

Pre-Post Differences Between the Groups

Table 2 summarizes the mean and SD for each group, along with preintervention and postintervention comparisons using 2-tailed paired *t* tests within each intervention group. In the SCM group, paired *t* tests revealed significant improvements in work performance, cognitive flexibility, self-compassion, common humanity (as measured by the SCS), and psychological safety. Significant reductions were observed in work inefficiency, perceived stress, self-judgment, and overidentification (as measured by the SCS). In the MM group,

XSL•FO
RenderX

2-tailed paired *t* tests showed a significant improvement in common humanity (as measured by the SCS). Table 3 presents the results of between-group comparisons using ANOVA. No significant interaction effect between group and time was found.

XSL•FO
RenderX

Table . Mean (SD) scores for intervention and control groups at preintervention and postintervention and changes in scores in each intervention group.

	Preintervention, mean (SD)	Postintervention, mean (SD)	*t* test (*df*)	*P* value	Cohen *d*
Psychological distress					
SCM[a] group	6.68 (4.27)	6.90 (4.40)	0.49 (77)	.62	0.06
MM[b] group	5.94 (4.92)	6.38 (5.08)	1.05 (77)	.30	0.09
Waitlist group	6.75 (4.76)	7.12 (5.44)	1.14 (73)	.26	0.09
Work performance					
SCM group	63.29 (16.93)	70.00 (15.85)	4.28 (78)	<.001	0.41
MM group	59.37 (17.27)	60.25 (18.81)	0.49 (78)	.62	0.05
Waitlist group	59.87 (21.07)	62.93 (20.58)	1.56 (74)	.12	0.15
Work inefficiency					
SCM group	26.22 (5.04)	24.85 (5.52)	–2.21 (75)	.03	–0.27
MM group	26.78 (5.99)	25.83 (5.81)	–1.34 (72)	.18	–0.14
Waitlist group	26.15 (5.02)	25.81 (6.00)	–1.11 (71)	.27	–0.09
Cognitive flexibility					
SCM group	47.69 (8.18)	49.04 (8.73)	2.27 (71)	.03	0.18
MM group	45.81 (8.49)	45.86 (9.09)	0.11 (76)	.91	0.01
Waitlist group	46.84 (9.31)	46.85 (9.32)	0.26 (73)	.80	0.01
Self-compassion					
SCM group	76.30 (17.94)	81.17 (18.06)	2.86 (65)	.006	0.24
MM group	74.55 (19.27)	77.32 (20.11)	1.91 (71)	.06	0.10
Waitlist group	79.42 (19.11)	79.70 (19.71)	1.33 (70)	.19	0.07
Self-kindness					
SCM group	15.79 (4.66)	16.56 (4.77)	1.68 (73)	.10	0.15
MM group	15.12 (4.91)	15.48 (4.99)	0.80 (77)	.42	0.05
Waitlist group	16.28 (4.90)	16.33 (5.01)	0.15 (75)	.88	0.01
Self-judgment					
SCM group	15.70 (4.96)	14.77 (4.38)	–2.36 (74)	.02	–0.21
MM group	15.68 (4.74)	15.42 (4.92)	–0.83 (77)	.41	–0.05
Waitlist group	15.13 (5.22)	14.69 (4.92)	–1.29 (74)	.20	–0.09
Common humanity					
SCM group	10.67 (3.78)	11.47 (3.98)	2.39 (75)	.02	0.19
MM group	10.97 (3.73)	11.48 (3.75)	2.19 (75)	.03	0.14
Waitlist group	11.56 (3.85)	11.63 (3.81)	0.21 (74)	.83	0.02
Isolation					
SCM group	10.64 (4.10)	10.14 (3.84)	–1.65 (74)	.10	–0.14
MM group	11.51 (3.93)	11.23 (3.81)	–0.85 (78)	.40	–0.07
Waitlist group	10.92 (4.47)	10.91 (4.27)	–0.04 (75)	.97	0.00
Mindfulness					
SCM group	12.38 (3.13)	12.78 (3.53)	1.19 (76)	.24	0.11
MM group	11.94 (3.34)	12.14 (3.55)	0.58 (77)	.56	0.05
Waitlist group	12.80 (3.24)	13.01 (3.16)	0.90 (74)	.37	0.07
Overidentification					

XSL•FO
RenderX

	Preintervention, mean (SD)	Postintervention, mean (SD)	*t* test (*df*)	*P* value	Cohen *d*
SCM group	13.42 (3.57)	12.74 (3.73)	–2.27 (76)	.03	–0.20
MM group	13.68 (3.89)	13.37 (3.86)	–1.19 (76)	.24	–0.08
Waitlist group	13.57 (3.73)	13.49 (4.12)	–0.31 (73)	.76	–0.02
Perceived stress					
SCM group	28.31 (9.86)	25.19 (8.52)	–2.65 (75)	.01	–0.31
MM group	28.46 (10.69)	27.05 (9.65)	–1.27 (77)	.21	–0.13
Waitlist group	28.99 (9.35)	27.61 (9.54)	–1.39 (67)	.17	–0.11
Work engagement					
SCM group	28.21 (13.55)	29.51 (13.29)	1.32 (73)	.19	0.07
MM group	25.58 (12.04)	26.03 (12.72)	0.52 (78)	.60	0.04
Waitlist group	27.61 (12.07)	28.08 (13.36)	0.79 (73)	.43	0.05
Vigor					
SCM group	8.54 (4.73)	8.96 (4.60)	1.17 (77)	.25	0.08
MM group	7.54 (4.41)	8.04 (4.54)	1.44 (78)	.15	0.11
Waitlist group	8.71 (4.46)	8.84 (4.97)	0.67 (74)	.51	0.05
Dedication					
SCM group	10.59 (4.58)	10.87 (4.55)	1.31 (77)	.19	0.07
MM group	9.84 (4.03)	9.87 (4.30)	0.13 (78)	.90	0.01
Waitlist group	10.28 (4.10)	10.50 (4.43)	0.82 (75)	.42	0.05
Absorption					
SCM group	9.22 (4.95)	9.72 (4.77)	1.27 (75)	.21	0.09
MM group	8.20 (4.35)	8.11 (4.62)	–0.25 (78)	.80	–0.02
Waitlist group	8.60 (4.43)	8.74 (4.78)	0.31 (74)	.76	0.02
Psychological safety					
SCM group	35.40 (6.83)	36.71 (6.73)	2.12 (75)	.04	0.17
MM group	34.34 (7.38)	34.23 (7.79)	–0.18 (78)	.86	–0.01
Waitlist group	33.57 (8.01)	33.80 (9.10)	0.48 (74)	.64	0.04
Creativity					
SCM group	12.17 (4.33)	12.45 (4.46)	1.09 (73)	.28	0.09
MM group	11.55 (4.38)	11.08 (4.26)	–1.18 (77)	.24	–0.10
Waitlist group	11.41 (4.18)	11.65 (4.73)	0.76 (74)	.45	0.06

[a]SCM: self-compassion meditation.

[b]MM: mindfulness meditation.

XSL•FO
RenderX

Table . Group × time interaction effect using ANOVA.

Variable	*F* test (*df*)	*P* value	General η^2
Psychological distress	0.09 (2, 227)	.91	0.000
Work performance	2.67 (2, 230)	.07	0.004
Work inefficiency	0.61 (2, 218)	.54	0.001
Cognitive flexibility	1.70 (2, 220)	.19	0.001
Self-compassion	1.73 (2, 206)	.18	0.001
Self-kindness	0.85 (2, 225)	.43	0.001
Self-judgment	1.19 (2, 225)	.31	0.001
Common humanity	1.36 (2, 224)	.26	0.001
Isolation	0.68 (2, 227)	.51	0.001
Mindfulness	0.15 (2, 227)	.86	0.000
OverIdentification	1.14 (2, 225)	.32	0.001
Perceived stress	1.07 (2, 219)	.34	0.002
Work engagement	0.11 (2, 224)	.90	0.000
Vigor	0.16 (2, 229)	.85	0.000
Dedication	0.28 (2, 230)	.76	0.000
Absorption	0.54 (2, 227)	.58	0.000
Psychological safety	0.99 (2, 227)	.37	0.001
Creativity	1.61 (2, 224)	.20	0.002

As a supplementary analysis, the results of the LMM for repeated measures are shown in Table S1 in Multimedia Appendix 1. A significant group × time interaction was detected for work performance ($F_{2,233}$=3.32; *P*=.04; partial η^2=0.028). A simple slope analysis revealed that participants in the SCM group showed a significant change (t_{237}=3.99; *P*<.001), whereas no significant difference was observed in the MM group (t_{237}=0.39; *P*=.70) and the WLC group (t_{237}=1.63; *P*=.11).

After adjusting for age and previous experience with meditation, significant group × time interactions were detected for work performance ($F_{2,230}$=3.22; *P*=.04; partial η^2=0.027) and cognitive flexibility ($F_{2,223}$=3.13; *P*=.046; partial η^2=0.027). Using simple slope analyses, participants in the SCM group showed a significant change in work performance (t_{234} = 3.93; *P*<.001) and cognitive flexibility (t_{224} = 2.87; *P*=.004), whereas no significant difference was observed in the MM and WLC groups.

Follow-Up Analysis

Table 4 summarizes the means and SDs of preintervention, postintervention, and follow-up measures in each intervention group. In the SCM group, significant main effects of time using ANOVA were observed for work performance ($F_{2,138}$=6.42; *P* = .002; η^2G=0.029), self-compassion ($F_{2,108}$=7.38; *P* <.001; η^2G=0.020), self-judgment ($F_{2,130}$=6.19; *P*=.003; η^2G=0.016), common humanity ($F_{2,134}$=5.72; *P*=.004; η^2G=0.012), overidentification ($F_{2,130}$=3.38; *P*=.04; η^2G=0.008), perceived stress ($F_{2,116}$=4.43; *P*=.01; η^2G=0.021), and psychological safety ($F_{2,118}$=4.39; *P*=.01; η^2G=0.013). In the SCM group, Bonferroni-adjusted simple comparisons revealed a significant increase in work performance from preintervention to postintervention (*P*<.001) and a significant decrease from postintervention to follow-up (*P*=.03). Significant increases were also observed in self-compassion from preintervention to postintervention (*P*=.03) and from preintervention to follow-up (*P*=.002), common humanity from preintervention to follow-up (*P*=.006), and psychological safety from postintervention to follow-up (*P*=.002). Significant decreases were found in self-judgment from preintervention to follow-up (*P*=.001), and perceived stress from preintervention to follow-up (*P*=.03). In the MM group, significant main effects of time using ANOVA were observed for self-compassion ($F_{2,108}$=3.35; *P*=.04; η^2G=0.004) and overidentification ($F_{2,126}$=4.28; *P*=.02; η^2G=.007). Bonferroni-adjusted simple comparisons indicated a significant decrease in overidentification from preintervention to follow-up (*P*=.01).

XSL•FO
RenderX

Table . Mean (SD) at follow-up and repeated-measures analyses across pre, post, and follow-up within the intervention groups.

	Follow-up assessment	Repeated measure ANOVA		
	Mean (SD)	*F* test (*df*)	*P* value	General η^2
Psychological distress				
SCM[a] group	5.93 (4.03)	1.98 (2, 134)	.14	0.008
MM[b] group	6.15 (5.21)	0.75 (2, 130)	.47	0.002
Work performance				
SCM group	66.71 (17.00)	6.42 (2, 138)	.002	0.029
MM group	62.42 (18.98)	0.38 (2, 130)	.68	0.002
Work inefficiency				
SCM group	26.21 (4.17)	1.68 (2, 116)	.19	0.013
MM group	25.52 (5.50)	1.02 (2, 114)	.36	0.007
Cognitive flexibility				
SCM group	48.57 (9.70)	2.00 (2, 108)	.14	0.006
MM group	46.86 (8.70)	1.97 (2, 126)	.14	0.004
Self-compassion				
SCM group	81.42 (18.05)	7.38 (2, 108)	<.001	0.020
MM group	77.35 (21.94)	3.35 (2, 108)	.04	0.004
Self-kindness				
SCM group	16.19 (5.12)	1.42 (2, 124)	.24	0.004
MM group	15.69 (5.21)	2.70 (2, 126)	.07	0.005
Self-judgment				
SCM group	14.30 (4.68)	6.19 (2, 130)	.003	0.016
MM group	15.18 (5.06)	2.45 (2, 128)	.09	0.004
Common humanity				
SCM group	11.77 (3.87)	5.72 (2, 134)	.004	0.012
MM group	15.69 (5.21)	2.37 (2, 124)	.10	0.006
Isolation				
SCM group	9.74 (3.69)	2.78 (2, 130)	.07	0.009
MM group	10.83 (3.97)	3.02 (2, 128)	.052	0.007
Mindfulness				
SCM group	13.11 (3.47)	2.70 (2, 134)	.07	0.007
MM group	11.84 (3.53)	0.22 (2, 124)	.80	0.001
Overidentification				
SCM group	12.79 (3.81)	3.38 (2, 130)	.04	0.008
MM group	13.17 (4.14)	4.28 (2, 126)	.02	0.007
Perceived stress				
SCM group	25.03 (9.88)	4.43 (2, 116)	.01	0.021
MM group	26.98 (10.42)	0.56 (2, 126)	.57	0.002
Work engagement				
SCM group	28.91 (13.99)	1.29 (2, 128)	.28	0.002
MM group	25.58 (13.63)	0.45 (2, 128)	.64	0.001
Vigor				

XSL•FO
RenderX

	Follow-up assessment	Repeated measure ANOVA		
	Mean (SD)	*F* test (*df*)	*P* value	General η^2
SCM group	8.60 (4.94)	1.44 (2, 136)	.24	0.003
MM group	7.83 (4.65)	0.64 (2, 128)	.53	0.001
Dedication				
SCM group	10.71 (4.63)	1.60 (2, 136)	.21	0.002
MM group	9.73 (4.49)	0.79 (2, 130)	.46	0.002
Absorption				
SCM group	9.60 (5.04)	1.48 (2, 132)	.23	0.002
MM group	8.14 (5.02)	0.26 (2, 130)	.77	0.001
Psychological safety				
SCM group	35.86 (7.37)	4.39 (2, 118)	.01	0.013
MM group	33.79 (7.91)	0.50 (2, 130)	.61	0.001
Creativity				
SCM group	12.19 (4.57)	0.71 (2, 130)	.49	0.001
MM group	11.44 (4.37)	1.23 (2, 128)	.29	0.003

[a]SCM: self-compassion meditation.

[b]MM: mindfulness meditation.

In the SCM group, significant main effects of time using LMM were observed for work performance ($F_{2,147}$=7.96; *P*<.001; partial η^2=0.098), work inefficiency ($F_{2,141}$=3.19; *P*=.04; partial η^2=0.043), cognitive flexibility ($F_{2,135}$=3.27; *P*=.04; partial η^2=0.046), self-compassion ($F_{2,134}$=8.66; *P*<.001; partial η^2=0.114), self-judgment ($F_{2,148}$=6.66; *P*=.002; partial η^2=0.082), common humanity ($F_{2,149}$=6.14; *P*=.003; partial η^2=0.076), isolation ($F_{2,148}$=3.74; *P*=.03; partial η^2=0.048), overidentification ($F_{2,147}$=4.08; *P*=.02; partial η^2=0.053), perceived stress ($F_{2,145}$=6.48; *P*=.002; partial η^2=0.082), and psychological safety ($F_{2,143}$=3.18; *P*=.04; partial η^2=0.043; Table S2 in Multimedia Appendix 2). In the SCM group, Bonferroni-adjusted simple comparisons revealed a significant increase in work performance from preintervention to postintervention (*P*<.001). Significant increases were also observed in self-compassion from preintervention to postintervention (*P*=.002) and to follow-up (*P*<.001), and in common humanity from preintervention to postintervention (*P*=.04) and to follow-up (*P*=.003). Significant decreases were found in self-judgment from preintervention to postintervention (*P*=.02) and to follow-up (*P*=.002), isolation from preintervention to follow-up (*P*=.02), overidentification from preintervention to follow-up (*P*=.02), and perceived stress from preintervention to postintervention (*P*=.008) and to follow-up (*P*=.008).

In the MM group, significant main effects of time using LMM were observed for self-compassion ($F_{2,133}$=4.22; *P*=.02; partial η^2=0.060), self-kindness ($F_{2,144}$=3.13; *P*=.047; partial η^2=0.042), common humanity ($F_{2,143}$=3.07; *P*=.049; partial η^2=0.041), isolation ($F_{2,145}$=3.62; *P*=.03; partial η^2=0.047), and overidentification ($F_{2,144}$=3.96; *P*=.02; partial η^2=0.052). Bonferroni-adjusted simple comparisons indicated a significant increase in self-compassion from preintervention to follow-up (*P*=.02) and self-kindness from preintervention to follow-up (*P*=.04). Significant decreases were found in isolation from preintervention to follow-up (*P*=.02) and overidentification from preintervention to follow-up (*P*=.02; Table S2 in Multimedia Appendix 2).

Table 5 summarizes the main effects for time and the interaction effects for group and time. Across all intervention groups, significant main effects for time were found for work performance ($F_{2,268}$=3.80; *P*=.02; η^2G=0.009), self-compassion ($F_{2,216}$=10.44; *P*<.001; η^2G=0.010), common humanity ($F_{2,258}$=7.54; *P*<.001; η^2G=0.008), self-judgment ($F_{2,258}$=8.35; *P*<.001; η^2G=0.009), overidentification (in the SCS; $F_{2,256}$=6.62; *P*=.001; η^2G=0.007), and perceived stress ($F_{2,242}$=4.26; *P*=.02; η^2G=0.008). Bonferroni-adjusted simple comparisons showed that in the intervention group, a significant increase was observed in work performance from preintervention to postintervention (*P*=.01), as well as significant increases in self-compassion from preintervention to postintervention (*P*=.006) and to follow-up (*P*<.001), and in common humanity from preintervention to postintervention (*P*=.02) and to follow-up (*P*=.002). Significant decreases were observed in self-judgment from preintervention to postintervention (*P*=.03) and to follow-up (*P*<.001), in isolation from preintervention to follow-up (*P*=.002), in overidentification from preintervention to postintervention (*P*=.02) and to follow-up (*P*=.002), and in perceived stress from preintervention to follow-up (*P*=.02).

XSL•FO
RenderX

Table . Main effects for time and interaction effects for group × time using ANOVA.

	Time			Interaction effects		
	F test (*df*)	*P* value	General η^2	*F* test (*df*)	*P* value	General η^2
Psychological distress	2.44 (2, 264)	.09	0.004	0.23 (2, 264)	.80	0.000
Work performance	3.80 (2, 268)	.02	0.009	2.32 (2, 268)	.10	0.005
Work inefficiency	2.37 (2, 230)	.10	0.008	0.26 (2, 230)	.77	0.001
Cognitive flexibility	2.48 (2, 234)	.09	0.003	1.53 (2, 234)	.22	0.002
Self-compassion	10.44 (2, 216)	<.001	0.010	0.92 (2, 216)	.40	0.001
Self-kindness	2.22 (2, 250)	.11	0.003	1.68 (2, 250)	.19	0.002
Self-judgment	8.35 (2, 258)	<.001	0.009	1.02 (2, 258)	.36	0.001
Common humanity	7.54 (2, 258)	<.001	0.008	0.32 (2, 258)	.73	0.000
Isolation	5.58 (2, 258)	.004	0.007	0.18 (2, 258)	.84	0.000
Mindfulness	1.57 (2, 258)	.21	0.002	1.12 (2, 258)	.33	0.002
Overidentification	6.62 (2, 256)	.002	0.007	0.95 (2, 256)	.39	0.001
Perceived stress	4.26 (2, 242)	.02	0.008	1.12 (2, 242)	.33	0.002
Work engagement	1.24 (2, 256)	.29	0.001	0.33 (2, 256)	.72	0.000
Vigor	1.94 (2, 264)	.15	0.002	0.10 (2, 264)	.90	0.000
Dedication	1.81 (2, 266)	.16	0.001	0.39 (2, 266)	.67	0.000
Absorption	0.53 (2, 262)	.59	0.001	0.86 (2, 262)	.42	0.001
Psychological safety	2.72 (2, 248)	.07	0.004	1.75 (2, 248)	.18	0.002
Creativity	0.07 (2, 258)	.94	0.000	1.93 (2, 258)	.15	0.002

As a supplementary analysis, the main effects for time and the interaction effects for group × time were examined using LMM (Table S3 in Multimedia Appendix 3). Across all intervention groups, significant main effects for time were found for work performance ($F_{2,292}$=5.52; *P*=.004; partial η^2=0.036), self-compassion ($F_{2,267}$=3.93; *P*=.02; partial η^2=0.029), self-judgment ($F_{2,293}$=3.84; *P*=.02; partial η^2=0.025), and perceived stress ($F_{2,290}$=3.35; *P*=.04; partial η^2=0.023). Bonferroni-adjusted simple comparisons showed that in the intervention group, a significant increase was observed in work performance from preintervention to postintervention (*P*=.008), and in self-compassion from preintervention to postintervention (*P*<.001) and follow-up (*P*<.001). Significant decreases were observed in self-judgment from preintervention to postintervention (*P*<.001) and follow-up (*P*=.02), and in perceived stress from preintervention to postintervention (*P*=.006) and follow-up (*P*=.003). A significant group × time interaction was detected for work performance ($F_{2,293}$=3.23; *P*=.04; partial η^2=0.022). A simple slope analysis revealed that participants in the SCM group showed a significant increase from preintervention to postintervention (t_{301}= 3.88; *P*<.001), whereas no other significant differences were observed.

After adjusting for age and previous experience with meditation, all outcomes that were statistically significant in the unadjusted analyses remained significant with respect to the main effect of time. In addition, cognitive flexibility also showed a significant main effect of time ($F_{2,277}$=3.09; *P*=.047; partial η^2=0.022). Bonferroni-adjusted simple comparisons showed that in the intervention group, a significant increase was also observed in cognitive flexibility from preintervention to follow-up (*P*=.046). A significant group × time interaction remained for work performance ($F_{2,291}$=3.16; *P*=.04; partial η^2=0.021). A simple slope analysis revealed that participants in the SCM group showed a significant increase from preintervention to postintervention (t_{298}=3.83; *P*<.001), whereas no other significant differences were observed.

Discussion

Principal Findings

This study aimed to develop a self-care smartphone app that enables individuals to independently practice mindfulness and self-compassion without the support of a facilitator, and to evaluate its effectiveness using psychological and work-related outcomes. Participants were randomized to SCM, MM, or WCG, and intervention effects were evaluated in an RCT.

SCM participants demonstrated improvements in work performance, cognitive flexibility, self-compassion (especially common humanity), and psychological safety, along with reductions in perceived stress, work inefficiency, self-judgment, and overidentification. In contrast, MM participants only showed

XSL•FO
RenderX

improvement in common humanity. No significant group × time interaction was observed in psychological outcomes. However, sensitivity analyses using LMMs revealed a significant group × time interaction for work performance, indicating a relatively greater improvement in the SCM group. Longitudinal analyses revealed that the SCM group maintained several improvements at follow-up, particularly in self-compassion, common humanity, and stress reduction. The MM group showed limited effects, mainly in overidentification. Taken together, within-group analyses suggested improvements in several psychological and work-related outcomes in the SCM group, while changes in the MM group were more limited. Significant reductions were also found in perceived stress, work inefficiency, self-judgment, and overidentification. The next section discusses these findings in detail, including their strengths, limitations, and implications for future research and applied settings.

A comparison of psychological indicators before and after the intervention in the intervention groups revealed that, in the SCM group, work performance, cognitive flexibility, self-compassion, self-esteem, and psychological safety improved compared to preintervention levels, while perceived stress, work inefficiency, self-judgment, and overidentification significantly decreased. In the MM group, a significant increase was observed only in the common humanity subscale of self-compassion. These findings are consistent with previous studies on self-compassion interventions via smartphone apps. For example, a 7-day intervention targeting young people in New Zealand also reported improvements in self-compassion [47]. Similarly, Beshai et al [48] reported that a 4-week guided self-help program (Mind-OP), which combined mindfulness, self-compassion, and goal-setting exercises and included video and audio guidance, resulted in a significant reduction in perceived stress compared to an active control group. This study implemented a 4-week, easy-to-follow guided meditation program without the assistance of a facilitator, targeting Japanese workers, and similarly demonstrated an improvement in self-compassion. Few studies on psychological support for workers have reported intervention effects that extend beyond mental health [11]. This study is one of the few to show that even a short-term self-help intervention among workers can have positive effects on workplace-related outcomes, such as improved work performance, enhanced psychological safety, and reduced productivity loss.

This study used an RCT design to examine the effects of guided meditation practiced individually, using a rigorous approach that compared 2 intervention groups—SCM and MM groups—with a control group. However, when comparing the intervention groups with the control group, no significant interaction effects between intervention and time were observed. One possible reason for the lack of intervention effects relative to the control group is the low intensity of the intervention, which may have resulted in only modest effects. Digital interventions targeting anxiety and depressive symptoms are generally known to yield small to moderate effect sizes [49]. Additionally, the shorter 4-week duration of this fully self-guided intervention, relative to the typical 8-week format of established programs, may account for the limited between-group effects.

Spijkerman et al [50] also pointed out a limitation of psychological interventions in populations with relatively high baseline mental health, similar to participants in this study. Specifically, they noted that individuals without clinical levels of depression or anxiety tend to have lower baseline scores on psychological symptom measures, resulting in limited scope for improvement compared to clinical populations. This may result in smaller effect sizes due to floor effects. In particular, previous studies on online self-compassion interventions have reported small effects on anxiety symptoms and no significant effects on depressive symptoms when compared to active control groups [48].

Moreover, previous research has suggested that therapist-guided interventions are more effective for treating depression than fully self-guided programs [51]. Given that this study used a completely self-directed format without any therapist contact, participants may have found it difficult to independently acquire the necessary skills. This could partly account for the limited improvement, especially in depressive symptoms. In this intervention, loving-kindness meditation was implemented as a form of meditation that incorporates elements of self-compassion. Loving-kindness meditation is also introduced as an optional practice during the day-long retreat in the Mindfulness-Based Stress Reduction program and is recognized as part of the broader mindfulness framework [52]. This type of meditation is designed to cultivate feelings of warmth and benevolence toward oneself and others, including all beings.

Neff and Dahm [53] argued that self-compassion can be more effectively cultivated through explicit rather than implicit methods. The inclusion of loving-kindness meditation, which explicitly integrates components of self-compassion, may have contributed to the effects observed in this intervention over standard mindfulness practices. This may be particularly relevant for workers who are highly self-critical, as the practice of self-compassion could have helped improve their perceived work performance.

Despite the notable findings, this study has several limitations. Specifically, no significant differences were observed in the magnitude of change between the intervention and control groups from preintervention to postintervention. This limitation may be attributed to the relatively low intensity of our intervention program, which may have produced smaller effect sizes insufficient to demonstrate statistically significant group differences.

In terms of feasibility, the study demonstrated better retention and adherence than previous research reporting dropout rates exceeding 50% [54]. Several factors may have contributed to this, including a brief and clearly structured app-based intervention that lasted approximately 6 to 12 minutes per session and allowed participants to complete it at their convenience in terms of time and location. Additionally, weekly email reminders were used. Recruitment via an online crowdsourcing platform—where participants may have approached the intervention as a structured “task”—and the provision of compensation (which was not contingent on the number of meditation days) may have further encouraged engagement. We did not collect or report detailed app usage

XSL•FO
RenderX

analytics, such as time-of-day, changes in use intensity over time, or module-level engagement; consequently, total practice time could not be summarized. Given the 1-session-per-day design with a fixed session type and duration (except for interruptions), variability in dose was primarily reflected in the number of completed practice days. Future studies should prospectively capture richer log-based metrics, such as continuity patterns and total practice time, to better characterize feasibility and inform intervention optimization. To enhance intervention effectiveness while maintaining feasibility, future studies could extend the practice period beyond 4 weeks without changing daily practice duration or combine meditation with other intervention techniques (modules). Given the typical trade-off between intervention intensity and accessibility, future research should focus on developing interventions that remain accessible and user-friendly for working adults while improving therapeutic outcomes.

Separately, work performance was assessed using brief self-report measures (a single WHO-HPQ presenteeism item and the SPS work impairment score) to minimize participant burden in this fully self-guided digital trial. Nevertheless, single-item and self-reported measures may be less sensitive to change and may be susceptible to reporting biases, potentially diminishing our ability to detect effects. Future studies should consider incorporating more comprehensive and/or objective indicators of productivity, such as supervisor-rated performance or administrative and behavioral metrics, where feasible.

Furthermore, we could not confirm improvement in depressive symptoms, which was one of our main outcome measures for mental health. It has been noted that unresolved trauma memories may emerge during self-compassion practice, representing a potential risk of such training. For participants with trauma memories practicing at home, it is recommended to limit sessions to about 10 to 15 minutes per day [12]. Therefore, to reduce risks in our fully self-care intervention format, we implemented a low-intensity intervention that could be completed in a short time and targeted only workers with relatively high baseline health status. However, as previous research has indicated, preventive interventions targeting healthy individuals tend to show smaller effects [50]. Future research could incorporate risk reduction strategies, such as regular follow-ups by facilitators based on participants' circumstances. This approach would potentially allow participation from workers experiencing moderate stress levels, enabling investigations not only into prevention effects for relatively healthy workers but also into supportive effects for workers currently experiencing mental health difficulties.

Finally, the mechanisms by which the intervention improved work performance remain insufficiently understood. Currently, few studies have examined the detailed mechanisms of the relationship between self-compassion practice and work performance. In previous research on mindfulness, a similar psychological intervention, and work performance, mindfulness practice has been shown to enhance attention functions and positively affect major domains of human functioning, including cognition, emotion, behavior, and physiology, thereby improving work performance [10]. Good et al [10] suggest that improved cognitive function and cognitive flexibility may allow sustained attention to work even in environments where attention is depleted or work is performed discontinuously. Additionally, Kirk et al [55], in a study examining emotional regulation in meditation practitioners using economic decision-making games, reported that meditation practitioners activate different brain regions during emotional regulation compared to nonmeditators, suggesting the possibility of rational engagement with others even in emotional situations. Furthermore, higher levels of self-compassion are associated with reduced fear of failure and an increased tendency to try again after failure [56,57]. Based on these findings, future research could enhance understanding of the intervention's mechanisms by incorporating measures of cognitive function, brain function measurements, and psychological constructs related to failure and goal setting.

Conclusions

The intervention program was designed to enable nonclinical working adults to practice self-care meditation using a smartphone app without facilitator involvement. The study findings support the feasibility of delivering a brief, low-burden, fully self-guided meditation program incorporating self-compassion elements in occupational settings, with participants able to practice at times and locations of their choosing. However, given the limited between-group effects observed across most outcomes, the improvements noted should be considered preliminary and do not substantiate strong causal claims regarding the superiority of the intervention. Clinical significance remains unclear in this nonclinical sample and should be evaluated in future research. These findings inform the development of scalable, low-burden digital self-care tools for occupational mental health and highlight the need for further studies to evaluate clinical significance.

Acknowledgments

The authors would like to thank all participants in this study. The authors also thank Kei Minoura for his technical assistance with data collection via smartphone apps. The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: publication support. The GAI tool used was ChatGPT 5.2. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by TK.

XSL•FO
RenderX

Funding

This work was supported by the Japan Society for the Promotion of Science Grant-in-Aid for Scientific Research (grant JP16H05653, JP19K03278, 22H01091, 22K18582, 23K22362, and 25K21957 to RT, 24K22798 to TK), the Japan Science and Technology Agency (grant JPMJSP2108 for KA), the Royal Society and British Academy (grant AL150003 to RT), and University of Tokyo Social Cooperation Program “Fulfillment through Work” (to RT). The funders had no role in data collection or analysis, decision to publish, or preparation of the manuscript.

Data Availability

The datasets generated and analyzed during this study are not publicly available because data sharing in publicly available repositories was not included in the informed consent. However, the datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

TK and RT conceptualized the study. TK and KA performed the formal analysis, investigation, and visualization. KA was responsible for data curation and resources. RT secured funding for the project and provided supervision. TK managed the project administration and performed the validation. TK wrote the original draft, and RT contributed to the review and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Group × time interaction effect using linear mixed-effects models (LMMs).
[DOCX File, 26 KB - jmir_v28i1e79991_app1.docx]

Multimedia Appendix 2
Mean (SD) at follow-up and repeated-measures analyses across pre, post, and follow-up within the intervention groups.
[DOCX File, 30 KB - jmir_v28i1e79991_app2.docx]

Multimedia Appendix 3
Main effects for time and interaction effects for group × time using linear mixed-effects models (LMMs).
[DOCX File, 28 KB - jmir_v28i1e79991_app3.docx]

Checklist 1
CONSORT-EHEALTH checklist [58].
[PDF File, 453 KB - jmir_v28i1e79991_app4.pdf]

References

1. Mental health at work. World Health Organization. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/mental-health-at-work [accessed 2024-12-26]
2. Yoshimoto T, Oka H, Fujii T, Nagata T, Matsudaira K. The economic burden of lost productivity due to presenteeism caused by health conditions among workers in Japan. J Occup Environ Med 2020 Oct;62(10):883-888. [doi: 10.1097/JOM.0000000000002001] [Medline: 32826548]
3. Hayes SC. Acceptance and commitment therapy, relational frame theory, and the third wave of behavioral and cognitive therapies. Behav Ther 2004;35(4):639-665. [doi: 10.1016/S0005-7894(04)80013-3]
4. Creswell JD. Mindfulness interventions. Annu Rev Psychol 2017 Jan 3;68(1):491-516. [doi: 10.1146/annurev-psych-042716-051139] [Medline: 27687118]
5. Neff KD, Germer CK. A pilot study and randomized controlled trial of the mindful self-compassion program. J Clin Psychol 2013 Jan;69(1):28-44. [doi: 10.1002/jclp.21923] [Medline: 23070875]
6. Bégin C, Berthod J, Martinez LZ, Truchon M. Use of mobile apps and online programs of mindfulness and self-compassion training in workers: a scoping review. J Technol Behav Sci 2022;7(4):477-515. [doi: 10.1007/s41347-022-00267-1] [Medline: 36091081]
7. Michel A, Groß C, Hoppe A, González‐Morales MG, Steidle A, O’Shea D. Mindfulness and positive activities at work: intervention effects on motivation-related constructs, sleep quality, and fatigue. J Occupat & Organ Psyc 2021 Jun;94(2):309-337. [doi: 10.1111/joop.12345]

XSL•FO
RenderX

8. Janssen M, Heerkens Y, Kuijer W, van der Heijden B, Engels J. Effects of mindfulness-based stress reduction on employees' mental health: a systematic review. PLoS One 2018;13(1):e0191332. [doi: 10.1371/journal.pone.0191332] [Medline: 29364935]
9. Coo C, Salanova M. Mindfulness can make you happy-and-productive: a mindfulness controlled trial and its effects on happiness, work engagement and performance. J Happiness Stud 2018 Aug;19(6):1691-1711. [doi: 10.1007/s10902-017-9892-8]
10. Good DJ, Lyddy CJ, Glomb TM, et al. Contemplating mindfulness at work: an integrative review. J Manag 2016 Jan;42(1):114-142. [doi: 10.1177/0149206315617003]
11. Kotera Y, Van Gordon W. Effects of self-compassion training on work-related well-being: a systematic review. Front Psychol 2021;12:630798. [doi: 10.3389/fpsyg.2021.630798] [Medline: 33967896]
12. Germer C, Neff K. Teaching the Mindful Self-Compassion Program: A Guide for Professionals: Guilford Press; 2019.
13. Wasson RS, Barratt C, O'Brien WH. Effects of mindfulness-based interventions on self-compassion in health care professionals: a meta-analysis. Mindfulness (N Y) 2020;11(8):1914-1934. [doi: 10.1007/s12671-020-01342-5] [Medline: 32421083]
14. Inwood E, Ferrari M. Mechanisms of change in the relationship between self-compassion, emotion regulation, and mental health: a systematic review. Appl Psychol Health Well Being 2018 Jul;10(2):215-235. [doi: 10.1111/aphw.12127] [Medline: 29673093]
15. Wakelin KE, Perman G, Simonds LM. Effectiveness of self-compassion-related interventions for reducing self-criticism: a systematic review and meta-analysis. Clin Psychol Psychother 2022 Jan;29(1):1-25. [doi: 10.1002/cpp.2586] [Medline: 33749936]
16. Reizer A. Bringing self-kindness into the workplace: Exploring the mediating role of self-compassion in the associations between attachment and organizational outcomes. Front Psychol 2019;10:1148. [doi: 10.3389/fpsyg.2019.01148] [Medline: 31164857]
17. Wouters-Soomers L, Van Ruysseveldt J, Bos AER, Jacobs N. An individual perspective on psychological safety: the role of basic need satisfaction and self-compassion. Front Psychol 2022;13:920908. [doi: 10.3389/fpsyg.2022.920908] [Medline: 36059778]
18. Li Y, Hu Y, Yang W, Wang Y. Daily interventions and assessments: the effect of online self-compassion meditation on psychological health. Applied Psych Health & Well 2021 Nov;13(4):906-921. [doi: 10.1111/aphw.12278] [Medline: 33890708]
19. Mitsea E, Drigas A, Skianis C. Mindfulness for anxiety management and happiness: the role of VR, metacognition, and hormones. Technium BioChemMed 2022 Sep;3(3):37-52. [doi: 10.47577/biochemmed.v3i3.7343]
20. Linardon J. Can acceptance, mindfulness, and self-compassion be learned by smartphone apps? A systematic and meta-analytic review of randomized controlled trials. Behav Ther 2020 Jul;51(4):646-658. [doi: 10.1016/j.beth.2019.10.002] [Medline: 32586436]
21. Musiat P, Johnson C, Atkinson M, Wilksch S, Wade T. Impact of guidance on intervention adherence in computerised interventions for mental health problems: a meta-analysis. Psychol Med 2022 Jan;52(2):229-240. [doi: 10.1017/S0033291721004621] [Medline: 34802474]
22. Cameron G, Mulvenna M, Ennis E, et al. Effectiveness of digital mental health interventions in the workplace: umbrella review of systematic reviews. JMIR Ment Health 2025 Jan 24;12:e67785. [doi: 10.2196/67785] [Medline: 39854722]
23. Schwartz K, Ganster FM, Tran US. Mindfulness-based mobile apps and their impact on well-being in nonclinical populations: systematic review of randomized controlled trials. J Med Internet Res 2023 Aug 4;25:e44638. [doi: 10.2196/44638] [Medline: 37540550]
24. Psychologists embrace telehealth to prevent the spread of COVID-19: a survey of APA members reveals how the pandemic has impacted clinicians and their practices. American Psychological Association Services, INC. URL: https://www.apaservices.org/practice/legal/technology/psychologists-embrace-telehealth [accessed 2024-12-26]
25. Kurosawa T, Adachi K, Takizawa R. Mindful self-compassion smartphone intervention for worker mental health in Japan: protocol for a randomized controlled trial. JMIR Res Protoc 2024 Jul 15;13:e53541. [doi: 10.2196/53541] [Medline: 39008345]
26. Kessler RC, Andrews G, Colpe LJ, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002 Aug;32(6):959-976. [doi: 10.1017/s0033291702006074] [Medline: 12214795]
27. Furukawa TA, Kawakami N, Saitoh M, et al. The performance of the Japanese version of the K6 and K10 in the World Mental Health Survey Japan. Int J Methods Psychiatr Res 2008;17(3):152-158. [doi: 10.1002/mpr.257] [Medline: 18763695]
28. Sakurai K, Nishi A, Kondo K, Yanagida K, Kawakami N. Screening performance of K6/K10 and other screening instruments for mood and anxiety disorders in Japan. Psychiatry Clin Neurosci 2011 Aug;65(5):434-441. [doi: 10.1111/j.1440-1819.2011.02236.x] [Medline: 21851452]
29. Kessler RC, Barber C, Beck A, et al. The World Health Organization Health and Work Performance Questionnaire (HPQ). J Occup Environ Med 2003 Feb;45(2):156-174. [doi: 10.1097/01.jom.0000052967.43131.51] [Medline: 12625231]

XSL•FO
RenderX

30. Kawakami N, Inoue A, Tsuchiya M, et al. Construct validity and test-retest reliability of the World Mental Health Japan version of the World Health Organization Health and Work Performance Questionnaire short version: a preliminary study. Ind Health 2020 Aug 7;58(4):375-387. [doi: 10.2486/indhealth.2019-0090] [Medline: 32173661]
31. Koopman C, Pelletier KR, Murray JF, et al. Stanford Presenteeism Scale: health status and employee productivity. J Occup Environ Med 2002 Jan;44(1):14-20. [doi: 10.1097/00043764-200201000-00004] [Medline: 11802460]
32. Wada K, Moriyama M, Narai R, et al. The effect of chronic health conditions on work performance in Japanese companies. J Occup Health 2007 May;49(3):103-109. [doi: 10.1539/sangyoeisei.49.103] [Medline: 17575410]
33. Martin MM, Rubin RB. A new measure of cognitive flexibility. Psychol Rep 1995 Apr;76(2):623-626. [doi: 10.2466/pr0.1995.76.2.623]
34. Oshiro K, Nagaoka S, Shimizu E. Development and validation of the Japanese version of Cognitive Flexibility Scale. BMC Res Notes 2016 May 17;9(1):275. [doi: 10.1186/s13104-016-2070-y] [Medline: 27188498]
35. Neff KD. The development and validation of a scale to measure self-compassion. Self Identity 2003 Jul;2(3):223-250. [doi: 10.1080/15298860309027]
36. Arimitsu K. Development and validation of the Japanese version of the Self-Compassion Scale. Shinrigaku Kenkyu 2014 Apr;85(1):50-59. [doi: 10.4992/jjpsy.85.50] [Medline: 24804430]
37. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [doi: 10.2307/2136404] [Medline: 6668417]
38. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Routledge Academic; 1988.
39. Sumi K. Reliability and validity of the Japanese version of the Perceived Stress Scale. Jpn J Health Psychol 2006;19(2):44-53. [doi: 10.11560/jahp.19.2_44]
40. Schaufeli WB, Salanova M, González-romá V, Bakker AB. The measurement of engagement and burnout: a two sample confirmatory factor analytic approach. J Happiness Stud 2002 Mar;3(1):71-92. [doi: 10.1023/A:1015630930326]
41. Schaufeli WB, Bakker AB, Salanova M. The measurement of work engagement with a short questionnaire: a cross-national study. Educ Psychol Meas 2006 Aug;66(4):701-716. [doi: 10.1177/0013164405282471]
42. Shimazu A, Schaufeli WB, Kosugi S, et al. Work engagement in Japan: validation of the Japanese version of the Utrecht Work Engagement Scale. Appl Psychol 2008 Jul;57(3):510-523. [doi: 10.1111/j.1464-0597.2008.00333.x]
43. Edmondson A. Psychological safety and learning behavior in work teams. Adm Sci Q 1999 Jun;44(2):350-383. [doi: 10.2307/2666999]
44. Matsuo A, Tsujita M, Kita K, Ayaya S, Kumagaya S. Developing and validating Japanese versions of Psychological Safety Scale, Knowledge Sharing Scale and Expressed Humility Scale. Manag Labour Stud 2024 Aug;49(3):375-388. [doi: 10.1177/0258042X231191871]
45. George JM, Zhou J. When openness to experience and conscientiousness are related to creative behavior: an interactional approach. J Appl Psychol 2001 Jun;86(3):513-524. [doi: 10.1037/0021-9010.86.3.513] [Medline: 11419810]
46. Olejnik S, Algina J. Generalized eta and omega squared statistics: measures of effect size for some common research designs. Psychol Methods 2003 Dec;8(4):434-447. [doi: 10.1037/1082-989X.8.4.434] [Medline: 14664681]
47. Thabrew H, Boggiss AL, Lim D, et al. Well-being app to support young people during the COVID-19 pandemic: randomised controlled trial. BMJ Open 2022 May 19;12(5):e058144. [doi: 10.1136/bmjopen-2021-058144] [Medline: 35589362]
48. Beshai S, Bueno C, Yu M, Feeney JR, Pitariu A. Examining the effectiveness of an online program to cultivate mindfulness and self-compassion skills (Mind-OP): randomized controlled trial on Amazon's Mechanical Turk. Behav Res Ther 2020 Nov;134:103724. [doi: 10.1016/j.brat.2020.103724] [Medline: 32942203]
49. Rauschenberg C, Schick A, Hirjak D, et al. Evidence synthesis of digital interventions to mitigate the negative impact of the COVID-19 pandemic on public mental health: rapid meta-review. J Med Internet Res 2021 Mar 10;23(3):e23365. [doi: 10.2196/23365] [Medline: 33606657]
50. Spijkerman MPJ, Pots WTM, Bohlmeijer ET. Effectiveness of online mindfulness-based interventions in improving mental health: a review and meta-analysis of randomised controlled trials. Clin Psychol Rev 2016 Apr;45:102-114. [doi: 10.1016/j.cpr.2016.03.009] [Medline: 27111302]
51. Gellatly J, Bower P, Hennessy S, Richards D, Gilbody S, Lovell K. What makes self-help interventions effective in the management of depressive symptoms? Meta-analysis and meta-regression. Psychol Med 2007 Sep;37(9):1217-1228. [doi: 10.1017/S0033291707000062] [Medline: 17306044]
52. Santorelli S, Meleo-Meyer F, Koerbel L, Kabat-Zinn J. Mindfulness-Based Stress Reduction (MBSR) Authorized Curriculum Guide.: University of Massachusetts Medical School, Center for Mindfulness in Medicine, Health Care, and Society; 2017. URL: https://www.rightdecisions.scot.nhs.uk/m/2105/mbsr-curriculum-guide-2017.pdf [accessed 2025-06-12]
53. Neff KD, Dahm KA. Self-compassion: what it is, what it does, and how it relates to mindfulness. In: Ostafin BD, Robinson MD, Meier BP, editors. Handbook of Mindfulness and Self-Regulation: Springer; 2015:121-137.
54. Carvalho SA, Skvarc D, Barbosa R, Tavares T, Santos D, Trindade IA. A pilot randomized controlled trial of online acceptance and commitment therapy versus compassion-focused therapy for chronic illness. Clin Psychol Psychother 2022 Mar;29(2):524-541. [doi: 10.1002/cpp.2643] [Medline: 34269493]
55. Kirk U, Downar J, Montague PR. Interoception drives increased rational decision-making in meditators playing the ultimatum game. Front Neurosci 2011;5:49. [doi: 10.3389/fnins.2011.00049] [Medline: 21559066]

XSL•FO
RenderX

56. Neff KD, Hsieh YP, Dejitterat K. Self-compassion, achievement goals, and coping with academic failure. Self Identity 2005 Jul;4(3):263-287. [doi: 10.1080/13576500444000317]
57. Neely ME, Schallert DL, Mohammed SS, Roberts RM, Chen YJ. Self-kindness when facing stress: the role of self-compassion, goal regulation, and support in college students' well-being. Motiv Emot 2009 Mar;33(1):88-97. [doi: 10.1007/s11031-008-9119-8]
58. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: Improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011;13(4):e126. [doi: 10.2196/jmir.1923]

Abbreviations

CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
LMM: linear mixed-effects model
MM: mindfulness meditation
RCT: randomized controlled trial
SCM: self-compassion meditation
SCS: Self-Compassion Scale
SPS: Stanford Presenteeism Scale
WCG: waitlist control group
WHO: World Health Organization
WHO-HPQ: World Health Organization Health and Work Performance Questionnaire

Edited by A Stone; submitted 05.Jul.2025; peer-reviewed by A Perez-Aranda, B Olaya, N Sripo; revised version received 08.Feb.2026; accepted 10.Feb.2026; published 17.Mar.2026.

Please cite as:
Kurosawa T, Adachi K, Takizawa R
Effects of Self-Compassion and Mindfulness Interventions on Mental Health and Work-Related Outcomes Among Japanese Workers: Randomized Controlled Trial
J Med Internet Res 2026;28:e79991
URL: https://www.jmir.org/2026/1/e79991
doi:10.2196/79991

© Takumu Kurosawa, Koichiro Adachi, Ryu Takizawa. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effects of Internet-Based Dementia Risk Reduction Education on Risk and Protective Factor Knowledge, Intentions, and Health Behaviors: Randomized Controlled Trial

Anthony J Levinson[1,2*], BA, MA, MSc, MD; Stephanie Ayers[1*], BA, BEd; Sandra Clark[1*], BSc, MSc; Rebekah Woodburn[1*], BA; Amy Schneeberg[1], PhD; Dima Hadid[1], MPH; Nick Kates[1], MD; Gillian Strudwick[3], RN, PhD; Roland Grad[4], MSc, MD; Alexandra Papaioannou[1,2], BScN, MSc, MD; Maureen Dobbins[1,5], RN, PhD; Henry Siu[1], MSc, MD; Dante Duarte[1,2], MD, PhD; Karen Saperson[1], MBChB; Sharon Marr[6], BSc, MEd, MD; Doug Oliver[1], BSc, MSc, MD; Sarah Neil-Sztramko[1,7], PhD

[1]Faculty of Health Sciences, McMaster University, 1280 Main Street West, Hamilton, ON, Canada

[2]Faculty of Health Sciences, Geras Centre for Aging Research, 88 Maplewood Avenue, Hamilton, ON, Canada

[3]Centre for Addiction and Mental Health, Toronto, Canada

[4]Department of Family Medicine, McGill University, Montreal, Canada

[5]Faculty of Health Sciences, National Collaborating Centre for Methods and Tools, Hamilton, Canada

[6]Department of Medicine, Unity Health Toronto, Toronto, Canada

[7]Faculty of Health Sciences, National Collaborating Centre for Methods and Tools, Hamilton, ON, Canada

[*]these authors contributed equally

Corresponding Author:
Anthony J Levinson, BA, MA, MSc, MD
Faculty of Health Sciences, McMaster University, 1280 Main Street West, Hamilton, ON, Canada

Abstract

Background: Dementia prevention through the reduction of modifiable risk factors is gaining attention as a public health strategy. However, public knowledge of dementia risk and protective factors remains low. Web-based education offers a potential solution to raise awareness and promote risk-reduction behaviors.

Objective: This randomized controlled trial evaluated the effectiveness of DementiaRisk.ca, an internet-based multimedia educational intervention, in increasing knowledge of dementia risk factors, intentions to engage in risk reduction behaviors, and changes in health behaviors.

Methods: A 2-arm randomized controlled trial was conducted with 510 participants (265 in the intervention group and 245 in the control group). Participants were randomized to receive either the e-learning about dementia risk and promoting brain health, which included a multimedia lesson and microlearning emails, or a control intervention focused on mild cognitive impairment. Outcomes included knowledge of dementia risk factors, intentions to engage in risk reduction, and health behaviors, measured at baseline (T1), 4 weeks (T2), and 2 months postintervention (T3). Outcomes were analyzed using linear mixed effects models with fixed effects for group, time, and their interaction, and a random intercept for participants.

Results: Of the 510 randomized participants, 405 (79.4%) completed all intervention components. Participants were predominantly female (n=309, 60.6%) and aged 55 years or older (n=284, 55.7%). Baseline mean dementia knowledge scores were 17.0 (SD 5.5) in the intervention group and 17.4 (SD 6.0) in the control group. At T2, scores increased to 25.8 (SD 4.5) and 23.6 (SD 5.1), respectively, yielding a between-group difference of 2.2 points (95% CI 1.2 - 3.2; *P*<.001), which was sustained at T3. Both groups showed significant improvements in knowledge, intentions, and health behaviors over time, with larger knowledge gains in the intervention group and particularly among participants with lower educational attainment. Intentions to engage in dementia risk reduction improved in both groups at T2 (intervention: +1.0, 95% CI 0.2 - 1.8; control: +1.4, 95% CI 0.5 - 2.3), with no significant between-group difference. Self-reported physical activity increased from 31.7 (SD 25.0) to 38.6 (SD 27.5) in the intervention group and from 29.9 (SD 23.5) to 32.5 (SD 26.6) in the control group, with a between-group difference of 5.4 points at T2 (95% CI 0.3 - 10.5; *P*=.04). No significant between-group differences were observed for diet, alcohol use, or other health behaviors. Qualitative findings indicated that participants valued the intervention for improving awareness of dementia risk factors, motivating proactive lifestyle changes, and enhancing confidence in applying prevention information.

Conclusions: This internet-based dementia risk reduction e-learning program improved dementia-related knowledge and increased self-reported physical activity, particularly among participants with lower educational attainment. Effects on other

XSL•FO
RenderX

health behaviors were limited. These findings support the use of well-designed e-learning as a scalable public health strategy to strengthen dementia risk reduction literacy and encourage selected healthy behaviors.

Trial Registration: ClinicalTrials.gov NCT05383118; https://clinicaltrials.gov/study/NCT05383118

International Registered Report Identifier (IRRID): RR2-10.2196/64718

(J Med Internet Res 2026;28:e79405) doi:10.2196/79405

KEYWORDS

web-based intervention; internet; eHealth; dementia risk; dementia prevention; Alzheimer disease; education and training; clinical trial; public health; digital health interventions

Introduction

Background

As our population ages, the prevalence of dementia is expected to continue to rise. Globally, dementia affects over 55 million people, with annual costs to health care, long-term care, and informal care partners exceeding US $1.3 trillion [1-4]. Without effective prevention strategies, the number of dementia cases is projected to triple by 2050, representing a significant global health and economic burden. In Canada, it is estimated that the number of people living with dementia will triple over the next 30 years [4]. However, increasing research highlights that there are many actions individuals can take to reduce their risk of dementia. According to the recent Lancet Commission report, up to 45% of dementias could be prevented through a variety of modifiable factors [4]. Similar estimates have been reported from the Canadian Longitudinal Study on Aging, for the impact of 12 risk factors on dementia cases, with physical inactivity, depression, and hypertension being particularly important in the Canadian context [5]. Other longitudinal studies have found that engaging in up to 4 - 5 healthy lifestyle behaviors may reduce the risk of developing Alzheimer dementia by 37% to 60% [6,7]. The Landmark Study, which developed evidence-based data modeling to demonstrate the impacts of improving risk reduction efforts to delay onset across the population, suggests that a delay of developing dementia by as little as one year could result in nearly 500,000 fewer cases of dementia by 2050 [8]. Additionally, if risk reduction efforts can delay the onset of dementia by 10 years, over 4 million new cases of dementia could be avoided by 2050 [8]. Initiatives to help prevent dementia may also prove to be cost-effective at the societal level while simultaneously allowing individuals to help themselves and for care partners to support their loved ones [9].

Gaps in Knowledge

Despite growing evidence about modifiable risk factors, public understanding of dementia prevention remains limited. A national public opinion survey conducted by EKOS Research Associates Inc for the Public Health Agency of Canada revealed that only 37% of Canadians recognized chronic conditions such as hypertension, heart disease, and diabetes as risk factors for dementia [10]. Awareness of other significant factors, such as hearing loss, was even lower, with only over 10% identifying it as a risk. Meanwhile, misconceptions, such as overestimating the role of toxic chemicals, for which there is limited evidence, persist [10].

Educational Interventions

Educational interventions have the potential to improve knowledge of dementia risk and promote health behavior change [11,12]. While much of the existing literature has focused on training health care providers, a smaller number of studies have examined public education initiatives, which typically involve synchronous, one-on-one delivery by professionals. Most existing resources are predominantly text-based and often lack interactivity or adherence to principles of instructional design [13].

Web-based education offers a promising alternative due to its scalability, flexibility, and accessibility. Evidence from other health domains shows that e-learning tools that incorporate multimedia elements, such as audio narration, visuals, and interactive quizzes, are more engaging and improve long-term knowledge retention compared to static, text-only formats [14-16]. However, most of the web-based content about modifying dementia risk factors is text-based, such as static web pages or online PDF pamphlets or booklets, often requires a substantial time commitment, is frequently associated with a commercial entity or product for sale, or requires real-time participation [14,17,18]. Moreover, few, if any, evidence-based multimedia interventions on this topic have been rigorously studied using randomized controlled trials (RCTs), particularly those freely accessible to the public in multiple languages.

Objectives

The objectives of this study were to: (1) evaluate whether exposure to an e-learning intervention (DementiaRisk.ca) changes knowledge of dementia risk factors, intention to engage in risk reduction activities, and health behaviors related to dementia risk reduction. The primary outcome was the change in these variables from baseline (T1) to 4 weeks postintervention (T2), with secondary analyses examining whether changes were sustained at 2 months postintervention (T3); and (2) explore qualitative aspects such as participants' engagement and satisfaction with the intervention, as well as barriers and facilitators to use and dissemination of the information provided by the intervention, in adults without cognitive impairment.

Methods

Study Design

This study used a sequential explanatory mixed methods design, consisting of a 2-arm RCT, followed by a qualitative thematic content analysis to explore the findings from the RCT in greater depth. Full details of the methods have been published [19]. No

XSL•FO
RenderX

changes were made from the published protocol. A mixed methods design was chosen to incorporate both quantitative and qualitative analysis, which allowed for a more comprehensive and nuanced understanding of the relationship between the intervention and its impacts.

Participants

Participants were recruited from March 2023 to April 2023 using AskingCanadians, a paid panel service to allow for the recruitment of a representative and diverse sample of Canadians. Eligible participants were English-speaking adults with access to email and the internet, aged 18 years or older who had never been diagnosed with dementia. Participants were provided with the study information and eligibility criteria through an online link. After meeting eligibility criteria, participants were redirected to our online research study platform to complete informed consent and baseline measures. Informed consent was obtained digitally, with participants required to review an online information page before consenting to participate. We attempted to recruit participants across a range of ages, as risk factors may impact people throughout their life course, and many risk factors often begin to accumulate in midlife, many years before cognitive signs or symptoms emerge [4]. Ethical approval was provided by the Hamilton Integrated Research Ethics Board (project ID 14886).

Study Procedures

Following completion of online informed consent, participants were randomized and directed to their assigned intervention or control group learning path. Participants were randomized in a 1:1 ratio via a computer-generated sequence through our research platform using a permuted block stratified randomization, with education and age as the stratification variables.

To the best of our ability, efforts were made to blind participants to the study hypothesis and allocation group by referring to the study in broad terms, such as “e-learning about cognitive impairment and promoting brain health” (rather than explicitly “dementia prevention,” for example). The principal investigator (AJL) and data analysts (DH and AS) remained blinded to group allocation throughout the data collection and analysis phases.

Intervention

The intervention group received e-learning about dementia risk reduction and promoting brain health consisting of the following components: (1) one 30-minute multimedia e-learning lesson (DementiaRisk.ca) about dementia risk and protective factors (Figure 1), and (2) a series of 12 prescheduled microlearning emails. The emails included small, focused segments of content taken directly from the e-learning lesson designed to reinforce key information from the e-learning lesson through spaced repetition.

Participants had 4 weeks to complete the e-learning lesson and received 3 emails per week during that period.

Content for the intervention was based predominantly on the World Health Organization (WHO) guidelines on risk reduction for cognitive decline and dementia and the Lancet Commission’s 2020 report on dementia prevention, intervention, and care [4,20]. Instructional design was guided by the Mayer Cognitive Theory of Multimedia Learning, ensuring that materials were delivered in manageable segments, using a combination of visuals, audio narration, and other evidence-based principles essential to effective e-learning that have been shown to enhance knowledge retention [13,14,16,17].

Other theoretical frameworks also informed the overall design of the intervention and its evaluation. The theory of planned behavior (TPB) [21,22] was used due to its extensive application with a number of dementia prevention risk factors (eg, smoking, diet, alcohol consumption, and physical activity) [23-27]. TPB suggests that intentions to engage in or change specific behaviors depend on motivation and ability across 6 constructs that collectively represent a person’s actual control over their behavior. DementiaRisk aimed to change participants’ attitudes and behavioral intentions by increasing knowledge related to dementia prevention and modifiable risk factors. The intervention and its evaluation were also informed by the Acquisition Cognition Application–Level of Outcomes model, which helps explain the value of the information from the user perspective, looking at the domains of situational relevance, cognitive impact, use of the information, and if used, any perceived health benefits [28].

XSL•FO
RenderX

Figure 1. Screenshot of the DementiaRisk e-learning intervention.

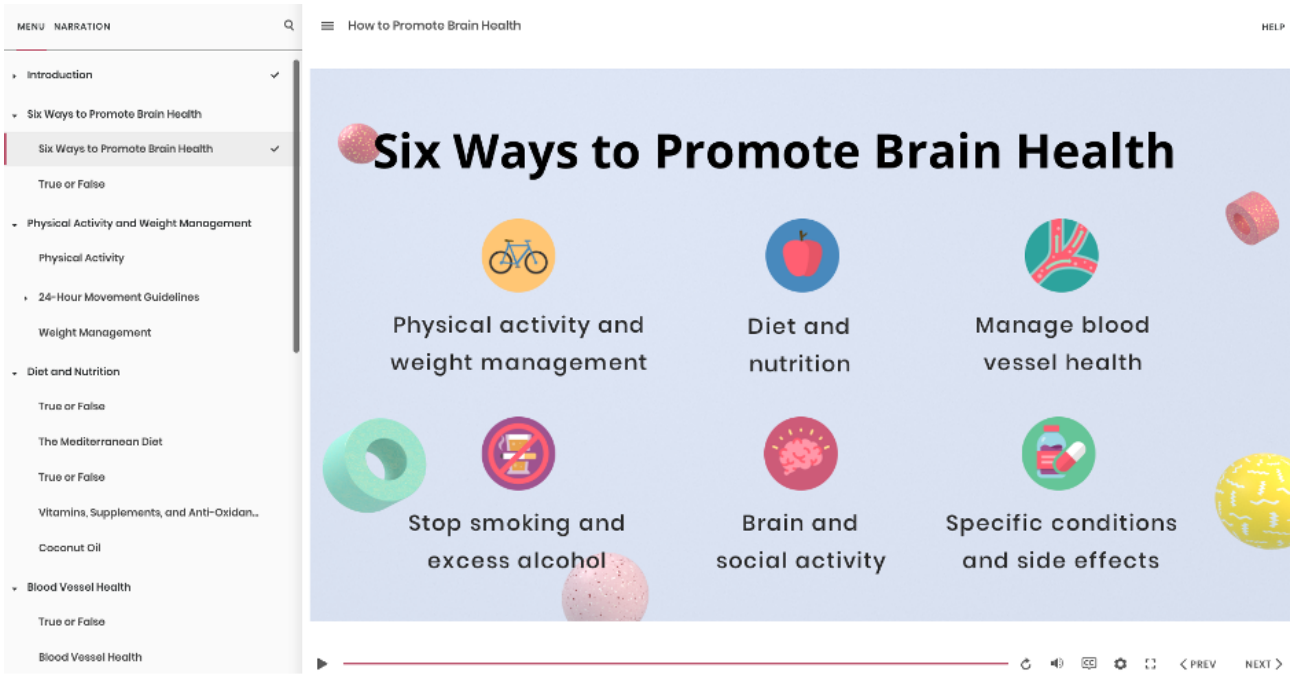

Control Condition

The control group participants received the following components: (1) 30-minute lesson on mild cognitive impairment (Figure 2), and (2) a series of 12 microlearning emails with small segments of content to reinforce material from the lesson. The design of the control group e-learning was also informed by similar principles and theoretical frameworks.

Figure 2. Screenshot from the control group e-learning.

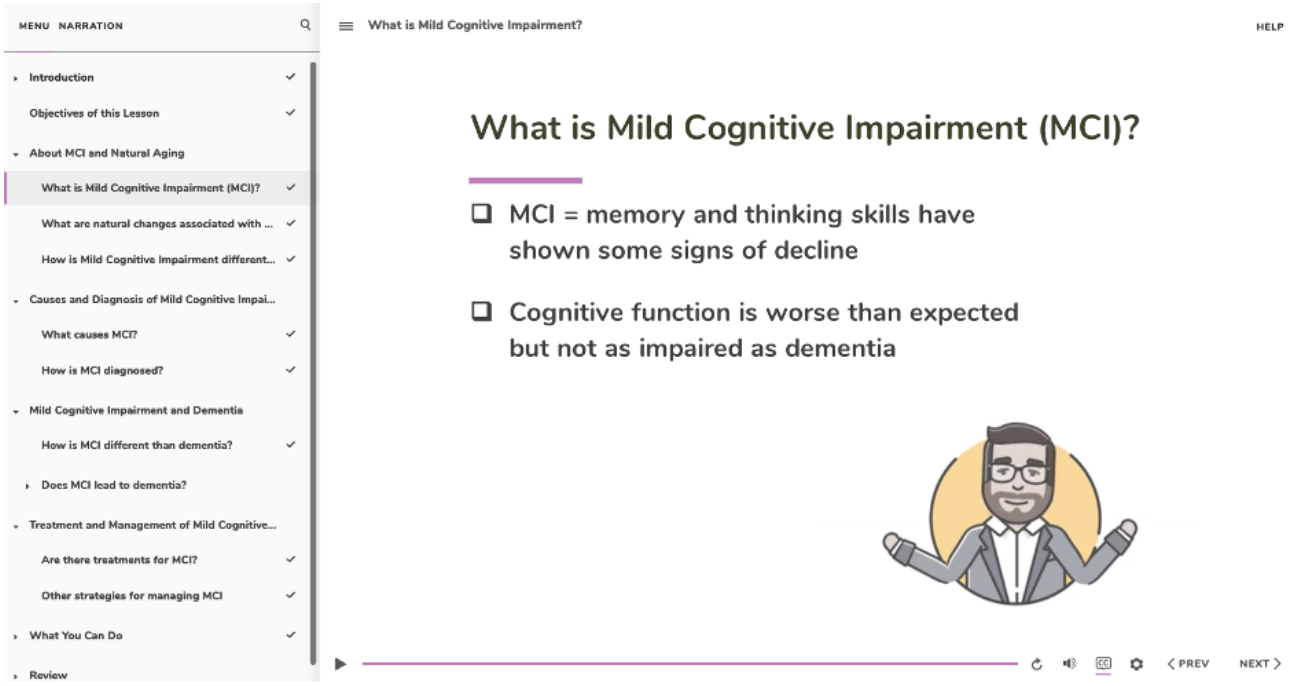

Outcome Measures

Quantitative data were collected through online questionnaires at baseline (T1), 4 weeks (T2), and at 2 months postintervention (T3). Knowledge of dementia risk factors was collected using a 15-item multiple choice assessment aligned with the content of the intervention as well as the Risks and Health Promotion subdomain of the Dementia Knowledge Assessment Scale [29]. The total maximum score for this assessment is 30 (maximum 2 points per question), with higher scores indicating higher levels of knowledge related to dementia risk factors. Intentions to engage in health behaviors were measured using a 10-item assessment. The total maximum score for this assessment is 60, with a higher score indicating higher levels of intention to engage. Current participant health behaviors related to dementia risk factors were assessed using the Godin-Shephard Leisure-Time Physical Activity Questionnaire for physical activity [30] along with additional questions related to diet,

XSL•FO
RenderX

smoking, alcohol consumption, social activity, traumatic brain injury, blood pressure, depression, air pollution, diabetes, sleep habits, brain activity, and hearing. Typical interpretation of total leisure activity scores uses a range of "insufficiently active or sedentary" (less than 14 points), "moderately active" (14 - 23 points), and "active" (24 points or more).

Data related to engagement with the e-learning programs were collected during the active study phase to gain a clearer understanding of how the use of web-based learning throughout the study (or fidelity of delivery) might have impacted changes in knowledge, intentions, or health behaviors. The following metrics were used: number of e-learning starts and completions, e-learning completion rates, time spent on lessons, and microlearning email open rates.

All data were collected through secure password-protected online platforms, which automatically recorded responses and stored data. Participant responses were anonymized prior to analysis. Missing data were handled using linear mixed effects models, which incorporate all available data under the assumption of missing at random (MAR). Given the structure of the data and the analytic approach, multiple imputation was not performed, as there is evidence that linear mixed models (LMMs) provide unbiased estimates without imputation when using full information maximum likelihood estimation [31].

Qualitative Data Collection

Qualitative data were collected from participants at T2 through both structured and open-ended questions using the Information Assessment Method for Patients and Consumers (IAM4all), a validated questionnaire that assesses outcomes of online consumer health information [19,32]. Qualitative questions explored participant satisfaction with the intervention, changes in attitudes, beliefs, and behaviors during the study, and feedback on proposed future dissemination methods. Data collection continued until no new themes emerged.

Data Analysis

Quantitative Data

RStudio (version 4.4.1; R Foundation) was used for quantitative data analysis [33-36]. LMMs were used to examine the relationship between group, time point, and each of the 3 outcomes.

All models included outcome scores from all 3 time points: baseline (T1), 4 weeks (T2), and 2 months post-intervention (T3) as the dependent variable. An interaction between time, group, and participant ID was included as a random intercept. Random slopes were not included due to limited time points and sample size constraints. The point estimates and 95% CIs comparing groups at T2 were used to address the primary objective. Exploratory analysis of whether changes observed at T2 were sustained at T3 was also addressed through the inclusion of the group by time point interaction.

Sex, age group (<45, 45 - 64, and ≥65 y), and education level (high school equivalent or less, some college or university, and college or university or graduate degree) were included as covariates in all models. Additionally, exploratory subgroup analyses of these were performed through testing interaction terms of group assignment and the stratification terms, as well as between time point and stratification terms. Nested models were compared using likelihood ratio tests (P<.05). For knowledge score only, having a family history of dementia or having been a care partner was combined into a single variable and tested as a potential effect modifier.

Missing data were explored across all 3 outcome variables: knowledge of dementia risk factors, intentions to engage in dementia risk reduction behaviors, and intentions to access dementia risk reduction information. For each outcome, participants with missing versus observed scores were compared at each timepoint across key demographic and exposure variables. Descriptive summaries and formal comparisons using chi-square or Fisher exact tests were conducted to assess potential differential attrition. Linear mixed effects models were used to incorporate all available data under the assumption of MAR, and multiple imputation was not performed.

All models were run for both intention-to-treat (ITT) and per-protocol (PP) group allotment. Effect sizes for group comparisons were calculated using the *emmeans* package in RStudio version 2023.12.1.402 [33-36]. Data analysts were blinded to participant grouping.

Using a conservative estimate of a small effect size (0.16, from a meta-analysis of internet health behavior change interventions), with a power of 0.80 and α of 0.05, we aimed to recruit a total of 388 participants in the study [37,38]. The sample size was increased by 25% to account for anticipated dropout, which brought the target recruitment number to 485 individuals [37,38]. These recruitment numbers and strategies have been used successfully in similar knowledge translation intervention studies and were a feasible target given the use of the paid panel [39].

Qualitative Data

For qualitative data, a conventional inductive content analysis approach was used [40]. All qualitative data were entered into NVivo (version 14; Lumivero) [41,42]. Data were systematically examined [19]. Codes and themes were developed based on the content, with no reliance on preconceived theories. The primary analyst (DH) conducted the initial coding and generated preliminary themes to ensure consistency and depth in the interpretation of the data. To strengthen rigor and minimize bias, the primary coder (DH) met weekly with the principal investigator (AJL) and other members of the research team to discuss developing codes and themes, confirm interpretations, and ensure analytic dependability. This approach has been shown to enhance analytic rigor through peer debriefing and triangulation [31]. The analysis focused on each theme separately. An iterative process of confirmation and reflection was used to analyze the data collaboratively. An audit trail was maintained, including memos from each stage of data collection and analysis. Findings of themes were triangulated between analysts. We ensured the study's trustworthiness by considering validity standards that ensure rigor, confirmability, credibility, dependability, and transferability.

XSL•FO
RenderX

Data Integration

Although the quantitative and qualitative components were analyzed separately, integration occurred during interpretation. The quantitative findings demonstrated changes in dementia risk–related knowledge, behavioral intentions, and selected health behaviors over time, while the qualitative findings helped contextualize these outcomes by elucidating participants' perceptions of the intervention's relevance, understandability, and anticipated benefits. In particular, qualitative data provided insight into how increased awareness translated into motivation for short-term lifestyle changes and highlighted perceived facilitators and barriers to applying dementia risk–reduction information. This sequential explanatory mixed methods approach enabled a more comprehensive understanding of the intervention's impact than either quantitative or qualitative findings alone.

Ethical Considerations

This study gained Hamilton Integrated Research Ethics Board approval on August 24, 2022 (project ID 14886). Participants completed all study components entirely online. Participants were required to provide informed consent and were informed of the length of time of the e-learning, surveys, and email campaign as well as informed about details surrounding data collection, storage, and investigator identities. Participants' identities and confidentiality were maintained throughout the research study. All participant data were deidentified, and all findings will be nonidentifiable. There was no known risk or harm to participating in this study or publicizing its results or findings. Participants were provided with AskingCanadians points that could be redeemed for various gift cards as compensation for participation in the study.

Results

Participant Characteristics

A total of 1319 individuals responded to the initial online recruitment message and 879 eligible participants were invited to complete informed consent. 534 individuals consented and were randomly assigned to the intervention (n=272) or control group (n=262). A total of 510 participants began the study (intervention n=265; control n=245). Retention at T3 was 79.4% (intervention n=211; control n=194). Participant flow is detailed in Figure 3. Participants were able to access all assessments at various time points, regardless of whether they had completed previous components and were therefore not considered lost to follow-up (see Checklist 1).

XSL•FO
RenderX

Figure 3. CONSORT (Consolidated Standards of Reporting Trials) study flow diagram. Note that participants were able to access all assessments at various time points, regardless of whether or not they completed previous components and were therefore not considered lost to follow-up. MCI: mild cognitive impairment.

XSL•FO
RenderX

Baseline characteristics were similar across all groups (all *P*>.05). Table 1 presents the demographic details of participants in each group. Participants were predominantly older (55.7%, aged ≥55 y), female (60.6%), and in good, very good, or excellent health (81.2%). The majority of participants identified as Caucasian (55.3%) followed by Chinese (10.4%) and South Asian or multiracial (each at 6.1%). Distribution patterns were relatively consistent across groups, though the control group had a slightly higher proportion of multiracial and Indigenous or First Nations participants. When asked about a family history of dementia, 31.4% of participants reported having a family history, 51.6% reported no family history, and 16.1% were unsure. The responses were consistent between groups. Finally, approximately one-quarter of participants (25.1%) identified as being a dementia care partner, with a higher proportion in the intervention group (27.9%) compared to the control group (22%).

XSL•FO
RenderX

Table . Participant demographics (N=510).

Characteristics	Intervention (n=265), n (%)	Control (n=245), n (%)	Overall (N=510), n (%)
Sex			
Male	107 (40.4)	89 (36.3)	196 (38.4)
Female	155 (58.5)	154 (62.9)	309 (60.6)
Missing	3 (1.1)	2 (0.8)	5 (1)
Age (y)			
18-24	1 (0.4)	3 (1.2)	4 (0.8)
25-34	26 (9.8)	23 (9.4)	49 (9.6)
35-44	41 (15.5)	37 (15.1)	78 (15.3)
45-54	51 (19.2)	44 (18)	95 (18.6)
55-64	53 (20)	58 (23.7)	111 (21.8)
65-74	59 (22.3)	60 (24.5)	119 (23.3)
≥75	31 (11.7)	18 (7.3)	49 (9.6)
Missing	3 (1.1)	2 (0.8)	5 (1)
Race and ethnicity			
Asian	7 (2.6)	10 (4.1)	17 (3.3)
Asian, East Asian	1 (0.4)	2 (0.8)	3 (0.6)
Black or African	8 (3)	6 (2.4)	14 (2.7)
Canadian	11 (4.2)	11 (4.5)	22 (4.3)
Caribbean	0 (0)	1 (0.4)	1 (0.2)
Chinese	28 (10.6)	25 (10.2)	53 (10.4)
East Asian	1 (0.4)	1 (0.4)	2 (0.4)
Eastern European	1 (0.4)	0 (0)	1 (0.2)
European	4 (1.5)	3 (1.2)	7 (1.4)
Filipino	1 (0.4)	2 (0.8)	3 (0.6)
Indigenous or First Nations	1 (0.4)	6 (2.4)	7 (1.4)
Jewish	3 (1.1)	3 (1.2)	6 (1.2)
Latino	2 (0.8)	0 (0)	2 (0.4)
Middle Eastern	1 (0.4)	0 (0)	1 (0.2)
Multiracial	11 (4.2)	20 (8.2)	31 (6.1)
South Asian	20 (7.5)	11 (4.5)	31 (6.1)
Western European	12 (4.5)	10 (4.1)	22 (4.3)
White or Caucasian	150 (56.6)	132 (53.9)	282 (55.3)
Missing	3 (1.1)	2 (0.8)	5 (1)
Health status			
Excellent	16 (6)	14 (5.7)	30 (5.9)
Very good	83 (31.3)	68 (27.8)	151 (29.6)
Good	124 (46.8)	109 (44.5)	233 (45.7)
Fair	38 (14.3)	48 (19.6)	86 (16.9)
Poor	1 (0.4)	4 (1.6)	5 (1)
Missing	3 (1.1)	2 (0.8)	5 (1)
Family history of dementia			
I am not sure	44 (16.6)	38 (15.5)	82 (16.1)

XSL•FO
RenderX

Characteristics	Intervention (n=265), n (%)	Control (n=245), n (%)	Overall (N=510), n (%)
No	136 (51.3)	127 (51.8)	263 (51.6)
Yes	82 (30.9)	78 (31.8)	160 (31.4)
Missing	3 (1.1)	2 (0.8)	5 (1)
Dementia care partner			
No	188 (70.9)	189 (77.1)	377 (73.9)
Yes	74 (27.9)	54 (22)	128 (25.1)
Missing	3 (1.1)	2 (0.8)	5 (1)

e-Learning Program Engagement

Metrics related to participant engagement with the e-learning programs are presented in Table 2. Overall, participants had a high e-learning lesson completion rate of 93% for both the intervention and control groups. Participants spent roughly the expected time on the e-learning lessons: an average of 31 minutes and 52 seconds (SD 17 minutes and 47 seconds) on the intervention lesson (estimated time to complete was 30 - 35 min), and an average of 26 minutes and 11 seconds (SD 10 minutes 37 seconds) on the control lesson (estimated time to complete was 25 - 30 min). The microlearning email campaign open rates were also strong (80.5% in the intervention and 78.6% in the control).

Table . e-Learning metrics.

e-Learning metric	Intervention (n=265)	Control (n=245)
Lesson starts	235	223
Lesson completions	218	207
Lesson completion rate (%)[a]	93	93
Average time spent on lesson	31 min, 52 s[b]	26 min 11 s
Micro-learning email open rate (%)	80.5	78.6

[a]Lesson completion rate is defined as lesson completions divided by lesson starts.

[b]Participants who spent over 1.5 hours on the module (n=20) who may have left their computer on after completing the lesson were removed from this average.

Changes in Knowledge, Intentions, and Health Behaviors

The results of the linear mixed effects model revealed no significant difference between groups in baseline knowledge, intentions to engage in health behaviors, or health behaviors' scores. The results of all models can be found in Multimedia Appendix 1.

Knowledge of Dementia Risk Factors

Knowledge scores did not differ between groups at baseline (estimate −0.39, 95% CI −1.29 to 0.51; *P*=.40). Both groups improved significantly from T1 to T2, but the intervention group showed a statistically significantly greater increase. The difference at T2 was 2.66 points in favor of the intervention (95% CI 1.70 - 3.63; *P*<.001). Knowledge gains were maintained at T3, with the intervention group continuing to exhibit significantly greater gain (estimate 1.60, 95% CI 0.65 - 2.55; *P*=.001). Crude scores are shown in Table 3 and marginal means in Figure 4.

Table . Crude knowledge scores across time points by group (N=510).

	Intervention (n=265)			Control (n=245)			Overall (N=510)		
	T1	T2	T3	T1	T2	T3	T1	T2	T3
Mean (SD)	17.0 (5.51)	25.8 (4.54)	25.6 (4.68)	17.4 (5.98)	23.6 (5.13)	24.2 (5.06)	17.2 (5.74)	24.8 (4.95)	24.9 (4.91)
Median (IQR)	17.0 (14.0-20.0)	28.0 (24.0-29.0)	27.0 (23.0-29.0)	17.0 (13.0-22.0)	25.0 (21.0-28.0)	25.0 (21.0-28.0)	17.0 (12.0-21.0)	26.0 (22.0-28.0)	26.0 (22.0-29.0)
Missing, n (%)	3 (1.1)	56 (21.1)	54 (20.4)	3 (1.2)	57 (23.3)	51 (20.8)	6 (1.2)	113 (22.2)	105 (20.6)

XSL•FO
RenderX

Figure 4. Marginal means from linear mixed model (LMM) with 95% CI for knowledge score by: (A) group, (B) level of education, and (C) dementia history or care partner.

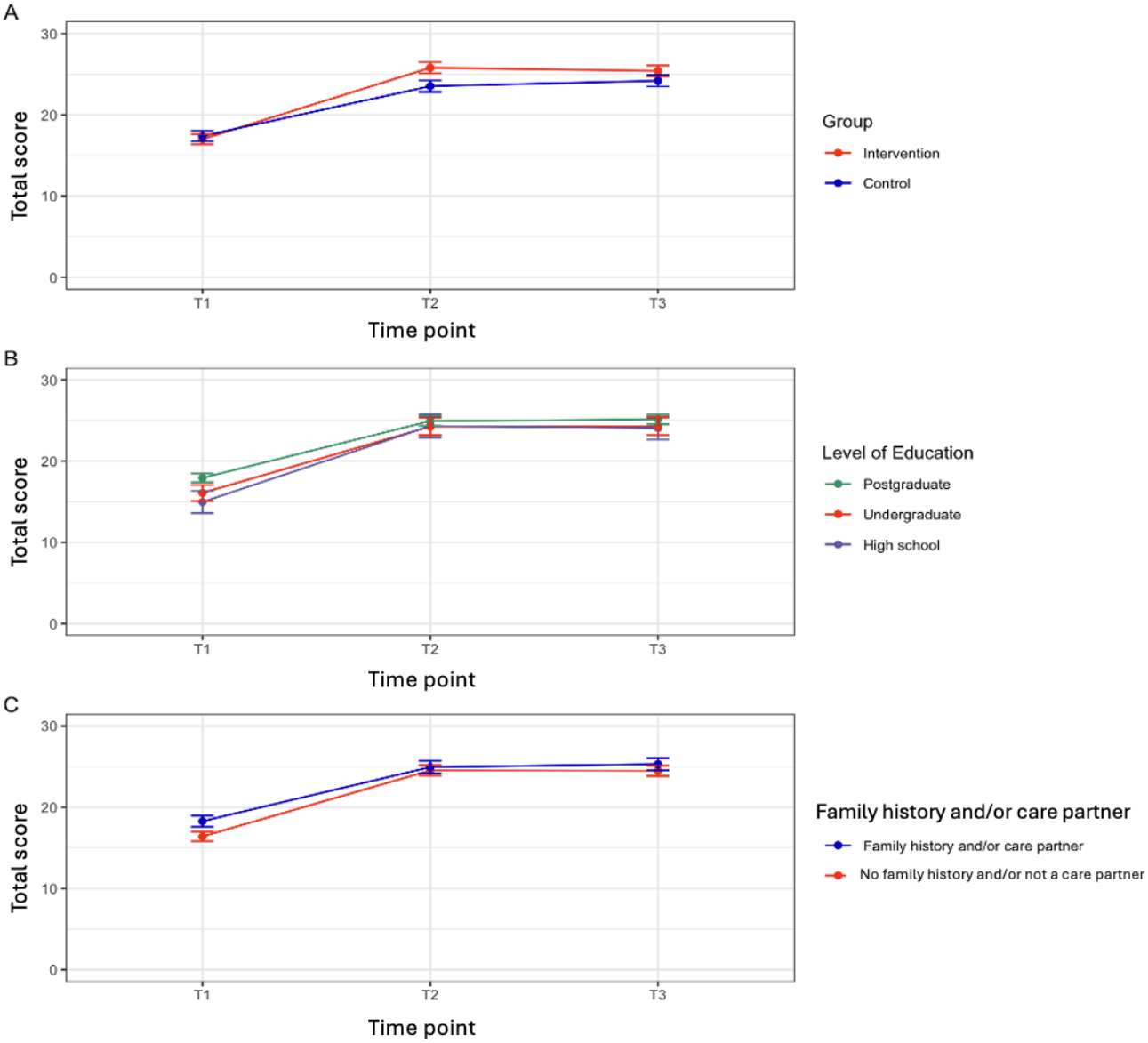

Subgroup analysis indicated that participants with lower levels of education achieved the greatest gains at both T2 (estimate 2.36, 95% CI 0.83 - 3.88; *P*=.003) and T3 (estimate 1.90, 95% CI 0.41 - 3.39; *P*=.01). Those with a family history of dementia or caregiving experience began with higher baseline knowledge (estimate 1.88, 95% CI 0.96 - 2.79; *P*<.001) and improved less over time, suggesting a ceiling effect. No significant interactions between group assignment and stratification variables were observed. Missing data analyses supported the assumption that data were missing at random.

Intentions to Engage in Risk Reduction

Intention scores differed at baseline, with higher scores in the intervention group. Both groups had a modest increase in their intentions from T1 to T2; the control group's improvement was significant (estimate 1.42, 95% CI 0.18 - 2.68; *P*=.03). However, there were no significant between-group differences at T2 (estimate −0.31, 95% CI −1.34 to 0.72; *P*=.55) or T3 (Multimedia Appendix 1; Figure 5). Table 4 provides crude intentions scores. Subgroup analyses did not identify significant interactions, and missingness followed an MAR pattern.

XSL•FO
RenderX

Figure 5. Intentions to engage in risk reduction score by level of education and group.

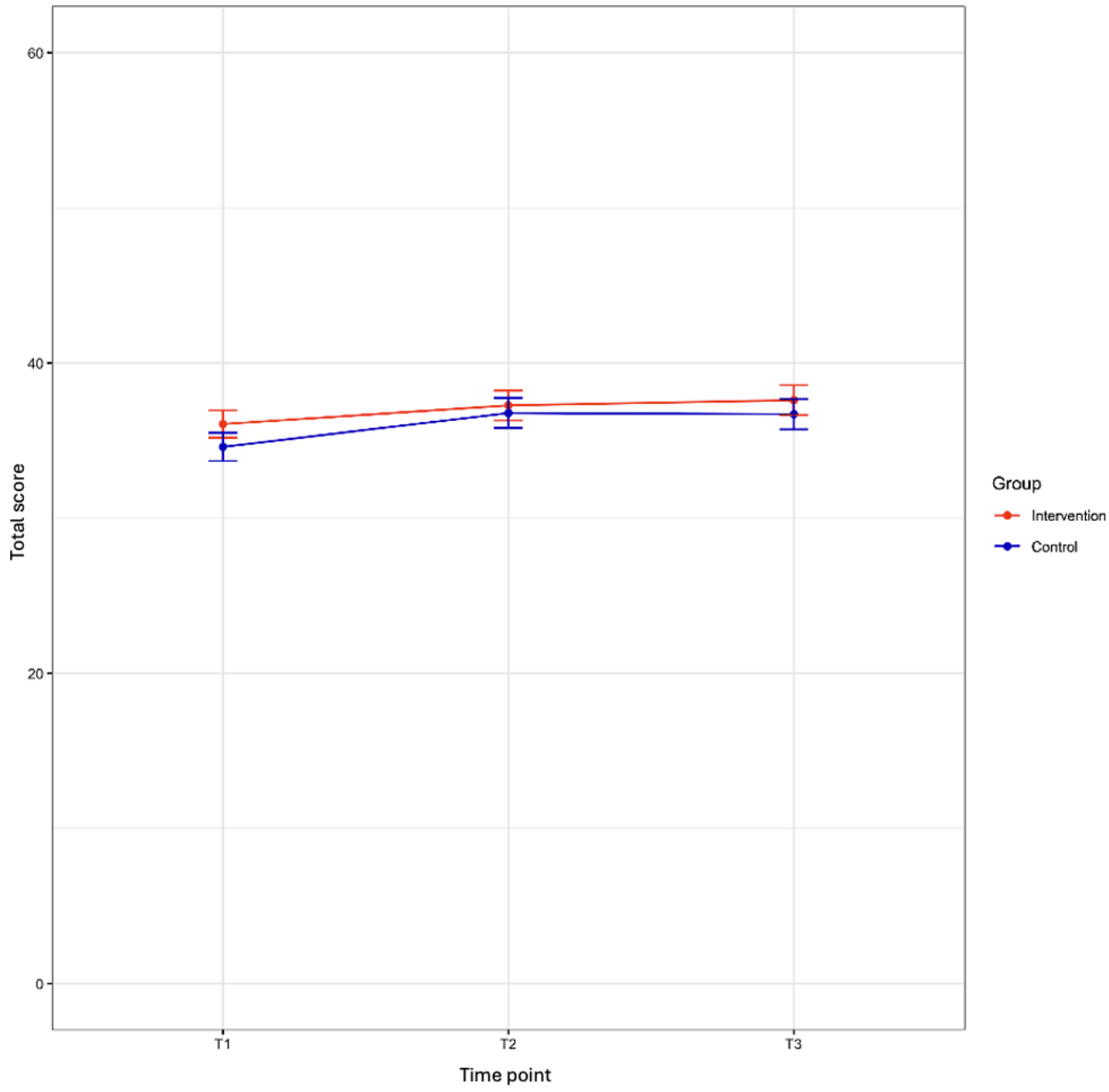

Table . Crude intentions scores across time points by group (N=510).

	Intervention (n=265)			Control (n=245)			Overall (N=510)		
	T1	T2	T3	T1	T2	T3	T1	T2	T3
Mean (SD)	37.5 (6.41)	39.2 (6.53)	39.3 (6.13)	36.2 (7.76)	37.6 (7.19)	38.1 (7.62)	36.9 (7.12)	38.4 (6.89)	38.7 (6.92)
Median (IQR)	38.0 (34.0-42.0)	40.0 (35.0-44.0)	40.0 (36.0-43.0)	37.0 (31.5-42.0)	39.0 (34.0-43.0)	40.0 (34-44.0)	38.0 (32.0-42.0)	40.0 (34.0-44.0)	40.0 (35.0-44.0)
Missing, n (%)	35 (13.2)	77 (29.1)	78 (29.4)	26 (10.6)	74 (30.2)	66 (26.9)	61 (12)	151 (29.6)	144 (28.2)

Health Behavior Changes: Godin-Shephard Leisure-Time Physical Activity Questionnaire

Self-reported physical activity levels did not differ between groups at baseline (estimate 1.87, 95% CI −2.52 to 6.25; *P*=.40). Both groups increased activity over time; the control group's activity rose significantly at T2 (estimate 3.23, 95% CI 0.81 - 5.64; *P*=.009) and was maintained at T3 (estimate 2.47, 95% CI 0.09 - 4.86; *P*=.04). Although the intervention group showed numerically greater gains, the group× time interactions were not statistically significant at either T2 (estimate 2.65, 95% CI −0.69 to 5.99; *P*=.12) or T3 (estimate 2.59, 95% CI −0.72 to 5.89; *P*=.13). Crude values are shown in Table 5, with

XSL•FO
RenderX

modeled estimates in Figure 6. Results of all models are shown in Multimedia Appendix 1.

Table . Crude Godin-Shephard Leisure-Time Physical Activity scores across time points by group (N=510).

	Intervention (n=265)			Control (n=245)			Overall (N=510)		
	T1	T2	T3	T1	T2	T3	T1	T2	T3
Mean (SD)	31.7 (25.0)	38.6 (27.5)	37.7 (27.1)	29.9 (23.5)	32.5 (26.6)	32.5 (26.5)	30.8 (24.3)	35.7 (27.2)	35.2 (26.9)
Median (IQR)	24.0 (14.0-46.8)	32.0 (19.0-56.0)	33.5 (17.0-50.8)	26.0 (14.0-40.0)	27.0 (12.0-44.2)	26.0 (15.0-42.0)	24.5 (14.0-44.0)	30.0 (15.0-49.0)	28.0 (15.8-48.0)
Missing, n (%)	3 (1.1)	56 (21.1)	55 (20.8)	3 (1.2)	57 (23.3)	51 (20.8)	6 (1.2)	113 (22.2)	106 (20.8)

Figure 6. Linear mixed model (LMM) estimates and 95% CI for Godin-Shephard Leisure-Time Physical Activity Questionnaire score by group.

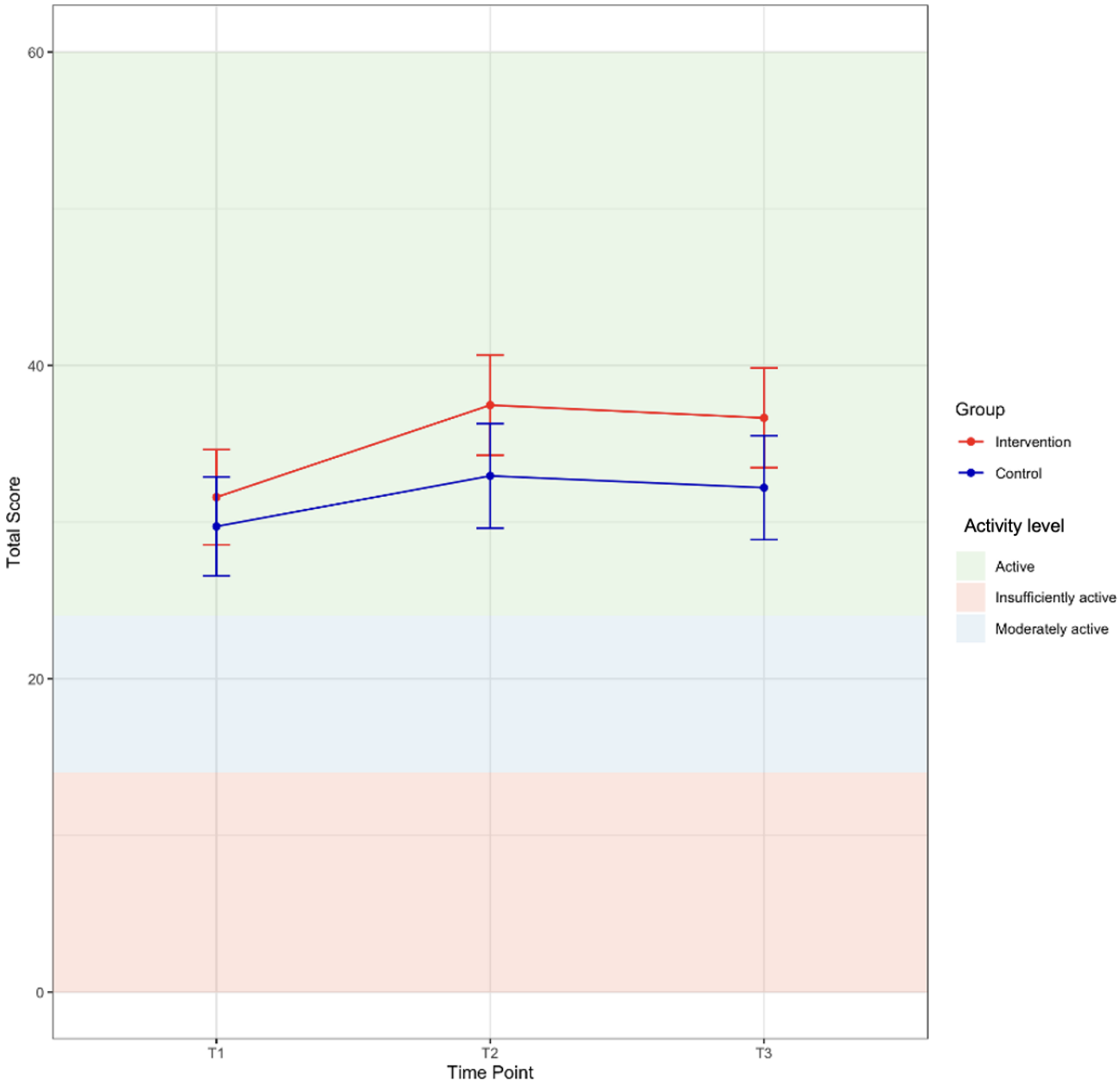

Other health behavior measures, including diet, alcohol use, sleep, and social activity, showed no significant between-group differences (Multimedia Appendix 2). PP analyses aligned with the ITT results.

IAM4all Outcomes

Using the IAM4all questionnaire at T2, participants in both groups rated the information as relevant or very relevant (intervention 94.1%, n=209; control 91.5%, n=194), easy to understand (intervention 100%, n=222; control 100%, n=212),

XSL•FO
RenderX

and beneficial (intervention 95.9%, n=213; control 91%, n=193). Nearly all participants indicated that they intend to use what they have learned (intervention 98.2%, n=218; control 97.6%, n=207), and anticipate using the information to better understand dementia prevention (intervention 60.3%, n=134; control 81.6%, n=173), discuss with someone else (intervention 63.5%, n=141; control 55.6%, n=118), or do things differently (intervention 62.1%, n=138; control 35.3%, n=75). These quantitative findings help to contextualize the qualitative themes reported in the "Qualitative Findings" section.

Qualitative Findings

Qualitative analysis was conducted inductively to generate themes and subthemes. A full list of qualitative themes and subthemes is provided in Multimedia Appendix 3. Overall, qualitative responses converged with data from the IAM4all, indicating that participants perceived the content as relevant, understandable, and actionable, with anticipated benefits for brain health and dementia risk reduction.

Relevance and Understandability

Participants commonly described the information as personally meaningful and focused on gaining a clearer understanding of dementia, including symptoms, early signs, and modifiable risk factors. Many emphasized the value of being able to recognize dementia symptoms and respond appropriately.

> *Recognize symptoms of dementia and learn the best ways to cope* [ID 551]

Participants also demonstrated a strong interest in recognizing the early signs of dementia to facilitate timely intervention. They displayed significant concern about the condition's risk factors and root causes, aiming to make informed lifestyle choices to mitigate or delay its onset.

> *Learn how to reduce the risk of dementia, strategies to delay the onset and minimize the symptoms of, dementia. Be able to recognize the signs of the onset of dementia.* [ID 298]

Intended Use and Action

Responses frequently demonstrated a proactive stance towards health, particularly in adopting lifestyle changes that could potentially delay or prevent the onset of dementia.

> *I will try to implement changes in my life based on what I learned in this.* [ID 248]

They were highly motivated to integrate new knowledge into their daily routines, making informed decisions related to diet, exercise, and mental health practices.

> *Motivates me to try and do more healthy things.* [ID 163]

> *Get motivated to do more walking, with the dogs. Maybe playing more cards with friends. I already play darts 2x a week and keep score in all games; adding and subtracting in my head while standing by the boards. Possibly learn how to play Sudoku, instead of just doing the crossword puzzles.* [ID 332]

Participants demonstrated a strong commitment to applying what they learned, with many initiating medical check-ups, adjusting their diets, and incorporating regular physical activity as direct responses to the information provided.

> *As a result of the information about the high correlation between hearing loss and dementia (higher than many other risk factors), I have made an appointment for ear wax removal, and will speak to the professional about hearing loss. Everyone should read the excellent information provided in the email campaign and surveys.* [ID 68]

> *Will go back on healthy diet, reduce sugar intake, get more exercise, more social activity.* [ID 328]

Participants frequently described increased awareness, motivation, and intentions to make lifestyle changes, and some reported early self-initiated actions.

> *This module did teach me somewhat more of what I need to do to keep myself and control my mental health, as well as physical health as I age. At 68, I do admit that I need to follow a better diet to help control my diabetes. As Emergency Medical Technicians, we were taught that we treated all causes of Dementia the same, as we could not differentiate between Diabetic, Alzheimer's or anything else...they were all forms of some mental health. Having seen the decline of my former Spouse into the Diabetic Dementia, and his refusal to even try to modify any of his routine: I know that I must make some changes to my lifestyle.* [ID 332]

Perceived Benefit

Beyond prevention, some participants valued learning about broader health connections, as well as current and emerging treatments. They appreciated the educational value of the study, which enhanced their understanding of brain health and its connections to other health issues to support more informed health and caregiving decisions.

> *I would never have believed that hearing loss might be a sign of dementia.* [ID 446]

Suggestions for Improvement

Participants also identified a desire for more comprehensive content, including practical steps for prevention and management, indicating areas for improving user experience and engagement in future studies.

> *I think that some of the suggestions for potential action could use more direction as to sources to engage in action - websites. Other information to enable people to follow up actively.* [ID 51]

The analysis further highlighted sex-specific interests, with female participants showing interest in caregiving strategies while male participants focused more on genetic risks. Additionally, across different age groups, there was a shift from a general focus on prevention in younger participants to practical management and caregiving in older adults, with a consistent appreciation for the informative impact of the study across all demographics.

XSL•FO
RenderX

Discussion

Principal Findings

The findings from this RCT demonstrate the effectiveness of the DementiaRisk internet-based education program in enhancing participants' knowledge of dementia risk and protective factors. These results were statistically significant at both the 4-week and 2-month postintervention follow-up assessments, reflecting sustained improvements post-intervention.

While both groups improved, the intervention group demonstrated significantly greater gains. This highlights the value of health education in general, but underscores the importance of providing well-designed dementia risk reduction-specific content.

The intervention improved knowledge particularly among participants with lower education levels, suggesting that e-learning may help bridge knowledge gaps among populations with potentially less prior exposure to dementia-related information. Participants who identified as having a family history of dementia or as a care partner had higher levels of baseline knowledge of dementia. Gains among those without such experience indicate that the intervention may help inform individuals who have not been previously motivated by personal exposure to dementia, thereby expanding its potential reach and impact on the general population.

Age Differences in Engagement With Preventative Measures

Older adults in our sample demonstrated greater intentions to engage in risk reduction behaviors compared to younger participants, which is consistent with prior evidence that perceived risk may increase motivation to adopt protective behaviors [43,44]. In contrast, younger adults may have lower perceived susceptibility to dementia, which can reduce their motivation to adopt lifestyle changes early in life [43,44].

Public health education efforts face unique challenges in engaging younger adults in reducing their risk of dementia. The concept of dementia as a 'disease of aging' contributes to a lack of urgency among younger populations [45]. Additionally, dementia prevention requires long-term behavior changes that may not provide immediate, tangible benefits, making it difficult to incentivize engagement in risk-reduction behaviors. Strategies such as emphasizing the broader benefits of lifestyle modifications (eg, improved cardiovascular health, mental well-being, and cognitive function) may help increase engagement among younger populations [44-47]. Framing dementia prevention within a broader "brain health" context, as suggested by the TPB, may also improve receptivity among younger individuals [48].

Interpreting Intentions to Engage in Risk Reduction

The increase in intentions observed across both groups indicates that exposure to health-related e-learning, even when not specific to dementia-risk reduction, may positively influence motivation to adopt healthier behaviors. However, measuring health behavior intentions remains a challenge due to the lack of a widely accepted, validated tool specifically designed for assessing intentions related to dementia risk reduction. The TPB suggests that intentions are strong predictors of behavior, but this relationship can be influenced by external factors such as perceived barriers or social norms [22,49,50]. Given the absence of a standardized instrument for measuring intentions in this context, our study relied on a custom-developed measure. This introduces potential limitations in comparability with other studies. Furthermore, it is important to acknowledge the "intention-behaviour gap," where individuals express a willingness to change but do not always follow through with action [50,51].

Health Behavior Change and Practical Implications

An interesting finding was the increase in self-reported physical activity levels among the participants, which suggests that the educational content may have prompted short-term behavior change. The sustained increase in physical activity observed at both follow-up points (4 wk and 2 mo postintervention) aligns with previous research suggesting that e-learning platforms, when well-designed, can encourage behavior changes beyond mere knowledge acquisition [52]. Given that physical activity is a key modifiable risk factor for dementia, this result has promising implications for the broader public health goal of dementia prevention. While the increase in physical activity levels among intervention participants is noteworthy, it should be interpreted within the context of a short-term study. Many health behaviors, particularly those related to physical activity, require sustained effort over time to produce meaningful long-term changes [53]. Some risk factors such as traumatic brain injury or hearing loss are not easily modifiable within the time frame of a 3-month study or may require additional resources or support from health care providers [54]. This distinction highlights the importance of differentiating between modifiable and nonmodifiable risk factors when designing and interpreting public health interventions.

Our study's findings align with existing research suggesting that short-term interventions are most effective for behaviors that require relatively low effort to change, such as increasing physical activity or making dietary improvements [55]. However, for more complex or chronic risk factors (eg, hypertension management and smoking cessation), longer-term interventions with ongoing reinforcement may be necessary to sustain behavior change.

Together, these results highlight that even brief, web-based interventions can produce measurable improvements in knowledge and short-term behavior, particularly when designed using evidence-based principles of instructional design. However, the practical significance and broader population impact of these changes warrant careful interpretation.

Interpretation of Findings

Across outcomes, the impact and practical relevance of observed changes varied. Knowledge of dementia risk factors increased by approximately 30% from baseline. Although modest, this gain likely reflects a meaningful increase in dementia literacy—particularly because small, scalable improvements in public awareness can produce significant population-level

XSL•FO
RenderX

benefits when applied broadly. While knowledge gains do not guarantee sustained behavior change, participants' qualitative comments indicate that many viewed the information as personally important and, in some cases, actionable in the short term, supporting the practical relevance of the knowledge gain as a precursor to informed risk-reduction decisions.

Behavioral intentions to engage in dementia risk-reduction activities also increased throughout the study, but the absolute change was small (approximately 2% - 3%) and occurred in both groups. The qualitative findings help explain these results by demonstrating that participants primarily described changes in awareness, motivation, and short-term action, particularly for behaviors perceived as immediately feasible, such as getting a hearing test, walking, or dietary adjustments, while expressing uncertainty about how to implement more complex or long-term risk-reduction strategies. This pattern suggests that exposure to health-related online education may positively influence motivation to adopt healthier behaviors, though the clinical importance of this change remains uncertain. The use of a novel intentions measure further limits comparability across studies, highlighting the need for validated scales tailored to dementia prevention contexts. In contrast, self-reported leisure-time physical activity increased by roughly 20% among participants, which may represent a meaningful short-term gain, even if participants largely remained within the same physical activity category. These effects may have been smaller in magnitude because participants were already relatively active and health conscious at baseline.

Instructional Design

One factor that may have contributed to the intervention's effectiveness is the use of evidence-based principles of instructional design. These principles, based on the Mayer Cognitive Theory of Multimedia Learning, have been shown to positively impact learning outcomes and retention [14,17]. Applying multimedia design principles such as segmenting, personalization, self-testing with feedback, and worked examples may strengthen comprehension and engagement in e-learning [15,56]. While the majority of research has examined these principles in health professionals, our findings extend this work to other general public populations.

The micro-learning email campaign also incorporated the principles of spaced education and repetition. These principles have also been shown to be effective in enhancing knowledge retention in health care professionals, but do not appear to have been systematically studied with respect to health education for the public [57,58].

Limitations and Future Directions

This study has several limitations. Although the sample was recruited from a representative panel of the Canadian population, it may not reflect populations with lower digital access or literacy. This could have introduced an element of selection bias, leading to higher baseline knowledge levels than anticipated. The sample also appeared relatively health-conscious, which may have limited the extent of behavior change. Additionally, all data were self-reported, which could introduce bias, especially in terms of behavior changes like physical activity. Finally, although a single coder performed the primary qualitative analysis, ongoing team-based verification and peer debriefing reduced potential bias and supported analytic consistency.

Conclusions

The findings from this randomized controlled trial indicate that a multimedia, instructionally designed, internet-based dementia-risk program can significantly improve dementia-related knowledge, particularly among participants with lower levels of education. We also observed modest, short-term increases in self-reported intention to engage in dementia risk–reducing behaviors and in physical activity levels in the intervention group, although effects on other behaviors were limited. Given its scalability and accessibility, this type of e-learning intervention represents a promising public health strategy to strengthen dementia risk-reduction literacy and support selected healthy behaviors at scale. Future research should broaden the program's reach and evaluate the durability of effects on knowledge, intentions, and health behaviors over time.

Acknowledgments

This work was supported by the research team at the Division of e-Learning Innovation at McMaster University. The authors would like to thank all who participated in this project for their time.

Funding

Funding was provided by the Public Health Agency of Canada Dementia Strategic Fund. The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript. AJL receives support through the John R Evans Chair in Health Sciences Educational Research and Instructional Development, Faculty of Health Sciences, McMaster University.

Data Availability

The datasets generated and analyzed during this study can be made available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

AJL and SA drafted the original study protocol. AJL, SC, and RW oversaw data collection, analysis, and interpretation. AS completed the primary quantitative analysis. DH completed the primary qualitative analysis. SA drafted the manuscript. All authors have reviewed the manuscript and approved the final version.

Conflicts of Interest

AJL, RS, and McMaster University are co-owners of DementiaRisk. The other authors declare no conflicts of interest.

Multimedia Appendix 1
Results of all models.
[XLSX File, 12 KB - jmir_v28i1e79405_app1.xlsx]

Multimedia Appendix 2
Health behavior data.
[XLSX File, 21 KB - jmir_v28i1e79405_app2.xlsx]

Multimedia Appendix 3
Qualitative themes and subthemes.
[DOCX File, 19 KB - jmir_v28i1e79405_app3.docx]

Checklist 1
CONSORT-EHEALTH checklist.
[PDF File, 1169 KB - jmir_v28i1e79405_app4.pdf]

References

1. Global status report on the public health response to dementia. : World Health Organization; 2021 URL: https://iris.who.int/server/api/core/bitstreams/9e8aab50-ca34-45e0-9fb0-26e675746f65/content [accessed 2026-02-14]
2. Prince M, Ali GC, Guerchet M, Prina AM, Albanese E, Wu YT. Recent global trends in the prevalence and incidence of dementia, and survival with dementia. Alzheimers Res Ther 2016 Jul 30;8(1):23. [doi: 10.1186/s13195-016-0188-8] [Medline: 27473681]
3. World Alzheimer Report 2021: Journey through the diagnosis of dementia. Alzheimer's Disease International. 2021 Sep. URL: https://www.alzint.org/resource/world-alzheimer-report-2021/ [accessed 2024-10-27]
4. Livingston G, Huntley J, Liu KY, et al. Dementia prevention, intervention, and care: 2024 report of the Lancet Standing Commission. The Lancet 2024 Aug;404(10452):572-628. [doi: 10.1016/S0140-6736(24)01296-0]
5. Dolatshahi Y, Mayhew A, O'Connell ME, et al. Prevalence and population attributable fractions of potentially modifiable risk factors for dementia in Canada: a cross-sectional analysis of the Canadian Longitudinal Study on Aging. Can J Public Health 2024 Dec;115(6):953-963. [doi: 10.17269/s41997-024-00920-7] [Medline: 39048849]
6. Dhana K, Evans DA, Rajan KB, Bennett DA, Morris MC. Healthy lifestyle and the risk of Alzheimer dementia: findings from 2 longitudinal studies. Neurology (ECronicon) 2020 Jul 28;95(4):e374-e383. [doi: 10.1212/WNL.0000000000009816] [Medline: 32554763]
7. Dhana K, Agarwal P, James BD, et al. Healthy lifestyle and cognition in older adults with common neuropathologies of dementia. JAMA Neurol 2024 Mar 1;81(3):233-239. [doi: 10.1001/jamaneurol.2023.5491] [Medline: 38315471]
8. The Landmark Study Report 1: Path to progress—navigating the path forward in dementia. : Alzheimer Society of Canada; 2022 URL: https://alzheimer.ca/sites/default/files/documents/Landmark-Study-Report-1-Path_Alzheimer-Society-Canada_0.pdf [accessed 2026-02-14]
9. Wimo A, Handels R, Antikainen R, et al. Dementia prevention: the potential long‐term cost‐effectiveness of the FINGER prevention program. Alzheimer's & Dementia 2023 Mar;19(3):999-1008. [doi: 10.1002/alz.12698]
10. Survey of Canadians regarding dementia prevention. Public Health Agency of Canada. 2022. URL: https://publications.gc.ca/collections/collection_2022/sc-hc/H14-414-2022-1-eng.pdf [accessed 2026-02-14]
11. Korthauer LE, Rosen RK, Tremont G, Davis JD. Intervention development for tailored education for aging and cognitive health (TEACH) for dementia prevention in midlife adults: protocol for a randomized controlled trial. JMIR Res Protoc 2024 Oct 16;13:e60395. [doi: 10.2196/60395]
12. Gong D, Peng Y, Liu X, et al. Dose health education on dementia prevention have more effects on community residents when a community physician/nurse leads it? A cross-sectional study. Front Public Health 2023;11:1101913. [doi: 10.3389/fpubh.2023.1101913] [Medline: 37206874]

XSL•FO
RenderX

13. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Instructional design variations in internet-based learning for health professions education: a systematic review and meta-analysis. Acad Med 2010 May;85(5):909-922. [doi: 10.1097/ACM.0b013e3181d6c319] [Medline: 20520049]
14. Clark R, Mayer R. e‐Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning, 5th edition: Wiley; 2023. [doi: 10.1002/9781119239086]
15. Issa N, Schuller M, Santacaterina S, et al. Applying multimedia design principles enhances learning in medical education. Med Educ 2011 Aug;45(8):818-826. [doi: 10.1111/j.1365-2923.2011.03988.x] [Medline: 21752078]
16. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: a meta-analysis. JAMA 2008 Sep 10;300(10):1181-1196. [doi: 10.1001/jama.300.10.1181] [Medline: 18780847]
17. Mayer RE. Multimedia Learning, 3rd edition: Cambridge University Press; 2020.
18. Belleville S, Anderson ND, Bherer L, et al. Brain health PRO/Santé cerveau PRO: the development of a web-based program for dementia literacy and risk factor reduction. J Prev Alzheimers Dis 2025 Jun;12(6):100134. [doi: 10.1016/j.tjpad.2025.100134] [Medline: 40090784]
19. Levinson AJ, Ayers S, Clark S, et al. Internet-based dementia prevention intervention (DementiaRisk): protocol for a randomized controlled trial and knowledge translation. JMIR Res Protoc 2025 Jan 27;14:e64718. [doi: 10.2196/64718] [Medline: 39869903]
20. Risk reduction of cognitive decline and dementia WHO guidelines. : World Health Organization; 2019 URL: https://iris.who.int/server/api/core/bitstreams/8fc586d7-b864-42eb-9274-95462496b328/content [accessed 2026-02-14]
21. Ajzen I. From intentions to actions: a theory of planned behavior. In: Action Control: Springer; 1985:11-39. [doi: 10.1007/978-3-642-69746-3_2]
22. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
23. Rise J, Kovac V, Kraft P, Moan IS. Predicting the intention to quit smoking and quitting behaviour: extending the theory of planned behaviour. Br J Health Psychol 2008 May;13(2):291-310. [doi: 10.1348/135910707X187245]
24. Conner M, Norman P, Bell R. The theory of planned behavior and healthy eating. Health Psychol 2002 Mar;21(2):194-201. [doi: 10.1037/0278-6133.21.2.194] [Medline: 11950110]
25. Cooke R, Dahdah M, Norman P, French DP. How well does the theory of planned behaviour predict alcohol consumption? A systematic review and meta-analysis. Health Psychol Rev 2016 Jun;10(2):148-167. [doi: 10.1080/17437199.2014.947547] [Medline: 25089611]
26. Chevance G, Caudroit J, Romain AJ, Boiché J. The adoption of physical activity and eating behaviors among persons with obesity and in the general population: the role of implicit attitudes within the theory of planned behavior. Psychol Health Med 2017 Mar;22(3):319-324. [doi: 10.1080/13548506.2016.1159705] [Medline: 26952746]
27. Blue CL. The predictive capacity of the theory of reasoned action and the theory of planned behavior in exercise research: an integrated literature review. Res Nurs Health 1995 Apr;18(2):105-121. [doi: 10.1002/nur.4770180205] [Medline: 7899566]
28. Pluye P, Grad R, Repchinsky C, et al. Four levels of outcomes of information‐seeking: a mixed methods study in primary health care. J Am Soc Inf Sci Tec 2013 Jan;64(1):108-125. [doi: 10.1002/asi.22793]
29. Annear MJ, Toye C, Elliott KEJ, McInerney F, Eccleston C, Robinson A. Dementia Knowledge Assessment Scale (DKAS): Confirmatory factor analysis and comparative subscale scores among an international cohort. BMC Geriatr 2017 Jul 31;17(1):168. [doi: 10.1186/s12877-017-0552-y] [Medline: 28760154]
30. Godin G, Shephard RJ. A simple method to assess exercise behavior in the community. Can J Appl Sport Sci 1985 Sep;10(3):141-146. [Medline: 4053261]
31. Pluye P, Granikov V, Bartlett G, et al. Development and content validation of the Information Assessment Method for Patients and Consumers. JMIR Res Protoc 2014 Feb 18;3(1):e7. [doi: 10.2196/resprot.2908] [Medline: 24550180]
32. RStudio integrated development environment (IDE). Posit. 2020. URL: https://posit.co/products/open-source/rstudio/?sid=1 [accessed 2026-03-02]
33. Bates D, Mächler M, Bolker BM, Walker SC. Fitting linear mixed-effects models using lme4. J Stat Softw 2015 Oct 7;67(1):1-48. [doi: 10.18637/jss.v067.i01]
34. Lenth RV. emmeans: estimated marginal means, aka least-squares means. The Comprehensive R Archive Network. 2024. URL: https://cran.r-project.org/web/packages/emmeans/index.html [accessed 2026-02-14]
35. Wickham H. Ggplot2 Elegant Graphics for Data Analysis, 2nd edition: Springer International Publishing; 2016. [doi: 10.1007/978-3-319-24277-4]
36. Norman G, Monteiro S, Salama S. Sample size calculations: should the emperor's clothes be off the peg or made to measure? BMJ 2012 Aug 23;345:e5278. [doi: 10.1136/bmj.e5278] [Medline: 22918496]
37. Lehr R. Sixteen S-squared over D-squared: a relation for crude sample size estimates. Stat Med 1992 Jun 15;11(8):1099-1102. [doi: 10.1002/sim.4780110811] [Medline: 1496197]
38. Neil-Sztramko SE, Belita E, Levinson AJ, Boyko J, Dobbins M. Evaluation of an online knowledge translation intervention to promote cancer risk reduction behaviours: findings from a randomized controlled trial. BMC Cancer 2019 Nov 21;19(1):1138. [doi: 10.1186/s12885-019-6361-2] [Medline: 31752751]

XSL•FO
RenderX

39. Vears DF, Gillam L. Inductive content analysis: a guide for beginning qualitative researchers. Focus Health Prof Educ 2022;23(1):111-127. [doi: 10.11157/fohpe.v23i1.544]
40. NVivo (version 14). Lumivero. 2023. URL: https://lumivero.com/ [accessed 2026-02-14]
41. Jackson K, Bazeley P. Qualitative Data Analysis with NVivo, 3rd edition: Sage Publications; 2019.
42. Leventhal H, Phillips LA, Burns E. The common-sense model of self-regulation (CSM): a dynamic framework for understanding illness self-management. J Behav Med 2016 Dec;39(6):935-946. [doi: 10.1007/s10865-016-9782-2] [Medline: 27515801]
43. Hagger MS, Orbell S. The common sense model of illness self-regulation: a conceptual review and proposed extended model. Health Psychol Rev 2022 Sep;16(3):347-377. [doi: 10.1080/17437199.2021.1878050] [Medline: 33461402]
44. Cations M, Radisic G, Crotty M, Laver KE. What does the general public understand about prevention and treatment of dementia? A systematic review of population-based surveys. PLoS One 2018;13(4):e0196085. [doi: 10.1371/journal.pone.0196085] [Medline: 29672559]
45. Norton S, Matthews FE, Barnes DE, Yaffe K, Brayne C. Potential for primary prevention of Alzheimer's disease: an analysis of population-based data. Lancet Neurol 2014 Aug;13(8):788-794. [doi: 10.1016/S1474-4422(14)70136-X] [Medline: 25030513]
46. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021 May 12;23(5):e17917. [doi: 10.2196/17917] [Medline: 33978589]
47. Farina FR, Booi L, Occhipinti JA, et al. Young adult brain capital: a new opportunity for dementia prevention. J Alzheimers Dis 2023;94(2):415-423. [doi: 10.3233/JAD-230260] [Medline: 37302036]
48. Ajzen I. The theory of planned behaviour: reactions and reflections. Psychol Health 2011 Sep;26(9):1113-1127. [doi: 10.1080/08870446.2011.613995] [Medline: 21929476]
49. Sheeran P, Webb TL. The intention-behavior gap. Soc Pers Psychol 2016 Sep;10(9):503-518. [doi: 10.1111/spc3.12265]
50. Rhodes RE, de Bruijn G. How big is the physical activity intention–behaviour gap? A meta‐analysis using the action control framework. Br J Health Psychol 2013 May;18(2):296-309. [doi: 10.1111/bjhp.12032]
51. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010 Feb 17;12(1):e4. [doi: 10.2196/jmir.1376] [Medline: 20164043]
52. Bauman AE, Reis RS, Sallis JF, et al. Correlates of physical activity: why are some people physically active and others not? Lancet 2012 Jul 21;380(9838):258-271. [doi: 10.1016/S0140-6736(12)60735-1] [Medline: 22818938]
53. Whiteneck GG, Cuthbert JP, Corrigan JD, Bogner JA. Prevalence of self-reported lifetime history of traumatic brain injury and associated disability: a statewide population-based survey. J Head Trauma Rehabil 2016;31(1):E55-E62. [doi: 10.1097/HTR.0000000000000140] [Medline: 25931187]
54. Kwasnicka D, Dombrowski SU, White M, Sniehotta F. Theoretical explanations for maintenance of behaviour change: a systematic review of behaviour theories. Health Psychol Rev 2016 Sep;10(3):277-296. [doi: 10.1080/17437199.2016.1151372] [Medline: 26854092]
55. Issa N, Mayer RE, Schuller M, Wang E, Shapiro MB, DaRosa DA. Teaching for understanding in medical classrooms using multimedia design principles. Med Educ 2013 Apr;47(4):388-396. [doi: 10.1111/medu.12127] [Medline: 23488758]
56. Matos J, Petri CR, Mukamal KJ, Vanka A. Spaced education in medical residents: an electronic intervention to improve competency and retention of medical knowledge. PLoS ONE 2017;12(7):e0181418. [doi: 10.1371/journal.pone.0181418] [Medline: 28759606]
57. Phillips JL, Heneka N, Bhattarai P, Fraser C, Shaw T. Effectiveness of the spaced education pedagogy for clinicians' continuing professional development: a systematic review. Med Educ 2019 Sep;53(9):886-902. [doi: 10.1111/medu.13895] [Medline: 31144348]
58. A dementia strategy for Canada: Together We Aspire. Public Health Agency of Canada. 2019. URL: https://www.canada.ca/en/public-health/services/publications/diseases-conditions/dementia-strategy.html [accessed 2026-02-14]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
CONSORT-eHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
IAM4all : Information Assessment Method for Patients and Consumers
ITT: intention-to-treat
LMM: linear mixed model
MAR: missing at random
PP: per protocol
RCT: randomized controlled trial
TPB: theory of planned behavior
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Stone; submitted 20.Jun.2025; peer-reviewed by N Coley, R Newton; revised version received 28.Jan.2026; accepted 28.Jan.2026; published 10.Mar.2026.

Please cite as:
Levinson AJ, Ayers S, Clark S, Woodburn R, Schneeberg A, Hadid D, Kates N, Strudwick G, Grad R, Papaioannou A, Dobbins M, Siu H, Duarte D, Saperson K, Marr S, Oliver D, Neil-Sztramko S
Effects of Internet-Based Dementia Risk Reduction Education on Risk and Protective Factor Knowledge, Intentions, and Health Behaviors: Randomized Controlled Trial
J Med Internet Res 2026;28:e79405
URL: https://www.jmir.org/2026/1/e79405
doi:10.2196/79405

© Anthony J Levinson, Stephanie Ayers, Sandra Clark, Rebekah Woodburn, Amy Schneeberg, Dima Hadid, Nick Kates, Gillian Strudwick, Roland Grad, Alexandra Papaioannou, Maureen Dobbins, Henry Siu, Dante Duarte, Karen Saperson, Sharon Marr, Doug Oliver, Sarah Neil-Sztramko. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Enhancing LGBTQ+ Inclusivity in an AI-Powered Sexual Health Chatbot: User-Centered Design Approach Through a Nonprofit and Academic Partnership

William Wibowo Liem[1], MSW; Elizabeth Casline[1], PhD; Julianna Lorenzo[1], BS; Jacob D Gordon[2], MSW, PhD; Andrés Alvarado Avila[1], MPH; Attia Taylor[3], BA; Nicole Levitz[3], MPH; Michael C O'Keefe[3], MA; Kathryn Macapagal[1,4], PhD

[1]Impact Institute, Northwestern University, Chicago, IL, United States

[2]College of Nursing, University of Cincinnati, Cincinnati, OH, United States

[3]Planned Parenthood Federation of America, New York, NY, United States

[4]Department of Medical Social Sciences, Impact Institute, Northwestern University, 625 North Michigan Avenue, Floor 14, Chicago, IL, United States

Corresponding Author:
Kathryn Macapagal, PhD
Impact Institute, Northwestern University, Chicago, IL, United States

Abstract

Background: Despite the growing use of digital platforms for sexual health education, many tools fail to meet the needs of LGBTQ+ (lesbian, gay, bisexual, transgender, and queer) teenagers, who often lack access to inclusive, affirming resources. Artificial intelligence (AI)–enabled chatbots have emerged as promising tools to address these gaps, but concerns remain around bias, usability, and trustworthiness—particularly for queer and transgender teenagers. Participatory design approaches centered around marginalized teenagers are critical to ensuring these tools are relevant, trustworthy, and equitable; yet, few studies have systematically engaged LGBTQ+ teenagers in the co-design of AI-powered sexual health interventions.

Objective: This paper examines LGBTQ+ teenagers' perceptions of Roo, Planned Parenthood Federation of America's (PPFA) AI-powered sexual education chatbot to identify opportunities and challenges in delivering LGBTQ+-inclusive, affirming sexual health information.

Methods: Embedded within Sharing Health Education Resources, a hybrid effectiveness implementation trial of a digital HIV prevention intervention for LGBTQ+ teenagers, we collaborated with PPFA to create a customized instance of Roo for integration into this study. We engaged a Youth Advisory Council comprising 15 LGBTQ+ teenagers to independently explore and interact with Roo, then gathered feedback through a week-long asynchronous discussion on a private Discord (Discord Inc) server. The research team posed open-ended questions prompting participants to reflect on Roo's inclusivity, usability, and content priorities. We used rapid qualitative analysis organized around our research questions.

Results: Participants expressed both skepticism and curiosity about AI's role in delivering sexual health information, offering critical insights on the chatbot's language, trustworthiness, and relevance. Teenagers identified key limitations in Roo's inclusivity, tone, and interface, particularly around transgender-specific content, conversational depth, and stigma reduction. These findings informed targeted content updates, interface refinements, and transparency improvements, implemented by PPFA to enhance Roo for broader use. Specific changes included expanding LGBTQ+ affirming content, revising language to eliminate gendered assumptions, incorporating concrete statistics and contextualized examples to reduce stigma, and adding clearer disclosures around Roo's AI capabilities and limitations.

Conclusions: Academic and nonprofit collaborations can leverage participatory methods to enhance digital health tools in real-world contexts. LGBTQ+ teenagers served not only as testers but as co-designers, shaping the chatbot's evolution and surfacing broader lessons about trust, AI literacy, and health equity. This study demonstrates that marginalized teenagers possess the critical insights needed to meaningfully shape AI-enabled health interventions when provided with structured opportunities for engagement. This partnership offers a scalable model for integrating community voice into the development, evaluation, and implementation of inclusive, AI-enabled health technologies.

(J Med Internet Res 2026;28:e78621) doi:10.2196/78621

KEYWORDS

sexual health; sex education; adolescents; sexual and gender minorities; artificial intelligence; stakeholder participation; digital health; AI; LGBTQ+; inclusivity; lesbian, gay, bisexual, transgender, and queer

XSL•FO
RenderX

Introduction

LGBTQ+ (lesbian, gay, bisexual, transgender, and queer) teenagers experience significant sexual and reproductive health disparities compared to their cisgender, heterosexual peers. They face higher rates of sexually transmitted infections (STIs) and HIV, with young men who have sex with men accounting for 81% of new HIV diagnoses among teenagers aged 13-19 years in the United States [1]. Sexual minority individuals who can become pregnant are also at increased risk for unintended pregnancies due to higher rates of sexual victimization and coerced sex [2]. Additionally, LGBTQ+ teenagers are more likely to experience intimate partner violence than their heterosexual peers, with transgender and nonbinary teenagers reporting particularly high rates of victimization [3]. Without access to inclusive, affirming sexual health resources designed for their needs, these disparities risk worsening. Accordingly, inclusivity, usability, and perceived helpfulness emerge as key criteria for evaluating digital sexual health tools—determining whether LGBTQ+ teenagers feel represented, can navigate the platform, and find information meaningful for decision-making.

Access to comprehensive, inclusive sexual health education plays a critical role in equipping teenagers with the knowledge and skills to navigate consent, boundaries, and healthy relationships. Yet, school-based sexual health education in the United States remains overwhelmingly focused on cisgender, heterosexual experiences, often omitting or stigmatizing LGBTQ+ identities [4,5]. As a result, many LGBTQ+ teenagers turn to the internet to fill these educational gaps [6], seeking inclusive and affirming resources that traditional channels fail to provide [7-9]. In the United States, digital interventions have emerged as powerful tools for delivering inclusive sexual health education to teenagers. Platforms ranging from mobile apps to interactive websites offer accessible, affirming alternatives [10]. For LGBTQ+ teenagers in particular, online platforms can serve as safer spaces to access accurate information, connect with peers navigating similar challenges, and build critical knowledge around consent, relationships, and HIV prevention [11]. Evidence from interventions such as TechStep (designed for transgender teenagers) and MyPEEPS Mobile (tailored to same-sex attracted teenage boys) demonstrates how well-designed, targeted content can successfully reduce condomless sex practices [12,13]. The success of these programs underscores the importance of thoughtful, community-informed design; features such as user-friendly interfaces, engaging and culturally relevant messaging, and interactive components (eg, reminder texts or skill-building exercises) can significantly enhance user engagement and educational outcomes [14]. As digital tools continue to evolve, these insights offer a critical foundation for the next generation of technology-enabled interventions, including those powered by artificial intelligence (AI).

Incorporating AI in personalized health care is emerging as a promising avenue to further enhance accessibility and engagement with users [15]. However, the adoption of AI-driven tools in sexual health education has sparked both optimism and skepticism, particularly among LGBTQ+ teenagers. While chatbots can effectively deliver health interventions when designed appropriately with LGBTQ+ individuals, significant concerns remain about information quality, reading level appropriateness, and inconsistent accuracy [16,17]. These concerns are particularly acute for Black LGBTQ+ youth facing health care access barriers. A scoping review of generative conversational AI apps for the LGBTQ community found that while initial deployments were feasible and well-received, substantial improvements are needed in content quality and engagement [18]. These findings underscore both AI's potential to address health disparities and the critical need for further development, formal evaluation, and community-informed design.

The "myth of AI," as described by McAra-Hunter [19], creates unrealistic expectations that AI can universally solve complex problems, often ignoring the nuanced needs of marginalized groups. This disconnect poses risks, as AI tools are frequently developed without sufficient input from LGBTQ+ communities, leading to unrealistic expectations and potentially harmful outcomes. For example, ChatGPT (OpenAI) has been shown to use outdated or exclusionary language—assuming people with vaginas are women, and those with penises are men—and to lack sociopolitical sensitivity in its responses [20]. These findings echo broader concerns raised by Tomasev et al [21], who caution that AI systems trained on large-scale datasets without community-informed oversight risk encoding and amplifying societal biases, particularly those affecting gender and sexual minorities. Moreover, a study by Hopelab found that LGBTQ+ teenagers are more likely than their cisgender and heterosexual peers to avoid AI tools due to concerns about inaccuracies and biases [22]. Integrating teenagers' input into the design process can enhance trust, usability, and the overall experience of AI-powered interventions, ensuring they align more closely with the needs of the communities they aim to serve. However, queer youth are rarely positioned as designers of AI health tools, despite their unique expertise in navigating heteronormative and cisnormative health care systems [23].

Despite the promise of digital and AI-powered tools, research-driven innovations often face a persistent research-to-adoption gap, where effective interventions fail to achieve sustained public health impact. It is estimated that it can take an average of 17 years for research findings to be translated into real-world practice, with only a small fraction ever implemented at scale [24]. To address this gap, it is critical to embed considerations of implementation and sustainability early in the design process, especially for teenager-centered tools. This is especially important for LGBTQ+ teenagers who often rely on digital tools to access affirming information and support that may be unavailable in their offline environments.

Participatory research approaches are particularly well-suited for engaging marginalized populations such as LGBTQ+ teenagers, as they center community expertise, address power imbalances inherent in research, and ensure interventions reflect the lived realities of those most affected [25]. By involving teenagers as co-designers rather than passive recipients, participatory methods not only improve intervention relevance and acceptability but also build trust with communities that have historically experienced harm or exclusion from medical institutions [26]. A key implementation is developing nonprofit

XSL•FO
RenderX

and academic partnerships that bridge the research-to-adoption gap while facilitating rigorous co-design and evaluation methods. Evidence shows that involving teenagers in intervention planning leads to better reflection of their needs and perspectives while addressing power imbalances in decision-making [27]. Coproducing interventions with teenagers increases their sense of ownership and supports adoption, with 1 review reporting more than a 2-fold increase in self-reported use of sexual and reproductive health services [28]. For these partnerships to maximize lasting public health benefit, there is a high need to prioritize cocreation, effectively manage expectations early, and integrate implementation planning from the outset.

To illustrate how aligning missions across sectors can accelerate impact, we describe the nonprofit and academic collaboration between Planned Parenthood Federation of America (PPFA) and Northwestern University aimed to enhance Roo (Figure 1), an existing, publicly available PPFA-operated AI-powered platform, through a week-long asynchronous discussion with LGBTQ+ teenagers on Discord.

Figure 1. Roo chatbot interface (Planned Parenthood, accessed May 31, 2025). Roo, Planned Parenthood Federation of America's free, web-based AI-powered sexual education chatbot. The interface features example prompts (eg, "What happens during puberty?" and "How long after unprotected sex can I know if I'm really pregnant?"), a privacy assurance message, and an open text input field for user-generated questions. Roo's database covers topics including abortion, birth control, relationships, LGBTQ+ health, and symptoms. AI: artificial intelligence; BTW: by the way; LGBTQ+: lesbian, gay, bisexual, transgender, and queer.

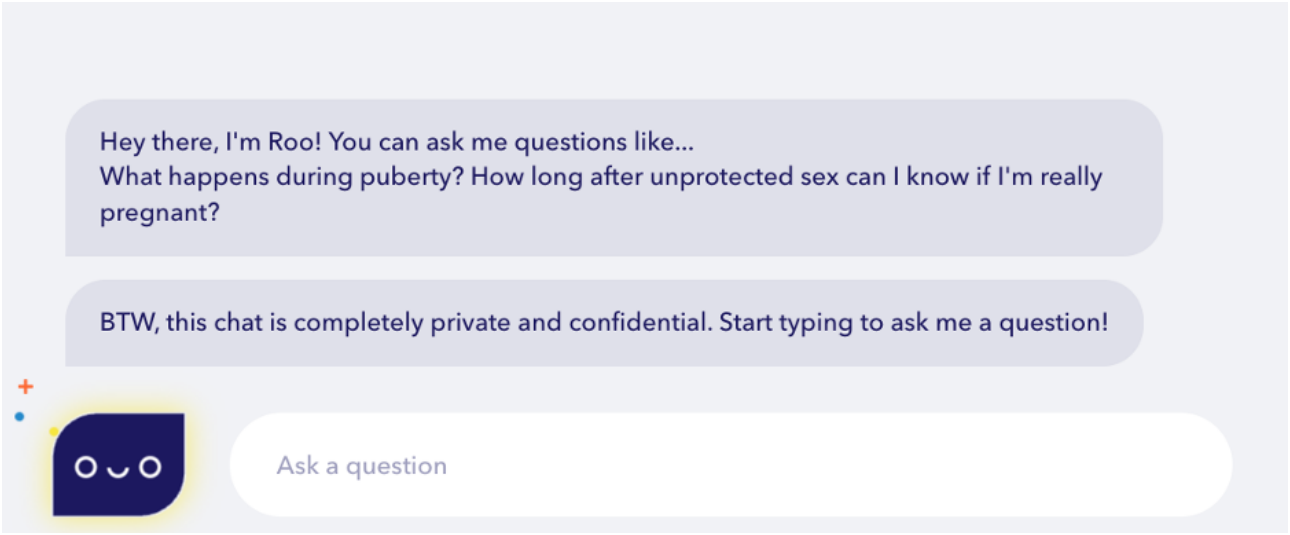

Methods

Study Design

This study was conducted as part of a hybrid type 1 effectiveness implementation randomized controlled trial examining the effectiveness of a digital text-based HIV prevention intervention called Sharing Health Education Resources (SHER) compared to an information-only control among LGBTQ+ teenagers. SHER builds on an earlier text-messaging intervention called Guy2Guy (G2G), a text messaging-based HIV prevention program to deliver educational and skill-oriented information to male teenagers who identify as gay, bisexual, and/or queer [29]. Within SHER, Roo, Planned Parenthood's AI-powered sexual education chatbot, was adapted and evaluated as a component intervention. For the purposes of this paper, we focus specifically on Roo and report findings from Youth Advisory Council (YAC) feedback about its inclusivity, usability, and trustworthiness.

The original G2G intervention featured an on-demand texting function, G2Genie, that allowed participants to request additional information by sending keywords (eg, texting "G2Genie relationships" would prompt a curated set of messages about relationship skills). Building on this concept—and recognizing the growing emphasis on personalization and rapid delivery in digital health interventions—SHER planned to adapt this feature by developing a bespoke sexual education chatbot feature tailored to the specific needs of LGBTQ+ teenagers [30,31]. The goal of the chatbot was to enable participants to ask follow-up questions and receive individually tailored responses across a broader range of sexual health topics than those included in the SHER text messages (Figure 2). By collaborating with a nonprofit resource already integrated into public health infrastructure, Northwestern University was able to accelerate translation, broaden reach among diverse teenage audiences, and leverage existing dissemination pathways to enhance both impact and equity. Rather than building a proprietary chatbot accessible only to study participants, our team strategically partnered with an existing, publicly available platform, PPFA's Roo.

XSL•FO
RenderX

Figure 2. Integration of the Roo chatbot into the SHER digital HIV prevention intervention for LGBTQ+ teenagers. The screenshot demonstrates how participants who enrolled in the SHER text messaging intervention (left panel) received a prompt directing them to Roo, Planned Parenthood's AI-powered sexual education chatbot (right panel). The customized Northwestern University instance of Roo was accessible via a dedicated microsite and provided on-demand sexual health information, complementing SHER's structured intervention content. AI: artificial intelligence; LGBTQ+: lesbian, gay, bisexual, transgender, and queer; PEP: post-exposure prophylaxis; PrEP: pre-exposure prophylaxis; SHER: Sharing Health Education Resources; TasP: Treatment as Prevention.

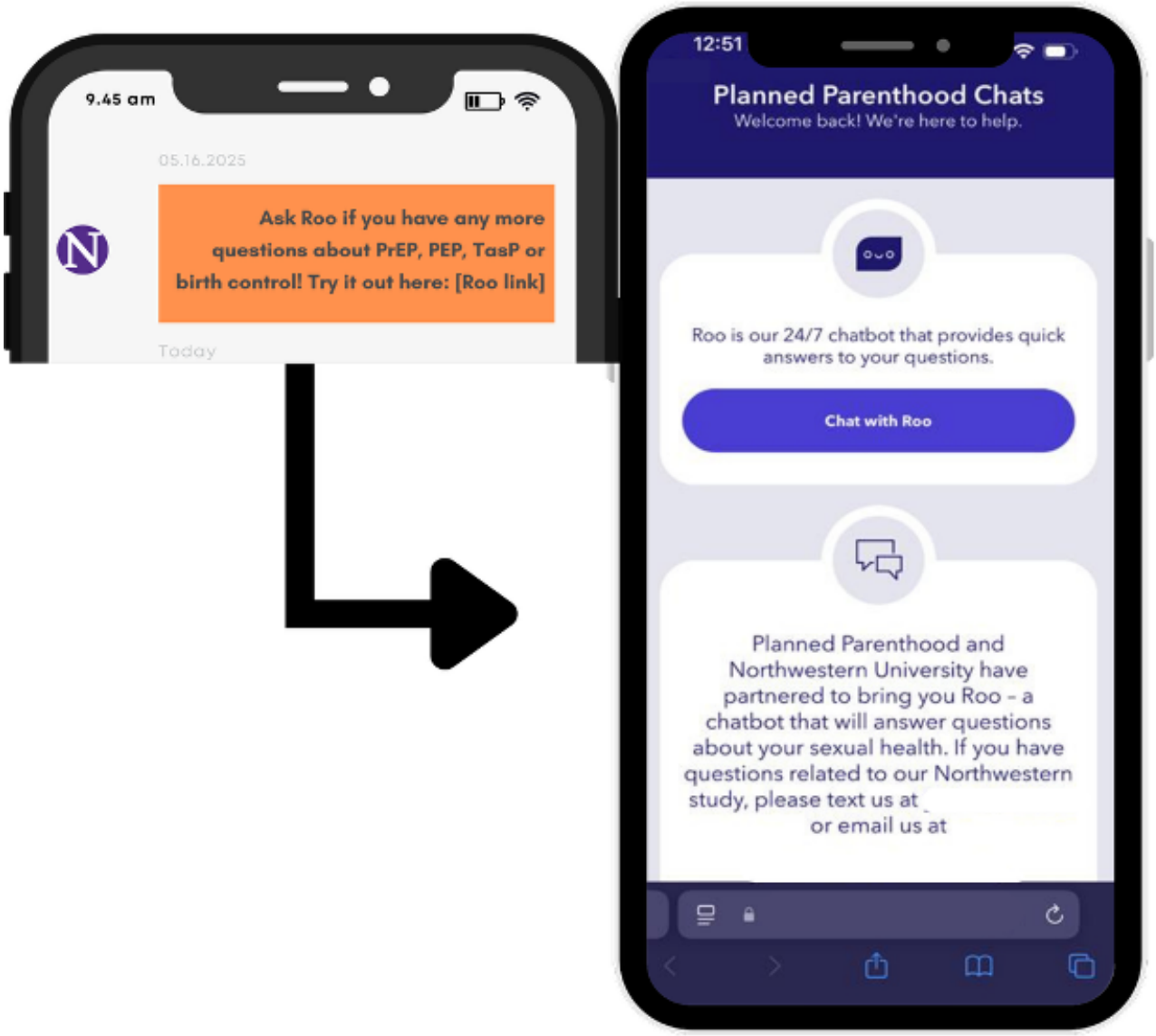

Roo is an AI-powered chatbot designed to provide confidential, accessible, and medically accurate sexual health information. Roo's responses are developed by Planned Parenthood's content and education teams to maintain a friendly, nonjudgmental tone through intentional strategies such as using conversational language (eg, "You deserve fun, comfortable, and pleasurable sex"), avoiding clinical jargon, affirming diverse experiences as normal, and framing sensitive topics without moral judgment. For instance, rather than emphasizing risk, Roo normalizes STI testing as routine care (eg, "Getting tested is often as easy as peeing in a cup at a health center") and highlights that "most of the time, STDs don't show symptoms, so the only way to know for sure is to get tested." Accessible via the PPFA website [32], Roo allows users to ask free-text questions and responds immediately with a nonjudgmental answer and a list of additional resources for the user to engage with.

Roo's development is grounded in the Unified Theory of Behavior, which integrates multiple behavioral theories to explain and influence health behaviors through attitudes, normative beliefs, self-efficacy, and behavioral intention [33]. Roo's content and education teams crafted responses to both provide information while supporting users in setting behavioral intentions. This application of the theory is operationalized through several strategies: presenting pros and cons to support informed decision-making, acknowledging emotions users may experience, connecting users to health care resources, and providing referrals to partner organizations. For example, when discussing contraception options, Roo provides balanced

XSL•FO
RenderX

information about effectiveness and side effects while prompting users to consider their personal priorities and next steps, such as scheduling a clinic appointment. All educational content on Planned Parenthood's website, including Roo, reflects this framework by helping users identify action plans and set intentions for next steps in their sexual health journey. To our knowledge, Roo is the only sexual and reproductive health chatbot that explicitly applies the Unified Theory of Behavior in its design and content strategy, distinguishing it from other AI-powered tools focused primarily on information delivery rather than behavioral support.

Roo uses natural language processing to scan each question for keywords to trigger the correct answer or send the user to another resource that can answer their question. Natural language processing uses machine learning and deep learning models to recognize patterns in the human language. Roo is a Planned Parenthood proprietary product that draws exclusively from a curated library of responses designed specifically for Roo users. All content is crafted and consistently iterated upon by Planned Parenthood's technology and sex education leadership teams, with all responses conforming to the federation's medical standards and guidelines. This human-curated approach ensures that responses maintain medical accuracy, age-appropriateness, and alignment with PPFA's mission of providing inclusive, affirming sexual health education. Topics in Roo cover a wide range of sexual and reproductive health topics. The questions that Roo can answer are limited by the database of content that Roo is able to pull from. The content team at Planned Parenthood uses data from Roo, other digital products, and user experience research to add common and frequently asked questions and their responses into Roo. If Roo cannot answer a user's question, it will let them know and offer alternatives, including PPFA's anonymous live chat with real health educators, as well as links to PPFA's blog and other educational resources.

In 2023, Roo facilitated approximately 640,000 sessions [34]. Roo offers a complementary approach to SHER's content and structure. Whereas SHER delivers structured, sequential educational messages with preplanned daily content, Roo is designed to offer on-demand, conversational access to medically accurate, affirming, and developmentally appropriate sexual health education. Roo's database covers a broad range of topics relevant to teenagers, including safe sex practices, consent, healthy relationships, LGBTQ+ health, and navigating health care systems.

This complementary functionality made Roo a promising addition to SHER, enhancing the program's ability to support LGBTQ+ teenagers' diverse information needs without duplicating content delivery. For PPFA, Roo has provided an opportunity to align its mission with innovative delivery models; however, implementing an AI-enabled tool that adequately meets the needs of LGBTQ+ teenagers has required technical and evaluation infrastructure beyond PPFA's in-house capacity. By collaborating with an academic institution, PPFA was able to benefit from Northwestern University's methodological expertise to co-design an intervention evaluation to receive actionable recommendations for iteratively refining Roo's responsiveness to LGBTQ+ teenagers' needs.

Academic and Nonprofit Partnership Collaborative Development Process

We sought to develop an academic and nonprofit partnership that advances both scientific rigor and real-world impact in the development of inclusive digital health innovations. Below, we describe the development of this academic and industry partnership using Proctor et al [35] domains for systematically specifying implementation strategies.

Speaking of actors, the partnership was initiated by the Northwestern University research team, who became aware of PPFA's Roo chatbot and were connected through a mutual colleague. Key actors included academic investigators, digital health researchers, product managers, and health education specialists from both institutions.

For action, we developed a collaborative relationship with PPFA that required iterative conversations to align organizational goals, resources, and capacities. Academic and industry collaborations in digital health often face challenges around aligning timelines, balancing rigor and innovation, and navigating competing priorities [30]. Early meetings focused on relationship-building, clarifying organizational priorities, and identifying areas of mutual benefit. Specific actions included assessing the fit between Roo and the SHER intervention, mapping resource needs, identifying technical constraints, establishing intervention goals across both teams, and drafting and negotiating formal agreements, including a nondisclosure agreement and data use agreement to facilitate ethical collaboration and data sharing. Given the reciprocal benefits identified, both parties agreed that there would be no financial exchange.

Subsequent meetings explored integration pathways, ultimately deciding to create a standalone customized microsite version of Roo for SHER to facilitate tracking participant engagement and allow key modifications to maintain intervention integrity. These changes included removing the option to talk to a live health educator and directing users who needed help with Roo to call the SHER study team rather than chat with PPFA. This instance was built and maintained by PPFA's product team, with all feedback integrated into both the SHER study instance and the publicly facing version of Roo.

Once the partnership was established, Northwestern University engaged its YAC—a diverse group of LGBTQ+ teenagers to test Roo and provide structured feedback on its content, user experience, and inclusivity.

Of temporality and dose, partnership development occurred over 6 months (February to July 2023) with monthly formal meetings supplemented by asynchronous collaboration, with higher-frequency early communication for trust-building and later focus on YAC feedback and platform customization.

Of target and justification, the overarching target was to integrate Roo into SHER in a way that enhanced accessibility, inclusivity, and personalization of sexual health resources for LGBTQ+ teenagers. PPFA aimed to expand its reach to LGBTQ+ teenagers, while Northwestern University aimed to apply participatory design principles centering LGBTQ+ youth voices. Partnering with an existing, trusted platform embedded

XSL•FO
RenderX

in public health infrastructure allowed the project to accelerate translation and broaden access beyond this study's sample.

Ethical Considerations

Concerning ethics approval, all study procedures were approved by the Northwestern University Institutional Review Board (study number: STU00217358) with a waiver of parental consent, given the potential risks of requiring guardian permission for LGBTQ+ minors.

Of informed consent, for minors, informed assent and a capacity-to-consent assessment were required. Eligibility was confirmed through screening forms and one-on-one video calls with study staff, which also served to review participation expectations and privacy protocols.

Of privacy and confidentiality, the YAC operates on a private, invite-only Discord server, where participants can use pseudonyms to protect their identity. Study staff moderated the server to ensure a safe, respectful environment. Participants could withdraw from this study at any time without penalty, and all data were stored securely in compliance with institutional data protection standards. The principal investigator is a licensed clinical psychologist with expertise in LGBTQ+ adolescent health, ensuring clinical oversight of all activities involving sensitive sexual health discussions.

Concerning compensation, participants received US $30 in digital gift cards per month for active participation, with opportunities to earn up to US $200 annually for optional activities.

Of the use of images, no identifiable participant images are included in this manuscript or supplementary materials.

Recruitment

As part of the SHER effectiveness implementation trial, LGBTQ+ teenagers were recruited through paid Meta (Facebook or Instagram) advertisements and an institutional participant registry. Interested teenagers were directed to a secure REDCap (Research Electronic Data Capture; Vanderbilt University) screening survey to complete eligibility questions and provide contact information. To reduce impostor enrollment, identity verification included a brief video call assessing personal information and comprehension of study details [36-38]. Following a virtual orientation session conducted via Zoom (Zoom Communications, Inc), eligible participants were invited to a private Discord server, an instant messaging platform that supports multimedia messaging services, voice, video, and multimedia communication.

Eligible participants were aged 13‐18 years, identified as sexual and/or gender minority, lived in the United States or US territories, and were able to read in English at an eighth-grade level. Recruitment sought to oversample transgender or gender-diverse and racial or ethnic minority teenagers to ensure diverse feedback. Recruitment occurred between January 2023 and July 2023 and was conducted entirely online with no in-person involvement. The YAC ultimately included 15 LGBTQ+ teenagers aged 15‐18 (mean 16.9) years (see Table 1 for demographics). For additional details on the YAC methods, see the study by Gordon et al [39].

XSL•FO
RenderX

Table . Demographic characteristics of the YAC[a] participants, including age, race or ethnicity, gender identity, and sexual orientation. The YAC comprised 15 LGBTQ+[b] adolescents (aged 15-18, mean 16.9 years) recruited from across the United States between January and July 2023 via paid social media advertisements and an institutional participant registry. The sample included 9 (60%) gender-diverse teenagers and 6 (40%) teenagers of color. All participants identified as sexual minorities, with the majority identifying as bisexual, pansexual, or gay.

Category	Count (n=15)
Age (years), n (%)	
15	1 (7)
16	3 (20)
17	8 (54)
18	3 (20)
Race or ethnicity, n (%)	
White	9 (60)
American Indian or Alaska Native	2 (13)
Black or African American	1 (7)
Asian	1 (7)
Mestizo Latin American	1 (7)
Not disclosed	1 (7)
Hispanic or Latino	3 (20)
Gender identity, n (%)	
Cisgender man or boy	6 (40)
Transgender man or boy	3 (20)
Transgender woman or girl	4 (27)
Gender nonconforming	1 (7)
Questioning gender identity	1 (7)
Sexual orientation, n (%)	
Bisexual or pansexual	7 (47)
Gay	5 (35)
Queer	2 (14)
Questioning	1 (7)

[a]YAC: Youth Advisory Council.

[b]LGBTQ+: lesbian, gay, bisexual, transgender, and queer.

Participatory Feedback Process With LGBTQ+ Teenagers

The YAC is a standing, Discord-based advisory group of LGBTQ+ teenagers who provide structured input across multiple digital health projects. Before the Roo evaluation, participants reviewed intervention materials and were advised on website features, familiarizing them with research goals and developing confidence as co-designers. The Roo evaluation was therefore embedded within a broader program of participatory engagement, with inclusivity, usability, and perceived helpfulness as central evaluation criteria.

The participatory feedback process was informed by a research-through-design orientation study, which treats interventions as iterative, generative artifacts shaped through ongoing engagement and inquiry [40]. Rather than approaching Roo as a static digital tool to be evaluated in isolation, we conceptualized it as a dynamic prototype whose development was guided by LGBTQ+ teenagers' feedback and contextualized within their lived experiences. This orientation supported creativity by encouraging teenagers to propose new content areas (eg, gender-affirming care and sex toy safety) and interface redesigns (eg, category-based navigation and dark mode), which moved beyond minor adjustments to reimagining Roo's functionality. It also enhanced responsiveness by enabling rapid incorporation of feedback into both the SHER-specific and public-facing versions of Roo, ensuring that participants saw their input translated into visible changes. Finally, research through design positioned social and structural factors (eg, stigma in school-based sex education, barriers to gender-affirming care, and mistrust of AI systems) as central design considerations, shaping not only what content was added but also how information was framed to reduce harm and foster trust.

XSL•FO
RenderX

To prepare for the use of Roo for the intervention, the Northwestern University team and PPFA were engaged over a period of 5 months to create a Northwestern University–specific instance of Roo that would be used in the intervention. This included establishing a data use agreement, removing the health educator feature, which could be considered contamination with RCT outcomes given its behavioral component, and adding study contact information. The Northwestern University instance of Roo was shared with the YAC.

The YAC provided feedback on all components of the SHER intervention design and implementation over approximately 18 months and was focused on providing Roo feedback over 1 week. As part of this evaluation, our analysis was guided by three research questions: (1) What do LGBTQ+ teenagers perceive as gaps and/or limitations in Roo's current content, particularly regarding inclusivity and relevance to their lived experiences? (2) How do LGBTQ+ teenagers evaluate the usability, tone, and format of Roo's chatbot interface? (3) Under what conditions do LGBTQ+ teenagers find chatbots such as Roo helpful or unhelpful for accessing sexual health information?

To address these questions, participants were invited to interact with Roo independently and provide feedback on its usability, inclusivity, and perceived helpfulness.

Over the course of 1 week, the research team posted 3 open-ended questions in the Discord server to guide feedback. These questions prompted participants to (1) explore Roo and share initial impressions; (2) reflect on inclusivity, especially for LGBTQ+ and transgender and gender-diverse users, and suggest improvements to Roo's interface and structure; and (3) identify additional topics Roo should cover more effectively, such as sexual assault, miscarriage, sex toys, and gender inclusivity. Participants had 7 days to respond to any or all questions at their own pace.

Each question was posted in the main research channel and was accompanied by a dedicated discussion thread to organize all responses. Participants provided written replies, reacted to peers' comments using emojis, and engaged in follow-up exchanges within threads. Emoji reactions were used as indicators of consensus or shared resonance across YAC members.

Data for this study consisted of text responses, emoji reactions, and threaded follow-up conversations. After data collection, raw Discord content was organized into summaries by question, capturing recurring themes and illustrative quotes. In line with member-checking practices, summaries were shared with YAC members for review before being provided to the Roo team [41].

Data Analysis

We used rapid qualitative analysis (RQA) to conduct content analysis to synthesize participant insights across our 3 research questions. RQA is an increasingly adopted approach in applied health research contexts, emphasizing structured data summarization over line-by-line coding to generate timely, actionable findings [42]. RQA is particularly well-suited for participatory projects that seek to elevate community voices while maintaining momentum in design iteration cycles [43].

Consistent with the content analysis approach, our RQA focused on a close reading of the text to identify substantive feedback that could inform actionable design improvements rather than developing emergent themes through open coding [44]. As such, our code development process prioritized capturing concrete, implementable recommendations within a time-bounded participatory design framework, an approach suitable for applied health intervention research where timely translation is essential.

The analysis was conducted collaboratively between members of the Northwestern University research team and PPFA's digital content strategy team to ensure findings were both academically rigorous and actionable for implementation. The analytic team included the lead author (WWL, MSW, and a doctoral student with training in qualitative methods and LGBTQ+ health), a research assistant professor with expertise in implementation science and LGBTQ+ adolescent health (EC, PhD), and a research project manager with experience in community-based participatory research (AAA, MPH).

The lead author conducted the initial analysis, organizing participant feedback into a matrix where each row represented a unique participant and each column corresponded to 1 of the 3 research questions. Feedback was organized to identify substantive codes such as specific critiques, suggestions, and design recommendations that could inform actionable improvements. Responses that addressed multiple research questions were mapped across domains to reflect the interrelated nature of the data. The lead author then conducted a systematic review of the coded content to identify recurring patterns, frequently mentioned issues, and high-priority recommendations, revisiting the data for each research question domain and identifying the most actionable codes grounded in participant language. This process emphasized iterative refinement, prioritizing content with the strongest support and clearest implementation pathways and was conducted until no new substantive actionable content codes were identified. Particular attention was given to points of consensus—where feedback was echoed or affirmed across participants—as well as divergence, where responses reflected differing or underrepresented perspectives. Agreement indicators such as "likes" and emoji reactions were systematically documented as they signaled areas of shared resonance and helped prioritize implementation efforts.

The preliminary analysis was reviewed collaboratively by the Northwestern University analytic team (WWL, EC, and AAA) to ensure accuracy, comprehensiveness, and alignment with the research questions. Following this internal review, the refined findings were shared with PPFA's Associate Director of Digital Content Strategy (AT), who brought expertise in sexual health education, digital product development, and youth engagement. The cross-organizational team then conducted a collaborative review to (1) validate themes and ensure they accurately reflected user feedback and (2) translate findings into concrete, implementable content and design changes. PPFA's implementation process systematically prioritized changes based on frequency of feedback, level of participant endorsement (indicated by likes or reactions), feasibility, and alignment with organizational capacity for rapid implementation. This

XSL•FO
RenderX

partnership approach directly informed real-world product improvements.

Results

Overview

The YAC provided feedback across three domains mapping to our research questions: (1) content gaps and relevance, (2) design and usability, and (3) trust and AI transparency.

RQ1: Content Gaps, Depth, and Relevance

Content Depth and Specificity

YAC participants consistently emphasized the need for Roo to go beyond foundational sexual health content and engage more meaningfully with topics specific to LGBTQ+ teenagers. While Roo provided accurate general information, participants found its responses insufficiently nuanced for addressing the kinds of "medium-depth" questions they were most likely to ask. These questions, while not highly clinical, often reflected real-life scenarios central to their identities and relationships. Several participants specifically flagged the absence of content on sexual assault, miscarriages, the safe use of sex toys, and queer-specific sexual practices. One highly endorsed comment by 5 other council members summarized this concern:

> *Why doesn't Roo have any clickable question buttons about sexual assault, sex toys, and miscarriages.* [P1, 5 likes]

This reflects a broader expectation among teenagers that a credible sexual health tool should anticipate sensitive topics and proactively make them accessible, rather than placing the burden on teenagers to generate the correct wording. Others echoed this sentiment, highlighting that Roo did not adequately respond to moderately detailed prompts related to LGBTQ+ health, with 1 participant noting, "For a lot of medium-depth questions even related [to sexual health], it didn't have a reply" (P4, 4 likes).

Gender-Affirming Care and Transgender Inclusion

A particularly salient gap concerned Roo's limited content on transgender and nonbinary health needs. Participants described the absence of information on gender-affirming care (eg, hormone therapy, safe binding practices, and navigating health care as a transgender person) as both alienating and harmful. As P9 explained,

> *It's supposed to be for transgender people like myself… but it's not so inclusive, which would deter me from reading further most times.* [P9]

This sense of exclusion was compounded by the broader difficulty teenagers described in accessing accurate, affirming sexual health information elsewhere. As 1 participant put it, "It's so hard to get accurate queer transgender sex ed" (P1, 6 likes), highlighting that Roo's shortcomings do not exist in isolation but within an already sparse landscape of transgender-inclusive health resources.

Taken together, these responses reveal that transgender-specific content is not a secondary feature but a core component of trust-building for this population. Its absence signals that the platform may not fully see or support them. Including gender-affirming information is therefore not only about comprehensiveness, but it is also about ensuring that transgender teenagers can see their realities reflected in the resource and feel affirmed and empowered to engage with it for their sexual health needs.

Reducing Stigma in Health Information

Participants emphasized the importance of presenting sexual health information in ways that actively reduce stigma, particularly around STIs and developmental norms. Teenagers called for content that normalizes diverse experiences rather than pathologizing them. As one participant noted, "We need info that gets rid of stigma on STIs" (P1, 2 likes), underscoring the need for nonjudgmental, affirming messaging.

At the same time, participants highlighted how even well-intentioned information could inadvertently reinforce stigma when framed without sufficient context. P3 reflected:

> *It just completely negates the ' oh it's okay you can take your time' message… it may also pressure someone to have [sex] by 17, which is awful.* [P3, 3 likes]

Rather than reassuring teenagers that there is no single timeline for sexual experiences, content that references normative milestones without careful framing was seen as inadvertently pressuring or alienating. Participants also expressed frustration with overly vague or evasive responses to personal questions. As P4 explained, "If you're gonna say something like 'everybody is different,' then I'd rather just not get an answer… I'd like to see statistics… instead of just being told it depends" (7 likes).

Such comments highlight the risk of vague responses not only being unhelpful but also eroding trust in the platform's authority. These insights suggest that effective stigma reduction requires more than inclusive language. It demands concrete, contextualized information that normalizes a range of sexual health experiences and empowers teenagers to feel validated, rather than judged, in their individual journeys.

RQ2: Design and Usability

Conversational Functionality and Interface

Participants had mixed reactions to Roo's chat-based format. While some appreciated its friendly, conversational tone, others found it cumbersome for retrieving information efficiently. Some teenagers likened the chatbot to a "complicated FAQ" (P8, 3 likes) where the conversational structure introduced unnecessary layers of interaction to access straightforward answers. As 1 participant explained:

> *It doesn't really feel like a chatbot if you're limited to pre-written questions.* [P4]

To address these concerns, participants proposed alternative interface designs that would offer greater flexibility, speed, and user control. Several teenagers suggested repositioning the chat input box to the bottom of the screen, aligning with familiar web design conventions and streamlining user interactions. Others advocated for adding features such as dark mode to

XSL•FO
RenderX

enhance visual accessibility, particularly for users who prefer reduced screen brightness during extended sessions. The idea of categorizing topics into a dropdown menu structure (rather than relying solely on "popular questions") was seen as critical to improving navigation and reducing the frustration associated with multiple conversational turns.

One participant (P9) proposed a comprehensive redesign that combined esthetic appeal with functionality, illustrating their vision through a self-created layout (Figure 3). Their design emphasized the importance of reducing friction in the user experience, advocating for organized topic categories (eg, "Sex>Vaginal sex>Resources"), easy-to-access FAQs, and a visually appealing, dark mode interface. Beyond esthetics, they stressed the need for Roo to maintain credibility by linking responses to trusted sources and providing external resources for users who wished to learn more. P9 explains:

> *People don't come to the AI like a social media platform—they come to ask questions about sexual health… a category dropdown that leads into a list of questions gives a more balanced system where everyone gets the same answer.* [P9]

XSL•FO
RenderX

Figure 3. Alternative interface mockup created by P9 for Roo. The proposed redesign features a hierarchical dropdown menu structure (eg, "sex"→"vaginal sex"→"resources") to improve navigation efficiency, dark mode for visual accessibility, a FAQ section, and an "ask Roo" chat function. This design reflects teenagers' critiques of Roo's original chat-based interface as cumbersome for information retrieval and their desire for categorized topic organization rather than conversational navigation. AI: artificial intelligence; FAQ: frequently asked question.



In addition to these interface suggestions, P9 proposed backend enhancements to strengthen Roo's reliability and responsiveness. These included using a GPT-4 model trained specifically on sexual health content, a "report AI" button for flagging problematic outputs, and an "ask an agent" feature to allow users to escalate complex or sensitive questions to trained human health educators. Collectively, these ideas reflect a broader, teenager-centered vision for Roo—not merely as a chatbot, but as an integrated, trustworthy health information platform designed to meet the expectations of digitally savvy teenagers. This emphasis on transparency, accessibility, and user agency echoed broader themes across YAC feedback, highlighting a shared desire for efficient, affirming, and user-friendly sexual health resources.

XSL•FO
RenderX

Inclusive and Accurate Language Use

Language emerged as a significant concern, particularly with some responses defaulting to cisgender or heteronormative assumptions. For instance, a phrase such as "Does it hurt for girls to have sex for the first time?" was flagged as exclusionary by participants, who noted that such wording could alienate gender-diverse users. Teenagers highlighted the importance of using gender-neutral language to ensure that Roo feels inclusive and affirming for all users, regardless of gender identity.

> *I think it's important it be able to answer questions even if they're phrased in a gender essentialist way–a young cis girl is probably going to ask something like that rather than ' does it hurt to have sex for the first time if you have a vagina.'* [P6, 4 likes]

As P6 noted, this included rephrasing responses to be both inclusive and accurate, such as changing the example to "Is it painful for people with vaginas to have sex for the first time?" Incorporating such thoughtful language choices was seen as a key step toward making the tool more accessible and welcoming for LGBTQ+ teenagers, while still providing accessible answers for cisgender heterosexual teenagers.

RQ3: Transparency, Trust, and AI Literacy

Transparency and Limitations of Roo's AI Capabilities

Participants expressed mixed perceptions of Roo's trustworthiness as an AI-driven sexual health tool. While many teenagers appreciated Roo's friendly, nonjudgmental tone—describing it as more approachable than other online sources—concerns emerged about its perceived lack of adaptability and clarity about its underlying technology. Some participants found responses too generic or static, raising doubts about Roo's ability to handle complex or highly individualized questions. Others were unsure whether they were interacting with an AI, a scripted FAQ, or a live human, which undermined their confidence in the chatbot. As 1 participant explained:

> *AI learning models are still young and may not always have what you're looking for, so I think that a live agent would be a good addition for more complex questions. They would be trained to provide accurate factual information about more complex questions that Roo could not.* [P9]

Teenagers emphasized the need for Roo to be more explicit about its AI nature and its limitations, advocating for clear disclosures, transparent sourcing, and pathways to escalate questions beyond the chatbot when needed. The ambiguity surrounding Roo's status, paired with inconsistent depth in responses, left some users hesitant to trust the platform fully.

AI Literacy and Critical Perspectives on AI Use

Beyond their critiques of Roo specifically, participants demonstrated a striking degree of AI literacy and sophistication in evaluating the role of conversational AI in sexual health education. Teenagers engaged in nuanced discussions comparing Roo to general-purpose large language models (LLMs) such as ChatGPT. Some participants expressed enthusiasm about ChatGPT's conversational style, with 1 participant remarking, "Yeah I think almost everyone has a positive attitude to ChatGPT. I like it so much. It's been really helpful" (P10, 2 likes). Others, however, voiced caution about LLMs' well-documented tendency to produce plausible-sounding but inaccurate information. As 1 participant warned, "It does literally just lie sometimes… if the AI gives out misleading information, it's going to have a lot of bad effects on the people who use the service" (P8, 6 likes). Rather than passively consuming information, participants positioned themselves as critical evaluators of AI outputs, demonstrating awareness of risks and suggesting safeguards. Several teenagers even articulated concrete design pathways for improving Roo, imagining how a model fine-tuned on credible sources could balance depth with accuracy.

> *If trained specifically on trusted sources… it could be, while not perfect, essentially very in-depth GPT—but rather than general questions, it could be way more in-depth on sexual health.* [P9]

These discussions revealed a nuanced tension: while teenagers were receptive to the idea of more dynamic, conversational AI, they also demanded safeguards such as expert oversight, transparent sourcing, and clear escalation pathways to human support. Their critical perspectives suggest that LGBTQ+ teenagers are not only capable of evaluating emerging technologies but are poised to play a vital role in shaping safer, more accountable AI-driven health resources.

These critiques not only highlighted teenagers' ability to interrogate the promises and risks of AI but also offered a roadmap for actionable improvements. In particular, their emphasis on the need for clear disclosures, sourcing, and escalation pathways informed Roo's subsequent revisions around transparency and live educator integration. Teenagers' calls for AI trained on trusted content also underscored the importance of expanding Roo's database and linking responses to vetted resources, changes that Planned Parenthood prioritized in the next iteration of the platform.

Implemented Changes

In response to YAC feedback, PPFA implemented targeted modifications to Roo between May 2023 and October 2024. Changes fell into 3 primary categories: content expansion, language and tone revision, and transparency and interface improvements.

Content Expansion

For transgender and gender-affirming care, PPFA added 15 new responses covering testosterone therapy, chest binding, and navigating health care as a transgender person. These included practical guidance (eg, "Taking testosterone can make your period lighter, less frequent, or make them totally go away") and access information (eg, "Contact your nearest Planned Parenthood health center to see if they offer gender-affirming hormone therapy").

For queer-specific sexual practices, new content addressed anal health and safety, clitoral anatomy and function, and sexual practices such as scissoring, providing both pleasure-focused information and safety guidance. Additional entries covered topics that teenagers flagged as absent or insufficient, including sex toy safety, miscarriage, and sexual assault resources.

XSL•FO
RenderX

To address concerns about vague or overly general responses, PPFA incorporated more concrete statistics and contextualized examples throughout Roo's responses. This shift responded directly to participants' calls for information that normalizes diverse sexual health experiences while providing actionable, specific guidance rather than platitudes such as "everybody is different."

Language and Tone Revisions

PPFA systematically revised Roo's language to eliminate gendered assumptions and heteronormative framing while emphasizing affirmation and stigma reduction. Gender-essentialist prompts were rewritten to use gender-neutral phrasing (eg, changing "Does it hurt for girls to have sex for the first time?" to "Is it painful for people with vaginas to have sex for the first time?"). This approach ensured that responses remained accessible to cisgender users while affirming gender-diverse teenagers.

STI-related content was updated to explicitly include diverse sexual practices and partnership types, moving beyond heteronormative assumptions. Tone modifications shifted from clinical risk language to affirming, nonjudgmental statements. For example, anal sex information was reframed to state "Anal sex is healthy when done safely. Using lube, going slow, and avoiding anal sex when you have hemorrhoids can help," positioning pleasure and safety as compatible rather than opposing concerns. Similarly, STI testing was normalized as routine self-care: "Getting tested is often as easy as peeing in a cup at a health center."

These revisions addressed participants' concerns that even well-intentioned content could inadvertently reinforce stigma through framing that emphasized risk, pathology, or normative developmental timelines without adequate context.

Transparency and Interface Improvements

To address concerns about Roo's ambiguous identity and perceived limitations, PPFA implemented several transparency-focused changes. Roo's introductory text was updated to more clearly identify it as an AI-powered tool with specific capabilities and constraints, directly responding to participants' confusion about whether they were interacting with a live person, a scripted FAQ, or an AI system.

Recognizing teenagers' desire for human escalation pathways, PPFA added alternative prompts (eg, "Can I talk to a real person?" and "I want to chat to a human") to facilitate seamless connection to live health educators through Planned Parenthood's chat services. While this live chat feature has always been part of Roo's broader infrastructure, it was temporarily removed from the version used during the Northwestern University study due to research constraints. Nevertheless, participants' strong interest in live support, evidenced by proposals for "ask an agent" features, reaffirms its value as a critical complement to AI-driven answers for handling complex or sensitive questions.

Further efforts were made to increase Roo's credibility and transparency through the inclusion of "learn more" links within responses, guiding users to vetted Planned Parenthood materials or trustworthy external sites. This addressed teenagers' concerns about insufficient sourcing and their desire to verify information or explore topics in greater depth.

While structural changes such as adding a full category-based dropdown system, integrating dark mode, or upgrading to more advanced backend AI models (as suggested by P9) remain under consideration for future development phases, PPFA prioritized user experience improvements that could be implemented rapidly. Conversational pathways were streamlined to reduce the number of steps needed to access frequently asked questions, and the language in suggested prompts was reviewed to avoid heteronormative assumptions.

These enhancements reflect Roo's continued evolution into a transparent and user-centered tool, one that better communicates its capabilities and limitations while expanding teenagers' access to reliable, affirming health information. Table 2 summarizes key areas of participant feedback alongside corresponding updates made to the platform, illustrating how LGBTQ+ teenager perspectives directly informed and improved this AI-driven health tool.

XSL•FO
RenderX

Table . Summary of YAC[a] feedback and corresponding updates to Roo. The table maps participant-identified limitations across 7 domains (content depth, transgender inclusion, stigma reduction, chatbot functionality, language inclusivity, AI[b] transparency, and AI literacy) to specific design and content changes implemented by PPFA[c].

Gap identified	Participant feedback	Implemented changes
Content depth and specificity	Need for more nuanced content on medium-depth questions (eg, sexual assault, sex toys, miscarriages, and queer-specific practices).	Added new entries on these topics to improve coverage of LGBTQ+[d] experiences.
Gender-affirming care and transgender inclusion	Lack of content on hormone therapy, binding, and transgender health care alienates transgender users.	Introduced content on testosterone, binding, and accessing affirming care.
Reducing stigma in health information	Content felt pathologizing or vague; need for nonjudgmental framing and clear statistics.	Revised tone to reduce stigma, included more examples or statistics to contextualize experiences.
Chatbot functionality	Chat felt like a complicated FAQ[e]; suggested dropdowns, dark mode, GPT-4 backend, and escalation to humans.	Streamlined conversational flow; future consideration for dropdown interface and backend upgrades.
Inclusivity and language use	Gendered or heteronormative language alienated users. Requested gender-neutral rephrasings.	Reviewed or revised language in prompts and answers to be gender-neutral and inclusive.
Transparency of AI identity	Confusion about whether Roo was AI or human reduced trust.	Updated introduction text to clarify Roo's AI identity and limitations.
AI literacy and oversight	Teenagers raised concerns about hallucinations and lack of sourcing; requested live agent escalation and better links.	Emphasized pathways to live health educators and added "learn more" links to vetted sources.

[a]YAC: Youth Advisory Council.

[b]AI: artificial intelligence.

[c]PPFA: Planned Parenthood Federation of America.

[d]LGBTQ+: lesbian, gay, bisexual, transgender, and queer.

[e]FAQ: Frequently Asked Questions.

These updates illustrate how teenager-led feedback can shape not just what information is delivered, but how it is structured, presented, and framed. Importantly, many of the changes implemented (eg, improvements to inclusive language, clearer sourcing, expanded topic coverage, etc) speak to both informational and relational aspects of trust-building in AI-based health communication.

Discussion

Principal Results

This study demonstrates how participatory design with LGBTQ+ teenagers can meaningfully shape the development of AI-powered sexual health tools. By partnering with PPFA to gather structured feedback on Roo, an AI-enabled chatbot, we engaged a YAC of LGBTQ+ teenagers in identifying gaps, proposing redesigns, and surfacing concerns about trust, inclusivity, and usability. Participants highlighted several key areas for improvement: the need for more in-depth and affirming content (particularly around transgender health, queer-specific sexual practices, and stigma reduction), more flexible and intuitive interface design, and greater transparency about Roo's limitations as an AI-driven system.

Teenagers also offered critical insight into the epistemic and ethical challenges posed by rule-based AI systems in health contexts, comparing Roo's limitations to the conversational fluidity and risks of LLMs such as ChatGPT. Rather than rejecting AI altogether, participants articulated a desire for hybrid systems that combined the responsiveness of AI with safeguards such as vetted content, human escalation pathways, and transparent sourcing. Their contributions not only informed direct improvements to Roo's content, interface, and tone but also pointed toward broader design principles for teenager-centered, accountable AI in health education.

Reimagining Trust and the Role of AI in Sexual Health Education

Teenagers' critiques of Roo centered on its perceived limitations as a rule-based system—an AI approach that relies on preprogrammed responses to anticipated queries rather than dynamically generating answers based on context. While many appreciated Roo's nonjudgmental tone, they described the platform as rigid, generic, and limited in depth, particularly when responding to nuanced or medium-complexity questions. Participants noted that while rule-based systems can offer predictability and safety, they often struggle to meet users' evolving or individualized information needs, especially in sensitive areas such as LGBTQ+ sexual health.

Several teenagers compared Roo to more dynamic tools such as ChatGPT, noting that while Roo felt safe and reliable, it sometimes lacked the adaptability or richness of a conversational AI. These critiques mirror broader limitations of digital health resources, which often lack the contextual responsiveness needed to address users' diverse needs, especially in areas such

XSL•FO
RenderX

as mental health and identity. Research has shown that many digital tools fail to provide accessible, personalized, or comprehensible information, limiting their effectiveness in real-world contexts [45]. Although static resources offer advantages in terms of accuracy and control, they often struggle to serve teenagers navigating sensitive, intersectional, or highly personalized sexual health concerns [46,47].

Yet, it is important to balance these concerns with evidence demonstrating AI's potential promise in health care. Systematic reviews show that conversational agents can effectively deliver health interventions when designed appropriately, with research indicating that chatbots can improve patient engagement, provide timely information access, and demonstrate acceptable accuracy [48]. Studies of user acceptance show that when people believe AI technology will be useful, this can serve as a strong predictor of adoption among LGBTQ+ individuals [49]. In contrast, dynamic and interactive tools (eg, chatbots or AI-based conversational agents) offer real-time responsiveness and perceived personalization, which may foster deeper engagement and affirmation, particularly for LGBTQ+ teenagers seeking information not readily reflected in mainstream health education [50].

Our findings extend this literature by demonstrating how trust is co-constructed through participatory design processes that enable users to contest and refine system outputs. While participants raised concerns about Roo's limitations, research demonstrates that conversational agents can be highly effective in health care when designed with transparency and accountability in mind [48,51]. The framework by Shneiderman [52] for trustworthy AI emphasizes the importance of human control and oversight, principles echoed in teenagers' proposals for "report AI" buttons and live agent escalation. Our findings extend this framework by demonstrating how trust emerges not from system perfection but from honest acknowledgment of constraints coupled with meaningful avenues for user input and recourse.

These expectations for responsiveness were closely tied to broader notions of trust. Existing research suggests that teenagers are more likely to trust AI health tools when systems clearly communicate their limitations, provide opportunities for human escalation, and demonstrate credibility through transparent sourcing [53]. Liao and Sundar [54] argue that designing for responsible trust in AI systems requires clear communication strategies that help users understand system capabilities and limitations. Trust was not predicated simply on whether a system used AI, but on how well it acknowledged uncertainty, supported user agency, and created channels for recourse when information felt insufficient or inaccurate. In this sense, trust and accountability were deeply intertwined: participants wanted tools they could rely on, not because they were infallible, but because they made space for users to push back, seek clarification, and surface unmet needs.

At the same time, participants demonstrated striking AI literacy. Teenagers critically discussed the strengths and risks of integrating LLMs such as ChatGPT into health tools, raising concerns about hallucinated facts, citation gaps, and the overconfidence of generative outputs. While some valued the conversational range offered by ChatGPT, others emphasized the potential harm of inaccurate information in high-stakes health contexts. Rather than rejecting AI altogether, participants proposed pragmatic solutions, such as fine-tuning models on verified health content, linking outputs to reputable sources, and integrating moderation or "report" features that would allow users to flag problematic responses and request additional review. These reflections signal that LGBTQ+ teenagers are not only capable of identifying epistemic risks but are actively imagining systems of accountability. Their feedback points toward a broader vision of participatory AI governance, one in which teenagers are positioned not as passive users, but as costewards of tools designed for their well-being.

Centering Teenage Voices to Advance Equity in AI-Enabled Health Tools

Roo is one of the few publicly available AI-powered chatbots designed to provide on-demand sexual health information to teenagers. Unlike static websites, Roo offers scalable, interactive engagement across a broad spectrum of questions, positioning it as a unique resource within an otherwise sparse landscape of LGBTQ+ inclusive digital health tools. Roo was developed in part to address the information marginalization faced by LGBTQ+ teenagers whose needs are often excluded from mainstream sexual health education. Through this project, Roo incorporated content specifically tailored to queer and transgender teenagers, including information on gender-affirming care, STI stigma, and mental health in sexual contexts. Yet despite these intentions, participants identified ongoing limitations in Roo's ability to represent and respond to the full scope of their lived experiences. Feedback highlighted not just content gaps, but structural constraints, such as rigid dialog flows and static language, that limited Roo's responsiveness, particularly for users with intersectional or stigmatized identities. These observations reinforced a critical lesson: content inclusivity is necessary, but not sufficient, for digital equity. The form and functionality of the system must also reflect the realities and needs of those it intends to serve.

To that end, Roo's participatory development process offers an important model for advancing equity in AI-based health interventions. Our engagement with the YAC gestures toward what Constanza-Chock [55] terms "design justice," in its effort to foreground the experiences of those most affected by the technology. LGBTQ+ teenagers were treated as co-designers whose insights materially shaped both the chatbot's interface and content. Their feedback drove refinements to language, layout, and tone, ensuring that Roo felt accessible, affirming, and useful to those it was built for. Participants' sophisticated critiques align with the observation by Yang et al [56] that human-AI interaction presents unique design challenges that require iterative, user-centered approaches. By positioning teenagers as epistemic contributors, we demonstrate how participatory methods can address the guidelines for human-AI interactions proposed by Amershi et al [57], particularly around transparency, user control, and system adaptability. These design contributions align with research showing that participatory methods, particularly when embedded early and iteratively, can produce tools that are not only more usable but also more trusted and ethically aligned [55]. Beyond improving Roo itself, the

XSL•FO
RenderX

process also demonstrated how co-design with marginalized teenagers can serve as a form of epistemic justice: making space for community knowledge to shape what counts as valid, visible, and valuable in digital systems.

Importantly, this process also cultivated a form of applied AI literacy among participants. Teenagers evaluated its limitations, questioned its logic, and proposed sophisticated alternatives, such as integrating GPT-4 with guardrails or flagging outputs for review. Engaging teenagers in the design process encouraged them to question the intelligence and trustworthiness of AI, equipping them with the skills to navigate the complexities of digital health technologies. This aligns with the growing recognition that AI literacy is essential for empowering marginalized populations to not only use AI tools effectively but also challenge their limitations [58]. By blending participatory design with transparent and inclusive AI practices, this study highlights a replicable model for fostering critical engagement with AI—one that enables LGBTQ+ teenagers to transition from passive users to informed, empowered stakeholders.

Privacy and Confidentiality Considerations

Interestingly, while participants raised concerns about content accuracy and trustworthiness, explicit concerns about data privacy were not prominent in their feedback. This finding contrasts with research suggesting that LGBTQ+ teenagers often prioritize anonymity and privacy in online health-seeking [59]. The relative absence of privacy concerns may reflect Roo's design as a web-based tool requiring no account creation or personal information, or it may suggest that content quality and inclusivity were more salient concerns than data practices. Roo's design incorporates privacy-protective features to encourage honest engagement: no account creation, login, or personal information is required, and users access the platform anonymously via web browser. However, this study did not directly assess whether teenagers experienced self-censorship despite these protections. Future research should examine whether privacy concerns shape the types of questions teenagers feel comfortable asking AI health tools.

However, this should not be interpreted as evidence that privacy is unimportant to LGBTQ+ teenagers accessing sexual health information. The broader literature on digital health emphasizes that privacy concerns intersect with identity safety and institutional trust, particularly for populations experiencing health inequities [60]. For LGBTQ+ teenagers who may face family rejection or school-based discrimination, privacy in health-seeking is not merely a preference but a safety imperative. Future research should explicitly probe privacy expectations and concerns among LGBTQ+ teenagers using AI health tools, particularly as these systems become more personalized and data-intensive.

Implications for Human-Centered AI, Health Informatics, and Public Health

This work contributes to human-computer interaction scholarship by demonstrating how participatory design with marginalized teenagers can surface unique requirements for AI systems that serve vulnerable populations. While guidelines for human-AI interaction emphasize transparency and user control, our findings reveal additional considerations specific to LGBTQ+ teenagers: the need for affirming language, intersectional content, and explicit acknowledgment of AI limitations [57]. These insights extend existing human-computer interaction frameworks and suggest that design guidelines must be iteratively refined through engagement with diverse user communities.

From a health informatics perspective, our findings corroborate observations that health chatbots require improvements in content quality and engagement depth, while extending this to show that effectiveness depends on attending to relational and cultural dimensions of care [61]. The emphasis participants placed on inclusive language, transgender-specific content, and stigma-reducing framing highlights that health information systems must address both informational accuracy and the social contexts in which health information is sought and used.

For public health practice, this study underscores AI's potential to address health disparities when development processes are actively centered around marginalized communities. However, realizing this potential requires sustained participatory engagement rather than 1-time consultation, as demonstrated by Madaio et al [62]. Co-designing checklists and interventions with affected communities helps organizations understand and address bias in AI systems while building trust directly reflected in our partnership model.

Lessons Learned From Academic and Nonprofit Partnership

This project exemplifies the value of cross-organizational collaboration in advancing equitable digital health tools, while also surfacing important lessons for future partnerships.

Complementary Organizational Capacities Accelerate Impact

Universities and nonprofits bring distinct but synergistic strengths. Academic institutions offer methodological expertise and research infrastructure; nonprofits bring public health infrastructure and capacity for rapid implementation. This partnership demonstrated how combining these capacities collapsed the research-to-practice timeline, with teenage feedback translated into platform updates within months rather than years.

Early Alignment on Values and Decision-Making Prevents Conflict

Establishing clear agreements about decision-making authority was essential. Northwestern University maintained control over research design and analysis; PPFA retained authority over content implementation. This division preserved both scientific integrity and organizational autonomy but required extensive upfront communication and shared commitment to health equity values. Organizations considering similar partnerships should assess both structural agreements and value compatibility early.

XSL•FO
RenderX

Partnership Infrastructure Requires Dedicated Resources and Sustained Engagement

While this model offers significant advantages, it demands substantial coordination that is often unfunded or underestimated. Faculty time for relationship-building, legal negotiations, and ongoing communication represents real costs, as does nonprofit capacity for research collaboration that may compete with direct service priorities.

Implications for Practice and Future Directions

This study highlights both the promise and complexity of developing AI-enabled health education tools for LGBTQ+ teenagers. Roo demonstrates how chatbots can expand access to affirming, immediate, and nonjudgmental information, particularly for teenagers navigating sensitive, stigmatized, or nuanced questions. For sexual health education in particular, questions often defy simple rule-based answers and demand context, empathy, and nuance. Our findings suggest that participatory design should not end once a tool is launched; instead, teenagers must be meaningfully engaged in ongoing feedback, evaluation, and refinement cycles. Designing for accountability means enabling mechanisms for teenagers to contest, correct, or expand system responses, and treating these interventions not as critiques to be managed, but as vital contributions to more inclusive, responsive AI systems.

This study also reinforces the value of embedding participatory design as a continuous, rather than episodic, element of tool development. LGBTQ+ teenagers brought critical perspectives that shaped Roo's language, tone, content depth, and usability, insights that could not have emerged without their sustained involvement. Importantly, the co-design process also cultivated AI literacy, empowering teenagers to assess the risks and capabilities of emerging technologies. As generative AI tools proliferate in health and education settings, there is an urgent need for implementation models that support participatory governance, foster critical engagement, and include marginalized teenagers not only as users but as epistemic contributors. Future efforts might explore how such participatory frameworks can be institutionalized across digital health product life cycles, from ideation to deployment to postlaunch evaluation.

Scalability will also require stronger cross-sector partnerships. This project benefited from the combined expertise of a public health nonprofit and a university-based research team, enabling agile iteration while preserving scientific rigor. Finally, more research is needed to evaluate the long-term impact of AI-based sexual health tools on teenage health outcomes, particularly among queer and transgender teenagers. Future studies should explore not only what works, but for whom, under what conditions, and with what ethical guardrails in place.

Limitations

While our YAC included participants across gender identities and races or ethnicities, the sample was limited. Of 15 participants, 9 (60%) identified as White, suggesting recruitment methods may have inadvertently excluded some racial and ethnic minority teenagers whose perspectives may differ regarding institutional trust and health care bias. The sample size was relatively small, and participants self-selected into this study, introducing potential bias toward those already motivated to engage with sexual health tools. While we examined how gender identity shaped experiences with Roo, we did not systematically analyze how histories of gender-based discrimination—including transphobia and transmisogyny—influenced teenagers' trust in AI health tools. Future research should intentionally oversample LGBTQ+ teenagers of color using culturally responsive methods and explicitly examine how systemic marginalization informs interactions with AI platforms.

The customized instance of Roo used during this study differed from the publicly available version—most notably by removing the live chat feature due to research-related monitoring constraints. This may have influenced perceptions of trust, usability, and comprehensiveness. Additionally, this study did not capture or analyze individual conversation logs between participants and Roo, instead relying on participants' reflections posted in Discord discussions. While this preserves privacy and reduces data burden, future studies might benefit from analyzing actual conversation data (with consent) to identify specific interaction patterns or content gaps not fully captured through retrospective self-report.

Although our YAC discussions did not capture which specific sources teenagers consider trustworthy, prior work suggests that teenagers often place greater confidence in well-established health organizations, including the Centers for Disease Control and Prevention and the World Health Organization, when information is perceived as transparent, relevant, and easy to understand [63]. At the same time, marginalized teenagers may be more likely to trust organizations with reputations for inclusive and affirming care, such as Planned Parenthood, which align more closely with their lived experiences and needs [64]. Future research should directly examine which health information sources LGBTQ+ teenagers consider credible, to better inform how chatbots such as Roo can direct users to external resources.

Finally, while our academic and nonprofit partnership enabled rapid feedback translation and impact, it also surfaced challenges around legal, operational, and infrastructural coordination. Negotiating data use, content ownership, and design authority required extensive collaboration, and the scalability of this model may depend on the capacity and alignment of future partnering institutions. Nonetheless, this case offers a valuable proof of concept for how cross-sector partnerships can meaningfully center community voices in AI health tool development.

Conclusions

This project illustrates the importance of designing AI-enabled health tools not just for marginalized teenagers, but with them—grounding development in participatory design, cross-sector collaboration, and iterative accountability. Roo's evolution through feedback from LGBTQ+ teenagers highlights the critical need to center those most affected by information gaps in shaping both the content and infrastructure of digital health tools. More than just expanding access to stigma-free sexual health information, Roo's development process surfaced deeper questions about how AI systems earn trust, respond to critique, and evolve alongside community needs. By treating

XSL•FO
RenderX

teenagers as epistemic contributors, this work offers a model for how AI can be governed more ethically, transparently, and equitably. The collaboration between Northwestern University and Planned Parenthood demonstrates how academic and public partnerships can meaningfully integrate community voices into scalable interventions that are not only technically sound but socially accountable.

Acknowledgments

We gratefully acknowledge our Teen Health Lab Youth Advisory Council members for their thoughtful insights and continued engagement with us.

Funding

Research reported here was supported by the National Institute of Mental Health under award numbers R01MH129207 (principal investigator: Macapagal). JDG's time was supported by T32MH130325 (principal investigator: Newcomb). The funding agencies had no role in the research or manuscript preparation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

Data from this study may be shared with qualified researchers upon reasonable request to the corresponding author, contingent upon execution of appropriate data sharing agreements that maintain participant privacy and confidentiality protections.

Conflicts of Interest

None declared.

References

1. CDC. Estimated HIV incidence and prevalence. HIV Data. 2025 Feb 7. URL: https://www.cdc.gov/hiv-data/nhss/estimated-hiv-incidence-and-prevalence.html [accessed 2026-01-30]
2. Charlton BM, Everett BG, Light A, et al. Sexual orientation differences in pregnancy and abortion across the lifecourse. Womens Health Issues 2020 Mar;30(2):65-72. [doi: 10.1016/j.whi.2019.10.007]
3. Langenderfer-Magruder L, Walls NE, Whitfield DL, Brown SM, Barrett CM. Partner violence victimization among lesbian, gay, bisexual, transgender, and queer youth: associations among risk factors. Child Adolesc Soc Work J 2016 Feb;33(1):55-68. [doi: 10.1007/s10560-015-0402-8]
4. Gowen LK, Winges-Yanez N. Lesbian, gay, bisexual, transgender, queer, and questioning youths' perspectives of inclusive school-based sexuality education. J Sex Res 2014;51(7):788-800. [doi: 10.1080/00224499.2013.806648] [Medline: 24003908]
5. Goldfarb ES, Lieberman LD. Three decades of research: the case for comprehensive sex education. J Adolesc Health 2021 Jan;68(1):13-27. [doi: 10.1016/j.jadohealth.2020.07.036] [Medline: 33059958]
6. Guse K, Levine D, Martins S, et al. Interventions using new digital media to improve adolescent sexual health: a systematic review. J Adolesc Health 2012 Dec;51(6):535-543. [doi: 10.1016/j.jadohealth.2012.03.014] [Medline: 23174462]
7. Proulx CN, Coulter RW, Egan JE, Matthews DD, Mair C. Associations of LGBTQ-inclusive sex education with mental health outcomes and school-based victimization in US high school students. J Adolesc Health 2019;64(5):608-614. [doi: 10.1016/j.jadohealth.2018.11.012]
8. Delmonaco D, Harper G, Li S, Paneda C, et al. Information needs for an online resource for LGBTQ+ young people: mental health, sexual health, and navigating services. Currents: J Diversity Scholarship Soc Change 2022;2(1). [doi: 10.3998/ncidcurrents.1774]
9. Hobaica S, Kwon P. "This Is how you hetero:" sexual minorities in heteronormative sex education. Am J Sex Educ 2017 Oct 2;12(4):423-450. [doi: 10.1080/15546128.2017.1399491]
10. Wadham E, Green C, Debattista J, Somerset S, Sav A. New digital media interventions for sexual health promotion among young people: a systematic review. Sex Health 2019 Apr;16(2):101-123. [doi: 10.1071/SH18127] [Medline: 30819326]
11. Steinke J, Root-Bowman M, Estabrook S, Levine DS, Kantor LM. Meeting the needs of sexual and gender minority youth: formative research on potential digital health interventions. J Adolesc Health 2017 May;60(5):541-548. [doi: 10.1016/j.jadohealth.2016.11.023] [Medline: 28108088]
12. Benkeser D, Horvath K, Reback CJ, Rusow J, Hudgens M. Design and analysis considerations for a sequentially randomized HIV prevention trial. Stat Biosci 2020 Dec;12(3):446-467. [doi: 10.1007/s12561-020-09274-3] [Medline: 33767798]
13. Schnall R, Kuhns LM, Pearson C, et al. Efficacy of MyPEEPS mobile, an HIV prevention intervention using mobile technology, on reducing sexual risk among same-sex attracted adolescent males: a randomized clinical trial. JAMA Netw Open 2022 Sep 1;5(9):e2231853. [doi: 10.1001/jamanetworkopen.2022.31853] [Medline: 36129712]

XSL•FO
RenderX

14. Martin P, Alberti C, Gottot S, Bourmaud A, de La Rochebrochard E. Young people's proposals for a web-based intervention for sexual health promotion: a French qualitative study. BMC Public Health 2023 Jul 19;23(1):1389. [doi: 10.1186/s12889-023-16257-8] [Medline: 37468863]
15. Hengstler M, Enkel E, Duelli S. Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices. Technol Forecast Soc Change 2016 Apr;105:105-120. [doi: 10.1016/j.techfore.2015.12.014]
16. de Cock C, Milne-Ives M, van Velthoven MH, Alturkistani A, Lam C, Meinert E. Effectiveness of conversational agents (virtual assistants) in health care: protocol for a systematic review. JMIR Res Protoc 2020 Mar 9;9(3):e16934. [doi: 10.2196/16934] [Medline: 32149717]
17. Darien K, Lee S, Knowles K, et al. Health information from web search engines and virtual assistants about pre-exposure prophylaxis for HIV prevention in adolescents and young adults: content analysis. JMIR Pediatr Parent 2023 Jul 18;6:e41806. [doi: 10.2196/41806] [Medline: 37463044]
18. Bragazzi NL, Crapanzano A, Converti M, Zerbetto R, Khamisy-Farah R. The impact of generative conversational artificial intelligence on the lesbian, gay, bisexual, transgender, and queer community: scoping review. J Med Internet Res 2023 Dec 6;25:e52091. [doi: 10.2196/52091] [Medline: 37864350]
19. McAra-Hunter D. How AI hype impacts the LGBTQ + community. AI Ethics 2024 Aug;4(3):771-790. [doi: 10.1007/s43681-024-00423-8]
20. De Vito A, Colpani A, Moi G, et al. Assessing ChatGPT's potential in HIV prevention communication: a comprehensive evaluation of accuracy, completeness, and inclusivity. AIDS Behav 2024 Aug;28(8):2746-2754. [doi: 10.1007/s10461-024-04391-2] [Medline: 38836986]
21. Tomasev N, McKee KR, Kay J, Mohamed S. Fairness for unobserved characteristics: insights from technological impacts on queer communities. arXiv. Preprint posted online on Apr 28, 2021. [doi: 10.48550/ARXIV.2102.04257]
22. Teen and young adult perspectives on generative AI. : Hopelab; 2024 URL: https://assets.hopelab.org/wp-content/uploads/2024/05/Teen-and-Young-Adult-Perspectives-on-Generative-AI.pdf [accessed 2026-01-30]
23. Liem W, Berry A, Macapagal K. Reclaiming power over AI: equipping queer teens as AI designers for HIV prevention. arXiv. Preprint posted online on Jun 18, 2024. [doi: 10.48550/arXiv.2406.13018]
24. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. J R Soc Med 2011 Dec;104(12):510-520. [doi: 10.1258/jrsm.2011.110180] [Medline: 22179294]
25. Vaughn LM, Whetstone C, Boards A, Busch MD, Magnusson M, Määttä S. Partnering with insiders: a review of peer models across community-engaged research, education and social care. Health Soc Care Community 2018 Nov;26(6):769-786. [doi: 10.1111/hsc.12562] [Medline: 29512217]
26. Flicker S. Who benefits from community-based participatory research? A case study of The Positive Youth Project. Health Educ Behav 2008 Feb;35(1):70-86. [doi: 10.1177/1090198105285927]
27. Freire K, Pope R, Jeffrey K, Andrews K, Nott M, Bowman T. Engaging with children and adolescents: a systematic review of participatory methods and approaches in research informing the development of health resources and interventions. Adolescent Res Rev 2022 Sep;7(3):335-354. [doi: 10.1007/s40894-022-00181-w]
28. Gilbey D, Morgan H, Lin A, Perry Y. Effectiveness, acceptability, and feasibility of digital health interventions for LGBTIQ+ young people: systematic review. J Med Internet Res 2020 Dec 3;22(12):e20158. [doi: 10.2196/20158] [Medline: 33270039]
29. Ybarra ML, Prescott TL, Philips GL, Bull SS, Parsons JT, Mustanski B. Iteratively developing an mHealth HIV prevention program for sexual minority adolescent men. AIDS Behav 2016 Jun;20(6):1157-1172. [doi: 10.1007/s10461-015-1146-3] [Medline: 26238038]
30. Hingle M, Patrick H, Sacher PM, Sweet CC. The intersection of behavioral science and digital health: the case for academic-industry partnerships. Health Educ Behav 2019 Feb;46(1):5-9. [doi: 10.1177/1090198118788600] [Medline: 30041556]
31. Austin D, May J, Andrade J, Jones R. Delivering digital health: the barriers and facilitators to university-industry collaboration. Health Policy Technol 2021 Mar;10(1):104-110. [doi: 10.1016/j.hlpt.2020.10.003]
32. Planned Parenthood Chat. URL: https://roo.plannedparenthood.org/ [accessed 2026-01-30]
33. Jaccard J. The prevention of problem behaviors in adolescents and young adults: perspectives on theory and practice. J Soc Social Work Res 2016 Dec 1;7(4):585-613. [doi: 10.1086/689354]
34. 2024 PPFA annual report. : Planned Parenthood Federation of America; 2024 URL: https://www.plannedparenthood.org/uploads/filer_public/bf/a3/bfa35e39-14e5-43c3-8192-de42ed9309f1/2024-ppfa-annualreport-c3-digital.pdf [accessed 2026-01-30]
35. Proctor EK, Powell BJ, McMillen JC. Implementation strategies: recommendations for specifying and reporting. Implement Sci 2013 Dec 1;8(1):139. [doi: 10.1186/1748-5908-8-139] [Medline: 24289295]
36. Bauermeister JA, Pingel E, Zimmerman M, Couper M, Carballo-Diéguez A, Strecher VJ. Data quality in HIV/AIDS web-based surveys: handling invalid and suspicious data. Field Methods 2012;24(3):272-291. [doi: 10.1177/1525822X12443097]
37. Grey JA, Konstan J, Iantaffi A, Wilkerson JM, Galos D, Rosser BRS. An updated protocol to detect invalid entries in an online survey of men who have sex with men (MSM): how do valid and invalid submissions compare? AIDS Behav 2015 Oct;19(10):1928-1937. [doi: 10.1007/s10461-015-1033-y] [Medline: 25805443]

XSL•FO
RenderX

38. Teitcher JEF, Bockting WO, Bauermeister JA, Hoefer CJ, Miner MH, Klitzman RL. Detecting, preventing, and responding to “fraudsters” in internet research: ethics and tradeoffs. J Law Med Ethics 2015;43(1):116-133. [doi: 10.1111/jlme.12200] [Medline: 25846043]
39. Gordon J, Lorenzo J, Avila AA, et al. Using discord to engage LGBTQ+ youth in human-centered design of a digital mental health intervention: a methods case study (preprint). JMIR Formative Res. Preprint posted online on Jul 18, 2025. [doi: 10.2196/preprints.80852]
40. Zimmerman J, Forlizzi J, Evenson S. Research through design as a method for interaction design research in HCI. Proc SIGCHI Conf Hum Factors Comput Syst 2007 Apr 29:493-502. [doi: 10.1145/1240624.1240704]
41. Birt L, Scott S, Cavers D, Campbell C, Walter F. Member checking: a tool to enhance trustworthiness or merely a nod to validation? Qual Health Res 2016 Nov;26(13):1802-1811. [doi: 10.1177/1049732316654870] [Medline: 27340178]
42. Gale RC, Wu J, Erhardt T, et al. Comparison of rapid vs in-depth qualitative analytic methods from a process evaluation of academic detailing in the Veterans Health Administration. Implement Sci 2019 Feb 1;14(1):11. [doi: 10.1186/s13012-019-0853-y] [Medline: 30709368]
43. Vindrola-Padros C, Johnson GA. Rapid techniques in qualitative research: a critical review of the literature. Qual Health Res 2020 Aug;30(10):1596-1604. [doi: 10.1177/1049732320921835] [Medline: 32667277]
44. Bengtsson M. How to plan and perform a qualitative study using content analysis. NursingPlus Open 2016;2:8-14. [doi: 10.1016/j.npls.2016.01.001]
45. Robillard JM, Feng TL, Sporn AB, et al. Availability, readability, and content of privacy policies and terms of agreements of mental health apps. Internet Interv 2019 Sep;17:100243. [doi: 10.1016/j.invent.2019.100243] [Medline: 30949436]
46. Yonker LM, Zan S, Scirica CV, Jethwani K, Kinane TB. “Friending” teens: systematic review of social media in adolescent and young adult health care. J Med Internet Res 2015 Jan 5;17(1):e4. [doi: 10.2196/jmir.3692] [Medline: 25560751]
47. Widman L, Nesi J, Kamke K, Choukas-Bradley S, Stewart JL. Technology-based interventions to reduce sexually transmitted infections and unintended pregnancy among youth. J Adolesc Health 2018 Jun;62(6):651-660. [doi: 10.1016/j.jadohealth.2018.02.007] [Medline: 29784112]
48. Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018 Sep 1;25(9):1248-1258. [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
49. Henkel T, Linn AJ, van der Goot MJ. Understanding the intention to use mental health chatbots among LGBTQIA+ individuals: testing and extending the UTAUT. Chatbot research and design. In: Følstad A, Araujo T, Papadopoulos S, Law ELC, editors. Chatbot Research and Design: 6th International Workshop, CONVERSATIONS 2022, Amsterdam, The Netherlands, November 22–23, 2022 Revised Selected Papers; Lecture Notes in Computer Science: Springer; 2023:83-100. [doi: 10.1007/978-3-031-25581-6_6]
50. Das RK, Drolet BC. Assessment of artificial intelligence chatbot attitudes toward LGBTQ+ individuals. J Adolesc Health 2024 Jun;74(6):1264-1266. [doi: 10.1016/j.jadohealth.2024.02.030] [Medline: 38762249]
51. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. Digit Health 2019;5:2055207619871808. [doi: 10.1177/2055207619871808] [Medline: 31467682]
52. Shneiderman B. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Trans Interact Intell Syst 2020;10(4):26. [doi: 10.1145/3419764]
53. Park JK, Singh VK, Wisniewski P. Current landscape and future directions for mental health conversational agents for youth: scoping review. JMIR Med Inform 2025 Feb 28;13:e62758. [doi: 10.2196/62758] [Medline: 40053735]
54. Liao QV, Sundar SS. Designing for responsible trust in AI systems: a communication perspective. Proc 2022 ACM Conf Fairness, Accountability, Transparency 2022 Jun 20:1257-1268. [doi: 10.1145/3531146.3533182]
55. Costanza-Chock S. Design Justice: Community-Led Practices to Build the Worlds We Need: The MIT Press; 2020.
56. Yang Q, Steinfeld A, Rosé C, Zimmerman J. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. Proc 2020 CHI Conf Hum Factors Comput Syst 2020 Apr 23:1-13. [doi: 10.1145/3313831.3376301]
57. Amershi S, Weld D, Vorvoreanu M, et al. Guidelines for human-AI interaction. Proc 2019 CHI Conf Hum Factors Comput Syst 2019 May 2:1-13. [doi: 10.1145/3290605.3300233]
58. Long D, Magerko B. What is AI literacy? Competencies and design considerations. Proc 2020 CHI Conf Hum Factors Comput Syst 2020 Apr 23:1-16. [doi: 10.1145/3313831.3376727]
59. Ybarra ML, Mitchell KJ. A national study of lesbian, gay, bisexual (LGB), and non-LGB youth sexual behavior online and in-person. Arch Sex Behav 2016 Aug;45(6):1357-1372. [doi: 10.1007/s10508-015-0491-7] [Medline: 25894645]
60. Baum F, Newman L, Biedrzycki K. Vicious cycles: digital technologies and determinants of health in Australia. Health Promot Int 2014 Jun;29(2):349-360. [doi: 10.1093/heapro/das062] [Medline: 23144236]
61. Abd-alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M. An overview of the features of chatbots in mental health: a scoping review. Int J Med Inform 2019 Dec;132:103978. [doi: 10.1016/j.ijmedinf.2019.103978]
62. Madaio MA, Stark L, Wortman Vaughan J, Wallach H. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. CHI ’20 2020 Apr 23:1-14. [doi: 10.1145/3313831.3376445]
63. Kowitt SD, Schmidt AM, Hannan A, Goldstein AO. Awareness and trust of the FDA and CDC: results from a national sample of US adults and adolescents. PLOS ONE 2017;12(5):e0177546. [doi: 10.1371/journal.pone.0177546] [Medline: 28520750]

XSL•FO
RenderX

64. Magee JC, Bigelow L, Dehaan S, Mustanski BS. Sexual health information seeking online: a mixed-methods study among lesbian, gay, bisexual, and transgender young people. Health Educ Behav 2012 Jun;39(3):276-289. [doi: 10.1177/1090198111401384] [Medline: 21490310]

Abbreviations

AI: artificial intelligence
G2G: Guy2Guy
LGBTQ+: lesbian, gay, bisexual, transgender, and queer
LLM: large language model
PPFA: Planned Parenthood Federation of America
REDCap: Research Electronic Data Capture
RQA: rapid qualitative analysis
SHER: Sharing Health Education Resources
STI: sexually transmitted infection
YAC: Youth Advisory Council

Edited by A Mavragani; submitted 06.Jun.2025; peer-reviewed by A Adepoju, CL Coleman, GH Koo, JK Park, N Sun, R Seretlo; revised version received 02.Dec.2025; accepted 03.Dec.2025; published 19.Feb.2026.

Please cite as:
Liem WW, Casline E, Lorenzo J, Gordon JD, Avila AA, Taylor A, Levitz N, O'Keefe MC, Macapagal K
Enhancing LGBTQ+ Inclusivity in an AI-Powered Sexual Health Chatbot: User-Centered Design Approach Through a Nonprofit and Academic Partnership
J Med Internet Res 2026;28:e78621
URL: https://www.jmir.org/2026/1/e78621
doi:10.2196/78621

© William Wibowo Liem, Elizabeth Casline, Julianna Lorenzo, Jacob D Gordon, Andrés Alvarado Avila, Attia Taylor, Nicole Levitz, Michael C O'Keefe, Kathryn Macapagal. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Impact of a Health Coaching App on the Subjective Well-Being of Individuals With Multimorbidity: Mixed Methods Study

Isabelle Symes[1,2], MSc; Alexandra Burton[3], PhD; Daniela Mercado[4], PhD; Feifei Bu[1], PhD

[1]Department of Behavioural Science and Health, Institute of Epidemiology and Health Care, University College London, 1-19 Torrington Place, London, United Kingdom

[2]Centre for Preventive Neurology, Wolfson Institute of Population Health, Queen Mary University of London, London, United Kingdom

[3]Centre for Psychiatry and Mental Health, Wolfson Institute of Population Health, Queen Mary University of London, London, United Kingdom

[4]Holly Health, London, United Kingdom

Corresponding Author:
Feifei Bu, PhD
Department of Behavioural Science and Health, Institute of Epidemiology and Health Care, University College London, 1-19 Torrington Place, London, United Kingdom

Abstract

Background: Multimorbidity, the coexistence of 2 or more chronic conditions, is associated with poor well-being. Health coaching apps offer cost-effective and accessible support. However, there is a lack of evidence of the impact of health coaching apps on individuals with multimorbidity.

Objective: This study aimed to assess the impact and acceptability of a health coaching app (the Holly Health [HH] app) on the subjective well-being (SWB) of adults with multimorbidity.

Methods: This study used an explanatory-sequential mixed methods design, with quantitative secondary data analysis in the first phase and qualitative interviews in the second phase. In the quantitative phase (n=565), pre- and post-SWB (Office for National Statistics' 4 personal well-being questions [ONS4]) scores from existing app users with multimorbidity were analyzed using Bayesian growth curve modeling to assess the impact of HH. In the qualitative phase (n=22), data were collected via semistructured interviews and analyzed using reflexive thematic analysis. Mechanisms of action that supported SWB were categorized using the Multi-Level Leisure Mechanisms Framework.

Results: There was a significant increase in life satisfaction (Coef.=0.71, 95% highest density interval [HDI] 0.52 - 0.89), worthwhileness (Coef.=0.62, 95% HDI 0.43 - 0.81), and happiness (Coef.=0.74, 95% HDI 0.54 - 0.92) and a decrease in anxiety (Coef.=−0.50, 95% HDI −0.74 to −0.25) before and after using the HH app. Overall, 8 acceptable app features activated 5 mechanisms of action, including behavioral, psychological, and social mechanisms. Three additional factors influenced the acceptability of the health coaching app: type of chronic condition, availability of time, and the use of other support tools.

Conclusions: The study demonstrates that health coaching apps could be effective and acceptable support tools for individuals with multimorbidity. This study contributes to understanding why health coaching apps support SWB and could be used to inform the development of future digital health interventions in multimorbidity.

(J Med Internet Res 2026;28:e78738) doi:10.2196/78738

KEYWORDS

multimorbidity; health coaching app; digital health intervention; subjective well-being; mixed methods; mechanisms of action

Introduction

By 2035, it is projected that two-thirds of adults in England aged 65 years and older will have multiple chronic conditions, with a 50% increase in people developing 4 or more chronic conditions [1]. Multimorbidity is the coexistence of 2 or more long-term physical or mental health conditions [2]. It has been associated with key disease clusters including hypertension, diabetes, asthma, cancer, depression, and anxiety [3]. Multimorbidity is more prevalent among older adults, females, and those living in lower-income areas [4,5]. It is also influenced by socioeconomic (eg, education), behavioral (eg, smoking), psychosocial (eg, loneliness), and biological factors [6].

Due to the complex nature of multimorbidity, consequences range in severity and scope. Associated outcomes include premature mortality [7], disability [8], elevated frailty risk [9], and hospitalization [10]. Multimorbidity also contributes to economic burden, including increased health care use and costs [11]. The complex health care needs associated with multimorbidity management cannot be met sufficiently by current health care systems with a single disease focus [12]. There is a general lack of holistic support options, especially

XSL•FO
RenderX

an inattention to well-being in primary care consultations [13]. Emphasis on patient-centered approaches, where individuals can establish and work toward their own health and well-being goals, is essential to the effective management of multimorbidity [14]. Patient-centered care is a key component of the National Institute for Health and Care Excellence (NICE) guidance on multimorbidity [15].

Subjective well-being (SWB) is defined as an individual's thinking and feeling toward life as desirable [16]. It encompasses 3 dimensions: evaluative, eudaemonic, and affective [17,18]. Evaluative well-being refers to the overall assessment of satisfaction with life, eudaemonic well-being links to accounts of the meaning of life, and affective well-being reflects feelings and moods experienced every day [19]. Previous research has consistently shown the association between multimorbidity and poor well-being [6]. This is concerning, given SWB could play an important role in health maintenance and management. For example, increased SWB has been found to predict greater engagement in health-promoting behaviors [20], which has significant implications for multimorbidity management [21].

In the last few decades, there has been an increasing recognition of engagement in leisure activities, such as physical activity, as a protective factor against multimorbidity [22]. The Multi-Level Leisure Mechanisms Framework [23] suggests that engaging in leisure activities elicits various health benefits via psychological (eg, improved well-being), biological (eg, increased physical fitness), social (eg, increased social support), and behavioral (eg, increased motivation) mechanisms. This is supported by empirical research showing that regular exercise (behavioral mechanism) and strong social networks (social mechanism) are associated with a lower risk of multimorbidity [24], while increased social support (social mechanism) and lifestyle changes (behavioral mechanism) are associated with slower illness progression [25] and improved life expectancy [26].

Digital solutions can be used to promote these positive behavioral changes, such as increased engagement in physical activities, which can then subsequently enhance subjective and physical well-being, with health coaching apps offering cost-effective and accessible support options [27]. Health coaching apps are a form of digital health intervention (DHI), which are tools that use information and communication technologies for the improvement of health management, monitoring, prevention, treatment, and lifestyle [28]. Various studies have demonstrated the feasibility, acceptability, and effectiveness of the use of health coaching apps to support the self-management of common chronic conditions such as depression [29], diabetes [30], and hypertension [31]. However, there is a paucity of evidence exploring the mechanisms of their impacts. Moreover, evidence of their impacts on individuals with multimorbidity remains preliminary and limited, particularly with interventions that explicitly target SWB as a primary outcome [32]. To our knowledge, there are currently no studies specifically focusing on the impact of digital interventions on the SWB of people with multimorbidity. Therefore, this study focused on well-being, given its established association with multimorbidity and as it remains underexplored compared to more commonly studied outcomes (eg, quality of life).

This study aimed to assess the impact of a health coaching app on the SWB of people with multimorbidity. More specifically, it addressed the following research questions: (1) Does using a health coaching app have a positive impact on the SWB of people with multimorbidity? (2) If a positive impact is observed, which app features contribute to improved SWB for people experiencing multimorbidity and how do they support SWB? (3) Which factors influence the perceived acceptability of the app?

Methods

Holly Health App

This study is partnered with Holly Health (HH), a health coaching app supported by iOS, Android, and web [33]. Key features include daily chatbot support, nudges and reminders, mood, stress, and energy tracking, habit tracking, health reviews, tailored educational content, and short-term challenges (see Multimedia Appendix 1 for app images). These challenges typically involve 1-week goals (eg, mindfulness or physical activity challenges) that are ranked from easy to hard. The app targets well-being, sleep, exercise, and eating to promote healthy aging and support chronic condition management. Engagement with the app is user-directed, allowing individuals to interact with different features according to their needs and preferences rather than through a prescribed pathway. HH is supported by the UK National Health Service (NHS) Innovation Accelerator, which aims to support high-impact innovations that address NHS priorities [34]. Currently, the app is partnered with more than 200 general practitioner (GP) practices in the United Kingdom. Users typically access the app through GP practices, where they are invited via an SMS text message to download and use the app free of charge. HH is grounded in cognitive behavioral therapy (CBT) [35], acceptance and commitment therapy [36], mindfulness, and the small habits approach, based on frameworks such as the Capability-Opportunity-Motivation Behavior (COM-B) model [37]. While HH incorporates features to support chronic condition management, it was not specifically developed for people with multimorbidity.

Study Design

This study used an explanatory sequential mixed methods design [38], comprising a quantitative phase followed by the qualitative phase (see Figure 1). First, the quantitative phase involved secondary data analysis of repeated survey data (n=565). This was followed by the qualitative phase, where semistructured interviews were conducted with 22 app users. The two phases were conducted and reported sequentially, with integration occurring during the interpretation of findings to explain and expand on the quantitative results. This study defines multimorbidity as 2 or more chronic conditions. Guidance for secondary data analysis (STROSA: Standardized Reporting of Secondary Data Analyses [39]) and qualitative research (COREQ: Consolidated Criteria for Reporting Qualitative Research [40]) were followed.

XSL•FO
RenderX

Figure 1. Study design flowchart.

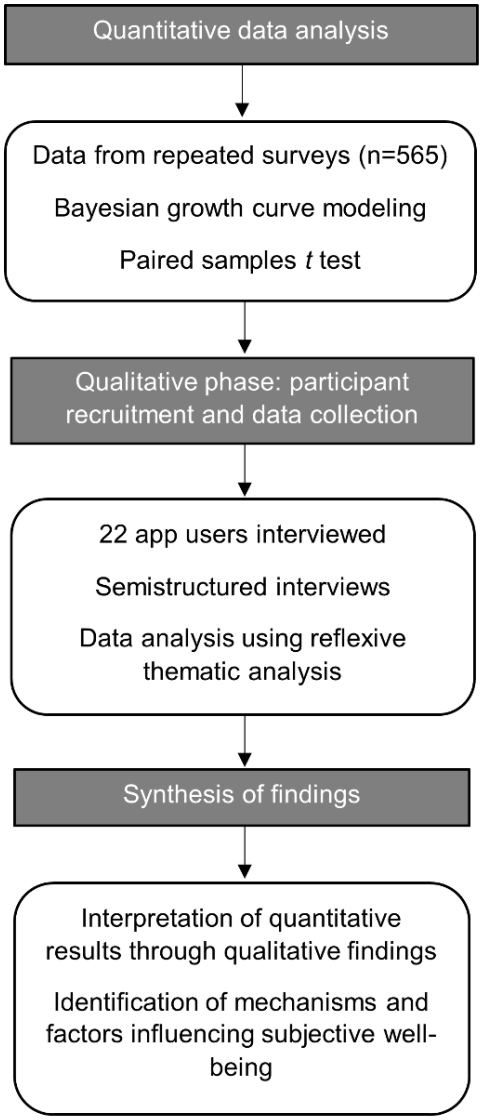

Ethical Considerations

The University College London (UCL) Research Ethics Committee approved this study (26737/001). All participants provided informed consent prior to participation, and all procedures were conducted in accordance with the ethical standards and institutional regulations. Secondary quantitative data were fully anonymized by HH before being transferred to the researchers for analysis. Data were stored on a secure research environment provided by UCL, accessible only to the research team. The interview recordings and transcripts from the qualitative phase were transferred to a secure research environment, ready for checking and deidentification. HH provided £5 (US $6.75) Amazon vouchers to those who completed an interview.

Participants

Quantitative Phase

Secondary data were obtained from health surveys of app users conducted by HH between March 2023 and March 2024. The surveys were originally collected outside the app but were later integrated within the app platform. Eligibility criteria for this study required participants to (1) report 2 or more chronic conditions during the onboarding health survey, (2) be aged 18 years or more, and (3) complete the baseline and follow-up (≥8 weeks) SWB questionnaires (n=565). Completion of the baseline questionnaire was compulsory; however, demographic information questions (eg, age, gender, and ethnicity) and the follow-up questionnaire were optional. Before applying the eligibility criteria, 882 app users had completed the baseline and follow-up questionnaire. Of this, 565 app users were eligible (64.1% of the available data). The sample size meets the minimum sample size to yield an alpha of .05, a statistical power of 0.8, and a small effect size (d=0.2) [41] in SWB outcomes. The assumption of a small effect size was based on previous literature that investigated the influence of a DHI on the well-being of individuals with a single chronic condition [42]. Demographics such as age, gender, and ethnicity were captured through the app, which were only available for 55.2% (312/565) of participants. Missingness occurred because the SWB survey was not initially integrated into the app; therefore, it cannot be linked to demographic information.

XSL•FO
RenderX

Qualitative Phase

A criterion-based purposive sampling technique was used to identify eligible app users with multimorbidity who had used the app for at least 2 weeks. Recruitment was conducted via an email campaign distributed by the HH team to eligible app users, and participants self-selected to take part by scheduling an interview through a Calendly link. The first email was sent to 2190 app users. A total of 30 app users replied, and interviews were booked; however, 9 did not go ahead due to participants' scheduling conflicts and technical difficulties. The email was then sent to 40 new app users who met the criteria, resulting in 1 more interview and 22 interviews in total. Information power was referenced during the planning and data collection phases to evaluate the sufficiency of the sample size [43]. This process was iterative, with no predefined number of interviews. Given the broad aim of the study and the use of cross-case analysis alongside the application of an established theory and a specific sample, a moderate sample size was deemed appropriate.

Materials and Procedure

Quantitative Phase

The primary outcome was measured by the Office for National Statistics' 4 personal well-being questions (ONS4) [44]. The 4 questions cover life satisfaction (evaluative), worthwhileness (eudemonic), happiness (affective), and anxiety (affective), each measured on an 11-point Likert scale (0 - 10) (see Multimedia Appendix 2 for questions and scoring). The ONS4 data were collected at baseline and the 8-week follow-up with no time limit for completion. The median number of days between baseline and follow-up completion was 84.5 (IQR 59-160) (~12 weeks). Other relevant data included age (18 - 34, 35 - 54, 55 - 64, and 65+ years), gender (male or female), ethnicity (White and other), number of daily habits completed (eg, 10 minutes of mindfulness=1 habit), number of medical conditions, and whether the app user felt they had developed automatic habits without relying on the app (yes or no). Chronic conditions were recorded through self-report using a predefined checklist. If a condition was not represented, app users had the option to select "something else." The checklist was developed based on publicly available NHS data on condition prevalence and the most commonly diagnosed chronic conditions identified among the first 20,000 HH users. They were not able to provide any additional free-text responses.

Qualitative Phase

Potential participants were sent an information sheet and a Calendly link to an interview booking page by HH via email. Participants could book 45-minute time slots for the interview. To book the time slot, participants were required to read the information sheet and check all the consent boxes, providing consent to take part in the study. The interviews were recorded and transcribed via Microsoft Teams. Due to confidentiality, we cannot confirm whether the interview participants were also part of the survey dataset. All interviews were conducted by the lead researcher (IS). The interviews were semistructured, following a topic guide (see Multimedia Appendix 3) discussing participants' multimorbidity and positive and negative experiences of using HH, with the flexibility of follow-up questions and prompts. The topic guide was developed by the lead researcher (IS), with supervision from 2 coauthors (FB and AB). It was informed by the study aims and the conceptual underpinnings of HH. The topic guide was piloted, with one question subsequently divided into two to improve clarity. This interview was included in the analysis. Demographic information, including self-reported chronic conditions, was collected during the interviews. The mean duration of interviews was 24.68 minutes (range 12.31-52.34 minutes). A distress protocol was in place due to potentially sensitive topics being disclosed, informed by the Qualitative Research Distress Protocol tool [45].

A reflective log was used by the lead researcher (IS) to guide self-reflection and transparency during the research process [46]. IS is a White female researcher (MSc) with experience in conducting research with people with long-term conditions and semistructured interviewing. Independence from HH was clarified to participants before the interview to ensure transparency and reduce potential bias. IS's values of respect, transparency, and evidence-based practice fostered rapport and shaped expectations of the app. IS's positionality, shared with many participants who were predominantly White and female, required reflexive attention to how shared identity could influence assumptions. Reflexive journaling and supervision supported critical reflection on how personal characteristics, interviewing style, and emotional responses influenced both data collection and analysis.

Analyses

Quantitative Phase

Descriptive analyses were conducted to summarize characteristics, including frequencies, percentages, means, and SDs. To examine changes in SWB before and after the intervention, quantitative data were analyzed using Bayesian growth curve modeling. It allowed us to examine person-specific and average changes between 2 time points, with random intercept and slope. For the main analysis, we fitted an unconditional linear growth model to the full sample (n=565) separately for each of the ONS4 SWB measures, using noninformative priors, 2000 iterations, 6 chains, a burn-in of 1000, and a thinning of 5. Please see Multimedia Appendix 4 for technical details. A paired samples *t* test was conducted as sensitivity analysis, using both the full sample and a conditional sample restricted to complete demographic data (n=312) to assess the potential influence of missing data. Sensitivity analyses were also conducted to examine if changes in SWB differed by individual characteristics including age, gender, ethnicity, number of medical conditions, automatic habits, and number of habit completions. This was done by fitting separate conditional growth models for each demographic covariate using data from a reduced sample with valid data for that specific covariate (n=292 to 312). The Bayesian growth curve models were fitted in R 4.3.2 (R Foundation for Statistical Computing) and using Markov Chain Monte Carlo algorithms, implemented in JAGS. Model convergence was assessed using the Gelman-Rubin statistic and visual inspection of the posterior distributions. Paired sample *t* tests were conducted in IBM SPSS Statistics 28.0 for Microsoft Windows.

XSL•FO
RenderX

Qualitative Phase

Transcripts were analyzed using NVivo 12 (Lumivero). Reflexive thematic analysis was used, a theoretically flexible methodology used to understand experiences and behaviors [47,48]. The analysis adopted an essentialist, experiential, inductive approach to capture and reflect participants' direct experiences and perspectives [49]. Once the themes were fully developed inductively, they were mapped onto the Multi-Level Leisure Mechanisms Framework [23] top-level categories and were interpreted in the context of participants' reported experiences with HH. This framework was selected as it offers a clear structure for understanding (eg, psychological, social, and behavioral mechanisms of change), which aligns with the study's aim of identifying which app features may support SWB and the factors linked to acceptability. The framework did not shape the coding or theme development process; rather, it was used post hoc to organize the themes within established mechanisms. This supported subsequent interpretation of findings, enabling connections to be drawn between themes and app features.

The lead researcher (IS) coded the transcripts twice, followed by discussion with two coauthors (FB and AB), who each read 2 transcripts. The reflective log was referred to throughout the analysis, with potential preconceptions about the familiarity of using technology considered.

Synthesis of Findings

This study followed an explanatory sequential mixed methods design, in which synthesis of findings occurs during the interpretive stage. In the quantitative phase, the analysis identified both the direction and significance of change in SWB. The qualitative phase was undertaken to provide explanatory depth and contextualization, focusing on potential mechanisms and specific app features that may be associated with the observed findings. The interpretation of the quantitative findings through consideration of the qualitative evidence is presented in the Discussion section.

Results

Quantitative Results

Sample Characteristics

Among app users with valid demographic information (n=312), 35.9% (n=112) were aged 65+ years, forming the largest age category (see Table 1). Over three-quarters (n=245, 78.5%) were female, and 94% (n=280) were White.

Table . Quantitative characteristics.

Variables	Values
Sex (n=312), n (%)	
Female	245 (78.5)
Male	67 (21.5)
Age in years (n=312), n (%)	
18 - 34	14 (4.5)
35 - 54	84 (26.9)
55 - 64	102 (32.7)
65+	112 (35.9)
Ethnicity (n=298), n (%)	
White	280 (94)
Other (Asian, Black, Mixed, or Other)	18 (6)
Automatic habits[a] (n=292), n (%)	
Yes	219 (75)
No	73 (25)
Number of medical conditions (n=312), mean (SD)	4.64 (3.36)
Habit completions[b] (n=312), mean (SD)	161.86 (229.51)

[a]Have you started to do any of your habits automatically? That is, without relying on the app to remind you?

[b]Number of habits recorded as completed.

The mean number of chronic conditions recorded was 4.64 (SD 3.36; Table 1). As shown in Table 2, the most common conditions were anxiety (126/312, 40.4%), hypertension (122/312, 39.1%), depression (108/312, 34.6%), and arthritis (96/312, 30.8%).

XSL•FO
RenderX

Table . Chronic condition frequency (n=312).

Chronic condition	Values, n (%)[a]
Mental health	
Anxiety	126 (40.4)
Depression	108 (34.6)
Cardiometabolic	
Hypertension	122 (39.1)
High cholesterol	85 (27.2)
Type 2 diabetes	43 (13.8)
Prediabetes	38 (12.2)
Heart disease	25 (8.0)
Musculoskeletal or pain	
Arthritis	96 (30.8)
Joint mobility issues	83 (26.6)
Chronic pain	41 (13.1)
Fibromyalgia	16 (7.1)
Osteoporosis	20 (6.4)
Severe mobility impairments	13 (4.2)
Respiratory	
Asthma	58 (18.6)
Chronic obstructive pulmonary disease	18 (5.8)
Gastrointestinal	
Irritable bowel syndrome	28 (17.2)
Irritable bowel syndrome or inflammatory bowel disease	19 (7.2)
Other	
Something else	104 (33.3)
Insomnia	52 (19.6)
Fatty liver	14 (9.2)
Stroke	15 (4.8)
Long COVID	12 (3.8)
Kidney disease	9 (2.9)

[a]Valid percentages calculated based on the number of app users who responded to each condition item. App users could report multiple chronic conditions.

Bayesian Growth Curve Model

The results from the unconditional Bayesian growth curve models are shown in Figure 2. On average, life satisfaction increased by 0.71 (95% highest density interval [HDI] 0.52-0.89) after using the app for 8 or more weeks, worthwhileness increased by 0.62 (95% HDI 0.43-0.81), and happiness increased by 0.74 (95% HDI 0.54-0.92). Alongside this, anxiety decreased by 0.50 (95% HDI −0.74 to −0.25).

XSL•FO
RenderX

Figure 2. Bayesian growth curve model mean trajectories.

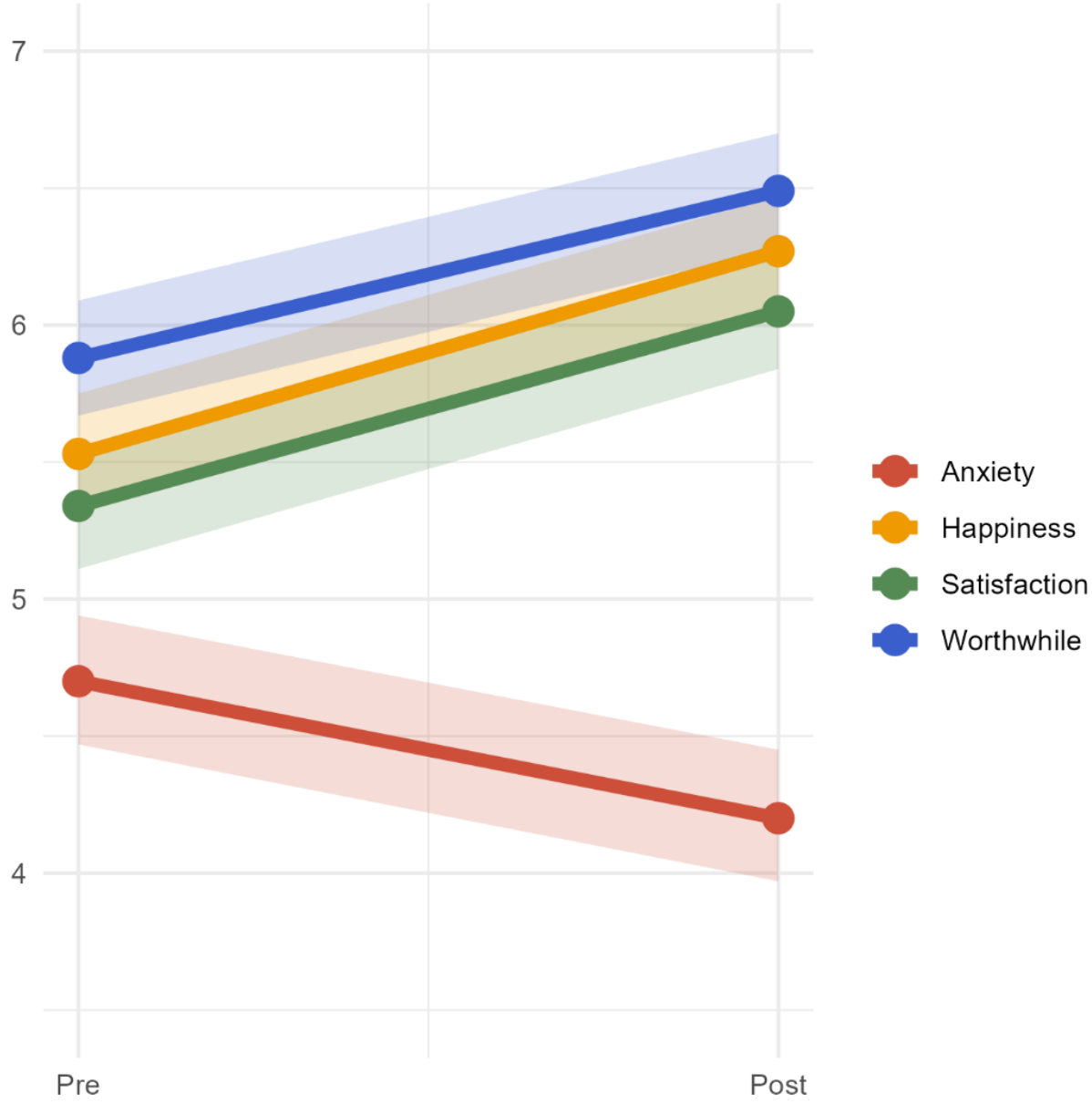

Sensitivity Analyses

As shown in Table 3, paired samples *t* tests revealed a significant difference in life satisfaction (t_{564}=−7.65, *P*<.001), worthwhileness (t_{564}=−6.58, *P*<.001), happiness (t_{564}=−7.46, *P*<.001), and anxiety scores (t_{564}=3.87, *P*<.001). This finding aligns with the Bayesian growth curve model, again suggesting a significant improvement in all ONS4 domains.

XSL•FO
RenderX

Table . Paired samples *t* test results (n=565).

ONS4[a] domains	Mean (SD)	*t* test[b]	*P* value	Cohen *d*[c]
Life satisfaction[d]		–7.65	<.001	–0.32
Prescore	5.34 (2.60)			
Postscore	6.06 (2.42)			
Worthwhileness[d]		–6.58	<.001	–0.28
Prescore	5.88 (2.55)			
Postscore	6.50 (2.46)			
Happiness[d]		–7.46	<.001	–0.31
Prescore	5.54 (2.64)			
Postscore	6.28 (2.53)			
Anxiety[e]		3.87	<.001	0.16
Prescore	4.70 (2.89)			
Postscore	4.20 (2.93)			

[a]ONS4: Office for National Statistics' 4 personal well-being questions.

[b]Two-tailed *t* test (*df*=564).

[c]Cohen *d* (1988) effect sizes: small (*d*=0.2), medium (*d*=0.5), and large (*d*=0.8).

[d]Life satisfaction, worthwhileness, and happiness thresholds: low (0-4), medium (5-6), high (7-8), and very high (9-10).

[e]Anxiety threshold: very low (0-1), low (2-3), medium (4-5), and high (6-10).

In the conditional Bayesian growth curve models (n=292 to 312), group effects were tested (see Multimedia Appendix 5). We found some evidence that age, automatic habits, and number of habit completions were associated with SWB outcomes at baseline. However, there was little evidence that these variables were related to the rate of change for any of the SWB measures.

Qualitative Results

Sample Characteristics

Participants in the qualitative sample (n=22) were aged 29 - 73 years (mean age 55.6 years). Most were female (n=18, 81.8%) and White (n=19, 86.4%). The most common chronic conditions reported were high blood pressure (n=7), anxiety (n=6), and type 2 diabetes (n=6) (see Table 4). Participants described varying levels of engagement with HH, from ongoing active use to reduced or discontinued use.

XSL•FO
RenderX

Table . Qualitative sample characteristics (n=22).

Characteristics	Values
Sex, n (%)	
Female	18 (81.8)
Male	4 (18.2)
Age (years), n (%)	
18 - 34	1 (4.6)
35 - 54	6 (27.3)
55 - 64	10 (45.5)
65+	5 (22.7)
Ethnicity, n (%)	
White	19 (86.4)
Other (Asian, Black, or prefer not to say)	3 (13.6)
Chronic conditions, n	
High blood pressure	7
Anxiety	6
Type 2 diabetes	6
Depression	4
High cholesterol	4
Fibromyalgia	3
Asthma	2
Chronic fatigue syndrome	3
Diverticulitis	2
Irritable bowel syndrome	2
Osteoarthritis	2
Prediabetes	2
Sciatica	2
Stroke	2
Other[a]	39

[a]Atrial fibrillation, attention-deficit/hyperactivity disorder, borderline osteoporosis, bowel cancer, bowel problems, chronic pain, cochlear hydrops, complex posttraumatic stress disorder, congenital heart disease, coronary heart disease, diabetic neuropathy, Graves' disease, gout, hard of hearing, heart disease, heart valve problems, hepatitis, hypermobility spectrum disorder, hypermobility syndrome, hypertension, inflammatory bowel disease, long COVID, low iron, mobility problems, multiple myeloma, musculoskeletal problems, perimenopausal disorder, peripheral neuropathy, rheumatoid arthritis, sleep apnea, tinnitus, thrombocytopenia, and unstable bladder.

Mechanisms of Action

Summary

Participants described 8 app features that activated 5 mechanisms of action leading to perceived improvements in SWB. App features included (1) those promoting action, such as "checking off daily goals," "setting small goals," "reminders and nudges," and "tracking and progress reports"; (2) features fostering reflective practice, including "mindfulness and affirmations" and "positive reinforcement"; and (3) interactive features such as "AI interface" (AI: artificial intelligence) and "constant support availability." Behavioral mechanisms included "maintenance of healthy behaviors" and "increased sense of achievement." The psychological mechanism was "improved self-compassion," and the social mechanisms included "increased feelings of social support" and "reduced feelings of loneliness." The acceptability of the app was impacted by "type of chronic condition," "availability of time," and "the use of other support tools." These relationships are presented in Figure 3. Participant quotes are used to illustrate the relationship between the app features and the activated mechanisms of action.

XSL•FO
RenderX

Figure 3. Thematic associations diagram. This figure illustrates the relationship between app features and the mechanisms of action identified in the study. The figure reflects both overarching categories from the Multi-Level Leisure Mechanisms Framework and participants' reported experiences. Dark pink denotes overarching mechanisms, while light pink indicates specific submechanisms. Similarly, dark blue denotes specific features within these groups. The black dotted arrows indicate the key associations between app features and mechanisms. Factors influencing app acceptability are shown in green, surrounding both features and mechanisms. AI: artificial intelligence.

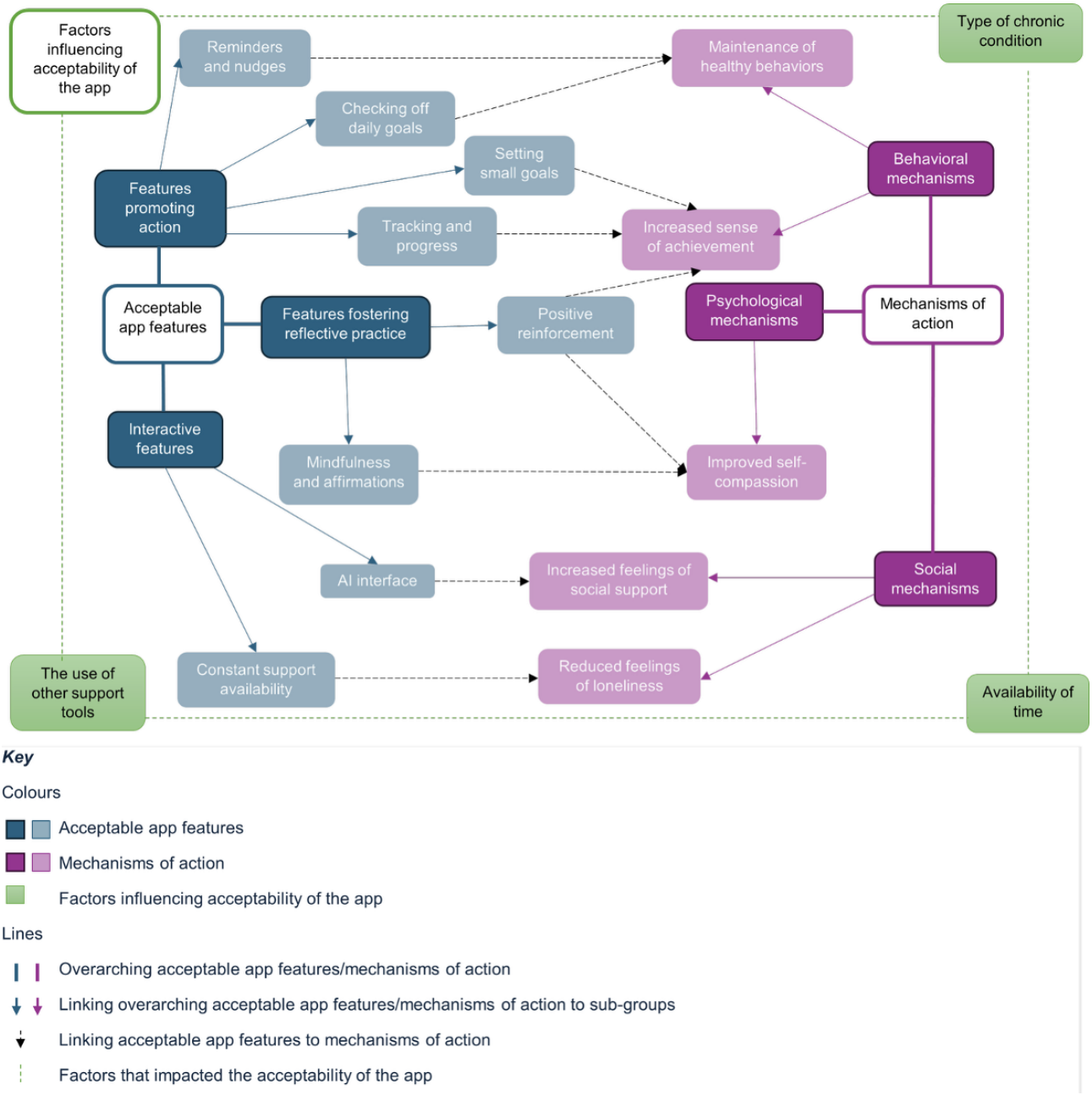

Behavioral Mechanisms

Maintenance of Healthy Behaviors

To maximize the benefits of engaging in healthy habits, including improved SWB, participants needed to reach a point of maintaining them. This process was supported by the app features "reminders and nudges" and "checking off daily goals." Participants felt that they had achieved maintenance when they no longer relied on the app to fulfill their daily goals:

> *The meditation, I just go and do that myself. Now I don't even [use the app] because I'm in the habit of doing it.* [Participant 13]

The app worked in the background to support participants to stay "on track" while maintaining their habits. Some participants reached a point where they felt confident enough to start introducing new habits due to the strong maintenance process:

> *I guess I really should be changing them, the ones that are now coming naturally to me.* [Participant 11]

Participants acknowledged the challenge of maintenance, whether that be due to forgetfulness, needing further nudges, or the nature of the habit itself. The success of the maintenance process varied, with some habits proving more difficult to maintain than others:

XSL•FO
RenderX

> *I try and respond to it daily. I don't always remember, but I try to. I would say that I've been less successful with the exercise things that I've set.* [Participant 11]

Increased Sense of Achievement

Often, participants attributed their progress to an increased sense of achievement, facilitated by the features "positive reinforcement," "tracking and progress reports," and "setting small goals." This "sense of accomplishment" served as a powerful motivator, encouraging them to continue progressing with their goals:

> *I do like the little confetti you get when you've done your habit. You know, I know it's just psychological, but I quite like the confetti. Yeah, there is a certain satisfaction.* [Participant 3]

Moreover, highlighting progress milestones provided further gratification, promoting positive reflection and motivation:

> *It's more encouraging. And I also like when it comes up with a kind of big explosion, that's 100 hours now that you've done in the brain studio.* [Participant 12]

Psychological Mechanisms: Increased Self-Compassion

Participants emphasized the importance of self-compassion in fostering a more benevolent outlook. This mechanism was linked to the features of "positive reinforcement" and "mindfulness and affirmations." Many participants described feeling discouraged about incomplete goals and restrictions on what they could achieve due to multimorbidity. This led some participants to "talk poorly" to themselves, using HH to help challenge these negative beliefs and "rationalize." This was especially effective in prompting participants to acknowledge their progress and establish realistic expectations:

> *Since I've had Holly Health, I've been able to put a lot of things into perspective. Because it's saying it's okay to have down days, it's okay not to feel your best. But it's also inspiring because it gives you the thought that you can do it another day when you feel better.* [Participant 6]

By adopting this mindset, HH helped participants cultivate greater self-compassion by not framing experience in absolute terms:

> *I think the problem is with my tick charts or where I've got a notebook and I'm checking a box or something, I can get down on myself if I've bitten off more than I can chew. I'm a bit of a perfectionist I think - avoiding that black and white, you either did it or you didn't, you were great, or you were rubbish kind of thing.* [Participant 4]

Social Mechanisms

Increased Feelings of Social Support

Participants described the human-like qualities of HH, which contributed to improving their SWB. This was directly attributed to the "AI interface" feature. Such support was particularly valuable when considering multimorbidity, as interactions with health care professionals are often related to a deterioration in health. The chatbot enabled participants to seek out and experience more preventative and proactive holistic support:

> *It almost feels like somebody's looking out for you, but not in a purely medical context. Because you end most of what you do with the healthcare professional is usually when something goes wrong. So, it's not when something's gone wrong, it's actually slightly helping things not to go wrong.* [Participant 3]

The impact of the support was especially strong when it was tailored; the "more holistic touch" of HH made the support feel "personal." This introduced feelings of having friend-like, reliable support:

> *It helps with depression because it can talk to you through times of no stimulation and depressive thoughts. It's like a friend.* [Participant 6]

Reduced Feelings of Loneliness

Living with multimorbidity often led participants to experience feelings of loneliness, with peers sometimes struggling to understand the associated challenges. Feeling listened to and acknowledged was crucial to participants, with some encountering dismissive interactions elsewhere. The "constant support availability" and "AI interface" of HH helped participants to feel less lonely:

> *I have chats with them maybe once a week. It does help knowing that somebody is at the other side, even though it's just a computer, and that somebody's listening.* [Participant 16]

Participants expressed that enduring support options were crucial for their well-being, particularly when they found it challenging to actively seek support:

> *I do feel Holly (Health) is what's carried me through that really tough spell. Where I genuinely didn't really have any support from anyone, and it is much because I couldn't ask for it.* [Participant 10]

Factors Influencing Acceptability of the App

Three factors that affected the acceptability of the app were identified: "type of chronic condition," "availability of time," and "using other support tools alongside HH."

Type of Chronic Condition

Many participants viewed HH as a "holistic" support option, capable of addressing a wide range of chronic conditions simultaneously:

> *It's quite universal though, I think there's so much in there that there'll be something that would work for anybody's condition.* [Participant 11]

However, others noted HH seemed more beneficial for certain conditions, with more advice and support available for general chronic conditions:

> *It seems to be designed for quite common conditions, anxiety, depression. I think there are more specific problems I had around rumination […] I don't think it was specifically good for that, I think it's good for general wellbeing.* [Participant 5]

XSL•FO
RenderX

Furthermore, participants with chronic fatigue syndrome (CFS) noted that the app provided limited information and advice on managing their condition, feeling it "needs something extra." Some felt their conditions were too complex or unique to be fully addressed by HH:

> *I don't really feel that I fit into an easy category. Particularly people with fibromyalgia. Your experience is very individual.* [Participant 21]

Availability of Time

Having time to fully engage in the app and complete the goals was a key factor in how suitable the app was for improving SWB. Many participants valued having the designated time to focus on their well-being goals daily:

> *There's notifications to give you that reminder, because let's be honest, well life gets in the way of everything, but sometimes you have to take that five-minutes for yourself.* [Participant 17]

Another participant described that "it's time for me," with the app allowing participants to adapt goals to fit within schedules. This was especially apparent in participants who were retired or semiretired, having more time to engage with the app and explore the activities:

> *But that's one of the things about being semiretired and having more time is trying these different things.* [Participant 9]

The Use of Other Support Tools

Many participants used a combination of support to ensure comprehensive care for their mental and physical health, with HH being one component of a broader support regimen:

> *Another tool in your general armoury of things. You're trying to look after yourself as best you can.* [Participant 15]

Some participants mentioned using multiple apps together, selecting the most suitable based on their current needs:

> *I have quite a few different apps on my phone. I suppose I'm always seeking which one feels right in the moment but also will give me longer-term support.* [Participant 8]

Others combined HH with more intensive forms of support, such as health coaching, adapting techniques from both face-to-face and digital formats to enhance their support plan and maximize the benefits:

> *I'm having counselling and stuff anyway, so obviously some techniques from there which I already knew anyway, but it opened my eyes to wider things.* [Participant 20]

HH was especially valuable when other support tools were not readily available, serving as an interim solution, particularly when waiting for support:

> *I was on the wait list again for CBT, so I thought some of the same techniques might be accessible in the app while I'm waiting.* [Participant 5]

However, some participants found using multiple support tools simultaneously challenging. This limited the extent of the progress made, with some participants perceiving the required commitment as excessive, especially when using multiple health apps together:

> *It just felt (it was) difficult to keep doing it because it's a big commitment to do any one of those apps for a long period.* [Participant 5]

Discussion

Principal Results

This study aimed to examine the impact of a health coaching app on the SWB of individuals with multimorbidity and understand how and why the app supported SWB. Our quantitative analyses revealed that using the app for 8 weeks or longer was associated with increased SWB, including life satisfaction, worthwhileness, happiness, and a reduction in anxiety. There was little evidence that these changes differed by individual characteristics.

Through qualitative analyses, 8 app features were identified as acceptable, categorized by features promoting action, reflective practice, and interaction. These features facilitated a positive impact on SWB through activating behavioral, psychological, and social mechanisms of action. Wider contexts that impact the app acceptability were also identified, including chronic condition type, having time to engage, and the availability of other support tools alongside HH. These findings suggest that a health coaching app can serve as an acceptable support tool within these contexts.

Comparison With Prior Work

These findings provide evidence for the SWB benefits of a health coaching app for individuals with multimorbidity after 8 weeks or longer. This aligns with previous research demonstrating the effectiveness of digital solutions in enhancing well-being and self-management of single chronic conditions [42,50] alongside previous internal reports from HH, indicating that app usage boosts confidence in managing multimorbidity [51]. Additionally, this is consistent with prior evidence indicating that behavioral interventions can produce measurable well-being improvements within as little as 4 weeks in a range of populations, including people with chronic conditions [52]. Although the estimated changes in SWB scores were modest (<1 point on the Likert scales), such differences are often regarded as meaningful at the population level. For example, the UK HM Treasury Green Book Wellbeing Guidance [53] equates a 1-point increase in life satisfaction with a monetary value of approximately £13,000 (US $17,556.65) per person per year. Furthermore, the observed effect sizes in this study align with those reported in comparable digital behavior change and well-being interventions [54]. While research on digital support tools for multimorbidity is sparse, these findings extend insights from single-condition literature [55], suggesting that this health coaching app may similarly promote self-management and SWB among people with multimorbidity.

The mapping of perceived mechanisms of action to the Multi-Level Leisure Mechanisms Framework [23] enabled the

XSL•FO
RenderX

connection between app use and improved SWB to be explored in-depth. These mechanisms often intersect, with behavioral, psychological, and social mechanisms reinforcing one another to support SWB outcomes. Identified behavioral mechanisms of action included maintenance of healthy behaviors and increased sense of achievement. These mechanisms were observed to be supported by reminders and nudges, checking off daily goals, setting small goals, and tracking progress reports. The process involved in checking off daily goals was highlighted as particularly impactful in promoting action, reflected in app users reaching, maintaining, and setting new goals. Positive reinforcement also played a key role in the sense of achievement, encouraging app users to accomplish new goals. This finding is consistent with previous research that suggests goal setting, action planning, and reinforcement are crucial for sustained user engagement and promoting behavior change [56]. Although evidence of the effectiveness of reminders and nudges has previously been inconsistent, the context-specific design of reminders (eg, those tailored to specific habits) was found to be effective in this study. This finding aligns with previous research showing that context-specific reminders are crucial in enabling habit formation [57,58].

Positive reinforcement was also reported to be linked to the psychological mechanism of improved self-compassion. Goal setting similarly facilitated improved self-compassion; app users were able to establish realistic expectations and recognize their progress. This mechanism of action was additionally facilitated by the mindfulness and affirmations content within the DHI, fostering reflective practice. Social mechanisms of action were identified as increased feelings of social support, observed to be activated by the AI interface feature, and reduced feelings of loneliness, reported to be activated by the constant support availability within the DHI. The AI interface was highly valued by app users, and the live chat supported app users through a dependable, personalized approach. These findings are consistent with previous research, with several features identified as valuable in DHIs, including relaxation, personalization, and live support [59]. Distinctly, previous research has recognized increased social support as a key mechanism of improved self-management in multimorbidity interventions [60]. This emphasizes the value of providing features that facilitate this mechanism of action, especially in multimorbidity interventions.

The effect of the DHI can also be strengthened by using evidence-based approaches, such as CBT [35] and the COM-B model [37]. Incorporating these approaches has been found to improve the effectiveness, engagement, perceived quality, and credibility of apps [61-63]. Findings in this study suggest that using app-based support alongside other support tools (eg, health coaching) enhances effectiveness and acceptability. This aligns with previous literature [64], emphasizing the importance of complementary principles to ensure consistency of support.

Novel findings from the qualitative interviews suggest that the complexity of multimorbidity plays a crucial role in the impact and acceptability of the health coaching app as a support tool. While the health coaching app demonstrated a positive impact on SWB, indicating its potential to enhance SWB in people with multimorbidity to some extent, qualitative findings identified that chronic condition type influenced the acceptability of the health coaching app, with lack of information and education on specific conditions consistently a key barrier to support [65]. CFS and fatigue-based symptoms were particularly less accommodated by the app content, which primarily targeted more general conditions such as anxiety. Previous research has emphasized the significance of fear-avoidance beliefs and determining personal thresholds in the efficacy of supporting CFS [66,67]. Therefore, acknowledging these factors is crucial to the acceptability of app-based support for individuals with a CFS comorbidity. Conversely, conditions such as depression and anxiety, which have the highest health care service use in individuals with multimorbidity [24], appear to benefit from the health coaching app as an effective support tool for improving SWB.

The value of DHIs in older adults has been recognized in previous literature, particularly given the high prevalence of long-term conditions in this demographic group [68]. However, there have been concerns that the digital exclusivity of these interventions may lead to higher attrition rates, along with decreased engagement and effectiveness [68]. Notably, our quantitative findings did not identify age as a barrier to SWB improvement, indicating the potential effectiveness of the health coaching app across age groups, including older adults. In fact, our qualitative findings identified that retirement appeared to facilitate greater app engagement, providing individuals with more time to achieve daily goals. This observation aligns with previous research that associates retirement with increased leisure activities, including personal growth and physical exercise [69].

Strengths and Limitations

The main strength of this study lies in its explanatory sequential mixed methods design, which allowed for a comprehensive understanding of the insights gained from the quantitative phase through qualitative work. This allowed an in-depth exploration of a complex, under-researched area. However, several limitations must be acknowledged. First, the sample in both phases predominantly consisted of White females aged 55 years and older. While this partly reflects the prevalence of multimorbidity in older adults and females [4,5], the lack of diversity in the sample raises concerns about the generalizability of the results to the wider population [70]. Second, confidentiality measures prevented verifying if the same participants were involved in both data collection phases, potentially introducing participant variability and affecting the consistency of results. Third, as participation in the qualitative phase was based on self-selection, the sample may have predominantly consisted of motivated or engaged app users. This self-selection could introduce volunteer bias and limit the generalizability of the findings. However, participants did report varying levels of engagement with the app and spoke about various challenges of use. Similarly, in the quantitative phase, completion of the follow-up questionnaire was optional, which may have resulted in overrepresentation of engaged app users. Additionally, the absence of a control group makes it challenging to establish a causal effect as there is no comparison group against the observed effects, ultimately reducing internal

XSL•FO
RenderX

validity. Due to these limitations, caution is advised when interpreting the study findings.

Conclusions

Our study provides empirical evidence that a health coaching app can be an effective and acceptable support tool to improve the SWB of individuals with multimorbidity. These effects were driven by specific app features promoting action, reflective practice, and interaction. These features led to improved SWB through the activation of reported behavioral, psychological, and social mechanisms. However, the magnitude of these effects could be affected by contextual factors, including users' time availability for engagement, specific chronic condition profiles, and concurrent use of other support tools. By elucidating the mechanisms and contextual nuances underlying app efficacy, this study provides critical insights to inform the refinement of existing interventions and the design of future DHIs tailored to the complex needs of individuals with multimorbidity.

Acknowledgments

The authors would like to thank Holly Health, particularly Liliana Chow, for their involvement and work toward the partnered research project, and for providing access to their data. The authors also thank the participants for their time and contribution to the study.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

Data will be shared upon reasonable request from the corresponding author.

Authors' Contributions

Conceptualization: FB, AB, IS
Data curation: IS, FB
Formal analysis: IS, FB
Methodology: IS, FB, AB
Supervision: FB, AB
Writing – original draft: IS, FB
Writing – review and editing: IS, FB, AB, DM

Conflicts of Interest

DM is employed by Holly Health. She was not involved in conducting the interviews or the data analysis. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Multimedia Appendix 1
Images of the Holly Health app.
[DOCX File, 941 KB - jmir_v28i1e78738_app1.docx]

Multimedia Appendix 2
ONS4 questions and thresholds. ONS4: Office for National Statistics' 4 personal well-being questions.
[DOCX File, 19 KB - jmir_v28i1e78738_app2.docx]

Multimedia Appendix 3
Interview topic guide.
[DOCX File, 17 KB - jmir_v28i1e78738_app3.docx]

Multimedia Appendix 4
Bayesian unconditional growth curve model technical specifications.
[DOCX File, 15 KB - jmir_v28i1e78738_app4.docx]

Multimedia Appendix 5
Bayesian growth curve model: group effects results.
[DOCX File, 22 KB - jmir_v28i1e78738_app5.docx]

XSL•FO
RenderX

Checklist 1
STROSA and COREQ checklists.
[DOCX File, 31 KB - jmir_v28i1e78738_app6.docx]

References

1. Kingston A, Robinson L, Booth H, Knapp M, Jagger C, MODEM project. Projections of multi-morbidity in the older population in England to 2035: estimates from the Population Ageing and Care Simulation (PACSim) model. Age Ageing 2018 May 1;47(3):374-380. [doi: 10.1093/ageing/afx201] [Medline: 29370339]
2. Johnston MC, Crilly M, Black C, Prescott GJ, Mercer SW. Defining and measuring multimorbidity: a systematic review of systematic reviews. Eur J Public Health 2019 Feb 1;29(1):182-189. [doi: 10.1093/eurpub/cky098] [Medline: 29878097]
3. Zemedikun DT, Gray LJ, Khunti K, Davies MJ, Dhalwani NN. Patterns of multimorbidity in middle-aged and older adults: an analysis of the UK Biobank data. Mayo Clin Proc 2018 Jul;93(7):857-866. [doi: 10.1016/j.mayocp.2018.02.012]
4. Agborsangaya CB, Lau D, Lahtinen M, Cooke T, Johnson JA. Multimorbidity prevalence and patterns across socioeconomic determinants: a cross-sectional survey. BMC Public Health 2012 Mar 19;12(1):201. [doi: 10.1186/1471-2458-12-201] [Medline: 22429338]
5. Kone AP, Mondor L, Maxwell C, Kabir US, Rosella LC, Wodchis WP. Rising burden of multimorbidity and related socio-demographic factors: a repeated cross-sectional study of Ontarians. Can J Public Health 2021 Aug;112(4):737-747. [doi: 10.17269/s41997-021-00474-y] [Medline: 33847995]
6. Skou ST, Mair FS, Fortin M, et al. Multimorbidity. Nat Rev Dis Primers 2022 Jul 14;8(1):1-22. [doi: 10.1038/s41572-022-00376-4]
7. Johnston MC, Black C, Mercer SW, Prescott GJ, Crilly MA. Prevalence of secondary care multimorbidity in mid-life and its association with premature mortality in a large longitudinal cohort study. BMJ Open 2020 May;10(5):e033622. [doi: 10.1136/bmjopen-2019-033622]
8. Rizzuto D, Melis RJF, Angleman S, Qiu C, Marengoni A. Effect of chronic diseases and multimorbidity on survival and functioning in elderly adults. J Am Geriatr Soc 2017 May;65(5):1056-1060. [doi: 10.1111/jgs.14868] [Medline: 28306158]
9. Lujic S, Randall DA, Simpson JM, Falster MO, Jorm LR. Interaction effects of multimorbidity and frailty on adverse health outcomes in elderly hospitalised patients. Sci Rep 2022 Aug 19;12(1):14139. [doi: 10.1038/s41598-022-18346-x]
10. Buja A, Rivera M, De Battisti E, et al. Multimorbidity and hospital admissions in high-need, high-cost elderly patients. J Aging Health 2020 Jun;32(5-6):259-268. [doi: 10.1177/0898264318817091]
11. Soley-Bori M, Ashworth M, Bisquera A, et al. Impact of multimorbidity on healthcare costs and utilisation: a systematic review of the UK literature. Br J Gen Pract 2021 Jan;71(702):e39-e46. [doi: 10.3399/bjgp20X713897]
12. Moffat K, Mercer SW. Challenges of managing people with multimorbidity in today's healthcare systems. BMC Fam Pract 2015 Oct 14;16(1):129. [doi: 10.1186/s12875-015-0344-4] [Medline: 26462820]
13. Rimmelzwaan LM, Bogerd MJL, Schumacher BMA, Slottje P, Van Hout HPJ, Reinders ME. Multimorbidity in general practice: unmet care needs from a patient perspective. Front Med 2020 Dec 22;7. [doi: 10.3389/fmed.2020.530085]
14. Wilkinson I, Preston J. Managing patients with multimorbidity. Clinics in Integrated Care 2021 Apr;5:100045. [doi: 10.1016/j.intcar.2021.100045]
15. Multimorbidity: clinical assessment and management. : NICE; 2016 Sep. URL: https://www.nice.org.uk/guidance/ng56/chapter/Recommendations#general-principles [accessed 2025-12-23]
16. Diener E, editor. Assessing Well-Being: Springer Netherlands; 2009. [doi: 10.1007/978-90-481-2354-4]
17. Hicks S, Tinkler L, Allin P. Measuring subjective well-being and its potential role in policy: perspectives from the UK Office for National Statistics. Soc Indic Res 2013 Oct;114(1):73-86. [doi: 10.1007/s11205-013-0384-x]
18. Personal well-being user guidance. Office for National Statistics. 2025 Jan. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/methodologies/personalwellbeingsurveyuserguide [accessed 2025-12-23]
19. Steptoe A, Deaton A, Stone AA. Subjective wellbeing, health, and ageing. Lancet 2015 Feb 14;385(9968):640-648. [doi: 10.1016/S0140-6736(13)61489-0] [Medline: 25468152]
20. Kushlev K, Drummond DM, Diener E. Subjective well-being and health behaviors in 2.5 million Americans. Appl Psychol Health Well Being 2020 Mar;12(1):166-187. [doi: 10.1111/aphw.12178] [Medline: 31482675]
21. Multimorbidity | Health topics A to Z. CKS | NICE. URL: https://cks.nice.org.uk/topics/multimorbidity/ [accessed 2025-12-23]
22. Feter N, Leite JS, Umpierre D, Caputo EL, Rombaldi AJ. Multimorbidity and leisure-time physical activity over the life course: a population-based birth cohort study. BMC Public Health 2021 Apr 9;21(1):700. [doi: 10.1186/s12889-021-10719-7] [Medline: 33836716]
23. Fancourt D, Aughterson H, Finn S, Walker E, Steptoe A. How leisure activities affect health: a narrative review and multi-level theoretical framework of mechanisms of action. Lancet Psychiatry 2021 Apr;8(4):329-339. [doi: 10.1016/S2215-0366(20)30384-9]
24. Multiple long-term conditions (multimorbidity): making sense of the evidence. : NIHR Evidence; 2021 Mar. [doi: 10.3310/collection_45881]

XSL•FO
RenderX

25. Schäfer I, Hansen H, Kaduszkiewicz H, et al. Health behaviour, social support, socio-economic status and the 5-year progression of multimorbidity: results from the MultiCare Cohort Study. J Comorb 2019;9:2235042X19883560. [doi: 10.1177/2235042X19883560] [Medline: 35174099]
26. Chudasama YV, Khunti K, Gillies CL, et al. Healthy lifestyle and life expectancy in people with multimorbidity in the UK Biobank: a longitudinal cohort study. PLOS Med 2020;17(9):e1003332. [doi: 10.1371/journal.pmed.1003332]
27. Eisenstadt M, Liverpool S, Infanti E, Ciuvat RM, Carlsson C. Mobile apps that promote emotion regulation, positive mental health, and well-being in the general population: systematic review and meta-analysis. JMIR Ment Health 2021 Nov 8;8(11):e31170. [doi: 10.2196/31170] [Medline: 34747713]
28. Digital health and care. European Commission. 2025. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/digital-health-and-care_en [accessed 2025-12-23]
29. Deady M, Glozier N, Calvo R, et al. Preventing depression using a smartphone app: a randomized controlled trial. Psychol Med 2022 Feb;52(3):457-466. [doi: 10.1017/S0033291720002081] [Medline: 32624013]
30. Bailey DP, Mugridge LH, Dong F, Zhang X, Chater AM. Randomised controlled feasibility study of the MyHealthAvatar-Diabetes smartphone app for reducing prolonged sitting time in type 2 diabetes mellitus.. Int J Environ Res Public Health 2020 Jun 19;17(12):4414. [doi: 10.3390/ijerph17124414] [Medline: 32575482]
31. Alessa T, Abdi S, Hawley MS, de Witte L. Mobile apps to support the self-management of hypertension: systematic review of effectiveness, usability, and user satisfaction. JMIR Mhealth Uhealth 2018 Jul 23;6(7):e10723. [doi: 10.2196/10723] [Medline: 30037787]
32. Irfan Khan A, Gill A, Cott C, Hans PK, Steele Gray C. mHealth tools for the self-management of patients with multimorbidity in primary care settings: pilot study to explore user experience. JMIR Mhealth Uhealth 2018 Aug 28;6(8):e171. [doi: 10.2196/mhealth.8593] [Medline: 30021707]
33. Holly Health | Home. URL: https://www.hollyhealth.io/ [accessed 2025-12-23]
34. Holly Health – digital support for women's health and wellbeing. NHS Innovation Accelerator. URL: https://nhsaccelerator.com/innovations/holly-health/ [accessed 2025-12-23]
35. Beck AT. Cognitive Therapy and the Emotional Disorders: Penguin; 1979.
36. Hayes SC, Pierson H. Acceptance and commitment therapy. In: Freeman A, Felgoise SH, Nezu CM, Nezu AM, Reinecke MA, editors. Encyclopedia of Cognitive Behavior Therapy: Springer US; 2005:1-4. [doi: 10.1007/0-306-48581-8_1]
37. Michie S, Atkins L, West R. The Behaviour Change Wheel: A Guide to Designing Interventions: Silverback Publishing; 2024. URL: https://www.behaviourchangewheel.com/ [accessed 2025-12-23]
38. Ivankova NV, Creswell JW, Stick SL. Using mixed-methods sequential explanatory design: from theory to practice. Field methods 2006 Feb;18(1):3-20. [doi: 10.1177/1525822X05282260]
39. Swart E, Schmitt J. Standardized Reporting of Secondary Data Analyses (STROSA)—a recommendation. Z Evid Fortbild Qual Gesundhwes 2014;108(8-9):511-516. [doi: 10.1016/j.zefq.2014.08.022] [Medline: 25523850]
40. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
41. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
42. Parks AC, Williams AL, Kackloudis GM, Stafford JL, Boucher EM, Honomichl RD. The effects of a digital well-being intervention on patients with chronic conditions: observational study. J Med Internet Res 2020 Jan 10;22(1):e16211. [doi: 10.2196/16211] [Medline: 31922491]
43. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
44. Surveys using our four personal well-being questions. Office for National Statistics. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/methodologies/surveysusingthe4officefornationalstatisticspersonalwellbeingquestions [accessed 2025-12-23]
45. Whitney C, Evered JA. The qualitative research distress protocol: a participant-centered tool for navigating distress during data collection. Int J Qual Methods 2022 Apr;21. [doi: 10.1177/16094069221110317]
46. Darawsheh W. Reflexivity in research: promoting rigour, reliability and validity in qualitative research. Int J Ther Rehabil 2014 Dec 2;21(12):560-568. [doi: 10.12968/ijtr.2014.21.12.560]
47. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
48. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qualitative Research in Sport, Exercise and Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
49. Byrne D. A worked example of Braun and Clarke's approach to reflexive thematic analysis. Qual Quant 2022 Jun;56(3):1391-1412. [doi: 10.1007/s11135-021-01182-y]
50. Bricca A, Pellegrini A, Zangger G, Ahler J, Jäger M, Skou ST. The quality of health apps and their potential to promote behavior change in patients with a chronic condition or multimorbidity: systematic search in App Store and Google Play. JMIR Mhealth Uhealth 2022 Feb 4;10(2):e33168. [doi: 10.2196/33168] [Medline: 35119367]

XSL•FO
RenderX

51. Long-term condition management across Holly Health users: a retrospective survey. Holly Health. 2023 Nov 30. URL: https://www.hollyhealth.io/articles/long-term-condition-management [accessed 2025-12-23]
52. Weiss LA, Westerhof GJ, Bohlmeijer ET. Can we increase psychological well-being? The effects of interventions on psychological well-being: a meta-analysis of randomized controlled trials. PLoS ONE 2016;11(6):e0158092. [doi: 10.1371/journal.pone.0158092] [Medline: 27328124]
53. Wellbeing guidance for appraisal: supplementary green book guidance. : HM Treasury; 2021 Jul URL: https://assets.publishing.service.gov.uk/media/60fa9169d3bf7f0448719daf/Wellbeing_guidance_for_appraisal_-_supplementary_Green_Book_guidance.pdf [accessed 2025-12-23]
54. Tiley K, Crellin R, Domun T, Harkness F, Blodgett JM. Effectiveness of 234 interventions to improve life satisfaction: a rapid systematic review. Soc Sci Med 2025;366:117662. [doi: 10.1016/j.socscimed.2024.117662]
55. Whitehead L, Seaton P. The effectiveness of self-management mobile phone and tablet apps in long-term condition management: a systematic review. J Med Internet Res 2016 May 16;18(5):e97. [doi: 10.2196/jmir.4883] [Medline: 27185295]
56. Szinay D, Perski O, Jones A, Chadborn T, Brown J, Naughton F. Perceptions of factors influencing engagement with health and well-being apps in the United Kingdom: qualitative interview study. JMIR Mhealth Uhealth 2021 Dec 16;9(12):e29098. [doi: 10.2196/29098] [Medline: 34927597]
57. Auf H, Dagman J, Renström S, Chaplin J. Gamification and nudging techniques for improving user engagement in mental health and well-being apps. Proc Des Soc 2021 Aug;1:1647-1656. [doi: 10.1017/pds.2021.426]
58. Pinder C, Vermeulen J, Cowan BR, Beale R. Digital behaviour change interventions to break and form habits. ACM Trans Comput-Hum Interact 2018 Jun 30;25(3):1-66. [doi: 10.1145/3196830]
59. Alqahtani F, Winn A, Orji R. Co-designing a mobile app to improve mental health and well-being: focus group study. JMIR Form Res 2021 Feb 26;5(2):e18172. [doi: 10.2196/18172] [Medline: 33635281]
60. Miller JJ, Pozehl BJ, Alonso W, Schmaderer M, Eisenhauer C. Intervention components targeting self-management in individuals with multiple chronic conditions: an integrative review. West J Nurs Res 2020 Nov;42(11):948-962. [doi: 10.1177/0193945920902146] [Medline: 32075541]
61. Rathbone AL, Clarry L, Prescott J. Assessing the efficacy of mobile health apps using the basic principles of cognitive behavioral therapy: systematic review. J Med Internet Res 2017 Nov 28;19(11):e399. [doi: 10.2196/jmir.8598] [Medline: 29187342]
62. Fitzgerald M, McClelland T. What makes a mobile app successful in supporting health behaviour change? Health Educ J 2017 Apr;76(3):373-381. [doi: 10.1177/0017896916681179]
63. Roberts AL, Potts HW, Koutoukidis DA, Smith L, Fisher A. Breast, prostate, and colorectal cancer survivors' experiences of using publicly available physical activity mobile apps: qualitative study. JMIR Mhealth Uhealth 2019 Jan 4;7(1):e10918. [doi: 10.2196/10918] [Medline: 30609982]
64. Obro LF, Heiselberg K, Krogh PG, et al. Combining mHealth and health-coaching for improving self-management in chronic care: a scoping review. Patient Educ Couns 2021 Apr;104(4):680-688. [doi: 10.1016/j.pec.2020.10.026] [Medline: 33143907]
65. Doyle J, Murphy E, Kuiper J, et al. Managing multimorbidity: identifying design requirements for a digital self-management tool to support older adults with multiple chronic conditions. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems: Association for Computing Machinery; 2019:1-14. [doi: 10.1145/3290605.3300629]
66. Stahl D, Rimes KA, Chalder T. Mechanisms of change underlying the efficacy of cognitive behaviour therapy for chronic fatigue syndrome in a specialist clinic: a mediation analysis. Psychol Med 2014 Apr;44(6):1331-1344. [doi: 10.1017/S0033291713002006]
67. Davies T, Jones SL, Kelly RM. Patient perspectives on self-management technologies for chronic fatigue syndrome. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems: Association for Computing Machinery; 2019:1-13. [doi: 10.1145/3290605.3300452]
68. Gulliford M, Alageel S. Digital health intervention at older ages. Lancet Digit Health 2019 Dec;1(8):e382-e383. [doi: 10.1016/S2589-7500(19)30194-3] [Medline: 33323216]
69. Tunney O, Henkens K, van Solinge H. A life of leisure? Investigating the differential impact of retirement on leisure activity.. J Gerontol 2023 Oct 9;78(10):1775-1784. [doi: 10.1093/geronb/gbad097]
70. Tindle R. Improving the global reach of psychological research. Discov Psychol 2021 Dec;1(1):5. [doi: 10.1007/s44202-021-00004-4]

Abbreviations

AI: artificial intelligence
CBT: cognitive behavioral therapy
CFS: chronic fatigue syndrome
COM-B: Capability-Opportunity-Motivation Behavior Model
COREQ: Consolidated Criteria for Reporting Qualitative Research
DHI: digital health intervention

XSL•FO
RenderX

GP: general practitioner
HDI: highest density interval
HH: Holly Health
NHS: UK National Health Service
NICE: National Institute for Health and Care Excellence
ONS4: Office for National Statistics' 4 personal well-being questions
STROSA: Standardized Reporting of Secondary Data Analyses
SWB: subjective well-being
UCL: University College London

Edited by N Cahill, TDA Cardoso; submitted 10.Jun.2025; peer-reviewed by C Yang, N O'Brien; revised version received 08.Dec.2025; accepted 10.Dec.2025; published 04.Feb.2026.

Please cite as:
Symes I, Burton A, Mercado D, Bu F
The Impact of a Health Coaching App on the Subjective Well-Being of Individuals With Multimorbidity: Mixed Methods Study
J Med Internet Res 2026;28:e78738
URL: https://www.jmir.org/2026/1/e78738
doi:10.2196/78738

© Isabelle Symes, Alexandra Burton, Daniela Mercado, Feifei Bu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Problems and Barriers Regarding the Admission, Financing, and Service Provision of Digital Health Apps: Qualitative Stakeholder Survey

Felix Plescher[1]; Godwin Denk Giebel[1], Dr rer medic; Carina Abels[1], Dr rer medic; Silke Neusser[1], DrPH; Bonnie Kampka[2]; Kirstin Börchers[2], Prof Dr Med; Jürgen Wasem[1], Prof Dr; Nikola Blase[1], Dr med

[1]Faculty of Economics, Institute for Healthcare Management and Research, University Duisburg-Essen, Thea-Leymann-Street 9, Essen, Germany
[2]QM Börchers Consulting +, Herne, Germany

Corresponding Author:
Felix Plescher
Faculty of Economics, Institute for Healthcare Management and Research, University Duisburg-Essen, Thea-Leymann-Street 9, Essen, Germany

Abstract

Background: Since their introduction with the Digital Care Act in 2019, selected digital health apps (DiGA) have been a part of the German statutory health care system. In order to become a DiGA, digital health apps have to complete a certification process covering both technical and evidence-related aspects. After completion, DiGA are added to the DiGA directory, containing a list of all reimbursable DiGA within German statutory health insurance. The first apps were added at the end of 2020, with the number steadily increasing. The novelty of this digital health care service and the fast implementation led to problems and barriers to optimal use along the way, which are studied from different stakeholder perspectives in this paper.

Objective: The aim of this survey was to identify problems and barriers in the context of the admission, financing, and service provision of DiGA in the statutory health care system in Germany.

Methods: We used semistructured expert interviews to evaluate the perspective of stakeholders of the German health care system on DiGA. The interview guide was developed according to Helfferich. The interviews were transcribed and analyzed using the qualitative content approach by Mayring, with the adjustments by Kuckartz. We conducted web-based interviews with stakeholders between February and June 2022. The stakeholder collective consisted of DiGA, statutory health insurance, physician, patient, technological, and quality assurance representatives.

Results: To identify problems from stakeholder perspectives regarding the admission, financing, and service distribution of DiGA, 21 interviews were conducted. The interviewed stakeholders reported problems with the authorization of DiGA and the corresponding process, for example, the duration of the DiGA Fast Track process. DiGA prices and the different negotiation positions were criticized, and financial challenges for smaller DiGA manufacturers were noted. Another problem was seen in the reimbursement of DiGA, independent of actual use by the patients. Within service provision, the participants reported superordinate aspects, for example, the negative public perception of DiGA and negative statements from stakeholders. In relation to the direct care process, technical problems (eg, with activation codes or software surrounding DiGA prescription) and insufficient knowledge and skills on the side of the patients, as well as the medical providers, were mentioned.

Conclusions: Digital health apps have the potential to improve health care by addressing health problems in new, innovative ways. Since the evidence-based and regulated use of this technology is relatively new, problems and barriers limiting the optimized, patient-centered use arose throughout the first years. This study provides an overview of problems and barriers in the context of DiGA in Germany from different stakeholder perspectives. Nevertheless, with these problems being continuously addressed, digital health apps are trending toward becoming a contributing factor to health care in Germany. Since other countries showed interest in implementing a federally regulated approach similar to Germany, valuable implications can be drawn from this survey.

(J Med Internet Res 2026;28:e73332) doi:10.2196/73332

KEYWORDS

mobile health apps; mHealth; stakeholder; DiGA; digitale Gesundheitsanwendungen; digital health apps; challenges; digital health care; health innovation; regulation; mobile health

XSL•FO
RenderX

Introduction

The inclusion and adoption of digitalized processes, products, and the integration and development of digital health care infrastructure is an ongoing process throughout health care systems around the world. Within the German health care system, the transition from analog to digital, connected structures is facing difficulties and still needs improvement (eg, [1,2]). For example, Germany is lagging behind other countries in the process of implementing an electronic health record [2]. Some solutions, for example, video consultation for statutory health care providers, were used frequently after implementation, though this might be partly due to contact restrictions during the COVID-19 pandemic. Other initiatives, like the above-mentioned electronic health record (ePA), are still in evolving stages [3].

German health legislation emphasized the focus on a more digitalized and connected health care system with the introduction of the Digital Healthcare Act. As part of this legislative agenda, selected digital health apps (DiGA) were added as health care services for statutorily insured people. Since the introduction in 2019, DiGA are reimbursable within the German statutory health insurance (SHI) system. As DiGA can be both mobile apps and web-based interventions, the term digital health apps also incorporates both these options. In general, after the completion of a certification process (DiGA Fast-Track process), DiGA are added to a directory, which lists all reimbursable apps. Once DiGA are listed, they are eligible for use within the German SHI system [4]. Over a year after the legislation was passed, the first DiGA were added to the DiGA directory in September 2020 [5]. There are a number of requirements for eligibility for reimbursement within the German SHI system. All listed applications need to be certified as medical devices according to the European Medical Device Regulation [6]. Furthermore, the listing requires the additional completion of a certification process, which is carried out by a federal institute (Federal Institute for Drugs and Medical Devices [BfArM: Bundesinstitut für Arzneimittel und Medizinprodukte]) covering both general criteria and a positive medical effect for the patients. The general requirements include safety, functionality, quality, as well as data security and protection [7]. The positive medical effect can either be proven with medical benefit, for example, improvement of the health status, or patient-relevant improvement of structures and processes, for example, adherence [8].

There are 2 different alternatives when applying for a listing in the DiGA directory, which have to be declared when starting the application process. All apps applying for a listing have to fulfill the general requirements, as described above. If the positive care effect has been proven in studies complying with the study requirements (§ 10 Digital Health Applications Regulation), apps can apply for a permanent listing. Nevertheless, if there is not enough evidence yet to prove the positive care effect, a preliminary listing for a period of up to 12 months (with a possible extension of 12 months) can be obtained. Systematic data evaluations indicating positive (medical) effects and an evaluation concept have to be submitted. The applying DiGA manufacturer has to prove the medical benefit within a period of up to 12 months, with a subsequent listing if successful [8].

Starting with the implementation of the DiGA directory, a total number of 71 DiGA have been listed (as of July 2025) so far. A total of 14 DiGA have been removed from the DiGA directory, leaving 57 currently listed. About 20% (13 of 57 currently listed) of DiGA are preliminarily added to the DiGA directory, with the final proof of a positive care effect still pending. The remaining 44 have proven the positive care effect through scientific studies and are therefore permanently listed. DiGA for mental diseases currently account for the largest share with almost 50% (28/57, 49%). Endocrine, nutritional, and metabolic diseases, diseases of the musculoskeletal systems, and diseases of the nervous system are other examples of indications that currently have multiple DiGA listed [9].

After successful completion of the certification process, manufacturers can set a price they consider suitable for the first 12 months. If DiGA for comparable indications are already added to the DiGA directory, the price cannot exceed the group price. The group price is calculated for DiGA of similar indications in order to ensure equal pricing [10]. After this, a price is negotiated between the National Association of Statutory Health Insurance Funds (GKV-SV: GKV-Spitzenverband) and the DiGA manufacturer, with repayments if the negotiated price is lower than initially set [11].

In order to have the cost reimbursed by the SHI funds, patients have 2 options. They can receive the application if it is prescribed by a physician or directly from the SHI fund without prior prescription (documented diagnosis, eg, in patients' electronic health record within last 6 mo [12]). Both options are fully reimbursed through SHI [13]. Giebel et al [14] described DiGA care and the corresponding processes in detail.

Patients obtain access to the respective apps through activation codes. About 81% of all activation codes have actually been used. The number of used activation codes increased from 50,122 DiGA in 2021 to 374,377 DiGA in total through 2023 to 861,053 at the end of 2024. But not all DiGA are used at the same rate. The top 6 DiGA account for almost two-thirds of the entered activation codes (approximately 500,000). This includes DiGA for metabolic diseases, diseases of the musculoskeletal system, diseases of the ear, as well as mental and behavioral diseases [15,16]. There are also differences when looking at the distribution of DiGA across indications. DiGA for mental diseases accounts for the largest share, with about 259,000 prescriptions. Metabolic diseases make up about 241,000 prescriptions, and diseases of the musculoskeletal system around 137,000 [16].

The introduction and corresponding processes of DiGA were a new and challenging task for the German health care system. As part of a larger research project (Refer to Acknowledgements), which is funded by the Federal Joint Committee (Gemeinsamer Bundesausschuss; Funding code: 01VSF20007), the presented qualitative analysis identified problems and barriers of DiGA regarding the regulation and organization, as well as the use in everyday practice. Even with increasing prescription numbers since the introduction of DiGA, skeptical public statements by different stakeholders remain

XSL•FO
RenderX

[17,18] (see Results for more), showing there is still disagreement about the disruption and success of DiGA. This poses the question of which problems are still persisting concerning DiGA and their corresponding processes.

With DiGA being evaluated for a lot of aspects prior to admission, we adjusted the aim of the study from DiGA itself—for example, problems with the technical components of DiGA—to certification, financing, and use of the same. Furthermore, since just a few countries have actually implemented digital health apps (similar to DiGA) in their health care system, we decided to focus on problems and barriers from the German infrastructural health care system perspective.

Prior research by Giebel et al [1] described problems with implementation of digital health apps "because of a lack of infrastructure […], [l]ack of access and organizational barriers, such as lack of preparedness of health care systems and reimbursement structures." A review investigating mHealth (mobile health) interventions during the COVID-19 pandemic identified barriers such as insufficient interaction, time, or privacy, but also the nonacademic nature of some apps [19]. This shows that there are studies investigating problems surrounding digital health apps themselves, yet they often include a wide range of interventions, potentially limiting the transferability of the generated results. At the same time—to the best of our knowledge—the perspective of a large collective of stakeholders concerning the process and infrastructure around federally regulated digital health apps has not had an exclusive focus in the existing literature. This is particularly interesting in the context of German DiGA, as different stakeholder groups with competing interests (eg, [20,21]) are involved with the implementation of this new federal health care service. Previous qualitative research in Germany covering DiGA has, for example, focused on certain aspects in a specific medical field (eg, [therapeutic] suitability in rheumatology) [22] or the perspective of a single stakeholder group (eg, physicians, patients) [23,24]. In addition, an article addressing a similar stakeholder group [25] has used a broader focus covering a wide range of aspects of DiGA (eg, potentials of DiGA for patients, suggestions for the improvement of DiGA care).

Our study is adding to current literature by focusing on digital health apps from an infrastructural standpoint. As described above, other studies have also mentioned infrastructural problems and barriers. However, these studies, for example, include external mHealth apps and therefore issues with the implementation into a health care system [1,26] or a broader scope [25]. Our approach covers mHealth apps, which have been implemented as a health care service in a statutory health care system, and the corresponding infrastructural issues from a stakeholder perspective. These include DiGA, SHI, physician, patient, technological, and quality assurance representatives, as well as experts from federal institutions. The stakeholder perspective is also important, as other research has suggested that more stakeholder collaboration is necessary in order to realize the full potential of mHealth apps [26].

This survey is being conducted shortly after the first DiGA were implemented; therefore, it will possibly identify first-hand problems during the implementation process. To provide perspective on how the DiGA market has evolved, the following paper also lays out what adaptations have been made to solve them and what problems may still persist. This study contributes to digital health app research by giving valuable insights for health care policy makers aiming to implement a system similar to the German DiGA approach.

Methods

Overview

The qualitative stakeholder survey was planned, conducted, and the results were reported in accordance with the Consolidated Criteria for Reporting Qualitative Studies (COREQ) checklist (refer to Checklist 1) [27]. We identified the perspective of the stakeholders using semistructured interviews. The study consisted of 3 main stages: In the first stage, the interviews were prepared by recruiting participants and developing an interview guide. The second stage included the actual interviews. During the third stage, interviews were transcribed, and data were analyzed and reported as presented below.

Ethical Considerations

The study was presented to and approved to be in accordance with the current Ethics Regulation by the Ethics Committee of the Medical Faculty of the University Duisburg-Essen (no conventional ethics approval vote was necessary). Additionally, the suitability of the data for publication was approved by the Ethics Committee of the Faculty of Economics of the University Duisburg-Essen.

An agreement covering the terms and conditions of participation and data protection was sent to potential candidates beforehand. All stakeholders had to sign the agreement before participating in the study (refer to the Recruitment section).

The interviews were recorded and pseudonymized during the subsequent transcription process. This ensured that transcripts did not contain any identifying information and the statements made cannot be traced back to the respective person (refer to the Interviews and Transcription sections).

Participants did not receive any compensation for participation in the study.

Recruitment

An initial list of stakeholders was assembled during the application for project funding in 2020 (see Acknowledgments for more). Before starting the recruitment process, this list was reviewed, and necessary adaptations were made. These adaptations were based on the DiGA legislation, which was not available during project application [7,8,13,15,16]. The documents were checked for additional stakeholder groups, and new, not yet included ones were added. The final list included the association of SHI physicians, representatives of patient organizations, the SHI funds, digital health app companies, and experts in the field of quality management, data security and protection, as well as medical device regulation. The recruitment process started in December 2021 and continued until May 2022. A total of 23 organizations were contacted. After each interview, the research team decided if sufficient information for the stakeholder group (in reliance on their involvement in

XSL•FO
RenderX

DiGA processes) was collected. Once saturation was reached for the stakeholder group, no further requests were sent out. However, if stakeholders, who were contacted before, decided to participate in the study, they were also included. Patients (representatives) were only represented by a small subsample, even though they are of particular importance in the context of DiGA, because we did an additional interview and focus group study prior to this study, exclusively covering patients' perspectives [28].

We approached the representatives of the organizations by direct phone call or e-mail. We requested the participation of the employee responsible for quality assurance, if available. If participation in the study was agreed, additional information about the topics of the discussion and the terms and conditions of participation and data protection was transmitted. We provided information on DiGA and the research project in order to ensure all participants had the same base knowledge before taking part in the interviews. Nonresponders to the telephone call were followed up by e-mail (directly after the phone call) containing the aforementioned information and additional general information about the study. An additional follow-up was sent out one week after the initial phone call. Participants had to sign the agreement about participation and data protection before the start of the interview.

Interview Guide

The interview guide was developed according to Helfferich et al [29]. We used a semistructured interview guide to provide guidance and cover the relevant topics, but also leave room to address aspects not included in the interview guide. As the interviews are part of a larger research project, the interview guide was based on results generated earlier in the project. A scoping review [1] covering general problems and barriers related to digital health apps provided the evidence base for the main topics. This was extended using results from a—as mentioned above—qualitative survey with DiGA users and nonusers [28]. As the previous study modules focused on the apps themselves, the interview guide aimed at identifying problems and barriers concerning the infrastructure and process surrounding DiGA. Open questions were used to identify aspects without influencing the interview participants. More precise follow-up questions were asked to inquire more deeply about certain aspects. The interview guide was not available to the study participants prior to the interviews. Only the topics of the interview guide were communicated. We carried out a pretest interview and made adaptations to the interview guide in order to improve comprehensibility and feasibility. As the following interviews were conducted within the planned time frame and the interview partners did not have problems answering the question, the adapted interview guide was used throughout the remaining study.

Interviews

The interviews were conducted by 4 of the authors of this article (NB, CA, GDG, and FP) between February and June 2022. Two of the interviewers are female and 2 male, with 2 senior researchers (PhD: Dr med/Dr rer medic) and 2 junior researchers (MSc/MA). All interviewers are part of the QuaSiApps project team (NB is project lead), with prior experience in qualitative research projects. During the interviews, only the interview study participants and part of the research team were present. There were no prior relationships between the researchers and any of the study participants. Participants did not receive any compensation for participation in the study.

The interviews were carried out using the videoconference platform ZOOM (Zoom Video Communications Inc). One of the researchers moderated the conversation with at least one additional researcher taking notes on the main talking points and overseeing the technical infrastructure. The interview sessions were recorded in accordance with the signed agreement using the integrated recording function of ZOOM. An additional disclaimer was stated by the moderator before starting the recording. The duration ranged from about 50 minutes to 1 hour and 45 minutes.

Transcription

After the conclusion, the recorded interviews were transcribed following the transcription rules by Claussen et al [30]. The transcript was revised by an independent reviewer from the research team, not involved in the transcription. After discussing the remarks and correcting mistakes, the transcript was finalized. All individuals taking part in the interviews, as well as organizations and institutions, were pseudonymized. Therefore, no connections can be made between the statements made and the participants.

Analysis

We analyzed the interviews in accordance with the qualitative content analysis approach by Mayring and the extensions by Kuckartz (content-structuring qualitative analysis) [31,32]. The analysis was carried out as described below using MAXQDA (VERBI), a software tool for qualitative research [33]. The transcripts were imported into MAXQDA, and a coding system containing deductive and inductive codes served as a tool for content extraction. The main topics of the interview guide provided the deductive codes. These included the codes "Problems: Admission," "Problems: Financing" and "Problems: Service Provision," which represent the primary outcomes of our study. We also added deductive codes covering problems regarding the main stakeholders in the context of DiGA, for example, DiGA manufacturers, SHI funds, or patients. Throughout the first coding process, inductive codes were added to the code system. The full coding scheme was reviewed, discussed, and finalized by the research team (Multimedia Appendix 1). After completion of the first coding process and the finalization of the coding scheme, all interview transcripts were coded again using the final code system, which included all inductive and deductive codes. This process was also carried out by 2 researchers independently, with a third researcher reviewing coded transcripts and deciding in case of disagreement. The codes were further grouped into the 3 categories: admission, financing, and service provision. This step was reviewed, discussed, and necessary adaptations were made by the research team. The results of the analysis and corresponding quotes were sent to the study participants for review. We did not receive any additional comments or requests for corrections.

XSL•FO
RenderX

Results

Interview Sample

Of the 23 contacted stakeholders, 2 declined and 21 agreed to participate in the study. None of the 21 stakeholders dropped out after further information about the study and the terms and conditions were sent out. The interviews were conducted with all 21 stakeholders. None of the stakeholders decided to withdraw after the interviews were completed. The characteristics of the 21 included participants are presented in Table 1.

Table . Surveyed stakeholder groups.

Stakeholder group	Participants, n (%)
DiGA representatives[a]	7 (33.3)
SHI representatives[b]	6 (28.6)
Physician representatives	2 (9.5)
Patient representatives	1 (4.8)
Representative from the Federal Joint Committee	1 (4.8)
Technological experts	2 (9.5)
Quality assurance experts	2 (9.5)

[a]DiGA : Digitale Gesundheitsanwendung (digital health application).

[b]SHI: statutory health insurance.

Interview Results

The results are presented for each of the 3 subtopics: admission, financing, and service provision. Anonymous quotes from the interviews are used to illustrate the mentioned aspects. In order to guarantee that, only the stakeholder group (eg, DiGA manufacturer) is mentioned. If not already described in the introduction, we provided additional context throughout the Results section about the specifications of the German DiGA systems in order to ensure a more comprehensive understanding of the presented problems and barriers.

As we aimed to investigate problems and barriers, we also wanted to mention that a large share of stakeholders expressed a generally positive attitude toward DiGA and DiGA processes during the early implementation phase. Better access to health care services for patients in rural regions (patient representative), bridging long waiting times (DiGA representative), and supporting general care processes (SHI representatives) are a few examples of potential benefits of DiGA, which were suggested during the interviews.

Admission

Even though the fast process for digital health apps to become DiGA was generally seen as a good innovation, different stakeholders mentioned problems in the way the DiGA Fast-Track is structured and the corresponding responsibilities. Especially the DiGA manufacturers expressed that there is not enough expertise to evaluate—in particular—the technical, nonevidence-related aspects.

> *So, the BfArM review is very focused on the whole topic of clinical evidence, e.g., study planning. That's their main task, you don't have to talk about it. Yes, so I think in the area of usability, user-friendliness, that could be-, so that's not their topic either. But I don't think it would do it any harm to go into it in more depth […].* [DiGA manufacturer]

Other problems were seen in the high volume of applications for the BfArM and the low level of flexibility (DiGA manufacturer), which was partly credited to the new and challenging nature of implementing a new digital health solution (SHI representative).

There was also criticism about the duration of the DiGA Fast-Track. Upon application by the DiGA manufacturer, a decision has to be made by the BfArM within 3 months [34]. The shift away from the regular admission process of medical products into SHI (quality assurance stakeholder) and the relatively short certification process raised concerns about the quality of the assessment (quality assurance stakeholder, SHI representative, physician representatives).

> *Because the assessment process has been rammed out of the ground by the legislator in a relatively short time, because it all wanted to happen very quickly and that is also - the terminology speaks for itself, fast-track procedure, eg, the fastest possible market entry is in our view a bit in the foreground and then naturally the proof of benefit comes a bit short. So that would be a structural problem or a very fundamental problem, which we still see.* [Physician representative]

A similar aspect with a shifted focus was mentioned by DiGA manufacturers. They criticized that the 3-month assessment period is not put on hold when, after all initial paperwork is turned in, the BfArM requires additional paperwork from the applying DiGA company. This led to very short deadlines to process, produce, and turn in the required documents (DiGA manufacturer).

A diverse range of perspectives was reported surrounding the authorization requirements for both general and evidence-related aspects. On the one hand, the requirements were seen as too low (SHI representative); on the other hand, the steadily

XSL•FO
RenderX

increasing requirements were described as limiting for the flexibility of innovation (DiGA manufacturers).

Since DiGA must be medical products and therefore are/were subject to external certification when obtaining the CE certification, problems surrounding the notified bodies (German: "Benannte Stellen," eg, Technical Inspection Association) were also mentioned. Especially in the early days of DiGA certification, the insufficient levels of knowledge about DiGA within these companies and the low availability of appointments for certification were described (certification companies, DiGA manufacturer):

> *And there are extreme waiting times and a kind of undersupply. So many people have a problem getting appointments at the designated places [certification companies]. So, there's already a big bottleneck and you have to look at how realistic it is to implement it.* [DiGA manufacturer]

There was also criticism concerning the innovation of the preliminary approved DiGA. Physician representatives admonished that the possibility of preliminary approved DiGA yields apps with low evidence to be reimbursed by SHI funds. The equivalent handling of preliminary and permanent approval may also distort the level of evidence between different DiGA (physician representative, SHI representative, DiGA manufacturer).

> *Yes, what is also problematic, for example, is the fact that there are two stages. On the one hand, there is the stage where you are finally included in the BfArM directory and on the other hand, there is the stage where you are temporarily included and then have to provide evidence.* [SHI representative]

> *What is viewed negatively, at least from my point of view in terms of quality, is that we have provisional approval that is not based on RCTs, which has a certain risk because it suggests that these products without evidence have the same efficacy as products with evidence and that is of course a problem, especially for the medical profession, which we have been trying to push into evidence-based medicine for 20 years.* [DiGA manufacturer]

A problem described by DiGA manufacturers regarding the conclusion of the Fast-Track process was the inconsistent handling of the market access. One manufacturer reported that after successfully becoming a DiGA, the first prescriptions are possible the next day, but in their case, no listing in the prescription software led to no prescriptions within the first days. A different manufacturer reported the same problem with no possibility of prescriptions due to the unavailability of a necessary interface port to the prescription software (DiGA manufacturer).

Financing

The interviewed stakeholders mostly talked about the price of DiGA, the process of negotiation between the SHI funds and the DiGA manufacturer, and the price-related policy. DiGA manufacturers said the price expectations of health insurance funds indirectly show low willingness to negotiate and lead to high initial prices (DiGA manufacturer).

> *The offers that the GKV-SV distributes here in the market mean that manufacturers anticipate that they will take the highest possible price at the time when it is possible, so that they can somehow reduce some of the debt that they have accumulated.* [DiGA manufacturer]

They also criticized the superior negotiation position of the GKV-SV—the head organization of all SHI funds in Germany, since they get to deal with multiple DiGA manufacturers, but a single DiGA manufacturer only gets to deal with the GKV-SV once (DiGA manufacturer). One health insurance fund representative mentioned that—even with this negotiation position—they have difficulties achieving favorable SHI positions in price negotiations (SHI representative).

Multiple stakeholders took a look at the finance aspect from the DiGA manufacturers' view and laid out that the financing of the app and corresponding activities (eg, evaluation studies, personnel) are challenging, especially for small companies (DiGA manufacturer, technological expert, patient representative).

> *However, due to the hurdles posed by the legislator, it is not always easy for start-ups to raise the basic capital to gain a foothold in the market, especially when it comes to financing.* [Technological expert]

One SHI representative mentioned that the refund of excess payments after price negotiations might lead to a market shrinkage, due to the superior financial resources of large companies compared with smaller DiGA manufacturers. Additionally, besides general criticism, a separate regulation for preliminary and permanently listed DiGA was called for, as the described regulation above is currently in place regardless of admission status. The necessity of usage and quality-related pricing was also pointed out (SHI representative).

As of now, DiGA are being reimbursed as soon as the patient downloads the app and enters the DiGA activation code, provided by the health insurance fund. Therefore, reimbursement is independent of usage. There were concerns among SHI representatives about the financing of DiGA without knowing if the app is actually used by the patient.

> *Problems are certainly, as an initial survey of insured persons by the health insurance fund [name anonymized] has also shown, problems are certainly that you don't know whether and how the insured person uses this DiGA in the end.* [SHI representatives]

Service Provision

In the early days of DiGA, there was disagreement among the questioned stakeholders about how well the communication between different stakeholders was going. Especially, the manufacturers emphasized that negative public reporting does not help DiGA flourish. According to them, different organizations made negative remarks about DiGA or even gave out "warnings" (DiGA manufacturer). General skepticism

XSL•FO
RenderX

among stakeholders when it comes to digitalization may also have an impact on how well DiGA are perceived. Multiple DiGA manufacturers added that this was accompanied by negative public statements and comments by SHI representatives about DiGA, which further hindered the development.

> *I often see very sweeping arguments being used when DiGA are criticized. Often the basics are not so well known, the legal and that is of course-, so DiGA are sometimes also seen as a symbol for digitization per se. And then perhaps no distinction is made as to whether it is a project such as an e-prescription or a DiGA, but it is simply a blanket term for digitization and there is a great deal of skepticism* [DiGA manufacturer]

The DiGA directory and the lack of (technical) information presented in it were also criticized among stakeholders. SHI and physician representatives both reported that they did not, or only in a limited fashion, receive access to selected DiGA (SHI representative, physician representative). The DiGA directory could not fill this knowledge gap sufficiently either, as especially SHI representatives pointed out that information about DiGA was too complex and no information on the quality of the applications was provided:

> *We would say that the issue of transparency, or rather the lack of transparency, is a major problem from our point of view. Also, in terms of the lack of comparison options for the apps. We are of the opinion - at the moment, the DiGA directory, the BfArM directory is basically the source that you can look at. […] In other words, to prepare it differently for patients and physicians. We still see a lot of room for improvement, especially as it is currently the case that - when you look at the DiGA directory, you can't really find anything about quality at first glance.* [SHI representatives]

Low cooperation was also seen in the work of medical software companies (German: “Praxisverwaltungssoftware [PVS] – Hersteller,” PVS manufacturer) responsible for technical solutions in SHI medical practices. This showed in slow adoption of existing practice software (Technological expert), establishment of own structures for app prescriptions (DiGA manufacturer), and little interest and knowledge in new DiGA policies:

> *And then I call a medium-sized PVS manufacturer and they say: "No, we don't make apps. “[…] They just didn't realize it and of course, the legislator doesn't have to do the work for us now and inform all the PVS manufacturers and message them to implement all this, but it’s important for those who want to make the DiGA project a success to understand: "The PVS manufacturers are sometimes to often not at all helpful to harmful."* [DiGA manufacturer]

Stepping away from the superordinate perspective and looking at problems and barriers to better use of DiGA within the direct care context, problems surrounding DiGA prescribing physicians, patients, and their interactions were reported. Especially in the early days of DiGA, the interviewed stakeholders reported insufficient knowledge about DiGA, their content, and the corresponding regulations among physicians (DiGA manufacturer, SHI representative, patient representative, physician representative). This was generally acknowledged by physicians while emphasizing the limited capacity for DiGA within a regular workday:

> *Today I had another Monday where [I] didn't even know where to stack the patients. And then when some people come in who have problems with their app and need technical support, I find it difficult in some cases how to accommodate that in an open consultation. That might be a bit out of touch with reality.* [Physician representative]

They also reported heterogeneous interests and skepticism among colleagues and concerns of (partially) being replaced by DiGA at some point (physician representative).

Different stakeholders agreed about insufficient awareness of the information services and content of the DiGA among patients. Patients did not know what DiGA were or how to claim them. The process of going through the SHI fund to receive the activation code hindered patients from using DiGA (DiGA manufacturer, patient representative). Patients also showed low compliance levels according to an SHI representative, which they found out through a company internal survey. There was also concern among patients that the “normal” treatment option may be withheld from them if they opt in to using a DiGA (patient representative).

> *And another problem for patients is that they sometimes have the feeling that if they use a DiGA, they will be deprived of another therapy, such as behavioral therapy. So, if you have the feeling that you have to choose either or, then in my experience most patients tend to go down the usual treatment routes.* [Patient representative]

Stakeholders also expressed similar opinions regarding problems with the prescription process. In some cases, low prescription quality (eg, missing diagnosis) was registered by SHI funds when processing DiGA prescriptions (SHI representative). As mentioned above, the prescription process of DiGA was also criticized, and a digitalized adaptation was called for (DiGA manufacturer). Closely interrelated with the prescription process, the production and distribution of DiGA activation codes were described as problematic in the early days of DiGA. Incorrect codes were sent out, which led to patients not being able to use the DiGA (DiGA manufacturer, SHI representative).

Two stakeholders criticized the missing combination of DiGA use and therapy by physicians. They reported that, especially for risk-associated DiGA (eg, DiGA for mental diseases), monitoring and guided app use might be necessary (SHI representative, patient representative).

> *There are also DiGA that are associated with risks, especially when I think of DiGA for depression. This prescription and use should at least be accompanied by a physician. So, I see problems here if the physician basically just prescribes and doesn't stay*

XSL•FO
RenderX

in conversation with the insured person about how the DiGA actually works. [SHI representative]

At the time of the study, DiGA were meant to be used mainly by the patients themselves. Active involvement of physicians was possible, but only a few DiGA incorporated such tools [9]. DiGA manufacturers described the integration of physicians with existing legal regulations as difficult and added that this may cut back on the potential of DiGA (DiGA manufacturer).

Discussion

Principal Findings

The aim of this study was to identify problems and barriers regarding the admission, financing, and use of DiGA. As other studies have focused on the apps themselves, we wanted to determine which problems concerning the infrastructure and processes influence optimal use of this new digital health care service. The findings show that some problems can be attributed to the early days of DiGA, while other issues are more prevalent for long-term success. Regarding the admission, the DiGA Fast-Track itself was criticized because of insufficient levels of expertise among the responsible institutions and the duration of the process. From a financial perspective, problems regarding the price negotiations and reimbursement were most frequently discussed. The superior negotiation position of the GKV-SV was pointed out, as well as criticism that reimbursement is independent from DiGA usage. DiGA service provision was criticized on a superordinate level because of, for example, the public perception of these new health care services and negative statements among involved stakeholders. As aforementioned, insufficient knowledge and skills surrounding DiGA were also a problem among patients and health care professionals, along with early issues of the prescription process and a lack of integration of physicians into DiGA use.

Admission

As mentioned before, some of the identified problems and barriers can be attributed to the short existence of DiGA at the time the survey was conducted (first DiGA: 09/2020, Interviews: 02-06/2022). DiGA manufacturers pointed out the challenges associated with a 3-month certification process, including review and additional paperwork. To this day (July 2025), only 25 applications were rejected after the DiGA Fast-Track process. Notably, 124 applications have been withdrawn from the certification process [34]. As we cannot view all reasons of withdrawal, we can assume that, at least to some extent, it may have been due to deadlines, which were not or could not be met. This leaves the question of whether just the quality of applications of digital health apps is not good enough, or if process adjustments may have the chance of allowing more qualitatively high digital health apps access to the DiGA directory.

In comparison to the 57 currently listed DiGA, over twice as many applications and digital health apps have not been included in the directory. In addition to withdrawals (n=124) and negative decisions (n=25), 14 DiGA have been deleted from the directory after originally being included (total number of DiGA not included: n=163). Again, as the reasons for unsuccessful applications cannot be viewed, this may also suggest that the DiGA Fast-Track with its requirements has the potential to ensure quality and reliability of DiGA [34].

This may be a small indicator that concerns about the DiGA Fast-Track allowing digital health apps to easily enter the German statutory health care market should be examined in more detail. The length of the process was not the only argument for this position. We identified a disagreement about the strength of the requirements of the DiGA Fast-Track as well. While DiGA manufacturers highlighted a high (but bearable) burden for successful completion of the DiGA Fast-Track, especially SHI-related stakeholders mentioned concerns about the (missing) strength of the requirements. Already in the early days of DiGA, the DiGA Fast-Track and its framework (eg, preliminary approved DiGA) were subject to SHI criticism [35]. As laid out above, the high number of withdrawn and negatively decided DiGA applications indicates that existing requirements cannot be met by a large share of applying companies. In addition, the underlying legislation (Digital Health Applications Regulation) has been updated frequently since its introduction in 2021 [36]. It has to be kept in mind that up until the “Act to accelerate the Digitization of the Healthcare System (Digital Act–DigiG; effective since March 2024)”, only DiGA of low-risk classes were reimbursable within SHI. The expansion to higher-risk classes within the act may influence the appropriateness of the DiGA Fast-Track requirements. For example, the act includes the requirement for DiGA of higher risk classes to prove the medical benefit through a prospective comparative study [37].

The preliminary approval of DiGA caused some controversy, as it presents a new and never-before-seen approach to introducing health care services into the SHI system. We like to point out that even if no final results are necessary to obtain preliminary approval, there are still requirements tied to this status. In addition to meeting all “general” requirements (eg, security, functionality, quality, data security), a preliminary systematic data evaluation as well as an evaluation concept is necessary, plausibly showing that the medical benefit can be proven within the preliminary approval period [4]. As this still bears the possibility for the evaluation studies to show the corresponding DiGA does not provide a medical benefit, it opens the door for an early market entry of new, innovative, potentially disruptive technologies. Nevertheless, DiGA are used by patients with diseases. The worst case for any health care services would be harming the patient. To ensure health care services are safe and beneficial to the patient, studies providing evidence of such effects with final, statistically significant information are conducted. These studies must also fulfill high-quality standards. Existing literature shows that some DiGA have deficiencies regarding the quality of the studies showing the medical benefit (for example, lack of blinding, inhomogeneous dropout rates) [38]. Therefore, in the case of DiGA, the balancing of innovation and (final) evidence may currently be prioritizing innovation.

The implementation of admission processes for DiGA-equivalent digital health apps in other European countries (eg, France and Belgium) shows that there is an agreement about DiGA being a necessary part of health care systems in the future.

XSL•FO
RenderX

This is underlined by ongoing discussions about the implementation of admission processes in additional European countries, for example, Spain, Austria, and Italy. However, some countries do not prioritize this topic as of now, which cannot be neglected [39].

Financing

Finally, one of the most talked-about topics in the public discourse surrounding DiGA is their price and the corresponding processes, for example, negotiations between DiGA manufacturers and the National Association of SHI funds. Some of the interviewed experts (eg, physician representatives, SHI representatives) criticized the possibility for DiGA providers to set their own price for the first 12 months of approval. To keep DiGA prices within a reasonable and comparable range, the legislator made adjustments regarding the pricing processes. A collective agreement between the National Association of SHI funds and leading DiGA manufacturer associations was put in place, introducing maximum prices for DiGA, setting the range for the price of the DiGA in relation to other DiGA treating the same or a comparable disease [40].

The opposing views by SHI representatives and DiGA manufacturers regarding the price of DiGA may be due to their competing interests. As the SHI funds are paying for the use of DiGA by their insured persons, lower and criteria-based (eg, quality, usage) prices can lead to lower costs for the corresponding SHI fund. At the same time, DiGA manufacturers are financing (to a significant extent) their company through the income generated from patients using their app, therefore having an interest in higher prices.

Service Provision

A problem medical providers mentioned was the current prescription process, calling for a more digitalized prescription process ("Digital health solution needs digital prescription" – DiGA manufacturer). This was shared by patients, as some of them did not want to use the DiGA because of the complicated prescription process. With the mandatory introduction of the e-prescription in Germany starting in January 2024, digital prescriptions are now the standard within German health care. They are currently only used for prescribable drugs, but are planned to be extended to DiGA in 2025 [41]. This has the potential of further boosting DiGA as the prescription process and, therefore, the access to DiGA can improve once the e-prescription is established [42]. With problems in the early days of e-prescription now solved, they are mainly seen as an improvement among representatives within the German SHI system [43] and may therefore contribute to more DiGA use when extended. However, even if they are not the main target group of DiGA, the preferences of people with low digital literacy and non–digital natives also need to be incorporated to create an option of low-barrier access to DiGA. This may be of particular relevance for indications that have a substantial number of patients in higher age groups.

A problem was also presented in the low knowledge and skills around DiGA by both the prescribing (medical providers) and the using parties (patients). One tool to tackle this challenge was the DiGA directory, providing a range of information separately presented for health care professionals and patients [9]. Our study showed that for a share of the involved stakeholders, this goal has not been met yet. As growing prescription and usage numbers indicate (see SHI Reports [15,16] for more), we assume the lack of knowledge was—at least to some part—due to the nature of the new digital health service and the introduction during the COVID-19 pandemic. The importance of knowledge and skills regarding digital health care services has also been emphasized by the World Health Organization. Borges do Nascimento et al [44] identified digital health services as possibly "a real asset to the health and care workforce." To realize these potentials, improving the digital health literacy among health care professionals is a fundamental step.

To further boost DiGA and enable it to thrive, more integration of medical providers in the use of DiGA might be a solution. International Studies have shown that this also has the potential to increase the adherence of the patients [45,46]—which was criticized by SHI representatives—assuring that reimbursed DiGA are being used as good and intended as possible to create a medical benefit for the patient. A large German health insurance fund reported that almost 40% of insured people with prescribed DiGA did not use the app over the intended time span. Of the people who discontinued using the DiGA, only about 20% reported it was due to improved health [47]. To improve adherence and guide the patient, some DiGA have integrated modules to include medical providers [9]. Monitoring of patient data and consultation of the patient at follow-up visits are some examples of these tasks. This appeal results in a call for reimbursement of the corresponding tasks (eg, monitoring of patient data, consulting services) carried out by the medical provider. Up to today, only a few DiGA have negotiated reimbursement for their DiGA-supporting tasks [48]. Applying similar structures to more DiGA may improve adherence among patients and, therefore, the conditions for DiGA to provide a medical benefit to the patient.

However, more integration of health care providers cannot be seen as an isolated option to improve DiGA use. One fundamental component of the long-term success of health care services is the acceptance among medical providers and patients. Since the evaluation reports [15,16] have shown that over 90% of DiGA are prescribed by medical providers, they play a key role as they are (for a large share) the gatekeepers in the distribution of this health care service. At the same time, acceptance of patients is also important, since they are the ones using DiGA. A web-based survey [49] covering potential users in Germany showed that there is a generally high willingness (76%) to use digital health apps. Still, the authors conclude that "targeted communication around effortless usage of mHealth services across age groups and demographics" is necessary to improve digital health app adoption. Another study [50] (covering mental illnesses) adds to this conclusion, for example, outlining that there is unequal access to digital health technologies for some patient groups, and this might lead to the worsening of health inequalities. This emphasizes the importance of addressing the acceptance of medical providers and patients in order to improve digital health app use.

XSL•FO
RenderX

Many of the presented problems show a strong reliance on the German health care system. Nevertheless, the results present an overview of relevant problems for researchers and policymakers from different health care system approaches when considering the implementation of digital health apps. The early days of DiGA contained a wide range of first impressions and statements from stakeholders. As we found out within this study, there was disagreement among involved parties about different aspects of DiGA within the German health care market. There are various reasons why stakeholders potentially disagreed, such as competing interests or system logic. On the one hand, DiGA are brought into the market by DiGA manufacturers, who get reimbursed for their service. On the other hand, as DiGA are part of the statutory health care system in Germany, they are reimbursed by SHI funds. These 2 stakeholders have competing interests, and this showed in opposing statements. One example is the authorization requirements for DiGA, where SHI representatives suggested the criteria are too low, when at the same time DiGA manufacturers described the increasing requirements as challenging (refer to Results: Admission). Therefore, it has to be kept in mind that some of these contradictory statements are also influenced by interests and system logic as described above.

Nevertheless, DiGA made their start in 2020 and have been growing since. As laid out throughout this article, this was accompanied by challenges along the way, not unexpectedly due to the iterative nature of the legislative process. On the side of the stakeholders involved in the actual distribution of DiGA, many challenges can be attributed to the early days. These aspects, therefore, should be addressed early on when implementing a DiGA-equivalent health care service in other countries.

Limitations

First, we interviewed stakeholders from relevant stakeholder groups. Nevertheless, within each stakeholder group, we could not completely ensure representativeness; for example, we did not interview a DiGA manufacturer for each disease in the DiGA directory. The sample is also not representative of factors like the number of insured people with the SHI funds or the share of the corresponding DiGA indication among all listed DiGA. In order to still cover a large share of the involved stakeholders, we included central associations of the corresponding stakeholder groups (eg, GKV-SV, DiGA-Spitzenverband [DiGA-SV]). Even though we tried to cover all relevant stakeholder groups by sending out multiple requests, some declined participation or opted for later involvement in the research project, for example, BfArM. We also only included one patient representative, which may lead readers to think this important stakeholder group is underrepresented in our study. We conducted a focus group and interview study exclusively covering the patient's perspective prior to this study and therefore refer to the corresponding article for more insights into patients' views [28].

We want to point out that all statements made are subjective and only represent the view of the interviewed stakeholder and may or may not translate to DiGA in general. The roles of the participants or the stakeholder group, which they represented, and the corresponding statements may have been influenced by personal interests or a political agenda. Statements were checked with existing evidence but remain subjective to some extent. Since this research paper aims to report problems surrounding DiGA from different stakeholder perspectives, this limitation has to be accepted in order to reach the research goal.

When conducting an interview study, the role and position of the involved researchers is an aspect that can generally influence the results of the study (eg, personal preconceptions regarding the research topic, prior collaboration between researchers and interview participants). To minimize the influence of this limitation, all steps of the study were generally carried out by at least 2 researchers. The generated data was coded and analyzed by 2 researchers independently, with subsequent cross-checking by a third experienced senior researcher. In addition, all transcripts and codes were quality assured by a researcher not primarily involved in the corresponding interview in order to identify any analytical results and conclusions influenced by preconceptions.

Another limitation is the period of time during which we conducted the interviews. Since we started conducting the interviews in February 2022, there have been only a few DiGA inducted into the DiGA directory. This ultimately led to a few prescribed DiGA, which decreased the representativeness of the generated evidence. To account for this limitation, we laid out which of the presented problems have already been solved and how processes and the general perception of DiGA may have improved since then. Still, we cannot fully guarantee that some of the aspects we described may be transitional or only related to early implementation efforts and may have been resolved after the study period. At the same time, we assume that some of the problems, for example, the imbalance between stakeholders during price negotiations or the different perceptions toward the preliminary approval of DiGA, may not be transitional and still have an influence on the current DiGA landscape.

The online format (interview via videocall) may have also impacted the quality of the generated data, as it was not possible to control, for example, for additional people in the room or technical problems. However, since the interviews were conducted almost 2 years after the COVID-19 pandemic started, online video tools were already common. There were also no major technical problems during the interviews. There is research indicating that there are no significant differences between online and face-to-face interviews [51,52].

Conclusions

With the introduction of DiGA to the German health care system, a step toward more digitization was taken, with the potential of improving care for patients. Nevertheless, there was also skepticism and risks to how well the new technology would improve and fit into the German health care system. The identified problems can be seen as "necessary" steps in the development and evolution of DiGA, for example, early problems with activation codes, insufficient knowledge about DiGA among involved actors, and different perceptions about the prices of DiGA. Some of the barriers have already been addressed along the way, for example, different adaptions with

XSL•FO
RenderX

pricing of DiGA. This goes along with new health policies further expanding DiGA to higher-risk classes, which opens up opportunities for new DiGA and, therefore, more patients to benefit from it. This also shows in a higher number of DiGA being prescribed and more DiGA being added to the DiGA directory.

Therefore, we are concluding that problems and barriers have been and are still impacting DiGA on the way to becoming a contributing component of patient care. To identify the current state of DiGA and which problems may have been solved and additionally arisen through the years, new research investigating the perspective of a similar group of stakeholders may be necessary. We also want to point out that although we aimed at identifying problems and barriers in the early days of German DiGA use, one result of our study was that a large share of the questioned stakeholders had a positive attitude toward the potential of DiGA and the processes in the early implementation phase. Yet, as our study has shown, there is still room for improvement so that DiGA can reach its proclaimed potential.

Finally, the identified problems and barriers can help other countries when implementing DiGA-equivalent digital health apps. With conclusions drawn from the presented challenges, an optimized implementation process may lead to an early realization of positive effects.

Acknowledgments

Generative artificial intelligence was not used for any part of this research article.

Funding

This qualitative stakeholder survey is part of the research project called “Continuous quality assurance of health apps used in SHI standard care (QuaSiApps)” funded by the German Federal Joint Committee (Gemeinsamer Bundesausschuss, G-BA; funding code: 01VSF20007). We acknowledge support from the Open Access Publication fund of the University of Duisburg-Essen.

Data Availability

Information and data generated or analyzed during the research project (QuaSiApps) are available from the corresponding author on reasonable request.

Authors' Contributions

All authors were involved in the conception and design of the study. Participants were recruited by FP and CA. The interview guide was constructed and redefined by FP, CA, GDG, and NB. FP, NB, CA, and GDG conducted the interviews. The transcription and qualitative analysis were performed by FP. FP wrote the first draft and the revision of the manuscript, which was critically revised by NB, CA, GDG, SN, BK, KB, and JW. The final version of the manuscript was read and approved by all authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Coding scheme.
[DOCX File, 19 KB - jmir_v28i1e73332_app1.docx]

Checklist 1
COREQ checklist.
[PDF File, 424 KB - jmir_v28i1e73332_app2.pdf]

References

1. Giebel GD, Speckemeier C, Abels C, et al. Problems and barriers related to the use of digital health applications: scoping review. J Med Internet Res 2023 May 12;25:e43808. [doi: 10.2196/43808] [Medline: 37171838]
2. Digitalisierung im Gesundheitswesen dringend weiterdenken [Urgent need to further develop digitization in health care]. Deutsches Ärzteblatt [German Medical Journal]. 2022. URL: https://www.aerzteblatt.de/nachrichten/135229/Digitalisierung-im-Gesundheitswesen-dringend-weiterdenken [accessed 2024-07-22]
3. Die elektronische Patientenakte (ePA) [The electronic health record]. Bundesministerium für Gesundheit [Federal Ministry of Health].: Federal Ministry of Health; 2024. URL: https://www.bundesgesundheitsministerium.de/elektronische-patientenakte.html [accessed 2025-01-06]
4. The Fast-Track-Process for digital health applications (DiGA) according to § 139e SGB V.: Federal Institute for Drugs and Medical Devices; 2023. URL: https://www.bfarm.de/SharedDocs/Downloads/DE/Medizinprodukte/diga_leitfaden.pdf?__blob=publicationFile [accessed 2025-12-27]

XSL•FO
RenderX

5. BfArM nimmt erste "Apps auf Rezept“ ins Verzeichnis digitaler Gesundheitsanwendungen (DiGA) auf [BfArM adds first “apps on prescription” to its directory of digital health applications (DiGA)].: Federal Institute for Drugs and Medical Devices; 2020. URL: https://www.bfarm.de/SharedDocs/Pressemitteilungen/DE/2020/pm4-2020.html [accessed 2024-07-22]
6. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. EUR-Lex. 2017. URL: http://data.europa.eu/eli/reg/2017/745/oj [accessed 2026-01-07]
7. § 139e Directory for Digital Health Applications. Social Security Code V. In: Statutory Health Insurance: Federal Law Gazette; 1988:2477.
8. Digital health applications regulation of April 8, 2020 (Federal Law Gazette). : BGBl; 2024 URL: https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/Gesetze_und_Verordnungen/GuV/D/DiGAV_RefE.pdf [accessed 2026-01-06]
9. Federal institute for drugs and medical devices. DiGA-directory. 2024. URL: https://diga.bfarm.de/de [accessed 2024-07-22]
10. Group prices and thresholds. National Association of Statutory Health Insurance Funds. 2024. URL: https://www.gkv-spitzenverband.de/krankenversicherung/digitalisierung/kv_diga/hoechstbetraege_und_schwellenwerte/hb_sw.jsp [accessed 2025-01-06]
11. § 134 Agreement between the National Association of Statutory Health Insurance Funds and the Manufacturers of Digital Health Applications on Remuneration Amounts; Regulation Authorization.: Statutory Health Insurance; 1988. URL: https://www.gesetze-im-internet.de/sgb_5/__134.html [accessed 2026-01-07]
12. Eine DiGA beantragen [Appyling for a DiGA].: Techniker Krankenkasse [Techniker health insurance fund]; 2024. URL: https://www.tk.de/techniker/krankheit-und-behandlungen/praxisbesuch-und-klinikaufenthalt/kompetent-als-patient/digitale-gesundheitsangebote-verstehen/digitale-gesundheitsanwendungen-antrag-stellen-2116854 [accessed 2025-08-20]
13. § 33a Digital Health Applications. Bundesamt für Justiz.: Statutory Health Insurance; 1988. URL: https://www.gesetze-im-internet.de/sgb_5/__33a.html [accessed 2026-01-19]
14. Giebel GD, Abels C, Börchers K, et al. Integration of digital health applications into the German healthcare system: development of “The DiGA-Care Path”. Front Health Serv 2024;4:1372522. [doi: 10.3389/frhs.2024.1372522] [Medline: 38545382]
15. Report by the National Association of Statutory Health Insurance Funds on the utilization and development of care with digital health applications (DiGA-Bericht). : National Association of Statutory Health Insurance Funds; 2023 URL: https://www.gkv-spitzenverband.de/media/dokumente/krankenversicherung_1/telematik/digitales/2023_DiGA_Bericht_GKV-Spitzenverband.pdf [accessed 2026-01-07]
16. Report by the National Association of Statutory Health Insurance Funds on the utilization and development of care with digital health applications (DiGA-Bericht). : National Association of Statutory Health Insurance Funds; 2024 URL: https://www.gkv-spitzenverband.de/media/dokumente/krankenversicherung_1/telematik/digitales/2024_DiGA-Bericht_final.pdf [accessed 2026-01-19]
17. Digitalisierung stößt bei Ärzteschaft auf große Skepsis [digitization met with great skepticism among doctors]. Deutsches Ärzteblatt [German Medical Journal]. 2022. URL: https://www.aerzteblatt.de/nachrichten/131027/Digitalisierung-stoesst-bei-Aerzteschaft-auf-grosse-Skepsis [accessed 2024-07-22]
18. KVB-Vorstand warnt vor Gesundheits-Apps [KVB board warns against health apps]. Ärzte Zeitung [The German Doctors Journal]. 2020. URL: https://www.aerztezeitung.de/Wirtschaft/KVB-Vorstand-warnt-vor-Gesundheits-Apps-413035.html [accessed 2024-07-22]
19. Amiri P, Nadri H, Bahaadinbeigy K. Facilitators and barriers of mHealth interventions during the Covid-19 pandemic: systematic review. BMC Health Serv Res 2023 Oct 28;23(1):1176. [doi: 10.1186/s12913-023-10171-w] [Medline: 37898755]
20. Stellungnahme des GKV-Spitzenverbandes vom 08.04.2021 zum Entwurf eines Gesetzes zur digitalen Modernisierung von Versorgung und Pflege [Statement by the GKV-spitzenverband (National Association of Statutory Health Insurance Funds) dated April 8, 2021, on the draft law on the digital modernization of health care and nursing care]. : National Association of Statutory Health Insurance Funds; 2021 URL: https://www.gkv-spitzenverband.de/media/dokumente/presse/p_stellungnahmen/2021_2023/210409_GKV-SV_SN_DVPMG.pdf [accessed 2026-01-07]
21. Geier AS. Digitale Gesundheitsanwendungen (DiGA) auf dem Weg zum Erfolg – die Perspektive des Spitzenverbandes Digitale Gesundheitsversorgung [Digital health applications (DiGA) on the road to success – the perspective of the German Digital Healthcare Association]. Bundesgesundheitsbl 2021 Oct;64(10):1228-1231. [doi: 10.1007/s00103-021-03419-5]
22. Boy K, May S, Labinsky H, et al. Back on track - digital health applications to treat back pain of rheumatic patients? Results of a qualitative interview study. Rheumatol Int 2024 Dec;44(12):2977-2987. [doi: 10.1007/s00296-024-05726-x] [Medline: 39340666]
23. Zugaj MR, Germeyer A, Kranz K, Züger A, Keßler J. Experiences of patients with endometriosis with a digital health application: a qualitative analysis. Arch Gynecol Obstet 2024 Oct;310(4):2253-2263. [doi: 10.1007/s00404-024-07651-7] [Medline: 39066952]
24. Schroeder T, Seaman K, Nguyen A, Gewald H, Georgiou A. Enablers and inhibitors to the adoption of mHealth apps by patients - a qualitative analysis of German doctors’ perspectives. Patient Educ Couns 2023 Sep;114:107865. [doi: 10.1016/j.pec.2023.107865] [Medline: 37356116]

XSL•FO
RenderX

25. Heidel A, Hagist C, Spinler S, Schoeneberger M. Removing dust from the German health care system by introducing health apps into standard care: semistructured interview study. JMIR Hum Factors 2023 May 4;10:e42186. [doi: 10.2196/42186] [Medline: 37140980]
26. Kao CK, Moawad A, Bhargava A. Chapter 4 - mobile health apps: current state, barriers, and future directions. In: The Digital Doctor: Academic Press; 2025. [doi: 10.1016/B978-0-443-15728-8.00019-7]
27. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
28. Giebel GD, Abels C, Plescher F, et al. Problems and barriers related to the use of mHealth apps from the perspective of patients: focus group and interview study. J Med Internet Res 2024 Apr 23;26:e49982. [doi: 10.2196/49982] [Medline: 38652508]
29. Helfferich C. Leitfaden- und Experteninterviews [Guideline and expert interviews]. In: Baur N, Blasius J, editors. Handbuch Methoden Der Empirischen Sozialforschung [Methods of Empirical Social Research] 2014. [doi: 10.1007/978-3-531-18939-0_39]
30. Claussen J, Jankowski D, Dawid F. Aufnehmen, Abtippen, Analysieren: Wegweiser zur Durchführung von Interview und Transkription [Recording, Typing, Analyzing: Guide to Conducting an Interview and Transcription]: Hannover; 2020.
31. Mayring P. Qualitative Inhaltsanalyse: Grundlagen und Techniken [Qualitative content analysis: basics and techniques]. In: Handbuch Methoden Der Empirischen Sozialforschung [Handbook of Methods in Empirical Social Research]: Beltz Publishing; 2015. [doi: 10.1007/978-3-658-21308-4_42]
32. Kuckartz U. Qualitative Inhaltsanalyse Methoden, Praxis, Umsetzung Mit Software Und Künstlicher Intelligenz [Qualitative Content Analysis: Methods, Practice, Computer Support]: Beltz Juventa; 2018.
33. MAXQDA. URL: https://www.maxqda.com/de/produkte/maxqda [accessed 2024-07-22]
34. Wissenswertes zu DiGA [Interesting facts about DiGA]. Bundesinstitut für Arzneimittel und Medizinprodukte [Federal Institute for Drugs and Medical Devices]. 2024. URL: https://www.bfarm.de/DE/Medizinprodukte/Aufgaben/DiGA-und-DiPA/DiGA/Wissenswertes/_node.html [accessed 2025-08-20]
35. Urbanek M. Kassen-Kritik and DiGA: Zentrale Anforderung ist medizinischer Nutzen [Health insurance criticism of DiGA: key requirement is medical benefit].: Ärzte Zeitung [The German Doctors Journal]; 2021. URL: https://www.aerztezeitung.de/Wirtschaft/Kassen-Kritik-an-DiGA-Zentrale-Anforderung-ist-medizinischer-Nutzen-416271.html [accessed 2024-07-22]
36. Regulation on the procedure and requirements for assessing the eligibility for reimbursement of digital health applications in statutory health insurance (Digital Health Applications Regulation - DiGAV). Bundesinstitut für Arzneimittel und Medizinprodukte [Federal Institute for Drugs and Medical Devices].: Federal Law Gazette I; 2020. URL: https://www.gesetze-im-internet.de/digav/BJNR076800020.html [accessed 2024-07-22]
37. Act to accelerate the digitization of the healthcare system (Digital Act – DigiG). Bundesinstitut für Arzneimittel und Medizinprodukte [Federal Institute for Drugs and Medical Devices].: Federal Law Gazette I; 2024. URL: https://www.recht.bund.de/bgbl/1/2024/101/VO.html [accessed 2025-01-06]
38. Kolominsky-Rabas PL, Tauscher M, Gerlach R, Perleth M, Dietzel N. Wie belastbar sind Studien der aktuell dauerhaft aufgenommenen digitalen Gesundheitsanwendungen (DiGA)? Methodische Qualität der Studien zum Nachweis positiver Versorgungseffekte von DiGA [How reliable are studies of currently approved digital health applications (DiGA)? Methodological quality of studies demonstrating positive healthcare effects of DiGA]. Z Evid Fortbild Qual Gesundhwes 2022 Dec;175:1-16. [doi: 10.1016/j.zefq.2022.09.008]
39. Bucksch M. Der DiGA-Fast-Track – bald auch europaweit? Zulassung von DiGA in der EU (2024): Überblick zum aktuellen Stand in allen Ländern der Europäischen Union [The DiGA fast track – soon to be available across Europe? Approval of DiGA in the EU (2024): overview of the current status in all countries of the European Union]. Quickbird Medical. 2023. URL: https://quickbirdmedical.com/diga-eu-europa-schweiz/ [accessed 2024-07-22]
40. Agreement according to § 134 Abs. 4 und 5 SGB V. National Association of Statutory Health Insurance Funds. 2023. URL: https://www.gkv-spitzenverband.de/media/dokumente/krankenversicherung_1/telematik/digitales/DiGA_Rahmenvereinbarung_nach_134_Abs_4_und_5_SGB_V_Stand_27.02.2024.pdf [accessed 2025-01-06]
41. Lau T. Gematik-Konzept: DiGA mit der E-Rezept-App verordnen [Gematik concept: prescribing DiGA with the e-prescription app].: Deutsches Ärzteblatt [German Medical Journal]; 2024 Aug. URL: https://www.aerzteblatt.de/archiv/240645/gematik-Konzept-DiGA-mit-der-E-Rezept-App-verordnen [accessed 2025-08-20]
42. Elektronisches Rezept (E-Rezept) [The electronic prescription]. Bundesministerium für Gesundheit [Federal Ministry of Health]. 2024. URL: https://www.bundesgesundheitsministerium.de/e-rezept [accessed 2025-08-20]
43. E-Rezept: Verbände ziehen überwiegend positive Bilanz [E-prescriptions: associations give a largely positive assessment].: Deutsches Ärzteblatt [German Medical Journal]; 2024. URL: https://www.aerzteblatt.de/nachrichten/152606/E-Rezept-Verbaende-ziehen-ueberwiegend-positive-Bilanz [accessed 2025-08-20]
44. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023 Sep 18;6(1):161. [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]

XSL•FO
RenderX

45. Kreyenbuhl J, Record EJ, Himelhoch S, et al. Development and feasibility testing of a smartphone intervention to improve adherence to antipsychotic medications. Clin Schizophr Relat Psychoses 2019;12(4):152-167. [doi: 10.3371/CSRP.KRRE.070816] [Medline: 27454213]
46. Nguyen AD, Frensham LJ, Baysari MT, Carland JE, Day RO. Patients' use of mobile health applications: what general practitioners think. Fam Pract 2019 Mar 20;36(2):214-218. [doi: 10.1093/fampra/cmy052] [Medline: 29873708]
47. Barmer Arztreport 2024 [Barmer physician report 2024]. Barmer. 2024. URL: https://www.barmer.de/resource/blob/1255894/d58c1bd7c1a99c78a54d410c490920c8/dl-arztreport-2024-data.pdf [accessed 2025-01-06]
48. Einheitlicher Bewertungsmaßstab (EBM). : National Association of Statutory Health Insurance Physicians; 2024 URL: https://www.kbv.de/documents/praxis/abrechnung/ebm/archiv/2024-4-ebm.pdf [accessed 2025-01-06]
49. Uncovska M, Freitag B, Meister S, Fehring L. Patient acceptance of prescribed and fully reimbursed mhealth apps in Germany: an UTAUT2-based online survey study. J Med Syst 2023 Jan 27;47(1):14. [doi: 10.1007/s10916-023-01910-x] [Medline: 36705853]
50. Machleid F, Jansky B, Wild V, Wiegmann C, Kaminski J, Schreiter S. Mobile Gesundheitstechnologien für eine gerechte Versorgung bei psychischen Erkrankungen [Mobile health technologies for equitable care in mental illness]. Nervenheilkunde [Ger J Neurol] 2024 Nov;43(12):688-700. [doi: 10.1055/a-2415-8433]
51. Peasgood T, Bourke M, Devlin N, Rowen D, Yang Y, Dalziel K. Randomised comparison of online interviews versus face-to-face interviews to value health states. Soc Sci Med 2023 Apr;323:115818. [doi: 10.1016/j.socscimed.2023.115818] [Medline: 36940582]
52. Guest G, Name E. Comparing Interview and Focus Group Data Collected in Person and Online [Internet]: Patient-Centered Outcomes Research Institute (PCORI); 2020. [doi: 10.25302/05.2020.ME.1403117064] [Medline: 36701499]

Abbreviations

BfArM: Bundesinstitut für Arzneimittel und Medizinprodukte (Federal Institute for Drugs and Medical Devices)
COREQ: Consolidated Criteria for Reporting Qualitative Studies
DiGA : Digitale Gesundheitsanwendung (digital health application)
GKV-SV: GK-Spitzenverband (National Association of Statutory Health Insurance Funds)
mHealth: mobile health
SHI: statutory health insurance

Edited by J Sarvestan; submitted 02.Mar.2025; peer-reviewed by M Guth, S Scheibe; revised version received 25.Nov.2025; accepted 25.Nov.2025; published 09.Feb.2026.

Please cite as:
Plescher F, Giebel GD, Abels C, Neusser S, Kampka B, Börchers K, Wasem J, Blase N
Problems and Barriers Regarding the Admission, Financing, and Service Provision of Digital Health Apps: Qualitative Stakeholder Survey
J Med Internet Res 2026;28:e73332
URL: https://www.jmir.org/2026/1/e73332
doi:10.2196/73332

© Felix Plescher, Godwin Denk Giebel, Carina Abels, Silke Neusser, Bonnie Kampka, Kirstin Börchers, Jürgen Wasem, Nikola Blase. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effects of Using a Smartphone App Combined With Behavior Change Techniques on the Level of Physical Activity Among Adults and Older Adults: Sequential Multiple Assignment Randomized Trial

Maria do Socorro Morais Pereira Simoes[1], PT, PhD; Neli Leite Proença[1], PT, PhD; Vinícius Tonon Lauria[1], BPE, PhD; Matheus Bibian do Nascimento[1], BPE; Ricardo da Costa Padovani[2], PsyD, PhD; Victor Zuniga Dourado[1,3], PT, BEng, PhD

[1]Departamento de Ciências do Movimento Humano, Instituto de Saúde e Sociedade, Universidade Federal de São Paulo, 136 Silva Jardim St, Room 338, Santos, Brazil

[2]Departamento de Saúde, Educação e Sociedade, Instituto de Saúde e Sociedade, Universidade Federal de São Paulo, Santos, Brazil

[3]Department of Global Health and Population, Bernard Lown Scholars in Cardiovascular Health Program, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, United States

Corresponding Author:
Victor Zuniga Dourado, PT, BEng, PhD
Departamento de Ciências do Movimento Humano, Instituto de Saúde e Sociedade, Universidade Federal de São Paulo, 136 Silva Jardim St, Room 338, Santos, Brazil

Abstract

Background: The use of tools, such as smartphone apps, to increase the level of physical activity (PA) decreases over time. Adaptive intervention trials have been recommended to test technology-based interventions owing to the possibility of adapting interventions based on individual responses.

Objective: This study aimed to investigate the effects of using a smartphone app combined with behavior change techniques on the PA level in adults and older adults (assessed using the step count). Moreover, the study investigated the time spent in sedentary behavior and time spent in moderate-to-vigorous PA (MVPA).

Methods: In this single-blinded, sequential multiple assignment randomized trial, participants were randomized into 3 groups during a 24-week intervention (group 1: app with tailored messages; group 2: app with tailored messages plus gamification I; and control group: educational information). In the sixth week, participants from groups 1 and 2 were classified as responders and nonresponders according to their average daily step count. Nonresponders were rerandomized among the other groups, adding a second type of gamification (group 3: app with tailored messages plus gamification II). After another 6 weeks, participants were reassessed and advised to keep monitoring their step count with the app, but without interference from the researchers. Face-to-face assessments were conducted. The behavior change techniques included app features (goal setting, auto-monitoring, ranking, and virtual badges) and researcher-provided resources (tailored messages and in-person sessions of PA). The intervention effects were analyzed using linear mixed models.

Results: The study included 53 participants (control group: n=17, group 1: n=17, group 2: n=19; mean age 44.0, SD 12.7 years). Groups 1 and 2 had 63% (10/16) and 47% (7/15) responders, respectively (P=.38). Regarding the PA level, participants from group 1 showed increases in the average daily step count at all assessments (final vs initial: B=797.2 steps/day, 95% CI 475.3-1119.1; P<.001; follow-up vs initial: B=2097.6 steps/day, 95% CI 1577.2-2618.1; P<.001). All participants showed a reduction in the time spent in sedentary behavior at the final assessment compared with the initial assessment (B=−70.8 min/week, 95% CI −88.8 to −52.9; P<.001), without differences among groups. The time spent in MVPA varied across time among all participants. Regardless of the initial group and allocation in the second randomization, responders from groups 1 and 2 showed a constant increase in the average daily step count (week 6 vs week 1: B=1548.0 steps/day, 95% CI 1407.4-1688.6; P<.001; week 12 vs week 1: B=1720.3 steps/day, 95% CI 1568.8-1871.7; P<.001; week 12 vs week 6: B=172.3, 95% CI 20.8-323.8; P=.03).

Conclusions: The adaptive intervention protocol using a smartphone app with behavior change techniques increased participants' PA levels. Stepping up behavior change techniques and progressively offering new stimuli may contribute to a change in behavior regarding PA.

Trial Registration: Brazilian Registry of Clinical Trials RBR-8xtc9c UTN: U1111-1218-1092; https://ensaiosclinicos.gov.br/rg/RBR-8xtc9c/

International Registered Report Identifier (IRRID): RR2-10.1186/s13063-019-3879-1

XSL•FO
RenderX

(J Med Internet Res 2026;28:e73388) doi:10.2196/73388

KEYWORDS

physical activity; adaptive clinical trial; transtheoretical model of behavior change; smartphone; app

Introduction

Despite the well-known benefits of physical activity (PA) in the prevention and treatment of several health conditions [1,2], a significant proportion of the global population does not meet its minimal recommended level [3]. The World Health Organization aims to reduce physical inactivity by 15% by 2030 [3]. Recent studies emphasize the complex interaction of individual [1,4,5], social [1,4,5], environmental [1,4,5], and political [4,5] factors influencing the practice of PA, suggesting that interventions should consider them [4,5]. Specifically, regarding individual and social factors, popular technologies, such as mobile devices and apps, have been used to increase the engagement of insufficiently active individuals.

Using smartphones as part of intervention programs can facilitate the auto-monitoring of PA, contributing to the behavior change process [6,7]. Duncan et al [6] compared auto-monitoring interventions for PA and eating habits using a website and printed material. They showed that both delivery methods improved the behavior, with no differences between groups. Evidence with moderate effects suggests that app-based interventions and pedometers may be effective [8]. Still, the association between app use and behavior change techniques may be more effective than app-based interventions alone [7].

Among the most commonly used behavior change techniques to promote PA are tailored messages, health education, gamification (the use of game elements in contexts other than games), and social support [7]. Regarding apps for PA and health, few are based on behavior change techniques. Among at least 25 behavior change techniques, only 1-8 techniques are offered by apps [9-12].

Despite the promising use of technologies, such as websites and smartphone apps, to deliver interventions targeted to increase the level of PA, studies show that the use of such devices decreases over time [6,7]. Interestingly, behavior change usually presents with the same pattern: individuals begin the program more engaged and active but do not maintain the new behavior for a long time [13]. That said, adaptive intervention trials, such as the sequential multiple assignment randomized trial (SMART), have been recommended to test technology-based interventions instead of classical clinical trials [14,15] due to the possibility of adapting interventions over time based on individuals' responses. It would be reasonable to extrapolate this rationale to interventions targeted to promote PA. Clinical trials with a SMART design have been shown to be feasible [16,17], but so far, this design has been barely used for PA interventions [17].

Our primary aim was to investigate the effects of using a smartphone app combined with behavior change techniques on the level of PA among adults and older adults, assessed by the average number of daily steps. Our secondary aim was to investigate the effects of the intervention on time spent in sedentary behavior and time spent in moderate-to-vigorous PA (MVPA), assessed by a triaxial accelerometer.

We hypothesized that the smartphone app combined with behavior change techniques offered in a SMART adaptive intervention design would be an effective tool to increase the level of PA among adults and older adults.

Methods

Study Design

This study had a single-blinded SMART design, in which the assessor was blinded regarding group allocations, with a 1:1 allocation ratio. The protocol was registered at the Brazilian Register of Clinical Trials (RBR-8xtc9c), and the study has been presented according to the CONSORT (Consolidated Standards of Reporting Trials) 2025 [18] and CONSORT-EHEALTH (V.1.6.1) [19] recommendations (Checklist 1). We enrolled participants continuously from November 2018 to February 2020. The study protocol was previously published in detail [20].

The primary outcome of the study was the average number of daily steps, and the secondary outcomes were the time spent in sedentary behavior and the time spent in MVPA, which were all assessed by a triaxial accelerometer.

Ethical Considerations

The project was approved by the Local Ethics Committee of the Federal University of São Paulo (CAAE 89112418.8.0000.5505), and all volunteers signed an informed consent form in person before participation. The informed consent form contained detailed information about the assessments and the group allocation, including a detailed description of the control and intervention groups. Participants did not receive any compensation before, during, or after the study. Only the principal researchers had access to participants' personal information. No identifiable information and/or images were or will be published.

Participants and Setting

We recruited participants through social media posts, distribution of printed material in different neighborhoods, and recommendations from other participants. The content of the recruitment material included the question "Do you want to increase your level of physical activity? We invite volunteers aged from 20 years, users of smartphones, that desire to move more. During 6 months, we will follow you to promote physical activity. Contact us!" or "Do you consider yourself a not very active person? Do you want to increase your level of physical activity? The EPIMOV Laboratory is looking for volunteers aged 20 years and older, able to walk without assistance from another person, and free of cardiac or pulmonary diseases. Contact us!" Those interested in participating were required to

XSL•FO
RenderX

call or text a number to schedule the assessment. Participants did not receive any type of incentive to cooperate in the study.

The inclusion criteria were as follows: (1) age 20 years or older; (2) absence of diagnosed cardiopulmonary, locomotor, or other conditions that could preclude the safe unsupervised performance of PA; and (3) having a smartphone and being familiar with its use, which was assessed by researcher observation and judgment during the enrollment phase.

The exclusion criteria were as follows: (1) basal level of PA of ≥10,000 average steps per day; (2) self-reported recent respiratory infection; (3) abnormalities on the cardiorespiratory fitness test that preclude the safe performance of unsupervised PA; and (4) refusal to participate by signing the informed consent. Although Tudor-Lock et al [21] suggested that interventions to increase the level of PA should mainly target individuals who perform less than 5000 steps per day, we decided not to restrict the sample to sedentary individuals and set the limit of 10,000 steps per day, which indicates a considerable amount of PA. This criterion was assessed using a triaxial accelerometer, worn by participants for 7 consecutive days. The procedure is described in detail in the Description of the Assessment Procedure section.

The assessments and in-person meetings for decision-making regarding the protocol took place at the Epidemiology and Human Movement (EPIMOV) Laboratory at the Federal University of São Paulo in Santos, Brazil. The target population was the residents in the metropolitan area of Santos, São Paulo, Brazil. The city of Santos is mainly flat and offers a gardened beachfront with an extension of approximately 7 km [22]. Cycle lanes and public equipment widely cover the city to practice PA. It is common to watch people exercising in different spaces of the city. Moreover, it is ranked 5th in Brazil in terms of quality of life [22].

Randomization Procedures

The randomization sequence (initial randomization and rerandomization) was generated in blocks of 6 participants, using a website [23], and was sealed in opaque envelopes numbered in the opening sequence. The envelopes were prepared by a third person who was not involved in any other study phase. The first envelope was opened at the time of enrollment after participants watched the animated video about the benefits of practicing PA. The second envelope, containing the new group allocation, was kept sealed until the sixth week of the intervention. We opened the second envelope only among nonresponder participants in their presence. After the onset of the COVID-19 pandemic and the suspension of in-person activities at the university, a research team member opened the envelopes and informed the nonresponder participants about the new group intervention, using the same text messaging app.

Description of Interventions

The adaptive intervention protocol was designed to present different stimuli in a stepwise manner, which could contribute to optimizing the cost-effectiveness of interventions delivered in health services. Based on the ecological model proposed by Bauman et al [4] and Sallis et al [5], the intervention was initially focused on individual (demographic and psychological) and social (social support) factors. By the end of the SMART protocol, it would be possible to analyze if there is an ideal sequence of stimuli to deliver, and thus, strategies in public health could be planned in a more cost-effective sequence, starting with those less demanding in financial and human resources and progressing to those more demanding in financial and human resources.

The intervention lasted a total of 24 weeks and was delivered in 2 phases. Initially, we randomized participants into 3 groups (group 1: app plus tailored messages; group 2: app plus tailored messages plus gamification; and control group: advice) (Figure 1). The first intervention phase lasted 12 weeks, during which the average number of participants' daily steps was registered weekly in groups 1 and 2. By the sixth week of the intervention, we assessed participants' performance in terms of daily steps and classified them into responders and nonresponders. We then rerandomized nonresponders, adding a third group (group 3: app plus tailored messages plus gamification II) (Figure 1). After completing 12 weeks, we advised participants to continue using the app and monitor daily steps for over 12 weeks. From that moment on, they did not receive any direct intervention from the researchers. We invited participants from all groups for reassessments at weeks 12 and 24.

XSL•FO
RenderX

Figure 1. Flowchart of the study protocol. R: randomization.

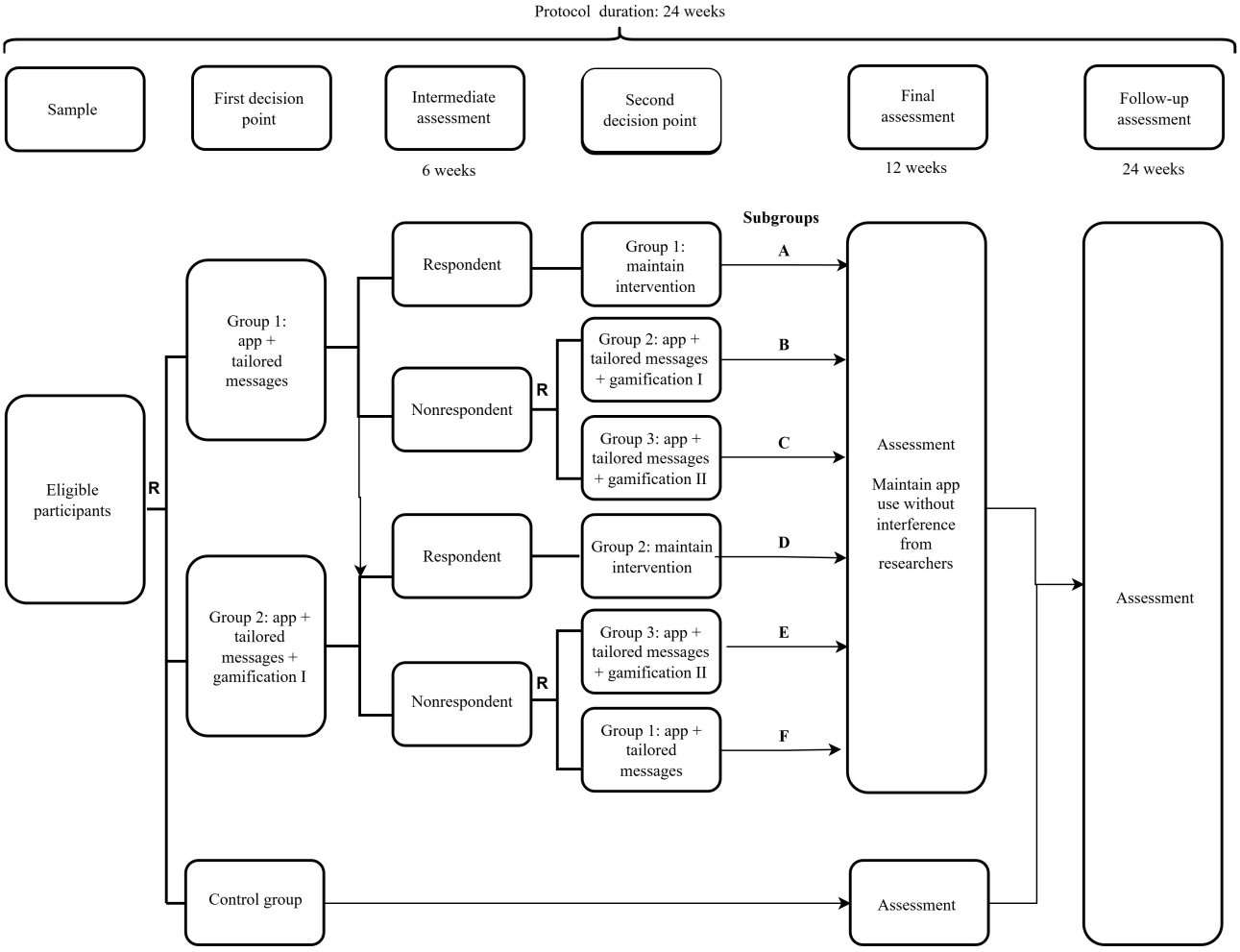

Initially, all participants watched an animated video about the benefits of practicing PA and received a printed booklet with the same content as the video. Moreover, we sent them a link so that they could watch the video again as often as they wanted. After this procedure, we opened the first randomization envelope containing the initial group assignment and explained the detailed intervention according to the group allocation.

A detailed description of the intervention has been published elsewhere [20]. Briefly, the approaches used in each group were as follows:

- Group 1: Participants used the app to self-monitor their daily steps. We sent tailored messages weekly using a text messaging app, according to their performance in the last week, and motivated them to increase their step count during the following week.
- Group 2: In addition to the self-monitoring and tailored messages, we instructed participants on how to access the gamification features the app offers, such as ranking, virtual challenges, virtual badges, and social media interaction.
- Control group: We advised participants to increase their PA levels, based on the information provided by the animated video and printed booklet. We did not inform them about using apps. Moreover, since these participants would not be rerandomized, we did not invite them to the intermediate assessment at week 6.
- Group 3: We included the second gamification phase at the intermediate assessment (week 6). Gamification II offered the same features as gamification I plus opportunities for socialization among participants and the research team. We planned in-person meetings every 2 weeks to practice PA in groups and created a group in a text messaging app to facilitate interaction.

Before beginning the study, researchers from the EPIMOV team used smartphone apps to track PA. After a period of use and comparing the features, easiness, and difficulties of the apps, they decided that Pacer (Pacer Health Inc) was the most suitable to use in the trial. It registers the step count and offers behavior change techniques (self-monitoring, goal setting, progress bars, competition, ranking, virtual badges, and social interaction). In addition, we tested the feasibility of the SMART design using Pacer [17].

We used the free version of the Pacer app in this study, providing the behavior change techniques according to group allocation. The participants downloaded the app on their phones with the aid of researchers. In addition to in-person instructions on using the app, we provided a printed booklet with the same instructions. In addition, at any time, participants could contact researchers to ask for help regarding the use of the app. There were no changes to the protocol after the trial commenced, including no changes in the version of the app.

XSL•FO
RenderX

The content of the weekly tailored text messages was developed by one of the researchers supervised by a sports psychologist, according to the stage of behavior change, and is available elsewhere [17]. We created a sequence of messages for precontemplation and contemplation, and another sequence for participants in preparation, action, and maintenance. For each sequence, we had messages for occasions when participants achieved or overcame their goals and messages for occasions when they did not reach their goals. Other than the tailored messages, participants did not receive reminders or prompts to use the app.

The opportunities for socialization offered to participants reallocated to group 3 occurred biweekly as in-person meetings. During these meetings, we practiced PA in groups (participants and researchers), which was planned and conducted by an experienced professional in physical education. All meetings took place on Saturdays at the beachfront, and at the end of the PA sessions, researchers remained available to interact with participants. The sessions lasted, on average, 1 hour and 30 minutes (exercise and postexercise interaction) and included a warm-up period (around 5 minutes) and planned walking or running exercises targeted (each session) at step count, activity duration, or pace. In addition, participants had the option to join a group in the text messaging app so that they could interact with each other and plan the PA meetings.

Goal Setting and Intermediate Assessment

We set individualized goals for each participant. We instructed them to maintain their routines during the first week of app use. Then, we registered the average number of daily steps collected by the app, and this step count was considered as the participants' goal. The goal-setting process adopted in this study has been published in another study [20]. Finally, we sent a text message informing participants that their goal was to overcome the initial step count from that day on.

We used the average number of daily steps collected by the app to determine responders and nonresponders at the intermediate assessment. We registered the weekly performance of participants and plotted it in individualized sheets. Then, we used linear regression lines along the performance data during the first 6 weeks of the intervention. Finally, responders were those participants with a positive slope on the regression line, while nonresponders were those with a zero or negative slope on the regression line. We set this assessment to consider the individualized performance of participants over time, attempting to set feasible goals for our sample. If we faced errors when extracting the step count from the app interface, such as a lack of performance updates or operational errors, participants remained in the same group for initial allocation.

After the intermediate assessment, we invited participants for an in-person meeting at the laboratory. In this meeting, we explained their performance during the first 6 weeks of the intervention, and nonresponders were rerandomized (we opened the second sealed envelope that remained stored since the first randomization). Moreover, we provided information about the different components of the intervention (Figure 1). After the onset of the COVID-19 pandemic and the restrictive measures, we conducted the intermediate assessment remotely (one of the researchers opened the second envelope and communicated with participants via text messages).

Study Measurements

We assessed participants at the beginning of the protocol (initial assessment) and reassessed them at weeks 12 (final assessment) and 24 (follow-up assessment). All of the assessors were blinded regarding the group allocation. The assessment protocol included demographic and socioeconomic information, general health condition, stages of behavior change for PA, level of PA, and anthropometric measures.

The assessors completed the protocol on 2 different days, 7 days apart. On the first day, participants collected their demographic and socioeconomic information, stages of behavior change for PA, and general health status. At the end of day 1, they received an accelerometer to wear during the subsequent 7 days when they returned to complete the assessments. On day 2, participants returned the accelerometer, and we assessed their level of PA and anthropometric measures.

By week 12 (the end of the intervention protocol), we reassessed participants and advised them to keep using the app to track their daily steps, but from this moment on, they did not receive tailored messages or any other direct intervention from the researchers. Finally, at week 24, we invited participants to return for the follow-up assessment, irrespective of whether they had presented for the 12-week evaluation.

At each assessment point, we presented the results of the previous evaluation to participants. At week 12, they received the results from the initial assessment, and at week 24, they received the results from the final evaluation. After the study ended at week 24, the results were provided according to their wish and availability. There were no consequences for participants who did not withdraw their results. The results were stored at a safe location and were available to participants at any time.

Description of the Assessment Procedure

The tests and interviews were standardized and conducted by trained personnel. Moreover, the equipment was periodically checked and calibrated according to the manufacturer's instructions. We invited all participants to the follow-up assessment regardless of whether they had presented for the final evaluation.

We collected the following demographic and socioeconomic information: age (years), sex, marital status, education, and occupational status. We used the *Critério Brasil* questionnaire to assess socioeconomic status [24]. This questionnaire contains 15 questions, divided into 3 categories: household items, educational level of the family's chief, and access to services. According to the total score, participants were classified into 6 socioeconomic strata, from A to D, where A represents the highest socioeconomic status and D represents the lowest socioeconomic status.

For assessing general health conditions, we registered participants' self-reported personal and family health precedents and medication use. Moreover, we registered the following self-reported cardiovascular risk factors: age 45 years or older

XSL•FO
RenderX

for men and 55 years or older for women, family history of cardiovascular diseases, diagnosis of arterial blood hypertension, diagnosis of hyperglycemia or diabetes, diagnosis of hypercholesterolemia or dyslipidemia, and smoking status. We complemented the assessment of cardiovascular risk factors with the evaluation of obesity and physical inactivity.

For evaluating the stages of behavior change for PA, we used the Brazilian version of the transtheoretical model [25], which considers that individuals may, according to the circumstances, transit nonlinearly among different behavior change stages [25]. According to the status of the current practice of PA, the willingness to change the behavior, and the fact that PA is essential to health, participants were classified into one of the following stages: precontemplation, contemplation, preparation, and maintenance [25].

For assessing the level of PA, participants received a triaxial accelerometer (Actigraph) at the end of the first day of assessments, which was required to be used for 7 consecutive days and worn over the hip according to the self-reported dominant side of the body. Data were considered valid when available for at least 4 days, including 1 weekend day, and for at least 10 hours per day [26]. When a participant presented with invalid data for the first time, we returned the accelerometer and asked them to use it for more than 7 days. We extracted accelerometer data using the manufacturer's software and registered the following information: average number of steps per day, time spent in sedentary behavior (0 to 99 step counts/min) [27], and time spent in MVPA (≥1952 step counts/min) [27].

With regard to anthropometric measures, we registered the height (m) and body mass (kg) of participants using a digital scale with a stadiometer (Toledo). We positioned participants standing barefoot over the equipment and asked them to keep looking forward during the measurement. For height measurement, participants kept their arms crossed over the trunk while sustaining maximal inspiration, and for body mass measurement, they kept their arms along the body.

Sample Size Calculation

Considering a fixed *k* probability, at least *m* participants would be allocated to subgroups B and E, as demonstrated in Figure 1. Moreover, based on a nonresponse rate *q* of 35% - 65% in SMART studies, we estimated that 3 participants per group would be sufficient to ensure familiarity with the protocol, allow identification, and deal with potential difficulties that could occur during the intervention period. Thus, for a probability *k* of 90% and a nonresponse rate *q* of 50% by the end of the sixth week of the protocol, the study should include at least 42 participants [28]. Due to the lack of a similar study to use data for estimating the sample size, after 42 participants completed the protocol, we planned to perform a new sample size calculation and continue the study until the sample size target was achieved. However, this approach was not possible because the protocol had to be interrupted after the onset of the COVID-19 pandemic. Therefore, we completed the study after including 53 participants.

Data Analysis

Participants' characteristics are presented using descriptive statistics, such as mean and SD for continuous variables, and absolute number and proportion for categorical variables. We compared the characteristics of participants among groups using 1-way ANOVA for continuous variables and the chi-square test for categorical variables.

We checked for participant performance in groups 1 and 2, using graphs with linear regression lines. For each group and participant, we plotted 3 graphs using the average number of daily steps: 1 with data from the 12 weeks of the intervention, 1 with data from the first 6 weeks (between the first and intermediate assessments), and 1 with data from the last 6 weeks (between the intermediate and final assessments).

To analyze the effects of the intervention on primary and secondary outcomes, we used repeated measures linear mixed models, considering the initial group allocation and time as fixed effects and the individual as a random effect. Moreover, we grouped participants from groups 1 and 2 according to their response to the intervention (responders and nonresponders) and compared their performance using repeated measures linear mixed models, considering response to the intervention and time as fixed effects and the individual as a random effect.

We managed missing data using intention-to-treat analysis by multivariate data imputation, adjusted by sex and age. The graphs for weekly performance were generated using Microsoft Excel (Microsoft Corporation), and data analysis was conducted using Stata version 14 (StataCorp LLC).

Results

Of 62 potential individuals who volunteered to participate, 53 met the inclusion criteria and were included in the protocol. Some participants were lost to follow-up, especially at the 24th-week assessment, mainly due to the COVID-19 pandemic. For the same reason, we had to interrupt the protocol and could not achieve a larger sample size. As shown in Figure 2, according to the intention-to-treat analysis, we included all participants in the data analysis.

XSL•FO
RenderX

Figure 2. Flowchart of participants in the study.

XSL•FO
RenderX

As we suddenly had to interrupt the inclusion of new participants, there were differences in the number of participants between group 2 and the other groups. Furthermore, our plan to conduct a second trial phase, using intermediate data to estimate sample size, could not be executed.

Protocol deviations were continuously checked, such as accessing app features that were not advised. Only 1 participant from group 1 explored the app and joined its gamification features, such as ranking and challenges of walking distance. There were no harmful events or unintended effects during the study. According to the taxonomy proposed by Michie et al [29], the behavior change techniques in our study included rewards (virtual badges provided by the app), incentives (tailored messages), graded tasks (overcoming the goal), feedback and monitoring (self-monitoring using the app and tailored messages), goals and planning (individual goal setting), social support (practical and general; provided by the app via teaming up with other participants, and for group 3, provided via in-person meetings and text messaging), comparison of behavior (social comparison provided by the ranking in the app), and instructions on how to perform a behavior (provided by the educational material).

In general, participants were middle-aged adults (mean age 44.0, SD 12.7 years), were mainly female (30/53, 57%), and were overweight or obese (mean BMI 29.8, SD 6.5 kg/m^2). Regarding the stage of behavior change for PA, almost half of the sample (24/53, 45%) was in the preparation stage. As shown in Table 1, there were no differences in participant characteristics, except for sex (P=.03) and the stage of behavior change (P=.047).

XSL•FO
RenderX

Table . Characteristics of participants at baseline.

Variable	All participants (n=53)	Control group (n=17)	Group 1 (n=17)	Group 2 (n=19)	*P* value[a]
Age (years), mean (SD)	44.0 (12.7)	45.2 (13.4)	47.7 (11.2)	39.6 (12.5)	.15
Sex (female), n (%)	30 (57)	9 (53)	6 (35)[b]	15 (79)	.03
Body mass (kg), mean (SD)	83.3 (21.9)	89.1 (28.8)	85.1 (14.7)	76.5 (19.3)	.21
Stature (m), mean (SD)	1.66 (0.08)	1.67 (0.08)	1.68 (0.09)	1.65 (0.07)	.49
BMI (kg/m^2), mean (SD)	29.8 (6.5)	31.5 (8.5)	30.2 (4.7)	28.0 (5.8)	.26
Race, n (%)					.15
White	33 (62)	12 (71)	7 (41)	14 (74)	
Black	6 (11)	0 (0)	4 (24)	2 (11)	
Brown	12 (23)	4 (24)	6 (35)	2 (11)	
Do not know/not reported	2 (4)	1 (6)	0 (0)	1 (5)	
Education, n (%)					.32
Middle school, incomplete	2 (4)	0 (0)	1 (6)	1 (5)	
Middle school, complete	2 (4)	1 (6)	1 (6)	0 (0)	
High school, complete	28 (53)	13 (77)	7 (41)	8 (42)	
Higher education, complete	13 (25)	3 (18)	4 (24)	6 (32)	
Postgraduation, complete	8 (15)	0 (0)	4 (24)	4 (21)	
Employed (yes), n (%)	39 (74)	11 (65)	14 (82)	14 (74)	.62
Socioeconomic classification[c], n (%)					.20
A	7 (13)	2 (12)	0 (0)	5 (26)	
B1	7 (13)	1 (6)	4 (24)	2 (11)	
B2	18 (34)	5 (29)	6 (35)	7 (37)	
C1	9 (17)	6 (35)	2 (12)	1 (5)	
C2	10 (19)	3 (18)	4 (24)	3 (16)	
D	2 (4)	0 (0)	1 (6)	1 (5)	
Stage of behavior change, n (%)					.047
Contemplation	1 (2)	0 (0)	0 (0)	1 (5)	
Preparation	24 (45)	6 (35)	9 (53)	9 (47)[d]	
Action	13 (25)	4 (24)	1 (6)[b]	8 (42)[d]	
Maintenance	15 (28)	7 (41)	7 (41)[b]	1 (5)[d]	

[a]*P* values refer to comparisons among groups.

[b]Significant difference between group 1 and group 2.

[c]For socioeconomic classification, A represents the highest socioeconomic status and D represents the lowest socioeconomic status.

[d]Significant difference between the control group and group 2.

All groups were similar when considering the outcome measures. The nonresponse rate was 53% (28/53), without differences in the proportion of respondents at week 6 (*P*=.38) (Table 2).

XSL•FO
RenderX

Table . Baseline assessment and comparison among groups.

Variable	All participants (N=53)	Control group (n=17)	Group 1 (n=17)	Group 2 (n=19)	*P* value[a]
Average daily step count, mean (SD)	6134 (1932)	6616 (2069)	5651 (1810)	6170 (1942)	.44
Average time spent in moderate-to-vigorous physical activity (min/day), mean (SD)	33 (16)	39 (14)	27 (18)	33 (13)	.17
Average time spent in sedentary behavior (min/day), mean (SD)	671 (130)	669 (77)	678 (154)	664 (150)	.96
Respondent at week 6 (yes), n (%)	—[b]	—	10[c] (63)	7[d] (47)	.38
Group allocation after rerandomization, n (%)					.003
Group 1	—	0 (0)	14 (82)	5 (26)	
Group 2	—	0 (0)	3 (18)	10 (53)	
Group 3	—	0 (0)	0 (0)	4 (21)	

[a]*P* values refer to comparisons among groups.

[b]Not applicable.

[c]Out of a total of 16.

[d]Out of a total of 15.

Table 3 presents the outcome measures in each group at each assessment (initial, final, and follow-up). We managed missing data using multivariate multiple imputation adjusted for sex and age.

Table . Effects of the intervention within and between groups.

Variable	Control group[a]	Group 1[a]	Group 2[a]
Average daily step count, mean (SD)			
Initial (baseline)	6606 (1989)[b]	5643 (1740)[b,c]	6180 (1887)[c]
Final (week 12)	7063 (2215)	7617 (1873)[d]	7524 (2256)[d]
Follow-up (week 24)	6271 (1574)	8480 (2387)	6350 (2413)
Average time spent in moderate-to-vigorous physical activity (min/day), mean (SD)			
Initial (baseline)	39 (14)[c]	27 (18)[b,e]	33 (13)
Final (week 12)	36 (20)[d]	35 (21)[d]	36 (14)
Follow-up (week 24)	41 (19)	44 (15)	33 (17)
Average time spent in sedentary behavior (min/day), mean (SD)			
Initial (baseline)	670 (74)[c]	678 (148)[c]	663 (144)[c]
Final (week 12)	569 (129)[d]	644 (137)[d]	613 (102)[d]
Follow-up (week 24)	694 (137)	677 (117)	663 (132)

[a]Missing data were treated using multivariate multiple imputation adjusted by sex and age.

[b]Significant difference within groups (initial vs follow-up assessment).

[c]Significant difference within groups (initial vs final assessment).

[d]Significant difference within groups (final vs follow-up assessment).

[e]Significant difference between groups (control group vs group 1).

For the primary outcome (the average count of daily steps), we observed differences in the factor *time* and at the individual level (Table 4). While participants from the control group showed a reduction in the step count over time (follow-up vs initial assessment: β=−575.8; *P*<.05) and participants from group 2 showed an increase in the step count between the initial and final assessments (β=897.4; *P*<.001), participants from group 1 showed an increase in the step count for all assessments (final vs initial assessment: β=650.2; *P*<.001; follow-up vs initial assessment: β=1521.9; *P*<.001).

XSL•FO
RenderX

Table . Average daily step count considering group, time, and group and time.

Variable[a]	Coefficient, value (95% CI)	*P* value
Group (Ref[b]: control group)		
Group 1	−895.2 (−2179.3 to 388.8)	.17
Group 2	−903.2 (−2158.7 to 352.2)	.16
Time (Ref: I[c])		
Fi[d]	−147.0 (−380.2 to 86.1)	.22
Fo[e]	−575.8 (−1045.3 to −106.2)	.02
Group 1 and time (Ref: control group and I)		
Fi	797.2 (475.3 to 1119.1)	<.001
Fo	2097.6 (1577.2 to 2618.1)	<.001
Group 2 and time (Ref: control group and I)		
Fi	1044.4 (747.0 to 1341.8)	<.001
Fo	982.8 (332.9 to 1632.6)	<.001
Contrasts, control group		
Fi and I	−147.0 (−380.2 to 86.1)	.22
Fo and I	−575.8 (−1045.3 to −106.2)	.02
Fo and Fi	−428.7 (−920.3 to 63.1)	.09
Contrasts, group 1		
Fi and I	650.2 (428.2 to 872.1)	<.001
Fo and I	1521.9 (1297.4 to 1746.4)	<.001
Fo and Fi	871.7 (640.4 to 1103.0)	<.001
Contrasts, group 2		
Fi and I	897.4 (712.8 to 1082.0)	<.001
Fo and I	407.0 (−42.2 to 856.2)	.08
Fo and Fi	−490.3 (−949.3 to −31.4)	.04

[a]Contrasts within groups over time (linear mixed model analysis). Fixed effects: initial group allocation and time of assessment; random effect: participant.

[b]Ref: reference.

[c]I: initial assessment.

[d]Fi: final assessment.

[e]Fo: follow-up assessment.

Results from the analysis of the time spent in sedentary behavior are presented in Table 5. We observed differences in the factor *time* in all groups (final vs initial assessment: β=−70.8; *P*<.001) and at the individual level (*P*<.001), without differences among groups.

XSL•FO
RenderX

Table . Time spent in sedentary behavior considering group, time, and group and time.

Variable[a]	Coefficient, value (95% CI)	*P* value
Group (Ref[b]: control group)		
Group 1	−6.9 (−81.9 to 68.1)	.86
Group 2	−20.3 (−93.8 to 53.3)	.59
Time (Ref: I[c])		
Fi[d]	−70.8 (−88.8 to −52.9)	<.001
Fo[e]	19.7 (−16.4 to 55.9)	.29
Group 1 and time (Ref: control group and I)		
Fi	26.0 (1.2 to 50.7)	.04
Fo	−23.8 (−63.9 to 16.2)	.24
Group 2 and time (Ref: control group and I)		
Fi	41.9 (19.0 to 64.8)	<.001
Fo	−10.3 (−60.3 to 39.7)	.69
Contrasts, control group and time		
Fi and I	−70.8 (−88.8 to −52.9)	<.001
Fo and I	19.7 (−16.4 to 55.9)	.29
Fo and Fi	90.5 (−88.8 to −52.9)	<.001
Contrasts, group 1 and time		
Fi and I	−44.9 (−61.9 to −27.8)	<.001
Fo and I	−4.1 (−21.4 to 13.2)	.64
Fo and Fi		
Contrasts, group 2 and time		
Fi and I	−29.0 (−43.2 to −14.7)	<.001
Fo and I	9.4 (−25.2 to 43.9)	.59
Fo and Fi	38.3 (3.04 to 73.6)	.03

[a]Contrasts within groups over time (linear mixed model analysis). Fixed effects: initial group allocation and time of assessment; random effect: participant.

[b]Ref: reference.

[c]I: initial assessment.

[d]Fi: final assessment.

[e]Fo: follow-up assessment.

We also observed differences at the individual level for the time spent in MVPA (Table 6). For the factors *group* and *time*, participants from the control group showed a reduction in the time spent in MVPA from the initial assessment to the final assessment (β=−7.9; *P*<.001) and then showed an increase in the time by the follow-up assessment (β=9.0; *P*<.001). On the other hand, participants from group 1 showed an increase in the time spent in MVPA at the follow-up assessment (β=7.8; *P*<.001).

XSL•FO
RenderX

Table . Time spent in moderate-to-vigorous physical activity considering group, time, and group and time.

Variable[a]	Coefficient, value (95% CI)	*P* value
Group (Ref[b]: control group)		
Group 1	−13.0 (−22.4 to −3.5)	.007
Group 2	−5.5 (−14.8 to 3.8)	.24
Time (Ref: I[c])		
Fi[d]	−7.9 (−9.9 to −5.9)	<.001
Fo[e]	1.1 (−2.9 to 5.2)	.59
Group 1 and time (Ref: control group and I)		
Fi	9.2 (6.4 to 11.9)	<.001
Fo	7.9 (3.4 to 12.4)	.001
Group 2 and time (Ref: control group and I)		
Fi	8.8 (6.2 to 11.4)	<.001
Fo	−2.6 (−8.2 to 3.0)	.37
Contrasts, control group and time		
Fi and I	−7.9 (−9.9 to −5.9)	<.001
Fo and I	1.1 (−2.9 to 5.2)	.59
Fo and Fi	9.0 (4.8 to 13.3)	<.001
Contrasts, group 1 and time		
Fi and I	1.2 (−0.7 to 3.2)	.20
Fo and I	9.1 (7.1 to 11.0)	<.001
Fo and Fi	7.8 (5.8 to 9.8)	<.001
Contrasts, group 2 and time		
Fi and I	0.9 (−0.7 to 2.5)	.27
Fo and I	−1.5 (−5.4 to 2.4)	.46
Fo and Fi	−2.4 (−6.3 to 1.6)	.25

[a]Contrasts within groups over time (linear mixed model analysis). Fixed effects: initial group allocation and time of assessment; random effect: participant.

[b]Ref: reference.

[c]I: initial assessment.

[d]Fi: final assessment.

[e]Fo: follow-up assessment.

Figures 3 and 4 present the performance of participants from groups 1 and 2 during the 12 weeks of the protocol and from weeks 1 to 6 and weeks 7 to 12 (before and after rerandomization, respectively).

XSL•FO
RenderX

Figure 3. Weekly performance of participants from group 1. (A) Performance from week 1 to week 12; (B) performance from week 1 to week 6 (before rerandomization); (C) performance from week 7 to week 12 (after rerandomization).

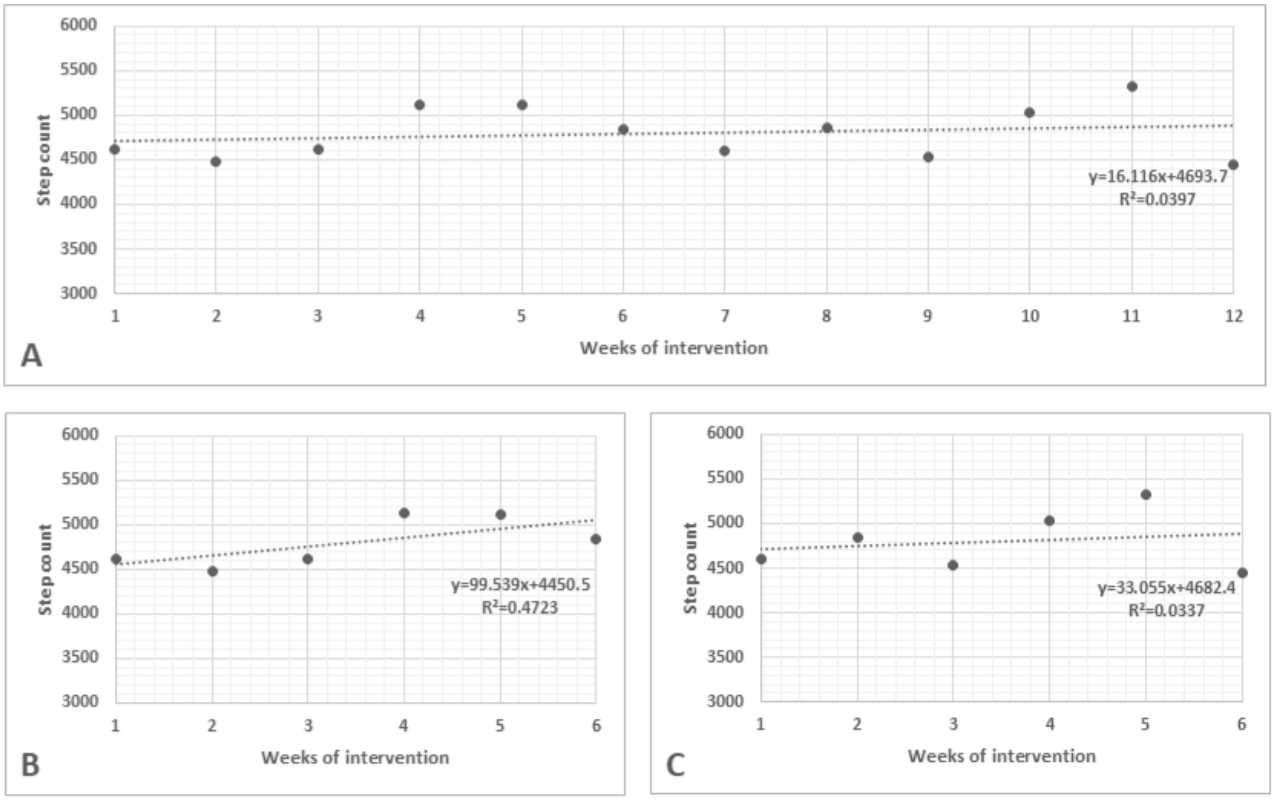

Figure 4. Weekly performance of participants from group 2. (A) Performance from week 1 to week 12; (B) performance from week 1 to week 6 (before rerandomization); (C) performance from week 7 to week 12 (after rerandomization).

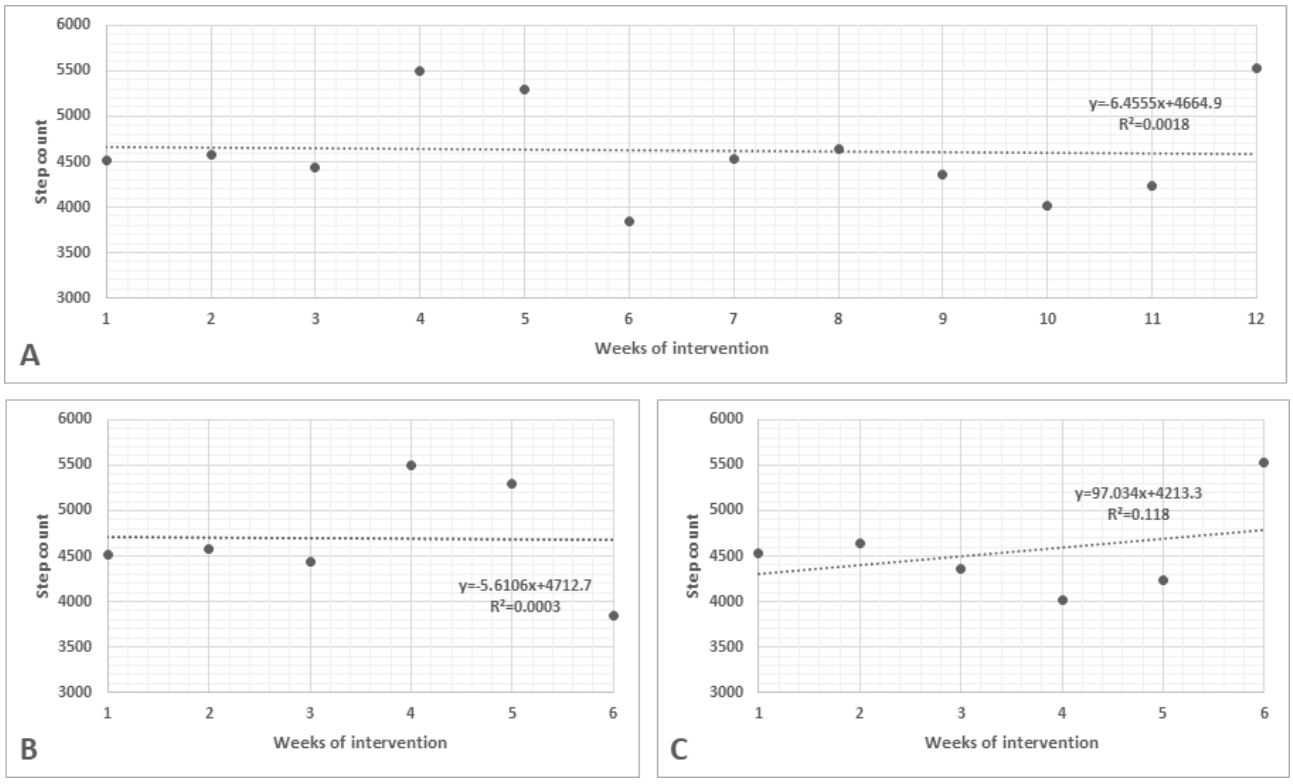

In Table 7, we present the analysis of the average daily step count of responders and nonresponders, regardless of group allocation. Using the Pacer app, we collected these data in weeks 1, 6, and 12. We observed differences at the individual level, and although all participants showed reductions in the average daily step count over time (week 6 vs week 1: β=−496.7; *P*<.001; week 12 vs week 1: β=−579.3; *P*<.001), among

XSL•FO
RenderX

responders, the step count constantly increased (week 6 vs week 1: β=2044.7; *P*<.001; week 12 vs week 1: β=2299.6; *P*<.001).

Table . Pacer app's average daily step count from responders and nonresponders, regardless of group allocation, according to the response to the intervention and time.

Variable[a]	Coefficient, value (95% CI)	*P* value
Response to the intervention (Ref[b]: no)		
Yes	−1588.8 (−3252.4 to 74.9)	.06
Time (Ref: week 1)		
Week 6	−496.7 (−651.6 to −341.8)	<.001
Week 12	−579.3 (−743.6 to −414.9)	<.001
Response and time (Ref: no and week 1)		
Yes and week 6	2044.7 (1835.5 to 2253.9)	<.001
Yes and week 12	2299.6 (2076.1 to 2523.0)	<.001
Contrasts, nonresponders and time		
Week 6 and week 1	−496.7 (−651.6 to −341.8)	<.001
Week 12 and week 1	−579.3 (−743.6 to −414.9)	<.001
Week 12 and week 6	−82.6 (−246.9 to 81.7)	.33
Contrasts, responders and time		
Week 6 and week 1	1548.0 (1407.4 to 1688.6)	<.001
Week 12 and week 1	1720.3 (1568.8 to 1871.7)	<.001
Week 12 and week 6	172.3 (20.8 to 323.8)	.03

[a]Contrasts between response and time (linear mixed model analysis). Fixed effects: response to the intervention and time of assessment; random effect: participant.

[b]Ref: reference.

Discussion

This study aimed to investigate the effects of using a smartphone app combined with behavior change techniques on the level of PA among adults and older adults. Due to the sudden interruption of the protocol after the onset of the COVID-19 pandemic, we completed the study with a small sample size. Overall, all participants presented with improvements in outcomes over time. Specifically for PA, all participants from group 1 showed a consistent increase in the average daily step count. According to the benchmarks for assessing the effectiveness of interventions to promote PA proposed by Wright et al [30], considering the digital intervention setting, participants from the control group presented a performance below the 25th percentile for step count, while participants from groups 1 and 2 performed greater than the 75th percentile and between the 50th and 75th percentiles, respectively.

Only participants from group 1 continued to increase the step count even after the end of the intervention. Despite the nonstatistical difference in the proportion of respondents and nonrespondents between groups, which may have occurred due to the small sample size and consequently low statistical power, it is possible that the stepwise offering of intervention components (initiation with less complex strategies and gradual increases according to participants' responses) may be more effective in supporting behavior change than offering more complicated strategies. Furthermore, gamification may be more effective for specific individual profiles and may not be as attractive to others. We support this hypothesis because participants from group 1 had positive regression lines over the 12 weeks of the intervention.

Interestingly, the time spent in sedentary behavior and the time spent in MVPA did not present the same trajectory as step count. The increase in the average daily step count was not accompanied by a reduction in the time spent in sedentary behavior or an increase in the time spent in MVPA, which can be explained by the outcome we measured. Participants may have altered their time for physical activities that we did not assess. Although the step count and the time spent in sedentary behavior are related [21], a systematic review and meta-analysis also found increases in step count not accompanied by reductions in sedentary behavior [31].

Our results are similar to those of Vandelanotte et al [32], who did not observe differences in the level of PA among 3 groups (1 control and 2 intervention groups). Despite differences in the study design and intervention protocol between the studies, both similarly associated the use of technology with behavior change techniques. The clinical trial conducted by Plotnikoff et al [33] found benefits in PA and general health among participants with type 2 diabetes, presenting a considerably higher effect size (Cohen *d*=0.67 for step count). Their protocol associated using a smartphone app with in-person meetings for group therapy and outdoor physical exercises [33]. It may be possible that, beyond the regular in-person meetings, a population

XSL•FO
RenderX

diagnosed with chronic conditions, such as diabetes, may be more motivated to become more active.

Likewise, other clinical trials aiming to increase the level of PA found better results in populations with clinical conditions than studies involving a healthy population [31]. In a systematic review and meta-analysis, Ringeval et al [31] showed that almost 60% of clinical trials that used an electronic device to monitor the level of PA included individuals at risk or already presenting with a clinical condition, and only 1 study included healthy individuals as the target population.

The nonresponse rate observed in our study (28/53, 53%) was lower than the expected rate (65%). This information may contribute to estimating the nonresponse rate in the planning of future clinical trials aiming to increase the level of PA in adaptive intervention designs, as well as the planning of adaptive intervention protocols with more decision points for rerandomization and changing the stimuli to favor behavior change in the practice of PA.

Respondents in week 6, independently of the initial or intermediate randomized allocation, constantly increased their average daily step count, while nonrespondents showed negative performance. Since our protocol did not include more decision points, it is impossible to know if these participants would benefit from more rerandomizations. These findings led us to reflect on the response at the intermediate assessment, which could be a prognostic indicator of the practice of PA. In this sense, Shang et al [34] pointed out the importance of deepening the investigation of fluctuating behavior in PA.

It is important for future studies to investigate if nonrespondents at the first phase of the intervention would benefit from new stimuli. In addition, a study with a larger sample size could investigate if nonrespondents present characteristics that could help health professionals and services to identify the population at higher risk for physical inactivity and thus design more personal approaches [34]. When planning interventions to increase PA, planning beyond the activity itself is necessary. As mentioned earlier, psychosocial, economic, and environmental factors are involved in this challenge [4]. Furthermore, there is a need to consider the dynamic characteristics of behavior change, which may include relapse periods and difficulties in maintaining the new behavior [34,35].

The loss to follow-up in our study, except for the control group, was consistent with the expectation for an interventional study to increase PA, which varies from 20% to 35% [32,33]. We faced difficulties following participants from the control group even before the onset of the COVID-19 pandemic. Our hypotheses for these losses are as follows: (1) it is possible that researchers and participants from groups 1 and 2 developed a bond due to the weekly text message interaction, which may have contributed to keeping the appointments for reassessments, and (2) it is possible that some of the participants may have volunteered for our study with the expectation of using an app for PA, which was not met for participants from the control group. For future studies, as a strategy to increase adherence among control participants, it may be helpful to offer an app without associating other behavior change techniques or even contact them periodically to develop a bond.

Many individuals have difficulty establishing effective strategies to change their behavior regarding PA. As pointed out by Warburton and Bredin [36], there is enough knowledge about the effects and benefits of the practice of PA, and the current challenge lies in translating knowledge into strategies to increase PA at a populational level. In our study, we tried to encompass different determinants of PA [4,5] at the individual level (eg, text messages and self-monitoring could support motivation) and the interpersonal level (eg, social support from the research team, in-person activities in group 3, and establishment of cultural norms among participants); provide a social or cultural environment through app features, such as ranking and groups/teams; and build an information environment through counseling and education.

After analyzing the performance of participants in this study, we suggest that intervention protocols aiming to increase the level of PA in research, clinical, and collective contexts should consider the adaptive intervention design as a viable and potentially effective alternative, as well as the association with environmental and institutional interventions [1]. The recommendations of the World Health Organization Global Action Plan [3] and ecological models [4,5] reinforce the complexity of this behavior change.

Finally, it is essential to highlight that recommendations about the ideal quantity of PA are being constantly revised. Considering the current recommendation of the World Health Organization for the adult population that PA of any amount and intensity is better than no PA [37] and the findings from Warburton and Bredin [36] that there are benefits of transitioning from a lower to a higher level of PA independently of the amount, we believe that our protocol was effective, as all participants presented some increase in the average daily step count.

As strengths of our study, we highlight the adaptive intervention design, which is still not explored in PA, and its association with behavior change techniques and complex technology. Moreover, despite the small sample size, we found relevant information that may contribute to advancing knowledge for PA promotion.

Our study had some limitations. Although we completed the first phase of the study, the second phase was not possible due to the onset of the COVID-19 pandemic. Data analysis was limited due to the small sample size and losses to follow-up, especially after the onset of the pandemic, which led to the sudden interruption of the protocol. Although we managed this limitation using adequate statistical treatment of the data, the analysis may be underpowered. Another limitation regarding the data analysis was the intention-to-treat approach adopted in our study, which could underestimate the impact of other factors influencing participation in the study and adherence to the protocol [38]. Finally, the characteristics of the sample, such as a higher level of education and having access to technologies, may limit the generalization of the results.

As observed in other PA studies, most participants completed at least high school [32,39,40]. Epidemiological studies have indicated that a higher education level and income are associated with higher practice of leisure-time PA [4]. One of the

XSL•FO
RenderX

challenges in designing programs for PA is engaging individuals who are less likely to participate in leisure-time PA due to the perception of limited time, a limited understanding of the effects of the practice of PA, or fatigue resulting from physically demanding work [40]. The socioeconomic characteristics of our sample may have contributed to our results.

In conclusion, our adaptive intervention protocol using a smartphone app combined with behavior change techniques led to increases in the level of PA over time, especially among participants from group 1. Participants from group 1 showed increases in PA levels at all assessments, while those from group 2 showed increases only at the final assessment when compared with the initial assessment. Moreover, there was a time effect for the increase in the time spent in MVPA. It is possible that offering stepwise new stimuli and behavior change techniques may contribute positively to the process of behavior change regarding PA. We acknowledge that the small sample size may limit the interpretation of the results.

Acknowledgments

We are thankful to all participants who dedicated their time to this study. We also appreciate the commitment of each member of the Epidemiology and Human Movement team during all phases of the study.

Funding

This study was funded by the Sao Paulo Research Foundation (grant 2016/50249–3) and by the Sao Paulo Research Foundation and the Brazilian Federal Agency for Support and Evaluation (grant 2018/21536).

Data Availability

Data may be available upon reasonable request to the corresponding author.

Authors' Contributions

Conceptualization: MSMPS, NLP, VTL, MBN, RCP, VZD
Data curation: MSMPS, VZD
Formal analysis: MSMPS, VZD
Funding acquisition: VZD
Investigation: MSMPS, NLP, VTL, MBN
Methodology: MSMPS, NLP, VTL, MBN, RCP, VZD
Project administration: MSMPS, VZD
Resources: VZD
Supervision: RCP, VZD
Visualization: MSMPS, NLP, VTL, MBN, RCP, VZD
Writing – original draft: MSMPS
Writing – review & editing: NLP, VTL, MBN, RCP, VZD

Conflicts of Interest

None declared.

Checklist 1
CONSORT - EHEALTH checklist (V 1.6.1).
[PDF File, 10034 KB - jmir_v28i1e73388_app1.pdf]

References

1. Heath GW, Parra DC, Sarmiento OL, et al. Evidence-based intervention in physical activity: lessons from around the world. The Lancet 2012 Jul;380(9838):272-281. [doi: 10.1016/S0140-6736(12)60816-2]
2. Salvador EP, Ribeiro EH, Garcia LM, et al. Interventions for physical activity promotion applied to the primary healthcare settings for people living in regions of low socioeconomic level: study protocol for a non-randomized controlled trial. Arch Public Health 2014 Mar 13;72(1):8. [doi: 10.1186/2049-3258-72-8] [Medline: 24624930]
3. Global action plan on physical activity 2018–2030: more active people for a healthier world. World Health Organization. 2018. URL: https://www.who.int/publications/i/item/9789241514187 [accessed 2025-11-21]
4. Bauman AE, Reis RS, Sallis JF, Wells JC, Loos RJ, Martin BW. Correlates of physical activity: why are some people physically active and others not? The Lancet 2012 Jul;380(9838):258-271. [doi: 10.1016/S0140-6736(12)60735-1]
5. Sallis JF, Cervero RB, Ascher W, Henderson KA, Kraft MK, Kerr J. An ecological approach to creating active living communities. Annu Rev Public Health 2006;27:297-322. [doi: 10.1146/annurev.publhealth.27.021405.102100] [Medline: 16533119]

XSL•FO
RenderX

6. Duncan M, Vandelanotte C, Kolt GS, et al. Effectiveness of a web- and mobile phone-based intervention to promote physical activity and healthy eating in middle-aged males: randomized controlled trial of the ManUp study. J Med Internet Res 2014 Jun 12;16(6):e136. [doi: 10.2196/jmir.3107] [Medline: 24927299]
7. Schoeppe S, Alley S, Van Lippevelde W, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2016 Dec 7;13(1):127. [doi: 10.1186/s12966-016-0454-y] [Medline: 27927218]
8. Fanning J, Mullen SP, McAuley E. Increasing physical activity with mobile devices: a meta-analysis. J Med Internet Res 2012 Nov 21;14(6):e161. [doi: 10.2196/jmir.2171] [Medline: 23171838]
9. Schoeppe S, Alley S, Rebar AL, et al. Apps to improve diet, physical activity and sedentary behaviour in children and adolescents: a review of quality, features and behaviour change techniques. Int J Behav Nutr Phys Act 2017 Jun 24;14(1):83. [doi: 10.1186/s12966-017-0538-3] [Medline: 28646889]
10. Sun L, Wang Y, Greene B, et al. Facilitators and barriers to using physical activity smartphone apps among Chinese patients with chronic diseases. BMC Med Inform Decis Mak 2017 Apr 19;17(1):44. [doi: 10.1186/s12911-017-0446-0] [Medline: 28420355]
11. Middelweerd A, Mollee JS, van der Wal CN, Brug J, Te Velde SJ. Apps to promote physical activity among adults: a review and content analysis. Int J Behav Nutr Phys Act 2014 Jul 25;11:97. [doi: 10.1186/s12966-014-0097-9] [Medline: 25059981]
12. Vandelanotte C, Müller AM, Short CE, et al. Past, present, and future of eHealth and mHealth research to improve physical activity and dietary behaviors. J Nutr Educ Behav 2016 Mar;48(3):219-228. [doi: 10.1016/j.jneb.2015.12.006] [Medline: 26965100]
13. van Stralen MM, Lechner L, Mudde AN, de Vries H, Bolman C. Determinants of awareness, initiation and maintenance of physical activity among the over-fifties: a Delphi study. Health Educ Res 2010 Apr;25(2):233-247. [doi: 10.1093/her/cyn045] [Medline: 18927443]
14. Kumar S, Nilsen WJ, Abernethy A, et al. Mobile health technology evaluation: the mHealth evidence workshop. Am J Prev Med 2013 Aug;45(2):228-236. [doi: 10.1016/j.amepre.2013.03.017] [Medline: 23867031]
15. Collins LM, Murphy SA, Strecher V. The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions. Am J Prev Med 2007 May;32(5 Suppl):S112-S118. [doi: 10.1016/j.amepre.2007.01.022] [Medline: 17466815]
16. Martin SS, Feldman DI, Blumenthal RS, et al. mActive: a randomized clinical trial of an automated mHealth intervention for physical activity promotion. J Am Heart Assoc 2015 Nov 9;4(11):e002239. [doi: 10.1161/JAHA.115.002239] [Medline: 26553211]
17. Gonze BDB, Padovani RDC, Simoes MDS, et al. Use of a smartphone app to increase physical activity levels in insufficiently active adults: feasibility sequential multiple assignment randomized trial (SMART). JMIR Res Protoc 2020 Oct 23;9(10):e14322. [doi: 10.2196/14322] [Medline: 33094733]
18. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. Lancet 2025 Apr 14:S0140-6736(25)00672-5. [doi: 10.1016/S0140-6736(25)00672-5] [Medline: 40245901]
19. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
20. Morais Pereira Simões MDS, de Barros Gonze B, Leite Proença N, et al. Use of a smartphone app combined with gamification to increase the level of physical activity of adults and older adults: protocol of a sequential multiple assignment randomized trial. Trials 2019 Dec 27;20(1):780. [doi: 10.1186/s13063-019-3879-1] [Medline: 31881987]
21. Tudor-Locke C, Craig CL, Thyfault JP, Spence JC. A step-defined sedentary lifestyle index: <5000 steps/day. Appl Physiol Nutr Metab 2013 Feb;38(2):100-114. [doi: 10.1139/apnm-2012-0235] [Medline: 23438219]
22. Conheça Santos [Article in Portuguese]. Prefeitura de Santos. URL: https://www.santos.sp.gov.br/?q=hotsite/conheca-santos [accessed 2022-03-23]
23. Dallal GE. Randomization. 2018. URL: http://www.randomization.com [accessed 2020-12-22]
24. Critério Brasil [Article in Portuguese]. Associação Brasileira de Empresas de Pesquisa. URL: https://abep.org/criterio-brasil/ [accessed 2025-11-21]
25. Guedes DP, Anira dos Santos C, Lopes CC. Estágios de mudança de comportamento e prática habitual de atividade física em universitários [Article in Portuguese]. Rev Bras Cineantropom Desempenho Hum 2006;8(4):5-15 [FREE Full text]
26. Troiano RP, Berrigan D, Dodd KW, Mâsse LC, Tilert T, McDowell M. Physical activity in the United States measured by accelerometer. Med Sci Sports Exerc 2008 Jan;40(1):181-188. [doi: 10.1249/mss.0b013e31815a51b3] [Medline: 18091006]
27. Freedson PS, Melanson E, Sirard J. Calibration of the Computer Science and Applications, Inc. accelerometer. Medicine & Science in Sports & Exercise 1998 May;30(5):777-781. [doi: 10.1097/00005768-199805000-00021]
28. Almirall D, Compton SN, Gunlicks-Stoessel M, Duan N, Murphy SA. Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy. Stat Med 2012 Jul 30;31(17):1887-1902. [doi: 10.1002/sim.4512] [Medline: 22438190]

XSL•FO
RenderX

29. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
30. Wright CE, Rhodes RE, Ruggiero EW, Sheeran P. Benchmarking the effectiveness of interventions to promote physical activity: a metasynthesis. Health Psychol 2021 Nov;40(11):811-821. [doi: 10.1037/hea0001118] [Medline: 34914485]
31. Ringeval M, Wagner G, Denford J, Paré G, Kitsiou S. Fitbit-based interventions for healthy lifestyle outcomes: systematic review and meta-analysis. J Med Internet Res 2020 Oct 12;22(10):e23954. [doi: 10.2196/23954] [Medline: 33044175]
32. Vandelanotte C, Short CE, Plotnikoff RC, et al. Are web-based personally tailored physical activity videos more effective than personally tailored text-based interventions? Results from the three-arm randomised controlled TaylorActive trial. Br J Sports Med 2021 Mar;55(6):336-343. [doi: 10.1136/bjsports-2020-102521] [Medline: 33144346]
33. Plotnikoff RC, Wilczynska M, Cohen KE, Smith JJ, Lubans DR. Integrating smartphone technology, social support and the outdoor physical environment to improve fitness among adults at risk of, or diagnosed with, Type 2 Diabetes: findings from the "eCoFit" randomized controlled trial. Prev Med 2017 Dec;105:404-411. [doi: 10.1016/j.ypmed.2017.08.027] [Medline: 28887192]
34. Shang B, Duan Y, Huang WY, Brehm W. Fluctuation - a common but neglected pattern of physical activity behaviour: an exploratory review of studies in recent 20 years. Eur J Sport Sci 2018 Mar;18(2):266-278. [doi: 10.1080/17461391.2017.1417486] [Medline: 29334317]
35. Kwasnicka D, Dombrowski SU, White M, Sniehotta F. Theoretical explanations for maintenance of behaviour change: a systematic review of behaviour theories. Health Psychol Rev 2016 Sep;10(3):277-296. [doi: 10.1080/17437199.2016.1151372] [Medline: 26854092]
36. Warburton DER, Bredin SSD. Health benefits of physical activity: a systematic review of current systematic reviews. Curr Opin Cardiol 2017 Sep;32(5):541-556. [doi: 10.1097/HCO.0000000000000437] [Medline: 28708630]
37. WHO guidelines on physical activity and sedentary behaviour: at a glance. World Health Organization. 2020. URL: https://www.who.int/publications/i/item/9789240014886 [accessed 2025-11-21]
38. Nich C, Carroll KM. Intention-to-treat meets missing data: implications of alternate strategies for analyzing clinical trials data. Drug Alcohol Depend 2002 Oct 1;68(2):121-130. [doi: 10.1016/s0376-8716(02)00111-4] [Medline: 12234641]
39. Lee CF, Ho JWC, Fong DYT, et al. Dietary and physical activity interventions for colorectal cancer survivors: a randomized controlled trial. Sci Rep 2018 Apr 10;8(1):5731. [doi: 10.1038/s41598-018-24042-6] [Medline: 29636539]
40. Biswas A, Dobson KG, Gignac MAM, de Oliveira C, Smith PM. Changes in work factors and concurrent changes in leisure time physical activity: a 12-year longitudinal analysis. Occup Environ Med 2020 May;77(5):309-315. [doi: 10.1136/oemed-2019-106158] [Medline: 32107318]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
EPIMOV: Epidemiology and Human Movement
MVPA: moderate-to-vigorous physical activity
PA: physical activity
SMART: sequential multiple assignment randomized trial

Edited by J Sarvestan; submitted 03.Mar.2025; peer-reviewed by A Tabaczynski, H Namba; revised version received 14.Jul.2025; accepted 16.Jul.2025; published 30.Jan.2026.

Please cite as:
Simoes MDSMP, Proença NL, Lauria VT, do Nascimento MB, Padovani RDC, Dourado VZ
Effects of Using a Smartphone App Combined With Behavior Change Techniques on the Level of Physical Activity Among Adults and Older Adults: Sequential Multiple Assignment Randomized Trial
J Med Internet Res 2026;28:e73388
URL: https://www.jmir.org/2026/1/e73388
doi:10.2196/73388

© Maria do Socorro Morais Pereira Simoes, Neli Leite Proença, Vinícius Tonon Lauria, Matheus Bibian do Nascimento, Ricardo da Costa Padovani, Victor Zuniga Dourado. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effect of Lung Cancer Screening, Smoking Cessation, and Cessation Smartphone App to Health-Related Quality of Life Among Heavy Smokers: Randomized Controlled Trial

Antti Kurtti[1,2,3], BMed; Sanna Iivanainen[1,2,3], MD, PhD; Riitta Kaarteenaho[3,4], MD, PhD; Heidi Andersen[5,6], MD, PhD; Antti Jekunen[5,6], MD, PhD; Tuula Vasankari[6,7], MD, PhD; Jussi Koivunen[1,2,3], MD, PhD

[1]Cancer Center, Oulu University Hospital, Kajaanintie 50, Oulu, Finland

[2]Faculty of Medicine, University of Oulu, Oulu, Finland

[3]Medical Research Center Oulu, Oulu, Finland

[4]Research Unit of Biomedicine and Internal Medicine, University of Oulu, Oulu, Finland

[5]Department of Oncology and Radiotherapy, Vaasa Central Hospital, Vaasa, Finland

[6]Faculty of Medicine, University of Turku, Turku, Southwest Finland, Finland

[7]FILHA, Helsinki, Finland

Corresponding Author:
Jussi Koivunen, MD, PhD
Cancer Center, Oulu University Hospital, Kajaanintie 50, Oulu, Finland

Abstract

Background: Lung cancer screening with low-dose computed tomography (LDCT) among heavy smokers can decrease lung cancer mortality. Smoking cessation intervention is recommended within the screening program; however, the methods for smoking cessation in the LDCT screening context are not well established. We have previously shown that a novel smartphone app can increase the chance for smoking cessation along with lung cancer screening. The effects of lung cancer screening, smoking cessation, and the use of smartphone apps on health-related quality of life (HRQoL) are widely unknown.

Objective: This study aims to investigate the effect of lung cancer screening, smoking cessation, and the use of smoking cessation app on HRQoL, an exploratory end point of the low-dose computed tomography screening for lung cancer combined to different smoking cessation methods in Finland (LDCT-SC-FI) study.

Methods: This study was conducted as a part of the LDCT-SC-FI (NCT05630950), which was a randomized controlled trial investigating different smoking cessation methods in participants undergoing lung cancer screening with LDCT. The main inclusion criteria included an age of 50 - 74 years, a marked smoking history (smoked ≥15 cigarettes per day for ≥25 years or smoked ≥10 cigarettes per day for ≥30 years), an active smoking status, and access to a smartphone. The recruitment was carried out by newspaper and internet advertisements and informing relevant health care units at hospital districts. The study participants (n=200), all at Oulu University Hospital, were randomized in 1:1 fashion to a yearly LDCT with standard smoking cessation (written material) or a stand-alone smartphone app–based cessation. HRQoL, an exploratory study end point, was assessed at baseline and at 1 year with Quality of Life Questionnaire Core 30 (QLQ-C30) and EQ-5D.

Results: In total, 199 and 186 individuals had both questionnaires completed at baseline and at 1 year, respectively. We did not detect a change in HRQoL between the time points using QLQ-C30 global health status score or EQ-5D index score. Smoking cessation at 1-year time did not affect QLQ-C30 global health status or EQ-5D. We observed improved quality of life scores by EQ-5D at 1 year (control: mean 0.720, SD 0.197 vs app: mean 0.799, SD 0.197; improved in 17/93, 18% of controls vs 29/93, 31% in app arm), while there was no difference in means at baseline. Smartphone app arm reported reduced pain (EQ-5D effect size [ES] 0.049, 95% CI 0.006 - 0.12; *P*=.01; adjusted ES 0.026; *P*=.007; QLQ-C30 ES 0.076, 95% CI 0.02 - 0.16; *P*<.001; adjusted ES 0.05; *P*=.02) and increased mobility (EQ-5D ES 0.031, 95% CI 0.01 - 0.09; *P*=.02; adjusted ES 0.037; *P*=.008) at 1 year. The number of completed questionnaires in the app was associated with improved HRQoL by EQ-5D (ES 0.073, 95% CI 0.00 - 0.180; *P*=.04; adjusted ES 0.071; *P*=.04).

Conclusions: This is the first study to test a smoking cessation smartphone app in the context of lung cancer screening. The use of the developed app correlated with improved HRQoL, mainly by decreased pain and fatigue. To conclude, the studied app provides a feasible and effective cessation intervention that is readily implementable in population-based lung cancer screening programs, with enhanced health benefits beyond smoking cessation.

Trial Registration: ClinicalTrial.gov NCT05630950; https://clinicaltrials.gov/study/NCT05630950

XSL•FO
RenderX

(J Med Internet Res 2026;28:e81687) doi:10.2196/81687

KEYWORDS

lung cancer screening; smoking cessation; mobile app; low-dose computed tomography; health-related quality of life

Introduction

The premier cause of cancer mortality in the Western world is lung cancer, of which smoking is the single most important risk factor [1]. Lung cancer is typically diagnosed at an advanced stage, preventing curative intent treatments. Using low-dose computed tomography (LDCT) for lung cancer screening in individuals with substantial smoking history can decrease lung cancer mortality [2,3]. Concordantly, smoking cessation interventions are recommended in lung cancer screening programs and may come with enhanced efficiency, regardless of screening results [4-7]. Approximately 7% - 23% of participants in LDCT programs achieve smoking cessation [8]. Still, optimal smoking cessation methods in the context of lung cancer screening are not well established.

Possible health-related quality of life (HRQoL) effects and losses are important to evaluate in cancer screening trials [9]. There is moderate evidence that, compared with no screening, persons receiving LDCT screening do not have worse general HRQoL or distress over 2 years of follow-up, and anxiety may even be lowered. However, consequences might differ based on screening results, at least in the short term [10-12]. Observed negative psychological effects diminished over time, and there were no known negative long-term effects on HRQoL [13,14].

Opportunities for using digital tools to access smoking cessation treatment are evolving rapidly due to the expansion in the proportion of the global population with access to a mobile phone. Data imply that apps that provide personalized and adaptive as well as interactive support may be more effective in promoting engagement [15-17]. Smartphone apps have reached abstinence odds ratios (ORs) of 1.25 - 1.51 (95% CI 0.99 - 1.56, 1.24 - 1.84) in meta-analyses [18]. Their efficacy can be enhanced with pharmacotherapy and physical participant recruitment [19]. We have recently shown that a novel smartphone app can increase smoking cessation with an OR of ~3 at 3 and 6 months when applied within LDCT screening for lung cancer [20].

Several studies have indicated that smoking is related to lower HRQoL and mental well-being [21-23]. However, studies on the impact of smoking cessation interventions on HRQoL have provided mixed results, some concluding that it may improve, yet others showing negative or no changes [24-26]. Smoking cessation may have a positive effect on physical and general health but no significant effects on mental aspects [24]. Nevertheless, it might take years of abstinence for smokers' HRQoL to be equal to nonsmokers' [24,27].

Low-dose computed tomography screening for lung cancer combined with different smoking cessation methods in Finland (LDCT-SC-FI) is a randomized controlled trial investigating smoking cessation with a smartphone app compared to written cessation materials in individuals participating in LDCT screening for lung cancer. The core concepts behind the developed app include cognitive behavioral (enhancing self-awareness, problem-solving skills, goal setting, and coping with cravings) and social cognitive theories and acceptance and commitment therapy as well as mindfulness. Therefore, the effects of smartphone apps could extend beyond smoking cessation.

This study aims to investigate changes in HRQoL, an exploratory end point of the LDCT-SCI-FI trial. We investigated the changes in HRQoL using 3 different patient-reported outcome measures (PROMs). Since the LDCT-SC-FI study focused on 2 main themes (feasibility of LDCT screening and smoking cessation with a smartphone app), the results were analyzed to ascertain what effects participation in screening, smoking cessation, and use of the interventional app might have on HRQoL outcomes.

Methods

Study Design

This study was conducted as a part of the LDCT-SC-FI (NCT05630950), which is a randomized controlled trial investigating different smoking cessation methods in participants undergoing lung cancer screening with LDCT. The study participants were randomized in a 1:1 fashion to a yearly LDCT with standard smoking cessation (written material) or the same LDCT screening approach with a smartphone app–based smoking cessation (experimental). The study was powered (80%) with 156 participants to detect a 15% difference with α of .1 (75% vs 90%) in the number of active smokers at 3 and 6 months after inclusion using an online calculator. With the expected dropout rate, the sample size was adjusted to 200.

Ethical Considerations

The study was approved by the ethics committee of Northern Ostrobothnia Hospital District (EETTKM 21/2022) and was prospectively registered at ClinicalTrials.gov (NCT05630950). All the participants signed an informed consent before any study procedures. The study participants were not compensated for their participation. Of note, LDCT lung cancer screening is not among the publicly funded cancer screenings in Finland. The study was conducted in accordance with the Declaration of Helsinki [28] and Good Clinical Practice guidelines [29]. Participant data were deidentified prior to analysis. Unique study codes were assigned to each participant, while identifying information was removed. All results are reported in aggregate form to protect participant confidentiality. Consent for publication has been granted by identifiable individuals.

Participants

Eligibility followed closely to the Nederlands-Leuvens Longkanker Screenings Onderzoek (NELSON) lung cancer screening trial criteria [2]. The main inclusion criteria included an age of 50 - 74 years, a marked smoking history (smoked ≥15 cigarettes per day for ≥25 years or smoked ≥10 cigarettes per

XSL•FO
RenderX

day for ≥30 years), an active smoking status (smoking during the last 2 weeks including regular [daily smoking] and occasional [nondaily smoking] habits), and access to a smartphone (iPhone or Android). The main exclusion criteria, as in the NELSON trial, included a moderate or bad self-reported health; current or past melanoma, lung, renal, or breast cancer; and a chest computed tomography (CT) examination within 1 year.

Recruitment and Randomization

All the participants were recruited at the Oulu University Hospital (from November 18, 2022, to April 14, 2023; the final study visit occurred on March 20, 2024). The recruitment was carried out by newspaper and internet advertisements and informing relevant health care units at the hospital district. A physical screening visit was performed at the site where participating individuals signed the informed consent. Eligibility was verified by a study nurse according to a checklist. Study participants did not receive any compensation for participation, and all the study procedures were free of charge. Eligible participants were randomized by study nurses with block method (sequentially numbered containers with a block size of 10, study arm written on a paper in an opaque, sealed envelope) in 1:1 fashion with stratification according to pack years (<30 or ≥30 pack years) and age (<65 or ≥65 years) to smartphone-based smoking cessation and control (written smoking cessation material) arms. The stratification factors were selected based on the assumption that bias could be generated by (1) adoption of smartphone use in older people and (2) higher pack years to be associated with a lesser likelihood of smoking cessation. The random allocation sequence (randomization envelopes and numbered blocks) was generated by the investigators (JK and SI) to ensure concealment. Because of an inability to blind the study participants from the intervention, as well as self-reported smoking cessation being the primary end point of the study, the study personnel were not blinded.

Outcomes

Data collected at baseline included standard demographics and detailed smoking history. Self-reported smoking cessation status was verified by phone at 3 and 6 months, and as a part of physical visit at 1 year. Another physical visit and LDCT screening investigation took place at 1 year. Quality of life (QoL) questionnaires were collected after randomization at baseline and at 1 year.

The primary outcomes of the study were self-reported smoking cessation at 3 and 6 months (±1 month), which has been recently reported, and study details are available in this publication [20]. Furthermore, secondary outcomes of the study included efficiency of different smoking cessation methods in the reduction of smoking, sensitivity and positive predictive value of CT screening, and costs related to CT screening, which are reported in previous publications. HRQoL assessed with different PROMs was an exploratory end point of the study and was not specified in detail in the protocol. Since the study focused on 2 main themes (smoking cessation with a smartphone app and feasibility of LDCT lung cancer screening), we planned to investigate HRQoL in the context of smoking cessation, study arm, and temporally. All the analysis was carried out according to specific instructions by the HRQoL questionnaire. ED-5D index scores and QLQ-C30 global health status (GHS) have standardized population-based cutoffs (5% for EQ-5D and 10% for the QLQ-C30), which were used in the main analysis of HRQoL. If any of the main analyses showed statistically significant differences temporally, or by study arm, or smoking status at 1 year, detailed analysis of specific subscales or scores would be carried out.

Procedures

The developed smoking cessation app called Suunta supports smokers in the cessation process and aids them to retain a smoking-free lifestyle (Multimedia Appendix 1). The theoretical and functional concept was created by the study team members, and the technical execution was done under subcontract by a company specialized in mobile app development (Techinspire). The stand-alone app is Android- and iPhone-compatible with cloud-based database back-up. The Suunta app was downloaded on participants' smartphones on the randomization visit, which was assisted by the study nurse. The Suunta smartphone app was offered for the individuals in the control arm after 6 months of follow-up.

The written materials used for smoking cessation are based on the Finnish Current Care Guideline for Prevention and Treatment of Smoking and Nicotine Addiction. A printed version of the patient guide was handed out to all the study participants. At the 6-month smoking status call, participants in the control arm were offered the possibility to start using the smoking cessation app. No other counseling for smoking cessation was provided to the participants during the study visits regardless of the study arm.

The LDCT-SC-FI study protocol for LDCT (Multimedia Appendix 2) follows the NELSON study protocol [2]. In brief, all the study participants undergo LDCT screening within 6 weeks from the randomization, and the next LDCT is scheduled for 1 year.

HRQoL Measurements

Study participants were assessed for HRQoL with European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire Core 30 (QLQ-C30)+Quality of Life Questionnaire Lung Cancer 13 (QLQ-LC13) and EuroQol EQ-5D-3L at baseline and at 1 year. The participants completed the questionnaires at a physical visit using paper format. HRQoL and changes therein were collected as the QLQ-C30 GHS and EQ-5D index scores at baseline and 1 year in all participants as well as according to smoking status at 1 year and study arms. HRQoL-related secondary outcomes included the means and changes in the QLQ-C30 functional scales as well as in the specific symptom scores of QLQ-C30, QLQ-LC13, and EQ-5D. These were carried out only if the overall score was found to be significant at a 1-year time point.

The EORTC QLQ-C30 (version 3.0) and QLQ-LC13 questionnaires were scored according to official EORTC scoring manuals. In brief, all the scales and single-item measures undergo linear transformation to range from 0 to 100, so that a high score for the functional and GHS scale represents a high level of functioning or QoL, while the score for the symptom

XSL•FO
RenderX

scales and single items represents a high level of symptomatology or problems. A 10-point change in scores is considered meaningful, so a 10% cutoff was selected for analysis [30].

With EQ-5D-3L questionnaires, single scores were reported by 3-grade answers representing no to severe problems. For the EQ-5D index score, all the scores were transformed to a single raw score between 0 and 1 according to the official scoring manual. Raw scores were transformed to a country-specific index score (Finland) using a Microsoft Excel-based calculator. A 5% change in EQ-5D index score is considered significant in Finland, and this was selected as a primary measure [31,32].

The use of the app and its association with HRQoL changes observed in the QLQ-C30 GHS and EQ-5D index scores were studied in the experimental arm. The smartphone app included weekly symptom questionnaires, and the frequency of app use was investigated by analyzing the number of completed symptom questionnaires by 24 weeks.

Statistical Analysis

Data analysis was carried out using SPSS (version 29.0.1; IBM Corp). All the statistical analyses were carried out blinded to the group allocation. Reliability of HRQoL data was evaluated with Cronbach α, with values of <0.6 considered poor or unacceptable. Paired 2-tailed *t* test was used to compare means of 2 related groups (change over time) and estimate *P* value and effect size (ES) with 95% CIs. For categorized variables, Pearson chi-square test was used to estimate *P* values, and 1-way ANOVA for ES with 95% CIs. For continuous variables, 1-way ANOVA was applied to estimate the means, *P* values, and ES with 95% CIs. One-way ANOVA was also applied in the analysis of variables with more than 2 categories to estimate *P* values and ES with 95% CIs. Adjusted analysis for *P* values and ES was carried out using analysis of covariance corrected for relevant baseline factors. Since the statistical software used does not provide CIs for adjusted analysis of covariance, these are not provided. Ordinal regression analysis with the proportional odds model was used to estimate ORs, 95% CIs, and *P* values. The generated model was evaluated with model fitting, goodness-of-fit, and test-of-parallel-lines test. *P* values of <.05 were considered statistically significant. For the ES, partial η^2 (small: 0.01; medium: 0.06; large: 0.14) or Cohen *d* (small: 0.2; medium: 0.5; large: 0.8) were used to estimate the magnitude of the effect. Missing data were not replaced.

Results

Demographics

The recruitment was initiated on November 18, 2022, the last participant was included on April 14, 2023, and the final study visit occurred on March 20, 2024. The median age was 60 (IQR 56-66) years, and 51% (102/201) of them were female. In smoking-related demographics, the median number of pack years was 31 (IQR 24-40), and the number of smoked cigarettes per day was 15. The detailed demographics are presented in Table S1 in Multimedia Appendix 3 and the study flowchart in Figure 1.

XSL•FO
RenderX

Figure 1. Study flowchart of the LDCT-SC-FI, which was a randomized controlled trial investigating different smoking cessation methods in participants undergoing lung cancer screening with low-dose computed tomography. LDCT-SC-FI: low-dose computed tomography screening for lung cancer combined with different smoking cessation methods in Finland.

XSL•FO
RenderX

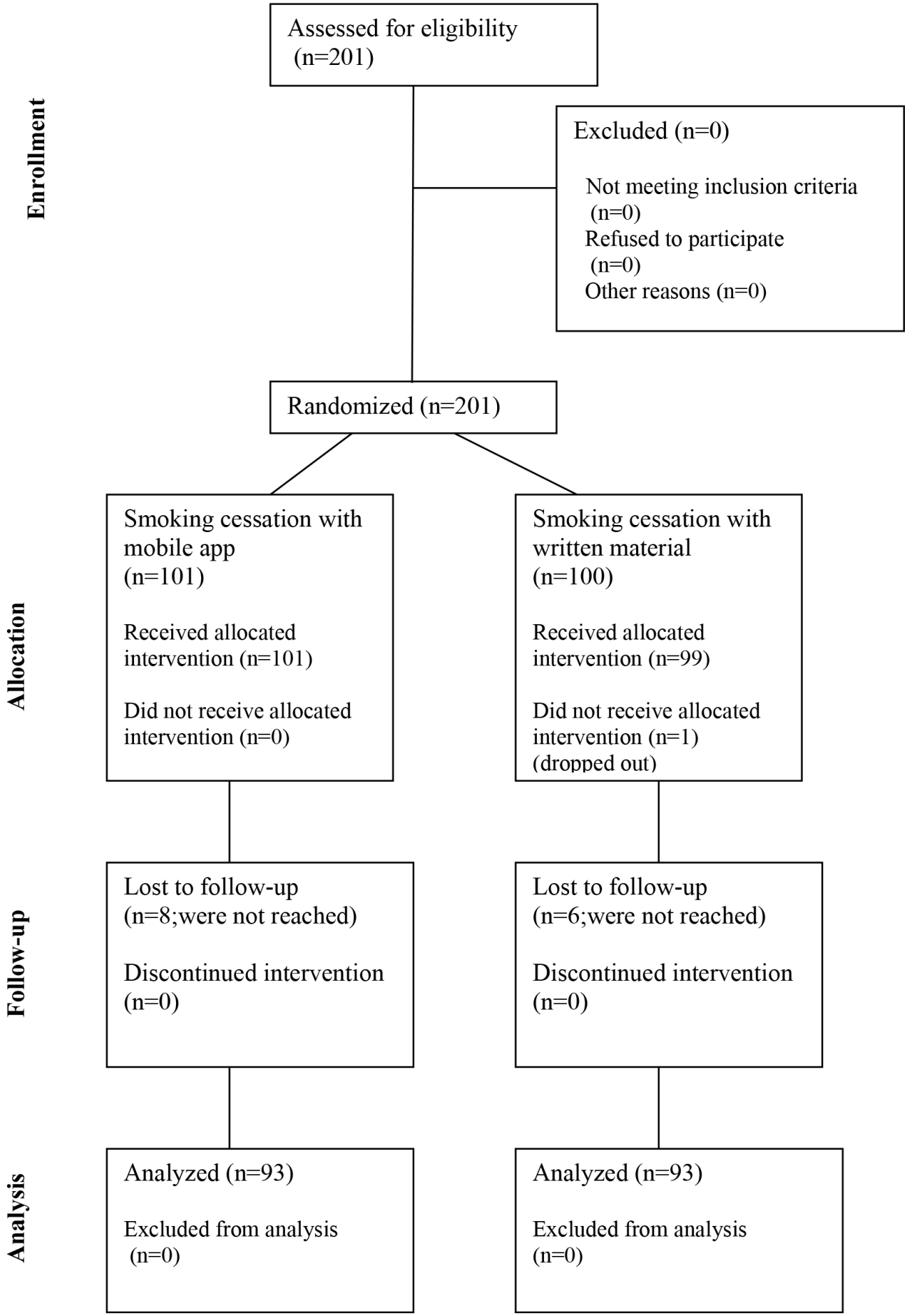

XSL•FO
RenderX

Temporal Changes in HRQoL During Screening

All the study participants (n=200) underwent LDCT screening at baseline and 183 of the eligible 196 (93%) at 1 year. At baseline, all the study participants (n=200) had 3 QoL (EQ-5D and QLQ-CO30+LC13) questionnaires filled, and all (n=186) came to the 1-year study visit (Figure 1). There were 14 dropouts for the 1-year visit (experimental: n=8; control: n=6), 4 by protocol-defined exclusion and 10 by self-willed withdrawal. The overall HRQoL scores and their changes between baseline and 1 year were analyzed. Data reliability was considered good or acceptable for overall HRQoL scores (Table S2 in Multimedia Appendix 3). Among all study participants, there was no statistically significant difference between the means of either score at baseline or at 1 year. However, there was a slight decrease in GHS over time. Most participants experienced no changes in either of their scores (GHS: 120/186, 65%, EQ-5D index score: 90/186, 48%). The subgroup of participants experiencing either improved or decreased HRQOL scores was quite balanced (Table 1). Furthermore, the association of EQ-5D index score change (5%) and GHS change (10%) was evaluated, and they showed a good correlation (Table S3 in Multimedia Appendix 3).

Table . Study participants were assessed for HRQoL[a] with European Organisation for Research and Treatment of Cancer QLQ-C30[b] and EuroQol EQ-5D-3L at baseline and at 1 year at a physical visit[c].

	Baseline	At 1 year	Effect size[d] (95% CI)	*P* value[e]
QLQ-C30 GHS[f], mean (SD)	73.21 (15.83)	71.64 (19.71)	0.09 (−0.6 to 0.23)	.23
QLQ-C30 GHS (10% change), n (%)	—[g]	186 (100)	—	—
Improved		33 (18)		
No change		120 (65)		
Decrease		33 (18)		
EQ-5D index score, mean (SD)	0.763 (0.195)	0.760 (0.201)	0.02 (−0.13 to 0.16)	.80
EQ-5D index score (5% change), n (%)	—	186 (100)	—	—
Improved		46 (25)		
No change		90 (48)		
Decrease		50 (27)		

[a]HRQoL: health-related quality of life.

[b]QLQ-C30: Quality of Life Questionnaire Core 30.

[c]HRQoL and changes therein were collected as the QLQ-C30 GHS and EQ-5D index scores. A 10-point change in scores of QLQ-C30 is considered meaningful; thus, a 10% cutoff was selected, while a 5% change in EQ-5D index score is considered significant in Finland.

[d]Cohen *d*.

[e]Paired samples *t* test.

[f]GHS: global health status.

[g]Not applicable.

HRQoL According to Smoking Status

Since smoking cessation was the primary end point of the study, we wanted to analyze whether smoking cessation would have an impact on HRQoL outcomes. Smoking status was available for all study participants (n=186) coming to the study visit at 1 year (Figure 1). We assessed the association of smoking cessation to GHS and EQ-5D scores. Smoking cessation status showed no statistically significant difference in either of the HRQoL measures at baseline or at 1 year. Furthermore, we detected no difference according to smoking status in participants whose scores improved, decreased, or remained the same (Table 2).

XSL•FO
RenderX

Table . Changes in quality of life based on QLQ-C30[a] and EQ-5D between baseline and at 1 year according to smoking status (at 1 year) irrespective of study arm.

	Smoker	Nonsmoker	Effect size (95% CI)[b]	*P* value[c]
QLQ-C30 GHS[d], mean (SD)				
Baseline	72.78 (15.97)	74.14 (16.72)	0.001 (0.00-0.003)	.70
1 year	71.44 (19.48)	72.70 (21.24)	0.001 (0.00-0.03)	.75
Mean change from baseline	−1.592 (17.09)	−1.437 (22.28)	0.00 (0.00-0.003)	.97
QLQ-C30 GHS (10% change), n (%)	157 (100)	29 (100)	0.006 (0.00-0.05)	.57
Improved	26 (17)	7 (24)		
No change	102 (65)	18 (62.1)		
Declined	29 (19)	4 (14)		
EQ-5D index score, mean (SD)				
Baseline	0.759 (0.196)	0.767 (0.204)	0.00 (0.00-0.02)	.86
1 year	0.760 (0.201)	0.758 (0.202)	0.00 (0.00-0.005)	.96
Mean change from baseline	−0.002 (0.17)	−0.009 (0.17)	0.00 (0.00-0.02)	.85
EQ-5D index score (5% change), n (%)	157 (100)	29 (100)	0.002 (0.00-0.04)	.80
Improved	40 (26)	6 (21)		
No change	76 (48)	14 (48)		
Declined	41 (26)	9 (31)		

[a]QLQ-C30: Quality of Life Questionnaire Core 30.

[b]Partial eta-square.

[c]ANOVA or Pearson chi-square test.

[d]GHS: global health status.

HRQoL According to Randomization Arm

Next, we assessed HRQoL outcomes according to study arm. There was no difference either in the GHS or EQ-5D index score at baseline. At 1 year, we observed a difference between the study arms in the means of both the GHS (68.46 vs 74.82; partial η^2=0.026, 95% CI 0.00 - 0.09; *P*=.03) and the EQ-5D index scores (0.720 vs 0.799; partial η^2=0.039, 95% CI 0.003 - 0.11; *P*=.007), with individuals randomized to the app arm showing higher scores. However, the statistical significance was lost for the GHS after adjusting for baseline scores. Using the 3-class classification of scores (improved, unchanged, or declined), there was a statistically significant difference between the study arms in the portion of participants whose scores improved and declined by EQ-5D (17/186, 18% vs 29/186, 31%; 31/186, 33% vs 19/186, 20%; partial η^2=0.032, 95% CI 0.001 - 0.10; *P*=.049). A similar trend was observed with GHS (improved: 11/186, 12% vs 22/186, 24%), but this did not reach statistical significance (Table 3).

XSL•FO
RenderX

Table . Changes in overall quality of life based on QLQ-C30[a] and EQ-5D questionnaires between randomization arms[b].

	Control[c]	App[d]	*P* value[e]	Effect size[f] (95% CI)	Adjusted *P* value[g]	Adjusted effect size[g] (95% CI)
QLQ-C30 GHS[h], mean (SD)						
Baseline	70.29 (15.85)	74.17 (16.31)	.09	0.015 (0.00-0.06)	—[i]	—
1 year	68.46 (19.77)	74.82 (19.23)	.03	0.026 (0.00-0.09)	.18	0.01
Mean change from baseline	−2.330 (16.41)	−0.807 (19.39)	.56	0.002 (0.00-0.03)	—	—
QLQ-C30 GHS (10% change), n (%)	93 (100)	93 (100)			—	—
Improved	11 (12)	22 (24)	.11	0.016 (0.00-0.07)		
No change	64 (69)	56 (60)	—	—		
Declined	18 (19)	15 (16)	—	—		
EQ-5D index score, mean (SD)						
Baseline	0.752 (0.197)	0.759 (0.206)	.82	0.000 (0.00-0.02)	—	—
1 year	0.720 (0.197)	0.799 (0.197)	.007	0.039 (0.003-0.11)	.003	0.047
Mean change from baseline	−0.033 (0.16)	0.027 (0.17)	.01	0.032 (0.001-0.10)	—	—
EQ-5D index score (5% change), n (%)	93 (100)	93 (100)			—	—
Improved	17 (18)	29 (31)	.049	0.032 (0.001-0.10)		
No change	45 (48)	45 (48)	—	—		
Declined	31 (33)	19 (20)	—	—		

[a]QLQ-C30: Quality of Life Questionnaire Core 30.

[b]Comparison is carried out based on means at individual time points, changes from baseline, means adjusted to baseline values, and by using clinically significant threshold changes (QLQ-C30 10% change; EQ-5D 5% change).

[c]Control: smoking cessation with written material.

[d]App: smartphone app–based smoking cessation.

[e]ANOVA or Pearson chi-square test.

[f]Partial eta-square.

[g]Analysis of covariance test adjusted for baseline QLQ-C30 GHS or EQ-5D index score.

[h]GHS: global health status.

[i]Not applicable.

We also carried out an ordinal regression analysis of EQ-5D index change (5%) with baseline factors and smoking status at 1 year. The model performance was poor with high *P* values for model fit, while goodness of fit and test of parallel lines showed both *P* values over >.5. Of the analyzed factors, only randomization arm (OR −0.668, 95% CI −1.227 to −0.109; *P*=.02) significantly predicted EQ-5D change (Table 4).

Since Suunta app was also offered to the control arm after 6 months, we wanted to assess whether this would affect the observed results. The app use was assessed at the 1-year study visit, and 14 participants (15% of the controls) reported having used the app. Removing these individuals from the analysis did not markedly alter the results compared to the intent-to-treat analysis. In brief, EQ-5D mean index scores and QLQ-C30 GHS means were similar at baseline, while there was a statistically significant difference at 1 year. Furthermore, according to 3-class classification of EQ-5D or QLQ-C30, very similar results favoring the app were observed while this was only statistically significant with QLQ-C30 (partial η^2=0.027, 95% CI 0.00 - 0.09; *P*=.09; partial η^2=0.023, 95% CI 0.00 - 0.08; *P*=.04, respectively; Table S4 in Multimedia Appendix 3).

XSL•FO
RenderX

Table . Ordinal regression analysis for clinically significant EQ-5D index change (5%) at 1 year.

	OR[a] (95% CI)	*P* value[b]
Age	0.034 (−0.018 to 0.087)	.20
Sex		
Female	−0.001 (−0.633 to 0.630)	.10
Male	0	—[c]
Relationship status		
Single	−0.082 (−0.713 to 0.548)	.83
In a relationship	0	—
ICT[d] skills		
Novice	−0.962 (−2.099 to 0.174)	.10
Average	−0.124 (−0.700 to 0.452)	.67
Experienced	0	—
How many cigarettes per day	0.004 (−0.072 to 0.080)	.92
Pack years	−0.003 (−0.041 to 0.035)	.88
Fagerström test	−0.23 (−0.337 to 0.291)	.89
Smoking status at 1 year		
Nonsmokers	−0.363 (−1.130 to 0.405)	.35
Smoker	0	—
Randomization arm		
Control	−0.668 (−1.227 to −0.109)	.02
App	0	—

[a]OR: odds ratio.

[b]Ordinal regression analysis with proportional odds model; model fitting: *P*=.36; goodness-of-fit: *P*=.38; test of parallel lines: *P*=.06.

[c]Not applicable.

[d]ICT: information and communication technology.

Data reliability of QLQ-C30-, LC13-, and EQ-5D–defined functioning or symptom subscales was considered good to questionable, and only a few symptoms (hemoptysis, sore mouth, and alopecia) produced poor or unacceptable scores (Table S2 in Multimedia Appendix 3). In the analysis for the QLQ-C30 functional scales, we observed no difference at baseline. At 1 year, there were statistically significant differences between study arms in role (mean 81.70, SD 25.20 vs mean 89.13, SD 19.83; partial η^2=0.026, 95% CI 0.00 - 0.09; *P*=.03) and social (mean 86.02, SD 19.24 vs mean 93.55, SD 15.74; partial η^2=0.04, 95% CI 0.01 - 0.11; *P*=.004) functioning scales, with higher means detected in the app arm. After adjusting for baseline scores, only social functioning scales retained the statistical significance (partial η^2=0.025; *P*=.03). Furthermore, improvements in means over time occurred only in the app arm (Table 5).

XSL•FO
RenderX

Table . QLQ-C30[a] functional scales of quality of life according to randomization arm[b].

	Control[c], mean (SD)	App[d], mean (SD)	*P* value[e]	Effect size[f] (95% CI)	Adjusted *P* value[g]	Adjusted effect size[f]
Physical functioning					—[h]	—
Baseline	80.49 (16.90)	82.60 (16.52)	.38	0.004 (0.00 to 0.04)		
1 year	79.20 (18.06)	81.54 (17.86)	.38	0.004 (0.00 to 0.04)		
Mean change	−2.097 (12.78)	−2.148 (13.40)	.98	0.00 (0.00 to 0.00)		
Role functioning						
Baseline	86.03 (21.65)	89.11 (18.62)	.28	0.006 (−0.00 to 0.04)	—	—
1 year	81.70 (25.20)	89.13 (19.83)	.03	0.026 (0.00 to 0.09)	.11	0.014
Mean change	−4.348 (20.36)	−1.268 (17.69)	.28	0.007 (0.00 to 0.05)	—	—
Emotional functioning					—	—
Baseline	83.93 (15.29)	85.81 (14.84)	.38	0.004 (0.00 to 0.04)		
1 year	82.97 (19.27)	87.72 (14.04)	.06	0.02 (0.00 to 0.08)		
Mean change	−0.815 (16.53)	0.807 (12.17)	.45	0.003 (0.00 to 0.04)		
Cognitive functioning					—	—
Baseline	87.71 (16.08)	88.78 (15.11)	.63	0.001 (0.00 to 0.03)		
1 year	85.90 (19.07)	90.50 (16.00)	.08	0.017 (0.00 to 0.07)		
Mean change	−1.648 (14.50)	0.000 (13.23)	.42	0.004 (0.00 to 0.04)		
Social functioning						
Baseline	90.57 (18.16)	94.72 (12.00)	.06	0.018 (0.00 to 0.07)	—	—
1 year	86.02 (19.24)	93.55 (15.74)	.004	0.044 (0.01 to 0.11)	.03	0.025
Mean change	−4.480 (19.06)	−1.971 (17.18)	.35	0.005 (0.00 to 0.04)	—	—

[a]QLQ-C30: Quality of Life Questionnaire Core 30.

[b]Comparison was carried out based on means at individual time points, changes from baseline, and means adjusted to baseline values.

[c]Control: smoking cessation with written material.

[d]App: smartphone app–based smoking cessation.

[e]ANOVA.

[f]Partial eta-square.

[g]Analysis of covariance test adjusted for baseline QLQ-C30 functional scales.

[h]Not applicable.

Of the specific symptom scores, fatigue, pain, insomnia, and financial difficulties were calculated based on the QLQ-C30, while dyspnea at rest and pain in other parts were based on the QLQ-LC13. In the analysis, only insomnia and dyspnea at rest showed differences at baseline, but differences were subtle. At 1 year, fatigue, pain, insomnia, financial difficulties, and pain in other parts showed lower means and symptom burden in the app arm. After adjusting for baseline scores, only pain and final difficulties retained statistical difference (partial η^2=0.05; *P*=.02; partial η^2=0.03; *P*=.02). Only the pain score showed a statistically significant difference in mean change (7.065 vs −0.538; partial η^2=0.026, 95% CI 0.00 - 0.09; *P*=.03) between the study arms (Table 6).

We also analyzed specific symptoms from the EQ-5D questionnaire. In the analysis, there was no difference at baseline according to study arm (not shown). At 1 year, we observed improvements in both mobility (partial η^2=0.031, 95% CI 0.001 - 0.09; *P*=.02) and pain or discomfort (partial η^2=0.049, 95% CI 0.006 - 0.12; *P*=.01) with individuals randomized to the app arm, and these retained in the adjusted analysis (partial η^2=0.037; *P*=.008; partial η^2=0.026; *P*=.007; Table 7).

XSL•FO
RenderX

Table . QLQ-C30[a]+QLQ-LC13[b]–specific symptom scores showing significant changes according to randomization arm[c].

	Control[d], mean (SD)	App[e], mean (SD)	*P* value[f]	Effect size[g] (95% CI)	Adjusted *P* value[h]	Adjusted effect size[g]
Fatigue						
Baseline	24.02 (18.08)	21.56 (19.93)	.36	0.004 (0.00-0.04)	—[i]	—
1 year	26.05 (21.27)	19.59 (17.75)	.03	0.027 (0.00-0.09)	.10	0.015
Mean change	2.031 (15.62)	−0.239 (17.91)	.36	0.005 (0.00-0.04)	—	—
Pain						
Baseline	24.32 (27.62)	18.65 (22.15)	.11	0.013 (0.00-0.06)	—	—
1 year	32.26 (32.68)	16.49 (21.35)	<.001	0.076 (0.02-0.16)	.02	0.05
Mean change	7.065 (26.29)	−0.538 (20.03)	.03	0.026 (0.00-0.09)	—	—
Insomnia						
Baseline	25.25 (27.80)	17.82 (25.19)	.049	0.019 (0.00-0.07)	—	—
1 year	22.58 (27.44)	15.05 (22.26)	.04	0.022 (0.00-0.08)	.25	0.07
Mean change	−1.792 (25.34)	−1.792 (23.24)	>.99	0.00 (0.00-0.00)	—	—
Financial difficulties						
Baseline	10.77 (19.53)	5.94 (17.25)	.06	0.017 (0.00-0.07)	—	—
1 year	13.26 (22.60)	3.94 (14.62)	.001	0.057 (0.01-0.13)	.02	0.03
Mean change	2.151 (18.26)	−0.717 (14.73)	.24	0.007 (0.00-0.05)	—	—
Dyspnea in rest					—	—
Baseline	1.35 (6.60)	4.29 (11.22)	.02	0.025 (0.00-0.008)		
1 year	3.23 (9.91)	2.51 (12.27)	.66	0.001 (0.00-0.03)		
Mean change	1.792 (9.02)	−1.075 (14.29)	.10	0.014 (0.00-0.07)		
Pain in other parts						
Baseline	33.33 (32.97)	26.53 (30.26)	.14	0.012 (0.00-0.06)	—	—
1 year	37.12 (35.53)	24.73 (30.26)	.01	0.035 (0.002-0.10)	.06	0.02
Mean change	2.811 (30.45)	−1.482 (35.30)	.40	0.004 (0.00-0.04)	—	—

[a]QLQ-C30: Quality of Life Questionnaire Core 30.

[b]QLQ-LC13: Quality of Life Questionnaire Lung Cancer 13.

[c]Comparison was carried out based on means at individual time points, changes from baseline, and means adjusted to baseline values.

[d]Control: smoking cessation with written material.

[e]App: smartphone app–based smoking cessation.

[f]ANOVA.

[g]Partial eta-square.

[h]Analysis of covariance adjusted for baseline QLQ-C30 symptom scores.

[i]Not applicable.

XSL•FO
RenderX

Table . EQ-5D–specific symptom scores and their change between baseline and 1 year according to prespecified classification of the questionnaire according to randomization arm[a].

	Control[b], n (%)	App[c], n (%)	*P* value[d]	Effect size[e] (95% CI)	Adjusted *P* value[f]	Adjusted effect size[f]
Mobility						
Baseline			.86	0.00 (0.00-0.017)	—[g]	—
No problems	61 (62)	61 (60)				
Some problems	38 (38)	40 (40)				
1 year			.02	0.031 (0.001-0.09)	.008	0.037
No problems	48 (52)	64 (69)				
Some problems	45 (48)	29 (31)				
Pain or discomfort						
Baseline			.46	0.006 (0.00-0.04)	—	—
No pain or discomfort	45 (46)	52 (52)				
Moderate pain or discomfort	51 (52)	58 (48)				
Extreme pain or discomfort	3 (3)	1 (1)				
1 year			.01	0.049 (0.006-0.12)	.007	0.026
No pain or discomfort	36 (39)	55 (59)				
Moderate pain or discomfort	50 (54)	36 (39)				
Extreme pain or discomfort	7 (8)	2 (2)				

[a]Comparison was carried out based on symptom scales.

[b]Control: smoking cessation with written material.

[c]App: smartphone app–based smoking cessation.

[d]Pearson chi-square test.

[e]ANOVA.

[f]Analysis of covariance adjusted for baseline EQ-5D symptom scores.

[g]Not applicable.

Frequency of App Use and Its Association to HRQoL

To further support that our observed improvements result from app use, we analyzed the connection between app use and changes in GHS and EQ-5D index scores. We assessed app use by the number of questionnaires the user had completed in the app over the course of the first 24 weeks. There was a statistically significant difference between the mean number of filled questionnaires by those whose EQ-5D index score showed no change or improved (partial η^2=0.073, 95% CI 0.00 - 0.18; *P*=.04; partial η^2=0.071; adjusted *P*=.04). In the analysis of GHS and app use, nonsignificant trends similar to those of the EQ-5D were observed (Table 8).

XSL•FO
RenderX

Table . The extent of app use in relation to health-related quality of life[a].

	Improved	No change	Declined	*P* value[b]	Effect size[b]	Adjusted *P* value[c]	Adjusted effect size[c]
QLQ-C30[d] GHS[e] (10% change), n (%)	21 (23)	55 (60)	15 (16.5)	—[f]	—	—	—
Completed questionnaires, mean (SD)	11.14 (14)	6.22 (7)	5.73 (6)	.08	0.056 (0.00-0.16)	—	—
EQ-5D index score (5% change), n (%)	29 (32)	43 (47)	19 (21)	—	—	—	—
Completed questionnaires, mean (SD)	9.86 (11)	7.40 (8)	3.05 (4)	.04	0.073(0.00-0.18)	.04	0.071

[a]An analysis of EQ-5D index score (5%) or QLQ-C30 (10%) change in the experimental arm. The use of the app was assessed by the number of questionnaires the user had completed in the smoking cessation app over the course of the first 24 weeks.

[b]One-way ANOVA.

[c]Analysis of covariance adjusted for age, sex, pack years.

[d]QLQ-C30: Quality of Life Questionnaire Core 30.

[e]GHS: global health status.

[f]Not applicable.

Discussion

Principal Findings

To our knowledge, LDCT-SC-FI was the first to report the results of a smartphone app–based smoking cessation in a randomized controlled trial setting among individuals participating in LDCT lung cancer screening. In this study, we investigated the changes in HRQoL in participants undergoing LDCT screening for lung cancer over a 1-year study period using 3 different PROMs. Additionally, we analyzed the effects of smoking cessation as well as the use of our interventional smoking cessation app on HRQoL outcomes. According to our results, neither LDCT screening for lung cancer nor smoking cessation was associated with changes in HRQoL, while the individuals randomized to the Suunta smoking cessation smartphone app reported improved HRQoL mainly by decreased pain and mobility. In general, observed changes in mean values of HRQoL are small, but these, however, fulfill the population norm for significant change (±5%) using the EQ-5D measure. Our results suggest that the positive effects of smoking cessation apps can be comprehensive and may not be limited to cessation only.

In addition to early detection of lung cancer, LDCT screening can have various health effects, which need to be considered when initiating screening programs [2,3,9,11,12]. There is some evidence that LDCT screening may induce negative effects on participants' HRQoL, especially psychosocial in nature. In most studies, these effects have been short-term and reversed over time [11,13,14]. Our results are in line with these findings, suggesting that lung cancer screening does not negatively impact HRQoL. Since our study had a limited number of participants, we were not able to study the correlation between screening results and HRQoL.

Smoking is the leading risk factor for lung cancer, and smoking cessation clearly prevents the disease as well as improves outcomes [1]. Lung cancer screening programs offer a superb opportunity for smoking cessation interventions and enhance their efficacy [4-6,8]. We have recently shown that smoking cessation can be enhanced 3-fold, and it is of interest if this would translate to alter HRQoL [20]. However, our results showed that smoking cessation does not have an impact on HRQoL or specific scales or symptoms, even though we had hypothesized the contrary. Reflecting on the previous literature, our negative results are not surprising since positive HRQoL effects of smoking cessation might require years to develop [24,27].

Comparison to Prior Work

Currently, smartphones are ubiquitous and provide an accessible platform for health interventions. In various medical conditions, including cancer, such interventions have shown beneficial effects on HRQoL among other study outcomes [33-35]. In recent years, various smartphone-based methods for smoking cessation have been developed, and some have proven to be effective [18,19]. To our knowledge, these studies have not assessed HRQoL, so our study brings novel information on this topic. Although the Suunta app was effective in smoking cessation and surpassed meta-analysis–based ORs for smoking cessation with mobile apps, we observed no association between cessation at 1 year and HRQoL [18,20]. Interestingly, our study showed that individuals randomized to the Suunta app arm presented with improved overall HRQoL as well as specific scales and symptoms. Improvements in HRQoL seemed to be driven by improvements in social function, mobility, as well as pain. Our findings suggest that the improved HRQoL is a result of a direct effect of Suunta app use, rather than this resulting due to smoking cessation. The effects of mobile health apps on

XSL•FO
RenderX

HRQoL have seldom been investigated, and it is possible that other apps would bear similar effects. Furthermore, HRQoL, or happiness, might also be measured with nonvalidated instruments [36]. These are interesting areas of future research.

The theoretical background of the Suunta app includes cognitive behavioral and social cognitive theories and commitment therapy as well as mindfulness. The participants may use the app for goal setting, decision-making, and personal empowerment in smoking cessation as well as for overall management of their health. Functionalities include a weekly symptom questionnaire with personalized feedback, mindfulness practices, self-reflection, and a guided smoking cessation plan. Observed improvements in functional scales might relate to these features, which aim to enhance self-reflection capabilities and induce empowerment.

It is unclear how the Suunta app relays its favorable effects on the pain symptom scores, observed with all 3 PROMs used in this study. To be precise, the pain symptom scores decreased in the app arm over the 1-year study period, whereas in the control arm, they increased. Since the sensation of pain is a complex phenomenon with a psychosocial component, features of the app, specifically mindfulness meditation–based exercises, might alter the overall experience of pain. There are evolving data that mindfulness meditation–based analgesia reduces pain through multiple, unique neural mechanisms and may lead to long-term effects [37]. Our results suggested that the app favorably affected generalized pain rather than regional or nociceptive pain, for example, chest pain and angina, which likely reflects its effect on the psychosocial, and maybe even neurophysiological components of pain.

We also analyzed the extent of app use in relation to HRQoL, and the more frequent use was associated with improved overall HRQoL. This favors the explanation that the features of and adherence to the app are involved in the changes observed. In addition, we have previously shown that adherence to app use was associated with increased smoking cessation [20]. Therefore, the means to increase app adherence could enhance the effects on both smoking cessation and HRQoL. Our results indicate that a mobile app can assist in achieving smoking cessation and retaining or even improving HRQoL during the LDCT screening for lung cancer.

Strengths and Limitations

Our study has some limitations. The study recruitment occurred only in a single center, and the number of participants was moderate, which might limit the generalizability of the results to a wider screening population. Smoking was assessed by self-reporting without biochemical verification, which may overestimate the number of nonsmokers. Yet, it has previously been shown that self-reported smoking cessation follows closely to biochemically verified data [38]. HRQoL was an exploratory end point of the trial, and analysis was not prespecified in detail, and the results should be interpreted with caution. In addition, HRQoL was analyzed only at 2 time points, which limits the time scope of the analysis. Nevertheless, in the context of the study population, the frequency of HRQoL assessments is reasonable. The credibility of our results is supported by using 3 different PROMs assessing HRQoL, all of which produced similar results. It has been suggested that combining generic HRQoL measures with more condition-specific ones may increase their overall usefulness [39]. In addition, the coverage of HRQoL questionnaires was very high because of good participant adherence. The smartphone app was offered to the controls at 6 months, which might have had an impact on the observed results. However, only 15% (14/93) of the controls reported app use at 1 year, and the observed HRQoL results were very similar to intention-to-treat when excluding these from analysis. As in all clinical trials, there were dropouts in our study. The dropout rate was 7.5%, which is very low compared to smoking cessation studies that often report rates as high as 50%. Furthermore, the dropout rates were very similar in both arms, and it is likely that dropouts have only a marginal impact on observed results.

Conclusions

We investigated the effect of LDCT lung cancer screening, smoking cessation, and a smoking cessation app on HRQoL. The LDCT-SC-FI trial was the first to investigate a smoking cessation mobile app in the context of LDCT; thus, the presented HRQoL results are unique. We were able to show that randomization to the smartphone app arm, as well as the frequency of app use, was associated with improved HRQoL, while participation in screening or smoking cessation had no effect. According to the study results, the beneficial impact of the app is driven by enhanced social functioning, mobility, and pain compared to controls. Our study suggests that a smartphone app aiming at smoking cessation is an effective intervention in the context of LDCT lung cancer screening since it can both enhance smoking cessation and improve HRQoL.

Acknowledgments

Generative artificial intelligence was not used in any portion of the manuscript.

Funding

This study was supported by AstraZeneca, Roche, and Cancer Foundation Finland. The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

Data collected for the study will be made available on a reasonable request to the corresponding author.

XSL•FO
RenderX

Authors' Contributions

JK, SI, HA, AJ, RK, and TV designed the study. JK, SI, and AK collected the data. JK, SI, and AK analyzed data and drafted the manuscript. All the authors read and approved the final version of the manuscript.

Conflicts of Interest

AK, HA, and AJ declare no conflict of interest. SI reports institutional grants from AstraZeneca and Roche for the conduct of this study. SI reports personal fees from MSD, Roche, BMS, AstraZeneca, Novartis, Takeda, and Eisai and lecture fees from Siemens Healthineers, all outside the submitted work. RK reports consulting, lecture, and advisory board fees from Boehringer Ingelheim and an advisory board fee from MSD outside the submitted work. TV reports advisory board fees from NordicInfu Care, MSD, and AstraZeneca outside the submitted work. JK reports institutional grants from AstraZeneca and Roche for the conduct of this study. JK reports a personal grant for the conduct of the study from Cancer Foundation Finland. JK reports personal fees from Roche, AstraZeneca, Janssen, BMS, Merck, Amgen, Novartis, Merck KgA, Sanofi, and Pfizer and lecturing fees from Siemens Healthineers, all outside of the submitted work. JK is a former part-time employee of Faron Pharmaceuticals.

Multimedia Appendix 1
Snapshots of the app.
[PDF File, 1108 KB - jmir_v28i1e81687_app1.pdf]

Multimedia Appendix 2
Study protocol.
[PDF File, 1516 KB - jmir_v28i1e81687_app2.pdf]

Multimedia Appendix 3
LDCT-SC-FI study demographics, and the integrity of the HRQoL questionnaire data.
[DOCX File, 39 KB - jmir_v28i1e81687_app3.docx]

Checklist 1
CONSORT-eHealth checklist (V 1.6.1).
[PDF File, 361 KB - jmir_v28i1e81687_app4.pdf]

References

1. Tesfaw LM, Dessie ZG, Mekonnen Fenta H. Lung cancer mortality and associated predictors: systematic review using 32 scientific research findings. Front Oncol 2023;13:1308897. [doi: 10.3389/fonc.2023.1308897] [Medline: 38156114]
2. de Koning HJ, van der Aalst CM, de Jong PA, et al. Reduced lung-cancer mortality with volume CT screening in a randomized trial. N Engl J Med 2020 Feb 6;382(6):503-513. [doi: 10.1056/NEJMoa1911793] [Medline: 31995683]
3. National Lung Screening Trial Research Team, Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011 Aug 4;365(5):395-409. [doi: 10.1056/NEJMoa1102873] [Medline: 21714641]
4. Brain K, Carter B, Lifford KJ, et al. Impact of low-dose CT screening on smoking cessation among high-risk participants in the UK Lung Cancer Screening Trial. Thorax 2017 Oct;72(10):912-918. [doi: 10.1136/thoraxjnl-2016-209690] [Medline: 28710339]
5. Kauczor HU, Baird AM, Blum TG, et al. ESR/ERS statement paper on lung cancer screening. Eur Radiol 2020 Jun;30(6):3277-3294. [doi: 10.1007/s00330-020-06727-7] [Medline: 32052170]
6. van der Aalst CM, van Klaveren RJ, van den Bergh KAM, Willemsen MC, de Koning HJ. The impact of a lung cancer computed tomography screening result on smoking abstinence. Eur Respir J 2011 Jun;37(6):1466-1473. [doi: 10.1183/09031936.00035410] [Medline: 21148233]
7. Wood DE, Kazerooni EA, Aberle D, et al. NCCN Guidelines® Insights: Lung Cancer Screening, Version 1.2022. J Natl Compr Canc Netw 2022 Jul;20(7):754-764. [doi: 10.6004/jnccn.2022.0036] [Medline: 35830884]
8. Moldovanu D, de Koning HJ, van der Aalst CM. Lung cancer screening and smoking cessation efforts. Transl Lung Cancer Res 2021 Feb;10(2):1099-1109. [doi: 10.21037/tlcr-20-899] [Medline: 33718048]
9. Ratushnyak S, Hoogendoorn M, van Baal PHM. Cost-effectiveness of cancer screening: health and costs in life years gained. Am J Prev Med 2019 Dec;57(6):792-799. [doi: 10.1016/j.amepre.2019.07.027] [Medline: 31753260]
10. Bonney A, Malouf R, Marchal C, et al. Impact of low-dose computed tomography (LDCT) screening on lung cancer-related mortality. Cochrane Database Syst Rev 2022 Aug 3;8(8):CD013829. [doi: 10.1002/14651858.CD013829.pub2] [Medline: 35921047]

XSL•FO
RenderX

11. Gareen IF, Duan F, Greco EM, et al. Impact of lung cancer screening results on participant health-related quality of life and state anxiety in the National Lung Screening Trial. Cancer 2014 Nov 1;120(21):3401-3409. [doi: 10.1002/cncr.28833] [Medline: 25065710]
12. Jonas DE, Reuland DS, Reddy SM, et al. Screening for lung cancer with low-dose computed tomography: updated evidence report and systematic review for the US Preventive Services Task Force. JAMA 2021 Mar 9;325(10):971-987. [doi: 10.1001/jama.2021.0377] [Medline: 33687468]
13. van den Bergh KAM, Essink-Bot ML, Borsboom G, Scholten ET, van Klaveren RJ, de Koning HJ. Long-term effects of lung cancer computed tomography screening on health-related quality of life: the NELSON trial. Eur Respir J 2011 Jul;38(1):154-161. [doi: 10.1183/09031936.00123410] [Medline: 21148229]
14. Wu GX, Raz DJ, Brown L, Sun V. Psychological burden associated with lung cancer screening: a systematic review. Clin Lung Cancer 2016 Sep;17(5):315-324. [doi: 10.1016/j.cllc.2016.03.007] [Medline: 27130469]
15. Naughton F, Nadkarni A. The contribution of digital treatment to efforts to reduce global tobacco use. N Engl J Med 2025 Oct 16;393(15):1449-1452. [doi: 10.1056/NEJMp2500683] [Medline: 41085052]
16. Perski O, Hébert ET, Naughton F, Hekler EB, Brown J, Businelle MS. Technology-mediated just-in-time adaptive interventions (JITAIs) to reduce harmful substance use: a systematic review. Addiction 2022 May;117(5):1220-1241. [doi: 10.1111/add.15687] [Medline: 34514668]
17. Szinay D, Cameron RA, Jones A, et al. Eliciting preferences for the uptake of smoking cessation apps: discrete choice experiment. J Med Internet Res 2025 Jan 14;27:e37083. [doi: 10.2196/37083] [Medline: 39808479]
18. Sha L, Yang X, Deng R, et al. Automated digital interventions and smoking cessation: systematic review and meta-analysis relating efficiency to a psychological theory of intervention perspective. J Med Internet Res 2022 Nov 16;24(11):e38206. [doi: 10.2196/38206] [Medline: 36383408]
19. Guo YQ, Chen Y, Dabbs AD, Wu Y. The effectiveness of smartphone app-based interventions for assisting smoking cessation: systematic review and meta-analysis. J Med Internet Res 2023 Apr 20;25:e43242. [doi: 10.2196/43242] [Medline: 37079352]
20. Iivanainen S, Kurtti A, Wichmann V, et al. Smartphone application versus written material for smoking reduction and cessation in individuals undergoing low-dose computed tomography (LDCT) screening for lung cancer: a phase II open-label randomised controlled trial. Lancet Reg Health Eur 2024 Jul;42:100946. [doi: 10.1016/j.lanepe.2024.100946] [Medline: 39070744]
21. Becoña E, Vázquez MI, Míguez MDC, et al. Smoking habit profile and health-related quality of life. Psicothema 2013;25(4):421-426. [doi: 10.7334/psicothema2013.73] [Medline: 24124772]
22. Coste J, Quinquis L, D'Almeida S, Audureau E. Smoking and health-related quality of life in the general population. Independent relationships and large differences according to patterns and quantity of smoking and to gender. PLoS ONE 2014;9(3):e91562. [doi: 10.1371/journal.pone.0091562] [Medline: 24637739]
23. Mesquita R, Gonçalves CG, Hayashi D, et al. Smoking status and its relationship with exercise capacity, physical activity in daily life and quality of life in physically independent, elderly individuals. Physiotherapy 2015 Mar;101(1):55-61. [doi: 10.1016/j.physio.2014.04.008] [Medline: 25108641]
24. Moayeri F, Hsueh YSA, Dunt D, Clarke P. Smoking cessation and quality of life: insights from analysis of longitudinal Australian data, an application for economic evaluations. Value Health 2021 May;24(5):724-732. [doi: 10.1016/j.jval.2020.11.022] [Medline: 33933242]
25. Taylor G, McNeill A, Girling A, Farley A, Lindson-Hawley N, Aveyard P. Change in mental health after smoking cessation: systematic review and meta-analysis. BMJ 2014 Feb 13;348(feb13 1):g1151. [doi: 10.1136/bmj.g1151] [Medline: 24524926]
26. Tomioka H, Sekiya R, Nishio C, Ishimoto G. Impact of smoking cessation therapy on health-related quality of life. BMJ Open Respir Res 2014;1(1):e000047. [doi: 10.1136/bmjresp-2014-000047] [Medline: 25478191]
27. Shields M, Garner RE, Wilkins K. Dynamics of smoking cessation and health-related quality of life among Canadians. Health Rep 2013 Feb;24(2):3-11. [Medline: 24257905]
28. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
29. ICH E6 (R3) guideline for good clinical practice (GCP). EMA/CHMP/ICH/135/1995. European Medicines Agency. URL: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e6-r3-guideline-good-clinical-practice-gcp-step-5_en.pdf [accessed 2026-01-13]
30. Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol 1998 Jan;16(1):139-144. [doi: 10.1200/JCO.1998.16.1.139] [Medline: 9440735]
31. Koskinen S, Lundqvist A, Ristiluoma N. Terveys, toimintakyky ja hyvinvointi suomessa 2011 [Article in Finnish]. : THL; 2012 68/2012 URL: https://www.julkari.fi/handle/10024/90832 [accessed 2025-01-13]
32. Ohinmaa A, Sintonen H. Inconsistencies and modelling of the Finnish EuroQol (EQ-5D) preference values. : EuroQol Plenary Meeting, Health Economics and Health System Research, University of Hannover; 1998 Jan 10 URL: https://eq-5dpublications.euroqol.org/download?id=0_53500&fileId=53923 [accessed 2025-01-13]

XSL•FO
RenderX

33. Chung C, Lee JW, Lee SW, Jo MW. Clinical efficacy of mobile app-based, self-directed pulmonary rehabilitation for patients with chronic obstructive pulmonary disease: systematic review and meta-analysis. JMIR Mhealth Uhealth 2024 Jan 4;12:e41753. [doi: 10.2196/41753] [Medline: 38179689]
34. Öztürk ES, Kutlutürkan S. The effect of the mobile application-based symptom monitoring process on the symptom control and quality of life in breast cancer patients. Semin Oncol Nurs 2021 Jun;37(3):151161. [doi: 10.1016/j.soncn.2021.151161] [Medline: 34088557]
35. Thiengwittayaporn S, Wattanapreechanon P, Sakon P, et al. Development of a mobile application to improve exercise accuracy and quality of life in knee osteoarthritis patients: a randomized controlled trial. Arch Orthop Trauma Surg 2023 Feb;143(2):729-738. [doi: 10.1007/s00402-021-04149-8] [Medline: 34453570]
36. Arigayota A, Duffek B, Hou C, Eisingerich AB. Effects of The Legend of Zelda: Breath of the Wild and Studio Ghibli Films on young people's sense of exploration, calm, mastery and skill, purpose and meaning, and overall happiness in life: exploratory randomized controlled study. JMIR Serious Games 2025 Aug 1;13(1):e76522. [doi: 10.2196/76522] [Medline: 40750097]
37. Eich W, Diezemann-Prößdorf A, Hasenbring M, et al. Psychosocial factors in pain and pain management: a statement. Schmerz 2023 Jun;37(3):159-167. [doi: 10.1007/s00482-022-00633-1] [Medline: 35303149]
38. Auer R, Schoeni A, Humair JP, et al. Electronic nicotine-delivery systems for smoking cessation. N Engl J Med 2024 Feb 15;390(7):601-610. [doi: 10.1056/NEJMoa2308815] [Medline: 38354139]
39. Payakachat N, Ali MM, Tilford JM. Can the EQ-5D detect meaningful change? A systematic review. Pharmacoeconomics 2015 Nov;33(11):1137-1154. [doi: 10.1007/s40273-015-0295-6] [Medline: 26040242]

Abbreviations

CT: computed tomography
EORTC: European Organisation for Research and Treatment of Cancer
ES: effect size
GHS: global health status
HRQoL: health-related quality of life
LDCT: low-dose computed tomography
LDCT-SC-FI: low-dose computed tomography screening for lung cancer combined to different smoking cessation methods in Finland
NELSON: Nederlands-Leuvens Longkanker Screenings Onderzoek
OR: odds ratio
PROM: patient-reported outcome measure
QLQ-C30: Quality of Life Questionnaire Core 30
QLQ-LC13: Quality of Life Questionnaire Lung Cancer 13
QoL: quality of life

Edited by S Brini; submitted 01.Aug.2025; peer-reviewed by A Eisingerich, W Li; revised version received 07.Nov.2025; accepted 11.Nov.2025; published 20.Jan.2026.

Please cite as:
Kurtti A, Iivanainen S, Kaarteenaho R, Andersen H, Jekunen A, Vasankari T, Koivunen J
Effect of Lung Cancer Screening, Smoking Cessation, and Cessation Smartphone App to Health-Related Quality of Life Among Heavy Smokers: Randomized Controlled Trial
J Med Internet Res 2026;28:e81687
URL: https://www.jmir.org/2026/1/e81687
doi:10.2196/81687

© Antti Kurtti, Sanna Iivanainen, Riitta Kaarteenaho, Heidi Andersen, Antti Jekunen, Tuula Vasankari, Jussi Koivunen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Patient and Care Team Perspectives of Barriers to and Facilitators for the Implementation of a Digital Health Program for Depression in Primary Care: Qualitative Study

Andrea Nederveld[1], MPH, MD; Elise A Robertson[2,3], MA; Angela M Lanigan[2], MPA, RD, LD; Elisabeth F Callen[2,3], PhD; Tarin L Clay[2,3], BA; Ben Fehnert[4], MA; Lambros Chrones[5], MD; Michael L Martin[5], MBA, MD; Margaret McCue[5], MS, RD; Christina M Hester[2,3], MPH, PhD; Melissa K Filippi[2,6,7], MPH, PhD

[1]University of Colorado Anschutz Medical Campus, Aurora, CO, United States

[2]National Research Network, American Academy of Family Physicians, Leawood, KS, United States

[3]DARTNet Institute, 12635 E Montview Blvd Ste 129, Aurora, CO, United States

[4]Ctrl Group/Fora Health, London, United Kingdom

[5]Takeda Pharmaceuticals U.S.A., Inc., Lexington, MA, United States

[6]Robert Graham Center, American Academy of Family Physicians, Washington, DC, United States

[7]Department of Family Medicine, Georgetown University, Washington, DC, United States

Corresponding Author:
Christina M Hester, MPH, PhD
National Research Network, American Academy of Family Physicians, Leawood, KS, United States

Abstract

Background: Depression is pervasive, and rates are rising in the United States. Most people with depression receive care from primary care clinicians, but gaps in the quality of care exist. Team-based approaches to depression care have been shown to aid in treatment and management; yet, challenges exist in implementation. Digital health apps have been shown to be effective in improving depression symptoms and enhancing patient engagement in some populations. Many, however, do not share data with clinical care teams.

Objective: This study aimed to understand the barriers to and facilitators for implementation of a digital health program that supports coordinated use by clinical care teams and patients, via a mobile app and care team–facing web interface, for depression in primary care.

Methods: This study was part of a larger intervention study that included 4 primary care practices: 2 intervention and 2 control sites. The intervention sites used a patient-facing mobile app and a care team–facing web interface, and the control sites continued usual care. The study team conducted interviews from May to October 2021. Patient and care team participants were recruited toward the end of their study involvement. Separate semistructured interview guides were developed for patient and care team participants. Interviews were recorded and transcribed. Data were coded using Atlas.ti.9, and data analysis was completed using a grounded theory approach.

Results: Interviews with patient (n=8) and care team (n=8) participants revealed 3 main topics for program implementation: app/interface usability, tracking, and program recommendations. For app/interface usability, overall, navigation for both patient and care team participants was simple and straightforward. Although app content was relevant, patient participants desired additional educational resources and information to aid in their depression treatment and management. In terms of tracking, care team participants indicated that data obtained via the interface enabled them to monitor patients in between visits; and in some circumstances, these data facilitated conversations with patients about treatment plans. Tracking medication adherence differed among patient participants due to established routines of taking medications consistently, lack of motivation to track, or lack of interest in tracking. Care team participants reported the ability to respond more quickly to side effects. Patients commented on tracking difficulties: confusing response options, insufficient goal attainment response options, not being able to provide details or write notes, and no ability to change goals. Some patient and care team participants perceived that tracking encouraged communication with one another; others perceived tracking as having no impact on shared decision-making.

Conclusions: Results suggest implementation of a digital health program for depression treatment and management in primary care practices could impact patient medication adherence, produce faster turnaround time for medication optimization, encourage goal setting, and foster communication between patients and care team members. Program enhancements could optimize patient and care team member engagement.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e72003) doi:10.2196/72003

KEYWORDS

depression treatment; digital health; family physicians; goal setting; depression; major depressive disorder; MDD; mobile health; primary care; qualitative; shared decision-making

Introduction

Primary care clinical teams play a critical role in depression diagnosis, treatment, and management. Rates of depression in the United States have risen since the beginning of the COVID-19 pandemic [1]. Up to 70% of patients being treated for behavioral and mental health conditions, including depression, receive care within primary care [2], and approximately 13% of primary care visits involved a patient with a depression diagnosis between 2010 and 2018 [3]. Clinician comfort with treating depression varies [4], and gaps in quality of care exist; notably in areas such as screening, diagnosis, patient engagement and education, clinician training, and clinician follow-up [5-10]. There is evidence that team-based care and care management between clinic visits, as well as patient self-management [11], can improve depression outcomes [12,13]. However, these strategies can be time and resource intensive for primary care practices, requiring regular check-ins with patients and providing education on self-management. Encouraging patients to use digital health apps is one strategy to assist patients in better self-management.

Health care providers can play an integral role in patient initiation of use of a digital or mobile app [14]. Digital health apps are acceptable to patients, easy to use, and potentially effective in supporting behavior change and self-management, particularly in improving depression and anxiety symptoms in various populations [15-19]. However, there is a lack of standardization in evaluating health apps [20,21], and many behavioral health apps are not able to share data with clinical care teams [20].

This pilot study examined a digital health program that included a patient-facing mobile app and care team web interface that support depression treatment and symptom management in relation to shared decision-making and goal setting in primary care practices. Here, we describe patient and care team participant perspectives regarding barriers to and facilitators for using the digital health program in their practices.

Methods

Study Design, Setting, and Recruitment

This study was part of a larger intervention study (currently under review) that included 4 primary care practices, located in Michigan (n=1), Florida (n=1), and New York (n=2), all with some level of behavioral health integration. The intervention arm (n=2 practices, 1 residency and 1 private practice) piloted the use of the Primary Care Path, a program with a patient-facing mobile app and accompanying care team–facing web interface, which was developed collaboratively by Takeda and American Academy of Family Physicians National Research Network (AAFP NRN) and powered by Fora Health (Ctrl Group Limited UK). The control arm (n=2 practices; 1 Community Health Center and 1 private practice) continued usual care for depression. Participating practices were recruited from the AAFP NRN. Patient interview participants were recruited from the patients enrolled in the intervention arm via email invitation in the final weeks of their study involvement. Practice champion clinicians, study coordinators, and clinicians who cared for patients enrolled in the study were invited to participate postintervention.

Ethical Considerations

This study was approved by the American Academy of Family Physicians Institutional Review Board (Protocol 20 - 389). All participants consented to be in the study. However, those participants who took part in the intervention were sent an email asking if they would be willing to participate in an interview. Those who responded received an invitation for a scheduled interview. Prior to the start of the interview, a study team member thoroughly described participation: expected interview length, questions asked, permission to record, and participant rights (voluntary participation, can terminate participation at any time without penalty, remuneration, and possible risks). Risks, such as discomfort answering certain questions and the potential loss of anonymity and confidentiality, were explained. Verbal consent was then obtained. On completion of the interview, participants received compensation; a US $100 e-gift card was sent to their email. All identifying information was removed from interviews and study data to protect participant anonymity.

Intervention

The Fora Health app supports goal setting and tracks daily medication adherence and side effects, weekly goal attainment, and behavioral health participation; assesses the 9-item Patient Health Questionnaire (PHQ-9) [22,23], 13-item Patient Activation Measure [24,25], World Health Organization-Five Well-Being Index [26,27], and Perceived Deficits Questionnaire [28,29] on a biweekly basis; and provides short educational tutorials about depression management on topics such as talking with your physician and goal setting. Patient participants began using the Primary Care Path app (powered by Fora Health) after meeting with the study coordinator to enroll and set goals. The Primary Care Path web interface allows staff and clinicians to see patient-reported data. The expectation was that the care team would use the information before visits or review between visits.

Analysis

Separate semistructured interview guides were developed by the AAFP NRN study team for patient and care team participants, according to the RE-AIM (reach, effectiveness, adoption, implementation, and maintenance) evaluation framework [30-32]. The interview guides were reviewed and revised by the broader study team. Interviews were conducted from May to October 2021 by a family physician (AN) and 2 research project managers (AML and EAR), each with a

XSL•FO
RenderX

master's degree and training and experience in qualitative methods. Interviews were conducted online and lasted approximately 30 minutes, and were audio recorded and transcribed. Four research team members (MKF, AN, AML, and EAR) met frequently over 13 weeks to develop the codebook, using a priori and emergent codes. Meaning, some codes were generated based on the discussion guide, and others were developed through open coding, where codes emerge from the data. After the research team members agreed on the codes and their definitions, each transcript was coded once (MKF and EAR) using Atlas.ti.9 (Berlin and Germany), and an audit was conducted (AML) to ensure intercoder reliability. This inductive approach to data analysis was completed by 4 members of the research team (MKF, AN, AML, and EAR), using a combination of grounded theory and constant comparative methods [33,34]. Themes emerged through repeated reading of the data. Themes were discussed and agreed upon by the broader research team.

Results

Comparing Patient and Care Team Perspectives

The study team interviewed intervention patients (n=8) and intervention care team participants (lead clinicians, study coordinators, and other clinicians; n=8, shown in Tables 1 and 2). The sample size was determined based on the 2 participant categories of the sampling frame and the sample's sufficiency to provide a meaningful perspective and nuanced detail [35]. Data comparisons between the patient and care team participant strata showed similarities and differences for topics of app/interface usability (navigation and content and usability), tracking (medication adherence, side effects, goals, suggestions for additional tracking features, and communication), and program recommendations. Saturation was reached within the patient stratum as there was consistency across all patient data. Even though much of the data indicates overall agreement, additional explanations are provided where discordance is present. Example quotations are displayed at the end of each section.

Table . Patient demographics.

Patient participant	Sex	Race	Ethnicity	Year of birth	Employment status	Living situation	Treatments used for depression
1	Female	White	Non-Hispanic or Latinx	1997	Part time	With spouse or other family	Antidepressant medication and behavioral health/talk therapy
2	Female	White	Non-Hispanic or Latinx	1986	Full time	Alone	Antidepressant medication, behavioral health/talk therapy, and self-help strategies (books, apps, or other tools)
3	Female	White	Non-Hispanic or Latinx	1982	Part time	With spouse or other family	Antidepressant medication
4	Female	White	Non-Hispanic or Latinx	1988	Part time	With friends and/or roommates	Antidepressant medication
5	Female	White	Non-Hispanic or Latinx	1957	Part time	Alone	Self-help strategies (books, apps, or other tools)
6	Male	No response	No response	1964	Retired	With spouse or other family	Antidepressant medication and self-help strategies (books, apps, or other tools)
7	Male	White	Hispanic or Latinx	1969	Full time	With spouse or other family	Antidepressant medication
8	Male	White	Non-Hispanic or Latinx	1962	Part time	Alone	No response

XSL•FO
RenderX

Table . Care team demographics.

Care team participant	Sex	Race	Ethnicity	Birth year	Years in practice	Role/credential
1	Female	White	Non-Hispanic or Latinx	1993	1 - 5	BH[a] care team
2	Female	White	Non-Hispanic or Latinx	1970	16 - 20	MD[b]
3	Female	White	Non-Hispanic or Latinx	1984	1 - 5	BH care team, PhD[c]
4	Female	White	Non-Hispanic or Latinx	1991	1 - 5	DO[d]
5	Male	White	Non-Hispanic or Latinx	1984	6 - 10	MD
6	Male	Missing	Missing	Missing	Missing	Coordinator, Practice Manager
7	Male	White	Non-Hispanic or Latinx	1963	≥21	MD
8	Male	White	Non-Hispanic or Latinx	1989	6 - 10	APRN[e]

[a]BH: behavioral health.

[b]MD: doctor of medicine.

[c]PhD: doctor of philosophy.

[d]DO: doctor of osteopathic medicine.

[e]APRN: advanced practice registered nurse.

App/Interface Usability

Navigation

Overall, both patient and care team participants reported that the app/interface was simple and easy to use. Patient participants said entering data did not consume much time, making it practical to use in daily and weekly routines. They also liked the convenience of being able to enter data at any time. Care team participants stated that reporting, monitoring, viewing, and inputting information (eg, joint goals) in the interface was straightforward. Particularly, care team participants liked the red flag next to the patient ID for the suicidality alert for patients who responded positively to the PHQ-9 question 9, making it easy to identify which patients needed follow-up.

> *I think how it emailed us about the [suicidality] alerts helped. Of course, as you would log on to the app, you would see that there was an alert on a specific patient's chart and then being able to dismiss it once that concern was addressed.* [Care team participant]
>
> *Oh, it was definitely easy. It was very well designed, very clear. It never crashed. I never had an issue with it on my phone.* [Patient participant]

Content and Usability

In terms of app content, patient participants thought the app was a dependable source of information. However, patient participants overwhelmingly stated they desired additional content such as videos and guides. They said the app was static, repetitive, and compliance-driven, with too much focus on medication adherence. They reported a desire for personalized insights regarding their own depression and targeted information on interpreting their results over time, as well as content options that evolve as their treatment changes. They also suggested displaying how much time an action takes to complete (watching videos, reading guides, etc). Care team participants who spoke with their patients specifically about the app responded similarly to patient participant impressions—that content was appropriate, but they wanted more substance. Care team participants also suggested expanding resources for patients beyond depression care and management to include topics such as substance use disorders and additional information for care team members, such as education on prescribing medications and using goal-setting techniques with patients. Some care team members liked that the program tenets of shared decision-making and goal setting were incorporated in the app/interface.

Some care team participants stated that the data and insights were helpful as they could monitor patients between visits via the interface. They said that seeing patterns and trends helped or could help facilitate conversations, particularly if a change was needed, such as medication, therapy, or goals. The residency practice specifically used the data in daily huddles to better prepare for appointments. At this practice, a few point people regularly logged into the interface, but many care team participants only reviewed printed reports. For those who did use the interface, a feature they appreciated was the ability to view the level of patient involvement and their activities, allowing insight into how patient participants were using and applying the tenets of the program. Some care team participants stated that they believed the content helped patients pay attention and be more mindful of their depression care and management; yet, they were also unsure if the data they viewed really

XSL•FO
RenderX

contributed to better patient outcomes, treatment, or symptom management.

Care team participants appreciated seeing the results of the PHQ-9, which initiated a specific alert to a designated care team member if the patient responded positively to question 9 (suicidal ideation), particularly if this was not a concern at the previous visit. However, they thought patient outreach could be more efficient. For example, PHQ-9 information was not in the medical chart, so they had to create a note and send it to other care team members. Even though care team participants thought the PHQ-9 alerts were appropriate, they were concerned about patient truthfulness, thinking patient participants may not want to be contacted about their answers in the future, which could impact how they answered questions.

> *My only thing that I couldn't figure out how to fix was every once in a while, there would be the small four-minute guides about setting goals with your caregiver, those ones. The last month and a half I was using it, it prompted me with the same guide every two days for a month and a half.* [Patient participant]

> *When it comes to features, about the only thing that I think would be useful is if there were any relevant articles, studies, videos or anything like that, that could be useful, that could be recommended or something like that. Just different ways to engage the user and to put more resources, basically.* [Patient participant]

Tracking

Medication Adherence

Patient participants who had never taken medication before stated that the daily medication question was particularly helpful, reporting that the daily check-in helped them think about goals and well-being more frequently. They reported being more mindful of their depression treatment and care throughout the week. However, for patient participants who already had a routine for taking medication, the reminders were not particularly helpful. Care team participants reiterated that for patients who need help establishing routines, reminders seemed to help with medication adherence.

> *I think it's really helpful if you're not good at taking your medication. I think that that is the number—I take it every night before bed, so I'm pretty strict in that routine. I can see that if it's something you're not used to doing, it being really helpful for that. It was really simple to use. If there were older users that weren't comfortable with technology, it would be really easy to navigate for them.* [Patient participant]

Overall, patient participants liked the option for tracking medication use, as it showed day-to-day adherence. Some patient participants viewed even small changes in their depression symptoms as motivation to continue taking medication. However, some patient participants did not track their medication use, or they only tracked it for a short period of time because they either already took medications consistently, lacked motivation to track, or were not interested in tracking. Patient and care team participants thought the act of tracking medication use contributed to a sense of accountability in depression treatment and care. However, patient participants mentioned that when they switched medications or changed dosage, it was difficult to account for those changes in the app. They also wanted questions streamlined to save time and avoid repetition. Finally, some care team participants questioned whether this approach changed medication adherence.

> *I think the biggest way it [tracking] probably affected their treatment was just actually (a) holding them a bit more accountable for what they were trying to do on their end, but (b) just seeing them check in daily with themselves or however often they were doing it...I think being able to be a little more retrospective more frequently helped them gain more insight too into how they were doing.* [Care team participant]

> *No, it [the program] didn't look like it [changed medication adherence]. It looked like we saw the same things. Some people were consistently taking their meds, some people who had some side effects stopped without telling us and would show up for the next visit. That didn't look like it changed much.* [Care team participant]

Side Effects

Tracking side effects was reported to be particularly helpful as patient participants often were not confident in differentiating between medication side effects and depression-related symptoms. For some patient participants, tracking side effects led to new discussions with their care team or therapist. Some care team participants stated they responded or could respond more quickly to side effects because patients were tracking them in the app, stating this was the most useful data collected because of in-depth, near real-time information, which allowed for timely adjustment of medications. Furthermore, in the residency practice, the data were used to help resident physicians make care decisions.

> *They really like the fact that if they were experiencing side effects that they could note it in the app, and their care team would be able to see it sooner before their next visit. It also allowed us to be aware of what's going on and make necessary changes if we needed to, if the side effects were that uncomfortable, or just to see if their PHQ was increasing that a change needed to be made.* [Care team participant]

> *It'd make you cognizant of the fact that it's something that you need to monitor, need to stay on top of. I think what it really helps is when it comes—'cause it makes you think about any side effects you my have had, and you may be having side effects but not necessarily attribute 'em to what you're takin' medication for, so it's a way of makes you think, and it's like, you know what? I did have this today, or I did go do that. Did that have something to do with it? At very least, it makes you think about.* [Patient participant]

XSL•FO
RenderX

Goals

In general, both patient and care team participants found value in setting and tracking goals. Patient participants stated it was helpful to set simple and reachable goals, and they could apply goal setting to other areas in their depression self-management beyond medication adherence (eg, exercising, spending time with family, reading, etc). In addition, patient participants appreciated seeing their progress displayed concretely in the app. Care team participants also thought that the tracking helped facilitate patient participants' depression self-management. Some care team participants reported that they had used goal setting as part of their normal care prior to study involvement, but the app/interface further facilitated the process of setting and tracking goals. They perceived the program, via the app/interface, as providing structure and consistency to ensure that goal setting was incorporated into discussions with patients. One care team participant stated that when patient participants used the app, it made conversations about setting goals easier, and it helped patient and care team participants to be on the same page.

Patient and care team participants also noted some hurdles with the goal-setting features. Patient participants discussed the difficulty of tracking goals within the app, including confusing responses (eg, goal benchmarks that did not make sense due to care team user error) or insufficient goal attainment responses. They wanted the ability to modify goals or remove outdated goals. Some patient participants reported feeling worse when reminded of a goal they were not meeting or that was no longer aligned with their current plan. Care team participants agreed that patient participants had difficulty managing goals within the app regarding frequency, milestones, and attainment selection.

> *Like I said, my favorite part of it is that you guys have those goals on there, which is typically something that I really like to do with patients and engaging them in shared decision-making, problem-solving with them, and goal setting to help accomplish their goals.* [Care team participant]

> *Yeah, one of our patient's weight was a big thing tied to her depression. She felt like the more she gained weight, the more depressed she felt. One of the things that really helped her was making goals in terms of exercising. She said having the goal of exercise by walking outside three times a week definitely really helped her become more engaged. She did lose, I believe, about 10 pounds, 15 pounds, something around there, which also improved her mood as well 'cause she was seeing that she was making these different changes.* [Care team participant]

> *That [questions about goals] was a little bit difficult to answer because my goal was to read twice a week for 15 minutes. The language of the app was, "Did you complete it once or did you complete it twice?" I wasn't sure because I was doing it twice already, should I say I did it once or should I say I did it twice? Sometimes just due to the language of the app, it didn't fit exactly what I was trying to do. Maybe a better way to see your progress with goals, rather than just checking off that you did it, maybe something like that.* [Patient participant]

Suggestions for Additional Tracking Features

Patient participants suggested streamlining questions to save time and avoid repetition. They also wanted to add some additional tracking features (eg, other self-management behavior that did not involve medication, such as physical activity, emotions, or well-being, with a place to write notes). Some desired a way to record more details on their progress. Care team participants said it was difficult to interpret the accuracy of the log (eg, medication compliance) and incomplete information (eg, missing data could indicate noncompliance or forgetfulness).

Communication

Patient and care team participants expressed that tracking encouraged or had the potential to encourage communication with one another. Patient participants liked the possibility of checking in with care team participants, especially without an appointment. Some felt that inputting data prior to appointments made the time together more efficient. Some care team participants also thought the appointments were easier and more effective when patients had tracked medication and goals. Care team participants stated that tracking enabled them to reference data before or during visits, highlight potential concerns, reassure alignment with the patient, and use information in follow-up calls. Some patient participants thought that the app had no impact on shared decision-making for depression treatment and management. Other patient participants stated it had made a difference, although some were unsure if a care team member viewed their information beforehand. Some patient and care team participants thought tracking impacted their relationship with one another; tracking reinforced feelings that the care team was invested in patient well-being, and the patient was invested in their own care. Care team participants thought that patient participants were more involved and proactive in their own care, and some believed patient engagement improved in the process.

> *I don't know that it cured the symptoms or anything like that. I just think that it was helpful and that I can keep that communication going with my team, so they knew what was going on.* [Patient participant]

> *This provided data points in between visits, especially while we're titrating medications because often we're starting a new [medication], changing something, changing a dosing, and this allowed communication points in between those visits or ways to track that data outside of these. We often will have them come back in six weeks after a medication change has been made. This can allow for interim information to help making a decision moving forward too.* [Care team participant]

App Use Recommendations

Patient and care team participants believed using the app could benefit several groups, namely those who are new to taking or switching medications, have trouble remembering to take

XSL•FO
RenderX

medications, are proactive in their care, want to make a behavioral change, need reminders, need accountability, and are seeking increased adherence. Some care team participants recommended targeting patients who needed closer follow-up and thought the app was more beneficial for those who were more diligent about their depression care and management, as the increased engagement gave them better insight into their health. Some care team participants indicated that using the app interface did not impact or change setting goals, reaching goals, making shared decisions, or patient outcomes. Reasons provided were either that their practice had implemented a patient-centered approach previously or that they needed data to compare before and after study participation to make a true assessment.

Care team participants thought the workflow around the interface could be improved. They overwhelmingly stated the need for electronic health record (EHR) integration because logging into the interface interrupted workflow, making it impractical during a patient appointment. Some care team participants admitted they did not access patient participant information: logging in was not well integrated into their workflow, they did not remember to log in, or they did not know how to log in. For example, one care team participant in the residency practice commented that it would have been helpful for the attending physician to see the data, and one patient participant stated that they wanted to share data with a psychiatrist outside of their primary care practice. Furthermore, some care team participants spoke about the need for app support within the practice. One care team participant stated that the practice had a dedicated behavioral health care specialist who carried out nonmedication interventions, and much of the program fell under their purview (eg, providing information to physicians and following up with patients). It was stressed that behavioral health is a team effort, and office staff are needed for integration.

Overall, patient participants who tracked and care team participants who used data to inform conversations saw value in shared decision-making and setting goals and thought these tenets could be applied to other conditions, such as anxiety, weight management, diabetes, high blood pressure, attention deficit disorder or attention-deficit/hyperactivity disorder, and medication compliance.

> *It really is about the EMR [electronic medical record], just because I'm spending so much time on it every day. It's always gonna be hard for me to have to login into something else outside of that when everything else is integrated.* [Care team participant]

Discussion

Principal Findings

Through this study, we identified barriers to and facilitators for implementation of a pilot program that incorporates a patient-facing mobile app and care team web interface to support depression treatment and management in primary care practices. These results on patient and care team participant perspectives of program implementation contribute to the rapidly expanding field of the use of digital health for depression care [17,36-45]. While some results between patient and care team participant strata were rather uniform, others differed. Our findings indicate that tracking medication adherence, side effects, and goals could allow both patient and care team participants greater insight into depression treatment and management. However, though the app may have helped some patient participants compile information for tracking, which may have led to more focused visits, not all care team participants thought app use improved patient outcomes. This may be more a matter of perception—perhaps patient participants who engaged regularly with the app thought they had better outcomes, possibly related to the perception of having more connectivity with their care team. Care team participants reported they could not definitively determine whether the app and program impacted patient outcomes, patient relationships, shared decision-making, goal setting, or goal attainment without tangible pre- and postintervention data. Nonetheless, medication adherence and goal setting are often related to depression outcomes and can be challenging [46-48]; this program could be another tool in their toolbox. Practices that already have patient-centered approaches to care may find aspects of this program duplicative; yet practices that do not prioritize approaches to goal setting and shared decision-making may experience more benefits if implementing this program, or a similar one.

Data and Action

App data in and of itself may not impact patient outcomes; it may be that the increased follow-up and touchpoints promote improvement in depression symptoms, as is evident in the literature [49,50]. The program was designed to connect patient participant data to the care team, allowing the care team to get real-time insight into patient symptoms and functioning between appointments. Thus, care team members who receive and review app data may be better able to recommend follow-up, suggest care changes, inform conversations, and shape next steps versus those who do not use additional data points collected outside of a care visit. The novel contribution of this study is that care team members who viewed near real-time information had an opportunity to make more timely decisions, especially for medication changes. Therefore, data insights provided to the care team via the interface may allow treatment pivots based on more accurate data on what is or is not working. This finding suggests that those care team members who accessed patient information in near real-time could make more prompt evidence-based decisions, which is particularly important for a patient with depression symptoms that impair day-to-day functioning.

Program Implementation

Even though care team participants appreciated aspects of the program, they agreed that successful implementation would need to be more integrated into existing workflows. It is unclear to what extent care team participants, especially physicians, accessed data prior to and during visits, and they often said that accessing a separate website was too time-consuming during a busy clinical day. This is consistent with other literature that demonstrates that app data integration into clinical workflows is often a requirement for use due to heavy workloads [51-55]. EHR integration would allow all care team members access, a

XSL•FO
RenderX

seamless transition for viewing data, and a comprehensive view of patient data. For example, if a patient scheduled an appointment for something unrelated to depression, a care team member would not necessarily review their depression data on a separate website. It is well documented that EHRs can help clinicians prepare for appointments, making them more efficient [56,57].

For uptake of this program, other workflow processes may need to be established depending on the practice, such as the level of behavioral health integration or resource support (eg, access to behavioral health professionals or referrals to community resources) [58-61]. Another consideration is designing the best approach for using data in conversations with patients [17]. Future iterations of this program could include additional education for care team members regarding roles, recommendations, and patient interactions.

Limitations

This study has limitations. First, our study sample was small and was racially and ethnically homogeneous, limiting the generalizability of the findings. Second, data saturation was reached in the patient stratum but not the care team member stratum. The care team contained diverse roles. We interviewed 8 primary care physicians and clinicians; their experiences varied. For example, a behavioral health professional or care coordinator may have had a touchpoint with the patient, and physicians may or may not have used the interface. Therefore, it is inconclusive as to whether physicians accessed and used the data before or during visits. To help remedy this situation, care team participants recommended embedding the interface data into the EHR. Third, to some extent, these data were influenced by patient and care team member engagement with the program. For example, if a patient used the app for a limited time span, they may not have tracked their behaviors long enough to see trends or for care team participants to make recommendations. Finally, a patient's depression cycle may have affected app use and program involvement. If depression symptoms were more prominent, patient participants may not have had the motivation to track medication adherence, side effects, or goals.

Conclusions

Implementation of a digital health program for depression treatment and management in primary care practices may help support patient medication adherence, facilitate timely medication changes, encourage goal setting, and foster communication between patients and care team members. While patient and care team participants valued program tenets, enhancements such as minimizing workflow interruptions, integrating data into the EHR, providing education and best practices for patient interactions, augmenting content with helpful resources, and adding personalized content that evolves alongside patient progression could increase program engagement.

Acknowledgments

The authors would like to acknowledge the collaboration of Takeda Pharmaceutical Company Limited with the primary care path developer Ctrl Group/Fora Health. They also want to acknowledge Takeda Pharmaceutical Company Limited for their collaborative study partnership, study design, and protocol development. The authors would also like to thank the practice team and patient participants who made this study possible. Generative artificial intelligence was not used in generating the study design, data collection, data analysis, or manuscript content.

Funding

Supported by funding from The Takeda Pharmaceutical Company Limited. This information or content and conclusions are those of the authors independent of the funder.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: LC, MM, CMH, MKF
Formal analysis: AN, EAR, AML, MKF
Funding acquisition: CMH
Investigation: EFC, TLC
Methodology: EFC, CMH, MKF
Project administration: AML
Supervision: CMH
Writing – original draft: AN, AML, MKF
Writing – review & editing: AN, EAR, EFC, TLC, BF, LC, MLM, MM, CMH, MKF

Conflicts of Interest

MM, MLM, and LC are or were employees of Takeda Pharmaceuticals at the time of the study. BF is an employee of Ctrl Group.

XSL•FO
RenderX

References

1. Ettman CK, Abdalla SM, Cohen GH, Sampson L, Vivier PM, Galea S. Prevalence of depression symptoms in US adults before and during the COVID-19 pandemic. JAMA Netw Open 2020 Sep 1;3(9):e2019686. [doi: 10.1001/jamanetworkopen.2020.19686] [Medline: 32876685]
2. Hunter CL, Goodie JL, Oordt MS, Dobmeyer AC. Integrated Behavioral Health in Primary Care: Step-by-Step Guidance for Assessment and Intervention: American Psychological Association; 2009. [doi: 10.1037/11871-000]
3. Jackson JL, Kuriyama A, Bernstein J, Demchuk C. Depression in primary care, 2010-2018. Am J Med 2022 Dec;135(12):1505-1508. [doi: 10.1016/j.amjmed.2022.06.022] [Medline: 35878693]
4. Chang ET, Magnabosco JL, Chaney E, et al. Predictors of primary care management of depression in the Veterans Affairs healthcare system. J Gen Intern Med 2014 Jul;29(7):1017-1025. [doi: 10.1007/s11606-014-2807-z] [Medline: 24567200]
5. Chew-Graham CA, Mullin S, May CR, Hedley S, Cole H. Managing depression in primary care: another example of the inverse care law? Fam Pract 2002 Dec;19(6):632-637. [doi: 10.1093/fampra/19.6.632] [Medline: 12429666]
6. Egede LE. Failure to recognize depression in primary care: issues and challenges. J Gen Intern Med 2007 May;22(5):701-703. [doi: 10.1007/s11606-007-0170-z] [Medline: 17370030]
7. Blackstone SR, Sebring AN, Allen C, Tan JS, Compton R. Improving depression screening in primary care: a quality improvement initiative. J Community Health 2022 Jun;47(3):400-407. [doi: 10.1007/s10900-022-01068-6] [Medline: 35076803]
8. Garcia ME, Hinton L, Neuhaus J, Feldman M, Livaudais-Toman J, Karliner LS. Equitability of depression screening after implementation of general adult screening in primary care. JAMA Netw Open 2022 Aug 1;5(8):e2227658. [doi: 10.1001/jamanetworkopen.2022.27658] [Medline: 35980633]
9. Kyanko KA, A Curry L, E Keene D, Sutherland R, Naik K, Busch SH. Does primary care fill the gap in access to specialty mental health care? a mixed methods study. J Gen Intern Med 2022 May;37(7):1641-1647. [doi: 10.1007/s11606-021-07260-z] [Medline: 34993864]
10. Samples H, Stuart EA, Saloner B, Barry CL, Mojtabai R. The role of screening in depression diagnosis and treatment in a representative sample of US primary care visits. J Gen Intern Med 2020 Jan;35(1):12-20. [doi: 10.1007/s11606-019-05192-3] [Medline: 31388917]
11. Dineen-Griffin S, Garcia-Cardenas V, Williams K, Benrimoj SI. Helping patients help themselves: a systematic review of self-management support strategies in primary health care practice. PLoS ONE 2019;14(8):e0220116. [doi: 10.1371/journal.pone.0220116] [Medline: 31369582]
12. Kappelin C, Carlsson AC, Wachtler C. Specific content for collaborative care: a systematic review of collaborative care interventions for patients with multimorbidity involving depression and/or anxiety in primary care. Fam Pract 2022 Jul 19;39(4):725-734. [doi: 10.1093/fampra/cmab079] [Medline: 34546354]
13. Menear M, Duhoux A, Roberge P, Fournier L. Primary care practice characteristics associated with the quality of care received by patients with depression and comorbid chronic conditions. Gen Hosp Psychiatry 2014;36(3):302-309. [doi: 10.1016/j.genhosppsych.2014.01.013] [Medline: 24629824]
14. Pung A, Fletcher SL, Gunn JM. Mobile app use by primary care patients to manage their depressive symptoms: qualitative study. J Med Internet Res 2018 Sep 27;20(9):e10035. [doi: 10.2196/10035] [Medline: 30262449]
15. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019 Jul 22;21(7):e12869. [doi: 10.2196/12869] [Medline: 31333198]
16. Wang K, Varma DS, Prosperi M. A systematic review of the effectiveness of mobile apps for monitoring and management of mental health symptoms or disorders. J Psychiatr Res 2018 Dec;107:73-78. [doi: 10.1016/j.jpsychires.2018.10.006] [Medline: 30347316]
17. McCue M, Blair C, Fehnert B, et al. Mobile app to enhance patient activation and patient-provider communication in major depressive disorder management: collaborative, randomized controlled pilot study. JMIR Form Res 2022 Oct 27;6(10):e34923. [doi: 10.2196/34923] [Medline: 36301599]
18. Rathbone AL, Prescott J. The use of mobile apps and SMS messaging as physical and mental health interventions: systematic review. J Med Internet Res 2017 Aug 24;19(8):e295. [doi: 10.2196/jmir.7740] [Medline: 28838887]
19. Milne-Ives M, Lam C, De Cock C, Van Velthoven MH, Meinert E. Mobile apps for health behavior change in physical activity, diet, drug and alcohol use, and mental health: systematic review. JMIR Mhealth Uhealth 2020 Mar 18;8(3):e17046. [doi: 10.2196/17046] [Medline: 32186518]
20. Carlo AD, Hosseini Ghomi R, Renn BN, Areán PA. By the numbers: ratings and utilization of behavioral health mobile applications. NPJ Digit Med 2019;2:54. [doi: 10.1038/s41746-019-0129-6] [Medline: 31304400]
21. Torous J, Andersson G, Bertagnoli A, et al. Towards a consensus around standards for smartphone apps and digital mental health. World Psychiatry 2019 Feb;18(1):97-98. [doi: 10.1002/wps.20592] [Medline: 30600619]
22. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
23. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatr Ann 2002 Sep;32(9):509-515. [doi: 10.3928/0048-5713-20020901-06]

XSL•FO
RenderX

24. Hibbard JH, Mahoney ER, Stockard J, Tusler M. Development and testing of a short form of the patient activation measure. Health Serv Res 2005 Dec;40(6 Pt 1):1918-1930. [doi: 10.1111/j.1475-6773.2005.00438.x] [Medline: 16336556]
25. Patient Activation Measure® (PAM®): learn more about the leading assessment of patient activation. Insignia Health. 2021. URL: https://www.insigniahealth.com/pam/ [accessed 2026-01-07]
26. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167-176. [doi: 10.1159/000376585] [Medline: 25831962]
27. Wellbeing measures in primary health care/the depcare project. : World Health Organization Regional Office for Europe; 1998 URL: https://iris.who.int/server/api/core/bitstreams/8af98105-30ec-4da4-9fe9-097f7459e6da/content [accessed 2026-01-07]
28. Sullivan MJ, Edgley K, Dehoux E. A survey of multiple sclerosis: I. Perceived cognitive problems and compensatory strategy use. Can J Rehabil 1990;4(2):99-105 [FREE Full text]
29. Lam RW, Lamy FX, Danchenko N, et al. Psychometric validation of the Perceived Deficits Questionnaire-Depression (PDQ-D) instrument in US and UK respondents with major depressive disorder. Neuropsychiatr Dis Treat 2018;14:2861-2877. [doi: 10.2147/NDT.S175188] [Medline: 30464471]
30. Klesges LM, Estabrooks PA, Dzewaltowski DA, Bull SS, Glasgow RE. Beginning with the application in mind: designing and planning health behavior change interventions to enhance dissemination. Ann Behav Med 2005 Apr;29 Suppl:66-75. [doi: 10.1207/s15324796abm2902s_10] [Medline: 15921491]
31. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
32. Gaglio B, Shoup JA, Glasgow RE. The RE-AIM framework: a systematic review of use over time. Am J Public Health 2013 Jun;103(6):e38-e46. [doi: 10.2105/AJPH.2013.301299] [Medline: 23597377]
33. Singer M, Baer H. Introducing Medical Anthropology: A Discipline in Action, 2nd edition: Lanham: AltaMira Press; 2012. URL: https://search.worldcat.org/title/introducing-medical-anthropology-a-discipline-in-action/oclc/800937988 [accessed 2026-01-07]
34. Patton MQ. Qualitative Evaluation and Research Methods, 2nd edition: Sage Publications; 1990. URL: https://openlibrary.org/books/OL2202510M/Qualitative_evaluation_and_research_methods [accessed 2026-01-07]
35. Corbin JM, Strauss AL. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, 4th edition: SAGE Publications; 2014. URL: https://collegepublishing.sagepub.com/products/basics-of-qualitative-research-4-235578 [accessed 2026-01-07]
36. Serrano-Ripoll MJ, Zamanillo-Campos R, Fiol-DeRoque MA, Castro A, Ricci-Cabello I. Impact of smartphone app-based psychological interventions for reducing depressive symptoms in people with depression: systematic literature review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2022 Jan 27;10(1):e29621. [doi: 10.2196/29621] [Medline: 35084346]
37. Graham AK, Greene CJ, Kwasny MJ, et al. Coached mobile app platform for the treatment of depression and anxiety among primary care patients: a randomized clinical trial. JAMA Psychiatry 2020 Sep 1;77(9):906-914. [doi: 10.1001/jamapsychiatry.2020.1011] [Medline: 32432695]
38. Leong QY, Sridhar S, Blasiak A, et al. Characteristics of mobile health platforms for depression and anxiety: content analysis through a systematic review of the literature and systematic search of two app stores. J Med Internet Res 2022 Feb 4;24(2):e27388. [doi: 10.2196/27388] [Medline: 35119370]
39. McCloud T, Jones R, Lewis G, Bell V, Tsakanikos E. Effectiveness of a mobile app intervention for anxiety and depression symptoms in university students: randomized controlled trial. JMIR Mhealth Uhealth 2020 Jul 31;8(7):e15418. [doi: 10.2196/15418] [Medline: 32735221]
40. Deady M, Glozier N, Calvo R, et al. Preventing depression using a smartphone app: a randomized controlled trial. Psychol Med 2022 Feb;52(3):457-466. [doi: 10.1017/S0033291720002081] [Medline: 32624013]
41. Broglia E, Millings A, Barkham M. Counseling with guided use of a mobile well-being app for students experiencing anxiety or depression: clinical outcomes of a feasibility trial embedded in a student counseling service. JMIR Mhealth Uhealth 2019 Aug 15;7(8):e14318. [doi: 10.2196/14318] [Medline: 31418424]
42. Teepe GW, Da Fonseca A, Kleim B, et al. Just-in-time adaptive mechanisms of popular mobile apps for individuals with depression: systematic app search and literature review. J Med Internet Res 2021 Sep 28;23(9):e29412. [doi: 10.2196/29412] [Medline: 34309569]
43. Myers A, Chesebrough L, Hu R, Turchioe MR, Pathak J, Creber RM. Evaluating commercially available mobile apps for depression self-management. AMIA Annu Symp Proc 2020;2020:906-914. [Medline: 33936466]
44. Lu SC, Xu M, Wang M, et al. Effectiveness and minimum effective dose of app-based mobile health interventions for anxiety and depression symptom reduction: systematic review and meta-analysis. JMIR Ment Health 2022 Sep 7;9(9):e39454. [doi: 10.2196/39454] [Medline: 36069841]
45. Zhang R, Nicholas J, Knapp AA, et al. Clinically meaningful use of mental health apps and its effects on depression: mixed methods study. J Med Internet Res 2019 Dec 20;21(12):e15644. [doi: 10.2196/15644] [Medline: 31859682]

XSL•FO
RenderX

46. Bosworth HB, Voils CI, Potter GG, Steffens DC. The effects of antidepressant medication adherence as well as psychosocial and clinical factors on depression outcome among older adults. Int J Geriatr Psychiatry 2008 Feb;23(2):129-134. [doi: 10.1002/gps.1852] [Medline: 17563920]
47. Coote HMJ, MacLeod AK. A self-help, positive goal-focused intervention to increase well-being in people with depression. Clin Psychol Psychother 2012;19(4):305-315. [doi: 10.1002/cpp.1797] [Medline: 22610936]
48. Jacob J, Stankovic M, Spuerck I, Shokraneh F. Goal setting with young people for anxiety and depression: what works for whom in therapeutic relationships? A literature review and insight analysis. BMC Psychol 2022 Jul 13;10(1):171. [doi: 10.1186/s40359-022-00879-5] [Medline: 35831897]
49. Simon GE, Ralston JD, Savarino J, Pabiniak C, Wentzel C, Operskalski BH. Randomized trial of depression follow-up care by online messaging. J Gen Intern Med 2011 Jul;26(7):698-704. [doi: 10.1007/s11606-011-1679-8] [Medline: 21384219]
50. Solberg LI, Trangle MA, Wineman AP. Follow-up and follow-through of depressed patients in primary care: the critical missing components of quality care. J Am Board Fam Pract 2005;18(6):520-527. [doi: 10.3122/jabfm.18.6.520] [Medline: 16322414]
51. Huckvale K, Nicholas J, Torous J, Larsen ME. Smartphone apps for the treatment of mental health conditions: status and considerations. Curr Opin Psychol 2020 Dec;36(65-70):65-70. [doi: 10.1016/j.copsyc.2020.04.008] [Medline: 32553848]
52. Dinkel D, Harsh Caspari J, Fok L, et al. A qualitative exploration of the feasibility of incorporating depression apps into integrated primary care clinics. Transl Behav Med 2021 Sep 15;11(9):1708-1716. [doi: 10.1093/tbm/ibab075] [Medline: 34231855]
53. Cohen DJ, Keller SR, Hayes GR, Dorr DA, Ash JS, Sittig DF. Integrating patient-generated health data into clinical care settings or clinical decision-making: lessons learned from project HealthDesign. JMIR Hum Factors 2016 Oct 19;3(2):e26. [doi: 10.2196/humanfactors.5919] [Medline: 27760726]
54. Ye J. The impact of electronic health record-integrated patient-generated health data on clinician burnout. J Am Med Inform Assoc 2021 Apr 23;28(5):1051-1056. [doi: 10.1093/jamia/ocab017] [Medline: 33822095]
55. Gordon WJ, Landman A, Zhang H, Bates DW. Beyond validation: getting health apps into clinical practice. NPJ Digit Med 2020;3(14):14. [doi: 10.1038/s41746-019-0212-z] [Medline: 32047860]
56. O'Malley AS, Draper K, Gourevitch R, Cross DA, Scholle SH. Electronic health records and support for primary care teamwork. J Am Med Inform Assoc 2015 Mar;22(2):426-434. [doi: 10.1093/jamia/ocu029] [Medline: 25627278]
57. Sinsky CA, Sinsky TA, Rajcevich E. Putting pre-visit planning into practice. Fam Pract Manag 2015;22(6):34-38. [Medline: 26761083]
58. Moon K, Sobolev M, Kane JM. Digital and mobile health technology in collaborative behavioral health care: scoping review. JMIR Ment Health 2022 Feb 16;9(2):e30810. [doi: 10.2196/30810] [Medline: 35171105]
59. Moon KC, Sobolev M, Grella M, Alvarado G, Sapra M, Ball T. An mHealth platform for augmenting behavioral health in primary care: longitudinal feasibility study. JMIR Form Res 2022 Jul 1;6(7):e36021. [doi: 10.2196/36021] [Medline: 35776491]
60. Hoffman L, Benedetto E, Huang H, et al. Augmenting mental health in primary care: a 1-year study of deploying smartphone apps in a multi-site primary care/behavioral health integration program. Front Psychiatry 2019;10:94. [doi: 10.3389/fpsyt.2019.00094] [Medline: 30873053]
61. Carleton KE, Patel UB, Stein D, Mou D, Mallow A, Blackmore MA. Enhancing the scalability of the collaborative care model for depression using mobile technology. Transl Behav Med 2020 Aug 7;10(3):573-579. [doi: 10.1093/tbm/ibz146] [Medline: 32766866]

Abbreviations

AAFP NRN: American Academy of Family Physicians National Research Network
EHR: electronic health record
PHQ-9: 9-item Patient Health Questionnaire
RE-AIM: reach, effectiveness, adoption, implementation, and maintenance

Edited by A Stone; submitted 31.Jan.2025; peer-reviewed by AJ Saxon, HV Marwijk; accepted 15.Dec.2025; published 29.Jan.2026.

Please cite as:
Nederveld A, Robertson EA, Lanigan AM, Callen EF, Clay TL, Fehnert B, Chrones L, Martin ML, McCue M, Hester CM, Filippi MK Patient and Care Team Perspectives of Barriers to and Facilitators for the Implementation of a Digital Health Program for Depression in Primary Care: Qualitative Study
J Med Internet Res 2026;28:e72003
URL: https://www.jmir.org/2026/1/e72003
doi:10.2196/72003

XSL•FO
RenderX

© Andrea Nederveld, Elise A Robertson, Angela M Lanigan, Elisabeth F Callen, Tarin L Clay, Ben Fehnert, Lambros Chrones, Michael L Martin, Margaret McCue, Christina M Hester, Melissa K Filippi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Internet Health Care Service Use Behavioral Pattern Among Older Adults and the Role of the Technology Acceptance and Social Ecological Theory Model: Cross-Sectional Survey

Rui Li[1*], MPH; Xinyu Xu[2,3*], MPH; Qingsong Li[2,3*], MPH; Haobiao Liu[2,3], PhD; Ting Ting Zhou[1], MPH; Abebe Feyissa Amhare[2,3], PhD; Peiyu Liu[2,3], MPH; Jing Tang[2,3], MPH; Wei Wang[4], MPH; Fuju Zheng[4], PhD; Jing Han[3,5], PhD

[1]Shandong Center for Disease Control and Prevention, Ji'nan, Shandong, China

[2]Key Laboratory of Environment and Genes Related to Disease, School of Public Health, Health Science Center, Xi'an Jiaotong University, Xi'an, Shaanxi, China

[3]Department of Occupational and Environmental Health, School of Public Health, Health Science Center, Xi'an Jiaotong University, No. 76, Yanta West Road, Xi'an, Shaanxi, China

[4]Department of General Dentistry, Jinan Stomatological Hospital, Jinan, China

[5]Department of Health Science Center Global Health Institute, Health Science Center, Xi'an Jiaotong University, Xi'an, Shaanxi, China

[*]these authors contributed equally

Corresponding Author:
Jing Han, PhD
Department of Occupational and Environmental Health, School of Public Health, Health Science Center, Xi'an Jiaotong University, No. 76, Yanta West Road, Xi'an, Shaanxi, China

Abstract

Background: The rapid growth of internet health care (IH) offers older adults convenient medical services like remote consultations and health monitoring. However, its adoption among this group remains low, highlighting a significant digital divide. Understanding the behavioral patterns and determinants of IH use in the older population is crucial for optimizing digital health design and improving service accessibility.

Objective: This study aimed to analyze the multidimensional influencing factors of Chinese older adults' use of IH services based on the integrated framework of the technology acceptance model and social ecological model, and explore their behavioral patterns and key driving factors.

Methods: A cross-sectional study design was adopted to conduct a multistage stratified cluster random sampling survey in 3 cities in Shandong Province from May 2024 to July 2024, with a total of 1828 older adults aged 60 to 75 years included. The study uses latent category analysis to classify the use of IH service behaviors and employs multiple logistic regression, decision tree models, and structural equation modeling to analyze influencing factors and mediating pathways.

Results: Five distinct user groups were identified: nonusers (n=911), registration-dominant users (n=286), low-activity users (n=320), moderate comprehensive users (n=288), and full-service users (n=23). Multinomial logistic regression with nonusers as the reference group identified key determinants: individuals with below primary education had 96% lower odds of membership (odds ratios [OR] 0.039, 95% CI 0.012 - 0.084) compared to the reference group with junior college education or above in moderate comprehensive users, while male participants had higher odds of being full-service (OR 1.980, 95% CI 1.126 - 3.514) or moderate comprehensive (OR 1.310, 95% CI 1.012 - 1.705) users. Older age was consistently associated with lower adoption across all classes. Full-service users exhibited exceptionally high social support (OR 4.502, 95% CI 3.601 - 5.627), while moderate comprehensive users showed the highest technology acceptance (OR 2.803, 95% CI 2.355 - 3.342). The decision tree model (area under the curve of 0.94) found the optimal path: sufficient social support (≥2), good health status (>5), and high technical acceptance (≥30) yield the highest use probability (92%→96%). Mediation analysis indicated that social support influences usage willingness through both direct and indirect pathways. The direct effect was 0.712 (95% CI 0.552 - 0.972; *P*<.001). Among indirect pathways, technology availability and practicality accounted for the largest proportion of mediation (19.7%, 95% CI 16.8% - 22.6%), followed by technology acceptance (13.7%, 95% CI 11.1% - 16.3%) and social influence (8.9%, 95% CI 6.9% - 10.9%).

Conclusions: Optimizing age-friendly design, strengthening social support networks, and improving technological usability are keys to increasing the adoption of IH services among the older population. Future policies should develop targeted intervention strategies for different user groups to narrow the digital health divide.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e78037) doi:10.2196/78037

KEYWORDS

aged; internet medicine; mediation effect; social ecological model; technology acceptance model

Introduction

The rapid development of internet health care (IH) has revolutionized health care delivery, particularly for aging populations [1,2]. IH enhances access to teleconsultations for chronic disease management and enables real-time remote monitoring of vital signs through wearable devices [3-5]. These innovations streamline access to everyday health care tasks, such as teleappointments, online prescription refills, and contactless payment, while enhancing medication adherence through user-friendly digital reminders, reducing the logistical barriers to care for older adults with complex needs [6,7]. However, older adults' adoption of IH services remains suboptimal, with usage rates and digital literacy gaps perpetuating health inequities [8].

Due to the decline of the older adults' physiological function, the decline of cognitive flexibility, the lack of digital literacy, and the transformation of social roles, their internet use research has a unique feature of multidimensional interaction and intergenerational cultural conflict [9-12]. These characteristics present particular challenges during the transition to digitization. A study by Yang [11] revealed that perceived usefulness, perceived ease of use, and social influence positively influence older adults' behavioral intention toward health care services. This finding is consistent with recent extensions of the technology acceptance model (TAM) in digital health contexts, such as the work by Mouloudj et al [12], which validated the core TAM constructs (perceived usefulness and ease of use) in predicting intentions to use digital dental health services, while also highlighting the additional roles of trust and social influence. Meraya's [13] findings indicated that the primary barriers to telemedicine use were disinclination towards the technology and difficulties in scheduling appointments. Tan's [14] study underscores the significance of addressing older adults' subjective well-being and enhancing the accessibility of IH services. While existing studies have significantly advanced our understanding of individual determinants in IH adoption, the complex interplay of technological, psychological, and social factors underscores the need for integrated frameworks to fully capture older adults-specific digital disparities. This multidimensional complexity requires a theoretical framework that simultaneously addresses individual technology perception and macro background influences. The TAM operates on an individual level of adoption through dimensions such as perceived technology usefulness and ease of use, capturing cognitive assessments of technology utility [15].

However, the adoption of digital health technologies among older adults is a process shaped by a complex interplay of individual cognition and multilayered environmental influences. While the TAM effectively explains individual-level cognitive mechanisms through perceived usefulness and ease of use, its limited consideration of broader social contexts necessitates integration with the social ecological model (SEM). The SEM systematically categorizes environmental determinants across interpersonal, community, and policy levels, and this integration creates a conceptually synergistic framework wherein factors at these outer levels actively shape the core TAM constructs. This is particularly salient for older adults who often face unique systemic and interpersonal barriers to technology adoption. Specifically, at the interpersonal level, social support and family recommendations not only enhance perceived usefulness through endorsement but also directly help overcome practical usage barriers, thereby reinforcing perceived ease of use [16,17]. At the community level, resources such as age-friendly training programs and accessible digital infrastructure systematically reduce technical and accessibility barriers, transforming potential obstacles into facilitators of technology acceptance [18]. This combined TAM-SEM framework thus unifies macro-level environmental embeddedness with microlevel cognitive processes, providing a comprehensive theoretical lens to understand and optimize IH accessibility for the older population.

This study synthesizes the TAM and the SEM into a unified theoretical framework, establishing a multidimensional analytical structure to systematically investigate intelligent health technology adoption among older adults. The proposed model systematically integrates individual-level determinants such as electronic health literacy and self-efficacy, interpersonal dynamics exemplified by social support networks, and environmental dimensions encompassing perceived security risks and service accessibility. This tripartite integration enables a stratified exploration of adoption mechanisms, effectively delineating how cognitive-behavioral predispositions, relational influences, and systemic enablers collectively shape technology adoption trajectories within geriatric populations.

The primary aim of this study was to characterize patterns of IH adoption among older adults and to determine the extent to which these patterns can be explained by an integrated framework combining constructs from the SEM and the TAM. We hypothesized that older adults would exhibit distinct, empirically identifiable adoption trajectories, ranging from nonuse to comprehensive engagement. The secondary aims were to examine how a combined SEM–TAM framework predicts IH adoption behavior among older adults. We hypothesized that integrating SEM-based contextual determinants (such as social support and health status) with TAM-based cognitive determinants (such as technology acceptance and perceived usability) would provide a coherent basis for understanding and predicting IH adoption patterns within this population.

XSL•FO
RenderX

Methods

Study Population

This study employed a multistage stratified cluster random sampling design to investigate IH service use among older adults aged 60 to 75 in Shandong Province, China. To ensure regional representation, 3 prefecture-level cities representing medium gross domestic product rankings were selected. From each city, we randomly selected 2 counties (representing rural areas) and 2 districts (representing urban areas), achieving a 20% sampling rate at this stage. Within each selected county or district, 2 villages or communities were further chosen through simple random sampling.

Potential participants were identified through local household registry systems. The inclusion criteria required participants to be (1) aged 60 to 75 years, (2) Chinese nationals, and (3) capable of reading and writing. Exclusion criteria included (1) transient residents and (2) individuals with cognitive impairments. Specifically, we excluded those who reported a physician diagnosis of dementia. Additionally, to ensure data quality, we implemented an objective cognitive assessment protocol adapted from standard cognitive screening principles. Exclusion was applied when participants demonstrated (a) failure to comprehend the study purpose and informed consent after 3 structured explanations; (b) inability to answer basic orientation questions (current age, season, or city); or (c) evident disorientation to time or place as evaluated by trained interviewers [19,20].

Between May and July 2024, trained investigators from the Shandong Provincial Center for Disease Control and Prevention conducted face-to-face household interviews. The investigation team consisted of public health professionals with backgrounds in epidemiology, preventive medicine, and health statistics, all of whom had received standardized training on the survey protocol, interview techniques, and ethical considerations specific to this study. A validated questionnaire integrating the TAM and SEM constructs was used. The questionnaire assessed IH service use patterns, perceived ease of use, social support, and environmental barriers. From 1950 eligible individuals approached, 1828 completed valid questionnaires, yielding a response rate of 93.7% (n=1828). The final sample consisted of 48.6% (890/1828) female participants and 52.7% (964/1828) rural residents.

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and was reviewed and approved by the Medical Ethics Review Committee of the Shandong Provincial Center for Disease Control and Prevention (approval SDJK (K) 2024-046-01). Written or verbal informed consent was obtained from all participants prior to their involvement in the study. As a token of appreciation for their time, participants were provided with a small financial compensation of 10 Chinese Yuan (US $ 1.4). To protect participants' privacy and ensure confidentiality, all collected data were anonymized and stored securely, with access restricted to the research team. The reporting of this cross-sectional research followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guideline [21] (Table S1 and Figure S1 in Multimedia Appendix 1).

Outcome

The primary outcome of IH services use was assessed through a multidimensional approach capturing different aspects of service adoption. Initial screening determined any prior IH use through a binary (yes or no) response to the question "have you used IH services before?" Further use was captured through 2 complementary measures: service breadth, represented by the count of different IH service types used from a comprehensive list of 10 specific services; and user typology, derived via latent class analysis of the specific service use patterns.

The secondary outcome, willingness to use IH in the future, was measured as an indicator of behavioral intention, an established precursor to technology adoption in acceptance theories. This construct was assessed using a 5-point Likert scale in response to the question, "Would you be willing to consider using IH services at your next medical visit?" This approach enables the examination of psychological precursors to adoption alongside actual use behavior within the observational study context.

Exposure Factors and Covariates

In this study, basic demographic information was collected from participants through a questionnaire. The participants' educational attainment was categorized as follows: below primary school, primary school, secondary school, high school or technical secondary school, junior college, or above. The marital status of the participants was also documented, with categories including married and unmarried. The participants' place of residence was categorized as either urban or rural. Key constructs, including usability, self-efficacy, eHealth literacy, perceived risks, social influence, and technology acceptance, were measured using validated scales adapted to the IH context. Usability was assessed using a 6-item adapted System Usability Scale, while self-efficacy was assessed using a 3-item adapted General Self-Efficacy Scale. The eHealth literacy construct was measured using the eHealth scale, and perceived risks were evaluated through a 4-item scale based on the Health Belief Model. Technology acceptance was captured using an 8-item scale derived from the TAM. Social influence was assessed as the degree to which people around me believe I should use IH services. It was measured using the Social Impact Scale [22]. All scales used 5-point Likert-type response formats. Complete measurement details, including all scale items, reliability metrics, and validity evidence, are provided in Table S2 in Multimedia Appendix 1.

Conceptual Framework

Our research used a conceptual framework that systematically integrated the TAM and the SEM, with variable selection directly informed by the theoretical constructs of both models. This integrated framework specified that technology adoption (the core outcome in TAM) was influenced not only by individual-level factors but also by multilevel social ecological determinants, thus providing the theoretical rationale for our variable selection. The chosen variables mapped directly onto specific theoretical constructs: usability and technology

XSL•FO
RenderX

acceptance represented core TAM dimensions capturing perceived ease of use and behavioral intention; self-efficacy and eHealth literacy constituted individual-level factors in the SEM that influenced technology adoption capabilities; perceived risks extended the TAM framework by incorporating threat appraisal mechanisms; social impact and social support operationalized the interpersonal level of the SEM, reflecting how social norms and relational resources affect adoption decisions; and health condition represented an individual-level factor in the SEM that shaped technology adoption capacity and motivation. Notably, "peer influence," a specific manifestation of social impact, theoretically operated at the interpersonal level of the SEM while simultaneously shaping the core TAM constructs of perceived usefulness and ease of use through social validation and practical assistance. This theoretical mapping explained why social factors demonstrated both direct effects on adoption and indirect effects through technology acceptance pathways in our analyses. The SEM's levels were intertwined and interactive, with interaction between the older adults and new medical technology occurring in societal (interpersonal) collaboration, all contextualized within health care system environments (Figure S2 in Multimedia Appendix 1) [15,23,24]. This theoretical grounding ensures our variable selection comprehensively captured the multidimensional nature of technology adoption across individual, interpersonal, and community levels.

Statistical Analysis

Latent class analysis (LCA) implemented in *R* classified participants into 5 IH service utilization profiles, optimized via Bayesian information criterion and Akaike information criterion [25]. The resulting classes were subsequently labeled according to their dominant use patterns for intuitive interpretation: class 1: "low-activity testers" (limited exploration of basic functions), class 2: "full-service users" (high probability of using all service types), class 3: "registration-dominant users" (highly focused on appointment services), class 4: "nonusers" (negligible use across all services), and class 5: "moderate comprehensive users" (balanced use of multiple core services). The 5-class solution was retained despite the small sample size (n=23) in class 2 ("full-service users") due to its strong theoretical relevance as a unique profile of early adopters, characterized by high education, income, and technological proficiency. This group is critical for understanding the full spectrum of digital health adoption. These descriptive labels were used in all subsequent tables and figures to replace numerical codes, enhancing the readability and direct interpretability of the findings. The characteristic use patterns of each class were visualized in the radar chart (Figure S3 in Multimedia Appendix 1). Multinomial logistic regression with elastic net regularization (λ=0.1, α=.5) identified predictors of class membership, contrasting nonusers (class 4) against other classes (classes 1 - 3, 5) while adjusting for age, health status, and behavioral covariates. A classification and regression tree algorithm with 10-fold cross-validation, trained on an 80:20 stratified split, predicted IH adoption using feature-selected predictors. Moderated mediation pathways were tested via structural equation modeling with robust maximum likelihood estimation, evaluating indirect effects and moderation by perceived risk. Analyses were supported by nonparametric bootstrap resampling (5000 iterations) for bias-corrected confidence intervals. Analyses used *R* packages tidy (LCA), caret classification and regression tree, and lavaan (structural equation modeling), with model diagnostics confirming adequacy (comparative fit index as 0.921, root mean square error of approximation as 0.043, and standardized root mean square residual as 0.049) [26] (Figure 1). All statistical analyses were conducted using R software (version 4.3.1).

Figure 1. Hypothesized pathways between high technical acceptance, mediating effects, and the health literacy group. (A) Conceptual diagram and (B) statistical diagram.

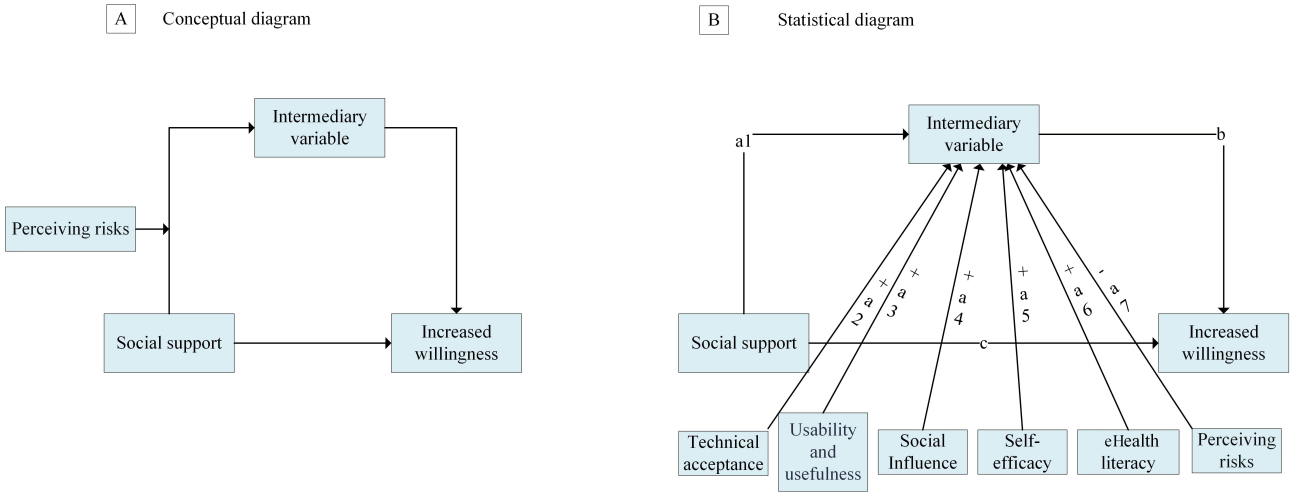

Results

Patterns and Characteristics of IH Service Adoption Among Older Adults

Among the 1828 respondents, 42.2% (n=772) reported using online appointment registration, 29.7% (n=543) used online payment, and 22.4% (n=410) checked test results online, whereas fewer engaged in online consultations (n=271, 14.8%), test scheduling (n=343, 18.8%), or advanced services such as psychological counseling (n=33, 1.8%), medication delivery (n=39, 2.1%), vaccination appointments (n=109, 6.0%), and in-home care (n=28, 1.5%).

XSL•FO
RenderX

Characteristics of Crowd Classification

We used LCA models, stratifying participants into 5 IH use classes (Tables S3 and S4 in Multimedia Appendix 1). Radar/Sankey plots visualized use patterns across clusters (Figures S3 and S4 in Multimedia Appendix 1). Low-activity users (class 1, n=320) demonstrated moderate education (n=161, 50.31%) and high technical acceptance (mean 33.72, SD 3.95), primarily engaging in basic functions like online payment. Compared to nonusers (class 4, n=911), they were younger (age: 65.9 y vs 69.0 y). Full-service users (class 2, n=23) featured high education (n=10, 43.48%), urban residence, male predominance (n=15, 65.22%), superior technical adaptability (mean 38.34, SD 4.77), service availability (mean 24.57, SD 3.25), and the highest median income. Registration-dominant users (class 3, n=286) focused on online appointments (>80%), showed moderate technical acceptance (mean 31.22, SD 3.64), lowest perceived risk (mean 10.98, SD 3.32), and male predominance (n=172, 60.14%). Moderate comprehensive users (class 5, N=288) displayed balanced multiservice engagement, with peak technology acceptance (mean 36.17, SD 5.89) and health literacy (mean 11.83, SD 1.67). Key intergroup distinctions involved technology acceptance, eHealth literacy, and urban-rural distribution. Nonusers (class 4, n=911) were predominantly rural (n=566, 62.13%), older (mean 69.0 y), and less educated (n=230, 25.25% ≤primary), and had elevated risk behaviors (n=107, 11.75% smoking every day), indicating a need for targeted interventions (Table 1).

XSL•FO
RenderX

Table . Baseline characteristics of the study population.

	Total (N=1828)	Class 1 (n=320)[a]	Class 2 (n=23)[a]	Class 3 (n=286)[a]	Class 4 (n=911)[a]	Class 5 (n=288)[a]	*P* value[b]	Relative rank across classes[c]
Sex, n (%)							<.001	
Male	938 (51.31)	158 (49.38)	15 (65.22)	172 (60.14)	426 (46.76)	167 (57.99)		4<1<5<3<2
Female	890 (48.69)	162 (50.63)	8 (34.78)	114 (39.86)	485 (53.24)	121 (42.01)		2<3<5<1<4
Age (years), average (range)	67.9 (60.0 - 75.0)	65.9 (60.0 - 75.0)	67.2 (60.0 - 73.0)	68.2 (60.0 - 75.0)	69.0 (60.0 - 65.0)	66.6 (60.0 - 74.0)	<.001	1<5<2<3<4
Education, n (%)							<.001	
Below primary school	303 (16.58)	22 (6.88)	2 (8.70)	38 (13.29)	230 (25.25)	11 (3.82)		5<1<2<3<4
Primary school	643 (35.18)	63 (19.69)	1 (4.35)	107 (37.41)	355 (38.97)	108 (37.5)		2<1<3<5<4
Secondary school	565 (30.91)	161 (50.31)	5 (21.74)	89 (31.12)	225 (24.7)	85 (29.51)		2<4<5<3<1
High school or technical secondary school	250 (13.68)	57 (17.81)	5 (21.74)	44 (15.38)	85 (9.33)	59 (20.49)		4<3<1<5<2
Junior college or above	67 (3.67)	17 (5.31)	10 (43.48)	8 (2.8)	16 (1.76)	25 (8.68)		4<3<1<5<2
Residency, n (%)							<.001	
Urban	864 (47.26)	165 (51.56)	23 (100)	120 (41.96)	345 (37.87)	211 (73.26)		4<3<1<5<2
Rural	964 (52.74)	155 (48.44)	0 (0)	166 (59.04)	566 (62.13)	77 (26.74)		2<5<1<3<4
Marriage, n (%)							<.001	
No	212 (11.60)	25 (7.81)	2 (8.70)	42 (14.69)	117 (12.84)	26 (9.03)		1<2<5<4<3
Yes	1616 (88.40)	295 (92.19)	21 (91.30)	244 (85.31)	794 (87.16)	262 (90.97)		3<4<5<2<1
Incomes[d]	4.34 (0.67 - 26.82)	5.89 (1.42 - 22.15)	8.73 (4.24 - 24.36)	4.96 (2.17 - 15.83)	2.87 (0.67 - 9.29)	6.92 (2.94 - 18.65)	.16	—[e]
Smoke, n (%)							<.001	
Everyday	196 (10.72)	48 (15.00)	2 (8.70)	18 (6.30)	107 (11.75)	21 (7.29)		3<5<2<4<1
Not everyday	58(3.17)	18 (5.63)	1 (4.35)	22 (7.69)	13 (1.42)	4 (1.39)		5<4<2<1<3
Never	1574 (86.11)	254 (79.37)	20 (86.95)	246 (86.01)	791 (86.83)	263 (91.32)		1<3<4<2<5
Drink, n (%)							<.001	
No	1333 (72.92)	222 (69.38)	20 (86.95)	207 (72.38)	662 (72.67)	222 (77.08)		1<3<4<5<2
Quit drinking	218 (11.93)	38 (11.86)	1 (4.35)	51 (17.82)	108 (11.86)	20 (6.95)		2<5<1<4<3
Yes	277 (15.15)	60 (18.76)	2 (8.70)	28 (9.80)	141 (15.47)	46 (15.97)		2<3<4<5<1
Usability and usefulness, mean (SD)	20.30 (2.47)	22.58 (1.83)	24.57 (3.25)	21.55 (1.98)	17.25 (2.61)	25.83 (2.34)	<.001	4<3<1<2<5
Self-efficacy, mean (SD)	9.53 (1.72)	11.14 (1.56)	11.61 (3.14)	9.67 (1.37)	7.92 (1.89)	12.51 (1.78)	<.001	4<3<1<2<5
Perceiving risks, mean (SD)	12.55 (3.01)	12.54 (2.98)	13.09 (4.56)	10.98 (3.32)	13.41 (3.17)	13.14 (2.89)	<.001	3<2<5<1<4
Electronic health literacy, mean (SD)	9.71 (1.65)	11.20 (1.79)	10.78 (3.82)	10.53 (1.46)	8.23 (2.15)	11.83 (1.67)	<.001	4<3<2<1<5
Social impact, mean (SD)	9.71 (1.88)	11.13 (1.92)	11.48 (3.63)	10.16 (1.54)	8.19 (2.27)	12.34 (1.81)	<.001	4<3<1<2<5

XSL•FO
RenderX

	Total (N=1828)	Class 1 (n=320)[a]	Class 2 (n=23)[a]	Class 3 (n=286)[a]	Class 4 (n=911)[a]	Class 5 (n=288)[a]	*P* value[b]	Relative rank across classes[c]
Social support, mean (SD)	2.94 (0.87)	2.35 (0.65)	6.07 (2.43)	2.91 (0.92)	2.00 (0.58)	2.03 (0.69)	<.001	4<5<1<3<2
Health status, mean (SD)	8.70 (1.34)	8.88 (1.17)	9.91 (2.75)	8.65 (1.28)	8.41 (1.43)	9.39 (1.56)	<.001	4<3<1<5<2
eHealth literacy, mean (SD)	9.92 (3.51)	10.78 (2.39)	11.83 (2.28)	10.53 (4.30)	8.23 (3.11)	11.20 (2.86)	<.001	4<3<1<2<5
Technology acceptance, mean (SD)	29.83 (4.12)	33.72 (3.95)	36.17 (5.89)	31.22 (3.64)	25.17 (3.38)	38.34 (4.77)	<.001	4<3<1<2<5

[a]A total of 5 user classes were identified: class 1 (low-activity tasters, n=320; minimal engagement with select internet health care services), class 2 (full-service users, n=23; high-frequency multifunctional platform use), class 3 (registration-dominant users, n=286; predominant appointment bookings), class 4 (nonusers, n=911; no internet health care adoption), and class 5 (moderate comprehensive users, n=288; intermediate engagement with integrated services, at lower intensity than class 2).

[b]*P* values were derived in accordance with established methodology, encompassing the chi-square test for categorical variables (eg, sex, education) and the Kruskal-Wallis test for continuous variables (eg, age, income). The significance level was set at .05.

[c]Classes ranked from lowest to highest value based on descriptive statistics; only shown for variables with $P<.05$.

[d]The data regarding income are expressed in RMB and the unit is 10,000 yuan (US $1428.6), the value is the median (range), rounded to 2 decimal places. Data processed by Winsorize (truncated by 1% extreme values at the beginning and end).

[e]Not available.

As shown in Table 2, the factors associated with membership in the distinct IH user typologies were derived from a multinomial logistic regression analysis with class 4 (nonusers) as the reference group. The analysis revealed a clear social gradient in IH adoption. This was most evident in class 5 (moderate comprehensive users), where individuals with below primary education had 96% lower odds of membership (OR 0.039, 95% CI 0.012 - 0.084) compared to the reference group with junior college education or above. Male gender was a significant predictor of membership in the more engaged classes, specifically class 2 (full-service users; OR 1.980, 95% CI 1.126 - 3.514) and class 5 (OR 1.310, 95% CI 1.012 - 1.705). Conversely, older age was consistently associated with significantly diminished odds of being in any user class relative to nonusers.

The psychographic and social determinants further delineated the user classes. Class 1 (low-activity tasters) was reported as having higher social impact (OR 1.221, 95% CI 1.125 - 1.325) and marginally positive technology acceptance (OR 1.951, 95% CI 1.630 - 2.332); their risk perception was not significantly different from nonusers, though they demonstrated moderate eHealth literacy (OR 1.905, 95% CI 1.597 - 2.279). Class 2 (full-service users) was distinguished by higher levels of social support (OR 4.502, 95% CI 3.601 - 5.627) and technology acceptance (OR 2.302, 95% CI 1.937 - 2.759), alongside elevated eHealth literacy (OR 1.729, 95% CI 1.443 - 2.060). Class 3 (registration-dominant users) also exhibited significant advantages in social support (OR 1.658, 95% CI 1.381 - 1.979) and technology acceptance (OR 1.652, 95% CI 1.314 - 1.908) compared to nonusers, complemented by higher eHealth literacy (OR 1.651, 95% CI 1.383 - 1.974). Class 5 (moderate comprehensive users) demonstrated the highest odds ratios (ORs) in eHealth literacy (OR 2.202, 95% CI 1.849 - 2.633), social impact (OR 1.320, 95% CI 1.194 - 1.459), and technology acceptance (OR 2.803, 95% CI 2.355 - 3.342).

XSL•FO
RenderX

Table . Correlates of user typology membership from multivariable logistic regression.[a]

Covariates	Class 1: low-activity tasters, OR[b] (95% CI)	Class 2: full-service users, OR (95% CI)	Class 3: registration-dominant users, OR (95% CI)	Class 5: moderate comprehensive users, OR (95% CI)
Sex				
Female (reference)	1.0	1.0	1.0	1.0
Male	0.696 (0.474 - 1.023)	1.980[c] (1.126 - 3.514)	1.463[c] (1.110 - 1.932)	1.310[c] (1.012 - 1.705)
Age	0.879[c] (0.846 - 0.919)	0.892[c] (0.790 - 0.982)	0.937[c] (0.903 - 0.973)	0.888[c] (0.846 - 0.931)
Residency				
Urban	1.404 (0.963 - 2.045)	—[d]	1.182 (0.897 - 1.574)	2.663[c] (1.724 - 4.113)
Rural (reference)	1.0	1.0	1.0	1.0
Education				
Below the primary school	0.111[c] (0.047 - 0.191)	0.073[c] (0.022 - 0.142)	0.282[c] (0.121 - 0.667)	0.039[c] (0.012 - 0.084)
Primary school	0.271[c] (0.151 - 0.497)	0.018[c] (0.002 - 0.059)	0.804 (0.367 - 1.778)	0.331[c] (0.202 - 0.552)
Secondary school	1.109 (0.611 - 3.728)	0.152[c] (0.095 - 0.231)	0.668 (0.297 - 1.499)	0.255[c] (0.157 - 0.424)
High school or technical secondary school	0.662 (0.364 - 1.211)	0.152[c] (0.095 - 0.231)	0.348[c] (0.159 - 0.786)	0.180[c] (0.108 - 0.329)
Junior college or above (reference)	1.0	1.0	1.0	1.0
Marriage				
No	0.584[c] (0.372 - 0.899)	0.453[c] (0.211 - 0.959)	1.508[c] (1.108 - 2.050)	0.651[c] (0.430 - 0.997)
Yes (reference)	1.0	1.0	1.0	1.0
Smoke				
Everyday	1.402 (0.973 - 2.020)	0.745 (0.167 - 3.416)	0.457[c] (0.242 - 0.861)	0.642 (0.393 - 1.057)
Not everyday	3.803 (2.051 - 7.061)	0.393 (0.055 - 3.113)	1.866[c] (1.169 - 8.042)	0.120[c] (0.004 - 0.346)
Never (reference)	1.0	1.0	1.0	1.0
Drink				
No	0.529[c] (0.315 - 0.889)	2.621 (0.566 - 12.300)	1.338 (0.772 - 2.320)	1.275 (0.853 - 1.894)
Quit drinking	0.687 (0.351 - 1.354)	0.314 (0.114 - 7.430)	1.809 (0.934 - 3.503)	0.654 (0.293 - 1.459)
Yes (reference)	1.0	1.0	1.0	1.0
Usability and usefulness	1.041 (0.994 - 1.090)	1.553[c] (1.107 - 2.184)	1.084[c] (1.038 - 1.132)	1.136[c] (1.077 - 1.199)
Self-efficacy	1.103 (0.853 - 1.388)	1.811[c] (1.253 - 2.601)	1.012 (0.934 - 1.097)	1.971[c] (1.372 - 2.879)
Perceiving risks	0.975 (0.931 - 1.022)	0.827[c] (0.619 - 0.981)	0.806[c] (0.730 - 0.904)	0.955[c] (0.913 - 0.999)
Electronic health literacy	1.905[c] (1.597 - 2.279)	1.729[c] (1.443 - 2.060)	1.651[c] (1.383 - 1.974)	2.202[c] (1.849 - 2.633)
Social impact	1.221[c] (1.125 - 1.325)	1.248[c] (1.071 - 1.603)	1.164[c] (1.078 - 1.258)	1.320[c] (1.194 - 1.459)
Social support	1.257[c] (1.055 - 1.496)	4.502[c] (3.601 - 5.627)	1.658[c] (1.381 - 1.979)	1.027 (0.857 - 1.229)
Health status	1.201[c] (1.014 - 1.430)	1.853[c] (1.555 - 2.214)	1.108 (0.927 - 1.319)	1.451[c] (1.224 - 1.727)
Technology acceptance	1.951[c] (1.630 - 2.332)	2.302[c] (1.937 - 2.759)	1.652[c] (1.314 - 1.908)	2.803[c] (2.355 - 3.342)

[a]With class 4 (nonusers) as a reference OR value (ratio), OR represents the probability ratio of a certain type of user possessing a certain feature relative to nonusers. OR>1: this type of user is more likely to possess the feature, OR<1: this type of user is less likely to possess the feature. The analysis used multinomial logistic regression with Bonferroni correction for multiple comparisons. Adjusted covariates included sex, age, education level, residency, marital status, income, smoking status, and drinking status. Five user classes were identified: class 1 (low-activity tasters, n=320; minimal engagement with select internet health care services), class 2 (full-service users, n=23; high-frequency multifunctional platform use), class 3 (registration-dominant users, n=286; predominant appointment bookings), class 4 (nonusers, n=911; no internet health care adoption), and class 5 (moderate comprehensive

XSL•FO
RenderX

users, n=288; intermediate engagement with integrated services, at lower intensity than class 2).

[b]OR: odds ratio.

[c]Statistically significant at the .05 level.

[d]OR and CI could not be calculated because the model did not converge, likely due to a lack of variation in residency (urban/rural) within class 2.

Predictive Factors

The model showed high accuracy in classifying future internet use intention, with area under the curve values of 0.94 (training set) and 0.93 (test set) (Figure 2B). Figure 2C and 2D showed the decision tree model's confusion matrices for the validation and training sets, respectively. Table S5 in Multimedia Appendix 1 shows high sensitivity in both test and training sets, with higher specificity, accuracy, precision, F1 score, lower false positive rate, and good overall test-set performance. The decision tree model revealed the recursive dichotomy of the multivariate feature space affecting IH service use probability. The variable importance order is: social support, health status, usability and usefulness, technology acceptance, perceiving risks, and eHealth literacy. In the left branch, future use probability dropped sharply to 23%. Further division showed that a technology usability and practicality score ≥21 (21% subsample) could partially offset insufficient social support, increasing probability to 55%; but with technology acceptance ≤29 (13% subsample), probability drops to 41%. When perceived health risk >13 (40% subsample), usage probability dropped to 7%, forming the lowest usage group. In the right branch, usage probability jumped to 90%, with health status >5 (38% subsample), further pushing it to 92%, forming a core high-usage group. For those with technology acceptance ≥30 and eHealth literacy ≥12 (34% subsample), use probability reached 96%, reflecting technology and literacy's multiplier effect. The optimal path: sufficient social support (≥2), good health status (>5), and high technical acceptance (≥30) yielded the highest use probability (92% →96%). Insufficient social support (<2) with high-risk perception (>13) led to extremely low use probability (7%), an 89% point difference from the optimal path (Figure 2).

The direct effect showed that for every 1 unit increase in social support, the willingness to use increased by 0.712 (95% CI 0.552 - 0.872; *P*<.001). Among the indirect effects, the mediating contribution of technology usability and practicality is the largest, accounting for 19.7% (95% CI 16.8% - 22.6%) of the total effect, followed by technology acceptance (13.7%, 95% CI 11.1% - 16.3%) and social influence (8.9%, 95% CI: 6.9% - 10.9%). Although self-efficacy (3.2%) and eHealth literacy (2.7%, 95% CI 0.9% - 4.5%) were significant, their contributions are relatively small, while perceived risk was weakly negatively mediated (−0.6%, β=−.007) (Table 3).

XSL•FO
RenderX

Figure 2. A decision tree model for predicting internet health care service use among older adults. AUC: area under the curve; ROC: receiver operating characteristic.

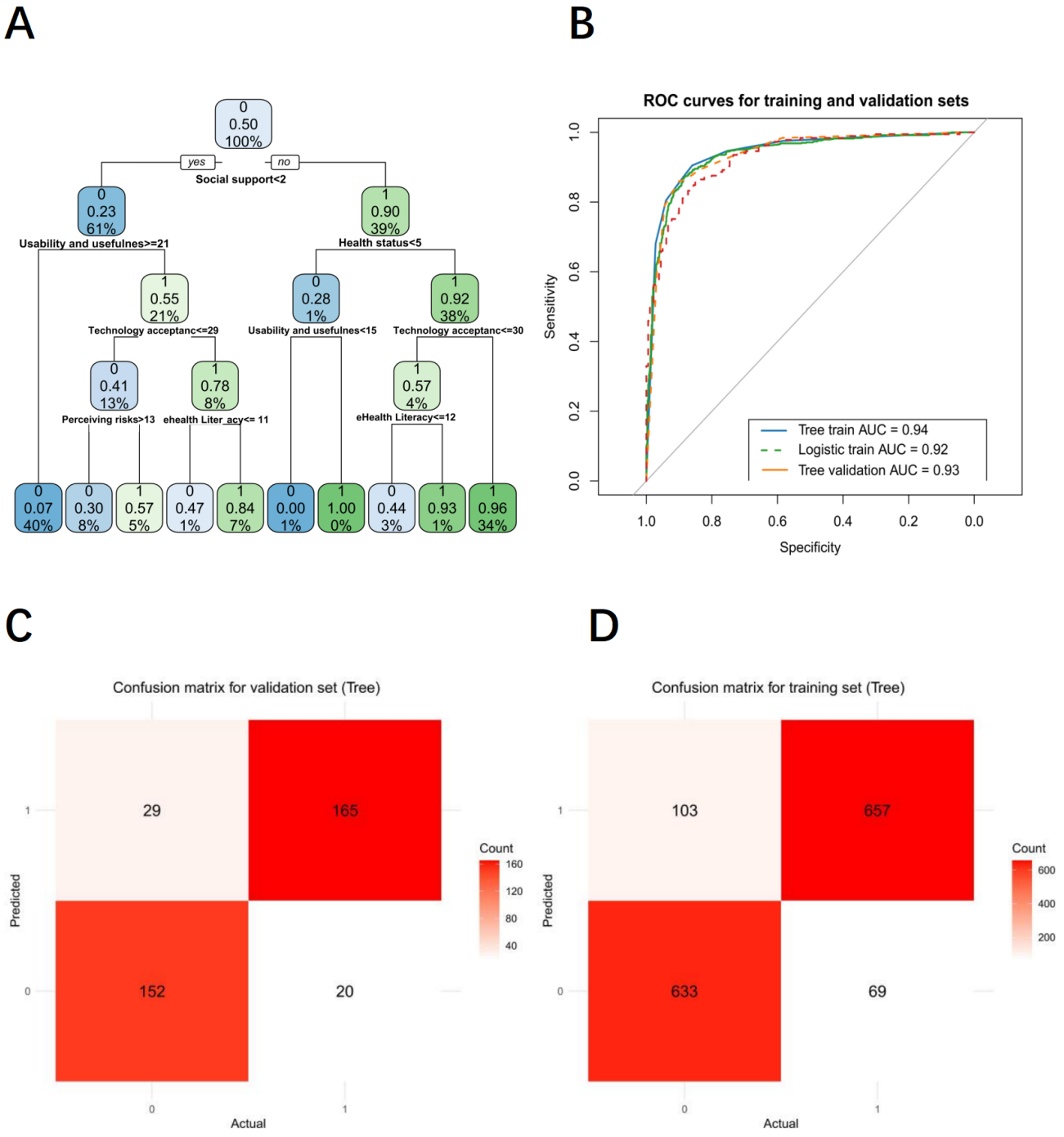

XSL•FO
RenderX

Table . Mediation analysis of the effect of social support on willingness for future internet health care use[a].

Pathway	β (95% CI)	*P* value	Mediation proportion[b], % (95% CI)
Total effect (social support → willingness to use)	1.198 (1.036 to 1.360)	<.001	100
Direct effect (social support → willingness to use)	0.712[c] (0.552 to 0.872)	<.001	59.4 (54.1 to 64.7)
Total indirect effect	0.486[c] (0.420 to 0.552)	<.001	40.6 (35.3 to 45.9)
Mediated pathways (social support → mediator → willingness)			
Usability and usefulness	0.236[c] (0.205 to 0.267)	<.001	19.7 (16.8 to 22.6)
Technology acceptance	0.164[c] (0.134 to 0.194)	<.001	13.7 (11.1 to 16.3)
Social influence	0.107[c] (0.085 to 0.129)	<.001	8.9 (6.9 to 10.9)
Self-efficacy	0.038[c] (0.021 to 0.055)	<.001	3.2 (1.7 to 4.7)
eHealth literacy	0.032[c] (0.011 to 0.053)	.004	2.7 (0.9 to 4.5)
Perceived risk	−0.007[c] (−0.010 to −0.004)	<.001	−0.6 (−0.9 to - 0.3)
Health status	0.009 (−0.002 to 0.020)	.12	—[d]

[a]Model adjusted variables: all analyses adjusted for sex, age, education level, place of residence (urban-rural), marital status, income level, and chronic disease status. Mediation effect calculation: the bias-corrected bootstrap method (5000 repeated samples) is used to estimate the indirect effect. Path explanation: Direct effect: independent impact of social support on usage intention (without mediating variables). Indirect effect: The chain effect of social support on usage intention through mediating variables.

[b]The formula for calculating the mediation ratio is: indirect effect/(direct effect + indirect effect) × 100%.

[c]*P*<.001; a 95% CI that is non-zero is considered significant.

[d]Not applicable.

Discussion

Principal Findings

This study divides the use of IH services by older adults into 5 groups through cluster analysis and constructs an integrated analysis framework based on ecological and technological models, providing a multidimensional perspective for understanding the complex mechanisms of older adults' use of IH services. Research has found that the overall adoption rate of IH services among the elderly population is low, with nearly half of the older adults never using related services and only a very small number becoming regular users. Multivariate analysis shows that advanced age, psychosocial factors, and ease of use of technology are key predictive factors affecting IH participation. Age growth and daily material use are negatively correlated with IH use, while technology usability awareness and electronic health literacy significantly promote usage behavior. Social support drives usage behavior through both direct effects (increasing willingness to use) and indirect effects (mediating effects of technology availability, technology acceptance, and social impact), with the mediating contribution of technology availability being the most prominent. The study emphasizes the need to restructure the design and policy framework of IH services to address the social and technological needs of the older population, in order to bridge the intergenerational gap in the use of health technology.

Theoretical Interpretation of User Patterns Through TAM-SEM Integration

The 5 distinct user patterns emerging from our analysis can be effectively interpreted through our integrated theoretical framework, with the SEM elucidating how macro-level social structures regulate microlevel behavioral patterns. Consistent with prior research on technology adoption barriers [11,27], the low-activity testers (class 1), despite demonstrating moderate technical acceptance, exhibited limited engagement, reflecting how high perceived risk (a TAM extension) and insufficient reinforcing factors at the interpersonal level of the SEM can inhibit adoption even when basic ease of use is acknowledged. This pattern aligns with studies highlighting the critical role of privacy concerns in limiting technology engagement among vulnerable populations [28]. Conversely, the full-service users (class 2), though small in number, exemplify the synergistic effect postulated by the SEM. However, the exceptionally low proportion of comprehensive users in our Chinese sample contrasts sharply with reports from high-income countries [29], suggesting contextual moderators, including structural barriers and cultural preferences for traditional care.

The registration-dominant users (class 3) present a compelling case where a specific, high-use function drives adoption despite moderate overall technical acceptance, echoing previous observations of "feature-specific adoption" in health technology literature [30]. This pattern provides empirical support for the "minimum viable product" approach advocated in implementation science, while simultaneously revealing its

XSL•FO
RenderX

limitations in fostering broader platform engagement. Meanwhile, the moderate comprehensive users (class 5) demonstrate how optimized perceptions across multiple TAM dimensions facilitate broader adoption, reinforcing established technology acceptance theories while extending them to the geriatric context [31]. The nonusers (class 4) represent a population where deficiencies across multiple SEM levels create compounded barriers, a finding that resonates with digital divide research across socioeconomic strata. This multilevel barrier pattern aligns with Bronfenbrenner's ecological systems theory while highlighting its specific manifestations in digital health contexts [32].

TAM-SEM Predictors

The introduction of technical models further reveals the data-driven prediction mechanism. The decision tree model identifies 6 core predictors of older adults' future willingness to use IH services. While perceived risk demonstrates a significant moderating effect that weakens the impact of technology usability on use frequency, usability and usefulness emerge as equally crucial factors that directly facilitate adoption through enhancing perceived ease of use. This finding extends beyond the risk-focused narrative prevalent in digital health literature, revealing a more balanced interplay between barriers and facilitators [33]. Cross-cultural comparisons further illuminate these relationships. Our findings regarding the strong predictive power of usability align with Möller et al's [34] European-Japanese study documenting how interface design fundamentally shapes older adults' digital engagement patterns. However, whereas their research emphasized technological coaching systems, our model reveals how usability interacts with locally specific factors like family support networks in the Chinese context. Similarly, Morishita-Suzuki et al [35] identified depression and leisure activities as crucial determinants of eHealth literacy across European Union and Japanese older adults, while our study extends this understanding by quantifying how eHealth literacy functions within a broader predictive framework alongside usability and social support. The identified predictor hierarchy—social support, health status, usability and usefulness, technology acceptance, eHealth literacy, and perceived risks—represents a significant departure from the conventional TAM that typically prioritizes cognitive perceptions over contextual factors. This corresponds to the "perceived usefulness → usage behavior" pathway in the TAM framework, where perceived usability directly affects technology adoption, while social support indirectly enhances usage intention through a mediating pathway that alleviates anxiety. The prominent role of usability in our model underscores its fundamental importance in gerontechnology design, particularly for populations with limited digital experience. This dual-framework approach reveals that social support not only directly encourages adoption but also indirectly facilitates use through enhancing technology acceptance and reducing anxiety, while usability serves as the foundational element that makes initial engagement possible for older users. This suggests that technology design needs to prioritize addressing pain points such as privacy protection and operational transparency to reduce decision-making barriers for older users [36-38]. This dual-framework approach reveals that social support not only directly encourages adoption but also indirectly facilitates usage through enhancing technology acceptance and reducing anxiety.

Digital Inequality and Policy Implications

This study also underscores the role of digital inequality and broader social determinants. Limited access to smartphones, poor connectivity in rural communities, and lower education levels all constrained adoption, reflecting entrenched disparities in digital health equity. Such inequalities suggest that IH adoption cannot be understood solely as an individual decision but rather as an outcome shaped by structural opportunities and constraints. Addressing these challenges requires coordinated policies that integrate infrastructure investment, targeted subsidies, and the inclusion of IH capacity indicators in routine health assessments [37,38].

The cross-analysis of the 2 models shows that moderate comprehensive service users (class 5) are more sensitive to the usability of the technical interface, while registered dominant users (class 3) are limited to the singular use of service functions, reflecting the gap between the functional integration of the IH platform and the matching needs of user segmentation. From the perspective of policy practice, the inspiration of the ecological model lies in constructing a 3-dimensional intervention system of "individual capability enhancement, social support strengthening, technological environment adaptation" [39,40]: for nonservice users (class 4), it is necessary to reduce perceived risks through community digital literacy training; for registered dominant users, the flow design of appointment services and other functions should be optimized to promote the transformation from a single service to comprehensive use. The successful cases of full-service users provide empirical evidence for the "technology feedback" family model, which amplifies the positive effects of social support through intergenerational collaboration mechanisms. The application value of technical models lies in accurately identifying high-risk groups, such as daily smokers and the older population, who account for a higher proportion of non–service users. Hierarchical intervention can be carried out through the risk scores output by decision tree models, combined with simulation training and privacy protection technology, gradually building an inclusive digital health ecosystem [41,42].

To implement these insights, we propose a further intervention strategy. At the microlevel, develop TAM-driven "cognitive scaffolding" interfaces, such as AI voice assistants with gerontologically optimized dialogue flow, to reduce navigation complexity. Enforce the IH platform to implement general data protection regulation privacy dashboards using intuitive icons, directly addressing the perceived risks we identified in risk literacy balance. At the macro level: standardize algorithm governance through the "elderly friction" policy—forcing IH platforms to disable autoplay functionality and insert reflective prompts (such as "You browsed for 10 min—take a break!") to prevent forced use [43]. Incorporate IH capability indicators into the national elderly health assessment, prioritizing infrastructure upgrades in areas with a lower prevalence of level 2 users [44]. Establish "Digital Health Hub" Peer-Learning Models: community centers (eg, libraries and senior clubs) should host regular peer-led training sessions, facilitated by

XSL•FO
RenderX

trained “Digital Health Champions” from among tech-savvy seniors [45].

The combined application of ecological and technological models also highlights the originality of this study. While ecological models emphasize external environments and TAM focuses on perceptions of usability, their integration offers a multidimensional explanation of adoption. This dual perspective enhances the explanatory power of our findings and demonstrates that interventions must simultaneously target cognitive, relational, and systemic barriers to promote digital health use among older adults.

Limitations

There are several limitations that should be acknowledged in this study. First, a cross-sectional design limits causal reasoning. Although the relationship between the identified factors (such as social support and technology acceptance) and the use of IH services has been observed, a longitudinal study is needed to establish a time relationship and determine whether the improvement of social support and technology acceptance will lead to an increase in the IH adoption rate over time. Second, due to regional sampling, its generalizability is limited. This regional focus may limit the generalizability of research results to other socio-economic backgrounds or geographical regions. Third, the screening of participants for cognitive impairment was based on prior research and clinical judgment rather than a standardized instrument like the Mini-Mental State Examination. While this approach was pragmatic for our large-scale study, it may have resulted in some residual confounding. Fourth, our scale was adapted from a standardized instrument based on the actual conditions of older adults in China, which may affect its external generalizability. These limitations highlight important directions for future research, including longitudinal design, broader geographic sampling, inclusion of objective use data, and a more comprehensive assessment of background factors influencing the adoption of IH in the older population.

Conclusions

This study uses an integrated TAM and SEM framework to examine IH service use among older adults in China. The identification of 5 distinct user profiles demonstrates substantial diversity in adoption patterns, ranging from complete nonuse to comprehensive adoption. Findings reveal that technology adoption is shaped by a multifaceted interplay of cognitive perceptions, relational resources, and systemic enablers. Willingness to use serves as a crucial psychological precursor to actual adoption, influenced by both technological factors and social-environmental contexts. Usability and perceived risk represent significant technological determinants, while social support facilitates adoption through multiple pathways. The integrated TAM-SEM framework provides theoretical value by demonstrating how factors across ecological levels collectively shape adoption behaviors. For practical application, technology design should prioritize aging-friendly interfaces and privacy protections, while community programs should strengthen social support through tailored digital literacy initiatives. Policy efforts must address structural barriers in underserved areas. Future research should explore the longitudinal evolution of user profiles across diverse contexts. By developing targeted strategies for different user segments, we can foster an inclusive digital health ecosystem that benefits all older adults regardless of technological proficiency or social resources.

Acknowledgments

The authors extend their gratitude to all the participants in this study. The authors also thanked the staff from the Shandong CDC and the local community workers for their invaluable support in data collection. In addition, FZ (email: 2778201941@qq.com) and WW (email: 2562109293@qq.com) acted as common corresponding authors. The authors thank them for their expert guidance, which has helped improve the research design and analysis framework.

Funding

This work was funded by Shandong Province Medical and Health Science and Technology Development Plan Project (202312071011) and Shandong Province Humanities and Social Sciences Research Projects (2023-JKZX-11). The funders had no involvement in data collection and analysis or the preparation of this article. The analysis and interpretation of the evidence was done by the authors themselves, not by the funders.

Data Availability

The datasets obtained during the current study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: RL, XX, QL, JH
Formal analysis: RL, XX, QL
Investigation: HL, TTZ, AFA, PL, JT, WW, FZ
Visualization: RL, XX, QL, JH
Writing – original draft: RL, XX, QL
Writing – review & editing: HL, TTZ, AFA, PL, JT, WW, FZ
All authors read and approved the final version.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Study design, measures, and analytic visuals related to internet use among older adults.
[DOCX File, 958 KB - jmir_v28i1e78037_app1.docx]

Checklist 1
STROBE checklist.
[DOCX File, 31 KB - jmir_v28i1e78037_app2.docx]

References

1. Yang F, Shu H, Zhang X. Understanding “internet plus healthcare” in China: policy text analysis. J Med Internet Res 2021 Jul 26;23(7):e23779. [doi: 10.2196/23779] [Medline: 34309581]
2. Yu J, Meng S. Impacts of the internet on health inequality and healthcare access: a cross-country study. Front Public Health 2022;10:935608. [doi: 10.3389/fpubh.2022.935608]
3. Olmedo-Aguirre JO, Reyes-Campos J, Alor-Hernández G, Machorro-Cano I, Rodríguez-Mazahua L, Sánchez-Cervantes JL. Remote healthcare for elderly people using wearables: a review. Biosensors (Basel) 2022 Jan 27;12(2):73. [doi: 10.3390/bios12020073] [Medline: 35200334]
4. Gill SK, Barsky A, Guan X, et al. Consumer wearable devices for evaluation of heart rate control using digoxin versus beta-blockers: the RATE-AF randomized trial. Nat Med 2024 Jul;30(7):2030-2036. [doi: 10.1038/s41591-024-03094-4] [Medline: 39009776]
5. Farsi D. Social media and health care, part I: literature review of social media use by health care providers. J Med Internet Res 2021 Apr 5;23(4):e23205. [doi: 10.2196/23205] [Medline: 33664014]
6. Slabodkin G. Text reminders help seniors improve medication adherence. Fierce Healthcare. 2013 May 20. URL: https://www.fiercehealthcare.com/mobile/text-reminders-help-seniors-improve-medication-adherence [accessed 2025-04-21]
7. Mansour F. Here are 6 ways mobile payments simplify transactions for seniors. Yves Brooks. 2024 Jan. URL: https://yves-brooks.com/here-are-6-ways-mobile-payments-simplify-transactions-for-seniors/ [accessed 2025-04-21]
8. Ausserhofer D, Piccoliori G, Engl A, et al. Community-dwelling older adults' readiness for adopting digital health technologies: cross-sectional survey study. JMIR Form Res 2024 Apr 30;8:e54120. [doi: 10.2196/54120] [Medline: 38687989]
9. Estrela M, Semedo G, Roque F, Ferreira PL, Herdeiro MT. Sociodemographic determinants of digital health literacy: a systematic review and meta-analysis. Int J Med Inform 2023 Sep;177:105124. [doi: 10.1016/j.ijmedinf.2023.105124] [Medline: 37329766]
10. Lin YC, Chung CP, Lee PL, et al. The flexibility of physio-cognitive decline syndrome: a longitudinal cohort study. Front Public Health 2022;10:820383. [doi: 10.3389/fpubh.2022.820383] [Medline: 35734760]
11. Yang HJ, Lee JH, Lee W. Factors influencing health care technology acceptance in older adults based on the technology acceptance model and the unified theory of acceptance and use of technology: meta-analysis. J Med Internet Res 2025 Mar 28;27:e65269. [doi: 10.2196/65269] [Medline: 40153796]
12. Mouloudj K, Bouarar AC, Martínez Asanza DA, Dimitrova M. Understanding customers’ intentions to use digital dental health services: an expanded technology acceptance model. In: Martínez Asanza D, editor. Advances in Medical Technologies and Clinical Practice IGI Global: IGI Global Scientific Publishing; 2024:189-210. [doi: 10.4018/979-8-3693-7165-7.ch007]
13. Meraya AM, Khardali A, Ahmad S, et al. Telehealth perceptions and associated factors among older adults with chronic conditions in Saudi Arabia: a comparative study of users and non-users. Front Public Health 2025;13:1542974. [doi: 10.3389/fpubh.2025.1542974] [Medline: 40144973]
14. Tan SH, Yap YY, Tan SK, Wong CK. Determinants of telehealth adoption among older adults: cross-sectional survey study. JMIR Aging 2025 Mar 24;8:e60936. [doi: 10.2196/60936] [Medline: 40126531]
15. Nadal C, Sas C, Doherty G. Technology acceptance in mobile health: scoping review of definitions, models, and measurement. J Med Internet Res 2020 Jul 6;22(7):e17256. [doi: 10.2196/17256] [Medline: 32628122]
16. De la Peña-López IJ, Acosta-Gonzaga E. Adoption of technology in older adults in Mexico City: an approach from the technology acceptance model. Brain Sci 2025 Jun 12;15(6):632. [doi: 10.3390/brainsci15060632] [Medline: 40563802]
17. Freeman S, Marston HR, Olynick J, et al. Intergenerational effects on the impacts of technology use in later life: insights from an international, multi-site study. Int J Environ Res Public Health 2020 Aug 7;17(16):5711. [doi: 10.3390/ijerph17165711] [Medline: 32784651]
18. Arthanat S. Promoting information communication technology adoption and acceptance for aging-in-place: a randomized controlled trial. J Appl Gerontol 2021 May;40(5):471-480. [doi: 10.1177/0733464819891045] [Medline: 31782347]
19. First MB. Diagnostic and statistical manual of mental disorders, 5th edition, and clinical utility. J Nerv Ment Dis 2013 Sep;201(9):727-729. [doi: 10.1097/NMD.0b013e3182a2168a] [Medline: 23995026]

XSL•FO
RenderX

20. Stephan BCM, Minett T, Pagett E, Siervo M, Brayne C, McKeith IG. Diagnosing mild cognitive impairment (MCI) in clinical trials: a systematic review. BMJ Open 2013;3(2):e001909. [doi: 10.1136/bmjopen-2012-001909] [Medline: 23386579]
21. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
22. Sifat MS, Saperstein SL, Tasnim N, Green KM. Motivations toward using digital health and exploring the possibility of using digital health for mental health in Bangladesh university students: cross-sectional questionnaire study. JMIR Form Res 2022 Mar 4;6(3):e34901. [doi: 10.2196/34901] [Medline: 35254267]
23. Holden RJ, Karsh BT. The technology acceptance model: its past and its future in health care. J Biomed Inform 2010 Feb;43(1):159-172. [doi: 10.1016/j.jbi.2009.07.002] [Medline: 19615467]
24. Davidson P, Rushton CH, Kurtz M, et al. A social-ecological framework: a model for addressing ethical practice in nursing. J Clin Nurs 2018 Mar;27(5-6):e1233-e1241. [doi: 10.1111/jocn.14158] [Medline: 29119653]
25. Grant RW, McCloskey J, Hatfield M, et al. Use of latent class analysis and k-means clustering to identify complex patient profiles. JAMA Netw Open 2020 Dec 1;3(12):e2029068. [doi: 10.1001/jamanetworkopen.2020.29068] [Medline: 33306116]
26. Choi DH, Ro YS, Park JH, et al. Evaluation of socioeconomic position and survival after out-of-hospital cardiac arrest in Korea using structural equation modeling. JAMA Netw Open 2023 May 1;6(5):e2312722. [doi: 10.1001/jamanetworkopen.2023.12722] [Medline: 37163262]
27. An J, Zhu X, Wan K, et al. Older adults' self-perception, technology anxiety, and intention to use digital public services. BMC Public Health 2024 Dec 19;24(1):3533. [doi: 10.1186/s12889-024-21088-2] [Medline: 39702092]
28. Liu H, Zhou Q, Liang S. Digital inclusion in public services for vulnerable groups: a systematic review for research themes and goal-action framework from the lens of public service ecosystem theory. Gov Inf Q 2025 Jun;42(2):102019. [doi: 10.1016/j.giq.2025.102019]
29. Tavares AI. Self-assessed health among older people in Europe and internet use. Int J Med Inform 2020 Sep;141:104240. [doi: 10.1016/j.ijmedinf.2020.104240] [Medline: 32739610]
30. Zhang Y, Wu P. Continuous adoption of online healthcare platforms: an extension to the expectation confirmation model and network externalities. BMC Public Health 2024 Sep 27;24(1):2630. [doi: 10.1186/s12889-024-20072-0]
31. Karahanna E, Straub DW, Chervany NL. Information technology adoption across time: a cross-sectional comparison of pre-adoption and post-adoption beliefs. MIS Q 1999 Jun;23(2):183-213. [doi: 10.2307/249751]
32. Ucar I, Gramaglia M, Fiore M, Smoreda Z, Moro E. News or social media? Socio-economic divide of mobile service consumption. J R Soc Interface 2021 Dec;18(185):20210350. [doi: 10.1098/rsif.2021.0350] [Medline: 34847793]
33. Tustin JL, Crowcroft NS, Gesink D, Johnson I, Keelan J. Internet exposure associated with Canadian parents' perception of risk on childhood immunization: cross-sectional study. JMIR Public Health Surveill 2018 Jan 19;4(1):e7. [doi: 10.2196/publichealth.8921] [Medline: 29351896]
34. Möller J, Stara V, Amabili G, et al. Toward innovation in healthcare: An analysis of the digital behavior of older people in Europe and Japan for the introduction of a technological coaching system. Healthcare (Basel) 2024 Jan 8;12(2):143. [doi: 10.3390/healthcare12020143] [Medline: 38255032]
35. Morishita-Suzuki K, Ogawa T, Bevilacqua R, et al. The impact of depression and leisure activities on e-health literacy among older adults: a cross-cultural study in the EU and Japan. Int J Environ Res Public Health 2025 Mar 10;22(3):403. [doi: 10.3390/ijerph22030403] [Medline: 40238518]
36. Wutz M, Hermes M, Winter V, Köberlein-Neu J. Factors influencing the acceptability, acceptance, and adoption of conversational agents in health care: integrative review. J Med Internet Res 2023 Sep 26;25:e46548. [doi: 10.2196/46548] [Medline: 37751279]
37. Zhang J, Gallifant J, Pierce RL, et al. Quantifying digital health inequality across a national healthcare system. BMJ Health Care Inform 2023 Nov 24;30(1):e100809. [doi: 10.1136/bmjhci-2023-100809] [Medline: 38007224]
38. Jahnel T, Dassow HH, Gerhardus A, Schüz B. The digital rainbow: digital determinants of health inequities. Digit Health 2022;8:20552076221129093. [doi: 10.1177/20552076221129093] [Medline: 36204706]
39. Ae-Ri J, Kowoon L, Eun-A P. Development and evaluation of the information and communication technology-based loneliness alleviation program for community-dwelling older adults: a pilot study and randomized controlled trial. Geriatr Nurs 2023;53:204-211. [doi: 10.1016/j.gerinurse.2023.07.011] [Medline: 37544264]
40. Sullivan J, Kosuth E. Technology use, barriers, and future needs among community-dwelling older adults. J Gerontol Nurs 2024 Feb;50(2):26-31. [doi: 10.3928/00989134-20240110-04] [Medline: 38290100]
41. Golinelli D, Pecoraro V, Tedesco D, et al. Population risk stratification tools and interventions for chronic disease management in primary care: a systematic literature review. BMC Health Serv Res 2025 Apr 10;25(1):526. [doi: 10.1186/s12913-025-12690-0] [Medline: 40205373]
42. Mishra US, Yadav S, Joe W. The Ayushman Bharat digital mission of India: an assessment. Health Syst Reform 2024 Dec 17;10(2):2392290. [doi: 10.1080/23288604.2024.2392290] [Medline: 39437234]
43. Liu M, Wang C, Hu J. Older adults' intention to use voice assistants: usability and emotional needs. Heliyon 2023 Nov;9(11):e21932. [doi: 10.1016/j.heliyon.2023.e21932]

XSL•FO
RenderX

44. Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med Ethics 2021 Sep 15;22(1):122. [doi: 10.1186/s12910-021-00687-3] [Medline: 34525993]
45. Lim GP, Appalasamy JR, Ahmad B, Quek KF, Shaharuddin S, Ramadas A. PEER-led digital health lifestyle intervention for a low-income community at risk for cardiovascular diseases (MYCardio-PEER): a quasi-experimental study protocol. Prim Health Care Res Dev 2025 Mar 3;26:e20. [doi: 10.1017/S1463423625000192] [Medline: 40025749]

Abbreviations

IH: internet health care
LCA: latent class analysis
OR: odds ratio
SEM: social ecological model
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TAM: technology acceptance model

Edited by A Stone; submitted 25.May.2025; peer-reviewed by G Amabili, J Kaner, K Mouloudj; accepted 26.Nov.2025; published 15.Jan.2026.

Please cite as:
Li R, Xu X, Li Q, Liu H, Zhou TT, Amhare AF, Liu P, Tang J, Wang W, Zheng F, Han J
Internet Health Care Service Use Behavioral Pattern Among Older Adults and the Role of the Technology Acceptance and Social Ecological Theory Model: Cross-Sectional Survey
J Med Internet Res 2026;28:e78037
URL: https://www.jmir.org/2026/1/e78037
doi:10.2196/78037

© Rui Li, Xinyu Xu, Qingsong Li, Haobiao Liu, Ting Zhou, Abebe Feyissa Amhare, Peiyu Liu, Jing Tang, Wei Wang, Fuju Zheng, Jing Han. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Therapeutic Effects of a WeChat Mini-Program on Metabolic Dysfunction–Associated Fatty Liver Disease: Randomized Controlled Trial

Chao Sun[1], MD, PhD; Guangyu Chen[2], MD, PhD; Cuicui Shi[1], MD, PhD; Haixia Cao[1], MD, PhD; Ruixu Yang[1], MD, PhD; Jing Zeng[1], MD, PhD; Xiaoyan Duan[1], MD, PhD; Xin Sun[2], MD; Jian-Gao Fan[1], MD, PhD

[1]Center for Fatty Liver Disease, Department of Gastroenterology, Xinhua Hospital, Shanghai Jiao Tong University School of Medicine, 1665 Kong Jiang Road, Shanghai, China

[2]Clinical Research and Innovation Unit, Xinhua Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Corresponding Author:
Jian-Gao Fan, MD, PhD
Center for Fatty Liver Disease, Department of Gastroenterology, Xinhua Hospital, Shanghai Jiao Tong University School of Medicine, 1665 Kong Jiang Road, Shanghai, China

Abstract

Background: For patients with metabolic dysfunction–associated fatty liver disease (MAFLD), weight loss is advised but challenging in practice. In China, there is a pronounced shortage of tailored digital lifestyle interventions for this population.

Objective: This study aimed to assess the effects of a WeChat mini-program-delivered lifestyle intervention on weight loss and hepatic steatosis among individuals with MAFLD who were overweight or obese.

Methods: Adults who are overweight or obese and have clinically diagnosed MAFLD with transient elastography examination were enrolled in this prospective randomized controlled trial. Patients were randomly assigned to receive either WeChat mini-program management (intervention group) or standard care (control group) at a 1:1 ratio. The intervention was structured around the development and implementation of personalized diet and exercise plans, supplemented by guided exercise video courses and reinforced through continuous monitoring and informational support. Body weight and clinical parameters were assessed at baseline and then at 6 months.

Results: A total of 89 patients met the inclusion criteria and were randomly assigned to the intervention group (n=45) or control group (n=44). Among the 89 patients with MAFLD, 60% (27/45) of them achieved a weight loss of ≥5%, and 24.4% (11/45) of them had a weight loss of ≥10% in the intervention group, which was greater than those in the control group (27/45 vs 7/44; relative risk [RR] 3.771, 95% CI 1.836 - 7.748; *P*<.001; 11/45 vs 3/44, RR 3.585, 95% CI 1.072 - 11.988; *P*=.02). Importantly, patients receiving the intervention were significantly more likely to achieve a ≥10% reduction or normalization of controlled attenuation parameter (CAP) than those in the control group (26/45 vs 14/44; RR 1.816, 95% CI 1.102 - 2.992; *P*=.01). After adjusting for key baseline covariates, multivariate analysis confirmed the intervention's positive effect on achieving a weight loss of ≥5% (OR [odds ratio] 8.380, 95% CI 2.886 - 24.331; *P*<.001) of ≥10% (OR 4.612, 95% CI 1.138 - 18.686; *P*=.03), as well as on CAP reduction of ≥10 % or normalization (OR 2.853, 95% CI 1.092 - 7.456; *P*=.03). In parallel, the intervention group presented greater reductions in liver enzymes (alanine aminotransferase, aspartate aminotransferase, and γ-glutamyl transpeptidase) and metabolic parameters (fasting insulin, hemoglobin A_{1c}, and triglyceride) than the control group (all *P*<.05). According to the fibrosis assessment, only the FibroScan-aspartate aminotransferase score decreased more in the intervention group than in the control group (median difference −0.06, 95% CI −0.13 to −0.01; *P*=.02), as compared to other non-invasive indicators.

Conclusions: Readily scalable in primary care and varied-resource settings, our WeChat mini-program-based intervention extends beyond weight loss to reduce hepatic steatosis and improve metabolic parameters, thereby addressing the critical gap in targeted MAFLD management in China with a low-cost model for high-burden populations. Nevertheless, larger future studies are needed to confirm these findings with greater precision and assess long-term sustainability.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2500100197; https://www.chictr.org.cn/showprojEN.html?proj=247294

(J Med Internet Res 2026;28:e76204) doi:10.2196/76204

XSL•FO
RenderX

KEYWORDS

metabolic dysfunction-associated fatty liver disease; WeChat mini-program; lifestyle intervention; weight loss; metabolic parameters; randomized controlled trial

Introduction

Background

Metabolic dysfunction–associated fatty liver disease (MAFLD) represents a significant and growing public health burden worldwide, paralleling the epidemics of obesity and type 2 diabetes mellitus [1]. It is a gradually progressive disease that evolves through a spectrum that begins with simple hepatic steatosis and advances to metabolic dysfunction-associated steatohepatitis (MASH), ultimately resulting in liver fibrosis, cirrhosis, and even hepatocellular carcinoma [1,2]. Despite its severe clinical sequelae, there are no pharmacologic therapies approved in China to reverse the histological progression of MAFLD, underscoring the critical role of nonpharmacological management [3].

Lifestyle intervention involving dietary adjustments and physical exercise is still the cornerstone of MAFLD management. Clinical guidelines confirm that a body weight reduction of 3%-5% can ameliorate hepatic steatosis, a 7%-10% loss can improve MASH, and over 10% may even lead to fibrosis regression [3]. However, the real-world implementation of these interventions faces significant challenges [4,5].

Traditional lifestyle interventions rely on face-to-face guidance from clinical doctors in hospital settings, which often results in poor execution and low compliance, limiting the coverage and sustainability of treatment and seriously affecting patient outcomes [6]. A cross-sectional survey from China revealed suboptimal self-management of lifestyle behaviors among adults with MAFLD [7]. Lack of motivation and real-time supervision are key contributing factors to the inadequate effectiveness of lifestyle interventions. In view of this, there is an urgent need for more convenient and effective measures to enhance the implementation and adherence of lifestyle changes for patients with MAFLD, thereby improving the prognosis.

With the popularization of the internet and the rapid development of artificial intelligence, digital therapeutics (DTx), such as mobile health apps or programs, have emerged as a promising solution to bridge this care gap [8,9]. By delivering personalized health information and professional guidance, DTx enables convenient, home-based self-management, which can correct unhealthy lifestyle habits and improve treatment adherence [10-12]. A meta-analysis of 8 studies demonstrated that DTx interventions lasting 4 - 24 months achieved a clinically significant weight loss at a rate of 33%, accompanied by improvements in liver enzymes and reduced liver fat [13]. In China, the near-ubiquitous use of WeChat offers a unique platform for such interventions [14,15]. Its widespread adoption eliminates the need for additional app downloads, fosters user trust through a familiar interface, and facilitates seamless communication between health care providers and patients, creating an ideal ecosystem for closed-loop chronic disease management [16,17].

Despite this potential, there is a notable scarcity of WeChat mini-programs specifically designed and rigorously evaluated for structured lifestyle intervention in Chinese patients with MAFLD. To address this gap, we developed an innovative interactive WeChat mini-program, named the therapeutic lifestyle changes (TLCs) program, which is designed to assist patients in implementing sustainable lifestyle changes for the management of MAFLD. The program offers a suite of features, including weight loss goal setting, diet records and recommendations, and instructional exercise videos.

Objectives

Therefore, the aim of this study was to evaluate the efficacy of this novel TLC program in managing MAFLD. We hypothesized that, compared to participants receiving standard care, those engaged in the TLC intervention would demonstrate significantly greater improvements in reduction of liver fat content and body weight, as well as glucose and lipid metabolic parameters.

Methods

Study Design and Setting

A 6-month parallel, randomized controlled trial was conducted at Xinhua Hospital, affiliated with Shanghai Jiao Tong University School of Medicine, a tertiary-care hospital in China. The study period for participant recruitment and data collection spanned from August 2022 to May 2024 at our hospital. All eligible participants were randomly assigned to either the TLC program management group (intervention group) or the standard care group (control group) at a 1:1 ratio. The protocol (which was retrospectively registered) with no deviations was approved by the Institutional Review Board (IRB) prospectively. The reporting of this trial adhered to the CONSORT (Consolidated Standards of Reporting Trials) statement [18], with the checklist presented in Checklist 1.

Participants

The participants with clinically diagnosed MAFLD were consecutively recruited by clinicians from our hepatology clinic. Those who were interested provided written informed consent. All participants underwent an initial, face-to-face eligibility assessment at Xinhua Hospital according to the study criteria. The inclusion criteria were as follows: (1) aged 18 - 65 years, (2) BMI between 24 and 35 kg/m^2, (3) clinically diagnosed with MAFLD by radiological features with ultrasound confirmation and a controlled attenuation parameter (CAP) value of ≥248 dB/m, and (4) able to learn and operate a mobile-based program. Patients were excluded if they had excessive alcohol consumption (weekly ethanol intake of ≥210 g for males and ≥140 g for females weekly), etiological evidence of an alternative liver disease, such as chronic viral hepatitis, autoimmune liver diseases, drug-induced hepatic steatosis, hepatic decompensation, severe comorbidities (severe cardiopulmonary disease, uncontrolled diabetes, chronic kidney

XSL•FO
RenderX

disease, psychiatric disease, and malignant tumors), recent weight loss, or weight loss medication intake.

Randomization

After the completion of baseline assessments, eligible participants were registered in the study. They were then automatically and randomly allocated to each study group at a 1:1 ratio by a computer-generated random system. The allocation was concealed until the moment of assignment, as the outcome was only revealed by the system after the participant's formal registration. Given the nature of the digital intervention, participants and health care providers were not blinded to group assignment, but outcome assessors were blinded during data analysis.

Lifestyle Intervention

"Design of the TLC program"

The TLC program-based lifestyle intervention was collaboratively designed by physicians from various departments, including Gastroenterology, Sport Rehabilitation, and Clinical Nutrition. Individuals in the control group were provided with standard care in the clinic according to the established treatment guidelines [19]. The TLC program intervention spanned 24 weeks and was divided into 4 distinct modules, which are described below.

Development of Diet and Exercise Plans

Individualized weight loss and exercise goals were set based on the baseline body weight, existing metabolic disorders, and severity of fatty liver disease. The target for weight reduction was established as 5%-10% of the initial body weight.

Dietary Intervention Plan

Based on preset weight loss goals, the program calculated daily dietary energy requirements, providing a recommended meal plan with the corresponding energy intake. If patients did not prefer the recipes suggested by the program, they could use the food exchange list to select alternative food with equivalent energy values. The program required patients to log and upload photos of their daily food intake, after which a human nutritionist analyzed the diet log and provided feedback via WeChat.

Exercise Video Courses

Participants were provided with a heart rate armband (HW702, YESOUL) to track heart rate during exercise. During each exercise session, data regarding heart rate were automatically uploaded by the armband. Each course lasted 30 minutes, and patients were to choose the training course at least 5 times per week. Additionally, patients could select other forms of exercise and record the duration of exercise, such as fast walking, running, cycling, and swimming.

Monitoring and Information Support

Patients were expected to complete their diet and exercise tasks through a check-in system every day. The research team sent tailored feedback messages weekly according to the status of task completion and offered regular suggestions through WeChat. Moreover, the system intelligently provided relevant educational articles, which provided professional information and support tailored to the patient's interests and concerns.

Measurements

Clinical Characteristics

Clinical characteristics regarding anthropometric and laboratory data were obtained at our hospital using a standard protocol at baseline and 6 months. Body composition was assessed via a bioimpedance analyzer (MC-980MA, TANITA). CAP and liver stiffness measurement (LSM) were performed using the FibroScan-502 device (Echosens) with an M or XL probe, following the manufacturer's guidelines. A reliable examination was defined as at least 10 valid measurements and an IQR to median ratio of LSM below 30%. Based on an individual patient data meta-analysis, CAP thresholds of 268 and 280 dB/m were identified as optimal cutoffs for diagnosing moderate and severe steatosis, respectively [20].

Dietary Nutrition and Physical Activity Assessment

Dietary nutrition was assessed through a face-to-face interview with a questionnaire developed by the Chinese Society of Health Management and the Chinese Nutrition Society, which consists of 24 questions with a total score of 100 [21]. A score below 60 indicates a risk of dietary nutrition issues; a score between 60 and 75 suggests a potential risk; and a score above 75 indicates no dietary nutritional risk. Physical activity levels were evaluated via the International Physical Activity Questionnaire-Short Form (IPAQ-SF), which captures data on activity intensity, frequency, and daily duration in a typical week. Energy expenditure was quantified by the metabolic equivalent of task (MET) [22]. The weekly physical activity level for a specific intensity was calculated as the MET value of the activity× weekly frequency (d/wk)× daily duration (min/d).

Diagnostic Criteria

Obesity was considered a BMI≥28 kg/m^2 [3]. The definition of MetS adhered to previously published criteria [23]. The homeostasis model assessment-insulin resistance (HOMA-IR) score was calculated as fasting insulin (mU/L)× fasting plasma glucose (FPG, mmol/L)/22.5. Insulin resistance was characterized by a HOMA-IR score of 2.5 or higher [1]. The fatty liver index was computed to indirectly assess the degree of hepatic steatosis using an established formula [24]. The FibroScan-aspartate aminotransferase (FAST) score was used to identify patients with MASH with significant fibrosis [25], whereas the aspartate aminotransferase-to-platelet ratio index (APRI), fibrosis-4 (FIB-4) index, nonalcoholic fatty liver disease fibrosis score (NFS), and Agile 3+ score were used to assess the risk of advanced fibrosis [26-29].

Outcome Measurements

The analyses included all participants who were randomized. The intention-to-treat (ITT) population for efficacy evaluation in this trial consisted of all cases that were randomized and had efficacy baseline records. The per-protocol (PP) population included participants who completed the entire planned observation period and had no major protocol deviations. The primary efficacy endpoint was the reduction of body weight by

XSL•FO
RenderX

at least 5%. The secondary endpoints included a CAP value of ≥10% reduction or normalization and changes in the levels of liver enzymes and lipid and glucose parameters.

Sample Size Calculation

The sample size calculation was based on the primary study outcome, which is the proportion of patients achieving a weight loss of ≥5% from baseline. It was postulated that the intervention would produce a 5-fold increase in the success rate from a baseline of 8% in the control group [30]. To detect this effect with 80% power at a 5% significance level and accounting for a 20% dropout rate, a total sample size of 84 participants was required.

Data Analysis

Statistical analyses were performed using SPSS 23.0 software (IBM Corp). Missing data were due to patient dropouts and imputed by the last observation carried forward. Regarding the primary endpoint, this approach was chosen as it provides a conservative estimate of the treatment effect by assuming no further improvement in their weight status due to the discontinuation of active intervention, thereby avoiding overoptimistic extrapolation [6,31,32]. Descriptive statistics were conducted for all variables, with means (95% CIs) or medians (IQRs) reported for continuous variables and percentages for categorical variables. Categorical variables were analyzed using the chi-square test or Fisher exact test between groups (reporting relative risk and 95% CI) and the marginal homogeneity test within each group. Continuous variables were assessed using 2-tailed paired *t* tests or Wilcoxon signed-rank tests within each group, whereas unpaired *t* tests (reporting mean difference and 95% CI) or Mann-Whitney *U* tests (reporting Hodges-Lehmann median difference and 95% CI) were used to compare the changes from the baseline between groups. Binary logistic regression was used to evaluate the effect of the intervention on weight and CAP changes. The results are presented as odds ratios (ORs) with 95% CIs. A 2-tailed *P* value of .05 was considered significant.

Ethical Considerations

This research was approved by the IRB of Xinhua Hospital (XHEC-C-2021-076-2) and was retrospectively registered with the Chinese Clinical Trial Registry (ChiCTR2500100197). This trial was registered retrospectively owing to disruptions in the research team's workflow and uncertainties about the study's continuity during the COVID-19 pandemic in China. We confirm that all data collection for this study was initiated subsequent to receiving approval from the IRB of Xinhua Hospital, with the IRB approval document provided in Multimedia Appendix 1. The study start was delayed due to COVID-19, and there were no other deviations or changes from the protocol after IRB approval and trial commencement. All eligible participants provided written informed consent. Participants were informed of their voluntary participation and their right to withdraw from the study at any time. It was clarified that all study procedures were offered at no cost, but no financial compensation would be made upon study completion. All personally identifiable information was coded with a unique study ID, and the anonymized data were securely stored in a password-protected cabinet.

Results

Participants and Baseline Characteristics

Figure 1 presents the study flowchart in accordance with the CONSORT guidelines. A total of 89 patients met the inclusion criteria and were randomly assigned to the intervention group (n=45) or control group (n=44); these patients comprised the ITT population. At the 6-month follow-up point, 8 participants were lost to follow-up in the intervention group and 8 in the control group. A total of 73/89 (82%) patients successfully completed the study and comprised the PP population, while the remaining 16/89 (18%) patients dropped out (Figure 1). The baseline characteristics of the patients are shown in Table 1. The mean age of the whole study population was 40.0 (SD 9.8) years. Around 61.8% (55/89) were males and 55.1% (49/89) were obese (BMI≥28 kg/m^2). Overall, 34.8% (31/89) had hypertension, 70.8% (63/89) had dyslipidemia, and 6.7% (6/89) had type 2 diabetes mellitus. There were no significant differences between the 2 groups in terms of age, sex, weight, presence of comorbidities, biochemical variables, diet, or activity-related parameters (*P*>.05). Additionally, no difference was found in the CAP or LSM between the groups at baseline (*P*>.05).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram.

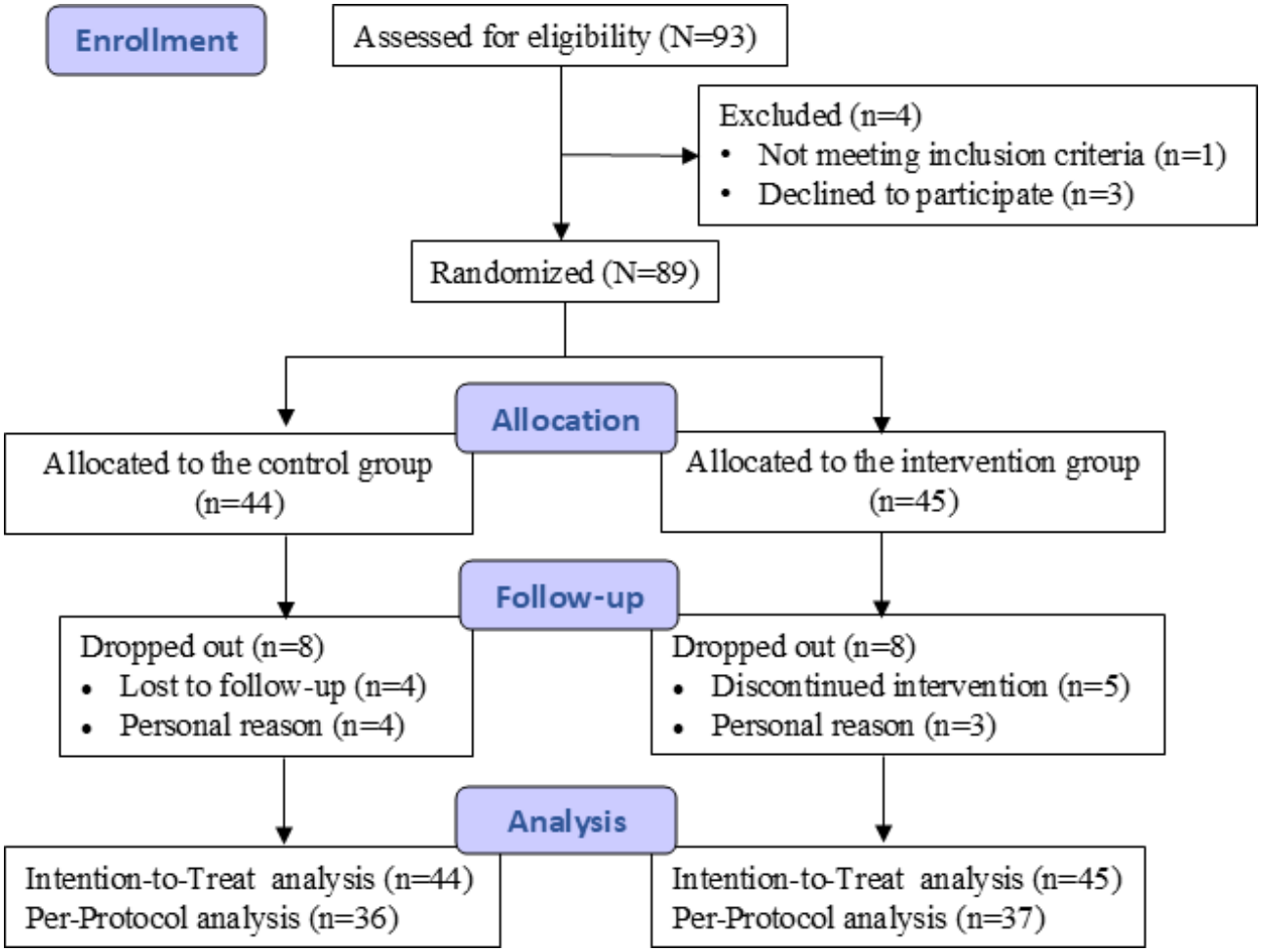

XSL•FO
RenderX

Table . Baseline characteristics of the total study population and the randomized allocation groups in intention-to-treat population.

Characteristic	Total (N=89)	Control (n=44)	Intervention (n=45)	*P* value[a]
Sex (male/female), n(%)	55 (61.8)/ 34(38.2)	31 (70.5)/ 13(29.5)	24 (53.3)/ 21(46.7)	.10
Age (years), mean (SD; 95% CI)	40.0 (9.8; 37.9-42.1)	40.7 (10.5; 37.5-43.9)	39.3 (9.1; 36.6-42.1)	.84
Weight (kg), mean (SD; 95% CI)	82.5 (11.6; 80.1-85.0)	82.3 (11.2; 78.9-85.7)	82.7 (12.1; 79.1-86.4	.92
BMI (kg/m^2), mean (SD; 95% CI)	28.6 (3.0; 27.9-29.2	28.2 (3.0; 27.2-29.1)	29.0 (2.9; 28.1-29.8)	.25
BMI≥28 (kg/m^2), n (%)	49 (55.1)	21 (47.7)	28 (50.9)	.75
WC[b] (cm), mean (SD; 95% CI)	92.8 (8.9; 90.9-94.7)	92.6 (8.9; 89.9-95.3)	93.0 (9.0; 90.3-95.7)	.97
Waist to hip ratio, mean (SD; 95% CI)	0.90 (0.06; 0.89-0.92)	0.91 (0.06; 0.89-0.93)	0.90 (0.07; 0.88-0.92)	.38
Waist to height ratio, mean (SD; 95% CI)	0.55 (0.05; 0.54-0.56)	0.54 (0.05; 0.53-0.56)	0.55 (0.04; 0.54-0.56)	.45
Systolic blood pressure (mmHg), mean (SD; 95% CI)	123.8 (6.8; 122.4-125.3)	124.4 (6.7; 122.4-126.5)	123.3 (7.0; 121.2-125.4)	.24
Diastolic blood pressure (mmHg), mean (SD; 95% CI)	79.3 (5.7; 78.1-80.5)	79.6 (5.9; 77.9-81.4)	79.0 (5.6; 77.3-80.7)	.35
Hypertension, n (%)	31 (34.8)	17 (38.6)	14 (31.1)	.46
Type 2 diabetes mellitus, n (%)	6 (6.7)	4 (9.1)	2 (4.4)	.43
Dyslipidemia, n (%)	63 (70.8)	33 (75)	30 (66.7)	.39
Hyperuricemia, n (%)	34 (38.2)	18 (40.9)	16 (35.6)	.60
Metabolic syndrome, n (%)	37 (41.6)	19 (43.2)	18 (40)	.76
Platelet count (10^9/L), mean (SD; 95% CI)	256.6 (51.7; 245.8-267.5)	254.5 (51.9; 238.7-270.3)	258.8 (51.9; 243.2-274.4)	.58
ALT[c] (U/L), median (IQR)	36.0 (26.0-59.0)	38.0 (24.5-57.3)	34.0 (26.5-61.0)	.90
AST[d] (U/L), median (IQR)	26.0 (21.0-36.0)	26.0 (21.3-35.5)	25.0 (19.5-36.0)	.68
GGT[e] (U/L), median (IQR)	35.0 (23.0-53.5)	39.0 (25.3-59.8)	34.0 (22.0-50.0)	.16
Albumin (g/L), mean (SD; 95% CI)	46.3 (2.4; 45.8-46.8)	46.6 (2.0; 46.0-47.2)	45.9 (2.6; 45.1-46.7)	.20
Urea (mmol/L), mean (SD; 95% CI)	4.9 (1.1; 4.7-5.1)	5.0 (1.1; 4.6-5.3)	4.9 (1.0; 4.6-5.2)	.75
Serum creatinine (μmol/L), mean (SD; 95% CI)	66.8 (14.4; 63.8-70.0)	68.9 (14.0; 64.6-73.1)	64.9 (14.6; 60.5-69.3)	.18
Uric acid (μmol/L), mean (SD; 95% CI)	410.7 (109.9; 387.5-433.8)	416.1 (112.5; 381.9-450.2)	405.4 (108.3; 372.9-437.9)	.62
Fasting plasma glucose (mmol/L), mean (SD; 95% CI)	5.3 (0.6; 5.2-5.5)	5.4 (0.6; 5.2-5.5)	5.3 (0.6; 5.1-5.5)	.54
Fasting insulin (pmol/L), mean (SD; 95% CI)	78.0 (39.3; 69.7-86.2)	80.1 (45.6; 66.2-94.0)	76.0 (32.4; 66.2-85.7)	.93
HOMA-IR[f], median (IQR)	2.4 (1.7-3.3)	2.4 (1.6-3.4)	2.4 (1.7-3.2)	.96
HOMA-IR≥2.5, n (%)	40 (44.9)	19 (43.2)	21 (46.7)	.74

XSL•FO
RenderX

Characteristic	Total (N=89)	Control (n=44)	Intervention (n=45)	*P* value[a]
Hemoglobin A_{1c} (%), median (IQR)	5.6 (5.4-5.8)	5.6 (5.2-5.8)	5.6 (5.4-5.9)	.36
TG[g] (mmol/L), median (IQR)	1.6 (1.0-2.0)	1.6 (1.0-2.2)	1.5 (1.0-1.8)	.43
TC[h] (mmol/L), median (IQR)	4.9 (4.2-5.3)	4.9 (4.2-5.4)	5.0 (4.3-5.4)	.78
HDL-C[i] (mmol/L), median (IQR)	1.2 (1.0-1.3)	1.1 (1.0-1.3)	1.2 (1.0-1.4)	.26
LDL-C[j] (mmol/L), median (IQR)	3.2 (2.6-3.7)	3.3 (2.6-3.8)	3.2 (2.5-3.6)	.68
CAP[k] (dB/m), median (IQR)	334 (295-365)	336 (288-362)	334 (297-369)	.25
Liver stiffness measurement (kPa), median (IQR)	5.1 (4.4-6.5)	5.2 (4.4-6.4)	4.9 (4.4-6.7)	.56
Diet quality score, median (IQR)	57 (50-66)	56 (49-66)	58 (51-66)	.64
Total MET[l]-minutes, median (IQR)	1116 (544-2057)	1118 (620-2068)	1102 (465-2088)	.59

[a]Continuous variables were assessed using unpaired t tests (normally distributed data) and Mann-Whitney *U* tests (non-normally distributed data) between groups. Categorical variables were analyzed using the chi-square test or Fisher exact test between groups.

[b]WC: waist circumference.

[c]ALT: alanine aminotransferase.

[d]AST: aspartate aminotransferase.

[e]GGT: γ-glutamyl transpeptidase.

[f]HOMA-IR: homeostasis model assessment-insulin resistance.

[g]TG: triglyceride.

[h]TC: total cholesterol.

[i]HDL-C: high-density lipoprotein-cholesterol.

[j]LDL-C: low-density lipoprotein-cholesterol.

[k]CAP: controlled attenuation parameter.

[l]MET: metabolic equivalent of task.

Changes in Clinical Parameters

Table 2 and Figure 2 present the clinical parameters at baseline and 24 weeks and their changes between groups in intention-to-treat population. There was a significant reduction in weight, waist circumference, waist-to-height ratio, and BMI from baseline within each group (*P*<.05). The reductions in these factors were also significantly greater in the intervention group than in the control group (*P*<.05). Patients in the TLC program intervention group had lower alanine aminotransferase (ALT), aspartate aminotransferase (AST), and γ-glutamyl transpeptidase (GGT) levels at 6 months than at baseline, and these liver enzyme levels decreased more in the intervention group than in the control group (*P*<.05). At 6 months, reductions in the levels of glucose metabolism–related factors, including fasting insulin, hemoglobin A_{1c} (HbA_{1c}), and HOMA-IR, were observed only within the intervention group (*P*<.05), and the levels of these factors showed significantly greater reductions in the intervention group than in the control group (*P*<.05). However, there was no between-group difference in FPG (*P*=.22). Patients in the intervention group had greater reductions in triglyceride (median difference −0.5, 95% CI −0.7 to −0.2; *P*=.001) and increases in high-density lipoprotein-cholesterol (HDL-C) levels than did those in the control group (median difference 0.1, 95% CI 0 - 0.1; *P*=.02). The diet quality score improved more in the TLC program group than in the control group (median difference 6, 95% CI 0 - 10; *P*=.04), whereas MET did not significantly differ between the groups (*P*=.92).

XSL•FO
RenderX

Table . Clinical parameters between the two study groups at baseline and 24 weeks in the intention-to-treat population.

Parameter	Control (n=44)			Intervention (n=45)			Mean or Hodges-Lehmann median difference (95% CI)[a]	*P* value[a]
	Baseline	24 weeks	*P* value[b]	Baseline	24 weeks	*P* value[b]		
Weight (kg), mean (SD; 95% CI)	82.3 (11.2; 78.9-85.7)	81.4 (12.2 ;77.7-85.0)	.02	82.7 (12.1; 79.1-86.4)	77.2 (11.7; 73.7-80.8)	<.001	–3.7 (–5.5 to –1.7)	<.001
BMI (kg/m^2), mean (SD; 95% CI)	28.2 (3.0; 27.2-29.1)	27.5 (3.4; 26.5-28.6)	.004	29.0 (2.9; 28.1-29.8)	26.9 (2.9; 26.0-27.8)	<.001	–1.4 (–2.1 to –0.7)	<.001
WC[c](cm), mean (SD; 95% CI)	92.6 (8.9; 89.9-95.3)	90.6 (8.6; 88.0-93.2)	.003	93.0 (9.0; 90.3-95.7)	85.3 (9.5; 82.7-88.4)	<.001	–5.0 (–7.0 to –2.0)	<.001
Waist to hip ratio, mean (SD; 95% CI)	0.91 (0.06; 0.89-0.93)	0.90 (0.06; 0.89-0.92)	.21	0.90 (0.07; 0.88-0.92)	0.87 (0.07; 0.85-0.89)	<.001	–0.02 (–0.04 to 0.01)	.06
Waist to height ratio, mean (SD; 95% CI)	0.54 (0.05; 0.53-0.56)	0.53 (0.05; 0.52-0.55)	.003	0.55 (0.04; 0.54-0.56)	0.50 (0.05; 0.49-0.52)	<.001	–0.03 (–0.04 to –0.02)	<.001
Systolic blood pressure (mmHg), mean (SD; 95% CI)	124.4 (6.7; 122.4-126.5)	125.6 (9.9; 122.6-128.6)	.98	123.3 (7.0; 121.2-125.4)	123.4 (11.3; 120.0-126.8)	.17	–1.0 (–6.0 to 4.0)	.87
Diastolic blood pressure (mmHg), mean (SD; 95% CI)	79.6 (5.9; 77.9-81.4)	81.1 (5.9; 79.3-82.9)	.18	79.0 (5.6; 77.3-80.7)	79.6 (5.8; 77.8-81.3)	.47	–0.9 (–3.7 to 1.9)	.62
Platelet count (10^9/L), mean (SD; 95% CI)	254.5 (51.9; 238.7-270.3)	256.0 (66.5; 235.7-276.2)	.92	258.8 (51.9; 243.2-274.4)	250.0 (44.3; 237.0-262.0)	.03	–12.8 (–26.7 to 1.1)	.07
ALT[d] (U/L), median (IQR)	38.0 (24.5-57.3)	30.0 (22.3-44.3)	.01	34.0 (26.5-61.0)	23.0 (17.0-33.5)	<.001	–8.0 (–15.0 to –1.0)	.02
AST[e] (U/L), median (IQR)	26.0 (21.3-35.5)	22.0 (19.0-27.8)	.003	25.0 (19.5-36.0)	19.0 (15.0-23.0)	<.001	–3.0 (–6.0 to 0)	.04
GGT[f] (U/L), median (IQR)	39.0 (25.3-59.8)	41.0 (25.3-57.5)	.76	34.0 (22.0-50.0)	23.0 (16.5-34.5)	<.001	–8.0 (–14.0 to –3.0)	<.001
Albumin (g/L), mean (SD; 95% CI)	46.6 (2.0; 46.0-47.2)	45.7 (2.1; 44.8-46.0)	.001	45.9 (2.6; 45.1-46.7)	45.1 (4.0; 43.9-46.3)	.15	0.4 (–0.8 to 1.6)	.06
Urea (mmol/L), mean (SD; 95% CI)	5.0 (1.1; 4.6-5.3)	5.0 (1.2; 4.6-5.4)	.40	4.9 (1.0; 4.6-5.2)	5.0 (1.1; 4.7-5.3)	.61	0.1 (–0.3 to 0.4)	.95
Serum creatinine (µmol/L), mean (SD; 95% CI)	68.9 (14.0; 64.6-73.1)	68.7 (13.8; 64.5-72.9)	.85	64.9 (14.6; 60.5-69.3)	64.5 (14.0; 60.3-68.7)	.78	–0.2 (–2.2 to 1.8)	.79

XSL•FO
RenderX

Parameter	Control (n=44)			Intervention (n=45)			Mean or Hodges-Lehmann median difference (95% CI)[a]	*P* value[a]
	Baseline	24 weeks	*P* value[b]	Baseline	24 weeks	*P* value[b]		
Uric acid (μmol/L), mean (SD; 95% CI)	416.1 (112.5; 381.9-450.2)	404.1 (75.2; 380.4-427.2)	.57	405.4 (108.3; 372.9-437.9)	386.8 (105.6; 355.0-418.5)	.008	–6.6 (–36.1 to 22.8)	.16
Fasting plasma glucose (mmol/L), mean (SD; 95% CI)	5.4 (0.6; 5.2-5.5)	5.5 (0.9; 5.3-5.8)	.39	5.3 (0.6; 5.1-5.5)	5.3 (0.5; 5.1-5.4)	.24	–0.1 (–0.5 to 0.3)	.22
Fasting insulin (pmol/L), mean (SD; 95% CI)	80.1 (45.6; 66.2-94.0)	81.0 (40.8; 68.6-93.4)	.35	76.0 (32.4; 66.2-85.7)	60.1 (34.8; 49.6-70.6)	.001	–16.8 (–30.0 to –3.5)	.002
HOMA-IR[g], median (IQR)	2.4 (1.6-3.4)	2.7 (1.9-3.7)	.25	2.4 (1.7-3.2)	1.7 (1.1-2.8)	.001	–0.7 (–1.1 to –0.2)	.002
HbA_{1c}[h] (%), median (IQR)	5.6 (5.2-5.8)	5.5 (5.3-5.7)	.58	5.6 (5.4-5.9)	5.4 (5.3-5.7)	.003	–0.1 (–0.2 to 0)	.047
TG[i] (mmol/L), median (IQR)	1.6 (1.0-2.2)	1.9 (1.2-2.4)	.09	1.5 (1.0-1.8)	1.0 (0.8-1.5)	.001	–0.5 (–0.7 to –0.2)	.001
TC[j] (mmol/L), median (IQR)	4.9 (4.2-5.4)	4.6 (4.0-5.2)	.15	5.0 (4.3-5.4)	4.5 (3.9-5.1)	.009	–0.1 (–0.4 to 0.1)	.32
HDL-C[k] (mmol/L), median (IQR)	1.1 (1.0-1.3)	1.1 (0.9-1.3)	.21	1.2 (1.0-1.4)	1.3 (1.0-1.5)	.07	0.1 (0-0.1)	.02
LDL-C[l] (mmol/L), median (IQR)	3.3 (2.6-3.8)	3.1 (2.6-3.6)	.06	3.2 (2.5-3.6)	3.1 (2.3-3.5)	.04	–0.1 (–0.2 to 0.1)	.75
CAP[m] (dB/m), median (IQR)	336 (288-362)	310 (288-342)	.04	334 (297-369)	291 (239-317)	<.001	–38 (–62 to –16)	.001
CAP<268, n (%)	4 (9.1)	6 (13.6)	.55	1 (2.2)	16 (35.6)	<.001	—[n]	—
268≤CAP<280, n (%)	4 (9.1)	3 (6.8)	—	1 (2.2)	4 (8.9)	—	—	—
CAP≥280, n (%)	36 (81.8)	35 (79.5)	—	43 (95.6)	25 (55.6)	—	—	—
Fatty liver index, mean (SD; 95% CI)	58.4 (22.3; 51.6-65.2)	57.1 (22.2; 50.3-63.8)	.38	56.7 (21.8; 50.2-63.3)	34.2 (22.3; 27.5-40.9)	<.001	–21.1 (–28.4 to –13.9)	<.001
Liver stiffness measurement (kPa), median (IQR)	5.2 (4.4-6.4)	5.1 (4.1-6.8)	.07	4.9 (4.4-6.7)	5.1 (4.0-6.0)	.16	–0.1 (–0.6 to 0.5)	.94
AST-to-platelet ratio index, median (IQR)	0.28 (0.20-0.37)	0.24 (0.18-0.30)	.008	0.27 (0.19-0.35)	0.21 (0.16-0.27)	<.001	–0.03 (–0.06 to 0.01)	.13
Fibrosis-4 index, median (IQR)	0.68 (0.50-1.00)	0.66 (0.46-0.99)	.20	0.64 (0.49-0.90)	0.62 (0.49-0.76)	.15	–0.05 (–0.12 to 0.03)	.85

XSL•FO
RenderX

Parameter	Control (n=44)			Intervention (n=45)			Mean or Hodges-Lehmann median difference (95% CI)[a]	*P* value[a]
	Baseline	24 weeks	*P* value[b]	Baseline	24 weeks	*P* value[b]		
NAFLD[o] fibrosis score, mean (SD; 95% CI)	–2.9 (1.2; –3.2 to –2.5)	–2.6 (1.2; –3.0 to –2.3)	.09	–2.9 (1.0; –3.2 to –2.6)	–2.7 (1.0; –3.0 to –2.4)	.15	–0.1 (–0.4 to 0.3)	.96
FAST[p], median (IQR)	0.20 (0.11-0.42)	0.11 (0.07-0.22)	.001	0.22 (0.09-0.43)	0.06 (0.03-0.15)	<.001	–0.06 (–0.13 to –0.01)	.02
Agile3+, median (IQR)	0.06 (0.02-0.15)	0.05 (0.02-0.15)	.96	0.05 (0.02-0.12)	0.05 (0.03-0.12)	.63	0.01 (–0.02 to 0.03)	.68
Diet quality score, median (IQR)	56 (49-66)	64 (55-71)	.01	58 (51-66)	66 (61-76)	<.001	6 (0-10)	.04
Total MET[q]-minutes, median (IQR)	1118 (620-2068)	1442 (620-2608)	.29	1102 (465-2088)	1102 (647-1725)	.44	–10 (–204 to 192)	.92

[a]Continuous variables were assessed using 2-tailed unpaired *t* tests (reporting mean difference and 95% CI) for normally distributed data and Mann-Whitney *U* tests (reporting Hodges-Lehmann median difference and 95% CI) for non-normally distributed data to compare the changes between the two groups.

[b]Continuous variables were assessed using paired *t* tests (normally distributed data) or Wilcoxon signed-rank tests (non-normally distributed data) within each group. Categorical variables were analyzed using marginal homogeneity tests within each group.

[c]WC: waist circumference.

[d]ALT: alanine aminotransferase.

[e]AST: aspartate aminotransferase.

[f]GGT: γ-glutamyl transpeptidase.

[g]HOMA-IR: homeostasis model assessment-insulin resistance.

[h]HbA_{1c}: hemoglobin A_{1c}.

[i]TG: triglyceride.

[j]TC: total cholesterol.

[k]HDL-C: high-density lipoprotein-cholesterol.

[l]LDL-C: low-density lipoprotein-cholesterol.

[m]CAP: controlled attenuation parameter.

[n]Not applicable.

[o]NAFLD: nonalcoholic fatty liver disease.

[p]FAST: FibroScan-aspartate aminotransferase.

[q]MET: metabolic equivalent of task.

XSL•FO
RenderX

Figure 2. Change in clinical parameters from baseline to 24 weeks between control group and intervention group. (A) Median change in weight between groups (–0.3 vs –4.4; median difference –3.7, 95% CI –5.5 to –1.7; *P*<.001). (B) Median change in waist circumference between groups (–1.5 vs –7.0; median difference –5.0, 95% CI –7.0 to –2.0; *P*<.001). (C) Median change in alanine aminotransferase values between groups (–2.5 vs –12.0; median difference –8.0, 95% CI –15.0 to –1.0; *P*=.02). (D) Median change in aspartate aminotransferase values between groups (–2.0 vs –6.0; median difference –3.0, 95% CI –6.0 to 0; *P*=.04). (E) Median change in hemoglobin A_{1c} between groups (0 vs –0.1; median difference –0.1, 95% CI –0.2 to 0; *P*=.047). (F) Median change in homeostasis model assessment-insulin resistance between groups (0 vs –0.6; median difference –0.7, 95% CI –1.1 to –0.2; *P*=.002). (G) Median change in triglyceride values between groups (0 vs –0.3; median difference –0.5, 95% CI –0.7 to –0.2; *P*=.001). (H) Median change in high-density lipoprotein-cholesterol values between groups (0 vs 0.1; median difference 0.1, 95% CI 0-0.1; *P* =.02). (I) Median change in controlled attenuation parameter values between groups (–4.0 vs –56.0; median difference –38, 95% CI –62 to –16; *P*=.001). (J) Median change in FibroScan-aspartate aminotransferase scores between groups (–0.04 vs –0.11; median difference –0.06, 95% CI –0.13 to –0.01; *P* =.02). ALT: alanine aminotransferase; AST: aspartate aminotransferase; CAP: controlled attenuation parameter; FAST: FibroScan–aspartate aminotransferase; HbA_{1c}: hemoglobin A_{1c}; HDL-C: high-density lipoprotein cholesterol; HOMA-IR: homeostasis model assessment–insulin resistance; TG: triglyceride; WC: waist circumference.

XSL•FO
RenderX

Changes in Noninvasive Factors of Hepatic Steatosis and Liver Fibrosis

At 6 months, both groups presented a reduction in the CAP value. The intervention group experienced a decrease to 291(239-317) dB/m, whereas the control group had a median CAP of 310 (288-342)dB/m (Table 2, *P*=.001). CAP values were significantly lower in the intervention group than in the control group (Figure 2, median difference −38, 95% CI −62 to −16; *P*=.001). In the subgroup analysis, the percentage of patients with severe steatosis decreased significantly in the intervention group from baseline at 6 months (*P*<.001). However, the LSM values were not significantly different between groups or within each group (*P*>.05). Based on the noninvasive assessment, patients in the TLC program intervention group had lower FAST scores than did those in the control group (median difference −0.06, 95% CI −0.13 to −0.01; *P*=.02). In the comparison of these 2 groups, no significant differences were found in noninvasive scores, including the APRI, FIB-4, NFS, and Agile 3+ score (*P*>.05).

Changes in Body Composition

According to the body composition analysis (Table 3), each group presented a decrease in total body fat mass, total body fat percentage, and visceral fat level, and these variables decreased more in the intervention group than in the control group across the 6-month period (*P*<.05). The basal metabolic rate underwent a greater reduction in the intervention group than in the control group (mean difference −42, 95% CI −62 to −22; *P*<.001). The intervention group also presented more pronounced decreases in fat-free mass, total lean body mass, and appendicular skeletal muscle mass (ASM; *P*<.05). The ratio of ASM to weight increased in the intervention group, but no significant difference was observed in this ratio between the groups (*P*>.05).

XSL•FO
RenderX

Table . Body composition analysis between the two study groups at baseline and 24 weeks in the intention-to-treat population.

Body composition	Control n=44			Intervention n=45			Mean or Hodges-Lehmann median difference (95% CI)[a]	*P* value[a]
	Baseline	24weeks	*P* value[b]	Baseline	24 weeks	*P* value[b]		
Total body fat mass (kg), mean (SD; 95% CI)	25.3 (6.8; 23.3-27.4)	24.3 (7.0; 22.2-26.4)	.006	28.3 (7.3; 26.2-30.6)	24.6 (7.3; 22.4-26.7)	<.001	–2.8 (–4.4 to –1.2)	<.001
Total body fat percentage (%), mean (SD; 95% CI)	30.7 (6.9; 28.6-32.8)	29.9 (6.8; 27.9-32.0)	.002	34.3 (7.8; 32.0-36.7)	30.9 (8.5; 28.3-33.5)	<.001	–2.6 (–4.1 to –1.2)	<.001
Visceral fat level, median (IQR)	13 (11-16)	12 (10-14)	.001	12 (10-15)	10 (8-13)	<.001	–1 (–2 to 0)	<.001
Fat-free mass (kg), mean (SD; 95% CI)	57.0 (9.1; 54.2-59.7)	57.1 (9.7; 54.1-60.0)	.44	54.3 (10.6; 51.1-57.5)	52.7 (10.1; 49.6-55.7)	<.001	–1.1 (–1.9 to –0.2)	.01
Total lean body mass (kg), mean (SD; 95% CI)	53.4 (9.2; 50.6-56.1)	52.7 (9.3; 49.9-55.5)	.29	51.0 (10.0; 48.0-54.0)	49.9 (9.9; 47.0-52.9)	<.001	–0.7 (–1.2 to –0.2)	.002
ASM[c] (kg), mean (SD; 95% CI)	25.5 (5.4; 23.8-27.1)	25.0 (5.8; 23.2-26.8)	.70	24.8 (5.9; 23.0-26.6)	23.9 (5.7; 22.2-25.6)	<.001	–0.6 (–1.0 to –0.3)	.001
ASM to weight ratio (%), mean (SD; 95% CI)	30.7 (3.8; 29.6-31.9)	30.8 (5.3; 29.1-32.3)	.009	29.8 (4.4; 28.5-31.1)	30.8 (4.5; 29.5-32.1)	<.001	0.5 (–0.1 to 1.1)	.08
Trunk muscle mass (kg), mean (SD; 95% CI)	27.9 (4.1; 26.6-29.1)	27.7 (4.1; 26.5-29.0)	.31	26.2 (4.5; 24.9-27.6)	26.0 (4.4; 24.7-27.3)	.06	–0.1 (–0.5 to 0.1)	.27
Total body water (kg), mean (SD; 95% CI)	38.1 (5.3; 36.5-39.7)	38.2 (5.4; 36.6-39.8)	.54	37.7 (5.6; 36.1-39.5)	36.9 (5.6; 35.2-38.6)	.001	–0.9 (–1.5 to –0.4)	.003
Total body water percentage (%), mean (SD; 95% CI)	47.0 (3.9; 45.8-48.1)	47.1 (5.1; 45.5-48.6)	.07	46.0 (4.0; 44.9-47.3)	47.8 (5.2; 46.3-49.4)	<.001	1.6 (0.2-3.0)	.001
Extracellular water percentage (%), mean (SD; 95% CI)	41.2 (2.1; 40.5-41.8)	40.9 (2.0; 40.3-41.5)	.003	41.4 (2.2; 40.7-42.0)	40.8 (2.0; 40.2-41.4)	.001	–0.2 (–0.6 to 0.1)	.10
BMC[d] (kcal), mean (SD; 95% CI)	1631 (252; 1555-1708)	1621 (251; 1545-1698)	.27	1597 (273; 1515-1679)	1545 (274; 1463-1627)	<.001	–42 (–62 to –22)	<.001

[a]Continuous variables were assessed using unpaired *t* tests (reporting mean difference and 95% CI) for normally distributed data and Mann-Whitney *U* tests (reporting Hodges-Lehmann median difference and 95% CI) for non-normally distributed data to compare the changes between the two groups.

[b]Continuous variables were assessed using paired *t* tests (normally distributed data) or Wilcoxon signed-rank tests (non-normally distributed data) within each group.

[c]ASM: appendicular skeletal muscle mass.

[d]BMC: basal metabolic rate.

XSL•FO
RenderX

Analysis of Primary and Secondary Efficacy Outcomes

As shown in Figure 3, the primary and secondary efficacy outcomes were compared between the control and intervention groups. According to the ITT analysis Figure 3, 60% (27/45) of patients achieved weight loss of ≥5% (27/45 vs 7/44; RR 3.771, 95% CI 1.836 - 7.748; *P*<.001), and 24.4 % (11/45) of patients attained weight loss of ≥10% among patients in the TLC program-delivered intervention group, which was greater than those in the control group (11/45 vs 3/44; RR 3.585, 95% CI 1.072 - 11.988; *P*=.02). A CAP value of ≥10% reduction or normalization was achieved in more patients who received the intervention than in those who received standard care (26/45 vs 14/44; RR 1.816, 95% CI 1.102 - 2.992; *P*=.01). Furthermore, patients in the intervention group were more likely than those in the control group to achieve an ALT reduction of ≥50% or normalization (30/45 vs 16/44; RR 1.833, 95% CI 1.178 - 2.853; *P*=.004), as well as triglyceride normalization (37/45 vs 20/44; RR 1.809, 95% CI 1.273 - 2.570; *P*<.001). Similarly, these results from the PP analysis were in line with those from the ITT analysis (Figure 3).

The effects of the TLC program on weight loss and the CAP value were also analyzed by multivariate regression (Table 4). Among the ITT population or PP population, the intervention had a positive effect on weight loss of ≥5% or ≥10% after adjusting for age, sex, and baseline weight (*P*<.05). Additionally, patients in the intervention group were more likely to achieve a CAP reduction of ≥10% or normalization after adjusting for age, sex, and baseline CAP in both the ITT population (OR 2.853, 95% CI 1.092 - 7.456; *P*=.03) and the PP population (OR 3.319, 95% CI 1.117 - 9.860; *P*=.03).

Figure 3. Analysis of primary and secondary efficacy outcome between control group and intervention group. (A) The percentage of patients with weight loss ≥5% between groups in the intention-to-treat population (27/45 vs 7/44, RR 3.771, 95% CI 1.836-7.748; *P*<.001) and the per-protocol population (26/37 vs 5/36, RR 5.059, 95% CI 2.184-11.719; *P* <.001) ; (B) The percentage of patients with weight loss ≥10% between groups in the intention-to-treat population (11/45 vs 3/44, RR 3.585, 95% CI 1.072-11.988; *P*=.02) and the per-protocol population (10/37 vs 3/36, RR 3.243, 95% CI 1.066-10.360; *P*=.04) ; (C) The percentage of patients with controlled attenuation parameter ≥10% reduction or normalization between groups in the intention-to-treat population (26/45 vs 14/44, RR 1.816, 95% CI 1.102-2.992; *P*=.01) and the per-protocol population (25/37 vs 13/36, RR 1.871, 95% CI: 1.148-3.050; *P*=.007; (D) The percentage of patients with alanine aminotransferase ≥50% reduction or normalization between groups in the intention-to-treat population (30/45 vs 16/44, RR 1.833, 95% CI 1.178-2.853; *P*=.004) and the per-protocol population (25/37 vs 16/36, RR 1.520, 95% CI 1.006-2.391; *P*=.047; (E) The percentage of patients with triglyceride normalization between groups in the intention-to-treat population (37/45 vs 20/44, RR 1.809, 95% CI 1.273-2.570; *P*<.001) and the per-protocol population (32/37 vs 17/36, RR 1.831, 95% CI 1.267-2.646; *P*<.001). ALT: aspartate aminotransferase; CAP: controlled attenuation parameter; ITT: intention-to-treat; PP: per-protocol; TG: triglyceride.

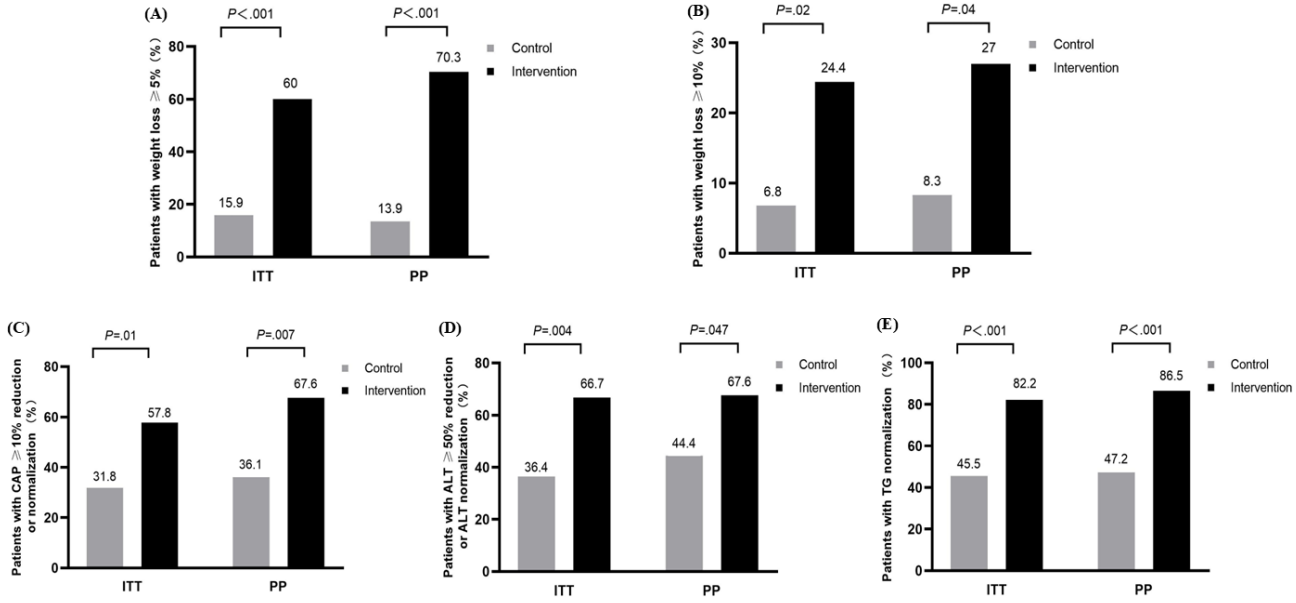

XSL•FO
RenderX

Table . Multivariate logistic analysis model of reduction in weight and controlled attenuation parameter value in intention-to-treat and per-protocol population.

Parameter	Unadjusted	Model 1[a]	Model 2[b]	Unadjusted	Model 1	Model 2
	ITT[c] population (N=89)			PP[d] population (N=73)		
Weight loss≥5%						
Control	reference	reference	reference	reference	reference	reference
Intervention, OR (95% CI)	7.929 (2.905-21.641)	8.537 (2.947-24.730)	8.380 (2.886-24.331)	14.655 (4.509-47.626)	14.534 (4.267-49.499)	15.231 (4.380-52.969)
P value	＜.001	＜.001	＜.001	＜.001	＜.001	＜.001
Weight loss ≥10%						
Control	reference	reference	reference	reference	reference	reference
Intervention, OR[e] (95% CI)	4.422 (1.140 - 17.144)	4.677 (1.160 - 18.855)	4.612 (1.138 - 18.686)	4.074 (1.018 - 16.305)	4.604 (1.093 - 19.386)	4.571 (1.078 - 19.387)
P value	.03	.03	.03	.047	.04	.04
CAP[f]≥10% reduction or normalization						
Control	reference	reference	reference	reference	reference	reference
Intervention, OR (95% CI)	2.932 (1.232-6.981)	2.843 (1.161-6.965)	2.853 (1.092-7.456)	3.686 (1.401-9.700)	3.399 (1.259-9.175)	3.319 (1.117-9.860)
P value	.02	.02	.03	.008	.02	.03

[a]Model 1: Adjusted for age, sex.

[b]Model 2: Adjusted for model 1 and baseline weight.

[c]ITT: intention-to-treat.

[d]PP: per-protocol population.

[e]OR: odds ratio.

[f]Model 2： Adjusted for model 1 and baseline controlled attenuation parameter value; CAP: controlled attenuation parameter.

Safety

During the study period, the intervention was safe in this cohort of patients with MAFLD, with no adverse events reported during the study period. The heart rate monitoring during the video exercises effectively protected participants from the risk of overexertion.

Discussion

Principal Findings

MAFLD is a public health concern and is associated with unhealthy lifestyles. Effective weight control is a crucial measure for reversing MASH and attenuating metabolic dysfunction [11,33]. While digital health tools offer a scalable solution, there is a notable scarcity of interventions specifically designed for the population of individuals with MAFLD in China. Therefore, we conducted a longitudinal study to evaluate an innovative WeChat mini-program-based lifestyle modification in patients with MAFLD. Our study revealed that the TLC program intervention was more effective for weight loss among patients with MAFLD who were overweight or obese than standard care was. Importantly, CAP values decreased significantly more in the intervention group than in the control group, suggesting that the TLC program intervention ameliorated hepatic steatosis. Along with weight reduction, the intervention group presented significant decreases in liver enzymes, HbA_{1c}, and triglyceride levels, indicating improved liver function and metabolic-related indicators. To investigate the changes in body composition, we found that the intervention group had a lower total body fat percentage and visceral fat level than did the control group, with parallel improvement in hepatic steatosis. In terms of liver fibrosis assessment, only the FAST score showed improvement in MASH with significant fibrosis after the intervention, whereas other fibrosis indicators, such as LSM and noninvasive scores (APRI, FIB-4, NFS, Agile 3+) showed no treatment effect.

Based on established evidence, DTx lifestyle interventions lead to significant weight loss in patients with MAFLD [6,30]. According to a meta-analysis of 8 cohort studies on DTx, 33% of patients with MAFLD experienced significant weight loss of ≥5% [13]. In our study, more than half (27/45, 60%) of the patients with MAFLD achieved a weight loss of ≥5% , demonstrating the efficacy of weight management over 6 months. The significant improvement in diet quality scores observed in the intervention group implies a potential key role of dietary modification within the TLC program. With the reduction in body weight, there was a significant improvement in liver chemistry variables, including ALT, AST, and GGT, which was consistent with published data [30,34]. In particular, approximately two-thirds of the patients with TLC intervention experienced an ALT reduction of more than 50% or

XSL•FO
RenderX

normalization, suggesting that this digital intervention improved liver function remarkably.

Importantly, 57.8% (26/45) of the patients in the intervention group presented a CAP value of more than a 10% reduction or normalization, indicating that the severity of hepatic steatosis was highly reduced among individuals after intervention. A similar study investigated a mobile technology–based intervention program for patients with MAFLD and demonstrated that 42.4% (14/33) of participants exhibited a reduction in CAP values, with 21.2% (7/33) achieving a significant decrease of ≥10% in the CAP value [35]. Based on our study, the reduction in body weight and visceral fat led to a marked decline in hepatic fat content and subsequent improvement in liver function.

Among patients with MAFLD, the severity of fibrosis is a major disease modifier and affects hepatic outcomes [36,37]. Gradual weight loss of more than 10% may ameliorate fibrosis in patients with MAFLD [38]. In our study, although the proportion of patients with weight loss greater than 10% reached 24.4% after intervention, the LSM value was not significantly different between the groups. Similarly, no significant difference was found in noninvasive scores, such as the APRI, FIB-4 score, NFS, and Agile 3+ score, except for the FAST score. With respect to noninvasive assessment of liver fibrosis, current evidence from different studies has shown inconsistencies. A German web-based exercise program intervention for patients with MAFLD reported improved APRI, FIB-4 score, and LSM, whereas a US mobile-based program intervention for patients with biopsy-confirmed MAFLD showed no changes in NFS or FIB-4 score [39,40]. The observed discrepancies across studies may stem from variations in intervention protocols, differential durations of implementation, and heterogeneity in patient adherence levels. Based on our data, the overall baseline LSM values of the patients enrolled seem to be remarkably low, with a median value of only 5.1 (IQR 4.4- 6.5) kPa. Furthermore, the duration of the intervention was only 6 months, which is too short for observing changes in fibrosis.

FAST scores reflect MASH with significant fibrosis and are calculated via the CAP, LSM, and AST. In our study, the observed reduction in the FAST score in the intervention group may be associated with the significant decreases in CAP and AST levels, suggesting potential alleviation of MASH to some extent by using the TLC program. Undoubtedly, the definitive evidence of MASH alleviation is contingent upon further study with patients who have undergone liver biopsies.

MAFLD results in impaired metabolic profiles in multiple organ systems, which not only promotes the progression of liver disease but also increases the risk of extrahepatic complications [41,42]. In the current investigation, we assessed key metrics of glucose metabolism, observing substantially greater reductions in fasting insulin, HbA_{1c}, and HOMA-IR levels among intervention group participants, which aligns with established findings in the literature [35,39]. Nevertheless, there was no significant difference in the FPG level between these two groups. It is important to emphasize that our cohort consisted predominantly of nondiabetic individuals, with an overall low baseline glucose (5.3 mmol/L), which limited the potential for a large absolute reduction. Interestingly, some studies concerning DTx-mediated lifestyle modifications in MAFLD management have reported similar results [34,40]. Indeed, insulin resistance is the key pathogenic link between obesity and associated metabolic disorders. Moreover, it is the primary driver in the development of MAFLD and cardiovascular disease. A significant reduction in body weight has favorable effects on increasing insulin sensitivity and improving insulin resistance, leading to a reduction in fasting insulin and glycemic fluctuations, substantially reducing cardiovascular events in patients with MAFLD [43]. A reduction in the HbA_{1c} level, reflecting good glycemic control over the past 2‐3 months, may lead to decreased hepatic fat accumulation, reduced liver fat synthesis, and increased fat breakdown [44].

For the lipid profile, decreased TG or elevated HDL-C levels were found in the intervention group, which was in line with previously published literature [35,45,46]. Favorable modifications in TG or HDL profiles have been correlated with the amelioration of insulin resistance, which not only upregulates lipoprotein lipase expression and accelerates TG hydrolysis but also downregulates hepatic lipase activity, reduces HDL-C degradation, and promotes HDL-C synthesis [41,47]. Overall, TLC digital intervention markedly improved metabolic indices, such as glucose and lipid levels, which are pivotal determinants of cardiovascular morbidity and mortality in patients with MAFLD.

To evaluate the quality of weight loss, we explored parameters involved in body composition. Compared with the control group, the intervention group presented greater reductions in total body fat percentage and visceral fat, as well as the alleviation of hepatic steatosis. Although the intervention group demonstrated significant reductions in total lean body mass and absolute ASM, the ratio of ASM to weight increased within the intervention group, suggesting that weight loss was driven primarily by fat reduction. As is well known, healthy weight loss involves fat loss with muscle gain. Furthermore, no significant difference in MET levels was detected between the groups in our study. To build muscle and improve body composition, future improvements to the TLC program should integrate progressive strength training, such as bodyweight exercises, resistance bands, and dumbbells. Monitoring changes in muscle mass using body fat scales can further support the personalization and optimization of these interventions.

Strengths and Limitations

This study has several strengths. Its innovation lies in being the first tailored digital tool to address the specific self-management challenges of this population in China, leveraging a widely accessible platform. Additionally, this mobile-based program offers customized dietary plans based on each patient's health status. Notably, this smartphone-delivered lifestyle program aids in weight control and liver fat reduction by facilitating healthier behaviors, which could be implemented in home settings. This high accessibility removes common barriers to care, such as transportation and time constraints, while its timely, personalized feedback promotes greater adherence and sustained habit formation. Together, these advantages make it

XSL•FO
RenderX

a scalable and cost-effective strategy for the long-term management of MAFLD.

This study has several limitations. First, all the patients enrolled were from a single tertiary hepatology clinic, and more motivated individuals tended to self-select into the program. This selection bias was inevitable. Second, hepatic steatosis and fibrosis were evaluated by FibroScan or noninvasive tests rather than liver biopsy. On the one hand, FibroScan is a widely used and noninvasive tool to detect hepatic steatosis and screen for fibrosis. On the other hand, it is extremely challenging to obtain liver biopsy specimens in clinical practice. Third, no postintervention follow-up was conducted to assess the maintenance of weight loss. The lack of long-term follow-up data limits the conclusions regarding the sustainability of the WeChat-based intervention effects. Future research will include a follow-up phase where the research team will formulate personalized weight-management plans and encourage the involvement of family or friends for support and supervision to facilitate long-term weight maintenance for participants. Additionally, the sample size herein was limited, and the dropout rate was relatively high (18%). The application of last observation carried forward for handling missing data may have introduced bias into the treatment effect estimate. These factors, collectively, may have contributed to the imprecision in our effect estimates, as evidenced by the wide CIs. Subsequent study should incorporate larger, multicenter cohorts and advanced modeling approaches. Furthermore, the mini-program itself will be further developed to integrate incentives, gamification, and social support. This comprehensive strategy aims to enhance user adherence, boost engagement, and reduce dropout rates.

Conclusions

Our study introduces a WeChat mini-program that delivers a tailored lifestyle intervention designed for patients with MAFLD in China, addressing a gap in scalable management. Our intervention is characterized by its ability to not only reduce weight but also to significantly improve hepatic steatosis and metabolic measures. These findings contribute to the field by establishing a feasible, low-cost model for managing MAFLD in high-burden populations, with strong potential for implementation in primary care and resource-limited settings. Given the imprecision of effect estimates, further larger-scale studies are essential to confirm these preliminary findings, obtain more precise estimates, and evaluate sustained long-term clinical outcomes with long-term follow-up.

Acknowledgments

This study was supported by grants from Noncommunicable Chronic Diseases-National Science and Technology Major Project (2023ZD0508704), the Construction Project of the “Discipline Peak-Climbing Plan” from Xinhua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine (XKPF2024B400). The authors thank all the participants and clinicians who participated in our study. We would like to thank Hangzhou Jianhai Technology Company for providing supportive digital technology.

Artificial intelligence (AI) was not used in any portion of the manuscript writing.

Funding

The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated and analyzed in the current study are available from the corresponding author upon reasonable request.

Authors' Contributions

SC and FJG contributed to the conceptualization of the study. SC, SCC, CHX, YRX, ZJ, and DXY were responsible for the investigation and data curation. CGY and SX guided the methodology. FJG and SX provided resources and secured funding. SC, SX, and FJG were responsible for the program development, project administration, and supervision. SC and FJG contributed to formal analysis and writing-original draft. SC, CGY, SCC, CHX, YRX, ZJ, DXY SX, and FJG carried out validation, writing-review, and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Institutional Review Board approval document.
[PDF File, 2513 KB - jmir_v28i1e76204_app1.pdf]

Checklist 1
CONSORT eHealth checklist (V1.6.1).
[PDF File, 37271 KB - jmir_v28i1e76204_app2.pdf]

XSL•FO
RenderX

References

1. Eslam M, Fan JG, Yu ML, et al. The Asian Pacific association for the study of the liver clinical practice guidelines for the diagnosis and management of metabolic dysfunction-associated fatty liver disease. Hepatol Int 2025 Apr;19(2):261-301. [doi: 10.1007/s12072-024-10774-3] [Medline: 40016576]
2. European Association for the Study of the Liver (EASL), European Association for the Study of Diabetes (EASD), European Association for the Study of Obesity (EASO). EASL-EASD-EASO Clinical Practice Guidelines on the Management of Metabolic Dysfunction-Associated Steatotic Liver Disease (MASLD). Obes Facts 2024;17(4):374-444. [doi: 10.1159/000539371] [Medline: 38852583]
3. Fan JG, Xu XY, Yang RX, et al. Guideline for the Prevention and Treatment of Metabolic Dysfunction-associated Fatty Liver Disease (Version 2024). J Clin Transl Hepatol 2024 Nov 28;12(11):955-974. [doi: 10.14218/JCTH.2024.00311] [Medline: 39544247]
4. Seeberg KA, Borgeraas H, Hofsø D, et al. Gastric bypass versus sleeve gastrectomy in type 2 diabetes: effects on hepatic steatosis and fibrosis: a randomized controlled trial. Ann Intern Med 2022 Jan;175(1):74-83. [doi: 10.7326/M21-1962] [Medline: 34843380]
5. Rinella ME, Neuschwander-Tetri BA, Siddiqui MS, et al. AASLD Practice Guidance on the clinical assessment and management of nonalcoholic fatty liver disease. Hepatology 2023 May 1;77(5):1797-1835. [doi: 10.1097/HEP.0000000000000323] [Medline: 36727674]
6. Pfirrmann D, Huber Y, Schattenberg JM, Simon P. Web-based exercise as an effective complementary treatment for patients with nonalcoholic fatty liver disease: intervention study. J Med Internet Res 2019 Jan 2;21(1):e11250. [doi: 10.2196/11250] [Medline: 30602434]
7. Zhou R, Zhang B, Zhang W, et al. Self-management behaviours in adults with non-alcoholic fatty liver disease: a cross-sectional survey from China. BMJ Open 2024 Feb 22;14(2):e078333. [doi: 10.1136/bmjopen-2023-078333] [Medline: 38388505]
8. Tincopa MA, Patel N, Shahab A, Asefa H, Lok AS. Implementation of a randomized mobile-technology lifestyle program in individuals with nonalcoholic fatty liver disease. Sci Rep 2024 Mar 28;14(1):7452. [doi: 10.1038/s41598-024-57722-7] [Medline: 38548875]
9. Zhou R, Gu Y, Zhang B, et al. Digital therapeutics: emerging new therapy for nonalcoholic fatty liver disease. Clin Transl Gastroenterol 2023 Apr 1;14(4):e00575. [doi: 10.14309/ctg.0000000000000575] [Medline: 36854062]
10. Kwon OY, Lee MK, Lee HW, Kim H, Lee JS, Jang Y. Mobile app-based lifestyle coaching intervention for patients with nonalcoholic fatty liver disease: randomized controlled trial. J Med Internet Res 2024 Feb 15;26:e49839. [doi: 10.2196/49839] [Medline: 38358794]
11. Sato M, Akamatsu M, Shima T, et al. Impact of a novel digital therapeutics system on nonalcoholic steatohepatitis: the NASH App clinical trial. Am J Gastroenterol 2023 Aug 1;118(8):1365-1372. [doi: 10.14309/ajg.0000000000002143] [Medline: 36656974]
12. Björnsdottir S, Ulfsdottir H, Gudmundsson EF, et al. User engagement, acceptability, and clinical markers in a digital health program for nonalcoholic fatty liver disease: prospective, single-arm feasibility study. JMIR Cardio 2024 Feb 15;8:e52576. [doi: 10.2196/52576] [Medline: 38152892]
13. Albhaisi S, Tondt J, Cyrus J, Chinchilli VM, Conroy DE, Stine JG. Digital therapeutics lead to clinically significant body weight loss in patients with metabolic dysfunction-associated steatotic liver disease: a systematic review and meta-analysis. Hepatol Commun 2024 Aug 1;8(8):e0499. [doi: 10.1097/HC9.0000000000000499] [Medline: 39082956]
14. Chen D, Shao J, Zhang H, et al. Development of an individualized WeChat mini program-based intervention to increase adherence to dietary recommendations applying the behaviour change wheel among individuals with metabolic syndrome. Ann Med 2023;55(2):2267587. [doi: 10.1080/07853890.2023.2267587] [Medline: 37898907]
15. Duan Y, Li X, Guo L, Liang W, Shang B, Lippke S. A WeChat mini program-based intervention for physical activity, fruit and vegetable consumption among Chinese cardiovascular patients in home-based rehabilitation: a study protocol. Front Public Health 2022;10:739100. [doi: 10.3389/fpubh.2022.739100] [Medline: 35392478]
16. Duan Y, Liang W, Guo L, et al. Effectiveness of a wechat mini program-based intervention on promoting multiple health behavior changes among Chinese patients with cardiovascular diseases in home-based rehabilitation: randomized controlled trial. J Med Internet Res 2025 Jun 3;27:e66249. [doi: 10.2196/66249] [Medline: 40460318]
17. Pan Y, Tang J, Lu B, et al. Effects of cognitive behavioral therapy for diet on postprandial glucose and pregnancy outcomes in gestational diabetes mellitus: multicenter randomized controlled trial. J Med Internet Res 2025 Jul 29;27:e71075. [doi: 10.2196/71075] [Medline: 40729762]
18. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
19. National Workshop on Fatty Liver and Alcoholic Liver Disease, Chinese Society of Hepatology, Chinese Medical Association, Fatty Liver Expert Committee, Chinese Medical Doctor Association. Guidelines of prevention and treatment for nonalcoholic fatty liver disease: a 2018 update. Zhonghua Gan Zang Bing Za Zhi 2018 Mar 20;26(3):195-203. [doi: 10.3760/cma.j.issn.1007-3418.2018.03.008] [Medline: 29804393]

XSL•FO
RenderX

20. Karlas T, Petroff D, Sasso M, et al. Individual patient data meta-analysis of controlled attenuation parameter (CAP) technology for assessing steatosis. J Hepatol 2017 May;66(5):1022-1030. [doi: 10.1016/j.jhep.2016.12.022] [Medline: 28039099]
21. Chinese Society of Health Management, Chinese Nutrition Society, Reproductive Medicine Branch of China International Exchange and Promotion Association for Medicine and Healthcare, China Health Promotion Foundation, Zhejiang Provincial Clinical Nutrition Center. Expert consensus & standard on weight management for overweight or obese people. Chin J Health Manage 2018;12(3):200-207. [doi: 10.3760/cma.j.issn.1674-0815.2018.03.003]
22. Craig CL, Marshall AL, Sjöström M, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003 Aug;35(8):1381-1395. [doi: 10.1249/01.MSS.0000078924.61453.FB] [Medline: 12900694]
23. Sun C, Goh GBB, Chow WC, et al. Prevalence and risk factors for impaired renal function among Asian patients with nonalcoholic fatty liver disease. Hepatobiliary Pancreat Dis Int 2024 Jun;23(3):241-248. [doi: 10.1016/j.hbpd.2023.08.004] [Medline: 37620227]
24. Bedogni G, Bellentani S, Miglioli L, et al. The Fatty Liver Index: a simple and accurate predictor of hepatic steatosis in the general population. BMC Gastroenterol 2006 Nov 2;6:33. [doi: 10.1186/1471-230X-6-33] [Medline: 17081293]
25. Newsome PN, Sasso M, Deeks JJ, et al. FibroScan-AST (FAST) score for the non-invasive identification of patients with non-alcoholic steatohepatitis with significant activity and fibrosis: a prospective derivation and global validation study. Lancet Gastroenterol Hepatol 2020 Apr;5(4):362-373. [doi: 10.1016/S2468-1253(19)30383-8] [Medline: 32027858]
26. Kruger FC, Daniels CR, Kidd M, et al. APRI: a simple bedside marker for advanced fibrosis that can avoid liver biopsy in patients with NAFLD/NASH. S Afr Med J 2011 Jun 27;101(7):477-480. [doi: 10.10520/EJC67626] [Medline: 21920102]
27. Angulo P, Hui JM, Marchesini G, et al. The NAFLD fibrosis score: a noninvasive system that identifies liver fibrosis in patients with NAFLD. Hepatology 2007 Apr;45(4):846-854. [doi: 10.1002/hep.21496] [Medline: 17393509]
28. Sterling RK, Lissen E, Clumeck N, et al. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. Hepatology 2006 Jun;43(6):1317-1325. [doi: 10.1002/hep.21178] [Medline: 16729309]
29. Sanyal AJ, Foucquier J, Younossi ZM, et al. Enhanced diagnosis of advanced fibrosis and cirrhosis in individuals with NAFLD using FibroScan-based Agile scores. J Hepatol 2023 Feb;78(2):247-259. [doi: 10.1016/j.jhep.2022.10.034] [Medline: 36375686]
30. Lim SL, Johal J, Ong KW, et al. Lifestyle intervention enabled by mobile technology on weight loss in patients with nonalcoholic fatty liver disease: randomized controlled trial. JMIR Mhealth Uhealth 2020 Apr 13;8(4):e14802. [doi: 10.2196/14802] [Medline: 32281943]
31. Björnsdottir S, Ulfsdottir H, Gudmundsson EF, et al. Long-term feasibility and outcomes of a digital health program to improve liver fat and cardiometabolic markers in individuals with nonalcoholic fatty liver disease: prospective single-arm feasibility study. JMIR Cardio 2025 Sep 12;9:e72074. [doi: 10.2196/72074] [Medline: 40939135]
32. Harrison SA, Ruane PJ, Freilich B, et al. A randomized, double-blind, placebo-controlled phase IIa trial of efruxifermin for patients with compensated NASH cirrhosis. JHEP Rep 2023 Jan;5(1):100563. [doi: 10.1016/j.jhepr.2022.100563] [Medline: 36644237]
33. Armandi A, Bugianesi E. Dietary and pharmacological treatment in patients with metabolic-dysfunction associated steatotic liver disease. Eur J Intern Med 2024 Apr;122:20-27. [doi: 10.1016/j.ejim.2024.01.005] [Medline: 38262842]
34. Mazzotti A, Caletti MT, Brodosi L, et al. An internet-based approach for lifestyle changes in patients with NAFLD: two-year effects on weight loss and surrogate markers. J Hepatol 2018 Nov;69(5):1155-1163. [doi: 10.1016/j.jhep.2018.07.013] [Medline: 30290973]
35. Tincopa MA, Lyden A, Wong J, Jackson EA, Richardson C, Lok AS. Impact of a pilot structured mobile technology based lifestyle intervention for patients with nonalcoholic fatty liver disease. Dig Dis Sci 2022 Feb;67(2):481-491. [doi: 10.1007/s10620-021-06922-6] [Medline: 33939147]
36. Lin H, Lee HW, Yip TCF, et al. Vibration-controlled transient elastography scores to predict liver-related Events in steatotic liver disease. JAMA 2024 Apr 16;331(15):1287-1297. [doi: 10.1001/jama.2024.1447] [Medline: 38512249]
37. Lee J, Vali Y, Boursier J, et al. Prognostic accuracy of FIB-4, NAFLD fibrosis score and APRI for NAFLD-related events: a systematic review. Liver Int 2021 Feb;41(2):261-270. [doi: 10.1111/liv.14669] [Medline: 32946642]
38. Romero-Gómez M, Zelber-Sagi S, Trenell M. Treatment of NAFLD with diet, physical activity and exercise. J Hepatol 2017 Oct;67(4):829-846. [doi: 10.1016/j.jhep.2017.05.016] [Medline: 28545937]
39. Huber Y, Pfirrmann D, Gebhardt I, et al. Improvement of non-invasive markers of NAFLD from an individualised, web-based exercise program. Aliment Pharmacol Ther 2019 Oct;50(8):930-939. [doi: 10.1111/apt.15427] [Medline: 31342533]
40. Stine JG, Rivas G, Hummer B, et al. Mobile health lifestyle intervention program leads to clinically significant loss of body weight in patients with NASH. Hepatol Commun 2023 Apr 1;7(4):e0052. [doi: 10.1097/HC9.0000000000000052] [Medline: 36930864]
41. Chew NWS, Mehta A, Goh RSJ, et al. Cardiovascular-liver-metabolic health: recommendations in screening, diagnosis, and management of metabolic dysfunction-associated steatotic liver disease in Cardiovascular disease via modified Delphi approach. Circulation 2025 Jan 7;151(1):98-119. [doi: 10.1161/CIRCULATIONAHA.124.070535] [Medline: 39723980]
42. Chan WK, Wong VWS, Adams LA, Nguyen MH. MAFLD in adults: non-invasive tests for diagnosis and monitoring of MAFLD. Hepatol Int 2024 Oct;18(Suppl 2):909-921. [doi: 10.1007/s12072-024-10661-x] [Medline: 38913148]

XSL•FO
RenderX

43. Li M, Cui M, Li G, et al. The Pathophysiological associations between obesity, NAFLD, and atherosclerotic cardiovascular diseases. Horm Metab Res 2024 Oct;56(10):683-696. [doi: 10.1055/a-2266-1503] [Medline: 38471571]
44. He S, Lu S, Yu C, et al. The newly proposed plasma-glycosylated hemoglobin A1c/High-Density lipoprotein cholesterol ratio serves as a simple and practical indicator for screening metabolic associated fatty liver disease: an observational study based on a physical examination population. BMC Gastroenterol 2024 Aug 19;24(1):274. [doi: 10.1186/s12876-024-03362-0] [Medline: 39160462]
45. Axley P, Kodali S, Kuo YF, et al. Text messaging approach improves weight loss in patients with nonalcoholic fatty liver disease: a randomized study. Liver Int 2018 May;38(5):924-931. [doi: 10.1111/liv.13622] [Medline: 29117472]
46. Motz V, Faust A, Dahmus J, Stern B, Soriano C, Stine JG. Utilization of a directly supervised telehealth-based exercise training program in patients with nonalcoholic steatohepatitis: feasibility study. JMIR Form Res 2021 Aug 17;5(8):e30239. [doi: 10.2196/30239] [Medline: 34402795]
47. Colantoni A, Bucci T, Cocomello N, et al. Lipid-based insulin-resistance markers predict cardiovascular events in metabolic dysfunction associated steatotic liver disease. Cardiovasc Diabetol 2024 May 20;23(1):175. [doi: 10.1186/s12933-024-02263-6] [Medline: 38769519]

Abbreviations

ALT: alanine aminotransferase
APRI: aspartate aminotransferase-to-platelet ratio index
ASM: appendicular skeletal muscle mass
AST: aspartate aminotransferase
CAP: controlled attenuation parameter
CONSORT: Consolidated Standards of Reporting Trials
DTx: digital therapeutics
FAST: FibroScan-aspartate aminotransferase
FIB-4: fibrosis-4
FPG: fasting plasma glucose
GGT: γ-glutamyl transpeptidase
HbA_{1c}: hemoglobin A_{1c}
HDL-C: high-density lipoprotein-cholesterol
HOMA-IR: homeostasis model assessment-insulin resistance
IPAQ-SF: International Physical Activity Questionnaire-Short Form
ITT: intention-to-treat
LSM: liver stiffness measurement
MAFLD: metabolic dysfunction–associated fatty liver disease
MASH: metabolic dysfunction-associated steatohepatitis
MET: metabolic equivalent of task
OR: odds ratio
PP: per-protocol
RR: relative risk
TLC: therapeutic lifestyle changes

Edited by S Brini; submitted 18.Apr.2025; peer-reviewed by J Zhang, N Maye; revised version received 26.Nov.2025; accepted 02.Dec.2025; published 27.Jan.2026.

Please cite as:
Sun C, Chen G, Shi C, Cao H, Yang R, Zeng J, Duan X, Sun X, Fan JG
Therapeutic Effects of a WeChat Mini-Program on Metabolic Dysfunction–Associated Fatty Liver Disease: Randomized Controlled Trial
J Med Internet Res 2026;28:e76204
URL: https://www.jmir.org/2026/1/e76204
doi:10.2196/76204

© Chao Sun, Guangyu Chen, Cuicui Shi, Haixia Cao, Ruixu Yang, Jing Zeng, Xiaoyan Duan, Xin Sun, Jian-Gao Fan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of

XSL•FO
RenderX

Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Cognitive-Behavioral Therapy–Based Massed Brief Psychoeducational Group via Videoconference for Social Anxiety: Randomized Controlled Trial

Lele Feng[1], MA; Wei Liu[1], MA; Liechuan Cui[1], MA; Deborah Dobson[2], PhD; Xinfeng Tang[1], PhD

[1]Department of Psychology, Renmin University of China, Huixian Building, 9th, 59 Zhongguancun St., Haidian District, Beijing, China

[2]Department of Psychology, University of Calgary, Calgary, AB, Canada

Corresponding Author:
Xinfeng Tang, PhD
Department of Psychology, Renmin University of China, Huixian Building, 9th, 59 Zhongguancun St., Haidian District, Beijing, China

Abstract

Background: Group cognitive-behavioral therapy (CBT), delivered through weekly videoconference sessions, has been shown to effectively reduce social anxiety. However, no studies have evaluated CBT delivered via videoconference in a 2-day massed brief psychoeducational group format.

Objective: This randomized controlled trial aimed to evaluate the efficacy of a videoconferencing CBT–based massed brief psychoeducational group among Chinese university students with social anxiety.

Methods: University students with social anxiety were recruited online and randomly assigned to an intervention group or a waitlist control group. Participants in the intervention group attended a 2-day workshop via videoconference. Assessments were conducted at baseline (T1), posttest (T2), 1-month follow-up (T3), and 3-month follow-up (T4), using the Social Phobia Inventory, Brief Fear of Negative Evaluation Scale, Depression Anxiety Stress Scales–Short Form, Social Anxiety Knowledge Test, Social Anxiety Stigma Inventory, and Attitudes Toward Seeking Professional Psychological Help Scale–Short Form.

Results: The intervention group showed significant reductions in Social Phobia Inventory scores (β=−4.00, 95% bootstrap CI −6.55 to −1.22; $d_{T2\text{-}4}$=−0.97 to −0.81) and Brief Fear of Negative Evaluation Scale scores (β=−1.37, 95% bootstrap CI −2.64 to −0.08; d_{T3}=−0.56), as well as significant increases in Social Anxiety Knowledge Test scores (β=.62, 95% bootstrap CI 0.05-1.17; $d_{T2\text{-}4}$=0.86-1.53). No significant changes were observed in Depression Anxiety Stress Scales–Short Form, Social Anxiety Stigma Inventory, or Attitudes Toward Seeking Professional Psychological Help Scale–Short Form scores.

Conclusions: The findings indicate that videoconferencing CBT–based massed brief psychoeducational group was effective in reducing social anxiety among university students. Future research with larger and more diverse samples is recommended to validate the efficacy and assess the scalability of this intervention format.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2400093444; www.chictr.org.cn/showprojEN.html?proj=235703

(J Med Internet Res 2026;28:e79825) doi:10.2196/79825

KEYWORDS

cognitive behavioral therapy; massed intervention; psychoeducational group; randomized controlled trial; social anxiety; workshop intervention; low-intensity intervention

Introduction

Social anxiety refers to the intense fear, worry, or avoidance of social interactions and situations in which one may be evaluated by others [1]. When social anxiety reaches a severity that impairs daily functioning, it is diagnosed as social anxiety disorder or social phobia [2]. A meta-analysis reported that the prevalence of social anxiety symptoms among Chinese adolescents and young adults aged 15 - 25 years is 29.8% [3]. Social anxiety disorder can lead to functional impairments across multiple aspects, including work, academics, social functioning, and cognitive processes [4]. Even subclinical symptoms, which do not meet diagnostic criteria, can negatively impact various aspects of life [5] and may progress into a chronic and debilitating condition if left unaddressed [6]. For example, social anxiety may hinder college students' learning, decrease their well-being [7], and even elevate their risk of suicidal thoughts or behaviors [8]. Early intervention is, therefore, critical to alleviating symptom severity and the likelihood of developing social anxiety disorder.

Cognitive-behavioral therapy (CBT) is widely recognized as the gold standard for treating social anxiety [9]. Meta-analytic

XSL•FO
RenderX

findings indicate that compared to pharmacological and alternative therapeutic approaches, CBT demonstrates superior efficacy, greater safety, and lower relapse rates in both the short and long term, making it the most highly recommended treatment [10,11]. CBT is typically delivered in individual or group formats, both of which have shown positive effects on social anxiety. However, group CBT has its unique advantages by allowing participants to engage with unfamiliar peers in socially relevant contexts while benefiting from mutual support. It also enables more individuals to receive treatment within a given timeframe [12]. Therefore, group interventions are also frequently used in both research and clinical practice.

In terms of therapist-guided delivery medium, psychological interventions can be categorized into face-to-face CBT and remote CBT. Traditional group CBT typically involves therapists working face-to-face with participants; however, limited mental health resources often prevent individuals from accessing CBT or other mental health services in their daily lives [13]. Moreover, traditional group CBT may entail several temporal and locational constraints, limiting the delivery of timely clinical interventions [14]. In this context, internet-based remote CBT offers a viable alternative [15]. Among remote modalities, *videoconferencing* is commonly used, as it enables synchronous, real-time communication through audio and video and allows therapists and participants to interact directly from different locations. Compared to other digital formats, such as message-based or app-based interventions, videoconferencing CBT maintains therapist visibility, facilitating greater interaction and engagement between participants and therapists [16]. Studies have shown that videoconferencing CBT yields significant therapeutic effects for social anxiety and social anxiety disorder, with symptom improvements maintained for up to 3 months postintervention [17,18] and effect sizes ranging from medium to large [19]. Moreover, compared to traditional CBT, videoconferencing CBT may alleviate the resistance to therapy triggered by face-to-face interaction in individuals with social anxiety [20].

Low-intensity interventions have become a common approach in the delivery of psychological interventions. The National Institute for Health and Care Excellence recommends that group CBT for social anxiety be delivered in weekly sessions, for a total of 8 - 12 sessions [21]. This implementation of full protocol requires substantial time and financial resources; as such, low-intensity CBT has emerged to meet the growing demand for mental health services while ensuring treatment effectiveness [22]. Within health care systems, low-intensity interventions are commonly employed in primary care settings for adults experiencing symptoms of depression or specific phobias [23]. A meta-analysis revealed that low-intensity CBT produced large effect sizes in the treatment of anxiety disorders (d=1.06) [24].

A common form of low-intensity intervention involves employing only select core therapeutic techniques rather than the full treatment protocol; this approach is often referred to as a *brief* intervention. For example, in social anxiety interventions, Clark's full protocol typically includes components, such as attention training, video feedback, behavioral experiments, and discrimination training [25]. However, Heimberg's intervention full protocol incorporates multiple sessions of cognitive restructuring and graduated exposure. Brief interventions for social anxiety based on these 2 protocols have also demonstrated promising outcomes [26]. A group CBT for children with social anxiety, delivered in three weekly 3-hour sessions, included psychoeducation, cognitive strategies, and behavioral exposure. Significant reductions in participants' social anxiety scores were observed at both posttest and 3-week follow-up [27]. A recent 7-day internet−based CBT program for social anxiety disorder also demonstrated substantial outcomes (Hedges gs=1.26 - 1.9). This program consisted of 6 online lessons accompanied by practice tasks. Participants were required to complete lessons and corresponding exercises, which included exposure tasks, cognitive challenges, and communication skills practice [28]. More recently, a 1-day CBT−based workshop consisting of cognitive restructuring and assertiveness training for secondary vocational students demonstrated moderate effect sizes in reducing social anxiety symptoms [29].

Another common type of low-intensity intervention is the psychoeducational group. According to Gladding's classification, group interventions are generally divided into *psychoeducational* groups and *counseling* groups [30]. Compared to counseling groups, psychoeducational groups are lower in intensity and emphasize using educational methods to acquire information and develop related meaning and skills [31]. These groups integrate both knowledge acquisition and skill development, efficiently delivering information about psychological disorders while also providing opportunities for practice and experiential learning. Accordingly, 1 major aim of psychoeducational groups is to enhance participants' mental health literacy, which is referred to as the ability to recognize a disorder, understand it, reduce stigma, and seek appropriate psychological help [32]. In addition, the skills training component can help alleviate psychological distress. Evidence indicates that psychoeducational interventions are effective in alleviating anxiety symptoms, including social anxiety [33,34]. A further advantage of psychoeducational groups is their ability to accommodate a larger number of participants. Counseling groups for social anxiety typically have limited membership, with recommended group sizes ranging from 6 to 12 participants [26,35]. In contrast, psychoeducational groups are often larger in scale, commonly ranging from 15 to 40 participants [30].

In terms of intervention frequency, the conventional model involves *spaced* interventions (eg, 1 session per week). Alternatively, *massed* intervention compresses the treatment timeline by delivering multiple sessions over a relatively short timeframe. Meta-analytic evidence indicates that for anxiety disorders in young adults, massed CBT can achieve intervention outcomes comparable to those of full-protocol CBT [36]. For example, Deacon and Abramowitz [37] conducted a 2-day CBT for panic disorder, condensing a 12-session protocol into 2 sessions of 6 and 3 hours, respectively. The intervention included psychoeducation, therapist-assisted exposure, fear hierarchy construction, and in vivo exposure. Following the treatment, all 10 participants showed significant reductions on the Panic Disorder Severity Scale, with large effect sizes (d_s=0.96 - 3.29). To date, spaced brief CBT has been shown to effectively reduce social anxiety and related negative beliefs

XSL•FO
RenderX

[38,39]; however, no empirical evidence currently exists on the effects of massed brief CBT for social anxiety.

The vast majority of existing remote interventions for social anxiety employ spaced, full-protocol treatment programs, typically consisting of 8 or 12 weekly sessions. These interventions are generally conducted in groups of 5 - 12 participants [40-42]. While this model offers several advantages, it also presents challenges. For example, participants may find it difficult to sustain long-term engagement remotely, which can result in them dropping out. A meta-analysis reported that the dropout rate for group CBT interventions targeting anxiety disorders is 24.6% [43]. To alleviate social anxiety more rapidly, efficiently, and conveniently, we developed a low-intensity intervention format—the massed brief psychoeducational group (MBPG)—delivered via videoconference. This study aimed to evaluate the effects of this delivery model on university students with social anxiety through a randomized controlled trial. We hypothesized that compared to the waitlist control group, participants in the intervention group would show significant reductions in social anxiety, fear of negative evaluation, and depressive symptoms following treatment. Additionally, we expected improvements in social anxiety literacy, including increased knowledge of social anxiety, reduced social anxiety stigma, and more positive attitudes toward seeking professional psychological help.

Methods

Participants

According to a meta-analysis by Mayo-Wilson et al [10], interventions with shortened sessions based on the Clark and Wells model for social anxiety have demonstrated large effect sizes. Based on these findings, the expected effect size for this study was conservatively set at 0.80. A priori power analysis using the G*Power 3.1 by Faul et al [44] indicated that a total sample size of 52 would be required. Accounting for an anticipated dropout rate of 20%, the target sample size was set at 65. Accordingly, a minimum of 65 participants were targeted.

Participants were recruited online and were required to meet the following eligibility criteria: (1) current enrollment in a university program, with no restrictions on degree level; (2) availability to attend a 2-day remote workshop and willingness to be randomly assigned; (3) no suicidal ideation and no formal diagnosis of any psychiatric disorder; (4) no psychotropic medication use within the past year and no history of psychological treatment or counseling; and (5) no specific cutoff on social anxiety scores was required for inclusion. As a psychoeducational group, our primary goal was to reach individuals who experienced social anxiety−related difficulties and were motivated to make changes, regardless of their symptom severity. Therefore, although a cutoff value of the Social Phobia Inventory (SPIN) is commonly used to indicate clinically significant symptoms, we did not adopt this cutoff during recruitment. Individuals were eligible for this workshop as long as they perceived themselves affected by social anxiety. Despite the absence of a formal cutoff, participants' baseline levels of social anxiety were relatively high. Among those enrolled, 62 (90%) of the 69 participants scored above the clinical threshold of 19 on the SPIN, with a mean score of 37.91 (SD=13.67).

Participants were randomly assigned to either the intervention group (n=39) or the waitlist control group (n=39). Prior to the start of the intervention, 7 participants from the intervention group and 2 from the control group withdrew. A total of 69 participants ultimately took part in the study (n=32 in the intervention group; n=37 in the control group).

Ethical Considerations

This study received approval from the Ethical Review Committee of the Department of Psychology at Renmin University of China (IRB-24 - 041). All participants provided written informed consent before participating. Individuals who completed all 4 assessments received both monetary compensation of 40 RMB (US $ 5.70) in total (10 RMB [US $1.42] per assessment) and a commemorative gift. To ensure the protection of participants' privacy, the raw data were accessible only to the research team and numerical IDs were assigned in lieu of personally identifiable information during data handling. All reported results were fully anonymized, and no identifiable personal features are presented in the manuscript.

Procedure

Simple randomization was employed to assign participants to groups. Random sequences were generated using Microsoft Excel, with half of the participants allocated to the intervention group and the other half to the waitlist control group.

Before the intervention, an online questionnaire was administered to all participants to collect baseline data (T1). The posttest (T2) was conducted 1 week after the intervention, with follow-up assessments at 1 month (T3) and 3 months (T4) postintervention. Participants in the waitlist control group received the same 2-day online workshop intervention as those in the intervention group after all data collection was completed.

Intervention

This study implemented a CBT-based MBPG intervention, delivered via videoconferencing. The intervention was delivered synchronously via Tencent Meeting, an online platform similar to Zoom. All needed materials were presented in the form of slides shared on screen during the online sessions. The intervention program was named *Joymaster Workshop* to engage participants' interest, reduce associated stigma, and symbolize mastery of social enjoyment. For CBT-based counseling groups targeting social anxiety, a recommended group size is approximately 6 participants [45], whereas psychoeducational groups are typically larger [30]. To avoid reductions in interaction, experiential engagement, and opportunities for practice associated with overly large groups, we set the size of the psychoeducational group at 15 - 20 participants. Therefore, we divided the 32 participants in the intervention condition into 2 workshops. In allocating participants, we considered their time availability and aimed to maintain balance across the 2 groups in terms of size and gender distribution to minimize potential biases arising from group composition. The 2 workshops ultimately included 17 and 15 participants, respectively. Both workshops were delivered by the same group

XSL•FO
RenderX

leaders on 2 consecutive weekends, with identical content to ensure treatment fidelity. The timing of the post-intervention and follow-up assessments was adjusted according to the specific workshop each participant attended.

The term *massed* refers to the delivery format, which consisted of 2 consecutive days, a total duration of 12 hours. As a *brief* intervention, the program selectively incorporated effective techniques from full-protocol CBT, focusing on cognitive restructuring and behavioral experiments. Contemporary CBT interventions for social anxiety typically follow either the Heimberg protocol or the Clark protocol [21]. Given the massed format and the aim of achieving effective outcomes within a short timeframe, this study's program prioritized cognitive change over habituation through exposure, as the latter generally requires more extended treatment. Therefore, we adapted the cognitive restructuring technique from Hope and Heimberg's model, which involves challenging automatic thoughts using supporting and opposing evidence [45].

Behavioral experiments were based on the approach by Leigh and Clark [46] as well as Hofmann and Otto [47], in which participants design and engage in social tasks to test and challenge their automatic thoughts. Behavioral experiments also enhanced the interactivity of the massed intervention and promoted participant engagement. A summary of the main intervention modules is presented in Table 1. Among these, the out-of-session behavioral experiments required participants to complete a selected task during a 2-hour lunch break, based on plans developed at the end of the morning session. Examples included asking a passerby to take a photo of them with a trash can or initiating a brief conversation with someone new in the campus canteen. All participants completed a behavioral experiment.

This intervention was conducted as a *psychoeducational* group rather than a counseling group. First, we employed online slide presentation as visual aids, combining didactic instruction for efficient knowledge delivery with opportunities for experiential practice of therapeutic techniques. Second, the intervention accommodated a larger number of participants than is common in counseling groups, with each workshop hosting 15 - 20 individuals.

Table . Intervention modules.

Module	Techniques	Contents
Day 1—Morning	Psychoeducation	1. Introduce the concept of social anxiety: definition, prevalence, diagnostic criteria, typical age of onset, and negative impacts 2. Provide an overview of cognitive models and contributing factors 3. Introduce treatment approaches for social anxiety
Day 1—Afternoon	Cognitive restructuring (Part 1)	1. Practice the first 2 steps of cognitive restructuring: • Identify automatic thoughts • Identify cognitive biases
Day 2—Morning	Behavioral experiments	1. Explain the purpose and steps of behavioral experiments 2. Conduct in-session behavioral experiments 3. Design out-of-session behavioral experiments for lunch break
Day 2—Afternoon	Cognitive restructuring (Part 2)	1. Reflect on behavioral experiments conducted during lunch break 2. Practice the final 2 steps of cognitive restructuring • Challenge automatic thoughts • Develop realistic, rational thoughts 3. Develop an action plan and conclude the session

Treatment Fidelity

The intervention was delivered by a leader and an assistant. The leader was a PhD-level counselor intensively trained in CBT, and the assistant was a master's student in psychology. Psychoeducational content was presented primarily through PowerPoint slides, with all procedures structured around the slide content. This ensured a highly standardized delivery and strong adherence to the intervention protocol.

Measurements

Primary Outcome—Social Anxiety

The Chinese version of the 17-item SPIN, developed by Connor et al [48] and revised by Xiao et al [49], was used to assess the severity of social anxiety. The scale comprises 3 subscales: fear (6 items; eg, fear of embarrassment or rejection), avoidance (7 items; eg, avoiding social situations because of fear), and physiological symptoms (4 items; eg, blushing or sweating in

XSL•FO
RenderX

social settings)—covering the core clinical manifestations of social anxiety. It uses a 5-point Likert rating scale from 0 (*not at all*) to 4 (*extremely*), with higher scores indicating greater levels of social anxiety. In this study, the SPIN had a Cronbach α of 0.94 at baseline.

Secondary Outcomes

Fear of Negative Evaluation

The Chinese version of the 12-item Brief Fear of Negative Evaluation Scale (BFNES), developed by Leary [50] and revised by Chen [51], was used to assess individuals' concerns about being negatively evaluated by others. It uses a 5-point Likert scale from 1 (*not at all characteristic of me*) to 5 (*extremely characteristic of me*), with higher scores indicating greater fear of negative evaluation. In this study, the BFNES had a Cronbach α of 0.95 at baseline.

Depression, Anxiety, and Stress

The Chinese version of the 21-item Depression Anxiety Stress Scales–Short Form (DASS-21), developed by Lovibond and Lovibond [52] and revised by Gong et al [53], was used to assess the severity of depressive symptoms, anxiety or autonomic arousal, and stress. It uses a 4-point Likert scale from 0 (*did not apply to me at all*) to 3 (*applied to me very much or most of the time*), with higher scores indicating greater symptom severity. In this study, the DASS-21 had a Cronbach α of 0.93 at baseline.

Social Anxiety Knowledge

The self-developed Social Anxiety Knowledge Test (SAKT) was used to assess participants' knowledge of the definition, symptoms, treatment, and related aspects of social anxiety. Based on a literature review, the content was categorized into 8 domains: basic knowledge, symptoms, prevalence, age of onset, gender differences, cultural differences, risk factors, and treatment. An initial pool of 23 single-choice items (with 4 answer options each) was rated and reviewed by 7 experts for content validity, and semistructured cognitive interviews were conducted with 5 nonspecialists. After revisions based on the expert ratings and cognitive interview results, 21 items were retained. Responses were scored as correct (1 point) or incorrect (0 points), with higher total scores (maximum=21) indicating better social anxiety literacy. The SAKT has been adopted in several interventions and shown good validity to test social anxiety−related knowledge [29,54].

Stigmatizing Attitudes Toward Social Anxiety

The self-developed Social Anxiety Stigma Inventory (SASI) was used to assess individuals' stigmatizing attitudes toward social anxiety. The 10-item questionnaire uses a 5-point Likert scale from 0 (*strongly disagree*) to 4 (*strongly agree*), with higher scores indicating stronger stigma. The scale demonstrated good structural validity (χ^{232}= 80.4, root mean square error of approximation=0.061, comparative fit index=0.933, Tucker-Lewis index=0.906, root mean square residual=0.048) and an internal consistency reliability of 0.75. In this study, the SASI had a Cronbach α of 0.86 at baseline. The SASI has been adopted in several interventions and shown good reliability and validity to assess the stigma toward social anxiety [29,54].

Attitudes Toward Seeking Professional Psychological Help

The Chinese version of the 10-item Attitudes Toward Seeking Professional Psychological Help Scale–Short Form (ATSPPH-SF), developed by Fischer and Farina [55] and revised by Fang et al [56], was used to assess individuals' attitudes toward psychological help-seeking. This instrument employs a 4-point Likert scale from 0 (*strongly disagree*) to 3 (*strongly agree*), with higher scores indicating a stronger willingness to seek help. In this study, the ATSPPH-SF had an acceptable Cronbach α of 0.73 at baseline.

Data Analysis

The data were analyzed following the intent-to-treat principle, including all randomized participants. The Shapiro-Wilk test was employed to assess residual normality for each scale across groups and time points. Intervention effects over time were examined using linear mixed models, which are less sensitive to missing data. Given that residuals for certain variables at specific time points were skewed and the sample size was limited, a bootstrap method with 1000 resamples was applied to estimate the linear mixed model fixed-effect parameters and their 95% CIs, thereby enhancing the robustness of statistical inferences and avoiding strict distributional assumptions [57].

Cohen *d* was calculated to determine both between-group and within-group effect sizes. Owing to baseline differences between the intervention and waitlist control groups, Morris' baseline-adjusted formula was applied to adjust the between-group effect sizes at posttest and follow-up [58]. The effect size was based on mean pre-post change ($M_{post,\ T} - M_{pre,\ T}$) in the intervention group minus the mean pre-post change ($M_{post,\ C} - M_{pre,\ C}$), divided by the pooled pretest standard deviation (SD_{pre}). All analyses were conducted using RStudio (version 4.5.0).

$$d=(Mpost,\ T-Mpre,\ T)-(Mpost,\ C-Mpre,\ C)SDpre$$

where the pooled standard deviation is defined as

$$SDpre=(nT-1)SDpre,T2+(nC-1)SDpre,C2nT+nC-2$$

Results

Participants' Demographic Characteristics and Baseline Scores

The detailed flow of participants' recruitment, allocation, and analysis is presented as Figure 1. No significant differences were found between the intervention and control groups in demographic variables or baseline outcome scores (see Table 2).

XSL•FO
RenderX

Figure 1. Participant flow.

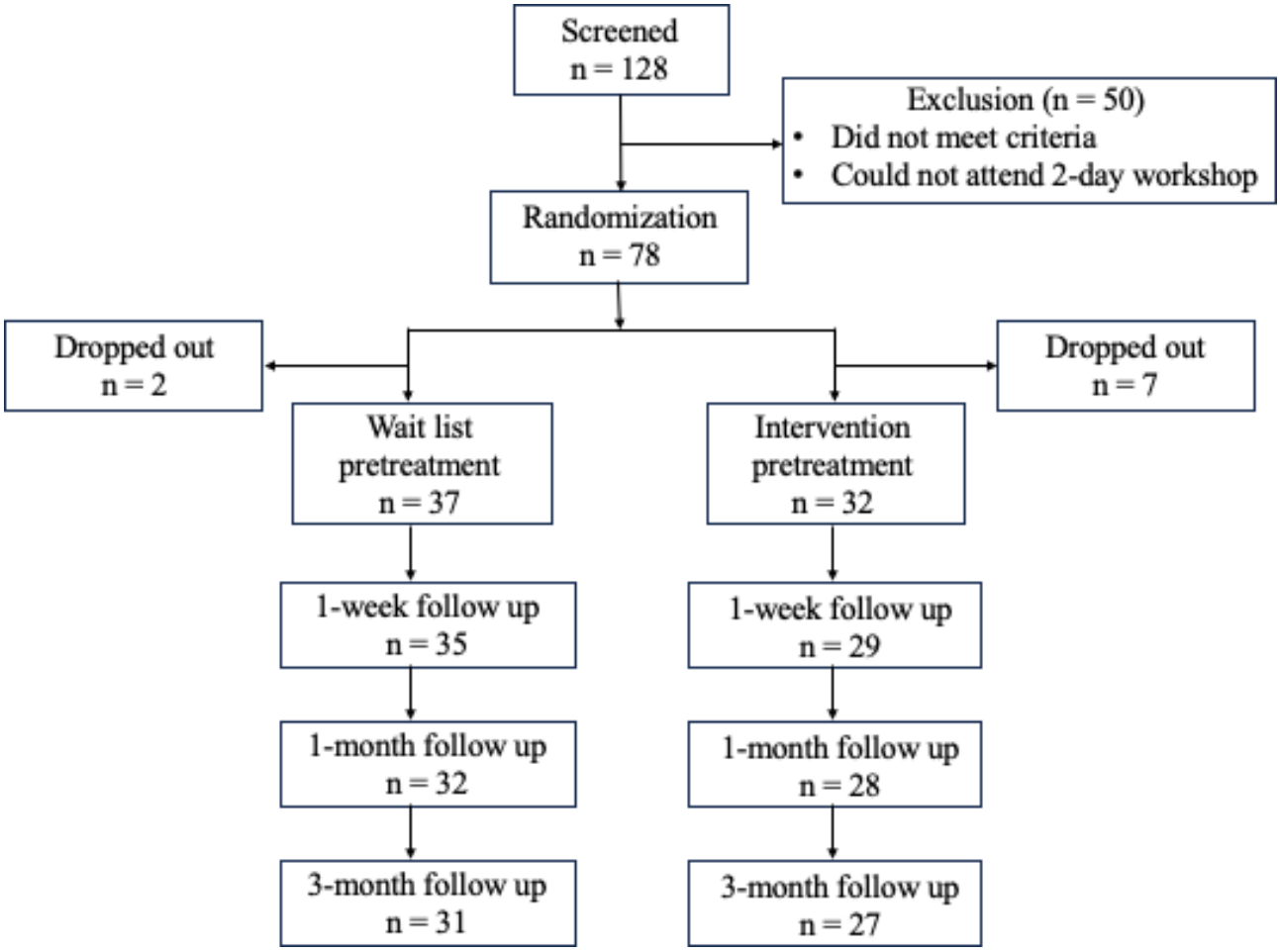

XSL•FO
RenderX

Table . Demographic characteristics and the baseline of participants.

Variable	All	Group		*t* test (*df*)	Chi-square (*df*)	*P* value
		Intervention (n=32)	Control (n=37)			
Age, mean (SD) (y)	23.16 (3.89)	22.47 (2.99)	23.76 (4.48)	−1.39 (67)	__[a]	.17
Gender, n (%)				—	0.39 (1)	.53
Man	17 (24.6)	9 (28.1)	8 (21.6)			
Woman	52 (75.4)	23 (71.9)	29 (78.4)			
Residence, n (%)				—	2.41 (2)	.30
Rural area	24 (34.8)	14 (43.8)	10 (27)			
Township	12 (17.4)	4 (12.5)	8 (21.6)			
City	33 (47.8)	14 (43.8)	19 (51.4)			
Education, n (%)				—	1.41 (2)	.495
Bachelor	40 (58)	18 (56.3)	22 (59.5)			
Master	25 (36.2)	11 (34.4)	14 (37.8)			
Doctor	4 (5.8)	3 (9.4)	1 (2.7)			
Baseline scores, mean (SD)						
SPIN[b]	—	38.97 (15.78)	37.00 (11.69)	−0.58 (67)	—	.56
BFNES[c]	—	46.19 (7.41)	45.51 (7.76)	−0.37 (67)	—	.71
DASS-21[d]	—	20.84 (10.16)	18.54 (10.53)	−0.92 (67)	—	.36
SAKT[e]	—	10.34 (2.67)	11.22 (2.18)	1.47 (67)	—	.14
SASI[f]	—	17.41 (6.71)	17.49 (6.06)	0.05 (67)	—	.96
ATSPPH-SF[g]	—	19.19 (4.64)	19.72 (4.15)	0.51 (67)	—	.61

[a]Not applicable.

[b]SPIN: Social Phobia Inventory.

[c]BFNES: Brief Fear of Negative Evaluation Scale.

[d]DASS-21: Depression Anxiety Stress Scales–Short Form.

[e]SAKT: Social Anxiety Knowledge Test.

[f]SASI: Social Anxiety Stigma Inventory.

[g]ATSPPH-SF: Attitudes Toward Seeking Professional Psychological Help Scale–Short Form.

Shapiro-Wilk Normality Test

Shapiro-Wilk normality tests were conducted separately for the intervention and waitlist control groups at each time point for all outcome variables. The results showed that for both groups, the residuals of the 5 outcome variables across 4 time points were generally normally distributed (W=0.93 - 0.98, P's=.052-.902, n=35). Exceptions were observed in a few cases: the BFNES in the waitlist group at T1 (W=0.93, P<.05) and T4 (W=0.89, P=.004) and the DASS-21 in the intervention group at T2 (W=0.91, P<.05) and T4 (W=0.90, P<.05) and in the waitlist control group at T4 (W=0.89, P=.003). Overall, these findings suggest that the residual distributions did not exhibit significant skewness.

Primary Outcomes

According to the results (Table 3), the main effects of group and time on SPIN scores were not significant. However, the group×time interaction was significant (β=−4.00, 95% bootstrap CI −6.55 to −1.22), indicating that compared to the waitlist control group, the intervention group experienced a significant reduction in social anxiety over time (see Figure 2 for the trend). Moreover, the intervention group exhibited significant reductions with large effect sizes at T2 (d=−0.81, 95% CI −1.28 to −0.33), T3 (d=−0.97, 95% CI −1.44 to −0.49), and T4 (d=−0.81, 95% CI −1.28 to −0.33; Table 4).

XSL•FO
RenderX

Table . Results of linear mixed model analysis.

Variable	Group		Time		Group×time	
	β	95% boot CI	β	95% boot CI	β	95% boot CI
SPIN[a]	2.76	−4.48 to 9.19	3.26	−1.71 to 6.80	−4.00	−6.55 to −1.22
BFNES[b]	1.22	−2.50 to 4.60	0.65	−1.31 to 2.64	−1.37	−2.64 to −0.08
DASS-21[c]	2.89	−2.18 to 7.37	1.20	−1.81 to 4.21	−1.77	−3.42 to 0.17
SAKT[d]	−0.42	−1.93 to 1.10	−0.43	−1.14 to 0.38	0.62	0.05 to 1.17
SASI[e]	0.46	−2.09 to 3.43	−0.85	−2.41 to 0.62	−0.43	−1.43 to 0.62
ATSPPH-SF[f]	−0.75	−2.87 to 0.94	0.02	−1.04, 1.03	0.32	−0.39 to 1.09

[a]SPIN: Social Phobia Inventory.

[b]BFNES: Brief Fear of Negative Evaluation Scale.

[c]DASS-21: Depression Anxiety Stress Scales–Short Form.

[d]SAKT: Social Anxiety Knowledge Test.

[e]SASI: Social Anxiety Stigma Inventory.

[f]ATSPPH-SF: Attitudes Toward Seeking Professional Psychological Help Scale–Short Form.

XSL•FO
RenderX

Figure 2. Change of outcomes. ATSPPH-SF: Attitudes Toward Seeking Professional Psychological Help Scale–Short Form; BFNES: Brief Fear of Negative Evaluation Scale; DASS-21: Depression Anxiety Stress Scales–Short Form; SAKT: Social Anxiety Knowledge Test; SASI: Social Anxiety Stigma Inventory; SPIN: Social Phobia Inventory.

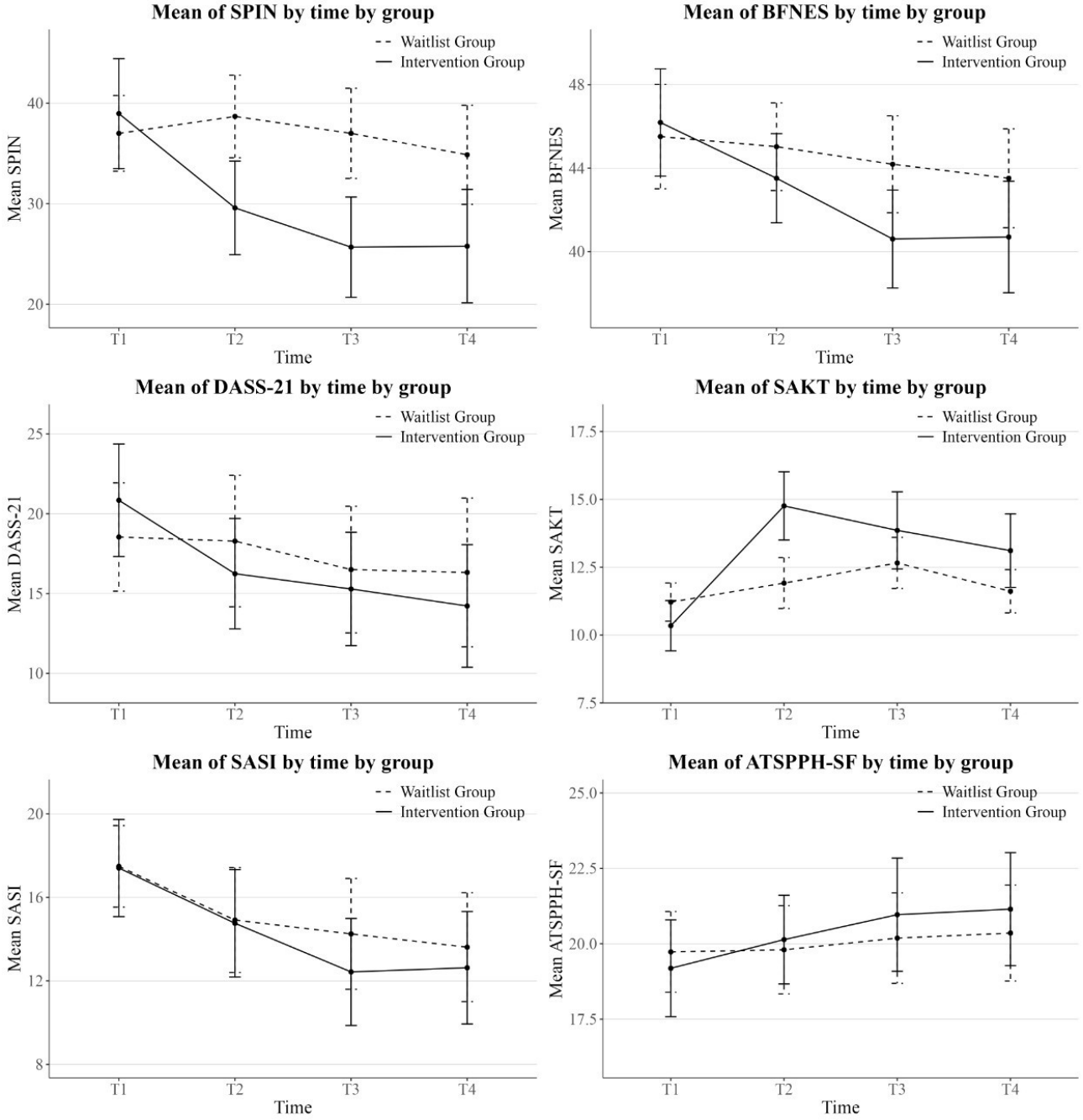

XSL•FO
RenderX

Table . Means, Standard deviations, and Cohen *d* for outcome variables by group and time[a].

Variable	Intervention group		Control group		$d_{b\text{-}g}$[c] (95% CI)
	Mean (SD)	$d_{w\text{-}g}$[b] (95% CI)	Mean (SD)	$d_{w\text{-}g}$	
SPIN[d]					
T1	38.97 (15.78)	—[e]	37.00 (11.69)	—	0.14 (–0.33 to 0.62)
T2	29.31 (13.16)	0.71 (–1.07 to –0.31)	37.59 (12.98)	0.10 (–0.27 to 0.41)	–0.81 (–1.28 to –0.33)
T3	26.03 (13.44)	–0.91 (–1.27 to –0.51)	36.46 (12.71)	–0.08 (–0.45 to 0.29)	–0.97 (–1.44 to –0.49)
T4	25.78 (14.96)	–0.89 (–1.24 to –0.53)	34.87 (14.00)	–0.11 (–0.48 to 0.25)	–0.81 (–1.28 to –0.33)
BFNES[f]					
T1	46.19 (7.41)	—	45.51 (7.76)	—	0.009 (–0.39 to 0.56)
T2	42.81 (6.19)	–0.52 (–0.81 to –0.17)	44.57 (6.48)	–0.01 (–0.44 to 0.32)	–0.29 (–0.76 to 0.19)
T3	40.16 (6.61)	–0.74 (–1.09 to –0.40)	43.00 (7.00)	–0.12 (–0.53 to 0.28)	–0.56 (–1.03 to –0.09)
T4	40.7 (7.08)	–0.76 (–1.04 to –0.42)	43.52 (6.72)	–0.20 (–0.52 to 0.24)	–0.46 (–0.93 to 0.01)
DASS-21[g]					
T1	20.84 (10.16)	—	18.54 (10.53)	—	0.22 (–0.25 to 0.70)
T2	16.66 (9.24)	–0.63 (–1.01 to –0.26)	17.97 (12.17)	0.02 (–0.35 to 0.32)	–0.42 (–0.89 to 0.05)
T3	15.41 (9.85)	–0.71 (–1.17 to –0.31)	16.14 (10.82)	–0.22 (–0.55 to 0.15)	–0.34 (–0.81 to 0.13)
T4	14.22 (10.18)	–0.78 (–1.19 to –0.36)	16.32 (13.23)	–0.07 (–0.41 to 0.31)	–0.43 (–0.90 to 0.05)
SAKT[h]					
T1	10.34 (2.67)	—	11.22 (2.19)	—	–0.36 (–0.84 to 0.12)
T2	14.19 (3.82)	1.21 (0.67 to 1.70)	11.65 (2.98)	0.23 (–0.12 to 0.50)	1.53 (1.06 to 2.01)
T3	13.31 (3.92)	0.98 (0.48 to 1.48)	12.46 (2.86)	0.40 (0.01 to 0.76)	0.86 (0.38 to 1.33)
T4	13.11 (3.59)	0.77 (0.37 to 1.18)	11.61 (2.26)	0.11 (–0.30 to 0.47)	0.98 (0.51 to 1.45)
SASI[i]					
T1	17.41 (6.71)	—	17.49 (6.06)	—	0.01 (–0.49 to 0.46)
T2	14.84 (6.73)	–0.44 (–0.85 to –0.05)	14.51 (7.76)	–0.55 (–0.86 to –0.16)	–0.01 (–0.49 to 0.46)
T3	12.88 (7.09)	–0.78 (–1.19 to –0.38)	14.27 (8.15)	–0.50 (–0.85 to –0.06)	0.27 (–0.75 to 0.20)
T4	12.63 (7.14)	–0.68 (–0.96 to –0.23)	13.61 (7.41)	–0.74 (–1.15 to –0.19)	0.14 (–0.62 to 0.33)
ATSPPH-SF[j]					
T1	19.19 (4.64)	—	19.73 (4.15)	—	0.12 (–0.60 to 0.35)
T2	19.84 (3.96)	0.19 (–0.19 to 0.56)	20.03 (4.41)	0.15 (–0.25 0.42)	0.20 (–0.27 to 0.67)
T3	20.69 (4.86)	0.34 (–0.10 to 0.68)	20.32 (4.36)	0.14 (–0.20, 0.51)	0.30 (–0.17 to 0.77)
T4	21.15 (4.97)	0.39 (0.01 to 0.75)	20.35 (4.52)	0.31 (–0.09, 0.70)	0.31 (–0.17 to 0.78)

[a]Negative values reflect reductions from the baseline. Bootstrapped 95% CIs are reported in parentheses.

[b]Within-group Cohen *d*.

[c]Between-group Cohen *d*.

[d]SPIN: Social Phobia Inventory.

[e]Not available.

[f]BFNES: Brief Fear of Negative Evaluation Scale.

[g]DASS-21: Depression Anxiety Stress Scales–Short Form.

[h]SAKT: Social Anxiety Knowledge Test.

[i]SASI: Social Anxiety Stigma Inventory.

[j]ATSPPH-SF: Attitudes Toward Seeking Professional Psychological Help Scale–Short Form.

XSL•FO
RenderX

Secondary Outcomes

The main effects of group and time on BFNES scores were not significant. However, the group×time interaction effect was significant (β=−1.37, 95% bootstrap CI −2.64 to −0.08), indicating that compared to the waitlist control group, the intervention group experienced a significant reduction in fear of negative evaluation over time. However, this reduction was statistically significant only at T3, with a moderate effect size (*d*=−0.56, 95% CI −1.03 to −0.09).

For DASS-21 scores, neither the main effects of group and time nor the group×time interaction effect was significant (β=-1.77, 95% bootstrap CI −3.42 to 0.17). This indicates that no significant differences existed between the groups in terms of depression, anxiety, or stress symptom trajectories over time.

The main effects of group and time on SAKT scores were not significant. However, the group×time interaction effect was significant (β=.62, 95% bootstrap CI 0.05 to 1.17), indicating that compared to the waitlist control group, the intervention group showed a significant increase in social anxiety knowledge over time. Large effect sizes were observed at T2 (*d*=1.53, 95% CI 1.06-2.00), T3 (*d*=0.86, 95% CI 0.38-1.33), and T4 (*d*=0.98, 95% CI 0.51-1.45).

For SASI scores, the main effects of group and time, as well as the group×time interaction (β=−.43, 95% bootstrap CI −1.43 to 0.62), were not significant. This suggests that there were no significant differences between the groups in terms of social anxiety stigma over time. Similarly, no significant effects were found for ATSPPH-SF scores. The pattern of results mirrored that of the SASI, indicating no significant between-group differences in attitudes toward seeking professional psychological help over time.

Discussion

General Findings

This study examined the effectiveness of a CBT-based MBPG delivered via videoconference for addressing social anxiety in university students. The results showed that the intervention led to significant improvements in social anxiety, including reductions in participants' fear of negative evaluation over a certain period and a marked increase in social anxiety knowledge. However, no significant improvements were observed in emotional distress (ie, depression, anxiety, and stress as measured by the DASS-21), social anxiety stigma, or attitudes toward seeking professional psychological help.

First, compared to the waitlist control group, participants in the intervention group reported significant reductions in social anxiety, with large effect sizes observed in both the short term (1-week posttest and 1-month follow-up) and longer term (3-month follow-up). Moreover, fear of negative evaluation significantly decreased 1 month after the intervention, consistent with the study's hypotheses. A recent meta-analysis reported that various forms of remote CBT yield medium-to-large effect sizes in treating social anxiety [19]. While this study did not directly compare videoconference delivery to other remote CBT modalities, the observed effect sizes align with those reported in prior research on remote CBT for social anxiety, with relatively durable effects. This further suggests that core CBT techniques, particularly cognitive restructuring and behavioral experiments targeting automatic thoughts, can be effectively delivered in both in-person and remote formats, likely without significant differences in efficacy.

Second, this study employed a massed brief intervention, delivering 12 hours of intervention intensively over 2 consecutive days. High-frequency delivery of this kind has been shown to sustain participants' motivation and engagement [36]. The selected intervention components in this study are well-established techniques for reducing social anxiety, and the massed format enables tighter integration across modules, allowing therapeutic gains to accumulate within a short period [59]. Therefore, despite involving fewer sessions than full-protocol interventions, this brief intervention yielded robust therapeutic outcomes. Given that fear of negative evaluation is a core cognitive mechanism in the development and maintenance of social anxiety [60], its reduction further supports the intervention's efficacy. However, in this study, improvement in fear of negative evaluation was significant only at the 1-month follow-up. This suggests that the observed reduction in social anxiety may involve additional mediating mechanisms beyond cognitive change alone, which warrants further investigation in future research.

However, the intervention did not have significant effects on anxiety, depression, or stress, which was inconsistent with the study's hypotheses. Cognitive restructuring, a core component of this intervention, is a widely applicable CBT technique and is theoretically effective for treating anxious or depressive symptoms as well [61]. Moreover, prior studies using videoconferencing group CBT for social anxiety have reported significant reductions in DASS-21 scores [40,41]. A 1-day CBT-based workshop for secondary vocational students with social anxiety also significantly reduced DASS-21 scores, with moderate effect sizes [29]. However, not all prior findings are consistent: an in-person group CBT intervention for social anxiety using a full treatment protocol found no improvements in depressive symptoms [62]. Taken together, the absence of significant change in DASS-21 scores in this study may suggest that the MBPG via videoconference has limitations in the intervention's depth. Compared to full-protocol CBT, the brief, massed format involves fewer techniques and a considerably shortened duration. Specifically, spaced CBT interventions commonly provide opportunities for participants to practice newly learned skills between weekly sessions through homework assignments, thereby consolidating gains and facilitating symptom improvement [63]. Although the MBPG also allowed participants to practice during sessions and lunch breaks, these chances were still limited, which in turn constrained the extent to which negative beliefs and cognitive biases could be corrected. This may inherently limit its capacity to address broader emotional symptom domains beyond social anxiety or reduce the extent of therapeutic change achievable within participants. In addition, the DASS-21 primarily assesses physiological arousal and subjective emotional experiences, rather than cognitive content. As this intervention targeted participants' cognitive biases, improvements in social anxiety were likely driven by cognition-focused changes. Such changes,

XSL•FO
RenderX

however, are not readily captured by the DASS-21. Although the DASS-21 is commonly used in social anxiety interventions, it may be less suitable for detecting the specific types of cognitive changes targeted in this study.

The outcomes related to mental health literacy on social anxiety were somewhat complex, with the corresponding hypotheses only partially supported. Participants' knowledge of social anxiety significantly improved, reflecting the most direct impact of the psychoeducational component and aligning with the study's hypotheses. However, no significant changes were observed in social anxiety stigma or in attitudes toward seeking professional help. Prior research has shown that when individuals' mental health literacy regarding social anxiety improves, they tend to adopt a more objective view of the condition, experience reduced stigma, and show a greater willingness to seek professional help [64]. Furthermore, Cui et al [54] conducted a 6-week, in-person psychoeducation group with the same measurements and found significant improvements in knowledge, stigma, and help-seeking attitudes. We believe several factors may account for the discrepancy between their findings and ours. First, disclosure among people with mental illness is an effective strategy to manage stigma [65], yet opportunities for interpersonal contact were highly constrained in the online format compared with in-person groups. This limitation likely hindered meaningful changes in stigma. Second, help-seeking attitudes are influenced by multiple factors, including mental health literacy and perceived service accessibility [66], and improving these attitudes requires shifts in participants' deeper beliefs and motivations [64]. However, participants in the online workshop were recruited from diverse regions across the country, resulting in variability in access to mental health resources. For participants from low-resource backgrounds, severely limited perceived accessibility of services may have constrained improvements in help-seeking attitudes. Therefore, no significant change in help-seeking attitudes was observed at the group level. Nonetheless, these findings tentatively suggest that improvements in social anxiety symptoms may occur independently of changes in stigmatizing attitudes toward social anxiety.

Implications, Limitations, and Future Research

This study has several notable strengths. First, it employs a videoconference platform to deliver a new format of intervention, that is, the MBPG, which offers numerous advantages. The videoconferencing delivery format overcomes geographical barriers, increasing access to evidence-based interventions. At the same time, however, online delivery inevitably reduces opportunities for in-person exposure. It must be carefully considered when selecting intervention techniques, as the effectiveness of exposure-based strategies may be more substantially limited. In this intervention, both the cognitive restructuring and behavioral experiments were cognition-targeted techniques that do not rely on exposure in vivo, thereby minimizing the potential impact of the videoconferencing format on treatment effects. The massed format—delivered over a short, tightly scheduled period—may improve efficiency and reduce dropout. Moreover, the psychoeducational group format accommodates more participants per session, and its highly structured, slide-based delivery reduces the demands on group leaders while ensuring treatment fidelity. Finally, the use of a randomized controlled trial design provides preliminary yet promising evidence for the efficacy of this intervention model.

The study also has some limitations. First, the sample size was relatively small, and we conducted our intervention only in the nonclinical sample. Future research should expand upon these findings by conducting large-scale, multicenter randomized controlled trials in both clinical and community samples to further evaluate the effects of the MBPG-based program for social anxiety. Second, videoconferencing-based CBT has inherent constraints compared to the face-to-face format. These include increased exposure to distractions, reduced attentional engagement, and limited ability for leaders to observe and monitor participant cues [67], which further reduce opportunities for direct contact and spontaneous communication both between the leaders and participants and among participants themselves. Such limitations may reduce intervention efficacy, warranting improved technological solutions and content organization to address these challenges. In addition, compared to traditional treatments, the massed brief intervention format offers limited opportunities for participants to practice newly acquired cognitive skills. As a result, these skills may not be sufficiently reinforced, which in turn constrains the depth of therapeutic change. Future iterations of the intervention protocol should incorporate additional strategies to address this limitation. Finally, the follow-up period was limited to 3 months. Longer-term follow-up (eg, 6 months, 1 year, or beyond) is necessary to assess the maintenance of treatment gains over time.

We finally derived several implications from this intervention for future research. First, although online delivery is convenient and lowers access barriers, it is important to select a platform that supports essential interactive features, particularly breakout rooms and spaces that allow participants to communicate more freely, which may partially mitigate the limitations associated with reduced in-person exposure. Second, because the content of this workshop was delivered in a highly structured, slide-based format, future dissemination could broaden not only the reach of the program but also the range of potential leaders. Beyond mental health professionals, trained paraprofessionals or nonspecialists (eg, peer counselors or students majoring in psychology or related fields) could deliver the intervention using standardized materials, such as guidance menus and slide decks, thereby enhancing the accessibility of mental health care.

Conclusion

This study demonstrates that a videoconferencing CBT-based massed brief psychoeducational group can produce meaningful and sustained improvements in social anxiety among university students. Future research should focus on scaling up this model through larger randomized controlled trials, adapting it for more diverse populations, and exploring the feasibility of delivery by nonspecialist leaders.

XSL•FO
RenderX

Funding

This research was funded by the Fundamental Research Funds for the Central Universities,and the Research Funds of Renmin University of China (22XNKJ27).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Data curation: LF, WL
Formal analysis: LF, WL
Methodology: LF
Investigation: WL, XT
Funding acquisition: XT
Conceptualization: DD
Visualization: LC, DD
Resources: XT
Supervision: XT
Writing-review & editing: LF, LC, XT
Writing-original drafting: LF

Conflicts of Interest

None declared.

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 1282 KB - jmir_v28i1e79825_app1.pdf]

References

1. Diagnostic and Statistical Manual of Mental Disorders (5th Ed): American Psychiatric Association; 2013. [doi: 10.1176/appi.books.9780890425596]
2. Morrison AS, Heimberg RG. Social anxiety and social anxiety disorder. Annu Rev Clin Psychol 2013;9(1):249-274. [doi: 10.1146/annurev-clinpsy-050212-185631] [Medline: 23537485]
3. Tang X, Liu Q, Cai F, Tian H, Shi X, Tang S. Prevalence of social anxiety disorder and symptoms among Chinese children, adolescents and young adults: a systematic review and meta-analysis. Front Psychol 2022;13:792356. [doi: 10.3389/fpsyg.2022.792356] [Medline: 36072051]
4. Aderka IM, Hofmann SG, Nickerson A, Hermesh H, Gilboa-Schechtman E, Marom S. Functional impairment in social anxiety disorder. J Anxiety Disord 2012 Apr;26(3):393-400. [doi: 10.1016/j.janxdis.2012.01.003] [Medline: 22306132]
5. Tonge NA, Lim MH, Piccirillo ML, Fernandez KC, Langer JK, Rodebaugh TL. Interpersonal problems in social anxiety disorder across different relational contexts. J Anxiety Disord 2020 Oct;75:102275. [doi: 10.1016/j.janxdis.2020.102275] [Medline: 32891027]
6. Hayward C, Wilson KA, Lagle K, Kraemer HC, Killen JD, Taylor CB. The developmental psychopathology of social anxiety in adolescents. Depress Anxiety 2008;25(3):200-206. [doi: 10.1002/da.20289] [Medline: 17348001]
7. Russell G, Topham P. The impact of social anxiety on student learning and well-being in higher education. J Ment Health 2012 Aug;21(4):375-385. [doi: 10.3109/09638237.2012.694505] [Medline: 22823093]
8. Seo EH, Yang HJ, Kim SG, Yoon HJ. Ego-resiliency moderates the risk of depression and social anxiety symptoms on suicidal ideation in medical students. Ann Gen Psychiatry 2022 Jun 18;21(1):19. [doi: 10.1186/s12991-022-00399-x] [Medline: 35717375]
9. Barkowski S, Schwartze D, Strauss B, Burlingame GM, Barth J, Rosendahl J. Efficacy of group psychotherapy for social anxiety disorder: a meta-analysis of randomized-controlled trials. J Anxiety Disord 2016 Apr;39:44-64. [doi: 10.1016/j.janxdis.2016.02.005] [Medline: 26953823]
10. Mayo-Wilson E, Dias S, Mavranezouli I, et al. Psychological and pharmacological interventions for social anxiety disorder in adults: a systematic review and network meta-analysis. Lancet Psychiatry 2014 Oct;1(5):368-376. [doi: 10.1016/S2215-0366(14)70329-3] [Medline: 26361000]
11. Sun L, Dai X, Zhu S, Liu Z, Zhongming Z. Psychotherapies for social anxiety disorder in adults: a systematic review and Bayesian network meta-analysis. J Affect Disord 2025 Jun 1;378:301-319. [doi: 10.1016/j.jad.2025.02.092] [Medline: 40023260]

XSL•FO
RenderX

12. Pilling S, Mayo-Wilson E, Mavranezouli I, et al. Recognition, assessment and treatment of social anxiety disorder: summary of NICE guidance. BMJ 2013 May 22;346:f2541. [doi: 10.1136/bmj.f2541] [Medline: 23697669]
13. Firth J, Torous J, Nicholas J, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017 Oct;16(3):287-298. [doi: 10.1002/wps.20472] [Medline: 28941113]
14. Hedman E, Andersson G, Ljótsson B, et al. Internet-based cognitive behavior therapy vs. cognitive behavioral group therapy for social anxiety disorder: a randomized controlled non-inferiority trial. PLoS One 2011 Mar 25;6(3):e18001. [doi: 10.1371/journal.pone.0018001] [Medline: 21483704]
15. Leichsenring F, Leweke F. Social anxiety disorder. N Engl J Med 2017 Jun 8;376(23):2255-2264. [doi: 10.1056/NEJMcp1614701] [Medline: 28591542]
16. Orsolini L, Pompili S, Salvi V, Volpe U. A systematic review on telemental health in youth mental health: focus on anxiety, depression and obsessive-compulsive disorder. Medicina (Kaunas) 2021 Jul 31;57(8):793. [doi: 10.3390/medicina57080793] [Medline: 34440999]
17. Ganho-Ávila A, Vieira Figueiredo D, Vagos P. Online cognitive therapy for social anxiety disorder in adolescence: a clinical case study using the CT@TeenSAD. Clin Case Stud 2022 Dec;21(6):533-551. [doi: 10.1177/15346501221091519]
18. Peros OM, Webb L, Fox S, Bernstein A, Hoffman L. Conducting exposure-based groups via telehealth for adolescents and young adults with social anxiety disorder. Cogn Behav Pract 2021 Nov;28(4):679-689. [doi: 10.1016/j.cbpra.2021.04.001] [Medline: 34690482]
19. Winter HR, Norton AR, Burley JL, Wootton BM. Remote cognitive behaviour therapy for social anxiety disorder: a meta-analysis. J Anxiety Disord 2023 Dec;100:102787. [doi: 10.1016/j.janxdis.2023.102787] [Medline: 37890219]
20. Vu BP, Dang HM, Andersen PN. Does videoconferencing-based cognitive behavioral therapy for anxious youth work? A systematic review of the literature. Curr Treat Options Psych 2023 Sep 2;10(4):511-533. [doi: 10.1007/s40501-023-00302-9]
21. Social anxiety disorder: recognition, assessment and treatment [clinical guideline no. CG159]. : National Institute for Health and Care Excellence; 2013 URL: https://www.nice.org.uk/guidance/cg159 [accessed 2025-12-19]
22. Bennett S, Myles-Hooton P, Schleider J, Shafran R, editors. Oxford Guide to Brief and Low Intensity Interventions for Children and Young People: Oxford University Press; 2022. [doi: 10.1093/med-psych/9780198867791.001.0001]
23. Shafran R, Myles-Hooton P, Bennett S, Öst LG. The concept and definition of low intensity cognitive behaviour therapy. Behav Res Ther 2021 Mar;138:103803. [doi: 10.1016/j.brat.2021.103803] [Medline: 33540242]
24. Cape J, Whittington C, Buszewicz M, Wallace P, Underwood L. Brief psychological therapies for anxiety and depression in primary care: meta-analysis and meta-regression. BMC Med 2010 Jun 25;8(1):38. [doi: 10.1186/1741-7015-8-38] [Medline: 20579335]
25. Heimberg RG, Liebowitz MR, Hope DA, Schneier FR, editors. Social Phobia: Diagnosis, Assessment, and Treatment: Guilford Press; 1995. [doi: 10.1192/S0007125000147543]
26. Heimberg RG, Becker RE. Cognitive-Behavioral Group Therapy for Social Phobia: Basic Mechanisms and Clinical Strategies: Guilford Press; 2002.
27. Gallagher HM, Rabian BA, McCloskey MS. A brief group cognitive-behavioral intervention for social phobia in childhood. J Anxiety Disord 2004;18(4):459-479. [doi: 10.1016/S0887-6185(03)00027-6] [Medline: 15149708]
28. Jain N, Stech E, Grierson AB, et al. A pilot study of intensive 7-day internet-based cognitive behavioral therapy for social anxiety disorder. J Anxiety Disord 2021 Dec;84:102473. [doi: 10.1016/j.janxdis.2021.102473] [Medline: 34534800]
29. Li H, Tang X. Effects of a cognitive behavioral therapy-based workshop intervention on social anxiety among secondary vocational students: a randomized controlled trial. Behav Ther 2025 Nov;56(6):1118-1132. [doi: 10.1016/j.beth.2025.05.005] [Medline: 41139107]
30. Gladding ST. Groups: A Counseling Specialty: Pearson; 2016.
31. Brown BM. Psychoeducation group work. Couns Human Dev 1997;29(7):1-14.
32. Kutcher S, Wei Y, Coniglio C. Mental health literacy: past, present, and future. Can J Psychiatry 2016 Mar;61(3):154-158. [doi: 10.1177/0706743715616609] [Medline: 27254090]
33. Houghton S, Saxon D. An evaluation of large group CBT psycho-education for anxiety disorders delivered in routine practice. Patient Educ Couns 2007 Sep;68(1):107-110. [doi: 10.1016/j.pec.2007.05.010] [Medline: 17582724]
34. Öst LG, Cederlund R, Reuterskiöld L. Behavioral treatment of social phobia in youth: does parent education training improve the outcome? Behav Res Ther 2015 Apr;67:19-29. [doi: 10.1016/j.brat.2015.02.001] [Medline: 25727679]
35. Yalom ID, Leszcz M. The Theory and Practice of Group Psychotherapy, 6th edition: Basic Books; 2020. [doi: 10.1176/appi.psychotherapy.20210007]
36. Öst LG, Ollendick TH. Brief, intensive and concentrated cognitive behavioral treatments for anxiety disorders in children: a systematic review and meta-analysis. Behav Res Ther 2017 Oct;97:134-145. [doi: 10.1016/j.brat.2017.07.008] [Medline: 28772195]
37. Deacon B, Abramowitz J. A pilot study of two-day cognitive-behavioral therapy for panic disorder. Behav Res Ther 2006 Jun;44(6):807-817. [doi: 10.1016/j.brat.2005.05.008] [Medline: 16084488]

XSL•FO
RenderX

38. Avramchuk O, Nizdran-Fedorovych O, Blozva P, Plevachuk O. Internet-delivered low-intensity CBT for people with social anxiety disorder in a period of COVID-19: results of pilot research. Wiad Lek 2022;75(12):3109-3114. [doi: 10.36740/WLek202212136] [Medline: 36723335]
39. Noda S, Shirotsuki K, Nakao M. Low-intensity mindfulness and cognitive-behavioral therapy for social anxiety: a pilot randomized controlled trial. BMC Psychiatry 2024 Mar 7;24(1):190. [doi: 10.1186/s12888-024-05651-0] [Medline: 38454396]
40. Nauphal M, Swetlitz C, Smith L, Rosellini AJ. A preliminary examination of the acceptability, feasibility, and effectiveness of a telehealth cognitive-behavioral therapy group for social anxiety disorder. Cogn Behav Pract 2021 Nov;28(4):730-742. [doi: 10.1016/j.cbpra.2021.04.011]
41. Shapiro IR, Boyd JE, McCabe RE, Rowa K. Lost connection? Comparing group cohesion and treatment outcomes between videoconference and in-person cognitive behavioural group therapy for social anxiety disorder and other anxiety disorders. Behav Cogn Psychother 2025 Jun;53(2):159-173. [doi: 10.1017/S1352465825000013] [Medline: 40077887]
42. Wang Y, Chen X, Chen X, et al. Intervention effects of Internet-based cognitive behavior therapy on social anxiety among medical students. Chin J Sch Health 2013;34(2):139-141. [doi: 10.16835/j.cnki.1000-9817.2013.02.006]
43. Fernandez E, Salem D, Swift JK, Ramtahal N. Meta-analysis of dropout from cognitive behavioral therapy: magnitude, timing, and moderators. J Consult Clin Psychol 2015 Dec;83(6):1108-1122. [doi: 10.1037/ccp0000044] [Medline: 26302248]
44. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
45. Hope DA, Heimberg RG, Turk CL. Managing Social Anxiety, Therapist Guide: A Cognitive-Behavioral Therapy Approach, 3rd edition: Oxford University Press; 2019. [doi: 10.1093/med-psych/9780190247591.001.0001]
46. Leigh E, Clark DM. Understanding social anxiety disorder in adolescents and improving treatment outcomes: applying the cognitive model of Clark and Wells (1995). Clin Child Fam Psychol Rev 2018 Sep;21(3):388-414. [doi: 10.1007/s10567-018-0258-5] [Medline: 29654442]
47. Hofmann SG, Otto MW. Cognitive Behavioral Therapy for Social Anxiety Disorder: Evidence-Based and Disorder-Specific Treatment Techniques: Routledge; 2018. [doi: 10.4324/9781315617039]
48. Connor KM, Davidson JR, Churchill LE, Sherwood A, Foa E, Weisler RH. Psychometric properties of the Social Phobia Inventory (SPIN). New self-rating scale. Br J Psychiatry 2000 Apr;176(4):379-386. [doi: 10.1192/bjp.176.4.379] [Medline: 10827888]
49. Xiao R, Wu W, Zhang W. The reliability and validity of the Chinese version of social phobia inventory. West China Med J 2007(3):477-479 [FREE Full text]
50. Leary MR. A brief version of the Fear of Negative Evaluation Scale. Pers Soc Psychol Bull 1983 Sep;9(3):371-375. [doi: 10.1177/0146167283093007]
51. Chen Z. Fear of negative evaluation and test anxiety in middle school students. Chinese Mental Health J 2002(12):855-857 [FREE Full text]
52. Lovibond PF, Lovibond SH. The structure of negative emotional states: comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. Behav Res Ther 1995 Mar;33(3):335-343. [doi: 10.1016/0005-7967(94)00075-u] [Medline: 7726811]
53. Gong X, Xie X, Xu R, Luo Y. Psychometric properties of the Chinese versions of DASS-21 in Chinese college students. Chinese J Clin Psychol 2010;18(4):443-446. [doi: 10.16128/j.cnki.1005-3611.2010.04.020]
54. Cui L, Zhang J, Hu H, Jia S, Tang X. Effectiveness and feasibility of a cognitive-behavioral therapy-based brief psychoeducational group intervention for social anxiety in university students. Res Soc Work Pract 2025 Nov 28. [doi: 10.1177/10497315251399894]
55. Fischer EH, Farina A. Attitudes toward seeking professional psychological help: a shortened form and considerations for research. J Coll Stud Dev 1995;36(4):368-373 [FREE Full text]
56. Fang S, Yang B, Yang F, Zhou Y. Study on reliability and validity of the Chinese version of the Attitudes toward Seeking Professional Psychological Help Scale-Short Form for community population in China. Chinese Nurs Res 2019;33(14):2410-2414. [doi: 10.12102/j.issn.1009-6493.2019.14.009]
57. Luke SG. Evaluating significance in linear mixed-effects models in R. Behav Res Methods 2017 Aug;49(4):1494-1502. [doi: 10.3758/s13428-016-0809-y] [Medline: 27620283]
58. Morris SB. Estimating effect sizes from pretest-posttest-control group designs. Organ Res Methods 2008 Apr;11(2):364-386. [doi: 10.1177/1094428106291059]
59. Pincus DB, Elkins RM, Hardway CL. Intensive treatments for adolescents with panic disorder and agoraphobia: helping youth move beyond avoidance. Psychopathol Rev 2014 Jul;a1(1):189-194. [doi: 10.5127/pr.033313]
60. Hofmann SG. Cognitive factors that maintain social anxiety disorder: a comprehensive model and its treatment implications. Cogn Behav Ther 2007;36(4):193-209. [doi: 10.1080/16506070701421313] [Medline: 18049945]
61. Santos B, Pinho L, Nogueira MJ, Pires R, Sequeira C, Montesó-Curto P. Cognitive restructuring during depressive symptoms: a scoping review. Healthcare (Basel) 2024 Jun 28;12(13):1292. [doi: 10.3390/healthcare12131292] [Medline: 38998827]

XSL•FO
RenderX

62. Rapee RM, McLellan LF, Carl T, et al. Comparison of transdiagnostic treatment and specialized social anxiety treatment for children and adolescents with social anxiety disorder: a randomized controlled trial. J Am Acad Child Adolesc Psychiatry 2023 Jun;62(6):646-655. [doi: 10.1016/j.jaac.2022.08.003] [Medline: 35987298]
63. McEvoy PM, Johnson AR, Kazantzis N, Egan SJ. Predictors of homework engagement in group CBT for social anxiety: client beliefs about homework, its consequences, group cohesion, and working alliance. Psychother Res 2024 Jan;34(1):68-80. [doi: 10.1080/10503307.2023.2286993] [Medline: 38109521]
64. Coles ME, Ravid A, Gibb B, George-Denn D, Bronstein LR, McLeod S. Adolescent mental health literacy: young people's knowledge of depression and social anxiety disorder. J Adolesc Health 2016 Jan;58(1):57-62. [doi: 10.1016/j.jadohealth.2015.09.017] [Medline: 26707229]
65. Corrigan PW, Larson JE, Michaels PJ, et al. Diminishing the self-stigma of mental illness by coming out proud. Psychiatry Res 2015 Sep 30;229(1-2):148-154. [doi: 10.1016/j.psychres.2015.07.053] [Medline: 26213379]
66. Hom MA, Stanley IH, Joiner TE. Evaluating factors and interventions that influence help-seeking and mental health service utilization among suicidal individuals: a review of the literature. Clin Psychol Rev 2015 Aug;40:28-39. [doi: 10.1016/j.cpr.2015.05.006] [Medline: 26048165]
67. Luo C, Sanger N, Singhal N, et al. A comparison of electronically-delivered and face to face cognitive behavioural therapies in depressive disorders: a systematic review and meta-analysis. EClinicalMedicine 2020 Jul;24:100442. [doi: 10.1016/j.eclinm.2020.100442] [Medline: 32775969]

Abbreviations

ATSPPH-SF: Attitudes Toward Seeking Professional Psychological Help Scale−Short Form
BFNES: Brief Fear of Negative Evaluation Scale
CBT: cognitive-behavioral therapy
DASS-21: Depression Anxiety Stress Scales
MBPG: Massed Brief Psychoeducational Group
SAKT: Social Anxiety Knowledge Test
SASI: Social Anxiety Stigma Inventory
SPIN: Social Phobia Inventory

Edited by A Schwartz, M Balcarras; submitted 30.Jun.2025; peer-reviewed by G Alcolado, S Chen; revised version received 15.Dec.2025; accepted 16.Dec.2025; published 06.Jan.2026.

Please cite as:
Feng L, Liu W, Cui L, Dobson D, Tang X
Cognitive-Behavioral Therapy−Based Massed Brief Psychoeducational Group via Videoconference for Social Anxiety: Randomized Controlled Trial
J Med Internet Res 2026;28:e79825
URL: https://www.jmir.org/2026/1/e79825
doi:10.2196/79825

© Lele Feng, Wei Liu, Liechuan Cui, Deborah Dobson, Xinfeng Tang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Smartphone-Based Digital Eczema Education Program for Atopic Dermatitis in Children Aged 0 to 6 Years: Multicenter, Randomized, Parallel Controlled Clinical Study

Huan Yang[1*], MD, PhD; Hong Shu[2*], MD; Liu-hui Wang[3*], MD; Ping Li[4*], MD, PhD; Yun-ling Li[5], MD; Qin-feng Li[6], MD; Xiu-ping Han[7], MD; Jing Tian[8], MD; Jing Chang[9], MD; Hua Qian[10], MD, PhD; Jing-ping Chen[11], MM; Xin-qiang Ding[12], MD; Pan-qian Wu[2], MD; Li-min Dou[3], MD; Zhen Luo[4], MD; Wei Li[5], MM; Yang-yang Lin[6], MD; Lin Li[7], MD; Shu-zhen Yue[9], MD; Yang Gu[10], MM; Li Yang[11], MD; Xiao-hong Sun[12], MD; Xiao-yan Luo[1*], MD, PhD; Lin Ma[8*], MD, PhD; Hua Wang[1*], MD, PhD

[1]Department of Dermatology, Children's Hospital of Chongqing Medical University, National Clinical Research Center for Child Health and Disorders, Ministry of Education Key Laboratory of Child Development and Disorders, NO.136 Zhongshan 2nd Road, Yuzhong District, Chongqing, China

[10]Department of Dermatology, Children's Hospital of Soochow University, Suzhou, China

[11]Department of Dermatology, Guangzhou Woman and Children's Medical Center, Guangzhou, China

[12]Department of Dermatology, Xi'an Children's Hospital, Xi'an, China

[2]Department of Dermatology, Kunming Children's Hospital, Kunming, China

[3]Department of Dermatology, Children's Hospital of Fudan University, Shanghai, China

[4]Department of Dermatology, Shenzhen Children's Hospital, Shenzhen, China

[5]Department of Dermatology, The Children's Hospital, Zhejiang University School of Medicine, National Clinical Research Center of Child Health, Hangzhou, China

[6]Department of Dermatology, Tianjin Children's Hospital, Tianjin, China

[7]Department of Dermatology, Shengjing Hospital of China Medical University, Shenyang, China

[8]Department of Dermatology, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing, China

[9]Department of Dermatology, Hunan Children's Hospital, Changsha, China

[*]these authors contributed equally

Corresponding Author:
Hua Wang, MD, PhD
Department of Dermatology, Children's Hospital of Chongqing Medical University, National Clinical Research Center for Child Health and Disorders, Ministry of Education Key Laboratory of Child Development and Disorders, NO.136 Zhongshan 2nd Road, Yuzhong District, Chongqing, China

Abstract

Background: Atopic dermatitis (AD) is a chronic, relapsing inflammatory skin condition that affects approximately 10% to 20% of children, imposing substantial health and economic burdens. Although education for patients and caregivers is acknowledged as a crucial element in the management of AD, conventional approaches, such as workshops or in-person consultations, are often resource intensive and face challenges related to scalability, personalization, and relapse prevention. Digital tools present promising alternatives; however, empirical evidence supporting their effectiveness in young children is currently limited.

Objective: This study aimed to evaluate whether a smartphone-based patient-caregiver educational program could reduce relapse rates in children aged 0 to 6 years with moderate-to-severe AD, compared with conventional outpatient consultation alone.

Methods: In this multicenter, randomized, parallel-controlled trial, 615 children were enrolled across 12 tertiary pediatric dermatology centers in China and randomized (1:1) to receive either a smartphone-based digital education program with standard care (intervention group) or conventional face-to-face consultation only (control group). The 12-week digital program, delivered via the WeChat-based *Skin Care E-Station* platform, included structured multimedia modules, interactive educational materials, and a dynamic electronic action plan tailored to the child's age and disease stage. The primary endpoint was the 12-week relapse rate after the acute treatment phase. The secondary endpoints included changes in disease severity (Scoring Atopic Dermatitis, Peak Pruritus Numerical Rating Scale, and Patient-Oriented Eczema Measure) and quality of life (Children's Dermatology Life Quality Index or Infant's Dermatitis Quality of Life Index and Dermatitis Family Impact) up to 52 weeks.

Results: Among 615 randomized participants (mean age 3.3, SD 1.7 y; n=317, 51.5% male), relapse at 12 weeks occurred significantly less frequently in the digital education group than in the control group (16.6% vs 24.0%; relative risk 0.69, 95% CI 0.50 - 0.96; *P*=.02). Kaplan-Meier analysis showed superior relapse-free survival over the first 100 days (hazard ratio 0.688,

XSL•FO
RenderX

95% CI 0.490 - 0.966; *P*=.03). Differences in relapse rates beyond 12 weeks and in secondary outcomes were not statistically significant. Engagement tracking indicated high adherence to the intervention, with 58.0% of caregivers maintaining regular weekly use of the digital platform.

Conclusions: A structured smartphone-based patient-caregiver educational intervention significantly reduced short-term relapse risk among young children with moderate-to-severe AD, likely through improved caregiver recognition and early management of disease flares. Although effects diminished beyond 12 weeks, this approach demonstrates that scalable digital education is a feasible and effective adjunct to standard care in pediatric AD. Future research should focus on sustaining engagement, optimizing long-term reinforcement, and assessing cost-effectiveness in diverse caregiver populations.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2000031474; https://www.chictr.org.cn/showproj.html?proj=32400

(J Med Internet Res 2026;28:e79559) doi:10.2196/79559

KEYWORDS

atopic dermatitis; digital education; smartphone-based intervention; caregiver engagement; relapse prevention; randomized controlled trial; pediatric dermatology; telemedicine portal; self-management; multicenter study

Introduction

Atopic dermatitis (AD) is a chronic inflammatory skin condition characterized by intense pruritus and eczematous lesions, accompanied by episodic exacerbations and persistent skin symptoms [1]. AD is recognized as a global health concern, affecting up to 20% of children and approximately 3% of adults worldwide [2]. In China, the prevalence of AD stands at approximately 30.48% among infants aged <1 year [3] and 12.94% among children aged 1 to 7 years [4], imposing substantial economic and public health burdens [5,6]. The recurrent nature of AD presents unique challenges, as patients often experience periods of exacerbation interspersed with phases of relative remission [1]. This cyclical pattern highlights the necessity for effective management strategies that not only address current symptoms but also equip patients and their caregivers with the knowledge and skills to prevent future recurrences [7].

Patient-caregiver education has emerged as a pivotal component of AD management, as highlighted in many international guidelines [8-12]. It is important to note that for pediatric AD populations, the primary focus of education is often on parents and caregivers, especially for infants and young children. As children grow older and their understanding develops, education can be progressively tailored and delivered directly to the patients themselves, in an age-appropriate manner. Providing patients and their caregivers with comprehensive information about the condition, its triggers, management strategies, and the importance of adherence to treatment regimens can empower them to take an active role in their own care [13]. Studies have shown that patients and families who receive structured educational programs are more likely to engage in treatment and preventive measures, leading to better overall outcomes and reduced disease severity, as evidenced by metrics such as Scoring Atopic Dermatitis (SCORAD) and Children's Dermatology Life Quality Index (CDLQI) [14]. Furthermore, patient-caregiver education may contribute to interrupting the "atopic march" toward comorbidities such as allergic rhinitis [15]. Despite the recognized importance of education, gaps remain in the current approaches used to educate patients with AD and their caregivers. Existing educational models, including workshops [16-19], eczema schools [20-23], and printed materials [19,24-26], are resource intensive, lack personalization, or fail to sufficiently address the specific concerns and experiences of patients and caregivers. As a result, patients and their families may feel overwhelmed or unsupported in managing their condition, leading to suboptimal adherence and an increased risk of recurrent flares [27,28]. In addition, prior investigations have predominantly focused on the short-term efficacy of educational interventions, whereas the impact of education on the prevention of AD relapse remains considerably underexplored [29].

The COVID-19 pandemic has instigated a paradigm shift toward digital health [30], resulting in an increased willingness among patients with AD and caregivers to use digital tools. Mobile platforms that integrate multimedia content, such as videos, interactive texts, and illustrated narratives, offer advantages in accessibility, engagement, and personalized learning experiences. These platforms theoretically address traditional limitations by providing scalable and adaptive interventions. However, empirical evidence supporting such innovations remains limited, particularly in children with AD.

In this multicenter, randomized controlled trial (RCT), we aim to assess the efficacy of a smartphone-based digital educational program in comparison to conventional outpatient consultation alone on the relapse rates of AD. By leveraging high smartphone penetration, this trial seeks to develop a more effective educational intervention that will ultimately reduce the burden of AD.

Methods

Study Design

This multicenter, parallel RCT (ChiCTR2000031474) was conducted within the pediatric dermatology departments of 12 tertiary public hospitals across China, including locations in Chongqing, Beijing, Liaoning, Shanghai, Shenzhen, Zhejiang, Hunan, Xi'an, Suzhou, Tianjin, Kunming, and Guangdong (Multimedia Appendix 1). The trial adhered to the CONSORT (Consolidated Standards of Reporting Trials) reporting guidelines (Checklist 1).

XSL•FO
RenderX

Participants

Children aged 0 to 6 years diagnosed with AD according to the American Academy of Dermatology Consensus criteria were eligible for inclusion if they presented with moderate-to-severe AD (SCORAD ≥25; Investigator's Global Assessment [IGA] ≥3) and had caregivers proficient in reading Chinese characters and using a smartphone. All participants were recruited through in-person clinical visits. Exclusion criteria included children with severe infections, psychiatric disorders, primary or secondary immune deficiencies, malignancies, or other medical conditions that significantly impair quality of life. Additionally, caregivers with mental disorders or cognitive impairment, prior participation in any patient-caregiver education program, and severe AD requiring systemic treatment were excluded.

Most caregivers were the patients' parents, and a minority were grandparents. All caregivers lived in the same household as the child, as families of "left-behind children," defined as children who remain in their hometowns in China for more than 6 months under the care of grandparents or other relatives while both parents migrate to urban areas for work, were excluded to avoid potential confounding due to social and environmental separation. Caregiver demographic details (eg, age and education level) were not collected, as the original study design focused on patient-level outcomes.

Randomization and Blinding

Eligible patients were randomly assigned in a 1:1 ratio to either the intervention group (smartphone-based digital education program plus standard care) or the control group (conventional outpatient consultation only). Randomization was conducted using a computer-generated number table with a block size of 4. The digital educational intervention protocol was developed and administered through a secure cloud-based platform operated by an independent clinical research organization (CRO) not involved in trial execution. This third-party CRO was responsible for maintaining the randomization database, delivering standardized digital educational content (including interactive modules, video tutorials, and self-assessment tools), and monitoring intervention adherence by automated adherence reminders (SMS text messaging or telephone) and digital footprint tracking (platform logins and content interaction time). Crucially, the CRO had no involvement in participant recruitment, clinical management, or outcome evaluation processes to ensure blinding integrity.

Intervention and Procedures

Overview

Before trial initiation, all participating dermatologists completed a 4-hour standardized training program on evidence-based patient-caregiver education.

At baseline enrollment, all patients were provided with a 1-page education leaflet (Multimedia Appendix 2) and completed disease severity assessments along with standardized questionnaires. Subsequently, participants entered a 2-week acute-phase treatment period, conducted strictly according to established clinical guidelines [10-12]. The acute-phase regimen included once-daily topical pharmacotherapy, 0.05% desonide or 0.1% hydrocortisone butyrate for those aged 0 to 2 years and 0.1% mometasone furoate for children older than 2 years, combined with nonpharmacological care. The nonpharmacological regimen consisted of twice-daily application of a ceramide-dominant emollient and daily bathing restricted to fewer than 10 minutes at water temperatures below 38°C. Patient status was reassessed every 14 (SD 2) days using the IGA. Patients achieving an IGA score ≤2 were advanced to the maintenance phase.

Digital Education Intervention

Upon entering the maintenance phase, participants in the intervention group were instructed to receive the digital health education program via the WeChat (Tencent Holdings Ltd)–based online platform "Skin Care E-Station" (Figure 1). This digital portal contained all educational components and could be accessed by scanning a QR code, without requiring any additional software installation. A procedural overview of patient enrollment and identity verification within the platform is provided in Multimedia Appendix 3. As WeChat is a widely used multifunctional communication, social, and payment application in China, this ensured high accessibility and user engagement. The platform consisted of three core components: (1) structured multimedia educational modules, (2) automated follow-up reminders, and (3) emergency response protocols for disease exacerbations.

XSL•FO
RenderX

Figure 1. Skin Care E-Station workflow. Participants in the digital education group accessed the clinical research organization–managed smartphone platform, receiving thrice-weekly education action plans (Mondays/Wednesdays/Fridays) for 12 wk alongside 54 structured multimedia modules (text/images/video/animations) distributed evenly throughout the 52-week program. The content covered AD pathophysiology, care routines, treatments, lifestyle, and mental health support, with real-time clinician access via an encrypted telemedicine portal (<8 h response) during flares. The illustration was created on the FIGDRAW 2.0 platform. AD: atopic dermatitis.

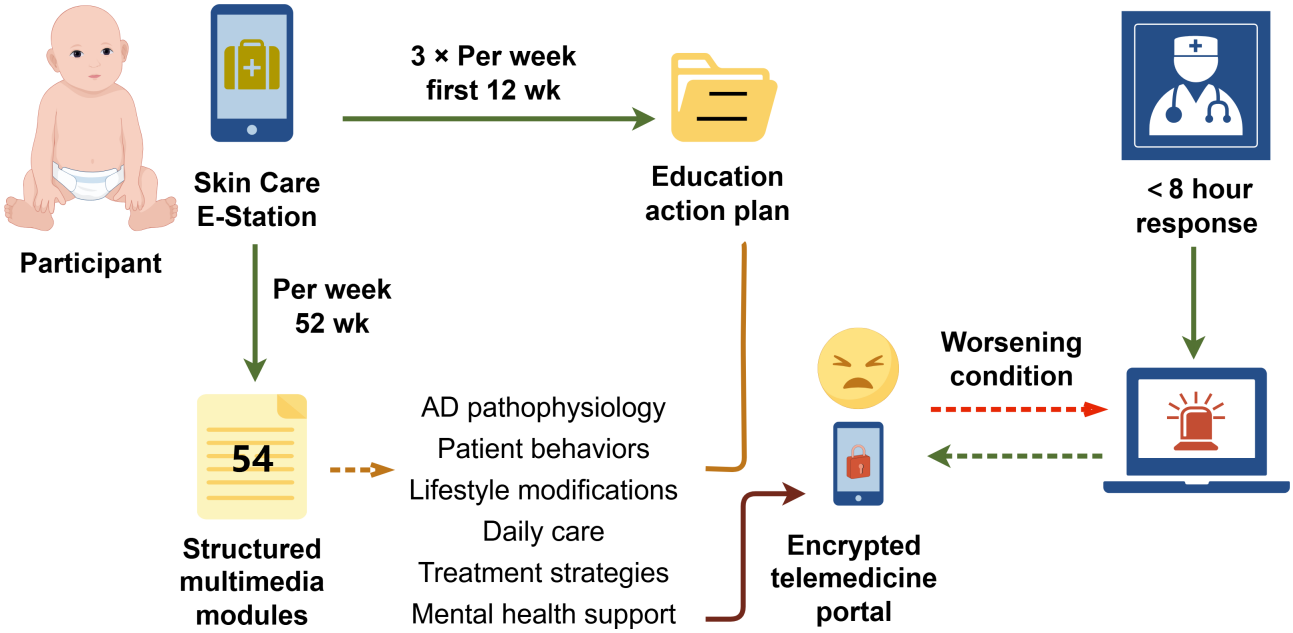

The digital education comprised 2 major elements. The first element was an electronic education action plan (EAP). Each participant in the intervention group received a digital, dynamically tailored EAP (Multimedia Appendix 4), customized according to the child's age group (1 - 2 y vs 2 - 6 y) and disease phase (acute flare vs maintenance). The EAP provided phase-specific guidance on bathing routines, emollient application, medication use, self-assessment of disease severity, and access to emergency consultations. The EAP was automatically delivered 3 times per week (Mondays, Wednesdays, and Fridays) for 12 consecutive weeks. The second element was structured multimedia educational modules. The core educational curriculum was collaboratively developed by the 12 participating centers, integrating an analysis of the most frequent caregiver concerns and key clinician recommendations from routine practice, supplemented by insights and experiences from other research teams [17-24]. This content was organized into 6 domains (AD pathophysiology, patient behaviors, lifestyle modifications, daily care, treatment strategies, and mental health support) and subsequently produced by an independent third-party company into 54 structured multimedia modules. These modules were delivered in formats tailored for children aged 0 to 6 years and their caregivers, including illustrated text (sample in Multimedia Appendix 5), images (sample in Multimedia Appendix 6), videos (sample in Multimedia Appendix 7), and animated stories (sample in Multimedia Appendix 8). The intervention group did not receive any additional systematic face-to-face patient education during scheduled follow-up visits or at unscheduled clinical encounters.

Control Group

Participants in the control group received only conventional education, consisting of a 15-minute face-to-face counseling session conducted at each scheduled follow-up visit. They were also provided with access to the "Skin Care E-Station" platform for noneducational purposes, including data collection and automated follow-up reminders, but did not receive any educational content, such as the EAP or structured multimedia modules.

Maintenance-Phase Treatment Regimen

Both the intervention and control groups received identical pharmacological treatments during the maintenance phase. In the 0 to 2 years age group, maintenance therapy comprised twice-weekly application of 0.05% desonide or 0.1% hydrocortisone butyrate cream, together with daily emollient use. In children aged >2 years, maintenance consisted of twice-weekly application of 0.03% tacrolimus ointment, alongside regular emollient therapy.

Assessments and Outcomes

Assessments were conducted at 4, 8, 12, 24, 36, and 52 weeks following randomization. Disease severity was evaluated using SCORAD, IGA, Peak Pruritus Numerical Rating Scale (PP-NRS) for pruritus, and the Patient-Oriented Eczema Measure (POEM). Quality of life was assessed using the Dermatitis Family Impact (DFI), the CDLQI, and the Infant's Dermatitis Quality of Life Index (IDQOL). The assessments were completed as follows: If the child was able to attend an in-person visit, the evaluations were performed by physicians not involved in the educational intervention; if not, caregivers uploaded photos and completed the PP-NRS, POEM, Quality of Life (QoL) questionnaires, and SCORAD self-assessment (Multimedia Appendix 9) on the platform. Subsequently, physicians not involved in the educational intervention performed the IGA and SCORAD assessments based on the uploaded photos and the SCORAD self-assessment form.

In addition, the timing of each relapse was meticulously recorded. Relapse was defined as an increase of ≥10 points in

XSL•FO
RenderX

the SCORAD score relative to the value recorded at the conclusion of the 2-week acute-phase treatment. The primary endpoint was the relapse rate at 12 weeks. The secondary endpoints included the changes in disease severity scores (SCORAD, PP-NRS, and POEM) and quality of life scores (CDLQI/IDQOL and DFI) from week 2 (end of the acute-phase treatment) to weeks 4, 8, 12, 24, 36, and 52 in both groups.

Statistical Analyses

Statistical analyses were performed using R (version 4.0.5; R Foundation for Statistical Computing). Sample size estimation was based on an expected 12-week relapse rate of 30% in the digital education group and 45% in the control group, which were derived from a multicenter randomized controlled clinical study on long-term intermittent maintenance therapy for children with AD conducted in China [31]. Assuming a 1:1 allocation ratio and a 30% expected dropout rate, an empirically supported adjustment commonly recommended in sample size estimation to maintain statistical power for both completer and intention-to-treat (ITT) analyses [32], a minimum of 229 patients per group was required to achieve 80% power with an α level of .05.

All analyses followed the ITT principle, which includes all randomized participants in the groups to which they were originally assigned. This approach maintains the integrity of randomization and provides a conservative estimate of treatment effectiveness in real-world clinical settings. To address missing data, multiple imputation by chained equations (MICEs) was performed under the assumption that the data were missing at random. Missing at random implies that the probability of missingness depends only on observed data and not on unobserved data. MICEs generate multiple plausible values for missing data based on observed variables, and we used 5 imputations (m=10) to account for uncertainty. The results from the imputed datasets were combined according to Rubin's rules. Normality of continuous variables was evaluated using the Kolmogorov-Smirnov test, a nonparametric test that compares the empirical distribution of the data with a theoretical normal distribution to assess deviations from normality. Data with normal distributions were presented as mean (SD) and compared using independent *t* tests. Nonnormally distributed data were presented as median (IQR) and analyzed using the Mann-Whitney test.

For the primary outcome, the 12-week relapse rate was compared between groups using the chi-square test. Time to relapse over 100 days was visualized using Kaplan-Meier curves and compared using the log-rank test. Cox proportional hazards regression was used to estimate hazard ratios (HRs) and 95% CIs, both unadjusted and adjusted for age and sex.

For continuous outcomes repeatedly measured over time (SCORAD, POEM, PP-NRS, IDQOL/CDLQI, and DFI), population-averaged generalized estimating equations were used with a Gaussian family, identity link, exchangeable correlation structure, and robust SEs to account for within-subject correlations. Each model was adjusted for baseline score, age, and sex. Missing data in these repeated measures were handled using MICEs (10 imputations), with final estimates combined using Rubin's rules.

A 2-sided *P* value of .04 was considered statistically significant.

Ethical Considerations

This study was reviewed and approved by the Medical Research Ethics Committee of the Children's Hospital of Chongqing Medical University (approval No. 2019-44-1). Written informed consent was obtained from all participants and their guardians before their involvement in the study. All collected data were deidentified to protect participant privacy and confidentiality. No compensation was provided to the participants.

Results

Participant Flow and Cohort Characteristics

The multicenter trial was conducted across 12 tertiary dermatology centers in China from July 2020 to March 2021, with a detailed CONSORT flow presented in Figure 2. Of 980 screened children with moderate-to-severe AD, 615 (62.8%) met inclusion criteria and were subsequently randomized into the digital education group (n=307, 49.9%) and the control group (n=308, 50.1%). Significant differences in attrition rates were observed at 12 weeks, with 20.2% (62/307) in the digital education group and 27.6% (85/308) in the control group (95% CI 0.66 - 0.98; χ^2_1=4.1; *P*=.04). Ultimately, 48.62% (299/615) completed the 52-week follow-up, comprising the per-protocol population: digital education (n=148, 24.1%) and control (n=151, 24.6%).

Baseline characteristics demonstrated adequate randomization balance (Table 1), with no significant differences observed in gender distribution (men: 52.1% vs 50.8%; *P*=.74), mean age (3.2, SD 1.8, vs 3.4, SD 1.7 y; *P*=.21), or disease severity measures, including SCORAD (49 [40, 59] vs 47.5 [37, 55]; *P*=.29), PP-NRS (6 [6, 8] vs 6 [5, 8]; *P*=.31) and POEM (15 [11, 20] vs 15 [11, 19]; *P*=.50). Similarly, no significant differences were noted in quality-of-life measures, including DFI (9 [4, 16] vs 10 [5,16]; *P*=.77) and QoL scores (13 [11, 15] vs 14 [11, 16]; *P*=.11). Quality of life was assessed using the CDLQI for children and IDQOL for infants.

XSL•FO
RenderX

Figure 2. CONSORT diagram of the randomization and follow-up of the study participants. AD: atopic dermatitis.

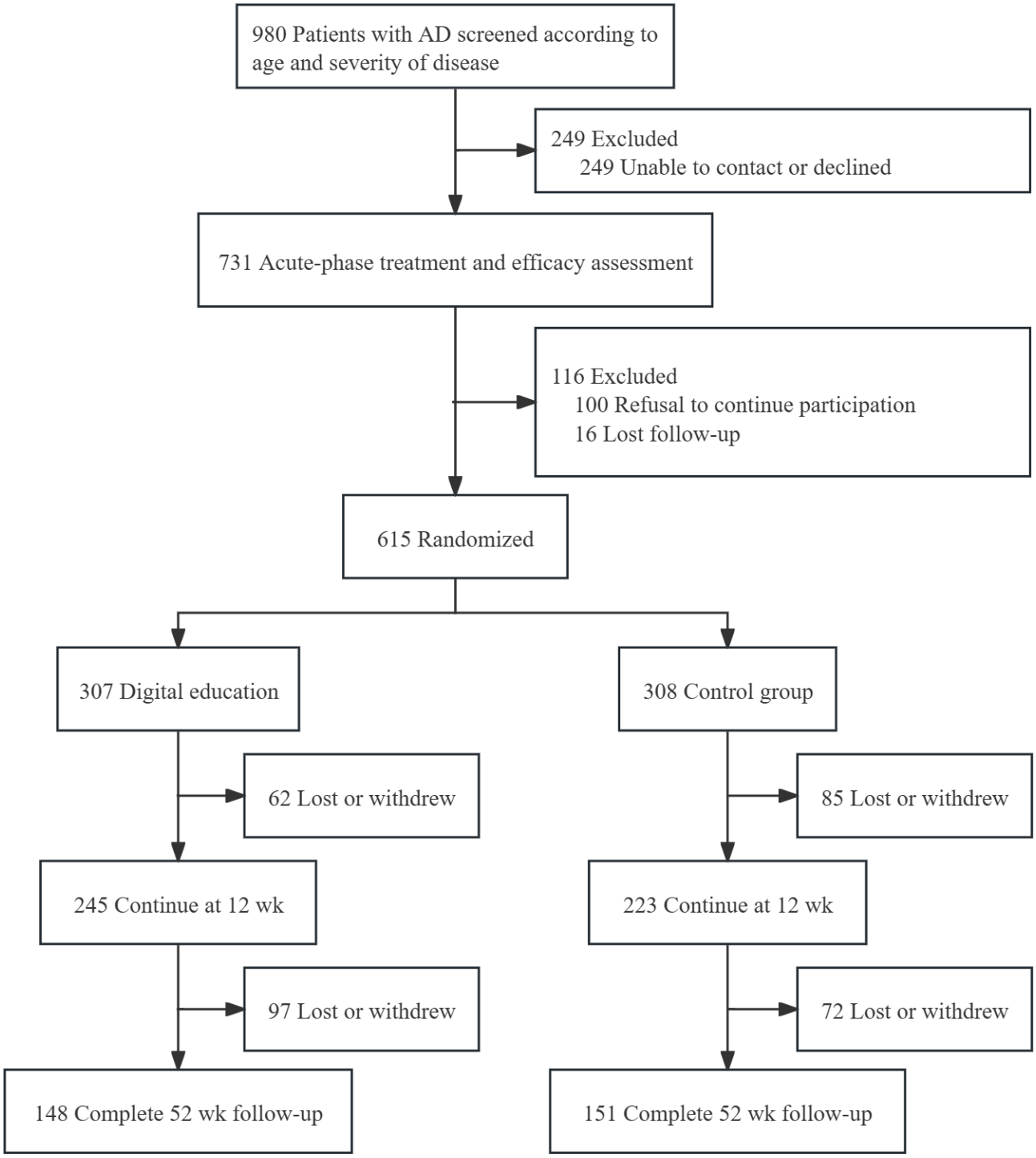

XSL•FO
RenderX

Table . The details of clinical characteristics of randomized patients. QoL was assessed using the Children's Dermatology Life Quality Index for children and the Infants' Dermatitis Quality of Life Index for infants.

Characteristic	Digital education (n=307)	Control (n=308)	*P* value
Gender, male (%)	52.1	50.8	.74
Age (y), mean (SD)	3.2 (1.8)	3.4 (1.7)	.21
SCORAD[a], median (Q1[b] ,Q3[c])	49 (40, 59)	47.5 (37, 55)	.29
PP-NRS[d], median (Q1, Q3)	6 (6, 8)	6 (5, 8)	.31
DFI[e], median (Q1, Q3)	9 (4, 16)	10 (5, 16)	.77
POEM[f], median (Q1, Q3)	15 (11, 20)	15 (11, 19)	.50
QoL[g], median (Q1, Q3)	13 (11, 15)	14 (11, 16)	.11

[a]SCORAD: Scoring Atopic Dermatitis.

[b]Q1: the first quartile.

[c]Q3: the third quartile.

[d]PP-NRS: Peak-Pruritus Numerical Rating Scale.

[e]DFI: Dermatitis Family Impact.

[f]POEM: Patient-Oriented Eczema Measure.

[g]QoL: quality of life.

Intervention Adherence and Participant Engagement

Intervention adherence, as measured by digital footprint tracking, demonstrated that 58.0% (178/307) of participants in the digital education group maintained regular engagement, averaging 1.64 (SD 0.38) sessions per week, and 26.7% (82/307) accessed the expedited physician consultation portals during disease flares. In the control arm, clinic attendance at the 52-week follow-up was 49.0 % (151/308). The 9.0% absolute difference in adherence between groups was statistically significant (Z=2.3, 95% CI 1.15%-16.85%; *P*=.03).

Primary Outcome: Relapse Prevention

The ITT analysis revealed a significant reduction in relapse rates in the digital education group at 12 weeks (16.6% [51/307] vs 24.0% [74/308]; relative risk [RR] 0.69, 95% CI 0.50 - 0.96; χ^2_1=5.1; *P*=.02). However, there was no statistically significant difference between the 2 groups at other time points: 4 weeks (8.79% vs 12.99%; RR 0.68, 95% CI 0.42 - 1.09; χ^2_1=2.8; *P*=.10), 8 weeks (15.64% vs 21.43%; RR 0.73, 95% CI 0.52 - 1.02; χ^2_1=3.4; *P*=.07), 24 weeks (21.82% vs 27.60%; RR 0.79, 95% CI 0.60 - 1.04; χ^2_1=2.8; *P*=.10), 36 weeks (23.13% vs 29.22%; RR 0.79, 95% CI 0.60 - 1.03; χ^2_1=3.0; *P*=.09), and 52 weeks (24.76% vs 31.17%; RR 0.79, 95% CI 0.62 - 1.03, χ^2_1=3.1; *P*=.08).

Kaplan-Meier analysis (100-d follow-up) further supported the 12-week finding, showing a significant difference in relapse risk between groups (HR] 0.688, 95% CI 0.490 - 0.966; χ^2_1=4.7; *P*=.03; Figure 3). In a Cox proportional hazards model adjusting for age and sex, the HR remained significant (adjusted HR 0.665, 95% CI 0.469 - 0.943; *P*=.02).

XSL•FO
RenderX

Figure 3. Kaplan-Meier estimates of relapse. Hundred-day relapse-free survival curves in the intention-to-treat set analysis.

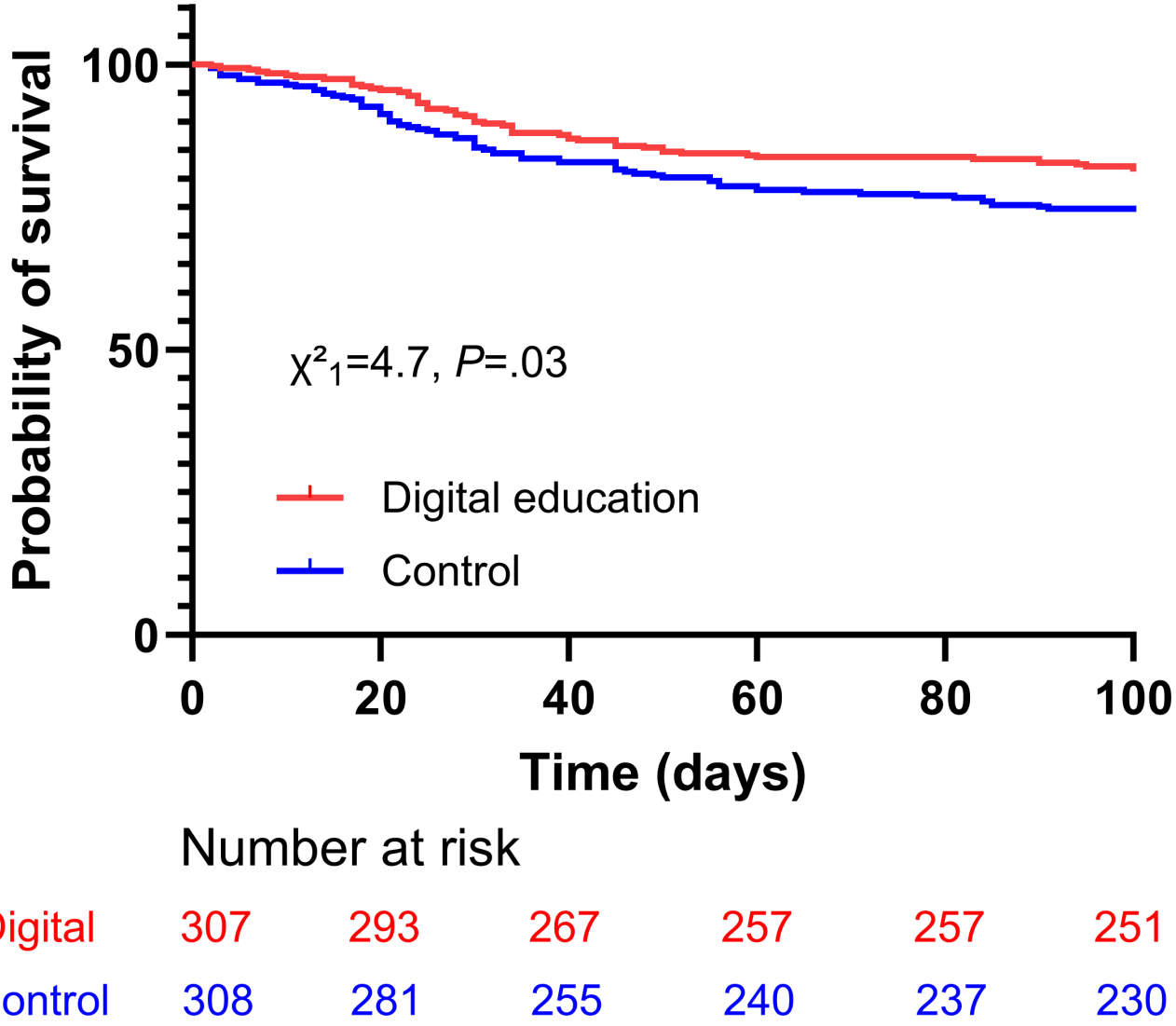

Secondary Outcomes: Disease Severity Trajectories

Changes in disease severity scores (including SCORAD, PP-NRS, and POEM) and QoL scores (IDQOL/CDLQI and DFI) from week 2 (the end of the acute treatment phase) were compared between the 2 groups at weeks 4, 8, 12, 24, 36, and 52. Using population-averaged GEE models adjusted for baseline score, age, and sex, the results demonstrated no statistically significant differences between the groups in the magnitude of score changes at any of the predefined follow-up time points (Multimedia Appendix 10).

Discussion

Principal Findings

This multicenter RCT demonstrated that a structured, smartphone-based multimodal digital education program for caregivers, when integrated with standard care, significantly reduced the substantial proportion of early relapses in children aged 0 to 6 years with moderate-to-severe AD, compared to conventional outpatient consultation alone. This early benefit was further supported by Kaplan-Meier survival analysis showing a clearer separation of relapse-free curves during the early follow-up period. Consistent with the study's primary objective, these results confirm that the digital educational intervention effectively achieved short-term relapse prevention. However, differences in relapse rates appeared to diminish over time, and no notable differences were observed in secondary end points assessing disease severity or quality of life across the 52-week follow-up.

The observed reduction in early relapse aligns with the core principle of AD management, which emphasizes the critical role of effective patient and caregiver education in improving clinical outcomes [8-12]. The success of our digital program likely stems from its multimodal design, which combined frequent, brief multimedia modules with a "Real-Time Clinician Access" pathway ("Green Channel"). This combination probably enhanced caregivers' ability to detect early signs of worsening and initiate prompt interventions, thereby preventing progression to full-blown relapse, even in the absence of measurable changes in objective severity indices. The platform's adherence rate suggested better engagement compared with what is often observed in traditional formats [33], indicating improved feasibility and user engagement.

XSL•FO
RenderX

Our findings are consistent with prior evidence demonstrating that structured education improves short-term management and adherence in pediatric AD [16-23,34-37]. However, they also highlight a common challenge: achieving long-term disease modulation with brief interventions is difficult. The diminishing effect on relapse prevention after the initial 12 weeks may be attributed to several factors. First, the structured program builds knowledge and habits, but its impact may wane without ongoing reinforcement [38]. Second, the natural relapsing-remitting course of AD may eventually overshadow the effects of a time-limited intervention. Furthermore, despite good initial adherence, participant engagement with digital health interventions often declines over time [39].

A critical component contributing to the early success was the integrated "Real-Time Clinician Access," used by a notable portion of participants during flares or uncertainty. This feature provided timely guidance, effectively replicating the benefit of rapid in-person intervention during exacerbations without the associated logistical burdens. This mechanism, improved early recognition and intervention facilitated by accessible education and on-demand clinical support, provides a plausible explanation for the significant reduction in early relapse rates, underscoring that the primary value of the intervention lies in optimizing the behavioral and self-management aspects of AD care.

The absence of significant differences in disease severity and QoL scores suggests that while the intervention improved caregivers' short-term flare management, it did not fundamentally alter the underlying core inflammatory process within the assessed follow-up period. This distinction has practical implications: digital education may be most valuable as a tool to reduce short-term exacerbations and health care utilization, complementing rather than replacing ongoing medical and pharmacological strategies needed to modify underlying disease activity.

Traditional educational formats (face-to-face counseling [34-37], workshops [16-19], and eczema schools [20-23] are effective but face significant scalability challenges due to their resource-intensive nature, geographical barriers, and time constraints for families and clinicians [35]. Passive educational materials, such as leaflets and videos, are limited by a lack of interactivity and personalization [26,36,37]. Our study addresses these limitations by leveraging widespread smartphone access to deliver a scalable, multimodal intervention. The platform's use of brief, engaging modules (animated stories, videos, and illustrated texts) delivered 3 times weekly made essential educational topics accessible to caregivers with varying literacy levels and busy schedules.

Limitations

Several limitations must be acknowledged. First, smartphone proficiency was an enrollment requirement, potentially excluding digitally underserved groups (eg, from rural or socioeconomically disadvantaged backgrounds) and limiting the generalizability of our findings to urban or digitally connected families. Future implementations should consider hybrid delivery models combining digital modules with printed materials or in-person sessions to enhance inclusivity. Second, understanding of the educational materials was not formally assessed, which may influence outcomes. Third, although early retention differed between groups, a considerable number of participants did not complete long-term follow-up, potentially affecting the accuracy of long-term outcome estimates. Fourth, although outcome assessors were blinded, caregivers and clinicians were not, creating potential for performance or reporting bias. Fifth, caregiver demographic variables (eg, age, education, and occupation) were not collected, limiting our ability to explore effect modification by caregiver characteristics. While the exclusion of 'left-behind children' and the predominance of parent caregivers likely reduced heterogeneity in living arrangements and direct caregiving availability, unmeasured socioeconomic or educational differences could still influence the uptake and use of digital education. Finally, while the control group also had access to the platform for noneducational purposes (eg, data entry, random allocation, and scheduling visits), they did not receive the structured educational modules or 'Green Channel' clinician access. Therefore, the observed effect likely reflects the combined impact of the educational content and the interactive digital delivery modality, rather than platform access alone. Future trials should include comparative arms using alternative digital formats (eg, app-based platforms and interactive chatbots) to disentangle which specific features most strongly influence adherence and relapse outcomes.

Conclusions and Broader Implications

In conclusion, this large multicenter RCT demonstrates that a structured, smartphone-based digital education program for caregivers, featuring interactive modules and real-time clinician support, can significantly reduce early relapse rates in young children with moderate-to-severe AD. This finding indicates that scalable digital education can strengthen short-term relapse prevention by optimizing caregiver empowerment and flare management, a strategy particularly relevant in regions with high smartphone penetration.

The broader implications of our study are 3-fold. First, given the high burden of AD and the scalability of digital tools, such interventions hold promise for improving access to quality education and potentially reducing health care disparities. Second, the lower dropout rate in the intervention group suggests that well-designed digital platforms may improve retention in long-term pediatric dermatology research and care. Finally, the transient nature of the benefit underscores that digital education should be viewed as a complementary adjunct to, not a replacement for, ongoing medical management.

Future research should prioritize (1) developing hybrid models that integrate digital education with targeted in-person support to enhance emotional connection and engagement; (2) designing strategies for sustained engagement, such as "booster" modules and adaptive content; (3) ensuring digital accessibility for underserved populations to address health equity; (4) conducting rigorous economic evaluations; and (5) conducting head-to-head comparisons of different digital features and mechanistic studies to identify the active ingredients of digital education.

XSL•FO
RenderX

Acknowledgments

The authors sincerely thank Mr. Lei Yan and Ms. Ping Yang for their kind support and assistance in the statistical analysis of the data. The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy (2025), translation was delegated to GAI tools under full human supervision. The GAI tool used was ChatGPT 5.

Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by Huan Yang.

Funding

This study was funded by Beijing Medical Award Foundation (2019 - 1115), the Joint Foundation of Chongqing Municipal Health Commission and Municipal Science and Technology Bureau (2021MSXM203), CQMU Program for Youth Innovation in Future Medicine (W0177), and Children's Hospital Affiliated to Chongqing Medical University (RC05036). The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Data Availability

The anonymized datasets generated and analyzed during this study are not publicly available due to participant privacy and institutional data protection policy but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: HY, HW, LM, XL
Methodology: HY, HW, LM, XL
Investigation: HY, HS, LW, PL, YL, QL, XH, JT, JC, HQ, JP Chen, XD, PW, LD, ZL, WL, YL, LL, SY, YG, LY, XS
Data Curation: HY, HS, LW, PL
Formal Analysis: HY, HS, LW, PL
Resources: HW, XL, HY
Software: HY
Validation: HY, HS, LW, PL
Funding Acquisition: HW, XL, HY
Project Administration: HW, LM, XL
Supervision: HW, LM, XL
Writing-Original Draft: HY
Writing-Review & Editing: HY, HS, LW
Visualization: HY, HS, LW, PL
All authors reviewed and approved the final manuscript and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of participating hospitals.
[PDF File, 66 KB - jmir_v28i1e79559_app1.pdf]

Multimedia Appendix 2
Atopic dermatitis awareness material.
[PDF File, 78 KB - jmir_v28i1e79559_app2.pdf]

Multimedia Appendix 3
Procedural overview illustrating the standardized workflow for patient enrollment and identity verification within the "Skin Care E-Station" platform, with representative interface screenshots supporting each operational step.
[PDF File, 294 KB - jmir_v28i1e79559_app3.pdf]

Multimedia Appendix 4
Education action plan.
[PDF File, 4038 KB - jmir_v28i1e79559_app4.pdf]

Multimedia Appendix 5

XSL•FO
RenderX

Sample illustrated educational text.
[PNG File, 537 KB - jmir_v28i1e79559_app5.png]

Multimedia Appendix 6
Sample caregiver-facing informational images.
[PNG File, 294 KB - jmir_v28i1e79559_app6.png]

Multimedia Appendix 7
Sample video.
[MP4 File, 15585 KB - jmir_v28i1e79559_app7.mp4]

Multimedia Appendix 8
Animated stories.
[PDF File, 121 KB - jmir_v28i1e79559_app8.pdf]

Multimedia Appendix 9
SCORAD self-assessment.
[PDF File, 121 KB - jmir_v28i1e79559_app9.pdf]

Multimedia Appendix 10
Comparison of changes in disease severity and quality of life scores between 2 groups over time.
[DOCX File, 16 KB - jmir_v28i1e79559_app10.docx]

Checklist 1
CONSORT checklist.
[PDF File, 1121 KB - jmir_v28i1e79559_app11.pdf]

References

1. Langan SM, Irvine AD, Weidinger S. Atopic dermatitis. Lancet 2020 Aug 1;396(10247):345-360. [doi: 10.1016/S0140-6736(20)31286-1] [Medline: 32738956]
2. Nutten S. Atopic dermatitis: global epidemiology and risk factors. Ann Nutr Metab 2015;66 Suppl 1:8-16. [doi: 10.1159/000370220] [Medline: 25925336]
3. Guo Y, Zhang H, Liu Q, et al. Phenotypic analysis of atopic dermatitis in children aged 1-12 months: elaboration of novel diagnostic criteria for infants in China and estimation of prevalence. J Eur Acad Dermatol Venereol 2019 Aug;33(8):1569-1576. [doi: 10.1111/jdv.15618] [Medline: 30989708]
4. Guo Y, Li P, Tang J, et al. Prevalence of atopic dermatitis in Chinese children aged 1–7 ys. Sci Rep 2016;6(1):29751. [doi: 10.1038/srep29751]
5. Silverberg JI. Public health burden and epidemiology of atopic dermatitis. Dermatol Clin 2017 Jul;35(3):283-289. [doi: 10.1016/j.det.2017.02.002] [Medline: 28577797]
6. Nicolescu AC, Strilciuc S, Lăpădatu R, Grad DA, Vlădescu C, Olteanu R. The economic burden of atopic dermatitis in Romania: a broad perspective. Front Public Health 2024;12:1531042. [doi: 10.3389/fpubh.2024.1531042] [Medline: 39850862]
7. Girolomoni G, Busà VM. Flare management in atopic dermatitis: from definition to treatment. Ther Adv Chronic Dis 2022 Jan;13. [doi: 10.1177/20406223211066728]
8. Lapeere H, Speeckaert R, Baeck M, et al. Belgian atopic dermatitis guidelines. Acta Clin Belg 2024 Feb;79(1):62-74. [doi: 10.1080/17843286.2023.2285576] [Medline: 37997950]
9. Wollenberg A, Kinberger M, Arents B, et al. European guideline (EuroGuiDerm) on atopic eczema - part II: non-systemic treatments and treatment recommendations for special AE patient populations. J Eur Acad Dermatol Venereol 2022 Nov;36(11):1904-1926. [doi: 10.1111/jdv.18429] [Medline: 36056736]
10. Saeki H, Ohya Y, Furuta J, et al. English version of clinical practice guidelines for the management of atopic dermatitis 2021. J Dermatol 2022 Oct;49(10):e315-e375. [doi: 10.1111/1346-8138.16527] [Medline: 35996152]
11. Panel A, Chu DK, Schneider L, et al. Atopic dermatitis (eczema) guidelines: 2023 American Academy of Allergy, Asthma and Immunology/American College of Allergy, Asthma and Immunology Joint Task Force on Practice Parameters GRADE- and Institute of Medicine-based recommendations. Ann Allergy Asthma Immunol Mar 2024;132(3):274-312. [doi: 10.1016/j.anai.2023.11.009]
12. Xu Y, Zhi-Qiang S, Wei L, et al. Guidelines for diagnosis and treatment of atopic dermatitis in China. Int J Derm Venereol 2021;4(1):1-9. [doi: 10.1097/JD9.0000000000000143]

XSL•FO
RenderX

13. Greenwell K, Sivyer K, Howells L, et al. “Eczema shouldn’t control you; you should control eczema”: qualitative process evaluation of online behavioural interventions to support young people and parents/carers of children with eczema. Br J Dermatol 2023 Mar 30;188(4):506-513. [doi: 10.1093/bjd/ljac115] [Medline: 36745562]
14. Andrade LF, Abdi P, Mashoudy KD, et al. Effectiveness of atopic dermatitis patient education programs - a systematic review and meta-analysis. Arch Dermatol Res 2024 Apr 25;316(5):135. [doi: 10.1007/s00403-024-02871-y] [Medline: 38662127]
15. Gabryszewski SJ, Hill DA. One march, many paths: insights into allergic march trajectories. Ann Allergy Asthma Immunol 2021 Sep;127(3):293-300. [doi: 10.1016/j.anai.2021.04.036] [Medline: 33971364]
16. Moore EJ, Williams A, Manias E, Varigos G, Donath S. Eczema workshops reduce severity of childhood atopic eczema. Australas J Dermatol 2009 May;50(2):100-106. [doi: 10.1111/j.1440-0960.2009.00515.x] [Medline: 19397561]
17. Heratizadeh A, Werfel T, Wollenberg A, et al. Effects of structured patient education in adults with atopic dermatitis: multicenter randomized controlled trial. J Allergy Clin Immunol 2017 Sep;140(3):845-853. [doi: 10.1016/j.jaci.2017.01.029] [Medline: 28242304]
18. Kupfer J, Gieler U, Diepgen TL, et al. Structured education program improves the coping with atopic dermatitis in children and their parents-a multicenter, randomized controlled trial. J Psychosom Res 2010 Apr;68(4):353-358. [doi: 10.1016/j.jpsychores.2009.04.014] [Medline: 20307702]
19. Liang Y, Tian J, Shen CP, et al. Therapeutic patient education in children with moderate to severe atopic dermatitis: a multicenter randomized controlled trial in China. Pediatr Dermatol 2018 Jan;35(1):70-75. [doi: 10.1111/pde.13362] [Medline: 29243849]
20. Grossman SK, Schut C, Kupfer J, Valdes-Rodriguez R, Gieler U, Yosipovitch G. Experiences with the first eczema school in the United States. Clin Dermatol 2018;36(5):662-667. [doi: 10.1016/j.clindermatol.2018.05.006] [Medline: 30217280]
21. Staab D, Diepgen TL, Fartasch M, et al. Age related, structured educational programmes for the management of atopic dermatitis in children and adolescents: multicentre, randomised controlled trial. BMJ 2006 Apr 22;332(7547):933-938. [doi: 10.1136/bmj.332.7547.933] [Medline: 16627509]
22. Broberg A, Kalimo K, Lindblad B, Swanbeck G. Parental education in the treatment of childhood atopic eczema. Acta Derm Venereol 1990;70(6):495-499. [Medline: 1981422]
23. Futamura M, Ito K, Otsuji K, et al. Effects of “Skin Care School,” a parental education program on childhood atopic dermatitis conducted during short hospitalization stays. Arerugi 2009 Dec;58(12):1610-1618. [Medline: 20220302]
24. LeBovidge JS, Timmons K, Delano S, et al. Improving patient education for atopic dermatitis: a randomized controlled trial of a caregiver handbook. Pediatr Dermatol 2021 Mar;38(2):396-404. [doi: 10.1111/pde.14519] [Medline: 33486817]
25. Park GY, Park HS, Cho S, Yoon HS. The effectiveness of tailored education on the usage of moisturizers in atopic dermatitis: a pilot study. Ann Dermatol 2017;29(3):360. [doi: 10.5021/ad.2017.29.3.360]
26. Armstrong AW, Kim RH, Idriss NZ, Larsen LN, Lio PA. Online video improves clinical outcomes in adults with atopic dermatitis: a randomized controlled trial. J Am Acad Dermatol 2011 Mar;64(3):502-507. [doi: 10.1016/j.jaad.2010.01.051] [Medline: 21236514]
27. Tier HL, Balogh EA, Bashyam AM, et al. Tolerability of and adherence to topical treatments in atopic dermatitis: a narrative review. Dermatol Ther (Heidelb) 2021 Apr;11(2):415-431. [doi: 10.1007/s13555-021-00500-4] [Medline: 33599887]
28. Patel N, Feldman SR. Adherence in atopic dermatitis. Adv Exp Med Biol 2017;1027:139-159. [doi: 10.1007/978-3-319-64804-0_12] [Medline: 29063437]
29. Wilken B, Zaman M, Asai Y. Patient education in atopic dermatitis: a scoping review. Allergy Asthma Clin Immunol 2023 Oct 13;19(1):89. [doi: 10.1186/s13223-023-00844-w] [Medline: 37833754]
30. Golinelli D, Boetto E, Carullo G, Nuzzolese AG, Landini MP, Fantini MP. Adoption of digital technologies in health care during the COVID-19 pandemic: systematic review of early scientific literature. J Med Internet Res 2020 Nov 6;22(11):e22280. [doi: 10.2196/22280] [Medline: 33079693]
31. Yuan L, Lingling L, Shan W, et al. Efficacy and safety of 0.03% tacrolimus ointment in the long-term intermittent maintenance treatment of atopic dermatitis in children: a multicenter randomized controlled clinical trial. Chin J Dermatol 2019;52(8):519-524. [doi: 10.3760/cma.j.issn.0412-4030.2019.08.001] [Medline: 30656725]
32. Overall JE, Shobaki G, Shivakumar C, Steele J. Adjusting sample size for anticipated dropouts in clinical trials. Psychopharmacol Bull 1998;34(1):25-33. [Medline: 9564195]
33. Bass AM, Anderson KL, Feldman SR. Interventions to increase treatment adherence in pediatric atopic dermatitis: a systematic review. J Clin Med 2015 Jan 27;4(2):231-242. [doi: 10.3390/jcm4020231] [Medline: 26239125]
34. Barbarot S, Aubert H, Vibet MA, et al. Effectiveness of a nurse-led one-to-one education programme in addition to standard care in children with atopic dermatitis: a multicentre randomized control trial. Br J Dermatol 2024 Jul 16;191(2):177-186. [doi: 10.1093/bjd/ljae111] [Medline: 38863109]
35. Cheong JYV, Hie SL, Koh EW, de Souza NNA, Koh MJA. Impact of pharmacists’ counseling on caregiver’s knowledge in the management of pediatric atopic dermatitis. Pediatr Dermatol 2019 Jan;36(1):105-109. [doi: 10.1111/pde.13708] [Medline: 30408232]

XSL•FO
RenderX

36. Chinn DJ, Poyner T, Sibley G. Randomized controlled trial of a single dermatology nurse consultation in primary care on the quality of life of children with atopic eczema. Br J Dermatol 2002 Mar;146(3):432-439. [doi: 10.1046/j.1365-2133.2002.04603.x] [Medline: 11952543]
37. Rolinck-Werninghaus C, Trentmann M, Reich A, Lehmann C, Staab D. Improved management of childhood atopic dermatitis after individually tailored nurse consultations: a pilot study. Pediatr Allergy Immunol 2015 Dec;26(8):805-810. [doi: 10.1111/pai.12338] [Medline: 25643831]
38. Boswell JF, Kraus DR, Miller SD, Lambert MJ. Implementing routine outcome monitoring in clinical practice: benefits, challenges, and solutions. Psychother Res 2015;25(1):6-19. [doi: 10.1080/10503307.2013.817696] [Medline: 23885809]
39. Cross SP, Alvarez-Jimenez M. The digital cumulative complexity model: a framework for improving engagement in digital mental health interventions. Front Psychiatry 2024;15:1382726. [doi: 10.3389/fpsyt.2024.1382726] [Medline: 39290300]

Abbreviations

AD: atopic dermatitis
CDLQI: Children's Dermatology Life Quality Index
CONSORT: Consolidated Standards of Reporting Trials
CRO: Clinical Research Organization
DFI: Dermatitis Family Impact
EAP: education action plan
HR: hazard ratio
IDQOL: Infant's Dermatitis Quality of Life Index
IGA: Investigator's Global Assessment
ITT: intention-to-treat
MICE: multiple imputation by chained equation
POEM: Patient-Oriented Eczema Measure
PP-NRS: Peak Pruritus Numerical Rating Scale
QoL: quality of life
RCT: randomized controlled trial
RR: relative risk
SCORAD: Scoring Atopic Dermatitis

Edited by N Cahill; submitted 24.Jun.2025; peer-reviewed by X Liang, Y Zhang; revised version received 14.Nov.2025; accepted 18.Nov.2025; published 07.Jan.2026.

Please cite as:
Yang H, Shu H, Wang LH, Li P, Li YL, Li QF, Han XP, Tian J, Chang J, Qian H, Chen JP, Ding XQ, Wu PQ, Dou LM, Luo Z, Li W, Lin YY, Li L, Yue SZ, Gu Y, Yang L, Sun XH, Luo XY, Ma L, Wang H
Smartphone-Based Digital Eczema Education Program for Atopic Dermatitis in Children Aged 0 to 6 Years: Multicenter, Randomized, Parallel Controlled Clinical Study
J Med Internet Res 2026;28:e79559
URL: https://www.jmir.org/2026/1/e79559
doi:10.2196/79559

© Huan Yang, Hong Shu, Liu-hui Wang, Ping Li, Yun-ling Li, Qin-feng Li, Xiu-ping Han, Jing Tian, Jing Chang, Hua Qian, Jing-ping Chen, Xin-qiang Ding, Pan-qian Wu, Li-min Dou, Zhen Luo, Wei Li, Yang-yang Lin, Lin Li, Shu-zhen Yue, Yang Gu, Li Yang, Xiao-hong Sun, Xiao-yan Luo, Lin Ma, Hua Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Bayesian-Based Pharmacokinetic Framework Integrated with Therapeutic Drug Monitoring for Assessing Adherence to Antiseizure Medications: A Clinical Trial Simulation Study

Xiao-Qin Liu[1], PhD; Zi-Ran Li[2], PhD; Wei-Wei Lin[3], PhD; Juan Wang[1], BSc; Fu-Qing Gu[1], BSc; Jun-Jie Ding[4], PhD; Zheng Jiao[1], PhD

[1]Department of Pharmacy, Shanghai Chest Hospital, Shanghai Jiao Tong University School of Medicine, 241 West Huaihai Road, Shanghai, China

[2]Department of Bioengineering & Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, United States

[3]Department of Pharmacy, The First Affiliated Hospital of Fujian Medical University, Fuzhou, Fujian, China

[4]Center for Tropical Disease and Global Health, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, United Kingdom

Corresponding Author:
Zheng Jiao, PhD
Department of Pharmacy, Shanghai Chest Hospital, Shanghai Jiao Tong University School of Medicine, 241 West Huaihai Road, Shanghai, China

Abstract

Background: Adherence to antiseizure medications (ASMs) is a cornerstone of effective epilepsy management. However, current consensus guidelines for assessing medication adherence via therapeutic drug monitoring (TDM) may neglect individual patient characteristics, thereby compromising the accuracy of adherence assessments.

Objective: This study proposed an innovative Bayesian–based pharmacokinetic (PK) framework integrated with TDM data to address the above limitations, with a focus on 14 widely prescribed ASMs, including brivaracetam, carbamazepine, clobazam, eslicarbazepine acetate, lacosamide, lamotrigine, levetiracetam, oxcarbazepine, perampanel, phenobarbital, topiramate, valproic acid, vigabatrin, and zonisamide.

Methods: Comprehensive clinical trial simulations were conducted to investigate the PK of ASMs in patients with epilepsy under conditions of full adherence and various nonadherent dosing behaviors, including omission of the last dose and consecutive missed doses. Bayesian posterior probabilities of these dosing behaviors were derived by integrating validated population PK models, individual patient demographics (eg, age, weight, creatinine clearance), dosing history, prior adherence probabilities and TDM measurements. Additionally, the influence of covariates on assessment outcomes was systematically evaluated.

Results: The Bayesian-based PK approach demonstrated robust discriminative ability. Under idealized simulation conditions with minimized variabilities, the approach achieved accurate retrodiction of the last 1 or 2 doses across all 14 ASMs and partial retrodiction of extended nonadherence trajectories for 6 ASMs. Concentration thresholds for adherence classification varied significantly across drugs and are influenced by patient-specific factors, comedications, formulation, sampling time, and prior probability. To translate these insights into practice, an adaptable web-based dashboard was developed using the *shiny* package in R software to enable precise and real-time assessments of medication adherence.

Conclusions: This study establishes a Bayesian-based PK approach to enhance the assessment of ASMs adherence. This approach facilitates a paradigm shift from population-based management to patient-specific adherence profiling, offering a practical methodology for the precise evaluation of medication-taking behaviors.

(J Med Internet Res 2026;28:e77917) doi:10.2196/77917

KEYWORDS

antiseizure medications; medication adherence; therapeutic drug monitoring; Bayesian theory; population pharmacokinetics

Introduction

Epilepsy is the second most common neurological disease globally. Antiseizure medications (ASMs) represent the cornerstone of treatment for epilepsy [1,2], with long-term medication adherence being critical to achieving successful therapeutic outcomes [3]. However, adherence to ASMs among people with epilepsy is often suboptimal [4-6], which is strongly associated with a range of adverse clinical outcomes, including increased mortality, heightened morbidity, greater health care utilization, and substantial economic burden [7,8]. Therefore, when evaluating treatment failures, it is imperative for health care providers to comprehensively assess patients' adherence,

XSL•FO
RenderX

to identify underlying issues and provide tailored support to improve seizure control and treatment efficacy.

In clinical practice, self-reported adherence is inherently subjective and prone to bias [9], in contrast to therapeutic drug monitoring (TDM), which offers a more objective measure of recent medication-taking behaviors [10,11]. Accurate TDM interpretation is straightforward in some situations, such as a notably low drug concentration suggesting nonadherence. However, it becomes much more complex in other cases due to various intrinsic and extrinsic confounders that affect drug concentrations, including organ function, drug-drug interactions and dosing intervals.

The Consensus Guidelines for TDM in Neuropsycho-pharmacology: 2017 update [12] (hereafter referred to as the 2017 Guidelines) provide reference ranges for commonly used ASMs, which are specifically tailored for steady-state trough concentrations (C_0) in adult patients undergoing monotherapy. The reference ranges for each ASM were determined by multiplying the daily dose by dose-related concentration factors, and then could be used to help identify nonadherence [12]. However, the reference ranges were based on average pharmacokinetic (PK) parameters from an adult population, and do not account for key subpopulations, such as pediatric, geriatric and pregnant patients, who exhibit clinically significant PK differences [13-15].

PK modeling and simulation approaches have been successfully used to evaluate the impact of medication nonadherence [16,17], and to design remedial dosing strategies for missed or delayed doses [16,18]. When combined with Bayesian principle, this methodology offers a powerful framework for integrating individual TDM data with population PK models [19,20]. By leveraging this approach, it becomes possible to infer posterior probabilities of different dosing patterns, thereby enabling a more refined and quantitative assessment of medication-taking behavior.

In light of the above, this study aims to characterize medication adherence patterns to ASMs using TDM measurements and a Bayesian-based PK approach. Additionally, a user-friendly dashboard is developed to offer health care providers an intuitive, practical tool for assessing individual adherence levels, thereby optimizing ASM therapy and improving the treatment outcomes of ASMs.

Method

Ethical Considerations

As this study exclusively used computational modeling and simulation techniques without involving direct human subject participation or personal data collection, it is exempt from institutional review board approval requirements in accordance with international ethical guidelines.

Rationale

When patients fully adhere to their medication regimens, the drug concentration fluctuates in a predictable manner. However, if patients miss any of their doses, the drug concentration will gradually decline to a suboptimal level, which may ultimately result in treatment failure. The differences in the probability distribution of drug concentration provides a valuable reference for differentiating between adherence to the prescribed medication and nonadherence.

In this study, the Bayesian-based PK approach, calculating the posterior probability of special dosing events, was used to assess medication adherence. The principle of the Bayesian approach is as follows [19]: given the probability of the occurrence of a specific scenario (ie, the prior probability, P(ωj)) and the probability of a particular drug concentration at a given scenario ωj (ie, conditional probability, P(C|ωj)) , the probability of the scenario at a given drug concentration (ie, posterior probability,P(ωjC) can be estimated, as presented in Equation 1.

$$(1)P(\omega j \mid C)=P(\omega j)\times P(C \mid \omega j)P(C)$$

Where P(C) is the full probability and could be calculated with Equation 2.

$$(2)P(C)=\sum jP(\omega j)\times P(C \mid \omega j)$$

The prior probability Pωj refers to the pre-existing or baseline probability estimate of a patient's likelihood to adhere to a prescribed medication regimen before any new, specific data related to that individual patient's adherence behavior in the current treatment course is considered. The conditional probability P(C|ωj) is calculated using Monte Carlo simulations based on population PK. Once these probabilities have been obtained, the posterior probability P(ωjC of each individual scenario is calculated using Equations 1; 2. The scenario with the highest posterior probability is considered the most likely to occur, while the one with the lowest posterior probability is deemed the least probable.

In our study, the dosing event scenario is defined by whether patients adhere to or miss their scheduled doses. There are 2^n possible scenarios when considering the last *n* dosing events prior to sampling. For instance, as depicted in Figure 1A, there are two scenarios (ω0 and ω1) when considering the last dosing instance. This can be expanded to four scenarios (ω00, ω01, ω10, and ω11) when the last two dosing instances are considered (Figure 1B), and eight scenarios (ω000, ω010, ω100, ω001, ω110, ω011, ω101, ω111) when the last three dosing events are taken into account (Figure 1C).

XSL•FO
RenderX

Figure 1. The dosing scenarios when the most recent one (A), two (B) or three (C) dosing events are considered. ω: medication-taking behavior, where the first, second and third digit after ω indicates the most recent one, two and three medication-taking events prior to sampling, respectively, where 1 indicates dose taken and 0 indicates dose missed.

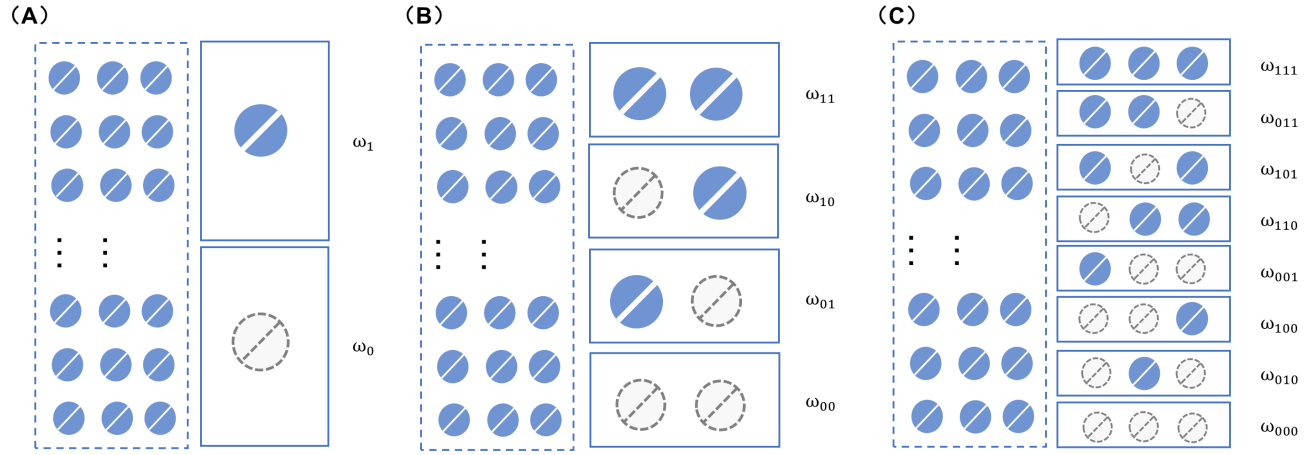

The workflow of adherence assessment is graphically represented with Figure 2, using the example of a 70 kg adult patient receiving oxcarbazepine 300 mg every 12 hours (q12h) and reached steady state. C_0 of oxcarbazepine was used to infer the patient's dosing behavior over the last two dosing intervals. When the C_0 approaches zero, the posterior probability of at least one missed dose is high. As C_0 increases, this probability decreases, while the posterior probability of complete adherence rises correspondingly, eventually approaching 100%.

XSL•FO
RenderX

Figure 2. The workflow of adherence assessment by therapeutic drug monitoring with Bayesian-based pharmacokinetic approach.

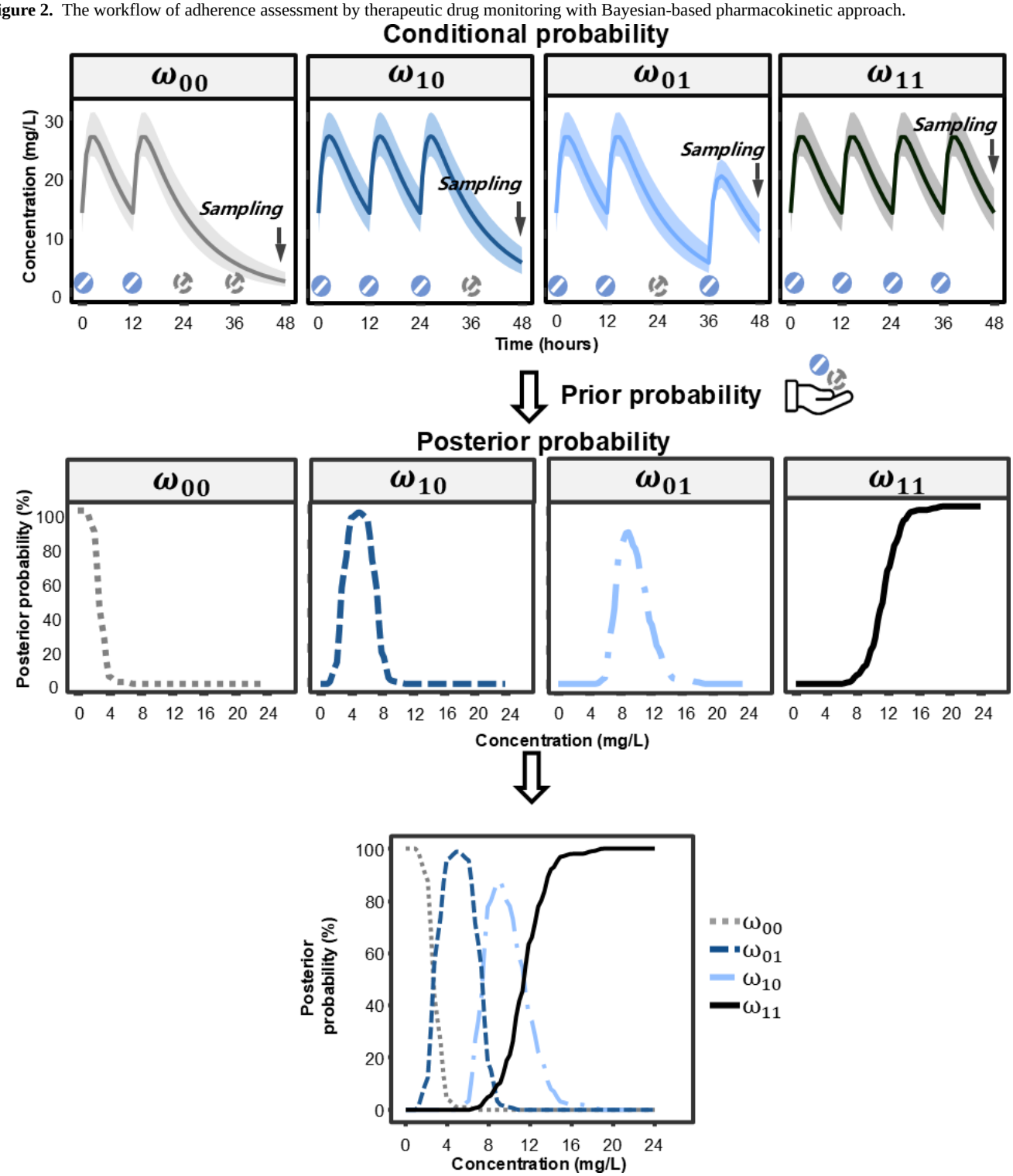

The C_0 value at which the posterior probabilities of two distinct dosing scenarios are equal, can be derived from Equations 1; 2, serving as a threshold to discriminate between these scenarios. As depicted in Figure 2, the posterior probabilities of missing two doses (ω00) and missing only the last dose while having taken the second last dose (ω01) are equal when the C_0 is approximately 3 mg/L. When C_0 is less than 3 mg/L, the probability of ω00 is the highest. Similarly, the C_0 range maps to the most probable scenario as follows: 3 - 7.5 mg/L to ω01, 7.5 - 11.5 mg/L to missing the second-last dose but taking the last dose (ω10) , and levels above 11.5 mg/L to taking both of the last two doses (ω11).

In this case, all dosing events could have a maximum posterior probability exceeding 80%, which was defined as complete retrodiction (Figure 3A). If only one or none dosing events had a maximum posterior probability exceeding 80%, it was defined as no retrodiction (Figure 3C). Other cases were defined as partial retrodiction (Figure 3B).

XSL•FO
RenderX

Figure 3. Illustration of retrodiction types based on posterior probabilities of dosing events prior to sampling. (A) complete retrodiction: the maximum posterior probabilities of all dosing events are ≥80%; (B) partial retrodiction: only events which are fully adherent and fully nonadherent have a maximum posterior probability ≥80%; and (C) no retrodiction: only one or no dosing events have a maximum posterior probability ≥80%.

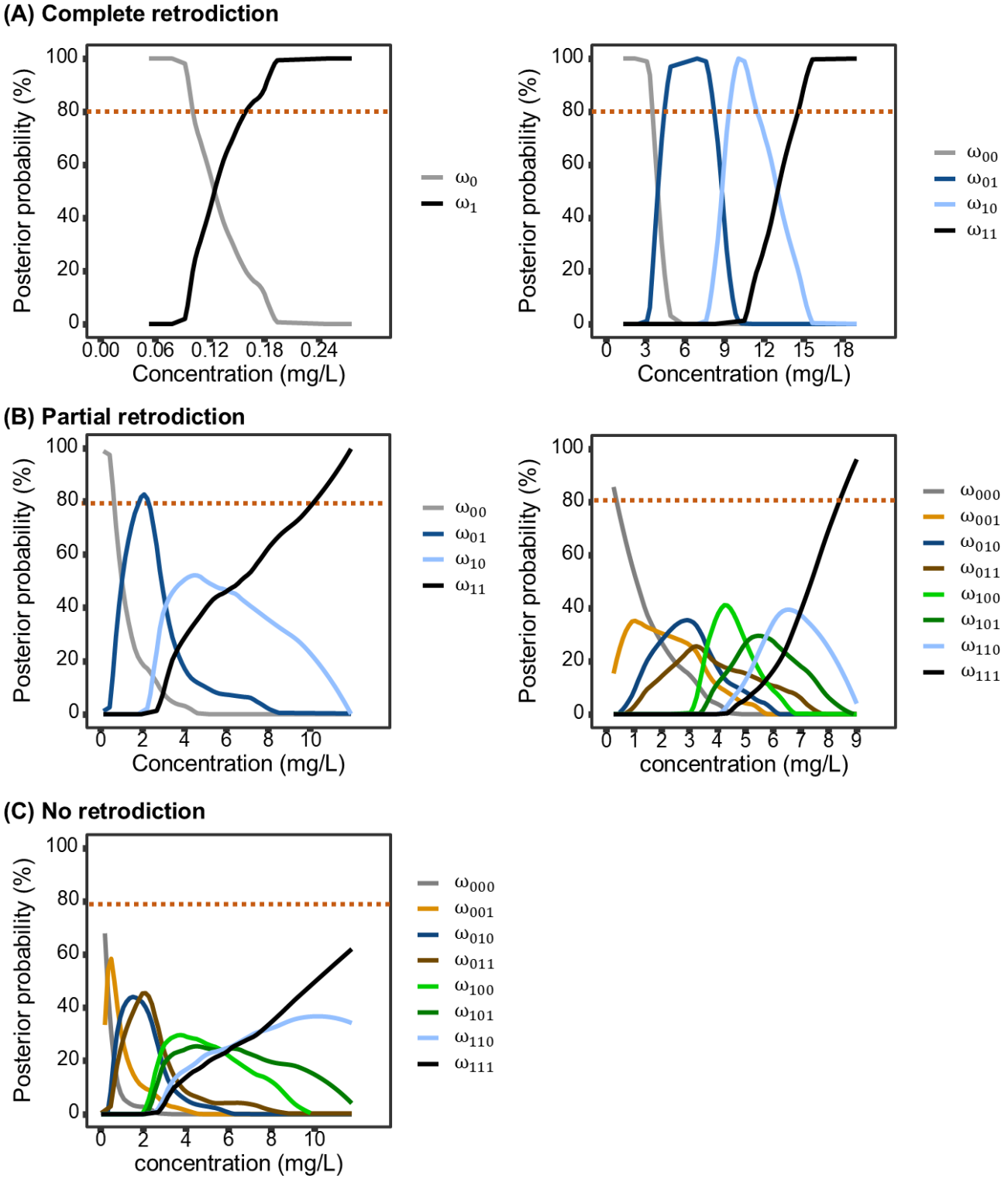

Population Pharmacokinetic Characteristics

To characterize the conditional probability associated with each dosing scenario, a systematic literature search was conducted in PubMed and Embase to collate available population PK parameters for ASMs across various formulations, including conventional tablets, oral solutions, suspensions, syrups, and extended-release (ER) formulations. In our previous study [18], population PK models for 10 commonly used ASMs were identified up to March 31, 2022, including carbamazepine, clobazam, eslicarbazepine acetate, lamotrigine, levetiracetam, oxcarbazepine, phenobarbital, topiramate, valproic acid, and zonisamide. An update search was then performed up to November 30, 2024. Additionally, the population PK characteristics of brivaracetam, lacosamide, perampanel and vigabatrin were also incorporated in the study. Details of the literature review were summarized in Multimedia Appendix 1.

XSL•FO
RenderX

Assessment of Adherence

During the adherence assessment, pediatric patients (aged 8 y, weighing 25 kg, and measuring 127 cm), adult patients (aged 40 y, weighing 70 kg, and measuring 180 cm), and pregnant women (aged 25 y, weighing 70 kg, measuring 160 cm, 30 wk pregnant) taking conventional tablet of ASMs were selected as typical patients. All typical patients had normal renal and liver function and were not taking any concomitant medications.

The evaluated dosing behaviors included the administration of the last one, two, and three doses prior to sampling at steady state. Sampling was conducted immediately before the subsequent dose. As for prior probability, it was assumed that each scenario had an equal chance of occurring. Specifically, this implies a probability of 50% for each scenario when considering only the last dosing behavior, 25% when considering the last two dosing behaviors, and 12.5% when considering the last three dosing behaviors.

For the Monte Carlo simulations, the parameters were fixed according to the final reported values, except for the residual unexplained variability (RUV) to obtain the "true" concentration-time profiles under various nonadherence scenarios. Consequently, RUV was set to negligible levels [21]—specifically 0.01 mg/L for additive error and 0.1% for proportional error—to minimize noise in the study. A total of 40,000 virtual patients were generated for each scenario. The Monte Carlo simulations were conducted using R programming (version 4.2.2; R Foundation for Statistical Computing) with the *rxode2* package (version 2.1.2). The results were plotted using the *ggplot2* package (version 3.5.1).

Evaluation of Critical Factors Affecting Adherence Assessment

The factors reported to significantly influence the PK of ASMs were investigated for their impact on adherence assessment, including renal function (estimated glomerular filtration rate, eGFR: 30, 60 and 90 mL/min/1.73m^2) and concomitant medications. Additionally, the effect of formulation (extending dosing interval to every 24 h for ER formulation), sampling time (2 h earlier or later), and prior probabilities (10%, 30%, 50%, 70%, and 90%) on medication adherence was also tested. The impact of these factors was assessed from two perspectives: the first was the ability to retrodict the number of the last scheduled doses, and the second was their influence on the concentration threshold used to distinguish between nonadherence patterns.

Development of Web-Based Dashboard

To facilitate quick calculation, an interactive online dashboard was developed to assess ASMs' medication adherence, informed by TDM results, and individual characteristics that were determined as significant factors on PK parameters in the included models. This tool was built using *rxode2* (version 2.1.2), *ggplot2* (version 3.5.1), and *shiny* (version 1.8.1.1) within the R framework (version 4.2.2; R Foundation for Statistical Computing).

Results

Population Pharmacokinetic Characteristics

A total of 23 population PK models encompassing 14 ASMs were ultimately included in the analysis [22-44]. Among these, models for adult [22-25,27-29,31,33,35,36,38,40,42] and pediatric patients [22,23,25-30,32,34,36,37,39,44] were available, while models specific to pregnant women were only available for lamotrigine [41] and levetiracetam [43]. Since age was consistently identified as a significant covariate for PK parameters in adults, elderly patients were thus grouped with the adult population. Models characterizing multiple formulations were identified for eslicarbazepine acetate [34], lamotrigine [36] and valproic acid [30,42]. Details of the identification of literature, included studies, and the final parameter estimates used in the analysis were comprehensively summarized in Figures S1-S9 in Multimedia Appendix 1 and Tables S1-S2 in Multimedia Appendix 1.

Assessment of Adherence

The posterior probabilities of various dosing behaviors when considering the last one, two, and three dosing behaviors for each ASM are detailed in Figure S10-S23 in Multimedia Appendix 1. Figure 4 demonstrates the ability to retrodict the number of the last scheduled doses for each ASM in typical patients under conditions of minimized RUV. Results indicated that when investigating the most recent dosing behavior, all ASMs can be fully retrodicted. Regarding the scenarios involving the last two dosing behaviors, complete retrodiction was achievable only for oxcarbazepine in pediatric patients, whereas other ASMs were partially retrodicted. When extending to the last three dosing behaviors, no ASMs can be fully retrodicted, and only carbamazepine, clobazam, eslicarbazepine acetate, oxcarbazepine, phenobarbital and zonisamide could be partially retrodicted across all investigated population. From a pharmacokinetic perspective, ASMs with higher clearance are eliminated rapidly, thereby diminishing the concentration "signal" necessary to distinguish earlier dosing events. Furthermore, the traceability may vary in clinical scenarios where patient characteristics significantly deviate from the typical population.

XSL•FO
RenderX

Figure 4. The ability to retrodict the last one, two and three dosing behaviors prior to sampling for typical patients in theoretical condition when minimizing residual unexplained variabilities. D1: the last dosing behavior; D2: the last two dosing behaviors; D3: the last three dosing behaviors; CR, complete retrodiction, which is defined as when the maximum posterior probabilities of all dosing events are ≥80%; PR, partial retrodiction, which is defined as when only events which are fully adherent and fully nonadherent (have a maximum posterior probability ≥80%; NR: no retrodiction, which is defined as when only one or no dosing events have a maximum posterior probability ≥80%. Adults: aged 40 y, weighing 70 kg, and measuring 180 cm; children: aged 8 y, weighing 25 kg, and measuring 127 cm; pregnant women: aged 25 y, weighing 70 kg, measuring 160 cm, and being 30 weeks pregnant. Residual unexplained variabilities were minimized by defining additive error as 0.01 mg/L and proportional error as 0.1%.

Antiseizure medication	Population	D1	D2	D3
Brivaracetam	Adults	CR	PR	NR
	Children	CR	PR	NR
Carbamazepine	Adults	CR	PR	PR
	Children	CR	PR	PR
Clobazam	Adults	CR	PR	PR
	Children	CR	PR	PR
Eslicarbazepine acetate	Adults	CR	PR	PR
	Children (tablet)	CR	PR	PR
	Children (oral suspension)	CR	PR	PR
Lacosamide	Adults	CR	PR	NR
	Children	CR	PR	NR
Lamotrigine	Adults (immediate-release/ extended-release formulation)	CR	PR	PR
	Children (immediate-release formulation)	CR	PR	NR
	Children (extended-release formulation)	CR	PR	PR
	Pregnant women	CR	PR	NR
Levetiracetam	Adults	CR	PR	NR
	Children	CR	PR	NR
	Pregnant women	CR	PR	NR
Oxcarbazepine	Adults	CR	PR	PR
	Children	CR	CR	PR
Perampanel	Adults	CR	PR	NR
	Children	CR	PR	PR
Phenobarbital	Adults	CR	PR	PR
	Children	CR	PR	PR
Topiramate	Adults	CR	PR	NR
	Children	CR	PR	NR
Valproic acid	Adults (tablet/oral solution/modified-release coated tablet)	CR	PR	NR
	Children (tablet/syrup/sustained-release)	CR	PR	NR
Vigabatrin	Adults	CR	PR	NR
	Children	CR	PR	NR
Zonisamide	Adults	CR	PR	PR
	Children	CR	PR	PR

XSL•FO
RenderX

Impact of Critical Factors on Adherence Assessment

It has been reported that renal function affects the apparent clearance (CL/F) of eslicarbazepine acetate [33], levetiracetam [24], oxcarbazepine [38] and vigabatrin [28]. As renal function decreases, there is no significant effect on the identification of nonadherence patterns. The effect of renal function on levetiracetam is illustrated in Figure S24 in Multimedia Appendix 1, with levetiracetam, oxcarbazepine, and vigabatrin demonstrating similar trends.

Pregnancy enhances the clearance of lamotrigine and levetiracetam, leading to lower systemic drug exposure and, consequently, may decrease the concentration thresholds (Figure S15-S16 in Multimedia Appendix 1). Similarly, pediatric patients show higher clearance per body weight compared with adults, leading to the decreased concentration thresholds (Figure S10-S23 in Multimedia Appendix 1).

The effects of concomitant inducers and inhibitors were also evaluated. Administration of inducers or inhibitors did not alter the fundamental distinguishability of nonadherence patterns but shifted the concentration threshold required for their discrimination. Specifically, the threshold was lowered by inducers and raised by inhibitors. The magnitude of these adjustments varied substantially across ASMs (Figure S25 in Multimedia Appendix 1). For instance, in a typical adult patient taking lamotrigine, coadministration with enzyme inducers (eg, carbamazepine, phenobarbital) lowered the threshold by approximately 67%, whereas the inhibitors valproic acid elevated it by approximately 83%. The effect on topiramate was less pronounced.

The impact of formulation on adherence assessment was evaluated for eslicarbazepine acetate, lamotrigine and valproic acid. At equivalent total daily doses, ER formulations with prolonged dosing intervals enhanced the discriminative capacity for dosing behaviors compared to immediate-release (IR) or other oral formulations requiring more frequent administration (Figure S15, S21 in Multimedia Appendix 1). In contrast, formulations such as oral suspensions and syrups exhibited minimal impact on the assessment (Figure S13, S21 in Multimedia Appendix 1).

Sampling time also influences adherence assessment (Figure S26 in Multimedia Appendix 1). Sampling 2 hours earlier or later than the scheduled time does not significantly influence the distinguishability of nonadherence patterns. However, compared to sampling just before administration, the concentration threshold for distinguishing nonadherence patterns increases when sampling is done earlier and decreases when sampling is done later. The magnitude of this change varies among different ASMs.

The impacts of prior probabilities on adherence assessment were also evaluated. The results indicated that prior probabilities could not only significantly affect the distinguishability of the nonadherence dosing scenarios, but also notably alter the concentration threshold for distinguishability (Figure S27 in Multimedia Appendix 1). The magnitude of the concentration threshold change was found to be dependent on the type of ASMs.

Application of Web-Based Dashboard

A web-based dashboard for assessing medication adherence has been developed and is freely accessible online [45]. After inputting the type of ASMs, patient characteristics (age, body weight, height, gender), scheduled dosing regimens, sampling time, TDM data, and prior probabilities for each scenario, the system estimates the posterior probabilities of each dosing scenario and plots them against the drug concentration. RUV are initialized with literature-reported values (as listed in Table S2 in Multimedia Appendix 1) when requiring consideration, but remain user-adjustable to accommodate specific clinical situations, thereby enabling the precise identification of medication adherence patterns.

Figure 5A presents a case of a 75-year-old male (70 kg) with epilepsy and impaired renal function (eGFR 40 mL/min/1.73m^2) who had remained seizure-free for over 3 years on oxcarbazepine 300 mg q12h. Following a recent increase in seizure frequency, TDM was performed to assess potential nonadherence. The measured C_0 of oxcarbazepine was 12 mg/L, which lies within the conventional therapeutic range of 10 - 35 mg/L and aligns with the recommended range of 6 - 24 mg/L as per the 2017 guidelines [12]. However, model-based estimates from the dashboard indicated a nearly negligible probability of full adherence and a high probability of having missed at least one dose. The posterior probability of full adherence remained at 0%, regardless of whether the prior probability was set as low as 1% (suggesting poor adherence) or as high as 99% (denoting high adherence) (Figure S28 in Multimedia Appendix 1), demonstrating the minimal influence of the prior in this case. By contrast, increasing the RUV from 0.1% to 30% increased the posterior probability of full adherence from 0% to 56% (Figure S29 in Multimedia Appendix 1), underscoring the substantial impact of RUV on the adherence assessment in this case.

Figure 5B illustrates another case of a 10-year-old (30 kg) boy with epilepsy and normal renal/hepatic function, treated with valproic acid tablet 500 mg and carbamazepine 150 mg q12h. His measured C_0 of valproic acid was 40 mg/L, below both the conventional therapeutic range (50 - 100 mg/L) and the 2017 guideline-recommended range (62.2 - 134.8 mg/L). Although subtherapeutic concentrations initially raised suspicion of nonadherence, model-based analysis estimated a probability of full adherence exceeding 80%, suggesting that the low valproic acid concentration likely resulted from carbamazepine-induced metabolic induction rather than missed doses. Varying the prior probability of adherence from 1% to 99% had minimal impact on this conclusion (Figure S30 in Multimedia Appendix 1). Similarly, increasing RUV from 0.1% to 30% altered the posterior probability of full adherence by only 5% (Figure S31 in Multimedia Appendix 1), demonstrating the robustness of the assessment against RUV variation in this clinical scenario.

XSL•FO
RenderX

Figure 5. Screenshot of the dashboard for adherence assessment. (A) Elderly patient, 75 years old, weighing 70 kg, measuring 180 cm, eGFR 40 mL/min/1.73m^2, taking oxcarbazepine 300 mg q12h; (B) pediatric patient, 10 years old, weighing 30 kg, measuring 130 cm, taking valproic acid 500 mg q12h and carbamazepine 100 mg q12h. $_{00}$: missing two continuous doses before sampling; $_{01}$, missing the second-to-last dose but taking the last dose; $_{10}$, missing the last dose but taking the second-to-last dose; $_{11}$, taking all doses.

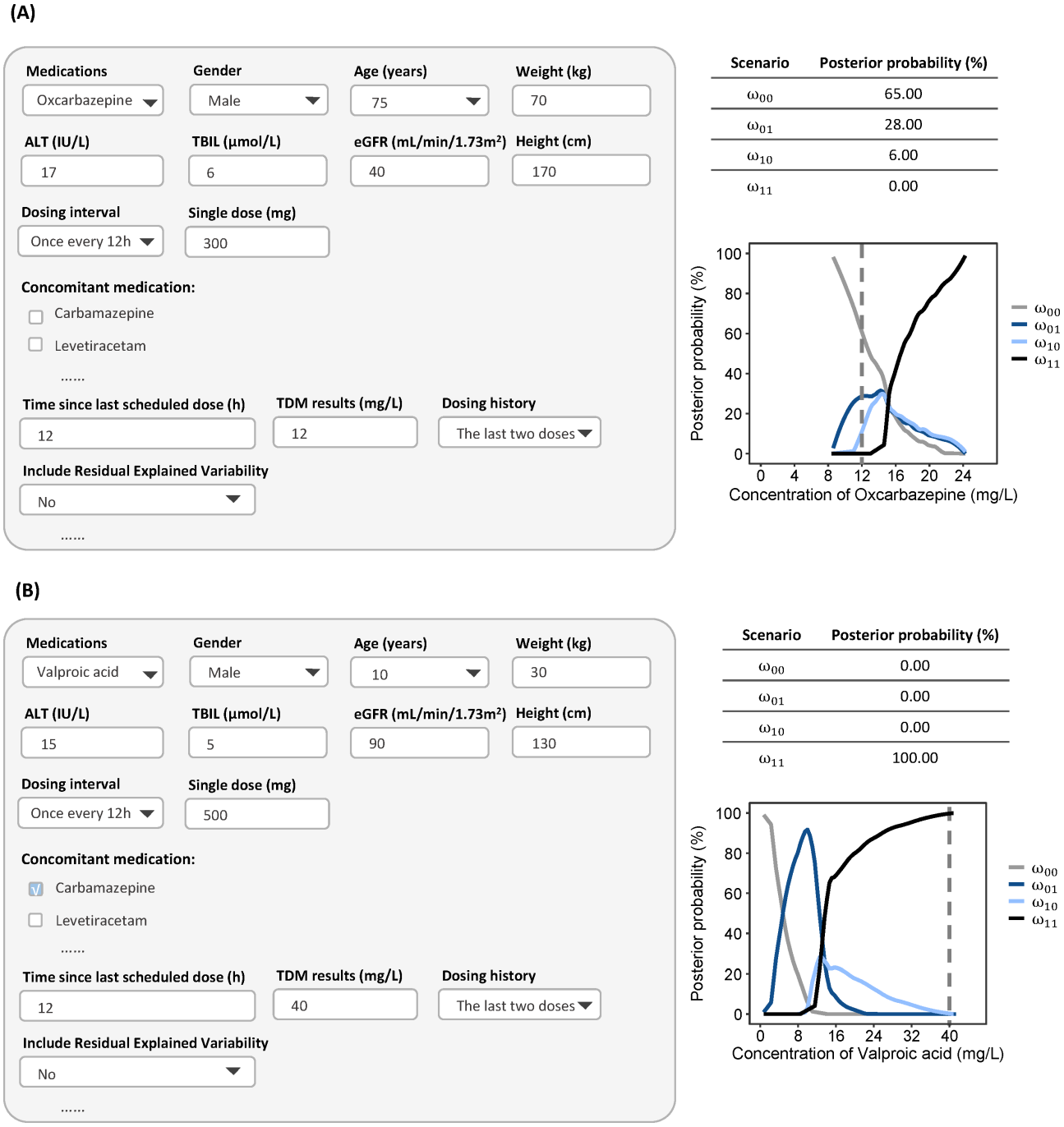

Discussion

Principal Findings

This study is the first to introduce a clinical framework to investigate the role of TDM in assessing medication adherence for 14 commonly used ASMs among diverse patients. By integrating Bayesian theory with population PK, we demonstrated that routine TDM, when combined with clinical factors, enables quantitative retrodiction of recent medication-taking behaviors for all investigated ASMs. The Bayesian-based PK framework can also be easily available through the open-access dashboard developed in this study.

Comparison to Prior Work

The 2017 Guideline established reference ranges encompassing approximately 66% of patients for commonly used ASMs, primarily derived from PK data obtained in adult patients receiving monotherapy [12]. While clinically useful, these population-derived thresholds have limited generalizability to special populations with distinct PK profiles. In contrast, the model-informed algorithm developed in this study enables a more personalized assessment of medication adherence. This

XSL•FO
RenderX

approach explicitly accounts for patient-specific factors including age (eg, pediatric and geriatric populations), pregnancy status, renal and hepatic function, concomitant use of enzyme inducers or inhibitors, formulation characteristics. It thereby provides a refined framework for evaluating medication-taking behavior across diverse clinical scenarios.

Interpretation of the Findings

We identified multiple critical factors that influence the adherence assessment, including intrinsic factors (physiological differences, concomitant medication, renal function, formulation , etc) and extrinsic factors (prior probability, RUV, etc).

The influence of intrinsic factors on adherence evaluation is primarily mediated through alterations in PK parameters, most notably systemic drug clearance. Enhanced drug clearance, commonly observed in pediatric and pregnant patients, as well as in those receiving enzyme inducers, reduces both the ability to differentiate adherence patterns and the corresponding concentration thresholds. In pediatric populations, the higher clearance per body weight results from ongoing organ maturation, larger organ size-to-body weight ratios, and increased metabolic enzyme activity [15,36,46]. In pregnant women, elevated clearance arises from increased cardiac output, enhanced renal blood flow, and hormonally mediated induction of metabolic pathways [14,41,47,48]. Conversely, reduced clearance, frequently encountered in patients with renal impairment or those receiving enzyme inhibitors, may improve differentiation ability or raise the concentration thresholds required for pattern discrimination.

Despite the increasing use of ER formulations of ASMs, conventional formulations continue to account for a substantial proportion of prescriptions due to their lower cost and wider availability [49,50]. Consequently, population PK studies have more frequently characterized conventional formulations. Based on the limited population PK data available for ER ASMs, our findings suggest that ER formulations may improve the differentiation of adherence behaviors, a finding attributable to the extended dosing interval (eg, from 12 to 24 h) and the resulting concentration-time fluctuation.

Prior probability is essential for estimating the posterior probability of dosing behaviors. In this study, we adopted an equiprobable prior probability to reflect a state of maximum uncertainty before considering the evidence (TDM measurements). This represents a conventional and conservative strategy in Bayesian modeling when reliable, specific prior knowledge is unavailable [19,51]. In real clinical settings, the prior probability can be informed by pharmacy refill data or population-average adherence estimates. When individual-level data are absent, population-based priors derived from patients with comparable covariates (eg, age, comorbidities, and socioeconomic status) may be applied. Although the impact of the prior was limited in the cases illustrated in Figure 5, its influence on adherence assessment can be substantial and depends on both the specific ASM and TDM measurements. Consequently, the ability for user-defined priors implemented in the dashboard remains highly valuable.

RUV in population PK analysis captures unexplained stochastic variations, including assay error, sampling inaccuracies, and model misspecification. As these elements may confound medication adherence assessments, RUV was intentionally minimized in the present analysis to reduce setting-specific noise and facilitate clearer characterization of covariate effects. To enhance real-world applicability, the accompanying dashboard allows users to adjust the RUV level based on reported values from source population PK studies (Table S2 in Multimedia Appendix 1), known assay variability, or clinical experience.

Limitations

The study has several limitations. First, there are numerous patterns of nonadherence and we only considered the scenario of missing doses. Other types of nonadherence, such as delayed doses, missed partial doses, and inadvertent overdoses, were not considered. Second, due to the lack of population-PK studies, specifically in pediatric patients, pregnant women, and for ER formulation, we did not include all these scenarios in our analysis. However, our dashboard can be readily extended to include these populations, or novel formulations once their population PK parameters become available. In addition, it is important to note that while we provide accurate estimation of probabilities for recent medication events, the clinical judgment should not be solely based on it. Comprehensive assessment must be performed to incorporate the patient's overall condition, medication history, and relevant information.

Future Directions

As epilepsy pharmacotherapy evolves, the dashboard will be updated to incorporate emerging population PK models for novel ASMs when available. Concurrently, it will extend to pharmacodynamic models to bridge the gap between PK and clinical outcomes, thereby quantifying the risks of breakthrough seizures from nonadherence trajectories. Leveraging prior work on remedial dosing regimens on ASMs [18], the tool can be expanded to provide remedial dosing strategies for clinicians to safely restore therapeutic concentrations after missed doses. Finally, the integration with multidimensional data, such as digital biomarkers and electronic health records, will be explored. The comprehensive strategy will ultimately facilitate the realization of precise, patient-specific life-cycle management in epilepsy treatment.

Conclusion

In conclusion, this study establishes a Bayesian-based PK approach to enhance the objective assessment of ASM adherence. By leveraging TDM data, the approach showed large improvement in assessing nonadherence patterns compared to the previous guidelines. In addition, to bridge the methodology with clinical practice, we developed an interactive dashboard that translates PK principles into visual and interpretable outputs. The work demonstrates the feasibility of transitioning from traditional population-based monitoring to individual-specific management for patients with epilepsy.

XSL•FO
RenderX

Funding

XQL was supported by the Elite Development Supporting Program from Shanghai Chest Hospital, Shanghai Jiao Tong University School of Medicine (RC-202407-076).

Data Availability

Authors can confirm that all relevant data are included in the article and its supplementary information files. The code used for data analysis is available from the corresponding author upon reasonable request

Authors' Contributions

XQL: Conceptualization, Software, Investigation, Formal analysis, Data Curation, Writing-Original Draft, Writing-Review & Editing, Funding acquisition; ZRL: Methodology, Formal analysis, Software, Writing-Review & Editing; WWL: Methodology, Software, Writing-Review & Editing; JW: Visualization, Writing-Review & Editing; FQG: Visualization, Writing-Review & Editing; JJD: Conceptualization, Validation, Writing-Review & Editing; ZJ: Conceptualization, Supervision, Project Administration, Writing-Original Draft, Writing-Review & Editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Literature identification, the posterior probabilities-concentration curves, sensitivity analysis results, and the summary of identified population pharmacokinetic models.
[DOCX File, 7346 KB - jmir_v28i1e77917_app1.docx]

References

1. Glauser T, Ben-Menachem E, Bourgeois B, et al. Updated ILAE evidence review of antiepileptic drug efficacy and effectiveness as initial monotherapy for epileptic seizures and syndromes. Epilepsia 2013 Mar;54(3):551-563. [doi: 10.1111/epi.12074] [Medline: 23350722]
2. Nevitt SJ, Sudell M, Cividini S, Marson AG, Tudur Smith C. Antiepileptic drug monotherapy for epilepsy: a network meta-analysis of individual participant data. Cochrane Database Syst Rev 2022 Apr 1;4(4):CD011412. [doi: 10.1002/14651858.CD011412.pub4] [Medline: 35363878]
3. Al-Aqeel S. Nonadherence to antiseizure medications: what have we learned and what can be done next? Expert Rev Pharmacoecon Outcomes Res 2024 Sep;24(7):791-798. [doi: 10.1080/14737167.2024.2349191] [Medline: 38676921]
4. Malek N, Heath CA, Greene J. A review of medication adherence in people with epilepsy. Acta Neurol Scand 2017 May;135(5):507-515. [doi: 10.1111/ane.12703] [Medline: 27781263]
5. Niriayo YL, Mamo A, Gidey K, Demoz GT. Medication belief and adherence among patients with epilepsy. Behav Neurol 2019 Apr 23;2019:2806341. [doi: 10.1155/2019/2806341] [Medline: 31178940]
6. Adisu MA, Zemariam AB, Derso YA, et al. Global patterns and predictors of anti-seizure medication adherence in pediatric epilepsy: A systematic review and meta-analysis. Epilepsy Behav 2025 Nov;172:110544. [doi: 10.1016/j.yebeh.2025.110544] [Medline: 40499262]
7. Faught E, Duh MS, Weiner JR, Guérin A, Cunnington MC. Nonadherence to antiepileptic drugs and increased mortality: findings from the RANSOM Study. Neurology 2008 Nov 11;71(20):1572-1578. [doi: 10.1212/01.wnl.0000319693.10338.b9] [Medline: 18565827]
8. Faught RE, Weiner JR, Guérin A, Cunnington MC, Duh MS. Impact of nonadherence to antiepileptic drugs on health care utilization and costs: findings from the RANSOM study. Epilepsia 2009 Mar;50(3):501-509. [doi: 10.1111/j.1528-1167.2008.01794.x] [Medline: 19183224]
9. Chan AHY, Wright DFB. Medication adherence-Everybody's problem but nobody's responsibility? Br J Clin Pharmacol 2025 Mar;91(3):681-683. [doi: 10.1111/bcp.16384] [Medline: 39734279]
10. Patsalos PN, Spencer EP, Berry DJ. Therapeutic drug monitoring of antiepileptic drugs in epilepsy: a 2018 update. Ther Drug Monit 2018 Oct;40(5):526-548. [doi: 10.1097/FTD.0000000000000546] [Medline: 29957667]
11. Stepanova D, Beran RG. The benefits of antiepileptic drug (AED) blood level monitoring to complement clinical management of people with epilepsy. Epilepsy Behav 2015 Jan;42:7-9. [doi: 10.1016/j.yebeh.2014.09.069] [Medline: 25499154]
12. Hiemke C, Bergemann N, Clement HW, et al. Consensus Guidelines for Therapeutic Drug Monitoring in Neuropsychopharmacology: Update 2017. Pharmacopsychiatry 2018 Jan;51(1-2):9-62. [doi: 10.1055/s-0043-116492] [Medline: 28910830]
13. Roberti R, Palleria C, Nesci V, et al. Pharmacokinetic considerations about antiseizure medications in the elderly. Expert Opin Drug Metab Toxicol 2020 Oct;16(10):983-995. [doi: 10.1080/17425255.2020.1806236] [Medline: 32757857]

XSL•FO
RenderX

14. Arfman IJ, Wammes-van der Heijden EA, Ter Horst PGJ, Lambrechts DA, Wegner I, Touw DJ. Therapeutic drug monitoring of antiepileptic drugs in women with epilepsy before, during, and after pregnancy. Clin Pharmacokinet 2020 Apr;59(4):427-445. [doi: 10.1007/s40262-019-00845-2] [Medline: 31912315]
15. Maglalang PD, Wen J, Hornik CP, Gonzalez D. Sources of pharmacokinetic and pharmacodynamic variability and clinical pharmacology studies of antiseizure medications in the pediatric population. Clin Transl Sci 2024 Apr;17(4):e13793. [doi: 10.1111/cts.13793] [Medline: 38618871]
16. Clark ED, Lawley SD. Should patients skip late doses of medication? A pharmacokinetic perspective. J Pharmacokinet Pharmacodyn 2022 Aug;49(4):429-444. [doi: 10.1007/s10928-022-09812-0] [Medline: 35726046]
17. Clark ED, Lawley SD. How drug onset rate and duration of action affect drug forgiveness. J Pharmacokinet Pharmacodyn 2024 Jun;51(3):213-226. [doi: 10.1007/s10928-023-09897-1] [Medline: 38198076]
18. Li ZR, Wang CY, Lin WW, Chen YT, Liu XQ, Jiao Z. Handling delayed or missed dose of antiseizure medications: a model-informed individual remedial dosing. Neurology 2023 Feb 28;100(9):e921-e931. [doi: 10.1212/WNL.0000000000201604] [Medline: 36450606]
19. Barrière O, Li J, Nekka F. A Bayesian approach for the estimation of patient compliance based on the last sampling information. J Pharmacokinet Pharmacodyn 2011 Jun;38(3):333-351. [doi: 10.1007/s10928-011-9196-2] [Medline: 21445612]
20. Ding J, Coldiron ME, Assao B, et al. Adherence and population pharmacokinetic properties of amodiaquine when used for seasonal malaria chemoprevention in african children. Clin Pharmacol Ther 2020 May;107(5):1179-1188. [doi: 10.1002/cpt.1707] [Medline: 31652336]
21. Nick H. Simulations with/without residual error. The Mail Archive. 2009. URL: https://www.mail-archive.com/nmusers@globomaxnm.com/msg01817.html [accessed 2025-11-30]
22. Jiao Z, Shi XJ, Zhao ZG, Zhong MK. Population pharmacokinetic modeling of steady state clearance of carbamazepine and its epoxide metabolite from sparse routine clinical data. J Clin Pharm Ther 2004 Jun;29(3):247-256. [doi: 10.1111/j.1365-2710.2004.00557.x] [Medline: 15153086]
23. Goto S, Seo T, Murata T, et al. Population estimation of the effects of cytochrome P450 2C9 and 2C19 polymorphisms on phenobarbital clearance in Japanese. Ther Drug Monit 2007 Feb;29(1):118-121. [doi: 10.1097/FTD.0b013e318030def0] [Medline: 17304159]
24. Pigeolet E, Jacqmin P, Sargentini-Maier ML, Stockis A. Population pharmacokinetics of levetiracetam in Japanese and Western adults. Clin Pharmacokinet 2007;46(6):503-512. [doi: 10.2165/00003088-200746060-00004] [Medline: 17518509]
25. Okada Y, Seo T, Ishitsu T, et al. Population estimation regarding the effects of cytochrome P450 2C19 and 3A5 polymorphisms on zonisamide clearance. Ther Drug Monit 2008 Aug;30(4):540-543. [doi: 10.1097/FTD.0b013e31817d842a] [Medline: 18641551]
26. Chhun S, Jullien V, Rey E, Dulac O, Chiron C, Pons G. Population pharmacokinetics of levetiracetam and dosing recommendation in children with epilepsy. Epilepsia 2009 May;50(5):1150-1157. [doi: 10.1111/j.1528-1167.2008.01974.x] [Medline: 19175400]
27. Girgis IG, Nandy P, Nye JS, et al. Pharmacokinetic-pharmacodynamic assessment of topiramate dosing regimens for children with epilepsy 2 to <10 years of age. Epilepsia 2010 Oct;51(10):1954-1962. [doi: 10.1111/j.1528-1167.2010.02598.x] [Medline: 20880232]
28. Nielsen JC, Kowalski KG, Karim A, Patel M, Wesche DL, Tolbert D. Population pharmacokinetics analysis of vigabatrin in adults and children with epilepsy and children with infantile spasms. Clin Pharmacokinet 2014 Nov;53(11):1019-1031. [doi: 10.1007/s40262-014-0172-z] [Medline: 25172554]
29. Saruwatari J, Ogusu N, Shimomasuda M, et al. Effects of CYP2C19 and P450 oxidoreductase polymorphisms on the population pharmacokinetics of clobazam and N-desmethylclobazam in japanese patients with epilepsy. Ther Drug Monit 2014 Jun;36(3):302-309. [doi: 10.1097/FTD.0000000000000015] [Medline: 24345815]
30. Ding J, Wang Y, Lin W, et al. A population pharmacokinetic model of valproic acid in pediatric patients with epilepsy: a non-linear pharmacokinetic model based on protein-binding saturation. Clin Pharmacokinet 2015 Mar;54(3):305-317. [doi: 10.1007/s40262-014-0212-8] [Medline: 25388986]
31. Schoemaker R, Wade JR, Stockis A. Brivaracetam population pharmacokinetics and exposure‐response modeling in adult subjects with partial‐onset seizures. J Clin Pharmacol 2016 Dec;56(12):1591-1602. [doi: 10.1002/jcph.761]
32. Schoemaker R, Wade JR, Stockis A. Brivaracetam population pharmacokinetics in children with epilepsy aged 1 month to 16 years. Eur J Clin Pharmacol 2017 Jun;73(6):727-733. [doi: 10.1007/s00228-017-2230-6] [Medline: 28280887]
33. Gidal BE, Jacobson MP, Ben-Menachem E, et al. Exposure-safety and efficacy response relationships and population pharmacokinetics of eslicarbazepine acetate. Acta Neurol Scand 2018 Sep;138(3):203-211. [doi: 10.1111/ane.12950] [Medline: 29732549]
34. Sunkaraneni S, Ludwig E, Fiedler-Kelly J, Hopkins S, Galluppi G, Blum D. Modeling and simulations to support dose selection for eslicarbazepine acetate therapy in pediatric patients with partial-onset seizures. J Pharmacokinet Pharmacodyn 2018 Aug;45(4):649-658. [doi: 10.1007/s10928-018-9596-7] [Medline: 29948795]
35. Takenaka O, Ferry J, Saeki K, Laurenza A. Pharmacokinetic/pharmacodynamic analysis of adjunctive perampanel in subjects with partial-onset seizures. Acta Neurol Scand 2018 Apr;137(4):400-408. [doi: 10.1111/ane.12874] [Medline: 29171002]

XSL•FO
RenderX

36. van Dijkman SC, de Jager NCB, Rauwé WM, Danhof M, Della Pasqua O. Effect of age-related factors on the pharmacokinetics of lamotrigine and potential implications for maintenance dose optimisation in future clinical trials. Clin Pharmacokinet 2018 Aug;57(8):1039-1053. [doi: 10.1007/s40262-017-0614-5] [Medline: 29363050]
37. Lin WW, Li XW, Jiao Z, et al. Population pharmacokinetics of oxcarbazepine active metabolite in Chinese paediatric epilepsy patients and its application in individualised dosage regimens. Eur J Clin Pharmacol 2019 Mar;75(3):381-392. [doi: 10.1007/s00228-018-2600-8] [Medline: 30456415]
38. Lin WW, Wang CL, Jiao Z, et al. Glomerular filtration rate is a major predictor of clearance of oxcarbazepine active metabolite in adult Chinese epileptic patients: a population pharmacokinetic analysis. Ther Drug Monit 2019 Oct;41(5):665-673. [doi: 10.1097/FTD.0000000000000644] [Medline: 31033858]
39. Winkler J, Schoemaker R, Stockis A. Population pharmacokinetics of adjunctive lacosamide in pediatric patients with epilepsy. J Clin Pharmacol 2019 Apr;59(4):541-547. [doi: 10.1002/jcph.1340] [Medline: 30427550]
40. Winkler J, Schoemaker R, Stockis A. Modeling and simulation for the evaluation of dose adaptation rules of intravenous lacosamide in children. Epilepsy Res 2019 Jan;149:13-16. [doi: 10.1016/j.eplepsyres.2018.10.011] [Medline: 30415109]
41. Wang ML, Tao YY, Sun XY, et al. Estrogen profile- and pharmacogenetics-based lamotrigine dosing regimen optimization: Recommendations for pregnant women with epilepsy. Pharmacol Res 2021 Jul;169:105610. [doi: 10.1016/j.phrs.2021.105610] [Medline: 33857625]
42. Teixeira-da-Silva P, Pérez-Blanco JS, Santos-Buelga D, Otero MJ, García MJ. Population Pharmacokinetics of Valproic Acid in Pediatric and Adult Caucasian Patients. Pharmaceutics 2022 Apr 7;14(4):811. [doi: 10.3390/pharmaceutics14040811] [Medline: 35456645]
43. Li Y, Wang ML, Guo Y, Cao YF, Zhao MM, Zhao LM. Population pharmacokinetics and dosing regimen optimization of levetiracetam in epilepsy during pregnancy. Br J Clin Pharmacol 2023 Mar;89(3):1152-1161. [doi: 10.1111/bcp.15572] [Medline: 36260320]
44. Li S, Yi J, Tuo Y, et al. Population pharmacokinetics and dosing optimization of perampanel in children with epilepsy: A real-world study. Epilepsia 2024 Jun;65(6):1687-1697. [doi: 10.1111/epi.17954] [Medline: 38572689]
45. Pharmacokinetic-based bayesian approach to assess medication adherence. MIPD. URL: https://mipd.shinyapps.io/adherence_ASM/ [accessed 2025-11-30]
46. Krekels EHJ, Calvier EAM, van der Graaf PH, Knibbe CAJ. Children are not small adults, but can we treat them as such? CPT Pharmacometrics Syst Pharmacol 2019 Jan;8(1):34-38. [doi: 10.1002/psp4.12366] [Medline: 30689298]
47. Pennell PB, Karanam A, Meador KJ, et al. Antiseizure medication concentrations during pregnancy: results from the Maternal Outcomes and Neurodevelopmental Effects of Antiepileptic Drugs (MONEAD) Study. JAMA Neurol 2022 Apr 1;79(4):370-379. [doi: 10.1001/jamaneurol.2021.5487] [Medline: 35157004]
48. Yin X, Liu Y, Guo Y, Zhao L, Li G, Tan X. Pharmacokinetic changes for newer antiepileptic drugs and seizure control during pregnancy. CNS Neurosci Ther 2022 May;28(5):658-666. [doi: 10.1111/cns.13796] [Medline: 35037389]
49. Gidal BE, Ferry J, Reyderman L, Piña-Garza JE. Use of extended-release and immediate-release anti-seizure medications with a long half-life to improve adherence in epilepsy: a guide for clinicians. Epilepsy Behav 2021 Jul;120:107993. [doi: 10.1016/j.yebeh.2021.107993] [Medline: 33971390]
50. Javarayee P, Mtchedlidze T, Snell W, et al. Pricing dynamics of anti-seizure medications in the U.S. Seizure 2024 Nov;122:26-33. [doi: 10.1016/j.seizure.2024.09.010] [Medline: 39306895]
51. Barrière O, Li J, Nekka F. Compliance spectrum as a drug fingerprint of drug intake and drug disposition. J Pharmacokinet Pharmacodyn 2013 Feb;40(1):41-52. [doi: 10.1007/s10928-012-9285-x] [Medline: 23250805]

Abbreviations

ASM: antiseizure medications
C_0: steady-state trough concentration
PK: pharmacokinetic
PPK: population pharmacokinetic
TDM: therapeutic drug monitoring

Edited by A Coristine; submitted 26.May.2025; peer-reviewed by J Liao, M Guignet, S Lawley, Y Wang; revised version received 09.Dec.2025; accepted 09.Dec.2025; published 02.Jan.2026.

Please cite as:
Liu XQ, Li ZR, Lin WW, Wang J, Gu FQ, Ding JJ, Jiao Z
Bayesian-Based Pharmacokinetic Framework Integrated with Therapeutic Drug Monitoring for Assessing Adherence to Antiseizure Medications: A Clinical Trial Simulation Study
J Med Internet Res 2026;28:e77917
URL: https://www.jmir.org/2026/1/e77917
doi:10.2196/77917

XSL•FO
RenderX

© Xiao-Qin Liu, Zi-Ran Li, Wei-Wei Lin, Juan Wang, Fu-Qing Gu, Jun-Jie Ding, Zheng Jiao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Real-Life Digital Intervention for Personalized Nutrition in Adults With Overweight or Obesity: Remote Randomized Controlled Trial

Jelle CBC de Jong[1*], PhD; Femke PM Hoevenaars[1*], PhD; Lotte GP Peters[1,2], MSc; Charlotte MM Berendsen[1], MSc; Wilrike J Pasman[1], PhD; Martien PM Caspers[1], BSc; Remon Dulos[1], BSc; Suzan Wopereis[1,2], PhD

[1]Department of Microbiology and Systems Biology, Netherlands Organisation for Applied Scientific Research, Leiden, The Netherlands

[2]Department of Biomedical Signals and Systems, University of Twente, Enschede, The Netherlands

[*]these authors contributed equally

Corresponding Author:
Suzan Wopereis, PhD
Department of Microbiology and Systems Biology
Netherlands Organisation for Applied Scientific Research
Sylviusweg 71
Leiden, 2333 BE
The Netherlands
Phone: 31 88 866 75 00
Email: suzan.wopereis@tno.nl

Abstract

Background: A digital-first strategy is increasingly implemented to reduce participant burden, accelerate recruitment, collect real-world data, and increase the diversity of the study population. However, fully remote studies lack face-to-face interaction, which may affect motivation, particularly in the delivery of personalized nutritional advice. Additionally, self-reported data may vary in terms of standardization and completeness.

Objective: The study's primary objective is to evaluate the feasibility of conducting a fully remote, fully digital randomized controlled nutritional intervention, including participant experience and the capability to perform do-it-yourself anthropometric measurements at home. Secondary objectives are to determine whether self-collected data could detect changes in body weight and other anthropometric outcomes, and to compare the effectiveness of generic versus personalized nutrition advice, with and without personalized food boxes.

Methods: We conducted a fully online, 3-arm randomized controlled trial including adults with overweight or obesity who were motivated to lose weight. Participants were assigned to a control group that received generic advice (n=43), a personalized intervention group that received personalized advice only (n=40), or a personalized intervention plus group that received personalized advice plus personalized food boxes (n=39). The 6-week intervention was delivered entirely digitally, and all anthropometric measurements, questionnaires, and dietary data were self-collected at home. Feasibility was assessed using adherence metrics, completion of self-measurements, and a user-experience questionnaire. Secondary analyses evaluated weight loss, changes in anthropometry, and exploratory associations, including sex differences.

Results: Feasibility was high—102 out of 122 (83.6%) participants found the self-measured anthropometric assessments easy to perform, and 112 (91.8%) participants reported that completing questionnaires from home was easy. For secondary outcomes, participants receiving personalized, but not generic, nutritional advice significantly lost body weight (–1.0 kg; *P*=.002). Participants receiving personalized food boxes in addition to personalized nutritional advice lost significantly more body weight than the other 2 groups (–2.5 kg; *P*=.001) and also showed a decrease in hip circumference (–2.9 cm; *P*=.01). Personalized advice was not easier or more enjoyable to implement than generic nutritional advice, whereas the addition of personalized food boxes improved the ease of implementing personalized nutritional advice (*P*<.001). All participants, irrespective of the intervention arm, reduced intake of unhealthy food groups, including ready-made meals (113.6 g vs 78.5 g, –30.9%); sauces and gravy (18.8 g vs 10.0 g, –46.8%); sweet snacks (84.8 g vs 64.1 g, –24.4%); savory snacks (50.5 g vs 40.0 g, –20.1%); bread, pasta, rice, and wraps (nutritional quality score of 1.9 vs 1.7, –10.5%); and vegetables (129.0 g vs 118.7 g, –8.0%); and replaced coffee with tea.

Conclusions: This study demonstrates that fully remote, participant-led nutritional intervention studies are feasible, with participants able to independently perform anthropometric measurements and self-report data of sufficient quality to detect meaningful effects. Personalized nutritional advice resulted in greater weight loss than generic advice, and the addition of

XSL•FO
RenderX

personalized food boxes further enhanced the beneficial anthropometric effects of the intervention. We conclude that such nutritional intervention studies can be conducted fully online, resulting in measurable anthropometric effects after 6 weeks.

Trial Registration: ClinicalTrials.gov NCT06547983; https://clinicaltrials.gov/ct2/show/NCT06547983

(J Med Internet Res 2026;28:e73367) doi:10.2196/73367

KEYWORDS

personalized nutrition; digital; dietary advice; behavior; feedback; digital first; real-world study; remote; do it yourself; DIY

Introduction

Decentralization of clinical trials is on the rise, meaning that participant inclusion and data collection are performed by the participants themselves, for example, from their homes. This may help attenuate the burden of trial participation, improve participant recruitment, and decrease study dropout [1]. In this way, decentralization helps to avoid insufficient recruitment and underpowered clinical trials [2,3], or complete trial failure [4,5]. Online recruitment and study inclusion may also create opportunities to reach a more diverse population, that is, a better representation of the complete population [6]. In addition, data collected through decentralized trials may represent real-world data, revealing whether interventions work when implemented in real-world settings.

Nutrition studies often require a large sample size, with possibly multiple follow-up measurements and multiple (control) groups. Therefore, decentralization is of interest in this field; however, this may also introduce novel caveats, since protocol adherence and, subsequently, data quality could be jeopardized, possibly leading to false conclusions. A limited number of nutritional studies have indicated that similar results can be produced at home compared with a controlled laboratory setting [7-9]. However, this topic remains understudied, and research is needed to assess participant experience and capability to perform simple measurements from home and fill in questionnaires.

Concomitantly, the paradigm of nutritional intervention studies aiming to improve dietary habits is shifting toward more personalized approaches. Personalized nutrition interventions are tailored for specific individuals based on their individual data, such as phenotype, nutritional habits, and biomarkers [10]. Ideally, personalized nutrition approaches also consider socioeconomic, behavioral, and cultural factors, as well as the food environment [11]. Decentralized studies conducted in real-world settings offer an opportunity to assess external validity and to explore how these contextual factors, such as the local food environment and broader food systems [12], influence intervention effectiveness. Moreover, they provide a unique opportunity to investigate behavioral patterns within the food environment and the food systems framework.

Studies have shown that such personalized nutritional advice is more effective than a generic one-size-fits-all approach in improving dietary habits [10,13] and in promoting weight loss in individuals with obesity [14,15]. However, more research is needed in this field to understand the contexts in which personalized nutrition is most effective [14]. Some studies have suggested that personalized nutrition is more effective than generic advice due to its face-to-face aspect and its perception as personally relevant [16,17]. Consequently, it can be questioned whether personalized nutrition interventions are as effective when delivered through the internet, possibly failing to provide recipients with the same level of motivation. Therefore, we chose to test the effectiveness of personalized versus generic nutritional interventions in this real-life, fully remote study.

In summary, the primary objective of this study was to evaluate the feasibility of conducting a fully remote, fully digital nutritional intervention, including participants' experience and their ability to perform do-it-yourself anthropometric measurements from home. In addition, feasibility was evaluated by examining the reliability of self-reported anthropometric data through analyses of correlations among changes in body weight, waist circumference, and hip circumference. Feasibility was defined as the capability of participants to independently perform all study procedures, including self-measured anthropometry and digital questionnaire completion, in a fully remote setting. Secondary objectives were to assess whether significant changes in body weight and other anthropometric outcomes could be detected using self-collected data, and to compare the effectiveness of generic versus personalized nutritional advice delivered digitally. In addition, we explored whether the provision of personalized food boxes enhanced adherence and led to greater weight loss, and we conducted exploratory analyses of sex differences.

Methods

Ethics Statement

This study was reviewed and approved by the independent Internal Ethical Review Board of The Netherlands Organization for Applied Scientific Research (TNO), registration number #2023-083, in September 2023. The study was registered at ClinicalTrials.gov (NCT06547983). All participants provided written informed consent before inclusion in the study. The study was conducted in accordance with the Declaration of Helsinki. All analyses presented in this paper were conducted in accordance with the primary approved study protocol. All data were anonymized. Each participant received a Christmas food box (valued at €90 [US $106]) as compensation for their participation in the trial. Additionally, depending on group allocation, some participants received free daily meals during the trial.

Study Population

Participants were 25-60 years of age and had a BMI between 25 and 40 kg/m^2. Participants were included only if they were motivated to lose weight; had the skills to complete digital

XSL•FO
RenderX

questionnaires; and had a computer and a smartphone, a weighing scale, and a measuring tape to perform the anthropometric measurements, as confirmed in the screening questionnaire (Multimedia Appendix 1). Participants were excluded if they had a food allergy, followed specific diets (eg, a keto or vegan diet), suffered from a chronic disease that influences food intake (eg, inflammatory bowel disease), or participated in another intervention study. Females in menopause transition, as defined through the screening questionnaire, were excluded as well. A formal power calculation was not feasible because reliable estimates of effect size and variance for this specific intervention and outcomes were not available at the time of study design. Instead, the sample size was based on the results of previous studies [13,18,19], in which 37, 59, or 82 participants were included and exposed to personalized nutritional advice for 16, 9, or 10 weeks, respectively. Consequently, a sample size of 150 at the start of the intervention was considered acceptable for this study. This number was higher than those in the cited previous studies due to the shorter intervention period (6 weeks), which was expected to require a larger sample size. To compensate for an anticipated 10%-20% dropout rate during the intervention, a total sample size of 180 participants was selected.

Participant recruitment was organized via advertisements on social media, as well as through the Dutch participant database of Norstat [20], focusing on zip codes to which food boxes could be delivered (Uitgekookt Meal Service). Only individuals who indicated that their motivation to participate in the study was to lose weight were selected (participants could select 1 of 4 predefined reasons). These participants were randomly allocated to 1 of 3 intervention groups, while taking into account the covariates biological sex, age, and BMI. Allocation was performed by calculating the Euclidean distance for the variable sex and the Kullback-Leibler divergence for the variables age and BMI for each permutation separately. The permutation with the best overall score (ie, the most balanced distribution among the groups) was selected. This allocation process was automated and blinded to both participants and investigators involved in data collection and analysis.

Study Design

The primary objective of this study was to evaluate the feasibility of conducting a fully remote, fully digital nutritional intervention trial, including participants' ability to perform do-it-yourself anthropometric measurements, adherence to the protocol, and user experience. To further evaluate feasibility, the reliability of self-reported anthropometric data was examined by analyzing correlations among changes in body weight, waist circumference, and hip circumference. Secondary objectives were to assess whether self-collected data could detect significant changes in body weight and other anthropometric measures, and to compare the effectiveness of generic versus personalized nutritional advice, with and without personalized food boxes. Exploratory analyses included examining sex differences to assess the reliability of the self-assessed anthropometric data. The study was a randomized controlled trial with 1 control intervention arm and 2 personalized intervention arms (Figure 1). The study duration was 6 weeks. Participants allocated to the control group received generic nutritional advice (see example in Multimedia Appendix 2), which was based on the Dutch national guidelines for healthy nutrition, assembled by the Dutch Health Council (Gezondheidsraad), an independent scientific advisory body for the Dutch government.

Figure 1. Overview of the study design. Self-measured anthropometric data, including body weight, and assessments of fiber intake and physical activity (via questionnaires) were collected through the online How Am I app at weeks 0, 2, 4, and 6. All intervention groups received feedback on their self-assessments at weeks 0, 2, 4, and 6. Nutritional advice provided to intervention groups P and PP was updated based on participants' self-assessments. Qualitative assessments of study experience were collected at weeks 2, 4, and 6. Habitual food intake was assessed at the beginning and end of the trial using the Nutri+ module. P-group: personalized intervention group; PP-group: personalized plus intervention group.

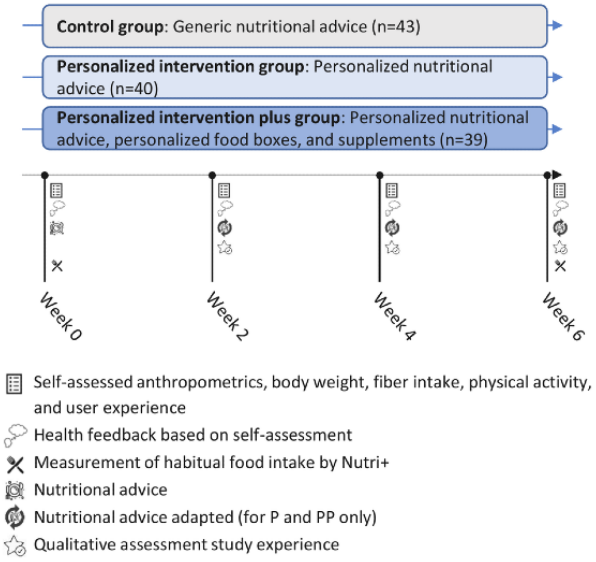

XSL•FO
RenderX

Study Intervention

Participants allocated to the personalized intervention group (P-group) received personalized advice instead of generic advice. The personalized advice entailed information on the specific number of calories to consume on a daily basis to meet their personal goal of weight loss. Furthermore, the proportion of macronutrients the participant should consume to reach their weight loss goal was also made explicit in the nutritional advice (see example in Multimedia Appendix 2). The prescribed caloric intake provided a daily energy deficit of 400 kcal below the daily energy requirement, as calculated using the Ten Haaf Formula [21] at the start of the trial, based on personal information including age, sex, and physical activity level (PAL). In addition, instructions were provided for the implementation of the Eetmeter application (Voedingscentrum [22]), a freely available digital application that participants were advised to use to check whether their daily meals matched their personalized nutritional advice according to the Dutch dietary guidelines.

Participants allocated to the personalized plus intervention group (PP-group) received the same personalized advice as the P-group; however, they also received personalized food boxes containing meals that were composed to match the personalized nutritional advice in terms of calories and macronutrients. These food boxes contained 3 main meals (breakfast, lunch, and dinner) and 3 snacks to be consumed between the main meals (Uitgekookt Meal Service, IJsselmuiden). The food boxes provided a daily energy deficit of 200-600 kcal below the daily energy requirement. The food boxes were provided on 5 out of 7 days per week. On the remaining 2 days, participants were instructed to use the online Eetmeter application to select and follow their diet based on the provided personalized advice.

Participants allocated to groups receiving personalized interventions received updated nutritional advice based on the outcomes of the self-measurements in terms of calories and macronutrients, whereas the control group received the same generic advice at weeks 2, 4, and 6.

Remote Data Collection

The study was fully digital, without any face-to-face contact between participants and researchers. To this end, participants were required to install the online research application, the How Am I app (TNO), through which all instructions, feedback, nutritional advice, questionnaires, and reminders were provided, and which served as the data collection platform [23]. At weeks 0, 2, 4, and 6, participants submitted data via the How Am I app on feasibility-related user experience items, self-measured anthropometric measurements, body weight, and brief questions related to fiber intake and physical activity (Multimedia Appendix 1). PAL and fiber intake were calculated according to validated methodologies, as described by Healey et al [24] and in the Food and Agriculture Organization (FAO)/World Health Organization (WHO)/United Nations University (UNU) Expert Consultation Report [25]. At weeks 2, 4, and 6, all participants received a reminder via the TNO How Am I app to complete all anthropometric measurements and questions and were provided with feedback on their measurements (see Multimedia Appendix 2). Researchers monitored data entry, issuing a reminder to participants who did not complete their questionnaires within 48 hours and instructing them to respond within an additional 24-hour period. Participants who remained noncompliant after this interval were excluded from the study.

Feasibility Outcomes

The primary outcome of this study was the feasibility of conducting a fully remote, fully digital nutritional intervention. Feasibility was assessed at weeks 2, 4, and 6 through participants' reported user experience, including ease of performing anthropometric measurements, clarity of instructions, and perceived burden, measured using multiple-choice questions, Likert scales, and visual analog scales (see Multimedia Appendix 1). Feasibility was further evaluated by examining the reliability of self-reported anthropometric data, based on correlations among changes in body weight, waist circumference, and hip circumference.

Self-Measured Anthropometrics and Body Weight

Secondary outcomes related to anthropometrics and body weight were also evaluated. At baseline and at weeks 2, 4, and 6, all participants were requested to self-measure body weight (kg). Using a measuring tape, participants were asked to measure height (cm), waist circumference (cm), and hip circumference (cm) according to standard operating procedures [26]. Detailed instructions on performing the anthropometric measurements were provided via the TNO How Am I app through videos containing step-by-step guidance. Body weight and height were used to calculate BMI (kg/m^2).

Measurement of Habitual Dietary Intake

At the start and end of the trial, all participants completed the Nutri+ module (TNO), a web-based questionnaire consisting of 89 questions that measures the dietary intake of 21 food groups for each participant and thereby estimates habitual food intake. Consequently, the average food intake for certain food groups (eg, fruit, meat, dairy) or the average frequency of consumption per week (eg, bread, potatoes, cooking fats) was calculated.

Statistical Analysis

All analyses were conducted using GraphPad Prism 10 (GraphPad Software) or IBM SPSS Statistics 29. Outlier analysis was performed, and data points greater than 3× the IQR were excluded. As the study aimed to evaluate intervention efficacy based on complete longitudinal data, analyses were performed per protocol. Participants with incomplete datasets (eg, due to dropping out) were not included in the analysis, as missing repeated-measures data would have violated model assumptions and precluded valid within-participant comparisons over time. Normality was confirmed using the D'Agostino and Pearson test. All data are reported as means (SD). Two-tailed *P* values were used, and *P* values <.05 were considered statistically significant.

User experience data collected via Likert scales, multiple-choice questions, and visual analog scales were analyzed descriptively. Group differences in categorical or ordinal feasibility responses were evaluated using Pearson chi-square tests, with Bonferroni-adjusted post hoc tests applied when appropriate.

XSL•FO
RenderX

The internal validity of self-reported anthropometric measurements was examined through Pearson correlations among changes in body weight, waist circumference, and hip circumference.

Differences between groups at baseline were tested using one-way ANOVA with Tukey post hoc tests. Anthropometric outcomes collected across multiple time points (weeks 0, 2, 4, and 6) were analyzed using repeated-measures ANOVA, with "time" as a within-participant factor and "group" as a between-participant factor. The Tukey post hoc tests were performed when main effects or interactions were detected. Categorical variables were compared using Pearson chi-square tests with Bonferroni corrections. Sex-stratified analyses were also performed to explore potential sex differences in anthropometric changes by testing for significant interaction effects with sex. Correlations between variables were assessed using the Pearson correlation coefficient, and statistical significance was defined as $P<.05$.

Results

Baseline Study Population Characteristics

A total of 315 participants were screened, and of these, 91 were excluded from participation for various reasons (Figure 2). The remaining 224 participants were invited to participate in the study, of whom 180 responded positively and were included in the study and randomly allocated to 1 of the 3 groups. Of the 180 recruited participants, 43 did not respond to the questionnaires and were consequently excluded from the study (Figure 2). Data from the remaining 137 participants were included in the analysis. An additional 15 participants were excluded based on outlier analysis, resulting in a final sample size of 122 participants. Results are reported in accordance with CONSORT (Consolidated Standards of Reporting Trials)-EHEALTH recommendations (Multimedia Appendix 3).

Figure 2. A flowchart of the recruitment and selection process, randomization, and group allocation of included participants.

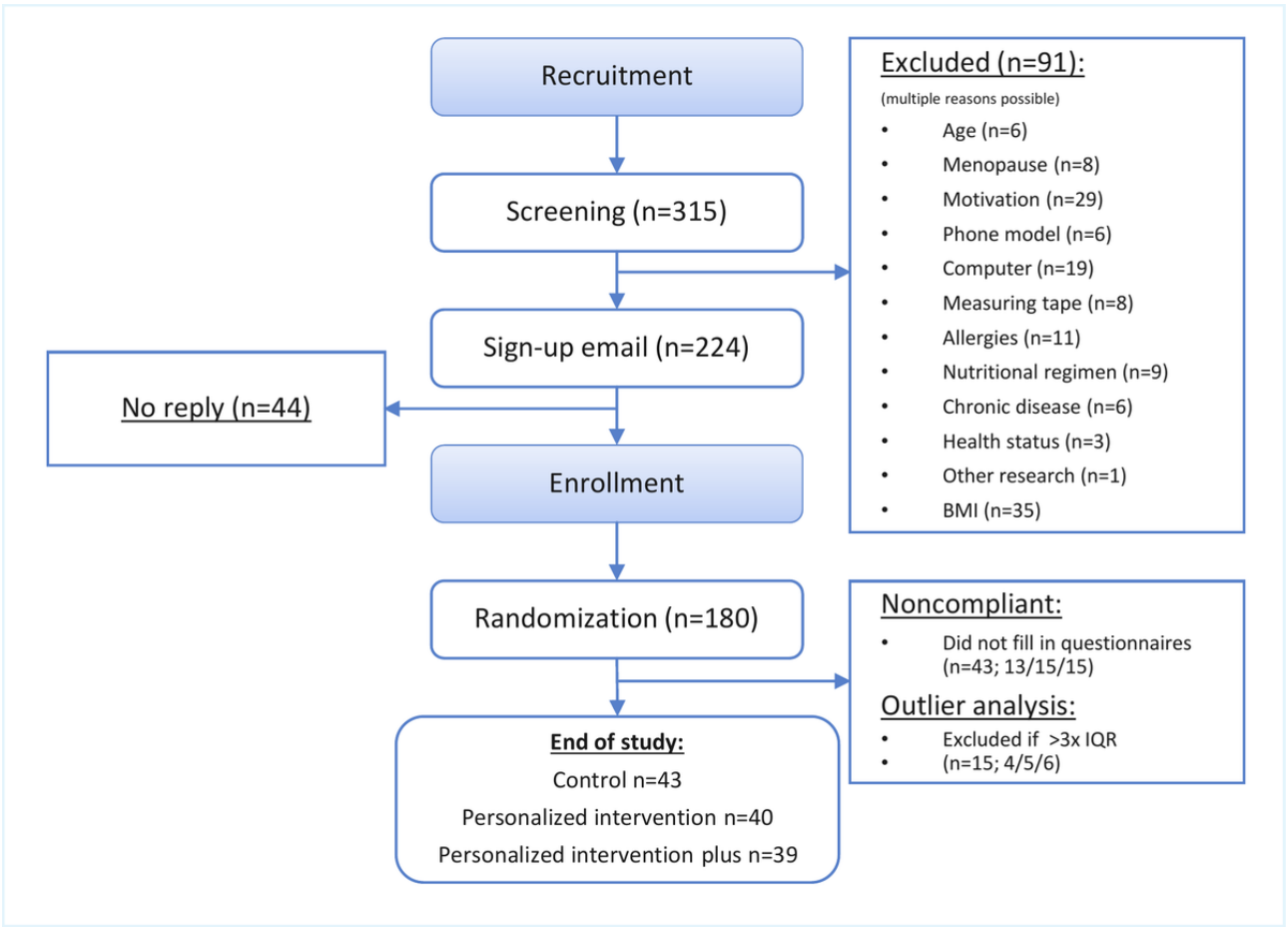

No statistically significant differences in any baseline characteristics were detected among the 3 intervention groups (Table 1). Habitual food intake was also measured for each food group before the intervention (Table 2). No statistically significant differences in habitual food intake were detected among the 3 intervention groups for any food group, except for sauces and gravy. Similarly, no statistically significant differences were found in baseline PAL among the 3 groups (Table 3; $P=.15$).

XSL•FO
RenderX

Table 1. Baseline characteristics of the study population.

Characteristics	Control (n=43)	Personalized intervention (n=40)	Personalized plus intervention (n=39)	*P* value
Age (years), mean (SD)	39.9 (6.1)	41.0 (6.0)	40.9 (5.9)	.60
Female sex, n (%)	27 (63)	27 (68)	19 (49)	.18
Weight (kg), mean (SD)	89.9 (12.0)	89.5 (13.3)	92.4 (14.1)	.65
Height (m), mean (SD)	1.75 (0.09)	1.75 (0.11)	1.75 (0.10)	.54
BMI (kg/m^2), mean (SD)	30.4 (4.1)	30.2 (4.5)	31.2 (4.8)	.81
Waist circumference (cm), mean (SD)	99.0 (10.1)	101.3 (10.0)	103.0 (12.4)	.26
Hip circumference (cm), mean (SD)	108.3 (10.1)	108.3 (7.5)	110.8 (8.9)	.36
Self-perceived health (out of 5 points, 5 being the highest score), mean (SD)	3.0 (0.6)	3.2 (0.7)	3.0 (0.8)	.48
Physical activity level, mean (SD)	1.40 (0.2)	1.44 (0.1)	1.40 (0.1)	.15
Fiber intake (g), mean (SD)	11.3 (5.2)	12.4 (8.4)	12.5 (10.6)	.99

Table 2. Baseline habitual food intake (based on the Nutri+ module) per food group of the study population.[a]

Food group	Control (n=43), mean (SD)	Personalized intervention (n=40), mean (SD)	Personalized plus intervention (n=39), mean (SD)	*P* value
Fruit (g/day)	141.6 (110.3)	136.7 (130.6)	113.3 (95.4)	.53
Vegetables (g/day)	122.3 (47.6)	132.6 (56.7)	125.2 (63.4)	.69
Legumes (g/week)	172.3 (284.0)	148.1 (202.1)	151.2 (161.6)	.91
Bread (nutritional quality score)	1.8 (0.6)	1.8 (0.7)	2.1 (0.7)	.09
Pasta, rice, and wraps (nutritional quality score)	1.8 (0.9)	1.9 (0.9)	2.1 (0.7)	.23
Potatoes (nutritional quality score)	2.4 (1.5)	2.7 (1.7)	2.2 (1.3)	.36
Meat (g/week)	2367.9 (1737.8)	2422.8 (1336.4)	2584.9 (1650.3)	.79
Fish (g/week)	348.8 (701.5)	223.1 (242.1)	218.9 (232.9)	.40
Eggs (g/week)	209.5 (305.1)	263.8 (210.6)	233.8 (163.3)	.60
Dairy (g/day)	314.6 (573.5)	308.3 (301.5)	197.3 (193.3)	.39
Nuts (g/day)	15.1 (22.7)	12.9 (16.4)	8.3 (10.3)	.24
Spreadable fats (nutritional quality score)	1.1 (0.9)	1.2 (1.0)	1.1 (0.8)	.94
Cooking fats (nutritional quality score)	1.5 (0.6)	1.3 (0.7)	1.4 (0.8)	.47
Sweet snacks (g/week)	77.6 (83.3)	83.4 (83.9)	94.3 (68.6)	.64
Savory snacks (g/week)	55.6 (74.0)	41.8 (37.4)	54.3 (49.4)	.39
Ready-made meals (g/week)	127.3 (218.5)	112.9 (85.9)	98.8 (97.2)	.73
Sauces and gravy (g/week)	15.0 (20.6)[b]	24.5 (22.8)[b]	17.0 (17.1)[b]	.02
Alcohol (units drank/day)	0.7 (1.1)	0.7 (1.2)	0.5 (0.8)	.82
Soft drinks and fruit juice (g/day)	209.4 (374.2)	176.1 (272.0)	278.3 (489.4)	.49
Tea (g/day)	453.1 (448.1)	360.7 (390.7)	400.2 (399.0)	.63
Coffee (g/day)	424.7 (476.0)	338.8 (378.1)	403.5 (451.5)	.67

[a]Data represent the average habitual food intake. Nutritional quality scores reflect factors such as whether whole wheat versus multigrain products were consumed or whether olive oil versus butter is used for frying. A lower score represents a healthier food choice. Extensive descriptions of these nutritional quality scores are attached as Multimedia Appendix 4.

[b]Control versus personalized intervention, *P*=.02; control versus personalized plus intervention, *P*=.54; personalized intervention versus personalized plus intervention, *P*=.23.

XSL•FO
RenderX

Table 3. Physical activity levels during the study period.

Physical activity level	Week 0	Week 2	Week 4	Week 6
Control group, mean (SD)	1.40 (0.2)	1.40 (0.2)	1.42 (0.2)	1.41 (0.2)
Personalized intervention, mean (SD)	1.44 (0.1)	1.45 (0.1)	1.44 (0.1)	1.45 (0.1)
Personalized plus intervention, mean (SD)	1.40 (0.1)	1.41 (0.1)	1.40 (0.1)	1.40 (0.1)

Participant-Reported Usability and Reliability of Measurements

The ease of collecting anthropometric data by participants themselves and completing the questionnaires digitally via the How Am I app was reported (Multimedia Appendix 1). The majority of participants indicated that they could perform the anthropometric measurements very easily or easily (102/122, 83.6%, participants across all groups; Figure 3A), with no differences among groups. In addition, most participants reported that they could understand the questionnaires (112/122, 91.8%, participants across all groups; Figure 3B). The quality of anthropometric data was cross-validated, as significant correlations were detected between changes in body weight, waist circumference, and hip circumference (week 6 to week 0; all $P<.001$, $r>0.34$; Figure 3C-3E).

Figure 3. Ease and ability to perform anthropometric measurements at home and to complete questionnaires. (A) Ease of independently performing anthropometric measurements. (B) Clarity of the questionnaires. (C) Correlation between Δ body weight and Δ waist circumference. (D) Correlation between Δ body weight and Δ hip circumference. (E) Correlation between Δ waist and Δ hip circumference. In all correlation analyses, Δ values were calculated as the difference between measurements at the beginning and end of the 6-week trial. P-group: personalized intervention group; PP-group: personalized plus intervention group.

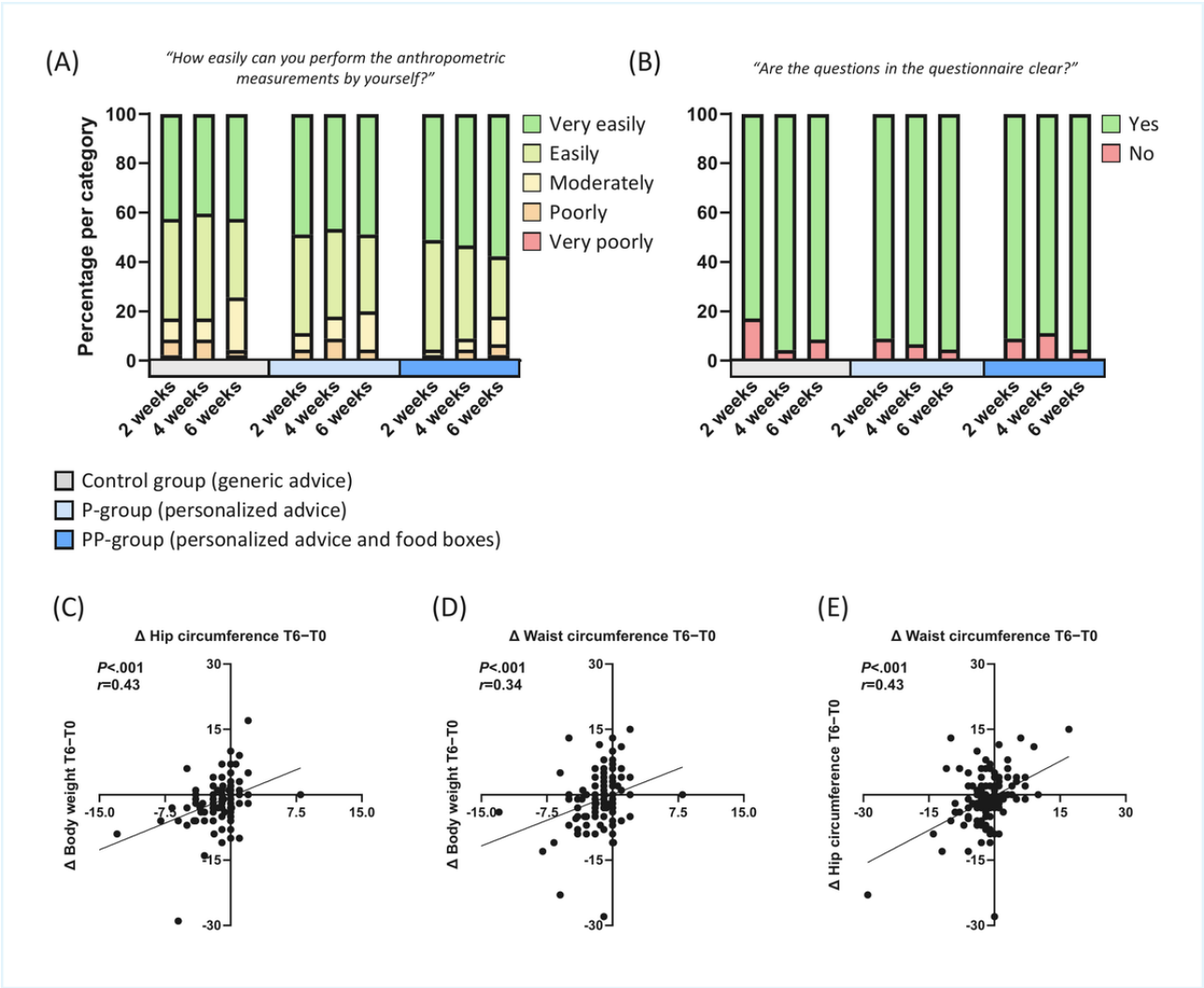

Trial Experience and Self-Reported Dietary Changes

No significant differences were detected in the experienced ease of implementing generic versus personalized nutritional advice during the trial ($P=.43$; Figure 4A). However, participants in the PP-group, who received personalized food boxes in addition to the personalized nutritional advice, reported that they could implement the nutritional advice more easily ($P<.001$; Figure 4A). At week 4 only, participants in the PP-group reported experiencing a greater amount of pleasure while implementing

XSL•FO
RenderX

their nutritional advice (P<.001 at week 4; Figure 4B), whereas no differences were observed between the groups receiving generic versus personalized nutritional advice. No differences in self-reported changes in diet were detected between the control group and the P-group (Figure 4C), while the PP-group consistently reported significant changes in their diet during the trial (P<.001; Figure 4C). Similarly, the PP-group reported significantly more often that they assessed their diet as being healthier at week 2 (P=.01; Figure 4D).

Figure 4. Participants' experience of participating in the fully digital trial. (A) Ease of implementing the nutritional advice during the trial. (B) Enjoyment of implementing the nutritional advice during the trial. (C) Self-reported changes in participants' diets during the trial. (D) Self-reported improvements in dietary healthfulness during the trial. *P<.05, **P<.01, ***P<.001. P-group: personalized intervention group; PP-group: personalized plus intervention group.

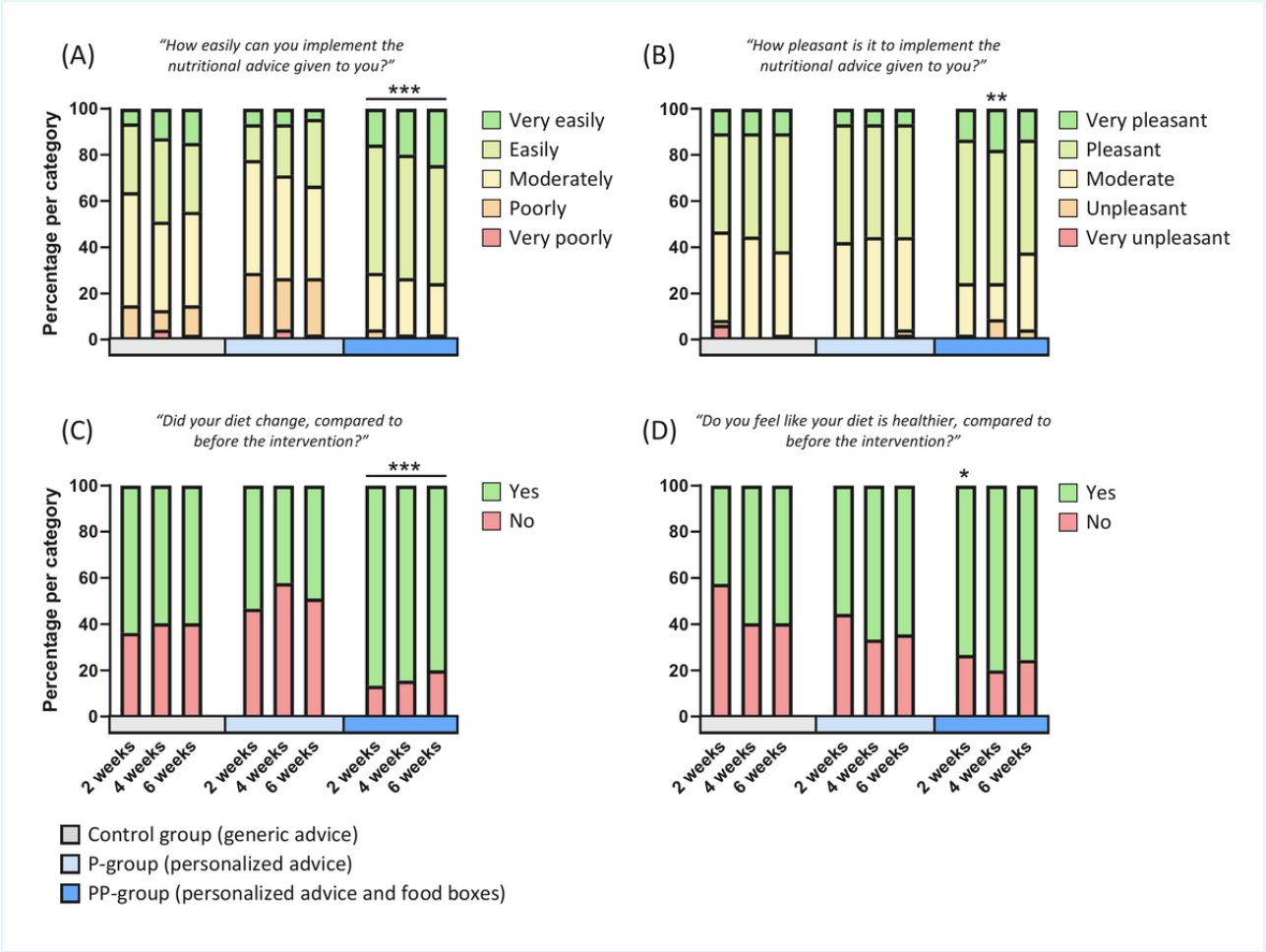

Changes in Body Weight, BMI, and Anthropometric Measures

After 6 weeks, the average body weight of the control group did not change compared with the start of the trial (P=.06; Figure 5A). In the P-group, body weight was significantly lower at week 4 (–0.8 kg; P=.04) and week 6 (–1 kg; P=.002) compared with baseline (Figure 5A). In the PP-group, body weight was significantly lower at week 2 (–0.9 kg; P=.004), week 4 (–1.6 kg; P=.002), and week 6 (–2.3 kg; P=.001) compared with baseline (Figure 5A). With respect to BMI, similar differences between the groups were observed (Figure 5B). BMI decreased significantly more in the PP-group (–0.8 kg/m^2, P=.001) compared with the P-group (–0.3 kg/m^2; P=.002) and the control group (–0.2 kg/m^2; P=.06) at week 6 (Figure 5B).

XSL•FO
RenderX

Figure 5. Effects of the different interventions on anthropometric measurements. (A) Body weight, (B) BMI, (C) hip circumference, and (D) waist circumference during the trial. Data are expressed as means (SDs). *P<.05 and **P<.01. The $ symbols indicate significant differences attributable to a statistically significant time effect in the absence of an interaction effect. P-group: personalized intervention group; PP-group: personalized plus intervention group.

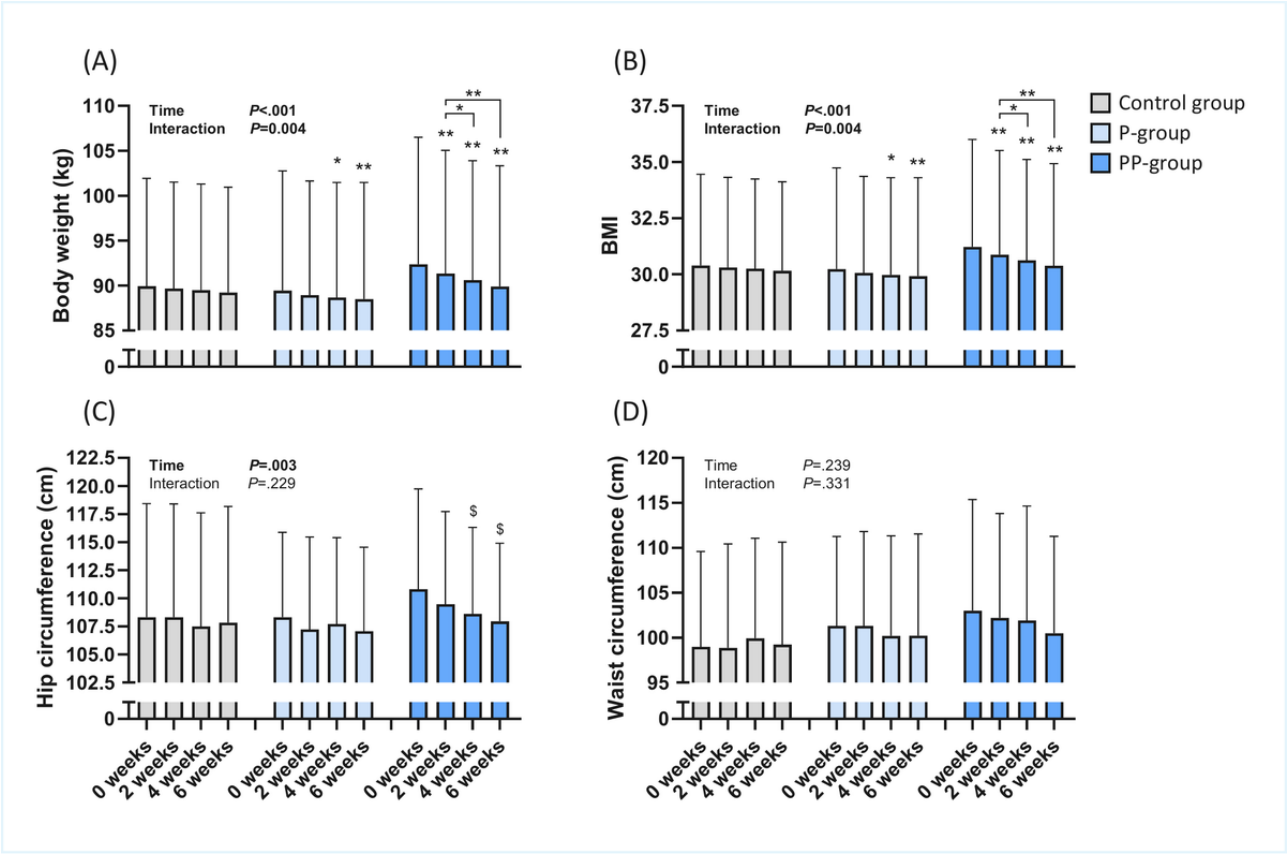

Hip circumference decreased significantly over time (P=.01), but no significant interaction between time and group was detected (P=.22; Figure 5C). Post hoc analysis revealed that the time effect was mainly driven by a significant loss of hip circumference in the PP-group, which decreased at week 4 (–2.2 cm; P=.04) and week 6 (–2.9 cm; P=.01) compared with baseline (Figure 5C). No significant effects were observed on waist circumference ($P_{interaction}$ =.33, P_{time}=.24; Figure 5D).

Changes in Habitual Food Intake and Physical Activity

Changes in intake of the food groups vegetables, bread, pasta/rice/wraps, sweet snacks, savory snacks, ready-made meals, sauces and gravy, tea, and coffee were observed (Table 4). The intake of all these food groups was lower after 6 weeks compared with baseline (vegetables, –9.0 g; bread, –0.2 times eaten per week; pasta/rice/wraps, –0.2 times eaten per week; sweet snacks, –21.1 g; savory snacks, –10.7 g; ready-made meals, –60.3 g; sauces and gravy, –8.7 g; coffee, –62.1 g), except tea, which significantly increased (+64.5 g; Table 4). No changes were detected for the food groups fruit, legumes, potatoes, meat, fish, eggs, dairy, nuts, spreadable and cooking fats, alcohol, or soft drinks and fruit juice. Only time effects were significant (see Table 4), without interaction effects, implying that changes in food intake occurred across all intervention groups, but no group-specific effects were observed. Only the food group legumes nearly reached statistical significance for both personalized interventions compared with the control group (P=.051), with increased intake postintervention for both the P- and PP-groups, whereas the control group, on average, showed decreased intake postintervention. Fiber intake was measured every 2 weeks, but no changes were detected (Multimedia Appendix 5). In addition, PALs did not change during the study (Table 3).

XSL•FO
RenderX

Table 4. Habitual food intake per food group before and after the intervention.a

Food group	Control (preintervention), mean (SD)	Control (postintervention), mean (SD)	Personalized intervention (preintervention), mean (SD)	Personalized intervention (postintervention), mean (SD)	Personalized plus intervention (preintervention), mean (SD)	Personalized plus intervention (postintervention), mean (SD)	Time effect (*P* value)	Interaction effect (*P* value)
Fruit (g/day)	141.6 (110.3)	142.5 (101.4)	136.7 (130.6)	128.8 (136.5)	113.3 (95.4)	105.5 (82.4)	.64	.92
Vegetables (g/day)	122.3 (47.6)	113.4 (44.0)	132.6 (56.7)	116 (52.0)	125.2 (63.4)	123.6 (57.6)	.03[b]	.35
Legumes (g/week)	172.3 (284.0)	121.7 (166.8)	148.1 (202.1)	172.0 (266.3)	151.2 (161.6)	183.6 (225.3)	.90	.051
Bread (nutritional quality score)	1.8 (0.6)	1.6 (0.7)	1.8 (0.7)	1.7 (0.7)	2.1 (0.7)	1.9 (0.6)	.02[b]	.37
Pasta, rice, and wraps (nutritional quality score)	1.8 (0.9)	1.6 (0.8)	1.9 (0.9)	1.5 (0.9)	2.1 (0.7)	1.9 (0.8)	<.01[b]	.57
Potatoes (nutritional quality score)	2.4 (1.5)	1.6 (1.6)	2.7 (1.7)	1.5 (1.5)	2.2 (1.3)	1.9 (1.1)	.71	.97
Meat (g/week)	2367.9 (1737.8)	2428.8 (1809.2)	2422.8 (1336.4)	2473.0 (1605.4)	2584.9 (1650.3)	2280.3 (1300.9)	.65	.50
Fish (g/week)	348.8 (701.5)	337.5 (710.7)	223.1 (242.1)	229.4 (267.8)	218.9 (232.9)	257.4 (237.1)	.61	.64
Eggs (g/week)	209.5 (305.1)	211.9 (265.7)	263.8 (210.6)	248.8 (250.0)	233.8 (163.6)	212.2 (200.1)	.40	.76
Dairy (g/day)	314.6 (573.5)	327.0 (800.3)	308.3 (301.5)	284.0 (383.7)	197.3 (193.3)	159.1 (172.0)	.49	.68
Nuts (g/day)	15.1 (22.7)	16.2 (19.9)	12.9 (16.4)	15.3 (18.3)	8.3 (10.3)	11.7 (13.1)	.11	.80
Spreadable fats (nutritional quality score)	1.1 (0.9)	1.1 (0.8)	1.2 (1.0)	1.1 (0.8)	1.1 (0.8)	1.0 (0.8)	.65	.91
Cooking fats (nutritional quality score)	1.5 (0.6)	1.2 (0.7)	1.3 (0.7)	1.3 (0.8)	1.4 (0.8)	1.3 (0.8)	.32	.41
Sweet snacks (g/week)	77.6 (83.3)	64.5 (89.6)	83.4 (83.9)	65.4 (56.4)	94.3 (68.6)	62.2 (60.6)	<.01[b]	.59
Savory snacks (g/week)	55.6 (74.0)	48.0 (67.6)	41.8 (37.4)	35.0 (37.0)	54.3 (49.4)	36.5 (47.8)	<.01[b]	.47
Ready-made meals (g/week)	127.3 (218.5)	77.4 (114.7)	112.9 (85.9)	78.2 (71.2)	98.8 (97.2)	80.0 (88.3)	.01[b]	.66
Sauces and gravy (g/week)	15.0 (20.6)	6.7 (9.1)	24.5 (22.8)	13.8 (18.2)	17.0 (17.1)	9.9 (14.3)	<.01[b]	.70
Alcohol (U/day)	0.7 (1.1)	0.7 (1.2)	0.7 (1.2)	0.5 (0.9)	0.5 (0.8)	0.5 (0.7)	.66	.66
Soft drinks and fruit juice (g/day)	209.4 (374.2)	212.1 (345.9)	176.1 (272.0)	204.8 (341.9)	278.3 (489.4)	218.3 (329.9)	.77	.66
Tea (g/day)	453.1 (448.1)	536.6 (504.6)	360.7 (390.7)	403.7 (433.0)	400.2 (399.0)	467.2 (464.2)	.03[b]	.95
Coffee (g/day)	424.7 (476.0)	396.3 (449.2)	338.8 (378.1)	303.6 (384.7)	403.5 (451.5)	280.9 (277.5)	.02[b]	.20

[a]Nutritional quality scores reflect factors such as whether whole wheat versus multigrain products were consumed or whether olive oil versus butter is used for frying. A lower score represents a healthier food choice. Extensive descriptions of these nutritional quality scores are attached as Multimedia Appendix 4.

[b]Significant values (ie, <.05).

XSL•FO
RenderX

Sex Differences and Correlations in Body Composition Changes

Exploratory analyses revealed that the waist-to-hip ratio was higher in males compared with females (Figure 6A). Furthermore, weight loss correlated similarly with loss in hip circumference in females (r=0.48, P<.001) and males (r=0.44, P<.001), as no significant interaction with biological sex was detected (P=.15; Figure 6B). Body weight loss correlated differently with loss in waist circumference in females (r=0.26, P=.02) and males (r=0.44, P<.001), as a significant interaction with biological sex was detected (P=.01; Figure 6C).

Figure 6. Sex differences in anthropometric measurements and fat mass loss. (A) Sex differences in waist-to-hip ratio at week 0, assessed fully online; data are expressed as means (SDs). Sex-specific correlations between Δ body weight and Δ waist circumference (B) and Δ hip circumference (C). In all correlation analyses, Δ values were calculated as the difference between measurements at the start and end of the 6-week trial. *P<.05.

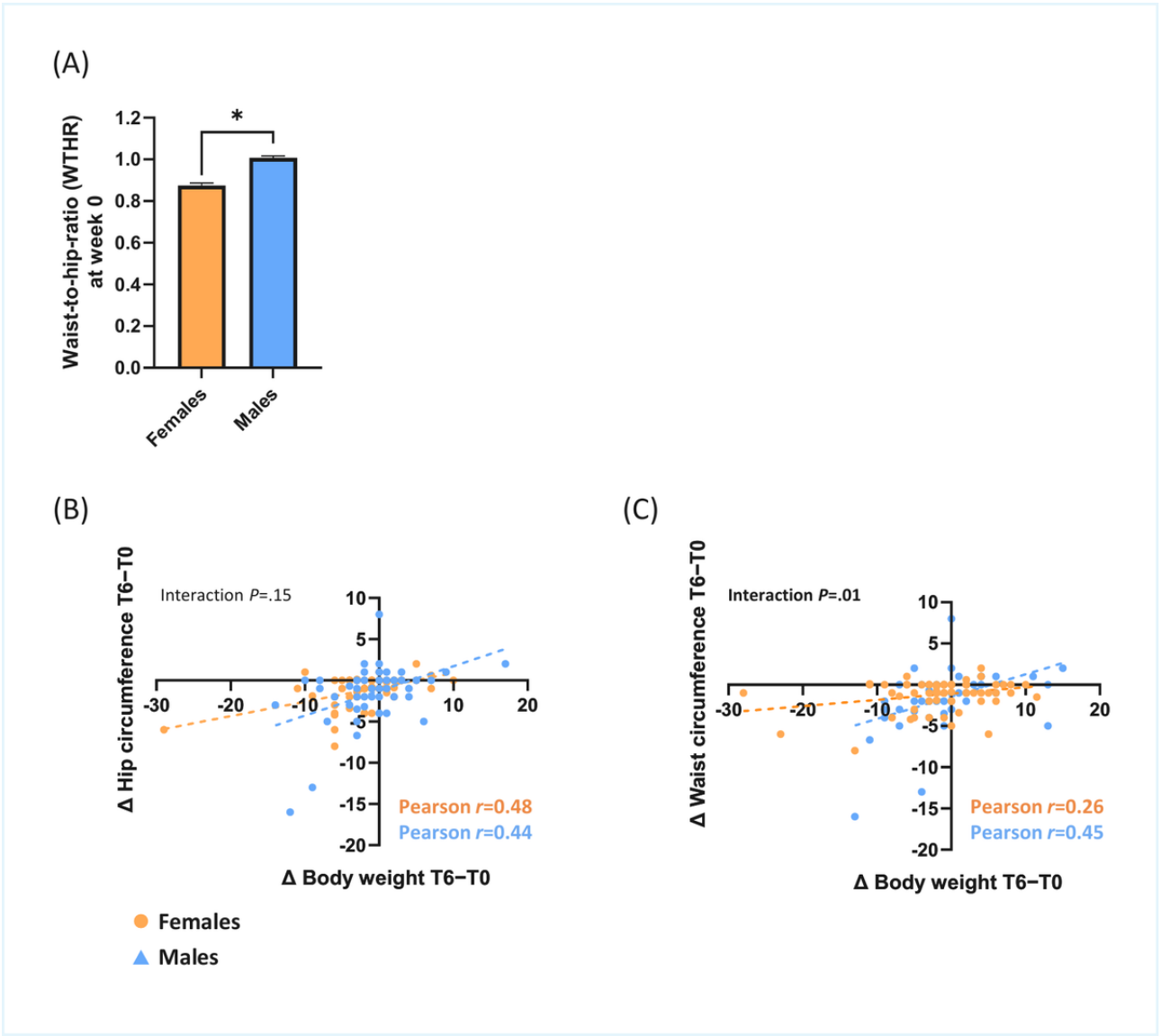

Discussion

Principal Findings

The primary aim of this study was to evaluate the feasibility of conducting a fully remote, fully digital nutritional intervention in adults with overweight or obesity. Our findings demonstrate that such a digital-first approach is feasible: most participants reported that anthropometric measurements were easy to perform at home, the digital questionnaires were clear and understandable, and the reliability of self-reported measurements was supported by consistent correlations among changes in body weight, waist circumference, and hip circumference (Figure 3). Together, these results indicate that participants were able to self-collect anthropometric data of sufficient quality for use in a randomized controlled trial.

The secondary aim of the study was to assess whether meaningful anthropometric changes could be detected using self-collected data. Participants who received personalized nutritional advice lost more body weight than those receiving generic advice (Figure 5), confirming the greater effectiveness of personalized advice even in a fully digital format. This effect was not explained by differences in the ease of implementing the advice, which increased only when personalized food boxes were provided in addition to the advice (Figure 4). The addition of personalized food boxes further increased weight loss and reduced hip circumference more than either form of advice

XSL•FO
RenderX

alone. Taken together, these results indicate that nutrition studies can be effectively conducted in a real-life, fully remote setting, achieving measurable intervention effects while maintaining ease of participation and high-quality data collection.

Exploratory analyses provided additional insights into sex differences in anthropometric responses. As expected from prior literature, males had higher waist-to-hip ratios and showed stronger correlations between weight loss and reductions in waist circumference compared with females. These findings support the biological plausibility of the self-measured data and highlight potential sex-specific patterns in responsiveness to nutritional interventions.

Comparison With Prior Work

An important factor underlying the success of personalized nutrition is the tailoring of the information provided to the recipient [27], which makes the advice perceived as personally relevant, for which face-to-face contact could be an important contributor [16,17]. Digital delivery of advice lacks this potentially important component, which may reduce its effectiveness. Importantly, in this study, we show that although all advice was provided digitally without any face-to-face contact, personalized nutritional advice still led to a significant loss of body weight, whereas generic nutritional advice did not produce a statistically significant effect on body weight or other anthropometric measurements (Figure 5A). This finding is consistent with other studies comparing digitally provided generic versus personalized advice in populations with obesity [28-31]. Together, the findings of these studies and our study provide evidence for the added value of personalized nutritional advice compared with generic advice, even when delivered online. Yet, it remains a topic of discussion whether the same personalized nutritional advice would have been more effective if delivered in person, rather than online. Interestingly, one study compared the effect of delivering personalized nutritional advice online with or without a personal coach [32]. Participants receiving similar advice through an online coach demonstrated greater engagement with the program and had an increased likelihood of achieving clinically relevant weight loss. This finding suggests that receiving feedback from an actual human can enhance protocol adherence, possibly due to an increased sense of accountability, motivation, and perception of receiving care tailored to an individual's specific needs. Similarly, another study found that offline nutritional advice was more effective than online advice in a small sample of rugby athletes [25], highlighting the potential added benefit of in-person delivery [33]. However, conventional face-to-face counseling also has its downsides, as it is more expensive, less flexible in terms of time and location, and ultimately less scalable compared with online methods [34]. Consequently, both online and in-person approaches have their own strengths, and one may be more appropriate than the other depending on an individual's personal preferences and needs.

Interestingly, participants in the PP-group, who received personalized food boxes based on their personalized advice, were able to lose more body weight compared with the other two groups. This effect was not due to differences in physical activity, as PAL did not change in any of the groups (Table 3). This outcome may reflect the difficulty of translating personalized nutritional advice into an actual diet over a prolonged period. Indeed, these participants reported greater ease in following the personalized nutritional advice, as well as a greater perceived dietary change, compared with participants in the control and P-groups (Figure 4A and 4C). This novel finding suggests that future interventions may benefit from placing particular emphasis on helping participants translate their advice into practice, for example, by simplifying meal planning and reducing decision fatigue. These data also revealed that participants in the P-group did not find it easier or more enjoyable to implement their personalized advice compared with participants in the control group receiving generic advice (Figure 4A and 4B). However, participants in the P-group did achieve a significant loss of body weight, which was not observed in the control group (Figure 5A). These findings suggest that the success of personalized nutritional advice is not necessarily due to greater ease or enjoyment of implementation, which would be expected to increase protocol adherence. Instead, a better fit and higher quality of the advice are likely the main contributors to the enhanced effectiveness of personalized nutritional advice.

Habitual food intake was also measured, which indicated that all intervention groups changed their dietary habits in a similar way. All groups reduced the intake of ready-made meals (113.6 g vs 78.5 g, –30.9%); sauces and gravy (18.8 g vs 10.0 g, –46.8%); sweet snacks (84.8 g vs 64.1 g, –24.4%); savory snacks (50.5 g vs 40.0 g, –20.1%); bread, pasta, rice, and wraps (nutritional quality score of 1.9 vs 1.7, –10.5%); and vegetables (129.0 g vs 118.7 g, –8.0%); and replaced coffee with tea (Table 4). These results indicate that the intake of unhealthy food groups was particularly reduced, while the intake of healthy food groups such as fruit, legumes, nuts, and fish was maintained. Only vegetable intake decreased, but to a lesser extent (129.0 g vs 118.7 g, –8.0%). Fiber intake was also maintained and did not change (Multimedia Appendix 5). This suggests that all 3 intervention groups shifted their dietary intake toward a healthier habitual pattern, in line with the nutritional advice provided. No significant time × group interaction effects were found in the habitual food intake data. This is noteworthy, as the diet of the PP-group changed substantially—they received personalized food boxes for all meals throughout the study—while the control group and P-group did not receive such food boxes. Based on this approach, we hypothesized that the PP-group would show greater changes in habitual food intake compared with the other groups; however, these changes were not detected. This suggests that participants in the PP-group may have had difficulty accurately recalling the foods they consumed. In contrast to the habitual food intake data, the anthropometric data revealed group-specific effects of the nutritional interventions. For these reasons, we suspect that the collected habitual food intake data were not accurate enough to detect group-specific effects. Therefore, future studies may benefit from incorporating new technology-based dietary assessment tools that combine web-based programs with mobile apps and wearable devices [35]. Interestingly, legume intake showed a divergent pattern between groups, with a decrease in the control group and an increase in both intervention groups. Although this difference did not reach conventional statistical

XSL•FO
RenderX

significance (interaction P=.051), the trend suggests a potential effect of the personalized interventions. Participants in the P- and PP-groups received targeted advice regarding protein and macronutrient intake, which may have encouraged higher legume consumption, whereas the control group received only general dietary guidance. Moreover, although the between-person variability in habitual food intake was considerable, this is not unexpected in nutritional intervention studies. Such variability likely reflects differences in individual dietary habits, adherence, and reporting accuracy. This observation emphasizes the importance of personalized dietary assessment and highlights the inherent heterogeneity in dietary behavior, even within relatively homogeneous study populations.

Sex differences in fat tissue distribution and fat loss across the waist and hip regions were observed in this fully remote study. Consistent with the literature, females had a lower waist-to-hip ratio compared with males (Figure 6A) [36]. Notably, body weight loss correlated more strongly with waist circumference loss in males (r=0.45) than in females (r=0.26, interaction P=.01; Figure 6B), which is also consistent with previous findings [37]. This sexual dimorphism is likely explained by the fact that males store more fat in the waist region compared with females, making the waist region more responsive to weight loss [37]. Together, these findings provide evidence for the accuracy of remotely collected data and demonstrate its potential to study sex differences in responses to nutritional interventions.

Strengths and Limitations

Strengths of the study included its 3-arm design, with a total of 122 participants who were randomized and completed the collection of real-world study data, providing an ecological sample. This design allowed for the evaluation of the effectiveness of the nutritional interventions, albeit in a setting that was less controlled compared with a classical nutritional intervention study. However, there were some limitations associated with this study. First, no comparison was made with nondigital nutritional interventions. Such a comparison would have been valuable, as it could reveal potential differences in the direct effect size of online versus offline personalized nutrition advice. Future studies could include a nondigital control group to address this. Second, the study duration was 6 weeks, reflecting only short-term effects. The long-term impact of the interventions remains unknown, and future research with longer follow-up is needed to assess sustained effects. Third, due to dropouts, the PP-group included a relatively higher proportion of males (Table 1). This likely resulted in a slightly higher baseline body weight in the PP-group compared with the other groups, although the difference was not statistically significant. Randomization and statistical adjustments were applied to mitigate this imbalance, but future studies could aim for larger sample sizes to avoid similar discrepancies. Fourth, although this study demonstrated short-term beneficial effects of personalized nutritional interventions, these results do not guarantee long-term benefits, such as sustained weight loss. Previous studies have shown that maintaining weight loss is challenging, with many individuals regaining weight within 2-5 years [38,39]. In addition, long-term success is influenced not only by individual behavior change but also by broader factors, such as the food environment and supportive public health policies [40], which were beyond the scope of this study. Future research should therefore investigate whether personalized digital interventions can contribute to sustainable behavior change when combined with strategies addressing environmental and policy-level determinants. Fifth, a potential bias in studies relying on self-reported outcomes is participants' tendency to report behaviors that align with perceived study goals, such as weight loss. To minimize this potential bias, we provided standardized instructions for anthropometric measurements, used validated dietary and activity questionnaires, and emphasized the importance of accurate reporting. The randomized design helped distribute any residual bias across study arms. In the food-provision arm, reporting bias may have been slightly higher due to participants' awareness of the intervention; however, randomization and the use of multiple anthropometric measures (body weight, waist, and hip circumference) enhanced the robustness of the results. Cross-validation analyses (Figure 3) and sex-specific effects (Figure 6), consistent with prior literature, further support the reliability of our findings. Sixth, a sample size of 150 at the start of the intervention was considered sufficient to demonstrate the effectiveness of the personalized nutritional interventions. However, due to a relatively high number of dropouts (n=43 due to noncompliance and an additional n=15 due to outlier exclusion), only 122 participants were included in the final data analysis. This reduction in sample size may have limited the statistical power to detect the intended effects. While a dropout rate of 10%-20% was anticipated, the actual dropout due to noncompliance was 23.9% (43/180 participants), with an additional 8.3% (15/180 participants) excluded as outliers. This relatively high attrition rate may be attributed to the remote digital study design without any face-to-face contact and may also reflect a relatively high burden of participation. Data from noncompliant participants were excluded from the analysis, and it is important to acknowledge that the exclusion of participants who did not complete the procedures may have led to an overestimation of feasibility outcomes. Moreover, the high number of outliers could reflect increased misreporting, possibly resulting from unclear instructions. In future studies, real-time plausibility checks could be helpful to flag implausible values and prompt participants to reenter their measurements. Additionally, digital-first studies should account for potentially higher dropout rates and provide clear participant instructions. Despite the attrition, the final sample remained balanced across the 3 study arms, with 39-43 participants per group, allowing for meaningful comparisons of intervention effects. Seventh, because 91 individuals were excluded before participation, the feasibility results reflect only those who entered and completed the study. This selective inclusion may limit generalizability and could lead to an overestimation of feasibility in the broader target population.

Future Directions

While personalized nutrition approaches hold promise in tailoring dietary advice to individual characteristics and improving adherence, it is important to recognize that individual choices are embedded within a broader socioecological context. Dietary behavior is shaped not only by biological and psychological factors but also by the surrounding food

XSL•FO
RenderX

environment, which includes the availability, affordability, accessibility, and cultural norms of food consumption. A growing body of public health literature emphasizes the importance of considering food systems and food environments when designing interventions, as these structural factors often constrain or facilitate the adoption of individualized recommendations [12]. Furthermore, diets are a key driver of environmental change, linking nutrition with planetary health challenges, such as climate change, biodiversity loss, and unsustainable land and water use [41]. Integrating personalized approaches with strategies that improve local food environments and align with sustainable food system goals may therefore offer a more comprehensive and equitable pathway for improving dietary behavior and long-term health outcomes. Moreover, factors such as socioeconomic status, language barriers, and digital literacy can influence both access to and adherence with personalized digital nutrition advice. For example, populations living in socioeconomically deprived neighborhoods often face cumulative disadvantages, including unhealthy food environments, financial constraints, and limited digital access, which can undermine the effectiveness of digital health tools [42]. These considerations suggest that personalized nutrition approaches are likely to be most effective when coupled with efforts to reduce structural barriers and enhance inclusivity, such as tailoring content to different languages, improving accessibility for individuals with lower digital skills, and embedding interventions within supportive community and policy contexts.

Conclusions

We conclude that it is feasible to conduct a fully remote, fully digital nutritional intervention study. Participants were able to independently perform anthropometric measurements at home, reported positive user experiences, and generated self-collected data of sufficient internal validity to detect meaningful changes over time. In addition, the study demonstrated that personalized nutritional advice led to greater weight loss than generic advice, even in the absence of face-to-face contact. The provision of personalized food boxes further facilitated the translation of advice into daily dietary behavior, resulting in the largest reductions in body weight and hip circumference. Taken together, our findings indicate that fully online nutritional intervention studies can be successfully implemented and offer a scalable approach for reaching broader populations while still producing reliable and actionable outcomes.

Acknowledgments

We thank all the participants who participated in this study.

Data Availability

Data are available from the corresponding author (SW) upon reasonable request.

Authors' Contributions

Conceptualization: SW
Data analysis: JCBCdJ
Data curation: MPMC, RD
Data interpretation: FPMH, WJP, LGPP, JCBCdJ, SW
Investigation: CMMB, FPMH, WJP
Methodology: FPMH, SW, WJP
Visualization: JCBCdJ, SW. All authors have read and agreed to the published version of the manuscript.
Writing – original draft: JCBCdJ
Writing– review & editing: all authors

Conflicts of Interest

This study was funded by MixMasters B.V., and Uitgekookt Meal Service provided the personalized food boxes to the participants. Study design, data collection, analyses, interpretation, and manuscript writing were conducted independently by the authors of this paper.

Multimedia Appendix 1
Questionnaires used in this study to (1) screen participants for inclusion and exclusion criteria, (2) collect baseline data, and (3) obtain feedback on the online intervention.
[XLSX File (Microsoft Excel File), 30 KB - jmir_v28i1e73367_app1.xlsx]

Multimedia Appendix 2
Examples of participant feedback, along with the generic and personalized nutritional advice provided.
[DOCX File , 40 KB - jmir_v28i1e73367_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 10789 KB - jmir_v28i1e73367_app3.pdf]

Multimedia Appendix 4
Overview of the calculation of nutritional quality scores.
[DOCX File , 31 KB - jmir_v28i1e73367_app4.docx]

Multimedia Appendix 5
Overall fiber intake during the study.
[DOCX File , 30 KB - jmir_v28i1e73367_app5.docx]

References

1. de Jong AJ, van Rijssel TI, Zuidgeest MGP, van Thiel GJMW, Askin S, Fons-Martínez J, Trials@Home Consortium. Opportunities and challenges for decentralized clinical trials: European regulators' perspective. Clin Pharmacol Ther 2022 Aug 17;112(2):344-352 [FREE Full text] [doi: 10.1002/cpt.2628] [Medline: 35488483]
2. Gul RB, Ali PA. Clinical trials: the challenge of recruitment and retention of participants. Journal of Clinical Nursing 2009 Dec 17;19(1-2):227-233. [doi: 10.1111/j.1365-2702.2009.03041.x]
3. Carlisle B, Kimmelman J, Ramsay T, MacKinnon N. Unsuccessful trial accrual and human subjects protections: An empirical analysis of recently closed trials. Clinical Trials 2014 Dec 04;12(1):77-83. [doi: 10.1177/1740774514558307] [Medline: 25475878]
4. Amstutz A, Schandelmaier S, Frei R, Surina J, Agarwal A, Olu KK, et al. Discontinuation and non-publication of randomised clinical trials supported by the main public funding body in Switzerland: a retrospective cohort study. BMJ Open 2017 Aug 01;7(7):e016216. [doi: 10.1136/bmjopen-2017-016216]
5. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. Contemporary Clinical Trials Communications 2018 Sep;11:156-164. [doi: 10.1016/j.conctc.2018.08.001]
6. Goodson N, Wicks P, Morgan J, Hashem L, Callinan S, Reites J. Opportunities and counterintuitive challenges for decentralized clinical trials to broaden participant inclusion. npj Digit. Med 2022 May 05;5(1):58. [doi: 10.1038/s41746-022-00603-y]
7. Pasman WJ, Hendriks HF, Minekus MM, de Ligt RA, Scholtes-Timmerman MJ, Clabbers ND, et al. Subjective feelings of appetite of wholegrain breakfasts evaluated under controlled, laboratory and ‘at home’ conditions. Physiology & Behavior 2018 Oct;194:285-291. [doi: 10.1016/j.physbeh.2018.06.024]
8. Pasman WJ, Boessen R, Donner Y, Clabbers N, Boorsma A. Effect of Caffeine on Attention and Alertness Measured in a Home-Setting, Using Web-Based Cognition Tests. JMIR Res Protoc 2017 Sep 07;6(9):e169. [doi: 10.2196/resprot.6727]
9. Dorhout BG, Wezenbeek N, de Groot LCPGM, Grootswagers P. Web-Based Exercise and Nutrition Intervention to Improve Leg Muscle Strength and Physical Functioning in Older Adults: Pre-Post Pilot Study. JMIR Form Res 2025 Jan 30;9:e54392-e54392. [doi: 10.2196/54392]
10. Jinnette R, Narita A, Manning B, McNaughton SA, Mathers JC, Livingstone KM. Does Personalized Nutrition Advice Improve Dietary Intake in Healthy Adults? A Systematic Review of Randomized Controlled Trials. Advances in Nutrition 2021 May;12(3):657-669. [doi: 10.1093/advances/nmaa144]
11. Livingstone KM, Love P, Mathers JC, Kirkpatrick SI, Olstad DL. Cultural adaptations and tailoring of public health nutrition interventions in Indigenous peoples and ethnic minority groups: opportunities for personalised and precision nutrition. Proc. Nutr. Soc 2023 Jun 19;82(4):478-486. [doi: 10.1017/s002966512300304x]
12. Fanzo J, Bellows AL, Spiker ML, Thorne-Lyman AL, Bloem MW. The importance of food systems and the environment for nutrition. The American Journal of Clinical Nutrition 2021 Jan;113(1):7-16. [doi: 10.1093/ajcn/nqaa313]
13. Doets EL, de Hoogh IM, Holthuysen N, Wopereis S, Verain MC, van den Puttelaar J, et al. Beneficial effect of personalized lifestyle advice compared to generic advice on wellbeing among Dutch seniors - An explorative study. Physiol Behav 2019 Oct 15;210:112642 [FREE Full text] [doi: 10.1016/j.physbeh.2019.112642] [Medline: 31394106]
14. Lau Y, Wong SH, Chee DGH, Ng BSP, Ang WW, Han CY, et al. Technology‐delivered personalized nutrition intervention on dietary outcomes among adults with overweight and obesity: A systematic review, meta‐analysis, and meta‐regression. Obesity Reviews 2024 Jan 31;25(5):e13699. [doi: 10.1111/obr.13699]
15. Kan J, Ni J, Xue K, Wang F, Zheng J, Cheng J, et al. Personalized Nutrition Intervention Improves Health Status in Overweight/Obese Chinese Adults: A Randomized Controlled Trial. Front. Nutr 2022 Jun 22;9:919882. [doi: 10.3389/fnut.2022.919882]
16. Al-Awadhi B, Fallaize R, Zenun Franco R, Hwang F, Lovegrove JA. Insights Into the Delivery of Personalized Nutrition: Evidence From Face-To-Face and Web-Based Dietary Interventions. Front. Nutr 2021 Jan 27;7:570531. [doi: 10.3389/fnut.2020.570531]

XSL•FO
RenderX

17. Griffiths F, Lindenmeyer A, Powell J, Lowe P, Thorogood M. Why Are Health Care Interventions Delivered Over the Internet? A Systematic Review of the Published Literature. J Med Internet Res 2006 Jun 23;8(2):e498. [doi: 10.2196/jmir.8.2.e10]
18. de Hoogh IM, Winters BL, Nieman KM, Bijlsma S, Krone T, van den Broek TJ, et al. A Novel Personalized Systems Nutrition Program Improves Dietary Patterns, Lifestyle Behaviors and Health-Related Outcomes: Results from the Habit Study. Nutrients 2021 May 22;13(6):1763. [doi: 10.3390/nu13061763]
19. van der Haar S, Hoevenaars FPM, van den Brink WJ, van den Broek T, Timmer M, Boorsma A, et al. Exploring the Potential of Personalized Dietary Advice for Health Improvement in Motivated Individuals With Premetabolic Syndrome: Pretest-Posttest Study. JMIR Form Res 2021 Jun 24;5(6):e25043. [doi: 10.2196/25043]
20. Norstat. URL: https://www.norstat.co [accessed 2025-12-22]
21. ten Haaf T, Weijs PJM. Resting Energy Expenditure Prediction in Recreational Athletes of 18–35 Years: Confirmation of Cunningham Equation and an Improved Weight-Based Alternative. PLoS ONE 2014 Oct 2;9(10):e108460. [doi: 10.1371/journal.pone.0108460]
22. Mijn Eetmeter. URL: https://mijn.voedingscentrum.nl/nl/eetmeter/ [accessed 2025-12-22]
23. van den Brink WJ, van den Broek TJ, Palmisano S, Wopereis S, de Hoogh IM. Digital Biomarkers for Personalized Nutrition: Predicting Meal Moments and Interstitial Glucose with Non-Invasive, Wearable Technologies. Nutrients 2022 Oct 24;14(21):4465. [doi: 10.3390/nu14214465]
24. Healey G, Brough L, Murphy R, Hedderley D, Butts C, Coad J. Validity and Reproducibility of a Habitual Dietary Fibre Intake Short Food Frequency Questionnaire. Nutrients 2016 Sep 10;8(9):558. [doi: 10.3390/nu8090558]
25. FAO/WHO/UNU Expert Consultation. Human energy requirements: report of a joint FAO/WHO/UNU Expert Consultation. In: FAO Food and Nutrition Technical Report Series. Rome: Food and Agriculture Organization of the United Nations; 2004:1-96.
26. World Health Organization. Waist circumference and waist-hip ratio: report of a WHO expert consultation. In: World Health Organization meeting report. Geneva: World Health Organization; 2011:1-39.
27. Ryan K, Dockray S, Linehan C. A systematic review of tailored eHealth interventions for weight loss. DIGITAL HEALTH 2019 Feb 05;5:e1. [doi: 10.1177/2055207619826685]
28. Cheng J, Costacou T, Sereika SM, Conroy MB, Parmanto B, Rockette-Wagner B, et al. Effect of an mHealth weight loss intervention on Healthy Eating Index diet quality: the SMARTER randomised controlled trial. Br J Nutr 2023 Jun 07;130(11):2013-2021. [doi: 10.1017/s0007114523001137]
29. Ambeba E, Ye L, Sereika S, Styn M, Acharya S, Sevick M, et al. The use of mHealth to deliver tailored messages reduces reported energy and fat intake. Journal of Cardiovascular Nursing 2015;30:35-43. [doi: 10.1097/jcn.0000000000000120]
30. Gonzalez-Ramirez M, Sanchez-Carrera R, Cejudo-Lopez A, Lozano-Navarrete M, Salamero Sánchez-Gabriel E, Torres-Bengoa MA, et al. Short-Term Pilot Study to Evaluate the Impact of Salbi Educa Nutrition App in Macronutrients Intake and Adherence to the Mediterranean Diet: Randomized Controlled Trial. Nutrients 2022 May 14;14(10):2061. [doi: 10.3390/nu14102061]
31. Kohl J, Brame J, Centner C, Wurst R, Fuchs R, Sehlbrede M, et al. Effects of a Web-Based Lifestyle Intervention on Weight Loss and Cardiometabolic Risk Factors in Adults With Overweight and Obesity: Randomized Controlled Clinical Trial. Journal of medical Internet research 2023;25:e43426. [doi: 10.2196/43426]
32. Beleigoli A, Andrade AQ, Diniz MDF, Ribeiro AL. Personalized Web-Based Weight Loss Behavior Change Program With and Without Dietitian Online Coaching for Adults With Overweight and Obesity: Randomized Controlled Trial. J Med Internet Res 2020 Nov 5;22(11):e17494. [doi: 10.2196/17494]
33. Junaidi, Apriyanto T, Laily I, Rizki P. The comparison of offline and online nutrition education on body mass index in rugby athletes during the Covid-19 Pandemic (The Body Mass Index profile of Jakarta athletes during Covid-19 Pandemic). Journal of Physical Education and Sport 2021 Aug 15:2295-2301.
34. Lau Y, Chee DGH, Chow XP, Cheng LJ, Wong SN. Personalised eHealth interventions in adults with overweight and obesity: A systematic review and meta-analysis of randomised controlled trials. Preventive Medicine 2020 Mar;132:106001. [doi: 10.1016/j.ypmed.2020.106001]
35. Eldridge AL, Piernas C, Illner A, Gibney MJ, Gurinović MA, De Vries JH, et al. Evaluation of New Technology-Based Tools for Dietary Intake Assessment—An ILSI Europe Dietary Intake and Exposure Task Force Evaluation. Nutrients 2018 Dec 28;11(1):55. [doi: 10.3390/nu11010055]
36. Karastergiou K, Smith SR, Greenberg AS, Fried SK. Sex differences in human adipose tissues – the biology of pear shape. Biol Sex Differ 2012 May 31;3(1):1-12. [doi: 10.1186/2042-6410-3-13]
37. Kuk JL, Ross R. Influence of sex on total and regional fat loss in overweight and obese men and women. Int J Obes 2009 Mar 10;33(6):629-634. [doi: 10.1038/ijo.2009.48]
38. Weiss EC, Galuska DA, Kettel Khan L, Gillespie C, Serdula MK. Weight Regain in U.S. Adults Who Experienced Substantial Weight Loss, 1999–2002. American Journal of Preventive Medicine 2007 Jul;33(1):34-40. [doi: 10.1016/j.amepre.2007.02.040]
39. Lowe MR, Kral TVE, Miller-Kovach K. Weight-loss maintenance 1, 2 and 5 years after successful completion of a weight-loss programme. Br J Nutr 2007 Nov 28;99(4):925-930. [doi: 10.1017/s0007114507862416]

XSL•FO
RenderX

40. Tewahade S, Berrigan D, Slotman B, Stinchcomb DG, Sayer RD, Catenacci VA, et al. Impact of the built, social, and food environment on long‐term weight loss within a behavioral weight loss intervention. Obesity Science & Practice 2022 Nov 03;9(3):261-273. [doi: 10.1002/osp4.645]
41. Willett W, Rockström J, Loken B, Springmann M, Lang T, Vermeulen S, et al. Food in the Anthropocene: the EAT–Lancet Commission on healthy diets from sustainable food systems. The Lancet 2019 Feb;393(10170):447-492. [doi: 10.1016/s0140-6736(18)31788-4]
42. Ter Ellen F, Oude Groeniger J, Stronks K, Hagenaars L, Kamphuis C, Mackenbach J, et al. Understanding the dynamics driving obesity in socioeconomically deprived urban neighbourhoods: an expert-based systems map. BMC Med 2025 Jan 07;23(1):2 [FREE Full text] [doi: 10.1186/s12916-024-03798-x] [Medline: 39762884]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
FAO: Food and Agriculture Organization
PAL: physical activity level
P-group: personalized intervention group
PP-group: personalized plus intervention group
TNO: Netherlands Organization for Applied Scientific Research
UNU: United Nations University
WHO: World Health Organization

Edited by N Cahill; submitted 04.Mar.2025; peer-reviewed by N Holliday, J Nielsen; comments to author 11.Sep.2025; revised version received 26.Nov.2025; accepted 26.Nov.2025; published 05.Jan.2026.

Please cite as:
de Jong JCBC, Hoevenaars FPM, Peters LGP, Berendsen CMM, Pasman WJ, Caspers MPM, Dulos R, Wopereis S
A Real-Life Digital Intervention for Personalized Nutrition in Adults With Overweight or Obesity: Remote Randomized Controlled Trial
J Med Internet Res 2026;28:e73367
URL: https://www.jmir.org/2026/1/e73367
doi:10.2196/73367
PMID:

©Jelle CBC de Jong, Femke PM Hoevenaars, Lotte GP Peters, Charlotte MM Berendsen, Wilrike J Pasman, Martien PM Caspers, Remon Dulos, Suzan Wopereis. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Web-Based Cancer Prevention Intervention for Rural Emerging Adults: Mixed Methods Development and Pilot-Testing Study

Echo L Warner[1], MPH, PhD; Alishia Kinsey[2], BS; Barbara J Walkosz[2], PhD; Julia Berteletti[2], MSW; Kayla Nuss[2], PhD; Annelise Small[2], BA; W Gill Woodall[2], PhD; Deanna Kepka[3], MPH, PhD; Douglas Taren[4], PhD; Meghan B Skiba[5], PhD; Dolores D Guest[6], PhD; Cindy K Blair[6], PhD; Judith S Gordon[5], PhD; David W Wetter[7], PhD; Evelinn A Borrayo[8], PhD; Kimberly L Henry[9], PhD; Andrew L Sussman[6*], PhD; David B Buller[2*], PhD

[1]College of Nursing, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, United States

[2]Klein Buendel (United States), Golden, United States

[3]Circle of Hope, College of Nursing; Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, United States

[4]Department of Pediatrics Nutrition Center, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, United States

[5]College of Nursing, University of Arizona, Tucson, AZ, United States

[6]Comprehensive Cancer Care Center, University of New Mexico, Albuquerque, United States

[7]Department of Population Health Sciences, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, United States

[8]University of Colorado Cancer Center, Aurora, United States

[9]Department of Psychology, Colorado State University, Fort Collins, CO, United States

[*]these authors contributed equally

Corresponding Author:
Echo L Warner, MPH, PhD
College of Nursing, Huntsman Cancer Institute
University of Utah
10 S 2000 E
Salt Lake City, UT, 84112
United States
Phone: 1 8012447040
Email: echo.warner@nurs.utah.edu

Abstract

Background: The rapid growth of user-generated web-based health information increases the complexity of cancer information seeking. One promising strategy for promoting high-quality cancer information consumption is through targeted interventions that are intentionally designed to reach individuals in the web-based spaces they occupy. However, there is a paucity of evidence-based information on the best strategies for designing and implementing web-based health behavior change interventions to improve individuals' cancer-related knowledge and prevent cancer.

Objective: This study aimed to develop and pilot test a theory-based intervention via the web to reduce 6 cancer risk factors among rural emerging adults (EAs) through community-engaged research.

Methods: This mixed methods evaluation describes the development of a web-based cancer prevention intervention aimed at rural EAs aged 18-26 years in the United States and delivered in Facebook private groups. The intervention was guided by behavior change theory and cocreated with EA and Stakeholder Organization Advisory Boards to ensure relevance, accessibility, and appropriateness. We report on 3 formative surveys, a pilot intervention, protocol development, and the community-engaged process for intervention development. Descriptive statistics were applied to the surveys and pilot intervention baseline results to produce means and SDs using R.

Results: We developed posts (n=400) for a Facebook feed aimed at reducing 6 cancer risk behaviors (unhealthy diet, lack of physical activity, tobacco use, alcohol use, sun exposure, and human papillomavirus infection) with iterative input from the EA and stakeholder advisory boards. Formative surveys with rural EAs (n=297) and a pilot study of the intervention with this population (n=26) were conducted. In the pilot study, the intervention reached participants across rural counties, with sustained engagement (post views=1060, reactions=346, comments=72) over a one-month period. Key modifications to the intervention content and design emerged from both advisory boards, the formative surveys, and the pilot intervention, focusing on using perceived reliable sources and direct links to source material.

XSL•FO
RenderX

Conclusions: This web-based cancer prevention intervention is scalable and delivers engaging, evidence-informed health information to rural EAs. We offer key insights into the design and implementation of web-based cancer prevention interventions for EAs by describing the resources, timelines, and expertise needed to design and implement the intervention. Considerations for fully engaging EA and community stakeholder partners are presented, and we discuss how their involvement resulted in modifications that strengthened the intervention. Finally, we highlight the importance of theory-based health-behavior messaging, digital messaging skillsets, and platform-tailored dissemination strategies for maximizing web-based intervention acceptability.

Trial Registration: ClinicalTrials.gov NCT05618158; https://classic.clinicaltrials.gov/ct2/show/NCT05618158

International Registered Report Identifier (IRRID): RR2-10.2196/50392

(J Med Internet Res 2026;28:e80803) doi:10.2196/80803

KEYWORDS

emerging adult; web-based; intervention development; cancer; prevention; rural; community engagement

Introduction

Emerging adults (EAs) in rural and remote areas of the United States face unique challenges that increase their vulnerability to cancer. Emerging adulthood, a critical life stage from ages 18 to 26 years, is marked by increased autonomy and transitions in financial, residential, and employment responsibilities [1]. However, many EAs establish unhealthy lifestyle patterns during this period, including reduced physical activity, poor dietary habits, nicotine and tobacco use, binge drinking, sporadic sun safety practices, and skipping human papillomavirus (HPV) vaccination [1,2]. Living in a rural setting can heighten feelings of isolation and limit access to resources that support healthy behaviors [3]. These modifiable risk factors contribute to premature cancer morbidity and mortality, making rural EAs a priority population for cancer prevention efforts.

Given its popularity among EAs [4], using a social web-based intervention offers a promising avenue for addressing cancer prevention among EAs in rural areas. Over 80% of individuals in this age group access the internet several times per day [4]. Despite known flaws in information quality, web-based interventions that are strategically and theoretically designed are promising strategies for providing high-quality health information from trusted voices [5]. Web-based educational interventions can also disseminate timely and relevant public health messages, leverage user-generated content to personalize information, engage audiences in 2-way communication, and be used to detect and respond to emerging trends [6,7]. The success of web-based interventions hinges on their ability to address the broader social determinants of health, community and cultural perceptions, and built environments in which individuals live. However, web-based media can also perpetuate harmful and misleading health information [8-11], and interventions that relate accurate and truthful cancer prevention strategies are needed [12-16]. Studies of web-based interventions have often lacked theoretical grounding, rigorous design and evaluation procedures, and guidelines on the development of content, limiting the rigor and reproducibility of this intervention approach.

Rural communities often face unique barriers to engaging in healthy lifestyles, such as limited health care access, socioeconomic challenges, and geographic isolation [17-19]. EAs have limited interaction with traditional community channels like schools, workplaces, and health care settings, so web-based interventions may provide unique opportunities for tailored, real-time engagement with EAs. Use of community engagement strategies in the development of both the content and structure of social media interventions can increase relevance to the specific needs of rural EAs [20,21]. A community-engaged approach to web-based intervention design should increase the relevance of interventions to the specific needs of rural EAs.

In this context, we developed a theory-based, web intervention to reduce 6 cancer risk factors common among rural EAs, including physical inactivity, unhealthy diets, nicotine/tobacco use, binge drinking, unprotected ultraviolet exposure, and preventing HPV infection. It was designed and pilot tested using community-engaged methods. Herein, we describe the development, pilot-testing, and refinement of our intervention prior to its launch in a randomized quasi-experimental trial. This developmental research project involved community-engaged research methods with a dynamic group of EAs and community stakeholders, a theory-informed process for creating content tailored to rural EAs, and formative research to refine and pilot test the intervention.

Methods

Setting and Study Design

The overall aim of the study was to create a web-based intervention, named PEAK Wellness Chat, and evaluate its effectiveness in a sample of EAs in a stepped wedge randomized quasi-experimental trial design (NCT05618158) [22]. The private-group function in Facebook was the platform for intervention delivery. Facebook is used by a large number of American adults, regardless of race/ethnicity, including EAs (67% of adults use it and 70% of rural adults) [23]. The private-group function cultivates the privacy of group members, which is essential to control experimental exposure to the intervention and avoid experimental contamination. Facebook also allows a variety of content delivery by length and type (eg, images and videos as well as links to other sites and sharing of Facebook posts from other organizations) and allows posts to remain in the group in perpetuity. Facebook also records engagement of each participant with posts (ie, reactions and comments) and tracks retention via group membership, to enable testing of intervention dose effects. Finally, the Facebook

XSL•FO
RenderX

algorithm promotes private-group posts in participants' feeds in terms of frequency and prominence when they engage more with group posts.

Ethical Considerations

This study was approved by the WCG Institutional Review Board (IRB20223673 and IRB20223673). Informed consent was obtained and documented electronically prior to enrollment in the formative surveys and pilot study described below. Privacy and confidentiality were maintained by storing all research subjects' data in password-protected and encrypted drives. No identifying information is presented herein. Participants were provided with a modest incentive for completing the formative survey (US $40) or pilot study (US $50).

Community-Engaged Strategy for Developing PEAK Wellness Chat

A library of Facebook posts for the PEAK Wellness Chat intervention was created by a core team of investigators and staff with expertise in cancer prevention, public health promotion, health communication, and emerging adulthood. The library was iteratively developed with structured community input from an Emerging Adult Advisory Board (EAAB) and Stakeholder Organization Advisory Board (SOAB), responses by EAs to formative surveys, and guidance by experts in the 6 cancer risk factors. Intervention management by the study's administrative team and engagement by EAs were assessed in a pilot study (Figure 1). The administrative team supported the overall study procedures, meeting facilitation, coordination and deployment of surveys, recruitment efforts, and software acquisition for post illustrations. Additionally, the administrative team produced and maintained a cancer information audit and a resource website. The cancer information audit was updated weekly with emergent cancer information and news about health and wellness, alongside any emerging internet trends and current events that should be incorporated into posts to promote relevance and relatability of post content. The resource website contained additional information about the cancer risk factors, community and national resources (eg, quit smoking hotline and food banks), and links to health education resources.

Figure 1. Community-engaged strategy for designing a Facebook post library. EA: emerging adult.

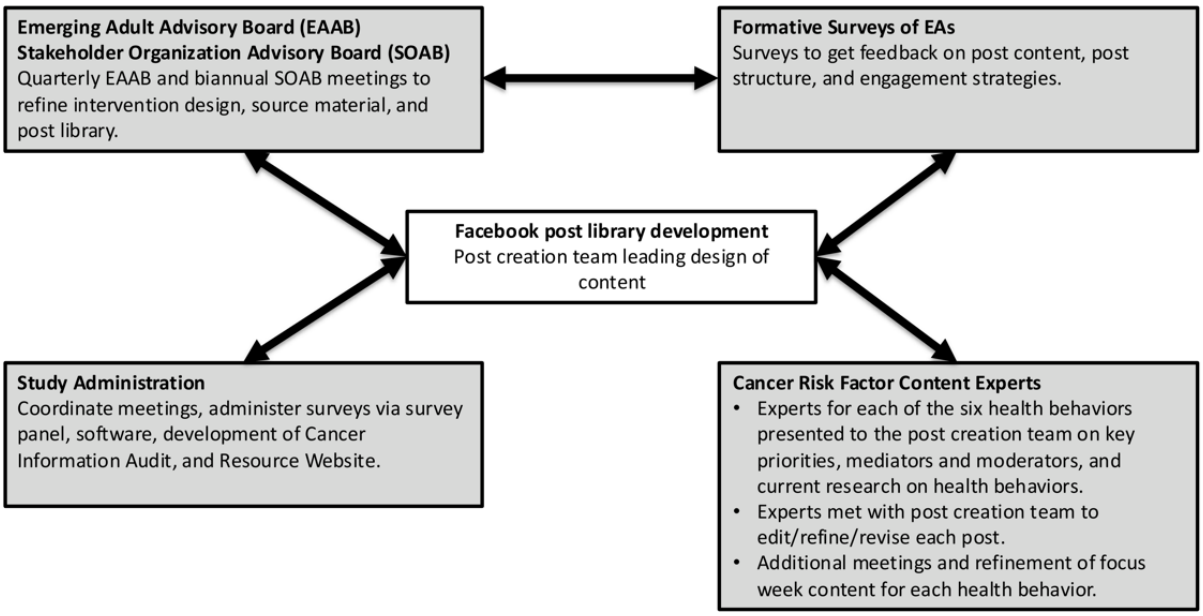

EAABs and SOABs

We used partnership processes aligned with the community-based participatory research model of Wallerstein et al [20] as a framework for the participatory approach. Through regular advisory board meetings, we addressed: (1) knowledge of rural "Contexts" that inform catchment area needs; (2) culturally informed "Partnership Processes" that facilitated a regular and iterative feedback cycle between the advisory boards, content experts, and post creation committee to refine the design features, language, and content of the Facebook posts; (3) development of responsive "Intervention & Research" protocols for rural conditions; and (4) participatory "Outcomes" that were iteratively disseminated to community partners through the intervention development and advisory board meetings (and are continuing during the intervention implementation in the trial).

EAAB and SOAB Composition

A total of 15 EAs, representing the composition of participating rural communities, agreed to serve on the EAAB. EAAB members were recruited through word-of-mouth referrals from the research teams' community networks. The EAAB held quarterly meetings to give input on the lived experiences of EAs, review the web-based intervention plans and protocols (eg, engagement strategies in posts), consider appropriateness, relevance, and engagement of proposed posts, and advise on recruiting EAs for the trial. Project investigators and staff at each participating Cancer Center also led direct engagement efforts with stakeholder organizations that serve rural EAs to convene members of the SOAB, which had 14 members. The

XSL•FO
RenderX

SOAB complemented the EAAB by identifying local resources for cancer risk reduction, reviewing intervention plans, protocols, and posts, and supporting recruitment and dissemination efforts in biannual meetings. Topics for discussion in the EAAB and SOAB meetings were jointly decided based on project needs, PEAK Wellness Chat content, and priorities of the board members.

Creation of the PEAK Wellness Post Library

The library of Facebook posts contained messages educating EAs about strategies to improve the 6 cancer risk behaviors. Messages were also created to increase EAs' media education about cancer and cancer prevention. Finally, posts were produced to enhance EAs' skills for communicating with family and friends because they reported: (1) living in a variety of arrangements, (2) relying on family and friends for acquiring and preparing food and paying for health care, and (3) living with others who engaged in the cancer risk behaviors. A focus week of up to 8 posts for each cancer risk factor was designed to emphasize messages on one important principle for improving the cancer risk factor. A focus week of posts on each of the 6 risk factors is being delivered during each 3-month period during the intervention.

Post Development Process

Each post was designed to address key features of health behavior change theories (self-determination theory, social cognitive theory, and diffusion of innovation theory [24-26]) to ensure the posts are conceptually driven to impact EAs' health behaviors. Specific principles incorporated into the posts included: intrinsic motivation, extrinsic motivation, social support, relatedness, personal goals, ability to control, self-efficacy, response efficacy, injunctive norms, descriptive norms, cancer risk perceptions, response cost, compatibility with values, observable benefits, autonomy, and simplicity. Table S1 in Multimedia Appendix 1 provides examples of posts incorporating each principle. Principles from self-determination theory, social cognitive theory, and diffusion of innovation were operationalized using messaging design and engagement strategies to ensure messages were not only informative but also encouraged motivation and engagement. Message design strategies included polls, narratives, calls for sharing, testimonials from EAs, source credibility, behavioral skills, and referrals to community resources. Engagement techniques included content from near peers, invitations to comment, cultural barriers and facilitators, images of ethnically diverse EAs, stories, videos or other visuals, question and answer formatting, and behavioral change techniques (ie, self-improvement and freedom to act), and high-interest, useful, and current event content. The framework also included message strategies to promote engagement (eg, visuals, current event content, cultural barriers/facilitators, polls, requests for comments, story requests, knowledge test, picture request). These features will be tracked and evaluated to ensure a variety of engagement strategies and best practices for health behavior change are integrated throughout the campaign. The post library was created and maintained in Microsoft Excel, where each post was coded for key theoretical concepts and engagement features (Table 1).

For this phase of intervention development, the EAAB and SOAB provided timely feedback on the content and design of posts, novel engagement strategies (ie, videos, use of polls), trustworthy and credible source material, and incorporation of local resources where possible. Along with the post library, the authors developed protocols for disseminating the intervention, moderating the Facebook group, and verifying the legitimacy of a participant's Facebook account. These included procedures for scheduling posts, responding to questions, and auditing the cancer prevention communication environment to identify emergent topics that should be included in the intervention to make messages relatable, relevant, and engaging. For example, the protocols included specific guidance for the moderator. Per the protocol, the moderator reacted (ie, like, love, and care) to every participant's comment left on posts. If a participant's comment resonated with the moderator, they would respond with their own comment to foster an authentic sense of community. Moderators will also respond to any direct Facebook messages participants send regarding the study (ie, questions about certain topics the participant does not want to share publicly). The protocol also includes boundaries for the moderator to model feasible and replicable interactions with the groups (eg, not to like comments that are misinformation).

XSL•FO
RenderX

Table 1. PEAK wellness post library.

Post features	Definition	Example
Post #	Indicates the order of the post and record number	1, 2, and 3
Date and time posted	Calendar date, time posted	26 February, 2025 9:00 AM
Day of week	Differentiate weekdays from weekends	Wednesday and Saturday
Topic	Identifies one of the 6 cancer prevention risk factors, media literacy, or family communication	Sun safety, HPV[a] prevention strategies, healthy diet
Message	Main content shared in the post	Knowing how to correctly use sunscreen to protect your skin is important. Here are some general guidelines.
Link	Source content link to be shared in the comments with viewers	https://www.cdc.gov/tobacco/campaign/tips/quit-smoking/index.html
Post visual suggestions	Indicates the type of visual content that should be included in the post	Team created/external content creator, TikTok video, YouTube video, infographic, or photo
Facebook post link	Link to the post on Facebook for tracking and review purposes	URL
Gender presentation	Indicate the visual gender presentation of people in the posts	Woman, man, N/A[b] if none
Character features	Indicate the body type, presenting race/ethnicity, and conventionally attractive features	Athletic build, Asian, conventionally attractive
Colors used in graphics	2-3 main colors	Red, blue, green
Engagement post	Indicator of whether or not the post is designed to promote engagement	Poll, "Tell us in the comments," share a story, knowledge test, picture request
Primary/secondary outcome addressed	Health behavior addressed with the post, also includes key topics addressed in the post	Cancer risk, autonomy, cooking skills, increased physical activity, preventing drunk driving, vaping knowledge, winter sun protection
Theoretic mediators	Key behavior mediators designed to promote behavior and health belief change	Intrinsic motivation, extrinsic motivation, social support, relatedness, personal goals, ability to control, self-efficacy, response efficacy, injunctive norms, descriptive norms, cancer risk perceptions, response cost, compatibility with values, observable benefits, simplicity
Message design features	Key design features based on health communication practices	Moderator instructions, referral to community resources, testimonials from EAs[c]
Engagement techniques	Strategies used in the post to promote engagement	Visuals, current event content, cultural barriers/facilitators, poll, "Tell us in the comments," share a story, knowledge test, picture request

[a]HPV: human papillomavirus.

[b]N/A: not applicable.

[c]EA: emerging adult.

Formative Surveys With Rural EAs

Overview

During the development of the post library, we administered 3 web-based surveys with rural EAs to evaluate initial reactions to the posts (Table 2). The surveys also collected information on EAs lived experiences, use of Facebook and other web-based channels, and perceived credibility of potential information sources used in the posts.

XSL•FO
RenderX

Table 2. PEAK Wellness Chat intervention development surveys.

Survey name	Purpose	Sample size (N)	Fielded dates
Formative survey 1	Health topics of interest, popular influencers or content creators, and credibility of health information sources.	100	February 2024
Formative survey 2	Discussing health behaviors with others, cancer risk behaviors by household, Facebook group name, and dietary experiences	100	July 2024
Formative survey 3	EAs'[a] engagement in physically active jobs and recreational activity, access to and perceptions of healthy and affordable food	97	January 2025
Pilot pretest survey	Baseline test of health behaviors, mediators, other covariates, and demographics	26	March 2024
Pilot posttest survey	Baseline test of health behaviors, mediators, other covariates, and demographics	23	April 2024

[a]EA: emerging adult.

Participants

Eligible individuals were ages 18-26 years, living in counties designated as Rural Urban Continuum Codes (RUCC) 4-9 [27] in Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming, and Rural Urban Commuting Area Codes 4-10, which classifies rurality by census tract [28] in Arizona where RUCC codes excluded several key rural communities. To facilitate adequate recruitment for the full trial, we expanded our recruitment to states beyond the original 4 corners states, and this modification was reflected in the formative survey participants. They also needed to be able to access a web-based survey and to use Facebook once or more times per week. Participants were identified through a survey panel company, Dynata [29]. The overall survey response rate was 302/1080 (28.0%), and the overall survey completion rate was 297/302 (98.3%).

Measures and Analysis

Posts were purposefully selected for inclusion in the surveys, focusing on issues that arose during post development and discussions with the EAAB, SOAB, and content experts. For example, the credibility of sources cited in the posts was a concern, so we selected posts with local and national source material (eg, Centers for Disease Control, newspapers, and local health departments). Respondents were asked to (1) rate each post for appropriateness, relevance, believability, amount of text, and trustworthiness on 5-point Likert scales; (2) indicate their potential engagement with the post through reading, scrolling past, reacting to, commenting on, and clicking on links (no, yes maybe, yes definitely); and (3) suggest ways for improving the post (open ended).

Surveys also contained questions on EAs lived experiences that informed postdevelopment. In the first survey (February 2024), participants were asked about topics of interest, popular influencers/content creators, and the credibility of information sources. In the second survey (July 2024), they were reported on discussing health behaviors with parents, siblings, partners, and friends, cancer risk behaviors by household members (use of nicotine products, consumption of alcohol, and intentional suntanning), preferences for the Facebook group name, and dietary experiences including sources of food, dependence on others for food purchasing, and sources of free food. In the third survey (January 2025), questions inquired about EAs' engagement in physically active jobs and recreational physical activity outside work, access to healthy and affordable food, and perceptions about healthy eating. Formative surveys are available upon reasonable request from the authors.

Analysis of EA Survey Responses

Summary statistics of participants' sociodemographics, lived experiences, and post feedback were calculated for each survey, using RStudio (Posit Software, PBC), 2024. Two open-ended questions ("What would you comment on this post?" and "How would you improve this post?") were coded to identify emergent themes. All responses were categorized into an individual theme. From the Comment question, 134 responses were analyzed, and 53 responses were excluded due to being out of context, incomplete statements, or illegible responses. The 81 responses coded resulted in the following themes: message-specific reactions and affirmative comments. From the Improvement question, 506 responses were analyzed, and 73 responses were excluded due to being out of context, incomplete statements, or illegible responses. The remaining 433 responses were coded by 2 coders in the following themes: suggestions on image, dislike or wrong audience, more information requested, reinforce source/suggestions on source, suggestion on content, too much information, unclear message, nothing to change, unsure.

Survey and open-ended feedback were iteratively reviewed by the full study team to improve the content and format of the posts in the intervention feed and ensure that posts were responsive, timely, and engaging for rural EAs. Summary reports of formative survey results were discussed in EAAB and SOAB meetings to review thematic findings and explore contextual and social considerations.

PEAK Wellness Chat Pilot Study

Overview

A 4-week pilot test of the intervention was conducted with a sample of rural EAs to refine procedures for the full trial pertaining to recruitment, baseline and posttest surveys, retention, and intervention protocols. Data were also obtained to confirm that posts were engaging for EAs. The pretest survey is available as a Multimedia Appendix 2.

XSL•FO
RenderX

Participants

Participants were recruited by Verasight [30], a research services company that recruits participants for studies through a web-based panel and advertising. Eligible participants for the pilot study were ages 18-26 years, living in counties designated as RUCC 4-9 in Arizona, Colorado, New Mexico, and Utah, and able to access the web-based survey. Pregnant participants were excluded because they may depart from their normal dietary and activity levels and consumption of alcohol and nicotine due to the pregnancy. Participants were screened for having an existing Facebook account for at least 1 year with weekly activity, with 2 EAs deemed ineligible. In addition, 3 EAs were excluded because they did not friend the Group Moderator and could not be joined to the Facebook private group for the pilot feed.

Pilot Study Procedures

Participants were enrolled in a single group, pretest-posttest design. Initially, they completed a baseline survey in REDCap (Research Electronic Data Capture; Vanderbilt University). The intervention was presented in a Facebook private group, with 2 posts made per day. During a 7-day period in the month, 8 posts in a focus week on physical activity were posted. This focus week strategy was designed to provide an in-depth intervention on improving self- and response-efficacy, increasing perceived risk associated with not being physically active [31-34], and linking cancer prevention to personal physical activity goals. During the 4-week campaign, 62 posts were made in the private group feed across domains related to physical activity (n=13), diet (n=7), sun safety (n=7), tobacco cessation (n=6), alcohol reduction (n=7), HPV prevention (n=7), media education (n=3), and family communication (n=2). Media literacy and family communication content were more general and not specifically related to the 6 health behaviors. At the end of the 4-week intervention period, participants were invited to complete a posttest survey in REDCap.

Measures and Analysis

The pretest and posttest surveys included questions about the 6 cancer risk factors: physical activity (Global Physical Activity Questionnaire [35]), diet (dietary screener; meal behaviors, food insecurity) [36,37], alcohol intake (Alcohol Use Disorders Identification Test–Consumption [38]), nicotine product use (30-day and 7-day smoking or vaping, quit ladder) [39,40], HPV prevention (initiation and completion of multi-shot vaccine series) [41], and ultraviolet protection (use of personal sun protection practices and sunburn) [42-45]. In addition, questions assessed basic needs, self-efficacy for cancer risk-reduction behaviors, cancer information overload, digital media use, health insurance coverage, last routine check-up, personal and family cancer history, and demographics.

Behavioral and experiential engagement with posts in the Facebook private group feed was measured. Behavioral engagement was assessed in two ways: (1) project staff kept a log of posts published to the Facebook private group that tracked date and time posted, topic, and post visual (graphic, gif, image, link, poll, or video); and (2) project staff extracted Facebook metadata reported in the private group platform, which included obtaining the total number of views, likes (eg, like and sad), and comments (comments from participants, comments from moderator) [46]. Facebook's reporting function permitted likes and comments but not views to be associated with specific users. The posttest survey contained questions assessing participants' experiential engagement with the intervention (ie, frequency of reading posts and sharing of content with others). Given the very small sample, the analysis involved descriptive statistics. All analyses were conducted in RStudio (version 4.4.2; Posit Software, PBC). The STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist is provided in Multimedia Appendix 3.

Results

Facebook Post Library

We produced a library of 400 Facebook posts on reducing the 6 cancer risk behaviors, media education, and family communication, using a real-time agile process to ensure posts engaged EAs. The development process followed an iterative feedback loop for each new post created. Content experts identified the key topics and information to address in the posts and essential mediators of behavior change for the cancer risk behaviors. The content from these presentations was adapted by a group of investigators and project staff into posts following the theoretical framework shown in Table 1. Content experts reviewed and approved the posts for inclusion in the final PEAK Wellness library, suggesting revisions that were incorporated to finalize the posts. Selected posts were pretested by having them reviewed by both the EAAB and SOAB, and included in the formative surveys for feedback as needed before finalization. Figure 2 illustrates the evolution of one post (about binge drinking) that was created using a photo visual and a local news source.

XSL•FO
RenderX

Figure 2. Evolution of a post through an iterative community-engaged development process. EA: emerging adult; EAAB: Emerging Adult Advisory Board; SOAB: Stakeholder Organization Advisory Board.

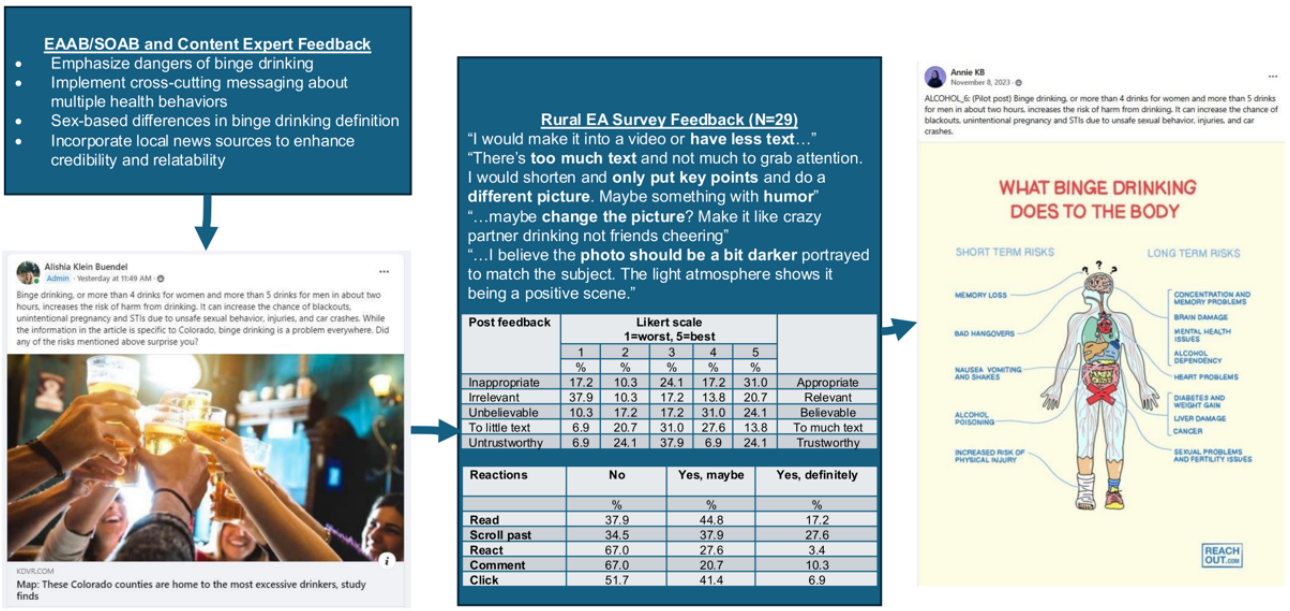

Formative Rural EA Surveys

Formative surveys with rural EAs provided insights about developing, revising, and finalizing posts, health behavior topics of importance, and contextual information to include to improve the posts. The 3 formative surveys with EAs obtained responses from 99, 98, and 100 EAs, for a total sample of 297 rural EA respondents. As shown in Table 3, the average age of respondents was 22.5 (SD 2.49) years. The sample of participants was primarily identified as 72.4% (n=215) female, 24.2% (n=72) Hispanic, 70.7% (n=210) White, 54.5% (n=54) high school educated or less, 43.2% (n=82) unemployed, 46.2% (n=90) living with a spouse/partner, and 34.3% (n=99) living with other relatives besides parents. The majority, 63.6% (n=63), had health insurance, and many EAs involved others in their health decisions, primarily parents (56/99, 55.6%) and a spouse/partner (40/99, 40.4%). Respondents resided in 8 states, with subsamples ranging from Nevada (26/297, 8.8%) to Idaho (61/297, 20.5%). Across the 3 formative surveys, respondents provided feedback about 61 posts (survey 1=21 posts, survey 2=20 posts, survey 3=20 posts).

Content analysis of the 2 open-ended questions asking about how participants would engage with the posts and improve the posts resulted in 2 themes for post comments and 9 themes for improving posts (Table 4). Content analysis was applied by one research team member, who met with 2 other researchers to discuss definitions, interpretations, and the application of codes. Discrepancies were resolved through consensus, after which coding was finalized. In terms of comments, themes focused on respondents' initial reaction to the post content and affirmations or approvals of the post content, creator, or message. Most posts received overwhelmingly positive feedback, but there were suggestions for improvements on visual appeal and less text across many posts. Due to this feedback, the post campaign library was revised to incorporate more video and photo content and eliminate text-only posts. Source material was a primary area of feedback that also resulted in changes to the campaign library, to provide greater transparency in primary source material and emphasize trusted organizations (local newspapers had lower levels of trust, while national organizations were seen as more trustworthy). Another unique theme related to the relevancy of the content, with sizable numbers of rural EAs reporting that they did not drink alcohol and thus felt alcohol-related posts were irrelevant. This reinforced the decision to focus the intervention on all 6 cancer risk behaviors rather than just one, so participants would likely receive at least some posts relevant to them. A final salient theme was the need for relatable characters in the posts. This request led us to partner with EAAB members and content experts to identify EAs who volunteered to record videos in which they talked about their experiences with each health behavior. These videos were developed following best practices for health communication via microvideos [47-49], with scripting designed to concisely convey the most important and relevant information and be emotionally engaging and authentic. Revised posts were reviewed by the health behavior content experts to ensure information accuracy.

XSL•FO
RenderX

Table 3. Formative rural emerging adult surveys respondent demographics.

Category	Total (N=297)
Age (years) (n=297), mean (SD)	22.5 (2.49)
State of residence (n=297), n (%)	
Arizona	29 (9.8)
Colorado	34 (11.4)
Idaho	61 (20.5)
Montana	37 (12.5)
New Mexico	48 (16.2)
Nevada	26 (8.8)
Utah	32 (10.8)
Wyoming	30 (10.1)
Gender identity (n=297), n (%)	
Women	215 (72.4)
Men	70 (23.6)
Transgender/gender fluid/nonbinary	11 (3.7)
Ethnicity (n=297), n (%)	
Not Hispanic/Latino	216 (72.7)
Hispanic/Latino	72 (24.2)
Race[a] (n=297), n (%)	
American Indian/Alaska Native	26 (8.8)
Asian	4 (1.3)
Black/African American	17 (5.7)
White	210 (70.7)
More than one group	19 (6.4)
Employment[a,b] (n=197), n (%)	
Full-time	66 (38.1)
Part-time	49 (24.1)
Looking for work	37 (13.1)
Unemployed, caregiver, student	82 (43.2)
Living situation[b] (n=197), n (%)	
Alone	20 (13.0)
With parents	57 (22.2)
With roommates/friends	27 (14.1)
With spouse/partner	90 (46.2)
With other relatives	99 (34.3)
Education [c] (n=99), n (%)	
High school or less	54 (54.5)
Some college/trade school	17 (17.2)
Four-year college degree	20 (20.2)
Graduate degree	6 (6.1)
Health insurance[c] (n=99), n (%)	
Yes	63 (63.6)

XSL•FO
RenderX

Category	Total (N=297)
No	15 (15.2)
Who helps with health decisions?[a,c] (n=99), n (%)	
I make them alone	14 (14.1)
Parents	56 (55.6)
Roommates/friends	10 (10.1)
Spouse/partner	40 (40.4)
Others	15 (15.2)

[a]Multiple responses could be selected, so totals may not equal 100%.

[b]Question not asked in the third round of surveys.

[c]Question not asked in the second or third round of surveys.

XSL•FO
RenderX

Table 4. Thematic categories of feedback with reactions and recommendations for posts by rural emerging adults (EAs) responding to the formative surveys (N=297), Emerging Adult Advisory Board (EAAB) members (N=17), and Stakeholder Organization Advisory Board (SOAB) members (N=16).

Theme	Definition	Example quote on how to improve posts	Resultant modifications
Comment			
Message-specific reaction	Unique specific reactions, opinions, or inquiries about the message or content.	• What product do you use to clear up skin? • Groceries are very expensive nowadays, you can just go shopping and take whatever you want. You need to think twice… • Wow, I didn't know this! • Maybe I should try this. • How urgently should I get it [HPV vaccine]?	Specific unclear messages were revised; additional information was provided
Affirmative comment	Positive receptivity or relatability, makes a direct affirmative comment, associates a positive reflection, or indicates a positive personal resonance on the message, content, or creator.	• That this statement is true • That i like the message • Thank you for trying to help society • How helpful this ad would be • This is great…how things like this can help people • I appreciate the vote [of] confidence • How I can relate	Prioritization of posts that received ample affirmations
Improvement			
Suggestions on image	Improving visual elements that would enhance the overall impact and reach of the posts. Selecting images that are, but not limited to, authentic, relevant to the topic, visually appealing, and tailored to look like and engage the intended audience effectively.	• Using a real image as opposed to clip-art/AI art style image will help, as I would immediately see this as an ad and scroll past it. Also, the tips at the top make the block of text too large, leading to less people reading it. • Change the stock photo to a clearer image, this one is a bit blurry • Since it is geared towards young adults, I would change the picture to be of a young adult rather than an older woman.	Inclusion of more realistic images, improvements in image resolution, and emphasis on younger appearing characters in posts
Dislike or wrong audience	The content, although potentially valuable or well-constructed, fails to engage them personally; misalignment with their interests; direct or indirect discontent with the post; lack of connection with the message or intent.	• I'm sure it's helpful to many people but for me it isn't • It's a great post just not for me because I don't drink. • It is well made, just irrelevant to me. • I don't really keep up with influencer content, I don't know why I should stop and listen or care about this person's experience with quitting	Revision of posts deemed irrelevant
More information requested	Content, while engaging, lacks adequate detail or depth.	• I would just add more facts of why you overeat when you're watching a show because it can happen • More facts to back up the information • Give a description about what HPV is for people who do not know and how one might contract this	Revision of posts to include additional detail where possible; additional posts created when more in-depth information was requested
Reinforce source or suggestions on source	Demands for greater transparency and verification concerning the origins or credibility of the information presented in post; any suggestion about the source materials or the need to further elaborate on them; avoid certain sources due to local political climate, testimonials must seem trustworthy, such as those from partnering with influencers.	• Make sure any links provided and information listed is factual and from a trusted source. • The "doctor" needs to have credentials posted so he is more believable. Anyone can say they are a doctor. • I would add sources that you could click on so you could read on more about it • Maybe linking sites to help people quit • We are just a very politically charged town, I know some will see CNN and just avoid it	Inclusion of source material from various outlets and improved transparency of source material across posts

XSL•FO
RenderX

Theme	Definition	Example quote on how to improve posts	Resultant modifications
Suggestions on content	Direct and postspecific suggestions that reflect a collection of feedback aimed at improving the effectiveness or engagement; recommendations on, but not limited to, enriching the informational depth, avoid use of scare tactics, shocking words, or condescension, keep tone positive, optimizing visual appeal, adding interactive content, and content's relevance and accuracy.	• Add subtitles • I would tell a story to help build common ground with people who are addicted • Maybe get a more attractive guy have him lose the hat, people stop scrolling for people pleasing to the eye • Explain why quitting nicotine is hard and not to beat yourself up if you can't quit immediately • If she was wearing reflective clothing it'd get the point across better. • I would show examples of what could happen, perhaps provide stories from real life people as well • I'd be more likely to click on a poll post to also see how other people responded	Subtitles added to video posts, emphasis on images that align with main messages, increased use of interactive posts like polls, addition of posts with real-world rural EAs for relatability
Too much information presented	Concerns about the overload of text or information, highlights the need for simplification and streamlining in content presentation, emphasizing that concise and clear communication often resonates better where attention spans are short. Brief posts and videos are preferred.	• I would make it a little more less detailed • Maybe make the video a little shorter, or have graphic on the screen with the overarching message (like the power of new habits or something) • This post is direct and straight to the point. It does have a lot of text, so I would find a way to shorten the text to make it more enticing to read. • Limit what is said before the post. I've recently looked up the information it is talking about but it says a bit too much before having to click the link.	Revision of posts to reduce text, shorten videos (limited to under 2 minutes), and abbreviate post captions
Unclear message	Content's intended message is not comprehensible or straightforward, leading to confusion and diminished engagement.	• I couldn't really tell if it was about limiting screen time or food intake, and didn't really get how it was connecting them, so maybe make the overall message/goal clearer • Use less biased language, as it is already painting these sponsorships in a bad light, which will affect who clicks on the link more than a neutral factual title • Starting with something beside the question, or maybe a different question. It just didn't feel particularly intriguing. • Make the message more clear. I do not really understand this video	Unclear messages were revised and reviewed with EAAB and SOAB members and content experts to improve clarity in messaging

EAAB and SOAB Feedback and Resultant Modifications

Finally, feedback from the EAAB and SOAB largely focused on using widely recognized sources instead of smaller, more local sources for posts due to the fact that EAAB members stated they would feel weary of information provided by a source they did not recognize and providing links to every source used in the messaging as EAAB members felt the need to vet the information for themselves in order to trust it (Table 4).

PEAK Wellness Pilot Study

There were 26 EAs who completed the pretest survey and took part in the pilot study. Three participants were lost to follow-up at posttest, with 23 (89%) participants completing posttests. Participants were 85% female and 35% Hispanic, with a median age of 23 years (SD 2.32). At pretest, many EAs reported having elevated cancer risk factors: 53% engaged in <150 minutes of moderate-to-vigorous physical activity weekly, 85% had low daily intake of fruits and vegetables, 35% used nicotine products, 58% had binged alcoholic beverages in the past 30 days, 65% were sunburned in the past 3 months, and 38% had not received the HPV vaccination.

Among all participants there were 1060 post views (mean per post: 16.6, mean per participant: 40.8), 346 reactions (mean per post: 5.4, mean per participant: 13.3), and 72 comments (mean per post: 1.2, mean per participant: 3.0). The participants viewed most posts and reacted to some posts (Table 5). Comments from participants on the posts included requests for more information, and the intervention manager posted 6 responses. EAs viewed posts on all topics a similar number of times, with most posts receiving 16-17 views. Posts on nicotine product use, alcohol consumption, diet, and physical activity received the most reactions (Table 6). The higher viewership of the physical activity posts was due in part to being the topic for the one focus week of posts in the pilot feed. Media education posts received the most comments. Most participants reported reading Facebook posts from private groups in general at least once or more per day, increasing from 58% at pretest to 77% at posttest. Feedback about the Facebook group was largely positive: at posttest 83% strongly agreed or agreed they would like to use the group, thought it was easy to use (91%), felt confident using the group (82%), and thought it was user-friendly (91.7% said it was good, excellent, or best imaginable, Table S2 in Multimedia Appendix 1).

XSL•FO
RenderX

Table 5. Aggregated feedback on posts reviewed in the formative surveys (64 posts, across 26 participants).

Engagement	Total	Mean per post	Mean per participant
Views	1060	16.6	40.8
Reactions (eg, likes)	346	5.4	13.3
Comments	72	1.2	3.0

Table 6. Aggregated engagement on posts reviewed in the formative surveys (64 posts, across 26 participants).

Post topic	Number of posts	Views, N (mean/post)	Reactions, N (mean/post)	Comments, N (mean/post)
Physical activity	13	216 (16.6)	77 (5.9)	14 (1.1)
Diet	7	118 (16.9)	44 (6.3)	10 (1.4)
Sun safety	7	115 (16.4)	30 (4.3)	8 (1.1)
Tobacco cessation	6	99 (16.5)	56 (9.3)	2 (0.3)
Alcohol reduction	7	125 (17.9)	52 (7.4)	5 (0.7)
HPV[a] vaccination	7	115 (16.4)	26 (3.7)	4 (0.6)
Media education	3	52 (17.3)	15 (5.0)	8 (2.7)
Family communication	2	28 (14.0)	2 (1.0)	0 (0.0)

[a]HPV: human papillomavirus.

Discussion

Principal Findings

The purpose of this paper was to illustrate the systematic, iterative process for developing and pilot-testing a theory-based, web intervention that engages rural EAs and promotes healthy behaviors for cancer prevention in the United States. The process described herein fills a gap in the existing literature on realistic timelines, resources, and staffing required to rigorously design a theory-based web intervention by including practical examples, clear iterative steps in developing the intervention content, and community engagement strategies. We demonstrated how to combine health behavior theory, feedback from EAAB and SOAB advisory boards, and web design to develop a robust health intervention. Below, we describe keys to success, lessons learned (Textbox 1), and limitations of this approach to add to the literature on how to rigorously design reproducible web-based interventions about healthy behaviors.

Textbox 1. Lessons learned in developing a social media–based health education intervention.

- Staffing and stakeholders: Coalescing and maintaining stakeholders required cultivating networks via dedicated personnel effort, and the strongest stakeholder engagement resulted from continuous quarterly engagement.
- Post creation: Developing theory-informed, community-engaged content for posts required an iterative design and revision process that included the community with stakeholders, content experts, and media experts.
- Protocols: In addition to post creation, which largely occurred before launching the intervention, protocols for handling questions and comments, promoting audience engagement, and scheduling posts were useful for delegating tasks and enhancing the reproducibility of dynamic interventions.

Designing a successful web-based intervention required an iterative and responsive approach to the development and testing of post content, delivery processes, and engagement strategies. Our process involved continuous refinement of Facebook posts, ensuring that the posts were relevant, engaging, and theoretically grounded. We built, tested, and revised messages based on feedback from participants, allowing for an agile development process that incorporated 4 groups of experts: a core intervention post development team, health behavior content experts, emerging adult and stakeholder organization advisory boards, and an administrative core. Organizing these expert stakeholders had implications for staffing and resources because it required defining roles, maintaining iterative communication, providing leadership, and delegating. Describing these components of the community-engaged research process expands existing literature by demonstrating realistic timelines, resources, and personnel requirements. Our multi-round, expert-informed development process reflects an agile, person-based approach consistent with recommendations from Yardley et al [50], who emphasize iterative refinement grounded in end user perspectives. Unlike traditional static digital interventions, our approach involved continuous testing and revision across multiple stakeholder groups to increase the ability of participants to enroll in and engage with our content, aligning with emerging literature calling for more engaging, adaptive, and co-designed digital health tools [51]. Our iterative process enabled us to balance the needs and priorities of the communities we aimed to engage—considering factors such as content sources, design elements, and stylistic choices—alongside insights from health behavior content experts and best practices for Facebook

XSL•FO
RenderX

engagement. While prior social media–based cancer prevention interventions have largely targeted older adult or urban populations [52,53], our focus on rural EAs (ages 18-26 years) is rare in digital intervention research. This population-specific focus expands the literature by demonstrating how social media messaging must be contextually adapted to audience characteristics, including rural identity and platform use patterns.

This intervention was not only about delivering information but also about motivating audiences to critically engage with web-based content, underscoring the importance of strategic engagement strategies. Given the widespread desire for reliable, expert-driven web-based health content, our strategy involved providing evidence-based information in the digital space. Unlike most interventions that provide one-time educational modules on media literacy [54-57], ours provides media literacy education via real-time sourcing and question-and-answer opportunities, improving responsiveness through timely information correction and availability of content experts to address information gaps. A unique contribution of our project is insight into the digital skillsets required for creating a web-based intervention (eg, web-based platform format, digital editing, algorithm literacy). The intervention was fundamentally community-driven, shaped by ongoing formative research, engagement with stakeholders, and direct interactions with EAs. Recognizing that there is no singular "perfect" message or approach, we focused on key considerations to guide content selection and refinement (eg, source material, visual appeal, representativeness of rurality). While refining our messaging, we also applied a system to assess message acceptability and audience reception through multiple surveys with the target population, whereas most social media-based health messaging studies lack sufficient survey pretesting with small samples [58,59]. This was of utmost importance because it helped us understand which messages were and were not engaging to rural EAs, beyond our EAAB.

Two additional strategies for web-based intervention development emerged. The first is to include information sources that feel reliable to participants, and here we acknowledge that participant perceptions of reliability are critical, particularly given that rural audiences likely differ from urban audiences in their trust patterns in health information sources [60]. A key component of our intervention was training users to assess the quality of web-based information and equipping them with the skills to evaluate content critically, which we refer to as media education. Ideally, this media education will foster the perception of our feed as a credible and dependable resource, as well as help EAs make informed decisions about information from competing sources. Second, including content experts in user-generated engagement videos was seen as an important strategy to enhance the credibility of the research team and the intervention health messages. User-generated engagement videos are a known strategy in advertising and web-based engagement to build trust and authenticity [61], but are scarcely accepted in academic research, which often values professionally designed video content featuring influencers or clinical experts [62].

Designing posts that are theoretically informed and community-engaged required an iterative process of review and revision. In our case, it took two and a half years to develop 400 theoretically informed posts by a team that included 7 media experts, 11 health behavior content experts, 29 community advisors, 26 pilot study participants, and 297 formative survey research participants. The extent of this effort is a likely reason most social media–based interventions fail to incorporate theory [63-66]. To balance structure with responsiveness, we adopted an agile development process, starting with an initial library of posts while continuously refining and expanding content in response to EAAB and SOAB discussions, emerging health information trends, and relevant current events. A yearlong intervention was launched in a randomized trial in March 2025 with twice-daily posts. At the time of the trial launch, nearly half of the posts were developed, with the remaining portion being produced in near real-time to maintain relevance and responsiveness to current events, and reactions and comments from EA users. Our description of the web-based intervention process provides rationale and justification for the inclusion of stakeholder groups, professional networks, and research personnel's effort to iteratively design a robust web-based intervention.

The pilot test provided valuable insights into the potential acceptability of our messaging and the demographics of our audience. A key focus was examining the prevalence of cancer risk factors and understanding how our messages resonated with a group of EAs by how they engaged with the messages. One of the key lessons learned during the pilot study was the necessity of engaging with audience comments in real-time—a departure from traditional public health communication approaches that historically emerge from larger organizations and government, with little back-and-forth interaction with individuals. This required the development of a structured moderator protocol, including guidelines for managing post schedules and crafting thoughtful reactions to audience engagement, including responding to questions. We relied on this dynamic process to uncover trends in user engagement and tailor our final posts and schedules accordingly. This approach should ensure that our intervention remains relevant, engaging, and impactful.

Interestingly, none of the comments we received during the pilot study actively disputed our messages. Instead, they reflected EAs' genuine desire for more information, highlighting the growing complexity of the web-based information ecosystem and the challenges EAs face in navigating it. Determining credible information has become increasingly challenging for the public, underscoring the need for clear, accessible, and responsive health communication strategies. Pretesting participants' views of the credibility of national and local sources was an essential early step in intervention development that should repeated as views of national and local health sources may evolve over time. Adaptability in web-based public health interventions was also essential to ensure that messaging remained both informative and engaging in an ever-evolving digital landscape. For example, health information from a local source may be viewed as credible because of its familiarity to participants or ability to relate it to a local context while information from a national source may be seen as credible because of name recognition and large resources to access key information, demonstrating the nuance of how participants

XSL•FO
RenderX

perceive credibility and the need to determine and respond to participants' perceptions.

Limitations

Several limitations to the generalizability of this analysis of creating web-based health communication should be acknowledged. Our study targeted EAs, and our analysis may be limited to this age group, which is highly media-savvy and among the heaviest users of digital platforms. Our development process may also apply to interventions for middle-aged and older adults who are also active on the web. However, it may not fully translate as these older groups tend to have different web-based engagement patterns (eg, higher use of different platforms). Additionally, this formative research focused on the Mountain West region, a geographic area with a highly rural demographic composition compared with the rest of the United States, which may additionally limit the generalizability of our findings. Furthermore, the intervention is currently limited to the Facebook platform for its wide use (even among EAs) and private group function that enables us to exercise experimental control in the randomized study design. However, many Americans use other platforms for different media formats and topics. For instance, YouTube and TikTok offer primarily video-based content that may appeal to different demographics or enhance engagement better than Facebook alone can achieve. Future research could explore multi-platform strategies to extend the reach and impact of web-based interventions across EAs and other audiences.

The methods also had 2 limitations. The sample size for the pilot intervention was generally small, as is common in intervention development research [67,68], limiting the representativeness of the EA population and the power of the statistical tests. While the measures used in the pilot study are well-established with good psychometric properties, they rely on self-report, which may produce social desirability biases or recall bias. The larger trial is designed to conduct validation methods of these measures as well. However, we did include observational measures (ie, Facebook metrics) of engagement with the intervention feed in the pilot study.

Conclusions and Implications

We described a theory-informed community-engaged process for developing a web-based health behavior change intervention, PEAK Wellness. The methods and results provide a novel roadmap for developing community-driven web-based interventions. Community-engaged strategies improve the relevance, feasibility, and cultural fit of health behavior interventions by incorporating community priorities and contextual knowledge into design and pilot-testing. Such approaches can increase recruitment, retention, and implementation potential, while promoting equity and shared ownership, which is especially important in social media interventions that are implemented in geographically diverse areas. These benefits, however, come with added demands for time, partnership building, and iterative adaptation. The resources and effort required to appropriately engage stakeholders, researchers, and the target population, rural EAs, were substantial but essential for developing an engaging and acceptable intervention. This study provides insight that can be used to enhance the rigor and reproducibility of social media intervention designs and facilitate realistic planning of procedures for future web-based health behavior interventions. Testing of the PEAK Wellness intervention has recently commenced in a rigorous pragmatic trial, using a randomized stepped-wedge design, enrolling EAs aged 18-26 years in rural counties in the Western United States. Our dynamic, iterative intervention development process will allow us to remain responsive to the evolving digital landscape while maintaining a foundation of evidence-based communication to promote cancer risk reduction among our sample of rural EAs.

Acknowledgments

We acknowledge the support of former faculty and staff doctors Cynthia A Thompson and Jenna Hatcher, as well as staff Christopher F Jones, Kaila Christini, John A Torres, Emilia Yessenya Barraza Perez, and Tatiana Gerena for their contributions to this project. The authors declare the use of generative artificial intelligence in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: Literature search to find and identify articles that are likely relevant to the research reported in the manuscript. The GAI tool used was ChatGPT 5.1. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This study was supported by grants from the National Institutes of Health National Cancer Institute (R01CA268037, P30CA118100, P30CA046934, P30CA023074, and P30CA042014). The opinions are those of the authors. The funding agency had no input on the contents of this paper.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to participant privacy and confidentiality, but are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

DBB, ALS, CAT, DK, DT, KLH, ELW, BJW, WGW, KN, CKB, DDG, EAB, JSG, JH, and DW conceptualized the study, designed methods, and secured extramural funding. DBB, ALS, JSG, DK, DT, and MS are supervising project activities. AK, DDG, AKY, TG, KC, JB, and AS are managing day-to-day study activities. All authors reviewed and approved the manuscript before submission.

Conflicts of Interest

DBB, BJW, WGW, KN, JB, AS, and AK receive a salary from Klein Buendel, Inc. DBB's spouse is an owner of Klein Buendel, Inc. DK is a consultant for Merck and has received 2 Merck Investigator Studies Program research awards. DT receives compensation from the International Life Sciences Institute (ILSI) as the Editor-in-Chief for Nutrition Reviews. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Supplemental tables.
[DOCX File , 34 KB - jmir_v28i1e80803_app1.docx]

Multimedia Appendix 2
PEAK Wellness Chat Pilot Pretest Survey.
[PDF File (Adobe PDF File), 376 KB - jmir_v28i1e80803_app2.pdf]

Multimedia Appendix 3
STROBE checklist.
[DOC File , 83 KB - jmir_v28i1e80803_app3.doc]

References

1. Arnett JJ. Emerging adulthood: a theory of development from the late teens through the twenties. Am Psychol 2000;55(5):469-480. [doi: 10.1037/0003-066x.55.5.469]
2. Daw J, Margolis R, Wright L. Emerging adulthood, emergent health lifestyles: sociodemographic determinants of trajectories of smoking, binge drinking, obesity, and sedentary behavior. J Health Soc Behav 2017;58(2):181-197 [FREE Full text] [doi: 10.1177/0022146517702421] [Medline: 28661779]
3. Chen X, Orom H, Hay JL, Waters EA, Schofield E, Li Y, et al. Differences in rural and urban health information access and use. J Rural Health 2019;35(3):405-417 [FREE Full text] [doi: 10.1111/jrh.12335] [Medline: 30444935]
4. Social Media Fact Sheet. Pew Research Center. 2025. URL: https://www.pewresearch.org/internet/fact-sheet/social-media/ [accessed 2025-03-04]
5. Sarkar U, Le GM, Lyles CR, Ramo D, Linos E, Bibbins-Domingo K. Using social media to target cancer prevention in young adults: viewpoint. J Med Internet Res 2018;20(6):e203 [FREE Full text] [doi: 10.2196/jmir.8882] [Medline: 29871850]
6. Veil SR, Buehner T, Palenchar MJ. A work ‐ in ‐ process literature review: incorporating social media in risk and crisis communication. J Conting Crisis Manage 2011;19(2):110-122. [doi: 10.1111/j.1468-5973.2011.00639.x]
7. Breland JY, Quintiliani LM, Schneider KL, May CN, Pagoto S. Social media as a tool to increase the impact of public health research. Am J Public Health 2017;107(12):1890-1891. [doi: 10.2105/AJPH.2017.304098] [Medline: 29116846]
8. Steffens MS, Dunn AG, Leask J, Wiley KE. Using social media for vaccination promotion: practices and challenges. Digit Health 2020;6:2055207620970785 [FREE Full text] [doi: 10.1177/2055207620970785] [Medline: 35173976]
9. Kata A. Anti-vaccine activists, web 2.0, and the postmodern paradigm--an overview of tactics and tropes used online by the anti-vaccination movement. Vaccine 2012;30(25):3778-3789. [doi: 10.1016/j.vaccine.2011.11.112] [Medline: 22172504]
10. Warner EL, Basen-Engquist KM, Badger TA, Crane TE, Raber-Ramsey M. The online cancer nutrition misinformation: a framework of behavior change based on exposure to cancer nutrition misinformation. Cancer 2022;128(13):2540-2548 [FREE Full text] [doi: 10.1002/cncr.34218] [Medline: 35383913]
11. Broniatowski DA, Jamison AM, Qi S, AlKulaib L, Chen T, Benton A, et al. Weaponized health communication: twitter bots and russian trolls amplify the vaccine debate. Am J Public Health 2018;108(10):1378-1384. [doi: 10.2105/AJPH.2018.304567] [Medline: 30138075]
12. Chou WS, Prestin A, Lyons C, Wen K. Web 2.0 for health promotion: reviewing the current evidence. Am J Public Health 2013;103(1):e9-e18. [doi: 10.2105/AJPH.2012.301071] [Medline: 23153164]
13. Vaccine safety communication in the digital age. World Health Organization. 2018. URL: https://apps.who.int/iris/handle/10665/311961 [accessed 2025-03-04]
14. Chou WS, Oh A, Klein WMP. Addressing health-related misinformation on social media. JAMA 2018;320(23):2417-2418. [doi: 10.1001/jama.2018.16865] [Medline: 30428002]

XSL•FO
RenderX

15. Yang YT, Broniatowski DA, Reiss DR. Government role in regulating vaccine misinformation on social media platforms. JAMA Pediatr 2019;173(11):1011-1012. [doi: 10.1001/jamapediatrics.2019.2838] [Medline: 31479099]
16. Sylvia Chou W, Gaysynsky A, Cappella JN. Where we go from here: health misinformation on social media. Am J Public Health 2020;110(S3):S273-S275. [doi: 10.2105/AJPH.2020.305905] [Medline: 33001722]
17. Probst J, Eberth JM, Crouch E. Structural urbanism contributes to poorer health outcomes for rural America. Health Aff (Millwood) 2019;38(12):1976-1984. [doi: 10.1377/hlthaff.2019.00914] [Medline: 31794301]
18. Cosby AG, McDoom-Echebiri MM, James W, Khandekar H, Brown W, Hanna HL. Growth and persistence of place-based mortality in the United States: the rural mortality penalty. Am J Public Health 2019;109(1):155-162. [doi: 10.2105/AJPH.2018.304787] [Medline: 30496008]
19. About rural health. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/rural-health/php/about/index.html [accessed 2025-07-02]
20. Wallerstein N, Duran B, Oetzel JG, Minkler M. Community-Based Participatory Research for Health: Advancing Social and Health Equity. Hoboken ,New Jersey: John Wiley & Sons; 2017.
21. Israel BA, Schulz AJ, Parker EA, Becker AB. Review of community-based research: assessing partnership approaches to improve public health. Annu Rev Public Health 1998;19:173-202. [doi: 10.1146/annurev.publhealth.19.1.173] [Medline: 9611617]
22. Buller DB, Sussman AL, Thomson CA, Kepka D, Taren D, Henry KL, et al. #4Corners4Health social media cancer prevention campaign for emerging adults: protocol for a randomized stepped-wedge trial. JMIR Res Protoc 2024;13:e50392 [FREE Full text] [doi: 10.2196/50392] [Medline: 38386396]
23. Gottfried J. Americans' social media use. Pew Research Center. URL: https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/ [accessed 2025-03-04]
24. Bandura A. Social Foundations of Thought and Action: A Social Cognitive Theory. Englewood Cliffs, NJ, US: Prentice-Hall, Inc; 1986:617.
25. Deci EL, Ryan RM. Self-determination theory: a macrotheory of human motivation, development, and health. Can Psychol 2008;49(3):182-185. [doi: 10.1037/a0012801]
26. Rogers EM. Diffusion of Innovations. New York City: Simon and Schuster; 2003.
27. Rural-urban continuum codes. Economic Research Service. URL: https://www.ers.usda.gov/data-products/rural-urban-continuum-codes [accessed 2025-07-02]
28. Rural-urban commuting area codes. Economic Research Service. URL: https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes [accessed 2025-05-21]
29. Dynata. 2025. URL: https://www.dynata.com/ [accessed 2025-03-04]
30. Verasight. 2025. URL: https://www.verasight.io/ [accessed 2025-03-04]
31. Fisher WR. Narration as a human communication paradigm: the case of public moral argument. Commun Monogr 2009;51(1):1-22. [doi: 10.1080/03637758409390180]
32. Baker M. Preventing preventing skin cancer in adolescent girls through intervention with their mothers. East Tennessee State University. 2013. URL: https://dc.etsu.edu/etd/1163 [accessed 2025-12-05]
33. Pagoto SL, Schneider KL, Oleski J, Bodenlos JS, Ma Y. The sunless study: a beach randomized trial of a skin cancer prevention intervention promoting sunless tanning. Arch Dermatol 2010;146(9):979-984 [FREE Full text] [doi: 10.1001/archdermatol.2010.203] [Medline: 20855696]
34. Hillhouse JJ, Turrisi R. Examination of the efficacy of an appearance-focused intervention to reduce UV exposure. J Behav Med 2002;25(4):395-409. [doi: 10.1023/a:1015870516460] [Medline: 12136499]
35. Global physical activity questionnaire (GPAQ). World Health Organization. URL: https://www.who.int/publications/m/item/global-physical-activity-questionnaire [accessed 2025-06-03]
36. Thompson FE, Midthune D, Subar AF, Kahle LL, Schatzkin A, Kipnis V. Performance of a short tool to assess dietary intakes of fruits and vegetables, percentage energy from fat and fibre. Public Health Nutr 2004;7(8):1097-1105. [doi: 10.1079/PHN2004642] [Medline: 15548349]
37. Hewawitharana SC, Thompson FE, Loria CM, Strauss W, Nagaraja J, Ritchie L, et al. Comparison of the NHANES dietary screener questionnaire to the automated self-administered 24-hour recall for children in the healthy communities study. Nutr J 2018;17(1):111 [FREE Full text] [doi: 10.1186/s12937-018-0415-1] [Medline: 30482218]
38. Viral hepatitis and liver disease. United States Department of Veterans Affairs. URL: https://www.hepatitis.va.gov/alcohol/treatment/audit-c.asp [accessed 2025-06-03]
39. Hughes J, Keely J, Niaura R, Ossip-Klein D, Richmond R, Swan G. Measures of abstinence in clinical trials: issues and recommendations. Nicotine Tob Res 2003;5(1):13-25. [Medline: 12745503]
40. Velicer WF, Prochaska JO. A comparison of four self-report smoking cessation outcome measures. Addict Behav 2004;29(1):51-60. [doi: 10.1016/s0306-4603(03)00084-4] [Medline: 14667420]
41. View HINTS questions. National Cancer Institute. URL: https://hints.cancer.gov/view-questions/all-hints-questions.aspx [accessed 2025-07-02]
42. Shoveller JA, Lovato CY. Measuring self-reported sunburn: challenges and recommendations. Chronic Dis Can 2001;22(3-4):83-98. [Medline: 11779421]

XSL•FO
RenderX

43. Lovato C, Shoveller J, Mills C. Canadian national workshop on measurement of sun-related behaviours [workshop report]. Chronic Dis Can 1999;20(2):96-100. [Medline: 10455042]
44. de Waal AC, van Rossum MM, Kiemeney L, Aben K. Reproducibility of self-reported melanoma risk factors in melanoma patients. Melanoma Res 2014;24(6):592-601. [doi: 10.1097/CMR.0000000000000089] [Medline: 24892955]
45. Hillhouse J, Turrisi R, Jaccard J, Robinson J. Accuracy of self-reported sun exposure and sun protection behavior. Prev Sci 2012;13(5):519-531 [FREE Full text] [doi: 10.1007/s11121-012-0278-1] [Medline: 22855253]
46. Moreno MA, D'Angelo J. Social media intervention design: applying an affordances framework. J Med Internet Res 2019;21(3):e11014 [FREE Full text] [doi: 10.2196/11014] [Medline: 30912754]
47. Zhu Z, Liu S, Zhang R. Examining the persuasive effects of health communication in short videos: systematic review. J Med Internet Res 2023;25:e48508 [FREE Full text] [doi: 10.2196/48508] [Medline: 37831488]
48. Emilia. The ultimate guide to microvideos – use cases, tools and tips for SaaS. Userpilot Blog. URL: https://userpilot.com/blog/microvideos-guide/ [accessed 2024-06-12]
49. Ester P, Morales I, Herrero L. Micro-videos as a learning tool for professional practice during the post-COVID era: an educational experience. Sustainability 2023;15(6):5596. [doi: 10.3390/su15065596]
50. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015;17(1):e30 [FREE Full text] [doi: 10.2196/jmir.4055] [Medline: 25639757]
51. O'Connor S, Hanlon P, O'Donnell CA, Garcia S, Glanville J, Mair FS. Barriers and facilitators to patient and public engagement and recruitment to digital health interventions: protocol of a systematic review of qualitative studies. BMJ Open 2016;6(9):e010895 [FREE Full text] [doi: 10.1136/bmjopen-2015-010895] [Medline: 27591017]
52. Valle CG, Tate DF, Mayer DK, Allicock M, Cai J. A randomized trial of a Facebook-based physical activity intervention for young adult cancer survivors. J Cancer Surviv 2013;7(3):355-368 [FREE Full text] [doi: 10.1007/s11764-013-0279-5] [Medline: 23532799]
53. Han CJ, Lee YJ, Demiris G. Interventions using social media for cancer prevention and management: a systematic review. Cancer Nurs 2018;41(6):E19-E31 [FREE Full text] [doi: 10.1097/NCC.0000000000000534] [Medline: 28753192]
54. Housten AJ, Gunn CM, Paasche-Orlow MK, Basen-Engquist KM. Health literacy interventions in cancer: a systematic review. J Cancer Educ 2021;36(2):240-252 [FREE Full text] [doi: 10.1007/s13187-020-01915-x] [Medline: 33155097]
55. Peipert JD, Lad T, Khosla PG, Garcia SF, Hahn EA. A low literacy, multimedia health information technology intervention to enhance patient-centered cancer care in safety net settings increased cancer knowledge in a randomized controlled trial. Cancer Control 2021;28:10732748211036783 [FREE Full text] [doi: 10.1177/10732748211036783] [Medline: 34565193]
56. Lazard AJ, Queen TL, Pulido M, Lake S, Nicolla S, Tan H, et al. Social media prompts to encourage intervening with cancer treatment misinformation. Soc Sci Med 2025;372:117950. [doi: 10.1016/j.socscimed.2025.117950] [Medline: 40096813]
57. Lyons B, King AJ, Kaphingst KA. A health media literacy intervention increases skepticism of both inaccurate and accurate cancer news among U.S. adults. Ann Behav Med 2024;58(12):820-831. [doi: 10.1093/abm/kaae054] [Medline: 39417815]
58. Tappin BM, Hewitt LB. Using survey experiment pretesting to support future pandemic response. PNAS Nexus 2024;3(11):pgae469 [FREE Full text] [doi: 10.1093/pnasnexus/pgae469] [Medline: 39484044]
59. Ricci-Cabello I, Bobrow K, Islam SMS, Chow CK, Maddison R, Whittaker R, et al. Examining development processes for text messaging interventions to prevent cardiovascular disease: systematic literature review. JMIR Mhealth Uhealth 2019;7(3):e12191 [FREE Full text] [doi: 10.2196/12191] [Medline: 30924790]
60. Lee W, Kim EM, Nemirovski EA, Kamprath S, Masel MC, Patel DI. Public trust in different sources of information: gaps in rural residents and cancer patients. Healthcare (Basel) 2025;13(6):640 [FREE Full text] [doi: 10.3390/healthcare13060640] [Medline: 40150490]
61. Ismail S, Abdul Latif R. Authenticity issues of social media: credibility, quality, and reality. World Acad Sci Eng Technol 2013;74:254-261 [FREE Full text]
62. Kaňková J, Binder A, Matthes J. Helpful or harmful? navigating the impact of social media influencers' health advice: insights from health expert content creators. BMC Public Health 2024 Dec 18;24(1):3511 [FREE Full text] [doi: 10.1186/s12889-024-21095-3] [Medline: 39696170]
63. Paul B, Headley-Johnson S. The impact of social media on health behaviors, a systematic review. Healthcare (Basel) 2025;13(21):2763 [FREE Full text] [doi: 10.3390/healthcare13212763] [Medline: 41228133]
64. Maher CA, Lewis LK, Ferrar K, Marshall S, De Bourdeaudhuij I, Vandelanotte C. Are health behavior change interventions that use online social networks effective? a systematic review. J Med Internet Res 2014;16(2):e40 [FREE Full text] [doi: 10.2196/jmir.2952] [Medline: 24550083]
65. Laranjo L, Arguel A, Neves AL, Gallagher AM, Kaplan R, Mortimer N, et al. The influence of social networking sites on health behavior change: a systematic review and meta-analysis. J Am Med Inform Assoc 2015;22(1):243-256 [FREE Full text] [doi: 10.1136/amiajnl-2014-002841] [Medline: 25005606]
66. Seiler J, Libby TE, Jackson E, Lingappa J, Evans W. Social media-based interventions for health behavior change in low- and middle-income countries: systematic review. J Med Internet Res 2022;24(4):e31889 [FREE Full text] [doi: 10.2196/31889] [Medline: 35436220]

XSL•FO
RenderX

67. Leon AC, Davis LL, Kraemer HC. The role and interpretation of pilot studies in clinical research. J Psychiatr Res 2011;45(5):626-629 [FREE Full text] [doi: 10.1016/j.jpsychires.2010.10.008] [Medline: 21035130]
68. Hertzog MA. Considerations in determining sample size for pilot studies. Res Nurs Health 2008;31(2):180-191. [doi: 10.1002/nur.20247] [Medline: 18183564]

Abbreviations

EA: emerging adult
EAAB: Emerging Adult Advisory Board
HPV: human papillomavirus
REDCap: Research Electronic Data Capture
RUCC: Rural Urban Continuum Codes
SOAB: Stakeholder Organization Advisory Board
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Stone; submitted 17.Jul.2025; peer-reviewed by B Uko, S Narayan, M Chukwuka, D Gyimah; comments to author 23.Sep.2025; revised version received 26.Nov.2025; accepted 02.Dec.2025; published 08.Jan.2026.

Please cite as:
Warner EL, Kinsey A, Walkosz BJ, Berteletti J, Nuss K, Small A, Woodall WG, Kepka D, Taren D, Skiba MB, Guest DD, Blair CK, Gordon JS, Wetter DW, Borrayo EA, Henry KL, Sussman AL, Buller DB
A Web-Based Cancer Prevention Intervention for Rural Emerging Adults: Mixed Methods Development and Pilot-Testing Study
J Med Internet Res 2026;28:e80803
URL: https://www.jmir.org/2026/1/e80803
doi:10.2196/80803
PMID:

©Echo L Warner, Alishia Kinsey, Barbara J Walkosz, Julia Berteletti, Kayla Nuss, Annelise Small, W Gill Woodall, Deanna Kepka, Douglas Taren, Meghan B Skiba, Dolores D Guest, Cindy K Blair, Judith S Gordon, David W Wetter, Evelinn A Borrayo, Kimberly L Henry, Andrew L Sussman, David B Buller. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effects of Artificial Intelligence Recognition–Based Telerehabilitation on Exercise Capacity in Patients With Hypertension: Randomized Controlled Trial

Qiuru Yao[1,2,3,4], MS; Baizhi Qiu[1,2,3,4], MS; Longlong He[5], MS; Qin Wang[1,2,3,4], MS; Jihua Zou[1,3,4,6,7], PhD; Donghui Liang[8], PhD; Shuyang Wen[1,2,3,4], MS; Yingchao Liu[1,3,4], MM; Gege Li[1,3,4,6], MS; Jinjing Hu[1,3,4,6], MS; Huan Ma[9*], PhD; Guozhi Huang[1,3,4,6*], PhD; Qing Zeng[1,3,4,6*], PhD

[1]Center of Rehabilitation Medicine, Zhujiang Hospital, Guangzhou, China

[2]School of Nursing, Southern Medical University, Guangzhou, China

[3]GuangDong Engineering Technology Research Center of Brain Function Assessment and Neuroregulation Rehabilitation, Guangzhou, China

[4]Institute of Exercise and Rehabilitation Science, Zhujiang Hospital, Guangzhou, China

[5]Department of Clinical Medicine, Xiamen Medical College, Xiamen, China

[6]School of Rehabilitation Sciences, Southern Medical University, Guangzhou, China

[7]Department of Rehabilitation Sciences, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)

[8]Department of Traditional Chinese Medicine, Zhujiang Hospital, Guangzhou, China

[9]Department of Cardiology, Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Southern Medical University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Qing Zeng, PhD
Center of Rehabilitation Medicine
Zhujiang Hospital
NO.253, Industrial Avenue Middle
Guangzhou, 510280
China
Phone: 86 15989189468
Email: zengqingyang203@126.com

Abstract

Background: Hypertension remains a major global health challenge, significantly increasing cardiovascular and all-cause mortality risks. While exercise therapy is effective, conventional approaches face limitations in accessibility and personalization, compromising adherence. Artificial intelligence (AI)–assisted remote rehabilitation enables real-time monitoring and personalized guidance, offering a promising alternative. Nevertheless, its clinical benefits and applicability require further systematic validation.

Objective: This study aimed to evaluate the efficacy of an 8-week AI-assisted telerehabilitation program on improving exercise capacity and related health outcomes in patients with hypertension.

Methods: This prospective, dual-arm, parallel, open-label, randomized controlled trial enrolled 62 patients with hypertension recruited via convenience sampling. Participants were adults aged between 18 and 75 years with a confirmed hypertension diagnosis who were excluded for severe cardiac complications, recent myocardial infarction, unstable angina, or physical disabilities preventing exercise. The participants were randomly assigned (1:1) to an intervention group that received AI-assisted remote rehabilitation plus routine health education, or a control group that received health education and conventional offline exercise guidance. The supervised exercise program included warm-up, cardiorespiratory endurance, strength resistance, balance, and flexibility training, followed by a cooldown. Sessions lasted between 30 and 50 minutes and were performed at least 3 times weekly for 8 weeks. Assessments at baseline and 8 weeks included the 6-minute walk test (6MWT), cardiopulmonary exercise testing (CPET), International Physical Activity Questionnaire (IPAQ), Short-Form Health Survey 12 (SF-12), Patient Health Questionnaire-9 (PHQ-9), Generalized Anxiety Disorder-7 (GAD-7), exercise self-efficacy, blood pressure (BP), body weight, handgrip strength, and other health-related indicators. The primary outcome was the change in 6-minute walk distance (6MWD). Data were analyzed according to the intention-to-treat principle.

XSL•FO
RenderX

Results: Throughout the 8-week intervention period, no serious adverse events related to the AI-assisted telerehabilitation intervention occurred. After 8 weeks, the intervention group demonstrated significantly greater improvements than the control group in 6-minute walk distance (6MWD; adjusted mean difference 62.77, 95% CI 26.33-99.22; *P*=.002), systolic BP reduction (adjusted mean difference 4.11, 95% CI 0.11-8.28; *P*=.046), IPAQ score (adjusted mean difference 658.96, 95% CI 159.23-1158.69; *P*=.011), exercise self-efficacy score (adjusted mean difference 21.71, 95% CI 13.59-29.82; *P*<.001), total exercise time (adjusted mean difference 98.24, 95% CI 49.39-147.08; *P*=.001) peak oxygen uptake (peak VO_2) (adjusted mean difference 3.39, 95% CI 0.49-6.29; *P*=.026), and peak oxygen uptake percent predicted (peak VO_2%pred) (adjusted mean difference 11.58, 95% CI 2.06-21.10; *P*=.021).

Conclusions: Compared with conventional exercise rehabilitation, AI-assisted remote rehabilitation was found to improve exercise capacity, boost regular physical activity and exercise self-efficacy, and aid in systolic BP control among patients with hypertension. This study positioned AI-assisted rehabilitation as a scalable and effective strategy for real-world hypertension management. It further contributes actionable guidance for developing effective home-based exercise strategies tailored to populations with hypertension.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2300076451; https://www.chictr.org.cn/showproj.html?proj=208353

(J Med Internet Res 2026;28:e81400) doi:10.2196/81400

KEYWORDS

hypertension; artificial intelligence; telerehabilitation; lifestyle change; exercise habit formation; randomized controlled trial

Introduction

Hypertension is a major public health concern with an enormous economic and social burden [1]. The World Health Statistics 2023 report by the World Health Organization (WHO) states that the worldwide prevalence of hypertension reached 33%, and it remains an upward trend in the context of global aging [2]. Given the high prevalence and severity of hypertension, there is an urgent need to implement effective, widely available, and sustainable strategies for its management.

Current interventions for hypertension are divided into pharmacological and nonpharmacological treatments. While pharmacological therapy effectively lowers blood pressure (BP) and remains the clinical mainstay [3,4], its benefits are often limited by poor adherence and the risk of BP rebound upon discontinuation, which elevates the risk of major cardiovascular events [5,6]. In contrast, nonpharmacological management focuses on lifestyle modifications, such as regular physical activity, dietary changes (eg, salt intake restriction), and weight management. The health benefits of exercise are well-established—regular moderate-to-high intensity aerobic activity can reduce BP by an average of 11/5 mm Hg [7]. Furthermore, its efficacy in lowering BP may be comparable to that of pharmacological interventions [8].

Regular exercise, a key nonpharmacological intervention, significantly reduces hypertension risk, improves BP control, and plays a vital role in its prevention and management [9]. Regular exercise not only significantly lowers BP but also helps regulate weight, improve lipid and blood sugar levels, and promote overall metabolic function. Engaging in moderate physical exercise before the onset of hypertension, combined with a healthy lifestyle, can reduce the incidence of primary hypertension by 54% [10]. Moderate exercise can help delay the progression of hypertension, while excessive exercise may cause BP to rise, thereby increasing the risk of cardiovascular and cerebrovascular diseases. However, the current state of physical activity in patients with hypertension is not encouraging. Patients with hypertension report being less physically active compared with individuals without hypertension, and they are less likely to meet physical activity recommendations [11]. Furthermore, they often encounter numerous obstacles when it comes to maintaining regular physical activity, such as a lack of self-efficacy, social support, and adequate supervision [12]. Therefore, selecting appropriate exercise methods and training approaches for patients with hypertension is of great significance for their BP management [13,14].

The rapid development of digital therapeutics in recent years has the potential to be an important part of BP management in the future [15]. It is an emerging, promising field of medicine that aims to implement lifestyle changes and ultimately facilitate disease management using software programs such as smartphone applications and device algorithms [16,17]. Digital therapeutics are gaining ground in the fields of medicine and health care, and the HERB Digital Hypertension 1 (HERB-DH1) pivotal trial conducted in Japan is one such success case [18]. Although there are many mobile technologies available for improving BP management, unfortunately, only a few of them have been developed with the involvement of health care professionals or medical organizations, and the evidence of their long-term benefits in terms of BP management is still scarce [19,20]. Meanwhile, existing digital therapeutics (such as the HERB-DH1 behavioral algorithm) face limitations due to the absence of real-time motion feedback mechanisms, making it challenging to effectively monitor and ensure patient adherence to exercise prescriptions. This technological gap highlights the medical potential of 3D skeletal pose recognition technology. By capturing and analyzing human motion data, this technology enables real-time corrective guidance and dynamically adaptive interventions, serving as an intelligent monitor that safeguards exercise safety and efficacy.

Among numerous artificial intelligence (AI) technologies, posture recognition, as a technology capable of understanding and interpreting human behavior, is gradually becoming a hot

XSL•FO
RenderX

topic of research. As an important branch of AI, posture recognition technology enables machines to better understand human behavior by capturing, analyzing, and interpreting human movements, thereby achieving a leap forward in human-machine interaction. Posture recognition technology detects the position and orientation of a person or object from a video or image and can significantly improve the accuracy and quality of telerehabilitation services [21]. Human posture estimation techniques use input data, such as images, to help with reconstructing human representations to assist in measuring changes in patient movement and assessing limb or joint function [22] and to provide feedback during the treatment process to help patients understand and correct postural errors. Research has shown [23] that an AI-based telerehabilitation system, combined with 3D posture recognition technology, can be a viable alternative to exercise intervention for patients with sarcopenia. In our previous work in the field of low back pain using AI-assisted telerehabilitation technology, we found that AI-assisted multimodal movement-based telerehabilitation could significantly improve patients' lower back pain and provide better therapeutic outcomes than traditional movement-based telerehabilitation [24]. This technology significantly improves the accuracy of training guidance through high-precision motion capture and real-time feedback mechanisms, and its intervention effect is equivalent to that of traditional face-to-face training and conventional telerehabilitation programs, while optimizing the quality of standardized execution of rehabilitation training. However, there is insufficient evidence in the field of hypertension.

Considering the above, we utilized a novel smartphone app, developed under the guidance of health care professionals, to address this need. This app was designed to help patients with hypertension develop a habit of regular physical activity and improve their overall quality of life. Its advantage lies in providing online, personalized exercise prescriptions, thereby enabling a scientific, effective, and targeted remote training intervention. Accordingly, this study aimed to examine the effects of this AI-assisted remote rehabilitation system on exercise capacity, self-efficacy, psychological well-being, and related health indicators in patients with hypertension. We hypothesized that this intervention would positively influence these treatment outcomes compared to usual care.

Methods

Study Design

This study was a multicenter, open-label randomized controlled trial conducted in Guangdong Provincial People's Hospital (Guangzhou, China) and Zhujiang Hospital, Southern Medical University (Guangzhou, China), 2 large tertiary care hospitals in southern China. The reporting of this trial conformed to the CONSORT (Consolidated Standards of Reporting Trials) guidelines for randomized controlled trials, and the study protocol adhered to the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) checklist (Multimedia Appendix 1) [25,26]. The study was prospectively registered in the Chinese Clinical Trial Registry (ChiCTR2300076451) on October 9, 2023.

Ethical Considerations

Overview

This study was conducted in accordance with the principles of the Declaration of Helsinki and relevant Chinese regulations governing clinical research. The study protocol was reviewed and approved by the ethics committee of Zhujiang Hospital, Southern Medical University (2023-KY-190-01). The study was also filed in the Medical Research Registration Information System of the National Health Commission of China. All participants provided written informed consent prior to enrollment. The consent form detailed the study objectives, procedures, potential risks and benefits, and participants' rights. For participants who were unable to continue the study due to health or personal reasons, data were retained and analyzed in accordance with the intention-to-treat principle, as approved by the ethics committee.

Privacy and Confidentiality

All participant data were deidentified prior to analysis. Personal information and medical records were stored securely and accessible only to authorized research personnel. Specifically, for the AI-assisted telerehabilitation, the technology platform was designed to safeguard privacy by not recording any personal identifiers (eg, real names or ID numbers). Unless users explicitly opted in for video recording, the system did not capture or store any recognizable video footage. Instead, it processed the video stream in real time to extract only the coordinate data of 17 skeletal key points, using this anonymized information to compute exercise completion and quality scores. Data anonymity was maintained throughout the study and in all publications.

Compensation

Participants did not receive financial compensation for their involvement in the study. However, they were provided with free health education materials and access to the AI-based telerehabilitation application during the intervention period.

Use of Images and Identifiable Information

The images included in this paper (eg, app interfaces, training scenarios) do not contain any identifiable images of the participants. All screenshots and illustrations are system display interfaces and have been obtained with the consent of the system owner and the person who took the photos. Therefore, no additional consent for image publication was required.

Participant Selection and Recruitment

Recruitment was conducted from November 2023 to October 2024. Given the pragmatic nature of this trial and the need to consecutively enroll eligible patients, a convenience sampling method was employed. Participants were recruited from the cardiology outpatient clinic and via community advertisements. Participants were primarily recruited from the cardiology outpatient clinics of Guangdong Provincial People's Hospital (Guangzhou, China) and Zhujiang Hospital, Southern Medical University (Guangzhou, China), and through community advertisements in the surrounding areas.

XSL•FO
RenderX

Recruitment information was disseminated through a combination of online and traditional channels. Online strategies included utilizing social media platforms, such as WeChat, to publish recruitment announcements and conduct preliminary online screenings. Traditional methods involved on-site activities (eg, setting up recruitment booths in hospitals and communities to provide briefings and answer questions directly), physician referrals from established partnerships, and the distribution of printed leaflets and posters.

Inclusion and Exclusion Criteria

Inclusion Criteria

The inclusion criteria were as follows: age 18 to 75 years; diagnosis of essential hypertension, defined as an office systolic blood pressure (SBP) ≥140 mmHg and/or diastolic blood pressure (DBP) ≥90 mmHg; no cognitive impairment or mental illness; passed the physical activity readiness questionnaire screening or approved by a physician prior to participation; ability to communicate normally and cooperate with the questionnaire survey; not participating in other research involving healthy lifestyle promotion, particularly physical exercise; no prior exercise habit, defined as exercising less than 3 times per week for under 30 minutes per session; and ownership of a smartphone with access to WeChat and the internet.

Exclusion Criteria

The exclusion criteria were as follows: patients with severe or ineffectively controlled hypertension (SBP ≥180 mmHg and/or DBP ≥110 mmHg at rest); patients with pulmonary hypertension, severe hypertension, hypertensive crisis, unstable stage III hypertension, hypertensive encephalopathy, acute hypertension, or other serious complications such as severe arrhythmia, tachycardia, heart failure, unstable angina pectoris, or obvious adverse reactions to antihypertensive drugs with failure to control; patients with malignant tumors or hypertension combined with left ventricular ejection fraction <55%, aortic, mitral, tricuspid, or pulmonary stenosis, severe mitral, tricuspid, or pulmonary valve insufficiency, hypertrophic cardiomyopathy, acute infection, fundus hemorrhage, diabetic acidosis, gangrene of the lower limbs, severe hypothyroidism, or renal insufficiency; patients with motor organ injury who underwent bone or joint surgery within the past year, such as hip or knee replacement or spine surgery; patients unable to understand or answer questions or complete questionnaires due to subjective or objective reasons; and participants considered unsuitable for clinical trials by the researchers.

Sample Size

In this study, the sample size was calculated using the 6-minute walk test (6MWT) distance as the primary outcome indicator. According to the relevant literature [27,28], telemedicine has been shown to improve exercise capacity and physical function in patients with cardiovascular disease. Setting the minimal clinical difference significance (>33 m), 2-sided α=.05, and the test power 1-β 0.90. Calculations were performed using PASS software (version 15.0; NCSS LLC). Considering the loss and refusal rate of 20%, a minimum of 29 participants for each group was required, for a total of at least 58 participants.

Randomization and Blinding

Block randomization was performed using SAS software (version 9.3; SAS Institute Inc) to ensure balanced group sizes and prevent the potential for imbalance inherent in simple randomization. The randomization process allocated participants to either the intervention or control group using a block size of 6, across a total of 10 blocks, to generate the random sequence and obtain group allocations. The allocation sequence was concealed using opaque, sealed envelopes. All researchers responsible for patient recruitment and data collection were blinded to the sequence generation and were not involved in preparing the concealment envelopes.

Interventions

Remote Exercise Rehabilitation

Based on hypertension health education and exercise guidance, the intervention group used the “rehabilitation exercise prescription” App (RaphAI Health Technologies Limited) to carry out AI-assisted remote rehabilitation exercise intervention (Figure 1). The specific procedures included are detailed as follows.

XSL•FO
RenderX

Figure 1. Interface demonstration of the "Rehabilitation Exercise Prescription" mobile app used for artificial intelligence (AI)–assisted telerehabilitation. A: main landing page of the app; B: interface showing a personalized exercise prescription generated for a participant; C: screen displayed upon completion of an exercise session, summarizing performance metrics.

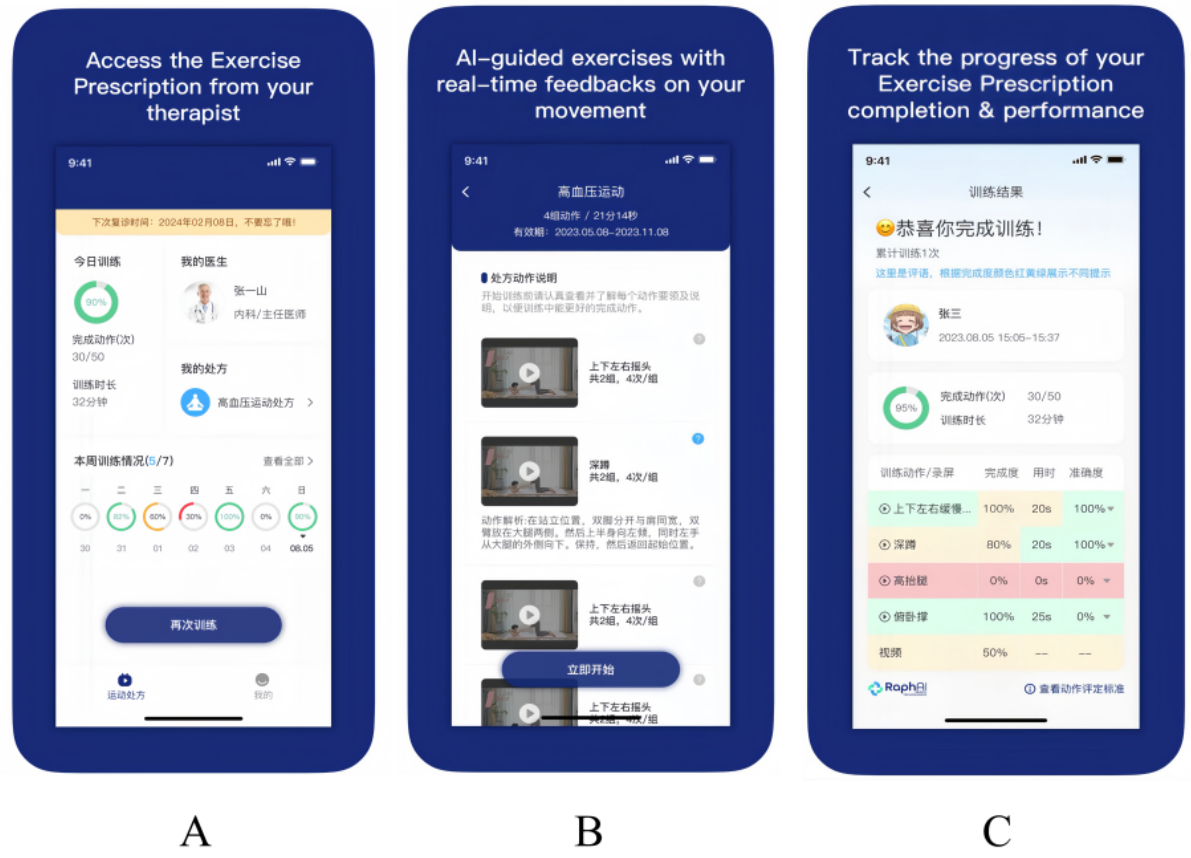

Assessment

Medical history was collected using the online assessment questionnaire, and each patient's general condition was evaluated through a combination of physical examination and questionnaire results. The assessment included (1) general information, including basic information (name, gender, age, contact information), behavioral data (smoking, drinking habits, eating habits), and medical history (past medical history, history of hypertension treatment, family history of hypertension, BP measurement); (2) current physical activity level (regular exercise, type and frequency of exercise); (3) anthropometric data (height, weight, waist circumference, and hip circumference); (4) the American Heart Association (AHA)/American College of Sports Medicine (ACSM) Health/Physical Fitness Preexercise Screening Questionnaire; (5) physical fitness assessments (cardiorespiratory endurance, strength, coordination, flexibility and balance assessment); and the (6) Physical Activity Readiness questionnaire (PAR-Q).

Exercise Prescription Generation and Delivery

For the prescription generation, 4 personalized cardiopulmonary rehabilitation prescriptions were customized according to the standards of the ACSM. After the assessment, the system determined the patient's risk stratification according to the self-assessment and offline assessment results and automatically generated exercise prescriptions.

The prescription was pushed to the patient 5 times a week for 30 to 50 minutes each time on a regular basis and was required to be completed at least 3 times a week. The content included warm-up, cardiopulmonary endurance training, strength resistance training, balance training, flexibility training, and cooldown.

AI Technology and Motion Capture Framework

The "Rehabilitation Exercise Prescription" app was built upon an AI engine that integrated TensorFlow, an open-source machine learning library, with the OpenPose architecture for real-time human pose estimation. This system employed a bottom-up approach using Part Affinity Fields to detect 17 key body joints (including eyes, nose, ears, shoulders, elbows, wrists, hips, knees, and ankles) and assembled them into a full-body skeleton without the need for physical markers. By processing video frames from the smartphone camera, the system tracked the coordinate changes of these skeletal key points across consecutive frames to quantify and analyze patient movement in real time.

Rehabilitation Training

The patients were trained according to the exercise prescription pushed by the app, and the action explanation was provided during the training. Computer AI visual recognition technology was used to identify 17 key bone points in the camera in a markerless way to monitor the training data in real time,

XSL•FO
RenderX

calculate the angle and displacement of different parts of the body in real time, and provide intelligent feedback and corrective guidance (Figure 2).

Figure 2. Examples of an artificial intelligence (AI)–assisted telerehabilitation process using the "Rehabilitation Exercise Prescription" app in patients with hypertension. A: scene during participant training; B: mobile interface showing real-time video and AI movement guidance.

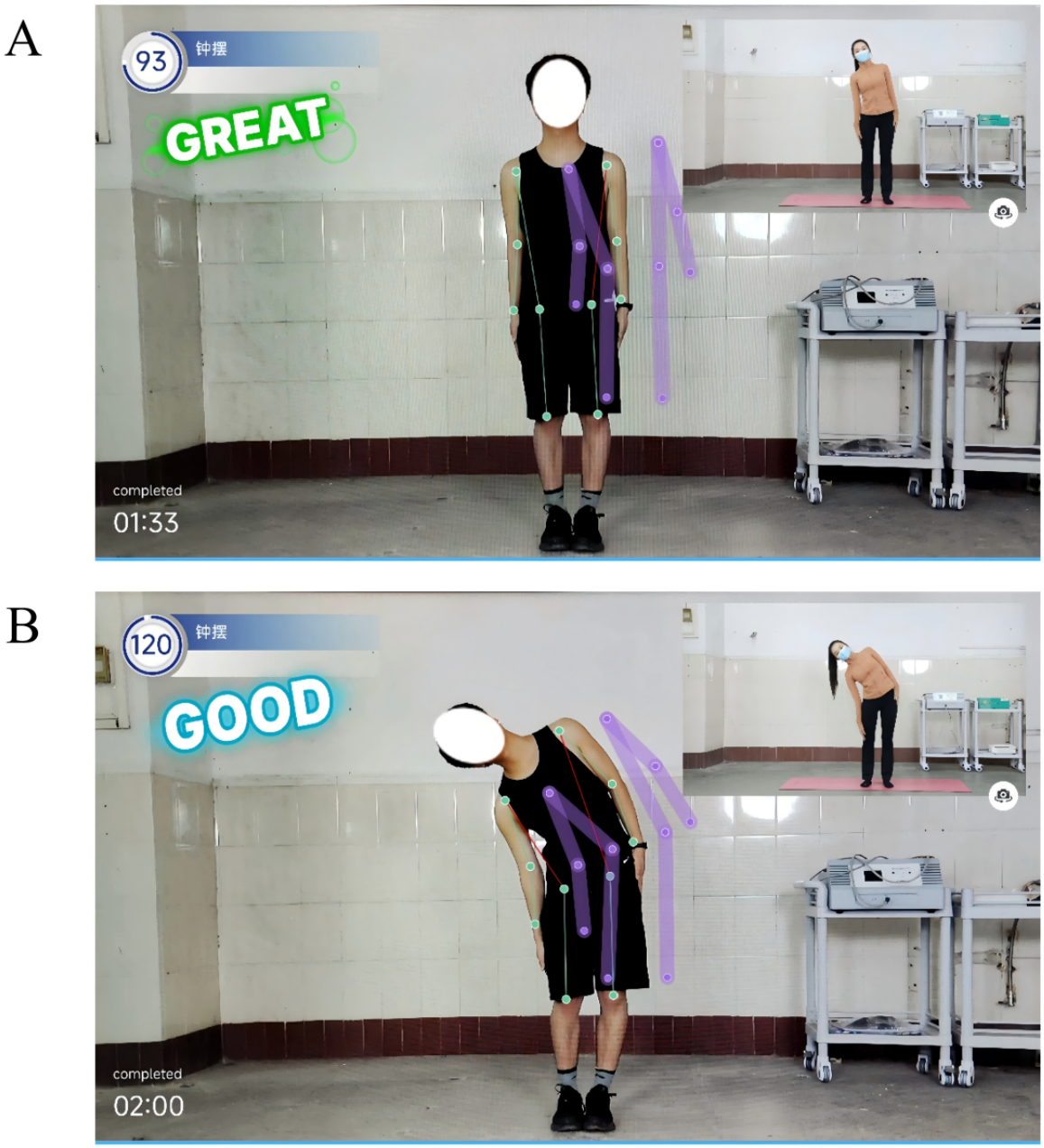

Follow-Up and Optimization

After each training, the system automatically sent the perceptual experience score and the training report. The report included the overall score (three grades: A, B, and C), completion degree, accuracy, and time for each action, and generated an exercise diary. To enhance adherence, the system also incorporated an automated reminder function that pushed notifications to prompt users to initiate their scheduled training sessions (Figure 3).

XSL•FO
RenderX

Figure 3. Postexercise evaluation and records via the "Rehabilitation Exercise Prescription" app. A: automatically generated assessment reports at the end of exercise; B: automated exercise diary recording interface.

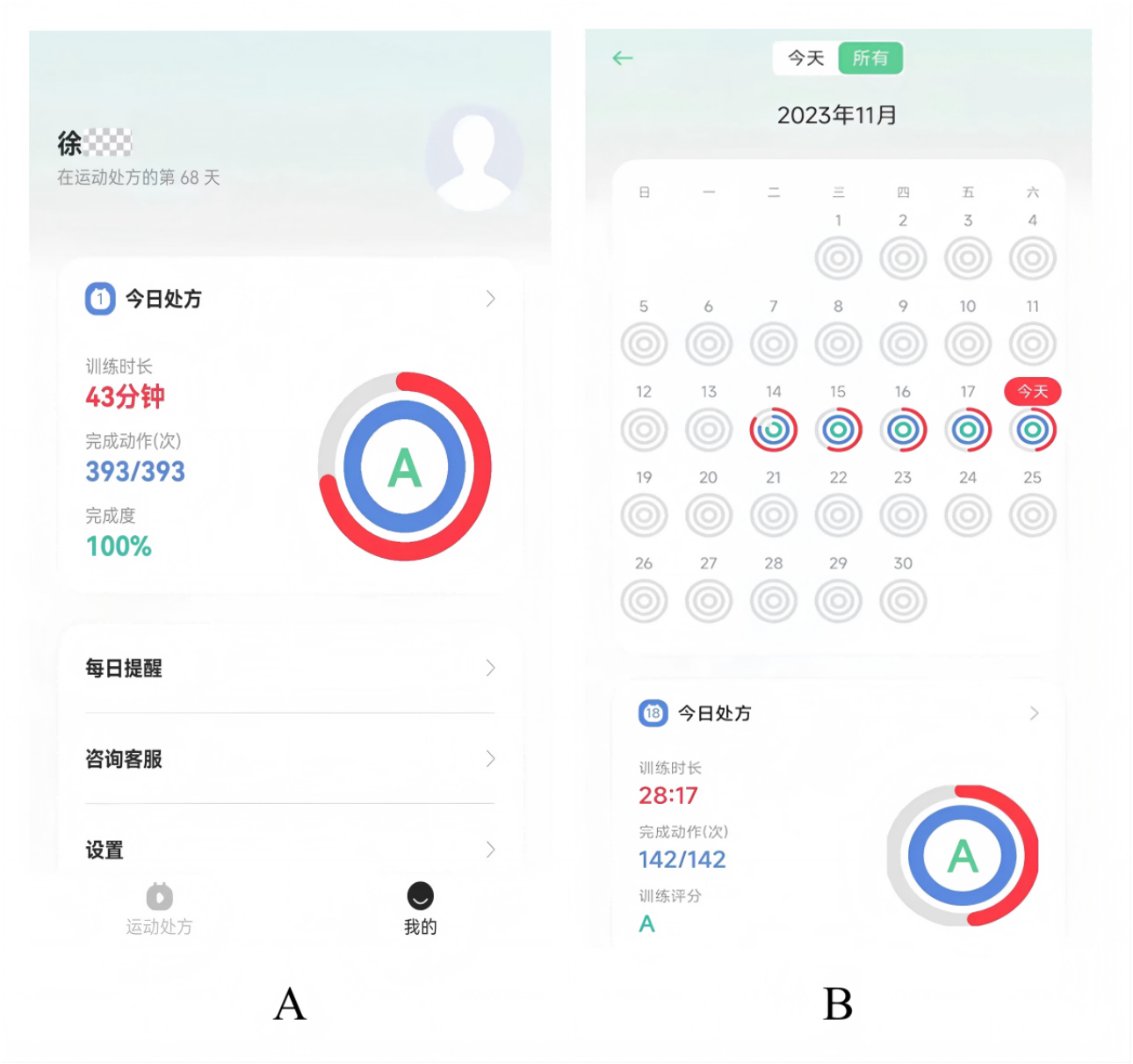

We provided immediate consultation and guidance to participants through online dialogue platforms such as WeChat or telephone follow-up, answered participants' questions during the exercise process, and adjusted the training program to suit each patient's actual situation. In addition to in-program exercise training, the suitability of higher intensity exercise for participants depended on patient feedback, remotely monitored exercise performance data, and the results of weekly online and telephone consultation feedback from the participants. The exercise program could also be adjusted, and specific exercises could be changed for individual participants if the current exercise intensity could not be achieved.

We conducted weekly online consultations or telephone follow-ups to quantify follow-up data and optimize exercise prescriptions. If the participants were unable to adapt to the current intensity, the exercise program was adjusted and replaced with a specific exercise.

Control Group

Participants in the control group received routine verbal health education on hypertension based on the current guidelines. This covered lifestyle guidance for hypertension, including dietary education emphasizing low-salt diets, balanced nutrition, alcohol restriction, adequate hydration, weight management, healthy food choices, psychological adjustment, smoking cessation guidance, sleep management, and advice on reducing disease risk and understanding medication. They also received exercise prescriptions, including warm-up, aerobics, strength, balance, and flexibility training types, intensity, and frequency, from the same movement library as the trial group. The prescription was 30 to 50 minutes per session, and patients were advised to complete it 5 times per week and required to complete it at least 3 times per week. Upon enrollment, the research team distributed the Hypertension Rehabilitation Handbook to all participants. This handbook was developed based on relevant guidelines [29-33] and covers the following topics: the definition and

XSL•FO
RenderX

classification of hypertension, common misconceptions, prevalence statistics, explanations of hazards, risk factor analysis, identification of susceptible populations, methods of early detection and diagnosis, techniques for self-BP measurement, prevention and treatment guidelines, recommendations for lifestyle adjustments, guidance on exercise rehabilitation, and guidance on medication. During the study period, control group participants only received necessary BP measurements and safety assessments at follow-up points, with no active exercise supervision or feedback on execution quality.

Statistical Analyses

All analyses were performed using SPSS statistical software (version 26.0; IBM Corp). Analyses were conducted according to the intention-to-treat principle and according to the intervention group to which participants were originally assigned, regardless of adherence to the intervention. Continuous variables are described as mean (SD) or median (IQR), and categorical variables are described as numbers (percentages). The baseline characteristics between groups were compared using independent *t* tests, Mann-Whitney U tests, or chi-square tests, as appropriate.

Missing data were handled using multiple imputations. The mechanism of missingness was evaluated using the missing completely at random (MCAR) test. The imputation procedure was performed using the fully conditional specification method, and we generated 50 imputed data sets to ensure efficiency and stability. The imputation model included all primary and secondary outcome variables, the treatment group, and auxiliary variables (eg, age and sex).

Between-group differences in primary and secondary outcomes were assessed using analysis of covariance (ANCOVA), with the 8-week measurement as the dependent variable and the baseline value as a covariate. Primary analysis was conducted on each of the 50 imputed data sets. The results were subsequently pooled to generate the final estimates, along with their 95% CIs and *P* values. The significance level was set at $P<.05$ for all statistical tests.

Results

Study Population

A total of 62 participants were included in this study, 31 each in the intervention and control groups. During the follow-up period, 2 people in the intervention group discontinued exercise intervention due to sudden illness, 1 person was unable to exercise for personal reasons, 1 person in the control group was unable to exercise for physical reasons, and 3 people were unable to adhere to exercise for personal reasons. Finally, according to the principle of intention-to-treat analysis, 31 people in the intervention group and 31 people in the control group were included (Figure 4).

Figure 4. Flowchart of the study process.

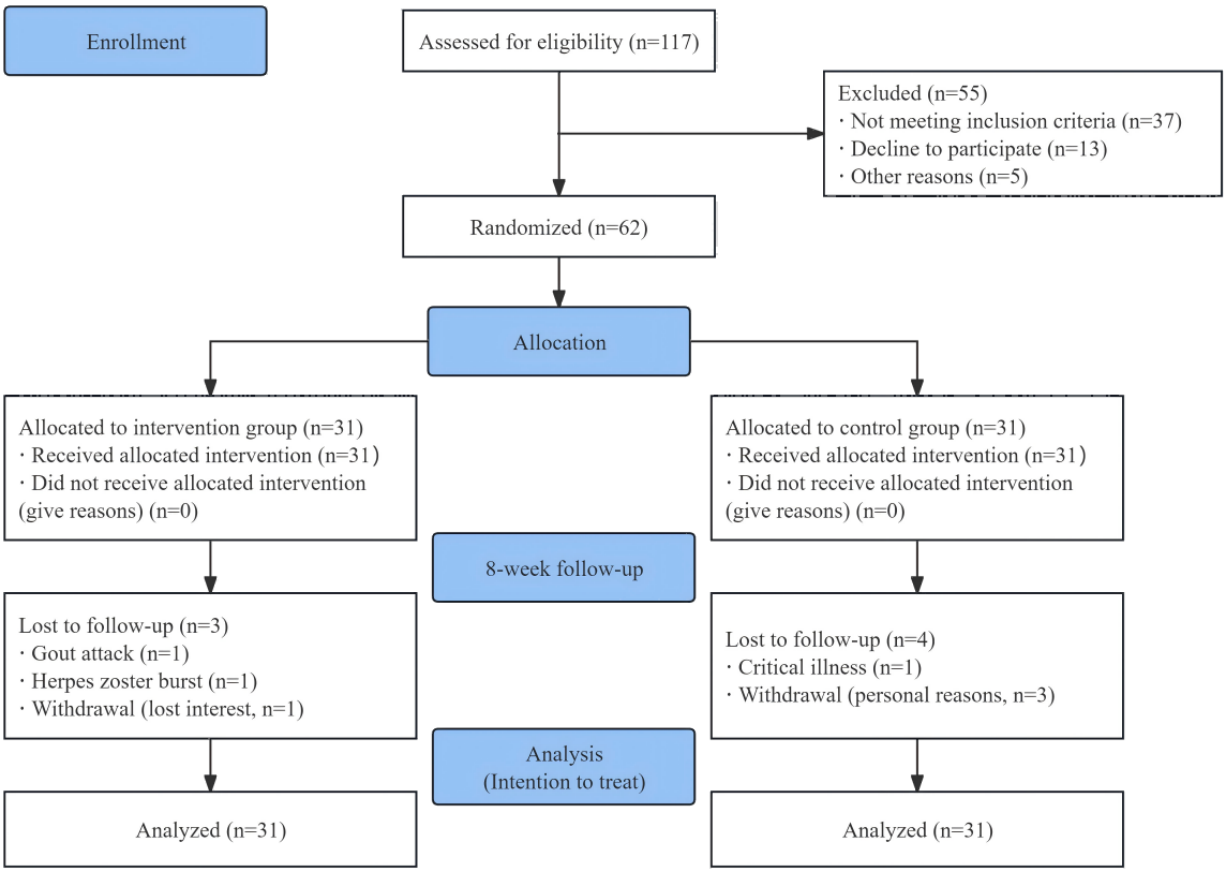

XSL•FO
RenderX

The average age of the participants overall was 52.95 (SD 11.46) years old. The average age of the intervention group was 50.48 (SD 9.44) years, and the average age of the control group was 55.42 (SD 12.86) years; there was no statistically significant difference between the 2 groups. Males accounted for 58% (n=36) and females accounted for 41.9% (n=26) of the participants. The average SBP was 128.87 (SD 12) mmHg, and the average DBP was 85.02 (SD 8.82) mmHg. The mean duration of hypertension was 7.14 (SD 7.5) years. The classification of hypertension was mainly grade 1 hypertension, accounting for 88.7% (n=55) of cases. The proportions of low-intensity, moderate-intensity, and high-intensity physical activity were 25.8% (n=16), 61.3% (n=38), and 12.9% (n=8), respectively. The results showed that the baseline data of the 2 groups were balanced, and there were no statistically significant differences (Table 1).

XSL•FO
RenderX

Table 1. Baseline characteristics of patients with hypertension in the artificial intelligence (AI)-assisted telerehabilitation randomized controlled trial (N=62).

Characteristic	All participants (n=62)	Control group (n=31)	Intervention group (n=31)	*P* value
Age (years), mean (SD)	52.95 (11.46)	55.42 (12.86)	50.48 (9.44)	.09
Gender, n (%)				.61
Male	36 (58)	17 (54.8)	19 (61.3)	
Female	26 (41.9)	14 (45.1)	12 (38.7)	
Height (m), mean (SD)	1.64 (0.07)	1.64 (0.09)	1.64 (0.06)	.77
Weight (kg), mean (SD)	65.42 (13.06)	66.32 (14.54)	64.53 (11.57)	.59
BMI (kg/m^2), mean (SD)	24.03 (3.35)	24.21 (3.37)	23.86 (3.40)	.68
Education, n (%)				.82
Primary and below	4 (6.5)	2 (6.5)	2 (6.5)	
Junior high school	9 (14.5)	3 (9.7)	6 (19.4)	
High school/technical secondary school	13 (21)	8 (25.8)	5 (16.1)	
College/bachelor's degree or above	36 (58.1)	18 (58.1)	18 (58.1)	
Smoking history, n (%)				.72
Yes	9 (14.5)	5 (16.1)	4 (12.9)	
No	53 (85.5)	26 (83.9)	27 (87.1)	
Drinking history, n (%)				.74
Yes	11 (17.7)	5 (16.1)	6 (19.4)	
No	51 (82.3)	26 (83.9)	25 (80.6)	
Operation history, n (%)				.59
Yes	42 (67.7)	22 (71)	20 (64.5)	
No	20 (32.3)	9 (29)	11 (35.5)	
Systolic BP[a] (mm/Hg), mean (SD)	128.87 (12)	129.48 (14.32)	128.26 (9.33)	.69
Diastolic BP (mm/Hg), mean (SD)	85.02 (8.82)	853 (9.60)	85 (8.13)	.99
Amount of medication, mean (SD)	2.73 (1.57)	2.62 (1.66)	2.85 (1.51)	.73
Number of comorbidities, mean (SD)	1.98 (0.96)	1.93 (0.92)	2.04 (1.02)	.74
Duration of hypertension (years), mean (SD)	7.14 (7.50)	6.57 (7.51)	7.80 (7.63)	.52
Grading of hypertension, n (%)				.14
1	55 (88.7)	28 (90.3)	27 (87.1)	
2	5 (8)	1 (3.2)	4 (12.9)	
3	2 (3.2)	2 (6.5)	0 (0)	
IPAQ[b] physical activity level, n (%)				.20
Low	16 (25.8)	7 (22.6)	9 (29)	
Middle	38 (61.3)	22 (71)	16 (51.6)	
High	8 (12.9)	2 (6.5)	6 (19.4)	

[b]BP: blood pressure.

[c]IPAQ: International Physical Activity Questionnaire.

Data Availability and Handling of Missing Data

The extent of missing data at the 8-week follow-up was detailed in Tables S1 and S2 in Multimedia Appendix 2. The proportion of missingness for the primary outcome, 6-minute walk distance (6MWD), was 6.5% (4/62). For secondary outcomes, the missingness ranged from 0% to 14.5%. A dedicated subsample

XSL•FO
RenderX

of 24 participants underwent cardiopulmonary exercise testing (CPET), among whom the missing rate for all CPET indexes was 12.5% (3/24; Table S2 in Multimedia Appendix 2). The MCAR test was performed on the data set for the entire cohort (χ^2_{596} = 557.85; *P*=.87) and separately on the CPET (χ^2_{12}=0; *P*>.99). Both results were statistically nonsignificant, which meant that the data were missing completely at random. Therefore, multiple imputation was deemed appropriate and employed to handle missing values in the primary analyses.

Comparison of Primary and Secondary Outcomes

There were no significant differences in waist circumference, hip circumference, grip strength, 6MWD, anxiety and depression, quality of life, and exercise self-efficacy scores between the 2 groups at baseline, indicating that the baseline levels of the indicators were comparable between the 2 groups before intervention.

The difference results showed that after the intervention, there were statistically significant differences in 6MWD (adjusted mean difference 62.77, 95% CI 26.33-99.22; *P*=.002), SBP reduction (adjusted mean difference 4.11, 95% CI 0.11-8.28, *P*=.046), IPAQ physical activity level (adjusted mean difference 658.96, 95% CI 159.23-1158.69; *P*=.011), exercise self-efficacy score (adjusted mean difference 21.71, 95% CI 13.59 to 29.82; *P*＜.001), between the 2 groups (Table 2).

XSL•FO
RenderX

Table 2. Comparison of primary and secondary health outcomes before and after the 8-week intervention in patients with hypertension.

Outcome	Intervention group (n=31)		Control group (n=31)		Between-group difference at 8 weeks	*P* value[a]
	Before (week 0)	After (week 8)	Before (week 0)	After (week 8)	Adjusted mean difference (95% CI)	
6MWD[b] (m), mean (SD)	472.05 (79.58)	528 (69.54)	448.57 (138.53)	449.43 (94.60)	62.77 (26.33-99.22)	*.002* [c]
IPAQ[d] (MET[e]-min/week), median (IQR)	1386 (510-2190)	1983 (1597.25-3037.50)	897 (636-1386)	1464 (1016.7-51924.50)	658.96 (159.23-1158.69)	.01
PHQ-9[f], mean (SD)	6.03 (3.91)	2.72 (3.31)	5.13 (4.49)	2.75 (1.82)	–0.14 (–1.64 to 1.36)	.85
GAD-7[g], mean (SD)	6.23 (4.38)	2.41 (2.20)	4.48 (4.21)	2.25 (1.68)	–0.23 (–1.26 to 0.80)	.65
SF-12[h], mean (SD)						
PCS[i], mean (SD)	41.38 (6.86)	47.51 (6.12)	42.64 (9.06)	47.26 (6.65)	0.66 (–2.64 to 3.96)	.69
MCS[j], mean (SD)	46.28 (10.17)	54.37 (5.20)	46.34 (10.16)	54.18 (5.04)	0.19 (–2.68 to 3.07)	.89
Exercise self-efficacy, mean (SD)	52.96 (30.02)	84.23 (15.84)	60.93 (28.37)	62.85 (12.67)	21.71 (13.59-29.82)	＜.001
Weight (kg), mean (SD)	64.53 (11.57)	63.92 (11.04)	66.32 (14.54)	66.66 (14.27)	–1.02 (–1.96 to 0.08)	.053
Girth (cm)	86.68 (9.76)	86.84 (9.80)	88.03 (9.31)	88 (9.40)	0.12 (–1.54 to 1.78)	.88
Hipline (cm)	97 (8.03)	96.48 (7.61)	98.45 (8.16)	97.71 (8.43)	–0.06 (–2.49 to 2.37)	.96
Right grip strength (kg), mean (SD)	32.87 (10.68)	35.89 (11.23)	26.49 (6.25)	27.70 (8.81)	4.39 (–4.67 to 13.44)	.33
Left grip strength (kg), mean (SD)	29.68 (10.16)	34 (11.24)	25.71 (6.27)	26.73 (9.71)	4.85 (–3.68 to 13.38)	.25
Systolic BP[k] (mm/Hg)	128.30 (9.28)	116.63 (7.64)	129.96 (15.21)	120.92 (7.15)	4.11 (0.11-8.28)	.046
Diastolic BP (mm/Hg)	84.52 (8.24)	75.56 (8.79)	85.25 (9.29)	77.71 (9.16)	1.81 (–2.72 to 6.35)	.43

[a]The significance level was set at $P<.05$.

[b]6MWD: 6-minute walk distance.

[c]The italicized values are the *P* value of the primary outcome.

[d]IPAQ: International Physical Activity Questionnaire.

[e]MET: metabolic equivalent of task.

[f]PHQ-9: Patient Health Questionnaire-9.

[g]GAD-7: Generalized Anxiety Disorder-7.

[h]SF-12: Short-Form Health Survey 12.

[i]PCS: Physical Component Summary.

[j]MCS: Mental Component Summary.

[k]BP: blood pressure.

After the intervention, the comparison of cardiopulmonary exercise test indexes between the 2 groups, total exercise time (adjusted mean difference 98.24, 95% CI 49.39-147.08; $P=.001$), peak oxygen uptake (peak VO2) (adjusted mean difference 3.39, 95% CI 0.49-6.29; $P=.026$) and peak oxygen uptake percent predicted (peak VO2%pred) (adjusted mean difference 11.58, 95% CI 2.06-21.10; $P=.021$) were statistically significant, and the exercise time of the intervention group was significantly longer than that of the control group. There was no significant difference in Watt, respiratory exchange rate, anaerobic threshold, anaerobic threshold %pred, HR rest, and HR peak between the two groups (Table 3).

XSL•FO
RenderX

Table 3. Comparison of cardiopulmonary exercise test (CPET) indices at baseline and after the 8-week intervention for both study subsample groups (N=24).

Outcome	Intervention group (n=12)		Control group (n=12)		Between-group difference at 8 weeks	*P* value[a]
	Before week 0), mean (SD)	After (week 8), mean (SD)	Before (week 0), mean (SD)	After (week 8), mean (SD)	Adjusted mean difference (95% CI)	
Exercise time (s)	472.17 (89.47)	564.29 (71.38)	481.67 (110.11)	457.63 (37.37)	98.24 (49.39-147.08)	.001
Maximum load (Watt)	112.17 (35.51)	131.71 (34.63)	97.33 (35.71)	93.35 (45.76)	17.08 (–18.29 to 52.44)	.31
Maximum load %pred	74.50 (11.29)	88.57 (8.42)	78.17 (19.72)	81.75 (12.43)	3.75 (–3.77 to 11.27)	.30
RER[b]	1.20 (0.13)	1.22 (0.09)	1.18 (0.10)	1.22 (0.11)	0.00 (–0.12 to 0.12)	.98
Peak VO_2[c] (ml/min/kg)	21.50 (4.80)	23.80 (2.89)	19.98 (4.44)	19.56 (4.35)	3.39 (0.49-6.29)	.03
Peak VO_2%pred[d]	66.75 (9.78)	82.57 (5.80)	76.25 (15.47)	73.25 (14.96)	11.58 (2.06-21.10)	.02
Anaerobic threshold (ml/min/kg)	13.63 (3.01)	14.13 (2.03)	12.70 (2.48)	12.60 (3.31)	1.54 (–1.08 to 4.16)	.23
Anaerobic threshold%pred	42.08 (7.35)	49.86 (7.90)	48.83 (10.24)	46.88 (8.92)	5.05 (–4.29 to 14.38)	.26
HR[e] at rest (bpm)	90.58 (17.90)	89.43 (13.45)	77.67 (15.71)	84.63 (17.35)	0.69 (–14.43 to 15.80)	.92
HR at peak (bpm)	155.50 (22.18)	154.29 (21.01)	134.42 (27.54)	138.13 (27.61)	–0.98 (–16.04 to 14.09)	.89

[a]The significance level was set at $P<.05$.

[b]RER: respiratory exchange rate.

[c]Peak VO_2: peak oxygen uptake.

[d]Peak VO_2%pred: peak oxygen uptake percent predicted.

[e]HR: heart rate.

The CPET data in this study were obtained from only a subset of participants. As shown in Table S3 in Multimedia Appendix 2, no statistically significant differences were observed between the 2 groups across all key baseline characteristics. This indicated that although only a subset of participants completed the CPET, they constituted a representative random sample from the overall population rather than a group with systematic bias. Throughout the 8-week intervention period, no serious adverse events related to the AI-assisted telerehabilitation intervention were reported.

Discussion

Principal Findings

The therapeutic effects of AI-assisted remote rehabilitation training on patients with hypertension were examined in this study. This study demonstrated that AI-assisted training based on 3D skeletal pose recognition significantly improved exercise capacity and BP and increased exercise self-efficacy in patients with hypertension, superior to usual care.

The Effect of AI-Assisted Remote Exercise Rehabilitation on Participants With Hypertension

Exercise rehabilitation for patients with hypertension is of great significance. It can not only effectively reduce BP level but also significantly improve cardiovascular function, thereby improving the quality of life for patients. However, there are some limitations in the current rehabilitation treatment of hypertension, such as the fact that patients frequently lack objective supervision and quality assurance during exercise implementation. This indicates that factors such as the specific execution methods of prescribed regimens, dosage standards, and the risk of muscle injury from nonstandardized exercises in home environments cannot be effectively safeguarded. This phenomenon not only significantly diminishes the effectiveness of interventions but also exacerbates issues related to individual variability. AI-based intelligent exercise rehabilitation is expected to solve these challenges.

In this study, the intervention group performed significantly better than the control group in terms of 6MWD, total exercise time, and peak VO_2 for cardiopulmonary exercise testing, indicating that AI-based remote rehabilitation has a significant effect on improving exercise endurance and cardiopulmonary function. Peak VO_2 is the core index of cardiopulmonary exercise testing, which reflects the maximum oxygen uptake capacity of patients during exercise and is an important standard to evaluate cardiopulmonary function. Although the intervention group showed significant improvement in 6MWD, the increase in peak VO_2 , although statistically significant, did not reach the minimal clinically important difference, suggesting the need for longer intervention duration or a more precise exercise prescription. For objective equipment reasons, cardiopulmonary exercise testing was performed in only 24 of 62 participants before and after the intervention. CPET data were obtained from only a subset of participants. Although baseline comparisons revealed no selection bias, future research should validate these findings across the entire sample. Nevertheless, some remarkable results were obtained. This study was similar to previous studies

XSL•FO
RenderX

on telerehabilitation in several ways. McDonagh et al [34,35] found that telecardiac rehabilitation and traditional rehabilitation had the same effect in restoring motor function, improving exercise capacity, and increasing physical activity; moreover, patients in the telecardiac rehabilitation group expressed higher satisfaction levels. There is some evidence to support greater adherence to telerehabilitation programs, which is particularly important in the context of health care crises (eg, due to pandemics) and in patients living in hard-to-reach, remote, and low- and middle-income areas. That said, the broader applicability of telerehabilitation in chronic disease management and remote health care still faces notable barriers—most prominently, difficulties in smartphone use among older adults [36]. As a key demographic for conditions like hypertension or heart disease, many older adult patients struggle with smartphone operation (eg, navigating apps, interpreting data feedback) due to limited digital literacy, sensory impairments, or tech unfamiliarity, which hinders engagement and risks widening health disparities. Therefore, future research should focus on assessing digital literacy's influence on telerehabilitation response and testing tailored strategies to mitigate these barriers, ensuring remote care benefits diverse older adult populations.

Lowering BP is a core objective in hypertension clinical research. Exercise therapy, a key adjunct to pharmacotherapy, is widely proven effective for BP reduction, with the specific magnitude dependent on the exercise program design [37-40]. In this study, the remote rehabilitation (intervention) group showed a significant SBP decrease (mean reduction >12 mmHg), while no significant change was observed in DBP, a finding that is consistent with previous studies and guidelines. The ACSM and AHA recommend moderate-intensity aerobic exercise for patients with hypertension, and existing research confirms that aerobic exercise and personalized prescriptions exert a more prominent regulatory effect on SBP [41]. The lack of significant DBP reduction may relate to patients' baseline BP, medication use, or exercise intensity. Physiologically, the significant SBP decrease can be attributed to regular exercise—particularly the "endurance and resistance" training used here—which regulates BP by improving endothelial function, reducing vascular stiffness, and mitigating oxidative stress. This aligns with findings by Guirado [42] and Kokkinos et al [43], who noted greater BP reductions with combined training. Additionally, no baseline intergroup BP differences existed, but postintervention SBP differed significantly, further validating the value of remote rehabilitation, as supported by prior research [44]. Clinically, it is meaningful that each 3 mmHg SBP reduction lowers coronary insufficiency risk by 5% to 9%, stroke risk by 8% to 14%, and all-cause mortality by 4% [45]. Notably, the impact of exercise interventions on SBP in populations with hypertension remains insufficiently understood. While exercise-induced SBP reduction is moderate but consistent, it is less pronounced than that from pharmacotherapy—although comparable to the effects of common antihypertensive drugs [37]. Future research should further explore the SBP-regulating mechanisms of exercise and assess the generalizability of such findings in real-world clinical settings.

Significant improvements in exercise self-efficacy and intervention adherence were observed with the AI-assisted remote rehabilitation program among patients with hypertension. Exercise self-efficacy is defined as an individual's confidence in their ability to successfully perform a specific physical activity. Enhanced exercise self-efficacy was achieved through multiple mechanisms. First, real-time monitoring of patient exercise data by the AI system enabled the generation of personalized feedback, allowing individuals to better understand their physiological responses and form realistic expectations about exercise outcomes. Second, a comprehensive collection of exercise demonstration videos and animated guides provided clear visual support, helping patients learn proper techniques and postures, which, in turn, boosted both skill proficiency and confidence. Third, exercise programs were dynamically adjusted using intelligent algorithms that modified intensity and difficulty based on individual health status and performance data, ensuring the intervention remained aligned with each patient's needs and enhancing overall engagement and experience. Compared with traditional face-to-face rehabilitation programs, a remote approach demonstrates higher adherence, as patients are not required to travel to hospitals or rehabilitation centers, significantly reducing time and transportation burdens and thereby increasing participation willingness. The AI system actively prompts patients to complete daily exercise tasks, establishing a closed-loop management process of "monitoring-feedback-prompting," which supports long-term engagement in physical activity and facilitates habit formation.

To our knowledge, this is the first randomized controlled trial to implement and evaluate an AI-assisted telerehabilitation system for patients with hypertension that utilizes real-time, AI-guided training based on 3D skeletal pose recognition. Our findings demonstrate that this technology can simultaneously and significantly improve objective exercise capacity, enhance psychological determinants of behavior such as self-efficacy, and aid in SBP control. This synergistic effect provides a new level of evidence, moving beyond remote monitoring to demonstrate the efficacy of interactive, form-correcting, and personalized exercise prescription. These results underscore the transformative potential of intelligent exercise rehabilitation for optimizing chronic disease management. Consequently, this study provides a foundation for developing individualized home-based rehabilitation programs, promoting sustained health behaviors, and advancing mobile health management models for hypertension and potentially other chronic conditions.

Limitations

Despite the positive results of this study, there are still some limitations. First, the study sample size was relatively small and focused on the Guangdong region in China, which may limit the generalizability of the results. Future studies could expand the sample size and validate it in different regions and populations to further confirm the effectiveness of the AI-assisted telerehabilitation program. Second, the relatively short 8-week intervention period may not capture the long-term effects of AI-driven telerehabilitation on exercise capacity and cardiovascular health. Moreover, the relatively short follow-up period in this study was limited to short-term postintervention outcome assessment. Future studies could extend the follow-up

XSL•FO
RenderX

period to assess the long-term effects and sustainability of the program. Finally, the lack of objective monitoring tools for participants' exercise intensity, such as using wearable devices to monitor heart rate to ensure training intensity and exercise safety, will be prioritized in future studies.

Conclusions

This study systematically evaluated the effects of AI-based telerehabilitation on exercise ability and related health indicators of patients with hypertension through a randomized controlled trial. The results showed that the intervention had certain effects in reducing SBP and improving exercise ability and exercise self-efficacy. Moreover, the patients' compliance was high, and the intervention process was safe and feasible. As a low-cost and accessible rehabilitation training method, the AI telerehabilitation model is expected to become a convenient and effective alternative in the future.

Acknowledgments

We would like to thank RaphAI Health Technologies Limited (Guangzhou, China) and Beijing Kangtang Medical Technology Co Ltd (Beijing, China) for their technical support. This study was supported by the National Natural Science Foundation of China (82472588), the Guangdong Natural Science Foundation of China (2025A1515012331 and 2025A1515012782), and the 2025 Guangdong Disability Cause Theory and Practice Research Project Grant. No generative AI was used in any part of the manuscript.

Data Availability

The raw data supporting the findings of this study will be made available by the corresponding author, without undue reservation.

Authors' Contributions

QZ, GH, and HM designed the study. QY, BQ, and LH drafted the manuscript. QY, LH, SW, and CL recruited participants. QW was responsible for the random assignment of participants. GL and JH were responsible for the intervention training. JZ collected the data, and DL analyzed the data. All authors agreed to the final version of the manuscript. QY, BQ, and LH contributed equally and share first authorship. HM, GH, and QZ are corresponding authors and contributed equally to this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 999 KB - jmir_v28i1e81400_app1.pdf]

Multimedia Appendix 2
Supplementary Tables.
[DOCX File , 25 KB - jmir_v28i1e81400_app2.docx]

References

1. Kearney P, Whelton M, Reynolds K, Muntner P, Whelton P, He J. Global burden of hypertension: analysis of worldwide data. Lancet 2005 Jan;365(9455):217-223 [FREE Full text] [doi: 10.1016/s0140-6736(05)17741-1]
2. Forouzanfar M, Liu P, Roth G, Ng M, Biryukov S, Marczak L, et al. Global Burden of hypertension and systolic blood pressure of at least 110 to 115 mm hg, 1990-2015. JAMA 2017 Jan 10;317(2):165-182 [FREE Full text] [doi: 10.1001/jama.2016.19043] [Medline: 28097354]
3. Kiviniemi A, Lepojärvi E, Tulppo P, Piira OP, Kenttä TV, Perkiömäki JS, et al. Prediabetes and risk for cardiac death among patients with coronary artery disease: the ARTEMIS study. Diabetes Care 2019 Jul;42(7):1319-1325 [FREE Full text] [doi: 10.2337/dc18-2549] [Medline: 31076416]
4. Lee S, Hwang S, Kang D, Yang H. Brain education-based meditation for patients with hypertension and/or type 2 diabetes: A pilot randomized controlled trial. Medicine (Baltimore) 2019 May;98(19):e15574 [FREE Full text] [doi: 10.1097/MD.0000000000015574] [Medline: 31083232]
5. Ettehad D, Emdin C, Kiran A, Anderson S, Callender T, Emberson J, et al. Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis. Lancet 2016 Mar;387(10022):957-967 [FREE Full text] [doi: 10.1016/s0140-6736(15)01225-8]
6. Burnier M. Drug adherence in hypertension. Pharmacol Res 2017 Nov;125(Pt B):142-149 [FREE Full text] [doi: 10.1016/j.phrs.2017.08.015] [Medline: 28870498]

XSL•FO
RenderX

7. Börjesson M, Onerup A, Lundqvist S, Dahlöf B. Physical activity and exercise lower blood pressure in individuals with hypertension: narrative review of 27 RCTs. Br J Sports Med 2016 Mar;50(6):356-361 [FREE Full text] [doi: 10.1136/bjsports-2015-095786] [Medline: 26787705]
8. Noone C, Leahy J, Morrissey E, Newell J, Newell M, Dwyer CP, et al. Comparative efficacy of exercise and anti-hypertensive pharmacological interventions in reducing blood pressure in people with hypertension: A network meta-analysis. Eur J Prev Cardiol 2020 Feb;27(3):247-255 [FREE Full text] [doi: 10.1177/2047487319879786] [Medline: 31615283]
9. Diaz K, Booth J, Seals S, Abdalla M, Dubbert PM, Sims M, et al. Physical activity and incident hypertension in African Americans. Hypertension 2017 Mar;69(3):421-427 [FREE Full text] [doi: 10.1161/hypertensionaha.116.08398]
10. Zhou K, Wang W, An J, Li M, Li J, Li X. Effects of progressive upper limb exercises and muscle relaxation training on upper limb function and health-related quality of life following surgery in women with breast cancer: a clinical randomized controlled trial. Ann Surg Oncol 2019 Jul;26(7):2156-2165 [FREE Full text] [doi: 10.1245/s10434-019-07305-y] [Medline: 30972655]
11. Churilla J, Ford E. Comparing physical activity patterns of hypertensive and nonhypertensive US adults. Am J Hypertens 2010 Sep;23(9):987-993 [FREE Full text] [doi: 10.1038/ajh.2010.88] [Medline: 20431526]
12. Williams N, Hendry M, France B, Lewis R, Wilkinson C. Effectiveness of exercise-referral schemes to promote physical activity in adults: systematic review. Br J Gen Pract 2007 Dec;57(545):979-986 [FREE Full text] [doi: 10.3399/096016407782604866] [Medline: 18252074]
13. Wang A, Zhang H. Efficacy of microsurgery for patients with cerebral hemorrhage secondary to gestational hypertension: a systematic review protocol of randomized controlled trial. Medicine 2019 [FREE Full text] [doi: 10.1097/md.0000000000017558]
14. Charidimou A, Turc G, Oppenheim C, Yan S, Scheitz JF, Erdur H, et al. Microbleeds, cerebral hemorrhage, and functional outcome after stroke thrombolysis. Stroke 2017 Aug;48(8):2084-2090. [doi: 10.1161/STROKEAHA.116.012992] [Medline: 28720659]
15. Ruilope L, Valenzuela P, Lucia A. Digital therapeutics and lifestyle: the start of a new era in the management of arterial hypertension? Eur Heart J 2021 Oct 21;42(40):4123-4125 [FREE Full text] [doi: 10.1093/eurheartj/ehab694] [Medline: 34571529]
16. Kario K. Digital hypertension towards to the anticipation medicine. Hypertens Res 2023 Nov;46(11):2503-2512 [FREE Full text] [doi: 10.1038/s41440-023-01409-5] [Medline: 37612370]
17. Nomura A, Tanigawa T, Kario K, Igarashi A. Cost-effectiveness of digital therapeutics for essential hypertension. Hypertens Res 2022 Oct;45(10):1538-1548 [FREE Full text] [doi: 10.1038/s41440-022-00952-x] [Medline: 35726085]
18. Kario K, Nomura A, Harada N, Okura A, Nakagawa K, Tanigawa T, et al. Efficacy of a digital therapeutics system in the management of essential hypertension: the HERB-DH1 pivotal trial. Eur Heart J 2021 Oct 21;42(40):4111-4122 [FREE Full text] [doi: 10.1093/eurheartj/ehab559] [Medline: 34455443]
19. Alessa T, Hawley M, Hock E, de Witte L. Smartphone apps to support self-management of hypertension: review and content analysis. JMIR Mhealth Uhealth 2019 May 28;7(5):e13645 [FREE Full text] [doi: 10.2196/13645] [Medline: 31140434]
20. McLean G, Band R, Saunderson K, Hanlon P, Murray E, Little P, et al. Digital interventions to promote self-management in adults with hypertension systematic review and meta-analysis. J Hypertens 2016:600-612 [FREE Full text] [doi: 10.1097/hjh.0000000000000859]
21. Maskeliūnas R, Kulikajevas A, Damaševičius R, Griškevičius J, Adomavičienė A. Biomac3D: 2D-to-3D human pose analysis model for tele-rehabilitation based on pareto optimized deep-learning architecture. Applied Sciences 2023 Jan 13;13(2):1116 [FREE Full text] [doi: 10.3390/app13021116]
22. Yang G, He S, Meng D, Wei M, Cao J, Guo H, et al. Body landmarks and genetic algorithm-based approach for non-contact detection of head forward posture among Chinese adolescents: revitalizing machine learning in medicine. BMC Med Inform Decis Mak 2023 Sep 11;23(1):179 [FREE Full text] [doi: 10.1186/s12911-023-02285-2] [Medline: 37697312]
23. Wei M, Meng D, He S, Lv Z, Guo H, Yang G, et al. Investigating the efficacy of AI-enhanced telerehabilitation in sarcopenic older individuals. Eur Geriatr Med 2025 Feb;16(1):115-123 [FREE Full text] [doi: 10.1007/s41999-024-01082-y] [Medline: 39453567]
24. Xiao C, Zhao Y, Li G, Zhang Z, Liu S, Fan W, et al. Clinical efficacy of multimodal exercise telerehabilitation based on AI for chronic nonspecific low back pain: randomized controlled trial. JMIR Mhealth Uhealth 2025 May 22;13:e56176 [FREE Full text] [doi: 10.2196/56176] [Medline: 40402551]
25. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340(mar23 1):c869-c869 [FREE Full text] [doi: 10.1136/bmj.c869] [Medline: 20332511]
26. Chan A, Tetzlaff J, Altman D, Laupacis A, Gøtzsche PC, Krleža-Jerić K, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. Ann Intern Med 2013 Feb 05;158(3):200-207 [FREE Full text] [doi: 10.7326/0003-4819-158-3-201302050-00583] [Medline: 23295957]
27. Piotrowicz E, Zieliński T, Bodalski R, Rywik T, Dobraszkiewicz-Wasilewska B, Sobieszczańska-Małek M, et al. Home-based telemonitored Nordic walking training is well accepted, safe, effective and has high adherence among heart failure patients,

XSL•FO
RenderX

including those with cardiovascular implantable electronic devices: a randomised controlled study. Eur J Prev Cardiol 2015 Nov;22(11):1368-1377 [FREE Full text] [doi: 10.1177/2047487314551537] [Medline: 25261268]
28. Butāne L, Spilva-Ekerte L, Šablinskis M, Skride A, Šmite D. Individually tailored home-based physiotherapy program makes sustainable improvement in exercise capacity and daily physical activity in patients with pulmonary arterial hypertension. Ther Adv Respir Dis 2022;16:17534666221132477 [FREE Full text] [doi: 10.1177/17534666221132477] [Medline: 36314474]
29. World Health Organization. Global Recommendations on Physical Activity for Health. Geneva: World Health Organization; 2010.
30. Thompson P, Arena R, Riebe D, Pescatello L. ACSM's new preparticipation health screening recommendations from ACSM's guidelines for exercise testing and prescription, ninth edition. Curr Sports Med Rep 2013;12(4):215-217 [FREE Full text] [doi: 10.1249/jsr.0b013e31829a68cf]
31. Thompson P, Buchner D, Pin˜a I, Balady G, Williams M, Marcus B, et al. Exercise and physical activity in the prevention and treatment of atherosclerotic cardiovascular disease. Circulation 2003 Jun 24;107(24):3109-3116 [FREE Full text] [doi: 10.1161/01.cir.0000075572.40158.77]
32. Balady G, Williams M, Ades P, Bittner V, Comoss P, Foody J, et al. Core components of cardiac rehabilitation/secondary prevention programs: 2007 update. Circulation 2007 May 22;115(20):2675-2682 [FREE Full text] [doi: 10.1161/circulationaha.106.180945]
33. Bairey Merz CN, Alberts M, Balady G, Ballantyne CM, Berra K, Black HR, et al. ACCF/AHA/ACP 2009 Competence and Training statement: a curriculum on prevention of cardiovascular disease. Circulation 2009 Sep 29;120(13):A [FREE Full text] [doi: 10.1161/circulationaha.109.192640]
34. McDonagh S, Dalal H, Moore S, Clark C, Dean S, Jolly K, et al. Home-based versus centre-based cardiac rehabilitation. Cochrane Db Syst Rev 2023;10(10) [FREE Full text] [doi: 10.1002/14651858.cd007130.pub5]
35. McDonagh S, Dalal H, Moore S, Clark C, Taylor R. Cochrane corner: centre versus telemedicine approaches to cardiac rehabilitation. Heart 2023 Dec 15;110(1):7-10 [FREE Full text] [doi: 10.1136/heartjnl-2023-322640] [Medline: 37487697]
36. Shibata S, Hoshide S. Current situation of telemedicine research for cardiovascular risk in Japan. Hypertens Res 2023 May;46(5):1171-1180 [FREE Full text] [doi: 10.1038/s41440-023-01224-y] [Medline: 36849580]
37. Naci H, Salcher-Konrad M, Dias S, Blum M, Sahoo S, Nunan D, et al. How does exercise treatment compare with antihypertensive medications? A network meta-analysis of 391 randomised controlled trials assessing exercise and medication effects on systolic blood pressure. Br J Sports Med 2019 Jul;53(14):859-869 [FREE Full text] [doi: 10.1136/bjsports-2018-099921] [Medline: 30563873]
38. Oliveira-Dantas F, Brasileiro-Santos M, Thomas S, Silva A, Silva D, Browne R, et al. Short-term resistance training improves cardiac autonomic modulationblood pressure in hypertensive older women: a randomized controlled trial. J Strength Cond Res 2020 [FREE Full text] [doi: 10.1519/jsc.0000000000003182]
39. Ruangthai R, Phoemsapthawee J. Combined exercise training improves blood pressure and antioxidant capacity in elderly individuals with hypertension. J Exerc Sci Fit 2019 Jan 20;17(2) [FREE Full text] [doi: 10.1016/j.jesf.2019.03.001] [Medline: 30949214]
40. Saco‐Ledo G, Valenzuela P, Ruiz‐Hurtado G, Ruilope L, Lucia A. Exercise reduces ambulatory blood pressure in patients with hypertension: a systematic review and meta‐analysis of randomized controlled trials. JAHA 2020 Dec 15;9(24):A [FREE Full text] [doi: 10.1161/jaha.120.018487]
41. Herrod P, Lund J, Phillips B. Time-efficient physical activity interventions to reduce blood pressure in older adults: a randomised controlled trial. Age Ageing 2021 May 05;50(3):980-984 [FREE Full text] [doi: 10.1093/ageing/afaa211] [Medline: 33068100]
42. Guirado GN, Damatto RL, Matsubara BB, Roscani MG, Fusco DR, Cicchetto LAF, et al. Combined exercise training in asymptomatic elderly with controlled hypertension: effects on functional capacity and cardiac diastolic function. Med Sci Monit 2012 Jul;18(7):CR461-CR465 [FREE Full text] [doi: 10.12659/msm.883215] [Medline: 22739737]
43. Kokkinos P, Narayan P, Colleran J, Pittaras A, Notargiacomo A, Reda D, et al. Effects of regular exercise on blood pressure and left ventricular hypertrophy in African American men with severe hypertension. N Engl J Med 1995 Nov 30;333(22):1462-1467 [FREE Full text] [doi: 10.1056/nejm199511303332204]
44. Rodrigues-da-Silva A, Suassuna J, Monteiro E, Borges de Lima I, Santos A, Brasileiro-Santos M. Effects of telerehabilitation on cardiac remodeling and hemodynamics parameters in hypertensive older adults: A randomized controlled trial. J Telemed Telecare 2024 Mar 14;31(7):1005-1013 [FREE Full text] [doi: 10.1177/1357633x241236572]
45. Pescatello L, Franklin B, Fagard R, Farquhar W, Kelley G, Ray C, American College of Sports Medicine. American College of Sports Medicine position stand. Exercise and hypertension. Med Sci Sports Exerc 2004 Mar;36(3):533-553 [FREE Full text] [doi: 10.1249/01.mss.0000115224.88514.3a] [Medline: 15076798]

Abbreviations

6MWD: 6-minute walk distance
6MWT: 6-minute walk test

XSL•FO
RenderX

ACSM: American College of Sports Medicine
AHA: American Heart Association
ANCOVA: analysis of covariance
BP: blood pressure
CONSORT: Consolidated Standards of Reporting Trials
CPET: cardiopulmonary exercise testing
DBP: diastolic blood pressure
HERB-DH1: HERB Digital Hypertension 1
MCAR: missing completely at random
PAR-Q: Physical Activity Readiness Questionnaire
peak VO2%pred: peak oxygen uptake percent predicted
peak VO2: peak oxygen uptake
SBP: systolic blood pressure
SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials
WHO: World Health Organization

Edited by S Brini; submitted 28.Jul.2025; peer-reviewed by L Zhang, A Higaki; comments to author 21.Sep.2025; accepted 11.Nov.2025; published 13.Jan.2026.

Please cite as:
Yao Q, Qiu B, He L, Wang Q, Zou J, Liang D, Wen S, Liu Y, Li G, Hu J, Ma H, Huang G, Zeng Q
Effects of Artificial Intelligence Recognition–Based Telerehabilitation on Exercise Capacity in Patients With Hypertension: Randomized Controlled Trial
J Med Internet Res 2026;28:e81400
URL: https://www.jmir.org/2026/1/e81400
doi:10.2196/81400
PMID:

©Qiuru Yao, Baizhi Qiu, Longlong He, Qin Wang, Jihua Zou, Donghui Liang, Shuyang Wen, Yingchao Liu, Gege Li, Jinjing Hu, Huan Ma, Guozhi Huang, Qing Zeng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Engagement Significantly Enhances Weight Loss Outcomes in Adults With Obesity Treated With Tirzepatide: Retrospective Cohort Study of a Digital Weight Loss Service

Hans Johnson[1,2,3], MSc, MRes, MBChB; Ashley Kieran Clift[2,4], MA, MBBS, DPhil; Daniel Reisel[2,5], MBBS, DPhil, MRCOG; David Huang[2], BSc, MBBS

[1]Engineering and Physical Sciences Research Council Centre for Doctorate Training in Digital Health and Care, University of Bristol, Bristol, United Kingdom

[2]Department of Clinical Innovation and Research, Voy (T/a Menwell LTD), London, United Kingdom

[3]Department of Education, University of Oxford, Oxford, United Kingdom

[4]Department of Surgery and Cancer, Imperial College London, London, United Kingdom

[5]Elizabeth Garrett Anderson Institute for Women's Health, University College London, London, United Kingdom

Corresponding Author:
Hans Johnson, MSc, MRes, MBChB
Engineering and Physical Sciences Research Council Centre for Doctorate Training in Digital Health and Care
University of Bristol
1 Cathedral Square, Trinity Street
Bristol, BS1 5DD
United Kingdom
Phone: 44 7916948134
Fax: 44 117 42 82343
Email: hj14789@bristol.ac.uk

Abstract

Background: The advent of tirzepatide has transformed obesity care; yet, real-world weight loss outcomes necessarily depend on patient engagement with behavioral support. Digital platforms offering coaching, self-monitoring, and automated feedback have the potential to further augment pharmacological efficacy.

Objective: The aim of the study is to examine associations between digital engagement and weight loss outcomes among adults prescribed tirzepatide in routine care over 12 months and to identify baseline correlates of engagement.

Methods: In this retrospective cohort study, we included adults (18-75 years; BMI ≥30 or ≥27.5 kg/m^2 with comorbidities) who initiated tirzepatide between February 2024 and August 2025 via a UK digital weight loss service. Engagement was defined by all 3: attendance at ≥1 coaching session AND ≥1 weekly weight log AND ≥1 app login over 12 months. Percent weight loss was analyzed at months 2, 4, 6, 8, 10, and 12 using a mixed model repeated measures adjusted for age, sex, baseline BMI, and comorbidities. Time-to-event analyses (Kaplan-Meier) assessed attainment of ≥5%, ≥10%, ≥15%, and ≥20% weight loss thresholds. Multivariable logistic regression identified correlates of engagement, reporting odds ratios (ORs) per decade of age and per 5 kg/m^2 BMI.

Results: Among 126,553 participants, 6746 (5.3%) were maximally engaged. Cohort demographics were a mean age of 42.3 (SD 12.4) years, 78.9% (99,905/126,553) female, and a mean BMI of 35.3 (SD 6.2) kg/m^2. Engaged users achieved greater adjusted weight loss at month 12 (–22.9%, 95% CI –23.2 to –22.6) versus nonengaged users (–17.5%, 95% CI –17.7 to –17.4), an absolute difference of 5.3 percentage points (P<.001; Cohen d=0.54). Differences emerged by month 2 (–7.4% vs –6.4%; P<.001) and widened steadily. Engaged participants reached all clinically significant weight loss thresholds faster (5%-20%; log-rank P<.001), and engaged participants were nearly 3 times more likely to achieve ≥20% weight loss compared to nonengaged participants (1079/6746, 16% vs 6710/119,807, 5.6%; risk ratio 2.88; P<.001). Older age (OR 1.18 per decade, 95% CI 1.15-1.20; P<.001), higher BMI (OR 1.14 per 5 kg/m^2, 95% CI 1.12-1.16; P<.001), and the presence of polycystic ovary syndrome (OR 1.59, 95% CI 1.45-1.74; P<.001) or fatty liver disease (OR 1.52, 95% CI 1.32-1.76; P<.001) correlated with engagement. Male sex (OR 0.86, 95% CI 0.81-0.92; P<.001) and diabetes (OR 0.83, 95% CI 0.73-0.95; P=.009) were associated with lower engagement.

XSL•FO
RenderX

Conclusions: Digital engagement was associated with substantially greater tirzepatide-associated weight loss in real-world practice. Integrating structured digital support with pharmacotherapy represents a promising strategy for optimizing obesity management.

(J Med Internet Res 2026;28:e83718) doi:10.2196/83718

KEYWORDS

obesity; weight loss; tirzepatide; digital health; engagement; behavior; coaching; retrospective study; GIP/GLP-1 RA; glucose-dependent insulinotropic polypeptide and glucagon-like peptide-1 receptor agonist

Introduction

Background

The global prevalence of obesity has tripled since 1975, and the economic burden of obesity and overweight is projected to reach 3.3% of global gross domestic product by 2060 [1]. Beyond economic considerations, obesity substantially increases risks for numerous chronic conditions including type 2 diabetes, cardiovascular disease, certain cancers, as well as overall mortality. Despite recognition of obesity as a complex chronic disease requiring multifaceted interventions, historically available treatments have shown limited long-term efficacy [2].

Recent advancements in pharmacotherapy have transformed the obesity treatment paradigm. Tirzepatide (Mounjaro) is a novel dual glucose-dependent insulinotropic polypeptide (GIP) and glucagon-like peptide-1 (GLP-1) receptor agonist (RA) that has demonstrated remarkable efficacy for weight management in randomized controlled trials (RCTs) [3]. The SURMOUNT-1 trial reported mean weight reductions of 15% to 20.9% at 72 weeks, significantly surpassing previous pharmacotherapies [4]. These results have generated substantial clinical interest; yet, questions remain regarding their generalization to real-world settings and the scope for adjuncts to support or optimize outcomes.

Digital weight loss services (DWLSs) combine pharmacotherapy with app-based tools, regular weight tracking, and professional health education and tailor nutrition-driven coaching. These services may address critical barriers to obesity treatment, including limited health care provider time, accessibility challenges, and inadequate behavioral support [5-7]. Understanding how digital engagement impacts pharmacotherapy outcomes could critically inform clinical recommendations and health service design. Furthermore, digital health interventions represent a promising complement to pharmacotherapy. Recent systematic reviews indicate that technology-based approaches can enhance weight loss outcomes by improving self-monitoring, accountability, personalized feedback, and remote support [8-10]. Studies of remotely delivered weight loss services have shown the efficacy of such digital intervention [11]. However, apart from a few recent studies [5], there is limited evidence regarding how digital engagement influences outcomes specifically in patients using dual GLP-1 and GIP RA for weight management.

Objectives

This study conducted a retrospective evaluation of the Voy DWLS, which is available throughout the United Kingdom. Our objectives were to characterize the real-world effectiveness of tirzepatide on weight loss, explore the associations between digital engagement and outcomes, and identify factors that correlate with engagement.

Methods

Study Design and Setting

This retrospective study used an open cohort approach. The study period spanned early February 2024 to early August 2025, with patients contributing follow-up from the date of first prescription to the earliest date of last weight measurement logged or reached 12 months of follow-up. The Voy digital health platform, a commercial telehealth DWLS for obesity management, was developed by a multidisciplinary team of clinicians, behavioral scientists, and software developers to provide remote behavioral support through live group video coaching sessions, text-based in-app support, dynamic educational content, and the direct supply of tirzepatide for weight management. Drawing upon established frameworks in behavior change and self-management support, the platform combined core digital tools (eg, real-time weight monitoring, medication adherence tracking, and personalized coaching sessions) into a single interface accessible via smartphone or web browser.

Procedure

Participants became aware of the Voy program through multiple channels: targeted social media campaigns, clinician referrals, word-of-mouth recommendations, and general web searches. They self-enrolled and self-paid via the Voy website, where they completed an online screening questionnaire covering medical history, BMI, and lifestyle factors. Thereafter, the participants interacted with the Voy digital health platform's weight management program, which integrates dual GLP-1 and GIP RA pharmacotherapies with digital behavioral support to enhance weight loss outcomes. Subsequently, participants self-enrolled via the Voy website, where they completed an online screening questionnaire covering medical history, BMI, and a range of lifestyle factors. Upon enrollment, participants underwent an initial assessment to confirm eligibility, including verification of age, BMI, and the absence of exclusion criteria (Figure 1). All participants underwent comprehensive clinical screening before treatment initiation. Structured safeguards ensured accuracy and safe prescribing: photo identification and full-body photographs for identity and BMI verification; detailed clinical questionnaires capturing medical history, contraindications, and current medications; and individual review by qualified licensed clinical prescribers with cross-checking for clinical red flags. The service operates under

XSL•FO
RenderX

Care Quality Commission registration with internal clinical audit processes, escalation protocols for safety concerns, and regular safety reviews. This approach adheres to General Medical Council and General Pharmaceutical Council standards for remote prescribing in the United Kingdom, with clinician-led prescribing decisions and independent verification of eligibility criteria to ensure patient safety.

The initial monthly cost of enrollment for tirzepatide was US $273.72. Participants receive tirzepatide (Mounjaro) via multidose prefilled KwikPen containing 4 weekly doses. Administration is a subcutaneous injection once weekly, with doses ranging from 2.5 to 15 mg. The starting dose is 2.5 mg, increased every 4 weeks in 2.5 mg increments as tolerability permits. New tirzepatide pens are prescribed every 28 days. Monthly costs cover medication, clinical support from qualified prescribing clinicians, personalized coaching from registered dietitians and nutritionists, and access to digital tracking tools and educational content (see Multimedia Appendix 1 for detailed program description). Although online weight tracking was mandatory for continued prescriptions, every other component of the Voy platform was optional.

Participants attended group onboarding sessions and were offered fortnightly coaching to enhance engagement and adherence. Coaches were trained based on principles from social cognitive theory, self-determination theory, the transtheoretical model, and the theory of planned behavior [12,13]. These techniques focused on fostering intrinsic motivation, goal setting, and problem-solving to promote sustainable lifestyle changes tailored to participants' individual progress.

Figure 1. Flowchart illustrating the critical steps of the cohort retrospective analysis conducted for the Voy digital weight management program. GIP: glucose-dependent insulinotropic polypeptide; GLP-1: glucagon-like peptide-1; RA: receptor agonist.

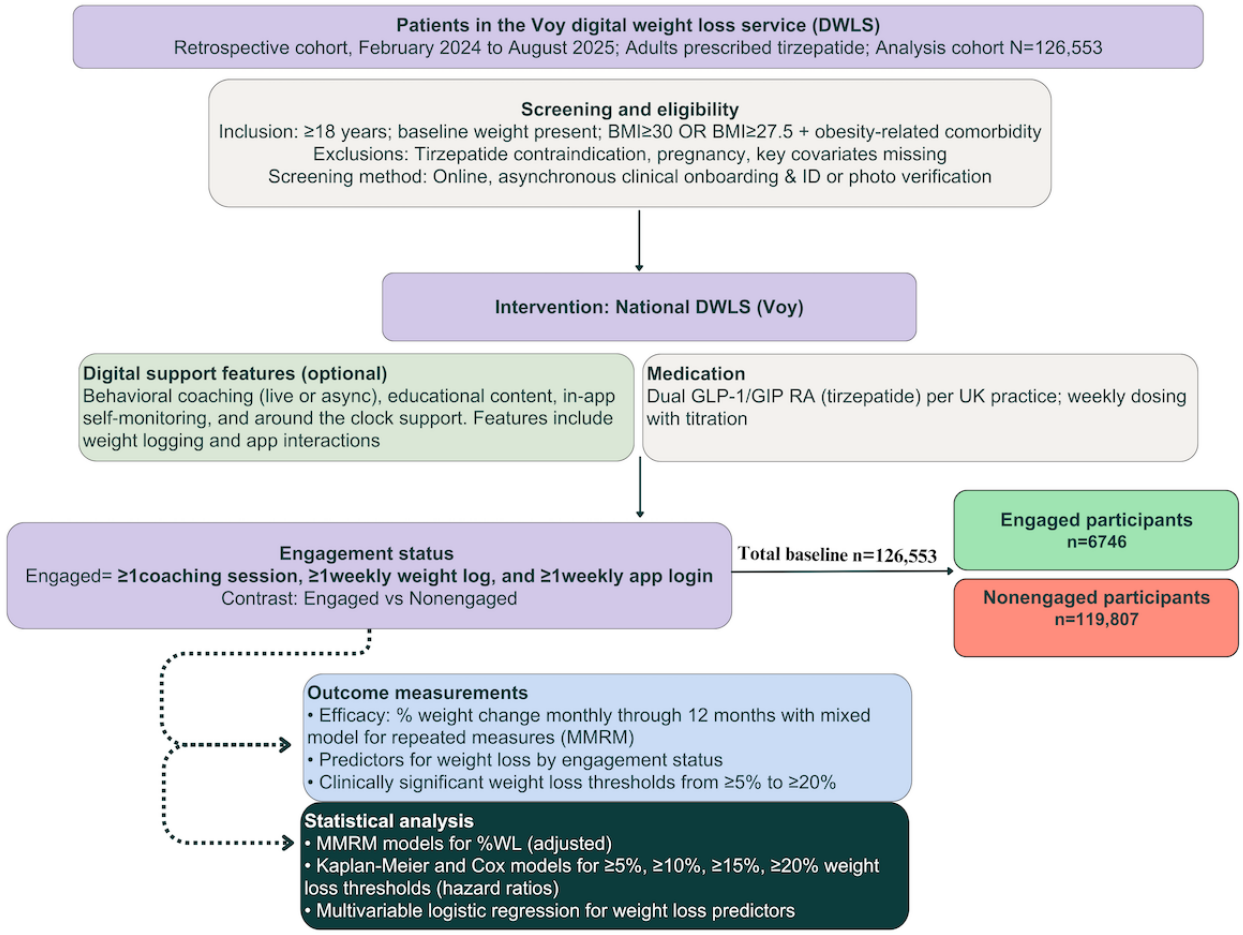

Participants

Participants were adult residents of the United Kingdom, aged between 18 and 75 years, with a BMI of 30 kg/m^2 or higher or >27.5 kg/m^2 with obesity-related comorbidities who were enrolled into the Voy digital weight loss service and were prescribed tirzepatide.

Eligibility Criteria

Eligibility required access to a smartphone or tablet. Exclusion criteria included history of self-reported eating disorders (eg, anorexia nervosa and bulimia nervosa), pregnancy or patients seeking fertility, known allergies or hypersensitivity to any components of tirzepatide, and severe medical conditions with established contraindications for tirzepatide as per Multimedia Appendix 2.

Data Completeness

Engagement metrics and weight loss outcomes were available for the entire sample of participants who used the digital platform during the study period at baseline. Complete demographic information (eg, age, sex, and comorbidities) was

XSL•FO
RenderX

available for the entire sample participants who enrolled during the initial phase of the service.

Defining Engagement and Outcome

Primary Outcome

The primary outcome was percentage weight loss from baseline. Weight measurements were entered directly in the app by participants.

Secondary Outcomes of Interest Included: Digital Engagement as an Exposure Variable

Engagement was defined based on a patient engaging in all 3 key behaviors identified as a priority by the Voy clinical and research teams as likely associated with weight loss outcomes. These behaviors are as follows:

- Coaching session attendance: participation in group or individual coaching sessions (video, audio, or text-based).
- Weight-tracking frequency: regular logging of body weight in the app.
- App use and logins: logging into the platform to view educational content, track health metrics, or interact with coaches.

As in our prior work [5], participants were classified as "engaged" if they met all of the following criteria: attended ≥1 coaching session AND tracked weight ≥1 per week AND logged into the app at least once during the study period. Those who did not fulfill all 3 of these criteria as a compound were classified as "nonengaged."

Statistical Analysis

Longitudinal Weight Loss Analysis

Our primary analysis used a mixed model for repeated measures (MMRM) approach, informed by the statistical methodology of the SURMOUNT-1 RCT study [4]. This approach accounted for repeated weight measurements for individuals, varying follow-up duration, and implicitly handled missing data without imputation under the missing at random assumption. The model incorporated fixed effects for digital engagement status, months, and the interaction between engagement and month. To adjust for potential confounding variables, we included sex, age, baseline BMI, and comorbid conditions as covariates (see Multimedia Appendix 3 for detailed methodology).

Sensitivity Analyses for Follow-Up Duration

To address the rolling enrollment design, we conducted sensitivity analyses restricted to subcohorts reaching specific follow-up milestones (3, 6, 9, and 12 months). Linear regression models were fitted for each subcohort with percentage weight change from baseline as the dependent variable, adjusted for the same confounders as the MMRM analyses. Marginal estimates (least squares means) were calculated from these models, with contrast *P* values comparing engaged versus nonengaged groups. These analyses assessed whether associations between engagement and weight loss remained consistent across different follow-up durations.

Postmodel Analyses and Effect Size Calculation

For postmodel analyses, we calculated adjusted means and conducted pairwise comparisons between engagement groups using the *emmeans* package. Significance testing of the engagement effect at each time point used 2-sided tests with α=.05, without adjustment for multiple comparisons, consistent with standard approaches for longitudinal repeated measures designs.

To quantify the magnitude and clinical relevance of the engagement effect, we calculated effect sizes using multiple approaches. Cohen *d* values were computed using the estimated difference divided by the pooled SD. Additionally, we computed both absolute differences (in percentage points) and relative differences (as a percentage of the nonengaged group's weight loss) between engaged and nonengaged tirzepatide users. To examine whether the impact of engagement on weight loss varied over time, we conducted a formal test of the engagement by time interaction.

Data Processing and Descriptive Statistics

All data processing and analyses were conducted using R software (version 4.3.1; R Foundation for Statistical Computing). Baseline characteristics were summarized using means and SDs for continuous variables, while frequencies and percentages were calculated for categorical variables. These statistics were calculated separately for engaged and nonengaged groups.

Baseline Correlates of Digital Engagement

Baseline characteristics associated with digital engagement were identified using multivariable logistic regression, where the dependent variable was digital engagement status (engaged vs not engaged, defined as mentioned earlier). Independent variables included demographic characteristics (age and sex), baseline BMI, and documented comorbidities (diabetes, hypertension, hypercholesterolemia, polycystic ovary syndrome [PCOS], and fatty liver disease). Age was modeled as a continuous variable, with additional transformations to express odds ratios (ORs) per decade increase. Similarly, BMI was analyzed with secondary calculations to express ORs per 5 kg/m^2 increase.

Achievement of Clinically Significant Weight Loss

Kaplan-Meier methods were used to estimate the proportions of users attaining ≥5%, ≥10%, ≥15%, and ≥20% weight loss during the study period with 95% CIs. This was performed for the cohort overall, then comparing engaged and nonengaged groups. Log-rank tests were used to compare cumulative incidence curves. Hazard ratios (HRs) with 95% CIs were calculated to quantify the relative rate of achieving weight loss thresholds between engagement groups, and risk ratios (RRs) were computed to assess the relative probability of threshold achievement.

Sample Size

A power analysis determined that a minimum of 118 participants per engagement group would provide 80% power to detect a 15% difference in the proportion of participants achieving ≥10% weight loss at a 5% significance level.

XSL•FO
RenderX

Bias and Missing Data

Self-reported weight measurements could introduce reporting bias. To mitigate this, participants were encouraged to provide accurate measurements through regular reminders and had the option to upload progress photographs, enhancing data validity. Additionally, data validation checks were performed internally and by statisticians to identify and address implausible values. Selection bias was minimized by including all eligible participants who initiated the program within the study period, ensuring the sample was representative of the population using the DWLS.

Missing data were addressed through the MMRM approach, which uses all available observations without requiring complete cases. This method operates under the missing at random assumption, whereby missingness may depend on observed covariates and previous measurements but not on unobserved values conditional on observed data. MMRM provides unbiased parameter estimates and maintains statistical efficiency in the presence of intermittent missing data, consistent with statistical methodology for landmark obesity trials [4]. For time-to-event analyses, participants who remained on treatment were censored at their last prescription date or study end, with censoring appropriately incorporated into Kaplan-Meier and Cox regression models.

Ethical Considerations

This retrospective open cohort study was an analysis of deidentified data collected during the routine clinical care of adults treated by Voy. The study was approved by the University College London Research Ethics Committee (Project ID 2025-0906-775). The study adhered to the principles outlined in the Declaration of Helsinki. Participants provided informed consent for their anonymized data to be used for ethically approved research and service improvement purposes upon enrollment in the program. To protect participant privacy, all data were deidentified prior to analysis, with direct identifiers (including names, contact details, and unique patient identifiers) removed and replaced with study-specific codes. Data were stored on secure servers with access restricted to authorized research personnel only. Participants did not receive any compensation for their participation in this study, as the analysis was conducted retrospectively using data collected during routine clinical care.

Adherence to STROBE Guidelines

This study adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [14] (Multimedia Appendix 4).

Results

Participant Characteristics

Among 126,553 participants who initiated tirzepatide, 6746 (5.3%) met criteria for maximally digital engagement. Baseline characteristics stratified by engagement status are presented in Table 1. The overall cohort had a mean age of 42.3 (SD 12.4) years, was predominantly female (99,905/126,553, 78.9%), and had a mean baseline BMI of 35.3 (SD 6.2) kg/m^2. Compared to nonengaged participants, digitally engaged individuals were significantly older (engaged: mean 44.9, SD 12.0 years vs nonengaged: mean 42.2, SD 12.5 years; P<.001), more likely to be female (engaged: 5486/6746, 81.3% vs nonengaged: 94,419/119,807, 78.8%; P<.001), and had higher baseline BMI (engaged: mean 36.4, SD 6.7 kg/m^2 vs nonengaged: mean 35.2, SD 6.2 kg/m^2; P<.001). Engaged participants also reported higher prevalence of comorbidities, including type 2 diabetes (engaged: 250/6746, 3.7% vs nonengaged: 3681/119,807, 3.1%; P<.001), hypertension (engaged: 978/6746, 14.5% vs nonengaged: 11,592/119,807, 9.7%; P<.001), hypercholesterolemia (engaged: 707/6746, 10.5% vs nonengaged: 7738/119,807, 6.5%; P<.001), and PCOS (engaged: 591/6746, 8.8% vs nonengaged: 7285/119,807, 6.1% of female participants; P<.001).

XSL•FO
RenderX

Table 1. Baseline demographic and clinical characteristics of participants by digital engagement status.

Characteristic	Overall (N=126,553)	Digitally engaged (n=6746)	Not digitally engaged (n=119,807)	*P* value[a]
Age (years)				<.001
Mean (SD)	42.3 (12.4)	44.9 (12.0)	42.2 (12.5)	
Age group (years), n (%)				<.001
18-24	6892 (5.4)	120 (1.8)	6772 (5.7)	
25-34	32,037 (25.3)	1399 (20.7)	30,638 (25.6)	
35-44	36,334 (28.7)	1977 (29.3)	34,357 (28.7)	
45-54	27,269 (21.5)	1643 (24.4)	25,626 (21.4)	
55+	24,016 (19)	1607 (23.8)	22,409 (18.7)	
Sex, n (%)				<.001
Female	99,905 (78.9)	5486 (81.3)	94,419 (78.8)	
Male	26,648 (21.1)	1260 (18.7)	25,388 (21.2)	
Anthropometric measures, mean (SD)				
Weight (kg)	98.3 (20.4)	101.9 (21.5)	98.1 (20.4)	<.001
Height (cm)	166.7 (9.1)	167.1 (8.7)	166.7 (9.2)	.003
BMI (kg/m^2)	35.3 (6.2)	36.4 (6.7)	35.2 (6.2)	<.001
BMI category, n (%)				<.001
Overweight (25-29.9 kg/m^2)	13,084 (10.4)	841 (12.5)	12,243 (10.2)	
Obese (≥30 kg/m^2)	113,246 (89.6)	5871 (87.5)	107,375 (89.8)	
Comorbidities, n (%)				
Diabetes mellitus	3931 (3.1)	250 (3.7)	3681 (3.1)	.009
Hypertension	12,570 (9.9)	978 (14.5)	11,592 (9.7)	<.001
Hypercholesterolemia	8445 (6.7)	707 (10.5)	7738 (6.5)	<.001
PCOS[b]	7876 (6.2)[c]	591 (10.8)[c]	7285 (6.1)[c]	<.001
Nonalcoholic fatty liver	2391 (1.9)	229 (3.4)	2162 (1.8)	<.001
Digital engagement metrics, n (%)				
Coaching sessions attended	37,224 (29.4)	6746 (100)	30,478 (25.4)	<.001
Weight tracking ≥1 per week readings	18,992 (15)	6746 (100)	12,246 (10.2)	<.001
Uses weight loss app[d]	93,627 (74)	6746 (100)	86,881 (72.5)	<.001

[a]*P* values calculated using independent 2-tailed *t* tests for continuous variables and chi-square tests for categorical variables.

[b]PCOS: polycystic ovary syndrome.

[c]Percentage calculated among female participants only.

[d]Defined as having logged into the app at least once during enrollment and used app features with the exception of weight measurement.

Engagement Characteristics

Among engaged participants (n=6746), all met the 3 required criteria by definition: coaching session attendance (mean 1.00, SD 0.00), app use (mean 1.00, SD 0.00), and weekly weight tracking (mean 1.00, SD 0.00). Among nonengaged participants (n=119,807), partial engagement was common: 86,881 (72.5%) used the app (mean 0.73, SD 0.45), 18,864 (15.7%) attended coaching (mean 0.16, SD 0.36), and tracking adherence (mean 0.09, SD 0.27) was lower. These patterns indicate that while most participants used some platform features, especially 93,627 of 126,553 (74%) of the overall cohort using the digital app, simultaneous engagement across all 3 modalities was achieved by 6,746 of 126,553 (5.3%) of users.

Weight Loss Outcomes

Overview

Figure 2 displays the MMRM model with compound symmetry covariance adjusted for confounders, representing percentage weight loss over 12 months by engagement status. Both digitally engaged and nonengaged groups demonstrated progressive weight loss throughout the observation period, but differences between groups emerged early and increased over time. Table 2 presents detailed weight loss outcomes at each time point.

XSL•FO
RenderX

Figure 2. Mean adjusted mixed model repeated measure derived and confounder-adjusted percentage weight loss from baseline among tirzepatide users stratified by digital engagement status.

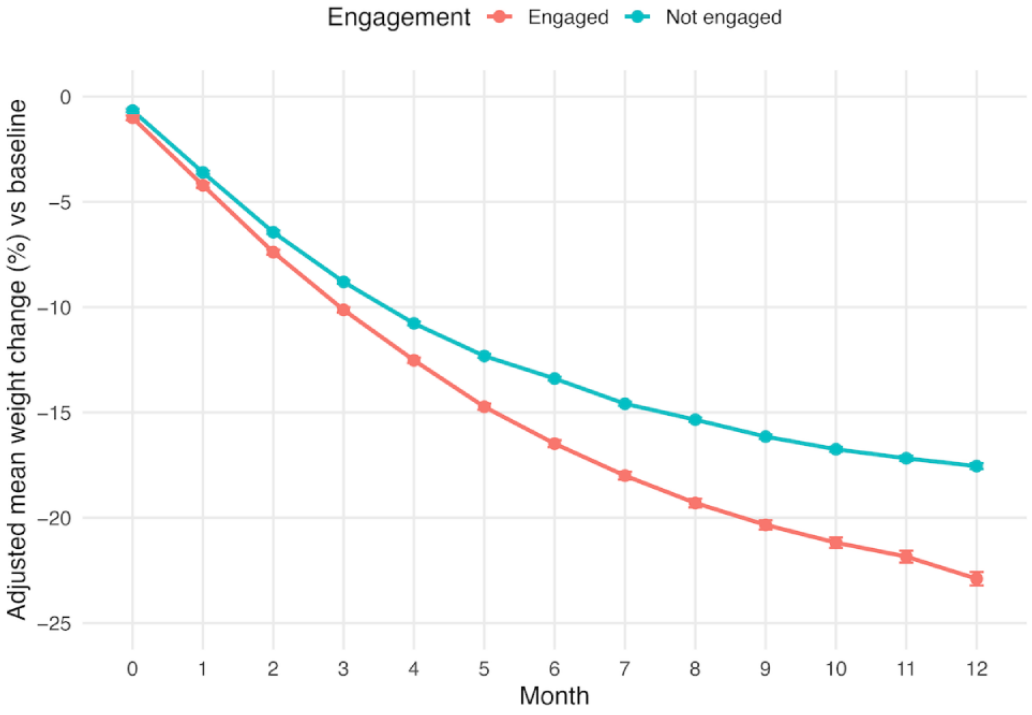

XSL•FO
RenderX

Table 2. Percentage weight loss by digital engagement status in participants taking tirzepatide (months 0-12), estimated using mixed model repeated measure (MMRM).

Month and engagement status	Participants, n	Adjusted mean % weight loss[a] (95% CI)	Absolute difference[b] (percentage points)	Relative difference[c] (%)	Effect size (Cohen *d*)	*P* value[d]
Month 0						
Not engaged	119,807	0	—[e]	—	—	—
Engaged	6746	0	—	—	—	—
Month 2						
Not engaged	40,582	–6.44 (–6.53 to –6.36)	0.95	14.8	0.083	<.001
Engaged	5751	–7.39 (–7.51 to –7.27)	—	—	—	—
Month 4						
Not engaged	21,762	–10.77 (–10.86 to –10.68)	1.75	16.2	0.207	<.001
Engaged	3420	–12.52 (–12.65 to –12.39)	—	—	—	—
Month 6						
Not engaged	13,559	–13.39 (–13.48 to –13.30)	3.09	23.1	0.422	<.001
Engaged	1724	–16.48 (–16.64 to –16.32)	—	—	—	—
Month 8						
Not engaged	6704	–15.34 (–15.45 to –15.24)	3.96	25.8	0.464	<.001
Engaged	896	–19.30 (–19.51 to –19.09)	—	—	—	—
Month 10						
Not engaged	3918	–16.75 (–16.87 to –16.63)	4.43	26.4	0.467	<.001
Engaged	567	–21.18 (–21.42 to –20.93)	—	—	—	—
Month 12						
Not engaged	2343	–17.55 (–17.69 to –17.41)	5.34	30.4	0.539	<.001
Engaged	310	–22.89 (–23.22 to –22.57)	—	—	—	—

[a]Adjusted means derived from MMRM with compound symmetry covariance structure, controlling for age, sex, baseline BMI, diabetes, hypertension, high cholesterol, polycystic ovary syndrome, and fatty liver disease.

[b]Absolute difference=engaged group mean–not engaged group mean (percentage points).

[c]Relative difference=(absolute difference/not engaged group mean)×100%.

[d]*P* values from the model time×engagement interaction terms.

[e]Not applicable.

Weight Loss by Engagement Status

By month 12, after adjusting for potential confounders (age, sex, baseline BMI, and comorbidities) in MMRM analysis, digitally engaged participants achieved a mean weight loss of 22.9% (95% CI 22.6-23.2) compared to 17.6% (95% CI 17.4-17.7) among nonengaged participants, representing an absolute difference of 5.3 percentage points (*P*<.001). The effect size (Cohen *d*=0.539) indicated a medium clinical effect of digital engagement on weight loss outcomes.

The divergence in adjusted weight loss trajectories between groups became statistically significant by month 2, with engaged participants achieving 7.4% weight loss compared to 6.4% among nonengaged participants (absolute difference 0.95 percentage points; *P*<.001). This difference progressively increased throughout the observation period. By month 6, the absolute difference reached 3.1 percentage points (16.5% vs 13.4%; *P*<.001), and by month 8, the difference widened to 4.0 percentage points (19.3% vs 15.3%; *P*<.001).

The MMRM analysis, which controlled for demographic factors and comorbidities while accounting for the correlation structure of repeated measurements, confirmed that digital engagement was consistently associated with enhanced weight loss outcomes at all time points. This statistical approach demonstrated that the engagement effect was not attributable to baseline differences or confounding factors but represented greater weight loss outcomes associated with higher engagement.

Sensitivity Analyses: Weight Loss by Follow-Up Duration

Among the cohort, 10,296 (8.1%) participants reached 3 months, 5349 (4.2%) reached 6 months, 2241 (1.8%) reached 9 months, and 2653 (2.1%) reached 12 months (see Multimedia Appendix 5 for discontinuation and censoring patterns).

Linear regression analyses adjusted for baseline weight and confounders demonstrated consistent associations between

XSL•FO
RenderX

engagement status and weight loss across all follow-up durations. At 3 months, nonengaged participants achieved 9.6% weight loss (95% CI 9.6-9.5) versus 10.3% (95% CI 10.5-10.2) for engaged participants (absolute difference 0.75 percentage points; *P*<.001). At 6 months, differences were 14.7% (95% CI 14.8-14.6) versus 17.6% (95% CI 17.9-17.3), respectively (absolute difference 2.9 percentage points; *P*<.001). At 9 months, nonengaged participants lost 17.8% (95% CI 18.1-17.6) compared to 21.8% (95% CI 22.4-21.2) for engaged participants (absolute difference 4.0 percentage points; *P*<.001). At 12 months, differences were 19.3% (95% CI 19.7-18.9) versus 24.2% (95% CI 25.3-23.2), respectively (absolute difference 4.9 percentage points; *P*<.001). These analyses confirmed that positive associations between engagement and weight loss were maintained regardless of the follow-up duration achieved.

Clinically Significant Weight Loss Thresholds

Kaplan-Meier survival analysis demonstrated that tirzepatide users achieved substantial weight loss milestones rapidly, with overall cumulative probabilities reaching high levels for clinically meaningful thresholds across all participants. When stratified by engagement status, digitally engaged participants achieved these clinically significant thresholds more rapidly and at higher rates than nonengaged participants (Figure 3, log-rank *P*<.001 for all comparisons). The association between engagement and increased weight loss was evident across all thresholds but became increasingly pronounced for more substantial weight loss targets.

Figure 3. Kaplan-Meier survival curves showing the probability of achieving clinically significant weight loss thresholds over time among tirzepatide users by digital engagement status. (A) ≥5% weight loss, (B) ≥10% weight loss, (C) ≥15% weight loss, and (D) ≥20% weight loss. Red lines represent digitally engaged participants; teal lines represent nonengaged participants. Shaded areas represent 95% CIs. *P* values determined by the log-rank test. χ is derived from the chi-square test comparing engaged versus nonengaged.

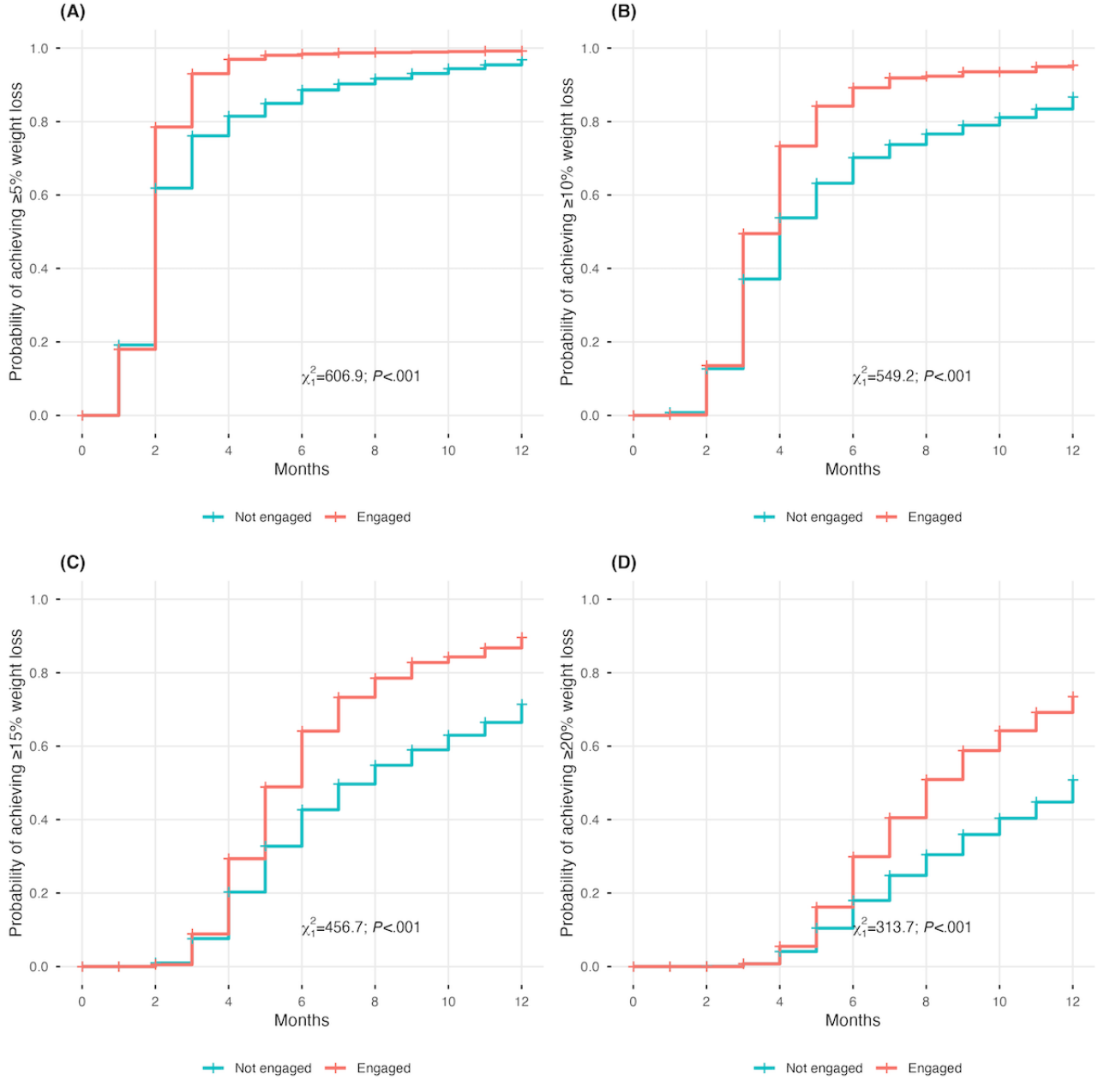

XSL•FO
RenderX

The cumulative probability of achieving ≥5% weight loss was substantially higher among engaged participants (5445/6746, 80.7% vs 44,329/119,807, 37%), demonstrating more than a 2-fold difference (RR 2.18; HR 1.49, 95% CI 1.45-1.53). This engagement advantage became even more pronounced for higher weight loss thresholds: ≥10% weight loss (3623/6746, 53.7% vs 26,240/119,807, 21.9%; RR 2.45; HR 1.55, 95% CI 1.50-1.60), ≥15% weight loss (2091/6746, 31% vs 14,017/119,807, 11.7%; RR 2.65; HR 1.67, 95% CI 1.60-1.75), and ≥20% weight loss (1079/6746, 16% vs 6710/119,807, 5.6%; RR 2.88; HR 1.79, 95% CI 1.68-1.91).

Longitudinal Participation Patterns

As expected, the number of participants with data at each time point decreased over the observation period, primarily reflecting the service's continuous enrollment pattern rather than attrition. Given the rolling nature of the service, participants joined at various time points throughout the evaluation period, with many not yet having the opportunity to reach later time points. The mean follow-up duration was 2.6 (SD 3.0) months, with substantial variation across participants reflecting the rolling enrollment design. Among the cohort, 2653 (2.1%) participants completed 12 months of follow-up. Of these, 310 (11.7%) were maximally engaged, and 2343 (88.3%) were nonengaged. Given the nature of the open cohort study, loss to follow-up could be due to discontinuation (stopped taking medication) or censoring (was still using medication at the cohort end date; Multimedia Appendix 5).

Baseline Correlates of Digital Engagement

Baseline characteristics associated with digital engagement are presented in Table 3. Male participants had significantly lower odds of engagement than female participants (adjusted OR for male: 0.86, 95% CI 0.81-0.92; $P<.001$). Each 10-year increase in age was associated with 18% higher odds of engagement (adjusted OR per decade 1.18, 95% CI 1.15-1.20; $P<.001$). Every 5 kg/m^2 increase in baseline BMI corresponded to 14% higher odds of engagement (adjusted OR 1.14, 95% CI 1.12-1.16; $P<.001$). Among comorbidities, participants with PCOS had 59% higher odds of engaging (adjusted OR 1.59, 95% CI 1.45-1.74; $P<.001$), and those with fatty liver disease had 52% higher odds (adjusted OR 1.52, 95% CI 1.32-1.76; $P<.001$). Hypercholesterolemia was associated with substantially higher engagement odds (adjusted OR 1.34, 95% CI 1.23-1.47; $P<.001$), while hypertension showed more modest associations (adjusted OR 1.16, 95% CI 1.07-1.25; $P<.001$). Notably, diabetes was associated with significantly lower odds of engagement (adjusted OR 0.83, 95% CI 0.73-0.95; $P=.009$).

Table 3. Baseline characteristics associated with digital engagement among tirzepatide users through multivariable logistic regression analysis[a].

Characteristic	Adjusted OR[b] (95% CI)	*P* value
Demographics		
Female sex	Reference (—[c])	—
Male sex	0.86 (0.81-0.92)	<.001
Age (per year)	1.02 (1.01-1.02)	<.001
Age (per decade)[d]	1.18 (1.15-1.20)	<.001
BMI (per unit)	1.03 (1.02-1.03)	<.001
BMI (per 5 kg/m^2)[d]	1.14 (1.12-1.16)	<.001
Comorbidities		
Polycystic ovary syndrome	1.59 (1.45-1.74)	<.001
Fatty liver disease	1.52 (1.32-1.76)	<.001
Hypertension	1.16 (1.07-1.25)	<.001
Hypercholesterolemia	1.34 (1.23-1.47)	<.001
Diabetes	0.83 (0.73-0.95)	.009

[a]The dependent variable was digital engagement status (engaged vs not engaged), defined as active use of digital coaching, tracking tools, or a mobile app. All characteristics were assessed at baseline (month 0) prior to tirzepatide initiation. Multivariable logistic regression model adjusted for all variables shown.

[b]OR: odds ratio.

[c]Not applicable.

[d]Derived values to express effect sizes for clinically meaningful increments.

Discussion

Principal Findings

This cohort study demonstrated the notable real-world effectiveness of tirzepatide in a UK-based DWLS and highlighted an association between higher digital engagement and greater weight loss. After controlling for demographic factors and comorbidities, digitally engaged participants achieved an absolute 5.3 percentage points greater weight loss at month 12 compared to nonengaged users ($P<.001$). This engagement effect emerged gradually, with trajectories

XSL•FO
RenderX

beginning to diverge around month 2 and differences progressively widening thereafter. By month 12, digitally engaged participants achieved a mean weight loss of 22.9% compared to 17.6% for nonengaged participants.

Nearly all tirzepatide users eventually achieved ≥5% weight loss regardless of engagement status, but more substantial differences were observed for the engaged group for higher weight loss thresholds. Notably, engaged participants demonstrated substantially higher cumulative probabilities of achieving ≥15% weight loss (2091/6746, 31% vs 14,017/119,807, 11.7%; HR 1.67, 95% CI 1.60-1.75) and ≥20% weight loss (1079/6746, 16% vs 6710/119,807, 5.6%; HR 1.79, 95% CI 1.68-1.91) by study end. The temporal pattern of weight loss differences is meaningful. While statistically significant differences emerged by month 2, the magnitude of difference progressively increased throughout the observation period. In other studies, a 21% weight loss end point was typically reached at week 72 as opposed to at week 52 (month 12) as evidenced in our findings [15].

The temporal differences in weight loss associated with engagement status were further highlighted in our analysis, which revealed significantly accelerated achievement of clinically meaningful weight loss thresholds. While engagement conferred substantial advantages for ≥5% weight loss achievement (5445/6746, 80.7% vs 44,329/119,807, 37%; RR 2.18; HR 1.49, 95% CI 1.45-1.53), more substantial differences emerged for more ambitious thresholds. Notably, engaged participants demonstrated substantially higher cumulative probabilities of achieving ≥10% weight loss (3623/6746, 53.7% vs 26,240/119,807, 21.9%; RR 2.45; HR 1.55, 95% CI 1.50-1.60), ≥15% weight loss (2091/6746, 31% vs 14,017/119,807, 11.7%; RR 2.65; HR 1.67, 95% CI 1.60-1.75), and ≥20% weight loss (1079/6746, 16% vs 6710/119,807, 5.6%; RR 2.88; HR 1.79, 95% CI 1.68-1.91).

Our analysis of baseline characteristics identified that female sex, older age, and higher BMI were associated with increased engagement likelihood, with male participants having 14% lower odds (OR 0.86, 95% CI 0.81-0.92), each decade of age conferring 18% higher odds (OR 1.18, 95% CI 1.15-1.20), and every 5 kg/m^2 BMI increase associated with 14% higher odds (OR 1.14, 95% CI 1.12-1.16; all P<.001). Among comorbidities, PCOS (59% higher odds; OR 1.59, 95% CI 1.45-1.74), fatty liver disease (52% higher odds; OR 1.52, 95% CI 1.32-1.76), and hypercholesterolemia (34% higher odds; OR 1.34, 95% CI 1.23-1.47) were positively associated with engagement, while diabetes was associated with 17% lower engagement odds (OR 0.83, 95% CI 0.73-0.95;; P=.009). These findings highlight important demographic factors influencing digital health use that may inform future program design and implementation.

Strengths and Limitations

This study has several strengths, including its large sample size, real-world setting, and comprehensive assessment of weight trajectories over time. The 126,000-participant cohort provides robust power at early time points (n=40,582 at 2 months); while 12-month completers (n=2653) match other digital studies, the real-world implementation enhances generalizability and ecological validity. Additionally, the operational definition of digital engagement captured meaningful interaction with service components while maintaining practical relevance.

While baseline differences between engaged and nonengaged participants reached statistical significance due to the large sample size (n=126,553), the absolute magnitude of these differences was modest and unlikely to be clinically meaningful. The mean age difference was less than 1 year (mean 42.3, SD 12.5 vs mean 44.9, SD 12.0 years), and the BMI difference was approximately 1 kg/m^2 (mean 35.2, SD 6.2 vs mean 36.4, SD 6.7 kg/m^2). These small baseline differences are insufficient to account for the substantial treatment effects observed, particularly the 7.4% absolute difference in weight loss at 2 months and the near 3-fold difference in treatment persistence. Furthermore, the consistency association between engagement and outcomes across all time points, combined with our multivariable adjustment for baseline characteristics, supports the conclusion that digital engagement genuinely enhances pharmacological efficacy rather than these findings being attributable to baseline confounding.

Several limitations warrant consideration. First, as a retrospective service evaluation, participants were not randomly assigned to engagement conditions, introducing potential selection bias, and we cannot ascertain the causality of this association. While we adjusted for observed confounders, unmeasured factors (eg, motivation and socioeconomic status) may influence both engagement and outcomes. Second, weight measurements were self-reported via home scales, potentially introducing measurement and reporting error. However, this limitation applied to both engaged and nonengaged groups, likely minimizing differential bias. We did not systematically capture concurrent participation in external nonpharmacological weight management programs; we note this as a limitation and as a priority for future data collection. Additionally, ethnicity data were not systematically collected during the study period, limiting our ability to examine potential disparities in engagement or outcomes across ethnic groups. Future implementations of DWLS should prioritize collecting comprehensive demographic data, including ethnicity and socioeconomic indicators, to ensure equity in access and outcomes and to identify populations that may benefit from targeted engagement strategies. The monthly cost for the service in this study service represents a significant financial consideration that may limit access primarily to individuals with higher socioeconomic status. This raises important considerations regarding health inequalities, as individuals who might benefit most from weight management interventions may face financial barriers to accessing digital health services that integrate pharmacotherapy with behavioral support.

Moreover, the rolling enrollment design means that participants were at different stages of their treatment journey during the evaluation period. While this reflects real-world implementation, it complicates the interpretation of longitudinal patterns. Our approach of examining both absolute outcomes at each time point and conducting subgroup analyses of participants with sufficient follow-up time helps address this limitation. Finally, our operational definition of digital engagement, while

XSL•FO
RenderX

evidence-informed, represents one approach among many possible definitions. Future research should explore alternative engagement metrics and potential dose-response relationships between engagement intensity and outcomes.

Comparison With Prior Work

The magnitude of additional weight loss associated with digital engagement was substantial in the context of standard obesity treatment. Previous research indicates that each 5% reduction in body weight confers significant cardiometabolic benefits, including improvements in glycemic control, blood pressure, and lipid profiles [16,17]. Our findings suggest that digital engagement may help individuals reach more substantial weight loss thresholds, potentially amplifying the health benefits of tirzepatide treatment. This suggests that digital engagement may have cumulative benefits, potentially by supporting medication adherence, dietary modifications, and behavioral changes that enhance long-term outcomes [18,19]. Similar patterns have been observed in other behavioral interventions for chronic disease management, where ongoing support provides incremental benefits over time [20].

The identified correlates of digital engagement offer insights for service optimization. The higher likelihood of engagement among female participants, older adults, and those with higher baseline BMI suggests that these demographic groups may find digital support particularly valuable. Conversely, younger male participants with lower BMI appear less likely to engage, potentially benefiting from targeted engagement strategies; yet, digital health engagement literature has shown that male participants benefit more than female participants [21]. Interestingly, studies have proposed that people with diabetes have a slower velocity of weight loss in both the GLP-1 [22] and GLP and GIP-1 [23] agonist exposure groups (albeit, faster velocity in tirzepatide vs semaglutide).

Furthermore, our study confirmed that diabetes status was significantly associated with a lower likelihood of digital engagement (adjusted OR 0.83, 95% CI 0.73-0.95; *P*=.009). This may reflect complex physiological and psychological mechanisms influencing weight loss velocity in people with diabetes. To address this, DWLS should consider implementing more intensive engagement and coaching interventions tailored for patients with diabetes. We propose that digital services pilot such enhanced approaches to test whether outcomes can be improved. By contrast, the positive association between comorbidities, particularly PCOS [24] and engagement, suggests that individuals with obesity-related health concerns may perceive greater value in comprehensive support services.

Several mechanisms may explain the enhanced outcomes associated with digital engagement. Health coaching provides accountability, problem-solving support, and personalized guidance that can address common barriers to weight management [25]. Regular weight tracking enhances self-monitoring, a behavioral strategy consistently associated with improved weight outcomes [8]. Additionally, digital platforms may facilitate greater medication adherence through reminders, side effect management, and addressing various concerns, particularly important for injectable medications like tirzepatide that require consistent administration [26].

Our findings align with previous digital weight management studies while revealing important distinctions. Xu et al [27] (n=153) defined engagement as any daily food tracking, identifying thresholds of 28.5%-39.4% of days for ≥3%-5% weight loss. W8Buddy used a liberal definition (any platform activity within 14 days), achieving 83.1% engagement with 0.74 kg per month additional weight loss. Our study used the strictest definition (coaching, weekly weight tracking, and app login combined), resulting in only 5.3% meeting criteria, yet demonstrating the largest effect (5.3 percentage points at 12 months) [28]. This contrast, where stricter criteria with fewer engaged participants yielding greater outcomes, suggests that multimodal engagement combining behavioral support, self-monitoring, and clinical oversight produces beneficial effects beyond single-modality interventions. Notably, partial engagement remained common (app use: 93,627/126,553, 74% and coaching: 37,224/126,553, 29.4%), and even nonengaged participants achieved substantial weight loss (17.6%), indicating that platform features provided benefit across engagement levels. These findings demonstrate that while engagement definitions substantially influence observed rates, the consistent association between digital engagement and enhanced outcomes persists across platforms and intervention designs.

The integration of digital services with pharmacotherapy represents an evolving care model addressing limitations of traditional obesity treatment approaches. By providing remote support, behavioral tools, and clinical monitoring, DWLSs may bridge critical gaps in obesity care while reducing barriers related to geography, provider availability, and stigma [29,30]. Our findings suggest that such integrated approaches may optimize the effectiveness of novel obesity pharmacotherapies.

Implications and Future Directions

Our findings have important implications for clinical practice, health service design, and research. Clinically, practitioners should consider recommending digital support services alongside tirzepatide prescriptions, particularly for patients who may benefit from additional accountability and behavioral guidance. From a service design perspective, our results support investment in research and development for digital infrastructure that facilitates engagement, particularly during early treatment phases when engagement patterns appear to be established.

Some self-funded DWLS can exacerbate health inequalities; therefore, future research should examine the cost-effectiveness of such integrated DWLS compared to medication-only approaches and explore alternative funding models, including potential National Health Service commissioning pathways, that could improve accessibility across socioeconomic groups while maintaining service quality and clinical outcomes [31]. Additionally, qualitative studies exploring patient experiences and clinician perspectives could identify specific digital components that contribute most to engagement and outcomes. Finally, RCTs comparing different digital support intensities would address causality questions and optimize resource allocation.

XSL•FO
RenderX

Conclusions

This study demonstrates that digital engagement was significantly associated with enhanced weight loss outcomes among individuals using tirzepatide for obesity management. By month 12, engaged participants achieved –22.9% weight loss and an absolute difference of –5.3% compared to nonengaged participants, with differences emerging early and increasing over time. These findings highlight the potential value of integrated care models combining pharmacotherapy with digital support services. As novel obesity medications continue to transform treatment possibilities, optimizing their implementation through complementary digital strategies represents a promising approach to addressing the global obesity epidemic.

Acknowledgments

The authors confirm that no generative artificial intelligence tools were used to generate the manuscript text, analyses, figures, or references.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the nature of the clinical data collected and the consent provided by patients; the individual participant data used in this study are not publicly available. The statistical code used in the analyses can be made available to researchers upon request to the corresponding author. Requests for access to aggregated high-level anonymized data for noncommercial projects should be directed to the corresponding author, will require appropriate ethics and legal approval, and are available on reasonable request.

Authors' Contributions

Conceptualization: HJ, AKC, DH
Methodology: HJ (lead), AKC (supporting)
Formal analysis: HJ (lead), AKC (supporting)
Validation: HJ, AKC, DR, DH
Investigation: HJ
Data curation: HJ (lead), AKC (supporting)
Visualization: HJ
Supervision: AKC, DH
Project administration: HJ
Writing—original draft: HJ (lead), AKC (supporting)
Writing—review and editing: HJ (lead), AKC (supporting), DR (supporting), DH (supporting)

Conflicts of Interest

The authors HJ, DH, AKC, and DR are members within the organization Voy, Menwell; HJ is the clinical researcher, AKC is the head of clinical research, DH is the innovation director, and DR is the research advisor.

Multimedia Appendix 1
Detailed program description.
[DOCX File , 24 KB - jmir_v28i1e83718_app1.docx]

Multimedia Appendix 2
Eligibility criteria and cohort derivation of the study.
[DOCX File , 17 KB - jmir_v28i1e83718_app2.docx]

Multimedia Appendix 3
Statistical methods and model specifications to run on R.
[DOCX File , 17 KB - jmir_v28i1e83718_app3.docx]

Multimedia Appendix 4
STROBE checklist.

XSL•FO
RenderX

[PDF File (Adobe PDF File), 184 KB - jmir_v28i1e83718_app4.pdf]

Multimedia Appendix 5
Participant flow diagram showing retention, discontinuation, and censoring patterns over 12 months of follow-up.
[DOCX File , 75 KB - jmir_v28i1e83718_app5.docx]

References

1. Okunogbe A, Nugent R, Spencer G, Ralston J, Wilding J. Economic impacts of overweight and obesity: current and future estimates for eight countries. BMJ Glob Health 2021;6(10):e006351 [FREE Full text] [doi: 10.1136/bmjgh-2021-006351] [Medline: 34737167]
2. Hajjaj AB, Guijarro PM, Khan KS, Bueno-Cavanillas A, Cano-Ibáñez N. Author correction: a systematic review and meta-analysis of weight loss in control group participants of lifestyle randomized trials. Sci Rep 2022;12(1):14444 [FREE Full text] [doi: 10.1038/s41598-022-18828-y] [Medline: 36002563]
3. Syed YY. Tirzepatide: first approval. Drugs 2022;82(11):1213-1220. [doi: 10.1007/s40265-022-01746-8] [Medline: 35830001]
4. Jastreboff AM, Aronne LJ, Ahmad NN, Wharton S, Connery L, Alves B, et al. Tirzepatide once weekly for the treatment of obesity. N Engl J Med 2022;387(3):205-216. [doi: 10.1056/NEJMoa2206038] [Medline: 35658024]
5. Johnson H, Huang D, Liu V, Ammouri MA, Jacobs C, El-Osta A. Impact of digital engagement on weight loss outcomes in obesity management among individuals using GLP-1 and dual GLP-1/GIP receptor agonist therapy: retrospective cohort service evaluation study. J Med Internet Res 2025;27:e69466 [FREE Full text] [doi: 10.2196/69466] [Medline: 40164173]
6. Tronieri JS, Wadden TA, Chao AM, Tsai AG. Primary care interventions for obesity: review of the evidence. Curr Obes Rep 2019;8(2):128-136 [FREE Full text] [doi: 10.1007/s13679-019-00341-5] [Medline: 30888632]
7. Richards R, Wren G, Whitman M. The potential of a digital weight management program to support specialist weight management services in the UK National Health Service: retrospective analysis. JMIR Diabetes 2024;9:e52987 [FREE Full text] [doi: 10.2196/52987] [Medline: 38265852]
8. Burke LE, Wang J, Sevick MA. Self-monitoring in weight loss: a systematic review of the literature. J Am Diet Assoc 2011;111(1):92-102 [FREE Full text] [doi: 10.1016/j.jada.2010.10.008] [Medline: 21185970]
9. Lau Y, Chee DGH, Chow XP, Cheng LJ, Wong SN. Personalised eHealth interventions in adults with overweight and obesity: a systematic review and meta-analysis of randomised controlled trials. Prev Med 2020;132:106001. [doi: 10.1016/j.ypmed.2020.106001] [Medline: 31991155]
10. Kouvari M, Karipidou M, Tsiampalis T, Mamalaki E, Poulimeneas D, Bathrellou E, et al. Digital health interventions for weight management in children and adolescents: systematic review and meta-analysis. J Med Internet Res 2022;24(2):e30675 [FREE Full text] [doi: 10.2196/30675] [Medline: 35156934]
11. Combet E, Haag L, Richardson J, Haig CE, Cunningham Y, Fraser HL, et al. Remotely delivered weight management for people with long COVID and overweight: the randomized wait-list-controlled ReDIRECT trial. Nat Med 2025;31(1):258-266. [doi: 10.1038/s41591-024-03384-x] [Medline: 39779922]
12. Cheung KL, Eggers SM, de Vries H. Combining the integrated-change model with self-determination theory: application in physical activity. Int J Environ Res Public Health 2020;18(1):28 [FREE Full text] [doi: 10.3390/ijerph18010028] [Medline: 33374522]
13. Farmanbar R, Niknami S, Lubans DR, Hidarnia A. Predicting exercise behaviour in Iranian college students: utility of an integrated model of health behaviour based on the transtheoretical model and self-determination theory. Health Educ J 2012;72(1):56-69. [doi: 10.1177/0017896911430549]
14. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007;335(7624):806-808 [FREE Full text] [doi: 10.1136/bmj.39335.541782.AD] [Medline: 17947786]
15. Tan B, Pan X, Chew HSJ, Goh RSJ, Lin C, Anand VV, et al. Efficacy and safety of tirzepatide for treatment of overweight or obesity. A systematic review and meta-analysis. Int J Obes (Lond) 2023;47(8):677-685. [doi: 10.1038/s41366-023-01321-5] [Medline: 37253796]
16. Ryan DH, Yockey SR. Weight loss and improvement in comorbidity: differences at 5%, 10%, 15%, and over. Curr Obes Rep 2017;6(2):187-194 [FREE Full text] [doi: 10.1007/s13679-017-0262-y] [Medline: 28455679]
17. Wing RR, Lang W, Wadden TA, Safford M, Knowler WC, Bertoni AG, et al. Benefits of modest weight loss in improving cardiovascular risk factors in overweight and obese individuals with type 2 diabetes. Diabetes Care 2011;34(7):1481-1486 [FREE Full text] [doi: 10.2337/dc10-2415] [Medline: 21593294]
18. Boucher EM, Raiker JS. Engagement and retention in digital mental health interventions: a narrative review. BMC Digit Health 2024;2(1):52. [doi: 10.1186/s44247-024-00105-9]
19. Lehmann M, Jones L, Schirmann F. App engagement as a predictor of weight loss in blended-care interventions: retrospective observational study using large-scale real-world data. J Med Internet Res 2024;26:e45469 [FREE Full text] [doi: 10.2196/45469] [Medline: 38848556]

XSL•FO
RenderX

20. Wadden TA, Tronieri JS, Butryn ML. Lifestyle modification approaches for the treatment of obesity in adults. Am Psychol 2020;75(2):235-251 [FREE Full text] [doi: 10.1037/amp0000517] [Medline: 32052997]
21. König LM, Western MJ, Denton AH, Krukowski RA. Umbrella review of social inequality in digital interventions targeting dietary and physical activity behaviors. NPJ Digit Med 2025;8(1):11 [FREE Full text] [doi: 10.1038/s41746-024-01405-0] [Medline: 39762352]
22. Jensterle M, Rizzo M, Haluzík M, Janež A. Efficacy of GLP-1 RA approved for weight management in patients with or without diabetes: a narrative review. Adv Ther 2022;39(6):2452-2467 [FREE Full text] [doi: 10.1007/s12325-022-02153-x] [Medline: 35503498]
23. Viljoen A, Pantalone KM, Galindo RJ, Cui X, Huh R, Hemmingway A, et al. Time to reach glycaemic and body weight loss thresholds with tirzepatide in patients with type 2 diabetes: a pre-planned exploratory analysis of SURPASS-2 and SURPASS-3. Diabetes Ther 2023;14(5):925-936 [FREE Full text] [doi: 10.1007/s13300-023-01398-1] [Medline: 37000390]
24. Monney M, Mavromati M, Leboulleux S, Gariani K. Endocrine and metabolic effects of GLP-1 receptor agonists on women with PCOS, a narrative review. Endocr Connect 2025;14(5):6975 [FREE Full text] [doi: 10.1530/EC-24-0529] [Medline: 40066975]
25. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
26. Oakley-Girvan I, Yunis R, Longmire M, Ouillon JS. What works best to engage participants in mobile app interventions and e-Health: a scoping review. Telemed J E Health 2022;28(6):768-780 [FREE Full text] [doi: 10.1089/tmj.2021.0176] [Medline: 34637651]
27. Xu R, Bannor R, Cardel MI, Foster GD, Pagoto S. How much food tracking during a digital weight-management program is enough to produce clinically significant weight loss? Obesity (Silver Spring) 2023 Jul;31(7):1779-1786 [FREE Full text] [doi: 10.1002/oby.23795] [Medline: 37271576]
28. Hanson P, Abdelhameed F, Sahir M, Parsons N, Panesar A, de la Fosse M, et al. Evaluation of the digital support tool Gro Health W8Buddy as part of tier 3 weight management service: observational study. J Med Internet Res 2025 May 16;27:e62661 [FREE Full text] [doi: 10.2196/62661] [Medline: 40378402]
29. Talay L, Alvi O. Digital healthcare solutions to better achieve the weight loss outcomes expected by payors and patients. Diabetes Obes Metab 2024;26(6):2521-2523. [doi: 10.1111/dom.15513] [Medline: 38379435]
30. Barron E, Bradley D, Safazadeh S, McGough B, Bakhai C, Young B, et al. Effectiveness of digital and remote provision of the Healthier You: NHS Diabetes Prevention Programme during the COVID-19 pandemic. Diabet Med 2023;40(5):e15028 [FREE Full text] [doi: 10.1111/dme.15028] [Medline: 36524707]
31. Thomsen RW, Mailhac A, Løhde JB, Pottegård A. Real-world evidence on the utilization, clinical and comparative effectiveness, and adverse effects of newer GLP-1RA-based weight-loss therapies. Diabetes Obes Metab 2025;27(Suppl 2):66-88. [doi: 10.1111/dom.16364] [Medline: 40196933]

Abbreviations

DWLS: digital weight loss service
GIP: glucose-dependent insulinotropic polypeptide
GLP-1: glucagon-like peptide-1
HR: hazard ratio
MMRM: mixed model for repeated measures
OR: odds ratio
PCOS: polycystic ovary syndrome
RA: receptor agonist
RCT: randomized controlled trial
RR: risk ratio
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by N Cahill; submitted 07.Sep.2025; peer-reviewed by G Foster, P Hanson; comments to author 27.Oct.2025; revised version received 11.Nov.2025; accepted 26.Nov.2025; published 15.Jan.2026.

Please cite as:
Johnson H, Clift AK, Reisel D, Huang D
Digital Engagement Significantly Enhances Weight Loss Outcomes in Adults With Obesity Treated With Tirzepatide: Retrospective Cohort Study of a Digital Weight Loss Service
J Med Internet Res 2026;28:e83718
URL: https://www.jmir.org/2026/1/e83718
doi:10.2196/83718
PMID:

XSL•FO
RenderX

©Hans Johnson, Ashley Kieran Clift, Daniel Reisel, David Huang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Relationship Between Physician Self-Disclosure and Patient Acquisition in Digital Health Markets: Cross-Sectional Study

Quanchen Liu[1], PhD; Pengqing Yin[1], MSc; Jing Fan[1], PhD

International Business School, Beijing Foreign Studies University, Beijing, China

Corresponding Author:
Jing Fan, PhD
International Business School
Beijing Foreign Studies University
No. 19 Xisanhuan North Road
Beijing, 100089
China
Phone: 86 (010) 88818121
Email: fanjing@bfsu.edu.cn

Abstract

Background: Online health communities have evolved into digital marketplaces where physicians have to compete for patients. Existing research examines physician-patient dynamics through a patient-centric lens, treating physicians as passive recipients of ratings and reviews, while the strategic role of physician self-disclosure remains unexamined. This gap constrains a comprehensive understanding of how physicians can actively shape patient decisions, making the investigation of strategic self-disclosure imperative.

Objective: This study aims to investigate the relationship between physician self-disclosure breadth (scope of information) and depth (detailed expertise) and patient decision-making, as well as whether regional digital health care level (DHL) moderates these relationships.

Methods: We conducted a cross-sectional analysis of observational data to test these relationships. Data were collected from China's online health care platform Haodf from September to December 2024. Self-disclosure breadth (including clinical performance, academic experience, and social reputation), self-disclosure depth (including expertise coverage, richness, and granularity), and patient decision-making (total visits) were captured through manual content coding and quantitative measurement. We used structured content analysis to extract the disclosure components, informational scope, and descriptive details of each profile. Then, using validated operational formulas, we calculated the composite indices for disclosure breadth and depth based on the coded dimensions. The study generated 1798 final physician samples with complete data across 14 focal variables. The hypotheses were tested using an ordinary least squares regression model, and 4 robustness checks were conducted, including variable substitution and different resampling techniques.

Results: In the primary ordinary least squares regression models, self-disclosure breadth was significantly and positively associated with patient visits (β=0.255, 95% CI 0.054-0.456; *P*=.01), as was self-disclosure depth (β=0.098, 95% CI 0.030-0.167; *P*=.005). The breadth×DHL interaction was positive and significant (β=0.261, 95% CI 0.061-0.461; *P*=.01). Similarly, the depth×DHL interaction was positive and significant (β=0.070, 95% CI 0.002-0.138; *P*=.045). It should be noted that the association for self-disclosure breadth was stronger than that of self-disclosure depth. DHL strengthened the relationship between the disclosure strategies with patient visits. This contextual amplification indicates that DHL serves as a critical boundary condition, determining the degree to which physician self-disclosure strategies translate into patient acquisition outcomes.

Conclusions: This study reconceptualizes physicians as strategic agents shaping patient decision-making through purposeful self-disclosure. Different from existing studies treating physicians as passive recipients of ratings and reviews, our research demonstrates that physicians can strategically shape patient acquisition through self-disclosure breadth and depth. This study brings new insights to digital health markets by demonstrating that self-disclosure operates as a viable patient acquisition mechanism, wherein the DHL acts as a critical boundary condition. The findings have real-world implications: (1) physicians can leverage evidence-based disclosure strategies, (2) platforms should implement context-adaptive features, and (3) policymakers should prioritize digital infrastructure investments to enhance physicians' competitive capabilities and patient decision-making quality.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e84963) doi:10.2196/84963

KEYWORDS

online health communities; physician online profile; physician self-disclosure; patient decision-making; physician-patient interaction; digital health care level

Introduction

Background

The health care landscape is experiencing an unprecedented digital transformation, with online health communities (OHCs) emerging as powerful intermediaries that fundamentally reshape patient-physician interactions. OHCs democratize medical information access and empower patients to actively evaluate health care providers before making consultation decisions [1]. Platforms, such as HealthTap [2] and China's Haodf, now serve millions of users globally, with the OHCs market projected to expand from US $13.3 billion in 2022 to US $42.9 billion by 2030 [3]. This revolution forces physicians to build compelling online presences beyond clinical excellence to compete effectively in an increasingly crowded digital marketplace.

Review of Relevant Scholarship

Existing research on OHCs has predominantly examined physician-patient dynamics through a patient-centric lens, treating physicians as passive recipients of online ratings and reviews rather than strategic actors capable of influencing patient decisions. Current studies focus extensively on how patients leverage physician profiles, ratings, and accumulated reviews to screen health care providers [4], effectively positioning physicians as static entities whose past digital footprints predetermine patient selection outcomes. This perspective fundamentally overlooks physicians' potential for active agency in patient acquisition. Moreover, while scholars acknowledge regional variations in digital health care level (DHL [5,6]), encompassing the sophistication of technological infrastructure, digital literacy capabilities, and information accessibility within specific health care environments), little attention has been paid to how these contextual differences might alter physicians' strategic opportunities and patients' information processing capabilities, creating a significant theoretical blind spot in understanding physician behavior within digitally heterogeneous health care environments.

In an increasingly crowded OHC landscape where patients can choose from hundreds of providers, physicians must now strategically differentiate themselves through deliberate self-presentation beyond clinical excellence and accumulated reviews. Within this context, self-disclosure theory from social psychology offers a promising framework, suggesting that strategic information disclosure across breadth (scope of information) and depth (level of detail) dimensions can enhance credibility, build trust, and reduce decision-making uncertainty [7], thereby influencing differential patient choices. However, the effectiveness of physician self-disclosure cannot be understood in isolation from regional digital health care contexts. In technologically advanced regions, physicians access sophisticated multimedia tools that enhance disclosure opportunities, while patients possess higher digital literacy, enabling effective interpretation of complex professional information [8]. Conversely, less digitally developed areas present technological constraints limiting physicians' ability to communicate expertise effectively, while patients may lack sufficient digital literacy to process and verify disclosed information [9]. This contextual complexity suggests that physician self-disclosure effectiveness may vary across digital health care environments, as strategies proving highly effective in advanced contexts might yield diminished returns in regions with limited infrastructure and lower digital literacy.

Aims, Objectives, and Hypotheses

Overview

Based on these research gaps and theoretical considerations, our research aims to contribute to the existing literature on physician strategic self-disclosure behavior in OHCs by addressing 2 primary research objectives. First, we aim to investigate how physician self-disclosure breadth and depth are associated with patient decision-making in OHCs. Second, we seek to determine whether and how the regional DHL moderates the relationship between these self-disclosure strategies and patient choices.

Physician Self-Disclosure and Patient Decision-Making

Literature review in Section 1 in Multimedia Appendix 1 highlights that physicians' self-disclosure significantly shapes patients' credibility and trust toward decision-making by providing multifaceted professional information. Within OHCs, both breadth and depth dimensions of physician self-disclosure systematically activate these cognitive evaluations, ultimately shaping patient consultation decisions.

Self-disclosure breadth enhances patient decision-making by providing comprehensive professional signals that directly build credibility. When physicians disclose extensive information across multiple professional dimensions, they create a rich tapestry of verifiable cues that patients can cross-reference and validate. This comprehensive presentation first enhances perceived source credibility, as patients can observe concrete evidence of qualifications across diverse professional domains, reducing concerns about physician competence. The breadth of disclosure subsequently fosters interpersonal trust by signaling transparency and professional openness, suggesting that physicians have "nothing to hide" and are confident in their professional standing. Finally, this extensive information scope significantly increases perceived diagnostic value by providing patients with sufficient data points to make informed assessments about physician-patient compatibility. Patients can evaluate whether the physician's experience, training, and achievements align with their specific medical needs and preferences, thereby reducing decision-making uncertainty and increasing consultation likelihood.

XSL•FO
RenderX

Self-disclosure depth is associated with patient decision-making through intensive information quality that demonstrates specialized expertise and professional communication. When physicians provide detailed expertise descriptions, they signal profound clinical knowledge and commitment to patient understanding. This depth reinforces perceived source credibility by showcasing mastery within specific medical domains, as detailed explanations indicate genuine expertise rather than superficial knowledge. It also then builds interpersonal trust by demonstrating physicians' investment in clear communication and patient education, suggesting benevolent intentions and professional dedication. Most critically, depth maximizes perceived diagnostic value by enabling patients to precisely evaluate treatment fit—detailed specialty descriptions allow patients to determine whether their specific conditions fall within the physician's demonstrated areas of expertise. This granular matching capability reduces ambiguity about treatment appropriateness and increases patients' confidence in scheduling consultations with physicians.

DHL as the Moderator

DHL systematically shapes the mechanisms through which self-disclosure breadth and depth operate, fundamentally altering how the same self-disclosure content is produced, transmitted, and interpreted.

Advanced DHL amplifies the credibility-building effects of self-disclosure breadth through enhanced verification mechanisms and seamless information processing. When DHL is high, physicians can populate comprehensive profile fields with verifiable credentials, embed direct links to official registries, and present information through user-friendly interfaces that facilitate patient navigation. Patients in these contexts possess the digital literacy to efficiently cross-validate credentials through integrated databases and verification systems, creating a low-friction pathway for establishing source credibility. This enhanced verification capability strengthens the breadth-credibility relationship posited by self-disclosure theory, thereby accelerating patients' cognitive progression from enhanced source credibility to ultimate consultation decisions. In contrast, regions with limited digital infrastructure constrain verification processes, weakening the credibility signals that breadth disclosure would otherwise provide.

Similarly, advanced DHL intensifies the trust-building effects of self-disclosure depth by enabling rich multimedia presentations and sophisticated patient interpretation capabilities. High-level digital health care environments allow physicians to create comprehensive and well-structured disclosure experiences. Patients with elevated digital literacy can effectively parse these complex multimedia presentations, interpreting granular clinical details and structured expertise descriptions as authentic signals of both professional competence and patient-centered communication. This enhanced processing capability amplifies the depth-trust relationship, as patients can fully appreciate the nuanced expertise demonstrations that deep disclosure provides. Conversely, in regions with poor connectivity and limited digital literacy, deep disclosure content may fail to render properly or overwhelm patients' interpretive capacities, potentially undermining rather than enhancing the intended trust-building effects. On the basis of the preceding discussion, we advance the following two hypotheses:

- Hypothesis 1: Physician self-disclosure breadth and depth are positively associated with patient decision-making.
- Hypothesis 2: DHL positively moderates the relationship between self-disclosure depth, breadth, and patient decision-making.

Methods

Sample Size, Power, and Precision

We conducted a cross-sectional study that was designed and reported in accordance with the JARS (Journal Article Reporting Standards) guidelines [10] to examine the relationship between physician self-disclosure and patient acquisition in digital health markets. Our cross-sectional secondary analysis aimed to estimate the association between physicians' online self-disclosure and patient acquisitions with an absolute precision of ±0.5 percentage points at a 95% CI. The required number of physician profiles was calculated with the single-proportion formula sample size (n)=$[(Z_{(1-\alpha/2)})^2 P(1-P)]/d^2$ [11-13], where $Z_{(1-\alpha/2)}$=the critical value with a corresponding standard level of confidence (1.96 at 95% CI), P was the conservative prevalence of 50% (as this value maximizes variance when the true prevalence is unknown and therefore yields the largest required sample size, ensuring adequate power and precision [14]), d=5% allowable margin of error or desired precision, indicating a minimum sample size of 384 physicians. To guard against unforeseen data-quality issues, we set a conservative target of at least 1000 evaluable records. Besides, this study is observational and contains no experimental aims, power calculations for between-group comparisons were unnecessary.

Data Collection

Haodf (established in 2006) is China's largest online health care platform, covering more than 10,000 hospitals and 900,000 physicians nationwide as of July 2023. Physician participation is exceptionally high, with 280,000 physicians registered under verified real names to deliver online consultation services, rendering the platform a unique and well-suited context for investigating physician–patient interactions [15].

Guided by the platform's interface, we deployed a crawler to systematically extract information from the public profiles of 2050 physicians from September to December 2024. After rigorous data cleaning and outlier removal, we followed the coding protocol established by Herzenstein et al [16] to quantify physicians' self-disclosure. Four trained research assistants independently coded fourteen dichotomous disclosure variables. The dataset was split into 2 equal batches, with each assigned to a distinct pair of coders who worked in parallel. Only observations with unanimous agreement were retained, and all cases of coder disagreement were excluded. This intercoder validation yielded 1798 reliable observations for subsequent empirical analysis, with no missing values in the final analytic dataset—comfortably exceeding our preregistered target of 1000 evaluable records.

XSL•FO
RenderX

Variable Measurements

Dependent Variable

We used "Total visits"—the cumulative number of consultations, calls, and bookings shown on each physician's profile—as the dependent variable. This variable, therefore, provides a comprehensive foundation for examining the relationship between physician self-disclosure and patient decision-making. All variable measurement details can be found in Table S1 in Section 2 of Multimedia Appendix 1.

Independent Variable

Self-disclosure breadth was indexed by entropic weighting of 3 public signals, namely clinical performance (cp), academic experience (ae), and social reputation (sr). This composite measure captures the comprehensiveness of physicians' self-presentation strategies by integrating these fundamental aspects of credible identity. Self-disclosure depth was also indexed by entropic weighting of 3 important cues, such as expertise coverage (ec), expertise richness (er), and expertise granularity (eg). These 3 dimensions collectively capture the multifaceted nature of disclosure depth, as physicians may vary in how extensively they elaborate (coverage), how comprehensively they describe (richness), and how specifically they detail their expertise (granularity). Specifically, self-disclosure breadth and depth were calculated using the following formulas:

$$\text{Self-disclosure breadth} = \log(\alpha_1 * cp + \alpha_2 * ae + \alpha_3 * sr + 1)$$

$$\text{Self-disclosure depth} = \log(\alpha_4 * ec + \alpha_5 * er + \alpha_6 * eg + 1)$$

Moderating Variable

DHL measures the extent of digital technology integration within a city's health care infrastructure, reflecting the adoption of telemedicine platforms and mobile clinical tools in routine medical practice [17]. We operationalized this variable using the China Urban Digital Economy Index (Medical Chapter) [18], which is the most up-to-date DHL-relevant index dataset we can find, as a comprehensive assessment jointly published by the School of Management at Zhejiang University and the Digital Economy Research Centre of New H3C Group. Based on this index, each city's DHL is assigned a score from 1 to 5, with higher numbers representing higher levels of digital integration and technological advancement in digital health care. We assigned each physician the DHL score corresponding to their practice location, thereby capturing the digital maturity of their local digital health care environment, as detailed in Table S2 in Section 2 of Multimedia Appendix 1.

Controls

To separate the associations of physician self-disclosure from other factors related to patient selection, we included three control variables—professional title (title), physician popularity (popularity), and gift (gift)—to account for alternative explanations of patient decision-making. A brief summary of variable measurements is presented in Table 1.

Table 1. Brief summary of measurement of core study variables.

Variables	Measurements
Dependent variable	
Total visits	The cumulative count of all patient-initiated interactions with each physician across all service channels (online consultations, telephone consultations, and appointment bookings), as recorded on the platform.
Independent variable	
Self-disclosure breadth	Coded as 1 if at least one clinical component (clinical experience, clinical effectiveness, and clinical manner) was disclosed, and 0 otherwise.
Clinical performance	Coded as 1 if at least one academic component (research productivity, international training, and educational credentials) was disclosed, and 0 otherwise.
Academic experience	Coded as 1 if at least one social component (part-time positions, honors and awards) was disclosed, and 0 otherwise.
Self-disclosure depth	
Expertise coverage	The total length of description text in vocabularies.
Expertise richness	The number of disease types explicitly mentioned as areas of expertise in the physician's profile.
Expertise granularity	Coded as 1 if the description uses delimiters to distinguish specialties, and 0 otherwise.
Moderator	
Digital health care level	Each physician is assigned a score from 1 to 5 corresponding to their practice location.
Controls	
Title	Chief physicians receive 4, deputy chief physicians 3, attending physicians 2, and resident physicians 1.
Popularity	A composite recommendation score (0-5 continuous scale) generated by the platform.
Gift	Patients' real payment to the physician after receiving services. Take the records on the platform.

XSL•FO
RenderX

Modeling and Statistical Analysis

Research Model

Figure 1 depicts our research model. This study examines the relationship between physicians' self-disclosure breadth and depth and patient decision-making in digital health markets (H1) and the moderating role of DHL in these relationships (H2).

Figure 1. Research model.

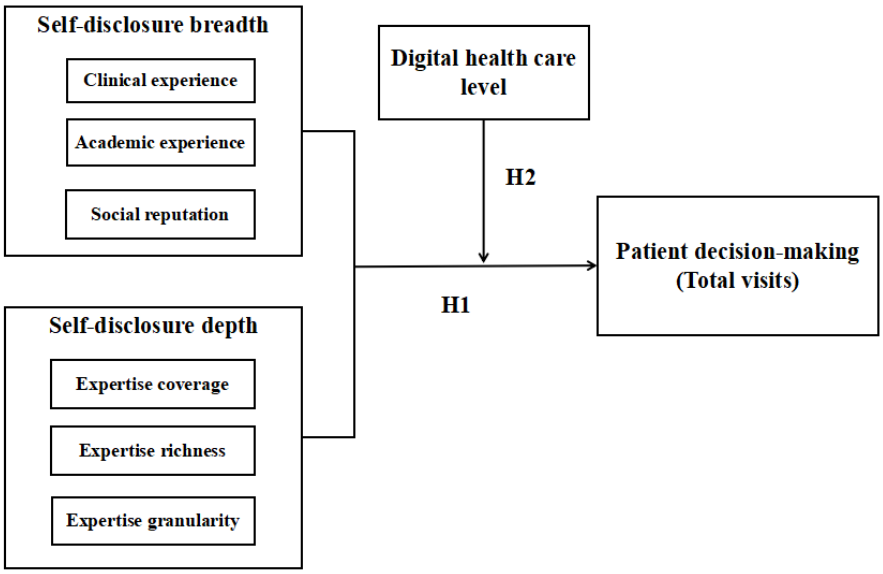

Ordinary Least Squares Model

Before testing the main and moderating effects, we first performed a descriptive statistical analysis to summarize the key characteristics of physicians and their self-disclosure behaviors. Subsequently, we conducted correlation analysis to assess the associations between key variables of interest. We then tested the main effects, focusing on how self-disclosure breadth and depth shape patient decision-making. Following this, we evaluated the moderating role of DHL, investigating whether regional digital health care conditions shape the effectiveness of these self-disclosure strategies. The empirical models pertaining to these tests were as follows:

- Model 1: Ln(Total visits) = β0 + β1Self-disclosure breadth + β2Title + β3Popularity + β4 Gift + ε
- Model 2: Ln(Total visits) = β0 + β1Self-disclosure depth + β2Title + β3Popularity + β4Gift + ε
- Model 3: Ln(Total visits) = β0 + β1Self-disclosure breadth + β2DHL + β3Self-disclosure breadth× DHL + β4Title + β5Popularity + β6Gift + ε
- Model 4: Ln(Total visits) = β0 + β1Self-disclosure depth + β2DHL + β3Self-disclosure depth × DHL + β4 Title + β5 Popularity + β6 Gift + ε

Our models include 3 control variables, including physician title, popularity metrics, and gift reception status, with ε representing the error term. We control for title because prior studies indicate that professional credentials significantly affect patient trust and provider selection [19]. Popularity serves as an important platform-based signal that may independently drive patient choices [20]. Gift reception reflects patient satisfaction from previous interactions, potentially influencing patients' future decision-making patterns [21].

We used ordinary least squares regression to estimate all models, with data analysis conducted in Stata (StataCorp LLC). To ensure comparability across variables with different scales, all variables were standardized before regression analysis.

Ethical Considerations

The Institutional Review Board of the International Business School, Beijing Foreign Studies University (001-2025-12-02) approved this study and granted an exemption from full human-subjects review. This study was conducted in accordance with the national ethical guidelines for research involving information data. Our use of legally obtained, fully anonymized public data, which contained no sensitive or commercial elements, qualified for an exemption from full ethics review as stipulated by the National Health Commission of the People's Republic of China [22]. No further approvals were required.

Researchers also confirmed that the original data-collection practices of Haodf are governed by the platform's user agreement and privacy notice, and private fields, such as physician names, clinical records, were anonymized and deidentified from the analytic file. No additional recruitment with physicians or patients occurred, no payments or incentives were offered to any physician or patient, and no identification of individual participants in any images of the manuscript or supplementary material is possible.

Results

Descriptive Statistics

Table 2 summarizes the central tendencies and dispersion of the key variables drawn from 1798 physician profiles on the Haodf platform. Min-Max denotes the actual minimum and maximum values in the sample. The IQR is calculated as the difference between the 75th and 25th percentiles and is reported

XSL•FO
RenderX

alongside the median (50th percentile). The results showed a skewed distribution of total visits per physician (mean 1471.83, SD 2486.38). In the subsequent data-processing pipeline, we applied appropriate normalization steps to mitigate the impact of extreme values and ensure the analytic dataset is well-behaved. Some physicians claim competence in clinical experience (mean 0.46, SD 0.50) and effectiveness (mean 0.32, SD 0.47), but only 4% (72/1798) explicitly mention clinical manner (mean 0.04, SD 0.20).

Table 2. Descriptive statistics of focal variables.

Variable	Mean (SD)	Median	Range
Total visits	1471.83 (2486.38)	680.5	1-27,352
Self-disclosure depth			
Expertise richness	3.76 (1.66)	4	0-8
Expertise coverage	8.16 (5.72)	7	0-58
Expertise granularity	0.51 (0.50)	1	0-1
Self-disclosure breadth			
Clinical experience	0.46 (0.50)	0	0-1
Clinical effectiveness	0.32 (0.47)	0	0-1
Clinical manner	0.04 (0.20)	0	0-1
Research productivity	0.75 (0.43)	1	0-1
International training	0.48 (0.50)	0	0-1
Educational credentials	0.60 (0.49)	1	0-1
Part-time positions	0.67 (0.47)	1	0-1
Honors and awards	0.44 (0.50)	0	0-1
Digital health care level	4.09 (1.01)	4	1-5
Title	3.35 (0.72)	3	1-4
Gift	128.65 (303.06)	39	0-4878
Popularity	4.14 (0.310)	4.1	3.4-5

Research productivity is reported by 75% (1349/1798) throughout the sample (mean 0.75, SD 0.43), whereas 48% (863/1798) list international training (mean 0.48, SD 0.50) and 60% (1079/1798) cite elite educational credentials (mean 0.60, SD 0.50). Roughly two-thirds hold part-time positions (mean 0.67, SD 0.47) and 44% (791/1798) have earned honors or awards (mean 0.44, SD 0.50). The average DHL is 4.09 (SD 1.01), and the mean title rank is 3.35 (SD 0.72), both approaching the upper end of their respective scales. Finally, popularity—an index computed by the platform—clusters tightly around 4.14 (SD 0.31), implying limited variance once the algorithmic score is normalized.

Pearson Correlation Analysis and Collinearity Testing

Table S3 in Section 2 of Multimedia Appendix 1 reports the Pearson correlations for 6 focal variables entering regression. Both self-disclosure depth and breadth are positively related to total visits (r=0.098; P<.001 and r=0.109; P<.001, respectively). The moderator DHL also positively relates to the dependent variable (r=0.119; P<.001). The VIF (variance inflation factor) analysis shows, afterwards in Table S4 in Section 2 of Multimedia Appendix 1, the largest value is 2.82 for gift, followed by 1.67 for popularity, whereas the remaining VIFs range only from 1.06 to 1.11. All VIF values are far below the conventional threshold of 5 (or 10) [23]. Taken together, the correlation matrix and the inflation factors jointly indicate that multicollinearity should not be a main concern for the stability or estimation of regression results.

Hypothesis Testing

We estimated 4 ordinary least squares models with heteroskedasticity-robust standard errors. The dependent variable, Total visits, was log-transformed to reduce skewness. All core continuous predictors, including self-disclosure breadth, self-disclosure depth, and DHL, were standardized (mean 0, SD 1) to facilitate coefficient comparability and to avert multicollinearity when interaction terms were introduced. Across all 4 models, the coefficients show the relative change in total visits associated with a one-standard-deviation shift in the focal variable. The results are reported in Table 3.

XSL•FO
RenderX

Table 3. Ordinary least squares regression results (Models 1-4) examining the relationship between physician self-disclosure breadth, depth, digital health care level, and patient decision-making.

Variable	Model 1	Model 2	Model 3	Model 4
Self-disclosure breadth, β (95% CI)	0.255[a] (0.054-0.456)	0.249[a] (0.047-0.450)	—[b]	—
Self-disclosure depth, β (95% CI)	—	—	0.098[c] (0.030-0.167)	0.092[c] (0.023-0.160)
DHL[d], β (95% CI)	—	–0.036 (–0.09 to 0.018)	—	–0.031 (–0.085 to 0.023)
Self-disclosure breadth×DHL, β (95% CI)	—	0.261[a] (0.061-0.461)	—	—
Self-disclosure depth×DHL, β (95% CI)	—	—	—	0.070[a] (0.002-0.138)
Title, β (95% CI)	0.248[c] (0.172-0.324)	0.242[c] (0.166-0.317)	0.276[c] (0.201-0.350)	0.275[c] (0.201-0.349)
Popularity, β (95% CI)	1.720[c] (1.50-1.94)	1.751[c] (1.532-1.970)	1.678[c] (1.460-1.900)	1.700[c] (1.480-1.921)
Gift, β (95% CI)	0.001[c] (0.001-0.001)	0.001[c] (0.001-0.001)	0.001[c] (0.001-0.001)	0.001[c] (0.001-001)
Constant, β (95% CI)	–1.825[c] (–2.720 to –0.930)	–1.938[c] (–2.843 to –1.031)	–1.748[c] (–2.648 to –0.847)	–1.838[c] (–2.749 to –0.926)
F test (*df*)	251.3 (6)	181.4 (8)	251.9 (6)	181.1 (8)
R^2	0.412	0.415	0.413	0.415

[a]The correlation is significant at a significance level of .05 (2-tailed).

[b]Not applicable.

[c]The correlation is significant at a significance level of .01 (2-tailed).

[d]DHL: digital health care level.

Model 1 establishes the baseline relationship for self-disclosure breadth, revealing a positive and statistically significant coefficient (β=0.255, 95% CI 0.054-0.456; *P*=.01), thereby confirming Hypothesis 1 that broader physician self-disclosure increases patient decision-making volume. Model 3 demonstrates a parallel finding for self-disclosure depth, with results showing a significant positive effect (β=0.098, 95% CI 0.030-0.167; *P*=.005), providing support for Hypothesis 1 that physician self-disclosure depth is positively associated with patient decision-making. Both findings confirm that comprehensive information disclosure, whether through diverse disclosure topics or detailed expertise presentation, enhances physician attractiveness to patients.

The interaction analyses reveal that DHL significantly amplifies self-disclosure effectiveness. Model 2 introduces the breadth×DHL interaction term, yielding a positive and significant coefficient (β=0.261, 95% CI 0.061-0.461; *P*=.01), which supports Hypothesis 2 that DHL strengthens the relationship between self-disclosure breadth and patient decision-making. Similarly, Model 4 demonstrates that the depth×DHL interaction is positive and significant (β=0.070, 95% CI 0.002-0.138; *P*=.045), corroborating Hypothesis 2 that DHL enhances the effectiveness of disclosure depth. Notably, the breadth interaction effect is substantially larger than the depth interaction effect, suggesting that DHL provides greater amplification benefits for diverse disclosure strategies compared to detailed expertise presentation.

Post Hoc Analysis

To better understand the nuanced mechanisms underlying the moderating effects of DHL, we conducted post hoc analysis using interaction plots. While our main regression results demonstrate statistically significant moderation effects, visualizing these interactions provides deeper insights into how DHL moderates the strength and nature of the relationships between physician self-disclosure strategies and patient decision-making. The visualized graphs are presented in Figures 2 and 3, where high and low DHL correspond to values of 1 SD above and below the mean (+1 and –1).

XSL•FO
RenderX

Figure 2. The moderating effect of digital health care level on the association between physician self-disclosure breadth and patient decision-making. DHL: digital health care level.

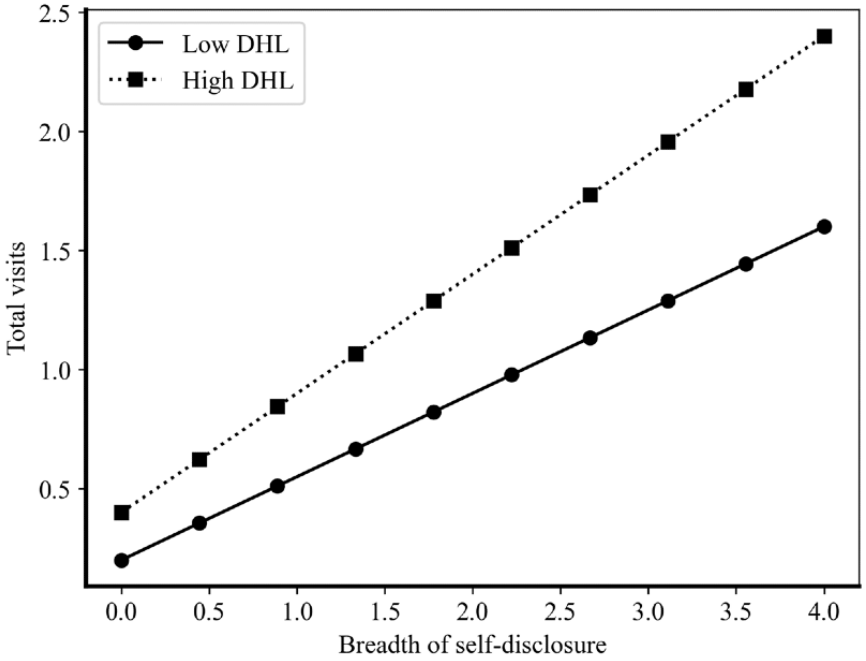

Figure 3. The moderating effect of digital health care level on the association between physician self-disclosure depth and patient decision-making. DHL: digital hhealth care level.

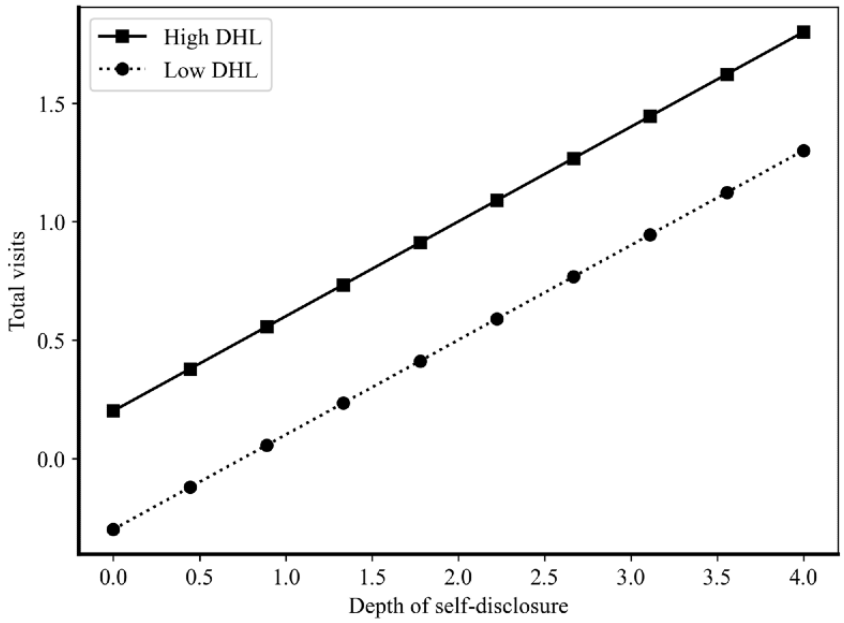

Figure 2 reveals particularly prominent insights about breadth moderation. Most notably, the gap between high and low DHL conditions grows increasingly larger as self-disclosure breadth increases, creating a pronounced divergence pattern that suggests breadth disclosure may be more sensitive to DHL than depth disclosure. In high DHL environments, patients appear exceptionally responsive to broad professional presentations. This amplified responsiveness may reflect enhanced information processing capabilities and greater appreciation for diverse professional information. Conversely, in low DHL environments, increasing breadth yields relatively restricted returns, potentially due to limited digital literacy or infrastructure constraints that prevent effective usage of comprehensive professional profiles. The steeper moderation gap for breadth compared to depth suggests that while diverse disclosure topics (breadth) become disproportionately valuable when supported by higher DHL, detailed expertise presentations (depth) maintain more consistent effectiveness across different DHL contexts. This creates a "digital divide" effect where technological advancement in health care becomes a critical prerequisite for breadth disclosure effectiveness.

Figure 3 provides compelling visual evidence for the DHL moderation effect on self-disclosure depth, revealing a striking divergence in self-disclosure effectiveness across DHL levels. In high DHL environments, the relationship between

XSL•FO
RenderX

self-disclosure depth and total visits exhibits a pronounced higher slope, demonstrating that each incremental increase in depth disclosure generates greater patient engagement gains. Conversely, in low DHL contexts, the relationship remains relatively lower, suggesting that detailed expertise presentations yield fewer additional benefits when digital health care is underdeveloped. This pattern indicates that physicians practicing in technologically advanced health care environments can leverage detailed self-disclosure strategies—spanning specialized medical capabilities, in-depth professional competencies, and granular clinical expertise—to achieve disproportionately greater patient attraction, while those in less digitized environments may find such detailed disclosure strategies less rewarding.

Robustness Checks

We used various methods for robustness checks to ensure the reliability and consistency of our research findings, including variable substitution, bootstrap resampling, subsampling, and the winsorizing technique. The core findings are summarized in Table 4 below. The procedures and interpretations for each approach are detailed below.

To ensure the reliability of our findings, we first conducted a robustness check by replacing our categorical DHL measurement with continuous digital health care scores, which were also presented in the report published by China Urban Digital Economy Index (Medical Chapter) [18], as shown in Table S5 of Section 3 in Multimedia Appendix 1. The detailed results of this first variable substitution check are shown immediately afterwards in Table S6, demonstrating consistent patterns across all model specifications. Self-disclosure breadth continues to be positively associated with total visits across Models 1 and 2 (β=0.255, 95% CI 0.054-0.456; *P*=.01 and β=0.258, 95% CI 0.057-0.459; *P*=.01, respectively), while self-disclosure depth demonstrates similar consistency in Models 3 and 4 (β=0.098, 95% CI 0.030-0.167; *P*=.005 and β=0.096, 95% CI 0.027-0.164; *P*=.006, respectively). The preservation of both effect magnitudes and significance levels indicates that our core findings are not artifacts of the initial categorical operationalization.

To rule out the possibility that physicians attract more consultations simply because they are more active or popular, we further re-estimated all models by replacing the dependent variable with Ln (total visits-per-popularity, ie, total visits divided by popularity), where Popularity is a platform-computed index that combines physician activity, patient recommendations, and review ratings to reflect overall physician popularity. Table S7 shows that the core disclosure variables remain positively significant across the 4 specifications, with Models 1 and 2 (β=0.494, 95% CI 0.290-0.699; *P*<.001 and β=0.482, 95% CI 0.277-0.687; *P*<.001, respectively); and Models 3 and 4 (β=0.157, 95% CI 0.086-0.228; *P*<.001 and β=0.151, 95% CI 0.080-0.222; *P*<.001, respectively). The persistent significance of self-disclosure after controlling for popularity corroborates our main results and demonstrates that, even after adjusting for physicians' baseline visibility and popularity, patients show a preference for physicians who provide greater breadth and depth of professional information.

To validate that our significance tests are not dependent on distributional assumptions, we re-estimated all models using bootstrap resampling with 1000 replications. Table S8 presents the bootstrapped results, demonstrating the robustness of our statistical inferences. First, self-disclosure breadth maintains its positive and significant relationship with total visits in both Models 1 and 2 (β=0.255, 95% CI 0.050-0.460; *P*=.02 and β=0.249, 95% CI 0.049-0.448; *P*=.02, respectively), while self-disclosure depth demonstrates identical significance patterns in Models 3 and 4 (β=0.098, 95% CI 0.294-0.167; *P*=.005 and β=0.092, 95% CI 0.026-0.258; *P*=.006, respectively). Besides, the moderation effects remain significant under bootstrap estimation. The breadth×DHL interaction retains its positive and significant coefficient (β=0.261, 95% CI 0.054-0.467; *P*=.01) in Model 2, while the depth×DHL interaction similarly maintains significance (β=0.070, 95% CI 0.007-0.133; *P*=.03) in Model 4.

To address potential concerns that our main findings might be driven by senior physicians who possess inherently greater credibility and resources, we conduct a robustness check by restricting our analysis to non–chief physicians only. As shown in Table S9, the detailed results of subsample analysis (non–chief physicians) show consistent effects across the restricted sample, confirming the robustness of our main findings. Self-disclosure breadth maintains its positive and significant relationship with total visits in both Models 1 and 2 (β=0.345, 95% CI 0.074-0.615; *P*=.01 and β=0.345, 95% CI 0.075-0.614; *P*=.01 respectively), while self-disclosure depth similarly shows robust positive effects in Models 3 and 4 (β=0.125, 95% CI 0.032-0.219; *P*=.009 and β=0.124, 95% CI 0.029-0.215; *P*=.01, respectively). Notably, the coefficient magnitudes are actually larger in this subsample compared to the full sample, suggesting that self-disclosure strategies may be even more crucial for physicians with lower hierarchies.

To address potential concerns about extreme values influencing our results, we re-estimated all models after winsorizing the top and bottom 10% of each variable. Table S10 presents the detailed results from this outlier-treatment approach. Self-disclosure breadth maintains its positive and significant effects across Models 1 and 2 (β=0.250, 95% CI 0.030-0.470; *P*=.03 and β=0.248, 95% CI 0.028-0.468; *P*=.03, respectively), while self-disclosure depth similarly preserves its significant positive relationship in Models 3 and 4 (β=0.106, 95% CI 0.032-0.180; *P*=.005 and β=0.101, 95% CI 0.027-0.175; *P*=.008, respectively). The coefficient magnitudes remain virtually identical to our original estimates, confirming that extreme values do not drive the main effect conclusions.

XSL•FO
RenderX

Table 4. A brief summary of the focal results of our 5 robustness tests.

Robust check	Self-disclosure breadth, β (95% CI)	Self-disclosure depth, β (95% CI)	Self-disclosure breadth×DHL[a], β (95% CI)	Self-disclosure depth×DHL, β (95% CI)
R1[b]				
Model 1	0.255[c] (0.054-0.456)	—[d]	—	—
Model 2	0.258[c] (0.057-0.459)	—	0.020[e] (0.006-0.034)	—
Model 3	—[c]	0.098[e] (0.030-0.167)	—	—
Model 4	—	0.096[e] (0.027-0.164)	—	0.006[c] (0.001-0.010)
R2[f]				
Model 1	0.494[e] (0.290-0.699)	—	—	—
Model 2	0.482[e] (0.277-0.687)	—	0.030[g] (0.002-0.063)	—
Model 3	—	0.157[e] (0.086-0.228)	—	—
Model 4	—	0.151[e] (0.080-0.222)	—	0.071[g] (0.001-0.142)
R3[h]				
Model 1	0.255[c] (0.050-0.460)	—	—	—
Model 2	0.249[c] (0.049-0.448)	—	0.261[c] (0.054-0.467)	—
Model 3	—	0.098[e] (0.029-0.167)	—	—
Model 4	—	0.092[e] (0.026-0.258)	—	0.070[c] (0.007-0.133)
R4[i]				
Model 1	0.345[c] (0.074-0.615)	—	—	—
Model 2	0.345[c] (0.075-0.614)	—	0.259[c] (0.001-0.518)	—
Model 3	—	0.125[e] (0.032-0.219)	—	—
Model 4	—	0.124[c] (0.029-0.215)	—	0.086[c] (0.001-0.173)
R5[j]				
Model 1	0.250[c] (0.030-0.470)	—	—	—

XSL•FO
RenderX

Robust check	Self-disclosure breadth, β (95% CI)	Self-disclosure depth, β (95% CI)	Self-disclosure breadth×DHL[a], β (95% CI)	Self-disclosure depth×DHL, β (95% CI)
Model 2	0.248[c] (0.028-0.468)	—	0.285[c] (0.066-0.504)	—
Model 3	—	0.106[e] (0.032-0.180)	—	—
Model 4	—	0.101[e] (0.027-0.175)	—	0.076[c] (0.003-0.149)

[a]DHL: digital health care level.

[b]Replace categorical DHL measurement with continuous digital health care scores.

[c]The correlation is significant at a significance level of .05 (2-tailed).

[d]Not applicable.

[e]The correlation is significant at a significance level of .01 (2-tailed).

[f]Replace the dependent variable with Log_e (total visits-per-popularity).

[g]The correlation is significant at a significance level of .1 (2-tailed).

[h]Re-estimated all models using bootstrap resampling with 1000 replications.

[i]Restricted our analysis to non–chief physicians only.

[j]Re-estimated all models after winsorizing the top and bottom 10% of each variable.

Discussion

Principal Findings

This study demonstrates that physicians can strategically shape patient decision-making through purposeful self-disclosure behaviors within OHCs. Our empirical analysis from China's leading online health platform reveals that both self-disclosure breadth and depth significantly increase patient visits, challenging the prevailing view of physicians as passive recipients of online reviews. Most critically, the regional DHL fundamentally moderates these relationships. In cities with advanced digital development, both breadth and depth effects are substantially amplified, while regions with limited digital development show diminished returns for the same disclosure strategies. These findings reconceptualize physician agency within digital health care platforms, demonstrating that strategic self-disclosure represents a viable patient acquisition mechanism rather than passive information provision.

Theoretical Implications

The study advances theoretical understanding of physician self-disclosure in digital health markets by embedding its 3 key contributions within—and explicitly contrasting them against—the extant literatures on physician agency, self-disclosure theory, and OHCs.

First, contrary to prior studies that predominantly frame physicians as passive recipients of online feedback [4,24,25], our reconceptualization of physician agency aligns with a growing but still underdeveloped stream of research that emphasizes the proactive role of service providers in digital platforms. For instance, recent work by Ouyang and Wang [26] and Lu and Wu [27] suggests that physicians can influence patient perceptions through profile customization and online engagement. This challenges the prevailing passive-physician paradigm in medical marketing research and establishes physicians as active agents capable of patient acquisition through evidence-based self-presentation strategies [28,29]. However, extant research has predominantly conceptualized online reputation management as a reactive endeavor—centering on prompt responses to negative feedback [30], remediation of service failures [31], or post hoc optimization of profile completeness [32]—while overlooking proactive self-disclosure breadth and depth as ex ante instruments that physicians can strategically deploy to attract new patients before any review is written. Our findings go beyond these accounts by revealing that physicians can deploy proactive, ex ante self-disclosure—systematically foregrounding personal credentials, institutional affiliations, and succinct expertise narratives—to pre-emptively sculpt patient trust and choice before any review is written or any service failure occurs. This reorients the theoretical lens from "damage control" to "impression engineering," recasting self-disclosure as a forward-looking signaling mechanism that anticipates patient heuristics rather than remedying prior dissatisfaction. By establishing strategic disclosure as a feasible alternative to reactive reputation management, we push the conceptual boundary of physician behavior in digital environments beyond traditional service-recovery frameworks and toward a predictive, marketing-as-signal paradigm.

Second, we extend self-disclosure theory beyond its traditional interpersonal communication context into health care settings while establishing critical boundary conditions [7,33,34]. We introduce the regional DHL as a critical boundary condition—a factor largely overlooked in prior self-disclosure research. While earlier studies have examined individual-level moderators, such as gender [35] or cultural orientation [36], our findings reveal that macrolevel technological development significantly moderates the disclosure-outcome relationship. This aligns with recent macrosociological perspectives on digital inequality [37], which argue that the same online behavior can yield divergent outcomes depending on the technological context in which it

XSL•FO
RenderX

is embedded. Thus, our study not only extends self-disclosure theory into a new domain but also redefines its boundary conditions by incorporating sociotechnological contingencies. Our findings confirm that self-disclosure breadth and depth operate in professional contexts, yet their effectiveness is contingent on the regional DHL. By demonstrating this macrolevel technological moderation, we extend the theory's scope and expose context boundaries previously overlooked [38,39].

Third, the prevailing view in the OHC literature maintains that these platforms mainly mitigate physician-patient information asymmetry by aggregating ratings, reviews, and outcome data [24,40,41]. We extend this perspective by demonstrating that physicians can actively reconfigure the information environment through strategic self-disclosure. Especially, our findings show that, within Chinese OHCs, breadth of disclosure markedly outweighs depth, as an expansive array of credential signals consistently exhibits a greater positive association with patient engagement than rich narrative detail does. Patients appear to follow a hierarchical signaling model in which credential heuristics operate as an initial, low-cognitive gatekeeper, with hospital tier, academic rank, and prestigious awards are rapidly recoded into a binary "pass-fail" filter that determines inclusion in the consideration set. Only after clearing this threshold do patients allocate scarce attentional resources to elaborately process narrative depth—articulations of treatment philosophy, detailed case histories, or other discursive evidence of clinical expertise. Empirically, the stronger main effect of self-disclosure breadth compared to depth evidences that, in the context of Chinese OHCs, patients may initially rely more heavily on credential heuristics—such as titles, affiliations, and awards [42,43]—as efficient signals of quality, before engaging with the more cognitively demanding narratives of detailed expertise. This insight complements and extends recent work by Wang et al [44], who find that physicians with more comprehensive profiles receive more appointment requests, thereby offering a mechanistic account of why breadth outperforms depth in Chinese OHCs.

Practical Implications

This study also provides several practical implications for multiple stakeholder groups seeking to enhance physician self-disclosure effectiveness and patient decision-making within OHCs.

For physicians, our findings emphasize a context-dependent approach to online self-disclosure [45-47]. Those in high-DHL regions should leverage comprehensive breadth—showcasing credentials across clinical, academic, and social domains—and enrich their profiles with multimedia depth, such as video introductions. In low-DHL regions, the priority shifts to maximizing clarity. Physicians should focus on core breadth elements like clinical experience and use concise, text-based depth with scannable lists of expertise to ensure accessibility.

For OHC platform designers, our results highlight the limitation of a uniform profile design and advocate for context-aware systems [48,49]. Actionable recommendations include implementing structured disclosure templates to guide physicians in highlighting decision-critical information, introducing visual verification badges for credentials to enhance trust, and developing tiered interface modes—a feature-rich version for high-DHL users and a streamlined, text-optimized version for regions with limited DHL.

For policymakers, our evidence on the moderating role of DHL underscores that digital infrastructure is a social determinant of health access. Specific interventions should prioritize closing the digital divide by investing in high-speed internet infrastructure in underserved areas, launching public health campaigns to improve patients' digital health literacy, and creating financial incentives for clinics and physicians in low-DHL regions to adopt and master digital consultation tools.

Limitations

This study acknowledges certain limitations. First, the cross-sectional design limits causal inference, leaving the temporal dynamics of these relationships unclear. Second, the DHL measurement uses a city-level categorization rather than granular technological indicators, which prevents identifying the specific infrastructure components that most critically moderate disclosure effectiveness. Third, owing to the absence of more recent publicly available datasets, the DHL indicators used in this study remain those published in 2021. The 3-year lag may introduce slight discrepancies with present-day infrastructure levels. Updated figures can further corroborate our findings. Finally, focus on China's online health platform may limit generalizability across different cultural and regulatory contexts. The findings may reflect sociocultural norms specific to Chinese health care markets, such as a pronounced hierarchy in physician-patient dynamics or a strong preference for credential-based trust signals.

Conclusions

This study is innovative in reconceptualizing physicians as strategic agents capable of actively shaping patient decision-making through purposeful self-disclosure in digital health markets. Different from existing studies that treat physicians as passive recipients of online ratings and reviews, our research demonstrates that physicians can strategically shape patient acquisition through deliberate self-disclosure breadth and depth. This study thus brings new insights to digital health markets by demonstrating that self-disclosure operates as a viable patient acquisition mechanism in professional health care contexts, wherein DHL acts as a critical boundary condition that fundamentally moderates the breadth-depth relationships. The findings have significant implications in the real world: (1) physicians can leverage evidence-based disclosure strategies for patient acquisition, (2) platform designers should implement context-adaptive features optimizing effectiveness across heterogeneous digital environments, and (3) policymakers should prioritize digital infrastructure investments to systematically enhance physicians' competitive capabilities and patient decision-making quality.

XSL•FO
RenderX

Acknowledgments

The authors gratefully acknowledge the guidance received from the editor and anonymous reviewers.

Funding

This study was supported by the Humanities and Social Science Fund of Ministry of Education of China (grant 22YJA630018); the Fundamental Research Funds for the Central Universities (grants 2022JJ007 and 2025ZZ022).

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: QL
Methodology: QL
Data curation: PY
Formal analysis: PY
Software: PY
Visualization: PY
Writing–original draft: QL
Writing–review & editing: QL, PY
Funding acquisition: QL, JF
Resources: JF
Supervision: JF

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details on literature review, variable measurements and robustness check results.
[DOCX File , 114 KB - jmir_v28i1e84963_app1.docx]

References

1. El-Tallawy SN, Pergolizzi JV, Vasiliu-Feltes I, Ahmed RS, LeQuang JK, Alzahrani T, et al. Innovative applications of telemedicine and other digital health solutions in pain management: a literature review. Pain Ther 2024;13(4):791-812. [doi: 10.1007/s40122-024-00620-7] [Medline: 38869690]
2. Digital health industry data book: Telehealth care, mHealth, health care analytics, and digital health systems—market size, share, trend analysis, and segment forecasts, 2022–2030. Grand View Research. 2022. URL: https://www.grandviewresearch.com/sector-report/digital-health-industry-data-book [accessed 2025-04-21]
3. Peng Y, Yin P, Deng Z, Wang R. Patient-physician interaction and trust in online health community: the role of perceived usefulness of health information and services. Int J Environ Res Public Health 2019;17(1):139 [FREE Full text] [doi: 10.3390/ijerph17010139] [Medline: 31878145]
4. Wu DC, Zhao X, Wu J. Online physician-patient interaction and patient satisfaction: empirical study of the internet hospital service. J Med Internet Res 2023;25:e39089 [FREE Full text] [doi: 10.2196/39089] [Medline: 37616031]
5. Fitzpatrick PJ. Improving health literacy using the power of digital communications to achieve better health outcomes for patients and practitioners. Front Digit Health 2023;5:1264780 [FREE Full text] [doi: 10.3389/fdgth.2023.1264780] [Medline: 38046643]
6. Ruangkanjanases A, Sivarak O, Jong D, Zhou Y. The effect of self-disclosure on mass trust through TikTok: an empirical study of short video streaming application users. Front Psychol 2022;13:968558 [FREE Full text] [doi: 10.3389/fpsyg.2022.968558] [Medline: 36059732]
7. Alrabiah S. Urrgh.. Will They Stop Going on About Their Relationships?: An Examination of Self-Disclosure by Travel Influencers on Consumer Outcomes. Edinburgh, Scotland: The University of Edinburgh; 2021.
8. Sundell E, Wångdahl J, Grauman Å. Health literacy and digital health information-seeking behavior - a cross-sectional study among highly educated Swedes. BMC Public Health 2022;22(1):2278 [FREE Full text] [doi: 10.1186/s12889-022-14751-z] [Medline: 36471284]
9. Saeed SA, Masters RM. Disparities in health care and the digital divide. Curr Psychiatry Rep 2021;23(9):61 [FREE Full text] [doi: 10.1007/s11920-021-01274-4] [Medline: 34297202]

XSL•FO
RenderX

10. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: the APA publications and communications board task force report. Am Psychol 2018;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
11. Sadiq IZ, Usman A, Muhammad A, Ahmad KH. Sample size calculation in biomedical, clinical and biological sciences research. J.Umm Al-Qura Univ. Appll. Sci 2024;11(1):133-141. [doi: 10.1007/s43994-024-00153-x]
12. Yu M, Shi X, Zou B, An S. [Comparison of 7 methods for sample size determination based on confidence interval estimation for a single proportion]. Nan Fang Yi Ke Da Xue Xue Bao 2023;43(1):105-110 [FREE Full text] [doi: 10.12122/j.issn.1673-4254.2023.01.14] [Medline: 36856217]
13. Li Y, Cen J, Wu J, Tang M, Guo J, Hang J, et al. The degree of anxiety and depression in patients with cardiovascular diseases as assessed using a mobile app: cross-sectional study. J Med Internet Res 2023;25:e48750 [FREE Full text] [doi: 10.2196/48750] [Medline: 37792455]
14. Zhang C, Zhou Y, Ke Y, Zhang W, Qin M, Alias H, et al. Fertility intentions in the era of the new three-child policy in China: a cross-sectional survey of married adults of reproductive age. Front Public Health 2025;13:1674687 [FREE Full text] [doi: 10.3389/fpubh.2025.1674687] [Medline: 41323596]
15. Li Y, Yan X, Song X. Provision of paid web-based medical consultation in China: cross-sectional analysis of data from a medical consultation website. J Med Internet Res 2019;21(6):e12126 [FREE Full text] [doi: 10.2196/12126] [Medline: 31162129]
16. Herzenstein M, Dholakia UM, Andrews RL. Strategic herding behavior in peer-to-peer loan auctions. J Interact Mark 2011;25(1):27-36. [doi: 10.1016/j.intmar.2010.07.001]
17. Yao L, Li Q, Li Q, Wang T, Peng S, Fu X, et al. Factors influencing the adoption of telemedicine services among middle-aged and older patients with chronic conditions in rural China: a multicentre cross-sectional study. BMC Health Serv Res 2025;25(1):775 [FREE Full text] [doi: 10.1186/s12913-025-12931-2] [Medline: 40448164]
18. Zhejiang University, New H3C Group. China Urban Digital Economy Index (Medical Chapter). URL: https://roadshow.h3c.com/zl/pdf/ylbps.pdf [accessed 2025-11-15]
19. Yu X, Wang H, Chen Z. The role of user-generated content in the sustainable development of online healthcare communities: exploring the moderating influence of signals. Sustainability 2024;16(9):3739. [doi: 10.3390/su16093739]
20. Yan J, Liang C, Gu D, Zhu K, Zhou P. Understanding patients' doctor choice behavior: elaboration likelihood perspective. Inf Dev 2024:02666669241259126. [doi: 10.1177/02666669241259126]
21. Wang Y, Wu H, Xia C, Lu N. Impact of the price of gifts from patients on physicians? service quality in online consultations: empirical study based on social exchange theory. J Med Internet Res 2019;22(5):e15685. [doi: 10.2196/15685]
22. Notice on issuing the measures for ethical review of life science and medical research involving human subjects. National Health Commission. 2025. URL: https://www.nhc.gov.cn/qjjys/c100016/202302/6b6e447b3edc4338856c9a652a85f44b.shtml [accessed 2025-12-11]
23. Kong M, Wang Y, Li M, Yao Z. Mechanism assessment of physician discourse strategies and patient consultation behaviors on online health platforms: mixed methods study. J Med Internet Res 2025;27:e54516 [FREE Full text] [doi: 10.2196/54516] [Medline: 40106798]
24. Song M, Elson J, Bastola D. Digital age transformation in patient-physician communication: 25-year narrative review (1999-2023). J Med Internet Res 2025;27:e60512 [FREE Full text] [doi: 10.2196/60512] [Medline: 39819592]
25. Guetz B, Bidmon S. The credibility of physician rating websites: a systematic literature review. Health Policy 2023;132:104821 [FREE Full text] [doi: 10.1016/j.healthpol.2023.104821] [Medline: 37084700]
26. Ouyang P, Wang J. Physician's online image and patient's choice in the online health community. Internet Res 2022;32(6):1952-1977. [doi: 10.1108/intr-04-2021-0251]
27. Lu W, Wu H. How online reviews and services affect physician outpatient visits: content analysis of evidence from two online health care communities. JMIR Med Inform 2019;7(4):e16185 [FREE Full text] [doi: 10.2196/16185] [Medline: 31789597]
28. Ouyang P, Wang J, Jasmine Chang A. Patients need emotional support: mmanaging physician disclosure information to attract more patients. Int J Med Inform 2022;158:104674. [doi: 10.1016/j.ijmedinf.2021.104674] [Medline: 34968960]
29. Duan Y, Gong S. Active medicine and passive medicine. Res Sq 2025:1-36. [doi: 10.13140/RG.2.2.22021.87527]
30. Han X, Lin Y, Han W, Liao K, Mei K. Effect of negative online reviews and physician responses on health consumers' choice: experimental study. J Med Internet Res 2024;26:e46713 [FREE Full text] [doi: 10.2196/46713] [Medline: 38470465]
31. van der Schyff EL, Ridout B, Amon KL, Forsyth R, Campbell AJ. Providing self-led mental health support through an artificial intelligence-powered chat bot (leora) to meet the demand of mental health care. J Med Internet Res 2023;25:e46448 [FREE Full text] [doi: 10.2196/46448] [Medline: 37335608]
32. Burke LE, Shiffman S, Music E, Styn MA, Kriska A, Smailagic A, et al. Ecological momentary assessment in behavioral research: addressing technological and human participant challenges. J Med Internet Res 2017;19(3):e77 [FREE Full text] [doi: 10.2196/jmir.7138] [Medline: 28298264]
33. Jo E, Jeong Y, Park S, Epstein D, Kim Y. Understanding the Impact of Long-Term Memory on Self-Disclosure with Large Language Model-Driven Chatbots for Public Health Intervention. 2024 Presented at: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems; May 11-16, 2024; New York, NY p. 1-21. [doi: 10.1145/3613904.3642420]

XSL•FO
RenderX

34. Kadji K, Schmid Mast M. The effect of physician self-disclosure on patient self-disclosure and patient perceptions of the physician. Patient Educ Couns 2021;104(9):2224-2231 [FREE Full text] [doi: 10.1016/j.pec.2021.02.030] [Medline: 33775499]
35. Nowell C, Pfeifer JH, Enticott P, Silk T, Vijayakumar N. Value of self-disclosure to parents and peers during adolescence. J Res Adolesc 2023;33(1):289-301. [doi: 10.1111/jora.12803] [Medline: 36166491]
36. Koch A, Rabins M, Messina J, Brennan-Cook J. Exploring the challenges of sexual orientation disclosure among lesbian, gay, bisexual, transgender, queer individuals. J Nurse Pract 2023;19(10):104765. [doi: 10.1016/j.nurpra.2023.104765]
37. Wu M, Xue Y, Ma C. The association between the digital divide and health inequalities among older adults in China: nationally representative cross-sectional survey. J Med Internet Res 2025;27:e62645 [FREE Full text] [doi: 10.2196/62645] [Medline: 39813666]
38. Liu J, He J, He S, Li C, Yu C, Li Q. Patients' self-disclosure positively influences the establishment of patients' trust in physicians: an empirical study of computer-mediated communication in an online health community. Front Public Health 2022;10:823692 [FREE Full text] [doi: 10.3389/fpubh.2022.823692] [Medline: 35145943]
39. Steele Gray C, Ramachandran M, Brinton C, Forte M, Loganathan M, Walsh R, et al. Digitally mediated relationships: how social representation in technology influences the therapeutic relationship in primary care. Soc Sci Med 2024;353:116962 [FREE Full text] [doi: 10.1016/j.socscimed.2024.116962] [Medline: 38908092]
40. Li H. Asymmetric information in the field of healthcare. Adv Econ Manag Polit Sci 2024;100(1):28-34. [doi: 10.54254/2754-1169/100/20241089]
41. Guo S, Wang K, Yang L, Dang Y. Extending signaling theory in online health communities to address medical information asymmetry: systematic review with narrative synthesis. J Med Internet Res 2025;27:e73208 [FREE Full text] [doi: 10.2196/73208] [Medline: 40749179]
42. Daraz L, Dogu C, Houde V, Bouseh S, Morshed KG. Assessing credibility: quality criteria for patients, caregivers, and the public in online health information-a qualitative study. J Patient Exp 2024;11:23743735241259440 [FREE Full text] [doi: 10.1177/23743735241259440] [Medline: 38827225]
43. van Velsen L, Flierman I, Tabak M. The formation of patient trust and its transference to online health services: the case of a Dutch online patient portal for rehabilitation care. BMC Med Inform Decis Mak 2021;21(1):188 [FREE Full text] [doi: 10.1186/s12911-021-01552-4] [Medline: 34118919]
44. Wang H, Jin J, Li L, Liu J, Wang D. Driving online healthcare growth amid the digital divide: how trust in professional signals from doctor biographies shapes patient decisions. Healthcare (Basel) 2025;13(12):1418 [FREE Full text] [doi: 10.3390/healthcare13121418] [Medline: 40565445]
45. Yuchao W, Ying Z, Liao Z. Health privacy information self-disclosure in online health community. Front Public Health 2020;8:602792 [FREE Full text] [doi: 10.3389/fpubh.2020.602792] [Medline: 33614566]
46. González-Pérez A, Matey-Sanz M, Granell C, Díaz-Sanahuja L, Bretón-López J, Casteleyn S. AwarNS: A framework for developing context-aware reactive mobile applications for health and mental health. J Biomed Inform 2023;141:104359 [FREE Full text] [doi: 10.1016/j.jbi.2023.104359] [Medline: 37044134]
47. Wyant K, Moshontz H, Ward SB, Fronk GE, Curtin JJ. Acceptability of personal sensing among people with alcohol use disorder: observational study. JMIR Mhealth Uhealth 2023;11:e41833 [FREE Full text] [doi: 10.2196/41833] [Medline: 37639300]
48. Tai-Seale M, Cheung M, Vaida F, Ruo B, Walker A, Rosen RL, et al. Patient-clinician communication interventions across multiple primary care sites: a cluster randomized clinical trial. JAMA Health Forum 2024;5(12):e244436. [doi: 10.1001/jamahealthforum.2024.4436] [Medline: 39671203]
49. Mostafapour M, Fortier JH, Garber G. Exploring the dynamics of physician-patient relationships: factors affecting patient satisfaction and complaints. J Healthc Risk Manag 2024;43(4):16-25. [doi: 10.1002/jhrm.21567] [Medline: 38706117]

Abbreviations

DHL: digital health care level
JARS: Journal Article Reporting Standards
OHC: online health community
VIF: variance inflation factor

XSL•FO
RenderX

Edited by S Brini; submitted 04.Oct.2025; peer-reviewed by Y Wan, Y Li; comments to author 30.Oct.2025; accepted 28.Dec.2025; published 29.Jan.2026.

Please cite as:
Liu Q, Yin P, Fan J
The Relationship Between Physician Self-Disclosure and Patient Acquisition in Digital Health Markets: Cross-Sectional Study
J Med Internet Res 2026;28:e84963
URL: https://www.jmir.org/2026/1/e84963
doi:10.2196/84963
PMID:

©Quanchen Liu, Pengqing Yin, Jing Fan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Website Use and Associations With Behavior Change and Weight Loss in Cancer Survivors and Their Partners: Secondary Analysis of a Randomized Controlled Trial

Harleen Kaur[1], PhD; Dori Pekmezi[2], PhD; Tracy E Crane[3], PhD, RD; David Farrell[4], MPH; Laura Q Rogers[5], MD, MPH; Wendy Demark-Wahnefried[6], PhD, RD

[1]Department of Psychology and Medicine, Division of Medical Oncology, Miller School of Medicine, University of Miami, Miami, FL, United States

[2]Department of Health Behavior, School of Public Health, University of Alabama at Birmingham, Birmingham, AL, United States

[3]Department of Medicine, Division of Medical Oncology, Miller School of Medicine, University of Miami, Miami, FL, United States

[4]People Designs (United States), Durham, NC, United States

[5]Department of Medicine, Division of General Internal Medicine and Population Science, School of Medicine, University of Alabama at Birmingham, Birmingham, AL, United States

[6]Department of Nutrition Sciences, School of Health Professions, University of Alabama at Birmingham, Birmingham, AL, United States

Corresponding Author:
Harleen Kaur, PhD
Department of Psychology and Medicine, Division of Medical Oncology
Miller School of Medicine
University of Miami
1425 NW Tenth Ave
Miami, FL, 33136
United States
Phone: 1 8323169470
Email: hxk629@miami.edu

Abstract

Background: Web-based lifestyle interventions to promote healthy diet and physical activity among cancer survivors and their partners are recent developments; therefore, few studies have reported patterns of website use or associations with behavior change.

Objective: The primary aim was to describe website use in the DUET (Daughters, Dudes, Mothers, and Others Together) trial and examine the associations between website use and changes in diet quality, moderate to vigorous physical activity (MVPA), and body weight.

Methods: This secondary analysis used data from 28 survivor-partner dyads (BMI ≥25 kg/m^2) randomized to the 6-month DUET web-based weight loss intervention, which released weekly e-learning sessions on diet and exercise. Website use was quantified as weeks of access, time spent, and frequency of page views. Diet quality was assessed using 2-day dietary recalls; MVPA was measured by the Godin Leisure-Time Exercise Questionnaire and accelerometry. Weight was measured on a scale. Website use was summarized descriptively, and associations were examined using Spearman partial correlations.

Results: Participants had a mean age of 58 (SD 12.5) years; 78.6% (44/56) identified as female, 66.1% (37/56) were non-Hispanic White, and 86% (24/28) were breast cancer survivors. On average, participants viewed 11.2 (SD 7.4) weeks of the 24-week intervention, or a total of 312.9 (SD 255.7) minutes per participant. *Sessions* (n=2736), *Home Page* (n=975), and *Tools* (n=967) features showed the highest activity (5885 total page views). Website use was higher among adults aged 65 years and older than younger participants, showcased by duration of use (mean 14.4, SD 7.4 weeks vs mean 9.2, SD 6.8 weeks; *P*=.009), time spent per week (mean 17.0, SD 9.7 minutes vs mean 10.5, SD 10.6 minutes; *P*=.01), and total number of page views (mean 135.7, SD 90 vs mean 85.3, SD 111.9; *P*=.008); higher website use was also reported among women versus men in terms of duration of use (mean 12.8, SD 7.1 weeks vs mean 5.6, SD 5.9 weeks; *P*=.003), time spent per week (mean 14.6, SD 10.3 minutes vs mean 7.4, SD 10.3 minutes; *P*=.02), and total number of page views (mean 120, SD 110.2 vs mean 50.3, SD 64.4; *P*=.01). Diet quality was positively associated with website use (weeks: *r*=0.50; *P*<.001; time: *r*=0.45; *P*<.001; total page views: *r*=0.46; *P*<.001; and sessions page views: *r*=0.39; *P*=.005). Self-reported MVPA was also positively associated with website use (weeks *r*=0.37;

XSL•FO
RenderX

P=.007; time: *r*=0.36; *P*=.009; total page views: *r*=0.36; *P*=.01; and sessions page views: *r*=0.35; *P*=.01). No significant associations were detected for accelerometry-measured MVPA or weight.

Conclusions: Cancer survivors and their partners engaged with the DUET web-based platform to support diet and physical activity (with use particularly high among older adults and females). However, larger, more diverse dyadic web-based lifestyle interventions are needed to confirm these findings.

Trial Registration: ClinicalTrials.gov NCT04132219; https://clinicaltrials.gov/study/NCT04132219

(J Med Internet Res 2026;28:e86908) doi:10.2196/86908

KEYWORDS

cancer survivors; diet; digital health; dyads; partners; physical activity; website

Introduction

Background

The number of cancer survivors in the United States is increasing, with over 18 million currently living with a history of cancer diagnosis [1,2]. While cure rates are encouraging, cancer survivors represent a population with a high burden of comorbidity, including cardiovascular disease, diabetes, and obesity [3,4]. Furthermore, most cancer survivors do not meet the nutrition and physical activity guidelines recommended by the World Cancer Research Fund/American Institute for Cancer Research (WCRF/AICR) and the American Cancer Society (ACS), often reporting low intake of vegetables and fruits (V&Fs) and insufficient physical activity [5-8]. These poor lifestyle practices can further exacerbate existing comorbidities, compounding their impact on long-term survivorship. Despite these challenges, cancer survivors frequently express a strong interest in improving their lifestyle behaviors [9]. Thus, lifestyle interventions that target diet, physical activity, and weight management have been implemented to promote healthier behaviors, support weight loss, and enhance the quality of life of cancer survivors [10-13].

Over the past decade, the landscape of lifestyle interventions for cancer survivors has evolved, with an increasing shift toward accessible and scalable digital delivery methods [14-16]. Among these, web-based lifestyle interventions have emerged as a promising approach to disseminate diet and physical activity guidance to cancer survivors [14,15]. Websites offer several advantages, such as scalability, personalization, cost-effectiveness, and the flexibility for participants to engage with content at their own pace [14,15,17]. To harness these advantages, the SurvivorSHINE pilot study, a 3-week, single-arm web-based lifestyle intervention, was implemented through an interactive website to promote diet and physical activity guidelines among 41 cancer survivors [18,19]. The website incorporated common strategies such as evidence-based diet and exercise knowledge, tools to facilitate behavior change, and resources for self-monitoring and goal setting [18]. The study reported that cancer survivors perceived the SurvivorSHINE website as a user-friendly platform that provided trustworthy information on diet and exercise, and reported improvements in knowledge related to diet and exercise [18].

Building upon the frameworks and strategies of the SurvivorSHINE intervention, a more refined website was developed that included 24 weekly serialized, interactive sessions, as well as additional tools to support behavior change not only among cancer survivors interested in cancer control, but also among their family members and friends for the purposes of cancer prevention. This newly developed website was then evaluated for feasibility and its impact on various health outcomes in a randomized controlled trial, known as DUET (Daughters, Dudes, Mothers, and Others Together; NCT04132219) that targeted cancer survivors and their chosen partners [20,21].

The DUET trial was 6 months in duration and evaluated the web-based weight loss intervention against a waitlist control among 112 participants (56 dyads, each consisting of a cancer survivor and a chosen partner) [20,21]. Results showed that dyads in the intervention arm achieved significant weight loss, along with improvements in diet quality and physical activity levels compared to the control arm [21]. A mediation analysis further revealed that reductions in perceived dietary barriers significantly contributed to the observed weight loss, highlighting the effectiveness of the intervention's strategies, which were based on social cognitive theory [22-24]. These findings underscore the potential of minimal-touch, web-based lifestyle interventions to facilitate meaningful behavior change and weight loss among cancer survivors and their partners. However, to optimize delivery and behavioral outcomes, it is important to understand how cancer survivors and their partners used the DUET website. While many web-based interventions focus on evaluating behavioral outcomes, relatively few have described website use or examined the association between website use and changes in diet, physical activity, and weight loss. Exploring patterns of use within the DUET website may provide insight into participant interaction with web-based platforms and inform refinements for future web-based lifestyle interventions for cancer survivors and their partners.

Aim and Objective

This secondary analysis aims to describe website use in the DUET trial, and also examine the associations between website use (as measured by number of weeks of website use, average time spent, total page views, and session page views) and changes in diet quality, moderate to vigorous physical activity (MVPA), and weight.

XSL•FO
RenderX

Methods

Study Design

The DUET study was a 2-arm single-blinded randomized controlled trial that enrolled 112 cancer survivors and their chosen partners (56 dyads). Dyads were randomized to either the 6-month web-based weight loss intervention or a waitlist control arm. This secondary analysis focuses on the subset of 56 participants (28 dyads) assigned to the DUET intervention arm, as website usability data were only collected for this group. Full details of the trial methods, primary outcomes, and mediation analysis have been published elsewhere [20-22].

Study Participants

Cancer survivors were identified through multiple recruitment strategies, including cancer registries, self-referrals, and curated lists of individuals who had previously expressed interest in lifestyle interventions. Recruitment efforts focused on cancer survivors who had completed treatment for obesity-related cancers with a 5-year survival rate of ≥70% (eg, localized renal, locoregional ovarian, colorectal, prostate, endometrial, or female breast cancer). Potential cancer survivors were contacted by mail with telephone follow-up, and, if interested, were screened by study staff for the following inclusion criteria: (1) BMI ≥25 kg/m^2; (2) V&F intake <2.5 cups/day; (3) engaging in <150 minutes/week of MVPA; and (4) routine access to the internet via a computer, tablet, or mobile phone. Eligible cancer survivors were asked to identify a partner who lived nearby (within 10 minutes by car) and interacted with them at least biweekly; partners were screened using the same criteria, excluding a cancer diagnosis. However, partners with a history of cancer were eligible.

Study Protocol

Eligible dyads were provided with an overview of the study protocol, and written informed consent was obtained electronically using Adobe Sign [25]. After completion of baseline assessments, dyads were randomized to either the 6-month web-based weight loss intervention or a waitlist control arm. Those assigned to the intervention arm were granted access to the DUET website via an electronic link and instructed to create a secure profile using a username and password unique for each survivor and partner [26]. Dyads were encouraged to log in weekly via text messages and engage with the key features of the DUET website throughout the 6-month intervention period. Follow-up assessments were conducted at 6 months, after which waitlist control dyads were provided access to the DUET intervention. Additional details on the study procedures have been published previously [20].

DUET Intervention Website

The DUET intervention was theoretically grounded in social cognitive theory and incorporated elements from interdependence theory and the theory of communal coping to support both individual and dyadic behavior change by targeting diet and exercise barriers, enhancing social support, and building self-efficacy [23,24,27,28]. DUET was adapted from 2 prior evidence-based lifestyle interventions, Daughters and Mothers Against Breast Cancer and SurvivorSHINE, and delivered via a secure, interactive website that served as the primary web-based platform for intervention delivery [18,26,29]. Details on the intervention development have been published previously [20].

The DUET website included a total of 9 key features: *Home Page, My Profile, Sessions, Healthy Weight, Healthy Eating, Exercise, Tools, News You Can Use, and Team Support.* Upon account creation, participants accessed the *My Profile* feature to enter demographic and lifestyle data, including their current weight, height, and responses to 1-item questions that assessed their dietary intake (eg, consumption of V&Fs, whole grains, red and processed meats, added sugars, and alcohol), and frequency of snacking and physical activity (both endurance and resistance exercise). Cancer survivors also provided details on diagnosis and treatment, which informed tailored feedback on overcoming treatment-related challenges and generating dietary, physical activity, and weight management goals, as described previously [20-22]. The *Home Page* displayed a "Tip of the Day" designed to encourage ongoing engagement with the website, along with visual indicators showing participants' completed weekly sessions and the next upcoming session. It also included direct links to other key features of the website (ie, *Healthy Weight, Healthy Eating, Exercise, Tools, News You Can Use, and Team Support).*

The *Sessions* feature included 24 weekly interactive e-learning modules (~15 minutes each), designed using Articulate Storyline software [30]. Sessions were released sequentially each week over the 24 weeks and introduced via a Monday SMS "push" text message, with additional text messages sent on Wednesdays and Fridays to reinforce continued engagement. These interactive e-learning sessions were designed to provide information on the WCRF/AICR and ACS diet and physical activity guidelines and equip participants with practical and actionable strategies to support behavior change and weight management [5,6]. Dietary recommendations were supported by sessions focused on promoting the consumption of V&Fs, whole grains, and legumes, and limiting red and processed meat, added sugar, alcohol, and snacking, with additional sessions focused on portion control, grocery shopping, and food preparation to support healthy eating habits. Physical activity recommendations were supported by sessions focused on aerobic, resistance, balance, and flexibility exercises, with emphasis on goal setting and problem solving to help participants gradually achieve the goal of 150 minutes of MVPA per week. Details on weekly topics have been published previously [20].

The *Healthy Weight* feature was designed to support self-monitoring of weight by providing an interactive bar graph that tracked participants' current weight and healthy weight. Accompanying educational materials helped contextualize these values by explaining the concept of a healthy weight and its relevance to cancer prevention and survivorship. The website delivered tailored guidance to help participants progress toward a healthy weight by supporting caloric restriction and strategies to promote gradual weight loss of approximately 0.5 kg/week [31].

XSL•FO
RenderX

The *Healthy Eating* feature supported self-monitoring of dietary goals aligned with the WCRF/AICR and ACS guidelines. Participants could record their current intake of key dietary components, that is, V&Fs (≥5 servings/day), whole grains (≥50% of total grain intake), added sugars (≤6 teaspoons/day), avoid snacks, red and processed meats (≤18 ounces/week), and alcohol (≤1 drink/day), through an interactive bar graph that visually displayed their reported intake alongside goal targets. To further support dietary changes, the feature provided educational resources on each dietary component.

The *Exercise* feature supported self-monitoring of physical activity through tailored recommendations based on participants' self-reported activity levels. Physical activity goals were aligned with WCRF/AICR and ACS guidelines, encouraging participants to achieve at least 150 minutes of MVPA and engage in strength training 2-3 times per week.

The *Tools* feature provided a centralized hub for participants to access downloadable materials that supported both dietary and physical activity behaviors aimed at achieving a healthy weight. This feature included 11 distinct resources: (1) BMI calculator, (2) calorie calculator, (3) sample meal plans, (4) food exchange lists, (5) SMART goal templates, (6) serving size guides, (7) fast food guide, (8) grocery lists and shopping tips, (9) calorie-burning guide, (10) exercise logs, and (11) tools for tracking Fitbit data (note: all DUET participants regardless of randomization status, received a Fitbit Inspire and were encouraged to use it during the study period; however, Fitbit data were not integrated into the DUET website) [32]. These resources were designed to offer participants additional support, practical strategies, and accessible tools to enhance self-efficacy throughout the intervention.

The *News You Can Use* feature provided brief, evidence-based summaries on current research related to cancer survivorship, diet, and exercise. The *Team Support* section provided strategies to strengthen dyadic communication, foster mutual goal setting, and enhance social support; it also allowed participants to directly connect with study staff for additional guidance and support.

Measures

Demographics

Cancer type and time since diagnosis were obtained from cancer registries or verified by treating physicians for self-referred participants. Demographic information, including age, sex, race, residence, educational status, employment, and income, was self-reported via electronic surveys completed at baseline. Cohabitation was assessed by comparing mailing addresses; dyads with the same address (0 miles) were classified as cohabitating, and those with different addresses (>0 miles) as noncohabitating.

Diet Quality, MVPA, and Weight

Dietary intake was assessed at baseline and 6 months via two 24-hour dietary recalls (1 weekday and 1 weekend day) conducted by a registered dietitian over the telephone. The Automated Self-Administered 24-hour Dietary Assessment Tool was used to capture dietary intake data. Diet quality was evaluated using the Healthy Eating Index 2015 [33,34].

MVPA was assessed both objectively and subjectively at baseline and 6 months. Participants wore ActiGraph accelerometers for 7 consecutive days. Data were then processed using ActiLife software following standardized procedures to calculate average weekly minutes of MVPA [35,36]. Self-reported MVPA was captured using the Godin Leisure-Time Exercise Questionnaire, a validated tool frequently used in cancer survivorship research [37].

Weight was measured remotely at baseline and 6 months. Each survivor-partner dyad completed the virtual assessment together via Zoom with study staff [38]. Participants used a digital bathroom scale to report their weight; a scale was provided with the assessment materials for those who did not own one. During the virtual assessment, trained staff instructed participants to wear light clothing and remove their shoes, and partners assisted in holding the camera and angling it so that study staff could verify the weight displayed on the digital scale. Additional details on the remote assessment protocol and its validity have been described previously [39].

Website Use Metrics

Website use was assessed using tracking data logged by the DUET website platform. Each participant was assigned a unique website username and password, which was linked to their website ID and further connected to their study ID, allowing for the tracking of individual-level website activity over the 24-week intervention period. Time-stamped data recorded the days, times, and pages accessed by participants (eg, *Home Page*, *My Profile*, *Sessions*, *Healthy Weight*, *Healthy Eating*, *Exercise*, *Tools*, *News You Can Use*, *and Team Support*). From these data, key website use metrics were derived, including (1) the total number of weeks participants accessed the website, defined as the number of distinct weeks during the 24-week intervention period in which any website activity was recorded; (2) the average time spent on the website per day of use, calculated using time-stamped activity logs that captured first and last activity on a given day; (3) the total number of page views, defined as the cumulative number of times participants navigated to individual pages within each of the website's key features; and (4) the total number of session page views, defined as the number of times participants accessed the e-learning *Sessions* feature.

Statistical Analysis

Website usability was analyzed using descriptive statistics, including means, SDs, and ranges for continuous variables (eg, number of weeks accessed, time spent on the website, and page views). Website usability metrics were described for the total sample and stratified by dyad member (cancer survivors or partners), clinical (ie, cancer type and time since diagnosis) and sociodemographic factors (ie, age, sex, race, residence, educational status, employment, and income) and cohabitation status (cohabitate or did not cohabitate), with all stratification variables dichotomized for analysis. To examine differences in website use between dyad members, clinical and sociodemographic factors, and cohabitation status, assumptions

XSL•FO
RenderX

of normality were assessed using Shapiro-Wilk tests and visual inspection of histograms and Q-Q plots. Given that usability metrics were not normally distributed, Wilcoxon rank-sum tests were conducted to compare median differences in website use between dyad members, clinical and sociodemographic factors, and cohabitation status. Assumptions of normality and linearity were assessed for the independent (website use metrics) and dependent variables (diet quality, MVPA, and weight). Given evidence of nonnormal distributions, likely influenced by the modest sample size and potential nonlinear relationships, nonparametric methods, such as bivariate Spearman partial rank correlation analyses, were conducted to examine associations between website use and diet quality, MVPA, and weight. Correlation coefficients were generated using 6-month outcome values as dependent variables, adjusting for baseline values of the respective outcome, as well as age, sex, and race to account for potential confounding and assess change over time. Given this was a secondary analysis and not prospectively powered for these aims, we conducted a post hoc power calculation to aid interpretation of the correlation analyses. With a total sample size of 56 participants, the study had ≥80% power to detect correlations of approximately $r \geq 0.40$. Missing data were handled using complete-case analysis. One participant had missing diet quality data at 6-month follow-up, and 13 participants had missing accelerometer-measured MVPA data at baseline (n=8) and 6-month follow-up (n=5); these cases were excluded from their respective models. No adjustments for multiple comparisons were made, given the exploratory nature of this secondary analysis. However, to address the increased type I error risk from multiple correlations, 95% CIs were reported with *P* values to assist in interpreting the precision of the coefficients. All analyses were conducted using SAS (version 9.4, SAS Institute Inc), and statistical significance was set at $P<.05$ [40].

Ethical Considerations

The DUET study was approved by the Institutional Review Board at the University of Alabama at Birmingham (IRB# 300003882) and was registered with ClinicalTrials.gov (NCT04132219). All participants provided written informed consent, and study procedures were conducted in accordance with the ethical standards of the Declaration of Helsinki to maintain participant confidentiality. No compensation was provided to study participants.

Results

Sample Characteristics

The average age of the sample was 58 (SD 12.5) years, with survivors averaging 60 (SD 11.2) years and partners 56 (SD 13.7) years. The majority of cancer survivors (24/28, 85.7%) had a breast cancer diagnosis, with an average time since diagnosis of approximately 71 (SD 80.4) months (6 years). Most participants identified as female (44/56, 78.6%), non-Hispanic White (37/56, 66.1%), and residents of urban areas (53/56, 94.6%). Employment status was evenly split between employed (30/56, 53.6%) and retired (26/56, 46.4%), and most (45/56, 80.4%) reported an annual household income above US $50,000 per year (Table 1).

XSL•FO
RenderX

Table 1. Characteristics of 56 cancer survivors and their chosen partners randomized to the intervention arm, stratified by dyad status.

Characteristics	Total sample	Survivor	Partner	*P* value[a]
Age (years), mean (SD; range)	58.1 (12.5; 23-78)	60.0 (11.2; 32-78)	56.3 (13.7; 23-74)	.28
Months from diagnosis, mean (SD; range)	71 (80.4; 10-303)	71 (80.4; 10-303)	—[b]	—[b]
BMI (kg/m^2), mean (SD; range)	31.4 (4.9; 25-45)	32.0 (5.4; 25-44)	30.9 (4.4; 25-45)	.41
Diet quality (HEI[c]), mean (SD; range)	53.1 (12.8; 29-87)	53.9 (13.7; 30-87)	52.2 (12; 29-81)	.62
MVPA[d] (min/week), mean (SD; range)	43.8 (60.5; 0-280)	48.5 (67.8; 0-280)	39.1 (52.9; 0-225)	.57
Cancer type[e], n (%)				<.001
Breast	25 (44.6)	24 (85.7)	1 (3.6)	
Other[f]	7 (12.5)	4 (14.3)	3 (10.7)	
Sex, n (%)				.05
Male	12 (21.4)	3 (10.7)	9 (32.1)	
Female	44 (78.6)	25 (89.3)	19 (67.9)	
Race, n (%)				.78
Non-Hispanic White	37 (66.1)	19 (67.9)	18 (64.3)	
Non-Hispanic Black or other[g]	19 (33.9)	9 (32.1)	10 (35.7)	
Residence, n (%)				.55
Urban	53 (94.6)	27 (96.4)	26 (92.9)	
Rural	3 (5.4)	1 (3.6)	2 (7.1)	
Educational status, n (%)				.13
High school or less	8 (14.3)	2 (7.1)	6 (21.4)	
Some college or more	48 (85.7)	26 (92.9)	22 (78.6)	
Employment, n (%)				>.99
Employed	30 (53.6)	15 (53.6)	15 (53.6)	
Retired or other[h]	26 (46.4)	13 (46.4)	13 (46.4)	
Income, n (%)				.09
Less than US $50,000/year	11 (19.6)	8 (28.6)	3 (10.7)	
More than US $50,000/year	45 (80.4)	20 (71.4)	25 (89.3)	

[a]*P* values were calculated using independent samples *t* tests for continuous variables (based on equal or unequal variances as appropriate) and chi-square tests for categorical variables, comparing survivors vs partners, significance set at $P<.05$.

[b]Data on months since diagnosis is unavailable for partners.

[c]HEI: Healthy Eating Index 2015.

[d]MVPA: moderate to vigorous physical activity.

[e]Total percentage does not sum to 100% due to the inclusion of nonsurvivors who were not diagnosed with cancer.

[f]Other cancer diagnoses include prostate, colorectal, gynecologic, and renal.

[g]Other race includes Hispanic ethnicity, accounting for 1%.

[h]Other employment includes student and disabled.

DUET Website Use

On average, participants accessed the DUET website for 11.2 (SD 7.4; range 0-24) weeks and spent a total of 312.9 (SD 255.7; range 0-1119) minutes on the platform, or about 13 (SD 10.7; range 0-46.6) minutes per week over the 24-week intervention period. A total of 5885 page views were recorded, with an average of 105.1 (SD 105.7; range 0-545) page views per participant. Cancer survivors used the website more frequently than their partners, accessing it for a longer duration (mean 13.0, SD 7.2 weeks vs mean 9.5, SD 7.4 weeks; $P=.08$), spending more time per week (mean 15.6, SD 11.4 minutes vs mean 10.5, SD 9.4 minutes; $P=.07$), and recording a higher average number of page views (mean 124.2, SD 114.2 vs mean 86.0, SD 94.5; $P=.07$); however, these differences were not statistically significant (Table 2).

XSL•FO
RenderX

Table 2. Daughters, Dudes, Mothers, and Others Together website usability over the 24-week intervention period for the total sample (n=56) and stratified by dyad status.

Engagement metrics	Total sample, mean (SD; range)	Survivor, mean (SD; range)	Partner, mean (SD; range)	*P* value[a]
Weeks participants accessed website	11.2 (7.4; 0-24)	13.0 (7.2; 0-24)	9.5 (7.4; 0-23)	.08
Total time spent on website (min)	312.9 (255.7; 0-1119)	373.8 (273.3; 0-1119)	252.0 (225.5; 0-670)	.07
Time spent on website per week (min)	13.0 (10.7; 0-46.6)	15.6 (11.4; 0-46.6)	10.5 (9.4; 0-27.9)	.07
Total page views per user (n=5885)	105.1 (105.7; 0-545)	124.2 (114.2; 0-545)	86.0 (94.5; 0-352)	.07

[a]*P* values represent comparisons between survivors and partners using Wilcoxon rank-sum tests due to the nonnormal distribution of website usability metrics, significance set at *P*<.05.

Website use differed significantly by age, time since diagnosis, and sex. Older adults aged ≥65 years demonstrated higher website use compared to younger participants (<65 years), accessing it for a longer duration (mean 14.4, SD 7.4 weeks vs 9.2, SD 6.8 weeks; *P*=.009), spending more time per week (mean 17.0, SD 9.7 minutes vs mean 10.5, SD 10.6 minutes; *P*=.01), and recording a higher average number of total page views (mean 135.7, SD 90 vs mean 85.3, SD 111.9; *P*=.008). Survivors who were 5 or more years post diagnosis also accessed the website for longer duration than those more recently diagnosed (<5 years; mean 18.4, SD 6.3 weeks vs mean 10.8, SD 6.4 weeks; *P*=.009). Additionally, females used the website significantly more than males, accessing it for a longer duration (mean 12.8, SD 7.1 weeks vs mean 5.6, SD 5.9 weeks; *P*=.003), spending more time per week (mean 14.6, SD 10.3 minutes vs mean 7.4, SD 10.3 minutes; *P*=.02), and recording a higher average number of total page views (mean 120, SD 110.2 vs mean 50.3, SD 64.4; *P*=.01; Table 3).

XSL•FO
RenderX

Table 3. Website use stratified by clinical characteristics, sociodemographic factors, and cohabitation status.

Measures	Weeks			Average time			Total page views			Session page views		
	Mean (SD)	Test statistic[a]	*P* value	Mean (SD)	Test statistic	*P* value	Mean (SD)	Test statistic	*P* value	Mean (SD)	Test statistic	*P* value
Age		6.83	.009		6.55	.01		7.03	.008		6.34	.01
<65 years	9.17 (6.8)			10.5 (10.6)			85.3 (111.9)			38.4 (35.2)		
≥65 years	14.4 (7.4)			17 (9.7)			135.7 (90)			65 (38.5)		
Time since diagnosis		6.88	.009		1.62	.20		2.57	.11		0.622	.43
<5 years	10.8 (6.4)			14 (11.9)			108.4 (117.8)			52.5 (35.1)		
≥5 years	18.4 (6.3)			19.4 (9.6)			163.6 (100.9)			65.3 (32.3)		
Cancer type		0.64	.42		0.03	.87		0.53	.47		1.15	.28
Breast	13.6 (7.2)			15.3 (11.3)			122.9 (119.5)			54.8 (37.1)		
Other	11.6 (4.9)			15.9 (9.7)			129.6 (81)			27 (36.7)		
Sex		8.82	.003		5.70	.02		6.69	.01		6.39	.01
Female	12.8 (7.1)			14.6 (10.3)			120 (110.2)			54.8 (37.1)		
Male	5.6 (5.9)			7.4 (10.3)			50.3 (64.4)			27 (36.7)		
Race		0.02	.89		0.08	.78		0.11	.75		0.01	.91
Non-Hispanic White	11.4 (6.8)			12.9 (9.6)			100.2 (86.2)			46.4 (33.2)		
Non-Hispanic Black	10.9 (8.6)			13.3 (12.5)			113.9 (136)			53.4 (47.1)		
Residence		0.01	.91		0.41	.52		0.28	.60		0.90	.34
Urban	11.2 (7.3)			12.7 (10.3)			104 (106.8)			47.2 (36.3)		
Rural	12.3 (11)			18.4 (17.1)			123.7 (98.1)			77.7 (70.3)		
Education		0.95	.33		0.01	.92		0.08	.78		0.00	.98
High school or less	8.9 (6.3)			12 (10.5)			116 (131.7)			45.9 (37.8)		
Some college or above	11.6 (7.6)			13.2 (10.8)			103.3 (102.2)			49.4 (38.9)		
Employment		1.29	.26		0.14	.71		0.35	.55		0.14	.71
Employed	10.4 (6.3)			12.9 (11.7)			101.8 (117.1)			47.6 (39.7)		
Retired or other	12.2 (8.5)			13.3 (9.5)			108.9 (92.8)			50.3 (37.7)		
Income		0.08	.78		0.45	.50		0.18	.67		0.81	.37
Less than US $50k/year	10.6 (5.9)			15.9 (13.9)			132 (152.8)			59.7 (44.9)		
More than US $50k/year	11.4 (7.8)			12.3 (9.8)			98.5 (91.7)			46.2 (36.8)		

XSL•FO
RenderX

Measures	Weeks			Average time			Total page views			Session page views		
	Mean (SD)	Test statistic[a]	*P* value	Mean (SD)	Test statistic	*P* value	Mean (SD)	Test statistic	*P* value	Mean (SD)	Test statistic	*P* value
Cohabitation		2.23	.14		1.54	.21		0.40	.53		2.01	.16
Cohabitate	9.5 (8.1)			11.1 (10.3)			93.1 (91)			42.2 (39.8)		
Do not co-habitate	12.5 (6.7)			14.5 (10.8)			114.1 (116)			53.8 (37.2)		

[a]Differences between groups were assessed using the Mann-Whitney *U* test, implemented via the Wilcoxon rank-sum approach. A chi-square approximation was used to derive the test statistics, with statistical significance defined as $P<.05$.

Website use was generally highest during the initial 9 weeks of the intervention, although some fluctuations were observed, particularly in weeks 4, 6, and 8. Website use gradually declined over time, with a slight increase observed in the final week of the intervention. Cancer survivors spent more time on the website each week compared to their partners, with notable peaks observed among cancer survivors at weeks 1, 2, 6, 8, 13, 16, 19, and 24 (Figure 1).

Figure 1. Average weekly time spent on the Daughters, Dudes, Mothers, and Others Together (DUET) website during the 24-week intervention period by the total sample, highlighting high and low engagement weeks stratified by dyad status. All weekly DUET sessions were designed to take 15 minutes or less, except for the initial onboarding session.

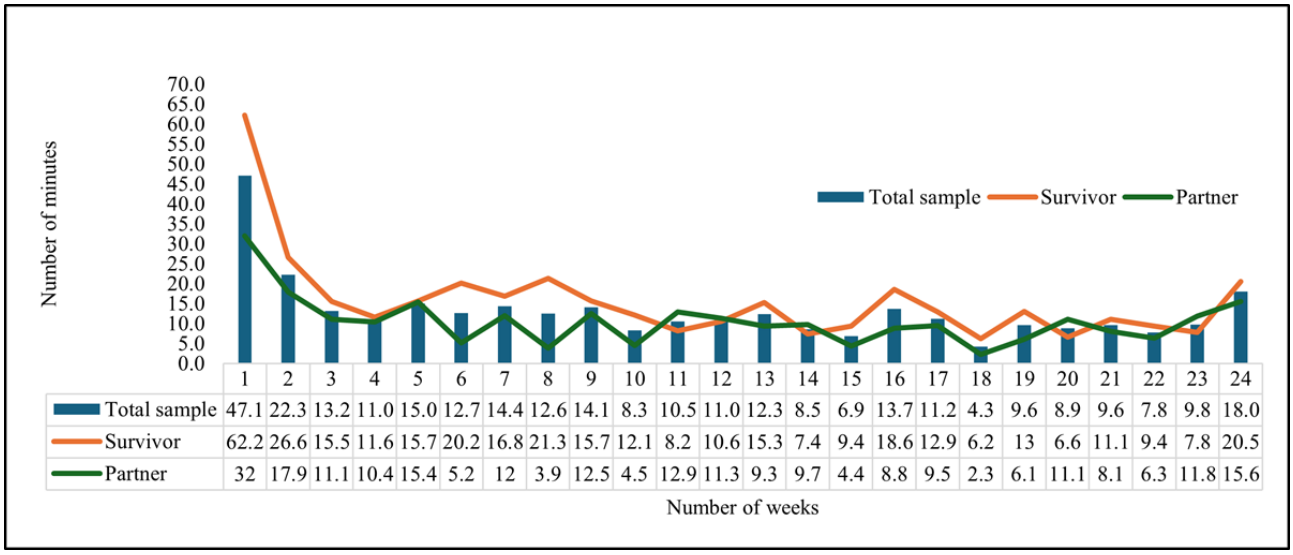

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Total sample	47.1	22.3	13.2	11.0	15.0	12.7	14.4	12.6	14.1	8.3	10.5	11.0	12.3	8.5	6.9	13.7	11.2	4.3	9.6	8.9	9.6	7.8	9.8	18.0
Survivor	62.2	26.6	15.5	11.6	15.7	20.2	16.8	21.3	15.7	12.1	8.2	10.6	15.3	7.4	9.4	18.6	12.9	6.2	13	6.6	11.1	9.4	7.8	20.5
Partner	32	17.9	11.1	10.4	15.4	5.2	12	3.9	12.5	4.5	12.9	11.3	9.3	9.7	4.4	8.8	9.5	2.3	6.1	11.1	8.1	6.3	11.8	15.6

A total of 5885 page views were recorded throughout the intervention period across the 9 key website features. The highest number of page views was observed for the weekly interactive e-learning *Sessions* feature (n=2736), followed by the *Home Page* (n=975) and *Tools* feature (n=967). The remaining features, *My Profile* (n=400), *Healthy Weight* (n=248), *Healthy Eating* (n=234), *Exercise* (n=162), *News You Can Use* (n=104), and *Team Support* (n=59), had comparatively fewer page views (Figure 2).

XSL•FO
RenderX

Figure 2. Page view distribution across key features of the Daughters, Dudes, Mothers, and Others Together website stratified by dyad status (n=5885 total page views).

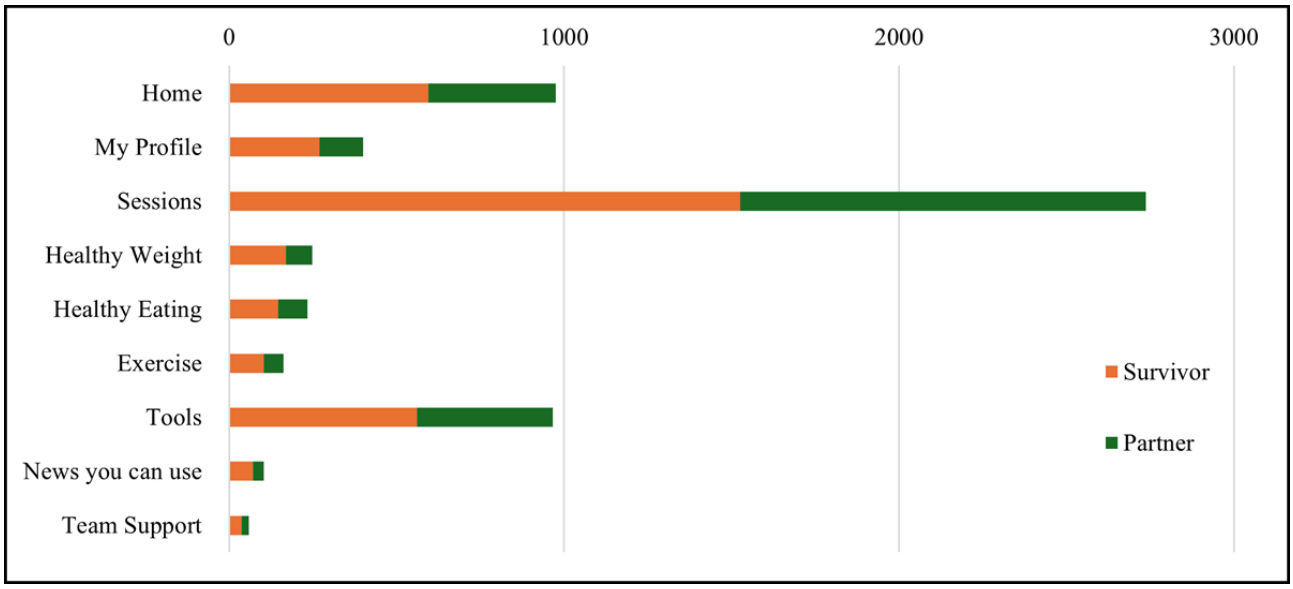

Figure 3 further illustrates the distribution of 2736 page views across the weekly e-learning *Sessions* feature over the 24-week intervention period. Page views for the weekly sessions were highest during the early weeks of the intervention, particularly sessions 1 through 8, and gradually declined over time. The onboarding session received the greatest number of views, and the three most frequently viewed sessions were (1) Get on Track for Success (session 2), (2) Moving Towards Better Health (session 4), and (3) Been Resisting "Resistance" Exercises (session 8). Conversely, the three least viewed sessions were (1) Want to Join the Party Without Blowing Your Diet? (session 20), (2) Why Am I Hungry All the Time? (session 22), (3) Are Supplements Really Good for You? (session 23).

Figure 3. Page view distribution of weekly released diet and exercise sessions accessed by cancer survivors and partners (n=2736 total session page views). Top 3 most viewed sessions for the total sample: Get on Track for Success (session 2), Moving Towards Better Health (session 4), and Been Resisting "Resistance" Exercises (session 8). Bottom 3 least viewed sessions for the total sample: Want to Join the Party Without Blowing Your Diet (session 20), Why Am I Hungry All the Time (session 22), Are Supplements Really Good for You (session 23).

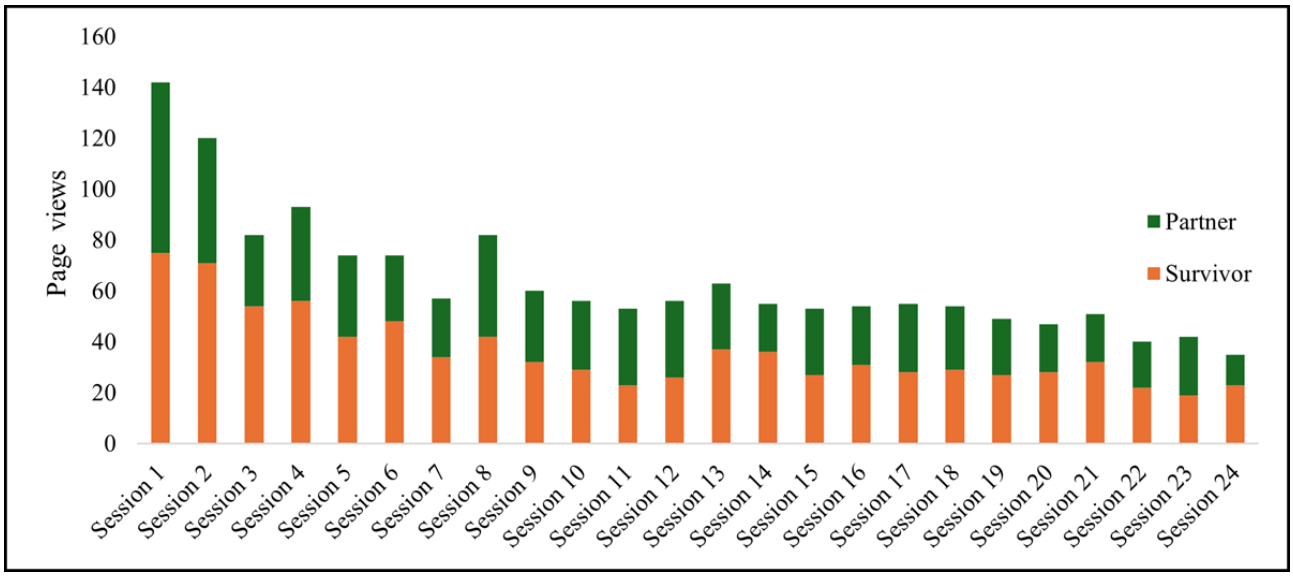

The *Tools* feature accounted for the third highest number of page views (n=967), following the *Home page* and *Sessions*, with participants most frequently viewing tools such as Sample Meal Plans (n=121 page views), Tracking with Fitbit (n=119), Exercise Logs (n=65), BMI Calculator (n=65), Fast Food Menu Maven (n=64), Calorie Calculator (n=62) and Calorie Burning Guide (n=60; Multimedia Appendix 1). The *Healthy Weight, Healthy Eating, and Exercise* features were also accessed throughout the intervention period. Within these features, the most commonly viewed content included "Common Questions About Weight Management" (n=20 page views) under the Healthy Weight category, "Increasing V&F Intake" (n=10 page views) under Healthy Eating, and "Leg Strengthening Exercises" (n=19 page views) under the Exercise category (Multimedia Appendix 2)

XSL•FO
RenderX

Website Use and Behavioral Associations

Diet quality was positively associated with website use, weeks (r=0.50; P<.001), time (r=0.45; P<.001), total page views (r=0.46; P<.001), and sessions page views (r=0.39; P=.005). Self-reported MVPA was also positively associated with website use, weeks (r=0.37; P=.007), time (r=0.36; P=.009), total page views (r=0.36; P=.01), and sessions page views (r=0.35; P=.01). However, no statistically significant associations were detected for accelerometry-measured MVPA or weight (P>.05; Table 4).

Table 4. Bivariate associations between website usability and Healthy Eating Index 2015 diet quality, moderate to vigorous physical activity (MVPA), and weight at 6-months. Bivariate associations were performed using Spearman partial correlations rank analysis for nonnormally distributed data. Adjusted for age, sex, race, and baseline outcome variables.

Measures	Weeks		Average time		Total page views		Session page views	
	r (95% CI)	*P* value	*r* (95% CI)	*P* value	*r* (95% CI)	*P* value	*r* (95% CI)	*P* value
Diet quality	0.50 (0.25 to 0.68)	<.001	0.45 (0.20 to 0.65)	<.001	0.46 (0.21 to 0.65)	<.001	0.39 (0.12 to 0.60)	.005
Self-reported MVPA	0.37 (0.10 to 0.58)	.007	0.36 (0.09 to 0.57)	.009	0.36 (0.09 to 0.57)	.01	0.35 (0.08 to 0.57)	.01
Accelerometer MVPA	0.15 (–0.16 to 0.44)	.33	0.06 (–0.25 to 0.37)	.68	0.03 (–0.28 to 0.34)	.82	0.04 (–0.27 to 0.35)	.79
Weight	–0.17 (–0.42 to 0.11)	.24	–0.06 (–0.32 to 0.22)	.67	–0.13 (–0.39 to 0.15)	.35	–0.11 (–0.38 to 0.16)	.41

Discussion

Primary Findings

There are very few web-based lifestyle interventions for cancer survivors and their support partners that use evidence-based theoretical constructs to promote healthful diet and physical activity behaviors [14]. As a result, no studies to date have described patterns of website use to inform the design and delivery of future dyadic, web-based lifestyle interventions. This study is among the first to address this gap by providing a detailed analysis of website use patterns among cancer survivors and their chosen partners randomized to the DUET intervention. Our findings suggest that cancer survivors and their partners used the DUET website to learn about healthy lifestyle guidelines, as reflected by website use metrics: on average, participants logged into roughly half of the 24-week content, interacting for a total of 313 minutes with the program, for which the highest page view activity was observed for *Sessions* (n=2736). Older adults and females engaged with the website to a significantly greater degree compared to younger participants and males, as reflected by weekly website activity (frequency, user time, and page views). Moreover, higher levels of website use were significantly associated with improvements in diet quality and self-reported MVPA. However, no significant associations were observed between website use and accelerometer-measured MVPA and weight.

Comparison With Previous Literature

Due to variability in how website usability metrics are reported, direct comparisons of website use across studies are challenging. However, findings from 3 pilot web-based lifestyle interventions implemented among cancer survivors generally align with our data that cancer survivors engage with a lifestyle website approximately once per week, particularly during the initial phase of the study. For example, the SurvivorSHINE study reported an average of 1.5 log ins per week over a 2-week period [41]. Similarly, in the A Lifestyle Intervention Via Email study, breast cancer survivors visited the website for an average of 9.6 out of 12 weeks in the physical activity arm and 10.7 out of 12 weeks in the diet arm [42]. A study by Blarigan and colleagues [43] reported that colorectal cancer survivors in the intervention arm accessed the website on a median of 13 out of 84 days. Despite the 24-week DUET intervention being roughly twice as long as web-based interventions used in previous studies, we observed comparable weekly website use during the initial 12 weeks of the program, suggesting sustained and moderate website use among dyads during the initial phases of the intervention. Our findings also showed that dyads spent a total of 313 minutes on the website, which differs from the SurvivorSHINE study (94 minutes) but aligns closely with the Breast Cancer eHealth Self-Management study (337 minutes) [41,44]. The difference in website use observed in our study compared to SurvivorSHINE is likely due to the longer intervention duration and the inclusion of weekly serialized e-learning sessions and videos, which were not part of the SurvivorSHINE's 2-week intervention.

This study also examined page view activity to assess participant usability with various DUET website features. The most frequently accessed feature was the interactive e-learning "*Sessions*," followed by the "*Home Page*" feature. Prior systematic reviews have emphasized the importance of interactive educational modules to promote real-time engagement with behavior change content [45-47]. The DUET "*Sessions*" were developed with this framework in mind, incorporating brief, skill-building activities designed to help both survivors and their support partners apply evidence-based knowledge to everyday challenges. These sessions encouraged teamwork and joint problem-solving to support healthier behaviors, a strategy supported by the theory of communal coping and prior studies [27,48,49]. Participants also received 3 weekly text messages reminding them to visit the website and engage with the weekly content, which likely contributed to the high usability observed with the "*Sessions*" feature. The brief, focused, and interactive nature of the sessions may have further

XSL•FO
RenderX

enhanced their appeal. Similar to findings from the SurvivorSHINE study, the "*Home Page*" was also commonly viewed, likely because it automatically loaded each time participants accessed the website, allowing them to engage with core behavior change constructs such as self-monitoring, goal setting, and motivation [41].

Consistent with prior research conducted in samples without a history of cancer, our findings indicate that website use was significantly higher among older adults (aged ≥65 years) and female participants. For example, Graham et al [50] reported that older adults using the commercially available Lark Health digital platform logged more meals (174 vs 89) and used more self-monitoring devices (39 vs 28) compared to younger adults. Relatedly, a scoping review found that among adults aged 50 years and older, structured, tailored digital programs were well-accepted [51]. Similarly, a growing body of literature has consistently shown that women are more likely to seek health-related information, engage with online health platforms, and participate in digital lifestyle interventions compared to men [14,52,53]. In our study, this trend may be amplified by the DUET program's structured, self-paced design and tailored content, which likely resonated with older and female participants who may prefer individualized guidance and flexibility with web-based programs [52,54].

Bivariate analyses from this study showed that website use was significantly associated with improvements in diet quality and self-reported MVPA at 6 months. The evidence on whether website use improves diet quality remains unclear, largely due to heterogeneity in how diet quality is measured across studies. For instance, our findings differ from those of a systematic review and meta-analysis of 29 studies in adults with chronic conditions, which found no evidence for website use to improve overall diet quality [55]. However, most studies in this review were short in duration (<3 months) and relied on self-reported measures; in contrast, our study used interviewer-administered 24-hour dietary recalls, the gold standard for diet quality assessment, which may be more sensitive to detecting change. Our findings are more consistent with systematic reviews and several studies that have reported moderate to strong associations between website use and increased physical activity using self-reported measures [41,55-57]. One possible reason the DUET website led to improvements in diet quality and self-reported MVPA is its inclusion of a range of relevant content for cancer survivors and their partners, such as guidance on V&F intake, reducing processed foods, portion control, setting SMART goals, and providing tools for tracking and self-monitoring diet and physical activity behaviors. However, it is important to note that diet quality and self-reported MVPA models were modestly significant, and may not withstand correction for multiple testing and should be interpreted as hypothesis-generating.

Despite these positive associations, our analysis did not find a significant association between website use and accelerometer-measured MVPA and weight. Our findings for not detecting significant associations with subjectively measured physical activity aligns with findings from previous studies [55,56], but, they contrast with findings from the Commonwealth Scientific and Industrial Research Organisation (CSIRO) Total Wellbeing Diet Online program and the Weight Loss Maintenance (WLM) trial, which found that website use was associated with greater weight loss and reduced weight regain, respectively [58,59]. However, the CSIRO program was commercially delivered and relied on participant self-reported weights within the platform, while the WLM trial focused on weight maintenance rather than weight loss [58]. Several methodological, individual, and behavioral level considerations may help explain the lack of association between website use and accelerometer-measured MVPA and weight [60]. Our power calculation confirmed that with 56 participants, the study was powered only to detect correlations of approximately $r \geq 0.40$. Given that the association between website use and accelerometer-measured MVPA and weight observed in our data was substantially smaller ($r < 0.20$), the study was likely underpowered to detect this relationship. Additionally, differences in how outcomes were measured may further clarify the findings. Website use, diet quality, and self-reported MVPA rely on participant interaction and reporting, which may naturally relate to one another and explain shared method variance. In contrast, accelerometer-measured MVPA and weight are objective measures that do not rely on participant reporting, and therefore their associations with website use may be smaller and more difficult to detect. Beyond methodological explanations, individual and behavioral factors may also play a role. For instance, website use may support behavior change but not be sufficient on its own to produce weight loss, as some participants may have relied less on the website once new habits were established. Additionally, weight loss may be influenced more by theoretical and behavioral mechanisms (ie, reduced perceived barriers, self-monitoring, calorie restriction, increased accountability from participating in the study, or lifestyle changes occurring outside the platform). These possibilities suggest that website use alone may not fully reflect the processes that contributed to weight loss in the parent trial [21].

Strengths and Limitations

This study had several notable strengths. It examined website use among cancer survivors and their partners, an area with limited prior research, as few lifestyle interventions have studied website use among dyads. Our analysis also provided detailed page view analytics across the entire DUET website, offering insight into which website features were most frequently used. Furthermore, the study used validated measures to assess both diet quality and MVPA, enhancing the accuracy of the findings. However, like all studies, there were limitations. Most importantly, website data were only available for the 56 participants in the intervention arm, resulting in a relatively small analytic sample. A post hoc power calculation indicated that with this sample size, the study was powered (≥80%) only to detect correlations of approximately $r \geq 0.40$, representing a moderate to large effect size. As a result, smaller associations, particularly for accelerometer-measured MVPA and weight change, may not have been detectable, and the null findings for these outcomes should be interpreted with caution. Future web-based trials should consider recruiting larger samples to ensure adequate power to detect smaller associations between website use and clinically relevant outcomes such as weight change. Additionally, accelerometer-measured MVPA had 20%

XSL•FO
RenderX

data missingness. As a result, complete-case analysis may introduce bias if participants with missing MVPA data differed meaningfully from those with complete data. Importantly, because multiple correlations were examined, there is an increased risk of type I error, and some associations may reflect spurious findings; therefore, the results should be interpreted within the exploratory, hypothesis-generating context of this secondary analysis. Similarly, it was not possible to determine whether participants viewed the full content of the weekly sessions they accessed, as data only represented whether participants accessed the sessions. Thus, it is likely that participants who briefly clicked into a session could therefore be coded similarly to those who reviewed the entire content, and time-stamp data may overestimate engagement if browser windows were left open. To partially address this limitation, we examined multiple measures of website use rather than relying on a single metric. Nevertheless, this measurement constraint may have led to imprecise estimates of website use and may help explain the modest strength of associations observed. Additionally, given the dyadic nature of the study, there may have been instances where dyads shared a single account and viewed website content together (as was documented in at least 1 case), potentially accounting for the difference between survivor and partner usage and underestimating individual-level usability. Another limitation is the demographic homogeneity of the sample, which primarily included female, non-Hispanic White breast cancer survivors residing in urban areas with higher socioeconomic status. As a result, these findings may not generalize to male survivors, racial and ethnic minority groups, individuals with lower income levels, or those living in rural or medically underserved settings.

Conclusions

Minimal-touch lifestyle interventions delivered through web-based platforms are being implemented to promote diet and physical activity behaviors for weight management among cancer survivors. However, patterns of website use are often understudied. Findings from our study suggest that cancer survivors and their partners (especially older adults and females) actively used the DUET website, particularly the interactive e-learning sessions, and higher levels of website use were significantly associated with improvements in diet quality and self-reported MVPA. However, no statistically significant improvements were detected for accelerometry-measured MVPA or weight. These results highlight that web-based platforms may serve as a promising, scalable approach for delivering diet and physical activity guidelines and promoting healthy behaviors for long-term older cancer survivors and their partners. However, larger, more diverse dyadic web-based lifestyle interventions, including male survivors, racial and ethnic minority populations, individuals with lower income, and survivors in rural or underserved areas, are needed to confirm and expand upon these findings. Importantly, future web-based lifestyle interventions should also use more detailed engagement tracking (ie, content completion indicators, differentiation between brief access and full interaction with embedded activities, and minimum and maximum time thresholds) to distinguish brief access from full content consumption to strengthen the validity of engagement measures.

Acknowledgments

First and foremost, we sincerely thank the cancer survivors and their partners who participated in the DUET intervention. We also gratefully acknowledge our funding sources: the American Institute for Cancer Research (585363); the American Cancer Society (CRP-19–175-06-COUN); the National Cancer Institute (P01 CA229997); O'Neal Comprehensive Cancer Center (P30 CA013148); the Cancer Prevention and Control Training Program (T32 CA047888); the Training to Reduce Burden across the Cancer Control Continuum multidisciplinary program (T32 CA251064-5); and the University of Arizona's Comprehensive Cancer Center (P30 CA023074). Lastly, we extend our appreciation to the clinical staff and students (Drs Teri Hoenemeyer and Amber Kinsey, J Ryan Buckman, Grey Freylersythe, Lauren King, Doctorre McDade, Christopher Reid, Abigail Sims, and Fariha Tariq) for their dedication and efforts in the design, implementation, and successful completion of the study. No generative artificial intelligence was used in any part of the manuscript creation.

Funding

This work was funded by the American Institute for Cancer Research (grant 585363); the American Cancer Society (grant CRP-19–175-06-COUN); the National Cancer Institute (grant P01 CA229997); O'Neal Comprehensive Cancer Center (grant P30 CA013148); the Cancer Prevention and Control Training Program (grant T32 CA047888); the Training to Reduce Burden across the Cancer Control Continuum multidisciplinary program (grant T32 CA251064-5); and the University of Arizona's Comprehensive Cancer Center (grant P30 CA023074). The funder was not involved in the conduct of the research, the decision to publish, or the approval of the publication.

Data Availability

The data reported in this study are available on request from the corresponding author.

Authors' Contributions

Conceptualization: HK (lead), WDW
Data curation: HK (lead), WDW

XSL•FO
RenderX

Formal analysis: HK (lead), WDW
Funding acquisition: WDW (lead)
Investigation: HK (lead), WDW (equal)
Methodology: WDW (lead), DF, DP, TEC, LQR, HK
Project administration: WDW (lead), DF, DP, HK
Resources: WDW (lead)
Supervision: WDW (lead)
Validation: WDW (lead)
Visualization: HK (lead), WDW
Writing – original draft: HK (lead)
Writing – review & editing: HK (lead), WDW (lead), DF, DP, LQR, TEC

Conflicts of Interest

None declared.

Multimedia Appendix 1
Distribution of DUET website page view activity for specific tools accessed by cancer survivors and their supportive partners, based on 967 total page views.
[PNG File , 55 KB - jmir_v28i1e86908_app1.png]

Multimedia Appendix 2
Distribution of DUET website page view activity for healthy eating, healthy weight, and exercise features accessed by cancer survivors and their chosen partners, based on 644 total page views.
[PNG File , 76 KB - jmir_v28i1e86908_app2.png]

References

1. Bluethmann SM, Mariotto AB, Rowland JH. Anticipating the "Silver Tsunami": prevalence trajectories and comorbidity burden among older cancer survivors in the United States. Cancer Epidemiol Biomarkers Prev 2016;25(7):1029-1036 [FREE Full text] [doi: 10.1158/1055-9965.EPI-16-0133] [Medline: 27371756]
2. Wagle NS, Nogueira L, Devasia TP, Mariotto AB, Yabroff KR, Islami F, et al. Cancer treatment and survivorship statistics, 2025. CA Cancer J Clin 2025;75(4):308-340 [FREE Full text] [doi: 10.3322/caac.70011] [Medline: 40445120]
3. Shahrokni A, Wu AJ, Carter J, Lichtman SM. Long-term toxicity of cancer treatment in older patients. Clin Geriatr Med 2016;32(1):63-80 [FREE Full text] [doi: 10.1016/j.cger.2015.08.005] [Medline: 26614861]
4. Heo J, Chun M, Oh Y, Noh OK, Kim L. Metabolic comorbidities and medical institution utilization among breast cancer survivors: a national population-based study. Korean J Intern Med 2020;35(2):421-428 [FREE Full text] [doi: 10.3904/kjim.2018.172] [Medline: 31480826]
5. Rock CL, Thomson CA, Sullivan KR, Howe CL, Kushi LH, Caan BJ, et al. American Cancer Society nutrition and physical activity guideline for cancer survivors. CA Cancer J Clin 2022;72(3):230-262 [FREE Full text] [doi: 10.3322/caac.21719] [Medline: 35294043]
6. Clinton SK, Giovannucci EL, Hursting SD. The world cancer research fund/american institute for cancer research third expert report on diet, nutrition, physical activity, and cancer: impact and future directions. J Nutr 2020;150(4):663-671 [FREE Full text] [doi: 10.1093/jn/nxz268] [Medline: 31758189]
7. Baughman C, Norman K, Mukamal K. Adherence to american cancer society nutrition and physical activity guidelines among cancer survivors. JAMA Oncol 2024;10(6):789-792. [doi: 10.1001/jamaoncol.2024.0470] [Medline: 38635238]
8. Kaur H, Pisu M, Pekmezi DW, Rogers LQ, Martin MY, Fontaine KR, et al. How healthy are the diets of cancer survivors? Characteristics of those most at risk and opportunities for improvement. J Natl Compr Canc Netw 2025;23(6):e257012. [doi: 10.6004/jnccn.2025.7012] [Medline: 40359987]
9. Demark-Wahnefried W, Aziz NM, Rowland JH, Pinto BM. Riding the crest of the teachable moment: promoting long-term health after the diagnosis of cancer. J Clin Oncol 2005;23(24):5814-5830 [FREE Full text] [doi: 10.1200/JCO.2005.01.230] [Medline: 16043830]
10. Xu J, Hoover RL, Woodard N, Leeman J, Hirschey R. A systematic review of dietary interventions for cancer survivors and their families or caregivers. Nutrients Dec 2023;16(1):56 [FREE Full text] [doi: 10.3390/nu16010056] [Medline: 38201886]
11. Sremanakova J, Sowerbutts AM, Todd C, Cooke R, Burden S. Systematic review of behaviour change theories implementation in dietary interventions for people who have survived cancer. Nutrients 2021;13(2):612. [doi: 10.3390/nu13020612]
12. Jung Y, Chung J, Son H. Physical activity interventions for colorectal cancer survivors: a systematic review and meta-analysis of randomized controlled trials. Cancer Nurs 2021;44(6):E414-E428 [FREE Full text] [doi: 10.1097/NCC.0000000000000888] [Medline: 34694086]

XSL•FO
RenderX

13. Ficarra S, Thomas E, Bianco A, Gentile A, Thaller P, Grassadonio F, et al. Impact of exercise interventions on physical fitness in breast cancer patients and survivors: a systematic review. Breast Cancer 2022;29(3):402-418 [FREE Full text] [doi: 10.1007/s12282-022-01347-z] [Medline: 35278203]
14. Williams V, Brown N, Becks A, Pekmezi D, Demark-Wahnefried W. Narrative review of web-based healthy lifestyle interventions for cancer survivors. Ann Rev Res 2020;5(4):555670 [FREE Full text] [doi: 10.19080/arr.2020.05.555670] [Medline: 33294850]
15. Lavoie A, Dubé V. Web-based interventions to promote healthy lifestyles for older adults: scoping review. Interact J Med Res 2022;11(2):e37315 [FREE Full text] [doi: 10.2196/37315] [Medline: 35998024]
16. Dee EC, Muralidhar V, Butler SS, Yu Z, Sha ST, Mahal BA, et al. General and health-related internet use among cancer survivors in the United States: a 2013-2018 cross-sectional analysis. J Natl Compr Canc Netw 2020;18(11):1468-1475. [doi: 10.6004/jnccn.2020.7591] [Medline: 33152707]
17. Holmes MM. Why people living with and beyond cancer use the internet. Integr Cancer Ther 2019;18:1534735419829830 [FREE Full text] [doi: 10.1177/1534735419829830] [Medline: 30741026]
18. Williams VA, Brown NI, Johnson R, Ainsworth MC, Farrell D, Barnes M, et al. A Web-based lifestyle intervention for cancer survivors: feasibility and acceptability of survivorSHINE. J Cancer Educ 2022;37(6):1773-1781 [FREE Full text] [doi: 10.1007/s13187-021-02026-x] [Medline: 34061334]
19. People Designs, Inc.. SURVIVORSHINE. URL: https://survivorshine.org/ [accessed 2025-05-29]
20. Pekmezi D, Crane T, Oster R, Rogers L, Hoenemeyer T, Farrell D, et al. Rationale and methods for a randomized controlled trial of a dyadic, web-based, weight loss intervention among cancer survivors and Partners: the DUET study. Nutrients 2021;13(10):3141 [FREE Full text] [doi: 10.3390/nu13103472] [Medline: 34684474]
21. Demark-Wahnefried W, Oster RA, Crane TE, Rogers LQ, Cole WW, Kaur H, et al. Results of DUET: a web-based weight loss randomized controlled feasibility trial among cancer survivors and their chosen partners. Cancers (Basel) 2023;15(5):1577 [FREE Full text] [doi: 10.3390/cancers15051577] [Medline: 36900368]
22. Kaur H, Pavela G, Pekmezi DW, Rogers LQ, Cole WW, Parrish KB, et al. Dietary barriers appear to influence the effects of a dyadic web-based lifestyle intervention on caloric intake and adiposity: a mediation analysis of the DUET trial. Nutrients 2023;15(23):4918 [FREE Full text] [doi: 10.3390/nu15234918] [Medline: 38068776]
23. Health Behavior and Health Education: Theory, Research, and Practice. 2nd ed. San Francisco: Jossey-Bass; 1997.
24. Bandura A. Health promotion from the perspective of social cognitive theory. Psychol Health 1998;13(4):623-649. [doi: 10.1080/08870449808407422]
25. Adobe Acrobat Sign computer software. Version 6.0. Adobe Inc. 2020. URL: https://acrobat.adobe.com/us/en/sign.html [accessed 2026-01-23]
26. People Designs, Inc. duet4health. URL: https://duet4health.org [accessed 2025-05-29]
27. Kelley HH. The "stimulus field" for interpersonal phenomena: the source of language and thought about interpersonal events. Pers Soc Psychol Rev 1997;1(2):140-169. [doi: 10.1207/s15327957pspr0102_3] [Medline: 15647123]
28. Lewis MA, McBride CM, Pollak KI, Puleo E, Butterfield RM, Emmons KM. Understanding health behavior change among couples: an interdependence and communal coping approach. Soc Sci Med 2006;62(6):1369-1380. [doi: 10.1016/j.socscimed.2005.08.006] [Medline: 16146666]
29. Demark-Wahnefried W, Jones LW, Snyder DC, Sloane RJ, Kimmick GG, Hughes DC, et al. Daughters and Mothers Against Breast Cancer (DAMES): main outcomes of a randomized controlled trial of weight loss in overweight mothers with breast cancer and their overweight daughters. Cancer 2014;120(16):2522-2534 [FREE Full text] [doi: 10.1002/cncr.28761] [Medline: 24804802]
30. Articulate G. Articulate Storyline [computer software]. Version 360. Articulate Global, LLC. 2017. URL: https://www.articulate.com [accessed 2026-01-14]
31. Frankenfield DC, Rowe WA, Smith J, Cooney R. Validation of several established equations for resting metabolic rate in obese and nonobese people. J Am Diet Assoc 2003;103(9):1152-1159. [doi: 10.1016/s0002-8223(03)00982-9] [Medline: 12963943]
32. Fitbit Inspire 2 [wearable activity tracker]. Google LLC. URL: https://www.fitbit.com/inspire [accessed 2026-01-14]
33. Krebs-Smith SM, Pannucci TE, Subar AF, Kirkpatrick SI, Lerman JL, Tooze JA, et al. Update of the healthy eating index: HEI-2015. J Acad Nutr Diet 2018;118(9):1591-1602 [FREE Full text] [doi: 10.1016/j.jand.2018.05.021] [Medline: 30146071]
34. Automated self-administered 24-hour dietary assessment tool. National Cancer Institute. 2025. URL: https://epi.grants.cancer.gov/asa24/ [accessed 2026-01-23]
35. Rogers LQ, McAuley E, Anton PM, Courneya KS, Vicari S, Hopkins-Price P, et al. Better exercise adherence after treatment for cancer (BEAT Cancer) study: rationale, design, and methods. Contemp Clin Trials 2012;33(1):124-137 [FREE Full text] [doi: 10.1016/j.cct.2011.09.004] [Medline: 21983625]
36. ActiGraph wGT3X-BT Activity Monitor. ActiGraph LLC. URL: https://theactigraph.com/actigraph-wgt3x-bt [accessed 2025-05-29]
37. Amireault S, Godin G, Lacombe J, Sabiston CM. Validation of the Godin-Shephard Leisure-Time Physical Activity Questionnaire classification coding system using accelerometer assessment among breast cancer survivors. J Cancer Surviv 2015;9(3):532-540. [doi: 10.1007/s11764-015-0430-6] [Medline: 25666749]

XSL•FO
RenderX

38. Inc. Zoom. Version 5.15.5. Zoom Video Communications, Inc. 2025. URL: https://zoom.us/ [accessed 2026-01-23]
39. Hoenemeyer TW, Cole WW, Oster RA, Pekmezi DW, Pye A, Demark-Wahnefried W. Test/retest reliability and validity of remote vs. in-person anthropometric and physical performance assessments in cancer survivors and supportive partners. Cancers (Basel) 2022;14(4):1075 [FREE Full text] [doi: 10.3390/cancers14041075] [Medline: 35205823]
40. SAS Software, Version 9.4. SAS Institute Inc. 2025. URL: https://www.sas.com/ [accessed 2025-05-29]
41. Williams V, Brown N, Moore JX, Farrell D, Perumean-Chaney S, Schleicher E, et al. Web-Based lifestyle interventions for survivors of cancer: usability study. JMIR Form Res 2022;6(2):e30974 [FREE Full text] [doi: 10.2196/30974] [Medline: 35188468]
42. Paxton RJ, Hajek R, Newcomb P, Dobhal M, Borra S, Taylor WC, et al. A lifestyle intervention via email in minority breast cancer survivors: randomized parallel-group feasibility study. JMIR Cancer 2017;3(2):e13 [FREE Full text] [doi: 10.2196/cancer.7495] [Medline: 28935620]
43. Van Blarigan EL, Kenfield S, Chan J, Van Loon K, Paciorek A, Zhang L, et al. Feasibility and acceptability of a web-based dietary intervention with text messages for colorectal cancer: a randomized pilot trial. Cancer Epidemiol Biomarkers Prev 2020;29(4):752-760 [FREE Full text] [doi: 10.1158/1055-9965.EPI-19-0840] [Medline: 31941707]
44. van den Berg SW, Peters EJ, Kraaijeveld JF, Gielissen MF, Prins JB. Usage of a generic web-based self-management intervention for breast cancer survivors: substudy analysis of the BREATH trial. J Med Internet Res 2013;15(8):e170 [FREE Full text] [doi: 10.2196/jmir.2566] [Medline: 23958584]
45. Fredericks S, Martorella G, Catallo C. A systematic review of web-based educational interventions. Clin Nurs Res 2015;24(1):91-113. [doi: 10.1177/1054773814522829] [Medline: 24571963]
46. Chatterjee A, Prinz A, Gerdes M, Martinez S. Digital interventions on healthy lifestyle management: systematic review. J Med Internet Res 2021;23(11):e26931. [doi: 10.2196/26931]
47. Shi Y, Wakaba K, Kiyohara K, Hayashi F, Tsushita K, Nakata Y. Effectiveness and components of web-based interventions on weight changes in adults who were overweight and obese: a systematic review with meta-analyses. Nutrients 2022;15(1):179 [FREE Full text] [doi: 10.3390/nu15010179] [Medline: 36615836]
48. Carr RM, Prestwich A, Kwasnicka D, Thøgersen-Ntoumani C, Gucciardi DF, Quested E, et al. Dyadic interventions to promote physical activity and reduce sedentary behaviour: systematic review and meta-analysis. Health Psychol Rev 2019;13(1):91-109 [FREE Full text] [doi: 10.1080/17437199.2018.1532312] [Medline: 30284501]
49. John JC, Ho J, Raber M, Basen-Engquist K, Jacobson L, Strong LL. Dyad and group-based interventions in physical activity, diet, and weight loss: a systematic review of the evidence. J Behav Med 2024;47(3):355-373. [doi: 10.1007/s10865-023-00457-z] [Medline: 38017250]
50. Graham SA, Stein N, Shemaj F, Branch OH, Paruthi J, Kanick SC. Older adults engage with personalized digital coaching programs at rates that exceed those of younger adults. Front Digit Health 2021;3:642818 [FREE Full text] [doi: 10.3389/fdgth.2021.642818] [Medline: 34713112]
51. De Santis KK, Mergenthal L, Christianson L, Busskamp A, Vonstein C, Zeeb H. Digital technologies for health promotion and disease prevention in older people: scoping review. J Med Internet Res 2023;25:e43542 [FREE Full text] [doi: 10.2196/43542] [Medline: 36951896]
52. Bidmon S, Terlutter R. Gender differences in searching for health information on the internet and the virtual patient-physician relationship in Germany: exploratory results on how Men and women differ and why. J Med Internet Res 2015;17(6):e156 [FREE Full text] [doi: 10.2196/jmir.4127] [Medline: 26099325]
53. Manierre MJ. Gaps in knowledge: tracking and explaining gender differences in health information seeking. Soc Sci Med 2015;128:151-158. [doi: 10.1016/j.socscimed.2015.01.028] [Medline: 25618604]
54. Kebede AS, Ozolins L, Holst H, Galvin K. Digital engagement of older adults: scoping review. J Med Internet Res 2022;24(12):e40192. [doi: 10.2196/40192]
55. Pogrebnoy D, Dennett AM, Simpson DB, MacDonald-Wicks L, Patterson AJ, English C. Effects of using websites on physical activity and diet quality for adults living with chronic health conditions: systematic review and meta-analysis. J Med Internet Res 2023;25:e49357 [FREE Full text] [doi: 10.2196/49357] [Medline: 37856187]
56. Linke SE, Dunsiger SI, Gans KM, Hartman SJ, Pekmezi D, Larsen BA, et al. Association between physical activity intervention website use and physical activity levels among Spanish-speaking Latinas: randomized controlled trial. J Med Internet Res 2019;21(7):e13063 [FREE Full text] [doi: 10.2196/13063] [Medline: 31342902]
57. Hachiya J, Oliveira R. The online community role in physical activity interventions. Procedia Computer Science 2024;239:781-789. [doi: 10.1016/j.procs.2024.06.236]
58. Funk KL, Stevens VJ, Appel LJ, Bauck A, Brantley PJ, Champagne CM, et al. Associations of internet website use with weight change in a long-term weight loss maintenance program. J Med Internet Res 2010;12(3):e29 [FREE Full text] [doi: 10.2196/jmir.1504] [Medline: 20663751]
59. Hendrie GA, Baird DL, Brindal E, Williams G, Brand-Miller J, Muhlhausler B. Weight loss and usage of an online commercial weight loss program (the CSIRO Total Wellbeing Diet Online) delivered in an everyday context: five-year evaluation in a community cohort. J Med Internet Res 2021;23(6):e20981 [FREE Full text] [doi: 10.2196/20981] [Medline: 34096869]

XSL•FO
RenderX

60. Dent R, McPherson R, Harper M. Factors affecting weight loss variability in obesity. Metabolism 2020;113:154388. [doi: 10.1016/j.metabol.2020.154388] [Medline: 33035570]

Abbreviations

ACS: American Cancer Society
CSIRO: Commonwealth Scientific and Industrial Research Organisation
DUET: Daughters, Dudes, Mothers, and Others Together
MVPA: moderate to vigorous physical activity
V&F: vegetable and fruit
WCRF/AICR: World Cancer Research Fund/American Institute for Cancer Research
WLM: Weight Loss Maintenance

Edited by A Mavragani; submitted 31.Oct.2025; peer-reviewed by O Ibikunle, E Madondo, A Famotire; comments to author 25.Nov.2025; revised version received 07.Dec.2025; accepted 23.Dec.2025; published 30.Jan.2026.

Please cite as:
Kaur H, Pekmezi D, E Crane T, Farrell D, Q Rogers L, Demark-Wahnefried W
Website Use and Associations With Behavior Change and Weight Loss in Cancer Survivors and Their Partners: Secondary Analysis of a Randomized Controlled Trial
J Med Internet Res 2026;28:e86908
URL: https://www.jmir.org/2026/1/e86908
doi:10.2196/86908
PMID:

©Harleen Kaur, Dori Pekmezi, Tracy E Crane, David Farrell, Laura Q Rogers, Wendy Demark-Wahnefried. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Traditional Rehabilitation Experiences, Unmet Needs, and Perspectives on Virtual Reality–Based Rehabilitation Among Patients With Stroke in China: Qualitative Thematic Analysis and Semistructured Interview Study

Xite Zheng[1], MS; Lu Xing[2], MSc; Haitao Lu[3], MD; Shimeng Hao[2*], PhD; Fen Liu[1*], PhD

[1]Department of Epidemiology and Health Statistics, School of Public Health, Beijing Key Laboratory of Environment and Aging, Capital Medical University, Beijing, China

[2]School of Architecture and Urban Planning, Beijing University of Civil Engineering and Architecture, Beijing, China

[3]Department of Neurorehabilitation, Beijing Bo'ai Hospital, Rehabilitation Research Center, School of Rehabilitation Medicine, Capital Medical University, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Fen Liu, PhD
Department of Epidemiology and Health Statistics
School of Public Health, Beijing Key Laboratory of Environment and Aging
Capital Medical University
No. 10 Xitoutiao, Youanmenwai Street
Fengtai District
Beijing, 100069
China
Phone: 86 010 83911497
Email: liufen05@ccmu.edu.cn

Abstract

Background: Traditional stroke rehabilitation is facing challenges, and virtual reality (VR)–based rehabilitation is a promising solution. However, results from studies focusing on VR-based stroke rehabilitation remain inconsistent, largely due to the use of noncustomized interventions in previous trials.

Objective: To enhance rehabilitation services and inform the development of patient-centered VR rehabilitation systems, this study aimed to (1) explore the experiences and unmet needs of survivors of stroke during current hospital rehabilitation, and (2) examine their perspectives on the use of VR technology in poststroke rehabilitation.

Methods: We conducted a qualitative thematic analysis based on descriptive phenomenology between January and July 2025 at the China Rehabilitation Research Center. Adult patients with a clinical diagnosis of stroke within the past 18 months were eligible. A total of 21 survivors of stroke (mean age, 52.7, SD 17.3 y; men, n=17) were included. Data were collected through face-to-face semistructured interviews, complemented by a short questionnaire on sociodemographic, clinical, and technology-use characteristics. All interviews were audio-recorded, transcribed verbatim, and analyzed using a thematic approach, with thematic saturation used to determine the sample size.

Results: After a stroke, patients experience significant physical and psychological changes. On the one hand, the sudden loss of abilities alters their perceived roles within the family and society; on the other hand, the sharp contrast between their desire for recovery and their current recovery limitations creates substantial psychological pressure. Accepting their condition and rebuilding confidence is a long-term process. Traditional rehabilitation is commonly described as burdensome, monotonous, and lacking continuity after discharge. Although patients desire a better rehabilitation approach and improved outcomes, attitudes toward VR-based rehabilitation vary. Some view VR as a convenient tool, while others express no interest or perceived need for technology-based rehabilitation. Patients indicated that serious games should be diversified to meet different individual and training needs, and should incorporate clearer feedback mechanisms, appropriate scoring systems, adjustable difficulty levels, and progressive game chapters. Functional expectations for VR systems included family involvement, access to personal rehabilitation data, telerehabilitation support, safety monitoring, and technical support.

XSL•FO
RenderX

Conclusions: Stroke rehabilitation services in China require improvement in the appeal of rehabilitation content, patient self-management, and continuity of care. Although patients desire better rehabilitation approaches and outcomes, the effective integration of VR technology must account for factors, such as personal characteristics and preferences, as well as socioeconomic status. Unlike previous studies that primarily examined user experiences with digital technologies or compared rehabilitation outcomes, our research contributes to the literature by linking the challenges and patient needs in conventional rehabilitation with concrete directions for the future design of VR rehabilitation. These insights deepen current understanding of how VR technologies can be meaningfully integrated into stroke care and provide a roadmap for developing patient-centered and culturally responsive VR solutions.

(J Med Internet Res 2026;28:e84532) doi:10.2196/84532

KEYWORDS

stroke patients; rehabilitation; digital health technology; virtual reality; qualitative study; user-requirements; patient-centered design

Introduction

Stroke is a leading cause of death and disability [1]. In 2021, there were an estimated 93.8 million survivors of stroke and 11.9 million new cases globally, including 26.3 million survivors and 4.1 million new cases in China [2]. Half of the survivors with stroke are left disabled, with one-third requiring assistance in daily activities [3]. In addition to physical impairments, cognitive and emotional disturbances, such as memory deficits, aphasia, and depression, are common and further reduce quality of life and hinder social reintegration [4]. Therefore, effective rehabilitation remains a critical component of stroke care in the long term.

While the rehabilitation research is moving toward the exploration of digital health technologies, current stroke rehabilitation in China primarily consists of physical therapy, occupational therapy, and speech therapy, with treatment plans tailored to individual conditions [5]. The demand for stroke rehabilitation services is increasing; however, substantial advances are yet to be made in stroke rehabilitation practice to meet this demand, including improving the effectiveness of current rehabilitation, enhancing adherence, and addressing limited rehabilitation resources, especially in home-based settings [6]. A deeper understanding of the real-world experiences and unmet needs of survivors of stroke throughout the rehabilitation process is critical for informing future service improvement and intervention development.

Virtual reality (VR)–based rehabilitation uses motion tracking, stereoscopic display, and real-time feedback to create immersive, interactive training environments, offering promising solutions for stroke rehabilitation [7]. Previous studies have shown that gamified training improves patient engagement and adherence [8]. Furthermore, the portability of the equipment and internet connectivity can improve access to treatment, promoting telerehabilitation, fostering equity, and supporting patient self-management [9,10]. VR-based rehabilitation has significant implications for improving current stroke care. Nevertheless, findings on its effectiveness remain mixed [11,12]. One major challenge is that many existing VR interventions rely on noncustomized content, such as commercial games, which fail to address the diverse functional needs, abilities, and preferences of survivors of stroke, ultimately limiting therapeutic effect [13]. Moreover, age, socioeconomic status, and educational differences influence individuals' access to and familiarity with VR technology, further shaping intervention acceptability and outcomes [14]. The success of VR-based rehabilitation depends not only on technological efficacy but also on its alignment with the lived acceptability, preferences, and perceived needs of survivors of stroke. Therefore, listening to patients' voices and developing patient-centered VR rehabilitation systems is crucial for addressing the current challenges [15].

Quantitative studies have provided important evidence on the effectiveness and feasibility of stroke rehabilitation interventions; however, they are limited in their ability to capture the rehabilitation experiences, unmet needs, and subjective perspectives of rehabilitation services among survivors of stroke. Qualitative approach and semistructured interviews were widely used to understand human phenomena, including patients' thoughts and experiences [16,17]. Such information is essential for informing the development of patient-centered and culturally appropriate VR-based rehabilitation interventions.

Therefore, this study used qualitative methods to (1) explore the experiences and unmet needs of Chinese survivors of stroke during current hospital rehabilitation and (2) examine their perspectives on the use of VR technology in poststroke rehabilitation. The overarching goal is to enhance rehabilitation services and inform the development of patient-centered VR rehabilitation systems.

Methods

Research Design Overview

This study adopted a qualitative thematic analysis based on descriptive phenomenology to explore the rehabilitation experiences, unmet needs, and perspectives of survivors of stroke on VR-based rehabilitation [17,18]. Data were collected through semistructured interviews and analyzed using thematic analysis. This study was conducted at the China Rehabilitation Research Center (Beijing Boai Hospital), as part of a larger project entitled Development of a Motor and Cognitive Rehabilitation System for Stroke Patients Based on Multisensory Virtual Reality Technology.

Researcher Description

This study was conducted through a collaboration among Capital Medical University, Beijing University of Civil Engineering

XSL•FO
RenderX

and Architecture, and the China Rehabilitation Research Center. The research team consists of university professors, chief physicians, and doctoral and master's students, all of whom have received professional training and possess extensive experience in stroke rehabilitation, qualitative research, and digital health. The primary interviewer and data analyst (XZ and LX) had received formal training in qualitative interviewing and thematic analysis, and possessed previous knowledge of stroke rehabilitation practices and the research landscape in China. XZ also had previous experience conducting and analyzing interviews with public health practitioners and primary and secondary school students. Reflexivity was maintained through regular team discussions, during which researchers reflected on their disciplinary backgrounds and potential assumptions to minimize interpretive bias. These experiences and foundation enabled the research team to successfully conduct this study and ensured the rigor of the methodology.

Participants

The final sample included 21 participants (4 women and 17 men) with a mean age of 52.7 (SD 17.3) years. For most participants (19/21, 90%), the time since stroke onset was less than 6 months. The majority (14/21, 67%) reported daily internet use, and nearly all (19/21, 90%) owned a smartphone or tablet. Detailed participant characteristics are presented in Table 1.

XSL•FO
RenderX

Table 1. Demographic characteristics of patients with stroke at the Department of Neurorehabilitation, China Rehabilitation Research Center (January to July 2025, N=21, qualitative thematic analysis based on descriptive phenomenology).

Patient	Sex	Age (y)	Education level	Marital status	Time since stroke (months)	Stroke type	Perceived health	Smartphone or tablet	Use of the internet
P01	Female	62	Senior high or vocational school	Divorced	0-3	Ischemic	Good	Smartphone	Every day
P02	Male	90	Senior high or vocational school	Widowed	12-18	Ischemic	Fair	No	No
P03	Male	76	Junior high school	Married	0-3	Ischemic	Good	Smartphone	Every day
P04	Female	60	Junior high school	Married	0-3	Ischemic	Poor	No	No
P05	Male	46	Senior high or vocational school	Married	3-6	Ischemic	Fair	Smartphone	Multiple times a week
P06	Male	38	College, University, or above	Divorced	3-6	Ischemic	Fair	Both	Every day
P07	Male	58	Junior high school	Married	3-6	Hemorrhagic	Fair	Smartphone	Multiple times a week
P08	Female	69	Primary school or below	Married	0-3	Ischemic	Fair	Smartphone	Every day
P09	Male	55	College, University, or above	Married	0-3	Ischemic	Good	Smartphone	Every day
P10	Male	46	College, University, or above	Single	3-6	Hemorrhagic	Fair	Smartphone	Every day
P11	Male	30	Senior high or vocational school	Single	3-6	Hemorrhagic	Fair	Both	Multiple times a week
P12	Male	71	Senior high or vocational school	Married	3-6	Hemorrhagic	Good	Both	Multiple times a week
P13	Male	40	College, University, or above	Married	0-3	Ischemic	Fair	Both	Every day
P14	Female	27	College, University, or above	Married	6-12	Hemorrhagic	Fair	Both	Every day
P15	Male	62	Junior high school	Married	0-3	Hemorrhagic	Fair	Smartphone	Every day
P16	Male	38	College, University, or above	Married	3-6	Hemorrhagic	Fair	Smartphone	Multiple times a month
P17	Male	62	Junior high school	Married	3-6	Hemorrhagic	Fair	Both	Every day
P18	Male	32	College, University, or above	Single	0-3	Hemorrhagic	Good	Smartphone	Every day
P19	Male	32	College, University, or above	Married	0-3	Hemorrhagic	Fair	Both	Every day
P20	Male	70	College, University, or above	Married	0-3	Ischemic	Fair	Smartphone	Every day
P21	Male	43	College, University, or above	Married	3-6	Hemorrhagic	Fair	Smartphone	Every day

Researcher–Participant Relationship

The researchers and participants had no previous relationship. At the scheduled interview time, a nurse introduced the researchers to the participants, and the participants felt at ease throughout the process.

Recruitment Process

Physician-researchers screened medical records to identify eligible participants. Patients meeting the inclusion criteria were approached during hospitalization, and those who provided verbal consent were later contacted by trained interviewers to schedule a face-to-face interview. The sample size was determined by the principle of data saturation, defined as the point at which no new themes or insights emerged during analysis [19]. A total of 26 survivors of stroke were interviewed, of whom 5 were excluded due to poor communication ability, leaving 21 participants in the final analysis.

XSL•FO
RenderX

Participant Selection

Participants were recruited through purposive sampling from the Department of Neurorehabilitation at the China Rehabilitation Research Center. Inclusion criteria were (1) age≥18 years, (2) first-ever stroke diagnosed within the past 18 months, and (3) ability to provide written informed consent. The 18-month time frame was chosen to capture both early and later recovery experiences and to examine the durability of these effects [20], while ensuring that participants were physically able to engage with the VR devices safely. To minimize unnecessary exclusion, patients with mild cognitive or communication impairments were permitted to receive support from informal caregivers when communication was slow. Exclusion criteria included inability to speak Chinese or insufficient communication capacity, as determined at the time of interview. Data collection took place between January and July 2025.

Data Collection

Data were collected through face-to-face semistructured interviews following a predefined sequence. At the scheduled interview time, a nurse led 2 researchers (XZ and LX), both trained in qualitative research, to the patient's bedside to confirm the appointment. After confirmation, the nurse left the room. The researchers then introduced themselves to the participant and explained the purpose and significance of the study. The participant then read and voluntarily signed the written informed consent. Participants were subsequently invited to complete 2 rounds of the Fruit Ninja (Halfbrick) game using a Pico 4 Enterprise headset (ByteDance; 1 round with visual access to the surrounding room environment and 1 without). Following the VR experience, participants completed a brief questionnaire collecting demographic and clinical information, including sex, age, marital status, education level, time since stroke, stroke type, self-perceived health, and use of mobile devices and the internet. Interviews were conducted after completion of the VR task and questionnaire, and took place either in the patient's room or a doctor's office to ensure a quiet, clean, and private environment. An interview guide (Multimedia Appendix 1), developed by the research team based on the study objectives and a review of relevant literature and refined through pilot interviews with 2 patients (data from the pilot interviews were not included in the final analysis), was used to facilitate the interviews. Interviews began with questions about stroke history and rehabilitation experiences, followed by exploration of perceived challenges, unmet needs, and suggestions for improving poststroke care. Participants were then asked to reflect on their VR experience, including attitudes, preferences, and perceptions of usability. Follow-up questions probed barriers to VR use and elicited recommendations for improvement. Open-ended prompts ("Is there anything else you would like to add regarding today's topic?") were used to encourage additional insights. Interviews lasted between 16 and 40 minutes (25 minutes on average), were audio-recorded, and transcribed verbatim.

Data-Analytic Strategies

Data were analyzed using thematic analysis following Braun and Clarke's 6-phase framework [21,22]. All interviews were transcribed within 24 hours by 2 researchers (XZ and LX). The first author completed a verbatim transcription, and the second author checked each transcript against the audio recordings to ensure accuracy and enhance familiarity with the dataset. Furthermore, the 2 researchers (XZ and LX) independently read and coded all of the transcripts using an inductive, data-driven approach, focusing on participants' rehabilitation experiences, needs, and perspectives on VR technology. After independent coding, the 2 researchers compared their coding and discussed similarities and discrepancies. Discrepancies were resolved through in-depth discussion, during which the researcher explained their coding decisions with reference to the original transcripts. When consensus could not be reached initially, the relevant data segments were re-examined, and codes were refined or merged as appropriate until agreement was achieved. Codes were then organized into potential themes and subthemes, which were reviewed for consistency and relevance across transcripts. Recoding was performed when necessary. Themes were iteratively refined and finalized in consultation with the broader research team. All qualitative data were managed and analyzed using NVivo software (version 1.2; Lumivero). Selected quotations were translated into English using a forward-backward translation process.

Methodological Integrity

Methodological integrity was ensured by aligning the study design, data collection, and analytic approach with the research aims. The study was conducted at the China Rehabilitation Research Center in Beijing, China, a leading institution with more than 30 years of experience in rehabilitation medicine and research, serving patients from across the country. A qualitative design informed by descriptive phenomenology was used to capture the rehabilitation experiences, unmet needs, and perspectives of survivors of stroke on VR-based rehabilitation. Semistructured interviews allowed participants to express their experiences, needs, and perspectives in their own words while ensuring coverage of key topics relevant to the research questions. Data collection continued until thematic saturation was reached, defined as the point at which no new themes emerged from successive interviews. Interviews were audio-recorded, transcribed verbatim, and analyzed using an inductive thematic analysis approach. Furthermore, 2 researchers (XZ and LX) independently coded the transcripts and engaged in iterative discussions to resolve discrepancies and refine the coding framework. Strategies to enhance methodological rigor, including reflexivity, trustworthiness, and ethical conduct, were integrated throughout the research process. These aspects are described in detail in the following sections: Researcher Description, Trustworthiness, and Ethical Considerations. This study adhered to the American Psychological Association's reporting standards for qualitative research [23] and the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist (Multimedia Appendix 2) [24].

Ethical Considerations

This study was reviewed and approved by the Medical Ethics Committee of Capital Medical University, Beijing, China (approval Z2025SY006). All procedures complied with the principles of the Declaration of Helsinki. Participants were

XSL•FO
RenderX

informed of the study's purpose, procedures, potential risks, and their right to withdraw at any time. Written informed consent was obtained before participation. To ensure privacy and confidentiality, data were deidentified through pseudonymization. Contact information was stored separately from research data. No images or personally identifiable information were included in the manuscript or supplementary materials. Participants received a small token of appreciation in the form of daily necessities valued at 20-30 Chinese yuan, equivalent to US $3-US $4.

Trustworthiness

Trustworthiness was established following Guba and Lincoln's criteria [25], encompassing credibility, transferability, dependability, and confirmability. Credibility was enhanced through timely verbatim transcription and cross-checking within 24 hours to ensure accurate and faithful representation of participants' narratives. All interviews and analyses were conducted by 2 researchers (XZ and LX), facilitating deep engagement with the data. Transferability was supported by providing detailed descriptions of the study setting, recruitment context, and participant characteristics, enabling readers to assess the relevance of the findings to other contexts. Dependability and confirmability were strengthened through transparent documentation of data collection and analysis procedures. Data were independently coded by 2 researchers, with discrepancies resolved through discussion and consensus in consultation with the wider research team. Thematic analysis followed Braun and Clarke's 6-phase framework, reducing potential researcher bias and enhancing analytic reliability.

Results

Overview

Thematic analysis generated six overarching themes: (1) changes following stroke and self-reconstruction, (2) effective yet challenging traditional rehabilitation, (3) unmet needs in the rehabilitation journey, (4) attitudes toward VR-based rehabilitation, (5) recommendations for serious game design, and (6) suggested features of VR systems (Figure 1). A summary of themes and representative quotations is provided in Multimedia Appendix 3.

XSL•FO
RenderX

Figure 1. Themes and subthemes identified during data analysis of patients with stroke at the Department of Neurorehabilitation, China Rehabilitation Research Center (January and July 2025, N=21, qualitative thematic analysis based on descriptive phenomenology). VR: virtual reality.



Changes Following Stroke and Self-Reconstruction

Shifts in Personal and Social Roles

All participants experienced varying degrees of physical impairment following stroke. Some reported relatively mild deficits, such as limb weakness or fatigue, whereas others described severe hemiplegia accompanied by cognitive impairments, including confusion and aphasia. These functional losses led to substantial disruptions in their independence and resulted in major shifts in their roles within both personal and social contexts.

> *Overnight, my entire left side became paralyzed. I can't do any work now. […] I've brought a lot of inconvenience to my family and close friends. That's what pains me the most.* [P09]
>
> *Half of my body doesn't move, so I rely on others for everything. At first, I couldn't even speak or recognize people. I couldn't do anything I was supposed to do. I became a burden to my family and brought misfortune to them.* [P13]

Inner Psychological Conflicts

The sudden loss of physical abilities and the associated role changes had profound psychological consequences. Participants described intense feelings of uselessness and a strong desire to recover bodily functions, which collectively created significant emotional pressure. Many reported experiencing anxiety, guilt, low self-esteem, irritability, and even despair.

> *After I got sick, I was constantly worried and sad. […] I felt irritable and depressed all the time.* [P08]
>
> *Right after it happened, I felt useless and even had thoughts of ending my life. But you cannot die even if you want to, because you cannot move. I couldn't*

XSL•FO
RenderX

even pick up a knife to hurt myself. It was a feeling of utter despair. [P19]

Coming to Terms With Changes

Despite these challenges, participants described conscious efforts to adopt a more positive mindset and gradually accept their condition. They highlighted the importance of patience, emotional stability, and realistic expectations to support long-term rehabilitation. Family encouragement and professional guidance played key roles in fostering acceptance and maintaining motivation.

Don't rush it, take your time. Since you already have this illness, you need to stay calm. Whether it takes one or two years, even four, it is still the same condition. Recovery is slow, so you need patience. Being too anxious is not good for your recovery; once your mindset becomes tense, nothing goes well. [P18]

Later on, with my family's encouragement and the doctors' support, I gained much more confidence in my recovery. You see, now I'm doing quite well. [P19]

Effective Yet Challenging Traditional Rehabilitation

Engagement Barriers

Many participants described current hospital rehabilitation as densely scheduled, repetitive, and lacking engagement. The monotony and high intensity of daily sessions often resulted in insufficient rest, leading to physical exhaustion and psychological strain. Over time, these factors diminished motivation, reduced adherence, and required substantial willpower to maintain participation.

The training is truly monotonous. I think it really depends on one's willpower and endurance. [P05]

My current rehabilitation includes OT, PT, robotic arm training, acupuncture, and massage. Except for massage, all the training is boring, just repeating the same movements every day. [P06]

I have trained every weekday, only resting on weekends. My daughter had to push me in a wheelchair, and we are always rushing to get there on time; otherwise, I won't be allowed to join the session. […] It's exhausting. Sometimes my legs are swollen by the time I return to the ward. [P08]

Lack of Continuity

Most participants reported performing little or no structured rehabilitation after hospital discharge. The absence of professional supervision, appropriate equipment, and adequate space at home posed major barriers to continuing rehabilitation outside the clinical environment.

At home, I can only go out for a walk and treat that as exercise. [P01]

After I went home, I didn't train at all. I think many people are like this. How can you train at home? There's no equipment, no one to guide you, and no conditions for proper training. [P03]

Perceived Benefits

Despite the challenges encountered, most participants expressed overall satisfaction with the effectiveness of rehabilitation. They also highlighted the emotional support provided by health care professionals, which contributed to their sense of comfort, motivation, and hope.

The therapists here are very experienced. The environment in the rehabilitation department is alsoheartwarming. Someone will hold my hand or greet me kindly with comforting words, this really touches me. [P02]

It's been almost a month, and I'll be discharged in a few days. My leg has improved a lot. [P03]

During the rehabilitation process, the doctors constantly encouraged me, saying things like 'You're making progress' or 'You're much better than a few days ago.' [P04]

Unmet Needs in the Rehabilitation Journey

Desire for Better Approaches

Several participants described transferring to the current hospital specifically to access more advanced technologies and effective rehabilitation interventions. Their accounts reflected a clear recognition of the limitations of previous treatments and a strong willingness to explore innovative strategies to optimize recovery.

This is already the third hospital I have been to. I started at the best hospital in my province, then went to the best local rehabilitation center, and now I am here at what they say is the top rehabilitation center in the country. I don't really know what treatment methods they use here, so I feel the need to try and find out. [P09]

Need for Rehabilitation Knowledge

Given the complexity and long duration of stroke recovery, many participants reported a lack of essential knowledge about their condition and appropriate rehabilitation strategies. They noted that their limited understanding of stroke mechanisms and treatment options hindered communication with health care providers and made it difficult to actively participate in decision-making.

We are not professionals, we don't understand how this illness occurs or what the whole process involves. […] We truly lack knowledge. Now they are giving us these treatments, but we don't know which ones are suitable for us. Even when doctors explain things, we still struggle to understand. [P10]

Demand for Social Respect

Participants emphasized that social acceptance and respect are crucial for psychological resilience, well-being, and successful social integration. They noted that the support and treatment they receive vary greatly both within rehabilitation facilities and in society at large, and expressed a desire to receive the same level of respect outside of rehabilitation institutions.

When I came to this hospital, I noticed that some of the staff members are also people with disabilities

XSL•FO
RenderX

> *like us. It shows we can still work and support ourselves even if we are disabled. This encourages me not to give up on life or my future. This is unlike the negative comments made by some people in society. The doctors and nurses are all very kind; they never show any prejudice because of our condition. Their respectful communication really helps with psychological healing.* [P19]

Attitudes Toward VR-Based Rehabilitation

Praised and Positive Perceptions

Most participants generally expressed a positive attitude toward VR-based rehabilitation. They highlighted that the gamified and immersive nature of VR could make rehabilitation more engaging, increase motivation, and potentially enhance adherence compared with conventional, repetitive training. Participants also noted that VR could help address resource limitations after discharge and improve access by enabling rehabilitation outside the hospital environment.

> *Once this device is fully developed and widely promoted, it could be placed at the nurses' station for patients to use in rotation. Patients could even purchase one to use at home.* [P03]

> *I think your VR device is very interesting, it's completely virtual. My arm still feels a bit tired afterward, but the experience was enjoyable. […] Compared with the usual hospital treatments like cycling or traditional therapy, which are quite dull, this was much more engaging, though of course I can't complain too much about hospital treatment either.* [P20]

Skepticism and Lack of Interest

In contrast, a subset of participants expressed hesitation or disinterest in using VR for rehabilitation. Many preferred conventional, face-to-face therapy due to its direct physical interaction and perceived reliability. Age-related limitations, physical discomfort, and economic concerns were also mentioned as barriers to adoption.

> *At my age, using computers is difficult. Sometimes I even feel dizzy.* [P01]

> *It depends on how much your device costs. I just tried it and my first impression was good, but if the price is too high, not everyone will be able to afford it, right?* [P10]

> *Honestly, I think what you are doing is great, but I just don't think it's necessary. No matter how good it is, I personally don't need it.* [P15]

> *I still prefer traditional rehabilitation. It feels more intuitive—you can see it and touch it.* [P17]

Recommendations for Serious Game Design

Diversified Game Types

Some participants suggested expanding the range of game genres, as people with different characteristics have different preferences. They emphasized that serious games should be tailored to users' age, preferences, and rehabilitation stages. Incorporating culturally familiar or age-appropriate activities, such as calligraphy, chess, or dancing, was viewed as essential to ensure wider appeal and encourage maintained engagement.

> *I think you should design more types of games. I prefer games set in natural scenery and am not interested in this sword-based game.* [P16]

> *I think age differences matter. For people in their fifties or sixties, slicing fruit may not interest them. You could include chess or similar games that are more suitable for that age group.* [P19]

> *There are gender differences. Women might prefer dancing games.* [P20]

Customized Training Content

Although enjoyment was appreciated, participants emphasized that rehabilitation should remain the primary focus of serious games. They recommended developing targeted training modules that correspond to specific functional impairments and rehabilitation goals, such as upper limb strength, hand dexterity, or fine-motor tasks. They suggest creating a library of customizable exercises that clinicians could assign based on individual needs, ensuring both personalization and clinical relevance.

> *To me, muscle and strength training are most important because I want to walk independently. The game is great, and the virtual environment feels real, but I hope for more specialized training. For example, if I want to practice my hand, then exercises targeting finger movement or using chopsticks should be available. You could develop a game library with many options, and clinicians can select the ones that suit us.* [P05]

Desired Game Functions

Participants also proposed functional enhancements to improve the usability and therapeutic value of serious games. Adjustable difficulty levels and sequential game chapters were recommended to accommodate different impairment levels and to maintain long-term engagement. Scoring systems were considered valuable for monitoring progress and promoting self-motivation. Additionally, immersive environments and clear feedback, through audiovisual prompts, tactile feedback, or visual cues, were highlighted as necessary features, especially for individuals with sensory limitations.

> *Is the vibration on the controller too weak? I didn't feel any vibration. You could add prompts or voice cues like 'please touch' or use arrows such as 'move the controller here.' That would be very helpful.* [P04]

> *I like a fully virtual environment because the scene feels realistic, and it gives you a sense of immersion. You're not distracted by the outside environment, so you stay more focused during training.* [P05]

> *If you add a scoring system—like getting 1000 points today and 1500 tomorrow—I would feel like I'm making progress. By the third day, I'd aim for an even*

XSL•FO
RenderX

higher score. That kind of thing motivates players. [P10]

Gradually increasing the difficulty, such as by speeding up or offering continuous stages, helps people adapt and stay interested in continuing. [P11]

Suggested Features of VR Systems

Family Involvement

Participants noted that VR rehabilitation devices resembled home gaming consoles, which made them especially appealing to family members, particularly children. Participants expressed a desire for family-connected gameplay, suggesting that allowing relatives to join the games could enhance enjoyment and strengthen emotional bonds during the recovery process.

This looks like a children's game, [...] can my grandson play it? [P01]

This is similar to the game consoles we have at home. Doesn't Xbox have a game like this? […] I go home mainly to spend time with my daughter. She loves these kinds of games, so if we could play together, or if she could play with me in some way, that would be wonderful. [P21]

Access to Personal Rehabilitation Data

Participants showed strong interest in accessing personal rehabilitation data directly through the VR system. They emphasized that personal medical and training records are crucial for understanding their treatment history, clarifying rehabilitation goals, and monitoring functional changes. Such transparency was viewed as essential for enhancing engagement, supporting self-management, and giving patients a sense of control over their rehabilitation journey.

Can you create a rehabilitation profile or progress tracking system? It looks like a game on the surface, but it's actually rehabilitation training. You need to know what you're training while playing and what you should do next. [P05]

Telerehabilitation Support

Participants recommended integrating educational materials and professional guidance into the VR system to support rehabilitation outside the hospital. They recognized the potential of VR technology in providing remote access to medical services and ensuring continuity of rehabilitation.

If, during the rehabilitation process, doctors could see my progress and send me some rehabilitation-related suggestions, that would definitely be meaningful. [P02]

The equipment in hospitals is very expensive and too large to use at home. If VR can be applied to home-based rehabilitation, I think it's an excellent method. It's lightweight, you can use it at home, and it still provides effective rehabilitation. That's very good. [P19]

Safety Monitoring

Safety emerged as a significant concern. Participants recommended incorporating physiological monitoring, such as blood pressure or fatigue indicators, and automated alerts to prevent overexertion or adverse events. They suggested implementing time limits, real-time feedback on training intensity, and emergency warnings. These proposals emphasize the need for built-in safeguards to protect vulnerable users, particularly older adults and those with cardiovascular conditions.

You have to limit the game time. What if someone plays all night? That would be dangerous. [P12]

You could add some features focused on health monitoring. Since we mainly track blood pressure, you could include blood pressure measurements and provide specific values. That would make it medically useful and safer. Otherwise, someone might get too excited while playing and suddenly feel dizzy or faint. You could set an alert when blood pressure reaches a certain level, reminding the player to stop. [P19]

Technical Support

Although participants found the current VR device easy to operate, many expressed concerns about potential technical issues or malfunctions, especially when used at home. They emphasized the need for accessible customer support and reliable maintenance to ensure confidence in long-term use. Suggested options included 24-hour service hotlines, troubleshooting through widely used communication platforms such as WeChat (Tencent Holdings Limited), and guaranteed repair.

If I buy it and take it home, I need help when the device has problems, something like a customer service number or a WeChat account. [P03]

If I am going to use it long-term, it needs to have warranty services. If it breaks, someone must be responsible for repairing it. [P12]

Discussion

Principal Findings

This qualitative study explored the experiences and unmet needs of survivors of stroke during current hospital rehabilitation, as well as their perspectives on integrating VR technology into stroke rehabilitation. Thematic analysis revealed six overarching themes: (1) changes following stroke and self-reconstruction, (2) effective yet challenging traditional rehabilitation, (3) unmet needs in the rehabilitation journey, (4) attitudes toward VR-based rehabilitation, (5) recommendations for serious game design, and (6) suggested features of VR systems. Unlike previous studies that primarily examined user experiences with digital technologies or compared rehabilitation outcomes, our research contributes to the literature by linking the challenges and patient needs in conventional rehabilitation with concrete directions for the future design of VR-based rehabilitation.

Survivors of stroke experience profound shifts in their roles within families and society, and the stark contrast between their

XSL•FO
RenderX

desire for recovery and the harsh realities of impairment generates substantial psychological pressure. Support from family members and health care professionals plays a critical role in helping patients regain confidence and rebuild their lives. Traditional rehabilitation programs are often described as burdensome and monotonous, and the near absence of home rehabilitation presents another major challenge. Although patients desire more engaging rehabilitation approaches and better functional outcomes, future VR rehabilitation devices must account for differences in personal characteristics and preferences, as well as socioeconomic factors. Moreover, to maximize both effectiveness and acceptability, VR rehabilitation systems should incorporate serious games tailored to diverse personal and training needs, and integrate functions such as family interaction, progress tracking, telehealth support, safety monitoring, and technical assistance.

Comparison With Previous Work

The level of exercise rehabilitation among survivors of stroke remains far below guideline recommendations, often characterized by the "three lows": low initiative, poor adherence, and reduced willingness to participate [26]. From a Chinese cultural perspective, rehabilitation often relies on rest or passive treatments, such as the concept of "Deqi" in traditional Chinese medicine [27], which may weaken motivation for active, repetitive training. Furthermore, access to rehabilitation services remains uneven across regions, and high out-of-pocket costs further reduce opportunities for maintained participation [28]. In our study, traditional rehabilitation exercises were frequently described as burdensome, monotonous, and lacking continuity. Previous studies have shown that repetitive movements without feedback can diminish motivation among patients with stroke [29]. After hospital discharge, many participants experienced a sharp decline in available resources and limited professional guidance [30]. In contrast, VR-based rehabilitation is reported to be more engaging than traditional face-to-face therapy with a therapist [31]. The portability of VR devices and the gamified approach to intervention can enhance motivation and extend rehabilitation into the community and home environments, thereby improving continuity of care [32].

Consistent with previous studies, survivors of stroke in our sample expressed varying attitudes toward VR-based rehabilitation [33]. Particularly in the context of traditional Chinese culture, many individuals still view games as purely for entertainment, which may lead to skepticism about game-based rehabilitation. This underscores the importance of individual readiness, digital literacy, and personal preferences in shaping acceptance and effectiveness of VR-based interventions. Notably, even participants who were initially skeptical of VR acknowledged the value of advanced hospitals and innovative rehabilitation technologies and expressed a desire for improved rehabilitation programs. Previous research indicates that patients are more likely to adopt digital health tools when they are simple, intuitive, and user-friendly [34], whereas poorly designed equipment can lead to negative experiences and reduced adherence [35]. Importantly, firsthand experience with the benefits of digital health technologies plays a key role in promoting acceptance and maintained use [36]. Therefore, when implementing VR in clinical practice, it is essential not only to clearly communicate the practical benefits of VR-based rehabilitation but also to address patients' concerns related to usability, safety, and reliability.

Serious games have been increasingly evaluated in clinical rehabilitation protocols, with studies reporting improved patient engagement and functional outcomes [37]. However, understanding how patients with diverse characteristics perceive different game elements is essential for maximizing acceptance, adherence, and therapeutic effectiveness. In our study, participants emphasized the importance of offering game types aligned with personal interests and selecting training content tailored to individual needs. In addition, due to decreased reaction times and slower cognitive processing, serious games should incorporate clearer feedback mechanisms, appropriate scoring systems to enhance motivation, and adjustable difficulty levels with progressively structured game chapters to accommodate varying motor abilities.

Survivors of stroke in our study also identified several desirable features for VR-based rehabilitation systems, including family involvement, access to personal rehabilitation data, telerehabilitation support, safety monitoring, and technical assistance. Family engagement has long been shown to enhance motivation, emotional support, and adherence in stroke rehabilitation [38]. Participants highlighted that family support from family members is critical for psychological healing, yet complex emotional challenges are often overlooked despite their substantial influence on rehabilitation outcomes [39]. A VR gaming platform that allows family members to participate together can, while providing entertainment, strengthen the bonds between family members and even promote intergenerational communication. Additionally, progress tracking can enhance self-efficacy, support goal-setting, and promote active self-management [40]. In our study, many participants expressed strong enthusiasm about VR's potential to increase engagement, maintain motivation, and improve accessibility, particularly for home-based rehabilitation. In a home setting, patients can independently complete structured therapeutic exercises through VR tutorials delivered via head-mounted displays and controllers [41,42]. Remote therapist supervision can be supported through real-time data transmission, video consultations, or automated progress reports generated by the system [43]. This potential became particularly evident during the pandemic, when remote technologies played an essential role in maintaining continuity of care. Built-in safety features, such as heart rate and blood pressure monitoring, usage timers, and fall detection alarms, can further enhance user motivation and reduce risks during home use [44]. Finally, reliable technical support is crucial for enhancing patient trust and ensuring long-term adherence.

The successful integration of VR technology into stroke rehabilitation, as well as equitable access to such innovations, requires not only technological advancement but also governmental support and a cultural shift within clinical practice toward embracing digital rehabilitation tools. Achieving this goal calls for interdisciplinary collaboration among rehabilitation clinicians, neuroscientists, engineers, designers, and patients to develop evidence-based systems and clinical protocols that ensure both effectiveness and usability [45]. Special attention

XSL•FO
RenderX

should be given to adapting VR interventions for diverse populations, particularly older adults with limited digital literacy and individuals living in resource-constrained environments. Critical challenges, including digital literacy, cost, security, and accessibility in rural areas [14,46], may be addressed through 2 complementary pathways: designing simple, intuitive systems supported by clinical validation, and reducing production costs through governmental funding while implementing pilot programs in community and rural settings. Additionally, VR systems must account for the evolving priorities of survivors of stroke throughout their recovery [47]. A phased, holistic intervention approach is needed to support the restoration of physical, psychological, and social functions. Using Maslow's hierarchy of needs as a conceptual framework [48,49], VR system design should adapt its focus at different recovery stages. In the early phase, rehabilitation focuses on stabilization, preventing complications, and gently initiating motor function while offering psychological support. In the subacute phase, it shifts to intensive task-specific training and cognitive and motivational interventions. In the long-term community or home setting, rehabilitation emphasizes daily-life reintegration, supported by family involvement, peer support, and continued physical and psychosocial training.

Strengths and Limitations

The strengths of this study include allowing participants to experience VR rehabilitation equipment before each interview, which enhanced their understanding and enabled them to provide more informed feedback. Additionally, we implemented rigorous qualitative methods appropriate for an exploratory study, allowing patients to express their rehabilitation experiences, unmet needs, and perspectives on VR freely and comprehensively. However, several limitations should be acknowledged. First, the sample was drawn from a single rehabilitation hospital and was predominantly male, which may limit generalizability. Second, all interviews were conducted in Chinese and subsequently translated into English, which may have introduced subtle linguistic or interpretive biases. Third, most participants were within 6 months of stroke onset, which may restrict insights into long-term or home-based rehabilitation experiences. Fourth, the relatively short interview duration and insufficient depth in probing questions may have constrained the richness of the data. Finally, the absence of longitudinal follow-up prevented examination of how patients' needs, self-management behaviors, and acceptance of VR may change over time. Future research should incorporate multicenter longitudinal studies with more diverse samples to explore how patients' needs and attitudes evolve across recovery stages.

Conclusions

This study explored the rehabilitation experiences and unmet needs of survivors of stroke, as well as their perspectives of VR-based rehabilitation. Survivors of stroke undergo significant changes in their family and social roles due to the sudden loss of physical abilities. Their strong desire for recovery stands in stark contrast to the slow progress they experience in reality, which causes them immense psychological stress. Traditional rehabilitation is described as arduous and monotonous, with little to no support for home-based rehabilitation. Consequently, stroke rehabilitation services in China require improvements in the attractiveness of rehabilitation content, patient self-management, and continuity of care. Although patients desire better rehabilitation approaches and outcomes, the effective integration of VR technology into stroke care must take into account factors, such as personal characteristics and preferences, as well as socioeconomic status. Future VR rehabilitation systems should incorporate serious games that address diverse personal and training needs and integrate functions, such as family involvement, progress tracking, telemedicine support, safety protection, and technical assistance. These findings deepen understanding of how VR technologies can be meaningfully embedded into stroke rehabilitation and offer a roadmap for developing patient-centered and culturally responsive VR solutions.

Acknowledgments

The authors gratefully acknowledge the support of all the participants for sharing their experiences and perspectives in this study.

Funding

This study was sponsored by Beijing Nova Program (20240484500). The funders had no role in the study design, data analysis or interpretation, manuscript drafting, or the decision to submit for publication.

Data Availability

The data supporting this study are available upon reasonable request and subject to approval by the corresponding authors. Requests should be directed to SH and FL.

Authors' Contributions

Conceptualization: SH, FL, XZ
Data curation: XZ, LX.
Formal analysis: XZ, LX
Funding acquisition: SH, FL
Investigation: XZ, LX, HL

XSL•FO
RenderX

Methodology: SH, FL, XZ
Project administration: SH, FL, XZ
Supervision: SH, FL
Validation: XZ
Writing – original draft: XZ
Writing – review & editing: XZ, FL
The co-corresponding author, SH, can be contacted at the School of Architecture and Urban Planning, Beijing University of Civil Engineering and Architecture, No. 1 Zhanlanguan Road, Xicheng District, Beijing 100044, China (Email: haoshimeng@bucea.edu.cn).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sample Interview Guide.
[DOCX File , 16 KB - jmir_v28i1e84532_app1.docx]

Multimedia Appendix 2
COREQ checklist.
[DOCX File , 206 KB - jmir_v28i1e84532_app2.docx]

Multimedia Appendix 3
Summary of Themes and Representative Quotations.
[DOCX File , 23 KB - jmir_v28i1e84532_app3.docx]

References

1. GBD 2023 DiseaseInjuryRisk Factor Collaborators. Burden of 375 diseases and injuries, risk-attributable burden of 88 risk factors, and healthy life expectancy in 204 countries and territories, including 660 subnational locations, 1990-2023: a systematic analysis for the Global Burden of Disease Study 2023. Lancet 2025;406(10513):1873-1922 [FREE Full text] [doi: 10.1016/S0140-6736(25)01637-X] [Medline: 41092926]
2. GBD 2021 Stroke Risk Factor Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. The Lancet Neurology 2024;23(10):973-1003 [FREE Full text] [doi: 10.1016/S1474-4422(24)00369-7] [Medline: 39304265]
3. Markus HS. Reducing disability after stroke. International journal of stroke 2022;17(3):249-250. [doi: 10.1177/17474930221080904] [Medline: 35191348]
4. Pinter D, Fandler-Höfler S, Fruhwirth V, Berger L, Bachmaier G, Horner S, et al. Relevance of cognition and emotion for patient-reported quality of life after stroke in working age: an observational cohort study. Front Neurol 2022;13:869550 [FREE Full text] [doi: 10.3389/fneur.2022.869550] [Medline: 35547373]
5. Liu L. Chinese Stroke Association. Clinical Management Guidelines for Cerebrovascular Diseases in China. 2 ed. In: The People's Health Press. Jan Swasthya Abhiyan: People's Health Press; 2020.
6. Stinear CM, Lang CE, Zeiler S, Byblow WD. Advances and challenges in stroke rehabilitation. The Lancet Neurology 2020;19(4):348-360. [doi: 10.1016/S1474-4422(19)30415-6] [Medline: 32004440]
7. Lu W, Shi M, Liu L, Wang S, Deng W, Ma Y, et al. Effect of virtual reality-based therapies on lower limb functional recovery in stroke survivors: systematic review and meta-analysis. J Med Internet Res 2025;27:e72364 [FREE Full text] [doi: 10.2196/72364] [Medline: 40737641]
8. Dawson J, Nee R, Ramirez C, Reyes S, Sanchez D, Sukhadia T, et al. Gamification in mHealth apps for rehabilitation: protocol for a scoping review. JMIR Res Protoc 2025;14:e63600 [FREE Full text] [doi: 10.2196/63600] [Medline: 40294403]
9. Verma A, Towfighi A, Brown A, Abhat A, Casillas A. Moving towards equity with digital health innovations for stroke care. Stroke 2022;53(3):689-697 [FREE Full text] [doi: 10.1161/STROKEAHA.121.035307] [Medline: 35124973]
10. Nikolaev VA, Nikolaev AA. Recent trends in telerehabilitation of stroke patients: A narrative review. NeuroRehabilitation 2022;51(1):1-22. [doi: 10.3233/NRE-210330] [Medline: 35527574]
11. Chen J, Or CK, Chen T. Effectiveness of using virtual reality-supported exercise therapy for upper extremity motor rehabilitation in patients with stroke: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2022;24(6):e24111 [FREE Full text] [doi: 10.2196/24111] [Medline: 35723907]
12. Laver KE, Lange B, George S, Deutsch JE, Saposnik G, Chapman M, et al. Virtual reality for stroke rehabilitation. Cochrane Database Syst Rev 2025;6(6):CD008349. [doi: 10.1002/14651858.CD008349.pub5] [Medline: 40537150]

XSL•FO
RenderX

13. Cucinella SL, de Winter JCF, Grauwmeijer E, Evers M, Marchal-Crespo L. Towards personalized immersive virtual reality neurorehabilitation: a human-centered design. J Neuroeng Rehabil 2025;22(1):7 [FREE Full text] [doi: 10.1186/s12984-024-01489-5] [Medline: 39833912]
14. National Institute for Health and Care Excellence: Clinical Guidelines. In: Stroke rehabilitation in adults. London: National Institute for Health and Care; 2023.
15. Ghosh R, Khan N, Migovich M, Tate JA, Maxwell CA, Newhouse PA, et al. Engaging older adults and staff in the Co-Design and evaluation of socially assistive robot and virtual reality activities for Long-Term Care: user-centered study. JMIR Aging 2025;8:e75288 [FREE Full text] [doi: 10.2196/75288] [Medline: 41330571]
16. PLoS Medicine Editors T. Qualitative research: understanding patients' needs and experiences. PLoS Med 2007;4(8):e258 [FREE Full text] [doi: 10.1371/journal.pmed.0040258] [Medline: 17760496]
17. Sundler AJ, Lindberg E, Nilsson C, Palmér L. Qualitative thematic analysis based on descriptive phenomenology. Nurs Open 2019;6(3):733-739 [FREE Full text] [doi: 10.1002/nop2.275] [Medline: 31367394]
18. Bradshaw C, Atkinson S, Doody O. Employing a qualitative description approach in health care research. Glob Qual Nurs Res 2017;4:2333393617742282 [FREE Full text] [doi: 10.1177/2333393617742282] [Medline: 29204457]
19. Fontanella BJB, Ricas J, Turato ER. [Saturation sampling in qualitative health research: theoretical contributions]. Cad Saude Publica 2008;24(1):17-27 [FREE Full text] [doi: 10.1590/s0102-311x2008000100003] [Medline: 18209831]
20. Tu WJ, Wang LD. China stroke surveillance report 2021. Military Medical Research 2023;10(1):33 [FREE Full text] [doi: 10.1186/s40779-023-00463-x] [Medline: 37468952]
21. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006;56(8):77-101. [doi: 10.1007/s10597-020-00591-x] [Medline: 32100154]
22. Braun V, Clarke V. Successful Qualitative Research. London: SAGE Publications; 2013.
23. Levitt HM, Bamberg M, Creswell JW, Frost DM, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA Publications and Communications Board task force report. Am Psychol 2018;73(1):26-46 [FREE Full text] [doi: 10.1037/amp0000151] [Medline: 29345485]
24. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
25. Guba EG, Lincoln Y. Competing paradigms in qualitative research. Handbook of qualitative research 1994;2:105. [doi: 10.2307/3121684]
26. Sun X, Shi Y, Liu C, Wang S, Li D, Zhu X, et al. Experience and needs of stroke patients in physical rehabilitation: a systematic review and meta-synthesis. BMC Health Serv Res 2025;25(1):1062. [doi: 10.1186/s12913-025-13213-7] [Medline: 40790477]
27. Hui KK, Sporko TN, Vangel MG, Li M, Fang J, Lao L. Perception of deqi by chinese and american acupuncturists: a pilot survey. Chin Med 2011;6(1):2 [FREE Full text] [doi: 10.1186/1749-8546-6-2] [Medline: 21251312]
28. Liang Y, Lu P. Medical insurance policy organized by Chinese government and the health inequity of the elderly: longitudinal comparison based on effect of New Cooperative Medical Scheme on health of rural elderly in 22 provinces and cities. Int J Equity Health 2014;13:37 [FREE Full text] [doi: 10.1186/1475-9276-13-37] [Medline: 24884944]
29. Ham Y, Shin J. Efficiency and usability of a modified pegboard incorporating computerized technology for upper limb rehabilitation in patients with stroke. Top Stroke Rehabil 2023;30(4):333-341. [doi: 10.1080/10749357.2022.2058293] [Medline: 35348031]
30. Lin S, Wang C, Wang Q, Xie S, Tu Q, Zhang H, et al. The experience of stroke survivors and caregivers during hospital-to-home transitional care: A qualitative longitudinal study. Int J Nurs Stud 2022;130:104213. [doi: 10.1016/j.ijnurstu.2022.104213] [Medline: 35378465]
31. Anwar N, Karimi H, Ahmad A, Gilani SA, Khalid K, Aslam AS, et al. Virtual reality training using nintendo wii games for patients with stroke: randomized controlled trial. JMIR Serious Games 2022;10(2):e29830 [FREE Full text] [doi: 10.2196/29830] [Medline: 35699989]
32. Everard G, Declerck L, Detrembleur C, Leonard S, Bower G, Dehem S. New technologies promoting active upper limb rehabilitation after stroke: an overview and network meta-analysis. Eur J Phys Rehabil Med 2022;58(4):530-548. [doi: 10.23736/s1973-9087.22.07404-4]
33. Chen J, Or CK, Li Z, Yeung EHK, Chen T. Perceptions of patients with stroke regarding an immersive virtual reality-based exercise system for upper limb rehabilitation: questionnaire and interview study. JMIR Serious Games 2025;13:e49847 [FREE Full text] [doi: 10.2196/49847] [Medline: 39742513]
34. Bally ELS, Cheng D, van Grieken A, Ferri Sanz M, Zanutto O, Carroll A, et al. Patients' perspectives regarding digital health technology to support self-management and improve integrated stroke care: qualitative interview study. J Med Internet Res 2023;25:e42556 [FREE Full text] [doi: 10.2196/42556] [Medline: 37014677]
35. Nedergård H, Sandlund M, Häger CK, Palmcrantz S. Users' experiences of intensive robotic-assisted gait training post-stroke - "a push forward or feeling pushed around?". Disabil Rehabil 2023;45(23):3861-3868 [FREE Full text] [doi: 10.1080/09638288.2022.2140848] [Medline: 36342771]

XSL•FO
RenderX

36. de Veer AJE, Peeters JM, Brabers AEM, Schellevis FG, Rademakers JJDJM, Francke AL. Determinants of the intention to use e-Health by community dwelling older people. BMC Health Serv Res 2015;15:103 [FREE Full text] [doi: 10.1186/s12913-015-0765-8] [Medline: 25889884]
37. Vieira C, Ferreira da Silva Pais-Vieira C, Novais J, Perrotta A. Serious game design and clinical improvement in physical rehabilitation: systematic review. JMIR Serious Games 2021;9(3):e20066 [FREE Full text] [doi: 10.2196/20066] [Medline: 34554102]
38. Eriksson G, Söderhielm K, Erneby M, Guidetti S. Family members' experiences of a person-centered information and communication technology-supported intervention for stroke rehabilitation (F@ce 2.0): qualitative analysis. JMIR Rehabil Assist Technol 2025;12:e69878 [FREE Full text] [doi: 10.2196/69878] [Medline: 40315427]
39. Kim JS. Post-stroke mood and emotional disturbances: pharmacological therapy based on mechanisms. J Stroke 2016;18(3):244-255 [FREE Full text] [doi: 10.5853/jos.2016.01144] [Medline: 27733031]
40. Chen Y, Li K, Lin C, Hung P, Lai H, Wu C. The effect of sequential combination of mirror therapy and robot-assisted therapy on motor function, daily function, and self-efficacy after stroke. Sci Rep 2023;13(1):16841 [FREE Full text] [doi: 10.1038/s41598-023-43981-3] [Medline: 37803096]
41. Toh SFM, Fong KNK, Gonzalez PC, Tang YM. Application of home-based wearable technologies in physical rehabilitation for stroke: a scoping review. IEEE Trans. Neural Syst. Rehabil. Eng 2023;31:1614-1623. [doi: 10.1109/tnsre.2023.3252880]
42. Laver KE, Adey-Wakeling Z, Crotty M, Lannin NA, George S, Sherrington C. Telerehabilitation services for stroke. Cochrane Database Syst Rev 2020;1(1):CD010255 [FREE Full text] [doi: 10.1002/14651858.CD010255.pub3] [Medline: 32002991]
43. Benadduci M, Franceschetti C, Marziali RA, Frese S, Sándor PS, Tombolesi V, et al. An integrated virtual reality-based telerehabilitation platform to support recovery and maintenance of functional abilities among older adults: protocol for a usability and acceptability study. JMIR Res Protoc 2025;14:e68358 [FREE Full text] [doi: 10.2196/68358] [Medline: 40729692]
44. Bargeri S, Scalea S, Agosta F, Banfi G, Corbetta D, Filippi M, et al. Effectiveness and safety of virtual reality rehabilitation after stroke: an overview of systematic reviews. EClinicalMedicine 2023;64:102220 [FREE Full text] [doi: 10.1016/j.eclinm.2023.102220] [Medline: 37745019]
45. Wankhede NL, Koppula S, Ballal S, Doshi H, Kumawat R, Raju S, et al. Virtual reality modulating dynamics of neuroplasticity: innovations in neuro-motor rehabilitation. Neuroscience 2025;566:97-111. [doi: 10.1016/j.neuroscience.2024.12.040] [Medline: 39722287]
46. Terp R, Kayser L, Lindhardt T. Older patients' competence, preferences, and attitudes toward digital technology use: explorative study. JMIR Hum Factors 2021;8(2):e27005 [FREE Full text] [doi: 10.2196/27005] [Medline: 33988512]
47. Purton J, Sim J, Hunter SM. Stroke survivors' views on their priorities for upper-limb recovery and the availability of therapy services after stroke: a longitudinal, phenomenological study. Disabil Rehabil 2023;45(19):3059-3069 [FREE Full text] [doi: 10.1080/09638288.2022.2120097] [Medline: 36111388]
48. Carpenito-Moyet LJ. Maslow's Hierarchy of Needs--revisited. Nurs Forum 2003;38(2):3-4. [doi: 10.1111/j.1744-6198.2003.tb01204.x] [Medline: 12894625]
49. Alam Z. Making international trainees feel welcome: don't forget Maslow's hierarchy of needs. BMJ (Clinical research ed) 2023;382:2087. [doi: 10.1136/bmj.p2087] [Medline: 37699640]

Abbreviations

VR: virtual reality
COREQ: Consolidated Criteria for Reporting Qualitative Research

Edited by S Brini; submitted 21.Sep.2025; peer-reviewed by S Wang, L Bulle, A Kaunnil; comments to author 25.Nov.2025; revised version received 08.Jan.2026; accepted 09.Jan.2026; published 02.Feb.2026.

Please cite as:
Zheng X, Xing L, Lu H, Hao S, Liu F
Traditional Rehabilitation Experiences, Unmet Needs, and Perspectives on Virtual Reality–Based Rehabilitation Among Patients With Stroke in China: Qualitative Thematic Analysis and Semistructured Interview Study
J Med Internet Res 2026;28:e84532
URL: https://www.jmir.org/2026/1/e84532
doi:10.2196/84532
PMID:

XSL•FO
RenderX

©Xite Zheng, Lu Xing, Haitao Lu, Shimeng Hao, Fen Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of the Mobile-Based Diabetes Little Helper Video Intervention on Medication Adherence Among Older Adults Living With Type 2 Diabetes Mellitus, Henan, China: Randomized Controlled Trial

Meng Wang[1], PhD; Khuan Lee[1], PhD; Hui Zhu Thew[2], MMed; Siti Noorkhairina Sowtali[3], PhD; Qiuhuan Jiang[4], MPH; Zhengyan Zhao[5], BSN; Yu Liu[6], MNS; Poh Ying Lim[7], PhD

[1]Department of Nursing, Faculty of Medicine & Health Sciences, Universiti Putra Malaysia, Selangor, Malaysia

[2]Department of Family Medicine, Faculty of Medicine & Health Sciences, Universiti Putra Malaysia, Selangor, Malaysia

[3]Department of Professional Nursing Studies, Kulliyyah of Nursing, International Islamic University Malaysia, Pahang, Malaysia

[4]Henan Provincial People's Hospital, Zhengzhou, China

[5]Zhengzhou Seventh People's Hospital, Zhengzhou, China

[6]The Second Mobile Contingent Hospital of the Chinese People's Armed Police Forces, Wuxi, China

[7]Department of Community Health, Faculty of Medicine & Health Sciences, Universiti Putra Malaysia, Selangor, Serdang, Malaysia

Corresponding Author:
Poh Ying Lim, PhD
Department of Community Health, Faculty of Medicine & Health Sciences
Universiti Putra Malaysia
Jalan Universiti 1
Selangor, Serdang, 43400
Malaysia
Phone: 60 397692950
Email: pohying_my@upm.edu.my

Abstract

Background: Medication adherence is vital for older adults living with type 2 diabetes mellitus (T2DM), but it remains low and needs improvement. Current interventions have limited effectiveness, while video-based interventions show promising potential for enhancing adherence.

Objective: To evaluate the impact of the “Diabetes Little Helper” video intervention, developed based on the information–motivation–behavioral skills model, on improving medication adherence in older adults living with T2DM in Henan.

Methods: This parallel-group randomized controlled trial was conducted in 2 hospitals in Zhengzhou, involving 68 patients from each hospital. The intervention group (IG) received standard care plus the video intervention for one month, while the control group (CG) received only standard care. The primary outcome was medication adherence, and secondary outcomes included medication knowledge, attitude, behavior, belief, and social support. Data were collected at baseline, postintervention, and at 3-month follow-up. Intention-to-treat analysis and the last observation carried forward method were applied for missing data, with the generalized estimating equation model used for effect assessment (P<.05 considered statistically significant).

Results: The average age of participants in the IG was 67.5 (SD 8.0) years, while in the CG, the average age was 66.0 (SD 7.0) years. Sex distribution was nearly identical, with 51.5% (n=35) of participants in the IG and 50.0% (n=34) in the CG being male. After the intervention, the IG and CG had retention rates of 95.6% (n=65) and 97.1% (n=66), respectively. At the 3-month follow-up, the retention rates for the IG and CG were 92.6% (n=63) and 92.2% (n=62), respectively. Both postintervention (β=4.956, 95% CI 3.702-6.210, P<.001) and at the 3-month follow-up (β=3.691, 95% CI 2.379-5.003, P<.001), medication adherence score for the IG was significantly higher than that for the CG. In addition, the IG showed significant improvements in secondary outcome, with medication knowledge (β=11.592, 95% CI 6.923-16.260, P<.001), attitude (β=5.467, 95% CI 4.531-6.763, P<.001), behavior (β=4.176, 95% CI 3.220-5.133, P<.001), and belief (β=2.882, 95% CI 1.990-3.775, P<.001) compared with the CG postintervention. However, there was no statistically significant difference in the secondary outcome of social support (β=0.008, 95% CI –1.834 to 2.011, P=.928).

XSL•FO
RenderX

Conclusions: The Diabetes Little Helper video intervention effectively improved medication adherence in older adults living with T2DM in Henan, highlighting the potential of digital health tools for managing chronic diseases in older adult populations.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2400083282; https://www.chictr.org.cn/showprojEN.html?proj=214847

(J Med Internet Res 2026;28:e78731) doi:10.2196/78731

KEYWORDS

older adults; video; type 2 diabetes mellitus; medication adherence; randomized controlled trial

Introduction

With the continuous improvement of economic levels, changes in lifestyle, and the accelerated process of population aging, the prevalence of diabetes mellitus (DM) has been steadily increasing, becoming a significant global public health challenge [1,2]. In 2021, the number of people aged 20-79 years with DM worldwide was approximately 536.6 million, and it is projected to rise to 783.2 million by 2045 [3]. The older adults, due to factors such as physiological decline, metabolic imbalance, and the coexistence of chronic diseases, represent a high-risk group for DM. In 2019, the number of people aged 65 years and older with DM worldwide reached 135.6 million, and it is expected to increase to 276.2 million by 2045 [4].

As one of the countries with the highest burden of DM, the trend of DM prevalence in China is consistent with the global pattern. The number of people aged 20-79 years with DM in China increased from 88.7 million in 2020 to over 108.4 million [5]. Among these, type 2 diabetes mellitus (T2DM) is the predominant form, accounting for 90%-95% of all DM cases, with the older adults being the primary affected group [6-9]. T2DM significantly shortens the lifespan of patients and can lead to severe complications, including cardiovascular diseases, kidney disease, retinopathy, and neuropathy, potentially resulting in organ failure [10-15]. Due to the long duration of the disease and the presence of multiple complications, older adults living with T2DM are at higher risk of developing complications such as diabetic foot, stroke, and nephropathy [16,17]. Medication treatment is crucial in the management of T2DM, and its effectiveness is highly dependent on patient adherence to medication, that is, whether patients follow the prescribed treatment regimen consistently [18,19]. Studies have shown that patients with high medication adherence not only experience more stable disease control but also significantly reduce mortality, hospitalization rates, and health care costs [20-22]. However, in China, more than half of older adults living with T2DM experience poor medication adherence [23,24], which severely limits the long-term effectiveness of DM management and urgently calls for effective intervention strategies. Research indicates that various factors contribute to poor medication adherence, including age [25], economic burden [26], and multiple medications [27]. To improve medication adherence among older adults, various intervention strategies have been attempted both domestically and internationally, including health management based on multidisciplinary collaborative models [28], face-to-face health education [29], mobile app support [30], and motivational interviewing [31]. However, these approaches generally face challenges such as high implementation costs, complex operations, and a strong reliance on health care resources. Furthermore, most lack systematic theoretical models, making them difficult to adapt to older adults with lower educational levels [32,33]. The information–motivation–behavioral skills (IMB) model provides a theoretical foundation for explaining and predicting individual health behaviors and has been widely applied in chronic disease self-management research [34,35]. Existing studies have confirmed that the IMB model can effectively predict medication adherence among older adults living with DM [36,37], offering scientific guidance for the development of intervention strategies. In recent years, video intervention has emerged as a new method of health communication. Existing studies have shown that videos not only effectively capture attention but also facilitate the understanding and retention of information [38]. Particularly within older adult populations, multimedia interventions have been demonstrated to overcome the limitations of traditional educational methods, improving the acquisition of health knowledge and enhancing patients' self-management capabilities [39,40]. Videos, delivered through convenient devices such as mobile phones, enable patients to view them anytime and anywhere, further reinforcing learning outcomes and improving medication adherence [41]. However, there is currently no research using video interventions to improve medication adherence among older adults living with T2DM. Therefore, exploring the application of IMB-based video interventions in this population and their potential to enhance medication adherence holds significant theoretical and practical importance. Based on this, we hypothesize that the Diabetes Little Helper video intervention will significantly improve medication adherence compared with standard care.

Methods

Ethical Considerations

This study received ethical approval from the Institutional Review Board of Universiti Putra Malaysia (JKEUPM-2023-1279) and Henan Provincial People's Hospital (20240107), and was registered with the Chinese Clinical Trial Registry (ChiCTR2400083282). Before data collection, the researcher obtained approval from the hospital and coordinated with relevant departments to ensure informed consent from participants and guide them to scan the QR code to follow the "Diabetes Little Helper" WeChat public account. All participant information was kept strictly confidential, and the questionnaire data were anonymized after collection to ensure that participants' identities could not be directly or indirectly identified. This study does not involve any monetary compensation or incentives for participants.

XSL•FO
RenderX

Study Design and Setting

This study was a parallel-group, cluster-randomized controlled trial (RCT) recruited older adults from 2 hospitals in Zhengzhou, Henan Province, China (Henan Provincial People's Hospital and Zhengzhou Central Hospital). Both hospitals are large, tertiary-level public general hospitals with an annual outpatient volume exceeding 1.1 million visits. Each hospital has over 70 clinical departments, including endocrinology, respiratory medicine, and cardiology, offering comprehensive medical services and a sufficient patient base for recruitment. The study spanned from April 2024 to June 2025, encompassing participant recruitment (from April to June 2024), intervention delivery, data collection, analysis, and manuscript preparation. The study was conducted and reported in accordance with the CONSORT (Consolidated Standards of Reporting Trials; checklist provided in Multimedia Appendix 1) guidelines. Further details of the study design and methodology are available in the published protocol in *BMJ Open* [42]. This study followed all specified procedures without any deviations from the protocol.

Participants

This study involved older adults diagnosed with T2DM who attended endocrinology outpatient clinics at 2 selected hospitals. Eligible individuals were identified from the outpatient population of each hospital and enrolled through simple random sampling. Participants must meet the following inclusion criteria: aged 60 years or older, diagnosed with T2DM [43], Chinese nationality, and able to understand Chinese, currently receiving at least one oral antidiabetic medication with a hemoglobin HbA_{1c} level of ≥7.0%, no cognitive impairment, and either the participant or their family member owns a smartphone and is able to use WeChat. Exclusion criteria include having severe comorbidities affecting major organs such as the heart, lungs, or kidneys, impairment in communication abilities, or participating in another study.

This cluster RCT study involves 2 hospitals, which were randomly assigned in a 1:1 allocation ratio to the control group (CG) and intervention group (IG) using a random number generator. Both hospitals used simple random sampling and a parallel design, recruiting 68 older adults living with T2DM. Randomization was conducted by the research coordinator using a computer-generated random number sequence, with concealed allocation to prevent selection bias. The CG received standard nursing care, while the IG received additional video intervention. Due to the visible nature of the video intervention, the study was conducted as an open-label trial, meaning blinding was not feasible for the participants or executors. However, evaluators and statistical analysts remained blind to minimize bias. To minimize contamination, participants were selected from separate hospitals, and communication between the groups was restricted. After a 3-month follow-up, the CG received the same video intervention as the IG to ensure equal treatment.

The sample size for this study was calculated using the 2-mean sample size formula [44], based on the mean and SD of medication adherence for both CG and IG at postintervention and 3-month follow-up (29.15, SD 3.10; 31.27, SD 3.35; and 26.25, SD 2.01; 30.19, SD 2.41) [45,46]. Assuming a 2-sided test with a type I error rate (α) of 0.05, a power (1-β) of 80%, an 80% retention rate, and a design effect of 1.5, the required total sample size was calculated to be 68 participants per group.

Diabetes Little Helper Intervention Development

A health education intervention program based on the IMB theory model was developed for this study. The research team conducted 3 rounds of expert consensus meetings involving a multidisciplinary expert panel, which included 2 nursing experts with extensive clinical experience and expertise in video production, a biostatistician with a background in statistics for health interventions, and a family medicine specialist with expertise in DM management. The team collaboratively established a comprehensive educational framework covering 8 core topics related to DM medication management. These topics include proper medication storage and transportation, methods for identifying medication quality, disease management strategies, the importance of medication adherence, key medication precautions, common misconceptions about medications, self-management empowerment for older adults living with DM, and guidelines for blood glucose monitoring. For each topic, a 2-5 minute educational animation was developed using the "WanCai Animation Master" tool to visualize professional medical knowledge into easy-to-understand dynamic images [47]. Table 1 shows the Diabetes Little Helper video intervention topic and plan.

Table 1. Topics and intervention plan of the Diabetes Little Helper video program for older adults living with type 2 diabetes mellitus in Henan, China.

Number	Week	Day of the week	IMB[a]	Topic	Video time
1	One	Tuesday	I	Storage & Traveling with Hypoglycemic Medications	3 min 31 s
2	One	Friday	I	How to Check Your Medications	3 min 20 s
3	Two	Tuesday	I	What to Do During Sick Days	2 min 57 s
4	Two	Friday	M	Importance of On-Time Medication	2 min 15 s
5	Three	Tuesday	M	Precautions for Taking Hypoglycemic Medication	1 min 53 s
6	Three	Friday	M	Correct Understanding of Hypoglycemic Medications	2 min
7	Four	Tuesday	B	Self-Empowerment & Support for Older Adults Living with DM	4 min 30 s
8	Four	Friday	B	Regular Blood Glucose Monitoring	3 min 6 s

[a]IMB: information–motivation–behavioral skills.

XSL•FO
RenderX

Five nursing experts were invited to assess the content validity of the program based on the standards developed by Siti Noorkhairina et al [48]. These standards were specifically designed for evaluating video-based educational materials and cover various aspects of content quality, including media characteristics (title relevance and audio clarity), educational design (plot logic and age-appropriate language), and technical specifications (appropriate duration and information density). The Siti Noorkhairina et al standards provide a comprehensive framework for assessing the effectiveness of educational videos, ensuring that the content is clear, relevant, and engaging for the target audience. Experts rated 24 evaluation items based on these criteria to determine the content validity of the program. A 4-point Likert scale was used for the ratings. The Item-level Content Validity Index (I-CVI) for each of the 24 items was calculated by determining the proportion of experts who rated the item as relevant, with scores of 3 or 4 indicating relevance. For calculation, scores of 3 and 4 were recoded as "1.00" (indicating relevant), and scores of 1 and 2 were recoded as "0" (indicating non-relevant). The I-CVI for each item was then calculated as the proportion of experts who rated the item as "relevant". A value of 1.00 for the I-CVI indicates unanimous agreement among the experts that each item was deemed relevant. The Scale-level Content Validity Index (S-CVI) was calculated by averaging the I-CVI scores across all items. A value of 1.00 for the S-CVI indicates that the overall scale was deemed fully valid by all experts. These results suggest that the content of the program was considered highly relevant and appropriate by all expert raters. Ten older adults living with T2DM were invited to participate in a face validity test, which focused on information reception indicators such as language comprehensibility, visual appeal, and pacing control. Based on patient feedback, the audiovisual elements of the videos were optimized.

Diabetes Little Helper Intervention Implementation

The intervention was implemented using a "standard care + digital health education" model, delivering content through the WeChat public account "Diabetes Little Helper." During the 4-week intervention period, educational videos were scheduled to be released at 8 AM every Tuesday and Friday, with a total of 8 episodes. Patients could watch the videos at their convenience.

To ensure that patients successfully watched the intervention videos and to improve intervention quality, the research team took the following measures: (1) real-time monitoring of video viewership to track whether patients watched the videos on time, ensuring timely completion and optimizing the intervention strategy, (2) weekly WeChat reminders for patients to watch the new videos, along with contact information for researchers to provide immediate assistance, ensuring patients received the necessary information and encouraging participation, (3) weekly phone follow-ups to check patients' video-watching progress, particularly focusing on those who missed videos, researchers would inquire about reasons for nonviewing and provide personalized help to ensure patient involvement, (4) personalized support for patients with special circumstances or frequent video nonviewing, including simple viewing guides and remote technical support to resolve technical issues, ensuring patients could watch the videos without difficulty.

Development of a Questionnaire to Evaluate Medication Adherence, Knowledge, Attitude, Behavior, Belief, and Social Support

A self-developed questionnaire was used to collect data from participants, covering 8 aspects. A detailed overview of all outcome measures used in this study can be found in the study protocol [42]. The measures analyzed in this manuscript are described below.

Sociodemographic and Disease-Related Characteristics

This study collected sociodemographic characteristics (such as age, sex, and marital status) as well as disease-related variables (such as the duration of T2DM, family history, and adverse medication reaction).

Medication Adherence

Medication adherence is assessed using the Self-Efficacy for Appropriate Medication Use Scale [49], which includes 13 items, rated on a 3-point Likert scale, with a total score range of 13 to 39. Higher scores indicate better medication adherence. The Cronbach α value of the scale is 0.934 [49,50].

Medication Knowledge, Attitude, and Behavior

The medication knowledge, attitude, and behavior questionnaires are used to measure patients' knowledge, attitudes, and behaviors regarding medication. The test-retest reliability for each dimension ranges from 0.854 to 0.920, with an overall test-retest reliability of 0.868 [51]. The knowledge questionnaire consists of 17 items, with a total score range of 0 to 17, where higher scores indicate a greater level of medication knowledge. The medication attitude questionnaire includes 10 items, with a total score range of 10 to 50, where higher scores reflect a more positive attitude toward safe medication use. The medication behavior questionnaire consists of 13 items, with scores ranging from 9 to 40, where higher scores indicate safer medication behaviors.

Medication Belief

The medication belief questionnaire is used to assess the personal medication beliefs of older adults living with DM [52]. The questionnaire consists of 10 items, divided into 2 dimensions, medication necessity and medication concern, with scores ranging from 1 to 5. Higher scores indicate stronger medication beliefs. The Cronbach α value for the questionnaire is 0.724, with Cronbach α values of 0.749 and 0.796 for the 2 dimensions, respectively.

Social Support

The social support questionnaire is used to assess the support patients receive across multiple dimensions [53]. This scale includes 4 dimensions, with a total of 19 items, and the total score ranges from 19 to 95. Higher scores indicate higher levels of social support among the older adults. The questionnaire demonstrates good reliability and validity, with a Cronbach α value of 0.803 and a content validity index of 0.958 [53].

XSL•FO
RenderX

To ensure the content validity of the questionnaire assessing medication adherence, knowledge, attitude, behavior, belief, and social support, 5 experts reviewed it to confirm that the content aligned with the research objectives and that the questions were relevant and representative of the study. Based on expert feedback, some questions were adjusted to ensure they accurately measured the intended content. The I-CVI and the S-CVI of the questionnaire were both calculated to be 1, indicating strong content validity. Additionally, to assess the face validity of the questionnaire, it was presented to 10 older adults living with T2DM, who provided feedback on the clarity and appropriateness of the language used in the questions. Based on their feedback, the wording of some questions was revised for better clarity and suitability. The revised questionnaire was found to be relevant, user-friendly, and suitable for this population.

Data Collection

Data collection occurred at 3 time points: baseline, postintervention, and 3-month follow-up. The questionnaire was available in both paper and electronic formats. Baseline data were collected between April and June 2024. Researchers conducted on-site checks to ensure the quality of questionnaire completion, confirming that all questions were fully answered and met the required standards. For older adult participants who had difficulty filling out or understanding the questions, standardized instructions were provided to help them interpret the content and complete the questionnaire correctly. During the postintervention and follow-up periods, the electronic questionnaire was sent to participants via a WeChat official account, with researchers assisting through WeChat or phone calls to ensure a high response rate and data quality. After collecting the questionnaires, researchers reviewed and processed any inconsistent or missing data to ensure completeness. Data were entered by 2 research assistants and imported into SPSS (version 29.0; IBM Corp) software for analysis. To reduce entry errors, all data were proofread twice after input to ensure consistency with the original questionnaires. Data from the electronic questionnaires were directly imported into SPSS.

Data Analysis

Data analysis was performed using SPSS Statistics, following the intention-to-treat principle. Missing data were handled using the last observation carried forward (LOCF) method. A significance level was set at $P<.05$. Histograms with normal curves were used to assess the normality of continuous variables. For normally distributed variables, data are expressed as mean (SD), while not normally distributed variables are presented as median and IQR. Categorical variables are reported as frequencies and percentages (%).

For baseline group comparisons, the Pearson chi-square test and chi-square test with continuity correction were applied for categorical variables. Independent samples *t* test and Mann-Whitney *U* test were used for continuous variables. To assess the intervention effects, generalized estimating equation models with an unstructured working correlation matrix were used to analyze changes in medication adherence, knowledge, attitude, behavior, belief, and social support scores. The models compared the changes between the IG and CG at baseline, postintervention, and at the 3-month follow-up, with time, group, and their interaction as independent variables. The interaction term was used to assess whether there were different changes between the groups at different time points. Significant differences between the groups were adjusted for as covariates in the final model.

Results

Participant Flow

The IG consisted of 68 participants randomly selected from a pool of 110 eligible patients in one hospital, while the CG included 68 participants randomly selected from a pool of 100 eligible patients in the other hospital. Compared with the web-based questionnaire, 26 (29.0%) participants in the IG and 20 (38.0%) participants in the CG chose to complete the paper questionnaire. Both groups had an initial retention rate of 100% (n=68). After the intervention, 3 participants dropped out of the IG, leading to a retention rate of 95.6% (n=65). In the CG, one participant did not participate due to health issues, and another withdrew, resulting in a retention rate of 97.1% (n=66). Following a 3-month follow-up, 2 participants left the IG, giving a retention rate of 92.6% (n=63). In the CG, 3 participants withdrew, and one could not be contacted, leading to a retention rate of 92.2% (n=62). Figure 1 shows the data collection flow according to the CONSORT guidelines.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram for a cluster randomized controlled trial on the effectiveness of the Diabetes Little Helper intervention video among older adults living with type 2 diabetes mellitus.

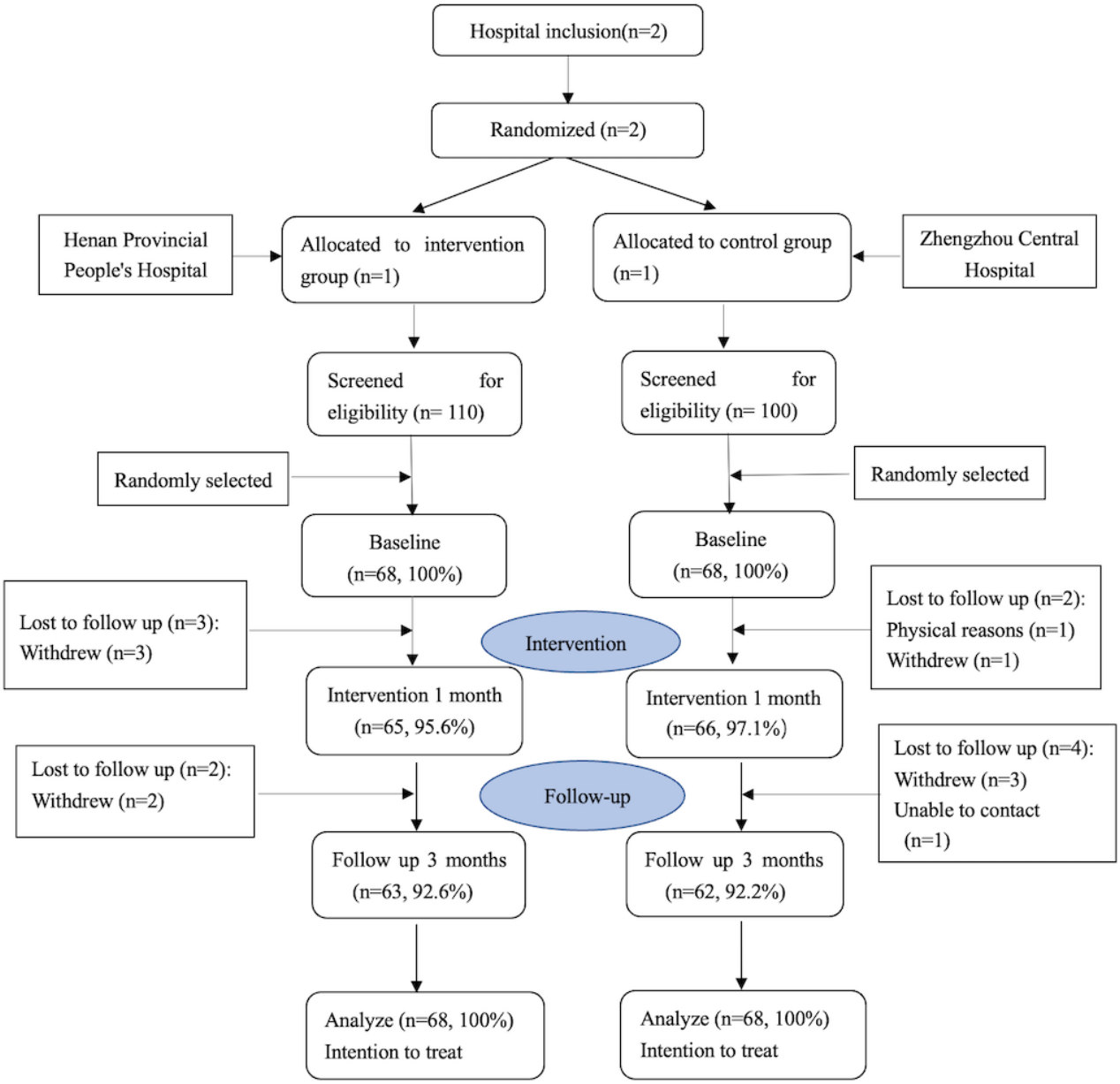

Participant Characteristics

Sociodemographic and Disease-Related Characteristics Between the IG and CG at Baseline

At baseline, a comparison of sociodemographic and disease-related characteristics between the IG and CG showed no statistically significant differences (P>.05). The sex distribution was nearly identical, with 51.5% (n=35) in the IG and 50.0% (n=34) in the CG being male. Most participants lived in urban areas (IG: 38/68, 55.9%; CG: 42/68, 61.8%) and resided with family members (IG: 62/68, 91.2%; CG: 60/68, 88.2%). A large portion of participants were married (IG: 50/68, 73.5%; CG: 52/68, 76.5%). All participants had children, and education levels were primarily at the junior high school (IG: 22/68, 32.4%; CG: 10/68, 14.7%) and high school (IG: 21/68, 30.9%; CG: 22/68, 32.4%) levels. Additionally, a significant portion of participants did not use smartphones independently (IG: 15/68, 22.1%; CG: 10/68, 14.7%). The mean and SD for BMI and the duration of T2DM were as follows: IG 24.04 (SD 3.475), CG 24.08 (SD 3.428); IG 14.76 (SD 8.381), CG 14.19 (SD 7.658). Most participants used oral medications (IG: 37/68, 54.4%; CG: 35/68, 53.3%), and the proportion of family members without diabetes was similar (IG: 26/68, 33.3%; CG: 23/68, 29.5%). The most common adverse reactions were hypoglycemia (IG: 21/68, 30.9%; CG: 16/68, 26.5%) and digestive system issues (IG: 13/68, 10.1%; CG: 10/68, 14.7%). Details are provided in Table 2.

XSL•FO
RenderX

Table 2. Comparison of sociodemographic characteristics between the intervention group (IG) and control group (CG) at baseline among older adults living with type 2 diabetes mellitus (T2DM) in Henan, China (N=136).

Variable	IG (n=68)	CG (n=68)	Chi-square (*df*)[a,b]	*t* test (*df*)[c]	Z[d]	*P* value
Age (years), median (IQR)	67.5 (63.3-71)	66.0 (69.8-63)	—[e]	—	–0.901[d]	.37
Sex, n (%)			0.029 (1)[a]	—	—	.86
Male	35 (51.5)	34 (50.0)				
Female	33 (48.5)	34 (50.0)				
Residence, n (%)			0.486 (1)[a]	—	—	.49
Countryside	30 (44.1)	26 (38.2)				
City	38 (55.9)	42 (61.8)				
Marital status, n (%)			2.039 (2)[a]	—	—	.36
Married	50 (73.5)	52 (76.5)				
Divorced	3 (4.4)	6 (8.8)				
Widowed	15 (22.1)	10 (14.7)				
Living arrangement, n (%)			0.319 (1)[a]	—	—	.57
Living alone	6 (8.8)	8 (11.8)				
Living with family	62 (91.2)	60 (88.2)				
Education level, n (%)			7.344 (5)[a]	—	—	.20
Illiterate	4 (5.9)	6 (8.8)				
Primary school	9 (13.2)	14 (20.6)				
Junior high school	22 (32.4)	10 (14.7)				
High school	21 (30.9)	22 (32.4)				
Bachelor degree	12 (17.6)	15 (22.1)				
Postgraduate and above	0 (0)	1 (1.5)				
Have children, n (%)			—	—	—	—
Yes	68 (100)	68 (100)				
No	0 (0)	0 (0)				
Smoke, n (%)			4.946 (3) [a]	—	—	.18
Yes, smoker	11 (16.2)	19 (27.9)				
No, never smoke	38 (55.9)	39 (57.4)				
Used smoke, quite now	13 (19.1)	7 (10.3)				
Occasional smoker	6 (8.8)	3 (4.4)				
Smoke year, median (IQR)	0 (0-30)	0 (0-34.3)	—	—	–0.024[d]	.98
Drink, n (%)			2.079 (3)[a]	—	—	.56
Yes, drinker	8 (11.8)	14 (20.6)				
No, never drink	35 (51.5)	33 (48.5)				
Used drink, quite now	9 (13.2)	7 (10.3)				
Occasional drinker	16 (23.5)	14 (20.6)				
Drink year, median (IQR)	0 (0-39.3)	15 (0-40)	—	—	–0.398[d]	.69
Personal income, n (%)			0.256 (1)[a]	—	—	.61
Have income	58 (85.3)	60 (88.2)				
No income	10 (14.7)	8 (11.8)				

XSL•FO
RenderX

Variable	IG (n=68)	CG (n=68)	Chi-square (*df*)[a,b]	*t* test (*df*)[c]	Z[d]	*P* value
Personal income (RMB)[f,g], median (IQR)	4000 (2800-5375)	4000 (2850-5375)	—	—	–0.264[d]	.79
Household income, n (%)			0.000 (1)[b]	—	—	>.99
Have income	66 (97.1)	66 (97.1)				
No income	2 (2.9)	2 (2.9)				
Household income (RMB)[f,g], median (IQR)	8000 (7000-10,750)	8000 (6000-10,000)	—	—	–0.389[d]	.70
Occupational status, n (%)			1.578 (3)[a]	—	—	.66
Employed	9 (13.2)	10 (14.7)				
Unemployed	6 (8.8)	8 (11.8)				
Retired	40 (58.8)	42 (61.8)				
Others	13 (19.1)	8 (11.8)				
Medical payment method, n (%)			0.342 (3)[a]	—	—	.93
Self-paid	1 (1.5)	2 (2.9)				
Medical insurance	61 (89.7)	60 (88.2)				
Commercial insurance	5 (7.4)	5 (7.4)				
Others	1 (1.5)	1 (1.5)				
BMI (kg/m^2), mean (SD)	24.04 (3.475)	24.08 (3.428)	—	–0.066 (134)[c]	—	.65
<18.5	4 (5.9)	4 (5.9)				
18.5-23.9	30 (44.1)	27 (39.7)				
24-27.9	27 (39.7)	29 (42.6)				
≥28	7 (10.3)	8 (11.8)				
Smartphone usage, n (%)			1.225 (1)[a]	—	—	.27
Yes	53 (77.9)	58 (85.3)				
No	15 (22.1)	10 (14.7)				
T2DM duration, mean (SD)	14.76 (8.381)	14.19 (7.658)	—	–0.063 (134)[c]	—	.55
Family history, n (%)			0.353 (2)[a]	—	—	.84
Yes	23 (33.8)	26 (38.2)				
No	28 (41.2)	25 (36.8)				
Do not know	17 (25.0)	17 (25.0)				
Family member has diabetes mellitus, n (%)			6.543 (6)[a]	—	—	.37
No	26 (33.3)	23 (29.5)				
Father	9 (11.5)	17 (21.8)				
Mother	10 (12.8)	10 (12.8)				
Sibling	11 (14.1)	5 (6.4)				
Grandparents	2 (2.6)	5 (6.4)				
Do not know	18 (23.1)	17 (21.8)				
Other relative	2 (2.6)	1 (1.3)				
Combination disease						
Hypertension, n (%)			0.034 (1)[a]	—	—	.85
No	21 (30.9)	22 (32.4)				
Yes	47 (69.1)	46 (67.6)				

XSL•FO
RenderX

Variable	IG (n=68)	CG (n=68)	Chi-square (*df*)[a,b]	*t* test (*df*)[c]	Z[d]	*P* value
Coronary heart disease, n (%)			0.873 (1)[a]	—	—	.35
No	50 (73.5)	45 (66.2)				
Yes	18 (26.5)	23 (33.8)				
Chronic heart failure, n (%)			0.831 (1)[b]	—	—	.36
No	64 (94.1)	67 (98.5)				
Yes	4 (5.0)	1 (1.5)				
Stroke, n (%)			0.133 (1)[b]	—	—	.71
No	63 (92.6)	65 (95.6)				
Yes	5 (7.4)	3 (4.4)				
Kidney disease, n (%)			0.831 (1)[b]	—	—	.36
No	64 (94.1)	67 (98.5)				
Yes	4 (5.9)	1 (1.5)				
Other diseases, n (%)			0.067 (1)[a]	—	—	.80
No	60 (88.2)	59 (86.8)				
Yes	8 (11.8)	9 (13.2)				
Type of medication, n (%)			0.216 (2)[a]	—	—	.90
Oral	37 (54.4)	35 (53.3)				
Insulin	0 (0)	0 (0)				
Both	28 (41.2)	29 (42.6)				
Others	3 (4.4)	4 (5.9)				
Kind of medication, median (IQR)	3.0 (3-4.8)	3.0 (2-5)	—	—	–0.401[d]	.69
Adverse reactions						
Hypoglycemia, n (%)			0.324 (1)[a]	—	—	.57
No	47 (69.1)	50 (73.5)				
Yes	21 (30.9)	16 (26.5)				
Digestive system, n (%)			0.471 (1)[a]	—	—	.49
No	55 (80.9)	58 (85.3)				
Yes	13 (10.1)	10 (14.7)				
Chronic heart failure, n (%)			0.133 (1)[b]	—	—	.72
No	63 (92.6)	65 (95.6)				
Yes	5 (7.4)	3 (4.4)				
Allergic reaction, n (%)			0.000 (1)[b]	—	—	>.99
No	66 (97.1)	65 (97.1)				
Yes	2 (2.9)	2 (2.9)				
Liver damage, n (%)			1.681 (1)[a]	—	—	.20
No	62 (83.8)	69 (92.0)				
Yes	12 (16.2)	6 (8.0)				
Kidney damage, n (%)			0.319 (1)[a]	—	—	.57
No	60 (88.2)	62 (91.2)				
Yes	8 (11.8)	6 (8.8)				

XSL•FO
RenderX

Variable	IG (n=68)	CG (n=68)	Chi-square (*df*)[a,b]	*t* test (*df*)[c]	Z[d]	*P* value
Others diseases, n (%)			0.000 (1)[b]	—	—	>.99
No	66 (97.1)	65 (95.6)				
Yes	2 (2.9)	3 (4.4)				

[a]Pearson chi-square.

[b]Continuity correction.

[c]Independent sample *t* test.

[d]Mann-Whitney *U* test.

[e]Not applicable.

[f]1 RMB ≈ RM 0.61 (November 08, 2024).

[g]1 RMB ≈ US $0.14 (January 07, 2026).

Medication Adherence, Medication Knowledge, Medication Attitude, Medication Behavior, Medication Belief, and Social Support Between the IG and CG at Baseline

There were no significant differences between the IG and CG of medication adherence (23.71, SD 5.725 vs 24.01, SD 5.700; t_{134}=–0.315; *P*=.820), medication knowledge (63.41, SD 19.145 vs 62.37, SD 20.202; t_{134}=0.308; *P*=.547), medication attitude (35.76, SD 3.801 vs 36.15, SD 4.129; t_{134}=–0.562; *P*=.568), medication behavior (29.62, SD 4.381 vs 29.13, SD 4.029; t_{134}=0.672; *P*=.468), medication belief (3.00, SD 3.515 vs 2.94, SD 3.777; t_{134}=0.094; *P*=.520) and social support (69.62, SD 11.402 vs 68.22, SD 10.935; t_{134}=0.729; *P*=.805) at baseline (Table 3).

Table 3. Total score of medication adherence, knowledge, attitude, behavior, belief, and social support between the intervention group (IG) and control group (CG) at baseline among older adults living with type 2 diabetes mellitus in Henan, China.

Outcomes (range)[a]	IG, mean (SD)	CG, mean (SD)	*t* test(*df*)	*P* value[b]
Medication adherence (13 to 39)	23.71 (5.725)	24.01 (5.700)	–0.315 (134)	.82
Medication knowledge (0 to 100)	63.41 (19.145)	62.37 (20.202)	0.308 (134)	.55
Medication attitude (10 to 50)	35.76 (3.801)	36.15 (4.129)	–0.562 (134)	.57
Medication behavior (9 to 40)	29.62 (4.381)	29.13 (4.029)	0.672 (134)	.47
Medication belief (–25 to +25)	3.00 (3.515)	2.94 (3.777)	0.094 (134)	.52
Social support (19 to 95)	69.62 (11.402)	68.22 (10.935)	0.729 (134)	.81

[a]Higher scores of outcomes indicate better medication adherence, knowledge, attitude, behavior, belief, and social support.

[b]*P* value with *P*>.05 indicates that there are no significant differences between the IG and CG for the measured outcomes.

Outcomes

Overview

This study includes 6 outcome variables, including medication adherence, medication knowledge, medication attitude, medication behavior, medication beliefs, and social support. Given the absence of significant effects, no covariates were incorporated into the final model. Generalized estimating equation was used to investigate the changes in outcomes between IG and CG across 3 time points. The results are shown in Table 4.

XSL•FO
RenderX

Table 4. Comparison of medication adherence, medication knowledge, medication attitude, medication behavior, medication belief, and social support scores between the intervention group (IG) and control group (CG) across time points using generalized estimating equations among older adults living with type 2 diabetes mellitus in Henan, China.

Variable	Medication knowledge[a]		Medication attitude[b]		Social support[c]		Medication belief[d]		Medication behavior[e]		Medication adherence[f]	
	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value
Group effect												
CG	Ref[g]	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
IG	7.887 (3.428 to 12.345)	.001[h]	2.465 (1.466 to 3.464)	<.001[h]	1.277 (–4.771 to 7.352)	.679	1.845 (1.078 to 2.613)	<.001[h]	2.731 (1.639 to 3.823)	<.001[h]	2.288 (0.681 to 3.894)	.005[h]
Time effect												
T0[i]	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
T1[j]	7.093 (4.564 to 9.623)	<.001[h]	3.103 (2.370 to 3.836)	<.001[h]	0.397 (–0.564 to 1.358)	.418	1.632 (1.125 to 2.140)	<.001[h]	2.441 (1.848 to 3.034)	<.001[h]	2.601 (1.858 to 3.363)	<.001[h]
T2[k]	4.844 (2.462 to 7.227)	<.001[h]	1.662 (1.059 to 2.264)	<.001[h]	–0.574 (–1.491 to 0.344)	.221	0.588 (0.062 to 1.114)	.028[h]	1.125 (0.586 to 1.664)	<.001[h]	1.684 (0.958 to 2.409)	<.001[h]
Timepoint×Group effect												
T0 × CG	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
T1 × IG	11.592 (6.923 to 16.260)	<.001[h]	5.467 (4.531 to 6.763)	<.001[h]	0.008 (–1.834 to 2.011)	.928	2.882 (1.990 to 3.775)	<.001[h]	4.176 (3.220 to 5.133)	<.001[h]	4.956 (3.702 to 6.210)	<.001[h]
T2× IG	6.055 (1.401 to 10.710)	.011[h]	3.059 (1.969 to 4.149)	<.001[h]	–0.147 (–1.982 to 1.688)	.875	1.882 (0.880 to 2.885)	<.001[h]	2.426 (1.429 to 3.424)	<.001[h]	3.691 (2.379 to 5.003)	<.001[h]

[a]Group: QIC=105,178.155, QICC=105,173.944, time: QIC=106,219.108, QICC=106,219.108, time × group: QIC=99,054.337, QICC=99,054.337.

[b]Group: QIC=5783.910, QICC=5780.419, time: QIC=5787.824, QICC=5787.824, time × group: QIC=4587.706, QICC=4687.706.

[c]Group: QIC=37,306.144, QICC=36,267.594, time: QIC=37,378.507, QICC=37,378.597, time × group: QIC=37,198.191, QICC=37,198.191.

[d]Group: QIC=3561.609, QICC=3558.406 time: QIC=3643.971, QICC=3643.971, time × group: QIC=3227.618, QICC=3227.618.

[e]Group: QIC=6009.207, QICC=6004.590, time: QIC=6301.456, QICC=6301.456, time × group: QIC=5304.176, QICC=5304.176.

[f]Group: QIC=11,557.678, QICC=11,551.984 time: QIC=11,731.963, QICC=11,731.963, time × group: QIC=10,617.505, QICC=10,617.505.

[g]Ref: Reference group.

[h]Significant at *P*<.05.

[i]T0: baseline.

[j]T1: postintervention.

[k]T2: 3-month follow up.

Primary Outcome

The IG exhibited a significantly higher medication adherence score compared with the CG, with a mean difference of 2.288 points (β=2.288, 95% CI 0.681-3.894, *P*=.005). Participants demonstrated significant improvements in medication adherence both immediately after the intervention (β=2.610, 95% CI 1.858-3.363, *P*<.001) and at the 3-month follow-up (β=1.684, 95% CI 0.958-2.409, *P*<.001) when compared with baseline. The IG showed a significantly higher medication adherence score postintervention, with a difference of 4.956 points compared with the CG at baseline (β=4.956, 95% CI 3.702-6.210, *P*<.001). At the 3-month follow-up, the IG maintained a significantly higher score than the CG at baseline, with a difference of 3.691 points (β=3.691, 95% CI 2.379-5.003, *P*<.001).

Secondary Outcomes

Compared with the CG, the IG showed a significant improvement in medication knowledge score, with a postintervention score higher by 7.887 points (β=7.887, 95% CI 3.428-12.345, *P*=.001). Compared with baseline, the

XSL•FO
RenderX

medication knowledge score of the participants increased by 7.093 points after the intervention and by 4.844 points at the 3-month follow-up (β=7.093, 95% CI 4.564-9.623, *P*<.001; β=4.844, 95% CI 2.462–7.227, *P*<.001). After the intervention, the medication knowledge score of the IG group was significantly higher than the CG group at baseline (β=11.592, 95% CI 6.923-16.260, *P*<.001). At the 3-month follow-up, the medication knowledge score of the IG group was 6.055 points higher than that of the CG group at baseline (β=6.055, 95% CI 1.401-10.710, *P*=.011).

The medication attitude score of the IG was significantly higher than that of CG, with a difference of 2.465 points (β=2.465, 95% CI 1.466-3.464, *P*<.001). Compared with the baseline, the medication attitude score of the participants significantly increased by 3.103 points after the intervention and by 1.662 points at the 3-month follow-up, respectively (β=3.103, 95% CI 2.370-3.836, *P*<.001; β=1.662, 95% CI 1.059-2.264, *P*<.001). After the intervention, the medication attitude score of the IG group was significantly higher than that of the CG group at baseline, with a difference of 5.467 points (β=5.467, 95% CI 4.531-6.763, *P*<.001). At the 3-month follow-up, the medication attitude score of the IG group was 3.059 points higher than that of the CG group at baseline (β=3.059, 95% CI 1.969-4.149, *P*<.001).

The IG showed a significantly higher medication behavior score compared with the CG, with a mean difference of 2.731 points (β=2.731, 95% CI 1.639-3.823, *P*<.001). Participants demonstrated significant improvements in medication behavior relative to baseline, both immediately after the intervention (β=2.441, 95% CI 1.848-3.034, *P*<.001) and at the 3-month follow-up (β=1.125, 95% CI 0.586-1.664, *P*<.001). Postintervention, the IG scored 4.176 points higher than the CG at baseline (β=4.176, 95% CI 3.220-5.133, *P*<.001). At the 3-month follow-up, the IG also outperformed the CG at baseline, with a difference of 2.426 points (β=2.426, 95% CI 1.429-3.424, *P*<.001).

The IG demonstrated a significantly higher medication belief score than the CG, with an estimated difference of 1.845 points (β=1.845, 95% CI 1.078-2.613, *P*<.001). Compared with baseline, participants showed a significant increase in medication belief score both immediately after the intervention (β=1.632, 95% CI 1.125-2.140, *P*<.001) and at the 3-month follow-up (β=0.588, 95% CI 0.062-1.114, *P*=.028). Furthermore, in the analysis of the interaction between time and group, the IG at postintervention scored significantly higher than the CG at baseline, with a difference of 2.882 points (β=2.882, 95% CI 1.990-3.775, *P*<.001). Similarly, at the 3-month follow-up, the IG maintained a significantly higher score than the CG at baseline (β=1.882, 95% CI 0.880-2.885, *P*<.001).

However, social support did not reach statistical significance in terms of time, group, and the interaction between time and group, in either the IG or CG group (*P*>.05).

Discussion

Principal Findings

Research has shown that the IMB model can effectively explain, predict, and improve medication adherence among older adults living with chronic diseases [54,55]. Based on previous research findings, this study combined the IMB model to develop an animated intervention video called “Diabetes Little Helper” which was disseminated through the WeChat platform, further validating the model’s applicability in older adults. As with the results of this study, a quasi-experimental study by Mirzaei-Alavijeh et al [56], based on the IMB model, significantly improved medication adherence in older adults living with T2DM through 6 face-to-face educational sessions, further supporting the effectiveness of the IMB model. This study, through the WeChat platform, disseminates educational information in video format, providing new evidence and intervention pathways for improving medication adherence among older adults living with T2DM.

A notable proportion of participants reported not using smartphones independently. This finding is significant, as it could influence the overall effectiveness of the intervention. However, it is important to note that all participants in this study either owned a smartphone and used WeChat or received assistance from family members to participate. Nonetheless, the involvement of family members could be seen as a limitation, as it may affect the autonomy of older adults and potentially influence their overall engagement with the intervention. Additionally, reliance on family assistance raises concerns about the generalizability and long-term feasibility of the intervention, particularly for older adults who may not have family members available to offer support. To address this, future research could explore alternative strategies to enhance technological support for older adults, such as simplified user interfaces or targeted training programs. These approaches could ensure broader accessibility and improve the effectiveness of similar interventions in diverse populations.

Participants’ medication adherence score showed significant improvement following the intervention, a result that mirrors outcomes seen in other studies using different methods. For instance, mobile apps were used to provide educational information to family members, who in turn helped older adults living with DM complete self-management tasks, pharmacy students conducted motivational interviewing through phone calls to enhance medication adherence, while pharmacists used apps, WeChat public accounts, WeChat groups, and both web-based and offline lectures to improve medication adherence among older adults living with DM [30,31,33]. These interventions effectively increased medication adherence in older adults living with DM, though they also presented certain challenges. Relying on family members to convey information poses the risk of information distortion and places psychological pressure on the family. Dependence on pharmacy students for motivational interviewing might lead to insufficient mastery of communication and problem-solving skills, potentially affecting the interview’s effectiveness. Although pharmacists use a combination of web-based and offline platforms to provide

XSL•FO
RenderX

comprehensive services, the use of multiple channels may confuse older adults, especially those less familiar with digital technologies. In contrast, this study directly engaged patients via WeChat, eliminating hierarchical errors in information transmission and avoiding family involvement in the medical process. The professionally designed animated videos ensured the accuracy and coherence of the information, minimizing human bias. Moreover, the approach did not require the development of complex programs or offline services, thus saving both manpower and financial resources.

This study demonstrated that the Diabetes Little Helper video significantly improved participants' medication knowledge score, which is consistent with previous research findings [57]. Compared with the one-on-one face-to-face educational model used by Yang et al [57], this study overcame the limitations of time and location while also saving human resources. Xiong and Wan [58] also improved older adult participants' medication knowledge score by sending text and images via the WeChat platform. In comparison, the animated video used in this study not only incorporated the advantages of text and images but also effectively captured patients' attention through its vivid and engaging format, helping viewers better understand concepts and stimulate cognitive responses. The use of animated videos has been shown to be a more attractive and dynamic method, aiding viewers in gaining a deeper understanding of health education content [59]. Future intervention research could explore how to combine face-to-face education with digital video interventions, leveraging the strengths of both traditional and modern educational methods to further enhance patients' medication knowledge scores.

Participants' medication attitudes were effectively improved, as agreed with Chen et al [60], who used a multidimensional, multiscenario intervention approach, which also successfully enhanced medication attitudes. However, their method requires substantial human resources and high costs, making it challenging to implement in regions with limited health care resources. Zhou [61] conducted an intervention through community health service centers, regularly monitoring patients' medication levels, metabolism, and relevant physiological indicators to guide medication adjustments, which improved medication attitudes in older adults. In contrast, the intervention model in this study overcomes regional and resource limitations through digital means, making it simple and feasible. Each intervention method has its unique advantages and suitable contexts. Future research should flexibly select intervention strategies based on the specific characteristics of hospitals and patients, and integrate multiple approaches to more effectively improve medication attitudes in older adults.

The intervention also had a positive impact on improving participants' medication behavior score, in line with the findings of Wang et al [62], whose study increased medication adherence through personalized plans and continuous optimization. However, their approach relied heavily on frequent face-to-face reminders, which may have provoked resistance from some patients. In this study, patients were able to choose their learning time and frequency, which reduced resistance and improved psychological acceptance. Moreover, the intervention program by Wan et al [63] required 50 participants to provide more tailored health education and medication management. Although it also led to some improvement in medication behavior, the management process was relatively complex and added additional costs. Moreover, staff turnover, absenteeism, and communication difficulties could negatively impact the effectiveness of the intervention.

The results of this study also indicate a significant improvement in participants' medication beliefs, aligning with findings from previous research [64,65]. These studies used interviews or multidisciplinary collaboration (involving family doctors, pharmacists, and nurses) to integrate and guide medication management. However, these approaches are typically limited to stable patients who can attend interviews and are not applicable to older adults living at home. Additionally, team collaboration demands more human resources. However, this study focused on broadening medication knowledge and promoting self-management, encouraging older adults to independently adjust their medication attitudes.

This study did not show a significant effect on improving participants' social support scores. This finding is consistent with the results of Yang et al [57], who also did not observe significant improvements in social support after using traditional intervention methods, such as face-to-face education, interviews, phone communication, and record-keeping. One possible reason for this is that the development of social support requires sustained intervention over more than 6 months [66]. Additionally, some studies have shown that including education for patients' family members and facilitating web-based interactive communication effectively improved social support among older adults [67,68]. Future research could integrate interactive modules, such as involving family members in the intervention plan, promoting web-based communication among patients, and combining community nursing, to further strengthen the patient's social support network and enhance the overall intervention effectiveness.

This study is the first to develop an animated video intervention to improve medication adherence in older adults living with T2DM, expanding the application of this model in medication management for older adults living with DM and providing new insights for chronic disease management. Moreover, this study innovatively used the WeChat public platform to deliver animated video intervention content to older adults living with DM. This approach not only overcomes the time and space limitations of traditional health education but also leverages the widespread use and interactivity of the WeChat platform, enabling continuous health interventions. As a result, it significantly increased the accessibility of intervention content and enhanced patient acceptance. Patients could repeatedly watch the intervention videos in their daily lives, which helped reinforce memory and promote behavior change.

Limitations and Future Research

While the findings of this study provide valuable insights into the effectiveness of the intervention, several limitations must be acknowledged. First, the reliance on self-reported adherence introduces potential biases, such as over-reporting, underreporting, and recall bias. These biases can compromise the accuracy of adherence measurements and may not fully

XSL•FO
RenderX

reflect the participants' true behaviors. Future studies could incorporate objective measures of adherence, such as electronic monitoring or biomarkers, to reduce these biases and provide a more accurate picture of adherence. Second, while the 3-month follow-up period allowed us to assess the short-term effects of the intervention, it was insufficient to evaluate the long-term sustainability of the intervention's impact. Future studies should aim for extended follow-up periods to gain a deeper understanding of the long-term benefits and challenges of such interventions. Furthermore, the inability to blind participants due to the visible nature of the intervention video may have introduced bias, as participants were aware of their intervention condition, which could have influenced their responses or behaviors. To mitigate this in future studies, an RCT design with blinding could be implemented, or a different form of intervention that is less visible could be explored.

Adherence monitoring measures, though designed to ensure compliance, may have inadvertently influenced engagement and adherence, thereby introducing additional intervention effects. Future studies should consider monitoring approaches that minimize such influences. In terms of data handling, this study used the LOCF method to address missing data. Although LOCF is computationally simple and commonly used in short-term studies, it assumes that missing data are equivalent to the last observed value, which may lead to an underestimation of variability in the results. If participants drop out due to deteriorating health, LOCF may fail to capture this change, potentially overestimating the intervention effect. Despite these limitations, LOCF was considered an appropriate choice for this study. Future research could consider exploring other advanced methods for handling missing data, such as multiple imputation, to improve the accuracy of data analysis.

Conclusions

This study evaluated the impact of a WeChat-based Diabetes Little Helper video intervention on medication adherence and medication-related outcomes in older adults living with T2DM. The results indicated that the intervention had a significant positive effect on improving patients' medication adherence, knowledge, attitude, behavior, and belief. However, it did not significantly enhance the patients' level of social support. Building on these findings, future research could expand to include more regions and health care institutions and conduct multicenter RCTs to verify the applicability and effectiveness of the Diabetes Little Helper video intervention across different geographic and cultural contexts. This would help to further improve the external validity and reliability of the study's findings. Moreover, future studies could explore the integration of more advanced health tools, such as smart wristbands and smart pillboxes, to monitor the health status and medication adherence of older adults living with T2DM in real time, thereby providing additional support for personalized interventions and data analysis.

Acknowledgments

The authors gratefully acknowledge the support of Universiti Putra Malaysia and 2 hospitals in Henan province, China, for their contributions to this research. The authors also thank Universiti Putra Malaysia for the publication fees support. We also extend our heartfelt thanks to all the experts who offered their valuable guidance, professional insights, and feedback, which were instrumental in enhancing the quality and impact of this intervention study. All authors approved this manuscript for publication.

Funding

This study did not receive research funding from any public, commercial, or nonprofit funding agencies. The publication fees are supported by Universiti Putra Malaysia.

Data Availability

The datasets generated or analyzed during this study are available in the Mendeley data repository.

Authors' Contributions

MW, PYL, KL, SNS, HZT, and QHJ contributed to the conceptualization and design of the study. Data collection was carried out by MW, QHJ, LLH, ZYZ, and YL. MW and PYL conducted the data analysis. Interpretation of the results was performed by MW, PYL, SNS, and HZT. MW was responsible for drafting and revising the manuscript. Supervision was provided by PYL, KL, SNS, and HZT. Project administration was managed by PYL, SNS, and HZT. Ethical approval was obtained by MW and QHJ.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1108 KB - jmir_v28i1e78731_app1.pdf]

XSL•FO
RenderX

References

1. Sun, Wen H, Liu X, Ma Y, Jang J, Yu C. Time trends in type 2 diabetes mellitus incidence across the BRICS from 1990 to 2019: an age-period-cohort analysis. BMC Public Health 2022;22(1):65 [FREE Full text] [doi: 10.1186/s12889-021-12485-y] [Medline: 35012503]
2. Hossain MJ, Al-Mamun M, Islam MR. Diabetes mellitus, the fastest growing global public health concern: early detection should be focused. Health Sci Rep 2024;7(3):e2004 [FREE Full text] [doi: 10.1002/hsr2.2004] [Medline: 38524769]
3. Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, et al. Erratum to "IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045" [Diabetes Res. Clin. Pract. 183 (2022) 109119]. Diabetes Res Clin Pract 2023;204:110945. [doi: 10.1016/j.diabres.2023.110945] [Medline: 37863776]
4. Sinclair A, Saeedi P, Kaundal A, Karuranga S, Malanda B, Williams R. Diabetes and global ageing among 65-99-year-old adults: findings from the International Diabetes Federation Diabetes Atlas, 9 edition. Diabetes Res Clin Pract 2020;162:108078. [doi: 10.1016/j.diabres.2020.108078] [Medline: 32068097]
5. Liu J, Liu M, Chai Z, Li C, Wang Y, Shen M, et al. Projected rapid growth in diabetes disease burden and economic burden in China: a spatio-temporal study from 2020 to 2030. Lancet Reg Health West Pac 2023;33:100700 [FREE Full text] [doi: 10.1016/j.lanwpc.2023.100700] [Medline: 36817869]
6. Babakhanian M, Razavi A, Rahimi Pordanjani S, Hassanabadi S, Mohammadi G, Fattah A. High incidence of type 1 diabetes, type 2 diabetes and gestational diabetes in Central Iran: a six years results from Semnan health cohort. Ann Med Surg (Lond) 2022;82:104749 [FREE Full text] [doi: 10.1016/j.amsu.2022.104749] [Medline: 36268322]
7. Baechle C, Stahl-Pehe A, Prinz N, Meissner T, Kamrath C, Holl R, in cooperation with the German Paediatric Surveillance Unit (ESPED), the DPV initiativethe German Center for Diabetes Research (DZD). Prevalence trends of type 1 and type 2 diabetes in children and adolescents in North Rhine-Westphalia, the most populous federal state in Germany, 2002-2020. Diabetes Res Clin Pract 2022;190:109995. [doi: 10.1016/j.diabres.2022.109995] [Medline: 35853531]
8. Li D, Luo M, Liu Y, Dong J, Geng W, Li X, et al. Increased rates of health management and health education on hypertension and diabetes in Inner Mongolia, China: 10-year population level trends (2009-2018). Int J Environ Res Public Health 2022;19(20) [FREE Full text] [doi: 10.3390/ijerph192013048] [Medline: 36293621]
9. Tomic D, Harding JL, Jenkins AJ, Shaw JE, Magliano DJ. The epidemiology of type 1 diabetes mellitus in older adults. Nat Rev Endocrinol 2025;21(2):92-104. [doi: 10.1038/s41574-024-01046-z] [Medline: 39448829]
10. Han J, Wang H, Liang HH, Guo JX. Correlation of the retinopathy degree with the change of ocular surface and corneal nerve in patients with type 2 diabetes mellitus. Int J Ophthalmol 2021;14(5):750-758 [FREE Full text] [doi: 10.18240/ijo.2021.05.17] [Medline: 34012892]
11. Kirthi V, Perumbalath A, Brown E, Nevitt S, Petropoulos IN, Burgess J, et al. Prevalence of peripheral neuropathy in pre-diabetes: a systematic review. BMJ Open Diabetes Res Care 2021;9(1) [FREE Full text] [doi: 10.1136/bmjdrc-2020-002040] [Medline: 34006607]
12. Wang L, Peng W, Zhao Z, Zhang M, Shi Z, Song Z, et al. Prevalence and treatment of diabetes in China, 2013-2018. JAMA 2021;326(24):2498-2506 [FREE Full text] [doi: 10.1001/jama.2021.22208] [Medline: 34962526]
13. Emerging Risk Factors Collaboration. Life expectancy associated with different ages at diagnosis of type 2 diabetes in high-income countries: 23 million person-years of observation. Lancet Diabetes Endocrinol 2023;11(10):731-742 [FREE Full text] [doi: 10.1016/S2213-8587(23)00223-1] [Medline: 37708900]
14. Jyotsna F, Ahmed A, Kumar K, Kaur P, Chaudhary M, Kumar S, et al. Exploring the complex connection between diabetes and cardiovascular disease: analyzing approaches to mitigate cardiovascular risk in patients with diabetes. Cureus 2023;15(8):e43882 [FREE Full text] [doi: 10.7759/cureus.43882] [Medline: 37746454]
15. Kumar M, Dev S, Khalid M, Siddenthi S, Noman M, John C, et al. The bidirectional link between diabetes and kidney disease: mechanisms and management. Cureus 2023;15(9):e45615 [FREE Full text] [doi: 10.7759/cureus.45615] [Medline: 37868469]
16. Kirya M, Bwayo D, Otim M, Mutoo P, Masaba J, Okibure A, et al. Prevalence of biomarkers and associated factors for chronic kidney disease in adult diabetic out-patients in a tertiary hospital in Eastern Uganda - a cross-sectional study. Res Sq 2024 [FREE Full text] [doi: 10.21203/rs.3.rs-3992049/v1] [Medline: 38496471]
17. Taimur H, Ahmad I, Khan H, Shirayama Y, Okamoto M, Aung MN, et al. A scoping review of type 2 diabetes mellitus in Pakistan investigating the status of glycemic control, awareness, treatment adherence, complications and cost. Front Endocrinol (Lausanne) 2024;15:1441591 [FREE Full text] [doi: 10.3389/fendo.2024.1441591] [Medline: 39649227]
18. Brown MT, Bussell J, Dutta S, Davis K, Strong S, Mathew S. Medication adherence: truth and consequences. Am J Med Sci 2016;351(4):387-399. [doi: 10.1016/j.amjms.2016.01.010] [Medline: 27079345]
19. Weinberg Sibony R, Segev O, Dor S, Raz I. Drug therapies for diabetes. Int J Mol Sci 2023 Dec 05;24(24). [doi: 10.3390/ijms242417147] [Medline: 38138975]
20. Pallangyo P, Millinga J, Bhalia S, Mkojera Z, Misidai N, Swai HJ, et al. Medication adherence and survival among hospitalized heart failure patients in a tertiary hospital in Tanzania: a prospective cohort study. BMC Res Notes 2020;13(1):89 [FREE Full text] [doi: 10.1186/s13104-020-04959-w] [Medline: 32085803]

XSL•FO
RenderX

21. Zhang, Nie X, Yang X, Mei Q, Xiang X, Cheng L. A meta-analysis of effectiveness of mobile health interventions on health-related outcomes in patients with heart failure. J Cardiovasc Med (Hagerstown) 2024;25(8):587-600. [doi: 10.2459/JCM.0000000000001631] [Medline: 38949146]
22. Rasu RS, Xavier C, Rianon N. Dynamic changes in medication burden leading to fall and hospital readmissions in older adults: toward a strategy for improving risk and managing costs. J Manag Care Spec Pharm 2025;31(1):96-100. [doi: 10.18553/jmcp.2025.31.1.96] [Medline: 39745841]
23. He HM, Cai H, Rong QF, Heng JQ. Investigation of factors influencing medication adherence among elderly patients with chronic diseases and strategies for dealing with them. Chin Remedies Clin 2020;20:1622-1624 [FREE Full text] [doi: 10.11655/zgywylc2020.10.013]
24. Zhou HB, Zhang YZ, Yang H, Xia Q, Su H. Investigation on medication compliance of elderly patients with type 2 diabetes mellitus. Diabetes New World 2022 Feb 16;25:13-16 [FREE Full text] [doi: 10.16658/j.cnki.1672-4062.2022.04.013]
25. Khan II, Pulock OS, Pinky SD, Barua B, Dola TA, Chowdhury P, et al. Treatment noncompliance level among patients with type 2 diabetes mellitus: a hospital based cross-sectional study in Bangladesh. PLoS One 2022;17(9):e0271107 [FREE Full text] [doi: 10.1371/journal.pone.0271107] [Medline: 36099268]
26. Zhou JY. Analysis of medication compliance and influencing factors among elderly patients with chronic diseases in Fangshan district, Beijing. Mod Dis Control Prev 2023;34(12):930-957 [FREE Full text] [doi: 10.13515/j.cnki.hnjpm.1006-8414.2023.12.012]
27. Zhang JY, Zhang LQ, Tao MX, Chen XL. Analysis of medication adherence status and influencing factors in elderly type 2 diabetes patients. Pract Prev Med 2024;31(11):1382-1385 [FREE Full text] [doi: 10.3969/j.issn.1006-3110.2024.11.025]
28. Yao YC, Zhong YJ, Yang L. Compliance health education for elderly patients with multidrug use management intervention based on multidisciplinary cooperation. J Prev Med 2023 Jan 23;39(12):1464-1469 [FREE Full text]
29. Li B, Zang S, Ma LP. Health education management in the follow-up of elderly patients with diabetes mellitus. J Qilu Nurs 2023 Feb 05;29(3):111-113 [FREE Full text] [doi: 10.3969/j.issn.1006-7256.2023.03.033]
30. Poonprapai P, Lerkiatbundit S, Saengcharoen W. Family support-based intervention using a mobile application provided by pharmacists for older adults with diabetes to improve glycaemic control: a randomised controlled trial. Int J Clin Pharm 2022;44(3):680-688. [doi: 10.1007/s11096-022-01389-5] [Medline: 35247147]
31. Mohan A, Majd Z, Johnson ML, Essien EJ, Barner J, Serna O, et al. A motivational interviewing intervention to improve adherence to ACEIs/ARBs among nonadherent older adults with comorbid hypertension and diabetes. Drugs Aging 2023;40(4):377-390 [FREE Full text] [doi: 10.1007/s40266-023-01008-6] [Medline: 36847995]
32. Wang ZM, Sun BY, Hu YY. Influence of remote clinical pharmacy service platform on drug use in older patients with type 2 diabetes mellitus. Int J Geriatr 2023 May 20;44(3):280-283 [FREE Full text] [doi: 10.3969/j.issn.1674-7593.2023.03.005]
33. Zhang XL, Liu HL, Wang S. Influence of family pharmacists′ service on medication adherence and efficacy of elderly patients with type 2 diabetes mellitus in community. Chin Pharm 2024;33(22):48-52 [FREE Full text] [doi: 10.3969/j.issn.1006-4931.2024.22.010]
34. Lee S, Kim B. Factors affecting nurses' health promotion behavior during the COVID-19 pandemic based on the information-motivation-behavioral skills model. Medicina (Kaunas) 2022;58(6) [FREE Full text] [doi: 10.3390/medicina58060720] [Medline: 35743983]
35. Park N, Jang S. Effects of mHealth practice patterns on improving metabolic syndrome using the information-motivation-behavioral skills model. Nutrients 2024;16(13) [FREE Full text] [doi: 10.3390/nu16132099] [Medline: 38999847]
36. Nelson LA, Wallston KA, Kripalani S, LeStourgeon LM, Williamson SE, Mayberry LS. Assessing barriers to diabetes medication adherence using the information-motivation-behavioral skills model. Diabetes Res Clin Pract 2018;142:374-384 [FREE Full text] [doi: 10.1016/j.diabres.2018.05.046] [Medline: 29879495]
37. Schoenthaler A, Leon M, Butler M, Steinhaeuser K, Wardzinski W. Development and evaluation of a tailored mobile health intervention to improve medication adherence in black patients with uncontrolled hypertension and type 2 diabetes: pilot randomized feasibility trial. JMIR Mhealth Uhealth 2020;8(9):e17135 [FREE Full text] [doi: 10.2196/17135] [Medline: 32965230]
38. Xu Q, Wu P, Du J, Zhuang W, He X, Ma Y, et al. Online short videos promoting public breast cancer literacy: a pretest-posttest control group trial on efficiency, attitude, and influencing factors. Front Public Health 2023;11:1198780 [FREE Full text] [doi: 10.3389/fpubh.2023.1198780] [Medline: 37397762]
39. Liang H, Pang H. The cognitive, behavioral and interpersonal impacts of virtual practice with short health videos on Chinese ageing women: a discursive approach. Int J Environ Res Public Health 2022;19(12) [FREE Full text] [doi: 10.3390/ijerph19127173] [Medline: 35742422]
40. Gould CE, Carlson C, Wetherell JL, Goldstein MK, Anker L, Beaudreau SA. Brief video-delivered intervention to reduce anxiety and improve functioning in older veterans: pilot randomized controlled trial. JMIR Aging 2024;7:e56959 [FREE Full text] [doi: 10.2196/56959] [Medline: 39652863]
41. Al-Kharouf MS, Abdeljalil MH, Obeidat NM, Oweidat KA, Awwad O. Video-based teach-to-goal intervention on inhaler technique on adults with asthma and COPD: a randomized controlled trial. PLoS One 2023;18(6):e0286870 [FREE Full text] [doi: 10.1371/journal.pone.0286870] [Medline: 37294823]

XSL•FO
RenderX

42. Wang M, Lim PY, Lee K, Sowtali SN, Jiang Q, Thew HZ. Effectiveness of diabetes little helper video intervention on medication adherence among elderly patients with type 2 diabetes mellitus in Henan, China: study protocol for a randomised controlled trial. BMJ Open 2025;15(4):e093191 [FREE Full text] [doi: 10.1136/bmjopen-2024-093191] [Medline: 40258641]
43. Chinese Elderly Type 2 Diabetes PreventionTreatment of Clinical Guidelines Writing Group. [Clinical guidelines for prevention and treatment of type 2 diabetes mellitus in the elderly in China (2022 edition)]. Zhonghua Nei Ke Za Zhi 2022 Jan 01;61(1):12-50. [doi: 10.3760/cma.j.cn112138-20211027-00751] [Medline: 34979769]
44. Lemeshow S, Hosmer DW. Adequacy of Sample Size in Health Studies. New Jersey: Wiley & Sons; 1992:120.
45. Cao Z. Analysis of the impact of clinical pharmacist guidance on rational medication and compliance in elderly patients with chronic diseases. Chin J Mod Drug Appl 2024;18(6):160-163 [FREE Full text] [doi: 10.14164/j.cnki.cn11-5581/r.2024.06.042]
46. Wu X, Qian J, Gong L. Targeted nursing interventions based on SEAMS assessment: effects on blood pressure control and medication adherence in elderly patients with hypertension. Heilongjiang Med Pharm 2024 Apr 8;47(2):84-89 [FREE Full text] [doi: 10.3969/j.issn.1008-0104.2024.02.026]
47. Yan C, Ding Y. Micro-lecture design and development based on Wondershare Animiz. Shoes Technol Des 2025 Jul 28:5-29 [FREE Full text] [doi: 10.31958/je.v3i1.10120]
48. Sowtali SN. Development, implementation and evaluation of family-based educational program on nutrient intake and stroke knowledge for first ever stroke patients. In: Universiti Sains Malaysia. Malaysia: Universiti Sains Malaysia; 2016.
49. Dong XF, Liu YJ, Wang A, Lv PH. Psychometric properties of the Chinese version of the self-efficacy for appropriate medication use scale in patients with stroke. Patient Prefer Adherence 2016;10:321-327 [FREE Full text] [doi: 10.2147/PPA.S101844] [Medline: 27042023]
50. Dong XF, Liu YJ, Wang AX. Reliability and validity of the Chinese version self-efficacy for appropriate medication use scale. J Nurs Sci 2015;30:47-49. [doi: 10.3870/hlxzz.2015.11.047]
51. Xiang GP. A study on the effect of health education on knowledge, attitude, behavior of safe medication for community-dwelling elderly with chronic diseases. In: Central South University. Changsha: Central South University; Nov 2012:1-55.
52. Tang J, Gao LL, Hu XL, Wu JP, Yu KH, Liu SH. The reliability and validity of the Chinese version of beliefs about medicines questionnaire-specific among patients with type 2 diabetes mellitus. J Nurs Sci 2019;34:25-28 [FREE Full text]
53. Huang SF, Chen LX, Ye YM, Gong GM, Shi GF. Preparation of empty nester social support scale and analysis of its reliability and validity. Nurs Res 2017 Mar 09;31(1):537-539 [FREE Full text] [doi: 10.3969/j.issn.1009-6493.2017.05.007]
54. Yang C, Lee DTF, Chair SY. Validation and adaption of an information-motivation-behavioral skills model of medication adherence in older patients with multimorbidity. Patient Educ Couns 2023;113:107756. [doi: 10.1016/j.pec.2023.107756] [Medline: 37156122]
55. Oh E, Kim C, Schlenk EA. A predictive model for medication adherence in older adults with heart failure. Eur J Cardiovasc Nurs 2024 Oct 05;23(6):635-643. [doi: 10.1093/eurjcn/zvae021] [Medline: 38408016]
56. Mirzaei-Alavijeh M, Yarmoradi S, Khashij S, Jalilian F. Application of the information-motivation-behavioral skills model to improve medication adherence among older adults with type 2 diabetes: findings from a quasi-experimental study in Kermanshah, Iran. BMC Geriatr 2025 May 13;25(1):328 [FREE Full text] [doi: 10.1186/s12877-025-06000-w] [Medline: 40361006]
57. Yang C, Lee DTF, Wang X, Chair SY. Effects of a nurse-led medication self-management intervention on medication adherence and health outcomes in older people with multimorbidity: a randomised controlled trial. Int J Nurs Stud 2022;134:104314. [doi: 10.1016/j.ijnurstu.2022.104314] [Medline: 35849886]
58. Xiong YY, Wan J. The influence of WeChat platform technology for the rational drug use and drug compliance of elderly hypertensive patients in the community. Shanghai Med Pharm J 2023 Aug 25;44(16):36-39 [FREE Full text] [doi: 10.3969/j.issn.1006-1533.2023.16.011]
59. Praveen CK, Srinivasan K. Psychological impact and influence of animation on viewer's visual attention and cognition: a systematic literature review, open challenges, and future research directions. Comput Math Methods Med 2022;2022:8802542 [FREE Full text] [doi: 10.1155/2022/8802542] [Medline: 36092789]
60. Chen J, Wei C, Deng L. Study on the influence of multi-faceted health education intervention inhospital-community-family model on the knowledg-attitude-practice of safe and rational drug use in elderly chronic co-patients. Chin J Hosp Pharm 2020 May 17;40(4):452-455 [FREE Full text] [doi: 10.13286/j.1001-5213.2020.04.19]
61. Zhou J. The impact of pharmaceutical care management on the rationality of medication use in elderly patients with multiple coexisting diseases. J North Pharm 2023 Jul 1;20(7):113-115 [FREE Full text] [doi: 10.3969/j.issn.1672-8351.2023.07.038]
62. Wang HH, Wang LY, Hou WY, Zhao YS. Effects of Motivational Interview on Medication Compliance?Self-management and Blood Glucose Control in Elderly Patieuts with Type 2 Diabetes. Henan Medical Research 2022 Mar 31;31(6):1023-1027 [FREE Full text] [doi: 10.3969/j.issn.1004-437X.2022.06.015]
63. Wan X, Cheng X, Wang Y, Li Y. Study on the influence of personalised health education on health behaviour, knowledge, attitude and practice of medication in elderly patients with chronic disease. Chin J Health Educ 2022 Apr 25;37(12):1135-1138 [FREE Full text] [doi: 10.16168/j.cnki.issn.1002-9982.2021.12.016]

XSL•FO
RenderX

64. Yang YX, Fang XX, Zhang HM, Wang H, Zhang CF, Wang ZP, et al. Effect of life retrospective intervention on belief about medication and perceived happiness in elderly patients with chronic kidney disease. Chin J Mod Nurs 2019 Jul 25;25(6):685-688 [FREE Full text] [doi: 10.3760/cma.j.issn.1674-2907.2019.06.006]
65. Li LS, Tang Q, Wang C, Wu WH, Wu HY, Gao ZQ, et al. The influence of rational drug use management mode of the elderly in community on patients' knowledge, attitude and practice of drug use. Chin Gen Pract 2024 Sep 5;27(25):3129-3135 [FREE Full text] [doi: 10.12114/j.issn.1007-9572.2023.0798]
66. Czaja S, Boot W, Charness N, Rogers W, Sharit J. Improving social support for older adults through technology: findings from the PRISM randomized controlled trial. Gerontologist 2018;58(3):467-477 [FREE Full text] [doi: 10.1093/geront/gnw249] [Medline: 28201730]
67. Hu L, Zhang XF, Song ZZ, Wang KJ, Song J, Jia YC, et al. The impact of family collaborative health education on the cancer-related fatigue, psychological resilience, and self-efficacy of elderly lung cancer patients. Chin J Gerontol 2023 Jun 25:3047-3050 [FREE Full text] [doi: 10.3969/j.issn.1005-9202.2023.12.059]
68. Gong YX, Li Y, Chen XL, Shang Y, Feng QQ. Effects of WeChat carrier information-knowledge-attitude-practice collaborative nursing intervention on health literacy and social support in elderly patients with chronic heart failure. Chin Evidence-Based Nurs 2025 Jan 22;11(2):277-280 [FREE Full text] [doi: 10.12102/j.issn.2095-8668.2025.02.016]

Abbreviations

CG: control group
CONSORT: Consolidated Standards of Reporting Trials
DM: diabetes mellitus
I-CVI: Item-level Content Validity Index
IG: intervention group
IMB: information–motivation–behavioral skills
LOCF: last observation carried forward
RCT: randomized controlled trial
S-CVI: Scale-level Content Validity Index
T2DM: type 2 diabetes mellitus

Edited by N Cahill; submitted 08.Jun.2025; peer-reviewed by ZR Fan, H Almomani, AN Ali; comments to author 20.Oct.2025; revised version received 28.Nov.2025; accepted 11.Dec.2025; published 17.Feb.2026.

Please cite as:
Wang M, Lee K, Thew HZ, Sowtali SN, Jiang Q, Zhao Z, Liu Y, Lim PY
Effectiveness of the Mobile-Based Diabetes Little Helper Video Intervention on Medication Adherence Among Older Adults Living With Type 2 Diabetes Mellitus, Henan, China: Randomized Controlled Trial
J Med Internet Res 2026;28:e78731
URL: https://www.jmir.org/2026/1/e78731
doi:10.2196/78731
PMID:

©Meng Wang, Khuan Lee, Hui Zhu Thew, Siti Noorkhairina Sowtali, Qiuhuan Jiang, Zhengyan Zhao, Yu Liu, Poh Ying Lim. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Added Technology for Weight Management in Cardiac Rehabilitation: Biopsychosocial Factors and Rehabilitees' Talk in a Mixed Method Study

Heli Lahtio[1,2*], MSc; Teemu Paajanen[3*], MSc; Hilkka Korpi[1,4*], PhD; Ari Heinonen[1*], PhD; Tuulikki Sjögren[1*], PhD

[1]Faculty of Sport and Health Sciences, University of Jyväskylä, Jyväskylä, Finland

[2]Physical Activity and Functional Capacity Research Group, LAB University of Applied Sciences, Lahti, Finland

[3]Finnish Institute for Health and Welfare, Helsinki, Finland

[4]Welfare and Culture, Oulu University of Applied Sciences, Oulu, Finland

[*]all authors contributed equally

Corresponding Author:
Heli Lahtio, MSc
Faculty of Sport and Health Sciences
University of Jyväskylä
P.O. Box 35
Seminaarinkatu 15
Jyväskylä, 40014
Finland
Phone: 358 14260 1211
Email: heli.lahtio@gmail.com

Abstract

Background: Hardly any previous studies have examined technology-based interventions among cardiac rehabilitees. The added value of technology and biopsychosocial factors in weight management should be studied more in cardiac rehabilitation.

Objective: This mixed method study aimed to examine what biopsychosocial factors predict WC reduction and what factors arise from the talk of cardiac rehabilitees during 12-month cardiac rehabilitation.

Methods: A total of 59 rehabilitees (mean age 60, SD 6 years; 48/59, 81.3% male) were randomly assigned in pairs into the experimental group (n=29) or the reference group (n=30). Biopsychosocial outcomes were measured at baseline and after 6 and 12 months with a 6-minute walk test (6MWT), the World Health Organization Quality of Life Brief Form questionnaire, body mass, BMI, WC, age, and sex. Rehabilitees' experiences were investigated using focus group interviews. Both groups received conventional cardiac rehabilitation. In addition, the experimental group used technology. Multiple linear regression was used to predict WC change in 0-6 months and 0-12 months. The qualitative data were analyzed with thematic analysis. Integrative analysis was used to create a research-based model.

Results: Group allocation predicted WC reduction (P=.007), with the experimental group showing a greater reduction (b=2.5 cm) than the reference group in the 0-6 month analysis. A significant interaction between the reference group and 6MWT was observed (P=.04), meaning that improvement in 6MWT was associated with WC reduction. Baseline WC was marginally associated with a greater reduction in WC (P=.05). None of the independent variables predicted WC reduction in 0- to 12-month analyses. Three themes arose from the interviews with the experimental group: meaningful factors related to weight management, a goal-oriented approach to weight loss, and monitoring body composition. Two themes emerged from the reference group: motivation for change and unstable weight management. According to the integrative analysis, belonging to the experimental group was important. The research-based model highlighted that the added value of remote technology on WC reduction was largest at the beginning of rehabilitation (0-6 months). The rehabilitees' talk in the experimental group was goal-oriented, and they talked about the achieved changes. In the reference group, the talk remained motivational, and their body composition changes were unstable.

Conclusions: The integrative model emphasizes the differences between the experimental and reference groups in weight management. The use of added remote technology seemed to explain the WC reduction. The rehabilitees reflected on behavioral changes in weight management. In the reference group, improved 6MWT distance was associated with decreased WC, but they

XSL•FO
RenderX

talked about motivation for weight management and unstable changes in body composition. In the future, results should be verified by larger research groups and individual interviews, which can be combined with personal quantitative data.

Trial Registration: ISRCTN Registry ISRCTN61225589; https://www.isrctn.com/ISRCTN61225589

(J Med Internet Res 2026;28:e78347) doi:10.2196/78347

KEYWORDS

weight management; waist circumference; cardiac rehabilitation; remote technology; overweight; obesity

Introduction

Background

Obesity is defined as a condition characterized by excess adiposity, with or without abnormal distribution or function of adipose tissue [1]. Obesity is not well understood. The causes are multifactorial, including genetic, metabolic, psychological, and environmental factors [1]. Obesity is a major risk factor of cardiovascular diseases (CVDs), with, for example, tobacco use, unhealthy diets, physical inactivity, and harmful use of alcohol [2]. Abdominal obesity with other risk factors, such as hypertension and diabetes, is associated with an increased risk of myocardial infarction [3]. Rehabilitees with obesity also experience events related to CVDs earlier, have longer years lived with CVD disability, and have shorter average life span [4].

CVDs are a leading cause of premature mortality worldwide. The incidence of CVDs has doubled between 1990 and 2019. Although CVDs can cause death, many rehabilitees continue to live with these conditions for many years [5]. The years lived with disability represent the number of years people live with a disease while experiencing a reduction in functional capacity. In 2019, the total years lived with disabilities attributed to CVDs amounted to 37 million years, reflecting a 106% increase from 1990 to 2019 [6].

Overweight or obesity is typically measured with BMI [1]. However, BMI does not differentiate fat and lean mass, reflect metabolic health, or consider body fat distribution [1]. Therefore, waist circumference (WC), or body fat percentage, might be more accurate for the evaluation of the perception of adiposity and obesity-related risks [1]. In every BMI category, the larger the WC, the higher the morbidity and mortality risk [1].

WC is a measurement to assess abdominal obesity [7,8]. It is a significant predictor of health [9] and cardiometabolic diseases [10]. High WC is a robust indicator of abdominal body fat, and thus is associated with CVDs [4], cardiovascular mortality [7], and recurrent atherosclerotic CVD, especially among men [11]. Increased WC is also associated with hypertension [9,12], diabetes mellitus, hypercholesterolemia, joint and low back pains, hyperuricemia, and obstructive sleep apnea syndrome [9]. Among people with CVDs, decreasing WC is an important target to prevent recurrence [11], morbidity and mortality [7], as well as decrease other well-known risk factors [7].

Weight loss is a complex and multidimensional issue, and therefore, obesity management needs comprehensive and multidimensional interventions [13]. Weight loss interventions should be individualized, multidisciplinary, and focused on realistic goals [14]. In the study by Rogerson et al [15], the rehabilitees experienced weight loss as an enduring and ongoing task. Weight loss involves physical, cognitive, behavioral, social, and environmental dimensions, which can, at the same time, either assist or inhibit weight loss. It has been stated that weight loss interventions concentrating only on the diet and increasing physical activity (PA) usually fail to achieve significant weight loss [16]. An individually designed diet, PA, and behavioral lifestyle interventions should be implemented to improve outcomes and adherence to weight loss interventions [17]. Psychological factors, such as motivation and self-efficacy, as well as behavioral factors, such as eating behavior, can affect the success of weight loss [18].

Technology-based weight loss interventions among rehabilitees with obesity have been studied widely [19-21], but the number of studies concerning technology-based weight loss interventions among cardiac rehabilitees is limited. However, single studies have been implemented. For example, den Uijl et al [22] found that the rehabilitees in cardiac rehabilitation involving activity trackers and behavioral change techniques lost significantly more weight than rehabilitees in standard cardiac rehabilitation. Technology-based interventions have been studied in preventing CVDs [23-25], in improving risk factors of CVDs [26], and in self-management of CVDs [27]. Positive results have been obtained in reducing CVD outcomes [25] and risk factors [24,26], as well as in the prevention and management of CVDs [23]. In health care, the role of artificial intelligence (AI) is increasing, for example, in medical image analysis, clinical decision support, and analyzing data from technological devices such as smartwatches [28]. In cardiac rehabilitation, AI may support the individuality, engagement, motivation, and self-direction of a cardiac rehabilitee [29,30]. In addition, AI-based interventions may be used in predicting CVDs, which enhance preventive strategies [31].

A systematic review, meta-analysis, and meta-regression study [32] found that web- and mobile-based interventions reduced BMI, WC, and body fat percentage. However, the interventions of the included studies were heterogeneous, and it was difficult to identify the real effects of technology-based interventions on body composition. Only 2 studies of 30 included cardiac rehabilitees, and 2 studies involved rehabilitees with cardiovascular risk factors [32]. A cluster randomized trial (CRT) study among cardiac rehabilitees [33] found that remote technology (Fitbit Charge HR activity monitor and Movendos mCoach) brought added value to WC reduction compared with conventional identical multi-professional rehabilitation without remote technology, but the reasons for this difference are unknown. It is important to obtain diverse and comprehensive knowledge about weight loss interventions among cardiac

XSL•FO
RenderX

rehabilitees and the factors influencing weight management in order to improve cardiac rehabilitation and the functional ability of cardiac rehabilitees in the future.

Objective

This mixed method study aimed to examine what biopsychosocial factors predict WC reduction and what factors arise from the speech of cardiac rehabilitees during 12-month cardiac rehabilitation. Based on the quantitative and qualitative results, an integrative research-based model of biopsychosocial factors and themes for weight management is created.

Methods

Study Design

This study (registration number ISRCTN61225589) was a mixed method study with a CRT and a qualitative study. The study design followed the convergent design, where the quantitative and qualitative data were collected at the same time, independently analyzed, and finally merged to combine the results [34].

The duration of the cardiac rehabilitation, implemented in a Finnish rehabilitation center, was 12 months in total. The measurements and interviews were carried out during in-rehabilitation periods at baseline, at 6 months, and at 12 months.

Recruitment of Rehabilitees

A total of 59 rehabilitees were recruited from the rehabilitees attending a coronary heart disease rehabilitation course between September 2015 and May 2017. Participation in a course, which was funded by the Social Insurance Institution of Finland [35], required a physician's referral. In Finland, rehabilitation courses for various diseases are available to all individuals diagnosed with the condition upon a physician's referral. The rehabilitation courses are nationwide, not limited to a single well-being services county. Generally, rehabilitees can choose the rehabilitation center whose courses they want to participate in. In this study, all applicants to a specific rehabilitation center who met the inclusion criteria were enrolled as participants. The eligibility criteria for cardiac rehabilitation in Finland are being older than 18 years, a doctor's statement recommending rehabilitation, and a diagnosed cardiac disease based on an appropriate diagnostic assessment. Further requirements include the following: the rehabilitee has limitations in work, study, or functional capacity caused by the disease, which impairs the ability to manage in everyday environments. The rehabilitee needs to have sufficient independent functional capacity to move in the rehabilitation environment. In addition, the rehabilitee's initial treatment period has been completed, and the condition is stabilized [34].

The eligibility criteria for this study were being older than 18 years, having a diagnosed coronary heart disease, and independent basic-level management of IT and remote technology applications. To ensure the capability of using technology, the exclusion criteria included serious musculoskeletal disorders, cognitive diseases, and memory diseases that affect essential functional abilities. All rehabilitees fulfilled the inclusion criteria.

The study was conducted in a Finnish Rehabilitation Center, where rehabilitees came from all over Finland. The researchers divided an incoming group into 6 groups. The 2 types of groups were a conventional cardiac rehabilitation group with added remote technology (experimental) and a conventional cardiac rehabilitation group (reference) without added remote technology. The technologies used in this study were Movendos mCoach internet software (Movendos Ltd) and Fitbit Charge HR (Fitbit) accelerometers. The remote technology intervention clusters and reference clusters started in autumn (September to November), winter (December to February), and spring (March to May). The study design has been described in more detail in previously published articles [33,36,37].

Rehabilitees

The mean age of the rehabilitees (n=59) was 60 (SD 6, range 41-66) years. A total of 77% (n=23) of the experimental group, and 86% (n=25) of the reference group were male. The rehabilitees had undergone a coronary angioplasty or coronary artery bypass 3-12 months before the study. At baseline, there were no statistically significant differences in the demographic and clinical characteristics between the experimental and reference groups.

Randomization

Randomization was accomplished within the rehabilitation groups with sealed envelopes. A researcher from the gerontological center of the University of Jyväskylä, who was outside the research group, accomplished the randomization 3 times for 2 consecutive groups, under the supervision of 2 researchers (TS, Heikki Kivistö). A total of 59 rehabilitees were randomly assigned in pairs into (1) the remote technology intervention group (experimental group; n=29) or (2) the reference group (n=30). At the beginning of the rehabilitation, the researchers (TS and Heikki Kivistö) informed the rehabilitees about their rehabilitation group. The rehabilitation was implemented in the rehabilitees' natural rehabilitation environment; it was not possible to blind the rehabilitees and caregivers due to the nature of the rehabilitation intervention and for ethical reasons. The outcome assessors were not blinded to the intervention, but the same educated persons who were not part of the research group performed the anthropometric measurements (a nurse specialized in heart diseases) and the 6-minute walk test (6MWT; a physiotherapist, specialized in heart diseases). The assessors came from outside the research group. The statistician (TP) was blinded.

Ethical Considerations

This study was approved by the Ethics Committee of the Central Finland Health Care District (Dnro: 12 U/2015) on October 15, 2015. After randomization, the rehabilitees signed a written consent form about their participation in the study at the rehabilitation center. Participating in the study was voluntary, and rehabilitees were allowed to withdraw from the rehabilitation program at any time. All collected data were anonymized to ensure participant privacy and confidentiality, and all identifiable information was removed from the data. The

XSL•FO
RenderX

rehabilitees did not receive any compensation for their involvement in the study. In Finland, rehabilitation is free for the rehabilitees, including free accommodation, program activities, and meals. In addition, rehabilitees who are employed receive a rehabilitation allowance for the loss of income due to the absence from work.

Intervention

The rehabilitation courses aimed to promote the rehabilitees' biopsychosocial functional ability and work ability [35]. Evidence-based rehabilitation methods and the Finnish Current Care Guidelines [38] were applied. Rehabilitation consisted of three 5-day in-rehabilitation periods in the rehabilitation center. Between these in-rehabilitation periods, the rehabilitees had two 6-month self-rehabilitation periods, when they followed home exercise programs in their own living environment.

Both the experimental and reference groups received conventional cardiac rehabilitation for 12 months, which included, for example, multidisciplinary rehabilitation, medical examination, physiotherapy, aqua therapy, gym training, and group discussions with a nutritionist, social worker, physiotherapist, psychologist, and physician. The rehabilitees received multi-professional information and pamphlets concerning, for example, CVDs and the management of daily activities, such as diets, PA, social security benefits, self-care, and self-rehabilitation while living with CVDs. From the experimental group rehabilitees, the researcher (Heikki Kivistö) collected individual and group-based PA information, stored in the database of Fitbit Charge HR (Fitbit) accelerometers, and sent the data to the physiotherapists. Based on that information, the rehabilitees received feedback on PA during the in-rehabilitation periods. The accelerometers were designed to be worn daily. After the intervention, all rehabilitees received personal information on PA and other biopsychosocial outcomes in paper reports. The only difference being the use of technology, the studied periods were equal in the experimental and reference groups, and the researchers (TS, Heikki Kivistö) from the University of Jyväskylä were in equal contact by phone with all rehabilitees during the different phases of the study.

The only difference between the reference and experimental groups was the use of technology. The reference groups did not have any technical devices, but they used printed materials and self-monitoring with paper and pen. The experimental groups used technological devices in monitoring their progress, communicating with the instructors, goal setting, self-monitoring, and giving feedback. The interventions have been described in more detail in our previous article [33].

Outcome Measurements

This study used both the primary and secondary outcomes [33,36] of the study project. The primary outcome was a PA assessment [36], and the secondary outcomes were based on the biopsychosocial model: 6MWT, the World Health Organization Quality of Life Brief Form (WHOQOL-BREF) questionnaire, weight, BMI, WC, age, and gender. Additional secondary outcomes included the rehabilitees' experiences collected with interviews. The outcome measurements were collected at baseline and after 6 and 12 months in the rehabilitation center.

6MWT is a reliable and valid field test [39] in which a rehabilitee walks at a self-selected pace as long as possible on a flat surface for a 6-minute period [40]. The heart rate, distance walked, and the symptoms were recorded. Anthropometric measurements included body mass, BMI, and WC. Body mass was measured by using a calibrated floor scale. The rehabilitees wore light clothes, such as T-shirts and shorts or tights, and they were without shoes. BMI was determined by dividing body mass by the height square in meters [41], and WC was measured on bare skin at the midpoint between the lowest rib and the iliac crest [7,42]. The WHOQOL-BREF questionnaire was used to measure quality of life. The questionnaire was performed on a paper format, and the researchers were responsible for completing it. The WHOQOL-BREF contains 26 questions. One question is related to overall health, and another question to overall quality of life. The other 24 questions are divided into 4 domains: physical (7 items), psychological (6 items), social (3 items), and environmental (8 items). Higher scores reflect a higher quality of life [43].

Interviews

Semistructured focus group interviews were performed in the rehabilitation center at every follow-up point. The interview themes were formed by a group of researchers (TS, Arja Piirainen), and the interviews were conducted by one researcher (Heikki Kivistö). The interviews at the 6-month follow-up point were themed around experiences of rehabilitation and technology in the rehabilitee's ordinary life and changes related to lifestyle. Additional themes concerned the personalization of rehabilitation and cooperation during rehabilitation. At the 12-month follow-up point, in addition to the previous themes, the interviews inquired about remote guidance and feedback, and meaningful experiences during rehabilitation. The full interview guides are presented in Multimedia Appendix 1. The interviews were implemented on the principle of an open interview, considering the themes. The interviews were recorded, and the recorded data were transcribed. Altogether, this study examined 312 pages of transcribed data, written in font size 12 and line spacing 1.0, including 4 pages of discussion about weight management in the experimental and reference groups. Saturation was not fully achieved because this study included rehabilitees from the rehabilitation course, where the interviews were part of the rehabilitation program. The schedule for all appointments was planned by the Social Insurance Institute of Finland, and the researchers had no possibility to extend the interviews. However, the rehabilitees in all target groups discussed the same themes, and no new themes emerged. In addition, several rehabilitees mentioned at the end of the focus groups that they had nothing further to add.

At the 6-month follow-up point, altogether, 47 rehabilitees participated (38 men and 9 women), and at the 12-month follow-up point, 30 rehabilitees (21 men and 9 women) participated in the interviews. Table 1 presents the number of rehabilitees in the focus groups.

XSL•FO
RenderX

Table 1. Group interviews with cardiac rehabilitees at 6 and 12 months follow-up points during their rehabilitation program.

	6 months		12 months	
	Men, n	Women, n	Men, n	Women, n
Group 1	7	1	2	1
Group 2	5	3	5	3
Group 3	7	2	7	2
Group 4	8	0	2	0
Group 5	7	1	3	2
Group 6	4	2	2	1
Total	38	9	21	9

Quantitative, Qualitative, and Integrative Analyses

Statistical Analysis

Multiple linear regression was the primary method used for quantitative analysis. Before performing that, the Shapiro-Wilk test was performed to ensure the normality of the data. Multicollinearity was tested with a variance-inflating factor. After these tests, 2 separate dependent variables were analyzed using multiple linear regression: WC change from 0 to 6 months and from 0 to 12 months.

Multiple linear regression was performed with R Studio software (version 2024.12.0+467). First, all independent variables were tested with a dependent variable (WC change). The independent variable with the highest *P* value was iteratively removed in a stepwise backward elimination process. Finally, all statistically significant independent variables were included in the final model. If $P<.05$, predictors were considered statistically significant. Separate regression models were constructed for the 0-6-months and 0-12-months. Positive values indicate reductions, as the change variables (0-6 months and 0-12 months) were computed by subtracting the follow-up measurement from the baseline, meaning that baseline values were higher than at follow-up.

The interaction between the experimental group and 6MWT and the interaction between the reference group and 6MWT were tested in both models. The interaction was included in the model because, in the analysis of individual factors (0-6 months), it seemed significant and interesting. In addition, previous studies have found a negative correlation between 6MWT and body weight or BMI [44,45]. However, previous studies have not examined the interaction between factors related to weight loss and biopsychosocial factors, although the interaction seems statistically and clinically relevant from the point of view of the complex topic.

Qualitative Analysis

Thematic analysis of the qualitative data aimed to discover what the rehabilitees talk about their weight management. According to Braun and Clarke [46], one option for thematic analysis is to provide a detailed and nuanced account of one theme, or a group of themes within the data. Typically, this is related to a specific research question or area of interest. This thematic analysis focused on weight management themes, was inductive, and strongly linked to data [46].

Thematic analysis comprises six phases: (1) familiarizing with the data, (2) generating initial codes, (3) searching for themes, (4) reviewing themes, (5) defining and naming themes, and (6) producing the report [46,47]. In this study, researcher triangulation (HL, HK, TS) was used in every phase of the thematic analysis to ensure the quality of the qualitative analysis. In the first phase, the researchers became familiar with the data by reading it carefully and repeatedly. In the second phase, the data was coded. In the third phase, the codes were combined, compared, and analyzed. The analysis was inductive, and the themes were formed from the data (Table 2). In the fourth phase, the themes were processed by reviewing the coded data and by deciding whether the codes formed a coherent pattern or whether they would need reworking. After that, the researchers evaluated the themes in relation to the entire data set and decided whether the themes fit into the thematic map reflecting the meaning of the data. In this phase, the researchers reread the data to examine the themes and recoded the data arising from the newly created themes. In the fifth phase, the researchers created a definition and description of each theme, and in the final phase, they wrote the final analysis and description of the findings. At the initial phase, AI was used for language proofreading.

XSL•FO
RenderX

Table 2. An example of generating themes in qualitative analysis.

Quotes	Coding	Subthemes	Themes	Group and pages
"I've even gotten my wife exercise quite well."	Has exercised with wife...	Weight management together	The meaningful factors related to weight management	G3, p. 15
"She's lost more weight than me."	...and wife has lost more weight than husband	Weight management together	The meaningful factors related to weight management	G3, p. 15
"It would be a personal (diet). So it would support everyone's exercise, and generally help with losing weight as well."	Personal diet in supporting exercising and weight loss	Personal diet in supporting exercise, weight loss and rehabilitation	The meaningful factors related to weight management	G3, p. 17
"But just that it would be personal and concrete – that would support this thing much more strongly."	Personal diet in supporting the rehabilitation	Personal diet in supporting exercise, weight loss and rehabilitation	The meaningful factors related to weight management	G3, p. 17
"Five days, what you get here. It's kind of... Should be more like rehabilitation, or kind of exercise."	There should be more physical activity in the rehabilitation	Increasing the amount of physical activity in rehabilitation to prevent weight gain caused by prolonged sitting	The meaningful factors related to weight management	G5, p. 19
"Well, you could send a message there, like why people are gaining weight. It's because we sit too much here. Yeah. We should just exercise more, and then..."	Sitting in the rehabilitation increases weight	Increasing the amount of physical activity in rehabilitation to prevent weight gain caused by prolonged sitting	The meaningful factors related to weight management	G5, p. 19

Integrative Analysis

In the integrative analysis, the main results from the quantitative and qualitative analyses were integrated. The integrative analysis was made following the steps of Skamagki et al [34]. Joint display was used in the first phase to organize the integration of quantitative and qualitative data. In the second phase, the linkages between quantitative and qualitative findings were defined. In the third phase, establishing relationships, the researchers looked for inconsistencies, alignments, or conflicting findings in the relationships between quantitative and qualitative data [34]. The fourth phase included interpreting and reporting the results.

Results

Overview

The results are presented in three phases: (1) quantitative, (2) qualitative, and (3) integrative study results. The results consider the outcomes from the experimental and the reference group, as well as the differences between them identified by both analysis methods. The results were used to develop a research-based model related to weight management and cardiac rehabilitation.

Biopsychosocial Factors Predicting WC Reduction

In multiple linear regression, group allocation to the experimental and the reference groups significantly predicted WC reduction (P=.007), with the experimental group showing a greater reduction (b=2.5 cm) compared with the reference group. A significant interaction between the reference group and 6MWT was observed (P=.04): in the reference group, greater improvement in 6MWT was associated with a greater reduction in WC. No such association was observed in the experimental group (Table 2).

In addition, baseline WC was marginally associated with a greater reduction of WC (P=.05) in the 0- to 6-month analysis. According to the model, each 1 cm higher baseline WC was associated with a 0.07 cm greater reduction in WC.

None of the independent variables significantly predicted WC reduction in the 0-12 month analyses. For comparison, Table 3 also presents the 0- to 12-month results using the same model that was found to be optimal in the 0- to 6-month analysis (Tables 3 and 4).

Table 3. Results of multiple linear regression on waist circumference (WC) at 0-6 and 0-12 months.

Predictor	Estimate (0-6 months)	*t* test (*df*)	*P* value	Estimate (0-12 months)	*t* test (*df*)	*P* value
Experimental group (vs reference group)	2.52	2.83 (42)	.007	2.88	1.35 (21)	.18
6MWT[a]	–0.02	–1.91 (42)	.06	–0.06	–1.55 (21)	.14
WC at baseline (cm)	0.07	1.98 (42)	.05	0.06	1.38 (21)	.18
Experimental * 6MWT[b]	0.04	2.16 (42)	.04	0.05	1.29 (21)	.21

[a]6MWT: 6-minute walk test.

[b]Experimental * 6MWT is an interaction term between the intervention and 6MWT.

XSL•FO
RenderX

Table 4. Regression model fit and predictive accuracy statistics.

Metric	0-6 months	0-12 months
R^2	0.23	0.19
Adjusted R^2	0.16	0.03
F test (*df*)	3.16 (4, 42)	1.22 (4, 21)
P value	<.02	.33
Residual standard error	2.64	2.91

Themes About Weight Management From the Talk of the Experimental Group

Three themes arose from the talk of the rehabilitees in the experimental group: the meaningful factors related to weight management, the goal-oriented approach to weight loss, and the monitoring of body composition. The group numbers and intervention follow-up points are presented in brackets after the quotations.

The talk about the meaningful factors related to weight management included the following subthemes: weight management together, personal diet instructions, and increasing PA in rehabilitation. The rehabilitees hoped for a personal diet which would support their weight loss, rehabilitation, and PA, as stated in the following example from a group interview:

> *It would be personalized (diet). So, it would still support everyone's physical activity and. And in general, support weight loss.* [Group 3; 6 months]

In addition, the rehabilitees talked about the amount of sitting during rehabilitation, resulting in a weight increase. Therefore, the rehabilitees hoped for more PA during the rehabilitation, as in the following example:

> *Well, you could send a message there asking why the fat keeps increasing. Because we sit here too much when we should... yeah... we should just move more, and then...* [Group 5; 12 months]

The goal-oriented approach to weight loss theme included the following subthemes: weight loss goals during rehabilitation, benefits of weight loss in lowering blood pressure, and the positive changes in muscle mass and body fat percentage. The rehabilitees mentioned that many rehabilitees aim for weight loss during rehabilitation, as in the following example:

> *Especially the goals—probably many have set a goal like 'I want to lose this much weight.* [Group 3; 6 months]

In addition, the rehabilitees remarked that as a result of weight loss, blood pressure decreases. This is illustrated by the following quote:

> *And of course, when the weight drops, it lowers blood pressure and everything else, and...* [Group 3; 6 months]

The monitoring body composition theme included 5 different perspectives. First, the rehabilitees mentioned that the weight remained the same because muscle weighs more than fat. Second, they discussed the results of their earlier body composition measurements. Third, they expressed a wish that measuring body composition would be included in the rehabilitation program. Fourth, they talked about their doubts about the reliability of the body composition measurement, and fifth, the positive changes in their body composition. This is illustrated by the following quote:

> *Why don't they measure at the beginning and the end of the Social Insurance Institute of Finland course? That would tell the truth, then.* [Group 2; 12 months]

Themes About Weight Management From the Talk of the Reference Group

Two themes arose from the talk of the rehabilitees in the reference group: motivation for change and unstable weight management. The motivation for change theme comprised 2 subthemes: the motivating effect of the disease on weight management and lifestyle changes. The following quote illustrates how the rehabilitees talked about the effect of a disease on motivation:

> *Yeah, the motivation came from the heart attack, so it sort of started from there. It was a bit like my own desire to lose some weight and pay attention to what I'm eating and so on.* [Group 4; 6 months]

The rehabilitees mentioned motivating issues that affect rehabilitation, such as a healthy body, family, and the ability to work. This is illustrated by the following example:

> *Well, the motivation comes from wanting to get this, uh, body in shape so that I could be around as long as possible with my family and spend time with the kids, and, you know, do a bit of work on the side too.* [Group 4; 6 months]

The second theme, unstable weight management, contained contradictory discussions about weight change. Some rehabilitees achieved positive changes in their weight, while others did not report any changes in their weight, which disappointed them. An example is presented in the following quote:

> *I've lost a little weight and, and my endurance was better, and…* [Group 1; 12 months]

Some rehabilitees gained more weight during the rehabilitation, and they talked about the hope of losing the weight they gained. This is illustrated by the following examples:

> *Eating habits have changed, but the weight hasn't gone down. That's disappointing, for sure.* [Group 1; 12 months]

XSL•FO
RenderX

I've of course had a bit of a setback, I've gained a couple of kilos. But maybe it'll come off during the fall. [Group 1; 12 months]

Integrative Results of the Quantitative and Qualitative Data

The integrative results of the quantitative and qualitative results are presented in Figure 1. The left side of the figure describes the results of the experimental group, and the right side describes the results of the reference group. The center of the figure describes the quantitative results at 0-6 and 0-12 month follow-up points. The sides of the figure describe the themes of the qualitative data.

The integrative model emphasizes that the experimental and reference groups differed in quantitative and qualitative results in terms of factors related to weight management. Whether the rehabilitee belonged to the experimental or the reference group was important in this study. The group was associated with WC reduction at 0-6 months (the center of Figure 1). The rehabilitees in the experimental group talked more about their weight management (left side of Figure 1) compared with the rehabilitees in the reference group (right side of Figure 1), and they seemed to have advanced in the process of change. The rehabilitees talked about elements of goal-oriented weight management, and they reflected on the changes they had achieved. In the reference group, improved 6MWT distance was associated with WC reduction at 0-6 months. In this group, the talk was more of a motivational speech, and the rehabilitees' body composition changes were unstable. Furthermore, body composition changes arose from the talk of both groups. In the experimental group, the changes in weight management were positive, but in the reference group, the changes were unstable.

Figure 1. Integrative results with the experimental group on the left and the reference group on the right. 6MWT: 6-minute walk test; WC: waist circumference.

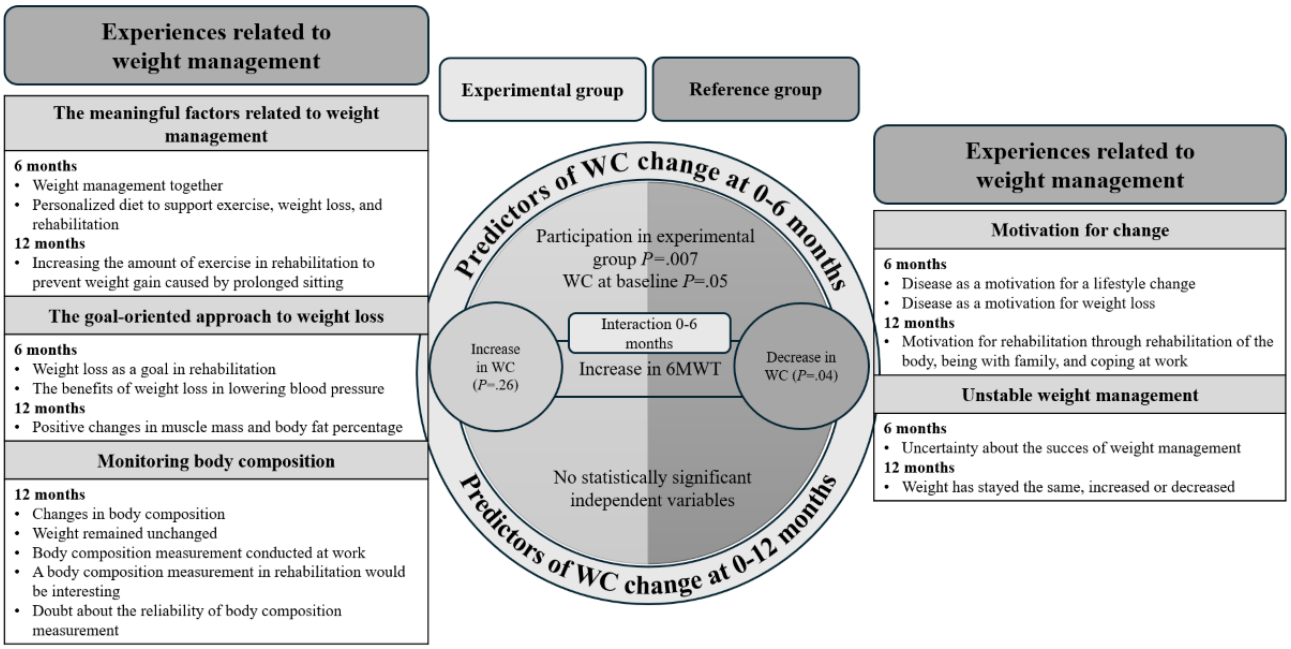

Discussion

Principal Findings

This mixed method study aimed to examine, during a 12-month cardiac rehabilitation, which biopsychosocial factors are related to WC reduction and what themes of weight management are meaningful for cardiac rehabilitees. A main finding was that allocation to the experimental group predicted greater WC reduction compared with the reference group. The integrative model indicated that the experimental and reference groups differed in the quantitative and qualitative results relating to weight management. Adding remote technology to conventional rehabilitation seemed to explain the WC reduction. The talk of rehabilitees was characterized by behavioral change and reflection on weight management. In the reference group, which did not use remote technology, an improved 6MWT distance in meters was associated with decreased WC. The rehabilitees' talk dealt with motivation for weight management and unstable changes in body composition.

Our findings indicate that technology-based interventions are associated with WC reduction. However, other biopsychosocial variables were not associated with WC reduction. A similarity between the groups was found when a larger WC at baseline might predict significant WC reduction at 0-6 months (P=.05). This is an important finding as decreasing WC is essential in weight management because abdominal obesity is associated with CVDs [3]. This finding is in line with previous studies of the importance of baseline value to weight management changes during intervention [48-51]. Iwayama et al [48] found that a higher baseline WC was associated with a larger reduction of WC. Other studies have examined the association between early weight loss and long-term weight loss. Al-Abdullah et al [49] discovered that higher baseline weight and early weight loss were predictors of successful weight loss during an intervention. Höchsmann et al [50] and Farage et al [51] observed that initial weight loss predicted long-term weight loss. However, only one of these previous studies considered WC [48], and none of the studies considered cardiac rehabilitees.

XSL•FO
RenderX

This study found a statistically significant interaction between increased walking distance in 6MWT and decreased WC in the reference group, but not in the experimental group. The change in 6MWT was not statistically significant on its own, meaning that within the groups, the change in the walking distance was not associated with the WC change. One possible explanation for this finding is that the rehabilitees in the experimental group improved their nutrition or other factors that could not be captured with the quantitative measurements. Increased WC has been associated with decreased walking distance in 6MWT [52,53] among rehabilitees with rheumatoid arthrosis [52] and type 2 diabetes [53]. Among cardiac rehabilitees, Lenasi et al [54] found that age, WC, and New York Heart Association classification were predictors of 6MWT in rehabilitees with coronary artery disease. Similarly, Pepera et al [55] found that BMI >25 kg/m^2 and age were independent predictors of poor 6MWT performance in rehabilitees with chronic heart failure. However, Sharma et al [56] studied healthy individuals and discovered a relation between 6MWT with BMI and waist-to-hip ratio. They found that when BMI increased, the 6MWT walking distance also increased. However, when the waist-to-hip ratio increased, the 6MWT walking distance decreased. Results regarding the interaction between 6MWT and other variables are contradictory, and more studies are needed to determine the relationship between 6MWT and weight management-related outcomes.

In this study, the significant predictors of WC change were statistically significant at the 0-6 month follow-up point, but none of the predictors were statistically significant at the 0-12 month follow-up point. Previous studies have found that the effects of interventions are initially more substantial but tend to diminish over the long term. In addition, typically engagement wanes over time as initial enthusiasm diminishes. Whiteley et al [57] found that during 6 months, engagement on technology-based weight loss intervention decreased 33%. In addition, Liljeroos et al [58] and Beleigoli et al [59] found that technology-based interventions were more effective in the short term, but the changes did not last in the long term.

The rehabilitees in the experimental group were more advanced in their weight management reflection compared with the rehabilitees in the reference group. The experimental group discussed the factors related to weight management, a goal-oriented approach to weight loss, and monitoring body composition. They talked about their body composition changes objectively, and their discussion was goal-oriented. The talk in the reference group was rather motivational, including motivation for change and unstable weight management. In line with the transtheoretical model, the rehabilitees in the reference group are in the preparation phase, in which the rehabilitees intend to take an action toward a change [60]. The rehabilitees in the experimental group are in the action phase. They have made changes, and their behavior change has been translated into action [61].

Previous studies have not compared the views of the rehabilitees in experimental and reference groups where the only difference was the use of technology. However, previous studies have examined the rehabilitees' experiences of technology in weight management. Our results have similarities with previous studies. Lambert et al [62] found that an e-coach was beneficial in motivating the rehabilitees for a behavior change at the initial phase. However, the engagement diminished over time. Følling et al [63] found that activity monitors can strengthen or undermine the attempt at weight loss. Some rehabilitees mentioned that activity monitors motivated them to maintain PA and diet, providing proof of their effort. However, some rehabilitees felt that activity monitors highlighted their failure to change their lifestyle.

The rehabilitees in the experimental group discussed personalized diets in supporting weight loss, exercise, and rehabilitation. Previous studies and guidelines have stated that individually tailored weight loss programs can be important in weight loss [21,64,65]. In addition, the rehabilitees in this study mentioned the importance of social support in weight loss. Previous studies have found that social support from family and friends [57,59] as well as peer support is important in the weight loss process [62,66,67]. In addition, both groups talked about the effect of being overweight on diseases. The rehabilitees in the experimental group discussed the benefit of weight loss in lowering blood pressure, and the rehabilitees in the reference group discussed the motivation for weight loss arising from the cardiac disease. This finding is consistent with the study by Spreckley et al [68], who found that intrinsic motivators include the desire to improve health and the effort to improve weight-related conditions.

The experimental group talked about a goal-oriented approach to weight loss and weight loss as a goal in their rehabilitation. The rehabilitees in the reference group did not discuss their weight loss goals. The goal of weight loss might be imposed from outside, for example, by health care professionals, and people with obesity might assume that their goal should be weight loss. According to the study by Liu et al [69], external motivation is the key to initiating dietary changes. However, internal motivation is important for sustaining the behavioral change. This is in line with the self-determination theory, which suggests that intrinsic motivation, which is the most autonomous form of motivation, comes from a rehabilitee's own interests, enjoyment, and satisfaction. Extrinsic motivation comes from outside (for example, from health care providers, social media, and society). When behavior is driven by external forces, long-term changes are typically unlike [70], whereas intrinsic goals result in improved self-regulation and long-term outcomes [71]. It might be important that the initial goal of weight loss comes, for example, from health care professionals, but for long-term changes, internal motivation is essential.

This study is the first mixed method study exploring technology-based weight loss intervention in cardiac rehabilitation. The study focuses on WC and incorporates several aspects that have not been studied before. For example, Dandeneau et al [72] performed a mixed method weight loss study on cardiac patients, but concentrated on adherence to the weight loss program as well as the strengths and limitations of the program. Liljeroos et al [58] conducted a mixed method study on the effects of a mHealth tool for rehabilitees with heart failure. However, they concentrated on the effectiveness and experiences of the mHealth tool with scales and questionnaires.

XSL•FO
RenderX

Leung et al [73] examined, with mixed method study, the psychological factors of dietary and PA adherence in the early weight maintenance phase. However, they concentrated on participants with overweight and obesity, not cardiac rehabilitees.

In health care, obesity-related measurements are universal, and they categorize the rehabilitees to certain classes. This study also used commonly used measures for assessing weight loss. However, these measurements can be very embarrassing for the rehabilitees [74,75]. In behavior change interventions, these measurements are not necessarily optimal because of the stress and anxiety they may cause to the rehabilitees. According to Phelan et al [76], a welcoming and less threatening environment can be created by reducing focus on body weight and concentrating on obesity-related diseases and conditions, as well as encouraging behaviors that improve health and well-being. Anttila et al [37] formed 4 biopsychosocial profiles in technology-based cardiac intervention. The "feeling outsider" group had a high-risk behavior related to inactive behavior and being overweight. The "being uninterested" group had low PA levels and a poor self-assessed physiological quality of life. Both groups had large WC compared with the recommended value, and 6MWT was below the recommended value. The "reflecting benefit" group had high self-efficacy for exercise and an interest in health technology. The "enthusiastic users" group had high self-determination in exercise behavior but lacked self-efficacy in PA. The WC of the rehabilitees in these groups was closer to the recommended value, and their 6MWT was very close to the recommended value [37]. Individual measurements that do not necessarily concentrate on weight should be developed. This view has not been studied, so future studies should evaluate different measurements for overweight rehabilitees.

Because CVDs are a public health challenge and obesity is a risk factor for CVDs [2], weight management should be included in cardiac rehabilitation. However, it has not been a primary focus in there [77]. Technology may increase the engagement in cardiac rehabilitation [78] and lifestyle changes [79] as well as weight management [80]. Therefore, it should be considered to include technological components in cardiac rehabilitation and weight management. Health care systems are undergoing transformation by increasing the usability of technological devices. With technological devices, it is possible to achieve clinical and environmental benefits such as improving disease management, adherence, and access to health care. From an environmental view, digital health technologies decrease carbon emissions, waste generation, and energy consumption [81].

Our studies [32,33] produce gradually deepening information on the effectiveness of the added value of remote technology. Our previous systematic review with meta-analysis and meta-regression highlighted that previous RCT studies have compared highly heterogeneous experimental and control groups [32]. They have not compared experimental and control groups where the only difference was the use of technology. According to the findings of our CRT study [33], rehabilitation interventions using technology decreased WC more noticeably compared with the reference group without technology. The results from this study might indicate that technology brings added value to conventional rehabilitation and may support weight management, but it requires more studies in the future. To our best knowledge, only a few studies have compared experimental and control groups where the only difference is the technology use [25,82]. However, other studies have compared technology-based interventions to very heterogeneous control groups, which do not consider the added value of technology.

Future studies should examine the added value of technology by comparing interventions that are otherwise similar, but one group uses technology, and the other group does not. In addition, future studies should investigate the predictors of WC reduction and weight management, as well as weight management among cardiac rehabilitees.

In the future, AI is predicted to have a major role in smart health care. Smart health care offers several benefits by enhancing diagnostic accuracy, automating clinical workflows, and enabling personalized treatment strategies [28]. AI apps may support the individuality of cardiac rehabilitees' rehabilitation, as well as their engagement, motivation, and self-direction [29,30]. In addition, AI-related solutions may enhance interaction between health care professionals and rehabilitees, support the monitoring of rehabilitation progress [29,30,83], and provide information on, for example, arrhythmias, heart rate variability, and PA [84,85]. However, research related to AI and rehabilitation is still largely limited to descriptive literature reviews [83-87] or studies conducted with non-experimental research designs [29,30,79]. At present, there is a substantial need for more precise evidence on effectiveness—obtained through randomized controlled trial designs—related to technology, including AI, and cardiac rehabilitation [88].

Future studies should investigate AI-based weight loss interventions among cardiac rehabilitees. With AI, it could be possible to, for example, personalize the interventions, provide real-time feedback, and improve adherence [28], which are key components in weight loss interventions. In cardiac rehabilitation, AI may be used in predicting CVDs [31], support early CVD detection, and enhance preventive strategies [31]. Furthermore, explainable AI may improve trust and accountability in health care, which are essential in technology-based interventions [89].

Strengths and Limitations

This study has several strengths. First, it analyzed biopsychosocial factors broadly with multiple linear regressions and deepened the understanding of weight management with interviews. In addition, quantitative and qualitative results were combined in an integrative analysis. Second, the comprehensive cardiac rehabilitation intervention was implemented by a multiprofessional team. Third, the analyses and the randomization process were carefully implemented. The quantitative analyses were performed in collaboration with a statistician (TP), a principal investigator (TS), and a researcher (HL). The qualitative analyses used researcher triangulation throughout the process (HL, HK, and TS).

XSL•FO
RenderX

The strength of this study is that the analyses were discussed collaboratively by the entire research team, and the decisions were reached collectively. In the first linear regression analyses, BMI was included. However, BMI overwhelmed the effects of all other dependent variables. Therefore, BMI was excluded from linear regression. WC was considered more important than BMI, as previous studies have shown that in every BMI category, the larger the WC, the greater the risk of morbidity or mortality [1]. Therefore, WC was considered a more important factor than BMI in this study. One more strength of the study was its CRT design, which was integrated with a longitudinal interview component. To our knowledge, no similar mixed method studies have been conducted before.

A weakness of this study was the relatively small sample size. The protocol for cardiac rehabilitation, which is standardized by the Social Insurance Institute of Finland, and which is also responsible for rehabilitation costs, determined the sample size. The data collection took place over a 1-year period and included all cardiac rehabilitation groups participating in rehabilitation during that time. Due to the study schedule, it was not possible to collect a larger sample. Additionally, the group size was limited to a maximum of 10 rehabilitees. Therefore, the number of rehabilitees was limited. Because of the small sample size, care must be taken in generalizing the results beyond the target population. The rehabilitees in this study were residents across Finland, which increases the generalizability of this study in Finland. Globally, the results can be generalized in high-income countries. However, in low-income countries, for example, access to medical treatment is not necessarily possible. In addition, the number of cardiac rehabilitation studies is limited in low-income countries [90].

The characteristics of the rehabilitees in this study are similar to those in previous studies. For example, the mean age of the rehabilitees in this study was 60 years, and most of the rehabilitees were male. In previous studies, the mean age of the participants varied from 55 to 63 years [25,82,91-93]. In all these studies, most of the rehabilitees were male [25,82,91-93]. The higher proportion of men effects the generalizability of the results, as these results cannot be directly generalized to women. Fat distribution typically differs between women and men. In addition, biopsychosocial factors behind weight management may differ. For example, the prevalence of obesity is typically higher among women. However, men typically have more visceral fat deposits than women [94].

This study applied a qualitative thematic analysis to group interviews, which enabled the inclusion of a wider range of rehabilitees compared with individual interviews. However, focus group interviews may have affected the level of the rehabilitees' talk by decreasing the discussion about weight management. A total of 312 pages of transcribed data were included in the qualitative study, but only 4 pages reflected weight management. Personal interviews would have been more reliable and made it possible to examine the experiences of the rehabilitees at an individual level. Despite the limited scope of discussion, several themes could be distinguished from the data.

The measurements used in this study are widely used but have some limitations. WC can vary across populations and between sexes. Another limitation is that WC does not reflect the accumulation of subcutaneous and visceral fat [1]. 6MWT and WHOQOL-BREF have limitations as well. Even though 6MWT is reliable and valid, small changes in the methodology of the test may affect the results [39]. The learning effect might be associated with improved walking distance. However, 6MWT has been validated for CVDs, and it can provide information about physiological reserve and prognosis [95]. WHOQOL-BREF is a reliable and valid questionnaire, but some studies have reported a ceiling effect [96-98]. However, the reliability of these measurements improved when the same nurse performed the body composition measurements at every measurement point, and one physiotherapist performed the measurements of 6MWT at every follow-up point.

Conclusions

Both quantitative and qualitative results showed that technology-based intervention supported weight management. The integrative model emphasizes that the experimental and reference groups differed in quantitative and qualitative results in terms of factors related to weight management. The added use of remote technology to conventional rehabilitation seemed to explain the observed positive changes in WC. In addition, rehabilitees' talk reflected behavioral changes in weight management, and the rehabilitees of the experimental group were further in their process of change when compared with the reference group. Technological devices may improve the engagement of cardiac rehabilitees in lifestyle changes and weight management and thus improve the well-being of the rehabilitees. In the future, the results should be verified for use by larger research groups, and interviews should be conducted as individual interviews that can be combined with personal quantitative data to achieve more personalized results. In addition, future studies should examine AI-based weight loss interventions for cardiac rehabilitees.

Acknowledgments

The authors would like to thank the Rehabilitation Center of Peurunka, all rehabilitees, and rehabilitation professionals. The authors also thank the project researcher, Heikki Kivistö, for enabling this study and Arja Piirainen for helping with forming the interview themes. Finally, the authors would like to thank Anneli Rinnevalli for language revision. This study was supported by the Social Insurance Institute of Finland. Copilot was used to check and correct English grammar. In addition, Copilot was used to translate previous studies in Finnish and to ensure the right understanding of the studies.

XSL•FO
RenderX

Data Availability

All data generated or analyzed during this study are included in this published article.

Funding

This study was funded by the Social Insurance Institute of Finland and the JYPE research grant.

Authors' Contributions

TS and AH contributed to the conceptualization of the study. TS was responsible for data curation, funding acquisition, project administration, and provision of resources. HL, TP, HK, AH, and TS conducted the formal analysis, investigation, methodology development, and validation. Supervision was provided by TS, HK, and AH. Visualization was carried out by HL, TP, and TS. HL prepared the original draft of the manuscript, and HL, TP, HK, AH, and TS contributed to the review and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Focus group themes at 6- and 12-months follow-up points.
[DOCX File , 42 KB - jmir_v28i1e78347_app1.docx]

References

1. Rubino F, Cummings DE, Eckel RH, Cohen RV, Wilding JPH, Brown WA, et al. Definition and diagnostic criteria of clinical obesity. Lancet Diabetes Endocrinol 2025 Mar;13(3):221-262. [doi: 10.1016/S2213-8587(24)00316-4] [Medline: 39824205]
2. Cardiovascular diseases (CVDs). World Health Organization. 2021. URL: https://www.who.int/europe/news-room/fact-sheets/item/cardiovascular-diseases [accessed 2024-10-10]
3. Teo KK, Rafiq T. Cardiovascular risk factors and prevention: a perspective from developing countries. Can J Cardiol 2021 May;37(5):733-743 [FREE Full text] [doi: 10.1016/j.cjca.2021.02.009] [Medline: 33610690]
4. Powell-Wiley TM, Poirier P, Burke LE, Després JP, Gordon-Larsen P, Lavie CJ, American Heart Association Council on LifestyleCardiometabolic Health; Council on CardiovascularStroke Nursing; Council on Clinical Cardiology; Council on EpidemiologyPrevention;Stroke Council. Obesity and cardiovascular disease: a scientific statement from the American heart association. Circulation 2021 May 25;143(21):e984-e1010 [FREE Full text] [doi: 10.1161/CIR.0000000000000973] [Medline: 33882682]
5. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, GBD-NHLBI-JACC Global Burden of Cardiovascular Diseases Writing Group. Global burden of cardiovascular diseases and risk factors, 1990-2019: update from the GBD 2019 study. J Am Coll Cardiol 2020 Dec 22;76(25):2982-3021 [FREE Full text] [doi: 10.1016/j.jacc.2020.11.010] [Medline: 33309175]
6. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the global burden of disease study 2019: a systematic analysis for the global burden of disease study 2019. Lancet 2021 Dec 19;396(10267):2006-2017 [FREE Full text] [doi: 10.1016/S0140-6736(20)32340-0] [Medline: 33275908]
7. Ross R, Neeland IJ, Yamashita S, Shai I, Seidell J, Magni P, et al. Waist circumference as a vital sign in clinical practice: a consensus statement from the IAS and ICCR working group on visceral obesity. Nat Rev Endocrinol 2020 Mar;16(3):177-189 [FREE Full text] [doi: 10.1038/s41574-019-0310-7] [Medline: 32020062]
8. Fang H, Berg E, Cheng X, Shen W. How to best assess abdominal obesity. Curr Opin Clin Nutr Metab Care 2018 Sep;21(5):360-365 [FREE Full text] [doi: 10.1097/MCO.0000000000000485] [Medline: 29916924]
9. Darsini D, Hamidah H, Notobroto HB, Cahyono EA. Health risks associated with high waist circumference: a systematic review. J Public Health Res 2020 Jul 03;9(2):1811 [FREE Full text] [doi: 10.4081/jphr.2020.1811] [Medline: 32728557]
10. Durrer Schutz D, Busetto L, Dicker D, Farpour-Lambert N, Pryke R, Toplak H, et al. European practical and patient-centred guidelines for adult obesity management in primary care. Obes Facts 2019;12(1):40-66 [FREE Full text] [doi: 10.1159/000496183] [Medline: 30673677]
11. Mohammadi H, Ohm J, Discacciati A, Sundstrom J, Hambraeus K, Jernberg T, et al. Abdominal obesity and the risk of recurrent atherosclerotic cardiovascular disease after myocardial infarction. Eur J Prev Cardiol 2020 Dec;27(18):1944-1952. [doi: 10.1177/2047487319898019] [Medline: 31958380]
12. Sun JY, Ma YX, Liu HL, Qu Q, Cheng C, Kong XQ, et al. High waist circumference is a risk factor of new-onset hypertension: evidence from the China health and retirement longitudinal study. J Clin Hypertens (Greenwich) 2022 Mar;24(3):320-328 [FREE Full text] [doi: 10.1111/jch.14446] [Medline: 35188335]

XSL•FO
RenderX

13. Khattab R. Weight loss programs: why do they fail? a multidimensional approach for obesity management. Curr Nutr Rep 2024 Sep;13(3):478-499. [doi: 10.1007/s13668-024-00551-x] [Medline: 38861120]
14. Yumuk V, Tsigos C, Fried M, Schindler K, Busetto L, Micic D, Obesity Management Task Force of the European Association for the Study of Obesity. European guidelines for obesity management in adults. Obes Facts 2015;8(6):402-424 [FREE Full text] [doi: 10.1159/000442721] [Medline: 26641646]
15. Rogerson D, Soltani H, Copeland R. The weight-loss experience: a qualitative exploration. BMC Public Health 2016 May 04;16:371 [FREE Full text] [doi: 10.1186/s12889-016-3045-6] [Medline: 27142984]
16. Welsh A, Hammad M, Piña IL, Kulinski J. Obesity and cardiovascular health. Eur J Prev Cardiol 2024 Jun 03;31(8):1026-1035 [FREE Full text] [doi: 10.1093/eurjpc/zwae025] [Medline: 38243826]
17. Cornier MA. A review of current guidelines for the treatment of obesity. Am J Manag Care 2022 Dec;28(15 Suppl):S288-S296 [FREE Full text] [doi: 10.37765/ajmc.2022.89292] [Medline: 36525676]
18. Pigsborg K, Kalea AZ, De Dominicis S, Magkos F. Behavioral and psychological factors affecting weight loss success. Curr Obes Rep 2023 Sep;12(3):223-230. [doi: 10.1007/s13679-023-00511-6] [Medline: 37335395]
19. Islam MM, Poly TN, Walther BA, Jack Li YC. Use of mobile phone app interventions to promote weight loss: meta-analysis. JMIR Mhealth Uhealth 2020 Jul 22;8(7):e17039 [FREE Full text] [doi: 10.2196/17039] [Medline: 32706724]
20. Rumbo-Rodríguez L, Sánchez-SanSegundo M, Ruiz-Robledillo N, Albaladejo-Blázquez N, Ferrer-Cascales R, Zaragoza-Martí A. Use of technology-based interventions in the treatment of patients with overweight and obesity: a systematic review. Nutrients 2020 Nov 26;12(12):3634 [FREE Full text] [doi: 10.3390/nu12123634] [Medline: 33255982]
21. Protano C, De Giorgi A, Valeriani F, Mazzeo E, Zanni S, Cofone L, et al. Can digital technologies be useful for weight loss in individuals with overweight or obesity? a systematic review. Healthcare (Basel) 2024 Mar 16;12(6):670 [FREE Full text] [doi: 10.3390/healthcare12060670] [Medline: 38540634]
22. den Uijl I, van den Berg-Emons RJG, Sunamura M, Lenzen MJ, Stam HJ, Boersma E, et al. Effects of a dedicated cardiac rehabilitation program for patients with obesity on body weight, physical activity, sedentary behavior, and physical fitness: the OPTICARE XL randomized controlled trial. Phys Ther 2023 Sep 01;103(9):pzad055 [FREE Full text] [doi: 10.1093/ptj/pzad055] [Medline: 37265452]
23. Kozik M, Isakadze N, Martin SS. Mobile health in preventive cardiology: current status and future perspective. Curr Opin Cardiol 2021 Sep 01;36(5):580-588 [FREE Full text] [doi: 10.1097/HCO.0000000000000891] [Medline: 34224437]
24. Burke LE, Ma J, Azar KMJ, Bennett GG, Peterson ED, Zheng Y, American Heart Association Publications Committee of the Council on EpidemiologyPrevention, Behavior Change Committee of the Council on Cardiometabolic Health, Council on CardiovascularStroke Nursing, Council on Functional GenomicsTranslational Biology, Council on Quality of CareOutcomes Research,Stroke Council. Current science on consumer use of mobile health for cardiovascular disease prevention: a scientific statement from the American heart association. Circulation 2015 Sep 22;132(12):1157-1213 [FREE Full text] [doi: 10.1161/CIR.0000000000000232] [Medline: 26271892]
25. Widmer RJ, Collins NM, Collins CS, West CP, Lerman LO, Lerman A. Digital health interventions for the prevention of cardiovascular disease: a systematic review and meta-analysis. Mayo Clin Proc 2015 Apr;90(4):469-480 [FREE Full text] [doi: 10.1016/j.mayocp.2014.12.026] [Medline: 25841251]
26. Tegegne TK, Shifti DM, Rawstorn JC, Jansons P, Zhang Y, Daryabeygikhotbehsara R, et al. Digital lifestyle interventions for cardiovascular risk reduction: a systematic review and network meta-analysis. Health Policy Technol 2024 Aug;13(3):100879. [doi: 10.1016/j.hlpt.2024.100879]
27. Hanlon P, Daines L, Campbell C, McKinstry B, Weller D, Pinnock H. Telehealth interventions to support self-management of long-term conditions: a systematic metareview of diabetes, heart failure, asthma, chronic obstructive pulmonary disease, and cancer. J Med Internet Res 2017 May 17;19(5):e172 [FREE Full text] [doi: 10.2196/jmir.6688] [Medline: 28526671]
28. Abbas SR, Seol H, Abbas Z, Lee SW. Exploring the role of artificial intelligence in smart healthcare: a capability and function-oriented review. Healthcare (Basel) 2025 Jul 08;13(14):1642 [FREE Full text] [doi: 10.3390/healthcare13141642] [Medline: 40724669]
29. Aharon KB, Gershfeld-Litvin A, Amir O, Nabutovsky I, Klempfner R. Improving cardiac rehabilitation patient adherence via personalized interventions. PLoS One 2022;17(8):e0273815 [FREE Full text] [doi: 10.1371/journal.pone.0273815] [Medline: 36037232]
30. Marketou ME, Anastasiou I, Fourlis A, Alevizaki A, Kochiadakis G. Efficacy of an internet of things-based system for cardiac rehabilitation monitoring: insights from the IntellIoT pivotal trial in heart failure patients. Eur Heart J Digit Health 2025 Mar;6(2):293-297. [doi: 10.1093/ehjdh/ztae093] [Medline: 40110225]
31. Fitriyani NL, Syafrudin M, Chamidah N, Rifada M, Susilo H, Aydin D, et al. A novel approach utilizing bagging, histogram gradient boosting, and advanced feature selection for predicting the onset of cardiovascular diseases. Mathematics 2025 Jul 04;13(13):2194 [FREE Full text] [doi: 10.3390/math13132194]
32. Lahtio H, Rintala A, Immonen J, Sjögren T. The effectiveness of physical activity-promoting web- and mobile-based distance weight loss interventions on body composition in rehabilitation settings: systematic review, meta-analysis, and meta-regression analysis. J Med Internet Res 2022 Mar 24;24(3):e25906 [FREE Full text] [doi: 10.2196/25906] [Medline: 35323126]

XSL•FO
RenderX

33. Lahtio H, Heinonen A, Paajanen T, Sjögren T. The added value of remote technology in cardiac rehabilitation on physical function, anthropometrics, and quality of life: cluster randomized controlled trial. J Med Internet Res 2023 Apr 12;25:e42455 [FREE Full text] [doi: 10.2196/42455] [Medline: 37043264]
34. Skamagki G, King A, Carpenter C, Wåhlin C. The concept of integration in mixed methods research: a step-by-step guide using an example study in physiotherapy. Physiother Theory Pract 2024 Feb;40(2):197-204 [FREE Full text] [doi: 10.1080/09593985.2022.2120375] [Medline: 36069530]
35. Sydänkuntoutuskurssi aikuiselle [cardiac rehabilitation course for adults]. Kela. 2021. URL: https://www.kela.fi/sydansairaudet-kuntoutuskurssit [accessed 2025-04-13]
36. Hakala S, Kivistö H, Paajanen T, Kankainen A, Anttila MR, Heinonen A, et al. Effectiveness of distance technology in promoting physical activity in cardiovascular disease rehabilitation: cluster randomized controlled trial, a pilot study. JMIR Rehabil Assist Technol 2021;8(2):e20299 [FREE Full text] [doi: 10.2196/20299] [Medline: 34142970]
37. Anttila MR, Söderlund A, Paajanen T, Kivistö H, Kokko K, Sjögren T. Biopsychosocial profiles of patients with cardiac disease in remote rehabilitation processes: mixed methods grounded theory approach. JMIR Rehabil Assist Technol 2021;8(4):e16864 [FREE Full text] [doi: 10.2196/16864] [Medline: 34730548]
38. Coronary heart disease. Käypä hoito. URL: https://www.kaypahoito.fi/hoi50130 [accessed 2022-03-23]
39. Holland AE, Spruit MA, Troosters T, Puhan MA, Pepin V, Saey D, et al. An official European respiratory society/American thoracic society technical standard: field walking tests in chronic respiratory disease. Eur Respir J 2014 Dec;44(6):1428-1446 [FREE Full text] [doi: 10.1183/09031936.00150314] [Medline: 25359355]
40. Giannitsi S, Bougiakli M, Bechlioulis A, Kotsia A, Michalis LK, Naka KK. 6-minute walking test: a useful tool in the management of heart failure patients. Ther Adv Cardiovasc Dis 2019;13:1753944719870084 [FREE Full text] [doi: 10.1177/1753944719870084] [Medline: 31441375]
41. Obesity and overweight. World Health Organization. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2025-11-30]
42. Body mass index and waist circumference. Finnish Medical Society Duodecim. URL: https://www.kaypahoito.fi/nix00163 [accessed 2025-11-30]
43. WHOQOL: measuring quality of life. World Health Organization. URL: https://www.who.int/tools/whoqol [accessed 2025-11-30]
44. Correia de Faria Santarém G, de Cleva R, Santo MA, Bernhard AB, Gadducci AV, Greve JMD, et al. Correlation between body composition and walking capacity in severe obesity. PLoS One 2015;10(6):e0130268 [FREE Full text] [doi: 10.1371/journal.pone.0130268] [Medline: 26098769]
45. Beriault K, Carpentier AC, Gagnon C, Ménard J, Baillargeon JP, Ardilouze JL, et al. Reproducibility of the 6-minute walk test in obese adults. Int J Sports Med 2009 Oct;30(10):725-727. [doi: 10.1055/s-0029-1231043] [Medline: 19585400]
46. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008 Jul 21;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
47. Kiger ME, Varpio L. Thematic analysis of qualitative data: AMEE guide no. 131. Med Teach 2020 Aug;42(8):846-854. [doi: 10.1080/0142159X.2020.1755030] [Medline: 32356468]
48. Iwayama Y, Shimba Y, Viswanathan CS, Yano Y. Effect and factors associated with weight and waist circumference reductions in information and communication technology-based specific health guidance. Environ Occup Health Pract 2023;5(1):0001-OA. [doi: 10.1539/eohp.2023-0001-OA] [Medline: 40059931]
49. Al-Abdullah L, Ahern A, Welsh P, Logue J. A predictive model for medium-term weight loss response in people with type 2 diabetes engaging in behavioural weight management interventions. Diabetes Obes Metab 2024 Sep;26(9):3653-3662. [doi: 10.1111/dom.15706] [Medline: 38874091]
50. Höchsmann C, Martin CK, Apolzan JW, Dorling JL, Newton RL, Denstel KD, PROPEL Research Group. Initial weight loss and early intervention adherence predict long-term weight loss during the promoting successful weight loss in primary care in Louisiana lifestyle intervention. Obesity (Silver Spring) 2023 Sep;31(9):2272-2282. [doi: 10.1002/oby.23854] [Medline: 37551762]
51. Farage G, Simmons C, Kocak M, Klesges RC, Talcott GW, Richey P, et al. Assessing the contribution of self-monitoring through a commercial weight loss app: mediation and predictive modeling study. JMIR Mhealth Uhealth 2021 Jul 14;9(7):e18741 [FREE Full text] [doi: 10.2196/18741] [Medline: 34259635]
52. Ferreira MB, Saraiva FA, Fonseca T, Costa R, Marinho A, Oliveira JC, et al. Clinical associations and prognostic implications of 6-minute walk test in rheumatoid arthritis. Sci Rep 2022 Nov 04;12(1):18672 [FREE Full text] [doi: 10.1038/s41598-022-21547-z] [Medline: 36333405]
53. Wang E, Abrahamson K, Liu PJ, Ahmed A. Can mobile technology improve weight loss in overweight adults? a systematic review. West J Nurs Res 2020 Sep;42(9):747-759. [doi: 10.1177/0193945919888224] [Medline: 31762402]
54. Lenasi H, Novak A, Jug B, Dervišević E, Karpljuk D, Videmšek M, et al. An alternative prediction equation for evaluation of six-minute walk distance in stable coronary artery disease patients. Front Physiol 2022;13:844847 [FREE Full text] [doi: 10.3389/fphys.2022.844847] [Medline: 35432004]
55. Pepera G, Ingle L, Sandercock GRH. Predictors of the 6-minute walk test in patients with chronic heart failure. Br J Card Nurs 2015 Sep 02;10(9):454-549. [doi: 10.12968/bjca.2015.10.9.454]

XSL•FO
RenderX

56. Sharma P, Sharma S, Chaudhary V. Relations of six minute walk distance with body mass index and wait hip ratio. Int J Sci Res 2018;7(1). [doi: 10.36106/ijsr]
57. Whiteley JA, Tjaden AH, Bailey CP, Faro J, DiPietro L, Hayman LL, et al. Engagement with digital weight loss intervention components and weight outcomes. J Technol Behav Sci 2024 Dec;9(4):714-723. [doi: 10.1007/s41347-024-00388-9] [Medline: 41122737]
58. Liljeroos M, Arkkukangas M, Strömberg A. The long-term effect of an m-health tool on self-care in patients with heart failure: a pre-post interventional study with a mixed-method analysis. Eur J Cardiovasc Nurs 2024;23(5):470-477 [FREE Full text] [doi: 10.1093/eurjcn/zvad107] [Medline: 38165027]
59. Beleigoli AM, Andrade AQ, Cançado AG, Paulo MN, Diniz MDFH, Ribeiro AL. Web-based digital health interventions for weight loss and lifestyle habit changes in overweight and obese adults: systematic review and meta-analysis. J Med Internet Res 2019;21(1):e298 [FREE Full text] [doi: 10.2196/jmir.9609] [Medline: 30622090]
60. Norcross JC, Krebs PM, Prochaska JO. Stages of change. J Clin Psychol 2011;67(2):143-154. [doi: 10.1002/jclp.20758] [Medline: 21157930]
61. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997;12(1):38-48. [doi: 10.4278/0890-1171-12.1.38] [Medline: 10170434]
62. Lambert JD, Dean SG, Terry RH, Charles N, Greaves C, Campbell JL, et al. Mechanisms of impact of web-based support and self-monitoring to augment and maintain physical activity levels: a qualitative study exploring participants' interactions with the e-coachER, a web-based support programme for people attending exercise referral schemes. BMJ Open 2024;14(10):e080472 [FREE Full text] [doi: 10.1136/bmjopen-2023-080472] [Medline: 39477258]
63. Følling IS, Oldervoll LM, Hilmarsen C, Ersfjord EMI. Exploring use of activity monitors for patients with obesity during weight-loss treatment - a qualitative study. BMC Sports Sci Med Rehabil 2021;13(1):25 [FREE Full text] [doi: 10.1186/s13102-021-00253-9] [Medline: 33731190]
64. Tay A, Hoeksema H, Murphy R. Uncovering barriers and facilitators of weight loss and weight loss maintenance: insights from qualitative research. Nutrients 2023;15(5):1297 [FREE Full text] [doi: 10.3390/nu15051297] [Medline: 36904294]
65. Raaijmakers LCH, Pouwels S, Berghuis KA, Nienhuijs SW. Technology-based interventions in the treatment of overweight and obesity: a systematic review. Appetite 2015;95:138-151. [doi: 10.1016/j.appet.2015.07.008] [Medline: 26165415]
66. Jøranli KT, Vefring LT, Dalen M, Garnweidner-Holme L, Molin M. Experiences of social support by participants with morbid obesity who participate in a rehabilitation program for health-behavior change: a qualitative study. BMC Nutr 2023;9(1):149 [FREE Full text] [doi: 10.1186/s40795-023-00810-0] [Medline: 38098104]
67. Morgan PJ, Scott HA, Young MD, Plotnikoff RC, Collins CE, Callister R. Associations between program outcomes and adherence to social cognitive theory tasks: process evaluation of the SHED-IT community weight loss trial for men. Int J Behav Nutr Phys Act 2014;11:89 [FREE Full text] [doi: 10.1186/s12966-014-0089-9] [Medline: 25011421]
68. Spreckley M, Seidell J, Halberstadt J. Perspectives into the experience of successful, substantial long-term weight-loss maintenance: a systematic review. Int J Qual Stud Health Well-being 2021 Dec;16(1):1862481 [FREE Full text] [doi: 10.1080/17482631.2020.1862481] [Medline: 33455563]
69. Liu K, Choi TST, Zhao L, Teong XT, Hutchison AT, Heilbronn LK. A qualitative exploration of behaviour change and maintenance experience in people with overweight or obesity in a dietary intervention. Nutr Diet 2024 Jun;81(3):296-305. [doi: 10.1111/1747-0080.12855] [Medline: 38057992]
70. Gillison FB, Rouse P, Standage M, Sebire SJ, Ryan RM. A meta-analysis of techniques to promote motivation for health behaviour change from a self-determination theory perspective. Health Psychol Rev 2019 Mar;13(1):110-130. [doi: 10.1080/17437199.2018.1534071] [Medline: 30295176]
71. Santos I, Silva MN, Teixeira PJ. A self-determination theory perspective on weight loss maintenance. EHP 2016;18(5) [FREE Full text]
72. Dandeneau DA, Leon SJ, Cameranesi M, Alexiuk MR, Protudjer JLP, Bohm CJ, et al. Evaluation of an online cardiometabolic and weight loss program: a mixed methods study. Appl Physiol Nutr Metab 2025 Jan 01;50:1-11. [doi: 10.1139/apnm-2024-0190] [Medline: 39854690]
73. Leung AWY, Chan RSM, Sea MMM, Woo J. Psychological factors of long-term dietary and physical activity adherence among Chinese adults with overweight and obesity in a community-based lifestyle modification program: a mixed-method study. Nutrients 2020;12(5):1379. [doi: 10.3390/nu12051379] [Medline: 32408509]
74. Puhl RM, Heuer CA. Obesity stigma: important considerations for public health. Am J Public Health 2010;100(6):1019-1028. [doi: 10.2105/AJPH.2009.159491] [Medline: 20075322]
75. Amy NK, Aalborg A, Lyons P, Keranen L. Barriers to routine gynecological cancer screening for White and African-American obese women. Int J Obes (Lond) 2006 Jan;30(1):147-155. [doi: 10.1038/sj.ijo.0803105] [Medline: 16231037]
76. Phelan SM, Burgess DJ, Yeazel MW, Hellerstedt WL, Griffin JM, van Ryn M. Impact of weight bias and stigma on quality of care and outcomes for patients with obesity. Obes Rev 2015 Apr;16(4):319-326 [FREE Full text] [doi: 10.1111/obr.12266] [Medline: 25752756]
77. Hussein N. Weight management in cardiac rehabilitation: evaluation, assessment, and interventions. J Clin Epigen 2022;7(9) [FREE Full text]

XSL•FO
RenderX

78. Imran TF, Wang N, Zombeck S, Balady GJ. Mobile technology improves adherence to cardiac rehabilitation: a propensity score-matched study. J Am Heart Assoc 2021 Aug 03;10(15):e020482 [FREE Full text] [doi: 10.1161/JAHA.120.020482] [Medline: 34278801]
79. Singh P, Kourav PS, Mohapatra S, Kumar V, Panda SK. Human heart health prediction using GAIT parameters and machine learning model. Biomed Signal Process Control 2024 Feb;88:105696 [FREE Full text] [doi: 10.1016/j.bspc.2023.105696]
80. Pujia C, Ferro Y, Mazza E, Maurotti S, Montalcini T, Pujia A. The role of mobile apps in obesity management: systematic review and meta-analysis. J Med Internet Res 2025 May 06;27:e66887 [FREE Full text] [doi: 10.2196/66887] [Medline: 40327853]
81. Faizan M, Han C, Lee SW. Policy-driven digital health interventions for health promotion and disease prevention: a systematic review of clinical and environmental outcomes. Healthcare (Basel) 2025 Sep 16;13(18):2319 [FREE Full text] [doi: 10.3390/healthcare13182319] [Medline: 41008449]
82. Pfaeffli Dale L, Whittaker R, Jiang Y, Stewart R, Rolleston A, Maddison R. Text message and internet support for coronary heart disease self-management: results from the Text4Heart randomized controlled trial. J Med Internet Res 2015 Oct 21;17(10):e237 [FREE Full text] [doi: 10.2196/jmir.4944] [Medline: 26490012]
83. Raja MA, Loughran R, Caffery FM. A review of applications of artificial intelligence in cardiorespiratory rehabilitation. Inform Med Unlocked 2023;41:101327 [FREE Full text] [doi: 10.1016/j.imu.2023.101327]
84. Marvasti TB, Gao Y, Murray KR, Hershman S, McIntosh C, Moayedi Y. Unlocking tomorrow's health care: expanding the clinical scope of wearables by applying artificial intelligence. Can J Cardiol 2024 Oct;40(10):1934-1945 [FREE Full text] [doi: 10.1016/j.cjca.2024.07.009] [Medline: 39025363]
85. Ahmad OM, Ibrahim R, Odunsi DI, Mohammed M, Mathew BG, Touny M, et al. Role of mobile health and wearable devices in cardiovascular disease prevention: a comprehensive review. Cureus 2025 May;17(5):e83673. [doi: 10.7759/cureus.83673] [Medline: 40486456]
86. Su J, Zhang Y, Ke QQ, Su JK, Yang QH. Mobilizing artificial intelligence to cardiac telerehabilitation. Rev Cardiovasc Med 2022;23(2):45 [FREE Full text] [doi: 10.31083/j.rcm2302045] [Medline: 35229536]
87. Narkhede M, Pardeshi A, Bhagat R, Dharme G. Review on emerging therapeutic strategies for managing cardiovascular disease. Curr Cardiol Rev 2024;20(4):e160424228949. [doi: 10.2174/011573403X299265240405080030] [Medline: 38629366]
88. Laine S, Heilala V, Säteri S, Sjögren T. Tekoälystä hyötyvät niin sydänkuntoutujat kuin fysioterapeutit [AI brings advantages to both cardiac rehab patients and physiotherapists]. Fysioterapia [Physiotherapy] 2025;72(6):36-41.
89. Abbas Q, Jeong W, Lee SW. Explainable AI in clinical decision support systems: a meta-analysis of methods, applications, and usability challenges. Healthcare (Basel) 2025 Aug 29;13(17):2154 [FREE Full text] [doi: 10.3390/healthcare13172154] [Medline: 40941506]
90. Taylor RS, Fredericks S, Jones I, Neubeck L, Sanders J, De Stoutz N, et al. Global perspectives on heart disease rehabilitation and secondary prevention: a scientific statement from the association of cardiovascular nursing and allied professions, European association of preventive cardiology, and international council of cardiovascular prevention and rehabilitation. Eur Heart J 2023;44(28):2515-2525 [FREE Full text] [doi: 10.1093/eurheartj/ehad225] [Medline: 37477626]
91. Varnfield M, Karunanithi M, Lee CK, Honeyman E, Arnold D, Ding H, et al. Smartphone-based home care model improved use of cardiac rehabilitation in postmyocardial infarction patients: results from a randomised controlled trial. Heart 2014 Nov;100(22):1770-1779 [FREE Full text] [doi: 10.1136/heartjnl-2014-305783] [Medline: 24973083]
92. Dorje T, Zhao G, Tso K, Wang J, Chen Y, Tsokey L, et al. Smartphone and social media-based cardiac rehabilitation and secondary prevention in China (SMART-CR/SP): a parallel-group, single-blind, randomised controlled trial. Lancet Digit Health 2019 Nov;1(7):e363-e374 [FREE Full text] [doi: 10.1016/S2589-7500(19)30151-7] [Medline: 33323210]
93. Huang K, Liu W, He D, Huang B, Xiao D, Peng Y, et al. Telehealth interventions versus center-based cardiac rehabilitation of coronary artery disease: a systematic review and meta-analysis. Eur J Prev Cardiol 2015 Aug;22(8):959-971. [doi: 10.1177/2047487314561168] [Medline: 25488550]
94. Kim H, Kim SE, Sung MK. Sex and gender differences in obesity: biological, sociocultural, and clinical perspectives. World J Mens Health 2025 Oct;43(4):758-772 [FREE Full text] [doi: 10.5534/wjmh.250126] [Medline: 40676890]
95. Casano HAM, Ahmed I, Anjum F. Six-minute walk test. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2023.
96. Kalfoss MH, Reidunsdatter RJ, Klöckner CA, Nilsen M. Validation of the WHOQOL-Bref: psychometric properties and normative data for the Norwegian general population. Health Qual Life Outcomes 2021 Jan 07;19(1):13 [FREE Full text] [doi: 10.1186/s12955-020-01656-x] [Medline: 33413455]
97. Archuleta AJ, Prost SG, Golder S. Psychometric properties of the WHOQOL-BREF in a population of older adults in prison: measuring quality of life in a carceral setting. Int J Prison Health 2023 Jul 25;ahead-of-print(ahead-of-print):668-681. [doi: 10.1108/IJPH-07-2022-0042] [Medline: 37480217]
98. Rosén H, Ahlström G, Lexén A. Psychometric properties of the WHOQOL-BREF among next of kin to older persons in nursing homes. Health Qual Life Outcomes 2020 Apr 19;18(1):103 [FREE Full text] [doi: 10.1186/s12955-020-01345-9] [Medline: 32306965]

XSL•FO
RenderX

Abbreviations

6MWT: 6-minute walk test
AI: artificial intelligence
CRT: cluster randomized trial
CVD: cardiovascular disease
PA: physical activity
WC: waist circumference
WHOQOL-BREF: World Health Organization Quality of Life Brief Form

Edited by A Stone; submitted 01.Jun.2025; peer-reviewed by SW Lee; comments to author 24.Nov.2025; accepted 07.Jan.2026; published 02.Mar.2026.

Please cite as:
Lahtio H, Paajanen T, Korpi H, Heinonen A, Sjögren T
Added Technology for Weight Management in Cardiac Rehabilitation: Biopsychosocial Factors and Rehabilitees' Talk in a Mixed Method Study
J Med Internet Res 2026;28:e78347
URL: https://www.jmir.org/2026/1/e78347
doi:10.2196/78347
PMID:

©Heli Lahtio, Teemu Paajanen, Hilkka Korpi, Ari Heinonen, Tuulikki Sjögren. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of the Components of a Digital Multiple Health Behavior Intervention Among University Students (Buddy): Factorial Randomized Trial

Katarina Åsberg[1], PhD; Oskar Lundgren[1,2], PhD; Hanna Henriksson[1,3], PhD; Pontus Henriksson[1], PhD; Ann Catrine Eldh[1,4], PhD; Preben Bendtsen[1,5], PhD; Marie Löf[6], PhD; Marcus Bendtsen[1], PhD

[1]Department of Health, Medicine and Caring Sciences, Linköping University, Linköping, Östergötland, Sweden

[2]Crown Princess Victoria Children's Hospital, Linköping University Hospital, Linköping, Sweden

[3]Unit for Strategic Healthcare, Region Östergötland, Linköping, Sweden

[4]Department of Public Health and Caring Sciences, Uppsala University, Uppsala, Sweden

[5]Medical Specialist Clinic, Motala Hospital, Motala, Sweden

[6]Department of Medicine Huddinge, Karolinska Institutet, Stockholm, Sweden

Corresponding Author:
Marcus Bendtsen, PhD
Department of Health, Medicine and Caring Sciences
Linköping University
Linköping, Östergötland, 581 83
Sweden
Phone: 46 13281000
Email: marcus.bendtsen@liu.se

Abstract

Background: Digital interventions have shown promise in supporting healthy behaviors among university students; however, few interventions support simultaneous change across multiple health behaviors. Moreover, behavioral interventions are typically evaluated as a whole, making it challenging to disentangle the contribution of individual components to the overall effects.

Objective: This study estimated the effects of the components of a digital behavior intervention on alcohol, diet, physical activity, and smoking outcomes among university students in Sweden.

Methods: A double-blind randomized factorial trial with 6 two-level factors was conducted. University students in Sweden were proactively recruited through student health care centers and social media. Participants were eligible if they were aged 18 years or older and had at least one health behavior classified as unhealthy. The effects of 6 components were estimated: screening and feedback; goal-setting and planning; motivation; skills and know-how; mindfulness; and self-authored SMS text messages. Primary outcomes were weekly alcohol consumption and frequency of heavy episodic drinking, average daily fruit and vegetable consumption, weekly sugary drink consumption, weekly moderate-to-vigorous physical activity (MVPA), and 4-week point prevalence of smoking.

Results: A total of 1704 students were randomized. The effectiveness of individual and pairwise components was estimated using available data from 1118 (65.61%) participants at 2 months and 874 (51.29%) at 4 months, with sensitivity analyses conducted using imputed missing data. Most consistently, the evidence indicated that screening and feedback affected fruit and vegetable consumption (2-month mean difference 0.11, compatibility interval [CoI] –0.02 to 0.24; probability of effect [POE] 94.7% and 4-month mean difference 0.12, CoI –0.03 to 0.26; POE 94.4%), as did skills and know-how (2-month mean difference 0.19, CoI 0.06-0.33; POE 99.8% and 4-month mean difference 0.14, CoI 0.01-0.28; POE 96.9%). The combination of these 2 components was even more effective (2-month mean difference 0.30, CoI 0.11-0.48; POE 99.9% and 4-month mean difference 0.26, CoI 0.05-0.46; POE 99.4%). The motivation and mindfulness components, both individually and in combination, increased MVPA at 2 months (combined mean difference 78.0, CoI 28.3-128.2; POE 99.9%); however, this effect was not observed at 4 months. Combining screening and feedback with skills and know-how increased MVPA at 4 months (mean difference 60.1, CoI 3.6-116.5; POE 98.2%). Heavy episodic drinking was reduced at 2 months by screening and feedback (incidence rate ratio 0.87, CoI 0.74-1.02; POE 95.2%), and the effect was greater when combined with goal-setting and mindfulness. There was some evidence that the motivation component was harmful with respect to heavy episodic drinking and that self-authored SMS text messages were harmful with respect to sugary drink consumption.

XSL•FO
RenderX

Conclusions: We dismantled a complex digital multiple behavior intervention and examined it using a factorial design to provide novel insights into the effectiveness of the intervention's different components. Both marginal and synergistic effects were observed across multiple behaviors, providing evidence regarding which components are most promising in complex interventions. These findings should be considered in light of the risk of bias introduced by attrition to follow-up, which was high in this effectiveness trial with low barriers to participation.

Trial Registration: International Standard Randomised Controlled Trial Number (ISRCTN) ISRCTN23310640; https://www.isrctn.com/ISRCTN23310640

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2021-051044

(J Med Internet Res 2026;28:e88884) doi:10.2196/88884

KEYWORDS

behavior change; digital interventions; multiple lifestyle; public health; mHealth intervention

Introduction

Unhealthy behaviors, including alcohol consumption, unhealthy diets, physical inactivity, and smoking, are leading causes of noncommunicable diseases, including cardiovascular diseases, cancers, and diabetes [1-5]. The young adult years are a critical period for the establishment of healthy behaviors [6-8], which may reduce future morbidity [9-14]. Engaging in healthy behaviors also provides several immediate positive effects on both physical and mental health [15-20]. These include fewer mental health problems, such as anxiety and depression [17-20]; improved cognitive function and academic achievement [21,22]; and various aspects of well-being [15,16]. Additionally, reduced alcohol consumption lowers the risk of alcohol-related harms or negative consequences, such as social problems, injuries, or risky sexual behaviors [23,24]. However, a study of university students in 24 countries found a high prevalence of unhealthy behaviors [25]. In addition, unhealthy behaviors tend to co-occur among students; for instance, low fruit and vegetable consumption and low levels of physical activity often occur together. This is concerning because the risks of negative consequences associated with multiple unhealthy behaviors compound [26].

In the digital era, new avenues to support individuals in improving their health behaviors are emerging. One form of support is individual-level digital behavior change interventions, which deliver support materials via personal web portals, SMS text messages, smartphone apps, etc. The effectiveness of digital interventions across various health behaviors has been demonstrated in a wide range of populations and settings [27-33]. As young adults frequently use technology to access health information [34], digital interventions have the potential to efficiently reach this target group, and numerous single–health behavior digital interventions have been evaluated among university and college students [35-38]. In Sweden, digital interventions among university students have shown promising results, particularly with respect to alcohol [39-42], smoking cessation [43], and mental health promotion [44]. However, only a few interventions that support changes in multiple health behaviors simultaneously have been studied [45-47], and even fewer are tailored specifically for students [48]. Therefore, while single-behavior digital interventions show promise, our understanding of the effects of digital interventions targeting multiple health behaviors remains limited, despite the prevalence of co-occurring unhealthy behaviors [49,50].

Most digital behavior change interventions are grounded in behavior change theory and utilize behavior change techniques (BCTs) [51]. Systematic reviews [52,53] support the use of theory and BCTs in developing effective digital interventions. However, findings regarding the effectiveness of BCTs have been heterogeneous [54], indicating that the ways in which BCTs are implemented may vary considerably across interventions. In addition, behavioral interventions are typically evaluated using designs that estimate the effects of interventions as a whole [55], rather than at the component level, making it challenging to disentangle the contribution of specific components and BCTs to the overall effects.

In response to the lack of evidence on the effectiveness of digital interventions targeting multiple health behaviors, and the limited evidence on the effectiveness of different components of digital interventions, we conducted a factorial trial among university and college students in Sweden. The trial aimed to estimate the effectiveness of 6 intervention components of a digital intervention in promoting behavior change related to alcohol consumption, diet, physical activity, and smoking.

Methods

Study Design

We conducted a factorial randomized trial of a digital multiple health behavior intervention among university and college students in Sweden. The factorial trial was designed with 6 two-level factors representing the presence or absence of 6 intervention components. The treatment effect being estimated was thus the effect of the presence versus absence of each specific component, adjusted for the presence and absence of other components, as well as pairwise interactions between components. The trial was part of the MoBILE research program, which includes studies of digital health behavior interventions across the lifespan [56]. A protocol for the trial was published prospectively [57], and the trial was preregistered in the International Standard Randomised Controlled Trial Number (ISRCTN) registry on January 28, 2021 (ISRCTN23310640). This report follows CONSORT (Consolidated Standards of Reporting Trials) guidelines (see Multimedia Appendix 1) [58], including extensions of CONSORT for factorial trials [59].

XSL•FO
RenderX

Students were included in the trial if they met at least one of six inclusion criteria based on unhealthy behaviors as defined by national guidelines. We made an adjustment to the inclusion criteria for sugary drinks after trial registration. Initially, we decided that participants who consumed 3 or more units (33 cl) of sugary drinks per week would be included in follow-up analyses of sugary drink consumption. However, this was changed to 2-3 or more units per week because the questionnaire used did not differentiate between those consuming 2 or more units per week and those consuming 3 or more units per week. No other changes were made to the trial protocol.

Study Setting, Recruitment, and Participants

Students at 18 geographically dispersed universities and colleges across Sweden (out of 50) were invited to participate in the trial. These institutions encompassed a variety of urban and rural settings and offered diverse educational profiles. Prospective participants were recruited using paper advertising (posters and leaflets), digital advertising (email, websites, and social media), and through student health care staff. Students registered their interest by sending an SMS text message to a dedicated phone number, after which they received an immediate reply with a hyperlink to a web page containing informed consent materials (see Multimedia Appendix 2). Consenting students were then asked to complete a baseline questionnaire (see Multimedia Appendix 3), the responses to which were used to assess eligibility. Included students were required to be 18 years or older and to fulfill at least one of the criteria presented in Textbox 1, which were based on Swedish national guidelines for healthy behaviors at the time the trial commenced.

Trial information and intervention materials were in Swedish and delivered to participants' mobile phones; thus, individuals who did not understand Swedish well enough to enroll or who did not have access to a mobile phone were not able to take part in the study. Notably, it has been estimated that 96% of 16-35-year-olds in Sweden own a smartphone, and this percentage is likely to be even higher among university and college students.

Textbox 1. Criteria for inclusion.

1. Weekly alcohol consumption: Consumed 10/15 (women/men) or more standard drinks of alcohol in the past week. A standard drink was defined as 12 g of pure alcohol.

2. Heavy episodic drinking: Consumed 4/5 (women/men) or more standard drinks of alcohol on a single occasion at least once in the past month.

3. Fruit and vegetables: Consumed less than 500 g of fruit and vegetables on average per day in the past week.

4. Sugary drinks: Consumed 2 or more units of sugary drinks in the past week. One sugary drink unit was defined as approximately 33 cl.

5. Moderate-to-vigorous physical activity: Spent less than 150 minutes on moderate-to-vigorous physical activity in the past week.

6. Smoking: Smoked at least one cigarette in the past week.

Interventions

The digital multiple health behavior intervention studied in this trial was a mobile phone–based intervention targeting alcohol, diet, physical activity, and smoking. The intervention, called Buddy, consisted of 6 components aimed at promoting behavior change: screening and feedback (C1); goal-setting and planning (C2); motivation (C3); skills and know-how (C4); mindfulness (C5); and self-authored SMS text messages (C6). The components were informed by social-cognitive models of health [60] and have been included in interventions that have proven successful in facilitating behavior change [61,62]. Table 1 provides a brief description of each of the 6 components, utilizing the BCTTv1 93-item taxonomy [63] to specify BCTs. Please see Multimedia Appendix 4 for full details of the intervention.

XSL•FO
RenderX

Table 1. Description of intervention components, conceptual framework, practical application, and factorial conditions.

Component	Conceptual framework	Practical application	Factorial conditions
Component 1: Screening and feedback	*Self-monitoring* has been shown to be a potentially effective strategy for reducing excessive alcohol consumption [39,64-66] and promoting healthy eating and physical activity [67,68]. This concerned the following BCT[a] components: Discrepancy between current behavior and goal(s) (BCT 1.6), Feedback on behavior (BCT 2.2), Self-monitoring of behavior (BCT 2.3), and Social comparison (BCT 6.2).	Every Sunday afternoon, participants received an SMS text message containing a hyperlink to a questionnaire regarding their current health behaviors. After completing the questionnaire, participants received feedback on their current behaviors in comparison with national guidelines. Thereafter, users were given access to the remaining components (depending on allocation).	When this component was absent, participants were not asked to complete the screening questionnaire but were instead shown the national guidelines without receiving personal feedback.
Component 2: Goal-setting and planning	*Self-regulatory* skills and capacity were addressed via goal-setting and planning. Planning-related activities, such as goal-setting, action planning, and practicing behavior, have been shown to be effective in lifestyle behavior interventions [67,69-74]. This concerned the following BCT components: Goal-setting (BCT 1.1), Problem solving (BCT 1.2), Action planning (BCT 1.4), Prompts/cues (BCT 7.1), and Behavior practice/rehearsal (BCT 8.1).	This component allowed participants to set 1 or more goals for their future behavior. It included action planning for how they would progress toward their goals, preparation for motivational struggles (coping planning), and strategies for rewarding themselves upon success. Participants could also create their own challenges or accept ready-made challenges for the coming week, such as walking for 15 minutes each day or not drinking any alcohol during the coming week. Reminders about their goals and challenges were sent to participants via SMS text messages throughout the week (up to 4 SMS text messages).	When absent, this component was not visible, and goal-setting reminders were not available.
Component 3: Motivation	Digital behavior change interventions have been shown to enhance *self-efficacy*; however, there is a lack of consensus across reviews regarding which content works best to facilitate increases in self-efficacy [75]. This concerned the following BCT components: Information about health consequences (BCT 5.1), Credible source (BCT 9.1), Pros and cons (BCT 9.2), and Comparative imagining of future outcomes (BCT 9.3).	This component contained information and tools to increase participants' awareness of their own motivation, encourage commitment, and boost self-efficacy. It included information on negative health consequences, costs induced by certain behaviors, and reflective tasks delivered via SMS text messages, videos, and exercises. If participants chose, they could also activate motivational SMS text messages derived from previously developed and evaluated interventions [40,43,76-80], which were sent to them throughout the week (a maximum of 8-10 SMS text messages per week).	When absent, this component was not visible, and motivational SMS text messages were not available.
Component 4: Skills and know-how	Intervention content based on shaping knowledge, aimed at increasing an individual's understanding and awareness to facilitate behavior change, has been shown to be effective [70,71,81,82]. This concerned the following BCT components: Instructions on how to perform a behavior (BCT 4.1), Behavior substitution (BCT 8.2), Habit formation (BCT 8.3), and Graded tasks (BCT 8.7).	This component aimed to increase participants' skills and know-how by providing concrete tips on how to initiate and maintain lasting changes in everyday life through repetition and substitution. Participants were given strategies, for example, how to say no to alcoholic beverages or how to introduce vegetables into their meals. If participants chose to do so, they could also activate SMS text messages with tips and know-how, which were sent to them throughout the week (a maximum of 8-10 SMS text messages per week).	When absent, this component was not visible, and SMS text messages with tips and know-how were not available.

XSL•FO
RenderX

Component	Conceptual framework	Practical application	Factorial conditions
Component 5: Mindfulness	The mindfulness exercises were based on previous research and were considered evidence-based methods to improve the mental well-being of clinical populations, whereas their effects on behavior change in nonclinical settings are less well studied [83-87].	This component aimed to increase users' awareness of their own lived experience and strengthen their capacity for a nonreactive, compassionate, and less stressful way of being in the world. Mindfulness exercises, including guided meditations, were offered to participants.	When absent, this component was not visible, and guided meditations were not available.
Component 6: Self-authored SMS text messages	SMS text message self-authorship is generally understudied in the literature, but was included in an effective digital alcohol intervention [88].	This component allowed participants to self-author up to 3 SMS text messages and schedule them to be sent to themselves throughout the week at times of their choosing. For example, a participant could write an SMS text message reminding themselves to eat 2 fruits each day, avoid drinking on Wednesdays, or go for a walk with a friend.	When absent, this component was not visible, and the ability to self-author SMS text messages was not available.

[a]BCT: behavior change technique.

As shown in Figure 1, the components were presented to participants in a menu, allowing access in any order and with any frequency. Participants could therefore navigate between components in ways typical of smartphone apps. Regardless of the inclusion criterion met, all participants received intervention materials regarding alcohol, diet, physical activity, and smoking behaviors. There was no difference in the materials provided based on risk profile, except for the feedback received after screening (C1), where reported behavior was visually matched against national guidelines. The components presented materials that were both behavior-specific and more general to behavior change. Screening and feedback (C1) was conducted for each behavior, with feedback provided separately (eg, showing how current heavy drinking patterns match recommendations). Screening and feedback (C1) was offered to participants each week, allowing them to receive feedback on their behavior multiple times throughout the intervention period. Goal-setting and planning (C2) guided participants to set goals without prompting any specific behavior and also allowed them to create personal challenges; in addition, preprepared challenges were available and categorized by behavior. Motivation (C3) included materials that were not behavior specific and aimed to increase motivation and commitment to behavior change. The C3 component also allowed participants to sign up to receive behavior-specific SMS text messages throughout the week to boost motivation and commitment (participants could sign up for multiple behaviors). Skills and know-how (C4) contained content and optional SMS text messages designed to increase participants' skills and knowledge about how to change, and was provided separately for each behavior, with loose references between behaviors (eg, recommending going for walks when feeling the urge to smoke). Mindfulness (C5) did not mention specific behaviors but instead aimed to strengthen the mental resources needed to change behavior. Self-authored SMS text messages (C6) provided a means for participants to send reminders to themselves throughout the week, with no content provided for any specific behavior.

XSL•FO
RenderX

Figure 1. Screenshot of the Buddy intervention displaying the components in order from top to bottom and left to right: (1) Screening and feedback, (2) Goal-setting and planning, (3) Motivation, (4) Skills and know-how, (5) Mindfulness, and (6) Self-authored SMS text messages.



Participants were randomly assigned to 1 of 64 conditions, each representing a unique combination of Buddy's 6 components. The components assigned to individuals, based on their factorial allocation, were available for use at their discretion over a 4-month period, with weekly SMS text message reminders sent on Sunday afternoons to encourage access to the intervention.

Outcomes and Measures

Outcomes are listed in Textbox 2 and described below. All questionnaires used in the trial can be found in Multimedia Appendix 3.

Textbox 2. Primary and secondary outcomes.

1. Primary outcomes

- Alcohol: weekly alcohol consumption; monthly frequency of heavy episodic drinking.
- Diet: average daily consumption of fruit and vegetables; weekly consumption of sugary drinks.
- Physical activity: weekly moderate-to-vigorous physical activity.
- Smoking: 4-week point prevalence of smoking abstinence.

2. Secondary outcomes

- Weekly consumption of candy and snacks.
- BMI.
- Weekly number of cigarettes smoked.
- Perceived stress.

XSL•FO
RenderX

Weekly alcohol consumption was assessed by asking participants to report the number of standard drinks of alcohol they consumed in the past week (short-term recall method [89]). The frequency of heavy episodic drinking was assessed by asking participants how many times they consumed more than 4/5 (women/men) standard drinks of alcohol on 1 occasion in the past month. These 2 outcomes are both part of the core outcome set for brief alcohol interventions [90,91].

Diet and physical activity were assessed using a questionnaire based on a previously published questionnaire by the National Board of Health and Welfare in Sweden [92], which was further modified to include portion sizes. Fruit and vegetable consumption was assessed using 2 questions on the number of portions (100 g) of fruit and vegetables, respectively, that participants consumed on average per day during the past week. Sugary drink consumption was assessed using a question on the number of units (33 cl) of sugary drinks participants consumed in the past week. Moderate-to-vigorous physical activity (MVPA) was assessed by summing responses to 2 questions on the number of minutes spent on MVPA, respectively, during the past week.

BMI was calculated using participants' self-reported weight (without further instructions). Baseline height was used, as it was deemed unlikely to have changed significantly.

Four-week point prevalence of smoking abstinence (no cigarettes in the past 4 weeks) was assessed as a binary question, a measure suggested by the Society for Research on Nicotine and Tobacco [93]. Participants who reported having smoked any cigarettes in the past 4 weeks were also asked to report the number of cigarettes smoked in the past week.

Perceived stress was assessed using the Short-Form Perceived Stress Scale [94].

Follow-Up

There were 2 follow-up intervals for primary and secondary outcomes: 2 and 4 months after randomization. The 2-month follow-up facilitated the assessment of the immediate effects of the components, while the 4-month follow-up facilitated estimation of effects after prolonged access. Additional longer follow-up intervals were not included to reduce participant burden and because attrition was expected to be much higher at longer intervals from baseline. An additional follow-up was conducted at 1 month after randomization to assess proposed mediators of the intervention; these results will be reported separately. All follow-ups were initiated by sending SMS text messages to participants with hyperlinks to questionnaires (see Multimedia Appendix 3). Additional attempts were made to collect data from nonresponders, including 2 SMS text message reminders sent 2 days apart, followed by phone calls with a maximum of 5 attempts.

Randomization and Blinding

We used block randomization with random block sizes of 64 and 128 to allocate participants equally among conditions, where each condition represented a unique combination of components. The process was automated, with the sequence computer-generated and the backend server allocating eligible participants after completion of the baseline. Neither research personnel nor participants were able to influence the allocation. Research personnel were blind to allocation throughout the trial. Participants were not aware of components other than those to which they were allocated and were therefore considered blind to allocation.

Statistical Analysis

We conducted analyses with participants in the conditions to which they were randomly allocated (intention-to-treat). We estimated models using both available data and data with missing values imputed. We used multiple imputation with chained equations, generating 100 datasets using 30 iterations with predictive mean matching. To account for the uncertainty inherent in missing data, imputed analyses were performed for all models as a sensitivity analysis, regardless of the proportion of missingness. All statistical analyses were prespecified in the trial protocol [57].

Longitudinal data were modeled using multilevel models with random intercepts for participants and time-by-component interactions. In addition to estimating the marginal effects of each component, we estimated the effects of pairwise interactions among components to examine potential synergistic effects. This was done by including pairwise interaction terms in the regression models for all outcomes. The treatment effect being estimated was thus the effect of the presence versus absence of each specific component, adjusted for the presence and absence of other components, as well as pairwise interactions between components. The models were adjusted for age, gender, and baseline measures of importance, confidence, and knowledge of how to change.

Bayesian inference was used to estimate the parameters of the models, using standard normal priors for coefficients to incorporate an a priori conservative view of the magnitudes of effects [95]. The standard normal priors induced shrinkage in the parameter estimates to protect against spurious findings. We report the marginal posterior probability of effect (POE), with the median of the posterior distribution representing a point estimate of the magnitude of the effect, along with 95% compatibility intervals (CoIs) defined by the 2.5th and 97.5th percentiles.

Primary and Secondary Outcomes

We analyzed primary outcomes at 2 and 4 months among participants who fulfilled the respective inclusion criteria. For example, the MVPA outcome was analyzed among those who reported less than 150 minutes of MVPA in the past week at baseline. Secondary outcomes at 2 and 4 months were analyzed among all participants, except for the number of cigarettes smoked in the past week, which was analyzed among baseline smokers who continued to smoke.

Weekly alcohol consumption, frequency of heavy episodic drinking per month, number of sugary drinks per week, weekly intake of candy and snacks, and cigarettes smoked in the past week were all analyzed using negative binomial regression. Average intake of fruit and vegetables per day, MVPA minutes per week, BMI, and stress were all analyzed using linear

XSL•FO
RenderX

regression. Point prevalence of smoking abstinence was analyzed using logistic regression.

Ancillary Analyses

We investigated attrition using 2 approaches. First, we estimated the odds of responding to follow-up, conditional on baseline characteristics and the presence or absence of the 6 components (factors), using logistic regression. One model included no interaction terms, while a second model included interaction terms between factors and baseline characteristics. Given the large number of covariates in these models, we used Cauchy priors to promote a parsimonious model (centered at 0 with a standard normal hyperprior for the scale). The use of Cauchy priors encodes an a priori conservative view and pulls estimated associations toward the null unless the data strongly suggest otherwise. Second, if data are missing systematically, nonresponders may be more similar to responders who require more attempts to collect data than to those who require fewer attempts. We therefore modeled primary outcome data conditional on the number of attempts required to collect follow-up data. Associations between the number of attempts and outcome measures may indicate systematic attrition.

Post Hoc Analyses

To investigate whether the number of components that participants had access to affected outcomes, we conducted post hoc analyses in which outcomes were modeled conditional on the number of components. The same model specifications were used as in the primary analyses, with dummy variables for components replaced by a variable representing the component count.

Sample Size

We used a Bayesian sequential design to monitor when to stop recruitment [96]. In such designs, targets are defined and evaluated throughout the course of the trial to determine whether the trial should stop or continue. As 4-month follow-up data became available, we modeled each of the primary outcomes according to the analysis plan, and coefficients for dummy variables representing the presence or absence of each component were compared against targets for effect, harm, and futility (see protocol for details of the criteria [57]).

We stopped recruitment after 30 months. While our prespecified limit was 24 months of recruitment irrespective of targets, we extended recruitment by 6 months because some colleges and universities began their recruitment procedures later than others. Overall, each participating college and university was expected to recruit for at least 24 months.

Ethical Considerations

The trial received ethical approval from the Swedish Ethical Review Authority on December 15, 2020 (Dnr: 2020-05496 and Dnr: 2022-03561-02). All participants were provided with study information and gave informed consent before responding to the baseline questionnaire and subsequently being randomized. Participants received no compensation for participating in the trial. Participants' phone numbers were encrypted and stored in the study database to link baseline and follow-up data. At project completion, the encrypted phone numbers will be deleted, and the data will be anonymized.

Results

Overview

A CONSORT participant flow diagram is presented in Figure 2 (sign up and randomization) and Figure 3 (follow-up). Between April 13, 2021, and October 18, 2023, 2468 individuals expressed interest in the study, of whom 1804 consented. The majority were recruited through the trial's information materials, either digital or leaflets provided by university student health care services (1999/2468, 81%), while the remainder were recruited through online advertisements on social media platforms (469/2468, 19%). A total of 1718 individuals completed the baseline questionnaire, of whom 14 were excluded for not fulfilling any of the behavioral inclusion criteria. The remaining 1704 participants were randomized. Baseline characteristics of the randomized participants were balanced across all factors (see Table 2).

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) participant flow diagram—sign up and randomization.

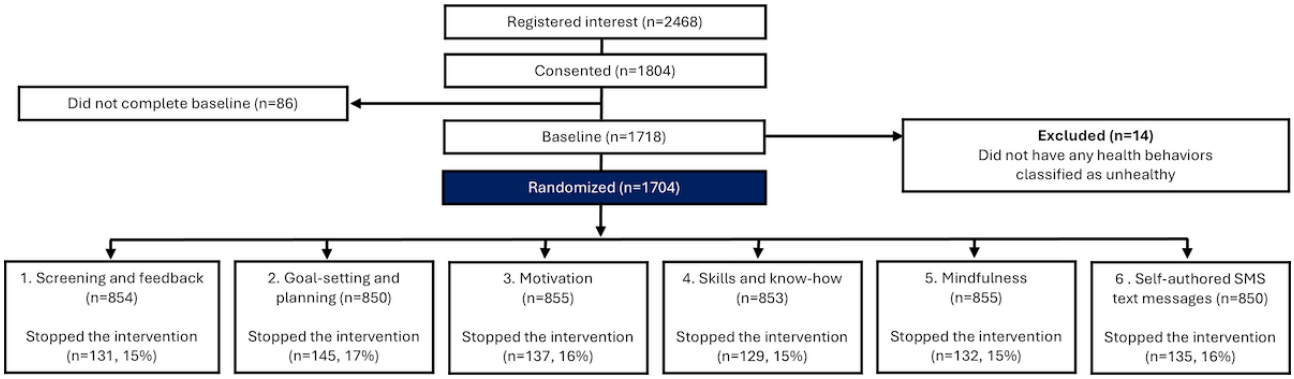

XSL•FO
RenderX

Figure 3. CONSORT (Consolidated Standards of Reporting Trials) participant flow diagram—follow-up.

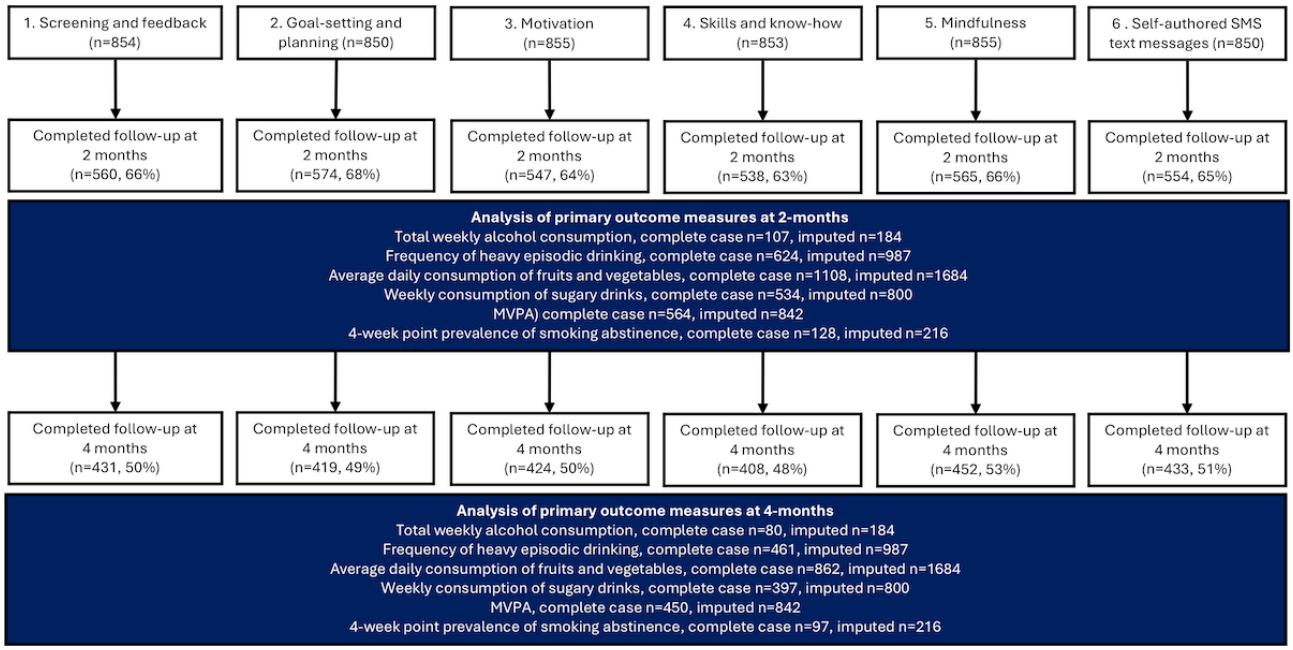

XSL•FO
RenderX

Table 2. Baseline characteristics of randomized participants.

Characteristics	Total (n=1704)	Screening and feedback (n=854)	Goal-setting and planning (n=850)	Motivation (n=855)	Skills and know-how (n=853)	Mindfulness (n=855)	Self-authored SMS text messages (n=850)
Sex, n (%)							
Women	1400 (82.16)	709 (83.02)	704 (82.82)	699 (81.75)	718 (84.17)	697 (81.52)	692 (81.41)
Men	304 (17.84)	145 (16.98)	146 (17.18)	156 (18.25)	135 (15.83)	158 (18.48)	158 (18.59)
Age (years), mean (SD)	28 (6.9)	28.3 (6.9)	28.5 (7.1)	28.2 (7.0)	27.8 (6.8)	28.5 (7.0)	28.4 (7.0)
BMI (kg/m^2), mean (SD)	25.7 (5.5)	25.9 (5.5)	25.6 (5.3)	25.7 (5.2)	25.8 (5.6)	25.7 (5.3)	25.8 (5.3)
Underweight (<18.5), n (%)	60 (3.52)	28 (3.28)	33 (3.88)	30 (3.51)	33 (3.87)	27 (3.16)	19 (2.24)
Normal weight (18.5-24.99), n (%)	844 (49.53)	417 (48.83)	413 (48.59)	415 (48.54)	429 (50.29)	426 (49.82)	432 (50.82)
Overweight (25-29.99), n (%)	458 (26.88)	234 (27.40)	236 (27.76)	239 (27.95)	214 (25.09)	241 (28.19)	227 (26.71)
Obese (≥30.0), n (%)	342 (20.07)	175 (20.49)	168 (19.76)	171 (20.00)	177 (20.75)	161 (18.83)	172 (20.24)
Alcohol, mean (SD)							
Total weekly alcohol consumption	3.7 (5.2)	3.7 (5.2)	3.8 (5.3)	3.7 (5.1)	3.6 (5.1)	3.8 (5.1)	3.7 (5.1)
Frequency of heavy episodic drinking	1.8 (2.7)	1.8 (2.7)	1.8 (2.8)	1.8 (2.7)	1.8 (2.7)	1.7 (2.4)	1.8 (2.5)
Dietary behavior, mean (SD)							
Average daily fruit and vegetable consumption, portions per week	1.3 (1.1)	1.3 (1.1)	1.3 (1.1)	1.3 (1.1)	1.3 (1.1)	1.3 (1.1)	1.3 (1.1)
Sugary drinks, cans per week	2.7 (4.1)	2.7 (4.1)	2.7 (4.2)	2.7 (4.1)	2.6 (3.9)	2.7 (4.0)	2.5 (3.9)
Candy and snacks, portions per week	6.5 (6.4)	6.6 (6.4)	6.4 (6.3)	6.6 (6.5)	6.5 (6.4)	6.5 (6.5)	6.3 (6.4)
Physical activity, mean (SD)							
MVPA (minutes)	207 (214)	209 (221)	214 (225)	213 (214)	211 (213)	207 (210)	196 (207)
Smoking							
Number of smokers, n (%)	218 (12.79)	106 (12.41)	119 (14.00)	106 (12.40)	105 (12.31)	104 (12.16)	101 (11.88)
Number of cigarettes last week, mean (SD)	3.5 (16.6)	3.0 (15.1)	4.2 (18.8)	3.6 (17.7)	2.6 (13.4)	3.3 (16.5)	2.8 (14.9)
Stress, mean (SD)							
Self-perceived stress[a]	7.8 (2.9)	7.8 (2.9)	7.8 (2.9)	7.7 (3.0)	7.8 (2.9)	7.7 (3.0)	7.7 (2.9)
Psychosocial measures, median (quartiles)							
Importance[b]	8 (7-10)	8 (7-10)	8 (7-10)	8 (7-10)	8 (7-10)	8 (7-10)	8 (7-10)
Confidence[b]	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)
Know-how[b]	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)

[a]The 4-item Perceived Stress Scale with total scores ranging from 0 to 16.

[b]Single item with 1-10 response options.

Among participants, 16% stopped the intervention before the 16-week program had ended. Consistent with the consent provided, we did not investigate the reasons why these individuals chose to stop receiving SMS text messages. The CONSORT flowchart in Figure 2 provides the numbers who stopped per component. All participants, including those who discontinued the intervention, were followed up and included in the main analysis. At the 2-month follow-up interval, primary outcome data were available for 1118 out of 1704 (65.61%)

XSL•FO
RenderX

participants, and for 874 out of 1704 (51.29%) participants at the 4-month follow-up interval.

Attrition

Attrition analyses revealed no marked associations between baseline characteristics and nonresponse at the 2-month follow-up interval. At the 4-month follow-up interval, there was evidence that older participants were more likely to respond to follow-up and that those smoking more cigarettes per week at baseline were less likely to respond. Notably, the association between nonresponse and smoking was driven by a few data points. Neither of these associations was found to differ between component conditions. Please see Multimedia Appendix 5 for full details.

There was some evidence that late responders, that is, those requiring more attempts to collect follow-up data, had healthier behaviors regarding fruit and vegetable intake and alcohol consumption. This association was attenuated among those with access to the screening and feedback (C1) component, who consumed more alcohol and sugary drinks at the 4-month follow-up and less fruit and vegetables at the 2-month follow-up. While there is considerable uncertainty surrounding these analyses, under the assumption that late responders are more similar to nonresponders than early responders, this may indicate systematic attrition. Please see Multimedia Appendix 6 for details.

Primary and Secondary Outcomes

Estimates and Main Effects

Table 3 contains the main effects of individual components on primary outcomes, and Multimedia Appendix 7 presents the main effects of individual components and 2-way interactions between components for the available data. The corresponding analyses with missing data imputed are presented in Multimedia Appendix 8. Here, we highlight findings for which the evidence of effects was strongest. We report estimates from analyses of the available data unless otherwise noted.

XSL•FO
RenderX

Table 3. Estimates of marginal effects of components on primary outcomes at 2- and 4-month follow-up.

Component	C1		C2		C3		C4		C5		C6	
	Estimate[a]	Probability[b], %	Estimate	Probability, %	Estimate	Probability, %	Estimate	Probability, %	Estimate	Probability, %	Estimate	Probability, %
Weekly alcohol consumption												
2 months	1.06 (0.77 to 1.50)	63.7	0.95 (0.68 to 1.32)	63.1	0.79 (0.58 to 1.08)	93.6	0.91 (0.66 to 1.25)	71.9	0.98 (0.72 to 1.34)	53.9	0.90 (0.65 to 1.23)	75.5
4 months	1.01 (0.71 to 1.45)	53.1	1.06 (0.74 to 1.53)	63.4	0.87 (0.62 to 1.22)	79.6	0.96 (0.67 to 1.36)	59.9	1.19 (0.84 to 1.69)	83.4	0.65 (0.45 to 0.93)	99.1
Heavy episodic drinking												
2 months	0.87 (0.74 to 1.02)	95.2	0.89 (0.75 to 1.04)	93.1	0.96 (0.82 to 1.13)	68.7	1.00 (0.86 to 1.18)	51.7	0.89 (0.76 to 1.04)	92.5	1.14 (0.98 to 1.34)	95.0
4 months	0.97 (0.80 to 1.16)	65.2	1.09 (0.91 to 1.32)	82.8	1.18 (0.99 to 1.42)	96.5	1.03 (0.86 to 1.25)	64.2	1.11 (0.92 to 1.33)	86.1	0.93 (0.77 to 1.12)	78.6
Fruit and vegetable consumption												
2 months	0.11 (–0.02 to 0.24)	94.7	–0.09 (–0.22 to 0.04)	90.2	0.09 (–0.04 to 0.23)	92.0	0.19 (0.06 to 0.33)	99.8	0.03 (–0.11 to 0.16)	64.7	–0.01 (–0.14 to 0.12)	54.9
4 months	0.12 (–0.03 to 0.26)	94.4	–0.01(–0.15 to 0.14)	53.6	0.12 (–0.02 to 0.27)	95.4	0.14 (–0.01 to 0.28)	96.9	–0.05 (–0.19 to 0.09)	74.9	–0.09 (–0.23 to 0.05)	88.2
Sugary drinks												
2 months	0.91 (0.77 to 1.06)	88.4	1.03 (0.87 to 1.21)	61.2	0.98 (0.84 to 1.15)	59.2	1.04 (0.88 to 1.22)	66.5	0.94 (0.80 to 1.10)	78.8	1.24 (1.05 to 1.46)	99.4
4 months	1.03 (0.86 to 1.24)	62.2	1.05 (0.87 to 1.26)	68.8	1.09 (0.90 to 1.31)	81.5	1.06 (0.88 to 1.27)	74.2	1.07 (0.89 to 1.29)	76.6	1.16 (0.96 to 1.39)	93.8
Moderate-to-vigorous physical activity												
2 months	9.34 (–25.6 to 45.5)	69.7	–5.73 (–41.9 to 29.5)	63.0	35.8 (–0.41 to 71.9)	97.4	17.1 (–19.2 to 54.2)	82.2	41.2 (5.04 to 77.6)	98.7	5.68 (–30.5 to 41.1)	62.2
4 months	28.1 (–11.3 to 67.6)	92.0	17.6 (–20.6 to 56.5)	81.4	6.63 (–33.1 to 45.2)	63.0	32.8 (–6.1 to 71.8)	95.2	3.02 (–36.3 to 42.0)	56.0	–19.2 (–57.6 to 20.1)	83.6
Smoking cessation												
2 months	1.12 (0.35 to 3.59)	57.4	0.56 (0.17 to 1.76)	84.3	1.43 (0.43 to 4.63)	72.3	0.68 (0.21 to 2.23)	74.5	0.46 (0.14 to 1.49)	90.4	2.15 (0.67 to 6.74)	90.4
4 months	1.45 (0.35 to 6.10)	69.2	0.48 (0.11 to 2.00)	84.3	2.30 (0.52 to 9.56)	87.3	0.77 (0.17 to 3.39)	64.0	0.55 (0.13 to 2.28)	79.6	0.94 (0.22 to 4.02)	53.0

[a]Estimates are presented as the median of the posterior distribution with 95% compatibility intervals. Incidence rate ratios are presented for weekly alcohol consumption, heavy episodic drinking, sugary drinks; odds ratios are presented for smoking cessation; and mean differences are presented for fruit and vegetable consumption and moderate-to-vigorous physical activity.

[b]Probability is the proportion of the posterior distribution above or below the null in the direction of the median.

XSL•FO
RenderX

Primary Outcomes

Total Weekly Alcohol Consumption

Self-authored SMS text messages (C6) reduced total weekly alcohol consumption at 4 months (incidence rate ratio [IRR] 0.65, CoI 0.45-0.93; POE 99.1%). Apart from weak evidence for an effect of motivation (C3) at 2 months, there was no marked evidence of effects for the other components individually. There was evidence that combining motivation (C3) and self-authored SMS text messages (C6) had a larger effect on alcohol consumption at 4 months (IRR 0.53, CoI 0.31-0.88; POE 99.4%). The evidence for the combined effect was also marked in the imputed analyses. Notably, relatively few participants drank above national guidelines with respect to weekly alcohol consumption, and therefore, the numbers included in these analyses were smaller than for other behaviors (see Figure 3 and Multimedia Appendices 7 and 8).

Frequency of Heavy Episodic Drinking

There was considerable variability in the evidence for the effectiveness of components on heavy episodic drinking, with evidence suggesting both benefit and harm. Screening and feedback (C1) reduced the frequency of heavy episodic drinking at 2 months (IRR 0.87, CoI 0.74-1.02; POE 95.2%), which was also evident in the imputed analyses. However, the effect was not sustained at the 4-month follow-up. The strongest evidence for benefit was observed for the combination of screening and feedback (C1) with goal-setting and planning (C2) (IRR 0.78, CoI 0.62-0.98; POE 98.4%), and screening and feedback (C1) with mindfulness (C5) (IRR 0.78, CoI 0.62-0.98; POE 98.4%) at 2 months. Similar evidence was found for the combination of goal-setting and planning (C2) with mindfulness (C5). Evidence of harm was most prominent for motivation (C3), both alone and in combination with goal-setting and planning (C2) or mindfulness (C5) (IRR 1.28, CoI 0.99-1.64; POE 97.1%) at the 4-month follow-up interval. These harmful effects were not evident at the 2-month interval. Notably, the evidence of harm was less prominent in the imputed analyses, whereas the evidence for benefit remained.

Daily Portions of Fruit and Vegetables

The strongest evidence for the beneficial effects of single components on fruit and vegetable consumption was found for screening and feedback (C1) and skills and know-how (C4) at both follow-up intervals. The largest effect was observed for the combination of these 2 components at 2 months (mean difference 0.30, CoI 0.11-0.48; POE 99.9%) and at 4 months (mean difference 0.26, CoI 0.05-0.46; POE 99.4%). While the evidence for motivation (C3) alone was not convincingly strong, in combination with screening and feedback (C1) or skills and know-how (C4), there was evidence of an effect at both the 2- and 4-month intervals on par with the effect observed for the combination of screening and feedback (C1) and skills and know-how (C4). These findings all persisted in the imputed analyses.

Weekly Consumption of Sugary Drinks

There was no strong evidence of the beneficial effects of the intervention components on sugary drink consumption. Rather, there was strong evidence of harm from self-authored SMS text messages (C6) at the 2- and 4-month intervals. The evidence suggested an increase of 15%-20% in weekly consumption of sugary drinks from this component alone, and the effect was larger in combination with goal-setting and planning (C2), motivation (C3), and skills and know-how (C4). These negative effects were partially sustained in the imputed analyses.

Weekly Moderate-to-Vigorous Physical Activity

There was evidence of beneficial effects on MVPA from several combinations of components at the 2-month interval; however, few of these effects persisted at the 4-month interval. The evidence for an effect of single components was strongest for motivation (C3), which increased weekly minutes of MVPA at 2 months (mean difference 35.8, CoI −0.41 to 71.9; POE 97.4%), and mindfulness (C5), which also increased minutes of MVPA at 2 months (mean difference 41.2, CoI 5.04-77.6; POE 98.7%). When motivation (C3) and mindfulness (C5) were combined, the effect on minutes of MVPA was even greater at 2 months (mean difference 78.0, CoI 28.3-128.2; POE 99.9%). Combining motivation (C3) with skills and know-how (C4) also had a positive effect on MVPA at 2 months. The imputed analyses produced similar findings.

While there was only weak evidence that screening and feedback (C1) alone affected MVPA, there was evidence that combining screening and feedback (C1) with skills and know-how (C4) increased minutes of MVPA at 4 months (mean difference 60.1, CoI 3.6-116.5; POE 98.2%), although the evidence was not as strong in the imputed analyses. In addition, there was evidence that screening and feedback (C1) combined with mindfulness (C5) increased minutes of MVPA at the 2-month interval (mean difference 50.6, CoI −0.55 to 101.9; POE 97.4%), with similar findings in the imputed analyses.

Smoking

There was weak evidence of both benefit and harm from the intervention components, and their combinations, on smoking cessation. The relatively strongest evidence for benefit was found for self-authored SMS text messages (C6) at 2 months (odds ratio 2.15, CoI 0.67-6.74; POE 90.4%), and for the combination of screening and feedback (C1) and motivation (C3) at 4 months (odds ratio 4.28, CoI 0.49-39.9; POE 90.7%). The strongest evidence for harm was found for the combination of goal-setting and planning (C2) and mindfulness (C5) at 2 months (odds ratio 0.23, CoI 0.04-1.27; POE 95.4%). The evidence for both benefit and harm persisted in the imputed analyses, although effect estimates were attenuated toward the null.

Secondary Outcomes

Weekly Portions of Candy and Snacks

At the individual component level, there was some evidence that motivation (C3) decreased candy and snack consumption at both 2 and 4 months. When motivation (C3) was combined with screening and feedback (C1), the strongest evidence for benefit was observed at 4 months (IRR 0.85, CoI 0.71-1.01; POE 96.8%). Increases in candy and snack consumption were observed for mindfulness (C5) alone at 4 months and self-authored SMS text messages (C6) alone at 2 months. While the evidence for these effects was not strong, it persisted in the

XSL•FO
RenderX

imputed analyses. In addition, combining mindfulness (C5) and self-authored SMS text messages (C6) increased candy and snack consumption at 4 months (IRR 1.21, CoI 1.01-1.44; POE 97.7%). These findings also persisted in the imputed analyses.

BMI

At the 2-month interval, there was evidence that mindfulness (C5) and self-authored SMS text messages (C6) increased BMI, both individually and when combined (combined mean difference 0.34, CoI 0.09-0.59; POE 99.6%). However, evidence for these effects was not observed at the 4-month interval. There was also evidence that combining screening and feedback (C1) with mindfulness (C5) resulted in higher BMI at both the 2- and 4-month intervals. These findings persisted in the imputed analyses.

Cigarettes Smoked Per Week

Regarding the number of cigarettes smoked per week among those who continued to smoke, there was evidence that the combination of screening and feedback (C1) and goal-setting and planning (C2) had a beneficial effect at 2 months, and that combining motivation (C3) and self-authored SMS text messages (C6) had a harmful effect at 2 months.

Perceived Stress

There was evidence of a beneficial effect on stress from the skills and know-how (C4) component alone at 4 months (mean difference −0.35, CoI −0.71 to 0.01; POE 97.1%). The effect size increased when skills and know-how (C4) was combined with screening and feedback (C1) or motivation (C3). These findings were not markedly different in the imputed analyses.

Post-Hoc Analyses

In Multimedia Appendix 9, we present estimates of the effects of the number of components participants had access to on primary and secondary outcomes. The relatively strongest evidence was found for fruit and vegetable consumption at 2 months (0.05 portion increase per module, CoI 0.001-0.11; POE 97.6%) and 4 months (0.04 portions per module, CoI −0.02 to 0.1; POE 90.7%), as well as for weekly MVPA at the 2-month follow-up (17.5 minutes per module, CoI 2.9-32.2; POE 99.0%) and 4 months (11.8 minutes per module, CoI −3.8 to 27.1; POE 93.1%).

Intervention Access

Of the 854 participants who had access to the screening and feedback component (C1), 807 (94.5%) screened themselves at least once during the intervention period. Among those who screened themselves at least once, 356 (44.1%) did so 4 or more times, while 129 (16.0%) screened themselves 3 times, 161 (20.0%) 2 times, and 161 (20.0%) 1 time.

Of the 850 participants who had access to the goal-setting and planning component (C2), 650 (76.5%) accessed the component at least once. Among all participants, 427 (50.2%) set at least one goal, and 403 (47.4%) selected or self-authored at least one challenge using the intervention's technical solution.

Of the 850 participants who had access to the motivation component (C3), 554 (65.2%) accessed the content at least once. Among all participants, 169 (19.9%) activated SMS text messages for physical activity, 166 (19.5%) for diet, 47 (5.5%) for alcohol, and 21 (2.5%) for smoking. The motivation component provided interactive content on motives, financial costs, and health consequences, with 294 recorded engagements using the intervention's technical solution.

Of the 855 participants who had access to the skills and know-how component (C4), 536 (62.7%) accessed the materials at least once. Among all participants, 166 (19.4%) activated SMS text messages for physical activity, 172 (20.1%) for diet, 51 (6.0%) for alcohol, and 28 (3.3%) for smoking. Most of the content in this module consisted of text and visual materials designed to build knowledge and skills for behavior change. A small number of interactive exercises were also included, with 283 recorded engagements using the intervention's technical solution.

Of the 855 participants who had access to the mindfulness component (C5), 435 (50.9%) accessed the component at least once. Meditations were accessed by 87 (10.2%) participants, and 156 (18.2%) completed at least one task. Among those who completed tasks, 76 (48.7%) completed 3 or more tasks, 33 (21.2%) completed 2 tasks, and 47 (30.1%) completed 1 task. Some tasks required participants to write longer texts, which could be completed either using the intervention's technical solution or offline; however, only activities completed using the technical solution could be measured.

Of the 850 participants who had access to the self-composed SMS text messages component (C6), 406 (47.8%) accessed the component at least once. Among those who accessed the component, a total of 570 SMS text messages were composed by 176 participants.

We had no means of accurately measuring the time spent on each component and therefore, cannot approximate the time spent reading the text and visual information provided in the components. Additionally, we had no means of measuring the extent to which participants comprehended and adopted the information provided. For instance, the skills and know-how component (C4) included mostly tips and information on concrete actions for participants to take, and we had no means of measuring how much time participants spent reading and engaging with the support provided.

Discussion

Principal Findings

This factorial randomized trial estimated the effectiveness of 6 behavior change components within a digital multiple health behavior intervention. We found evidence that both individual components and 2-way interactions between components led not only to notable improvements in health behaviors, but also to notable harm. The components screening/feedback (C1), motivation (C3), and skills/know-how (C4) each demonstrated individual and interacting improvements across various behavioral outcomes, including reductions in alcohol consumption, smoking, and stress, as well as increases in fruit and vegetable consumption and physical activity. Most effects were observed at 4 months; however, changes in heavy episodic

XSL•FO
RenderX

drinking and physical activity were noted only at the 2-month follow-up interval.

Previous factorial trials of health behavior change interventions have shown mixed results regarding the effectiveness of individual components. In general, combining components tends to lead to more substantial improvements in health behaviors compared with individual components alone. For instance, Crane et al [97] found that although individual modules of the Drink Less app did not affect weekly alcohol consumption or Alcohol Use Disorders Identification Test scores, 2-way interactions between components did. Specifically, the combination of normative feedback and cognitive bias retraining reduced weekly alcohol consumption, whereas self-monitoring, feedback, and action planning improved Alcohol Use Disorders Identification Test scores. Similarly, in our study, screening/feedback (C1) reduced heavy episodic drinking at 2 months, with enhanced effects when combined with goal-setting/planning (C2) or mindfulness (C5). Notably, the study by Crane et al [97] was not restricted to university students and focused solely on alcohol consumption.

Schroé et al [98] found that the MyPlan2.0 digital intervention increased MVPA and reduced sedentary behavior through self-monitoring. Additionally, coping planning further boosted MVPA, with the combination of action planning, coping planning, and self-monitoring proving most effective at 5 weeks. Similarly, Fanning et al [99] evaluated a 12-week physical activity app incorporating self-monitoring, goal-setting, and feedback, resulting in an 11-minute increase in daily MVPA (from 34 to 46 minutes). Participants who received points-based feedback engaged in an additional 6 minutes of MVPA per day. Our findings indicated that combining components such as screening/feedback (C1), motivation (C3), skills/know-how (C4), and mindfulness (C5) enhanced short-term physical activity. However, this increase was not sustained over time, suggesting that although the components were effective at 2 months, they did not support the longer-term maintenance of increased physical activity at 4 months. The dissimilarities in findings may be due to both Schroé et al [98] and Fanning et al [99] focusing exclusively on physical activity and not restricting their samples to university students.

The evidence showed no effect for the goal-setting/planning component (C2) when delivered alone. Harms were observed for the interaction between goal-setting/planning (C2) and mindfulness (C5) on smoking cessation at 2 months. Additionally, self-authored SMS text messages (C6) increased sugary drink consumption, with the effect amplified when combined with goal-setting/planning (C2), motivation (C3), or skills/know-how (C4). These findings regarding sugary drinks may be related to the observed decrease in alcohol consumption, as individuals might substitute sugary drinks for alcohol, and behavior substitution was part of the advice provided to participants. Additionally, students may consume energy drinks during studying and physical activity, further contributing to increased sugary beverage intake. Previous research [100] highlighted the complexity of young adults' perceptions of beverage healthfulness, which are influenced by factors including sugar content, packaging, marketing, and context of consumption. For example, a drink may be perceived as healthier if consumed during physical activity or as part of a balanced meal.

Although beneficial, health behavior interventions can increase stress and health anxiety due to heightened health consciousness, dissonance with desired behaviors, or pressure to meet goals, all of which may negatively affect mental well-being [101]. However, our study found that perceived psychological stress decreased at both the 2- and 4-month intervals when participants were provided with the skills/know-how component (C4). This effect became more pronounced over time, highlighting the importance of equipping individuals with practical skills and knowledge to manage their health in daily life.

Taking a broader perspective beyond the scope of factorial trials, our findings differ from previous digital multiple health interventions for young adults and university students, which have reported small effect sizes when evaluating interventions as a whole. For instance, Champion et al [102] reported that a digital multiple health intervention showed no evidence of effect in modifying lifestyle behaviors among adolescents in Australia; however, this study included a much younger demographic (ages 11-13) than our study. The online intervention U@Uni by Cameron et al [103], which targeted new university students, found no evidence of effect on the 4 primary outcomes: portions of fruit and vegetables per day, physical activity in the previous week, alcohol consumption in the previous week, and smoking status at the 6-month follow-up. However, U@Uni also targeted a younger demographic of incoming university students who may not have had the same motivation to change as students who had already established themselves. By contrast, similar to our study, Duan et al [104] demonstrated positive effects on physical activity and fruit and vegetable consumption after 8 weeks of a sequentially delivered web-based intervention among Chinese college students.

Generalizability and Limitations

This trial employed a pragmatic recruitment strategy, reaching students at universities across Sweden through channels typically used by Student Health Care Centers to provide health information and services. This open invitation, with broad inclusion criteria and low barriers to participation, targeted students with at least one health behavior categorized as a significant risk factor for their physical or mental health. Although the validity of trials is not contingent on the sample being representative of the target population, in our case, we anticipate that those who would use the intervention in a real-world rollout would be similar to those who participated in this trial. Therefore, we argue that this trial estimates the effectiveness, rather than the efficacy, of the intervention components. Relatedly, we do not have data to determine which recruitment source led to each individual's participation; thus, a limitation of the study design is that we cannot assess whether effectiveness was moderated by mode of recruitment.

In line with our aim to estimate effectiveness, we designed the study to gain knowledge about how access to components affects outcomes in a large population, rather than to examine the efficacy of different components in controlled environments. Effectiveness, as a concept, seeks to understand what happens under real-world conditions; for this reason, we neither expected

XSL•FO
RenderX

nor observed that all participants would use all of the intervention materials available to them. Nevertheless, in this study, which proactively recruited university students, 768 of 854 (89.9%) with access to the screening component used it at least once, and approximately one-half of participants with access to the other components used them at least once. Based on the measures available to us, the 2 components with the lowest usage were mindfulness (C5) and self-authored SMS text messages (C6). This is not surprising, as both require a stronger cognitive commitment, and mindfulness was not advertised to participants before enrollment; thus, there may have been limited capacity or interest to engage with these components. However, even if a smaller proportion of individuals engage with specific components, it may still be beneficial to include them in interventions, as they may have positive effects from a population perspective. Overall, the findings from this study should be understood as generalizing to the proactive dissemination of digital interventions rather than as controlled treatment of specific individuals.

The invitation reached individuals who were not necessarily actively seeking help but who could benefit from it and were willing to participate when approached. Although recruitment did not target students who had explicitly requested support for change, baseline measures indicated a gap between participants' ratings of importance and their perceived capacity to change, with importance rated, on average, 8 out of 10 and confidence and know-how rated 6 out of 10. This suggests that our proactive recruitment strategy can reach those who may benefit from, and be interested in, support. The relatively low baseline ratings of know-how and confidence also suggest potential to deliver effective behavior change support, as previous research shows that enhancing individuals' confidence in their ability to change and improving their knowledge of how to change can lead to positive behavior modifications, such as reduced alcohol consumption and smoking cessation [105,106]. It should be noted that these 3 measures did not constitute a primary objective of this trial, and to reduce participant burden, we used single-item measures, which cannot capture a more nuanced understanding of confidence, importance, and know-how. A related limitation is that, because participants were not necessarily actively seeking help, baseline screening leading to inclusion may have prompted reflection on behavior for some participants, which in itself may have induced change. Notably, this occurred before randomization; thus, the consequence may be that estimates are biased toward the null.

Attrition rates were high in this study; however, they were consistent with what can be expected in remote trials with low barriers to participation [107]. Attrition analyses did not indicate marked differences across factorial conditions, thereby strengthening the internal validity of the findings. However, there was some evidence that late responders and nonresponders exhibited healthier behaviors compared with early responders. This underscores the importance of the imputed analyses in supporting the main findings.

The factorial design ensured blinding by providing every participant with some form of intervention. Although social desirability is always a concern in studies in which the objectives are clear to participants—and possibly more so when outcomes are self-reported—blinding reduces the risk of compensatory rivalry and other research participation effects [108,109]. Therefore, the risk of bias stemming from the use of self-reported measures in this study is somewhat tempered. Regarding blinding, there was a risk that participants could have deliberately or inadvertently learned about others' participation and shared intervention content, thereby becoming partially aware of other conditions. Although we cannot determine the extent to which this occurred, we judge the risk of bias to be low. If such contamination did occur, our estimates of effectiveness are likely biased toward the null.

Finally, the factorial design used in this trial allowed us to study the individual components of Buddy; however, it limited our ability to estimate the total effect of the intervention as a whole. Therefore, future studies are needed to investigate the overall effects of the Buddy intervention. Indeed, post hoc analyses suggested that access to a greater number of components increased effectiveness with respect to fruit and vegetable consumption and physical activity.

Conclusions

Using a randomized factorial design, we dismantled a complex digital multiple behavior intervention and provided novel insights into the effectiveness of its different components. Both marginal and synergistic effects were estimated across multiple behaviors. Our results highlight which components of digital interventions contribute to effectiveness and demonstrate that component effectiveness is heterogeneous across the targeted behaviors. The findings provide an evidence base for designing more effective digital interventions. These findings should be considered in light of the risk of bias introduced by attrition to follow-up, which was high in this effectiveness trial with low barriers to participation. The trial also relied on self-reported measures, although the risk of bias associated with this design choice may have been partially mitigated by participant blinding.

Funding

This trial was funded by the Swedish Research Council for Health, Working Life and Welfare (grant 2018-01410; principal investigator: ML). The trial was also supported by funding from the Swedish Cancer Society (Cancerfonden; grants 20 0883 Pj and 23 2888 Pj; principal investigator: MB). The funders had no role in the planning, design, conduct, analysis, or reporting of the study.

XSL•FO
RenderX

Data Availability

The datasets used or analyzed in this study are available from the corresponding author (MB) upon reasonable request. Ethical approval is required for data sharing.

Authors' Contributions

Conceptualization: PB, MB, ML
Data curation: KÅ, MB
Formal analysis: KÅ, MB
Investigation: KÅ, MB
Methodology: MB, ML, PH
Project administration: KÅ, MB
Resources: KÅ, OL, HH, PH, ACE, PB, ML, MB
Software: MB
Writing – original draft: KÅ, MB
Writing – review & editing: KÅ, OL, HH, PH, ACE, PB, ML, MB

Conflicts of Interest

MB and PB own a private company (Alexit AB) that develops and distributes digital solutions to the public and private sectors. Alexit AB played no role in the development of the intervention, study design, data analysis, data interpretation, or writing of this report. All other authors declare no conflicts of interest.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 984 KB - jmir_v28i1e88884_app1.pdf]

Multimedia Appendix 2
Informed consent form.
[PDF File (Adobe PDF File), 48 KB - jmir_v28i1e88884_app2.pdf]

Multimedia Appendix 3
Questionnaires.
[PDF File (Adobe PDF File), 75 KB - jmir_v28i1e88884_app3.pdf]

Multimedia Appendix 4
Intervention description and allocation.
[PDF File (Adobe PDF File), 1567 KB - jmir_v28i1e88884_app4.pdf]

Multimedia Appendix 5
Attrition nonresponse.
[PDF File (Adobe PDF File), 841 KB - jmir_v28i1e88884_app5.pdf]

Multimedia Appendix 6
Attrition attempts.
[PDF File (Adobe PDF File), 95 KB - jmir_v28i1e88884_app6.pdf]

Multimedia Appendix 7
Complete-case estimates of effect.
[PDF File (Adobe PDF File), 265 KB - jmir_v28i1e88884_app7.pdf]

Multimedia Appendix 8
Effect estimates based on imputed data.
[PDF File (Adobe PDF File), 267 KB - jmir_v28i1e88884_app8.pdf]

Multimedia Appendix 9
Estimates of the effects of component count.

XSL•FO
RenderX

[PDF File (Adobe PDF File), 87 KB - jmir_v28i1e88884_app9.pdf]

References

1. GBD 2020 Alcohol Collaborators. Population-level risks of alcohol consumption by amount, geography, age, sex, and year: a systematic analysis for the Global Burden of Disease Study 2020. Lancet 2022 Jul 16;400(10347):185-235 [FREE Full text] [doi: 10.1016/S0140-6736(22)00847-9] [Medline: 35843246]
2. GBD 2019 Tobacco Collaborators. Spatial, temporal, and demographic patterns in prevalence of smoking tobacco use and attributable disease burden in 204 countries and territories, 1990-2019: a systematic analysis from the Global Burden of Disease Study 2019. Lancet 2021 Jun 19;397(10292):2337-2360 [FREE Full text] [doi: 10.1016/S0140-6736(21)01169-7] [Medline: 34051883]
3. GBD 2021 Diabetes Collaborators. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2023 Jul 15;402(10397):203-234 [FREE Full text] [doi: 10.1016/S0140-6736(23)01301-6] [Medline: 37356446]
4. GBD 2021 Risk Factors Collaborators. Global burden and strength of evidence for 88 risk factors in 204 countries and 811 subnational locations, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2024 May 18;403(10440):2162-2203 [FREE Full text] [doi: 10.1016/S0140-6736(24)00933-4] [Medline: 38762324]
5. World Health Organisation (WHO). Fact sheet Noncommunicable diseases. WHO. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases [accessed 2025-01-01]
6. Ashdown-Franks G, Vancampfort D, Firth J, Smith L, Sabiston CM, Stubbs B, et al. Association of leisure-time sedentary behavior with fast food and carbonated soft drink consumption among 133,555 adolescents aged 12-15 years in 44 low- and middle-income countries. Int J Behav Nutr Phys Act 2019 Apr 23;16(1):35 [FREE Full text] [doi: 10.1186/s12966-019-0796-3] [Medline: 31014359]
7. Patrick ME, Evans-Polce RJ, Arterberry BJ, Terry-McElrath Y. Initiation of and escalation to high-intensity drinking in young adults. JAMA Pediatr 2023 Mar 01;177(3):286-293 [FREE Full text] [doi: 10.1001/jamapediatrics.2022.5642] [Medline: 36716022]
8. Poole R, Carver H, Anagnostou D, Edwards A, Moore G, Smith P, et al. Tobacco use, smoking identities and pathways into and out of smoking among young adults: a meta-ethnography. Subst Abuse Treat Prev Policy 2022 Mar 28;17(1):24 [FREE Full text] [doi: 10.1186/s13011-022-00451-9] [Medline: 35346260]
9. Donat-Vargas C, Guerrero-Zotano, Casas A, Baena-Cañada JM, Lope V, Antolín S, et al. Trajectories of alcohol consumption during life and the risk of developing breast cancer. Br J Cancer 2021 Oct 06;125(8):1168-1176 [FREE Full text] [doi: 10.1038/s41416-021-01492-w] [Medline: 34483338]
10. Rodriguez-Ayllon M, Cadenas-Sánchez C, Estévez-López F, Muñoz NE, Mora-Gonzalez J, Migueles JH, et al. Role of physical activity and sedentary behavior in the mental health of preschoolers, children and adolescents: a systematic review and meta-analysis. Sports Med 2019 Sep 16;49(9):1383-1410. [doi: 10.1007/s40279-019-01099-5] [Medline: 30993594]
11. Guasch-Ferré M, Willett WC. The Mediterranean diet and health: a comprehensive overview. J Intern Med 2021 Sep 23;290(3):549-566 [FREE Full text] [doi: 10.1111/joim.13333] [Medline: 34423871]
12. Li Y, Schoufour J, Wang DD, Dhana K, Pan A, Liu X, et al. Healthy lifestyle and life expectancy free of cancer, cardiovascular disease, and type 2 diabetes: prospective cohort study. BMJ 2020 Jan 08;368:l6669 [FREE Full text] [doi: 10.1136/bmj.l6669] [Medline: 31915124]
13. Nyberg ST, Singh-Manoux A, Pentti J, Madsen IEH, Sabia S, Alfredsson L, et al. Association of healthy lifestyle with years lived without major chronic diseases. JAMA Intern Med 2020 May 01;180(5):760-768 [FREE Full text] [doi: 10.1001/jamainternmed.2020.0618] [Medline: 32250383]
14. Steffen LM, Van Horn L, Daviglus ML, Zhou X, Reis JP, Loria CM, et al. A modified Mediterranean diet score is associated with a lower risk of incident metabolic syndrome over 25 years among young adults: the CARDIA (Coronary Artery Risk Development in Young Adults) study. Br J Nutr 2014 Sep 19;112(10):1654-1661. [doi: 10.1017/s0007114514002633]
15. Alosaimi N, Sherar LB, Griffiths P, Pearson N. Clustering of diet, physical activity and sedentary behaviour and related physical and mental health outcomes: a systematic review. BMC Public Health 2023 Aug 18;23(1):1572 [FREE Full text] [doi: 10.1186/s12889-023-16372-6] [Medline: 37596591]
16. Kemel PN, Porter JE, Coombs N. Improving youth physical, mental and social health through physical activity: a systematic literature review. Health Promot J Austr 2022 Jul 15;33(3):590-601. [doi: 10.1002/hpja.553] [Medline: 34735738]
17. Solomou S, Logue J, Reilly S, Perez-Algorta G. A systematic review of the association of diet quality with the mental health of university students: implications in health education practice. Health Educ Res 2023 Jan 20;38(1):28-68 [FREE Full text] [doi: 10.1093/her/cyac035] [Medline: 36441584]
18. Bailey AP, Hetrick SE, Rosenbaum S, Purcell R, Parker AG. Treating depression with physical activity in adolescents and young adults: a systematic review and meta-analysis of randomised controlled trials. Psychol Med 2017 Oct 10;48(7):1068-1083. [doi: 10.1017/s0033291717002653]
19. Yonek JC, Meacham MC, Shumway M, Tolou-Shams M, Satre DD. Smoking reduction is associated with lower alcohol consumption and depressive symptoms among young adults over one year. Drug Alcohol Depend 2021 Oct 01;227:108922 [FREE Full text] [doi: 10.1016/j.drugalcdep.2021.108922] [Medline: 34364192]

XSL•FO
RenderX

20. Taylor G, Lindson N, Farley A, Leinberger-Jabari A, Sawyer K, Te Water Naudé R, et al. Smoking cessation for improving mental health. Cochrane Database Syst Rev 2021 Mar 09;3(3):CD013522 [FREE Full text] [doi: 10.1002/14651858.CD013522.pub2] [Medline: 33687070]
21. Haverkamp BF, Wiersma R, Vertessen K, van Ewijk H, Oosterlaan J, Hartman E. Effects of physical activity interventions on cognitive outcomes and academic performance in adolescents and young adults: a meta-analysis. J Sports Sci 2020 Dec 12;38(23):2637-2660 [FREE Full text] [doi: 10.1080/02640414.2020.1794763] [Medline: 32783695]
22. Seilo N, Paldanius S, Autio R, Kunttu K, Kaila M. Associations between e-health questionnaire responses, health checks and graduation: Finnish register-based study of 2011-2012 university entrants. BMJ Open 2020 Dec 16;10(12):e041551 [FREE Full text] [doi: 10.1136/bmjopen-2020-041551] [Medline: 33328260]
23. Cousins K, Connor JL, Kypri K. Effects of the Campus Watch intervention on alcohol consumption and related harm in a university population. Drug Alcohol Depend 2014 Oct 01;143:120-126. [doi: 10.1016/j.drugalcdep.2014.07.015] [Medline: 25108584]
24. Cho H, Yang Y. Relationship between alcohol consumption and risky sexual behaviors among adolescents and young adults: a meta-analysis. Int J Public Health 2023;68:1605669 [FREE Full text] [doi: 10.3389/ijph.2023.1605669] [Medline: 37153699]
25. Pengpid S, Peltzer K. Prevalence and correlates of multiple behavioural risk factors of non-communicable diseases among university students from 24 countries. J Public Health (Oxf) 2021 Dec 10;43(4):857-866. [doi: 10.1093/pubmed/fdaa138] [Medline: 34918087]
26. Zhang Y, Pan X, Chen J, Cao A, Xia L, Zhang Y, et al. Combined lifestyle factors, all-cause mortality and cardiovascular disease: a systematic review and meta-analysis of prospective cohort studies. J Epidemiol Community Health 2021 Jan 05;75(1):92-99. [doi: 10.1136/jech-2020-214050] [Medline: 32892156]
27. Oosterveen E, Tzelepis F, Ashton L, Hutchesson MJ. A systematic review of eHealth behavioral interventions targeting smoking, nutrition, alcohol, physical activity and/or obesity for young adults. Prev Med 2017 Jun;99:197-206. [doi: 10.1016/j.ypmed.2017.01.009] [Medline: 28130046]
28. Schoeppe S, Alley S, Van Lippevelde W, Bray NA, Williams SL, Duncan MJ, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2016 Dec 07;13(1):127 [FREE Full text] [doi: 10.1186/s12966-016-0454-y] [Medline: 27927218]
29. Mönninghoff A, Kramer JN, Hess AJ, Ismailova K, Teepe GW, Tudor Car L, et al. Long-term effectiveness of mHealth physical activity interventions: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2021 Apr 30;23(4):e26699. [doi: 10.2196/26699]
30. Kaner E, Beyer F, Muirhead C, Campbell F, Pienaar E, Bertholet N, et al. Effectiveness of brief alcohol interventions in primary care populations. Cochrane Database Syst Rev 2018 Feb 24;2(2):CD004148 [FREE Full text] [doi: 10.1002/14651858.CD004148.pub4] [Medline: 29476653]
31. O Rourke L, Humphris G, Baldacchino A. Electronic communication based interventions for hazardous young drinkers: a systematic review. Neurosci Biobehav Rev 2016 Sep;68:880-890. [doi: 10.1016/j.neubiorev.2016.07.021] [Medline: 27453570]
32. Taylor G, Dalili M, Semwal M, Civljak M, Sheikh A, Car J. Internet-based interventions for smoking cessation. Cochrane Database Syst Rev 2017 Sep 04;9(9):CD007078 [FREE Full text] [doi: 10.1002/14651858.CD007078.pub5] [Medline: 28869775]
33. McIntosh J, Jay S, Hadden N, Whittaker P. Do E-health interventions improve physical activity in young people: a systematic review. Public Health 2017 Jul;148:140-148. [doi: 10.1016/j.puhe.2017.04.001]
34. Montagni I, Cariou T, Feuillet T, Langlois E, Tzourio C. Exploring digital health use and opinions of university students: field survey study. JMIR Mhealth Uhealth 2018 Mar 15;6(3):e65 [FREE Full text] [doi: 10.2196/mhealth.9131] [Medline: 29549071]
35. Berman AH, Gajecki M, Sinadinovic K, Andersson C. Mobile interventions targeting risky drinking among university students: a review. Curr Addict Rep 2016 Apr 5;3(2):166-174 [FREE Full text] [doi: 10.1007/s40429-016-0099-6] [Medline: 27226948]
36. Prosser T, Gee KA, Jones F. A meta-analysis of effectiveness of E-interventions to reduce alcohol consumption in college and university students. J Am Coll Health 2018 Apr 19;66(4):292-301. [doi: 10.1080/07448481.2018.1440579] [Medline: 29452058]
37. Dute DJ, Bemelmans WJE, Breda J. Using mobile apps to promote a healthy lifestyle among adolescents and students: a review of the theoretical basis and lessons learned. JMIR Mhealth Uhealth 2016 May 05;4(2):e39 [FREE Full text] [doi: 10.2196/mhealth.3559] [Medline: 27150850]
38. Brown J. A review of the evidence on technology-based interventions for the treatment of tobacco dependence in college health. Worldviews Evid Based Nurs 2013 Aug 19;10(3):150-162. [doi: 10.1111/wvn.12000] [Medline: 23421669]
39. McCambridge J, Bendtsen M, Karlsson N, White IR, Nilsen P, Bendtsen P. Alcohol assessment and feedback by email for university students: main findings from a randomised controlled trial. Br J Psychiatry 2013 Nov 02;203(5):334-340 [FREE Full text] [doi: 10.1192/bjp.bp.113.128660] [Medline: 24072758]

XSL•FO
RenderX

40. Thomas K, Müssener U, Linderoth C, Karlsson N, Bendtsen P, Bendtsen M. Effectiveness of a text messaging-based intervention targeting alcohol consumption among university students: randomized controlled trial. JMIR Mhealth Uhealth 2018 Jun 25;6(6):e146 [FREE Full text] [doi: 10.2196/mhealth.9642] [Medline: 29941417]
41. Bendtsen M. An electronic screening and brief intervention for hazardous and harmful drinking among Swedish university students: reanalysis of findings from a randomized controlled trial using a Bayesian framework. J Med Internet Res 2019 Dec 17;21(12):e14420 [FREE Full text] [doi: 10.2196/14420] [Medline: 31845903]
42. Berman AH, Andersson C, Gajecki M, Rosendahl I, Sinadinovic K, Blankers M. Smartphone apps targeting hazardous drinking patterns among university students show differential subgroup effects over 20 weeks: results from a randomized, controlled trial. J Clin Med 2019 Oct 28;8(11):1807 [FREE Full text] [doi: 10.3390/jcm8111807] [Medline: 31661868]
43. Müssener U, Bendtsen M, Karlsson N, White IR, McCambridge J, Bendtsen P. Effectiveness of short message service text-based smoking cessation intervention among university students: a randomized clinical trial. JAMA Intern Med 2016 Mar 01;176(3):321-328 [FREE Full text] [doi: 10.1001/jamainternmed.2015.8260] [Medline: 26903176]
44. Bendtsen M, Müssener U, Linderoth C, Thomas K. A mobile health intervention for mental health promotion among university students: randomized controlled trial. JMIR Mhealth Uhealth 2020 Mar 20;8(3):e17208 [FREE Full text] [doi: 10.2196/17208] [Medline: 32196462]
45. Chow CK, Redfern J, Hillis GS, Thakkar J, Santo K, Hackett ML, et al. Effect of lifestyle-focused text messaging on risk factor modification in patients with coronary heart disease: a randomized clinical trial. JAMA 2015 Sep 22;314(12):1255-1263. [doi: 10.1001/jama.2015.10945] [Medline: 26393848]
46. Palmer M, Sutherland J, Barnard S, Wynne A, Rezel E, Doel A, et al. The effectiveness of smoking cessation, physical activity/diet and alcohol reduction interventions delivered by mobile phones for the prevention of non-communicable diseases: a systematic review of randomised controlled trials. PLoS One 2018 Jan 5;13(1):e0189801 [FREE Full text] [doi: 10.1371/journal.pone.0189801] [Medline: 29304148]
47. Milne-Ives M, Lam C, De Cock C, Van Velthoven MH, Meinert E. Mobile apps for health behavior change in physical activity, diet, drug and alcohol use, and mental health: systematic review. JMIR Mhealth Uhealth 2020 Mar 18;8(3):e17046 [FREE Full text] [doi: 10.2196/17046] [Medline: 32186518]
48. Atorkey P, Paul C, Wiggers J, Bonevski B, Nolan E, Oldmeadow C, et al. Clustering of multiple health-risk factors among vocational education students: a latent class analysis. Transl Behav Med 2021 Oct 23;11(10):1931-1940. [doi: 10.1093/tbm/ibab068] [Medline: 34155507]
49. Noble N, Paul C, Turon H, Oldmeadow C. Which modifiable health risk behaviours are related? A systematic review of the clustering of Smoking, Nutrition, Alcohol and Physical activity ('SNAP') health risk factors. Prev Med 2015 Dec;81:16-41. [doi: 10.1016/j.ypmed.2015.07.003] [Medline: 26190368]
50. Jonsson KR, Corell M, Löfstedt P, Adjei NK. The clustering of multiple health and lifestyle behaviors among Swedish adolescents: a person-oriented analysis. Front Public Health 2023;11:1178353 [FREE Full text] [doi: 10.3389/fpubh.2023.1178353] [Medline: 37538263]
51. Mair J, Salamanca-Sanabria A, Augsburger M, Frese B, Abend S, Jakob R, et al. Effective behavior change techniques in digital health interventions for the prevention or management of noncommunicable diseases: an umbrella review. Ann Behav Med 2023 Sep 13;57(10):817-835 [FREE Full text] [doi: 10.1093/abm/kaad041] [Medline: 37625030]
52. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010 Feb 17;12(1):e4 [FREE Full text] [doi: 10.2196/jmir.1376] [Medline: 20164043]
53. Dugas M, Gao G, Agarwal R. Unpacking mHealth interventions: a systematic review of behavior change techniques used in randomized controlled trials assessing mHealth effectiveness. Digit Health 2020;6:2055207620905411 [FREE Full text] [doi: 10.1177/2055207620905411] [Medline: 32128233]
54. Aguiar M, Trujillo M, Chaves D, Álvarez R, Epelde G. mHealth apps using behavior change techniques to self-report data: systematic review. JMIR Mhealth Uhealth 2022 Sep 09;10(9):e33247 [FREE Full text] [doi: 10.2196/33247] [Medline: 36083606]
55. Pham Q, Wiljer D, Cafazzo JA. Beyond the randomized controlled trial: a review of alternatives in mHealth clinical trial methods. JMIR Mhealth Uhealth 2016 Sep 09;4(3):e107 [FREE Full text] [doi: 10.2196/mhealth.5720] [Medline: 27613084]
56. Bendtsen M, Bendtsen P, Henriksson H, Henriksson P, Müssener U, Thomas K, et al. The Mobile Health Multiple Lifestyle Behavior Interventions Across the Lifespan (MoBILE) research program: protocol for development, evaluation, and implementation. JMIR Res Protoc 2020 Apr 20;9(4):e14894 [FREE Full text] [doi: 10.2196/14894] [Medline: 32310147]
57. Åsberg K, Lundgren O, Henriksson H, Henriksson P, Bendtsen P, Löf M, et al. Multiple lifestyle behaviour mHealth intervention targeting Swedish college and university students: protocol for the Buddy randomised factorial trial. BMJ Open 2021 Dec 30;11(12):e051044. [doi: 10.1136/bmjopen-2021-051044]
58. Schulz K, Altman D, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340:c332 [FREE Full text] [doi: 10.1136/bmj.c332] [Medline: 20332509]
59. Kahan BC, Hall SS, Beller EM, Birchenall M, Chan A, Elbourne D, et al. Reporting of factorial randomized trials: extension of the CONSORT 2010 statement. JAMA 2023 Dec 05;330(21):2106-2114. [doi: 10.1001/jama.2023.19793] [Medline: 38051324]

XSL•FO
RenderX

60. Conner M, Norman P. Predicting and Changing Health Behaviour: Research and Practice with Social Cognition Models. Berkshire, UK: McGraw-Hill Education (UK); 2015.
61. Bendtsen M, Åsberg K, McCambridge J. Effectiveness of a digital intervention versus alcohol information for online help-seekers in Sweden: a randomised controlled trial. BMC Med 2022 May 17;20(1):176 [FREE Full text] [doi: 10.1186/s12916-022-02374-5] [Medline: 35578276]
62. Blomqvist J, Gunnarsson KU, Bendtsen P, Bendtsen M. Effects of a text messaging smoking cessation intervention amongst online help-seekers and primary health care visitors: findings from a randomised controlled trial. BMC Med 2023 Oct 04;21(1):382 [FREE Full text] [doi: 10.1186/s12916-023-03073-5] [Medline: 37794399]
63. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The Behavior Change Technique Taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Mar 20;46(1):81-95. [doi: 10.1007/s12160-013-9486-6]
64. Michie S, Whittington C, Hamoudi Z, Zarnani F, Tober G, West R. Identification of behaviour change techniques to reduce excessive alcohol consumption. Addiction 2012 Apr 17;107(8):1431-1440. [doi: 10.1111/j.1360-0443.2012.03845.x]
65. Bendtsen P, McCambridge J, Bendtsen M, Karlsson N, Nilsen P. Effectiveness of a proactive mail-based alcohol internet intervention for university students: dismantling the assessment and feedback components in a randomized controlled trial. J Med Internet Res 2012 Oct 31;14(5):e142 [FREE Full text] [doi: 10.2196/jmir.2062] [Medline: 23113955]
66. Bendtsen P, Bendtsen M, Karlsson N, White IR, McCambridge J. Online alcohol assessment and feedback for hazardous and harmful drinkers: findings from the AMADEUS-2 randomized controlled trial of routine practice in Swedish universities. J Med Internet Res 2015 Jul 09;17(7):e170 [FREE Full text] [doi: 10.2196/jmir.4020] [Medline: 26159179]
67. Michie S, Abraham C, Whittington C, McAteer J, Gupta S. Effective techniques in healthy eating and physical activity interventions: a meta-regression. Health Psychol 2009 Nov;28(6):690-701. [doi: 10.1037/a0016136] [Medline: 19916637]
68. Murray JM, Brennan SF, French DP, Patterson CC, Kee F, Hunter RF. Effectiveness of physical activity interventions in achieving behaviour change maintenance in young and middle aged adults: a systematic review and meta-analysis. Soc Sci Med 2017 Nov;192:125-133. [doi: 10.1016/j.socscimed.2017.09.021] [Medline: 28965003]
69. Knittle K, Nurmi J, Crutzen R, Hankonen N, Beattie M, Dombrowski SU. How can interventions increase motivation for physical activity? A systematic review and meta-analysis. Health Psychol Rev 2018 Sep 15;12(3):211-230. [doi: 10.1080/17437199.2018.1435299] [Medline: 29385950]
70. Howlett N, Trivedi D, Troop N, Chater A. Are physical activity interventions for healthy inactive adults effective in promoting behavior change and maintenance, and which behavior change techniques are effective? A systematic review and meta-analysis. Transl Behav Med 2019 Jan 01;9(1):147-157 [FREE Full text] [doi: 10.1093/tbm/iby010] [Medline: 29506209]
71. Ashton LM, Sharkey T, Whatnall MC, Williams RL, Bezzina A, Aguiar EJ, et al. Effectiveness of interventions and behaviour change techniques for improving dietary intake in young adults: a systematic review and meta-analysis of RCTs. Nutrients 2019 Apr 11;11(4):825 [FREE Full text] [doi: 10.3390/nu11040825] [Medline: 30979065]
72. Garnett C, Crane D, Brown J, Kaner E, Beyer F, Muirhead C, et al. Behavior change techniques used in digital behavior change interventions to reduce excessive alcohol consumption: a meta-regression. Ann Behav Med 2018 May 18;52(6):530-543 [FREE Full text] [doi: 10.1093/abm/kax029] [Medline: 29788261]
73. McCrabb S, Baker A, Attia J, Skelton E, Twyman L, Palazzi K, et al. Internet-based programs incorporating behavior change techniques are associated with increased smoking cessation in the general population: a systematic review and meta-analysis. Ann Behav Med 2019 Feb 01;53(2):180-195. [doi: 10.1093/abm/kay026] [Medline: 29750240]
74. Ek A, Alexandrou C, Söderström E, Bergman P, Delisle Nyström C, Direito A, et al. Effectiveness of a 3-month mobile phone-based behavior change program on active transportation and physical activity in adults: randomized controlled trial. JMIR Mhealth Uhealth 2020 Jun 08;8(6):e18531 [FREE Full text] [doi: 10.2196/18531] [Medline: 32510462]
75. Newby K, Teah G, Cooke R, Li X, Brown K, Salisbury-Finch B, et al. Do automated digital health behaviour change interventions have a positive effect on self-efficacy? A systematic review and meta-analysis. Health Psychol Rev 2021 Mar 20;15(1):140-158. [doi: 10.1080/17437199.2019.1705873] [Medline: 31847702]
76. Müssener U, Bendtsen M, Karlsson N, White IR, McCambridge J, Bendtsen P. SMS-based smoking cessation intervention among university students: study protocol for a randomised controlled trial (NEXit trial). Trials 2015 Apr 08;16:140 [FREE Full text] [doi: 10.1186/s13063-015-0640-2] [Medline: 25872503]
77. Thomas K, Bendtsen M, Linderoth C, Karlsson N, Bendtsen P, Müssener U. Short message service (SMS)-based intervention targeting alcohol consumption among university students: study protocol of a randomized controlled trial. Trials 2017 Apr 04;18(1):156 [FREE Full text] [doi: 10.1186/s13063-017-1898-3] [Medline: 28372563]
78. Müssener U, Bendtsen M, McCambridge J, Bendtsen P. User satisfaction with the structure and content of the NEXit intervention, a text messaging-based smoking cessation programme. BMC Public Health 2016 Nov 22;16(1):1179 [FREE Full text] [doi: 10.1186/s12889-016-3848-5] [Medline: 27876031]
79. Müssener U, Thomas K, Linderoth C, Leijon M, Bendtsen M. A text message-based intervention targeting alcohol consumption among university students: user satisfaction and acceptability study. JMIR Hum Factors 2018 Jul 10;5(3):e23 [FREE Full text] [doi: 10.2196/humanfactors.9641] [Medline: 29991469]

XSL•FO
RenderX

80. Thomas K, Linderoth C, Bendtsen M, Bendtsen P, Müssener U. Text message-based intervention targeting alcohol consumption among university students: findings from a formative development study. JMIR Mhealth Uhealth 2016 Oct 20;4(4):e119 [FREE Full text] [doi: 10.2196/mhealth.5863] [Medline: 27765732]
81. Kaner E, Beyer F, Garnett C, Crane D, Brown J, Muirhead C, et al. Personalised digital interventions for reducing hazardous and harmful alcohol consumption in community-dwelling populations. Cochrane Database Syst Rev 2017 Sep 25;9(9):CD011479 [FREE Full text] [doi: 10.1002/14651858.CD011479.pub2] [Medline: 28944453]
82. Black N, Johnston M, Michie S, Hartmann-Boyce J, West R, Viechtbauer W, et al. Behaviour change techniques associated with smoking cessation in intervention and comparator groups of randomized controlled trials: a systematic review and meta-regression. Addiction 2020 Nov 16;115(11):2008-2020. [doi: 10.1111/add.15056] [Medline: 32196796]
83. Crane RS, Brewer J, Feldman C, Kabat-Zinn J, Santorelli S, Williams JMG, et al. What defines mindfulness-based programs? The warp and the weft. Psychol Med 2016 Dec 29;47(6):990-999. [doi: 10.1017/s0033291716003317]
84. Creswell JD. Mindfulness interventions. Annu Rev Psychol 2017 Jan 03;68(1):491-516. [doi: 10.1146/annurev-psych-042716-051139] [Medline: 27687118]
85. Wong SYS, Chan JYC, Zhang D, Lee EKP, Tsoi KKF. The safety of mindfulness-based interventions: a systematic review of randomized controlled trials. Mindfulness 2018 Feb 2;9(5):1344-1357. [doi: 10.1007/s12671-018-0897-0]
86. Galante J, Friedrich C, Dawson AF, Modrego-Alarcón M, Gebbing P, Delgado-Suárez I, et al. Mindfulness-based programmes for mental health promotion in adults in nonclinical settings: a systematic review and meta-analysis of randomised controlled trials. PLoS Med 2021 Jan 11;18(1):e1003481 [FREE Full text] [doi: 10.1371/journal.pmed.1003481] [Medline: 33428616]
87. Neff KD, Germer CK. A pilot study and randomized controlled trial of the mindful self-compassion program. J Clin Psychol 2013 Jan 15;69(1):28-44. [doi: 10.1002/jclp.21923] [Medline: 23070875]
88. Crawford J, Collier E, Bendtsen M. Individualized treatment effects of a digital alcohol intervention and their associations with participant characteristics and engagement. Alcohol Alcohol 2024 Jul 21;59(5):agae049 [FREE Full text] [doi: 10.1093/alcalc/agae049] [Medline: 39034147]
89. Rehm J. Measuring quantity, frequency, and volume of drinking. Alcohol Clin Exp Res 1998 Apr;22(2 Suppl):4S-14S. [doi: 10.1097/00000374-199802001-00002] [Medline: 9603301]
90. Shorter GW, Bray JW, Giles EL, O'Donnell AJ, Berman AH, Holloway A, et al. The variability of outcomes used in efficacy and effectiveness trials of alcohol brief interventions: a systematic review. J Stud Alcohol Drugs 2019 May;80(3):286-298. [doi: 10.15288/jsad.2019.80.286]
91. Shorter GW, Bray JW, Heather N, Berman AH, Giles EL, Clarke M, et al. The "Outcome Reporting in Brief Intervention Trials: Alcohol" (ORBITAL) core outcome set: international consensus on outcomes to measure in efficacy and effectiveness trials of alcohol brief interventions. J Stud Alcohol Drugs 2021 Sep;82(5):638-646 [FREE Full text] [doi: 10.15288/jsad.2021.82.638] [Medline: 34546911]
92. Swedish National Board of Health and Welfare. Swedish National Guidelines for Prevention and Treatment of Unhealthy Lifestyle Behaviours [Nationella Riktlinjer: Prevention och Behandling vid Ohälsosamma Levnadsvanor]. Stockholm, Sweden: Swedish National Board of Health and Welfare; 2018. URL: https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/artikelkatalog/nationella-riktlinjer/2018-6-24.pdf [accessed 2026-02-24]
93. SRNT Subcommittee on Biochemical Verification. Biochemical verification of tobacco use and cessation. Nicotine Tob Res 2002 May 1;4(2):149-159. [doi: 10.1080/14622200210123581] [Medline: 12028847]
94. Vallejo MA, Vallejo-Slocker L, Fernández-Abascal EG, Mañanes G. Determining factors for stress perception assessed with the Perceived Stress Scale (PSS-4) in Spanish and other European samples. Front Psychol 2018;9:37 [FREE Full text] [doi: 10.3389/fpsyg.2018.00037] [Medline: 29434563]
95. Bendtsen M. A Gentle Introduction to the Comparison Between Null Hypothesis Testing and Bayesian Analysis: Reanalysis of Two Randomized Controlled Trials. J Med Internet Res 2018 Oct 24;20(10):e10873. [doi: 10.2196/10873]
96. Bendtsen M. Avoiding under- and overrecruitment in behavioral intervention trials using Bayesian sequential designs: tutorial. J Med Internet Res 2022 Dec 16;24(12):e40730 [FREE Full text] [doi: 10.2196/40730] [Medline: 36525297]
97. Crane D, Garnett C, Michie S, West R, Brown J. A smartphone app to reduce excessive alcohol consumption: identifying the effectiveness of intervention components in a factorial randomised control trial. Sci Rep 2018 Mar 12;8(1):4384 [FREE Full text] [doi: 10.1038/s41598-018-22420-8] [Medline: 29531280]
98. Schroé H, Van Dyck D, De Paepe A, Poppe L, Loh WW, Verloigne M, et al. Which behaviour change techniques are effective to promote physical activity and reduce sedentary behaviour in adults: a factorial randomized trial of an e- and m-health intervention. Int J Behav Nutr Phys Act 2020 Oct 07;17(1):127 [FREE Full text] [doi: 10.1186/s12966-020-01001-x] [Medline: 33028335]
99. Fanning J, Roberts S, Hillman CH, Mullen SP, Ritterband L, McAuley E. A smartphone "app"-delivered randomized factorial trial targeting physical activity in adults. J Behav Med 2017 Oct;40(5):712-729. [doi: 10.1007/s10865-017-9838-y] [Medline: 28255750]
100. Brownbill AL, Braunack-Mayer AJ, Miller CL. What makes a beverage healthy? A qualitative study of young adults' conceptualisation of sugar-containing beverage healthfulness. Appetite 2020 Jul 01;150:104675. [doi: 10.1016/j.appet.2020.104675] [Medline: 32184093]

XSL•FO
RenderX

101. Cho H, Salmon C. Unintended effects of health communication campaigns. Journal of Communication 2006;57(2):317. [doi: 10.1111/j.1460-2466.2007.00344.x]
102. Champion KE, Newton NC, Gardner LA, Chapman C, Thornton L, Slade T, et al. Health4Life eHealth intervention to modify multiple lifestyle risk behaviours among adolescent students in Australia: a cluster-randomised controlled trial. The Lancet Digital Health 2023 May;5(5):e276-e287. [doi: 10.1016/s2589-7500(23)00028-6]
103. Epton T, Norman P, Dadzie AS, Harris PR, Webb TL, Sheeran P, et al. A theory-based online health behaviour intervention for new university students (U@Uni): results from a randomised controlled trial. BMC Public Health 2014 Jun 05;14:563 [FREE Full text] [doi: 10.1186/1471-2458-14-563] [Medline: 24903620]
104. Duan Y, Liang W, Wang Y, Lippke S, Lin Z, Shang B, et al. The effectiveness of sequentially delivered web-based interventions on promoting physical activity and fruit-vegetable consumption among Chinese college students: mixed methods study. J Med Internet Res 2022 Jan 26;24(1):e30566 [FREE Full text] [doi: 10.2196/30566] [Medline: 35080497]
105. Bendtsen M, Åsberg K, McCambridge J. Mediators of effects of a digital alcohol intervention for online help-seekers: findings from an effectiveness trial. Drug Alcohol Depend 2023 Oct 01;251:110957 [FREE Full text] [doi: 10.1016/j.drugalcdep.2023.110957] [Medline: 37688979]
106. Crawford J, Blomqvist J, Gunnarsson KU, Bendtsen P, Bendtsen M. Mediated effects of a randomised control trial for a text messaging smoking cessation intervention for online help-seekers and primary care visitors. BMC Public Health 2024 Jul 09;24(1):1824 [FREE Full text] [doi: 10.1186/s12889-024-19273-4] [Medline: 38977972]
107. Gold N, Yau A, Rigby B, Dyke C, Remfry EA, Chadborn T. Effectiveness of digital interventions for reducing behavioral risks of cardiovascular disease in nonclinical adult populations: systematic review of reviews. J Med Internet Res 2021 May 14;23(5):e19688 [FREE Full text] [doi: 10.2196/19688] [Medline: 33988126]
108. MacNeill V, Foley M, Quirk A, McCambridge J. Shedding light on research participation effects in behaviour change trials: a qualitative study examining research participant experiences. BMC Public Health 2016 Jan 29;16:91 [FREE Full text] [doi: 10.1186/s12889-016-2741-6] [Medline: 26825583]
109. Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. Lancet 2002 Feb 23;359(9307):696-700. [doi: 10.1016/S0140-6736(02)07816-9] [Medline: 11879884]

Abbreviations

BCT: behavior change technique
CoI: compatibility interval
CONSORT: Consolidated Standards of Reporting Trials
IRR: incidence rate ratio
ISRCTN: International Standard Randomised Controlled Trial Number
MVPA: moderate-to-vigorous physical activity
POE: probability of effect

Edited by A Stone; submitted 03.Dec.2025; peer-reviewed by S Ondersma; comments to author 05.Jan.2026; revised version received 16.Feb.2026; accepted 17.Feb.2026; published 09.Mar.2026.

Please cite as:
Åsberg K, Lundgren O, Henriksson H, Henriksson P, Eldh AC, Bendtsen P, Löf M, Bendtsen M
Effectiveness of the Components of a Digital Multiple Health Behavior Intervention Among University Students (Buddy): Factorial Randomized Trial
J Med Internet Res 2026;28:e88884
URL: https://www.jmir.org/2026/1/e88884
doi:10.2196/88884
PMID:

©Katarina Åsberg, Oskar Lundgren, Hanna Henriksson, Pontus Henriksson, Ann Catrine Eldh, Preben Bendtsen, Marie Löf, Marcus Bendtsen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Telepathology and Mobile Health System for Province-Wide Pathology Consultation in Henan, China: Retrospective Evaluation Study

Jinming Shi[1*], MEng; Ming Ye[1*], MEng; Dongxu Sun[1*], MEng; Xianying He[1], MSc; Yaoen Lu[1], MEng; Linlin Wang[1], MSc; Haotian Chen[1], MEng; Wenchao Wang[1], MEng; Jie Zhao[1,2], PhD; Fangfang Cui[1,3], MSc

[1]National Engineering Laboratory for Internet Medical Systems and Applications, First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan, China

[2]Shanghai Artificial Intelligence Laboratory, Shanghai, China

[3]Institute of Intelligent Medicine, Henan Academy of Innovations in Medical Sciences, Zhengzhou, Henan, China

[*]these authors contributed equally

Corresponding Author:
Fangfang Cui, MSc
National Engineering Laboratory for Internet Medical Systems and Applications
First Affiliated Hospital of Zhengzhou University
No. 3, Jianshe East Road
Zhengzhou, Henan, 450000
China
Phone: 86 15093159748
Email: fcccuiff@zzu.edu.cn

Abstract

Background: Telepathology has emerged as a transformative digital health solution to address the global shortage of pathologists and the unequal distribution of diagnostic services, particularly in underserved and rural areas. In Henan Province, China, high diagnostic demand, rapid population growth, and limited pathology expertise exacerbate regional health care inequities, leading to delayed diagnoses and restricted access to specialist care.

Objective: This study aimed to design, implement, and evaluate a province-wide telepathology system integrating web and mobile platforms to enhance diagnostic quality, efficiency, and equitable access across health care tiers.

Methods: We conducted a retrospective, multicenter observational study using deidentified data from 120 health care institutions between 2016 and 2024. The system used a 3-tier architecture with virtual private network–secured transmission and a Browser-Server framework, supporting standardized whole-slide image acquisition, remote review, and reporting via web interfaces and a WeChat (Tencent) mini-program. System performance was assessed by consultation volume, turnaround time, concurrency, and diagnostic concordance in a subset of 1027 cases with paired tertiary-hospital expert diagnoses. Economic impact was estimated using previously published per-case savings, reflecting patient travel and ancillary cost reductions. Additional assessments included workflow integration, mobile platform use, and system stability under peak load.

Results: Over 8 years, the network processed 72,916 consultations encompassing 355,104 whole-slide images, supporting 220-300 concurrent users with stable performance. Median turnaround time was 10.06 (IQR 1.63-29.10) hours, with 96.41% (70,298/72,916) of cases completed within 72 hours. County-level hospitals contributed 77.63% (56,603/72,916) of consultations, demonstrating substantial engagement from lower-tier institutions. In the diagnostic subset, originating-site preliminary classifications achieved 0.90 sensitivity and 0.75 specificity relative to expert reference diagnoses, with 17.2% discordance corrected through remote expert review. Estimated annual direct cost savings ranged from US $0.14 to $0.63 million. Mobile-enabled access facilitated remote review and reporting without compromising data security, supporting integration into routine clinical workflows across diverse hospital settings.

Conclusions: The Henan Province telepathology system demonstrates that a centrally coordinated, scalable digital health platform can improve diagnostic efficiency, quality, and equity in resource-constrained settings. High county-level hospital use highlights its potential to reduce geographic and structural diagnostic inequities. Future work should explore formal cost-effectiveness evaluation, artificial intelligence–assisted diagnostic support, and cross-regional interoperability to enable broader adoption and sustainable integration into health care systems.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e75172) doi:10.2196/75172

KEYWORDS

telepathology; diagnostic equity; digital pathology; remote diagnostics; rural health; health systems innovation

Introduction

The digitization of pathology workflows has revolutionized diagnostic practices. High-throughput slide scanning and standardized data management have enabled unprecedented improvements in service accessibility and interinstitutional collaboration [1]. Digital pathology (DP) offers multiple benefits, including enhanced diagnostic accuracy, faster turnaround times (TAT), and increased efficiency in pathology workflows [2-5]. However, translating these technological advancements into equitable real-world health care delivery remains constrained by 2 systemic barriers: the geographic concentration of pathology expertise and the lack of scalable infrastructure to support cross-regional coordination [6]. In Henan Province, China, a region with a population exceeding 100 million and reporting approximately 279,740 new cancer cases annually [7], these challenges are particularly pronounced. The province faces a critical shortage of pathologists, who constitute only 0.4% of the medical workforce [8,9]. This shortage leads to significant delays in diagnosis, especially in rural and underserved areas, where patients often face prolonged wait times for pathology reports. These diagnostic delays directly contribute to adverse patient outcomes, including delayed treatment initiation and reduced survival rates, while exacerbating health care inefficiencies [7,10]. These diagnostic delays are symptomatic of deeper structural inequities in health care access between urban and rural populations.

Telepathology has emerged as a transformative solution to decentralize diagnostic services, offering a viable strategy to overcome these systemic barriers [11-14]. By leveraging digital slide scanning technologies [15], secure data transmission protocols [16], and remote diagnostic interfaces, telepathology enables cross-regional and cross-institutional collaboration among pathologists [17,18]. Numerous studies have demonstrated that telepathology systems can effectively reduce TAT, enhance diagnostic accuracy, and optimize resource distribution, making them valuable tools for both routine diagnostics and specialized consultations [19-24]. For instance, a telepathology diagnostic network established in Tanzania showed TAT performance comparable to conventional intraoperative consultations using optical microscopes [23]. Similarly, the Clinical Telepathology Service at the University of California, Los Angeles, demonstrated improved diagnostic efficiency and cost savings through its cloud-based infrastructure [25].

However, the scalability of telepathology systems remains a critical barrier to widespread adoption. A recent systematic review revealed that the majority of telepathology initiatives are limited to single institutions or pilot programs, with relatively few achieving province- or state-wide coverage [26]. This limitation arises from multiple technical and organizational challenges, including inconsistent standards for image acquisition, storage, and sharing, which hinder interoperability across regions and institutions [26]. Additionally, the large volumes of data generated by whole slide image (WSI) systems contribute to computational bottlenecks, complicating data processing and analysis [15,16]. The absence of tiered governance models, which are essential for coordinating multilevel health care systems, further constrains the effective management and deployment of telepathology solutions [27,28]. Moreover, without equitable deployment strategies, such innovations risk widening existing gaps in diagnostic access.

To address these challenges, this study presents the design, deployment, and evaluation of a province-wide telepathology system across 120 health care institutions in Henan Province, China. Enabled by dedicated funding from the Henan Provincial Health Commission, the system facilitated the deployment of DP scanners in county- and city-level hospitals. Led by the First Affiliated Hospital of Zhengzhou University, the largest medical institution in China, the system operates under a centralized framework for case management, expert consultation, and data integration, supported by a mobile-enabled interface to expand diagnostic access to remote and rural regions.

This study aims to (1) design and implement a scalable province-wide telepathology system to enhance equitable access to diagnostic services across health care tiers; (2) evaluate its performance in terms of diagnostic efficiency, accuracy, and economic impact (direct patient cost savings); (3) demonstrate the feasibility of large-scale telepathology expansion in resource-constrained settings; and (4) offer a scalable model that can be replicated in other health care systems facing similar challenges. These findings contribute to the growing body of evidence supporting digital health solutions in bridging health care gaps and optimizing pathology services worldwide.

Methods

Overview

This Methods section describes the study design, data sources, evaluation methods, and data protection measures used to assess a province-wide telepathology system.

Study Design

This study adopted a multicenter retrospective observational design to evaluate the real-world implementation and performance of a province-wide telepathology system across 120 health care institutions in Henan, China. Routinely collected, deidentified system data generated between January 2016 and June 2024 were analyzed.

The evaluation comprised three components aligned with the study aims: (1) technical performance, assessed using Apache Benchmark stress testing to measure response time and concurrent user capacity; (2) operational use and efficiency, evaluated descriptively using system-generated metrics including consultation volume, contributing hospital level, and diagnostic TAT; and (3) diagnostic accuracy, evaluated in a retrospective

XSL•FO
RenderX

subset of 1027 cases by comparing originating-site preliminary diagnoses with tertiary-hospital expert reference diagnoses recorded in the platform.

Data Collection and Sources

All analyses used routinely collected, deidentified data automatically recorded by the telepathology platform between January 2016 and June 2024. System logs and administrative exports were used to summarize consultation volume, contributing hospital level, annual trends, and TAT, defined as the interval from WSI submission to report issuance.

For frozen-section consultations, the request-to-diagnosis time was summarized based on real-time notification and reporting timestamps recorded in the platform. Patient and diagnostic characteristics (age, sex, benign vs malignant classification, and specimen anatomical site) were obtained from system-generated pathology reports. Platform metadata and storage reports were used to describe supported WSI formats, per-slide file size range, and cumulative storage volume.

Diagnostic Accuracy Assessment

Diagnostic accuracy was evaluated in a retrospective subset of 1027 telepathology consultations that had both (1) an originating-site preliminary diagnosis recorded at submission and (2) a corresponding tertiary-hospital expert diagnosis recorded in the platform for the same case. Cases with missing or noncomparable diagnostic categories were excluded. Diagnoses were dichotomized as malignant (positive) versus non-malignant (negative) according to the recorded report impression. The tertiary-hospital expert diagnosis served as the reference standard. Sensitivity, specificity, and overall agreement were calculated using a confusion matrix.

Economic Impact Estimation

Per-case cost savings were not recalculated from primary economic data in this study. Instead, we adopted the previously published per-case savings estimate for the same provincial telepathology program (378.5 RMB per case; US $1=6.46 RMB) reported by He et al [10], which was based on local consultation fees and assumed travel and food expenses avoided when patients no longer needed to travel to tertiary hospitals for expert consultation. To derive an annual estimate, we multiplied the per-case savings (378.5 RMB; US $1=6.46 RMB) by the number of completed telepathology consultations in each calendar year (Table 1). All RMB values were converted to US dollars using the exchange rate reported in He et al [10].

Table 1. Annual volume of completed telepathology consultations by hospital level in a province-wide observational study conducted in Henan Province, China.

Hospital level	2016	2017	2018	2019	2020	2021	2022	2023	2024
City-level	240	657	1060	923	2100	3503	2970	2439	2421
County-level	2095	3673	6202	8317	7490	6536	6198	8267	7825

System Architecture Design

To address the shortage of pathologists and the uneven distribution of diagnostic resources in Henan Province, China, we designed an integrated province-wide telepathology system optimized for scalability, diagnostic accuracy, and operational efficiency. The system follows a 3-tiered hospital architecture, consisting of tertiary-level hubs (provincial hospitals), secondary-level nodes (city-level hospitals), and primary-level health care institutions (county-level hospitals). This hierarchical structure enables standardized workflows for digital slide acquisition, data transmission, remote diagnosis, and reporting across 120 health care institutions, ensuring equitable access to high-quality pathology services.

The system adopts a Browser-Server architecture, chosen for its flexibility, ease of deployment, and compatibility with diverse IT infrastructures in both urban and rural health care settings. To support real-time access in resource-constrained environments, we incorporated both web-based platforms and mobile access via a WeChat mini-program. This mobile interface was particularly critical for ensuring remote diagnostic access in rural areas where conventional IT infrastructure is limited. The mini-program was selected due to its widespread adoption in China and seamless integration with existing health care communication workflows, eliminating the need for standalone app installations. The system consists of 4 core modules: image acquisition module, transmission link module, data storage module, and teleconsultation platform module. These modules are tightly integrated to support high-throughput data processing, secure transmission, and efficient telepathology consultations (Figure 1).

XSL•FO
RenderX

Figure 1. Architecture of the province-wide telepathology system evaluated in a retrospective multicenter observational study across 120 health care institutions in Henan Province, China.

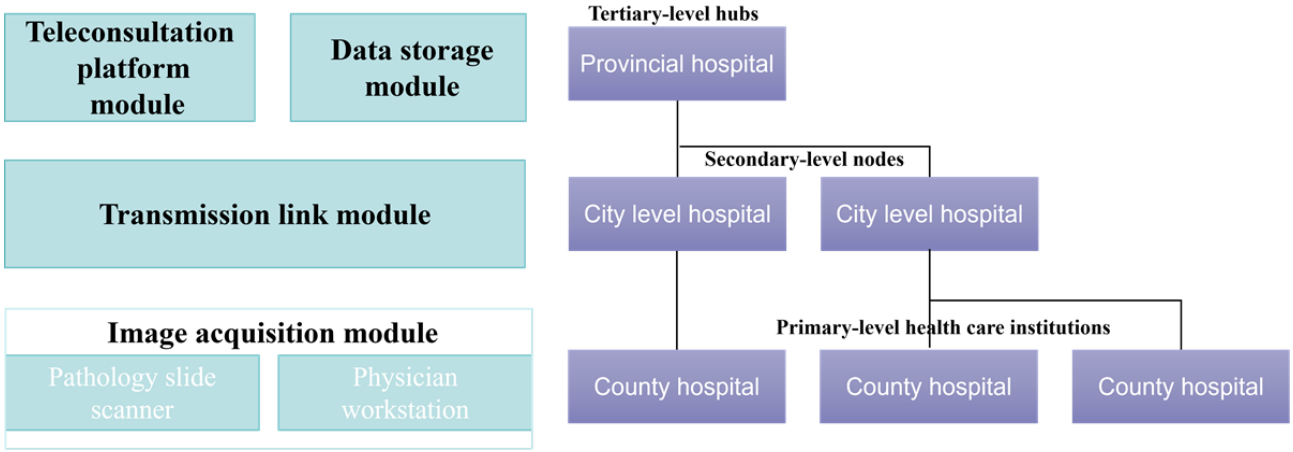

Image Acquisition Module

This module digitizes traditional pathology slides into high-resolution WSIs and prepares them for secure transmission. Each participating hospital is equipped with (1) pathology slide scanner Model KF-PRO-005 (Konfoong Biotech International Co., Ltd.), capable of scanning slides within 40 seconds at 0.47 µm per pixel (20 magnification) and within 100 seconds at 0.5 µm per pixel (40 magnification); and (2) physician workstation powered by Intel Core i5 processors with 8 GB RAM, preloaded with K-Scanner 1.6.0.14 and K-Viewer 1.5.3.1 (Konfoong Bioinformation Tech Co, Ltd., Ningbo, China) for scanner control, image browsing, and management.

The scanned WSIs are processed using a multiresolution pyramid structure, optimizing image loading speed and reducing data size without compromising diagnostic quality.

In addition, a standardized scanning workflow and a 2-step image quality control (QC) procedure were applied across participating sites to ensure that WSIs met diagnostic requirements. Before scanning, slides were checked for correct labeling and basic preparation adequacy, and scanners were routinely calibrated according to the manufacturer's recommendations. After scanning, technicians performed an on-screen QC review to confirm tissue coverage completeness, focus adequacy across the tissue area, and absence of major scanning artifacts (eg, excessive blur, stripe or banding, or large occlusions), as well as reasonable stain and color appearance. Slides failing QC were rescanned, and if persistent issues were attributed to slide preparation, the originating laboratory was notified for recutting or restaining. A second, preanalytical QC check was conducted by triage personnel during case allocation; cases with insufficient image quality were returned for resubmission prior to expert review.

Transmission Link Module

The transmission link module ensures that digitized pathology images can be transmitted from the applying hospital to higher-level hospitals. Considering data security and transmission rates, the system uses the 5G fixed-mobile convergence medical private network, previously established by the Henan Telemedicine Center of China and covering Henan Province, as the transmission link. Digitized pathology slides are scanned and uploaded to the central server for storage and diagnosis through this private network. The architecture of this private network is illustrated in Figure 2.

The telemedicine private network can be classified into wired transmission networks, 5G telemedical transmission networks, and Internet of Things transmission networks. The wired network is primarily used in fixed office scenarios and easily accessible medical institutions, such as large hospitals, for teleconsultations, telepathology, remote education, and applications in medical institutions at the county level and above. The 5G telemedical transmission network mainly supports dynamic scenarios, off-hospital scenes, and the access of medical institutions in remote areas, including remote ward rounds, remote emergency services, remote electrocardiography, remote ultrasound, and access for village and town health clinics. The construction and application of this private network provide effective network support for the transmission of remote pathological slides.

XSL•FO
RenderX

Figure 2. Architecture of the medical private network supporting province-wide telepathology consultations in Henan Province, China.

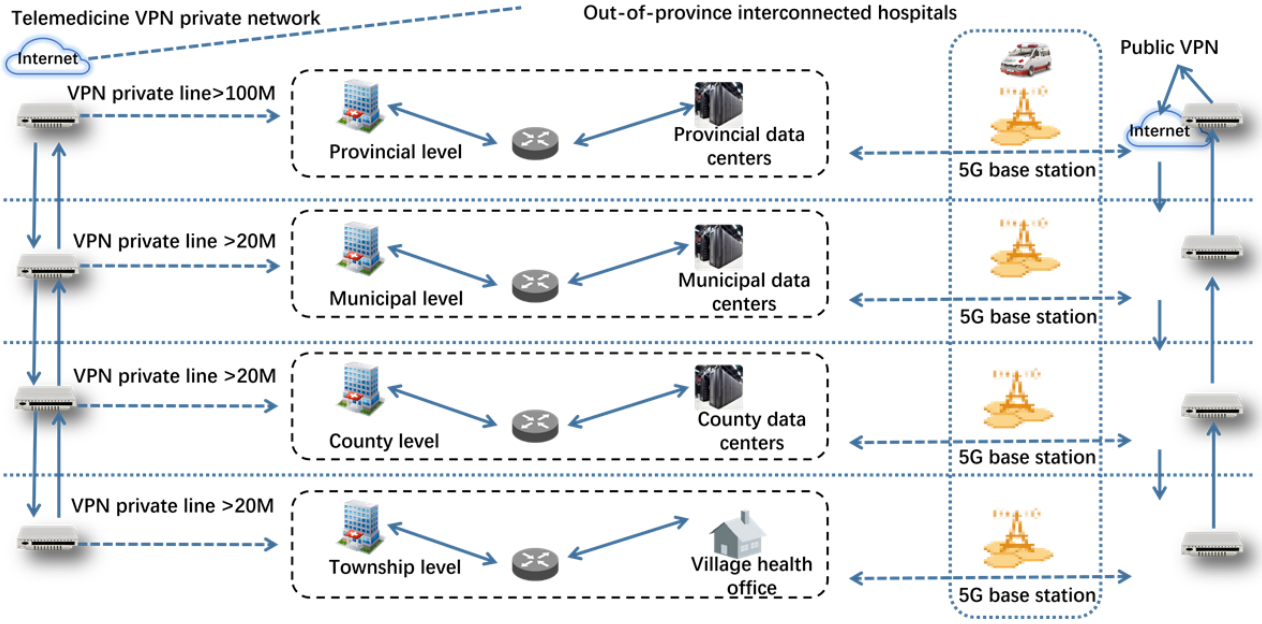

Data Storage Module

The data storage module is designed to ensure the secure, scalable, and efficient storage of WSIs, diagnostic reports, and patient records. To enhance storage efficiency and fault tolerance, we use erasure coding technology, which works by dividing data into multiple fragments and generating parity blocks distributed across different storage nodes. In the event of data corruption or hardware failure, erasure coding enables data reconstruction from the remaining fragments, ensuring high availability and data integrity with minimal storage overhead [29].

For data backup, the system adopts a master-slave data synchronization strategy, implementing asynchronous replication to ensure the consistency and integrity of data. Additionally, binary log (Binlog) technology is used to record all database changes, allowing for rapid data recovery in the event of a system failure. The backup mechanism uses a mixed-based replication model, combining the advantages of statement-level and row-based replication, ensuring efficient and stable data backups.

Teleconsultation Platform Module

Overview

The teleconsultation platform is the core of the system, supporting remote pathology consultations, diagnostic reporting, and interinstitutional case discussions. The software architecture (Figure 3) is organized into 4 layers.

XSL•FO
RenderX

Figure 3. Software architecture of the teleconsultation platform used for province-wide telepathology services in Henan Province, China. AI: artificial intelligence; MQ: message queue; VPN: virtual private network.

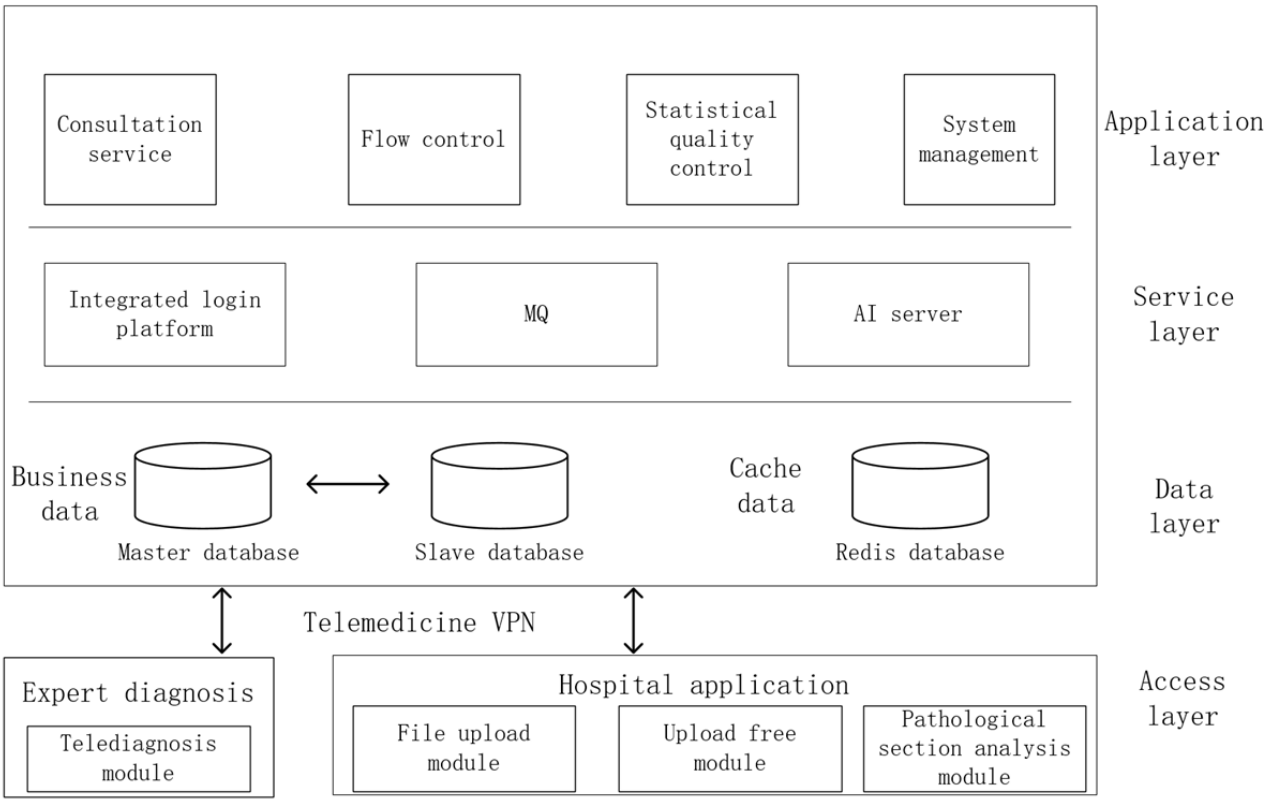

Access Layer

The access layer connects requesting hospitals with expert terminals, enabling seamless interaction and workflow integration. Within this layer, the upload free module eliminates the need for manual file uploads by integrating directly with local DP scanners, ensuring automatic and efficient transmission of WSIs. Additionally, the Pathological Section Analysis Module performs preliminary analysis, including image quality assessment, segmentation, and lesion detection, aiding triage personnel in prioritizing cases. To optimize performance, WSI images undergo segmentation, compression, and pyramid layer processing, allowing for rapid loading and smooth scaling of high-definition images. Experts can access, analyze, and report cases either through a web browser or via the WeChat mini-program, enhancing flexibility and accessibility.

Data Layer

The data layer is responsible for managing both business-related data and cache data to ensure system stability and efficiency. Business data is stored within a MySQL relational database, maintaining structured and secure data management, while cache data, such as user authentication tokens and system configuration information, is stored using Redis, an in-memory database designed to facilitate high-speed data access and enhance system performance.

Service Layer

The service layer provides the system's core functionalities, encompassing consultation process management, diagnostic reporting, and QC statistics. To ensure seamless communication between different modules, message queue technology is implemented, allowing for asynchronous data transmission. This enables efficient handling of high-volume concurrent requests, ensuring that teleconsultation processes remain reliable and scalable without system bottlenecks.

Application Layer

The application layer provides various application functionalities and permissions tailored for triage personnel, system administrators, primary hospital doctors, and others, catering to their specific roles and access levels.

Data protection and privacy. To support province-wide deployment and mobile access, the platform implemented the following safeguards. The WeChat mini-program served only as a mobile access interface. WSIs and clinical records remained in the telemedicine center’s private cloud and were accessed through the secured telepathology platform via a dedicated application programming interface gateway (HTTPS with Transport Layer Security encryption, authentication, role-based access control, and an allowlist of permitted application programming interfaces). Access events were logged for audit and monitored with rate limiting when needed. The WeChat-based support chat group was limited to technical troubleshooting and operational coordination; no patient-identifiable information or diagnostic images were shared, and all clinical case review and reporting were performed within the secured telepathology platform.

XSL•FO
RenderX

System Functionality Overview

Overview

During the functional design phase of the system, a multidisciplinary design discussion group was established to ensure that the system effectively met the needs of different stakeholders. This group consisted of 6 members, including 1 pathologist from the First Affiliated Hospital of Zhengzhou University, representing the requirements of tertiary hospitals, 2 pathologists from primary health care institutions who articulated user needs from a clinical perspective, 1 project manager overseeing system development, and 2 software developers responsible for technical implementation. To validate the system's diagnostic reliability, a clinical study was conducted in which pathologists compared WSI-based diagnoses with conventional glass slide diagnoses across various case types. Expert feedback was collected on image interpretability, usability, and diagnostic confidence using both the browser-based platform and the WeChat mini-program. The results demonstrated a high concordance between digital and conventional diagnoses, confirming the system's suitability for clinical application.

Based on functional requirements, users were categorized into 4 distinct roles within the telepathology framework, ensuring an optimized workflow that facilitates seamless collaboration between requesting personnel, triage personnel, diagnostic experts, and system administrators (Figure 4).

Figure 4. Use case diagram illustrating roles and interactions within the province-wide telepathology system in Henan Province, China.

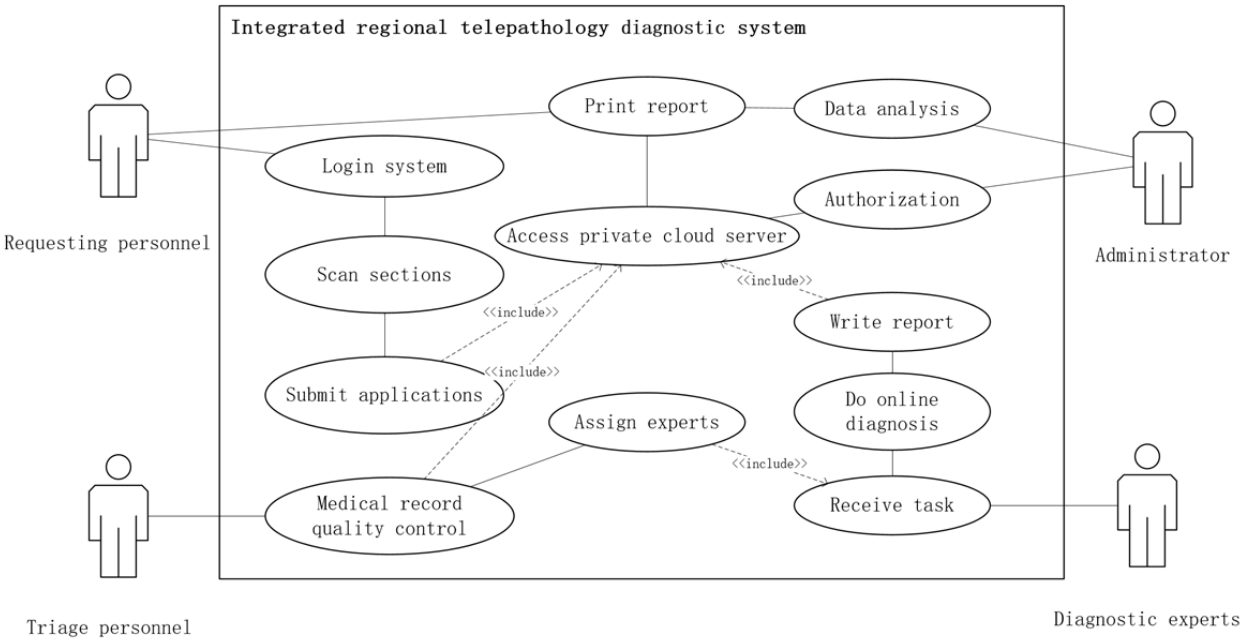

Requesting Personnel

Primarily from county-level (primary) and city-level (secondary) hospitals, initiate telepathology processes by digitizing and scanning pathology slides, completing patient medical records, and submitting telepathology requests to higher-level institutions. This decentralized design enabled even low-resource county hospitals to initiate telepathology consultations independently, contributing to more equitable diagnostic access.

Triage Personnel

As pathologists within city-level (secondary) and provincial-level (tertiary) hospitals, they are responsible for preanalytical QC and case allocation. By leveraging their in-depth knowledge of their hospital's pathology experts and their areas of specialization, they ensure that each case is assigned to the most suitable specialist within the institution.

Diagnostic Experts

Consisting of pathologists within city-level and provincial-level hospitals, analyze the assigned digital slides and generate diagnostic reports, either providing a direct diagnosis or escalating particularly complex cases to a higher-level institution when necessary.

Administrators

They oversee the integrated telepathology system, managing user roles and permissions, ensuring data security, and maintaining system functionality to guarantee a seamless telepathology workflow.

Operational Workflow and User Interaction

Overview

The integrated telepathology system is designed to streamline the end-to-end diagnostic workflow, encompassing 3 primary functionalities: application initiation, triage, and diagnostic review. These functionalities are supported by an intuitive user interface that facilitates efficient interaction between primary health care providers, triage coordinators, and diagnostic experts. This section provides a detailed description of the system’s operational flow, user roles, and key technical features.

XSL•FO
RenderX

Application Initiation Workflow

The initiation of a consultation application is a core function performed by requesting personnel at primary health care institutions. This process involves several key steps: logging into the system, scanning pathology slides, and completing the consultation application form. The output of this function is the transmission of consultation application data via the virtual private network (VPN) to the telemedical private cloud server at the Henan Telemedicine Center of China. This information is then made accessible to triage personnel and diagnostic experts at tertiary medical institutions for review and further processing. The use case diagram for this function is illustrated in Figure 5.

Figure 5. Workflow for telepathology consultation request submission in a province-wide telepathology network in Henan Province, China.

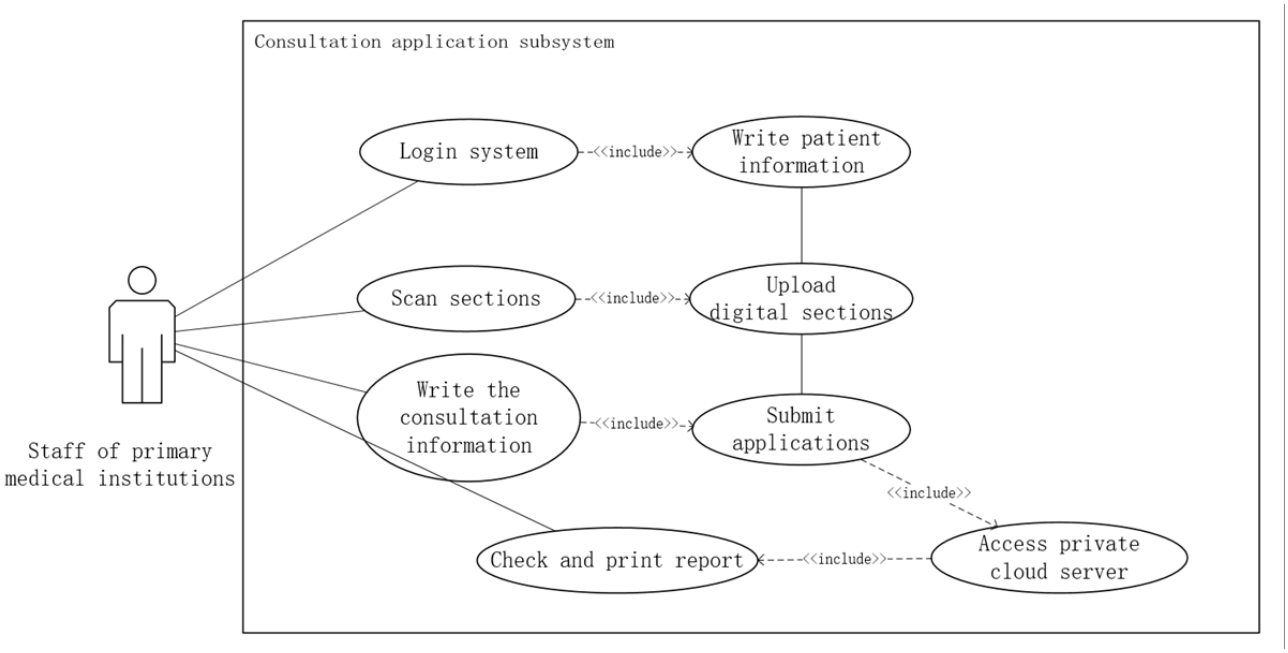

Triage Workflow and QC

The triage function is performed by pathologists serving as triage personnel within tertiary medical institutions, playing a pivotal role in ensuring the integrity of the diagnostic process. This function involves conducting preanalytical QC on received telepathology applications, which includes the following:

1. Assessing the adequacy of scanned WSIs for diagnostic interpretation, including sufficient tissue coverage, effective resolution, and focus across the tissue area, and absence of major artifacts (eg, excessive blur, stripe or banding, or large occlusions) that may compromise interpretation; staining or color issues were assessed in terms of image readability.
2. Verifying the completeness of clinical information, ensuring that relevant patient history, laboratory findings, and any necessary supporting documentation are provided to facilitate accurate diagnosis.

If the WSI quality is insufficient for diagnostic interpretation or the clinical information is incomplete, triage personnel may request additional data or resubmission (eg, rescanning, recutting, or restaining when preparation-related issues are suspected) before assigning the case to the most appropriate diagnostic expert based on the specific pathology subspecialty required. The workflow diagram illustrating this process is depicted in Figure 6.

When resubmission was required, the case was returned within the platform with standardized reasons, and the requesting site was notified through in-platform status updates and SMS alerts to registered staff; unresolved returns were escalated to the site administrator for follow-up.

XSL•FO
RenderX

Figure 6. Workflow for case triage and preanalytical quality control in the province-wide telepathology system in Henan Province, China. QC: quality control.

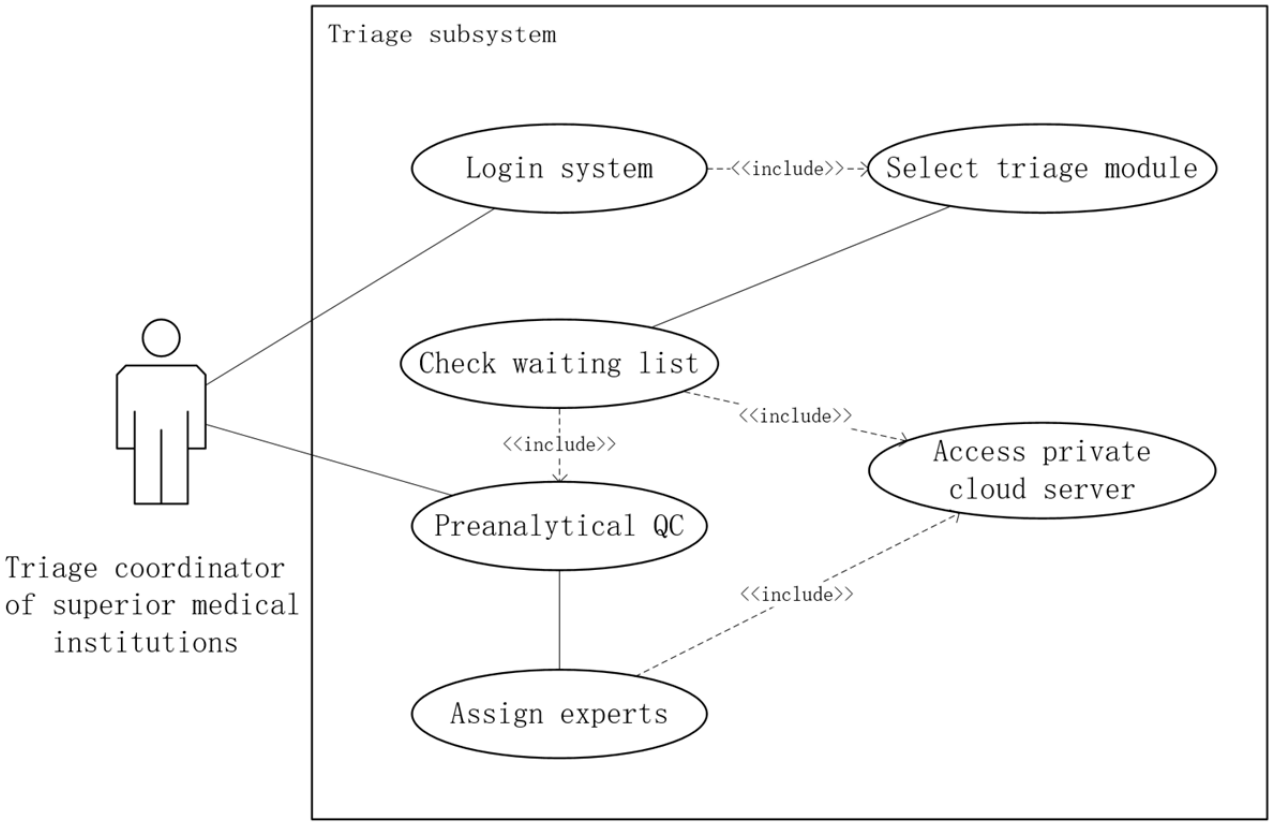

Diagnostic Function Design

The diagnostic function is executed by experts at tertiary hospitals, specifically pathologists, who access the system via a computer interface or the WeChat mini-program. Upon logging in, they review the assigned telepathology cases, analyzing DP slides and accompanying clinical data to provide a comprehensive pathological diagnosis. The diagnostic report includes a pathological description, grading, staging, results of immunohistochemistry, and additional relevant findings (Figure 7).

When an expert at a tertiary hospital determines that additional immunohistochemistry staining is required, they can submit an immunohistochemistry request through the system, specifying the recommended immunohistochemistry markers for further testing. This request is automatically recorded in the system and made accessible to the originating hospital. The originating hospital performs the immunohistochemistry staining locally, digitizes the stained slides, and uploads them back to the original case record within the platform. The tertiary hospital expert then reviews the newly uploaded immunohistochemistry-stained slides, incorporating the findings into a supplementary diagnostic report to ensure a comprehensive pathology assessment. The diagnostic report is completed using a semistructured interface, where predefined sections (eg, “Microscopic Findings,” “Diagnosis,” and “Comments”) guide report writing, while allowing free-text input to maintain clinical flexibility.

XSL•FO
RenderX

Figure 7. Workflow for remote pathology diagnosis and reporting within the province-wide telepathology system in Henan Province, China.

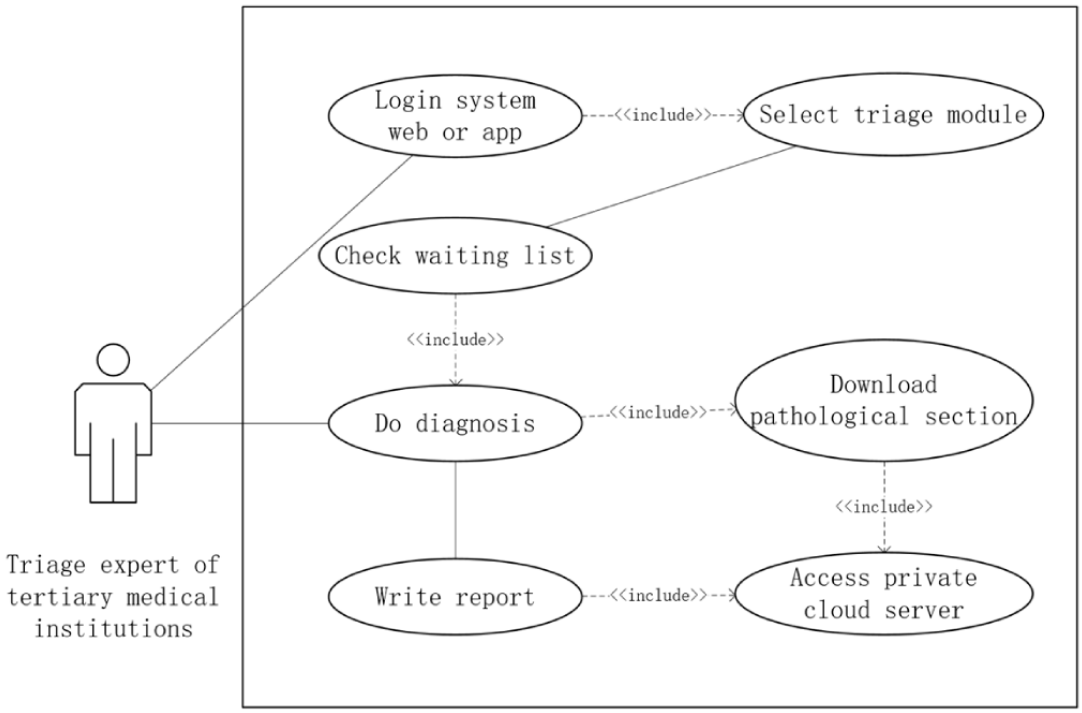

Ethical Considerations

This retrospective multicenter study was approved by the Ethics Committee of the First Affiliated Hospital of Zhengzhou University (approval number 2025-KY-1532-002). The requirement for informed consent was waived because the study involved secondary analysis of deidentified routine telepathology data without direct participant contact.

All data were anonymized prior to analysis, and no personally identifiable information was accessible to the research team. Data were stored on secure institutional servers with restricted access in accordance with applicable data protection regulations. No compensation was provided, as this study involved retrospective analysis of existing data. All procedures adhered to institutional policies and the principles of the Declaration of Helsinki.

Results

System Deployment and Operational Integration

The telepathology system was successfully deployed within the Henan Telemedicine Center's private cloud infrastructure, leveraging a secure VPN to facilitate encrypted data transmission across 120 integrated health care institutions at the provincial, municipal, and county levels. This architecture ensures robust data security, transmission stability, and interoperability, which are critical for maintaining diagnostic accuracy during remote consultations.

The deployment of DP scanners followed a stepwise approach to ensure systematic integration, validation, and training across hospitals. In 2015, the Henan Provincial Health Commission initiated a province-wide infrastructure assessment to evaluate hospital readiness for telepathology implementation. The procurement and phased installation of 120 scanners began in 2016, prioritizing tertiary hospitals first, followed by city and county-level hospitals, based on workload demand and digital infrastructure readiness. Pathologists and technicians were trained in scanner operation, QC, and troubleshooting. By 2017, scanners were fully integrated into routine workflows, with ongoing monitoring and performance optimization to maintain consistency in image acquisition.

To accommodate various clinical scenarios, the system supports both web-based interfaces and mobile access via a WeChat mini-program. The mobile integration significantly enhances accessibility, particularly in rural and resource-limited regions, where mobile technology adoption is high. The user interface was optimized for both platforms to streamline diagnostic workflows, reduce user friction, and improve the TAT for pathology consultations (Figure 8).

Since deployment, the system has facilitated 72,916 telepathology consultations, averaging approximately 22 cases per day. A total of 355,104 WSIs have been processed. The cumulative data volume has reached 220.1 TB, with each slide averaging 0.62 GB. However, slide storage size varies depending on scanning resolution, typically ranging from 500 MB to 2 GB, with 20× scans occupying 500 MB to 1 GB and 40× scans ranging from 1 GB to 2 GB. The system supports multiple whole-slide image formats, including Tagged Image File Format (TIFF) and NDPI (NanoZoomer format), as well as other vendor-specific formats that use pyramidal storage for efficient multiresolution viewing. Unlike Digital Imaging and Communications in Medicine (DICOM), which is primarily used for radiology, these formats support multilayered, ultra-high-resolution whole-slide imaging, ensuring optimal fidelity for pathological assessment.

The system is hosted on Henan Telemedicine Center’s private cloud, using a distributed storage system for high availability,

XSL•FO
RenderX

redundancy, and secure access. Diagnosed slides are archived rather than deleted, ensuring long-term availability for clinical review, QC, and research. To optimize storage, frequently accessed images remain in high-performance storage, while older slides are moved to cost-efficient archival storage, maintaining scalability as telepathology adoption grows.

Figure 8. User interfaces of the web-based platform and WeChat mini-program used for telepathology consultations in Henan Province, China.

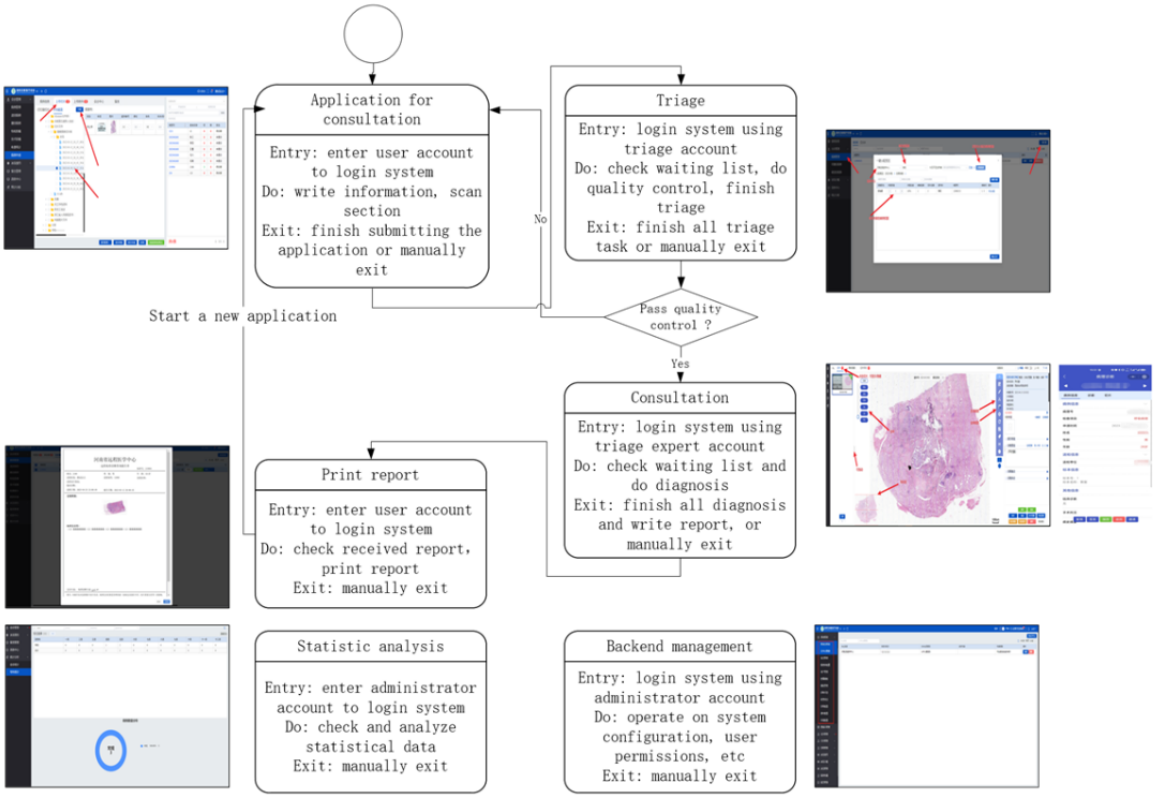

Diagnostic Process Implementation

Since deployment, the telepathology system has streamlined diagnostic workflows across 120 health care institutions, enabling seamless collaboration between county, city, and provincial-level hospitals. The system has facilitated the efficient submission, triage, and assignment of pathology cases, ensuring that cases requiring expert consultation are promptly routed.

The WeChat mini-program has been widely adopted to facilitate remote access and case review, allowing pathologists to view WSIs and provide feedback even outside traditional workstations. However, medical-grade monitors remain the preferred choice for primary diagnosis, while the mobile platform serves as a supplementary tool for preliminary assessments and report verification.

Health care professionals receive ongoing training and technical support to ensure effective system use. Annual training conferences provide updated guidance and case discussions. A comprehensive operations manual is available for reference, covering image acquisition, QC, and reporting workflows.

To optimize medical resource distribution, the system uses a tiered health care architecture, incorporating provincial, municipal, and county-level medical institutions. County-level institutions may escalate complex cases directly to provincial-level institutions or route them through municipal-level institutions. Municipal hospitals act both as requestors and providers, while provincial hospitals primarily manage complex case diagnostics.

System Performance Evaluation

To evaluate the system's technical performance, a stress test was conducted using the Apache Benchmark Tool. The detailed performance metrics are presented in Table 2. Simulated environments with 100 concurrent users performing 100 requests each were used to assess system responsiveness, scalability, and load-bearing capacity under both typical and peak usage conditions.

XSL•FO
RenderX

Table 2. Technical performance metrics of the province-wide telepathology platform under concurrent user load, evaluated using stress testing.

Requests	Response times (ms)					Throughput (transactions/s)	Network (KB/sec)	
Label	Min	Max	90th percentile	95th percentile	99th percentile		Received	sent
Consultation request interface	2568	7619	7013.10	7144.85	7497.85	25.78	114.72	48.81
Application list view page	190	3237	2859.50	2988.70	3206.30	58.88	646.77	30.24
Pending diagnosis list page	200	3017	2838.30	2891.75	3001.79	59.00	648.10	30.59
Diagnosis page	423	4359	3793.00	4031.70	4353.06	43.87	15.72	50.08

The data derived from the system performance evaluation reveals significant capabilities in handling concurrent operations across various interfaces:

1. Consultation request interface: the system can respond to 25.78 requests per second, which supports approximately 125 users operating online simultaneously.
2. Application list view page: the system is capable of handling 58.88 requests per second, allowing for about 300 users to be active online at the same time.
3. Pending diagnosis list page: the system's response rate is 59 requests per second, indicating capacity for approximately 300 users to operate online simultaneously.
4. Diagnosis page: the system manages 43.87 requests per second, suggesting it can accommodate an estimated 220 users online simultaneously.

The results demonstrate that the telepathology system can maintain high performance even under substantial user loads. The integration of a dedicated VPN, coupled with the optimized Browser-Server architecture, ensures that diagnostic workflows are not compromised during peak operational periods. These capabilities are critical for large-scale deployments in health care systems with fluctuating demands.

Real-World Impact and Clinical Outcomes

System Use and Expansion

The scanners deployed in hospitals were used for both routine in-house pathology workflows and telepathology consultations, but only selected cases were uploaded for remote expert review. The telepathology system was not intended to replace all local diagnostic workflows; rather, it served as a platform for second-opinion consultations and specialized expert input when necessary. Cases uploaded for telepathology consultations primarily included diagnostically complex cases requiring additional expert interpretation, subspecialist consultation requests for specific pathology types, and frozen section cases where rapid intraoperative decision-making was required. This selective approach ensured efficient use of DP resources while allowing hospitals to retain routine diagnostics locally. By integrating telepathology into the existing diagnostic framework, hospitals could leverage the expertise of higher-level institutions without overloading the digital infrastructure with routine cases that did not require external review.

Since its implementation, the telepathology system has been integrated into 120 health care institutions across Henan Province. From 2016 to 2024, a total of 72,916 diagnostic cases were completed, with municipal hospitals contributing 16,313 (22.37%) cases and county-level hospitals accounting for 56,603 (77.63%) cases. Provincial hospitals primarily participated as expert consultation centers and triage hubs; although technically able to submit cases, they rarely acted as requestors in routine practice. This hospital-level distribution was used as an indirect indicator of access to expert pathology support for lower-tier institutions. In China's tiered health care system, county-level hospitals typically have more limited access to subspecialty pathology expertise than higher-level centers. Therefore, the predominance of county-level submissions indicates that the platform was primarily used by lower-tier institutions and facilitated routine access to tertiary-hospital expertise for remote sites.

Analysis of annual service volumes revealed a significant upward trend. Municipal hospitals experienced a growth from 240 cases in 2016 to 3503 cases in 2021, representing a 1360% increase. Similarly, county-level hospitals saw an increase from 2095 cases in 2016 to 8317 cases in 2019, reflecting a 297% growth. Despite fluctuations in service volumes in later years, the overall trend indicates sustained adoption and use across health care tiers (Table 1).

These findings suggest that the telepathology system has effectively enhanced diagnostic capacity, particularly in county-level institutions where pathology services were previously limited.

Diagnostic Efficiency and TAT

The system's efficiency was assessed based on TAT, defined as the interval from whole-slide image submission to diagnostic report issuance. The median TAT was 10.06 hours (IQR 1.63-29.10), with 54.28% (39,579/72,916) diagnosed within 12 hours and 96.41% (70,298/72,916) completed within 72 hours.

For frozen section consultations, where rapid diagnostic turnaround is critical, the system facilitated real-time notifications to pathologists upon case submission. From 2016 to 2024, a total of 508 frozen section cases were processed through the telepathology system, with an average time from request to diagnosis of approximately 16 minutes and 6 seconds. This highlights the system's ability to support time-sensitive

XSL•FO
RenderX

intraoperative pathology decisions and enhance surgical workflow efficiency.

Patient Demographics and Diagnostic Distribution

A total of 72,916 patient cases were analyzed. Patient age ranged from 2 to 91 years (mean 51.8, SD 16.9, median 52.6, IQR 45-67 years). Age-group distribution was as follows: 18 years, n=2187 (3%); 19-40 years, n=14,583 (20%); 41-60 years, n=38,652 (53.01%); and >60 years, n=17,494 (23.99%). Overall, 43.58% (n=31,777) of cases were male, and 56.42% (n=41,139) were female. Regarding diagnostic outcomes, 48.80% (35,583 cases) were benign, while 51.20% (37,333 cases) were malignant. Compared with a baseline dataset from 2020 [10], there was an observed increase in the proportion of male patients by 2.49% (from 41.09%) and a decrease in benign lesion diagnoses by 3.38% (from 52.18%).

The most frequently reported anatomical sites were uterine, gastrointestinal, and esophageal tissues. Other common specimen categories included breast, lung, mediastinum, and male genital organs. Additional sites included thyroid, head and neck, and respiratory tract. Although biopsy specimens were included, cases were categorized by anatomical site rather than procedural type (biopsy or resection), and detailed specimen-type classification was not available in the current platform; future updates may incorporate specimen-type fields to improve data granularity.

Economic Impact

Before the implementation of telepathology services, patients in county hospitals seeking expert pathology consultation typically had to travel to tertiary hospitals with their glass slides and medical records, incurring transportation costs, lost wages, and long wait times. Alternatively, some hospitals mailed slides for consultation, but this method was slow and inefficient, often leading to delayed diagnoses.

With telepathology, biopsy slides are now digitized locally and reviewed remotely by experts at tertiary institutions, eliminating the need for physical transport. Using the previously published per-case savings estimate of 378.5 RMB (approximately US $50) [10], and applying it to the annual consultation volumes observed in our network (Table 1), the estimated direct cost savings were approximately US $0.14 to $0.63 million per year across 2016-2024.

In the retrospective subset of 1027 consultations with paired diagnoses, the tertiary-hospital expert diagnosis recorded in the platform was treated as the reference standard (malignant vs nonmalignant). Originating-site preliminary diagnoses showed a sensitivity of 0.90 (481/535) and a specificity of 0.75 (369/492), with an overall agreement of 0.83 (850/1027). Overall, 177 (17.2%) cases were discordant, including 54 false negatives and 123 false positives (Table 3).

Table 3. Confusion matrix comparing originating-site preliminary diagnoses with tertiary-hospital expert reference diagnoses in telepathology consultations.

Originating-site preliminary diagnosis	Expert reference malignant (n=535), n	Expert reference nonmalignant (n=492), n
Originating-site malignant	481[a]	123[b]
Originating-site nonmalignant	54[c]	369[d]

[a]TP: true positive.

[b]FP: false positive.

[c]FN: false negative.

[d]TN: true negative.

Discussion

Principal Results

This study presents the design, deployment, and retrospective evaluation of a province-wide telepathology system implemented across multiple health care institutions in Henan, China. In line with the study aims, the system demonstrated (1) robust technical performance under concurrent load, (2) sustained real-world use with timely reporting, and (3) improved diagnostic accuracy in a retrospective subset compared with originating-site preliminary diagnoses. Overall, these findings suggest that a centrally coordinated telepathology network, supported by secure transmission and optional mobile access, can improve diagnostic access, operational efficiency, and quality assurance at scale.

Interpretation and Implications

A key finding of this study was the high level of engagement from county-level hospitals. This pattern of use suggests that the telepathology network effectively expanded access to specialist pathology expertise for lower-tier and predominantly rural institutions, where in-house subspecialty support is often limited within China's tiered health care system. By lowering geographic and institutional barriers, the system enabled routine consultation with tertiary-hospital experts and may help reduce long-standing disparities in diagnostic capacity.

From an operational perspective, the system supported timely reporting for routine consultations and rapid turnaround for time-sensitive frozen-section cases, indicating that telepathology can be integrated into both routine and intraoperative workflows without compromising clinical responsiveness. The availability of mobile access via a WeChat mini-program further enhanced flexibility for case review and report verification, particularly for clinicians working outside traditional workstations, while maintaining compatibility with established diagnostic practices.

Expert review conducted through the telepathology network corrected a meaningful proportion of preliminary diagnoses originating from lower-tier institutions. This finding underscores the role of telepathology not only in expanding access but also

XSL•FO
RenderX

in strengthening diagnostic quality assurance by enabling second opinions and subspecialty input for complex or ambiguous cases.

In addition to clinical benefits, the system demonstrated potential economic value by reducing the need for patients to travel to tertiary hospitals for expert consultation. These cost-related benefits may be particularly important for rural populations, for whom travel expenses and delays can pose substantial barriers to timely diagnosis and care.

Comparison With Prior Work

Prior studies have shown that telepathology can improve access to subspecialty expertise, shorten diagnostic turnaround, and support pathology services in resource-limited or geographically dispersed settings [19,23,24]. However, most reported implementations are confined to single institutions, pilot programs, or short-term projects. As a result, their impact on system-level diagnostic inequities remains limited [26].

In this context, this study differs in both scale and duration. We report a province-wide telepathology network involving 120 health care institutions across multiple levels of care, sustained over nearly a decade. Previous regional or academic center–based networks, including the California Telepathology Service, have demonstrated feasibility and operational benefits but were implemented within more restricted organizational or geographic boundaries [25,30]. By integrating telepathology into a tiered provincial health care system, our network primarily served county and city-level hospitals while using tertiary centers as expert hubs, a service and governance model that has been less frequently described.

Earlier work has emphasized that successful telepathology implementation relies on standardized image acquisition, secure data transmission, reliable storage, and structured quality assurance workflows [20,31,32]. The system evaluated in this study follows these principles through province-wide scanner deployment, centralized triage, and private-network transmission. At the same time, it addresses practical challenges related to large-scale WSIs storage and concurrent access, which have been identified as key barriers to broader adoption [26,28].

Remote access has become increasingly important for pathology workflows, particularly during periods of workforce shortage or limited on-site availability [33,34]. Most prior studies have focused on web-based or remote desktop solutions for diagnostic review. In our network, the WeChat mini-program was introduced as a complementary tool, supporting case monitoring, preliminary review, and report verification.

Regarding diagnostic quality, previous validation and consultation studies have reported that telepathology-supported second opinions can improve diagnostic confidence and help resolve challenging cases [19,25,34,35]. Our findings are consistent with this literature and indicate that expert review through the telepathology network corrected a substantial proportion of preliminary classifications. This supports the role of telepathology as a quality assurance and capacity building tool, rather than a replacement for local diagnostic workflows.

Economic analyses of telepathology have mainly focused on reductions in patient travel, slide transport, and associated delays [10]. The estimated cost savings in this study, derived by applying a previously published per-case savings estimate to observed consultation volumes, are consistent with prior reports and suggest that large-scale telepathology deployment may yield meaningful aggregate benefits for patients and health care systems. Nonetheless, as highlighted in prior studies, comprehensive cost-effectiveness evaluations that incorporate long-term outcomes and system-level resource use remain scarce and should be addressed in future research [4,26].

Limitations

This study has several limitations. First, use and workflow metrics were derived from routinely collected platform data and reflect cases submitted for telepathology consultation rather than all pathology cases processed locally, which may limit representativeness. Second, diagnostic accuracy was assessed retrospectively in a subset of cases using expert diagnoses as the reference standard, and therefore does not capture interobserver variability or replace prospective multireader validation. Third, the economic impact was estimated rather than assessed through a formal cost-effectiveness analysis that incorporates long-term outcomes and system-level resource use. Specifically, we adopted a published per-case savings estimate for the same program [10] and extrapolated annual savings using observed consultation volumes, which may not capture variation in travel distance, local costs, or indirect costs. Finally, the study was conducted in a single province, and generalizability may vary in regions with different infrastructure and workforce distribution.

Clinical Prospects

This province-wide implementation provides a practical blueprint for scaling telepathology in resource-limited settings by combining standardized workflows, secure transmission, and multiterminal access. Future work should include prospective validation across broader case mixes, formal economic evaluation, and continued investment in training and quality assurance to sustain diagnostic reliability across health care tiers. Incorporating AI-assisted decision support and interoperability standards may further enhance efficiency and consistency as adoption expands [36].

Conclusions

A province-wide telepathology network can improve access to expert pathology services, support timely reporting, and enhance diagnostic performance in real-world practice. The high use of county-level hospitals suggests potential to reduce diagnostic inequity by extending specialist capacity to lower-tier institutions. This model may inform regional pathology capacity building and scalable telepathology deployment in other settings.

XSL•FO
RenderX

Acknowledgments

The authors declare the use of generative artificial intelligence (GenAI) during manuscript preparation. According to the GAIDeT taxonomy (2025), the following tasks were supported by a GenAI tool under full human supervision: English language editing and phrasing optimization, translation, and reformatting. The tool used was ChatGPT (OpenAI; accessed on 2025-03-29). The authors reviewed and verified all AI-assisted edits and are fully responsible for the final manuscript. No GenAI was used for data analysis, statistical modeling, or result generation.

Funding

This study was supported by the Henan Provincial Natural Science Foundation for Young Scholars (No. 252300423337), the Henan Provincial Key Medical Science and Technology Project (Jointly Constructed by the Ministry of Science and Technology and Henan Province; No. SBGJ202402097), and the National Key R&D Program of China (Grant No. 2022ZD0160704). The funders had no role in the study design, data collection, data analysis, interpretation of results, or manuscript preparation.

Data Availability

The study data are not publicly available because of legal, ethical, and privacy restrictions related to provincial health information regulations. The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

JS led the conceptualization of the study, with MY and DS providing supporting input. JS, MY, and DS were responsible for the study methodology and investigation. JS and DS conducted data curation and formal analysis. XH, YL, and LW contributed to system implementation and resources. HC, WW, and JZ contributed to data interpretation, manuscript review, and editing. JS drafted the original manuscript. FC supervised the study, managed project administration, and led funding acquisition.

Conflicts of Interest

None declared.

References

1. Bu H, Li Y. Pathology. Beijing: People's Medical Publishing House; Jul 01, 2018.
2. Hanna MG, Ardon O. Digital pathology systems enabling quality patient care. Genes Chromosomes Cancer 2023;62(11):685-697. [doi: 10.1002/gcc.23192] [Medline: 37458325]
3. Eloy C, Vale J, Curado M, Polónia A, Campelos S, Caramelo A, et al. Digital pathology workflow implementation at IPATIMUP. Diagnostics (Basel) 2021;11(11):2111 [FREE Full text] [doi: 10.3390/diagnostics11112111] [Medline: 34829458]
4. Ho J, Ahlers SM, Stratman C, Aridor O, Pantanowitz L, Fine JL, et al. Can digital pathology result in cost savings? A financial projection for digital pathology implementation at a large integrated health care organization. J Pathol Inform 2014;5(1):33 [FREE Full text] [doi: 10.4103/2153-3539.139714] [Medline: 25250191]
5. Schüffler PJ, Geneslaw L, Yarlagadda DVK, Hanna MG, Samboy J, Stamelos E, et al. Integrated digital pathology at scale: a solution for clinical diagnostics and cancer research at a large academic medical center. J Am Med Inform Assoc 2021;28(9):1874-1884 [FREE Full text] [doi: 10.1093/jamia/ocab085] [Medline: 34260720]
6. Montezuma D, Monteiro A, Fraga J, Ribeiro L, Gonçalves S, Tavares A, et al. Digital pathology implementation in private practice: specific challenges and opportunities. Diagnostics (Basel) 2022;12(2):529 [FREE Full text] [doi: 10.3390/diagnostics12020529] [Medline: 35204617]
7. Zhang J, Zhang S. Cancer Report in Henan. Henan, Zhengzhou: Henan Science and Technology Press; Apr 19, 2023.
8. Andre K. The importance of health information on the internet: how it saved my life and how it can save yours. J Med Internet Res 2019;21(10):e16690 [FREE Full text] [doi: 10.2196/16690] [Medline: 31661084]
9. National Health Commission. China Health Statistics Yearbook. Beijing: China Union Medical University Press; May 17, 2023.
10. He X, Wang L, Wang L, Gao J, Cui F, Ma Q, et al. The effectiveness of a cloud-based telepathology system in China: large-sample observational study. J Med Internet Res 2021;23(7):e23799 [FREE Full text] [doi: 10.2196/23799] [Medline: 34326037]
11. Rodriguez JA, Charles JP, Bates DW, Lyles C, Southworth B, Samal L. Digital healthcare equity in primary care: implementing an integrated digital health navigator. J Am Med Inform Assoc 2023;30(5):965-970 [FREE Full text] [doi: 10.1093/jamia/ocad015] [Medline: 36795062]

XSL•FO
RenderX

12. Verma N, Lehmann H, Alam AA, Yazdi Y, Acharya S. Development of a digital assistant to support teleconsultations between remote physicians and frontline health workers in India: user-centered design approach. JMIR Hum Factors 2023;10:e25361 [FREE Full text] [doi: 10.2196/25361] [Medline: 36729578]
13. Weinstein RS, Holcomb MJ, Krupinski EA. Invention and early history of telepathology (1985-2000). J Pathol Inform 2019;10(1):1 [FREE Full text] [doi: 10.4103/jpi.jpi_71_18] [Medline: 30783545]
14. Orah N, Rotimi O. Telepathology in low resource African settings. Front Public Health 2019;7:264 [FREE Full text] [doi: 10.3389/fpubh.2019.00264] [Medline: 31572705]
15. Brunyé TT, Booth K, Hendel D, Kerr KF, Shucard H, Weaver DL, et al. Machine learning classification of diagnostic accuracy in pathologists interpreting breast biopsies. J Am Med Inform Assoc 2024;31(3):552-562 [FREE Full text] [doi: 10.1093/jamia/ocad232] [Medline: 38031453]
16. Kothari S, Phan JH, Stokes TH, Wang MD. Pathology imaging informatics for quantitative analysis of whole-slide images. J Am Med Inform Assoc 2013;20(6):1099-1108 [FREE Full text] [doi: 10.1136/amiajnl-2012-001540] [Medline: 23959844]
17. Kaplan KJ, Burgess JR, Sandberg GD, Myers CP, Bigott TR, Greenspan RB. Use of robotic telepathology for frozen-section diagnosis: a retrospective trial of a telepathology system for intraoperative consultation. Mod Pathol 2002;15(11):1197-1204. [doi: 10.1097/01.mp.0000033928.11585.42] [Medline: 12429799]
18. Ahuja S, Zaheer S. Advancements in pathology: digital transformation, precision medicine, and beyond. J Pathol Inform 2025;16:100408 [FREE Full text] [doi: 10.1016/j.jpi.2024.100408] [Medline: 40094037]
19. Bashshur RL, Krupinski EA, Weinstein RS, Dunn MR, Bashshur N. The empirical foundations of telepathology: evidence of feasibility and intermediate effects. Telemed J E Health 2017;23(3):155-191 [FREE Full text] [doi: 10.1089/tmj.2016.0278] [Medline: 28170313]
20. Lennerz JK, Pantanowitz L, Amin MB, Eltoum I, Hameed MR, Kalof AN, et al. Ensuring remote diagnostics for pathologists: an open letter to the US Congress. Nat Med 2022;28(12):2453-2455. [doi: 10.1038/s41591-022-02040-6] [Medline: 36266514]
21. Weinstein RS, Descour MR, Liang C, Bhattacharyya AK, Graham AR, Davis JR, et al. Telepathology overview: from concept to implementation. Hum Pathol 2001;32(12):1283-1299. [doi: 10.1053/hupa.2001.29643] [Medline: 11774159]
22. Weinstein RS, Graham AR, Richter LC, Barker GP, Krupinski EA, Lopez AM, et al. Overview of telepathology, virtual microscopy, and whole slide imaging: prospects for the future. Hum Pathol 2009;40(8):1057-1069. [doi: 10.1016/j.humpath.2009.04.006] [Medline: 19552937]
23. Voelker HU, Strehl A, Stuefe A, Stauch G, Mueller-Hermelink H. Re-evaluation of challenging telepathological diagnoses in support of a hospital in Tanzania. J Telemed Telecare 2019;27(3):183-190. [doi: 10.1177/1357633x19866564] [Medline: 31370738]
24. Dietz RL, Hartman DJ, Pantanowitz L. Systematic review of the use of telepathology during intraoperative consultation. Am J Clin Pathol 2020;153(2):198-209 [FREE Full text] [doi: 10.1093/ajcp/aqz155] [Medline: 31618416]
25. Chong T, Palma-Diaz MF, Fisher C, Gui D, Ostrzega NL, Sempa G, et al. The California telepathology service: UCLA's experience in deploying a regional digital pathology subspecialty consultation network. J Pathol Inform 2019;10(1):31 [FREE Full text] [doi: 10.4103/jpi.jpi_22_19] [Medline: 31620310]
26. Jahn SW, Plass M, Moinfar F. Digital pathology: advantages, iimitations and emerging perspectives. J Clin Med 2020;9(11):3697 [FREE Full text] [doi: 10.3390/jcm9113697] [Medline: 33217963]
27. Mori I, Ozaki T, Muragaki Y, Ibata T, Ueda H, Shinagawa T, et al. Construction of web-based remote diagnosis system using virtual slide for routine pathology slides of the rural hospital in Japan. Diagn Pathol 2013;8(S1):1-4. [doi: 10.1186/1746-1596-8-s1-s4]
28. Subramanian H, Subramanian S. Improving diagnosis through digital pathology: proof-of-concept implementation using smart contracts and decentralized file storage. J Med Internet Res 2022;24(3):e34207 [FREE Full text] [doi: 10.2196/34207] [Medline: 35343905]
29. Li X, Cheng K, Shen Z, Lee PPC. Fast proactive repair in erasure-coded storage: analysis, design, and implementation. IEEE Trans Parallel Distrib Syst 2022;33(12):3400-3414. [doi: 10.1109/tpds.2022.3152817]
30. Chong T, Palma-Diaz MF, Fisher C, Gui D, Ostrzega NL, Sempa G, et al. The california telepathology service: UCLA's experience in deploying a regional digital pathology subspecialty consultation network. J Pathol Inform 2019;10(1):31 [FREE Full text] [doi: 10.4103/jpi.jpi_22_19] [Medline: 31620310]
31. Hanna MG, Ardon O, Reuter VE, Sirintrapun SJ, England C, Klimstra DS, et al. Integrating digital pathology into clinical practice. Mod Pathol 2022;35(2):152-164 [FREE Full text] [doi: 10.1038/s41379-021-00929-0] [Medline: 34599281]
32. Evans AJ, Vajpeyi R, Henry M, Chetty R. Establishment of a remote diagnostic histopathology service using whole slide imaging (digital pathology). J Clin Pathol 2021;74(7):421-424. [doi: 10.1136/jclinpath-2020-206762] [Medline: 32611763]
33. Rojansky R, Jhun I, Dussaq AM, Chirieleison SM, Nirschl JJ, Born D, et al. Rapid deployment of whole slide imaging for primary diagnosis in surgical pathology at stanford medicine: responding to challenges of the COVID-19 pandemic. Arch Pathol Lab Med 2023;147(3):359-367 [FREE Full text] [doi: 10.5858/arpa.2021-0438-OA] [Medline: 35802938]
34. Hanna MG, Reuter VE, Ardon O, Kim D, Sirintrapun SJ, Schüffler PJ, et al. Validation of a digital pathology system including remote review during the COVID-19 pandemic. Mod Pathol 2020;33(11):2115-2127 [FREE Full text] [doi: 10.1038/s41379-020-0601-5] [Medline: 32572154]

XSL•FO
RenderX

35. Costello SSP, Johnston DJ, Dervan PA, O'Shea DG. Development and evaluation of the virtual pathology slide: a new tool in telepathology. J Med Internet Res 2003;5(2):e11 [FREE Full text] [doi: 10.2196/jmir.5.2.e11] [Medline: 12857667]
36. Cui M, Zhang DY. Artificial intelligence and computational pathology. Lab Invest 2021;101(4):412-422 [FREE Full text] [doi: 10.1038/s41374-020-00514-0] [Medline: 33454724]

Abbreviations

DICOM: Digital Imaging and Communications in Medicine
DP: digital pathology
QC: quality control
TAT: turnaround time
TIFF: tagged image file format
VPN: virtual private network
WSI: whole-slide image

Edited by A Stone; submitted 29.Mar.2025; peer-reviewed by J Li, W Zhang; comments to author 09.Dec.2025; revised version received 05.Feb.2026; accepted 10.Feb.2026; published 09.Mar.2026.

Please cite as:
Shi J, Ye M, Sun D, He X, Lu Y, Wang L, Chen H, Wang W, Zhao J, Cui F
Telepathology and Mobile Health System for Province-Wide Pathology Consultation in Henan, China: Retrospective Evaluation Study
J Med Internet Res 2026;28:e75172
URL: https://www.jmir.org/2026/1/e75172
doi:10.2196/75172
PMID:

©Jinming Shi, Ming Ye, Dongxu Sun, Xianying He, Yaoen Lu, Linlin Wang, Haotian Chen, Wenchao Wang, Jie Zhao, Fangfang Cui. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Creation of EmpowerMe Website to Promote Self-Efficacy in Survivors of Stroke: Co-Design Study

Elizabeth A Lynch[1], BPTh, PhD; Adrian O'Malley[2]; Zoe Adey-Wakeling[3,4], BMed, PhD; Billie Bonevski[4], BA, PhD; Dominique A Cadilhac[5,6], BSN, PhD; Saran Chamberlain[1,7]; Leonid Churilov[8], BSc, PhD; Robyn Clark[1], BSN, PhD; Richard Cullen[9], BMMC; Erin Godecke[10,11], BSpP, PhD; Gillian Harvey[1], BSN, PhD; Fiona Jones[12], BPTh, PhD; Natasha A Lannin[13,14], BOccThy, PhD; Stacy Larcombe[1], BPTh; Jarrad Law[15,16], BEng; Annette McGrath[17]; Lisa Murphy[9], BMed, PhD; Katie Nesbitt[1], BSN, PhD; Karly Zacharia[18], BNAD, PhD; Niranjan Bidargaddi[15*], BEng, PhD; Coralie English[18,19,20*], BPTh, PhD

[1]Caring Futures Institute, College of Nursing and Health Sciences, Flinders University, College of Nursing and Health Sciences, Flinders University, Adelaide, Australia

[2]Lived Experience of Stroke, Sydney, Australia

[3]Flinders Medical Centre, Adelaide, Australia

[4]Flinders Health and Medical Research Institute, College of Medicine and Public Health, Flinders University, Adelaide, Australia

[5]Stroke and Ageing Research, Department of Medicine, School of Clinical Sciences, Monash University, Melbourne, Australia

[6]Public Health and Health Services Research, Stroke and Critical Care, The Florey, The University of Melbourne, Melbourne, Australia

[7]Lived Experience of Stroke, Adelaide, Australia

[8]Melbourne Medical School, The University of Melbourne, Melbourne, Australia

[9]National Stroke Foundation, Melbourne, Australia

[10]Stroke Recovery and Rehabilitation, Perron Institute for Neurological and Translational Science, Perth, Australia

[11]Allied Health Research, Sir Charles Gairdner Osborne Park Health Care Group, Perth, Australia

[12]Population Health Research Institute, City St George's University of London, London, United Kingdom

[13]Department of Neurosciences, School of Translational Medicine, Monash University, Melbourne, Australia

[14]Alfred Health, Melbourne, Australia

[15]Flinders Digital Health Research Centre, College of Medicine and Public Health, Flinders University, Adelaide, Australia

[16]CareMappr, Adelaide, Australia

[17]Lived Experience of Stroke, Port Elliot, Australia

[18]School of Health Sciences, University of Newcastle Australia, Newcastle, Australia

[19]Heart and Stroke Program, Hunter Medical Research Institute, Newcastle, Australia

[20]Centre for Research Excellence to Accelerate Stroke Trial Innovation and Translation, The University of Sydney, Sydney, Australia

[*]these authors contributed equally

Corresponding Author:
Elizabeth A Lynch, BPTh, PhD
Caring Futures Institute, College of Nursing and Health Sciences, Flinders University
College of Nursing and Health Sciences
Flinders University
GPO Box 2100
Adelaide, 5001
Australia
Phone: 61 0882017647
Email: elizabeth.lynch@flinders.edu.au

Abstract

Background: Digital health innovations are frequently used to support people in managing chronic health conditions. Stroke is common, and people who have survived a stroke and live in the community must learn to manage their health independently. Digital tools can help, but only if designed to match survivors' specific needs. In response to a need expressed by people living

XSL•FO
RenderX

with chronic health conditions, the Australian government created a funding stream to support the development of a digital resource to help individuals gain confidence in managing their health.

Objective: This study aimed to co-design a digital resource to promote self-efficacy to manage life after stroke in community-dwelling survivors of stroke.

Methods: Co-design methodology, which emphasized meaningful engagement with intended end-users throughout the design of the website, was used. The project steering group comprised health professional researchers, digital designers, people with lived experience of stroke (survivors and carers), and representatives from the Stroke Foundation (Australia). A systematic review was conducted to inform the core components of the resource. A lived experience workgroup was convened to advise on features of the digital resource and aspects of its evaluation. Iterative review stages and frequent consultation between the steering group and the lived experience workgroup occurred. Online and in-person usability testing was conducted with survivors and carers.

Results: The lived experience workgroup (workgroup) initially comprised 14 survivors and 1 carer. In total, 11 survivors and 1 carer remained engaged throughout the co-design period. One workgroup member was invited to join the steering group and co-facilitated all co-design workgroup meetings. Defining features of the digital resource were identified by the steering group and the workgroup, including that the resource would be a website that augmented (not reproduced) existing resources, and it needed to be accessible to survivors of stroke with communication changes. Website specifications were determined by the workgroup and included that information needed to be tailored to the individual user, and specific accessibility features were recommended. The workgroup prioritized what content to include on the website and recommended the creation of video stories by Australian survivors and carers. A resource-tailoring tool was created so information could be individualized to the interests of the website user. Overall, 9 web pages containing high-priority content were created, comprising text, video stories, and a downloadable PDF that summarized key information for that page. More than 150 short video stories were created by 26 survivors of stroke and 10 carers for the website. Usability testing indicated that the website was more usable than 83% of all websites.

Conclusions: Authentic co-design with inclusion of people with lived experience of stroke at all stages of development enabled the successful build of a digital resource (website) to improve self-efficacy. An evaluation of the website is underway.

Trial Registration: Australia New Zealand Clinical Trials Registry ACTRN12624001018505; https://anzctr.org.au/Trial/Registration/TrialReview.aspx?id=388255

(J Med Internet Res 2026;28:e76756) doi:10.2196/76756

KEYWORDS

stroke; website; self-efficacy; co-design; digital health

Introduction

Chronic health conditions are the leading causes globally of death and disability [1,2]. Effective ways to support people to manage their chronic health conditions have long been recognized as important to maximize health, quality of life, and reduce health care use [3,4]. As populations have become more familiar with using the internet to find health advice [5,6], and advances have been made in available digital health technologies [7,8], there has been a concurrent growth in digital health care delivery, which expanded exponentially with the COVID-19 pandemic [9,10]. Accordingly, digital health innovations are now commonly used to support people to manage their chronic health conditions [11,12].

In 2023, stroke was the second most common cause of death and disability worldwide [1,2]. Despite reductions in the incidence of stroke in high-income countries, the prevalence of people living with stroke is increasing because of aging populations and better survival rates [13]. Affecting more than 100 million people worldwide [14], stroke can affect an individual's physical ability, energy levels, mood, as well as their ability to communicate, concentrate, and understand [15]. Changes to cognition, communication, and stroke-related fatigue (all affecting at least half of all survivors) [16-18] can present difficulties for survivors in accessing digital health technologies [19,20]. Further, survivors of stroke with communication difficulties and people with cognitive changes are frequently excluded from participating in research [21,22], and little guidance exists in the published literature on how to work with survivors of stroke when designing a digital health innovation.

People living with chronic health conditions were consulted about research priorities to address in Australia and prioritized the development of digital resources to promote health self-efficacy [23]. Self-efficacy is an individual's confidence in their abilities [24], and when teamed with health, it relates to an individual's confidence in their ability to manage their health effectively. A research team was convened and sought to develop a digital resource to build health self-efficacy in survivors of stroke.

For any digital innovation, it is recommended that designers work with intended users during the development phase to optimize acceptability and uptake [25]. A co-design approach was chosen so survivors of stroke and other key stakeholders could be meaningfully involved and engaged throughout the entire research process [26], from conceptualizing and building the digital resource, through to its subsequent evaluation. This study describes how the team co-designed an accessible digital resource to promote self-efficacy in survivors of stroke.

XSL•FO
RenderX

Methods

Overview

The principles of design science in information systems research were used to guide the development of an innovative, purposeful digital resource [27]. The design process was planned to systematically incorporate the needs and preferences of survivors of stroke and carers through each planning and design stage of the digital resource [26]. In total, 4 steps were followed: determining what was required from the resource, determining specifications of the resource, building the resource to the specifications, and testing the resource to determine if it met requirements. These are presented in Figure 1.

Figure 1. Steps followed to design the EmpowerMe website and content.

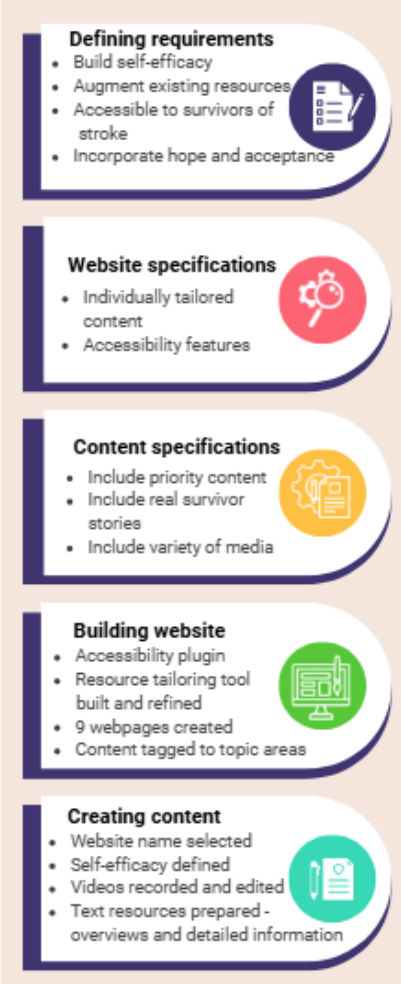

Formation of the Steering Group

The team that developed the research proposal and coordinated the study (hereafter referred to as the steering group) comprised 20 members (16 women), invited by Author EAL for their complementary expertise relevant to the project. The steering group included 14 academic researchers from health, bioinformatics, and biostatistics disciplines, 2 survivors of stroke, 1 carer, health professionals, and representatives from Stroke Foundation (Australia), the national stroke advocacy body. The steering group determined the features of the digital resource that were essential, and the adaptable features that should be determined by people with lived experience of stroke (Figure 2). The steering group met online monthly and had oversight of the resource design, timelines, and budget management.

XSL•FO
RenderX

Figure 2. Core and adaptable features of the digital resource.



Formation of Lived Experience Workgroup

A lived experience workgroup (hereafter referred to as the Workgroup) was convened, comprising survivors of stroke and carers, to advise on the adaptable features of the digital resource. Workgroup members were recruited by advertising expressions of interest via social media (Facebook [Meta] and X [formerly Twitter]). Interested individuals contacted a project team member (KZ), who then met online to explain and answer any questions about the project and discuss skills each individual could contribute to the project, in addition to their lived experience. Written information was provided about the project, and all Workgroup members provided informed consent to contribute their ideas and have online meetings video-recorded. Overall, 14 survivors and 1 carer (10 women) were recruited to the workgroup, representing community-dwelling people of different ages (ranging from 43 to 76 years), stroke-related changes, and time since stroke (6 months to 17 years at project commencement). Workgroup members were from 6 Australian states or territories. Some had extensive experience in research and co-design, whereas 3 had no prior research experience. Over the course of the 12-month co-design phase, 3 workgroup members withdrew due to overseas travel, illness, and being too busy.

Lived Experience Workgroup Meetings

Workgroup meetings were run online via Microsoft Teams and recorded. All Workgroup members were offered support to familiarize themselves with Microsoft Teams at the start of the project. Meetings were co-facilitated by an academic researcher (author EAL) and a survivor of stroke (authors SC or AO'M). Discussion points with or without preliminary reading were sent to attendees before each meeting. Each meeting was attended by a maximum of 5 workgroup members to allow time for contributions from each attendee, so 3 meetings (attended by different workgroup members) were held to discuss each topic.

At the start of each meeting, EAL presented relevant background to the intended discussions and the anticipated outcomes from each meeting. Author SC then facilitated discussions according to the predetermined discussion guide, inviting each person to contribute their ideas. Workgroup members were invited to "think big" and to challenge the status quo when discussing the digital resource design. At the conclusion of each meeting, EAL verbally summarized the key points raised in discussion, and Workgroup members provided agreement or clarification. Discussion points from 1 meeting were presented at subsequent meetings attended by different Workgroup members, to further explore these ideas and determine consensus or conflict. Meeting recordings were reviewed by author EAL, and key points of discussion, consensus, or conflict were documented and distributed to all attendees within a week of the final meeting. Attendee names were not included in meeting summaries. Attendees were invited to provide further details if they considered that information was missing. Facilitators SC and EAL shared key points from workgroup meetings at the bi-monthly steering group meetings, where final decisions about the digital resource design were made (in all instances, Workgroup recommendations were approved). Workgroup members were remunerated for their time at rates recommended by the New South Wales Consumer Alliance.

Overall, 6 rounds of workgroup co-design meetings were conducted between June and December 2023, interspersed with 3 online surveys between meetings, which followed up on workgroup discussions regarding content to include and how to measure the efficacy of the developed digital resource.

Online Surveys

Survey 1 comprised 19 questions, and Workgroup members were asked to rate the importance of including content about 18 different health-related needs (identified from the Australian Stroke Needs Assessment Survey [28]) using a 5-point Likert Scale (definitely include; probably include; consider including;

XSL•FO
RenderX

probably DO NOT include; definitely DO NOT include). An additional question allowed free-text responses seeking any other information to include. In Survey 2, workgroup members were asked to rate the importance (using a slider from 0 not at all important to 100 extremely important) of measuring 28 items related to self-efficacy that were identified in Workgroup discussions and from the Australian Stroke Needs Assessment Survey [28]. Respondents were also asked to nominate their top 3 items from the included list, and a final question allowed free text for any additional comments to consider about measuring the efficacy of the digital resource. In Survey 3, 2 shortlists of 10 items from Surveys 1 and 2 were presented. Workgroup members were asked to rank items in order of importance regarding the priority content to include and the priority aspects of self-efficacy to measure.

Survey Data Analysis

Data from Surveys 1 and 2 were analyzed descriptively. The frequency of "definitely include" responses for items in Survey 1 was calculated, and the means and frequency in the "top 3" were calculated for each item in Survey 2. Data from ranking exercises in Survey 3 were analyzed using a graph theory-based voting system [29], wherein group-based ordinal rankings were generated from the individual ordinal rankings.

Determining Technological Features

A further 3 rounds of co-design meetings with the Workgroup were co-facilitated by a digital designer and a survivor of stroke (author SC) to determine the technological features of the digital resource. The workgroup was reconvened between March and April 2024 to review the digital prototype. The topics discussed at the meetings are presented in Table 1. The project timeline is illustrated in Figure 3.

Table 1. Focus of workgroup meetings.

Meeting number	Main agenda items
1	• Introduction to the project • Discuss ways of working • Definition of self-efficacy • Personal experience of building self-efficacy after a stroke
2	• Terms of reference • Definition of self-efficacy • Do different groups of people have different self-efficacy needs? • Resources to support self-efficacy • Selection of a digital platform
3	• Measuring self-efficacy • Content to include on website
4	• Content to include on website • Existing resources that should be included • Preferred format • Previous (good and bad) learning experiences using digital technology
5	• Recap of progress to date • Format of resources • Features to include on website
6	• Experiences using websites for information about health or well-being • Features of websites that make them harder or easier to use, more or less enjoyable • Opinions on • Interactive features on websites • Sharing personal information on websites • Receiving "push" emails or text messages • Any changes to the way information is processed since the stroke
7	• Identifying an individual's information-seeking behaviors and digital preferences
8	• Reviewing the prototype of the quiz to tailor content
9	• Reviewing and refining the prototype of the quiz
10	• Reviewing website prototype • Feedback on layout and color palette • Name of website

XSL•FO
RenderX

Figure 3. Timeline of website co-design and usability testing.



Usability Testing

The tailoring of usability testing procedures and methods to meet the needs of people with language and cognitive changes has been presented elsewhere (Manuscript submitted to *Journal of Medical Internet Research* concurrently). A total of 3 groups of participants were recruited for usability testing: survivors of stroke with no self-reported language changes (n=4), survivors with moderately severe aphasia (a language disorder that affects communication; n=4), and carers of survivors of stroke (n=3). None had been involved in the design or development stage. Participants were offered an Aus $50 (US $35.45) gift voucher on completion of the testing session to thank them for participating.

Usability Testing Procedures

Testing was conducted in a one-to-one moderated format by author KN, with support from author EG for participants with aphasia. In-person testing was organized for all participants with aphasia, and in-person or online participation was offered to other participants. Usability scripts were prepared to guide participants to test navigation around the website and evaluate comprehension of the content (Multimedia Appendix 1).

Data Collection

Demographic data, time taken to complete tasks (for each task and total cumulative time), percentage of tasks completed, knowledge acquisition (questions about definition and importance of self-efficacy), System Usability Scale (SUS), and user experience (survey developed by the research team incorporating specifics of the EmpowerMe website, based on the Website Evaluation Questionnaire).

Usability Data Analysis

Participant demographic data and usability task-time data were analyzed descriptively using IBM SPSS Statistics 28.0.0.0. Knowledge acquisition was analyzed using the chunk and check technique [30,31] and reported narratively. Mean (SD) for overall, individual, and group SUS scores were analyzed using the SUS calculator package [32], and using this package, the raw mean score was normalized and converted to a percentile rank. The percentile rank indicated the relative usability of the website compared to other digital products. Group mean scores for survivors of stroke with and without aphasia were compared using a between-subject 2-sample *t* test (2-tailed), with statistical significance set at $P \leq .05$. Data from the user experience survey were analyzed descriptively and tabulated against task-time and task-completion data to examine usability from different perspectives, providing an opportunity to make targeted website changes when issues were identified.

Ethical Considerations

The project had ethical approval (5886) from Flinders University's Human Research Ethics Committee. People in the lived experience workgroup provided written informed consent for their contributions to be used to shape the digital resource design and for meetings to be video-recorded. Due to the group nature of workgroup meetings, the ideas generated were not confidential and were shared with other workgroup members. Anonymized summaries were shared with the project steering group. Workgroup members were offered Aus $100 (US $70) per meeting, inclusive of preparation time, paid via invoice or electronic gift cards.

Usability testing participants provided written informed consent to participate. The research speech pathologist (EG) prepared an aphasia-friendly participant information sheet and explained the information to each participant with aphasia, using aphasia-friendly language and allowing time for processing of information. She checked the participant's understanding of the information as part of the consent process. Data from usability testing participants were collected anonymously to protect participants' confidentiality, and only the research team had access to the data. Each participant was offered an Aus $50 (US $35) gift card as remuneration for completing the usability testing.

Results

Defining Requirements of the Digital Resource

The lived experience workgroup recommended that a website would be the most acceptable digital resource to support survivors of stroke develop their self-efficacy. Essential requirements determined by the Steering Group before project commencement were that the website would be free to access and needed to augment, not reproduce, information and other

XSL•FO
RenderX

supports already available (such as available through Stroke Foundation and its website EnableMe [33]). The website needed to be accessible to non-experts in stroke, with low or medium levels of health literacy, and accessible to people with stroke-related changes, including changes to cognition, communication, or reading ability. The Workgroup highlighted the importance of including representation of people of different ages, genders, and cultural backgrounds.

Another essential requirement of the website was that it needed to support self-efficacy in Australian survivors of stroke and people in their support network (family members, friends, and carers), so it should incorporate features such as learning from or observing other survivors and carers, and provide information. Elements that were identified by workgroup members as being important to building self-efficacy and a sense of well-being after stroke were having a purpose in life, having self-compassion, having goals to work toward, and being realistic while also having hope. Acceptance that things have changed and having resilience to bounce back if experiencing a setback were also recognized as intertwined with a person's confidence in their abilities. These elements guided the development of resources hosted on the website.

Setting Specifications

Specifications of Website

Workgroup members expressed that the website needed to incorporate a feature so that information could be tailored to each user, who could then "choose their own adventure" to find information or other resources relevant to them. Interactivity in the form of a chatbot was universally rejected by workgroup members. However, a feature where users were asked a series of questions about areas of interest or information needs that would then guide the user to a curated list of relevant resources was deemed to be of benefit. The Workgroup also recommended that the tailoring feature be optional, so users who preferred not to provide information about themselves, or people who prefer to "browse," could look through all the available resources.

In terms of accessibility features for survivors of stroke, the use of consistent layout and spacing was suggested to improve the accessibility of information for people with visual field loss. Use of sans serif font, having aphasia-friendly text options, and availability of captions for videos and adjustable playing speed were recommended to meet the needs of people with communication impairments. The Workgroup recommended having a text-to-voice option so users could choose to listen to rather than read the written information. Use of large boxes for clickable links (rather than underlined text) was recommended to improve functionality for people with motor control difficulties. Avoiding any spontaneous animations, background music, and moderating the tone of voice and volume on videos was important for people who experience sensory overload.

Other recommended features were not specific to the needs of survivors of stroke, including having options to adjust the website colors and having clear navigation pathways.

Specifications of Content to Include

Priority content to include on the website was determined by the workgroup following discussions and surveys. The list of priority content topics is presented in Table 2. The most important topics to address were "What is self-efficacy and why is it important?" and "How to be involved in health care decisions." Due to substantial overlap in content, 2 topics ("Managing concentration problems" and "Managing cognitive changes") were grouped, leaving 9 priority topics to address.

Table 2. Priority content to include and the number of videos created.

Topic number	Topics, prioritized by the lived experience workgroup	Number of videos	
		Survivors	Carers
1	Defining self-efficacy and its importance	50	9
2	Being involved in decisions	7	4
3	Seeking the information you need	10	6
4	Advocating for yourself	12	2
5	Managing fatigue	11	2
6	Emotional problems	10	3
7	Managing cognitive changes[a]	6	3
8	Managing mobility changes	11	2
9	Managing speaking difficulties	10	9
10	Managing concentration problems[a]	—[b]	—[b]

[a]Due to substantial overlap in content, "managing cognitive changes" and "managing concentration problems" were combined into one topic.

[b]Not available.

Results from a systematic review conducted by the research team about self-management interventions to improve self-efficacy in survivors of stroke [34] did not highlight any further components that needed to be incorporated into the website.

XSL•FO
RenderX

Specifications of Content Format

To meet different users' learning styles, the website needed to host content in a variety of media, including video, text (including downloadable information sheets), and audio resources. For written information, having both "bite-size" information chunks and "deep-dive" detailed information for topics was preferred, so a user could browse headlines of topics and then go into detail on topics of interest to them.

To enhance self-efficacy via vicarious learning from peers, the steering group and workgroup planned to include content delivered via testimonials from survivors of stroke and carers. There was a strong preference from workgroup members for these to be real, unscripted video-stories of Australian survivors and carers who shared their struggles and how these were managed, as well as their experiences of recovery. Specifically, Workgroup members recommended using home recordings (eg, via smartphone camera or Zoom videoconferencing) rather than professional videographers to prepare the video stories, as they felt this improved the relatability of the person sharing their story and made the stories and messages more credible. The preferred length of videos was less than 3 minutes.

For the delivery of information prepared by health professionals, Workgroup members preferred to have real people in any videos rather than using graphics or animations, and the use of Australian-specific information when available.

Building the Website and Developing Content

Incorporating Accessibility Features

An accessibility plugin (UserWay) was used to allow users to customize the site to their needs. This included options like increasing contrast, enlarging the text, increasing the text spacing, highlighting links so clickable elements are more prominent, and other options.

Building the Resource Tailoring Tool

The Workgroup members met with a digital designer and shared their poststroke information experiences, device usage patterns, content preferences, and strategies for validating health information. Following this session, the designer developed distinct user personas (Table 3) and used this information to develop a quiz-based tailoring prototype to accommodate all identified personas. The subsequent 2 sessions were used to refine the prototype through iterative feedback, emphasizing content relevance and quiz structure. Workgroup members provided critical input on optimal questionnaire length, preferred response formats, and content delivery mechanisms.

Following development and feedback, the resource tailoring tool was trialed and finalized. It comprised 2-3 (depending on responses) multiple-choice questions that subsequently became more precise to identify topics of interest to each user. Additional questions were added to the start of the tool to collect demographic data of users for research purposes (Textbox 1).

Table 3. Personas of survivors of stroke accessing digital information emerging from the design sessions.

Characteristics	Tech savvy (n=4)	Traditional learner (n=3)	Supported learner (n=3)
Digital skills	High confidence with multiple devices	Prefers structured content	Confident but needs accessibility features
Learning style	Self-directed research and critical evaluation	Relies on medical professionals	Institution-guided learning
Information sources	Multiple sources, fact-checking, and research background	Printed materials from traditional health channels	Verified institutional sources
Preferred content format	Short videos, blogs, and interactive content	Academic content and structured text	Short videos (<15 min) and articles
Device usage	Multiple devices (phone, tablet, and laptop)	Preference for larger screens	Phone and laptop primarily
Support needs	Independent learners	Guided learning approach	Accessibility adaptations

XSL•FO
RenderX

Textbox 1. Questions to guide content curation. Topic was a high priority and had a web page developed as part of this project.

What topics would you like to explore today?

How to build my confidence after a stroke

- Finding the information I need
- Learning from survivors or carers
- Practice
- Support from others
- Skill development

How to speak up for what I want or need

- Self-advocacy
- Involvement in decisions about my care

How to manage the effects of the stroke

- Managing physical changes
  - Using my arm
  - Bladder and bowel problems
  - Falls prevention
  - Mobility changes
  - Swallowing
- Managing changes that are hard to see
  - Vision
  - Cognitive (thinking and memory) changes
  - Speaking difficulties
  - Pain and sensation changes
  - Emotional changes
  - Fatigue

The resource tailoring tool was a plugin to the website built using Flutter Web, which sat on top of the website to add the quiz-based tailoring information. The structure of the Tailoring Tool and the content provided were stored in a Webiny Content Management System. Upon using the Tailoring Tool, relevant resources were pushed to website users – see Organizing Resources within the Website section.

Preferred Language and Definitions

Terminology was considered and workshopped in both workgroup and steering group meetings. Most workgroup members (except for individuals with occupational therapy or education backgrounds) were unfamiliar with the word “self-efficacy” at the start of the project, and some found it hard to pronounce. Many steering group members and workgroup members felt that the word “self-efficacy” was not part of general day-to-day language and were concerned that use of it might be off-putting. However, some workgroup members felt that it was important to educate survivors and carers about self-efficacy, and familiarization with the word “self-efficacy” was an important educational component that could empower survivors and carers. The Steering Group then decided to use lay language (avoiding the word “self-efficacy”) in headings and titles and include the word “self-efficacy” in second-tier text designed to inform and educate users.

After several iterations, the preferred definition of self-efficacy, determined by the Workgroup, was “Belief in my abilities to achieve the things that matter to me.”

Naming the Website

Initially, the Steering group anticipated naming the website ASSET-Stroke according to the project name (derived from co-designing a Digital Solution for Self-Efficacy After Stroke). However, both the steering group and workgroup agreed that this name did not encapsulate the philosophy and purpose of the website and was unlikely to attract the attention of survivors of stroke unfamiliar with the project. Accordingly, discussions were facilitated, and voting was conducted with the workgroup, and the domain name EmpowerMe was selected. This name aligns with others curated by Stroke Foundation, such as EnableMe [33] for survivors of stroke and carers and InformMe [35] for health professionals working with survivors of stroke.

XSL•FO
RenderX

Creating Website Content

Project team members created content and accompanying downloadable PDFs to address each priority topic area. All text was edited until it was deemed to be understandable by people who had completed grade 7 at school by Microsoft Copilot. Text resources for "Managing speaking difficulties" were edited by a speech pathologist with expertise in aphasia (author EG) and then checked and further edited with a survivor of stroke with moderate aphasia.

Expressions of interest were advertised for survivors of stroke and carers to create video stories to be hosted on the website about building self-efficacy after stroke, with a focus on the prioritized topic areas. Twenty-six survivors of stroke and 10 carers recorded video stories, usually over Zoom with a project team member, lasting up to 30 minutes. The project team edited the recordings into numerous short (30-s to 3-min) videos, creating more than 150 videos (Table 2). Each edited video was sent to the individual for approval or further editing before uploading to the Stroke Foundation YouTube channel, and being allocated to a topic area (one of the 9 priority topics when appropriate). A thumbnail with a brief title was provided for each video. Video contributors were reimbursed for their time.

Creating Web Pages

The main website was developed on WordPress, which allowed content to be easily updated by nontechnical team members, if needed, and was compatible with Stroke Foundation's systems to allow Stroke Foundation to manage the website on project completion. The 9 priority topics were organized into 9 web pages. The website home page presented information from the highest priority topic, "Defining self-efficacy and its importance." The resource tailoring tool was presented at the top of the page, presented as a prominent click box to "Find resources relevant to you."

The subsequent 8 topic web pages were created using a standard template, the use of icons, and formatting to ensure consistent layout. Each page had a lay description of the topic, a video about the topic, text about how to build self-efficacy in the topic, written content for carers about the importance of self-efficacy for carers or the carer's role in supporting self-efficacy of survivors, and thumbnails and titles of relevant videos created by survivors and carers relevant to the topic.

Organizing Resources Within the Website

Resources created as part of the project, and existing resources relevant to the priority content areas, were tagged by project team members regarding topic, format, and weighting in the Webiny Content Management System. Resources developed as part of the project were weighted higher than preexisting resources, and Australian resources were weighted higher than internationally produced resources. High-priority topics (identified during the co-design phase) had whole web pages and embedded resources developed, and the web page received top-weighting for the topic.

On using the resource tailoring tool, a curated list of relevant resources was created for each individual based on their responses regarding their areas of interest and associated resources' tags and weightings.

Usability Testing of the Website

Navigation of Website: Task Completion (Time Taken and Percentage Completed)

Survivors of stroke without aphasia (communication difficulties) and carers completed their tasks more quickly and completely than survivors with aphasia. Activities where survivors of stroke with aphasia struggled were reviewed by the team, and adaptations were made to enhance accessibility for this group of survivors.

Comprehension of Content

Overall, 6 participants (4 survivors with aphasia and 2 survivors without aphasia) participated in the evaluation of comprehensibility of website content. In total, 4 participants (2 [50%] survivors with aphasia and 2 [100%] survivors without aphasia) were able to understand and explain 100% of the content included in the knowledge assessment. One survivor with aphasia completed the "chunk and check" for 1 task only (20% of content included in knowledge assessment), and the remaining survivor with aphasia was unable to complete any knowledge tasks. These 2 survivors had more severe aphasia, so it was deemed that the website content was suitable for survivors with mild-moderate aphasia, and more work was required to meet the needs of survivors with severe post-stroke communication difficulties.

SUS

The SUS data were normally distributed, so overall mean scores, group mean scores, and between-subject comparison were calculated (comparing survivors with aphasia to survivors without aphasia, comparing survivors without aphasia to carers). SUS scores of survivors of stroke with aphasia were significantly lower than scores of survivors without aphasia (mean 61.3, SD 6.6 vs mean 86.8, SD 2.4; 95% CI –36.8 to –14.4). There was no statistically significant difference in SUS scores between carers and survivors without aphasia (mean 88.3, SD 11.5 vs mean 86.9, SD 2.4; 95% CI –23.3 to 26.3).

The normalized raw mean SUS score for the website was 78, and following the calculation of the percentile ranking, the usability of the EmpowerMe website was deemed better than 83% of all websites.

User Experience

Suggestions to improve the website from participants, which were also noted by the usability session moderator, included improving the visibility of resources after use of the resource tailoring tool; addressing layouts that were difficult for people with poststroke visual inattention; and improving problematic layouts (text heavy) for people with aphasia. These were subsequently addressed by the design team.

Website Efficacy

The potential of the digital resource "EmpowerMe" to influence self-efficacy, health-related quality of life, participation, and anxiety and depression in different cohorts of survivors

XSL•FO
RenderX

(stratified by age and time since stroke) is currently being evaluated as part of a phase II trial.

Discussion

Principal Results

We have presented how a digital resource in the form of a website can be built specifically to meet the needs of a client group with heterogeneous challenges, including cognitive, language, and motor changes. Rather than retrofit an existing website to the needs of people with poststroke visual, motor, cognitive, or sensory changes, we designed the website proactively to meet the diverse needs of the intended user group. People with lived experience of stroke were included at all stages, from preparation and submission of the grant, being members of the steering group, to facilitating and contributing in lived experience workgroup meetings and contributing to this article.

The Steering Group actively sought to balance the power dynamic between the academic researchers and people with lived experience. Following project commencement, the team recognized the need for more lived experience leadership to facilitate the lived experience workgroup meetings. One workgroup member (SC, survivor of stroke) with experience in both research and in facilitating online peer groups accepted an invitation to join the steering group, cofacilitate workgroup meetings, and was employed as a research assistant for the life of the project to assist in data analysis.

Within a research framework, we were successfully able to embed systems to support people with a lived experience of stroke to contribute meaningfully alongside a team of multidisciplinary experts to co-design a digital resource to address self-efficacy after stroke. Recognizing the different expertise of each party allowed the team to determine who decided each feature of the digital resource, to simultaneously address the needs and preferences of survivors of stroke and carers, and maintain the rigor and aims of the research process.

Survivors of stroke and carers, who comprised the lived experience workgroup, determined the digital platform (website with a feature to direct the user to individually relevant information), the priority content to include, and the priority format of resources (home-made video stories) to be developed by the project team. In contrast, the project Steering Group (which included 2 survivors of stroke and 1 carer) made fewer decisions but worked consistently to incorporate feedback and suggestions from the lived experience workgroup, positioning these within best-practice frameworks of adult learning, self-efficacy, self-management, and digital design.

Comparison With Prior Work

There is a growing body of literature documenting how to facilitate co-design and interactive workshops online, particularly since the COVID-19 pandemic [36-39]. Now that social distancing measures have lifted, online methods of co-designing have remained popular, with obvious benefits of reducing travel time, costs, and increasing accessibility for people who live in regional and remote areas [39,40]. While some researchers in previous work have described using interactive features of online platforms to enhance activity within online co-design meetings [37,41], interactions with the lived experience working group in the current project were mostly dialogue-based, although they were augmented with online surveys between meetings, similar to what others have also described [41].

The project reported in this article is far from the first website to have involved intended users in the design. Supporters of collaborative design methods argue that involving target users when designing a digital resource will improve the resource's impact through better user engagement and satisfaction [40,42-44]. However, the level of involvement of the intended users can vary greatly between studies that report using co-design methods. While co-design has been described as “meaningful end-user engagement in research design and includes instances of engagement that occur across all stages of the research process” [26], it is also acknowledged that the engagement activities “range in intensity from relatively passive to highly active and involved [26]. In a systematic review which tabled the most common co-design activities in studies reporting co-design of digital health solutions, 2 of the most common activities reported were interviews and surveys [45]. Implicit in these activities are the research team setting the agenda, seeking a one-way transference of “data” from participants, and controlling how the data are analyzed, which can be disempowering (rather than engaging) for participants [46]. Other commonly reported co-design activities that may promote more engagement were workshops [45], which tend to be more collaborative and allow for a two-way exchange of information, having been described as occasions in which a group of people can acquire new knowledge or perform creative problem-solving [47].

Intended users have been variably involved in determining whether a website would be of benefit, ranging from not being involved in the decision, (for instance when the rationale of developing a website is based on disease incidence and rates of technology use [43]), to contributing data that is used by researchers to support the project idea (eg, rationale is underpinned by interviews of participants living with the condition under investigation [48]) to sharing responsibility for the decision to develop the website (eg, when people with lived experience are members of the research team and help shape the research proposal [49,50]).

Similarly, the level of input and decision-making responsibility of representatives of the intended user group about features of the co-designed websites varies. Intended users could contribute data to be used by the design team to make decisions (such as people being recruited as participants in interviews, focus groups or co-design sessions) [42,48], they could advise on features (eg, via lived experience advisory groups) [42,48,50] or be decision-making partners (eg, via being members of the research or design team) [49,50]. Just as in our project, several studies reported involving people from the intended user group in different roles within the project [42,48,50], for instance, with some people recruited as participants, others consulting in an advisory capacity, and others partnering on the research team. We have advanced the field beyond existing data tools by explicitly working with the intended end users to embed features

XSL•FO
RenderX

to tailor content and accessibility to survivors of stroke. Unique to this project was the need to develop a website that was accessible to people with language, visual, motor, cognitive, or sensory changes from their stroke. By proactively incorporating accessibility features into the design (rather than retrofitting these to a standard website design), we anticipate the website will address the variety of challenges faced by survivors of stroke. Further, by co-creating content, the voices and preferences of survivors of stroke have been prioritized to support peer learning and peer support.

Compared to prior work, the current project used more active and involved methods of working with people with lived experience of stroke throughout the design of the website than has been reported in most other publications reporting on co-design of websites to support health or well-being. We identified 1 published work presenting the co-design of a website to support physical activity and diet to reduce secondary stroke [50]. This was also conducted in Australia, with 4 academic researchers and 1 team member from the Stroke Foundation also involved in the current project. These 2 Australian projects involved contributions from people with lived experience of stroke and included survivors of stroke in the authorship team; the current project convened a Lived Experience workgroup, and Pogrebnoy et al [50] consulted people with lived experience via a Consumer Advisory Group (1 survivor of stroke advised on both projects). Differences include that the website developed by Pogrebnoy was an adaptation of a newly created telehealth service, whereas in this study, the program and content were co-designed and integrated as part of the website build. Together, the projects demonstrate that collaborative ways of working are feasible, even when working with people who experience communication, cognitive, or physical challenges after their stroke. These ways of working are transferable to other clinical fields, and the methods reported in this study could be used by other research and design teams.

Limitations

By nature of engagement, wherein meetings were conducted online via Microsoft Teams, the lived experience workgroup members who contributed over the duration of the project tended to have high levels of digital literacy. This was a proof-of-concept project, and there was no specific strategy to recruit people from linguistically or culturally diverse backgrounds. Several team members (on the Steering group and Lived Experience Workgroup) were born in countries other than Australia, but most team members spoke English as a first language. Should the website indicate promise at improving outcomes for survivors of stroke, future work will be undertaken to engage with culturally and linguistically diverse groups of survivors and carers, and with people with lower digital literacy to determine their interest in accessing and using a website to promote self-efficacy. Partnership with the community leaders would be sought so that similar methods as have been described in this study could be used to understand and address the needs and preferences of groups of people that have not yet been represented on the EmpowerMe website.

Conclusions

We have demonstrated how a multidisciplinary team of academic researchers, people with lived experience of stroke, and Stroke Foundation representatives can work collaboratively from the inception of a project to co-design a digital resource known as the “EmpowerMe” website, which hosts content to meet the needs of survivors with stroke to support self-efficacy. Our team was convened to answer a question raised by people with lived experience of chronic health conditions, the team comprised people with varied skills and experiences (including people with lived experience of stroke and Stroke Foundation representatives), we convened a lived experience advisory group which we consulted over the course of the project to ensure the resource met the needs of the intended user group. Finally, we had clear roles and responsibilities for the steering group, the lived experience workgroup, and the usability testing participants. In this way, this project can serve as a reference for other co-designed digital health resources and offer practical lessons for similar initiatives.

Acknowledgments

The authors acknowledge the input of members of the Lived Experience Workgroup: Toni Arfaras, Kim Beesley, Brian A. Beh, Saran Chamberlain, Alison Douglas, Emma Gee, Clive Kempson, Letisha Living, Peter Meers, Susan Mueller, Veraina Ryan, and Mabel Ventura.

Funding

This project was funded by a grant from the Australian Government (National Health and Medical Research Council 2021 Medical Research Future Fund Consumer-Led Application 2022828). The sponsor determined the project would involve the development of a digital resource to promote health self-efficacy, but has had no input into the research design or conduct.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

Conceptualization (EAL, AO'M, BB, DC, LC, EG, NL, LM, NB, CE, ZA-W, Robyn Clark, Richard Cullen, GH, FJ, and AM), data curation (EAL, SC, AO'M, KN, SL, KN, and KZ), formal analysis (LC, EAL, and KN), funding acquisition (EAL, AO'M, BB, DC, LC, EG, NL, LM, NB, CE, ZA-W, Robyn Clark, Richard Cullen, GH, FJ, and AM), investigation (EAL, SC, AO'M, KN, EG, and NB), methodology (EAL, NL, LC, CE, Richard Cullen, NB, KN, and EG), project administration (KN and EAL), resources (SL, KN, EAL, KZ, and SC), software (NB and JL), supervision (EAL, CE, and Richard Cullen) visualization (AO'M, SC, Richard Cullen, NB, JL, EAL, and KN) writing – original draft (EAL, NB, and KN), writing – review and editing (AO'M, BB, DC, LC, EG, NL, LM, CA, ZA-W, RC, RC, GH, FJ, and AM). All authors have contributed to the shaping of the website through Steering group and project group discussions.

Conflicts of Interest

NB is cofounder and shareholder of CareMappr Pty Ltd. The other authors declare no conflicts of interest.

Multimedia Appendix 1
Usability testing script and list of customizable features.
[DOCX File , 22 KB - jmir_v28i1e76756_app1.docx]

References

1. GBD 2023 Causes of Death Collaborators. Global burden of 292 causes of death in 204 countries and territories and 660 subnational locations, 1990-2023: a systematic analysis for the global burden of disease study 2023. Lancet 2025;406(10513):1811-1872 [FREE Full text] [doi: 10.1016/S0140-6736(25)01917-8] [Medline: 41092928]
2. GBD 2023 DiseaseInjuryRisk Factor Collaborators. Burden of 375 diseases and injuries, risk-attributable burden of 88 risk factors, and healthy life expectancy in 204 countries and territories, including 660 subnational locations, 1990-2023: a systematic analysis for the global burden of disease study 2023. Lancet 2025;406(10513):1873-1922 [FREE Full text] [doi: 10.1016/S0140-6736(25)01637-X] [Medline: 41092926]
3. Barlow J, Wright C, Sheasby J, Turner A, Hainsworth J. Self-management approaches for people with chronic conditions: a review. Patient Educ Couns 2002;48(2):177-187. [doi: 10.1016/s0738-3991(02)00032-0] [Medline: 12401421]
4. Ahn S, Basu R, Smith ML, Jiang L, Lorig K, Whitelaw N, et al. The impact of chronic disease self-management programs: healthcare savings through a community-based intervention. BMC Public Health 2013;13:1141 [FREE Full text] [doi: 10.1186/1471-2458-13-1141] [Medline: 24314032]
5. Rice RE. Influences, usage, and outcomes of Internet health information searching: multivariate results from the Pew surveys. Int J Med Inform 2006;75(1):8-28 [FREE Full text] [doi: 10.1016/j.ijmedinf.2005.07.032] [Medline: 16125453]
6. Cheng C, Dunn M. Health literacy and the internet: a study on the readability of Australian online health information. Aust N Z J Public Health 2015;39(4):309-314 [FREE Full text] [doi: 10.1111/1753-6405.12341] [Medline: 25716142]
7. Yeung AWK, Torkamani A, Butte AJ, Glicksberg BS, Schuller B, Rodriguez B, et al. The promise of digital healthcare technologies. Front Public Health 2023;11:1196596 [FREE Full text] [doi: 10.3389/fpubh.2023.1196596] [Medline: 37822534]
8. Madanian S, Nakarada-Kordic I, Reay S, Chetty T. Patients' perspectives on digital health tools. PEC Innov 2023;2:100171 [FREE Full text] [doi: 10.1016/j.pecinn.2023.100171] [Medline: 37384154]
9. Senbekov M, Saliev T, Bukeyeva Z, Almabayeva A, Zhanaliyeva M, Aitenova N, et al. The recent progress and applications of digital technologies in healthcare: a review. Int J Telemed Appl 2020;2020:8830200 [FREE Full text] [doi: 10.1155/2020/8830200] [Medline: 33343657]
10. Golinelli D, Boetto E, Carullo G, Nuzzolese AG, Landini MP, Fantini MP. Adoption of digital technologies in health care during the COVID-19 pandemic: systematic review of early scientific literature. J Med Internet Res 2020;22(11):e22280 [FREE Full text] [doi: 10.2196/22280] [Medline: 33079693]
11. Pong C, Tseng RMWW, Tham YC, Lum E. Current implementation of digital health in chronic disease management: scoping review. J Med Internet Res 2024;26:e53576 [FREE Full text] [doi: 10.2196/53576] [Medline: 39666972]
12. Bashi N, Fatehi F, Mosadeghi-Nik M, Askari MS, Karunanithi M. Digital health interventions for chronic diseases: a scoping review of evaluation frameworks. BMJ Health Care Inform 2020;27(1):e100066 [FREE Full text] [doi: 10.1136/bmjhci-2019-100066] [Medline: 32156751]
13. Li L, Scott CA, Rothwell PM, Oxford Vascular Study. Trends in stroke incidence in high-income countries in the 21st century: population-based study and systematic review. Stroke 2020;51(5):1372-1380 [FREE Full text] [doi: 10.1161/STROKEAHA.119.028484] [Medline: 32208842]
14. Global stroke fact sheet 2022. World Stroke Organisation. 2022. URL: https://www.world-stroke.org/assets/downloads/WSO_Global_Stroke_Fact_Sheet.pdf [accessed 2026-02-10]
15. Langhorne P, Bernhardt J, Kwakkel G. Stroke rehabilitation. Lancet 2011;377(9778):1693-1702. [doi: 10.1016/S0140-6736(11)60325-5] [Medline: 21571152]

XSL•FO
RenderX

16. El Husseini N, Katzan IL, Rost NS, Blake ML, Byun E, Pendlebury ST, et al. Cognitive impairment after ischemic and hemorrhagic stroke: a scientific statement from the American Heart Association/American Stroke Association. Stroke 2023;54(6):e272-e291 [FREE Full text] [doi: 10.1161/STR.0000000000000430] [Medline: 37125534]
17. De Cock E, Batens K, Hemelsoet D, Boon P, Oostra K, De Herdt V. Dysphagia, dysarthria and aphasia following a first acute ischaemic stroke: incidence and associated factors. Eur J Neurol 2020;27(10):2014-2021. [doi: 10.1111/ene.14385] [Medline: 32515514]
18. Alghamdi I, Ariti C, Williams A, Wood E, Hewitt J. Prevalence of fatigue after stroke: a systematic review and meta-analysis. Eur Stroke J 2021;6(4):319-332 [FREE Full text] [doi: 10.1177/23969873211047681] [Medline: 35342803]
19. Lee J, Worthy P, Deslandes R, Burton B, Copland DA, Jamieson P, et al. Bridging the digital divide for people with aphasia: a study protocol for codesigning web accessibility tools and guidelines. BMJ Open 2025;15(8):e099273 [FREE Full text] [doi: 10.1136/bmjopen-2025-099273] [Medline: 40784777]
20. Niyomyart A, Ruksakulpiwat S, Benjasirisan C, Phianhasin L, Nigussie K, Thorngthip S, et al. Current status of barriers to mHealth access among patients with stroke and steps toward the digital health era: systematic review. JMIR Mhealth Uhealth 2024;12:e54511 [FREE Full text] [doi: 10.2196/54511] [Medline: 39173152]
21. Brady MC, Fredrick A, Williams B. People with aphasia: capacity to consent, research participation and intervention inequalities. Int J Stroke 2013;8(3):193-196. [doi: 10.1111/j.1747-4949.2012.00900.x] [Medline: 23130972]
22. Taylor JS, DeMers SM, Vig EK, Borson S. The disappearing subject: exclusion of people with cognitive impairment and dementia from geriatrics research. J Am Geriatr Soc 2012;60(3):413-419. [doi: 10.1111/j.1532-5415.2011.03847.x] [Medline: 22288835]
23. Archived grant opportunity view - GO5353. Grant Connect. 2025. URL: https://www.grants.gov.au/Go/Show?GoUuid=00ba3a66-86cd-4b6d-b880-3f00cb9e5152 [accessed 2026-02-10]
24. Bandura A. Social cognitive theory: an agentic perspective. Annu Rev Psychol 2001;52:1-26. [doi: 10.1146/annurev.psych.52.1.1] [Medline: 11148297]
25. Seixas AA, Olaye IM, Wall SP, Dunn P. Optimizing healthcare through digital health and wellness solutions to meet the needs of patients with chronic disease during the COVID-19 era. Front Public Health 2021;9:667654 [FREE Full text] [doi: 10.3389/fpubh.2021.667654] [Medline: 34322469]
26. Slattery P, Saeri AK, Bragge P. Research co-design in health: a rapid overview of reviews. Health Res Policy Syst 2020;18(1):17 [FREE Full text] [doi: 10.1186/s12961-020-0528-9] [Medline: 32046728]
27. Hevner A, March S, Park J, Ram S. Design science in information systems research. MIS Quarterly 2004;28:75-105 [FREE Full text]
28. Andrew NE, Kilkenny M, Naylor R, Purvis T, Lalor E, Moloczij N, et al. Understanding long-term unmet needs in Australian survivors of stroke. Int J Stroke 2014;9Suppl A100:106-112. [doi: 10.1111/ijs.12325] [Medline: 25042019]
29. Utley M, Gallivan S, Mills M, Mason M, Hargraves C. A consensus process for identifying a prioritised list of study questions. Health Care Manag Sci 2007;10(1):105-110. [doi: 10.1007/s10729-006-9003-6] [Medline: 17323658]
30. Maqbool B, Herold S. Potential effectiveness and efficiency issues in usability evaluation within digital health: a systematic literature review. Journal of Systems and Software 2024;208:111881 [FREE Full text] [doi: 10.1016/j.jss.2023.111881]
31. Talevski J, Wong Shee A, Rasmussen B, Kemp G, Beauchamp A. Teach-back: a systematic review of implementation and impacts. PLoS One 2020;15(4):e0231350 [FREE Full text] [doi: 10.1371/journal.pone.0231350] [Medline: 32287296]
32. Sauro J. A Practical Guide to Measuring Usability. Denver, USA: Measuring Usability LLC; 2011.
33. Find everything you need to live well after stroke. Stroke Foundation, Enable me. 2025. URL: https://enableme.org.au/ [accessed 2026-02-10]
34. Lynch EA, Nesbitt K, Gulyani A, Chan RJ, Bidargaddi N, Cadilhac DA, et al. Do self-management interventions improve self-efficacy and health-related quality of life after stroke? A systematic review. Int J Stroke 2025;20(7):786-800 [FREE Full text] [doi: 10.1177/17474930251340286] [Medline: 40275483]
35. Delivering quality stroke care. Stroke Foundation, Inform Me. 2025. URL: https://informme.org.au/ [accessed 2026-02-10]
36. Kennedy A, Cosgrave C, Macdonald J, Gunn K, Dietrich T, Brumby S. Translating co-design from face-to-face to online: an Australian primary producer project conducted during COVID-19. Int J Environ Res Public Health 2021;18(8):4147 [FREE Full text] [doi: 10.3390/ijerph18084147] [Medline: 33919920]
37. Fails JA, Ratakonda DK, Koren N, Elsayed-Ali S, Bonsignore E, Yip J. Pushing boundaries of co-design by going online: lessons learned and reflections from three perspectives. Int J Child Comput Interact 2022;33:100476 [FREE Full text] [doi: 10.1016/j.ijcci.2022.100476] [Medline: 36034969]
38. Istanboulian L, Rose L, Yunusova Y, Dale C. Adapting co-design methodology to a virtual environment: co-designing a communication intervention for adult patients in critical care. Res Involv Engagem 2023;9(1):103 [FREE Full text] [doi: 10.1186/s40900-023-00514-6] [Medline: 37957776]
39. Sanders EBN, Shen Y. Direct vs. remote data collection methods in co-design: navigating the evolving landscape. CoDesign 2025:1-15 [FREE Full text] [doi: 10.1080/15710882.2025.2472172]
40. Ramage ER, Burke M, Galloway M, Graham ID, Janssen H, Marsden DL, et al. Fit for purpose. Co-production of complex behavioural interventions. A practical guide and exemplar of co-producing a telehealth-delivered exercise intervention for

XSL•FO
RenderX

people with stroke. Health Res Policy Syst 2022;20(1):2 [FREE Full text] [doi: 10.1186/s12961-021-00790-2] [Medline: 34980156]
41. Mallakin M, Dery C, Vaillancourt S, Gupta S, Sellen K. Web-based co-design in health care: considerations for renewed participation. Interact J Med Res 2023;12:e36765 [FREE Full text] [doi: 10.2196/36765] [Medline: 36595738]
42. Morse B, Soares A, Ytell K, DeSanto K, Allen M, Holliman BD, et al. Co-design of the transgender health information resource: web-based participatory design. J Particip Med 2023;15:e38078 [FREE Full text] [doi: 10.2196/38078] [Medline: 36626222]
43. Wright M, Thomas KC, Carpenter D, Lee C, Coyne I, Garcia N, et al. Co-designing a website with and for youth, so they can better manage their health. PEC Innov 2023;2:100164 [FREE Full text] [doi: 10.1016/j.pecinn.2023.100164] [Medline: 37384159]
44. Brown M, Lord E, John A. Adaptation of ACTivate your wellbeing, a digital health and well-being program for young persons: co-design approach. JMIR Form Res 2023;7:e39913 [FREE Full text] [doi: 10.2196/39913] [Medline: 37052994]
45. Sanz MF, Acha BV, García MF. Co-design for people-centred care digital solutions: a literature review. Int J Integr Care 2021;21(2):16 [FREE Full text] [doi: 10.5334/ijic.5573] [Medline: 33981193]
46. Kvale S. Dominance through interviews and dialogues. Qualitative Inquiry 2006;12(3):480-500 [FREE Full text] [doi: 10.1177/1077800406286235]
47. Ørngreen R, Levinsen K. Workshops as a research methodology. Electron J e-Learn 2017;15(1):70-81 [FREE Full text]
48. Santin O, McShane T, Hudson P, Prue G. Using a six-step co-design model to develop and test a peer-led web-based resource (PLWR) to support informal carers of cancer patients. Psychooncology 2019;28(3):518-524. [doi: 10.1002/pon.4969] [Medline: 30597666]
49. Donald M, Beanlands H, Straus SE, Smekal M, Gil S, Elliott MJ, et al. A web-based self-management support prototype for adults with chronic kidney disease (My Kidneys My Health): Co-design and usability testing. JMIR Form Res 2021;5(2):e22220 [FREE Full text] [doi: 10.2196/22220] [Medline: 33560245]
50. Ashton L, Beh BA, Burke M, Cullen R, Czerenkowski J, Davey J, et al. Adapting a telehealth physical activity and diet intervention to a co-designed website for self-management after stroke: tutorial. J Med Internet Res 2024;26:e58419 [FREE Full text] [doi: 10.2196/58419] [Medline: 39437389]

Abbreviations

SUS: System Usability Scale

Edited by A Stone; submitted 30.Apr.2025; peer-reviewed by J Hao, Y Qian, T DÜzcÜ; comments to author 01.Dec.2025; revised version received 03.Feb.2026; accepted 04.Feb.2026; published 12.Mar.2026.

Please cite as:
Lynch EA, O'Malley A, Adey-Wakeling Z, Bonevski B, Cadilhac DA, Chamberlain S, Churilov L, Clark R, Cullen R, Godecke E, Harvey G, Jones F, Lannin NA, Larcombe S, Law J, McGrath A, Murphy L, Nesbitt K, Zacharia K, Bidargaddi N, English C
Creation of EmpowerMe Website to Promote Self-Efficacy in Survivors of Stroke: Co-Design Study
J Med Internet Res 2026;28:e76756
URL: https://www.jmir.org/2026/1/e76756
doi:10.2196/76756
PMID:

©Elizabeth A Lynch, Adrian O'Malley, Zoe Adey-Wakeling, Billie Bonevski, Dominique A Cadilhac, Saran Chamberlain, Leonid Churilov, Robyn Clark, Richard Cullen, Erin Godecke, Gillian Harvey, Fiona Jones, Natasha A Lannin, Stacy Larcombe, Jarrad Law, Annette McGrath, Lisa Murphy, Katie Nesbitt, Karly Zacharia, Niranjan Bidargaddi, Coralie English. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Insights and Recommendations From Moderators and Community Members for Keeping Online Peer Support Safe: Thematic Analysis

Hannah Grace Jones[1,2,3], BSc; Grace Lavelle[1], PhD; Elly Aylwin-Foster[1], BA; Ciara Regan[4], BSc; Alan Simpson[4,5], PhD; Ewan Carr[6], PhD; Matthew Hotopf[1*], PhD; Vanessa Lawrence[4*], PhD

[1]Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[2]Lifelong Health, Persistent Pain Research Group, Hopwood Centre for Neurobiology, South Australian Health and Medical Research Institute, Adelaide, South Australia, Australia

[3]IIMPACT (Innovation, IMPlementation and Clinical Translation) in Health, School of Allied Health and Human Performance, Adelaide University, Adelaide, South Australia, Australia

[4]Health Service and Population Research, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[5]Florence Nightingale Faculty of Nursing, Midwifery & Palliative Care, King's College London, United Kingdom, London, United Kingdom

[6]Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, United Kingdom, London, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Hannah Grace Jones, BSc
Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience, King's College London
16 De Crespigny Park
London, SE5 8AF
United Kingdom
Phone: 44 02078480002
Email: hannah.g.jones@kcl.ac.uk

Abstract

Background: Online peer support can help people living with long-term physical health conditions to manage their mental well-being. Although the potential negative events that can occur and risks associated with web-based peer communities are well recognized, our understanding of how best to moderate these spaces is relatively limited, particularly with regard to new communities. Previous work has focused on the experiences of either moderators or community members.

Objective: This study aims to explore the perspectives of both members and moderators of a new online peer support community to evaluate the moderation procedures and inform recommendations for best practice.

Methods: Community members (n=39) who participated in a research trial of a new online peer community, CommonGround, were interviewed. The moderation team (n=5) was invited to a focus group. Community member interviews explored their opinions of moderation policies and the behavior of the moderation team. The moderator focus group explored their experiences of moderating the community, including perceived benefits, common challenges, and areas for improvement. All interviews and the focus group were conducted online, audio-recorded, and transcribed verbatim. An inductive thematic analysis was conducted to sort the data into overarching themes through an iterative process.

Results: Effective moderation was considered critical in creating a safe space that members wanted to engage with and for mitigating any risks, particularly around the spread of medical misinformation. Both moderators and community members felt that the moderation policies and practices were appropriate and applicable to the community. Moderators found navigating the moderation threshold, where they balanced safety against free speech, challenging when determining whether to intervene or not. Being part of a team with mixed clinical expertise helped moderators build confidence in navigating this threshold and also presented other benefits of easy access to support and improving the consistency of their moderation practices. It was suggested that in order for a community to flourish, community members would self-moderate. However, moderators and members felt that the strong community culture and high levels of member engagement that are needed to support self-moderation had not yet evolved. Proposed improvements to moderation included new features to support the efficiency of identifying new content for review and reviewing the rule of anonymity.

XSL•FO
RenderX

Conclusions: Moderation is critical in making online peer communities feel safe and engaging. Moderation practices should be co-produced with the target audience to ensure that they are aligned with the community's unique moderation wants and needs, including clear escalation pathways, transparent communication patterns, and plans to review and update policies or procedures as the community evolves. There should be technological features that promote self-moderation, as the community may shift towards self-moderation as it matures. It is also critical to ensure that moderators feel supported so that they are best placed to support the broader community.

Trial Registration: ClinicalTrials.gov NCT06222346; https://clinicaltrials.gov/study/NCT06222346

International Registered Report Identifier (IRRID): RR2-10.2196/71513

(J Med Internet Res 2026;28:e81943) doi:10.2196/81943

KEYWORDS

chronic conditions; depression; long-term physical health conditions; mental wellbeing; moderation; online peer support; online safety

Introduction

Connecting with peers or other people "like me" who have similar life experiences can be a powerful way of receiving social support and managing one's mental well-being [1,2]. By engaging with peers, people can exchange emotional support and their experiential knowledge [3]. In the last 10 years, there has been a rise in the use of online peer support, with the web-based setting facilitating convenient and flexible engagement that overcomes any physical barriers to accessing in-person support groups [4]. In particular, asynchronous peer community forums have become popular, as they provide the opportunity to exchange support with a potentially unlimited number of peers around the clock [5,6]. Those living with mental health conditions who have engaged with web-based peer support have experienced improved mental health symptoms, quality of life, and medical self-management skills [2,7,8]. Web-based peer support, therefore, offers a highly scalable, low-cost intervention that has the potential to promote mental well-being and reduce the burden of negative mental health on the individual, society, and health care system [4,9].

One population who may particularly benefit from online peer support is people living with long-term physical health conditions, as this population often reports experiencing psychological distress, loneliness, and disconnection from people in their social circles [10-12]. Around 15 million people in the United Kingdom live with a long-term physical health condition [13]. Of these, approximately 3 million are also living with symptoms of low mood that negatively impact their physical health [14], hinder effective self-management, and impair quality of life [15,16]. These symptoms of low mood also put them at risk of developing major depressive disorder, highlighting the critical need to intervene early to mitigate the impacts of mental-physical comorbidities on individuals, families, society, and health services [17-19]. Despite initial evidence documenting the psychosocial benefits of engaging with online peer support for people with long-term conditions [1,20-24], the provision of peer support specifically tailored to the needs of this population is relatively limited [25,26]. Our own qualitative research with people with diverse long-term physical health conditions found that traditional condition-specific communities can limit their ability to connect on other commonalities beyond their diagnosis and may not be appropriate for those with rare or multiple conditions [27,28]. To address this unmet need, we worked alongside people living with long-term health conditions, clinicians, and software developers in an iterative coproduction process to create CommonGround: an online peer support and psychoeducation platform that aims to intervene early to prevent the progression of symptoms of low mood to major depressive disorder [29]. By placing the voices of people with long-term health conditions in the foreground, we have tailored the intervention design and features specifically to their needs, including the all-diagnoses-welcome ethos that reflects the view that many experiences and challenges transcend diagnoses.

While online peer support communities have the potential to enhance the well-being of their members, there are challenges and risks associated with them. People with long-term health conditions have shared concerns about the potential negative outcomes from engaging digitally with their peers and the need for web-based spaces to feel safe [27,30,31]. Web-based forums carry inherent risks of medical misinformation and the sharing or encouragement of risky and unsafe behaviors [9,31]. Hearing about peers' experiences can potentially exacerbate low mood and promote symptom-comparisons that generate anxiety about one's condition [27,32-34]. Behind the screen of anonymity, people can argue, have hostile interactions, and experience cyberbullying or trolling [35,36]. It is well acknowledged that appropriate community governance is essential in reducing the risk of these negative outcomes arising, particularly among communities of people with long-term health conditions who are likely to discuss sensitive, personal, and medical-related topics.

One means of mitigating the risks of online peer support is through moderation by professional moderators. Clear, consistent, and comprehensive moderation policies play an essential role in transparently curating trust and safety in the community. By clearly outlining boundaries and acceptable behavior, community members can make judgments on whether the community is a space they wish to attend and have a clear understanding of how to behave within the community. Professional moderators can play an essential role in implementing these policies, including identifying and removing inappropriate content, preventing conflicts between members, enforcing the community guidelines, and supporting those in potential crisis [37-39]. Often, these moderators consider

XSL•FO
RenderX

themselves as a safety net and draw on their expertise to identify and eliminate hazardous scenarios that could cause harm, such as the sharing of unsafe or fraudulent medical advice and promotions of self-harm or suicide [9,37]. In addition to their safeguarding role, moderators can help establish a supportive, thriving community by creating a safe space for personal disclosures and fostering positive relations between members [40]. Indeed, moderated communities have been found to have higher engagement, perceived social support, and acceptability to members than unmoderated communities [41]. The impact of moderators within a community has also been noted in real-world settings where shifts in community moderation happen. For instance, following the departure of moderators from an established web-based health community, community members felt reluctant to engage with the community and recognized that the professional moderators had played a unique role in fairly and consistently implementing the rules and regulations [31]. Appropriate moderation is therefore often considered essential in laying the foundations of safety and facilitating the success of web-based communities. In the creation of CommonGround, the voices of lived experience experts were central to determining the Community Principles ("terms of use" community members agree to) and the associated governing policies to ensure that they remained aligned with their expectations of moderation and priorities for safety.

Much work has focused on how specific moderation issues are handled in well-established communities [9,30,31], yet there is limited guidance or recommendations on how to integrate moderation practices into a newly founded peer support community [37]. In addition, even though their experiences are inextricably intertwined, many studies consider the experiences of either the moderator [9,30] or the community member in isolation [31], and do not synthesize these perspectives to create a comprehensive understanding of what best practice might look like. Whilst moderators and community members participate with communities for different reasons, they coexist in the same space and their experiences are consistently shaped by the presence and behavior of one another. Moderators help establish the community culture and the boundaries for participation, while the behavior and responses of community members, in turn, influence what moderators do (eg, the content they have to moderate) and how they adapt their moderation responsively to suit the needs of the community. In this way, the experiences and behaviors of moderators and communities are continually shaping one another over time and cannot occur independently of one another. We therefore explored moderation experiences within the newly formed CommonGround community from the perspective of both community members and moderators. Our objective was to evaluate the feasibility and acceptability of the moderation procedures and inform recommendations for best practice in moderating newly forming web-based peer support communities.

Methods

Design

This work exists within the wider context of the CommonGround project, which aimed to coproduce an online peer support and psychoeducation platform and assess the feasibility and acceptability of the intervention in a randomized controlled trial (fRCT [feasibility randomized controlled trial] [42]; ClinicalTrials.gov: NCT06222346). The CommonGround project is underpinned by a pragmatic mixed methods research paradigm with community participation, which prioritizes what works, for whom, and in what context. The fRCT was delivered remotely and randomized 83 participants with long-term physical health conditions and subthreshold depression to the CommonGround intervention (ie, community members) and 42 participants to the control condition of National Health Service Mental Health webpages for 3 months (findings published in full elsewhere; in press). A group of 187 people with long-term physical health conditions who were interested in managing their mental health better, but did not have subthreshold depression, were also onboarded onto the CommonGround platform to increase the size of the community (community engagement cohort [29]). A nested qualitative process evaluation examined the experiences of community members with long-term physical health conditions and moderators. The qualitative process evaluation adopts an interpretivist-constructivist approach that assumes realities are socially constructed through interactions with others in the cultural context in which they are situated. This paper is reported within the SRQR (Standards for Reporting Qualitative Research) guidelines as per EQUATOR (Enhancing the Quality and Transparency of Health Research) recommendations (Multimedia Appendix 1 [43]).

Ethical Considerations

This study involves data from the nested qualitative study of part of a larger feasibility randomized control trial that was conducted remotely at King's College London (KCL). The trial and nested qualitative study received ethical approval from the South Central - Oxford C Research Ethics Committee (IRAS ID 328175). The protocol was preregistered (ClinicalTrials.gov: NCT06222346) [29] and outlined the larger feasibility randomized controlled trial and the nested qualitative study (ie, community member and control group interviews, moderator focus group). This project is sponsored and organized by King's College London. The Research Governance Office can be contacted via email at rgo@kcl.ac.uk.

The data from the nested qualitative study were collected in alignment with the original protocol aims of understanding participant perspectives of the intervention, control condition, and trial procedures generally (published elsewhere [29]) and for understanding perspectives and experiences of moderation specifically (this study).

Informed consent was obtained from all community members at two stages: (1) at T_1 when completing the 2 eligibility screenings, and if found to be eligible for the trial, (2) at T_2 when they consented to participate in the trial. All community members had provided consent to take part in the feasibility randomized controlled trial at enrolment. At enrolment, they were informed that they may be invited to interview after the 3-month intervention period, and that additional consent would be sought at that time if selected. Consent was recapped for

XSL•FO
RenderX

those who were invited to interview prior to their agreement to interview.

Community members and moderators were invited to be interviewed or take part in the focus group, respectively, after the intervention period. Five days after the invite, they were contacted again to discuss any questions about what taking part would involve before completing a web-based consent form (via Qualtrics) and scheduling the interview. Participants were allowed more time to consider their participation if needed and asked to contact the research team should they decide they wished to participate.

At the beginning of the interviews and focus group, the information about what the session would involve was repeated verbally, with participants reminded of their right to withdraw and to decline to answer any question, at any time. Community members were reimbursed £20 (US $26.93) for their participation, and moderators were reimbursed in alignment with their hourly wage.

The CommonGround Platform

CommonGround was codesigned in collaboration with people with lived experience of long-term conditions or physical-mental health comorbidities, clinicians, academics, and software developers [28]. The anonymous web-based platform was open to all diagnoses and could be accessed via desktop or mobile devices. The key features included a discussion forum with different posting options, reactions, and the ability to comment (peer support component), alongside a resources page of evidence-based self-help information (psychoeducation component). The self-help information consisted of 91 resources that were hand-selected by the research team from reliable, reputable sources and covered various topics ranging from keeping active, managing mental well-being, and managing finances. The resources were selected for their relevance to people living with any long-term physical health condition (ie, were not condition-specific). The CommonGround platform features, functions, and layout remained static throughout the intervention period with one minor exception. During the first week, the software developers made a minor edit to how the latest activity in the community forum appeared. Initially, at launch, the newest post was displayed at the top of the community forum feed. This was changed so that the post with the most recent activity displayed at the top of the feed (ie, existing post with the most recent comment or new post). Further details of the features of CommonGround and screenshots of the platform are available in Multimedia Appendix 2.

Moderation Policies and Practices

The CommonGround platform was governed by the Community Principles, Moderation Policy, Safeguarding Policy, and Privacy Policy, which were all published on the platform (and are available on the CommonGround projects Open Science Framework webpage [44]). These policies were coproduced in collaboration with our research advisory group (10 members living with a range of long-term health conditions, including 6 females and 4 males (self-reported sex), aged between 25 and 73 years), software developers, and the Samaritans Online Harms Team [45]. Key principles of the moderation approach included that the platform was anonymous, was not moderated 24/7, and did not provide one-to-one or crisis support. The moderation team also had access to a moderator handbook, developed in consultation with a leading disability charity [46], which included guidance on how to moderate different scenarios. The CommonGround platform had an admin backend that allowed moderators to edit and delete posts, delete comments, review flagged posts, and escalate posts for further action. Community members were able to view all policies and “flag” posts for review by moderators. Prior to the launch of the community, moderators were trained by the research team on how to use CommonGround and how to implement the governing policies. The moderator training was developed based on consultations with the community manager at SCOPE [47]. Each moderation session was expected to take approximately 60 minutes per day, and to occur at some time between the hours of 8 AM and 8 PM. Moderators were able to choose whether to moderate across a single or multiple sessions, and exactly when to moderate, dependent on the needs of the community (eg, when there was a moderation concern to monitor) and their personal schedules. Moderators were required to complete a handover after each moderation session of outstanding tasks, posts, or comments to watch, and any other comments. Moderators were instructed to review the previous moderation handover at the beginning of their session and to complete any outstanding tasks as a priority. Then moderators were instructed to review new content and complete any moderation tasks as needed (ie, in response to what had been posted). These moderation tasks may have involved any (or all) of the following: edit posts, delete posts, delete comments, review posts with attachments, review flagged posts, comment on posts to guide discussions toward a productive versus harmful narrative, and contact users to issue warnings. A Microsoft Teams chat was also created to support shared decision-making and provide updates to the moderation team. All moderators were of equal status and consulted with the research team (HGJ, GL, and MH) as required for additional support or to resolve any queries they had on the best course of action. For instance, if a moderator was unsure about whether to moderate or not (eg, whether to delete a comment, an action that could not be reversed), they were encouraged to discuss this with the team to come to a collective decision. If required, any decisions to issue warnings to community members (part of the “three strike” policy) or to remove community members from CommonGround would be made by joint decision of the moderators and research team. At any time, moderators could request an informal debrief with the research team, which included a liaison psychiatrist. A feedback session for moderators to share experiences was conducted halfway through the 3-month intervention period.

There was a separate engagement team whose exclusive focus was on cultivating engagement on CommonGround by sharing discussion topics to stimulate conversation and weekly circular emails with community updates and reminders to log in. The engagement team had no moderating or safeguarding responsibilities, but could use the “flag this post” button if they came across something they felt required attention (as a regular community member could also do). The engagement team consisted of our co-researcher with lived experience and 2 other research team members. The profiles of the Engagement Team

XSL•FO
RenderX

on the platform were clearly identifiable as being distinct from both regular community members and the Moderation Team via their avatar and account username. As the engagement team involved different team members to the moderation team, had unique roles and responsibilities, and were not involved in any moderation activity, the engagement team will not be discussed further in this manuscript.

Participants

CommonGround Community Members

CommonGround community members were recruited in the United Kingdom through multiple channels, including outpatient clinics within King's College Hospital and Guy's and St Thomas's Hospital, public advertisements via charities, and word-of-mouth. The inclusion/exclusion criteria for the feasibility trial are listed in Table 1 and were all self-reported by participants. As mentioned above, to increase the size of the community at the time that CommonGround was launched, people who did not have subthreshold depression (ie, scored 0-4 or 10-14 on 8-item Patient Health Questionnaire) but met all remaining inclusion criteria were also offered the opportunity to join CommonGround as part of a community cohort study (study protocol contains full details [29]). The only difference between participants of the feasibility trial and the community cohort was their 8-item Patient Health Questionnaire scores: feasibility trial participants scored 5-9, whereas community cohort scored either 0-4 or 10-14. Approximately 40 community members who had access to CommonGround (including ≈30 fRCT participants and ≈10 community engagement cohort) were invited to participate in an interview via email with a follow-up phone call. Community members were approached purposefully based on their long-term condition(s), age, gender, ethnicity, and their engagement level with the community. The engagement levels were defined as follows: (1) never joined (never registered their account or logged in), (2) never logged in (created an account but never logged in), (3) readers (logged in but never commented, reacted, or posted), (4) acknowledgers (logged in and commented or reacted, but did not post), and (5) content creators (logged in and created posts).

Table 1. Inclusion and exclusion criteria for the feasibility randomized controlled trial (fRCT) and community cohort.

	Inclusion criteria	Exclusion criteria
fRCT	• Currently living with one or more long-term conditions. A long-term condition was defined as a diagnosed physical medical condition with a duration of ≥6 months that demonstrates recurrence or deterioration and is typically associated with a poor prognosis [48]. There were no restrictions on the type of long-term condition. • Probable subthreshold depression, defined as scoring 5-9 on the PHQ-8[a]. • Age ≥18 years. • Have access to the internet (via phone, smart device, desktop, or laptop at any location, including at home, at the workplace, or free Wi-Fi in public spaces). • Have sufficient English proficiency to engage with the platform. • Ability to give informed consent (ie, no known circumstances/conditions that may affect the capacity to consent).	• Has ever received a clinical diagnosis of severe mental illness, including bipolar disorder, psychosis, posttraumatic stress disorder (PTSD), and schizophrenia. • Has ever received a clinical diagnosis of dementia.
Community cohort	• Currently living with one or more long-term conditions. A long-term condition was defined as a diagnosed physical medical condition with a duration of ≥6 months that demonstrates recurrence or deterioration and is typically associated with a poor prognosis [48]. There were no restrictions on the type of long-term condition. • Scoring 0-4 or 10-14 on the PHQ-8a. • Age ≥18 years. • Have access to the internet (via phone, smart device, desktop, or laptop at any location, including at home, at the workplace, or free Wi-Fi in public spaces). • Have sufficient English proficiency to engage with the platform. • Ability to give informed consent (ie, no known circumstances/conditions that may affect the capacity to consent).	• Has ever received a clinical diagnosis of severe mental illness, including bipolar disorder, psychosis, PTSD, and schizophrenia. • Has ever received a clinical diagnosis of dementia.

[a]PHQ-8: 8-item Patient Health Questionnaire.

Moderators

Moderators were recruited through KCL PhD cohorts and through professional connections of the research team, based on their clinical expertise in mental health treatment and experience of attending or moderating peer support groups. At the end of the trial period, all moderators (n=5) were invited via email to participate in a focus group.

Data Collection

Interviews

Demographic data were collected at baseline as part of the fRCT via anonymous Qualtrics questionnaires. Interviews were conducted after the 3-month intervention period (September 28, 2024, to December 8, 2024) by a member of the research team (CR, HGJ, GL) to explore individual experiences of taking part in the trial and of the CommonGround intervention. Topic

XSL•FO
RenderX

guides were coproduced with the research advisory group (Multimedia Appendix 3). The topic guides included the following open-ended questions around moderation: "What were your experiences, if any, of the Moderation Team on CommonGround?" and "What did you think about CommonGround being an anonymous platform?" All interviews lasted approximately 30-90 minutes and were conducted online via Microsoft Teams and audio-recorded, before being transcribed verbatim and anonymized for analysis.

Focus Group

Demographic data were collected via Qualtrics. A focus group was conducted (by a member of the research team, CR, who had no prior contact with the moderators during the trial) to stimulate debate and discussion around their experiences of moderating the platform, moderation policies and training, the positive and negative experiences they had observed community members have, and how moderation could be improved in the future. Moderators were given the opportunity to elaborate on any topic they felt was relevant that may not have been covered by the topic guides and were prompted to provide specific examples. The focus group was conducted via Microsoft Teams, audio-recorded, and transcribed verbatim and anonymized for analysis. The focus group lasted for approximately 90 minutes.

Data Analysis

Data from the community members (interviews) and moderators (focus group) were analyzed together using NVivo software (version 14, released 2023; Lumivero). An inductive thematic analysis was performed to sort the data into overarching themes that were developed inductively and from an interpretivist-constructivist position [49,50]. The first author (HGJ) familiarized herself with the entire dataset by reading and rereading the transcripts. Once immersed in the data, an initial list of codes was generated, which were constantly compared with one another to identify patterns of shared meaning that formed provisional themes. An experienced qualitative researcher (VL) reviewed the initial coding framework to help refine the codes and provide an additional perspective on interpreting the data. VL and HGJ also met regularly to discuss the analysis and explore different interpretations. The transcripts were then reviewed, and relevant extracts of the data were coded into NVivo nodes (the refined codes). Through an iterative process, codes were refined, added, and condensed as needed before being organized into themes. Following further discussions, a consensus was reached on the analysis, and the themes were named and defined. Throughout, our lived experience co-researcher (EA-F) ensured the coding framework was reflective of lived experience terminologies and definitions.

Results

Participant Characteristics

Community Members

A total of 39 participants who had been provided access to the CommonGround intervention were interviewed (29 fRCT participants, 10 Community Engagement participants), including 21 women, including transgender women, and 18 men, including transgender men, living with a wide variety of long-term health conditions (Table 2). Data saturation was reached after the 39 interviews, and no further interviews were conducted.

XSL•FO
RenderX

Table 2. Demographics of the community members interviewed.

Category	Overall (n=39)
Age (years), mean (SD)	53.2 (13.7)
Gender (n=39), n (%)	
Women, including transgender women	21 (53.8)
Men, including transgender men	18 (46.2)
Ethnicity, n (%)	
Asian or Asian British	—[a]
Black, Black British, Caribbean or African	—
Other ethnic group	—
Prefer not to say	—
White	31 (79.5)
Mixed or multiple ethnic groups	—
Engagement level with CommonGround (n=39), n (%)	
Never joined: never registered their account or logged in	6 (15.4)
Never logged in: created an account but never logged in	5 (12.8)
Readers: logged in but never comments, reacted, or posted	5 (12.8)
Acknowledgers: logged in and commented or reacted, but did not post	7 (17.9)
Content creators: logged in and created posts	16 (41.1)
Long-term physical health conditions: total, n	
Total number of unique LTCs[b] represented	31
Long-term physical health conditions: categories, n	
Autoimmune and inflammatory	11
Cancer	—
Cardiovascular	9
Gastrointestinal and digestive	18
Metabolic and endocrine	6
Musculoskeletal and chronic pain	9
Neurological (eg, Parkinsons)	13
Other	8
Reproductive and gynecological	—
Respiratory	—

[a]Represents <5 participants, removed to protect anonymity.

[b]LTC: long-term condition.

Moderators

The moderation team included 4 women and 1 man (self-reported gender), aged between 26 and 45 years. Three were clinical research fellows. One was a mental health nurse, one a pharmacist, one a clinical psychologist, and 2 were adult psychiatrists. Four had clinical experience in working with people with mental and physical health conditions, and one had experience working with people with mental health conditions only. Their years of clinical experience ranged from 6 to 19 years (mean 10, SD 5.1 years). All the moderators were currently completing PhDs.

Themes and Subthemes

Overview

Data from both groups of participants converged around five key themes: (1) moderation as a pillar of community safety, (2) the moderation blueprint, (3) the moderation threshold, (4) self-moderation is the (achievable) dream, and (5) being part of the moderation team. Table S1 in Multimedia Appendix 4 provides an overview of the themes and illustrative quotations.

Theme 1: Moderation as a Pillar of Community Safety

A widely shared opinion among moderators and community members was that moderation is essential: “moderation, it’s

XSL•FO
RenderX

just important. It's got to be there" (community member 6, acknowledger). In particular, moderation was thought to play a critical role in making the community feel safe; a foundational pillar needed for community members to share, interact, and, ultimately, provide peer support. Regardless of how aware members were of the specific moderation policies or day-to-day activities, the simple presence of the moderation team engendered a sense of safety that was considered unique to CommonGround.

> *It was being overseen or looked after, where, sometimes, on [well-established social media platform], there are some very good support groups, and there are some that are just absolute wild west, and there's people arguing and fighting on them, and they're not healthy places to be. So, there was a sense of safety.* [Community member 28, content creator]

The fact that the community was for people living with long-term physical health conditions augmented the need for moderation and safety. Community members wished to protect the space from people who did not have long-term conditions, as they felt that they were likely to share sensitive and personal information and experiences related to their health.

> *[Moderation is] really important, because we're all quite vulnerable. Those of us with chronic illnesses especially, we're all quite vulnerable, and of course, it comes under the whole Equalities Act 2011 and all that, with hate against disabled people being a hate crime, and all that kind of stuff.* [Community member 9, content creator]

The specific needs of this community also influenced moderator and community members' opinions on the key areas of moderation. For instance, they shared the view that potentially dangerous topics included the spread of medical misinformation and expressions of emotional crisis or self-harm that required follow-up.

> *There would be some things where you'd want to be on it really quickly, like if someone's posting about self-harm methods or how they want to end their life. Something like that, you don't want to wait.* [Moderator 3]

While it was felt that the community was a place to discuss medical and emotional experiences, there was shared concern that these topics require careful review by moderators to ensure that they are discussed in a safe way.

> *I think for a platform offering peer support, there needs to be some moderation to just make sure that, A, you don't go off down wormholes, that it's not taken over by individuals, the information that's posted is accurate, and that any conflict is dealt with quickly and effectively.* [Community member 16, acknowledger]

Theme 2: The Moderation Blueprint

Overview

This theme relates to the policies, community principles, moderator guidance, and training that underpinned the governance of CommonGround. Two subthemes emerged: (1) the importance of this blueprint and (2) the moderator's experience of translating the blueprint into practice.

Subtheme 2a: The Moderation Blueprint Is Essential

In general, community members were satisfied with the moderation and safeguarding policies that were on the platform, recognizing that they were critical for transparently outlining what is and is not acceptable and what to expect in different circumstances. They also provided reassurance that there was a safety net for worst-case scenarios. Typically, community members briefly reviewed the policies at sign-up, often indicating that their review was a relatively light touch and purely to ensure that there were no 'red flags' in the policy. Others made assumptions about the type of content based on their previous experiences or because they considered the owners of the platform (KCL) to be trustworthy.

> *So, the moderation rules are there, and they should be abided by, and people know if they don't, they're parked, and that's how it should be.* [Community member 11, content creator]

The moderators also felt that the comprehensive policies laid out clear, consistent boundaries for how the platform would be moderated. The moderator handbook, which provided detailed guidance on how to implement the policies, was valued for its specific examples of potential events, as this reduced some of the burden in working out how to respond, including the specific language and tone to use. The moderator handovers were also valuable for keeping up to date on recent activity and highlighting points to review in their moderation shift.

> *The guidance that we had was quite helpful in that it had quite specific examples for a lot of those scenarios that might come up. And then an example of how you might respond. I found that was quite helpful. And then I guess, as we started doing more of it, you could also see what other people had posted to similar things from the handover document. And then that would, I guess, we'd all respond in a similar way a lot of the time.* [Moderator 5]

One aspect of the moderation blueprint that split opinions among community members related to the anonymity of CommonGround, with mixed views on whether this constituted a necessary safety feature or stifled members' ability to establish genuine connections. Other community members felt indifferent, or that the best option would be allow members to select their level of anonymity and give individuals the power to disclose their personal information as they see fit. This call for greater flexibility in anonymity was also recognized by the moderators, who suggested that the moderation blueprint could, and should, be reviewed as the community evolves.

> *I think maybe going forward is probably about giving people a choice if they want to stay anonymous or not* [Community member 44, reader]

XSL•FO
RenderX

Subtheme 2b: Translating the Moderation Blueprint Into Practice

In general, the moderators agreed that their moderation shifts were quick, manageable, and quite light touch with minimal need to intervene, edit, or delete posts:

> *It was quite a quick process. Had a checkbox of, check this, check this, check this. And most things had nothing, you know, the flag post, delete posts. Most of them, you're just checking that there was nothing there.* [Moderator 1]

The moderators found the guidance in the moderation handbook useful and applicable to the content that came up, and were confident in applying it (Multimedia Appendix 5 provides an illustration of how comments posted by moderators mapped onto advice in the moderation handbook). It was clear that for moderators to successfully do their job, they needed to feel that it was easy to translate the moderation blueprint into the real-world setting of the web-based community. In general, the moderators felt that this translation was easy to do. The main challenge that moderators faced related to the ease and efficiency with which they identified what content was new since the last moderation session. As noted in the methods, during the trial itself, the software developers updated the forum feed layout to display the most recent activity (comment or post) at the top of the feed. The moderation team shared that this technical update helped to reduce the burden of identifying the new content during the trial and that it could be further developed to improve efficiency in a larger, future community. This initial frustration that moderators expressed at the platform-specific issue provided insight into the fact that if their basic need of easy navigation around the community content is unsatisfied, any form of moderation becomes difficult.

> *Initially, it was more of a system issue. So, we would have to look at comments and threads within the last 24 hours. But initially, the messages were coming up when they were posted, so we wouldn't see the new things immediately. So, having to monitor and try to look at what was new, was really difficult.* [Moderator 5]

Theme 3: The Moderation Threshold

A key experience that moderators spoke about related to navigating the moderation threshold: when to intervene and when to "wait and see" how the situation played out. As time went on, the moderators felt more confident in where the moderation threshold sat, as they learnt how the community worked and reflected on previous moderator activity.

> *I think when we started to notice that certain moderation styles would actually stifle the engagement from the community, it became quite clear that if you come back to what is the purpose of the platform, that then helps you to make a decision about whether you should do something or not. I guess, like a risk, reward thought process.* [Moderator 1]

Moderators and community members shared the view that navigating this threshold was particularly difficult around medical information.

> *[…] By giving space to talk about it, but also disclaiming that that's not a good source, and here's a better source, and having a more balanced opinion, contemplating what they're saying, but also say… It is a fine line, but by acknowledging what they're saying, whilst also posting the counterargument in, it allows people to see.* [Community member 6, acknowledger]

While the moderators considered their clinical experience useful overall, at times, their experience of dealing with clinical populations and fear of "worst case" scenarios could feel unhelpful. Some spoke of the need to recalibrate their threshold for stepping in to suit the current audience. A couple suggested that including people with lived experience of chronic illness in the moderation team could be beneficial, as they may be better placed to judge when an expression of emotion constitutes emotional distress that requires signposting to one-to-one support.

> *I wondered sometimes if our professional backgrounds made us more anxious about some of the comments that people made on the platform. As [moderator] said, sometimes you'd get such a small amount of information indicating that somebody is struggling with something. Then because, I guess, my professional background as a clinical psychologist, you see the real peak of distress in terms of the clients that you might work with, particularly when you're working in a physical health context. So then, I suppose, my mind might go to, this person's really distressed. And then you start thinking of all the people that you've worked with clinically, who've been distressed and whether you might need to intervene, and then in terms of managing risk. Whereas actually, it might be that the likelihood is that person's distressed, but it's a really a much lower level of distress than what I might see professionally, that's therefore, less in need of some intervention from the moderators.* [Moderator 1]

Moderators and community members shared the concern that intervening too early could stifle the community and prevent people from sharing freely. Moderators observed that when they commented, discussion among peers often stopped, leading to a hesitancy to intervene "too early" or unnecessarily. Community members suggested that over-moderation could create a strange dynamic in the community and stagnate conversation. It was also felt to have potential to propel some community members toward "rebellion," creating further issues.

> *I think stifling conversations just makes people more likely to try… Well, not even to try and cause trouble, but that's the end result, but it makes them more likely to have extreme views and to go further down the rabbit hole.* [Community member 6, acknowledger]

Despite the moderators' anxiety about finding the right threshold, community members often felt that the moderators

XSL•FO
RenderX

had struck the right balance. Some suggested that a more authoritarian dynamic may be needed, or necessary, at these earlier stages of community formation to ensure that ground rules are understood and the community grows in a safe manner.

> *I also felt that there was that kind of safeguarding element with the moderators just being in the background, but they weren't, it wasn't like a very controlling element to it. And they almost showed up as equal participants, almost, but just with that, oh, just a reminder of… It didn't feel like an enforcement group or anything.* [Community member 27, content creator]

Theme 4: Self-Moderation Is the (Achievable) Dream

This theme is related to the shared view that self-moderation was the ideal and the hope that this would evolve in the community over time. Self-moderation would include community members jumping in to moderate potentially problematic conversations and preventing any arguments from emerging. Community members and moderators observed that some members of the community not only regulated what they shared on CommonGround but also intervened in situations where people seemed to be entering potential conflict, share incorrect information, or be in need of emotional support. In interviews, community members commented on their responsibility to think before they post or jump into discussions.

> *We found after a while, that other users were self-moderating anyway, so we really didn't have to intervene that much, because by the end of the thread, it would have resolved.* [Moderator 2]

One feature that was built into the platform to support self-moderation was the "flag this post" function that prompted members to determine what they thought was inappropriate and flag it to the attention of the moderators for priority review. This feature was relatively unused by the community, but moderators recognized it as important infrastructure to allow the community to move toward self-moderation if there were to be a larger volume of member-generated content:

> *[...] One of the functions that was on the platform that wasn't used very much, was users at the platform actually flagging posts that they thought were inappropriate or distressing. That's one way that if it [CommonGround] got bigger and someone wasn't looking at every single post, things would still get identified and then would help to reinforce that ownership of the platform. Because then it's like, what do they [users] find unacceptable? What do they think is going against the principles? […] I think that would be good, maybe to see more of that.* [Moderator 5]

The opinion that self-moderation would be a superior model of moderation was in part driven by the moderators' view that hearing from peers is much more powerful than hearing from a moderator.

> *I felt like the self-moderation was much more powerful in terms of influencing and changing opinion or getting people to think about things in a different way, because it came from a peer.* [Moderator 3]

However, handing the baton of responsibility over to the community members required careful navigation as moderators balanced their responsibility to keep the community safe and the desire to promote members to have ownership over the community:

> *[…] The learning from trial and error. Okay, like, this is something that we were concerned about. We were thinking that possibly it could go down that avenue of concern. And actually, what we saw was someone else commented. That ended up fine, we didn't need to do anything. So you learn from the experience, which I think helped us to feel more confident in leaving things be, a bit more, as well.* [Moderator 5]

Both community members and moderators recognized that a high level of engagement and strong bonds between members were necessary for the community to self-moderate. However, it was recognized that the CommonGround community did not seem to reach this, with more time needed for it to become established and flourish.

> *I think it's quite a short amount of time to form a community, especially online. And I think maybe there was the start of some of that. At times, certainly. Like what [moderator] was referring to earlier, when there were quite, very supportive responses to people who were clearly distressed and sharing of similar experiences. There were moments where it felt like there was good community support.* [Moderator 5]

One moderator noted that the inclusion of lived experience experts as part of the moderation team might help to encourage this shift toward self-moderation. When speaking about the future of moderation on the platform, community members noted that the moderation team would likely still be needed in the background. Some community members explicitly identified a continued responsibility of the professional moderators in ensuring platform safety, highlighting that certain scenarios demand their intervention:

> *Obviously, you're relying on a moderator to block trolls, or whatever you call them.* [Community member 39, never logged in]

Theme 5: Being Part of the Moderation Team

This theme relates to the positive consequences moderators experienced as a result of being part of a team. The moderators described a sense of shared endeavor in moderating the CommonGround community and had a strong understanding of what their role as a team was, and how their role as a moderator (to keep people safe) was distinct from the engagement team (to promote peer support discussions). Being part of a team helped the moderators to balance the position alongside their other commitments and allow downtime away from the community.

> *I think we had the right number of people as well. It wasn't that you did it all the time. You did it a couple of times a week at most, which I think was about right. It'd be hard to do it every day, I think.* [Moderator 4]

XSL•FO
RenderX

The benefit of being part of a moderation team went beyond the practical aspects, allowing for emotional support and encouragement, and opportunities to debrief after tricky situations. One moderator shared that "[…] It felt very open door, in terms of people [the research team and other moderators] being approachable. If we had any questions, it felt very safe and supported" (Moderator 3). In some way, the moderation team's use of Microsoft Teams chat and handover documents became a form of peer support, with moderators able to seek advice and share how they felt about the role or the community at any time. One moderator recalled feeling well supported by fellow moderators and the research team after they received some negative feedback from a community member whose post they moderated:

> *[…] Then they wrote another one [post] saying, […] and was a bit shirty. But it was nice because I had a chat with the KCL research team as well, and they said, yes, you did the right thing, which was quite reassuring.* [Moderator 4]

Being part of a team also helped to build the moderators' confidence in the role, as they could learn from one another as they implemented moderation policies in a new community. Shared decision-making and reflection helped the team move toward a more efficient method of moderation.

> *[…] It was quite helpful to have other people dipping in, in between your moderation sessions, because then you get a different perspective on what's already come up. And if there are things that are on that boundary of, is this something we need to do something about or not. Then I found it quite helpful to have someone else's eyes on that. Just my own, I guess, reassurance that we're doing the right thing.* [Moderator 1]

The range of expertise within the moderator team was also beneficial, particularly when it came to certain topics that moderators were most concerned about, such as diagnosis-related information or medication. Here, it was understood that it was important for the community to be able to speak about it in a safe way.

> *I do think it was really helpful having some medically trained moderators amongst the group who could be a bit more aware of the comments about medications and treatments and things. Because that was something that I didn't really feel able to moderate that well, because I don't know very much about the medications and the medical treatments for these different health problems.* [Moderator 1]

Future Improvements for Moderation

As moderators reflected on their experiences of moderating the CommonGround platform, they highlighted areas where moderation could be improved in the future. Areas for potential improvements and evidencing quotes are presented in Table 3 below.

XSL•FO
RenderX

Table 3. Potential areas where the CommonGround moderation could be improved, with corresponding evidencing quotations.

Potential improvement to moderation	Quotes
General moderation improvements	
Include lived experience experts, or clinicians with relevant lived experience, on the moderation team	• I think it could have been quite good to have lived experience moderators as well, who would maybe... And maybe that would make a difference to the power dynamic a little bit, in terms of when there were moderator comments, it might be perceived differently? [Moderator 3] • I think having people who aren't healthcare professionals, who are more coming from the perspective of that lived experience might help to get a bit more balance there, possibly, in terms of, I guess, the psychological distress part of it. [Moderator 1] • If anything, I think getting more community ownership over the platform would be a positive, and maybe having lived experience moderators. But still, the community principles are great, and I think those things are really important. But I think to encourage more community ownership of the platform would help make it a more active platform. [Moderator 2]
Explore whether AI[a] or algorithms could aid moderation	• There are lots of massive platforms online that manage to moderate huge amounts of posts. And I guess the way they do that is using algorithms. And they must, because there's no way someone's looking at everything. So, if you scale up, that might be one way to do it. [Moderator 5]
Encourage the community towards self-moderation	• ...One of the functions that was on the platform that wasn't used very much, was users at the platform actually flagging posts that they thought were inappropriate or distressing. That's one way that if it [CommonGround] got bigger and someone wasn't looking at every single post, things would still get identified and then would help to reinforce that ownership of the platform. Because then it's like, what do they [users] find unacceptable? What do they think is going against the principles? […] I think that would be good, maybe to see more of that. [Moderator 4]
Platform-specific improvements	
Improve the feature(s) for identifying new content on the platform since the last moderation session	• I think the efficiency of the moderation could have been way better if you're flagged to anything new that's come up within 24 hours. […] I think it could be improved by having some way that the platform can flag all of these things are new in the last 24 hours. And then you just go through them, almost like a notification system. [Moderator 1] • […] You can almost just have all of them and then you tick them [new comments/posts] off, and you know that they've definitely been moderated. That would be the most efficient way, probably. [Moderator 5] • I think with the addition of the notification system, making it super, super clear, what's new, would help it to be moderated. It would help the moderator, if there were more users. I don't feel like it's necessary for the moderator to see everything, because so much of it is just really positive and helpful and supportive, and day to day chats about books and movies and things like that that don't really need a moderator. [Moderator 1]
Review the community principle of anonymity	• It felt to me that there were a few people who wanted it to be less anonymous. And what "moderator" said earlier, about somebody who wasn't very happy about their personal identifying information being taken out of their posts. I wonder if creating that sense of community for some people is easier, if you are able to share a little bit more information about who you are and what part of the country you live in, and things like that. [Moderator 1]

[a]AI: artificial intelligence.

Discussion

Principal Findings

Online peer support promises a low-cost, easily accessible, and scalable way to support mental well-being and self-management among people living with long-term physical health conditions [1,24]. The risks and potential adverse events that can occur when engaging with peers via web-based communities are well recognized [9,27,31-34]; yet, research into how best to moderate communities to keep them safe is limited, particularly in the context of new web-based health communities [37]. In this work, we explored the perspectives and experiences of both community members and moderators together to evaluate the acceptability and feasibility of our co-designed moderation procedures and identify areas for future improvements. Overall, community members and moderators both had positive views of our moderation procedures. Moderators found it feasible to implement the polices but identified opportunities to improve moderation as the community culture evolves. Based on these findings, we provide recommendations to guide the development and implementation of moderation practices for new communities (Table 4).

XSL•FO
RenderX

Table 4. Recommendations on how to moderate web-based peer support communities.

Recommendation	Description
1. Coproduce clear moderation policies that reflect best practice	• Develop moderation in collaboration with people who have lived experience to ensure alignment with the values of the target end user. Clearly define acceptable and unacceptable behaviors and identify key topics of moderation concern that should be prioritized. Be transparent in why, when, and how any moderation will happen. Lived experience experts should define what is acceptable and appropriate around topics that are important to them, such as crisis support and anonymity. The tone and language that will be used should be agreed upon enhance the acceptability of moderation practices.
2. Establish clear escalation and referral pathways	• Clearly define what support is available for community members in crisis. Ensure moderators are trained in handling high-risk situations in accordance with the escalation and referral pathways. Consider the pathways to be taken when handling complex or ambiguous cases.
3. Determine how the moderators "speak" and be consistent	• Tone and language can dramatically affect how moderation is received in the community, with inappropriate tone or language stifling conversation or creating a "big brother" culture. Consistency is key for moderation to perceived as fair and reduces the risk of members developing a specific attachment to individual moderators. Sharing template phrases and responses to common situations can help reduce moderation burden. In health-related settings, tone and language is particularly important in relation to moderating medical topics and emotional distress.
4. Communicate moderation and moderation decisions transparently	• Share moderation policies transparently in easily accessible formats. Provide clear and respectful feedback to members who are moderated, referring to the relevant sections of the policies. This helps members understand boundaries, reduce hostility toward moderators and minimize the likelihood of future moderation issues.
5. Create features that encourage self-moderation among the community	• Having a community that keeps itself safe and moderates itself (eg, encouraging peers to seek support from their general practice doctor if medical advice is being discussed) with minimal intervention from the moderation team is the ideal. When creating the interface of the online peer community, co-create features that encourage self-moderation, such as "flag this post".
6. Be responsive and evolve moderation processes as the community evolves	• Moderation should evolve with the community. Regularly review what is working well, what is burdensome, and how moderation is impacting members and moderators. Early on, there is a greater need to establish clear boundaries. Over time, community members will grow ownership over the community, and moderator behavior should shift to nurture and accommodate self-moderation. • Adjust the features of the interface and moderation practices as needed. Ensure to seek the perspectives of both community members and moderators to ensure a comprehensive understanding of what is working well and what needs adapting. For example, the moderation team told us that they found it challenging to track new content since their last session. In response, we improved the visibility of new content with new timestamps and post ordering in the community forum. • Guidance on gathering feedback: Moderators are encouraged to reflect on how different moderation approaches and practices are being perceived by the community by reflecting on the community engagement/comments that follow any intervention by moderators. By considering what was done and how the community reacted to the moderation behavior, moderators can adapt their approach to improve how it is delivered and received. Additionally, by reflecting on commonalities among posts that users "flag" to the attention for the moderations, key areas that need further development or greater moderation attention can be detected. With the emergence of AI[a] and algorithms that can monitor community activity, in the future, there could be scope to explore how AI could assist moderation and help identify where moderation is working well or needs adjusting. Whilst further work is needed to explore the usefulness and appropriateness of AI-assisted moderation, this emerging area could offer exciting opportunities for streamlining and facilitating community moderation.
7. Ensure moderators feel supported	• Have ways for the moderators to connect outside of the peer community, for instance, through a Microsoft Teams chat. Creating a space for moderators to discuss situations and come to a collective decision, decompress from emotionally-taxing situations, and share learnings or updates can be beneficial. Provide regular training specifically related to the community's policies, particularly when these change over time.

[a]AI: artificial intelligence.

To our knowledge, this is the first study to integrate the experiences of community members and moderators to understand moderation in a newly launched peer support community. It was universally agreed that moderation is a critical pillar of community safety, reflecting the perspectives of members [31] and moderators [9] of other established communities that moderation is essential for creating a safe space and mitigating risks. As in other web-based health communities, we also found that the community members and moderators had particular concerns about the need to moderate medical misinformation [9,31,37]. This highlights the importance of tailoring moderation practice to the unique aims and sensitivities of the specific community, and the need to identify these through collaboration with the target audience

XSL•FO
RenderX

(Recommendation 1, Table 3). While moderators reported a few instances where they were required to enforce the moderation policies, they felt that the moderation blueprint with the clear escalation procedures was appropriate and applicable to the community. The minimal need for moderator intervention may reflect the relatively low activity levels on CommonGround, where fewer user-generated posts naturally reduced the need for moderation. Equally, it may be because members understood and were happy to abide by the moderation policies [30]. Community members found "no red flags" in the moderation policies, suggesting that these were aligned with their expectations and moderation needs.

The apparent success of moderation also suggests that co-designing moderation practices is critical in ensuring their acceptability among the community, and that having clear boundaries and explicit guidance for moderators can lay the foundations for successful moderation from "Day One" of the community (Recommendations 1-3, Table 4). Community members valued being able to easily access the moderation policies when needed, underlining the need to share moderation practices transparently, which can enhance legitimacy, perceived consistency, and accountability of moderation ([51]; Recommendation 4, Table 4). Improving the communication of reactive moderation decisions may help to increase understanding, promote adherence to norms, and improve future behavior [51,52]. Together, these findings underscore the importance of having clear, comprehensive moderation practices tailored to the community, including the specific content of policies and the transparency, tone, and language of communications. Where the capacity to codesign moderation policies and practices is limited, examples from existing communities can be drawn upon, particularly as key moderation practices and experiences are likely common to any online peer community, such as the use of peer moderators [30].

By integrating the perspectives of community members and moderators, we gained unique insight into areas of moderation that may require review and adaptation as the newly formed CommonGround community evolves. The anonymous nature of web-based communities is often seen as a double-edged sword: anonymity can facilitate stigma-free and honest posting for some [53], whilst others feel that anonymity hinders the ability to form the meaningful and enduring social connections that are the cornerstone of peer support [54,55]. We found similar mixed opinions among our community members. The desire for reduced anonymity may grow as community members become more willing to share personal information if it makes accessing social support from peers easier, even if this could introduce other risks [56,57]. The reputable KCL ownership and branding positively influenced perceptions of safety. However, as some members appreciated the anonymity or acknowledged its value for others, offering members the ability to choose their preferred level of anonymity may be appropriate. Future work should explore with lived experience experts how to implement features that allow personal control over anonymity to ensure that anonymity can serve as a facilitator to engagement for those who desire it, while also removing this potential barrier to social connection for others [58,59]. These divergent views on a core aspect of CommonGround's moderation practice highlight the importance of reflecting on moderation practices by listening to the community and adapting as necessary (Recommendation 6, Table 4).

The moderation practice and style needed for CommonGround as a new community were recognized to be different from those of larger, well-established communities. In well-established communities, community members can be seen to be actively self-moderating by detecting medical misinformation, questioning information shared, reminding peers not to give medical advice, and reporting trolls to moderators [31,37,60]. While CommonGround moderators and community members agreed that peer-to-peer moderation could be more powerful and feel less punitive, they recognized that CommonGround could not self-moderate yet. The limited volume of user-generated content and the weak sense of community and ownership among members is likely to have hindered the emergence of self-moderation. The shift toward self-moderation must occur organically, as the premature withdrawal of moderators could harm community cohesion and reduce perceptions of safety if members are not comfortable acquiring moderation responsibilities [31]. Indeed, as prior work often cites "super users" (ie, those who engage with the community often, and are often observed to be generating larger volumes of content than the average user) in contributing to how this culture evolves, who often only emerge as a community evolves and may also facilitate the shift toward self-moderation [61]. As voiced in our research and prior work [31], there may always be a need for professional moderators to deal with high-risk or serious issues, even if the community takes ownership over other moderation activities. Although used infrequently on CommonGround, the moderators valued features like "flag this post" as critical tools for encouraging a community towards self-moderation. Indeed, moderators of other web-based health communities have urged members to use such self-moderation tools more to help shape moderation norms by defining what content they deem unacceptable [30]. Therefore, while the community culture needed for self-moderation cannot be artificially created, the features and functions that support self-moderation can be implemented early (Recommendation 5, Table 4). Future work should evaluate member and moderator perspectives as a community evolves through different stages of self-moderation to determine what additional practices may help support a successful transition.

Few studies have explored the experiences of moderators, but some key challenges consistently emerge [37]. First, a common challenge lies in efficiently identifying new content to reduce the risk of unmoderated posts escalating into worst-case scenarios [9,37]. Moderators were also mindful of how this challenge could intensify as the community grows, underscoring the need to develop and test novel tools that enhance moderation efficiency as CommonGround is scaled. Another challenge that our moderators faced was navigating the moderation threshold where they must evaluate whether to take action [9,30,51]. The views of the moderators echo those of moderators in other communities, who want to prioritize safety but are concerned that premature or overbearing moderation can stifle conversation and potentially damage community culture [62]. Moderators observed that once they commented, the community ceased to

XSL•FO
RenderX

comment further. This pattern has been observed in other communities [62], with moderators interpreting this to reflect disengagement from the community. However, members typically felt that the moderators had struck the right balance, suggesting that the quiet comment section may instead reflect the generally low engagement we had on CommonGround, where most members were "readers" preferring to observe posts rather than actively create content. This disconnect between how moderators believe their actions are perceived and how the community actually experiences them highlights the importance of considering both perspectives, particularly if informing changes to moderation practices (Recommendation 6, Table 4).

The benefits that the CommonGround moderators experienced as a result of operating as a team have been documented among other moderation teams. For instance, the CommonGround moderators shared how it became easier to navigate the moderation threshold over time and as a result of seeking support from their team, as they were able to exchange experiences to align their moderation practice, seek emotional support, and access the expertise of other members for second opinions on complex situations [63]. This adds to the literature highlighting how the moderation team becomes a form of peer support in itself that can help improve moderation efficiency, consistency, and reduce the moderation burden [9,64]. This highlights how it is important to consider how best to support moderators whilst they are supporting the community and give them a dedicated space to connect and support one another (Recommendation 7, Table 4).

Limitations

There are some limitations that should be considered for the interpretation and transferability of our findings and recommendations. First, this study explored the perspectives of community members and moderators who participated in the new community that was live for only 3 months. Many community members shared that they did not engage with CommonGround as they would have liked to due to personal circumstances or having limited time available during the 3-month trial. While lower levels of engagement and forum activity could be expected during early stages of community development, the experiences shared in this work are contextualized within a small community with relatively low engagement and sense of community in a research setting. As such, our findings and recommendations may not apply to communities launched with larger populations or with high volumes of user-generated content from the outset. Second, our findings should be interpreted relative to the nature of our moderators: professional, paid moderators with clinical expertise who were trained specifically about the moderation practices for CommonGround. Prior research suggests that professional moderators are often perceived as more trustworthy and credible than informal moderators, particularly around their moderation of medical topics [31]. Indeed, our moderators were aware of how their clinical expertise influenced their approach to moderation, and therefore, our findings may not generalize to moderation teams of volunteers, peers, or those without clinical or moderation backgrounds. Third, there is a potential for selection bias in the community members' interviews, as those who consented to participate may have more positive views of web-based peer forums or the moderation and governance of CommonGround. However, we did purposefully sample community members based on their engagement levels, probing those who had never logged in or only logged in a few times for their underlying reasons, motivations for low (or no) engagement, to unpack whether moderation or governance played a role in their disengagement. Fourth, although community members were anonymous to one another, they were aware that the moderators could link their pseudonyms to their personal details if required for moderation purposes. This awareness that they could be identified in exceptional circumstances may have influenced what members posted, meaning that the CommonGround environment may not directly reflect non-research anonymous communities.

All moderators and community members needed internet access to participate in the web-based community and interviews or focus groups. While 96% of households in the United Kingdom have internet access [65], for those without or who have limited access, their ability to engage with the community and interviews may have been limited. Additionally, some participants may have experienced barriers in accessing CommonGround privately, which may have had a downstream impact on the content of the community forum, and therefore the moderation needed [66]. Future work should consider how digital exclusion and privacy influence the nature of web-based communities, in particular their moderation.

Conclusion

Our work adds to our understanding of the critical role moderation has in shaping safety in online peer support communities designed to support mental well-being, and the importance of creating appropriate and trustworthy governance that is specifically tailored to the needs of the community members. By exploring perspectives on moderation in a new community and by integrating the experiences of both community members and moderators, we have gained unique insights into how to enhance our codesigned moderation practices going forward. We have provided a transparent account of the creation, implementation, and reflection on our moderation policies and governance, and recommendations on how to approach the moderation of new communities in the future.

Acknowledgments

We would like to acknowledge our software developers at TOAD and BitJam, who were instrumental in helping to co-design the CommonGround platform and delivering the intervention to participants. We would also like to acknowledge the Online Harms Team at the Samaritans, who kindly reviewed our moderation and safeguarding policies and procedures. And finally, thank you to all the members of our public patient involvement groups and research advisory group who contributed at various

XSL•FO
RenderX

stages throughout the development and delivery of this trial. CommonGround is a co-produced intervention, and this research would not have been possible without their dedication, support, and insightful contributions.

Funding

This project has been funded by King's Health Partners and Guy's and St Thomas Charity from the Multiple Long Term Conditions Challenge Fund (EIC180702 [MH]). The funders had no role in study design, data collection and analysis, decision to publish, or manuscript preparation. The research was part-funded by the National Institute for Health and Care Research (NIHR) Maudsley Biomedical Research Centre (BRC). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Authors' Contributions

All authors listed meet the International Committee of Medical Journal Editors criteria for authorship. MH, AS, VL, and GL conceived the study. All authors co-designed the study. HGJ, CR, and GL collected qualitative data. HGJ completed the qualitative analysis under the supervision of VL and AS, who are experienced qualitative researchers in the areas of mental health and chronic illness. HGJ prepared the manuscript. All authors have read and approved the final manuscript. All the research team members directly involved in data collection and analysis adopt an interpretivist-constructivist approach for this work. Throughout the analysis, collaboration through discussions ensured that the analysis remained grounded in the lived experience and voices of the participants and was not subject to any biases the team may have from their experience in research exploring the interplay between physical and mental health. E-AF has lived experience of multiple long-term conditions, and ensured that this work remained rooted in the lived experience of those with chronic conditions.

Conflicts of Interest

None declared.

Multimedia Appendix 1
SRQR checklist.
[PDF File (Adobe PDF File), 86 KB - jmir_v28i1e81943_app1.pdf]

Multimedia Appendix 2
Screenshots of the final prototype of CommonGround.
[DOCX File , 2293 KB - jmir_v28i1e81943_app2.docx]

Multimedia Appendix 3
Interview and focus group topic guides.
[DOCX File , 36 KB - jmir_v28i1e81943_app3.docx]

Multimedia Appendix 4
Theme descriptions and further evidencing quotes.
[DOCX File , 32 KB - jmir_v28i1e81943_app4.docx]

Multimedia Appendix 5
Comparison of moderator comments on CommonGround to moderation guidance.
[DOCX File , 28 KB - jmir_v28i1e81943_app5.docx]

References

1. Hossain SN, Jaglal SB, Shepherd J, Perrier L, Tomasone JR, Sweet SN, et al. Web-based peer support interventions for adults living with chronic conditions: scoping review. JMIR Rehabil Assist Technol 2021;8(2):e14321 [FREE Full text] [doi: 10.2196/14321] [Medline: 34032572]
2. Fortuna KL, Naslund JA, LaCroix JM, Bianco CL, Brooks JM, Zisman-Ilani Y, et al. Digital peer support mental health interventions for people with a lived experience of a serious mental illness: systematic review. JMIR Ment Health 2020;7(4):e16460 [FREE Full text] [doi: 10.2196/16460] [Medline: 32243256]
3. Wood S, Finnis A, Khan H. At the Heart of Health: Realising the Value of People and Communities. London: Health Foundation and Nesta; 2016:15.
4. Wright KB. Communication in health-related online social support groups/communities: a review of research on predictors of participation, applications of social support theory, and health outcomes. Rev Commun Res 2016;4:65-87 [FREE Full text] [doi: 10.12840/issn.2255-4165.2016.04.01.010] [Medline: 2255-4165]

XSL•FO
RenderX

5. Reidy C, Doble E, Robson A, Kar P. Peer support in chronic health conditions. BMJ 2024 Jul 01;386:e070443. [doi: 10.1136/bmj-2022-070443] [Medline: 38950946]
6. Malik S, Coulson NS. The therapeutic potential of the internet: exploring self-help processes in an internet forum for young people with inflammatory bowel disease. Gastroenterol Nurs 2011;34(6):439-448. [doi: 10.1097/SGA.0b013e318237a9ba] [Medline: 22129797]
7. Druss BG, Singh M, von Esenwein SA, Glick GE, Tapscott S, Tucker SJ, et al. Peer-led self-management of general medical conditions for patients with serious mental illnesses: a randomized trial. Psychiatr Serv 2018;69(5):529-535 [FREE Full text] [doi: 10.1176/appi.ps.201700352] [Medline: 29385952]
8. Cooper RE, Saunders KRK, Greenburgh A, Shah P, Appleton R, Machin K, et al. The effectiveness, implementation, and experiences of peer support approaches for mental health: a systematic umbrella review. BMC Med 2024;22(1):72 [FREE Full text] [doi: 10.1186/s12916-024-03260-y] [Medline: 38418998]
9. Deng D, Rogers T, Naslund JA. The role of moderators in facilitating and encouraging peer-to-peer support in an online mental health community: a qualitative exploratory study. J Technol Behav Sci 2023;8(2):128-139 [FREE Full text] [doi: 10.1007/s41347-023-00302-9] [Medline: 36810998]
10. Tran VT, Barnes C, Montori VM, Falissard B, Ravaud P. Taxonomy of the burden of treatment: a multi-country web-based qualitative study of patients with chronic conditions. BMC Med 2015;13:115 [FREE Full text] [doi: 10.1186/s12916-015-0356-x] [Medline: 25971838]
11. Demain S, Gonçalves AC, Areia C, Oliveira R, Marcos AJ, Marques A, et al. Living with, managing and minimising treatment burden in long term conditions: a systematic review of qualitative research. PLoS One 2015;10(5):e0125457 [FREE Full text] [doi: 10.1371/journal.pone.0125457] [Medline: 26024379]
12. Riegel B, Jaarsma T, Lee CS, Strömberg A. Integrating symptoms into the middle-range theory of self-care of chronic illness. ANS Adv Nurs Sci 2019;42(3):206-215 [FREE Full text] [doi: 10.1097/ANS.0000000000000237] [Medline: 30475237]
13. Long term conditions compendium of information. GOV.UK.: Department of Health; 2012. URL: https://assets.publishing.service.gov.uk/media/5a7c638340f0b62aff6c154e/dh_134486.pdf [accessed 2025-08-08]
14. Gold SM, Köhler-Forsberg O, Moss-Morris R, Mehnert A, Miranda JJ, Bullinger M, et al. Comorbid depression in medical diseases. Nat Rev Dis Primers 2020;6(1):69. [doi: 10.1038/s41572-020-0200-2] [Medline: 32820163]
15. Schmitz N, Gariépy G, Smith KJ, Clyde M, Malla A, Boyer R, et al. Recurrent subthreshold depression in type 2 diabetes: an important risk factor for poor health outcomes. Diabetes Care 2014;37(4):970-978. [doi: 10.2337/dc13-1832] [Medline: 24198303]
16. Yohannes AM, Baldwin RC, Connolly MJ. Prevalence of sub-threshold depression in elderly patients with chronic obstructive pulmonary disease. Int J Geriatr Psychiatry 2003;18(5):412-416. [doi: 10.1002/gps.851] [Medline: 12766917]
17. Hance M, Carney RM, Freedland KE, Skala J. Depression in patients with coronary heart disease. a 12-month follow-up. Gen Hosp Psychiatry 1996;18(1):61-65. [doi: 10.1016/0163-8343(95)00100-x] [Medline: 8666215]
18. Bot M, Pouwer F, Ormel J, Slaets JPJ, de Jonge P. Predictors of incident major depression in diabetic outpatients with subthreshold depression. Diabet Med 2010;27(11):1295-1301. [doi: 10.1111/j.1464-5491.2010.03119.x] [Medline: 20950389]
19. Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B. Depression, chronic diseases, and decrements in health: results from the World Health Surveys. Lancet 2007;370(9590):851-858. [doi: 10.1016/S0140-6736(07)61415-9] [Medline: 17826170]
20. Iliffe LL, Thompson AR. Investigating the beneficial experiences of online peer support for those affected by alopecia: an interpretative phenomenological analysis using online interviews. Br J Dermatol 2019;181(5):992-998 [FREE Full text] [doi: 10.1111/bjd.17998] [Medline: 30972732]
21. Berkanish P, Pan S, Viola A, Rademaker Q, Devine KA. Technology-based peer support interventions for adolescents with chronic illness: a systematic review. J Clin Psychol Med Settings 2022;29(4):911-942 [FREE Full text] [doi: 10.1007/s10880-022-09853-0] [Medline: 35147830]
22. Gerritzen EV, Lee AR, McDermott O, Coulson N, Orrell M. Online peer support for people with parkinson disease: narrative synthesis systematic review. JMIR Aging 2022;5(3):e35425 [FREE Full text] [doi: 10.2196/35425] [Medline: 35896025]
23. Maclachlan LR, Mills K, Lawford BJ, Egerton T, Setchell J, Hall LM, et al. Design, delivery, maintenance, and outcomes of peer-to-peer online support groups for people with chronic musculoskeletal disorders: systematic review. J Med Internet Res 2020 Apr 24;22(4):e15822 [FREE Full text] [doi: 10.2196/15822] [Medline: 32329746]
24. Kingod N, Cleal B, Wahlberg A, Husted GR. Online peer-to-peer communities in the daily lives of people with chronic illness: a qualitative systematic review. Qual Health Res 2017 Jan;27(1):89-99. [doi: 10.1177/1049732316680203] [Medline: 27956659]
25. The living better project: addressing mental health and wellbeing in people living with long-term condition. Mental Health Foundation. 2011. URL: https://www.mentalhealth.org.uk/publications/living-better-report [accessed 2026-02-12]
26. NHS. The Five Year Forward View for Mental Health. England: Mental Health Taskforce NHS England; 2016.
27. Rowlands H, Aylwin-Foster E, Hotopf M, Rayner L, Simpson A, Lavelle G, et al. A qualitative exploration of the psychosocial needs of people living with long-term conditions and their perspectives on online peer support. Health Expect 2023 Oct;26(5):2075-2088 [FREE Full text] [doi: 10.1111/hex.13814] [Medline: 37458403]

XSL•FO
RenderX

28. Jones HL, Lavelle G, Aylwin-Foster E, Carr E, Simpson A, Hotopf M, et al. Co-producing an online platform for people with long-term physical health conditions: a development and usability study. JMIR Preprints . [doi: 10.2196/preprints.79666]
29. Lavelle G, Jones HG, Carr E, Aylwin-Foster E, Lawrence V, Simpson A, et al. Online peer support for long-term conditions: protocol for a feasibility randomized controlled trial. JMIR Res Protoc 2025 Jul 23;14:e71513 [FREE Full text] [doi: 10.2196/71513] [Medline: 40700748]
30. Wood HE, Karampatakis GD, Coulson NS, Sastry N, Li X, Taylor SJC, et al. Moderators' experiences of the safety and effectiveness of patient engagement in an asthma online health community: exploratory qualitative interview study. J Med Internet Res 2025;27:e58167 [FREE Full text] [doi: 10.2196/58167] [Medline: 40279577]
31. Huh J, Marmor R, Jiang X. Lessons learned for online health community moderator roles: a mixed-methods study of moderators resigning from webMD communities. J Med Internet Res 2016;18(9):e247 [FREE Full text] [doi: 10.2196/jmir.6331] [Medline: 27608721]
32. Easton K, Diggle J, Ruethi-Davis M, Holmes M, Byron-Parker D, Nuttall J, et al. Qualitative exploration of the potential for adverse events when using an online peer support network for mental health: cross-sectional survey. JMIR Ment Health 2017;4(4):e49 [FREE Full text] [doi: 10.2196/mental.8168] [Medline: 29084710]
33. Ziebland S, Wyke S. Health and illness in a connected world: how might sharing experiences on the internet affect people's health? Milbank Q 2012;90(2):219-249 [FREE Full text] [doi: 10.1111/j.1468-0009.2012.00662.x] [Medline: 22709387]
34. Lobban F, Coole M, Donaldson E, Glossop Z, Haines J, Johnston R, et al. Improving peer online forums (iPOF): protocol for a realist evaluation of peer online mental health forums to inform practice and policy. BMJ Open 2023;13(7):e075142 [FREE Full text] [doi: 10.1136/bmjopen-2023-075142] [Medline: 37518092]
35. Young C. Community management that works: how to build and sustain a thriving online health community. J Med Internet Res 2013;15(6):e119 [FREE Full text] [doi: 10.2196/jmir.2501] [Medline: 23759312]
36. Kraut RE, Resnick P. Building Successful Online Communities: Evidence-Based Social Design. Cambridge: Mit Press; 2012.
37. Skousen T, Safadi H, Young C, Karahanna E, Safadi S, Chebib F. Successful moderation in online patient communities: inductive case study. J Med Internet Res 2020;22(3):e15983 [FREE Full text] [doi: 10.2196/15983] [Medline: 32181743]
38. Matzat U, Rooks G. Styles of moderation in online health and support communities: an experimental comparison of their acceptance and effectiveness. Comput Hum Behav 2014;36:65-75. [doi: 10.1016/j.chb.2014.03.043]
39. Grimmelmann J. The virtues of moderation. Yale J Law Technol 2015;17:42. [doi: 10.31228/osf.io/qwxf5]
40. Seering J, Wang T, Yoon J, Kaufman G. Moderator engagement and community development in the age of algorithms. New Media Soc 2019;21(7):1417-1443. [doi: 10.1177/1461444818821316]
41. Biagianti B, Quraishi SH, Schlosser DA. Potential benefits of incorporating peer-to-peer interactions into digital interventions for psychotic disorders: a systematic review. Psychiatr Serv 2018;69(4):377-388 [FREE Full text] [doi: 10.1176/appi.ps.201700283] [Medline: 29241435]
42. Lavelle G, Jones HG, Carr E, Aylwin-Foster E, Lawrence V, Simpson A, et al. Online peer support for long-term conditions: protocol for a feasibility randomized controlled trial. JMIR Res Protoc 2025 Jul 23;14:e71513 [FREE Full text] [doi: 10.2196/71513] [Medline: 40700748]
43. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
44. Jones HG, Lavelle G, Aylwin-Foster E, Carr E, Simpson A, Hotopf M, et al. CommonGround: online peer support in long-term conditions: a feasibility randomised controlled trial. Open Science Framework (OSF). 2025. URL: https://osf.io/zhncp/ [accessed 2025-04-08]
45. Samaritans T. Online harms advisory service. Samaritans. 2024. URL: https://www.samaritans.org/about-samaritans/research-policy/internet-suicide/samaritans-online-harms-advisory-service/ [accessed 2026-02-12]
46. Scope. 2025. URL: https://www.scope.org.uk/ [accessed 2025-11-07]
47. Scope Online Forum. 2025. URL: https://forum.scope.org.uk/ [accessed 2025-07-22]
48. O'Halloran J, Miller GC, Britt H. Defining chronic conditions for primary care with ICPC-2. Fam Pract 2004;21(4):381-386. [doi: 10.1093/fampra/cmh407] [Medline: 15249526]
49. Braun V, Clarke V. One size fits all? what counts as quality practice in (reflexive) thematic analysis? Qual Res Psychol 2021;18(3):328-352. [doi: 10.1080/14780887.2020.1769238]
50. Successful qualitative research: A practical guide for beginners. In: Successful Qualitative Research: A Practical Guide for Beginners. California: Sage; Mar 01, 2016:48-50.
51. Jiang JA, Nie P, Brubaker JR, Fiesler C. A trade-off-centered framework of content moderation. ACM Trans Comput-Hum Interact 2023 Mar 07;30(1):1-34. [doi: 10.1145/3534929]
52. Jhaver SA, Bruckman A, Gilbert E. Does transparency in moderation really matter? Proc ACM Hum-Comput Interact 2019;3(CSCW):1-27. [doi: 10.1145/3359252]
53. Saha K, Kiranmai Ernala S. Understanding moderation in online mental health communities. In: Social Computing and Social Media. Participation, User Experience, Consumer Experience, and Applications of Social Computing. Cham, Switzerland: Springer International Publishing; 2020.

XSL•FO
RenderX

54. Strand M, Eng LS, Gammon D. Combining online and offline peer support groups in community mental health care settings: a qualitative study of service users' experiences. Int J Ment Health Syst 2020;14(1):39 [FREE Full text] [doi: 10.1186/s13033-020-00370-x] [Medline: 32514303]
55. Smith-Merry J, Goggin G, Campbell A, McKenzie K, Ridout B, Baylosis C. Social connection and online engagement: insights from interviews with users of a mental health online forum. JMIR Ment Health 2019;6(3):e11084 [FREE Full text] [doi: 10.2196/11084] [Medline: 30912760]
56. Tseng HT, Ibrahim F, Hajli N, Nisar TM, Shabbir H. Effect of privacy concerns and engagement on social support behaviour in online health community platforms. Technol Forecast Soc Change 2022;178:121592. [doi: 10.1016/j.techfore.2022.121592]
57. Zhu M, Wu C, Huang S, Zheng K, Young SD, Yan X, et al. Privacy paradox in mHealth applications: an integrated elaboration likelihood model incorporating privacy calculus and privacy fatigue. Telemat Inform 2021;61:101601. [doi: 10.1016/j.tele.2021.101601]
58. Adjekum A, Blasimme A, Vayena E. Elements of trust in digital health systems: scoping review. J Med Internet Res 2018;20(12):e11254 [FREE Full text] [doi: 10.2196/11254] [Medline: 30545807]
59. Alhammad N, Alajlani M, Abd-Alrazaq A, Epiphaniou G, Arvanitis T. Patients' perspectives on the data confidentiality, privacy, and security of mHealth apps: systematic review. J Med Internet Res 2024;26:e50715 [FREE Full text] [doi: 10.2196/50715] [Medline: 38820572]
60. Maloney-Krichmar D, Preece J. A multilevel analysis of sociability, usability, and community dynamics in an online health community. ACM Trans Comput-Hum Interact 2005;12(2):201-232. [doi: 10.1145/1067860.1067864]
61. De Simoni A, Shah AT, Olivia F, Jasmine P, Aziz S, Pietro P, et al. Superusers' engagement in asthma online communities: asynchronous web-based interview study. J Med Internet Res 2020;22(6):e18185 [FREE Full text] [doi: 10.2196/18185] [Medline: 32573463]
62. Huh J. Clinical questions in online health communities: the case of "See your doctor" threads. 2015 Presented at: Proceedings of the Conference on Computer-Supported Cooperative Work. Conference on Computer-Supported Cooperative Work; 2025 Oct 18; Bergen, Norway p. 1488-1499. [doi: 10.1145/2675133.2675259]
63. Seering J. Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation. Proc ACM Hum-Comput Interact 2020;4(CSCW2):1-28. [doi: 10.1145/3415178]
64. Gibson AD. What teams do: exploring volunteer content moderation team labor on Facebook. Soc Media Soc 2023 Jul 19;9(3):20563051231186109. [doi: 10.1177/20563051231186109]
65. Statistics OFN. Internet access – households and individuals, Great Britain Statistical bulletins. Office for National Statistics. 2020. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/previousreleases [accessed 2026-02-12]
66. Schlief M, Saunders KRK, Appleton R, Barnett P, Vera San Juan N, Foye U, et al. Synthesis of the evidence on what works for whom in telemental health: rapid realist review. Interact J Med Res 2022;11(2):e38239 [FREE Full text] [doi: 10.2196/38239] [Medline: 35767691]

Abbreviations

EQUATOR: Enhancing the Quality and Transparency of Health Research
fRCT: feasibility randomized controlled trial
KCL: King’s College London
SRQR: Standards for Reporting Qualitative Research

Edited by A Stone; submitted 08.Aug.2025; peer-reviewed by H Wood, E Kalles, A Gupta, S Mittal; comments to author 02.Oct.2025; accepted 30.Jan.2026; published 12.Mar.2026.

Please cite as:
Jones HG, Lavelle G, Aylwin-Foster E, Regan C, Simpson A, Carr E, Hotopf M, Lawrence V
Insights and Recommendations From Moderators and Community Members for Keeping Online Peer Support Safe: Thematic Analysis
J Med Internet Res 2026;28:e81943
URL: https://www.jmir.org/2026/1/e81943
doi:10.2196/81943
PMID:

©Hannah Grace Jones, Grace Lavelle, Elly Aylwin-Foster, Ciara Regan, Alan Simpson, Ewan Carr, Matthew Hotopf, Vanessa Lawrence. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effect of a Digital-Driven Physician-Pharmacist Collaborative Model for Diabetes in Primary Health Care: Cluster Randomized Trial

Jie Xiao[1,2], MM; Qing Wang[1,2], MM; Shenglan Tan[1,2], PhD; Lei Chen[1,2], PhD; Daxiong Xiang[1,2], PhD; Haiyan Yuan[1,2], PhD; Xia Li[3], MD; Shuting Huang[1,2], MM; Bingjie Tang[1,2], MM; Yan Guo[4], MS; Haiying Huang[5], MS; Danhui Zhao[6], BS; Yue Li[7], BS; Li Wang[8], BS; Qun Li[9], MS; Juan Liu[10], MS; Ping Xu[1,2], PhD

[1]Department of Pharmacy, Second Xiangya Hospital, Central South University, Changsha, China
[2]Institute of Clinical Pharmacy, Second Xiangya Hospital, Central South University, Changsha, China
[3]Department of Endocrine, Second Xiangya Hospital, Central South University, Changsha, China
[4]Department of Pharmacy, Taoyuan People's Hospital, Changde, China
[5]Department of Pharmacy, The People's Hospital of Liuyang, Changsha, China
[6]Department of Pharmacy, The First People's Hospital of Pingjiang, Yueyang, China
[7]Department of Pharmacy, People's Hospital of Ningxiang, Changsha, China
[8]Department of Pharmacy, Yueyang Central Hospital, Yueyang, China
[9]Department of Pharmacy, The Second People's Hospital of Huaihua, Huaihua, China
[10]Department of Pharmacy, Yiyang Central Hospital, Yiyang, China

Corresponding Author:
Ping Xu, PhD
Department of Pharmacy
Second Xiangya Hospital
Central South University
No.139 Renmin Road, Furong District
Changsha, 410011
China
Phone: 86 731 85292128
Email: xuping1109@csu.edu.cn

Abstract

Background: Evidence-based physician-pharmacist collaborative clinics have demonstrated significant short-term benefits for patients with type 2 diabetes (T2D), but their long-term effectiveness remains unclear, especially in primary health care settings.

Objective: This study aimed to explore the long-term effectiveness and cost-effectiveness of a novel, digital-driven, multifaceted physician-pharmacist collaborative model for managing patients with T2D in underresourced settings.

Methods: We conducted a 12-month cluster randomized controlled trial from May 2021 to December 2022 across 6 primary health care settings in China. Guided by the theory of planned behavior, the intervention involved routine therapy from physicians along with pharmaceutical interventions from pharmacists. These were delivered through a combination of face-to-face visits and mobile health care. The intervention group received 4 face-to-face visits and biweekly remote education sessions over the 12 months. We conducted intention-to-treat analyses to estimate differences in clinical and behavior indicators between the intervention and control groups. Primary outcomes included glycosylated hemoglobin and 10-year atherosclerotic cardiovascular risk. Data were analyzed using adjusted generalized estimation equations.

Results: This study included 574 patients (291 in the intervention group and 283 in the control group). Over 12 months, patients in the intervention group had significant reductions in hemoglobin A_{1c} (–2.57 vs –1.96, respectively; P<.001; 95% CI –1.027 to –0.238) and 10-year atherosclerotic cardiovascular risk (–1.35 vs 0.01, respectively; P<.001; 95% CI –1.690 to –0.630) compared with the control group. Substantial improvements were also observed in several secondary outcomes, including fasting blood glucose, 2-hour postprandial blood glucose, waist circumference, waist-to-hip ratio, blood pressure, triglyceride, and total cholesterol. Total diabetes-related costs decreased, and patient satisfaction improved significantly in the intervention group. There were no significant differences in BMI, high-density lipoprotein, or low-density lipoprotein.

XSL•FO
RenderX

Conclusions: These findings suggest that the physician-pharmacist collaborative model could improve the long-term quality and efficiency of T2D management and reduce medical costs in underresourced areas globally. Patients with T2D, especially those with central obesity or high cardiovascular risk, may benefit more from collaborative clinics.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2000031839; https://www.chictr.org.cn/showproj.html?proj=51910

(J Med Internet Res 2026;28:e77470) doi:10.2196/77470

KEYWORDS

diabetes mellitus; physician-pharmacist collaboration; central obesity; glycemic control; atherosclerotic cardiovascular risk; cost-effectiveness analysis

Introduction

The global epidemic of diabetes, particularly type 2 diabetes (T2D), has caused severe public health challenges worldwide [1]. The International Diabetes Federation estimates that the prevalence of diabetes worldwide was 10.5% in 2021 and would increase to 12.2% by 2045 due to rapid lifestyle changes and aging of the population [2]. China has the largest population of people living with diabetes worldwide, with an estimated 1.39 million deaths attributable to diabetes or its complications in 2021, and this number continues to rise rapidly [2]. As in other countries, most of the population with diabetes in China has T2D, and half of them present with complications, leading to a huge economic burden [3]. This burden is exacerbated in resource-limited settings, where health care systems are often underequipped to manage the long-term complexities of the disease [1].

Effective, long-term management of T2D is therefore crucial to mitigating these complications and associated costs [4]. Diabetes management is multidimensional, including a low-energy diet, physical activity, medication, diabetes education, and continuous glucose monitoring [4-6]. In vast rural areas of China, primary health care centers are the preferred institutions for more than 900 million people, where diabetes management is still dominated by physicians [7]. However, diabetes management in rural areas in China is currently insufficient due to limited resources and staff shortage in primary health care centers [8]. Moreover, patients in rural areas generally have lower educational background and socioeconomic status, which necessitates more detailed treatment and management support, which puts further strain on the capacity of the current physician-led care model [9].

To address these issues, innovative team-based care models have shown promise [10-13]. One well-established model is the physician-pharmacist collaboration, which improves the quality of outpatient services and reduces physicians' workload by involving pharmacists in medication counseling and patient education [14,15]. However, due to the limited reach and accessibility of in-person pharmaceutical services, it could be difficult to implement established pharmaceutical intervention programs in underresourced areas [16]. Mobile health (mHealth) interventions offer a promising way to enhance and sustain collaborative care in these contexts [17]. Evidence shows that mHealth tools, including medication reminders, mobile apps for tracking, and remote monitoring, can improve medication adherence, self-management behaviors, and clinical outcomes in diabetes care, thereby alleviating the workload of health care providers [18,19]. An mHealth-based physician-pharmacist collaboration model could be adopted in underresourced settings.

Previous studies have demonstrated that the intervention programs based on the theory of planned behavior (TPB) contribute to change behavior patterns, but few studies have tailored the method to comprehensive intervention for diabetes management in underresourced settings [20,21]. Therefore, we have developed a digital-driven, multifaceted intervention program based on physician-pharmacist collaborative models, combining face-to-face visits and mHealth support. We conducted a cluster randomized controlled trial to evaluate the long-term effectiveness of the intervention on hemoglobin A_{1c} (HbA_{1c}), 10-year atherosclerotic cardiovascular disease (ASCVD) risk, body weight, blood pressure, lipids, and cost-effectiveness of the physician-pharmacist collaborative model for T2D management in primary health care centers in China.

Methods

Study Design

The cluster randomized controlled trial was conducted from May 2021 to December 2022 in 6 health care centers in Hunan province, China. A trial protocol was reported in a previous study [22]. Hunan province is located in the center of China, with middle-level economy and general medical condition. To ensure representativeness, all 88 Hunan counties were divided into 3 subgroups by per capita income level, namely, high, medium, and low incomes. We then sent invitations to 27 primary health care centers (the largest by outpatient volume in their respective counties) across these strata. Of the 18 centers that agreed to participate, we randomly selected 2 centers from each income stratum. Within each stratum, the 2 selected centers were then randomly allocated to the intervention and control groups (1 per group), resulting in 3 centers in the intervention group and 3 in the control group. All eligible patients attending these centers received care according to their center's allocation. This study was reported following the CONSORT (Consolidated Standards of Reporting Trials) extension for cluster trials (Multimedia Appendix 1). We held progress meetings monthly and used fidelity checklists to monitor implementation (Multimedia Appendix 2).

Randomization and Masking

Randomization was performed at the level of the health care center (cluster). Using a standard randomization protocol with a computer-based random sequence generation, an independent

XSL•FO
RenderX

statistician performed the randomization [23]. This involved randomly selecting the 6 study centers from the 18 consenting centers, followed by randomly allocating the selected centers to the intervention and control groups with balanced allocation. Because the intervention required knowledge of patient allocation, blinding of participants was not feasible; however, data processors were blinded to group assignment.

Participants

The study sample included patients with T2D who were aged ≥18 years, had HbA_{1c} levels >7.5% in the past 2 months, and provided contact information. Patients were excluded if they were pregnant or had end-stage renal failure, dementia, psychiatric disorders, cancer, congestive heart failure, or pancreatitis. All eligible patients were required to provide written informed consent before participation in the study.

Intervention and Procedures

Pharmacist Training and Qualification

Pharmacists participating in the study had worked as clinical pharmacists for >2 years, held standardized training qualifications, and completed a 3-month comprehensive training course at the Second Xiangya Hospital. The training courses mainly cover 2 aspects: diabetes management and cooperative skills. Topics included knowledge of diabetes, hypoglycemic agents, treatment target, interpretation of laboratory tests, medication adherence, diet, exercise, and lifestyle. Detailed training content is available in Table S1 in Multimedia Appendix 3 [24]. Intermediate and final tests are used to assess learning, and those who fail the tests were excluded from the study [22].

Intervention Group

Centers in the intervention group provided routine physician-led therapy combined with pharmacist-delivered pharmaceutical care through outpatient visits at baseline and at the third, sixth, ninth, and 12th months, along with biweekly remote education via WeChat (the most popular social media platform in China [25,26]) throughout the 12-month study period. The pharmaceutical intervention program was based on the TPB framework and included 3 areas: attitude toward the behavior, subjective norm, and perceived behavior control (Textbox 1). Well-trained pharmacists provided multifaceted pharmaceutical care at each follow-up visit, which lasted at least 20 minutes; handed out health education manuals at the end of visits; and conducted biweekly remote health education through WeChat to regulate the patient attitude toward treatment. Multifaceted pharmaceutical care included diabetes education, medication guidance, lifestyle intervention, treatment of adverse drug reactions, and complication identification, while remote health education involved pharmaceutical consultation, structured educational messages, and regular medication reminders (Table S2 in Multimedia Appendix 3). Subjective norm was achieved through peer support and family support to strengthen patient behavior intention. Meanwhile, pharmacists set health goals for patients and strengthened their medication knowledge to improve the perception of treatment effects.

Textbox 1. The intervention program based on the theory of planned behavior.

Attitude toward the behavior

- Face-to-face pharmaceutical care at each visit, including diabetes education, medication guidance, lifestyle intervention, adverse drug reaction management, and complication identification
- Health education manuals were distributed, covering diabetes knowledge, hazards of complications, introduction of commonly used drugs, healthy lifestyle, and other related knowledge
- Health education was delivered through the WeChat group, including answering questions from patients and their families; providing guidance on daily management; offering disease education; and sharing pharmaceutical knowledge, dietary advice, and exercise plans

Subjective norm

- Peer support, which allows patients to interact with each other within their peer group in WeChat
- Family support, including diabetes care, rational drug use, healthy lifestyle, and other related knowledge taught by pharmacists

Perceived behavior control

- Formulation of health goals, including short-term and long-term health goals
- Medication consultation services, including dose adjustment, medication duration, drug interactions, and adverse drug reactions

Control Group

Patients in the control centers received usual treatment and management by physicians, with no pharmacist available to provide additional patient education. After the study, all patients received a diabetes education manual.

Outcomes

The primary outcomes were HbA_{1c} and 10-year ASCVD risk over 12 months, while glycemic measurements, weight-related measurements, blood pressure, lipid indicators, patient satisfaction, and costs were secondary outcomes. HbA_{1c} provides a better reflection of the patient's average blood glucose level over a 3-month period. The 10-year ASCVD risks were calculated by the Prediction for atherosclerotic cardiovascular disease risk in China model [27]. The biomarkers of glycemic measurements include fasting blood glucose (FBG) and 2-hour postprandial blood glucose (PBG2h). Weight-related measurements included BMI, waist circumference (WC), and waist-to-hip ratio (WHR). The indicators of blood pressure

XSL•FO
RenderX

included systolic blood pressure and diastolic blood pressure. The biomarkers of blood lipid status include triglyceride, total cholesterol (TC), high-density lipoprotein cholesterol, and low-density lipoprotein cholesterol (LDL-C). Patient satisfaction was assessed by the Diabetes Treatment Satisfaction Questionnaire, with a Cronbach alpha index of 0.717. The total cost of a patient's treatment included the costs of drugs, registration, examinations, hospitalization, lost wages, and transportation related to diabetes care.

Sample Size

On the basis of previous studies, we calculated that at 90% power and 5% significance level, the minimum detectable difference in HbA_{1c} in the intervention group was 0.76% with an SD of 1.0%, while minimum detectable difference in 10-year ASCVD risk was 3.98 [28,29]. The intraclass correlation coefficients at the center level were 0.03 for HbA_{1c} and 0.05 for 10-year ASCVD risk [30,31]. Considering the potential effects of clustering and loss to follow-up over 12 months, the study required a sample size of at least 288 patients in this study (144 in each group).

Statistical Analysis

All analyses were conducted using a modified intention-to-treat approach, which excluded 26 patients who withdrew after randomization but before trial implementation. Measurement data were presented as means with SDs or median with IQRs, while enumeration data were presented as counts with percentages. Descriptive statistics were used to summarize the baseline demographic characteristics, while independent sample 2-tailed *t* tests and Mann-Whitney *U* tests were used to compare the differences between groups at baseline. Differences in measurements and costs over 12 months between the groups were detected by generalized estimating equation (GEE) analysis. GEE provides stable estimates of time, group, and interaction effects for repeated measurement data, while GEE model gave valid estimation under the assumption of missing at random. GEE models were adjusted for center, gender, age, comorbidity, smoking history, alcohol history, family history, education, income, occupation, and residence. The standard rate of control was referred to the comprehensive control target of T2D in China (Table S3 in Multimedia Appendix 3). Sensitivity analysis was performed using the per-protocol population, which included patients who adhered to the protocol and were not lost to follow-up. Missing data were minimal (<10% for primary outcomes) and were handled using GEE, which provides valid estimates under the missing-at-random assumption by using all available data without imputation. Additionally, 2-sided *P* values <.05 were considered significant for all analyses. SPSS Statistics (version 25.0; IBM Corp) and SAS (version 9.4; SAS Institute) were used.

Ethical Considerations

This study was approved by the Clinical Research Ethics Committee of the Second Xiangya Hospital of Central South University (2019-213). All participants were informed of the study's purpose, procedures, potential risks, and benefits, and they provided written informed consent prior to enrollment. Strict measures were taken to protect participant privacy and confidentiality; all data were anonymized before analysis. As compensation for their participation, all enrolled patients received free laboratory examinations related to the study.

Results

Baseline Characteristics

The study was conducted in 6 centers (Table S4 in Multimedia Appendix 3) and included 574 patients, of whom 50.7% (n=291) participated in physician-pharmacist collaborative clinics (Table 1). The mean age of the enrolled patients was 55.0 (48.0 to 64.0) years and 52.4% (n=301) were male. Overall, 34% (n=195) of the patients had a history of smoking and 27.2% (n=156) a history of alcohol use. Baseline characteristics of patients were generally comparable between the 2 groups. Patient enrollment is shown in Figure 1.

XSL•FO
RenderX

Table 1. Patients' demographic characteristics (N=574).

Characteristics	Overall	Intervention (n=291)	Control (n=283)
Age (years), median (IQR)	55.0 (48.0-64.0)	54.0 (49.0-60.0)	55.0 (46.0-67.0)
Sex, n (%)			
Male	301 (52.4)	139 (47.8)	162 (57.2)
Female	273 (47.6)	152 (52.2)	121 (42.8)
Comorbidity, n (%)			
Hypertension			
Yes	90 (15.7)	50 (17.2)	40 (14.1)
Hyperlipidemia			
Yes	26 (4.5)	10 (3.4)	16 (5.7)
Coronary heart disease			
Yes	20 (3.5)	11 (3.8)	9 (3.2)
History of smoking, n (%)			
Yes	195 (34)	84 (28.9)	111 (39.2)
History of alcohol intake, n (%)			
Yes	156 (27.2)	83 (28.5)	73 (25.8)
Family history, n (%)			
Diabetes			
Yes	164 (28.6)	89 (30.6)	75 (26.5)
Hypertension			
Yes	94 (16.4)	54 (18.6)	40 (14.1)
Hyperlipidemia			
Yes	45 (7.8)	23 (7.9)	22 (7.7)
Education, n (%)			
Junior high school and below	318 (55.4)	168 (57.7)	150 (53)
Senior high school	152 (26.5)	78 (26.8)	74 (26.1)
College and above	104 (18.1)	45 (15.5)	59 (20.9)
Income per month (CNY)[a], n (%)			
<1000	96 (16.7)	54 (18.5)	42 (14.8)
1000-3000	191 (33.3)	98 (33.7)	93 (32.9)
>3000	287 (50)	139 (47.8)	148 (52.3)
Occupation, n (%)			
Full-time job	395 (68.8)	215 (73.9)	180 (65.9)
Part-time job or no job	179 (31.2)	76 (26.1)	103 (34.1)
Residence, n (%)			
Rural area	321 (55.9)	151 (51.9)	170 (62.3)
Town or urban area	253 (44.1)	140 (48.1)	113 (37.7)
HbA_{1c}[b], mean (SD)	9.83 (2.02)	9.65 (1.87)	10.02 (2.15)
10-year ASCVD[c] risk, mean (SD)	7.56 (5.87)	7.05 (4.92)	8.06 (6.65)
FBG[d] (mmol/L), mean (SD)	10.65 (4.16)	10.60 (3.60)	10.69 (4.66)
PBG2h[e] (mmol/L), mean (SD)	15.19 (5.27)	15.34 (4.64)	15.03 (5.84)
BMI (kg/m^2), mean (SD)	24.87 (3.33)	24.73 (3.38)	25.02 (3.28)

XSL•FO
RenderX

Characteristics	Overall	Intervention (n=291)	Control (n=283)
WC[f] (cm), mean (SD)			
Male	91.92 (9.34)	91.17 (8.89)	92.60 (9.69)
Female	88.73 (10.04)	88.62 (10.23)	88.88 (9.86)
WHR[g], mean (SD)			
Male	0.927 (0.062)	0.924 (0.061)	0.930 (0.062)
Female	0.927 (0.078)	0.927 (0.076)	0.927 (0.078)
SBP[h] (mm Hg), mean (SD)	133.38 (18.33)	132.53 (16.05)	134.26 (20.41)
DBP[i] (mm Hg), mean (SD)	81.20 (9.90)	80.47 (9.03)	81.96 (10.71)
Triglyceride (mmol/L), mean (SD)	2.80 (3.15)	3.09 (3.87)	2.53 (2.20)
TC[j] (mmol/L), mean (SD)	4.98 (1.56)	5.12 (1.74)	4.84 (1.36)
HDL-C[k] (mmol/L), mean (SD)	1.30 (0.57)	1.30 (0.53)	1.30 (0.60)
LDL-C[l] (mmol/L), mean (SD)	2.94 (1.11)	2.96 (1.25)	2.92 (0.94)

[a]CNY 1=US $0.145.

[b]HbA_{1c}: hemoglobin A_{1c}.

[c]ASCVD: atherosclerotic cardiovascular disease.

[d]FBG: fasting blood glucose.

[e]PBG2h: 2-hour postprandial blood glucose.

[f]WC: waist circumference.

[g]WHR: waist-to-hip ratio.

[h]SBP: systolic blood pressure.

[i]DBP: diastolic blood pressure.

[j]TC: total cholesterol.

[k]HDL-C: high-density lipoprotein cholesterol.

[l]LDL-C: low-density lipoprotein cholesterol.

XSL•FO
RenderX

Figure 1. Flowchart of the trial.

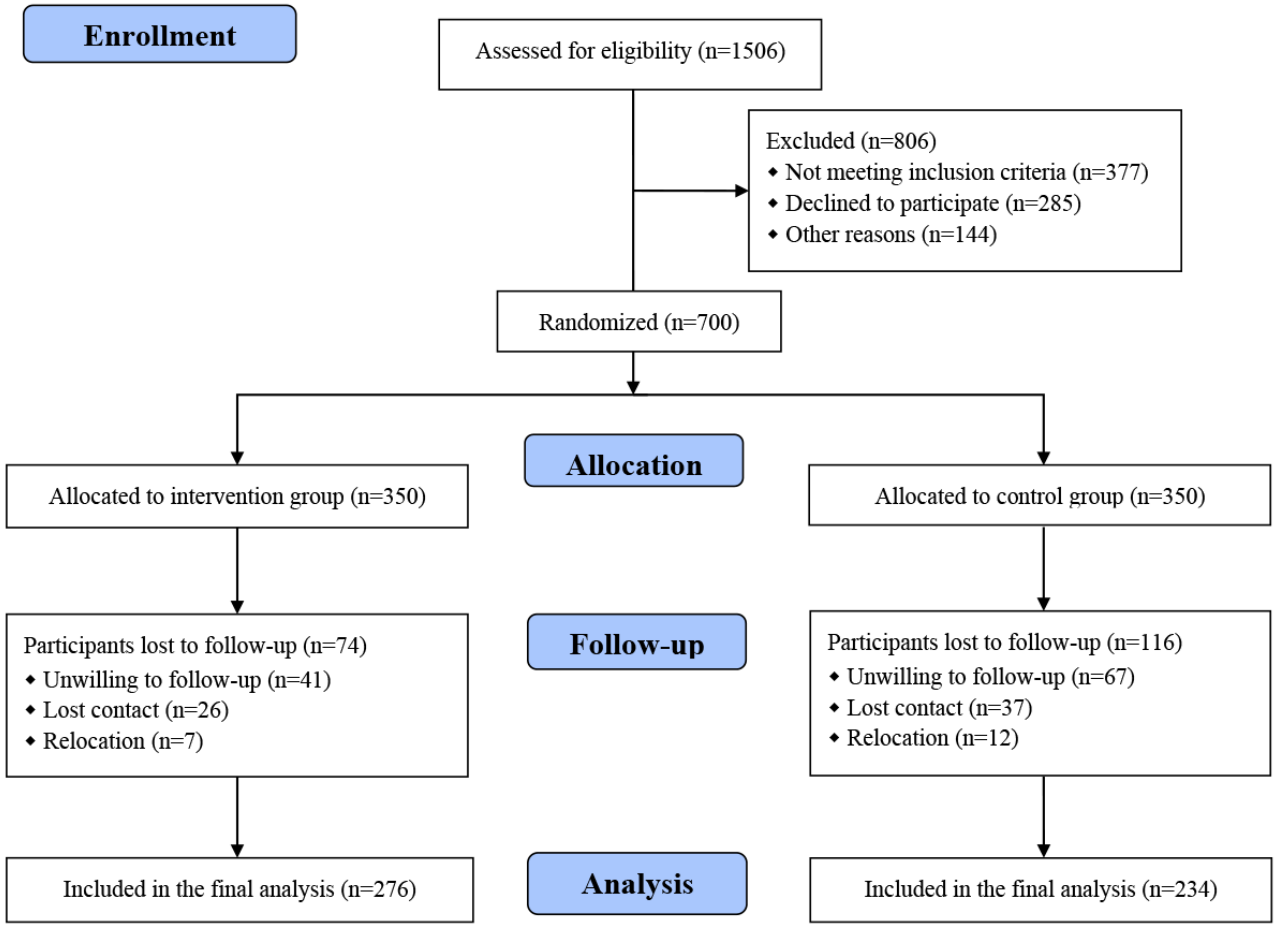

Primary Outcomes

At 12 months, the reduction in HbA_{1c} was significantly greater in the intervention group (−2.57%) than in the control group (−1.96%; β=−0.632; *P*<.001; 95% CI −1.027 to −0.238). Detailed information is shown in Table 2. Correspondingly, the HbA_{1c} control rates increased markedly from baseline to 12 months in both groups, but the improvement was significantly greater in the intervention group (from 0% to 53.61%) than in the control group (from 1.77% to 38.52%; *P*=.001). Refer to Table S5 in Multimedia Appendix 3 for full details. The 10-year ASCVD risk decreased significantly in the intervention group compared with the control group over the study period (β=−1.206; *P*<.001; 95% CI −1.730 to −0.68).

XSL•FO
RenderX

Table 2. The 12-month effectiveness of the physician-pharmacist collaborative clinics on clinical outcomes.

Variables	Intervention group, mean (SD)	Control group, mean (SD)	Unadjusted effect		Adjusted effect[a]	
			P value[b]	β (95% CI)	*P* value	β (95% CI)
HbA_{1c}[c]			<.001	−0.632 (−1.026 to −0.237)	<.001	−0.632 (−1.027 to −0.238)
Baseline	9.65 (1.87)	10.02 (2.15)				
3 months	7.93 (1.53)	8.43 (2.19)				
6 months	7.76 (1.63)	8.15 (1.86)				
9 months	7.55 (1.48)	8.18 (1.87)				
12 months	7.08 (1.13)	8.06 (2.02)				
10-year ASCVD[d] risk			<.001	−1.160 (−1.690 to −0.630)	<.001	−1.206 (−1.730 to −0.682)
Baseline	7.05 (4.92)	8.06 (6.65)				
12 months	5.70 (4.05)	8.07 (6.37)				
FBG[e]			.29	−0.592 (−1.363 to 0.178)	.27	−0.602 (−1.372 to 0.169)
Baseline	10.60 (3.60)	10.69 (4.66)				
3 months	8.38 (2.54)	9.03 (3.32)				
6 months	8.40 (2.77)	8.92 (2.67)				
9 months	8.02 (2.43)	8.31 (2.28)				
12 months	7.58 (2.09)	8.22 (2.67)				
PBG2h[f]			.03	−1.234 (−2.338 to −0.129)	.03	−1.210 (−2.315 to −0.106)
Baseline	15.34 (4.64)	15.03 (5.84)				
3 months	12.38 (3.89)	12.03 (4.54)				
6 months	11.60 (3.67)	11.45 (4.63)				
9 months	11.36 (3.42)	11.06 (4.07)				
12 months	10.44 (3.66)	11.17 (4.10)				
BMI			.21	−0.238 (−0.528 to 0.051)	.23	−0.236 (−0.527 to 0.054)
Baseline	24.73 (3.38)	25.02 (3.28)				
3 months	24.49 (3.22)	24.68 (3.38)				
6 months	24.37 (3.13)	24.89 (3.16)				
9 months	24.23 (3.12)	24.73 (3.29)				
12 months	24.28 (3.14)	24.77 (3.36)				
WC[g]			<.001	−3.570 (−4.200 to −2.940)	<.001	−3.557 (−4.185 to −2.929)
Baseline	89.83 (9.68)	90.98 (9.89)				
3 months	89.06 (9.67)	91.46 (9.83)				
6 months	87.09 (9.21)	91.82 (9.46)				
9 months	87.02 (9.42)	92.50 (9.13)				
12 months	86.84 (9.24)	91.92 (9.85)				
WHR[h]			<.001	−0.020 (−0.025 to −0.015)	<.001	−0.020 (−0.024 to −0.013)
Baseline	0.926 (0.069)	0.929 (0.070)				
3 months	0.920 (0.073)	0.930 (0.080)				

XSL•FO
RenderX

Variables	Intervention group, mean (SD)	Control group, mean (SD)	Unadjusted effect		Adjusted effect[a]	
			P value[b]	β (95% CI)	*P* value	β (95% CI)
6 months	0.911 (0.068)	0.932 (0.073)				
9 months	0.909 (0.070)	0.938 (0.073)				
12 months	0.910 (0.068)	0.933 (0.067)				
SBP[i]			.007	−3.614 (−6.239 to −0.989)	.008	−3.556 (−6.187 to −0.925)
Baseline	132.53 (16.05)	134.26 (20.41)				
3 months	130.21 (13.26)	132.67 (16.37)				
6 months	128.53 (12.14)	132.92 (14.10)				
9 months	127.28 (11.24)	134.30 (17.08)				
12 months	127.30 (11.70)	132.54 (15.20)				
DBP[j]			.01	−2.683 (−4.342 to −1.024)	.01	−2.639 (−4.295 to −0.983)
Baseline	80.47 (9.03)	81.96 (10.71)				
3 months	78.72 (9.04)	81.77 (10.64)				
6 months	77.51 (8.16)	81.53 (9.45)				
9 months	77.12 (8.99)	82.09 (9.94)				
12 months	77.06 (7.94)	81.26 (8.52)				
Triglyceride			.01	−0.815 (−1.293 to −0.337)	.01	−0.814 (−1.293 to −0.335)
Baseline	3.09 (3.87)	2.53 (2.20)				
3 months	2.52 (2.54)	2.40 (2.37)				
6 months	2.37 (2.24)	2.44 (2.70)				
9 months	2.19 (2.30)	2.20 (1.53)				
12 months	1.93 (1.40)	2.24 (1.53)				
TC[k]			.001	−0.544 (−0.804 to 0.284)	.001	−0.540 (−0.798 to −0.282)
Baseline	5.12 (1.74)	4.84 (1.36)				
3 months	4.71 (1.18)	4.60 (1.09)				
6 months	4.43 (0.96)	4.47 (1.14)				
9 months	4.16 (0.77)	4.28 (1.12)				
12 months	4.12 (0.86)	4.37 (0.96)				
HDL-C[l]			.24	0.109 (0.014 to 0.205)	.28	0.109 (0.013 to 0.204)
Baseline	1.30 (0.53)	1.30 (0.60)				
3 months	1.33 (0.66)	1.26 (0.37)				
6 months	1.32 (0.45)	1.24 (0.49)				
9 months	1.33 (0.49)	1.25 (0.49)				
12 months	1.36 (0.39)	1.25 (0.56)				
LDL-C[m]			.40	−0.043 (−0.228 to 0.143)	.42	−0.048 (−0.232 to 0.137)
Baseline	2.96 (1.25)	2.92 (0.94)				
3 months	2.74 (0.96)	2.85 (0.93)				
6 months	2.78 (0.95)	2.78 (0.81)				

XSL•FO
RenderX

Variables	Intervention group, mean (SD)	Control group, mean (SD)	Unadjusted effect		Adjusted effect[a]	
			P value[b]	β (95% CI)	*P* value	β (95% CI)
9 months	2.79 (0.93)	2.77 (0.77)				
12 months	2.78 (0.83)	2.78 (0.80)				
Satisfaction			<.001	−6.271 (−7.191 to −5.352)	<.001	−4.314 (−5.176 to −3.451)
12 months	32.43 (4.38)	26.16 (5.59)				

[a]Adjusted for center, gender, age, comorbidity, history of smoking and alcohol, family history, education, income, occupation, and residence.

[b]Estimated group-by-time interaction effects from generalized estimation equation; time effect and group effect are displayed in Table S6 in Multimedia Appendix 3.

[c]HbA_{1c}: hemoglobin A_{1c}.

[d]ASCVD: atherosclerotic cardiovascular disease.

[e]FBG: fasting blood glucose.

[f]PBG2h: 2-hour postprandial blood glucose.

[g]WC: waist circumference.

[h]WHR: waist-to-hip ratio.

[i]SBP: systolic blood pressure.

[j]DBP: diastolic blood pressure.

[k]TC: total cholesterol.

[l]HDL-C: high-density lipoprotein cholesterol.

[m]LDL-C: low-density lipoprotein cholesterol.

Secondary Outcomes

FBG levels decreased significantly more in the intervention group (mean 7.58, SD 2.09 mmol/L) than in the control group (mean 8.22, SD 2.67 mmol/L; *P*<.001), although the group-by-time interaction effect was not significant (β=−0.602; *P*=.27; 95% CI −1.372 to 0.169). Several other secondary outcomes showed significant group-by-time interaction effects favoring the intervention group, including PBG2h (β=−1.210; *P*=.03; 95% CI −2.315 to −0.106), systolic blood pressure (β=−3.556; *P*=.008; 95% CI −6.187 to −0.925), diastolic blood pressure (β=−2.639; *P*=.01; 95% CI −4.295 to −0.983), triglyceride (β=−0.814; *P*=.01; 95% CI −1.293 to −0.335), and TC (β=−0.540; *P*=.001; 95% CI −0.798 to −0.282). However, there were no significant differences in high-density lipoprotein cholesterol (β=0.109; *P*=.28; 95% CI 0.013-0.204) and LDL-C (β=−0.048; *P*=.42; 95% CI −0.232 to 0.137) between the 2 groups in the long-term interventions. Detailed data are presented in Table S6 in Multimedia Appendix 3.

Significant improvements in indicators of central obesity were also observed. WC (β=−3.557; *P*<.001; 95% CI −4.185 to −2.929) and WHR (β=−0.020; *P*<.001; 95% CI −0.024 to −0.013) decreased significantly in the intervention group, showing a significant effect on weight loss in obese people. However, the change in BMI (β=−0.236; *P*=.23; 95% CI −0.527 to 0.054) was not significant between groups. Patients in the intervention group had a mean satisfaction score of 32.43 (SD 4.38), significantly higher than those in the control group with a mean score of 26.16 (SD 5.59; β=−4.314; *P*<.001; 95% CI −5.176 to −3.451).

Economic Evaluation

Over the 12 months, the cost of diabetes care was significantly lower in the intervention group (mean 5505.86, SD 3939.86 CNY) than in the control group (mean 7902.61, SD 11,218.51 CNY; *P*=.01; Table 3). The costs of registration, examination, lost wages, and transportation of patients were significantly higher in the intervention group than in the control group, while the hospitalization costs were significantly lower, indicating a shift from inpatient care toward more routine outpatient management.

XSL•FO
RenderX

Table 3. The 12-month total costs (in CNY[a]) of patients in the 2 groups (N=574).

Variables	Overall, mean (SD)	Intervention (n=291), mean (SD)	Control (n=283), mean (SD)	*P* value
Drug cost	3986.64 (2268.41)	3835.58 (2006.62)	4141.97 (2503.41)	.12
Registration cost	47.40 (31.40)	56.01 (29.01)	38.54 (31.34)	<.001
Examination cost	520.67 (741.48)	550.88 (323.56)	490.14 (999.86)	<.001
Hospitalization cost	1935.12 (7983.22)	849.72 (3299.63)	3051.21 (10,762.68)	.01
Lost wages	160.87 (276.69)	167.56 (301.06)	154.00 (249.52)	.02
Transportation cost	41.85 (63.14)	56.54 (69.86)	26.75 (51.28)	<.001
Total cost	6687.53 (8440.00)	5505.86 (3939.86)	7902.61 (11,218.51)	.01

[a]CNY 1=US $0.145.

Sensitivity Analysis

The per-protocol population included 510 patients who adhered to the protocol and were not lost to follow-up. There were no significant differences in demographic characteristics between the modified intention-to-treat and per-protocol populations (Table S7 in Multimedia Appendix 3). Notably, significant improvements in primary and secondary outcomes were detected in per-protocol population, including HbA_{1c}, 10-year ASCVD risk, glycemic measurements, weight-related outcomes, blood pressure, and blood lipid parameters (P<.05; Tables S8-S10 in Multimedia Appendix 3).

Discussion

Principal Findings

This study reported a digital-driven, multifaceted physician-pharmacist collaborative model for diabetes treatment and management, adapted for underresourced settings. The study also evaluated the long-term effectiveness and cost-effectiveness of implementing physician-pharmacist collaborative models on participants with diabetes in primary health care. The intervention was found to significantly improve long-term clinical efficacy and associated costs for patients over 12 months in primary health care settings, with the positive effects becoming more significant over time. In all cases, the magnitude of the improvements implied benefits for patients and public health.

The intervention program, based on the TPB framework, aims to enhance self-management in patients with diabetes by facilitating the establishment of behavioral intentions and motivation, was well as stimulating positive therapeutic behaviors [21,32]. Previous studies have found that TPB serves to predict individuals' motivation to engage in health behaviors, including exercise, health checks, disease screening, and health management behaviors [33]. However, most of the available studies are derived from high-income countries and regions, leaving very limited research on TPB-based interventions in rural populations with low education and economic status [21,33,34]. Our study suggests that the TPB-based intervention program contributes to the improvement of health behaviors among patients with diabetes in resource-limited areas, providing a strong basis for the widespread use of TPB in primary health care settings.

Our findings indicated a long-term advantage of physician-pharmacist collaborative models in improving HbA_{1c} and glycemic outcomes. The HbA_{1c} control rate in the intervention group increased to 51.45%, remarkably higher than the average control rate in rural China of 44.1% and close to the urban rate of 54.1%, suggesting a narrowing of the rural-urban disparities in diabetes treatment and management [35]. Moreover, there was an improvement in FBG at the early intervention, whereas the improvement in PBG2h was not observed until 12 months. This is supported by previous studies, showing that postprandial glycemic fluctuations tend to occur in patients with well-controlled diabetes, whereas patients with worsening diabetes have higher FBG [36-38]. Therefore, physician-pharmacist collaborative clinics have shown excellent management of glycemic outcomes in the early stages of diabetes, contributing to the delay of disease progression.

The results were consistent with other studies in high-income countries and regions, which have reported comprehensive improvements in blood glucose, blood pressure, and lipids in patients with diabetes [39,40]. Pharmacist-delivered services improved the quality of health care and reduced the burden of care, generating an annual return on investment of US $1.2 million to US $2.9 million [39]. A study of 2480 patients demonstrated that the pharmacist-physician collaborative care model resulted in significant improvements in HbA_{1c}, blood pressure, and lipid levels [40]. However, it is controversial for improving cardiovascular outcomes, while long-term cardiovascular risk was unclear [15,41]. Few retrospective multicenter cohort studies showed that the collaborative care model effectively decreased the HbA_{1c} levels but did not improve cardiovascular risk in primary care settings [42,43]. The potential explanation is the absence of an explicit theoretical framework to guide interventions, resulting in a lack of long-term motivation for patients to maintain healthy behaviors.

Health economic evaluation illustrated that physician-pharmacist collaborative models significantly reduced the total cost of diabetes care for patients while improving patient outcomes. This provided solid evidence to support the implementation of physician-pharmacist collaborative models in primary health care. Combined with care-seeking behavior, we found that the

XSL•FO
RenderX

number of outpatient visits increased by 78.7% and emergency visits decreased by half in physician-pharmacist collaborative models, indicating that the direct costs due to hospitalization for diabetes were greatly reduced [44]. In conclusion, it is economically feasible to implement physician-pharmacist collaborative clinics for diabetes management in primary health care settings.

In contrast to the previous study, significant reductions in WC and WHR, rather than BMI, were identified over 12 months in physician-pharmacist collaborative care models [45]. This inconsistency could be attributed to the peculiar physical conditions of the Chinese population, which leads to excessive fat deposition in the abdominal region instead of the peripheral region [46]. In China, the obese population primarily exhibits central obesity despite having a normal BMI, and waist-related indicators more accurately reflect obesity trends in this population [47,48]. In our study, the mean BMI at baseline was close to the normal range (BMI<24), and excessive WC was more common (WC≥90 cm for men and WC≥85 cm for women). This study suggests that physician-pharmacist collaborative models achieved significant improvements in WC and WHR, indicating that patients with diabetes and central obesity may benefit more from collaborative clinics.

In addition, our findings supported that physician-pharmacist collaborative models improve blood pressure, TC, and triglyceride levels and significantly reduce the 10-year ASCVD risk. However, we failed to find a significant enhancement in LDL-C, which is inconsistent with previous studies [49]. This discrepancy may be attributed to the interaction of our intervention focus and population characteristics. First, the primary focus of our multifaceted intervention was comprehensive glycemic control; lipid management was addressed through education but was not part of a protocol-driven intervention. Second, consistent with this context, the enrolled population in these primary care settings had a relatively favorable lipid profile, with mean LDL-C already near the recommended target for low-risk patients (LDL-C [without ASCVD]<2.6 mmol/L) [24,50,51]. Studies suggested that pharmacist interventions have more modest effects on LDL-C in such populations [52]. Overall, the collaborative models led to integrated improvements and facilitated effective, comprehensive health management.

Strengths, Generalizability, and Implementation Considerations

This study has several strengths. First, as one of the few implementation studies on physician-pharmacist collaborative clinics with a long-term follow-up, we developed and evaluated a collaborative model for underresourced settings. A key innovation was the integration of the mHealth component, which enabled continuous support and education, overcoming geographical barriers and facilitating sustained patient engagement beyond traditional, infrequent clinic visits. While this model proved effective in primary care settings in China, its broader application requires consideration of local contexts. Key factors may include the existing health care system, cultural acceptance of team-based care, and digital tools. Second, the physician-pharmacist collaborative clinics potentially bridge the rural-urban gap in medical resources disparities for diabetes management. Moreover, the physician-pharmacist collaborative care model could serve as an evidence-based management model for patients with chronic diseases and a collaborative model for team-based care in underresourced settings. In future implementations, a pilot phase should be given priority so that the core components, structured collaboration and digitally augmented support, can be adapted to local resources and patient needs.

Limitations

This study had several limitations. First, although the rate was comparable between the intervention and control groups, and baseline characteristics were similar between intention-to-treat and per-protocol populations, attrition bias cannot be entirely ruled out. Second, unobserved confounding bias could not be excluded from this study. Potential sources include differences in medication adherence or undisclosed lifestyle changes, as well as unmeasured variations in local health care resources or practice patterns. Third, this study was affected by the COVID-19 pandemic, which reduced patients' willingness to seek care and introduced unpredictable bias. Fourth, the sample size was calculated on the strength of the primary outcome; therefore, we may not have sufficient power to detect other meaningful, significant outcomes. Blinding was not feasible in this study, and a Hawthorne effect may have occurred. Finally, while the final 6 study centers were randomly selected and allocated, the initial identification of candidate centers relied on a pragmatic criterion, which may introduce selection bias.

In conclusion, this trial provides strong evidence to support the implementation of the physician-pharmacist collaborative care model in long-term diabetes treatment and management in primary health care settings in underresourced areas. Patients with T2D, especially those with central obesity or high 10-year ASCVD risk, may benefit more from physician-pharmacist collaborative clinics and should be the focus of further widespread implementation. This theory-based intervention program is a promising multidisciplinary model for improving the quality and efficiency of diabetes treatment and management, as well as other chronic diseases in underresourced settings globally.

Acknowledgments

The authors would like to thank all patients, physicians, and pharmacists for participating in the trial.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Funding

This study was funded by the Chronic Disease Management Research Project of the National Health Commission Capacity Building and Continuing Education Center (YXFSC2022JJSJ010), China Medical Board (grant 18-292), and Hunan Provincial Health Commission (grant 2019158). The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Authors' Contributions

PX, DX, HY, and XL conceptualized and designed this study; YG, HH, DZ, YL, LW, QL, and JL collected the data; JX, QW, SH, and BT performed the statistical analysis; and PX, JX, ST, and LC contributed to the interpretation of the results and drafted the manuscript. All authors have reviewed, revised, and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT checklist.
[PDF File (Adobe PDF File), 209 KB - jmir_v28i1e77470_app1.pdf]

Multimedia Appendix 2
Fidelity checklist.
[PDF File (Adobe PDF File), 195 KB - jmir_v28i1e77470_app2.pdf]

Multimedia Appendix 3
Methods and results.
[PDF File (Adobe PDF File), 469 KB - jmir_v28i1e77470_app3.pdf]

References

1. GBD 2016 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet 2017 Sep 16;390(10100):1211-1259 [FREE Full text] [doi: 10.1016/S0140-6736(17)32154-2] [Medline: 28919117]
2. Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, et al. IDF Diabetes Atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Res Clin Pract 2022 Jan;183:109119 [FREE Full text] [doi: 10.1016/j.diabres.2021.109119] [Medline: 34879977]
3. Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nat Rev Endocrinol 2018 Feb;14(2):88-98. [doi: 10.1038/nrendo.2017.151] [Medline: 29219149]
4. Carlson AL, Mullen DM, Bergenstal RM. Clinical use of continuous glucose monitoring in adults with type 2 diabetes. Diabetes Technol Ther 2017 May;19(S2):S4-11 [FREE Full text] [doi: 10.1089/dia.2017.0024] [Medline: 28541137]
5. Kanaley JA, Colberg SR, Corcoran MH, Malin SK, Rodriguez NR, Crespo CJ, et al. Exercise/physical activity in individuals with type 2 diabetes: a consensus statement from the American College of Sports Medicine. Med Sci Sports Exerc 2022 Feb 01;54(2):353-368. [doi: 10.1249/MSS.0000000000002800] [Medline: 35029593]
6. Nauck MA, Wefers J, Meier JJ. Treatment of type 2 diabetes: challenges, hopes, and anticipated successes. Lancet Diabetes Endocrinol 2021 Aug;9(8):525-544. [doi: 10.1016/S2213-8587(21)00113-3] [Medline: 34181914]
7. Lin CW, Abdul SS, Clinciu DL, Scholl J, Jin X, Lu H, et al. Empowering village doctors and enhancing rural healthcare using cloud computing in a rural area of mainland China. Comput Methods Programs Biomed 2014 Feb;113(2):585-592. [doi: 10.1016/j.cmpb.2013.10.005] [Medline: 24296075]
8. Zhao X, Zhang Y, Yang Y, Pan J. Diabetes-related avoidable hospitalisations and its relationship with primary healthcare resourcing in China: a cross-sectional study from Sichuan Province. Health Soc Care Community 2022 Jul;30(4):e1143-e1156. [doi: 10.1111/hsc.13522] [Medline: 34309097]
9. Wang Y, Li Y, Qin S, Kong Y, Yu X, Guo K, et al. The disequilibrium in the distribution of the primary health workforce among eight economic regions and between rural and urban areas in China. Int J Equity Health 2020 Feb 26;19(1):28 [FREE Full text] [doi: 10.1186/s12939-020-1139-3] [Medline: 32102655]
10. Firkus D, McCoy RG, Matulis J3, Kessler M, Mara K, Herges J. Evaluation of pharmacist consults within a collaborative enhanced primary care team model to improve diabetes care. PLoS One 2023;18(1):e0280654 [FREE Full text] [doi: 10.1371/journal.pone.0280654] [Medline: 36662741]
11. Lee JY, Lee SW. Telemedicine cost-effectiveness for diabetes management: a systematic review. Diabetes Technol Ther 2018 Jul;20(7):492-500. [doi: 10.1089/dia.2018.0098] [Medline: 29812965]

XSL•FO
RenderX

12. Munshi MN, Sy S, Lekarcyk J, Sullivan E. A successful diabetes management model of care in long-term care facilities. J Am Med Dir Assoc 2021 Jun;22(6):1322-6.e2. [doi: 10.1016/j.jamda.2020.06.046] [Medline: 32753320]
13. Anderegg MD, Gums TH, Uribe L, MacLaughlin EJ, Hoehns J, Bazaldua OV, et al. Pharmacist intervention for blood pressure control in patients with diabetes and/or chronic kidney disease. Pharmacotherapy 2018 Mar;38(3):309-318 [FREE Full text] [doi: 10.1002/phar.2083] [Medline: 29331037]
14. Pardo A, Ntabaza V, Rivolta M, Goulard A, Sténuit S, Demeester R, et al. Impact of collaborative physician-pharmacist stewardship strategies on prophylactic antibiotic practices: a quasi-experimental study. Antimicrob Resist Infect Control 2022 Jul 26;11(1):100 [FREE Full text] [doi: 10.1186/s13756-022-01138-3] [Medline: 35883189]
15. Norton MC, Haftman ME, Buzzard LN. Impact of physician-pharmacist collaboration on diabetes outcomes and health care use. J Am Board Fam Med 2020;33(5):745-753 [FREE Full text] [doi: 10.3122/jabfm.2020.05.200044] [Medline: 32989069]
16. Wu S, Tannous E, Haldane V, Ellen ME, Wei X. Barriers and facilitators of implementing interventions to improve appropriate antibiotic use in low- and middle-income countries: a systematic review based on the Consolidated Framework for Implementation Research. Implement Sci 2022 May 12;17(1):30 [FREE Full text] [doi: 10.1186/s13012-022-01209-4] [Medline: 35550169]
17. Miles LM, Hawkes RE, French DP. How the behavior change content of a nationally implemented digital diabetes prevention program is understood and used by participants: qualitative study of fidelity of receipt and enactment. J Med Internet Res 2023 Jan 11;25:e41214 [FREE Full text] [doi: 10.2196/41214] [Medline: 36630165]
18. Jiang H, Mi Z, Xu W. Online medical consultation service-oriented recommendations: systematic review. J Med Internet Res 2024 Jul 30;26:e46073 [FREE Full text] [doi: 10.2196/46073] [Medline: 38777810]
19. Seng JJ, Gwee MF, Yong MH, Kwan YH, Thumboo J, Low LL. Role of caregivers in remote management of patients with type 2 diabetes mellitus: systematic review of literature. J Med Internet Res 2023 Sep 11;25:e46988 [FREE Full text] [doi: 10.2196/46988] [Medline: 37695663]
20. Karloh M, Matias TS, de Oliveira JM, de Lima FF, Araújo Pinheiro DH, Barbosa GB, et al. Breaking barriers to rehabilitation: the role of behavior change theories in overcoming the challenge of exercise-related behavior change. Braz J Phys Ther 2023;27(6):100574 [FREE Full text] [doi: 10.1016/j.bjpt.2023.100574] [Medline: 38056192]
21. Paul B, Kirubakaran R, Isaac R, Dozier M, Grant L, Weller D, RESPIRE Collaboration. A systematic review of the theory of planned behaviour interventions for chronic diseases in low health-literacy settings. J Glob Health 2023 Sep 08;13:04079 [FREE Full text] [doi: 10.7189/jogh.13.04079] [Medline: 37681679]
22. Tan SL, Xiao J, Yuan HY, Chen L, Wang Q, Xiang DX, et al. Establishing and evaluating physician-pharmacist collaborative clinics to manage patients with type 2 diabetes in primary hospitals in Hunan province: study protocol of a multi-site randomized controlled trial in the era of COVID-19 pandemic. BMC Health Serv Res 2022 Mar 04;22(1):299 [FREE Full text] [doi: 10.1186/s12913-022-07653-8] [Medline: 35246117]
23. Bellary S, Kyrou I, Brown JE, Bailey CJ. Type 2 diabetes mellitus in older adults: clinical considerations and management. Nat Rev Endocrinol 2021 Sep;17(9):534-548. [doi: 10.1038/s41574-021-00512-2] [Medline: 34172940]
24. Chinese Diabetes Society. Guidelines for the prevention and treatment of type 2 diabetes in China (2020 edition). Chin J Endocrinol Metab 2021;37(4):311-398 [FREE Full text] [doi: 10.3760/cma.j.cn311282-20210304-00142]
25. Bao Y, Wang C, Xu H, Lai Y, Yan Y, Ma Y, et al. Effects of an mHealth intervention for pulmonary tuberculosis self-management based on the integrated theory of health behavior change: randomized controlled trial. JMIR Public Health Surveill 2022 Jul 14;8(7):e34277 [FREE Full text] [doi: 10.2196/34277] [Medline: 35834302]
26. Luo T, Li MS, Williams D, Fritz J, Beiter K, Phillippi S, et al. A WeChat-based smoking cessation intervention for Chinese smokers: a feasibility study. Transl Behav Med 2022 Nov 16;12(10):1018-1027. [doi: 10.1093/tbm/ibac067] [Medline: 36130313]
27. Yang X, Li J, Hu D, Chen J, Li Y, Huang J, et al. Predicting the 10-year risks of atherosclerotic cardiovascular disease in Chinese population: the China-PAR project (Prediction for ASCVD Risk in China). Circulation 2016 Nov 08;134(19):1430-1440. [doi: 10.1161/CIRCULATIONAHA.116.022367] [Medline: 27682885]
28. Olomu A, Tikaria R, Kelly-Blake K, Hart-Davidson W, Wang L, Alroshood Z, et al. Type 2 diabetes patient activation and mHealth interventions decreased cardiovascular disease risk. Am J Manag Care 2022 Nov 01;28(11):e392-e398 [FREE Full text] [doi: 10.37765/ajmc.2022.89263] [Medline: 36374656]
29. Collins C, Limone BL, Scholle JM, Coleman CI. Effect of pharmacist intervention on glycemic control in diabetes. Diabetes Res Clin Pract 2011 May;92(2):145-152. [doi: 10.1016/j.diabres.2010.09.023] [Medline: 20961643]
30. Gold R, Larson AE, Sperl-Hillen JM, Boston D, Sheppler CR, Heintzman J, et al. Effect of clinical decision support at community health centers on the risk of cardiovascular disease: a cluster randomized clinical trial. JAMA Netw Open 2022 Feb 01;5(2):e2146519 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.46519] [Medline: 35119463]
31. Lumu W, Kibirige D, Wesonga R, Bahendeka S. Effect of a nurse-led lifestyle choice and coaching intervention on systolic blood pressure among type 2 diabetic patients with a high atherosclerotic cardiovascular risk: study protocol for a cluster-randomized trial. Trials 2021 Feb 11;22(1):133 [FREE Full text] [doi: 10.1186/s13063-021-05085-z] [Medline: 33573687]

XSL•FO
RenderX

32. Armitage CJ, Conner M. Efficacy of the theory of planned behaviour: a meta-analytic review. Br J Soc Psychol 2001 Dec;40(Pt 4):471-499. [doi: 10.1348/014466601164939] [Medline: 11795063]
33. Khani Jeihooni A, Layeghiasl M, Yari A, Rakhshani T. The effect of educational intervention based on the theory of planned behavior on improving physical and nutrition status of obese and overweight women. BMC Womens Health 2022 Jan 15;22(1):13 [FREE Full text] [doi: 10.1186/s12905-022-01593-5] [Medline: 35033045]
34. Khani Jeihooni A, Jormand H, Saadat N, Hatami M, Abdul Manaf R, Afzali Harsini P. The application of the theory of planned behavior to nutritional behaviors related to cardiovascular disease among the women. BMC Cardiovasc Disord 2021 Dec 07;21(1):589 [FREE Full text] [doi: 10.1186/s12872-021-02399-3] [Medline: 34876014]
35. Wang L, Peng W, Zhao Z, Zhang M, Shi Z, Song Z, et al. Prevalence and treatment of diabetes in China, 2013-2018. JAMA 2021 Dec 28;326(24):2498-2506. [doi: 10.1001/jama.2021.22208] [Medline: 34962526]
36. Moon J, Kim JY, Yoo S, Koh G. Fasting and postprandial hyperglycemia: their predictors and contributions to overall hyperglycemia in Korean patients with type 2 diabetes. Endocrinol Metab (Seoul) 2020 Jun;35(2):290-297 [FREE Full text] [doi: 10.3803/EnM.2020.35.2.290] [Medline: 32615713]
37. Riddle M, Umpierrez G, DiGenio A, Zhou R, Rosenstock J. Contributions of basal and postprandial hyperglycemia over a wide range of A1C levels before and after treatment intensification in type 2 diabetes. Diabetes Care 2011 Dec;34(12):2508-2514 [FREE Full text] [doi: 10.2337/dc11-0632] [Medline: 22028279]
38. Monnier L, Lapinski H, Colette C. Contributions of fasting and postprandial plasma glucose increments to the overall diurnal hyperglycemia of type 2 diabetic patients: variations with increasing levels of HbA(1c). Diabetes Care 2003 Mar;26(3):881-885. [doi: 10.2337/diacare.26.3.881] [Medline: 12610053]
39. Lehn JM, Gerkin RD, Kisiel SC, O'Neill L, Pinderhughes ST. Pharmacists providing palliative care services: demonstrating a positive return on investment. J Palliat Med 2019 Jun;22(6):644-648. [doi: 10.1089/jpm.2018.0082] [Medline: 30570421]
40. Matzke GR, Moczygemba LR, Williams KJ, Czar MJ, Lee WT. Impact of a pharmacist-physician collaborative care model on patient outcomes and health services utilization. Am J Health Syst Pharm 2018 Jul 15;75(14):1039-1047. [doi: 10.2146/ajhp170789] [Medline: 29789318]
41. Aguiar PM, da Silva CH, Chiann C, Dórea EL, Lyra DP, Storpirtis S. Pharmacist-physician collaborative care model for patients with uncontrolled type 2 diabetes in Brazil: results from a randomized controlled trial. J Eval Clin Pract 2018 Feb;24(1):22-30. [doi: 10.1111/jep.12606] [Medline: 27452158]
42. Hinds A, Lopez D, Rascati K, Jokerst J, Srinivasa M. Adherence to the 2013 blood cholesterol guidelines in patients with diabetes at a PCMH: comparison of physician only and combination physician/pharmacist visits. Diabetes Educ 2016 Apr;42(2):228-233. [doi: 10.1177/0145721716631431] [Medline: 26902526]
43. Peterson J, Hinds A, Garza A, Barner J, Hill L, Nguyen M, et al. Impact of physician-pharmacist covisits at a primary care clinic in patients with uncontrolled diabetes. J Pharm Pract 2020 Jun;33(3):321-325. [doi: 10.1177/0897190018807374] [Medline: 30428760]
44. Xiao J, Wang Q, Tan S, Chen L, Tang B, Huang S, et al. Analysis of patient medication compliance and quality of life of physician-pharmacist collaborative clinics for T2DM management in primary healthcare in China: a mixed-methods study. Front Pharmacol 2023 Mar 24;14:1098207 [FREE Full text] [doi: 10.3389/fphar.2023.1098207] [Medline: 37033638]
45. Ledford JL, Hess R, Johnson FP. Impact of clinical pharmacist collaboration in patients beginning insulin pump therapy: a retrospective and cross-sectional analysis. J Drug Assess 2013;2(1):81-86 [FREE Full text] [doi: 10.3109/21556660.2013.815624] [Medline: 27536441]
46. Du T, Sun X, Yin P, Huo R, Ni C, Yu X. Increasing trends in central obesity among Chinese adults with normal body mass index, 1993-2009. BMC Public Health 2013 Apr 10;13:327 [FREE Full text] [doi: 10.1186/1471-2458-13-327] [Medline: 23575244]
47. Wu J, Zou L, Liu Y, Yu H, Yin H, Zhong L, et al. Association between obesity indicators and cardiometabolic disease in Chinese adults. PLoS One 2023 Jan 20;18(1):e0273235 [FREE Full text] [doi: 10.1371/journal.pone.0273235] [Medline: 36662790]
48. Jia Z, Zhou Y, Liu X, Wang Y, Zhao X, Wang Y, et al. Comparison of different anthropometric measures as predictors of diabetes incidence in a Chinese population. Diabetes Res Clin Pract 2011 May;92(2):265-271. [doi: 10.1016/j.diabres.2011.01.021] [Medline: 21334088]
49. Elnaem MH, Rosley NF, Alhifany AA, Elrggal ME, Cheema E. Impact of pharmacist-led interventions on medication adherence and clinical outcomes in patients with hypertension and hyperlipidemia: a scoping review of published literature. J Multidiscip Healthc 2020;13:635-645 [FREE Full text] [doi: 10.2147/JMDH.S257273] [Medline: 32764955]
50. Ruixing Y, Qiming F, Dezhai Y, Shuquan L, Weixiong L, Shangling P, et al. Comparison of demography, diet, lifestyle, and serum lipid levels between the Guangxi Bai Ku Yao and Han populations. J Lipid Res 2007 Dec;48(12):2673-2681 [FREE Full text] [doi: 10.1194/jlr.M700335-JLR200] [Medline: 17890682]
51. Deng B, Luo T, Huang Y, Shen T, Ma J. Prevalence and determinants of hyperlipidemia in moderate altitude areas of the Yunnan-Kweichow plateau in Southwestern China. High Alt Med Biol 2012 Mar;13(1):13-21 [FREE Full text] [doi: 10.1089/ham.2011.1037] [Medline: 22429228]

XSL•FO
RenderX

52. Dixon DL, Baker WL, Buckley LF, Salgado TM, Van Tassell BW, Carter BL. Effect of a physician/pharmacist collaborative care model on time in target range for systolic blood pressure: post hoc analysis of the CAPTION trial. Hypertension 2021 Sep;78(4):966-972 [FREE Full text] [doi: 10.1161/HYPERTENSIONAHA.121.17873] [Medline: 34397278]

Abbreviations

ASCVD: atherosclerotic cardiovascular disease
CONSORT: Consolidated Standards of Reporting Trials
FBG: fasting blood glucose
GEE: generalized estimating equation
HbA_{1c}: glycosylated hemoglobin
LDL-C: low-density lipoprotein cholesterol
mHealth: mobile health
PBG2h: 2-hour postprandial blood glucose
T2D: type 2 diabetes
TC: total cholesterol
TPB: theory of planned behavior
WC: waist circumference
WHR: waist-to-hip ratio

Edited by A Stone; submitted 13.May.2025; peer-reviewed by G Ni, Y Shin; comments to author 02.Jan.2026; accepted 18.Feb.2026; published 13.Mar.2026.

Please cite as:
Xiao J, Wang Q, Tan S, Chen L, Xiang D, Yuan H, Li X, Huang S, Tang B, Guo Y, Huang H, Zhao D, Li Y, Wang L, Li Q, Liu J, Xu P
Effect of a Digital-Driven Physician-Pharmacist Collaborative Model for Diabetes in Primary Health Care: Cluster Randomized Trial
J Med Internet Res 2026;28:e77470
URL: https://www.jmir.org/2026/1/e77470
doi:10.2196/77470
PMID:

©Jie Xiao, Qing Wang, Shenglan Tan, Lei Chen, Daxiong Xiang, Haiyan Yuan, Xia Li, Shuting Huang, Bingjie Tang, Yan Guo, Haiying Huang, Danhui Zhao, Yue Li, Li Wang, Qun Li, Juan Liu, Ping Xu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Coproducing an Online Platform for People With Long-Term Physical Health Conditions: Development and Usability Study

Hannah Grace Jones[1,2,3*], BSc; Grace Lavelle[1,4*], PhD; Elly Aylwin-Foster[1], BA; Ewan Carr[5], PhD; Alan Simpson[6,7], PhD; Matthew Hotopf[1], PhD; Vanessa Lawrence[7], PhD

[1]Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[2]Lifelong Health, Persistent Pain Research Group, Hopwood Centre for Neurobiology, South Australian Health and Medical Research Institute, Adelaide, South Australia, Australia

[3]IIMPACT (Innovation, Implementation and Clinical Translation) in Health, School of Allied Health and Human Performance, Adelaide University, Adelaide, South Australia, Australia

[4]CP-Life Research Centre, School of Physiotherapy, Royal College of Surgeons in Ireland, Dublin, Leinster, Ireland

[5]Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[6]Florence Nightingale Faculty of Nursing, Midwifery & Palliative Care, King's College London, London, United Kingdom

[7]Health Service and Population Research, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Hannah Grace Jones, BSc
Department of Psychological Medicine
Institute of Psychiatry, Psychology and Neuroscience
King's College London
16 De Crespigny Park
London, SE5 8AB
United Kingdom
Phone: 44 02078480002
Email: hannah.g.jones@kcl.ac.uk

Abstract

Background: There is relatively limited psychological support dedicated to people living with long-term physical health conditions and subthreshold depressive disorder. Online peer support may be an appropriate intervention to help bolster patients' mental well-being to prevent progression of their symptoms to major depressive disorder. For interventions to be successfully integrated into the self-management routines of people with long-term physical health conditions, they should be co-designed to ensure that they align with the wants and needs of the target audience.

Objective: This study aims to coproduce an online peer support intervention with people with lived experience, software experts, clinicians, and academics through an iterative process of co-design and subsequent co-validation through usability testing.

Methods: We followed a 4-stage coproduction process: co-assess, co-design, co-validate, and co-deliver. Our research advisory group was actively involved in all stages, consisting of 1 coinvestigator and 6 people with lived experience of long-term physical and/or mental health comorbidities. The co-assess and co-design stages involved our participatory design panel, which included 10 members living with various long-term conditions. The participatory design panel participated in online focus groups to assess their unmet psychosocial needs and then co-designed the intervention prototype through online workshops with software developers. The co-validation stage involved an additional group of participants (n=12) with long-term physical health conditions. During co-validation, the prototype underwent usability testing, including think-aloud exercises and semistructured interviews. Content analysis identified the priorities for the iterative development that formed the basis of further research advisory group co-design workshops. The next stage, co-delivery, involved coproducing the protocol of a feasibility and acceptability randomized controlled trial.

Results: Participants highlighted that a platform must feel safe and trustworthy for the space to support the mental well-being of those living with long-term health conditions. The participatory design panel co-designed a platform prototype to meet this need. During the co-validation stage, the think-aloud exercises identified common issues related to navigation challenges and feature glitches. Content analysis of the semistructured interviews confirmed that the community forum, resources, and other

XSL•FO
RenderX

platform pages were appropriate and acceptable, but revealed usability concerns. Participants stressed the need for intuitive navigation and suggested new features that would enhance user experience. Facilitators and barriers to engagement were also noted, including the importance of fostering trust in the platform's ethos and branding to create a safe space. Through iterative development and subsequent usability testing, the final prototype was approved.

Conclusions: We have provided a worked example of a comprehensive, coproduction process where we worked alongside people with lived experience to successfully design an online peer support platform with embedded psychoeducation. The platform, called CommonGround, is ready to be evaluated in a feasibility randomized controlled trial.

(J Med Internet Res 2026;28:e79666) doi:10.2196/79666

KEYWORDS

long-term conditions; peer support; coproduction, physical health; mental health; psychosocial support; online peer support; usability testing

Introduction

An estimated 15 million people are living with a long-term physical health condition in the United Kingdom [1]. Up to 20% are living with comorbid major depressive disorder [2,3], and a further 20% are experiencing subthreshold depressive disorder [4]. Subthreshold depressive disorder is characterized by experiencing between 2 and 5 depressive symptoms for at least 2 weeks, one of which must be either low mood or loss of interest/pleasure [5,6]. Although living with subthreshold depressive disorder is a key risk factor for developing major depressive disorder [7,8], the provision of psychological support to help prevent the worsening of these existing depressive symptoms among people with long-term physical health conditions is relatively limited. Targeted approaches to support the mental well-being of this at-risk population are therefore urgently needed.

A critical component in the effective management of long-term physical health conditions is self-management, whereby patients take an active role in the day-to-day management of their symptoms, treatments, and necessary lifestyle modifications [9,10]. As such, in the development of any tool to support the well-being of people with long-term physical health conditions, it is critical that intervention design and implementation recognize this need to self-manage effectively and should not add undue burden [11,12]. Indeed, empowering patients to address their own needs by providing tools that support their self-management has been associated with improved health outcomes [13,14] and health behaviors [15].

One potential solution to support patients' self-management of their psychosocial well-being could be peer support, where people with similar life experiences or characteristics (ie, living with a long-term physical health condition) are able to share emotional, informational, and appraisal support with one another to improve their physical and psychosocial health [16-18]. Peer support can be delivered in a variety of formats, from one-to-one to group-based support, and in-person or online. Recently, the provision of online peer support has increased, as it is typically low-cost, easily scalable, and offers a flexible, convenient solution that overcomes the physical barriers people may face in accessing in-person groups [19]. The positive effects of engaging with online peer support resemble those of offline peer support [20], including reciprocal emotional and social support, patient empowerment, improved quality of life, and mental well-being [17,21-24].

Online peer support interventions therefore have the potential to reduce the burden of long-term conditions and the associated psychological distress [16] but also reduce the burden of self-management by improving the tools one has available to self-manage [25]. However, to date, the development of online peer support interventions has largely been driven by academic researchers and industry with limited input from the intended end user [26]. As a result, interventions may not meet end users' needs and perspectives, and be inappropriately designed, with missing features or poor interfaces [27]. Such poor design can ultimately limit users' long-term motivation to engage with an intervention, for instance, if they encounter difficulties in trying to log-in [27]. The lack of involvement of end users in design is particularly problematic when the end users are people living with long-term physical health conditions, as this population has complex needs and requires interventions that can be easily incorporated into their existing self-management routines [28]. Indeed, in our prior work, people living with a variety of long-term physical health conditions emphasized that this would be essential to ensure that the intervention feels safe, credible, and appropriate for their needs [29]. Coproduction is therefore vital, whereby an online peer support platform is created through an equal partnership between lived experience experts, researchers, and other partners with the underlying ethos of developing *with* and not *for* the users [30].

Coproduction can enhance the relevance, acceptability, and usability of interventions, thereby facilitating patients' long-term engagement. The diversity of perspectives fosters innovative and creative interventions, ensuring that features and functions are uniquely designed and rooted in lived experience [27,31]. A critical aspect of coproduction is ensuring the perspectives of lived experience experts have been accurately translated into components of the intervention [27]. One means of ensuring this is by evaluating the usability and acceptability of the co-designed prototype and allowing time for further iterative development as required [32]. Think-aloud exercises are valuable in usability testing for identifying specific usability issues in common user tasks, for instance, glitches encountered when creating an account or creating a forum post [33]. The analysis of think-aloud exercises can also reveal unmet user needs and highlight areas for improvement, guiding future prototype modification [34,35]. To assess acceptability,

XSL•FO
RenderX

qualitative interviews are useful as users' perceptions of acceptability are deeply contextualized (eg, in relation to their long-term physical health condition) and shaped by social and cultural factors [36]. Together, the findings from usability testing can inform the iterative development of the intervention and its ability to satisfy user needs. Ultimately, this process leads to a final approved prototype, which can then undergo real-world evaluation of safety and efficacy in randomized controlled trials.

Although the benefits of coproduction are widely recognized, there is a significant lack of research evidence documenting the coproduction of real-world interventions [37], particularly those dedicated to supporting psychological well-being in the context of physical health [38]. In this paper, we report how we coproduced "CommonGround," our online peer support and psychoeducation platform dedicated to people living with long-term physical health conditions. This study aimed to coproduce an online peer support platform with embedded psychoeducation for people living with long-term conditions and subthreshold depression to support their psychological well-being. To fulfill this aim, the following objectives were set: (1) to identify the psychosocial needs of people living with long-term physical health conditions, (2) to co-design an online peer support platform prototype that aligns with their wants and needs, and (3) to co-validate the prototype through usability testing and iterative development work to produce a final platform prototype that can be evaluated in a future feasibility randomized controlled trial.

Methods

Design

The CommonGround project followed a 4-stage coproduction framework adapted from Elwyn and colleagues [39] comprising (1) co-assess, (2) co-design, (3) co-validate, and (4) co-deliver (Figure 1).

Figure 1. The coproduction cycle of CommonGround, adapted from Elwyn et al [39]. RCT: randomized controlled trial.

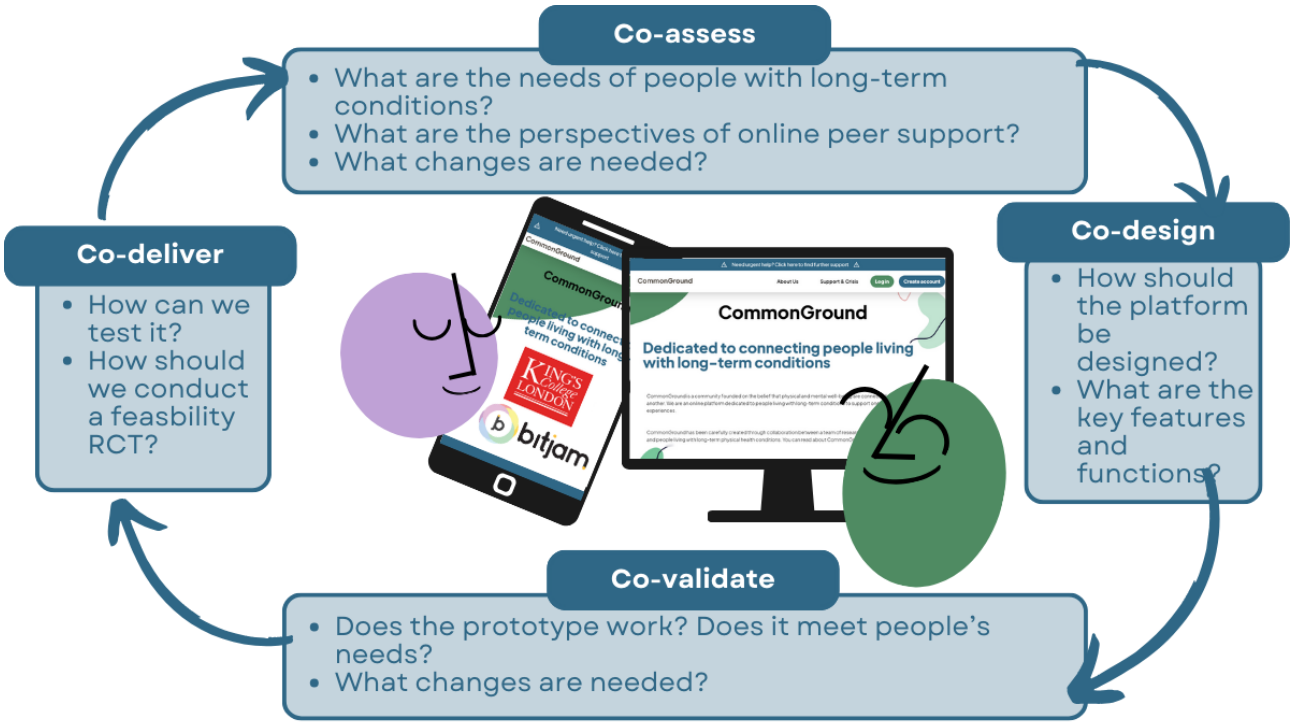

Ethical Considerations

Full ethical approval for the online focus groups and usability testing was sought and granted by King's College London Research Ethics Office, PNM Research Ethics Subcommittee (HR-19/20-14938). Electronic informed consent was obtained before taking part in the relevant research activities (see Multimedia Appendix 1). The procedures to maintain participant confidentiality and anonymity were outlined in the consent forms. All data were pseudoanonymized using unique study IDs, password-protected, and stored on secure servers. Audio-recordings from focus groups and interviews were shared securely with a transcription service, who have signed nondisclosure agreements and deleted audio-recordings immediately following transcription. Participants received a £20 (US $26.9) voucher for participation in the focus groups. The same reimbursement was provided for participation in the usability testing.

Patient and Public Involvement

At each stage of the coproduction cycle, different stakeholders and groups of people with lived experience were involved (Figure 2). One of the authors and a coinvestigator on the project has lived experience of multiple long-term health conditions (cystic fibrosis, type 1 diabetes, and adrenal insufficiency) and, throughout the project, has been a core member of the research team and remunerated for their time. Their role included chairing our independent research advisory group (RAG) of people with lived experience who guided the whole research process, including the coproduction of study documents and key decision-making processes. Our RAG involved 6 members (gender identity: 4 women and 2 men; age range: 31-71 years;

XSL•FO
RenderX

ethnicities: Black British, Caribbean, White British, and White European) living with various long-term physical health conditions. Three members were living with multiple long-term conditions, and 4 had comorbid mental health diagnoses. RAG members were identified and invited to join through our coinvestigator with lived experience and through patient representative groups based within Guy's and St Thomas Hospital and King's College Hospital, and King's College London's Integrated Care Consultation Partners Group, and via snowballing through recruited members. All recruitment materials emphasized the focus on preventing depression in the context of chronic physical illness, and any individual that felt that they had relevant lived experience regarding psychological distress and well-being in the context of their physical health was able to take part (ie, with or without a current or past clinical diagnosis of depression; no formal assessment of depressive symptoms was administered). This inclusive approach not only sought to be representative of a range of lived experiences but was also sensitive to the fact that many people may not recognize that they are experiencing subthreshold depressive symptoms and that these symptoms can also fluctuate over time. Terms of reference were agreed in accordance with UK standards for public involvement [40]. Our RAG members were reimbursed at an hourly rate based on patient and public involvement (PPI) reimbursement recommendations for their time preparing for meetings, attending meetings, and reviewing study documents [41,42]. Members were invited to share feedback outside of any meetings according to their personal preference (eg, via email).

Figure 2. The different stakeholders and groups of people with lived experience involved in the CommonGround coproduction cycle. LTC: long-term condition; RAG: research advisory group.

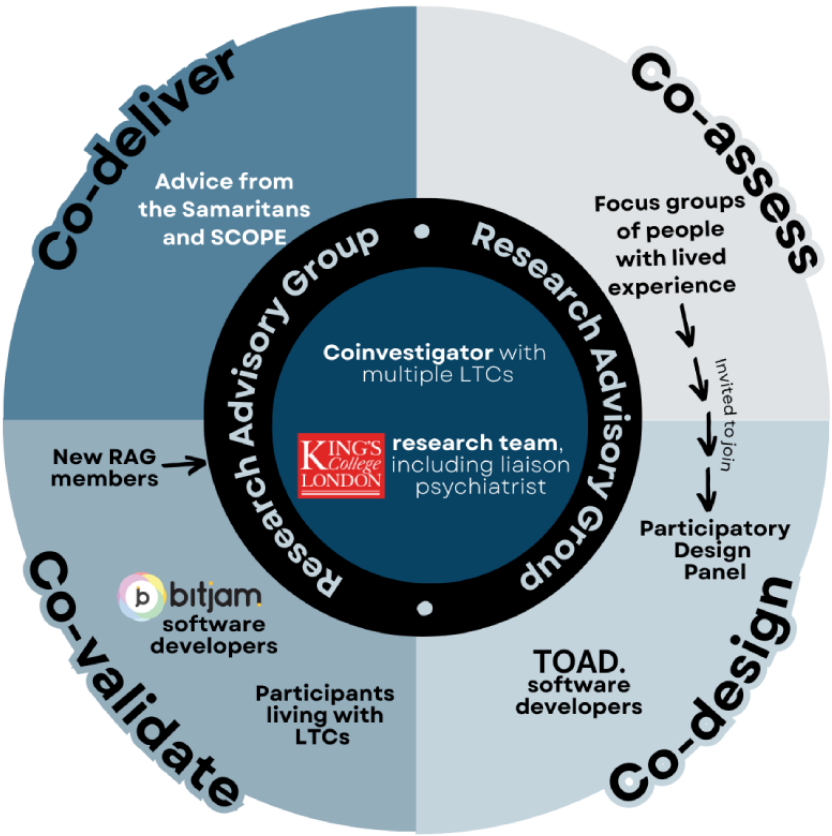

The Coproduction Cycle

The coproduction cycle had 4 stages: co-assess, co-design, co-validate, and co-deliver (Figure 1) [39], which are outlined below.

Co-Assess

The co-assess stage addressed the question: "What are the unmet psychosocial needs of people living with long-term physical health conditions, and how might an online peer support platform address these needs?"

Three online focus groups were conducted, following a topic guide codeveloped with our RAG. Reflexive thematic analysis was conducted using a critical-realist approach [43,44]. The full methodology and findings for this phase have been published elsewhere [29] and laid the foundations for co-designing the online peer support platform. In brief, online focus groups followed a semistructured interview schedule and were transcribed verbatim for analysis. Early stages of the focus group explored their psychosocial needs and how these were linked to their psychological well-being before exploring how these identified needs could be addressed. Participants were asked if, why, and how an online community that connects people living with a variety of long-term physical health conditions would be something that they would use to support the unmet psychosocial needs they identified in the earlier stages. Transcripts were analyzed using reflexive thematic analysis,

XSL•FO
RenderX

adopting a critical-realist approach. Following the analysis, the 10 participants who took part in these interviews were invited to be lived experience experts as part of the participatory design panel alongside other stakeholders, for the co-design stage (described below).

Co-Design

The co-design stage involved the participatory design panel, which was made up of an external software development agency (TOAD), researchers and clinicians from King's College London (KCL), our coinvestigator with lived experience, and lived experience experts from the focus groups (n=10) in the co-assess stage [29]. The lived experience experts were originally recruited from King's College London's Integrated Care Consultation Partners Group, the Guy's and St Thomas' PPI group, and the King's College Hospital PPI group. Many were living with multiple conditions, and a range of musculoskeletal, neurological, cardiovascular, respiratory, gastrointestinal, gynecological, autoimmune, and pain-related disorders were represented by the lived experience experts of our participatory design panel. The majority (8/10, 80%) of the lived experience experts used technology daily; however, only 30% (3/10) had used internet support groups before. No formal assessment of depressive symptoms was administered to participants, but recruitment materials emphasized the focus on preventing depression in the context of chronic illness and invited people who felt that they had relevant lived experience of low mood and psychological distress (current or past, with or without a formal clinical diagnosis) to participate. This inclusive, open approach was considered critical in this coproduction process, particularly with regard to identifying a breadth of different perspectives and experiences.

Six online workshops were conducted over 6 weeks, each lasting approximately 2 hours. The KCL team and software developers guided the workshops following specific agendas to maximize the discussions' applicability for creating the prototype (see Table 1). The panel was invited to share any additional suggestions after the sessions via email. The software developers incorporated feedback from each workshop to create the online peer support platform prototype.

Table 1. Agenda of the 6 development workshops with the participatory design panel to co-design the platform prototype.

Workshop	Agenda
1: How will the platform help people?	This workshop adopted a "users first" approach. This involved exploring questions such as "What relationships in your life are important to you, and how do you navigate them?" and "Have you looked for information on living with a long-term health condition?"
2: What are users' needs and wants for the platform?	This workshop focused on assessing user needs in a features-driven approach. This included brainstorming the features and functions that users would like the platform to have. The associated features or factors that may enhance the benefit of and/or reduce barriers of using the platform were also explored.
3: Platform features and solutions	Establishing the core features and how these should be prioritized in the development of the prototype. This involved asking people to rank their preference for different types of social interaction on a platform, from posting content anonymously to commenting in response to other content, and whom they prefer to socially interact with (eg, people with other long-term conditions, whoever is online at that time, and people in close geographical proximity).
4: Psychoeducation and moderation	Identify the type of educational content that the platform should host. Also identify key moderation principles and rules for communication on the platform by asking "What is the platform?" versus "What the platform isn't?"
5: What is the creative concept of the platform?	Three different creative concepts (eg, color, schemes, and layouts) were presented. Aimed to discuss what will work best and why, to allow further development of the platform features and creative brand.
6: Finalizing the creative bread and prototype	Finalizing the creative brand and prototype, including refining the brand position and name, alongside the visual look and feel.

Co-Validate

Overview of Co-Validation Stage

In the co-validation stage, we sought to review and finalize the platform prototype by answering the following questions: (1) "Have we correctly translated what people wanted into the platform itself?" (2) "Does everything work as we expect it to?" (3) "What, if any, improvements are needed?". This was an iterative process that involved the following steps (see Figure 3): (1) usability testing: think-aloud task and semistructured interviews, (2) iterative platform development, and (3) further prototype testing, final development work, and final prototype approval.

XSL•FO
RenderX

Figure 3. The steps involved in the co-validation stage of the CommonGround project. KCL: King's College London; LTC: long-term condition; RAG: research advisory group; UT: usability testing.

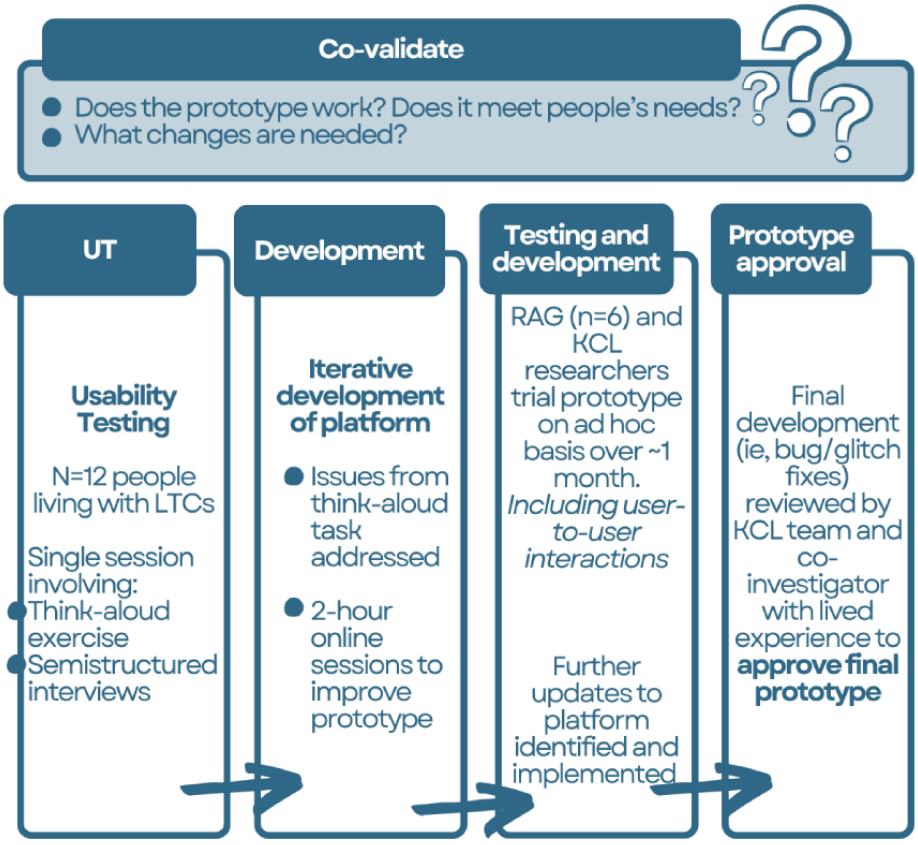

Usability Testing: Think-Aloud Task and Semistructured Interviews

Overview

The usability testing involved participants with different long-term health conditions who had no prior involvement in the project. Participants were recruited through national charities and were reimbursed with a £20 (US $26.6) voucher; ethical approval received from King's College Research Ethics Committee: HR-19/20-14938). The usability testing sessions were conducted via Zoom during the COVID-19 pandemic, as the planned in-person sessions were not possible. Each session included the participant, a research assistant, and a member of the software team. Participants screenshared their view of the platform throughout the session, which involved (1) a think-aloud session (~30 minutes [34]) and (2) a semistructured interview (30-60 minutes).

Think-Aloud Session

Participants were asked to verbalize their thoughts as they explored the platform prototype and worked through a series of tasks that a typical user may complete (see Multimedia Appendix 2 for think-aloud activity tasks). For example, they had to create an account, voicing everything that they were thinking as they read through the pages and completed each step. If participants encountered issues in completing a task, the research assistant and software developer encouraged them to troubleshoot themselves, rather than providing explicit guidance on how to overcome the issues.

Semistructured Interviews

Participants were then interviewed to explore views on the design and usability of the platform prototype, and potential facilitators and barriers to accessing or using peer support platforms in general. The topic guide (see Multimedia Appendix 3) included questions like "How did you find navigating the platform and finding what you were looking for?" and "Was there anything you looked for on the platform that you could not find?" Interviews were conducted until data saturation was reached: when no new themes or insights were emerging during data collection [45].

All usability testing sessions were audio-recorded and transcribed verbatim. The anonymized transcripts were analyzed in NVivo (version 14, released 2023; Lumivero) using content analysis [46] to organize participants' views and experiences into 5 predetermined themes and draw realistic conclusions from the data that could be implemented through further co-design workshops to finalize the prototype. Content analysis involved an iterative process of systematically organizing and condensing the data into a coding framework involving the overarching themes of (1) features and functions, (2) usability, (3) branding, (4) barriers and facilitators, and (5) ethos and values. These overarching themes were selected through discussions with our RAG for their ability to organize the data in accordance with the overall objective of coproducing an online peer support platform. Within the themes, codes were developed inductively in response to emerging patterns on a manifest level (ie, taking the meaning of the text at what the text says). Three coders (HGJ, GL, and VL) independently coded 3 transcripts before meeting to ensure consistency in

XSL•FO
RenderX

interpretations and make final updates to the coding framework. Once the final coding framework was confirmed, the coders analyzed the remaining data. Bugs and glitches were identified (and coded under a separate code) by extracting statements where participants expressed difficulties or problems when completing each think-aloud task.

Iterative Platform Development

First, the software developers from TOAD resolved the glitches and bugs identified in the think-aloud task analysis. The software code was then transferred to the software developers at Bitjam for the remainder of the coproduction. Through a series of 7 two-hour online discussion sessions, the KCL team, our coinvestigator with lived experience, and the RAG agreed on the developments required to improve the existing prototype, and designed the new features and functions that could solve issues or fulfill unmet needs identified in the first stages of usability testing. Based on the RAG discussions, the KCL team prepared user stories and visual mock-ups for all proposed developments to ensure the vision of the RAG was implemented correctly by Bitjam. In these discussions, safety and moderation policies, alongside the related functional features on the platform, were also finalized.

Further Prototype Testing, Final Development Work, and Prototype Approval

Following the iterative development of the platform, a review of the refined prototype was completed by the RAG (who had not accessed the "live" platform during development sessions) and KCL researchers (who had no other involvement in this project). This testing occurred over a 1-month period, where testers were encouraged to interact with the prototype as they would with existing social media platforms or forums, logging in on multiple occasions, exploring all the available features, and interacting with other testers. This usability testing aimed to review the new features and functions and identify any bugs or glitches that occur when multiple users are simultaneously accessing the platform and interacting with each other. All testers shared written feedback about bugs and glitches from which the required edits were extracted. The final edits were then implemented by Bitjam. The final prototype then underwent extensive testing by our research assistant and coinvestigator with lived experience before final platform approval was granted by the KCL team.

Co-Deliver

Following the successful codevelopment of the peer support platform, the next stages involved coproducing a protocol for a feasibility and acceptability randomized controlled trial. The co-deliver stage involved online group discussions with our RAG and coinvestigator with lived experience to guide study design, the review of study documents, questionnaire testing, and cocreating our adverse events protocol. The Samaritans and SCOPE charity organizations were also consulted with regard to moderation and safeguarding policies, and in the development of moderator training. The protocol of the feasibility and acceptability randomized controlled trial is available elsewhere [47].

Results

Co-Assess

The full findings of this work that laid the foundations for co-designing our online platform have been published elsewhere [29]. In brief, findings from 3 online focus groups identified 3 key themes regarding how an online peer support platform could address the unmet psychosocial needs of people living with long-term physical health conditions: (1) relationship between self and the outside world, (2) past experiences of peer support, and (3) philosophy and vision of peer support. Although participants spoke of their mind and body being entwined, they felt this was often neglected in their health care, with their mental and physical health treated as separate entities. The need for their mental health to be supported in the context of their physical health was clear to them. Participants felt that peer support offered both an opportunity to share knowledge and a sense of mutual validation, particularly as people with similar experiences are best placed to understand their circumstances. However, it was clear that online peer support platforms needed to be safe and credible spaces and that to achieve this, platforms must be co-designed with the people who will ultimately use them. Participants also explored the potential benefits and concerns of using online communities, which helped to shape the ethos and philosophy of our platform. Online communities were valued for being easily accessible and easy to participate in; however, it was considered key that a "code of conduct" should be shared and enforced for the space to feel safe and credible. Although many participants were accustomed to the transitions of face-to-face activities to online settings, there were concerns related to how digital interactions may be impaired if equipment or technology does not operate properly. Together, these findings provided the basis for the following co-design stage.

Co-Design

A prototype of the peer support and psychoeducation platform was successfully developed in the co-design phase with the following features and functions:

- A landing page that introduced the platform
- The community principles that govern the community, presented as "terms of use" that people must agree to before registration
- A registration page, where individuals could create their account and set up their profile
- The community feed which hosted a forum page where users could create posts, add comments, and react to posts
- A resource page where users could view, download, react to, and comment on resources relevant to people living with long-term physical health conditions
- The ability for users to save community feed posts and resources to their private "My Garden" page

The web-based prototype was accessible on desktop and mobile devices. During the co-design phase, the platform name "CommonGround" and the "all diagnoses welcome" ethos were developed. This ethos and related manifesto created by our coinvestigator with lived experience (see Multimedia Appendix 4) reflects the view that people living with any long-term

XSL•FO
RenderX

physical health condition have shared lived experiences that they can relate to, from adapting their work or social lives to taking medication and daily self-management. The slogan "share, support, learn, cultivate" was also developed to reflect the main purposes of CommonGround. The term "cultivate" was later replaced with "grow" as our RAG felt this was a more familiar term and better reflected the CommonGround ethos of supporting someone to grow on their self-management journey.

Co-Validate

Usability Testing: Think-Aloud Task and Semistructured Interviews

Overview of Findings

The first stage of usability testing included 12 people living with at least one long-term physical health condition (see Table 2) who had not been involved in the co-assess or co-design stages. The range of diagnoses included asthma, chronic kidney disease, diabetes, epilepsy, fibromyalgia, hypertension, multiple sclerosis, myalgic encephalomyelitis, and rheumatoid arthritis. Participants were aged 25-71 years (median 53.5, IQR 29 years), had access to their own internet-connected device, had internet at home, and used the internet daily. Five (42%) participants had used internet support groups before.

Table 2. Characteristics of participants involved in usability testing through think-aloud exercise and interviews (N=12).

Characteristics	Values, n (%)
Gender identity	
Woman, including transgender woman	11 (92)
Man, including transgender man	1 (8)
Number of participants living with multiple long-term conditions	
1 long-term condition	5 (42)
2 long-term conditions	4 (33)
3+ long-term conditions	4 (33)
Number of participants living with comorbid mental health conditions	3 (25)
Do you have access to the internet at home?	
Yes	12 (100)
No	0 (0)
How frequently do you use the internet?	
Daily	12 (100)
Weekly	0 (0)
Monthly	0 (0)
Devices used to access the internet? (more than one device could be selected)	
Laptop	7 (58)
Desktop	1 (8)
Phone	7 (58)
Tablet device	3 (25)
Have you used internet support groups before?	
Yes	5 (42)
No	7 (58)

From the think-aloud task, explicit verbal indications of bugs and glitches were successfully identified and consolidated (see Table 3 for full results). In total, 9 usability issues were identified. The most common usability issues were related to navigation, layout, understanding instructions, and the multiple steps required to complete a task. The most usability issues were encountered for the "create an account" task. Across ages, gender, and long-term physical health conditions, experiences of issues were similar, with the majority of issues experienced by most participants.

XSL•FO
RenderX

Table 3. Think-aloud tasks and usability testing results for each task.

Usability issues	Description of issues	Example Quotes
Task 1: browse around for 1-2 minutes, providing your feedback [logged out state]		
Navigation	Users had difficulty moving around the home page. Some experienced a sticky scroll. The arrow at the bottom of the page that intended to move the user down did not function properly.	• "So, basically, it doesn't scroll. It doesn't scroll so easy. I think when, when I put in my cursor on the blank space on the side, then it tends to scroll a bit. So, it's a bit erratic, the way it scrolls." [Participant 1] • "Once you click to go back up it's, kind of, doing that thing that if you try and go to go down with the arrow it, kind of, keeps doing that jump thing." [Participant 8]
Layout	Users could not read all the text as some was obscured or not displayed properly.	• "See, I can't read, I can't read that because the pages kind of separates, and you're kind of missing a little bit of something there." [Participant 1] • "The first thing I've noticed, I can't read the top line there." [Participant 6]
Task 2: create an account		
Navigation	Users had difficulty in identifying the button to create an account, often selecting the wrong button.	• "Right, so where, where do I go to join an account, to make an account?" [Participant 2] • "Where do I create an account? Or do I have to login? Right. So, do...? Am I using? Do I put me email address in and stuff?" [Participant 3]
Understanding instructions	Users were confused about what to input for various fields of the registration form and desired further instructions.	• "I think what's really interesting there is it wasn't really clear for you where to start. It's the question's so broad. So, we can look at, you know, maybe having a couple of kind of tips underneath or something." [Participant 1]
Glitchy function	Users struggled with the calendar date selector when inputting date of birth.	• "That was a bit tricky writing down a date of birth... but I did about...I think it was the third time I typed it in. Like I said, here I think a drop-down menu would be easier." [Participant 7] • "Mm, don't even know how to put the date in." [Participant 11]
Workflow	Users often missed a step in the registration form, having to review closely to resolve the issue and proceed.	• "What's that? I've got a hand, so that's saying that's a live link, but it's not allowing me to (click)...Oh, okay. So I've gotta put something else then?" [Participant 7]
Task 3: browse around for 1-2 minutes, providing your feedback [logged in – community feed]		
Visuals	Users were unsure what some of the graphics and labels of buttons meant, and it created confusion about the potential functions of different buttons.	• "I'm not sure what the Add to Garden means. I might be a bit dumb [laughing] if I don't know, but I'm not sure, um, about that, if I'm honest." [Participant 12] • "The, you know, the things you have, the images that you have above the post a thought, share knowledge, ask a question, I was just wondering what that is, what those images depict." [Participant 1]
Task 4: navigating between pages for tasks: (1) find a psychoeducational resource and view it and (2) view posts from other community members		
Navigation	Users often did not notice the navigation bar as it did not stand out on the page. This limited their ability to switch between pages, including to find a resource or to go view their "My Garden."	• "Okay. Okay, so it took me a while to see that thing at the top where I've got the options for, Community Resources In "My Garden". Um, I didn't see that until I'd scrolled all the way down to the bottom." [Participant 11] • "I think it's because I completely didn't realise that those were up there. If I maybe did. Maybe if they were bigger. I'm, I'm trying to...I don't know what other things you could do to, um, make it more noticeable." [Participant 12]
Task 5: create a post		

XSL•FO
RenderX

Usability issues	Description of issues	Example Quotes
Workflow	Some users had trouble in completing the different fields for creating a post.	• "Mmm. I don't know. Is there a way to, sorry, I don't know how to work this. Is there a way to like, type to here, or..." [Participant 4]
Task 6: view posts from other community members		
Navigation	Some users did not know how to comment on a post.	• "I'd be looking for a, um, respond button to be able to get back to [username], you know?" [Participant 10]

The content analysis of the usability testing transcripts grouped the data into the 5 themes predetermined from the agendas of the co-design workshops. The first 4 themes—features and functions, usability, branding, and barriers and facilitators—are presented in full in Table S1 in Multimedia Appendix 5. Table S1 in Multimedia Appendix 5 also describes the platform modifications and subsequent development work that were considered and/or implemented to address participants' feedback. The fifth theme—ethos and values—is depicted in Table S2 in Multimedia Appendix 5. Within each theme, codes and subcodes were also added as appropriate during analysis (see Tables S1 and S2 in Multimedia Appendix 5).

Theme 1: Features and Functions

This theme included the codes of additional features, content and materials, community page, "My Garden," and joining form. In addition to the bugs and glitches participants experienced when creating an account, upon reflection, some expressed apprehension about signing up and agreeing to the community principles (terms of use). For instance,

> *It's almost like these are all our rules and regulations, and you have to agree to all of these rules and regulations before we will even tell you anything. That was, that put me on my guard, really.* [Participant 10]

Participants expressed the need to fully understand "what am I agreeing to," suggesting that background information about CommonGround and who owns it is needed to help people decide whether to join. However, participants shared that the personal information required matched their expectations of what creating an account typically involves. Throughout the feedback, it was clear that making the account creation process as easy and as informative as possible was important for facilitating registration on CommonGround.

> *To be honest, it's, it's really straight forward and it kind of makes it easier because it's, it goes along with what the majority of the registered, sort of, websites go for.* [Participant 8]

Participants generally had positive views of the community, resources, and "My Garden" pages. Specifically, people appreciated the unique features of the community forum, stating that the different types of posts you could create were "something I've not seen ever before" [Participant 2], and that the "reactions are great... the 'I hear you' and this (thinking of you) is helpful, it's quite nice to see that" [Participant 8]. Some participants were unsure what the "My Garden" page was for, but were keen to use it once they understood that they could save posts of interest to their own private "garden": "that really appeals to me, because I can, I can, yeah, grow my own library, oh, that's heaven [laughs]" [Participant 12]. Participants suggested that additional text should be added to the "My Garden" page, and the "save post" button on forum posts should be changed to "add to My Garden," to clearly demonstrate how the page can be used.

Participants also shared the view that new additional features could improve the platform. For instance, participants expressed a clear desire for a search bar to fulfill their need of easily finding specific content of interest and overcoming issues in navigating the platform: "I would seriously just give me a search bar and let me write in what I want to look for" [Participant 10]. Participants also suggested adding pages that share "about us" information and the full moderation policy.

Theme 2: Usability

This theme included the codes of platform navigation, visual design, visual layout, and tone and language. Overall, participants indicated that the platform "is quite user-friendly" [Participant 3], and that they felt comfortable using the platform, even if they did not feel technologically adept themselves:

> *I wouldn't say that I'm fantastic in technology. So, it was...It's, it's...I think it's quite technically friendly, and quite easy to access, to be honest.* [Participant 9]

Participants stressed the importance of the platform being intuitive to navigate with a typical website layout. For instance, participants expected there to be a menu bar as they have "got so used to seeing a menu thing at the top that when I go on to any website, that's the first thing I look for, is that menu" [Participant 10]. Although the prototype had a menu bar, participants "didn't initially notice that they (the menu bar and page tabs) were there at all" [Participant 12], until asked to complete a task involving switching to a different page that forced them to search around the page for it.

Participants appreciated the simple, clear, uniform layout of the community feed, with posts that are "a nice size to be able to digest" [Participant 11]. While many participants liked the colors and graphics of the platform, some of the graphics used to indicate the functionality of buttons were not sufficiently clear: "the images that you have above the post a thought, share knowledge, ask a question, I was just wondering what that is, what those images depict" [Participant 1]. Participants also wanted some clarity on the language used within the site copy, particularly for terms relating to peer support and mental health that may be unfamiliar:

> *Peer to peer...I think at times, it could be slightly wordy...It's a term that I've come across, yes*

XSL•FO
RenderX

> *definitely. I'm not sure it's term that everybody will know, though.* [Participant 1]

Participants also shared that extra explanations would be helpful on the resources page, when introducing the resources overall ("Um, you say end, endorsed by experts...But who, who exactly?" [Participant 2]) and on specific topics such as "keeping active" ("I'm a little bit confused on when you say active, is that more like exercising keeping active sort of thing or...?" [Participant 8]).

Theme 3: Branding

This theme related to how participants felt an online platform should be presented to the world, including the codes of the platform name, platform branding, and advertising. Participants felt that the name "CommonGround" clearly communicated a safe space for people who have the shared experience of living with long-term health conditions to connect. However, participants were unsure whether the name was able to communicate the fact that the platform has a focus on supporting mental well-being:

> *I do think that's a really good name, but, um, I don't know whether it says enough...It depends how people are introduced to CommonGround...If it's something you find, you know, if you Google, if you Googled psychological services, and CommonGround came up, then I don't think it would really matter.* [Participant 8]

These views were also expressed in relation to the logo:

> *I know CommonGround is probably the main logo. Um, reading CommonGround for the first time you wouldn't exactly think it's, like, health-related or anything.* [Participant 6]

Having a brand, including the name and logo, that communicates the purpose of CommonGround was suggested to facilitate engagement with the platform.

The importance of being able to trust the online platforms was emphasized. Participants wanted to know who owns the platform and if their motivation is to genuinely improve the well-being of users or to generate profit. In the case of CommonGround, the KCL logo on the landing page helped develop trust in our motivations and governance: "you're under King's College so that gives you that credibility anyway" [Participant 11]. Participants suggested that advertising the platform through avenues that people with long-term conditions trust, including their health care teams or charities, would also foster trustworthiness in CommonGround:

> *I think it carry, it would carry more weight. You know, if a, if a professional has recommended something.* [Participant 2]

Theme 4: Barriers and Facilitators

During their reflections of exploring CommonGround, participants identified some barriers and facilitators that might affect someone's interest in, and engagement with, a peer support platform. For instance, someone's desire to use the platform might be limited by their perceptions of peer support, and whether peer support is suitable for them or how to engage with it appropriately: "(being) on this platform practically every day. That's not healthy" [Participant 7]. For those who were motivated to use CommonGround, participants believed that a user's technological competency would not be a barrier. Participants who self-identified as "not very tech-savvy" felt "like it's very open to a lot of people" [Participant 8] and "it's quite technically friendly, and quite easy to access" [Participant 9]. However, there were concerns related to the safety of the online community, which may be vulnerable to the sharing of misinformation or malicious links, and infiltration by rogue users:

> *It's just you know how there's like a lot of spam, a lot of scams, especially when it comes to links sort of thing. And you never know what you're clicking on. So, somebody who's got good intention and coming on the website, you don't know if that's going to be everybody's intention.* [Participant 8]

Theme 5: Ethos and Values

The ethos and values theme had codes including platform purpose, the meaning of coproduction, and the power of peer support (see Table S2 in Multimedia Appendix 5). The fact that CommonGround was rooted in coproduction was considered essential, with participants emphasizing the importance of combining lived experience alongside academic experts who bring scientific rigour:

> *Co-design is essential if you're doing anything for a, a particular audience. You, you, you must include representatives of that audience if you want to get a, a, a fully rounded, um, end product.* [Participant 7]

That CommonGround would fulfill a current gap in support for people with long-term conditions was also noted, with participants sharing that CommonGround is "much needed" [Participant 1], particularly for its dual focus on the mind and body: "I mean, as well as your physical health conditions, I think it would help tremendously with my mental health" [Participant 6].The power of peer support was recognized through participants' previous experiences, where the ability to connect anonymously with people "like me" is powerful, as sometimes there was hesitancy to speak with people in their existing social circles:

> *A lot of people I, I, I would imagine would, would talk to somebody on the bus or whatever, before they'd talk to family, or friends even...The good thing about that is nobody knows who I am.* [Participant 5]

Although sharing experiences can help alleviate feelings of "why me," some concerns were shared of becoming overfocused on the effects of their long-term conditions during these exchanges, when they might rather "just get up and get on with my life" [Participant 7]. Participants recognized that a peer support platform can be beneficial at times, but also something to take some space from at other times dependent on their personal circumstances and well-being needs.

Iterative Platform Development

Table S1 in Multimedia Appendix 5 outlines the platform development opportunities identified through content analysis

XSL•FO
RenderX

of the think-aloud exercise and interviews that guided the iterative development process. The software developers were successful in completing the development work agreed by the RAG and KCL team. For instance, in response to the need for improved navigation, the RAG designed a search function. This included adding a search button within the navigation menu and allowing users to search by free text or content tags (similar to hashtags), with the option to filter results by recency or relevance.

Throughout the iterative process, the RAG also created additional features to enhance user experience. One key addition was the "How I Am Feeling Now" feature. The RAG recognized the importance of being able to communicate one's emotional context in forum posts to help their peers understand the type of support they are looking for. The RAG therefore suggested that when users are creating their post, there should be an option to select an emoji that represents their current emotional state or mood which then appears alongside the post's text to give this emotional context.

Further Prototype Testing, Final Development Work, and Final Prototype Approval

The testing by the RAG and KCL volunteers identified minor bugs, slow loading times, and that notifications were not being sent or received properly. Other feedback included suggestions to reorder or reformat text, for instance, by making key information more prominent or highlighted to catch the attention of end users who may skim through the content. The final required edits were successfully implemented, with the final prototype for the feasibility randomized controlled trial completed in May 2024. Screenshots of the final platform prototype are provided in Multimedia Appendix 6.

Discussion

Principal Findings

There is a current lack of psychological support to help people with long-term physical health conditions manage their mental well-being that is easily accessible and tailored to their specific psychosocial needs, particularly for those at risk of developing major depression [48,49]. In response to this gap in psychological support, we sought to coproduce an online peer support platform following a 4-stage, iterative coproduction cycle. While coproduction has become the gold standard for intervention development [28], to date, end users have rarely been involved in all aspects of intervention design, development, and evaluation [27]. Our work serves as an exemplar of how to coproduce an online peer support and psychoeducation platform by creating partnerships between people with long-term physical health conditions, academics, clinicians, and software experts. In the co-assess stage, people living with long-term physical health conditions expressed that an online intervention that provides a safe and trustworthy space for connecting with others like me, while acknowledging the mind-body interplay, could support their psychosocial well-being [29]. Following the co-design phase, which resulted in the successful creation of an initial prototype, usability testing during the co-validation stage confirmed that CommonGround included features and functions that were appropriate and acceptable. Critically, areas for improvement were also identified, such as making platform navigation more intuitive, resolving platform glitches, and building new features to enhance user experiences. Ultimately, through the iterative coproduction cycle that consistently places the voices of lived experience experts at the foreground, we created an intervention that has been specifically tailored to our target audiences' unique characteristics and needs, from the specific features and functions to the "all diagnoses welcome" ethos and platform governance. The CommonGround platform is now viable and ready to be evaluated in a future feasibility randomized controlled trial.

As noted in the development of other mobile health applications (eg, [50]), we found that iterative usability testing was critical for resolving lingering bugs and glitches to minimize the impact of these problems on user engagement [51]. Notably, many of the usability issues our participants experienced are commonly encountered during usability testing, such as navigational barriers and errors in site copy [52]. Our participants' emphasis on the importance of CommonGround being easy to understand and intuitive to navigate offers further evidence of the critical role that language and usability play in facilitating engagement with peer support platforms [16]. This finding also aligns with the technology acceptance model, which states that whether a user will accept or reject an online platform is partially dependent on how easy it is to use [53]. Together, these insights highlight the importance of incorporating structured, iterative usability testing throughout the coproduction of novel online interventions to ensure that they are technically functional, accessible, and aligned with the anticipated end user's needs.

An individual's digital literacy is often cited as a potential barrier to engaging with online interventions [19], yet interestingly, our participants did not identify technological competency as a potential barrier to accessing CommonGround. As participants described the platform as relatively intuitive to navigate, this could suggest that CommonGround is easy to use and does not require a significant level of digital literacy. However, it is important to note that although some participants identified as "not very tech-savvy," they were all daily internet users and may not represent the views of those with limited internet access or digital experience. This highlights the importance of actively recruiting people who may not typically engage with online interventions to comprehensively explore potential barriers to using online peer support forums [16].

Developing an online intervention, however, goes beyond creating the features and functions that may improve end users' experiences. For instance, the importance of CommonGround being a safe space, and underpinned by consistent, reputable, and trustworthy branding, was emphasized throughout the co-design and co-validation stages. This echoes prior work highlighting how the inclusion of reputable branding, logos, and declarations of ownership can positively influence the trustworthiness and credibility of web-based health information [53,54]. Another facet of safety that our participants highlighted was technological security. Concerns about security and data protection are commonly cited as barriers to engaging with online interventions [55], with older users, those from high-income backgrounds, and those who had a previous

XSL•FO
RenderX

experience of a data breach expressing the most apprehension [56]. This highlights the importance of collaborating with software experts to build a robust, protected online interface with security features to minimize potential threats, assure data confidentiality, and maintain privacy. However, given that concerns about security vary between users, it is pertinent to develop security features alongside lived experience experts to strike a balance between allaying fears (a facilitator to engagement) and avoiding excessive measures (a barrier to engagement) [57,58].

The "power of peer support" emanated as a central theme throughout our coproduction process, reinforcing previous findings among people with long-term conditions that online peer support provides a valued space to strengthen social ties [21] and can improve an individual's emotional self-management and self-efficacy [16]. Throughout this work, our PPI and research participants echoed these findings, emphasizing the unique value of peers in providing valuable information and support, independent of their existing social ties [16,59,60] and reducing loneliness by providing connections with "someone like me" [21,49]. However, while the power of peer support was well recognized among our participants, some expressed awareness of potential negative events that can occur in online peer support and warned that not everyone may have such favorable opinions or understand exactly what peer support is. It has been recognized that the term "peer support" has been inconsistently defined in research [16] and is often not colloquially used in patient communities. There is often a lack of clarity or misperceptions of what peer support involves that might serve as a barrier to engagement [61]. Indeed, evidence suggests that some people are deterred from online peer support as they believe that they will not fit in, peer support will be unable to support their well-being, their posts will remain unanswered, or that they will be exposed to incorrect or misleading information [16,59,62]. Given that perceived usefulness of an app can predict engagement, considering the end users' opinion and understanding of peer support may be an important barrier or facilitator to engagement [63]. It will therefore be important to provide clear information to users on what peer support is and how potential adverse effects will be mitigated on CommonGround. Future work should also evaluate participants' views of peer support before and after engaging with online peer support to develop an understanding of the relationship between perceptions of peer support and engagement with interventions.

Limitations

The design of the CommonGround platform and the findings from the formal usability testing (think-aloud exercises, interviews) should be considered within the context of our participant demographics. The majority identified as women, were daily internet users, and had access to the internet at home. Our sample did however have varying levels of prior experience engaging with online forums (30% [3/10] of participatory design panel lived experience experts and 42% [5/12] of usability testing participants had no prior experience of using online forums), representing a range of different baseline expectations of what online forums should be like. Nonetheless, as frequent internet users, they may perceive and prioritize issues differently than those who use the internet less often, particularly as they have greater experience to compare the prototype to. Although 96% of UK households have access to the internet [64], internet usage is lower among those in older adults (eg, 99.5% of those aged 25-34 years versus 85.5% aged 65-74 years using the internet in the last 3 months [65]). As living with (multiple) long-term conditions is more common among older people, who therefore represent a significant proportion of those that could benefit from CommonGround, future work should prioritize exploring the perspectives of this group and investigating the potential role of digital exclusion. By recruiting a larger, more diverse sample as part of a future feasibility randomized controlled trial, we will capture a broader range of experiences, usability concerns, and insights into barriers, facilitators, and intervention improvements. We will purposively interview participants based on various factors, including age and engagement with CommonGround, with topic guides including focus on exploring barriers, such as digital exclusion, in greater depth. Findings from this future work will ensure that CommonGround can be developed further to align with any specific needs identified, including those of older adults and those who may experience digital exclusion.

Our participatory design panel and subsequent RAG also represent a group of patients who are actively engaged in managing their physical-mental well-being, have an interest in peer support, and are frequently online. As such, their perspectives and associated intervention development are potentially rooted in positive experiences and opinions of peer support. Further exploration of the views of individuals who are new to peer support and/or less active in their self-management would provide broader insights into their psychosocial needs and expectations of how a peer support platform should function.

It should also be noted that no formal assessment of depressive symptoms was conducted among participants or RAG members; however, all involved self-identified as having lived experience relevant to preventing depression among people with long-term physical health conditions. While this broad, inclusive approach enabled a range of different perspectives to be incorporated throughout the coproduction process, we cannot conclude that our findings are specifically applicable to those living with subthreshold depression. Future work could explore capturing formal measures of mental health and well-being but should be mindful of excluding the voices of lived experience experts solely based on score cutoffs. Our future feasibility and acceptability trial will offer opportunity to explore the perceived acceptability, feasibility, and appropriateness of the CommonGround platform exclusively among a sample of those living with subthreshold depression.

Due to the COVID-19 pandemic, usability testing and subsequent RAG meetings for iterative development were conducted online via videoconferencing software. For participants unfamiliar with Microsoft Teams, this unfamiliarity may have contributed to nervousness or influenced how they interacted with the platform prototype during the think-aloud exercise. Resource restrictions also meant that usability sessions were conducted by members of the research team involved in prototype development. In future usability testing, involving an

XSL•FO
RenderX

independent facilitator could enhance objectivity and encourage more candid feedback. However, to mitigate any bias as best as possible during the think-aloud exercises, the research team avoided prompting or guiding participants, allowing them to navigate and troubleshoot issues independently.

Finally, our findings may not be applicable beyond the cultural context of the United Kingdom. The willingness of people to discuss their mental and physical health with other people, particularly those who are not included in their existing social circles, likely varies cross-culturally due to different norms around expressing emotions and the role of family/community, and stigma.

Conclusions

For interventions to be successfully integrated into the self-management routines of people with long-term physical health conditions, it is critical that they are developed in collaboration with individuals who have lived experience. By dedicating significant time and resources throughout the coproduction cycle, we have successfully created an online peer support platform with embedded psychoeducation that is designed to support the mental well-being of people living with long-term physical health conditions. Through an iterative process of usability testing and prototype development, we refined the platform to ensure that it aligns with the needs and preferences of the target audience.

The CommonGround platform is now ready to be further evaluated with continued input from our RAG during the delivery of a trial to explore the feasibility and acceptability of CommonGround. The findings from the next stages will be critical in informing the preparation of a future large-scale efficacy trial. If the intervention is found to be feasible, acceptable, and clinically effective at preventing the progression of subthreshold depressive symptoms to comorbid major depressive disorder, CommonGround may offer a low-cost, scalable solution to bridge the current gap in psychological support for people with long-term physical health conditions who are at risk of developing comorbid major depressive disorder.

Acknowledgments

We would like to acknowledge our software developers at TOAD and Bitjam who were instrumental in helping to co-design the CommonGround platform. We also extend our thanks to Hannah Rowlands for their support and time dedicated to the co-assess stage. We would also like to acknowledge the time and contributions of all of the participants who made this work possible. And finally, thank you to all the members of our Public Patient Involvement groups and Research Advisory Group who contributed at various stages throughout the development of this protocol. CommonGround is a coproduced intervention, and this research would not be possible without their dedication, support, and insightful contributions.

Data Availability

Data are available upon request to the corresponding author.

Funding

This project is being funded by King’s Health Partners and Guy’s and St Thomas Charity from the Multiple Long Term Conditions Challenge Fund [EIC180702 (to MH)]. The funders had no role in study design, data collection and analysis, decision to publish, or manuscript preparation. The research is part-funded by the National Institute for Health and Care Research (NIHR) Maudsley Biomedical Research Centre (BRC). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Authors' Contributions

All authors listed meet the International Committee of Medical Journal Editors criteria for authorship. MH, AS, VL, and GL conceived the study. All authors co-designed the study. HGJ led data analysis alongside supervision from VL and GL. HGJ and GL wrote the manuscript. All authors reviewed and approved the final manuscript.

Conflicts of Interest

EA-F declares voluntary and unpaid advisory to Someone Who Knows, an early-stage technical initiative seeking to create an online platform that matches people with diagnosed long-term health conditions with peers for informal support and virtual conversations. Someone Who Knows is registered as a business in the Netherlands. EA-F has never received payment, shares, or any other form of exchange for this advice. All other authors have no conflicts of interest to declare.

Multimedia Appendix 1
Consent forms.
[DOCX File , 67 KB - jmir_v28i1e79666_app1.docx]

Multimedia Appendix 2
Think-aloud activity tasks.

XSL•FO
RenderX

[DOCX File , 21 KB - jmir_v28i1e79666_app2.docx]

Multimedia Appendix 3
Topic guide for semistructured interviews.
[DOCX File , 25 KB - jmir_v28i1e79666_app3.docx]

Multimedia Appendix 4
The CommonGround platform manifesto written by our coinvestigator with lived experience.
[DOCX File , 21 KB - jmir_v28i1e79666_app4.docx]

Multimedia Appendix 5
The results tables of the content analysis of transcripts of the think-aloud task and semistructured interviews.
[DOCX File , 91 KB - jmir_v28i1e79666_app5.docx]

Multimedia Appendix 6
Screenshots of the final prototype of CommonGround.
[DOCX File , 1104 KB - jmir_v28i1e79666_app6.docx]

References

1. Long Term Conditions Compendium of Information. 3rd ed. UK Department of Health. 2012 May 30. URL: https://assets.publishing.service.gov.uk/media/5a7c638340f0b62aff6c154e/dh_134486.pdf [accessed 2026-03-14]
2. Egede LE. Major depression in individuals with chronic medical disorders: prevalence, correlates and association with health resource utilization, lost productivity and functional disability. Gen Hosp Psychiatry 2007;29(5):409-416. [doi: 10.1016/j.genhosppsych.2007.06.002] [Medline: 17888807]
3. Depression in adults with a chronic physical health problem: recognition and management. National Institute for Health and Care Excellence. 2009 Oct 29. URL: https://www.nice.org.uk/guidance/cg91 [accessed 2026-03-06]
4. Rayner L, Matcham F, Hutton J, Stringer C, Dobson J, Steer S, et al. Embedding integrated mental health assessment and management in general hospital settings: feasibility, acceptability and the prevalence of common mental disorder. Gen Hosp Psychiatry 2014;36(3):318-324 [FREE Full text] [doi: 10.1016/j.genhosppsych.2013.12.004] [Medline: 24630892]
5. Littlewood E, Chew-Graham C, Coleman E, Gascoyne S, Sloan C, Ali S. A psychological intervention by community pharmacies to prevent depression in adults with subthreshold depression and long-term conditions: the CHEMIST pilot RCT. Public Health Res 2022;10(5). [doi: 10.3310/ekze0617] [Medline: 35353469]
6. Rodríguez MR, Nuevo R, Chatterji S, Ayuso-Mateos JL. Definitions and factors associated with subthreshold depressive conditions: a systematic review. BMC Psychiatry 2012 Oct 30;12(1):1-7. [doi: 10.1186/1471-244x-12-181]
7. Hance M, Carney RM, Freedland KE, Skala J. Depression in patients with coronary heart disease: a 12-month follow-up. Gen Hosp Psychiatry 1996;18(1):61-65. [doi: 10.1016/0163-8343(95)00100-x] [Medline: 8666215]
8. Bot M, Pouwer F, Ormel J, Slaets JPJ, de Jonge P. Predictors of incident major depression in diabetic outpatients with subthreshold depression. Diabet Med 2010;27(11):1295-1301. [doi: 10.1111/j.1464-5491.2010.03119.x] [Medline: 20950389]
9. Bodenheimer T, Wagner EH, Grumbach K. Improving primary care for patients with chronic illness. JAMA 2002;288(14):1775-1779. [doi: 10.1001/jama.288.14.1775] [Medline: 12365965]
10. Miller WR, Lasiter S, Bartlett Ellis RB, Buelow JM. Chronic disease self-management: a hybrid concept analysis. Nurs Outlook 2015;63(2):154-161 [FREE Full text] [doi: 10.1016/j.outlook.2014.07.005] [Medline: 25241136]
11. Five year forward view. NHS England. 2014 Oct. URL: https://www.england.nhs.uk/wp-content/uploads/2014/10/5yfv-web.pdf [accessed 2026-03-06]
12. Taylor S, Pinnock H, Epiphanou E, Pearce G, Parke H, Schwappach A. A rapid synthesis of the evidence on interventions supporting self-management for people with long-term conditions: PRISMS (Practical Systematic Review of Self-Management Support) for long-term conditions. Health Serv Deliv Res 2014;2(53). [doi: 10.3310/hsdr02530] [Medline: 25642548]
13. Bodenheimer T, Wagner EH, Grumbach K. Improving primary care for patients with chronic illness: the chronic care model, Part 2. JAMA 2002;288(15):1909-1914. [doi: 10.1001/jama.288.15.1909] [Medline: 12377092]
14. Franek J. Self-management support interventions for persons with chronic disease: an evidence-based analysis. Ont Health Technol Assess Ser 2013;13(9):1-60 [FREE Full text] [Medline: 24194800]
15. Kim S, Park M, Song R. Effects of self-management programs on behavioral modification among individuals with chronic disease: a systematic review and meta-analysis of randomized trials. PLOS One 2021;16(7):e0254995 [FREE Full text] [doi: 10.1371/journal.pone.0254995] [Medline: 34297741]
16. Hossain SN, Jaglal SB, Shepherd J, Perrier L, Tomasone JR, Sweet SN, et al. Web-based peer support interventions for adults living with chronic conditions: scoping review. JMIR Rehabil Assist Technol 2021;8(2):e14321 [FREE Full text] [doi: 10.2196/14321] [Medline: 34032572]

XSL•FO
RenderX

17. Lauckner HM, Hutchinson SL. Peer support for people with chronic conditions in rural areas: a scoping review. Rural Remote Health 2016;16(1):3601 [FREE Full text] [Medline: 26943760]
18. Cooper RE, Saunders KRK, Greenburgh A, Shah P, Appleton R, Machin K, et al. The effectiveness, implementation, and experiences of peer support approaches for mental health: a systematic umbrella review. BMC Med 2024 Feb 29;22(1):72 [FREE Full text] [doi: 10.1186/s12916-024-03260-y] [Medline: 38418998]
19. Wright KB. Communication in health-related online social support groups/communities: a review of research on predictors of participation, applications of social support theory, and health outcomes. Rev Commun Res 2016;4:65-87. [doi: 10.12840/issn.2255-4165.2016.04.01.010] [Medline: 2255-4165]
20. Rafieifar M, Hanbidge AS, Lorenzini SB, Macgowan MJ. Comparative efficacy of online vs. face-to-face group interventions: a systematic review. Res Soc Work Pract 2024 Mar 18;35(5):524-542. [doi: 10.1177/10497315241236966]
21. Kingod N, Cleal B, Wahlberg A, Husted GR. Online peer-to-peer communities in the daily lives of people with chronic illness. Qual Health Res 2016 Dec 12;27(1):89-99. [doi: 10.1177/1049732316680203] [Medline: 27956659]
22. Johansson V, Islind AS, Lindroth T, Angenete E, Gellerstedt M. Online communities as a driver for patient empowerment: systematic review. J Med Internet Res 2021 Feb 09;23(2):e19910. [doi: 10.2196/19910] [Medline: 33560233]
23. Cooper K, Kirkpatrick P, Wilcock S. The effectiveness of peer support interventions for community-dwelling adults with chronic non-cancer pain: a systematic review. JBI Database System Rev Implement Rep 2014;12(5):319-348. [doi: 10.11124/jbisrir-2014-1590]
24. McCaughan E, Parahoo K, Hueter I, Northouse L, Bradbury I. Online support groups for women with breast cancer. Cochrane Database Syst Rev 2017;3(3):CD011652 [FREE Full text] [doi: 10.1002/14651858.CD011652.pub2] [Medline: 28278559]
25. Jennings H, Slade M, Bates P, Munday E, Toney R. Best practice framework for Patient and Public Involvement (PPI) in collaborative data analysis of qualitative mental health research: methodology development and refinement. BMC Psychiatry 2018 Jun 28;18(1):213 [FREE Full text] [doi: 10.1186/s12888-018-1794-8] [Medline: 29954373]
26. Hollis C, Sampson S, Simons L, Davies EB, Churchill R, Betton V, et al. Identifying research priorities for digital technology in mental health care: results of the James Lind Alliance Priority Setting Partnership. Lancet Psychiatry 2018 Oct;5(10):845-854 [FREE Full text] [doi: 10.1016/S2215-0366(18)30296-7] [Medline: 30170964]
27. Brotherdale R, Berry K, Branitsky A, Bucci S. Co-producing digital mental health interventions: a systematic review. Digit Health 2024;10:20552076241239172 [FREE Full text] [doi: 10.1177/20552076241239172] [Medline: 38665886]
28. Thompson DM, Booth L, Moore D, Mathers J. Peer support for people with chronic conditions: a systematic review of reviews. BMC Health Serv Res 2022 Mar 31;22(1):427 [FREE Full text] [doi: 10.1186/s12913-022-07816-7] [Medline: 35361215]
29. Rowlands H, Aylwin-Foster E, Hotopf M, Rayner L, Simpson A, Lavelle G, et al. A qualitative exploration of the psychosocial needs of people living with long-term conditions and their perspectives on online peer support. Health Expect 2023 Oct;26(5):2075-2088 [FREE Full text] [doi: 10.1111/hex.13814] [Medline: 37458403]
30. Sanders EBN, Stappers PJ. Probes, toolkits and prototypes: three approaches to making in codesigning. CoDesign 2014 Mar 06;10(1):5-14. [doi: 10.1080/15710882.2014.888183]
31. Berry N, Machin M, Ainsworth J, Berry K, Edge D, Haddock G, et al. Developing a theory-informed smartphone app for early psychosis: learning points from a multidisciplinary collaboration. Front Psychiatry 2020;11:602861 [FREE Full text] [doi: 10.3389/fpsyt.2020.602861] [Medline: 33362612]
32. Callan JA, Dunbar Jacob J, Siegle GJ, Dey A, Thase ME, DeVito Dabbs A, et al. CBT MobileWork©: user-centered development and testing of a mobile mental health application for depression. Cogn Ther Res 2020 Sep 25;45(2):287-302. [doi: 10.1007/s10608-020-10159-4]
33. Jaspers MWM, Steen T, van den Bos C, Geenen M. The think aloud method: a guide to user interface design. Int J Med Inform 2004 Nov;73(11-12):781-795. [doi: 10.1016/j.ijmedinf.2004.08.003] [Medline: 15491929]
34. Eccles DW, Arsal G. The think aloud method: what is it and how do I use it? Qual Res Sport Exerc Health 2017 May 26;9(4):514-531. [doi: 10.1080/2159676x.2017.1331501]
35. Crane D, Garnett C, Brown J, West R, Michie S. Factors influencing usability of a smartphone app to reduce excessive alcohol consumption: think aloud and interview studies. Front Public Health 2017;5:39 [FREE Full text] [doi: 10.3389/fpubh.2017.00039] [Medline: 28421175]
36. Perski O, Short CE. Acceptability of digital health interventions: embracing the complexity. Transl Behav Med 2021 Jul 29;11(7):1473-1480 [FREE Full text] [doi: 10.1093/tbm/ibab048] [Medline: 33963864]
37. Filipe A, Renedo A, Marston C. The co-production of what? Knowledge, values, and social relations in health care. PLOS Biol 2017;15(5):e2001403 [FREE Full text] [doi: 10.1371/journal.pbio.2001403] [Medline: 28467412]
38. Lambert N, Carr S. 'Outside the original remit': co-production in UK mental health research, lessons from the field. Int J Ment Health Nurs 2018;27(4):1273-1281 [FREE Full text] [doi: 10.1111/inm.12499] [Medline: 29920906]
39. Elwyn G, Nelson E, Hager A, Price A. Coproduction: when users define quality. BMJ Qual Saf 2020;29(9):711-716 [FREE Full text] [doi: 10.1136/bmjqs-2019-009830] [Medline: 31488570]
40. UK standards for public involvement. NIHR. 2019 Nov. URL: https://www.cso.scot.nhs.uk/wp-content/uploads/UK-Standards-for-Public-Involvement_Nov19.pdf [accessed 2026-03-06]

XSL•FO
RenderX

41. Payment guidance for researchers and professionals. NIHR. 2024 Dec 5. URL: https://www.nihr.ac.uk/about-us/who-we-are/policies-and-guidelines/payment-guidance-researchers-and-professionals [accessed 2026-03-06]
42. Payment for public involvement in health and care research: a guide for organisations on employment status and tax. NHS Health Research Authority. 2023 Apr 12. URL: https://www.hra.nhs.uk/planning-and-improving-research/best-practice/public-involvement/resources/payment-public-involvement-health-and-care-research-guide-organisations-employment-status-and-tax/ [accessed 2026-03-06]
43. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
44. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019;11(4):589-597. [doi: 10.1080/2159676x.2019.1628806]
45. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907 [FREE Full text] [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
46. Bengtsson M. How to plan and perform a qualitative study using content analysis. NursingPlus Open 2016;2:8-14. [doi: 10.1016/j.npls.2016.01.001]
47. Lavelle G, Jones HG, Carr E, Aylwin-Foster E, Lawrence V, Simpson A, et al. Online peer support for long-term conditions: protocol for a feasibility randomized controlled trial. JMIR Res Protoc 2025 Jul 23;14:e71513 [FREE Full text] [doi: 10.2196/71513] [Medline: 40700748]
48. Akyirem S, Forbes A, Wad JL, Due-Christensen M. Psychosocial interventions for adults with newly diagnosed chronic disease: a systematic review. J Health Psychol 2022 Jun;27(7):1753-1782 [FREE Full text] [doi: 10.1177/1359105321995916] [Medline: 33586486]
49. Marsh M, Shah SR, Munce SEP, Perrier L, Lee TSJ, Colella TJF, et al. Characteristics of existing online patient navigation interventions: scoping review. JMIR Med Inform 2024;12:e50307 [FREE Full text] [doi: 10.2196/50307] [Medline: 39159443]
50. Ben-Zeev D, Kaiser SM, Brenner CJ, Begale M, Duffecy J, Mohr DC. Development and usability testing of FOCUS: a smartphone system for self-management of schizophrenia. Psychiatr Rehabil J 2013;36(4):289-296 [FREE Full text] [doi: 10.1037/prj0000019] [Medline: 24015913]
51. Jaspers MWM. A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence. Int J Med Inform 2009;78(5):340-353. [doi: 10.1016/j.ijmedinf.2008.10.002] [Medline: 19046928]
52. Broekhuis M, van Velsen L, Peute L, Halim M, Hermens H. Conceptualizing usability for the eHealth context: content analysis of usability problems of eHealth applications. JMIR Form Res 2021 Jul 27;5(7):e18198 [FREE Full text] [doi: 10.2196/18198] [Medline: 34313594]
53. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. Manag Inf Syst Q 1989;13(3):319-340. [doi: 10.2307/249008]
54. Sbaffi L, Rowley J. Trust and credibility in web-based health information: a review and agenda for future research. J Med Internet Res 2017;19(6):e218 [FREE Full text] [doi: 10.2196/jmir.7579] [Medline: 28630033]
55. Adjekum A, Blasimme A, Vayena E. Elements of trust in digital health systems: scoping review. J Med Internet Res 2018;20(12):e11254 [FREE Full text] [doi: 10.2196/11254] [Medline: 30545807]
56. Alhammad N, Alajlani M, Abd-Alrazaq A, Epiphaniou G, Arvanitis T. Patients' perspectives on the data confidentiality, privacy, and security of mHealth apps: systematic review. J Med Internet Res 2024;26:e50715 [FREE Full text] [doi: 10.2196/50715] [Medline: 38820572]
57. Tseng HT, Ibrahim F, Hajli N, Nisar TM, Shabbir H. Effect of privacy concerns and engagement on social support behaviour in online health community platforms. Technol Forecast Soc Chang 2022;178:121592. [doi: 10.1016/j.techfore.2022.121592]
58. Zhu M, Wu C, Huang S, Zheng K, Young SD, Yan X, et al. Privacy paradox in mHealth applications: an integrated elaboration likelihood model incorporating privacy calculus and privacy fatigue. Telemat Inform 2021;61:101601. [doi: 10.1016/j.tele.2021.101601]
59. Kemp E, Koczwara B, Butow P, Turner J, Girgis A, Schofield P, et al. Online information and support needs of women with advanced breast cancer: a qualitative analysis. Support Care Cancer 2018;26(10):3489-3496. [doi: 10.1007/s00520-018-4206-1] [Medline: 29693203]
60. Brady E, Segar J, Sanders C. Accessing support and empowerment online: the experiences of individuals with diabetes. Health Expect 2017 Oct;20(5):1088-1095 [FREE Full text] [doi: 10.1111/hex.12552] [Medline: 28718928]
61. Taylor F, Gutteridge R, Willis C. Peer support for CKD patients and carers: overcoming barriers and facilitating access. Health Expect 2016;19(3):617-630 [FREE Full text] [doi: 10.1111/hex.12348] [Medline: 25649115]
62. Griffiths KM, Reynolds J, Vassallo S. An online, moderated peer-to-peer support bulletin board for depression: user-perceived advantages and disadvantages. JMIR Ment Health 2015;2(2):e14 [FREE Full text] [doi: 10.2196/mental.4266] [Medline: 26543919]
63. Lazard AJ, King AJ. Objective design to subjective evaluations: connecting visual complexity to aesthetic and usability assessments of eHealth. Int J Hum Comput Interact 2019;36(1):95-104. [doi: 10.1080/10447318.2019.1606976]

XSL•FO
RenderX

64. Internet access – households and individuals, Great Britain statistical bulletins. Office for National Statistics. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/previousreleases [accessed 2026-03-06]
65. Internet users – Table 2b: Recent and lapsed internet users and internet non-users, by age group (years) and sex, UK, 2013 to 2020. Office for National Statistics. 2021 Apr 6. URL: https://www.ons.gov.uk/businessindustryandtrade/itandinternetindustry/datasets/internetusers [accessed 2026-03-06]

Abbreviations

KCL: King's College London
PPI: patient and public involvement
RAG: research advisory group

Edited by A Stone; submitted 25.Jun.2025; peer-reviewed by A Barooah, E Littlewood; comments to author 20.Oct.2025; revised version received 02.Feb.2026; accepted 18.Feb.2026; published 24.Mar.2026.

Please cite as:
Jones HG, Lavelle G, Aylwin-Foster E, Carr E, Simpson A, Hotopf M, Lawrence V
Coproducing an Online Platform for People With Long-Term Physical Health Conditions: Development and Usability Study
J Med Internet Res 2026;28:e79666
URL: https://www.jmir.org/2026/1/e79666
doi:10.2196/79666
PMID:

©Hannah Grace Jones, Grace Lavelle, Elly Aylwin-Foster, Ewan Carr, Alan Simpson, Matthew Hotopf, Vanessa Lawrence. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of the Components of a Digital Multiple Health Behavior Change Intervention Among Individuals Seeking Help Online (Coach): Factorial Randomized Trial

Joel Crawford[1], PhD; Katarina Åsberg[1], PhD; Jenny Blomqvist[1], MSc; Oskar Lundgren[1,2], MD, PhD; Hanna Henriksson[1,3], PhD; Pontus Henriksson[1], PhD; Preben Bendtsen[1,4], MD, PhD; Marie Löf[5], PhD; Marcus Bendtsen[1], PhD

[1]Department of Health, Medicine and Caring Sciences, Linköping University, Linköping, Östergötland, Sweden

[2]Crown Princess Victoria Children's Hospital, Linköping University Hospital, Linköping, Östergötland, Sweden

[3]Unit for Strategic Healthcare, Region Östergötland, Linköping, Östergötland, Sweden

[4]Motala Hospital, Motala, Östergötland, Sweden

[5]Department of Medicine Huddinge, Karolinska Institutet, Stockholm, Stockholm, Sweden

Corresponding Author:
Marcus Bendtsen, PhD
Department of Health, Medicine and Caring Sciences
Linköping University
Campus US
Linköping, Östergötland, 581 83
Sweden
Phone: 46 13281000
Email: marcus.bendtsen@liu.se

Abstract

Background: Extant digital multiple health behavior change interventions have shown promise in various populations; however, evidence for a broader approach among the general population is lacking. Moreover, existing interventions often contain several components but are typically assessed as a whole, meaning it remains unclear to what extent individual components contribute to intervention effects and how they may interact to influence health outcomes.

Objective: This study estimates the effects of 6 components of a digital health behavior change intervention on alcohol, diet, physical activity, and smoking outcomes among individuals searching for help online.

Methods: A double-blind randomized factorial trial design with 6 two-level factors was used. Adults from the general public in Sweden who were seeking help to change their behaviors were recruited through web searches and social media. Participants were eligible if they were 18 years or older and had at least one health behavior classified as unhealthy. Effects of 6 components were estimated: screening/feedback, goal-setting/planning, motivation, skills/know-how, mindfulness, and self-authored SMS text messages. Primary outcomes were weekly alcohol consumption and frequency of heavy episodic drinking, average daily fruit and vegetable consumption, weekly moderate-to-vigorous physical activity, and 4-week point-prevalence smoking.

Results: A total of 5419 individuals were randomized. Overall, the screening/feedback component was the most effective for changing health behaviors, along with goal-setting/planning and motivation to change. In particular, there was evidence that screening/feedback increased average daily portions of fruit and vegetables at 2 months (mean difference 0.17, compatibility interval [CoI] 0.09-0.25, probability of effect [POE] >99.9%) and at 4 months (mean difference 0.13, CoI 0.04-0.21, POE 99.9%) and reduced the frequency of heavy episodic drinking at 4 months (incidence rate ratio 0.91, CoI 0.81-1.03, POE 94.2%). Components also interacted to further improve health outcomes, most notably the combination of screening/feedback with motivation to change, which further increased fruit and vegetable consumption (2 months: mean difference 0.20, CoI 0.09-0.30, POE >99.9%; 4 months: mean difference 0.17, CoI 0.05-0.29, POE 99.8%).

Conclusions: The results from this study contribute to the development of more effective interventions by providing novel insights into the effects of individual and pairwise components of complex digital health behavior change interventions.

Trial Registration: ISRCTN Registry ISRCTN16420548; http://www.isrctn.com/ISRCTN16420548

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2022-061024

XSL•FO
RenderX

(J Med Internet Res 2026;28:e88881) doi:10.2196/88881

KEYWORDS

behavior change; digital interventions; multiple lifestyle; public health

Introduction

Alcohol consumption, diet, physical activity, and smoking are leading factors determining general health and susceptibility to disease [1] and are prominent causes of noncommunicable diseases (eg, cardiovascular disease, respiratory disease, cancer, and diabetes) [1]. The increasing prevalence of chronic disease is one of the major challenges to global health, and unlike intrinsic etiologies of noncommunicable diseases, health behaviors are modifiable [2]. In light of this, the World Health Organization has set a charter for member states to lower the incidence of noncommunicable diseases by implementing measures to reduce engagement in health risk behaviors while increasing engagement in health promotion behaviors [2]. To achieve this, there is a need to develop and implement effective measures to help individuals improve their health behaviors.

In Sweden, individuals regularly engage in unhealthy behaviors; national surveys have shown that 40% of Swedish individuals drink at levels that indicate risky alcohol use, 66% fail to eat an adequate amount of vegetables, 32% report insufficient physical activity, 5.8% smoke daily, and 51% have obesity or are overweight [3,4]. Despite macro-level interventions aimed at curbing the incidence of unhealthy behaviors [5-7], the prevalence of these behaviors has remained consistent over the last 10 years, with the exception of tobacco smoking [4]. Overall, unhealthy lifestyle behaviors present a significant health burden for premature mortality and morbidity in Sweden, with 15% of all deaths and 2974 disability-adjusted life years (DALYs) attributed to tobacco use, 15% of all deaths and 2204 DALYs attributed to poor dietary behaviors, 5% of all deaths and 1292 DALYs attributed to alcohol use, and 2% of all deaths and 266 DALYs attributed to insufficient physical activity in 2019 (DALYs provided per 100,000 population) [8,9]. Thus, new means of intervening in the general public are warranted to avoid another decade of stagnating or increasing prevalence of unhealthy behaviors and subsequent maladies.

Current intervention efforts have relied on public health recommendations, national ad campaigns, and policy measures such as restrictions on availability and advertising [5-7]. While macro-level approaches do have some merit (ie, the ability to reach a significant portion of the populace), they are unable to address the barriers and intricacies of health behavior change; for example, they provide little support on how to change and do not consider that health behaviors often cluster and interact [10,11]. Evidence indicates that health behaviors consistently co-occur: for example, smoking and alcohol use, smoking and unhealthy diet, and physical inactivity and unhealthy diet [12]. A recent study conducted in Sweden found that 36% of adults had a co-occurrence of at least 2 unhealthy behaviors, with 10% having a co-occurrence of 3 behaviors and 2% having a co-occurrence of 4 behaviors [13]. The clustering of unhealthy behaviors significantly increases the risk of all-cause mortality and morbidity compared with a single behavior alone [14-16]. Current intervention efforts typically focus on a single behavior, meaning they are unable to target underlying clusters that considerably increase health risk [12]; hence, there is a need for interventions that can address the intricacies and barriers this presents to health behavior change.

One method to overcome issues with current prevention efforts is to offer digital support tools for those seeking help online. This approach has the potential to be widely disseminated while also addressing the intricacies of health behavior change. For instance, content can be tailored to personal dimensions of behavior and can also address barriers to health (eg, reducing stigma associated with seeking help) [17]. In addition, a digital approach may be particularly apt for Sweden due to the near-ubiquitous ownership of smartphones and the high prevalence of seeking help online—66% of individuals aged 16-84 years sought health advice online in 2023 [18]. While effectiveness trials for digital behavior change interventions targeting the Swedish general population have shown effectiveness for smoking cessation [19] and reducing alcohol consumption [20], there is currently a lack of digital interventions targeting multiple health behaviors [21]. Existing multiple health behavior change interventions have shown promise for various subpopulations, including university students, patients with a specific malady, and different professions [22-25]; however, evidence for a broader approach for the general population is lacking.

In addition, many current interventions use a sociocognitive framework to target modifiable determinants of health, such as intentions, attitudes, beliefs, self-efficacy, risk perceptions, motivation, planning, and expectancies [26]. Sociocognitive models of health and behavior change (eg, Theory of Planned Behavior, Health Action Process Approach, Capability-Opportunity-Motivation-Behavior [COM-B] model [27-29]) highlight that health behavior is seldom influenced by a single factor, but rather by a configuration of constructs that interact to influence health outcomes; they also indicate that targeting one factor without addressing deficits in others may not result in meaningful change [27,28]. Current interventions that target improvements in these determinants often use various components to instill change; however, they typically assess the effectiveness of the components as a whole [26]. This means that we are unsure to what degree separate components contribute to the effects of the interventions, how they interact to influence health outcomes, or whether certain components are effective only in conjunction with other commensurate components; for example, action planning may only be effective when self-efficacy or motivation is high [27,28]. Hence, gaining knowledge regarding the separate and interactive effects of intervention components would enable a more nuanced understanding of the effects of health behavior change interventions.

We conducted this study in response to the sparsity of studies investigating multiple health behavior change interventions for

XSL•FO
RenderX

the general public and the need to better understand the effects of individual components of digital behavior change interventions. We used a factorial trial design to estimate the effects of 6 separate intervention components of a novel digital intervention targeting multiple health behaviors. The study was carried out among individuals seeking help online. Specifically, the study aimed to achieve the following:

- Estimate the effects of 6 individual intervention components on reducing alcohol use, improving dietary behaviors, increasing physical activity, and increasing smoking cessation.
- Estimate the pairwise interaction effects of the 6 intervention components on reducing alcohol use, improving dietary behaviors, increasing physical activity, and increasing smoking cessation.

Methods

Study Design

We conducted a factorial randomized trial (6 factors representing 6 components, C1-C6, each with 2 levels representing the absence or presence of the component) [30] to estimate the effects of a digital multiple health behavior change intervention among individuals seeking help online. The trial was an extension of the MoBILE (Mobile Health Multiple Lifestyle Behavior Interventions Across the Lifespan) research program [31] and was prospectively registered (ISRCTN16420548). A full protocol and analysis plan were published prospectively [32]. This report follows CONSORT (Consolidated Standards of Reporting Trials; Multimedia Appendix 1) guidelines [33], including extensions for factorial trials [34].

Study Setting, Recruitment, and Participants

The target population was Swedish adults seeking help online to change their health behaviors. Participants were recruited through advertisements on Google (Alphabet Inc), Bing (Microsoft Corporation), Facebook (Meta Platforms, Inc), and websites related to lifestyle and behavior change. Individuals interested in participating sent an SMS text message to a dedicated phone number and received a reply with a link to a page containing informed consent materials (see Multimedia Appendix 2). All individuals who provided consent were then asked to complete a baseline survey, which was also used to assess eligibility (see Multimedia Appendix 3). To be included, individuals had to be 18 years of age or older and fulfill at least one of the conditions presented in Textbox 1. Note that, while there is no safe limit for alcohol consumption [35], we used Swedish national guidelines for alcohol to identify individuals who could benefit greatly from reducing their consumption.

As the study was conducted via participants' smartphones, individuals without access to one were unable to take part. This also applied to those who did not have sufficient understanding of Swedish to comprehend the study materials.

Textbox 1. Criteria for inclusion.

- **Weekly alcohol consumption:** Consumed 10 or more (for women) or 15 or more (for men) standard drinks of alcohol in the past week. In Sweden, a standard drink is defined as 12 g of pure alcohol.
- **Heavy episodic drinking:** Consumed 4 or more (for women) or 5 or more (for men) standard drinks of alcohol on a single occasion at least once in the past month.
- **Fruit and vegetables:** Consumed less than 500 g of fruit and vegetables on average per day in the past week.
- **Moderate-to-vigorous physical activity:** Spent less than 150 minutes on moderate-to-vigorous physical activity in the past week.
- **Smoking:** Smoked at least one cigarette in the past week.

Interventions

The Coach intervention studied in this trial was a smartphone-based intervention comprising 6 components: (C1) screening/feedback, (C2) goal-setting/planning, (C3) motivation to change, (C4) skills/know-how, (C5) mindfulness, and (C6) self-authored SMS text messages. The components were informed by sociocognitive models of health [36] and have been included in interventions that have proven successful in facilitating behavior change [19,20]. While built on what is already known about digital health behavior change interventions, the novelty of the Coach intervention lies in its approach of including multiple behaviors within a single intervention. The intervention was designed so that each component could be presented to participants in a menu list, making it possible to add or remove components based on factorial conditions.

A core feature of the intervention was an SMS text message sent to participants each Sunday afternoon, containing a link and reminder to access the intervention materials. The intervention components, depending on factorial allocation, and weekly reminders were available to participants over 4 months. Table 1 provides an overview of each component, along with its relevant behavior change techniques [37]. Please see Multimedia Appendix 4 for full details of the intervention. As Coach was designed to be a multiple health behavior change intervention, all participants, regardless of their baseline screening, received intervention materials for alcohol, diet, physical activity, and smoking. There was no fixed pathway through the materials; rather, they were presented as a toolbox. Therefore, participants could engage with materials concerning different behaviors according to their own preferences and were not required to engage with materials for behaviors they did not find relevant at the time.

XSL•FO
RenderX

Table 1. Description of intervention components, conceptual framework, practical application, and factorial conditions.

Component	Conceptual framework	Practical application	Factorial conditions
Component 1: screening and feedback	*Self-monitoring* has been shown to be a potentially effective strategy for reducing excessive alcohol [37-40] consumption and promoting healthy eating and physical activity [41,42]. These include BCT[a] 1.6 (Discrepancy between current behavior and goals), BCT 2.2 (Feedback on behavior), BCT 2.3 (Self-monitoring of behavior), and BCT 6.2 (Social comparison).	Every Sunday afternoon, participants received an SMS text message with a hyperlink to a questionnaire regarding their current health behaviors. Once the questionnaire was completed, participants received feedback on their current behaviors in comparison to national guidelines. Thereafter, users were given access to the rest of the components (depending on allocation).	When this component was absent, participants were not asked to respond to the screening questionnaire but were instead shown national guidelines without any personal feedback.
Component 2: goal-setting and planning	*Self-regulatory* skills and capacity via goal-setting and planning. Planning-related activities such as goal-setting, action planning, and practicing behavior have been shown to be effective in lifestyle behavior interventions [41,43-48]. These include BCT 1.1 (Goal-setting), BCT 1.2 (Problem solving), BCT 1.4 (Action planning), BCT 7.1 (Prompts/cues), and BCT 8.1 (Behavior practice/rehearsal).	This component allowed participants to set 1 or more goals for their future behavior. It included action planning for how they were to progress toward their goals, preparation for motivational struggles (coping planning), and strategies for rewarding themselves upon success. Participants could also create their own or accept ready-made challenges for the coming week, such as walking for 15 minutes each day or not drinking any alcohol the coming week. Reminders were sent via SMS text messages to participants about their goals and challenges throughout the week (up to 4 messages).	When absent, this component was not visible, and goal-setting reminders were not available.
Component 3: motivation	Digital health behavior change interventions have been shown to enhance s*elf-efficacy*; however, there is a lack of consensus across reviews as to which content works best to facilitate an increase in self-efficacy [49]. These include BCT 5.1 (Information about health consequences), BCT 9.1 (Credible source), BCT 9.2 (Pros and cons), and BCT 9.3 (Comparative imagining of future outcomes).	This component contained information and tools to increase participants' awareness of their own motivation, encourage commitment, and boost self-efficacy. This included information on negative health consequences, costs induced by certain behaviors, and reflective tasks via texts, videos, and exercises. If participants chose, they could also activate motivational SMS text messages (derived from previously developed and evaluated interventions [50-56]) sent to them throughout the week (a maximum of 8-10 messages per week).	When absent, this component was not visible, and SMS text messages were not available.
Component 4: skills and know-how	Intervention content based on shaping knowledge, aimed at increasing an individual's understanding and awareness to facilitate behavior change, has been shown to be effective [44,45,57,58]. These include BCT 4.1 (Instructions on how to perform a behavior), BCT 8.2 (Behavior substitution), BCT 8.3 (Habit formation), and BCT 8.7 (Graded tasks).	This component aimed to increase participants' skills and know-how by providing concrete tips on how to initiate and maintain lasting changes in everyday life by repetition and substitution. Participants were given strategies (eg, how to say no to alcoholic beverages or how to introduce vegetables into their meals). If participants chose to, they could also activate SMS text messages with tips and know-how, which were sent to them throughout the week (a maximum of 8-10 messages per week).	When absent, this component was not visible, and SMS text messages were not available.
Component 5: mindfulness	The mindfulness exercises were based on previous research and considered to be evidence-based methods to improve the mental well-being of clinical populations, while effects on behavior change in nonclinical settings are less well-studied [59-63].	This component aimed to increase users' awareness of their own lived experience and strengthen their capacity for a nonreactive, compassionate, and less stressful way of being in the world. Mindfulness exercises were offered to participants, including guided meditations.	When absent, this component was not visible, and guided meditations were not available.

XSL•FO
RenderX

Component	Conceptual framework	Practical application	Factorial conditions
Component 6: self-authored SMS text messages	SMS text message self-authorship is generally understudied in the literature but was included in an effective digital alcohol intervention [64].	This component allowed participants to self-author up to 3 SMS text messages and schedule them to be sent to themselves throughout the week at times of their choosing. For example, a participant could write an SMS text message reminding themselves to eat 2 fruits each day, avoid drinking on Wednesdays, or go for a walk with a friend.	When absent, this component was not visible, and the ability to self-author SMS text messages was not available.

[a]BCT: behavior change technique.

Outcomes and Measures

Outcomes are listed in Textbox 2 and subsequently explained. All questionnaires used in the trial can be found in Multimedia Appendix 3.

Weekly alcohol consumption was assessed using a short-term recall method that asked participants to record the number of standard drinks consumed in the past week [65]. The frequency of heavy episodic drinking was assessed by asking participants to report how many times they consumed 4 or more (for women) or 5 or more (for men) standard drinks per session over the past month. These 2 outcomes are both part of the core outcome set for brief alcohol interventions [66,67].

Diet and physical activity were assessed using items based on a questionnaire from the Swedish National Board of Health and Welfare [21]. The items were adapted to include portion sizes (100 g of fruit or vegetables, respectively) and asked how many portions were consumed per day in the preceding week. Sugary drink consumption was assessed by asking participants to report the number of units (33 cl) of sugary drinks consumed in the past week. Moderate-to-vigorous physical activity (MVPA) was assessed with 2 items asking for the number of minutes spent in both moderate and vigorous activity in the past week; responses were summed to provide a total MVPA value. BMI was measured by asking participants to report their height (cm) and weight (kg).

Four-week point-prevalence smoking abstinence (no cigarettes in the past 4 weeks) was assessed using a binary item. Those reporting nonabstinence were asked to record the number of cigarettes smoked in the past week. These items were based on suggestions from the Society for Research on Nicotine [68].

Perceived stress was assessed using the short-form Perceived Stress Scale [69], and quality of life (QoL) was measured using Patient-Reported Outcomes Measurement Information System (PROMIS) Global 10 [70].

Textbox 2. Primary and secondary outcomes.

Primary outcomes

- **Alcohol:** Weekly alcohol consumption; monthly frequency of heavy episodic drinking
- **Diet:** Average daily consumption of fruit and vegetables
- **Physical activity:** Weekly moderate-to-vigorous physical activity
- **Smoking:** Four-week point prevalence of smoking abstinence

Secondary outcomes

- Weekly consumption of sweets and snacks
- Weekly consumption of sugary drinks
- BMI
- Weekly number of cigarettes smoked
- Perceived stress
- Quality of life

Follow-Up

Primary and secondary outcomes were assessed at 2 and 4 months after randomization. The 2-month follow-up facilitated the assessment of the immediate effects of the components, while the 4-month follow-up facilitated the estimation of effects after prolonged access. Additional longer follow-up intervals were not included to reduce participant burden, and because attrition was expected to be much higher the longer the time from baseline. An additional follow-up was conducted at 1 month after randomization to assess the proposed mediators of the intervention. Mediator analyses will be reported separately. Follow-ups were initiated by SMS text messages containing hyperlinks to questionnaires (see Multimedia Appendix 3). A total of 2 reminder SMS text messages were sent, 2 days apart, to nonresponders. If no response was received, participants were contacted by phone and asked to respond to the primary outcome measures only. A total of 5 call attempts were made.

XSL•FO
RenderX

Randomization and Blinding

We used block randomization with random block sizes of 64 and 128 to allocate participants equally among factors. The process was automated, with the sequence computer-generated and the backend server allocating eligible participants after completion of baseline assessments. Neither research personnel nor participants were able to influence allocation. Research personnel were blinded to allocation throughout the trial. Participants were also blinded, as they had access to different components of the trial but were unaware of the other available conditions. All measurements, including baseline and follow-up, were conducted through digital questionnaires completed by participants independently on their smartphones, thereby preserving blinding during measurement. Attempts to collect data from nonresponders via smartphones introduced a risk of disclosure of allocation for some participants; however, research personnel were experienced and instructed to ask only about the outcome measures. In addition, given that there were 64 conditions in the trial, it would have been difficult for participants to inadvertently reveal their allocation.

Statistical Analysis

Intention-to-Treat Analysis and Bayesian Multilevel Modeling

We conducted analyses keeping participants in the conditions to which they were randomly allocated (intention to treat). We estimated models using both available data and data with missing values imputed (multiple imputation with chained equations [71], with 100 datasets and 30 iterations using predictive mean matching). All statistical analyses were prespecified in the study protocol [32].

We contrasted the presence versus absence of components using multilevel regression models with covariates for time-by-component interactions and participant-level adaptive intercepts. We also estimated models with pairwise interactions between components. All models were adjusted for the presence or absence of all components, as well as baseline measures of age, sex, importance, confidence, and know-how. Primary outcomes, perceived stress, and weekly number of cigarettes smoked were additionally adjusted for their respective baseline measures. BMI, sugary drink consumption, and sweets and snacks were adjusted for baseline MVPA minutes per week and average daily intake of fruit and vegetables. QoL was adjusted for perceived stress at baseline.

Bayesian inference was used to estimate the parameters of the models [72]. We used standard normal priors for all covariates and random intercepts to reflect a conservative view of effect magnitudes; thus, the priors induced shrinkage to protect against spurious findings. For each coefficient of interest, we report the marginal posterior probability of effect (POE), using the median as a point estimate of the magnitude of the effect. We also report 95% compatibility intervals (CoIs), defined by the 2.5% and 97.5% percentiles of the posterior distributions.

Primary and Secondary Outcomes

We analyzed primary outcomes among participants who fulfilled each respective inclusion criterion; for example, the MVPA outcome was analyzed among those with less than 150 minutes of MVPA in the past week at baseline. Secondary outcomes were analyzed among all participants, except for the number of cigarettes smoked in the past week, which was analyzed among baseline smokers who continued to smoke.

All count variables, including total weekly alcohol consumption, frequency of heavy episodic drinking, consumption of sugary drinks, sweets and snacks, and cigarettes smoked per week, were analyzed using negative binomial regression models. Average daily intake of fruit and vegetables, minutes of MVPA per week, BMI, QoL, and perceived stress were analyzed using linear regression. Point-prevalence smoking abstinence was analyzed using logistic regression.

Ancillary Analyses

We studied attrition by estimating the odds of responding to follow-up, conditional on baseline characteristics and the presence or absence of the 6 components (factors), using logistic regression. We estimated 1 model without interaction terms and a second model including interaction terms between factors and baseline characteristics. We used Cauchy priors to promote a parsimonious model (centered at 0 with a standard normal hyperprior for scale). We also modeled primary outcome data conditional on the number of attempts to collect follow-up. Under the assumption that late responders are more similar to nonresponders than early responders are, any association between the number of attempts and outcomes may indicate systematic attrition.

Post Hoc Analyses

To investigate whether the number of components to which participants had access affected outcomes, we post hoc modeled outcomes conditional on the number of components. The same model specifications were used as in the primary analyses, with dummy variables for components replaced by a variable for component count.

Sample Size

We used a Bayesian sequential design to monitor when to stop recruitment [73]. As 4-month follow-up data became available, we modeled each of the primary outcomes according to the analysis plan, and effect estimates for each component were assessed against targets for effect, harm, and futility (see protocol for details of the criteria [32]). We anticipated that recruitment would not last more than 12 months; however, we allowed recruitment to continue for a total of 24 months and stopped recruitment at that point, despite targets not being met for all outcomes.

Ethical Considerations

Ethical approval was received on August 11, 2021, from the Swedish Ethical Review Authority (approval Dnr 2021-02855). All participants were provided with study information and gave informed consent before responding to the baseline questionnaire and subsequently being randomized. Participants received no compensation for participating in the trial. Participants' phone numbers were encrypted and stored in the study database to link baseline and follow-up data. At project completion, the

XSL•FO
RenderX

encrypted phone numbers will be deleted and the data will be anonymized.

Results

Overview

A CONSORT participant flow diagram is presented in Figure 1 (sign-up and randomization) and Figure 2 (follow-up). Table 2 provides an overview of participants' baseline characteristics. Between October 19, 2021, and October 19, 2023, a total of 6405 individuals showed interest in the study, of whom 5574 consented. A total of 5445 completed the baseline questionnaire, revealing that 26 participants were not eligible for inclusion. The remaining 5419 individuals were randomized. Among the 5419 participants, 650 (11.99%) decided to stop the reminder SMS text messages before the 16-week program had ended. We did not investigate the reasons why individuals chose to stop the reminders; however, all randomly allocated participants were sent follow-up questionnaires. Primary outcome data were available for 3371 (62.21%) participants at the 2-month follow-up interval and for 2786 (51.41%) at the 4-month interval.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) participant flow diagram—sign-up and randomization.

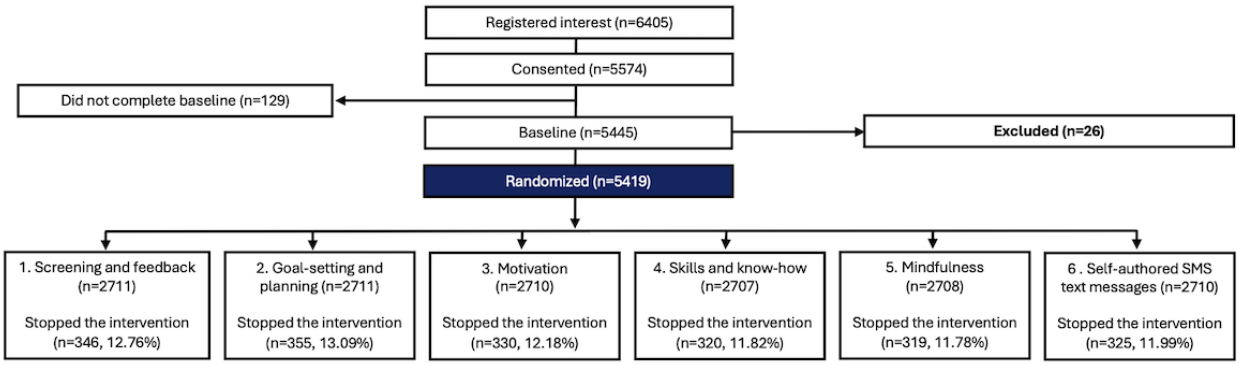

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) participant flow diagram—follow-up.

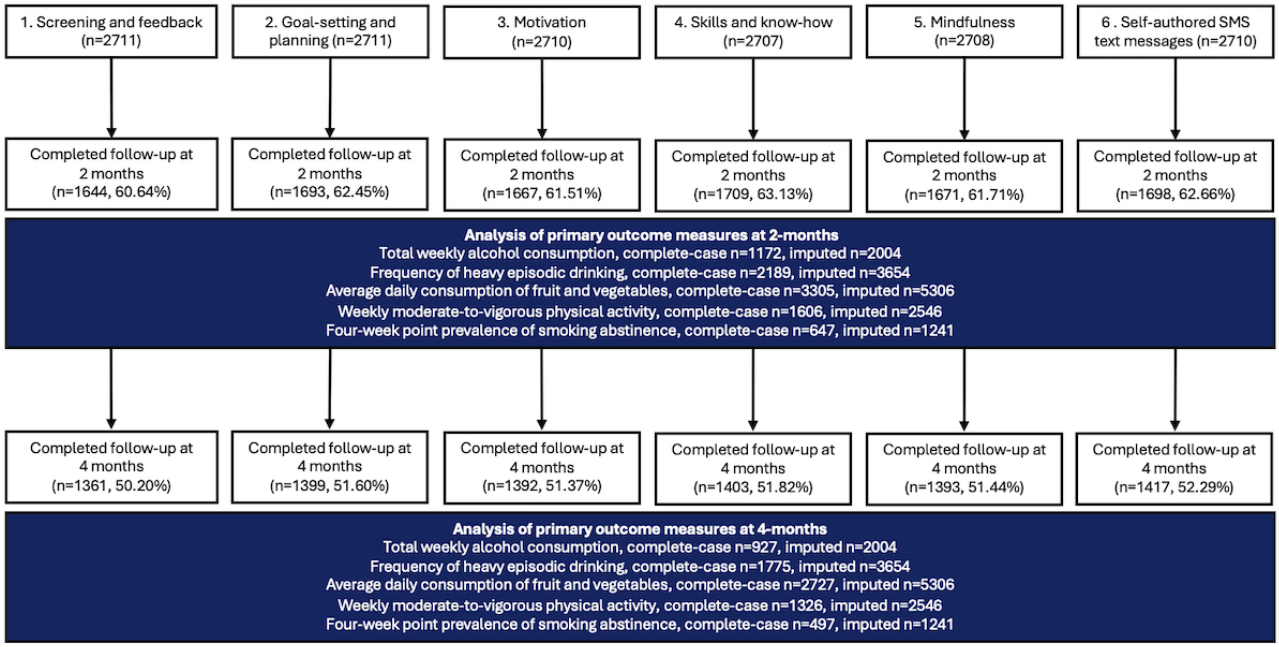

XSL•FO
RenderX

Table 2. Baseline characteristics of randomly allocated participants.

Characteristics	Total (N=5419)	Screening/feedback (n=2711)	Goal-setting/ planning (n=2711)	Motivation (n=2710)	Skills/know-how (n=2707)	Mindfulness (n=2708)	Self-authored SMS text message (n=2710)
Sex, n (%)							
Women	4301 (79.37)	2140 (78.94)	2141 (78.97)	2151 (79.37)	2172 (80.24)	2164 (79.91)	2136 (78.82)
Men	1118 (20.63)	571 (21.06)	570 (21.03)	559 (20.63)	535 (19.76)	544 (20.09)	574 (21.18)
Age (years), mean (SD)	50 (11.7)	49.8 (11.8)	50 (11.7)	50.6 (11.6)	50 (11.4)	50.4 (11.6)	51 (11.5)
Alcohol, mean (SD)							
Total weekly alcohol consumption (standard drinks)	8.89 (8.94)	8.88 (8.99)	8.91 (9.00)	8.54 (8.58)	9.04 (8.67)	8.77 (8.97)	9.48 (8.87)
Frequency of heavy episodic drinking	4.80 (7.10)	4.77 (7.07)	4.91 (7.27)	4.54 (6.57)	4.76 (6.57)	4.46 (6.19)	4.22 (5.54)
Smoking							
Number of smokers, n (%)	1240 (22.88)	636 (23.46)	626 (23.09)	632 (23.32)	633 (23.38)	615 (22.71)	636 (23.47)
Number of cigarettes last week, mean (SD)	15.27 (38.0)	15.28 (37.5)	15.04 (38.2)	16.01 (39.9)	15.26 (38.2)	11.89 (34.2)	11.72 (32.3)
Physical activity							
Moderate-to-vigorous physical activity (minutes), mean (SD)	220.4 (226.9)	226.5 (231.4)	208.7 (221.5)	222.0 (232.1)	225.3 (220.3)	210.2 (201.2)	215.3 (234.2)
Dietary behavior							
Average daily fruit and vegetable consumption (portions), mean (SD)	1.49 (1.17)	1.53 (1.18)	1.39 (1.12)	1.47 (1.18)	1.47 (1.17)	1.61 (1.17)	1.34 (1.04)
Stress							
Self-perceived stress[a], mean (SD)	7.62 (2.80)	7.60 (2.77)	7.59 (2.89)	7.67 (2.92)	7.82 (2.69)	7.51 (2.65)	7.55 (2.44)
Psychosocial measures, median (IQR)							
Importance[b]	10 (8-10)	10 (8-10)	10 (8-10)	10 (8-10)	10 (8-10)	10 (8-10)	10 (9-10)
Confidence[b]	7 (5-8)	7 (5-8)	7 (5-8)	7 (5-8)	7 (5-8)	6 (5-8)	6 (5-8)
Know-how[b]	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)	6 (5-8)	6 (4-8)

[a]The 4-item Perceived Stress Scale with total scores ranging from 0 to 16.

[b]Single item with 1-10 response options.

Attrition

Results from the attrition analyses can be found in Multimedia Appendices 5 and 6. In summary, we found evidence that older participants were more likely to respond to follow-up than younger participants at both follow-up intervals. In addition, those with more frequent episodes of heavy drinking and those who smoked more at baseline were less likely to respond to follow-up. Notably, these associations were not moderated by differential access to components.

We found evidence suggesting that late responders differed from early responders with respect to fruit and vegetable consumption, MVPA, and smoking. The evidence also suggested that these associations were moderated by differential access to components. Under the assumption that late responders are more similar to nonresponders than early responders are, this provides evidence of systematic attrition.

XSL•FO
RenderX

Primary and Secondary Outcomes

Overview

Multimedia Appendix 7 contains tables reporting the main effects and 2-way interactions for the intervention components with available data, whereas Multimedia Appendix 8 contains the corresponding effect estimates with missing data imputed. Here, we highlight findings for which the evidence of effects was strongest, quoting estimates from analyses of available data.

Primary Outcomes

Total Weekly Alcohol Consumption

There was no consistent evidence of effects from any of the components on weekly alcohol consumption, either individually or in combination. The relatively strongest evidence was found for an effect of screening/feedback (C1) at 4 months (incidence rate ratio [IRR] 0.94, CoI 0.83-1.07, POE 82.4%). In addition, there was some evidence of an effect of goal-setting/planning (C2) at 2 months (IRR 0.95, CoI 0.85-1.06, POE 81.8%), but this effect was not observed at 4 months. The evidence suggested that combining screening/feedback (C1) with skills/know-how (C4) resulted in even lower alcohol consumption at 4 months (IRR 0.92, CoI 0.76-1.11, POE 81.4%). These findings persisted when imputing missing data.

Frequency of Heavy Episodic Drinking

Among individual components, the relatively strongest evidence for reduced frequency of heavy episodic drinking was found for screening/feedback (C1) at 4 months (IRR 0.91, CoI 0.81-1.03, POE 94.2%). However, there was considerable evidence of harm from multiple combinations of components at the 4-month follow-up interval. For instance, combining goal-setting/planning (C2) with motivation (C3) (IRR 1.14, CoI 0.97-1.31, POE 93.4%), mindfulness (C5) (IRR 1.15, CoI 0.97-1.37, POE 94.6%), or self-authored SMS text messages (C6) (IRR 1.17, CoI 0.98-1.38, POE 96.3%) was associated with increased frequency of heavy episodic drinking. These findings persisted in analyses with missing data imputed.

Daily Portions of Fruit and Vegetables

Individually, there was strong evidence that screening/feedback (C1) increased daily portions of fruit and vegetables at 2 months (mean difference 0.17, CoI 0.09-0.25, POE >99.9%) and at 4 months (mean difference 0.13, CoI 0.04-0.21, POE 99.9%). There was also evidence that combining screening/feedback (C1) with other components increased effects, most prominently when combined with motivation (C3) at 2 months (mean difference 0.20, CoI 0.09-0.30, POE >99.9%) and at 4 months (mean difference 0.17, CoI 0.05-0.29, POE 99.8%). There was evidence that goal-setting/planning (C2) decreased fruit and vegetable consumption at 4 months (mean difference −0.08, CoI −0.16 to 0.00, POE 97.1%), and even more so when combined with skills/know-how (C4) (mean difference −0.14, CoI −0.26 to −0.02, POE 99.1%). These findings persisted in the imputed data.

Weekly MVPA

There was no consistent evidence of effects from any of the components, nor from combinations of components, on MVPA. The relatively strongest evidence in the available-data analyses suggested that mindfulness increased minutes of MVPA at 4 months (mean difference 10.3, CoI −11.3 to 32.1, POE 82.3%), and even more so when combined with screening/feedback (C1), goal-setting/planning (C2), and motivation (C3). However, in the imputed analyses, there was no evidence of these effects.

Smoking

There were several individual components for which there was evidence of effects on the point prevalence of smoking cessation in the available-data analyses. Notably, motivation (C3) increased the odds of smoking cessation at 4 months (odds ratio [OR] 2.4, CoI 0.98-5.99, POE 97.2%), whereas mindfulness (C5) decreased the odds of smoking cessation at 2 months (OR 0.35, CoI 0.15-0.78, POE 99.5%). These findings persisted in the imputed data, whereas evidence for other individual component effects did not. Evidence for effects of combined components was most consistent across available-data and imputed analyses for the combination of motivation (C3) and self-authored SMS text messages (C6) at 4 months (OR 2.42, CoI 0.70-8.72, POE 91.9%), and the combination of mindfulness (C5) and self-authored SMS text messages (C6) at 2 months (OR 0.23, CoI 0.07-0.71, POE 99.4%). The harmful effects of mindfulness that were prominent at 2 months were less evident at 4 months. There was no consistent evidence suggesting that any of the components, individually or in combination, affected the number of cigarettes smoked per week among those who continued smoking.

Secondary Outcomes

Weekly Sugary Drink Consumption

There was evidence that screening/feedback (C1) initially reduced weekly sugary drink consumption both individually (IRR 0.87, CoI 0.72-1.10, POE 92.2%) and in combination with skills/know-how (C4) and mindfulness (C5). However, there was no strong evidence that these effects persisted at 4 months, and findings were attenuated in the imputed analyses. By contrast, evidence based on available-data and imputed analyses suggested that skills/know-how (C4) reduced sugary drink consumption at 4 months (IRR 0.85, CoI 0.70-1.05, POE 93.8%), and increasingly so when combined with goal-setting/planning (C2) and motivation (C3). There was also evidence that combining motivation (C3) and self-authored SMS text messages (C6) increased sugary drink consumption at 2 months (IRR 1.22, CoI 0.92-1.59, POE 91.6%); however, there was no evidence of this harmful effect at 4 months.

Weekly Portions of Candy and Snacks

There was evidence that screening/feedback (C1) reduced weekly portions of candy and snack consumption at 2 months (IRR 0.93, CoI 0.85-1.02, POE 95.1%) and 4 months (IRR 0.92, CoI 0.83-1.01, POE 96.0%). In addition, there was evidence that motivation (C3) also reduced consumption at 4 months (IRR 0.85, CoI 0.77-0.94, POE 99.9%). Combining screening/feedback (C1) and motivation (C3) led to even greater reductions at 4 months (IRR 0.77, CoI 0.67-0.89, POE >99.9%).

XSL•FO
RenderX

There was also evidence of increased effects when screening/feedback (C1) was combined with goal-setting/planning (C2), as well as with skills/know-how (C4). These findings persisted in the imputed-data analyses.

BMI

The relatively strongest evidence of effects on BMI was found for mindfulness (C5), which reduced BMI at both 2 months (mean difference −0.24, CoI −0.70 to 0.18, POE 86.4%) and 4 months (mean difference −0.23, CoI −0.69 to 0.20, POE 85.0%). There was also evidence that when mindfulness (C5) was combined with skills/know-how (C4), BMI was further reduced at both 2 months (mean difference −0.41, CoI −1.00 to 0.19, POE 90.7%) and 4 months (mean difference −0.47, CoI −1.07 to 0.14, POE 92.9%). These findings persisted in the imputed-data analyses.

Perceived Stress

The most consistent evidence of effects on perceived stress scores was found for the mindfulness (C5) component, both at 2 months (mean difference −0.17, CoI −0.37 to 0.03, POE 95.4%) and at 4 months (mean difference −0.22, CoI −0.44 to 0.0, POE 97.4%). Evidence of additional benefit at the 4-month follow-up interval was found when mindfulness was combined with screening/feedback (C1), as well as with motivation (C3). These findings persisted in the imputed analyses. There was also evidence that skills/know-how (C4) reduced perceived stress scores at 2 months (mean difference −0.25, CoI −0.45 to −0.05, POE 99.2%); however, the evidence suggested that the benefit did not persist at 4 months.

Quality of Life

There was no strong evidence of effects of the components, either individually or in combination, on QoL. Relatively, the strongest evidence was found for motivation (C3), which increased QoL scores both individually (mean difference 0.26, CoI −0.31 to 0.84, POE 81.7%) and in combination with skills/know-how (C4) (mean difference 0.54, CoI −0.28 to 1.35, POE 90.5%). These findings persisted but were attenuated in the imputed-data analyses.

Post Hoc Analyses

In Multimedia Appendix 9, we present estimates of the effects of the number of components to which participants had access on primary and secondary outcomes. The relatively strongest evidence for effects was found for increasing component count with regard to fruit and vegetable consumption and perceived stress at the 2-month follow-up. There was also evidence of reduced candy and snack consumption with increased component count at the 4-month follow-up interval.

Discussion

Primary Findings

This study used a factorial design to estimate the effectiveness of intervention components on multiple health behaviors, along with proxies for health and well-being. The results highlight novel findings regarding both beneficial and harmful effects of specific components, as well as how these components interact to affect health outcomes.

Screening/feedback (C1) was most consistent in reducing alcohol consumption, in terms of both weekly consumption and heavy episodic drinking. Further reductions in weekly consumption were observed when combined with skills/know-how (C4). These findings are consistent with extant digital alcohol interventions focused on screening and feedback, which have been shown to reduce alcohol consumption [20,39,74]. In addition, evidence from a study examining the mediated effects of an alcohol reduction app, “Drink Less,” found that tracking (ie, self-monitoring behavior) mediated the effects of the intervention [75]. However, in this study, notable evidence of harm, in terms of increased frequency of heavy episodic drinking, was found for combinations of components, particularly those including self-authored SMS text messages. We can only speculate why this was the case, but conflicting priorities in behavior change may have partially contributed to these harms. Notably, the intervention content was based on a digital alcohol intervention previously shown to be effective among people seeking help online [20]; however, when this content was combined with materials targeting other behaviors, it may not have been sufficiently tailored to individuals who typically engage in heavy episodic drinking. This highlights the need for caution when offering multiple behavior interventions to individuals whose primary focus should be on reducing heavy episodic drinking rather than improving overall lifestyle.

For smoking cessation, motivation (C3) was the most effective component, along with its interaction with other components, most notably with self-authored SMS text messages. Motivation reflects the importance of quitting smoking, suggesting that motivation to change is influenced by the perceived value of the outcome; the importance of the decision to quit is an integral part of cessation [76]. Evidence indicates that perceived benefits of cessation are associated with increased motivation to quit [77], and that positive abstinence outcome expectancies are associated with greater intentions to quit [78]. An analysis of individualized effects from a recent smoking cessation trial in Sweden showed that those who considered it highly important to quit reported better outcomes than those who considered it less important [79]. By contrast, we found that mindfulness (C5) was harmful in the short term, reducing the odds of smoking cessation at 2 months. Although speculative, many smokers use cigarettes as a stress reliever, and initiating a mindfulness program may initially heighten awareness of previously suppressed feelings, which can feel overwhelming. Previous research has shown that mindfulness interventions may be associated with increased stress during an initial phase before new coping strategies are established [80,81]. Furthermore, this study used self-directed practice, which lacks the emotional support of a group setting and guidance from an experienced teacher. Thus, individuals who smoked may have been more likely to continue smoking in the short term as a coping strategy. Notably, these harmful effects were not observed at the 4-month follow-up.

The screening/feedback component was effective in improving dietary outcomes, including increasing fruit and vegetable consumption, decreasing candy and snack consumption, and, to a lesser extent, reducing sugary drink consumption. This finding is consistent with a review by König et al [82], which

XSL•FO
RenderX

reported that feedback and self-monitoring of dietary behavior may be key to intervention effectiveness. Nonetheless, the literature has yet to establish a dose-response relationship between feedback effects and engagement. Some studies suggest that daily tracking and feedback are effective [83,84], whereas others indicate that this may be too intrusive and therefore less effective [82]; by contrast, the current trial demonstrated that weekly screening and feedback can improve outcomes. Future research is needed to establish consensus on the appropriate level of screening and feedback, potentially by examining how self-perception interacts with external feedback to influence behavior [85]. Screening/feedback also interacted with various components to further improve dietary outcomes, most prominently with motivation (C3). Goal-setting/planning (C2) was also effective in improving candy and snacks consumption, supporting findings that demonstrate action planning and goal-setting as effective approaches [45,86,87]. The impact on dietary outcomes may be explained by how components interact to influence factors necessary for change, as outlined in various sociocognitive models of health [29,88,89]. For example, motivation may increase intentions, goal-setting and planning may help translate intentions into action, and feedback on performance may further enhance efficacy. This suggests that changing dietary behaviors may be best achieved through interventions that target motivation, facilitate the translation of intentions into action, and provide performance feedback. An extant intervention incorporating these components has been shown to improve fruit and vegetable consumption [90], while the current results suggest they may also help reduce consumption of calorie-dense foods and sugary drinks.

There was no consistent evidence that the intervention components, either individually or in pairwise combinations, affected physical activity. This is not consistent with existing literature suggesting that goal-setting and planning are effective intervention approaches for increasing engagement in physical activity [91,92]. There was some evidence indicating that mindfulness was effective in increasing physical activity, which aligns with prior evidence highlighting the effectiveness of this approach [93]. Nonetheless, in the imputed analyses, there was limited evidence of effects on physical activity. There was, however, relatively stronger evidence that mindfulness reduced BMI, supporting findings across trials that highlight the efficacy of mindfulness-based interventions for weight loss [94]. Evidence for skills/know-how (C4) also suggested effectiveness in reducing BMI, consistent with studies demonstrating the efficacy of skills training for weight loss [95,96]. Furthermore, an interaction between mindfulness and skills/know-how resulted in the largest reduction in BMI. This interaction may have promoted more conscious decision-making regarding diet and physical activity. However, it is also possible that the observed effects on BMI were secondary to the components' effects on dietary behaviors, leading to reduced caloric intake.

Finally, mindfulness (C5) was found to be effective in reducing stress, both individually and in combination with other components. Mindfulness-based stress reduction is an established approach for reducing psychological stress and can be effectively delivered either in person or online [97,98]. While there was no strong evidence of effects on QoL, motivation (C3), in combination with skills/know-how (C4), showed some evidence of improving QoL. Extant interventions for improving QoL generally target individuals with chronic health conditions and are typically administered as part of integrated care interventions [99]. This study suggests that QoL may potentially be improved in the general population using digital interventions, possibly as a result of increased perceived self-efficacy arising from improvements in the 4 main health behavior domains.

Generalizability and Limitations

We used a pragmatic approach in this trial and kept barriers to participation low. This strengthens the interpretation of the estimated effects as measures of effectiveness rather than efficacy. However, this comes with a trade-off with respect to attrition. Participants may have enrolled out of curiosity, and dropping out was easy; thus, attrition was, as expected, high. We found notable associations between nonresponse and age, as well as baseline smoking and more frequent episodes of heavy drinking. However, none of these associations were moderated by component access. Furthermore, although the evidence for these associations was relatively strong, the ORs were small. Thus, while the most conservative estimates of effects may be more relevant for older individuals, less frequent heavy drinkers, and smokers, overall, the evidence for systematic attrition with respect to responders versus nonresponders was not sufficiently strong to raise major concern. By contrast, there was evidence that late responders had higher fruit and vegetable consumption at follow-up compared with early responders, which, under the assumption that late responders and nonresponders are similar, suggests a risk of attrition bias. However, this association was not observed among those with access to the screening/feedback component (C1), which showed the most promising effects on fruit and vegetable consumption. Thus, the risk of bias for this finding in particular may be considered lower. Additionally, late responders had more MVPA minutes and were less likely to have quit smoking than early responders. Overall, this suggests that our estimates for these outcomes may be biased due to attrition; however, we do not have sufficient data to determine the direction or magnitude of this bias. Although this limitation should be considered when interpreting the findings, the primary aim of this study should also be emphasized: to examine the relative contributions of different intervention components to changes in health behaviors. The overall conclusions regarding which components are relatively more promising from an effectiveness perspective are less likely to be affected by attrition in this study than the precise effect estimates themselves.

The study used self-report measures to record outcomes. To minimize recall bias, we focused on behavior in the past week (except for heavy episodic drinking and smoking, which considered the past month). This avoided asking participants to recall behavior over several months or a year, which would not only be difficult to recall accurately but also less relevant given the time frame of the intervention. Although the validity of online measures has been established [100], they remain prone to social desirability or impression management biases [101]. However, in our study, blinding reduced the likelihood that these biases differed by allocation. For smoking and

XSL•FO
RenderX

physical activity outcomes, intervention trials have used objective measures such as pedometers or salivary nicotine tests. Evidence suggests that self-reports may be less accurate than direct measures [102,103]; however, in this study, administering direct measures—particularly for smoking—was not feasible. Additionally, the study did not specify a particular type of physical activity; rather, participants reported minutes of activity. The Society for Research on Nicotine and Tobacco also recommends not using salivary tests in population studies to avoid introducing selection bias due to unequal opportunities to complete and return tests [68]. Conducting scientific studies requires decisions about which biases can be controlled and which must be accepted as limitations. In the case of self-report versus device-based measurement, there are trade-offs that are not uniform across studies. Rather than focusing solely on the accuracy of measured behavior—which is subject to bias regardless of method—future work should consider simulation studies to examine the impact of varying levels and types of bias. Estimating how measurement decisions influence effect estimates may help inform and justify study design choices.

An important consideration when assessing the generalizability of the findings from this factorial trial is the level of service utilization in a real-world implementation. If different groups in society are more or less likely to use digital interventions, the health benefits of implementation may disproportionately affect certain groups. Furthermore, the intervention materials may have differential effects across population subgroups, and not all groups may have been adequately represented in our study sample. Our estimates of effects are generalizable to the wider population insofar as it can be assumed that effects are not moderated by factors associated with nonparticipation. However, it is not possible to assess such moderation in this study, and dedicated studies are needed to examine implementation and moderation effects in greater detail.

Finally, this study targeted individuals who had sought help online; thus, participants were expected to have some motivation to change at study entry. This was supported by the high importance reported at baseline. Providing support to individuals actively seeking help is important; however, engaging those who could benefit from behavior change but are not actively seeking support remains a challenge. Future research should focus on proactive approaches to community outreach for multiple behavior support, potentially drawing on brief alcohol intervention screening strategies used in higher education settings [39,40]. This may include offering digital interventions to all individuals visiting primary health care clinics or other central community locations, such as pharmacies, grocery stores, and sports arenas. Relatedly, maintaining motivation throughout the behavior change process is critical for promoting long-term, sustained change. In this study, we provided motivational and tip-based reminders throughout the week; however, the optimal approach to sustaining motivation in digital interventions remains unclear. Strategies such as gamification [104] or habit formation [105] may be effective for some individuals. Future research should explore heterogeneity in sustained motivation and behavior change to better understand what drives long-term success.

Conclusion

This study used a factorial design to estimate the effects of a digital multiple health behavior change intervention at the component level. Studying effects at the component level provided novel insights into the building blocks of complex digital health behavior change interventions. In addition, these effects were examined across multiple behaviors. Overall, the screening/feedback component was the most effective for changing health behaviors, along with goal-setting/planning and motivation to change. Several components interacted to further improve health outcomes, most notably screening/feedback with motivation to change, screening/feedback with skills/know-how, and motivation to change with skills/know-how. By contrast, some components also had detrimental effects on certain behaviors, suggesting that not all components are appropriate for all behaviors. The results from this study contribute to the development of more effective interventions by identifying which components of complex interventions are most effective for specific behaviors.

Acknowledgments

Generative artificial intelligence was not used in this study.

Funding

This trial was funded by the Swedish Cancer Society (Cancerfonden; 20 0883 Pj, 23 2888 Pj; principal investigator: MB) and is an extension of the MoBILE (Mobile Health Multiple Lifestyle Behavior Interventions Across the Lifespan) research program, which is funded by the Swedish Research Council for Health, Working Life and Welfare (grant 2018-01410; principal investigator: ML). The funders had no role in the study planning, design, execution, analysis, or presentation of the findings.

Data Availability

The datasets used or analyzed during this study are available from the corresponding author (MB) on reasonable request. Ethical approval is required for data sharing.

Authors' Contributions

Conceptualization: PB, MB, ML

XSL•FO
RenderX

Data curation: JC, MB
Formal analysis: JC, MB
Intervention content and resources: KÅ, JB, OL, HH, PH, PB, ML, MB
Investigation: JB, KÅ, MB
Methodology: MB, ML, PH
Project administration: JB, KÅ, MB
Software: MB
Writing – original draft: JC, MB
Writing – review & editing: All authors

Conflicts of Interest

MB and PB own a private company (Alexit AB) that develops and distributes digital solutions to the public and private sectors. Alexit AB played no role in the development of the intervention, study design, data analysis, data interpretation, or writing of this report. All other authors declare no conflicts of interest.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 992 KB - jmir_v28i1e88881_app1.pdf]

Multimedia Appendix 2
Informed consent materials.
[PDF File (Adobe PDF File), 61 KB - jmir_v28i1e88881_app2.pdf]

Multimedia Appendix 3
Questionnaires.
[PDF File (Adobe PDF File), 75 KB - jmir_v28i1e88881_app3.pdf]

Multimedia Appendix 4
Intervention description and factorial conditions.
[PDF File (Adobe PDF File), 1873 KB - jmir_v28i1e88881_app4.pdf]

Multimedia Appendix 5
Attrition nonresponse analyses.
[PDF File (Adobe PDF File), 150 KB - jmir_v28i1e88881_app5.pdf]

Multimedia Appendix 6
Attrition attempts analyses.
[PDF File (Adobe PDF File), 103 KB - jmir_v28i1e88881_app6.pdf]

Multimedia Appendix 7
Estimates of effects (available data).
[PDF File (Adobe PDF File), 274 KB - jmir_v28i1e88881_app7.pdf]

Multimedia Appendix 8
Estimates of effects (imputed data).
[PDF File (Adobe PDF File), 309 KB - jmir_v28i1e88881_app8.pdf]

Multimedia Appendix 9
Estimates of effects of number of components.
[PDF File (Adobe PDF File), 94 KB - jmir_v28i1e88881_app9.pdf]

References

1. World Health Organization (WHO). Noncommunicable diseases - fact sheet. WHO. 2025. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases [accessed 2026-03-01]
2. World Health Organization (WHO). Noncommunicable Diseases - Progress Monitor 2022. Geneva, Switzerland: World Health Organization; 2022.

XSL•FO
RenderX

3. Ramstedt M, Guttormsson U. Riskbruk av alkohol i Sverige enligt nya riktlinjer. Kortfakta 04. Centralförbundet för alkohol- och narkotikaupplysning. Stockholm, Sweden: Centralförbundet för alkohol- och narkotikaupplysning; 2024 Feb 22. URL: https://www.can.se/app/uploads/2024/02/can-kortfakta-04-riskbruk-av-alkohol-i-sverige-enligt-nya-riktlinjer.pdf [accessed 2026-03-24]
4. Folkhälsan i sverige - arsrapport 2023. Folkhälsomyndigheten. Stockholm, Sweden: Folkhälsomyndigheten; 2023. URL: https://www.folkhalsomyndigheten.se/publikationer-och-material/publikationsarkiv/f/folkhalsan-i-sverige-arsrapport-2023/ [accessed 2026-03-24]
5. Alkohol ökar risken för sjukdomar och skador Internet. Folkhälsomyndigheten. 2024. URL: https://tinyurl.com/95eehuet [accessed 2026-03-01]
6. Alkohollag. Sveriges Riksdag. URL: https://www.riksdagen.se/sv/dokument-och-lagar/dokument/svensk-forfattningssamling/alkohollag-20101622_sfs-2010-1622/ [accessed 2026-03-01]
7. En hållbar och hälsosam livsmedelskonsumtion. Folkhälsomyndigheten och Livsmedelsverket. Stockholm, Sweden: Folkhälsomyndigheten och Livsmedelsverket; 2024. URL: https://www.folkhalsomyndigheten.se/publikationer-och-material/publikationsarkiv/e/en-hallbar-och-halsosam-livsmedelskonsumtion-aterredovisning-av-regeringsuppdrag/ [accessed 2026-03-24]
8. Organisation for Economic Co-operation and Development (OECD). State of Health in the EU - Sweden Country Health Profile 2023. Brussels, Belgium: European Commission; 2023.
9. Disability-adjusted life years per 100 000 population attributable to dietary risks, low physical activity, tobacco and alcohol use in the EU in 2019. European Commission. URL: https://knowledge4policy.ec.europa.eu/health-promotion-knowledge-gateway/eu-burden-non-communicable-diseases-key-risk-factors-3_en [accessed 2026-03-01]
10. Schuit AJ, van Loon AJM, Tijhuis M, Ocké M. Clustering of lifestyle risk factors in a general adult population. Prev Med 2002 Sep;35(3):219-224. [doi: 10.1006/pmed.2002.1064] [Medline: 12202063]
11. Rabel M, Laxy M, Thorand B, Peters A, Schwettmann L, Mess F. Clustering of health-related behavior patterns and demographics. results from the population-based KORA S4/F4 cohort study. Front Public Health 2018;6:387 [FREE Full text] [doi: 10.3389/fpubh.2018.00387] [Medline: 30723712]
12. Meader N, King K, Moe-Byrne T, Wright K, Graham H, Petticrew M, et al. A systematic review on the clustering and co-occurrence of multiple risk behaviours. BMC Public Health 2016 Jul 29;16:657 [FREE Full text] [doi: 10.1186/s12889-016-3373-6] [Medline: 27473458]
13. Klompstra L, Löf M, Björkelund C, Hellenius M, Kallings LV, Orho-Melander M, et al. Co-occurrence of unhealthy lifestyle behaviours in middle-aged adults: findings from the Swedish CArdioPulmonary bioImage Study (SCAPIS). Sci Rep 2024 Oct 01;14(1):22853 [FREE Full text] [doi: 10.1038/s41598-024-71092-0] [Medline: 39353999]
14. Khaw K, Wareham N, Bingham S, Welch A, Luben R, Day N. Combined impact of health behaviours and mortality in men and women: the EPIC-Norfolk prospective population study. PLoS Med 2008 Jan 08;5(1):e12 [FREE Full text] [doi: 10.1371/journal.pmed.0050012] [Medline: 18184033]
15. Kvaavik E, Batty GD, Ursin G, Huxley R, Gale CR. Influence of individual and combined health behaviors on total and cause-specific mortality in men and women: the United Kingdom health and lifestyle survey. Arch Intern Med 2010 Apr 26;170(8):711-718. [doi: 10.1001/archinternmed.2010.76] [Medline: 20421558]
16. Loef M, Walach H. The combined effects of healthy lifestyle behaviors on all cause mortality: a systematic review and meta-analysis. Prev Med 2012 Sep;55(3):163-170. [doi: 10.1016/j.ypmed.2012.06.017] [Medline: 22735042]
17. Riper H, Hoogendoorn A, Cuijpers P, Karyotaki E, Boumparis N, Mira A, et al. Effectiveness and treatment moderators of internet interventions for adult problem drinking: an individual patient data meta-analysis of 19 randomised controlled trials. PLoS Med 2018 Dec;15(12):e1002714 [FREE Full text] [doi: 10.1371/journal.pmed.1002714] [Medline: 30562347]
18. ICT usage in households and by individuals. Statistiska Centralbyrån. 2023. URL: https://www.scb.se/en/finding-statistics/statistics-by-subject-area/living-conditions/living-conditions/ict-usage-in-households-and-by-individuals/ [accessed 2026-03-01]
19. Blomqvist J, Gunnarsson KU, Bendtsen P, Bendtsen M. Effects of a text messaging smoking cessation intervention amongst online help-seekers and primary health care visitors: findings from a randomised controlled trial. BMC Med 2023 Oct 04;21(1):382 [FREE Full text] [doi: 10.1186/s12916-023-03073-5] [Medline: 37794399]
20. Bendtsen M, Åsberg K, McCambridge J. Effectiveness of a digital intervention versus alcohol information for online help-seekers in Sweden: a randomised controlled trial. BMC Med 2022 May 17;20(1):176 [FREE Full text] [doi: 10.1186/s12916-022-02374-5] [Medline: 35578276]
21. Nationella riktlinjer för prevention och behandling vid ohälsosamma levnadsvanor Internet. Socialstyrelsen. 2018. URL: https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/artikelkatalog/nationella-riktlinjer/2018-6-24.pdf [accessed 2026-03-01]
22. Deitz D, Cook RF, Hersch RK, Leaf S. Heart healthy online: an innovative approach to risk reduction in the workplace. J Occup Environ Med 2014 May;56(5):547-553 [FREE Full text] [doi: 10.1097/JOM.0000000000000148] [Medline: 24806568]

XSL•FO
RenderX

23. Oenema A, Brug J, Dijkstra A, de Weerdt I, de Vries H. Efficacy and use of an internet-delivered computer-tailored lifestyle intervention, targeting saturated fat intake, physical activity and smoking cessation: a randomized controlled trial. Ann Behav Med 2008 Apr;35(2):125-135. [doi: 10.1007/s12160-008-9023-1] [Medline: 18363076]
24. Duncan MJ, Vandelanotte C, Trost SG, Rebar AL, Rogers N, Burton NW, et al. Balanced: a randomised trial examining the efficacy of two self-monitoring methods for an app-based multi-behaviour intervention to improve physical activity, sitting and sleep in adults. BMC Public Health 2016 Jul 30;16:670 [FREE Full text] [doi: 10.1186/s12889-016-3256-x] [Medline: 27473327]
25. Schulz DN, Kremers SPJ, Vandelanotte C, van Adrichem MJG, Schneider F, Candel MJJM, et al. Effects of a web-based tailored multiple-lifestyle intervention for adults: a two-year randomized controlled trial comparing sequential and simultaneous delivery modes. J Med Internet Res 2014 Jan 27;16(1):e26 [FREE Full text] [doi: 10.2196/jmir.3094] [Medline: 24472854]
26. Pham Q, Wiljer D, Cafazzo JA. Beyond the randomized controlled trial: a review of alternatives in mHealth clinical trial methods. JMIR Mhealth Uhealth 2016 Sep 09;4(3):e107 [FREE Full text] [doi: 10.2196/mhealth.5720] [Medline: 27613084]
27. Schwarzer R. Health action process approach (HAPA) as a theoretical framework to understand behavior change. AP 2016 Dec 05;30(121):119. [doi: 10.15517/ap.v30i121.23458]
28. West R, Michie S. A brief introduction to the COM-B model of behaviour and the PRIME theory of motivation. Qeios 2020 Apr 07:1-1. [doi: 10.32388/ww04e6]
29. Ajzen I. The theory of planned behavior. Organizational Behavior and Human Decision Processes 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
30. Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomised controlled trials. BMC Med Res Methodol 2003 Nov 24;3:26 [FREE Full text] [doi: 10.1186/1471-2288-3-26] [Medline: 14633287]
31. Bendtsen M, Bendtsen P, Henriksson H, Henriksson P, Müssener U, Thomas K, et al. The mobile health multiple lifestyle behavior interventions across the lifespan (MoBILE) research program: protocol for development, evaluation, and implementation. JMIR Res Protoc 2020 Apr 20;9(4):e14894 [FREE Full text] [doi: 10.2196/14894] [Medline: 32310147]
32. Åsberg K, Blomqvist J, Lundgren O, Henriksson H, Henriksson P, Bendtsen P, et al. Digital multiple health behaviour change intervention targeting online help seekers: protocol for the COACH randomised factorial trial. BMJ Open 2022 Jul 26;12(7):e061024 [FREE Full text] [doi: 10.1136/bmjopen-2022-061024] [Medline: 35882466]
33. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMC Med 2010 Mar 24;8:18 [FREE Full text] [doi: 10.1186/1741-7015-8-18] [Medline: 20334633]
34. Kahan BC, Hall SS, Beller EM, Birchenall M, Chan A, Elbourne D, et al. Reporting of factorial randomized trials: extension of the CONSORT 2010 statement. JAMA 2023 Dec 05;330(21):2106-2114. [doi: 10.1001/jama.2023.19793] [Medline: 38051324]
35. GBD 2016 Alcohol Collaborators. Alcohol use and burden for 195 countries and territories, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet 2018 Sep 22;392(10152):1015-1035 [FREE Full text] [doi: 10.1016/S0140-6736(18)31310-2] [Medline: 30146330]
36. Conner M, Norman P. Predicting and Changing Health Behaviour: Research and Practice with Social Cognition Models. London, England: McGraw-Hill Education (UK); 2015.
37. Michie S, Whittington C, Hamoudi Z, Zarnani F, Tober G, West R. Identification of behaviour change techniques to reduce excessive alcohol consumption. Addiction 2012 Aug;107(8):1431-1440. [doi: 10.1111/j.1360-0443.2012.03845.x] [Medline: 22340523]
38. Bendtsen P, McCambridge J, Bendtsen M, Karlsson N, Nilsen P. Effectiveness of a proactive mail-based alcohol internet intervention for university students: dismantling the assessment and feedback components in a randomized controlled trial. J Med Internet Res 2012 Oct 31;14(5):e142 [FREE Full text] [doi: 10.2196/jmir.2062] [Medline: 23113955]
39. McCambridge J, Bendtsen M, Karlsson N, White IR, Nilsen P, Bendtsen P. Alcohol assessment and feedback by email for university students: main findings from a randomised controlled trial. Br J Psychiatry 2013 Nov;203(5):334-340 [FREE Full text] [doi: 10.1192/bjp.bp.113.128660] [Medline: 24072758]
40. Bendtsen P, Bendtsen M, Karlsson N, White IR, McCambridge J. Online alcohol assessment and feedback for hazardous and harmful drinkers: findings from the AMADEUS-2 randomized controlled trial of routine practice in Swedish universities. J Med Internet Res 2015 Jul 09;17(7):e170 [FREE Full text] [doi: 10.2196/jmir.4020] [Medline: 26159179]
41. Michie S, Abraham C, Whittington C, McAteer J, Gupta S. Effective techniques in healthy eating and physical activity interventions: a meta-regression. Health Psychol 2009 Nov;28(6):690-701. [doi: 10.1037/a0016136] [Medline: 19916637]
42. Murray JM, Brennan SF, French DP, Patterson CC, Kee F, Hunter RF. Effectiveness of physical activity interventions in achieving behaviour change maintenance in young and middle aged adults: a systematic review and meta-analysis. Soc Sci Med 2017 Nov;192:125-133. [doi: 10.1016/j.socscimed.2017.09.021] [Medline: 28965003]
43. Knittle K, Nurmi J, Crutzen R, Hankonen N, Beattie M, Dombrowski SU. How can interventions increase motivation for physical activity? A systematic review and meta-analysis. Health Psychol Rev 2018 Sep;12(3):211-230. [doi: 10.1080/17437199.2018.1435299] [Medline: 29385950]
44. Howlett N, Trivedi D, Troop NA, Chater AM. Are physical activity interventions for healthy inactive adults effective in promoting behavior change and maintenance, and which behavior change techniques are effective? A systematic review

XSL•FO
RenderX

and meta-analysis. Transl Behav Med 2019 Jan 01;9(1):147-157 [FREE Full text] [doi: 10.1093/tbm/iby010] [Medline: 29506209]
45. Ashton LM, Sharkey T, Whatnall MC, Williams RL, Bezzina A, Aguiar EJ, et al. Effectiveness of interventions and behaviour change techniques for improving dietary intake in young adults: a systematic review and meta-analysis of RCTs. Nutrients 2019 Apr 11;11(4) [FREE Full text] [doi: 10.3390/nu11040825] [Medline: 30979065]
46. Garnett CV, Crane D, Brown J, Kaner EFS, Beyer FR, Muirhead CR, et al. Behavior change techniques used in digital behavior change interventions to reduce excessive alcohol consumption: a meta-regression. Ann Behav Med 2018 May 18;52(6):530-543 [FREE Full text] [doi: 10.1093/abm/kax029] [Medline: 29788261]
47. McCrabb S, Baker AL, Attia J, Skelton E, Twyman L, Palazzi K, et al. Internet-based programs incorporating behavior change techniques are associated with increased smoking cessation in the general population: a systematic review and meta-analysis. Ann Behav Med 2019 Feb 01;53(2):180-195. [doi: 10.1093/abm/kay026] [Medline: 29750240]
48. Ek A, Alexandrou C, Söderström E, Bergman P, Delisle Nyström C, Direito A, et al. Effectiveness of a 3-month mobile phone-based behavior change program on active transportation and physical activity in adults: randomized controlled trial. JMIR Mhealth Uhealth 2020 Jun 08;8(6):e18531 [FREE Full text] [doi: 10.2196/18531] [Medline: 32510462]
49. Newby K, Teah G, Cooke R, Li X, Brown K, Salisbury-Finch B, et al. Do automated digital health behaviour change interventions have a positive effect on self-efficacy? A systematic review and meta-analysis. Health Psychol Rev 2021 Mar;15(1):140-158. [doi: 10.1080/17437199.2019.1705873] [Medline: 31847702]
50. Müssener U, Bendtsen M, Karlsson N, White IR, McCambridge J, Bendtsen P. Effectiveness of short message service text-based smoking cessation intervention among university students: a randomized clinical trial. JAMA Intern Med 2016 Mar;176(3):321-328 [FREE Full text] [doi: 10.1001/jamainternmed.2015.8260] [Medline: 26903176]
51. Müssener U, Bendtsen M, Karlsson N, White IR, McCambridge J, Bendtsen P. SMS-based smoking cessation intervention among university students: study protocol for a randomised controlled trial (NEXit trial). Trials 2015 Apr 08;16:140 [FREE Full text] [doi: 10.1186/s13063-015-0640-2] [Medline: 25872503]
52. Thomas K, Bendtsen M, Linderoth C, Karlsson N, Bendtsen P, Müssener U. Short message service (SMS)-based intervention targeting alcohol consumption among university students: study protocol of a randomized controlled trial. Trials 2017 Apr 04;18(1):156 [FREE Full text] [doi: 10.1186/s13063-017-1898-3] [Medline: 28372563]
53. Thomas K, Müssener U, Linderoth C, Karlsson N, Bendtsen P, Bendtsen M. Effectiveness of a text messaging-based intervention targeting alcohol consumption among university students: randomized controlled trial. JMIR Mhealth Uhealth 2018 Jun 25;6(6):e146 [FREE Full text] [doi: 10.2196/mhealth.9642] [Medline: 29941417]
54. Müssener U, Bendtsen M, McCambridge J, Bendtsen P. User satisfaction with the structure and content of the NEXit intervention, a text messaging-based smoking cessation programme. BMC Public Health 2016 Nov 22;16(1):1179 [FREE Full text] [doi: 10.1186/s12889-016-3848-5] [Medline: 27876031]
55. Müssener U, Thomas K, Linderoth C, Leijon M, Bendtsen M. A text message-based intervention targeting alcohol consumption among university students: user satisfaction and acceptability study. JMIR Hum Factors 2018 Jul 10;5(3):e23 [FREE Full text] [doi: 10.2196/humanfactors.9641] [Medline: 29991469]
56. Thomas K, Linderoth C, Bendtsen M, Bendtsen P, Müssener U. Text message-based intervention targeting alcohol consumption among university students: findings from a formative development study. JMIR Mhealth Uhealth 2016 Oct 20;4(4):e119 [FREE Full text] [doi: 10.2196/mhealth.5863] [Medline: 27765732]
57. Kaner EF, Beyer FR, Garnett C, Crane D, Brown J, Muirhead C, et al. Personalised digital interventions for reducing hazardous and harmful alcohol consumption in community-dwelling populations. Cochrane Database Syst Rev 2017 Sep 25;9(9):CD011479 [FREE Full text] [doi: 10.1002/14651858.CD011479.pub2] [Medline: 28944453]
58. Black N, Johnston M, Michie S, Hartmann‐Boyce J, West R, Viechtbauer W, et al. Behaviour change techniques associated with smoking cessation in intervention and comparator groups of randomized controlled trials: a systematic review and meta‐regression. Addiction 2020 Apr 16;115(11):2008-2020. [doi: 10.1111/add.15056]
59. Crane RS, Brewer J, Feldman C, Kabat-Zinn J, Santorelli S, Williams JMG, et al. What defines mindfulness-based programs? The warp and the weft. Psychol Med 2017 Apr;47(6):990-999. [doi: 10.1017/S0033291716003317] [Medline: 28031068]
60. Creswell JD. Mindfulness interventions. Annu Rev Psychol 2017 Jan 03;68:491-516. [doi: 10.1146/annurev-psych-042716-051139] [Medline: 27687118]
61. Wong SYS, Chan JYC, Zhang D, Lee EKP, Tsoi KKF. The safety of mindfulness-based interventions: a systematic review of randomized controlled trials. Mindfulness 2018 Feb 2;9(5):1344-1357. [doi: 10.1007/s12671-018-0897-0]
62. Galante J, Friedrich C, Dawson AF, Modrego-Alarcón M, Gebbing P, Delgado-Suárez I, et al. Mindfulness-based programmes for mental health promotion in adults in nonclinical settings: a systematic review and meta-analysis of randomised controlled trials. PLoS Med 2021 Jan;18(1):e1003481 [FREE Full text] [doi: 10.1371/journal.pmed.1003481] [Medline: 33428616]
63. Neff KD, Germer CK. A pilot study and randomized controlled trial of the mindful self-compassion program. J Clin Psychol 2013 Jan;69(1):28-44. [doi: 10.1002/jclp.21923] [Medline: 23070875]
64. Crawford J, Collier ES, Bendtsen M. Individualized treatment effects of a digital alcohol intervention and their associations with participant characteristics and engagement. Alcohol Alcohol 2024 Jul 21;59(5):agae049 [FREE Full text] [doi: 10.1093/alcalc/agae049] [Medline: 39034147]

XSL•FO
RenderX

65. Rehm J. Measuring quantity, frequency, and volume of drinking. Alcohol Clin Exp Res 1998 Apr;22(2 Suppl):4S-14S. [doi: 10.1097/00000374-199802001-00002] [Medline: 9603301]
66. Shorter GW, Bray JW, Giles EL, O'Donnell AJ, Berman AH, Holloway A, et al. The variability of outcomes used in efficacy and effectiveness trials of alcohol brief interventions: a systematic review. J Stud Alcohol Drugs 2019 May;80(3):286-298. [Medline: 31250793]
67. Shorter GW, Bray JW, Heather N, Berman AH, Giles EL, Clarke M, et al. The "Outcome Reporting in Brief Intervention Trials: Alcohol" (ORBITAL) core outcome set: international consensus on outcomes to measure in efficacy and effectiveness trials of alcohol brief interventions. J Stud Alcohol Drugs 2021 Sep;82(5):638-646 [FREE Full text] [doi: 10.15288/jsad.2021.82.638] [Medline: 34546911]
68. Benowitz NL, Bernert JT, Foulds J, Hecht SS, Jacob P, Jarvis MJ, et al. Biochemical verification of tobacco use and abstinence: 2019 update. Nicotine Tob Res 2020 Jun 12;22(7):1086-1097 [FREE Full text] [doi: 10.1093/ntr/ntz132] [Medline: 31570931]
69. Vallejo MA, Vallejo-Slocker L, Fernández-Abascal EG, Mañanes G. Determining factors for stress perception assessed with the Perceived Stress Scale (PSS-4) in Spanish and other European samples. Front Psychol 2018;9:37 [FREE Full text] [doi: 10.3389/fpsyg.2018.00037] [Medline: 29434563]
70. Hays RD, Bjorner JB, Revicki DA, Spritzer KL, Cella D. Development of physical and mental health summary scores from the Patient-Reported Outcomes Measurement Information System (PROMIS) global items. Qual Life Res 2009 Sep;18(7):873-880 [FREE Full text] [doi: 10.1007/s11136-009-9496-9] [Medline: 19543809]
71. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med 2011 Feb 20;30(4):377-399. [doi: 10.1002/sim.4067] [Medline: 21225900]
72. Bendtsen M. A gentle introduction to the comparison between null hypothesis testing and bayesian analysis: reanalysis of two randomized controlled trials. J Med Internet Res 2018 Oct 24;20(10):e10873 [FREE Full text] [doi: 10.2196/10873] [Medline: 30148453]
73. Bendtsen M. Avoiding under- and overrecruitment in behavioral intervention trials using Bayesian sequential designs: tutorial. J Med Internet Res 2022 Dec 16;24(12):e40730 [FREE Full text] [doi: 10.2196/40730] [Medline: 36525297]
74. Suffoletto B, Pacella M, Huber J, Chung T. Effectiveness of text message interventions with different behavior change techniques on alcohol consumption among young adults: a five-arm randomized controlled trial. Addiction 2023 Feb;118(2):265-275. [doi: 10.1111/add.16074] [Medline: 36305694]
75. Garnett C, Dinu L, Oldham M, Perski O, Loebenberg G, Beard E, et al. Do engagement and behavioural mechanisms underpin the effectiveness of the Drink Less app? NPJ Digit Med 2024 Jun 29;7(1):174 [FREE Full text] [doi: 10.1038/s41746-024-01169-7] [Medline: 38951560]
76. DiClemente CC, Prochaska JO. Self-change and therapy change of smoking behavior: a comparison of processes of change in cessation and maintenance. Addict Behav 1982;7(2):133-142. [doi: 10.1016/0306-4603(82)90038-7] [Medline: 7102444]
77. McKee SA, O'Malley SS, Salovey P, Krishnan-Sarin S, Mazure CM. Perceived risks and benefits of smoking cessation: gender-specific predictors of motivation and treatment outcome. Addict Behav 2005 Mar;30(3):423-435. [doi: 10.1016/j.addbeh.2004.05.027] [Medline: 15718060]
78. Kaufmann A, Malloy EJ, Haaga DAF. Examining outcome expectancies for smoking vs. abstinence among adult daily smokers. Addict Behav 2020 Mar;102:106140 [FREE Full text] [doi: 10.1016/j.addbeh.2019.106140] [Medline: 31706142]
79. Crawford J, Blomqvist J, Ulfsdotter Gunnarsson K, Bendtsen P, Bendtsen M. Individualized treatment effects of a digital smoking cessation intervention among individuals looking online for help: secondary analysis of a randomized controlled trial. JMIR Mhealth Uhealth 2026 Feb 11;14:e63578 [FREE Full text] [doi: 10.2196/63578] [Medline: 41671427]
80. Baer RA, Carmody J, Hunsinger M. Weekly change in mindfulness and perceived stress in a mindfulness-based stress reduction program. J Clin Psychol 2012 Jul;68(7):755-765. [doi: 10.1002/jclp.21865] [Medline: 22623334]
81. Lundgren O, Garvin P, Kristenson M, Jonasson L, Thylén I. A journey through chaos and calmness: experiences of mindfulness training in patients with depressive symptoms after a recent coronary event - a qualitative diary content analysis. BMC Psychol 2018 Sep 13;6(1):46 [FREE Full text] [doi: 10.1186/s40359-018-0252-1] [Medline: 30213276]
82. König LM, Attig C, Franke T, Renner B. Barriers to and facilitators for using nutrition apps: systematic review and conceptual framework. JMIR Mhealth Uhealth 2021 Apr 01;9(6):e20037 [FREE Full text] [doi: 10.2196/20037] [Medline: 34254938]
83. Ambeba EJ, Ye L, Sereika SM, Styn MA, Acharya SD, Sevick MA, et al. The use of mHealth to deliver tailored messages reduces reported energy and fat intake. J Cardiovasc Nurs 2015;30(1):35-43 [FREE Full text] [doi: 10.1097/JCN.0000000000000120] [Medline: 24434827]
84. Turk MW, Elci OU, Wang J, Sereika SM, Ewing LJ, Acharya SD, et al. Self-monitoring as a mediator of weight loss in the SMART randomized clinical trial. Int J Behav Med 2013 Dec;20(4):556-561 [FREE Full text] [doi: 10.1007/s12529-012-9259-9] [Medline: 22936524]
85. Eva KW, Armson H, Holmboe E, Lockyer J, Loney E, Mann K, et al. Factors influencing responsiveness to feedback: on the interplay between fear, confidence, and reasoning processes. Adv Health Sci Educ Theory Pract 2012 Mar;17(1):15-26 [FREE Full text] [doi: 10.1007/s10459-011-9290-7] [Medline: 21468778]
86. Lara J, Evans EH, O'Brien N, Moynihan PJ, Meyer TD, Adamson AJ, et al. Association of behaviour change techniques with effectiveness of dietary interventions among adults of retirement age: a systematic review and meta-analysis of

XSL•FO
RenderX

randomised controlled trials. BMC Med 2014 Oct 07;12:177 [FREE Full text] [doi: 10.1186/s12916-014-0177-3] [Medline: 25288375]
87. Livingstone KM, Rawstorn JC, Partridge SR, Godrich SL, McNaughton SA, Hendrie GA, et al. Digital behaviour change interventions to increase vegetable intake in adults: a systematic review. Int J Behav Nutr Phys Act 2023 Mar 27;20(1):36 [FREE Full text] [doi: 10.1186/s12966-023-01439-9] [Medline: 36973716]
88. Bandura A. Social foundations of Thoughtaction: A Social Cognitive Theory. Englewood Cliffs, NJ: Prentice-Hall, Inc; 1986.
89. Prochaska JO, DiClemente CC. Stages and processes of self-change of smoking: toward an integrative model of change. Journal of Consulting and Clinical Psychology 1983;51(3):390-395. [doi: 10.1037/0022-006X.51.3.390]
90. de Vries H, Kremers SPJ, Smeets T, Brug J, Eijmael K. The effectiveness of tailored feedback and action plans in an intervention addressing multiple health behaviors. Am J Health Promot 2008;22(6):417-425. [doi: 10.4278/ajhp.22.6.417] [Medline: 18677882]
91. Garstang KR, Jackman PC, Healy LC, Cooper SB, Magistro D. What effect do goal setting interventions have on physical activity and psychological outcomes in insufficiently active adults? A systematic review and meta-analysis. J Phys Act Health 2024 Jun 01;21(6):541-553. [doi: 10.1123/jpah.2023-0340] [Medline: 38560998]
92. Rhodes R, Grant S, De BG. Planning and implementation intention interventions. In: Hagger MS, Cameron LD, Hamilton K, Hankonen N, Lintunen T, editors. The Handbook of Behavior Change. Cambridge, UK: Cambridge University Press; 2020:572.
93. Sala M, Rochefort C, Lui PP, Baldwin AS. Trait mindfulness and health behaviours: a meta-analysis. Health Psychol Rev 2020 Sep 11;14(3):345-393. [doi: 10.1080/17437199.2019.1650290] [Medline: 31362588]
94. Carrière K, Khoury B, Günak MM, Knäuper B. Mindfulness-based interventions for weight loss: a systematic review and meta-analysis. Obes Rev 2018 Feb;19(2):164-177. [doi: 10.1111/obr.12623] [Medline: 29076610]
95. Turk MW, Yang K, Hravnak M, Sereika SM, Ewing LJ, Burke LE. Randomized clinical trials of weight loss maintenance: a review. J Cardiovasc Nurs 2009;24(1):58-80 [FREE Full text] [doi: 10.1097/01.JCN.0000317471.58048.32] [Medline: 19114803]
96. Greaves CJ, Poltawski L, van Beurden SB, Price L, Taylor RS, Merrifield R, et al. Addressing the psychology of weight loss and maintenance: a feasibility study of the Skills for Weight Loss and Maintenance Weight Management Programme. Br J Health Psychol 2023 Nov;28(4):1261-1282 [FREE Full text] [doi: 10.1111/bjhp.12683] [Medline: 37487646]
97. Kriakous SA, Elliott KA, Lamers C, Owen R. The effectiveness of mindfulness-based stress reduction on the psychological functioning of healthcare professionals: a systematic review. Mindfulness (N Y) 2021;12(1):1-28 [FREE Full text] [doi: 10.1007/s12671-020-01500-9] [Medline: 32989406]
98. Bossi F, Zaninotto F, D'Arcangelo S, Lattanzi N, Malizia AP, Ricciardi E. Mindfulness-based online intervention increases well-being and decreases stress after Covid-19 lockdown. Sci Rep 2022 Apr 20;12(1):6483 [FREE Full text] [doi: 10.1038/s41598-022-10361-2] [Medline: 35444198]
99. Flanagan S, Damery S, Combes G. The effectiveness of integrated care interventions in improving patient quality of life (QoL) for patients with chronic conditions. An overview of the systematic review evidence. Health Qual Life Outcomes 2017 Sep 29;15(1):188 [FREE Full text] [doi: 10.1186/s12955-017-0765-y] [Medline: 28962570]
100. Kypri K, Wilson A, Attia J, Sheeran P, Miller P, McCambridge J. Social desirability bias in the reporting of alcohol consumption: a randomized trial. J Stud Alcohol Drugs 2016 May;77(3):526-531. [doi: 10.15288/jsad.2016.77.526] [Medline: 27172587]
101. Miller TM, Abdel-Maksoud MF, Crane LA, Marcus AC, Byers TE. Effects of social approval bias on self-reported fruit and vegetable consumption: a randomized controlled trial. Nutr J 2008 Jun 27;7:18 [FREE Full text] [doi: 10.1186/1475-2891-7-18] [Medline: 18588696]
102. Prince SA, Adamo KB, Hamel ME, Hardt J, Connor Gorber S, Tremblay M. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. Int J Behav Nutr Phys Act 2008 Nov 06;5:56 [FREE Full text] [doi: 10.1186/1479-5868-5-56] [Medline: 18990237]
103. Williams J, Rakovac I, Loyola E, Sturua L, Maglakelidze N, Gamkrelidze A, et al. A comparison of self-reported to cotinine-detected smoking status among adults in Georgia. Eur J Public Health 2020 Oct 01;30(5):1007-1012 [FREE Full text] [doi: 10.1093/eurpub/ckaa093] [Medline: 32588045]
104. El-Hilly AA, Iqbal SS, Ahmed M, Sherwani Y, Muntasir M, Siddiqui S, et al. Game on? Smoking cessation through the gamification of mHealth: a longitudinal qualitative study. JMIR Serious Games 2016 Oct 24;4(2):e18 [FREE Full text] [doi: 10.2196/games.5678] [Medline: 27777216]
105. Zhu Y, Long Y, Wang H, Lee KP, Zhang L, Wang SJ. Digital behavior change intervention designs for habit formation: systematic review. J Med Internet Res 2024 May 24;26:e54375 [FREE Full text] [doi: 10.2196/54375] [Medline: 38787601]

Abbreviations

CoI: compatibility interval
COM-B: Capability-Opportunity-Motivation-Behavior

XSL•FO
RenderX

CONSORT: Consolidated Standards of Reporting Trials
DALY: disability-adjusted life year
IRR: incidence rate ratio
MoBILE: Mobile Health Multiple Lifestyle Behavior Interventions Across the Lifespan
MVPA: moderate-to-vigorous physical activity
OR: odds ratio
POE: probability of effect
PROMIS: Patient-Reported Outcomes Measurement Information System
QoL: quality of life

Edited by A Stone; submitted 03.Dec.2025; peer-reviewed by A Eisingerich, L Cheng; comments to author 05.Jan.2026; revised version received 15.Mar.2026; accepted 16.Mar.2026; published 02.Apr.2026.

Please cite as:
Crawford J, Åsberg K, Blomqvist J, Lundgren O, Henriksson H, Henriksson P, Bendtsen P, Löf M, Bendtsen M
Effectiveness of the Components of a Digital Multiple Health Behavior Change Intervention Among Individuals Seeking Help Online (Coach): Factorial Randomized Trial
J Med Internet Res 2026;28:e88881
URL: https://www.jmir.org/2026/1/e88881
doi:10.2196/88881
PMID:

©Joel Crawford, Katarina Åsberg, Jenny Blomqvist, Oskar Lundgren, Hanna Henriksson, Pontus Henriksson, Preben Bendtsen, Marie Löf, Marcus Bendtsen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Preferences for Personalized Text Message Appointment Reminders Among Outpatients in a Universal Health System: Cross-Sectional Study

Shih-Huai Hsiao[1,2,3], MPH; Yu-Chen Chuang[4*], MSc; Shu-Ling Ma[5], MSc; Hui-Tzu Lin[6], MPH; I-Chen Lee[7*], PhD; Pei-Shih Chen[2], PhD

[1]Tele-Healthcare Center, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan

[2]Department of Public Health, College of Health Science, Kaohsiung Medical University, No. 100 Shih-Chuan 1st Road, Sammin District, Kaohsiung, Taiwan

[3]Taiwan Health Insurance Association, Taipei, Taiwan

[4]Cardiovascular Division, Department of Internal Medicine, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan

[5]Department of Nursing, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan

[6]Clinical Skill Training Center, Department of Clinical Education and Skill Training, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan

[7]Department of Healthcare Administration and Medical Informatics, College of Health Sciences, Kaohsiung Medical University, Kaohsiung, Taiwan

[*]these authors contributed equally

Corresponding Author:
Pei-Shih Chen, PhD
Department of Public Health, College of Health Science, Kaohsiung Medical University, No. 100 Shih-Chuan 1st Road, Sammin District, Kaohsiung, Taiwan

Abstract

Background: SMS text messaging reminders are widely used to reduce missed outpatient appointments; however, evidence remains limited regarding which types of reminder content patients prefer, particularly within East Asian universal health systems. In Taiwan, minimal financial barriers to care and unrestricted access to secondary and tertiary hospitals contribute to high outpatient visit volumes and persistent no-show rates. These contextual features underscore the need for behaviorally informed and demographically tailored reminder strategies rather than uniform messaging approaches.

Objective: This study aimed to examine patient preferences for 6 theory-guided SMS appointment reminder types and to identify the predictors of reminder preference related to demographic characteristics and health care utilization, with the goal of informing personalized reminder design for a forthcoming randomized controlled trial.

Methods: We conducted a cross-sectional online survey among adults in Taiwan with prior outpatient experience. Six SMS reminder prototypes were developed based on behavioral communication principles and validated by a multidisciplinary expert panel using item-level content validity indices. Participants selected their preferred SMS reminder type and reported sociodemographic characteristics and recent health care utilization. Bivariate associations were examined using chi-square tests and one-way ANOVA, with Benjamini-Hochberg false discovery rate correction applied to control for multiple testing. To identify independent predictors of SMS reminder preference while adjusting for potential confounding, we fitted a multinomial logistic regression model with all covariates entered simultaneously.

Results: A total of 1095 respondents completed the survey. General reminders and messages referencing prior missed appointments were most frequently preferred, whereas empathy-based or relationally framed messages were selected less often. In false discovery rate–adjusted univariate analyses, both age and sex were associated with SMS reminder preference. However, in the fully adjusted multinomial logistic regression model, age emerged as the only statistically significant independent predictor. Participants younger than 50 years were significantly more likely to prefer alternative reminder message types compared with the general reminder (adjusted odds ratio 1.64, 95% CI 1.18 - 2.28; *P*=.003). Sex did not retain statistical significance after multivariable adjustment. Other sociodemographic characteristics and health care utilization variables, including education level, employment status, residential region, outpatient visit frequency, and recent missed appointments history, were not independently associated with reminder preference.

Conclusions: Preferences for outpatient SMS reminder content vary systematically, with age representing the most robust independent predictor. Across the sample, concise and behavior-focused reminders were preferred over empathy-oriented or relational formats. These findings support age-informed tailoring of SMS reminder content and provide content-validated SMS prototypes for use in subsequent interventional research. The results offer formative evidence to guide the design of randomized

XSL•FO
RenderX

trials aimed at reducing outpatient no-shows and improving the efficiency of ambulatory care delivery in Taiwan's universal health care system.

(J Med Internet Res 2026;28:e81324) doi:10.2196/81324

KEYWORDS

SMS reminders; outpatient no-shows; patient preferences; behavioral design; health communication; personalized messaging; universal health coverage

Introduction

Timely attendance at outpatient appointments is fundamental to maintaining care continuity, optimizing health care resource utilization, and achieving favorable clinical outcomes [1]. However, missed appointments (MAs), commonly referred to as "no-shows," remain a persistent global challenge, with no-show rates in outpatient settings averaging approximately 23% [2]. These disruptions not only compromise the efficiency of health care delivery but also contribute to inequities in access and outcomes, particularly among patients facing socioeconomic disadvantages, transportation barriers, or prolonged wait times in specialty care services, such as cardiology and dermatology [3-6].

To address this issue, SMS reminders have been widely implemented and demonstrated to be effective in reducing no-show rates across various clinical contexts, including gastrointestinal procedures [7-9] and surgeries requiring anesthesia [10]. Compared with traditional telephone calls, which are labor-intensive and less scalable, SMS reminders offer a cost-effective and scalable solution for promoting appointment adherence [11]. Moreover, behaviorally informed message designs, such as emotionally salient language, appeals to social responsibility, or moral framing, have been shown to enhance patient engagement and accountability [7,12-14]. The inclusion of personalized content, such as references to prior MAs, has also been associated with further reductions in no-show rates [10,15,16].

Despite the growing body of evidence supporting the efficacy of SMS reminders, limited research has examined the optimal design of message content. Specifically, there is a lack of systematic investigation into which core components should be included, how linguistic and emotional framing influences patient behavior, and to what extent content should be tailored to individual characteristics. These characteristics may include insurance type, age, occupation, socioeconomic status, prior attendance history, disease severity, or the need for chronic follow-up care. Personalization strategies of this nature are likely critical to maximizing the impact of reminder interventions but have been insufficiently explored in the literature.

This gap is particularly salient in East Asian health care systems, such as Taiwan's, where universal health coverage and unrestricted access to secondary and tertiary care may lead to distinct patterns of health care utilization. Under Taiwan's National Health Insurance program, patients can access higher-level hospitals directly without referral and with minimal out-of-pocket expense, often resulting in frequent outpatient visits for relatively minor conditions [17]. The ease of access and low perceived urgency of such visits may contribute to elevated no-show rates; one tertiary medical center in Taiwan reported a no-show rate as high as 16.8% [18]. These contextual factors highlight the need for more effective and targeted reminder strategies that are responsive to patient behavior and health care system dynamics.

Specifically, the study aimed to (1) assess whether there was significant heterogeneity in preferences for various types of SMS content and (2) examine the extent to which these preferences were associated with patients' sociodemographic characteristics, socioeconomic status, and recent health care utilization patterns.

Methods

Study Design and Participant Recruitment

This cross-sectional study used a nonprobability convenience sampling method. Participants were recruited through online platforms (eg, Facebook, LINE) and professional networks, including outreach by health care professionals.

Participants were eligible if they were adults aged 20 years or older, resided in Taiwan, had experience using SMS-capable mobile phones, and completed the online questionnaire with informed consent. Individuals were excluded if they lacked SMS access, had known cognitive impairments, or reported working in health care–related professions (eg, hospitals or clinics) to minimize potential response bias. The structured questionnaire was distributed electronically, and the data were collected anonymously over a 14-day period from February 25 to March 10, 2025. A total of 1152 responses were received, of which 1095 were valid, yielding a valid response rate of 95.05%.

Instrument Development and Content Validity

SMS Reminder Messages Development

Grounded in behavioral theory and prior empirical research, this study developed 6 SMS appointment reminder prototypes that incorporate cognitive, emotional, and social cues to enhance patient relevance and behavioral efficacy. Message design was guided by the SMS content framework proposed by Acharya et al [19], with 4 core principles applied: the use of plain language, a neutral tone, explicit appointment details, and clear sender identification. Three principles, namely the inclusion of external links, preparation instructions, and emphasis on the consequences of nonattendance, were considered less applicable to general outpatient contexts. While the prototypes largely conformed to established design standards, contextual adaptation remains essential when applying screening-focused frameworks

XSL•FO
RenderX

to broader clinical settings. The 6 SMS message types are summarized in Table 1 and are described as follows:

1. General reminder: a neutral notification containing appointment details [9-11,20].
2. Prior MA: referencing the patient's history of missed visits and institutional policies. This type can further reduce the no-show rate by approximately 1% compared to standard reminders [14] and promote behavioral change through increased awareness of institutional norms [10].
3. Provider-focused empathy: highlighting the emotional burden and workload of health care staff, aiming to evoke moral responsibility and cooperation, especially in high-burden clinical settings [12].
4. Patient-focused empathy: emphasizing how MA may limit others' access to care, particularly under Taiwan's hospital-specific global budgeting system, where physician time and service volume are constrained. Such awareness may foster social responsibility and encourage patients to proactively cancel or reschedule. Empathy-based reminders have also been shown to reduce provider-patient tension [21].
5. Physician follow-up: simulating concern from the attending physician by referencing the patient's upcoming visit and inviting them to return on the scheduled date. This tone, regardless of prior clinical contact, may enhance engagement and continuity of care, especially for patients with chronic conditions or long-term follow-up needs [22].
6. Supportive reminder: using friendly, conversational language to reduce psychological barriers and ease anxiety related to clinic visits. Warm, supportive messaging may improve follow-up motivation and engagement, particularly for patients who perceive medical visits as stressful [14].

The 6 SMS messages in their original Chinese form have been included in Multimedia Appendix 1.

Table . Theory-guided SMS reminder types with message content and technical specifications for outpatient appointment adherence.

Type of SMS reminder	Wording of SMS content in this study
General reminders	Hi! You have an appt. with [Dr Name] on the [Morning/Afternoon/Evening] of [Date]. To cancel or reschedule, please visit [clinic or hospital] booking system. We care about your health!
Prior missed appointment behavior	Hi! You have an appt. with [Dr Name] on the [Morning/Afternoon/Evening] of [Date]. Your Appt. No. is [Appt. No.]. If this MA results in a total of three no-shows without prior cancellation, you will no longer be eligible for remote booking. Future appt. must be made in person at the [clinic or hospital]. To cancel or change, please visit [clinic or hospital] booking system. We care about your health!
Joint empathy communication type 1: provider-focused	Hi! You have an appt. with [Dr Name] on the [Morning/Afternoon/Evening] of [Date]. Your Appt. No. is [Appt. No.]. If you don't show up and don't cancel, it may cause trouble for medical staff. To cancel or reschedule, please visit [clinic or hospital] booking system. We care about your health!
Joint empathy communication type 2: patient-focused	Hi! You have an appt. with [Dr Name] on the [Morning/Afternoon/Evening] of [Date]. Your Appt. No. is [Number]. If you miss your appt. without cancellation, it may affect other patients' access to care. To cancel or reschedule, please visit [clinic or hospital booking system]. We care about your health!
Gentle and friendly (physician follow-up)	Hi! [Dr Name] cares about you. You have a [specialty] appt. on the [Morning/Afternoon/Evening] of [Date]. To cancel or reschedule, please visit [clinic or hospital] booking system. We care about your health!
Supportive reminder	Hope you're doing well! [Dr Name] cares about your health. You have a [specialty] appt. on the [Morning/Afternoon/Evening] of [Date]. Your Appt. No. is [number]. To cancel or reschedule, please visit [clinic or hospital] booking system. We care about your health!

Content Validity Examination of SMS Message

To evaluate the content validity of the SMS messages, a multidisciplinary panel of 12 experts was convened. The panel included 2 clinicians, 3 managers from patient registration and cashier services, 2 language educators, 3 nursing supervisors responsible for outpatient care, and 2 professionals from insurance and customer service departments. All panel members had prior personal experience as patients or caregivers and demonstrated strong communication competencies, ensuring a broad range of perspectives from clinical, administrative, linguistic, and consumer viewpoints. No significant demographic differences were found among the expert subgroups (P>.05).

Each SMS prototype was assessed using a 5-point Likert scale across 3 domains: semantic clarity, content relevance, and ease of comprehension. The item-level content validity index for all 6 messages exceeded the standard threshold of 0.78 [17,23], confirming adequate content validity.

Statistical Analysis

Descriptive statistics summarized participant characteristics. Chi-square tests and one-way ANOVA were used to examine bivariate associations between participant characteristics and

XSL•FO
RenderX

SMS reminder preferences. To control for inflated type 1 error due to multiple comparisons, the Benjamini-Hochberg false discovery rate (FDR) procedure was applied to the bivariate tests [24].

To adjust for potential confounding and identify independent predictors of SMS reminder preference, we fitted a multinomial logistic regression model with SMS reminder preference as the outcome. The general reminder SMS was specified as the reference category. Demographic characteristics and health care utilization variables were entered simultaneously as covariates, and the results are reported as adjusted odds ratios with 95% CI. Statistical significance was set at $P<.05$ (2-tailed). All analyses were conducted using SPSS (version 29.0.1.0).

Ethical Considerations

This study was reviewed and approved by the Institutional Review Board of Kaohsiung Medical University Chung-Ho Memorial Hospital (IRB number: KMUHIRB-E(I)-20240047; approval date: February 3, 2025). Electronic informed consent was obtained from all participants prior to survey submission. Participation was voluntary, anonymity was ensured, and no identifiable personal information was collected. All study procedures adhered to institutional ethical guidelines and the principles of the Declaration of Helsinki.

Results

Participant Characteristics

Demographic and Socioeconomic Status

Among the 1095 participants, female participants predominated (n=851, 77.7%). The mean age was 43.2 years (SD 12.37), with 65.8% (n=720) aged below 50 years. Most (n=726, 66.3%) respondents held a university or college degree, followed by postgraduate qualifications (n=272, 24.8%), and secondary-level education or below (n=97, 8.9%).

The largest occupational groups were professionals (n=352, 32.1%) and clerical and administrative staff (n=254, 23.2%), with smaller proportions in sales or service roles, technical support, senior management, and unemployed categories. Nearly half (n=515, 47%) resided in the Kaohsiung-Pingtung region, and the majority (n=800, 73.1%) lived with others. A minority (n=216, 19.7%) were unmarried or living alone (n=79, 7.2%; Table 2).

XSL•FO
RenderX

Table . Demographic and socioeconomic distribution of the respondents.

Variable	Respondents
Sex, n (%)	
Male	244 (22.3)
Female	851 (77.7)
Age (y), mean (SD)	43.2 (12.38)
Age group (y), n (%)	1095 (100)
<50	720 (65.8)
≥50	375 (34.2)
Education level, n (%)	
Senior high school or lower education	97 (8.9)
College or university education	726 (66.3)
Postgraduate degree	272 (24.8)
Occupation category, n (%)	
Senior executives and decision-makers	114 (10.4)
Professionals	352 (32.1)
Technicians and assistants	96 (8.8)
Clerical and administrative staff	254 (23.2)
Sales and service workers	147 (13.4)
Other occupations	54 (4.9)
Unemployed	78 (7.1)
Employment status, n (%)	
Retired	109 (10)
Employed	897 (81.9)
Unemployed	89 (8.1)
Region of residence, n (%)	
Kaohsiung-Pingtung region	515 (47)
Outside Kaohsiung-Pingtung region	580 (53)
Living arrangement, n (%)	
Unmarried	216 (19.7)
Living alone	79 (7.2)
Living with others	800 (73.1)

XSL•FO
RenderX

Variable	Respondents
Outpatient visits in the past 3 months, n (%)	
0 visit	192 (17.5)
1 - 3 visits	658 (60.1)
4 or more visits	245 (22.4)
Missed outpatient appointment in the past 3 months, n (%)	
No missed appointment	923 (84.3)
Missed appointment at least once	172 (15.7)
Frequency of missed appointments in the past 3 months, n (%)	
Fewer than 3 times	1074 (98.1)
Three times or more	21 (1.9)

Health Care Utilization

In the preceding 3 months, 60.1% (n=658) of the participants had 1 - 3 outpatient visits, 22.4% (n=245) reported 4 or more visits, and 17.5% (n=192) had not accessed outpatient services. Additionally, 15.7% (n=172) reported at least 1 MA, while 84.3% (n=923) had no such record.

Of note, 17.5% (n=192) of the participants reported no outpatient visits in the past 3 months, yet 15.7% (n=172) indicated at least 1 MA. This discrepancy may reflect recall bias or differences in participants' interpretation of MAs, such as counting canceled or forgotten appointments that were never formally registered as visits.

Distribution of SMS Reminder Preferences

As shown in Table 3, nearly half (n=520, 47.5%) of the respondents preferred reminders referencing "Prior MAs" behavior, followed by general reminders (n=341, 31.1%). Fewer participants selected messages with an empathetic tone, including provider-focused empathy (n=48, 4.4%) and patient-focused empathy (n=60, 5.5%). Relational reminders, such as physician follow-up (n=53, 4.8%) and supportive reminders (n=73, 6.7%), accounted for a combined total of 11.5% of the preferences. These results indicate a predominant preference for practical and behavior-linked reminder formats over affective or relational messaging.

Because 4 SMS categories (provider-focused empathy, patient-focused empathy, physician follow-up, and supportive reminders) were selected by relatively few participants, the corresponding subgroup cell sizes in the chi-square analyses were small. Therefore, statistical comparisons involving these categories should be interpreted with caution.

Male respondents showed relatively higher proportions selecting supportive or relational message types, although these differences were not statistically significant after FDR correction.

Table . Distribution of preferred SMS reminder types (N=1095).

SMS reminder type	Values, n (%)
Prior missed appointment reminder	520 (47.5)
General reminder	341 (31.1)
Supportive reminder	73 (6.7)
Patient-focused empathy	60 (5.5)
Physician follow-up	53 (4.8)
Provider-focused empathy	48 (4.4)
Total	1095 (100)

XSL•FO
RenderX

Associations Between Demographic, Socioeconomic, and Health Care Utilization Characteristics and Preferred SMS Reminder Type

Analyses of demographic and health care utilization characteristics showed several initial associations with SMS reminder type preferences. As summarized in the cross-tabulations, gender, age group, education level, occupation, residential region, and living arrangement each demonstrated significant differences across message categories in unadjusted chi-square and ANOVA analyses (Table 4). For example, women and younger adults more frequently selected general or prior missed appointment reminders, whereas older adults showed comparatively higher proportions selecting empathetic or supportive formats. Occupational variation was also observed, with clerical and professional workers tending toward standard reminders, while senior executives reported a greater share of emotionally supportive preferences. These subgroup differences, however, did not persist following correction for multiple comparisons.

Health care utilization variables exhibited similar patterns. Participants with 1 to 3 outpatient visits in the past 3 months tended to prefer neutral, task-oriented reminders, while those with at least 1 recent MA were more likely to choose messages explicitly referencing prior absences (Table 5).

Table . Associations between demographic characteristics and preferred SMS reminder type.

Variable	General reminder, n (%)	Prior missed appointment behavior, n (%)	Provider-focused empathy, n (%)	Patient-focused empathy, n (%)	Physician follow-up, n (%)	Supportive reminder, n (%)	Chi-square or *F* test (*df*)	*P* value
Sex							26.554[a] (5)	.001
Male	67 (19.6)	102 (19.6)	12 (25.0)	21 (35.0)	11 (20.8)	31 (42.5)		
Female	274 (80.4)	418 (80.4)	36 (75.0)	39 (65.0)	42 (79.2)	42 (57.5)		
Age (y)								
Mean age (SD)	45.82 (12.43)	40.05 (12.33)	43.13 (12.41)	45.62 (11.93)	50.45 (11.19)	45.53 (13.39)	19.54[b]	.001
<50, n (%)	197 (57.8)	395 (76.0)	30 (62.5)	32 (53.3)	24 (45.3)	42 (57.5)		
≥50, n (%)	144 (42.2)	125 (24.0)	18 (37.5)	28 (46.7)	29 (54.7)	31 (42.5)		
Education								
Senior high or less	28 (8.2)	46 (8.8)	3 (6.3)	8 (13.3)	8 (15.1)	8 (11.0)	5.977[b] (10)	.82
College or university	225 (66.0)	356 (68.5)	31 (64.6)	35 (58.3)	35 (66.0)	44 (60.3)		
Postgraduate	88 (25.8)	118 (22.7)	14 (29.2)	17 (28.3)	14 (26.4)	21 (28.8)		

[a]Chi-square test.

[b]*F* test.

XSL•FO
RenderX

Table . Associations between health care utilization characteristics and preferred SMS reminder type.

Variable	General reminder, n (%)	Prior MA[a] behavior, n (%)	Provider-focused empathy, n (%)	Patient-focused empathy, n (%)	Physician follow-up, n (%)	Supportive reminder, n (%)	Chi-square (*df*)	*P* value
Outpatient visits (past 3 months)							5.112 (10)	.88
0 visits	56 (5.1)	89 (8.1)	10 (9.4)	12 (11.0)	11 (10.4)	14 (13.3)		
1 - 3 visits	201 (58.9)	315 (60.6)	31 (64.6)	37 (61.7)	33 (62.3)	41 (56.2)		
≥4 visits	84 (24.6)	116 (22.3)	7 (14.6)	11 (18.3)	9 (17.0)	18 (24.7)		
MAs (past 3 months)							3.267 (5)	.66
None	286 (83.9)	446 (85.8)	41 (85.4)	50 (83.3)	44 (83.0)	57 (78.1)		
≥1 MA	56 (16.4)	74 (14.2)	7 (14.6)	10 (16.7)	9 (17.0)	16 (21.9)		
Frequency of MAs							10.464 (5)	.06
<3 MAs	331 (97.1)	490 (94.2)	47 (97.9)	59 (98.3)	52 (98.1)	69 (94.5)		
≥3 MAs	10 (2.9)	4 (0.8)	1 (2.1)	1 (1.7)	1 (1.9)	4 (5.5)		

[a]MA: missed appointment.

To address the potential inflation of type 1 error due to multiple comparisons, the Benjamini-Hochberg FDR correction was applied across all demographic and utilization variables. After adjustment, only age and gender remained statistically significant predictors of SMS preference. All other variables, including education, occupation, region, living arrangement, outpatient visit frequency, and missed appointment history, no longer met the adjusted significance threshold. Multinomial logistic regression yielded consistent results, confirming age and sex as the only independent predictors after controlling for all covariates.

Although initial chi-square and ANOVA results suggested that several demographic and health care utilization variables were associated with SMS reminder preferences, these patterns did not remain statistically stable after controlling for multiple comparisons. After applying the Benjamini-Hochberg FDR correction, only age and sex remained significant predictors, whereas education level, occupation, residential region, living arrangement, outpatient visit frequency, and missed appointment history were no longer statistically significant (Table 6).

Table . Summary of false discovery rate (FDR)–corrected significance across all demographic and health care utilization variables.

Variable	Chi-square and *F* test (*df*)	Original *P* value	FDR-adjusted *P* value	Significant after FDR?
Sex	26.554 (df)[a]	.001	.006	Yes
Age (y; continuous)	19.54 (df)[b]	.001	.003	Yes
Education	5.977 (df) [a]	.82	.98	No
Outpatient visits (past 3 months)	5.112 (df)[a]	.88	.88	No
Missed appointments (past 3 months)	3.267 (df)[a]	.66	.99	No
Frequency of missed appointments	10.464 (df)[a]	.06	.13	No

[a]Chi-square.

[b]*F* test.

To identify independent predictors of SMS reminder preference while simultaneously adjusting for all covariates, a multinomial logistic regression model was constructed. The full results of this multivariable analysis are presented in Table 7, including adjusted odds ratios (aORs), 95% CI, and *P* values.

As shown in Table 7, the multinomial logistic regression analysis identified age as the only statistically significant independent predictor of SMS reminder preference after simultaneous adjustment for all covariates. Participants younger than 50 years were significantly more likely to prefer alternative reminder message types compared with the general reminder SMS (aOR 1.64, 95% CI 1.18 - 2.28; *P*=.003). In contrast, sex did not remain independently associated with reminder preference after multivariable adjustment (aOR 1.23, 95% CI 0.87 - 1.73; *P*=.24). Other sociodemographic characteristics and health care utilization variables, including education level, employment status, residential region, living arrangement, outpatient visit frequency, and recent MAs history, were not statistically significant in the fully adjusted model.

XSL•FO
RenderX

Table . Multinomial logistic regression analysis of factors associated with SMS reminder preference[a].

Predictor and category	aOR[b]	95% CI	*P* value
Sex			
Male (vs female)	1.23	0.87-1.73	.24
Age group			
<50 years (vs ≥50 years)	1.64	1.18-2.28	.003
Education level			
Less than or equal to high school (vs graduate+)	1.58	0.90-2.78	.11
College (vs graduate+)	1.12	0.81-1.55	.50
Employment status			
Retired (vs unemployed)	1.33	0.64-2.78	.44
Employed (vs unemployed)	1.65	0.90-3.01	.11
Residential region			
Kaoping region (vs non-Kaoping)	0.79	0.60-1.03	.08
Living arrangement			
Unmarried (vs with others)	1.31	0.91-1.89	.15
Living alone (vs with others)	1.21	0.72-2.06	.47
Outpatient visits (past 3 months)			
0 visits (vs ≥4)	1.19	0.78-1.81	.42
1-3 visits (vs ≥4)	1.15	0.84-1.58	.40
Missed appointments (past 3 months)			
≤3 times (vs ≥3)	2.08	0.86-5.04	.10

[a]Outcome variable: preference for SMS reminder type. Reference category: general reminder SMS. Model: multinomial logistic regression (all covariates entered simultaneously).

[b]aOR: adjusted odds ratio.

Discussion

Principal Findings

The primary objective of this study was to examine public preferences regarding SMS reminder content in preparation for a forthcoming large-scale randomized controlled trial. In the planned randomized controlled trial, participants will be randomly allocated to either an intervention group, which will receive a standardized SMS reminder seven days prior to their scheduled appointment containing details such as the appointment date, time, department, and queue number, or a control group that receives no reminder. Insights from this study will inform the development of effective SMS content strategies to enhance outpatient appointment adherence. The following sections address several key issues that merit further discussion.

Sample Representativeness

The observed preference for concise, behaviorally framed SMS reminders, particularly the general reminder and prior MA formats, was most prominent among women, younger adults, and highly educated respondents. These patterns align with the demographic profile of the study sample, in which the majority were female (n=851, 77.7%), under age 50 (n=720, 65.8%), and held a college degree or above (91.1%).

Prior literature indicates that women are more likely than men to assume caregiving responsibilities within households, maintain greater engagement with preventive health services, and demonstrate higher willingness to participate in self-administered online questionnaires related to health behaviors and health care use [25-27]. These factors may contribute to their increased representation in digital health surveys.

Furthermore, women consistently report higher health literacy and mobile engagement for health-related purposes, including appointment scheduling, medication reminders, and symptom monitoring, compared with men of the same age groups [28]. Studies have also shown that women exhibit stronger preferences for receiving health information via SMS or other asynchronous communication channels, which may further elevate their likelihood of responding to an online survey focused on SMS reminder content. In contrast, men tend to participate at lower rates in health communication surveys unless participation is tied to an active medical condition or clinical follow-up requirement [29].

Collectively, the predominance of women in this sample likely reflects well-documented patterns in digital health engagement rather than a study-specific recruitment artifact. Nevertheless,

XSL•FO
RenderX

the skewed sex distribution requires the cautious interpretation of aggregated preference patterns. To address this concern, sex was explicitly incorporated as a key covariate in the multinomial logistic regression analysis. After multivariable adjustment, sex did not remain independently associated with SMS reminder preference, indicating that the observed preference patterns were not solely driven by sample composition.

Potential Data Inconsistency in Appointment and Missed Visit Reporting

A minor discrepancy was observed between outpatient visit reports and MA data. Specifically, 17.5% (n=192) of the respondents reported no outpatient visits in the past 3 months, but 15.7% (n=172) reported at least 1 MA. This inconsistency may reflect recall bias or variable interpretations of MAs, such as canceled or forgotten visits not formally captured by hospital systems. Our findings echo those of de Reuver and Bouwman [30], who observed that type 1 and type 2 reporting errors are common in self-reported utilization data.

Preference Heterogeneity

The chi-square analyses indicated that participants most frequently selected general reminders and prior MA messages, while empathy-oriented or relationally framed messages were selected far less often. This suggests that patients prioritize clarity and actionable information over affective framing when receiving appointment-related communication. These findings are consistent with Hallsworth et al [16], who demonstrated that reminders containing cost-related information and clear behavioral cues were more effective than emotionally framed messages.

Preferences Differ by Sociodemographic and Health Care Utilization Characteristics

The results demonstrated meaningful variation in SMS reminder preferences by age and sex, whereas other demographic and behavioral characteristics did not show stable associations after correction for multiple testing.

After correction for multiple testing, age and sex remained statistically significant in FDR-adjusted univariate analyses. However, when all covariates were entered simultaneously into the multinomial logistic regression model, age emerged as the only independent predictor of SMS reminder preference. Sex did not retain statistical significance after multivariable adjustment.

After multivariable adjustment, age emerged as the only independent predictor of SMS reminder preference, whereas sex did not retain statistical significance. Other variables, including education, occupation, residential region, living arrangement, outpatient visit frequency, and recent MAs, lost statistical significance following the FDR adjustment.

Descriptively, men demonstrated a more even distribution of preferences across message types; however, these differences did not remain statistically significant after multivariable adjustment and should be interpreted cautiously.

These findings indicate that designing personalized SMS strategies based primarily on age and sex may be more appropriate than tailoring messages by education level or utilization history. This pattern is consistent with Crutchfield et al [31] and Mohammed Selim et al [32], who observed that sociodemographic characteristics influence responsiveness to reminder messages. An Israeli randomized controlled trial further found that older adults and those with prior MAs were more responsive to emotionally framed messages [33]. Junod Perron et al [34], however, emphasized that communication mode (SMS vs telephone) may influence engagement differently than message framing.

Because multiple chi-square tests were conducted, the Benjamini-Hochberg FDR procedure was applied to reduce the risk of type 1 error. After correction, only age and sex remained statistically significant predictors of SMS preference. Residential region, living arrangement, and all health care utilization variables, including outpatient visit frequency and recent MAs, were not significant after correction.

These results were corroborated by the multinomial logistic regression model, in which age and sex remained the only independent predictors after mutual adjustment. Overall, these findings indicate that although unadjusted analyses suggested broader subgroup differences, only age- and sex-based differences demonstrated sufficient stability to warrant consideration in tailored SMS design.

It should be noted that several message types, particularly empathy-based formats, had limited subgroup counts. The observed associations for these categories may therefore lack statistical robustness and should be interpreted conservatively. The revised text avoids overstating preference differences in low-frequency groups.

Content Validity Assessment

To ensure the appropriateness and clarity of SMS content prior to trial implementation, this study used item-level content validity indices based on 9 communication design principles derived from existing literature [17,19,23]. A multidisciplinary expert panel of 12 evaluators assessed the messages, exceeding recommended minimum panel sizes for content validation.

Although the panel did not include formal patient representatives, all members had extensive experience interacting with patients in outpatient or hospital settings. These perspectives provided practical insights grounded in real-world communication processes. This limitation is acknowledged, yet expert-based content validation remains standard practice in early formative research.

Rationale for Age Cutoff at 50 Years

The decision to dichotomize age at 50 years was justified by both statistical and contextual considerations. Statistically, this threshold yielded significant differences in SMS preferences. Contextually, adults around age 50 years in Taiwan frequently assume dual caregiving responsibilities for both aging parents and children, which may influence their engagement with health care communication [35].

XSL•FO
RenderX

Limitations

This study relied on convenience sampling through an online survey, which may limit representativeness. The predominance of female respondents reflects a well-documented pattern in online and digital health surveys, in which women are more likely than men to participate, particularly in studies related to health behaviors and health care communication [36,37]. To address this imbalance, sex was explicitly incorporated as a covariate in both univariate analyses and the multinomial logistic regression model. After multivariable adjustment, sex did not remain independently associated with SMS reminder preference, suggesting that the observed patterns were not solely driven by sample composition. Nevertheless, absolute preference proportions should be interpreted with caution, particularly for male patients. Future studies may enhance representativeness through targeted recruitment strategies and the inclusion of more inclusive gender response options, in line with emerging best practices in survey methodology [38].

Conclusion

This study demonstrates that preferences for SMS appointment reminder content vary systematically among outpatients, with age emerging as the most robust independent predictor. General reminders and prior missed appointment messages were most consistently preferred, whereas empathy-based or relational formats were selected far less frequently. After applying FDA correction and conducting multinomial logistic regression, age remained the only stable independent predictor of SMS reminder preference, whereas sex did not retain statistical significance after multivariable adjustment. Beyond predictor effects, the observed preference for concise, action-oriented messages underscores the potential importance of clarity and behavioral specificity when designing SMS reminders in Taiwan's universal health coverage context. Content validity assessment further confirmed that the 6 SMS prototypes possess adequate semantic clarity and cultural appropriateness for use in subsequent interventional research.

Although convenience sampling limits generalizability, the results offer valuable formative evidence to guide the development of SMS interventions aimed at improving appointment adherence. Future randomized controlled trials should evaluate the real-world effectiveness of these demographically tailored and behaviorally informed reminder formats in reducing outpatient no-show rates and supporting more efficient use of ambulatory care resources.

Acknowledgments

The authors would like to thank the Kaohsiung Medical University Hospital and Kaohsiung Medical University for their generous support.

Funding

This study was supported by a research grant from Kaohsiung Municipal Ta-Tung Hospital, Taiwan (Grant No. KMTTH-112-036).

Authors' Contributions

Conceptualization: P-SC, S-HH
Data curation: H-TL, S-HH, S-LM, Y-CC
Formal analysis: S-HH, Y-CC
Funding acquisition: P-SC
Investigation: H-TL, S-HH, S-LM, Y-CC
Methodology: P-SC, S-HH
Project administration: P-SC, S-HH
Resources: P-SC
Software: S-HH, Y-CC
Supervision: P-SC
Validation: H-TL, S-HH, S-LM, Y-CC
Visualization: P-SC, S-HH, Y-CC
Writing–original draft: S-HH, Y-CC
Writing–review and editing: I-CL, P-SC, S-HH

Conflicts of Interest

The authors declare no conflicts of interest.

Multimedia Appendix 1
The original 6 prototypes of SMS reminders (Chinese and English).
[DOCX File, 20 KB - jmir_v28i1e81324_app1.docx]

References

XSL•FO
RenderX

1. Li DG, Pournamdari AB, Liu KJ, Laskowski K, Joyce C, Mostaghimi A. Evaluation of point-of-care decision support for adult acne treatment by primary care clinicians. JAMA Dermatol 2020 May 1;156(5):538-544. [doi: 10.1001/jamadermatol.2020.0135] [Medline: 32129799]
2. Creadore A, Desai S, Li SJ, et al. Insurance acceptance, appointment wait time, and dermatologist access across practice types in the US. JAMA Dermatol 2021 Feb 1;157(2):181-188. [doi: 10.1001/jamadermatol.2020.5173] [Medline: 33439219]
3. Ayanian JZ, Sequist TD, Zaslavsky AM, Johannes RS. Physician reminders to promote surveillance colonoscopy for colorectal adenomas: a randomized controlled trial. J Gen Intern Med 2008 Jun;23(6):762-767. [doi: 10.1007/s11606-008-0576-2] [Medline: 18386103]
4. Lam TYT, Hui AJ, Sia F, et al. Short Message Service reminders reduce outpatient colonoscopy nonattendance rate: a randomized controlled study. J Gastroenterol Hepatol 2021 Apr;36(4):1044-1050. [doi: 10.1111/jgh.15218] [Medline: 32803820]
5. Deng X, Wang Y, Zhu T, Zhang W, Yin Y, Ye L. Short message service (SMS) can enhance compliance and reduce cancellations in a sedation gastrointestinal endoscopy center: a prospective randomized controlled trial. J Med Syst 2015 Jan;39(1):169. [doi: 10.1007/s10916-014-0169-z] [Medline: 25476268]
6. Kooij FO, Vos N, Siebenga P, Klok T, Hollmann MW, Kal JE. Automated reminders decrease postoperative nausea and vomiting incidence in a general surgical population. Br J Anaesth 2012 Jun;108(6):961-965. [doi: 10.1093/bja/aes024] [Medline: 22382396]
7. Chen LC, Schafheutle EI, Noyce PR. The impact of nonreferral outpatient co-payment on medical care utilization and expenditures in Taiwan. Res Social Adm Pharm 2009 Sep;5(3):211-224. [doi: 10.1016/j.sapharm.2008.08.003] [Medline: 19733822]
8. Sallis A, Sherlock J, Bonus A, et al. Pre-notification and reminder SMS text messages with behaviourally informed invitation letters to improve uptake of NHS Health Checks: a factorial randomised controlled trial. BMC Public Health 2019 Aug 22;19(1):1162. [doi: 10.1186/s12889-019-7476-8] [Medline: 31438908]
9. Lacy NL, Paulman A, Reuter MD, Lovejoy B. Why we don't come: patient perceptions on no-shows. Ann Fam Med 2004;2(6):541-545. [doi: 10.1370/afm.123] [Medline: 15576538]
10. Hwang AS, Atlas SJ, Cronin P, et al. Appointment "no-shows" are an independent predictor of subsequent quality of care and resource utilization outcomes. J Gen Intern Med 2015 Oct;30(10):1426-1433. [doi: 10.1007/s11606-015-3252-3] [Medline: 25776581]
11. Husain-Gambles M, Neal RD, Dempsey O, Lawlor DA, Hodgson J. Missed appointments in primary care: questionnaire and focus group study of health professionals. Br J Gen Pract 2004 Feb;54(499):108-113. [Medline: 14965389]
12. Smith A. In: Haakonssen K, editor. Adam Smith: The Theory of Moral Sentiments: Boston University; 2012. [doi: 10.1017/CBO9780511800153]
13. Berliner Senderey A, Kornitzer T, Lawrence G, et al. It's how you say it: systematic A/B testing of digital messaging cut hospital no-show rates. PLoS One 2020;15(6):e0234817. [doi: 10.1371/journal.pone.0234817] [Medline: 32574181]
14. Hu M, Xu X, Li X, Che T. Managing patients' no-show behaviour to improve the sustainability of hospital appointment systems: exploring the conscious and unconscious determinants of no-show behaviour. J Clean Prod 2020 Oct;269:122318. [doi: 10.1016/j.jclepro.2020.122318]
15. Smith A. The Theory of Moral Sentiments: Penguin; 2010.
16. Hallsworth M, Berry D, Sanders M, et al. Stating appointment costs in SMS reminders reduces missed hospital appointments: findings from two randomised controlled trials. PLoS One 2015;10(9):e0137306. [doi: 10.1371/journal.pone.0137306] [Medline: 26366885]
17. Haynes SN, Richard DCS, Kubany ES. Content validity in psychological assessment: a functional approach to concepts and methods. Psychol Assess 1995;7(3):238-247. [doi: 10.1037/1040-3590.7.3.238]
18. Tsai WC, Lee WC, Chiang SC, Chen YC, Chen TJ. Factors of missed appointments at an academic medical center in Taiwan. J Chin Med Assoc 2019 May;82(5):436-442. [doi: 10.1097/JCMA.0000000000000068] [Medline: 30907780]
19. Acharya A, Judah G, Ashrafian H, et al. Investigating the national implementation of SMS and mobile messaging in population screening (the SIPS study). EBioMedicine 2023 Jul;93:104685. [doi: 10.1016/j.ebiom.2023.104685] [Medline: 37384997]
20. Neal RD, Hussain-Gambles M, Allgar VL, Lawlor DA, Dempsey O. Reasons for and consequences of missed appointments in general practice in the UK: questionnaire survey and prospective review of medical records. BMC Fam Pract 2005 Nov 7;6:47. [doi: 10.1186/1471-2296-6-47] [Medline: 16274481]
21. Barrera Ferro D, Brailsford S, Bravo C, Smith H. Improving healthcare access management by predicting patient no-show behaviour. Decis Support Syst 2020 Nov;138:113398. [doi: 10.1016/j.dss.2020.113398]
22. Watson J, Woodward T. Jean Watson's theory of human caring. In: Nursing Theories and Nursing Practice, 3rd edition: Davis Company; 2010:351-369 URL: https://www.fadavis.com/product/nursing-research-theory-practice-parker-smith-3 [accessed 2026-02-03]
23. Lynn MR. Determination and quantification of content validity. Nurs Res 1986;35(6):382-385. [Medline: 3640358]
24. Sarkar SK, Tang CY. Adjusting the Benjamini–Hochberg method for controlling the false discovery rate in knockoff-assisted variable selection. Biometrika 2022 Nov 29;109(4):1149-1155. [doi: 10.1093/biomet/asab066]

XSL•FO
RenderX

25. DiMatteo MR. Social support and patient adherence to medical treatment: a meta-analysis. Health Psychol 2004 Mar;23(2):207-218. [doi: 10.1037/0278-6133.23.2.207] [Medline: 15008666]
26. Dobransky K, Hargittai E. The disability divide in internet access and use. Inf Commun Soc 2006 Jun;9(3):313-334. [doi: 10.1080/13691180600751298]
27. Xie B. Effects of an eHealth literacy intervention for older adults. J Med Internet Res 2011 Nov 3;13(4):e90. [doi: 10.2196/jmir.1880] [Medline: 22052161]
28. Bidmon S, Terlutter R. Gender differences in searching for health information on the internet and the virtual patient-physician relationship in Germany: exploratory results on how men and women differ and why. J Med Internet Res 2015 Jun 22;17(6):e156. [doi: 10.2196/jmir.4127] [Medline: 26099325]
29. Terrero Ledesma NE, Nájera A, Reolid-Martínez RE, Escobar-Rabadán F. Internet health information seeking by primary care patients. Rural Remote Health 2022 Oct;22(4):6585. [doi: 10.22605/RRH6585] [Medline: 36192837]
30. de Reuver M, Bouwman H. Dealing with self-report bias in mobile Internet acceptance and usage studies. Inf Manage 2015 Apr;52(3):287-294. [doi: 10.1016/j.im.2014.12.002]
31. Crutchfield TM, Kistler CE. Getting patients in the door: medical appointment reminder preferences. Patient Prefer Adherence 2017;11:141-150. [doi: 10.2147/PPA.S117396] [Medline: 28182131]
32. Mohammed Selim S, Senanayake S, McPhail SM, Carter HE, Naicker S, Kularatna S. Consumer preferences for a healthcare appointment reminder in Australia: a discrete choice experiment. Patient 2024 Sep;17(5):537-550. [doi: 10.1007/s40271-024-00692-9] [Medline: 38605246]
33. Teo AR, Niederhausen M, Handley R, et al. Using nudges to reduce missed appointments in primary care and mental health: a pragmatic trial. J Gen Intern Med 2023 Jul;38(Suppl 3):894-904. [doi: 10.1007/s11606-023-08131-5] [Medline: 37340264]
34. Junod Perron N, Dao MD, Righini NC, et al. Text-messaging versus telephone reminders to reduce missed appointments in an academic primary care clinic: a randomized controlled trial. BMC Health Serv Res 2013 Apr 4;13(1):125. [doi: 10.1186/1472-6963-13-125] [Medline: 23557331]
35. Do EK, Cohen SA, Brown MJ. Socioeconomic and demographic factors modify the association between informal caregiving and health in the Sandwich Generation. BMC Public Health 2014 Apr 15;14(1):362. [doi: 10.1186/1471-2458-14-362] [Medline: 24735495]
36. Becker R. Gender and survey participation: an event history analysis of the gender effects of survey participation in a probability-based multi-wave panel study with a sequential mixed-mode design. Methods Data Anal 2022;16(1):3-32. [doi: 10.12758/mda.2021.08]
37. Weber AM, Gupta R, Abdalla S, Cislaghi B, Meausoone V, Darmstadt GL. Gender-related data missingness, imbalance and bias in global health surveys. BMJ Glob Health 2021 Nov;6(11):e007405. [doi: 10.1136/bmjgh-2021-007405] [Medline: 34836912]
38. Albaugh QM, Harell A, Loewen PJ, Rubenson D, Stephenson LB. Measuring transgender and nonbinary identities in online surveys: evidence from two national election studies. PS: Polit Sci Polit 2025 Jul;58(3):406-412. [doi: 10.1017/S1049096525000137]

Abbreviations

aOR: adjusted odds ratio
FDR: false discovery rate
MA: missed appointment

Edited by A Schwartz, M Balcarras; submitted 26.Jul.2025; peer-reviewed by K Le, YW Chiu; accepted 18.Dec.2025; published 07.Apr.2026.

Please cite as:
Hsiao SH, Chuang YC, Ma SL, Lin HT, Lee IC, Chen PS
Preferences for Personalized Text Message Appointment Reminders Among Outpatients in a Universal Health System: Cross-Sectional Study
J Med Internet Res 2026;28:e81324
URL: https://www.jmir.org/2026/1/e81324
doi:10.2196/81324

© Shih-Huai Hsiao, Yu-Chen Chuang, Shu-Ling Ma, Hui-Tzu Lin, I-Chen Lee, Pei-Shih Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet

XSL•FO
RenderX

Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Gamified Mobile Health Intervention to Promote Physical Activity, Executive Function, and Mental Health in College Students: Randomized Controlled Trial

Ming Yang[1,2], PhD; Yingjun Guo[3], MEd; Zhen Li[4], PhD; Lei Jiang[1], PhD; Yanan Gao[4], PhD; Sen Yang[1], PhD; Zhixiong Zhou[1], PhD

[1]Institute of Artificial Intelligence in Sports, Capital University of Physical Education and Sports, No.11, North Third Ring West Road, Haidian District, Beijing, China

[2]School of Physical Education, Yantai University, Yantai, Shandong, China

[3]Beijing Sport University Press, Beijing Sport University, Beijing, Beijing, China

[4]School of Physical Education and Sport Science, Fujian Normal University, Fuzhou, Fujian, China

Corresponding Author:
Zhixiong Zhou, PhD
Institute of Artificial Intelligence in Sports, Capital University of Physical Education and Sports, No.11, North Third Ring West Road, Haidian District, Beijing, China

Abstract

Background: College students commonly experience suboptimal health conditions, including insufficient physical activity (PA), excessive body weight, and declining physical fitness. Traditional interventions face low adherence, while gamified mobile health (mHealth) programs may improve engagement and outcomes.

Objective: This study aimed to evaluate the feasibility and effectiveness of a novel gamified, incentive-based mHealth intervention on primary outcomes (PA and adherence) and secondary outcomes (physical fitness, body composition, executive function [EF], and mental health).

Methods: A 2-arm parallel-group randomized controlled trial (RCT) was conducted in 2025 at Yantai University with 160 college students (18 - 25 years; BMI 18.5 - 30.0) who were randomized 1:1 (computer-generated, sex-stratified blocks of 4; concealed allocation) to the intervention group (IG) or control group (CG; n=80 each); major exclusions were contraindications to exercise, severe physical/mental illness, recent PA interventions, or psychotropic medication use. Both used the same fitness watch–app system and identical PA targets (≥150 min moderate-to-vigorous physical activity [MVPA] per week or ≥900 metabolic equivalent-minutes [MET-min] per week); IG additionally received team-based gamification (competition, points/leaderboards, feedback, and rewards), while CG received monitoring only. PA and adherence were monitored throughout the 8-week intervention; other outcomes were assessed at baseline and 8 weeks (fitness, body composition, EF, and mental health). Open-label with blinded outcome assessors/analysts; intention-to-treat (ITT) with multiple imputation.

Results: At 8 weeks, data were available for 154 participants (IG 78; CG 76); all 160 were analyzed per ITT. Compared to the CG, the IG demonstrated significantly higher mean levels in all primary PA outcomes over 8 weeks (daily steps: mean 10,356, SD 1245 versus 8242, SD 1087; Δ=2114; d=1.81, 95% CI 1.44 - 2.18; P<.001; daily MVPA: mean 71, SD 15 versus 43, SD 12 min; Δ=28 min; d=2.06, 95% CI 1.68 - 2.45; P<.001; and weekly MET-min: mean 1650, SD 310 versus 1340, SD 285; Δ=310; d=1.04, 95% CI 0.71 - 1.37; P<.001). Adherence was also higher in the IG (both P<.001). Key secondary outcomes, including skeletal muscle mass (Δ +0.54 kg), body fat percentage (Δ −0.46 percentage points), 800/1000 m run time (Δ −6.5 s), and depressive symptoms (Δ −2.0 points), significantly favored the IG (all P<.001). No serious adverse events occurred.

Conclusions: This RCT demonstrated that integrating gamification into an mHealth platform significantly enhanced PA, intervention adherence, and selected health outcomes among college students. The key innovation lies in the trial design, which held the technology platform and activity goals constant across groups, thereby allowing a clearer estimate of the incremental contribution of gamification. Unlike many existing studies, this approach reduces confounding from co-occurring intervention components in between-group comparisons. It therefore provides more comparable and methodologically robust evidence on the specific efficacy of gamification in mHealth interventions. For real-world implementation, this model leverages the existing smartphone and wearable ecosystem, offering a low-cost, scalable strategy for university health-promotion programs.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2500108950; https://www.chictr.org.cn/showproj.html?proj=285460

(J Med Internet Res 2026;28:e82769) doi:10.2196/82769

XSL•FO
RenderX

KEYWORDS

mHealth; eHealth; gamification; physical fitness; psychological well-being; cognitive function; adherence

Introduction

Physical inactivity has become a prevalent issue among today's college student population, posing a significant threat to both their physical and mental well-being as well as their overall quality of life. Recent national estimates indicate that the combined prevalence of overweight and obesity among Chinese adults aged ≥18 years exceeds 50% [1], with college students being particularly affected. This cluster of risks, including insufficient physical activity (PA), imbalanced diet, excess weight, poor sleep, and psychological distress, not only undermines individual development but also poses challenges to national public health policies, such as the "Healthy China 2030" and "Weight Management Year" campaigns [2,3].

Traditional exercise programs often underperform because they are monotonous, provide delayed feedback, and lack sustained motivational scaffolding, which collectively erode adherence and limit scalability [4]. By contrast, mobile health (mHealth) leverages smartphones, wearables, and connected data to enable real-time, personalized monitoring and feedback on activity, heart rate, sleep, and related indicators [5]. Recent reviews and empirical studies have confirmed that mHealth interventions can produce small-to-moderate improvements in PA and related health behaviors [6]. In parallel, gamification is the purposeful use of points, badges, leaderboards, challenges, and achievement systems in nongame contexts and has gained traction in health promotion for strengthening motivation and social engagement [7]. However, evidence for the impact of gamification is mixed: while some studies report significant benefits, others show minimal effects, highlighting the need for further research on the mechanisms and optimal design of gamified interventions [8].

Self-determination theory (SDT) provides a framework for designing more effective interventions. SDT posits that satisfying the psychological needs for autonomy, competence, and relatedness supports motivation internalization and sustained behavior change. In our design, autonomy was supported via self-selection and personalized goals, competence via real-time feedback and progress dashboards, and relatedness via team-based collaboration and competition [9]. Similarly, the theory of planned behavior (TPB) contends that behavior is driven by intention, which is shaped by attitude, subjective norms, and perceived behavioral control [10].

In our gamification context, these determinants were strengthened as follows: positive outcome expectancies aimed to improve attitudes; peer norms embedded in teams and leaderboards helped establish subjective norms; and clear weekly targets were designed to enhance participants' perceived behavioral control [11].

Recent studies and the World Health Organization (WHO) Global Action Plan on Physical Activity 2018‐2030 both emphasize that such theory-driven, cross-sectoral approaches are essential for promoting sustainable PA [12]. However, there remains a scarcity of research that systematically integrates mHealth with gamification under a combined SDT-TPB framework and rigorously evaluates its efficacy among college students. Furthermore, many existing studies rely on self-reported outcomes or focus on a narrow set of physiological measures, lacking comprehensive evaluation across physical, cognitive, and mental health domains within a single trial [13].

Integrating gamification with mHealth, therefore, offers a promising approach that combines real-time, data-driven monitoring with motivational and socially interactive elements to transform health behaviors into a more visualized, task-oriented, and engaging process. Within this framework, TPB helps explain the formation of behavioral motives and intentions, whereas SDT elucidates how short-term motivation is converted into long-term engagement [14]. This synergy is expected not only to promote PA but also to support mental well-being and, via goal setting and self-monitoring demands, to engage core domains of executive function [15]. College students, who have high smartphone penetration and digital literacy, are therefore an ideal population for such an approach [16].

Therefore, the aim of this study was to evaluate the feasibility and effectiveness of a novel, theory-based gamified mHealth intervention incorporating an incentive mechanism. This intervention integrates real-time monitoring, team-based competition, and personalized feedback and was comprehensively evaluated using a randomized controlled trial (RCT) design [17]. We prespecified two primary outcomes, (1) PA levels and (2) intervention adherence, to capture the core efficacy of the model. Secondary outcomes included physical fitness, body composition, mental health, and executive function (EF) to further assess the potential broader benefits for physical and psychological well-being. We hypothesized that, relative to control, participants receiving the gamified intervention would demonstrate higher PA levels and better adherence after the intervention and would show more favorable changes across secondary outcomes [18]. By addressing these objectives, this study aims to provide a scientific basis and practical strategies for college health promotion in the context of the "Healthy China" initiative and contributes to the innovation of health education systems in higher education institutions.

Methods

Trial Design

This study used a 2-arm, parallel-group (RCT) design to evaluate the feasibility and effectiveness of a gamified, competition-based intervention using mHealth technology (intervention group [IG]) versus a conventional intervention approach (control group [CG]) in promoting physical and mental health among college students aged 18‐25 years. The formal intervention period lasted for 8 weeks. Baseline assessments (T0) were conducted one week prior to the intervention, and postintervention assessments (T1) were carried out one week after the intervention concluded.

XSL•FO
RenderX

Prior to the formal intervention, all participants underwent a 1-week preintervention phase, during which the accuracy of the wearable smartwatches was validated (tolerance ≤3%), and participants were familiarized with the "Shouti Fitness" app (Tianbao Zhu) testing procedures and measurement protocols.

The study was designed and reported in accordance with the CONSORT (Consolidated Standards Of Reporting Trials) 2025 statement for randomized trials, the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth; Checklist 1) extension for web- and mobile-based health interventions, and the CONSORT extension for abstracts [19-21].

Trial Setting and Eligibility Criteria

Participants were recruited between February 04 and 10, 2025, at Yantai University, Shandong Province, through campus announcements, social media platforms, and health education lectures. A total of 224 volunteers were initially screened. Eligibility criteria were the following: full-time college students aged 18 - 25 years owning a smartphone; BMI between 18.5 and 30.0; and willingness to provide written informed consent. Exclusion criteria included the presence of severe physical or mental illnesses or contraindications to exercise; participation in other PA interventions within the past 3 months; and the use of psychotropic medications at the time of this writing. After screening, 160 eligible participants were enrolled.

Eligibility Criteria for Sites and Individuals Delivering the Interventions

No eligibility criteria were prespecified for individuals delivering the interventions. The interventions were primarily technology-mediated and self-administered by participants. The role of the research team was limited to initial instruction, technical support, and data monitoring, rather than direct delivery of the exercise intervention itself.

Intervention and Comparator

The intervention system comprised the following hardware and software components.

- Hardware: participants wore fitness watches equipped with high-precision accelerometers to monitor key metrics such as step count (Figure 1A-1B), duration of moderate-to-vigorous physical activity (MVPA), and metabolic equivalent (MET).
- Software: a customized mobile app, the "Shouti Fitness" app (compatible with Android and iOS; Figure 1C), was employed. The app features AI-driven exercise guidance, personalized exercise prescription delivery, real-time feedback, and synchronization of PA data, supporting both intervention implementation and adherence monitoring (Figure 1D).

XSL•FO
RenderX

Figure 1. Technical platform used in this study. (**A**) High-precision fitness watch. (**B**) Data transmission station. (**C**) "Shouti Fitness" App. (**D**) Health behavior monitoring system. (**E**) InBody H30 body composition analyzer. (**F**) NIH Toolbox cognitive function assessment tool on the iPad platform.

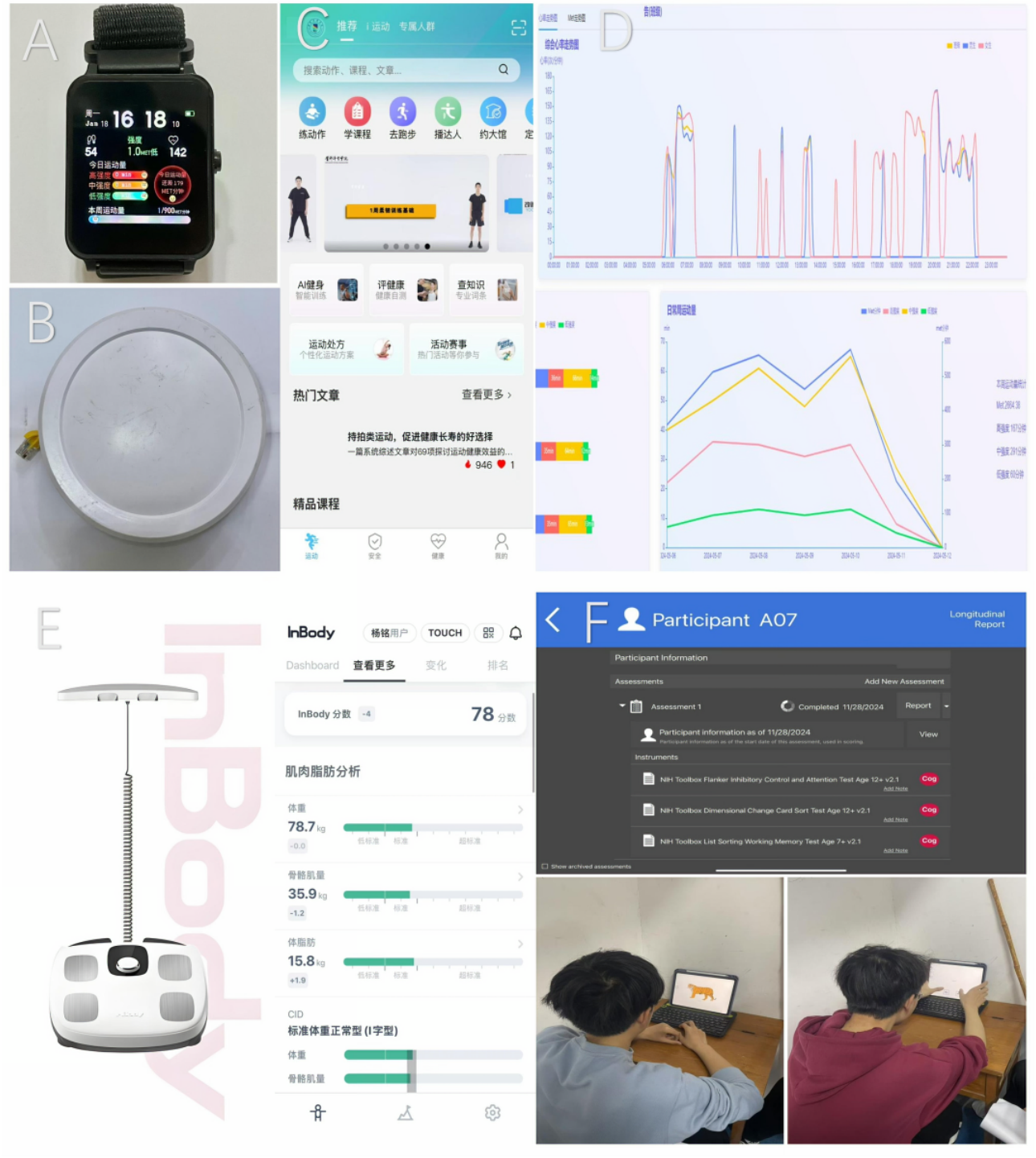

All participants were instructed to download and register on the "Shouti Fitness" app and were provided with a study-issued fitness watch. They then completed a 1-week familiarization phase (February 17 - 23, 2025) to learn the procedures and adapt to the device. The formal intervention ran for 8 weeks, from March 3 to April 27, 2025 (Figure 2).

XSL•FO
RenderX

Figure 2. Workflow diagram of the intervention mechanism used in this study. Overview of the 8-week randomized controlled trial (RCT) workflow in college students, including T0/T1 assessments, physical activity targets, and intervention group-only gamified components.

XSL•FO
RenderX

During the intervention, the fitness watches estimated exercise intensity and energy expenditure (metabolic equivalent-minutes [MET-min]) using a convolutional neural network-long short-term memory (CNN-LSTM) model and recorded durations of MVPA [22]. Participants were required to complete daily recommended PA targets (≥30 minutes of MVPA or ≥180 MET-min), with weekly goals set at ≥150 minutes of MVPA or ≥900 MET-min [23]. Data synchronization occurred twice daily at fixed data stations (07:30 AM and 11:30 PM), with activity data automatically uploaded to the exercise behavior monitoring system.

In IG (n=80) participants were assigned to 4 teams, 2 single-sex teams (1 all-male or 1 all-female), and 2 mixed-sex teams (10 males and 10 females each). Beyond meeting the baseline activity goals, a team-based competition was implemented; weekly rankings of teams based on average activity levels awarded points as follows. 1st place: 40 points, 2nd: 30 points, 3rd: 20 points, and 4th: 10 points. Additionally, the top 10% of individuals within each team (overall top 8 performers) contributed an extra 20 points to their team. Real-time activity progress bars (MVPA duration and MET-min) were displayed via the fitness watch and monitoring platform. Weekly midweek updates announced leading teams and individual champions every Wednesday, with complete team and individual leaderboards published each Sunday. Feedback and incentives included cumulative volunteer hours, honorary certificates, academic credit coordination, and material rewards of 400 Chinese yuan (Renminbi; approximately US$55.53 or €49.31, based on the exchange rate in April 2025) for individuals ranking in the top 10%.

CG (n=80) received an active, technology-based PA program without any gamification or social competition. Participants used the same technical platform, fitness watches, and app system as the IG and received identical daily and weekly personalized activity goal prompts. This design controlled for the nonspecific effects of platform use, goal setting, and researcher attention. The key distinction was the absence of all gamification and team-based competition elements, including team assignments; the point-based scoring/ranking system; real-time progress bars; midweek updates; and weekly leaderboards; as well as any performance-based feedback, incentives, or rewards.

Outcomes

The outcome measures were prespecified as primary and secondary. Primary outcomes directly assess the core efficacy of the mHealth-gamification intervention; secondary outcomes examine potential downstream health benefits.

Primary Outcomes

The primary outcomes were PA levels and intervention adherence. These indicators are selected to test the core hypothesis of this study, that "mHealth+gamification" intervention model would effectively increase participants' PA and adherence.

- PA levels: daily steps, daily minutes of MVPA, and weekly MET-min were derived from the self-developed wearable device and a companion sports application. Intensity classification followed contemporary MET cut-points (moderate 3.0 - 5.9 METs; vigorous ≥6.0 METs), and weekly behavioral targets were benchmarked to the 2020 WHO adult guidelines (≥150 - 300 min per week of moderate activity, or ≥75 - 150 min per week of vigorous activity, or an equivalent combination) [23].
- Adherence: calculated from the health behavior monitoring system, including daily MVPA≥30 min or MET-min≥180; weekly MVPA≥150 min or MET-min≥900, aligning with WHO guidelines [23]. The data synchronization rate was the percentage of weeks with successful device-data synchronization on ≥5 days. Group-level results are presented as the mean (SD) of these individual percentages.

Secondary Outcomes

The secondary outcomes evaluated the potential far-end health benefits of increased PA, including physical fitness, mental health, EF, and body composition. We prespecified mental health and EF as secondary outcomes because (1) increasing PA is associated with lower depressive and anxiety symptoms in adult and college populations, and exercise shows small-to-moderate benefits in randomized trials [24] and (2) team-based, goal-driven gamification requires planning, monitoring, inhibition, and cognitive flexibility, core executive domains plausibly sensitive to change during an 8-week intervention [25].

- Physical fitness: physical fitness was field-tested on the Changxiang 100 platform in accordance with the National Student Physical Fitness Standards (GB/T 18041 - 2022). The assessment comprised 7 tests to evaluate multiple facets of physical fitness: 1000 m run (men) / 800 m run (women), vital capacity, sit-and-reach test, 1-min sit-ups (women) / pull-ups (men), 50 m sprint, and standing long jump. This standardized national protocol ensures the assessments' widespread applicability and comparability.
- Mental health: depressive and anxiety symptoms were assessed online (Wenjuanxing) using the Zung Self-Rating Depression Scale (SDS) and the Zung Self-Rating Anxiety Scale (SAS). These scales are well-established tools with demonstrated good reliability and validity (eg, SAS Cronbach α=.84; SDS Cronbach α=.88). Recent studies in Chinese student samples and broader Chinese cohorts support acceptable internal consistency and the applicability of Chinese norms for SDS/SAS. The SDS comprises 20 items, each rated on a 4-point scale, with the total raw score converted to an index. Similarly, the SAS contains 20 items rated on a 4-point scale. For both scales, higher standardized scores indicate a greater severity of symptoms. We report raw and standardized scores; higher scores indicate greater symptom burden [26].
- EFs: EFs were measured on iPad using the National Institutes of Health (NIH) Toolbox Cognition Battery (Figure 1F). This computerized assessment battery was selected due to its strong psychometric properties, standardization, and ability to efficiently measure core components of EF. The battery specifically targeted 3 key domains: the dimensional change card sort test, which assesses cognitive flexibility by requiring individuals to shift between matching rules; the flanker inhibitory control

XSL•FO
RenderX

and attention test, which measures inhibitory control and attention by assessing the ability to suppress responses to distracting stimuli; and the List Sorting Working Memory Test, which evaluates working memory capacity through the temporary storage and manipulation of information. Scoring used age-adjusted T-scores per the latest NIH Toolbox V3 technical guidance (renormed to the 2020 US Census and validated against gold-standard measures). Recent meta-analyses and reviews indicate that persistent PA and movement-based programs can yield small-to-moderate improvements in EF in youth and young adults, especially in inhibitory control and cognitive flexibility; acute-exercise effects are mixed, underscoring the focus on multiweek training [27].

- Body composition: body fat percentage, skeletal muscle mass, BMI, and basal metabolic rate (BMR) were obtained with a multifrequency bioelectrical impedance analyzer (InBody H30 (InBody Co, Ltd; Figure 1E). We adhered to standard pretest controls (eg, no vigorous exercise or large meals prior to testing). This objective bioelectrical impedance analysis ensured accurate tracking of changes in body composition resulting from the PA intervention.

Sample Size

Based on findings from previous comparable studies [28-30], the sample size was determined using G*Power 3.1 (Heinrich Heine University Düsseldorf). A 2-tailed significance level of α=.05, statistical power (1-β)=.80, and an anticipated medium effect size (d=0.5) were specified. For an independent samples 2-tailed t test comparing 2 groups, using a 2-tailed test, the minimum required sample size was 64 participants per group (total N=128). Allowing for an estimated 20% attrition, the final target was set at 80 participants per group (total N=160).

Randomization

Random allocation sequences were generated using the PROC PLAN procedure in SAS 9.4 (SAS Institute Inc) by an independent researcher who was not involved in participant recruitment, outcome assessment, or data analysis. We used sex-stratified block randomization with a block size of 4 and a 1:1 allocation ratio to the IG (n=80) and CG (n=80) to ensure balance in sex distribution. Allocation concealment was maintained using sequentially numbered, opaque, sealed envelopes prepared off-site; after completion of baseline assessments, research staff opened the next envelope in sequence to assign participants to groups. Recruitment staff had no access to the allocation sequence; assignment was revealed when the next sealed envelope was opened after baseline assessments.

For the IG only, a postrandomization team structure was created to implement the gamified competition mechanics. The IG was subdivided into 4 equal-size teams (n=20 each). An independent coordinator used constrained random allocation based only on sex and headcount (no outcome data) to form 2 single-sex teams (Team A: 20 males; Team B: 20 females) and 2 mixed-sex teams (Teams C-D: 10 males and 10 females each). Each team functioned as an independent gamification unit for scoring and ranking. In contrast, CG participants were not assigned to any teams and did not participate in gamified competitions; they only completed the baseline PA tasks set on the same platform. This subdivision served two purposes: (1) to create the stable in-app subgroups required for points, leaderboards, and weekly challenges and (2) to promote competition fairness and mitigate structural influences arising from sex-related differences in PA and fitness [31,32].

Blinding

Given the visible nature of the intervention components, namely the gamified team competition and group structure, this study adopted an open-label design, and participant blinding was not feasible. Participants were informed that they would be randomly assigned either to the IG, which involved team-based gamified competition, or to the CG, which involved only meeting the basic PA targets. Outcome assessors were blinded. Physical-fitness testers, InBody H30 technicians, and NIH Toolbox/SDS/SAS administrators used deidentified testing schedules, were trained not to discuss allocation, and logged any accidental unmasking. Data analysts worked on a deidentified dataset with groups masked as A/B; the statistical analysis plan was finalized and code frozen before unmasking, which occurred only after database lock and completion of primary analyses.

Statistical Methods

Data analysis followed the intention-to-treat (ITT) principle. Continuous variables were expressed as mean (SD), and categorical variables as frequency and percentage (n%). Baseline differences were assessed using independent-samples t tests (or Mann-Whitney U tests, as appropriate) for continuous variables and χ^2 tests (or Fisher exact tests) for categorical variables.

The primary outcomes (PA and adherence) were summarized as mean values over the 8-week intervention period. Between-group differences in these postintervention means were then assessed using ANCOVA, with the corresponding baseline values included as covariates. For secondary outcomes (physical fitness, body composition, SDS/SAS, and EF), ANCOVA with baseline adjustment was likewise used, with mixed effects models as sensitivity analyses. Adherence-related indicators (eg, goal completion and data synchronization) were summarized descriptively and compared between groups; when applicable, ANCOVA with baseline adjustment was applied. Sex-based comparisons were not prespecified, and the trial was not powered for sex-related effect modification. To avoid multiplicity and over-interpretation, no inferential subgroup analyses (eg, group×sex interactions) were conducted; inference focuses on overall between-group effects.

To address missing data, Little test for missing completely at random (MCAR) was applied to all outcomes measured at T0 and T1. The results showed no significant departure from MCAR (χ^2_{15}=11.236; P=.73). Given the low proportion of missing values (<5%) and to preserve statistical power and reduce potential bias associated with listwise deletion, multiple imputation (m=20) using chained equations was performed under a missing-at-random assumption. The imputation model included group assignment (IG vs CG), sex, age, baseline values of each outcome, and adherence indicators (eg, percentage of days meeting the PA goal) as predictors. Robustness was confirmed by reporting both complete-case and multiple

XSL•FO
RenderX

imputation–pooled estimates for primary outcomes and through sensitivity analyses that excluded algorithm-flagged suspect data, which yielded consistent conclusions. All statistical tests were 2-sided, with $P<.05$ considered statistically significant.

Analyses were performed using SPSS (version 26.0; IBM Corp).

Ethical Considerations

The study received ethics approval (Approval No. 2025A125, approved on January 15, 2025) from the Ethics Committee of Capital University of Physical Education and Sports. The study protocol was finalized before the start of participant recruitment (February 4, 2025) and is identical to the protocol retrospectively registered in the Chinese Clinical Trial Registry (ChiCTR2500108950) with respect to trial design, eligibility criteria, prespecified primary and secondary outcomes, and planned statistical analyses. No changes were made to these elements after trial commencement.

Trial registration was completed retrospectively in the Chinese Clinical Trial Registry (ChiCTR2500108950) on September 9, 2025, after enrollment had begun, because prospective registration was inadvertently not completed during trial setup due to an administrative oversight; this was not an attempt to avoid registration. To mitigate concerns about selective reporting, the institutional review board (IRB)–approved protocol was finalized prior to recruitment and remained unchanged throughout the trial with respect to the trial design, eligibility criteria, prespecified outcomes, and planned statistical analyses. An official letter from the IRB confirming that the submitted study design is identical to the protocol assessed and approved before trial initiation has been provided to the editorial office for review.

All procedures were conducted in accordance with the principles of the Declaration of Helsinki. Before enrollment, all participants received written and verbal information about the study purpose, procedures, potential risks, and anticipated benefits and subsequently provided written informed consent. Participants were informed that participation was voluntary and that they could withdraw from the study at any time without providing a reason. This withdrawal would incur no penalty or detriment to their academic standing. To ensure privacy and confidentiality, all participant data were deidentified upon collection and assigned a unique study ID. The analytical dataset contained only these anonymized IDs. The master list linking IDs to personal information was stored separately on a secure, password-protected server with access restricted to the principal investigator and the data manager. Participants who completed the entire intervention cycle received nonmonetary compensation (accumulated volunteer hours, a certificate of honor, and coordination with academic credit where applicable). Additionally, the top 10% of performers received a material reward of 400 Chinese yuan (Renminbi; approximately US $55.53 or €49.31, based on the exchange rate in April 2025). No identifiable images of participants are presented in this manuscript or in the supplementary materials; the participants pictured in Figure 1F provided written consent for publication of the deidentified photographs for academic purposes.

Data Security and Risk Management

To prevent cheating and ensure data integrity, each smartwatch (serial number/MAC) was identity-verified and bound 1:1 to a unique participant account. Weekly, a random 10% sample of participants underwent raw-data audits in which 7-day device traces (steps, moderate-to-vigorous physical activity [MVPA], metabolic equivalent-minutes [MET-min], and wear time with timestamps) were cross-checked against server sync logs; discrepant days were documented and flagged for sensitivity analyses. For privacy and confidentiality, analytical datasets were deidentified, access followed role-based access control under a least-privilege policy, and data were used solely for academic research as approved by the ethics committee; all access and exports were audit-logged. To reduce attrition, participants received a standardized briefing on study requirements before randomization; withdrawals were allowed at any time, and reasons were recorded and summarized in the CONSORT flow.

Harms

Safety was monitored through weekly checks and audits using predefined triggers; events were logged by a blinded assessor. Serious adverse events required immediate IRB reporting and discontinuation as prespecified in the protocol. Patients or members of the public were not involved in this trial. No important changes to the trial protocol, outcomes, or analyses were made after trial initiation.

Results

Participant Flow

A total of 224 volunteers were screened for eligibility. After excluding 64 individuals who did not meet the inclusion criteria, 160 participants were enrolled and randomly assigned in a 1:1 ratio to the IG and CG, with 80 participants in each group (Figure 3).

XSL•FO
RenderX

Figure 3. CONSORT (Consolidated Standards of Reporting Trials) flow diagram, participant flow through the randomized trial. CG: control group; IG: intervention group; PA: physical activity;

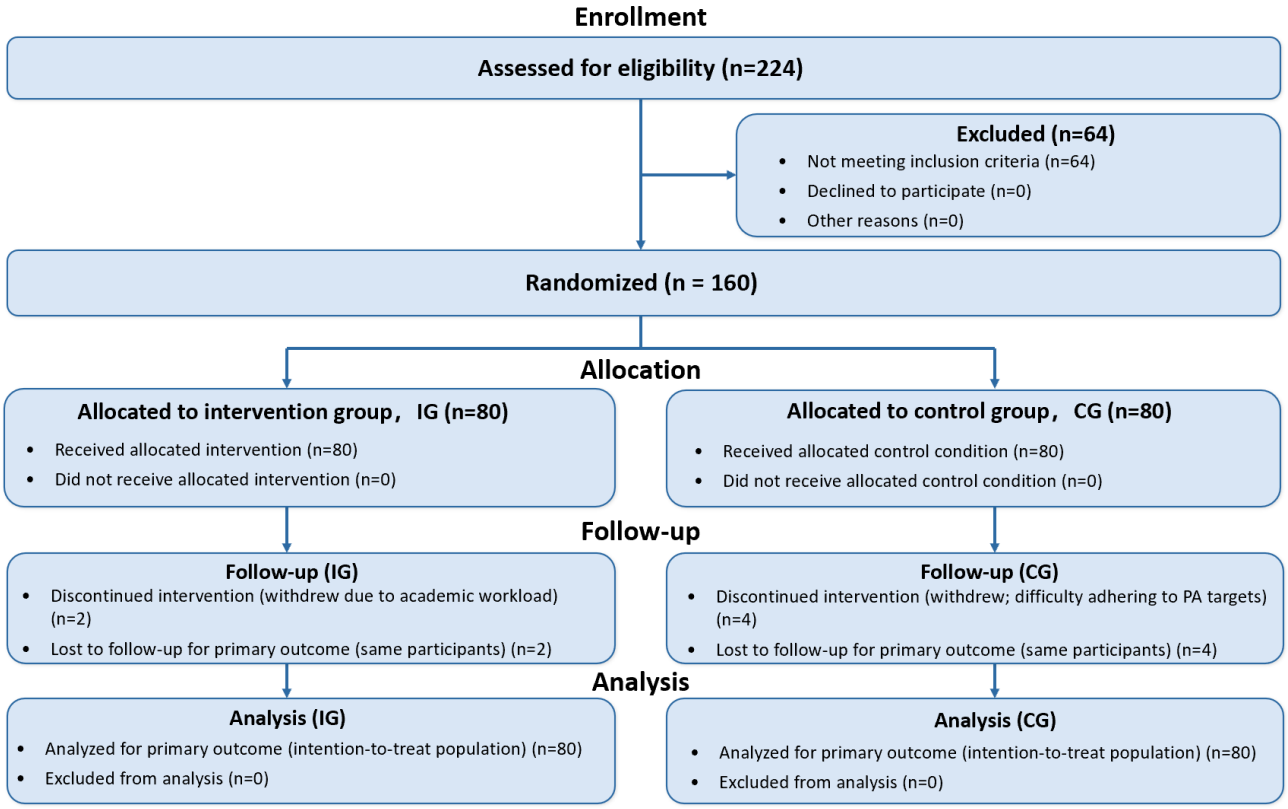

During the 8-week intervention period, 6 participants withdrew from the study (IG: 2 participants, 2.5%; CG: 4 participants, 5%). The primary reasons for dropout were academic workload (n=2) and failure to meet PA targets (n=4). Overall, this resulted in 3.8% missingness in post-intervention outcomes. Little MCAR test across all primary and secondary outcomes indicated that this missingness did not significantly depart from MCAR (χ^2_{15}=11.236; *P*=.73).

The trial ended as planned after completion of the prespecified 8-week intervention and postintervention assessments and was not stopped early.

Baseline Data

There were no statistically significant differences between the 2 groups in baseline demographic characteristics (age, sex, and BMI) or psychological scale scores (*P*>.05), indicating good comparability between group (Table 1).

XSL•FO
RenderX

Table . Baseline characteristics of participants.

Characteristic	IG[a] (n=80)	CG[b] (n=80)	Total (N=160)	*P* value
Age (years), mean (SD)	20.5 (1.2)	20.8 (1.3)	20.6 (1.3)	.15
Sex, n (%)				
Female	40 (50.0)	40 (50.0)	80 (50.0)	>.99
Male	40 (50.0)	40 (50.0)	80 (50.0)	>.99
Body composition, mean (SD)				
Body weight, (kg)	66.4 (12.5)	65.8 (11.8)	66.1 (12.1)	.76
BMI (kg/m²)	22.4 (2.5)	22.1 (2.3)	22.2 (2.4)	.42
Mental health, mean (SD)				
SAS[c] score	44.3 (7.2)	45.1 (5.9)	44.7 (6.6)	.43
SDS[d] score	46.4 (7.7)	45.9 (8.1)	46.2 (7.9)	.67
PA[e]				
Daily steps, mean (SD)	7842 (1,603)	7921 (1,425)	7882 (1,514)	.74
MVPA[f] (min/day), mean (SD)	37.2 (11.5)	37.4 (10.7)	37.3 (11.1)	.91
Weekly MET-min[g], mean (SD)	1020 (205)	1035 (220)	1028 (212)	.66

[a]IG: intervention group.

[b]CG: control group.

[c]SAS: Self-Rating Anxiety Scale.

[d]SDS: Self-Rating Depression Scale.

[e]PA: physical activity.

[f]MVPA: moderate-to-vigorous physical activity.

[g]MET-min: metabolic equivalent-minute.

Intervention and Comparator Delivery

Both arms were technology-mediated and primarily self-administered; the study team provided standardized onboarding, technical support, and data monitoring. No crossovers or major deviations from the intended delivery were identified, and adherence indicators (goal completion and data synchronization) are summarized in Table 2.

XSL•FO
RenderX

Table . Outcome during the 8-week intervention and change from T0 to T1.

Domain and outcome[a]	IG[b] (n=80), mean (SD)	CG[c] (n=80), mean (SD)	Δ (IG-CG)	Δ IG-CG (%)	*P* value	Effect size, *d*[d] (95% CI)
PA[e] levels[f]						
Daily steps	10,356 (1245)	8242 (1087)	+2114	25.6	<.001	1.81 (1.44 to 2.18)
Daily MVPA[g] (min)	71 (15)	43 (12)	+28	65.1	<.001	2.06 (1.68 to 2.45)
Weekly MET-min[h]	1650 (310)	1340 (285)	+310	23.1	<.001	1.04 (0.71 to 1.37)
Adherence (%)						
Goal completion rate	88.9 (8.5)	77.5 (10.2)	+11.4	14.7	<.001	1.21 (0.88 to 1.55)
Data sync rate	89.1 (7.7)	75.5 (11.3)	+13.6	18.0	<.001	1.41 (1.06 to 1.75)
Body composition						
Δ Skeletal muscle (kg)	+0.86 (0.23)	+0.32 (0.11)	+0.54	168.8	<.001	3.00 (2.54 to 3.45)
Δ Body fat (%)	−1.46 (0.51)	−1.00 (0.32)	−0.46	46.0	<.001	−1.08 (−1.41 to −0.75)
Δ BMR[i] (kcal/d)	+20.7 (4.9)	+12.9 (4.8)	+7.8	60.5	<.001	1.61 (1.25 to 1.97)
Δ Body weight (kg)	−1.17 (1.45)	−0.91 (1.45)	−0.26	28.6	.26	−0.18 (−0.49 to 0.13)
Δ BMI (kg/m²)	−0.45 (1.10)	−0.26 (1.04)	−0.19	73.1	.26	−0.18 (−0.49 to 0.13)
Physical fitness						
Δ 800/1000 m run (s)	−13.0 (6.0)	−6.5 (4.0)	−6.5	100.0	<.001	−1.27 (−1.62 to −0.93)
Δ Vital capacity (mL)	+470 (110)	+280 (80)	+190	67.9	<.001	1.98 (1.60 to 2.35)
Δ Sit-and-reach (cm)	+5.2 (1.1)	+3.4 (1.2)	+1.8	52.9	<.001	1.56 (1.21 to 1.92)
Δ Sit-ups (female), counts (n=40 per group)	+10.5 (2.0)	+6.0 (1.5)	+4.5	75.0	<.001	2.55 (1.95 to 3.14)
Δ Pull-ups (male, counts, n=40 per group)	+0.6 (0.5)	+0.4 (0.5)	+0.2	50.0	.08	0.40 (−0.04 to 0.84)
Δ 50 m sprint (s)	−0.30 (0.30)	−0.25 (0.30)	−0.05	20.0	.28	−0.17 (−0.48 to 0.14)
Δ Standing long jump (cm)	+8.5 (13.1)	+4.5 (14.5)	+4.0	88.9	.07	0.29 (−0.02 to 0.60)
EF[j]						
Δ Cognitive flexibility (ms)	−104 (26)	−75 (32)	−29	38.7	<.001	−0.99 (−1.32 to −0.67)
Δ Inhibitory control (errors)	−1.4 (2.5)	−1.0 (2.5)	−0.4	40.0	.31	−0.16 (−0.47 to 0.15)
Δ Working memory (correct)	+0.5 (1.5)	+0.3 (1.3)	+0.2	66.7	.38	0.14 (−0.17 to 0.45)
Mental health						
Δ SDS[k] (depression)	−4.5 (2.0)	−2.5 (1.5)	−2.0	80.0	<.001	−1.13 (−1.47 to −0.80)

XSL•FO
RenderX

Domain and outcome[a]	IG[b] (n=80), mean (SD)	CG[c] (n=80), mean (SD)	Δ (IG-CG)	Δ IG-CG (%)	*P* value	Effect size, *d*[d] (95% CI)
Δ SAS[l] (anxiety)	−2.2 (3.5)	−1.3 (3.1)	−0.9	69.2	.09	−0.27 (−0.58 to 0.04)

[a]Values are mean (SD). Analyses followed intention-to-treat with multiple imputation

[b]IG: intervention group.

[c]CG: control group.

[d]Effect sizes are Cohen *d* (pooled SD); the sign of *d* follows Δ (IG−CG).

[e]PA: physical activity.

[f]PA and adherence outcomes were summarized over the 8-week intervention period; outcomes prefixed with Δ are change scores (T1−T0).

[g]MVPA: moderate-to-vigorous physical activity.

[h]MET-min: metabolic equivalent-minutes.

[i]BMR: basal metabolic rate.

[j]EF: executive function.

[k]SDS: Self-Rating Depression Scale.

[l]SAS: Self-Rating Anxiety Scale.

No concomitant care was provided by the study team, and concomitant interventions outside the trial that could influence outcomes were not systematically monitored.

Numbers Analyzed

Primary and secondary outcomes were analyzed according to the ITT principle (IG n=80; CG n=80). Missing postintervention (T1) values were handled using multiple imputation as prespecified. Outcome data were available at T1 for 78/80 participants in the IG (97.5%) and 76/80 in the CG (95.0%). Table 2 presents MI-pooled descriptive statistics and between-group inferences based on the primary ITT analysis.

Outcomes and Estimation

Primary Outcomes

PA Levels

Across the 8-week period, both arms showed the typical early rise-mid plateau-late rebound pattern. Adjusted between-group differences consistently favored IG for steps, daily MVPA, and weekly MET-min throughout the intervention (all *P*<.001). These results indicate that the gamified mHealth arm elicited additional PA beyond baseline platform targets; see Table 2 for detailed outcomes and 95% CIs and Figure 4 for effect sizes.

XSL•FO
RenderX

Figure 4. Effect sizes (*d*) and 95% CIs for each outcome measure (IG vs CG): positive values favor IG (directions standardized). The dashed vertical line indicates no between-group difference (*d*=0); CIs crossing 0 indicate nonsignificance. MET-min: metabolic equivalent-minute; MVPA: moderate-to-vigorous physical activity; SAS: Self-Rating Anxiety Scale; SDS: Self-Rating Depression Scale.

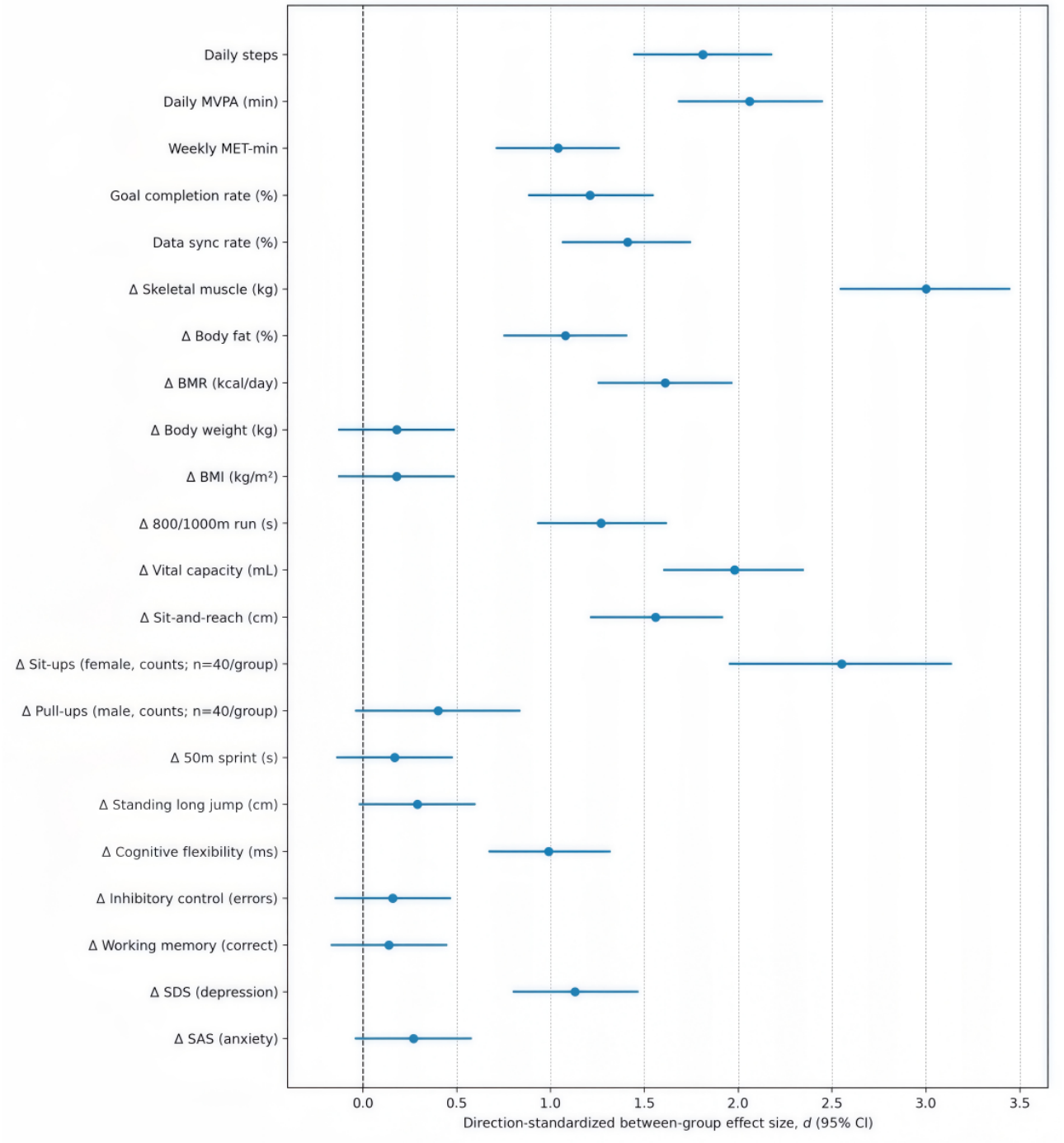

Intervention Adherence

IG demonstrated significantly higher adherence—for both goal completion and data synchronization—than CG (all $P<.001$). Exact means, between-group differences (Δ), and effect sizes with 95% CIs are reported in Table 2.

Secondary Outcomes

Physical Fitness

The intervention produced differential effects across fitness levels. For physical fitness, IG improved more than CG in aerobic endurance (800/1000 m run), vital capacity, muscular endurance (female sit-ups), and flexibility (sit-and-reach) (all $P<.001$), while no between-group differences emerged for 50 m sprint, pull-ups, or standing long jump ($P>.05$). See Table 2 and Figure 4 for detailed outcomes and 95% CIs.

Body Composition

Relative to CG, IG showed larger body-composition gains, higher skeletal muscle mass, lower body fat percentage, and higher BMR (all $P<.001$), while changes in body weight and BMI showed no significant between-group differences ($P>.05$). See Table 2 for detailed outcomes and 95% CIs and Figure 4 for effect sizes.

XSL•FO
RenderX

EF

EF showed a domain-specific pattern, cognitive flexibility improved in IG (P<.001), whereas working memory and inhibitory control showed no between-group differences (P>.05). See Table 2 and Figure 4 for detailed outcomes and 95% CIs.

Mental Health Status

Depressive symptoms (SDS) decreased more in IG than in CG (P<.001), whereas anxiety (SAS) changes were small and not different between group (P>.05). Individual trajectories varied; see Table 2 and Figure 4 for detailed outcomes and 95% CIs.

Ancillary Analyses

Given the lack of prespecification for sex-based subgroup inference, no inferential sex comparisons are reported; primary conclusions rely on overall effects.

Harms

No study-related serious adverse events; no device-related dermatologic events requiring care; and no major technical failures or data losses affecting outcome interpretation were recorded.

Discussion

Principal Results

This RCT evaluated the incremental effect of adding team-based gamification and incentives to a standard wearable- and app-based PA program for college students. Compared with an active control using the identical platform and activity goals, the IG showed greater improvements in the prespecified primary outcomes—objectively measured PA and intervention adherence—and in selected secondary outcomes related to physical fitness and body composition, whereas effects on EF and anxiety outcomes were limited. This overall pattern is consistent with meta-analytic evidence that gamified mHealth interventions tend to increase daily steps/MVPA and yield small-to-moderate improvements in adiposity-related indicators compared with nongamified comparators [33,34]. Notably, by holding the app-wearable platform and activity goals constant across groups, this study may help isolate the incremental contribution of gamification and reduce confounding that is common when platforms and components differ across conditions [35]. The findings indicate that the added value of gamification in this context may operate primarily through strengthening adherence and engagement, leading to downstream benefits for selected physical outcomes [36].

Mechanisms

Improvements in PA and adherence may be interpreted within SDT and TPB, in that need-supportive gamification could facilitate motivational internalization and help translate intention into action [37]. First, SDT posits that autonomy, competence, and relatedness are core components of intrinsic motivation [38,39]. In this study, team-based competition, real-time feedback, and visible progress cues may have supported competence and relatedness while maintaining task structure to promote more sustained engagement in PA.

In addition, the intervention combined support for basic psychological needs (autonomy, competence, and relatedness) with extrinsic incentives (points and leaderboards) [40]. From an SDT perspective, a key issue is how such external incentives are framed and delivered [41]. When points and leaderboards are presented as informational signals of personal progress rather than controlling demands, they may better support autonomy and align with the regulatory continuum of motivational internalization [42], which could help explain the observed improvements in adherence during the intervention period.

More specifically, short-term gains in PA and adherence may reflect a synergy between extrinsic prompts and need-supportive features: the former may catalyze initial participation and short-term persistence, whereas the latter may enhance self-efficacy and belonging by satisfying competence and relatedness needs [43]. Nevertheless, the salience and material nature of certain incentives may slow internalization toward more integrated regulation [44]. In this RCT, in addition to designing nonmaterial rewards such as cumulative volunteer hours, honorary certificates, and academic credit coordination, we also incorporated a conditional monetary reward for the IG, tied to leaderboard performance. This external economic incentive served as a salient extrinsic goal and could bolster adherence and help sustain PA, particularly during the mid-to-late stages of the intervention. However, its effect may attenuate after the incentive is withdrawn and could further limit motivational internalization [45]. Such incomplete internalization could partly limit the transfer of benefits to domains that typically require greater cognitive engagement, longer accumulation, or additional components beyond increasing activity volume (eg, strength-related fitness and selected EF domains) [46]. Evidence from multi-arm randomized trials directly comparing gamification, financial incentives, and their combination supports the plausibility that each component can increase PA and that combining them may yield larger improvements than either component alone, consistent with additive or synergistic effects [47]. Therefore, the intervention benefits observed in the present study are more likely the result of the combined action of the engaging environment created by gamification and the salient extrinsic goals set by monetary rewards.

Overall, the SDT-based account primarily addresses motivation formation and internalization. TPB provides a complementary perspective by clarifying how attitudes, subjective norms, and perceived behavioral control jointly shape the translation of behavioral intention into actual behavior [48]. In the present intervention, digital reminders, social-comparison signals, and personalized feedback may have strengthened perceived control and normative support, thereby facilitating daily goal execution and adherence [49].

Interpretation and Boundaries

Unlike prior physical fitness–promotion interventions that typically report small-to-moderate effect sizes [7], the present mHealth-based intervention with added gamification showed comparatively large standardized effect sizes for objectively measured physical activity and selected fitness-related outcomes. This pattern may reflect the intervention's efficacy under

XSL•FO
RenderX

relatively controlled trial conditions. First, the single-university student cohort was relatively homogeneous in age range and daily routines, which may reduce within-sample outcome variability and thereby inflate standardized effect size estimates for a given raw mean difference [50]. Second, delivery was supported by a technology platform with high implementation intensity (eg, frequent monitoring, scheduled synchronization, and staged feedback) and high adherence during the intervention period (Table 2), which likely improved intervention fidelity and strengthened the effective "dose contrast" between groups. Prior evidence suggests that higher engagement/adherence in digital interventions is associated with larger behavioral effects [51]. However, when such interventions are scaled from controlled trial settings to more heterogeneous real-world populations with lower monitoring/support intensity, observed effect sizes may attenuate [52]. Building on this overall interpretation, we now turn to the domain-specific findings and key boundaries.

The higher frequency of MVPA in IG may have contributed to improvements in aerobic capacity [53]. This is plausible because the intervention primarily promoted cardiovascular adaptation by increasing activity volume, and the achieved activity pattern was broadly aligned with WHO-recommended PA levels [23]. In contrast, between-group differences were not statistically significant for strength- or power-related outcomes (eg, 50-m sprint, standing long jump, pull-ups). A primary explanation may lie in the structural emphasis of the intervention: the program was designed to promote MVPA via mHealth, with common modalities such as jogging and cycling that primarily load cardiorespiratory and energy-metabolic systems, but it did not include a systematic resistance-training component targeting large muscle groups [54]. Improvements in strength/power generally require more targeted resistance training, and baseline capability (eg, the ability to perform pull-ups) may also influence the magnitude of improvement [55,56].

Although the intervention did not yield statistically significant changes in body weight or BMI, IG showed significant improvements in body-composition indicators (eg, skeletal muscle mass, body fat percentage, and BMR) relative to CG. This pattern may be explained by the fact that, in the absence of strict dietary control and a standardized resistance-training dose, increasing PA can preferentially improve body composition [57]. Moreover, because adipose tissue has a lower density than muscle tissue, concurrent fat loss and muscle gain may result in a "weight substitution effect," whereby meaningful changes in body shape occur with minimal change in total body weight [58]. In addition, reductions in fat mass together with increases in lean mass may elevate resting metabolic rate, consistent with the metabolic benefits associated with increases in fat-free mass [59].

For EF, IG showed a statistically significant improvement only in cognitive flexibility, whereas changes in working memory and inhibitory control did not reach statistical significance. This suggests that cognitive flexibility may be a domain more readily influenced by this PA-focused intervention. One possible mechanism is the acute and cumulative effects of moderate-intensity aerobic exercise on prefrontal hemodynamics and neural signaling [60]. By contrast, limited changes in working memory and inhibitory control are also consistent with the general "dose-response" and domain-specific nature of EF adaptation [61], whereby these domains may require higher intensity, longer duration, or more cognitively engaging activity to change meaningfully [62,63].

The intervention showed limited effects on anxiety symptoms, which may reflect that the protocol primarily targeted PA behavior and did not incorporate psychological components (eg, cognitive behavioral therapy, mindfulness, or stress-management modules) that more directly modulate emotional processes [64]. In addition, gamification mechanisms may be double-edged: competitive elements can enhance motivation, but they may also thwart autonomy and relatedness for some participants, potentially constraining psychological benefits [65]. Prior evidence also suggests substantial heterogeneity in mental health responses to gamified interventions, which may relate to individual sensitivity to social comparison and competitive pressure [66]. Moreover, PA tends to show stronger effects on depressive symptoms than on anxiety, and benefits may be more apparent among participants with milder baseline symptom levels, implying conditional efficacy [67].

Taken together, changes in EF and mental health may be interpreted as potential, conditional downstream benefits of PA-focused interventions rather than inevitable effects of gamification per se [62]. This further suggests that when interventions aim to target mental health or EF, it is important to align mechanisms, dose/intensity, and the target population to achieve more precise matching and maximize the likelihood of benefit [68].

Limitations

This study used an RCT design and leveraged mHealth technology to enable real-time tracking and feedback of intervention behaviors. Nevertheless, several limitations should be acknowledged.

First, the 8-week intervention duration, while sufficient for assessing short-term efficacy, may be insufficient to observe long-term behavioral maintenance and sustained neural or physiological adaptations. Future studies with longer follow-up periods are needed to evaluate the sustainability of the effects [69]. Second, the absence of systematic control over dietary intake and sleep patterns represents a potential confounding factor for the observed changes in body composition, as these lifestyle components are known to interact with PA [70]. Third, all participants were recruited from a single university, which may limit generalizability to other collegiate populations and demographic groups; the relatively homogeneous cohort may also have reduced outcome variability [71]. Fourth, the team sex composition was implemented for fairness and feasibility rather than for testing sex differences; the study was neither powered to detect effect modification by sex, nor were such comparisons prespecified. Additionally, because conditional financial incentives were bundled with gamification only in the IG, we could not disentangle the independent effects of incentives versus gamification [72]. Moreover, although the protocol had been finalized before recruitment and remained unchanged throughout the trial, the retrospective registration

XSL•FO
RenderX

of the trial due to an administrative oversight may still be considered a limitation in terms of methodological transparency and potential risk of bias. Finally, the evaluation of EF and mental health, although we cited literature [73], is limited by the lack of baseline impairment, which limits sensitivity to change and raises the possibility of ceiling effects.

Implications

Overall, this 8-week RCT showed that, compared with a nongamified (monitoring only) intervention with identical activity goals, a gamified mHealth exercise intervention delivered within the same app-wearable system yielded clear incremental benefits: it increased PA and intervention adherence and improved selected physical fitness and body-composition indicators. The key innovation lies in quantifying the "incremental effect" attributable to gamification on top of a unified activity goal and the same technical platform, thereby allowing a clearer estimate of the incremental contribution of gamification. This approach differed from much of the prior literature, where between-group comparisons were often confounded by co-occurring intervention components [35], and it reduced such confounding by holding the technology platform and activity goals constant across groups. In terms of evaluation, by integrating objective behavioral data with a multidimensional outcome set, the study offered more robust and comparable evidence and helped delineate where gamification is most likely to be effective and where its effects may be bounded. From a real-world perspective, the model could be scaled at low cost by leveraging existing smartphone apps and wearable devices, supporting feasible implementation in university health-promotion and campus health management programs and offering an accessible digital health strategy to address physical inactivity among college students and inform related public health practice and policy [74].

Future Directions

Building on our findings and limitations, we propose several avenues for future research: (1) investigate integration of multimodal components, such as the integration of structured resistance training components to more effectively target muscular strength and power, which showed limited improvement here [75]; (2) extend intervention exposure (eg, >12 - 16 wk) and prespecify dose-response targets for PA to test whether benefits for domains like working memory and inhibition emerges with a greater cumulative dose [76]; (3) explore synergies with evidence-based psychological approaches. This includes embedding digital cognitive behavioral therapy elements or mindfulness training within the mHealth platform to directly target mechanisms underlying anxiety, depression, and cognitive regulation [77]; (4) testing the adapted intervention in more diverse populations, including individuals with baseline subclinical symptoms or lower digital literacy, will be crucial for determining its broader public health applicability [78]; (5) add pragmatic diet/sleep capture (eg, 1 - 2 24-hour dietary recalls per week including one weekend day, photo-assisted intake logs, and 7-day sleep diary + actigraphy/wearable-derived sleep/heart rate) to strengthen causal attribution for body-composition outcomes [79]; (6) preregister sex-related hypotheses, power accordingly, and conduct planned interaction/subgroup analyses to evaluate potential heterogeneity of effects by sex and to align analytic decisions with study aims [80]; and (7) use multiarm or factorial designs (eg, gamification-only, incentive-only, both, and control), with postincentive follow-up, to isolate component-specific effects and assess maintenance [47].

Conclusions

This RCT demonstrates that an mHealth intervention integrating gamification effectively enhanced physical fitness, body composition, and adherence to PA among college students. While its impact on psychological health and EF was limited, the approach demonstrates strong potential for scalable, technology-driven health promotion in young adult populations.

However, because this study did not include post-intervention follow-up, the sustainability of the effects was not assessed. Future research should evaluate maintenance and scalability in more diverse real-world settings, ideally using multiarm/factorial designs with postincentive follow-up. Studies should also test longer intervention exposure, multimodal components, and dietary strategies.

Acknowledgments

This work was supported by the Emerging Interdisciplinary Platform for Medicine and Engineering in Sports (EIPMES).

Funding

This study was funded by the Emerging Interdisciplinary Platform for Medicine and Engineering in Sports (EIPMES) and the teaching reform program of Yantai University (Grant No. JYXM2024025Y).

Data Availability

The trial protocol, statistical analysis plan, de-identified individual participant data (including the data dictionary), statistical code, and other study materials are available from the corresponding author on reasonable request.

Authors' Contributions

MY and ZZ performed conceptualization. MY, Y Guo, ZZ, SY, ZL, Y Gao, and JL performed methodology. Y Gao conducted the literature search. MY and ZZ wrote the original draft. ZZ performed visualization. All authors contributed to writing – review & editing.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Editorial Notice

This randomized study was only retrospectively registered because the lack of prospective registration was due to an administrative oversight during trial setup, which resulted in the registration being omitted from the trial-start checklist. Once the omission was identified while checking the journal's submission and reporting requirements for randomized controlled trials, the authors retrospectively registered the trial in the Chinese Clinical Trial Registry. The editor granted an exception from ICMJE rules mandating prospective registration of randomized trials because the risk of bias appears low and because the authors provided an official letter from the institutional review board confirming that the submitted study design is identical to the protocol assessed and approved before trial initiation. However, readers are advised to carefully assess the validity of any potential explicit or implicit claims related to primary outcomes or effectiveness, as retrospective registration does not prevent authors from changing their outcome measures retrospectively.

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 245 KB - jmir_v28i1e82769_app1.pdf]

References

1. Jia X, Hu C, Xu Y, et al. Revisiting obesity thresholds for cardiovascular disease and mortality risk in Chinese adults: age- and gender-specific insights. Cell Rep Med 2025 Sep 16;6(9):102309. [doi: 10.1016/j.xcrm.2025.102309] [Medline: 40858104]
2. Sun Q, Yu D, Fan J, et al. Healthy lifestyle and life expectancy at age 30 years in the Chinese population: an observational study. Lancet Public Health 2022 Dec;7(12):e994-e1004. [doi: 10.1016/S2468-2667(22)00110-4] [Medline: 35926549]
3. Sun Z, Sun M, Wei W, Peng W, Wang Y. China launches National Obesity Campaign. Lancet Diabetes Endocrinol 2025 Jun;13(6):465-466. [doi: 10.1016/S2213-8587(25)00131-7] [Medline: 40404276]
4. Lee SA, Park JH. Systematic review and meta analysis of standalone digital behavior change interventions on physical activity. NPJ Digit Med 2025 Jul 14;8(1):436. [doi: 10.1038/s41746-025-01827-4] [Medline: 40659761]
5. Beratarrechea A, Lee AG, Willner JM, Jahangir E, Ciapponi A, Rubinstein A. The impact of mobile health interventions on chronic disease outcomes in developing countries: a systematic review. Telemed J E Health 2014 Jan;20(1):75-82. [doi: 10.1089/tmj.2012.0328] [Medline: 24205809]
6. Mönninghoff A, Kramer JN, Hess AJ, et al. Long-term effectiveness of mHealth physical activity interventions: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2021 Apr 30;23(4):e26699. [doi: 10.2196/26699] [Medline: 33811021]
7. Mazeas A, Duclos M, Pereira B, Chalabaev A. Evaluating the effectiveness of gamification on physical activity: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2022 Jan 4;24(1):e26779. [doi: 10.2196/26779] [Medline: 34982715]
8. Looyestyn J, Kernot J, Boshoff K, Ryan J, Edney S, Maher C. Does gamification increase engagement with online programs? A systematic review. PLoS ONE 2017;12(3):e0173403. [doi: 10.1371/journal.pone.0173403] [Medline: 28362821]
9. Autin KL, Herdt ME, Garcia RG, Ezema GN. Basic psychological need satisfaction, autonomous motivation, and meaningful work: a self-determination theory perspective. J Career Assess 2022 Feb;30(1):78-93. [doi: 10.1177/10690727211018647]
10. Abou Kamar M, Maher A, Salem IE, Elbaz AM. Gamification impact on tourists' pro-sustainability intentions: integration of technology acceptance model (TAM) and the theory of planned behaviour (TPB). TR 2024 Feb 6;79(2):487-504. [doi: 10.1108/TR-04-2023-0234]
11. Girelli L, Hagger M, Mallia L, Lucidi F. From perceived autonomy support to intentional behaviour: testing an integrated model in three healthy-eating behaviours. Appetite 2016 Jan 1;96:280-292. [doi: 10.1016/j.appet.2015.09.027] [Medline: 26423363]
12. Jang D, Kim I, Kwon S. Motivation and intention toward physical activity during the COVID-19 pandemic: perspectives from integrated model of self-determination and planned behavior theories. Front Psychol 2021;12:714865. [doi: 10.3389/fpsyg.2021.714865] [Medline: 34393954]
13. Wang Q, Liu J, Zhou L, et al. Usability evaluation of mHealth apps for elderly individuals: a scoping review. BMC Med Inform Decis Mak 2022 Dec 2;22(1):317. [doi: 10.1186/s12911-022-02064-5] [Medline: 36461017]
14. Schmidt-Kraepelin M, Warsinsky S, Thiebes S, Sunyaev A. The role of gamification in health behavior change: a review of theory-driven studies. Presented at: Hawaii International Conference on System Sciences; Jan 7-10, 2020. [doi: 10.24251/HICSS.2020.155]
15. Damghanian M, Alijanzadeh M. Theory of planned behavior, self-stigma, and perceived barriers explains the behavior of seeking mental health services for people at risk of affective disorders. Asian J Soc Health Behav 2018;1(2):54-61. [doi: 10.4103/SHB.SHB_27_18]

XSL•FO
RenderX

16. Kim HM, Cho I, Kim M. Gamification aspects of fitness apps: implications of mHealth for physical activities. Int J Hum Comput Interact 2023 Jun 15;39(10):2076-2089. [doi: 10.1080/10447318.2022.2073322]
17. Yu T, Parry M, Yu T, et al. Effectiveness of mobile health-based gamification interventions for improving physical activity in individuals with cardiovascular diseases: systematic review and meta-analysis of randomized controlled trials. JMIR Serious Games 2025 Jan 24;13:e64410. [doi: 10.2196/64410] [Medline: 39854099]
18. Xu L, Li J, Zhang X, et al. Mobile health-based gamification intervention to increase physical activity participation among patients with coronary heart disease: study protocol of a randomised controlled trial. BMJ Open 2022 Jan 31;12(1):e054623. [doi: 10.1136/bmjopen-2021-054623] [Medline: 35105640]
19. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025 Apr 14;389:e081123. [doi: 10.1136/bmj-2024-081123] [Medline: 40228833]
20. Eysenbach G, Group CE. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
21. Hopewell S, Clarke M, Moher D, et al. CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. PLoS Med 2008 Jan 22;5(1):e20. [doi: 10.1371/journal.pmed.0050020] [Medline: 18215107]
22. Zhang J, Li Z, Li M, et al. High-precision monitoring of fitness behavior and intelligent recommendation of personalized fitness programs. J Capital Univ Phys Educ Sports 2023;35(3):276-283. [doi: 10.14036/j.cnki.cn11-4513.2023.03.004]
23. WHO Guidelines on Physical Activity and Sedentary Behaviour: World Health Organization; 2020. URL: https://www.who.int/publications/i/item/9789240015128 [accessed 2025-03-20]
24. Contreras-Osorio F, Ramirez-Campillo R, Cerda-Vega E, et al. Effects of physical exercise on executive function in adults with depression: a systematic review and meta-analysis. Int J Environ Res Public Health 2022 Nov 18;19(22):15270. [doi: 10.3390/ijerph192215270] [Medline: 36429985]
25. Ángel Latorre-Román P, Berrios-Aguayo B, Aragón-Vela J, Pantoja-Vallejo A. Effects of a 10-week active recess program in school setting on physical fitness, school aptitudes, creativity and cognitive flexibility in elementary school children. A randomised-controlled trial. J Sports Sci 2021 Jun 3;39(11):1277-1286. [doi: 10.1080/02640414.2020.1864985]
26. Yue T, Li Q, Wang R, et al. Comparison of Hospital Anxiety and Depression Scale (HADS) and Zung Self-Rating Anxiety/Depression Scale (SAS/SDS) in evaluating anxiety and depression in patients with psoriatic arthritis. Dermatology (Basel) 2020;236(2):170-178. [doi: 10.1159/000498848] [Medline: 31434087]
27. Bollimbala A, James PS, Ganguli S. The impact of physical activity intervention on creativity: role of flexibility vs persistence pathways. Think Skills Creat 2023 Sep;49:101313. [doi: 10.1016/j.tsc.2023.101313]
28. Schmidt-Kraepelin M, Toussaint PA, Thiebes S, Hamari J, Sunyaev A. Archetypes of gamification: analysis of mHealth apps. JMIR Mhealth Uhealth 2020 Oct 19;8(10):e19280. [doi: 10.2196/19280] [Medline: 33074155]
29. Zolfaghari M, Shirmohammadi M, Shahhosseini H, Mokhtaran M, Mohebbi SZ. Development and evaluation of a gamified smart phone mobile health application for oral health promotion in early childhood: a randomized controlled trial. BMC Oral Health 2021 Jan 7;21(1):18. [doi: 10.1186/s12903-020-01374-2] [Medline: 33413304]
30. Cechetti NP, Bellei EA, Biduski D, Rodriguez JPM, Roman MK, De Marchi ACB. Developing and implementing a gamification method to improve user engagement: a case study with an m-Health application for hypertension monitoring. Telematics and Informatics 2019 Aug;41:126-138. [doi: 10.1016/j.tele.2019.04.007]
31. Wang M, Xu J, Zhou X, Li X, Zheng Y. Effectiveness of gamification interventions to improve physical activity and sedentary behavior in children and adolescents: systematic review and meta-analysis. JMIR Serious Games 2025 Sep 18;13:e68151. [doi: 10.2196/68151] [Medline: 40966596]
32. Fernandez-Rio J, de las Heras E, González T, Trillo V, Palomares J. Gamification and physical education. Viability and preliminary views from students and teachers. Phys Educ Sport Pedagogy 2020 Sep 2;25(5):509-524. [doi: 10.1080/17408989.2020.1743253]
33. Gao Y, Zhang J, He Z, Zhou Z. Feasibility and usability of an artificial intelligence-powered gamification intervention for enhancing physical activity among college students: quasi-experimental study. JMIR Serious Games 2025 Mar 24;13:e65498. [doi: 10.2196/65498] [Medline: 40127464]
34. Patel MS, Small DS, Harrison JD, et al. Effectiveness of behaviorally designed gamification interventions with social incentives for increasing physical activity among overweight and obese adults across the United States: the STEP UP randomized clinical trial. JAMA Intern Med 2019 Dec 1;179(12):1624-1632. [doi: 10.1001/jamainternmed.2019.3505] [Medline: 31498375]
35. Romeo A, Edney S, Plotnikoff R, et al. Can smartphone apps increase physical activity? Systematic review and meta-analysis. J Med Internet Res 2019 Mar 19;21(3):e12053. [doi: 10.2196/12053] [Medline: 30888321]
36. Nishi SK, Kavanagh ME, Ramboanga K, et al. Effect of digital health applications with or without gamification on physical activity and cardiometabolic risk factors: a systematic review and meta-analysis of randomized controlled trials. EClinicalMedicine 2024 Oct;76:102798. [doi: 10.1016/j.eclinm.2024.102798] [Medline: 39764571]
37. Li L, Hew KF, Du J. Gamification enhances student intrinsic motivation, perceptions of autonomy and relatedness, but minimal impact on competency: a meta-analysis and systematic review. Education Tech Research Dev 2024 Apr;72(2):765-796. [doi: 10.1007/s11423-023-10337-7]

XSL•FO
RenderX

38. Patricio R, Moreira AC, Zurlo F. Gamification in innovation teams. Int J Innov Stud 2022 Sep;6(3):156-168. [doi: 10.1016/j.ijis.2022.05.003]
39. Wang T, Fan L, Zheng X, et al. The impact of gamification-induced users' feelings on the continued use of mHealth apps: a structural equation model with the self-determination theory approach. J Med Internet Res 2021 Aug 12;23(8):e24546. [doi: 10.2196/24546] [Medline: 34387550]
40. Sotos-Martínez VJ, Ferriz-Valero A, García-Martínez S, Tortosa-Martínez J. The effects of gamification on the motivation and basic psychological needs of secondary school physical education students. Phys Educ Sport Pedagogy 2024 Mar 3;29(2):160-176. [doi: 10.1080/17408989.2022.2039611]
41. Rigby CS, Ryan RM. Self-determination theory in human resource development: new directions and practical considerations. Adv Dev Hum Resour 2018 May;20(2):133-147. [doi: 10.1177/1523422318756954]
42. Yang Y, Goh KY, Teo HH, et al. Compete with me? The impact of online gamified competition on exercise behavior. JAIS 2023;24(3):912-935. [doi: 10.17705/1jais.00806]
43. Zainuddin Z, Chu SKW, Perera CJ. Gamification in education. In: Gamification in a Flipped Classroom: Pedagogical Methods and Best Practices: Springer Nature; 2024:67-113. [doi: 10.1007/978-981-97-2219-8_3]
44. Zhang Q, Yu Z. Meta-analysis on investigating and comparing the effects on learning achievement and motivation for gamification and game-based learning. Education Research International 2022 Aug 1;2022:1-19. [doi: 10.1155/2022/1519880]
45. Luong MLN, Hall M, Bennell KL, Kasza J, Harris A, Hinman RS. The impact of financial incentives on physical activity: a systematic review and meta-analysis. Am J Health Promot 2021 Feb;35(2):236-249. [doi: 10.1177/0890117120940133] [Medline: 32672050]
46. Shi P, Feng X. Motor skills and cognitive benefits in children and adolescents: relationship, mechanism and perspectives. Front Psychol 2022;13:1017825. [doi: 10.3389/fpsyg.2022.1017825] [Medline: 36478944]
47. Fanaroff AC, Patel MS, Chokshi N, et al. Effect of gamification, financial incentives, or both to increase physical activity among patients at high risk of cardiovascular events: the BE ACTIVE randomized controlled trial. Circulation 2024 May 21;149(21):1639-1649. [doi: 10.1161/CIRCULATIONAHA.124.069531] [Medline: 38583084]
48. Egashira M, Son D, Ema A. Serious game for change in behavioral intention toward lifestyle-related diseases: experimental study with structural equation modeling using the theory of planned behavior. JMIR Serious Games 2022 Feb 21;10(1):e28982. [doi: 10.2196/28982] [Medline: 35188465]
49. Zhu Y, Long Y, Wang H, Lee KP, Zhang L, Wang SJ. Digital behavior change intervention designs for habit formation: systematic review. J Med Internet Res 2024 May 24;26:e54375. [doi: 10.2196/54375] [Medline: 38787601]
50. Lakens D. Sample size justification. Collabra 2022 Mar 22;8(1):33267. [doi: 10.1525/collabra.33267]
51. Akinosun AS, Polson R, Diaz-Skeete Y, et al. Digital technology interventions for risk factor modification in patients with cardiovascular disease: systematic review and meta-analysis. JMIR Mhealth Uhealth 2021 Mar 3;9(3):e21061. [doi: 10.2196/21061] [Medline: 33656444]
52. Moshe I, Terhorst Y, Philippi P, et al. Digital interventions for the treatment of depression: a meta-analytic review. Psychol Bull 2021 Aug;147(8):749-786. [doi: 10.1037/bul0000334] [Medline: 34898233]
53. Mora-Gonzalez J, Pérez-López IJ, Esteban-Cornejo I, Delgado-Fernández M. A gamification-based intervention program that encourages physical activity improves cardiorespiratory fitness of college students: "The Matrix rEFvolution Program". Int J Environ Res Public Health 2020 Jan 30;17(3):877. [doi: 10.3390/ijerph17030877] [Medline: 32019259]
54. Chu XL, Zhao XX, Liu SY, et al. Research progress in different physical therapies for treating peripheral nerve injuries. Front Neurol 2025;16:1508604. [doi: 10.3389/fneur.2025.1508604] [Medline: 40260135]
55. Santos PDG, Vaz JR, Correia J, Neto T, Pezarat-Correia P. Long-term neurophysiological adaptations to strength training: a systematic review with cross-sectional studies. J Strength Cond Res 2023 Oct 1;37(10):2091-2105. [doi: 10.1519/JSC.0000000000004543] [Medline: 37369087]
56. Lu Y, Wiltshire HD, Baker JS, Wang Q. The effects of running compared with functional high-intensity interval training on body composition and aerobic fitness in female university students. IJERPH 2021;18(21):11312. [doi: 10.3390/ijerph182111312]
57. Dakin C, Beaulieu K, Hopkins M, Gibbons C, Finlayson G, Stubbs RJ. Do eating behavior traits predict energy intake and body mass index? A systematic review and meta-analysis. Obes Rev 2023 Jan;24(1):e13515. [doi: 10.1111/obr.13515] [Medline: 36305739]
58. Fu J, Sun S, Zhu S, et al. Relationship between 24-h activity behavior and body fat percentage in preschool children: based on compositional data and isotemporal substitution analysis. BMC Public Health 2024 Apr 16;24(1):1063. [doi: 10.1186/s12889-024-18570-2] [Medline: 38627707]
59. Jayedi A, Soltani S, Emadi A, Zargar MS, Najafi A. Aerobic exercise and weight loss in adults: a systematic review and dose-response meta-analysis. JAMA Netw Open 2024 Dec 2;7(12):e2452185. [doi: 10.1001/jamanetworkopen.2024.52185] [Medline: 39724371]
60. Shen QQ, Hou JM, Xia T, et al. Exercise promotes brain health: a systematic review of fNIRS studies. Front Psychol 2024;15:1327822. [doi: 10.3389/fpsyg.2024.1327822] [Medline: 38659667]

XSL•FO
RenderX

61. Huang TY, Chen FT, Li RH, et al. Effects of acute resistance exercise on executive function: a systematic review of the moderating role of intensity and executive function domain. Sports Med Open 2022 Dec 8;8(1):141. [doi: 10.1186/s40798-022-00527-7] [Medline: 36480075]
62. Feng X, Zhang Z, Jin T, Shi P. Effects of open and closed skill exercise interventions on executive function in typical children: a meta-analysis. BMC Psychol 2023 Nov 30;11(1):420. [doi: 10.1186/s40359-023-01317-w] [Medline: 38037184]
63. Kolovelonis A, Goudas M. Acute enhancement of executive functions through cognitively challenging physical activity games in elementary physical education. Eur Phy Educ Rev 2023 May;29(2):268-285. [doi: 10.1177/1356336X221135139]
64. Gkintoni E, Vantaraki F, Skoulidi C, Anastassopoulos P, Vantarakis A. Gamified health promotion in schools: the integration of neuropsychological aspects and CBT—a systematic review. Med Bogota Colomb 2024;60(12):2085. [doi: 10.3390/medicina60122085]
65. Cao C, Cheng S. Does competition in online games foster green behaviour in life? The relationship between gamification affordances and engagement in green behaviours. Int J Hum Comput Interact 2025;41(5):3491-3508. [doi: 10.1080/10447318.2024.2338657]
66. Singh B, Olds T, Curtis R, et al. Effectiveness of physical activity interventions for improving depression, anxiety and distress: an overview of systematic reviews. Br J Sports Med 2023 Sep;57(18):1203-1209. [doi: 10.1136/bjsports-2022-106195] [Medline: 36796860]
67. Wanjau MN, Möller H, Haigh F, et al. Physical activity and depression and anxiety disorders: a systematic review of reviews and assessment of causality. AJPM Focus 2023 Jun;2(2):100074. [doi: 10.1016/j.focus.2023.100074] [Medline: 37790636]
68. Recchia F, Bernal JDK, Fong DY, et al. Physical activity interventions to alleviate depressive symptoms in children and adolescents: a systematic review and meta-analysis. JAMA Pediatr 2023 Feb 1;177(2):132-140. [doi: 10.1001/jamapediatrics.2022.5090] [Medline: 36595284]
69. Jeong J, Pitchik HO, Fink G. Short-term, medium-term and long-term effects of early parenting interventions in low- and middle-income countries: a systematic review. BMJ Glob Health 2021 Mar;6(3):e004067. [doi: 10.1136/bmjgh-2020-004067] [Medline: 33674266]
70. Chaput JP, McHill AW, Cox RC, et al. The role of insufficient sleep and circadian misalignment in obesity. Nat Rev Endocrinol 2023 Feb;19(2):82-97. [doi: 10.1038/s41574-022-00747-7] [Medline: 36280789]
71. Almeqdad QI, Alodat AM, Alquraan MF, Mohaidat MA, Al-Makhzoomy AK. The effectiveness of universal design for learning: A systematic review of the literature and meta-analysis. Cogent Education 2023 Dec 31;10(1):2218191. [doi: 10.1080/2331186X.2023.2218191]
72. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine--reporting of subgroup analyses in clinical trials. N Engl J Med 2007 Nov 22;357(21):2189-2194. [doi: 10.1056/NEJMsr077003] [Medline: 18032770]
73. Lund JI, Toombs E, Radford A, Boles K, Mushquash C. Adverse childhood experiences and executive function difficulties in children: a systematic review. Child Abuse Negl 2020 Aug;106:104485. [doi: 10.1016/j.chiabu.2020.104485] [Medline: 32388225]
74. Damaševičius R, Maskeliūnas R, Blažauskas T. Serious games and gamification in healthcare: a meta-review. Information 2023;14(2):105. [doi: 10.3390/info14020105]
75. Treacy C, Levenstein JM, Jefferies A, et al. The LEISURE study: a longitudinal randomized controlled trial protocol for a multi-modal lifestyle intervention study to reduce dementia risk in healthy older adults. J Alzheimers Dis 2023;94(2):841-856. [doi: 10.3233/JAD-230193] [Medline: 37334601]
76. Ludyga S, Held S, Rappelt L, Donath L, Klatt S. A network meta-analysis comparing the effects of exercise and cognitive training on executive function in young and middle-aged adults. Eur J Sport Sci 2023 Jul;23(7):1415-1425. [doi: 10.1080/17461391.2022.2099765] [Medline: 35848068]
77. Gkintoni E, Vassilopoulos SP, Nikolaou G. Next-generation cognitive-behavioral therapy for depression: integrating digital tools, teletherapy, and personalization for enhanced mental health outcomes. Medicina (Kaunas) 2025 Feb 28;61(3):431. [doi: 10.3390/medicina61030431] [Medline: 40142242]
78. Victoria LW, Oberlin LE, Ilieva IP, et al. A digital intervention for cognitive deficits following COVID-19: a randomized clinical trial. Neuropsychopharmacology 2024 Dec;50(2):472-479. [doi: 10.1038/s41386-024-01995-z] [Medline: 39358543]
79. Wilson K, St-Onge MP, Tasali E. Diet composition and objectively assessed sleep quality: a narrative review. J Acad Nutr Diet 2022 Jun;122(6):1182-1195. [doi: 10.1016/j.jand.2022.01.007] [Medline: 35063665]
80. Tschon M, Contartese D, Pagani S, Borsari V, Fini M. Gender and sex are key determinants in osteoarthritis not only confounding variables. A systematic review of clinical data. J Clin Med 2021 Jul 19;10(14):3178. [doi: 10.3390/jcm10143178] [Medline: 34300344]

Abbreviations

BMR: basal metabolic rate
CG: control croup
EF: executive function
IG: intervention group

XSL•FO
RenderX

IRB: institutional review board
ITT: intention-to-treat
MCAR: missing completely at random
MET-min: metabolic equivalent-minute
mHealth: mobile health
MVPA: moderate-to-vigorous physical activity
NIH: National Institutes of Health
PA: physical activity
RCT: randomized controlled trial
SAS: Self-Rating Anxiety Scale
SDS: Self-Rating Depression Scale
SDT: self-determination theory
TPB: theory of planned behavior
WHO: World Health Organization

Edited by S Brini; submitted 21.Aug.2025; peer-reviewed by F Coelho, K Peng, W Vieira; revised version received 04.Feb.2026; accepted 06.Mar.2026; published 07.Apr.2026.

Please cite as:
Yang M, Guo Y, Li Z, Jiang L, Gao Y, Yang S, Zhou Z
A Gamified Mobile Health Intervention to Promote Physical Activity, Executive Function, and Mental Health in College Students: Randomized Controlled Trial
J Med Internet Res 2026;28:e82769
URL: https://www.jmir.org/2026/1/e82769
doi:10.2196/82769

© Ming Yang, Yingjun Guo, Zhen Li, Lei Jiang, Yanan Gao, Sen Yang, Zhixiong Zhou. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Peer Support to Increase Walking Among Older Adults: Cluster Randomized Trial

Atsushi Nakagomi[1], MD, PhD; Noriyuki Abe[1], PhD; Takayuki Ueno[2], PhD; Gemmei Izuka[1], MD, PhD; Yohei Kawasaki[3], PhD; Katsunori Kondo[1,4], MD, PhD

[1]Department of Social Preventive Medical Sciences, Center for Preventive Medical Sciences, Chiba University, 1-33 Yayoicho, Inage-ku, Chiba, Japan

[2]Research Team for Social Participation and Healthy Aging, Tokyo Metropolitan Institute for Geriatrics and Gerontology, Tokyo, Japan

[3]Department of Biostatistics, Graduate School of Medicine, Saitama Medical University, Saitama, Japan

[4]Research Department, Institute for Health Economics and Policy, Tokyo, Japan

Corresponding Author:
Atsushi Nakagomi, MD, PhD
Department of Social Preventive Medical Sciences, Center for Preventive Medical Sciences, Chiba University, 1-33 Yayoicho, Inage-ku, Chiba, Japan

Abstract

Background: As the population ages, older adults face an increasing risk of physical inactivity and related health complications, highlighting the need for scalable interventions. Smartphone-based programs have emerged as a promising strategy to support sustained physical activity among older adults.

Objective: This study aimed to evaluate whether a smartphone lecture program incorporating a digital peer support app would increase physical activity among older adults, compared to a conventional smartphone lecture program.

Methods: This 2-arm, 1:1 parallel-arm, cluster-randomized trial was conducted in 2 urban regions of Japan (Sumida Ward, Tokyo, and Chiba City, Chiba). Eligible participants were community-dwelling adults aged ≥60 years, able to walk independently, and smartphone users; exclusion criteria included prior use of the peer support app or medical restrictions on walking. Participants were recruited offline during community smartphone lectures (closed-group recruitment). The intervention combined face-to-face lectures with app-based peer support, while outcomes were assessed both objectively (via smartphones) and through self-administered paper questionnaires. All participants received a baseline smartphone lecture. Intervention participants attended 2 additional sessions using a digital peer support app (Minchalle; A10 Lab Inc), which included features such as daily step goals, peer sharing, and group encouragement. Control participants attended 2 standard follow-up smartphone lectures. The primary outcome was the change in weekly average daily step count from baseline to Week 12. Secondary outcomes included total metabolic equivalent of task (MET)–minutes per week (assessed via the International Physical Activity Questionnaire), walking time (≥30 minutes per day), daily smartphone use, and number of smartphone use purposes.

Results: A total of 156 community-dwelling older adults were grouped into 40 clusters and randomized (20 intervention clusters, n=80 and 20 control clusters, n=76). In total, 124 participants (79.5%) completed the follow-up, and valid step data were available for 117 participants, with missing data ranging from 5.1% to 29.1%. Baseline daily steps averaged 3951 (SD 1686) in controls versus 4583 (SD 1973) in the intervention arm. An unadjusted mixed model for repeated measures showed significantly higher step changes for intervention participants at Week 12 (difference=579, 95% CI 36-1123; *P*=.04). No significant differences emerged for total METs (difference=646 MET-min per week, 95% CI –12 to 1303; *P*=.054) or walking ≥30 minutes per day (odds ratio [OR] 1.56, 95% CI 0.63-3.90; *P*=.33). However, the intervention arm demonstrated a significant increase in daily smartphone use (OR 4.10, 95% CI 1.15-14.6; *P*=.03) and in the number of smartphone use purposes (difference=0.58, 95% CI 0.12-1.05; *P*=.01).

Conclusions: A smartphone lecture program integrated with app-based peer support led to modest but meaningful improvements in step counts among older Japanese adults, at Week 12 of the 12-week intervention. Future research should investigate long-term maintenance, additional measures of physical activity, and subpopulation responses to optimize digital health programs for older adults.

Trial Registration: UMIN-CTR UMIN000051904; https://tinyurl.com/4m4zm99v

(J Med Internet Res 2026;28:e75708) doi:10.2196/75708

KEYWORDS

peer support; smartphone; digital health; physical activity; older adults; digital divide

XSL•FO
RenderX

Introduction

Physical activity is a key factor in healthy aging, helping prevent chronic diseases, maintain functional independence, and improve quality of life [1,2]. In Japan, where a rapidly aging population has raised concerns about inactivity and health care costs, the Ministry of Health, Labor, and Welfare (MHLW) has taken steps to address the issue. As part of the Healthy Japan 21 (3rd term) initiative, the MHLW set a target of 6000 steps per day for older adults by 2032, based on 2019 data showing averages of 5396 steps for men and 4656 for women [3]. The MHLW also released the Physical Activity and Exercise Guide for Health Promotion 2023 [4], emphasizing that even small increases in light-intensity activity yield significant health benefits, particularly for sedentary individuals [5]. As Japan's older population continues to grow, scalable strategies to promote physical activity are an urgent public health need.

Evidence suggests behavior change strategies can promote sustained physical activity among older adults [6,7]. Intrapersonal strategies—such as action planning, goal setting, commitment, and self-monitoring—are commonly used in physical activity interventions [8,9]. Interpersonal strategies, especially peer support, are also gaining traction [10-12]. Social cognitive theory (SCT) provides a strong theoretical basis for these strategies, as it emphasizes the dynamic interaction between individual, behavioral, and social-environmental factors [13]. Key constructs of SCT include self-efficacy (confidence in one's ability to maintain behavior), observational learning (acquiring new behaviors by observing others), and reinforcement (sustaining behaviors through social and environmental feedback). By drawing attention to how individual agency and social context interact, SCT offers a useful framework for understanding how both intrapersonal and interpersonal strategies can promote lasting engagement in physical activity among older adults.

With the growth of digital health technologies, many mobile apps now integrate behavior change techniques [14]. Alongside intrapersonal features, peer-support elements such as group challenges and shared goals are increasingly common [15-17]. For example, peer support can be structured in a smartphone app (Minchalle; A10 Lab Inc) through small groups of 3-5 participants who collaboratively select step goals, commit to posting updates at regular intervals, and encourage each other's progress [17,18]. Integrating such peer support with intrapersonal strategies within the app has been shown to enhance user engagement and adherence to health goals [17,18].

However, older adults often face barriers—known as the "digital divide"—that hinder their use of technology-based interventions [19]. In Japan, local governments and community groups have introduced smartphone lectures to bridge this gap, yet sustained smartphone use remains a challenge. Digital peer support may offer a more lasting solution by providing ongoing group-based reinforcement, promoting both technology use and healthier behaviors. Still, few studies have compared smartphone lectures with and without digital peer support, highlighting a gap in empirical evidence.

This study evaluates whether adding digital peer support to smartphone lectures increases physical activity among older Japanese adults. Using a cluster-randomized controlled trial, we compared a conventional smartphone lecture to one combined with a peer support app grounded in behavioral science. A cluster design was required because the intervention was implemented within peer groups of 3-5 participants; randomizing individuals within the same group would have compromised the integrity of the peer support mechanism and risked contamination between trial arms. The primary objective of the trial was to evaluate changes in daily step counts at the individual participant level, while randomization occurred at the cluster (group) level to preserve the integrity of the peer support intervention. Participants were recruited in community settings and randomly assigned to either group. This design allowed us to examine whether ongoing digital peer support enhances motivation, sustains smartphone engagement, and improves physical activity outcomes.

Methods

Overview

The study was conducted in 2 urban municipalities within the Tokyo metropolitan region, Sumida Ward (Tokyo) and Chiba City (Chiba Prefecture). The trial was first implemented in Sumida Ward from September to December 2023 in collaboration with the municipal government and subsequently replicated in Chiba City from June to September 2024 through collaborations with local senior clubs. Although implemented in different municipalities, both settings are urban communities located near central Tokyo, allowing the study procedures to be applied consistently. This manuscript follows the 2010 CONSORT (Consolidated Standards of Reporting Trials) guidelines for cluster randomized trials (see Table S1 in Multimedia Appendix 1) [20] and the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidelines (Checklist 1) [21].

Design

This was a 2-arm, 1:1 parallel-arm, cluster-randomized trial that included a 2-week baseline measurement period followed by a 12-week intervention period. The unit of randomization (the cluster) was a peer support group consisting of 3-5 participants. In the intervention arm, these clusters functioned as peer groups within the Minchalle app, while in the control arm, participants were grouped to maintain comparability but did not engage in peer support during the trial. Control participants were informed that they would be offered the opportunity to form peer support groups after the trial (waiting-list design). Randomization at the cluster level was required to preserve the integrity of arm allocation.

Participants and Recruitment

Recruitment took place from September 2023 to June 2024, in partnership with the Sumida Ward local government and senior clubs in Chiba City. Individuals were recruited in community venues during smartphone lectures hosted by the local municipal government or senior clubs, which aimed to bridge the digital

XSL•FO
RenderX

divide and promote physical activity to prevent frailty among older adults. The local government and senior clubs outsourced lecture services to A10 Lab Inc., developer of the "Minchalle" app (Challenge with All), and Active Senior Outreach Nonprofit Organization, a nonprofit organization specializing in smartphone lectures for older adults.

These smartphone lectures were implemented as part of the annual programs conducted by the municipalities and senior clubs and were held in public community venues, typically accommodating 15 - 30 older adults per session. Participants were approached in person at the beginning of the first session. Researchers informed all attendees about the study using slides and printed information sheets. Attendees were eligible to participate regardless of prior interest, and every individual attending the session was approached, ensuring a consistent and nonselective recruitment process.

Participants were eligible if they (1) were ≥60 years old, (2) could communicate in Japanese, (3) could read and provide informed consent, and (4) owned a smartphone at the time of consent. Participants were excluded if they (1) had any condition making participation infeasible (eg, inability to walk with a smartphone), (2) had a medical restriction on walking (eg, per doctor's advice), or (3) had already used the Minchalle app.

Interested individuals were then screened by the research team based on the exclusion criteria. Those who met all criteria received a detailed explanation of the study procedures and provided written informed consent on-site. No participant was recruited through individual referrals, advertisements, or convenience sampling outside of the lecture setting.

Clusters were formed by grouping 3-5 participants who attended the same lecture session. No eligibility criteria were applied at the cluster level, and only individuals who provided informed consent were included. Individuals who did not meet the inclusion criteria or declined to participate were allowed to attend the lecture activities but were not included in the study cohort.

Initial Session and Assessment

Figure 1 outlines the study procedures. Consenting participants were organized into groups of 3 - 5 and attended a smartphone lecture, followed by completion of baseline questionnaires. The lecture included a 15-minute presentation by a medical doctor on the health benefits of walking and a one-hour training session on installing and using health-tracking apps—Google Fit (Google LLC; Android) and Apple Health app (Apple Inc; iOS). Participants were encouraged to use basic app features, such as step counting and daily goal setting, incorporating intrapersonal behavior change strategies (action planning, goal setting, commitment, and self-monitoring). Participants were instructed to carry their smartphones daily to ensure continuous step count recording.

Figure 1. Timeline of study procedures.

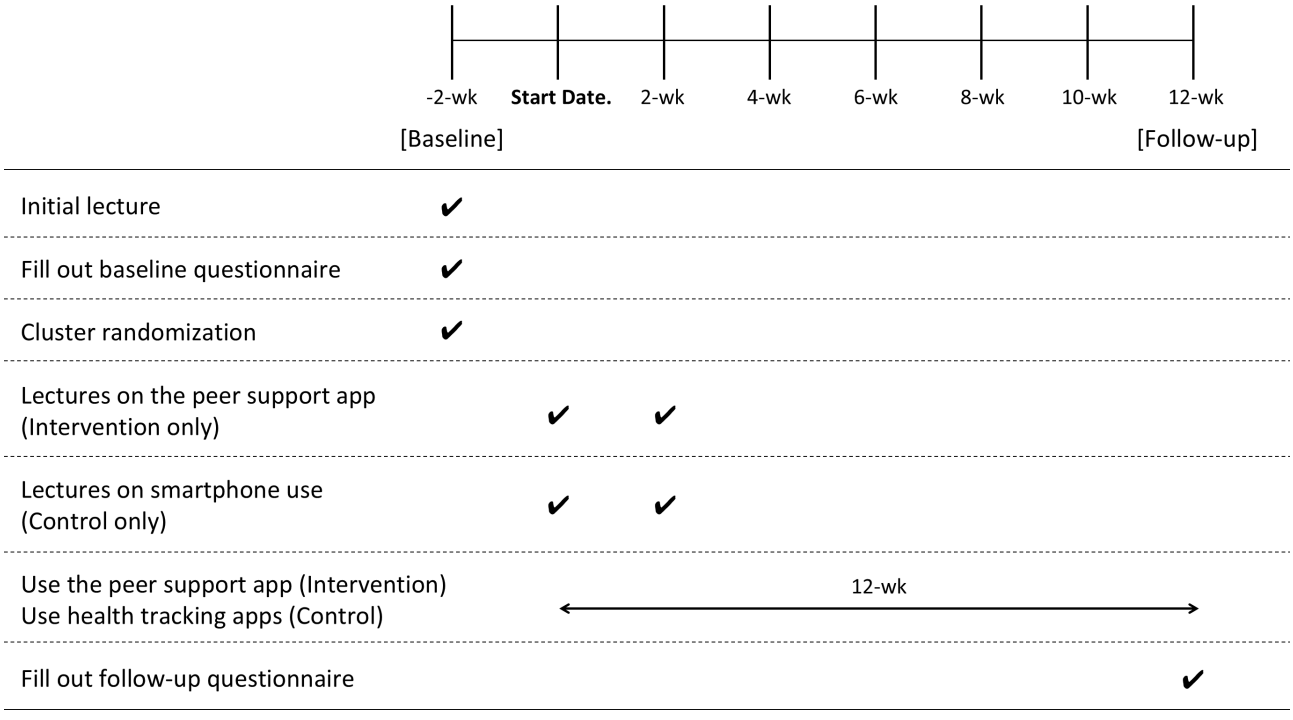

Before leaving the session, research staff verified successful installation of the app and confirmed that step-count tracking was active. Baseline paper questionnaires were completed on-site and collected immediately by trained staff. Baseline questionnaire data were linked to each participant using a deidentified study ID.

Follow-Up Data Collection

At the 12-week follow-up session, participants completed follow-up paper questionnaires. During this session, research staff guided participants through the in-app procedures required to export and send their accumulated step count data to the Minchalle server using each participant's deidentified study ID. Only participants who attended the follow-up session and completed these in-app actions were able to transmit their step

XSL•FO
RenderX

data. The app provider then securely transferred the deidentified step-count dataset to the research team, where it was merged with questionnaire responses using the study ID. No personally identifiable information was shared, and all data were stored on secure, access-restricted servers.

Randomization

Following the initial session and baseline assessment, randomization was conducted at the cluster level, with each group of 3-5 participants serving as the unit of allocation. The random allocation sequence was generated electronically using a random number generator with block sizes of 2-4 clusters. An independent allocation officer from the Institute for Assistance of Academic and Education implemented the sequence. The allocation sequence and block size were concealed from study personnel during enrollment, and arm assignments were revealed only after enrollment was finalized.

Individual participants were enrolled during community smartphone lectures and grouped pragmatically into clusters of 3-5 based on attendance at the same session. Randomization was therefore performed at the cluster level, while allocation concealment and consent were handled at the individual participant level.

Control arm participants were informed they would receive the intervention content after the trial (ie, a waiting-list control design).

Blinding

Because of the nature of the intervention, participants and local staff could not be blinded to arm assignment. However, outcome assessment was based on step count data automatically recorded by participants' smartphones, and data analysts were blinded to allocation, reducing the risk of bias.

The 2 study arms differed in that only the intervention arm received access to the Minchalle app and related training. Therefore, the interventions could not be made similar in appearance or content.

Intervention

The intervention was a smartphone lecture program incorporating the digital peer support app Minchalle. All smartphone lecture programs were conducted in community venues in Sumida Ward (Tokyo) and Chiba City (Chiba Prefecture), which are typical venues for older adult smartphone lectures. The app, developed by A10 Lab Inc. and launched in 2015, was provided free of charge. Illustrations of the app are shown in Figure S1 in Multimedia Appendix 1. A structured description of the intervention is also provided according to the Template for Intervention Description and Replication (TIDieR) checklist (see Table S2 in Multimedia Appendix 1).

Minchalle integrates several behavior change techniques (BCTs) as defined by the BCT Taxonomy v1 [22]. The core element of Minchalle was the peer support component, in which groups of 3-5 members supported one another to achieve their goals by posting at least one photo and one message daily in the group chat (social support: general [BCT 3.1]). In addition, groups of 3-5 members selected daily step goals of 2500, 5000, or 10,000 steps (goal setting [BCT 1.1] and action planning [BCT 1.4]). Groups also selected a posting interval (every 4, 8, or 15 days). Members who failed to post within the agreed interval were removed (with the option to rejoin), reinforcing commitment (BCT 1.9: commitment). Step counts were automatically tracked and displayed in the chat, enabling self-monitoring of behavior (BCT 2.3) and promoting feedback and social comparison (BCT 6.2, 6.3) through the visibility of group achievements. This design was grounded in SCT, where peer encouragement, observational learning, and reinforcement are expected to enhance self-efficacy and sustain physical activity.

Two weeks after the initial baseline session, participants in the intervention arm attended 2 additional weekly lectures led by A10 Lab Inc, the developer of Minchalle. In these lectures, participants were instructed on how to install the app and configure initial settings, including selecting group step goals and defining posting intervals, after which members who failed to post within the interval would be removed. They also learned how to take photos, type messages, and post them in the group chat, supported by manuals provided during the sessions.

The intervention was delivered primarily at the cluster level (peer support groups of 3-5 participants), but individual participants used the app within their assigned peer support groups.

Control

Participants in the control arm also attended 2 smartphone lectures, akin to typical sessions for older adults in Japan. These lectures covered basic smartphone functions, including email, web browsing, use of the messenger app LINE, and official local government apps or accounts. As in the initial lecture, they were encouraged to continue using Google Fit or Apple Health for step tracking. The final assessment for the control arm was conducted 3 months later, simultaneously with the intervention arm but in different locations.

Outcomes

The primary outcome was the change in the weekly average of daily step counts from baseline to Week 12 (3 months after installation of the peer support app). After the baseline assessment, participants were asked to carry their smartphones at all times to capture daily steps. During the 2-week baseline period leading up to the second lecture, baseline step counts were measured. To minimize mismeasurement, any day with fewer than 1000 steps was treated as missing [23]. No changes to the prespecified primary outcome were made after trial commencement. All primary outcome data pertained to the individual participant level, although randomization was conducted at the cluster level.

Secondary outcomes included weekly step counts at other intervention weeks (Weeks 1 - 11) and monthly averages (Months 1 - 3), total physical activity (metabolic equivalent of task [MET] assessed by the International Physical Activity Questionnaire [IPAQ] short form) [24], self-reported walking time (≥30 minutes per day), daily smartphone use, and the number of smartphone use purposes at follow-up. The IPAQ short form was administered at baseline and follow-up to assess frequency and duration of vigorous, moderate, and walking

XSL•FO
RenderX

activities (≥10 min bouts). Walking time was dichotomized as <30 or ≥30 minutes per day. Smartphone use was dichotomized as daily (almost daily) versus not daily. Smartphone use purposes were assessed from a checklist and summed to yield a total score from information search and browsing, work-related tasks, communication (eg, email, LINE, or video calls), taking photos or videos, using maps or traffic information, online banking, shopping, social media (eg, Facebook [Meta], Twitter [subsequently rebranded X; X Corp]), health apps (eg, step counters), entertainment (eg, video streaming, e-books, or music), gaming, and other uses.

Statistical Analysis

The sample size was calculated assuming a 1-tailed test based on preliminary usage data with the same app; however, all subsequent analyses were conducted using 2-tailed tests at the 0.05 significance level. From pilot data provided by the app developer, we anticipated a medium effect size (Cohen d≈0.52) for changes in step counts. With α=.05, an intraclass correlation coefficient of 0.05 to account for clustering (mean cluster size of 4), and an expected attrition rate of 30%, we estimated that a total of 200 participants (100 per arm) would provide 80% power to detect a significant effect.

Baseline characteristics were summarized by treatment arm. Continuous variables were reported as mean (SD) and compared via 2-tailed 2 - sample t tests, while categorical variables were shown as frequencies (percentages) and analyzed via chi - square or Fisher exact tests, using the *CreateTableOne* package in R (R Core Team).

Days with <1000 steps and other missing data were imputed; based on evidence, these often reflect incomplete capture [23,25]. Multiple imputation was performed using the package in R, incorporating baseline covariates and primary and secondary outcomes. Ten imputed datasets were generated, and estimates were combined using Rubin rules.

Separate imputation models were created for step count data (n=117) and questionnaire-based outcomes (n=124).

The primary analysis estimated the between-arm difference in the preregistered primary outcome—the change in weekly average daily step count from baseline to Week 12—using an unadjusted mixed model for repeated measures (MMRM). This model included treatment arm, week (categorical: baseline and Weeks 1 - 12), and their interaction, with an unstructured covariance matrix at the participant level and cluster-robust standard errors to account for clustering of participants within groups.

Several sensitivity analyses were conducted to examine the robustness of the findings. These included (1) adjusted MMRM, (2) unadjusted, and (3) adjusted linear mixed-effects models with random intercepts at the participant and group levels, (4) unadjusted and (5) adjusted ANCOVA models with the change from baseline to Week 12 as the dependent variable, and (6) unadjusted MMRM using unimputed data, excluding days with missing or <1000 steps. For adjusted models, baseline covariates were selected a priori based on theoretical relevance and prior literature. Operating system (Android or iOS), study field (Sumida or Chiba), gender (men or women), education (high school or less vs college or more), marital status (married or unmarried), baseline smartphone use (daily vs not daily), and walking ≥30 minutes per day were treated as categorical variables. Age, self-rated health (0 - 10 scale), and total MET-minutes per week were treated as continuous variables and modeled linearly.

Analyses of weekly step counts for Weeks 1 - 11 and monthly averages (Months 1 - 3) were conducted as secondary outcomes using the same unadjusted MMRM framework as the primary analysis. These analyses were not adjusted for multiple comparisons and were not part of the primary estimand, but are included to aid interpretation of temporal patterns.

Secondary outcomes obtained from the Week-12 follow-up questionnaire were analyzed using unadjusted models to maintain consistency with the primary analytic approach. Total physical activity (MET-minutes per week) and the number of smartphone use purposes were examined using ANCOVA, adjusting only for the baseline value of each outcome. Walking ≥30 minutes per day and daily smartphone use were analyzed using logistic regression with the same minimal adjustment.

Because a baseline age imbalance between arms was observed, we conducted an exploratory subgroup analysis stratified at the median age of 80 years. For each age stratum, the primary outcome was re-estimated using the unadjusted MMRM model described above. An interaction model including treatment × week × age (binary: ≥80 vs <80 years) was further examined to assess whether age modified the intervention effect.

Ethical Considerations

This study received ethical approval from the Chiba University Ethical Committee (protocol number: M10626) and was registered with the University Hospital Medical Information Network (UMIN000051904). Written informed consent was obtained from all participants before randomization. All procedures adhered to the principles of the Declaration of Helsinki. Participant privacy and confidentiality were protected by deidentifying all data prior to analysis. Data were stored securely on password-protected servers accessible only to authorized study personnel. Only aggregated results are reported in this manuscript. Participants did not receive monetary compensation; however, they were allowed to use the smartphone app provided as part of the intervention, and no additional incentives were offered.

Results

Overview

Recruitment ended when the smartphone lecture series provided by the municipality and senior clubs—our sole recruitment venue—was completed, making additional enrollment infeasible. At that point, 177 participants had been enrolled. The analysis was conducted at this stage because recruitment had necessarily concluded, although the final sample size did not reach the originally planned statistical power.

At the analysis, 40 groups comprising 156 participants were allocated either to an intervention arm (n=80, 20 groups) or a control arm (n=76, 20 groups; see Figure 2). Of the 156 enrolled

XSL•FO
RenderX

participants, 124 (79.5%) completed the 3-month follow-up evaluation (64 in the intervention arm and 60 in the control arm), with step-count data successfully linked for 117 participants.

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram for study clusters and participants.

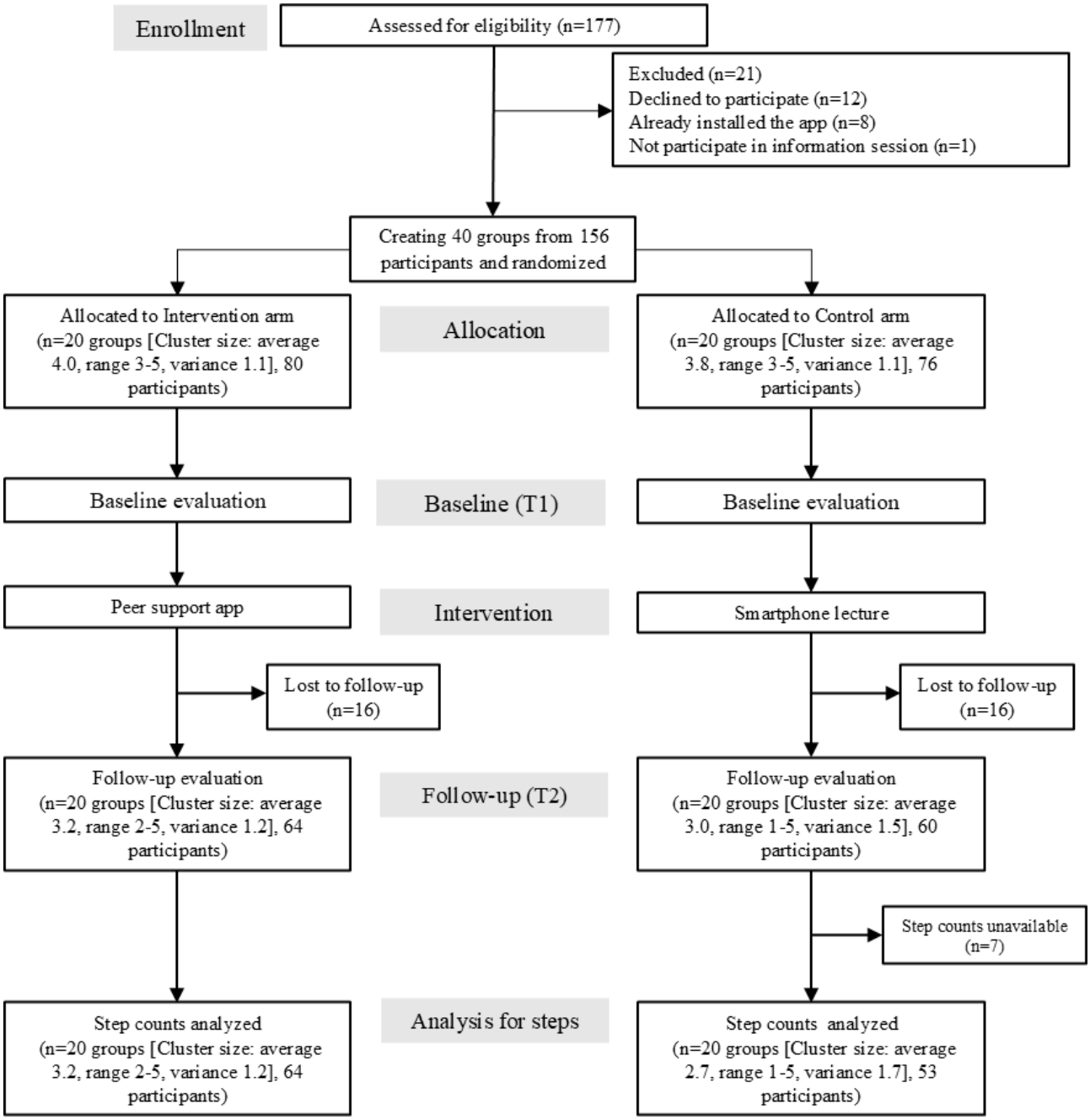

Missing step data ranged from 5.1% to 29.1%. As shown in Table 1, baseline characteristics were generally well balanced between arms, except for a significant difference in mean age (P=.009). Mean (SD) baseline daily steps were 3951 (1686) for the control arm and 4583 (1976) for the intervention arm.

XSL•FO
RenderX

Table . Characteristics of study participants.

Variables	Control (n=53)	Intervention (n=64)
Age (year), mean (SD)	81.2 (4.7)	79.0 (4.5)
Gender (Men), n (%)	17 (32.1)	18 (28.1)
OS (iOS), n (%)	9 (17.0)	18 (28.1)
Daily use of smartphone, n (%)	41 (77.4)	55 (85.9)
Education (college or more), n (%)	25 (47.2)	20 (31.7)
Married, n (%)	26 (49.1)	31 (48.4)
Self-rated health, mean (SD)	6.8 (2.0)	6.8 (1.9)
Total MET[a], mean (SD)	3274 (1850)	3372 (2373)
Walking more than 30 minutes, n (%)	37 (69.8)	43 (67.2)
Baseline weekly average steps, mean (SD)	3951 (1686)	4583 (1976)

[a]MET: metabolic equivalent of task.

Primary Outcome

The unadjusted differences in mean daily steps by week and study arm are illustrated in Figure 3. Across most time points, the intervention arm recorded higher step counts than the control.

Figure 3. Unadjusted differences in mean daily steps by week and study arm.

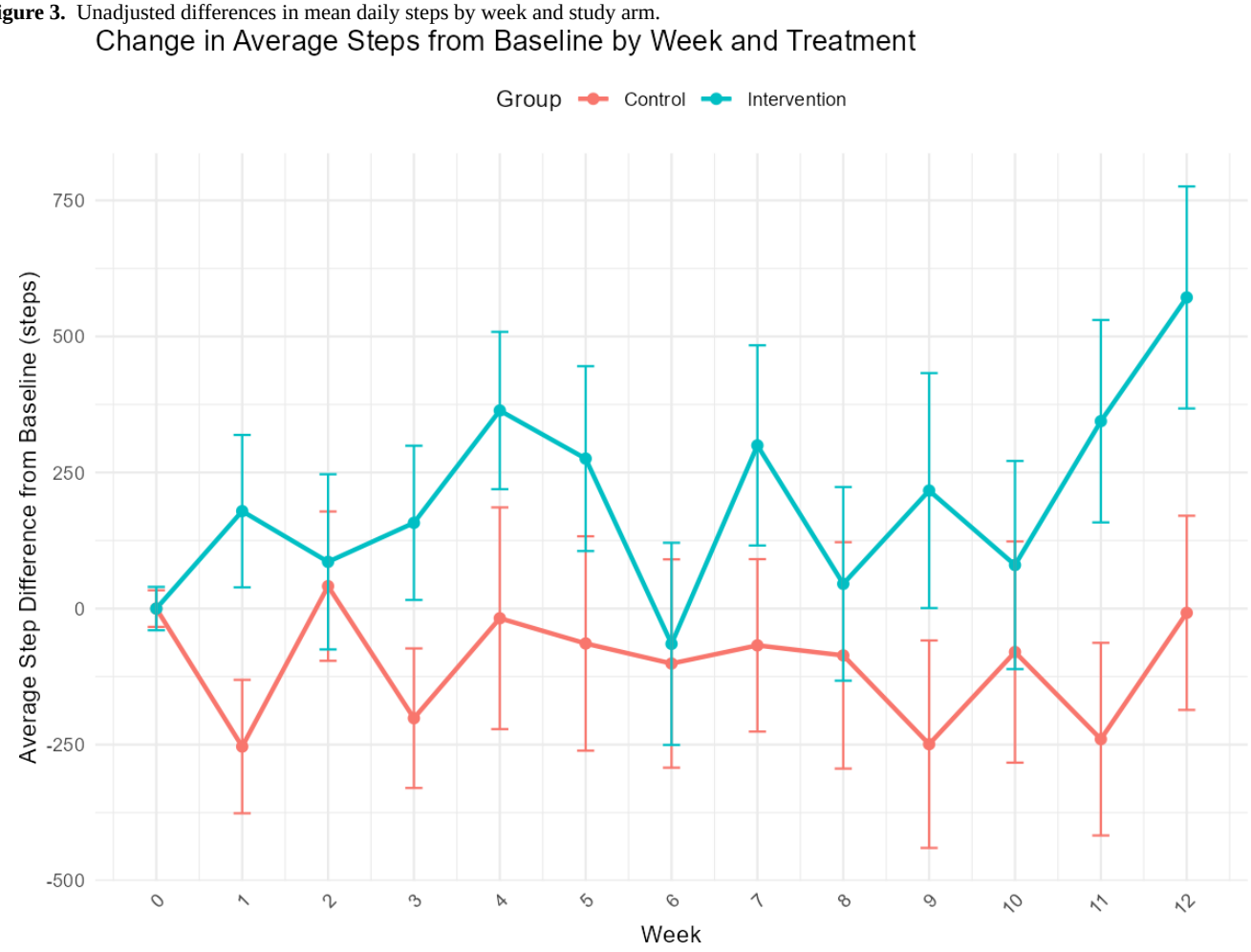

The primary outcome was the change in average daily steps at week 12 compared with baseline. In the unadjusted MMRM model, the intervention arm recorded significantly higher steps than the control arm at Week 12 (difference=579 steps, 95% CI 36-1123**;***P*=.04) (Table 2). The estimated cluster-level variance for the primary outcome was essentially zero (intraclass correlation coefficient ≈ 0.00), indicating minimal between-cluster variability after adjustment.

XSL•FO
RenderX

Table . Estimated between-arm differences in step counts by week (intervention minus control), unadjusted mixed-effects model for repeated measures.

Variables	Estimate[a], β (95% CI)	*P* value
Intervention × Week 1[b]	433 (74 to 791)	.02
Intervention × Week 2	45 (−385 to 474)	.84
Intervention × Week 3	359 (−26 to 744)	.07
Intervention × Week 4	382 (−180 to 943)	.18
Intervention × Week 5	340 (−273 to 953)	.28
Intervention × Week 6	36 (−564 to 637)	.91
Intervention × Week 7	367 (−173 to 908)	.18
Intervention × Week 8	131 (−420 to 683)	.64
Intervention × Week 9	466 (−222 to 1154)	.18
Intervention × Week 10	160 (−434 to 754)	.60
Intervention × Week 11	584 (17 to 1151)	.04
Intervention × Week 12	579 (36 to 1123)	.04

[a]Estimates represent unadjusted between-arm differences at each week, derived from arm×week interaction terms.

[b]Intervention×week=product terms in the model.

The overall pattern remained consistent with the main analysis in sensitivity analysis: adjusted MMRM model (difference=539, 95% CI 92-985, *P*=.02); unadjusted mixed effects model (difference=589, 95% CI 110-1069; *P*=.02); adjusted mixed effects model (difference=526, 95% CI 65-987; *P*=.03); unadjusted ANCOVA model (difference=589, 95% CI 76-1101; *P*=.02); adjusted ANCOVA model (difference=489, 95% CI 33-946; *P*=.04); unadjusted MMRM model with complete case analysis (difference=1295, 95% CI 107-2483; *P*=.03).

Secondary Outcomes

Exploratory analyses of weekly step counts (Weeks 1 - 11) showed that the intervention arm had higher steps at Week 1 and Week 11 (Week 1: difference=433, 95% CI 74-791; *P*=.02; Week 11: difference=584, 95% CI 17-1151; *P*=.04, respectively) (Table 2). In a further exploratory analysis aggregating daily steps by month, the intervention arm showed greater, but not significant, differences at each month (Table S3 in Multimedia Appendix 1).

No significant arm differences were observed for physical activity measured by total METs or for self-reported walking time (≥30 minutes per day). The unadjusted difference in total METs was 646 MET-minutes per week (95% CI −12 to 1303; *P*=.054), and the odds ratio (OR) for walking ≥30 minutes per day was 1.56 (95% CI 0.63-3.90; *P*=.33; Table S4 in Multimedia Appendix 1). However, there was a suggestive trend toward an increase in daily smartphone use (OR 4.10, 95% CI 1.15-14.6; *P*=.03) and the number of smartphone use purposes (mean difference 0.58, 95% CI 0.12-1.05; *P*=.01) among participants in the intervention arm.

Exploratory Post-Hoc Analyses

In exploratory subgroup analyses stratified by the median age of 80 years (younger: n=57; older: n=58), baseline age was well balanced within subgroups (younger: mean 76.2 vs 76.3 years; older: mean 84.2 vs 83.1 years). Intervention effects varied by age. Among older participants (≥80 years; n=58), significant treatment effects were observed, with between-arm differences exceeding 1000 steps at Week 12. In contrast, younger participants (<80 years; n=57) showed inconsistent and occasionally negative effects, with no sustained significant benefits. Consistent with these findings, the treatment × week × age interaction model indicated that intervention effects became stronger among participants aged 80 years or older in multiple weeks. Detailed estimates from the subgroup and interaction analyses are provided in Table S5 in Multimedia Appendix 1.

Discussion

Short Summary

This cluster-randomized controlled trial demonstrates that a smartphone lecture program, combined with a mobile app–based peer support component, may produce modest but meaningful increases in step counts among older adults in Japan compared to a conventional smartphone lecture program. The intervention arm recorded a statistically significant increase in average daily steps at Week 12 (the preregistered primary end point), with supportive differences observed at Weeks 1 and 11. All other weekly comparisons and monthly averages were exploratory and should be interpreted with caution, as no adjustment for multiple testing was applied. In addition, smartphone use and purposes were greater in the intervention arm. These results suggest that integrating peer support features into smartphone-based health initiatives could be a promising strategy for sustaining engagement and encouraging physical activity and smartphone use over time.

Principal Findings

The peer support app appeared to stimulate progressive increases in step counts, particularly noticeable during the latter weeks. An important question is why intervention effects became most apparent in the latter part of the program (Weeks 11 - 12). Novelty effects typically lead to high engagement at the outset

XSL•FO
RenderX

of an intervention but tend to diminish over time, making it unlikely that novelty alone explains the late increase in steps. A more plausible explanation is that repeated goal setting, self-monitoring, and peer reinforcement gradually strengthened self-efficacy and supported the formation of walking habits, consistent with SCT. In this framework, the app's design may have enhanced self-efficacy (through repeated goal attainment and encouragement), promoted observational learning (by making peers' behaviors visible), and provided reinforcement (through accountability and supportive feedback).

In addition, seasonal and contextual factors may have contributed. The study was conducted in Sumida Ward, Tokyo, from September to December 2023. December includes cultural and social events that may have motivated participants to walk more and share photos with peers in the app. Also, in Chiba City, these were held from June to September 2024. Older adults might avoid walking outside in hot weather in July and August, while cooler weather in September may have encouraged outdoor activity. These contextual influences, alongside habit formation, may help explain why intervention effects were observed more strongly toward the end of the study period.

Comparison to Prior Work

Our findings align with and extend prior evidence that interpersonal and peer-support components are particularly important for promoting and sustaining physical activity among older adults. McMahon et al [12] conducted a factorial randomized trial comparing intrapersonal and interpersonal behavior change strategies among community-dwelling adults aged ≥70 years and demonstrated that interventions including interpersonal components produced greater improvements in objectively measured physical activity, whereas intrapersonal components alone were less effective. Our results are consistent with this framework, while extending it to a pragmatic community education setting by embedding peer support within routinely delivered smartphone lectures rather than a research-intensive intervention.

Importantly, Tabira et al [17] evaluated the same digital peer-support smartphone app (Minchalle) used in our study in a nonrandomized controlled design among adults aged ≥65 years. They reported high feasibility, retention, and increases in physical activity, supporting the app's acceptability and potential effectiveness in older populations. However, their study focused on feasibility and short-term behavioral change under a nonrandomized design, leaving uncertainty regarding causal effects. By contrast, our randomized controlled trial using the same app extends this prior work by strengthening causal inference and demonstrating effectiveness when Minchalle is implemented as part of existing smartphone lecture programs that are already widely disseminated across Japan.

In addition, Patel et al [18] showed in a large randomized trial that socially structured incentives embedded in behaviorally designed gamification can increase step counts, although the magnitude and persistence of effects varied by incentive type. While conducted primarily among working-age adults in a remote, workplace-adjacent context, their findings reinforce the broader principle that social and peer-based design features meaningfully influence physical activity behavior. Our study complements this literature by demonstrating that similar social mechanisms can be effectively operationalized in older adults using commercially available apps without requiring specialized devices or intensive program infrastructure.

Exploratory Post-Hoc Analyses

This study suggests that older participants (≥80 years) may derive greater benefit from app-based peer support, a group often most vulnerable to the digital divide [26]. By facilitating engagement with smartphones through peer support, our approach may help promote physical activity while also addressing inequities in digital participation. These findings are exploratory and not prespecified, and given the limited sample size, should be interpreted with caution.

Secondary Findings

Despite the significant gains in step counts, we observed no significant arm differences in total METs or self-reported walking time (≥30 minutes per day). These discrepancies may stem from the inherently coarse nature of subjective physical activity measures, which can miss nuances that more objective step-tracking captures [12,27]. Indeed, correlations between self-report and objectively measured physical activity in older adults are positive, yet often moderate at best [28]. Future studies integrating more comprehensive, sensor-based activity assessments will help clarify whether peer support can robustly influence other domains of physical activity beyond daily step counts.

There was also a suggestive trend toward increased daily smartphone use in the intervention arm, implying that digital peer support may facilitate broader technology adoption. This is consistent with the primary aim of smartphone lectures for older adults and suggests that digital peer support apps can effectively enhance participants' confidence and willingness to explore additional smartphone features.

Limitations

Several limitations merit consideration. First, the final sample size was smaller than planned, which reduced statistical power, particularly for secondary outcomes. Attrition (≈20%) further reduced the analytic sample, mainly due to nonattendance at the 3-month follow-up lecture, and step-count data from 7 participants were lost because of a technical linkage error. Second, missing data and imputation procedures, while consistent with prior interventions, could introduce bias if data were not missing at random. Third, digital peer support may not uniformly appeal to all participants, potentially limiting generalizability. Future studies should examine heterogeneous responses to peer support for physical activity and smartphone adoption. Fourth, although we reported step count differences at multiple weekly and monthly time points, only the between-arm difference at Week 12 was prespecified as the confirmatory primary end point. All other weekly and monthly comparisons were exploratory and not adjusted for multiplicity; therefore, these findings should be interpreted with caution. Fifth, no formal arm sequential testing procedure (eg, alpha-spending function) was applied to adjust for multiple looks at the data. Recruitment was concluded after 177 participants had been enrolled, based on the completion of the

XSL•FO
RenderX

smartphone lectures organized by the local municipal government and senior clubs. The absence of a formal stopping rule represents a limitation with respect to strict error rate control. Sixth, SCT constructs were not directly measured in this trial, and the hypothesized mechanisms (eg, self-efficacy, observational learning, and reinforcement) remain speculative. Seventh, no process measures such as outcome expectancies or social support perceptions were collected, and detailed app usage metrics (eg, posting volume, response rates, and network analyses) were not available, limiting the ability to fully evaluate the peer support component. Eighth, baseline age differed modestly between arms (79.0 years in intervention vs 81.2 years in control). However, exploratory analyses suggested that older participants benefited more from the intervention. Therefore, this imbalance would likely have biased estimates toward underestimation of the true effect rather than exaggeration. These age-stratified analyses were exploratory, not prespecified, and should be interpreted with caution given the limited sample size. Ninth, although the Minchalle app interface and features were identical on Android and iOS, potential platform-specific differences in step data capture (eg, Google Fit vs Apple Health) cannot be excluded. Due to the small number of iOS users (n=27), subgroup analyses by operating system were not feasible, and future studies should address this more explicitly. Tenth, individual-level attrition resulted in smaller cluster sizes at follow-up, particularly in the control arm, where some clusters ultimately had only one participant contributing data. Participants who did not attend the follow-up session were unable to provide questionnaire or step-count data, and in the intervention arm, we were also unable to monitor their continued use of the peer-support app. As a result, it remains uncertain whether any intervention clusters functionally shrank to very small active groups during the trial period. This uncertainty may influence the interpretation of peer-related effects, as reduced group size could weaken the social facilitation or mutual-accountability mechanisms central to the intervention. Finally, the external validity of our findings should be interpreted with caution. This trial was conducted in 2 urban areas of Japan, where walkable environments and access to digital infrastructure are relatively well developed. The results may not directly translate to rural areas with limited walkability or to contexts with lower smartphone and internet penetration. In addition, cultural differences in peer support and technology use may influence the effectiveness of similar interventions in international settings. These contextual factors should be carefully considered when applying our findings to broader populations.

Conclusions

In conclusion, our results suggest that a mobile peer support intervention combined with a smartphone lecture can yield significant increases in step counts among older Japanese adults. While secondary measures of total physical activity and self-reported walking time did not show a comparable effect, the observed gains in daily steps underscore the potential of app-based peer support for promoting sustained walking behavior. Moreover, a marginally significant increase in daily smartphone use and a significant increase in the number of smartphone use purposes highlight the possibility that a digital peer support program can enhance technology adoption among older adults. Future research should explore longer-term maintenance and identify which segments of the older population stand to benefit the most in order to optimize digital health interventions for this rapidly growing demographic.

Acknowledgments

We would like to thank Sumida-Ward and senior clubs in Chiba city for giving us opportunities to recruit participants. We would like to thank Active Senior Outreach Nonprofit Organization for holding smartphone lectures. We would like to thank the Institute for Assistance of Academic and Education for carrying out the randomization.

During the preparation of this work, the authors used ChatGPT in order to improve language and readability. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This study was funded by A10 Lab Inc. AN and KK have received research funding from A10 Lab Inc. to evaluate the Minchalle app and provide feedback on service improvement based on study findings. The sponsor's role was limited to introducing municipalities for recruitment and providing raw step count data. The sponsor had no involvement in the study design, data analysis, data interpretation, manuscript writing, or the decision to submit the article for publication. To safeguard independence, all statistical analyses were conducted solely by academic investigators (NA and AN), with independent oversight by YK.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: AN, TU, NA
Methodology: AN, NA, TU
Data Curation: TU, NA, GI
Formal Analysis: NA, AN
Validation: YK

XSL•FO
RenderX

Supervision: AN, KK
Writing – Original Draft Preparation: AN
Writing – Review & Editing: AN, NA, TU, GI, YK, KK
Funding Acquisition: AN, KK
All authors had full access to the dataset, contributed to the interpretation of the data, approved the final version of the manuscript, and agree to be accountable for all aspects of the work.

Conflicts of Interest

AN and KK have received research funding from A10 Lab Inc. The sponsor had no involvement in the design of the study; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Multimedia Appendix 1
Supplementary figure and additional statistical analysis tables.
[DOCX File, 4631 KB - jmir_v28i1e75708_app1.docx]

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1)
[PDF File, 57248 KB - jmir_v28i1e75708_app2.pdf]

References

1. Lee IM, Shiroma EJ, Lobelo F, et al. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. Lancet 2012 Jul 21;380(9838):219-229. [doi: 10.1016/S0140-6736(12)61031-9] [Medline: 22818936]
2. Bull FC, Al-Ansari SS, Biddle S, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br J Sports Med 2020 Dec;54(24):1451-1462. [doi: 10.1136/bjsports-2020-102955] [Medline: 33239350]
3. Healthy Japan 21. Ministry of Health Law. 2024 Mar 26. URL: https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/kenkou/kenkounippon21_00006.html [accessed 2025-03-26]
4. Physical activity and exercise guide for health promotion 2023. Ministry of Health Law. 2023. URL: https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/kenkou/undou/index.html [accessed 2025-03-26]
5. Murakami H, Tripette J, Kawakami R, Miyachi M. "Add 10 min for your health": the new Japanese recommendation for physical activity based on dose-response analysis. J Am Coll Cardiol 2015 Mar 24;65(11):1153-1154. [doi: 10.1016/j.jacc.2014.10.080] [Medline: 25790890]
6. García-Hermoso A, Ramirez-Vélez R, Sáez de Asteasu ML, et al. Safety and effectiveness of long-term exercise interventions in older adults: a systematic review and meta-analysis of randomized controlled trials. Sports Med 2020 Jun;50(6):1095-1106. [doi: 10.1007/s40279-020-01259-y] [Medline: 32020543]
7. Piercy KL, Troiano RP, Ballard RM, et al. The physical activity guidelines for Americans. JAMA 2018 Nov 20;320(19):2020-2028. [doi: 10.1001/jama.2018.14854] [Medline: 30418471]
8. French DP, Olander EK, Chisholm A, Mc Sharry J. Which behaviour change techniques are most effective at increasing older adults' self-efficacy and physical activity behaviour? A systematic review. Ann Behav Med 2014 Oct;48(2):225-234. [doi: 10.1007/s12160-014-9593-z] [Medline: 24648017]
9. Zubala A, MacGillivray S, Frost H, et al. Promotion of physical activity interventions for community dwelling older adults: a systematic review of reviews. PLoS ONE 2017;12(7):e0180902. [doi: 10.1371/journal.pone.0180902] [Medline: 28700754]
10. Dennis CL. Peer support within a health care context: a concept analysis. Int J Nurs Stud 2003 Mar;40(3):321-332. [doi: 10.1016/s0020-7489(02)00092-5] [Medline: 12605954]
11. Ginis KAM, Nigg CR, Smith AL. Peer-delivered physical activity interventions: an overlooked opportunity for physical activity promotion. Transl Behav Med 2013 Dec;3(4):434-443. [doi: 10.1007/s13142-013-0215-2] [Medline: 24294332]
12. McMahon SK, Lewis BA, Guan W, et al. Effect of intrapersonal and interpersonal behavior change strategies on physical activity among older adults: a randomized clinical trial. JAMA Netw Open 2024 Feb 5;7(2):e240298. [doi: 10.1001/jamanetworkopen.2024.0298] [Medline: 38421648]
13. Bandura A. Social cognitive theory: an agentic perspective. Annu Rev Psychol 2001;52:1-26. [doi: 10.1146/annurev.psych.52.1.1] [Medline: 11148297]
14. Laranjo L, Ding D, Heleno B, et al. Do smartphone applications and activity trackers increase physical activity in adults? Systematic review, meta-analysis and metaregression. Br J Sports Med 2021 Apr;55(8):422-432. [doi: 10.1136/bjsports-2020-102892] [Medline: 33355160]
15. Zhang J, Jemmott Iii JB. Mobile app-based small-group physical activity intervention for young African American women: a pilot randomized controlled trial. Prev Sci 2019 Aug;20(6):863-872. [doi: 10.1007/s11121-019-01006-4] [Medline: 30788692]

XSL•FO
RenderX

16. Xu L, Shi H, Shen M, et al. The effects of mHealth-based gamification interventions on participation in physical activity: systematic review. JMIR Mhealth Uhealth 2022 Feb 3;10(2):e27794. [doi: 10.2196/27794] [Medline: 35113034]
17. Tabira K, Oguma Y, Yoshihara S, et al. Digital peer-supported app intervention to promote physical activity among community-dwelling older adults: nonrandomized controlled trial. JMIR Aging 2024 May 30;7:e56184. [doi: 10.2196/56184] [Medline: 38814686]
18. Patel MS, Small DS, Harrison JD, et al. Effectiveness of behaviorally designed gamification interventions with social incentives for increasing physical activity among overweight and obese adults across the United States: the STEP UP randomized clinical trial. JAMA Intern Med 2019 Dec 1;179(12):1624-1632. [doi: 10.1001/jamainternmed.2019.3505] [Medline: 31498375]
19. Deursen AJ, Helsper EJ. The third-level digital divide: who benefits most from being online? Communication and information technologies annual: Emerald Group Publishing Limited; 2015:29-52. [doi: 10.1108/S2050-206020150000010002]
20. Campbell MK, Piaggio G, Elbourne DR, Altman DG, CONSORT Group. Consort 2010 statement: extension to cluster randomised trials. BMJ 2012 Sep 4;345:e5661. [doi: 10.1136/bmj.e5661] [Medline: 22951546]
21. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
22. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
23. Kang M, Rowe DA, Barreira TV, Robinson TS, Mahar MT. Individual information-centered approach for handling physical activity missing data. Res Q Exerc Sport 2009 Jun;80(2):131-137. [doi: 10.1080/02701367.2009.10599546] [Medline: 19650377]
24. Lee PH, Macfarlane DJ, Lam TH, Stewart SM. Validity of the International Physical Activity Questionnaire Short Form (IPAQ-SF): a systematic review. Int J Behav Nutr Phys Act 2011 Oct 21;8(115):115. [doi: 10.1186/1479-5868-8-115] [Medline: 22018588]
25. Bassett DR, Wyatt HR, Thompson H, Peters JC, Hill JO. Pedometer-measured physical activity and health behaviors in U.S. adults. Med Sci Sports Exerc 2010 Oct;42(10):1819-1825. [doi: 10.1249/MSS.0b013e3181dc2e54] [Medline: 20305579]
26. OECD. OECD digital economy outlook 2024 (volume 1): digital technology diffusion and data. : OECD Publishing; 2024 URL: https://www.oecd.org/en/publications/oecd-digital-economy-outlook-2024-volume-1_a1689dc5-en/ [accessed 2026-02-28]
27. Grande GD, Oliveira CB, Morelhão PK, et al. Interventions promoting physical activity among older adults: a systematic review and meta-analysis. Gerontologist 2020 Nov 23;60(8):583-599. [doi: 10.1093/geront/gnz167] [Medline: 31868213]
28. Skender S, Ose J, Chang-Claude J, et al. Accelerometry and physical activity questionnaires - a systematic review. BMC Public Health 2016 Jun 16;16:515. [doi: 10.1186/s12889-016-3172-0] [Medline: 27306667]

Abbreviations

BCT: behavior change technique
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
IPAQ: International Physical Activity Questionnaire
MET: metabolic equivalent of task
MHLW: Ministry of Health, Labor, and Welfare
MMRM: mixed model for repeated measures
OR: odds ratio
SCT: social cognitive theory
TIDieR: Template for Intervention Description and Replication

Edited by A Stone, N Cahill; submitted 09.Apr.2025; peer-reviewed by M Dolezel, P Liv; revised version received 05.Feb.2026; accepted 10.Feb.2026; published 10.Mar.2026.

Please cite as:
Nakagomi A, Abe N, Ueno T, Izuka G, Kawasaki Y, Kondo K
Digital Peer Support to Increase Walking Among Older Adults: Cluster Randomized Trial
J Med Internet Res 2026;28:e75708
URL: https://www.jmir.org/2026/1/e75708
doi:10.2196/75708

XSL•FO
RenderX

© Atsushi Nakagomi, Noriyuki Abe, Takayuki Ueno, Gemmei Izuka, Yohei Kawasaki, Katsunori Kondo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Enhancing Adherence to Home-Based Expiratory Muscle Strength Training in Parkinson Disease: Randomized Controlled Trial of an mHealth Intervention

Martin Srp[1*], PhD; Martina Hoskovcova[1*], PhD; Rebeka Lagnerova[1], BSc; Katerina Dvorakova[1], MSc; Radim Kliment[2], PhD; Jan Muzik[2], PhD; Radim Krupicka[3], PhD; Ota Gal[1], PhD; Evzen Ruzicka[1], MD, DrSc

[1]Department of Neurology and Center of Clinical Neuroscience, General University Hospital and First Faculty of Medicine, Charles University, Katerinska 30, Prague, Czech Republic

[2]Department of Information and Communication Technologies in Medicine, Faculty of Biomedical Engineering, Czech Technical University in Prague, Prague, Czech Republic

[3]Department of Biomedical Informatics, Faculty of Biomedical Engineering, Czech Technical University in Prague, Kladno, Czech Republic

[*]these authors contributed equally

Corresponding Author:
Evzen Ruzicka, MD, DrSc
Department of Neurology and Center of Clinical Neuroscience, General University Hospital and First Faculty of Medicine, Charles University, Katerinska 30, Prague, Czech Republic

Abstract

Background: Aspiration pneumonia is a leading cause of death in Parkinson disease (PD). Expiratory muscle strength training (EMST) is a promising intervention for respiratory and swallowing dysfunction. However, long-term EMST adherence is frequently poor in PD.

Objective: This study aims to determine whether mobile health (mHealth)–assisted EMST with the *SpiroGym* app (Czech Technical University) improves long-term adherence and physiological outcomes versus conventional EMST among participants at risk for nonadherence.

Methods: In this single-center, parallel, phase 2 randomized controlled trial, 75 individuals with PD were randomized 1:1 to conventional EMST (control; n=38) or the same protocol enhanced with the *SpiroGym* app (experimental; n=37), using a simple computer-generated randomization sequence. The *SpiroGym* is an mHealth app that provides real-time performance monitoring, direct visual feedback, and longitudinal progress tracking. All participants completed 8 weeks of semisupervised intensive EMST with biweekly in-person reassessments, followed by 16 weeks of unsupervised maintenance training. The primary outcome was adherence during weeks 8 to 24 among participants at risk for nonadherence, defined a priori at week 8 as Self-Efficacy for Home Exercise Program Scale (SEHEPS) less than 59. Because risk status was determined at week 8 and all participants subsequently entered the unsupervised phase, individuals not classified as at-risk were not excluded. Their data from week 8 onward were reported alongside the at-risk group. Secondary outcomes were changes in maximum expiratory pressure and SEHEPS.

Results: No study-related adverse events occurred. Groups were well matched at baseline (control vs experimental: mean disease duration 7.0 (SD 5.7) vs 7.3 SD 4.7) y; mean Hoehn-Yahr 1.97 (SD 0.6) vs 2.0 (SD 0.5)). The mixed-effects model showed no significant 3-way interaction (group×interval×SEHEPS risk; *P*=.14). At week 24, the at-risk category for the nonadherence cohort comprised 34 participants (control, n=17; experimental, n=17). In this at-risk cohort, the experimental group demonstrated a smaller decline in adherence during weeks 8 to 24 than controls (β=496.9, 95% CI 130.7 - 863.3; *P*=.008), completing 1073 (95% CI 643 - 1502) expiratory maneuvers versus 525 (95% CI 358 - 692). Maximum expiratory pressure increased in both groups from weeks 0 to 24, with larger gains in the experimental group (+43.1, 95% CI 32.4 - 53.8 cmH_2O) than in controls (+22.8, 95% CI 13.8 - 31.8 cmH_2O; *P*=.006; Cohen *d*=0.74). SEHEPS improved after intensive training in both groups, but only the experimental group exceeded the 12-point minimal detectable change at the 95% confidence limit.

Conclusions: This is the first randomized controlled trial to integrate mHealth with EMST. Unlike prior studies in the EMST field, we focused on sustaining long-term exercise adherence. *SpiroGym*-assisted EMST resulted in higher long-term adherence and greater gains in expiratory muscle strength than conventional EMST. In real-world PD care, assessing self-efficacy after the supervised EMST phase may help identify individuals who would benefit from digital support, making mHealth-assisted EMST a practical approach for maintaining exercise adherence.

Trial Registration: ClinicalTrials.gov NCT05728099; https://clinicaltrials.gov/study/NCT05728099

XSL•FO
RenderX

(J Med Internet Res 2026;28:e78022) doi:10.2196/78022

KEYWORDS

mobile health; randomized controlled trial; Parkinson disease; expiratory muscle strength training; exercise adherence

Introduction

Neurodegenerative disorders are now the leading cause of disability worldwide [1]. Among them, Parkinson disease (PD) is the fastest-growing condition [2]. PD is a complex, progressive neurodegenerative disorder, with symptoms primarily arising from the loss of dopaminergic neurons in the brain [3]. Aspiration pneumonia is the leading cause of death in PD [4-6]. Among nonpharmacological interventions, expiratory muscle strength training (EMST) has shown promise in improving airway protection [7,8]. However, long-term adherence to EMST, which is critical to sustaining therapeutic benefits [9], is often poor in people with chronic conditions such as PD [10]. This challenge is compounded by the home-based, unsupervised nature of EMST, limiting opportunities for professional feedback and monitoring [11]. Additionally, dopaminergic dysfunction may reduce motivation, interest in pleasurable activities, and confidence in one's ability to exercise, further potentiating nonadherence to EMST [12].

The widespread use of smartphones has enabled the development of health apps that track and manage symptoms, thereby strengthening self-care interventions for people with chronic illness [13-15]. Digital health technologies can expand access to home-based therapy and extend the benefits of in-person physiotherapy by providing real-time feedback, goal tracking, and performance summaries [14,16]. Feedback on past performance is particularly effective in strengthening exercise self-efficacy [17], which is defined as an individual's belief in their ability to successfully perform goal-directed behaviors and stands out as a consistent predictor of exercise adherence across diverse populations [18-20], including people with PD [21]. Digital health features may be especially relevant for EMST, where patients may have limited insight into whether maneuvers are performed with adequate effort and technique, and where reinforcement is minimal when training alone at home.

To address barriers to long-term EMST, we developed the *SpiroGym* app using mobile health (mHealth) technologies to provide real-time performance monitoring, direct visual feedback, and longitudinal progress tracking. A pilot study [22] demonstrated the feasibility of *SpiroGym*-assisted EMST in PD. Just 2 weeks of training significantly improved maximum expiratory pressure (MEP) [22], achieving results comparable to longer EMST protocols [23,24]. These rapid gains were likely facilitated by increased training effort driven by real-time feedback. Further validation came from a recent multicenter study showing high usability and acceptance of *SpiroGym* across diverse international PD cohorts [25]. A subsequent proof-of-concept study reported excellent adherence during a 16-week unsupervised home-based EMST program using *SpiroGym*, even among participants unaccustomed to strength training [26]. That study also reported significant improvements in the Self-Efficacy for Home Exercise Program Scale (SEHEPS) [26]. Greater confidence on this scale correlated with long-term EMST adherence, highlighting self-efficacy's critical role in sustained engagement [21,27,28]. While these findings underscore the potential of *SpiroGym*-assisted EMST, previous feasibility and proof-of-concept studies did not include an active-controlled randomized comparison against conventional EMST. Therefore, an active-controlled randomized controlled trial (RCT) is needed to determine whether mHealth-supported EMST improves adherence beyond conventional home-based EMST.

The primary objective of this RCT was to compare adherence during the unsupervised 16-week training period between conventional EMST and *SpiroGym*-assisted EMST in participants at risk for nonadherence. Secondary objectives were to evaluate changes in MEP and SEHEPS. Based on prior findings [22,26], we hypothesized that *SpiroGym*-assisted EMST would demonstrate superior results across these domains.

Methods

Study Design

We conducted a single-center, parallel, phase 2, randomized, active-controlled clinical trial that was retrospectively registered at ClinicalTrials.gov (NCT05728099) in 2023, after recruitment began in 2022 but before trial completion in 2025, and reported in accordance with the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidelines (Checklist 1) [29]. The institutional review board (IRB; the ethics committee of the General University Hospital in Prague) confirmed that the protocol and study design reported in this paper are consistent with the study protocol and design that were reviewed and approved by the ethics committee before the trial started. A side-by-side comparison of key protocol elements across the IRB submission, the ClinicalTrials.gov record, and this paper is provided in Multimedia Appendix 1. This trial was conducted at the Department of Neurology, First Faculty of Medicine, Charles University and General University Hospital in Prague, Czech Republic, between March 2022 and January 2025. Participants were randomly assigned, with allocation concealment, to either the control group following the conventional EMST protocol or to the experimental group, following the same protocol enhanced with the *SpiroGym* app.

Participants

Participants were recruited through convenience sampling from the Movement Disorders Center, Department of Neurology, First Faculty of Medicine, Charles University and General University Hospital in Prague, Czech Republic, between March 2022 and August 2024. The inclusion criteria were: (1) a diagnosis of PD confirmed by movement disorder specialist using the UK Brain Bank criteria [30], (2) a modified Hoehn and Yahr stage of I-IV in ON medication state, (3) an age of 40 to 80 years, and (4) stable dopaminergic medication (stable dose

XSL•FO
RenderX

for at least 1 mo). The exclusion criteria were: (1) suspected parkinsonism due to causes other than idiopathic PD, (2) significant cognitive impairment (Montreal Cognitive Assessment score <19), (3) respiratory disorders or diseases, (4) current or previous history of head and neck cancer, (5) smoking within the past 5 years, (6) uncontrolled hypertension, and (7) previous experience with EMST.

Randomization and Blinding

A simple computer-generated random allocation sequence (1:1 ratio) was generated by the trial statistician prior to study initiation using a web-based randomization tool. All outcome assessors and data analysts were blinded to group allocation. However, the physiotherapists overseeing the EMST training could not be blinded, as the use of the *SpiroGym* app in the experimental group was apparent. Participants were informed that the study compared 2 EMST programs. They were not given specific details regarding the nature of each program or the hypothesis that one approach might be superior, thereby maintaining participant blinding as effectively as possible under these circumstances.

Study Outcome Parameters

The primary aim of the study was to compare EMST adherence during the unsupervised home-training phase (weeks 8 - 24) between a control group and an experimental group among patients at risk for nonadherence, defined as a SEHEPS score below 59 points at the 8-week visit. The decision to identify at-risk participants at week 8 was based on previous research [26], demonstrating no correlation between baseline SEHEPS scores and later adherence, indicating that week 8 is a more reliable point to detect adherence risk. Because the SEHEPS classification was obtained at week 8, randomization at baseline could not be stratified by SEHEPS. Allocation was not altered thereafter, and participants remained in their originally randomized arms. Secondary outcomes included changes in MEP after the intensive EMST phase (weeks 0 - 8) and the maintenance phase (weeks 8 - 24), as well as changes in SEHEPS scores following the intensive phase (weeks 0 - 8). Since the risk of nonadherence was defined at week 8, and all participants proceeded from that visit into the unsupervised phase, we did not exclude those classified as not at risk. Their data from week 8 onward are reported alongside the at-risk group to provide a comprehensive view.

Assessment Visits

The same assessment protocol was completed at pre- and posttreatment visits. Outcome examinations were carried out at the same time of day in the ON medication state (1 h after regular dopaminergic medication) at the Department of Neurology, First Faculty of Medicine, Charles University and General University Hospital in Prague, Czech Republic.

Adherence (Primary Outcome)

Adherence during the unsupervised home-training phase (weeks 8 - 24) was defined as the cumulative number of completed expiratory maneuvers. The prescribed total was 800 maneuvers. In the experimental group, adherence was monitored using the *SpiroGym* app, which automatically recorded the number of completed EMST maneuvers. In the control group, adherence was tracked via self-reported training diaries (Multimedia Appendix 2), where participants recorded the number of EMST maneuvers completed.

MEP and SEHEPS (Secondary Outcomes)

MEP assessments were performed using a flanged rubber mouthpiece connected to a pressure manometer (Micro RPM, Vyaire Medical). The procedure followed standards for respiratory muscle testing [31]. Seated upright, participants inhaled to total lung capacity and then exhaled forcefully into the manometer. Participants received verbal encouragement throughout the MEP assessment. Nose clips were applied to prevent air leakage through the nose. Participants were explicitly instructed to support their cheeks and lips with their hands during the MEP maneuver, as any cheek bulging and lip leakage can dissipate pressure and result in artificially reduced values [32]. One week before baseline testing, a training MEP session was conducted to minimize learning effects [33]. Before each formal assessment, participants completed a warm-up of 6 MEP maneuvers at 50% maximal effort [34] to account for neural facilitation from repeated efforts [33]. At least 3 measurements with a less than 10% variation between them were taken, and the highest value (cmH_2O) was recorded as the MEP.

The exercise self-efficacy was assessed using the SEHEPS [35]. The test-retest reliability of SEHEPS has been demonstrated as excellent with an intraclass correlation coefficient (ICC) of 0.96 (95% CI 0.91 - 0.98) in patients with PD [26]. The minimal detectable change at the 95% confidence limit (MDC_{95}) for SEHEPS is 12 points [35]. A total self-efficacy score is derived by summing responses to 12 items, resulting in a range from 0 to 72, where higher scores indicate greater exercise self-efficacy. A SEHEPS score of less than 59 points identifies individuals at risk of nonadherence (defined as <70% of exercise adherence) to home exercise programs [35]. At baseline, patients completed the SEHEPS based on their confidence in undertaking any long-term home exercise program. At week 8, they completed the SEHEPS again, focusing specifically on their confidence in continuing long-term EMST.

Training Protocols

Conventional EMST

Patients performed EMST at home using an expiratory muscle trainer (the EMST150; Aspire Products, LLC). During the intensive phase, therapy sessions consisted of 5 sets of 5 forceful expirations, completed on 5 self-selected days per week. The EMST150 was set at 75% of each participant's MEP value. If 75% was not initially achievable, participants were instructed to temporarily lower the threshold to a tolerable level and progressively readjust it upward across subsequent sessions, with the aim of reaching 75% as soon as feasible. Participants were instructed to occlude their nose with nose clips, take a big breath in, and blow as forcefully as possible into the device to open the valve. The expiratory effort should last a couple of seconds for the air to move through the device. A typical training session lasted approximately 10 minutes. Every 2 weeks during the intensive phase, participants attended in-person visits. At each visit, a physiotherapist reassessed MEP, readjusted the device to 75% of the participant's current MEP, and briefly

XSL•FO
RenderX

verified technique (1 set of 5 expiratory maneuvers). A full 25-repetition training session was not conducted. The same procedure was applied in both groups (control and experimental). After the initial 8-week period, patients continued EMST for an additional 16 weeks (maintenance phase) without scheduled visits. In the maintenance phase, the prescription was 5 sets of 5 forceful expirations twice weekly, with optional extra sessions to mirror real-world practice. In our clinical experience, some patients prefer to exercise more often than twice a week. During the maintenance phase, participants were instructed to maintain the threshold set at the week-8 visit. Self-adjustment was not permitted. If they felt they could tolerate a higher resistance or if resistance needed to be decreased (eg, after an intercurrent illness), they were instructed to contact the study team for an in-person visit to reassess MEP and have the device reset by the physiotherapist. The same guidance applied to both groups. This protocol was adapted from our earlier proof-of-concept study [26].

SpiroGym-Assisted EMST

Participants in the experimental group followed the same EMST schedule and intensity as the control group, but their EMST150 was coupled with the *SpiroGym* app on a study smartphone (Samsung Galaxy A20e) preloaded with the app. The app monitors expiratory training in real time by analyzing acoustic signals captured via an external microphone mounted to the EMST device using a custom 3D-printed ring fabricated by our team. A download link to the ring's build instructions and a photo of the correct attachment on the EMST150 are available in [36]. The attachment does not alter resistance, airflow, or the valve and standardizes the microphone's position and angle. The microphone was wired to the phone via a 3.5-mm jack. When the training is performed correctly, the trainer's valve opens, generating increased airflow and sound. These signals are converted by the app into a visual feedback curve displayed on the participant's smartphone (Multimedia Appendix 3). Additionally, the app automatically records session data, creating a digital training diary that enables patients to review their performance over time. Although *SpiroGym* can send exercise reminders, this feature was not enabled in the present trial because participants were loaned a study device and were unlikely to carry it during the day, making push notifications unreliable. Technical details of the app have been described in detail elsewhere [22].

Sample Size

Sample size calculations were based on the primary outcome (adherence to unsupervised EMST). Because no prior controlled studies had compared EMST alone with *SpiroGym*-assisted EMST, we estimated the effect size from an internal pilot study. We analyzed week 8 to 24 adherence in the first 10 PD participants in each arm with SEHEPS less than 59 at week 8. Their adherence data yielded a Cohen *d* of 0.92 (unpublished internal data). A power analysis using G*Power (version 3.1.9.3; Heinrich Heine University Düsseldorf) [37] indicated that 16 participants at risk for nonadherence per group would achieve 80% power at a 1-sided 5% type 1 error rate to detect this effect size. Accounting for a 20% attrition rate, each group was thus targeted to include 19 participants.

Safety

Safety was assessed by monitoring adverse events throughout the study. During the intensive phase, safety was monitored via bi-weekly in-person visits. During the maintenance phase, participants were instructed to contact the study team by phone if any adverse effects occurred.

Statistical Analysis

Baseline variables were tested for normality (Kolmogorov-Smirnov test). Normally distributed data were compared with independent 2-tailed *t* tests; nonnormal data with the Mann-Whitney *U* test. Sex distribution was analyzed by the Fisher exact test. A linear mixed-effects model was used to analyze adherence across time intervals, groups, and SEHEPS subgroups, accounting for repeated measures within individuals. The model formula was:

$$\text{adherence} \sim \text{group} \times \text{interval} \times \text{SEHEPS group} + (1 \mid \text{ID})$$

where group (control/experimental), interval (wk 0-8/8-24), and SEHEPS group (<59/≥59) were included as fixed effects and their interactions, and participant ID was included as a random intercept to account for within-individual correlation. Model assumptions were checked using residual diagnostics (Shapiro-Wilk and Breusch-Pagan tests). Post hoc analyses were performed to assess group differences at each interval. Adherence was compared between groups for weeks 0 to 8 and 8 to 24, and within SEHEPS subgroups (<59 vs ≥59) using the Mann-Whitney *U* test. Four-week interval (8-12, 12-16, 16-20, and 20-24) adherences were also analyzed using the same approach. MEP changes (0 - 8, 8 - 24, and 0 - 24 wk) were assessed with paired *t* tests within groups and independent *t* tests between groups. Effect size was expressed as Cohen *d* (95 % CI), interpreted as moderate (≥0.5) and large (≥0.8). Associations between adherence and SEHEPS were examined with Spearman ρ. Our sample-size planning used a 1-sided hypothesis based on prior evidence [26] supporting a directional expectation (*SpiroGym* >control). However, for the inferential analyses, we adopted 2-sided tests to provide a more conservative interpretation. Missing data were examined for randomness using the Little Missing Completely at Random test to determine whether the missingness was random or systematic. The missing data originated from participants who did not complete the therapy. $P<.05$ was considered statistically significant. All analyses were performed in Python (version 3.12; Python Software Foundation) using the *scipy* and *statsmodels* packages.

Ethical Considerations

This study was conducted in accordance with the principles of the Declaration of Helsinki. The study was approved prospectively by the IRB of the General University Hospital in Prague (protocol 223/20 S-IV). All participants provided written informed consent prior to any study procedures, including consent for collection, analysis, and publication of deidentified data. Participation was entirely voluntary, and participants could withdraw at any time without consequence. Participants did not receive any financial compensation. All data were deidentified before analysis and stored in password-protected files accessible only to the research team in accordance with institutional data

XSL•FO
RenderX

protection policies. The paper and supplementary materials contained no identifiable images of participants. However, Multimedia Appendix 3 shows an identifiable study researcher (not a participant). Written consent for publication from this individual has been obtained.

Results

Participants

No study-related adverse events occurred during the study period. To achieve the target of 19 participants with PD in each arm with low self-efficacy, recruitment continued until this quota was achieved, which did not occur until enrolling 75 individuals. Patients in the control group (n=38) and *SpiroGym*-assisted EMST group (n=37) were well-matched at baseline. Baseline demographics also did not differ significantly between the at-risk (SEHEPS <59) and not-at-risk (SEHEPS ≥59) subgroups. A summary of demographic and clinical characteristics is provided in Table 1. A total of nine participants from the control group and seven participants from the experimental group did not complete the study (detailed reasons are provided in Figure 1), resulting in missing data. The Little Missing Completely at Random test was nonsignificant (P>.05), indicating that the data were missing completely at random. Over the 24-week period, the mean MDS-UPDRS-III motor score increased by 1.8 (SD 5.2) points in the experimental group and by 1.7 (SD 3.9) points in the control group, with no significant difference between the groups (P=.95).

XSL•FO
RenderX

Table . Baseline demographic characteristics for all randomized participants and subgroups defined at week 8 by risk for nonadherence using the Self-Efficacy for Home Exercise Program Scale (SEHEPS): at-risk (SEHEPS <59) and not at-risk (SEHEPS ≥59).

Measure	All patients			At-risk patients, SEHEPS <59			Not-at-risk patients, SEHEPS ≥59		
	Control group	Experimental group	*P* value	Control group	Experimental group	*P* value	Control group	Experimental group	*P* value
Sex (male/female), n	21/17	24/13	.48	10/8	10/9	>.99	7/6	13/3	.15
Age (y), mean (SD)	65.4 (9.5)	65.5 (8.9)	.99	66.2 (9.0)	64.8 (8.4)	.62	67.8 (9.8)	66.2 (9.6)	.68
Height (cm), mean (SD)	174 (8.9)	173 (9.3)	.68	174.8 (8.8)	171.0 (9.1)	.21	171.8 (8.4)	175.2 (9.0)	.31
Weight (kg), mean (SD)	78.3 (19.2)	79.8 (18.5)	.74	76.2 (17.0)	79.0 (17.7)	.63	78.9 (17.5)	82.6 (20.2)	.61
BMI, mean (SD)	25.8 (5.0)	26.5 (4.4)	.51	24.8 (4.4)	26.7 (4.0)	.17	26.7 (5.4)	26.7 (4.8)	.99
Education (y), mean (SD)	14.9 (2.9)	15.6 (2.9)	.25	15.6 (3.0)	15.1 (2.7)	.60	14.3 (2.6)	15.8 (2.7)	.10
Disease duration (y), mean (SD)	7 (5.7)	7.3 (4.7)	.58	6.2 (6.2)	7.5 (4.5)	.49	6.9 (4.3)	7.6 (5.1)	.72
Hoehn-Yahr, mean (SD)	1.97 (0.6)	2 (0.5)	.84	2.0 (0.5)	2.1 (0.3)	.46	2.2 (0.6)	1.9 (0.7)	.38
MDS-UPDRS III[a] (ON), mean (SD)	17.1 (8.8)	17.5 (9.5)	.86	17.3 (9.3)	20.2 (7.5)	.30	17.2 (9.4)	15.1 (11.2)	.61
MoCa[b], mean (SD)	26.3 (3.0)	25.9 (2.9)	.64	25.7 (3.2)	26.3 (2.4)	.56	27.4 (3.1)	25.5 (3.3)	.13
LEDD[c] (mg/d), mean (SD)	672 (300)	696 (295)	.75	690.0 (313.6)	673.2 (347.9)	.89	697 (280.0)	726.4 (253.0)	.78
SEHEPS, mean (SD)	40.7 (15.4)	39.8 (17.3)	.81	32.8 (12.3)	30.0 (13.8)	.52	48.3 (16.5)	49.6 (14.6)	.83
Maximum expiratory pressure (cmH_2O), mean (SD)	130 (50)	137 (42.9)	.56	133.5 (44.7)	130.9 (45.8)	.86	128.7 (54.5)	141.4 (41.1)	.48

[a]MDS-UPDRS III: Movement Disorder Society—Unified Parkinson's Disease Rating Scale, Part III: Motor Examination

[b]MoCA: Montreal Cognitive Assessment.

[c]LEDD: levodopa equivalent daily dose.

XSL•FO
RenderX

Figure 1. Flowchart of the patients with Parkinson disease in the study. SEHEPS: Self-Efficacy for Home Exercise Program Scale.

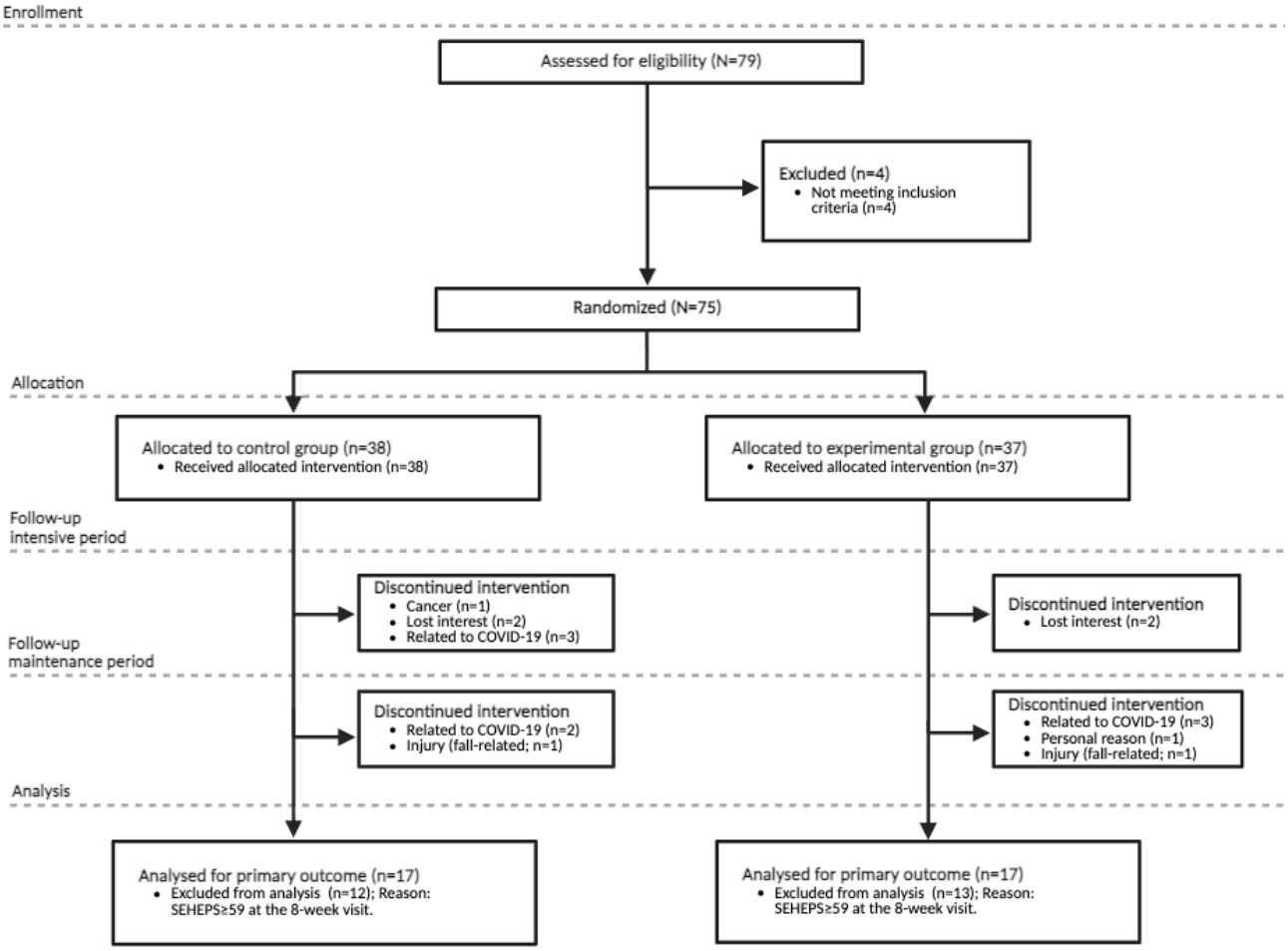

Primary Outcome: Adherence

The primary analysis using a linear mixed-effects model did not show a significant 3-way interaction between group, time interval, and baseline risk for nonadherence (group×interval×SEHEPS group; *P*=.14; Multimedia Appendix 4). However, a planned secondary analysis of participants at high risk for nonadherence (SEHEPS <59 at week 8) revealed a significant group×interval interaction. In this subgroup, the intervention group showed a significantly smaller decline in adherence during weeks 8 to 24 compared to the control group, as confirmed by a post hoc test (β=496.9, 95% CI 130.7-863.3; *P*=.008), indicating that the intervention was effective in sustaining adherence specifically in this at-risk population.

Post Hoc Analyses: Adherence

Among participants at risk for nonadherence (n=34), adherence during the maintenance phase (wk 8 - 24) was significantly higher in the experimental group than in the control group (*P*=.04; Cohen *d*=0.80; Figure 2 and Table 2). Within this group, 47% (8/17) of *SpiroGym* participants versus 18% (3/17) of controls achieved the minimum prescribed 800 or more repetitions during weeks 8 to 24. A 4-week interval analysis revealed that between-group differences in adherence reached significance during weeks 12 to 16 (*P*=.04; Cohen *d*=0.84) and weeks 20 to 24 (*P*=.01; Cohen *d*=1.03), with no significant differences during weeks 8 to 12 or 16 to 20 (Figure 3). Among patients (n=25) without risk for nonadherence, no difference in adherence was observed during the maintenance period between the control and experimental groups (*P*=.34; Cohen *d*=0.22; Table 2; Multimedia Appendix 5). For the overall sample (n=59), adherence during the intensive phase (wk 0 - 8) did not differ significantly between groups (*P*=.05; Cohen *d*=0.49), while during the maintenance phase (wk 8 - 24), the experimental group achieved significantly greater adherence (*P*=.02; Cohen *d*=0.58; Table 2).

XSL•FO
RenderX

Figure 2. Adherence to expiratory muscle strength training during weeks 8 to 24 in participants at risk for nonadherence based on the Self-Efficacy for Home Exercise Program Scale (SEHEPS <59). The values are presented as mean (SD).

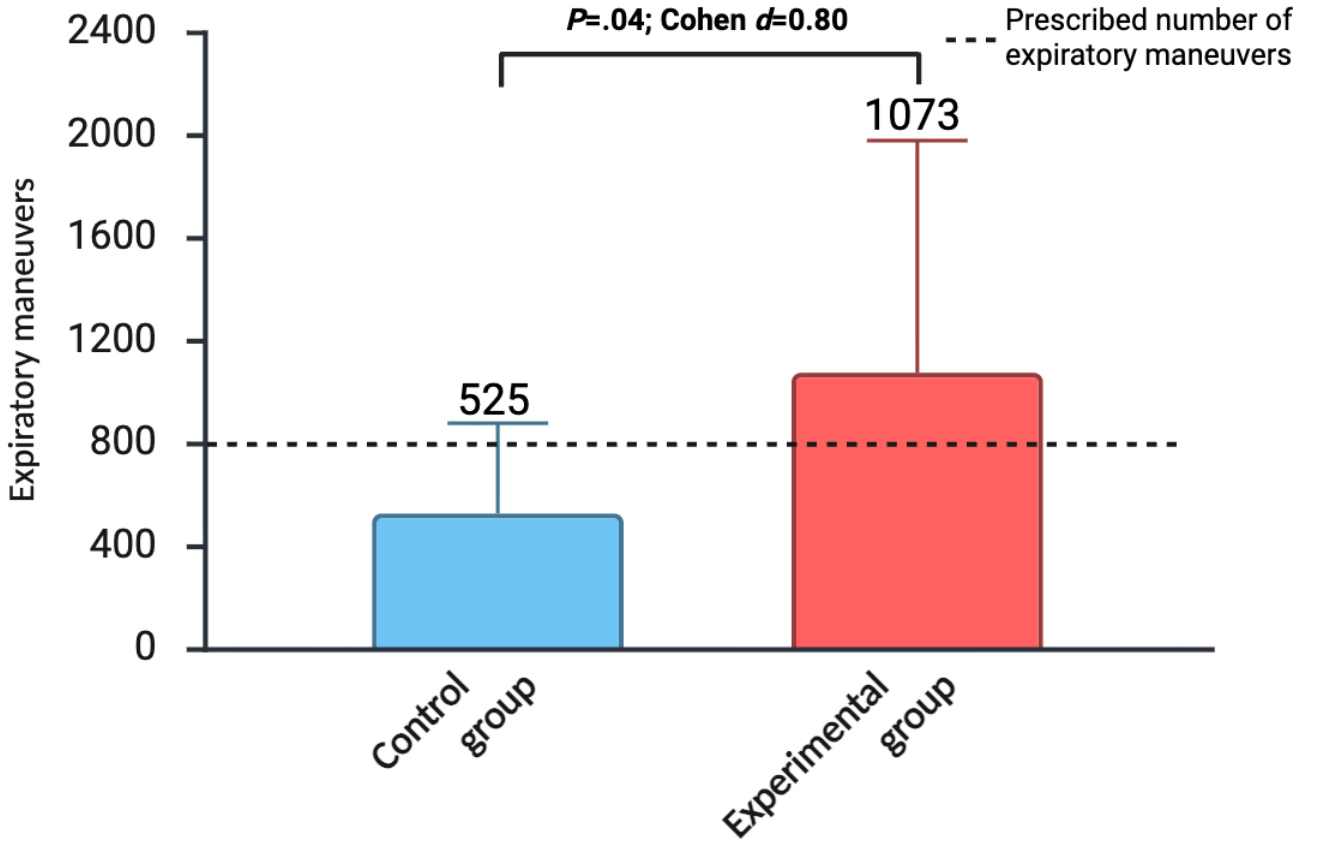

XSL•FO
RenderX

Table . Adherence to expiratory muscle strength training by study interval: weeks 0 to 8 (intensive phase) and weeks 8 to 24 (maintenance), for the full sample and for subgroups defined at week 8 by Self-Efficacy for Home Exercise Program Scale (SEHEPS) risk (at-risk: SEHEPS <59; not at-risk: SEHEPS ≥59).

Time interval and group	Participants, n	Repetitions, mean (SD)	95% CI	Between-group effect	
				P value	Cohen *d* (95% CI)
Overall adherence (all patients)					
Weeks 0-8				.05	0.49 (0.01 to 0.97)
CG[a]	32	953 (83.2)	924 to 982		
EG[b]	35	998 (95.7)	966 to 1030		
Weeks 8-24				.02	0.58 (0.06 to 1.10)
CG	29	771 (538)	575 to 967		
EG	30	1139 (712)	884 to 1393		
Adherence analysis based on SEHEPS at week 8					
Weeks 8-24				.04	0.80 (0.10 to 1.50)
CG SEHEPS <59	17	525 (351)	358 to 692		
EG SEHEPS <59	17	1073 (903)	643 to 1502		
Weeks 8-24				.34	0.22 (–0.57 to 1.0)
CG SEHEPS ≥59	12	1120 (577)	794 to 1446		
EG SEHEPS ≥59	13	1225 (352)	1034 to 1416		
Four-week interval adherence analysis in patients at risk of nonadherence					
Weeks 8-12				.15	0.74 (–0.01 to 1.4)
CG	17	165 (98.4)	118 to 212		
EG	17	277 (204)	180 to 374		
Weeks 12-16				.04	0.84 (0.14 to 1.54)
CG	17	135 (108)	84 to 186		
EG	17	301 (258)	178 to 423		
Weeks 16-20				.06	0.75 (0.05 to 1.45)
CG	17	107 (118)	51 to 163		
EG	17	255 (252)	135 to 375		
Weeks 20-24				.01	1.03 (0.32 to 1.75)
CG	17	74 (90.3)	31 to 116		
EG	17	240 (208)	141 to 339		

[a]CG: control group.

[b]EG: experimental group.

XSL•FO
RenderX

Figure 3. Four-week interval adherence to expiratory muscle strength training during weeks 8 to 24 in patients at risk for nonadherence (Self-Efficacy for Home Exercise Program Scale [SEHEPS] <59). The values are presented as mean (SD). CG: control group; EG: experimental group.

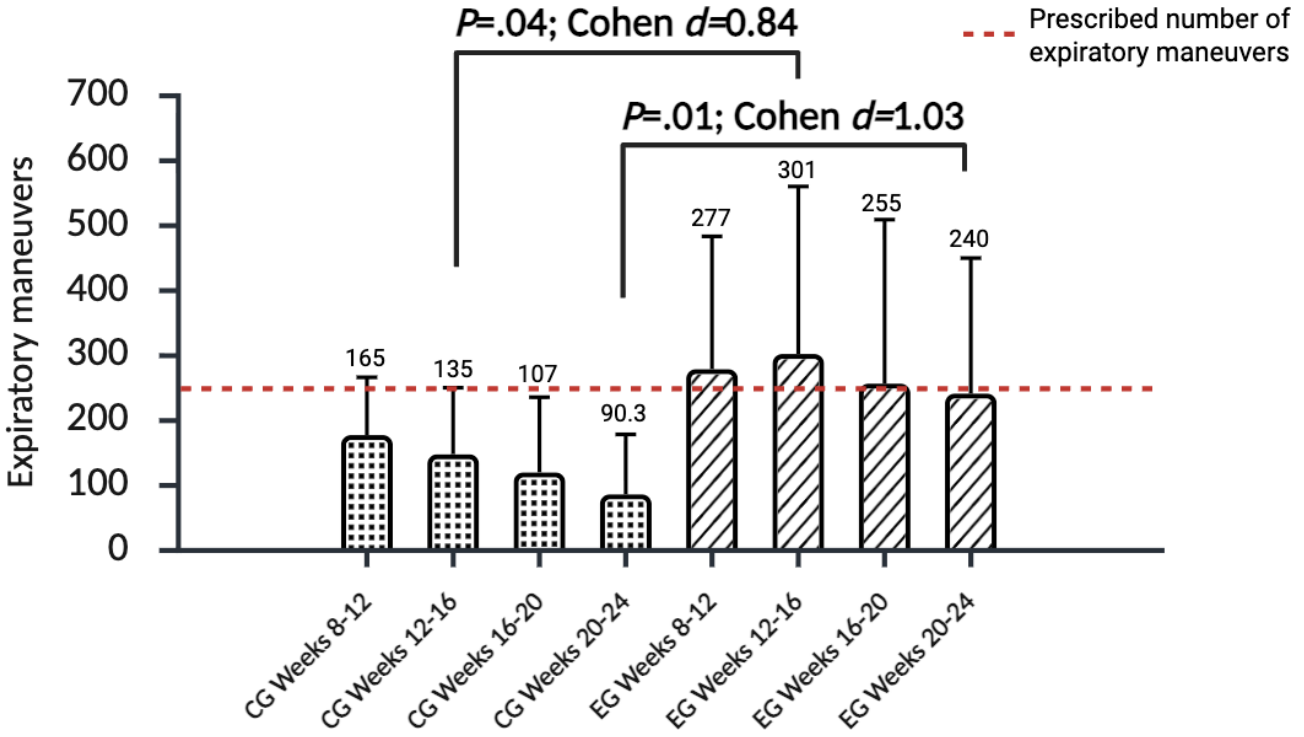

Secondary Outcomes

Maximum Expiratory Pressure

From baseline to week 8, MEP increased by 28.4 cmH_2O in the control group (P<.001; Cohen d=0.58) and by 42.3 cmH_2O in the experimental group (P<.001; Cohen d=0.89), with the experimental group showing significantly greater improvement (P=.03; Cohen d=0.54). During the unsupervised period (wk 8 - 24), MEP decreased by 6.8 cmH_2O in the control group (P=.11; Cohen d=–0.15) and increased by 0.1 cmH_2O in the experimental group (P=0.96, Cohen d=0.004), with no significant between-group difference (P=.16; Cohen d=0.37). Over the entire 24-week study period, MEP increased by 22.8 cmH_2O in the control group (P<.001; Cohen d=0.50) and by 43.1 cmH_2O in the experimental group (P<.001; Cohen d=0.91), with the experimental group demonstrating significantly greater improvements (P=.006; Cohen d=0.74). Table 3 summarizes MEP outcomes.

XSL•FO
RenderX

Table . Overall maximum expiratory pressure changes by study interval: weeks 0 to 8 (intensive phase) and weeks 8 to 24 (maintenance), for the full sample and for subgroups defined at week 8 by Self-Efficacy for Home Exercise Program Scale (SEHEPS) risk (at-risk: SEHEPS <59; not at-risk: SEHEPS ≥59).

Time interval and group	Participants (n)	Preinterval, mean (SD)	Postinterval, mean (SD)	Within-group change		Between-group effect	
				Mean (SD) Δ[a]	95% CI Δ	*P* value	Cohen *d* (95% CI)
All patients							
Weeks 0-8						.03	0.54 (0.05 to 1.03)
CG[b]	31	131.5 (48.2)	159.9 (50.3)	+28.4 (25.8)	19.3 to 37.5		
EG[c]	35	135.7 (43.4)	178.1 (51.7)	+42.3 (25.9)	33.8 to 50.9		
Weeks 8-24						.16	0.37 (–0.14 to 0.89)
CG	29	161.3 (50.4)	154.6 (42.6)	–6.8 (22.4)	–15 to 1.4		
EG	30	179.7 (51.1)	179.9 (52.8)	+0.1 (13.7)	–4.8 to 5.0		
Weeks 0-24						.006	0.74 (0.21 to 1.27)
CG	29	131.8 (48.1)	154.6 (42.6)	+22.8 (24.7)	13.8 to 31.8		
EG	30	136.8 (41.4)	179.9 (52.8)	+43.1 (29.9)	32.4 to 53.8		
Subgroup analysis based on exercise self-efficacy (SEHEPS at wk 8)							
Weeks 8-24 (SEHEPS <59)						.04	0.73 (0.04 to 1.43)
CG	17	168.1 (53.4)	153.5 (45.4)	–14.6 (21)	–24.6 to –4.6		
EG	17	180.1 (49.9)	177.9 (50.7)	–2.2 (11.3)	–7.6 to 3.1		
Weeks 8-24 (SEHEPS ≥59)						.90	–0.06 (–0.84 to 0.73)
CG	12	151.8 (46.4)	156.1 (40)	+4.2 (20.4)	–7.3 to 15.8		
EG	13	179.2 (54.5)	182.5 (57.5)	+3.2 (16.3)	–5.6 to 12.1		

[a]Δ: change.

[b]CG: control group.

[c]EG: experimental group.

Self-Efficacy for Home Exercise Program Scale

From baseline to week 8, the mean SEHEPS score significantly increased by 11.2 (SD 13.6, 95% CI 6.36-15.96; *P*<.001) points in the control group and by 14.5 (SD 14.0, 95% CI 9.91-19.2; *P*<.001) points in the experimental group. Furthermore, at week 8, SEHEPS scores correlated positively with adherence during weeks 8 to 24 in both groups (control: ρ=0.60, *P*=.001; experimental: ρ=0.44, *P*=.01).

Discussion

Principal Findings

This phase 2 RCT compared the efficacy of a conventional EMST protocol with a *SpiroGym*-assisted protocol. For the primary objective, participants at risk for EMST nonadherence who trained with *SpiroGym* sustained greater adherence during the unsupervised phase than those following conventional EMST. Regarding secondary objectives, the *SpiroGym* group achieved greater improvements in MEP and a clinically meaningful increase in self-efficacy, whereas the control group resulted in smaller but still significant gains.

Building on earlier findings [26] demonstrating sustained user engagement with the *SpiroGym* app during long-term EMST, our study adds further evidence that *SpiroGym* can effectively support adherence over time. Notably, participants with sufficient self-efficacy (SEHEPS ≥59) performed 30% more expiratory maneuvers than the minimum prescribed target during the maintenance period in both groups. In contrast, among those with low self-efficacy (SEHEPS <59), adherence in the control group dropped from 83% (wk 8 - 12) to 37% (wk 20 - 24), while the *SpiroGym* group consistently (every 4-wk interval) maintained adherence 20% above the prescribed maneuver target. These findings align with Chung et al [38], who reported higher adherence and enhanced self-efficacy in an mHealth-supported exercise program compared to a paper-based approach. Similarly, Sørensen et al [39] reported that feedback-based inspiratory muscle training yielded better

XSL•FO
RenderX

adherence than self-reported protocols over a 12-week unsupervised period.

In line with our prior study [26], we observed a positive correlation between SEHEPS scores at week 8 and adherence during weeks 8 to 24 in both groups, suggesting that self-efficacy assessed after the intensive phase may predict long-term EMST compliance. Although SEHEPS scores improved significantly in both protocols, only the *SpiroGym* group exceeded the 12-point MDC_{95} threshold [35], indicating a true increase in exercise self-efficacy beyond measurement error—a finding supported by previous research [40].

These findings have direct clinical implications for tailoring EMST. Patients should be assessed for self-efficacy following the intensive phase. Those with low self-efficacy may benefit from enhanced supervision or motivational support, such as the *SpiroGym* app, to sustain adherence during the maintenance period. Conversely, conventional EMST may be sufficient for patients with higher self-efficacy. Beyond self-efficacy, other factors such as apathy or caregiver support may plausibly influence adherence and treatment outcomes and may guide treatment selection in practice [41,42]. Screening for these factors alongside SEHEPS may help triage patients to conventional versus supported EMST.

Another key finding relates to real-time visual feedback and its impact on MEP. After the 8-week intensive phase, both groups achieved meaningful MEP gains, but the increase was significantly larger in the *SpiroGym* group (P=.03; Cohen d=0.54). On average, MEP increased by 22% in the control group and 31% in the experimental group, consistent with prior reports of approximately 24% MEP gains in PD [8]. Based on the observation that, during weeks 0 - 8, the experimental group showed greater MEP gains despite comparable adherence, we hypothesize that real-time visual feedback may enhance motivational aspects of learning and thereby facilitate better skill acquisition. This interpretation is consistent with prior studies showing amplified respiratory muscle activation under visual feedback conditions [43]. Given the role of EMST in airway protection [4,6], these superior MEP outcomes suggest that SpiroGym-assisted EMST may translate into improved clinical outcomes for individuals with dysphagia or impaired cough, although this remains to be confirmed in future research.

A further notable result pertains to long-term MEP maintenance in patients at risk for nonadherence (SEHEPS <59). While the control group showed an average 9% decline in MEP between weeks 8 and 24, the *SpiroGym* group maintained stable values. Previous studies report detraining-related MEP declines of 2% to 20% over 1 to 18 months in PD [9,44,45]. Long-term maintenance training at least twice per week is generally recommended [46]. In our study, control group participants missed an average of 11 out of 32 scheduled sessions, falling below the minimum recommended dose and likely contributing to the observed MEP reduction. In contrast, adherence in the experimental group was sufficient to prevent detraining effects. On the other hand, resistance-training adaptations typically extend beyond 8 weeks in both healthy adults and people with PD. One likely explanation for the plateau in MEP gains is that no participant requested recalibration during the maintenance period. As a result, training thresholds remained fixed, reducing progressive resistance. This contrasts with evidence from a 2-year randomized trial, in which Corcos et al [47] showed that people with PD who trained twice weekly with progressively increased resistance continued to improve upper-limb strength up to the 24-month visit, whereas those following a nonprogressive program did not. A second contributor may be the device load ceiling. In a subset of experimental participants (n=11), week-8 MEP exceeded 200 cmH_2O, placing the 75% training target above the EMST150's maximum load. To sustain MEP progression, future protocols should implement routine recalibration during the maintenance period and, for participants approaching the ceiling, employ devices with a higher resistance range.

Limitations

This study has several limitations. First, adherence in the control group was based on self-reported training diaries, which might be prone to inaccuracy [48]. Second, we did not include objective measures of swallowing function or cough efficacy, limiting conclusions about the broader clinical impact of the observed increase in the MEP. Third, the sample primarily consisted of individuals with mild-to-moderate PD, which limits the generalizability of findings to those with more advanced PD or cognitive impairment. Fourth, we did not collect person-centered outcomes (eg, treatment burden and quality of life). Although a separate study reports acceptability of *SpiroGym*, it does not provide a head-to-head comparison with conventional EMST [25,26]. Consequently, it remains unclear whether SpiroGym-supported EMST reduces perceived burden or improves quality of life relative to standard care. Fifth, during the maintenance phase, participants were instructed to perform five sets of five expirations twice weekly, with the option to train more frequently to reflect real-world practice. This flexibility allowed some individuals to accumulate higher totals of expiratory maneuvers, which may have influenced the effect size.

Future studies should involve larger cohorts, longer follow-up periods, and objective assessments of airway protection. In addition, qualitative and mixed methods studies comparing *SpiroGym* with traditional EMST are warranted to understand patient experience (eg, satisfaction, motivation, enjoyment, and usability).

To support further research and clinical adoption, the *SpiroGym* app is freely available at [36], along with detailed hardware specifications. A user manual is included in Multimedia Appendix 6.

Conclusions

This is the first RCT to integrate mHealth with EMST. Unlike prior studies in the EMST field, we focused on long-term exercise adherence, which is critical for sustaining therapeutic effects. Compared with conventional EMST, *SpiroGym*-assisted training led to higher adherence, most notably among patients prone to nonadherence, and produced greater expiratory muscle strength gains. In routine PD care, evaluating self-efficacy after the supervised EMST phase may help identify patients who are likely to benefit from digital support, positioning

XSL•FO
RenderX

mHealth-assisted EMST as a practical strategy to sustain exercise adherence.

Acknowledgments

The authors thank the individuals with Parkinson disease and their families who participated in this study for their immense dedication and contribution to this work. The authors also thank Pavel Dusek for his assistance with patient recruitment and his methodological guidance. The authors declare the use of generative artificial intelligence in the research and writing process. According to GAIDeT (Generative AI Delegation Taxonomy; 2025) [49], the following tasks were delegated to generative artificial intelligence tools under full human supervision: proofreading and editing. ChatGPT 5.2 (GPT; OpenAI's large-scale language model) was used for these purposes.

Funding

The study was supported by the Czech Health Research Council (project NW24-04-00259) and by General University Hospital in Prague (project MH CZ-DRO-VFN64165).

Data Availability

The authors have deposited an open-access deidentified participant-level dataset along with the source code for statistical analysis on GitHub [50].

Authors' Contributions

MS was involved in the conception, organization, and execution of the research project; design, execution, review, and critique of statistical analysis; and writing the original draft of the manuscript. MH was involved in the conception, organization, and execution of the research project; design, execution, review, and critique of statistical analysis; and writing the original draft of the manuscript. RL was involved in the conception and organization of the research project and reviewing and critiquing the manuscript. KD was involved in the execution of the research project, reviewing and critiquing the statistical analysis, and reviewing and critiquing the manuscript. R Kliment was involved in the conception of the research project and reviewing and critiquing the manuscript. JM was involved in the conception of the research project and reviewing and critiquing the manuscript. R Krupicka was involved in the conception and organization of the research project and reviewing and critiquing the manuscript. OG was involved in the conception of the research project, reviewing and critiquing statistical analysis, and reviewing and critiquing the manuscript. ER was involved in the conception of the research project, reviewing and critiquing statistical analysis, and reviewing and critiquing the manuscript.

Conflicts of Interest

MS, RL, R Kliment, and ER are creators of the *SpiroGym* app. All other authors declare no conflicts of interest.

Editorial Notice

This randomized study was only retrospectively registered because of an initial lack of awareness of the requirement for prospective registration prior to study commencement, with the need for registration recognized only during the course of the study. The editor granted an exception from the International Committee of Medical Journal Editors' rules mandating prospective registration of randomized trials because the risk of bias appears low and because the authors provided an official letter from the institutional review board confirming that the submitted study design is identical to the protocol assessed and approved before trial initiation. However, readers are advised to carefully assess the validity of any potential explicit or implicit claims related to primary outcomes or effectiveness, as retrospective registration does not prevent authors from changing their outcome measures retrospectively.

Multimedia Appendix 1
A side-by-side comparison of key protocol elements across the institutional review board submission, the ClinicalTrials.gov record (NCT05728099), and published manuscript.
[DOCX File, 18 KB - jmir_v28i1e78022_app1.docx]

Multimedia Appendix 2
Self-monitoring training diaries.
[DOCX File, 17 KB - jmir_v28i1e78022_app2.docx]

Multimedia Appendix 3
Visual feedback provided by the SpiroGym app during expiratory muscle strength training.
[MP4 File, 10843 KB - jmir_v28i1e78022_app3.mp4]

XSL•FO
RenderX

Multimedia Appendix 4
Linear mixed-effects model for adherence.
[DOCX File, 18 KB - jmir_v28i1e78022_app4.docx]

Multimedia Appendix 5
Four-week interval adherence analysis in patients without risk of nonadherence.
[DOCX File, 17 KB - jmir_v28i1e78022_app5.docx]

Multimedia Appendix 6
SpiroGym user manual.
[DOCX File, 438 KB - jmir_v28i1e78022_app6.docx]

Checklist 1
CONSORT-EHEALTH checklist.
[PDF File, 1247 KB - jmir_v28i1e78022_app7.pdf]

References

1. Dorsey ER, Bloem BR. The Parkinson pandemic—a call to action. JAMA Neurol 2018 Jan 1;75(1):9-10. [doi: 10.1001/jamaneurol.2017.3299] [Medline: 29131880]
2. Steinmetz JD, Seeher KM, Schiess N, et al. Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet Neurol 2024 Apr;23(4):344-381. [doi: 10.1016/S1474-4422(24)00038-3]
3. Magrinelli F, Picelli A, Tocco P, et al. Pathophysiology of motor dysfunction in Parkinson's disease as the rationale for drug treatment and rehabilitation. Parkinsons Dis 2016;2016(1):9832839. [doi: 10.1155/2016/9832839] [Medline: 27366343]
4. D'Amelio M, Ragonese P, Morgante L, et al. Long-term survival of Parkinson's disease: a population-based study. J Neurol 2006 Jan;253(1):33-37. [doi: 10.1007/s00415-005-0916-7] [Medline: 16021349]
5. Fall PA, Saleh A, Fredrickson M, Olsson JE, Granérus AK. Survival time, mortality, and cause of death in elderly patients with Parkinson's disease: a 9-year follow-up. Mov Disord 2003 Nov;18(11):1312-1316. [doi: 10.1002/mds.10537] [Medline: 14639673]
6. Pennington S, Snell K, Lee M, Walker R. The cause of death in idiopathic Parkinson's disease. Parkinsonism Relat Disord 2010 Aug;16(7):434-437. [doi: 10.1016/j.parkreldis.2010.04.010]
7. Claus I, Muhle P, Czechowski J, et al. Expiratory muscle strength training for therapy of pharyngeal dysphagia in Parkinson's disease. Mov Disord 2021 Aug;36(8):1815-1824. [doi: 10.1002/mds.28552] [Medline: 33650729]
8. Troche MS, Curtis JA, Sevitz JS, et al. Rehabilitating cough dysfunction in Parkinson's disease: a randomized controlled trial. Mov Disord 2023 Feb;38(2):201-211. [doi: 10.1002/mds.29268] [Medline: 36345090]
9. Troche MS, Rosenbek JC, Okun MS, Sapienza CM. Detraining outcomes with expiratory muscle strength training in Parkinson disease. J Rehabil Res Dev 2014;51(2):305-310. [doi: 10.1682/JRRD.2013.05.0101] [Medline: 24933728]
10. Ellis T, Cavanaugh JT, Earhart GM, et al. Factors associated with exercise behavior in people with Parkinson disease. Phys Ther 2011 Dec;91(12):1838-1848. [doi: 10.2522/ptj.20100390] [Medline: 22003171]
11. Isernia S, Pagliari C, Jonsdottir J, et al. Efficiency and patient-reported outcome measures from clinic to home: the human empowerment aging and disability program for digital-health rehabilitation. Front Neurol 2019;10:1206. [doi: 10.3389/fneur.2019.01206] [Medline: 31824398]
12. Stevens A, Stanton R, Rebar AL. Helping people with Parkinson disease build exercise self-efficacy. Phys Ther 2020 Feb 7;100(2):205-208. [doi: 10.1093/ptj/pzz160] [Medline: 31665447]
13. Lee J, Yeom I, Chung ML, Kim Y, Yoo S, Kim E. Use of mobile apps for self-care in people with Parkinson disease: systematic review. JMIR Mhealth Uhealth 2022 Jan 21;10(1):e33944. [doi: 10.2196/33944] [Medline: 35060910]
14. Templeton JM, Poellabauer C, Schneider S, et al. Modernizing the staging of Parkinson disease using digital health technology. J Med Internet Res 2025 Apr 4;27:e63105. [doi: 10.2196/63105] [Medline: 40184612]
15. Willemse IHJ, Schootemeijer S, van den Bergh R, Dawes H, Nonnekes JH, van de Warrenburg BPC. Smartphone applications for movement disorders: towards collaboration and re-use. Parkinsonism Relat Disord 2024 Mar;120:105988. [doi: 10.1016/j.parkreldis.2023.105988] [Medline: 38184466]
16. Litman L, Rosen Z, Spierer D, Weinberger-Litman S, Goldschein A, Robinson J. Mobile exercise apps and increased leisure time exercise activity: a moderated mediation analysis of the role of self-efficacy and barriers. J Med Internet Res 2015 Aug 14;17(8):e195. [doi: 10.2196/jmir.4142] [Medline: 26276227]
17. Ashford S, Edmunds J, French DP. What is the best way to change self-efficacy to promote lifestyle and recreational physical activity? A systematic review with meta-analysis. Br J Health Psychol 2010 May;15(Pt 2):265-288. [doi: 10.1348/135910709X461752] [Medline: 19586583]

XSL•FO
RenderX

18. Jancey J, Lee A, Howat P, Clarke A, Wang K, Shilton T. Reducing attrition in physical activity programs for older adults. J Aging Phys Act 2007 Apr;15(2):152-165. [doi: 10.1123/japa.15.2.152] [Medline: 17556782]
19. Oman RF, King AC. Predicting the adoption and maintenance of exercise participation using self-efficacy and previous exercise participation rates. Am J Health Promot 1998;12(3):154-161. [doi: 10.4278/0890-1171-12.3.154] [Medline: 10176088]
20. Rhodes RE, Martin AD, Taunton JE. Temporal relationships of self-efficacy and social support as predictors of adherence in a 6-month strength-training program for older women. Percept Mot Skills 2001 Dec;93(3):693-703. [doi: 10.2466/pms.2001.93.3.693] [Medline: 11806588]
21. Schootemeijer S, van der Kolk NM, Ellis T, et al. Barriers and motivators to engage in exercise for persons with Parkinson's disease. J Parkinsons Dis 2020;10(4):1293-1299. [doi: 10.3233/JPD-202247] [Medline: 32925106]
22. Srp M, Korteová R, Kliment R, Jech R, Růžička E, Hoskovcová M. Expiratory muscle strength training in patients with Parkinson's disease: a pilot study of mobile monitoring application. Mov Disord Clin Pract 2021 Oct;8(7):1148-1149. [doi: 10.1002/mdc3.13313] [Medline: 34631956]
23. Reyes A, Castillo A, Castillo J. Effects of expiratory muscle training and air stacking on peak cough flow in individuals with Parkinson's disease. Lung 2020 Feb;198(1):207-211. [doi: 10.1007/s00408-019-00291-8] [Medline: 31720808]
24. Reyes A, Castillo A, Castillo J, Cornejo I. The effects of respiratory muscle training on peak cough flow in patients with Parkinson's disease: a randomized controlled study. Clin Rehabil 2018 Oct;32(10):1317-1327. [doi: 10.1177/0269215518774832] [Medline: 29756459]
25. Pérez-Rangel D, Monje MHG, Grosjean S, et al. Insights from a mixed methods analysis of 3 health technologies used in patients with Parkinson disease: mixed methods study. J Med Internet Res 2025 Aug 1;27:e67986. [doi: 10.2196/67986] [Medline: 40749223]
26. Srp M, Ferro Á, Ferreira J, et al. mHealth-assisted expiratory muscle strength training in Parkinson's disease patients: a proof-of-concept study. J Parkinsons Dis 2024 Nov;14(8):1623-1630. [doi: 10.1177/1877718X241296013] [Medline: 39957190]
27. Brassington GS, Atienza AA, Perczek RE, DiLorenzo TM, King AC. Intervention-related cognitive versus social mediators of exercise adherence in the elderly. Am J Prev Med 2002 Aug;23(2 Suppl):80-86. [doi: 10.1016/s0749-3797(02)00477-4] [Medline: 12133741]
28. Neupert SD, Lachman ME, Whitbourne SB. Exercise self-efficacy and control beliefs: effects on exercise behavior after an exercise intervention for older adults. J Aging Phys Act 2009 Jan;17(1):1-16. [doi: 10.1123/japa.17.1.1] [Medline: 19299835]
29. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
30. Hughes AJ, Daniel SE, Kilford L, Lees AJ. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. J Neurol Neurosurg Psychiatry 1992 Mar 1;55(3):181-184. [doi: 10.1136/jnnp.55.3.181]
31. Laveneziana P, Albuquerque A, Aliverti A, et al. ERS statement on respiratory muscle testing at rest and during exercise. Eur Respir J 2019 Jun;53(6):1801214. [doi: 10.1183/13993003.01214-2018] [Medline: 30956204]
32. Rubinstein I, Slutsky AS, Rebuck AS, et al. Assessment of maximal expiratory pressure in healthy adults. J Appl Physiol 1988 May;64(5):2215-2219. [doi: 10.1152/jappl.1988.64.5.2215] [Medline: 3391920]
33. Windisch W, Criée CP. Maximum respiratory pressure assessment: standardization is required. Chest 2022 Oct;162(4):738-739. [doi: 10.1016/j.chest.2022.05.033] [Medline: 36210100]
34. Marion R, Niebuhr BR. Effect of warm-up prior to maximal grip contractions. J Hand Ther 1992 Jul;5(3):143-146. [doi: 10.1016/S0894-1130(12)80349-8]
35. Picha KJ, Lester M, Heebner NR, et al. The self-efficacy for home exercise programs scale: development and psychometric properties. J Orthop Sports Phys Ther 2019 Sep;49(9):647-655. [doi: 10.2519/jospt.2019.8779]
36. SpiroGym. GitHub. URL: https://fbme-ictm.github.io/spirogym [accessed 2026-02-13]
37. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007 May;39(2):175-191. [doi: 10.3758/bf03193146] [Medline: 17695343]
38. Chung BPH, Chiang WKH, Lau H, et al. Pilot study on comparisons between the effectiveness of mobile video-guided and paper-based home exercise programs on improving exercise adherence, self-efficacy for exercise and functional outcomes of patients with stroke with 3-month follow-up: a single-blind randomized controlled trial. Hong Kong Physiother J 2020 Jun;40(1):63-73. [doi: 10.1142/S1013702520500079] [Medline: 32489241]
39. Sørensen D, Svenningsen H. Adherence to home-based inspiratory muscle training in individuals with chronic obstructive pulmonary disease. Appl Nurs Res 2018 Oct;43:75-79. [doi: 10.1016/j.apnr.2018.07.005] [Medline: 30220368]
40. Theis S, Schäfer D, Haubrich C, et al. Perceived self-efficacy in Parkinson's disease through mobile health monitoring. In: Stephanidis C, Antona M, Gao Q, Zhou J, editors. HCI International 2020 – Late Breaking Papers: Universal Access and Inclusive Design: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 2020:749-762. [doi: 10.1007/978-3-030-60149-2_57]

XSL•FO
RenderX

41. Cools CI, Kotz SA, Bloem BR, Duits AA, de Vries NM. Motivation matters: elucidating factors driving exercise in people with Parkinson disease. Phys Ther 2025 Jun 2;105(6):pzaf048. [doi: 10.1093/ptj/pzaf048] [Medline: 40207869]
42. Kim M, Kim E, Kim M, et al. Motivators and barriers affecting exercise in patients with Parkinson's disease. J Clin Neurol 2025 Jan;21(1):13-20. [doi: 10.3988/jcn.2024.0328] [Medline: 39778563]
43. Park HK, Kim YJ, Kim TH. The role of visual feedback in respiratory muscle activation and pulmonary function. J Phys Ther Sci 2015 Sep;27(9):2883-2886. [doi: 10.1589/jpts.27.2883] [Medline: 26504317]
44. Huang CC, Lai YR, Wu FA, et al. Detraining effect on pulmonary and cardiovascular autonomic function and functional outcomes in patients with Parkinson's disease after respiratory muscle training: an 18-month follow-up study. Front Neurol 2021;12:735847. [doi: 10.3389/fneur.2021.735847] [Medline: 34744975]
45. Saleem AF, Sapienza CM, Okun MS. Respiratory muscle strength training: treatment and response duration in a patient with early idiopathic Parkinson's disease. NeuroRehabilitation 2005;20(4):323-333. [Medline: 16403998]
46. Sapienza C, Hoffman B. Respiratory Muscle Strength Training: Plural Publishing; 2020.
47. Corcos DM, Robichaud JA, David FJ, et al. A two-year randomized controlled trial of progressive resistance exercise for Parkinson's disease. Mov Disord 2013 Aug;28(9):1230-1240. [doi: 10.1002/mds.25380] [Medline: 23536417]
48. Schmidt M, Paul SS, Canning CG, et al. The accuracy of self-report logbooks of adherence to prescribed home-based exercise in Parkinson's disease. Disabil Rehabil 2022 Apr;44(8):1260-1267. [doi: 10.1080/09638288.2020.1800106] [Medline: 32762573]
49. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2025 Aug 8:1-27. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]
50. Expiratory muscle strength training in Parkinson's disease - analysis. GitHub. URL: https://github.com/kbi-fbmi/articles--2025expiratory [accessed 2026-02-26]

Abbreviations

CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
EMST: expiratory muscle strength training
ICC: intraclass correlation coefficient
IRB: institutional review board
MDC95: minimal detectable change at the 95% confidence level
MEP: maximum expiratory pressure
mHealth: mobile health
PD: Parkinson disease
RCT: randomized controlled trial
SEHEPS: Self-Efficacy for Home Exercise Program Scale

Edited by S Brini; submitted 24.May.2025; peer-reviewed by A Dakin, J Sevitz; revised version received 17.Jan.2026; accepted 18.Jan.2026; published 11.Mar.2026.

Please cite as:
Srp M, Hoskovcova M, Lagnerova R, Dvorakova K, Kliment R, Muzik J, Krupicka R, Gal O, Ruzicka E
Enhancing Adherence to Home-Based Expiratory Muscle Strength Training in Parkinson Disease: Randomized Controlled Trial of an mHealth Intervention
J Med Internet Res 2026;28:e78022
URL: https://www.jmir.org/2026/1/e78022
doi:10.2196/78022

© Martin Srp, Martina Hoskovcova, Rebeka Lagnerova, Katerina Dvorakova, Radim Kliment, Jan Muzik, Radim Krupicka, Ota Gal, Evzen Ruzicka. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Comparing the Associations of Internet Addiction and Internet Gaming Disorder With Psychopathological Symptoms: Cross-Sectional Study of Three Independent Adolescent Samples

Ying-ying Li[1*], BMBS; A-qian Hu[1*], MM; Ling-li Yi[2,3*], MM; Zi-xin Mao[3,4], MM; Qiu-yue Lü[3], MD, PhD; Juan Wang[1], PhD; Wei Wei[1], MD, PhD; Yue-qi Huang[1], MM; Shu Huang[1], BMBS; Wen-jing Dai[1,5], BMBS; Meng-xuan Qiao[1], MM; Jia-jun Xu[3], MD, PhD; Qiang Wang[3], MD, PhD; Xiao-jing Li[1,3,6,7], MD, PhD; Fu-gang Luo[1], BMBS; Wei Deng[1,7], MD, PhD; Yu-zheng Hu[6,7], PhD; Tao Li[1,3,5,6,7], MD, PhD; Wan-jun Guo[1,3,5,6,7], MD, PhD

[1]Affiliated Mental Health Center & Hangzhou Seventh People's Hospital, School of Brain Science and Brain Medicine, Zhejiang University School of Medicine, 305 Tiamnyshan Rd, Hangzhou, Zhejiang, China

[2]MOE Key Lab for Neuroinformation, The Clinical Hospital of Chengdu Brain Science Institute, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

[3]Mental Health Center and Psychiatric Laboratory, State Key Laboratory of Biotherapy, West China Hospital of Sichuan University, Chengdu, Sichuan, China

[4]Department of Psychiatry, Yichang Mental Health Center, Hubei, China

[5]School of Mental Health, Wenzhou Medical University, 325035, Wenzhou, China

[6]Department of Psychology and Behavioral Science, Zhejiang University, Hangzhou, 148 Tianmushan Road, Xihu District, Hangzhou, China

[7]Liangzhu Laboratory, MOE Frontier Science Center for Brain Science and Brain-Machine Integration, State Key Laboratory of Brain-Machine Intelligence, Zhejiang University, Hangzhou, China

[*]these authors contributed equally

Corresponding Author:
Wan-jun Guo, MD, PhD
Affiliated Mental Health Center & Hangzhou Seventh People's Hospital, School of Brain Science and Brain Medicine, Zhejiang University School of Medicine, 305 Tiamnyshan Rd, Hangzhou, Zhejiang, China

Abstract

Background: Both internet gaming disorder (IGD) and internet addiction (IA) have been associated with diverse psychopathological symptoms. However, how the 2 conditions relate to each other and which is more strongly associated with psychopathology remain unclear.

Objective: This study aimed to examine the association between IGD and IA and compare the strength of their associations with various types of psychopathological symptoms.

Methods: This cross-sectional study surveyed 3 independent samples of Chinese adolescents: the first sample (S1) comprised 8194 first-year undergraduates at a comprehensive university in Chengdu, the second sample (S2) comprised 1720 students from a high school in Hangzhou, and the third sample (S3) comprised 551 inpatients aged 13 to 19 years recruited from 2 tertiary psychiatric hospitals in Hangzhou and Chengdu. IGD was defined as a score of 22 or more on the Internet Gaming Disorder Scale–Short Form (IGDS9-SF), whereas IA was defined as a score of 50 or more on Young's 20-item Internet Addiction Test (IAT-20). Symptoms of depression, anxiety, psychoticism, paranoid ideation, and attention-deficit or hyperactivity were assessed using internationally validated scales including 9-item the Patient Health Questionnaire, 7-item Generalized Anxiety Disorder, psychoticism and paranoid ideation subscales of the Symptom Checklist 90 (absent for S2), and Adult ADHD Self-Report Scale (absent for S1), through online surveys in S1 (October 2020) and S3 (January 2022 to February 2025) and via an offline survey in S2 (March 2024).

Results: The prevalence estimates (95% CI) of IGD were 4.8% (4.3% - 5.2%) in S1, 15.8% (14.0% - 17.5%) in S2, and 32.3% (28.4% - 36.2%) in S3, whereas prevalence estimates (95% CI) of IA were consistently higher across samples, ranging from 7.3% (6.8% - 7.9%) in S1 and 18.8% (17.0% - 20.6%) in S2 to 45.9% (41.8% - 50.1%) in S3. The IGDS9-SF and the IAT-20 were moderately correlated (Pearson r=0.51 - 0.57; all P<.001) and were associated with the severity of most psychopathological symptom domains, with consistently stronger associations observed for IAT-20 scores. In multivariate models including all psychopathological symptoms as independent variables, the coefficients of determination (R^2, 95% CIs) were consistently higher for the IAT-20 than for the IGDS9-SF in S1 (0.33, 0.30 - 0.35 vs 0.13, 0.11 - 0.16) and S2 (0.44, 0.39 - 0.49 vs 0.23, 0.18 - 0.27), with a similar but nonsignificant pattern observed in S3 (0.13, 0.06 - 0.26 vs 0.06, 0.03 - 0.16). Post hoc analyses indicated that

XSL•FO
RenderX

psychopathological symptoms were generally more severe in individuals with IA, either alone or comorbid with IGD, than in those with IGD only.

Conclusions: This study provides additional evidence that IGD and IA are distinct yet interrelated constructs, and further demonstrates that IA consistently exhibits stronger associations with the severity of psychopathological symptoms than IGD. These findings underscore the importance of recognizing and addressing compulsive and problematic online behaviors that extend beyond gaming, highlighting the need to refine diagnostic frameworks and prioritize targeted clinical interventions.

(J Med Internet Res 2026;28:e82414) doi:10.2196/82414

KEYWORDS

adolescent; attention deficit; hyperactivity; internet addiction; internet gaming disorder; psychopathology

Introduction

In 2023, approximately 5.4 billion people were using the internet [1], many of whom were teenagers and young adults. These groups are at a critical stage of growth and development, which can be endangered by excessive or inappropriate use of the internet, potentially leading to internet gaming disorder (IGD) and internet addiction (IA) [2,3]. Recent systematic reviews and meta-analyses indicate that IA and IGD are increasingly prevalent worldwide, particularly among adolescents and young adults, and are associated with a growing burden of mental health and functional impairments [4,5].

IGD has been proposed in the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* as a condition warranting further study [6], which specifies 9 dimensions of preoccupation, withdrawal, tolerance, unsuccessful attempts to stop or limit gaming, loss of interest in previous activities, continued use despite harm, avoidance, deception, and harm. The 11th edition of the *International Classification of Diseases, Eleventh Revision* (*ICD-11*) has defined gaming disorder, including predominantly online gaming, emphasizing impaired control, increasing priority given to gaming over other activities, and continued gaming despite negative consequences [7]. Both definitions imply significant social dysfunction arising from excessive online gaming. Studies based on one of the most widely used validated instruments for assessing internet gaming disorder that was proposed by a meta-analysis for screening internet gaming disorder, the Internet Gaming Disorder Scale–Short Form (IGDS9-SF) [8], suggest that the condition affects 9.3%-18.2% of adolescents in various countries [9,10]. Recent epidemiological and longitudinal studies have consistently shown that IGD in adolescents is associated with a wide range of adverse outcomes, including depressive and anxiety symptoms, attention-deficit or hyperactivity problems, aggression, sleep disturbances, poorer quality of life, and suicidality [11-13].

In contrast to IGD, IA has yet to be recognized as a bona fide disorder in the *Diagnostic and Statistical Manual* or other consensus guidelines [14]. It is typically defined as a behavioral addiction characterized by impaired control over internet use or online behaviors that results in clinically significant distress and functional impairment [15]. Studies using the most widely used instrument for assessing IA, the 20-item Young's Internet Addiction Test (IAT-20) [16], suggest that up to 18% of adolescents may suffer from moderate IA and up to 1.5% may suffer from severe IA [17]. Studies based on the IAT-20 have linked IA to diverse psychopathological symptoms including depression, anxiety, obsessive-compulsive disorder, and suicidal ideation [18,19].

How IA and IGD relate to each other is unclear, with studies proposing that they are one and the same, that they are entirely separate from each other, or that one is a subtype of the other [20,21]. Although recent work has begun to assess generalized problematic internet use, gaming-specific problems, and other online behaviors within the same samples, highlighting both shared risk factors (eg, negative affect, maladaptive coping) and partially distinct symptom profiles across these conditions [22,23], their associations with psychopathological symptoms have not been systematically compared between IGD and IA within the same samples. Therefore, this study employed internationally validated instruments to assess IA and IGD in 3 adolescent samples from distinct settings in China, with the goal of examining the association between the 2 conditions and determining which is more strongly related to psychopathological symptom severity.

Methods

Participants

This study involved participants from 3 independent samples. Sample 1 (S1) comprised first-year undergraduates enrolled in 2020 at Sichuan University, a large comprehensive university in southwestern China that recruits students from all provincial-level administrative regions nationwide. As part of the Development of Psychological Health Assessment and Crisis Alarm and Intervention System project, all freshmen (n=9409) were invited to complete an online, self-administered psychosomatic health questionnaire via the university's Online Psychosomatic Health Survey in October 2020. This study used the data from 8125 freshmen who provided valid responses to the Online Psychosomatic Health Survey, representing a response rate of 86.4%. Sample 2 (S2) comprised 1720 respondents (response rate: 78.2%) out of 2200 students invited to participate in a school-wide mental health survey at a high school in Hangzhou, China, conducted in March 2024 using paper-based questionnaires. Sample 3 (S3) comprised 551 respondents (response rate: 86.8%) out of 635 psychiatric inpatients aged 13 - 19 years recruited from 2 psychiatric hospitals in Hangzhou and Chengdu, China, to complete online self-report scales of psychiatric symptoms between January 2022 and February 2025.

XSL•FO
RenderX

Individuals were excluded from the study if they were younger than 13 or older than 19 years, if they did not complete all questionnaire items, if their responses seemed unreliable (eg, endorsing the same symptom severity across all items), or if they used the same personal identification number as another participant. Undergraduates were excluded if they failed to submit their questionnaires properly through the online system or submitted them after the designated deadline.

Measurements

The questionnaire packets contained queries on sociodemographic background along with various internationally validated assessment scales, including the IGDS9-SF, the IAT-20, the Patient Health Questionnaire-9 (PHQ-9), Generalized Anxiety Disorder-7 (GAD-7), psychoticism and paranoid Ideation Subscales of the Symptom Checklist 90 (administered only for S1 and S3), and Adult ADHD Self-Report Scale (ASRS; administered only for S2 and S3). These instruments were used to assess IGD, IA, and 5 domains of psychopathology (depression, anxiety, psychoticism, paranoid ideation, and attention-deficit or hyperactivity symptoms), respectively.

Internet Gaming Disorder Scale–Short Form

The IGDS9-SF [8] assesses the frequency of symptoms of IGD during the previous 12 months. The symptoms correspond to the diagnostic criteria of the 5th edition of the *Diagnostic and Statistical Manual* and the features defined by the 11th edition of the *International Classification of Diseases*. The IGDS9-SF is one of the few questionnaires assessing IGD that includes the criterion of "continued internet use despite harm." Participants respond to each of the 9 items on a 5-point Likert scale from 1 ("rarely") to 5 ("always"), yielding a total score ranging from 0 to 45. Scores ≥22 were defined as indicative of IGD [24]. Cronbach α in all 3 samples of this study ranged from 0.89 to 0.93.

Twenty-Item Young's Internet Addiction Test

The IAT-20 [25] is the most frequently used self-report measure of problematic internet use [26]. Participants respond to each of the 20 items on a 5-point Likert scale from 1 ("rarely") to 5 ("always") or by entering the value 0 ("not applicable"), yielding a total score from 0 to 100. Scores ≥50 were considered indicative of IA [27]. In this study, Cronbach α across the 3 samples ranged from 0.93 to 0.94.

Nine-Item Patient Health Questionnaire

The PHQ-9 assesses the severity of depressive symptoms over the past 2 weeks. Participants respond to each of the 9 items using a 4-point scale ranging from 0 ("not at all") to 3 ("nearly every day"), yielding a total score range of 0 - 27. Higher scores indicate greater depressive symptom severity. The scale has demonstrated satisfactory psychometric properties among Chinese populations [28]. In this study, Cronbach α across the 3 samples ranged from 0.86 to 0.91.

Seven-Item Generalized Anxiety Disorder

The GAD-7 [29] assesses the severity of anxiety symptoms during the previous 2 weeks. Participants respond to each of 7 items on a 4-point scale from 0 ("not at all") to 3 ("nearly every day"), yielding a total score range of 0-21. Higher scores indicate greater anxiety symptom severity. The Chinese version has demonstrated satisfactory validity and reliability. Cronbach α in all 3 samples of this study ranged from 0.90 to 0.93.

Psychoticism and Paranoid Ideation Subscales of the Symptom Checklist-90

The severity of psychoticism and paranoid ideation was assessed using the corresponding subscales of the Symptom Checklist-90. Participants respond to each of the 10 items on the psychoticism subscale or 6 items on the paranoid ideation subscale using a 5-point Likert scale ranging from 0 ("not at all") to 4 ("extremely"). The SCL-90 has demonstrated robust validity and reliability in Chinese populations [30]. These subscales were administered only in S1 (undergraduates) and S3 (inpatients). Cronbach α was 0.83 (S1) and 0.79 (S3) for psychoticism and 0.87 (S1) and 0.84 (S3) for paranoid ideation.

Adult ADHD Self-Report Scale

The ASRS, widely used clinically to screen adults for attention-deficit or hyperactivity disorder, is based on the 18 criteria in the "TR" revision of the 4th edition of the *Diagnostic and Statistical Manual of Mental Disorders* [31]. It examines symptoms over the past 6 months. Participants respond to 18 items on a Likert scale from 0 ("never") to 4 ("very often"), giving a total score from 0 to 72. Higher scores indicate greater ADHD symptom severity. The Chinese version has demonstrated both reliability and validity in young adults in Taiwan [32]. This survey was administered only in S2 (high school students) and S3 (inpatients). Cronbach α was 0.93 in S2 and 0.91 in S3.

Analysis

The data were analyzed using SPSS 27.0 (IBM) and R (version 4.4.3; R Core Team). Descriptive statistics, correlation analyses, independent-samples *t* tests, ANOVA, and post hoc comparisons were conducted in SPSS, whereas regression-related analyses, including the computation of confidence intervals for R^2 (not available in SPSS), were performed in R. The significance of univariate associations was assessed using, as appropriate, Pearson correlation coefficients (*r*), 2-tailed *t* tests, chi-square tests, or ANOVA. ANOVA results were adjusted using Tukey honest significant difference procedure to control for multiple comparisons. Potential relationships among variables were explored using multiple linear and logistic regression. To control for the confounding effects of demographics, we used standardized residuals (ZREs) of psychopathology scores adjusted for age and sex in regression and post hoc analyses; descriptive statistics were based on raw scores to facilitate comparability with prior studies. To evaluate whether multicollinearity among independent variables (eg, $r>0.7$ between the PHQ-9 and the GAD-7) affected the reliability of our regression models, we obtained tolerance and variance inflation factors before multivariate regression analyses. Where appropriate, associations were expressed as odds ratios (ORs) with corresponding 95% CIs. The results were considered statistically significant at a 2-tailed $P<.05$. The missing data in S2 were handled using complete-case analysis without multiple imputation after testing for missing completely at random, as

XSL•FO
RenderX

item-level missingness was minimal (0.2% - 3.0% per item). S1 and S3 used forced-response online questionnaires, resulting in no item-level missingness. The ASRS data in S3 were available only for 241 participants because the scale was introduced midway through the study. Accordingly, analyses involving the ASRS in S3 included only these 241 participants.

Ethical Considerations

The study procedures were carried out in accordance with the Declaration of Helsinki, and the study was conducted and reported in accordance with the Journal Article Reporting Standards [33]. The study was approved by the Ethics Committee of West China Hospital, Sichuan University (approval No. 2016 - 171), and the Ethics Committee of Hangzhou Seventh People's Hospital (approval No. 2023 - 064). All S1 participants provided electronic informed consent; for S2 and S3, both participants and their parents received detailed study information and provided written informed consent. All data were anonymized and stored securely on password-protected servers, and only aggregated results were reported. Participants did not receive any financial or material compensation for their participation in this study.

Results

Demographics

Male participants constituted a slight majority in S1 (n=8194, 54.5%), were slightly underrepresented in S2 (n=1720, 43.8%), and were substantially underrepresented in S3 (n=551, 26.0%). Mean ages (95% CI) in years were 18.1 (18.1-18.1) among undergraduates, 17.3 (17.0-17.6) among high school students, and 15.6 (15.4-15.7) among inpatients.

Measurement Scores of IGD and IA and Their Correlation With Demographics

The mean IGDS9-SF score (95% CI) in S1 was 11.8 (11.7-12.0), which was significantly lower than in S2 (15.1, 14.8-15.4), and S2 was significantly lower than S3 (18.8, 17.8-19.7). Similar trends were observed across samples for IA. The mean IAT-20 score (95% CI) in S1 was 31.7 (31.5-32.0), which was significantly lower than in S2 (32.4, 31.5-33.3), and S2 was significantly lower than S3 (47.7, 45.5-49.9; Table 1).

XSL•FO
RenderX

Table . Mean scores of measurements of internet gaming disorder (IGD), internet addiction (IA), and psychopathology, and the Pearson correlation matrix among demographic variables and these measures, across the 3 samples[a].

Measure	Scores, mean (95% CI)	Pearson correlation coefficients (*r*) among demographic variables and measures of IGD, IA, and psychopathology								
		Age	Gender	IAT-20[b]	IGDS9-SF[c]	PHQ-9[d]	GAD-7[e]	SCL90-psychoticism[f]	SCL90–paranoid ideation[g]	ASRS[h]
Sample 1 (n=8194)										
IGDS9-SF	11.8 (11.7-12.0)	0.02	–0.28[i]	0.51[i]	1.00	0.27[i]	0.24[i]	0.32[i]	0.27[i]	—[j]
IAT-20	31.7 (31.5-32.0)	–0.02[k]	0.08[i]	1.00	0.51[i]	0.50[i]	0.46[i]	0.52[i]	0.48[i]	—
PHQ-9	3.0 (2.9-3.1)	–0.00	0.12[i]	0.50[i]	0.27[i]	1.00	0.77[i]	0.63[i]	0.56[i]	—
GAD-7	2.0 (1.9-2.1)	–0.00	0.10[i]	0.46[i]	0.24[i]	0.77[i]	1.00	0.61[i]	0.56[i]	—
SCL90-psychoticism	1.2 (1.2-1.3)	–0.01	0.03[l]	0.52[i]	0.32[i]	0.63[i]	0.61[i]	1.00	0.80[i]	—
SCL90–paranoid ideation	1.3 (1.2-1.3)	–0.02	0.06[i]	0.48[i]	0.27[i]	0.56[i]	0.56[i]	0.80[i]	1.00	—
Sample 2 (n=1720)										
IGDS9-SF	15.1 (14.8-15.4)	–0.02	–0.24[i]	0.55[i]	1.00	0.35[i]	0.30[i]	—	—	0.41[i]
IAT-20	32.4 (31.5-33.3)	0.05[k]	0.09[i]	1.00	0.55[i]	0.54[i]	0.48[i]	—	—	0.62[i]
PHQ-9	6.3 (6.0-6.5)	0.03	0.11[i]	0.54[i]	0.35[i]	1.00	0.69[i]	—	—	0.55[i]
GAD-7	4.1 (3.9-4.2)	0.02	0.07[l]	0.48[i]	0.30[i]	0.69[i]	1.00	—	—	0.55[i]
ASRS	22.8 (22.3-23.3)	–0.01	0.05[k]	0.62[i]	0.41[i]	0.55[i]	0.55[i]	—	—	1.00
Sample 3										
IGDS9-SF (n=551)	18.8 (17.8-19.7)	–0.10[k]	–0.14[l]	0.57[i]	1.00	0.13	0.08	0.14[k]	0.16[k]	0.24[i]
IAT-20 (n=551)	47.7 (45.5-49.9)	–0.09[k]	0.03	1.00	0.57[i]	0.24[i]	0.16[k]	0.26[i]	0.18[l]	0.35[i]
PHQ-9 (n=551)	18.2 (17.4-19.1)	–0.16[i]	0.13[l]	0.24[i]	0.13	1.00	0.77[i]	0.67[i]	0.57[i]	0.57[i]
GAD-7 (n=551)	13.5 (12.9-14.1)	–0.14[l]	0.11[l]	0.16[k]	0.08	0.77[i]	1.00	0.69[i]	0.58[i]	0.53[i]
SCL90-psychoticism (n=551)	2.5 (2.3-2.6)	–0.18[i]	0.06	0.26[i]	0.14[k]	0.67[i]	0.69[i]	1.00	0.77[i]	0.67[i]
SCL90–paranoid ideation (n=551)	2.5 (2.3-2.7)	–0.17[i]	0.07	0.18[l]	0.16[k]	0.57[i]	0.58[i]	0.77[i]	1.00	0.59[i]
ASRS (n=241)[m]	54.0 (51.8-56.2)	–0.17[l]	0.12	0.35[i]	0.24[i]	0.57[i]	0.53[i]	0.67[i]	0.59[i]	1.00

[a]Sample 1 comprises undergraduate freshmen enrolled at Sichuan University. Sample 2 comprises students recruited from a high school in Hangzhou.

XSL•FO
RenderX

Sample 3 comprises inpatients aged 13 - 19 years recruited from 2 tertiary mental health centers in Hangzhou and Chengdu. The Internet Gaming Disorder Scale–Short Form, 20-item Young's Internet Addiction Test, Patient Health Questionnaire-9, and Generalized Anxiety Disorder-7 were administered in all 3 samples; the SCL-90 Psychoticism and Paranoid Ideation subscales were administered only in S1 and S3, and the Adult ADHD Self-Report Scale was administered only in S2 and S3.

[b]IAT-20: 20-item Young's Internet Addiction Test.

[c]IGDS9-SF: Internet Gaming Disorder Scale–Short Form.

[d]PHQ-9: 9-item Patient Health Questionnaire.

[e]GAD-7: 7-item Generalized Anxiety Disorder.

[f]SCL90-psychoticism: subscale of the Symptom Checklist-90 to measure psychoticism.

[g]SCL90–paranoid ideation: subscale of the Symptom Checklist-90 to measure paranoid ideation.

[h]ASRS: Adult ADHD Self-Report Scale.

[i]P<.001.

[j]Not available.

[k]P<.05.

[l]P<.01.

[m]The Adult ADHD Self-Report Scale data in sample 3 were available only for 241 participants because the scale was introduced midway through the study.

IGDS9-SF and IAT-20 scores showed either statistically significant but modest correlations with age or nonsignificant associations across all 3 samples (r=–0.10 to 0.10; Table 1). The scores of the IGDS9-SF were significantly higher in male than in female participants in all 3 samples (t_{8192}=26.2, P<.001 in S1; t_{1718}=9.92, P<.001 in S2; and t_{549}=3.18, P=.002 in S3). Conversely, the scores of the IAT-20 were significantly higher in female than in male participants in S1 (t_{8192}=–7.00; P<.001) and S2 (t_{1718}=–3.67; P<.001), whereas no significant sex differences were observed in S3 (t_{549}=0.61; P=.55).

Prevalence Rates of IGD and IA

The prevalence (95% CI) of IGD in S1, based on the predefined cutoff (IGDS9-SF ≥22), was 4.8% (4.3%-5.2%), which was significantly lower than in S2 (15.8%, 14.0%-17.5%), and in turn significantly lower than in S3 (32.3%, 28.4%-36.2%). The prevalence (95% CI) of IA in S1, based on the predefined cutoff (IAT-20 ≥50), was 7.3% (6.8%-7.9%), which was significantly lower than in S2 (18.8%, 17.0%-20.6%), and in turn was significantly lower than that in S3 (45.9%, 41.8%-50.1%).

In other words, in S1 (n=8194), most undergraduates (n=7408, 90.4%, 95% CI 89.8%-91.1%) had neither IGD nor IA, whereas small proportions (n=184, 2.24%, 95% CI 1.94%-2.59%) had IGD only, IA only (n=396, 4.83%, 95% CI 4.38%-5.32%), or both (n=206, 2.51%, 95% CI 2.19%-2.88%). In S2 (n=1720), a smaller majority (n=1287, 74.83%, 95% CI 72.75%-76.83%) of the high school students had neither disorder, with correspondingly higher prevalence of only IGD (n=109, 6.34%, 95% CI 5.27%-7.59%), only IA (n=163, 9.48%, 95% CI 8.17%-10.94%), and both (n=161, 9.36%, 95% CI 8.06%-10.83%). In contrast, in S3 (n=551), a slight majority of the inpatients had one or both disorders: only 253 (45.92%, 95% CI 41.71%-50.18%) had neither disorder, whereas 45 (8.17%, 95% CI 6.08%-10.86%) had IGD only, 120 (21.78%, 95% CI 18.45%-25.51%) had IA only, and 133 (24.14%, 95% CI 20.67%-27.98%) had both (Table 1).

Associations of Measurements and Prevalence Between IGD and IA

Univariate analyses indicated that IGDS9-SF and IAT-20 scores were moderately correlated in all 3 samples (r=0.51, P<.001 in S1; r=0.55, P<.001 in S2; and r=0.51, P<.001 in S3). These correlations remained significant after controlling for sex and age (Table 1). Similarly, multivariable analyses controlling for sex and age indicated that the odds of IA were significantly higher among participants with IGD in all 3 samples, with adjusted ORs (95% CI) of 29.6 (23.2-37.9) in S1, 14.1 (10.3-19.6) in S2, and 6.5 (4.3-9.9) in S3.

Scores of Psychopathological Symptoms

The mean PHQ-9 score (95% CI) in S1 was 3.0 (2.9-3.1), which was significantly lower than in S2 (6.3, CI 6.0-6.5) and in turn significantly lower than in S3 (18.2, 17.4-19.1). Similarly, the mean GAD-7 score (95% CI) in S1 was 2.0 (1.9-2.1), which was significantly lower than in S2 (4.1, 3.9-4.2), and in turn was significantly lower than in S3 (13.5, 12.9-14.1). The mean SCL-90 psychoticism subscale score (95% CI) in S1 was 1.2 (1.2-1.3), which was significantly lower than in S3 (2.5, 2.3-2.6). Similarly, the mean SCL-90 paranoid ideation subscale score (95% CI) in S1 was 1.3 (1.2-1.3), which was significantly lower than that in S3 (2.5, 2.3-2.7). The mean ASRS score (95% CI) in S2 was 22.8 (22.3-23.3), which was significantly lower than that in S3 (54.0, 51.8-56.2).

Psychopathology severity was not associated with age in S1 or S2 but showed small negative correlations with age in S3 (r=–0.09 to –0.18). All measured types of psychopathological symptoms in S1 and S2 were significantly more severe in female than in male participants, whereas only depression (PHQ-9) and anxiety (GAD-7) exhibited this sex difference in S3 (Table 1).

Associations of IGD or IA With Severity of Psychopathological Symptoms

To examine whether IGD or IA was associated with psychopathology, we performed pairwise comparisons between participants with and without IGD and between participants with and without IA; these analyses revealed statistically significant differences across all psychopathological variables

XSL•FO
RenderX

in each sample. Notably, participants with IA exhibited significantly higher dimensional psychopathology scores than those with IGD on the PHQ-9, GAD-7, and the SCL-90 paranoid ideation subscale in S1. A similar trend of differences was observed in S2 and S3, although the differences did not reach statistical significance (Table 2).

Table . Mean scores (95% CI) of measurements for internet gaming disorder (IGD), internet addiction (IA), and other psychopathologies among all participants, those with IGD and IA, and those without IGD or IA, across the 3 samples[a].

Sample and psychopathologies	Scores among participants with IGD[b], mean (95% CI)	Scores among participants with IA[c], mean (95% CI)	Scores among participants without IGD[d], mean (95% CI)	Scores among participants without IA[e], mean (95% CI)
Sample 1				
IGDS9-SF[f]	25.2 (24.7-25.7)	17.6 (17.0-18.2)	11.2 (11.1-11.2)	11.4 (11.3-11.5)
IAT-20[g]	50.8 (49.3-52.3)	58.2 (57.5-58.8)	30.8 (30.5-31.0)	29.6 (29.4-29.8)
PHQ-9[h]	6.0 (5.5-6.5)	7.2 (6.9-7.6)	2.8 (2.8-2.9)	2.7 (2.6-2.7)
GAD-7[i]	4.6 (4.1-5.1)	5.6 (5.3-6.0)	1.9 (1.9-2.0)	1.8 (1.7-1.8)
SCL90-psychoticism[j]	1.6 (1.5-1.7)	1.7 (1.6-1.7)	1.2 (1.2-1.2)	1.2 (1.2-1.2)
SCL90–paranoid ideation[k]	1.7 (1.6-1.7)	1.8 (1.7-1.8)	1.3 (1.3-1.3)	1.2 (1.2-1.3)
Sample 2				
IGDS9-SF	26.5 (25.9-27.0)	21.3 (20.4-22.2)	13.0 (12.8-13.2)	13.6 (13.4-13.8)
IAT-20	53.4 (51.3-55.4)	62.8 (61.8-63.9)	28.4 (27.5-29.4)	32.4 (31.9-32.8)
PHQ-9	9.8 (9.1-10.5)	10.6 (10.0-11.2)	5.7 (5.4-5.9)	5.2 (5.0-5.4)
GAD-7	6.8 (6.1-7.5)	7.4 (6.9-7.9)	3.5 (3.3-3.8)	3.2 (3.1-3.3)
ASRS[l]	31.9 (30.4-33.5)	33.9 (32.7-35.0)	21.0 (20.3-21.6)	20.0 (19.4-20.6)
Sample 3				
IGDS9-SF	30.0 (29.0-30.9)	23.1 (21.9-24.4)	13.5 (13.0-13.9)	15.1 (14.3-15.8)
IAT-20	60.9 (58.1-63.8)	67.2 (65.6-68.8)	41.4 (39.4-43.5)	31.2 (29.7-32.6)
PHQ-9	17.9 (17.0-18.9)	18.9 (17.8-19.9)	18.4 (17.7-19.1)	17.7 (16.9-18.5)
GAD-7	13.2 (12.4-14.1)	14.0 (13.3-14.7)	13.6 (13.0-14.3)	13.1 (12.4-13.8)
SCL90-psychoticism	2.6 (2.4-2.7)	2.7 (2.6-2.9)	2.4 (2.3-2.5)	2.2 (2.1-2.3)
SCL90–paranoid ideation	2.7 (2.5-2.9)	2.8 (2.7-2.9)	2.4 (2.3-2.5)	2.3 (2.1-2.4)
ASRS[m]	56.5 (54.0-59.0)	56.7 (54.9-58.5)	52.4 (50.3-54.4)	50.0 (46.3-53.7)

[a]Internet gaming disorder was defined as a total score ≥22 on the IDSG9-SF, and internet addiction was defined as a total score ≥50 on Young's 20-item Internet Addiction Test (IAT-20).

[b]Sample sizes for participants with internet gaming disorder: sample 1: n=393; sample 2: n=271; sample 3: n=178.

[c]Sample sizes for participants with internet addiction: sample 1: n=598; sample 2: n=323; sample 3: n=253.

[d]Sample sizes for participants without internet gaming disorder: sample 1: n=7801; sample 2: n=1,449; sample 3: n=373.

[e]Sample sizes for participants without internet addiction: sample 1: n=7596; sample 2: n=1397; sample 3: n=298.

[f]IDSG9-SF: Internet Gaming Disorder Scale–Short Form.

[g]IAT-20: 20-item Young's Internet Addiction Test.

[h]PHQ-9: 9-item Patient Health Questionnaire.

[i]GAD-7: 7-item Generalized Anxiety Disorder.

[j]SCL90-psychoticism: subscale of the Symptom Checklist-90 to measure psychoticism.

[k]SCL90–paranoid ideation: subscale of the Symptom Checklist-90 to measure paranoid ideation.

[l]ASRS: Adult ADHD Self-Report Scale.

[m]For the ASRS in sample 3, there were 88 participants with IGD, 10 with IA, 59 without IGD, and 84 without IA.

Next, we examined whether IGDS9-SF and IAT-20 scores were associated with psychopathological symptom scores. In these analyses, the independent variables were the ZREs of psychopathological symptom scores, adjusted for age and sex; the dependent variable was the ZRE of the IGDS9-SF or IAT-20 score. Univariate models indicated that both IGDS9-SF and

XSL•FO
RenderX

IAT-20 scores were significantly associated with all psychopathological symptom scores. These associations were significantly stronger for the IAT-20 than for the IGDS9-SF, as indicated by higher R^2 values, in S1 and S2. A similar trend was observed in S3, though most differences in associations did not reach statistical significance in this sample. Consistently, multivariate models, in which multicollinearity was acceptable (all tolerances >0.20, 95% CI 0.28 - 0.53, and all variance inflation factors <5, 95% CI 1.55 - 3.55), showed that psychopathological symptom scores were related to a larger proportion of changes in IAT-20 scores than IGDS9-SF scores in S1 (0.33, 95% CI 0.30-0.35 vs 0.13, 95% CI 0.11-0.16) and S2 (0.44, 95% CI 0.39-0.49 vs 0.23, 95% CI 0.18-0.27). A similar trend was observed in S3, although it did not reach statistical significance.

We also performed regression analyses in which IGD or IA status (present vs absent), defined using the prespecified IGDS9-SF and IAT-20 cutoff scores, served as the dependent variables, and the ZREs of psychopathological symptom scores served as the independent variables. In univariate models, both IGD and IA were significantly associated with all psychopathological symptom scores in S1 and S2, and with some of these symptom scores in S3. These associations were significantly stronger for IA than for IGD, as indicated by larger ORs, in S1 and S2. A similar trend was observed in S3, although most differences in ORs did not reach statistical significance in this sample. Multivariate analyses showed a similar pattern in which ORs were generally larger for IA than for IGD, particularly in S1 and S2, although most differences did not reach statistical significance.

Notably, in S2, IAT-20 scores showed a stronger univariate association with attention-deficit or hyperactivity symptom severity (ASRS scores) (R^2, 95% CI; 0.39, 0.33-0.43) than with depression (0.28, 0.23-0.32) or anxiety (0.22, 0.18-0.26). IGDS9-SF scores (R^2, 95% CI) also showed a stronger univariate association with attention deficit or hyperactivity symptom severity (0.19, 0.14-0.23) than with anxiety (0.10, 0.07-0.14). Similar trends were observed in S3, although they did not reach statistical significance. In the logistic regression analyses, univariate models showed that ASRS scores had the strongest associations with IGD and IA among all psychopathological measures in both S2 and S3, although differences in association strength, in terms of ORs, did not reach statistical significance. In multivariate models adjusting for the intercorrelation among other psychopathological measures, higher ASRS scores were independently associated with increased odds of IGD (OR, 95% CI; S2: 2.17, 1.80 - 2.65; S3: 1.43, 0.99-2.09) and IA (S2: 3.18, 2.57 - 3.96; S3: 1.75, 1.19-2.64). By contrast, most associations between other psychopathological symptoms and IGD or IA that were significant in univariate models were no longer statistically significant after multivariate adjustment (Table 3).

XSL•FO
RenderX

Table . Associations of Internet Gaming Disorder Scale–Short Form (IGDS9-SF), 20-item Young's Internet Addiction Test (IAT-20), and identified internet gaming disorder (IGD) and internet addiction (IA) with the severities of other psychopathologies across three samples[a].

Sample and psychopathology	IGDS9-SF[b]		IAT-20[b]		IGD		IA	
	R^2 (95% CI) in univariate model	R^2 (95% CI) in multivariate model[c]	R^2 (95% CI) in univariate model	R^2 (95% CI) in multivariate model[c]	OR[d] (95% CI)	Adjusted OR[c] (95% CI)	OR (95% CI)	Adjusted OR[c] (95% CI)
Sample 1		0.13 (0.11-0.16)		0.33 (0.30-0.35)				
PHQ-9[b,e]	0.10 (0.08-0.12)		0.25 (0.23-0.27)		1.89 (1.76-2.04)	1.33 (1.16-1.53)	2.50 (2.33-2.69)	1.50 (1.34-1.69)
GAD-7[b,f]	0.08 (0.06-0.10)		0.21 (0.19-0.23)		1.76 (1.64-1.89)	1.09 (0.95-1.23)	2.23 (2.09-2.39)	1.22 (1.09-*1.35)*
SCL90-psychoticism[b,g]	0.12 (0.10-0.14)		0.27 (0.25-0.30)		1.87 (1.75-2.00)	1.24 (1.08-1.42)	2.40 (2.25-2.58)	1.54 (1.36-1.74)
SCL90–paranoid ideation[b,h]	0.09 (0.07-0.11)		0.23 (0.21-0.25)		1.88 (1.75-2.01)	1.32 (1.16-1.51)	2.21 (2.07-2.36)	1.16 (1.03-1.30)
Sample 2		0.23 (0.18-0.27)		0.44 (0.39-0.49)				
PHQ-9[b]	0.15 (0.11-0.19)		0.28 (0.23-0.32)		2.20 (1.93-2.52)	1.56 (1.28-1.91)	3.08 (2.66-3.58)	1.89 (1.54-2.31)
GAD-7[b]	*0.10 (0.07-0.14)*		0.22 (0.18-0.26)		1.90 (1.68-2.15)	1.02 (0.84-1.23)	2.47 (2.17-2.83)	1.09 (0.90-1.31)
ASRS[b,i]	0.19 (0.14-0.23)		0.39 (0.33-0.43)		2.85 (2.43-3.38)	2.17 (1.80-2.65)	4.60 (3.80-5.64)	3.18 (2.57-3.96)
Sample 3		0.06 (0.03-0.16)		0.13 (0.06-0.26)				
PHQ-9[b]	0.00 (0.00-0.01)		0.01 (0.00-0.03)		0.93 (0.78-1.12)	0.92 (0.60-1.41)	1.16 (0.98-1.37)	1.21 (0.79-1.85)
GAD-7[b]	0.00 (0.00-0.00)		0.00 (0.00-0.02)		0.93 (0.78-1.11)	0.89 (0.57-1.40)	1.13 (0.95-1.34)	0.85 (0.54-1.33)
SCL90-psychoticism[b]	0.02 (0.00-0.05)		0.10 (0.06-0.16)		1.16 (0.97-1.40)	0.98 (0.59-1.63)	1.71 (1.43-2.06)	1.17 (0.69-1.99)
SCL90–paranoid ideation[b]	0.03 (0.01-0.07)		0.11 (0.06-0.16)		1.32 (1.09-1.59)	1.21 (0.78-1.87)	1.65 (1.37-1.99)	0.80 (0.50-1.24)
ASRS[b]	0.06 (0.01-0.14)		0.11 (0.02-0.22)		1.40 (1.07-1.85)	1.43 (0.99-2.09)	1.73 (1.30-2.36)	1.75 (1.19-2.64)

[a]IGD was defined based on the total score of the IGDS9-SF, with a cutoff value of ≥21. IA was defined based on the total score of IAT-20, with a cutoff value of ≥50. R^2 represents the square of the correlation coefficient (ie, *r*).

[b]Standardized residuals (ZREs) adjusted for the confounding effects of age and sex of measurement scores were used as dependent or independent variables in the regression models.

[c]These multivariate models included ZREs of other psychopathological symptom measures as independent variables.

[d]ORs: odds ratios.

[e]PHQ-9: 9-item Patient Health Questionnaire.

[f]GAD-7: 7-item Generalized Anxiety Disorder.

[g]SCL90-psychoticism: subscale of the Symptom Checklist-90 to measure psychoticism.

[h]SCL90–paranoid ideation: subscale of the Symptom Checklist-90 to measure paranoid ideation.

[i]ASRS: Adult ADHD Self-Report Scale. The ASRS data in S3 were available for only 241 participants because the scale was added midway through the study.

Finally, we classified participants into 4 groups according to IGD and IA status: both IGD and IA, neither condition, IGD only, or IA only. ANOVA tests indicated that ZREs of all measured psychopathological symptoms differed significantly

XSL•FO
RenderX

across the 4 groups in all 3 samples. Post hoc analyses in S1 revealed a significant stepwise increase in depressive, anxiety, psychoticism, and paranoid ideation symptom severity from participants with neither disorder to those with IGD only, then to those with IA only, and finally to those with both disorders. Similar trends were observed in S2 for the severity of depression, anxiety, and attention-deficit or hyperactivity, although the differences in depressive and anxiety symptom severity between the IA-only group and the comorbid group were not statistically significant. In S3, participants with IA (with or without comorbid IGD) tended to have higher psychopathological symptom scores than those with neither disorder nor IGD only, although some post hoc comparisons did not reach statistical significance (Figure 1).

Figure 1. Post hoc comparisons of psychopathological symptom severity (z-standardized residuals adjusted for age and sex) among participants with neither internet gaming disorder (IGD) nor internet addiction (IA; group A), IGD only (group B), IA only (group C), or comorbid IGD and IA (group D) across 3 independent samples: S1 (first-year undergraduates at Sichuan University), S2 (high school students recruited in Hangzhou), and S3 (inpatients aged 13 - 19 years recruited from 2 tertiary mental health centers in Hangzhou and Chengdu). ASRS: Adult ADHD Self-Report Scale; GAD-7: 7-item Generalized Anxiety Disorder; PHQ-9: 9-item Patient Health Questionnaire; SCL-90-paranoid: subscale of the Symptom Checklist-90 to measure paranoid ideation; SCL-90-psychotism: subscale of the Symptom Checklist-90 to measure psychoticism. ***P*<.01; ****P*<.001.

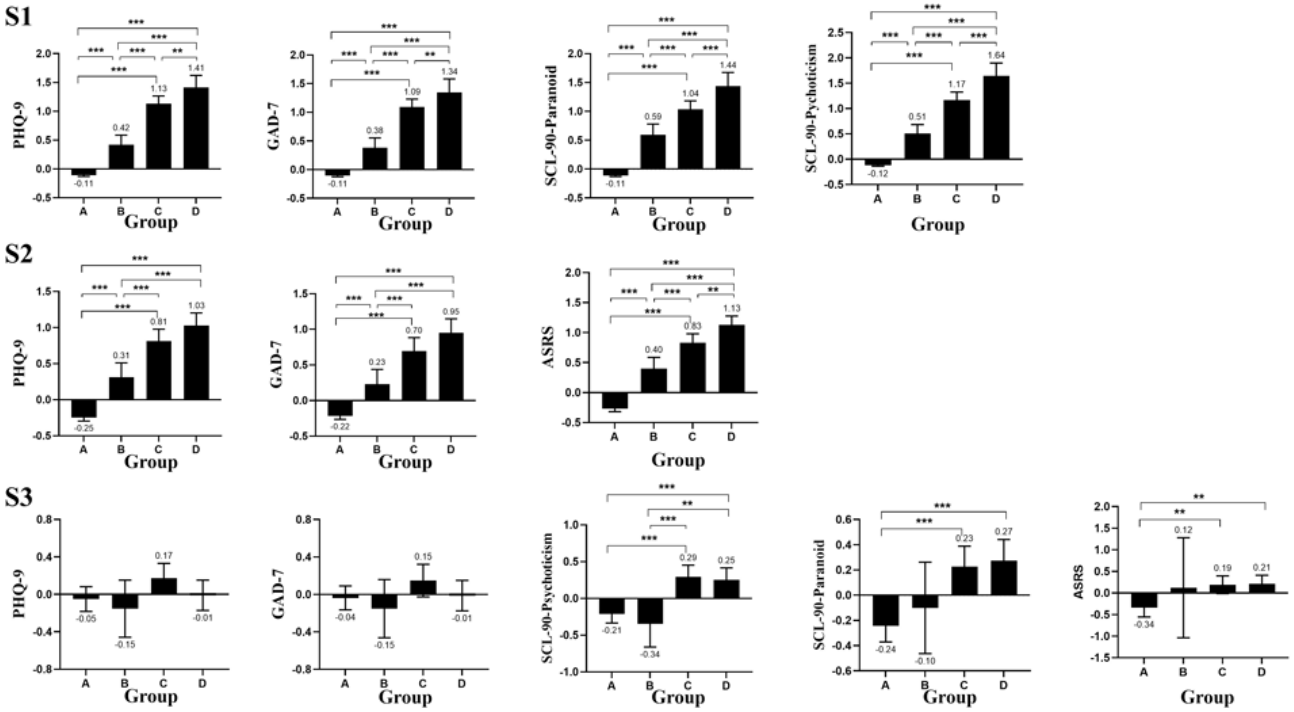

Discussion

Findings

To our knowledge, this is the first study to examine a potential relationship between IGD and IA, as well as to compare their associations with a broad range of psychopathological symptoms within the same samples. In addition, we cross-validated the main findings across 3 independent samples. Across samples, we found that IGD and IA were distinct yet moderately correlated and that IA was more prevalent and more strongly associated with psychopathological symptom severity.

These key findings have important implications for clarifying the nosological relationship between IGD and IA. The moderate association we detected between the 2 disorders was similar to that reported between so-called "problematic internet use" and "problematic online gaming" [20], but it was weaker than the associations between depressive and anxiety symptoms observed in this study and in previous work in other Chinese samples [34,35]. These observations support considering IGD and IA as distinct yet correlated entities, rather than as a single entity or as a subtype of a broader construct [21]. Consistent with this interpretation, we identified participants who met the most widely accepted definition of IGD but not IA, and vice versa. In our sample, IA, whether occurring alone or comorbid with IGD, was associated with more severe psychopathological symptoms than IGD alone. Furthermore, our univariate correlation and multivariate regression analyses indicated that IA severity was consistently more strongly associated with psychopathological symptom severity than IGD severity. These findings suggest that compulsive, problematic online behaviors extending beyond gaming warrant greater attention. Although "gaming" is currently the only type of online activity codified in the *International Classification of Diseases* and listed as a condition for further study in the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (*DSM-5*), other forms of online entertainment, such as social media and algorithm-driven short-video platforms, may pose comparable or even greater risks to mental health and may undermine psychological well-being, academic achievement, and family dynamics [36,37]. It may be more appropriate to refer to broader constructs such as "internet entertainment disorder" or "internet entertainment addiction" to better capture the spectrum of addictive behaviors related to internet use. These constructs may require further expansion or adaptation to reflect the growing use of immersive technologies, such as virtual or

XSL•FO
RenderX

augmented reality and artificial intelligence, which may further facilitate addictive behaviors.

In our study, both IGD and IA were more strongly associated with symptoms of attention-deficit or hyperactivity symptoms than with depressive or anxiety symptoms, consistent with findings from a study of adults aged 20 - 40 years in Taiwan [38]. These findings suggest that ADHD-related inattention and impulsivity, as indexed by ASRS scores, may partially account for the observed associations between IGD or IA and other psychopathological symptoms. Evidence from adolescents in several countries further indicates that attention-deficit or hyperactivity disorder frequently co-occurs with IGD and IA and may even predict their onset [39]. Core features of attention-deficit or hyperactivity, including impulsivity, inattention, heightened sensation seeking, and poor regulatory control, may increase the risk of problematic internet use [40-42]. More broadly, both disorders may be more strongly associated with externalizing symptoms characteristic of attention-deficit or hyperactivity than with internalizing symptoms, such as depression or anxiety, a pattern supported by recent meta-analytic evidence, although the pooled effect sizes were modest [43]. This closer association with externalizing symptoms may be related to the immediate gratification and heightened sensory stimulation inherent in gaming and other forms of online entertainment [44,45]. In contrast, internalizing symptoms may instead reflect maladaptive coping mechanisms for psychological distress [46,47].

We consider our data to be reliable because we assessed IGD and IA using widely validated, commonly used instruments and because we were able to replicate key associations across the 3 independent samples of adolescents. In addition, the differences in psychopathological symptoms observed across samples were consistent with expectations: hospitalized adolescents exhibited high levels of psychopathological symptoms similar to those reported in a study of Caucasian adolescents [48], and these levels were higher than those in high school students, which in turn were higher than those observed in undergraduates. Our high school students may have experienced heightened stress due to the upcoming national university entrance examination (gaokao), which may help explain their higher levels of depression and anxiety [49]. By contrast, our first-year undergraduates had already performed sufficiently well on the entrance examination to secure admission to a top-tier university. This may have contributed to their lower levels of depressive and anxiety symptoms. More generally, students admitted to top-tier universities in China may possess more effective coping strategies and problem-solving skills, which could buffer against psychopathological symptoms [50].

In our samples, male sex appeared to be associated with a higher risk of internet gaming disorder but a lower risk of IA, consistent with previous work in Chinese and US populations in which the 2 disorders were conceptualized as separate constructs [51,52]. Future research should examine whether and through what mechanisms sex influences the risk of either disorder, particularly given well-documented sex differences in internet use: male participants tend to engage more in computer gaming, whereas female participants tend to engage more in social networking and social media [53]. Previous studies have reported inconsistent findings regarding whether sex influences the risk of IGD or IA [54,55]. These inconsistencies may reflect differences in the relative proportions of individuals with IGD only, IA only, or comorbid IGD and IA.

Limitations

Our findings should be interpreted with caution in light of several limitations. First, the cross-sectional design of this study precludes causal inference. Second, all data were obtained from self-report questionnaires, which may increase the risk of social desirability and recall biases. Third, the sample comprised exclusively Chinese participants residing in China, which may limit the generalizability of our findings to other cultural and geographic contexts. In addition, the inpatient sample size was relatively small, and many participants scored near the maximum possible values on several symptom scales, a phenomenon known as the ceiling effect. These factors may have reduced statistical power and, together with other sample-specific characteristics (eg, differences in sex ratios across samples), may help explain why several associations observed in the university and high school samples did not reach statistical significance in the inpatient sample. Nevertheless, most analyses conducted among inpatients showed patterns similar to those observed in the other 2 samples.

It should also be noted that our findings reflect IGD as a global construct without differentiating between specific game genres (eg, real-time strategy, massively multiplayer online role-playing games, sports games, or first-person shooters). Given evidence that the prevalence and psychological correlates of IGD may vary by game genre [56,57], future studies should systematically characterize predominant game types and examine genre-specific associations with psychopathology. In addition, the measures of psychopathological symptoms included in the 3 samples were not entirely consistent across the 3 samples (eg, the absence of SCL-90 in S2 and ASRS in S1), due to considerations such as survey timing constraints and primary study objectives. This inconsistency may have reduced the comparability of certain results (eg, associations involving ASRS) across samples. However, the key findings—namely, that IGD and IA are distinct yet moderately correlated constructs, with IA more strongly associated with the severity of psychopathological symptoms—were robust and consistent across all 3 samples.

Conclusions

This study provides additional evidence that IGD and IA represent distinct yet interrelated constructs and further demonstrates that IA consistently exhibits a stronger association with the severity of psychopathological symptoms than IGD. These findings underscore the importance of recognizing and addressing compulsive and problematic online behaviors that extend beyond gaming, contributing to ongoing debates regarding the classification and clinical significance of behavioral addictions related to internet use, and highlighting the need for further refinement of diagnostic frameworks and the prioritization of targeted, evidence-based clinical interventions.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank all the patients who participated in this study. The authors declare the use of generative artificial intelligence (GAI) in the writing process. According to GAIDeT (Generative AI Delegation Taxonomy; 2025), translation was delegated to GAI tools under full human supervision. The GAI tool used was ChatGPT (OpenAI, GPT-4.5). Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was supported by the National Natural Science Foundation of China (grant No. 82171487), “Pioneer” and “Leading Goose” R&D Program of Zhejiang (2024C03006), and the Leading Innovation and Entrepreneurship Team of Hangzhou (TD2024003). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

YYL and WJG conceived and designed the study and drafted the manuscript. YYL, AQH, and LLY curated and analyzed the data. AQH and WW performed data visualization, YYL, LLY, AQH, JW, and ZXM collected the data. QYL, WJD, MXQ, and JJX validated the results. YQH, SH, FGL, and XJL provided resources. JJX, QW, XJL, TL, and WJG supervised the project. TL and WJG acquired funding and administered the project. QW, WD, YZH, TL, and WJG reviewed and edited the manuscript. All authors have full access to the study dataset and are responsible for ensuring its integrity and the accuracy of the resulting analyses.

Conflicts of Interest

None declared.

References

1. Statistics. International Telecommunication Union. 2023. URL: https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx [accessed 2026-01-20]
2. León Méndez M, Padrón I, Fumero A, Marrero RJ. Effects of internet and smartphone addiction on cognitive control in adolescents and young adults: a systematic review of fMRI studies. Neurosci Biobehav Rev 2024 Apr;159:105572. [doi: 10.1016/j.neubiorev.2024.105572] [Medline: 38320657]
3. Bersani FS, Barchielli B, Ferracuti S, et al. The association of problematic use of social media and online videogames with aggression is mediated by insomnia severity: a cross-sectional study in a sample of 18- to 24-year-old individuals. Aggress Behav 2022 May;48(3):348-355. [doi: 10.1002/ab.22008] [Medline: 34870339]
4. Kim HS, Son G, Roh EB, et al. Prevalence of gaming disorder: a meta-analysis. Addict Behav 2022 Mar;126:107183. [doi: 10.1016/j.addbeh.2021.107183] [Medline: 34864436]
5. Soriano-Molina E, Limiñana-Gras RM, Patró-Hernández RM, Rubio-Aparicio M. The association between internet addiction and adolescents’ mental health: a meta-analytic review. Behav Sci (Basel) 2025 Jan 23;15(2):116. [doi: 10.3390/bs15020116] [Medline: 40001747]
6. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5): American Psychiatric Association; 2013. [doi: 10.1176/appi.books.9780890425596]
7. ICD-11: International Classification of Diseases 11th Revision. 2022. URL: https://icd.who.int/en/ [accessed 2026-01-20]
8. King DL, Chamberlain SR, Carragher N, et al. Screening and assessment tools for gaming disorder: a comprehensive systematic review. Clin Psychol Rev 2020 Apr;77:101831. [doi: 10.1016/j.cpr.2020.101831] [Medline: 32143109]
9. Rahman M, Sarkar P, Haque SE, et al. Technology addiction: effects of electronic games and social media use on academic performance and symptoms of psychiatric disorders among school-age adolescents. Health Sci Rep 2025 Jul;8(7):e71045. [doi: 10.1002/hsr2.71045] [Medline: 40687541]
10. Severo RB, Soares JM, Affonso JP, et al. Prevalence and risk factors for internet gaming disorder. Braz J Psychiatry 2020;42(5):532-535. [doi: 10.1590/1516-4446-2019-0760] [Medline: 32785455]
11. Wartberg L, Kriston L, Kramer M, Schwedler A, Lincoln TM, Kammerl R. Internet gaming disorder in early adolescence: associations with parental and adolescent mental health. Eur Psychiatry 2017 Jun;43:14-18. [doi: 10.1016/j.eurpsy.2016.12.013] [Medline: 28365463]
12. Wang CY, Wu YC, Su CH, Lin PC, Ko CH, Yen JY. Association between Internet gaming disorder and generalized anxiety disorder. J Behav Addict 2017 Dec 1;6(4):564-571. [doi: 10.1556/2006.6.2017.088] [Medline: 29280398]

XSL•FO
RenderX

13. Falcione K, Weber R. Psychopathology and gaming disorder in adolescents. JAMA Netw Open 2025 Jul 1;8(7):e2528532. [doi: 10.1001/jamanetworkopen.2025.28532] [Medline: 40728787]
14. Petry NM, Zajac K, Ginley MK. Behavioral addictions as mental disorders: to be or not to be? Annu Rev Clin Psychol 2018 May 7;14:399-423. [doi: 10.1146/annurev-clinpsy-032816-045120] [Medline: 29734827]
15. Young KS. Caught in the Net: How to Recognize the Signs of Internet Addiction—and a Winning Strategy for Recovery: John Wiley & Sons; 1998. URL: https://www.wiley.com/en-us/Caught+in+the+Net%3A+How+to+Recognize+the+Signs+of+Internet+Addiction–and+a+Winning+Strategy+for+Recovery-p-9780471191599 [accessed 2026-01-20]
16. Meng SQ, Cheng JL, Li YY, et al. Global prevalence of digital addiction in general population: a systematic review and meta-analysis. Clin Psychol Rev 2022 Mar;92:102128. [doi: 10.1016/j.cpr.2022.102128] [Medline: 35150965]
17. Guo W, Tao Y, Li X, et al. Associations of internet addiction severity with psychopathology, serious mental illness, and suicidality: large-sample cross-sectional study. J Med Internet Res 2020 Aug 11;22(8):e17560. [doi: 10.2196/17560] [Medline: 32780029]
18. Kitazawa M, Yoshimura M, Murata M, et al. Associations between problematic internet use and psychiatric symptoms among university students in Japan. Psychiatry Clin Neurosci 2018 Jul;72(7):531-539. [doi: 10.1111/pcn.12662] [Medline: 29652105]
19. Thorens G, Achab S, Billieux J, et al. Characteristics and treatment response of self-identified problematic internet users in a behavioral addiction outpatient clinic. J Behav Addict 2014 Mar;3(1):78-81. [doi: 10.1556/JBA.3.2014.008] [Medline: 25215217]
20. Király O, Griffiths MD, Urbán R, et al. Problematic internet use and problematic online gaming are not the same: findings from a large nationally representative adolescent sample. Cyberpsychol Behav Soc Netw 2014 Dec;17(12):749-754. [doi: 10.1089/cyber.2014.0475] [Medline: 25415659]
21. Young KS. Psychology of computer use: XL. Addictive use of the Internet: a case that breaks the stereotype. Psychol Rep 1996 Dec;79(3 Pt 1):899-902. [doi: 10.2466/pr0.1996.79.3.899] [Medline: 8969098]
22. Moreno M, Riddle K, Jenkins MC, Singh AP, Zhao Q, Eickhoff J. Measuring problematic internet use, internet gaming disorder, and social media addiction in young adults: cross-sectional survey study. JMIR Public Health Surveill 2022 Jan 27;8(1):e27719. [doi: 10.2196/27719] [Medline: 34081596]
23. Sánchez-Fernández M, Borda-Mas M, Horvath Z, Demetrovics Z. Similarities and differences in the psychological factors associated with generalised problematic internet use, problematic social media use, and problematic online gaming. Compr Psychiatry 2024 Oct;134:152512. [doi: 10.1016/j.comppsych.2024.152512] [Medline: 38955108]
24. Severo RB, Barbosa APPN, Fouchy DRC, et al. Development and psychometric validation of Internet Gaming Disorder Scale-Short-Form (IGDS9-SF) in a Brazilian sample. Addict Behav 2020 Apr;103:106191. [doi: 10.1016/j.addbeh.2019.106191] [Medline: 31887719]
25. Young KS, de Abreu CN, editors. Internet Addiction: A Handbook and Guide to Evaluation and Treatment: John Wiley & Sons P&T; 2011. URL: https://www.vitalsource.com/products/internet-addiction-v9780470892244?srsltid=AfmBOor1sLx0xQsKzvwxfzRwAyYkPEBTF3lqPhLP6JTMTZvhFugoBrS_ [accessed 2026-01-20]
26. Durkee T, Kaess M, Carli V, et al. Prevalence of pathological internet use among adolescents in Europe: demographic and social factors. Addiction 2012 Dec;107(12):2210-2222. [doi: 10.1111/j.1360-0443.2012.03946.x] [Medline: 22621402]
27. Endomba FT, Demina A, Meille V, et al. Prevalence of internet addiction in Africa: a systematic review and meta-analysis. J Behav Addict 2022 Sep 26;11(3):739-753. [doi: 10.1556/2006.2022.00052] [Medline: 35984734]
28. Yu X, Tam WWS, Wong PTK, Lam TH, Stewart SM. The Patient Health Questionnaire-9 for measuring depressive symptoms among the general population in Hong Kong. Compr Psychiatry 2012 Jan;53(1):95-102. [doi: 10.1016/j.comppsych.2010.11.002] [Medline: 21193179]
29. Löwe B, Decker O, Müller S, et al. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. Med Care 2008 Mar;46(3):266-274. [doi: 10.1097/MLR.0b013e318160d093] [Medline: 18388841]
30. Preti A, Carta MG, Petretto DR. Factor structure models of the SCL-90-R: replicability across community samples of adolescents. Psychiatry Res 2019 Feb;272:491-498. [doi: 10.1016/j.psychres.2018.12.146] [Medline: 30611969]
31. Diagnostic and Statistical Manual of Mental Disorders, 4th Edition: American Psychiatric Association; 2022. [doi: 10.1176/appi.books.9780890425787]
32. Yeh CB, Gau SSF, Kessler RC, Wu YY. Psychometric properties of the Chinese version of the adult ADHD Self-report Scale. Int J Methods Psychiatr Res 2008;17(1):45-54. [doi: 10.1002/mpr.241] [Medline: 18286465]
33. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
34. Xie Y, Tang L. The symptom network of internet gaming addiction, depression, and anxiety among children and adolescents. Sci Rep 2024 Nov 29;14(1):29732. [doi: 10.1038/s41598-024-81094-7] [Medline: 39614079]

XSL•FO
RenderX

35. Teng Z, Pontes HM, Nie Q, Griffiths MD, Guo C. Depression and anxiety symptoms associated with internet gaming disorder before and during the COVID-19 pandemic: a longitudinal study. J Behav Addict 2021 Mar 10;10(1):169-180. [doi: 10.1556/2006.2021.00016] [Medline: 33704085]
36. Chao M, Lei J, He R, Jiang Y, Yang H. TikTok use and psychosocial factors among adolescents: comparisons of non-users, moderate users, and addictive users. Psychiatry Res 2023 Jul;325:115247. [doi: 10.1016/j.psychres.2023.115247] [Medline: 37167877]
37. Feng T, Wang B, Mi M, et al. The relationships between mental health and social media addiction, and between academic burnout and social media addiction among Chinese college students: a network analysis. Heliyon 2025 Feb 15;11(3):e41869. [doi: 10.1016/j.heliyon.2025.e41869] [Medline: 39959490]
38. Yen JY, Király O, Griffiths MD, Demetrovics Z, Ko CH. A case-control study for psychiatric comorbidity and associative factors of gaming disorder and hazardous gaming based on ICD-11 criteria: cognitive control, emotion regulation, and reinforcement sensitivity. J Behav Addict 2024 Dec 30;13(4):1014-1027. [doi: 10.1556/2006.2024.00066] [Medline: 39636323]
39. Peeters M, Koning I, van den Eijnden R. Predicting internet gaming disorder symptoms in young adolescents: a one-year follow-up study. Comput Human Behav 2018 Mar;80:255-261. [doi: 10.1016/j.chb.2017.11.008]
40. Dullur P, Krishnan V, Diaz AM. A systematic review on the intersection of attention-deficit hyperactivity disorder and gaming disorder. J Psychiatr Res 2021 Jan;133:212-222. [doi: 10.1016/j.jpsychires.2020.12.026] [Medline: 33360866]
41. Wang BQ, Yao NQ, Zhou X, Liu J, Lv ZT. The association between attention deficit/hyperactivity disorder and internet addiction: a systematic review and meta-analysis. BMC Psychiatry 2017 Jul 19;17(1):260. [doi: 10.1186/s12888-017-1408-x] [Medline: 28724403]
42. Koronczai B, Kökönyei G, Griffiths MD, Demetrovics Z. The relationship between personality traits, psychopathological symptoms, and problematic internet use: a complex mediation model. J Med Internet Res 2019 Apr 26;21(4):e11837. [doi: 10.2196/11837] [Medline: 31025955]
43. Eirich R, McArthur BA, Anhorn C, McGuinness C, Christakis DA, Madigan S. Association of screen time with internalizing and externalizing behavior problems in children 12 years or younger: a systematic review and meta-analysis. JAMA Psychiatry 2022 May 1;79(5):393-405. [doi: 10.1001/jamapsychiatry.2022.0155] [Medline: 35293954]
44. Wang L, Wu L, Lin X, et al. Dysfunctional default mode network and executive control network in people with Internet gaming disorder: independent component analysis under a probability discounting task. Eur Psychiatry 2016 Apr;34:36-42. [doi: 10.1016/j.eurpsy.2016.01.2424] [Medline: 26928344]
45. Ko CH, Wang PW, Liu TL, Chen CS, Yen CF, Yen JY. The adaptive decision-making, risky decision, and decision-making style of Internet gaming disorder. Eur Psychiatry 2017 Jul;44:189-197. [doi: 10.1016/j.eurpsy.2017.05.020] [Medline: 28646731]
46. Kardefelt-Winther D. A conceptual and methodological critique of internet addiction research: towards a model of compensatory internet use. Comput Human Behav 2014 Feb;31:351-354. [doi: 10.1016/j.chb.2013.10.059]
47. Miao S, Xu L, Gao S, Bai C, Huang Y, Peng B. The association between anxiety and internet addiction among left-behind secondary school students: the moderating effect of social support and family types. BMC Psychiatry 2024 May 30;24(1):406. [doi: 10.1186/s12888-024-05855-4] [Medline: 38811914]
48. Torres-Rodríguez A, Griffiths MD, Carbonell X, Oberst U. Internet gaming disorder in adolescence: psychological characteristics of a clinical sample. J Behav Addict 2018 Sep 1;7(3):707-718. [doi: 10.1556/2006.7.2018.75] [Medline: 30264606]
49. Zhou J, Liu Y, Ma J, et al. Prevalence of depressive symptoms among children and adolescents in China: a systematic review and meta-analysis. Child Adolesc Psychiatry Ment Health 2024 Nov 19;18(1):150. [doi: 10.1186/s13034-024-00841-w] [Medline: 39563377]
50. Zhang W, Gao W, Liu X. Does attending elite colleges matter in the relationship between self-esteem and general self-efficacy of students in China? Heliyon 2022 Jun;8(6):e09723. [doi: 10.1016/j.heliyon.2022.e09723] [Medline: 35756109]
51. Peng C, Guo T, Cheng J, et al. Sex differences in association between Internet addiction and aggression among adolescents aged 12 to 18 in mainland of China. J Affect Disord 2022 Sep 1;312:198-207. [doi: 10.1016/j.jad.2022.06.026] [Medline: 35728679]
52. Ohayon MM, Roberts L. Internet gaming disorder and comorbidities among campus-dwelling U.S. university students. Psychiatry Res 2021 Aug;302:114043. [doi: 10.1016/j.psychres.2021.114043] [Medline: 34129998]
53. Chen B, Liu F, Ding S, Ying X, Wang L, Wen Y. Gender differences in factors associated with smartphone addiction: a cross-sectional study among medical college students. BMC Psychiatry 2017 Oct 10;17(1):341. [doi: 10.1186/s12888-017-1503-z] [Medline: 29017482]
54. Zhuang X, Zhang Y, Tang X, Ng TK, Lin J, Yang X. Longitudinal modifiable risk and protective factors of internet gaming disorder: a systematic review and meta-analysis. J Behav Addict 2023 Jun 29;12(2):375-392. [doi: 10.1556/2006.2023.00017] [Medline: 37224007]
55. Kiviruusu O. Excessive internet use among Finnish young people between 2017 and 2021 and the effect of COVID-19. Soc Psychiatry Psychiatr Epidemiol 2024 Dec;59(12):2291-2301. [doi: 10.1007/s00127-024-02723-0] [Medline: 38985326]

XSL•FO
RenderX

56. Na E, Choi I, Lee TH, et al. The influence of game genre on Internet gaming disorder. J Behav Addict 2017 Jun 29;6(2):1-8. [doi: 10.1556/2006.6.2017.033] [Medline: 28658960]
57. Kim D, Nam JK, Keum C. Adolescent Internet gaming addiction and personality characteristics by game genre. PLoS ONE 2022;17(2):e0263645. [doi: 10.1371/journal.pone.0263645]

Abbreviations

ASRS: Adult ADHD Self-Report Scale
***DSM-5*:** *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*
GAD-7: 7-item Generalized Anxiety Disorder
IA: internet addiction
IAT-20: 20-item Young's Internet Addiction Test
***ICD-11*:** *International Classification of Diseases, 11th Revision*
IGD: internet gaming disorder
IGDS9-SF: Internet Gaming Disorder Scale–Short Form
OR: odds ratio
PHQ-9: 9-item Patient Health Questionnaire
S1: sample 1
S2: sample 2
S3: sample 3
ZRE: standardized residual

Edited by S Brini; submitted 14.Aug.2025; peer-reviewed by KW Tay, X Xiang, Y Feng; accepted 07.Jan.2026; published 03.Feb.2026.

Please cite as:
Li YY, Hu AQ, Yi LL, Mao ZX, Lü QY, Wang J, Wei W, Huang YQ, Huang S, Dai WJ, Qiao MX, Xu JJ, Wang Q, Li XJ, Luo FG, Deng W, Hu YZ, Li T, Guo WJ
Comparing the Associations of Internet Addiction and Internet Gaming Disorder With Psychopathological Symptoms: Cross-Sectional Study of Three Independent Adolescent Samples
J Med Internet Res 2026;28:e82414
URL: https://www.jmir.org/2026/1/e82414
doi:10.2196/82414

© Ying-ying Li, A-qian Hu, Ling-li Yi, Zi-xin Mao, Qiu-yue Lü, Juan Wang, Wei Wei, Yue-qi Huang, Shu Huang, Wen-jing Dai, Meng-xuan Qiao, Jia-jun Xu, Qiang Wang, Xiao-jing Li, Fu-gang Luo, Wei Deng, Yu-zheng Hu, Tao Li, Wan-jun Guo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 3.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Phenotyping for Adolescent Mental Health: Feasibility Study Using Machine Learning to Predict Mental Health Risk From Active and Passive Smartphone Data

Balasundaram Kadirvelu[1*], PhD; Teresa Bellido Bel[2*], MD; Aglaia Freccero[2], MSc; Martina Di Simplico[2], PhD, MD; Dasha Nicholls[2], MD; A Aldo Faisal[1,3], PhD

[1]Brain & Behaviour Lab, Department of Computing and Department of Bioengineering, Imperial College London, Royal School of Mines, London, United Kingdom

[2]Division of Psychiatry, Department of Brain Sciences, Imperial College London, London, England, United Kingdom

[3]Chair in Digital Health & Data Science, Faculty of Life Sciences, University of Bayreuth, Bayreuth, Bavaria, Germany

[*]these authors contributed equally

Corresponding Author:
A Aldo Faisal, PhD
Brain & Behaviour Lab, Department of Computing and Department of Bioengineering, Imperial College London, Royal School of Mines, London, United Kingdom

Abstract

Background: Adolescents are particularly vulnerable to mental disorders, with over 75% of lifetime cases emerging before the age of 25 years. Yet most young people with significant symptoms do not seek support. Digital phenotyping, leveraging active (self-reported) and passive (sensor-based) data from smartphones, offers a scalable, low-burden approach for early risk detection. Despite this potential, its application in school-going adolescents from general (nonclinical) populations remains limited, leaving a critical gap in community-based prevention efforts.

Objective: This study evaluated the feasibility of using a smartphone app to predict mental health risks in nonclinical adolescents by integrating active and passive data streams within a machine learning (ML) framework. We examined the utility of this approach for identifying risks related to internalizing and externalizing difficulties, eating disorders, insomnia, and suicidal ideation.

Methods: Participants (n=103; mean age 16.1 years, SD 1.0) from 3 UK secondary schools used the Mindcraft app (Brain and Behaviour Lab) for 14 days, providing daily self-reports (eg, mood, sleep, and loneliness) and continuous passive sensor data (eg, location, step count, and app usage). We developed a deep learning model incorporating contrastive pretraining with triplet margin loss to stabilize user-specific behavioral patterns, followed by supervised fine-tuning for binary classification of 4 mental health outcomes, namely, the Strengths and Difficulties Questionnaire (SDQ)-high risk, insomnia, suicidal ideation, and eating disorder. Performance was assessed using leave-one-subject-out cross-validation (LOSO-CV), with balanced accuracy as the primary metric. Comparative analyses were conducted using CatBoost (Yandex) and multilayer perceptron (MLP) models without pretraining. Feature importance was assessed using Shapley Additive Explanations (SHAP) values, and associations between key digital features and clinical scales were analyzed.

Results: Integration of active and passive data outperformed single-modality models, achieving mean balanced accuracies of 0.71 (0.03) for SDQ-high risk, 0.67 (0.04) for insomnia, 0.77 (0.03) for suicidal ideation, and 0.70 (0.03) for eating disorder. The contrastive learning approach improved representation stability and predictive robustness. SHAP analysis highlighted clinically relevant features, such as negative thinking and location entropy, underscoring the complementary value of combining subjective and objective data. Correlation analyses confirmed meaningful associations between key digital features and mental health outcomes. Performance in an independent external validation cohort (n=45) achieved balanced accuracies of 0.63 - 0.72 across outcomes, suggesting generalizability to new settings.

Conclusions: This study demonstrates the feasibility and utility of smartphone-based digital phenotyping for predicting mental health risks in nonclinical, school-going adolescents. By integrating active and passive data with advanced machine modeling techniques, this approach shows promise for early detection and scalable intervention strategies in community settings.

(J Med Internet Res 2026;28:e72501) doi:10.2196/72501

XSL•FO
RenderX

KEYWORDS

youth mental health; digital health; mobile applications; mobile health; mHealth; smartphone sensing; ecological momentary assessment; EMA; artificial intelligence; early intervention

Introduction

Children and young people are particularly vulnerable to mental health problems due to critical developmental changes in emotion, behavior, and cognition [1], with over 75% of mental disorders emerging before the age of 25 years [2]. The World Health Organization estimates that one in 7 adolescents aged 10 - 19 years at the time of this writing lives with a diagnosable mental disorder [3].

Adolescent mental health symptoms are typically classified as internalizing (eg, anxiety, depression, and suicidal thoughts) or externalizing (eg, aggression and impulsivity) [4,5]. These symptoms can impair social, academic, and interpersonal functioning, and if left unaddressed, may lead to long-term psychiatric disorders [6]. Anxiety, depression, and eating disorders are among the most prevalent conditions, with sleep problems often acting as both a symptom and a risk factor [7,8]. Among the most severe manifestations of internalizing psychopathology is suicidal ideation, a particularly serious concern during adolescence; when unaddressed, it may progress to suicide, which ranks as the third leading cause of death globally among individuals aged 15-29 years [3]. Yet only 18% - 34% of young people with significant symptoms seek professional help [9]. This critical gap in receiving care underscores an urgent need for scalable, accessible, and youth-friendly solutions in mental health care.

Given the central role of schools in adolescents' daily lives, digital health interventions in schools offer a promising, cost-effective, and scalable way to support mental well-being in children and young people [10,11]. The proliferation of smartphones has enabled a new class of digital mental health interventions that leverage their unique data collection capabilities [12-14]. Smartphones can gather active data (subjective self-reports of behaviors and experiences) and passive data (sensor-based measures such as GPS, step count, and ambient light). These behavioral markers, collectively termed "digital phenotyping," reveal dynamic interactions between individuals and their environments [15-17].

Active data capture internal states such as mood, sleep quality, or loneliness, but it depends on user engagement and is vulnerable to recall bias. In contrast, passive data offer continuous, objective insights into daily routines, capturing behavioral patterns that may reflect underlying mental health states. For instance, lower location entropy and reduced movement are associated with depression [18], fewer steps with lower mood [19,20], and variations in ambient light with circadian disruption [21]. However, passive data alone may miss crucial psychological context and can be challenging to interpret in isolation [22].

Although several studies [23-41] have explored either active or passive data for mental health monitoring, few have examined their integration, particularly in community-based adolescent populations. Combining these modalities provides a more comprehensive view of mental health, connecting how young people feel with how they behave in real-world settings [42]. This fusion can enhance model robustness, uncover hidden patterns, and address mismatches between self-report and behavior. For example, a young person may report feeling fine (active data) while showing signs of social withdrawal or sleep disruption (passive data), a critical challenge for unimodal approaches. Multimodal approaches are better suited to detecting such discrepancies, enabling earlier and more nuanced detection in nonclinical settings [42,43].

The complexity of digital phenotyping data, such as its high dimensionality, multimodal nature, and nonlinear interaction patterns, poses challenges for traditional statistical approaches, which rely on strong parametric assumptions and may require extensive a priori feature engineering. In contrast, machine learning (ML) can model complex, heterogeneous data and automatically learn nonlinear latent patterns without predefined hypotheses [44,45]. This makes ML particularly well-suited to adolescent mental health prediction using digital data, where subtle behavioral signals may be distributed across multiple features and modalities [17].

Despite growing interest in applying ML to digital mental health [23-41], existing studies often present one or more limitations in the context of adolescent-focused research: (1) they are primarily conducted in adult populations, limiting relevance to younger age groups such as adolescents; (2) they typically involve clinical samples, reducing generalizability to community-based or non–help-seeking populations; (3) they focus on a single condition (eg, depression or anxiety); and (4) they rely on either active or passive data, but not both. These limitations constrain the generalizability, robustness, and real-world utility of current approaches for children and young people.

This study advances the field in several ways. First, it uses a nonclinical adolescent sample, enabling evaluation in a real-world prevention context, which is critical for early detection and scalable screening. Second, it investigates 4 distinct mental health outcomes (internalizing and externalizing difficulties, insomnia, eating disorder risk, and suicidal ideation) rather than focusing on a single condition. Third, it integrates both subjective (active) and objective (passive) smartphone data to develop multimodal predictive models that capture patterns that may be missed when using either modality alone. Finally, it uses a contrastive learning framework, a novel ML technique for learning stable behavioral representations, enhancing generalizability.

Specifically, this study addresses the following research questions: (1) Can integrating active and passive smartphone data improve the prediction of mental health risks in a nonclinical adolescent population using ML? (2) How well does this multimodal approach generalize across different mental health outcomes? (3) Which digital features are most relevant to predicting specific mental health outcomes?

XSL•FO
RenderX

Methods

Recruitment and Data Collection

Participants were recruited from secondary schools in northwest London between November 2022 and July 2023. We contacted schools via email and followed up via telephone. Three schools that expressed interest in taking part in our study were recruited. The inclusion criteria were young people aged 14 - 18 years attending 10-13 years with a sufficient level of English to respond to the study instrument and use the app and who had access to an iOS- or Android-compatible smartphone.

Students initially completed an online survey accessed via a Qualtrics link included in the promotional materials. This survey began with the Strengths and Difficulties Questionnaire (SDQ) [46], a screening tool whose predictions have been largely consistent with clinical diagnoses with good levels of internal consistency and test-retest stability. To detect eating disorders, we included the Eating Disorder-15 Questionnaire (ED - 15), which has been described as a valuable tool to assess eating disorder psychopathology in young individuals quickly [47]. We excluded the compensatory behaviors section to simplify the data collection process. Its ability to detect changes early in treatment means that it could be used as a routine outcome measure within therapeutic contexts. We also incorporated a question from the Patient Health Questionnaire version 9 (PHQ-9) [48], which is validated for young people, to identify suicidal ideation [49]. Finally, we used the Sleep Condition Indicator (SCI), a brief scale to evaluate insomnia disorder in everyday clinical practice [50,51].

Upon completing the online survey, participants received a link to download the Mindcraft app (Brain and Behaviour Lab) [52] from the App Store or Play Store, along with a unique login. Participants were asked to use the app for at least 2 weeks. The Mindcraft app is a user-friendly mobile app designed to collect self-reported well-being updates (active data) and phone sensor data (passive data). Participants set their data-sharing preferences during onboarding and can adjust them at any time through the app's settings. Detailed technical specifications of the Mindcraft app are available in the reference [52].

Active and Passive Data Features

Once participants began using the app, we gathered active data and 8 categories of raw passive data sourced from phone sensors and usage metrics. Active data responses (eg, mood, sleep quality, and loneliness), scored on a scale of 1-7, were directly incorporated as features for the ML model.

Passive sensing data were collected via the Mindcraft mobile app, which continuously monitored phone-based sensors and system events in the background. Sensor sampling frequencies were set to balance data resolution with device energy efficiency. All sensors except the location sensor were collected at 15-minute intervals. GPS location data were triggered by significant location changes (as determined by the operating system). From the passive data, we engineered 92 distinct features (Table 1). Location-derived features such as total distance, radius of gyration, and maximum distance from the day's center of mass were derived from GPS logs using Haversine distance metrics and filtered for spurious location jumps. Features were aggregated over 2 time windows, the full 24-hour day and the nighttime period, defined as 10 PM to 6 AM.

A complete list of all active and passive features, along with descriptions and correlations with mental health scales, is provided in Multimedia Appendices 1 and 2, respectively. We highlight the key active and passive data features most strongly associated with each mental health outcome (the union of the top 3 per modality per outcome, ranked by absolute Spearman correlation) in Table 2.

Passive features that were unavailable due to user permission settings or OS-level constraints were set to a sentinel value of –1, indicating "sensor unavailable." This allowed the model to learn platform- or user-specific absence patterns without biasing distributions. All continuous features were *z* score normalized using training set means and SDs. Binary indicator features were kept in their original form.

XSL•FO
RenderX

Table . Summary of features engineered from passive data sensors.

Passive sensor	Number of features	Feature list
Ambient light (Android only)	8	Total, mean, median, and SD of ambient light reading in the day; total, mean, median, and SD of ambient light reading during night hours
App usage (Android only)	36	Total app usage count; unique apps; total, mean, and median time usage in the day; total app usage count; unique apps; total, mean, and median time usage during the night hours; total time in app categories of camera, communication, entertainment, gaming, physical health, mental health, Mindcraft, news, productivity, and social media; percentage time in app categories of camera, communication, entertainment, gaming, physical health, mental health, Mindcraft, news, productivity, and social media
Background noise level (Android only)	10	Total, median, mean, maximum, and SD of background noise levels in the day; total, median, mean, maximum, and SD of background noise levels during night hours
Battery	8	Min, maximum, mean, and median of battery level; number of charges per day; mean battery use per hour; time below 20 percent; nighttime usage count
Location	15	Mean latitude; mean longitude; total distance traveled in a day; location count; maximum distance from home; mean distance from home; median distance from home; nighttime movement; radius of gyration; SD of latitude; SD of longitude; location entropy; time spent at home
Mindcraft usage	3	First hour of use; last hour of use; nighttime usage;
Screen brightness (iOS only)	8	Total, mean, median, SD of screen brightness sensor reading in the day; total, mean, median, SD of screen brightness sensor reading during night hours
Step count	4	Daily step count; is daily step count greater than 5000 steps or 7500 steps or 10,000 steps?

XSL•FO
RenderX

Table . Key active data and passive data features and their Spearman correlations with mental health outcomes (Strengths and Difficulties Questionnaire [SDQ], Sleep Condition Indicator [SCI], Eating Disorder-15 Questionnaire [ED-15], and suicidal ideation).

Feature type	Feature	SDQ[a]	SCI[b]	Suicidal ideation	ED-15[c]
Active	Loneliness - "How lonely are you feeling today?"	0.48[d]	–0.42[d]	0.46[d]	0.31[d]
Active	Negative thinking - "How negative do you think today?"	0.48[d]	–0.47[d]	0.57[d]	0.46[d]
Active	Racing thoughts - "Are you experiencing racing thoughts today?"	0.45[d]	–0.44[d]	0.52[d]	0.48[d]
Active	Sleep quality - "How did you sleep last night?"	–0.37[d]	0.44[d]	–0.34[d]	–0.20[d]
Active	Self-care - "How is your self-care today?"	–0.37[d]	0.33[d]	–0.24[d]	–0.42[d]
Passive	Max background noise levels over the full day	–0.49[e]	0.36	–0.39	–0.1
Passive	Mean latitude of GPS samples over the full day	0.38[f]	–0.23	0.44[e]	–0.03
Passive	Number of entertainment app usage events over the full day	0.37[e]	–0.37[f]	0.51[d]	0.41[e]
Passive	Median ambient light at night	0.06	–0.39[e]	0.14	0.27
Passive	Mean longitude of GPS samples over the full day	0.19	–0.38[f]	0.25	–0.01
Passive	SD of background noise levels over the full day	–0.43	0.36	–0.63[d]	–0.31
Passive	Median app session duration over the full day	0.37[f]	–0.19	0.44[d]	0.19
Passive	Mean background noise level at night	0.2	–0.23	0.42	0.46[f]
Passive	Total number of app usage events at night	0.09	–0.06	0.27	0.39[e]

[a]SDQ: Strengths and Difficulties Questionnaire.

[b]SCI: Sleep Condition Indicator.

[c]ED-15: Eating Disorder-15 Questionnaire.

[d]$P<.001$.

[e]$P<.01$.

[f]$P<.05$.

To reduce day-to-day variability and enhance the stability of daily feature measurements, we computed the cumulative median of each feature for every participant, that is, the median of all values up to each day. This approach progressively aggregates behavioral signals over time, dampening the influence of short-lived anomalies (eg, a one-day spike due to sensor glitches or atypical behavior, which are common in smartphone data) while preserving sustained trends. Multimedia Appendix 3 illustrates how the cumulative median stabilizes noisy input without suppressing genuine behavioral shifts, such as a consistent drop in activity due to worsening mood. In preliminary analyzes, using the same model architecture, hyperparameters, and evaluation protocol, we compared models trained on raw daily features versus those trained on cumulative median-aggregated features. We found that the latter consistently improved balanced accuracy across all outcomes (Multimedia Appendix 4). We further compared the cumulative median and the cumulative mean aggregation. While both methods yielded comparable predictive accuracy, we selected the cumulative median as the primary aggregation function due to its superior

XSL•FO
RenderX

robustness to outliers, a common artifact in passive mobile sensing, thereby ensuring greater stability in user representation.

Using the engineered features, we developed an ML model for each of the 4 mental health outcomes, namely SDQ risk, insomnia, suicidal ideation, and eating disorders. We used 3 distinct feature sets, including active data, passive data, and a combination of both. This design enabled us to assess the predictive strength of each feature set individually and in combination, allowing us to systematically quantify each modality's contribution to prediction performance across mental health outcomes.

Participants were classified as high-risk or low-risk for each outcome using validated thresholds specific to each mental health measure, framing the prediction task as a binary classification problem. Each scale's total score range is as follows: SDQ score 0‐40, SCI 0‐32, ED-15 0‐6, and suicidal ideation 0‐4 (based on response frequency to the question, "Over the last two weeks, how often have you been bothered by thoughts that you would be better off dead or of hurting yourself in some way?"). High-risk classifications were defined as follows: for SDQ, a self-reported score of ≥16 [53]; for insomnia, if their SCI score was ≤16 [51]; for suicidal ideation, if they responded at least once to the question regarding frequency of suicidal thoughts over the last 2 weeks, and for eating disorders, an ED-15 total score exceeded 2.69, which corresponds to the mean plus one SD in a nonclinical population [54]. The proportions of participants classified as high-risk for each mental health outcome are summarized in Table 3.

Table . Demographics and mental health measures of the study population (N=103).

Variables	Values
Sex (Female), n (%)	73 (70.9)
Age (years), mean (SD)	16.1 (1)
Strengths and Difficulties Questionnaire (SDQ) score, mean (SD)	12.8 (6.2)
High-risk SDQ category (SDQ score≥16), n (%)	31 (30.1)
Eating Disorder (ED-15) scale, mean (SD)	2.2 (1.8)
High-risk eating disorder category (ED-15 score≥2.7), n (%)	38 (36.9)
Sleep Condition Indicator (SCI) score, mean (SD)	19.9 (7.8)
High-risk insomnia category (SCI score<17), n (%)	34 (33)
"Over the last two weeks, how often have you been bothered by thoughts that you would be better off dead or of hurting yourself in some way?" mean (SD)	0.6 (0.9)
High-risk suicidal ideation category (≥1), n (%)	38 (36.9)

ML Workflow and Model Development

Figure 1A outlines our ML pipeline, starting with active and passive data collection via the Mindcraft app. The data were preprocessed and engineered to create a comprehensive feature set, which was subjected to a pretraining phase with contrastive learning using triplet margin loss. This pretraining step clustered user-specific features from different days, minimizing day-to-day variability and preserving individual behavioral patterns. The resulting stable embeddings were fine-tuned on labeled data in a supervised setting to predict mental health outcomes. This end-to-end pipeline, combining pretraining and fine-tuning, enabled the development of a predictive model evaluated using balanced accuracy and additional metrics.

XSL•FO
RenderX

Figure 1. Overview of the machine learning (ML) framework for predicting adolescent mental health outcomes using smartphone-based digital phenotyping. (A) Workflow of the ML pipeline, from data acquisition to mental health outcome prediction, incorporating contrastive pretraining with triplet loss and fine-tuning. (B) t-SNE visualization of feature embeddings for a sample of 10 test users before (left) and after (right) contrastive pretraining, showing enhanced user-specific clustering following pretraining. t-SNE: t-distributed Stochastic Neighbor Embedding.

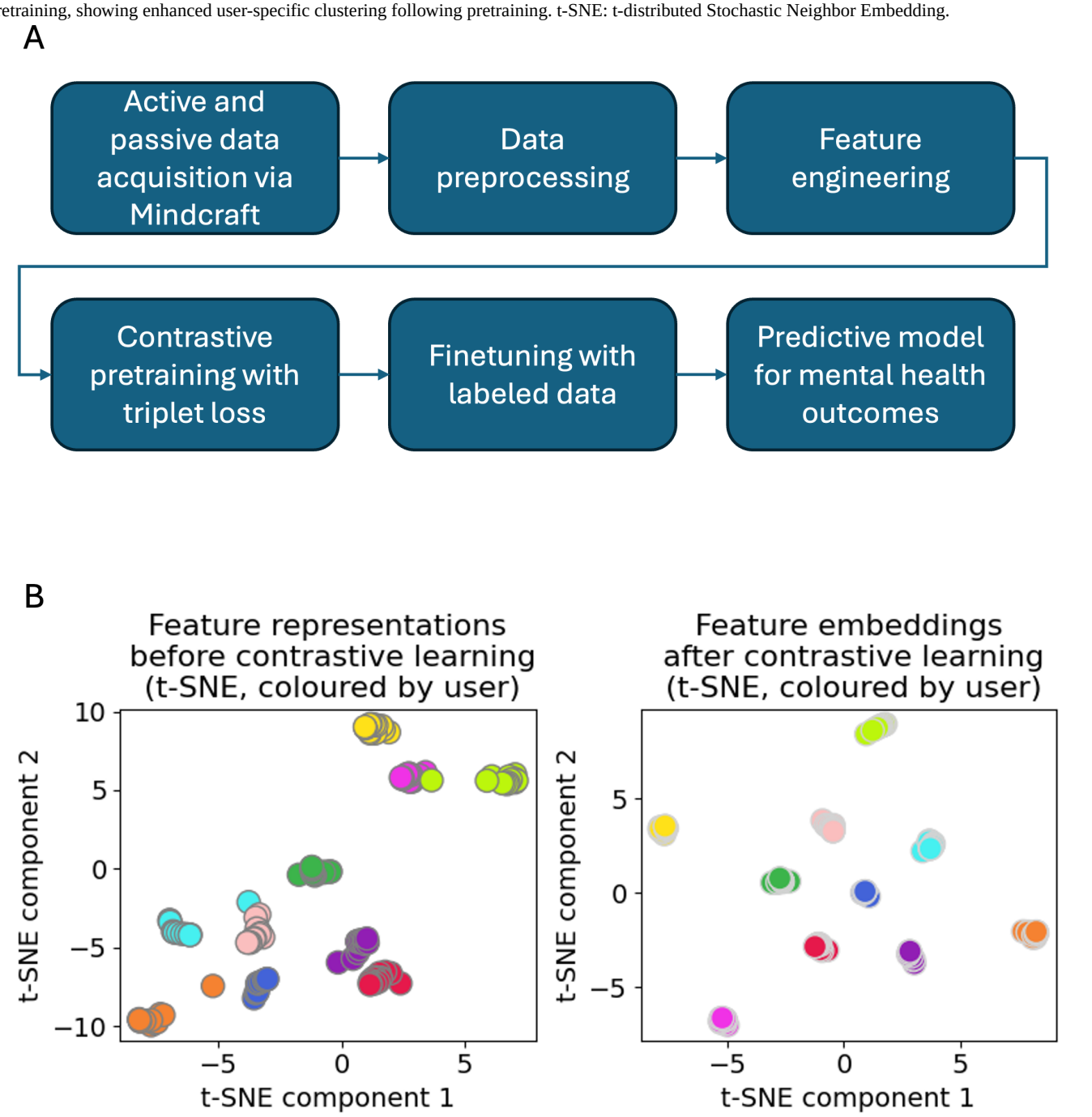

Contrastive Pretraining Model Architecture and Training

To accurately predict mental health risks, it is essential to distinguish between a user's stable behavioral patterns and random daily fluctuations. Raw smartphone data can vary significantly day to day due to external factors (such as school holidays) unrelated to mental health. To address this, we used an ML technique known as contrastive learning with triplet margin loss. In simple terms, this technique teaches the model to recognize a user's unique digital fingerprint. It does so by examining 3 data points (a "triplet") simultaneously:

- Anchor: data from a specific user on a particular day (eg, User A, Monday).
- Positive: data from the same user on a different day (eg, User A, Thursday).
- Negative: data from a different user entirely (eg, User B, Monday).

The model is trained to minimize the distance between the Anchor and the Positive (pulling them together in mathematical space) while maximizing the distance between the Anchor and the Negative (pushing them apart). Repeating this process across thousands of triplets many times helps the model learn to ignore irrelevant daily noise and form a stable, underlying behavioral

XSL•FO
RenderX

fingerprint unique to that user. These stable representations are then used for the subsequent supervised prediction of mental health risk. A simplified visual explanation of this idea is provided in Figure 1B, which helps understand why contrastive pretraining is valuable in datasets with naturally high daily variability, such as smartphone-based behavioral data.

We implemented a custom contrastive learning pipeline in PyTorch (Meta AI), based on established principles of triplet-based metric learning [55,56], with the objective of learning robust, user-specific feature embeddings that remain consistent across days, thereby reducing intrasubject variability while maximizing intersubject separability. For this pretraining phase, we selected triplet margin loss over other contrastive objectives (eg, InfoNCE) due to its suitability for instance-level metric learning, where the aim is to ensure that embeddings from the same user are closer in the latent space than those from different users. Compared to InfoNCE, triplet loss is more stable with moderate mini-batch sizes and is better suited to high-dimensional tabular data [55,57].

Triplets were constructed without using mental health outcome labels, as the pretraining phase was fully self-supervised. A user was first randomly selected from the training set, and a single day from their data was sampled as the Anchor. The Positive sample was randomly chosen from a different day for the same user, while the Negative was drawn from a random day of a different user. The triplet loss requires that an Anchor data point is closer to a Positive data point than it is to a Negative data point, by at least a specified margin m.

The contrastive learning model consisted of a 2-layer multilayer perceptron (MLP) encoder, which mapped input features to a latent embedding space, followed by a 2-layer MLP projection head. The projection head was trained using the triplet loss, allowing the encoder to retain generalizable behavioral representations while the projection space focused on the contrastive objective [58]. The Adam optimizer was used for pretraining with a fixed learning rate of 1×10^{-3}. The model was pretrained for 3 epochs with a batch size of 256 triplets. The margin m was set to 1.0 in our experiments.

Supervised Fine-Tuning Model Architecture and Training

Following contrastive pretraining, the learned encoder was used as a fixed base for the downstream classification tasks. Its weights were frozen to preserve the stable user-specific embeddings. A new 2-layer MLP classification head was added on top of the frozen encoder to perform binary classification for each mental health outcome. The fine-tuning model was trained using the Adam optimizer with a learning rate of 1×10^{-3} for 20 epochs and a batch size of 1024. Binary cross-entropy with logits loss was used as the training objective, with class weights applied to account for imbalance. These weights were calculated based on the inverse frequency of each class within the training set of each cross-validation fold.

Key architectural hyperparameters (eg, number of layers, dimensions of embeddings, and projections) and training hyperparameters (eg, learning rate, batch size, triplet margin, and number of epochs) were selected based on common practice in contrastive learning and refined through exploratory tuning on a validation split of the training data within the first cross-validation fold. No test user data from any fold was used during hyperparameter selection. All selected hyperparameters were then fixed across folds to ensure methodological consistency and reproducibility and to prevent data leakage.

To demonstrate the effect of contrastive pretraining, Figure 1B shows t-distributed Stochastic Neighbor Embedding visualizations of feature embeddings for a sample of 10 test users. Before pretraining (left plot), user-specific data points were scattered with minimal clustering, reflecting high day-to-day variability. After applying contrastive learning with triplet loss (right plot), data points from the same user formed tighter clusters, indicating enhanced user-specific feature stability.

Evaluation and Benchmarking

Day-wise prediction probabilities were averaged for each test user to obtain a single, user-level prediction. Model performance was evaluated using balanced accuracy as the primary metric, along with other relevant metrics such as area under the receiver operating characteristic curve (AUC) and F_1-score, to assess classification outcomes comprehensively. For interpretability, Shapley additive explanations (SHAP) values [59] were computed for test folds using DeepExplainer [60], offering insight into the contributions of specific features to classification outcomes.

We validated the model's performance using leave-one-subject-out cross-validation (LOSO-CV), where data from all but one user were used for training and validation, with the excluded user's data serving as the test set. This approach ensures the model's generalizability to new individuals, closely simulating real-world applications where accurate predictions for unseen users are critical.

To benchmark our model, we compared it with a CatBoost (Yandex) classifier [61] and an MLP network that had a similar number of parameters but was trained without pretraining. Both benchmarks used class-weight balancing to address the class imbalance in the dataset. CatBoost was selected for its strong performance on tabular data and its built-in capability to handle class imbalance [62], making it well-suited for datasets like ours. These comparisons isolated the effect of contrastive pretraining on model performance.

External Validation on an Independent Cohort

We collected data from an additional cohort of 90 adolescents across 5 new secondary schools in London. To create an independent external validation set, we randomly split this new cohort into 2 halves. One half (n=45) was added to the original dataset to form an expanded training set, while the remaining half (n=45) served as a fully held-out external validation cohort. Importantly, no model architecture, hyperparameters, preprocessing, or feature engineering steps were modified after adding the new data; the complete pipeline, including contrastive pretraining and supervised fine-tuning, remained fixed. This design allowed us to assess generalizability to users from different schools, collected at a later time point, with different demographic characteristics and smartphone-sensor enabling

XSL•FO
RenderX

patterns. Model performance on this external cohort was evaluated using balanced accuracy following the same user-level prediction protocol used in the original LOSO-CV analysis.

Ethical Considerations

Ethics approval for the study was granted by the Imperial College London Research Ethics Committee (reference number: ICREC 20IC6132). All procedures complied with relevant national and institutional ethical standards and with the Declaration of Helsinki. Informed consent was obtained digitally from all participants at the start of the survey; for participants younger than 16 years, parental consent was obtained before participation. Participants could withdraw from the study at any time before the start of data analysis. Study-specific identification numbers were used to maintain participant anonymity, and data handling complied with the General Data Protection Regulation (GDPR) for health and care research. Participation was voluntary and uncompensated. Multimedia Appendix 5 provides a CONSORT (Consolidated Standards of Reporting Trials)-style flow diagram outlining the subject inclusion process from the initial 205 survey respondents.

Results

Recruitment and App Usage

Figure 2A provides a conceptual overview of the study, demonstrating how active data (eg, sleep quality, mood, and loneliness) and passive data (eg, location, app usage, and noise levels) collected via the Mindcraft app are integrated into a contrastive learning-based deep neural network to predict mental health outcomes, including SDQ risk, insomnia, suicidal ideation, and eating disorders.

A total of 103 students from 3 London schools downloaded and installed the Mindcraft app. The average age was 16.1 years (SD 1), with 71% identifying as female, 25% as male, and 4% as other or nonbinary. The skew in gender distribution is partially due to one of the participating schools being girls-only. Of the participants, 78 used the app on iPhones, and 25 used Android phones. Table 3 provides demographic information and mental health outcome scores, and Multimedia Appendix 6 illustrates the distribution of the different mental health measures across our study population.

Participants contributed active data via self-reported measures and passive data through smartphone sensors. Active data included daily ratings of mental well-being measures such as sleep quality, mood, confidence, and loneliness on a 1 - 7 scale. Passive data comprised data from phone sensors like location, app usage, ambient noise, and step count. Figure 2B shows user engagement patterns over the 14-day study period. Initial engagement was high, with all participants contributing at baseline. However, active data engagement declined more rapidly than passive data, with 14 users contributing active data and 36 users contributing passive data on day 14.

Engagement with active data measures (Figure 2C) remained consistent across users, with slight variations reflecting individual preferences. In contrast, passive data collection exhibited substantial variability (Figure 2D). While 36 users opted not to enable any sensors, others enabled multiple categories. The most frequently enabled sensors were step count and battery usage, followed by Mindcraft usage and screen brightness (Figure 2E). The heatmap visualization of passive data coverage by users and sensor type (Figure 2F) underscores substantial interuser variability, with some users providing comprehensive coverage across multiple sensors and others contributing sporadically.

XSL•FO
RenderX

Figure 2. Study overview and participant engagement with active and passive data collection using the Mindcraft app. (A) Conceptual overview of the study. (B) User engagement trends for active and passive data over the 14-day period. (C) User participation across active data questions. (D) Distribution of enabled passive sensors among users. (E) User participation across different passive sensor types. (F) Heatmap of passive data completeness by user and sensor type. SDQ: Strengths and Difficulties Questionnaire.

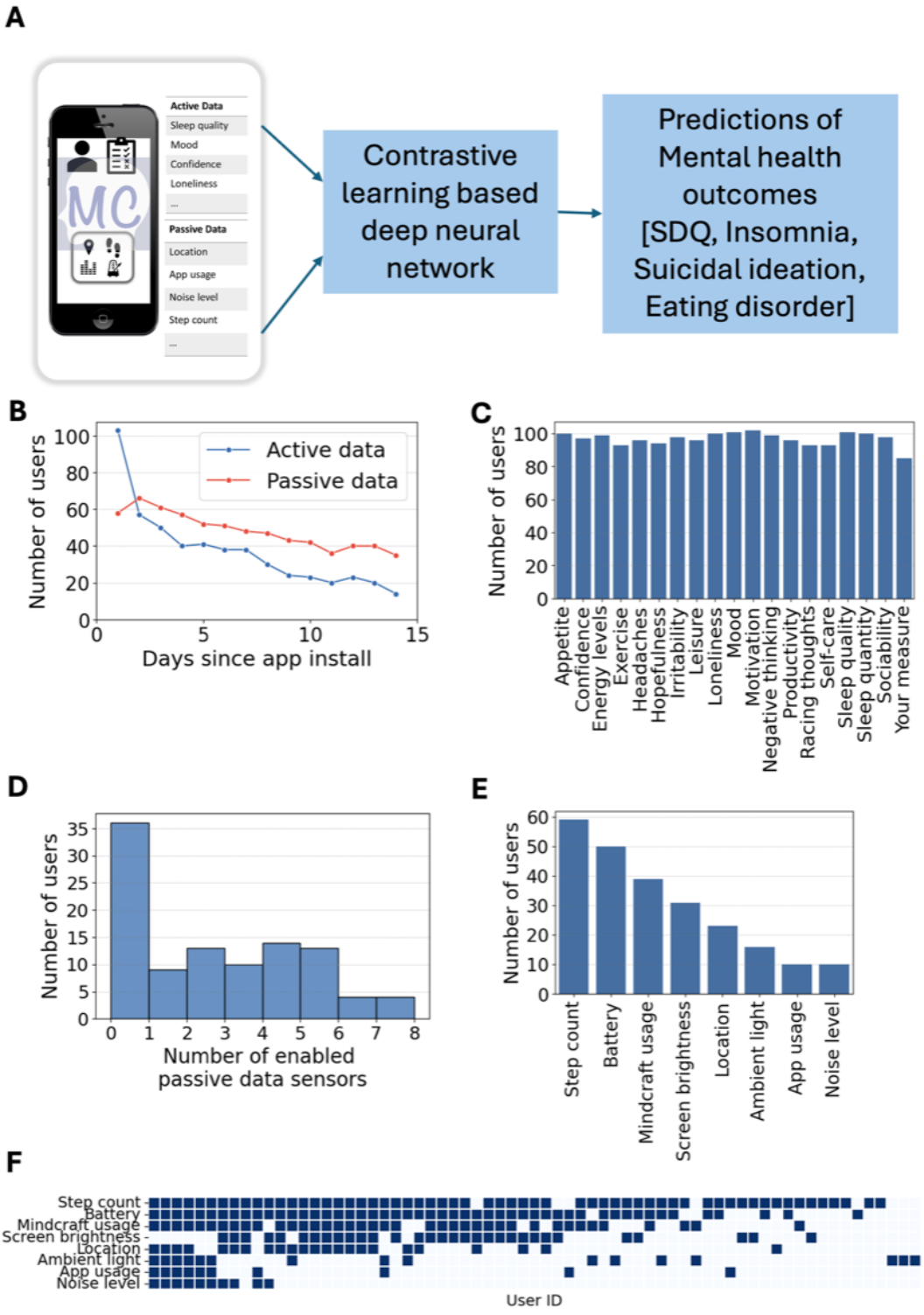

Exploratory Analysis of Active and Passive Data Features

Figure 3 provides an overview of descriptive statistics and correlations among active and passive data features collected from users. Figure 3A illustrates the distribution of self-reported active data features on a scale of 1-7. Positive indicators, such as mood, motivation, and confidence, had higher mean values than negative indicators, such as negative thinking, racing thoughts, and irritability.

XSL•FO
RenderX

Figure 3. Descriptive statistics of active and passive digital phenotyping features. (A) Distribution of responses across active data features. (B) Correlation heatmap of active data features. (C) Frequency distributions of passive data features: daily step count. (D) Number of unique apps opened per day. (E) Location entropy reflecting movement variability. (F) Mean background noise levels at night.

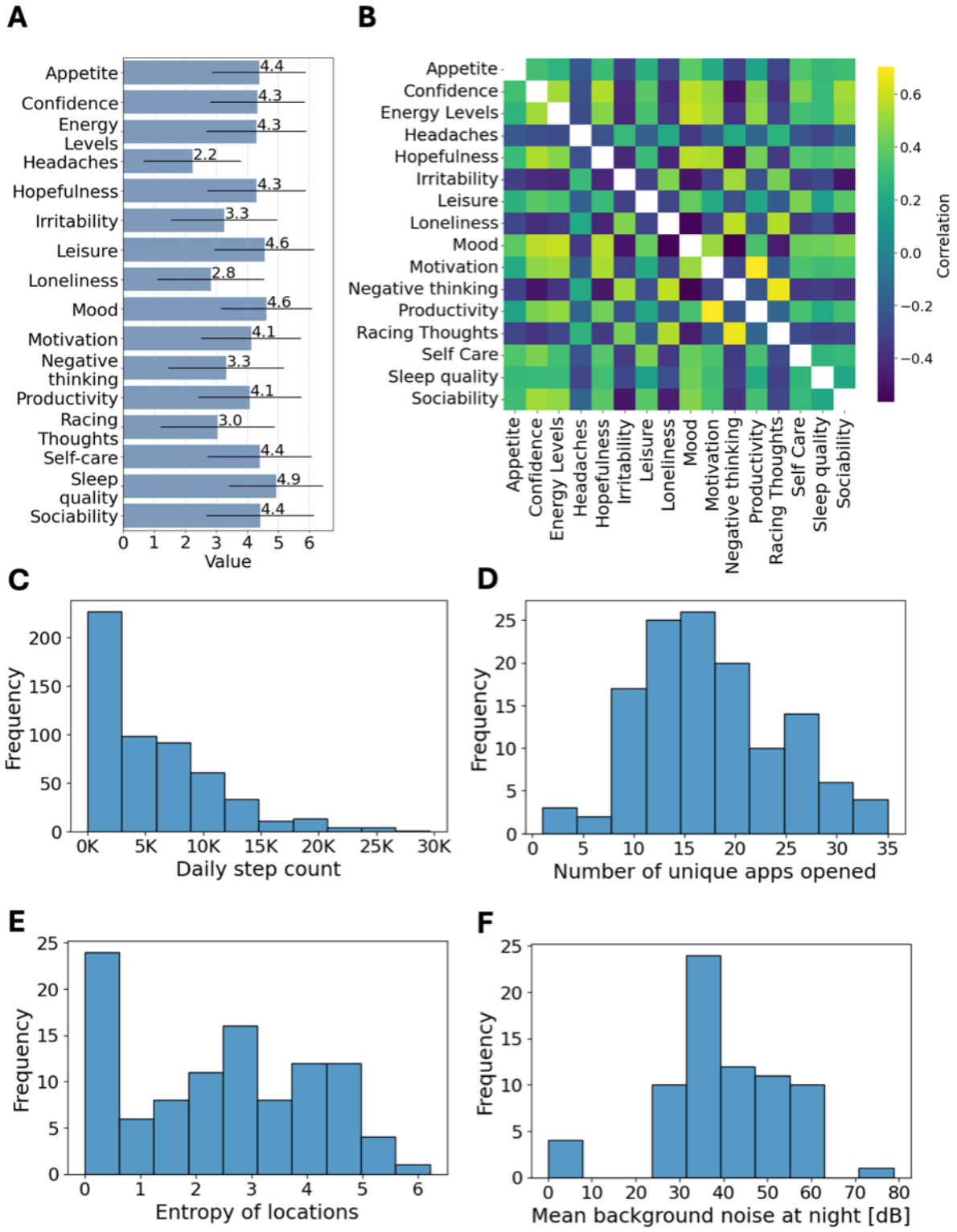

The correlation heatmap (Figure 3B) highlights relationships among active data features. A fully annotated, high-resolution version of this heatmap is provided in Multimedia Appendix 7. The strongest correlation (r=0.7) was observed between motivation and productivity, followed by a strong association between negative thinking and racing thoughts (r=0.66). Positive correlations were also observed between 2 well-being indicators, energy levels and mood (r=0.58), and between 2 distress indicators, loneliness and negative thinking (r=0.57). Conversely, negative correlations were seen, such as between mood and negative thinking (r=–0.56) and irritability and sociability (r=–0.49).

Figures 3C-3F illustrate 4 of the 92 engineered passive data features. The distribution of daily step counts (Figure 3C) is

XSL•FO
RenderX

right-skewed, with most users taking fewer than 10,000 steps per day. Figure 3D shows the frequency of unique apps opened daily, peaking at 15 - 20 apps, indicating varying levels of mobile engagement. The entropy of locations visited (Figure 3E) reflects movement variability, with higher values suggesting diverse activity patterns. Finally, Figure 3F highlights the distribution of mean background noise levels at night, clustering between 30 and 50 decibels.

Associations Between Active and Passive Features and Clinical Outcomes

To assess the alignment between digital phenotyping features and clinical mental health symptoms, we examined Spearman correlations between active and passive features and continuous scores on the SDQ, SCI, ED-15, and suicidal ideation frequency. Among active features, negative thinking consistently showed the strongest associations across all outcomes (ρ=0.48 with SDQ, ρ=–0.47 with SCI, ρ=0.57 with suicidal ideation, ρ=0.48 with ED-15, all P<.001). Higher levels of racing thoughts and loneliness were associated with more severe mental health symptoms, whereas greater confidence and hopefulness were linked to reduced risk. Key passive features also demonstrated moderate associations with outcomes. For instance, greater entertainment app usage consistently showed associations across all outcomes (ρ=0.37 with SDQ, ρ=–0.37 with SCI, ρ=0.51 with suicidal ideation, ρ=0.41 with ED-15, all P<.05). Nighttime ambient light exposure and location variability (latitude and longitude) were also relevant across multiple outcomes, particularly for insomnia and eating disorder symptoms. Key active and passive features, along with their correlations with each outcome, are summarized in Table 2. Multimedia Appendices 1 and 2 list detailed descriptions and Spearman correlations for all active and passive features used in the predictive modeling pipeline, offering an overview of their associations with SDQ, SCI, ED-15, and suicidal ideation.

To further illustrate the discriminative capacity of key digital phenotyping features, we compared the top 5 active and passive features between high- and low-risk groups for each mental health outcome. Multimedia Appendix 8 shows the variability in these features across the 4 outcomes (SDQ, insomnia, suicidal ideation, and eating disorder). The observed group differences are consistent with the correlation-based findings and reinforce the predictive relevance of both active and passive data streams in distinguishing individuals at elevated mental health risk.

Performance of Models Predicting Mental Health Outcomes

Building on these associations, we evaluated how well ML models could predict mental health outcomes using active, passive, and combined data. Figure 4A illustrates the balanced accuracy of predictive models for the 4 mental health outcomes (SDQ-high risk, insomnia, suicidal ideation, and eating disorder) evaluated across 10 repetitions of LOSO-CV. The analysis involved 3 feature sets, including passive data, active data, and a combination of both. This evaluation was restricted to the 67 participants who provided both active and passive data to ensure a fair comparison. The red dashed line indicates chance-level performance, and statistically significant differences are denoted by asterisks (*P<.05, **P<.01, ***P<.001; Wilcoxon signed-rank test).

XSL•FO
RenderX

Figure 4. Model performance for predicting mental health outcomes using active and passive digital phenotyping data. (A) Balanced accuracy of mental health outcome predictions (SDQ-high risk, insomnia, suicidal ideation, and eating disorder) using passive, active, and combined data. (B) Comparison of balanced accuracy for models with contrastive pretraining, without pretraining, and a CatBoost model, showing the performance benefit of pretraining (P<.001, paired t test). (C) Confusion matrices for the combined data model's predictions, showing true-positive and true-negative classifications across mental health outcomes. *: P<.05; **: P<.01; ***: P<.001; SDQ: Strengths and Difficulties Questionnaire.

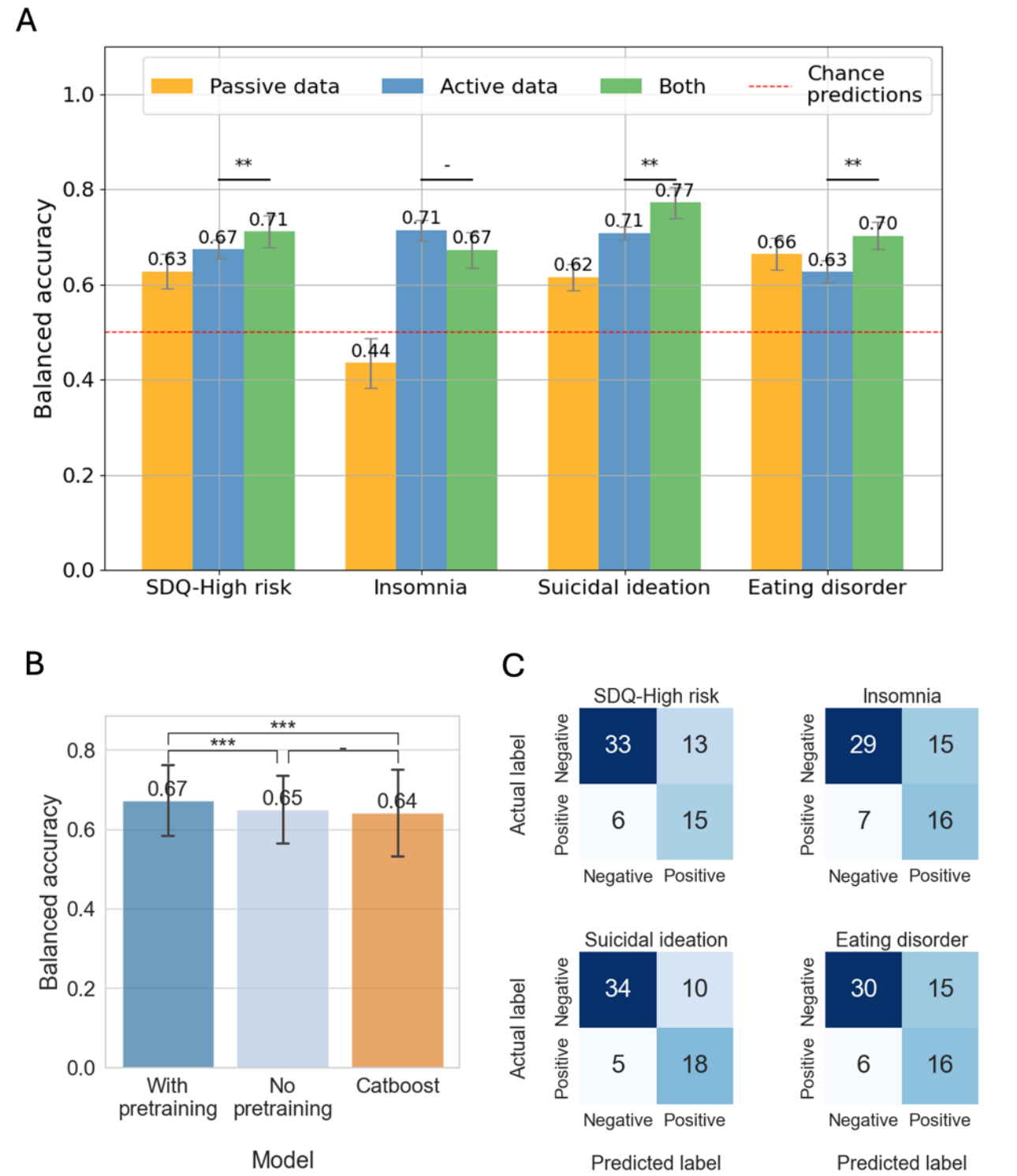

For SDQ-high risk, the model using passive data achieved a balanced accuracy of 0.63, while active data alone reached 0.67 (Multimedia Appendices 1 and 2). The combined model, leveraging both data types, achieved a significantly higher balanced accuracy of 0.71 compared to active data alone (P=.003, Wilcoxon signed-rank test). Similarly, the combined data model outperformed the active data alone for eating disorder predictions, with balanced accuracies of 0.70 and 0.63, respectively (P=.003; Wilcoxon signed-rank test). In predicting insomnia, the combined model achieved a balanced accuracy of 0.67, while passive data alone performed below the chance level (0.44). For suicidal ideation, the combined model achieved

XSL•FO
RenderX

the highest balanced accuracy of 0.77, significantly outperforming both active data (0.71) and passive data (0.62; P=.003; Wilcoxon signed-rank test). Table 4 summarizes additional performance metrics, including the AUC, the area under the precision-recall curve, F_1-scores, sensitivity, specificity, precision, and recall for each mental health outcome.

Table . Detailed performance metrics for mental health outcome predictions.

Metric	SDQ[a]-high risk, mean (SD)	Insomnia, mean (SD)	Suicidal ideation, mean (SD)	Eating disorder, mean (SD)
Balanced accuracy	0.71 (0.03)	0.67 (0.04)	0.77 (0.03)	0.70 (0.03)
AUC[b]	0.77 (0.03)	0.74 (0.02)	0.82 (0.03)	0.73 (0.02)
AUC-PR[c]	0.53 (0.04)	0.52 (0.05)	0.64 (0.05)	0.52 (0.03)
*F*1-score	0.61 (0.04)	0.59 (0.04)	0.70 (0.04)	0.61 (0.03)
F1 macro	0.69 (0.03)	0.66 (0.04)	0.76 (0.03)	0.68 (0.03)
Sensitivity	0.71 (0.06)	0.68 (0.03)	0.78 (0.05)	0.74 (0.03)
Specificity	0.71 (0.05)	0.66 (0.07)	0.77 (0.04)	0.67 (0.04)
Precision	0.53 (0.04)	0.52 (0.05)	0.64 (0.05)	0.52 (0.03)
Recall	0.71 (0.06)	0.68 (0.03)	0.78 (0.05)	0.74 (0.03)

[a]SDQ: Strengths and Difficulties Questionnaire.

[b]AUC: area under the receiver operating characteristic curve.

[c]AUC-PR: area under the precision-recall curve.

Figure 4B demonstrates the effectiveness of contrastive pretraining. Models with pretraining achieved the highest balanced accuracy (0.67), significantly outperforming both the model without pretraining (0.65; P<.001; paired 2-tailed t test) and the CatBoost model (0.64; P<.001; paired 2-tailed t test).

Figure 4C presents the confusion matrices for the combined data models. For SDQ-high risk, the model correctly identified 33 negatives and 15 positives, with 6 false negatives and 13 false positives. The model had higher misclassification rates for insomnia, with 7 false negatives and 15 false positives. In predicting suicidal ideation, the model demonstrated strong performance, correctly classifying 34 negatives and 18 positives, with only 5 false negatives and 10 false positives. Similarly, for eating disorders, the model accurately identified 30 negatives and 16 positives, with 6 false negatives and 15 false positives.

To assess whether the active-only model's predictive performance generalizes across user subgroups, we compared balanced accuracy for the subset (n=67) who provided both active and passive data with that of the full cohort (n=103), which includes participants who shared only active data (Multimedia Appendix 9). The consistent results across these 2 groups indicate that the active-only model's performance is stable and not restricted to a specific subgroup, thereby confirming that restricting the analysis to the subset of users with both active and passive data (as done in Figure 4A) has not introduced bias in model performance.

External Validation Performance on an Independent Cohort

Model performance on the held-out external validation cohort (n=45 adolescents from 5 additional schools; none of these adolescents' data were used for model development) showed a similar pattern to the LOSO-CV results, albeit with reduced accuracy for some outcomes (Figure 5). Figure 5 shows balanced accuracy for models trained using passive data only, active data only, and combined active and passive data; error bars indicate standard deviation across 10 repeated runs. The red dashed line indicates chance-level performance, and statistically significant differences are denoted by asterisks (*P<.05, **P<.01, ***P<.001; Wilcoxon signed-rank test). Using combined active and passive data, balanced accuracy in the external cohort was 0.72 for SDQ-high risk, 0.63 for insomnia, 0.63 for suicidal ideation, and 0.63 for eating disorder risk (vs 0.71, 0.67, 0.77, and 0.70, respectively, in the LOSO-CV analysis). Models trained using passive data only and active data only showed the same qualitative pattern, with active and combined data generally outperforming passive data alone, but with lower accuracies than in LOSO-CV, particularly for suicidal ideation and eating disorders. A full set of evaluation metrics for the combined model in the external validation sample is provided in Multimedia Appendix 10. Overall, the external validation results showed a similar performance pattern to the LOSO-CV analysis, with reduced accuracy for suicidal ideation and eating disorder.

XSL•FO
RenderX

Figure 5. External validation performance on an independent cohort. Balanced accuracy for predicting SDQ high risk, insomnia, suicidal ideation, and eating disorder risk in an external validation sample (n=45 adolescents not used for model development). Error bars indicate SD across 10 repeated runs. *: *P*<.05; **: *P*<.01; ***: *P*<.001; SDQ: Strengths and Difficulties Questionnaire.

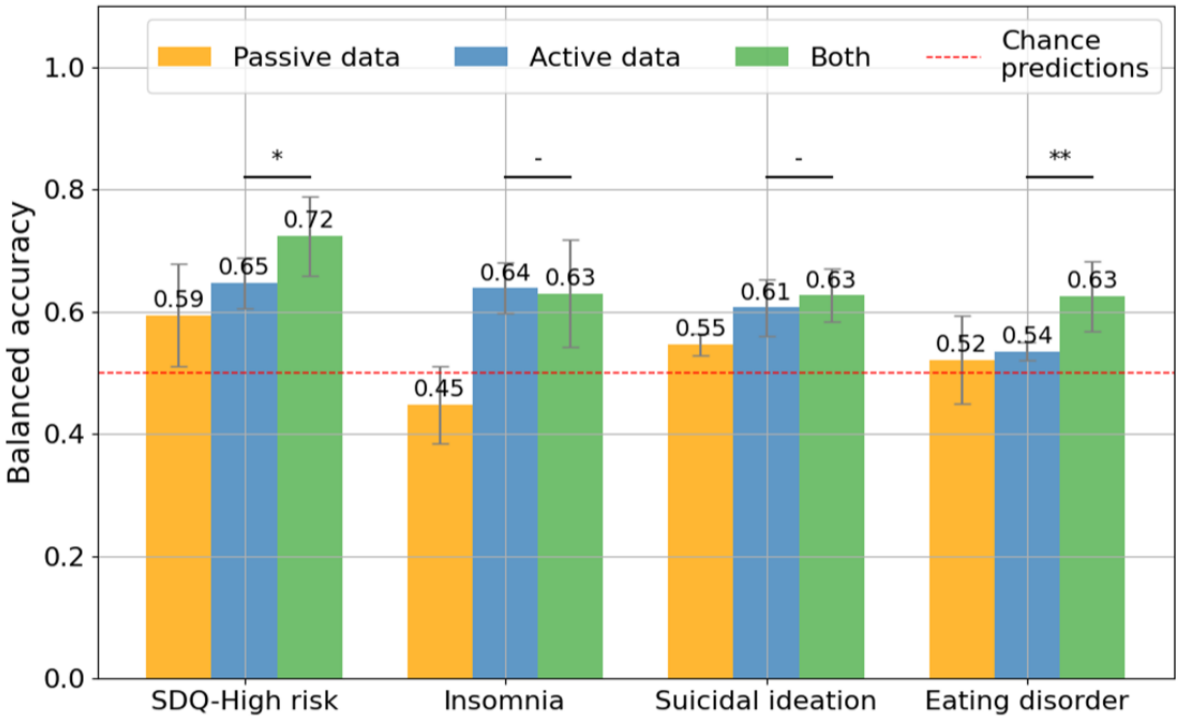

Model Fairness Across Gender and School Contexts

Given known differences in smartphone use and mental health between demographic groups, we examined whether model performance varied systematically by gender (Multimedia Appendix 11). Participants were grouped into female (n=46) and male and other (n=21). For the combined active and passive model, balanced accuracy was broadly comparable across gender for all 4 outcomes, with differences modest in magnitude and inconsistent in direction (eg, suicidal ideation: 0.73 vs 0.74). Error bars overlapped for all outcomes, and no subgroup was consistently disadvantaged, suggesting no evidence of systematic gender-related performance degradation in this sample.

Because individual-level socioeconomic status (SES) data were unavailable, we used school as a contextual proxy for both SES and the institutional environment. The 3 participating schools differed in selectivity, gender composition, and catchment-area deprivation: School 1 (girls-only, partially selective grammar in a mid-SES area), School 2 (mixed-gender community school in a deprived urban area), and School 3 (mixed-gender selective sixth-form college in a mid-to-higher SES area). For each school, we compared balanced accuracy for students from that school with that for students from the other 2 schools combined (Multimedia Appendix 11). Across all 4 outcomes, differences were varied in direction rather than systematically favoring or disadvantaging any single school (eg, School 1 vs others: SDQ 0.71 vs 0.73; insomnia 0.76 vs 0.59; suicidal ideation 0.68 vs 0.76; eating disorder 0.65 vs 0.71), with overlapping CIs in all cases. Overall, these analyses provide no evidence of systematic performance degradation associated with school context or school-level SES, although the study was not powered to detect more subtle or intersectional fairness effects.

Robustness to Heterogeneous Sensor Activation

To evaluate whether missing passive-sensor streams introduced systematic bias, we stratified participants based on the number of passive sensors enabled. The median number of sensors in the dataset was 3; therefore, we compared model performance between users with ≤3 sensors enabled and those with >3 sensors enabled (Multimedia Appendix 12). Model performance was evaluated separately for each subgroup for all 4 outcomes using the combined active-and-passive model. Differences in balanced accuracy were inconsistent in direction across outcomes, indicating no evidence of systematic performance degradation for users with more missing sensor data. Notably, for suicidal ideation, the 2 subgroups performed nearly identically (0.74 vs 0.72). For eating disorder risk, the ≤3-sensor group performed better (0.80 vs 0.56), suggesting that the core predictive signals for this outcome are captured within the most commonly enabled sensors and that additional sensors may introduce noise for this specific target. These results provide no evidence that heterogeneous sensor activation introduced systematic group-level bias against users with fewer sensors. They also suggest that the combined model is reasonably robust to the observed patterns of missing passive data. However, our sample size limits the detection of more subtle fairness effects related to specific sensor combinations or intersectional subgroups.

XSL•FO
RenderX

Predictive Accuracy Across Mental Health Risk Groups

Figure 6 illustrates model accuracy in predicting mental health risks using combined active and passive data, segmented by risk levels for various mental health measures. The model performed exceptionally well at extreme risk levels, achieving near-perfect accuracy for high-risk groups (eg, SCI scores 0 - 8 and EDQ scores 4 - 6) and low-risk groups (eg, SDQ scores 1 - 8). However, accuracy decreased significantly in ranges near thresholds (eg, SDQ scores 9 - 16 and SCI scores 9 - 16).

Figure 6. Accuracy of mental health risk prediction across different levels of (A) Strengths and Difficulties Questionnaire (SDQ) total score. (B) Sleep Condition Indicator (SCI) score. (C) Frequency of suicidal ideation thoughts. (D) Eating Disorder (ED-15) total score.

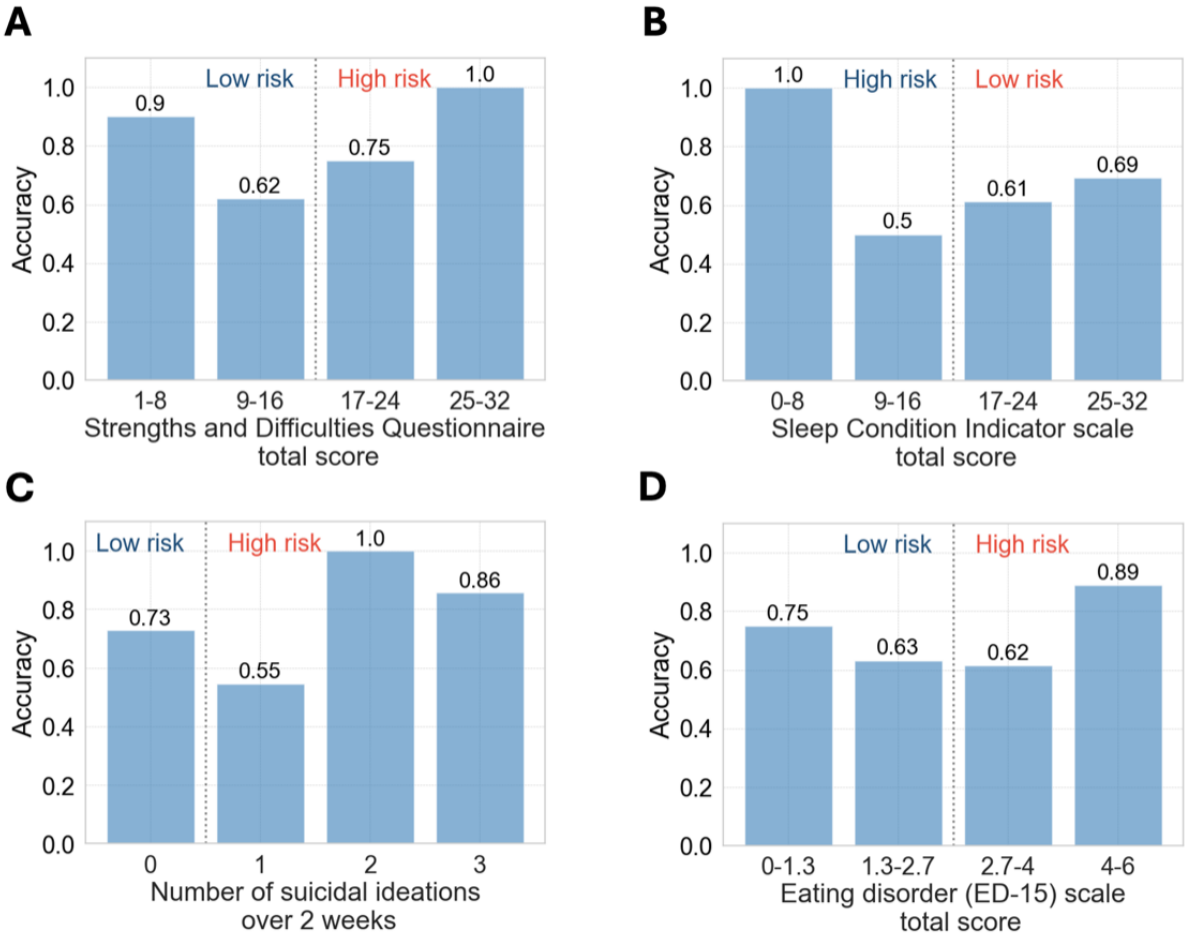

Model Interpretability: Active and Passive Data Contributions

Figure 7A illustrates the feature importance calculated using SHAP values for predicting the SDQ high-risk category using a combination of both active and passive data, with passive data aggregated by sensor type and active data shown individually. The top predictors included negative thinking, location features, app usage, racing thoughts, and self-care. Cognitive and emotional indicators (eg, negative thinking and racing thoughts) ranked highest among active data features, while movement and environmental stability (eg, location entropy and step count) dominated passive data contributions. Corresponding SHAP-based feature importance plots for predicting insomnia, suicidal ideation, and eating disorders are provided in Multimedia Appendices 13-15, respectively.

XSL•FO
RenderX

Figure 7. Feature importance analysis for predicting the SDQ high-risk category using both active and passive data. (A) SHAP-based feature importances, with passive data aggregated by sensor type and active data shown individually. (B) Distribution of the top five active data features across SDQ risk categories. (C) Distribution of the top five passive data features across SDQ risk categories. Statistically significant differences between low-risk and high-risk groups are indicated (*: $P<.05$, **: $P<.01$, ***: $P<.001$, t test). SDQ: Strengths and Difficulties Questionnaire; SHAP: Shapley Additive Explanations.

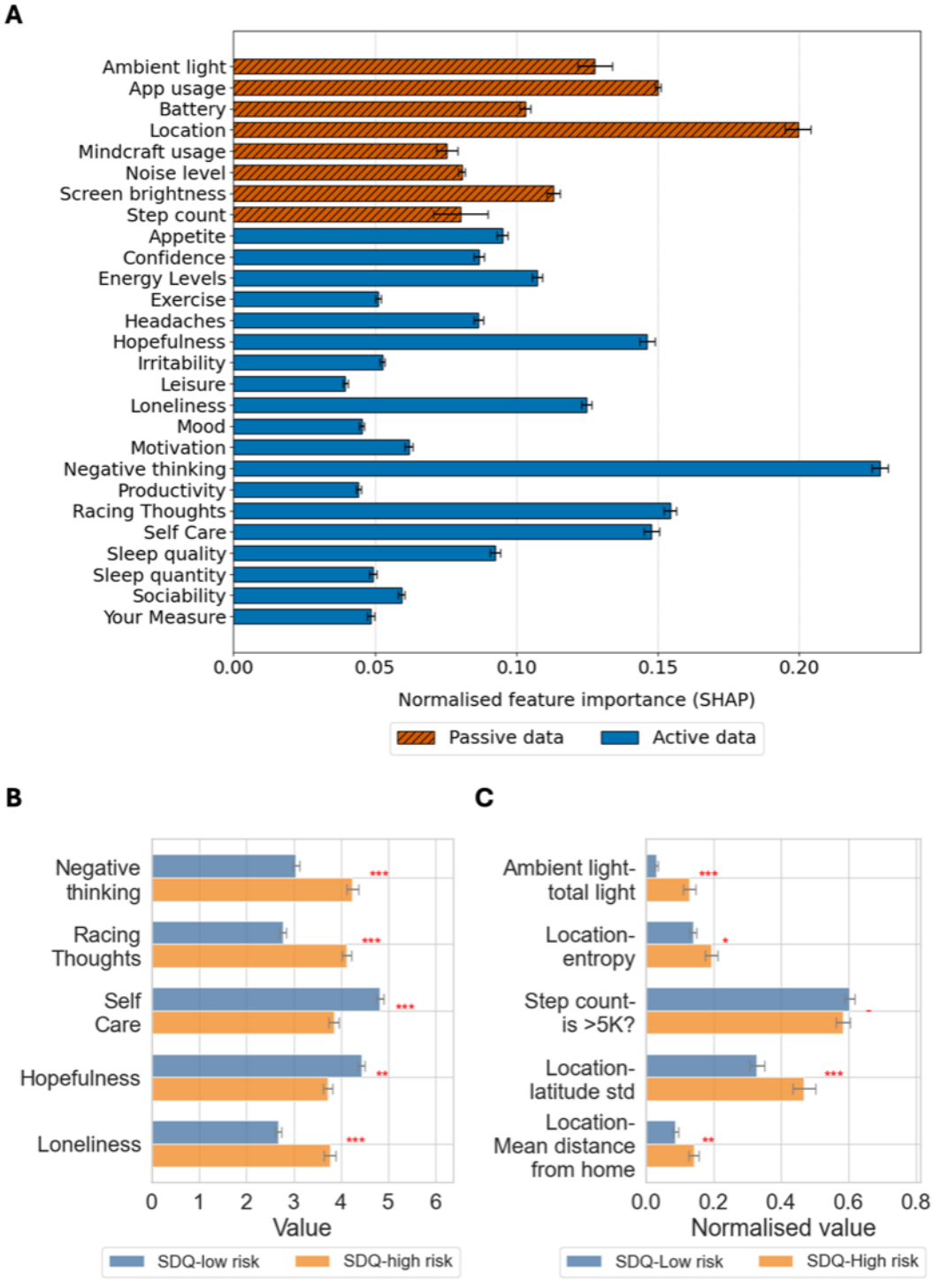

The distribution of the top five active data features (negative thinking, racing thoughts, self-care, hopefulness, and loneliness) showed clear distinctions between low- and high-risk SDQ groups (Figure 7B). Negative thinking and racing thoughts were significantly higher in the high-risk group ($P<.001$; t test). Conversely, self-care ($P<.001$; t test) and hopefulness ($P<.001$; t test) were significantly lower. Loneliness was also notably higher in the high-risk group ($P<.001$; t test).

The distribution of the top 5 passive data features (ambient light, location entropy, step count, latitude SD, and mean distance from home) highlighted significant differences between risk

XSL•FO
RenderX

groups (Figure 7C). High-risk individuals showed greater ambient light exposure (P<.001; t test), potentially reflecting greater exposure to light at night and sleep disruptions. They also exhibited higher location entropy (P=.02; t test) and latitude variability (P<.001; t test). Additionally, fewer high-risk individuals exceeded 5000 daily steps (P=.002; t test).

Discussion

Principal Findings

Our study demonstrated the effectiveness of integrating active self-reported and passive smartphone sensor data to predict adolescent mental health risks using a novel ML framework. By leveraging data collected via the Mindcraft app, we evaluated predictions across 4 critical mental health outcomes, namely SDQ-high risk, insomnia, suicidal ideation, and eating disorders. Combined models consistently outperformed unimodal approaches, achieving competitive balanced accuracies across all outcomes (eg, 0.77 for suicidal ideation and 0.71 for SDQ-high risk), even in a broad, nonclinical adolescent sample (Figure 4A). These results highlight the complementary value of passive data, which unobtrusively captures continuous behavioral patterns that enrich the subjective insights provided by active data.

Importantly, these results were obtained in a nonclinical, school-based adolescent cohort that was deliberately broad, behaviorally diverse, and not selected for help-seeking status. Such populations introduce substantial intragroup variability, natural fluctuations in engagement, and noise in both active and passive measures, making predictive modeling especially challenging. Previous digital phenotyping studies in non-clinical adolescent or student samples have reported the AUC or F_1-scores typically ranging from 0.60 to 0.80 when predicting stress, depression, or anxiety using mobile sensing [35-41]. Many of these studies also involved smaller or more homogeneous samples and often focused on a single outcome. Our balanced accuracies of 0.70‐0.77 are therefore competitive, especially given that our study simultaneously addressed 4 distinct mental health outcomes in a naturalistic, non–help-seeking adolescent sample. The external validation results suggest that the learned behavioral representations generalize reasonably well to new schools and time periods, while also highlighting some loss of accuracy for the more challenging outcomes.

User engagement patterns indicated sustained utility of passive data collection, underscoring its lower participant burden and feasibility in scalable longitudinal mental health monitoring. Participants preferred less intrusive metrics, including step count, battery usage, and screen brightness, emphasizing the importance of prioritizing user-friendly data collection methods. Our innovative contrastive learning approach effectively addressed variability inherent in daily behavioral data, stabilizing user-specific feature representations. This methodological advancement yielded improved performance and increased confidence in the model's applicability to the real world.

Comparison With Prior Work

Previous work has largely focused on adults or clinical populations [23-40], limiting its applicability to adolescents in community settings. Even among adolescent-focused studies [25,33,34,63,64], most relied solely on passive sensing and targeted depression or anxiety, limiting the generalizability of their findings to broader, nonclinical groups.

Digital self-monitoring has a potential role in multiple stages of the clinical pathway, from prevention to clinical intervention. Our work addresses a broader range of mental health outcomes of internalizing and externalizing disorders, eating disorders, insomnia, and the presence of suicidal ideation in a nonclinical, non–help-seeking school-going adolescent population. MacLeod et al [63], the closest study to ours, included younger adolescents from clinical and nonclinical settings but relied solely on passive sensing. To our knowledge, this study is the first to use ML to accurately predict mental health risk across a broad range of outcomes in low- and higher-risk school-going adolescents, using a combination of active and passive data in a general, nonclinical population.

To interpret our results in light of previous literature, we discuss the findings for each mental health outcome in our study individually below.

SDQ

SHAP analysis highlighted key passive features (Figure 7C), including lower step count, increased location entropy, and elevated ambient light exposure, which were linked to behavioral and environmental patterns associated with mental health risk, consistent with prior studies on physical inactivity, disrupted routines, and light exposure [18-21]. Fewer high-risk individuals exceeded 5000 daily steps, reinforcing associations with sedentary behavior [19,20]. Location entropy, capturing variability in movement, may signal a lack of daily structure [18], while ambient light patterns could reflect disturbed circadian rhythms [21]. In parallel, active features such as negative thinking, racing thoughts, and poor self-care also ranked highly (Figure 7B) and aligned with literature on internalizing and externalizing symptomatology [65-67].

Insomnia

Predictions were driven by active features such as self-care, negative thinking, and appetite, and by passive measures such as app usage hours, nighttime movement, and ambient light (Multimedia Appendix 13). This aligns with earlier research highlighting subjective perceptions, such as increased negative thinking, loneliness, and decreased self-care and hopefulness, as critical factors in sleep disturbances [68,69]. The significance of ambient light exposure further supports evidence linking circadian disruptions to poor sleep quality in adolescents [70,71].

Suicidal Ideation

Key active features (Multimedia Appendix 10), including negative thinking, loneliness, and reduced hopefulness, reflect core cognitive-affective vulnerabilities and are consistent with prior evidence linking these traits to elevated suicide risk in adolescents [72,73]. Passive features (Multimedia Appendix 13) such as screen brightness variability and late-night Mindcraft

XSL•FO
RenderX

app usage may indicate disrupted circadian rhythms, which have been associated with poorer mental health outcomes and suicidal ideation [70,74-76].

Eating Disorder

High-risk individuals reported lower appetite and energy, poorer self-care, and more negative thinking (Multimedia Appendix 11), consistent with links between cognitive-emotional dysregulation, somatic symptoms, and disordered eating in adolescents [77,78]. Passive data (Multimedia Appendix 11) showed consistently elevated screen brightness and earlier app use in the high-risk group, possibly indicating compulsive nighttime device use and disrupted sleep-wake cycles, both associated with emotional dysregulation and body image concerns [79,80].

Strengths and Limitations

This study offers several strengths that advance the field of adolescent digital mental health. By integrating active self-reports and passive smartphone sensor data via the Mindcraft app, we provide a scalable, unobtrusive, and practical framework for early risk detection. Notably, our models maintained strong performance despite high attrition in active data, underscoring the robustness and low participant burden of passive sensing. Additionally, our use of contrastive learning to stabilize day-to-day behavioral features enhanced model robustness. SHAP-based interpretability increased transparency and clinical relevance, both of which are key attributes for adoption in real-world settings.

Several limitations warrant consideration. First, the sample was relatively small and drawn from 3 London-based schools, which may limit the generalizability of our findings, particularly regarding socioeconomic and regional representativeness. Prior research has shown that digital phenotyping features, such as smartphone usage patterns and affective expression, as well as the manifestation and reporting of mental health symptoms, can differ by gender, geography, and cultural context [81-83]. Our geographic concentration in an urban, high-resource setting may therefore not capture behavioral or contextual variability observed in rural or culturally distinct environments, where access to technology, school structures, and sociocultural norms may differ substantially. Furthermore, the gender imbalance driven by the inclusion of a girls-only school may have biased the learned representations toward behavioral and emotional patterns more characteristic of female adolescents, potentially reducing performance for male or nonbinary young people. Consequently, while our findings provide novel insights into adolescents' behavioral monitoring, caution is warranted when extrapolating these results to nonurban, gender-balanced, and resource-limited populations. Future work should therefore include socioeconomically and demographically diverse samples to assess the generalizability of this approach more robustly.

Second, passive sensor data quality varied across device types, operating systems, permission settings, and user engagement, with some participants not enabling key sensors, such as location or app usage, resulting in heterogeneous data completeness. We applied sentinel values to flag missing sensor inputs, enabling the model to learn from patterns of missingness. However, this may not fully eliminate systematic differences, as participants with more complete data may differ meaningfully from those with limited data, potentially skewing model learning. Future work should incorporate fairness-aware modeling and stratified evaluation to ensure equitable performance across subgroups [84]. While cumulative median aggregation improves robustness to short-term noise, it may dampen sensitivity to clinically meaningful abrupt behavioral changes, such as those observed during acute mood episodes or crisis events. Hybrid strategies such as combining cumulative medians with volatility-sensitive features may better capture both stable patterns and sudden shifts.

Finally, this study did not include broader biological and environmental factors, such as genetic risk, socioeconomic status, family history, adverse childhood experiences, or neurodevelopmental profiles, that are known to influence adolescent mental health [81,85]. The absence of such contextual and historical information may limit the model's ability to capture all relevant sources of variance, potentially constraining predictive accuracy. Future studies should consider incorporating these factors to enhance explanatory power and clinical utility.

Implications and Recommendations

Active data engagement declined markedly by day 14 (Figure 2B), underscoring a key longitudinal feasibility challenge. Future iterations of Mindcraft will therefore incorporate specific design adaptations to reduce burden and sustain engagement. These include shifting from fixed-time prompts to context-aware adaptive sampling triggered by behavioral anomalies detected through passive data; incorporating light gamification (such as streaks) to maintain motivation; and closing the feedback loop by providing personalized behavioral insights and just-in-time recommendations informed by each user's active and passive data. Together, these adaptations aim to shift Mindcraft from a one-way data collection tool to a personalized support platform, thereby improving long-term engagement and feasibility.

Real-time digital phenotyping at a population level can complement traditional screening methods by identifying and prioritizing high-risk individuals and enabling tailored prevention and early intervention strategies [76,86]. The use of a mobile app for digital phenotyping is particularly valuable for children and young people, for whom early identification and intervention are essential to prevent the onset of more severe mental health issues in adulthood. When implemented in schools, it addresses barriers such as stigma and accessibility, offering adolescents a preventive tool that empowers them to manage their mental health. Digital phenotyping provides the opportunity to inform school-based digital interventions that might be central to early intervention and prevention of mental health problems in the community [87]. While adherence to GDPR and secure data protocols provides a legal baseline, the ethical landscape of adolescent digital phenotyping extends far beyond regulatory compliance [88,89]. Collecting continuous passive data from minors introduces distinct tensions regarding autonomy and informed consent; specifically, the invisible nature of passive sensing means adolescents may habituate to the monitoring, potentially eroding their ongoing awareness of data sharing [89]. Furthermore, the deployment of predictive

XSL•FO
RenderX

risk models carries the risk of digital labeling, where algorithmic outputs, if misinterpreted or generating false positives, could lead to stigma, unnecessary anxiety, or oversurveillance [88]. Addressing these complexities demands more than static consent forms; it requires ongoing participatory approaches involving adolescents, parents, and clinicians, alongside governance mechanisms that ensure transparency, prioritize interpretability over black-box predictions, and maintain human clinical oversight of algorithmic outputs [88]. To this end, future work must empirically evaluate how adolescents understand, experience, and respond to continuous passive sensing, and determine how ethical frameworks can best support safe, acceptable integration in school settings.

Achieving real-world feasibility requires distinguishing between scientific interpretability and user-centric explainability [90,91]. While the SHAP values presented in this study provide necessary transparency for model validation, raw feature importance scores are unlikely to be meaningful to non-expert stakeholders such as adolescents, parents, or educators. For broad deployment, these technical outputs must be bridged by a translation layer that converts granular risk estimates into accessible, actionable narratives. For instance, rather than displaying a high SHAP value for "location entropy," a user-facing interface should translate this into an intuitive insight, such as detecting "significant changes in your daily routine." Future implementation work must prioritize the co-design of these explanatory interfaces to ensure that algorithmic transparency supports understanding rather than overwhelming users.

Digital biomarkers from sensors and ML have shown accuracy in predicting disease progression [92,93], underscoring the broader potential of sensor-based technologies for personalized healthcare. Smartphones, being both ubiquitous and affordable, enable continuous, real-time data collection even in low-resource settings, reducing reliance on clinical supervision [34,94]. Digital phenotyping offers a scalable mechanism for continuous, context-aware monitoring [94], which could feed into well-being dashboards accessible to pastoral staff, school counselors, or clinical teams. In school settings, such dashboards could support early identification of distress and enable stepped-care approaches that match support intensity to need. In clinical pathways, risk predictions could support triage decisions and monitoring between appointments, complementing existing services rather than replacing them [86]. Achieving this will require implementation research on workflow integration, governance, and alignment with school well-being policies and child mental health services.

Traditional platforms such as Childline rely on proactive engagement from children and young people, creating barriers for disengaged users. In contrast, Mindcraft's passive tracking capabilities offer a proactive approach by identifying early signs of poor mental health and prompting timely professional interventions. Building on this pilot work, we have developed a school-based intervention study that evaluates the effectiveness of personalized artificial intelligence recommendations delivered through the Mindcraft app across schools in the United Kingdom (ISRCTN11686798). With this further development, Mindcraft is evolving into a comprehensive platform that delivers in-app recommendations informed by active and passive data, leveraging user profiles [95,96] to tailor suggestions to individual needs, enhance engagement, and improve intervention effectiveness. Subsequent phases will focus on scaling and validation across diverse school settings and on integrating Mindcraft with existing education and mental health pathways to support sustainable, real-world implementation. This integration of proactive detection and tailored intervention could help address significant gaps in traditional mental health support systems.

Conclusion

In conclusion, this study underscores the transformative potential of integrating active and passive smartphone data to predict adolescent mental health. By leveraging innovative ML techniques, such as contrastive learning, and the scalability of tools like the Mindcraft app, we present a robust framework for early risk detection across diverse mental health outcomes. These findings lay the groundwork for more inclusive, accessible, and personalized early detection and intervention strategies in adolescent mental health.

Funding

TBB is supported by a Fellowship funded by the Fundación Alicia Koplowitz. MDS is supported by the NIHR Imperial Biomedical Research Collaboration and NIHR Mental Health Translational Research Collaboration. DN is supported by the National Institute for Health Research (NIHR) Applied Research Collaboration Northwest London and NIHR Imperial Biomedical Research Collaboration. AAF acknowledges support from the United Kingdom Research and Innovation Turing AI Fellowship (EP/V025449/1).

Data Availability

The study data are not publicly available because of the General Data Protection Regulation restrictions and privacy policies outlined in the study information provided to participants. The data sets generated during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

DN, MDS, and AAF conceptualized the study. TBB, DN, and MDS gained ethical approval and conducted recruitment and data collection. BK curated the data and performed data analysis and machine learning under the supervision of AAF. BK, TBB, and

XSL•FO
RenderX

AF drafted the original manuscript. MDS, DN, and AAF critically reviewed and edited the manuscript. All authors reviewed and approved the final version of the manuscript for publication.

BK and TBB contributed equally as joint first authors; additionally, DN and AAF contributed equally as joint last authors. DN, the clinical corresponding author, can be reached at d.nicholls@imperial.ac.uk.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Active data features, including feature descriptions and Spearman correlations with mental health outcomes (SDQ, SCI, ED-15, and suicidal ideation).
[DOCX File, 47 KB - jmir_v28i1e72501_app1.docx]

Multimedia Appendix 2
Passive data features engineered from smartphone sensors, including feature descriptions and Spearman correlations with mental health outcomes (SDQ, SCI, ED-15, and suicidal ideation).
[DOCX File, 55 KB - jmir_v28i1e72501_app2.docx]

Multimedia Appendix 3
Example illustrating the behavior of the cumulative median compared to raw daily step count values for a single participant. The cumulative median smooths isolated anomalies (eg, a one-day spike on Day 3) while remaining sensitive to sustained behavioral shifts (eg, a persistent drop beginning on Day 7).
[PNG File, 186 KB - jmir_v28i1e72501_app3.png]

Multimedia Appendix 4
Balanced accuracy (mean [SD] across 10 different runs of the experiment) for models trained on raw daily features, cumulative mean-aggregated and cumulative median–aggregated features of the combined active and passive data for the mental health outcomes.
[DOCX File, 45 KB - jmir_v28i1e72501_app4.docx]

Multimedia Appendix 5
CONSORT (Consolidated Standards of Reporting Trials)-style flow diagram showing participant inclusion for the main analysis. The main analysis included 67 users who provided at least one active self-report and enabled at least one passive data stream.
[PNG File, 174 KB - jmir_v28i1e72501_app5.png]

Multimedia Appendix 6
Distribution of mental health assessment scores across participants. (A) Strengths and Difficulties Questionnaire (SDQ) scores, representing overall mental health and behavioral difficulties. (B) Sleep Condition Indicator (SCI) scores, assessing sleep quality and potential insomnia. (C) Frequency of self-reported suicidal ideations over a 2-week period. (D) Eating Disorder Examination Questionnaire (ED-15) scores, evaluating symptoms associated with eating disorders.
[PNG File, 151 KB - jmir_v28i1e72501_app6.png]

Multimedia Appendix 7
Fully annotated correlation heatmap of active data features, with Spearman correlation coefficients displayed in each cell. This supplementary version is provided to support the complete numerical transparency of the simplified heatmap shown in Figure 3B of the main manuscript.
[PNG File, 539 KB - jmir_v28i1e72501_app7.png]

Multimedia Appendix 8
Mean (SE) *z* score–normalized values of the top five active and passive features that significantly differed between low- and high-risk groups across 4 mental health outcomes: SDQ, insomnia, suicidal ideation, and eating disorder. Group differences were assessed using one-sided t-tests with Benjamini–Hochberg correction for multiple comparisons. Features were selected based on adjusted p-values. Red asterisks indicate statistical significance (*$P \leq .05$, **$P \leq .01$, ***$P \leq .001$). Normalization was applied to enable comparison across features measured on different scales.
[PNG File, 241 KB - jmir_v28i1e72501_app8.png]

XSL•FO
RenderX

Multimedia Appendix 9
Comparison of balanced accuracy of mental health outcome predictions using active data only for all users in the study (N=103) vs users who also enabled passive data collection (N=67).
[PNG File, 156 KB - jmir_v28i1e72501_app9.png]

Multimedia Appendix 10
Detailed performance metrics for predicting SDQ high risk, insomnia, suicidal ideation, and eating disorder risk using combined active and passive data in an external validation sample (N=45 adolescents not used for model development).
[DOCX File, 46 KB - jmir_v28i1e72501_app10.docx]

Multimedia Appendix 11
Fairness analysis across gender and school context for the combined active and passive data model. Balanced accuracy for predicting SDQ high risk, insomnia, suicidal ideation, and eating disorder risk stratified by (A) gender (female vs male/other) and (B–D) school context. Error bars represent SD across 10 different runs of the experiment.
[PNG File, 225 KB - jmir_v28i1e72501_app11.png]

Multimedia Appendix 12
Fairness analysis of the combined active and passive data model across passive-sensor availability. Balanced accuracy for users with less than or equal to 3 enabled sensors and greater than 3 enabled sensors for each mental health outcome. Error bars represent the SD across 10 different runs of the experiment.
[PNG File, 98 KB - jmir_v28i1e72501_app12.png]

Multimedia Appendix 13
Feature importance analysis for predicting the insomnia high-risk category using both active and passive data. (A) SHAP-based feature importances, with passive data aggregated by sensor type and active data shown individually. (B) Distribution of the top five active data features across insomnia risk categories. (C) Distribution of the top five passive data features across insomnia risk categories. Statistically significant differences between low-risk and high-risk groups are indicated (*P<.05, **P<.01, ***P<.001, t test).
[PNG File, 261 KB - jmir_v28i1e72501_app13.png]

Multimedia Appendix 14
Feature importance analysis for predicting the suicidal ideation high-risk category using both active and passive data. (A) SHAP-based feature importances, with passive data aggregated by sensor type and active data shown individually. (B) Distribution of the top five active data features across suicidal ideation risk categories. (C) Distribution of the top five passive data features across suicidal ideation risk categories. Statistically significant differences between low-risk and high-risk groups are indicated (*P<.05, **P<.01, ***P<.001, t test).
[PNG File, 276 KB - jmir_v28i1e72501_app14.png]

Multimedia Appendix 15
Feature importance analysis for predicting the eating disorder high-risk category using both active and passive data. (A) SHAP-based feature importances, with passive data aggregated by sensor type and active data shown individually. (B) Distribution of the top five active data features across eating disorder risk categories. (C) Distribution of the top five passive data features across eating disorder risk categories. Statistically significant differences between low-risk and high-risk groups are indicated (*P<.05, **P<.01, ***P<.001, t test).
[PNG File, 271 KB - jmir_v28i1e72501_app15.png]

References

1. Mughal F, England E. The mental health of young people: the view from primary care. Br J Gen Pract 2016 Oct;66(651):502-503. [doi: 10.3399/bjgp16X687133] [Medline: 27688487]
2. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005 Jun;62(6):593-602. [doi: 10.1001/archpsyc.62.6.593] [Medline: 15939837]
3. Mental health of adolescents. World Health Organization. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health [accessed 2025-05-22]
4. Mathiesen KS, Sanson A, Stoolmiller M, Karevold E. The nature and predictors of undercontrolled and internalizing problem trajectories across early childhood. J Abnorm Child Psychol 2009 Feb;37(2):209-222. [doi: 10.1007/s10802-008-9268-y] [Medline: 18766436]

XSL•FO
RenderX

5. Samek DR, Hicks BM. Externalizing disorders and environmental risk: mechanisms of gene-environment interplay and strategies for intervention. Clin Pract (Lond) 2014;11(5):537-547. [doi: 10.2217/CPR.14.47] [Medline: 25485087]
6. Kim-Cohen J, Caspi A, Moffitt TE, Harrington H, Milne BJ, Poulton R. Prior juvenile diagnoses in adults with mental disorder: developmental follow-back of a prospective-longitudinal cohort. Arch Gen Psychiatry 2003 Jul;60(7):709-717. [doi: 10.1001/archpsyc.60.7.709] [Medline: 12860775]
7. Polanczyk GV, Salum GA, Sugaya LS, Caye A, Rohde LA. Annual research review: a meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. J Child Psychol Psychiatry 2015 Mar;56(3):345-365. [doi: 10.1111/jcpp.12381] [Medline: 25649325]
8. Murray CJL. The global burden of disease study at 30 years. Nat Med 2022 Oct;28(10):2019-2026. [doi: 10.1038/s41591-022-01990-1] [Medline: 36216939]
9. Gulliver A, Griffiths KM, Christensen H. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. BMC Psychiatry 2010 Dec 30;10:1-9. [doi: 10.1186/1471-244X-10-113] [Medline: 21192795]
10. Bonell C, Jamal F, Harden A, et al. Systematic review of the effects of schools and school environment interventions on health: evidence mapping and synthesis. Public Health Research 2013;1(1):1-320. [doi: 10.3310/phr01010]
11. Anderson JK, Ford T, Soneson E, et al. A systematic review of effectiveness and cost-effectiveness of school-based identification of children and young people at risk of, or currently experiencing mental health difficulties. Psychol Med 2019 Jan;49(1):9-19. [doi: 10.1017/S0033291718002490] [Medline: 30208985]
12. Larsen ME, Huckvale K, Nicholas J, et al. Using science to sell apps: evaluation of mental health app store quality claims. NPJ Digit Med 2019;2(1):18. [doi: 10.1038/s41746-019-0093-1] [Medline: 31304366]
13. Donker T, Petrie K, Proudfoot J, Clarke J, Birch MR, Christensen H. Smartphones for smarter delivery of mental health programs: a systematic review. J Med Internet Res 2013 Nov 15;15(11):e247. [doi: 10.2196/jmir.2791] [Medline: 24240579]
14. Eisenstadt M, Liverpool S, Infanti E, Ciuvat RM, Carlsson C. Mobile apps that promote emotion regulation, positive mental health, and well-being in the general population: systematic review and meta-analysis. JMIR Ment Health 2021 Nov 8;8(11):e31170. [doi: 10.2196/31170] [Medline: 34747713]
15. Oudin A, Maatoug R, Bourla A, et al. Digital phenotyping: data-driven psychiatry to redefine mental health. J Med Internet Res 2023 Oct 4;25:e44502. [doi: 10.2196/44502] [Medline: 37792430]
16. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA 2017 Oct 3;318(13):1215-1216. [doi: 10.1001/jama.2017.11295] [Medline: 28973224]
17. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacol 2016 Jun;41(7):1691-1696. [doi: 10.1038/npp.2016.7]
18. Shin J, Bae SM. A systematic review of location data for depression prediction. Int J Environ Res Public Health 2023 May 29;20(11):5984. [doi: 10.3390/ijerph20115984] [Medline: 37297588]
19. Kandola A, Lewis G, Osborn DPJ, Stubbs B, Hayes JF. Depressive symptoms and objectively measured physical activity and sedentary behaviour throughout adolescence: a prospective cohort study. Lancet Psychiatry 2020 Mar;7(3):262-271. [doi: 10.1016/S2215-0366(20)30034-1] [Medline: 32059797]
20. Brown HE, Pearson N, Braithwaite RE, Brown WJ, Biddle SJH. Physical activity interventions and depression in children and adolescents: a systematic review and meta-analysis. Sports Med 2013 Mar;43(3):195-206. [doi: 10.1007/s40279-012-0015-8] [Medline: 23329611]
21. Obayashi K, Saeki K, Iwamoto J, Ikada Y, Kurumatani N. Exposure to light at night and risk of depression in the elderly. J Affect Disord 2013 Oct;151(1):331-336. [doi: 10.1016/j.jad.2013.06.018] [Medline: 23856285]
22. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu Rev Clin Psychol 2017 May 8;13(1):23-47. [doi: 10.1146/annurev-clinpsy-032816-044949] [Medline: 28375728]
23. Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. Psychiatr Rehabil J 2015 Sep;38(3):218-226. [doi: 10.1037/prj0000130] [Medline: 25844912]
24. Rickard N, Arjmand HA, Bakker D, Seabrook E. Development of a mobile phone app to support self-monitoring of emotional well-being: a mental health digital innovation. JMIR Ment Health 2016 Nov 23;3(4):e49. [doi: 10.2196/mental.6202] [Medline: 27881358]
25. Cao J, Truong AL, Banu S, Shah AA, Sabharwal A, Moukaddam N. Tracking and predicting depressive symptoms of adolescents using smartphone-based self-reports, parental evaluations, and passive phone sensor data: development and usability study. JMIR Ment Health 2020 Jan 24;7(1):e14045. [doi: 10.2196/14045] [Medline: 32012072]
26. Faurholt-Jepsen M, Vinberg M, Frost M, Christensen EM, Bardram J, Kessing LV. Daily electronic monitoring of subjective and objective measures of illness activity in bipolar disorder using smartphones--the MONARCA II trial protocol: a randomized controlled single-blind parallel-group trial. BMC Psychiatry 2014 Nov 25;14(1):309. [doi: 10.1186/s12888-014-0309-5] [Medline: 25420431]
27. Berrouiguet S, Ramírez D, Barrigón ML, et al. Combining continuous smartphone native sensors data capture and unsupervised data mining techniques for behavioral changes detection: a case series of the evidence-based behavior (eB2) study. JMIR Mhealth Uhealth 2018 Dec 10;6(12):e197. [doi: 10.2196/mhealth.9472] [Medline: 30530465]

XSL•FO
RenderX

28. Wang R, Aung MS, Abdullah S, et al. CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia. 2016 Presented at: Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing; Sep 12-16, 2016. [doi: 10.1145/2971648.2971740]
29. Beiwinkel T, Kindermann S, Maier A, et al. Using smartphones to monitor bipolar disorder symptoms: a pilot study. JMIR Ment Health 2016 Jan 6;3(1):e2. [doi: 10.2196/mental.4560] [Medline: 26740354]
30. Jacobson NC, Chung YJ. Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. Sensors (Basel) 2020 Jun 24;20(12):3572. [doi: 10.3390/s20123572] [Medline: 32599801]
31. Opoku Asare K, Visuri A, Ferreira DST. Towards early detection of depression through smartphone sensing. 2019 Sep 9 Presented at: UbiComp '19 URL: https://dl.acm.org/doi/proceedings/10.1145/3341162 [doi: 10.1145/3341162.3347075]
32. Dargél AA, Mosconi E, Masson M, et al. Toi Même, a mobile health platform for measuring bipolar illness activity: protocol for a feasibility study. JMIR Res Protoc 2020 Aug 18;9(8):e18818. [doi: 10.2196/18818] [Medline: 32638703]
33. Mullick T, Radovic A, Shaaban S, Doryab A. Predicting depression in adolescents using mobile and wearable sensors: multimodal machine learning-based exploratory study. JMIR Form Res 2022 Jun 24;6(6):e35807. [doi: 10.2196/35807] [Medline: 35749157]
34. Maharjan SM, Poudyal A, van Heerden A, et al. Passive sensing on mobile devices to improve mental health services with adolescent and young mothers in low-resource settings: the role of families in feasibility and acceptability. BMC Med Inform Decis Mak 2021 Apr 7;21(1):117. [doi: 10.1186/s12911-021-01473-2] [Medline: 33827552]
35. Thakur SS, Roy RB. Predicting mental health using smart-phone usage and sensor data. J Ambient Intell Human Comput 2021 Oct;12(10):9145-9161. [doi: 10.1007/s12652-020-02616-5]
36. Wang W, Nepal S, Huckins JF, et al. First-gen lens: assessing mental health of first-generation students across their first year at college using mobile sensing. Proc ACM Interact Mob Wearable Ubiquitous Technol 2022 Jul;6(2):1-32. [doi: 10.1145/3543194] [Medline: 36561350]
37. Acikmese Y, Alptekin SE. Prediction of stress levels with LSTM and passive mobile sensors. Procedia Comput Sci 2019;159:658-667. [doi: 10.1016/j.procs.2019.09.221]
38. Rhim S, Lee U, Han K. Tracking and modeling subjective well-being using smartphone-based digital phenotype. In: Rhim S, Lee U, Han K, editors. 2020 Jul 7 Presented at: UMAP '20 URL: https://dl.acm.org/doi/proceedings/10.1145/3340631 [accessed 2025-12-23] [doi: 10.1145/3340631.3394855]
39. Xu X, Liu X, Zhang H, Wang W, Nepal S, Sefidgar Y, et al. GLOBEM: cross-dataset generalization of longitudinal human behavior modeling. Proc ACM Interact Mob Wearable Ubiquitous Technol 2022;6(4):1-34. [doi: 10.1145/3569485]
40. Currey D, Torous J. Digital phenotyping correlations in larger mental health samples: analysis and replication. BJPsych Open 2022 Jun 3;8(4):e106. [doi: 10.1192/bjo.2022.507] [Medline: 35657687]
41. Xu X, Chikersal P, Doryab A, et al. Leveraging routine behavior and contextually-filtered features for depression detection among college students. Proc ACM Interact Mob Wearable Ubiquitous Technol 2019 Sep 9;3(3):1-33. [doi: 10.1145/3351274]
42. Cornet VP, Holden RJ. Systematic review of smartphone-based passive sensing for health and wellbeing. J Biomed Inform 2018 Jan;77:120-132. [doi: 10.1016/j.jbi.2017.12.008] [Medline: 29248628]
43. Choo M, Park D, Cho M, Bae S, Kim J, Han DH. Exploring a multimodal approach for utilizing digital biomarkers for childhood mental health screening. Front Psychiatry 2024;15:1348319. [doi: 10.3389/fpsyt.2024.1348319]
44. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. Biol Psychiatry Cogn Neurosci Neuroimaging 2018 Mar;3(3):223-230. [doi: 10.1016/j.bpsc.2017.11.007] [Medline: 29486863]
45. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychol Med 2019 Jul;49(9):1426-1448. [doi: 10.1017/S0033291719000151] [Medline: 30744717]
46. Goodman R. The strengths and difficulties questionnaire: a research note. J Child Psychol Psychiatry 1997 Jul;38(5):581-586. [doi: 10.1111/j.1469-7610.1997.tb01545.x] [Medline: 9255702]
47. Tatham M, Turner H, Mountford VA, Tritt A, Dyas R, Waller G. Development, psychometric properties and preliminary clinical validation of a brief, session-by-session measure of eating disorder cognitions and behaviors: the ED-15. Int J Eat Disord 2015 Nov;48(7):1005-1015. [doi: 10.1002/eat.22430] [Medline: 26011054]
48. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
49. Fonseca-Pedrero E, Díez-Gómez A, Pérez-Albéniz A, Al-Halabí S, Lucas-Molina B, Debbané M. Youth screening depression: validation of the Patient Health Questionnaire-9 (PHQ-9) in a representative sample of adolescents. Psychiatry Res 2023 Oct;328:115486. [doi: 10.1016/j.psychres.2023.115486] [Medline: 37738682]
50. Espie CA, Kyle SD, Hames P, Gardani M, Fleming L, Cape J. The sleep condition indicator: a clinical screening tool to evaluate insomnia disorder. BMJ Open 2014 Mar 18;4(3):e004183. [doi: 10.1136/bmjopen-2013-004183] [Medline: 24643168]
51. Espie CA, Farias Machado P, Carl JR, et al. The sleep condition indicator: reference values derived from a sample of 200 000 adults. J Sleep Res 2018 Jun;27(3):e12643 [FREE Full text] [doi: 10.1111/jsr.12643]

XSL•FO
RenderX

52. Kadirvelu B, Bellido Bel T, Wu X, et al. Mindcraft, a mobile mental health monitoring platform for children and young people: development and acceptability pilot study. JMIR Form Res 2023 Jun 26;7:e44877. [doi: 10.2196/44877] [Medline: 37358901]
53. Goodman R, Ford T, Simmons H, Gatward R, Meltzer H. Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. Br J Psychiatry 2000 Dec;177(6):534-539. [doi: 10.1192/bjp.177.6.534] [Medline: 11102329]
54. Rodrigues T, Vaz AR, Silva C, Conceição E, Machado PPP. Eating Disorder-15 (ED-15): factor structure, psychometric properties, and clinical validation. Eur Eat Disord Rev 2019 Nov;27(6):682-691. [doi: 10.1002/erv.2694] [Medline: 31257707]
55. Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering. In: Schroff F, Kalenichenko D, Philbin J, editors. 2015 Presented at: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [doi: 10.1109/CVPR.2015.7298682]
56. Hoffer E, Ailon N, editors. Deep metric learning using triplet network. 2015 Presented at: Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015; Oct 12-14, 2015. [doi: 10.1007/978-3-319-24261-3_7]
57. Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification. arXiv. Preprint posted online on 2017. [doi: 10.48550/arXiv.1703.07737]
58. Chen T, Kornblith S, Norouzi M. A simple framework for contrastive learning of visual representations. In: Hinton G, editor. 2020 Presented at: Proceedings of the 37th International Conference on Machine Learning (ICML 2020); Jul 13-18, 2020.
59. Lundberg S. A unified approach to interpreting model predictions. . Preprint posted online on May 22, 2017. [doi: 10.48550/arXiv.1705.07874]
60. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):56-67. [doi: 10.1038/s42256-019-0138-9]
61. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst 2018;31:6638-6648. [doi: 10.5555/3327757.3327770]
62. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. J Big Data 2020;7(1):94. [doi: 10.1186/s40537-020-00369-8] [Medline: 33169094]
63. MacLeod L, Suruliraj B, Gall D, et al. A mobile sensing app to monitor youth mental health: observational pilot study. JMIR Mhealth Uhealth 2021 Oct 26;9(10):e20638. [doi: 10.2196/20638] [Medline: 34698650]
64. Ware S, Yue C, Morillo R, et al. Predicting depressive symptoms using smartphone data. Smart Health (2014) 2020 Mar;15:100093. [doi: 10.1016/j.smhl.2019.100093]
65. LeMoult J, Gotlib IH. Depression: a cognitive perspective. Clin Psychol Rev 2019 Apr;69:51-66. [doi: 10.1016/j.cpr.2018.06.008] [Medline: 29961601]
66. Piguet C, Dayer A, Kosel M, Desseilles M, Vuilleumier P, Bertschy G. Phenomenology of racing and crowded thoughts in mood disorders: a theoretical reappraisal. J Affect Disord 2010 Mar;121(3):189-198. [doi: 10.1016/j.jad.2009.05.006] [Medline: 19515428]
67. Town R, Hayes D, March A, Fonagy P, Stapley E. Self-management, self-care, and self-help in adolescents with emotional problems: a scoping review. Eur Child Adolesc Psychiatry 2024 Sep;33(9):2929-2956. [doi: 10.1007/s00787-022-02134-z] [Medline: 36641785]
68. Di Benedetto M, Towt CJ, Jackson ML. A cluster analysis of sleep quality, self-care behaviors, and mental health risk in Australian University students. Behav Sleep Med 2020;18(3):309-320. [doi: 10.1080/15402002.2019.1580194] [Medline: 30821507]
69. Kurina LM, Knutson KL, Hawkley LC, Cacioppo JT, Lauderdale DS, Ober C. Loneliness is associated with sleep fragmentation in a communal society. Sleep 2011 Nov 1;34(11):1519-1526. [doi: 10.5665/sleep.1390] [Medline: 22043123]
70. Chang AM, Aeschbach D, Duffy JF, Czeisler CA. Evening use of light-emitting eReaders negatively affects sleep, circadian timing, and next-morning alertness. Proc Natl Acad Sci USA 2015 Jan 27;112(4):1232-1237. [doi: 10.1073/pnas.1418490112]
71. Auger RR, Burgess HJ, Dierkhising RA, Sharma RG, Slocumb NL. Light exposure among adolescents with delayed sleep phase disorder: a prospective cohort study. Chronobiol Int 2011 Dec;28(10):911-920. [doi: 10.3109/07420528.2011.619906] [Medline: 22080736]
72. Franklin JC, Ribeiro JD, Fox KR, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol Bull 2017 Feb;143(2):187-232. [doi: 10.1037/bul0000084] [Medline: 27841450]
73. Gallagher M, Prinstein MJ, Simon V, Spirito A. Social anxiety symptoms and suicidal ideation in a clinical sample of early adolescents: examining loneliness and social support as longitudinal mediators. J Abnorm Child Psychol 2014 Aug;42(6):871-883. [doi: 10.1007/s10802-013-9844-7] [Medline: 24390470]
74. Littlewood DL, Kyle SD, Carter LA, Peters S, Pratt D, Gooding P. Short sleep duration and poor sleep quality predict next-day suicidal ideation: an ecological momentary assessment study. Psychol Med 2019 Feb;49(3):403-411. [doi: 10.1017/S0033291718001009] [Medline: 29697037]

XSL•FO
RenderX

75. Pizzoli SFM, Monzani D, Conti L, Ferraris G, Grasso R, Pravettoni G. Issues and opportunities of digital phenotyping: ecological momentary assessment and behavioral sensing in protecting the young from suicide. Front Psychol 2023;14:1103703. [doi: 10.3389/fpsyg.2023.1103703]
76. Melcher J, Hays R, Torous J. Digital phenotyping for mental health of college students: a clinical review. Evid Based Mental Health 2020 Nov;23(4):161-166. [doi: 10.1136/ebmental-2020-300180]
77. Paulus FW, Ohmann S, Möhler E, Plener P, Popow C. Emotional dysregulation in children and adolescents with psychiatric disorders. A narrative review. Front Psychiatry 2021;12:628252. [doi: 10.3389/fpsyt.2021.628252] [Medline: 34759846]
78. Lavender JM, Wonderlich SA, Engel SG, Gordon KH, Kaye WH, Mitchell JE. Dimensions of emotion dysregulation in anorexia nervosa and bulimia nervosa: a conceptual review of the empirical literature. Clin Psychol Rev 2015 Aug;40:111-122. [doi: 10.1016/j.cpr.2015.05.010] [Medline: 26112760]
79. Allison KC, Spaeth A, Hopkins CM. Sleep and eating disorders. Curr Psychiatry Rep 2016 Oct;18(10):1-8. [doi: 10.1007/s11920-016-0728-8] [Medline: 27553980]
80. Yang H, Wang JJ, Tng GYQ, Yang S. Effects of social media and smartphone use on body esteem in female adolescents: testing a cognitive and affective model. Children (Basel) 2020 Sep 21;7(9):148. [doi: 10.3390/children7090148] [Medline: 32967376]
81. Reiss F. Socioeconomic inequalities and mental health problems in children and adolescents: a systematic review. Soc Sci Med 2013 Aug;90:24-31. [doi: 10.1016/j.socscimed.2013.04.026]
82. Salk RH, Hyde JS, Abramson LY. Gender differences in depression in representative national samples: meta-analyses of diagnoses and symptoms. Psychol Bull 2017 Aug;143(8):783-822. [doi: 10.1037/bul0000102] [Medline: 28447828]
83. Zainal NH, Wang V, Garthwaite B, Curtiss JE. What factors are related to engagement with digital mental health interventions (DMHIs)? A meta-analysis of 117 trials. Health Psychol Rev 2025 Aug 26:1-21. [doi: 10.1080/17437199.2025.2547610] [Medline: 40857362]
84. Buolamwini J, Gebru T, editors. Gender shades: intersectional accuracy disparities in commercial gender classification. Presented at: Conference on fairness, accountability and transparency; Feb 23-24, 2018 URL: https://proceedings.mlr.press/v81/buolamwini18a.html [accessed 2026-01-22]
85. Patel V, Flisher AJ, Hetrick S, McGorry P. Mental health of young people: a global public-health challenge. The Lancet 2007 Apr;369(9569):1302-1313. [doi: 10.1016/S0140-6736(07)60368-7]
86. Bufano P, Laurino M, Said S, Tognetti A, Menicucci D. Digital phenotyping for monitoring mental disorders: systematic review. J Med Internet Res 2023 Dec 13;25:e46778. [doi: 10.2196/46778] [Medline: 38090800]
87. Cohen KA, Ito S, Ahuvia IL, et al. Brief school-based interventions targeting student mental health or well-being: a systematic review and meta-analysis. Clin Child Fam Psychol Rev 2024 Sep;27(3):732-806. [doi: 10.1007/s10567-024-00487-2] [Medline: 38884838]
88. Martinez-Martin N, Greely HT, Cho MK. Ethical development of digital phenotyping tools for mental health applications: delphi study. JMIR Mhealth Uhealth 2021 Jul 28;9(7):e27343. [doi: 10.2196/27343] [Medline: 34319252]
89. Wies B, Landers C, Ienca M. Digital mental health for young people: a scoping review of ethical promises and challenges. Front Digit Health 2021;3:697072. [doi: 10.3389/fdgth.2021.697072] [Medline: 34713173]
90. Fleming W, Coutts A, Pochard D, Trivedi D, Sanderson K. Human-centered design and digital transformation of mental health services. JMIR Hum Factors 2025 Aug 11;12:e66040. [doi: 10.2196/66040] [Medline: 40789172]
91. Miller T. Explanation in artificial intelligence: insights from the social sciences. Artif Intell 2019 Feb;267:1-38. [doi: 10.1016/j.artint.2018.07.007]
92. Kadirvelu B, Gavriel C, Nageshwaran S, et al. A wearable motion capture suit and machine learning predict disease progression in Friedreich's ataxia. Nat Med 2023 Jan;29(1):86-94. [doi: 10.1038/s41591-022-02159-6] [Medline: 36658420]
93. Ricotti V, Kadirvelu B, Selby V, et al. Wearable full-body motion tracking of activities of daily living predicts disease trajectory in Duchenne muscular dystrophy. Nat Med 2023 Jan;29(1):95-103. [doi: 10.1038/s41591-022-02045-1] [Medline: 36658421]
94. Khosravi M, Azar G. A systematic review of reviews on the advantages of mHealth utilization in mental health services: a viable option for large populations in low-resource settings. Glob Ment Health (Camb) 2024;11:e43. [doi: 10.1017/gmh.2024.39] [Medline: 38690573]
95. Pieritz S, Khwaja M, Faisal AA. In: Matic A, editor. Personalised Recommendations in Mental Health Apps: The Impact of Autonomy and Data Sharing: Association for Computing Machinery (ACM); 2021. [doi: 10.1145/3411764.3445523]
96. Khwaja M, Pieritz S. In: Matic A, editor. Personality and Engagement with Digital Mental Health Interventions: Association for Computing Machinery (ACM); 2021. [doi: 10.1145/3450613.3456823]

Abbreviations

AUC: area under the receiver operating characteristic curve
CONSORT: Consolidated Standards of Reporting Trials
ED-15: Eating Disorder-15 questionnaire
GDPR: General Data Protection Regulation

XSL•FO
RenderX

LOSO CV: leave-one-subject-out cross-validation
ML: machine learning
MLP: multilayer perceptron
PHQ-9: Patient Health Questionnaire version 9
SCI: Sleep Condition Indicator
SDQ: Strengths and Difficulties Questionnaire
SES: socioeconomic status
SHAP: Shapley Additive Explanations

Edited by J Sarvestan; submitted 11.Feb.2025; peer-reviewed by CY Chen, JH Song; revised version received 04.Dec.2025; accepted 05.Dec.2025; published 04.Feb.2026.

Please cite as:
Kadirvelu B, Bellido Bel T, Freccero A, Di Simplico M, Nicholls D, Faisal AA
Digital Phenotyping for Adolescent Mental Health: Feasibility Study Using Machine Learning to Predict Mental Health Risk From Active and Passive Smartphone Data
J Med Internet Res 2026;28:e72501
URL: https://www.jmir.org/2026/1/e72501
doi:10.2196/72501

© Balasundaram Kadirvelu, Teresa Bellido Bel, Aglaia Freccero, Martina Di Simplico, Dasha Nicholls, A Aldo Faisal. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Effectiveness of the Headspace App for Improving Sleep: Randomized Controlled Trial

Zoltan Andre Torok, PhD; Larisa Gavrilova, PhD; Amish Patel, BA; Matthew Jason Zawadzki, PhD

Department of Psychological Sciences, University of California, 5200 North Lake Road, Merced, CA, United States

Corresponding Author:
Zoltan Andre Torok, PhD
Department of Psychological Sciences, University of California, 5200 North Lake Road, Merced, CA, United States

Abstract

Background: Improving sleep is critical for optimizing short-term and long-term health. Although in-person meditation training has been shown to impact sleep positively, there is a gap in our understanding of whether apps that teach self-guided meditation are also effective.

Objective: This study aims to test whether Headspace (Headspace, Inc) improves sleep quality, tiredness, sleep duration, and sleep efficiency.

Methods: Staff employees (N=135; mean age 38.1, SD 10.9; 75.0% female; 59.3% non-Hispanic White; 27.1% Hispanic) from a university in California's San Joaquin Valley participated in the study. Participants were randomized to complete 10 minutes of daily meditation via the Headspace app for 8 weeks or waitlist control. Sleep assessments were taken for 4 consecutive days at baseline, and then for 4-day bursts at 2, 5, and 8 weeks after randomization. Sleep quality and subjective sleep duration were assessed each morning with a sleep diary, tiredness was assessed throughout the day using ecological momentary assessment, and objective sleep duration and efficiency were measured using a Fitbit Charge 2.

Results: Both subjective and objective sleep outcomes improved. For subjective sleep outcomes, multilevel modeling revealed that those in the Headspace condition, compared to the control group, reported better sleep quality at sessions 2 (β=0.48, SE=0.12; *P*<.001), 5 (β=0.91, SE=0.13; *P*<.001), and 8 (β=0.69, SE=0.15; *P*<.001) compared to baseline, and a decrease in tiredness at session 5 (β=−0.58, SE=0.19; *P*=.001) compared to baseline, but not at sessions 2 or 8. For objective sleep outcomes, those in the Headspace condition compared to the control group had longer sleep durations at session 5 (β=23.96, SE=12.19; *P*=.04) compared to baseline, but not at sessions 2 or 8. There were no significant effects for sleep efficiency.

Conclusions: This study continues adding to the ever-developing field of mobile health apps by demonstrating that Headspace can positively impact sleep quality, tiredness, and duration.

Trial Registration: ClinicalTrials.gov NCT03652168; https://clinicaltrials.gov/study/NCT03652168

(J Med Internet Res 2026;28:e56287) doi:10.2196/56287

KEYWORDS

Headspace; sleep quality; tiredness; sleep duration; sleep efficiency; ecological momentary assessment; app-based; sleep; subjective; objective; outcome; meditation; randomized controlled trial; RCT; application; self-guided; female; women; United States; mHealth; mobile health; Fitbit; wearable

Introduction

Overview

Training in mindfulness meditation has shown the ability to improve sleep outcomes, such as sleep quality and efficiency [1,2]. Most of this research has tested meditation as traditional in-person mindfulness meditation practices, including mindfulness-based stress reduction and mindfulness-based therapy for insomnia [3,4]. There has been a shift in recent years toward using mobile health (mHealth) apps, such as Headspace (Headspace, Inc) and Calm (Calm.com, Inc), to learn meditation and mindfulness. Therefore, because of their increased popularity, these apps are critical to study as users are seeking these apps to improve their health, including sleep [5]. In addition, they have the potential to be more accessible and available than in-person practices. However, these benefits may come with a trade-off, as these apps are generally self-guided and may not be as effective in teaching meditation or mindfulness practices. App designers often focus on usability and acceptability rather than effectiveness in improving health outcomes [6]. Thus, it is necessary to rigorously examine whether mHealth apps for meditation can effectively improve sleep outcomes. Furthermore, there is a need to research when these positive sleep effects emerge, enabling the ability to track the efficacy of interventions during engagement rather than waiting until a postintervention evaluation period.

XSL•FO
RenderX

Sleep Is a Complex Construct

Sleep is a complex, multidimensional construct that can be measured subjectively and objectively. A well-supported approach to studying sleep identifies 5 dimensions: sleep quality, sleep duration, sleep continuity, sleep timing, and sleepiness [7]. Given that each dimension of sleep can change independently of the others depending on the intervention, it is advisable to measure multiple dimensions. Subjective sleep is self-reported and evaluated using retrospective surveys such as the Pittsburgh Sleep Quality Index and sleep diaries [7]. Objective sleep measurements are assessed using behavioral and physiological technology, and thus are generally outside the participant's control regarding the data collected [7]. Both objective and subjective measures tap into different parts of the sleep experience. For example, several studies have shown modest to null correlations between objective and subjective assessments of the same dimension [8-10]. Therefore, collecting both can provide a more complete picture of how sleep is affected by engaging in mindfulness meditation.

mHealth Apps and Sleep

Research examining the impact of mHealth apps on meditation is in its early stages, with only a few randomized controlled trials testing the efficacy of these apps. Most of this work has been tested using the Calm app. For example, in a sample of adults with sleep disturbance, researchers found that Calm users had reduced self-reported daytime fatigue, sleepiness, and presleep arousal compared to a control group [11]. In a study among employees of a large retail company, researchers found that participants using the Calm app reported decreased daytime sleepiness compared to a control group [12]. Cross-sectional studies indicate that individuals who use Calm report longer and better sleep than those who do not use the app [13].

Yet, there is a dearth of knowledge about whether the Headspace app would also effectively improve sleep outcomes. Headspace is vital to study on its own, given its widespread use. For example, in September 2023 alone, Headspace was downloaded 400,000 times and generated $4 million in revenue [14]. Yet, while Calm is pitched as a multimodal app aimed at improving sleep using techniques that include meditation, Headspace operates from a different perspective. Namely, it was developed to provide step-by-step guidance on the basic principles of mindfulness meditation practice. In this way, Headspace serves as a mHealth corollary to the in-person Mindfulness-Based Stress Reduction training program. Indeed, research has demonstrated that the beneficial changes that occur directly from learning mindfulness meditation, such as decentering [15], acceptance [16], and nonreactivity to inner experience [17,18], are present after using Headspace [19]. As such, it is plausible that improvements in sleep may also ensue after using Headspace. Some initial evidence from quasi-experimental research has shown that children with attention-deficit hyperactivity disorder who used Headspace had reduced self-reported sleep disturbance after a 4-week intervention compared to baseline [20]. However, there is a need to rigorously investigate the impact of Headspace on various measures of sleep to further inform prospective users of its effectiveness.

Additional gaps exist in the current literature on mHealth meditation apps and sleep. Most research has focused on how mHealth meditation apps affect individuals with diagnosable sleep disorders. Yet, the usage rates of mHealth apps indicate that they are being used by the general population, who are likely not meeting clinical levels of sleep disturbance. Yet, poor sleep outcomes are still regularly reported among this population daily. For example, according to the 2020 Behavioral Risk Factor Surveillance System, 33.2% of adults report getting less than 7 hours of sleep per 24 hours [21]. Second, there is a need to test the effectiveness of Headspace using subjective and objective sleep measures to reduce the potential for expectancy effects to influence self-reports, which could account for significant effects. To this end, testing the effects of Headspace in daily life at several time points during the intervention can provide a richer picture of when mHealth apps affect sleep outcomes.

The Present Study

This study tested whether the mHealth app Headspace affects several dimensions of sleep, including self-reported quality via a morning diary and daily tiredness via ecological momentary assessment (EMA), and objectively assessed duration and efficiency via a Fitbit Charge 2 wearable. We hypothesize that the use of a mHealth app will have a positive impact on both sleep quality and sleep duration. We also hypothesize that sleep quality and duration improvements will be recognized early on, possibly by the week 2 assessment. Finally, we do not have a clear hypothesis on how the improvements will last, as there is very little research on how sleep improvements change over time.

Methods

Participants

In this randomized controlled trial, participants were staff employees at a university in the San Joaquin Valley of California. Participants were excluded if they were not university employees or if they were university faculty, younger than 18 years old, not fluent in English, did not have access to a smartphone, or had prior experience in meditation, defined as having participated in mindfulness meditation 2 times per week for 10 minutes over the previous 3 months. To determine the sample size, we conducted a power analysis to compare 2 groups, assuming a moderate effect size, an alpha of .05, and a power of 0.80. The moderate effect size was used, given prior research suggesting such effects could be observed with this kind of mHealth app [22]. The power analysis results indicated that the required sample size to detect the effect was 128 participants. We aimed to enroll 140 participants to account for potential dropouts.

Procedure

Participants were recruited through posted flyers on campus, presentations at departmental and university staff assembly meetings, and word-of-mouth referrals. Interested participants were directed to a secure website, where they read more about the study and were directed to a screener delivered through Qualtrics.

XSL•FO
RenderX

Participants attended an in-lab orientation lasting about 60 minutes. This orientation introduced participants to the assessment procedure. First, they downloaded the RealLife Exp app (Life Data Corporation) on their smartphone and loaded the survey. Participants were guided through a start-up session that demonstrated how to navigate the app and view examples of each question they would encounter during the study. Participants practiced how to answer and were able to ask questions about the process. After practicing, participants were instructed that when a survey was ready, it would appear as a push notification.

To assess subjective sleep quality, participants completed a daily diary upon waking. The survey was available for completion by 6:00 AM and remained open for 4 hours. Participants completed an EMA protocol throughout the day to assess tiredness. EMA surveys were completed randomly within each block: 8:00-10:00 AM, 10:30 AM-12:30 PM, 1:00-3:00 PM, 3:30-5:30 PM, and 6:00-8:00 PM. If participants did not immediately respond to a survey, they had up to 1 hour to complete it, with a reminder notification 20 minutes after the initial prompt. Surveys took between 3 and 4 minutes to complete on average.

Four consecutive days of data were collected via daily diary (quality) and EMA (tiredness). Each orientation took place on a Monday, Tuesday, or Wednesday morning. All data were then scheduled to begin data collection on Wednesday to ensure relative commonality across participants and to ensure both working (Wednesday through Friday) and nonworking (Saturday) days. Participants completed the 4 days of data collection (daily diary and EMA) as part of the baseline session and then at 2, 5, and 8 weeks after the first day of data collection at baseline.

Until this point, the researcher and participant were masked as to their condition. After randomization, both the researcher and the participant were unmasked for the duration of the study. Participants needed to be unmasked as they were instructed to engage with the Headspace app or not. Although researchers were also unmasked, they did not interact with participants outside of standardized messages sent during the study. Per instructions described during the in-lab orientation, participants were only randomized into the Headspace or control condition after the initial 4 days of data collection. Thus, weeks 2, 5, and 8 of data are postrandomization. Allocation to condition occurred at a 2:1 ratio, with more people in the Headspace condition, to account for the potential higher dropout rate in the intervention group as observed in mHealth studies [23]. A random number generator in Microsoft Excel was used to develop the order in which participants would be randomly allocated to a condition. To mimic the experience participants would normally encounter when downloading Headspace, minimal training was provided to participants by research staff. Participants in the Headspace condition were sent an email with download instructions for the app and a code to enter that would grant 12 months of access to Headspace. Headspace provided all the codes but had no say over data collection procedures, analyses, or dissemination. Participants were instructed to use Headspace for 10 minutes daily for the 8-week intervention period. The instructions specified that participants should complete each of the 3 Basic packs (with 10 units each) and then complete the Stress pack (with 30 units). To ensure compliance, we tracked downloads and initial use of the app, ensuring that all participants completed their first session within the first week postrandomization.

For sleep quality assessed via diary, at week 0, there were 462 assessments (mean 4.90, SD 0.56), 418 assessments at week 2 (mean 4.93, SD 0.44), 330 assessments at week 5 (mean 4.85, SD 0.76), and 237 assessments at week 8 (mean 4.85, SD 0.74). For tiredness assessed via EMA, at week 0, there were 2556 assessments (mean 18.66, SD 2.82), 2256 assessments at week 2 (mean 18.05, SD 3.83), 1832 assessments at week 5 (mean 18.87, SD 3.40), and 1326 assessments at week 8 (mean 18.46, SD 4.45). For sleep duration and efficiency assessed via Fitbit at week 0, there were 681 assessments (mean 4.80, SD 0.83), 487 assessments at week 2 (mean 4.87, SD 0.63), 315 assessments at week 5 (mean 4.50, SD 1.32), and 159 assessments at week 8 (mean 4.68, SD 1.17).

Participants in the control condition were given a 1-year subscription to Headspace after completing the 4-month waitlist period. During these 4 months on the waitlist, they were asked not to participate in any mindfulness activities such as yoga or meditation during this time. They completed the same daily diary and EMA protocol as participants in the Headspace condition.

Measures

Baseline Measures

During the baseline assessment, participants first completed demographic information, including their gender (coded for analysis as 0=male, 1=female), age (in years), and ethnicity (coded for analysis as 0=non-Hispanic and Latino, 1=Hispanic and Latino), as well as other measures not relevant to this study.

Daily Diary and EMA Measures

Subjective sleep quality and tiredness were measured using EMA, based on methods from previous research that used EMA to assess physical well-being [24] and developed a standard daily sleep diary [25]. To assess subjective sleep quality, each morning, participants were asked, “How well did you sleep?,” on a scale from 0 (not at all well) to 10 (extremely well). To assess subjective tiredness, participants were asked, “Right now, how tired do you feel?”, rated from 0 (not at all) to 10 (extremely).

Fitbit

A Fitbit Charge 2 wearable was used to estimate sleep duration and efficiency. Fitbit is a popular fitness tracker with a microelectronic triaxial accelerometer to capture body motion in 3-dimensional space. These motion data are then analyzed using proprietary algorithms to identify motion patterns. The Fitbit Charge 2 also monitors heart rate activity through a patented photoplethysmography technology called PurePulse. PurePulse uses light-emitting diodes on the skin-facing surface to continuously estimate heart rate by monitoring blood volume changes [26]. Fitbit estimates steps, calories burned, and sleep through the estimated heart rate. Participants received training on the proper use of the Fitbit and the Fitbit app (Google LLC).

XSL•FO
RenderX

To standardize this, participants were asked to wear the device continuously (including sleep) throughout the study. In addition, they were instructed to sync and charge it on the same days, Tuesday and Sunday, and to ensure the device was placed back on their wrist by the evening of each of these days. Fitabase (Small Steps Labs LLC) was used to extract and process the Fitbit data for the study. Fitabase measures duration as the total number of minutes asleep. Fitabase estimates efficiency by dividing the total time in bed by the total number of minutes asleep. To ensure data were valid, any sleep duration that was less than 2 hours or more than 12 hours was deleted from analyses (although patterns are similar when either or both criteria are not applied), as well as any efficiency scores greater than 100.

Headspace

The intervention was delivered through the commercially available mindfulness meditation app Headspace, widely used in previous intervention studies [27,28]. Headspace provides a variety of formal guided and unguided meditation practices, with instructions delivered through short, animated training videos. The intervention group was instructed to start meditating using the Basic pack. This pack is designed as an introduction to mindfulness meditation and can be used as an opportunity for participants to get familiar with the Headspace teaching style. Each lesson in the Basic pack is approximately 11 minutes and is led by a teacher of the user's choosing. In the first part of the session, the teacher gives the user a short tutorial on the concept of mindfulness. Then, the teacher proceeds with a guided meditation. There are 3 total Basic packs, and each one has 10 sessions, with each session lasting approximately 11 minutes. If participants follow the directions of the study, they will finish all the Basic packs within 30 days. Once participants completed the Basic, they were instructed to move on to the Stress pack, which lasted for 30 sessions. Like the Basic pack, the Stress pack used a teacher and combined visualization and body scanning to help users learn to accept their emotions and pay close attention to the present moment.

Participants were encouraged to complete the meditation quietly for each session without distraction. They were instructed to use the app to meditate for 10 minutes a day for 8 weeks. This duration was chosen based on prior work, which showed that just 10 days of practicing mindfulness for 10 minutes a day successfully reduced stress, negative affect, and improved well-being among a range of sample types [29-32].

Analytic Plan

We used multilevel modeling to account for the 3-level structure of our data, with multiple daily diary and EMA observations nested within assessment sessions (0, 2, 5, or 8) nested within participants. Models were tested using the PROC MIXED command in SAS v.9.4 (SAS Institute Inc), which tests a linear mixed model analysis. We used restricted maximum likelihood estimates to handle missing data, which is the recommended approach given its robustness in addressing missing data that is often nested (eg, a participant who dropped out in session 8 will have missing data for all diary and EMA observations) [33]. This procedure does not impute missing data but uses available data to calculate maximum likelihood estimates.

Analyses followed an intention-to-treat approach, assuming those randomized to the condition completed the intervention material as instructed [34]. As such, we did not factor in specific measures of usage of Headspace in analyses (an issue we return to in the "Discussion"). To test our main research question, we included the following variables as predictors: session, condition (0=control group, 1=Headspace group), and the interaction of session by condition. In the presented results, the terms labeled session refer to the effects of the control condition at that session relative to baseline. The interaction term of session by condition refers to the effects of comparing the Headspace to the control group at that session relative to baseline. Sleep was the outcome variable; each dimension of sleep was tested in a separate model. The session was modeled as a categorical variable to interpret whether the sleep data from sessions 2, 5, and 8 were statistically different than the session 0 data.

Finally, at the between-person level, models controlled for gender and age, and at the within-person level, models controlled for the day of the week as either a nonweekend (0) or weekend (1) day.

Ethical Considerations

The local Institutional Review Board approved all the study procedures and measures (IRB #UCM2018). This study was also registered on clinicaltrials.gov (NCT03652168). There were no deviations from the preregistered protocol, and the analyses presented are those for the secondary set of outcomes proposed. Data from the primary set of outcomes testing the effect of Headspace on mechanisms of mindfulness have previously been reported [19]. The participant completed informed consent at the initial intake session. Participants were reminded at all sessions that they could refrain from answering any questions they wished or could opt out of study procedures without penalty. All participants were given a study code to allow deidentification in the datasets. As compensation, participants received a 1-year subscription to Headspace. For each weekly survey (weeks 0, 2, 5, and 8), participants received US $15. In addition, participants could receive up to a US $20 bonus for a high completion rate (ie, over 80% of surveys completed) across the study.

Results

Table 1 provides descriptive statistics for demographics and the sleep variables. The final sample (N=135) was between the ages of 21 and 65 years old and self-identified as primarily White or Hispanic, and female. Participants in the control condition (mean 40.88, SD 10.84, t_{137}= 2.02; *P*=.05) were slightly older than participants in the Headspace condition (mean 36.91, SD 10.84, t_{137}=2.02; *P*=.05). There were similar proportions of participants identified as Hispanic and female across the 2 conditions, χ^2=2.29; *P*=.13, and χ^2=0.28; *P*=.60, respectively. On average, participants reported sleeping moderately well each morning and feeling somewhat tired. They were recorded as sleeping about 5 and a half hours a night, and their sleep efficiency was below normal.

XSL•FO
RenderX

Table . Descriptive statistics for sleep outcomes.

Variable	Headspace group	Control group	Overall
Age (years), mean (SD)	36.91 (10.84)	41.12 (10.71)	38.19 (10.94)
Ethnicity, n (%)			
American Indian or Alaska Native	1 (1.04)	2 (4.76)	3 (2.17)
Asian	7 (7.29)	4 (9.52)	11 (7.97)
Black	3 (3.13)	1 (2.38)	4 (2.90)
Hispanic	30 (31.25)	8 (19.05)	38 (27.54)
Pacific Islander	2 (2.08)	1 (2.28)	3 (2.17)
White	57 (59.38)	26 (61.90)	83 (60.14)
Gender, n (%)			
Women	74 (77.08)	30 (71.43)	104 (75.36)
Men	22 (22.92)	12 (28.57)	34 (24.64)
Sleep quality, mean (SD)	5.87 (1.40)	5.31 (1.29)	5.68 (1.40)
Tiredness, mean (SD)	4.08 (1.65)	4.51 (1.44)	4.19 (1.56)
Total minutes asleep, mean (SD)	329.54 (63.70)	312.28 (44.34)	322.78 (71.52)
Efficiency, mean (SD)	74.74 (9.33)	75.03 (8.30)	73.93 (10.32)
Total time in bed, mean (SD)	449.36 (50.51)	417.13 (57.22)	447.70 (70.22)

Figure 1 provides a diagram of the process used to determine the final participant number for the study. An initial 291 employees were screened, with 271 eligible. Of the 271 eligible participants, 186 people provided informed consent. Before the study began, 43 participants dropped out, mainly citing time demands. One hundred forty-three participants were randomized, but 8 additional participants were dropped for not having any data. Randomization was completed using a random number generator via Excel to create the order in which participants would be enrolled and allocated to a condition.

At week 0, our sample consisted of 135 participants, with 93 randomized into the Headspace group and 42 into the waitlist control group using a planned 2:1 allocation strategy. At week 2, 87 (93.5% of the original Headspace group) and 40 (95.2% of the original control group) participants completed at least one daily diary and EMA survey. At week 5, 68 (71.5% of the original Headspace group) and 35 (83.3% of the original control group) participants completed at least one daily diary and EMA survey. At week 8, 49 (54.3% of the original Headspace group) and 26 (61.9% of the original control group) participants completed at least one daily diary and EMA survey.

We tested whether participants in the Headspace condition experienced improved subjective and objective sleep outcomes compared to those in the control condition, examining when these effects emerged over the 8-week intervention period.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram of participant flow.

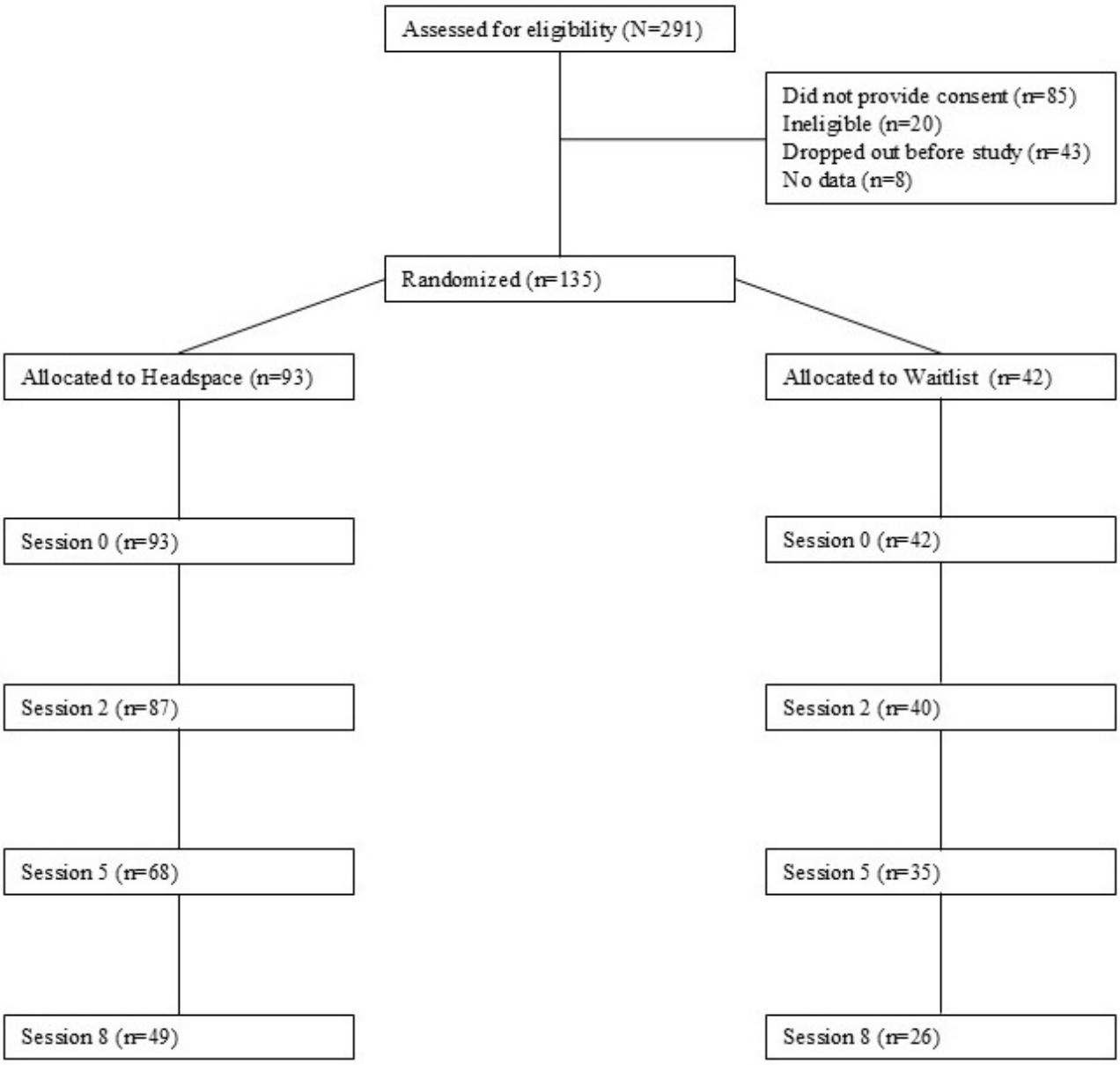

We tested quality and tiredness for the subjective sleep outcomes with results reported in Table 2. The effects for the control group over time are represented as the session terms, whereas the effects for the Headspace condition compared to the control group over time are represented as session x Headspace terms. For quality, compared to the baseline session, those in the control condition reported worse sleep at session 5 (P<.001), with no differences at sessions 2 (P=.41) or 8 (P=.99). There were significant interaction effects of condition by session. Those in the Headspace condition, compared to the control group, reported better sleep at sessions 2 (P<.001), 5 (P<.001), and 8 (P<.001), with each session compared to baseline. For tiredness, those in the control condition reported, compared to the baseline session, an increase in tiredness at session 5 (P=.04), but not at sessions 2 (P=.85) or 8 (P=.79). There was a significant interaction effect of condition by session for session 5. Compared to the control group, those in the Headspace condition reported a decrease in tiredness at session 5 (P=.001), but no difference at sessions 2 (P=.97) or 8 (P=.97), with each session compared to baseline.

XSL•FO
RenderX

Table . Unstandardized beta estimates (SEs) for subjective sleep outcomes with Headspace, session, and interaction as predictors.

Subjective sleep outcomes	Subjective quality		Subjective tiredness	
	β (SE)	*P* value	β (SE)	*P* value
Random effects				
Intercept	1.63[a] (0.21)[a]	<.001	2.29[a] (0.30[a])	<.001
Residual	3.33[a] (0.06)[a]	<.001	4.41[a] (0.09[a])	<.001
Fixed effects				
Intercept	5.08[a] (0.51[a])	<.001	4.76[a] (0.61[a])	<.001
Female	−0.40 (0.26)	.12	−0.16 (0.31)	.61
Age	0.01 (0.01)	.18	−0.003 (0.01)	.79
Weekend	0.35[a] (0.05[a])	<.001	−0.32[a] (0.07[a])	<.001
Session 2	−0.08 (0.10)	.41	−0.03 (0.13)	.85
Session 5	−0.41[a] (0.11[a])	<.001	0.30[a] (0.15[a])	.04
Session 8	−0.002 (0.12)	.99	−0.04 (0.16)	.79
Headspace	0.24 (0.25)	.34	−0.34 (0.31)	.27
Session 2x Headspace	0.48[a] (0.12[a])	<.001	0.01 (0.16)	.97
Session 5x Headspace	0.91[a] (0.13[a])	<.001	−0.58[a] (0.18[a])	.001
Session 8x Headspace	0.69[a] (0.15[a])	<.001	0.01 (0.19)	.97
Model effects				
Pseudo r^2	0.034	—[b]	0.017	—

[a]Coefficients significant at *P*<.05.

[b]Not applicable.

For the objective sleep outcomes, we tested duration and efficiency with results reported in Table 3. For duration, compared to the baseline session, those in the control condition had shorter sleep durations at session 5 (*P*=.03) compared to baseline, but not at sessions 2 (*P*=.94) or 8 (*P*=.20). There was a significant interaction effect of condition by session for session 5. Compared to the control group, those in the Headspace condition had longer sleep durations at session 5 (*P*=.04), but no difference at sessions 2 (*P*=.59) or 8 (*P*=.88), with each session compared to baseline. For sleep efficiency, compared to the baseline session, those in the control condition had a higher efficiency at sessions 2 (*P*=.04) and 5 (*P*=.002) compared to baseline, but not at session 8 (*P*=.33). There were no significant interaction effects of condition by session for sessions 2 (*P*=.37), 5 (*P*=.36), and 8 (*P*=.11).

XSL•FO
RenderX

Table . Unstandardized beta estimates (SEs) for objective sleep outcomes with Headspace, session, and interaction as predictors.

Objective sleep outcomes	Objective duration		Objective efficiency	
	β (SE)	*P* value	β (SE)	*P* value
Random effects				
Intercept	2404.70[a] (571.52[a])	<.001	4.19[a] (0.99[a])	<.001
Residual	4011.98[a] (151.83[a])	<.001	9.95[a] (0.41[a])	<.001
Fixed effects				
Intercept	353.93[a] (32.26[a])	<.001	89.70[a] (1.39[a])	<.001
Female	29.27 (15.04)	.05	0.80 (0.65)	.21
Age	0.57 (0.67)	.40	−0.04 (0.03)	.12
Weekend	11.20[a] (4.12[a])	.007	−0.37 (0.23)	.11
Session 2	−0.59 (8.16)	.94	0.98[a] (0.46[a])	.04
Session 5	−23.27[a] (10.55[a])	.03	1.77[a] (0.58[a])	.002
Session 8	−14.37 (11.21)	.20	0.59 (0.60)	.33
Headspace	23.32 (16.81)	.17	−0.69 (0.74)	.35
Session 2x Headspace	−5.26 (9.68)	.59	−0.49 (0.54)	.37
Session 5x Headspace	23.96[a] (12.19[a])	.04	−0.61 (0.66)	.36
Session 8x Headspace	2.12 (14.58)	.88	1.26 (0.79)	.11
Model effects				
Pseudo r^2	0.058	—[b]	0.050	—

[a]Coefficients significant at $P<.05$.

[b]Not applicable.

Discussion

Principal Findings

This study used daily diaries, EMA, and Fitbit wearables to investigate the impact of Headspace on several dimensions of sleep. Our findings indicate that Headspace had the most significant effect on sleep quality, with some impact on tiredness and sleep duration, and no significant effects on sleep efficiency. Notably, improvements in subjective sleep (quality and tiredness) and objective sleep outcomes (duration) did not occur simultaneously but rather at different times during the study. These findings suggest the continued viability of using mHealth meditation apps to improve sleep and the likelihood that subjective and objective benefits may not happen in tandem. Furthermore, they are consistent with previous research on the Calm app [11,13] that illustrates sleep is multidimensional, with some dimensions improving while others fail to reach significance.

Sleep Quality

The most consistent finding of this study was the rapid improvement in sleep quality. Participants reported better sleep quality within 2 weeks of using Headspace. This finding is impactful because there is emerging evidence that sleep quality is considered a superior sleep index to sleep quantity for assessing sleep [35]. Getting enough sleep is crucial for psychological and physical health, as there are negative consequences when sleep duration decreases past a certain amount of time [36]. What is interesting about sleep quality is that several studies have demonstrated that even if a person gets a whole night's sleep, they can still report their sleep quality as poor and say they were unsatisfied with the night's sleep [37-39]. These findings suggest that objective sleep measures do not fully account for subjective sleep quality. Studies have shown that sleep quality is a good indicator of psychological and overall health [40-42]. This also demonstrates that a person's perceived efficacy of an intervention (which we suggest is as essential a dimension as objective measures) is a significant component that impacts adherence to the intervention [43]. As a future research direction, testing how the perceived efficacy of Headspace impacts adherence is of specific interest.

The improvements in sleep quality continued throughout the study. This promising finding shows that Headspace can alter a person's thinking outside of novel effects. This may not seem profound, but if one considers that it is not uncommon for a person to claim an intervention changed them, then to only renounce that claim quickly thereafter. This is an example of the novelty effect, a well-known phenomenon further discussed below [44-46]. The fact that Headspace users reported continued improvements in sleep quality for 8 weeks indicates that this finding is more than likely not because of novelty and that Headspace does alter a person's quality of sleep positively. This study did not investigate possible mechanisms; however, it has been suggested that meditation may improve sleep quality by

XSL•FO
RenderX

reducing sleep-interfering cognitive processes [47], altering sleep architecture [48], changing perceptions of stress [49], and morphometric and connectivity alterations in sleep-related brain regions [50,51]. Taken together, these findings raise many unanswered questions, fueling future investigation.

Intervention Length

A surprising finding is the convergence of improvements that peaked at week 5 in both subjective measurements, sleep quality and tiredness, and in one of the objective measurements, duration. This finding is novel in suggesting that there may be a previously unidentified point in time at which effectiveness peaks. It has been shown that 8 weeks, but not 4 weeks, of meditation is needed to see benefits in sleep quality and mood [52]. Equally, in non–sleep-related studies, 8 weeks of meditation increased brain structure and function [53] and mental health in university students [54], while 12 weeks of meditation improved depression, anxiety, and stress [55]. However, these studies measure 1 time point or compare 2 time points and cannot accurately pinpoint when effects emerge.

Although the 5-week time point may seem brief, research has shown the potential of mindfulness to have an acute positive effect even after a single use [56]. By allowing users to engage in meditation precisely when needed, such as before bed on a stressful day, and to use it frequently throughout the day in small but frequent doses (thus allowing microdose training), mHealth apps may accelerate the accumulation process. Indeed, other work in this sample found changes to the mindfulness mechanisms of acceptance, attention, and nonreactivity in as little as 2 weeks [19]. More broadly, these findings show the benefit of this study design to test for effectiveness during the engagement with Headspace and not simply wait for a postintervention evaluation period. This approach could be expanded in future studies to determine, with greater detail, the acute and longer-term effects of Headspace in each dimension of sleep. In the long term, this additional information will hopefully replace the “do ten minutes a day of Headspace” with a more nuanced approach.

It was unexpected that Headspace’s effects appeared to wane after 5 weeks. As suggested by other research, it was assumed that sleep benefits would continue throughout the 8-week intervention [51]. A possible explanation for this waning could be that the improvement was a placebo effect. A recent study showed that a sham meditation could increase state mindfulness, mindful observing, state decentering, and mindful nonreacting, similar to an actual meditation condition [57]. Yet, simply chalking up findings to a placebo effect seems unsatisfying. The continuous measurement of sleep daily, within and across days, would suggest an incredibly robust and omnipresent effect. Moreover, it is unclear why the placebo would wear off after 5 weeks. Nevertheless, future research should continue to probe how expected benefits from meditation may prime certain short-term perceived benefits.

There is another, broader hypothesis that may better explain the post-5-week waning: the well-known phenomenon that user engagement with mHealth apps decreases over time. For example, among users measured, 53% uninstall the app within 30 days [58]. Although determining the exact reasons for waning or discontinuing is challenging, research has shown that the main reasons are a loss of interest or declining motivation, a lack of desired features, and the app not being fun [59,60]. Ultimately, the novelty effect significantly influences user abandonment; users may be drawn to the new app out of curiosity but lose interest once it wears off [44-46]. Finally, attrition is a common phenomenon in longitudinal studies, with the literature suggesting that rates from 30% to 70% are not uncommon [61,62].

Of concern is that this pattern of waning, discontinued use, or attrition was observed despite the initial robust effects. The ability of these apps to run autonomously allows for their incredible reach [63]. However, without continued user involvement and self-management, behavioral changes and health improvements tend to dissipate [64]. Therefore, there is a need to continue developing features and behavior change techniques that sustain usage levels for long-term maintenance. In a systematic review of sleep apps, feedback and monitoring were used most often as a behavior change technique [65]. Yet, a recent meta-analysis identified 7 main areas that impact user engagement, with personalization—the ability to make technology act in a particular way depending on user preferences—being the element that users cited as most influencing their engagement [66]. Thus, there may be a disconnect between what app developers pursue as features to implement and what techniques may benefit long-term outcomes. App designers often focus on usability and acceptability rather than effectiveness in improving health outcomes [6]. Due to the increasing shift toward the use of mHealth apps and the vast number of apps available, user engagement is likely to become more critical as app developers strive to attract users. This will further the field of behavior change techniques and is a possible direction for future research.

Limitations and Future Directions

A limitation of this study is that we were unable to track Headspace usage accurately and reliably. Therefore, we have yet to determine whether people in the Headspace group were meditating and are only relying on random assignments to the experimental condition to drive expected changes (ie, intention-to-treat analyses). In this way, we may be underestimating the effect of Headspace, as there may be nonusers in the Headspace group. In addition, tracking engagement would have allowed us to identify a type of person who was more likely to continue completing the intervention and to tailor the intervention content better to these individuals. Research from clinical psychology has shown that younger age, lower education, poorer social problem-solving, lower levels of persistence, and greater avoidance coping are factors of noncompletion [67]. Although not all these factors apply to the current intervention, it demonstrates distinct factors when examining noncompletion and why it is essential to track engagement.

As with most mHealth experimental research, attrition was a limitation, which may have led to the selection of certain traits related to study completion and sleep behaviors. For example, a recent meta-analysis found that the pooled attrition rate in young adults across 15 mHealth trials was 26% [67], and another

XSL•FO
RenderX

found a pooled attrition rate of 43% [22]. This study's attrition rate aligns with the attrition rates identified in these meta-analyses. Therefore, resolving or limiting attrition remains an unresolved issue, and addressing this would greatly benefit research aimed at determining interventions to reduce attrition.

Conclusion

This study demonstrated that Headspace can improve sleep, showing that sleep dimensions do not all change with the same amount of meditation. Some sleep dimensions improved quickly, while others took longer to change. There was a significant convergence of benefits at the 5-week time point and increased attrition. These convergences uncovered 2 critical pieces of information: that there might be a lower-than-expected threshold of meditation needed to experience sleep benefits and that future research using mHealth apps must include factors that improve adherence and engagement. The current study showed that mHealth apps, specifically Headspace, are viable for those seeking to enhance sleep with minimal life disruption. The results from this study can be used by future scientists wanting to learn how to dose mHealth meditation interventions better and practitioners looking to use mHealth meditation to improve patient sleep outcomes.

Acknowledgments

Portions of this work were presented at the American Psychosomatic Society conference in March 2023 [68].

Funding

Funding for data collection was partially supported by a seed grant from the Healthy Campus Network at the University of California, San Francisco (PI: JZ). ZAT's time was supported by the National Heart, Lung, and Blood Institute under the award number 3R15HL161681-01S1 (PI: JZ).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: ZAT, LG, MJZ
Data curation: LG, MJZ
Formal analysis: ZAT, MJZ
Funding acquisition: ZAT, MJZ
Investigation: ZAT, LG, MJZ
Methodology: LG, MJZ
Project administration: LG, MJZ
Resources: MJZ
Supervision: MJZ
Validation: MJZ
Visualization: MJZ
Writing – original draft: ZAT, LG, AP, MJZ
Writing – review & editing: ZAT, AP, MJZ

Conflicts of Interest

LG is now employed at Service Titan. No conflicts of interest were declared for the remaining authors.

Checklist 1
CONSORT checklist.
[DOCX File, 33 KB - jmir_v28i1e56287_app1.docx]

References

1. Kanen J, Nazir R, Sedky K, Pradhan B. The effects of mindfulness-based interventions on sleep disturbance: a meta-analysis. . 2015(2) p. 105-115. [doi: 10.2174/2210676605666150311222928]
2. Chen TL, Chang SC, Hsieh HF, Huang CY, Chuang JH, Wang HH. Effects of mindfulness-based stress reduction on sleep quality and mental health for insomnia patients: a meta-analysis. J Psychosom Res 2020 Aug;135:110144. [doi: 10.1016/j.jpsychores.2020.110144] [Medline: 32590218]
3. Ong JC, Manber R, Segal Z, Xia Y, Shapiro S, Wyatt JK. A randomized controlled trial of mindfulness meditation for chronic insomnia. Sleep 2014 Sep 1;37(9):1553-1563. [doi: 10.5665/sleep.4010] [Medline: 25142566]

XSL•FO
RenderX

4. Rusch HL, Rosario M, Levison LM, et al. The effect of mindfulness meditation on sleep quality: a systematic review and meta-analysis of randomized controlled trials. Ann N Y Acad Sci 2019 Jun;1445(1):5-16. [doi: 10.1111/nyas.13996] [Medline: 30575050]
5. Lee E, Han S, Jo SH. Consumer choice of on-demand mHealth app services: context and contents values using structural equation modeling. Int J Med Inform 2017 Jan;97:229-238. [doi: 10.1016/j.ijmedinf.2016.10.016] [Medline: 27919381]
6. Chib A, Lin SH. Theoretical advancements in mHealth: a systematic review of mobile apps. J Health Commun 2018;23(10-11):909-955. [doi: 10.1080/10810730.2018.1544676] [Medline: 30449261]
7. Buysse DJ. Sleep health: can we define it? Does it matter? Sleep 2014 Jan 1;37(1):9-17. [doi: 10.5665/sleep.3298] [Medline: 24470692]
8. Argyropoulos SV, Hicks JA, Nash JR, et al. Correlation of subjective and objective sleep measurements at different stages of the treatment of depression. Psychiatry Res 2003 Sep 30;120(2):179-190. [doi: 10.1016/s0165-1781(03)00187-2] [Medline: 14527649]
9. Armitage R, Trivedi M, Hoffmann R, Rush AJ. Relationship between objective and subjective sleep measures in depressed patients and healthy controls. Depress Anxiety 1997;5(2):97-102. [doi: 10.1002/(sici)1520-6394(1997)5:2<97::aid-da6>3.0.co;2-2] [Medline: 9262940]
10. Rotenberg VS, Indursky P, Kayumov L, Sirota P, Melamed Y. The relationship between subjective sleep estimation and objective sleep variables in depressed patients. Int J Psychophysiol 2000 Sep;37(3):291-297. [doi: 10.1016/s0167-8760(00)00110-0] [Medline: 10858574]
11. Huberty JL, Green J, Puzia ME, et al. Testing a mindfulness meditation mobile app for the treatment of sleep-related symptoms in adults with sleep disturbance: a randomized controlled trial. PLoS One 2021;16(1):e0244717. [doi: 10.1371/journal.pone.0244717] [Medline: 33411779]
12. Huberty JL, Espel-Huynh HM, Neher TL, Puzia ME. Testing the pragmatic effectiveness of a consumer-based mindfulness mobile app in the workplace: randomized controlled trial. JMIR Mhealth Uhealth 2022 Sep 28;10(9):e38903. [doi: 10.2196/38903] [Medline: 36169991]
13. Huberty J, Puzia ME, Larkey L, Vranceanu AM, Irwin MR. Can a meditation app help my sleep? A cross-sectional survey of Calm users. PLoS One 2021;16(10):e0257518. [doi: 10.1371/journal.pone.0257518] [Medline: 34679078]
14. Headspace revenue & app download estimates from sensor tower - Apple app store. Headspace Inc. URL: https://app.sensortower.com/ios/publisher/publisher/384434796 [accessed 2024-10-03]
15. Shapiro SL, Carlson LE, Astin JA, Freedman B. Mechanisms of mindfulness. J Clin Psychol 2006 Mar;62(3):373-386. [doi: 10.1002/jclp.20237] [Medline: 16385481]
16. Brown KW, Ryan RM. The benefits of being present: mindfulness and its role in psychological well-being. J Pers Soc Psychol 2003 Apr;84(4):822-848. [doi: 10.1037/0022-3514.84.4.822] [Medline: 12703651]
17. Baer RA, Carmody J, Hunsinger M. Weekly change in mindfulness and perceived stress in a mindfulness-based stress reduction program. J Clin Psychol 2012 Jul;68(7):755-765. [doi: 10.1002/jclp.21865] [Medline: 22623334]
18. Bishop SR. What do we really know about mindfulness-based stress reduction? Psychosom Med 2002;64(1):71-83. [doi: 10.1097/00006842-200201000-00010] [Medline: 11818588]
19. Gavrilova L, Zawadzki MJ. Examining how headspace impacts mindfulness mechanisms over an 8-week app-based mindfulness intervention. Mindfulness (N Y) 2023 Sep;14(9):2236-2249. [doi: 10.1007/s12671-023-02214-4]
20. Fried R, DiSalvo M, Farrell A, Biederman J. Using a digital meditation application to mitigate anxiety and sleep problems in children with ADHD. J Atten Disord 2022 May;26(7):1033-1039. [doi: 10.1177/10870547211025616] [Medline: 34865550]
21. Pankowska MM, Lu H, Wheaton AG, et al. Prevalence and geographic patterns of self-reported short sleep duration among US adults, 2020. Prev Chronic Dis 2023 Jun 29;20:E53. [doi: 10.5888/pcd20.220400] [Medline: 37384831]
22. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567. [doi: 10.2196/14567] [Medline: 31573916]
23. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020 Sep 29;22(9):e20283. [doi: 10.2196/20283] [Medline: 32990635]
24. Zawadzki MJ, Hussain M, Kho C. Comparing multidimensional facets of stress with social, emotional, and physical well-being using ecological momentary assessment among a Hispanic sample. Stress Health 2022 Apr;38(2):375-387. [doi: 10.1002/smi.3098] [Medline: 34494721]
25. Carney CE, Buysse DJ, Ancoli-Israel S, et al. The consensus sleep diary: standardizing prospective sleep self-monitoring. Sleep 2012 Feb 1;35(2):287-302. [doi: 10.5665/sleep.1642] [Medline: 22294820]
26. Fitbit charge: wireless activity wristband product manual. Fitbit Inc.: N.p. Fitbit, Inc. n.d.
27. Bostock S, Crosswell AD, Prather AA, Steptoe A. Mindfulness on-the-go: effects of a mindfulness meditation app on work stress and well-being. J Occup Health Psychol 2019 Feb;24(1):127-138. [doi: 10.1037/ocp0000118] [Medline: 29723001]
28. Champion L, Economides M, Chandler C, ed. The efficacy of a brief app-based mindfulness intervention on psychosocial outcomes in healthy adults: a pilot randomised controlled trial. PLoS One 2018;13(12):e0209482. [doi: 10.1371/journal.pone.0209482] [Medline: 30596696]

XSL•FO
RenderX

29. Economides M, Martman J, Bell MJ, Sanderson B. Improvements in stress, affect, and irritability following brief use of a mindfulness-based smartphone app: a randomized controlled trial. Mindfulness (N Y) 2018;9(5):1584-1593. [doi: 10.1007/s12671-018-0905-4] [Medline: 30294390]
30. Foley T, Lanzillotta-Rangeley J. Stress reduction through mindfulness meditation in student registered nurse anesthetists. AANA J 2021 Aug;89(4):284-289. [Medline: 34342565]
31. Zollars I, Poirier TI, Pailden J. Effects of mindfulness meditation on mindfulness, mental well-being, and perceived stress. Curr Pharm Teach Learn 2019 Oct;11(10):1022-1028. [doi: 10.1016/j.cptl.2019.06.005] [Medline: 31685171]
32. Schwartz JE, Stone AA. Strategies for analyzing ecological momentary assessment data. Health Psychol 1998;17(1):6-16. [doi: 10.1037/0278-6133.17.1.6]
33. Altman DG. Missing outcomes in randomized trials: addressing the dilemma. Open Med 2009 May 12;3(2):e51-e53. [Medline: 19946393]
34. Kohyama J. Factors affecting the quality of sleep in children. Children (Basel) 2021 Jun 12;8(6):499. [doi: 10.3390/children8060499] [Medline: 34204755]
35. Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem: National Academies Press; 2006:11617. [doi: 10.17226/11617]
36. Faerman A, Kaplan KA, Zeitzer JM. Subjective sleep quality is poorly associated with actigraphy and heart rate measures in community-dwelling older men. Sleep Med 2020 Sep;73:154-161. [doi: 10.1016/j.sleep.2020.04.012] [Medline: 32836083]
37. Kaplan KA, Hardas PP, Redline S, Zeitzer JM, Sleep Heart Health Study Research Group. Correlates of sleep quality in midlife and beyond: a machine learning analysis. Sleep Med 2017 Jun;34:162-167. [doi: 10.1016/j.sleep.2017.03.004] [Medline: 28522086]
38. Kaplan KA, Hirshman J, Hernandez B, et al. When a gold standard isn't so golden: lack of prediction of subjective sleep quality from sleep polysomnography. Biol Psychol 2017 Feb;123:37-46. [doi: 10.1016/j.biopsycho.2016.11.010] [Medline: 27889439]
39. Kohyama J. Which is more important for health: sleep quantity or sleep quality? Children (Basel) 2021 Jun 24;8(7):542. [doi: 10.3390/children8070542] [Medline: 34202755]
40. Lao XQ, Liu X, Deng HB, et al. Sleep quality, sleep duration, and the risk of coronary heart disease: a prospective cohort study with 60,586 adults. J Clin Sleep Med 2018 Jan 15;14(1):109-117. [doi: 10.5664/jcsm.6894] [Medline: 29198294]
41. Moore PJ, Adler NE, Williams DR, Jackson JS. Socioeconomic status and health: the role of sleep. Psychosom Med 2002;64(2):337-344. [doi: 10.1097/00006842-200203000-00018] [Medline: 11914451]
42. Al-Hashmi I, Hodge F, Nandy K, Thomas E, Brecht ML. The effect of a self-efficacy-enhancing intervention on perceived self-efficacy and actual adherence to healthy behaviours among women with gestational diabetes mellitus. Sultan Qaboos Univ Med J 2019;18(4):e513-e319. [doi: 10.18295/squmj.2018.18.04.014]
43. Bai S, Hew KF, Huang B. Does gamification improve student learning outcome? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts. Educational Research Review 2020 Jun;30:100322. [doi: 10.1016/j.edurev.2020.100322]
44. Sailer M, Sailer M. Gamification of in‐class activities in flipped classroom lectures. Brit J Educational Tech 2021 Jan;52(1):75-90. [doi: 10.1111/bjet.12948]
45. Schmidt-Kraepelin M, Thiebes S, Stepanovic S, Mettler T, Sunyaev A. Gamification in health behavior change support systems - a synthesis of unintended side effects. Presented at: 52nd Hawaii International Conference on System Sciences (HICSS 2019); Jan 8-11, 2019.
46. Winbush NY, Gross CR, Kreitzer MJ. The effects of mindfulness-based stress reduction on sleep disturbance: a systematic review. Explore (NY) 2007;3(6):585-591. [doi: 10.1016/j.explore.2007.08.003] [Medline: 18005910]
47. Nagendra RP, Maruthai N, Kutty BM. Meditation and its regulatory role on sleep. Front Neurol 2012;3:54. [doi: 10.3389/fneur.2012.00054] [Medline: 22529834]
48. Omidi A, Zargar F. Effects of mindfulness-based stress reduction on perceived stress and psychological health in patients with tension headache. J Res Med Sci 2015 Nov;20(11):1058-1063. [doi: 10.4103/1735-1995.172816] [Medline: 26941809]
49. Fox KCR, Nijeboer S, Dixon ML, et al. Is meditation associated with altered brain structure? A systematic review and meta-analysis of morphometric neuroimaging in meditation practitioners. Neurosci Biobehav Rev 2014 Jun;43:48-73. [doi: 10.1016/j.neubiorev.2014.03.016] [Medline: 24705269]
50. Hasenkamp W, Barsalou LW. Effects of meditation experience on functional connectivity of distributed brain networks. Front Hum Neurosci 2012;6:38. [doi: 10.3389/fnhum.2012.00038] [Medline: 22403536]
51. Basso JC, McHale A, Ende V, Oberlin DJ, Suzuki WA. Brief, daily meditation enhances attention, memory, mood, and emotional regulation in non-experienced meditators. Behav Brain Res 2019 Jan 1;356:208-220. [doi: 10.1016/j.bbr.2018.08.023] [Medline: 30153464]
52. Kral TRA, Schuyler BS, Mumford JA, Rosenkranz MA, Lutz A, Davidson RJ. Impact of short- and long-term mindfulness meditation training on amygdala reactivity to emotional stimuli. Neuroimage 2018 Nov 1;181:301-313. [doi: 10.1016/j.neuroimage.2018.07.013] [Medline: 29990584]
53. Cassar L, Fischer M, Valero V. Keep calm and carry on: the short- vs. long-run effects of mindfulness meditation on (academic) performance. 2022.

XSL•FO
RenderX

54. Nikkhah Ravari O, Mousavi SZ, Babak A. Evaluation of the effects of 12 weeks mindfulness-based stress reduction on glycemic control and mental health indices in women with diabetes mellitus type 2. Adv Biomed Res 2020;9(1):61. [doi: 10.4103/abr.abr_133_20] [Medline: 33457344]
55. Johnson KT, Merritt MM, Zawadzki MJ, Di Paolo MR, Ayazi M. Cardiovascular and affective responses to speech and anger: proactive benefits of a single brief session of mindfulness meditation. J Appl Biobehavioral Res 2019 Sep;24(3) [FREE Full text] [doi: 10.1111/jabr.12167]
56. Davies JN, Colagiuri B, Sharpe L, Day MA. Placebo effects contribute to brief online mindfulness interventions for chronic pain: results from an online randomized sham-controlled trial. Pain 2023 Oct 1;164(10):2273-2284. [doi: 10.1097/j.pain.0000000000002928] [Medline: 37310492]
57. The uninstall threat: 2020 app uninstall benchmarks. AppsFlyer. URL: https://www.appsflyer.com/infograms/2019-app-uninstall-benchmarks/ [accessed 2023-12-13]
58. Attig C, Franke T. I track, therefore I walk – exploring the motivational costs of wearing activity trackers in actual users. Int J Hum Comput Stud 2019 Jul;127:211-224. [doi: 10.1016/j.ijhcs.2018.04.007]
59. Krebs P, Duncan DT. Health app use among US mobile phone owners: a national survey. JMIR Mhealth Uhealth 2015 Nov 4;3(4):e101. [doi: 10.2196/mhealth.4924] [Medline: 26537656]
60. Miller RB, Wright DW. Detecting and correcting attrition bias in longitudinal family research. J Marriage Fam 1995 Nov;57(4):921. [doi: 10.2307/353412]
61. Tambs K, Rønning T, Prescott CA, et al. The Norwegian Institute of Public Health twin study of mental health: examining recruitment and attrition bias. Twin Res Hum Genet 2009 Apr;12(2):158-168. [doi: 10.1375/twin.12.2.158] [Medline: 19335186]
62. Ben-Zeev D, Scherer EA, Gottlieb JD, et al. mHealth for schizophrenia: patient engagement with a mobile phone intervention following hospital discharge. JMIR Ment Health 2016 Jul 27;3(3):e34. [doi: 10.2196/mental.6348] [Medline: 27465803]
63. Bakker D, Rickard N. Engagement in mobile phone app for self-monitoring of emotional wellbeing predicts changes in mental health: MoodPrism. J Affect Disord 2018 Feb;227:432-442. [doi: 10.1016/j.jad.2017.11.016] [Medline: 29154165]
64. Arroyo AC, Zawadzki MJ. The implementation of behavior change techniques in mHealth apps for sleep: systematic review. JMIR Mhealth Uhealth 2022 Apr 4;10(4):e33527. [doi: 10.2196/33527] [Medline: 35377327]
65. Wei Y, Zheng P, Deng H, Wang X, Li X, Fu H. Design features for improving mobile health intervention user engagement: systematic review and thematic analysis. J Med Internet Res 2020;22(12):e21687. [doi: 10.2196/21687] [Medline: 33295292]
66. McMurran M, Huband N, Overton E. Non-completion of personality disorder treatments: a systematic review of correlates, consequences, and interventions. Clin Psychol Rev 2010 Apr;30(3):277-287. [doi: 10.1016/j.cpr.2009.12.002] [Medline: 20047783]
67. Prior E, Dorstyn D, Taylor A, Rose A. Attrition in psychological mHealth interventions for young people: a meta-analysis. J technol behav sci 2023;9(4):639-651. [doi: 10.1007/s41347-023-00362-x]
68. Torok Z, Gavrilova L, Patel A, Zawadzki M. Does headspace improve sleep quality in daily life: a daily diary study of app-based mindfulness meditation and sleep. Presented at: American Psychosomatic Society 80th Annual Scientific Meeting Challenging the Future: Towards a Better Biopsychosocial Health; Mar 8-11, 2023 URL: https://thesbsm.org/wp-content/uploads/2023/03/APS-2023-Abstract-Book.pdf [accessed 2026-02-02]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
EMA: ecological momentary assessment
mHealth: mobile health

Edited by TDA Cardoso; submitted 11.Jan.2024; peer-reviewed by D Wilson, SAA Massar, T Nakamura; revised version received 02.Jul.2025; accepted 07.Jul.2025; published 04.Feb.2026.

Please cite as:
Torok ZA, Gavrilova L, Patel A, Zawadzki MJ
The Effectiveness of the Headspace App for Improving Sleep: Randomized Controlled Trial
J Med Internet Res 2026;28:e56287
URL: https://www.jmir.org/2026/1/e56287
doi:10.2196/56287

© Zoltan Andre Torok, Larisa Gavrilova, Amish Patel, Matthew Jason Zawadzki. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 4.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and

XSL•FO
RenderX

reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Associations Between Smartphone-Based Finger Tapping and Cognitive Performance in Older Adults: Observational Study

Huitong Ding[1,2], PhD; Taylor A Orwig[3], MS; Jian Rong[4], PhD; Heaven Y Tatere[3], MD; Chunyu Liu[5], PhD; Eric Schramm[6], PharmD; Chathurangi H Pathiravasan[7], PhD; Joanne M Murabito[2,8*], MD, ScM; Honghuang Lin[3*], PhD

[1]Department of Anatomy and Neurobiology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, United States

[2]The Framingham Heart Study, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, United States

[3]Department of Medicine, University of Massachusetts Chan Medical School, 55 Lake Avenue North, Worcester, MA, United States

[4]Department of Neurology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, United States

[5]Department of Biostatistics, Boston University School of Public Health, Boston, MA, United States

[6]Care Evolution, Ann Arbor, MI, United States

[7]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

[8]Department of Medicine, Section of General Internal Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, United States

[*]these authors contributed equally

Corresponding Author:
Honghuang Lin, PhD
Department of Medicine, University of Massachusetts Chan Medical School, 55 Lake Avenue North, Worcester, MA, United States

Abstract

Background: Finger tapping tasks assess fine motor control and have been proposed as potential markers of cognitive function. With smartphones widely available, these tasks can be easily administered at home or in other nonclinical settings. However, the relationship between smartphone-based finger tapping measures and cognitive performance is still not well understood.

Objective: This study aimed to examine the association between smartphone digital finger tapping features and cognitive performance in an aging population.

Methods: Participants were enrolled in the study as part of the electronic Framingham Heart Study. They were instructed to perform a 2-finger tapping task every 2 months over a 1-year period. In each session, they tapped with 2 fingers of the left hand for 10 seconds and then with 2 fingers of the right hand for 10 seconds, for a total of 20 seconds. Global cognition performance and 4 cognitive domains, including memory, executive function, language, and visuospatial function, were assessed using a standardized neuropsychological battery. Fifteen tapping features were extracted to capture aspects of motor performance, including the mean, SD, skewness, and slope of the intertap interval (ITI). Associations between tapping-derived features and cognitive performance were assessed using linear regression models, adjusting for age, sex, handedness, education, cohort, and the time interval between cognitive and tapping assessments.

Results: A total of 302 participants (mean age 74.7, SD 6.3 years; n=169, 56% female participants, and n=40, 13.2% non-White participants) completed the digital finger tapping tasks. Eleven tapping features such as basic temporal properties (eg, number of taps and mean ITI), temporal variability (eg, SD and coefficient of variation of ITI, ITI range, and Microfluctuation Index), and fatigue/temporal drift (ITI slope) were significantly associated with global cognitive function (all $P<.001$). Each 1 SD increase in the number of taps was associated with a 0.14-unit higher global cognitive function score ($\beta=.14$, 95% CI 0.07-0.21). Furthermore, finger tapping features were significantly associated with multiple cognitive domains, with the greatest number of associations observed for executive function (11 significant features), including a strong association with mean ITI ($\beta=-0.27$, 95% CI −0.33 to −0.20). Stratified analyses by hand showed consistent effect directions across both hands. The aggregated composite scores derived from finger tapping features demonstrated significant associations with their respective cognitive test scores, with partial correlation coefficients ranging from 0.21 for memory function to 0.41 for global cognitive function.

Conclusions: Digital finger tapping features, particularly those reflecting ITI variability, are significantly associated with cognitive performance. These findings suggest that finger tapping tasks may serve as noninvasive, scalable tools for assessing cognitive performance. Further research is warranted to validate their use for monitoring cognitive health in aging populations.

(J Med Internet Res 2026;28:e82463) doi:10.2196/82463

XSL•FO
RenderX

KEYWORDS

finger tapping test; smartphone; digital features; cognitive performance; cohort

Introduction

Motor performance has emerged as a promising indicator of cognitive aging. Growing evidence suggests its potential utility as an early marker of cognitive decline [1]. While gait-related metrics have been extensively explored and consistently associated with impairments in memory, executive function, and visuospatial function [2,3], there is increasing recognition that motor biomarkers should also include upper-limb function. Recent studies have suggested that abnormalities in upper-limb motor function, such as slower movement speed and irregular rhythm, may reflect early neurodegenerative changes and have been associated with mild cognitive impairment (MCI) and dementia [4-6].

Among upper-limb assessments, the finger tapping test (FTT) is a simple and widely used measure of psychomotor speed, motor coordination, and central nervous system integrity [4,7]. Early versions of the FTT used a telegraph key device [8]. Later adaptations moved to devices such as the computer mouse [9] and keyboard [10]. With the widespread availability of smartphones, the FTT can now be administered on touchscreen interfaces [11,12]. This approach is low-cost and scalable. And it allows fine motor function to be assessed outside traditional clinical settings. Smartphones are also well-suited for population-level digital phenotyping because they can support at-home assessments that are more convenient and accessible for diverse and underserved groups. In addition, digital platforms capture high-resolution temporal data. This makes it possible to derive a wide range of tapping features that may help identify subtle motor changes related to early neurodegenerative processes.

Neuropsychological testing remains the gold standard for assessing domain-specific cognitive performance including memory and executive function [13]. However, high-resolution digital motor measures may provide complementary information about brain-behavior relationships. Examining how fine motor measures derived from widely accessible technologies relate to cognitive performance could improve the way we characterize cognitive aging and support the development of scalable digital biomarkers. Linking accessible, real-world digital tools with rigorous neuropsychological assessment is an important step toward more precise evaluation of cognitive function in older adults and may ultimately contribute to earlier identification of subtle changes.

In this study, we examined the relationship between smartphone-based finger tapping performance and multiple cognitive domains in participants from the Framingham Heart Study (FHS), a well-characterized multigenerational cohort. We hypothesized that poorer tapping performance, as captured by digital tapping-derived metrics, would be associated with lower cognitive performance. By linking these digital motor measures with standard neuropsychological tests, our findings may help support the use of smartphone-based tapping tasks as scalable tools for characterizing cognitive function in aging.

Methods

Study Population

This study included Offspring and Omni 1 cohort participants from the FHS, a longitudinal and multigenerational cohort study initiated in 1948 [14-16]. Offspring participants were enrolled from 1971 to 1975 (n=5124), and the multiethnic Omni 1 participants were enrolled from 1994 to 1998 (n=506) to reflect the increasing diversity of the city of Framingham, MA. Both cohorts were examined in the FHS research center every 4 to 8 years using the same protocols. Eligible participants were invited to enroll in the electronic FHS (eFHS) at the time of the last research exam (Offspring examination 10 and Omni 1 examination 5, 2019 - 2022). English-speaking participants who owned a smartphone (iPhone with iOS 10 or higher or Android version 7 or higher) were invited to download the MyDataHelps app, which included a number of surveys and tasks. A research assistant assisted participants with app download in the research center and provided written instructions if participants preferred to download the app at home. Further details of the eFHS protocol have been previously described [17]. A total of 620 participants enrolled in the eFHS study. Of these, 9 participants who enrolled using a computer and 133 participants who enrolled using an Android device were excluded, as the 2-finger tapping test was only available on the iPhone platform. Among the remaining participants, 445 completed at least one 2-finger tapping test. For the present analysis, we further restricted the sample to individuals who had completed at least one 2-finger tapping assessment and at least one neuropsychological assessment within a 5-year window. This additional criterion yielded a final analytic sample of 302 participants. Participants provided electronic consent within the eFHS mobile app and gave written consent for the parent FHS at each in-person exam visit and neuropsychological testing session. Figure 1 presents the graphical abstract of this study.

Cognitive and clinical characteristics were assessed using standardized procedures. Cognitive status was adjudicated by a review committee that included at least 1 neurologist and 1 neuropsychologist [18]. Dementia was diagnosed according to the criteria outlined in the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition [19]. MCI was defined as cognitive decline in one or more domains without functional impairment and not meeting criteria for dementia [20]. Details of the FHS diagnostic procedures have been described in a prior study [21]. Parkinson disease (PD) status was determined using 3 sources: (1) self-report to the question “have you been told by a doctor you have PD?” (2) examiner-administered health history interviews; and (3) responses to the question “for what medical conditions do you use cannabis?” where PD was listed as an option. Self-rated health was measured using a question from the self-administered Short Form-12 Health Survey. In this survey, participants rated their overall health on a 4-point scale ranging from “poor” to “excellent” [22].

XSL•FO
RenderX

Figure 1. Graphical abstract.

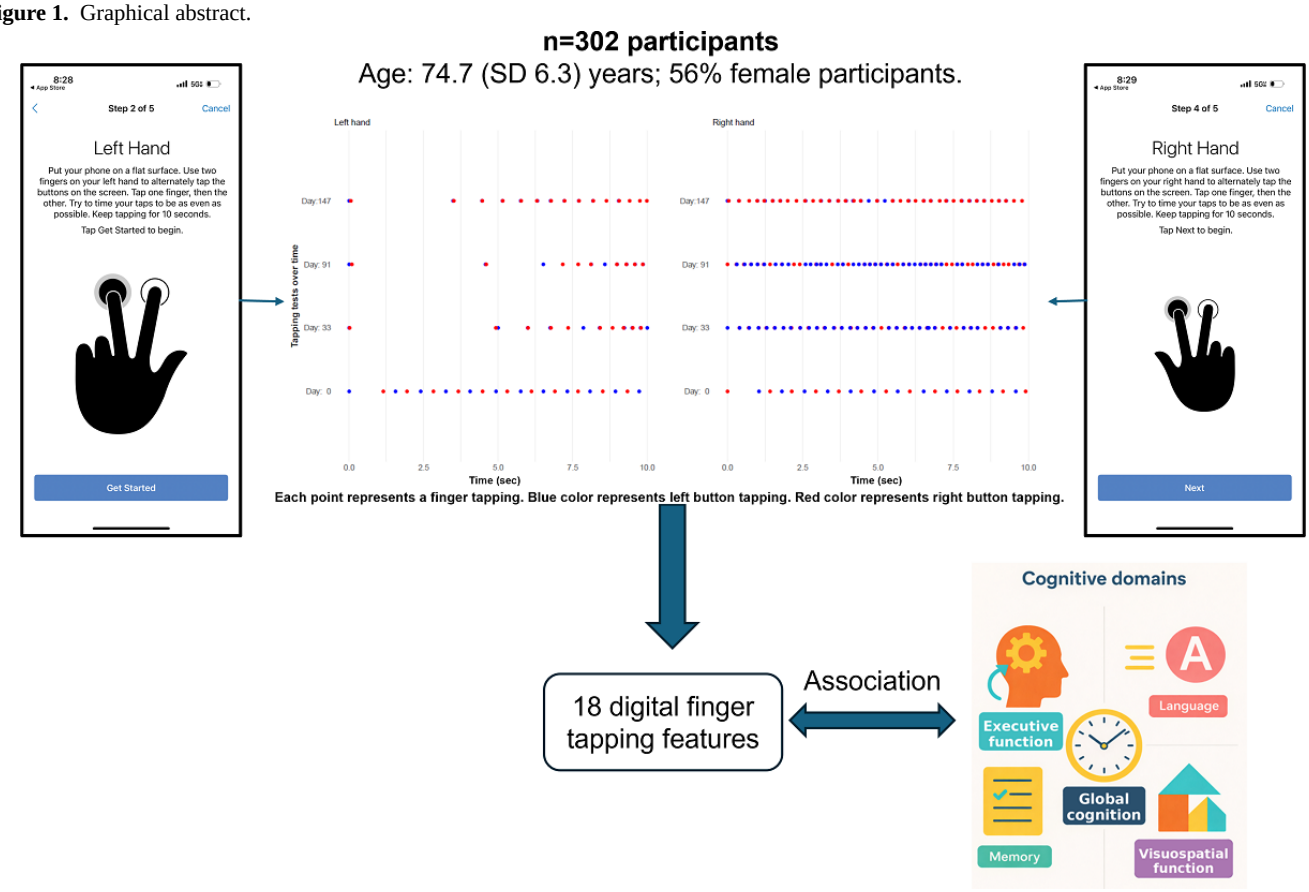

Ethical Considerations

This research received ethical approval from the institutional review board of Boston University Medical Center (H-32132). Participants provided electronic consent within the eFHS mobile app and gave written consent for the parent FHS at each in-person exam visit and neuropsychological testing session. The original informed consent and institutional review board approval allowed secondary analysis of the data without requiring additional participant consent. All data analyzed in this study were deidentified. No financial compensation was provided to participants.

Two-Finger Tapping Test and Feature Extraction

As part of the eFHS Offspring and Omni 1 protocol, participants were sent the 2-finger tapping test at enrollment and then every 2 months over the course of 1 year. The 2-finger tap task informed participants that the task measured tapping speed and instructed participants to put their phone on a flat surface to use 2 fingers to alternately tap the buttons on the screen while trying to time the taps to be as even as possible (Figure S1 in Multimedia Appendix 1, screenshot of task). Each test session lasted 20 seconds, during which participants alternated tapping on designated areas of the smartphone screen with 2 fingers of the left hand for 10 seconds, followed by 2 fingers of the right hand for 10 seconds, as quickly and consistently as possible. Data were included if the app recorded a valid 10-second trial with at least 1 tap event. We did not set additional performance-based exclusion thresholds, such as a minimum number of taps. Our goal was to capture the full range of tapping behavior, including slow, irregular, or instruction-nonadherent tapping, because these patterns may reflect cognitive function. A few examples of finger tapping results are shown in Figure S2 in Multimedia Appendix 1. A comprehensive set of features was extracted from the intertap interval (ITI) data to capture various aspects of motor performance. (1) Basic temporal features included the total number of taps per session as an indicator of effort, endurance, and overall speed, the mean ITI as measures of central tendency, and the SD of ITI as a measure of temporal variability. We also calculated the coefficient of variation of ITI to account for differences in overall tapping speed. (2) Asymmetry features included the alternating tap ratio (the proportion of left-right alternating taps), and the mean and SD of ITI for each finger. Skewness and kurtosis of the ITI distribution were also calculated to capture irregular tapping patterns. (3) Consistency metrics included ITI range. This feature reflected the stability of performance. (4) Fatigue and temporal drift were characterized using the ITI slope (the change in ITI over time during the trial) and the last-to-first ITI ratio. These features indicated whether tapping slowed or changed over the course of the session. (5) Microfluctuation feature included the Microfluctuation Index. This index was defined as the SD of successive ITI differences to capture subtle motor instability.

Neuropsychological Assessment

The methodology for administering neuropsychological tests in FHS has been previously described [23]. Cognitive testing was first introduced in 1976, through brief cognitive screenings and expanded to a more comprehensive neuropsychological battery beginning in 1999 [13,24]. In the present study, we used 21 commonly administered neuropsychological tests to evaluate

XSL•FO
RenderX

cognitive function across 4 domains (Table S1 in Multimedia Appendix 1) [25]. To address missing values of neuropsychological tests, we used multiple imputation by chained equations [26]. Distributions of the Hooper Visual Organization Test, Trail Making Test part A, and Trail Making Test part B were normalized using natural logarithmic transformation. To aid interpretation, the transformed scores were directionally adjusted so that higher values reflected better cognitive performance. All neuropsychological test scores were subsequently standardized to *z*-scores (mean 0, SD 1). These neuropsychological test scores were then summarized into 4 cognitive domains by averaging *z*-scores of neuropsychological tests within each domain: memory, executive function, language, and visuospatial function. A global cognitive score was further derived based on the mean values of all 4 cognitive domain scores.

Statistical Analyses

Linear regression models were used to assess the associations between finger tapping performance and cognitive domain scores. For each participant, the mean value of each finger tapping feature across all available trials was calculated and used as the independent variable. Global cognitive function served as the primary dependent variable. All finger tapping features were standardized to *z*-scores to enable comparability across metrics. Models were adjusted for age at the time of the tapping assessment, sex, education, cohort (Offspring or Omni 1), handedness, and the time interval (in years) between the neuropsychological assessment and the baseline tapping test. As secondary analyses, we further explored associations with 4 specific cognitive domains: memory, executive function, language, and visuospatial function.

To further characterize hand-specific relationships, we performed stratified analyses for right- and left-hand tapping data. Sensitivity analyses were also conducted within the subset of participants identified as right-handed to reduce heterogeneity due to hand preference. Within this group, we reassessed the associations between each finger tapping metric and cognitive domain scores, additionally stratifying by the hand used during the tapping assessment (right vs left). In addition, we grouped the trials according to whether they were performed with the dominant or nondominant hand and repeated the association analyses within these groups. To explore the impact of noncompliant trials, we conducted a sensitivity analysis excluding trials (3729/4277, 12.8%) with an alternating tap ratio less than 0.80. To evaluate the potential impact of practice effects arising from different numbers of completed sessions, we performed a further sensitivity analysis to examine the associations between finger tapping measures and cognitive domain scores using each participant's first 3 sessions. Bonferroni adjustment was used to correct multiple testing, and a significance threshold was determined with $P<.05/15=.003$, with the total number of finger tapping features. As an exploratory analysis, we examined whether cohort modified the association between finger tapping measures and cognition. In each linear regression model, we included an interaction term between the tapping feature and cohort.

We further used the least absolute shrinkage and selection operator (LASSO) regression to derive a composite score for global cognitive function based on finger tapping features. Specifically, we implemented a 5-fold cross-validation loop. For each training set, an inner 5-fold cross-validation was used to select the optimal regularization parameter (α). The composite score was defined as a weighted linear combination of the tapping features. The weights were the LASSO-derived coefficients within each training fold. The fitted model was then applied to the corresponding held-out test fold to obtain out-of-sample predicted global cognitive scores. Then, we assessed the association between the composite score and global cognitive function using partial correlation analysis adjusting for age, sex, and education. In addition, domain-specific composite tapping scores were constructed for each cognitive domain following the same procedure. All analyses were performed using *R* (version 4.0.3; R Core Team) or Python (version 3.6.8; Python Software Foundation).

Results

Cohort Descriptive

This study included 302 eFHS participants (mean age 74.7, SD 6.3 years; n=169, 56%, were female participants, n=192, 63.6%, had college degree or higher, and n=40, 13.2%, were non-White participants; Table 1). As expected, the majority (247/302, 81.8%) were right-handed, while 11.9% (36/302) were left-handed and 5% (15/302) were ambidextrous. On average, participants had completed 7 (SD 4) FTT sessions with both hands administered. Neurologically, the sample was generally healthy—the prevalence of PD was 0.3% (1/302). The average Mini-Mental State Exam (MMSE) score was 28.8 (SD 1.3), with 2.3% (7/302) of participants meeting criteria for MCI and 0.3% (1/302) for dementia. Self-rated health status was favorable—74.1% (224/302) of participants rated their general health as "very good" or "excellent."

XSL•FO
RenderX

Table . Demographics of the study participants[a].

Variable	Total (N=302)	Global cognitive performance		
		First tertile (n=101)	Second tertile (n=100)	Third tertile (n=101)
Age (y), mean (SD)	74.7 (6.3)	76.0 (7.0)	75.0 (6.0)	73.3 (5.5)
Sex, n (%)				
Females	169 (56)	52 (51.5)	52 (52)	65 (64.4)
Males	133 (44)	49 (48.5)	48 (48)	36 (35.6)
Omni cohort, n (%)	40 (13.2)	27 (26.7)	7 (7)	6 (5.9)
Education, n (%)				
High school	30 (9.9)	15 (14.9)	10 (10)	5 (5)
Some college	80 (26.5)	39 (38.6)	23 (23)	18 (17.8)
College and higher	192 (63.6)	47 (46.5)	67 (67)	78 (77.2)
Handedness, n (%)				
Right	247 (81.8)	81 (80.2)	84 (84)	82 (81.2)
Left	36 (11.9)	14 (13.9)	11 (11)	11 (10.9)
Ambidextrous	15 (5)	5 (5)	4 (4)	6 (5.9)
Unknown/missing	4 (1.3)	1 (1)	1 (1)	2 (2)
Time interval between the finger tapping task and the neuropsychological tests, mean (SD)	1.7 (1.2)	1.7 (1.2)	1.6 (1.1)	1.8 (1.3)
Number of finger tapping test session administered, mean (SD)	7 (4)	6 (4)	7 (4)	8 (3)
MMSE[b], mean (SD)	28.8 (1.3)	28.1 (1.6)	29.0 (1.1)	29.4 (0.8)
Prevalent MCI[c], n (%)	7 (2.3)	7 (6.9)	0	0
Prevalent dementia, n (%)	1 (0.3)	1 (1)	0	0
Prevalent Parkinson disease, n (%)	1 (0.3)	0	1 (1)	0
General health SF[d]-12, n (%)				
Poor	1 (0.3)	1 (1)	0	0
Fair	10 (3.3)	6 (5.9)	2 (2)	2 (2)
Good	67 (22.2)	26 (25.7)	25 (25)	16 (15.8)
Very good	146 (48.3)	45 (44.6)	47 (47)	54 (53.5)
Excellent	78 (25.8)	23 (22.8)	26 (26)	29 (28.7)

[a]Categorical variables are presented as counts (proportions). Continuous variables are presented as mean (SD).

[b]MMSE: Mini-Mental State Exam.

[c]MCI: mild cognitive impairment.

[d]SF: Short Form.

When stratified by tertiles of global cognitive function, participants in the third tertile (n=101) were younger (mean age 73.3 vs 76 years in the first tertile), more likely to be female participants (n=65, 64.4% vs n=52, 51.5%), and more highly educated (n=78, 77.2% with college or higher vs n=47, 46.5% in the first tertile). The proportion of participants from the Omni cohort (representing greater racial/ethnic diversity) was higher in the first tertile (n=27, 26.7%) compared to the highest (n=6, 5.9%). MMSE scores followed the expected gradient, increasing from 28.1 (SD 1.6) in the first tertile to 29.4 (SD 0.8) in the third tertile. Similarly, MCI and dementia cases were observed only in the first tertile group. Participants in the third tertile also completed more FTTs on average (17 vs 13). This suggested possible differences in adherence or engagement.

XSL•FO
RenderX

Participants who completed the digital finger tapping assessment were younger and more highly educated compared to those in the broader Offspring/Omni 1 cohort at Exam 10/5. Additional demographic comparisons with the full eFHS and Offspring/Omni 1 samples are presented in Table S2 in Multimedia Appendix 1. The distributions of finger tapping features and cognitive domain scores, including 25th percentile, median, and 75th percentile values, are shown in Tables S3 and S4**,** respectively, in Multimedia Appendix 1.

Association Between Digital Finger Tapping Features and Cognitive Performance

Among the 15 finger tapping features evaluated, 11 were significantly associated with global cognitive function (*P* values<.001; Table 2). The strongest associations were observed for tapping variability metrics derived from ITI, such as SD, range, and Microfluctuation Index. For example, a 1 SD increase in mean ITI (average time differences between successive taps) was associated with a 0.23 decrease in global cognitive function score (β=−.23; 95% CI −0.30 to −0.17; *P*<.001). In addition to variability-based features, performance-related features such as number of taps (β=.14; 95% CI 0.07-0.21) were positively associated with global cognitive function.

Table . The association between digital finger tapping features and global cognitive function[a].

Group and finger tapping features	Description	β (95% CI)	*P* value
Basic temporal features			
Number of taps	Indicator of effort, endurance, and speed	0.14 (0.07 to 0.21)	<.001
Mean ITI[b]	Average time differences between successive taps	–0.23 (–0.30 to –0.17)	<.001
SD of ITI	Measures tapping variability	–0.22 (–0.28 to –0.15)	<.001
CV[c] of ITI	CV of ITI	–0.14 (–0.20 to –0.07)	<.001
Asymmetry features			
Skewness of ITI	Skewness of ITI	0.00 (–0.07 to 0.06)	.95
Kurtosis of ITI	Kurtosis of ITI	0.03 (–0.04 to 0.09)	.38
Alternating tap ratio	Fraction of alternating left-right taps	0.05 (–0.01 to 0.12)	.11
Mean of left ITI	Mean of tap times for the left finger	–0.22 (–0.29 to –0.16)	<.001
SD of left ITI	SD of tap times for the left finger	–0.18 (–0.25 to –0.11)	<.001
Mean of right ITI	Mean of tap times for the right finger	–0.20 (–0.27 to –0.14)	<.001
SD of right ITI	SD of tap times for the right finger	–0.12 (–0.19 to –0.05)	<.001
Consistency metrics			
ITI range	Difference between the longest and shortest ITI	–0.23 (–0.30 to –0.17)	<.001
Fatigue and temporal drift			
ITI slope	Change in ITI overtime	–0.11 (–0.17 to –0.05)	<.001
Last to first ITI	Ratio of last ITI to first ITI	–0.06 (–0.13 to 0.00)	.05
Microfluctuation feature			
Microfluctuation Index	SD of successive ITI differences	–0.16 (–0.22 to –0.09)	<.001

[a]Significant associations were claimed if *P*<.05/15=.0033. Features without laterality labels were computed using all taps across both buttons within each trial. Features labeled “left” or “right” were calculated separately for taps on the left and right buttons, respectively.

[b]ITI: intertap interval.

[c]CV: coefficient of variation.

Beyond global cognitive function, we further examined the associations between finger tapping features and 4 domain-specific cognitive outcomes: memory, executive function, language, and visuospatial function (Table 3, Table S5 in Multimedia Appendix 1). Similar to findings for global cognition, variability in ITI emerged as a consistent predictor across domains. Executive function demonstrated the strongest associations, with 11 out of 15 finger tapping features reaching significance. The most significant association was observed for mean ITI (β=−0.27; 95% CI −0.33 to −0.20; *P*<.001). Visuospatial function showed fewer associations (6 significant features), yet ITI range again emerged as the strongest predictor (β=−0.26; 95% CI −0.38 to −0.15; *P*<.001). When we applied a false discovery rate correction across all 75 finger

XSL•FO
RenderX

tapping-cognition tests, the overall pattern remained similar. Ten features remained significant for memory, 11 for executive function, 10 for language, 10 for visuospatial function, and 11 for global cognitive function (all with false discovery rate <.05). After excluding trials with an alternating tap ratio less than 0.8, the pattern of the associations between tapping metrics and cognitive domains remained similar (Table S6 in Multimedia Appendix 1). In additional sensitivity analysis, we used the first 3 sessions of each participant. The directions of the associations between finger tapping measures and cognitive domains were largely consistent, but fewer associations were statistically significant (Table S7 in Multimedia Appendix 1). Across all domains, 10 (13.3%) interactions reached nominal significance (Table S8 in Multimedia Appendix 1). For global cognitive function, the most significant interaction was for ITI slope (P=.03). We conducted a sensitivity analysis using complete-case data by excluding participants with missing neuropsychological test scores (Table S9 in Multimedia Appendix 1). Most associations remained significant with the same direction of effects.

Table . The most significant digital finger tapping feature for each cognitive domain[a].

Cognitive domain	Number of significant finger tapping features, n	Most significant association		
		Finger tapping features	β (95% CI)	*P* value
Memory	9	ITI[b] range	–0.19 (–0.26 to –0.12)	<.001
Executive function	11	Mean ITI	–0.27 (–0.33 to –0.20)	<.001
Language	9	Mean of right ITI	–0.24 (–0.32 to –0.16)	<.001
Visuospatial function	6	ITI Range	–0.26 (–0.38 to –0.15)	<.001

[a]Significant associations were claimed if P<.05/15=.003.

[b]ITI: intertap interval.

Given the presence of multiple finger tapping features associated with cognitive performance, we constructed a composite tapping score for each cognitive test domain using LASSO regression. The finger tapping composite scores were significantly correlated with cognitive domain scores (all with P<.001). The strength of these correlations differed by domain. Partial correlation coefficients were 0.21 for memory, 0.22 for executive function, 0.38 for language, 0.23 for visuospatial function, and 0.41 for global cognition.

We further examined the associations between finger tapping features and cognitive function by analyzing FTTs administered separately to the left and right hands (Tables S5, S10, and S11 in Multimedia Appendix 1). Overall, the direction and significance of associations observed in the combined-hand models were largely consistent with those identified in hand-specific analyses, supporting the robustness of the findings across tapping conditions. For global cognitive function, 9 features remained significantly associated in both left-hand and right-hand models. For example, 2 key features of temporal control, including the mean ITI and ITI range, were consistently associated with lower global cognition across all 3 models. Across domain-specific outcomes, effect directions were again broadly consistent. For executive function, the number of significant tapping features was similar across hands (10 in left-hand and 8 in right-hand models), with mean ITI or mean right-hand ITI consistently emerging as the most predictive features. Language and memory domains also demonstrated robust associations, particularly with right-hand tapping metrics such as ITI range and mean ITI. In contrast, associations for visuospatial function were more pronounced in the right-hand model (6 significant features) than the left-hand model (1 significant feature). This asymmetry may reflect functional lateralization of spatial processing or task engagement differences between hands. Additionally, the predominance of right-handed individuals in the sample (247/302, 81.8%) may have contributed to the stronger associations observed in right-hand tapping performance.

To further examine the robustness of associations and minimize potential confounding from handedness, we conducted subgroup analyses restricted to right-handed participants (n=247; Tables S12-S14 in Multimedia Appendix 1). While differences in the number of significant tapping features emerged across cognitive domains, the direction of effects, particularly for temporal variability features, remained consistent. When we reanalyzed the data by dominant versus nondominant hand, the overall pattern of associations was similar (Tables S15 and S16 in Multimedia Appendix 1). For executive function, however, more finger tapping measures were significantly associated with performance when tests were administered with the nondominant hand (n=10) than with the dominant hand (n=8). In dominant-hand trials, mean ITI showed the most significant associations across all cognitive domains. In nondominant hand trials, mean ITI again showed the strongest associations with memory, executive function, and language. However, the ITI range, reflecting temporal variability, was most significantly associated with visuospatial and global cognitive function.

Discussion

Principal Results

In this study of predominantly cognitively intact, community-based older adults, several digital finger tapping features were significantly associated with cognitive performance. These results support the use of motor-based digital measures as a tool for evaluating cognitive health. Notably, variability features such as the SD, coefficient of variation, ITI range, and the Microfluctuation Index were consistently associated with cognitive domains. These results

XSL•FO
RenderX

suggest that temporal features derived from simple tapping tasks may serve as sensitive and scalable indicators of general cognitive functioning.

Executive function emerged as the most consistently associated domain, showing strong relationships with 13 finger tapping features. Given that executive function is a multidimensional construct encompassing key cognitive processes such as processing speed, planning, working memory, and inhibitory control [27], this finding provides important insight into the potential cognitive mechanisms underlying fine motor performance. Although the finger tapping task appears simple, its successful execution in a timed context requires the ability to sustain and update task goals (working memory) [28], initiate and maintain a consistent motor rhythm (planning and motor initiation) [29], and suppress irrelevant motor impulses (inhibitory control) [30]. The strong associations between temporal tapping features, particularly variability-based features, and executive function suggest that digital motor timing tasks may serve as sensitive proxies for detecting early executive dysfunction. Interestingly, language performance was most strongly associated with mean tap times for the right finger of both right and left hands. This suggested that beyond just timing or variability, the dynamics of motor execution may have a specific connection with language processing abilities. This finding is consistent with prior research indicating shared neural substrates for motor and language functions [31]. And it provides behavioral evidence for the functional interdependence between these domains.

Stratified analyses by hand and by handedness revealed generally consistent associations between finger tapping features and cognitive performance. Across both left- and right-hand administered tests, key temporal metrics such as SD of ITI, ITI range, and Microfluctuation Index remained associated with global cognitive function. However, subtle differences emerged in the number and strength of significant associations across hands and cognitive domains. For instance, visuospatial associations were more prominent in right-hand tests, while left-hand results showed fewer significant features overall. Analyses restricted to right-handed participants largely confirmed the main findings, with core tapping features such as mean ITI and ITI range consistently emerging as the informative predictors across domains. These results reinforce the generalizability of tapping-derived digital biomarkers, while also suggesting that motor laterality and hand-specific administration may modestly influence the association. While many tapping features showed associations with cognition, 4 features, including skewness and kurtosis of ITI, alternating tap ratio, and last-to-first ITI ratio, were not significantly related to any cognitive domain. The alternating tap ratio is primarily an index of adherence to the alternating-tap instruction. However, maintaining an alternating pattern relies on attention and executive function. Therefore, variation in this measure may carry cognitively relevant information rather than simply indicating noncompliance to following instructions. For this reason, we included it in our analyses, although we did not observe significant associations with any cognitive domain. These findings highlight the need to prioritize the most informative digital biomarkers when evaluating cognitive health.

Although these associations reached statistical significance, the effect sizes were small. This suggests that finger tapping measures explain only a modest proportion of variation in cognitive performance. From a clinical standpoint, these findings also underscore the growing relevance of digital phenotyping as a complementary approach to traditional cognitive screening [32]. Early signs of executive dysfunction and cognitive decline are often subtle and can be overlooked, especially in high-functioning or highly educated individuals. The associations identified between fine motor irregularities and cognitive performance suggest that brief, app-based tapping tasks may serve as passive indicators of early neurocognitive vulnerability.

This study offers several notable strengths. First, it leverages a comprehensive battery of neuropsychological assessments covering four major cognitive domains, enabling robust evaluation of cognitive function. Second, the repeated administration of digital FTTs at regular intervals provides a more detailed characterization of motor performance over time and its association with cognitive outcomes. While digital finger tapping tasks have been explored in prior studies, they have predominantly focused on targeted clinical populations, such as individuals with PD or other neurodegenerative disorders [10,33,34]. In contrast, our study is among the first to investigate smartphone-based finger tapping performance in a large, community-based sample of older adults. The majority of participants were subjectively healthy and free of diagnosed cognitive impairment. This distinction enhances the generalizability of our findings and supports the broader application of digital motor assessments for population-level cognitive monitoring. The use of a smartphone-based tapping task, coupled with the derivation of multiple temporal and variability-based digital metrics, represents a novel, noninvasive, and scalable approach for assessing fine motor control in real-world settings. Notably, a high proportion of older iPhone users completed and returned the tapping assessments. Many participants completed multiple administrations. Given the well-documented challenges of adherence in digital health studies, this level of engagement suggests that the task was both accessible and acceptable to participants. Additionally, stratification of tapping performance by left and right hand allows for the exploration of potential lateralization effects. This offers further insight into the relationship between motor function and cognitive domains.

Comparison With Prior Work

Our findings align with prior research using the traditional method of assessing finger tapping speed [35]. In the earlier study, finger tapping was featured by instructing participants to repeatedly press a mechanical lever with an attached counter using the index finger of each hand over a fixed 10-second interval. It demonstrated that individuals who later converted to MCI exhibited a faster decline in tapping speed compared to nonconverters. Similarly, in our study, a lower number of taps (slower tapping speed) was significantly associated with lower global and domain-specific cognitive performance, supporting motor slowing as a behavioral marker of cognition.

XSL•FO
RenderX

Limitations

First, this study was cross-sectional and conducted in a sample consisting predominantly of cognitively intact older adults. Although finger tapping features were collected longitudinally, the cognitive assessments were not temporally aligned with each tapping session. Therefore, we summarized finger tapping performance by averaging repeated trials into a single mean value per participant. This approach allowed us to focus on between-person differences in mean tapping performance but did not capture within-person variability or change over time. In addition, the time interval between neuropsychological test and finger tapping (mean 1.7, SD 1.2 y) limited inferences about how well tapping performance reflects real-time cognitive status. As longitudinal follow-up data become available, future work will evaluate whether tapping metrics can predict subsequent cognitive decline or discriminate between cognitively impaired and unimpaired individuals.

Second, the first step of the 2-finger tap app entitled tapping speed informed the participant that "this activity measures your tapping speed." However, the instructions for the left hand and right hand in steps 2 and 4 include the statement "try to make your taps as even as possible" (Figure S1 in Multimedia Appendix 1). These instructions were designed to encourage fast tapping with a stable rhythm. In practice, this may have led some participants to focus more on speed and others more on regularity. In this context, ITI variability may reflect both intrinsic motor control and individual task strategies. Our findings should be interpreted accordingly. Future studies could use separate speed-focused and regularity-focused conditions to better disentangle these components as both may reflect early neurodegenerative processes. Besides, we examined only the alternating 2-finger version of the FTT with a fixed 10-second trial duration. Because of this short duration, the ITI slope may capture rhythm establishment rather than physiological fatigue. It should be interpreted with caution. Future work should compare alternating and single-finger tapping protocols and systematically vary test duration to better separate rhythm establishment from physiological fatigue. This also can help determine the optimal design for digital tapping-based cognitive assessment.

Third, while the smartphone-based tapping task provides a standardized and practical assessment platform, performance may still be influenced by external factors such as participants' familiarity with digital devices or technical variability in smartphone hardware and software. Although this study adjusted for age, sex, education, and handedness, conditions such as arthritis, motor fatigue, medication use, or depression may independently affect tapping speed, rhythm, or consistency. In some cases, motor slowing may reflect physical rather than cognitive decline. Therefore, these factors should be considered when interpreting tapping metrics in clinical practice, especially in people with multiple comorbidities.

Finally, the tapping task was only available on iOS devices. Therefore, 133 Android users were excluded. Because iOS and Android ownership differs by socioeconomic and demographic factors, this restriction may introduce socioeconomic selection bias. In addition, participants in our sample were generally well-educated and reported good to excellent subjective health. These characteristics may limit the generalizability of our findings to more diverse populations with different demographic or clinical profiles. Moreover, cognitive performance was high in this cohort (mean MMSE 28.8, SD 1.3; dementia prevalence 0.3%), suggesting possible ceiling effects in the cognitive measures. Such ceiling effects may attenuate observed associations and underestimate the true relationships that may be more apparent in clinically diverse populations. In addition, the iPhone models and variation in touchscreen sampling rates or input latency across devices may introduce measurement noise into fine-grained variability metrics (eg, SD of ITI and Microfluctuation Index). Further research in larger and more diverse populations with standardized hardware is needed to confirm these findings.

Conclusions

In summary, this study identified several finger tapping features that were significantly associated with performance across multiple cognitive domains and highlighted the potential of digital finger tapping tasks as informative biomarkers for assessing cognitive function. These findings suggest that finger tapping tasks may be used as noninvasive, scalable tools to assess cognitive performance.

Acknowledgments

The authors sincerely thank the Framingham Heart Study participants for their dedication and invaluable contribution, without which this study would not have been possible. They also appreciate the Framingham Heart Study researchers and staff for their continued efforts in conducting participant examinations.

Funding

This work was supported by grants from the National Institute on Aging (R01AG076757 and R03AG095992) and National Heart, Lung, and Blood Institute (R01HL141434). The Framingham Heart Study is funded through the National Heart, Lung, and Blood Institute Contract (N01-HC-25195, HHSN268201500001I, 75N92019D00031, and 75N92025D00012), and the neuropsychological testing data and brain health data is supported by grants R21HL175584 and by grants from the National Institute on Aging (R01AG083735, R01AG080670, and U01AG068221). The funding agencies had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

XSL•FO
RenderX

Data Availability

The data used in this study can be accessed through the Framingham Heart Study upon reasonable request and in compliance with their data sharing guidelines [36].

Authors' Contributions

HD, JMM, and HL conceptualized and designed the study. HD and HL performed the analysis. HD and HL drafted the initial manuscript. TAO, JR, HYT, CL, ES, and CHP critically reviewed the final draft. All authors reviewed and edited the manuscript and have read and agreed to the version for publication.

Conflicts of Interest

ES is an employee of CareEvolution. The other authors have no conflicts of interest to report.

Multimedia Appendix 1
Supplemental tables and figures.
[DOCX File, 650 KB - jmir_v28i1e82463_app1.docx]

References

1. Yaqub A, Darweesh SKL, Dommershuijsen LJ, et al. Risk factors, neuroimaging correlates and prognosis of the motoric cognitive risk syndrome: a population-based comparison with mild cognitive impairment. Eur J Neurol 2022 Jun;29(6):1587-1599. [doi: 10.1111/ene.15281] [Medline: 35147272]
2. Collyer TA, Murray AM, Woods RL, et al. Association of dual decline in cognition and gait speed with risk of dementia in older adults. JAMA Netw Open 2022 May 2;5(5):e2214647. [doi: 10.1001/jamanetworkopen.2022.14647] [Medline: 35639376]
3. Grande G, Rizzuto D, Vetrano DL, et al. Cognitive and physical markers of prodromal dementia: a 12-year-long population study. Alzheimers Dement 2020 Jan;16(1):153-161. [doi: 10.1002/alz.12002] [Medline: 31914224]
4. Rudd KD, Lawler K, Callisaya ML, Alty J. Investigating the associations between upper limb motor function and cognitive impairment: a scoping review. Geroscience 2023 Dec;45(6):3449-3473. [doi: 10.1007/s11357-023-00844-z] [Medline: 37337026]
5. Kueper JK, Speechley M, Lingum NR, Montero-Odasso M. Motor function and incident dementia: a systematic review and meta-analysis. Age Ageing 2017 Sep 1;46(5):729-738. [doi: 10.1093/ageing/afx084] [Medline: 28541374]
6. Miyazaki A, Ito Y, Okuyama T, et al. Association between upper limb movements during drumming and cognition in older adults with cognitive impairment and dementia at a nursing home: a pilot study. Front Rehabil Sci 2023;4:1079781. [doi: 10.3389/fresc.2023.1079781] [Medline: 37347105]
7. Gopal A, Hsu WY, Allen DD, Bove R. Remote assessments of hand function in neurological disorders: systematic review. JMIR Rehabil Assist Technol 2022 Mar 9;9(1):e33157. [doi: 10.2196/33157] [Medline: 35262502]
8. Cousins MS, Corrow C, Finn M, Salamone JD. Temporal measures of human finger tapping: effects of age. Pharmacol Biochem Behav 1998 Feb;59(2):445-449. [doi: 10.1016/s0091-3057(97)00443-7] [Medline: 9476994]
9. Hubel KA, Reed B, Yund EW, Herron TJ, Woods DL. Computerized measures of finger tapping: effects of hand dominance, age, and sex. Percept Mot Skills 2013 Jun;116(3):929-952. [doi: 10.2466/25.29.PMS.116.3.929-952] [Medline: 24175464]
10. Akram N, Li H, Ben-Joseph A, et al. Developing and assessing a new web-based tapping test for measuring distal movement in Parkinson's disease: a Distal Finger Tapping test. Sci Rep 2022 Jan 10;12(1):386. [doi: 10.1038/s41598-021-03563-7] [Medline: 35013372]
11. Bot BM, Suver C, Neto EC, et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. Sci Data 2016 Mar 3;3:160011. [doi: 10.1038/sdata.2016.11] [Medline: 26938265]
12. Omberg L, Chaibub Neto E, Perumal TM, et al. Remote smartphone monitoring of Parkinson's disease and individual response to therapy. Nat Biotechnol 2022 Apr;40(4):480-487. [doi: 10.1038/s41587-021-00974-9] [Medline: 34373643]
13. Ding H, Wang B, Hamel AP, et al. Exploring cognitive progression subtypes in the Framingham Heart Study. Alzheimers Dement (Amst) 2024;16(1):e12574. [doi: 10.1002/dad2.12574] [Medline: 38515438]
14. Feinleib M, Kannel WB, Garrison RJ, McNamara PM, Castelli WP. The Framingham offspring study. Design and preliminary data. Prev Med 1975 Dec;4(4):518-525. [doi: 10.1016/0091-7435(75)90037-7] [Medline: 1208363]
15. Andersson C, Johnson AD, Benjamin EJ, Levy D, Vasan RS. 70-year legacy of the Framingham Heart Study. Nat Rev Cardiol 2019 Nov;16(11):687-698. [doi: 10.1038/s41569-019-0202-5] [Medline: 31065045]
16. Splansky GL, Corey D, Yang Q, et al. The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. Am J Epidemiol 2007 Jun 1;165(11):1328-1335. [doi: 10.1093/aje/kwm021] [Medline: 17372189]

XSL•FO
RenderX

17. Rong J, Pathiravasan CH, Zhang Y, et al. Baseline smartphone app survey return in the electronic Framingham Heart Study Offspring and Omni 1 Study: eCohort study. JMIR Aging 2024 Dec 31;7(1):e64636. [doi: 10.2196/64636] [Medline: 39740111]
18. Seshadri S, Wolf PA. Lifetime risk of stroke and dementia: current concepts, and estimates from the Framingham Study. Lancet Neurol 2007 Dec;6(12):1106-1114. [doi: 10.1016/S1474-4422(07)70291-0] [Medline: 18031707]
19. Guze SB. Diagnostic and Statistical Manual of Mental Disorders, (DSM-IV). Am J Psychiatry 1995 Aug;152(8):1228-1228. [doi: 10.1176/ajp.152.8.1228]
20. Albert MS, DeKosky ST, Dickson D, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011 May;7(3):270-279. [doi: 10.1016/j.jalz.2011.03.008] [Medline: 21514249]
21. Yuan J, Maserejian N, Liu Y, et al. Severity distribution of Alzheimer's disease dementia and mild cognitive impairment in the Framingham Heart Study. J Alzheimers Dis 2021;79(2):807-817. [doi: 10.3233/JAD-200786] [Medline: 33361590]
22. Lee J, Walker ME, Gabriel KP, Vasan RS, Xanthakis V. Associations of accelerometer-measured physical activity and sedentary time with chronic kidney disease: the Framingham Heart Study. PLoS One 2020;15(6):e0234825. [doi: 10.1371/journal.pone.0234825] [Medline: 32542048]
23. Au R, Seshadri S, Wolf PA, et al. New norms for a new generation: cognitive performance in the Framingham Offspring cohort. Exp Aging Res 2004;30(4):333-358. [doi: 10.1080/03610730490484380] [Medline: 15371099]
24. De Anda-Duran I, Sunderaraman P, Searls E, et al. Comparing cognitive tests and smartphone-based assessment in 2 US community-based cohorts. J Am Heart Assoc 2024 Jan 16;13(2):e032733. [doi: 10.1161/JAHA.123.032733] [Medline: 38226519]
25. Chen J, Doyle MF, Fang Y, et al. Eripheral inflammatory biomarkers are associated with cognitive function and dementia: Framingham Heart Study Offspring cohort. Aging Cell 2023 Oct;22(10):e13955. [doi: 10.1111/acel.13955] [Medline: 37584418]
26. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011 Mar;20(1):40-49. [doi: 10.1002/mpr.329] [Medline: 21499542]
27. Karr JE, Areshenkoff CN, Rast P, Hofer SM, Iverson GL, Garcia-Barrera MA. The unity and diversity of executive functions: a systematic review and re-analysis of latent variable studies. Psychol Bull 2018 Nov;144(11):1147-1185. [doi: 10.1037/bul0000160] [Medline: 30080055]
28. de Paula JJ, Malloy-Diniz LF, Romano-Silva MA. Reliability of working memory assessment in neurocognitive disorders: a study of the Digit Span and Corsi Block-Tapping tasks. Braz J Psychiatry 2016;38(3):262-263. [doi: 10.1590/1516-4446-2015-1879] [Medline: 27579598]
29. Omata K, Ito S, Takata Y, Ouchi Y. Similar neural correlates of planning and execution to inhibit continuing actions. Front Neurosci 2018;12:951. [doi: 10.3389/fnins.2018.00951] [Medline: 30631263]
30. Prigatano GP, de Oliveira SB, Goncalves CWP, Denucci SM, Pereira RM, Braga LW. Inhibitory control of adjacent finger movements while performing a modified version of the Halstead Finger Tapping Test: effects of age, education and sex. J Int Neuropsychol Soc 2021 Sep;27(8):813-824. [doi: 10.1017/S1355617720001101] [Medline: 33190661]
31. Brozzoli C, Roy AC, Lidborg LH, Lövdén M. Language as a tool: motor proficiency using a tool predicts individual linguistic abilities. Front Psychol 2019 Jul 16;10:1639. [doi: 10.3389/fpsyg.2019.01639] [Medline: 31379674]
32. Insel TR. Digital phenotyping: a global tool for psychiatry. World Psychiatry 2018 Oct;17(3):276-277. [doi: 10.1002/wps.20550] [Medline: 30192103]
33. Lee CY, Kang SJ, Hong SK, Ma HI, Lee U, Kim YJ. A validation study of a smartphone-based finger tapping application for quantitative assessment of bradykinesia in Parkinson's disease. PLoS ONE 2016 Jul 28;11(7):e0158852. [doi: 10.1371/journal.pone.0158852] [Medline: 27467066]
34. Roalf DR, Rupert P, Mechanic-Hamilton D, et al. Quantitative assessment of finger tapping characteristics in mild cognitive impairment, Alzheimer's disease, and Parkinson's disease. J Neurol 2018 Jun;265(6):1365-1375. [doi: 10.1007/s00415-018-8841-8] [Medline: 29619565]
35. Buracchio T, Dodge HH, Howieson D, Wasserman D, Kaye J. The trajectory of gait speed preceding mild cognitive impairment. Arch Neurol 2010 Aug;67(8):980-986. [doi: 10.1001/archneurol.2010.159] [Medline: 20697049]
36. For researchers. Framingham Heart Study. URL: https://www.framinghamheartstudy.org/fhs-for-researchers [accessed 2026-01-24]

Abbreviations

eFHS: electronic Framingham Heart Study
FHS: Framingham Heart Study
FTT: finger tapping test
ITI: intertap interval
LASSO: least absolute shrinkage and selection operator
MCI: mild cognitive impairment

XSL•FO
RenderX

MMSE: Mini-Mental State Exam
PD: Parkinson disease

Edited by A Coristine; submitted 15.Aug.2025; peer-reviewed by G Souza, W Myung; revised version received 20.Dec.2025; accepted 20.Dec.2025; published 18.Feb.2026.

Please cite as:
Ding H, Orwig TA, Rong J, Tatere HY, Liu C, Schramm E, Pathiravasan CH, Murabito JM, Lin H
Associations Between Smartphone-Based Finger Tapping and Cognitive Performance in Older Adults: Observational Study
J Med Internet Res 2026;28:e82463
URL: https://www.jmir.org/2026/1/e82463
doi:10.2196/82463

© Huitong Ding, Taylor A Orwig, Jian Rong, Heaven Y Tatere, Chunyu Liu, Eric Schramm, Chathurangi H Pathiravasan, Joanne M Murabito, Honghuang Lin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Impact of Push Notifications on Physical Activity and Sodium Intake Among Patients with Hypertension: Microrandomized Trial of a Just-in-Time Adaptive Intervention

Jessica R Golbus[1,2*], MD, MS; Michael P Dorsch[3*], MS, PharmD; Yuxuan Chen[4], MS; Tanima Basu[1], MA, MS; Evan Luff[1], MS; Predrag Klasnja[5], PhD; Mark W Newman[5], PhD; Lesli E Skolarus[6], MD, MS; Walter Dempsey[4], PhD; Brahmajee K Nallamothu[1,2], MPH, MD

[1]Division of Cardiovascular Medicine, Department of Internal Medicine, University of Michigan, 2723 Cardiovascular Center, 1500 E Medical Center Dr. SPC 5853, Ann Arbor, MI, United States

[2]Michigan Integrated Center for Health Analytics and Medical Prediction (MiCHAMP), University of Michigan, Ann Arbor, MI, United States

[3]Department of Clinical Pharmacy, College of Pharmacy, University of Michigan, Ann Arbor, MI, United States

[4]Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, United States

[5]School of Information and Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, United States

[6]Division of Stroke Vascular Neurology, Davee Department of Neurology, Northwestern University Feinberg School of Medicine, Chicago, IL, United States

[*]these authors contributed equally

Corresponding Author:
Jessica R Golbus, MD, MS
Division of Cardiovascular Medicine, Department of Internal Medicine, University of Michigan, 2723 Cardiovascular Center, 1500 E Medical Center Dr. SPC 5853, Ann Arbor, MI, United States

Abstract

Background: Achieving adequate blood pressure control is challenging for patients and clinicians. Digital hypertension management solutions that use push notifications to promote lifestyle management have been proposed as an approach, but their effectiveness remains unknown.

Objective: This analysis was designed to interrogate the independent and short-term effects of push notifications, tailored to participant and environmental factors, and on physical activity levels and sodium intake among individuals with hypertension.

Methods: The myBPmyLife study was a 6-month randomized controlled trial of participants with self-reported hypertension recruited from an academic medical center and federally qualified health centers. A core component of the intervention consisted of microrandomized push notifications promoting lifestyle modifications that were randomly delivered at 4 daily time points and focused on physical activity and dietary sodium intake. Our primary outcome for this secondary analysis was the step count 60 minutes after a physical activity notification and lower-sodium food choices 24 hours after a dietary notification. This analysis focuses on the results of the microrandomized trial and used a centered and weighted least squares method adapted for 2 or more treatments.

Results: A total of 298 participants were randomized to the intervention arm, of whom 287 had data available for analysis. Participants' mean age was 59.5 (SD 13.1) years, 138 (48.1%) were women, and 206 (71.8%) were White. Participants were randomized at 187,517 time points over 6 months, which led to 0.96 (SD 0.86) push notifications per day divided between activity (50.4%; SD 0.4) and dietary (49.8%; SD 0.4) notifications. Activity notifications did not increase step count in the 60 minutes after a notification (estimate 1.01, 95% CI 0.98 - 1.04; *P*=.40). Similarly, dietary notifications did not impact the number of lower-sodium food choices in the subsequent 24 hours (estimate 0.93, 95% CI 0.83 - 1.04; *P*=.23), but in exploratory post hoc analyses, did increase mobile app use by 95.5% (95% CI 1.81 - 2.10; *P*<.001), mobile app clicks or searches by 93.7% (95% CI 1.72% - 2.16%; *P*<.001), and low sodium searches by 113.0% (95% CI 1.73 - 2.53; *P*<.001), all within 60 minutes.

Conclusions: In patients with hypertension, push notifications did not impact short-term physical activity levels or dietary sodium intake but did improve intervention engagement.

Trial Registration: ClinicalTrials.gov NCT05154929; https://clinicaltrials.gov/study/NCT05154929

International Registered Report Identifier (IRRID): RR2-10.1161/JAHA.123.031234

(J Med Internet Res 2026;28:e78218) doi:10.2196/78218

XSL•FO
RenderX

KEYWORDS

digital health; hypertension; mHealth; mobile health; physical activity; just-in-time adaptive intervention

Introduction

Nearly half of all US adults have hypertension, though only 1 in 4 have their blood pressure (BP) adequately treated [1]. Achieving adequate BP control is challenging for patients and clinicians given the episodic nature of clinical encounters, high patient volumes, and the silent nature of the disease process, promoting clinical inertia [2]. Digital hypertension management solutions aim to improve BP control by promoting patient self-management and expanded access to care. These solutions often center around connected BP cuffs for self-monitoring, with or without medication management, and may be paired with behavior and lifestyle change interventions in the form of mobile apps, text messages or push notifications, and phone calls [2]. Although these methods could enhance BP control, the effectiveness of each digital intervention component—especially within a multicomponent framework—continues to be a significant and unresolved issue.

We recently completed the myBPmyLife study, which evaluated a mobile health (mHealth) intervention designed to improve BP control by promoting increased physical activity and the selection of lower-sodium food choices. The study recruited participants from an academic medical center and federally qualified health centers [3]. The intervention consisted of a mobile app designed to facilitate goal setting and feedback and contextually tailored push notifications delivered as a part of a microrandomized trial, a novel experimental design in which participants are serially randomized to different types or levels of an intervention (eg, push notifications) [4]. While the intervention reduced sodium intake and improved physical activity levels, it failed to lower systolic BP (SBP) compared to a control group that engaged in BP self-monitoring alone [5].

One important question that remained was whether push notifications led to immediate short-term effects or mediated behavior change that occurred on a longer timescale. To address this question, we present the myBPmyLife study's 6-month microrandomized trial results, which focuses on important secondary, mechanistic outcomes of this trial. This analysis was designed to interrogate the independent and short-term effects of push notifications, tailored to participant and environmental factors, on physical activity levels and sodium intake. Such an approach allows us to isolate the causal effects of push notifications relative to the larger intervention package. We hypothesized that tailored push notifications would increase participants' step counts in the 60 minutes following delivery and decrease sodium intake in the subsequent 24 hours.

Methods

Study Design and Participants

The myBPmyLife study was a prospective, remotely administered, randomized-controlled trial (ClinicalTrials.gov NCT05154929) of 602 participants with self-reported hypertension (CONSORT [Consolidated Standards of Reporting Trials] guidelines; Checklist 1). Participants were recruited from the University of Michigan Health and the Hamilton Community Health Network, a series of federally qualified health centers in Flint, Michigan. The analysis described here focuses on the results of the microrandomized trial, which was embedded within the intervention arm of the randomized controlled trial, and which was composed of push notifications promoting increased physical activity and the selection of lower-sodium food choices. The data and code that support this study are openly available through GitHub.

Participants were eligible for the study if they were 18 years or older with self-reported hypertension, had no changes in their antihypertensive therapies in the preceding 4 weeks (if on medical therapies), consumed >1500 mg of daily sodium, and owned a compatible smartphone. Sodium intake was assessed by the Block Sodium Screener [6] after informed consent was obtained, and those who consumed <1500 mg/d of sodium were ineligible. Participants with contraindications to physical activity or to following a low-sodium diet were excluded from the study. The full inclusion and exclusion criteria have been previously published [3].

Ethical Considerations

The study was approved by the University of Michigan Health IRB (HUM00205845). All participants participated in an informed consent process and signed consent forms. All study data are deidentified and remain anonymous. Participants received up to US $100 over 6 months for their time completing study tasks. All participants were provided with a Bluetooth-connected smartwatch (Fitbit Versa 2) and a Bluetooth BP monitor (Omron Evolv BP7000) for the purposes of study participation.

Study Procedures

The myBPmyLife study was conducted between December 2021 and July 2023. Study procedures have been previously published. Briefly, participants underwent remote screening, recruitment, and consent processes. Consent forms were signed using the mobile study app myDataHelps (CareEvolution, LLC). Following the completion of the Block Sodium Screener, eligible participants were randomized 1:1 to the intervention arm, which received a BP monitor and an mHealth intervention delivered through a mobile app and a smartwatch, or to the enhanced usual arm, which received a BP monitor and smartwatch alone. Randomization schemes were created by study statisticians in Randomize.net and performed by study staff using a permuted block design with variable block sizes of 2-6, stratified by study site. Researchers and participants were unblinded to allocation assignment given the nature of the intervention.

Following informed consent, participants in both study arms were mailed a Fitbit Versa 2 and a Bluetooth BP monitor (Omron Evolv BP7000) and provided instructions on downloading the associated mobile apps to enable data sharing. All participants subsequently underwent remote enrollment appointments, at which time they were assisted with pairing

XSL•FO
RenderX

their smartwatches and smartphones as needed. Participants randomized to the intervention arm additionally answered a series of questions to facilitate intervention tailoring, including those on mobility, confidence in selecting lower-sodium food choices, preferred name, times of day for push notification (ie, morning, lunch, afternoon, and evening times), and preferred day for grocery shopping. These preferences could be changed during the study upon participant request. Following enrollment but before intervention receipt, participants experienced a baseline period in which no interventions were delivered to establish baseline step counts.

Description of the Intervention

The myBPmyLife study was designed to increase physical activity levels and promote the selection of lower-sodium food choices through an mHealth intervention grounded in the behavior change strategies of goal setting, prompts, visualizations, and feedback. The intervention consisted of both static (ie, mobile app with a central visualization promoting strategic feedback and goal setting) and dynamic components, with the latter referring to push notifications delivered as part of the microrandomized trial [5]. Microrandomized trials are an experimental approach to guide the design of just-in-time adaptive interventions. Just-in-time interventions, within the context of health behavior change, aim to provide support at times of opportunity or risk when individuals are receptive to change [7]. As such, just-in-time interventions aim to shape behavior change both in the moment and over time.

Microrandomized trials are an experimental design in which participants are serially randomized to receive (or not receive) different types or levels of an intervention (eg, push notifications) at different time points (ie, decision points) over the length of a study [4]. Intervention efficacy is ascertained by 1 or more proximal outcomes, which refer to short-term outcomes hypothesized to mediate a desired long-term effect. By leveraging intraindividual contrasts, microrandomization is a powerful experimental design for determining causal effects. In the myBPmyLife trial, microrandomized push notifications comprised a core component of the intervention (Figure S1 in Multimedia Appendix 1). They were designed to promote low-level physical activity and the selection of lower-sodium food choices. Activity notifications were tailored based on the time of day (ie, morning, lunch, afternoon, and evening), day of week (ie, weekend vs weekday), weather, and mobility. Similarly, dietary notifications were tailored according to the time of day, day of the week (ie, weekend, weekday, or grocery day), and participants' confidence in selecting lower-sodium food choices. Both notification types consisted of expert-generated and community-generated notifications, with the latter tailored based on participants' site of enrollment (ie, University of Michigan Health or Hamilton Community Health Network). A subset of notifications was personalized using participants' preferred names [8].

The study was designed so that participants would receive 1 activity or dietary notification per day, on average, throughout the study. Thus, participants had a 25% probability of receiving a notification at each decision point, divided equally between activity and dietary notifications. The proximal outcomes for the microrandomized trial serve as key secondary outcomes from the overarching clinical trial. The proximal outcome for activity notifications was step count 60 minutes after a decision point as determined by the smartwatch or mobile phone, as these messages were intended to be actionable in real time. For dietary notifications, the proximal outcome was self-reported lower-sodium food choices in the mobile app within 24 hours of a decision point, as these messages were intended to be applicable at a future time around meals or snacks and when at restaurants or grocery stores. For the dietary analysis, the decision was made post hoc to explore a series of more proximal engagement measures, as dietary notifications encouraged participants to use the mobile app to identify lower-sodium alternatives to commonly consumed or purchased foods. These exploratory outcomes included mobile app use (yes or no), the number of mobile app searches or clicks, and the number of low salt searches within the mobile app, all within 60 minutes of a decision point.

Statistical Analysis

Baseline clinical characteristics are described as means and SD for continuous symmetric variables and median with IQR for skewed continuous variables. Categorical variables are presented as counts and percentages. We performed Student 2-tailed *t* tests for bivariate comparisons between continuous variables and chi-square tests for comparisons across categorical variables.

For both the physical activity and dietary analyses, the intervention period was defined as the time of participants' first decision point at or beyond day 8 and lasting until day 180 (ie, maximum study period of 173 d). This was defined to account for variability in the duration of the baseline period (as determined by smartwatch wear time) and in the completion of the final surveys.

For the physical activity analysis, we evaluated the effect of delivering an activity notification compared to no notification on step count 60 minutes after a decision point. As step counts were positively skewed, 0.5 was added to all counts before natural log transformation. Analyses used a version of the weighted and centered least squares (WCLS) method. The WCLS approach is designed for microrandomized trials, where treatments occur frequently and moderators may be affected by prior interventions. This method involves centering time-varying treatments using their conditional means based on past history and applying weights similar to inverse probability of treatment weighting in causal inference. By combining centering and weighting, WCLS offers robustness against the misspecification of the working model of weighted outcome condition on history [9]. Our analyses used a version of the WCLS method adapted to account for 2 or more treatments [9]. Such an approach increases statistical power by avoiding treatment-specific models and guarantees the robust, unbiased inclusion of covariates to reduce noise. Models were adjusted for age (<65 vs >65 y), gender, race (White vs non-White), baseline mean daily step count (dichotomized on mean value), step count 30 minutes before a decision point (standardized as mean or SD), and time (d). At each decision point, participants were considered available to receive the intervention if they were wearing their smartwatches consistent with prior studies [10]. This was

XSL•FO
RenderX

operationalized by requiring that participants have at least 1 heart rate measurement recorded on their smartwatches in the 30 minutes before a decision point. Models included decision points at which participants were available to receive notifications, and step count data were available in the 30 minutes before and 60 minutes after a decision point (ie, excluded 5807, 3.1% of decision points). Subsequently, we conducted a series of exploratory analyses to understand the impact of key demographic (ie, age, gender, and community) and clinical (ie, baseline step count) characteristics on our primary outcome by evaluating the interaction between these covariates and treatment. An additional exploratory model evaluated the impact between time as a quadratic term and treatment. Finally, as outcome data can be present in instances of missing wear time as data can come from both the mobile phone and smartwatch, we conducted a sensitivity analysis in which we excluded the data from participants with zero or missing wear time 3 hours after a decision point. This allowed us to disentangle instances of smartwatch nonwear from zero step counts.

For the dietary analysis, we evaluated the effect of delivering a dietary notification compared to the composite of an activity notification or no notification on lower-sodium choices (as denoted in the mobile app) within 24 hours of a decision point. Given the potential overlap in push notifications over 24 hours, the assessment of treatment effect included only decision points at which participants received a dietary notification, and no additional dietary notifications were sent in the subsequent 24 hours (excluded 42,313, 22.6% decision points). These were contrasted to decision points at which no dietary notification was sent at that time or in the subsequent 24 hours. As lower-sodium choices were zero-inflated, we used the method proposed by Liu et al [11] and used the R (R Foundation for Statistical Computing) function *emee* in package *MRTAnalysis*. Models were adjusted for age (<65 vs >65 y), gender, race (White vs non-White), baseline sodium intake as measured by the Block Sodium Screener (dichotomized on mean value), low salt choices 30 minutes before a decision point, and time (d). Participants were considered available at all decision points given the extended time window for outcome assessment. As with the physical activity analysis, we also conducted a series of exploratory analyses to understand the potential impact of key demographics (ie, age, gender, and community) and baseline characteristics (eg, baseline sodium intake) on sodium intake by evaluating the interaction between these covariates and treatment. An additional exploratory model evaluated the impact between time as a quadratic term and treatment. Similar models were created for our exploratory outcomes of mobile app use (yes or no), number of mobile app clicks or searches, and number of low sodium searches within the mobile app, all within 60 minutes of a decision point. These similarly used the method proposed by Liu et al [11], though it included all decision points (N=187,517), unlike the prior analysis. Covariates in these models included age (<65 vs >65 y), gender, race (White vs Non-white), and either mobile app use, number of search or clicks, or number of low sodium choices 30 minutes before a decision point, respectively. In all cases, $P<.05$ was considered statistically significant. All analyses were conducted in R (version 4.3.1; R Foundation for Statistical Computing).

Results

Study Population

Between December 2021 and July 2023, 912 participants were screened for eligibility, of whom 602 were ultimately enrolled. A total of 298 participants were randomized to the intervention arm of the study, 4 of whom withdrew before intervention receipt during the initial 5 days of the study (Figure 1). An additional 15 participants withdrew from the study, though they allowed their data to be used up until their withdrawal date (Figure 1). Post hoc, an additional 7 participants were excluded from the analysis (6 due to missing baseline step count and 1 missing step count 30 min prior to the decision point), leaving 287 participants available for the primary analysis.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram. Between December 2021 and July 2023, 912 participants were screened for eligibility. Following informed consent, participants were randomized 1:1 to the intervention and enhanced usual care arms with 298 participants randomized to the intervention arm of the trial. Of these, 4 participants withdrew before intervention receipt (postrandomization exclusions), and an additional 15 participants withdrew from the study but allowed their data to be used up until their withdrawal date. A total of 287 participants were included in the analysis of the microrandomized trial.

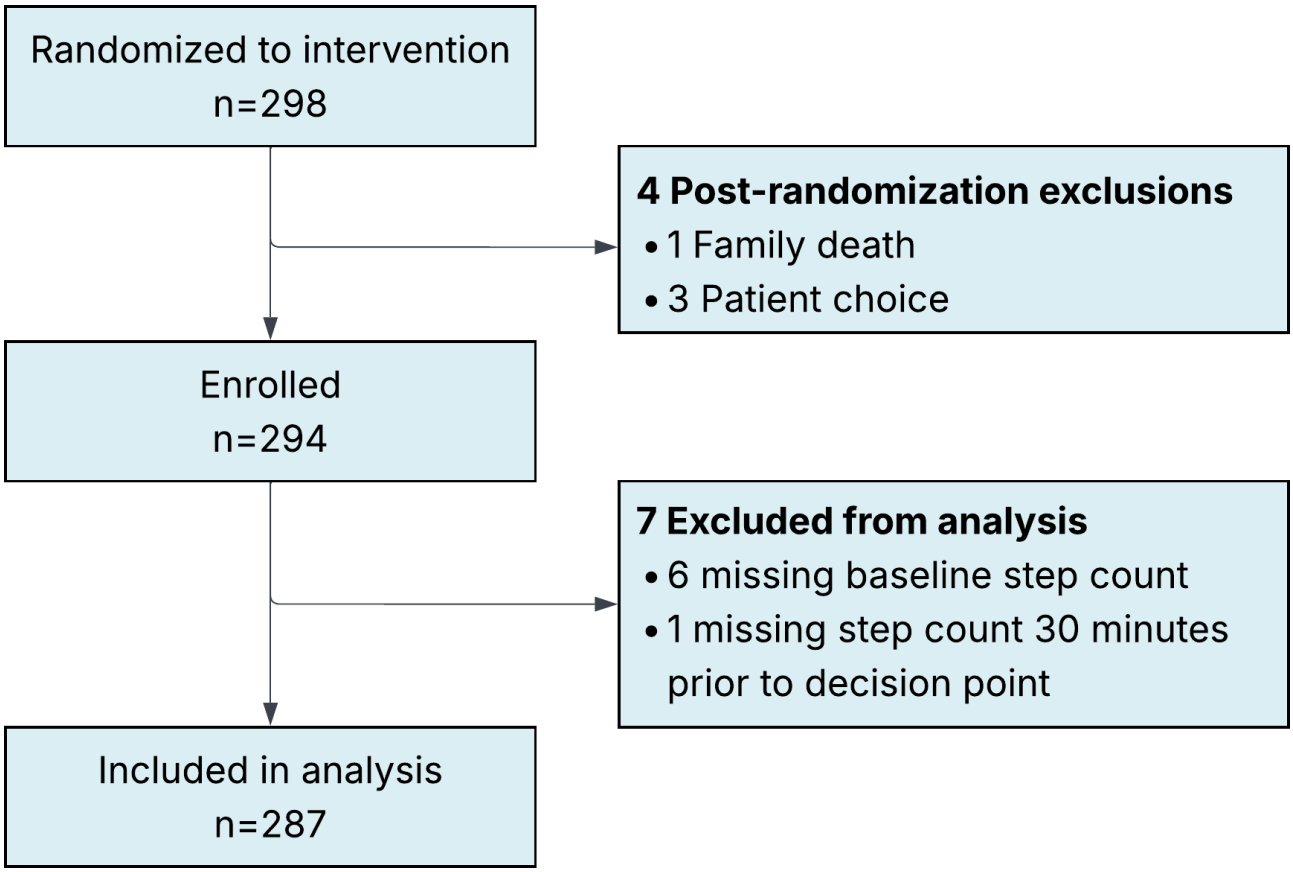

Participants were enrolled in the study for a median of 180 (IQR 37 - 180) days with a median intervention duration of 171 (IQR 25 - 173) days. The baseline characteristics of the population are displayed in Table 1. Participants had a mean age of 59.5 (SD 13.1) years, 138 (48.1%) were women, and 206 (71.8%) were White. Most were from the University of Michigan Health (242/287, 84.3%). Participants' baseline step count was 7408.1 (SD 3611.9) steps per day, and baseline sodium intake was 3072.7 (SD 1049.9) mg/d.

XSL•FO
RenderX

Table . Demographic and clinical characteristics of study population (N=287).

Characteristics	Ann Arbor (n=242)	Flint (n=45)	Overall
Age (y), mean (SD)	61.35 (12.5)	49.24 (11.6)	59.5 (13.1)
Gender, n (%)			
Woman	106 (43.8)	32 (71.1)	138 (48.1)
Man	136 (56.2)	13 (28.9)	149 (51.9)
Race, n (%)			
Asian	28 (11.6)	0 (0)	28 (9.8)
Black	20 (8.3)	20 (44.4)	40 (13.9)
Multiple	2 (0.8)	4 (8.9)	6 (2.1)
Other[a]	5 (2.1)	2 (4.4)	7 (2.4)
White	187 (77.3)	19 (42.2)	206 (71.8)
Ethnicity, n (%)			
Hispanic	11 (4.5)	1 (2.2)	12 (4.2)
Non-Hispanic	231 (95.5)	44 (97.8)	275 (95.8)
Baseline step count (steps/d) mean (SD)	7633.09 (3564.1)	6197.87 (3666.5)	7408.06 (3611.9)
Baseline estimated dietary sodium intake (mg/d), mean (SD)	3054.95 (1016.0)	3167.96 (1223.4)	3072.67 (1049.7)
Self-reported comorbid conditions, n (%)			
Chronic kidney disease	12 (5.0)	4 (8.9)	16 (5.6)
High cholesterol	109 (45.0)	16 (35.6)	125 (43.6)
Depression	39 (16.1)	13 (28.9)	52 (18.1)
Coronary artery disease	9 (3.7)	1 (2.2)	10 (3.5)
Stroke	7 (2.9)	2 (4.4)	9 (3.1)
Diabetes mellitus	57 (23.6)	9 (20.0)	66 (23.0)

[a]Other race: American Indian, Native Hawaiian or Other Pacific Islander, or Other or refused to answer race question

Study Execution

Over the duration of the intervention, participants were randomized to receive or not receive a push notification at 187,517 decision points (ie, time points; median 679, IQR 89 - 694 decision points per participant). On average, participants had 4.0 (SD 0.2) decision points per day and were randomized to receive a push notification at 24.9% (SD 0.2) of decision points. These were nearly equally split between activity notifications (50.4%; SD 0.4) and dietary notifications (49.8%; SD 0.4), consistent with the study design. On average, participants were randomized to receive 0.96 (SD 0.86) push notifications per day, distributed nearly equally over the 4 time points (Tables S1-S3 in Multimedia Appendix 1). After accounting for participants' availability (considered available at mean 81.3%, SD 0.3 of decision points), participants were presumed to have received 0.79 (SD 0.83) push notifications each day.

Activity Intervention

The mean step count in the 60 minutes following a decision point was 456.1 (SD 707.0) steps. In a multivariable model accounting for the demographic and baseline characteristics of participants, activity notifications did not significantly impact 60-minute step count (estimate 1.01, 95% CI 0.98 - 1.04; *P*=.40; Tables S4 and S5 in Multimedia Appendix 1; Figure 2). In a subsequent sensitivity analysis excluding decision points at which participants had zero hour or missing wear time in the 3 hours after a decision point (17,452, 9.6% of decision points), the results were similar with no significant change in 60-minute step count (Table S6 in Multimedia Appendix 1). To understand potential time-varying effects, we performed an exploratory analysis in which we modeled the intervention effect as a quadratic term over time, which did not significantly impact intervention efficacy (Figure 2). Similarly, in a sequence of univariable moderator analyses, age, community, and gender did not significantly impact intervention efficacy. In an exploratory analysis, however, push notifications were more effective in the subgroup of participants that were less active at baseline (dichotomized based on mean daily step count), increasing 60-minute step count by 4% (estimate 1.04, 95% CI 1.00 - 1.08; *P*=.03).

XSL•FO
RenderX

Figure 2. Impact of receiving (A) an activity notification on step count 60 minutes after a decision point or (B) a dietary notification on lower-sodium food choices 24 hours after a decision point. The models show treatment effect over time, with time modeled as a quadratic term. Activity notifications and dietary notifications did not significantly impact step count or the number of lower-sodium food choices, respectively. The results were exponentiated for interpretability.

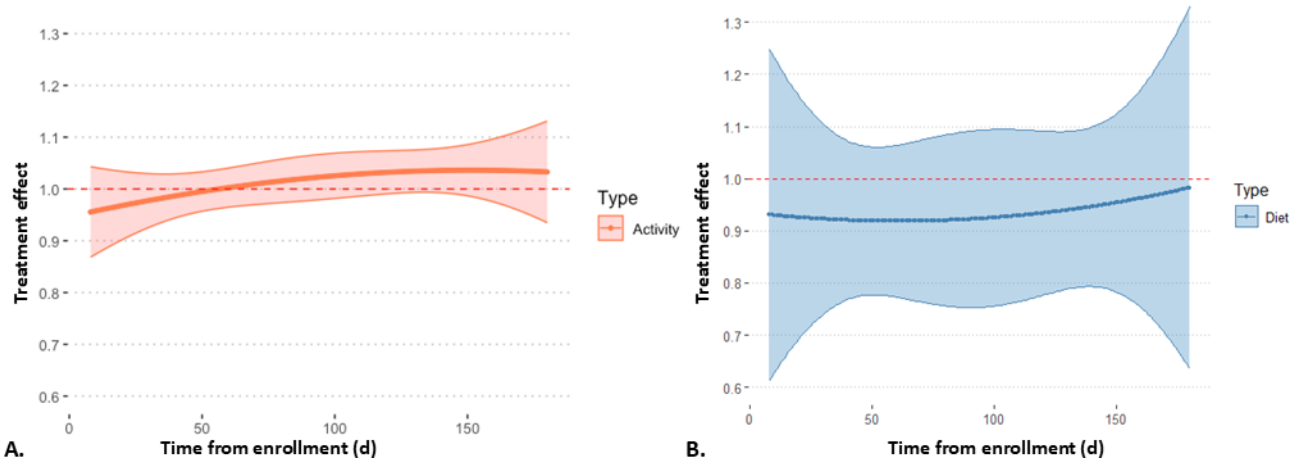

Dietary Intervention

In a multivariable model accounting for demographic and baseline characteristics, dietary notifications did not change the number of lower-sodium food choices in the 24 hours after a decision point (estimate 0.93, 95% CI 0.83 - 1.04; *P*=.23; Tables S4 and S7 in Multimedia Appendix 1; Figure 2). The results were similar in all subgroups of the population and did not change over time.

Subsequently, we conducted an exploratory analysis to evaluate the impact of push notifications on immediate measures of intervention engagement. Dietary notifications increased mobile app use in the subsequent 60 minutes by 95.5% (estimate 1.96, 95% CI 1.81 - 2.10; *P*< .001; Figure 3). Similarly, dietary push notifications increased the number of mobile app clicks or searches by 93.7% (estimate 1.94, 95% CI 1.72% - 2.16%; *P*<.001) and increased the number of low sodium food searches by 113% (estimate 2.13, 95% CI 1.73 - 2.53; *P*<.001), 60 minutes after a decision point (Figure 3). In all cases, the effect of dietary push notifications was the highest at the beginning of the study and decreased over time, though it remained significant throughout the study period.

Figure 3. Intervention engagement after dietary notifications. Dietary notifications significantly increased (**A**) mobile app use, (**B**) number of mobile app searches/clicks, and (**C**) number of low sodium searches, all within 60 minutes of a decision point. The models show treatment effect over time, with time modeled as a quadratic term. The results were exponentiated for interpretability.

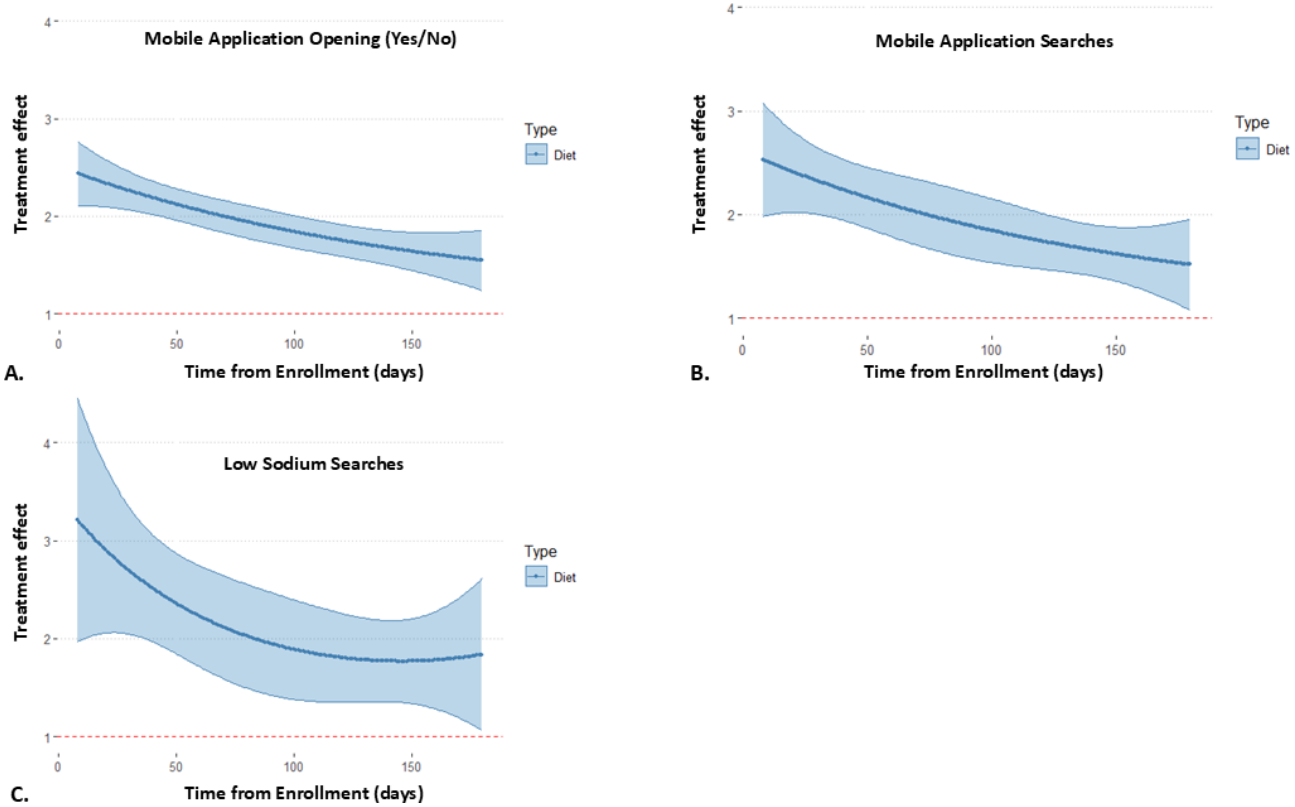

XSL•FO
RenderX

Discussion

Principal Findings

Digital hypertension solutions, typically delivered as multicomponent interventions, have shown promise for enhancing BP control [2,12,13]. However, the independent effects of push notifications when offered as part of a comprehensive digital hypertension intervention remain unknown. This is especially true with regard to the impact of push notifications on behavior change. We found through this trial that push notifications, tailored to participant (eg, community, mobility, and time) and environmental (eg, weather) factors, did not increase short-term physical activity levels or reduce sodium intake despite improving overall measures of these outcomes at 6 months. In an exploratory analysis, however, push notifications promoting physical activity were more effective for less active individuals, though this effect was overall very small, and for the overall cohort, the intervention was insufficient to overcome a participant's existing behavioral inertia. We also identified through an exploratory analysis that notifications that encouraged participants to interact with the mobile app and select lower-sodium food choices led to a nearly 2-fold increase in mobile app engagement, a measure of mechanical engagement though not necessarily effective behavioral engagement. While this effect persisted over the 6-month study period, the magnitude of effect decremented over time, suggesting either internalization of the behavior of interest (ie, engagement with the mobile app was no longer necessary for promoting adherence to a lower-sodium diet) or habituation.

Interpreting these results is best done within the context of those from the larger randomized controlled trial in which we found that the digital intervention package, consisting of both a mobile app and push notifications, increased physical activity levels (mean difference 489 steps, 95% CI 22-956) and reduced sodium intake (mean difference –285 mg, 95% CI −462 to –108) at 6 months [5]. One possible explanation for these results is that the interaction with the mobile app, which included multiple behavior change techniques (ie, goal setting, self-monitoring, and feedback) but not necessarily the content of the push notifications themselves, mediated the observed effects. This is supported by prior literature demonstrating positive associations between engagement with digital health interventions and health outcomes [14-16]. An alternative explanation, however, would be that push notifications did mediate the observed results, but that the proximal outcomes did not adequately capture their long-term impact. This could be due to lagged effects (ie, increase in physical activity beyond 60 min), incompletely captured effects (ie, failure to record lower-sodium food choices within the mobile app), or slower-developing mediators of the distal outcome (ie, knowledge about lower-sodium food choices, salience of activity goals). Future interventions should consider incorporating both proximal and intermediate outcomes within their study design, with the latter potentially capturing lagged effects (eg, 24 h step count and lower-sodium food choices over 48 h) or more slowly developing mediators of the target behavior.

Comparison With Prior Work

In general, digital hypertension solutions have focused on promoting BP self-monitoring, medication management, and lifestyle modification [2]. However, a challenge within the field has been in understanding the relative contributions of different digital intervention components, particularly when delivered as 1 part of a comprehensive intervention. In the pragmatic, randomized controlled trial BP Home, for example, over 2000 patients were randomized to BP self-monitoring using a standard device or to enhanced self-monitoring using a connected smartphone app [17]. Both groups experienced similarly large reductions in SBP (>10 mm Hg), without significant differences between the 2 groups. A second study found no significant differences in SBP reduction at 6 months using an artificial intelligence–enhanced conversational smartphone app promoting self-management, BP measurement, and lifestyle management, compared to a regular smartphone app paired with a home BP monitor [18]. There was, however, greater self-confidence in controlling BP in the intervention group and a trend toward greater self-reported physical activity.

To disentangle the effects of varying intervention components, Tucker et al [12] conducted an individual participant meta-analysis with over 7000 patients from 15 studies. Self-monitoring was associated with reduced SBP at 12 months, though this effect was strongly influenced by the intensity of co-interventions, with no effect with self-monitoring alone and a greater effect when combined with more intensive interventions, such as lifestyle counseling or systematic medication titration. These results have been supported by those from other studies, which, while mixed, suggest in aggregate that successful interventions are those with more comprehensive functionality and that incorporate medication management in conjunction with a care team [2,13]. A limitation of these and other studies, however, has been their focus on evaluating the distal effects of an intervention package, often delivered through a mobile app with multiple behavioral components. Our microrandomized trial analysis overcomes these limitations by isolating the impact of push notifications on shorter-term mediators of the desired long-term effect (eg, lifestyle management) and suggests how the mobile intervention package may impact long-term behavior change.

Study Strengths

Our study has several strengths. First, this is one of the first microrandomized trials that we are aware of among individuals with hypertension. Microrandomized trials are a novel experimental design, and as such, few studies have been performed among individuals with cardiovascular disease or cardiovascular risk factors. By incorporating serial randomization, microrandomized trials can provide insight into the causal effects of an intervention component over time in such patients. This serves to advance the science of behavioral interventions, which have traditionally been optimized using a series of randomized controlled trials. Randomized controlled trials, however, are designed to assess the average effect of an intervention package on a behavior of interest and not to investigate which components of an intervention are most efficacious, their time-varying effects, or what psychosocial or

XSL•FO
RenderX

contextual factors impact their efficacy [4,19]. Second, we delivered the trial remotely and enrolled participants from 2 sites, including a series of federally qualified health centers, enhancing the diversity of the study population and the generalizability of our findings. In general, the data around mHealth interventions for individuals with hypertension with digital barriers or from underrepresented groups have been limited [20]. Finally, we followed participants for 6 months, addressing a critical limitation of many mHealth interventions, namely the short time horizon.

Limitations

This study should be interpreted within the context of its limitations, which relate both to the study population and the intervention design and analysis. First, we only enrolled participants with a compatible smartphone. Smartphone ownership, however, is common across age, race, and socioeconomic groups, and we ensured the use of a broad range of possible devices [21]. Second, participants were active at baseline (mean 7408 [SD 3611.9] steps per day), had only mildly elevated BP, and were connected to their health care system. Given that this population was already active at baseline, this may have led to a ceiling effect. Similarly, the intervention may have been perceived as less salient to participants with relatively well-controlled BPs. Whether the results would be similar with less active individuals or to those with poorly controlled BP is unknown. Third, push notifications were tailored on a limited number of environmental and participant-level factors. It is unknown whether notifications would be more effective in promoting lifestyle modification if tailored to psychosocial constructs or an extended set of participant factors (eg, recent physical activity). We also did not evaluate notification efficacy based on message framing, nor did we evaluate treatment moderation based on attributes of the notifications (eg, time of day, weather). For example, in a 2×2 randomized experiment of over 500 participants designed to evaluate message framings, participants preferred the ability to choose message framing (autonomy-supportive vs controlling language) over the presumed preference for autonomy-supportive language [22]. Future studies should consider an expanded set of tailoring variables and evaluate for treatment moderation by the characteristics of the notifications. Fourth, we required that participants self-report low sodium choices within the mobile app. It is thus possible that dietary notifications increased the number of low sodium food choices, though these were not reported by participants in the mobile app. Fifth, it is possible that the effect size of notifications may have been less than anticipated, leading us to underpower the study for our proximal outcomes. Finally, we required that participants be wearing their smartwatches to be considered available for the physical activity intervention. We assumed that participants were wearing their smartwatches if they had 1 or more heart rate measurements in the 30 minutes before a message was sent. Although this increased the rigor of our analyses, we cannot confirm that participants were wearing their smartwatches when push notifications were sent. It is also possible that push notifications had a delayed or unmeasured effect for participants not wearing their smartwatches or who received the push notifications at a later time.

Conclusions and Future Directions

In conclusion, in this microrandomized trial, we demonstrated that tailored push notifications did not increase short-term physical activity levels or reduce sodium intake among individuals with hypertension. We did observe greater mobile app engagement following push notifications, which may have mediated the observed longer-term effects on sodium intake and physical activity levels. These results suggest that push notifications may be effective in promoting intervention engagement. Additional studies, however, are needed to identify which individuals benefit most from push notifications and how to optimally tailor push notifications based on environmental and psychosocial factors, particularly when delivered as part of multicomponent interventions. Furthermore, future studies may consider mediation analyses to enhance our understanding of the impact of intervention engagement on longer-term measures of behavior change and which evaluate the impact of novel methodological approaches such as reinforcement learning algorithms on notification efficacy.

Funding

This study received funding from American Heart Association, Inc (AWD014891).

Data Availability

The data and code that support this study are openly available through GitHub [23].

Authors' Contributions

JRG, MPD, PK, MWN, LES, WD, and BKN contributed to the conceptualization of the project, investigation, and methodology. MPD, MWN, LES, and BKN were responsible for funding acquisition. EL was responsible for data curation. YC and TB were responsible for the formal analysis under the oversight of JRG, WD, and BKN. JRG and MPD drafted the original manuscript (writing – original draft) which was reviewed and edited by all co-authors (writing – review & editing).

Conflicts of Interest

MPD receives funding from the Agency for Health Research and Quality, the National Heart, Lung, and Blood Institute, and the American Heart Association. JRG receives funding from the National Institutes of Health (NIH; 1K23HL168220) and the Patient-Centered Outcomes Research Institute and from Blue Cross and Blue Shield of Michigan for her role in the Blue Cross

XSL•FO
RenderX

Blue Shield of Michigan Cardiovascular Consortium. LES receives funding from the National Institute of Neurologic Diseases and Stroke, the National Institute of Minority Health and Health Disparities, and the American Heart Association. BKN receives funding from the NIH, VA HSR&D, American Heart Association, and Janssen. He also receives compensation as Editor-in-Chief of Circulation: Cardiovascular Quality & Outcomes, a journal of the American Heart Association. Finally, he is a co-inventor on US Utility Patent US15/356,012 (US20170148158A1) entitled “Automated Analysis of Vasculature in Coronary Angiograms,” which uses software technology with signal processing and machine learning to automate the reading of coronary angiograms, held by the University of Michigan. The patent is licensed to AngioInsight, Inc., where BKN holds ownership shares and receives consultancy fees. MWN receives funding from the National Heart, Lung, and Blood Institute and the National Institute on Drug Abuse. PK is a principal investigator or a co-investigator on research grants from the NIH.

Multimedia Appendix 1
Supplemental tables and figures.
[DOCX File, 238 KB - jmir_v28i1e78218_app1.docx]

Checklist 1
CONSORT 2025 checklist.
[PDF File, 216 KB - jmir_v28i1e78218_app2.pdf]

References

1. Martin SS, Aday AW, Almarzooq ZI, et al. 2024 heart disease and stroke statistics: a report of US and global data from the American Heart Association. Circulation 2024 Feb 20;149(8):e347-e913. [doi: 10.1161/CIR.0000000000001209] [Medline: 38264914]
2. Digital hypertension management solutions. Peterson Health Technology Institute. 2024. URL: https://phti.org/assessment/digital-hypertension-management-solutions/ [accessed 2025-12-26]
3. Golbus JR, Jeganathan VSE, Stevens R, et al. A physical activity and diet just-in-time adaptive intervention to reduce blood pressure: the myBPmyLife study rationale and design. J Am Heart Assoc 2024 Jan 16;13(2):e031234. [doi: 10.1161/JAHA.123.031234] [Medline: 38226507]
4. Golbus JR, Dempsey W, Jackson EA, Nallamothu BK, Klasnja P. Microrandomized trial design for evaluating just-in-time adaptive interventions through mobile health technologies for cardiovascular disease. Circ Cardiovasc Qual Outcomes 2021 Feb;14(2):e006760. [doi: 10.1161/CIRCOUTCOMES.120.006760] [Medline: 33430608]
5. Dorsch MP, Golbus JR, Stevens R, et al. Physical activity and diet just-in-time adaptive intervention to reduce blood pressure: a randomized controlled trial. NPJ Digit Med 2025 Jul 14;8(1):438. [doi: 10.1038/s41746-025-01844-3] [Medline: 40659778]
6. Tangney CC, Rasmussen HE, Richards C, Li M, Appelhans BM. Evaluation of a brief sodium screener in two samples. Nutrients 2019 Jan 14;11(1):166. [doi: 10.3390/nu11010166] [Medline: 30646541]
7. Nahum-Shani I, Hekler EB, Spruijt-Metz D. Building health behavior models to guide the development of just-in-time adaptive interventions: a pragmatic framework. Health Psychol 2015 Dec;34S:1209-1219. [doi: 10.1037/hea0000306] [Medline: 26651462]
8. Hellem AK, Casetti A, Bowie K, et al. A community participatory approach to creating contextually tailored mHealth notifications: myBPmyLife project. Health Promot Pract 2024 May;25(3):417-427. [doi: 10.1177/15248399221141687] [Medline: 36704967]
9. Boruvka A, Almirall D, Witkiewitz K, Murphy SA. Assessing time-varying causal effect moderation in mobile health. J Am Stat Assoc 2018;113(523):1112-1121. [doi: 10.1080/01621459.2017.1305274] [Medline: 30467446]
10. Golbus JR, Shi J, Gupta K, et al. Text messages to promote physical activity in patients with cardiovascular disease: a micro-randomized trial of a just-in-time adaptive intervention. Circ Cardiovasc Qual Outcomes 2024 Jul;17(7):e010731. [doi: 10.1161/CIRCOUTCOMES.123.010731] [Medline: 38887953]
11. Liu X, Qian T, Bell L, Chakraborty B. Incorporating nonparametric methods for estimating causal excursion effects in mobile health with zero-inflated count outcomes. Biometrics 2024 Mar 27;80(2):ujae054. [doi: 10.1093/biomtc/ujae054] [Medline: 38837902]
12. Tucker KL, Sheppard JP, Stevens R, et al. Self-monitoring of blood pressure in hypertension: a systematic review and individual patient data meta-analysis. PLoS Med 2017 Sep;14(9):e1002389. [doi: 10.1371/journal.pmed.1002389] [Medline: 28926573]
13. Alessa T, Abdi S, Hawley MS, de Witte L. Mobile apps to support the self-management of hypertension: systematic review of effectiveness, usability, and user satisfaction. JMIR mHealth uHealth 2018 Jul 23;6(7):e10723. [doi: 10.2196/10723] [Medline: 30037787]
14. Mclaughlin M, Delaney T, Hall A, et al. Associations between digital health intervention engagement, physical activity, and sedentary behavior: systematic review and meta-analysis. J Med Internet Res 2021 Feb 19;23(2):e23180. [doi: 10.2196/23180] [Medline: 33605897]

XSL•FO
RenderX

15. Alshurafa N, Jain J, Alharbi R, Iakovlev G, Spring B, Pfammatter A. Is more always better?: Discovering incentivized mHealth intervention engagement related to health behavior trends. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018 Dec;2(4):153. [doi: 10.1145/3287031] [Medline: 32318650]
16. Schoeppe S, Alley S, Van Lippevelde W, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2016 Dec 7;13(1):127. [doi: 10.1186/s12966-016-0454-y] [Medline: 27927218]
17. Pletcher MJ, Fontil V, Modrow MF, et al. Effectiveness of standard vs enhanced self-measurement of blood pressure paired with a connected smartphone application: a randomized clinical trial. JAMA Intern Med 2022 Oct 1;182(10):1025-1034. [doi: 10.1001/jamainternmed.2022.3355] [Medline: 35969408]
18. Persell SD, Peprah YA, Lipiszko D, et al. Effect of home blood pressure monitoring via a smartphone hypertension coaching application or tracking application on adults with uncontrolled hypertension: a randomized clinical trial. JAMA Netw Open 2020 Mar 2;3(3):e200255. [doi: 10.1001/jamanetworkopen.2020.0255] [Medline: 32119093]
19. Klasnja P, Hekler EB, Shiffman S, et al. Microrandomized trials: an experimental design for developing just-in-time adaptive interventions. Health Psychol 2015 Dec;34S:1220-1228. [doi: 10.1037/hea0000305] [Medline: 26651463]
20. Khoong EC, Olazo K, Rivadeneira NA, et al. Mobile health strategies for blood pressure self-management in urban populations with digital barriers: systematic review and meta-analyses. NPJ Digit Med 2021 Jul 22;4(1):114. [doi: 10.1038/s41746-021-00486-5] [Medline: 34294852]
21. Mobile fact sheet. Pew Research Center. 2025. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2025-12-30]
22. Smit ES, Zeidler C, Resnicow K, de Vries H. Identifying the most autonomy-supportive message frame in digital health communication: a 2x2 between-subjects experiment. J Med Internet Res 2019 Oct 30;21(10):e14074. [doi: 10.2196/14074] [Medline: 31670693]
23. MRT_myBPmyLife. GitHub. URL: https://github.com/WIRED-L-um/MRT_myBPmyLife/blob/main/README.md [accessed 2025-12-29]

Abbreviations

BP: blood pressure
CONSORT : Consolidated Standards of Reporting Trials
mHealth: mobile health
SBP: systolic blood pressure
WCLS: weighted and centered least square

Edited by A Coristine; submitted 29.May.2025; peer-reviewed by AA Lopez-Gonzalez, RJ Katz; revised version received 10.Dec.2025; accepted 10.Dec.2025; published 07.Jan.2026.

Please cite as:
Golbus JR, Dorsch MP, Chen Y, Basu T, Luff E, Klasnja P, Newman MW, Skolarus LE, Dempsey W, Nallamothu BK
Impact of Push Notifications on Physical Activity and Sodium Intake Among Patients with Hypertension: Microrandomized Trial of a Just-in-Time Adaptive Intervention
J Med Internet Res 2026;28:e78218
URL: https://www.jmir.org/2026/1/e78218
doi:10.2196/78218

© Jessica Rachel Golbus, Michael P Dorsch, Yuxuan Chen, Tanima Basu, Evan Luff, Predrag Klasnja, Mark W Newman, Lesli E Skolarus, Walter Dempsey, Brahmajee K Nallamothu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effects of an eHealth Cardiac Exercise Rehabilitation Platform for Patients After Percutaneous Coronary Intervention Based on the Persuasive Systems Design Model: Randomized Controlled Trial

Yang Liu[1], MS; Xiting Huang[1,2], MS; Ziying Dai[1], MS; Zhili Jiang[3], MS; Wenxiao Wu[1], MS; Jing Wang[1], MS; Zhiqian Wang[1], MS; Luyao Yu[1], MS; Hanyu Li[1], MS; Lihua Huang[1], PhD

[1]The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China

[2]Department of Nursing, Sun Yat-sen University Cancer Center, Guangzhou, China

[3]School of Humanities and Management, Zhejiang Chinese Medical University, Hangzhou, China

Corresponding Author:
Lihua Huang, PhD
The First Affiliated Hospital, Zhejiang University School of Medicine
79 Qingchun Road
Shangcheng District
Hangzhou, 310003
China
Phone: 86 13867129329
Email: lihuahuang818@zju.edu.cn

Abstract

Background: Cardiac exercise rehabilitation is an important intervention for disease management of patients with coronary heart disease (CHD) after percutaneous coronary intervention (PCI). Still, the participation and compliance with exercise rehabilitation remain suboptimal. Mobile health technology is a promising approach to promoting involvement in cardiac exercise rehabilitation. Remote rehabilitation can overcome the problems existing in traditional rehabilitation.

Objective: This study aimed to evaluate the effects of an eHealth cardiac rehabilitation (CR) platform based on the persuasive systems design model in addition to standard CR after PCI on physical activity (PA), exercise endurance, self-perceived fatigue, exercise self-efficacy (ESE), and quality of life for patients after PCI.

Methods: A single-blinded, parallel, randomized controlled trial design was used. The study was conducted in the Department of Cardiology of a tertiary hospital in Hangzhou, China. A total of 180 eligible patients with CHD were enrolled from June to December 2023. Participants were randomly assigned (1:1) to the intervention group or the control group, with 90 patients in each group. The study is a 24-week eHealth CR program. The primary outcome was PA level; the secondary outcomes included exercise endurance, self-perceived fatigue, ESE, and quality of life. Data on the primary and secondary outcome measures were collected at baseline (T0), at 12 weeks of intervention (T1), and at 4 (T2), 8 (T3), and 12 (T4) weeks of follow-up. The generalized estimating equation model was used to examine changes in the outcome variables between the 2 groups across the study end points.

Results: Generalized estimating equation analyses revealed significant group-by-time interactions for all outcome measures (all $P<.001$). At T4, compared with the control group, the intervention group demonstrated statistically significant improvements in the following outcomes: PA: median 1723.00 versus 805.50 Metabolic Equivalent Task minutes per week (β coefficient=937.29, 95% CI 867.61-1006.97); 6-minute walk distance: median 436.00 versus 405.00 m (β coefficient=31.00); self-perceived fatigue: median 9.00 versus 10.00 (β coefficient=–1.00, indicating reduced fatigue); ESE: 61.11 versus 27.78 (β coefficient=33.33); Short Form of 36 Health Survey Questionnaire score: 91.19 versus 84.13 (β coefficient=7.06; all $P<.001$). Notably, there was no significant difference in self-perceived fatigue between the 2 groups at T1 ($P=.50$).

Conclusions: The findings of this study demonstrate the effectiveness of the eHealth CR based on the persuasive systems design model in addition to standard CR after PCI in improving the PA level, exercise endurance, ESE, quality of life, and self-perceived fatigue of patients. These findings also provide insights into the application of an eHealth cardiac exercise rehabilitation interventions to enhance the rehabilitation of patients with CHD.

XSL•FO
RenderX

Trial Registration: China Clinical Trials Registry (ChiCTR) ChiCTR2300071666; https://www.chictr.org.cn/showprojEN.html?proj=197908

(J Med Internet Res 2026;28:e71450) doi:10.2196/71450

KEYWORDS

coronary heart disease; percutaneous coronary intervention; PCI; exercise rehabilitation; persuasive systems design; eHealth

Introduction

Background

Cardiovascular diseases (CVDs) have long been a global health scourge, exacting a heavy toll on both individual well-being and health care systems worldwide. The World Health Organization (WHO) highlights that CVDs claim approximately 17.9 million lives annually, constituting a staggering 32% of all global deaths [1]. In China, the situation is equally disconcerting. With the rapid pace of urbanization, changes in lifestyle, and an aging population, the prevalence of CVDs, especially coronary heart disease (CHD), has been on an upward trajectory [2,3]. This has led to a substantial increase in the number of patients undergoing percutaneous coronary intervention (PCI), a common and effective treatment for CHD.

Despite the remarkable advancements in PCI technology, which significantly improve the acute condition of patients by restoring blood flow to the heart, it is not a cure-all solution. Patients post PCI often face a plethora of challenges. Exercise intolerance is a prevalent issue, as a large proportion of these patients experience a decline in their physical capacity, which restricts their daily activities and quality of life [4]. Self-perceived fatigue is another common complaint, which can be attributed to the physiological stress of the intervention, underlying cardiac damage, and the body's recovery process [5]. Moreover, many patients struggle with suboptimal exercise self-efficacy (ESE), lacking the confidence to engage in regular physical activity (PA), which is crucial for their long-term cardiac health [6].

Cardiac rehabilitation (CR) has emerged as an essential component of post-PCI care. It is a comprehensive, multidisciplinary program that encompasses exercise training, risk factor modification, psychological counseling, and patient education [7]. Extensive research has demonstrated its effectiveness in improving exercise capacity, reducing cardiovascular risk factors, enhancing psychological well-being, and ultimately decreasing mortality and morbidity rates among patients post PCI [8]. For instance, a network meta-analysis found that multiple exercise interventions enhanced peak oxygen consumption, with high-intensity interval training showing the greatest effect, followed by combined water-based and moderate-intensity continuous training, and other interventions like combined aerobic and resistance exercise also demonstrated benefits [9].

However, traditional CR programs, which are typically centered around in-hospital or clinic-based services, are plagued with limitations. Geographical barriers pose a significant challenge, especially for patients living in rural or remote areas, who may have to travel long distances to access these services [10]. Time constraints are another hurdle, as many patients, particularly those who are still working or have family responsibilities, find it difficult to fit regular rehabilitation sessions into their busy schedules. Additionally, the cost associated with in-person rehabilitation, including transportation and potential loss of income during treatment, can be a deterrent for some patients [11]. The proportion of patients participating in the CR program is only 25% to 35% [12].

In recent years, remote cardiac rehabilitation (RCR) has been advancing rapidly, and the potential exists for the challenges of traditional facility-based CR programs to be addressed by delivering care to patients in the convenience of their own homes with real-time, personalized support [11]. The RCR uses different methods such as the internet, wearable devices, and mobile apps [13]. Despite the positive outcomes observed in the application of digital health interventions for CR, such as SMS, remote electrocardiographic monitoring, and mobile or web portal tools, these advancements have predominantly remained in research settings and have not yet been adopted widely in clinical practice [14]. A systematic review by Duff et al [15] assessed the application of behavior change techniques (BCTs) in eHealth interventions designed to increase PA in CVD, and identified feedback and monitoring as the most common BCT category implemented in these interventions. Behavior change theory proponents note that these theories explain how behavior change happens [16]. However, information system developers often prioritize using BCTs over understanding the underlying theories [17]. Systems lacking a strong theoretical basis may have conflicting mechanisms, harming long-term effectiveness.

Such systems are designed to form, alter, or reinforce the attitudes, behaviors, or compliance of their users voluntarily. A key element in behavior and attitude change is persuasion. The persuasive systems design (PSD) model addresses this gap by guiding the analysis of the persuasion context, including recognizing the intent of the persuasion, understanding the persuasion event, and defining the strategies in use [18]. It describes how to inject persuasive characteristics into the system to stimulate behavior change in the design process. In the functional design of the persuasive system, the characteristics of the system are divided into main task support, dialogue support, system credibility support, and social support [19], and a total of 28 persuasive principles are included. It fully explains how the design principles are translated into software requirements and further manifests the system characteristics. The information system developed and designed under the guidance of the framework can motivate users to implement self-management and trigger health behavior change [20].

XSL•FO
RenderX

Aims of This Research

In the context of post-PCI CR, there is a paucity of well-designed mobile health (mHealth) interventions that fully leverage the PSD model to comprehensively address patients' multifaceted needs. Existing studies either focus on single-aspect interventions or fail to fully capitalize on the potential of the PSD model. Therefore, this study aimed to develop and evaluate an eHealth CR platform grounded in the PSD model in addition to standard CR for patients post PCI. We hypothesized that such a platform could effectively improve patients' PA levels, exercise endurance, ESE, and quality of life, while reducing self-perceived fatigue.

Methods

Study Design

The study was a single-blinded, parallel, randomized controlled clinical trial. The study protocol complied with the Declaration of Helsinki. The report was in accordance with the CONSORT (Consolidated Standards of Reporting Trials) guidelines.

Participants

Participants were recruited in the Department of Cardiology of a tertiary hospital in Hangzhou, a provincial capital city in eastern China, from June to December 2023.

The inclusion criteria were as follows: (1) patients were older than 18 years old; (2) met the WHO diagnostic criteria of CHD, underwent PCI with 1 or 2 stents at first time, and were diagnosed as low-risk factors according to the risk classification of exercise rehabilitation of CHD; (3) patients with a left ventricular ejection fraction of 50% to 70%; (4) patients with cardiac function grade Ⅰ or Ⅱ; (5) transradial or ulnar artery puncture; (6) patients and their families agreed and actively cooperated with the exercise rehabilitation treatment; and (7) no language disorder, could read and speak Chinese, had no prescribed PA restrictions. Participants were excluded from the study if they (1) had presenting with cardiogenic shock, severe arrhythmia and cardiac function grade Ⅲ or above; (2) had severe liver and kidney dysfunction (liver dysfunction of Child-Pugh class B or C; estimated glomerular filtration rate <30 ml/[min·1.73m^2]), severe anemia (hemoglobin<60 g/L); (3) had severe pulmonary disease, such as chronic obstructive pulmonary disease, emphysema, bronchiectasis, and pulmonary heart disease; (4) had a history of falls within half a month, or were unable to exercise independently; (5) took antianxiety or depression drugs; (6) had previous history of cerebral hemorrhage; and (7) could not use WeChat (Tencent Holdings Limited) skillfully.

Sample Size

The sample size was determined by PA level based on a pilot trial, in which the SD was 133.81, and the mean difference between the 2 groups was 131. A total of 150 participants were required to detect a difference between the 2 groups at a 5% (2-sided) significant level with a power of 80%. Considering the loss to follow-up, the sample size of each group was at least 90, and the total sample size was at least 180, allowing a 20% dropout.

Randomization and Blinding

A total of 180 eligible participants were randomly assigned to the intervention or control group at a 1:1 ratio via block randomization (block size=4) to ensure balanced group sizes. The randomization was implemented by an independent statistician (not part of the core study team, nor involved in recruitment, assessment, or intervention) to avoid selection bias.

The statistician used SPSS (version 26.0; IBM Corp) to generate a random number sequence, pairing each number with a unique participant ID. Group assignment was predefined: even numbers for the intervention group and odd numbers for the control group. For allocation concealment, each ID, random number, and group assignment was sealed in a sequentially numbered opaque envelope, stored in a locked cabinet. Upon recruitment, the research coordinator, unaware of the sequence, retrieved the next envelope to inform the participant of their group.

Finally, 90 participants were allocated to each group. The intervention lasted for 12 weeks, with an additional 12-week follow-up period. Notably, an eHealth cardiac exercise rehabilitation based on the PSD model was provided in addition to standard CR; the volume and content of standard rehabilitation were consistent between the 2 groups to ensure that group differences in outcomes could be attributed to the eHealth platform intervention. Due to the distinguishable eHealth platform intervention, blinding of participants and intervention deliverers was unfeasible, but outcome assessors remained blinded. Thus, this was a single-blinded, randomized controlled trial.

Interventions

Both groups received standard CR to isolate the effect of the additional mHealth intervention in the intervention group. Standard care included in-hospital education, exercise guidance, and follow-up support.

In in-hospital education, a 60-minute group session was conducted by a cardiac nurse, covering PCI postcare knowledge included medication adherence, dietary management, warning signs of adverse events (AEs), and basic exercise principles.

In exercise guidance, a 30-minute one-on-one session was conducted with a physiotherapist to teach low-to-moderate intensity aerobic exercises, including brisk walking, stationary cycling, and resistance training, such as hand grip exercises suitable for home practice, with a recommended weekly volume of 150 minutes of moderate-intensity activity.

In the follow-up support, monthly 15-20 minutes each telephone check-ins were conducted to address patient questions, remind them of exercise and medication adherence, and collect brief feedback on physical status.

Intervention Group: Standard Care+mHealth Platform

Overview

The intervention group received a PSD model–based eHealth exercise rehabilitation in addition to standard CR. We established a development team composed of cardiology experts, physiotherapists, clinical nursing experts, evidence-based nursing experts, software engineers, and artists. According to

XSL•FO
RenderX

the design content, we constructed an exercise rehabilitation WeChat applet for patients with CHD after PCI. The specific construction process of the WeChat mini program is detailed in Figure 1.

Figure 1. Schematic diagram of the persuasive systems design–based eHealth application development process for patients with coronary heart disease after percutaneous coronary intervention.

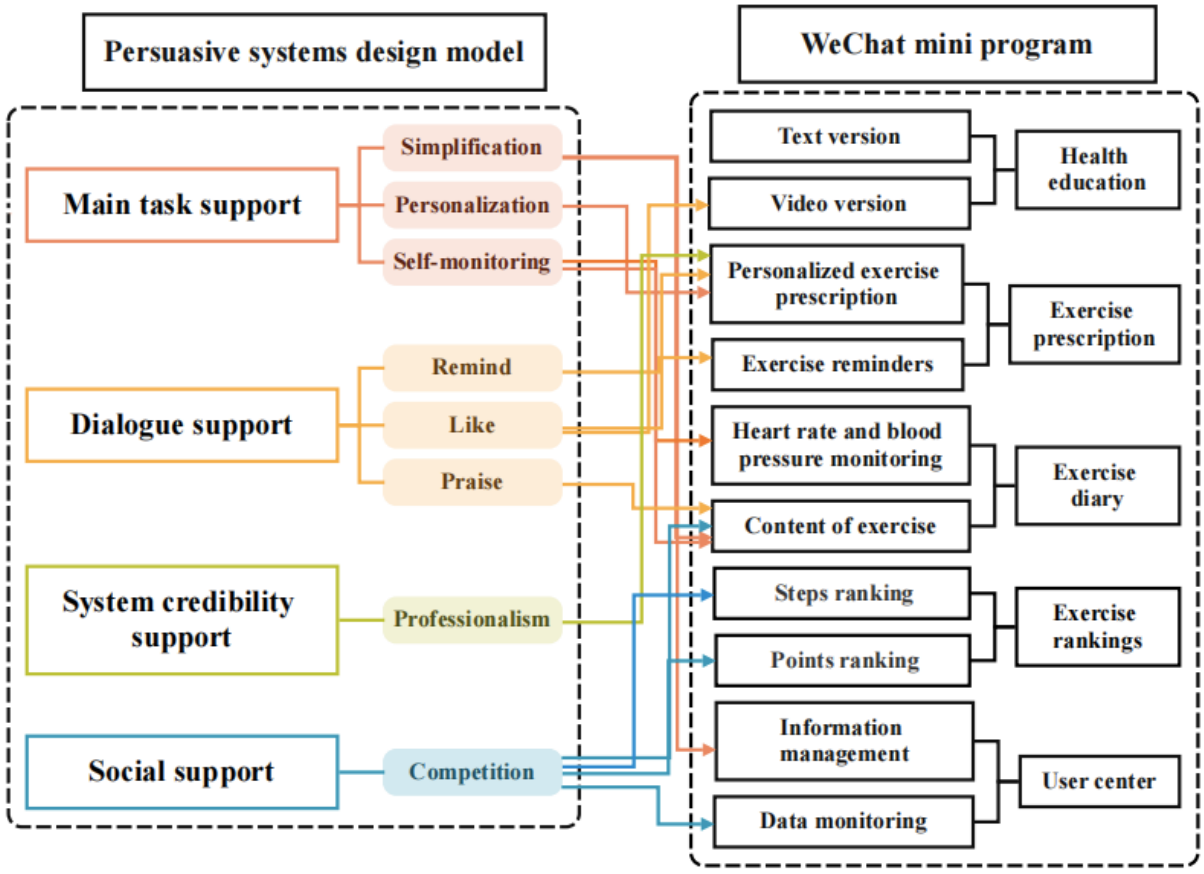

The eHealth platform was developed exclusively under the guidance of the PSD model—a theoretical framework specifically tailored for designing digital tools to stimulate behavior change. The platform's functional modules and design logic (schematic diagram illustrated in Figure 1) directly align with the 4 core components of the PSD model (main task support, dialogue support, system credibility support, and social support). The detailed integration of the PSD model into the intervention, its connection to global eHealth behavior change frameworks, and how it guided the platform's design are described further in this study.

Main Task Support: Simplifying Exercise Rehabilitation Adherence

PSD Theoretical Basis

The PSD model defines "main task support" as designing features to reduce user effort in completing core goals via principles like simplification and personalization. This component addresses a key barrier to digital health adherence—complex or nontailored tasks that increase user burden.

Integration Into the Platform

Simplification Principle

The home page was designed with a single "Exercise Prescription" button as the core entry point, reducing navigation steps from an average of 4 clicks to 1. Postexercise logging only requires 3 inputs (duration, heart rate, and discomfort), avoiding redundant data entry.

Personalization Principle

Based on baseline 6-minute walk distance (6MWD), left ventricular ejection fraction, and exercise preferences, the platform automatically generates individualized exercise plans (eg, 30-minute brisk walking for patients with 6MWD > 350 m vs 20-minute slow walking for 6MWD < 300 m)—aligned with PSD's focus on "matching tasks to user capabilities." This personalization is distinct from generic plans in non-PSD eHealth tools, which often use fixed templates.

Self-Monitoring Principle

Patients could record the data before and after each exercise in a diary, and track the progress of setting goals over time, so that patients could clearly understand their performance or status in the exercise rehabilitation process, and the rehabilitation

XSL•FO
RenderX

platform can support patients to achieve self-reporting and motivate patients to achieve exercise goals.

Dialogue Support: Enhancing Real-Time Interaction and Feedback

PSD Theoretical Basis

"Dialogue support" in the PSD model refers to interactive features that guide behavior via reminders, praise, and preference matching—critical for maintaining user engagement in long-term rehabilitation.

Integration Into the Platform

Reminder Principle

Patients set customizable exercise reminders (eg, "7 PM daily walk") with pop-up and vibration alerts. The platform also sends adaptive reminders (eg, "You haven't exercised in 3 days—start with a 15-minute walk today") if adherence drops, addressing the "forgetfulness" barrier common in patients post PCI.

Praise Principle

After each exercise session, the platform generates immediate positive feedback (eg, "Great job! You completed 100% of today's plan") and weekly "adherence badges" (eg, "3-week Consistent Exercise Award"). This aligns with PSD's goal of reinforcing desired behaviors through positive reinforcement.

Like Principle

Patients can also receive "likes" from the research team and anonymous peers (via the "Exercise Ranking" module) for completing exercise plans or achieving personal goals. The platform sends real-time notifications of likes (eg, "3 peers praised your consistent exercise!"), leveraging social approval to enhance motivation—consistent with PSD's dialogue support, focus on interactive feedback.

System Credibility Support: Building Trust in Digital Guidance

PSD Theoretical Basis

"System credibility support" ensures users perceive the platform as reliable via professionalism and transparency—a prerequisite for accepting digital health advice, especially in cardiac care.

Integration Into the Platform: Professionalism Principle

All exercise prescriptions and educational content were co-developed by a team of 2 cardiologists and 1 CR nurse, with references to the guidelines of the European Society of Cardiology [21] displayed in the "Health Education" section.

Social Support: Leveraging Social Influence for Adherence

PSD Theoretical Basis

"Social support" in the PSD model uses competitive or cooperative features to motivate behavior via social comparison—proven effective in digital health interventions for exercise adherence.

Integration Into the Platform: Competitive Principle

The "Exercise Ranking" module displays anonymized weekly exercise completion rates of other participants (eg, "You are in the top 25% of users this week")—without personal identifiers to protect privacy. This feature leverages mild social comparison to boost motivation, consistent with PSD's focus on positive social influence [22] and avoiding the pressure of direct competition.

Iterative Phases

The PSD model–based eHealth platform was developed in 4 iterative phases, with pilot testing integrated to refine functionality:

Phase 1: Framework Design (Month 1-2)

Guided by the PSD model, we identified four core persuasive features for CR:

1. Primary task support: personalized exercise prescriptions aligned with frequency, intensity, time, and type principles to simplify task completion.

2. Dialogue support: real-time feedback and reminders to enhance user engagement, including praise, likes, and adaptive reminders.

3. Credibility support: evidence-based educational content cited from the guidelines of the European Society of Cardiology [21] to build trust.

4. Social support: using competitive features to motivate behavior, a multidisciplinary team collaborated to draft the platform's functional modules (health education, exercise prescription, exercise diary, and exercise rankings) and user interface wireframes.

Phase 2: Prototype Development (Month 3-4)

Based on the framework, a prototype version (V1.0) was developed with basic functionalities—exercise plan generation based on baseline data input, manual heart rate logging (preliminary version before integrating wearable device connectivity), and static educational articles (n=10) on post-PCI care.

Phase 3: Pilot Testing and Revision (Month 5)

To validate usability and feasibility, we conducted a pilot test with 20 patients post PCI (mean age 62.3, SD 7.5 years; 12/20, 60% male) who met the study's inclusion criteria.

Pilot Testing Procedures

Usability assessment: Patients used V1.0 for 4 weeks, then completed the System Usability Scale (SUS) [23].

Functional evaluation: The research team recorded technical issues, such as loading delays and calculation errors relating to exercise volume, as well as adherence metrics, such as the daily login and module completion rates.

Key Revisions After Pilot Testing

Technical fixes: In total, 2 critical bugs were resolved (eg, an incorrect MET [Metabolic Equivalent Task]-min/week calculation), and the loading speed was optimized by 40%.

XSL•FO
RenderX

Usability improvements: Simplified the exercise logging interface (reduced input fields from 5 to 3) and added video tutorials for first-time users.

Content expansion: Increased educational materials to 20 (added 5 videos on exercise form correction based on patient feedback).

The revised version (V2.0) achieved a posttest SUS score of 78.5 (SD 6.2) versus 62.3 (SD 8.1) for V1.0 (P<.01), indicating acceptable usability.

Phase 4: Finalization (Month 6)

Incorporated feedback from the pilot test to finalize the platform (V3.0), which included all features personalized, simplification, self-monitoring, and so on. A 1-day training session was conducted for researchers to ensure consistent data collection during the formal study.

Establishment of the Exercise Prescription Library

To ensure the safety, feasibility, and effectiveness of exercise interventions for patients post PCI, we developed a tailored exercise prescription library based on evidence summary and multidimensional considerations, as illustrated in Figure 2.

Figure 2. Exercise prescription library for cardiac rehabilitation in patients with coronary heart disease after percutaneous coronary intervention in a tertiary hospital in Hangzhou, China.

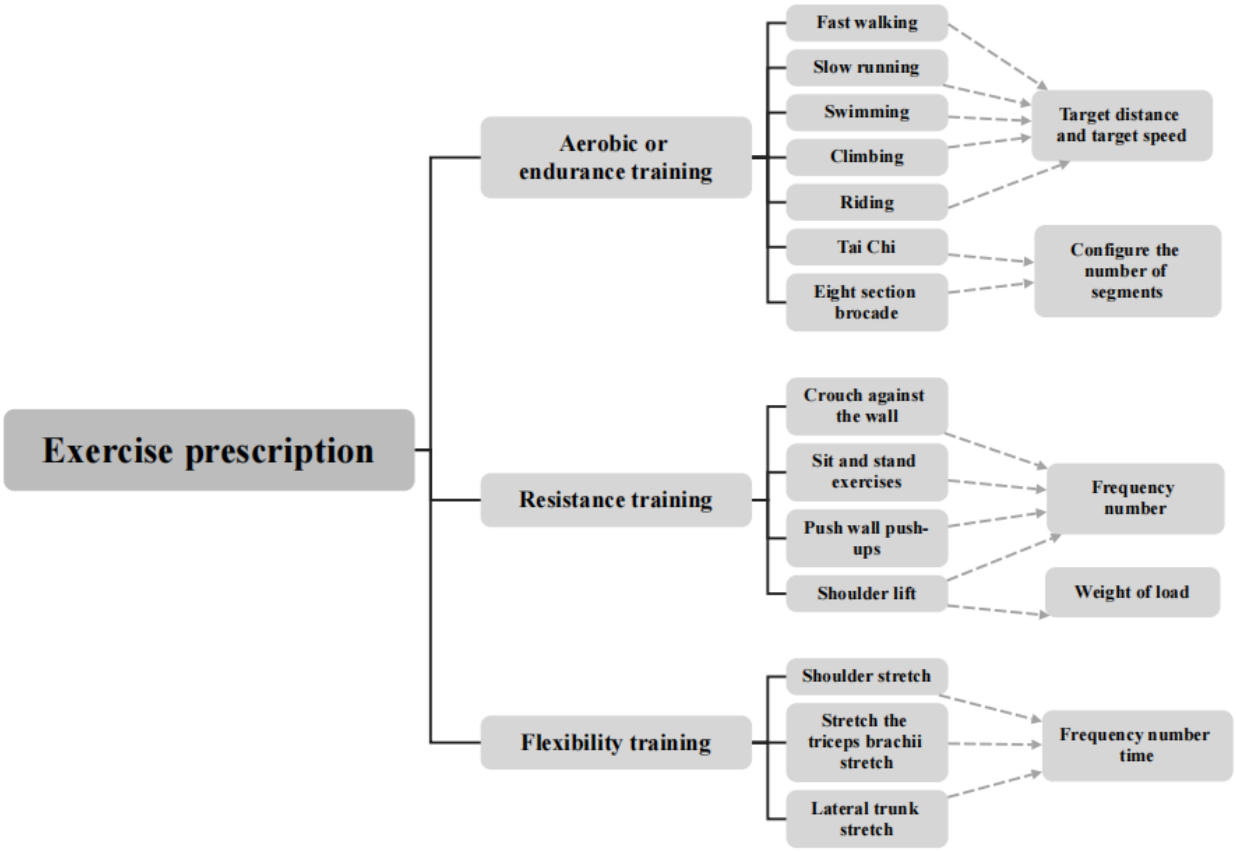

Evidence Summary for Prescription Development

First, we conducted a systematic review of guidelines and studies on post-PCI exercise rehabilitation, including (1) 2021 ESC Guidelines for CR [21], which recommend 150 minutes/week of moderate-intensity aerobic exercise for patients with CHD; (2) a network meta-analysis by Gomes-Neto et al [9], confirming that structured exercise improves exercise capacity in patients post PCI; and (3) Chinese expert consensus on CR [24], emphasizing gradual intensity progression and individualization for Chinese populations.

From these, we extracted core parameters, that is, suitable intensity (40-60% of maximal oxygen uptake), optimal frequency (3-5 sessions/week), and safe duration (20-40 minutes/session for moderate-intensity activity).

The details of the multidimensional considerations in library construction are presented in Multimedia Appendix 1.

The library was reviewed and approved by a panel of 3 CR specialists (with more than 10 years of experience) to ensure clinical appropriateness.

There were 5 basic modules in the program, namely, health education, exercise prescription, exercise diary, exercise ranking, and personal center. The user interface design diagram is illustrated in Figure 3.

XSL•FO
RenderX

Figure 3. User interface diagram of the persuasive systems design–based WeChat mini program for exercise rehabilitation in patients with coronary heart disease after percutaneous coronary intervention.

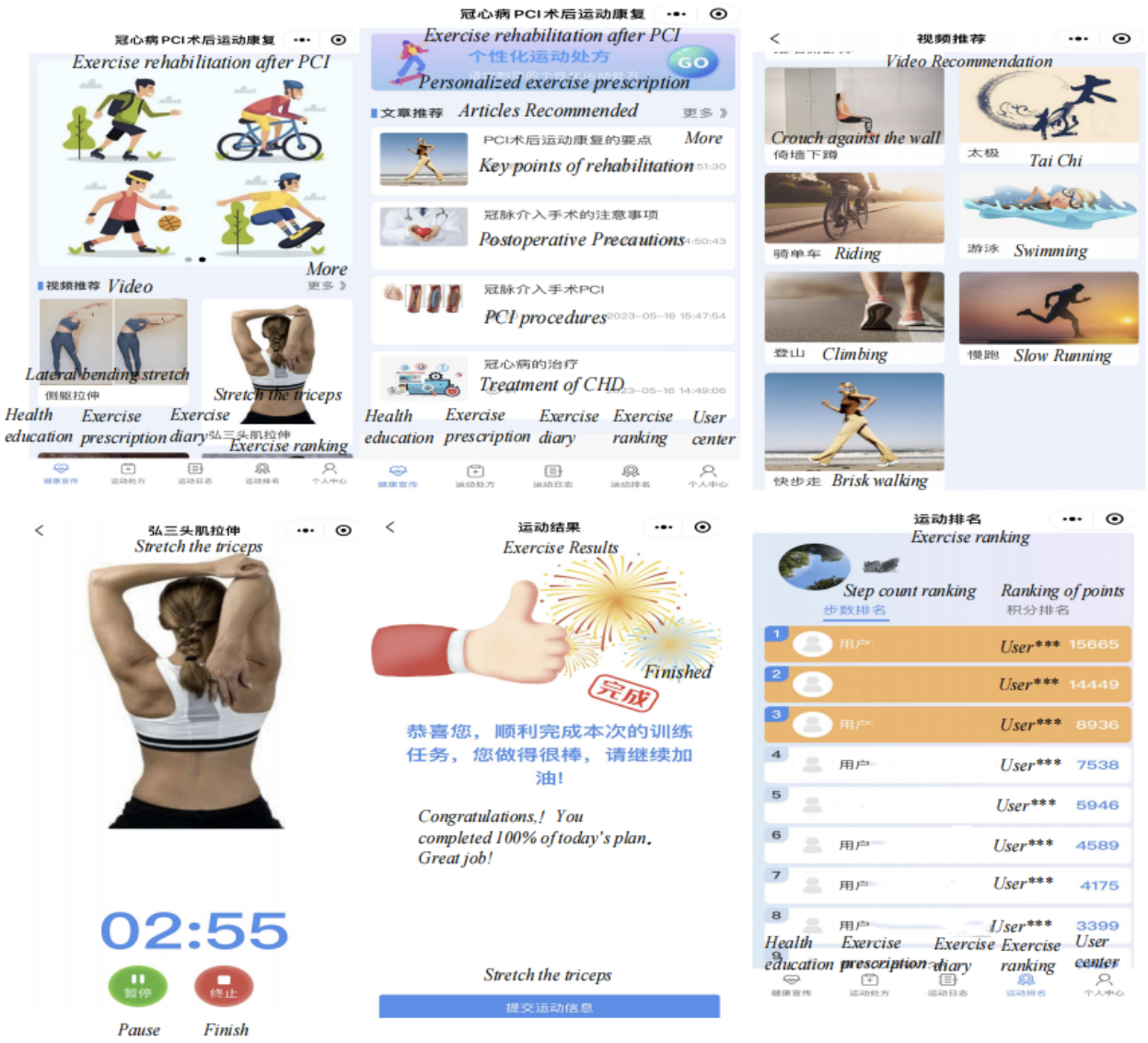

We introduced the use of a WeChat mini program to patients before discharge, ensuring that patients could successfully operate the mini program. The mini program's health education module allows patients to access disease knowledge through text and video formats, complete cardiac function grading assessments, 6-minute walking tests, Borg fatigue ratings, and monitor relevant vital signs within 1 week after discharge. Personalized exercise prescriptions are then tailored for patients based on comprehensive data evaluation. Patients can set weekly exercise reminders and follow the guidance process of the mini program to start their exercise rehabilitation. After each session, vital signs and discomfort symptoms during exercise are recorded in the exercise log module. Medical staff can timely adjust patients' exercise prescriptions according to their weekly situation and provide feedback on problems. Participants were required to continue to use the WeChat mini program for 12 weeks, and the step tracking function of the WeChat mini program was active during the 12-week intervention period. The duration of using the WeChat mini program, the frequency of weekly login to the platform, and the specific exercise data recorded in the exercise log could also be monitored through the data background of the medical and nursing side.

According to the frequency, intensity, time, and type principle, the specific structure of the selected exercise items, namely the frequency, intensity, and time of exercise, can be adjusted, which can realize the personalized adjustment of exercise tasks and task details, and provide a basis for the customization of personalized exercise prescription. Low-intensity exercise was defined as 30%-49% of maximal heart rate (or 3-5 METs), requiring aerobic or endurance training twice a week; moderate-intensity exercise as 50%-69% of maximal heart rate (or 6-8 METs), with aerobic or endurance training 3 times a week, resistance training once a week, and flexibility training once a week. High-intensity exercise, as ≥70% of maximal heart rate (or ≥9 METs), consisted of aerobic or endurance training 5 times a week, resistance training twice a week, and flexibility training twice a week.

XSL•FO
RenderX

Measurements

Overview

A self-designed demographic questionnaire was used to collect baseline sociodemographic data, including age, sex, marital status, education, employment status, living conditions, course of disease, and so on. Clinical data, including comorbidities, number of diseased vessels, cardiac function classification, PA level, scores of exercise endurance, self-perceived fatigue, ESE, and quality of life, were retrieved from the medical records. Primary and secondary outcomes were assessed with validated questionnaires. Questionnaires were collected by trained researchers through a pencil-and-paper survey. A previous intervention study [25] suggested that the exercise rehabilitation effect of patients with CHD after PCI should be intervened for at least 12 weeks, and the difference was significant. Therefore, the intervention period of this study was 12 weeks, and follow-up was conducted at 4 weeks, 8 weeks, and 12 weeks after the end of the intervention. Data collection time points were: within 3 days of admission (T0), at 12 weeks of intervention (T1), 4 weeks of follow-up (T2), 8 weeks of follow-up (T3), and 12 weeks of follow-up (T4).

Primary Outcome

The primary outcome focused on increased PA, which was assessed at baseline (T0), at 12 weeks of intervention (T1), 4 weeks of follow-up (T2), 8 weeks of follow-up (T3), and 12 weeks of follow-up (T4). PA comprises body movements that use energy. PA level was defined as weekly exercise volume, quantified by MET-min/week via the International Physical Activity Questionnaire (IPAQ; Long Form) [26]. The 27-item IPAQ-Long Form evaluates 4 domains (walking, moderate- to vigorous-intensity activity, and sitting time). Participants reported activity frequency (days/week) and duration (minutes/day), which was converted to MET-min/week using standard values (3.5 METs for walking, 4.0 for moderate, and 8.0 for vigorous activity). It has cross-cultural reliability (test-retest r=0.75) and validity across 12 countries. The Chinese version showed internal consistency (Cronbach α=0.78) and criterion validity (r=0.42 with accelerometer data, P<.01), suitable for Chinese patients post PCI [27].

Secondary Outcomes

The secondary outcomes focused on exercise endurance, self-perceived fatigue, ESE, and quality of life, which were also measured at baseline (T0), at 12 weeks of intervention (T1), and at 4 weeks of follow-up (T2), 8 weeks of follow-up (T3), and 12 weeks of follow-up (T4) weeks of follow-up.

Exercise Endurance

Exercise endurance was measured by the 6-minute walk test (6MWT), a standardized field test originally developed by Guyatt et al [28] to assess functional exercise capacity in patients with chronic respiratory or cardiac problems. Following American Thoracic Society guidelines [29], participants walked back-and-forth along a 30 m flat corridor (red turn markers); researchers instructed “fastest comfortable pace” and gave standardized encouragement every 1 minute. The total 6MWD was recorded. A portable pulse oximeter monitored peripheral oxygen saturation (SPO_2)—test paused or terminated if SPO_2<88% or severe discomfort occurred. The 6MWT correlates with maximal oxygen uptake (r=0.6-0.8) in patients with cardiac problems [29].

Self-Perceived Fatigue

Self-perceived fatigue was assessed using the Borg fatigue rating scale. According to the Borg fatigue rating scale, 6-8 points indicate very very easy, 9-10 points indicate very easy, 11-12 points indicate easy, 13-14 points indicate slight exertion, 15-16 points indicate exertion, 17-18 points indicate very exerted, and 19-20 points indicate very very exerted [30].

ESE

Measured with the Exercise Self-Efficacy Scale (ESES), developed by Resnick and Jenkins [31] for older adults and validated in CR. The 10-item scale assesses confidence in exercise tasks, such as “walking 1 km” (1=“not at all confident” to 10=“completely confident”). Total score=sum of items; higher scores=greater self-efficacy. The ESES had passed the reliability and validity test, with a Cronbach α of 0.92. The reliability coefficients of each dimension ranged from 0.72 to 0.91, indicating good reliability. The structural validity of factor analysis showed that the cumulative variance of the 4 factors was 70.19%, and the factor load of each item was 0.606-0.909, indicating good validity [31].

Quality of Life

Quality of life was assessed using the Short Form of 36 Health Survey Questionnaire (SF-36), which evaluates the influence of CHD on individuals’ physical, emotional, and social well-being. The SF-36 questionnaire has a total of 8 dimensions and 36 items. Except for physiological function and emotional function, which are answered with “yes” and “no,” the other items are scored according to 3-6 levels, and the scores of each dimension are finally converted to 0-100 points. Cronbach α coefficient of the total volume table is 0.91. The scale has been validated in Chinese populations, demonstrating good reliability and validity [32].

Usability Indicator

Usability evaluation forms an integral part of the electronic product development process. It acts as a crucial factor in ensuring the successful implementation of telemedicine programs and provides a vital means to promote user acceptance and improve compliance among users [33]. System usability was measured at 12 weeks of intervention (T1) using the SUS. The 10-item scale includes 5 positive (eg, “I would like to use this system frequently”) and 5 negative (eg, “I found the system unnecessarily complex”) statements; 5-point Likert scale (1=“strongly disagree” to 5=“strongly agree”) [34]. Scoring criteria: (1) for odd-numbered items (1, 3, 5, 7, and 9): assign a score equal to the raw item score (ie, score= raw score); (2) for even-numbered items (2, 4, 6, 8, and 10): reverse score by calculating 5 minus the raw item score (ie, score=5–raw score). Compute the total sum of all item scores, then multiply by 2.5 to obtain the final score (sum×2.5=final score; range 0-100). Higher final scores indicate better usability of the platform. A

XSL•FO
RenderX

score above 68 points is considered to be above average, a score below 68 is considered below average [23].

Safety Indicator

The incidence of AEs was used to evaluate the exercise safety of patients during the whole study period. If the patient has chest tightness, chest pain, arrhythmia, pale face, dizziness, gait instability, soft tissue injury, dislocation of bone and joint, fracture, and other conditions during exercise, the occurrence of one or more symptoms is considered as an AE, and the incidence of AEs is as follows: (Number of AEs÷total number of exercises)×100%. If there are adverse reactions, stop exercising immediately and seek help from medical personnel in time. The cause of the event will be analyzed by the research team, and all AEs will be reported to the Ethics Committee as required.

Statistical Analysis

Data analysis was performed using SPSS (version. 26.0), with 2-sided $P<.05$ considered statistically significant. An intention-to-treat approach was used. All randomly assigned participants (n=180, 90 per group) were included in the analysis, regardless of whether they completed the intervention or withdrew from follow-up. Missing data were handled via the generalized estimating equation (GEE) model, which inherently accounts for longitudinal data dependency and preserves the intention-to-treat principle. Descriptive statistics were used to summarize the participants' characteristics—continuous variables as mean (SD) or median (IQR; Shapiro-Wilk test for normality) and categorical variables as n (%). Baseline between-group comparisons used independent samples *t* test (normal continuous data), chi-square test (categorical data), or Mann-Whitney *U* test (nonnormal continuous data); intragroup nonnormal data comparisons used paired Wilcoxon rank sum test. The GEE model analyzed all primary and secondary outcomes across T0-T4, adjusting for baseline outcome values, age, sex, and education level with an exchangeable correlation structure. Given the nonnormal distribution of outcome data (confirmed via the Shapiro-Wilk test), we initially considered distributional assumptions for skewed continuous data. However, after sensitivity analysis, a Gaussian distribution with an identity link function was ultimately selected for all outcomes (PA, 6MWD, ESES, SF-36, and Borg fatigue scale). This choice was justified by two key considerations. First, the large sample size (n=180) provides robustness to violations of normality via the central limit theorem; and second, the identity link yields interpretable additive β coefficients (directly reflecting between-group mean differences), which align with clinical relevance for rehabilitation outcomes (eg, absolute changes in MET-min/week or meters). The reference category was defined as "baseline (T0)+control group" for clear effect interpretation [35]. Bonferroni correction adjusted for multiple comparisons to ensure rigor.

Ethical Considerations

This study was conducted in compliance with the Declaration of Helsinki and approved by the Hospital Institutional Review Board (approval IIT20230069B-R2). Written, fully informed consent was obtained from all participants before their enrollment, detailing the study purpose, procedures, potential risks, and benefits. Participants retained the right to withdraw from the study at any time without penalty or impact on their subsequent medical care.

All participant data were managed in strict adherence to privacy and confidentiality protocols. Identifiable personal information was deidentified during data processing and storage, and access to the dataset was restricted to authorized research personnel only.

Participants were provided with a digital health management tool as compensation for their time and participation. The compensation arrangement was fully disclosed in the informed consent document, and participants voluntarily agreed to the terms before study participation.

Results

Patient Flow and Baseline Characteristics

A total of 180 participants were randomly allocated to the intervention group (n=90) or control group (n=90). Overall, 159 (88.3%) participants completed the follow-up. The CONSORT flow diagram is presented in Figure 4. As illustrated in Table 1, no statistically significant differences were observed between the intervention group and control group in all baseline characteristics ($P>.05$), including sociodemographic variables (age: mean 63.86, SD 9.17 y vs mean 62.68, SD 8.58 y; $P=.21$; and sex (male): 73/90, 81.1% vs 67/90, 74.4%; $P=.28$), clinical indicators (cardiac function grade I: 46/90, 51.1% vs 38/90, 42.2%; $P=.23$), and baseline values of outcome measures (6MWD: median 383.50, IQR 370.00-405.00 m vs median 380.00, IQR 373.00-386.00 m; $P=.16$; and PA: median 693.00, IQR 396.00-1031.25 MET-min/week vs median 648.00, IQR 495.00-1041.75 MET-min/week; $P=.60$). Consistent with the CONSORT guidelines, this baseline balance indicates that the randomization process was effective, and any subsequent differences in outcome measures between the 2 groups can be attributed to the intervention rather than preexisting group disparities.

XSL•FO
RenderX

Figure 4. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of the randomized controlled trial involving patients with coronary heart disease after percutaneous coronary intervention over the 24-week study period (12-week intervention+12-week follow-up) in Hangzhou, China. Analysis was conducted using an intention-to-treat approach, including all 180 randomly assigned participants regardless of intervention completion or follow-up status. Missing data were handled via the generalized estimating equation model.

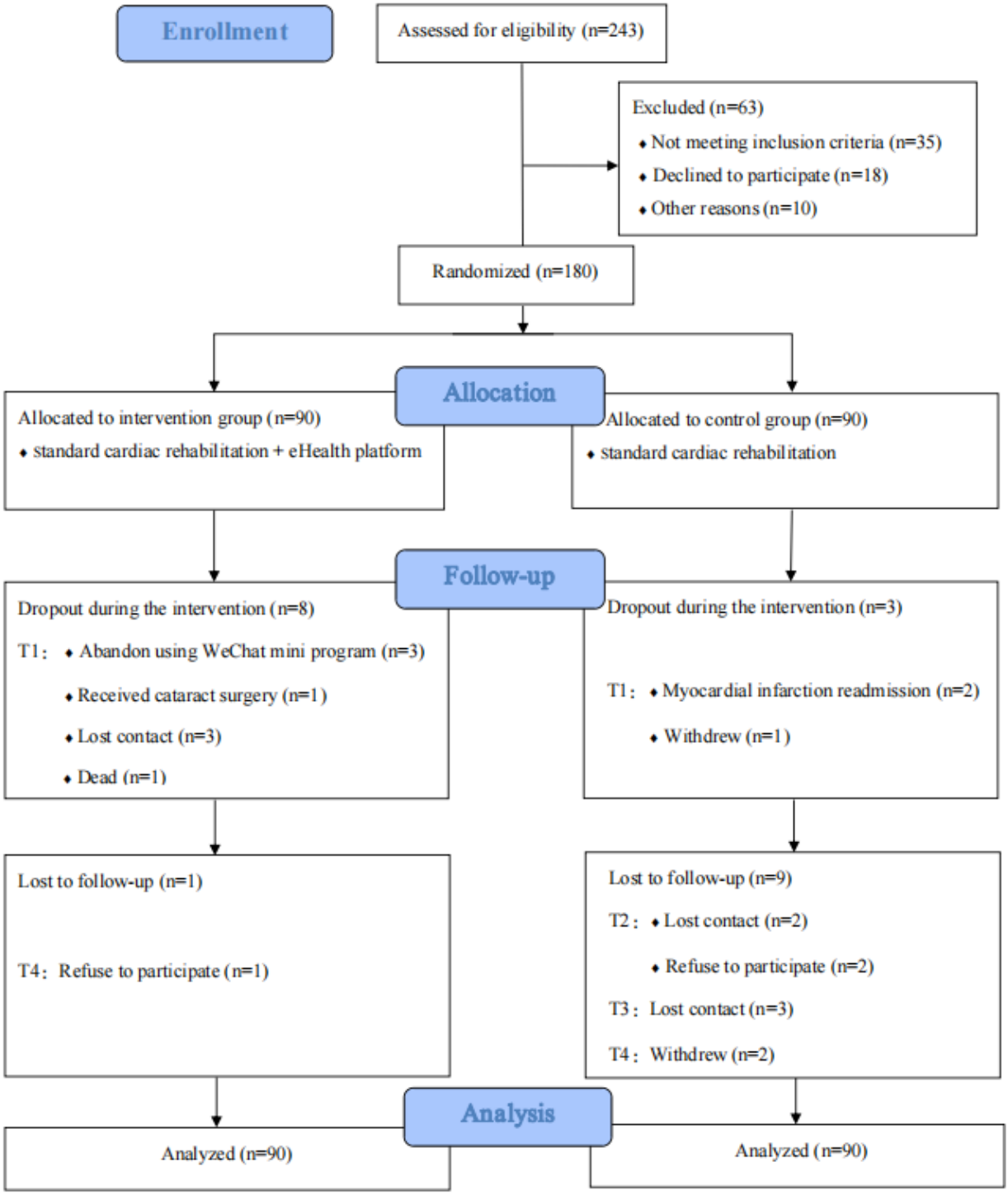

XSL•FO
RenderX

Table 1. Baseline demographic, clinical characteristics, and outcome variables of patients with coronary heart disease after percutaneous coronary intervention in the intervention and control groups (Hangzhou, China).

Variable	Control group (n=90)	Intervention group (n=90)	*t* test (*df*), Chi-square (*df*), or *z* score	*P* value
Age (y), mean (SD)	63.86 (9.17)	62.68 (8.58)	–1.266 (157)[a]	.21[b]
BMI (kg/m^2), mean (SD)	23.85 (2.37)	24.35 (2.07)	–1.425 (157)[a]	.16[b]
Sex, n (%)			1.2 (1)[c]	.28[d]
Male	73 (81.1)	67 (74.4)		
Female	17 (18.9)	23 (25.6)		
Marital status, n (%)			—[e]	.12[f]
Married	88 (97.8)	87 (96.7)		
Unmarried	2 (2.2)	0 (0)		
Widowed	0 (0)	3 (3.3)		
Residence, n (%)			2.3 (1)[c]	.13[d]
Town	31 (34.4)	41 (45.6)		
Countryside	59 (65.6)	49 (54.4)		
Education level, n (%)			0.5 (1)[c]	.98[d]
Primary school or below	40 (44.4)	38 (42.2)		
Junior high school	28 (31.1)	30 (33.3)		
High school or technical secondary school	9 (10)	11 (12.2)		
Junior college	7 (7.8)	6 (6.7)		
Undergraduate or above	6 (6.7)	5 (5.6)		
Professional types, n (%)			3.7 (4)[c]	.46[d]
Enterprises or institutions	7 (7.8)	2 (2.2)		
Farmer or worker	36 (40)	38 (42.2)		
Freelance	9 (10)	13 (14.4)		
Retired	32 (35.5)	32 (35.6)		
Unemployed	6 (6.7)	5 (5.6)		
Health insurance, n (%)			—	.44[f]
Medical insurance for employees	23 (25.6)	25 (27.8)		
Medical insurance for urban and rural residents	66 (73.3)	61 (67.8)		
Other payment methods	1 (1.1)	4 (4.4)		
Duration of illness (y), n (%)			0.8 (1)[c]	.93[d]
≤1	54 (60)	55 (61.1)		
1-3	19 (21.1)	18 (20)		
3-5	4 (4.4)	6 (6.7)		
5-10	7 (7.8)	7 (7.8)		
>10	6 (6.7)	4 (4.4)		
Monthly income (¥; ¥1=US $0.14), n (%)			1.6 (2)[c]	.46[d]
≤3000	22 (24.4)	20 (22.2)		

XSL•FO
RenderX

Variable	Control group (n=90)	Intervention group (n=90)	*t* test (*df*), Chi-square (*df*), or *z* score	*P* value
3001-4999	38 (42.2)	32 (35.6)		
≥5000	30 (33.3)	38 (42.2)		
Smoking , n (%)			2.0 (2)[c]	.37[d]
Yes	25 (27.8)	22 (24.4)		
Ever	48 (53.3)	43 (47.8)		
No	17 (18.9)	25 (27.8)		
Drinking , n (%)			2.4 (2)[c]	.30[d]
Yes	8 (8.9)	15 (16.7)		
Ever	59 (65.6)	54 (60)		
No	23 (25.6)	21 (23.3)		
Comorbidities, n (%)			6.2 (4)[c]	.18[d]
Hypertension	64 (71.1)	62 (68.9)		
Diabetes	31 (34.4)	33 (36.7)		
Dyslipidemia	5 (5.6)	17 (18.9)		
Others	67 (74.4)	71 (78.9)		
No	5 (5.6)	7 (7.8)		
Blocked vessels, n (%)			0.4 (1)[c]	.54[d]
Single	13 (14.4)	16 (17.8)		
Multiple	77 (85.6)	74 (82.2)		
Cardiac function grade, n (%)			1.4 (1)[c]	.23[d]
I	46 (51.1)	38 (42.2)		
II	44 (48.9)	52 (57.8)		
6MWD[g], median (IQR)	383.50 (370.00-405.00)	380.00 (373.00-386.00)	–1.412[h]	.16[i]
Borg fatigue rating, median (IQR)	12.00 (10.00-12.00)	12.00 (10.75-13.00)	–1.376[h]	.17[i]
ESES[j], median (IQR)	25.00 (22.22-27.78)	25.00 (22.22-31.25)	–0.173[h]	.86[i]
PA[k], median (IQR)	693.00 (396.00-1031.25)	648.00 (495.00-1041.75)	–0.522[h]	.60[i]
SF-36[l], median (IQR)	57.13 (53.44-67.09)	60.09 (50.94-66.72)	–0.963[h]	.34[i]

[a]*t* test.
[b]Independent samples *t* test.
[c]Chi-square.
[d]Pearson chi-square.
[e]Not applicable.
[f]Fisher exact test.
[g]6MWD: 6-minute walk distance.
[h]*z* score.
[i]Mann-Whitney *U* test.
[j]ESES: Exercise Self-Efficacy Scale.
[k]PA: physical activity.
[l]SF-36: Short Form of 36 Health Survey Questionnaire.

XSL•FO
RenderX

Primary Outcome

The primary outcome, PA level, was demonstrated in Figure 5 and Table 2. For the primary outcome (PA level), GEE analysis was performed with baseline (T0)+control group as the reference category to clarify group differences and time-dependent intervention effects. The model adjusted for baseline PA level and accounted for the correlation of repeated measurements within participants using an exchangeable correlation structure. Baseline PA levels were balanced between groups in unadjusted analysis (Table 1; *P*=.60), confirming effective randomization. The GEE model's "group (baseline, adjusted)" effect (β coefficient=48.254; *P*=.002) reflects the residual group difference after adjusting for age, sex, and education level—not an inherent baseline imbalance. This adjusted effect does not undermine randomization validity but enhances the model's precision by accounting for potential confounding. The control group showed a modest natural increase in PA over time (eg, T4 vs T0: β coefficient=125.358, 95% CI 101.798-148.918; *P*<.001). In contrast, the intervention group exhibited significant additional improvements at all postbaseline time points, with the strongest effect at T4 (group×time interaction: β coefficient=937.288, 95% CI 867.609-1006.967; *P*<.001). The consistent significance of group×time interactions (all *P*<.001) confirmed that the intervention's PA-enhancing effect was maintained throughout the study period.

Figure 5. Changes in physical activity levels (MET-min/week) over the 24-week study period in patients with coronary heart disease after percutaneous coronary intervention.

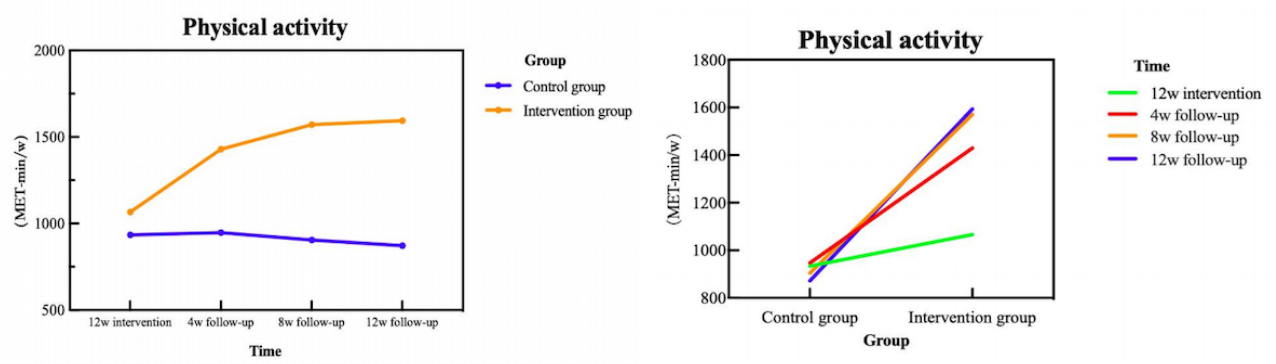

Table 2. Generalized estimating equation analysis for physical activity levels (MET-min/week) over the 24-week study period in patients with coronary heart disease after percutaneous coronary intervention.

Parameter category[a] (main effects) and parameter description	β coefficient (95% CI)	SE	*P* value
Time (control group)			
T1 (12-week intervention) vs T0	23.521 (–0.149 to 47.191)	11.683	.05
T2 (4-week follow-up) vs T0	46.857 (23.819 to 69.895)	11.748	<.001
T3 (8-week follow-up) vs T0	31.523 (8.412 to 54.634)	11.809	.008
T4 (12-week follow-up) vs T0	125.358 (101.798 to 148.918)	12.032	<.001
Group (baseline, adjusted)			
Intervention Group vs Control Group (T0, adjusted for covariates)	48.254 (18.021 to 78.487)	15.389	.002
Group×time interaction[b] (intervention additional effect)			
T1 (12-week intervention)	861.109 (783.934 to 938.284)	39.376	<.001
T2 (4-week follow-up)	899.386 (826.259 to 972.514)	37.311	<.001
T3 (8-week follow-up)	879.404 (811.213 to 947.596)	34.792	<.001
T4 (12-week follow-up)	937.288 (867.609 to 1006.967)	35.551	<.001

[a]Reference category: baseline (T0)+control group.

[b]β coefficients for group×time interactions represent the additional intervention effect (beyond the control group's natural change) in the intervention group at each time point relative to baseline.

The GEE model used a Gaussian distribution with an identity link function, yielding additive β coefficients (absolute between-group differences). "Group (baseline, adjusted)" represents the adjusted group difference in PA at T0 after accounting for covariates (age, sex, and education level). This adjusted effect does not contradict the unadjusted baseline balance (Table 1, *P*=.60), as it reflects the residual group difference after covariate adjustment (a common observation in longitudinal models) and confirms no failure of randomization.

Secondary Outcomes

Secondary outcomes are illustrated in Table 3. GEE analysis confirmed significant group-time interaction effects (all *P*<.001).

XSL•FO
RenderX

Compared with the control group, the intervention group had significantly longer 6MWD at all time points (eg, T4: β coefficient=45.278, 95% CI 41.084-49.472; *P*<.001), lower self-perceived fatigue from T2 onwards (T4: β coefficient=–1.058, 95% CI –1.243 to –0.873; *P*<.001), and higher ESE (T4: β coefficient=31.015, 95% CI 29.019-33.010; *P*<.001) and quality of life (T4: β coefficient=6.857, 95% CI 6.171-7.543; *P*<.001) across all postbaseline assessments. No significant difference in self-perceived fatigue was observed at T1 (β coefficient=0.062, 95% CI –0.116 to 0.240; *P*=.50).

Table 3. Effects of the persuasive systems design–based exercise rehabilitation platform on secondary outcomes in patients with coronary heart disease after percutaneous coronary intervention over the 24-week study period.

Outcome measures	Control group, median (IQR)	Intervention group, median (IQR)	GEE[a] β coefficient[b] (95% CI)	*P* value
6MWD[c]				
12 weeks of intervention	410.00 (389.00-420.00)	435.00 (423.00-448.00)	33.760 (29.311 to 38.208)	<.001
4 weeks of follow-up	413.00 (395.00-421.00)	440.00 (430.00-448.00)	37.577 (33.346 to 41.808	<.001
8 weeks of follow-up	406.00 (385.25-416.00)	439.50 (431.50-448.25)	44.495 (40.334 to 48.656)	<.001
12 weeks of follow-up	405.00 (385.00-413.50)	436.00 (429.00-449.00)	45.278 (41.084 to 49.472)	<.001
Borg fatigue rating				
12 weeks of intervention	10.00 (9.00-11.00)	10.00 (10.00-11.00)	0.062 (–0.116 to 0.240)	0.50
4 weeks of follow-up	10.00 (9.00-10.00)	9.50 (9.00-10.00)	–0.331 (–0.512 to –0.150)	<.001
8 weeks of follow-up	10.00 (9.00-10.00)	9.00 (9.00-9.00)	–0.707 (–0.881 to –0.533)	<.001
12 weeks of follow-up	10.00 (9.00-10.00)	9.00 (8.00-9.00)	–1.058 (–1.243 to –0.873)	<.001
ESES[d]				
12 weeks of intervention	33.33 (27.78-33.33)	58.33 (50.00-66.67)	23.355 (20.865 to 25.845)	<.001
4 weeks of follow-up	30.56 (25.00-30.56)	61.11 (52.78-66.67)	28.865 (26.534 to 31.195)	<.001
8 weeks of follow-up	30.56 (25.00-33.33)	61.11 (54.87-66.67)	28.012 (25.965 to 30.060)	<.001
12 weeks of follow-up	27.78 (25.00-30.56)	61.11 (54.17-66.67)	31.015 (29.019 to 33.010)	<.001
SF-36[e]				
12 weeks of intervention	80.19 (77.58-81.88)	89.44 (87.84-90.69)	9.520 (8.732 to 10.308)	<.001
4 weeks of follow-up	81.13 (78.71-83.00)	90.56 (89.31-91.31)	9.443 (8.733 to 10.153)	<.001
8 weeks of follow-up	82.56 (80.75-85.05)	90.88 (89.94-91.55)	8.138 (7.390 to 8.885)	<.001
12 weeks of follow-up	84.13 (82.38-86.63)	91.19 (90.41-91.81)	6.857 (6.171 to 7.543)	<.001

[a]GEE: generalized estimating equation.

[b]β coefficients represent the estimated difference in outcome values between the intervention and control groups (positive β=higher outcome in the intervention group and negative β=lower outcome in the intervention group). All generalized estimating equation models were adjusted for baseline values of the respective outcome and sociodemographic covariates (age, sex, and education level).

[c]6MWD: 6-minute walk distance.

[d]ESES: Exercise Self-efficacy Scale.

[e]SF-36: Short Form of 36 Health Survey Questionnaire.

Usability Indicator

The average SUS score of the WeChat mini program was 85.72 (SD 4.26) points (the highest score was 92 and the lowest score was 74.6, n=82), and its usability score was higher than the average level (≥68), indicating that the usability, ease of use, and learnability of the system were acceptable. The usability scores for the 10 SUS items and the average ratings for all items are summarized in Figure 6. Lighter shades of green indicate negative responses to the items (eg, “strongly disagree” for negative items and “strongly agree” for positive items), while darker shades of green indicate positive responses to the items (eg, “strongly agree” for negative items and “strongly disagree” for positive items).

XSL•FO
RenderX

Figure 6. Summary plot of the System Usability Scale scores for the persuasive systems design–based WeChat mini program in patients with coronary heart disease after percutaneous coronary intervention (12-week intervention period). SUS: System Usability Scale.

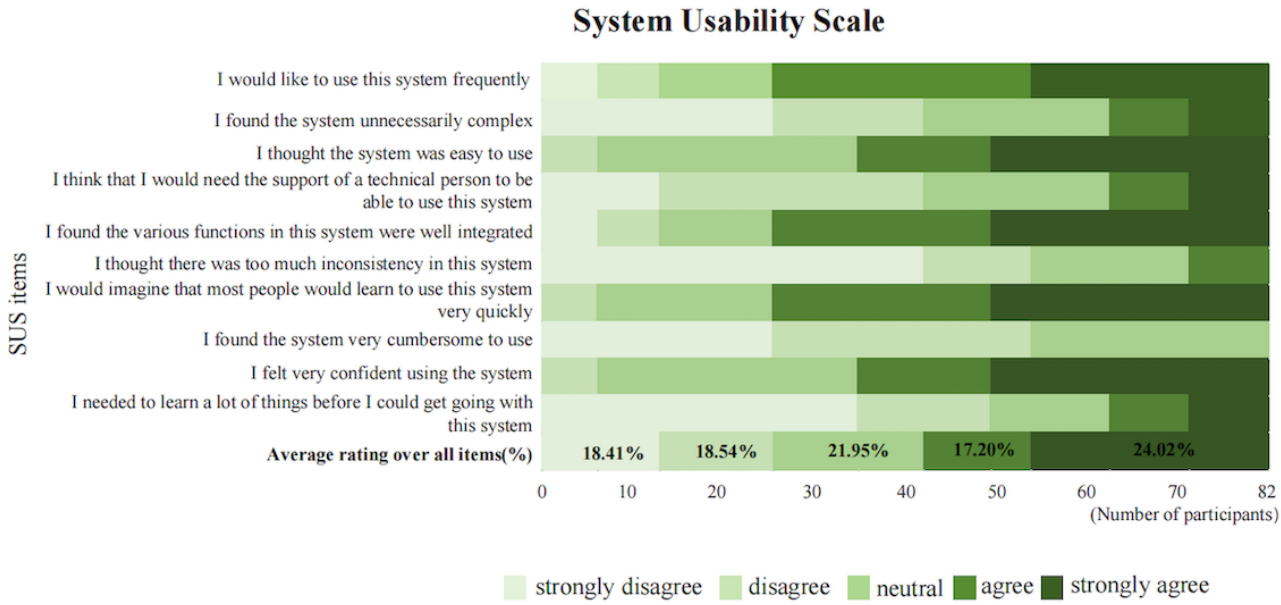

Safety Indicator

AEs during the study period were recorded comprehensively. As illustrated in Figure 4, three serious AEs were reported—2 cases of myocardial infarction readmission in the control group and 1 death in the intervention group. The 2 readmissions were unrelated to standard CR, and the death (acute cerebral hemorrhage in an 82-year-old male with hypertension and diabetes) was attributed to underlying comorbidities, not the exercise intervention. No other AEs (eg, chest tightness and musculoskeletal injury) occurred in either group. Mild muscle soreness in the early intervention stage resolved spontaneously. All serious AEs were documented, reported to the Ethics Committee, and managed per protocol.

Discussion

Principal Findings

This study confirmed that the PSD model–based mHealth platform provides added value beyond standard CR: both groups received consistent usual care, but the intervention group showed greater improvements in PA, exercise endurance, and ESE.

A key strength is the consistent delivery of standard care across both groups, which eliminated confounding from unequal usual care and clearly isolated the effect of the mHealth intervention. Our exclusive use of the PSD model addresses a key gap in CR eHealth studies. The platform's strength lies in strict alignment between PSD principles and every functional module, rather than combining multiple theories such as social cognitive theory, self-efficacy theory, theory of planned behavior, and control theory [17]. For example, our platform embeds persuasive principles within intervention designs, ensuring theoretical consistency. This focused approach avoids the "theoretical fragmentation" common in multitheory interventions. Globally, RCR interventions primarily fall into 3 categories. Basic telemonitoring systems rely on wearable devices (eg, Google Fitbit or Apple Watch) to track steps and heart rate. These systems provide limited active guidance, focusing more on data collection than behavior change [36]. App-based educational tools provide static exercise videos and reminders, but they lack personalization and include fixed exercise plans regardless of the patient's functional capacity [37]. Telehealth with human coaching combines weekly video consultations with physiotherapists and basic app tracking, but it is resource-intensive, limiting scalability [38]. In contrast, our intervention integrates PSD model–driven persuasive design with dynamic personalization—a gap in most RCR tools. Unlike basic telemonitoring or static apps, our WeChat mini program uses PSD-derived BCTs (self-monitoring via exercise diaries, social support via exercise ranking, reminders, and praise feedback) to actively motivate adherence.

The findings of PA are consistent with a previous study by Sankaran et al [39], which selected personalized and persuasive design principles to develop the HeartHab, a smartphone-based mobile app (developed by Hasselt University, Diepenbeek, Belgium) for telerehabilitation, by incorporating novel persuasive techniques to motivate patients with CHD to reach personalized PA targets. Compared with conventional rehabilitation, the effect of telerehabilitation treatment decreased at a lower rate after completion. A systematic review by Aldenaini et al [40] evaluated the effectiveness of persuasive techniques used to promote PA and reduce sedentary behavior and found that mobile and handheld devices were the most commonly used platforms, while the persuasive strategies most commonly used to achieve desired behavioral outcomes in the PA domain were self-monitoring and reminder. Salvi et al [41] found that by using the HeartCycle mobile medical system (a remote cardiovascular monitoring system developed by the research team of the University of Coimbra, Coimbra, Portugal; under a European Union–funded project coordinated by Philips, with participation of 17 partners including communication enterprises, hospitals, and universities, eg, Jessa Hospital,

XSL•FO
RenderX

Hasselt, Belgium) to guide exercise, taking into account the design principles of the persuasive system and incorporating the principles of reward, reminder, suggestion, and rehearsal into the implementation process of the system, the exercise monitoring, guidance, and feedback are provided to patients to motivate them to persist in completing the rehabilitation plan, and the exercise time of patients is significantly increased. Exercise habits also improved after 6 months of follow-up, with high user acceptance and perceived usefulness. Psychological research shows that effective implementation of the behavior change intervention helps habit formation and helps for a long time to maintain motivation [42-44]. Moreover, when it comes to technology-supported interventions, it is important to design user-friendly and accessible systems. By considering the needs and perspectives of the user to enhance the user experience, human-computer interaction research has mainly focused on usability and interaction techniques when designing technology-supported rehabilitation systems. Consider a patient's specific needs and ideas to help adjust these interventions based on technology, to accurately meet the target user, and to minimize the loss [42,45].

The 6MWT is a safe and effective tool to test exercise endurance, which has been widely used in clinical trials and the outcome measurement of cardiopulmonary rehabilitation [46]. The 6MWD was the main outcome index of the 6MWT. The study by Taylor et al [47] proposed that a within-group 6MWD improvement of at least 54 m constitutes a minimum clinically important difference (MCID) for patients with cardiac disease, reflecting meaningful functional recovery. It is important to distinguish this within-group change from the net intervention effect (between-group difference) reported in our study. For the intervention group, the within-group improvement in 6MWD from baseline (T0: 380.00 m) to T4 (436.00 m) was 56.00 m, exceeding the 54 m MCID threshold and indicating clinically meaningful functional gain for individual patients; the GEE β coefficients in Table 3 represent the net intervention effect (additional benefit of the PSD-based platform beyond the control group's natural recovery), which ranged from 33.76 m (T1) to 45.28 m (T4). While this net effect did not reach the 54 m MCID for between-group comparisons, it remains clinically relevant for several reasons. First, the control group also exhibited natural recovery (within-group improvement of 21.50 m from T0 to T4), and the intervention group's net gain still contributed to meaningful functional improvement at the individual level. Second, previous studies in CR have noted that even modest 6MWD improvements (30-45 m) are associated with better long-term prognosis [48,49]. Third, the intervention's net effect was maintained across follow-up, indicating durable functional benefits. A systematic review found that the 6MWT could be used to determine clinical response to cardiac interventions and is likely a good marker of prognosis, and demonstrated an inverse relationship between 6MWD and the likelihood of a major adverse cardiac event (MACE) [48]. Moreover, a study demonstrated that each 50 m increase in 6MWT distance was associated with a lower odds ratio and hazard ratio for the MACE. The 6MWT has a prognostic value for predicting MACE in patients with prevalent CVDs [49]. Conversely, another researcher [50] reported that 6MWT may not be sensitive enough, unable to monitor early intervention for patients with normal functions. They indicated that familiarity with walking routes or a good walking pace could lead to an artificially increased walking distance. They suggested that a familiar experiment should be carried out, the second test to establish a baseline value. In future studies, we will consider multiple repeated measurements in more populations and incorporate familiarization tests into the 6MWT to further improve the accuracy of assessment and testing and to present more precise clinical value and significance.

A study by Cavalheri et al [51] has pointed out that fatigue is one of the most frequently cited barriers for patients to adhere to exercise programs, and that exercises or tools to alleviate fatigue symptoms are very important for patients and health care workers and should be the target of exercise interventions. Exercise interventions in populations, such as lung transplantation [52] and rheumatoid arthritis [53], have also been found to have a good effect on improving the degree of fatigue in patients. However, in a systematic review of the effects of exercise rehabilitation or exercise rehabilitation on health-related quality of life and fatigue in patients with lung cancer [54], none of the studies reported statistically significant reductions in fatigue. This may be due to the large heterogeneity of the physical exercise program, the short duration of the intervention in some studies, and the poor treatment effect in most studies. This may explain why self-perceived fatigue in this study did not improve significantly during the intervention period but decreased significantly during the follow-up period.

The exercise rehabilitation platform based on the PSD model significantly improves the ESE of patients after PCI during the intervention period and the follow-up period, and has a good maintenance effect. This is consistent with the intervention effect of ESE of patients in the sedentary behavior intervention program based on the behavior change wheel theory [55]. In the intervention process, the research participants are guided to actively participate in the change of sedentary behavior, and self-monitoring is carried out, so that their real sense is affected by the positive change after the interruption of sedentary behavior, and the self-efficacy, participation intention, and compliance of sedentary patients are improved. Alhasani et al [22] found that self-monitoring to support self-management of health problems related to the application is crucial; it enhances the user's insight into their health, inspires them to seek medical help when necessary, and enhances patient confidence in the disease intervention process. In addition, the persuasive principle of praise is also selected, using incentives as a way of feedback to users. By evaluating the completion of exercise tasks, patients are sent praise messages, and certain points are rewarded, which will also strengthen the action of patients to take health behavior changes to a certain extent. Success depends on a digital self-management system related to health-promoting factors, including self-adjusting strategies, such as cognitive and emotional adjustment, goal setting, effective response, and problem-solving skills, and enhanced confidence in self-adjusting [56]. With time, changing the established behaviors and habits may be challenging, but in the future, eHealth interventions, combining health behaviors with persuasive strategies, may contribute to a sustainable behavior change in health management.

XSL•FO
RenderX

In Europe, especially Nordic countries, physiotherapists have long been core leaders in CR—driving clinical practice, program innovation, and research as key members of multidisciplinary teams, a role rooted in regional health care traditions [57]. In China, physiotherapists' involvement in CR is evolving, but is more focused on clinical execution rather than leadership or research. This is partly due to a shortage of specialized rehabilitation professionals and underdeveloped training systems, with CR programs still predominantly being physician-led [58]. This difference reflects varied health care models. Nordic systems prioritize physiotherapists as CR innovators, while China's developing CR infrastructure currently positions them as implementers—highlighting a need for workforce development to fully leverage their expertise.

Limitations

This study has limitations that should be considered. First, participants were recruited from a single tertiary hospital in Hangzhou, which may limit the generalizability of findings to patients from primary or secondary hospitals or rural areas. Second, PA was assessed by the self-reported IPAQ, and while outcome assessors were blinded to group assignments, participants knew their group allocation due to the unblinded exercise intervention. This awareness may have influenced self-reporting (eg, overestimation or underestimation of activity), a bias that cannot be fully eliminated by only blinding data collectors [59-61]. Third, the study design compared "standard CR + eHealth platform" versus "standard CR alone" and thus cannot evaluate the standalone effectiveness of the eHealth platform. Fourth, the GEE model assumes that missing data are missing at random. However, some dropouts in the control group were due to myocardial infarction (a serious AE related to the underlying disease), which constitutes informative censoring (missing not at random). This assumption may introduce bias, as the reasons for dropout are associated with the outcome of interest (cardiac function and PA). Future studies could use sensitivity analyses (eg, inverse probability weighting) or alternative statistical models (eg, joint models for longitudinal data and survival) to account for informative censoring and enhance the robustness of the results. Other limitations include the application environment, cost budget, time, and the ability of the technical team [62], which also become the obstacle factors for the application of the WeChat mini program of exercise rehabilitation to a certain extent.

Future Work

Future research directions should address these limitations. Priority could be given to direct comparative studies between center-based CR and the PSD model–based eHealth platform—a design that would clarify whether the eHealth tool can serve as a viable alternative for patients unable to attend in-person rehabilitation. Future studies could use objective tools (eg, accelerometers) to improve outcome reliability. Additionally, multicenter trials with larger sample sizes are needed to validate the effectiveness of the platform across diverse patient populations. Long-term follow-up should also be incorporated to assess the sustainability of outcomes, such as exercise adherence and quality of life.

Conclusions

The findings of this study demonstrate the effectiveness of the eHealth CR platform based on the PSD model in improving key rehabilitation outcomes for patients after PCI. Specifically, when provided in addition to standard CR, the PSD model–based eHealth platform further enhanced patients' PA level, exercise endurance, ESE, and quality of life, while also reducing self-perceived fatigue during the follow-up period. These findings provide practical insights for optimizing CR delivery in clinical settings. For health care providers, integrating PSD model–based eHealth tools into existing CR programs may serve as a feasible strategy to address suboptimal adherence and enhance long-term rehabilitation outcomes. For patients, the platform offers a flexible, accessible means to reinforce in-person rehabilitation guidance, particularly during postintervention follow-up.

Funding

This work was supported by a grant from the General Public Welfare Projects of the Science and Technology Department of Zhejiang Province, China (LGF20G030009).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author (LH) on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Multi-dimensional Considerations in Exercise Prescription Library Construction.
[DOCX File , 13 KB - jmir_v28i1e71450_app1.docx]

Multimedia Appendix 2
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1253 KB - jmir_v28i1e71450_app2.pdf]

XSL•FO
RenderX

References

1. Cardiovascular diseases (CVDs). World Health Organization. 2025. URL: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [accessed 2025-10-08]
2. Liu MB, He XY, Yang XH, Wang ZW. Interpretation of Annual Report on Cardiovascular Health and Diseases in China 2024. Biomed Environ Sci 2025 Aug 20;38(8):893-917 [FREE Full text] [doi: 10.3967/bes2025.099] [Medline: 40928269]
3. China Health Statistics Yearbook 2021. 2021. URL: http://www.tjnjw.com/hangyefb/w/zhongguo-weishengjiankang-tongjinianjian-2022.html [accessed 2025-12-09]
4. Myers J, Prakash M, Froelicher V, Do D, Partington S, Atwood JE. Exercise capacity and mortality among men referred for exercise testing. N Engl J Med 2002;346(11):793-801. [doi: 10.1056/NEJMoa011858] [Medline: 11893790]
5. Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C, Catapano AL, ESC Scientific Document Group. 2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts) Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). Eur Heart J 2016;37(29):2315-2381 [FREE Full text] [doi: 10.1093/eurheartj/ehw106] [Medline: 27222591]
6. Howarter AD, Bennett KK, Barber CE, Gessner SN, Clark JMR. Exercise self-efficacy and symptoms of depression after cardiac rehabilitation: predicting changes over time using a piecewise growth curve analysis. J Cardiovasc Nurs 2014;29(2):168-177. [doi: 10.1097/JCN.0b013e318282c8d6] [Medline: 23416936]
7. Fletcher GF, Ades PA, Kligfield P, Arena R, Balady GJ, Bittner VA, American Heart Association Exercise, Cardiac Rehabilitation,Prevention Committee of the Council on Clinical Cardiology, Council on Nutrition, Physical ActivityMetabolism, Council on CardiovascularStroke Nursing,Council on EpidemiologyPrevention. Exercise standards for testing and training: a scientific statement from the American Heart Association. Circulation 2013;128(8):873-934. [doi: 10.1161/CIR.0b013e31829b5b44] [Medline: 23877260]
8. Dibben GO, Faulkner J, Oldridge N, Rees K, Thompson DR, Zwisler A, et al. Exercise-based cardiac rehabilitation for coronary heart disease: a meta-analysis. Eur Heart J 2023;44(6):452-469 [FREE Full text] [doi: 10.1093/eurheartj/ehac747] [Medline: 36746187]
9. Gomes-Neto M, Durães AR, Conceição LSR, Saquetto MB, Alves IG, Smart NA, et al. Some types of exercise interventions are more effective than others in people with coronary heart disease: systematic review and network meta-analysis. J Physiother 2024;70(2):106-114 [FREE Full text] [doi: 10.1016/j.jphys.2024.02.018] [Medline: 38503676]
10. Kotseva K, Rydén L, De Backer G, De Bacquer D, Wood D. EURObservational research programme: EUROASPIRE. Eur Heart J 2015;36(16):950-951 [FREE Full text] [doi: 10.1093/eurheartj/ehv047] [Medline: 25899410]
11. Wongvibulsin S, Habeos EE, Huynh PP, Xun H, Shan R, Porosnicu Rodriguez KA, et al. Digital health interventions for cardiac rehabilitation: systematic literature review. J Med Internet Res 2021;23(2):e18773 [FREE Full text] [doi: 10.2196/18773] [Medline: 33555259]
12. Thakker R, Khan M, Al-Hemyari B. Cardiac rehabilitation after hospitalization for acute coronary syndrome. Curr Cardiol Rep 2023;25(12):1699-1703. [doi: 10.1007/s11886-023-02010-5] [Medline: 38063996]
13. Masterson Creber R, Dodson JA, Bidwell J, Breathett K, Lyles C, Harmon Still C, American Heart Association Cardiovascular Disease in Older Populations Committee of the Council on Clinical Cardiologythe Council on CardiovascularStroke Nursing; Council on Quality of CareOutcomes Research;Council on Peripheral Vascular Disease. Telehealth and health equity in older adults with heart failure: a scientific statement from the American Heart Association. Circ Cardiovasc Qual Outcomes 2023;16(11):e000123 [FREE Full text] [doi: 10.1161/HCQ.0000000000000123] [Medline: 37909212]
14. Bairey Merz CN, Alberts MJ, Balady GJ, Ballantyne CM, Berra K, Black HR, American Academy of Neurology, American Association of CardiovascularPulmonary Rehabilitation, American College of Preventive Medicine, American College of Sports Medicine, American Diabetes Association, American Society of Hypertension, Association of Black Cardiologists, Centers for Disease ControlPrevention, National Heart, Lung,Blood Institute, National Lipid Association, Preventive Cardiovascular Nurses Association. ACCF/AHA/ACP 2009 competence and training statement: a curriculum on prevention of cardiovascular disease: a report of the American College of Cardiology Foundation/American Heart Association/American College of Physicians Task Force on Competence and Training (Writing Committee to Develop a Competence and Training Statement on Prevention of Cardiovascular Disease): developed in collaboration with the American Academy of Neurology; American Association of Cardiovascular and Pulmonary Rehabilitation; American College of Preventive Medicine; American College of Sports Medicine; American Diabetes Association; American Society of Hypertension; Association of Black Cardiologists; Centers for Disease Control and Prevention; National Heart, Lung, And Blood Institute; National Lipid Association; and Preventive Cardiovascular Nurses Association. Circulation 2009;120(13):e100-e126. [doi: 10.1161/CIRCULATIONAHA.109.192640] [Medline: 19770387]
15. Duff OM, Walsh DM, Furlong BA, O'Connor NE, Moran KA, Woods CB. Behavior change techniques in physical activity eHealth interventions for people with cardiovascular disease: systematic review. J Med Internet Res 2017;19(8):e281 [FREE Full text] [doi: 10.2196/jmir.7782] [Medline: 28768610]
16. Johnston M, Carey RN, Connell Bohlen LE, Johnston DW, Rothman AJ, de Bruin M, et al. Development of an online tool for linking behavior change techniques and mechanisms of action based on triangulation of findings from literature synthesis

XSL•FO
RenderX

and expert consensus. Transl Behav Med 2021;11(5):1049-1065 [FREE Full text] [doi: 10.1093/tbm/ibaa050] [Medline: 32749460]
17. Agyei EEYF, Ekpezu A, Oinas-Kukkonen H. Persuasive systems design trends in coronary heart disease management: scoping review of randomized controlled trials. JMIR Cardio 2024;8:e49515 [FREE Full text] [doi: 10.2196/49515] [Medline: 38896840]
18. Chatterjee S, Price A. Healthy living with persuasive technologies: framework, issues, and challenges. J Am Med Inform Assoc 2009;16(2):171-178 [FREE Full text] [doi: 10.1197/jamia.M2859] [Medline: 19074300]
19. Oyibo K, Morita PP. Designing better exposure notification apps: the role of persuasive design. JMIR Public Health Surveill 2021;7(11):e28956 [FREE Full text] [doi: 10.2196/28956] [Medline: 34783673]
20. Hayes Watson C, Nuss H, Celestin M, Tseng TS, Parada N, Yu Q, et al. Health beliefs associated with poor disease self-management in smokers with asthma and/or COPD: a pilot study. J Asthma 2019;56(9):1008-1015. [doi: 10.1080/02770903.2018.1509990] [Medline: 30285498]
21. Authors/Task Force Members:, McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, ESC Scientific Document Group. 2021 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure: developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). With the special contribution of the heart failure association (HFA) of the ESC. Eur J Heart Fail 2022;24(1):4-131 [FREE Full text] [doi: 10.1002/ejhf.2333] [Medline: 35083827]
22. Alhasani M, Mulchandani D, Oyebode O, Baghaei N, Orji R. A systematic and comparative review of behavior change strategies in stress management apps: opportunities for improvement. Front Public Health 2022;10:777567 [FREE Full text] [doi: 10.3389/fpubh.2022.777567] [Medline: 35284368]
23. Kortum P, Acemyan CZ, Oswald FL. Is it time to go positive? Assessing the positively worded System Usability Scale (SUS). Hum Factors 2021;63(6):987-998. [doi: 10.1177/0018720819881556] [Medline: 31913715]
24. He J, Jia J, Chen X, Dong A, Ding Y, Liao W. Expert consensus on full-cycle rehabilitation of elderly patients with chronic heart failure. Journal of Rehabilitation 2025;35(02):112-123 [FREE Full text] [doi: 10.3724/sp.j.1329.2025.02002]
25. Widmer RJ, Senecal C, Allison TG, Lopez-Jimenez F, Lerman LO, Lerman A. Dose-response effect of a digital health intervention during cardiac rehabilitation: subanalysis of randomized controlled trial. J Med Internet Res 2020;22(2):e13055 [FREE Full text] [doi: 10.2196/13055] [Medline: 32130116]
26. Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, Ainsworth BE, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003;35(8):1381-1395. [doi: 10.1249/01.MSS.0000078924.61453.FB] [Medline: 12900694]
27. Qu N, Li J. Study on the reliability and validity of international physical activity questionnaire (Chinese Vision, IPAQ). Zhonghua Liu Xing Bing Xue Za Zhi 2004 Mar;25(3):265-268. [Medline: 15200945]
28. Guyatt GH, Sullivan MJ, Thompson PJ, Fallen EL, Pugsley SO, Taylor DW, et al. The 6-minute walk: a new measure of exercise capacity in patients with chronic heart failure. Can Med Assoc J 1985;132(8):919-923 [FREE Full text] [Medline: 3978515]
29. Enright PL, Sherrill DL. Reference equations for the six-minute walk in healthy adults. Am J Respir Crit Care Med 1998;158(5 Pt 1):1384-1387. [doi: 10.1164/ajrccm.158.5.9710086] [Medline: 9817683]
30. Morishita S, Tsubaki A, Nashimoto S, Fu JB, Onishi H. Face scale rating of perceived exertion during cardiopulmonary exercise test. BMJ Open Sport Exerc Med 2018;4(1):e000474 [FREE Full text] [doi: 10.1136/bmjsem-2018-000474] [Medline: 30622732]
31. Resnick B, Jenkins LS. Testing the reliability and validity of the self-efficacy for exercise scale. Nurs Res 2000;49(3):154-159. [doi: 10.1097/00006199-200005000-00007] [Medline: 10882320]
32. Li J, Wang Y, Shen L. Development and psychometric tests of a Chinese version of the SF-36 Health Survey Scales. Zhonghua Yu Fang Yi Xue Za Zhi 2002 Mar;36(2):109-113. [Medline: 12410965]
33. Leng M, Sun Y, Li C, Han S, Wang Z. Usability evaluation of a knowledge graph-based dementia care intelligent recommender system: mixed methods study. J Med Internet Res 2023;25:e45788 [FREE Full text] [doi: 10.2196/45788] [Medline: 37751241]
34. Hyzy M, Bond R, Mulvenna M, Bai L, Dix A, Leigh S, et al. System Usability Scale benchmarking for digital health apps: meta-analysis. JMIR Mhealth Uhealth 2022;10(8):e37290 [FREE Full text] [doi: 10.2196/37290] [Medline: 35980732]
35. de Melo MB, Daldegan-Bueno D, Menezes Oliveira MG, de Souza AL. Beyond ANOVA and MANOVA for repeated measures: advantages of generalized estimated equations and generalized linear mixed models and its use in neuroscience research. Eur J Neurosci 2022;56(12):6089-6098. [doi: 10.1111/ejn.15858] [Medline: 36342498]
36. Thorup C, Hansen J, Grønkjær M, Andreasen JJ, Nielsen G, Sørensen EE, et al. Cardiac patients' walking activity determined by a step counter in cardiac telerehabilitation: data from the intervention arm of a randomized controlled trial. J Med Internet Res 2016;18(4):e69 [FREE Full text] [doi: 10.2196/jmir.5191] [Medline: 27044310]
37. Xu L, Li F, Zhou C, Li J, Hong C, Tong Q. The effect of mobile applications for improving adherence in cardiac rehabilitation: a systematic review and meta-analysis. BMC Cardiovasc Disord 2019;19(1):166 [FREE Full text] [doi: 10.1186/s12872-019-1149-5] [Medline: 31299903]

XSL•FO
RenderX

38. Rosenstrøm S, Cecilie Tjustrup N, Kallemose T, Risom SS, Hove JD, Brødsgaard A. Evaluating a co-created model for video consultations in cardiac rehabilitation: impact on health literacy, quality of life and family support-a study protocol. BMJ Open 2025;15(10):e101099 [FREE Full text] [doi: 10.1136/bmjopen-2025-101099] [Medline: 41043840]
39. Sankaran S, Dendale P, Coninx K. Evaluating the impact of the HeartHab app on motivation, physical activity, quality of life, and risk factors of coronary artery disease patients: multidisciplinary crossover study. JMIR Mhealth Uhealth 2019;7(4):e10874 [FREE Full text] [doi: 10.2196/10874] [Medline: 30946021]
40. Aldenaini N, Alqahtani F, Orji R, Sampalli S. Trends in persuasive technologies for physical activity and sedentary behavior: a systematic review. Front Artif Intell 2020;3:7 [FREE Full text] [doi: 10.3389/frai.2020.00007] [Medline: 33733127]
41. Salvi D, Ottaviano M, Muuraiskangas S, Martínez-Romero A, Vera-Muñoz C, Triantafyllidis A, et al. An m-Health system for education and motivation in cardiac rehabilitation: the experience of HeartCycle guided exercise. J Telemed Telecare 2018;24(4):303-316 [FREE Full text] [doi: 10.1177/1357633X17697501] [Medline: 28350282]
42. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. J Med Internet Res 2017;19(6):e232 [FREE Full text] [doi: 10.2196/jmir.7126] [Medline: 28663162]
43. Adkisson RV. Nudge: improving decisions about health, wealth and happiness. The Social Science Journal 2019;45(4):700-701. [doi: 10.1016/j.soscij.2008.09.003]
44. Stawarz K, Cox A, Blandford A, Assoc CM. Beyond self-tracking and reminders: designing smartphone apps that support habit formation. In: editors. 2015 Presented at: CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems; 2015 April 18 - 23; Seoul Republic of Korea p. 2653-2662. [doi: 10.1145/2702123.2702230]
45. Schweitzer R, Schlögl S, Schweitzer M. Technology-supported behavior change-applying design thinking to mHealth application development. Eur J Investig Health Psychol Educ 2024;14(3):584-608 [FREE Full text] [doi: 10.3390/ejihpe14030039] [Medline: 38534900]
46. Du H, Wonggom P, Tongpeth J, Clark RA. Six-minute walk test for assessing physical functional capacity in chronic heart failure. Curr Heart Fail Rep 2017;14(3):158-166. [doi: 10.1007/s11897-017-0330-3] [Medline: 28421409]
47. Taylor RS, Long L, Mordi IR, Madsen MT, Davies EJ, Dalal H, et al. Exercise-based rehabilitation for heart failure: cochrane systematic review, meta-analysis, and trial sequential analysis. JACC Heart Fail 2019;7(8):691-705 [FREE Full text] [doi: 10.1016/j.jchf.2019.04.023] [Medline: 31302050]
48. Coulshed A, Coulshed D, Pathan F. Systematic review of the use of the 6-minute walk test in measuring and improving prognosis in patients with ischemic heart disease. CJC Open 2023;5(11):816-825 [FREE Full text] [doi: 10.1016/j.cjco.2023.08.003] [Medline: 38020329]
49. Sohn S, Jeon J, Lee JE, Park SH, Kang DO, Park EJ, et al. Prognostic value of the six-minute walk test in patients with cardiovascular disease. Sci Rep 2025;15(1):20817 [FREE Full text] [doi: 10.1038/s41598-025-04480-9] [Medline: 40593905]
50. Schmidt K, Vogt L, Thiel C, Jäger E, Banzer W. Validity of the six-minute walk test in cancer patients. Int J Sports Med 2013;34(7):631-636. [doi: 10.1055/s-0032-1323746] [Medline: 23444095]
51. Cavalheri V, Burtin C, Formico VR, Nonoyama ML, Jenkins S, Spruit MA, et al. Exercise training undertaken by people within 12 months of lung resection for non-small cell lung cancer. Cochrane Database Syst Rev 2019;6(6):CD009955 [FREE Full text] [doi: 10.1002/14651858.CD009955.pub3] [Medline: 31204439]
52. Diamond JM, Courtwright AM, Balar P, Oyster M, Zaleski D, Adler J, et al. Mobile health technology to improve emergent frailty after lung transplantation. Clin Transplant 2021;35(4):e14236. [doi: 10.1111/ctr.14236] [Medline: 33527520]
53. Azeez M, Clancy C, O'Dwyer T, Lahiff C, Wilson F, Cunnane G. Benefits of exercise in patients with rheumatoid arthritis: a randomized controlled trial of a patient-specific exercise programme. Clin Rheumatol 2020;39(6):1783-1792. [doi: 10.1007/s10067-020-04937-4] [Medline: 32036584]
54. Voorn MJJ, Driessen EJM, Reinders RJEF, van Kampen-van den Boogaart VEM, Bongers BC, Janssen-Heijnen MLG. Effects of exercise prehabilitation and/or rehabilitation on health-related quality of life and fatigue in patients with non-small cell lung cancer undergoing surgery: a systematic review. Eur J Surg Oncol 2023;49(10):106909 [FREE Full text] [doi: 10.1016/j.ejso.2023.04.008] [Medline: 37301638]
55. Chen D, Zhang H, Wu J, Xue E, Guo P, Tang L, et al. Effects of an individualized mhealth-based intervention on health behavior change and cardiovascular risk among people with metabolic syndrome based on the behavior change wheel: quasi-experimental study. J Med Internet Res 2023;25:e49257 [FREE Full text] [doi: 10.2196/49257] [Medline: 38019579]
56. Phelan S, Halfman T, Pinto AM, Foster GD. Behavioral and psychological strategies of long-term weight loss maintainers in a widely available weight management program. Obesity (Silver Spring) 2020;28(2):421-428 [FREE Full text] [doi: 10.1002/oby.22685] [Medline: 31970912]
57. Terbraak M, Major M, Jørstad H, Scholte Op Reimer W, van der Schaaf M. Home-based cardiac rehabilitation in older adults: expert-recommendations for physiotherapist-led care to improve daily physical functioning and reduce comorbidity-related barriers. Eur J Physiother 2024;26(5):288-298. [doi: 10.1080/21679169.2023.2276712] [Medline: 39380594]
58. Terbraak M, Verweij L, Jepma P, Buurman B, Jørstad H, Scholte Op Reimer W, et al. Feasibility of home-based cardiac rehabilitation in frail older patients: a clinical perspective. Physiother Theory Pract 2023;39(3):560-575 [FREE Full text] [doi: 10.1080/09593985.2022.2025549] [Medline: 35068322]

XSL•FO
RenderX

59. Healey EL, Allen KD, Bennell K, Bowden JL, Quicke JG, Smith R. Self-report measures of physical activity. Arthritis Care Res (Hoboken) 2020;72 Suppl 10(Suppl 10):717-730 [FREE Full text] [doi: 10.1002/acr.24211] [Medline: 33091242]
60. Adams SA, Matthews CE, Ebbeling CB, Moore CG, Cunningham JE, Fulton J, et al. The effect of social desirability and social approval on self-reports of physical activity. Am J Epidemiol 2005;161(4):389-398 [FREE Full text] [doi: 10.1093/aje/kwi054] [Medline: 15692083]
61. Prince SA, Adamo KB, Hamel ME, Hardt J, Connor Gorber S, Tremblay M. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. Int J Behav Nutr Phys Act 2008;5:56 [FREE Full text] [doi: 10.1186/1479-5868-5-56] [Medline: 18990237]
62. Sporrel K, De Boer RDD, Wang S, Nibbeling N, Simons M, Deutekom M, et al. The design and development of a personalized leisure time physical activity application based on behavior change theories, end-user perceptions, and principles from empirical data mining. Front Public Health 2020;8:528472 [FREE Full text] [doi: 10.3389/fpubh.2020.528472] [Medline: 33604321]

Abbreviations

6MWD: 6-minute walk distance
6MWT: 6-minute walk test
AE: adverse event
BCT: behavior change technique
CHD: coronary heart disease
CONSORT: Consolidated Standards of Reporting Trials
CR: cardiac rehabilitation
CVD: cardiovascular disease
ESE: exercise self-efficacy
ESES: Exercise Self-Efficacy Scale
GEE: generalized estimating equation
IPAQ: International Physical Activity Questionnaire
MACE: major adverse cardiac event
MCID: minimum clinically important difference
MET: Metabolic Equivalent Task
mHealth: mobile health
PCI: percutaneous coronary intervention
PSD: persuasive systems design
PA: physical activity
RCR: remote cardiac rehabilitation
SUS: System Usability Scale
SF-36: Short Form of 36 Health Survey Questionnaire
SPO2: peripheral oxygen saturation
WHO: World Health Organization

Edited by A Coristine; submitted 20.Jan.2025; peer-reviewed by M Hollings, T Sjögren; comments to author 11.Jul.2025; revised version received 02.Dec.2025; accepted 03.Dec.2025; published 14.Jan.2026.

Please cite as:
Liu Y, Huang X, Dai Z, Jiang Z, Wu W, Wang J, Wang Z, Yu L, Li H, Huang L
Effects of an eHealth Cardiac Exercise Rehabilitation Platform for Patients After Percutaneous Coronary Intervention Based on the Persuasive Systems Design Model: Randomized Controlled Trial
J Med Internet Res 2026;28:e71450
URL: https://www.jmir.org/2026/1/e71450
doi:10.2196/71450
PMID:

©Yang Liu, Xiting Huang, Ziying Dai, Zhili Jiang, Wenxiao Wu, Jing Wang, Zhiqian Wang, Luyao Yu, Hanyu Li, Lihua Huang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the

XSL•FO
RenderX

Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Feasibility, Acceptability, and Perspectives Regarding the Use of Activity Tracking Wearable Devices Among Home Health Aides: Mixed Methods Study

Ian René Solano-Kamaiko[1], MS; Michael Dicpinigaitis[2], BS; Melissa Tan[1], MS; Irene Yang[1], MS; Kexin Cheng[1], MS; Ronica Peramsetty[2], BS; Michelle Shum[2], BS; Yanira Escamilla[3], LCSW; Jennifer Bayly[2], MPH, MD; Meghan Reading Turchioe[4], RN, MPH, PhD; Ariel Avgar[5], PhD; Aditya Vashistha[5], PhD; Nicola Dell[1,5], PhD; Madeline R Sterling[2], MPH, MD, MS, FAHA

[1]Cornell Tech, New York, NY, United States

[2]Weill Cornell Medicine, New York, NY, United States

[3]1199SEIU Funds, New York, NY, United States

[4]Columbia University, New York, NY, United States

[5]Cornell University, Ithaca, NY, United States

Corresponding Author:
Madeline R Sterling, MPH, MD, MS, FAHA
Weill Cornell Medicine
420 East 70th Street, LH-357
New York, NY
United States
Phone: 1 646 962 5029
Email: mrs9012@med.cornell.edu

Abstract

Background: Home health aides and attendants (HHAs) provide in-home care to the growing population of older adults who want to age in place. Despite their vital role in patient care, HHAs are an underserved and vulnerable population of health care professionals who often experience poor health themselves. Activity tracking devices offer a promising way to improve HHAs' health-related awareness and promote health behavior change, particularly regarding physical activity and sleep quality, 2 areas in which the workforce struggles.

Objective: This study aimed to understand how feasible it is for HHAs to use activity tracking devices and assess their perceptions of such devices for improving their health. Specifically, we conducted (1) a field study to assess the use, feasibility, and acceptability of these devices among HHAs and (2) a qualitative study to understand HHAs' perspectives on and reactions to activity trackers on and off the job.

Methods: We partnered with the 1199 Service Employees International Union Training and Employment Fund to conduct a field study with home care agency–employed HHAs working in New York City, New York. Participants wore activity tracking devices for 4 weeks that collected data on physical activity and sleep. The HHAs were subsequently interviewed on their experiences with and attitudes toward the devices and asked to reflect on personalized visualizations of their data to prompt them to think aloud. Quantitative data were analyzed using descriptive statistics. Qualitative data were analyzed using grounded theory.

Results: A total of 17 HHAs participated; their mean age was 48.7 (SD 12.2) years, 15 (88%) were women, 11 (65%) identified as Black, 5 (29%) identified as Hispanic or Latinx, and they had worked as HHAs for a mean of 11.7 (SD 7.5) years. In total, 94% (n=16) of the HHAs wore their activity trackers for the full 28-day study period. Participants took a mean of 10,230 (SD 3586) daily steps during the study period and slept for a mean of 6.27 (SD 0.58) hours per night. Overall, 4 key themes emerged: (1) activity tracking devices enhanced participants' health awareness by providing empirical data for self-reflection; (2) this increased awareness led to positive behavior changes, including setting and achieving health-related goals; (3) HHAs believed that these devices could improve not only their own health but also that of their patients through positive behavior changes; and (4) despite this optimism, participants emphasized that their ability to modify sleep and activity patterns was constrained by social and occupational determinants, with sleep improvements being particularly challenging.

XSL•FO
RenderX

Conclusions: Our findings suggest that appropriately designed personal tracking interventions could offer a promising approach to supporting positive health-related changes in this historically overlooked workforce, potentially improving their well-being and the quality of care they provide to their patients.

(J Med Internet Res 2026;28:e77510) doi:10.2196/77510

KEYWORDS

home health aides and attendants; home health care; home care; worker well-being; frontline work; low-wage work; data governance; data advocacy; wearable activity trackers; passive sensing

Introduction

Home health aides and attendants (HHAs) are an essential frontline workforce providing in-home care to a growing number of older adults who wish to age in place [1-3]. HHAs provide personal care to older adults, including assistance with activities of daily living and instrumental activities of daily living as well as medically oriented care (monitoring symptoms and vital signs and counseling patients on physical activity [PA], diet, and medication adherence), all while providing emotional support [4].

Despite their critical role in patient care, HHAs are an underserved and vulnerable group of health care professionals who often have poor health themselves [5-7]. Predominantly comprising middle-aged women from racial minority groups earning substandard wages (national median of US $12 per hour), this workforce numbers over 3.5 million in the United States and has a high burden of cardiovascular disease risk factors, including obesity, hypertension, hyperlipidemia, physical inactivity, smoking, and poor sleep [4,8,9]. Our recent national study revealed that HHAs' health and health behaviors are notably worse than those of other frontline health care workers not employed in the home (ie, nursing homes and hospitals), with over one-third reporting fair or poor general health, nearly 40% being obese, and more than half being physically inactive [10]. Contributing factors to this include limited health insurance and opportunities to seek medical care, as well as little time to engage in positive health-promoting behaviors themselves. Left unaddressed, these occupational health disparities among HHAs could potentially harm them and compromise the quality of care they provide.

Activity tracking devices, also known as wearables, may be uniquely positioned to promote positive health behavior change in a workforce that has rarely been the focus of interventions and technology innovation. These devices, which include Fitbit trackers, Oura Rings, and others, show promise in potentially improving PA and sleep, 2 behaviors HHAs struggle to achieve optimal levels of, based on our prior studies [4,10]. Consistent with social cognitive theory and the theory of planned behavior, wearable devices allow users to visualize and reflect on their daily activity and sleep data, enhancing self-efficacy and promoting positive behavior change [11,12]. By improving their own health behaviors, HHAs may also model and encourage similar changes among their patients. To date, research on activity trackers in this context has primarily focused on their use among older adults [13,14], family caregivers [15,16], and facility-based care workers in international settings [17-19] and has never focused on HHAs, a growing workforce in constant contact with patients who also struggle to reach sufficient levels of PA and sleep. Furthermore, while these adjacent groups share caregiving roles, HHAs differ in that they are paid, agency-employed workers who face unique structural and occupational barriers to prioritizing their own health [20,21]. Therefore, studying HHAs extends existing work on caregiver- and worker-focused interventions by addressing a critical and understudied population within the US health care system.

To understand the potential for activity trackers to promote positive health behavior change, we conducted (1) a field study to assess the use, feasibility, and acceptability of these devices for daily use and (2) a qualitative study to understand HHAs' perspectives on and reactions to activity trackers on and off the job.

Methods

Overview and Study Setting

We conducted this field and qualitative study from July 19, 2024, to October 11, 2024, in collaboration with 1199 Service Employees International Union Training and Employment Fund (1199SEIU-TEF), a labor management fund of the 1199 Service Employees International Union (SEIU) United Healthcare Workers East, the largest health care union in the United States [22]. The 1199SEIU-TEF provides education and training to more than 55,000 HHAs who are employed by over 50 licensed and certified home care agencies in New York City, New York [23].

Study Participants and Recruitment

To be eligible, HHAs had to (1) be aged ≥18 years, (2) be English speaking, (3) own a smartphone (iOS or Android), and (4) be employed as an HHA in New York City. To recruit our intended sample of 25 to 30 study participants, 1199SEIU-TEF staff reached out to HHAs who had previously expressed interest in participating in research initiatives and invited them to take part in our study using standardized recruitment materials (email script and flyer). If potential participants were interested, staff referred them to the study team to be screened for eligibility. HHAs who were eligible for the study provided written consent. All aspects of this study took place in person at 1199SEIU-TEF headquarters in New York City.

Data Collection and Study Procedures

Before conducting the field study, the research team met with participants to elicit their initial perspectives on activity tracking devices and the device form factor (wristband vs ring). From these conversations, we found that HHAs were enthusiastic about the potential benefits of activity trackers for increasing

XSL•FO
RenderX

their health awareness. HHAs preferred wristband devices over rings as they felt that the wristband's small screen would be easier to check discreetly during the workday than their phone and its form factor would not interfere with their daily tasks.

The field study involved giving HHAs Fitbit Charge 6 devices, which they were asked to wear for 28 days. Before providing the devices, we conducted a 1-hour onboarding session where participants were introduced to the research team; asked to respond to a few self-reported demographic questions (age, gender, race, Hispanic or Latinx ethnicity, educational level, and years of experience as an HHA); and guided through setting up the Fitbit device and mobile app, highlighting key features including activity and sleep tracking. After the participants created their Fitbit accounts, we walked them through the authorization process to allow the research team remote access to their data through the Fitbit application programming interface.

Participants were then asked to wear the Fitbit daily for 4 weeks, during which the research team contacted them weekly via their preferred method (SMS text message). Before the field study, we confirmed participants' contact preferences (eg, SMS text message, email, or phone), and all preferred SMS text messages. Prior work has also supported this modality for communicating with HHAs [24].

SMS text message reminders prompted participants to wear the wristband, review their data, and contact the team with questions or issues. However, these reminders did not substantially affect engagement or acceptance. Most HHAs either ignored the messages or briefly acknowledged their receipt. In interviews, participants reported wearing the devices consistently because they valued the insights into their health and well-being, with SMS text messages serving primarily as a channel for troubleshooting or procedural clarifications (eg, how to access participant gift cards). Following the 28-day field study, we used the Fitbit application programming interface to collect participants' data to create personalized visualizations showing each participant's activities in relation to (1) the National Institutes of Health's (NIH) recommended PA and sleep guidelines for adults and (2) their broader group of peers in the study.

Four investigators from the research team (IRS-K, MT, IY, and KC) then conducted 1-hour semistructured interviews with each participant to understand their experiences, data preferences, and insights into the wearable devices (Multimedia Appendix 1). Along with participants' general perspectives, the investigators focused on device acceptability—as illustrated by participants' attitudes toward the devices' usability and perceived value. Following these discussions, we introduced the personalized visualizations of their data and prompted participants to engage in a think-aloud process [25]. Participants reflected from multiple perspectives: as individuals balancing complex personal and work lives, as members of a distributed peer group, and as part of a broader community advocating for improved working conditions. Interviews ended when thematic saturation was reached.

Quantitative Data Analysis

Participant characteristics were analyzed using descriptive statistics. We collected participants' Fitbit data, comprising 187 JSON files containing metrics on activity levels (sedentary, light, fairly active, and very active), step count, distance, heart rate, and sleep duration and stages (wake, light, rapid eye movement, and deep), each tagged with precise date and time stamps. We focused on sleep duration and step count and assessed use feasibility, defined as participants' ability to consistently engage with the devices, by examining the date and time stamps associated with these metrics to determine whether the devices captured participants' data on a daily basis.

We selected sleep duration and step count because these metrics were intuitive for participants (ie, required little explanation, were immediately interpretable, and were clearly actionable) and have clear policy relevance, such as the No More 24 campaign [26] advocating to end 24-hour HHA shifts. Furthermore, prior research indicates that Fitbit-derived sleep duration and step counts are reasonably reliable and accurate; systematic reviews and validation studies have found that Fitbit devices provide sleep duration estimates comparable to those of research-grade accelerometers [27,28], perform well in distinguishing wake from sleep [28,29], and meet acceptable validity and reliability standards for tracking PA in free-living conditions [30].

Participants' Fitbit data (steps and sleep duration) were processed using exploratory data analysis methods in Jupyter Notebook [31]. As mentioned above, during the exploratory data analysis process, we generated data visualizations for the reflection interviews, which also incorporated auxiliary data such as participants' sleep and work schedules along with the NIH's PA and sleep recommendations to create benchmarks for optimal levels.

Qualitative Data Analysis

Each participant interview lasted approximately 60 minutes, resulting in 17 hours of audio-recorded interview data. The interviews were professionally transcribed using NoScribe (noScribe.ai) [32], an open-source artificial intelligence tool run on the research team's computers. We verified each transcript against the original recordings, correcting errors and redacting identifying information. Our analysis used grounded theory and, as such, used an inductive open coding approach [25] with 4 investigators (IRS-K, MT, IY, and KC). We established a baseline code set by jointly coding one interview and then validated our approach by coding a second interview in 2 separate pairs (IRS-K and KC as well as MT and IY), meeting to reconcile differences. The remaining 15 interviews were coded independently, with regular meetings to resolve disagreements and refine the coding scheme. Finally, we conducted affinity diagramming to synthesize these codes into high-level themes with representative quotations.

Ethical Considerations

All participants gave informed consent to take part in the study, including consent for note taking, photo taking, audio recording of interviews, and collection of participants' Fitbit data. Identifiable details in the data and features in participant images

XSL•FO
RenderX

were removed to ensure participant privacy and confidentiality. Participants were compensated for their time with US $25 gift cards for each of the 3 components of data collection (preliminary discussions, field study, and final interview), for a total possible compensation of US $75 in addition to the Fitbit device, which they were allowed to keep. Cornell University's institutional review board classified the study as expedited and reviewed and approved the project under protocol IRB0148598.

Results

Participant Characteristics

A total of 17 HHAs participated in this study. They had a mean age of 48.7 (SD 12.2) years, 88% (n=15) were women, 65% (n=11) identified as Black, 29% (n=5) identified as Hispanic or Latinx, and they had worked as HHAs for a mean of 11.6 (SD 7.5) years (Table 1).

Table 1. Sample characteristics (N=17).

Characteristic	Values
Age (years), mean (SD)	48.7 (12.2)
Gender, n (%)	
Women	15 (88)
Men	2 (12)
Race, n (%)	
African American or Black	11 (65)
White	1 (6)
Ethnicity, n (%)	
Hispanic or Latinx	5 (29)
Educational level, n (%)	
High school diploma	3 (18)
Some college	3 (18)
Associate's degree	4 (24)
Bachelor's degree	5 (29)
Graduate degree	2 (12)
Experience (years), mean (SD)	11.6 (7.5)

Field Study Findings

Overall, 94% (16/17) of the HHAs wore their activity trackers for the full 28-day study period, with only 6% (1/17) wearing the device for 26 days, and 82% (14/17) indicated continued use after the study's conclusion. As shown in Table 2, participants took a mean of 10,230 (SD 3586) daily steps during the study period, ranging from 3855 to 20,528 steps. In total, 53% (9/17) of the participants surpassed the NIH recommendations for 10,000 daily steps for healthy adults [33]. Table 2 shows that participants slept for a mean of 6.27 (SD 0.58) hours per night, ranging from 5.40 to 7.38 hours, with 18% (3/17) exceeding the NIH recommendations for 7 hours or more of daily sleep duration for healthy adults [34].

XSL•FO
RenderX

Table 2. Wearable device-measured mean steps and sleep duration among study participants (N=17). Comparison of mean daily step count and mean daily sleep (in hours) by participant. The National Institutes of Health recommends that healthy adults should walk 10,000 steps [33] and sleep 7 hours or more per day [34].

Participants	Daily step count, mean (SD)	Daily sleep (hours), mean (SD)
P1	3855 (2345)	6.59 (2.36)
P2	12,683 (4577)	6.58 (2.05)
P3	11,680 (2552)	5.82 (1.87)
P4	6867 (3705)	7.38 (1.58)
P5	7935 (5063)	5.75 (3.04)
P6	7133 (2782)	5.87 (1.95)
P7	20,528 (6384)	7.30 (0.99)
P8	10,262 (3442)	5.79 (1.56)
P9	10,739 (3386)	5.74 (1.66)
P10	11,928 (3351)	6.46 (1.30)
P11	11,133 (3217)	6.13 (1.95)
P12	9212 (3556)	6.36 (0.80)
P13	12,980 (7196)	5.40 (1.84)
P14	7026 (2396)	7.15 (1.85)
P15	7863 (1894)	5.64 (1.23)
P16	9063 (3976)	6.55 (1.17)
P17	13,024 (3123)	6.09 (1.76)

Major Themes From Qualitative Interviews

A total of 4 major themes arose. Each theme is presented below with accompanying quotations that represent key concepts.

Theme 1: Activity Tracking Devices Were Feasible to Wear and Improved HHAs' Health Awareness

Most HHAs found the use of the devices to be seamless and feasible to incorporate into their day-to-day activities. Many cited the personal value they derived from using the devices. Participant P9 shared their experience:

> *It's becoming part of me, since one month I've been on it, so it's like I'm getting accustomed to it.... I love wearing it because the information it provided...was very helpful to my life.*

While many participants reported having a general awareness of their own health and well-being, a number stated how the activity tracker helped enhance their understanding of their own health. For instance, P4 said:

> *The difference is that with the Fitbit, I can see exactly how long I sleep or walk.... I know that sometimes I wasn't sleeping, but I can [only] say maybe I slept five or six hours when I count. But with the Fitbit, it's going to let you know exactly how long was the deeper sleep, the lighter sleep, and that stuff.*

Participants also reflected on the fact that, after being shown their data, this informed their perceptions of their own health behaviors. For example, P5 reflected:

> *A lot of us are going to be surprised because in our mind, we're sleeping a lot.... If this didn't come about, I might have said, "oh, yes, I get enough sleep." But this now is in fact showing me, "oh, girl, you're not sleeping."*

This revelation was common among participants, who found that the tracking data offered a concrete way to reflect on their habits and gain a deeper understanding of how their work and home life impacted aspects of their well-being, such as sleep. P16 reflected:

> *When I look, sometimes I see I have four hours, 40 minutes of sleep.... I'm tired and everything is bothering me. But prior to the Fitbit, I didn't realize what was bothering me.*

Theme 2: Activity Tracking Devices Helped HHAs Set and Achieve Health Goals

Many participants shared how increased awareness of their health led to positive changes in behavior. Several reported adopting healthier habits because of heightened awareness of their sleep patterns. Examples they shared included limiting television and phone use (ie, screen time), adjusting their bedtime and wake-up times, and incorporating relaxation techniques such as warm showers and listening to calming music. For example, P6 explained:

> *When I watched my sleep time (on my device), I was sleeping between three and five [hours].... [Now I] try and see if I could fall to sleep earlier.... I started getting eight hours, seven hours. I started trying to focus on going to sleep now, [no] TV, [no] phone.*

XSL•FO
RenderX

Reflecting on their activity data also had a motivating effect for many participants, helping them re-evaluate their existing behaviors and adopt new habits. Several mentioned that these changes had a compounding effect, where improvements in one area, such as increased walking, made them feel better and motivated them to walk even more. P5 shared their experience:

> *If I'm out there...for five, six minutes and I don't see the bus, I'm going down the road. [Before] I wouldn't do that. I'd stay there and wait for the bus. But it [the Fitbit] really encourages me now to walk. And as I said, the more I walk, the lighter I feel. The more encouraged I feel.*

Other participants highlighted how greater health awareness helped them set and achieve specific health goals. For example, P3 discussed how she aimed to improve her PA at work to meet the step count goal (which was preprogrammed into the device). She said:

> *Sometimes I go to work, when I see I didn't have the number of steps required, when I put my patient to bed, I went out to finish the steps. To have the score. So that was very interesting because I never did it before.*

Theme 3: Activity Tracking Devices May Have Potential to Improve the Health of HHAs and Their Patients

HHAs were acutely aware of the negative impacts of poor sleep on their health and well-being. Beyond personal health improvements, they expressed concerns about their ability to perform their high-stakes job, which demands constant vigilance, when sleep deprived. For instance, P5 highlighted this challenge, stating:

> *If you don't get enough rest, you cannot function effectively.... So, you as a [home care agency] admin, you want your workers to perform effectively. What will you do now for us so we can get adequate sleep?*

At the same time, several HHAs were optimistic about the potential of activity tracking devices to enhance both their personal health and the health of their patients while on the job. For example, P8 said:

> *If you get a case that will let you [and] the patient walk. So the two of you can walk around and do more exercise.*

Theme 4: HHAs' Abilities to Make Sleep and Activity Changes Were Impacted by Social and Occupational Determinants

HHAs reported that certain factors related to their social environments or working conditions impacted their health and made it challenging to make desired health changes. For example, participants working live-in cases or overnight shifts mentioned that their sleep schedules were dictated by patients' demands. P17 explained:

> *In live-in case we, by the law, [we] are supposed to sleep five hours not interrupted, but it's not possible.... You can't sleep this amount of hours.... They [patients] put TV on high volume and you can't sleep.*

This demonstrates the limited control that HHAs have over their sleep schedules. Additionally, HHAs expressed the need to remain vigilant, attending to patients' needs and staying alert for emergencies. This constant hypervigilance negatively affected their ability to relax and sleep during overnight shifts.

HHAs also noted the challenges of nontraditional work schedules. Shift work, subject to frequent changes based on patient needs, made it difficult for many to regulate their sleep, even on days off. For instance, P4 described the stability of working with the same patient vs their current unpredictable schedule:

> *Like last year or years before, I have a stable patient. I know that I was working four days a week with the same patient.... I know exactly what was my schedule at that time. But right now, I have no fixed schedule.*

Others, such as P15, faced difficulties regardless of their work schedule (daytime or nighttime), expressing difficulties balancing responsibilities in addition to work, such as school, family, and more, which impacted their ability to stay healthy:

> *Once I'm at home, I want to like catch up [on] my [school] assignments or cause currently I'm doing my internship as well. I still have to keep up with my personal life, my kids, my husband.... I have three kids and they are just like babies.... So when I'm at home, I think I tend to do more cause, if I can sleep, they have to like sleep before I can get to do anything.*

Beyond sleep, HHAs also felt that their PA was often dictated by patients' needs. P14 explained:

> *My patients, some of them, they want company.... They say, "let's watch a movie, or let's go to the theater," and they call an Uber. So you don't walk that much. You walk during the time that you're going to work, and then you go home.*

While some HHAs found that their PA was limited by their cases, others felt that their activity levels were more malleable. Many developed strategies to increase exercise, such as walking instead of taking the bus, neighborhood walks on days off, and encouraging patients to go outside for exercise with them during downtime.

Discussion

Principal Findings

To our knowledge, this is the first study to examine how activity tracking devices influence HHAs' awareness of their health and well-being using empirical data from the devices and qualitative reflection interviews with participants. Our findings demonstrate that the use of activity trackers was both feasible and acceptable among HHAs. Feasibility was reflected quantitatively by participants' ability to consistently engage with the devices, with 94% (16/17) wearing them for the full 4-week field study period. Acceptability was shown qualitatively through participants' reflections during follow-up interviews, where they expressed positive attitudes toward the devices' usability and perceived value for increasing health awareness. This level of engagement is noteworthy as previous research has shown

XSL•FO
RenderX

lower participant adherence to consumer wearables [35]. The high engagement observed among HHAs in this study suggests that activity tracking interventions may be particularly effective for supporting the health and well-being of HHAs more broadly, although further research is necessary to validate this assumption.

Additionally, the Fitbit Charge 6's wristband form factor and small screen provided participants with a discreet, functional user experience. Beyond use and feasibility, HHAs not only found that the devices offered them more awareness of key aspects of health—PA and sleep—but even in a short time, they reflected that they were empowered to make positive behavior changes because of wearing the devices. Finally, we found that the devices may have health benefits not only for HHAs but also for the patients for whom they care.

Our study builds on scientific literature in a few key ways. First, while many studies have focused on the acceptance and impact of wearables among older adults [13,14], few have addressed their use among family caregivers and facility-based health care workers [15-19], and even fewer have examined these applications with HHAs [36].

For example, several studies conducted in Finland and Norway have examined the use of wearables among family caregivers and facility-based home care nurses, nursing assistants, and occupational therapists [17-19]. However, these studies often relied on technologies such as accelerometers [17,18] and electrode-based monitors [19], which can be difficult to set up and may not offer real-time or easily interpretable health feedback for lay users, unlike consumer-friendly devices such as Fitbit trackers.

To date, there has been one prior case study in which Fitbits were deployed among caregivers at the SEIU 775 Benefits Group in Washington state [36]. The study found that using Fitbit devices, combined with health coaching, resulted in improvements in health behaviors over a 4-month period. These improvements included increased PA (measured via step counts and active minutes), weight loss, and reduced blood pressure. The findings were based on multiple blood pressure readings, self-reported weight loss, and recorded activity data from the devices. However, this was a pilot program conducted by SEIU 775 Benefits Group in collaboration with Fitbit, Inc (a Google subsidiary), resulting in a self-authored "case study" rather than a peer-reviewed journal article. While the case study presents promising findings, its details and depth are limited due to its intended audience. Our study builds on this preliminary work by outlining a detailed and robust study methodology focused on Fitbit devices for improving HHAs' awareness of their health and well-being. Additionally, our study incorporates both wearable data and qualitative reflections, aimed at providing a deeper understanding of how HHAs navigate barriers to improving their health and how tracking devices may help support these efforts.

Prior research has shown that wearable devices support health monitoring and facilitate behavior change, particularly in PA and sleep [13,16], and our study extends the literature to HHAs. In doing so, it highlights how these devices have the potential to improve the health of a vulnerable population and shape workplace dynamics and patient care within home care settings. Importantly, we found that wearables can not only track behavior but also offer a simple and feasible way to readily influence workplace routines, patient interactions, and broader occupational health strategies in home care settings. This finding is notable as previous research has shown that influencing behavior change is often challenging and may reflect HHAs' perception of being in control of their own health [37].

In recent years, there has been a growing call for better understanding and improving the health of HHAs [4,7,9,10]. However, empirical studies that aim to improve their health behaviors and health are scarce. Our study is a first step toward this through harnessing technology and offering HHAs real-time access to their own health data. While we found striking benefits, it is also important to note the challenges. For example, while these devices can offer HHAs personalized feedback and enhance user motivation and self-efficacy [13,16], we also found that external and structural barriers in the environments in which the HHAs live and work can limit the extent to which HHAs can act on their health feedback, making it difficult to prioritize meaningful and sustainable behavioral health changes. For example, unpredictable work schedules, multiple jobs, and familial obligations all left them with limited time and resources to obtain sufficient sleep.

Nevertheless, HHAs found the intervention valuable, using wearable devices to self-monitor, reflect on current health habits, and make changes to their behaviors. Some participants adjusted their sleep routines, such as modifying bedtimes and reducing screen time before sleep, whereas others increased their PA in response to real-time device feedback. Consistent with social cognitive theory and the theory of planned behavior, these self-directed changes suggest that wearables can increase health awareness and motivation to change, representing a promising approach to promote health behavior change among HHAs [11,12].

Although structural barriers such as power asymmetries among workers, patients, and agencies; low wages; and isolated work environments limit HHAs' ability to improve their health and well-being, wearable data could be integrated into existing peer support programs that aim to reduce worker isolation and strengthen community ties [38,39]. For example, in our prior studies, we have found that HHAs are more likely to engage in PA if they know peers or their clients are engaged as well. Our findings in this study may help bolster and extend these efforts, which have largely relied on HHAs' self-reported experiences. Aggregated wearable data could more concretely illustrate the challenges that HHAs face, enrich qualitative accounts, and guide the design of peer support programs by grounding discussions and activities in workers' collective experiences to identify strategies for navigating or mitigating these structural constraints.

Notably, the findings also suggest that PA may be more modifiable than sleep for the HHAs and have clearer positive implications for their patients. That is, several participants found that wearing their activity trackers reminded them of the importance of PA on the job and motivated them to spend more time walking with their patients to increase both their step

XSL•FO
RenderX

counts. This is an important observation as previous research has identified challenges in improving PA among older adults receiving home care [40] and researchers have suggested that activities such as walking are valuable for promoting PA outside of structured exercise programs [41].

In future studies, this relationship warrants investigation. Key questions include the following: does increasing HHAs' step count in turn increase their patients' PA and mobility and reduce the risk of falls? Does increasing time spent together moving foster closer relationships between HHAs and their patients through shared activity? Our work has previously found that higher levels of mutuality between HHAs and patients result in higher job satisfaction among HHAs [42], and this too could be formally examined after introducing technology. Another important area for exploration is the affordability of wearable devices for a workforce with limited disposable income (eg, the Fitbit Charge 6 costs approximately US $150 as of this writing). If these devices are shown to significantly impact worker and patient outcomes, they may be justified for inclusion in future workforce wellness programs. Alternatively, existing devices such as smartphones could offer a more cost-effective means of collecting data on PA and sleep. However, unlike wearable devices, smartphones are not continuously worn on the body, which may limit the accuracy of the data they capture. Regardless, the feasibility of using smartphones should be further examined.

Beyond improving personal health, these devices may help further center workers' experiences by introducing an additional layer of granular quantitative data to complement workers' qualitative narratives. This collection of data sources may help bolster ongoing advocacy efforts, such as supporting improving wages or better working conditions. For example, collective data can strengthen labor union and worker advocacy campaigns such as the push for Fair Pay for Home Care [43], which advocates for livable wages and job stability for HHAs. Similarly, the No More 24 campaign [26], which seeks to end the harmful practice of requiring HHAs to work 24-hour shifts, where HHAs work for 24 hours but are only paid for 13 hours under the premise that they receive at least 5 hours of uninterrupted sleep and other breaks, is directly relevant to HHAs' poor sleep levels. Our research suggests that tracking HHAs' activities could generate data-driven evidence on sleep patterns and quality, which, when combined with worker-centered reflections that contextualize these data, could provide compelling arguments for policymakers and bolster organizational advocacy efforts. Future research should explore how to responsibly manage and share HHAs' collective data in ways that center their needs and experiences to improve their health and well-being.

Limitations

We acknowledge that our study has several limitations. We conducted a small-scale, short-term empirical mixed methods study in a single geographic location, New York City. Further research is needed to understand how our findings might generalize to larger samples and other locales, including rural and non-US settings, particularly those where driving is a more frequent mode of transportation than walking or public transit. Participants were recruited through their affiliation with a labor management sector of a health care union and may have been more motivated to engage with the wearable devices than nonunionized HHAs, which could confound our findings and limit generalizability to the broader HHA workforce.

Our study did not ask separate questions regarding sex and gender or race and ethnicity, which may obscure participants' intersectional experiences. Future work should aim to include these perspectives. Additionally, researchers should investigate the results of a longitudinal study to understand how participants' experiences might change over time and how they objectively impact the patients they serve. Our study also has limitations inherent to qualitative interviews, including potential participant response bias [44]. Further research might explore techniques such as journaling and remotely distributed surveys to decrease the impact of researchers' positionality.

Conclusions

Our findings suggest that personal tracking devices offer a feasible and acceptable approach to supporting positive health-related changes in this historically overlooked workforce. These devices have the potential to improve both the health behaviors and well-being of HHAs and, potentially, of the patients for whom they care. These concepts should be formally tested in future clinical trials.

Acknowledgments

The authors would like to thank the 1199 Service Employees International Union Training and Employment Fund team for their assistance with recruitment. They would also like to thank the study participants for sharing their time and experience to make this study possible.

Funding

This work was supported in part by a Google Cyber NYC Research Award, an American Heart Association grant (23SCISA1142170), and the Initiative on Home Care Work at the Center for Applied Research at Cornell.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: IRS-K, AV, ND, MRS
Data curation: IRS-K, IY
Formal analysis: IRS-K, MD, MT, IY, KC, ND, MRS
Funding acquisition: AA, AV, ND, MRS
Investigation: IRS-K, MD, MT, IY, KC, RP, ND, MRS
Methodology: IRS-K, MD, MT, IY, KC, YE, AA, AV, ND, MRS
Project administration: IRS-K, MD, RP, MS, YE, ND, MRS
Resources: IRS-K, IY, YE
Software: IRS-K, IY
Supervision: IRS-K, JB, MRT, AA, AV, ND, MRS
Validation: IRS-K, MT, IY, RP, MS, ND, MRS
Visualization: IRS-K, MD, MT, IY, AV, ND, MRS
Writing—original draft: IRS-K, MD, RP, JB, MRT, ND, MRS
Writing—review and editing: IRS-K, MD, RP, ND, MRS

Conflicts of Interest

MRT would like to disclose her start-up, Iris OB Health Inc. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Qualitative interview guide.
[DOCX File , 1902 KB - jmir_v28i1e77510_app1.docx]

References

1. Markkanen P, Quinn M, Galligan C, Chalupka S, Davis L, Laramie A. There's no place like home: a qualitative study of the working conditions of home health care providers. J Occup Environ Med 2007 Mar;49(3):327-337. [doi: 10.1097/JOM.0b013e3180326552] [Medline: 17351519]
2. Occupational outlook handbook: home health aides and personal care aides. U.S. Bureau of Labor Statistics. URL: https://www.bls.gov/ooh/healthcare/home-health-aides-and-personal-care-aides.htm [accessed 2025-05-29]
3. Sterling MR, Shaw AL. Sharing the care-a patient and her caregivers. JAMA Intern Med 2019 Dec 01;179(12):1617-1618. [doi: 10.1001/jamainternmed.2019.4231] [Medline: 31566655]
4. Cho J, Toffey B, Silva AF, Shalev A, Safford MM, Phillips E, et al. To care for them, we need to take care of ourselves: a qualitative study on the health of home health aides. Health Serv Res 2023 Jun;58(3):697-704 [FREE Full text] [doi: 10.1111/1475-6773.14147] [Medline: 36815290]
5. Sterling MR, Kern LM, Safford MM, Jones CD, Feldman PH, Fonarow GC, et al. Home health care use and post-discharge outcomes after heart failure hospitalizations. JACC Heart Fail 2020 Dec;8(12):1038-1049 [FREE Full text] [doi: 10.1016/j.jchf.2020.06.009] [Medline: 32800510]
6. Thompson A, Fleischmann KE, Smilowitz NR, de Las Fuentes L, Mukherjee D, Aggarwal NR, Peer Review Committee Members. 2024 AHA/ACC/ACS/ASNC/HRS/SCA/SCCT/SCMR/SVM guideline for perioperative cardiovascular management for noncardiac surgery: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Circulation 2024 Nov 05;150(19):e351-e442 [FREE Full text] [doi: 10.1161/CIR.0000000000001285] [Medline: 39316661]
7. Sterling MR, Silva AF, Leung PB, Shaw AL, Tsui EK, Jones CD, et al. "It's like they forget that the word 'health' is in 'home health aide'": understanding the perspectives of home care workers who care for adults with heart failure. J Am Heart Assoc 2018 Dec 04;7(23):e010134 [FREE Full text] [doi: 10.1161/JAHA.118.010134] [Medline: 30571599]
8. Bercovitz A, Moss A, Sengupta M, Park-Lee EY, Jones A, Harris-Kojetin LD. An overview of home health aides: United States, 2007. Natl Health Stat Report 2011 May 19(34):1-31 [FREE Full text] [Medline: 21688727]
9. Silver S, Boiano J, Li J. Patient care aides: differences in healthcare coverage, health-related behaviors, and health outcomes in a low-wage workforce by healthcare setting. Am J Ind Med 2020 Jan;63(1):60-73 [FREE Full text] [doi: 10.1002/ajim.23053] [Medline: 31631375]
10. Sterling MR, Li J, Cho J, Ringel JB, Silver SR. Prevalence and predictors of home health care workers' general, physical, and mental health: findings from the 2014 2018 behavioral risk factor surveillance system. Am J Public Health 2021 Dec;111(12):2239-2250. [doi: 10.2105/AJPH.2021.306512] [Medline: 34878879]
11. Bandura A. Health promotion from the perspective of social cognitive theory. Psychol Health 1998 Jul;13(4):623-649. [doi: 10.1080/08870449808407422]
12. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]

XSL•FO
RenderX

13. Moore K, O'Shea E, Kenny L, Barton J, Tedesco S, Sica M, et al. Older adults' experiences with using wearable devices: qualitative systematic review and meta-synthesis. JMIR Mhealth Uhealth 2021 Jun 03;9(6):e23832 [FREE Full text] [doi: 10.2196/23832] [Medline: 34081020]
14. Kononova A, Li L, Kamp K, Bowen M, Rikard RV, Cotten S, et al. The use of wearable activity trackers among older adults: focus group study of tracker perceptions, motivators, and barriers in the maintenance stage of behavior change. JMIR Mhealth Uhealth 2019 Apr 05;7(4):e9832 [FREE Full text] [doi: 10.2196/mhealth.9832] [Medline: 30950807]
15. Aparicio A, Inostroza-Correa MA, Miranda P, Salinas C, Herskovic V, Chackiel C, et al. The impact of technology use for care by informal female caregivers on their well-being: a scoping review. Syst Rev 2025 Apr 17;14(1):89 [FREE Full text] [doi: 10.1186/s13643-025-02817-z] [Medline: 40241162]
16. Mahmood A, Kim H, Kedia S, Dillon P. Wearable activity tracker use and physical activity among informal caregivers in the United States: quantitative study. JMIR Mhealth Uhealth 2022 Nov 24;10(11):e40391 [FREE Full text] [doi: 10.2196/40391] [Medline: 36422886]
17. Mänttäri S, Säynäjäkangas P, Selander K, Laitinen J. Increased physical workload in home care service is associated with reduced recovery from work. Int Arch Occup Environ Health 2023 Jul 18;96(5):651-660 [FREE Full text] [doi: 10.1007/s00420-023-01960-1] [Medline: 36808567]
18. Lohne FK, Fimland MS, Rasmussen CL, Liaset IF, Fischer H, Redzovic S. Is patients' activities of daily living self-care score in Norwegian home care a proxy for workers standing at work? BMC Health Serv Res 2024 May 09;24(1):565 [FREE Full text] [doi: 10.1186/s12913-024-10897-1] [Medline: 38724977]
19. Tjøsvoll SO, Wiggen Ø, Gonzalez V, Seeberg TM, Elez Redzovic S, Frostad Liaset I, et al. Assessment of physical work demands of home care workers in Norway: an observational study using wearable sensor technology. Ann Work Expo Health 2022 Nov 15;66(9):1187-1198 [FREE Full text] [doi: 10.1093/annweh/wxac052] [Medline: 35959647]
20. Grønoset Grasmo S, Frostad Liaset I, Redzovic SE. Home health aides' experiences of their occupational health: a qualitative meta-synthesis. Home Health Care Serv Q 2021 May 05;40(2):148-176 [FREE Full text] [doi: 10.1080/01621424.2021.1921650] [Medline: 33949920]
21. Yanez Hernandez M, Kuo EF, Henriquez Taveras Y, Lee A, Ramos A, Ringel J, et al. Mental health and well-being among home health aides. JAMA Netw Open 2024 Jun 03;7(6):e2415234 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.15234] [Medline: 38842806]
22. 1199SEIU benefit and pension funds. 1199SEIU Funds. URL: https://www.1199seiubenefits.org [accessed 2025-04-28]
23. MyTEF sign-up: 1199SEIU benefit and pension funds. 1199SEIU Funds. URL: https://www.1199seiubenefits.org/mytef-sign-up/ [accessed 2025-04-28]
24. Sterling M, Espinosa C, Vergez S, McDonald MV, Ringel J, Tobin J, et al. Home health aides caring for adults with heart failure: a pilot randomized clinical trial. JAMA Netw Open 2025 Nov 03;8(11):e2548121 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.48121] [Medline: 41213039]
25. Charters E. The use of think-aloud methods in qualitative research an introduction to think-aloud methods. Brock Educ J 2003 Jul 01;12(2):33. [doi: 10.26522/brocked.v12i2.38]
26. Home page. No More 24!. URL: https://nomore24.org/ [accessed 2025-05-29]
27. Diaz KM, Krupka DJ, Chang MJ, Peacock J, Ma Y, Goldsmith J, et al. Fitbit®: an accurate and reliable device for wireless physical activity tracking. Int J Cardiol 2015 Apr 15;185:138-140 [FREE Full text] [doi: 10.1016/j.ijcard.2015.03.038] [Medline: 25795203]
28. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Accuracy of wristband Fitbit models in assessing sleep: systematic review and meta-analysis. J Med Internet Res 2019 Nov 28;21(11):e16273 [FREE Full text] [doi: 10.2196/16273] [Medline: 31778122]
29. Feehan LM, Geldman J, Sayre EC, Park C, Ezzat AM, Yoo JY, et al. Accuracy of Fitbit devices: systematic review and narrative syntheses of quantitative data. JMIR Mhealth Uhealth 2018 Aug 09;6(8):e10527 [FREE Full text] [doi: 10.2196/10527] [Medline: 30093371]
30. Lee XK, Chee NI, Ong JL, Teo TB, van Rijn E, Lo JC, et al. Validation of a consumer sleep wearable device with actigraphy and polysomnography in adolescents across sleep opportunity manipulations. J Clin Sleep Med 2019 Sep 15;15(9):1337-1346 [FREE Full text] [doi: 10.5664/jcsm.7932] [Medline: 31538605]
31. Project Jupyter. Jupyter. URL: https://jupyter.org [accessed 2025-04-28]
32. Dröge K. kaixxx / noScribe. GitHub. URL: https://github.com/kaixxx/noScribe [accessed 2025-04-28]
33. How many steps for better health? National Institutes of Health. 2019. URL: https://www.nih.gov/news-events/nih-research-matters/how-many-steps-better-health [accessed 2025-04-28]
34. How much sleep do you need? National Center for Chronic Disease Prevention and Health Promotion. 2019. URL: https://magazine.medlineplus.gov/article/how-much-sleep-do-you-need [accessed 2025-04-28]
35. Beukenhorst AL, Howells K, Cook L, McBeth J, O'Neill TW, Parkes MJ, et al. Engagement and participant experiences with consumer smartwatches for health research: longitudinal, observational feasibility study. JMIR Mhealth Uhealth 2020 Jan 29;8(1):e14368 [FREE Full text] [doi: 10.2196/14368] [Medline: 32012078]
36. Caring for caregivers: a case study. Fitbit. URL: https://assets.ctfassets.net/0ltkef2fmze1/6Ja9Y4zhx0ubn0SFuNHkfQ/aac0b22552e9803009596c8e1e408bd0/FE_Customer-Case-Study_SEIU.pdf [accessed 2025-04-28]

XSL•FO
RenderX

37. Reading Turchioe M, Burgermaster M, Mitchell EG, Desai PM, Mamykina L. Adapting the stage-based model of personal informatics for low-resource communities in the context of type 2 diabetes. J Biomed Inform 2020 Oct;110:103572 [FREE Full text] [doi: 10.1016/j.jbi.2020.103572] [Medline: 32961309]
38. Olson R, Elliot D, Hess J, Thompson S, Luther K, Wipfli B, et al. The COMmunity of Practice And Safety Support (COMPASS) Total Worker Health™ study among home care workers: study protocol for a randomized controlled trial. Trials 2014 Oct 27;15:411 [FREE Full text] [doi: 10.1186/1745-6215-15-411] [Medline: 25348013]
39. Poon A, Guerrero L, Loughman J, Luebke M, Lee A, Sterling M, et al. Designing for peer-led critical pedagogies in computer-mediated support groups for home care workers. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023 Presented at: CHI '23; April 23-28, 2023; Hamburg, Germany p. 1-18 URL: https://dl.acm.org/doi/10.1145/3544548.3580881 [doi: 10.1145/3544548.3580881]
40. Burton E, Lewin G, Boldy D. Physical activity preferences of older home care clients. Int J Older People Nurs 2015 Sep;10(3):170-178. [doi: 10.1111/opn.12065] [Medline: 25400233]
41. Burton E, Farrier K, Galvin R, Johnson S, Horgan NF, Warters A, et al. Physical activity programs for older people in the community receiving home care services: systematic review and meta-analysis. Clin Interv Aging 2019;14:1045-1064 [FREE Full text] [doi: 10.2147/CIA.S205019] [Medline: 31239654]
42. Shalev A, Ringel JB, Riegel B, Vellone E, Stawnychy MA, Safford M, et al. Does connectedness matter? The association between mutuality and job satisfaction among home health aides caring for adults with heart failure. J Appl Gerontol 2023 Apr;42(4):747-757 [FREE Full text] [doi: 10.1177/07334648221146772] [Medline: 36541188]
43. Fair pay for home care. New York Caring Majority. URL: https://www.nycaringmajority.org [accessed 2025-04-28]
44. Dell N, Vaidyanathan V, Medhi I, Cutrell E, Thies W. "Yours is better!": participant response bias in HCI. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2012 Presented at: CHI '12; May 5-10, 2012; Austin, TX p. 1321-1330 URL: https://dl.acm.org/doi/10.1145/2207676.2208589 [doi: 10.1145/2207676.2208589]

Abbreviations

1199SEIU-TEF: 1199 Service Employees International Union Training and Employment Fund
HHA: home health aide or attendant
NIH: National Institutes of Health
PA: physical activity
SEIU: Service Employees International Union

Edited by A Stone; submitted 14.May.2025; peer-reviewed by F Materia, LF Belo; comments to author 14.Oct.2025; accepted 23.Dec.2025; published 26.Jan.2026.

Please cite as:
Solano-Kamaiko IR, Dicpinigaitis M, Tan M, Yang I, Cheng K, Peramsetty R, Shum M, Escamilla Y, Bayly J, Turchioe MR, Avgar A, Vashistha A, Dell N, Sterling MR
Feasibility, Acceptability, and Perspectives Regarding the Use of Activity Tracking Wearable Devices Among Home Health Aides: Mixed Methods Study
J Med Internet Res 2026;28:e77510
URL: https://www.jmir.org/2026/1/e77510
doi:10.2196/77510
PMID:

©Ian René Solano-Kamaiko, Michael Dicpinigaitis, Melissa Tan, Irene Yang, Kexin Cheng, Ronica Peramsetty, Michelle Shum, Yanira Escamilla, Jennifer Bayly, Meghan Reading Turchioe, Ariel Avgar, Aditya Vashistha, Nicola Dell, Madeline R Sterling. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Characteristics Influencing Support for the National Health Service COVID-19 App in England and Wales: Findings From a Longitudinal Survey

Josephine Exley[1], PhD; Paul Boadu[1], PhD; Kasim Allel[2], PhD; Bob Erens[1], MA; Nicholas Mays[1], MA; Mustafa Al-Haboubi[1], PhD

[1]London School of Hygiene & Tropical Medicine, London, United Kingdom

[2]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom

Corresponding Author:
Mustafa Al-Haboubi, PhD
London School of Hygiene & Tropical Medicine
15-17 Tavistock Place
London, WC1H 9SH
United Kingdom
Phone: 44 (0)20 7636 ext 8636
Email: Mustafa.Al-Haboubi@lshtm.ac.uk

Abstract

Background: The use of proximity (contact) tracing mobile phone apps during the COVID-19 pandemic to support manual contact tracing was novel. Uptake of the app was lower than expected.

Objective: We sought to identify distinct subgroups of individuals based on their level of support for the National Health Service (NHS) COVID-19 app in the first 15 months of the app's implementation, and to identify the attitudes and characteristics associated with membership of more and less supportive groups.

Methods: We conducted 8 waves of a longitudinal survey data of smartphone users, recruited from an online panel (n=2023 at baseline and n=1198 at survey wave 6) between October 14, 2020, and December 13, 2021. We used latent class analysis to identify subgroups of individuals with different inclinations of support for the NHS COVID-19 app. Sankey diagram analysis was used to assess individuals whose subgroup changed over the study period. We estimated population-weighted multinomial logistic regression models using sociodemographic characteristics as independent variables.

Results: We identified 4 subgroups in survey waves 1 to 4—"not supportive" (1765/7210, 25%), "ambivalent" (2124/7210, 30%), "somewhat supportive" (1421/7219, 20%), and "completely supportive" (1900/7210, 26%). At wave 5, a total of 3 subgroups of support for the app emerged—"not supportive" (549/1613, 34%), "ambivalent" (497/1613, 31%), and "supportive" (567/1613, 35%). From wave 6 onward, the results showed 4 subgroups emerging—"least supportive" (1568/6952, 23%), "less supportive" (1179/6952, 17%), "ambivalent" (2105/6952, 30%), and "supportive" (2100/6952, 29%). The majority of respondents remained within their identified subgroups between survey waves. Among those who moved into different subgroups, most moved into a less supportive subgroup. Exceptions to this were from waves 2 to 3 and from waves 3 to 4, when higher percentages of respondents moved into more supportive subgroups. The biggest movement to less supportive subgroups occurred after wave 1 (October 2020), when 38% (2740/7210) of respondents moved into a less supportive subgroup. The biggest movement to more supportive subgroups, on the other hand, occurred after wave 2, when 22% (1586/7210) of respondents moved into more supportive subgroups. Over the course of the 8 waves, the percentage of respondents in supportive subgroups declined from 56% (3353/5988) to 29% (1737/5988). Key characteristics of more supportive individuals included having higher levels of trust in the government to control the spread of COVID-19 and having the app installed, while those less concerned about the risk COVID-19 posed to the country were more likely to be unsupportive (P<.05).

Conclusions: When the app was launched, just over half of respondents were supportive, but this declined over the following 15 months. The attrition in support poses important challenges for governments to the use of apps in future pandemics. A potential reason was mistrust in the government's handling of the pandemic.

(J Med Internet Res 2026;28:e76863) doi:10.2196/76863

XSL•FO
RenderX

KEYWORDS

contact tracing apps; COVID-19; infection prevention and control; public health; test and trace

Introduction

As part of the response to the COVID-19 pandemic, several countries deployed novel proximity (contact) tracing mobile phone apps to supplement manual contact tracing efforts. They keep a temporary record of app users who have been in close contact with each other to alert them in case one user subsequently reports through the app contracting the virus. Evidence highlights the potential for apps to reduce the spread of COVID-19 [1-7]. For example, the National Health Service (NHS) COVID-19 app was estimated to have averted approximately 1 million cases and 10,000 deaths in England and Wales in its first year [8], and for every percentage increase in uptake the number of cases was estimated to have been reduced by up to 2% [9].

The effectiveness of such technology depends on the level of uptake among the population [1,10-12]. In settings where their use was not mandated, uptake was reliant on individuals' willingness to install and use the app. However, uptake of contact tracing apps in many countries was found to be lower than anticipated [13-17], although surveys conducted before the introduction of apps suggested high levels of public support for the use of these apps as part of the pandemic response [18,19]. Evidence from our own work following a cohort of smartphone users in England and Wales during the first year of implementation of the NHS COVID-19 app found that despite changes in policy and case numbers, installation of the app remained relatively stable at around 50% [20]. The majority of those who had ever installed the app did so soon after the app's launch.

With scientists predicting a growing probability of future pandemics [21], contact tracing apps are likely to become an increasingly ubiquitous public health tool. As such, there is a need to learn how to improve the implementation and uptake of such apps in the face of likely future pandemics. In this analysis of a longitudinal 8-wave survey of a representative sample of adult smartphone users in England and Wales, we aimed to identify distinct subgroups of individuals based on their level of support for the NHS COVID-19 app in the first 15 months of the app's implementation and to identify the attitudes and characteristics associated with membership of more and less supportive groups.

Methods

This study draws on longitudinal (prospective cohort) survey data of smartphone users aged 18 to 79 years in England and Wales. Details of the study have been described previously [20].

Setting

The NHS COVID-19 app became part of the NHS Test and Trace Programme in July 2020 and was launched in England and Wales on September 24, 2020 (Textbox 1). It alerted close contacts of individuals who later tested positive for coronavirus to self-isolate. Additional features allowed users to check their symptoms, book a COVID-19 test, and check in at a venue. The survey started 2 weeks after the app was launched and coincided with rising COVID-19 cases and the tightening of COVID-19 restrictions. Multimedia Appendix 1 maps the 8 rounds of the online survey against the number of new COVID-19 cases reported and the key policy changes introduced by the governments in England and Wales during the study period. COVID-19 cases were highest during survey wave 3 (conducted between December 28, 2020, and January 6, 2021) and survey wave 8 (conducted between November 25, 2021, and December 13, 2021). Cases were lowest during survey wave 5 (conducted between March 15 and 31, 2021).

Textbox 1. Overview of the National Health Service (NHS) COVID-19 app and the key COVID-19 restrictions in place in England and Wales during the first year of the app's launch.

The NHS COVID-19 contact tracing app, using a decentralized model, was launched in England and Wales on September 24, 2020, for users aged 16 years and older and was made available in 12 languages.

It required a recent smartphone operating system and only collected users' postcode areas to provide local risk information. By July 2022, it had over 31 million downloads [22]. The app sent anonymous alerts to close contacts (determined by proximity and duration algorithms) of users who reported positive COVID-19 tests in the app [23]. Unlike public health authorities' self-isolation instructions (which ended in February 2022), app guidance was not legally binding. Users needed Bluetooth enabled for contact tracing, with an option within the app to temporarily disable it. Other features included symptom checking, test ordering, venue check-ins by QR codes (mandatory in England from September 2020 to February 2022 and a recommendation in Wales), and a self-isolation timer.

The app's launch coincided with rising COVID-19 cases and the tightening of restrictions [24]. Wales introduced a 2-week firebreak on October 23, while England entered a second national lockdown on November 5. Both countries briefly returned to tiered restrictions in December before reimposing lockdowns by year-end. In early 2021, England entered a third lockdown, and Wales maintained the highest restriction level ("stay at home"). Restrictions were slowly eased throughout spring and summer 2021, with most lifted by July 19, including the need to self-isolate if double vaccinated. The emergence of the Omicron variant in November 2021 once again marked the tightening of restrictions.

As shown in Multimedia Appendix 1, survey wave 1 (October 14-22, 2020) took place after the introduction of the local 3-tier system of COVID-19 restrictions; at that time, it was a legal requirement for venues in England to take the contact details of visitors either manually or by scanning a QR code, while in Wales this was a recommendation only. Survey wave 2 (November 12-23, 2020) was conducted during the second national lockdown, introduced on November 5, and followed a 2-week firebreak in Wales from October 23 to November 8 in which the public were instructed to stay at home and the

XSL•FO
RenderX

hospitality industry and nonessential restaurants had to close. Survey wave 3 (December 28, 2020, to January 6, 2021) was conducted over the Christmas period when the country had returned to the tier system; the southeast of England and the whole of Wales moved to a newly created tier 4 ("stay at home"), the first vaccine was administered in the United Kingdom on December 8, 2020, and England entered the third national lockdown on January 6, 2021. Survey wave 4 (February 1-15, 2021) occurred while England remained in the third national lockdown and Wales remained at tier 4. Survey wave 5 (March 15-31, 2021) took place at the start of the easing of restrictions, after the government in England outlined a 4-step plan; by the start of the survey, schools had reopened and outdoor mixing in groups of up to 6 was allowed from March 29. Survey wave 6 (July 1-18, 2021) occurred ahead of the final step of easing restrictions in both England and Wales, when indoor venues had reopened and the public were allowed to meet in groups of up to 30 outdoors and 6 indoors. Survey wave 7 (August 31-September 13, 2021) took place after all legal limits on social contact had been lifted, including the requirement to check in to venues in England, and those who had been double vaccinated were no longer required to self-isolate if they had come into contact with someone who tested positive for COVID-19, provided they did not have any symptoms. Survey wave 8 (November 25-December 13, 2021) occurred at the same time as the Omicron variant was first being reported; a total of 6 southern African countries were placed on the travel red list and the first cases were detected in the United Kingdom, and on December 8 the government announced that Plan B measures were to be reintroduced, including mandatory face coverings in most indoor settings, working from home, and use of the NHS COVID Pass for entry into nightclubs and settings where large crowds gather. Data on daily new cases were obtained from the UK Coronavirus Dashboard, and information on lockdown and restriction timelines from the Institute for Government's coronavirus timeline.

Participants

We were interested in looking at differences in the usage of the app between different sociodemographic groups, such as age, level of education, social grade, health status, and minoritized ethnic groups. We expected to find differences in attitudes between members of our sample in different social grades (nonmanual vs manual). In order to detect a 6% difference (at 80% power and a 95% significance level) in attitudes between these 2 subgroups, a sample size of 1657 is required. More importantly, we wanted to detect differences in app use over time. Assuming about two-thirds of the sample at baseline were using the app, we would be able to detect about a 3% change in the overall sample between baseline/wave 1 and survey wave 2 with 95% confidence.

A representative sample of smartphone users aged 18 to 79 years was recruited through YouGov's volunteer online panel, with quotas set on age, gender, and social grade. Panel members are recruited from different sources, including through advertising and partnerships with a range of websites. Sociodemographic data are collected when they join the panel. Participants are invited from the panel in a way that seeks to generate a nationally-representative sample [25].

In total, 2023 participants were recruited at the start of the study (referred to as the "baseline sample"). Panel members who completed the baseline survey were invited to all follow-up waves. After the fifth round of data collection, we recruited an additional 1198 participants to increase the sample size (referred to as the "additional sample"). At survey wave 8, 61% (1233/2023) of the main sample and 71% (848/1198) of the additional sample responded. The response rate for each sample at each survey wave is provided in Table S1 in Multimedia Appendix 2.

Data Collection and Variables of Interest

Online surveys were undertaken roughly every 6 weeks between October 14, 2020, and December 13, 2021 (Multimedia Appendix 1). The survey was developed by the study team and adapted from previous surveys. It included questions on attitudes toward the app, including the level of support for it (measured on a 5-point Likert scale from 1 "not at all supportive" to 5 "completely supportive"), as well use of the app, perceptions of COVID-19, governments' response to the pandemic, and demographic information (ie, age, sex, ethnicity, income, employment, etc). The core content of the questionnaire remained the same across all waves, but some questions were added or removed in response to the evolving pandemic.

The survey was administered by YouGov and sent to respondents by email. Participants who completed the first survey (survey wave 1 for the baseline sample or wave 6 for the additional sample) were invited to participate in all subsequent waves of the survey.

Analysis

Determining Latent Classes of Support for NHS COVID-19 App

We used latent class analysis (LCA) to identify distinct subgroups of individuals with similar levels of support for the app at each survey wave [26] using the *gsem* function in Stata [27]. We considered LCA an optimal strategy due to its robustness against potential measurement errors, such as potential biases introduced by the timing of the questionnaire, and its ability to parsimoniously model restricted latent groups across multiple levels of support. This approach also helps prevent measurement errors that may arise from assigning respondents to groups based on their survey responses without recourse to statistical justification. To determine the optimal number of classes to include in the LCA model at each survey wave, we estimated the Bayesian information criteria (BIC) and Akaike information criteria (AIC) maximum likelihood values, with lower values indicating a better fit for the number of subgroups [28,29]. The results were similar when the test was conducted with and without covariates, including general health, disability, vulnerability to COVID-19, region, household income, tenure, app installation, trust in government, age, ethnicity, and sex. The results of the maximum likelihood test are presented in Table S2 in Multimedia Appendix 2.

Sankey Diagram Analysis

To assess the movement of individuals between subgroups of support for the app over survey waves, we constructed a Sankey

XSL•FO
RenderX

diagram of the wave-specific estimated classes and examined individual transitions between identified support subgroups over the survey period [30,31]. The analysis was restricted to individuals who responded to 2 consecutive waves.

Regression Analysis

To examine the factors associated with belonging to a given subgroup of support for the app, we estimated population-weighted multinomial logistic regression models at three timepoints: survey wave 1, survey wave 5, and survey wave 8. We opted to estimate the regression model at the study start and endpoints, but additionally included wave 5 as it represents a notable change in the distribution of LCA results, in that the number of subgroups reduced from 4 to 3. The definition of dependent and independent variables included in the multinomial logistic regression model are presented in Table S3 in Multimedia Appendix 2.

To select the independent variables for the model and mitigate multicollinearity, a correlation matrix (Tables S4-S6 in Multimedia Appendix 2) and variance inflation factors (VIF) (Tables S7-S9 in Multimedia Appendix 2) were generated. Based on the results, the following explanatory variables were included in the model: age, gender, ethnicity, self-reported health status, whether day-to-day activities were limited because of a health problem or disability, perceived vulnerability to COVID-19, region of residence, household income, household ownership, whether the app was installed at the current survey wave, whether the respondent had a COVID-19 infection since the previous survey, extent of trust in the government to control the spread of COVID-19, extent of concern about the risk COVID-19 poses to oneself, and extent of concern about the risk COVID-19 poses to the country. Fisher statistics are reported for measuring the model's goodness-of-fit. All statistical analyses were performed using Stata Standard Edition (version 18; StataCorp LLC) software. The research is reported in line with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) Statement for cohort studies [32] (Multimedia Appendix 3).

Ethical Considerations

Consent was sought before starting the first survey but not at subsequent waves. Responses were anonymized by YouGov before being passed to the researchers. Participants were free to withdraw at any time without needing to provide a reason. Ethics approval was obtained from the London School of Hygiene & Tropical Medicine Research Ethics Committee (reference number 22483). Participants receive "points" for completing surveys, which can subsequently be converted into cash payments.

Results

Overview

Overall, a total of 3221 participants were recruited to the study (2023 at survey wave 1 and 1198 at survey wave 6). The characteristics of participants are presented in Table 1. About 53% (1118/2023) of participants at wave 1 and 53% (619/1198) of participants at wave 6 were female, 42% (810/2023) and 39% (467/1198) were aged 40 years or younger, 86% (2775/2023) and 87% (1045/1198) were White, and 61% (1235/2023) and 57% (679/1198) had received higher education in the baseline and additional samples, respectively.

XSL•FO
RenderX

Table 1. Characteristics of respondents to an 8-wave longitudinal survey on use and views on the National Health Service (NHS) COVID-19 app (October 2020 to December 2021). A currency exchange rate of GBP £1=US $1.32 is applicable.

Characteristics	Baseline sample (n=2023), n (%)	Additional sample (n=1198), n (%)
Sex		
Male	905 (47.2)	579 (48.3)
Female	1118 (52.8)	619 (51.7)
Age group (years)		
18-29	439 (23.1)	218 (18.2)
30-39	371 (18.4)	249 (20.8)
40-49	381 (20.8)	220 (18.4)
50-59	381 (17.5)	210 (17.5)
60-69	298 (12.1)	192 (16.0)
70-79	153 (8.1)	109 (9.1)
Ethnicity		
White	1775 (86.1)	1045 (87.2)
All other ethnic groups	248 (13.9)	153 (12.8)
Region		
North East	103 (4.4)	30 (2.5)
North West	252 (12.4)	102 (8.5)
Yorkshire and the Humber	185 (9.1)	94 (7.8)
East Midlands	174 (8.0)	74 (6.2)
West Midlands	176 (9.9)	63 (5.3)
East of England	209 (10.4)	84 (7.0)
London	283 (15.2)	163 (13.6)
South East	320 (15.8)	101 (8.4)
South West	197 (9.5)	69 (5.8)
Wales	124 (5.4)	418 (34.9)
Highest level of education attainment		
No formal qualifications	78 (3.5)	42 (3.5)
GCSE[a] or equivalent	232 (11.3)	143 (11.9)
A-level or equivalent	293 (15.6)	200 (16.7)
Higher education	1235 (60.8)	679 (56.7)
Other	137 (6.4)	92 (7.7)
Prefer not to say or did not answer	48 (2.5)	42 (3.5)
Employment status		
Currently working	1151 (56.8)	693 (57.9)
Not currently working	92 (4.7)	33 (2.8)
Voluntary work or carer	96 (4.6)	58 (4.8)
Unemployed or permanent sick leave	222 (11.0)	102 (8.5)
Education	81 (5.7)	80 (6.7)
Retired	344 (15.6)	215 (18.0)
Other	37 (1.7)	17 (1.4)
Household income (£)		
<14,999	224 (10.8)	130 (10.9)

XSL•FO
RenderX

Characteristics	Baseline sample (n=2023), n (%)	Additional sample (n=1198), n (%)
15,000-24,999	285 (14.0)	197 (16.4)
25,000-34,999	251 (12.7)	154 (12.9)
35,000-60,000	431 (20.9)	227 (19.0)
>60,000	346 (17.5)	188 (15.7)
Prefer not to say or did not answer	486 (24.1)	302 (25.2)
Personal income (£)		
<14,999	510 (25.7)	300 (25.0)
15,000-24,999	382 (18.4)	237 (19.8)
25,000-34,999	319 (15.6)	157 (13.1)
35,000-60,000	267 (12.7)	154 (12.9)
>60,000	107 (5.5)	43 (3.6)
Prefer not to say or did not answer	438 (22.0)	307 (25.6)
IMD[b]		
1–most deprived	351 (17.9)	218 (18.2)
2	373 (18.0)	246 (20.6)
3	408 (20.3)	233 (19.5)
4	442 (21.3)	231 (19.3)
5–least deprived	448 (22.5)	269 (22.5)
Living arrangements		
lives alone	372 (18.3)	211 (17.6)
other adult(s), no children	1151 (55.4)	633 (52.8)
children, no other adults	49 (2.4)	32 (2.7)
other adult(s) and children	451 (23.9)	322 (26.9)
Household ownership		
Own	1169 (55.9)	683 (60.1)
Rent	589 (29.5)	292 (25.7)
Live with friends or family	233 (13.0)	133 (11.7)
Other	32 (1.6)	28 (2.5)
Self-reported health status		
Very good	535 (27.1)	297 (24.8)
Good	988 (49.2)	572 (47.8)
Fair	352 (16.6)	234 (19.5)
Bad or very bad	117 (5.4)	72 (6.0)
Not answered	31 (1.7)	23 (1.9)
Day-to-day activities limited because of a health problem or disability lasting (or expected to last) at least 12 months		
Limited a lot	159 (7.3)	108 (9.0)
Limited a little	325 (15.5)	195 (16.3)
No	1498 (74.9)	863 (72.0)
Not answered	41 (2.2)	32 (2.7)

[a]GCSE: General Certificate of Secondary Education.

[b]IMD: index of multiple deprivation based on respondent's usual place of residence [33].

XSL•FO
RenderX

Latent Classes of Support for the App

The test results (BIC/AIC) of the number of latent classes show 4 subgroups of population with differences in support for the NHS COVID-19 app for all survey waves, except for wave 5, where 3 subgroups were identified (Table S2 in Multimedia Appendix 2). The distribution of the identified subgroups in each survey wave is presented in Table 2.

Table 2. Distribution of subgroups of support for the National Health Service (NHS) COVID-19 app across 8 survey waves (W1–W8) from an 8-wave longitudinal survey conducted between October 2020 and December 2021 (n=number of observations per wave; W=survey wave).

Level of support	W1 (Oct 14-22, 2020; n=2023), %	W2 (Nov 12-23, 2020; n=1781), %	W3 (Dec 28, 2020-Jan 6, 2021; n=1732), %	W4 (Feb 1-15, 2021; n=1674), %	W5 (Mar 15-31, 2021; n=1613), %	W6 (Jul 1-18, 2021; n=2667), %	W7 (Aug 31-Sep 13, 2021; n=2204), %	W8 (Nov 25-Dec 13, 2021; n=2081), %
Least supportive	16[a]	28[a]	28[a]	27[a]	25[a]	20	23	25
Less supportive	—	—	—	—	—	17	18	16
Ambivalent	28	32	29	29	36	32	29	30
Somewhat supportive	20	19	20	19	39[b]	32[b]	29[b]	29[b]
Completely supportive	36	21	22	25	—	—	—	—

[b]This percentage value applies to both the “least supportive” and “less supportive” subgroups; in Waves 1-5 these categories were statistically measured as a single latent class.

[b]This percentage value applies to both the “somewhat supportive” and “completely supportive” subgroups; in Waves 5-8 these categories were statistically measured as a single latent class.

This percentage value applies to both the “least supportive” and “less supportive” subgroups; in Waves 1-5 these categories were measured as a single latent class.

The characteristic of individuals belonging to the identified subgroups were similar for survey waves 1-4 and survey waves 6-8, so they were presented and analyzed together.

From survey waves 1 to 4 (October 14, 2020, to February 15, 2021), we identified 4 subgroups of support for the app. Table 3 shows the characteristics of the 4 population subgroups with different levels of support for the NHS COVID-19 app. Individuals in subgroup 1 were either not supportive at all or least supportive of the app, with good health and no disability, not vulnerable to COVID-19, generally equally distributed across all income groups and regions, and the majority were homeowners. Most had never installed the app and had no trust in the government to control the spread of COVID-19. The average age of individuals in subgroup 1 was 46 years, and they comprised similar proportions of individuals from all ethnic groups, with higher proportions of males. Based on these characteristics, we labeled this population subgroup as “not supportive.”

XSL•FO
RenderX

Table 3. Summary of characteristics of individuals belonging to the 4 subgroups of support for the National Health Service (NHS) COVID-19 app (survey wave 1 to wave 4 [October 14, 2020, to February 15, 2021]). Table S10 in Multimedia Appendix 2 provides underlying statistics.

Characteristic	Subgroup 1: Not Supportive	Subgroup 2: Ambivalent	Subgroup 3: Somewhat Supportive	Subgroup 4: Completely Supportive
Support	Not at all supportive, 51% (901/1765) or least supportive, 49% (863/1765)	Indifferent about support for the app, 100% (2124/2124)	Somewhat supportive, 100% (1421/1421)	Completely supportive of the app, 100% (1900/1900)
General health	Majority in good health, 52% (813/1765)	Majority in good health, 53% (1134/2124)	Majority in good health, 53% (754/1421)	Higher proportion with good health, 46% (871/1900)
Disability	Majority without disability, 75% (1317/1765)	Majority without disability, 75% (1598/2124)	Majority without disability, 75% (1060/1421)	Majority without disability, 72% (1359/1900)
Vulnerability to COVID-19	Majority not vulnerable, 59% (1039/1765)	Majority not vulnerable, 59% (1258/2124)	Majority not vulnerable, 57% (815/1421)	Majority vulnerable, 51% (960/1900)
Region	Generally evenly spread across all regions	Generally evenly spread across all regions	Generally evenly spread across all regions	Generally evenly spread across all regions
Household income	Generally evenly spread across income groups	Generally evenly spread across income groups	Generally evenly spread across income groups	Generally evenly spread across income groups
Tenure	Majority homeowners, 58% (1015/1765)	Majority homeowners, 56% (1181/2124)	Majority homeowners, 58% (831/1421)	Majority homeowners, 64% (1213/1900)
App installation	Majority never installed, 62% (1093/1765)	Higher proportion never installed, 44% (935/2124)	Majority installed, 59% (845/1421)	Majority installed, 74% (1409/1900)
Trust	Higher proportion with not at all trust, 84% (854/1765)	Higher proportion with little trust, 38% (803/2124)	Higher proportion with little trust, 37% (521/1421)	Higher proportion with fair amount of trust, 34% (646/1900)
Age (average)	46 years	45 years	45 years	49 years
Ethnicity	Similar proportion across all ethnic groups	Higher proportions among other ethnic groups, 35% (298/855)	Similar proportion across all ethnic groups	Higher proportions among White individuals, 27% (1714/6355)
Sex	Higher proportion of males, 27% (859/3232)	Higher proportion of females, 33% (1281/3978)	Similar proportions of males and females	Higher proportion of males, 28% (890/3232)

Most of the individuals in subgroup 2 neither supported nor opposed the NHS COVID-19 app and indicated that they had good health and were not vulnerable to COVID-19. Most were homeowners, had never installed the app, and did not have very much trust in the government to control the spread of COVID-19. Most were from ethnic groups other than White, with a higher proportion being female, and their average age was 45 years.

They were equally distributed across regions and household incomes, with a higher proportion being female. We labeled this subgroup “Ambivalent” based on their characteristics.

The third subgroup was labeled “Somewhat supportive” because individuals in this group had some level of support for the app. Most were in good health, had no disability, and did not consider themselves vulnerable to COVID-19. A higher proportion had never installed the app and did not have much trust in the government to control the spread of COVID-19. Individuals in this subgroup were generally equally distributed across regions, income groups, sex, and ethnic groups, and the majority were homeowners, with an average age of 45 years.

Individuals in the fourth subgroup were completely supportive of the NHS COVID-19 app, a higher proportion were in good health, and the majority had no disability. However, the majority also indicated they were vulnerable to COVID-19, were homeowners, had installed the app, and had a fair amount of trust in the government to control the spread of COVID-19. Similar to the other subgroups, individuals in this subgroup were generally equally distributed across all income groups and regions. A higher proportion of individuals in this group were of White ethnic group and were male. The average age of this subgroup was 49 years.

The number of subgroups identified in wave 5 (March 15-31, 2021) was reduced by one to 3. Table 4 shows the characteristics of individuals in each of the identified subgroups.

XSL•FO
RenderX

Table 4. Summary of characteristics of individuals belonging to the 3 subgroups of support for the National Health Service (NHS) COVID-19 app (survey wave 5 [March 15-31, 2021]). Table S11 in Multimedia Appendix 2 provides underlying statistics.

Characteristic	Subgroup 1: Not Supportive	Subgroup 2: Ambivalent	Subgroup 3: Supportive
Support	Not at all supportive, 52% (288/549) or least supportive, 48% (261/549)	Indifferent about support for app 100% (497/497)	Somewhat supportive, 46% (258/567) or completely supportive of the app, 54% (309/567)
App installation	Majority never installed, 56% (308/549)	Higher proportion either installed, 41% (205/497) or never uninstalled, 37% (186/497)	Majority installed, 73% (413/567)
Trust	Higher proportion with no trust at all, 35% (193/549)	Higher proportion with a fair amount of trust, 41% (202/497)	Higher proportion with a fair amount of trust, 42% (236/567)
Age (average)	47 years	45 years	48 years
Sex	Higher proportion of males, 37% (268/721)	Higher proportion of females, 34% (306/892)	Higher proportion of males, 36% (262/721)

Unlike the subgroups identified from study waves 1-4 (October 14, 2020, to February 15, 2021), characteristics such as general health, disability status, vulnerability to COVID-19, region, household income, and tenure were not statistically different among the 3 identified subgroups in wave 5. Based on the characteristics of individuals in subgroup 1, we labeled it "Not supportive." Those in this subgroup were not at all or least supportive of the NHS COVID-19 app, the majority had never installed the app, and a higher proportion had no trust at all in the government to control the spread of COVID-19. The average age of this subgroup was 47 years, and a higher proportion were male. Following a similar approach, subgroups 2 and 3 were labeled "Ambivalent" and "Supportive," respectively.

Furthermore, the results from survey waves 6-8 (July 1, 2021, to December 13, 2021) show the reemergence of 4 subgroups of support for the NHS COVID-19 app, with some subgroups showing different characteristics than those identified in survey waves 1-4 (Table 5). Based on the characteristics of the individuals in the identified subgroups, we labeled subgroups 1, 2, 3, and 4 as "Least supportive," "Less supportive," "Ambivalent," and "Supportive," respectively (Table 5).

Table 5. Summary of characteristics of individuals belonging to the 4 subgroups of support for the National Health Service (NHS) COVID-19 app (survey waves 6-8 [July 1, 2021, to December 13, 2021]). Table S12 in Multimedia Appendix 2 provides underlying statistics.

Characteristic	Least Supportive	Less Supportive	Ambivalent	Supportive
Support	Not at all supportive, 100% (1568/1568)	Least supportive, 100% (1179/1179)	Indifferent about support for app, 100% (2105/2105)	Somewhat supportive, 48% (999/2100) or completely supportive of app, 52% (1101/2100)
General health	Higher proportion in good health, 48% (745/1568)	Higher proportion in good health, 51% (607/1179)	Higher proportion in good health, 50% (1043/2105)	Higher proportion in good health, 48% (1016/2100)
Disability	Majority without disability, 71% (1109/1568)	Majority without disability, 77% (903/1179)	Majority without disability, 70% (1484/2105)	Majority without disability, 69% (1444/2100)
Region	Higher proportion in Wales, 24% (318/1301)	Higher proportions in London and South and the North	Higher proportion in Midlands and East of England	Higher proportion in London and South and Wales
Household income	Generally evenly spread across income groups	Generally evenly spread across income groups	Generally evenly spread across income groups	Generally evenly spread across income groups
Tenure	Majority homeowners, 64% (1008/1568)	Majority homeowners, 61% (721/1179)	Majority homeowners, 59% (1246/2105)	Majority homeowners, 64% (1335/2100)
App installation	Majority never installed, 68% (1074/1568)	Higher proportion never installed, 47% (549/1179)	Higher proportion installed, 43% (901/2105)	Majority installed, 76% (1604/2100)
Trust	Higher proportion with no trust at all, 40% (630/1568)	Higher proportion with little trust, 37% (436/1179)	Higher proportion with a fair amount of trust, 37% (781/2105)	Higher proportion with a fair amount of trust, 40% (834/2100)
Age (average)	49 years	46 years	48 years	49 years
Sex	Higher proportion of males, 26% (856/3289)	Higher proportion of females, 18% (675/3663)	Higher proportion of females, 33% (1216/2105)	Higher proportion of males, 32% (1040/3289)

The Sankey diagram presented in Multimedia Appendix 4 shows the movement of respondents between classes over the course of the surveys. Between each survey wave, more than half of respondents did not change class. Among those who did, most moved into a less supportive class (measured by the thickness of the palette), except between survey waves 2 to 3 and 3 to 4,

XSL•FO
RenderX

where more individuals moved upwards into more supportive classes than downwards. The biggest movement downwards occurred after survey wave 1 (October 2020) when 38% (275/723) of individuals moved into a less supportive class, half of whom (130/723, 18%) moved out of the completely supportive class to the somewhat supportive class, and over a third (108/723, 15%) moved into the Ambivalent subgroup. The biggest movement upwards occurred after survey wave 2 (November 2020) with 22% (233/1060) of individuals moving to more supportive classes. As a result of the greater number moving into less supportive classes, the percentage of individuals in "supportive" subgroups decreased over time, from 56% (1131/2023; "completely" or "somewhat supportive") at survey wave 1 to 29% (611/2081; "supportive") by survey wave 8.

Factors Associated With Class Membership

To quantify the factors associated with the likelihood of an individual's membership in a particular subgroup at survey waves 1 (October 2020), 5 (March 2021), and 8 (December 2021), we calculated the relative risk ratio (RRR) of variables being present in a particular subgroup relative to the reference group (completely supportive). An RRR<1 suggests a lower likelihood relative to the reference group, while an RRR>1 indicates a higher likelihood of the variable's presence. The full results of the multinomial logistic regression are presented in Tables S13-S15 in Multimedia Appendix 2 and summarized in Multimedia Appendix 5.

At all 3 waves, individuals who never had the app installed and had less trust in government to control the spread of COVID-19 were associated with a relative reduction in the likelihood of being in the most supportive subgroups compared with all other subgroups. For example, never having the app installed reduced the likelihood of being in the most supportive group compared with the least supportive group (RRR 0.03, 95% CI 0.02-0.05 at survey wave 1; RRR 0.07, 95% CI 0.06-0.10 at survey wave 5; and RRR 0.03, 95% CI 0.02-0.05 at survey wave 8). Similarly, having little or no trust in the government to control the spread of COVID-19 reduced the likelihood of being in the most supportive group compared with the least supportive group (RRR 0.19, 95% CI 0.12-0.29 at survey wave 1; RRR 0.27, 95% CI 0.20-0.38 at survey wave 5; and RRR 0.30, 95% CI 0.19-0.47 at survey wave 8).

At all 3 waves, not being at all concerned about the risks COVID-19 posed to the country was associated with a reduced likelihood of being in the most supportive subgroups compared with the least supportive subgroup. Compared with individuals who were very concerned about the risks COVID-19 posed to the country, being not at all concerned increased the likelihood of being in the "not supportive" group at survey wave 1 (RRR 10.87, 95% CI 2.63-44.85) and survey wave 5 (RRR 9.92, 95% CI 3.03-32.48) and increased the likelihood of being in the "least supportive" subgroup at survey wave 8 (RRR 13.72, 95% CI 2.56-73.62).

Discussion

Overview

We examined how levels of support for the COVID-19 app changed in the 15 months following its launch in England and Wales. We found just over half of respondents were supportive of the app around the time of its launch and that support declined over time, with a notable drop-off in those who were completely supportive occurring between the first (October 14-22, 2020) and second (November 12-23, 2020) survey wave. From survey wave 6 (July 1-18, 2021), more respondents were unsupportive than supportive of the app. Key characteristics of individuals who were more supportive included having higher levels of trust in the government to control the spread of COVID-19, having the app installed, and being more concerned about the risk COVID-19 posed to the country.

The level of support observed in this study was lower than might have been expected based on studies conducted before the app was launched [18,19]. A contact tracing app was initially framed as a central component of the government's strategy to control the spread of COVID-19, and media coverage in early 2020 was positive [34,35]. Yet following technical issues during piloting, including abandoning the first design [36], the government's framing of the app changed from a technological solution to a more experimental technology [35]. The development process received prominent critical media coverage, which often presented contact tracing apps as controversial [37] and narrowed the debate to privacy concerns around what access authorities would have to personal data on users' phones [38,39]. Among the cohort included in this study, being worried about privacy was the key reason for not downloading the app [20].

Lower levels of support by the time of the national launch also likely reflect wider public perceptions of the government's handling of the pandemic. By autumn 2020, trust in the government to handle the pandemic was low [40,41], and its launch coincided with rising COVID-19 cases and the tightening of restrictions (Multimedia Appendix 1). Soon after the app's launch, a range of new restrictions were introduced, including a second national lockdown and a modified tiered system [42]. The government's approach was widely characterized as confusing and chaotic [43,44], exemplified by the last-minute scrapping of plans in England to relax restrictions over the 2020 Christmas period, and likely contributed to negative attitudes toward the government's handling of the pandemic. In turn, this may have contributed to the perception that the app was ineffective at controlling the spread of COVID-19 and did not live up to earlier positive expectations [45,46].

Unlike levels of trust, which were found to fluctuate over the study period [41], levels of support consistently declined. Following the app's launch, individuals' mobility and social contact were highly restricted, with nonessential indoor leisure facilities and outdoor hospitality only reopening in April 2021. This might have fed perceptions that the app had limited utility at an individual level. Evidence on uptake indicates that the venue check-in feature was a driver of uptake [20]. Yet by July 2021, most social distancing requirements (including the

XSL•FO
RenderX

requirement to check in to venues) were dropped. The shift observed at survey wave 6 (July 1-18, 2021), with more people reporting they were unsupportive than supportive for the first time, potentially indicates that individuals were no longer placing much value on the app. Among those who uninstalled the app, not finding it useful was a key reason for doing so [20]. When restrictions were introduced again in autumn 2021 in response to the Omicron variant, the requirement to check in to venues was not reintroduced, which potentially reinforced perceptions that the app did not have a major role to play in controlling the spread of COVID-19.

Declining support might also reflect perceptions around the reliability of the app. Issues experienced by users in the weeks following the app's launch, including "ghost notifications," were widely reported in the mainstream media [47], and analysis of reviews posted in app stores showed that problems associated with the app's functionality were a key driver of negative comments [48,49]. As England and Wales approached so-called "freedom day" in July 2021, there was an increase in app users being told to self-isolate, coined the "pingdemic" in the media, and often reported as a shortcoming of the app rather than the direct result of a spike in infections due to the newly circulating Delta variant [50]. In response, the government decreased the sensitivity of the app [51]. This decision was criticized by the Opposition [52], and it likely reinforced the (incorrectly held) perception that the app was not working effectively.

Individuals who used the app were found to be more supportive throughout the study period. Based on our findings, it is not possible to determine whether being more supportive was a reason for installing the app or whether using the app increased support, though the fact that most people installed the app for the first time around the app's launch in October 2020 suggests that the main direction of effect is likely from support in principle to take up the app. Since individuals who were more concerned about the risk COVID-19 posed to the country were more supportive of the app [53], these individuals potentially prioritized steps to control the spread of the virus over other concerns [54]. Evidence from elsewhere indicates that individuals who were more concerned about the risk of COVID-19 were more likely to engage in protective behaviors, including app use [20,55,56].

Strengths and Limitations

The strength of this study is that it includes 8 waves of data collected over a 15-month period after the launch of the NHS COVID-19 app. At each timepoint, participants were asked about their level of support for the app, allowing us to examine changing trends in support.

The sample was representative of the general population of smartphone users in terms of age (up to 79 years), gender, social grade, and region. The sample were younger and more highly educated than the general population. This partly reflects smartphone ownership, which tends to be higher among younger individuals and those from higher-income households [57], but is also a reflection of the population that participates in nonprobability online panels [58]. One particular weakness of the sample is the underrepresentation of some minoritized groups. This is particularly important in light of evidence that some groups at higher risk of contracting COVID-19 were less likely to use the app [20].

The attrition rate was relatively high, with 61% of the original sample responding at survey wave 8. Nevertheless, we recruited new panel members ahead of survey wave 6 to top up the sample.

The question on support for the app could have been interpreted in different ways, and the results likely capture respondents' support in principle for apps to control the spread of COVID-19 as well as specific support for the app as deployed. Using mixed methods that incorporate qualitative components would enhance surveillance systems by providing deeper insights into the public's perceptions of contact tracing apps and which aspects of the app people are more or less supportive of.

Implications

Evidence demonstrates that apps had a positive impact on controlling the spread of COVID-19 [8,9]. However, uptake of the app was disappointing compared to expectations based on reports from the start of the pandemic, and by the time the app launched, only 36% of our sample were completely supportive. As with many other countries, the deployment of a contact tracing app was novel in England and Wales. As such, some of the teething problems experienced during the development of the app are unlikely to be repeated in the future. However, while most countries, including England and Wales, have ended their COVID-19 contact tracing programs, it is likely that apps will continue to be included in future pandemic preparedness plans.

The results presented in this study provide important lessons and potential strategies as to how the governments in England and Wales might increase the impact of apps in the future. First, there is a need to develop a communication strategy that considers the different phases from app development through deployment and maintenance. In the future, when relying on untested technology, there is a need to be open and transparent during the development phase about the potential risks and benefits [45], avoiding overly optimistic messaging before the technology is proven. To improve public perceptions of effectiveness, an initial step would be to publicize the growing body of evidence that apps made a positive contribution to reducing transmission in the COVID-19 pandemic and be very clear from the outset about the functionality of the app, including what data the app collects and who these are shared with [59]. Given support was strongly influenced by trust in government, messaging might be better delivered by individuals and organizations in whom the public places greater trust [45,60]. More widely, support for an app is likely to depend on the degree to which the government of the day is trusted to be doing a reasonable job in responding to any pandemic or public health emergency.

There is a need to understand the optimal timing for launching an app to increase public perceptions that apps are an effective strategy to control the spread of COVID-19 at the population level, while also benefiting individuals in safely managing their daily lives. Given the wider contextual factors in autumn 2020, the launch of the app may have been mistimed. It might have been easier for the public to understand its benefit and utility if

XSL•FO
RenderX

it had it been introduced as part of the government's strategy to exit the lockdown between March and July 2021. Future research should examine how similar apps could be adapted to improve user experience and satisfaction [48].

Conclusions

In this study among smartphone users in England and Wales, we found that the level of support for the NHS COVID-19 app was highest at launch and declined over the 15 months after rollout. A potential reason for the low and declining level of support was mistrust in both the Welsh and Westminster government's handling of the pandemic, which likely contributed to a lack of support for the technology. The attrition in levels of support observed poses important challenges for the use of apps in future pandemics. However, our findings also show that individuals who installed the app and were more concerned about the risk of COVID-19 to the country were more supportive, suggesting that there is room to build support for apps especially among those concerned about the potential harm of a pandemic to others.

Acknowledgments

The authors would like to thank the survey respondents for their time and effort. No generative AI was used in any portion of the manuscript writing.

Data Availability

The datasets generated or analyzed during this study are available on reasonable request. Requests should be directed to PIRU-pm@lshtm.ac.uk.

Funding

This research was funded by the National Institute for Health and Care Research (NIHR) Policy Research Programme through the Policy Innovation and Evaluation Research Unit (PR-PRU-1217-20602). The views expressed are those of the authors and are not necessarily those of the NIHR or the Department of Health and Social Care.

Authors' Contributions

JE contributed to data curation, formal analysis, investigation, methodology, project administration, visualization, writing—original draft, and writing—review and editing. PB contributed to data curation, formal analysis, methodology, visualization, writing—original draft, and writing—review and editing. KA contributed to data curation, formal analysis, methodology, project administration, visualization, writing—original draft, and writing—review and editing. BE contributed to conceptualization, data curation, formal analysis, methodology, writing—original draft, and writing—review and editing. NM contributed to conceptualization, data curation, formal analysis, funding acquisition, methodology, writing—original draft, and writing—review and editing. MAH contributed to conceptualization, data curation, funding acquisition, methodology, project administration, writing—original draft, and writing—review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of number of daily new cases of COVID-19 and key changes in policy measures in England and Wales (September 2020 to December 2021) [61,62].
[PNG File , 73 KB - jmir_v28i1e76863_app1.png]

Multimedia Appendix 2
Additional tables.
[DOCX File , 155 KB - jmir_v28i1e76863_app2.docx]

Multimedia Appendix 3
STROBE checklist.
[PDF File (Adobe PDF File), 199 KB - jmir_v28i1e76863_app3.pdf]

Multimedia Appendix 4
Sankey diagram of respondents' movement between subgroups over the course of 8 waves of a longitudinal survey on use and views of the National Health Service (NHS) COVID-19 app (October 2020 to December 2021), with n=1781 from wave 1 to wave 2, n=1612 from wave 2 to wave 3, n=1584 from wave 3 to wave 4, n=1509 from wave 4 to wave 5, n=1353 from wave 5 to wave 6, n=1249 from wave 6 to wave 7, and n=1124 from wave 7 to wave 8.

XSL•FO
RenderX

[PNG File , 343 KB - jmir_v28i1e76863_app4.png]

Multimedia Appendix 5
Summary of multinomial logistic regression model.
[DOCX File , 22 KB - jmir_v28i1e76863_app5.docx]

References

1. Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler-Dörner L. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. Science 2020;368(6491):eabb6936 [FREE Full text] [doi: 10.1126/science.abb6936] [Medline: 32234805]
2. Leung KY, Metting E, Ebbers W, Veldhuijzen I, Andeweg SP, Luijben G, et al. Effectiveness of a COVID-19 contact tracing app in a simulation model with indirect and informal contact tracing. Epidemics 2024;46:100735 [FREE Full text] [doi: 10.1016/j.epidem.2023.100735] [Medline: 38128242]
3. Kretzschmar ME, Rozhnova G, Bootsma MCJ, van Boven M, van de Wijgert JHHM, Bonten MJM. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. Lancet Public Health 2020;5(8):e452-e459 [FREE Full text] [doi: 10.1016/S2468-2667(20)30157-2] [Medline: 32682487]
4. Burdinski A, Brockmann D, Maier BF. Understanding the impact of digital contact tracing during the COVID-19 pandemic. PLOS Digit Health 2022;1(12):e0000149 [FREE Full text] [doi: 10.1371/journal.pdig.0000149] [Medline: 36812611]
5. Unim B, Zile-Velika I, Pavlovska Z, Lapao L, Peyroteo M, Misins J. The role of digital tools and emerging devices in COVID-19 contact tracing during the first 18 months of the pandemic: a systematic review. Eur J Public Health 2024;34(Supplement_1):i11-i28. [doi: 10.1093/eurpub/ckae039] [Medline: 38946444]
6. Elder AS, Arrouzet CJ, Miljacic L, Karras BT, Higgins A, West LM. Evaluation of the effectiveness of Washington State's digital COVID-19 exposure notification system over one pandemic year. Front Public Health 2024;12:1408178 [FREE Full text] [doi: 10.3389/fpubh.2024.1408178] [Medline: 39206001]
7. Daniore P, Nittas V, Ballouz T, Menges D, Moser A, Höglinger M. Performance of the Swiss digital contact-tracing app over various SARS-CoV-2 pandemic waves: repeated cross-sectional analyses. JMIR Public Health Surveill 2022;8(11):e41004 [FREE Full text] [doi: 10.2196/41004] [Medline: 36219833]
8. Kendall M, Tsallis D, Wymant C, Di Francia A, Balogun Y, Didelot X. Epidemiological impacts of the NHS COVID-19 app in England and Wales throughout its first year. Nat Commun 2023;14(1):858 [FREE Full text] [doi: 10.1038/s41467-023-36495-z] [Medline: 36813770]
9. Wymant C, Ferretti L, Tsallis D, Charalambides M, Abeler-Dörner L, Bonsall D. The epidemiological impact of the NHS COVID-19 app. Nature 2021;594(7863):408-412. [doi: 10.1038/s41586-021-03606-z] [Medline: 33979832]
10. Hinch R, Probert W, Nurtay A, Kendall M. Effective Configurations of a Digital Contact Tracing App: A report to NHSX. 2020. URL: https://cdn.theconversation.com/static_files/files/1009/Report_-_Effective_App_Configurations.pdf [accessed 2025-11-15]
11. Yasaka TM, Lehrich BM, Sahyouni R. Peer-to-peer contact tracing: development of a privacy-preserving smartphone app. JMIR Mhealth Uhealth 2020;8(4):e18936 [FREE Full text] [doi: 10.2196/18936] [Medline: 32240973]
12. Elmokashfi A, Sundnes J, Kvalbein A, Naumova V, Reinemo S, Florvaag PM. Nationwide rollout reveals efficacy of epidemic control through digital contact tracing. Nat Commun 2021;12(1):5918 [FREE Full text] [doi: 10.1038/s41467-021-26144-8] [Medline: 34635661]
13. von Wyl V, Höglinger M, Sieber C, Kaufmann M, Moser A, Serra-Burriel M. Drivers of acceptance of COVID-19 proximity tracing apps in Switzerland: panel survey analysis. JMIR Public Health Surveill 2021;7(1):e25701 [FREE Full text] [doi: 10.2196/25701] [Medline: 33326411]
14. Horvath L, Banducci S, Blamire J, Degnen C, James O, Jones A. Adoption and continued use of mobile contact tracing technology: multilevel explanations from a three-wave panel survey and linked data. BMJ Open 2022;12(1):e053327 [FREE Full text] [doi: 10.1136/bmjopen-2021-053327] [Medline: 35039293]
15. Grill E, Eitze S, De Bock F, Dragano N, Huebl L, Schmich P. Sociodemographic characteristics determine download and use of a corona contact tracing app in Germany-results of the COSMO surveys. PLoS One 2021;16(9):e0256660 [FREE Full text] [doi: 10.1371/journal.pone.0256660] [Medline: 34473733]
16. Johnson B. Nearly 40% of Icelanders are using a covid app—and it hasn't helped much. MIT Technology Review. 2020. URL: https://www.technologyreview.com/2020/05/11/1001541/iceland-rakning-c19-covid-contact-tracing/ [accessed 2025-11-15]
17. Chen AT, Thio KW. Exploring the drivers and barriers to uptake for digital contact tracing. Soc Sci Humanit Open 2021;4(1):100212 [FREE Full text] [doi: 10.1016/j.ssaho.2021.100212] [Medline: 34642660]
18. Guillon M, Kergall P. Attitudes and opinions on quarantine and support for a contact-tracing application in France during the COVID-19 outbreak. Public Health 2020;188:21-31 [FREE Full text] [doi: 10.1016/j.puhe.2020.08.026] [Medline: 33059232]

XSL•FO
RenderX

19. Altmann S, Milsom L, Zillessen H, Blasone R, Gerdon F, Bach R. Acceptability of app-based contact tracing for COVID-19: cross-country survey study. JMIR Mhealth Uhealth 2020;8(8):e19857 [FREE Full text] [doi: 10.2196/19857] [Medline: 32759102]
20. Al-Haboubi M, Exley J, Allel K, Erens B, Mays N. One year of digital contact tracing: Who was more likely to install the NHS COVID-19 app? Results from a tracker survey in England and Wales. Digit Health 2023;9:20552076231159449. [doi: 10.1177/20552076231159449]
21. Marani M, Katul GG, Pan WK, Parolari AJ. Intensity and frequency of extreme novel epidemics. Proc Natl Acad Sci U S A 2021;118(35):e2105482118 [FREE Full text] [doi: 10.1073/pnas.2105482118] [Medline: 34426498]
22. NHS COVID-19 app statistics. NHS Test and Trace. 2022. URL: https://stats.app.covid19.nhs.uk/ [accessed 2022-08-09]
23. NHS COVID-19 app support: what the app does. NHS Test and Trace. 2022. URL: https://www.covid19.nhs.uk/what-the-app-does.html [accessed 2022-03-07]
24. Coronavirus timeline: Welsh and UK governments' response. Welsh Parliament. 2021. URL: https://research.senedd.wales/research-articles/coronavirus-timeline-welsh-and-uk-governments-response/ [accessed 2022-03-20]
25. Methodology. YouGov. 2025. URL: https://yougov.co.uk/about/panel-methodology [accessed 2025-10-10]
26. Goodman LA. Latent class analysis: the empirical study of latent types, latent variables, and latent structures. In: Hagenaars JA, McCutcheon AL, editors. Applied Latent Class Analysis. Cambridge: Cambridge University Press; 2002:3-55.
27. Bartus T. Multilevel multiprocess modeling with Gsem. Stata J 2017;17(2):442-461. [doi: 10.1177/1536867x1701700211]
28. Boxall PC, Adamowicz WL. Understanding heterogeneous preferences in random utility models: a latent class approach. Environ Resour Econ 2002;23(4):421-446. [doi: 10.1023/a:1021351721619]
29. Swait J. A structural equation model of latent segmentation and product choice for cross-sectional revealed preference choice data. J Retail Consum Serv 1994;1(2):77-89. [doi: 10.1016/0969-6989(94)90002-7]
30. Jann B. Color palettes for Stata graphics: an update. Stata J 2023;23(2):336-385. [doi: 10.1177/1536867x231175264]
31. Schonlau M. The clustergram: A graph for visualizing hierarchical and nonhierarchical cluster analyses. Stata J 2002;2(4):391-402. [doi: 10.1177/1536867x0200200405]
32. Cuschieri S. The STROBE guidelines. Saudi J Anaesth 2019;13(5):31. [doi: 10.4103/sja.sja_543_18]
33. English indices of deprivation 2019. Ministry of Housing. 2019. URL: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019 [accessed 2025-11-15]
34. NHS COVID-19 app launches across England and Wales. Department of Health and Social Care. 2020. URL: https://www.gov.uk/government/news/nhs-covid-19-app-launches-across-england-and-wales [accessed 2025-11-15]
35. Samuel G, Sims R. The UK COVID-19 contact tracing app as both an emerging technology and public health intervention: the need to consider promissory discourses. Health (London) 2023;27(4):625-644 [FREE Full text] [doi: 10.1177/13634593211060768] [Medline: 34812092]
36. Kelion L. UK Virus-Tracing App Switches to Apple-Google Model. 2020. URL: https://www.bbc.co.uk/news/technology-53095336 [accessed 2025-11-15]
37. Amann J, Sleigh J, Vayena E. Digital contact-tracing during the Covid-19 pandemic: an analysis of newspaper coverage in Germany, Austria, and Switzerland. PLoS One 2021;16(2):e0246524 [FREE Full text] [doi: 10.1371/journal.pone.0246524] [Medline: 33534839]
38. Lueks W, Benzler J, Bogdanov D, Kirchner G, Lucas R, Oliveira R. Toward a common performance and effectiveness terminology for digital proximity tracing applications. Front Digit Health 2021;3:677929 [FREE Full text] [doi: 10.3389/fdgth.2021.677929] [Medline: 34713149]
39. Samuel G, Lucivero F. Framing ethical issues associated with the UK COVID-19 contact tracing app: exceptionalising and narrowing the public ethics debate. Ethics Inf Technol 2022;24(1):5 [FREE Full text] [doi: 10.1007/s10676-022-09628-z] [Medline: 35110970]
40. Davies B, Lalot F, Peitz L, Heering MS, Ozkececi H, Babaian J. Changes in political trust in Britain during the COVID-19 pandemic in 2020: integrated public opinion evidence and implications. Humanit Soc Sci Commun 2021;8(1):166. [doi: 10.1057/s41599-021-00850-6]
41. Boehm C, Boadu P, Exley J, Al-Haboubi M, Mays N. Public trust in the government to control the spread of COVID-19 in England after the first wave-a longitudinal analysis. Eur J Public Health 2023;33(6):1155-1162 [FREE Full text] [doi: 10.1093/eurpub/ckad148] [Medline: 37579239]
42. Timeline of UK government coronavirus lockdowns and restrictions. Institute for Government. 2022. URL: https://www.instituteforgovernment.org.uk/charts/uk-government-coronavirus-lockdowns [accessed 2025-11-15]
43. Haddon C, Sasse T, Nice A. Science advice in a crisis. Institute of Government. 2020. URL: https://www.instituteforgovernment.org.uk/publication/report/science-advice-crisis [accessed 2025-11-15]
44. Coronavirus: lessons learned to date report published. House of Commons Health and Social Care. 2021. URL: https://committees.parliament.uk/committee/81/health-and-social-care-committee/news/157991/coronavirus-lessons-learned-to-date-report-published [accessed 2025-11-15]
45. Samuel G, Lucivero F, Johnson S, Diedericks H. Ecologies of public trust: the NHS COVID-19 contact tracing app. J Bioeth Inq 2021;18(4):595-608 [FREE Full text] [doi: 10.1007/s11673-021-10127-x] [Medline: 34609676]

XSL•FO
RenderX

46. Pepper C, Reyes-Cruz G, Pena AR, Dowthwaite L, Babbage CM, Wagner H. Understanding trust and changes in use after a year with the NHS covid-19 contact tracing app in the united kingdom: longitudinal mixed methods study. J Med Internet Res 2022;24(10):e40558 [FREE Full text] [doi: 10.2196/40558] [Medline: 36112732]
47. Cellan-Jones R. Disappearing Covid-19 App Alerts Cause Alarm. 2020. URL: https://www.bbc.co.uk/news/technology-54389083 [accessed 2022-07-08]
48. Garousi V, Cutting D. What do users think of the UK's three COVID-19 contact-tracing apps? A comparative analysis. BMJ Health Care Inform 2021;28(1):E100320 [FREE Full text] [doi: 10.1136/bmjhci-2021-100320] [Medline: 34281994]
49. Dixon EL, Joshi SM, Ferrell W, Volpp KG, Merchant RM, Guntuku SC. COVID-19 contact tracing app reviews reveal concerns and motivations around adoption. PLoS One 2022;17(9):e0273222 [FREE Full text] [doi: 10.1371/journal.pone.0273222] [Medline: 36084078]
50. Callaway E. Delta coronavirus variant: scientists brace for impact. Nature 2021;595(7865):17-18. [doi: 10.1038/d41586-021-01696-3] [Medline: 34158664]
51. Davies C, Stewart H. NHS Covid App to Be Tweaked to Cut Need for Self-Isolation. 2021. URL: https://www.theguardian.com/world/2021/jul/08/sunak-hints-at-changes-to-test-and-trace-app-to-cut-numbers-told-to-isolate [accessed 2025-11-15]
52. Stewart H. Keir Starmer: tweaking NHS Covid app ‘like taking batteries out of smoke alarm’. 2021. URL: https://www.theguardian.com/world/2021/jul/09/keir-starmer-tweaking-nhs-covid-app-taking-batteries-smoke-alarm [accessed 2025-11-15]
53. Dowthwaite L, Fischer J, Perez Vallejos E, Portillo V, Nichele E, Goulden M. Public adoption of and trust in the NHS COVID-19 contact tracing app in the United Kingdom: quantitative online survey study. J Med Internet Res 2021;23(9):e29085 [FREE Full text] [doi: 10.2196/29085] [Medline: 34406960]
54. Williams SN, Armitage CJ, Tampe T, Dienes K. Public attitudes towards COVID-19 contact tracing apps: a UK-based focus group study. Health Expect 2021;24(2):377-385 [FREE Full text] [doi: 10.1111/hex.13179] [Medline: 33434404]
55. Smith LE, Potts HWW, Aml t R, Fear NT, Michie S, Rubin GJ. Patterns of social mixing in England changed in line with restrictions during the COVID-19 pandemic (September 2020 to April 2022). Sci Rep 2022;12(1):10436 [FREE Full text] [doi: 10.1038/s41598-022-14431-3] [Medline: 35729196]
56. Zabel S, Schlaile MP, Otto S. Breaking the chain with individual gain? Investigating the moral intensity of COVID-19 digital contact tracing. Comput Human Behav 2023;143:107699 [FREE Full text] [doi: 10.1016/j.chb.2023.107699] [Medline: 36818428]
57. OFCOM Nations & Regions Technology Tracker - 2020, 9th January to 7th March 2020. 2020. URL: https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/data/statistics/2020/technology-tracker-2020/technology-tracker-2020-uk-data-tables?v=324794 [accessed 2025-11-15]
58. Erens B, Burkill S, Couper MP, Conrad F, Clifton S, Tanton C. Nonprobability web surveys to measure sexual behaviors and attitudes in the general population: a comparison with a probability sample interview survey. J Med Internet Res 2014;16(12):e276 [FREE Full text] [doi: 10.2196/jmir.3382] [Medline: 25488851]
59. Sujarwoto S, Maharani A. Facilitators and barriers to the adoption of mHealth apps for COVID-19 contact tracing: a systematic review of the literature. Front Public Health 2023;11:1222600 [FREE Full text] [doi: 10.3389/fpubh.2023.1222600] [Medline: 38145061]
60. Liccardi I, Alekseyev J, Woltz VLA, McLean JE, Zurko ME. Public willingness to engage with COVID-19 contact tracing, quarantine, and exposure notification. Public Health Rep 2022;137(2_suppl):90S-95S [FREE Full text] [doi: 10.1177/00333549221125891] [Medline: 36255241]
61. UKHSA data dashboard. URL: https://coronavirus.data.gov.uk/details/cases [accessed 2025-11-20]
62. Timeline of UK government coronavirus lockdowns and restrictions. URL: https://www.instituteforgovernment.org.uk/data-visualisation/timeline-coronavirus-lockdowns [accessed 2025-11-20]

Abbreviations

AIC: Akaike information criteria
BIC: Bayesian information criteria
LCA: latent class analysis
NHS: National Health Service
RRR: relative risk ratio
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
VIF: variance inflation factors

XSL•FO
RenderX

Edited by A Mavragani, T de Azevedo Cardoso; submitted 02.May.2025; peer-reviewed by J Igboanugo, V Surasani; comments to author 10.Jun.2025; revised version received 16.Oct.2025; accepted 16.Oct.2025; published 28.Jan.2026.

Please cite as:
Exley J, Boadu P, Allel K, Erens B, Mays N, Al-Haboubi M
Characteristics Influencing Support for the National Health Service COVID-19 App in England and Wales: Findings From a Longitudinal Survey
J Med Internet Res 2026;28:e76863
URL: https://www.jmir.org/2026/1/e76863
doi:10.2196/76863
PMID:

©Josephine Exley, Paul Boadu, Kasim Allel, Bob Erens, Nicholas Mays, Mustafa Al-Haboubi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patterns and Characteristics of Mobile App Use to Promote Wellness and Manage Illness: Cross-Sectional Study

Hayriye Gulec[1,2], PhD; David Smahel[1], PhD; Yi Huang[1,3], PhD

[1]Interdisciplinary Research Team on Internet and Society, Faculty of Social Studies, Masaryk University, Brno, Czech Republic

[2]Department of Psychology, Faculty of Arts and Sciences, Bursa Uludag University, Bursa, Turkey

[3]Centre for Research and Development, Pan-European University, Bratislava, Slovakia

Corresponding Author:
Hayriye Gulec, PhD
Interdisciplinary Research Team on Internet and Society
Faculty of Social Studies
Masaryk University
Joštova 10
Brno, 620 00
Czech Republic
Phone: 420 549 49 4180
Email: hayriyegulec@uludag.edu.tr

Abstract

Background: Mobile health (mHealth) apps target diverse health behaviors, but engagement may vary by purpose.

Objective: This study examined the prevalence, usage patterns, and user characteristics of mHealth apps among Czech adults with internet access, focusing on sociodemographics, digital knowledge and use, and health indicators predicting wellness- and illness-related app use.

Methods: Overall, 4775 Czech adults (2365/4775, 49.53% women) aged 18-95 (mean 45.37, SD 16.40) years completed an online survey. Sociodemographic factors included age, gender, education, and income. Digital knowledge and use were measured using the eHealth Literacy Scale and the passive/active use of social networking sites (SNS) for health information. Health indicators covered symptom severity, physical activity, BMI, and eating disorder–related risk propensity (body dissatisfaction, dietary restraint, and weight/shape overvaluation). Participants reported app use for sports, number of steps, nutrition, vitals, sleep, diagnosed conditions, reproductive health, diagnosis assistance, mood and mental well-being, and emergency care guidance. Multivariate hierarchical binary logistic regression analysis identified characteristics of app users. Exploratory structural equation modeling (ESEM) clustered apps into "promoting wellness" and "managing illness" and examined the predictors of frequency of use.

Results: Of 4440 respondents, 2172 (48.92%) used mHealth apps. Users were younger (odds ratio [OR] 0.98, 95% CI 0.98-0.99, *P*<.001), had a monthly income more than 50,000 CZK (1 CZK=US $0.048; vs ≤20,000 CZK: OR 0.54, 95% CI 0.41-0.7, *P*<.001; 20,001-35,000 CZK: OR 0.78, 95% CI 0.65-0.93, *P*=.006; 35,001-50,000 CZK: OR 0.83, 95% CI 0.7-0.99, *P*=.03), were more eHealth literate (OR 1.17, 95% CI 1.06-1.3, *P*=.003), used SNS passively for health information (OR 1.35, 95% CI 1.21-1.51, *P*<.001), and had higher eating disorder risk (OR 1.18, 95% CI 1.12-1.25, *P*<.001) and physical activity (OR 1.18, 95% CI 1.13-1.23, *P*<.001) than nonusers. Step-counting apps were most common; 65.99% (1430/2167) used them daily or several times a day, followed by apps for sleep (691/2163, 31.95%), vitals (611/2165, 28.22%), and sports (407/2158, 18.86%). ESEM confirmed a 2-factor structure ("promoting wellness" and "managing illness"; χ^2_{26}=71.9, comparative fit index=0.99, Tucker-Lewis index=0.99, root-mean-square error of approximation=0.03, and standardized root-mean-square residual=0.03). Frequent use of wellness apps was associated with younger age (standardized β=–0.22, *P*<.001), higher eHealth literacy (standardized β=0.10, *P*<.001), and physical activity (standardized β=0.15, *P*<.001). Illness-management app use was associated with active use of SNS for health information (standardized β=0.62, *P*<.001) and eating disorder risk (standardized β=0.11, *P*<.001). Digital knowledge, digital use, and health indicators mediated the association between age and mHealth app use.

Conclusions: mHealth app engagement reflects broader social, digital, and psychological inequalities rather than individual preferences alone. Encouraging digital inclusion and addressing body image- and diet-related use may help ensure that mHealth technologies do not exacerbate existing health inequalities across age and user groups.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e71363) doi:10.2196/71363

KEYWORDS

mobile health; mHealth; eHealth literacy; eating disorder; ESEM; exploratory structural equation modeling; SNS; social networking site

Introduction

Background

The rapid advancements in mobile health (mHealth) app technologies and the continued improvement to content have resulted in the proliferation of health-related areas addressed via mHealth apps and wearables in the last decade [1-3]. Now, smartphone users can monitor and regulate daily routines (eg, number of steps and circadian rhythms) and autonomic indicators (eg, heart rate, blood pressure, body temperature, and respiratory rate). They can use apps to modify lifestyle behaviors (eg, physical activity, nutritional intake, and weight management), improve mental health and well-being (eg, meditation and mindfulness), get health assistance (eg, diagnosis assistance, reproductive health, and emergency care guidance), and manage diagnosed conditions (eg, diabetes and depression).

Two users who download the same app to track and regulate a particular health behavior (eg, nutritional intake) may use it for separate purposes: one to maintain a healthy lifestyle and the other to manage a chronic health condition (eg, diabetes and obesity). The characteristics of users might differ depending on their purpose of use. Although previous literature has linked mHealth app usage to user intentions related to self-management of wellness and illness [4-7], empirical studies have primarily focused on app adoption in general, overlooking user characteristics associated with the purpose of use [8]. Additionally, only a few studies have examined the prevalence and characteristics of mHealth app usage in large-scale, representative samples [9-12].

Therefore, the aim of this study is 2-fold. First, we examine the prevalence and characteristics of app users within a representative sample of Czech adults with internet access. We focus on sociodemographics, digital knowledge and use, and health indicators as user characteristics. Second, we cluster mHealth apps using a data-driven approach under the "promoting wellness" and "managing illness" dimensions. Finally, we determine the user characteristics that predict the app usage frequency for each dimension. This study reports on a large representative adult sample concerning the prevalence, usage patterns, and user characteristics of various types of apps. It also provides initial evidence for user characteristics as predictors of wellness- and illness-related app usage.

Prior large-scale studies examining adult mHealth app users were primarily conducted in the United States using different waves of data from the National Cancer Institute's Health Information National Trends Survey. The prevalence of mHealth app use among smartphone owners ranged between 34% and 56%, and they consistently indicated younger age and higher educational status as significant predictors [9-12]. Females were more likely to own and use mHealth apps in general [10-12], although no gender difference was reported in one study [9]. A recent scoping review identified age, gender, educational level, and income status as significant factors [8]. This study examines the prevalence and sociodemographic characteristics of mHealth app use among a representative sample of Czech adults with internet access. It also reveals whether these characteristics differ when using apps to promote wellness and manage illness.

eHealth literacy is the perceived knowledge and skills to obtain, understand, and evaluate digital health information [13]. Previous studies reported eHealth literacy as a significant predictor of mHealth app use in general adult samples [14,15] and individuals with chronic health conditions [16,17]. In addition, higher eHealth literacy was associated with the perception of mHealth apps as effective and their frequent usage for health-related behavioral change [18]. This study provides initial insights into whether app usage to promote wellness and manage illness differs as a function of eHealth literacy.

Digital platforms, including social networking sites (SNS; eg, Facebook, Instagram, and X, formerly known as Twitter), are frequently used for health-related purposes [19,20]. Individuals who use social media for health information are more likely to seek health information through mHealth apps, indicating that both media might be used complementarily [21]. Moreover, prior research has shown that health messages delivered via SNS can shape users' health-related decisions and behaviors [21,22]. Therefore, a positive relationship between SNS activity and the use of mobile apps for health purposes is plausible. This study aims to provide initial evidence on whether health-related SNS use is linked to the adoption of health apps and the frequency of their use for promoting wellness and managing illness.

Prior studies have examined general health status as a predictor of mHealth app use; however, these studies have assessed health status using a single item, yielding mixed findings [9-12]. This study adopts a more nuanced approach to address this limitation, assessing perceived somatic-symptom severity across multiple domains, including gastrointestinal, cardiopulmonary, musculoskeletal, and sleep/energy areas. In parallel, an active lifestyle—as reflected in physical activity levels—has received growing attention in mHealth research, with evidence suggesting that app use may support health promotion among physically active individuals [10,18,23]. By jointly examining both physical activity and detailed health status indicators, this study contributes to understanding how these factors shape the adoption and use of apps for promoting wellness and managing illness.

Lastly, only a few studies examined BMI and eating disorder–related risk factors as correlates of mHealth app use. This study focused on BMI because it is a significant predictor of health outcomes, including diabetes and obesity [24]. Evidence suggests that individuals with a higher BMI are more likely to use mHealth apps [10,18,25] and to use them for a

XSL•FO
RenderX

health behavior goal [9,26]. A higher BMI is also a significant predictor of eating disorder symptoms, including body dissatisfaction, dietary restraint, and weight and shape overvaluation [27]. Previous research has reported a connection between using mHealth apps to track diet and physical activity and their associations with eating disorder–related risks and symptoms, including BMI [28-30]. These findings underscore the importance of evaluating core risk factors for eating disorders together with BMI to understand how they might play roles in using apps that address behaviors related to diet and physical exercise (ie, health promotion) but also those related to monitoring health (eg, vitals) and managing disease (eg, obesity and diabetes). In this study, we evaluate body dissatisfaction, dietary restraint, and weight and shape overvaluation to determine participants' risk propensity for an eating disorder.

This Study

This study investigated the prevalence and frequency of mHealth app use in a representative sample of Czech adults with internet access. Our first objective was to identify user characteristics that predict app adoption and usage. We examined a comprehensive set of predictors, including sociodemographic variables (age, gender, education, and income), digital knowledge and use (eHealth literacy and SNS use for health information), and health indicators (somatic symptom severity, physical activity, BMI, and eating disorder risk propensity). This multidimensional approach enabled us to move beyond basic demographic correlates and gain a deeper understanding of the factors influencing engagement with mHealth apps. Our second objective was to cluster mHealth apps into two key dimensions—promoting wellness and managing illness—as suggested by the prior literature [4-7]. Rather than relying solely on predefined app classifications, we used a data-driven approach to cluster mHealth app usage into categories related to wellness and illness. Furthermore, we determined user characteristics associated with the frequency of use within each category. This approach provided a more nuanced understanding of user characteristics connected with using wellness and illness app types.

During the analyses, we observed that the effect size of age on app use was consistently reduced after entering the variables related to digital knowledge, digital use, and health indicators. This finding prompted us to conduct additional exploratory mediation analyses to test whether these factors mediated the association between age and app use. These exploratory analyses provide insight into potential mechanisms underlying age-related differences in mHealth engagement, informing future strategies to reduce disparities in app adoption and use for wellness promotion and illness management.

Methods

Recruitment

This study used cross-sectional data collected through an online survey targeting adult Czech internet users. The study design adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for observational cross-sectional research. Eligible participants were adults aged 18 years or older who reported using the internet. Data collection was conducted by STEM/MARK, a Czech market research agency that manages the Czech National Panel (part of National Sample s.r.o.). STEM/MARK is a member of the European Society for Opinion and Market Research and follows its data protection and panel management standards. Data collection occurred between October 2 and 16, 2023, as part of the NPO "Systemic Risk Institute" project.

The Czech National Panel includes 64,000 individuals and is designed to capture a broad spectrum of the population, including harder-to-reach subgroups. Eligible participants were selected from an initial pool of 58,000 participants. A quota-sampling strategy was applied with respect to age, educational attainment, household income, municipality size, and administrative region, following the Nomenclature of Territorial Units for Statistics (NUTS3) classification, as per Eurostat data. The gender distribution was balanced, with a tolerance of ±7% based on the registered gender within the panel. Our recruitment goal was to maximize sample size within the project's funding constraints, with a minimum target of 3500 respondents to ensure meaningful representation even from the smallest regions.

Participants were recruited through online and face-to-face invitations, and the questionnaire was delivered via computer-assisted web interviewing. The survey was opened by 5480 panelists. Of these, 5135 filled in the items to verify the respondent's suitability for data collection (eg, quotas). Due to noncompliance with research requirements or failure to fulfill quotas, 214 respondents were rejected. The number of panelists who completed the questionnaire was 4921 individuals. An additional 146 questionnaires were discarded due to excessive missing data (>10%), unrealistically fast completion times, or logically inconsistent answers. The final sample consisted of 4775 adult Czech internet users (2365/4775, 49.53% women) aged 18-95 years (mean 45.37, SD 16.40 years).

All survey instruments underwent cognitive pretesting before data collection. Feedback was obtained from 5 individuals representing diverse genders, ages, and educational backgrounds (aged 26-73 years). Their input focused on evaluating the clarity, comprehensibility, and overall content of the questionnaire.

Measures

Sociodemographic Characteristics

Participants reported their gender and responded to an open-ended question about their age. They indicated their highest education with the following response options: (1) no education or primary school, (2) secondary vocational school, (3) secondary school, and (4) university (or higher vocational school). The income status was assessed by asking, "In which category would you put the average monthly income of your household for all its members? Please consider the income from employment, part-time jobs, business, and social security." The response options were (1) up to 20,000 CZK, (2) 20,001-35,000 CZK, (3) 35,001-50,000 CZK, and (4) 50,001 CZK and more. [A currency exchange rate of CZK 1=US $0.048 is applicable.] Higher scores were indicative of better financial status.

XSL•FO
RenderX

Digital Knowledge and Use Factors

eHealth Literacy

eHealth literacy was assessed by the eHealth Literacy Scale [31]. The items inquired about knowledge of online health information sources (eg, I know what health resources are available on the internet), how to navigate the internet to obtain answers to health-related questions (eg, I know where to find helpful health resources on the internet), the perceived skills to evaluate the quality of online health information (eg, I can tell high-quality health resources from low-quality health resources on the internet), and ability to apply health information for health purposes (eg, I know how to use the health information I find on the internet to help me). The response options ranged from 1 (strongly disagree) to 5 (strongly agree). Higher scores indicated better eHealth literacy skills. The internal consistency was adequate (Cronbach α=0.89).

Health-Related SNS Use

The health-related use of SNS was assessed using items developed in a previous study to measure both active and passive social media use [32]. We adapted the items to tap into passive and active use of SNS for health information. Passive health-related SNS use refers to the passive engagement with health information. Participants responded to how often they "watched videos or pictures posted about health or illness," "followed health- or illness-related pages," and "read health- or illness-related discussions, user comments, ratings, or reviews" on SNS. Active health-related SNS use items evaluated active engagement with the health information and asked how often participants "posted health- or illness-related content," "chatted or interacted with others about health or illness," and "commented on, liked, or shared others' health- or illness-related posts" on SNS. The response options were (1) never, (2) a few times at most, (3) several times a month, (4) several times a week, (5) every day, and (6) several times a day. The scores were calculated separately for passive and active use; the Cronbach alphas were 0.9 for both scales.

Health Indicators

Symptom Severity

The perceived severity of symptoms was assessed using items from the Patient Health Questionnaire-15, which evaluates 15 somatic symptoms encountered in outpatient settings [33]. It is a widely used brief screening instrument with good psychometric properties for assessing somatic symptoms and screening for somatization [34]. The original scale asks participants to rate how much they were bothered by each symptom in the preceding month. For brevity, we grouped the symptoms under gastrointestinal (ie, stomach pain, constipation, loose bowels, diarrhea, nausea, gas, and indigestion), cardiopulmonary (ie, chest pain, feeling your heart pound or race, and shortness of breath), musculoskeletal (ie, back pain, pain in your arms or legs, and pain in joints like knees and hips), and sleep/energy (ie, headaches, dizziness, fainting spells, feeling tired, having low energy, and trouble sleeping) areas. For each symptom area, the participants evaluated the extent to which they were bothered by any of the symptoms in the preceding 4 weeks. The items were rated on a 5-point Likert scale that ranged from 1 (never) to 5 (always). Higher scores were indicative of more severe symptoms. The internal consistency of the scale was satisfactory (Cronbach α=0.72).

Physical Activity

Participants responded to the question—"How many hours a week do you usually exercise to the extent that you sweat and feel shortness of breath?"—to evaluate their physical activity. The response options included (1) less than half an hour per week, (2) about half an hour per week, (3) about 1 hour per week, (4) about 2-3 hours per week, (5) about 4-6 hours per week, and (6) about 7 hours or more per week.

BMI

Participants responded to open-ended questions about their height (in centimeters) and weight (in kilograms). The BMI was calculated by weight (kg)/height (m^2)×10,000 formula.

Eating Disorder–Related Risk Propensity

Eating disorder–related risk propensity was assessed with the modified 7-item version [35] of the Eating Disorder Examination Questionnaire [36], a widely used measure of eating pathology. It consists of 3 factors that evaluate dietary restraint, shape/weight overvaluation, and body dissatisfaction. Body dissatisfaction (ie, dissatisfaction with weight and shape) and weight/shape overvaluation (ie, the influence of weight and shape on self-worth) are concerned with the cognitive-evaluative aspects of body image. Dietary restraint assesses the frequency of rigid behaviors to influence weight and shape. The scale is rated with a 7-point forced-choice format that considers the preceding 28 days. Higher scores indicate a greater frequency of dietary restraint (ie, from zero to every day) and a greater severity of body dissatisfaction and weight/shape overvaluation (ie, from "not at all" to "extremely"). The internal consistencies of the factors ranged from 0.89 to 0.91, and the factor loadings and item intercepts were invariant for sex and overweight status in a previous study with adults [35]. The scale's internal consistency was satisfactory in this study (Cronbach α=0.87).

mHealth App Use

App use was determined by asking, "You can use various applications on your phone, tablet, or other mobile devices, like smartwatches. Have you used applications to monitor health and exercise (eg, counting steps, tracking calories, weight, sports activities, eating/drinking, stress, or sleep) in the last year?" The response options were (1) no and (2) yes. mHealth app users responded to an additional question about the frequency of using different types of apps, which included sports (eg, workouts, exercise, running, and strengthening), number of steps, nutrition (eg, calorie intake or expenditure, weight, and nutritional facts), vitals (eg, blood pressure, heart rate, body temperature, and respiratory rate), sleep (eg, sleep patterns), mood and mental well-being (eg, mindfulness, meditation, mental health, and self-esteem), support or control of a diagnosed health condition (eg, diabetes management and depression management), reproductive health (eg, pregnancy, ovulation, menstruation, and sexual health), diagnosis assistance (eg, symptom checking), and emergency care guidance. The response options included (1) never, (2) a few times at most, (3) several times a month, (4) several times a week, (5) every

XSL•FO
RenderX

day, and (6) several times a day. The scale's internal consistency was satisfactory (Cronbach α=0.82).

Statistical Analysis

Descriptive statistics were run for the sociodemographic characteristics, digital knowledge and use factors, and health indicators. They included means, SDs, and frequencies for the entire sample, app users, and app nonusers. Participants who reported using an mHealth app in the previous year were identified as app users. A multivariate hierarchical binary logistic regression analysis was conducted to examine the characteristics of app users. Sociodemographic factors (age, gender, education, and income) were included in the first step. The second step added the digital knowledge and use factors (eHealth literacy, active and passive use of SNS) and health indicators (perceived symptom severity, physical activity, BMI, and eating disorder–related risk propensity).

The remaining analyses were based on app users' responses regarding their app use patterns. First, we presented usage frequencies for each app type. Then, we examined the factor structure of app types based on usage frequencies. Guided by the mHealth app categories suggested in the previous literature [4-7], we assumed that mHealth app usage could be divided into two dimensions: mHealth app usage for "promoting wellness" and "managing illness." We initially ran the exploratory factor analysis (EFA), and based on the factor loading patterns (see the "mHealth_EFA" worksheet in Multimedia Appendix 1), we found that specific mHealth apps (related to nutrition and mental health) appeared to load on two dimensions. Furthermore, the strictly constrained confirmatory factor analysis (CFA) model (see the "mHealth_CFA" worksheet in Multimedia Appendix 1), where, psychometrically speaking, each item is specified to load exclusively on a single factor, yielded an unsatisfactory model fit (comparative fit index=0.86, Tucker-Lewis index=0.81, and root-mean-square error of approximation=0.13). Thus, we subsequently decided to apply the exploratory structural equation modeling (ESEM) method to test this hypothesis. The ESEM used usage frequencies of each app type to determine the underlying factors of use. Although ESEM operates within the traditional CFA framework, it notably allows for the free estimation of cross-loadings while still displaying targeted and constrained factors [37]. This less restrictive approach provides an alternative when CFA does not achieve an adequate model fit. The flexibility of ESEM is advantageous because it more effectively captures the multidimensional aspects of a measure [37,38]. In this study, the ESEM enabled us to assess the model's fit for our targeted factors (ie, promoting wellness and managing illness). In addition, it could capture the multidimensional aspects of app use by facilitating the cross-loading of app types for both dimensions.

The ESEM approach was also used to predict user characteristics associated with mHealth app usage for our targeted factors (ie, wellness and illness). The dependent variables in the first ESEM model were the frequency of mHealth app usage for promoting wellness and the frequency of mHealth app usage for managing illness. The independent variables were sociodemographics, including age, gender, educational level, and income status. In the second model, we added the predictors of digital knowledge and use factors (eHealth literacy, active and passive use of SNS) and health indicators (symptom severity, physical activity, BMI, eating disorder–related risk propensity).

Lastly, we conducted two separate mediation analyses to test whether digital knowledge, digital use, and health indicators mediated the association between age and app usage. The first mediation analysis examined the relationship between age and wellness app use frequency, and the second examined the relationship between age and illness app use frequency. The analyses were conducted using SPSS software (version 28.0, IBM Corp), the R package "lavaan" (The R Foundation), and the open-source software Jamovi (version 2.6), which provides an R-based graphical user interface. Within Jamovi, the analyses used the R packages "psych," "lavaan," and "lm." Mahalanobis distances were computed to identify multiple outliers, and data from 1.11% (53/4775) of the participants with significant Mahalanobis distance values at $P<.001$ were deleted.

Ethical Considerations

This study was reviewed and approved by the Research Ethics Committee of Masaryk University (EKV-2023-102). The study complied with ethical standards for human subjects research. Before completing the questionnaires, participants were informed about the purpose of the survey, its anonymity, and their right to refuse participation or skip any question by selecting "I don't know" or "I prefer not to say." Written informed consent was obtained before participation. All data were collected and analyzed anonymously, with no personally identifiable information retained. Respondents received monetary compensation for participation, determined by the panel's standard reward system based on the questionnaire length; the exact amount was not disclosed to the researchers. No images or other materials that could reveal participant identities are included in this manuscript.

Results

Data Availability

The sample size consisted of 4775 adults. Of these, data on mHealth app usage were available for 4440 participants and were included in the analyses. Those who did not provide data on mHealth app use (n=282) were more likely to report an older age ($t_{4720}=17.18$, $P<.001$; Cohen $d=1.055$). All t values are 2-tailed. There were also significant differences regarding educational level ($\chi^2_{3,4722}=19.5$, $P<.001$; $\phi_c=0.064$) and income status ($\chi^2_{3,4722}=117.2$, $P<.001$; $\phi_c=0.158$). Data providers were more likely to have a university or higher vocational school education and a monthly income of more than 50,000 CZK than data nonproviders. On the other hand, data nonproviders reported secondary vocational school education more frequently, and they were more likely to have an annual income of less than 20,000 CZK than data providers. No significant gender differences were observed ($\chi^2_{1,4718}=3.04$, $P=.08$; $\varphi=0.025$).

XSL•FO
RenderX

Prevalence of mHealth App Use and Sample Characteristics

Around half of the participants reported using mHealth apps (2172/4440, 48.92%). The characteristics of the total sample, mHealth app users, and nonusers on sociodemographics, digital knowledge and use factors, and health indicators are shown in Table 1.

Table 1. Sample characteristics of Czech internet users participating in this cross-sectional study[a].

Variables	Total sample (N=4440)	App users (n=2172)	App nonusers (n=2268)
Sociodemographics			
Age (years), mean (SD)	44.43 (15.97)	41.36 (15.01)	47.38 (16.31)
Gender, woman, n (%)	2185 (49.26)	1108 (51.08)	1077 (47.51)
Educational level, n (%)			
None/primary education	286 (6.44)	141 (6.49)	145 (6.39)
Secondary vocational	1324 (29.82)	568 (26.15)	756 (33.33)
Secondary school	1733 (39.03)	857 (39.46)	876 (38.62)
University (or higher vocational)	1097 (24.71)	606 (27.90)	491 (21.65)
Income status, n (%)			
Up to 20,000 CZK[b]	454 (10.23)	175 (8.06)	279 (12.30)
20,001-35,000 CZK	1364 (30.72)	623 (28.68)	741 (32.67)
35,001-50,000 CZK	1455 (32.77)	715 (32.92)	740 (32.63)
50,001 CZK and more	1167 (26.28)	659 (30.34)	508 (22.40)
Digital knowledge and use, mean (SD)			
eHealth literacy	3.73 (0.67)	3.81 (0.65)	3.65 (0.68)
Passive SNS use[c]	1.9 (0.94)	2.07 (1)	1.74 (0.83)
Active SNS use[d]	1.49 (0.82)	1.59 (0.9)	1.4 (0.73)
Health indicators, mean (SD)			
Symptom severity	2.55 (0.83)	2.6 (0.81)	2.5 (0.84)
Physical activity	2.82 (1.63)	3.10 (1.59)	2.52 (1.62)
BMI	27.66 (5.7)	27.49 (5.49)	27.82 (5.9)
ED-related risk[e]	2.73 (1.4)	2.97 (1.41)	2.5 (1.34)

[a]All values are calculated based on available (nonmissing) data; denominators may therefore vary across variables.

[b]A currency exchange rate of 1 CZK=US $0.048 is applicable.

[c]Passive SNS use: passive use of social networking sites for health information.

[d]Active SNS use: active use of social networking sites for health information.

[e]ED-related risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).

Characteristics of mHealth App Users

The multivariate hierarchical binary logistic regression analysis to evaluate the factors associated with mHealth app use is shown in Table 2. Model 1 reports the sociodemographic variables—age, gender, educational level, and income status. Model 2 tests the direct effects of digital knowledge and use factors (eHealth literacy and active and passive use of SNS for health information) and health indicators (symptom severity, physical activity, BMI, and eating disorder–related risk propensity). All models report odds ratios with 95% CIs.

XSL•FO
RenderX

Table 2. Sociodemographic, digital, and health-related predictors of mHealth app use based on multivariate hierarchical binary logistic regression analysis in a cross-sectional sample of Czech internet users (n=3923).

Predictors	Model 1[a]		Model 2[b]	
	OR[c] (95% CI)	*P* value	OR (95% CI)	*P* value
Sociodemographics				
Age (years)	0.98 (0.97-0.98)	<.001	0.98 (0.98-0.99)	<.001
Gender (reference: male)	1.24 (1.09-1.42)	.001	1.13 (0.98-1.3)	.09
Educational level				
None/primary education	0.75 (0.56-1.01)	.06	0.8 (0.59-1.09)	.15
Secondary vocational	0.87 (0.73-1.04)	.14	0.93 (0.77-1.12)	.44
Secondary school	0.92 (0.78-1.08)	.29	0.96 (0.81-1.13)	.61
University (or higher vocational)	Reference	Reference	Reference	Reference
Income status				
Up to 20,000 CZK[d]	0.56 (0.43-0.71)	<.001	0.54 (041-0.7)	<.001
20,001-35,000 CZK	0.78 (0.66-0.93)	<.001	0.78 (0.65-0.93)	.006
35,001-50,000 CZK	0.83 (0.7-0.98)	.006	0.83 (0.7-0.99)	.03
50,001 CZK and more	Reference	Reference	Reference	Reference
Digital knowledge and use				
eHealth literacy	—[e]	—	1.17 (1.06-1.3)	.003
Passive SNS use[f]	—	—	1.35 (1.21-1.51)	<.001
Active SNS use[g]	—	—	0.91 (0.81-1.02)	.10
Health indicators				
Symptom severity	—	—	1.09 (0.99-1.19)	.07
Physical activity	—	—	1.18 (1.13-1.23)	<.001
BMI	—	—	1 (0.99-1.02)	.57
ED-related risk[h]	—	—	1.18 (1.12-1.25)	<.001

[a]Nagelkerke R^2=0.056.

[b]Nagelkerke R^2=0.128.

[c]OD: odds ratio.

[d]A currency exchange rate of CZK 1=US $0.048 is applicable.

[e]Not applicable.

[f]Passive SNS use: passive use of social networking sites for health information.

[g]Active SNS use: active use of social networking sites for health information.

[h]ED-related risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).

The significant predictors of having an app were age, gender, and income status in model 1. Participants who were younger, female, and had a monthly income of more than 50,000 CZK (vs up to 20,000 CZK, between 20,001 and 35,000 CZK, or between 35,001 and 50,000 CZK) were more likely to use mHealth apps. In model 2, digital knowledge, use factors, and health indicators were entered. The role of gender became insignificant when additional variables were included in model 2. After adjusting for the roles of sociodemographic characteristics, the results demonstrated that app users had higher eHealth literacy and were more likely to engage in passive use of SNS for health information. Among the health indicators, higher physical activity and a propensity for risk related to eating disorders were predictive of mHealth app use.

mHealth App Use by Type of App

The frequency of use for each mHealth app type among app users is shown in Table 3. The most frequently used mHealth app type was the number of steps. The percentage of participants who count steps daily or several times a day was 65.99% (1430/2167). It was followed by mHealth apps to track sleep, vitals, and sports. The percentage of participants who used apps daily or several times a day for monitoring sleep was 31.95% (691/2163), 28.22% (611/2165) for monitoring vitals, and 18.86% (407/2158) for sports. App users were least likely to have ever used apps for emergency care guidance, diagnosis

XSL•FO
RenderX

assistance, or to manage diagnosed conditions. The percentages of participants who reported nonuse of apps for emergency care guidance, diagnosis assistance, and diagnosed conditions were 76.72% (1654/2156), 71.62% (1542/2153), and 70.49% (1517/2155), respectively.

Table 3. Frequency of using different types of mHealth apps in a cross-sectional sample of Czech internet users (n=2175).

Type of app	Never, n (%)	A few times at most, n (%)	Several times a month, n (%)	Several times a week, n (%)	Daily, n (%)	Several times a day, n (%)
Sports (n=2158)	449 (20.83)	489 (22.66)	394 (18.26)	419 (19.42)	328 (15.2)	79 (3.66)
Number of steps (n=2167)	60 (2.77)	189 (8.72)	224 (10.34)	264 (12.18)	1174 (54.18)	256 (11.81)
Nutrition (n=2161)	803 (37.16)	565 (26.15)	299 (13.84)	219 (10.13)	222 (10.27)	53 (2.45)
Vitals (n=2165)	382 (17.64)	418 (19.31)	392 (18.11)	362 (16.72)	503 (23.23)	108 (4.99)
Sleep (n=2163)	530 (24.5)	408 (18.86)	277 (12.81)	257 (11.88)	644 (29.77)	47 (2.17)
Mood and well-being (n=2156)	1103 (51.16)	447 (20.73)	222 (10.3)	177 (8.21)	180 (8.35)	27 (1.25)
Diagnosed conditions (n=2155)	1517 (70.49)	266 (12.36)	136 (6.32)	104 (4.83)	105 (4.88)	24 (1.12)
Reproductive health (n=2152)	1327 (61.52)	233 (10.8)	314 (14.56)	125 (5.8)	131 (6.07)	27 (1.25)
Diagnosis assistance (n=2153)	1542 (71.62)	323 (15)	134 (6.22)	81 (3.76)	58 (2.69)	15 (0.7)
Emergency care guidance (n=2156)	1654 (76.72)	267 (12.38)	90 (4.17)	78 (3.62)	56 (2.6)	11 (0.51)

Factor Analysis for mHealth Apps

The results confirmed the 2-factor structure of the mHealth app measure with good construct validity (χ^2_{26}=71.9, comparative fit index=0.99, Tucker-Lewis index=0.99, root-mean-square error of approximation=0.03, and standardized root-mean-square residual=0.03). The factor loadings are shown in Figure 1. The results revealed that mHealth usage could be categorized into two dimensions: mHealth app usage for promoting wellness and mHealth app usage for managing illness.

XSL•FO
RenderX

Figure 1. Two-factor structure of mobile health app usage identified using exploratory structural equation modeling in a cross-sectional sample of Czech internet users (n=2152). The numerical values represent factor loadings. **P<.001.

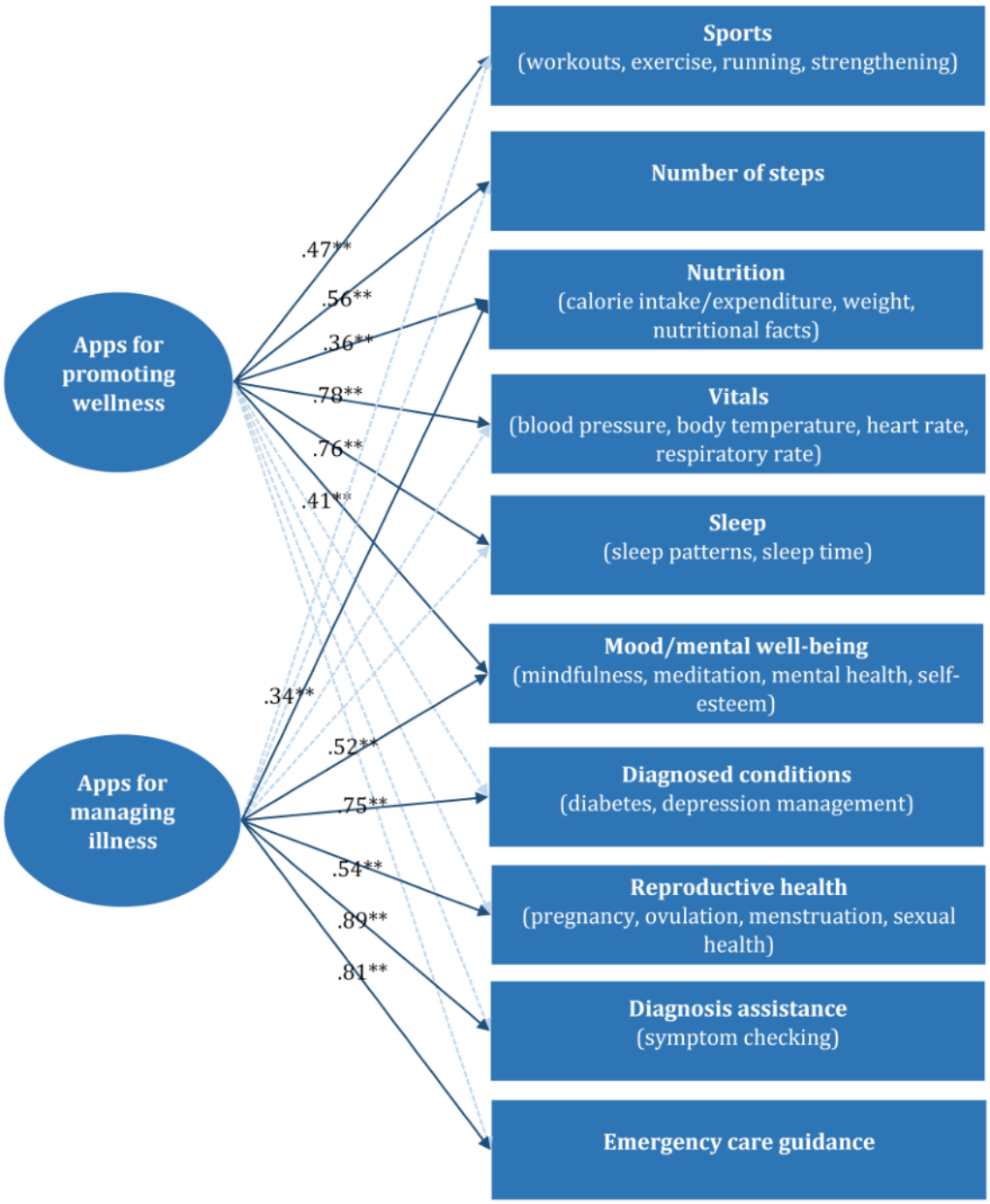

Characteristics of Wellness and Illness App Users

Table 4 presents the correlations between mHealth app usage and the predictors.

XSL•FO
RenderX

Table 4. Spearman correlations between mHealth app use and sociodemographic, digital, and health-related variables in a cross-sectional sample of Czech internet users[a].

Variables	Wellness apps[b]	Illness apps[c]	Age	Gender	Education	Income	eHealth literacy	SNS active[d]	SNS passive[e]	Sx[f]	BMI	PA[g]	ED risk[h]
Wellness apps													
ρ [i]	1												
P value	—[j]												
n	—												
Illness apps													
ρ	0.645	1											
P value	<.001	—											
n	2154	—											
Age													
ρ	–0.207	–0.27	1										
P value	<.001	<.001	—										
n	2178	2161	—										
Gender													
ρ	0.031	–0.08	0.009	1									
P value	0.15	<.001	0.55	—									
n	2175	2159	4771	—									
Education													
ρ	0.007	–0.078	–0.077	–0.053	1								
P value	0.75	<.001	<.001	<.001	—								
n	2178	2161	4775	4771	—								
Income													
ρ	0.043	–0.042	–0.169	0.148	0.273	1							
P value	0.05	0.05	<.001	<.001	<.001	—							
n	2178	2161	4775	4771	4775	—							
eHealth literacy													
ρ	0.123	0.086	–0.193	0.006	0.183	0.155	1						
P value	<.001	<.001	<.001	0.69	<.001	<.001	—						
n	2174	2157	4756	4752	4756	4756	—						
SNS active													
ρ	0.202	0.47	–0.298	–0.07	–0.057	–0.051	0.1	1					
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—					
n	2164	2147	4732	4728	4732	4732	4715	—					
SNS passive													
ρ	0.223	0.462	–0.243	–0.155	–0.012	–0.04	0.165	0.668	1				
P value	<.001	<.001	<.001	<.001	0.4	0.005	<.001	<.001	—				
n	2165	2148	4740	4736	4740	4740	4723	4731	—				
Sx													
ρ	0.099	0.242	0.009	–0.146	–0.106	–0.147	–0.032	0.245	0.304	1			
P value	<.001	<.001	0.52	<.001	<.001	<.001	0.03	<.001	<.001	—			
n	2178	2161	4766	4762	4766	4766	4747	4724	4732	—			

XSL•FO
RenderX

Variables	Wellness apps[b]	Illness apps[c]	Age	Gender	Education	Income	eHealth literacy	SNS active[d]	SNS passive[e]	Sx[f]	BMI	PA[g]	ED risk[h]
BMI													
ρ	0.005	–0.047	0.279	0.105	–0.112	–0.019	–0.085	–0.06	–0.057	0.086	1		
P value	0.83	0.03	<.001	<.001	<.001	0.2	<.001	<.001	<.001	<.001	—		
n	2178	2161	4775	4771	4775	4775	4756	4732	4740	4764	—		
PA													
ρ	0.179	0.093	–0.19	0.072	0.13	0.128	0.16	0.088	0.108	–0.055	–0.137	1	
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	
n	2034	2019	4299	4295	4299	4299	4283	4259	4265	4295	4299	—	
ED risk													
ρ	0.199	0.307	–0.116	–0.14	0.02	–0.009	0.053	0.227	0.296	0.307	0.276	0.154	1
P value	<.001	<.001	<.001	<.001	0.18	0.52	<.001	<.001	<.001	<.001	<.001	<.001	—
n	2164	2147	4731	4727	4731	4731	4712	4689	4697	4726	4731	4273	—

[a]Sample sizes vary by correlation pair; Ns are shown within the table.

[b]Wellness apps: frequency of using mHealth apps for promoting wellness.

[c]Illness apps: frequency of using mHealth apps for managing illness.

[d]SNS active: active use of social networking sites (SNS) for health information.

[e]SNS passive: passive use of social networking sites (SNS) for health information.

[f]Sx: perceived severity of symptoms.

[g]PA: physical activity.

[h]ED risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).

[i]ρ: Spearman rank order correlation.

[j]Not applicable.

Aligned with the ESEM method for factor analysis, the current model did not restrict the factor loadings of mHealth apps to a particular factor. Table 5 shows the results. When interpreting the results, we focused not only on statistical significance but also on the effect sizes. Following widely acknowledged guidelines, standardized path coefficients (β) between 0.1 and 0.3 were interpreted as indicating a small effect, those between 0.3 and 0.5 as medium, and those above 0.5 as large effects [39]. In the first model, which included only sociodemographic independent variables, app use to promote wellness was predicted by age (standardized β=–0.22, *P*<.001), indicating that older age was correlated with less mHealth app use to promote wellness. Except for gender, sociodemographic characteristics, including age, educational level, and income status, were significantly linked to app use in managing illness. However, the effect size of income was negligible (retaining three decimal places, standardized β=0.097). Likewise, the effect of education on illness-related app use was also at the boundary of a negligible effect size (retaining three decimal places, standardized β=–0.106). Furthermore, after including health-, digital knowledge-, and use-related variables, none of the sociodemographic variables remained significant. We did not interpret them further because both education level and income were already at the boundary of a negligible effect size when only sociodemographic variables were considered, and adding additional variables merely pushed them fully into the negligible range. Notably, however, the reduction in effect size of age was considerable for both app types, which prompted us to conduct further analyses to examine the mediating roles of digital knowledge, digital use, and health indicators in the association between age and wellness- and illness-related app use.

XSL•FO
RenderX

Table 5. Predictors of the frequency of using mHealth apps to promote wellness and manage illness based on exploratory structural equation modeling in a cross-sectional sample of Czech internet users[a].

	Model 1 (n=2152)			Model 2 (n=1985)		
Predictors	Standardized β	SE	*P* value	Standardized β	SE	*P* value
Apps to promote wellness						
Age	–0.22	0.00	<.001	–0.18	0.00	<.001
Gender	0.03	0.03	.07	0.04	0.05	.04
Education	0.03	0.02	.15	0.02	0.03	.24
Income	0.02	0.02	.27	0.02	0.02	.42
eHealth literacy	—[b]	—	—	0.10	0.02	<.001
Active SNS use[c]	—	—	—	0.07	0.04	.12
Passive SNS use[d]	—	—	—	0.08	0.05	.07
Symptom severity	—	—	—	0.04	0.03	.08
BMI	—	—	—	0.08	0.03	<.001
Physical activity	—	—	—	0.15	0.02	<.001
ED-related risk[e]	—	—	—	0.09	0.03	<.001
Apps to manage illness						
Age	–0.23	0.00	<.001	–0.04	0.00	.07
Gender	0.02	0.03	.31	0.01	0.08	.75
Education	–0.11	0.02	<.001	–0.05	0.04	.05
Income	–0.10	0.02	<.001	–0.02	0.04	.40
eHealth literacy	—	—	—	0.00	0.04	.93
Active SNS use[c]	—	—	—	0.62	0.15	<.001
Passive SNS use[d]	—	—	—	0.06	0.11	.49
Symptom severity	—	—	—	0.04	0.04	.20
BMI	—	—	—	–0.07	0.04	.01
Physical activity	—	—	—	0.02	0.03	.46
ED-related risk[e]	—	—	—	0.11	0.04	<.001

[a]Sample sizes vary across models; Ns are reported in the table.

[b]Not applicable.

[c]SNS active: active use of social networking sites for health information.

[d]SNS passive: passive use of social networking sites for health information.

[e]ED-related risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).

The results also suggested that eHealth literacy and physical activity were significantly correlated with the frequency of mHealth app usage to promote wellness. Regarding the use of mHealth apps to manage illness, the active use of SNS for health-related information and eating disorder–related risk propensity were significant predictors of the frequency of use.

For a more robust interpretation, we additionally performed basic linear regression analyses using the ordinary least squares method (see Table 6). The standardized estimates did not differ substantially from the main analyses, except that active use of SNS for health-related information (SNS active) was significant in the simple linear regression model but was not significantly associated with mHealth app usage for promoting wellness in the full model. It is worth noting, however, that the effect size was small (standardized β=0.13, *P*<.001). Similarly, in the ESEM model, the effect size of eating disorder risk on wellness-related mHealth app usage was negligible (standardized β=0.09), and in the ordinary least squares regression model, it was 0.12—still very small, bordering on a negligible effect size. Such a minor difference cannot be considered meaningful. Considering that the ESEM model accounted for factor loadings and that the effect size in the simple regression remained small, we regarded this discrepancy as acceptable. Therefore, the results from the ESEM model (Table 5) should be prioritized for interpretation.

XSL•FO
RenderX

Table 6. Predictors of the frequency of using mHealth apps to promote wellness and manage illness based on simple regression models in a cross-sectional sample of Czech internet users[a].

	Model 1			Model 2		
Predictors	Standardized β	SE	*P* value	Standardized β	SE	*P* value
Apps to promote wellness (n=2005)						
Age	–0.22	0.00	<.001	–0.15	0.00	<.001
Gender (2-1)[b]	0.07	0.04	.08	0.10	0.04	.02
Education						
3-4[c]	–0.01	0.05	.91	0.01	0.05	.81
2-4[d]	0.05	0.06	.41	0.05	0.06	.37
1-4[e]	–0.12	0.10	.23	–0.13	0.09	.18
Income	0.00	0.02	.88	0.02	0.02	.48
eHealth literacy	—[f]	—	—	0.08	0.03	<.001
Active SNS use[g]	—	—	—	0.13	0.03	<.001
Passive SNS use[h]	—	—	—	0.09	0.03	.004
Symptom severity	—	—	—	0.04	0.03	.12
BMI	—	—	—	0.04	0.00	.12
Physical activity	—	—	—	0.15	0.01	<.001
ED-related risk[i]	—	—	—	0.12	0.02	<.001
Apps to manage illness (n=1991)						
Age	–0.26	0.00	<.001	–0.09	0.00	<.001
Gender (2-1)[b]	–0.02	0.04	.64	0.00	0.03	.85
Education						
3-4[c]	0.03	0.05	.62	0.02	0.04	.62
2-4[d]	0.26	0.06	<.001	0.14	0.05	.004
1-4[e]	0.13	0.09	.17	0.02	0.07	.84
Income	–0.08	0.02	<.001	–0.01	0.02	.50
eHealth literacy	—	—	—	0.02	0.02	.19
Active SNS use[g]	—	—	—	0.51	0.02	<.001
Passive SNS use[h]	—	—	—	0.07	0.02	.01
Symptom severity	—	—	—	0.04	0.02	.05
BMI	—	—	—	–0.04	0.00	.03
Physical activity	—	—	—	0.05	0.01	.003
ED-related risk[i]	—	—	—	0.14	0.01	<.001

[a]Sample sizes vary across models; Ns are reported in the table.

[b]2-1: male-female.

[c]3-4: secondary school – university (or higher vocational school).

[d]2-4: secondary vocational school – university (or higher vocational school).

[e]1-4: none/primary education – university (or higher vocational school).

[f]Not applicable.

[g]SNS active: active use of social networking sites for health information.

[h]SNS passive: passive use of social networking sites for health information.

[i]ED-related risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).

XSL•FO
RenderX

Mediation Analysis

Due to the decrease in the effect size of age on two dependent variables after the addition of digital knowledge, digital use, and health indicators, it was worthwhile to explore further whether these variables mediated the relationship between age and mHealth usage. The mediation analysis suggested that older age correlated with more frequent use of wellness apps through "eHealth literacy," "active SNS use," "passive SNS use," "physical activity," and "eating disorder–related risk propensity" (see Figure 2). On the other hand, "active SNS use," "passive SNS use," and "eating disorder–related risk propensity" mediated the relationship between age and frequency of mHealth app use to manage illness (see Figure 3). It is worth noting that the correlations between active and passive SNS use and illness-related mHealth app usage were more robust than those of wellness-related mHealth app usage.

Figure 2. Mediation model depicting associations between age and wellness-related mobile health app use based on exploratory structural equation modeling in a cross-sectional sample of Czech internet users (n=1985). Only statistically significant paths (P<.05) are shown. Standardized coefficients are displayed. *SNS active: active use of social networking sites for health information. **SNS passive: passive use of social networking sites for health information. ***ED-related risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).

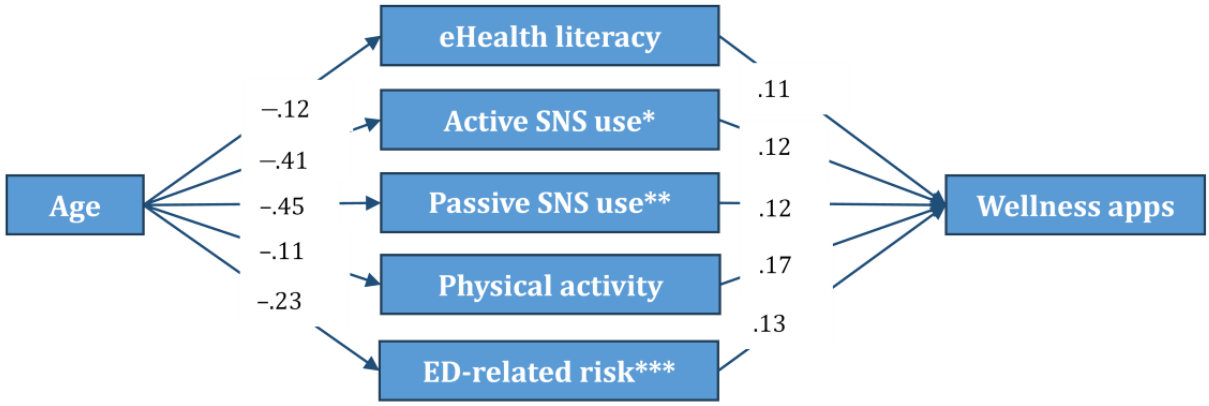

Figure 3. Mediation model depicting associations between age and illness-related mobile health app use based on exploratory structural equation modeling in a cross-sectional sample of Czech internet users (n=1985). Only statistically significant paths (P<.05) are shown. Standardized coefficients are displayed. *SNS active: active use of social networking sites for health information. **SNS passive: passive use of social networking sites for health information. ***ED-related risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).

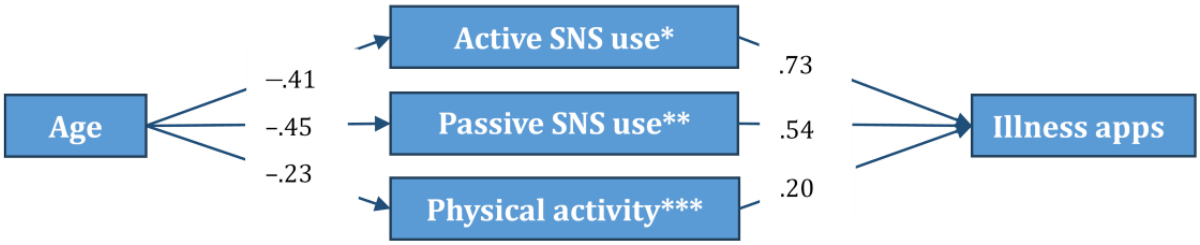

Discussion

Overview

This study evaluated the prevalence, user characteristics, and usage patterns of various mHealth apps in a representative sample of Czech adults with internet access. Furthermore, it distinguished between wellness and illness apps using a data-driven approach and identified user characteristics associated with the frequency of use of each dimension. This study provided significant insights into the relationship between user characteristics, intentions, and engagement with mHealth apps.

Principal Results

Prevalence of mHealth App Use

Around half of the participants with internet access (48.92%) had been using mHealth apps. Previous studies with adults have reported prevalence rates ranging from 34.1% to 56.6% in the United States [9-12]. Lower rates were reported in European and Far Eastern countries. For instance, 20.5% of German

XSL•FO
RenderX

smartphone owners in 2015 [18] and 24.1% of Hong Kongian smartphone or tablet owners in 2016 [23] used mHealth apps in population-based samples. More up-to-date findings are needed for a more accurate comparison of the prevalence rates across countries.

Characteristics of mHealth App Users

Sociodemographic Characteristics

Age was significantly associated with using mHealth apps. In line with earlier studies [8], younger adults were more likely to use them. This finding highlights that older adults are underserved in terms of mHealth app use. Lack of awareness, lack of motivation, low self-efficacy, mistrust of technology, and the absence of required technical skills are frequently reported barriers to using mHealth apps among older adults [40,41]. Additionally, cognitive impairment, reduced physical ability, and visual changes associated with aging can compromise the usability of app interfaces in this population [42-45]. Dispositional, attitudinal, and aging-related factors must be considered when addressing this population via apps.

Participants with monthly incomes exceeding 50,000 CZK were more likely to have mHealth apps than those in other income categories (ie, up to 20,000 CZK, between 20,001 and 35,000 CZK, and between 35,001 and 50,000 CZK). This finding is consistent with previous studies that have reported higher annual income as a significant predictor of mHealth app adoption and usage [9,11,12,23]. Higher-income status may facilitate access to apps that require payment to download and use, whereas such apps may be less affordable for low-income individuals. The association between educational level and mHealth app usage was not statistically significant. This finding contrasts with previous studies conducted in the United States [9-12] and Hong Kong [23], but it is similar to an earlier study conducted in Germany [18]. Future studies will clarify whether educational background will influence the adoption and use of apps as they become more affordable, user-friendly, and accessible.

Digital Knowledge and Use Factors

Adults with higher eHealth literacy were more likely to have mHealth apps. This finding aligns with previous studies [14,46] and confirms that eHealth literacy is a crucial digital skill for understanding mHealth app use behaviors. Individuals with higher eHealth literacy tend to seek online health information more frequently [47], are more health information-oriented [48], and place greater importance on health-related information [49]. They are also more likely to report competence in self-care [50] and their intention to promote health [51]. These findings underscore that perceived digital skills in finding, understanding, evaluating, and applying online health information in a health context might be precursors to having and using apps.

This study provided initial evidence in support of the association between SNS activity for health information and having mHealth apps. We found that passive use (ie, reading, watching, or following health content) was significantly associated with the use of apps. However, the association between active use (ie, creating, sharing, and commenting on health content) and app use was insignificant. Our results reveal that adopting mHealth apps might be more closely related to consuming health-related information on SNS (ie, passive use) than engaging with online networks for health-related information (ie, active use). This finding aligns with a previous study that reported a significant association between health information-seeking behaviors on social media and the use of mHealth apps for health information [52]. Health information–oriented individuals may use SNS and mHealth apps in a complementary manner to meet their health information needs. Additionally, SNS provides a platform for promoting various functions of mHealth apps [53], which may help meet the personal health care needs of consumers and contribute to the adoption of mHealth apps. These findings suggest that SNS could be a significant promotional channel.

Health Indicators

In line with previous research [10,18,23], participants with higher physical activity were more likely to have mHealth apps, which provides further evidence for the association between a physically active lifestyle and mHealth app adoption. We found a significant association between eating disorder–related risks and mHealth app use. Participants who reported higher dietary restraint, weight and shape overvaluation, and body dissatisfaction were more likely to have and use mHealth apps. Prior studies reporting an association between eating disorder–related risks and app usage have primarily focused on apps that promote weight loss and physical activity [28-30]. This study shows that body dissatisfaction, weight and shape overvaluation, and restrictive eating behaviors correlate with overall app adoption.

BMI was not associated with using apps, contrasting with earlier studies that reported a significant association between app usage and a higher BMI [10,18,25]. It is possible that eating disorder–related risks (ie, body dissatisfaction, weight/shape overvaluation, and dieting behaviors) capture app usage behaviors better than body weight. We did not observe a significant association between the use of apps and symptom severity. This finding aligns with studies that found no significant association between health status and mHealth app usage [10-12]. This study focused on a detailed measurement of perceived health status by assessing somatic symptom severity in multiple domains.

mHealth App Use by Type of App

Apps that count the number of steps were most frequently used. This finding aligns with global statistics, which indicate that the most downloaded app in 2022 was a step counter [54]. One reason for the frequent use of step counter apps could be related to their simplicity. Previous research showed this to be a significant factor in determining the frequency of using apps for physical activity [25]. Apps that count the number of steps may take less effort from consumers to integrate them into everyday life. These apps are often built-in apps delivered in the latest releases of major smartphone companies, and their availability could enhance increased adoption and use. Apps that track sleep, vitals, and sports activities were also popular. In a recent qualitative study, app users reported that their continued use of apps with tracking functions (eg, tracking steps, exercise, heart rate, and sleep) was associated with the perceived usefulness of these apps for maintaining and improving their health and well-being [55].

XSL•FO
RenderX

Factor Analysis for mHealth Apps

Using a data-driven approach with ESEM, we confirmed the 2-factor structure of mHealth app usage in the analyses. The apps under the "promoting wellness" factor were related to lifestyle behaviors (eg, sports apps, nutrition apps, and number of steps apps), monitoring of health indicators (eg, vitals apps and sleep apps), and the enhancement of mental vitality (eg, mood and mental well-being apps). This factor consisted of apps focusing on preventive health behaviors to promote wellness and reduce the risk of illness. The apps under the "managing illness" factor focused on maintaining health (eg, apps for reproductive health and diagnosis assistance) and providing support for illnesses (eg, apps for diagnosed conditions and emergency care guidance). We found that nutrition apps and mood and mental well-being apps cross-loaded on both factors. Several health conditions (eg, diabetes) require close monitoring of nutritional intake [56]. Therefore, it is reasonable that app users might monitor nutrition not only for promoting wellness but also for managing diagnosed health conditions. Similarly, mood and mental well-being are broad domains, and app users' intentions may have dual relevance, encompassing both the promotion of mood and mental well-being for wellness purposes and the management of mental health problems as part of a treatment regimen [57]. In conclusion, the factor structure of our mHealth measure was feasible. It might guide future research studies on app usage for health promotion and illness management.

Characteristics of Wellness and Illness App Users

Age was significantly associated with the frequency of using wellness apps. Older adults used them less frequently to promote wellness. This finding is disappointing because older adults are more susceptible to chronic conditions like diabetes, hypertension, and cardiovascular disease [58]. Effective characteristics of wellness apps that enhance the adherence of older adults should be identified to combat their digital exclusion.

eHealth literacy was significantly associated with more frequent use of apps to promote wellness, but no such association was observed for apps that manage illness. Previous research has confirmed that higher eHealth literacy is associated with perceiving mHealth apps as effective and using them more frequently for health-related behavioral change [14,15,45]. Our findings indicate that this association might be particularly relevant for behaviors that promote health. Individuals with higher eHealth literacy are more motivated to engage in behaviors to regulate their diet, sleep, and physical activity [59]. They are more likely to adhere to recommendations to reduce the risk of illness [60] and to manage stress [51]. Having the necessary digital skills to navigate within apps might reinforce the regular use of wellness apps in this population, which is inherently more motivated to engage in health-promoting behaviors.

We found a significant association between active SNS use for health information and the more frequent usage of apps to manage illness. This finding suggests that proactive engagement with online social networks for health information plays a crucial role in the frequent use of illness apps. A recent study reported a significant association between having a friend to discuss health-related issues and the use of mHealth apps [12]. We consider that our finding aligns with this previous study because the active use of SNS promises a platform for individuals to interact with their close networks about health and connect with others who might have similar health concerns. Users of illness apps might use SNS complementarily to exchange health information, share experiences, discuss health-related decisions, and provide and receive support from their online social networks [61]. The active use of these networks can synergistically encourage compliance with apps to manage illness, as they offer real-time feedback and a direct line of communication to those who may face similar health concerns.

Higher physical activity was associated with a more frequent use of wellness apps. Wellness apps (eg, Fitbit) offer a wide range of functions that physically active individuals can use to monitor and regulate their physical activity, conditioning, and fitness [62]. They also offer a platform to manage their diet, vitals, energy levels, and mental state [62], which addresses well-being holistically. These apps may be particularly appealing to physically active adults who may be inherently more motivated to pursue a healthy lifestyle [63,64] and thus use them more regularly to enhance their well-being.

This study provided initial evidence to indicate that individuals with elevated risks of eating disorders use mHealth apps more frequently to manage illness. Body image concerns and dieting behaviors may be influenced by various health conditions beyond eating disorders (eg, obesity, diabetes, and depression) [65,66]. The patterns and consequences of using apps can differ substantially depending on the health domain for which they are used. Previous studies focused on weight loss and physical activity apps to examine the association between unintended uses and the relative consequences of using these apps in exacerbating or triggering adverse outcomes among individuals with body image concerns and dieting behaviors [67-69]. These previous findings underscored that underweight BMI goals had a detrimental impact on eating disorder–related symptoms and behaviors, such as more restrictive eating behaviors, control, and obsession with eating and exercise. In contrast, more realistic goals regarding weight and shape, along with motivation toward healthy lifestyle choices, were associated with improved food choices, exercise behaviors, and health outcomes. Therefore, future research should specify health behavior goals and the consequences of using apps to manage illness in individuals with elevated body image and eating concerns.

Mediation Analysis

As mentioned earlier, while conducting the analyses, we observed that the effect size of age was considerably reduced for both app types after including digital knowledge, digital use, and health-related variables in the equation. Therefore, although it was not among our initial research aims, we decided to explore whether any of these variables mediated the association between age and app usage. The results revealed important insights into the association between older age and the less frequent use of apps for wellness and illness purposes.

The association between age and mHealth app use was negatively mediated by the active and passive use of SNS for

XSL•FO
RenderX

health information. In other words, older adults were less likely to use SNS to seek health information (ie, passive use) and to interact with online networks about health (ie, active use), which was significantly associated with less frequent app use. This association was significant for both app types and notably stronger for apps to manage illness. Compared with younger adults, older adults use social media and search engines less frequently for health information [52]. They are also more likely to prefer traditional media (eg, television and newspaper) [70]. Our findings indicate that a lack of orientation toward seeking and exchanging health information on SNS could hinder older adults' use of mHealth apps. SNS may provide know-how about app functions and facilitate social support for sustained use of mHealth apps.

The association between age and wellness app use was negatively mediated by eHealth literacy. Older adults had lower eHealth literacy, which was significantly associated with less frequent use of wellness apps. Previous studies have confirmed the associations between older age, lower eHealth literacy, and the use of mHealth apps [8,71]. Our finding shows that enhancing eHealth literacy skills could significantly increase app compliance for health behaviors that promote well-being in this population.

The association between age and wellness app use was negatively mediated by physical activity. Older adults reported lower physical activity levels, which were significantly associated with less frequent use of wellness apps. Older age is characterized by reduced physical activity [72], and older adults are less likely to adhere to recommendations for physical exercise [73,74]. However, a physically active lifestyle is associated with a reduced risk of all-cause mortality, a better quality of life, and improved cognitive functioning in older age [75]. Wellness apps offer novel opportunities to promote physical activity, reduce sedentary behaviors, and enhance physical and mental well-being in this population [76]. Thus, our finding highlights a significant gap in their dissemination to older adults, who might need such strategies most. The obstacles associated with lower acceptance and the usage of wellness apps and strategies to empower older adults' self-efficacy regarding healthy aging should be identified.

Finally, the association between age and mHealth app use was negatively mediated by eating disorder–related risk propensity. This association was significant for both app types. In other words, older adults were less likely to report body image concerns and dieting behaviors, which, in turn, were significantly associated with less frequent use of apps for wellness and illness purposes. The importance placed on physical appearance is reduced by increasing age. Body dissatisfaction, weight and shape overvaluation, and dieting behaviors are reported less frequently after the age of 40 [77]. In general, lower body image preoccupation is associated with healthier lifestyle behaviors, psychological adjustment, and better management of health conditions that require close monitoring of health indicators and dietary behaviors [65,78,79]. Therefore, it is interesting to find that a lower eating disorder–related risk is associated with less frequent use of apps in older adults. Further research is needed to understand the body image perceptions, dieting behaviors, and weight- and shape-related concerns of older adults, as well as their associations with the use of apps.

Limitations and Future Research

The findings of this study should be interpreted in light of its limitations. Due to the cross-sectional study design, temporal and causal associations between the variables could not be assessed. For instance, we cannot ascertain whether eating disorder–related risks determine mHealth app use or whether they are exacerbated due to the use of mHealth apps. Our mHealth app measure consisted of apps frequently downloaded in app stores for health purposes. However, it is not exhaustive and can be improved in later studies. We used a data-driven approach (ESEM) to categorize app usage into wellness and illness dimensions. This approach enabled us to demonstrate the multidimensional structure of mHealth app usage. Wellness and illness dimensions also drew parallels between preventive health behaviors and illness behaviors in the health context [80]. Although ESEM allows for more flexible estimation of latent constructs compared with traditional CFA, its results can still be sensitive to model specification and sample characteristics. Thus, despite ESEM being a better option in this study, the findings should be interpreted with caution, and future studies are encouraged to replicate the model using different samples or model structures.

We determined the usage patterns of apps based on their frequency of use. This approach limited our ability to identify the healthy and unhealthy uses of apps. For instance, excessive use of sports apps can be dysfunctional when users engage with them to the point of overexercising. Weight loss apps might have adverse consequences for at-risk individuals if they are used frequently due to an obsession with eating and body image concerns.

The assessments were based on self-reports. Thus, we could not control the response characteristics of the participants. We used adapted versions of frequently used questionnaires to assess symptom severity and SNS use for health information. Additionally, eating disorder–related risk propensity and eHealth literacy measures have not been previously adequately validated in the Czech context. Nevertheless, all the measures had adequate internal consistency in this study. Besides, we conducted cognitive testing with a subsample of participants to ensure the comprehensibility of items and the quality of the assessment procedure.

Future research is needed to investigate the health behaviors and outcomes associated with the use of apps. Usage intentions and motivations are essential to understanding the healthy and unhealthy uses of mHealth apps. Therefore, future studies should consider the health behavior goals associated with using apps and identify the factors that contribute to unintended uses and consequences.

Conclusions

This study demonstrates that sociodemographic, digital, and health-related factors might jointly shape engagement with mHealth apps. eHealth literacy and health-related SNS activity emerged as key digital correlates of use, highlighting the potential of promoting digital inclusion to support equitable

XSL•FO
RenderX

participation in digital health. Eating disorder–related risk propensity—reflecting body image concerns and dietary restraint—was associated with the overall app adoption and frequent use of apps to manage illness, underscoring the importance of examining app use purposes and health outcomes among individuals with elevated body image and dietary concerns, and the need for future research and app development to promote health-supportive use while minimizing appearance-driven behaviors. Indirect associations of age through eHealth literacy, health-related SNS use, physical activity, and eating disorder–related risk further highlight multidimensional barriers older adults face in adopting and maintaining mHealth app use. Collectively, these findings indicate that mHealth engagement reflects broader social, digital, and psychological inequalities rather than individual preferences alone. Enhancing digital skills, promoting inclusive app design, and addressing body image- and diet-related use may foster equitable mHealth participation and help prevent widening health disparities across age and user groups.

Data Availability

The data and the analysis scripts are shared via the Open Science Framework [81].

Funding

The research reported in this publication was supported by the NPO “Systemic Risk Institute” grant LX22NPO5101, funded by the European Union—Next Generation EU (Ministry of Education, Youth and Sports, NPO: EXCELES). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional analysis of mobile health app usage.
[XLSX File (Microsoft Excel File), 20 KB - jmir_v28i1e71363_app1.xlsx]

References

1. Wells C, Spry C. CADTH Horizon Scan: an overview of smartphone apps. Can J Health Technol 2022;2(2):1-26. [doi: 10.51731/cjht.2022.269]
2. Peng C, He M, Cutrona SL, Kiefe CI, Liu F, Wang Z. Theme trends and knowledge structure on mobile health apps: bibliometric analysis. JMIR Mhealth Uhealth 2020;8(7):e18212 [FREE Full text] [doi: 10.2196/18212] [Medline: 32716312]
3. Wu CC, Huang CW, Wang YC, Islam M, Kung W, Weng Y, et al. mHealth research for weight loss, physical activity, and sedentary behavior: bibliometric analysis. J Med Internet Res 2022;24(6):e35747 [FREE Full text] [doi: 10.2196/35747] [Medline: 35675126]
4. Sama PR, Eapen ZJ, Weinfurt KP, Shah BR, Schulman KA. An evaluation of mobile health application tools. JMIR Mhealth Uhealth 2014;2(2):e19 [FREE Full text] [doi: 10.2196/mhealth.3088] [Medline: 25099179]
5. Stec M, Arbour MW. Wellness and disease self-management mobile health apps evaluated by the Mobile Application Rating Scale. Adv Fam Pract Nurs 2020;2:87-102. [doi: 10.1016/j.yfpn.2020.01.003]
6. Xu W, Liu Y. mHealth apps: a repository and database of mobile health apps. JMIR Mhealth Uhealth 2015;3(1):e28 [FREE Full text] [doi: 10.2196/mhealth.4026] [Medline: 25786060]
7. Iribarren SJ, Akande TO, Kamp KJ, Barry D, Kader YG, Suelzer E. Effectiveness of mobile apps to promote health and manage disease: systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2021;9(1):e21563 [FREE Full text] [doi: 10.2196/21563] [Medline: 33427672]
8. Bao H, Lee EWJ. Examining theoretical frameworks and antecedents of health apps and wearables use: a scoping review. Health Commun 2024;39(12):2671-2681. [doi: 10.1080/10410236.2023.2283655] [Medline: 37968803]
9. Bhuyan SS, Lu N, Chandak A, Kim H, Wyant D, Bhatt J, et al. Use of mobile health applications for health-seeking behavior among US adults. J Med Syst 2016;40(6):153. [doi: 10.1007/s10916-016-0492-7] [Medline: 27147516]
10. Carroll JK, Moorhead A, Bond R, LeBlanc WG, Petrella RJ, Fiscella K. Who uses mobile phone health apps and does use matter? A secondary data analytics approach. J Med Internet Res 2017;19(4):e125 [FREE Full text] [doi: 10.2196/jmir.5604] [Medline: 28428170]
11. Rising CJ, Jensen RE, Moser RP, Oh A. Characterizing the US population by patterns of mobile health use for health and behavioral tracking: analysis of the National Cancer Institute's Health Information National Trends Survey data. J Med Internet Res 2020;22(5):e16299 [FREE Full text] [doi: 10.2196/16299] [Medline: 32406865]
12. Tundealao S, Titiloye T, Sajja A, Egab I, Odole I, Alufa O, et al. Factors associated with the non-use of mobile health applications among adults in the United States. J Public Health (Berl.) 2023;33(7):1575-1581. [doi: 10.1007/s10389-023-02132-8]

XSL•FO
RenderX

13. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
14. Vervier LS, Valdez AC, Ziefle M. "Attitude"- mHealth apps and users' insights: an empirical approach to understand the antecedents of attitudes towards mHealth applications. In: Proceedings of the 5th International Conference on Information and Communication Technologies for Ageing Well and e-Health (ICT4AWE 2019). Setúbal, Portugal: SciTePress; 2019:213-221.
15. Durmuş A. The influence of digital literacy on mHealth app usability: the mediating role of patient expertise. Digit Health 2024;10:20552076241299061 [FREE Full text] [doi: 10.1177/20552076241299061] [Medline: 39600388]
16. Shaw G, Castro BA, Gunn LH, Norris K, Thorpe RJ. The association of eHealth literacy skills and mHealth application use among US adults with obesity: analysis of health information national trends survey data. JMIR Mhealth Uhealth 2024;12:e46656 [FREE Full text] [doi: 10.2196/46656] [Medline: 38198196]
17. Guo SH, Hsing HC, Lin JL, Lee CC. Relationships between mobile eHealth literacy, diabetes self-care, and glycemic outcomes in Taiwanese patients with type 2 diabetes: cross-sectional study. JMIR Mhealth Uhealth 2021;9(2):e18404 [FREE Full text] [doi: 10.2196/18404] [Medline: 33544088]
18. Ernsting C, Dombrowski SU, Oedekoven M, O Sullivan JL, Kanzler M, Kuhlmey A, et al. Using smartphones and health apps to change and manage health behaviors: a population-based survey. J Med Internet Res 2017;19(4):e101 [FREE Full text] [doi: 10.2196/jmir.6838] [Medline: 28381394]
19. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
20. Zhao Y, Zhang J. Consumer health information seeking in social media: a literature review. Health Info Libr J 2017;34(4):268-283 [FREE Full text] [doi: 10.1111/hir.12192] [Medline: 29045011]
21. Liu PL. COVID-19 information seeking on digital media and preventive behaviors: the mediation role of worry. Cyberpsychol Behav Soc Netw 2020;23(10):677-682. [doi: 10.1089/cyber.2020.0250] [Medline: 32498549]
22. Iftikhar R, Abaalkhail B. Health-seeking influence reflected by online health-related messages received on social media: cross-sectional survey. J Med Internet Res 2017;19(11):e382 [FREE Full text] [doi: 10.2196/jmir.5989] [Medline: 29146568]
23. Shen C, Wang MP, Chu JT, Wan A, Viswanath K, Chan SSC, et al. Health app possession among smartphone or tablet owners in Hong Kong: population-based survey. JMIR Mhealth Uhealth 2017;5(6):e77 [FREE Full text] [doi: 10.2196/mhealth.7628] [Medline: 28583905]
24. Kramer H, Cao G, Dugas L, Luke A, Cooper R, Durazo-Arvizu R. Increasing BMI and waist circumference and prevalence of obesity among adults with type 2 diabetes: the National Health and Nutrition Examination Surveys. J Diabetes Complications 2010;24(6):368-374. [doi: 10.1016/j.jdiacomp.2009.10.001] [Medline: 19914095]
25. Oba T, Takano K, Katahira K, Kimura K. Use patterns of smartphone apps and wearable devices supporting physical activity and exercise: large-scale cross-sectional survey. JMIR Mhealth Uhealth 2023;11:e49148 [FREE Full text] [doi: 10.2196/49148] [Medline: 37997790]
26. Liu D, Maimaitijiang R, Gu J, Zhong S, Zhou M, Wu Z, et al. Using the unified theory of acceptance and use of technology (UTAUT) to investigate the intention to use physical activity apps: cross-sectional survey. JMIR Mhealth Uhealth 2019;7(9):e13127 [FREE Full text] [doi: 10.2196/13127] [Medline: 31507269]
27. Rø Ø, Reas DL, Rosenvinge J. The impact of age and BMI on Eating Disorder Examination Questionnaire (EDE-Q) scores in a community sample. Eat Behav 2012;13(2):158-161. [doi: 10.1016/j.eatbeh.2011.12.001] [Medline: 22365803]
28. Elavsky S, Smahel D, Machackova H. Who are mobile app users from healthy lifestyle websites? Analysis of patterns of app use and user characteristics. Transl Behav Med 2017;7(4):891-901 [FREE Full text] [doi: 10.1007/s13142-017-0525-x] [Medline: 28929368]
29. Hahn SL, Hazzard VM, Larson N, Klein L, Loth KA, Neumark-Sztainer D. Correlates of weight-related self-monitoring application use during emerging adulthood in a population-based sample. Eat Weight Disord 2022;27(6):2107-2119 [FREE Full text] [doi: 10.1007/s40519-021-01349-4] [Medline: 35040079]
30. Linardon J, Messer M. My fitness pal usage in men: associations with eating disorder symptoms and psychosocial impairment. Eat Behav 2019;33:13-17. [doi: 10.1016/j.eatbeh.2019.02.003] [Medline: 30772765]
31. Norman CD, Skinner HA. eHEALS: the eHealth literacy scale. J Med Internet Res 2006;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
32. Escobar-Viera CG, Shensa A, Bowman ND, Sidani JE, Knight J, James AE, et al. Passive and active social media use and depressive symptoms among United States adults. Cyberpsychol Behav Soc Netw 2018;21(7):437-443. [doi: 10.1089/cyber.2017.0668] [Medline: 29995530]
33. Kroenke K, Spitzer RL, Williams JBW. The PHQ-15: validity of a new measure for evaluating the severity of somatic symptoms. Psychosom Med 2002;64(2):258-266. [doi: 10.1097/00006842-200203000-00008] [Medline: 11914441]
34. Kroenke K, Spitzer RL, Williams JBW, Löwe B. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen Hosp Psychiatry 2010;32(4):345-359. [doi: 10.1016/j.genhosppsych.2010.03.006] [Medline: 20633738]

XSL•FO
RenderX

35. Grilo CM, Reas DL, Hopwood CJ, Crosby RD. Factor structure and construct validity of the eating disorder examination-questionnaire in college students: further support for a modified brief version. Int J Eat Disord 2015;48(3):284-289 [FREE Full text] [doi: 10.1002/eat.22358] [Medline: 25346071]
36. Fairburn CG, Beglin SJ. Assessment of eating disorders: interview or self-report questionnaire? Int J Eat Disord 1994;16(4):363-370. [Medline: 7866415]
37. van Zyl LE, Olckers C, Roll LC. The psychometric properties of the Grit-O scale within the Twente region in Netherlands: an ICM-CFA vs. ESEM approach. Front Psychol 2020;11:796 [FREE Full text] [doi: 10.3389/fpsyg.2020.00796] [Medline: 32457679]
38. Jordan SL, Ferris GR, Hochwarter WA, Wright TA. Toward a work motivation conceptualization of Grit in organizations. Group Organ Manag 2019;44(2):320-360. [doi: 10.1177/1059601119834093]
39. Nieminen P. Application of standardized regression coefficient in meta-analysis. BioMedInformatics 2022;2(3):434-458. [doi: 10.3390/biomedinformatics2030028]
40. Aranha M, James K, Deasy C, Heavin C. Exploring the barriers and facilitators which influence mHealth adoption among older adults: a literature review. Gerontechnology 2021;20(2):1-16. [doi: 10.4017/gt.2021.20.2.424.06]
41. Ahmad NA, Mat Ludin AF, Shahar S, Mohd Noah SA, Mohd Tohit N. Willingness, perceived barriers and motivators in adopting mobile applications for health-related interventions among older adults: a scoping review. BMJ Open 2022;12(3):e054561 [FREE Full text] [doi: 10.1136/bmjopen-2021-054561] [Medline: 35264349]
42. Wildenbos GA, Peute LW, Jaspers MWM. A framework for evaluating mHealth tools for older patients on usability. Stud Health Technol Inform 2015;210:783-787. [Medline: 25991261]
43. Wildenbos GA, Peute L, Jaspers M. Aging barriers influencing mobile health usability for older adults: a literature based framework (MOLD-US). Int J Med Inform 2018;114:66-75. [doi: 10.1016/j.ijmedinf.2018.03.012] [Medline: 29673606]
44. Wildenbos GA, Jaspers MWM, Schijven MP, Dusseljee-Peute LW. Mobile health for older adult patients: using an aging barriers framework to classify usability problems. Int J Med Inform 2019;124:68-77. [doi: 10.1016/j.ijmedinf.2019.01.006] [Medline: 30784429]
45. Ramdowar H, Khedo KK, Chooramun N. A comprehensive review of mobile user interfaces in mHealth applications for elderly and the related ageing barriers. Univ Access Inf Soc 2023;23(4):1613-1629. [doi: 10.1007/s10209-023-01011-z]
46. Cho J, Park D, Lee HE. Cognitive factors of using health apps: systematic analysis of relationships among health consciousness, health information orientation, eHealth literacy, and health app use efficacy. J Med Internet Res 2014;16(5):e125 [FREE Full text] [doi: 10.2196/jmir.3283] [Medline: 24824062]
47. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012;14(1):e19 [FREE Full text] [doi: 10.2196/jmir.1619] [Medline: 22357448]
48. Lwin MO, Panchapakesan C, Sheldenkar A, Calvert GA, Lim LK, Lu J. Determinants of eHealth literacy among adults in China. J Health Commun 2020;25(5):385-393. [doi: 10.1080/10810730.2020.1776422] [Medline: 32552607]
49. Zakar R, Iqbal S, Zakar MZ, Fischer F. COVID-19 and health information seeking behavior: digital health literacy survey amongst university students in Pakistan. Int J Environ Res Public Health 2021;18(8):4009 [FREE Full text] [doi: 10.3390/ijerph18084009] [Medline: 33920404]
50. Kim S, Oh J. The relationship between e-Health literacy and health-promoting behaviors in nursing students: a multiple mediation model. Int J Environ Res Public Health 2021;18(11):5804 [FREE Full text] [doi: 10.3390/ijerph18115804] [Medline: 34071469]
51. Cho H, Han K, Park BK. Associations of eHealth literacy with health-promoting behaviours among hospital nurses: a descriptive cross-sectional study. J Adv Nurs 2018;74(7):1618-1627. [doi: 10.1111/jan.13575] [Medline: 29575085]
52. Zhang L, Qin Y, Li P. Media complementarity and health information acquisition: a cross-sectional analysis of the 2017 HINTS-China survey. J Health Commun 2020;25(4):291-300. [doi: 10.1080/10810730.2020.1746868] [Medline: 32255740]
53. Moungui HC, Nana-Djeunga HC, Anyiang CF, Cano M, Ruiz Postigo JA, Carrion C. Dissemination strategies for mHealth apps: systematic review. JMIR Mhealth Uhealth 2024;12:e50293 [FREE Full text] [doi: 10.2196/50293] [Medline: 38180796]
54. Ceci L. Leading health and fitness apps worldwide in 2022, by downloads. Statista. 2024 Feb 5. URL: https://www.statista.com/statistics/1284875/global-top-health-and-fitness-apps/ [accessed 2025-01-07]
55. El-Gayar O, Elnoshokaty A. Factors and design features influencing the continued use of wearable devices. J Healthc Inform Res 2023;7(3):359-385. [doi: 10.1007/s41666-023-00135-4] [Medline: 37637719]
56. Malik VS, Hu FB. The role of sugar-sweetened beverages in the global epidemics of obesity and chronic diseases. Nat Rev Endocrinol 2022;18(4):205-218 [FREE Full text] [doi: 10.1038/s41574-021-00627-6] [Medline: 35064240]
57. Anthes E. Mental health: there's an app for that. Nature 2016;532(7597):20-23. [doi: 10.1038/532020a] [Medline: 27078548]
58. Maresova P, Javanmardi E, Barakovic S, Barakovic Husic J, Tomsone S, Krejcar O, et al. Consequences of chronic diseases and other limitations associated with old age – a scoping review. BMC Public Health 2019;19(1):1431 [FREE Full text] [doi: 10.1186/s12889-019-7762-5] [Medline: 31675997]
59. Britt RK, Collins WB, Wilson K, Linnemeier G, Englebert AM. eHealth literacy and health behaviors affecting modern college students: a pilot study of issues identified by the American College Health Association. J Med Internet Res 2017;19(12):e392 [FREE Full text] [doi: 10.2196/jmir.3100] [Medline: 29258979]

XSL•FO
RenderX

60. Guo Z, Zhao SZ, Guo N, Wu Y, Weng X, Wong JY, et al. Socioeconomic disparities in eHealth literacy and preventive behaviors during the COVID-19 pandemic in Hong Kong: cross-sectional study. J Med Internet Res 2021;23(4):e24577 [FREE Full text] [doi: 10.2196/24577] [Medline: 33784240]
61. Zhang L, Jung EH. WeChatting for health: an examination of the relationship between motivations and active engagement. Health Commun 2019;34(14):1764-1774. [doi: 10.1080/10410236.2018.1536942] [Medline: 30358416]
62. Duncan M, Murawski B, Short CE, Rebar AL, Schoeppe S, Alley S, et al. Activity trackers implement different behavior change techniques for activity, sleep, and sedentary behaviors. Interact J Med Res 2017;6(2):e13 [FREE Full text] [doi: 10.2196/ijmr.6685] [Medline: 28807889]
63. Hong H. An extension of the extended parallel process model (EPPM) in television health news: the influence of health consciousness on individual message processing and acceptance. Health Commun 2011;26(4):343-353. [doi: 10.1080/10410236.2010.551580] [Medline: 21416420]
64. Pu B, Zhang L, Tang Z, Qiu Y. The relationship between health consciousness and home-based exercise in China during the COVID-19 pandemic. Int J Environ Res Public Health 2020;17(16):5693 [FREE Full text] [doi: 10.3390/ijerph17165693] [Medline: 32781751]
65. Himmerich H, Mirzaei K. Body image, nutrition, and mental health. Nutrients 2024;16(8):1106 [FREE Full text] [doi: 10.3390/nu16081106] [Medline: 38674797]
66. Chao H, Lao I, Hao L, Lin C. Association of body image and health beliefs with health behaviors in patients with diabetes: a cross-sectional study. Diabetes Educ 2012;38(5):705-714. [doi: 10.1177/0145721712452796] [Medline: 22814357]
67. Eikey EV. Unintended users, uses, and consequences of mobile weight loss apps: using eating disorders as a case study. In: Sezgin E, Yildirim S, Özkan-Yildirim S, Sumuer E, editors. Current and Emerging mHealth Technologies: Adoption, Implementation, and Use. Cham, Switzerland: Springer International Publishing; 2018:119-133.
68. Eikey EV. Effects of diet and fitness apps on eating disorder behaviours: qualitative study. BJPsych open 2021;7(5):e176. [doi: 10.1192/bjo.2021.1011]
69. Eikey EV, Reddy MC, Booth KM, Kvasny L, Blair JL, Li V, et al. Desire to be underweight: exploratory study on a weight loss app community and user perceptions of the impact on disordered eating behaviors. JMIR Mhealth Uhealth 2017;5(10):e150 [FREE Full text] [doi: 10.2196/mhealth.6683] [Medline: 29025694]
70. Lin J, Dutta MJ. A replication of channel complementarity theory among internet users in India. Health Commun 2017;32(4):483-492. [doi: 10.1080/10410236.2016.1140268] [Medline: 27301884]
71. Jiang X, Wang L, Leng Y, Xie R, Li C, Nie Z, et al. The level of electronic health literacy among older adults: a systematic review and meta-analysis. Arch Public Health 2024;82(1):204 [FREE Full text] [doi: 10.1186/s13690-024-01428-9] [Medline: 39511667]
72. Bauman AE, Reis RS, Sallis JF, Wells JC, Loos RJF, Martin BW, Lancet Physical Activity Series Working Group. Correlates of physical activity: why are some people physically active and others not? Lancet 2012;380(9838):258-271. [doi: 10.1016/S0140-6736(12)60735-1] [Medline: 22818938]
73. Nelson ME, Rejeski WJ, Blair SN, Duncan PW, Judge JO, King AC, et al. Physical activity and public health in older adults: recommendation from the American College of Sports Medicine and the American Heart Association. Med Sci Sports Exerc 2007;39(8):1435-1445. [doi: 10.1249/mss.0b013e3180616aa2] [Medline: 17762378]
74. Jefferis BJ, Sartini C, Lee I, Choi M, Amuzu A, Gutierrez C, et al. Adherence to physical activity guidelines in older adults, using objectively measured physical activity in a population-based study. BMC Public Health 2014;14:382 [FREE Full text] [doi: 10.1186/1471-2458-14-382] [Medline: 24745369]
75. Cunningham C, O'Sullivan R, Caserotti P, Tully MA. Consequences of physical inactivity in older adults: a systematic review of reviews and meta-analyses. Scand J Med Sci Sports 2020;30(5):816-827. [doi: 10.1111/sms.13616] [Medline: 32020713]
76. Sohaib Aslam A, van Luenen S, Aslam S, van Bodegom D, Chavannes NH. A systematic review on the use of mHealth to increase physical activity in older people. Clin eHealth 2020;3:31-39. [doi: 10.1016/j.ceh.2020.04.002]
77. Peat CM, Peyerl NL, Muehlenkamp JJ. Body image and eating disorders in older adults: a review. J Gen Psychol 2008;135(4):343-358. [doi: 10.3200/GENP.135.4.343-358] [Medline: 18959226]
78. Paxton SJ, Phythian K. Body image, self-esteem, and health status in middle and later adulthood. Aust Psycho 2007;34(2):116-121. [doi: 10.1080/00050069908257439]
79. Chatelan A, Carrard I. Diet quality in middle-aged and older women with and without body weight dissatisfaction: results from a population-based national nutrition survey in Switzerland. J Nutr Sci 2021;10:e38 [FREE Full text] [doi: 10.1017/jns.2021.32] [Medline: 34367623]
80. Glanz K. Health behavior and risk factors. In: Quah SR, editor. International Encyclopedia of Public Health. 3rd ed. New York: Academic Press; 2025:175-181.
81. Gulec H, Smahel D, Huang Y. Open Science Framework. 2024 Dec 23. URL: https://osf.io/ubazy/?view_only=ab5bbaad18824c248d6227fb02b1653340 [accessed 2026-01-08]

XSL•FO
RenderX

Abbreviations

CFA: confirmatory factor analysis
EFA: exploratory factor analysis
ESEM: exploratory structural equation modeling
mHealth: mobile health
NUTS3: Nomenclature of Territorial Units for Statistics
OR: odds ratio
SNS: social networking site
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 22.Jan.2025; peer-reviewed by A Eisingerich; comments to author 03.Sep.2025; revised version received 17.Nov.2025; accepted 30.Dec.2025; published 30.Jan.2026.

Please cite as:
Gulec H, Smahel D, Huang Y
Patterns and Characteristics of Mobile App Use to Promote Wellness and Manage Illness: Cross-Sectional Study
J Med Internet Res 2026;28:e71363
URL: https://www.jmir.org/2026/1/e71363
doi:10.2196/71363
PMID:

©Hayriye Gulec, David Smahel, Yi Huang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effects of an 8-Week App-Based Mindfulness Intervention on Mental Health in Working Women: Randomized Controlled Trial

Riko Uwagawa[1*], MA; Koichiro Adachi[1*], MA; Mariko Shimoda[1*], MA; Ryu Takizawa[1,2*], MD, PhD

[1]Department of Clinical Psychology, Graduate School of Education, The University of Tokyo, Tokyo, Japan

[2]MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[*]all authors contributed equally

Corresponding Author:
Ryu Takizawa, MD, PhD
Department of Clinical Psychology
Graduate School of Education
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku
Tokyo, 1130033
Japan
Phone: 81 3 5841 1397
Email: takizawar-tky@umin.ac.jp

Abstract

Background: Although working women experience increased work-related stress, preventive interventions to reduce its negative effects on their mental health are insufficient.

Objective: This study evaluated the effectiveness of an 8-week mindfulness-based self-help intervention via a smartphone app across 4 domains (general psychological, work-related, family-related, and work-to-conflict) among working women.

Methods: This study recruited women workers via various media sources, such as crowdsourcing sites and social networking services. Participants were randomly assigned to the intervention (n=106) or waitlist control groups (n=107). Participants in the intervention group practiced guided mindfulness meditation every day at their convenience via an app on their cell phones for 8 weeks. The app provides an 8-week program with 4 meditation contents per 2 weeks. Participants in the waitlist control group lived as usual for 8 weeks. We conducted web-based questionnaires to assess participants' general psychological (life satisfaction, perceived stress, depressive and anxiety symptoms, trait anger, and mindfulness), work-related (work performance, job satisfaction, quantitative job overload, and job control), family-related (family satisfaction and partner satisfaction), and work-to-family conflict indicators.

Results: An analysis of covariance, controlled for preintervention scores, revealed that the intervention significantly increased life satisfaction (b=1.47, β=0.11; P=.005) and decreased perceived stress (b=–2.00, β=–0.17; P=.01), depressive and anxiety symptoms (b=–1.24, β=–0.15; P=.02), and trait anger (reaction; b=–0.59, β=–0.11; P=.04). The intervention group demonstrated significantly increased life satisfaction (t_{93}=–3.36; P=.001) and decreased depressive and anxiety symptoms (t_{93}=2.35; P=.02).

Conclusions: The app was effective in reducing perceived stress, depressive and anxiety symptoms, and trait anger (reaction), and in improving life satisfaction among working women. However, to improve work- and family-related indicators, higher-intensity interventions may be required, such as modifying the intervention content or extending its duration.

Trial Registration: University Hospital Medical Information Network Clinical Trials Registry (UMIN-CTR) UMIN000051796; https://center6.umin.ac.jp/cgi-open-bin/ctr_e/ctr_view.cgi?recptno=R000059110

(J Med Internet Res 2026;28:e62814) doi:10.2196/62814

KEYWORDS

mindfulness; mobile apps; randomized controlled trial; women's health; mental health; subjective well-being; health promotion; mHealth; application; applications; work-related stress; stress; intervention; interventions; women; mobile phone

XSL•FO
RenderX

Introduction

The impact of work-related stress on workers' mental health has been recently investigated, and its significant social impact has become an issue [1]. According to the World Health Organization, work-related stress refers to "the response people may have when presented with work demands and pressures that are not matched to their knowledge and abilities and which challenge their ability to cope" [2].

Working women experience increased work-related stress compared with working men. The American Psychological Association found that women consistently exhibited higher levels of stress than men and had additional difficulty in coping [3]. Furthermore, women are more likely to develop stress-related symptoms owing to neurobiological differences, a sense of burden from the dual roles of balancing work and family, and exposure to job insecurity [4-6]. Work-family conflict of working women has a negative impact on their stress and on their physical and mental health, and a framework regarding the relationship between these is presented [7]. Work-related factors may affect women and men differently, with women possibly being further affected owing to their work and family roles. With the global aim of gender parity in the labor market [8], the number of women in the working population is expected to increase. Therefore, preventive interventions to reduce the negative effects of work-related stress on women's mental health are required. However, such support is insufficient [9,10].

Traditionally, psychiatry has focused on the treatment of mental disorders rather than prevention. However, mental health is more than the absence of mental illness [10,11]. Therefore, interventions that focus on preventing mental health problems among women workers before they worsen and improving positive aspects, such as life satisfaction, could have positive effects on women's well-being and their work, family, and society as a whole.

Mindfulness meditation is an effective intervention strategy for improving mental health and well-being. Mindfulness is the awareness that emerges from deliberate, nonjudgmental attention to experiences as they unfold moment-by-moment [12]. As mindfulness-based interventions reduce symptoms of depression, anxiety, and perceived stress and improve sleep quality and well-being [13-15], they are attracting attention as a preventive intervention strategy.

Additionally, the effectiveness of mindfulness meditation provided by smartphone apps has been recently highlighted. According to the International Telecommunication Union, there are over 8.89 billion mobile subscriptions worldwide [16]. Therefore, mobile technology can be used to provide preventive health care interventions to numerous people.

A traditional mindfulness-based program is high-intensity (8 weekly sessions of 2.5 hours per session and 30-40 minutes of practice per day) and time-constrained, which creates a participation barrier for nonclinical working women. The mobile-based mindfulness intervention is an app-based, voice-guided meditation practice that allows users to practice at their own convenience, which offers the advantages of high convenience and low cost [17-19]. Furthermore, online mindfulness interventions are effective in improving depression, anxiety, stress, rumination, and well-being [20,21].

However, no studies have examined the effects of mobile-based mindfulness meditation on working women from work, family, and work–family conflict aspects, as well as general measures. To our knowledge, only two studies have examined the effectiveness of mobile-based mindfulness meditation among working women. Santos et al [10] found that an app-based mindfulness and positive psychology intervention effectively reduced perceived stress and anxiety symptoms in working women. Coelhoso et al [11] revealed that a well-being mobile app designed to handle psychological stress based on relaxation training, breathing techniques, meditation (mindfulness, loving meditation, such as mindfulness, loving, kindness, and empathetic joy), and positive psychology principles improved working women's work-related well-being and reduced their work-related and overall stress.

Conversely, no study has examined working women's well-being from the 4 aspects of general psychological, work-related, family-related, and work-family conflict indicators. Examining their effects is essential for the future applications of mindfulness meditation, as it will help us comprehensively understand how mindfulness meditation works for women workers.

Therefore, this study aimed to evaluate the effectiveness of an 8-week mindfulness meditation intervention via a smartphone app among women workers through a randomized controlled trial (RCT). Effectiveness was examined via 4 indicators: general psychological, work-related, family-related, and work-to-conflict measures. Furthermore, we examined the measures that would effectively influence. We hypothesized that participants in the intervention group (self-care mindfulness meditation via the smartphone app) would have a higher level of general psychological (life satisfaction, perceived stress, depressive and anxiety symptoms, trait anger, and mindfulness), work-related (work performance, job satisfaction, quantitative job overload, and job control), family-related (family satisfaction and partner satisfaction), and work-to-family conflict indicators compared with those in the waitlist control group.

Methods

Participants

A power analysis was conducted to determine the sample size needed for this study (significance=.05; statistical power=.8; effect size=0.4), and a sample size of 100 participants per group, for a total of 200 participants, was needed. The effect size demonstrated in the meta-analysis of the effects of online mindfulness-based interventions on mental health was used as reference (depression Hedges g=0.34; stress Hedges g=0.44) [21].

This study recruited 397 women workers via various media sources, such as crowdsourcing sites and social networking services. Inclusion criteria included those who were (1) biologically female, (2) employed for at least 20 hours per week,

XSL•FO
RenderX

(3) owned an iPhone (for convenience of the app used), and (4) aged 18-64 years. Exclusion criteria included those who (1) received treatment for a mental disorder, (2) scored ≥13 on the 6-item Kessler Psychological Distress Scale (K6) Japanese version, (3) were on leave, and (4) were currently pregnant or likely to become pregnant within six months. Among the participants, 95 did not meet the inclusion and exclusion criteria. Hence, 302 women workers who met the criteria were asked to respond to the preintervention assessment, and 215 who completed the assessment were randomly assigned to the intervention (n=107) or waitlist control group (n=108). Randomization was computerized using a blocked randomization scheme (block size 10). A total of 8 working women dropped out. Of the 8 participants, 2 participants (intervention group, n=1; wait-list control group, n=1) declined to participate in this study, 2 participants in the intervention group opted out of the intervention, and 4 participants (intervention group, n=2; wait-list control group, n=2) could not be contacted. After 8 weeks, the participants were asked to respond to the postintervention assessment, and 196 women workers completed the assessment (intervention group, n=95; waitlist control group, n=101). Of 215 participants who completed the preintervention assessment, 4 who worked <19 hours per week on average in the preintervention assessment were excluded from analysis (intervention group, n=1; waitlist control group, n=3). Therefore, of the 215 participants who were randomized, data from 209 participants (intervention group, n=105; waitlist control group, n=104) were finally analyzed, excluding 2 participants who declined to participate in this study and 4 participants who worked <19 hours per week on average in the preintervention assessment.

Procedure

Overview

This study was designed as a parallel-design RCT. Randomization was computerized independently by research staff using a blocked randomization scheme (block size 10). Participants were expected to be randomized in a ratio of 1:1 to the intervention or waitlist control group. This study was an open-label RCT as it was not possible to blind the allocation.

This study was conducted from July 2023 to January 2024 via web forms. Participants in the intervention group installed the app for meditation after the preintervention assessment. Participants practiced guided mindfulness meditation via the app on their cell phones every day at their convenience for 8 weeks. After 8 weeks, the participants received the postintervention questionnaire via the app and email. The participants in the waitlist control group lived as usual for 8 weeks after the preintervention assessment. After 8 weeks, they also responded to the postintervention questionnaire via email.

This study was registered in the University Hospital Medical Information Network (UMIN) Clinical Trials Registry (UMIN000051796).

8-Week Mindfulness-Based Self-Help Intervention via the Smartphone App

Mindfulness meditation was conducted via the iOS app, with the content changed every 2 weeks (Table 1). The app displayed the day's meditation content and explanation on the home screen. After viewing this screen, the participants pressed the play button to hear the guided audio and practiced meditation. Figure 1 illustrates the display of the app. In addition, the psychoeducation pages on mindfulness and self-compassion were created and inserted on the app (Figure 2).

Table 1. Content of the 8-week self-help mindfulness-based meditation.

Week	Types of meditation	Duration (minutes)
1 and 2	Meditation of breath	7
3 and 4	Body scan	7
5 and 6	Meditation of breath, sound, and body	12
7 and 8	Loving-kindness meditation	12

XSL•FO
RenderX

Figure 1. Display of the smartphone app.



Figure 2. Display of the psychoeducation.

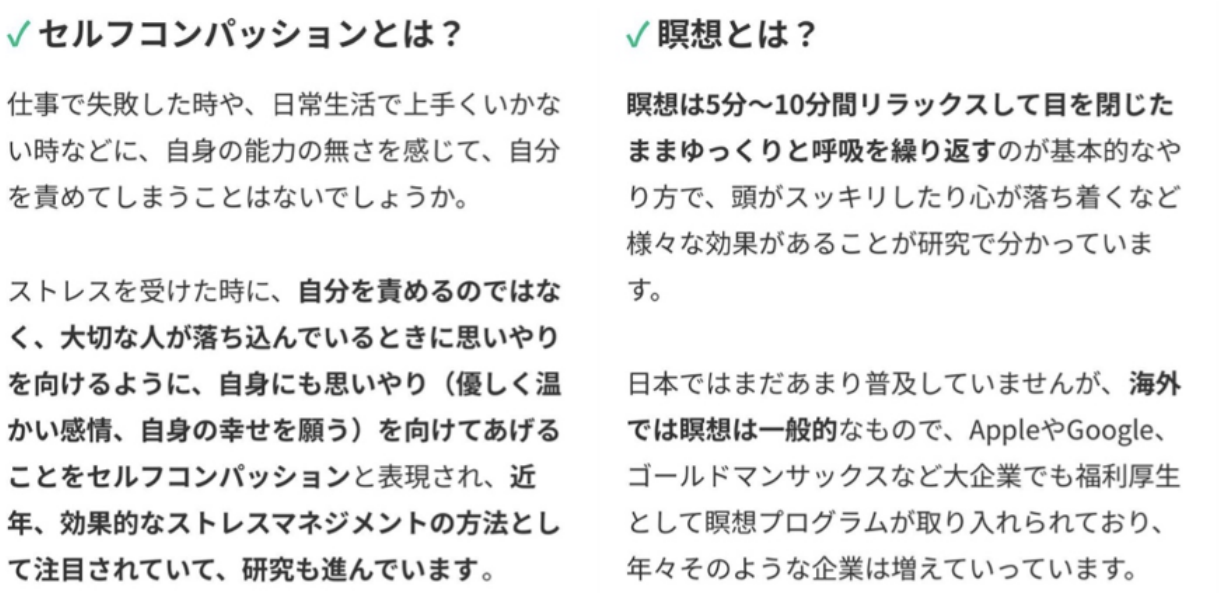

The content included “meditation of breath,” “meditation of breath, sound, and body,” and “body scan meditation,” based on previous studies [22]. As the “body scan” was partially included in “meditation of breath, sound, and body,” in this study, the latter was conducted after the former. Furthermore, as the effectiveness of interventions that incorporated elements of self-compassion was recently highlighted, “loving-kindness meditation” was ultimately added. As a daily 13-minute meditation was effective after 8 weeks [23], the intervention period was designed to be 8 weeks.

Measurements

General Psychological Domain: Well-Being

Well-being was assessed as life satisfaction using the 5-item Satisfaction with Life Scale. Participants evaluated their subjective life satisfaction on a 7-point Likert scale that ranged from 1 (strongly disagree) to 7 (strongly agree) [24,25]. This measurement was developed by Diener et al [24]. The development of the Japanese version used in this study and its validity and reliability were studied by Sumino [25]. Sample items included “In most ways, my life is close to my ideal.” The total score was a sum of all the individual item scores, and

XSL•FO
RenderX

higher scores indicated greater life satisfaction. In this study, Cronbach α was 0.85 and 0.81 for the pre- and postintervention assessments, respectively.

Mental Health Outcomes

Perceived Stress

The 10-item Perceived Stress Scale was used to assess perceived stress. Participants rated how unpredictable, uncontrollable, and overloaded they found their lives on a 5-point Likert scale that ranged from 0 (never) to 4 (very often) [26,27]. This measurement was developed by Cohen et al [26]. The development of the Japanese version used in this study and its validity and reliability were studied by Sumi [27]. Sample items included "How often have you been upset because of something that happened unexpectedly?" The total score was a sum of the individual item scores, and higher scores indicated greater perceived stress. Cronbach α was 0.69 and 0.79 for pre- and postintervention, respectively.

Depressive and Anxiety Symptoms

K6 was used to assess depression and anxiety symptoms. Participants described how often they experienced depressive symptoms in the past 30 days on a 5-point Likert scale that ranged from 0 (none of the time) to 4 (all of the time) [28-30]. This measurement was developed by Kessler et al [28], and the Japanese version of it used in this study was developed by Furukawa et al [29]. The validity and reliability were studied by Furukawa et al [29] and by Sakurai et al [30]. Sample items included "How often did you feel nervous?" and "How often did you feel restless or fidgety?" The total score was a sum of all the individual item scores, and higher scores indicated a greater severity of depression and anxiety. Cronbach α was 0.83 and 0.80 for pre- and postintervention, respectively.

Trait Anger

"Trait anger (T-Ang; 10-item)," a subscale of the 57-item State-Trait Anger Expression Inventory 2 (STAXI-2), was used to assess the traits of anger reaction [31-33]. This measurement was developed by Spielberger [31]. The development of the Japanese version used in this study and its reliability were studied by Mine and Ohki [32] and by Mine and Sato [33]. Participants evaluated their perceptions of anger proneness on a 4-point Likert scale that ranged from 1 (strongly disagree) to 4 (strongly agree). Sample items included "I am quick-tempered." T-Ang included two subfactors: T-Ang/Temperament (T-Ang/T; trait of feeling anger with or without stimulus) and T-Ang/Reaction (T-Ang/R; frequency of experiencing feelings of anger in situations involving irritation or negative evaluation). The total score within each subfactor and all items was calculated by summing the item scores. Higher scores indicated greater trait anger. Cronbach α for preintervention was T-Ang Cronbach α=0.84, T-Ang/T Cronbach α=0.79, and T-Ang/R Cronbach α=0.77, and for postintervention was T-Ang Cronbach α=0.83, T-Ang/T Cronbach α=0.84, and T-Ang/R Cronbach α=0.76.

Mindfulness

The 15-item Mindful Attention Awareness Scale was used to assess dispositional mindfulness [34,35]. This measurement was developed by Brown and Ryan [34]. The development of the Japanese version used in this study and its validity and reliability were studied by Fujino et al [35]. Participants rated the degree to which they functioned without awareness of the present experience in daily life on a 6-point scale that ranged from 1 (almost never) to 6 (almost always). Sample items included "I could be experiencing some emotion and not be conscious of it until sometime later." All items were reversed as they assessed the lack of mindful attention and awareness. The total score was a sum of all the reversed-item scores, and higher scores indicated greater mindful attention and awareness. Cronbach α was 0.81 and 0.87 for pre- and postintervention, respectively.

Work-Related Domain

Work Performance

The World Health Organization Health and Work Performance Questionnaire Short Form was used to assess work performance. The questions included: "On a scale of 0-10, where 0 is the worst job performance anyone could have at your job, and 10 is the performance of a top worker, how would you rate the usual performance of most workers in a job similar to yours?" (possible performance) and "Using the same 0-10 scale, how would you rate your overall job performance on the days you worked during the past four weeks?"(actual performance) [36-38]. This measurement was developed by Kessler et al [36]. The development of the Japanese version used in this study and its validity and reliability were studied by Kawakami et al [38]. Participants evaluated the workplace costs of health problems regarding self-reported sickness leaves and reduced job performance (presenteeism). Presenteeism was assessed by "absolute" and "relative presenteeism." "Absolute presenteeism" was calculated by multiplying the score of actual performance by 10. Higher scores indicated greater performance. "Relative presenteeism" was calculated by the ratio of actual performance to possible performance (restricted to the range of 0.25–2.0, where values <0.25 and >2.0 were converted to 0.25 and 2.0, respectively). Higher scores indicated greater performance.

Job Satisfaction

Job satisfaction was assessed via a single item from the Brief Job Stress Questionnaire (BJSQ) [39]. The development of this measurement used in this study and its validity and reliability were studied by Inoue et al [39]. Participants rated the degree to which they agreed with the item, "I am satisfied with my job," on a 4-point Likert scale that ranged from 1 (satisfied) to 4 (dissatisfied). The item was reversed as it assessed the high level of job satisfaction. Higher scores indicated greater job satisfaction.

Quantitative Job Overload

"Quantitative job overload (3-item)," a subscale of the BJSQ, was used to assess job overload [39]. Participants rated the degree of their job overload on a 4-point Likert scale that ranged from 1 (agree) to 4 (disagree). Sample items included "I have a lot of work to do." All items were reversed as they assessed the high level of job overload. The total score was a sum of all the reversed-item scores, and higher scores indicated a greater

XSL•FO
RenderX

job overload. Cronbach α was 0.64 and 0.56 for pre- and postintervention, respectively.

Job Control

"Job control (3-item)," a subscale of the BJSQ, was used to assess job control [39]. Participants rated the degree of their job control on a 4-point Likert scale that ranged from 1 (agree) to 4 (disagree). Sample items included "I can work at my own pace." All items were reversed as they assessed the high level of job control. The total score was a sum of all the reversed-item scores, and higher scores indicated a greater sense of job control. Cronbach α was 0.60 and 0.64 for pre- and postintervention, respectively.

Family-Related Domain

Family Satisfaction

Family satisfaction was assessed via a single item from the BJSQ [39]. Participants rated the degree to which they agreed with the item, "I am satisfied with my family life," on a 4-point Likert scale that ranged from 1 (satisfied) to 4 (dissatisfied). The item was reversed as it assessed the high level of family satisfaction. Higher scores indicated greater family satisfaction.

Partner Satisfaction

Partner satisfaction was assessed via a single item: "Using the 10-point scale, how would you rate your current level of satisfaction with your relationship with your partner?" Only participants who lived with their partners were asked to respond. Participants rated the degree of their satisfaction with their partner on a 10-point Likert scale that ranged from 1 (dissatisfied) to 10 (satisfied). Higher scores indicated greater satisfaction.

Work-to-Family Conflict Domain

The 22-item Survey Work-Home Interaction-Nijmegen was used to assess the 4 subscales that reflected the underlying dimensions of work–family spillover: (1) work-family negative spillover (WFNS, 8 items; eg, "You do not have the energy to engage in leisure activities with your spouse/family/friends because of your job."), (2) family-work negative spillover (FWNS, 4 items; eg, "You do not feel like working because of problems with your spouse/family/friends."), (3) work-family positive spillover (WFPS, 5 items; eg, "You fulfill your domestic obligations better because of the things you have learned on your job."), (4) family-work positive spillover (FWPS, 5 items; eg, "You have greater self-confidence at work because you have your home life well organized") [40,41]. This measurement was developed by Geurts et al [40] in 2005. The development of the Japanese version used in this study and its validity and reliability were studied by Shimada et al [41]. Responses were rated on a 4-point Likert scale that ranged from 0 (never) to 3 (always). The total score of each subscale was calculated as a sum of all the individual item scores. Higher scores on the positive (WFPS and FWPS) and negative spillover subscales (WFNS and FWNS) indicated greater positive and negative impacts, respectively. For preintervention, the Cronbach α were WFNS Cronbach α=0.88, FWNS Cronbach α=0.79, WFPS Cronbach α=0.73, and FWPS Cronbach α=0.78, and for postintervention, it was WFNS Cronbach α=0.89, FWNS Cronbach α=0.76, WFPS Cronbach α=0.79, and FWPS Cronbach α=0.83.

Statistical Analysis

We conducted Chi-squared, *t* tests, and the Fisher exact test in order to examine whether there are differences in demographic variables and psychological indices between the intervention and control groups. Subsequently, we conducted 2-tailed *t* tests to examine whether there were differences in demographic variables and psychological indices of participants in the intervention and waitlist control groups, respectively.

For the intervention effects, we conducted an analysis of covariance (ANCOVA; independent variables: intervention group=1 and waitlist control group=0) that used the least squares method as an estimation method, controlled for preintervention scores. We conducted an ANCOVA that used the least squares estimation method, controlled for preintervention scores, age, employment status (regular employment: employed full time with no fixed term of employment; nonregular employment: not regular employment), psychiatric history, education, and marital status. In this study, the participants were randomly assigned to the intervention and control groups. However, because of the possibility that the intervention effect might not be properly detected due to group differences in preintervention scores and demographic data, we controlled for them. Additionally, paired *t* tests were conducted to determine any differences in the pre- and postintervention assessments within each group. An intention-to-treatment analysis was used. R (version 4.3.2; R Foundation for Statistical Computing) was used for statistical analysis.

Ethical Considerations

This study was approved by the Life Science Research Ethics and Safety Committee, the University of Tokyo (23-144, 23-227, and 24-020).

Results

Baseline

Figure 3 illustrates the CONSORT (Consolidated Standards for Reporting Trials) flow diagram (the CONSORT checklist is provided in Multimedia Appendix 1).

Table 2 shows the participants' demographic characteristics. Chi-squared and *t* tests revealed no differences in demographic variables and psychological indices between the intervention and waitlist control groups ($P>.05$).

XSL•FO
RenderX

Figure 3. CONSORT (Consolidated Standards for Reporting Trials) flowchart for participants.

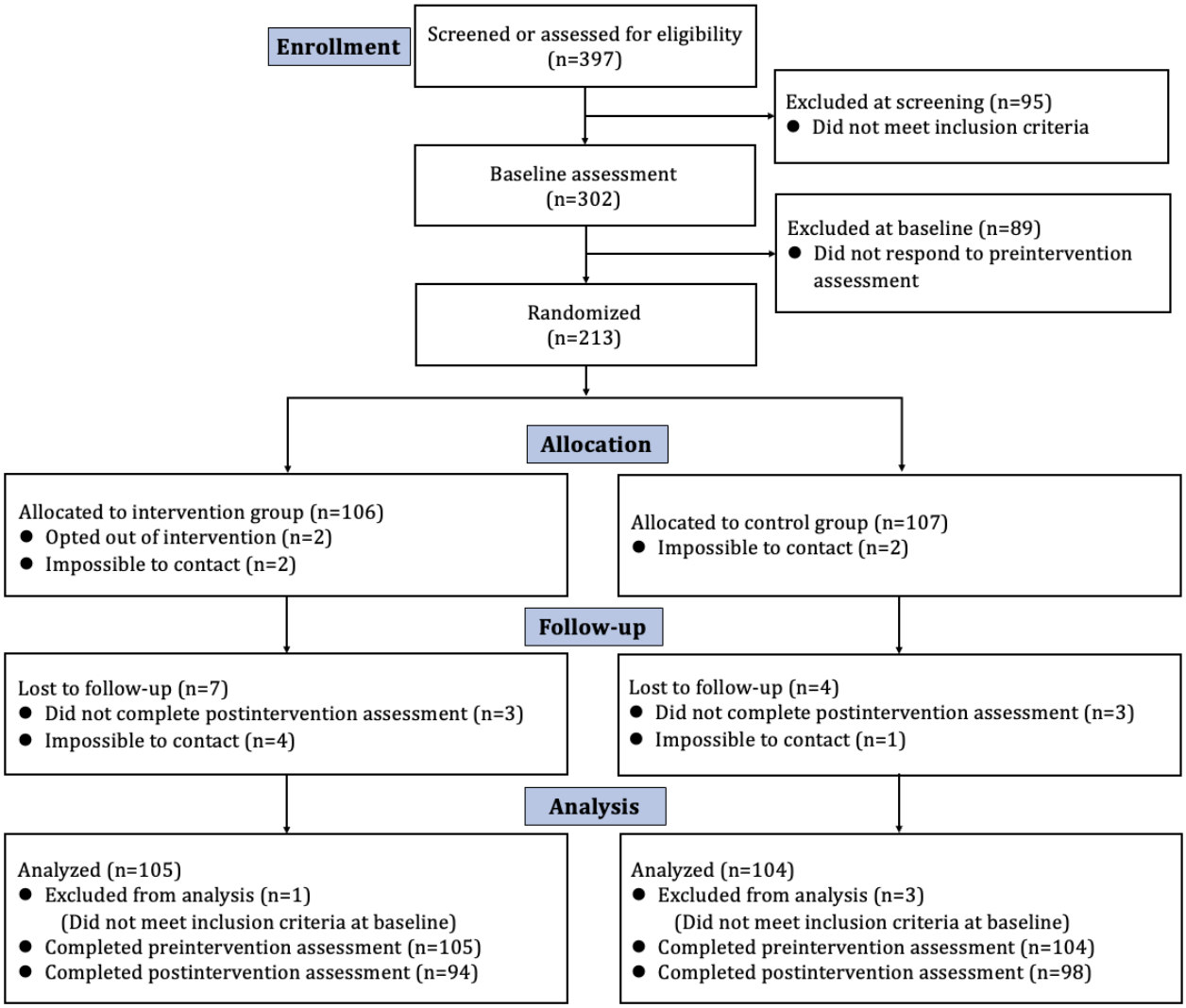

XSL•FO
RenderX

Table 2. Participants' demographic information.

Participant characteristics	Intervention group (n=105)	Waitlist control group (n=104)	Difference statistic: t test (*df*) or chi-square (*df*)	*P* value
Age (years), mean (SD)	36.81 (10.82)	36.81 (10.70)	0.0 (207)[a]	.99
Education level, n (%)			Fisher exact test	.09
Less than a bachelor's degree	46 (43.6)	31 (29.8)		
Bachelor's degree	48 (45.7)	60 (57.7)		
Master's degree	11 (10.6)	11 (10.6)		
Doctoral degree	0 (0)	2 (1.9)		
Marital status, n (%)			Fisher exact test	.44
Married	46 (46.8)	50 (48.1)		
Single	51 (44.7)	42 (40.4)		
Divorced	6 (6.4)	11 (10.6)		
Widowed	2 (2.1)	1 (1)		
Employment status[b], n (%)			2.1 (1)[c]	.14
Regular employment	66 (62.9)	54 (51.9)		
Nonregular employment	39 (37.1)	50 (48.1)		
Psychiatric history[d], n (%)			2.6 (1)[c]	.11
Yes	7 (6.7)	15 (14.4)		
No	98 (93.3)	89 (85.6)		
Living with a partner, n (%)			0.0 (1)[c]	.83
Yes	52 (49.5)	54 (51.9)		
No	53 (50.5)	50 (48.1)		
Youngest child age in years, n (%)			2.1 (3)[c]	.56
0-2	8 (7.6)	5 (4.8)		
3-18	30 (28.6)	30 (28.8)		
19+	4 (3.8)	8 (7.7)		
None	63 (6)	61 (58.7)		

[a]*t* test (*df*).

[b]Employment status indicates whether the individual is a regular employee.

[c]Chi-square (*df*).

[d]Psychiatric history indicates whether the individual has a history of visiting a psychosomatic medicine or psychiatric clinic.

Comparing Completers and Dropouts Within Each Group

In the intervention group, no statistically significant differences were observed in demographic information and psychological measurement scores between the dropouts and participants who completed the postintervention assessment (*P*>.05). In the waitlist control group, there were differences in age (t_{102}=2.66; *P*=.009), T-Ang (t_{102}=–2.22; Cohen *d*=0.93; *P*=.03), T-Ang/R (t_{102}=–2.26; Cohen *d*=0.95; *P*=.03), and job control (t_{102}=–2.57; Cohen *d*=1.08; *P*=.01). Dropouts were significantly younger (mean 25.83, SD 4.62), more angry (T-Ang: mean 24.00, SD 9.06; T-Ang/R: mean 11.00, SD 4.65), and perceived an additional sense of job control (mean 9.17, SD 1.60) compared with the retained participants (age mean 37.48, SD 10.61; T-Ang: mean 18.76, SD 5.39; T-Ang/R: mean 8.34, SD 2.68; job control: mean 6.56, SD 2.45).

Practice Frequency

Participants in the intervention group used the app for a mean of 42.32 days (75.57%, SD 15.63) in 8 weeks.

Outcomes

Group Effects

Table 3 presents the scores of the pre- and postintervention assessments. Table 4 presents the results of ANCOVA. The ANCOVA, controlled for preintervention scores, revealed significant group effects on life satisfaction (*b*=1.47, β=0.11; *P*=.005), perceived stress (*b*=–2.00, β=–0.17; *P*=.01), depressive and anxiety symptoms (*b*=–1.24, β=–0.15; *P*=.02), and T-Ang/R

XSL•FO
RenderX

(*b*=–0.59, β=–0.11; *P*=.04). The ANCOVA, controlled for pre-intervention scores and demographic data (age, employment status, psychiatric history, education, marital status), revealed significant group effects on life satisfaction (*b*=1.35, β=0.10; *P*=.02), perceived stress (*b*=–1.91, β=–0.16; *P*=.02), depressive and anxiety symptoms (*b*=–1.13, β=–0.13; *P*=.03), and T-Ang/R (*b*=–0.71, β=–0.13; *P*=.02).

Table 3. Scores of the pre- and postintervention assessments.

	Intervention group				Waitlist control group				*t* test	
	Mean (SD)		*P* value	Cohen *d* (95% Cl)	Mean (SD)		*P* value	Cohen *d* (95% Cl)	*P* value	Cohen *d* (95% Cl)
	Pre	Post			Pre	Post				
General psychological domain										
Life satisfaction	17.70 (6.68)	18.80 (6.76)	.001	0.22 (0.09 to 0.35)	18.62 (6.89)	18.58 (6.32)	.38	0.05 (–0.06 to 0.15)	.82	0.03 (–0.25 to 0.31)
Perceived stress	18.71 (5.14)	18.65 (5.65)	.67	0.05 (–0.19 to 0.29)	17.62 (5.95)	19.94 (6.38)	<.001	0.40 (0.21 to 0.60)	.14	0.21 (–0.07 to 0.50)
Depressive and anxiety symptoms	6.16 (4.06)	5.37 (3.61)	.02	0.27 (0.04 to 0.50)	6.20 (4.89)	6.41 (4.71)	.33	0.08 (–0.08 to 0.25)	.09	0.25 (–0.04 to 0.53)
Trait anger	18.50 (5.16)	18.27 (5.31)	.17	0.10 (–0.04 to 0.23)	19.06 (5.73)	18.92 (5.58)	.70	0.03 (–0.12 to 0.18)	.41	0.12 (–0.17 to 0.40)
Trait anger (temperament)	7.09 (2.62)	6.94 (2.54)	.06	0.12 (–0.01 to 0.24)	7.49 (2.78)	7.30 (2.95)	.56	0.04 (–0.10 to 0.19)	.37	0.13 (–0.15 to 0.41)
Trait anger (reaction)	8.50 (2.65)	8.28 (2.71)	.24	0.10 (–0.07 to 0.27)	8.49 (2.86)	8.71 (2.74)	.07	0.14 (–0.01 to 0.29)	.27	0.16 (–0.12 to 0.44)
Mindfulness	43.36 (10.65)	44.05 (11.31)	.70	0.03 (–0.10 to 0.07)	43.74 (9.85)	44.33 (11.75)	.72	0.02 (–0.11 to 0.16)	.87	0.02 (–0.26 to 0.31)
Work										
Work performance										
Absolute presenteeism	61.43 (19.24)	62.02 (18.64)	.73	0.04 (–0.19 to 0.27)	61.06 (19.5)	61.84 (18.69)	.67	0.05 (–0.17 to 0.27)	.95	0.01 (–0.27 to 0.29)
Relative presenteeism	1.02 (0.32)	1.05 (0.27)	.50	0.09 (–0.17 to 0.34)	1.00 (0.33)	1.01 (0.32)	.67	0.05 (–0.19 to 0.30)	.35	0.14 (–0.15 to 0.42)
Job satisfaction	2.70 (0.72)	2.67 (0.79)	.90	0.01 (–0.21 to 0.24)	2.81 (0.87)	2.81 (0.83)	.53	0.05 (–0.10 to 0.20)	.25	0.17 (–0.12 to 0.45)
Quantitative job overload	8.10 (2.36)	8.11 (2.43)	.82	0.02 (–0.17 to 0.22)	7.68 (2.59)	7.85 (2.30)	.43	0.07 (–0.10 to 0.23)	.45	0.11 (–0.17 to 0.39)
Job control	8.36 (2.33)	8.54 (2.23)	.17	0.11 (–0.05 to 0.26)	8.29 (2.48)	8.40 (2.29)	.84	0.02 (–0.15 to 0.18)	.66	0.06 (–0.22 to 0.35)
Family										
Family satisfaction	3.00 (0.77)	3.06 (0.81)	.18	0.13 (–0.06 to 0.33)	2.99 (0.82)	2.96 (0.88)	.45	0.07 (–0.12 to 0.26)	.40	0.12 (–0.16 to 0.41)
Partner satisfaction	7.52 (2.14)	7.39 (2.25)	.82	0.02 (–0.15 to 0.19)	7.69 (2.05)	7.15 (2.43)	.02	0.25 (0.04 to 0.45)	.62	0.10 (–0.29 to 0.49)
Work-to-family conflict										
Work-family negative spillover	5.62 (4.92)	5.63 (4.48)	.84	0.02 (–0.14 to 0.17)	5.51 (5.51)	5.18 (5.23)	.20	0.08 (–0.04 to 0.20)	.53	0.09 (–0.19 to 0.37)
Family-work negative spillover	1.25 (1.71)	1.20 (1.72)	.64	0.05 (–0.16 to 0.26)	1.19 (1.66)	1.36 (1.85)	.33	0.09 (–0.09 to 0.28)	.55	0.09 (–0.20 to 0.37)
Work-family positive spillover	7.08 (3.04)	7.02 (3.05)	.92	0.01 (–0.18 to 0.20)	6.77 (3.30)	7.13 (3.72)	.34	0.08 (–0.09 to 0.25)	.82	0.03 (–0.25 to 0.32)
Family-work positive spillover	7.09 (3.51)	7.22 (3.55)	.66	0.04 (–0.13 to 0.21)	7.19 (3.94)	7.28 (4.06)	.98	0.00 (–0.17 to 0.16)	.92	0.01 (–0.27 to 0.30)

XSL•FO
RenderX

Table 4. Comparison between the control and the intervention groups.

	Controlling for prescores					Controlling for prescores and demographic data				
	b	β	SE	*t* test (*df*)	*P* value	*b*	β	SE	*t* test (*df*)	*P* value
General psychological domain										
Life satisfaction	1.47	0.11	0.52	2.82 (188)	.005	1.35	0.10	0.55	2.47 (181)	.02
Perceived stress	–2.00	–0.17	0.79	–2.55 (188)	.01	–1.91	–0.16	0.82	–2.34 (181)	.02
Depressive and anxiety symptoms	–1.24	–0.15	0.51	–2.43 (188)	.02	–1.13	–0.13	0.53	–2.14 (181)	.03
Trait anger	–0.66	–0.06	0.53	–1.26 (188)	.21	–0.88	–0.08	0.54	–1.64 (181)	.10
Trait anger (temperament)	–0.22	–0.04	0.25	–0.85 (188)	.40	–0.23	–0.04	0.26	–0.89 (181)	.37
Trait anger (reaction)	–0.59	–0.11	0.29	–2.03 (188)	.04	–0.71	–0.13	0.29	–2.41 (181)	.02
Mindfulness	–0.02	0.00	1.03	–0.01 (188)	.99	0.11	0.00	1.08	0.10 (181)	.92
Work										
Work performance										
Absolute presenteeism	0.05	0.00	2.49	0.02 (188)	.98	–0.33	–0.01	2.60	–0.13 (181)	.90
Relative presenteeism	0.03	0.06	0.04	0.80 (188)	.42	0.02	0.04	0.04	0.53 (181)	.60
Job satisfaction	–0.03	–0.02	0.10	–0.26 (188)	.79	–0.01	0.00	0.10	–0.07 (181)	.94
Quantitative job overload	–0.01	0.00	0.28	–0.05 (188)	.96	0.00	0.00	0.29	0.01 (181)	.99
Job control	0.24	0.05	0.24	0.99 (188)	.32	0.25	0.06	0.24	1.03 (181)	.30
Family										
Family satisfaction	0.14	0.08	0.10	1.37 (188)	.17	0.14	0.08	0.10	1.32 (181)	.19
Partner satisfaction	0.50	0.11	0.30	1.64 (188)	.10	0.38	0.08	0.32	1.20 (181)	.23
Work-to-family conflict										
Work-family negative spillover	0.37	0.04	0.44	0.84 (188)	.40	0.28	0.03	0.46	0.60 (181)	.55
Family-work negative spillover	–0.21	–0.06	0.22	–0.93 (188)	.35	–0.23	–0.06	0.23	–1.00 (181)	.32
Work-family positive spillover	–0.21	–0.03	0.40	–0.53 (188)	.60	–0.06	–0.01	0.41	–0.15 (181)	.88
Family-work positive spillover	0.07	0.01	0.42	0.16 (188)	.87	–0.04	0.00	0.43	–0.09 (181)	.93

Differences Between Pre- and Postintervention Assessment Within Each Group

Regarding the intervention group, the postintervention scores of life satisfaction were significantly higher ($mean_{pre}$ 17.70, SD_{pre} 6.68; $mean_{post}$ 18.80, SD_{post} 6.76; t_{93}=–3.36; Cohen *d*=0.22, 95% CI 0.09-0.35; *P*=.001) and those of depressive and anxiety symptoms were significantly lower ($mean_{pre}$ 6.16, SD_{pre} 4.06; $mean_{post}$ 5.37, SD_{post} 3.61; t_{93}=2.35; Cohen *d*=0.27, 95% CI 0.04-0.50; *P*=.02) than the preintervention scores.

Regarding the waitlist control group, the postintervention scores of perceived stress were significantly higher ($mean_{pre}$ 17.62, SD_{pre} 5.95; $mean_{post}$ 19.94, SD_{post} 6.38; t_{97}=–4.20; Cohen *d*=0.40, 95% CI 0.21-0.60; *P*<.001) and those of partner satisfaction were significantly lower than the preintervention scores ($mean_{pre}$ 7.69, SD_{pre} 2.05; $mean_{post}$ 7.15, SD_{post} 2.43; t_{50}=2.41; Cohen *d*=0.25, 95% CI 0.04-0.45; *P*=.02).

Discussion

Principal Findings

This study examined the effectiveness of a mindfulness meditation intervention via a smartphone app among healthy women workers. To our knowledge, this was the first study that examined the effects of the mindfulness meditation intervention via a smartphone app on 4 domains (psychological, work, family, and work-to-family conflict) among women workers. Women workers who received the intervention demonstrated higher postintervention scores on the general psychological indicators (life satisfaction, perceived stress, depressive and anxiety symptoms, and trait anger (reaction) than those in the waitlist control group, controlled for preintervention scores as well as age, employment status, psychiatric history, education, and marital status. However, the intervention was not effective in the other 3 domains (work, family, and work-to-family

XSL•FO
RenderX

conflict). In particular, life satisfaction and depression, and anxiety symptoms significantly improved in the intervention group.

Our results corroborated the findings of Santos et al [10] and Coelhoso et al [11] that app-based mindfulness interventions reduced perceived stress and anxiety symptoms and improved subjective well-being among working women. Additionally, we found that app-based mindfulness interventions were useful for reducing reactive anger in working women. Working women are more likely to experience stress owing to neurobiological differences and balancing work and family than men, which may impair their well-being [3-7]. Mindfulness interventions enhance acceptance and observation skills by halting in daily life, paying attention to what is happening "here and now," and observing and accepting things as they are [12,42,43]. Therefore, acceptance and observation skills enable working women to pause and look at things as they are without being overwhelmed by negative thoughts and feelings when they are burdened by work and family in their daily lives. This is likely to calm their anger, lower their subjective stress, and increase their sense of well-being.

Conversely, this study observed no improvements in work-related, family-related, and work-to-family conflict indicators after the intervention. Previous studies have reported that mindfulness interventions increase family satisfaction among elementary and secondary school teachers, partner satisfaction among participants in a romantic relationship, and work satisfaction and performance among workers, and decrease the work-to-family conflict among workers [44-50]. The inconsistency of our results with those of previous studies could be owing to differences in sex, intervention duration, and meditation time per session. Previous studies examining the effects of preventive online mindfulness interventions for nonclinical populations on perceived stress and mindfulness have shown substantial differences across studies regarding design, setting, participants' age, gender ratio, intervention characteristics, and outcome measures [51]. These factors have been suggested to potentially influence the magnitude of the observed effects. Moreover, this preventive intervention may have been too short to reduce burden in the 3 work- or family-related domains. Furthermore, the intervention content was aimed at general meditation ("mindfulness of breath," "body scan," "mindfulness of breath, sound, and body," and "loving-kindness meditation"), rather than work- or family-specific content, and was implemented in a specific order. Therefore, the 8-week low-intensity meditation intervention could have led to an improvement in the individual's general well-being; however, the effect on work- or family-related indicators may have occurred after a few months. Alternatively, higher-intensity interventions may be required, such as modifying the intervention's contents or extending its duration. Additionally, only 1 item was used to measure work performance and job, family, and partner satisfaction, whereas 3 items were used to measure job overload and control. Therefore, the number of questions may have been too small to detect significant differences.

Limitations and Future Directions

This study has some limitations: a lack of subgroup analysis, an intervention not designed specifically for the target or context, and problems with generalizability and variability in the intensity of the intervention due to the application and problems with the scales used. First, in this study, subgroup analyses were not conducted to examine the impact of the subjects' traits on intervention effects. The effects of our mindfulness intervention on general psychological, work-related, family-related, and work-to-family conflict indicators may differ based on other factors. Some participants may have benefited from work-related, family-related, or work-to-conflict indicators. Therefore, it is necessary to examine the factors that moderate the effect of mindfulness interventions.

Second, the mindfulness intervention used in this study was not designed as target- and context-specific. Previous studies have developed target- and context-specific mindfulness interventions, such as for the workplace and parenting. Therefore, future studies should be designed specifically for working women, with an aim to increase the effects on work- and family-related indicators.

Third, there are two limitations of using an app for the intervention: the quality of the intervention cannot be assessed, and generalizability is limited due to restrictions on participant conditions. Since a self-help app was used as the intervention in this study, it was not possible to assess how well participants were focused on meditation, which may have resulted in variability in the effectiveness of the intervention. In addition, the limitations of the app used for the intervention limited the participants in this study to iPhone users, which may have biased the sample and limited generalizability. Therefore, future studies should address compliance issues to address these limitations caused by the app without limiting participants to iPhone users.

Fourth, some of the scales used had few items. The small number of items might have prevented the detection of significant differences. Therefore, future research should increase the number of items used in the survey.

Conclusion

This study examined the effects of mindfulness interventions via a smartphone app on women workers' general psychological, work-related, family-related, and work-to-family conflict indicators through an RCT. Our results revealed that the intervention increased life satisfaction and reduced perceived stress, depressive and anxiety symptoms, and anger reactions.

Acknowledgments

We would like to thank the participants for their time and effort, and Mai Sugie, Kohki Kaji, Yukari Kimura, Takumu Kurosawa, and Kei Minoura for their contributions and technical assistance in data collection. RU, KA, and MS were supported in part by Japan Science and Technology Agency SPRING (JPMJSP2108).

XSL•FO
RenderX

Funding

This work was supported by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (JP16H05653, JP19K03278, 22H01091, and 22K18582 to RT), the Royal Society and the British Academy (AL150003 to RT), and the University of Tokyo Social Cooperation Program “Fulfillment through Work” (to RT). The funders had no role in data collection or analyses, the decision to publish, or preparation of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-EHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 480 KB - jmir_v28i1e62814_app1.pdf]

References

1. Hassard J, Teoh KRH, Visockaite G, Dewe P, Cox T. The cost of work-related stress to society: a systematic review. J Occup Health Psychol 2018;23(1):1-17. [doi: 10.1037/ocp0000069] [Medline: 28358567]
2. Occupational health: stress at the workplace. World Health Organization. 2020. URL: https://tinyurl.com/dkrj3f2c [accessed 2024-04-24]
3. Stress in America 2023. A nation recovering from collective trauma. American Psychological Association. 2023. URL: https://www.apa.org/news/press/releases/stress/2023/collective-trauma-recovery [accessed 2024-04-24]
4. Kawakami N, Haratani T, Kobayashi F, Ishizaki M, Hayashi T, Fujita O, et al. Occupational class and exposure to job stressors among employed men and women in Japan. J Epidemiol 2004;14(6):204-211 [FREE Full text] [doi: 10.2188/jea.14.204] [Medline: 15617394]
5. Savic I, Perski A, Osika W. MRI shows that exhaustion syndrome due to chronic occupational stress is associated with partially reversible cerebral changes. Cereb Cortex 2018;28(3):894-906. [doi: 10.1093/cercor/bhw413] [Medline: 28108490]
6. Peristera P, Westerlund H, Magnusson Hanson LL. Paid and unpaid working hours among Swedish men and women in relation to depressive symptom trajectories: results from four waves of the Swedish longitudinal occupational survey of health. BMJ Open 2018;8(6):e017525 [FREE Full text] [doi: 10.1136/bmjopen-2017-017525] [Medline: 29880559]
7. Poms L, Fleming L, Jacobsen K. Work-family conflict, stress, and physical and mental health: a model for understanding barriers to and opportunities for women's well-being at home and in the workplace. World Med Health Policy 2016;8(4):444-457. [doi: 10.1002/wmh3.211]
8. Global gender gap report 2023. World Economic Forum. 2023. URL: https://www3.weforum.org/docs/WEF_GGGR_2023.pdf [accessed 2024-04-24]
9. Ornek OK, Esin MN. Effects of a work-related stress model based mental health promotion program on job stress, stress reactions and coping profiles of women workers: a control groups study. BMC Public Health 2020;20(1):1658 [FREE Full text] [doi: 10.1186/s12889-020-09769-0] [Medline: 33148247]
10. Santos FRMD, Lacerda SS, Coelhoso CC, Barrichello CR, Tobo PR, Kozasa EH. The integration of meditation and positive psychology practices to relieve stress in women workers (Flourish): effects in two pilot studies. Behav Sci (Basel) 2021;11(4):43 [FREE Full text] [doi: 10.3390/bs11040043] [Medline: 33810304]
11. Coelhoso CC, Tobo PR, Lacerda SS, Lima AH, Barrichello CRC, Amaro E, et al. A new mental health mobile app for well-being and stress reduction in working women: randomized controlled trial. J Med Internet Res 2019;21(11):e14269 [FREE Full text] [doi: 10.2196/14269] [Medline: 31697244]
12. Kabat-Zinn J. Mindfulness-based interventions in context: past, present, and future. Clinical Psychology: Science and Practice 2003;10(2):144-156. [doi: 10.1093/clipsy.bpg016]
13. Khoury B, Sharma M, Rush SE, Fournier C. Mindfulness-based stress reduction for healthy individuals: a meta-analysis. J Psychosom Res 2015;78(6):519-528. [doi: 10.1016/j.jpsychores.2015.03.009] [Medline: 25818837]
14. Galante J, Friedrich C, Dawson AF, Modrego-Alarcón M, Gebbing P, Delgado-Suárez I, et al. Mindfulness-based programmes for mental health promotion in adults in nonclinical settings: a systematic review and meta-analysis of randomised controlled trials. PLoS Med 2021;18(1):e1003481 [FREE Full text] [doi: 10.1371/journal.pmed.1003481] [Medline: 33428616]
15. Rusch H, Rosario M, Levison L, Olivera A, Livingston WS, Wu T, et al. The effect of mindfulness meditation on sleep quality: a systematic review and meta-analysis of randomized controlled trials. Ann N Y Acad Sci 2019;1445(1):5-16 [FREE Full text] [doi: 10.1111/nyas.13996] [Medline: 30575050]
16. Key ICT indicators for the world and special regions (totals and penetration rates). The International Telecommunication Union. 2024. URL: https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ITU_regional_global_Key_ICT_indicator_aggregates_Nov_2023.xlsx [accessed 2024-04-24]

XSL•FO
RenderX

17. Plaza I, Demarzo MMP, Herrera-Mercadal P, García-Campayo J. Mindfulness-based mobile applications: literature review and analysis of current features. JMIR Mhealth Uhealth 2013;1(2):e24 [FREE Full text] [doi: 10.2196/mhealth.2733] [Medline: 25099314]
18. Flynn S, Hastings RP, Burke C, Howes S, Lunsky Y, Weiss JA, et al. Online mindfulness stress intervention for family carers of children and adults with intellectual disabilities: feasibility randomized controlled trial. Mindfulness 2020;11(9):2161-2175. [doi: 10.1007/s12671-020-01436-0]
19. Liu C, Chen H, Zhou F, Long Q, Wu K, Lo L, et al. Positive intervention effect of mobile health application based on mindfulness and social support theory on postpartum depression symptoms of puerperae. BMC Womens Health 2022;22(1):413 [FREE Full text] [doi: 10.1186/s12905-022-01996-4] [Medline: 36217135]
20. Rigabert A, Motrico E, Moreno-Peral P, Resurrección DM, Conejo-Cerón S, Cuijpers P, et al. Effectiveness of online psychological and psychoeducational interventions to prevent depression: Systematic review and meta-analysis of randomized controlled trials. Clin Psychol Rev 2020;82:101931 [FREE Full text] [doi: 10.1016/j.cpr.2020.101931] [Medline: 33137611]
21. Sommers-Spijkerman M, Austin J, Bohlmeijer E, Pots W. New Evidence in the booming field of online mindfulness: an updated meta-analysis of randomized controlled trials. JMIR Ment Health 2021;8(7):e28168 [FREE Full text] [doi: 10.2196/28168] [Medline: 34279240]
22. Armstrong L. An investigation into the effects of a short-term mindfulness intervention on stress and emotion regulation in undergraduate students: understanding mechanisms of action. Manchester Metropolitan University. 2012. URL: https://www.semanticscholar.org/paper/An-investigation-into-the-effects-of-a-short-term-Armstrong/96cbd0ad6b2a2c1f19fb88d76fa247626d1bc860 [accessed 2026-01-08]
23. Basso JC, McHale A, Ende V, Oberlin DJ, Suzuki WA. Brief, daily meditation enhances attention, memory, mood, and emotional regulation in non-experienced meditators. Behav Brain Res 2019;356:208-220. [doi: 10.1016/j.bbr.2018.08.023] [Medline: 30153464]
24. Diener E, Emmons RA, Larsen RJ, Griffin S. The satisfaction with life scale. J Pers Assess 1985;49(1):71-75. [doi: 10.1207/s15327752jpa4901_13] [Medline: 16367493]
25. Sumino Z. Development of the Japanese version of the satisfaction with life scale. 1994 Presented at: Proceedings of the 36th Annual Meeting of the Japanese Association of Educational Psychology; September 28-30, 1994; Kyoto, Japan p. 28-30.
26. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983;24(4):385. [doi: 10.2307/2136404]
27. Sumi K. Reliability and validity of the Japanese version of the perceived stress scale [Article in Japanese]. Jpn J Health Psychol 2006;19(2):44-53. [doi: 10.11560/jahp.19.2_44]
28. Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SLT, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002;32(6):959-976. [doi: 10.1017/s0033291702006074] [Medline: 12214795]
29. Furukawa TA, Kawakami N, Saitoh M, Ono Y, Nakane Y, Nakamura Y, et al. The performance of the Japanese version of the K6 and K10 in the world mental health survey Japan. Int J Methods Psychiatr Res 2008;17(3):152-158 [FREE Full text] [doi: 10.1002/mpr.257] [Medline: 18763695]
30. Sakurai K, Nishi A, Kondo K, Yanagida K, Kawakami N. Screening performance of K6/K10 and other screening instruments for mood and anxiety disorders in Japan. Psychiatry Clin Neurosci 2011;65(5):434-441 [FREE Full text] [doi: 10.1111/j.1440-1819.2011.02236.x] [Medline: 21851452]
31. Spielberger CD. State-Trait Anger Expression Inventory (Staxi): Professional Manual. Tampa, FL: Psychological Assessment Resources, Inc; 1999.
32. Mine H, Ohki M. Applicability of anger expression inventory (STAXI-2) to Japanese. 2001 Presented at: Proceedings of the 14th Annual Convention of the Japanese Association of Health Psychology; November 3-4, 2001; Miyagi, Japan.
33. Mine H, Sato S. An attempt for the assessment of anger expression behavior (STAXI-2)--- comparison of anger expression behavior with U.S. and Japan ---. 2005 Presented at: The 18th Annual Convention of the Japanese Association of Health Psychology; September 1-3, 2005; Hyogo, Japan.
34. Brown KW, Ryan RM. The benefits of being present: mindfulness and its role in psychological well-being. J Pers Soc Psychol 2003;84(4):822-848. [doi: 10.1037/0022-3514.84.4.822] [Medline: 12703651]
35. Fujino M, Kajimura S, Nomura M. Development and validation of the Japanese version of the mindful attention awareness scale using item response theory analysis [Article in Japanese]. Jpn J Personal 2015;24(1):61-76. [doi: 10.2132/personality.24.61]
36. Kessler RC, Barber C, Beck A, Berglund P, Cleary PD, McKenas D, et al. The world health organization health and work performance questionnaire (HPQ). J Occup Environ Med 2003;45(2):156-174. [doi: 10.1097/01.jom.0000052967.43131.51] [Medline: 12625231]
37. Kessler R, Ames M, Hymel P, Loeppke R, McKenas DK, Richling DE, et al. Using the world health organization health and work performance questionnaire (HPQ) to evaluate the indirect workplace costs of illness. J Occup Environ Med 2004;46(6 Suppl):S23-S37. [doi: 10.1097/01.jom.0000126683.75201.c5] [Medline: 15194893]

XSL•FO
RenderX

38. Kawakami N, Inoue A, Tsuchiya M, Watanabe K, Imamura K, Iida M, et al. Construct validity and test-retest reliability of the world mental health japan version of the world health organization health and work performance questionnaire short version: a preliminary study. Ind Health 2020;58(4):375-387 [FREE Full text] [doi: 10.2486/indhealth.2019-0090] [Medline: 32173661]
39. Inoue A, Kawakami N, Shimomitsu T, Tsutsumi A, Haratani T, Yoshikawa T, et al. Development of a short version of the new brief job stress questionnaire. Ind Health 2014;52(6):535-540 [FREE Full text] [doi: 10.2486/indhealth.2014-0114] [Medline: 24975108]
40. Geurts SAE, Taris TW, Kompier MAJ, Dikkers JSE, Van Hooff MLM, Kinnunen UM. Work-home interaction from a work psychological perspective: development and validation of a new questionnaire, the SWING. Work & Stress 2005;19(4):319-339. [doi: 10.1080/02678370500410208]
41. Shimada K, Shimazu A, Geurts SAE, Kawakami N. Reliability and validity of the Japanese version of the survey work–home interaction – NijmeGen, the SWING (SWING-J). Community, Work & Family 2018;22(3):267-283. [doi: 10.1080/13668803.2018.1471588]
42. Siegel RD, Germer CK, Olendzki A. Mindfulness: What is it? Where did it come from? In: Clinical Handbook of Mindfulness. New York, NY: Springer New York; 2009.
43. Hölzel BK, Lazar SW, Gard T, Schuman-Olivier Z, Vago DR, Ott U. How does mindfulness meditation work? Proposing mechanisms of action from a conceptual and neural perspective. Perspect Psychol Sci 2011;6(6):537-559. [doi: 10.1177/1745691611419671] [Medline: 26168376]
44. Crain TL, Schonert-Reichl KA, Roeser RW. Cultivating teacher mindfulness: effects of a randomized controlled trial on work, home, and sleep outcomes. J Occup Health Psychol 2017;22(2):138-152. [doi: 10.1037/ocp0000043] [Medline: 27182765]
45. Espel-Huynh H, Baldwin M, Puzia M, Huberty J. The indirect effects of a mindfulness mobile app on productivity through changes in sleep among retail employees: secondary analysis. JMIR Mhealth Uhealth 2022;10(9):e40500 [FREE Full text] [doi: 10.2196/40500] [Medline: 36169994]
46. Hülsheger UR, Alberts HJEM, Feinholdt A, Lang JWB. Benefits of mindfulness at work: the role of mindfulness in emotion regulation, emotional exhaustion, and job satisfaction. J Appl Psychol 2013;98(2):310-325. [doi: 10.1037/a0031313] [Medline: 23276118]
47. Kappen G, Karremans JC, Burk WJ. Effects of a short online mindfulness intervention on relationship satisfaction and partner acceptance: the moderating role of trait mindfulness. Mindfulness 2019;10(10):2186-2199. [doi: 10.1007/s12671-019-01174-y]
48. Kiburz KM, Allen TD, French KA. Work–family conflict and mindfulness: investigating the effectiveness of a brief training intervention. J Organ Behavior 2017;38(7):1016-1037. [doi: 10.1002/job.2181]
49. Nicklin JM, Shockley KM, Dodd H. Self-compassion: implications for work-family conflict and balance. J Vocat Behav 2022;138:103785. [doi: 10.1016/j.jvb.2022.103785]
50. Slutsky J, Chin B, Raye J, Creswell JD. Mindfulness training improves employee well-being: a randomized controlled trial. J Occup Health Psychol 2019;24(1):139-149. [doi: 10.1037/ocp0000132] [Medline: 30335419]
51. Jayewardene WP, Lohrmann DK, Erbe RG, Torabi MR. Effects of preventive online mindfulness interventions on stress and mindfulness: a meta-analysis of randomized controlled trials. Prev Med Rep 2017;5:150-159 [FREE Full text] [doi: 10.1016/j.pmedr.2016.11.013] [Medline: 28050336]

Abbreviations

ANCOVA: analysis of covariance
BJSQ: Brief Job Stress Questionnaire
CONSORT: Consolidated Standards for Reporting Trials
FWNS: family-work negative spillover
FWPS: family-work positive spillover
K6: 6-item Kessler Psychological Distress Scale
RCT: randomized controlled trial
STAXI-2: State-Trait Anger Expression Inventory 2
T-Ang/R: Trait-anger/reaction
T-Ang/T: Trait-anger/temperament
UMIN: University Hospital Medical Information Network
WFNS: work-family negative spillover
WFPS: work-family positive spillover

XSL•FO
RenderX

Edited by T McCall; submitted 01.Jun.2024; peer-reviewed by U Sinha, D Abdel-Hady, ID Yucel; comments to author 13.Jan.2025; revised version received 07.Nov.2025; accepted 24.Nov.2025; published 02.Feb.2026.

Please cite as:
Uwagawa R, Adachi K, Shimoda M, Takizawa R
Effects of an 8-Week App-Based Mindfulness Intervention on Mental Health in Working Women: Randomized Controlled Trial
J Med Internet Res 2026;28:e62814
URL: https://www.jmir.org/2026/1/e62814
doi:10.2196/62814
PMID:41627885

©Riko Uwagawa, Koichiro Adachi, Mariko Shimoda, Ryu Takizawa. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Changing Habits With the Happy Hands App: Qualitative Focus Group Study of a Hand Osteoarthritis Self-Management Intervention

Kristine Aasness Fjeldstad[1,2], MSc; Anne Therese Tveter[1], PhD; Eivor Rasmussen[3], BSc; Lena Olden[4], BSc; Sissel Nyheim[5], MSc; Thalita Blanck[1], MSc; Rikke Munk Killingmo[2], PhD; Ingvild Kjeken[1], PhD

[1]Health Service Research and Innovation Unit, Center for Treatment of Rheumatic and Musculoskeletal Diseases (REMEDY), Diakonhjemmet Hospital, Oslo, Norway

[2]Faculty of Health Sciences, Department of Rehabilitaion Science and Health Technology, OsloMet – Oslo Metropolitan University, Oslo, Norway

[3]Hand Therapy Clinic, Kirkenes Hospital, Kirkenes, Norway

[4]Department of Physiotherapy and Occupational Therapy, Levanger Hospital, Levanger, Norway

[5]Norwegian Rheumatism Association, Oslo, Norway

Corresponding Author:
Kristine Aasness Fjeldstad, MSc
Health Service Research and Innovation Unit
Center for Treatment of Rheumatic and Musculoskeletal Diseases (REMEDY)
Diakonhjemmet Hospital
Postboks 23 Vinderen
Oslo, 0319
Norway
Phone: 47 22451500
Email: KristineAasness.Fjeldstad@diakonsyk.no

Abstract

Background: People with hand osteoarthritis represent a large patient group with limited access to recommended treatment. In recent years, there has been a notable shift in health care delivery, with increased use of digital technologies. The Happy Hands app (The University Information Technology Center [USIT]) is a digital self-management intervention developed to provide evidence-based treatment for people with hand osteoarthritis, with the goal of empowering them to self-manage their disease. Participants' experiences and perceptions of using this digital intervention are crucial for the adoption and continued use of the Happy Hands app.

Objective: The objective of this qualitative study was to explore participants' experience with using the Happy Hands app, focusing on whether and how it empowered them to self-manage their hand osteoarthritis.

Methods: The study is embedded within a randomized controlled trial (RCT). The participants were recruited from the intervention group in the RCT, who got access to the Happy Hands app. The 12-week self-management intervention included a hand exercise program and informational videos about hand osteoarthritis. Focus groups were conducted in various geographical areas in Norway. The focus groups were transcribed verbatim, coded, and analyzed inductively using reflexive thematic analysis.

Results: Seven focus groups, with a total of 26 participants, were recruited from both specialist and primary health care. The mean age was 67 years. Three themes were developed from the analysis. The first theme, "Being acknowledged," highlights the essential role of recognition for people with hand osteoarthritis. It suggests that the Happy Hands app provided participants with a sense of validation and support. The second theme, "Changed perception of hand osteoarthritis," indicates that participants gained insights and knowledge about their condition. This new understanding empowered them to make more informed decisions about their care, fostering a sense of hope and motivation by demonstrating that effective measures are available to manage the disease. The third theme, "Changing habits with the Happy Hands app," describes how participants developed new habits after using the self-management intervention delivered through the app. The exercise program was experienced as motivating, flexible, well-structured, and committing. Some challenges were reported, including experiencing pain during or after exercising. The new habits included performing hand exercises and implementing ergonomic working methods, which were tailored to meet the individual needs and integrated into the participants' daily lives and routines.

XSL•FO
RenderX

Conclusions: The findings suggest that the Happy Hands app is a valuable tool for supporting people with hand osteoarthritis in managing their disease by helping them integrate hand osteoarthritis management into their daily lives.

Trial Registration: ClinicalTrials.gov NCT05568875; https://clinicaltrials.gov/study/NCT05568875

(J Med Internet Res 2026;28:e82773) doi:10.2196/82773

KEYWORDS

mHealth; eHealth; osteoarthritis; hand exercises; self-management; social cognitive theory; qualitative research

Introduction

Hand osteoarthritis is a prevalent and debilitating condition. Nearly half of all women and a quarter of all men will develop hand osteoarthritis during their lifetime [1], with symptoms often emerging around middle adulthood. People with hand osteoarthritis typically experience pain, stiffness, and reduced grip strength [2]. These symptoms challenge individuals' ability to perform everyday activities, such as household tasks, hobbies, and work, and may reduce health-related quality of life. Although there is no cure for hand osteoarthritis, various effective measures can help alleviate symptoms [3]. Access to treatment is, however, often limited, and the quality of care is suboptimal [4], with patients often being referred to surgery without having received recommended treatment [5].

First-line treatments for people with hand osteoarthritis include patient education, hand exercises, and the use of assistive devices [3,6]. Patient education is central, as much of the management of hand osteoarthritis relies on actions individuals have to perform themselves. However, supporting individuals in self-managing a chronic condition like hand osteoarthritis can be challenging [4], particularly in maintaining adherence [7].

There has been a significant shift in the delivery of health care in recent years, with an increased emphasis on the use of digital devices in the provision of care. Government documents have highlighted eHealth as a key strategy to optimize resource use, streamline patient pathways, and ensure that individuals get access to health information [8].

Previous studies have demonstrated that digital delivery for hand osteoarthritis is a good alternative to traditional in-person consultations [9] and effective in improving hand function and reducing pain compared to usual care [10]. Qualitative studies have described participants being mostly positive to digital delivery of osteoarthritis management [11-13].

The use of mobile apps offers a valuable tool for ensuring access to treatment and supporting individuals in self-managing their hand osteoarthritis [6]. Mobile apps can offer tailored information and guided exercise routines, which may also improve exercise adherence [14,15].

The Happy Hands app (The University Information Technology Center [USIT]) contains a 12-week self-management intervention for people with hand osteoarthritis, developed to ensure access to evidence-based treatment. The goal of the app is to empower people with hand osteoarthritis to self-manage their disease [15]. Before this study, the app underwent feasibility testing aimed at refining and improving its design. The results indicated that the Happy Hands app contributed to reductions in pain and stiffness, as well as improvements in activity performance and grip strength. In focus group interviews, participants reported that the app was useful and highlighted several areas for further enhancement [15]. In this study, the aim was to explore participants' experience with using the Happy Hands app, focusing on whether and how it empowered them to self-manage their hand osteoarthritis.

Methods

Study Design

This qualitative study is conducted as part of a randomized controlled trial (RCT) investigating the effect and cost-effectiveness of the Happy Hands app. Focus groups were conducted with participants from the intervention group in the RCT to explore their experiences of using a digital intervention for hand osteoarthritis.

Happy Hands App

The Happy Hands app was developed by a research group at Diakonhjemmet hospital [15]. The development of the app was guided by social cognitive theory [16]. The behavior change taxonomy, a classification system categorizing different techniques used to change behavior [17], was used to classify the elements in the app.

The app contains a self-management intervention with informational videos and a hand exercise program. The informational videos include different themes, such as information about hand osteoarthritis, hand exercises, use of assistive devices and orthoses, medication and surgical options, and how to cope with everyday life. Furthermore, the self-management program includes a hand exercise program with videos showing how the patients should warm up and perform exercises to improve mobility, strength, and coordination, and a stretching exercise. The informational videos and exercise program are delivered in a progressive order across 12 weeks. The participants could choose 3 days a week to use the app. On these designated days they received notifications on their smartphones with that week's informational videos and hand exercises. The participants had the opportunity to tick off each completed exercise. Encouragement, motivational messages, and quizzes were provided each week to enhance continued adherence to the intervention [18].

Setting

This qualitative study was nested within the Happy Hands study, a multicenter RCT that evaluated whether a self-management intervention delivered through the Happy Hands app, in addition to usual care, was more effective than usual care alone for

XSL•FO
RenderX

people with hand osteoarthritis [18]. The RCT was conducted between November 2022 and February 2024. Participants in the RCT were recruited from 14 hospitals, 2 rehabilitation centers, 3 physiotherapy clinics, and 1 private rheumatology center, following their regular consultations at these sites. In total, 376 participants were enrolled and randomly allocated to either the control group (n=185) or the intervention group (n=191).

The control group received a one-page information sheet and usual care. Usual care for people with hand osteoarthritis varies across settings and can include no treatment, a one-day educational course, or consultations with an occupational therapist and/or a rheumatologist. The intervention group received the same one-page information sheet and usual care; in addition, they were given access to the Happy Hands app, which provided them with a 12-week self-management program.

Participants and Recruitment

In the RCT, eligible participants had symptomatic hand osteoarthritis diagnosed by health care personnel, owned a smartphone, and could read and understand Norwegian. Participants were excluded from the RCT if they had cognitive deficits, were scheduled for hand surgery within 3 months after inclusion, or had serious comorbidities or inflammatory rheumatic disease. Participants were recruited through clinical practice. Clinicians applied the inclusion and exclusion criteria to determine eligibility. Information on serious comorbidities was obtained through participant self-report, and cognitive function was assessed through clinical evaluation.

Participants in the focus groups were recruited from the intervention group in the RCT. When they were enrolled in the RCT and signed the consent form, participants could indicate their willingness to be contacted for an interview about their experiences with the app. A purposive sampling strategy was used to ensure diversity in geographic location and health care setting. Participants were recruited from various regions across Norway to ensure geographic diversity and from both primary and specialist care. Individuals in the intervention group who consented to an interview were contacted by phone. In total, 28 participants agreed to participate. However, 2 of the participants declined before the focus groups were conducted due to unknown reasons. Thus, a total of 26 participants participated in 7 focus groups.

The focus groups took place between 7 and 11 months after recruitment for the RCT began. All focus groups were conducted in 2023, the first 6 in June and the last one in October. We chose this timing to ensure that participants had completed the intervention, enabling us to explore their reflections on the full intervention period. One participant, however, joined a focus group after one month because of late enrollment, whereas the rest were interviewed after 3 or more months. Adherence to app use throughout the 12-week program period was not an inclusion criterion.

Data Collection

Focus groups were chosen as the data collection method in this study because they are useful to obtain in-depth understanding of a topic by taking advantage of the group dynamics and discussions that can occur in a group setting [19].

All 7 focus groups were conducted by the first author (KAAF). Either the last author (IK), a master student associated with the project, or one of the clinicians working at the recruitment sites assisted as moderators in the focus groups by taking notes and asking follow-up questions. KAAF is a female PhD candidate and nurse with some experience in qualitative research, while IK is an occupational therapist with extensive experience in both clinical practice and research. The master student is also an occupational therapist and was writing a thesis concerning the Happy Hands app. The clinicians were working as either a physiotherapist or occupational therapist at the recruitment sites and assisted in recruiting patients in their clinical practice to the RCT and have experience working with individuals with hand osteoarthritis. KAAF had previously met a few of the participants in connection with their inclusion in the RCT, while IK and the master student had no prior relationship with the participants before the study. The clinicians had previously met some of the participants during consultations.

The focus groups were held at the site where the participants were recruited to the RCT, at a hospital or clinic, except for one focus group that was held in a conference room at a hotel. Before the focus groups began, participants were informed about the aim of the study. They were told that any experience, both positive and negative, they had had with the Happy Hands app was welcomed and were encouraged to discuss and ask each other questions. This aligns with Liamputtong [20], who describes that focus groups allow for more topics to be initiated by participants, as they steer the conversation more than the researcher does in individual interviews.

A first version of the interview guide was developed by KAAF, ATT, and IK and focused on participants' experiences, problems with using the app, perceived benefits, and future use (Multimedia Appendix 1). It was reviewed by patient research partners (SN and TB), who suggested adding a question about initial experiences, including whether users received advice and support. They also recommended using the term "challenging" instead of "difficult" to frame questions more positively. The interview guide was thereafter tested in a pilot focus group. As no major revisions were necessary, the data from this session was included in the final dataset.

At the end of the focus group, the participants were asked if they had anything to add. Each focus group consisted of 3-5 participants and lasted between 50 minutes and one and a half hours. The audio recordings were transcribed verbatim by KAAF and a research assistant. Notes were taken by the moderator during each focus group. Following each session, the moderator and interviewer (KAAF) discussed their impressions of the content and clarified any points that required further understanding.

We applied the concept of information power to determine whether we had a sufficient sample size. This approach suggests that factors such as the study aim, participant specificity, theoretical framework, interview quality, and analysis strategy should guide decisions about when sufficient information has been obtained [21]. Our study sought to explore user experience

XSL•FO
RenderX

in a broad context, which typically necessitates a diverse sample. We therefore decided to include focus groups from different regions in Norway, covering various levels of care. Accordingly, 6 focus groups were conducted in different geographical areas, encompassing both primary and specialist health care settings. We anticipated that this strategy would yield sufficient data to address our research questions. Following the initial round of analysis, we noted a lack of information regarding negative experiences with the app and how it would be used in the future. To strengthen our understanding of this aspect, we conducted one additional focus group. We then considered that the focus groups provided a comprehensive understanding of the participants' experiences and perspectives. The data had sufficient depth and richness to address the study aim, and thus, we concluded that information power was achieved.

Theoretical Framework

Following an initial phase of inductive analysis and coding, social cognitive theory (SCT) [16] was applied to explore whether participants had benefited from the intervention and achieved behavior change. SCT outlines key determinants for behavior change: knowledge about health risks, outcome expectations, goal-setting and strategies to achieve them, and perceived facilitators and barriers. A crucial aspect is self-efficacy, which is the belief that individuals have in their ability to achieve change through their actions [16]. Additionally, observational learning, learning a new behavior by watching others perform it, is a central concept [22].

To complement this behavioral perspective, the common-sense model of illness representations developed by Leventhal et al [23], was also applied to the analysis in this study. While SCT focuses on the mechanisms underlying behavior change, the Common-Sense Model helps to explain how individuals understand and make sense of their illness. It posits that illness representations are shaped by various sources of information, such as societal lay knowledge, advice from health care professionals, and personal experiences with the illness. Illness representation can be divided into 5 different dimensions. "Cause" considers beliefs about what is causing the disease. "Consequences" addresses beliefs regarding its impact, "identity" is beliefs regarding the symptoms, and "timeline" refers to beliefs about the duration and progress. "Cure or control" is the belief about the potential for recovery or management and the individual's ability to influence their condition. These dimensions collectively shape how patients perceive their illness and have significance for the way individuals will seek help and for adopting a coping strategy for the disease [24].

Involvement of Patient Research Partners

Two patient research partners (SN and TB), both of whom completed the 12-week app program, were involved from the outset of the study and contributed to the review of the interview guide. Following a presentation of the results, a discussion was held between them and IK and KAAF. The patient research partners indicated that they identified with the results and the developed themes. They are both coauthors of this article and have reviewed and provided feedback on the manuscript.

Data Analysis

Reflexive thematic analysis was applied to analyze the interviews, aiming to explore, interpret, and develop patterns of meaning across the dataset. This method involves 6 key phases, including familiarizing oneself with the data, generating initial codes, constructing themes, reviewing themes, defining and naming themes, and producing the final report. A key concept in this method is reflexivity, which involves continuously reflecting on our assumptions and practices and how these affect the research [25]. Analysis was performed by the first author (KAAF), the second author (ATT), and the last author (IK). Throughout the research process, we actively engaged in reflexivity through several discussions. These discussions enabled us to examine our perceptions and how they might have influenced the research.

After conducting the first 6 focus groups, we performed an initial analysis by reading through the transcripts multiple times. The aim of this familiarization phase was to gain an overall understanding of the data. Notes were written throughout this process to identify interesting elements and possible patterns across the dataset and as a tool to be able to reflect on interpretations [25]. Following this familiarization phase, we also decided to conduct one additional focus group.

Each transcript was coded by KAAF using NVivo software (QSR International Pty Ltd). Text segments relevant to the aim of the study were tagged with distinct codes for different meanings. The initial analysis was conducted inductively. Coding was performed at a semantic level, staying close to the participants' language**.** In a second round of coding, a deductive approach was applied, looking for segments of text relevant to the theories we planned to include. The second and last authors (ATT and IK) each read half of the transcripts, and the team met to discuss reflections and interpretations on content in data and potential themes.

The codes were gathered, and 3 initial themes were developed from them. The tentatively developed themes were reviewed against all the codes clustered around each theme. A discussion was held with the patient research partners, IK and KAAF, following a presentation about the results. They expressed recognition of the developed themes and highlighted the importance of feeling seen and heard, access to information, and how exercises had become a natural part of their daily routines, facilitated by the app. Quotes were selected from the transcripts to illustrate the findings. Finally, the themes were defined and given names. All coauthors reviewed the drafts and provided feedback.

Reflexive notes were written throughout the research process to increase awareness of preconceptions and to enhance transparency. Initial preconceptions were documented at the beginning of the study and later compared with the final results. The considerable differences between them can be understood as an indication that the first author's preconceptions did not prevent new insights from emerging.

Ethical Considerations

This study was approved by the Regional Committees for Health Research Ethics (477746), the Data Protection Officer (00660),

XSL•FO
RenderX

and the local research committee at Diakonhjemmet hospital. Informed consent was obtained from all participants. The interviews were audio recorded by the Nettskjema-Diktafon app (University of Oslo), which safely transfers encrypted audio files to Services for Sensitive data (TSD) at the University of Oslo. TSD is a platform for collecting and storing sensitive data in compliance with the Norwegian privacy regulation. The qualitative study is briefly described in the trial description for the RCT registered on ClinicalTrials.gov (NCT05568875). We followed the Consolidated Criteria for Reporting Qualitative Research (COREQ; Multimedia Appendix 2) [26]. Participants did not receive compensation for taking part in the focus groups but were offered reimbursement for travel expenses.

Results

Participants' Characteristics

A total of 26 participants (18 women and 8 men) participated in 7 focus groups. The mean age was 67 years. Participants had hand osteoarthritis diagnosis for between one and 25 years (Table 1).

Table 1. Participant characteristics (N=26).

Characteristic	Participants
Age (years), mean (SD)	67 (7)
Gender, n (%)	
Women	18 (69)
Men	8 (31)
Years living with hand osteoarthritis, n (%)	
0-9	13 (50)
10-19	11 (42)
More than 19	2 (8)
Geographical area, n (%)	
Southeast	15 (58)
West	4 (15)
Central	3 (12)
North	4 (15)

Through the analysis we developed 3 themes. The theme "Being acknowledged" highlights the importance informants placed on having their hand osteoarthritis recognized and validated. Meanwhile, the theme "Changed perception of hand osteoarthritis" reflects how their perceptions of hand osteoarthritis changed after gaining insights from the self-management intervention in the app. Finally, "Changing habits with the Happy Hands" app reports how people with hand osteoarthritis integrated hand exercising and ergonomic working methods into their daily life and routines.

Being Acknowledged

A discussion about a feeling of not getting support and recognition for their hand osteoarthritis emerged in one of the focus groups. A few participants described a feeling of not receiving adequate health care. For example, one participant described having had pain for several years and experiencing a prolonged wait before finally receiving a diagnosis:

> *I've actually had pain for many years, but no one understood what it was. [...] I go to the doctor, no, it's not that, I take tests, nothing happens, but this year I got the diagnosis.* [Woman, 57 years]

In one focus group it was discussed that osteoarthritis is a disease that is given little attention in society, even though it affects many people. The participants described lacking support from the workplace and that they did not talk much about their disease with family or friends. Moreover, participants described a sense of dealing with the burden of hand osteoarthritis alone:

> *It is not talked about a lot, you go around with pains and you don't talk about it. We're supposed to manage it, it's not a problem, we women can fix it.* [Woman, 72 years]

Being part of the project and getting access to the app was therefore perceived as meaningful, as these experiences provided a sense of validation and acknowledgment. This was described in one of the focus groups.

I want to take a slightly different approach, which has actually been my problem. I've had pain and limitations in activity for a very long time, and that leads to, now I'm almost about to cry, poor sleep, and a low mood for a long time […] you feel like it's just something we have to live with […]So when [the doctor] said that I could be part of this, I was really happy […] That someone cares and someone sees you. Yes. [Woman, 75 years]

Furthermore, the focus group revealed that after getting the app and becoming part of the Happy Hands study, participants were increasingly willing to discuss hand osteoarthritis with their friends and family, having previously avoided doing so.

> *But what is currently the most important thing is all my female friends […] it has started to become a topic*

XSL•FO
RenderX

> *of conversation. […] There have been times when we've sat together and done some exercises at gatherings, saying, 'Now we're stretching our fingers,' and turning it into a little show.* [Woman, 72 years]

This willingness to share experiences with others reflects an increased openness about the condition, which may contribute to a sense of acknowledgment and support.

The theme highlights how some participants perceive a lack of support for their hand osteoarthritis, making their involvement in the study and access to the app feel like a genuine acknowledgment of their condition. This experience encouraged them to open up about their struggles with friends and family, indicating that it developed a sense of community and support.

Changed Perception of Hand Osteoarthritis

Through the self-management intervention delivered through the app, participants gained knowledge and insights about hand osteoarthritis, which seemed to have changed their perception of the disease in different ways. Participants described that the app gave them hope and motivation by demonstrating that different measures could be taken to manage the disease effectively.

> *It's important that one can do something about it, because I think it is what has been most discouraging in a way, that there isn't anything to do about it. It just has to run its course. And then, just the fact that you see that there is something you can do.* [Woman, 65 years]

Moreover, knowledge about the disease and its mechanisms reduced the feeling of uncertainty, for example, by getting an understanding of why they had pain and why some activities suddenly were difficult to perform.

> *You are now more aware of why it is there, right [...] what is it now, what have I done, so the uncertainty is gone, now you sort of know what you have to deal with.* [Man, 66 years]

The participants emphasized that it was important to receive knowledge about recommended treatment options for patients with hand osteoarthritis. This could be information about surgery, medication, use of orthoses and how to modify activity performance. This information could empower them to make informed decisions about their care based on updated knowledge. For example, in one of the focus groups, it was a consensus that they appreciated knowing that surgery is recommended as the last treatment option, since they did not want to have surgery. One participant was referred to surgery but decided to wait and see if he would benefit from doing hand exercises.

> *So, there were a number of things that became clear now […] it was like, you were given the steps and that surgery was presented as the last option, right. […] And... I liked that, because I don't want to be cut open.* [Man, 66 years]

A few participants described gaining knowledge that using their hands can have a preventive effect and that it is not harmful to use their hands even when experiencing pain:

> *I was a bit skeptical at first because I was afraid that doing something painful might be harmful. I thought I was causing damage, so I was very cautious. But now, I am willing to tolerate some pain as long as it doesn't get worse. In a way, it's reassuring to know this, and it has helped me normalize the experience.* [Woman, 64 years]

This theme describes how participants gained valuable insights about hand osteoarthritis, which shifted their perceptions of the disease. They expressed that increased knowledge about treatment options and their understanding of their condition alleviated uncertainty, supported them to make more informed decisions about their care, and provided hope and motivation by demonstrating that there exist measures that can help manage the disease.

Changing Habits With the Happy Hands App

This theme explores how participants changed their habits after completing the self-management intervention in the Happy Hands app, particularly in exercising and activity performance.

Developing New Habits: Exercising

The app's exercise component helped participants learn hand exercises and integrate them into their daily lives. Participants unanimously agreed that it was easy to use the app, with no one reporting any difficulties in understanding its technical aspects. Many participants emphasized the flexibility of the app as a key factor in its success, appreciating the ability to perform the exercises anywhere and at a time that was convenient for them.

> *You can exercise almost anywhere. I have exercised outdoors, I have exercised indoors in different places, I've kind of brought along a small box with equipment. It is very practical not to be tied to a specific table or chair or anything like that.* [Woman, 74 years]

The participants described that the app instilled a sense of commitment by requiring them to complete specific exercises within a designated timeframe. This sense of obligation was further reinforced by their participation in a research project and the requirement to tick off each completed exercise within the app. Several participants expressed that receiving notifications on their smartphones served as helpful reminders to complete their exercises. Moreover, scheduling 3 fixed days per week provided structure and consistency.

The app includes videos that demonstrate how to perform the exercises, which participants noted improved their ability to execute them correctly. A few even described feeling a sense of connection with the person in the videos, as if they were exercising alongside someone, which they found motivating.

> *I think it's very important that you see a person demonstrating the exercises, rather than just sitting and reading something. For me, there's a significant difference between the two. […] The physical or psychological—I'm not sure exactly what to call*

XSL•FO
RenderX

> *it—connection makes a big difference when it comes to actually doing those exercises.* [Woman, 72 years]

The possibility to track their progress in the app, complete quizzes, and receive feedback was also described as a motivation by some of the participants. Moreover, these features encouraged them to put in extra effort to both understand and retain the app' content.

Participants in the focus groups reported experiencing various improvements. When they experienced improvement, for instance, in improved strength, mobility, or ability to perform activities, they were motivated to carry out the training program.

> *I found it very motivating when I noticed that, wow, I have much more mobility, it didn't come the first week, not the second week, and not the third week, but maybe in the fourth and fifth week I started to get better. I can bend all my fingers now, which I couldn't do before, and then you understand, this is working, and it is also motivating to continue.* [Woman, 57 years]

There were, however, also some challenges described when it came to using the Happy Hands app. These included experiencing pain during or after exercising. Some participants said that their hand osteoarthritis was too advanced for effective exercising. Not everyone reported improvement. It was suggested that the app should include guidance on how to manage pain during exercise.

> *But my experience has been, in short, that I have a few fingers that have very mild osteoarthritis. And I can feel how good this app is for those fingers. At the same time, I notice that it's not very good for my bad fingers, as they actually get provoked and irritated by some of them [the exercises]*. [Woman, 65 years]

There were also other reasons for not continuing to perform hand exercises. For example, one participant said she had stopped doing exercises simply because she had more going on in her life and she had forgotten about it. Other diseases or injuries could also take away the focus of conducting hand exercises. Additionally, some participants expressed that they wished that new exercises had appeared in the app, as this would have been motivating.

The results suggest that the Happy Hands app enabled participants to learn an exercise program for hand osteoarthritis, to integrate it into their daily routines, and ultimately to establish it as a habit. Once the participants had memorized the exercises, it was no longer necessary to watch the app while performing them. This made it possible to integrate exercising into their daily life by performing them while engaged in other activities.

> *Now I notice that I sit and watch the news while doing it with a tennis ball and perform many of the exercises a bit automatically in my daily life. When you go for a walk, […] many of them are easy to continue with.* [Woman, 67 years]

They could also adjust the exercising to when and how it suited the individual. For example, some preferred following the exercise program less structured. Other participants preferred to continue to exercise in a structured way, 3 times a week with the app in front of them.

Moreover, the participants adjusted the exercises because of other factors. For example, conducting exercises with less intensity when experiencing pain or conducting exercises differently when finding the exercise difficult to carry out.

> *… squeezing that ball has also started to hurt more, as I said you should squeeze as hard as you can for five seconds, five times, per hand. So, I'm trying to be a bit more careful; I'm still trying to do it and hold it, but maybe I'm not squeezing as hard simply because it hurts more than it did before.* [Man, 59 years]

The majority of the participants said they planned to continue to exercise. These results suggest that the app facilitated making exercising a habit.

This theme discusses how the app helped participants integrate hand exercises into their daily routines. Participants found the app easy to use and appreciated its flexibility. A sense of commitment was supported by the need to complete the exercise on time, reminders from the app, and scheduled sessions. Seeing a person demonstrate the exercises gave participants motivation, as well as making them easier to understand. Many experienced improvements that enhanced motivation to continue. However, some participants faced challenges, such as pain during or after exercising or osteoarthritis too advanced to benefit from exercise. The app facilitated the development of new habits for many participants, enabling them to integrate exercises into their daily life and adjust them to their individual needs and circumstances.

Developing New Habits: Adapting Daily Activities

The findings also reveal that participants made changes to how they performed everyday activities, including ergonomic working methods and use of assistive devices as suggested in the app. The adaptations they learned through the app made some activities easier to perform. For example, participants reported learning new techniques for holding and lifting objects, such as avoiding carrying items solely with their fingers or warming up their hands before use. These changes seemed to have had a significant impact on their daily life, as one participant described,

> *I do notice a difference, yes, [...] and then I think about warming up when I'm going to write, for example. Once I've warmed up, it's not so embarrassing to write my name anymore. It was embarrassing [...] it was almost like being naked, I would say, sixty-seven, and I could hardly write my name.* [Man, 67 years]

Additionally, participants found the information about different assistive devices to be helpful. For example, they mentioned tools like a nutcracker for opening soda bottles, as well as other devices such as a cheese cutter and a bread knife with ergonomic handles. However, a few participants expressed some reluctance to using assistive devices, as they associated them with "being old."

XSL•FO
RenderX

This theme described how participants also changed their performance of everyday activities after insights they had gained from the app.

Discussion

Principal Findings

The aim of this study was to explore participants' experience with using the Happy Hands app, focusing on whether and how it empowered them to self-manage their hand osteoarthritis. Our findings indicate that participants previously had experienced a lack of support and recognition for their hand osteoarthritis. Hence, to become part of the study and get access to the app provided them with a sense of validation and acknowledgment. Furthermore, the results suggest that participants' illness perception changed after gaining insights from the informational videos in the app, from viewing their condition as something insignificant and beyond their control to recognizing its importance, taking their own experiences seriously, and gaining insights into strategies they could use to alleviate pain and improve function.

Participants perceived the app as flexible, motivating, and well-structured, which facilitated both their learning and adherence to the exercise program as well as alternative working methods. As a result, they developed new habits, such as integrating hand exercises into their daily life and changing the way they perform everyday activities. However, for some participants, their condition hindered their ability to complete the exercises, often due to pain.

The findings in our study reveal that some participants felt unsupported and lacked recognition in their struggles with hand osteoarthritis by family members, health care providers, and society at large. This lack of understanding and acknowledgment is consistent with findings from other qualitative studies. Hill et al [27] noted that individuals with hand osteoarthritis perceive a lack of empathy from health professionals regarding the impact of the disease, whereas Gignac et al [28] found that symptoms were often dismissed as a normal part of aging. In our study, participants reported minimizing the severity of their condition, rarely discussing their hand osteoarthritis despite its significant impact on their daily life. This tendency to downplay symptoms aligns with the findings of Magnussen et al [29], who reported that people with hand osteoarthritis often feel undeserving of health care. Bukhave and Huniche [30] also found that participants did not seek medical care, despite experiencing a wide range of activity limitations. According to the common-sense model of illness representation [23], people's perceptions and understanding of their disease influence their behaviors and coping strategies. It is therefore essential for health care professionals to explore patients' previous experiences with the health care system and to actively acknowledge their perceived symptoms and functional challenges. Feeling that their condition is validated may have shifted participants' views from considering their hand osteoarthritis as insignificant and something they simply had to endure to recognizing its importance and the impact it has on their daily lives.

As outlined in the common-sense model of illness representations, various kinds of knowledge contribute to a person's perception of their illness [24]. The findings in our study suggest that participants changed their illness perception by acquiring new knowledge. They gained expert knowledge through the app, where health professionals provided valuable information about hand osteoarthritis. Additionally, they developed new personal knowledge through their own experiences, such as noticing improvement and regaining the ability to perform activities. This suggests that the newly acquired knowledge enabled participants to develop a new understanding of their illness, resulting in a shift in their illness perception.

According to the common-sense model of illness representations, illness perception can be categorized into different dimensions, including the "cause" dimension, which refers to what people believe to be the underlying reason for their illness [24]. Our findings suggest that participants altered their perception of this dimension by using the app, transitioning from uncertainty and fear about the origins of their pain, worried that activity might worsen their condition, to a more informed understanding. They came to recognize the actual cause of their illness and understood that using their hands would not exacerbate their pain.

Our results indicate that participants' perceptions shifted from believing nothing could be done about the disease to recognizing that effective measures are available. This aligns with the "cure or control" dimension of the common-sense model of illness representations, which describes a person's belief about how a disease can be managed [24]. The initial perception that nothing can be done aligns with findings from other studies, which report that individuals often view hand osteoarthritis as a natural result of wear and tear or aging [27,31], and believe that the condition is untreatable [29]. Participants in our study also changed their perception of their own ability to take action. For instance, information about treatment options helped them understand that delaying surgery is often a recommended approach. This shift in perception, from viewing hand osteoarthritis as untreatable to acknowledging the availability of effective management strategies, empowered participants to begin adopting these strategies.

An important finding in our study is that participants developed new habits by integrating hand exercises into their daily routines and adjusting them to their individual needs. The results further suggest that the Happy Hands app played a key role in facilitating this behavioral change. SCT, which is widely used in the development of health interventions [22], may help explain how the Happy Hands app contributed to this change. By supporting participants to gradually learn and implement the exercise program, the app enabled them to master activities that were previously difficult.

The findings further reveal that due to new behaviors, such as hand exercising, participants experienced positive outcomes, including reduced pain and an improved ability to perform previously challenging activities. In SCT, self-efficacy, a person's belief in their ability to make necessary changes to achieve a desired outcome, is a central concept. Self-efficacy

XSL•FO
RenderX

can be enhanced by empowering individuals to succeed with achievable actions that progressively become challenging [22]. Thus, the Happy Hands app may have played a significant role in enhancing participants' self-efficacy, which is critical for driving sustained behavior change [16].

According to SCT, self-regulation is a key strategy for facilitating successful behavior change. It comprises several elements, including self-monitoring, goal setting, feedback, and self-reward, and often includes the possibility to observe and record one's behavior [22]. These elements were integrated into the app and were highlighted by participants as important in helping them adhere to the exercise program. Participants emphasized the value of tracking their progress by ticking off completed exercises, which fostered a sense of commitment. Moreover, having 3 designated exercise days per week provided structure, while app notifications served as helpful reminders to complete the exercises. This aligns with the goal-setting element of SCT, which suggests that setting short-term, achievable goals motivates effort and guides actions [16]. Furthermore, participants found motivation in the app's feedback features, such as being able to track their progress and answer quizzes. Feedback is recognized in SCT as a crucial component in facilitating behavior change, reinforcing effort, and sustaining motivation.

The app includes videos that demonstrate how to perform the exercises, which was appreciated by the participants. First, observing someone performing the exercises helped them understand how they should be performed correctly. Second, participants expressed that doing the exercises while watching the videos felt as though they were training alongside someone, which increased their motivation. This finding can be further explained through SCT, where observational learning is a key concept, suggesting that seeing others perform a specific behavior encouraged app users to replicate that behavior [22].

Most participants expressed a positive attitude toward the intervention delivered through the Happy Hands app. However, some challenges were reported. These included experiencing pain during or after exercising, and for some participants, their hand osteoarthritis seemed too severe for them to benefit from the exercises. Such individuals may require individual face-to-face guidance, either as a substitute for or in combination with digitally delivered interventions. While this raises concerns about the usability of digital interventions for all patients, it is important to recognize that widespread use of digital solutions could free up valuable health care resources for those who require more intensive follow-up care from health professionals.

People with hand osteoarthritis are a large patient group with limited treatment options. To ensure a sustainable health care service, the integration of digital solutions and technologies is essential. The Happy Hands app has the potential to provide people with hand osteoarthritis access to reliable information, guidance, and recommended treatment.

Strengths and Limitations

A key strength of this study is that participants were recruited from both primary and specialist care and from various regions across Norway, ensuring representation across a broad geographical area and levels of care. Furthermore, we collaborated with patient research partners with hand osteoarthritis who had used the Happy Hands app for 12 weeks. These partners provided feedback on the interview guide and participated in a meeting to discuss the study results. They expressed that they could personally relate to the findings. Respondent validation, that is, sharing results with people who have experienced the phenomena being studied, serves as an important strategy to help prevent misinterpretations [32].

The focus groups consisted of participants who shared their perceptions and experiences, with the interviews being transcribed verbatim. This approach facilitated the collection of rich data, characterized by depth and diversity, thus providing a comprehensive understanding of the issues being studied [32]. Another strength was the collaborative analysis process, where both the PhD candidate and supervisors read through transcripts, participated in coding, and engaged in discussions regarding the content in the interviews. Such discussions and feedback are beneficial for identifying potential flaws in logic or methodology, thereby enhancing the rigor of the research [32].

Our study also has some limitations. One is that participants voluntarily chose to participate in the Happy Hands study and agreed to be interviewed. This may have introduced a selection bias, as they might have had a more positive attitude toward using an app and were more receptive of digitally delivered interventions. This could also indicate that the participants in our study had a higher level of eHealth literacy compared to the general population. Individuals who did not use the app may have been more likely to decline participation in the focus group interviews. Future studies should aim to include their perspectives and experiences, as this could yield valuable insights into whether and how the app could be improved to meet their needs, as well as inform alternative approaches to care delivery.

While focus groups are useful for exploring collective experiences and generating rich, interactive data, they also have limitations. Group dynamics can influence what is shared, with dominant voices potentially influencing the discussion. Some individuals may hesitate to express dissenting or personal views, particularly when discussing sensitive topics. Additionally, the time constraints of group discussions may limit the depth of individual narratives [20].

Conclusions

The findings indicate that the Happy Hands app initiated and facilitated a process that resulted in behavior change among the participants. The app provided validation and acknowledgment of the disease. Furthermore, insights from the app contributed to a more informed understanding of hand osteoarthritis, its causes, and its consequences. Features of the app supported participants in learning hand exercises and new working methods. Together with altered illness perception, this enabled participants to develop new behaviors. For a few participants, challenges such as pain during or after exercise were noted, indicating that some individuals may require more individualized face-to-face follow-up.

XSL•FO
RenderX

Patients with hand osteoarthritis represent a large patient group with limited access to recommended treatment. There is, therefore, a need for a new model for hand osteoarthritis care. The findings of this study demonstrate that participants benefited from the Happy Hands app. The app proved to be a valuable tool in empowering participants to better self-manage their condition. The result of this study suggests that the app can serve as a component in a treatment pathway for people with hand osteoarthritis.

Acknowledgments

We would like to thank all the participants who took the time to participate in the focus groups. We also thank Christine Hillestad Hestevik for reviewing the manuscript and for providing valuable feedback. GPT UiO (OpenAI's GPT models within University of Oslo's privacy requirements) [33] was used to translate quotes from Norwegian to English. The authors carefully reviewed the quotes afterwards to ensure they were translated correctly. GPT UiO was also used to improve English language. All suggested changes were evaluated by authors.

Data Availability

The dataset generated in this study is not available due to ensure participants anonymity and to not reveal sensitive information.

Funding

The project is funded by Foundation Dam (2022/FO387170). The funder was not involved in the study.

Authors' Contributions

ATT and IK planned the study. KAAF, ATT, SN, TB and IK developed the interview guide. KAAF recruited participants and conducted the focus groups. ER and LO assisted during focus groups. KAAF, ATT and IK performed the main analysis. SN and TB provided feedback on the results. KAAF, ATT, ER, LO, SN, TB, RMK and IK provided feedback on drafts of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[DOCX File , 18 KB - jmir_v28i1e82773_app1.docx]

Multimedia Appendix 2
COREQ checklist.
[DOC File , 72 KB - jmir_v28i1e82773_app2.doc]

References

1. Qin J, Barbour KE, Murphy LB, Nelson AE, Schwartz TA, Helmick CG, et al. Lifetime Risk of Symptomatic Hand Osteoarthritis: The Johnston County Osteoarthritis Project. Arthritis Rheumatol 2017 Jun;69(6):1204-1212 [FREE Full text] [doi: 10.1002/art.40097] [Medline: 28470947]
2. Marshall M, Watt FE, Vincent TL, Dziedzic K. Hand osteoarthritis: clinical phenotypes, molecular mechanisms and disease management. Nat Rev Rheumatol 2018;14(11):641-656. [doi: 10.1038/s41584-018-0095-4] [Medline: 30305701]
3. Kloppenburg M, Kroon FP, Blanco FJ, Doherty M, Dziedzic KS, Greibrokk E, et al. 2018 update of the EULAR recommendations for the management of hand osteoarthritis. Ann Rheum Dis 2019;78(1):16-24. [doi: 10.1136/annrheumdis-2018-213826] [Medline: 30154087]
4. Dziedzic KS, Allen KD. Challenges and controversies of complex interventions in osteoarthritis management: recognizing inappropriate and discordant care. Rheumatology (Oxford) 2018;57(suppl_4):iv88-iv98 [FREE Full text] [doi: 10.1093/rheumatology/key062] [Medline: 29684219]
5. Gravås EMH, Tveter AT, Nossum R, Eide REM, Klokkeide Å, Matre KH, et al. Non-pharmacological treatment gap preceding surgical consultation in thumb carpometacarpal osteoarthritis - a cross-sectional study. BMC Musculoskelet Disord 2019;20(1):180 [FREE Full text] [doi: 10.1186/s12891-019-2567-3] [Medline: 31039774]
6. Kjeken I, Bordvik DH, Osteras N, Haugen IK, Aasness Fjeldstad K, Skaalvik I, et al. Efficacy and safety of non-pharmacological, pharmacological and surgical treatments for hand osteoarthritis in 2024: a systematic review. RMD Open 2025;11(1):e004963 [FREE Full text] [doi: 10.1136/rmdopen-2024-004963] [Medline: 39793978]
7. Pisters MF, Veenhof C, Schellevis FG, Twisk JWR, Dekker J, De Bakker DH. Exercise adherence improving long-term patient outcome in patients with osteoarthritis of the hip and/or knee. Arthritis Care Res (Hoboken) 2010;62(8):1087-1094 [FREE Full text] [doi: 10.1002/acr.20182] [Medline: 20235201]

XSL•FO
RenderX

8. Meld. St. 9 Nasjonal helse- og samhandlingsplan 2024–2027 - Vår felles helsetjeneste. Norwegian Ministry of Health and Care Services. 2024. URL: https://www.regjeringen.no/no/dokumenter/meld.-st.-9-20232024/id3027594/ [accessed 2025-04-01]
9. Walter MM, Sirard P, Nero H, Hörder H, Dahlberg LE, Tveter AT, et al. Digitally delivered education and exercises for patients with hand osteoarthritis-an observational study. Musculoskeletal Care 2023;21(4):1154-1160. [doi: 10.1002/msc.1796] [Medline: 37421256]
10. Rodríguez Sánchez-Laulhé P, Biscarri-Carbonero Á, Suero-Pineda A, Luque-Romero LG, Barrero García FJ, Blanquero J, et al. The effects of a mobile app-delivered intervention in people with symptomatic hand osteoarthritis: a pragmatic randomized controlled trial. Eur J Phys Rehabil Med 2023;59(1):54-64 [FREE Full text] [doi: 10.23736/S1973-9087.22.07744-9] [Medline: 36633498]
11. Cronström A, Dahlberg LE, Nero H, Ericson J, Hammarlund CS. 'I would never have done it if it hadn't been digital': a qualitative study on patients' experiences of a digital management programme for hip and knee osteoarthritis in Sweden. BMJ Open 2019;9(5):e028388 [FREE Full text] [doi: 10.1136/bmjopen-2018-028388] [Medline: 31129601]
12. Nelligan RK, Hinman RS, Teo PL, Bennell KL. Exploring attitudes and experiences of people with knee osteoarthritis toward a self-directed ehealth intervention to support exercise: qualitative study. JMIR Rehabil Assist Technol 2020;7(2):e18860 [FREE Full text] [doi: 10.2196/18860] [Medline: 33242021]
13. Ezzat AM, Bell E, Kemp JL, O'Halloran P, Russell T, Wallis J, et al. "Much better than I thought it was going to be": telehealth delivered group-based education and exercise was perceived as acceptable among people with knee osteoarthritis. Osteoarthr Cartil Open 2022;4(3):100271 [FREE Full text] [doi: 10.1016/j.ocarto.2022.100271] [Medline: 36474949]
14. Gell NM, Smith PA, Wingood M. Physical therapist and patient perspectives on mobile technology to support home exercise prescription for people with arthritis: a qualitative study. Cureus 2024;16(3):e55899 [FREE Full text] [doi: 10.7759/cureus.55899] [Medline: 38601402]
15. Tveter AT, Varsi C, Maarnes MK, Pedersen SJ, Christensen BS, Blanck TB, et al. Development of the happy hands self-management app for people with hand osteoarthritis: feasibility study. JMIR Form Res 2024;8:e59016 [FREE Full text] [doi: 10.2196/59016] [Medline: 39470716]
16. Bandura A. Health promotion by social cognitive means. Health Educ Behav 2004;31(2):143-164. [doi: 10.1177/1090198104263660] [Medline: 15090118]
17. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
18. Tveter AT, Fjeldstad KA, Varsi C, Maarnes MK, Pedersen SJ, Christensen BS, et al. Evaluation of an e-self-management intervention (Happy Hands app) for hand osteoarthritis: study protocol for a multicentre randomised controlled trial. Rheumatol Int 2025;45(1):30 [FREE Full text] [doi: 10.1007/s00296-025-05787-6] [Medline: 39820549]
19. Bowling A. Research Methods in Health: Investigating Health and Health Services. Maidenhead: McGraw-Hill Education; 2014.
20. Liamputtong P. Focus Group Methodology : Principles and Practice. London: SAGE; 2011.
21. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
22. McAlister AL, Perry CL, Parcel GS. How individuals, environments, and health behaviors interact: social cognitive theory. In: Glanz K, Rimer BK, Viswanath K, editors. Health Behavior and Health Education : Theory, Research, and Practice. San Francisco, CA: Jossey-Bass; 2008:169-188.
23. Leventhal H, Meyer D, Nerenz D. The common sense representation of illness danger. In: Rachman S, editor. Medical Psychology. Elmsford, NY: Pergamon Press; 1980:7-30.
24. Hagger MS, Orbell S. A meta-analytic review of the common-sense model of illness representations. Psychology Health 2003;18(2):141-184. [doi: 10.1080/0887044031000081321]
25. Braun V, Clarke V. Thematic Analysis: A Practical Guide. Los Angeles, CA: SAGE; 2022.
26. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
27. Hill S, Dziedzic KS, Nio Ong B. Patients' perceptions of the treatment and management of hand osteoarthritis: a focus group enquiry. Disabil Rehabil 2011;33(19-20):1866-1872. [doi: 10.3109/09638288.2010.550381] [Medline: 21859421]
28. Gignac MAM, Davis AM, Hawker G, Wright JG, Mahomed N, Fortin PR, et al. "What do you expect? You're just getting older": a comparison of perceived osteoarthritis-related and aging-related health experiences in middle- and older-age adults. Arthritis Rheum 2006;55(6):905-912. [doi: 10.1002/art.22338] [Medline: 17139636]
29. Magnussen HJ, Kjeken I, Pinxsterhuis I, Sjøvold TA, Hennig T, Thorsen E, et al. Participation in healthcare consultations: a qualitative study from the perspectives of persons diagnosed with hand osteoarthritis. Health Expect 2023;26(3):1276-1286 [FREE Full text] [doi: 10.1111/hex.13744] [Medline: 36916677]
30. Bukhave EB, Huniche L. Activity problems in everyday life--patients' perspectives of hand osteoarthritis: "try imagining what it would be like having no hands". Disabil Rehabil 2014;36(19):1636-1643. [doi: 10.3109/09638288.2013.863390] [Medline: 24308906]

XSL•FO
RenderX

31. Grime J, Richardson JC, Ong BN. Perceptions of joint pain and feeling well in older people who reported being healthy: a qualitative study. Br J Gen Pract 2010;60(577):597-603 [FREE Full text] [doi: 10.3399/bjgp10X515106] [Medline: 20822692]
32. Maxwell JA. Qualitative Research Design: An Interactive Approach. Los Angeles, CA: Sage; 2013.
33. GPT UiO-UiOs privacy friendly GPT chat. University of Oslo. 2025. URL: https://www.uio.no/english/services/it/ai/gpt-uio/ [accessed 2025-08-01]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
RCT: randomized controlled trial
SCT: social cognitive theory
USIT: The University Information Technology Center

Edited by A Stone; submitted 25.Aug.2025; peer-reviewed by W Ogundare, GA McHugh; comments to author 17.Oct.2025; accepted 24.Dec.2025; published 02.Feb.2026.

Please cite as:
Fjeldstad KA, Tveter AT, Rasmussen E, Olden L, Nyheim S, Blanck T, Killingmo RM, Kjeken I
Changing Habits With the Happy Hands App: Qualitative Focus Group Study of a Hand Osteoarthritis Self-Management Intervention
J Med Internet Res 2026;28:e82773
URL: https://www.jmir.org/2026/1/e82773
doi:10.2196/82773
PMID:

©Kristine Aasness Fjeldstad, Anne Therese Tveter, Eivor Rasmussen, Lena Olden, Sissel Nyheim, Thalita Blanck, Rikke Munk Killingmo, Ingvild Kjeken. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Facilitators and Barriers to Implementing Mobile Mental Health Interventions: Qualitative Study of the Consolidated Framework for Implementation Research in Pediatric Oncology Providers

Shannon J H Hong[1], BS; Michaela Patton[2], PhD; Krysta S Barton[3], PhD; Tonya M Palermo[2,4], PhD; Karen Mulholland[2]; Eric J Chow[5,6], MPH, MD; Nancy Lau[2,7], PhD

[1]Department of Psychology, University of California, Berkeley, Berkeley, CA, United States

[2]Center for Child Health, Behavior and Development, Seattle Children's Research Institute, Seattle, WA, United States

[3]Stanford Center for Biomedical Ethics, Stanford University School of Medicine, Stanford, CA, United States

[4]Department of Anesthesiology and Pain Medicine, University of Washington School of Medicine, Seattle, WA, United States

[5]Department of Pediatrics, University of Washington School of Medicine, Seattle, WA, United States

[6]Clinical Research and Public Health Sciences Divisions, Fred Hutchinson Cancer Center, Seattle, WA, United States

[7]Department of Psychiatry and Behavioral Sciences, University of Washington School of Medicine, Seattle, WA, United States

Corresponding Author:
Shannon J H Hong, BS
Department of Psychology
University of California, Berkeley
2121 Berkeley Way
Berkeley, CA, 94704
United States
Phone: 1 510 642 5292
Email: s.hong@berkeley.edu

Abstract

Background: Adolescent and young adult (AYA) cancer survivors experience unique psychosocial needs during and after treatment. Mobile health (mHealth) interventions are an emerging area of research to help address unmet psychosocial needs. However, few studies have examined provider perspectives on the design-to-implementation pipeline.

Objective: Guided by the Consolidated Framework for Implementation Research (CFIR), our study aimed to examine provider perspectives on facilitators and barriers to implementing mHealth apps in routine clinical care.

Methods: AYA oncology providers participated in a semistructured 1:1 interview on facilitators and barriers to incorporating mHealth apps as psychosocial standard of care. We conducted a directed content analysis of the interviews using a standardized CFIR codebook and construct definitions, with codebook adaptations for mHealth innovations and the population of AYAs with cancer.

Results: A total of 20 providers (mean 39, SD 7.0 years; 80% female and 70% non-Hispanic White) representing various medical and psychosocial roles participated in the interviews. The data were analyzed with 16 CFIR constructs. We identified the following facilitators to mHealth implementation across four CFIR domains: (1) Innovation: alignment with patient needs, patient-centered co-design, strong research evidence, and user-friendly design; (2) Outer Setting: shared commitment to addressing mental health needs and openness to mHealth use; (3) Inner Setting: openness to training on mHealth use; and (4) Individuals: engaging key implementation partners such as bedside nurses and social workers and strong clinical team buy-in. We identified the following barriers to mHealth implementation across three CFIR domains: (1) Innovation: associated costs, (2) Outer Setting: heavy clinical workloads, and (3) Inner Setting: lack of cross-team collaboration and communication and clinical workflow integration.

Conclusions: Our findings highlight key considerations for mHealth co-design, the adoption of mHealth apps into routine care, implementation strategies, and provider training opportunities in the context of AYA cancer care. Partnering with AYA cancer survivors, families, and providers will be crucial for developing and implementing mHealth apps with the ultimate goal of advancing universally accessible evidence-based digital health care.

(J Med Internet Res 2026;28:e87533) doi:10.2196/87533

XSL•FO
RenderX

KEYWORDS

adolescents; young adults; cancer; digital health; mHealth; mental health; mobile phone, psychosocial intervention; implementation science

Introduction

Approximately 80,000 adolescents and young adults (AYAs) are diagnosed with cancer annually in the United States [1]. Cancer disrupts key developmental milestones for AYAs, such as identity formation, transitioning to independence, relationship development, and achieving educational and vocational goals [2,3]. AYAs living with cancer and AYA cancer survivors (hereafter referred to as AYA cancer survivors) also have unique psychosocial needs that may be unmet throughout their cancer experience [4,5]. Consequently, AYA cancer survivors experience a myriad of mental health issues, including anxiety, depression, and distress, which often persist into long-term survivorship [6,7].

Evidence-based interventions for AYA cancer survivors are a new and emerging area of research [8-11]. The number of AYA-specific psychosocial intervention trials conducted annually has increased significantly from 2 during the years 2007-2014 to 11 during the years 2015-2021 [9]. This growing research activity highlights a commitment to developing evidence-based, developmentally tailored interventions that address the unique needs of this population [12]. As new interventions continue to emerge, leveraging technology has the potential to address barriers that AYAs face in accessing (eg, lack of service availability and financial challenges) and using (eg, time limitations and discomfort in sharing personal problems) face-to-face services [13,14]. As digital natives who have been immersed in technology from a young age, AYAs frequently use digital tools in their daily lives for communication, entertainment, and to access health information and support [15,16]. In recent years, there has been growing research on digital health interventions designed for AYA cancer survivors, with most studies focusing on symptom management, behavior change, self-care, and emotional health [17]. Digital health interventions have demonstrated feasibility and acceptability for AYA cancer survivors [18]. Compared with standard care, digital psychosocial interventions have been found to significantly reduce anxiety and depression and increase resilience in children and AYAs undergoing cancer treatment [10,19]. With 95% of teenagers having access to smartphones [20], mobile health (mHealth) interventions have the potential to expand evidence-based psychosocial care.

On average, it takes 17 years for research evidence to be translated into clinical practice, and only 1 in 5 evidence-based interventions are eventually implemented into routine clinical care [21-23]. While mHealth apps align with AYAs' digital preferences and hold promise in expanding access to psychosocial care, their integration into clinical care remains a significant research-to-practice gap. Although it is well recognized that successful mHealth implementation depends on the commitment and support of both patients and providers, little prior research has examined provider perspectives [24]. Recent implementation research studies have examined provider perspectives on mHealth tools such as symptom-tracking apps for patients, decision support apps for providers, and cancer triage systems for the general public [25-27]. However, studies have not yet examined provider perspectives on the implementation of psychosocial mHealth interventions. Providers are key stakeholders and partners whose insights are essential for understanding what influences the uptake and implementation of mHealth apps as standard of care within clinical settings.

To address these gaps, we conducted a directed content analysis of qualitative interview data from AYA oncology providers. The primary objectives of our study were to identify facilitators and barriers to implementing new and emerging mHealth apps into routine clinical care from the perspective of AYA oncology providers. Our analysis was guided by the Consolidated Framework for Implementation Research (CFIR), a widely used determinant framework to examine contextual factors influencing implementation effectiveness [28].

Methods

Ethical Considerations

This study was approved by the Seattle Children's Hospital (SCH) Institutional Review Board (number STUDY00004006). All participants provided written informed e-consent via REDCap (Research Electronic Data Capture; Vanderbilt University). Any and all quotes have been deidentified. Providers received a small token of appreciation (ie, water tumbler) for participation in the study. No other monetary incentives were provided.

Design, Setting, and Participants

Eligible participants were health care providers in the SCH Cancer and Blood Disorders Clinic. All patient-facing providers in the clinic, regardless of role, were eligible to participate in the study. This included psychosocial clinicians (social workers, therapists, and psychologists), oncologists (physicians and fellows), and nurses (nurse practitioners, inpatient, and outpatient) who would potentially integrate mHealth apps into their clinical workflows. Past studies have focused primarily on the perspectives of oncologists. In the interest of inclusivity, we used provider networks across medical and psychosocial staff to try to get equal representation across 6 different patient-facing roles (occupational therapist, pediatric oncologist, psychologist, nurse, social worker, or mental health therapist). We advertised the study on SCH oncology clinic listservs and newsletters and used snowball sampling to identify additional potential participants for recruitment. Participants were queried for convenient appointment times and scheduled for participation in a 1:1 interview. Before the interview, participants completed a brief demographics survey via REDCap.

Research Team

Authors' backgrounds included health services research (TMP and NL), implementation science (NL), digital health (SJHH, MP, TMP, and EJC), clinical and pediatric psychology (SJHH,

XSL•FO
RenderX

MP, TMP, and NL), pediatric oncology (EJC), psycho-oncology research (MP and NL), and qualitative research (KSB and NL). Authors have extensive background and experience in use of the CFIR framework and qualitative methods (SJHH, MP, TMP, KSB, and NL), which informed study design and research objectives. Authors also have extensive clinical experience with AYA cancer survivors (MP, EJC, and NL), which informed the interview guide, codebook development, and clinical implications drawn from the study.

Interviews

Semistructured 1:1 interviews were conducted with 20 health care providers via videoconference and were audio-recorded. This sample size was determined based on qualitative research guidelines recommending group sizes of 15-20 to achieve thematic saturation (no new themes or qualifying patterns) [29,30]. Interviews were deidentified and transcribed verbatim. A clinical psychologist trained in qualitative research methods (NL) conducted interviews from July 2024 to September 2024. Individual interviews were 30-50 minutes in length (average 40 minutes) and focused on facilitators and barriers to incorporating emerging mHealth apps as psychosocial standard of care. The interview script is shown in Multimedia Appendix 1. The interviewer memoed during the data collection process. Interviews were conducted at a single cross-sectional timepoint. We adapted a standardized CFIR interview guide for mHealth and the population of AYAs with cancer focused on the CFIR Innovation, Outer Setting, and Inner Setting domains [28].

Qualitative Analysis

We followed the SRQR (Standards for Reporting Qualitative Research) guidelines for complete reporting of qualitative research data collection and analysis procedures [31]. The completed SRQR checklist is provided in Multimedia Appendix 2. The overall analysis approach was a directed content analysis aimed at identifying facilitators and barriers to implementation of mHealth apps in routine care. We used a standardized CFIR codebook with CFIR domains and constructs, and CFIR definitions. We then developed coding definitions that were specific to mHealth innovations and the population of AYAs with cancer. Using this initial codebook, SJHH, MP, and NL independently coded the first 3 interview transcripts to determine relevance, validity, and fit of each relevant CFIR domain and construct. Over consecutive coding meetings, the analysis team (SJHH, MP, and NL) independently coded transcripts in sets of 3 and iteratively discussed and addressed codebook discrepancies and refined definitions until the codebook was finalized.

The final version of the codebook consisted of 16 of the 67 constructs from the CFIR that were most appropriate to digital mental health interventions. Standardized protocol for applying the CFIR framework is to select a small subset of relevant constructs and domains, specifically operationalized for each research project in accordance with the CFIR qualitative study design guidelines [32]. The most relevant 16 constructs operationalized for our study spanned across all 4 CFIR domains: Innovation (Source, Evidence-Base, Relative Advantage, Adaptability, Design, and Cost), Outer Setting (Local Attitudes and Local Conditions), Inner Setting (Information Technology Infrastructure, Work Infrastructure, Relational Connections, Communications, Recipient Centeredness, Compatibility, and Access to Knowledge and Information), and Individuals (Implementation Team Members). A primary coder (SJHH) and secondary coders (MP and NL) independently coded all transcripts, and all coders met twice monthly to discuss discrepancies and reach consensus. Interview transcripts were managed and analyzed using DeDoose software (SocioCultural Research Consultants).

The interviewer and coder (NL) had prior research relationships with some providers who participated in the study. The interview guide was designed so that questions and probes placed equal emphasis on facilitators and barriers, ensuring that participant feedback covered both domains. Research relationships and solicitations for feedback have historically fostered constructive dialogue, including weighing the pros and cons of different approaches. Dissenting opinions were shared openly and were explicitly encouraged. Reflexive memos and extensive audit trails documenting coding decisions were further used to enhance analytic rigor.

Results

Demographics

Participants were 20 oncology providers (mean 39, SD 7.0, range 29-56 years; Table 1). Most of the sample was female (80%) and self-identified as non-Hispanic White (85%). In terms of provider roles, 60% were medical providers and 40% were psychosocial providers. Provider demographics were representative of the overall staff within the clinic. The mean number of years participants had provided clinical care for oncology patients was 10 (SD 9.2, range 1-24) years. On average, participants reported providing care for 20 AYA patients per month (SD 21.0; range 4-100).

XSL•FO
RenderX

Table 1. Participant demographic characteristics (N=20).

Participant characteristics	Values, n (%)
Sex	
Female	16 (80)
Male	4 (20)
Race	
Asian	3 (15)
White	14 (70)
Multiracial	3 (15)
Ethnicity	
Hispanic or Latino	1 (5)
Not Hispanic or Latino	19 (95)
Role	
Occupational therapist	1 (5)
Pediatric oncologist	8 (40)
Psychologist	2 (10)
Nurse (registered nurse or advanced practice provider)	4 (20)
Mental health therapist	2 (10)
Social worker	3 (15)

CFIR Framework

We identified facilitators and barriers to mHealth implementation by using constructs across CFIR domains (Innovation, Outer Setting, Inner Setting, and Individuals) and indicated whether each CFIR construct was a facilitator or barrier using (+) or (–), respectively. Thematic saturation was reached with our study sample. A visual representation of the interrelationships among the Innovation, Outer Setting, Inner Setting, and Individuals domains is shown in Figure 1. We demarcated each CFIR construct as a facilitator or barrier to provide a holistic synthesis of our findings. All findings were relevant to AYA cancer survivors across the cancer care continuum from diagnosis to long-term survivorship. Exemplar quotes for each CFIR construct are shown in Table 2.

XSL•FO
RenderX

Figure 1. Identified Consolidated Framework for Implementation Research facilitators (+) and barriers (–) to implementing mHealth apps in pediatric oncology settings across the cancer care continuum.

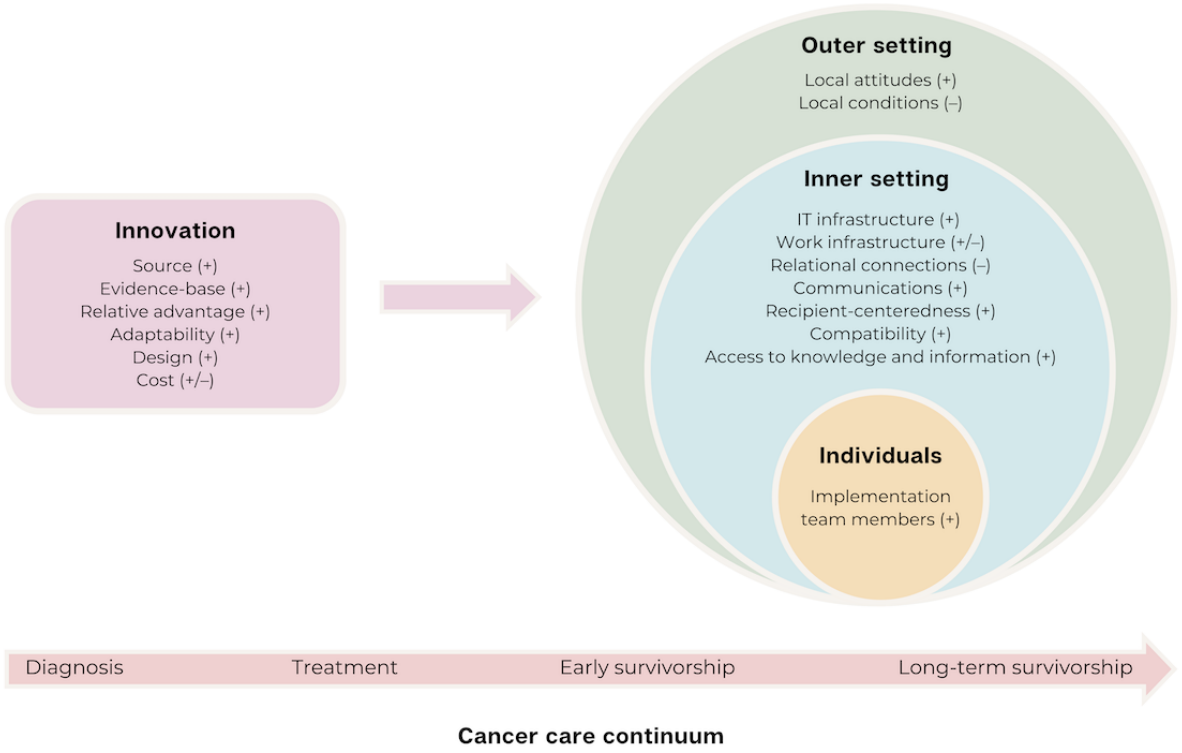

XSL•FO
RenderX

Table 2. Exemplar quotes of Consolidated Framework for Implementation Research constructs by domain.

Domain and construct	Representative Quotes
Innovation	
Source	• "As a professional recommending something to patients, I would want to do better than just generating the type of recommendation patients could find...just Googling. So I would want for someone with some expertise in the area to have evaluated the apps...Good resources would be things [that are] evidence based, sound, scientific foundations, have been used with [an] AYA[a] cancer population." [P9][b] • "I think about programs...that were born out of things that patients were involved in. I think about our AYA, advisory board. People get excited when our own patients are involved in the creation and implementation of things...if patients could take some ownership of this as well and be involved in the beginning stages." [P14]
Evidence-base	• "I would want a product that was vetted by psychologists and other clinicians and had some sort of proven efficacy...I just want to make sure that anything that I am recommending...is something that is going to provide some sort of medical benefit." [P1] • "A lot of it, of course, is dependent on how engaged the patient is...There are a lot of apps out there that are not evidence based, and so whether, you know, patients are being exposed to concepts or interventions that are anecdotally useful for people but not really evidence based." [P3]
Relative advantage	• "I think a digital app can be presented and interacted with much more frequently than a face-to-face interaction with a clinician. And you know, because of that, you can get maybe a bit more longitudinal interaction and tracking." [P1] • "We provide really good psychosocial care to our AYA population, which something like this would just enhance. But of the 80,000-ish AYAs that are diagnosed in the US annually...I think something like this app is particularly valuable for reaching those patients who are a sizable proportion of the AYA patient population nationally." [P9]
Adaptability	• "One of the things that I think is always beneficial are those med reminders [and] symptom management tools where they can document things. Just different tools to help...address that...developmental stage that they're in and...seeking out more of their independence, autonomy, all those different types of things...plays really deeply into their mental health." [P7] • "It's very scalable. It would allow AYAs to move through the material at their own pace, according to their own interests on their own time...It potentially could be delivered in a variety of languages with the content potentially tailored to a variety of cultures. So I think it's very scalable, which is important for potential use outside of just our patient population in our institution." [P9]
Design	• "It's my clinical experience that [AYAs] like to receive information by video rather than by spoken word or written word. They like content that is short snippets, rapidly paced, attention getting, highly personalized to them in terms of their interests, their issues, their questions...Something that feels more like that is probably going to be more acceptable to them, and so they'd probably be more likely to use it." [P9] • "I feel receptive to an app, especially if it's well designed and relatively intuitive." [P12]
Cost	• "I would consider cost and what kind of burdens [exist] for signing up. You know, everybody always has to create a username and all this stuff...but I would want a product that was really easy to sign up for and register and wasn't collecting tons of overly invasive stuff or posing extra burdens for an AYA to sign up for." [P1] • "I think the cost and availability of the app would definitely be...in the forefront. If a patient has to commit or get a subscription, I may be more hesitant to recommend that." [P4]
Outer setting	
Local attitudes	• "As I've discussed with other nurses and we do see a huge need for mental health support for our patient population and I think that the hospital and the [clinic] program have done a good job of trying to improve upon that and we've had more resources and more people join that team." [P17] • "A gap in our psychosocial services that maybe a tool could help us with is transition planning for survivors [with] assessments and then also like resources and interventions for survivors, I don't think we do a great job." [P12]
Local conditions	• "Nurses do feel like they're overburdened with tasks to complete throughout the day and so I do worry that some would and kind of cast it off to the side...It's just a reminder fatigue, I guess in that sense." [P17] • "One [barrier] is definitely the time and the fact that there's many different providers involved and things can fall through the cracks when they don't feel like the top priority to everyone involved, as there's a lot of things going on." [P4]
Inner setting	

XSL•FO
RenderX

Domain and construct	Representative Quotes
IT infrastructure	• "We have the Get Well Network which is part of our TV system...that included videos...about...what it means to have your counts dropped, what it means to get a central line, all of those things. And alerts pop up on the TV for families to watch those videos...So I'm not sure how you could embed it into that, but I think that could be really helpful to say "reminder, don't forget to do your [app]." [P17] • "Making it a standardized thing in our clinic, whether it's on the iPads that are in the rooms or...just having the visibility of...the [app] in our clinic, I think would be helpful, especially if it's proven to be effective. I think having easier access to it, within our clinical space and showing that the clinic backs it up, maybe that would help more patients download it and utilize it at home and take out that a little bit of that stigma piece to it." [P7] • "When you talk to someone about an app, it's harder to really understand what it looks like or what it entails...[Using TVs] could be a nice opportunity to reach families...while they're inpatient." [P18]
Work infrastructure	• "How can there be something that feels...automated and part of the system that then can be sort of supported for patients who have questions about it or who bring it up rather than relying on one person remembering to mention this at a visit or at a different time point." [P4] • "I'm learning a lot more in terms of being a supervisor and how to support patients and family or support my team...when you find the people who have the passion behind the work, things tend to happen more rather than just kind of sitting in this creative process." [P7]
Relational connections	• "There is a lack of collaboration between our social work and our embedded behavioral and mental health teams. And so, because of that, it's made a little bit challenging to understand who's doing what and how and where there's reason for overlap...and best practices of how things are delivered and who is delivering it or who's responsible for it." [P2] • "Factors include, I guess a lack of clarity with differences in clarity with roles and scope of practice. I really think it's like we're in such extreme silos, if there is just kind of this often this lack of understanding of what other services do in comparison to your own." [P14]
Communications	• "Half the battle is gonna be making people aware of [the apps] and buying into their benefit. I think any kind of 1 pager or a few minutes at a meeting, but that both talks about the benefits and the research behind why this is a good approach." [P5] • "The psychosocial group has tried to incorporate...a monthly email that comes out, that's just a reminder...And then I can refer back to the email and it has bullet points of how to refer and what the criteria are and the dates of the groups. And so something like that could also be helpful." [P4]
Recipient centeredness	• "I always think about independence and helping [AYAs] have some autonomy in their decision making aspects and or learning the skills so that when they reach the age 18 that it's not such a shock and preparing them for that transition. So I think helping them create a sense of identity or independence helps their mental health, and especially in this kind of this disease based populations, specifically when they rely so much on their caregivers and their parents or any other adults in their lives." [P7] • "We transition to patients about like a year and a half to two years after therapy to the survivorship for long term follow up team and that might be another point where [the resource] could be helpful...some patients find challenging reintegrating into school or thinking about the fact that scans are upcoming, and each time scans come up that can induce some anxiety. So maybe highlighting some things that other patients or families have cited as stressful and thinking about how these apps could support patients and sort of coping or having increased resilience through that." [P4] • "There's so much isolation...lack of autonomy, and a lot of issues with self-image that I just would want there to be some sort of other avenue for them to be able to engage in [apps] and be like 'Hey, you're not the only person who's feeling this' and 'Here's some other people that you can talk to about this.'" [P11]
Compatibility	• "For a lot of the patients that I meet with, they're sort of at the top of the pyramid in terms of their clinical need. Those types of tools could be especially helpful sort of at that universal tier where they're not necessarily connected to an individual mental health provider, but could get some support that way or do some more kind of self directed skills based learning." [P8] • "In a survivorship clinic visit...it's been my experience [that] almost universally the first six months off therapy are actually harder than on therapy cancer experience itself. So my thoughts are that almost immediately after that end of therapy visit and scans, that is where...these tools from digital health are potentially most critical and allow for kind of continued way to kind of check in and communicate back. And for the team to get an insight on what's going on such that when patients come back, they...haven't been suffering from mental health issues and stress and decreased resilience for three to six months by the time they come back to see us for scans." [P2]
Access to knowledge and information	

XSL•FO
RenderX

Domain and construct	Representative Quotes
	• "[What] would be the most helpful is if [innovation developers] came probably to one of our meetings and just presented the information to our team. And just kind of let us know what you need from us, and how we can best promote this to the patients. I think that would probably be the first step. I could bring it back to our team and then see if there's interest, which I'm assuming there would be. And then [innovation developers] could come to our team meeting, and we could talk about it as a as a group about how to go about it." [P15] • "I think better training on [app] functionality and kind of the interventions that they utilize. As well as I think more, kind of research data around their efficacy would be really helpful." [P7]
Individuals	
Implementation team members	• "Social workers for sure often have a really good finger on the pulse and you know, I also think potentially some bedside nursing members or nursing champions on the inpatient unit around that new diagnosis time period could be people who could potentially help in a way." [P1] • "I think people would be incredibly receptive. I know that my nursing colleagues at the bedside would be really excited by something like this. I think that my medical colleagues would also be, you know, social work and psychology, obviously. I think child life is another group who would be very excited. I think a group that we may not sometimes think of as being invested in the mental health of AYAs, but that truly are is PT and OT...we think of them obviously as being involved in promoting physical health, but working with them closely on the inpatient unit, so much of their work focuses on mental health...so much of what PT OT does is really promoting mental health that I think very group who would be very excited by this as well." [P9]

[a]AYA: adolescent and young adult.

[b]P: provider.

Innovation

Source (+)

Most providers emphasized the importance of mHealth apps being recommended by trusted sources such as colleagues on their clinical team or national medical specialty associations. Providers also highlighted the value of AYA involvement in developing these apps and incorporating their perspectives and feedback, as "people get excited when our own patients are involved in the creation and implementation of [programs and initiatives]" [P14].

Evidence-Base (+)

Many providers valued robust scientific evidence demonstrating the effectiveness of mHealth apps and being able to see whether apps would be most beneficial for AYA cancer survivors. One provider stated, "[It would] feel like a wasted opportunity to be able to introduce a useful tool if it's not grounded in any good evidence and theory" [P3].

Relative Advantage (+)

Providers reported several advantages to using mHealth apps compared with traditional clinical practice. Providers emphasized the developmental appropriateness of apps and stated that AYA cancer survivors can "engage within the app in a way that may actually speak their language a bit better than an oncologist sitting in front of them" and that can "facilitate them being able to express themselves in a different kind of way that is probably a little bit more fitting to them" [P3]. Providers also described how mHealth apps could expand access to psychosocial care for AYA cancer survivors following cancer treatment as they transitioned back to a new normal.

Adaptability (+)

All providers discussed the adaptability of mHealth apps. Given that AYAs have unique needs and interests, "an app that has really good uptake is [going to be] one that is able to address [what] an AYA is most interested in" [P1]. A common suggestion from providers for tailoring mHealth app content to AYAs' unique psychosocial needs was incorporating social connection features (eg, connecting with AYA peers with cancer).

Design (+)

Providers stated that mHealth apps should be well designed and intuitive to use. One provider stated, "I think it's a great way to get health content...to this population as long as A: the content is sound and B: the delivery is engaging to them. I think those are the two key ingredients" [P9]. Based on their experiences with AYA cancer survivors, providers suggested that content should be presented in an interactive and time-efficient format and be highly personalized to their interests. Some app features of interest include short informational videos, gamification elements (eg, leaderboards and streaks), and mood and symptom monitoring.

Cost (+/–)

Nearly half of the sample stated that they were concerned about cost and exclusively interested in recommending mHealth apps that were available for free. Providers emphasized the importance of affordability, with one provider stating, "I wouldn't recommend ones that are not free, especially if there were a large charge" [P20].

Outer Setting

Local Attitudes (+)

Providers shared a commitment to meeting AYA cancer survivors' mental health needs and were receptive to using mHealth apps to support clinical needs. One provider stated, "A lot of providers probably...sense the impact of their patients' mental health on care and may not have the tools...that may be able to help their patients in this way. So [apps] may actually help them feel sort of empowered to at least provide something

XSL•FO
RenderX

that the patient could use in the interim before meeting with a mental health provider" [P4].

Local Conditions (–)

Some providers identified challenges within the current conditions of the clinic that could hinder the implementation of mHealth apps. A common challenge, particularly among nurses and social workers, was the time constraints in their clinical workload. To address this, providers emphasized the need for systemic support of clinic-wide mHealth implementation efforts. One provider stated, "Some of the buy-in would also have to be systemic support...It would have to come through management to say we want to provide this opportunity for anyone that's interested to have this time carved out of their clinical day to put this work into it" [P19].

Inner Setting

IT Infrastructure (+)

Many providers recommended integrating information from mHealth apps (eg, data that show that the patient is improving or struggling) into electronic medical record systems. One provider stated, "Incorporating the results or critical information from that tool back into the environment and EMR...is a critical element for the clear communication and for longevity of the app" [P2].

Providers also suggested leveraging existing technological tools in patient rooms, such as iPads and TVs, to support adoption. For example, they described how educational videos and alerts are preloaded onto TVs for patients and their families. Similarly, TVs could be used to display tutorials of the functions and features of mHealth apps and reminders to download and use them.

Work Infrastructure (+/–)

Providers highlighted how mHealth apps could help with their limited bandwidth, as "having another tool that doesn't actually require really additional provider time that we may not have would be super useful" [P3]. However, they emphasized that tasks and responsibilities would need to be integrated or automated into their existing workflow to streamline the adoption of mHealth apps.

Relational Connections (–)

Providers noted operating in silos and a lack of awareness about what other clinical teams do and how they collaborate. One provider described having "a lack of clarity...with roles and scope of practice" and "how "we're in such extreme silos...[that] there is just kind of this lack of understanding of what other services do in comparison to your own" [P14]. As a result, it has been "challenging to understand who's doing what and how, where there's reason for overlap...and best practices of how things are delivered and who is delivering it or who's responsible for it" [P2].

Communications (+)

Although there was no consensus about the frequency of sharing information about mHealth apps among providers, they mentioned that a combination of face-to-face and written communication methods, such as email, one-pagers, and presence at team meetings, would be helpful. One provider stated, "Those Friday AM talks...have always been really well-received...I just feel more visibility about these things and the work that we're doing or that is being done to support our patients from the psychosocial mental health perspective...I feel like the providers are craving it and needing it...And I think it would actually create a lot of buy-in" [P7].

Recipient Centeredness (+)

Providers demonstrated shared values around supporting and tailoring mHealth apps to the needs and interests of AYA cancer survivors. A commonly perceived need was social connection. Providers also described how AYAs have previously engaged in gamified activities. For example, one provider shared that AYAs "would get so competitive and see that another teen had gotten a certain number of points and they wanted to try to beat them" [P20]. They suggested that a similar leaderboard feature could be incorporated into mHealth apps to enhance engagement.

Compatibility (+)

Providers stated that mHealth apps could "play a role across [the] treatment continuum," particularly for AYAs getting off treatment and into survivorship as "mental health concerns persist or can absolutely worsen during that time" [P8]. Providers stated that mHealth apps could be "a tool to use especially in moments of distress, which are common in survivorship" or when patients are "coming for scans or labs and have a really high increase of anxiety or stress" [P19]. AYA cancer survivors receiving active treatment also "have a lot of downtime" and mHealth apps would be "a really great way to...harness some of that when they're between appointments or waiting after labs and before the provider comes in the clinic" [P5]. Providers also described how mHealth apps could be universally used with all AYA cancer survivors, which would "be less of a strain on...the more limited resources that we have for individual patients" [P8].

Access to Knowledge and Information (+)

Providers expressed a need for guidance on the benefits of mHealth apps and how to promote them to AYA cancer survivors. One provider stated that "very basic education about...what it's for and why it's important" would be needed so that providers "feel engaged and invested" [P4].

Individuals: Implementation Team Members (+)

Providers recommended bedside nurses and social workers as the clinical team members who would need to be involved as the most ideal potential clinic partners for the successful implementation of mHealth apps. Providers also stated that buy-in and strong engagement from clinical team members would be needed. One provider stated, "You could have created the best thing ever and it won't make a difference. What you need is a boots-on-the-ground clinic champion to make an impact at the bedside" [P9].

XSL•FO
RenderX

Discussion

Principal Findings

To our knowledge, this is the first study to examine facilitators and barriers to implementing mHealth apps in routine clinical care from the perspective of AYA oncology providers. Constructs across the CFIR Innovation, Outer Setting, Inner Setting, and Individuals domains were identified as facilitators and barriers to mHealth implementation. Our study addresses a key gap in the design-to-implementation pipeline of emerging mHealth interventions.

In terms of facilitators to mHealth implementation, a central theme was strong alignment with patient needs. Providers were interested in collaborating on the development of mHealth apps and emphasized that they should be co-designed with AYAs. Strong research evidence and user-friendly design were other key facilitators, with providers stressing that mHealth apps needed to be both effective and engaging for AYA cancer survivors. Providers viewed mental health care as a valuable component of care in oncology and felt optimistic about the potential of mHealth apps to improve patient outcomes and advance mental health equity. They also expressed openness to receiving guidance and training on using and recommending mHealth apps to AYA cancer survivors. Providers identified bedside nurses and social workers as key implementation partners and highlighted the importance of strong clinical team engagement and buy-in.

Providers also identified several implementation barriers. Cost and accessibility were common concerns, with providers expressing hesitancy in recommending subscription-based or high-fee apps. Usability was another common barrier—providers emphasized that AYA cancer survivors would not want to use mHealth apps with a slow user interface. In terms of organizational barriers, providers reported that heavy clinical workloads and limited cross-team collaboration would make it challenging to implement mHealth apps. Providers further acknowledged the potential challenges of integrating these practices into existing clinic workflows, often referencing the high task demands and the time burden associated with supporting mHealth app implementation.

Comparison With Prior Work

Our findings align with prior research with health care providers who have found positive attitudes and receptivity toward mHealth apps across various health care conditions and disease management functions. A growing body of research also supports the use of mHealth apps in oncology care, which provide opportunities to address patient needs and help them feel engaged throughout treatment [24]. However, to our knowledge, this is the first qualitative study to examine oncology provider perspectives on mHealth implementation. It is also the first to focus specifically on the AYA age group and addressing psychosocial needs in this population. Previous studies that examined provider perspectives have focused on co-designing mHealth interventions with patients and families. Features such as social connection, symptom tracking, and appointment management have emerged as important considerations for these interventions [33,34]. With regard to facilitators of mHealth adoption, our findings align with prior research that has found that providers emphasize the need for evidence-based content and user-friendly design [35]. Our findings support previous recommendations to involve both patients and providers throughout the app development process, ensuring a balance of user and clinical expertise [36,37]. Beyond intervention design, providers offered insights on the integration of mHealth tools and interventions into clinical workflows through electronic health records and existing electronic systems [38]. Providers also shared suggestions for establishing partnerships with boots-on-the-ground clinical champions for large-scale adoptions.

With regard to barriers, our findings align with existing literature on concerns about data privacy, provider liability, and time burden associated with adopting new digital tools [24,34,39]. Our study extends this work by identifying previously unreported barriers, namely, fragmented communication between clinical teams and unclear ownership of implementation tasks. Our findings also highlight opportunities to integrate both medical and psychosocial provider roles into the implementation process and the use of mHealth tools across the cancer care continuum. This includes using mHealth tools during inpatient or outpatient care and throughout the active treatment and posttreatment phases. Additionally, a qualitative review conducted by Jacob et al [40] found that only 3 out of 50 studies examining clinician adoption of mHealth tools used the CFIR framework. This highlights the added value of our study in applying a rigorous implementation science approach to understanding factors that enable and hinder mHealth implementation. While the studies that used the CFIR framework primarily identified structural and organizational factors influencing implementation [41-43], our findings uniquely underscore the value of examining team-level processes when implementing mHealth tools in complex, multidisciplinary care environments.

Limitations

The findings of our study should be interpreted in the context of several limitations. First, our study was conducted at a single academic hospital, which may limit the generalizability of our results. Providers at this institution may have more experience with or interest in digital health than providers in other clinical settings. Second, our sample lacked demographic diversity, with most participants identifying as female and non-Hispanic White. Although this reflects the provider population in the geographic location in which our study was conducted, it may limit the applicability of our findings to more diverse clinical settings. Third, we intentionally predefined 16 CFIR constructs that were applicable to facilitators and barriers within our settings. With the CFIR having 67 constructs total, future research may explore additional factors that could impact implementation, such as policies, partnerships, or external pressures. Fourth, implementation contexts are dynamic. Our interviews captured the perspectives of providers on mHealth implementation as of 2024.

Conclusions

Our findings synthesized key implementation considerations for incorporating mHealth apps as a standard of care in

XSL•FO
RenderX

psychosocial support for AYA cancer survivors. Based on these findings, recommendations for practice include co-designing mHealth apps with both patients and providers to ensure alignment with user preferences and clinical needs. These tools should then be embedded within existing clinical workflows such as electronic health records to reduce provider burden and ensure seamless integration. Improvements in cross-team collaboration and communication can also be achieved by clearly delineating roles and responsibilities across clinical teams. Additionally, concerns around data privacy, provider liability, usability, and cost should be addressed to support the sustained use and implementation of mHealth apps. Implementation strategies should also include provider training on the benefits of mHealth apps and best practices for recommending these tools to AYA cancer survivors.

Our findings have important clinical implications. Strong partnerships among AYA cancer survivors, families, and providers will be essential for developing mHealth tools that are patient-centered and evidence-based [44]. Importantly, providers of all represented clinical roles believed that mHealth tools had clinical utility and wanted to be involved in the innovation development through implementation pipeline. They were also ready and willing to partner with researchers in early adoption efforts and had suggestions on how to do so. Future research should focus not only on evaluating the effectiveness of mHealth apps but also on implementing strategies that integrate into clinical workflows and minimize provider burden [45,46]. Understanding how these strategies impact provider uptake and patient outcomes will be crucial to moving toward universal evidence-based digital health care using innovative digital health technologies.

Acknowledgments

The authors thank the providers who participated in the study and generously shared their insights.

Funding

This work is supported by the National Cancer Institute at the National Institutes of Health (grant 1K08CA263474). The opinions herein represent those of the authors and not necessarily the funders.

Data Availability

The data used in the research cannot be publicly posted and cannot be shared privately with any person because participants did not consent to data sharing. Any shared data, even if deidentified, may result in breach of confidentiality.

Authors' Contributions

Conceptualization was done by SJHH, KSB, and NL. Data were curated by SJHH, KM, and NL. Formal analysis was handled by SJHH, MP, and NL. Investigation was done by SJHH, KSB, TMP, EJC, and NL. Methodology was handled by SJHH, KSB, and NL. Project administration and supervision were done by NL. Visualization was done by SJHH and KM. Writing of the original draft was done by SJHH. Review and editing of the writing were done by SJHH, MP, KSB, TMP, KM, EJC, and NL.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Provider interview guide.
[DOCX File , 18 KB - jmir_v28i1e87533_app1.docx]

Multimedia Appendix 2
SRQR checklist.
[DOCX File , 17 KB - jmir_v28i1e87533_app2.docx]

References

1. US cancer statistics data visualizations tool. US Cancer Statistics Working Group. 2024. URL: https://www.cdc.gov/united-states-cancer-statistics/dataviz/index.html [accessed 2025-05-30]
2. Bibby H, White V, Thompson K, Anazodo A. What are the unmet needs and care experiences of adolescents and young adults with cancer? A systematic review. J Adolesc Young Adult Oncol 2017;6(1):6-30. [doi: 10.1089/jayao.2016.0012] [Medline: 27454408]
3. Itzep N, Roth M. Psychosocial distress due to interference of normal developmental milestones in AYAs with cancer. Children (Basel) 2022;9(3):309 [FREE Full text] [doi: 10.3390/children9030309] [Medline: 35327680]

XSL•FO
RenderX

4. Zebrack BJ, Block R, Hayes-Lattin B, Embry L, Aguilar C, Meeske KA, et al. Psychosocial service use and unmet need among recently diagnosed adolescent and young adult cancer patients. Cancer 2013;119(1):201-214 [FREE Full text] [doi: 10.1002/cncr.27713] [Medline: 22744865]
5. Avutu V, Lynch KA, Barnett ME, Vera JA, Glade Bender JL, Tap WD, et al. Psychosocial needs and preferences for care among adolescent and young adult cancer patients (ages 15-39): a qualitative study. Cancers (Basel) 2022;14(3):710 [FREE Full text] [doi: 10.3390/cancers14030710] [Medline: 35158975]
6. Hall AE, Boyes AW, Bowman J, Walsh RA, James EL, Girgis A. Young adult cancer survivors' psychosocial well-being: a cross-sectional study assessing quality of life, unmet needs, and health behaviors. Support Care Cancer 2012;20(6):1333-1341. [doi: 10.1007/s00520-011-1221-x] [Medline: 21720746]
7. Barnett M, McDonnell G, DeRosa A, Schuler T, Philip E, Peterson L, et al. Psychosocial outcomes and interventions among cancer survivors diagnosed during adolescence and young adulthood (AYA): a systematic review. J Cancer Surviv 2016;10(5):814-831 [FREE Full text] [doi: 10.1007/s11764-016-0527-6] [Medline: 26920873]
8. Rosenberg AR, Bradford MC, McCauley E, Curtis JR, Wolfe J, Baker KS, et al. Promoting resilience in adolescents and young adults with cancer: results from the PRISM randomized controlled trial. Cancer 2018;124(19):3909-3917 [FREE Full text] [doi: 10.1002/cncr.31666] [Medline: 30230531]
9. Murphy KM, Siembida E, Lau N, Berkman A, Roth M, Salsman JM. A systematic review of health-related quality of life outcomes in psychosocial intervention trials for adolescent and young adult cancer survivors. Crit Rev Oncol Hematol 2023;188:104045. [doi: 10.1016/j.critrevonc.2023.104045] [Medline: 37269881]
10. Salsman JM, McLouth LE, Tooze JA, Little-Greene D, Cohn M, Kehoe MS, et al. An eHealth, positive emotion skills intervention for enhancing psychological well-being in young adult cancer survivors: results from a multi-site, pilot feasibility trial. Int J Behav Med 2023;30(5):639-650. [doi: 10.1007/s12529-023-10162-5] [Medline: 36890329]
11. Salsman JM, Rosenberg AR. Fostering resilience in adolescence and young adulthood: considerations for evidence-based, patient-centered oncology care. Cancer 2024;130(7):1031-1040 [FREE Full text] [doi: 10.1002/cncr.35182] [Medline: 38163249]
12. Thornton CP, Ruble K, Kozachik S. Psychosocial interventions for adolescents and young adults with cancer: an integrative review. J Pediatr Oncol Nurs 2020;37(6):408-422. [doi: 10.1177/1043454220919713] [Medline: 32452711]
13. Holland LR, Walker R, Henney R, Cashion CE, Bradford NK. Adolescents and young adults with cancer: barriers in access to psychosocial support. J Adolesc Young Adult Oncol 2021;10(1):46-55. [doi: 10.1089/jayao.2020.0027] [Medline: 32486879]
14. Lau N, Parsa AG, Walsh C, Yi-Frazier JP, Weiner BJ, Curtis JR, et al. Facilitators and barriers to utilization of psychosocial care in adolescents and young adults with advanced cancer: integrating mobile health perspectives. J Adolesc Young Adult Oncol 2021;10(4):476-482 [FREE Full text] [doi: 10.1089/jayao.2020.0129] [Medline: 32985937]
15. van de Graaf DL, Vlooswijk C, Bol N, Krahmer EJ, Bijlsma R, Kaal S, et al. AYAs' online information and eHealth needs: a comparison with healthcare professionals' perceptions. Cancer Med 2023;12(2):2016-2026 [FREE Full text] [doi: 10.1002/cam4.5048] [Medline: 35879825]
16. Abrol E, Groszmann M, Pitman A, Hough R, Taylor RM, Aref-Adib G. Exploring the digital technology preferences of teenagers and young adults (TYA) with cancer and survivors: a cross-sectional service evaluation questionnaire. J Cancer Surviv 2017;11(6):670-682 [FREE Full text] [doi: 10.1007/s11764-017-0618-z] [Medline: 28634734]
17. McCann L, McMillan KA, Pugh G. Digital interventions to support adolescents and young adults with cancer: systematic review. JMIR Cancer 2019;5(2):e12071 [FREE Full text] [doi: 10.2196/12071] [Medline: 31368438]
18. Devine KA, Viola AS, Coups EJ, Wu YP. Digital health interventions for adolescent and young adult cancer survivors. JCO Clin Cancer Inform 2018;2:1-15 [FREE Full text] [doi: 10.1200/CCI.17.00138] [Medline: 30652583]
19. Ozturk CS, Toruner EK. Effectiveness of technology-based psychosocial interventions for child, adolescents and young adults undergoing cancer treatment: a meta-analysis of randomised controlled trials. Eur J Cancer Care (Engl) 2022;31(1):e13515. [doi: 10.1111/ecc.13515] [Medline: 34632662]
20. Vogels EA, Gelles-Watnick R, Massarat N. Teens, social media and technology 2022. Pew Research Center. 2022. URL: https://www.pewresearch.org/internet/2022/08/10/teens-social-media-and-technology-2022/ [accessed 2025-06-13]
21. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. J R Soc Med 2011;104(12):510-520 [FREE Full text] [doi: 10.1258/jrsm.2011.110180] [Medline: 22179294]
22. Kilbourne AM, Glasgow RE, Chambers DA. What can implementation science do for you? Key success stories from the field. J Gen Intern Med 2020;35(Suppl 2):783-787 [FREE Full text] [doi: 10.1007/s11606-020-06174-6] [Medline: 33107001]
23. Rubin R. It takes an average of 17 years for evidence to change practice-the burgeoning field of implementation science seeks to speed things up. JAMA 2023;329(16):1333-1336. [doi: 10.1001/jama.2023.4387] [Medline: 37018006]
24. Berkowitz CM, Zullig LL, Koontz BF, Smith SK. Prescribing an App? Oncology providers' views on mobile health apps for cancer care. JCO Clin Cancer Inform 2017;1:1-7. [doi: 10.1200/CCI.17.00107] [Medline: 30657404]
25. Gregory ME, Cao W, Rahurkar S, Jonnalagadda P, Stock JC, Ghazi SM, et al. Exploring the incorporation of a novel cardiotoxicity mobile health app into care of patients with cancer: qualitative study of patient and provider perspectives. JMIR Cancer 2023;9:e46481 [FREE Full text] [doi: 10.2196/46481] [Medline: 38085565]

XSL•FO
RenderX

26. Jacob C, Sanchez-Vazquez A, Ivory C. Clinicians' role in the adoption of an oncology decision support app in europe and its implications for organizational practices: qualitative case study. JMIR Mhealth Uhealth 2019;7(5):e13555 [FREE Full text] [doi: 10.2196/13555] [Medline: 31066710]
27. Sangers TE, Wakkee M, Moolenburgh F, Nijsten T, Lugtenberg M. Mobile health apps for skin cancer triage in the general population: a qualitative study on healthcare providers' perspectives. BMC Cancer 2025;25(1):851 [FREE Full text] [doi: 10.1186/s12885-025-14244-3] [Medline: 40346534]
28. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated consolidated framework for implementation research based on user feedback. Implement Sci 2022;17(1):75 [FREE Full text] [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
29. Creswell JW. Qualitative Inquiry and Research Design: Choosing Among Five Approaches. Cham: SAGE; 2013.
30. Moser A, Korstjens I. Series: practical guidance to qualitative research. Part 3: sampling, data collection and analysis. Eur J Gen Pract 2018;24(1):9-18 [FREE Full text] [doi: 10.1080/13814788.2017.1375091] [Medline: 29199486]
31. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
32. Reardon CM, Damschroder LJ, Ashcraft LE, Kerins C, Bachrach RL, Nevedal AL, et al. The Consolidated Framework for Implementation Research (CFIR) user guide: a five-step guide for conducting implementation research using the framework. Implement Sci 2025;20(1):39 [FREE Full text] [doi: 10.1186/s13012-025-01450-7] [Medline: 40819067]
33. Hou SHJ, Henry B, Drummond R, Forbes C, Mendonça K, Wright H, et al. Co-designing priority components of an mhealth intervention to enhance follow-up care in young adult survivors of childhood cancer and health care providers: qualitative descriptive study. JMIR Cancer 2025;11:e57834 [FREE Full text] [doi: 10.2196/57834] [Medline: 40279633]
34. Jibb LA, Sivaratnam S, Hashemi E, Chu CH, Nathan PC, Chartrand J, et al. Parent and clinician perceptions and recommendations on a pediatric cancer pain management app: a qualitative co-design study. PLOS Digit Health 2023;2(11):e0000169 [FREE Full text] [doi: 10.1371/journal.pdig.0000169] [Medline: 38019890]
35. Ardito V, Golubev G, Ciani O, Tarricone R. Evaluating barriers and facilitators to the uptake of mHealth apps in cancer care using the consolidated framework for implementation research: scoping literature review. JMIR Cancer 2023;9:e42092 [FREE Full text] [doi: 10.2196/42092] [Medline: 36995750]
36. Hartzler A, Pratt W. Managing the personal side of health: how patient expertise differs from the expertise of clinicians. J Med Internet Res 2011;13(3):e62 [FREE Full text] [doi: 10.2196/jmir.1728] [Medline: 21846635]
37. Tarricone R, Cucciniello M, Armeni P, Petracca F, Desouza KC, Hall LK, et al. Mobile health divide between clinicians and patients in cancer care: results from a cross-sectional international survey. JMIR Mhealth Uhealth 2019;7(9):e13584 [FREE Full text] [doi: 10.2196/13584] [Medline: 31493318]
38. Upadhyay VA, Landman AB, Hassett MJ. Landscape analysis of oncology mobile health applications. JCO Clin Cancer Inform 2021;5:579-587. [doi: 10.1200/CCI.20.00156] [Medline: 34033510]
39. Kessel KA, Vogel MME, Schmidt-Graf F, Combs SE. Mobile apps in oncology: a survey on health care professionals' attitude toward telemedicine, mHealth, and oncological apps. J Med Internet Res 2016;18(11):e312 [FREE Full text] [doi: 10.2196/jmir.6399] [Medline: 27884810]
40. Jacob C, Sanchez-Vazquez A, Ivory C. Understanding clinicians' adoption of mobile health tools: a qualitative review of the most used frameworks. JMIR Mhealth Uhealth 2020;8(7):e18072 [FREE Full text] [doi: 10.2196/18072] [Medline: 32442132]
41. Lord S, Moore SK, Ramsey A, Dinauer S, Johnson K. Implementation of a substance use recovery support mobile phone app in community settings: qualitative study of clinician and staff perspectives of facilitators and barriers. JMIR Ment Health 2016;3(2):e24 [FREE Full text] [doi: 10.2196/mental.4927] [Medline: 27352884]
42. Possemato K, Kuhn E, Johnson EM, Hoffman JE, Brooks E. Development and refinement of a clinician intervention to facilitate primary care patient use of the PTSD Coach app. Transl Behav Med 2017;7(1):116-126 [FREE Full text] [doi: 10.1007/s13142-016-0393-9] [Medline: 27234150]
43. Varsi C, Ekstedt M, Gammon D, Ruland CM. Using the consolidated framework for implementation research to identify barriers and facilitators for the implementation of an internet-based patient-provider communication service in five settings: a qualitative study. J Med Internet Res 2015;17(11):e262 [FREE Full text] [doi: 10.2196/jmir.5091] [Medline: 26582138]
44. Abernethy A, Adams L, Barrett M, Bechtel C, Brennan P, Butte A, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022;2022 [FREE Full text] [doi: 10.31478/202206e] [Medline: 36177208]
45. Ramsey WA, Heidelberg RE, Gilbert AM, Heneghan MB, Badawy SM, Alberts NM. eHealth and mHealth interventions in pediatric cancer: a systematic review of interventions across the cancer continuum. Psychooncology 2020;29(1):17-37. [doi: 10.1002/pon.5280] [Medline: 31692183]
46. Parikh RB, Basen-Enquist KM, Bradley C, Estrin D, Levy M, Lichtenfeld JL, et al. Digital health applications in oncology: an opportunity to seize. J Natl Cancer Inst 2022;114(10):1338-1339 [FREE Full text] [doi: 10.1093/jnci/djac108] [Medline: 35640986]

XSL•FO
RenderX

Abbreviations

AYA: adolescent and young adult
CFIR: Consolidated Framework for Implementation Research
mHealth: mobile health
REDCap: Research Electronic Data Capture
SCH: Seattle Children's Hospital
SRQR: Standards for Reporting Qualitative Research

Edited by A Stone; submitted 10.Nov.2025; peer-reviewed by T Jimoh, WNA Wan Adnan; comments to author 04.Dec.2025; revised version received 07.Jan.2026; accepted 14.Jan.2026; published 23.Feb.2026.

Please cite as:
Hong SJH, Patton M, Barton KS, Palermo TM, Mulholland K, Chow EJ, Lau N
Facilitators and Barriers to Implementing Mobile Mental Health Interventions: Qualitative Study of the Consolidated Framework for Implementation Research in Pediatric Oncology Providers
J Med Internet Res 2026;28:e87533
URL: https://www.jmir.org/2026/1/e87533
doi:10.2196/87533
PMID:

©Shannon J H Hong, Michaela Patton, Krysta S Barton, Tonya M Palermo, Karen Mulholland, Eric J Chow, Nancy Lau. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Mobile Health App Attitudes and Adoption Among Oncology Providers: Cross-Sectional National Survey

Nancy Lau[1,2], PhD; Kavin Srinakarin[1], BS; Shannon JH Hong[3], BS; Homer Aalfs[1], BS; Michael E Roth[4], MD; Karly M Ingram[5], PhD; Amy Berkman[6], MD; Sherilynn Chan[1,2,8], PhD; Joanna Patten[7], PsyD; Wade Iwata[8], LICSW; Mallory R Taylor[9,10], MD, MS; Jesse R Fann[2], MD, MPH; Eric J Chow[09,10,11], MD, MPH; Tonya M Palermo[1,12], PhD

[1]Center for Child Health, Behavior and Development, Seattle Children's Research Institute, Seattle, WA, United States

[2]Department of Psychiatry and Behavioral Sciences, University of Washington School of Medicine, Seattle, WA, United States

[3]Department of Psychology, University of California, Berkeley, Berkeley, CA, United States

[4]Department of Pediatrics, MD Anderson Cancer Center, Houston, TX, United States

[5]Department of Psychology, East Carolina University, Greenville, NC, United States

[6]Department of Epidemiology and Cancer Control, St Jude Children's Research Hospital, Memphis, TN, United States

[7]Division of Psychiatry and Behavioral Medicine, Seattle Children's Hospital, Seattle, WA, United States

[8]Cancer and Blood Disorders Center, Seattle Children's Hospital, Seattle, WA, United States

[9]Ben Towne Center for Childhood Cancer and Blood Disorders Research, Seattle Children's Research Institute, Seattle, WA, United States

[10]Department of Pediatrics, University of Washington School of Medicine, Seattle, WA, United States

[11]Clinical Research and Public Health Sciences Divisions, Fred Hutch Cancer Center, Seattle, WA, United States

[12]Department of Anesthesiology & Pain Medicine, University of Washington School of Medicine, Seattle, WA, United States

Corresponding Author:
Nancy Lau, PhD
Center for Child Health, Behavior and Development
Seattle Children's Research Institute
1920 Terry Ave
Seattle, WA, 98101
United States
Phone: 1 206 884 8238
Email: nancy.lau@seattlechildrens.org

Abstract

Background: Mobile health (mHealth) apps can address health inequities and enhance access to care for individuals with immunocompromising conditions. Although hundreds of oncology apps exist, research on provider perspectives regarding their use in clinical care remains limited.

Objective: This study aimed to describe oncology providers' recommended apps, mHealth attitudes and beliefs, and perceived barriers to and facilitators of mHealth adoption. Exploratory aims examined differences based on provider type (medical vs psychosocial), provider age (<45 vs ≥45 years old), and patient population (pediatric vs adult).

Methods: We conducted a cross-sectional survey administered via REDCap (Research Electronic Data Capture) to oncology providers across the United States between June and November 2024. Data were summarized using descriptive statistics. Pearson's chi-square analyses examined exploratory group differences based on provider type, provider age, and patient age.

Results: Of 188 respondents, the majority self-identified as female (150/188, 79.8%), White (161/188, 85.6%), and non-Hispanic/Latino (174/188, 92.6%). Nearly all providers (178/188, 94.7%) reported either recommending or using mHealth apps with their patients, with primary use for patient-provider communication (139/188, 73.9%). Providers perceived potential benefit across a broad spectrum of holistic care functions. Providers, on average, reported a growth mindset and confidence in their ability to learn mHealth tools and in its potential to improve care access. Key facilitators included alignment with patient needs, increased accessibility, and cost-effectiveness, while barriers included disparities in technology access, digital health literacy, and data security and privacy. Exploratory analyses showed some significant group differences by provider role, provider age, and patient age. Psychosocial providers were significantly more likely to recommend or use apps for pain management (χ^2_1=14.34, P<.001, ϕ=0.28), mental health (χ^2_1=50.54, P<.001, ϕ=0.53), and sleep health (χ^2_1=25.47, P<.001, ϕ= 0.38).

XSL•FO
RenderX

Psychosocial providers also perceived higher benefit for sleep health apps (χ^2_1=6.40, *P*=.01, ϕ=0.19). Medical providers were significantly more likely to perceive medication management apps as potentially beneficial (χ^2_1=10.93, *P*<.001, ϕ=0.25). Older providers (16/88, 18.2%) and adult care providers (8/32, 25%) were significantly more likely to recommend or use disease management apps compared to younger providers (5/100, 5%; χ^2_1=8.20, *P*=.004, ϕ=0.21) and pediatric care providers (6/101, 5.9%; χ^2_2=9.22, *P*=.01, Cramer V=0.22), respectively. Pediatric care providers (83/101, 82.2%) were more likely to recommend or use medical team communication apps compared to adult care providers (15/32, 46.9%; χ^2_2=15.66, *P*<.001, Cramer V=0.29).

Conclusions: Our study underscores the opportunity to develop inclusive mHealth solutions tailored to the diverse needs of individuals across the cancer care continuum, including those in active treatment and survivorship care. Engaging diverse medical and psychosocial providers is essential to inform clinical integration of mHealth technologies in oncology care.

(J Med Internet Res 2026;28:e85583) doi:10.2196/85583

KEYWORDS

digital health technology; digital health technologies; mobile application; mobile app; mobile health; mHealth; health personnel; oncology, psycho-oncology; delivery of health care; health care delivery

Introduction

Over the past decade, the worldwide adoption and use of smartphone technologies has grown exponentially. The total number of smartphone users exceeds 6.7 billion worldwide [1]. Smartphones have become ubiquitous, regardless of gender, race, ethnicity, level of education, and socioeconomic status [2-5]. Over 350,000 mobile health (mHealth) apps (software apps downloadable on mobile devices) are available for public download on Google Play and Apple App Store [6]. mHealth apps have the potential to reduce health disparities for underresourced populations and to improve accessibility for those with immunocompromised or debilitating health conditions [7-15] .mHealth apps may perform a wide range of health care delivery and information dissemination functions.

A 2016 systematic review identified 539 mHealth oncology apps available for public download [16]. Content categories across the oncology apps included teaching and information, medication help, monitoring, prevention, screening, smoking cessation, and patient support. A more recent 2021 systematic review identified 794 mHealth oncology apps available for public download [17], a 47% increase in new apps created over a 5-year time frame. Content categories across oncology apps included education, clinical decision support, and patient support. However, most available oncology apps have not been developed or tested in research settings, and few have Food and Drug Administration (FDA) approval as “digital therapeutics” [16-23]. A 2020 systematic review identified only nine publicly available oncology apps with scientific publications supporting their feasibility or efficacy, suggesting an industry-research gap [18].

Recent research suggests that patients with cancer are interested in using mHealth apps. Studies have found positive patient attitudes toward mHealth apps for a range of functions over the cancer care continuum [24-27], such as medication adherence, early detection and diagnostics, patient-provider communication, lifestyle interventions, weight and diet management, exercise facilitation, symptom monitoring, medical health literacy, disease self-management, and medical decision-making [28-38]. Results of a 2018 oncology patient survey showed that over 60% of survey respondents regularly used the internet to research information about disease prevention, healthy lifestyles, health care, and their cancer treatment [39]. Over 80% of respondents stated that they would download and use an app if their health care provider recommended it to them, highlighting the importance of health care provider involvement in patient adoption of mHealth tools [39].

To date, relatively few studies have examined oncology health care provider attitudes toward mHealth apps [21,40,41]. An international survey of oncologists showed that 80% of respondents used mHealth, primarily for the following functions: to conduct literature reviews, to interact with colleagues, for patient communication, to access patient electronic health records, to collect test results, and for follow-up care [40]. A single-site provider survey of oncologists and nurses in Germany found that the mHealth functions rated as most useful were appointment and medication reminders, assessment of side effects, and collecting patient-reported outcomes [41]. Most survey respondents considered mHealth apps to be a welcome complement to traditional health care and perceived benefits to implementing mobile apps as part of the standard of care. Provider concerns endorsed by a minority of respondents included legal/medical responsibility, data privacy, and data security. A qualitative interview study [21] found that oncology providers expressed openness to mHealth apps and acknowledged their potential to fill care gaps. Perceived implementation barriers included limited access to technologies, time burden, and concerns about the trustworthiness and credibility of app developers [21]. Providers prioritized app usability and patient preferences over app efficacy or effectiveness [21].

Early research on provider attitudes toward mHealth apps focused primarily on oncologist perspectives and disease management functions. Here, we filled a research gap by incorporating perspectives from providers across diverse health care roles, including medical and psychosocial providers. We included providers who work with patients across the age continuum who may have different attitudes toward and familiarity with mHealth. We explored feedback on mHealth content that extends beyond disease management to include

XSL•FO
RenderX

behavioral health and holistic patient care. Holistic patient care is a "whole person" approach to medicine that encompasses physical, mental, emotional, lifestyle, and social needs—not just treating symptoms of disease—in order to promote overall well-being and improve long-term health outcomes [42-44]. This broader focus is critical for understanding how technology can support a more comprehensive approach to patient outcomes. Consequently, this study sought to report an update and extension of prior research by answering the following research questions (RQs): (1) What types of apps have oncology providers used with their patients or would consider potentially beneficial? (2) What are the attitudes, beliefs, facilitators, and barriers related to the adoption of mHealth apps in oncology settings? (3) (Exploratory) How do provider preferences differ based on provider role (medical vs psychosocial), provider age (<45 vs ≥45 years), and patient populations served (pediatric vs adult)? The selection of exploratory variables was informed by existing mHealth and technology acceptance literature [45-50]. Findings will inform recommendations for apps that providers perceive to have clinical utility, future directions for app development, and methods to promote implementation and uptake of mHealth apps.

Methods

Ethical Considerations

All study procedures were approved by the Seattle Children's Hospital Institutional Review Board (IRB approval number STUDY00004962; approval date June 11, 2024). The study was conducted under IRB-exempt status in which survey information obtained is recorded in a manner that the identity of the survey respondents cannot be ascertained. Survey participation was voluntary and anonymous, and no compensation was offered for completing the survey. A one-page REDCap (Research Electronic Data Capture) Study Information Sheet containing purpose of the study, investigator information, study description, confidentiality statement, data storage details, and contact information was provided on the main screen of the online questionnaire. Survey respondents were informed via the Study Information Sheet that their responses would remain anonymous and research records would be stored exclusively within secure institutional intranet databases. The survey was administered through Seattle Children's Hospital REDCap, a secure web-based survey platform that uses password-protected access and role-based permissions to prevent unauthorized data access.

Study Design

We disseminated an online survey called the "Oncology Provider Experiences National Survey (OPENS): Mobile Health" (hereafter referred to as "OPENS") requesting responses on provider attitudes and experiences toward the use of mHealth apps in oncology. The target population for this study consisted of health care professionals currently practicing in oncology settings within the United States. Participation was open to all who self-identified as an oncology provider. Although we did not include an eligibility screener, eligibility was controlled with targeted recruitment through oncology provider list servers (listservs). Our survey was distributed nationally across the United States to members of the American Psychosocial Oncology Society (APOS) and provider listservs at the home institutions of several of the coauthors (eg, Seattle Children's Hospital Cancer and Blood Disorders Center, Fred Hutchinson Cancer Center, MD Anderson Cancer Center, St Jude Children's Research Hospital). We distributed the survey to several organizations that are pediatric-focused: Children's Oncology Group (COG) Behavioral Science Committee, Society of Pediatric Psychology Hematology/Oncology/Bone Marrow Transplant Special Interest Group, and Association of Pediatric Oncology Social Workers (APOSW). Permission to distribute our online survey was obtained by research committees or administrators, as appropriate for each respective listserv. Each listserv was sent a single request with a designated survey end date and with no follow-up reminders. We staggered survey distribution pending submission and review of materials by each listserv's respective research committee and administrators. Prospective participants were informed via the survey dissemination email and the REDCap Study Information Sheet that the survey was intended for oncology providers only. The survey took approximately 10 minutes to complete. The survey remained open from June 2024 to December 2024. This study was reported in accordance with CHERRIES (Checklist for Reporting Results of Internet E-Surveys). See Multimedia Appendix 1 for the complete checklist.

Oncology Provider Experiences National Survey: Mobile Health

To the best of our knowledge, no validated measure currently exists to assess oncology provider attitudes toward mHealth apps. We developed the 30-item OPENS to describe mHealth-related usage, attitudes, perceived benefit, facilitators, and barriers. OPENS was designed as a descriptive survey to assess provider perspectives and usage patterns rather than as a psychometric scale intended to measure a single latent construct. These survey items directly address our study aims to inform the clinical integration of mHealth technologies in oncology care. Internally developed surveys are a standard approach in digital health research when established measures are unavailable [51-54]. Survey items were collaboratively developed by a multidisciplinary research team (including experts in psychology, psychiatry, social work, oncology, and digital health) following a targeted literature review. We established content validity using a modified Delphi approach [55-58]. This process involved asynchronous feedback in which collaborators independently reviewed survey items for clarity and relevance to both the study aims and the oncology clinical context. Facilitated group consensus meetings were held with external oncology experts to refine survey items. Eight items assessing mHealth attitudes and beliefs were adapted from the "Questionnaire to Assess Public Health Staff Viewpoints about Mobile Phone Use in the Delivery of Their Services" [59]. These items were updated for oncology relevance. Survey items were finalized asynchronously among the research team members.

Before national dissemination, the survey underwent pilot-testing with a small group of oncology providers to assess technical functionality within REDCap, item comprehension, and user burden. All 33 items were distributed across 4 pages and administered in a forced-response format to ensure complete

XSL•FO
RenderX

data capture. Questions included single-choice items, multiple-choice items, "5-point Likert scale" items (eg, strongly disagree to strongly agree), and "select all that apply" items (including "none of the above" and "other" options). Respondents were able to review and modify their responses prior to survey submission, with no time limits or automatic expiration imposed for completion. Consistent with the scoring method of the validated instrument from which eight survey items were adapted [48], Likert-scale responses were summarized using means (SDs) to characterize the strength and variability of providers' beliefs and attitudes on mHealth apps. The brief close-ended survey was designed to reduce participant burden and maximize response rates for uncompensated national listserv recruitment methods, targeting a completion time of approximately 10 minutes. To ensure that our study met IRB-exempt criteria, we did not include open-ended survey questions in which human subjects' identifying information may have inadvertently been obtained. See Multimedia Appendix 2 for all survey items.

Statistical Analysis

We summarized provider demographics and other survey items using descriptive statistics and data visualizations created in Canva. Exploratory statistical analyses were performed using IBM SPSS Statistics (version 29). Study results are presented as frequency counts, percentages, and means (SDs). We also conducted exploratory analyses on survey responses broadly categorized by provider type (medical providers vs psychosocial providers) who may differ in their attitudes toward and the use of mHealth tools for disease management and behavioral/emotional support based on their training backgrounds and areas of expertise. For the purposes of this study, medical providers were defined as those whose primary training and clinical focus address disease treatment and the management of physiological symptoms. This group included those who identified their role as oncologist, nurse or advanced practice provider, physician, or health care administrator. Psychosocial providers were defined as clinicians whose primary training and clinical focus address the psychological, social, and emotional aspects of the cancer experience. This group included those who identified their role as psychologist, social worker, mental health therapist, or occupational therapist. This categorization also allowed for a balanced distribution of providers for comparison (90/188, 47.9%, medical vs 89/188, 47.3%, psychosocial). We performed chi-square analyses to examine differences between medical providers and psychosocial providers on the use of mHealth apps in clinical practice, mHealth apps endorsed as potentially helpful or beneficial, and the perceived facilitators of and barriers to mHealth app adoption. We also performed chi-square analyses to examine differences between younger providers (age≤45 years) vs older providers (age≥46 years) and pediatric care providers (children, teens, or young adults) vs adult care providers (adults, older adults) on the use of mHealth apps in clinical practice. Given that the study used categorical data, traditional assumptions of normality were not applicable. As all chi-square analyses were exploratory and intended for hypothesis generation rather than formal hypothesis testing, we did not control for multiple comparisons.

Results

Demographic and Provider Role Characteristics

A total of 188 providers responded to the survey (Table 1). Participants ranged in age from under 24 to over 65 years. Over one-third (70/188, 37.2%) were between the ages of 36 and 45 years. Most providers identified as female (150/188, 79.8%), White (161/188, 85.6%), and not Hispanic or Latino (174/188, 92.6%).

XSL•FO
RenderX

Table 1. Demographic characteristics of oncology providers (N=188).

Characteristics	Respondents, n (%)
Age group (years)	
<24	2 (1.1)
24-35	28 (14.9)
36-45	70 (37.2)
46-55	42 (22.3)
56-65	29 (15.4)
>65	17 (9.0)
Gender identity	
Female	150 (79.8)
Male	37 (19.7)
Prefer not to respond	1 (0.5)
Race	
American Indian or Alaska Native	2 (1.1)
Asian	13 (6.9)
Black or African American	0 (0.0)
Hispanic or Latino	7 (3.7)
Middle Eastern or North African	1 (0.5)
Native Hawaiian or Pacific Islander	0 (0.0)
White	161 (85.6)
Not listed	2 (1.1)
Prefer not to respond	4 (2.1)
Ethnicity	
Hispanic or Latino	12 (6.4)
Not Hispanic or Latino	174 (92.6
Prefer not to respond	2 (1.1)

Of the 188 respondents, 47.9% (90/188) were medical providers, 47.3% (89/188) were psychosocial providers, and 4.8% (9/188) had administrative staff or research roles. Psychologists represented the largest group of survey respondents (66/188, 35.1%), followed by oncologists (54/188, 28.7%) and nurses or advanced practice providers (32/188, 17%), as shown in Table 2. Most providers practiced in academic or teaching hospitals (155/188, 82.4%) and urban settings (130/188, 69.1%). Over half of the sample provided care for children (120/188, 63.8%), teens (124/188, 66%), and young adults (143/188, 76.1%).

XSL•FO
RenderX

Table 2. Role characteristics of oncology providers (N=188).

Characteristics	Respondents, n (%)
Current role[a]	
Oncologist	54 (28.7)
Nurse or advanced practice provider	32 (17.0)
Psychologist	66 (35.1)
Social worker	13 (6.9)
Health care administrator	2 (1.1)
Researcher	33 (17.6)
Physician (other specialty)	4 (2.1)
Other	14 (7.4)
Practice setting[a]	
Private practice	11 (5.9)
Academic or teaching hospital	155 (82.4)
Community hospital	16 (8.5)
Hospice care	0 (0.0)
Other	15 (8.0)
Practice location	
Rural	15 (8.0)
Suburban	35 (18.6)
Urban	130 (69.1)
Other	8 (4.3)
Duration of experience providing care (years)	
<1	1 (0.5)
1-5	39 (20.7)
6-10	41 (21.8)
11-15	38 (20.2)
>16	68 (36.2)
Prefer not to respond	1 (0.5)
Patient age demographic[a]	
Children	120 (63.8)
Teens	124 (66.0)
Young adults	143 (76.1)
Adults	82 (43.6)
Older adults	50 (26.6)

[a]Questions on the respondents' current role, practice setting, and patient age demographic were multiple-response items; providers could select all applicable options.

mHealth Apps Used in Clinical Practice

Nearly all providers (178/188, 94.7%) reported recommending or using at least one mHealth app with their patients (Figure 1). Most providers recommended or used telehealth and virtual consultation apps (139/188, 73.9%). Other types of mHealth apps that were commonly recommended to or used with patients included communication with the medical team (136/188, 72.3%) and mental health and wellness (98/188, 52.1%) apps. Only a small proportion of providers reported recommending or using cancer and disease management (21/188, 11.2%) or health literacy (15/188, 8%) mHealth apps with their patients. Both medical and psychosocial providers predominantly used mHealth apps for medical team communication or telehealth (Figure 2 and Figure 3, respectively). Psychosocial providers (25/89, 28.1%) were significantly more likely than medical providers (6/90, 6.7%) to recommend or use pain management

XSL•FO
RenderX

apps (χ^2_1=14.34, *P*<.001, ϕ=0.28). Psychosocial providers (70/89, 78.7%) were significantly more likely than medical providers (23/90, 25.6%) to recommend or use mental health apps (χ^2_1=50.54, *P*<.001, ϕ=0.53). Psychosocial providers (48/89, 53.9%) were significantly more likely than medical providers (16/90, 17.8%) to recommend or use sleep health apps with their patients (χ^2_1=25.47, *P*<.001, ϕ=0.38). All other chi-square analyses comparing medical and psychosocial provider groups were nonsignificant.

Figure 1. Comparison of mHealth apps used in clinical practice vs potentially useful or beneficial mHealth apps for patients. mHealth: mobile health.

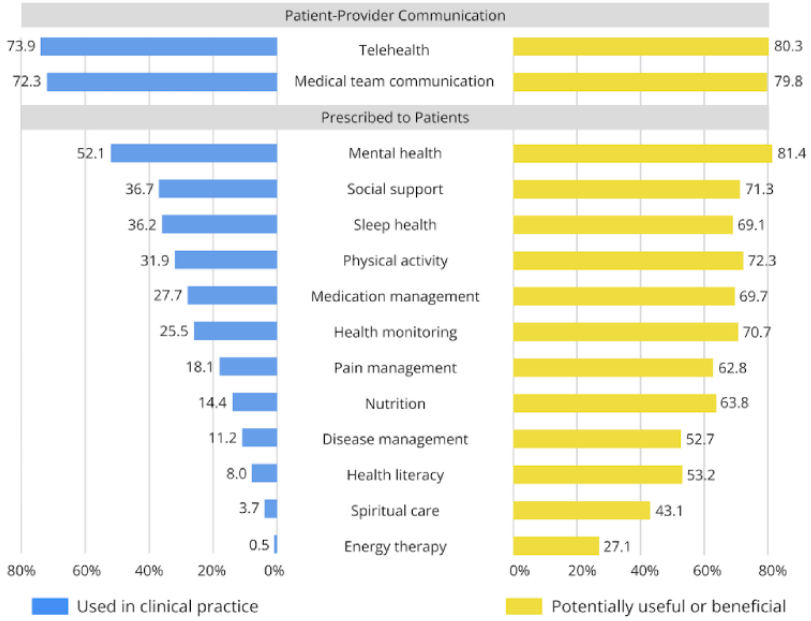

Figure 2. Comparison of mHealth apps used in clinical practice vs potentially useful or beneficial mHealth apps for medical providers. mHealth: mobile health.

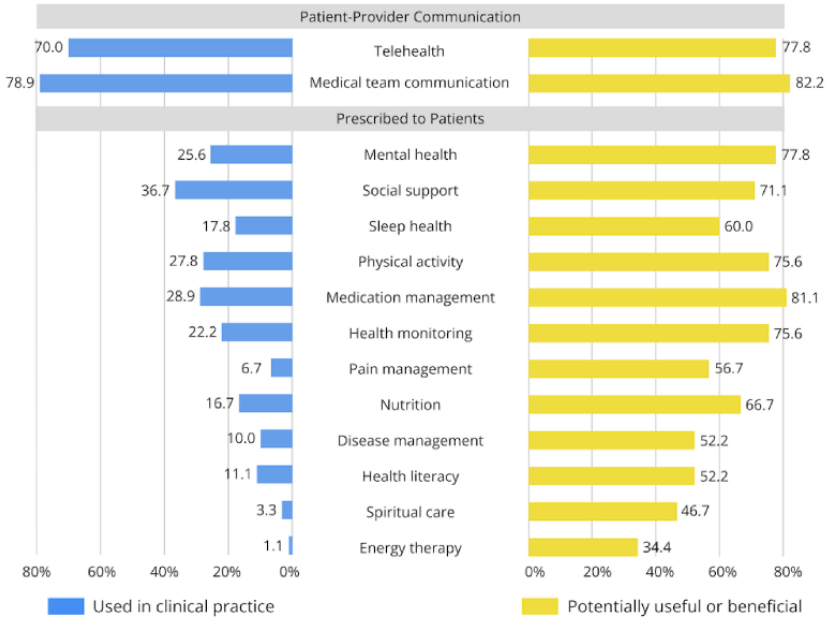

XSL•FO
RenderX

Figure 3. Comparison of mHealth apps used in clinical practice vs potentially useful or beneficial mHealth apps for psychosocial providers. mHealth: mobile health.

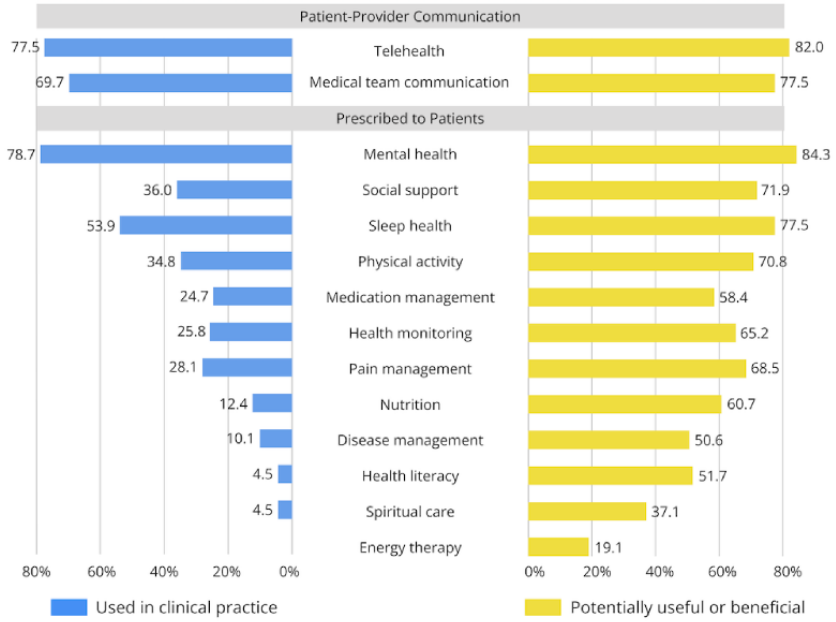

Providers stratified by age (≤45 and ≥46 years) recommended or used similar types of mHealth apps with patients (Multimedia Appendix 3). Older providers (16/88, 18.2%) were significantly more likely than younger providers (5/100, 5%) to recommend or use disease management apps with their patients (χ^2_1=8.20, *P*=.005, ϕ=0.21). All other chi-square analyses comparing older and younger providers were nonsignificant. Providers working with pediatric and adult patients also recommended or used similar types of mHealth apps (Multimedia Appendix 4). Adult care providers (8/32, 25%) were significantly more likely than pediatric care providers (6/101, 5.9%) to recommend or use disease management apps (χ^2_2=9.22, *P*=.01, Cramer V=0.22). Pediatric care providers (83/101, 82.2%) were more likely than adult care providers (15/32, 46.9%) to recommend or use medical team communication apps (χ^2_2=15.66, *P*<.001, Cramer V=0.29). All other chi-square analyses comparing adult and pediatric care providers were nonsignificant.

Potentially Useful or Beneficial mHealth Apps

All providers endorsed mHealth apps to be potentially useful or beneficial for patients (Figure 1). Most providers perceived mental health and wellness apps to be potentially useful (153/188, 81.4%). Other types of mHealth apps that were perceived as potentially useful were telehealth and virtual consultation (151/188, 80.3%), communication with the medical team (150/188, 79.8%), exercise and physical activity (136/188, 72.3%), social support (134/188, 71.3%), health monitoring and medical symptom tracking (133/188, 70.7%), medication management (131/188, 69.7%), and sleep health (130/188, 69.1%). Although providers predominantly used mHealth apps for patient communication (eg, medical team communication, telehealth), they endorsed interest in prescribing mHealth apps for a broad range of holistic care functions.

Interest in a wide array of psychosocial support functions was fairly consistent across medical and psychosocial providers (Figure 2 and Figure 3, respectively). Medical providers (73/90, 81.1%) were significantly more likely than psychosocial providers (52/89, 58.4%) to perceive medication management apps as potentially beneficial for their patients (χ^2_1=10.93, *P*<.001, ϕ=0.25). Medical providers (31/90, 34.4%) were significantly more likely than psychosocial providers (17/89, 19.1%) to perceive energy therapy apps as potentially beneficial (χ^2_1=5.37, *P*=.02, ϕ=0.17). In contrast, psychosocial providers (69/89, 77.5%) were significantly more likely than medical providers (54/90, 60%) to endorse sleep health apps as potentially helpful for their patients (χ^2_1=6.40, *P*=.01, ϕ=0.19). All other chi-square analyses comparing medical and psychosocial provider groups were nonsignificant.

mHealth Attitudes and Beliefs

Providers had positive attitudes and beliefs toward specific facets of mHealth apps (Table 3). Providers had the highest mean scores for specific items pertaining to health equity, indicating that they “agree” or “strongly agree” that mHealth apps reduce travel costs for services (mean 4.18/5, SD 0.68) and expands the range of health services (mean 4.18/5, SD 0.69). Providers had the highest mean scores for specific items pertaining to confidence and a growth mindset, indicating that they “agree” or “strongly agree” that they are confident in their ability to navigate mobile apps (mean 4.14/5, SD 0.83) and can adapt to health technology changes over time (mean 4.18/5, SD 0.70). Providers had the lowest mean scores for specific items pertaining to the patient-provider relationship and training, indicating that they “neither agree nor disagree” that mHealth improves provider-patient interactions (mean 3.21/5, SD 0.95) or are trained and informed about mHealth regulatory and ethical issues in health care (mean 3.24/5, SD 1.07).

XSL•FO
RenderX

Table 3. Attitudes and beliefs toward mHealth[a] apps among oncology providers.

Item	Mean (SD)
Training	
I feel that I am trained and informed about regulatory and ethical issues related to technology use in health care.	3.24 (1.07)
I can effectively teach patients how to use mHealth technologies for their care.	3.44 (1.03)
I am interested in receiving more training or specialized training in the use of digital tools in my health care practice.	3.88 (0.96)
Confidence and growth mindset	
I can adapt to health technology changes over time.	4.18 (0.70)
I feel confident in my ability to navigate mobile apps.	4.14 (0.83)
I am comfortable incorporating technology into patient care practices.	3.74 (0.90)
Quality of care	
The use of mobile technology improves the quality of public health services.	3.92 (0.68)
Mobile technology can help improve patients' quality of life.	3.94 (0.66)
Mobile technology improves self-management in public health care professionals.	3.65 (0.76)
Mobile technology can improve patients' health literacy and encourage them to actively engage in their care.	3.86 (0.73)
Health equity	
Using mobile technology reduces the cost of providing public health services.	3.64 (0.83)
The use of mobile technology expands the range of health services.	4.18 (0.69)
The use of mobile technology creates equal access to facilities and services for the general public.	3.31 (0.98)
Using mobile technology reduces the number of visits to health centers.	3.77 (0.83)
The use of mobile technology reduces travel costs for services.	4.18 (0.68)
Patient-provider relationship	
Mobile technology improves interpersonal interactions between the provider and the recipient.	3.21 (0.95)
Integrating mobile technology can improve patient-provider interactions.	3.81 (0.70)

[a]mHealth: mobile health.

Facilitators of mHealth Adoption

The top five perceived facilitators of mHealth adoption in clinical settings were "alignment with patients' expressed interest and needs" (147/188, 78.2%), "improved health care accessibility" (135/188, 71.8%), "cost-effectiveness for patients and health care systems" (127/188, 67.6%), "seamless integration with existing health care systems and/or electronic medical records" (113/188, 60.1%), and "improved job performance and efficiency" (109/188, 58%); see Multimedia Appendix 5. Medical and psychosocial providers reported similar top facilitators overall (Figure 4). However, medical providers (47/90, 52.2%) were significantly more likely than psychosocial providers (23/89, 25.8%) to endorse patient health literacy as a perceived facilitator (χ^2_1=13.08, P<.001, ϕ=0.27).

Figure 4. Common facilitators of adopting mHealth apps in health care settings among medical and psychosocial providers. mHealth: mobile health.

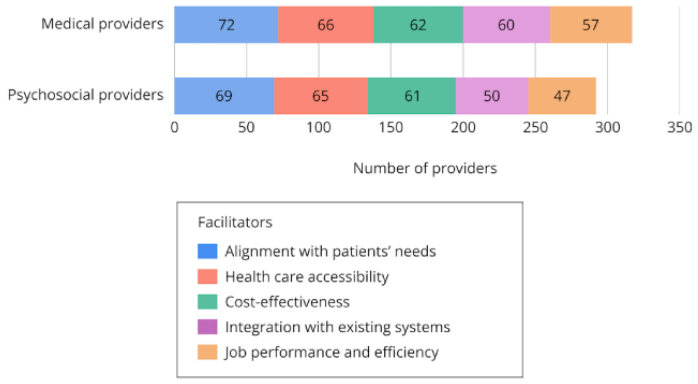

XSL•FO
RenderX

Barriers to mHealth Adoption

The top five perceived barriers to mHealth adoption in clinical settings were "disparities in access to technology" (132/188, 70.2%), "digital literacy" (ie, technical knowledge to access, understand, and use digital resources and tools; 117/188, 62.2%), "data privacy and safety concerns" (114/188, 60.6%), "lack of knowledge regarding which digital tools to recommend or where to find relevant information" (110/188, 58.5%), and "being overwhelmed by too many digital health technology options" (105/188, 55.9%); see Multimedia Appendix 6. Medical and psychosocial providers reported similar top barriers overall (Figure 5). However, medical providers (52/90, 57.8%) were more likely than psychosocial providers (38/89, 42.7%) to endorse cost as a perceived barrier (χ^2_1=4.07, *P*=.04, ϕ=0.15).

Figure 5. Common barriers to adopting mHealth apps in health care settings among medical and psychosocial providers. mHealth: mobile health.

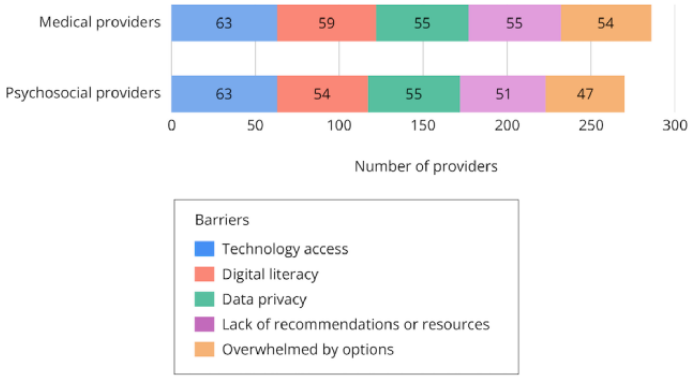

Discussion

Principal Findings

This study provided an update to and extension of prior research regarding the integration of mHealth apps in oncology care. Regarding mHealth app usage and its perceived benefit (RQ1), oncology providers commonly use mHealth apps, with an emphasis on patient-provider communication tools. Although current usage is primarily logistical, there is clear interest in whole person health–supporting apps that providers consider potentially beneficial but that currently see lower rates of adoption. With regard to provider attitudes, facilitators, and barriers related to mHealth adoption (RQ2), providers consistently endorsed a growth mindset, reporting high confidence in their ability to learn and adopt digital tools. Providers recognized the potential for mHealth to improve health equity by expanding access to care. The top perceived facilitators of mHealth adoption were alignment with patient needs, improved health care accessibility, and cost-effectiveness. The top perceived barriers to mHealth adoption were disparities in access to technology, digital health literacy, and data privacy.

With regard to exploratory differences by clinical role, provider age, and patient age (RQ3), we found several group differences. Reflecting their respective training backgrounds, medical providers were more likely to endorse the potential benefit of medication management apps, whereas psychosocial providers were more likely to use and/or perceive potential benefit in pain management, mental health, and sleep health apps. The use of mHealth apps was generally consistent across oncology provider age groups (≤45 and ≥46 years) and patient population (pediatric care providers and adult care providers). However, older providers (age≥46 years) and those providing care to adult patients used more disease management tools, whereas pediatric providers used more medical team communication apps.

Comparison With Prior Work

Our study expands on the background literature on oncology providers' attitudes toward mHealth apps by soliciting perspectives of those in various clinical roles, not exclusively oncologists; expanding mHealth app functions beyond disease management; and identifying perceived barriers to and facilitators of implementation in oncology care settings. Our findings are consistent with previous studies that have reported that oncologists primarily use mHealth tools for telehealth and medical team communication and rarely use mHealth tools for disease management and follow-up care [40]. Of note, we also found that medical and psychosocial providers alike are interested in using mHealth apps focusing on a wide range of psychosocial care functions to their patients and perceived them to have clinical utility. This was consistent with the literature on patients with cancer and those who have survived cancer, showing high rates of acceptance of mHealth apps but low rates of adoption [60-64]. Specific patient facilitators and barriers beyond demographic and disease characteristics have not yet been examined. Studies examining patient preferences have predominantly focused on disease management apps [60,62-66]. Consistent with the background literature, we found that a primary barrier identified by providers is disparities in access to technology [21,40,67]. However, research has shown that mHealth technologies are universally accessible to individuals regardless of income level, rurality, or marginalized status [3,4,68]. In the United States, 79%-94% of individuals from low-income households (<US $30,000) have access to smartphones, computers, and tablets [3,4,69]. Digitally deployed health interventions have also been proposed for populations/persons experiencing homelessness as viable alternatives to in-person services to overcome traditional access care barriers [68,70,71]. In the United States, 91% of individuals own a smartphone, 95% use the internet, and 90% have a high-speed internet subscription at home [1,3,4,69]. Universal use of smartphones holds across the age continuum, with 95%

XSL•FO
RenderX

of teens/young adults, 95%-97% of adults aged>45 years [2-4,72], and 76%-79% of older adults aged>65 years owning a smartphone [4,69,72]. Digital technologies (smartphones, computers, tablets) have become significantly more affordable in recent years, and their rapid adoption has been further accelerated by the COVID-19 pandemic.

Consistent with the background literature, a primary barrier to mHealth adoption, as identified by providers, was limited digital health literacy [61]. Previous studies have shown that health literacy and digital health literacy are significant mediators of the efficacy of digital health interventions [73] and are not fixed but malleable patient characteristics [74]. Thus, providers may be more likely to recommend and use mHealth apps with intuitive user interfaces, accessibility for those with impairments, provision of health information at lower reading levels, and multimedia options (eg, text-to-speech and animations) [61,75,76]. Recommendations for mHealth may also be paired with apps that identify a user's health literacy level and provide tailored definitions of medical terminologies to their reading level (eg, MediReader [77]).

Previous research has found that younger oncologists (<45 years old) and younger cancer patients and survivors (<45 years old) are more likely to use internet-enabled mHealth tools for disease management and cancer care [40]. Our results showed consistent use of mHealth apps in younger as well as in older providers (>45 years old) and among providers of patients across the age continuum. For younger pediatric patients, caregivers may primarily interact with mHealth apps, while teens and young adults may be more independent app users [78]. Digital inclusion for older adults is recognized as an important social determinant of health and a crucial component of a healthy aging society [79-88]. These findings highlight the importance of developing, testing, and implementing mHealth apps that are tailored to the diverse needs of patients with cancer and those who have survived cancer, of all ages.

The Technology Acceptance Model (TAM) proposes that perceived usefulness (ie, the belief that using a particular technology enhances job performance) and perceived ease of use (ie, the belief that using a particular technology is effort free) are the primary drivers of user adoption of new technologies [89,90]. Interpreted through the TAM, our findings suggest high readiness for mHealth adoption across the oncology workforce. Specifically, providers' endorsement of mHealth apps for reducing costs and expanding care access aligns with high perceived usefulness. Providers' reported growth mindset and confidence in their ability to learn and adapt to mHealth technologies reflect high self-efficacy, which aligns with high perceived ease of use. Although TAM pathways were not formally tested in our study, the alignment between our findings and these TAM constructs suggest that oncology providers have the positive attitudes and behavioral intentions necessary to facilitate mHealth adoption.

Future Directions

mHealth apps that address holistic patient care have the potential to improve patient outcomes across the cancer care continuum and patient age continuum. Despite the narrow use of mHealth apps in current practice, the perceived utility of mHealth across medical and psychosocial domains suggests important opportunities for broader integration. Guided by the RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) framework [91-93], future research should focus on expanding *reach* by partnering with medical and psychosocial providers on codesign strategies of mHealth tools in the innovation design-to-implementation pipeline. Future research should also enhance *effectiveness* via rigorous clinical testing and addressing the complex ethical and regulatory landscape of mHealth apps. Although oncology providers are interested in using mHealth tools, the lack of standardized clinical guidelines for recommending non-FDA-approved apps remains an important barrier. Endorsing or recommending mHealth clinical tools that lack rigorous scientific testing poses risks to patient safety and data confidentiality. Future research and policy work should focus on developing clinical guidelines to identify evidence-based mHealth apps and pathways for determining their clinical efficacy/effectiveness. In addition, future research should support *adoption* and *implementation* by examining providers' nuanced perspectives on the perceived barrier of technology access and investigate the specific sources of these concerns (eg, financial burden of cancer, lack of reliable cellular or internet service in rural areas). Efforts to mitigate barriers, such as concerns about data privacy and security, will also be critical to ensuring adoption of mHealth solutions. Successful integration of mHealth apps into oncology care will require accessible, user-centered design (eg, intuitive interfaces, appropriate reading levels, multimedia supports) and shared decision-making informed by ongoing examination of concordant and discordant beliefs and attitudes toward mHealth apps among providers and patients. Long-term *maintenance* will depend on sustained institutional and policy support (reimbursement pathways, standardized evaluation criteria, regulatory guidance on data privacy and security), engagement of local providers or stakeholder champions to promote uptake through education and shared learning in navigating digital health systems, and preservation of alternative care options (eg, in-person services) as technologies continues to evolve. Future research using a mixed methods approach is also warranted to provide richer context to these exploratory findings.

Limitations

This study has several limitations. First, this is a convenience sample of oncology providers, which introduces the potential for selection and response bias due to voluntary participation. It is possible that providers with a pre-existing interest in mHealth were more motivated to participate in the study, potentially leading to nonresponse bias where favorable attitudes are overrepresented. In addition, social desirability bias may have influenced participants to overreport favorable attitudes to align with current digital health trends. To mitigate potential bias, we distributed the survey through multiple national professional listservs representing diverse professional roles. In addition, survey questions assessed diverse app functions, barriers, and facilitators, which resulted in variability in study findings. This suggests that the survey captured an array of provider perspectives. Second, most respondents identified as White and female, practiced in urban settings, and did not see older adults in their clinical practice. This limits the

XSL•FO
RenderX

generalizability and transferability of our findings to underrepresented providers because of a lack of diversity in our sample. Future studies should implement recruitment strategies to target provider demographic diversity, rural health settings, and those in older adult and geriatric care to ensure findings are representative of all clinical contexts. Third, this is a small sample relative to the number of oncology providers who practice nationally, and we relied on our institutional and national professional organization listservs to distribute the survey. Sample size was determined by convenience sampling within the study recruitment time frame. As the primary aim of this study was descriptive, we did not establish a predetermined target sample size. Fourth, as there are no well-validated measures of provider attitudes toward mHealth interventions with established cutoff scores, we used an internally developed survey. This influenced the study by limiting our ability to categorize and compare scores with existing studies. To mitigate the lack of a pre-existing validated scale, we used a modified Delphi process involving a multidisciplinary team and external oncology experts to ensure content validity. Fifth, to comply with IRB-approved protocols and IRB-exempt status, we did not collect identifiers from survey respondents, so it is not possible to determine response rates from specific organizations or listservs or analyze group differences based on affiliations. Future research could include such identifiers in order to assess representativeness and potential cohort effects. Sixth, due to the exploratory nature of our analyses, we did not control for multiple comparisons or include demographic covariates. Although this approach increases the risk of type 1 errors, we mitigated the risk of overinterpretation by explicitly framing these analyses as hypothesis generating, and our findings should be viewed as preliminary. Finally, this was a cross-sectional study who findings provide a snapshot in time of provider attitudes in a post-COVID-19 pandemic landscape that has normalized digital health care. Because attitudes and perspectives may continue to shift over time and technology evolves rapidly, future longitudinal research is warranted.

Conclusion

The findings of this study reveal a significant “readiness gap” in oncology care. Although current use remains largely focused on telehealth services and patient-provider communication, providers expressed interest in integrating a wider array of mHealth apps encompassing health information, disease and symptom monitoring, mental health resources, and psychosocial support. Our findings suggest that the oncology workforce is prepared to move beyond logistical communication tools toward an integrated holistic model of oncology care that encompasses broad disease management and psychosocial support functions. Because these positive attitudes are consistent across clinical roles and patient populations, mHealth interventions have broad appeal and should be implemented as interdisciplinary standards of care. Health care systems can leverage existing provider buy-in to create a more equitable and accessible oncology care model.

Acknowledgments

We thank the providers who participated in the study and shared their insights.

Data Availability

The study was conducted under Institutional Review Board (IRB)-exempt status, and data generated are not publicly available. Data may be made available from the corresponding author (NL) upon reasonable request and subject to a formal data use agreement.

Funding

NL is supported by the National Cancer Institute at the National Institutes of Health (grant number 1K08CA263474). The opinions herein represent those of the authors and not necessarily the funders.

Authors' Contributions

Conceptualization was handled by NL, KS, HA, MER, KMI, AB, EJC, and TMP; data curation by KS, SJHH, and HA; formal analysis and writing—original draft by NL, KS, and SJHH; funding acquisition by NL; methodology by NL, KS, and TMP; visualization by SJHH; and investigation and writing—review and editing by all authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Checklists for Reporting Results of Internet E-Surveys (CHERRIES).
[DOCX File , 26 KB - jmir_v28i1e85583_app1.docx]

Multimedia Appendix 2
Oncology Provider Experiences National Survey (OPENS): Mobile Health.
[DOCX File , 27 KB - jmir_v28i1e85583_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Types of mobile health (mHealth) apps providers prescribed to or used with patients stratified by provider age (≤45 vs ≥46 years).
[PNG File , 135 KB - jmir_v28i1e85583_app3.png]

Multimedia Appendix 4
Types of mobile health (mHealth) apps providers prescribed to or used with patients stratified by patient age (pediatric vs adult).
[PNG File , 136 KB - jmir_v28i1e85583_app4.png]

Multimedia Appendix 5
Facilitators of adopting mobile health (mHealth) apps in health care settings.
[PNG File , 165 KB - jmir_v28i1e85583_app5.png]

Multimedia Appendix 6
Barriers to adopting mobile health (mHealth) apps in health care settings.
[PNG File , 120 KB - jmir_v28i1e85583_app6.png]

References

1. Laricchia F. Smartphone penetration worldwide as share of global population 2016-2023. Statista. 2024. URL: https://www.statista.com/statistics/203734/global-smartphone-penetration-per-capita-since-2005/ [accessed 2024-11-24]
2. Bashir U. Americans owning a smartphone in the U.S., by age 2024. Statista. 2024. URL: https://www.statista.com/statistics/231612/number-of-cell-phone-users-usa/ [accessed 2025-01-08]
3. Bazen A, Schmitt K. Cell phone statistics 2025. ConsumerAffairs. 2024. URL: https://www.consumeraffairs.com/cell_phones/cell-phone-statistics.html [accessed 2025-01-08]
4. Gelles-Watnick R. Americans' use of mobile technology and home broadband. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/2024/01/31/americans-use-of-mobile-technology-and-home-broadband/ [accessed 2025-01-08]
5. Laricchia F. Smartphone ownership in the U.S. 2024, by gender. Statista. 2024. URL: https://www.statista.com/forecasts/1363291/smartphone-ownership-by-gender-us [accessed 2025-01-08]
6. Digital health trends 2021. IQVIA. 2021. URL: https://www.iqvia.com/insights/the-iqvia-institute/reports-and-publications/reports/digital-health-trends-2021 [accessed 2024-11-24]
7. Buneviciene I, Mekary RA, Smith TR, Onnela JP, Bunevicius A. Can mHealth interventions improve quality of life of cancer patients? A systematic review and meta-analysis. Crit Rev Oncol Hematol 2021 Jan;157:103123 [FREE Full text] [doi: 10.1016/j.critrevonc.2020.103123] [Medline: 33190065]
8. Qudah B, Luetsch K. The influence of mobile health applications on patient - healthcare provider relationships: a systematic, narrative review. Patient Educ Couns 2019 Jun;102(6):1080-1089 [FREE Full text] [doi: 10.1016/j.pec.2019.01.021]
9. Rivers BM, Bernhardt JM, Fleisher L, Green BL. Opportunities and challenges of using technology to address health disparities. Future Oncol 2014 Mar;10(4):519-522 [FREE Full text] [doi: 10.2217/fon.14.17] [Medline: 24754582]
10. Saevarsdottir SR, Gudmundsdottir SL. Mobile apps and quality of life in patients with breast cancer and survivors: systematic literature review. J Med Internet Res 2023 Jul 26;25:e42852 [FREE Full text] [doi: 10.2196/42852] [Medline: 37494111]
11. Schliemann D, Tan MM, Hoe WMK, Mohan D, Taib NA, Donnelly M, et al. mHealth interventions to improve cancer screening and early detection: scoping review of reviews. J Med Internet Res 2022 Aug 15;24(8):e36316 [FREE Full text] [doi: 10.2196/36316] [Medline: 35969450]
12. Sibelius K. 2011 mHealth summit. 2011 Presented at: 3rd Annual mHealth Summit; December 6-8, 2011; Washington, DC URL: https://conference.researchbib.com/view/event/10146 [doi: 10.5040/9798765195659]
13. Steinhubl SR, Muse ED, Topol EJ. Can mobile health technologies transform health care? JAMA 2013 Dec 11;310(22):2395-2396 [FREE Full text] [doi: 10.1001/jama.2013.281078] [Medline: 24158428]
14. Stiles-Shields C, Reyes KM, Archer J, Lennan N, Zhang J, Julion WA, et al. mHealth uses and opportunities for teens from communities with high health disparities: a mixed-methods study. J Technol Behav Sci 2022 Sep 13:1-13 [FREE Full text] [doi: 10.1007/s41347-022-00278-y] [Medline: 36117748]
15. van Veen T, Binz S, Muminovic M, Chaudhry K, Rose K, Calo S, et al. Potential of mobile health technology to reduce health disparities in underserved communities. West J Emerg Med 2019 Aug 06;20(5):799-802 [FREE Full text] [doi: 10.5811/westjem.2019.6.41911] [Medline: 31539337]
16. Brouard B, Bardo P, Bonnet C, Mounier N, Vignot M, Vignot S. Mobile applications in oncology: is it possible for patients and healthcare professionals to easily identify relevant tools? Ann Med 2016 Nov 27;48(7):509-515 [FREE Full text] [doi: 10.1080/07853890.2016.1195010] [Medline: 27348761]
17. Upadhyay VA, Landman AB, Hassett MJ. Landscape analysis of oncology mobile health applications. JCO Clin Cancer Inform 2021 May;5:579-587 [FREE Full text] [doi: 10.1200/CCI.20.00156] [Medline: 34033510]

XSL•FO
RenderX

18. Ana F, Loreto M, José LMM, Pablo SM, María Pilar MJ, Myriam SA. Mobile applications in oncology: a systematic review of health science databases. Int J Med Inform 2020 Jan;133:104001 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.104001] [Medline: 31706229]
19. Pandey A, Hasan S, Dubey D, Sarangi S. Smartphone apps as a source of cancer information: changing trends in health information-seeking behavior. J Cancer Educ 2013 Mar;28(1):138-142 [FREE Full text] [doi: 10.1007/s13187-012-0446-9] [Medline: 23275239]
20. Bender JL, Yue RYK, To MJ, Deacken L, Jadad AR. A lot of action, but not in the right direction: systematic review and content analysis of smartphone applications for the prevention, detection, and management of cancer. J Med Internet Res 2013 Dec 23;15(12):e287 [FREE Full text] [doi: 10.2196/jmir.2661] [Medline: 24366061]
21. Berkowitz CM, Zullig LL, Koontz BF, Smith SK. Prescribing an app? Oncology providers' views on mobile health apps for cancer care. JCO Clin Cancer Inform 2017 Nov;1(1):1-7 [FREE Full text] [doi: 10.1200/CCI.17.00107] [Medline: 30657404]
22. Nasi G, Cucciniello M, Guerrazzi C. The role of mobile technologies in health care processes: the case of cancer supportive care. J Med Internet Res 2015 Feb 12;17(2):e26 [FREE Full text] [doi: 10.2196/jmir.3757] [Medline: 25679446]
23. Wolf JA, Moreau JF, Akilov O, Patton T, English JC, Ho J, et al. Diagnostic inaccuracy of smartphone applications for melanoma detection. JAMA Dermatol 2013 Apr;149(4):422-426 [FREE Full text] [doi: 10.1001/jamadermatol.2013.2382] [Medline: 23325302]
24. Blair CK, Harding E, Wiggins C, Kang H, Schwartz M, Tarnower A, et al. A home-based mobile health intervention to replace sedentary time with light physical activity in older cancer survivors: randomized controlled pilot trial. JMIR Cancer 2021 Apr 13;7(2):e18819 [FREE Full text] [doi: 10.2196/18819] [Medline: 33847588]
25. Chow EJ, Doody DR, Di C, Armenian SH, Baker KS, Bricker JB, et al. Feasibility of a behavioral intervention using mobile health applications to reduce cardiovascular risk factors in cancer survivors: a pilot randomized controlled trial. J Cancer Surviv 2021 Aug 10;15(4):554-563 [FREE Full text] [doi: 10.1007/s11764-020-00949-w] [Medline: 33037989]
26. King-Dowling S, Psihogios AM, Hill-Kayser C, Szalda D, O'Hagan B, Darabos K, et al. Acceptability and feasibility of survivorship care plans and an accompanying mobile health intervention for adolescent and young adult survivors of childhood cancer. Pediatr Blood Cancer 2021 Mar;68(3):e28884 [FREE Full text] [doi: 10.1002/pbc.28884] [Medline: 33416214]
27. Quintiliani LM, Mann DM, Puputti M, Quinn E, Bowen DJ. Pilot and feasibility test of a mobile health-supported behavioral counseling intervention for weight management among breast cancer survivors. JMIR Cancer 2016;2(1):e4 [FREE Full text] [doi: 10.2196/cancer.5305] [Medline: 27761518]
28. Allicock M, Kendzor D, Sedory A, Gabriel KP, Swartz MD, Thomas P, et al. A pilot and feasibility mobile health intervention to support healthy behaviors in African American breast cancer survivors. J Racial Ethn Health Disparities 2021 Feb 8;8(1):157-165 [FREE Full text] [doi: 10.1007/s40615-020-00767-x] [Medline: 32385847]
29. Beck SL, Smith R, Mindes J, Beck K, Leah Kim J, Weitzman M, et al. Feasibility and usability of EnergyPoints: a mobile health app to guide acupressure use for cancer symptom management. Integr Cancer Ther 2024;23:15347354231223965 [FREE Full text] [doi: 10.1177/15347354231223965] [Medline: 38284345]
30. Birkhoff S, Cantrell M, Moriarty H, Lustig R. The usability and acceptability of a patient-centered mobile health tracking app among a sample of adult radiation oncology patients. ANS Adv Nurs Sci 2018;41(3):243-259 [FREE Full text] [doi: 10.1097/ANS.0000000000000202] [Medline: 29474226]
31. Gelgoot EN, Kruglova K, Chan P, Lo K, Rosberger Z, Chown P, et al. Evaluation of a mobile health app offering fertility information to male patients with cancer: usability study. JMIR Cancer 2022 May 04;8(2):e33594 [FREE Full text] [doi: 10.2196/33594] [Medline: 35507408]
32. Lee H, Uhm KE, Cheong IY, Yoo JS, Chung SH, Park YH, et al. Patient satisfaction with mobile health (mHealth) application for exercise intervention in breast cancer survivors. J Med Syst 2018 Nov 06;42(12):254 [FREE Full text] [doi: 10.1007/s10916-018-1096-1] [Medline: 30402781]
33. Lee H, Mtengezo JT, Makin MS, Shi L, Malata A, Fitzpatrick J, et al. Mobile health-delivered narrative intervention to increase cervical cancer screening among Malawian women living with HIV: a pilot randomized controlled trial. Asia Pac J Oncol Nurs 2024 May;11(5):100448 [FREE Full text] [doi: 10.1016/j.apjon.2024.100448] [Medline: 38784066]
34. Pereira-Salgado A, Westwood JA, Russell L, Ugalde A, Ortlepp B, Seymour JF, et al. Mobile health intervention to increase oral cancer therapy adherence in patients with chronic myeloid leukemia (the REMIND system): clinical feasibility and acceptability assessment. JMIR Mhealth Uhealth 2017 Dec 06;5(12):e184 [FREE Full text] [doi: 10.2196/mhealth.8349] [Medline: 29212628]
35. Schrier E, Xiong N, Thompson E, Poort H, Schumer S, Liu JF, et al. Stepping into survivorship pilot study: harnessing mobile health and principles of behavioral economics to increase physical activity in ovarian cancer survivors. Gynecol Oncol 2021 May;161(2):581-586 [FREE Full text] [doi: 10.1016/j.ygyno.2021.02.023] [Medline: 33637350]
36. Somers TJ, Abernethy AP, Edmond SN, Kelleher SA, Wren AA, Samsa GP, et al. A pilot study of a mobile health pain coping skills training protocol for patients with persistent cancer pain. J Pain Symptom Manage 2015 Oct;50(4):553-558 [FREE Full text] [doi: 10.1016/j.jpainsymman.2015.04.013] [Medline: 26025279]

XSL•FO
RenderX

37. Li Y, Chen W, Liang Y, Yang L, Hou L. Evaluation of mobile health technology interventions for the postdischarge management of patients with head and neck cancer: scoping review. JMIR Mhealth Uhealth 2023 Oct 23;11:e49051 [FREE Full text] [doi: 10.2196/49051] [Medline: 37870887]
38. Wang T, Huang R, Yang S, Chou C, Chen L. Evaluating the effects of a mobile health app on reducing patient care needs and improving quality of life after oral cancer surgery: quasiexperimental study. JMIR Mhealth Uhealth 2020 Jul 27;8(7):e18132 [FREE Full text] [doi: 10.2196/18132] [Medline: 32716303]
39. Collado-Borrell R, Escudero-Vilaplana V, Calles A, Garcia-Martin E, Marzal-Alfaro B, Gonzalez-Haba E, et al. Oncology patient interest in the use of new technologies to manage their disease: cross-sectional survey. J Med Internet Res 2018 Oct 23;20(10):e11006 [FREE Full text] [doi: 10.2196/11006] [Medline: 30355554]
40. Tarricone R, Cucciniello M, Armeni P, Petracca F, Desouza KC, Hall LK, et al. Mobile health divide between clinicians and patients in cancer care: results from a cross-sectional international survey. JMIR Mhealth Uhealth 2019 Sep 06;7(9):e13584 [FREE Full text] [doi: 10.2196/13584] [Medline: 31493318]
41. Kessel KA, Vogel MME, Schmidt-Graf F, Combs SE. Mobile apps in oncology: a survey on health care professionals' attitude toward telemedicine, mHealth, and oncological apps. J Med Internet Res 2016 Nov 24;18(11):e312 [FREE Full text] [doi: 10.2196/jmir.6399] [Medline: 27884810]
42. Ventegodt S, Kandel I, Ervin D, Merrick J. Concepts of holistic care. In: Rubin IL, Merrick J, Greydanus DE, Patel DR, editors. Health Care for People with Intellectual and Developmental Disabilities across the Lifespan. Cham: Springer; 2016:1935-1941.
43. Ayash EJ, Alasiri MA, Alfarhan FAH, Alshammmari HHR, Alharfi HHK, Alazmi SI, et al. Improving patient outcomes through holistic approaches and collaboration between different specialties in hospital settings. J Popul Ther Clin Pharmacol 2022 Oct 10;29(4):3209-3218. [doi: 10.53555//jptcp.v29i04.5828]
44. Strandberg EL, Ovhed I, Borgquist L, Wilhelmsson S. The perceived meaning of a (w)holistic view among general practitioners and district nurses in Swedish primary care: a qualitative study. BMC Fam Pract 2007 Mar 08;8(1):8 [FREE Full text] [doi: 10.1186/1471-2296-8-8] [Medline: 17346340]
45. Kuhn E, Eftekhari A, Hoffman JE, Crowley JJ, Ramsey KM, Reger GM, et al. Clinician perceptions of using a smartphone app with prolonged exposure therapy. Adm Policy Ment Health 2014 Nov;41(6):800-807 [FREE Full text] [doi: 10.1007/s10488-013-0532-2] [Medline: 24398700]
46. Castor C, Lindkvist R, Hallström IK, Holmberg R. Health care professionals' experiences and views of eHealth in pediatric care: qualitative interview study applying a theoretical framework for implementation. JMIR Pediatr Parent 2023 Oct 18;6(1):e47663 [FREE Full text] [doi: 10.2196/47663] [Medline: 37851500]
47. Lieser T, Huang Y, Sezgin E. The current state of mobile apps owned by large pediatric hospitals in the United States: systematic search and analysis on Google Play and Apple App Stores. JMIR Pediatr Parent 2022 Oct 06;5(4):e38940 [FREE Full text] [doi: 10.2196/38940] [Medline: 36201385]
48. Nievas Soriano BJ, Uribe-Toril J, Ruiz-Real JL, Parrón-Carreño T. Pediatric apps: what are they for? A scoping review. Eur J Pediatr 2022 Apr 20;181(4):1321-1327 [FREE Full text] [doi: 10.1007/s00431-021-04351-1] [Medline: 35048178]
49. Ramsey WA, Heidelberg RE, Gilbert AM, Heneghan MB, Badawy SM, Alberts NM. eHealth and mHealth interventions in pediatric cancer: a systematic review of interventions across the cancer continuum. Psychooncology 2020 Jan;29(1):17-37 [FREE Full text] [doi: 10.1002/pon.5280] [Medline: 31692183]
50. Wen X, Nguyen L, Kowalski-Dobson T, Mohammed L, Woolford S. Employing smartphone health apps in weight management for adolescents: analysis on experiences and perspectives from pediatric providers. Clin Pediatr (Phila) 2022 Oct;61(10):692-698 [FREE Full text] [doi: 10.1177/00099228221097277] [Medline: 35581724]
51. Burmann A, Tischler M, Faßbach M, Schneitler S, Meister S. The role of physicians in digitalizing health care provision: web-based survey study. JMIR Med Inform 2021 Nov 11;9(11):e31527 [FREE Full text] [doi: 10.2196/31527] [Medline: 34545813]
52. Kong T, Scott MM, Li Y, Wichelman C. Physician attitudes towards-and adoption of-mobile health. Digit Health 2020;6:2055207620907187 [FREE Full text] [doi: 10.1177/2055207620907187] [Medline: 32128235]
53. Machleid F, Kaczmarczyk R, Johann D, Balčiūnas J, Atienza-Carbonell B, von Maltzahn F, et al. Perceptions of digital health education among European medical students: mixed methods survey. J Med Internet Res 2020 Aug 14;22(8):e19827 [FREE Full text] [doi: 10.2196/19827] [Medline: 32667899]
54. Tajirian T, Stergiopoulos V, Strudwick G, Sequeira L, Sanches M, Kemp J, et al. The influence of electronic health record use on physician burnout: cross-sectional survey. J Med Internet Res 2020 Jul 15;22(7):e19274 [FREE Full text] [doi: 10.2196/19274] [Medline: 32673234]
55. Robles N, Puig EP, Gómez-Calderón C, Saigí-Rubió F, Cambra GC, Zamora A, et al. Evaluation criteria for weight management apps: validation using a modified Delphi process. JMIR Mhealth Uhealth 2020 Jul 22;8(7):e16899 [FREE Full text] [doi: 10.2196/16899] [Medline: 32706689]
56. Potter A, Munsch C, Watson E, Hopkins E, Kitromili S, O'Neill IC, et al. Identifying research priorities in digital education for health care: umbrella review and modified Delphi method study. J Med Internet Res 2025 Feb 19;27:e66157 [FREE Full text] [doi: 10.2196/66157] [Medline: 39969988]

XSL•FO
RenderX

57. de la Torre E, Montane C, Rubio O, Sampietro-Colom L, Camacho-Mahamud A, Grau-Corral I. Integrating patient perspectives into the digital health technology readiness framework: Delphi study. J Med Internet Res 2025 Aug 20;27(1):e71600 [FREE Full text] [doi: 10.2196/71600] [Medline: 40835414]
58. Nwosu AC, McGlinchey T, Sanders J, Stanley S, Palfrey J, Lubbers P, et al. Identification of digital health priorities for palliative care research: modified Delphi study. JMIR Aging 2022 Mar 21;5(1):e32075 [FREE Full text] [doi: 10.2196/32075] [Medline: 35311674]
59. Erfannia L, Barman MP, Hussain S, Barati R, Arji G. How mobile health affects primary healthcare? Questionnaire design and attitude assessment. Digit Health 2020;6:2055207620942357 [FREE Full text] [doi: 10.1177/2055207620942357] [Medline: 32742715]
60. Jiang Y, West BT, Barton DL, Harris MR. Acceptance and use of eHealth/mHealth applications for self-management among cancer survivors. Stud Health Technol Inform 2017;245:131-135 [FREE Full text] [Medline: 29295067]
61. Kemp E, Trigg J, Beatty L, Christensen C, Dhillon HM, Maeder A, et al. Health literacy, digital health literacy and the implementation of digital health technologies in cancer care: the need for a strategic approach. Health Promot J Austr 2021 Feb;32 Suppl 1:104-114 [FREE Full text] [doi: 10.1002/hpja.387] [Medline: 32681656]
62. Abdelwhab Z, Elshwey R, Adel N, Eid R, Shash E. Evaluation of acceptance and perceived barriers to mobile health applications among breast cancer patients and caregivers in a low- to middle-income country. BMC Digit Health 2025 Sep 05;3(1):58 [FREE Full text] [doi: 10.1186/s44247-025-00203-2]
63. Richards R, Kinnersley P, Brain K, Staffurth J, Wood F. The preferences of patients with cancer regarding apps to help meet their illness-related information needs: qualitative interview study. JMIR Mhealth Uhealth 2019 Jul 31;7(7):e14187 [FREE Full text] [doi: 10.2196/14187] [Medline: 31368446]
64. Steeb T, Wessely A, Mastnik S, Brinker TJ, French LE, Niesert A, et al. Patient attitudes and their awareness towards skin cancer-related apps: cross-sectional survey. JMIR Mhealth Uhealth 2019 Jul 02;7(7):e13844 [FREE Full text] [doi: 10.2196/13844] [Medline: 31267978]
65. Vaffis S, Whaley S, Axon DR, Hall-Lipsy E, Hincapie A, Slack M, et al. Features of cancer mHealth apps and evidence for patient preferences: scoping literature review. JMIR Cancer 2023 Apr 28;9(1):e37330 [FREE Full text] [doi: 10.2196/37330] [Medline: 37115587]
66. Armbruster C, Knaub M, Farin-Glattacker E, von der Warth R. Predictors of adherence to cancer-related mHealth apps in cancer patients undergoing oncological or follow-up treatment—a scoping review. Int J Environ Res Public Health 2022 Oct 21;19(20):13689 [FREE Full text] [doi: 10.3390/ijerph192013689] [Medline: 36294265]
67. LeBaron V, Adhikari A, Bennett R, Chapagain Acharya S, Dhakal M, Elmore CE, et al. A survey of cancer care institutions in Nepal to inform design of a pain management mobile application. BMC Palliat Care 2021 Nov 05;20(1):171 [FREE Full text] [doi: 10.1186/s12904-021-00824-0] [Medline: 34740339]
68. Heaslip V, Richer S, Simkhada B, Dogan H, Green S. Use of technology to promote health and wellbeing of people who are homeless: a systematic review. Int J Environ Res Public Health 2021 Jun 25;18(13):6845 [FREE Full text] [doi: 10.3390/ijerph18136845] [Medline: 34202242]
69. Sidoti O, Dawson W, Gelles-Watnick R, Faverio M, Atske S, Radde K, et al. Mobile fact sheet. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2025-02-25]
70. Kershaw K, Martelly L, Stevens C, McInnes DK, Silverman A, Byrne T, et al. Text messaging to increase patient engagement in a large health care for the homeless clinic: results of a randomized pilot study. Digit Health 2022;8:20552076221129729 [FREE Full text] [doi: 10.1177/20552076221129729] [Medline: 36238754]
71. Polillo A, Gran-Ruaz S, Sylvestre J, Kerman N. The use of eHealth interventions among persons experiencing homelessness: a systematic review. Digit Health 2021 Feb 02;7:2055207620987066 [FREE Full text] [doi: 10.1177/2055207620987066] [Medline: 33598308]
72. Laricchia F. Smartphone ownership in the US by age 2023. Statista. 2024. URL: https://www.statista.com/statistics/489255/percentage-of-us-smartphone-owners-by-age-group/ [accessed 2025-02-25]
73. Jiang S, Hong YA. Mobile-based patient-provider communication in cancer survivors: the roles of health literacy and patient activation. Psychooncology 2018 Mar;27(3):886-891 [FREE Full text] [doi: 10.1002/pon.4598] [Medline: 29193503]
74. Gilstad H, Sand K, Solbjør M, Melby L. Deconstructing (e)health literacy: aspects that promote and inhibit understanding of health information in breast cancer patient pathways. Int J Qual Stud Health Well-being 2022 Dec;17(1):2137961 [FREE Full text] [doi: 10.1080/17482631.2022.2137961] [Medline: 36268568]
75. Kim H, Xie B. Health literacy in the eHealth era: a systematic review of the literature. Patient Educ Couns 2017 Jun;100(6):1073-1082 [FREE Full text] [doi: 10.1016/j.pec.2017.01.015] [Medline: 28174067]
76. Smith B, Magnani JW. New technologies, new disparities: the intersection of electronic health and digital health literacy. Int J Cardiol 2019 Oct 01;292:280-282 [FREE Full text] [doi: 10.1016/j.ijcard.2019.05.066] [Medline: 31171391]
77. Hendawi R, Alian S, Li J. A smart mobile app to simplify medical documents and improve health literacy: system design and feasibility validation. JMIR Form Res 2022 Apr 01;6(4):e35069 [FREE Full text] [doi: 10.2196/35069] [Medline: 35363142]

XSL•FO
RenderX

78. Mueller EL, Cochrane AR, Bennett WE, Carroll AE. A survey of mobile technology usage and desires by caregivers of children with cancer. Pediatr Blood Cancer 2018 Nov 17;65(11):e27359 [FREE Full text] [doi: 10.1002/pbc.27359] [Medline: 30015371]
79. Hasnan S, Aggarwal S, Mohammadi L, Koczwara B. Barriers and enablers of uptake and adherence to digital health interventions in older patients with cancer: a systematic review. J Geriatr Oncol 2022 Nov;13(8):1084-1091 [FREE Full text] [doi: 10.1016/j.jgo.2022.06.004] [Medline: 35752605]
80. Hwang M, Jiang Y. Personalization in digital health interventions for older adults with cancer: a scoping review. J Geriatr Oncol 2023 Nov;14(8):101652 [FREE Full text] [doi: 10.1016/j.jgo.2023.101652] [Medline: 37866009]
81. Nahm E, McQuaige M, Steacy K, Zhu S, Seong H. The impact of a digital cancer survivorship patient engagement toolkit on older cancer survivors' health outcomes. Comput Inform Nurs 2025 Jan 01;43(1):e01199 [FREE Full text] [doi: 10.1097/CIN.0000000000001199] [Medline: 39365650]
82. Newton L, Monkman H, Fullerton C. Exploring older adult cancer survivors' digital information needs: qualitative pilot study. JMIR Cancer 2025 Feb 27;11:e59391 [FREE Full text] [doi: 10.2196/59391] [Medline: 40014827]
83. Paillaud E, Galvin A, Doublet S, Poisson J, Gay P, Perrin C, et al. Health literacy and the use of digital tools in older patients with cancer and their younger counterparts: a multicenter, nationwide study. Patient Educ Couns 2025 Jan;130:108420 [FREE Full text] [doi: 10.1016/j.pec.2024.108420] [Medline: 39303502]
84. Rivière C, Brureau L, Parnot C, Becherirat S, Duverger C, Picchi H, et al. Effectiveness of a digital telemonitoring platform for cancer care of older patients: the ConnectElderlyPatientToDoctor study. Int J Cancer 2023 Feb 01;152(3):504-510 [FREE Full text] [doi: 10.1002/ijc.34196] [Medline: 35770377]
85. Sanders AB, Conroy DE, Schmitz KH, Gusani N. Physical activity and sedentary behavior in older gastrointestinal cancer survivors: need and acceptability of digital health interventions. J Gastrointest Cancer 2018 Jun 18;50:703-708 [FREE Full text] [doi: 10.1007/s12029-018-0128-x] [Medline: 29911290]
86. Verma R, Saldanha C, Ellis U, Sattar S, Haase KR. eHealth literacy among older adults living with cancer and their caregivers: a scoping review. J Geriatr Oncol 2022 Jun;13(5):555-562 [FREE Full text] [doi: 10.1016/j.jgo.2021.11.008] [Medline: 34810146]
87. Wankah P, Chandra S, Lofters A, Mohamednur N, Osei B, Makuwaza T, et al. Improving digital cancer care for older black adults: qualitative study. J Med Internet Res 2025 Feb 19;27:e63324 [FREE Full text] [doi: 10.2196/63324] [Medline: 39970422]
88. Zhou W, Cho Y, Shang S, Jiang Y. Use of digital health technology among older adults with cancer in the United States: findings from a national longitudinal cohort study (2015-2021). J Med Internet Res 2023 May 31;25:e46721 [FREE Full text] [doi: 10.2196/46721] [Medline: 37256672]
89. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989 Sep;13(3):319-340 [FREE Full text] [doi: 10.2307/249008]
90. Nadal C, Sas C, Doherty G. Technology acceptance in mobile health: scoping review of definitions, models, and measurement. J Med Internet Res 2020 Jul 06;22(7):e17256 [FREE Full text] [doi: 10.2196/17256] [Medline: 32628122]
91. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327 [FREE Full text] [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
92. Glasgow RE, Harden SM, Gaglio B, Rabin B, Smith ML, Porter GC, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019 Mar;7:64 [FREE Full text] [doi: 10.3389/fpubh.2019.00064] [Medline: 30984733]
93. Kwan BM, McGinnes HL, Ory MG, Estabrooks PA, Waxmonsky JA, Glasgow RE. RE-AIM in the real world: use of the RE-AIM framework for program planning and evaluation in clinical and community settings. Front Public Health 2019 Nov 22;7:345 [FREE Full text] [doi: 10.3389/fpubh.2019.00345] [Medline: 31824911]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
FDA: Food and Drug Administration
IRB: Institutional Review Board
listserv: list server
mHealth: mobile health
OPENS: Oncology Provider Experiences National Survey
REDCap: Research Electronic Data Capture
RQ: research question
TAM: Technology Acceptance Model

XSL•FO
RenderX

Edited by A Stone, N Cahill; submitted 09.Oct.2025; peer-reviewed by S Ajayi, M Lemonde; comments to author 18.Nov.2025; revised version received 13.Feb.2026; accepted 18.Feb.2026; published 23.Mar.2026.

Please cite as:
Lau N, Srinakarin K, Hong SJH, Aalfs H, Roth ME, Ingram KM, Berkman A, Chan S, Patten J, Iwata W, Taylor MR, Fann JR, Chow EJ, Palermo TM
Mobile Health App Attitudes and Adoption Among Oncology Providers: Cross-Sectional National Survey
J Med Internet Res 2026;28:e85583
URL: https://www.jmir.org/2026/1/e85583
doi:10.2196/85583
PMID:

©Nancy Lau, Kavin Srinakarin, Shannon JH Hong, Homer Aalfs, Michael E Roth, Karly M Ingram, Amy Berkman, Sherilynn Chan, Joanna Patten, Wade Iwata, Mallory R Taylor, Jesse R Fann, Eric J Chow, Tonya M Palermo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patient Sharing of Digital Health Data in the Veterans Health Administration: Cross-Sectional Analysis

Mark S Zocchi[1,2,3], PhD; Jessica M Lipschitz[1,4,5], PhD; Felicia R Bixler[1,6], MS; Stephanie A Robinson[1,2,7], PhD; Bella Etingen[1,8,9], PhD; Timothy P Hogan[1,2,9], PhD; Ndindam Ndiwane[1,2], MPH; Saige Calkins[1,2], MA; Terry Newton[10], MD; Nilesh Shah[10,11], MD; Ben Kragen[1,2], MBA, PhD; Alexander S Young[1,12,13], MD; Stephanie L Shimada[1,2,3,14], PhD

[1]eHealth Partnered Evaluation Initiative, VA Bedford Healthcare System, Bedford, MA, United States

[2]Center for Health Optimization and Implementation Research, VA Bedford Healthcare System, Bedford, MA, United States

[3]Division of Health Informatics and Implementation Science, Department of Population and Quantitative Health Sciences, UMass Chan Medical School, Worcester, MA, United States

[4]Department of Psychiatry, Brigham and Women's Hospital, Boston, MA, United States

[5]Department of Psychiatry, Harvard Medical School, Boston, MA, United States

[6]Center of Innovation for Complex Chronic Healthcare, Edward Hines Jr. VA Hospital, Hines, IL, United States

[7]The Pulmonary Center, Boston University School of Medicine, Boston, MA, United States

[8]Research and Development Service, Dallas VA Medical Center, Dallas, TX, United States

[9]Department of Health Economics, Systems, and Policy, Peter O'Donnell Jr School of Public Health, The University of Texas Southwestern Medical Center, Dallas, TX, United States

[10]Office of Connected Care, Veterans Health Administration, Washington, DC, United States

[11]School of Medicine, University of California San Diego, La Jolla, CA, United States

[12]Mental Illness Research, Education, and Clinical Center, Veterans Affairs Greater Los Angeles Healthcare System, Los Angeles, CA, United States

[13]Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, United States

[14]Department of Health Law, Policy, and Management, Boston University School of Public Health, Boston, MA, United States

Corresponding Author:
Mark S Zocchi, PhD
Center for Health Optimization and Implementation Research
VA Bedford Healthcare System
200 Springs Road, Bldg. 70
Bedford, MA, 01730
United States
Phone: 1 781 687 4737
Email: Mark.Zocchi@va.gov

Abstract

Background: The integration of patient-generated health data (PGHD) into health care has the potential to significantly transform patient care and clinical practice. PGHD includes health-related data created by patients, enabling the collection of health data beyond traditional health care settings. The Veterans Health Administration (VA) has taken proactive steps to incorporate PGHD into health care through its Share My Health Data (SMHD) mobile app. Launched in 2023, the SMHD app allows veterans to securely share data from their personal digital health devices with the VA for clinical and research use. However, data characterizing patients who use such tools in real-world health care systems are lacking, creating an evidence gap for implementing PGHD-informed care equitably.

Objective: This study aimed to identify the characteristics of patients using the VA SMHD mobile app, which allows veterans to share PGHD with the VA.

Methods: We conducted a cross-sectional analysis of veterans who began using SMHD between October 2023 and September 2024 (n=3157, "SMHD users"). We collected demographic information, including age, sex, race/ethnicity, and rurality, and clinical information, including physiological and mental health conditions, from VA administrative data. We compared characteristics of SMHD users to a 10% random sample of veterans from the same underlying administrative data cohort that

XSL•FO
RenderX

had never used the app (n=632,187, “nonusers”). Statistical analyses were performed using chi-square tests, independent t tests, and multivariable regression to assess the relationship between use and key characteristics.

Results: Middle-aged veterans were more likely to be SMHD users (40-49 years: odds ratio [OR] 1.55, *P*<.001; 50-59 years: OR 1.37, *P*<.001), while veterans aged 60 years and over were less likely (60-69 years: OR 0.72, *P*<.001; ≥70 years: OR 0.24, *P*<.001). Female (OR 1.23, *P*<.001) and married (OR 1.31, *P*<.001) veterans were more likely to be SMHD users than male and unmarried veterans. In contrast, Black or African American (OR 0.62, *P*<.001) and rural (OR 0.82, *P*<.001) veterans were less likely to be SMHD users than White and urban veterans. Veterans in higher-income zip codes (OR 1.36, *P*<.001) were more likely to have used the app than those in lower-income zip codes. Clinically, SMHD users were more likely to have a service-connected disability (OR 1.81, *P*<.001), multiple physiological conditions (OR 1.86, *P*<.001), and multiple mental health diagnoses (OR 1.35, *P*<.001) versus none.

Conclusions: Veterans who used the SMHD app differed significantly from nonusers across several demographic and clinical characteristics. These insights identify specific demographic and clinical subgroups with higher and lower app adoption, providing an evidence base to inform targeted implementation and outreach and support strategies to promote enhanced engagement in PGHD-informed care.

(J Med Internet Res 2026;28:e80517) doi:10.2196/80517

KEYWORDS

patient-generated health data; veterans; mobile health; mHealth; digital phenotyping; wearable electronic devices

Introduction

The health care landscape is rapidly evolving with the integration of digital patient-generated health data (PGHD) into medical practice [1]. The Office of the National Coordinator for Health Information Technology defines PGHD as “health-related data created, recorded, or gathered by or from patients (or family members or other caregivers) to help address a health concern” [2]. PGHD can include health-related data, such as physical activity, vital signs, sleep, and other physiological metrics collected through wearable devices (eg, smartwatches), mobile health apps (eg, Apple Health), and other digital devices (eg, Bluetooth-enabled blood pressure cuffs).

The use of digital PGHD is part of a broader movement, sometimes called digital phenotyping [3] or personal sensing [4], to understand how to use data from personal digital devices to improve health care outcomes and personalization of treatment. Capturing health data in patients’ everyday environments has the potential to positively impact care delivery and health outcomes. These data provide frequent, objective insights into patients’ health status, offering a more complete view of health and well-being than periodic clinical encounters. This may allow for more personalized, proactive, and responsive care, especially for management of chronic conditions and preventive care [5]. In addition, PGHD has the potential to enhance patient engagement in self-management. Researchers have found many use cases for PGHD, such as predicting onset of new symptoms [6], monitoring chronic disease [7], detection of physiological and psychological conditions [8], and predicting response to treatment [9].

Despite the many use cases and opportunities offered by PGHD, health care system implementation of PGHD-informed care has been limited. Survey research suggests that many patient populations (eg, psychiatric, arthritis) have interest in digital symptom monitoring [10-12]. However, we do not yet know which patients may be most likely to actually use digital tools that support the collection of PGHD. Understanding the characteristics of these users could help app developers and clinical leaders identify what clinical populations to target for initial PGHD-informed care initiatives and also offer insight into ways to increase uptake in patient populations that show lower propensity to use these kinds of tools. Therefore, characterizing patients who adopt PGHD-sharing tools within health care systems is a critical piece of implementation science necessary for guiding rollout and tailoring support strategies.

To move toward implementing PGHD-informed care, the Veterans Health Administration (VA) launched the Share My Health Data (SMHD) mobile app in 2023 [13]. The SMHD app allows veterans to voluntarily share data from personal digital health devices, such as Bluetooth-enabled glucometers and third-party health apps, such as Apple Health, Fitbit, and Garmin [14]. Once paired with the SMHD app, veterans’ PGHD are securely stored in a dedicated VA PGHD database [15]. The contents of this database can, in turn, be applied in support of clinical care and related research. For example, the VA is currently pilot-testing the integration of these data into clinical dashboards, allowing providers to view PGHD directly from their patients during or between appointments.

VA’s SMHD app offers a unique opportunity to examine the passive digital health–monitoring initiative within a large, integrated health care system. This study compared the demographic characteristics, clinical conditions, and health care costs of SMHD users to a nationally representative random sample of veterans who have not used the app. By describing the demographic and clinical profile of SMHD users and comparing them to the broader VA population, we aimed to generate foundational evidence to guide targeted engagement strategies, identify clinical populations for pilot programs, and highlight potential disparities in digital health tool adoption. These insights can inform efforts to broaden patient participation in PGHD-informed care and support decisions about which clinical populations may be most appropriate for early implementation within the VA and in other health care systems.

XSL•FO
RenderX

Methods

Design

This was a retrospective, observational, cross-sectional analysis using routinely collected data from national VA data sources. This evaluation focused on SMHD users and compared characteristics of those who used the app to those who did not. The main objective was descriptive, with no intervention assignment or randomization.

Setting

VA's Office of Connected Care launched the SMHD app on May 1, 2023. The app was added to the VA App Store, featured on its home page, and made available to download on iOS and Android devices [16]. After downloading and accepting the end-user agreement, users can pair their digital health devices through VA's secure authentication process. Other than instructional materials provided on the VA Mobile App Store, no additional training, outreach, or incentives were provided.

Cohort Selection

We examined veterans who began using the app shortly after it was launched, between October 1, 2023, and September 30, 2024 ("SMHD users") and compared them to a random sample of veterans who did not use the app prior to or during this same time period ("nonusers").

Our analytic cohort included SMHD users, who we identified as having at least one measurement in the PGHD database submitted through the SMHD app (N=4202). This definition of "use" indicates SMHD app download, linkage to a digital health device, and initiation of data sharing; it does not measure intensity or continuity of use over time. Exclusions were nonveterans (n=62, 1.5%), those not included in the fiscal year 2023 VA Nosos risk score dataset (n=318, 7.6%) [17], or participants in early field tests of the SMHD app (n=665, 15.8%). Early field test participants were excluded to focus this evaluation on patients who discovered, downloaded, and began using the app on their own, independent of recruitment into a controlled field test.

The comparison group of nonusers was selected using a 10% random sample of veterans who were included in the fiscal year 2023 VA Nosos risk score dataset and who had never used the SMHD app prior to September 30, 2024, the end of the evaluation period. Of note, all veterans who received care in VA during fiscal year 2023 (October 1, 2022-September 30, 2023) are included in the fiscal year 2023 VA Nosos risk score dataset. This sampling approach ensured all veterans in the sample received VA or VA-paid care and allowed for sufficient statistical power to detect differences between SMHD users and nonusers, while maintaining feasibility for data processing and analysis. The final cohort consisted of 3157 SMHD users and 632,187 nonusers.

Data Sources

We obtained SMHD app usage data from VA's PGHD database. PGHD were extracted via secure SQL queries and validated for appropriate SMHD source linkage and timestamp accuracy. PGHD with missing timestamps or insufficient source metadata were removed prior to cohort selection. Patient demographics and clinical encounter data were obtained from the VA's Corporate Data Warehouse (CDW). Zip code–level household income data came from the US Census Bureau American Community Survey (2023 5-year estimates) [18].

Preuse "baseline" measures of comorbidity and cost came from the fiscal year 2023 VA Nosos risk score dataset, which includes indicators for chronic physiological and mental health conditions based on the Centers for Medicare and Medicaid Services' Hierarchical Condition Categories (HCC) [19] and a psychiatric case-mix system (PsyCMS) developed by Sloan et al [20] and later updated for *International Classification of Diseases* (ICD)-10 diagnosis codes [21]. The Nosos dataset uses diagnosis data (ICD-9 or ICD-10) from VA outpatient encounters and inpatient stays during the entire fiscal year, including VA-purchased community care data, to calculate comorbidities. Comorbidity data are combined with use data to calculate an overall "score," which is standardized to center around 1, so a value of 1 means that the veteran is expected to have costs that are equivalent to the national average for VA patients.

Variables

Variables were selected based on availability in the electronic medical record and prior work examining adoption and use of other digital health technologies in VA [22-24]. Demographic variables included age, race/ethnicity, sex, marital status, rurality, zip code–level household income, and service-connected disability. A service-connected disability refers to a physical or mental health condition that was incurred or aggravated during active military service and is formally recognized by VA. Clinical variables included counts of physiological and mental health conditions (eg, 0, 1, ≥2). To avoid overlap with the mental health counts, we excluded HCC codes from the overall physiological condition counts that were for mental health conditions (eg, alcohol use disorder, schizophrenia, major depressive disorder). For the mental health counts, we excluded PsyCMS codes indicating conditions in remission (eg, alcohol use remission). Relative cost outcomes were estimated using the fiscal year 2023 Nosos concurrent risk score [19].

Data Analysis

We compared SMHD users with the random sample of nonusers across all variables of interest using means (SDs), frequencies, and percentages. Frequencies of categorical variables were compared using chi-square tests. Means of continuous variables were compared using independent *t* tests. Standardized mean differences (SMDs) were calculated for all variables to aid interpretation of the magnitude of differences between groups, with an absolute value of >0.1 conventionally considered a small but meaningful difference.

A multivariable logistic regression was used to estimate odds ratios (ORs) associated with being an SMHD user, with SEs adjusted for potential clustering of veterans within VA regions. Because our objective was descriptive and did not have an a priori hypothesis about effect moderation, we did not include any interaction terms. To avoid multicollinearity, we checked covariates' variance inflation factor and included a term for

XSL•FO
RenderX

total chronic condition counts (eg, 0, 1, ≥2) but excluded individual comorbidities in the model. As a conventional check for model fit, we reported the Hosmer-Lemeshow goodness-of-fit test (10 groups). A complete case analysis was used for the multivariable model. The proportion of missing data was small (5%), and sensitivity analyses using alternative missing data methods were not conducted.

All analyses were conducted using R version 4.4.1 (R Foundation for Statistical Computing). Key packages included *stats* for regression modeling and *dplyr* for data manipulation.

Ethical Considerations

This work was reviewed by the Institutional Review Boards of the VA Bedford Healthcare System in Bedford, Massachusetts, and the Edward Hines Jr VA Hospital in Hines, Illinois, and designated as program evaluation for quality improvement purposes, exempting it from further oversight (VA Handbook 1058.05).

Results

Characteristics of SMHD Users

SMHD users differed from nonusers in several ways (Table 1). SMHD users were younger than nonusers by 9 years (users: mean 53.2, SD 13.4; nonusers: mean 62.3, SD 17.2, years), and a greater proportion were female (users: n=539, 17.1%; nonusers: n=67,401, 10.7%). In addition, race/ethnicity differed between SMHD users and nonusers. A larger proportion of SMHD users were Hispanic or Latino (users: n=322, 10.2%; nonusers: n=47,265, 7.5%), while a smaller proportion were non-Hispanic Black or African American (users: n=477, 15.1%; nonusers: n=109,459, 17.3%). Furthermore, a smaller proportion of SMHD users lived in rural areas (users: n=914, 29%; nonusers: n=214,608, 33.9%), while a larger proportion were married (users: n=1878, 59.5%; nonusers: n=349,908, 55.3%), had a service-connected disability (users: n=2791, 88.4%; nonusers: n=460,129, 72.8%), and, on average, lived in higher-income zip codes (mean household income for users: US $79,731; nonusers: US $75,071).

Table 2 presents a comparison of health conditions between SMHD users and nonusers. Clinically, SMHD users had a significantly higher burden of mental health conditions, such as posttraumatic stress disorder, depression, anxiety, adjustment disorder, sleep disorder, attention deficit disorder, and mood disorder. In contrast, they had a lower incidence of several major physiological conditions, such as chronic obstructive pulmonary disorder, heart arrhythmias, and congestive heart failure (all $P<.001$), though they showed higher rates of obesity and arthritis. Rates of alcohol dependence or abuse did not differ significantly between SMHD users and nonusers; however, rates of nicotine dependence were slightly lower among SMHD users (users: n=358, 11.3%; nonusers: n=86,180, 13.6%; $P<.001$). In terms of total VA health care costs, SMHD users had higher mean Nosos risk scores than nonusers (users: 1.13, SD 1.34; nonusers: 1.05, SD 1.64; $P=.005$), indicating VA health care costs were about 13% higher than the VA population average for SMHD users.

XSL•FO
RenderX

Table 1. Unadjusted differences in demographic characteristics among SMHD[a] users and nonusers.

Characteristics	SMHD users (n=3157)	Nonusers (n=632,187)	SMD[b]	*P* value
Age (years), mean (SD)	53.2 (13.4)	62.3 (17.2)	0.589	<.001
Age group (years), n (%)			0.716	<.001
<40	538 (17.0)	86,458 (13.7)	—[c]	—
40-49	765 (24.2)	72,881 (11.5)	—	—
50-59	852 (27.0)	90,594 (14.3)	—	—
60-69	561 (17.8)	117,407 (18.6)	—	—
>69	441 (14.0)	264,842 (41.9)	—	—
Female sex, n (%)	539 (17.1)	67,401 (10.7)	0.186	<.001
Race/ethnicity, n (%)			0.116	<.001
White, not Hispanic	2016 (63.9)	402,055 (63.6)	—	—
Black or African American, not Hispanic	477 (15.1)	109,459 (17.3)	—	—
Hispanic	322 (10.2)	47,265 (7.5)	—	—
Other race, not Hispanic	107 (3.4)	19,222 (3.0)	—	—
Unknown	235 (7.4)	54,186 (8.6)	—	—
Rural residence n (%)	914 (29.0)	214,608 (33.9)	0.122	<.001
Zip code median income (US $), mean (SD)	79,731 (27,953)	75,071 (26,646)	0.171	<.001
Zip code median income (US $) category, n (%)			0.169	<.001
<50,000	328 (10.4)	83,355 (13.2)	—	—
50,000-69,999	1011 (32.0)	221,976 (35.1)	—	—
70,000-89,999	828 (26.6)	157,757 (25.0)	—	—
≥90,000	948 (30.0)	146,026 (23.1)	—	—
Married, n (%)	1878 (59.5)	349,908 (55.3)	0.083	<.001
Service-connected disability, n (%)	2791 (88.4)	460,129 (72.8)	0.403	<.001

[a]SMHD: Share My Health Data.

[b]SMD: standardized mean difference.

[c]Not applicable.

XSL•FO
RenderX

Table 2. Unadjusted differences in health conditions among SMHD[a] users and nonusers.

Health conditions	SMHD users (n=3157), n (%)	Nonusers (n=632,187), n (%)	SMD[b]	*P* value
Nosos Risk Score, Mean (SD)	1.13 (1.34)	1.05 (1.64)	0.055	.005
Physiological conditions (15 most common)				
Diabetes with chronic complications	450 (14.3)	88,718 (14.0)	0.006	.74
Diabetes without complication	360 (11.4)	65,279 (10.3)	0.035	.05
Chronic obstructive pulmonary disease	185 (5.9)	63,720 (10.1)	0.156	<.001
Specified heart arrhythmias	235 (7.4)	62,671 (9.9)	0.088	<.001
Congestive heart failure	154 (4.9)	45,954 (7.3)	0.100	<.001
Vascular disease	158 (5.0)	44,802 (7.1)	0.087	<.001
Breast, prostate, and other cancers and tumors	96 (3.0)	25,993 (4.1)	0.058	.003
Chronic kidney disease, moderate (stage 3)	94 (3.0)	24,418 (3.9)	0.049	.01
Morbid obesity	308 (9.8)	24,068 (3.8)	0.238	<.001
Acute renal failure	68 (2.2)	20,267 (3.2)	0.065	.001
Rheumatoid arthritis and inflammatory connective tissue disease	149 (4.7)	20,011 (3.2)	0.080	<.001
Cardiorespiratory failure and shock	43 (1.4)	17,325 (2.7)	0.097	<.001
Coagulation defects and other specified hematological disorders	80 (2.5)	17,161 (2.7)	0.011	.57
Other significant endocrine and metabolic disorders	80 (2.5)	15,280 (2.4)	0.008	.71
Ischemic or unspecified stroke	56 (1.8)	14,715 (2.3)	0.039	.05
Mental health conditions (15 most common)				
Posttraumatic stress disorder	775 (24.5)	103,665 (16.4)	0.203	<.001
Nicotine dependence	358 (11.3)	86,180 (13.6)	0.069	<.001
Major depressive disorder, recurrent	474 (15.0)	56,132 (8.9)	0.19	<.001
Anxiety, not elsewhere classified	316 (10.0)	41,227 (6.5)	0.127	<.001
Adjustment disorder	339 (10.7)	39,001 (6.2)	0.165	<.001
Generalized anxiety	306 (9.7)	32,517 (5.1)	0.174	<.001
Other mental health code	124 (3.9)	22,660 (3.6)	0.018	.32
Major depressive disorder, single episode	160 (5.1)	20,080 (3.2)	0.095	<.001
Alcohol dependence	90 (2.9)	19,126 (3.0)	0.01	.60
Alcohol abuse	69 (2.2)	16,206 (2.6)	0.025	.20
Dementia	8 (0.3)	14,143 (2.2)	0.18	<.001
Sleep disorders	138 (4.4)	13,647 (2.2)	0.125	<.001
Attention deficit disorder	216 (6.8)	13,488 (2.1)	0.229	<.001
Mood disorder, not elsewhere classified	103 (3.3)	12,369 (2.0)	0.082	<.001
Bipolar disorders	104 (3.3)	10,730 (1.7)	0.103	<.001
Number of physiological conditions			0.065	.001
0	1500 (47.5)	306,388 (48.5)	—[c]	—
1	831 (26.3)	149,436 (23.6)	—	—
≥2	826 (26.2)	176,363 (27.9)	—	—
Number of mental health conditions			0.324	<.001
0	1171 (37.1)	323,307 (51.1)	—	—
1	695 (22.0)	138,703 (21.9)	—	—

XSL•FO
RenderX

Health conditions	SMHD users (n=3157), n (%)	Nonusers (n=632,187), n (%)	SMD[b]	*P* value
≥2	1291 (40.9)	170,177 (26.9)	—	—

[a]SMHD: Share My Health Data.

[b]SMD: standardized mean difference.

[c]Not applicable.

Predictors of SMHD App Use

The results from the multivariable logistic regression model are presented in Table 3. Of the 635,344 patients in the cohort, 95% (n=600,590) had complete data for the multivariable model. The Hosmer-Lemeshow test indicated no evidence of lack of fit (*P*=.19), and the variance inflation factor values ranged from 1.1 to 1.5, which is significantly below the commonly used threshold of 5 to identify multicollinearity [25]. Several characteristics were significantly associated with reduced odds of being an SMHD user. Compared to younger veterans (<40 years), older veterans (>60 years) were less likely to be SMHD users (60-69 years: OR 0.72, 95% CI 0.63-0.83; ≥70 years: OR 0.24, 95% CI 0.21-0.28). Compared to White non-Hispanic veterans, Black non-Hispanic (OR 0.62, 95% CI 0.56-0.69) and other race/ethnicity (OR 0.79, 95% CI 0.65-0.97) veterans were less likely to be SMHD users. Veterans living in rural areas (OR 0.82, 95% CI 0.75-0.89) were less likely to be SMHD users compared to veterans living in more urban areas.

Conversely, several characteristics were found to be associated with increased odds of being an SMHD user. Compared to younger veterans (<40 years), middle-aged veterans were more likely to be SMHD users (40-49 years: OR 1.55, 95% CI 1.38-1.73; 50-59 years: OR 1.37, 95% CI 1.22-1.54). Female veterans (OR 1.23, 95% CI 1.11-1.35) were more likely to be SMHD users compared to male veterans, and married veterans (OR 1.31, 95% CI 1.22-1.41) were more likely to use SMHD than unmarried veterans (ie, divorced, separated, widowed, or single). In addition, compared to veterans living in zip codes with mean incomes less than US $50,000, veterans living in higher-income zip codes were more likely to be SMHD users (US $70,000-$89,999: OR 1.15, 95% CI 1.01-1.31; ≥US $90,000: OR 1.36, 95% CI 1.19-1.55). Finally, having a service-connected disability (OR 1.81, 95% CI 1.61-2.04), having one (OR 1.54, 95% CI 1.41-1.68) or two or more (OR 1.86, 95% CI 1.68-2.05) physiological conditions (vs none), and having two or more mental health conditions (OR 1.35, 95% CI 1.24-1.47) versus none were all associated with increased odds of being an SMHD user.

XSL•FO
RenderX

Table 3. Multivariable logistic regression model predicting odds of being an SMHD[a] user (N=600,590).[b]

Characteristics	OR[c] (95% CI)	*P* value
Age group (years; reference: <40 years)		
40-49	1.55 (1.38-1.73)	<.001
50-59	1.37 (1.22-1.54)	<.001
60-69	0.72 (0.63-0.83)	<.001
>69	0.24 (0.21-0.28)	<.001
Female (reference: male)	1.23 (1.11-1.35)	<.001
Race/ethnicity (reference: White)		
Black or African American	0.62 (0.56-0.69)	<.001
Hispanic	0.94 (0.83-1.07)	.35
Other	0.79 (0.65-0.97)	.02
Unknown	0.82 (0.71-0.95)	.01
Rural (reference: urban)	0.82 (0.75-0.89)	<.001
Zip code median income category (reference: <US $50,000)		
50,000-69,999	1.06 (0.93-1.20)	.36
70,000-89,999	1.15 (1.01-1.31)	.04
≥90,000	1.36 (1.19-1.55)	<.001
Married (reference: not married)	1.31 (1.22-1.41)	<.001
Service-connected disability	1.81 (1.61-2.04)	<.001
Number of physiological conditions (reference: 0)		
1	1.54 (1.41-1.68)	<.001
≥2	1.86 (1.68-2.05)	<.001
Number of mental health conditions (reference: 0)		
1	1.09 (0.99-1.20)	.09
≥2	1.35 (1.24-1.47)	<.001

[a]SMHD: Share My Health Data.

[b]Note that 34,754 (5%) patients were excluded from the model due to missing data. SEs were adjusted for clustering at the Veterans Health Administration (VA) regional level. Hosmer-Lemeshow goodness-of-fit test: χ^2_8=11.2, *P*=.19.

[c]OR: odds ratio.

Discussion

Principal Findings

PGHD-guided care has the potential to be the future of medicine. Effective future implementation requires understanding which patients are open to using digital devices and other tools to collect and share PGHD. To the best of our knowledge, this is the first evaluation to look at users of a mobile app intended to support the collection and sharing of PGHD across a health care system. We found that SMHD users are more likely to be female, be married, and live in a higher-income zip code and are less likely to be aged 60 years and over, be Black or from another race, or live in a rural area. We also found that SMHD users have higher health care costs and more medical and mental health comorbidities.

Our findings about the demographic characteristics of SMHD users are similar to the literature on users of other digital health technologies. Studies on the use of other digital health technologies have also found that women and higher-income and urban-dwelling patients are more likely to use digital health technologies [26-30]. Studies have also found that patients who are from minority races (in this case non-Hispanic Black and other races), older, and divorced or separated are less likely to use digital health technologies [31-33]. These patterns may reflect broader determinants, such as differential access to technology and broadband [29,30], variations in digital health literacy, and potential questions of trust in data-sharing initiatives [31]. As PGHD tools and PGHD-informed care initiatives are designed and implemented over the coming years, it will be important to consider how to make these tools accessible to all patients to promote health and reduce health care disparities.

In addition, these findings could inform future decisions regarding what clinical populations could be appropriate targets for initial pilots of PGHD-guided care models. In this study,

XSL•FO
RenderX

SMHD users were more likely to have, for example, arthritis, morbid obesity, multiple comorbidities, and mental health diagnoses. These findings corroborate those from survey studies indicating high interest in monitoring apps and sensing technologies among these populations [10-12], but until now, data on actual uptake have been limited. Thus, our findings can help inform health care systems about what patient populations to consider first, as they start to develop and test PGHD-informed care models.

Our findings indicate that SMHD users have higher health care costs and greater rates of both mental and physiological conditions compared to the general VA population. The association between app use and higher clinical complexity suggests that SMHD adopters represent a subgroup that could stand to benefit most from more precise, personalized, and proactive PGHD-informed care models. This presents a compelling opportunity to evaluate whether integrating PGHD into routine care can improve outcomes and optimize resource use within this more complex patient population.

As health care systems increasingly adopt PGHD tools, it will be important to ensure that implementation strategies account for differences in patient characteristics and technology use. Our findings highlight demographic and clinical differences between SMHD users and nonusers, which can inform targeted outreach and support efforts. Future work should explore barriers that may limit use among certain patient groups, and identify approaches to encourage broader engagement with PGHD tools across the VA population. For example, future implementation strategies for SMHD and similar tools could integrate targeted support, such as peer-led digital coaching for older adults, partnerships with rural community centers to facilitate access, and explicit messaging addressing data privacy concerns.

Limitations

This analysis was limited to characteristics found in the VA electronic health record. For the multivariable models, we only used complete case data, which may introduce bias if data are not completely missing at random; however, the amount of missing data was relatively small (5%). Our focus was on SMHD users for sharing PGHD, and we did not evaluate the frequency, quality, or validity of the shared data.

In addition, this paper describes the use of the SMHD app before clinical displays or broader PGHD-informed care models have been systematically implemented in VA clinical settings. The connection to care received, and, consequently, the value of sharing PGHD, is likely different at this stage of implementation than it would be in the context of provider recommendation and systematic care team engagement with the PGHD shared by patients. In addition, as implementation efforts move forward, organizational factors may play a role in adoption and use and will be important to consider in future work.

It is also important to keep in mind that this was an exploratory cross-sectional analysis, and as such, it does not make any causal claims about SMHD use and subsequent health outcomes. By design, the comorbidity and cost measures represent a preuse baseline status and therefore do not capture any changes that may have occurred after the study period began. Although we used a large nationwide sample and included several relevant covariates in our model, we cannot eliminate the possibility of unobserved confounders. For example, smartphone ownership, internet access, and perceptions of data privacy and security are not available in VA administrative data. Feasibility and acceptability were inferred from use of the SMHD app; however, no direct user feedback was collected. Finally, although this evaluation draws on a large, nationwide sample of veterans, the findings may have limited generalizability to non-VA health care settings or populations.

Conclusion

This national evaluation of the VA's SMHD app reveals that app adopters are more likely to be middle-aged, female, married, urban-dwelling veterans with service-connected disabilities and multiple health conditions, while being less likely to be veterans who are older, Black, or from rural or lower-income areas. These findings provide the first large-scale, real-world evidence of who actively shares PGHD within an integrated health care system, moving beyond prior studies that relied primarily on patient self-report.

Understanding these adoption patterns is critical for implementation initiatives. They identify natural adopter populations for pilot programs and highlight demographic groups at risk for digital exclusion, informing the design of equitable engagement strategies. As PGHD integration evolves, ongoing monitoring of adoption trends and focused research on barriers will be essential to ensure these transformative tools benefit all patients.

Finally, organizational factors, such as local clinic promotion strategies, provider endorsement, and the degree of PGHD integration into clinical workflows, may significantly influence future adoption. Addressing these factors alongside patient-level disparities will be key to achieving scalable PGHD-informed care across VA and other health care systems.

Acknowledgments

The views expressed in this paper are those of the authors and do not necessarily reflect the position or policy of the US Department of Veterans Affairs or of the US government. This work was supported by the US Department of Veterans Affairs, Office of Connected Care and Office of Research and Development, Health Services Research and Development Service, Quality Enhancement Research Initiative Program (PEC 15–470; principal investigator: TPH). The authors of this paper are affiliated with both the Office of Connected Care and the Quality Enhancement Research Initiative Program.

XSL•FO
RenderX

Data Availability

The US Department of Veterans Affairs prohibits unauthorized sharing of data. The data used for this evaluation are not permitted to be shared outside the Veterans Health Administration (VA) firewall without a Data Use Agreement. This limitation is consistent with other studies based on VA data; however, VA data are available to researchers behind the VA firewall with an approved VA study protocol. For more information, please visit VA Information Resource Center (VIReC) [34].

Authors' Contributions

All authors have made substantial contributions to the conception or design of the work. MSZ and JML wrote the initial draft, and FRB, SAR, BE, TPH, SC, TN, NS, and SLS contributed to substantive revisions. MSZ and NN conducted the statistical analysis and managed the data. TPH, BE, TN, NS, BK, ASY, and SLS critically reviewed and provided important intellectual content to the final draft. All authors were involved in data interpretation and approved the final version submitted for publication. MSZ had access to the data in the evaluation and takes responsibility for data integrity and accuracy.

Conflicts of Interest

JML has a consulting relationship with Sway Behavioral Health. Additionally, she receives grant funding from Alkermes, Inc. and BD2 (Breakthrough Discoveries for thriving with Bipolar Disorder).

References

1. Smuck M, Odonkor CA, Wilt JK, Schmidt N, Swiernik MA. The emerging clinical role of wearables: factors for successful implementation in healthcare. NPJ Digit Med 2021 Mar 10;4(1):45 [FREE Full text] [doi: 10.1038/s41746-021-00418-3] [Medline: 33692479]
2. Patient-generated health data. Office of the National Coordinator for Health Information Technology. URL: https://www.healthit.gov/topic/scientific-initiatives/pcor/patient-generated-health-data-pghd [accessed 2025-05-14]
3. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. NPJ Digit Med 2019 Sep 6;2(1):88 [FREE Full text] [doi: 10.1038/s41746-019-0166-1] [Medline: 31508498]
4. Mohr DC, Shilton K, Hotopf M. Digital phenotyping, behavioral sensing, or personal sensing: names and transparency in the digital age. NPJ Digit Med 2020 Mar 25;3(1):45 [FREE Full text] [doi: 10.1038/s41746-020-0251-5] [Medline: 32219186]
5. Nazi KM, Newton T, Armstrong CM. Unleashing the potential for patient-generated health data (PGHD). J Gen Intern Med 2024 Feb 22;39(Suppl 1):9-13 [FREE Full text] [doi: 10.1007/s11606-023-08461-4] [Medline: 38252246]
6. Quer G, Radin JM, Gadaleta M, Baca-Motes K, Ariniello L, Ramos E, et al. Wearable sensor data and self-reported symptoms for COVID-19 detection. Nat Med 2021 Jan 29;27(1):73-77 [FREE Full text] [doi: 10.1038/s41591-020-1123-x] [Medline: 33122860]
7. Gashi S, Oldrati P, Moebus M, Hilty M, Barrios L, Ozdemir F, PHRT Consortium, et al. Modeling multiple sclerosis using mobile and wearable sensor data. NPJ Digit Med 2024 Mar 11;7(1):64 [FREE Full text] [doi: 10.1038/s41746-024-01025-8] [Medline: 38467710]
8. Chakrabarti S, Biswas N, Jones LD, Kesari S, Ashili S. Smart consumer wearables as digital diagnostic tools: a review. Diagnostics (Basel) 2022 Aug 31;12(9):2110 [FREE Full text] [doi: 10.3390/diagnostics12092110] [Medline: 36140511]
9. Masoumian Hosseini M, Masoumian Hosseini ST, Qayumi K, Hosseinzadeh S, Sajadi Tabar SS. Smartwatches in healthcare medicine: assistance and monitoring; a scoping review. BMC Med Inform Decis Mak 2023 Nov 03;23(1):248 [FREE Full text] [doi: 10.1186/s12911-023-02350-w] [Medline: 37924029]
10. Torous J, Chan SR, Yee-Marie Tan S, Behrens J, Mathew I, Conrad EJ, et al. Patient smartphone ownership and interest in mobile apps to monitor symptoms of mental health conditions: a survey in four geographically distinct psychiatric clinics. JMIR Ment Health 2014 Dec;1(1):e5 [FREE Full text] [doi: 10.2196/mental.4004] [Medline: 26543905]
11. Barrera A, Ghenciulescu A, Owens N, Cortez A, De Giorgi R, Cowen P, et al. A survey of preferences for sensing technologies in people with severe mental illness admitted to an acute psychiatric unit. Ann Biomed Eng 2025 Sep 25;53(9):2372-2383 [FREE Full text] [doi: 10.1007/s10439-025-03769-0] [Medline: 40560456]
12. Bergmann J, Chandaria V, McGregor A. Wearable and implantable sensors: the patient's perspective. Sensors 2012 Dec 05;12(12):16695-16709 [FREE Full text] [doi: 10.3390/s121216695] [Medline: 23443394]
13. Hogan TP, Sherman SE, Dardashti N, McMahon N, Slightam C, Zulman DM. Realizing virtual care in VA: supporting the healthcare system's journey towards enhanced access, engagement, and outcomes. J Gen Intern Med 2024 Feb 23;39(Suppl 1):1-4 [FREE Full text] [doi: 10.1007/s11606-024-08618-9] [Medline: 38393612]
14. Share My Health Data. U.S. Department of Veterans Affairs. URL: https://mobile.va.gov/app/share-my-health-data [accessed 2025-05-14]
15. Newton TJ, Shah N, Lewis K, Zocchi MS, Bixler FR, Etingen B, et al. Developing infrastructure to realize the value of patient-generated health data in a large integrated health care system: the Veterans Health Administration experience. J Med Internet Res 2025 Jun 06;27:e70755-e70755 [FREE Full text] [doi: 10.2196/70755] [Medline: 40479588]
16. VA App Store. U.S. Department of Veterans Affairs. URL: https://mobile.va.gov/appstore [accessed 2025-11-20]

XSL•FO
RenderX

17. Wagner T, Moran E, Shen M, Gehlert E. HERC's guide to the Nosos risk adjustment score. Health Economics Resource Center, U.S. Department of Veterans Affairs. 2024 May. URL: https://www.herc.research.va.gov/include/page.asp?id=guidebook-nosos [accessed 2025-05-14]
18. American community survey 5-year data (2009-2023). United States Census Bureau. URL: https://www.census.gov/data/developers/data-sets/acs-5year.html [accessed 2025-05-19]
19. Nosos risk adjustment. U.S. Department of Veterans Affairs. URL: https://www.herc.research.va.gov/include/page.asp?id=risk-adjustment [accessed 2025-05-19]
20. Sloan K, Montez-Rath M, Spiro A, Christiansen CL, Loveland S, Shokeen P, et al. Development and validation of a psychiatric case-mix system. Med Care 2006 Jun;44(6):568-580 [FREE Full text] [doi: 10.1097/01.mlr.0000215819.76050.a1] [Medline: 16708006]
21. Wagner T, Gehlert E, Rosen A. Updating the psychiatric case mix system (PsyCMS) mental health and substance use grouper for ICD-10-CM. Technical Report 31. Health Economics Resource Center, U.S. Department of Veterans Affairs. 2016 Nov. URL: https://www.herc.research.va.gov/include/page.asp?id=technical-report-psycms-icd10 [accessed 2025-05-14]
22. Hogan TP, Etingen B, McMahon N, Bixler FR, Am L, Wacks RE, et al. Understanding adoption and preliminary effectiveness of a mobile app for chronic pain management among US military veterans: pre-post mixed methods evaluation. JMIR Form Res 2022 Jan 20;6(1):e33716 [FREE Full text] [doi: 10.2196/33716] [Medline: 35049515]
23. Lipschitz J, Miller CJ, Hogan TP, Burdick KE, Lippin-Foster R, Simon SR, et al. Adoption of mobile apps for depression and anxiety: cross-sectional survey study on patient interest and barriers to engagement. JMIR Ment Health 2019 Jan 25;6(1):e11334 [FREE Full text] [doi: 10.2196/11334] [Medline: 30681968]
24. Griffin A, Troszak L, Van Campen J, Midboe A, Zulman D. Tablet distribution to veterans: an opportunity to increase patient portal adoption and use. J Am Med Inform Assoc 2022 Dec 13;30(1):73-82 [FREE Full text] [doi: 10.1093/jamia/ocac195] [Medline: 36269168]
25. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: With Applications in R. 1st ed. Corr. 7th printing 2017 edition. Cham: Springer; 2013.
26. Lindsay JA, Caloudas A, Hogan J, Ecker AH, Day S, Day G, et al. Getting connected: a retrospective cohort investigation of video-to-home telehealth for mental health care utilization among women veterans. J Gen Intern Med 2022 Sep 30;37(Suppl 3):778-785 [FREE Full text] [doi: 10.1007/s11606-022-07594-2] [Medline: 36042096]
27. Ferguson JM, Goldstein KM, Zullig LL, Zulman DM. Gender differences in adoption and frequency of virtual primary care among men and women veterans. J Womens Health (Larchmt) 2024 Jun 01;33(6):749-757. [doi: 10.1089/jwh.2023.0781] [Medline: 38629443]
28. Liu SK, Osborn AE, Bell S, Mecchella JN, Hort S, Batsis JA. Patient characteristics and utilization of an online patient portal in a rural academic general internal medicine practice. BMC Med Inform Decis Mak 2022 Feb 16;22(1):42 [FREE Full text] [doi: 10.1186/s12911-022-01778-w] [Medline: 35172805]
29. Greenberg AJ, Haney D, Blake KD, Moser RP, Hesse BW. Differences in access to and use of electronic personal health information between rural and urban residents in the United States. J Rural Health 2018 Feb 11;34(Suppl 1):s30-s38 [FREE Full text] [doi: 10.1111/jrh.12228] [Medline: 28075508]
30. Vogels E. Digital divide persists even as Americans with lower incomes make gains in tech adoption. Pew Research Center. 2021 Jun. URL: https://pewrsr.ch/2TRM7cP [accessed 2025-05-19]
31. Guajardo E, Amspoker AB, Stanley MA, Hogan J, Ecker AH, Lindsay JA. Patterns of telehealth use for mental health treatment among Hispanic veterans. Telemed J E Health 2023 May 01;29(5):788-792. [doi: 10.1089/tmj.2022.0300] [Medline: 36282802]
32. Javier SJ, Carey EP, Gunzburger E, Chen HP, Zeliadt SB, Williams EC, et al. Pain care disparities and the use of virtual care among racial-ethnic minority groups during COVID-19. J Gen Intern Med 2024 Feb 22;39(Suppl 1):68-78 [FREE Full text] [doi: 10.1007/s11606-023-08473-0] [Medline: 38252245]
33. Day G, Ecker AH, Amspoker AB, Dawson DB, Walder A, Hogan JB, et al. Black veteran use of video telehealth for mental health care. Psychol Serv 2024 Aug;21(3):478-488. [doi: 10.1037/ser0000827] [Medline: 38059987]
34. The researcher's guide to VA data. VA Information Resource Center. URL: https://www.virec.research.va.gov [accessed 2026-01-25]

Abbreviations

HCC: Hierarchical Condition Categories
ICD: International Classification of Diseases
OR: odds ratio
PGHD: patient-generated health data
PsyCMS: psychiatric case-mix system
SMD: standardized mean difference
SMHD: Share My Health Data
VA: Veterans Health Administration

XSL•FO
RenderX

Edited by J Sarvestan; submitted 14.Jul.2025; peer-reviewed by C Turvey, CK Yoo, S Hosseini; comments to author 01.Aug.2025; revised version received 30.Jan.2026; accepted 31.Jan.2026; published 26.Mar.2026.

Please cite as:
Zocchi MS, Lipschitz JM, Bixler FR, Robinson SA, Etingen B, Hogan TP, Ndiwane N, Calkins S, Newton T, Shah N, Kragen B, Young AS, Shimada SL
Patient Sharing of Digital Health Data in the Veterans Health Administration: Cross-Sectional Analysis
J Med Internet Res 2026;28:e80517
URL: https://www.jmir.org/2026/1/e80517
doi:10.2196/80517
PMID:

©Mark S Zocchi, Jessica M Lipschitz, Felicia R Bixler, Stephanie A Robinson, Bella Etingen, Timothy P Hogan, Ndindam Ndiwane, Saige Calkins, Terry Newton, Nilesh Shah, Ben Kragen, Alexander S Young, Stephanie L Shimada. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Leveraging Social Media to Achieve Population-Level Reach of Lung Cancer Screening-Eligible Individuals: A RE-AIM Framework Perspective

Lisa Carter-Bawa[1,2], MSN, MPH, PhD; Jamie S Ostroff[3], PhD; Susan M Rawl[4], PhD; Erin A Hirsch[1], PhD; Smita C Banerjee[3], PhD; Andrew Ciupek[5], PhD; Robert Skipworth Comer[6], MS; Minal Kale[7], MPH, MD; Katherine T Leopold[8], MD; Patrick O Monahan[9], PhD; James E Slaven Jr[9], MS; Francis Valenzona[1], BS; Renda Soylemez Wiener[10,11,12], MPH, MD; Ana Guadalupe Vielma[1], PhD

[1]Cancer Prevention Precision Control Institute, Center for Discovery & Innovation, Hackensack Meridian Health, 123 Metro Blvd, 6th Floor, 6400 Pod, Nutley, NJ, United States

[10]Center for Health Optimization & Implementation Research, VA Boston Healthcare System, Boston, MA, United States

[11]National Center for Lung Cancer Screening, Veterans Health Administration, Washington, DC, United States

[12]The Pulmonary Center, Boston University Chobanian & Avedesian School of Medicine, Boston, MA, United States

[2]Cancer Prevention & Control Program, Georgetown Lombardi Comprehensive Cancer Center, Washington, DC, United States

[3]Department of Psychiatry & Behavioral Science, Memorial Sloan Kettering Cancer Center, New York, NY, United States

[4]School of Nursing, Indiana University, Indianapolis, United States

[5]Go2 for Lung Cancer, Washington, United States

[6]School of Informatics, Indiana University, Indianapolis, United States

[7]Department of Medicine, Internal Medicine, Icahn School of Medicine at Mount Sinai, New York, United States

[8]Hackensack Meridian School of Medicine, Hackensack Meridian Health, Nutley, United States

[9]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, United States

Corresponding Author:
Lisa Carter-Bawa, MSN, MPH, PhD
Cancer Prevention Precision Control Institute, Center for Discovery & Innovation, Hackensack Meridian Health, 123 Metro Blvd, 6th Floor, 6400 Pod, Nutley, NJ, United States

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2026/1/e94664

Abstract

Background: Annual lung cancer screening (LCS) can decrease lung cancer–related mortality by finding cancer at earlier, more treatable stages, yet uptake remains abysmally low in the United States, especially among adults who seldom interact with the health system. Many eligible individuals are unaware that LCS exists, underscoring the critical need for scalable, population-level communication strategies that increase awareness and engagement.

Objective: The aim of this study was to evaluate *reach*, as defined by the Reach, Effectiveness, Adoption, Implementation, and Maintenance framework, as the extent to which the target population comes in contact with a social media–based strategy, Facebook-targeted advertisement (FBTA), designed to connect LCS-eligible individuals in the United States with a digital health communication message. The advertisement served as a digital outreach strategy for promoting engagement with *LungTalk*, an evidence-based intervention aimed at increasing awareness and informed decision-making about LCS.

Methods: As part of the INSPIRE-Lung Study (INnovating Social Media for Prevention: LUNG Cancer Screening Awareness, Knowledge, and Uptake), 5 FBTA campaigns were launched over a 79-day period throughout the United States. Advertisements targeted adults aged 50 - 80 years with interests related to smoking or smoking cessation and linked to a study website where participants could complete an eligibility screener and learn more about the trial. Facebook analytics were used to assess *reach*, defined by the number, proportion, and demographic characteristics of individuals exposed to and interacting with FBTA content. Key metrics included total reach, impressions, link clicks, and cost-efficiency.

Results: The FBTA campaigns reached 1,048,191 unique users and generated 3,109,482 impressions (total advertisement displays, including repeat exposures to the same user). A total of 24,816 individuals clicked on the advertisements (2.37% click-through rate), and 7117 completed the eligibility screener. Of those eligible, 1272 (17.9%) met lung screening criteria, and

XSL•FO
RenderX

of these, 483 (38% participation rate) enrolled in the trial. The cost per click was US $0.40, and the cost per enrolled participant was US $19.46. Individuals reached via FBTA were demographically diverse and included many who may be disconnected from traditional health care systems.

Conclusions: FBTA is a scalable, cost-effective strategy to achieve population-level *reach* of LCS-eligible adults. By conceptualizing *reach* as exposure to an upstream digital message rather than enrollment alone, this study illustrates how social media can broaden population access to evidence-based cancer prevention tools such as *LungTalk*. Future research should explore embedding intervention content directly into social media platforms and tracking downstream clinical outcomes.

Trial Registration: ClinicalTrials.gov NCT05824273; https://clinicaltrials.gov/study/NCT05824273

International Registered Report Identifier (IRRID): RR2-10.1186/s12889-023-15857-8

(J Med Internet Res 2026;28:e80281) doi:10.2196/80281

KEYWORDS

lung cancer screening; RE-AIM; Reach, Effectiveness, Adoption, Implementation, and Maintenance; digital health communication; Facebook-targeted advertisement; implementation science; public health outreach

Introduction

Lung cancer remains the leading cause of cancer-related death in the United States, claiming more lives each year than breast, cervical, colorectal, and prostate cancers combined [1]. Although low-dose computed tomography of the chest is an effective screening tool that can detect lung cancer at early, more treatable stages [2], uptake of lung cancer screening (LCS) remains low, particularly among populations at highest risk [3,4]. Many eligible individuals are unaware that LCS exists or do not know that they meet the criteria, underscoring the urgent need for scalable, population-level communication strategies that increase awareness and engagement [5-7].

The RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) framework provides a widely accepted lens for evaluating the population impact of health interventions. Within this framework, *reach* is defined as the absolute number, proportion, and representativeness of individuals who are willing to participate in a given initiative, intervention, or program [8]. However, in digital communication contexts, *reach* should also include those exposed to messaging that serves as a gateway to an intervention, particularly when the goal is to raise awareness and to generate activation toward health care engagement upstream from the health system. Upstream exposure is critical, as prior research demonstrates that initial exposure to preventive health information can subsequently enhance readiness for clinical engagement [9,10].

Social media platforms such as Facebook offer a powerful and cost-effective channel for health communication outreach. Facebook-targeted advertisement (FBTA) enables precise delivery of content to users based on demographic characteristics and behavioral interests, allowing public health practitioners to connect with specific high-risk audiences in real time. Although FBTA has often been used for participant recruitment [11,12], its potential to achieve population-level exposure to health communication interventions has not been fully explored within an implementation science framework.

The objective of this study was to evaluate the *reach* of an FBTA campaign designed to connect LCS-eligible individuals with a digital health communication intervention. Specifically, we examined (1) the number and demographic characteristics of individuals exposed to the FBTA; (2) engagement metrics including click-through rates (CTRs) and progression through the study enrollment pathway; and (3) cost efficiency of FBTA as a population-level outreach strategy. To our knowledge, this is the first study to examine FBTA as a scalable public health outreach strategy for connecting high-risk individuals with a health communication intervention designed to promote awareness of, and informed decision-making about, LCS within an implementation science framework.

Methods

Conceptual Framework

This study was guided by the RE-AIM framework, a widely used implementation science model for evaluating the public health impact of interventions delivered in real-world settings [8]. In digital outreach efforts, conceptually, *reach* can be broadened to account for upstream, preintervention exposure that can set the stage for subsequent engagement and trial enrollment [13,14]. Previous studies using digital platforms indicate that broadening the conceptualization of *reach* to include initial digital exposure is important to accurately capture the public health potential of such interventions [15]. In this study, we operationalized *reach* as (1) the number of unique individuals exposed to the FBTA campaign (total reach); (2) the proportion who engaged with the content (CTR); and (3) the demographic representativeness of those exposed compared to the target population for LCS. This operationalization extends the traditional clinic-based conceptualization to include upstream digital exposure as a meaningful public health end point.

This campaign was designed to reach adults likely to meet LCS eligibility criteria by leveraging Facebook's advertisement targeting features (eg, age, interests, and keywords related to smoking or tobacco use). The ads delivered brief, visually engaging content, with images that served as proxy exposures to the visuals used in the health communication intervention to which participants would later be randomized. For example, some ads featured images of clinical encounters, such as a patient speaking with a clinician or undergoing a lung exam with a stethoscope—images that mirrored scenes from the health communication intervention called *LungTalk* (Table 1). Briefly, *LungTalk* is a computer-tailored health communication and

XSL•FO
RenderX

decision support tool designed to educate users about lung health, lung cancer risks, and LCS. The accompanying advertisement text was designed to recruit eligible participants into the study and included statements such as "seeking 50 - 80-year-old men and women for online study opportunity about lung health," and "Are you between 50 and 80 years old and smoke cigarettes or used to smoke? Click to learn more and help improve patient education around lung screening!" Second, we assess user engagement with the FBTA content using Facebook's built-in analytics to measure visibility (impressions), interaction (click-throughs), and efficiency (cost per click and cost per consent).

XSL•FO
RenderX

Table . Demographics and Facebook analytics across 5 Facebook-targeted advertisement campaigns for the INSPIRE-Lung Study (INnovating Social Media for Prevention: LUNG Cancer Screening Awareness, Knowledge, and Uptake), a randomized controlled trial to increase lung cancer screening awareness among adults aged 50 - 80 years with smoking-related interests in the United States from August 2023 to February 2024 (n=1,048,191 unique users reached). Images in this table were sourced from the Microsoft 365 stock image library and are reproduced under the Microsoft 365 license for stock content for documentation of study materials.

Facebook metrics	Campaign #1	Campaign #2	Campaign #3	Campaign #4	Campaign #5	Overall
Campaign dates	Aug 1 - 19, 2023	Sep 4 - 15, 2023	Oct 16-Nov 27, 2023	Nov 28, 2023-Jan 28, 2024	Dec 9 - 16, 2023	—[a]
Duration, d	19	12	43	62	8	144[b]
Total reach, n (%[c])	402,817 (38.4)	343,043 (32.7)	75,742 (7.2)	186,544 (17.8)	40,045 (3.8)	1,048,191 (100)
Female	216,513 (53.7)	193,090 (56.3)	55,502 (73.3)	108,804 (58.3)	8175 (20.4)	582,084 (55.5)
Male	180,992 (44.9)	145,921 (42.5)	19,679 (26)	77,162 (41.4)	31,806 (79.4)	455,560 (43.5)
Unknown	5312 (1.3)	4032 (1.2)	560 (0.7)	578 (0.3)	64 (0.2)	10,546 (1.0)
Link clicks, n (%[c])	7871 (31.7)	5889 (23.7)	7906 (31.9)	2659 (10.7)	491 (2)	24,816 (100)
Female: 45 - 54 y	276 (3.5)	267 (4.5)	353 (4.5)	141 (5.3)	22 (4.5)	1059
Female: 55 - 64 y	1112 (14.1)	852 (14.5)	1773 (22.4)	471 (17.7)	47 (9.6)	4255
Female: 65+ y	3785 (48.1)	2623 (44.5)	3761 (47.6)	1014 (38.1)	73 (14.9)	11,256
Male: 45 - 54 y	132 (1.7)	147 (2.5)	102 (1.3)	118 (4.4)	44 (9)	543
Male: 55 - 64 y	505 (6.4)	466 (7.9)	511 (6.5)	333 (12.5)	109 (22.2)	1924
Male: 65+ y	1809 (23)	1476 (25.1)	1319 (16.7)	566 (21.3)	107 (21.8)	5277
Unknown	252 (3.2)	58 (0.98)	87 (1.1)	16 (0.6)	89 (18.1)	502
Impressions, n (%[c])	751,251 (36.07)	692,496 (33.25)	183,078 (8.79)	361,983 (17.38)	93,682 (4.5)	2,082,490 (100)
Female: 45 - 54 y	27,127 (3.6)	27,734 (4)	7216 (3.9)	18,505 (5.1)	4473 (4.8)	85,055
Female: 55 - 64 y	98,284 (13.1)	96,094 (13.9)	44,591 (24.4)	73,422 (20.3)	8455 (9)	320,846
Female: 65+ y	285,392 (38)	270,208 (39)	78,966 (43.1)	123,969 (34.2)	12,632 (13.5)	771,166
Male: 45 - 54 y	35,629 (4.7)	28,269 (4.1)	2795 (1.5)	20,038 (5.5)	20,339 (21.7)	107,070
Male: 55 - 64 y	99,562 (13.3)	81,489 (11.8)	15,279 (8.3)	54,944 (15.2)	27,618 (29.5)	278,892
Male: 65+ y	205,257 (27.3)	188,702 (27.2)	34,231 (18.7)	71,105 (19.6)	20,165 (21.5)	519,460
CTR[d], %	1.96	1.72	10.44	1.43	1.23	2.37
Cost per click, US $	0.32	0.35	0.38	0.63	1.43	0.40
Total spent, US $	2499.55	2084.82	3000.00	1676.76	701.85	9962.98
Advertisement image						—
Advertisement text	Seeking 50-80-year-old men and women for a study opportunity about lung health.	Are you between 50-80 years old and smoke cigarettes or used to smoke?	Seeking 50-80-year-old men and women to learn more about lung health and screening.	Are you between 50-80 years old and interested in learning more about lung health and screening?	Are you between 50-80 years old and smoke cigarettes or used to smoke?	—

[a]Not applicable.

[b]Total days reflect 79 unique recruitment days; some campaigns overlapped (campaign 4 and 5 ran concurrently).

XSL•FO
RenderX

[c]The percentages in this main row were calculated by using the "Overall" value as the denominator. The percentages for the subcategories were calculated by using the corresponding n values from this row as the denominator.

[d]CTR: click-through rate (CTR = link clicks/reach ×100).

Most important, FBTA was not the intervention in itself. Rather, it was a communication medium for a study advertisement that enabled scalable outreach to high-risk individuals—adults aged 50‐80 years with a history of smoking—prior to their engagement with the health care system. These individuals represent a population that may have limited engagement with traditional clinical touchpoints (eg, annual physical exam and wellness visits) [16]. Given that LCS awareness remains below 10% nationally [17] and that eligible individuals who smoke are less likely to engage with preventive care [18], FBTA provides an opportunity to deliver targeted health messaging outside traditional clinical encounters—meeting individuals where they already spend time online and potentially catalyzing initial awareness that may precede health care engagement. As such, exposure to the FBTA served as a critical first step in a targeted health communication intervention engagement pathway.

Anchoring this analysis in the RE-AIM framework allows us to expand beyond basic metrics of trial enrollment. Instead, we offer a nuanced understanding of population-level *reach* by quantifying how digital advertising can expand the frontiers of public health communication, particularly for LCS, where knowledge gaps and access barriers persist. While the RE-AIM framework encompasses 5 dimensions, researchers have noted that these components can be examined individually when the research question warrants focused analysis of a specific dimension [14,19]. The present analysis focuses exclusively on *reach* because (1) this paper evaluates the FBTA outreach strategy as a discrete implementation component; and (2) establishing population-level *reach* is a necessary precursor to evaluating downstream dimensions. The remaining RE-AIM dimensions—effectiveness, adoption, implementation, and maintenance—will be evaluated in subsequent publications from the INSPIRE-Lung Study (INnovating Social Media for Prevention: LUNG Cancer Screening Awareness, Knowledge, and Uptake).

Study Overview

The INSPIRE-Lung Study (ClinicalTrials.gov NCT05824273) is a national, community-based randomized controlled trial designed to increase awareness, knowledge, and uptake of LCS among high-risk individuals [20]. To evaluate reach, we measured exposure to an FBTA campaign that served as a proxy for the health communication intervention.

Ethical Considerations

This study was approved by the Hackensack Meridian Health Institutional Review Board (Protocol: Pro2022-0860). Informed consent was obtained electronically from all the participants who enrolled in the trial. For aggregated Facebook analytics data (reach, impressions, and clicks), a waiver of informed consent was granted, as these data are anonymized and do not contain personally identifiable information. No personal identifying information was retained from users who only viewed advertisements. All data were deidentified prior to analysis to protect participant privacy and confidentiality. Participants received US $50 for completing baseline and 1-week follow-up surveys and US $25 for completing the 6-month follow-up survey.

FBTA Campaigns

Between August 2023 and February 2024, 5 discrete FBTA campaigns were used to reach adults aged 50‐80 years in the United States. Campaign durations ranged from 8 to 62 days (Table 1). Campaigns varied in messaging and imagery to test different engagement approaches; however, the primary focus of this analysis is cumulative *reach* rather than comparative campaign effectiveness. Advertisements were developed using Facebook's advertisement manager platform and tailored to target users based on age, geographic location, and interests related to smoking (eg, cigarettes, vaping, smoking cessation, and nicotine replacement therapy).

Each advertisement linked to the study landing page that briefly described the purpose of the trial and invited users to complete a screening questionnaire to determine LCS eligibility. Although users were not exposed to the full health communication intervention unless they enrolled in the trial, the FBTA messaging was designed to reflect the tone and content focus of health messaging around LCS, emphasizing that screening can save lives and is available to eligible adults with a history of smoking. Thus, FBTA served as a population-level outreach tool and an upstream proxy exposure to the intervention.

Campaigns varied in wording, imagery, and call-to-action phrasing to optimize engagement. Prior to deployment, all ads were reviewed for cultural appropriateness, literacy level, and ethical compliance with best practices for digital recruitment. Real-time monitoring allowed for iterative campaign refinements based on audience response.

Reach: Definition and Metrics

To evaluate *reach* in alignment with the RE-AIM framework, we used both Facebook analytics [21] and internal study data. Facebook analytics [21] provided the real-time performance metrics as follows:

- *Total reach:* The number of unique users who viewed the FBTA content at least once
- *Impressions:* The total number of times the FBTA appeared on users' screens (which may include multiple exposures per user)
- *Link clicks:* The number of users who clicked the FBTA and were directed to the study website
- *CTR:* The percentage of viewers who clicked on the advertisement
- *Cost metrics:* Cost per click and cost per consented participant

An important consideration is that Facebook's targeting capabilities allow selection based on age and behavioral interests (eg, smoking-related content) but cannot verify clinical eligibility criteria such as the minimum 20 pack-year smoking

XSL•FO
RenderX

history. Therefore, the total reach metric represents exposure among individuals likely but not confirmed to meet screening eligibility. Most importantly, for the purpose of this aim, we conceptualized *reach* as population-level exposure to a public-facing health communication message (delivered via FBTA), independent of whether individuals proceeded to trial enrollment. This approach aligns with RE-AIM's emphasis on assessing exposure and potential engagement among the broader target population.

Analytic Strategy

Descriptive statistics were used to summarize campaign-level and cumulative metrics across the 5 FBTA campaigns. We calculated total reach, impressions, CTRs, and cost-efficiency measures. We also examined FBTA impressions by age and sex distribution (Table 1). To assess representativeness of digital engagement, we compared the demographic distribution of link clicks to impressions using chi-square tests. Because Facebook analytics provide aggregate demographic data on impressions rather than unique viewers, we used impression distribution as a proxy for viewer demographics in these comparisons. Additionally, we examined user progression through the digital pathway—from advertisement exposure to click-through to screener completion and trial enrollment—to contextualize how FBTA contributed to downstream engagement with the full health communication intervention.

Results

Over a 79-day recruitment period between August 2023 and February 2024, 5 US-based FBTA campaigns were launched nationally to reach high-risk individuals eligible for LCS. Across all campaigns, a total of 1,048,191 unique Facebook users were exposed to at least 1 advertisement, representing the total reach of the FBTA strategy. These users generated 3,109,482 total impressions, indicating repeated exposure to the content over time. Among those reached, 24,816 (2.37%) users clicked on the advertisement, which directed them to the study landing page. Of these, 7117 (28.6%) individuals completed the brief eligibility screening questionnaire. A total of 1272 (17.9% of screeners) individuals met LCS eligibility criteria, and of these, 483 (38%) enrolled in the randomized trial. Table 1 details demographic characteristics across the 5 campaigns. Table 2 details the demographic characteristics of participants enrolled in the study.

XSL•FO
RenderX

Table . Demographic characteristics of participants enrolled in the INSPIRE-Lung Study (INnovating Social Media for Prevention: LUNG Cancer Screening Awareness, Knowledge, and Uptake), a community-based randomized controlled trial to increase lung cancer screening awareness, recruited via Facebook-targeted advertisement in the United States from August 2023 to February 2024 (n=483).

Characteristic	Participants
Age (y), mean (SD)	63.3 (6.7)
Sex, n (%)	
Male	236 (48.9)
Female	246 (50.9)
Nonbinary	1 (0.2)
Race, n (%)	
White	376 (79.3)
Black	70 (14.8)
Asian	7 (1.5)
American Indian	12 (2.5)
Native Hawaiian/Alaskan	2 (0.4)
Multiracial	7 (1.5)
Hispanic, n (%)	
Yes	23 (4.8)
No	454 (95.2)
Education, n (%)	
Less than high school	10 (2.1)
High school graduate	56 (11.6)
Some college	144 (29.9)
College graduate or higher	272 (56.4)
Income, n (%)	
<US $25,000/y	102 (21.2)
$US 25,000-US $50,000/y	141 (29.3)
>US $50,000/y	238 (49.5)
Smoking history (y), mean (SD)	35.5 (10.3)
Currently smokes, n (%)	
Yes	313 (64.8)
No	170 (35.2)

These data demonstrated the substantial *reach* of the FBTA campaign as an upstream population-level exposure strategy. Table 3 presents key campaign metrics across the 5 FBTA iterations. The average cost per link click was US $0.40, and the total campaign cost was US $9962.98, yielding a cost per enrolled participant of US $19.46. Performance varied across campaigns, with Campaign #1 generating the highest total reach (38.4%) and the lowest cost per click (US $0.32). Campaign #3 generated the most link clicks (7906; 31.9%) despite lower total reach. These variations may reflect differences in messaging, imagery, timing, or audience saturation and will be explored in future work.

XSL•FO
RenderX

Table . Summary of Facebook-targeted advertisement campaign performance metrics for the INSPIRE-Lung Study (INnovating Social Media for Prevention: LUNG Cancer Screening Awareness, Knowledge, and Uptake), a randomized controlled trial targeting lung cancer screening–eligible adults in the United States from August 2023 to February 2024.

Metric	Value[a]
Total reach, number of unique Facebook users	1,048,191
Total impressions, n	3,109,482
Total link clicks, n	24,816
Click-through rate, %	2.36
Screener completers, n	7117
Screen-eligible respondents, n	1272
Enrolled participants, n	483
Total campaign cost, US $	9962.98
Cost per click, US $	0.40
Cost per consent, US $	19.46

[a]These values are total social media and study metrics.

Participants reached through the FBTA campaign were geographically and demographically diverse, with representation across all 50 US states (Figure 1). Among those who enrolled, the average age was 63.3 (SD 6.7) years, and 50.9% identified as female participants.

Figure 1. Geographic distribution of INSPIRE-Lung Study participants enrolled via Facebook-targeted advertisement across all 50 US states from August 2023 to February 2024 (n=483). INSPIRE-Lung Study: INnovating Social Media for Prevention: LUNG Cancer Screening Awareness, Knowledge, and Uptake.

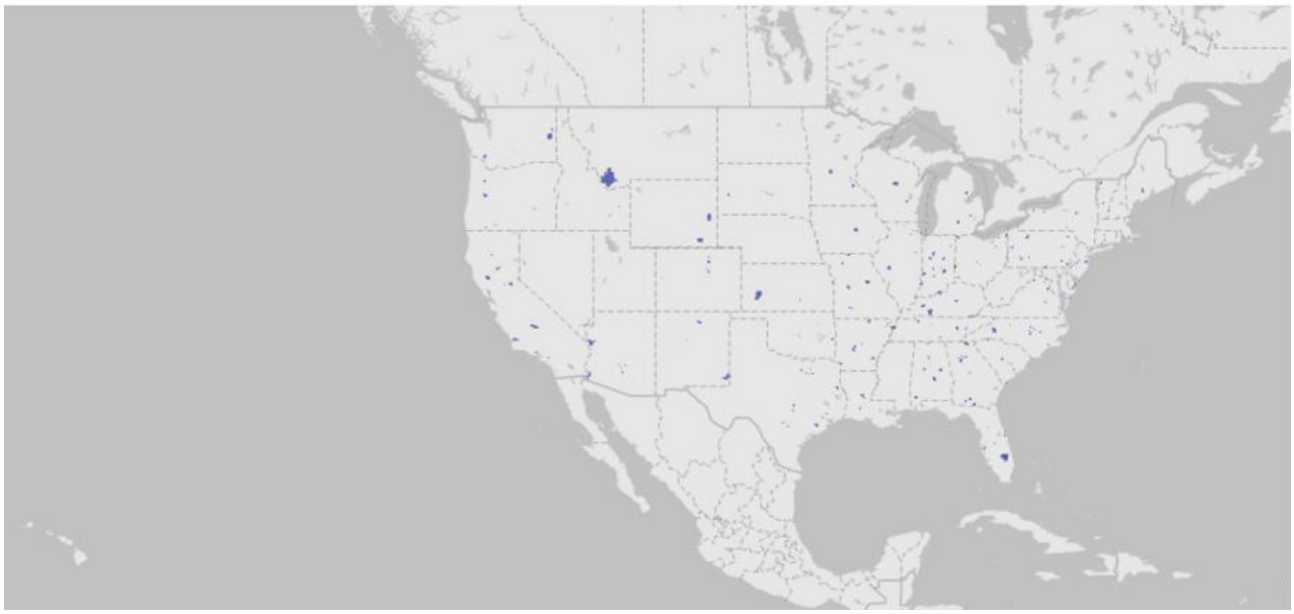

To examine representativeness, we compared the demographic distribution of individuals who clicked on the FBTA to those who viewed the advertisements (using impressions as a proxy for exposure). Chi-square tests revealed significant differences in engagement patterns by age, sex, and geography. Regarding age, individuals aged 65 years and older comprised 61.9% of total impressions but 68% of link clicks, indicating higher engagement among older viewers (χ^2_2=436.86; *P*<.001). CTRs increased with age: 0.83% for ages 45 - 54 years, 1.03% for ages 55 - 64 years, and 1.28% for ages 65+ years. Regarding sex, female participants comprised 56.5% of impressions but 68.2% of link clicks, demonstrating significantly higher engagement compared to male participants (χ^2_1=1337.74; *P*<.001). The CTR for female participants (1.41%) was notably higher than for male participants (0.86%). Geographic distribution also differed significantly across US Census regions (χ^2_3=26.89; *P*<.001), with the South accounting for the highest proportion of both impressions (36.2%) and clicks (37.1%).

However, even among individuals who did not proceed to enrollment, exposure to the FBTA message about LCS represents a meaningful first step toward increasing awareness and consideration of screening. This broader lens on *reach* captures the use of digital messaging strategies upstream from health care system engagement.

XSL•FO
RenderX

Discussion

Principal Findings

This study demonstrates the feasibility and use of FBTA as a digital strategy to achieve broad, population-level *reach* to LCS-eligible individuals using tailored health messaging. Guided by the RE-AIM framework [8], we operationalized *reach* not as trial enrollment but as exposure to a communication message that served as an upstream proxy for a health communication intervention. Across 5 campaigns in the United States, we reached over 1 million unique individuals within the target age range and smoking history risk profile for LCS. These findings highlight the potential of FBTA to serve as a scalable public health outreach tool that promotes engagement with evidence-based communication interventions.

Most importantly, this analysis distinguishes between *exposure* and *enrollment*. While a subset of users exposed to the FBTA ultimately enrolled in the trial and received the *LungTalk* intervention or control condition depending on randomization, the majority did not progress beyond the initial digital touchpoint. Nonetheless, this first contact with a tailored screening message represents a meaningful public health outcome, particularly in a context where lack of awareness is a persistent barrier to LCS uptake. Although FBTA effectively reached individuals who may not have engaged with traditional health care outreach, the extent to which initial digital exposure translates into sustained clinical behavior change remains uncertain and warrants further longitudinal investigation. In addition, given the fact that LCS awareness is so low in general (<10%) [17], increasing population level awareness regardless of screening eligibility is of great value. In line with RE-AIM, conceptualizing *reach* in this broader way captures the real-world potential of digital tools to connect individuals with preventive health information upstream of clinical interaction. The remaining RE-AIM dimensions—effectiveness, adoption, implementation, and maintenance—will be evaluated in subsequent publications from the INSPIRE-Lung Study. Specifically, effectiveness will be assessed through the primary RCT outcomes examining *LungTalk's* impact on LCS awareness, knowledge, and screening uptake. Adoption and implementation metrics will examine health care system and provider engagement, while maintenance will assess sustained behavior change at 6- and 12-month follow-ups.

To better understand the representativeness of the FBTA campaigns, we compared the demographics of enrolled participants with the screening population in the American College of Radiology's Lung Cancer Screening Registry (ACR LCSR), as reported by Silvestri et al [22]. In that study of over 1 million individuals screened for lung cancer with low-dose computed tomography between 2015 and 2019, the population was predominantly White (91.6%), with only 7.4% Black and 0.9% Asian individuals among those with known race/ethnicity data. In contrast, our FBTA-recruited sample was notably more diverse: 79.3% White, 14.8% Black (double the ACR LCSR proportion), and 4.8% Hispanic. This suggests that FBTA may be particularly effective at reaching racial and ethnic minority populations who have historically been underrepresented in clinic-based LCS.

Additional demographic patterns also differed from the ACR LCSR. While the registry population skewed older adults (42.6% aged 65+ y) and female participants (approximately 54%), our enrolled participants showed a more balanced sex distribution (50.9% female participants, 48.9% male participants) and a mean age of 63.3 (SD 6.7) years. The ACR LCSR also reported that 61.4% of screened individuals currently smoked; similarly, 64.8% of our enrolled participants currently smoked, indicating that FBTA successfully reached individuals with active smoking behaviors who represent a priority population for LCS outreach. Notably, Black and Hispanic individuals demonstrated lower adherence to annual screening in the ACR LCSR, suggesting that alternative outreach strategies like FBTA may help address these disparities at the initial awareness and engagement stage.

However, certain groups, such as those with lower educational attainment or without internet access, may still be underrepresented in digital outreach, underscoring the importance of multichannel approaches. In our sample, 56.4% had a college degree or higher, which exceeds national averages and suggests potential selection bias toward more educated individuals.

The FBTA strategy also proved highly cost-efficient. While the average cost per link click was US $0.40 and the cost per enrolled participant was US $19.46, the cost per view (or impression) was substantially lower, at just over US $0.003 per exposure. This underscores the potential of FBTA to deliver high-volume, targeted health communication at minimal per-person cost. Our cost per enrolled participant (US $19.46) compares favorably to published estimates for traditional recruitment methods in cancer prevention research. Community-based recruitment strategies have reported costs ranging from US $50 to over US $500 per participant, depending on intensity and population [23,24]. Direct mail campaigns for health research typically cost US $100 - 200 per enrolled participant [25]. Our prior work using FBTA for LCS research demonstrated similar cost efficiencies [11]. Compared to traditional outreach strategies such as print mailings, community events, telephone outreach, or provider referrals, which often require substantial personnel time and logistical coordination, FBTA offers real-time precision targeting and scalability. This is especially advantageous for reaching populations who may have limited engagement with the health care system. According to the Pew Research Center (2024), approximately 68% of the US adults use Facebook, with relatively consistent usage across income levels and geographic regions [26]. Most importantly, Facebook penetration among adults aged 50 - 64 (70%) years and 65+ (50%) years approaches or mirrors general population estimates, suggesting potential to reach individuals across the socioeconomic spectrum. While Facebook users as a whole may not be systematically disconnected from health care, our targeting strategy—focusing on individuals with smoking-related interests—may identify a subpopulation with elevated barriers to care, as prior research indicates that individuals who currently smoke are less likely to have a usual source of care and less likely to receive preventive services [18]. Moreover, the low rate of LCS uptake nationally (<6%) suggests substantial

XSL•FO
RenderX

numbers of eligible individuals are not engaging with screening through traditional clinical pathways. FBTA may be particularly valuable for reaching individuals living in rural areas, those without consistent primary care access, and those less likely to encounter LCS information through conventional clinical channels.

Limitations

Nevertheless, several limitations should be acknowledged. First, while FBTA allowed for targeted delivery of content based on self-reported profile characteristics and behavioral interests, we could not verify screening eligibility (eg, 20+ packs per year smoking history) until individuals clicked through to the screening questionnaire. As such, Facebook's *reach* metric represents exposure among individuals likely but not confirmed to meet screening eligibility criteria and may include a significant proportion who are not personally eligible for LCS. The fact that only 17.9% of those who completed the screener met eligibility criteria illustrates this gap between interest-based targeting and clinical eligibility. However, we conceptualize this exposure as meaningful from a public health perspective: raising awareness about LCS, even among some individuals who may not personally meet eligibility criteria, has value, as these individuals may share information with eligible family members or friends. Given the inherently social nature of Facebook, this kind of peer-sharing and diffusion of information could amplify the campaign's impact within networks of individuals with shared interests in smoking or cessation.

Second, although the study provides valuable insights into digital *reach*, it does not assess the quality or depth of engagement with the FBTA message among those who viewed but did not interact with the advertisement. Future studies could incorporate passive analytics (eg, time spent viewing or scrolling behavior) or embed brief *LungTalk*-derived content within the advertisement itself to assess this dimension.

Third, there are broader ethical and equity considerations when leveraging commercial social media platforms for health communication. While Facebook offers powerful precision-targeting capabilities, Facebook algorithms may inherently exclude or underrepresent certain demographic groups, potentially biasing exposure toward users already digitally active or engaged, which could inadvertently widen health disparities [27]. Ensuring equitable digital engagement will be essential as these strategies are scaled [27]. Future research should systematically monitor and adjust targeting criteria to mitigate this risk.

Fourth, we cannot definitively rule out exposure to fake or bot accounts, though Facebook's platform integrity measures are designed to minimize such accounts. Additionally, our downstream metrics (screener completion, eligibility verification, and trial enrollment) require human engagement that automated accounts cannot replicate, providing confidence that our enrollment figures represent genuine participants.

Despite these limitations, our findings have important implications for cancer prevention and public health communication. Reaching high-risk individuals who are not yet engaged with the health care system remains a major barrier to improving LCS rates. FBTA offers a powerful, scalable method to extend the reach of evidence-based interventions like *LungTalk* and to promote awareness among populations traditionally missed by clinic-based strategies.

Future efforts should explore how social media–based outreach can be integrated with other implementation strategies, such as patient navigation or community partnerships, to support screening uptake. Additionally, embedding the intervention content itself directly into the FBTA (rather than linking out) will allow for even greater scalability and immediacy of impact.

Next Steps and Future Directions

Building on the demonstrated feasibility and population-level *reach* of FBTA to connect with LCS-eligible individuals, several strategic opportunities emerge for advancing digital public health communication and implementation efforts.

First, future research should explore embedding core components of the intervention, such as tailored *LungTalk* messaging or *LungTalk* itself, directly within the FBTA interface. This would allow for immediate engagement with decision support content at the moment of exposure, reducing drop-off and increasing the likelihood that viewers absorb key information, even if they do not proceed to enrollment. Embedding brief video clips, testimonials, or tailored visual messages adapted from *LungTalk* could help bridge the gap between exposure and education. For example, a randomized comparative effectiveness study could assess whether embedding *LungTalk* intervention components directly within FBTA enhances immediate knowledge gains compared to ads linking externally to a study website.

Second, optimizing message design and delivery remains critical. Future campaigns may consider using systematic A or B testing of message framing, imagery, tone, and calls to action, with attention to behavioral science principles and cultural relevance. Tailoring content for priority populations, such as Black, Hispanic or Latino, or rural adults with elevated lung cancer risk, can further improve engagement and equity.

Third, future studies should examine how digital *reach* translates into downstream behavioral and clinical outcomes. For example, FBTA-driven exposure could be linked to subsequent engagement with patient navigation, telehealth consults, or direct referral systems to LCS programs. Implementation strategies that combine digital outreach with systems-level linkage to care hold promise for closing the gap between awareness and screening uptake.

Fourth, collaboration across outreach and communication sectors will be essential for scaling impact. Partnering with health systems, public health departments, advocacy organizations, and community-based coalitions can support broader dissemination, localized adaptation, and sustainable infrastructure for ongoing FBTA campaigns.

Finally, ensuring equity in digital health communication is imperative. As social media becomes an increasingly common tool for public health messaging, attention must be paid to disparities in internet access, digital literacy, and algorithmic exposure. Mixed methods evaluations, community-engaged

XSL•FO
RenderX

research, and monitoring of campaign delivery across demographic groups can help identify and address gaps. Most importantly, even users who are not personally eligible for LCS may serve as information conduits within their social networks, amplifying campaign effects through peer sharing and expanding awareness among others who may have missed the original exposure.

In addition, FBTA offers a scalable, adaptable, effective, and cost-efficient platform for delivering tailored health communication content upstream of clinical care. By refining and integrating this strategy with broader implementation efforts, future research and practice can more effectively promote equitable access to LCS information and ultimately reduce disparities in early detection and outcomes.

Conclusions

This study demonstrates the potential of FBTA to achieve meaningful population-level *reach* to individuals eligible for LCS, as defined by the RE-AIM framework. By delivering a tailored communication message that served as a proxy exposure to a health communication intervention that promotes awareness and informed decision-making about LCS, FBTA enabled contact with over 1 million adults within the eligible age range, many of whom are unlikely to encounter such information through traditional health care pathways.

By broadening *reach* to encompass initial exposure to health-promoting digital messages, this analysis highlights the importance of upstream engagement in implementation science and public health practice. FBTA proved to be a cost-efficient and scalable outreach strategy capable of penetrating social networks, initiating awareness, and opening pathways for subsequent engagement with LCS services.

As public health increasingly embraces digital platforms to extend its reach, integrating social media strategies with evidence-based interventions offers a promising direction for reducing disparities in screening and early detection. Future efforts should build on this foundation by embedding intervention content directly into advertisement formats, testing links to downstream care, and ensuring equitable access across diverse communities.

Funding

This work was supported by grant no. R01CA263662 (principal invesigator [PI]: LC-B) received from the National Cancer Institute. LC-B is supported in part by Georgetown University Lombardi CCSG P30CA051008 (PI: Louis Weiner). RSW is supported in part by resources from the VA Boston Healthcare system. JSO and SCB are supported in part by MSK CCSG P30CA008748 (PI: Selwyn Vickers).

Disclaimer

The views expressed herein are those of the authors and do not necessarily reflect the views of the Department of Veterans Affairs or the US government.

Data Availability

The Facebook analytics data presented in this study are available from the corresponding author upon reasonable request. Individual-level participant data from the INSPIRE-Lung Study (INnovating Social Media for Prevention: LUNG Cancer Screening Awareness, Knowledge, and Uptake) are not publicly available due to privacy and ethical restrictions but may be available upon reasonable request and with appropriate data use agreements.

Authors' Contributions

Conceptualization: LC-B, JSO, SMR
Data curation: LC-B, RSC, FV
Formal analysis: LC-B, POM, JES Jr
Funding acquisition: LC-B
Investigation: LC-B, EAH, AGV
Methodology: LC-B, JSO, SMR, SCB
Project administration: LC-B, FV
Resources: LC-B, AC
Software: RSC
Supervision: LC-B
Validation: LC-B, EAH
Visualization: LC-B, RSC
Writing – original draft: LC-B
Writing – review & editing: LC-B, JSO, SMR, EAH, SCB, AC, RSC, MK, KTL, POM, JES Jr, FV, RSW, AGV

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Cancer facts & figures 2023. : American Cancer Society, Inc; 2023 URL: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2023/2023-cancer-facts-and-figures.pdf [accessed 2026-02-03]
2. National Lung Screening Trial Research Team, Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011 Aug 4;365(5):395-409. [doi: 10.1056/NEJMoa1102873] [Medline: 21714641]
3. Rivera MP, Katki HA, Tanner NT, et al. Addressing disparities in lung cancer screening eligibility and healthcare access. An official American Thoracic Society statement. Am J Respir Crit Care Med 2020 Oct 1;202(7):e95-e112. [doi: 10.1164/rccm.202008-3053ST] [Medline: 33000953]
4. Racial and ethnic disparities. American Lung Association, State of Lung Cancer. 2023. URL: https://www.lung.org/research/state-of-lung-cancer/racial-and-ethnic-disparities [accessed 2026-02-03]
5. Carter-Harris L, Ceppa DP, Hanna N, Rawl SM. Lung cancer screening: what do long-term smokers know and believe? Health Expect 2017 Feb;20(1):59-68. [doi: 10.1111/hex.12433] [Medline: 26701339]
6. Carter-Harris L, Tan ASL, Salloum RG, Young-Wolff KC. Patient-provider discussions about lung cancer screening pre- and post-guidelines: Health Information National Trends Survey (HINTS). Patient Educ Couns 2016 Nov;99(11):1772-1777. [doi: 10.1016/j.pec.2016.05.014] [Medline: 27241830]
7. Williamson TJ, Walsh LE, Rawl SM, Carter-Bawa L. Slipping through the cracks: who is eligible but does not receive a healthcare provider recommendation for lung cancer screening? Lung Cancer (Auckl) 2023 May;179:107185. [doi: 10.1016/j.lungcan.2023.107185] [Medline: 37023535]
8. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
9. Hibbard JH, Greene J. What the evidence shows about patient activation: better health outcomes and care experiences; fewer data on costs. Health Aff (Millwood) 2013 Feb;32(2):207-214. [doi: 10.1377/hlthaff.2012.1061] [Medline: 23381511]
10. Prochaska JO, DiClemente CC. Stages and processes of self-change of smoking: toward an integrative model of change. J Consult Clin Psychol 1983 Jun;51(3):390-395. [doi: 10.1037//0022-006x.51.3.390] [Medline: 6863699]
11. Carter-Harris L, Bartlett Ellis R, Warrick A, Rawl S. Beyond traditional newspaper advertisement: leveraging Facebook-targeted advertisement to recruit long-term smokers for research. J Med Internet Res 2016 Jun 15;18(6):e117. [doi: 10.2196/jmir.5502] [Medline: 27306780]
12. Walsh LE, Carter-Bawa L. Using social media to recruit individuals for health-related research: feasibility and lessons learned. J Health Psychol 2023 Jun;28(7):599-606. [doi: 10.1177/13591053221136928] [Medline: 36416197]
13. Kessler RS, Purcell EP, Glasgow RE, Klesges LM, Benkeser RM, Peek CJ. What does it mean to "employ" the RE-AIM model? Eval Health Prof 2013 Mar;36(1):44-66. [doi: 10.1177/0163278712446066] [Medline: 22615498]
14. Glasgow RE, Estabrooks PE. Pragmatic applications of RE-AIM for health care initiatives in community and clinical settings. Prev Chronic Dis 2018 Jan 4;15:E02. [doi: 10.5888/pcd15.170271] [Medline: 29300695]
15. Eppes EV, Augustyn M, Gross SM, Vernon P, Caulfield LE, Paige DM. Engagement with and acceptability of digital media platforms for use in improving health behaviors among vulnerable families: systematic review. J Med Internet Res 2023 Feb 3;25:e40934. [doi: 10.2196/40934] [Medline: 36735286]
16. Haddad DN, Sandler KL, Henderson LM, Rivera MP, Aldrich MC. Disparities in lung cancer screening: a review. Ann Am Thorac Soc 2020 Apr;17(4):399-405. [doi: 10.1513/AnnalsATS.201907-556CME] [Medline: 32017612]
17. Williamson TJ, Brymwitt WM, Hirsch EA, Reese MT, Carter-Bawa L. Patient-clinician discussions on lung cancer screening in the United States before and after 2021 guidelines. Transl Behav Med 2025 Jan 16;15(1):ibaf039. [doi: 10.1093/tbm/ibaf039] [Medline: 40803694]
18. Cornelius ME, Loretan CG, Jamal A, et al. Tobacco product use among adults - United States, 2021. MMWR Morb Mortal Wkly Rep 2023 May 5;72(18):475-483. [doi: 10.15585/mmwr.mm7218a1] [Medline: 37141154]
19. Holtrop JS, Estabrooks PA, Gaglio B, et al. Understanding and applying the RE-AIM framework: clarifications and resources. J Clin Transl Sci 2021;5(1):e126. [doi: 10.1017/cts.2021.789] [Medline: 34367671]
20. Carter-Bawa L, Banerjee SC, Comer RS, et al. Leveraging social media to increase lung cancer screening awareness, knowledge and uptake among high-risk populations (the INSPIRE-Lung Study): study protocol of design and methods of a community-based randomized controlled trial. BMC Public Health 2023 May 26;23(1):975. [doi: 10.1186/s12889-023-15857-8] [Medline: 37237339]
21. Meta. About reach. Facebook Business Help Center. URL: https://www.facebook.com/business/help/710746785663278 [accessed 2026-02-03]
22. Silvestri GA, Goldman L, Tanner NT, et al. Outcomes from more than 1 million people screened for lung cancer with low-dose CT imaging. Chest 2023 Jul;164(1):241-251. [doi: 10.1016/j.chest.2023.02.003] [Medline: 36773935]
23. Treweek S, Pitkethly M, Cook J, et al. Strategies to improve recruitment to randomised trials. Cochrane Database Syst Rev 2018 Feb 22;2(2):MR000013. [doi: 10.1002/14651858.MR000013.pub6] [Medline: 29468635]
24. UyBico SJ, Pavel S, Gross CP. Recruiting vulnerable populations into research: a systematic review of recruitment interventions. J Gen Intern Med 2007 Jun;22(6):852-863. [doi: 10.1007/s11606-007-0126-3] [Medline: 17375358]

XSL•FO
RenderX

25. Coley HL, Sadasivam RS, Williams JH, et al. Crowdsourced peer- versus expert-written smoking-cessation messages. Am J Prev Med 2013 Nov;45(5):543-550. [doi: 10.1016/j.amepre.2013.07.004] [Medline: 24139766]
26. Social media fact sheet. Pew Research Center. 2025 Nov 20. URL: https://www.pewresearch.org/internet/fact-sheet/social-media/ [accessed 2026-02-03]
27. Arigo D, Pagoto S, Carter-Harris L, Lillie SE, Nebeker C. Using social media for health research: methodological and ethical considerations for recruitment and intervention delivery. Digit Health 2018;4:2055207618771757. [doi: 10.1177/2055207618771757] [Medline: 29942634]

Abbreviations

ACR LCSR: American College of Radiology's Lung Cancer Screening Registry
CTR: click-through rate
FBTA: Facebook targeted advertisement
INSPIRE-Lung Study: INnovating Social Media for Prevention: LUNG Cancer Screening Awareness, Knowledge, and Uptake
LCS: lung cancer screening
RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance

Edited by A Mavragani; submitted 07.Jul.2025; peer-reviewed by I -Wen Pan, R Williams; revised version received 02.Jan.2026; accepted 20.Jan.2026; published 03.Mar.2026.

Please cite as:
Carter-Bawa L, Ostroff JS, Rawl SM, Hirsch EA, Banerjee SC, Ciupek A, Comer RS, Kale M, Leopold KT, Monahan PO, Slaven Jr JE, Valenzona F, Wiener RS, Vielma AG
Leveraging Social Media to Achieve Population-Level Reach of Lung Cancer Screening-Eligible Individuals: A RE-AIM Framework Perspective
J Med Internet Res 2026;28:e80281
URL: https://www.jmir.org/2026/1/e80281
doi:10.2196/80281

© Lisa Carter-Bawa, Jamie S Ostroff, Susan M Rawl, Erin A Hirsch, Smita C Banerjee, Andrew Ciupek, Robert Skipworth Comer, Minal Kale, Katherine T Leopold, Patrick O Monahan, James E Slaven Jr, Francis Valenzona, Renda Soylemez Wiener, Ana Guadalupe Vielma. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 3.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Changes in Workplace Productivity and Estimated Cost Savings During Internet-Based Cognitive Behavioral Therapy in the Irish National Health Service: Naturalistic, Repeated-Measures, Retrospective Survey Study

Chi Tak Lee[1,2,3,4*], PhD; Siobhán Harty[1,2,3,4*], PhD; Angel Enrique[4], PhD; Alba Jiménez-Díaz[5], MS; Garrett Hisler[4*], PhD; Daniel Duffy[4], PhD; Derek Richards[1,4], PhD

[1]School of Psychology, Trinity College Dublin, Dublin, Ireland

[2]Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland

[3]Global Brain Health Institute, Trinity College Dublin, Dublin, Ireland

[4]Amwell Science, Amwell, 1 Stephen Street Upper, Dublin, Ireland

[5]Department of Personality, Evaluation and Psychological Treatment, Faculty of Psychology, University of Valencia, Valencia, Spain

[*]these authors contributed equally

Corresponding Author:
Garrett Hisler, PhD
Amwell Science, Amwell, 1 Stephen Street Upper, Dublin, Ireland

Abstract

Background: Depression and anxiety can significantly impact workplace productivity, for instance, by increasing absenteeism and presenteeism. This loss of productivity leads to diminished workplace economic outcomes. Internet-based cognitive behavioral therapy (iCBT) has emerged as a cost-effective intervention within workplace settings that improves workplace productivity loss due to depression and anxiety, but more generalizable evidence beyond the workplace, such as in a national health service setting, is lacking.

Objective: This naturalistic, repeated-measures, retrospective study investigated the impact of iCBT on work productivity metrics using nationally representative data from patients enrolled in the Irish national health service (ie, the Health Service Executive).

Methods: We analyzed repeated measures retrospective data from 7125 employed patients enrolled in iCBT at the Health Service Executive between March 2023 and May 2024. The Work Productivity and Activity Impairment questionnaire was used to measure absenteeism, presenteeism, overall productivity loss, and activity impairment. Secondary outcomes included depression (Patient Health Questionnaire-9) and anxiety (Generalized Anxiety Disorder-7). Patients were primarily 25 to 64 years old (n=5578, 78%), female (n=4956, 70%), and met clinical scoring criteria on the Patient Health Questionnaire-9 or Generalized Anxiety Disorder-7 (n=4774, 67%). Missing data were handled using multiple imputation. We used mixed-effects models to assess pre-post treatment changes in outcomes and then utilized Irish national salary estimates from 2022 to derive cost savings (in 2022 € values; €1=approximately US $1.05) based on productivity improvement during use of the iCBT program.

Results: From baseline to follow-up, absenteeism reduced by 6.85% (P<.001, 95% CI 5.79% - 7.91%, Cohen d=0.21), presenteeism reduced by 5.84% (P<.001, 95% CI 4.59% - 7.09%, Cohen d=0.18), productivity loss reduced by 9.48% (P<.001, 95% CI 8.30% - 10.66%, Cohen d=0.27), and activity impairment reduced by 8.34% (P<.001, 95% CI 7.07% - 9.61%, Cohen d=0.30). Depression symptom scores reduced by 2.70 points (P<.001, 95% CI 2.50 - 2.90, Cohen d=0.51) and anxiety symptom scores reduced by 2.71 points (P<.001, 95% CI 2.51 - 2.91, Cohen d=0.52) scores posttreatment. Larger reductions in depression and anxiety symptoms were linked to greater improvements in workplace functioning outcomes (r=0.06-0.34, all P<.001). Patients with higher baseline clinical severity experienced approximately 6% greater improvements in workplace productivity than subclinical patients (P<.001). These improvements in work productivity from baseline to follow-up corresponded to more than €4000 annual savings per patient treated, which equated to an estimated annual savings of €29 million in the sample (in 2022 € values). The average currency exchange rate in 2022 would be €1 to approximately US $1.05.

Conclusions: Building on work that primarily focused on work productivity improvements from iCBT in occupational contexts, this study provides evidence that iCBT for the general adult population in a national routine health care setting is associated with improvements in both Common mental disorders (CMD) and workplace productivity losses associated with CMD. These productivity improvements were further associated with sizable cost savings. These findings suggest that integrating iCBT into

XSL•FO
RenderX

national mental health infrastructure can be a feasible, scalable solution, which could generate clinical and economic benefits at a population level.

(J Med Internet Res 2026;28:e80689) doi:10.2196/80689

KEYWORDS

digital mental health; internet-based cognitive behavioral therapy; iCBT; workplace productivity; depression; anxiety; cost effectiveness; health services

Introduction

Common mental disorders (CMDs) like depression and anxiety impact millions of people worldwide [1], resulting in negative social, functional, and health outcomes for individuals [2], as well as significant economic costs. Indeed, the global economy suffers an annual loss of US $1.15 trillion due to CMD, with one-third of this loss attributed to lost employment and productivity [3]. Compared to healthy individuals, 4.7 billion additional workdays are lost, amounting to a total cost of US $592 billion per year [3]. These significant implications warrant cost-effective interventions for CMD, with cognitive behavioral therapy (CBT) showing promise as an affordable, accessible solution when delivered online. Internet-based cognitive behavioral therapy (iCBT) has consistently demonstrated effectiveness in treating CMD across various settings [4-7], with cost-effectiveness [8] superior to other treatments or no treatment [9,10]. One key area to determine the cost-effectiveness of iCBT is through the indirect costs of CMD, such as the costs of decreased workplace productivity due to CMD [11]. Interestingly, recent systematic reviews on economic evaluations of iCBT for CMD suggest iCBT economic evaluations typically focus on direct costs (eg, costs of treatment, prescriptions, medical care), and indirect costs are frequently lacking or understudied [8,12].

CMD can accrue indirect costs through the impairment of workplace productivity, such as by increasing absenteeism (increased work absences), presenteeism (reduced effectiveness at work), higher rates of disability leaves, and early retirement [13-15]. These effects are observed at both clinical and subclinical levels of CMD [16,17]. A common method of measuring workplace productivity loss due to health conditions is through the Work Productivity and Activity Impairment (WPAI) questionnaire [18-20]. This questionnaire allows for the assessment of absenteeism, presenteeism, and overall productivity loss (the combination of absenteeism and presenteeism), as well as general activity impairment (the degree to which a health problem impairs daily activities outside of work), though the activity impairment index is less relevant for examining economic consequences. Focusing on the WPAI, a recent systematic review found that depression was associated with approximately 26 work absences per year and a 37% to 52% reduction in effectiveness at work [21].

Research conducted in occupational contexts suggests that iCBT can also improve absenteeism, presenteeism, and overall productivity [22-28], which, in turn, generates substantial financial returns for employers [22-24,29]. However, access to effective interventions through an employer is not always guaranteed. Despite growing awareness and actions addressing employee well-being, adoption and utilization rates of workplace mental health programs are low and dependent on company structure and culture [30,31]. Alternatively, iCBT administered to the general population through a public health service could provide a different approach to improving work productivity losses due to CMD. However, whether the positive impact of iCBT on work-related outcomes can generalize beyond workplace settings remains unclear.

The integration of iCBT into public health services can greatly increase the reach and cost-effectiveness of mental health care nationwide [25]. This expanded reach should potentially amplify any associated benefits from improved workplace productivity. In Ireland, the Health Service Executive (HSE) provides comprehensive, publicly funded mental health services for treating CMD. The HSE incorporated iCBT from *SilverCloud Health* by Amwell as part of their offering in 2021, making it one of the first nationally implemented iCBT services. Previous work on SilverCloud has shown that it is effective in treating CMD within the HSE and has yielded acceptable service-level metrics and high levels of user satisfaction [32]. Outside of the HSE, previous work has supported SilverCloud as being cost-effective within the United Kingdom's Improving Access to Talking Therapies, though this prior study was not able to incorporate or specifically examine costs associated with workplace productivity loss [33].

Building on this, we investigated the impact of providing iCBT in a national health service on workplace functioning, including absenteeism, presenteeism, productivity loss, and activity impairment. To do so, we analyzed mental health and work productivity routine care data collected from patients enrolled in SilverCloud at the HSE. We hypothesized that both mental health symptoms and workplace functioning would improve posttreatment. We further estimated the cost savings incurred by observed changes in work-related outcomes.

Methods

Study Setting and Participants

This naturalistic study analyzed repeated measures retrospective data from n=11,131 patients enrolled in SilverCloud within Ireland's HSE during the time period in which the WPAI questionnaire was administered (from March 8, 2023 to May 27, 2024). Note that this study adhered to CHEERS (Consolidated Health Economic Evaluation Reporting Standards) 2022 guidelines for economic evaluations (see Checklist 1) [34]. Within the HSE, patients are referred to SilverCloud through health professionals across 5 referral groups (ie, General Practitioners, Primary Care Psychology, Counseling in Primary Care, Community Mental Health, and an HSE-funded

XSL•FO
RenderX

mental health charity called Jigsaw). The HSE and SilverCloud staff both provided communications to these referral groups detailing the availability of this iCBT service, who would be suitable for it, and how to have a referral conversation with their patients.

When considering a patient for SilverCloud, health professionals perform an initial screening to assess whether they may be suitable for the intervention. Eligible patients for the treatment (and the study) were at least 18 years old and presented with common mental health issues (eg, low mood, anxiety, feelings of isolation, stress, poor sleep). Patients were ineligible if they had severe mental health symptoms such as manic or psychotic episodes and/or suicidal ideation. If a patient is suitable, the health professional submits a referral to SilverCloud, which triggers an email invite sent to the patient through the SilverCloud platform. Within the email, an invitation link redirects the client to the sign-up page on the SilverCloud website. Patients who do not activate their account receive weekly reminder emails for the first 30 days following the initial invitation. During the sign-up process, patients are asked to provide demographic information, complete baseline clinical questionnaires, and can choose to consent for their pseudoanonymized data to be used for clinical research and service evaluation purposes. Once sign-up is complete, patients are directed to the SilverCloud homepage where they can begin the program.

Intervention

Within the HSE, SilverCloud's *Space from Depression*, *Space from Anxiety*, *Space from Depression and Anxiety*, and *Space from Generalized Anxiety Disorder* programs were offered during the study time period. These are general purpose iCBT programs that consist of a series of modules aligned with evidence-based CBT principles for the treatment of depression and anxiety [33,35]. Specifically, each program aims to help the patient develop and increase an understanding of the relationship between one's thoughts, feelings, and behaviors, alongside information about coping strategies and therapeutic techniques such as graded exposure, behavioral activation, cognitive restructuring, problem solving, and activity scheduling. This content is delivered through various interactive media and tools, including quizzes, videos, activities, and personal stories. During program use, patients also complete routine outcome measures (ie, Patient Health Questionnaire-9 [PHQ-9], Generalized Anxiety Disorder-7 [GAD-7], WPAI, see below) at intake, week 2, week 4, and week 8 of treatment. Risk is monitored continuously through item 9 on the PHQ-9.

Patients are also assigned a supporter during their use of SilverCloud. All supporters are graduate psychologists with at least a master's degree or doctorate in progress who have been trained in the delivery of online support. Supporters also receive training specific to providing support on SilverCloud and receive supervision from clinical supervisors. Supporters provide an asynchronous written "review" in the platform to the patient. The primary aim of the review is to provide the patient with personalized guidance and feedback based on their progress and clinical symptomology. Supporter reviews are provided weekly for the first 6 weeks, though additional reviews may be provided at the discretion of the supporter and their clinical supervisor. Once all the reviews are provided, or if the patient does not engage with the program for 3 consecutive review periods, the supporter discharges the patient from the supported service and no longer provides reviews. After patients are discharged, they still have full access to the platform in an unsupported capacity.

Assessments

Demographic Questionnaire

During intake, clients completed demographic questions about gender, education, race, ethnicity, age, work status, and marital status. They are also asked whether they are in concurrent mental health treatment and to rate how likely they think SilverCloud will work for them.

Routine Outcome Measures

The following questionnaires are part of the HSE's routine outcome measures that were administered at intake, week 2, week 4, and week 8 of treatment.

WPAI: Specific Health Problem

The WPAI: Specific Health Problem is a 6-item questionnaire that probes the impact of a specified health problem on workplace functioning over the past 7 days [19]. Outcomes include absenteeism (work time missed), presenteeism (impairment while working), overall productivity loss (absenteeism and presenteeism combined), and nonwork activity impairment (see Multimedia Appendix 1). Each outcome is expressed as a percentage, with higher numbers indicating greater impairment. The measure has shown good construct validity and test-retest reliability [19].

Patient Health Questionnaire-9

The PHQ-9 is a 9-item self-report questionnaire for depressive symptoms, with good reliability, sensitivity, and specificity [36]. The total score ranges from 0 to 27, with each item rated on a 4-point Likert scale (0=*not at all* to 3=*nearly every day*). The scale has 4 cut-off points of 5, 10, 15, and 20, which represent mild, moderate, moderately severe, and severe levels of depression, respectively.

Generalized Anxiety Disorder-7

The GAD-7 is a 7-item self-report questionnaire that identifies individuals with generalized anxiety disorder [37]. Items are scored on a 4-point Likert scale (0=*not at all* to 3=*nearly every day*). The total score can range from 0 to 21, with three cut-off points (5, 10, and 15) representing mild, moderate, and severe anxiety levels, respectively. The scale has also demonstrated good reliability, sensitivity, and specificity [37].

Ethical Considerations

All procedures of this study adhered to the ethical standards of the relevant national and institutional committees on human experimentation and the Declaration of Helsinki (1975, revised in 2008). The protocol-related documents and informed consent were reviewed and approved by the Trinity College Dublin School of Psychology Research Ethics Committee (Approval ID: SPREC112021-04). Approval for users to not provide informed consent was granted, as consent had already been

XSL•FO
RenderX

given for treatment and for therapists to gather and examine clinical outcome measures as part of normal service evaluation procedures. Service users who refused consent to having their data passed on for these purposes were not included in this study. Study data were collected and stored in a deidentified manner and complied with General Data Protection Regulations. Patients were not compensated for this study, as the data were derived from routine service assessments administered during routine treatment. No images of patients were used in this study or are presented in this manuscript or supplemental materials.

Data Analysis

The study sample size was based on all available data from all patients who reported being employed via the WPAI questionnaire. Each patient's last assessment prior to treatment discharge was used as their follow-up measure. We included all patients who completed the baseline assessment and used multiple imputation to estimate follow-up CMD and workplace functioning scores for those missing follow-up data. Performing the Little Missing Completely At Random test suggested WPAI data were not missing at random (χ^2_{27}=3921.24, P<.001). Multiple imputation is a conservative method for handling missing data as it incorporates natural variability into the missing values [38]. Pooled estimates from the imputation are reported. For sensitivity purposes, core analyses were repeated using complete-case data (Multimedia Appendix 2). Descriptive analyses were conducted on patients' demographics, baseline clinical and work-related outcomes, as well as treatment characteristics.

First, univariate regressions were used to test whether CMD (depression and anxiety) (independent variable) and workplace functioning (dependent variable) were associated at baseline. To identify potential confounds, we tested the association between changes in WPAI outcomes and covariates, including demographics (age, gender, education, ethnicity, marital status, long-term conditions) and treatment-related characteristics (concurrent treatment, expectation), which were coded according to Table 1. Due to our naturalistic study design, the time from baseline to follow-up varied considerably (mean 34.89, SD 14.68 d). We controlled for this by evaluating the assessment interval as a potential covariate, which captured the treatment duration relevant to the measured effects. Next, to assess whether workplace functioning and CMD improved posttreatment, we employed mixed-effects models, regressing workplace functioning and CMD outcomes (dependent variable) on time as a fixed effect (independent variable), including significant covariates previously identified. To address repeated observations nested within patients across time, we incorporated random intercepts for patients in our models. To quantify the pre-post change in WPAI outcomes, we calculated Cohen *d* effect sizes [39]. We explored baseline CMD severity and comorbidity as moderators to examine their effects on changes in workplace functioning.

To calculate productivity-related cost savings from the intervention, we adopted the human capital approach, a widely established method for estimating the monetary value associated with productivity losses. The method values 1 hour of productivity loss based on the individual's hourly wages [40]. We obtained median hourly salary estimates (stratified by gender and education) from available Irish census data in the 2022 Structure of Earnings Survey [41]. Thus, estimates are in 2022 € values (€1=approximately US $1.05) [42]. We then multiplied these estimates by the percentages of absenteeism, presenteeism, and overall productivity loss to determine the cost, scaling it accordingly to weekly and yearly estimates, assuming a full-time 39-hour work week. We did not apply a discount rate given the short timeframe of the study. We calculated this per patient and used the mean for analysis.

All analyses were conducted in R (version 4.4.0), and all tests were 2-sided, with statistical significance set at α=.05.

XSL•FO
RenderX

Table . Baseline characteristics of the study sample[a].

Characteristics	Participants
Age (y), n (%)	
18 - 24	1461 (20.64)
25 - 64	5578 (78.82)
65+	38 (0.54)
Gender, n (%)	
Female	4956 (70.23)
Male	2067 (29.29)
Other/Prefer not to say	34 (0.48)
Race and ethnicity, n (%)	
White Irish	6023 (85.32)
Other White European	652 (9.24)
Asian	99 (1.40)
Mixed	74 (1.05)
Black	71 (1.01)
Latino	62 (0.88)
Other	78 (1.10)
Education, n (%)	
Primary to secondary	2411 (34.41)
College/university	3614 (51.58)
Postgraduate	982 (14.01)
Marital status, n (%)	
Single	2134 (30.30)
In a relationship	2332 (33.12)
Married	2155 (30.60)
Separated/divorced	376 (5.34)
Widowed	45 (0.64)
Long-term condition, n (%)	
Yes	1605 (22.87)
No	5413 (77.13)
Concurrent treatment, n (%)	
Yes	3766 (53.34)
No	3295 (46.66)
Depression severity (PHQ-9)[b], mean (SD)	12.82 (6)
Minimal, n (%)	546 (7.66)
Mild, n (%)	1832 (25.71)
Moderate, n (%)	1982 (27.82)
Moderately severe, n (%)	1621 (22.75)
Severe, n (%)	1144 (16.06)
Anxiety severity (GAD-7)[c], mean (SD)	12.62 (5.38)
Minimal, n (%)	457 (6.41)
Mild, n (%)	1830 (25.68)

XSL•FO
RenderX

Characteristics	Participants
Moderate, n (%)	1903 (26.71)
Severe, n (%)	2935 (41.19)
WPAI[d] absenteeism (%), mean (SD)	19.45 (33.88)
WPAI presenteeism (%), mean (SD)	34.17 (30.04)
WPAI productivity loss (%), mean (SD)	49.19 (34.44)
WPAI activity impairment (%), mean (SD)	49.74 (27.87)
Treatment duration (d), mean (SD)	34.89 (14.68)
Treatment expectation (0 - 4), mean (SD)	2.17 (0.65)

[a]Summary descriptives were calculated based on the total responses per variable, which may differ due to missing data. The proportion of missingness for all variables was less than 2.5%, except for baseline presenteeism (6%), productivity loss (6%), and activity impairment (8%).

[b]PHQ-9: Patient Health Questionnaire-9.

[c]GAD-7: Generalized Anxiety Disorder-7.

[d]WPAI: Workplace Productivity and Activity Impairment.

Results

Patient Characteristics

The participant flow chart is shown in Figure 1. Of the 11,810 patients enrolled in the national iCBT service, 11,131 completed the baseline assessment fully or partially, of whom 7125 (64%) reported being employed. These 7125 participants formed the final sample for analyses.

XSL•FO
RenderX

Figure 1. Participant flow chart. Participants may have completed their baseline and follow-up assessments fully or partially, as completion of the assessment was not compulsory for them to progress in the treatment. WPAI: Work Productivity and Activity Impairment.

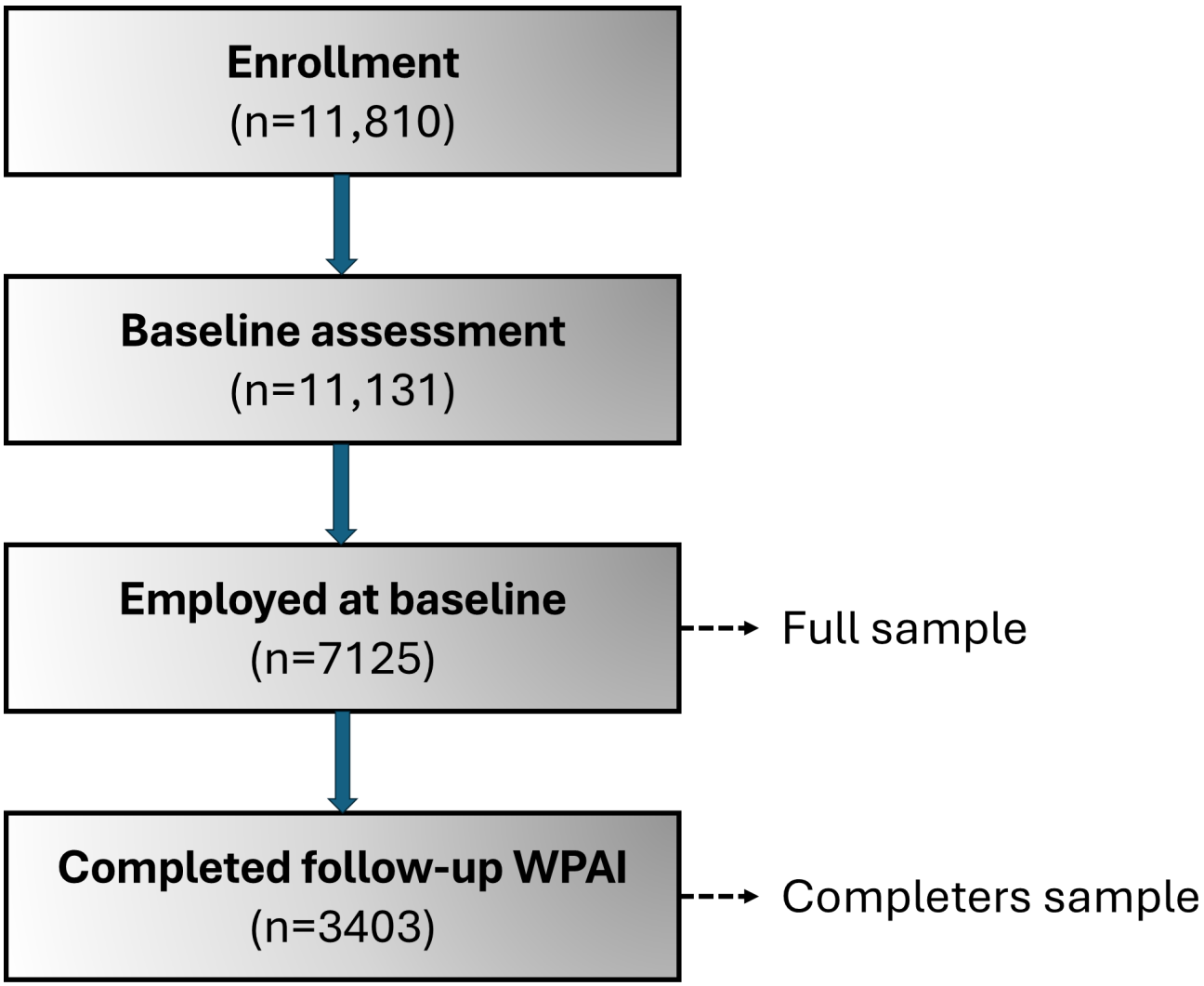

Baseline Analyses

Table 1 outlines the characteristics of our sample. Patients were primarily 25 to 64 years old, female, White Irish, had college or university education, in romantic relationships, and not presenting with a physical long-term condition. Their baseline severity on average fell within the moderate classification of depression (PHQ-9) and anxiety (GAD-7), respectively. Over half were receiving concurrent treatment in addition to iCBT, which most were undertaking for approximately 35 days. Overall, there was a positive expectation regarding the benefit that users believed they would gain from the treatment, with approximately 90% indicating that they believed that it was at least 'somewhat likely' to work for them. Regarding workplace functioning, patients were on average absent from work 20% of the week and had 34% reduced performance while working during the week. Both productivity loss and activity impairment averaged near 50%.

A series of univariate regressions revealed positive associations between both PHQ-9 and GAD-7 and all four WPAI outcomes at baseline (Table 2), where greater levels of depression and anxiety predicted increased absenteeism, presenteeism, overall productivity loss, and activity impairment (all P<.001). More specifically, every 1-point increase on the PHQ-9 and GAD-7 predicted approximately a 1% to 2.6% decrease in workplace functioning.

XSL•FO
RenderX

Table . Baseline associations between clinical and work-related outcomes at baseline, as assessed using univariate linear regressions.

Work-related outcomes and CMD[a]	Univariate analyses			
	Unstandardized β (95% CI)	SE	*t* test (*df*)	*P* value
Absenteeism				
PHQ-9[b]	1.36 (1.24 - 1.48)	0.06	20.89 (7119.64)	<.001
GAD-7[c]	1.01 (0.87 - 1.15)	0.07	13.76 (7120.21)	<.001
Presenteeism				
PHQ-9	1.43 (1.31 - 1.54)	0.06	24.28 (2451.73)	<.001
GAD-7	1.56 (1.42 - 1.70)	0.07	23.34 (1358.33)	<.001
Productivity loss				
PHQ-9	2.30 (2.18 - 2.42)	0.06	36.38 (3893.65)	<.001
GAD-7	2.13 (1.99 - 2.27)	0.07	28.84 (1972.90)	<.001
Activity impairment				
PHQ-9	2.55 (2.45 - 2.65)	0.05	52.26 (1090.67)	<.001
GAD-7	2.26 (2.14 - 2.38)	0.06	39.88 (3198.07)	<.001

[a]CMD: common mental disorder.

[b]PHQ-9: Patient Health Questionnaire-9.

[c]GAD-7: Generalized Anxiety Disorder-7.

Pre-Post Changes in Workplace Functioning

Over an average treatment period of 35 days, we observed significant improvements across all WPAI outcomes. Specifically, there was an average reduction of 6.85% in absenteeism (b=−6.85, $t_{71.47}$=−12.75, P<.001, 95% CI −5.79 to −7.91, Cohen d=0.21), 5.84% in presenteeism (b=−5.84, $t_{38.51}$=−9.15, P<.001, 95% CI −4.59 to −7.09, Cohen d=0.18), 9.48% in productivity loss (b=−9.48, $t_{61.18}$=−15.70, P<.001, 95% CI −8.30 to −10.66, Cohen d=0.27), and 8.34% in activity impairment (b=−8.34, $t_{29.13}$=−12.88, P<.001, 95% CI −7.07 to −9.61, Cohen d=0.30) (Figure 2). Covariates such as treatment duration, treatment expectation, gender, concurrent treatment, and marital status were significant predictors of change in one or more WPAI outcomes. Importantly, all observed pre-post changes in WPAI outcomes remained statistically significant after adjusting for these confounders (all P<.001) (Multimedia Appendix 3).

XSL•FO
RenderX

Figure 2. Pooled mean impairment at baseline and follow-up for absenteeism, presenteeism, productivity loss, and activity impairment. Error bars denote the 95% CI.

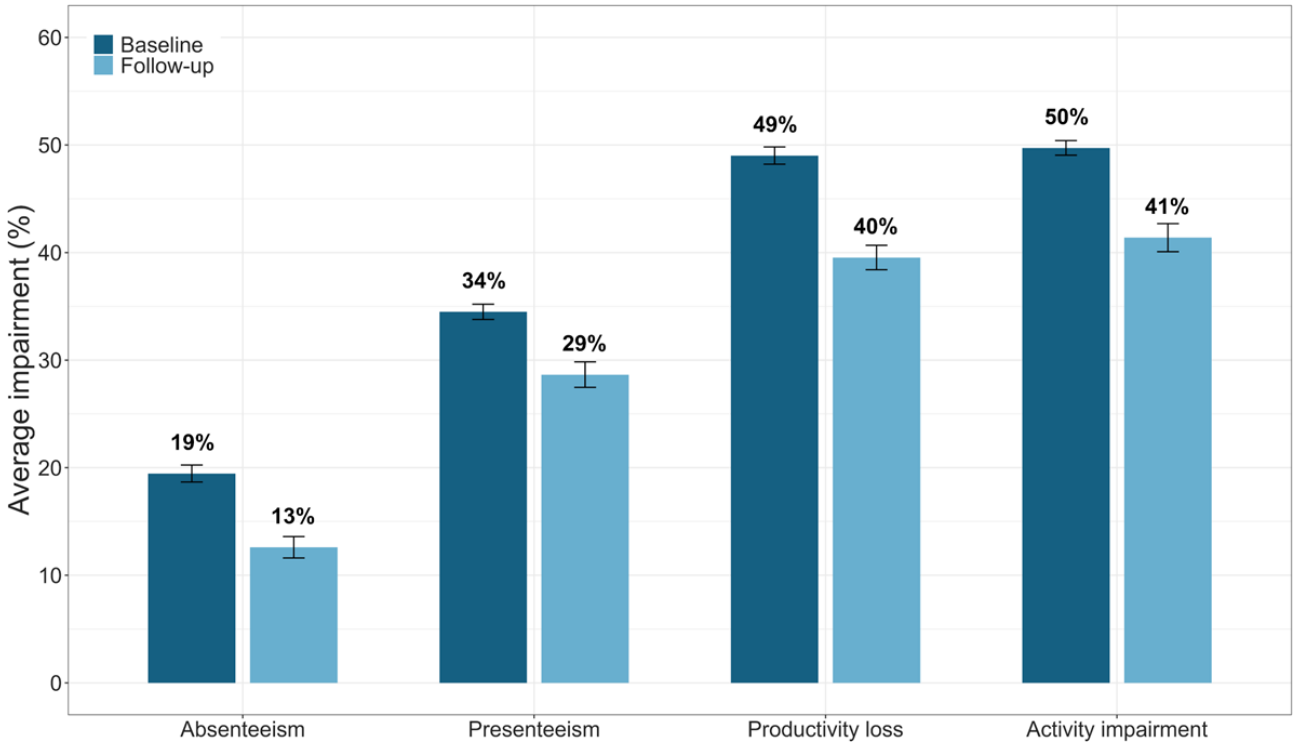

For depression and anxiety, we further observed a significant reduction in both PHQ-9 (b=−2.70, $t_{42.71}$=-27.11, P<.001, 95% CI −2.50 to −2.90, Cohen d=0.51) and GAD-7 (b=−2.71, $t_{39.36}$=−26.52, P<.001, 95% CI −2.51 to −2.91, Cohen d=0.52) scores posttreatment. Partial correlations accounting for significant covariates previously identified revealed that these clinical symptom reductions were also positively associated with changes in absenteeism (PHQ-9: r=0.10, 95% CI 0.07 to 0.13; GAD-7: r=0.06, 95% CI 0.03 to 0.09), presenteeism (PHQ-9: r=0.18, 95% CI 0.14 to 0.21; GAD-7: r=0.21, 95% CI 0.18 to 0.24), productivity loss (PHQ-9: r=0.21, 95% CI 0.18 to 0.24; GAD-7: r=0.21, 95% CI 0.18 to 0.24), and activity impairment (PHQ-9: r=0.34, 95% CI 0.31 to 0.37; GAD-7: r=0.33, 95% CI 0.30 to 0.36; all P<.001). This pattern indicates that greater improvements in depression and anxiety symptoms were associated with greater improvements in each WPAI outcome.

To examine whether baseline CMD severity, that is, clinical (moderate to severe symptoms) vs subclinical (minimal to mild symptoms), influences treatment effects on workplace functioning, we included baseline PHQ-9 and GAD-7 scores as moderators, along with covariates previously identified for each outcome (see Multimedia Appendix 4 for demographic characteristics by CMD severity). Significant time × baseline PHQ-9 interaction effects suggested that patients with higher baseline depression experienced larger reductions in absenteeism (b=−5.68, $t_{130.37}$=-5.55, 95% CI −7.68 to −3.68; P<.001), presenteeism (b=−2.48, $t_{50.02}$=−2.03, 95% CI −4.87 to −0.09; P=.048), productivity loss (b=−5.32, $t_{80.23}$=−4.43, 95% CI −6.35 to −2.87; P<.001), and activity impairment (b=−4.61, $t_{107.97}$=−5.17, 95% CI −9.61 to −7.07; P<.001). More specifically, those who were clinical at baseline experienced larger reductions than their subclinical counterparts in absenteeism (subclinical: b=−3.06, 95% CI −1.59 to −4.53, P<.001; clinical: b=−8.74, 95% CI −7.35 to −10.13, P<.001), presenteeism (subclinical: b=−4.18, 95% CI −2.55 to −5.81, P<.001; clinical: b=−6.67, 95% CI −4.95 to −8.39, P<.001), productivity (subclinical: b=−5.93, 95% CI −4.13 to −7.73, P<.001; clinical: b=−11.25, 95% CI −9.72 to −12.78, P<.001), and activity impairment (subclinical: b=−5.27, 95% CI −3.80 to −6.74, P<.001; clinical: b=−9.88, 95% CI −8.35 to −11.41, P<.001; Figure 3 and Multimedia Appendix 4).

A similar trend was also observed for anxiety, where those entering treatment with higher GAD-7 scores experienced larger reductions in workplace functioning (absenteeism: b=−4.87, $t_{380.35}$=−5.21, 95% CI −6.69 to −3.05, P<.001; presenteeism: b=−4.62, $t_{76.03}$=−4.17, 95% CI −6.80 to −2.44, P<.001; productivity loss: b=−6.61, $t_{102.34}$=−5.70, 95% CI −8.88 to −4.34, P<.001; activity impairment: b=−5.67, $t_{75.85}$=−5.90, 95% CI −7.55 to −3.79, P<.001). Patients with clinical score levels on the GAD-7 had larger decreases than subclinical patients in absenteeism (subclinical: b=−3.54, 95% CI −1.99 to −5.09, P<.001; clinical: b=−8.41, 95% CI −7.16 to −9.66, P<.001), presenteeism (subclinical: b=−2.70, 95% CI −0.96 to −4.44, P=.02; clinical: b=−6.19, 95% CI −4.64 to −7.74, P<.001), productivity loss (subclinical: b=−4.99, 95% CI −3.09 to −6.89, P<.001; clinical: b=−11.60, 95% CI −10.19 to −13.01, P<.001), as well as activity impairment (subclinical: b=−4.50, 95% CI −2.79 to −6.21, P<.001; clinical: b=−10.20, 95% CI −8.73 to −11.67, P<.001) (Figure 3 and Multimedia Appendix 4).

Importantly, we found that patients experiencing comorbid clinical levels of depression and anxiety at baseline also improved more in workplace functioning than those who were noncomorbid. This was evidenced by significant time ×

XSL•FO
RenderX

comorbidity interaction effects observed for all WPAI outcomes (all $P<.001$, Multimedia Appendix 5).

Figure 3. Changes in work-related outcomes from baseline to follow-up, moderated by baseline depression severity (top row) and baseline anxiety severity (bottom row). The points represent unadjusted means, and the error bars represent SE. GAD7: Generalized Anxiety Disorder-7; PHQ9: Patient Health Questionnaire-9.

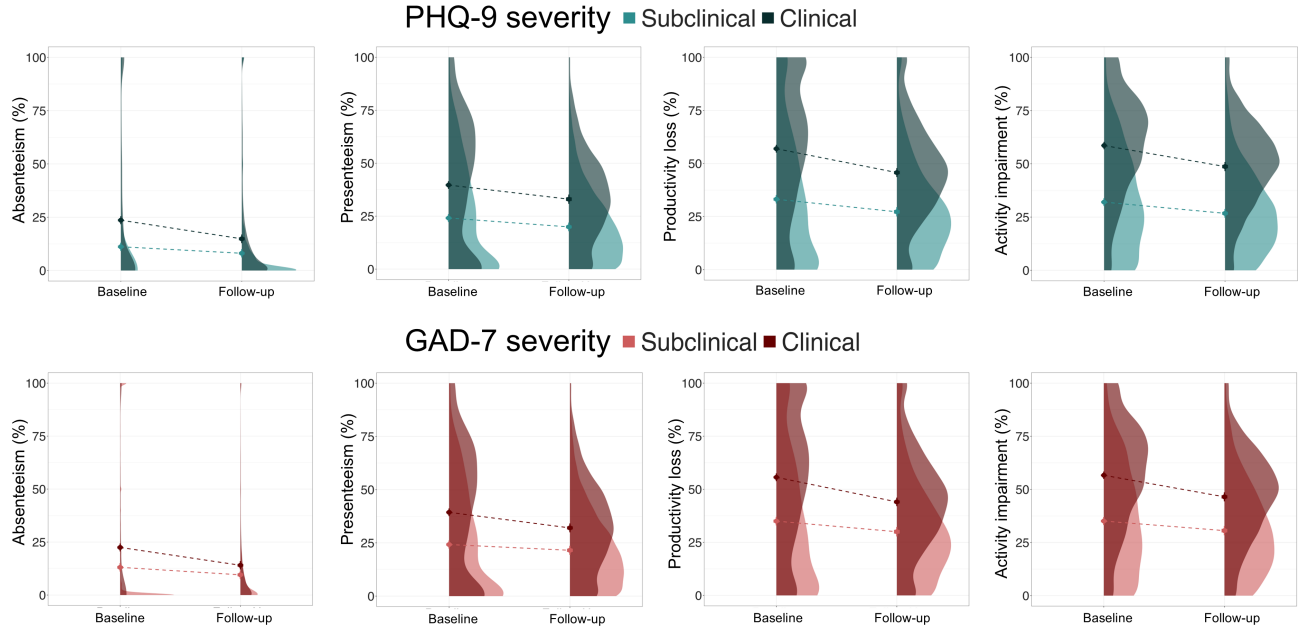

Cost Analysis

Based on 2022 Irish census median salary estimates by gender and educational attainment, the average median salary in the sample was estimated to be €836.06 (95% CI €830.29-€841.83, SD €248.62) per week, which is, €43,475.34 (95% CI €43,175.15-€43,775.43, SD €12,928.16) per year. At baseline, we calculated the cost induced by deficits in workplace functioning for each patient (WPAI % multiplied by salary), before totaling these values to produce an average. The average costs of absenteeism and presenteeism per patient per week were estimated to be €159.82 (95% CI €153.00-€166.64, SD €293.59) and €290.80 (95% CI €284.30-€297.30, SD €280.04), totaling to an annual monetary loss of €8310.64 (95% CI €7955.52-€8664.48, SD €15,266.39) and €15,121.60 (95% CI €14,782.87-€15,459.13, SD €14,562.22). For overall productivity loss, the total cost induced is nearly half of each patient's earnings, estimated at €409.87 (95% CI €402.41-€417.33, SD €321.29) per week and €21,313.26 (95% CI €20,925.32-€21,701.20, SD €16,707.25) per year.

To estimate the annual cost savings due to improvements in workplace functioning, we multiplied the adjusted mean reduction in absenteeism (6.85%), presenteeism (5.84%), and productivity loss (9.48%) by the average salary of our sample. With each percent change in WPAI outcomes, annual costs were estimated to decline by €434.75 per patient. This results in annual cost savings of €2978.06 for absenteeism (95% CI €2546.94-€3409.18), €2538.96 for presenteeism (95% CI €1989.67-€3088.25), and €4121.11 for overall productivity loss (95% CI €3589.11-€4645.11) for each patient treated (Figure 4A). In the context of our study sample (n=7125), this improvement equates to an annual cost saving of €29.33 million.

Importantly, when we consider CMD severity and comorbidity of patients at baseline, cost savings were more substantial for those presenting with more severe symptomatology (Figure 4). Considering overall productivity loss, annual cost savings for patients with clinical depression (11.25% improvement) were 1.90 times higher compared to subclinical patients (5.93% improvement) (average cost savings for clinical depression: €4892.98, 95% CI €4233.28-€5550.72 vs subclinical depression: €2579.04, 95% CI €1762.23-€3395.77; Figure 4B). For anxiety, annual cost savings for clinical patients (11.60% improvement) were 2.33 times higher than subclinical patients (4.99% improvement) for overall productivity loss (average cost savings for clinical anxiety: €5042.36, 95% CI €4402.67-€5682.71 vs subclinical anxiety: €2169.02, 95% CI €1317.28-€3020.76; Figure 4C). Similarly, cost savings are 2.05 times higher for comorbid patients (12.23% improvement) compared to their counterparts (5.98% improvement) (average cost savings for non-comorbid: €2601.92, 95% CI €1890.44-€3313.40 vs comorbid: €5318.41, 95% CI €4623.67-€6013.15; Figure 4D).

XSL•FO
RenderX

Figure 4. (A) Estimated average annual cost savings (in 2022 € values; €1=approximately US $1.05) postintervention induced by improvements in absenteeism, presenteeism, and overall productivity loss for patients with clinical levels versus subclinical levels of (B) depression, (C) anxiety, and for (D) patients with comorbid clinical levels of depression and anxiety vs those without. Monetary figures were rounded to the nearest integer. Error bars denote the 95% CI.

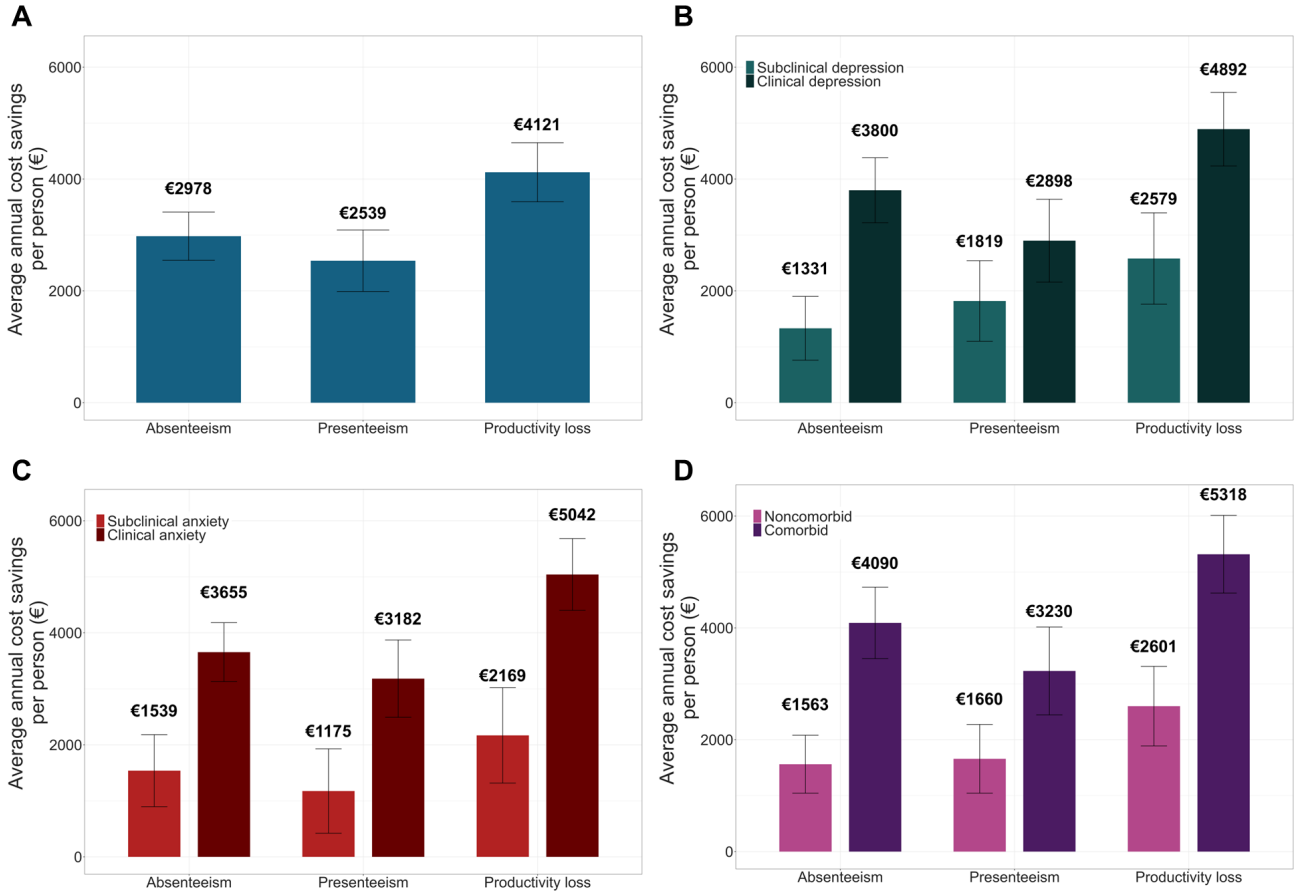

Sensitivity Analysis

To determine whether multiple imputation introduced bias into our findings, we repeated the core analysis, including only patients who completed both baseline and follow-up assessments (n=3402) (Multimedia Appendix 2). The results remained consistent, showing significant improvements in all CMD and WPAI outcomes, with greater improvements among patients with clinical levels of CMD and comorbid CMD compared to their counterparts.

Discussion

This study is the first to use nationwide routine care data to investigate the effectiveness of iCBT, delivered as a public health initiative for treating CMD, in improving workplace functioning in the general population. These improvements in work productivity were estimated to lead to substantial cost savings. As hypothesized, patients reported fewer missed workdays, less impairment during work and activities, and higher overall productivity after 5 weeks of treatment. Although the effects observed here (6% - 10% reductions) were smaller than those in prior studies that included small, completer-only samples with limited generalizability [22,29], they align more closely with those reported in meta-analyses of work-focused CBT or iCBT interventions on work absences and effectiveness (Hedges g=0.22 - 0.25) [25,43]. Work-related improvements were observed alongside reductions in depression and anxiety symptoms, which were consistent with our previous study [32].

Notably, iCBT improved workplace functioning for both clinical and subclinical patients, with greater benefits for those with more severe symptomatology at baseline. This finding not only highlights iCBT as an effective intervention for improving workplace functioning in individuals with clinically apparent symptoms but also for subclinical individuals, helping to prevent their symptoms from reaching clinical levels. Note that these subclinical individuals likely represent the typical working population [22]. Future research should explore whether engagement with standard vs work-focused content differentially impacts improvements in workplace functioning, again using more representative data from the general working population. This is particularly pertinent given the limited evidence supporting the superiority of tailored over nontailored digital interventions [26,27,44].

The consequences and scale of impaired workplace functioning due to mental illness present major public health challenges, necessitating the implementation of cost-effective and accessible interventions [3,44]. Our study builds on the promising evidence of iCBT interventions in improving clinical, work-related, and financial outcomes in occupational contexts [22-24] and demonstrates the scalability of these benefits when iCBT is provided in a national health service. By expanding the intervention's reach beyond industry-specific settings and

XSL•FO
RenderX

barriers, these outcomes have significant implications for the economic health of the nation. This is evident in the measurable cost savings demonstrated in our study, estimated at more than €4000 over the course of a year per patient treated, totaling €29 million annually within our sample. These findings are aligned with the conclusions of the few economic evaluations on workplace iCBT interventions [22-24,29]. The findings also dovetail with other mental health and economic evaluations focusing on iCBT implemented within national health care systems, such as MindSpot in Australia and SilverCloud within Talking Therapies in the United Kingdom, though these prior studies were not focused on cost-effectiveness through workplace productivity [33,45]

These cost-savings estimates should also be seen as conservative, given that they focus on costs due to workplace impairment and do not capture medical costs, costs of providing the treatment, or broader societal benefits (eg, reduced welfare and health care expenditure, better health care utilization, and positive spillover effects on patients' support networks) [3,46]. While it is possible that expanding treatment coverage in the population may increase the average cost per person initially, particularly after reaching previously inaccessible patients in remote areas, the expected financial returns from improved health and work-related outcomes have been shown to far outweigh the treatment costs for depression and anxiety (return ratio of 2.3-3:1 considering only economic benefits, 3.3 - 5.7:1 when including health returns) [3]. Furthermore, our findings highlight that iCBT is uniquely positioned to bridge the substantial treatment gap between demand and availability on a national level. As a feasible, scalable solution, iCBT can alleviate the strain of CMDs on overburdened health care systems and improve workforce productivity, all the while allowing greater access to care at lower cost compared to in-person approaches [47,48].

This study has several limitations. The naturalistic design of our study, without comparison groups, prevents us from determining the causality of the observed treatment effects. Estimates from this study could serve as benchmarks for future studies focused on evaluating outcomes and the cost-effectiveness of mental health treatments within the HSE. Outcomes were also collected at fixed intervals set by the service, with the latest assessment administered at week 8 of treatment. Thus, the extent to which treatment effects on symptoms persist over longer time frames (eg, 6 mo, 1 y) is unknown in this sample. This may particularly bias study estimates projecting annual cost savings. Though prior work investigating the treatment durability of the SilverCloud iCBT program suggests that approximately 70% of patients who recover from symptoms remain in recovery 9 months later [49]. The study also relies on self-report data, uncorroborated by objective and formal assessments, which may be more accurate. Another limitation is the precision of our estimates due to the nature of the available data. As WPAI outcomes can only be calculated for employed individuals [19], this fails to account for those unemployed due to their condition (ie, complete loss of productivity).

Additionally, while using national salary estimates is a common procedure for estimating cost savings [22-24,29], collecting and incorporating data on patient-specific salary, employment sector, and typical hours worked would produce financial return estimates produced directly from actual participant wage and job data. It is also important to note that we assumed a linear transformation of productivity loss to monetary loss, and it is possible that the association may be nonlinear. Data on intervention costs and other medical costs were also unavailable, which prevented a more comprehensive cost-effectiveness analysis. Future studies should incorporate income and medical cost assessments with comparator treatments using representative data to enable a national economic evaluation.

Depression and anxiety significantly impact labor participation and productivity, but this can be effectively managed with accessible treatments through public health services. Building on work that primarily focused on work productivity improvements from iCBT in occupational contexts, this study provides evidence that iCBT for the general adult population in a national routine health care setting is associated with improvements in both CMD and workplace productivity losses associated with CMD. These productivity improvements were further associated with sizable cost savings. These findings suggest that integrating iCBT into national mental health infrastructure can be a feasible, scalable solution, which could generate clinical and economic benefits at a population level.

Acknowledgments

We extend our gratitude to the Health Service Executive leadership, all relevant stakeholders, and clinicians for their roles in the success of this collaboration. We thank our patients for providing their data for this project. Additionally, we would like to thank all the people who played a role in the dissemination and promotion of this collaboration, specifically Elaine Martin, Hannah Farrell, Paulina Chmara, and Liam Ryan. Artificial Intelligence was not used in the writing of this manuscript.

Funding

This study was not supported by a grant from any funding agency, commercial or not-for-profit sectors; however, CTL, SH, AE, GH, DD, and DR are current employees of SilverCloud Health.

Data Availability

The dataset, along with the analysis scripts for the current study, is available upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

CTL, DR, SH, DD, and AE contributed to the conceptualization of the overall study. CTL, SH, GH, and AE analyzed and interpreted the data. CTL, AJ-D, SH, and AE produced the first draft of the manuscript. All authors edited, revised, and approved the final manuscript.

Conflicts of Interest

CTL, SH, AE, GH, DD, and DR are or were employees of SilverCloud Health during the production of this manuscript. SilverCloud Health is a subsidiary of Amwell. SH, GH, and DR hold shares in Amwell.

Multimedia Appendix 1
Calculation of work productivity activity and impairment outcomes.
[DOCX File, 17 KB - jmir_v28i1e80689_app1.docx]

Multimedia Appendix 2
Core data analyses for completers sample.
[DOCX File, 21 KB - jmir_v28i1e80689_app2.docx]

Multimedia Appendix 3
Covariate analyses examining pre-post changes in Work Productivity and Activity Impairment (WPAI).
[DOCX File, 17 KB - jmir_v28i1e80689_app3.docx]

Multimedia Appendix 4
Pre-post analyses of changes in Work Productivity and Activity Impairment (WPAI) for patients with clinical and subclinical common mental disorders.
[DOCX File, 20 KB - jmir_v28i1e80689_app4.docx]

Multimedia Appendix 5
Pre-post analyses of changes in Work Productivity and Activity Impairment (WPAI) for patients with comorbid and noncomorbid common mental disorders.
[DOCX File, 17 KB - jmir_v28i1e80689_app5.docx]

Checklist 1
CHEERS 2022 checklist.
[DOCX File, 19 KB - jmir_v28i1e80689_app6.docx]

References

1. World mental health report: transforming mental health for all. : World Health Organization; 2022 URL: https://www.who.int/publications/i/item/9789240049338 [accessed 2024-08-05]
2. National Collaborating Centre for Mental Health (UK). Common Mental Health Disorders: Identification and Pathways to Care: British Psychological Society; 2011. [Medline: 22536621]
3. Chisholm D, Sweeny K, Sheehan P, et al. Scaling-up treatment of depression and anxiety: a global return on investment analysis. Lancet Psychiatry 2016 May;3(5):415-424. [doi: 10.1016/S2215-0366(16)30024-4] [Medline: 27083119]
4. Andrews G, Basu A, Cuijpers P, et al. Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: an updated meta-analysis. J Anxiety Disord 2018 Apr;55:70-78. [doi: 10.1016/j.janxdis.2018.01.001] [Medline: 29422409]
5. Cuijpers P, Noma H, Karyotaki E, Cipriani A, Furukawa TA. Effectiveness and acceptability of cognitive behavior therapy delivery formats in adults with depression: a network meta-analysis. JAMA Psychiatry 2019 Jul 1;76(7):700-707. [doi: 10.1001/jamapsychiatry.2019.0268] [Medline: 30994877]
6. Hadjistavropoulos HD, Peynenburg V, Thiessen DL, et al. Utilization, patient characteristics, and longitudinal improvements among patients from a provincially funded transdiagnostic internet-delivered cognitive behavioural therapy program: observational study of trends over 6 years. Can J Psychiatry 2022 Mar;67(3):192-206. [doi: 10.1177/07067437211006873] [Medline: 33840264]
7. Titov N, Hadjistavropoulos HD, Nielssen O, Mohr DC, Andersson G, Dear BF. From research to practice: ten lessons in delivering digital mental health services. J Clin Med 2019 Aug 17;8(8):1239. [doi: 10.3390/jcm8081239] [Medline: 31426460]

XSL•FO
RenderX

8. Mitchell LM, Joshi U, Patel V, Lu C, Naslund JA. Economic evaluations of internet-based psychological interventions for anxiety disorders and depression: a systematic review. J Affect Disord 2021 Apr 1;284:157-182. [doi: 10.1016/j.jad.2021.01.092] [Medline: 33601245]
9. Ahern E, Kinsella S, Semkovska M. Clinical efficacy and economic evaluation of online cognitive behavioral therapy for major depressive disorder: a systematic review and meta-analysis. Expert Rev Pharmacoecon Outcomes Res 2018 Feb;18(1):25-41. [doi: 10.1080/14737167.2018.1407245] [Medline: 29145746]
10. Hedman E, Ljótsson B, Lindefors N. Cognitive behavior therapy via the internet: a systematic review of applications, clinical efficacy and cost-effectiveness. Expert Rev Pharmacoecon Outcomes Res 2012 Dec;12(6):745-764. [doi: 10.1586/erp.12.67] [Medline: 23252357]
11. Smith AF, Brown GC. Understanding cost effectiveness: a detailed review. Br J Ophthalmol 2000 Jul;84(7):794-798. [doi: 10.1136/bjo.84.7.794] [Medline: 10873998]
12. König H, König HH, Konnopka A. The excess costs of depression: a systematic review and meta-analysis. Epidemiol Psychiatr Sci 2020;29:e30. [doi: 10.1017/S2045796019000180]
13. Dewa CS, Lin E. Chronic physical illness, psychiatric disorder and disability in the workplace. Soc Sci Med 2000 Jul;51(1):41-50. [doi: 10.1016/S0277-9536(99)00431-1]
14. Hendriks SM, Spijker J, Licht CMM, et al. Long-term work disability and absenteeism in anxiety and depressive disorders. J Affect Disord 2015 Jun 1;178:121-130. [doi: 10.1016/j.jad.2015.03.004] [Medline: 25805404]
15. Knudsen AK, Harvey SB, Mykletun A, Øverland S. Common mental disorders and long-term sickness absence in a general working population. The Hordaland Health Study. Acta Psychiatr Scand 2013 Apr;127(4):287-297. [doi: 10.1111/j.1600-0447.2012.01902.x] [Medline: 22775341]
16. Beck A, Crain AL, Solberg LI, et al. Severity of depression and magnitude of productivity loss. Ann Fam Med 2011;9(4):305-311. [doi: 10.1370/afm.1260] [Medline: 21747101]
17. Erickson SR, Guthrie S, Vanetten-Lee M, et al. Severity of anxiety and work-related outcomes of patients with anxiety disorders. Depress Anxiety 2009;26(12):1165-1171. [doi: 10.1002/da.20624] [Medline: 19842165]
18. Walker-Bone K. The Work Productivity and Activity Impairment (WPAI) questionnaire. Occup Med (Lond) 2025;14:kqaf097. [doi: 10.1093/occmed/kqaf097] [Medline: 40971698]
19. Reilly MC, Zbrozek AS, Dukes EM. The validity and reproducibility of a work productivity and activity impairment instrument. Pharmacoeconomics 1993 Nov;4(5):353-365. [doi: 10.2165/00019053-199304050-00006] [Medline: 10146874]
20. de Oliveira C, Saka M, Bone L, Jacobs R. The role of mental health on workplace productivity: a critical review of the literature. Appl Health Econ Health Policy 2023 Mar;21(2):167-193. [doi: 10.1007/s40258-022-00761-w] [Medline: 36376610]
21. Rojanasarot S, Bhattacharyya SK, Edwards N. Productivity loss and productivity loss costs to United States employers due to priority conditions: a systematic review. J Med Econ 2023;26(1):262-270. [doi: 10.1080/13696998.2023.2172282] [Medline: 36695516]
22. Attridge M. Internet-based cognitive-behavioral therapy for employees with anxiety, depression, social phobia, or insomnia: clinical and work outcomes. Sage Open 2020 Jan;10(1):2158244020914398. [doi: 10.1177/2158244020914398]
23. Bondar J, Babich Morrow C, Gueorguieva R, et al. Clinical and financial outcomes associated with a workplace mental health program before and during the COVID-19 pandemic. JAMA Netw Open 2022 Jun 1;5(6):e2216349. [doi: 10.1001/jamanetworkopen.2022.16349] [Medline: 35679044]
24. Ward EJ, Fragala MS, Birse CE, et al. Assessing the impact of a comprehensive mental health program on frontline health service workers. PLOS ONE 2023;18(11):e0294414. [doi: 10.1371/journal.pone.0294414] [Medline: 37988363]
25. Carolan S, Harris PR, Cavanagh K. Improving employee well-being and effectiveness: systematic review and meta-analysis of web-based psychological interventions delivered in the workplace. J Med Internet Res 2017 Jul 26;19(7):e271. [doi: 10.2196/jmir.7583] [Medline: 28747293]
26. Moe-Byrne T, Shepherd J, Merecz-Kot D, et al. Effectiveness of tailored digital health interventions for mental health at the workplace: a systematic review of randomised controlled trials. PLOS Digit Health 2022 Oct;1(10):e0000123. [doi: 10.1371/journal.pdig.0000123] [Medline: 36812547]
27. Persson Asplund R, Asplund S, von Buxhoeveden H, et al. Work-focused versus generic internet-based interventions for employees with stress-related disorders: randomized controlled trial. J Med Internet Res 2023 Apr 25;25:e34446. [doi: 10.2196/34446] [Medline: 37097739]
28. Kechagias EP, Papadopoulos GA, Rokai I. Evaluating the impact of digital health interventions on workplace health outcomes: a systematic review. Adm Sci 2024;14(6):131. [doi: 10.3390/admsci14060131]
29. Forman-Hoffman VL, Flom M, Montgomery R, Robinson A. Improvements in work productivity and activity impairment among adults with anxiety or depressive symptoms participating in a relational agent-delivered digital mental health intervention. J Occup Environ Med 2024 Mar 1;66(3):e99-e105. [doi: 10.1097/JOM.0000000000003038] [Medline: 38242139]
30. Paterson C, Leduc C, Maxwell M, et al. Barriers and facilitators to implementing workplace interventions to promote mental health: qualitative evidence synthesis. Syst Rev 2024 Jun 7;13(1):152. [doi: 10.1186/s13643-024-02569-2] [Medline: 38849924]

XSL•FO
RenderX

31. Evans-Lacko S, Knapp M. Global patterns of workplace productivity for people with depression: absenteeism and presenteeism costs across eight diverse countries. Soc Psychiatry Psychiatr Epidemiol 2016 Nov;51(11):1525-1537. [doi: 10.1007/s00127-016-1278-4] [Medline: 27667656]
32. Harty S, Enrique A, Akkol-Solakoglu S, et al. Implementing digital mental health interventions at scale: one-year evaluation of a national digital CBT service in Ireland. Int J Ment Health Syst 2023 Oct 10;17(1):29. [doi: 10.1186/s13033-023-00592-9] [Medline: 37817270]
33. Richards D, Enrique A, Eilert N, et al. A pragmatic randomized waitlist-controlled effectiveness and cost-effectiveness trial of digital interventions for depression and anxiety. NPJ Digit Med 2020;3(1):85. [doi: 10.1038/s41746-020-0293-8] [Medline: 32566763]
34. Husereau D, Drummond M, Augustovski F, et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations. MDM Policy Pract 2022;7(1). [doi: 10.1177/23814683211061097] [Medline: 35036563]
35. Richards D, Richardson T, Timulak L, McElvaney J. The efficacy of internet-delivered treatment for generalized anxiety disorder: a systematic review and meta-analysis. Internet Interv 2015 Sep;2(3):272-282. [doi: 10.1016/j.invent.2015.07.003]
36. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
37. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
38. Wijesuriya R, Moreno-Betancur M, Carlin JB, White IR, Quartagno M, Lee KJ. Multiple imputation for longitudinal data: a tutorial. Stat Med 2025 Feb 10;44(3-4):e10274. [doi: 10.1002/sim.10274] [Medline: 39846338]
39. Cohen J. Statistical Power Analysis for the Behavioral Sciences: Routledge Academic; 1988.
40. Lofland JH, Pizzi L, Frick KD. A review of health-related workplace productivity loss instruments. Pharmacoeconomics 2004;22(3):165-184. [doi: 10.2165/00019053-200422030-00003] [Medline: 14871164]
41. Structure of earnings survey. Eurostat. URL: https://ec.europa.eu/eurostat/web/microdata/structure-of-earnings-survey [accessed 2026-03-23]
42. US dollar (USD). European Central Bank. URL: https://www.ecb.europa.eu/stats/policy_and_exchange_rates/euro_reference_exchange_rates/html/eurofxref-graph-usd.en.html [accessed 2026-03-30]
43. Finnes A, Enebrink P, Ghaderi A, Dahl J, Nager A, Öst LG. Psychological treatments for return to work in individuals on sickness absence due to common mental disorders or musculoskeletal disorders: a systematic review and meta-analysis of randomized-controlled trials. Int Arch Occup Environ Health 2019 Apr;92(3):273-293. [doi: 10.1007/s00420-018-1380-x] [Medline: 30474734]
44. Joyce S, Modini M, Christensen H, et al. Workplace interventions for common mental disorders: a systematic meta-review. Psychol Med 2016 Mar;46(4):683-697. [doi: 10.1017/S0033291715002408] [Medline: 26620157]
45. Lee YC, Gao L, Dear BF, Titov N, Mihalopoulos C. The Cost-effectiveness of the online MindSpot Clinic for the treatment of depression and anxiety in Australia. J Ment Health Policy Econ 2017 Dec 1;20(4):155-166. [Medline: 29300702]
46. Penev T, Zhao S, Lee JL, Chen CE, Metcalfe L, Ozminkowski RJ. The impact of a workforce mental health program on employer medical plan spend: an application of cost efficiency measurement for mental health care. Popul Health Manag 2023 Feb;26(1):60-71. [doi: 10.1089/pop.2022.0240] [Medline: 36799934]
47. Torous J, Jän Myrick K, Rauseo-Ricupero N, Firth J. Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. JMIR Ment Health 2020 Mar 26;7(3):e18848. [doi: 10.2196/18848] [Medline: 32213476]
48. Webb CA, Rosso IM, Rauch SL. Internet-based cognitive-behavioral therapy for depression: current progress and future directions. Harv Rev Psychiatry 2017;25(3):114-122. [doi: 10.1097/HRP.0000000000000139] [Medline: 28475503]
49. Palacios JE, Enrique A, Mooney O, et al. Durability of treatment effects following internet-delivered cognitive behavioural therapy for depression and anxiety delivered within a routine care setting. Clin Psychol Psychother 2022 Sep;29(5):1768-1777. [doi: 10.1002/cpp.2743] [Medline: 35466486]

Abbreviations

CBT: cognitive behavioral therapy
CHEERS: Consolidated Health Economic Evaluation Reporting Standards
CMD: common mental disorder
GAD-7: Generalized Anxiety Disorder-7
HSE: Health Service Executive
iCBT: internet-based cognitive behavioral therapy
PHQ-9: Patient Health Questionnaire-9
WPAI: Work Productivity and Activity Impairment

XSL•FO
RenderX

Edited by S Brini; submitted 16.Jul.2025; peer-reviewed by A Adepoju, R Singh; accepted 05.Jan.2026; published 07.Apr.2026.

Please cite as:
Lee CT, Harty S, Enrique A, Jiménez-Díaz A, Hisler G, Duffy D, Richards D
Changes in Workplace Productivity and Estimated Cost Savings During Internet-Based Cognitive Behavioral Therapy in the Irish National Health Service: Naturalistic, Repeated-Measures, Retrospective Survey Study
J Med Internet Res 2026;28:e80689
URL: https://www.jmir.org/2026/1/e80689
doi:10.2196/80689

© Chi Tak Lee, Siobhán Harty, Angel Enrique, Alba Jiménez-Díaz, Garrett Hisler, Daniel Duffy, Derek Richards. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Optimization of University Counseling Consent Forms With Large Language Models: Multidimensional Comparative Evaluation

Jianchen Luo[1,2*], MD; Jing Ma[3*], MM; Danni Zhan[4], MEd; Yuhong Zhou[5], PhD; Jiayu Li[6], PhD; Lan Zhang[2,3], MD; Wentao Wang[1,2], MD, PhD

[1]Department of Liver Surgery, West China Hospital of Sichuan University, 37 Guoxue Alley, Wuhou District, Chengdu, China

[2]West China School of Clinical Medicine, Sichuan University, Chengdu, China

[3]Mental Health Center, West China Hospital of Sichuan University, Chengdu, China

[4]Department of Vocational Education, The Open University of Mianyang, Mianyang, China

[5]Psychological Research and Counseling Center, Southwest Jiaotong University, Chengdu, China

[6]Department of Psychology, Hebei Normal University, Shijiazhuang, China

[*]these authors contributed equally

Corresponding Author:
Wentao Wang, MD, PhD
Department of Liver Surgery, West China Hospital of Sichuan University, 37 Guoxue Alley, Wuhou District, Chengdu, China

Abstract

Background: Mental health problems among university students are a growing global concern, yet limited counseling resources and inadequate understanding of counseling procedures often delay timely help-seeking. Informed consent forms (ICFs) are essential for safeguarding autonomy and clarifying counseling procedures, but many universities' counseling ICFs are incomplete, ambiguous, or overly technical. Large language models (LLMs) may offer scalable assistance for improving clarity and accessibility.

Objective: This study aimed to evaluate whether LLM-based rewriting could improve the structure, readability, content quality, and comprehensibility of university counseling ICFs, and compared 2 advanced models (ChatGPT [GPT-5] and Grok-4).

Methods: We conducted a comparative evaluation of counseling ICFs collected from 33 Chinese universities (original texts) and generated 2 rewritten versions for each ICF using ChatGPT (GPT-5) and Grok-4. A multidimensional framework assessed (1) textual structure and readability, (2) expert-rated content quality from a counselor perspective, and (3) volunteer-rated reading comprehension from a client perspective. Comparisons between original and rewritten texts were performed using Wilcoxon signed rank tests, with linear mixed-effects models used to validate results while accounting for rater variability.

Results: Compared with the originals, both LLM-rewritten ICFs showed significant improvements across all evaluated dimensions. The mean Lee-Yang Readability Index decreased from 28.68 (SD 5.69) to 22.39 (SD 2.13) with ChatGPT (GPT-5) and 24.37 (SD 2.32) with Grok-4 (both P<.001), and mean tone friendliness increased from 2.57 (SD 0.29) to 2.67 (SD 0.12) and 2.67 (SD 0.13), respectively. The mean expert-rated content quality improved from 45.33 (SD 8.74) to 52.54 (SD 7.92) and 55.49 (SD 7.81) (P<.001), driven mainly by higher completeness and specificity of key information. The mean volunteer-rated reading comprehension scores increased from 19.02 (SD 1.32) to 22.33 (SD 0.81) and 22.05 (SD 0.90) (P<.001), indicating improved clarity, readability, and acceptability. Across structural features, Grok-4 tended to produce longer rewritten forms than the originals, highlighting a potential trade-off between added informational content and document length.

Conclusions: In this comparative evaluation of 33 Chinese university counseling ICFs, LLM-based rewriting was associated with improved readability, expert-rated content quality, and volunteer-rated comprehension relative to original forms. These findings suggest that LLMs can support the optimization of counseling documentation; however, implementation should consider practical constraints (eg, document length) and retain human oversight.

(J Med Internet Res 2026;28:e86502) doi:10.2196/86502

KEYWORDS

large language models; informed consent forms; university counseling; mental health accessibility; higher education

Introduction

Mental health problems among adolescents and young adults represent a growing global public health concern, particularly in the context of higher education [1]. Estimates suggest that up to 50% of college freshmen worldwide experience at least 1 mental health problem during their time at university [2]. With the rising prevalence of mental health challenges in this

XSL•FO
RenderX

population, psychological counseling services have become an indispensable component of campus mental health systems [3]. As the initial step in delivering such services, informed consent forms (ICFs) not only embody fundamental ethical and legal principles but also serve as a critical safeguard for students' right to information and autonomy. Nevertheless, substantial variability exists in the structure, content, and language of ICFs across universities [4]: some are excessively lengthy and filled with technical jargon, making them difficult for students to understand, while others are incomplete or ambiguously worded, undermining their protective function [5,6].

The clarity and standardization of ICFs are directly linked to recipients' levels of understanding and the quality of their decision-making [5]. Although relevant efforts have been made in the medical and mental health fields, systematic research specifically targeting ICFs for university counseling services remains scarce [7]. At the institutional level in higher education, many universities lack the capacity to continuously refine these forms due to constraints in manpower and resources, as well as the absence of unified standards and practical technical approaches [8,9].

In recent years, large language models (LLMs) have demonstrated notable strengths in text generation and rewriting [10-13]. Their natural language processing capabilities not only enhance readability and structural clarity but also offer universities lacking professional support a low-cost and scalable auxiliary approach to gradually establish more standardized consent form systems [14-16]. However, empirical evidence remains limited regarding whether LLMs may inadvertently omit critical information during rewriting [17], how to balance professionalism with comprehensibility [18], and whether they can genuinely contribute to the standardization of such documents [19].

To address these gaps, we collected counseling consent forms from 33 universities in China and applied 2 state-of-the-art LLMs, ChatGPT (GPT-5) and Grok-4, to optimize and standardize the texts. The evaluation was conducted across three dimensions: textual structure and readability, content quality, and reading comprehension. By comparing original and LLM-optimized versions, this study aims to examine the feasibility and limitations of using LLMs to improve counseling consent forms and to provide insights for strengthening youth mental health services and exploring human-artificial intelligence (AI) collaboration in clinical communication.

Methods

Study Materials and Sources

This study collected 33 psychological counseling ICFs, including 32 from the official websites of Chinese universities and 1 template issued by the Clinical Psychology Registration Work Committee of the Chinese Psychological Society [20]. The template, initially used for pilot testing of LLM rewriting and prompt design, was later included in the final analysis. The sampled universities covered China's eastern, central, western, and northeastern regions, ensuring geographic representativeness, and encompassed national key, provincial key, and local undergraduate institutions, reflecting variation across higher education tiers. To maintain objectivity and validity, all original documents were analyzed without modification, other than the removal of university identifiers. A full list of universities is provided in Table S1 (see Multimedia Appendix 1).

Model Selection and Text Generation

Two representative LLMs were examined: ChatGPT (GPT-5) (OpenAI; released August 7, 2025) and Grok-4 (xAI; released July 9, 2025) [21,22]. Both are recently released, widely adopted, and high-performing closed-source systems with strong capabilities in Chinese language processing and text generation [23,24]. All model outputs were generated using the publicly available versions of these models accessible during a defined time window (August 21-September 8, 2025), and the reported results reflect model behavior at the time of data generation. Prior to formal experimentation, we developed and tested a structured prompt defining requirements for contextual framing, preservation and supplementation of key information, linguistic style, and textual organization (Supplementary Information S1 in Multimedia Appendix 2). The structured prompts were natively drafted in Chinese to align with the language of the source documents, and no cross-linguistic translation was involved. The template ICF was used for preliminary trials to verify that the models generated content that was both complete and structurally consistent. Following this procedure, each original text was optimized independently by ChatGPT (GPT-5) and Grok-4, yielding three versions—Original, ChatGPT (GPT-5), and Grok-4—for subsequent comparison and evaluation (all documents are provided in Multimedia Appendix 3). When generating optimized ICFs, both ChatGPT (GPT-5) and Grok-4 were prompted using the same instruction template to ensure fair comparison. The prompts required adherence to both international ethical guidelines and domestic regulations relevant to counseling practice, with Chinese laws prioritized in cases of discrepancy. Retrieval-augmented generation or external document retrieval was not used; all outputs were generated through the models' standard web-based interfaces.

Textual Structure and Readability Analysis

To evaluate differences among the 3 ICF versions in structural complexity and linguistic expression, 8 metrics were applied. Five basic structural features were measured: character count, word count, sentence count, words per sentence, and characters per word [4,6,25,26]. Syntactic complexity was captured by the nested sentence ratio [27], reading difficulty by the Lee-Yang Readability Index [28], and tone friendliness was added to reflect approachability and acceptability [29]. All features were derived through a standardized processing pipeline.

The Lee-Yang Readability Index is a weighted composite of average sentence length, word length, and lexical complexity, widely used to assess the readability of Chinese texts [30], especially in science communication, education, and health contexts [28,31]. Tone friendliness was operationalized using a lexical counting approach. Two predefined dictionaries—politeness- or empathy-related words and harsh words—were applied (Supplementary Information S2 in Multimedia Appendix 2) [32]. For each text, the normalized

XSL•FO
RenderX

frequency difference between the 2 word categories (adjusted by sentence count) was linearly mapped from −1 to 1 onto a 0 to 5 scale, with values truncated if outside this range. This metric was intended as a surface-level lexical proxy of linguistic tone rather than a contextual or pragmatic assessment of emotional friendliness. Higher scores indicate a greater relative use of polite or empathic wording and fewer harsh or legalistic terms. The method is interpretable, reproducible, and adaptable through dictionary refinement [33].

Content Quality Assessment

To assess the completeness and standardization of ICFs, we drew on major international and national ethical guidelines. These included the American Psychological Association Ethical Principles and Code of Conduct (2016) [34], the British Psychological Society Code of Ethics and Conduct (2018) [35], the Code of Ethics for Clinical and Counseling Psychology (2018) [36], and the World Health Organization Comprehensive Mental Health Action Plan 2013 - 2030 (2021) [37]. Together, these documents provided the ethical and methodological foundation for the content quality evaluation framework.

Building on this foundation, we established an evaluation framework of 20 core indicators widely considered essential for ICFs: confidentiality; exceptions to confidentiality; client rights; guardian consent; goals and scope; format and frequency; fees and cancelation policy; recording methods; authorization and revocation; crisis procedures; complaints and appeals; data protection; disclaimer of boundaries; language clarity; voluntariness; client obligations; counseling limitations; counselor qualifications; counseling modalities; target population (detailed descriptions and rationales in Multimedia Appendix 1, Table S2). Each indicator was assessed across three dimensions: existence (0 - 1), defined as whether the element was explicitly presented as an independent item (eg, a distinct heading or clearly delineated paragraph); specificity (0 - 2), reflecting whether concrete details were provided; and operability (0 - 2), indicating whether actionable procedures were specified (detailed in Multimedia Appendix 1**,** Table S3). When a relevant indicator was mentioned within other sections without being explicitly listed as an independent item, existence was scored as 0, while specificity and operability were still assessed based on the content provided. Each indicator had a maximum of 5 points, for a total of 100 across all 20 indicators, with subscores of 20 for existence, 40 for specificity, and 40 for operability.

Prior to formal application, 3 senior experts in psychology and ethics reviewed and revised the proposed framework to ensure scientific rigor, comprehensiveness, and operability. A pilot test using the template ICF was then conducted to examine applicability and discriminative capacity. Once feasibility was confirmed, the finalized framework was applied to the remaining 32 ICFs. Scoring was performed independently by 5 experts with backgrounds in psychology or mental health. To maximize fairness and consistency, all texts were randomized and anonymized so that raters were blinded to both the source and version.

Reading Comprehension Assessment

To assess differences in comprehensibility among ICF versions, 10 university volunteers from nonpsychology and nonhealth-related majors participated as readers simulating clients; none had received formal training in counseling or mental health disciplines. They evaluated the texts using a 6D framework: comprehensibility—ease of reading and understanding [8,38]; clarity—whether key information was clearly and unambiguously expressed [39,40]; trustworthiness—perceived reliability of content [41,42]; friendliness—supportive, respectful, and approachable tone [32]; professionalism—appropriate formality and expertise [43]; and acceptability—whether content and style were acceptable to clients [44]. Each dimension was rated on a 5-point Likert scale (1=strongly disagree, 5=strongly agree) [45,46]. The maximum per-dimension score was 5, yielding a total score of 6 - 30. A seventh composite dimension ("dimension means"), defined as the average of the six dimensions (range 1–5), was calculated. Prior to the formal study, a pilot test with the template ICF was conducted to examine the scale's applicability and discriminative capacity. During formal rating, all texts were randomized and anonymized to blind raters to both source and version, ensuring objectivity.

Statistical Analysis

Differences among the 3 ICF versions were examined across textual structure and readability, content quality, and reading comprehension. For textual structure and readability, values for 8 indicators were calculated for each document. For content quality and reading comprehension, ratings from multiple evaluators were averaged per document and then summarized by version. In all cases, means and SDs were reported.

All between-version comparisons were conducted using the Wilcoxon signed rank test. Effect sizes (r) were reported, and P-values were adjusted for multiple testing using the false discovery rate approach (Benjamini-Hochberg procedure). Interrater reliability for total scores of content quality and reading comprehension was assessed using intraclass correlation coefficients (ICCs), based on ratings from 5 experts and 10 readers, respectively.

In addition, to account for individual differences among raters and documents, linear mixed-effects models (LMMs) were fitted with "version" as a fixed effect and "document ID" and "rater" as random effects, allowing robust estimation of version effects under a repeated-measures design. Because content quality and reading comprehension were evaluated by different groups, separate models were constructed for each outcome. Estimation used the MixedLM module. All tests were 2-tailed with significance set at $P<.05$. P values from the LMMs were not adjusted for multiple comparisons; therefore, the possibility of type I error due to multiple testing cannot be excluded. Analyses were performed in Python 3.11 using pandas (v2.2.2; pandas development team), NumPy (v1.26.4; NumPy developers), SciPy (v1.11.4; SciPy developers), jieba (v0.42.1; Jieba development team), statsmodels (v0.14.1; statsmodels developers), scikit-learn (v1.4.0; scikit-learn developers), and Pingouin (v0.5.4; Pingouin developers). All codes are available in Multimedia Appendix 4 for reproducibility.

XSL•FO
RenderX

Ethical Considerations

This study involved human participants and was reviewed and approved by the Ethics Committee of West China Hospital, Sichuan University (Approval 2024 - 1431). The study was classified as minimal risk and conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants prior to participation, including expert raters and student volunteers who evaluated deidentified university counseling ICFs; no clinical intervention or therapeutic interaction was involved. To protect privacy and confidentiality, all documents were anonymized before analysis, with institutional identifiers removed, and no personally identifiable information was collected or analyzed. Participants received no financial or material compensation, and the manuscript and multimedia appendices do not contain any images or content that could enable identification of individual participants.

Results

Textual Structure and Readability

Compared with the original texts, ChatGPT (GPT-5)–optimized ICFs showed more sentences (mean 51.30, SD 12.27 vs mean 36.79, SD 13.29; *P*<.001), fewer words per sentence (mean 20.23, SD 2.29 vs mean 28.05, SD 6.45; *P*<.001), more characters per word (mean 1.16, SD 0.04 vs mean 1.12, SD 0.03; *P*<.001), and a lower nested sentence ratio (mean 0.21, SD 0.06 vs mean 0.32, SD 0.14; *P*<.001). Character count (mean 1206.42, SD 353.00 vs mean 1117.61, SD 376.58; *P*=.17) and word count (mean 1039.33, SD 297.52 vs mean 996.24, SD 338.14; *P*=.44) did not differ significantly. The Lee-Yang Readability Index was lower (mean 22.39, SD 2.13 vs mean 28.68, SD 5.69; *P*<.001), while tone friendliness was higher (mean 2.67, SD 0.12 vs mean 2.57, SD 0.29; *P*=.04).

Grok-4–optimized ICFs were significantly longer, with higher character counts (mean 1669.52, SD 234.09; *P*<.001) and word counts (mean 1393.21, SD 208.82; *P*<.001). Sentence count also increased (mean 65.33, SD 10.15; *P*<.001), words per sentence decreased (mean 21.49, SD 2.56; *P*<.001), characters per word increased (mean 1.20, SD 0.04; *P*<.001), and the nested sentence ratio decreased (mean 0.23, SD 0.06; *P*<.001). The Lee-Yang Index also decreased (mean 24.37, SD 2.32; *P*<.001), and tone friendliness improved (mean 2.67, SD 0.13; *P*=.02).

Direct comparisons showed that Grok-4 had higher character counts, word counts, sentence counts, and characters per word (*P*<.001). Differences in words per sentence were modest (*P*=.02), nested sentence ratios did not differ (*P*=.09), and Grok-4 had a higher Lee-Yang Readability Index (*P*<.001). Tone friendliness was comparable (*P*=.76). Full results are shown in Table 1 and Figure 1. Detailed statistical analysis results are provided in Table S4 (Multimedia Appendix 1).

XSL•FO
RenderX

Table . Comparative analysis of multidimensional evaluation results across all versions.

Indicator	Original, mean (SD)	ChatGPT-5, mean (SD)	Grok-4, mean (SD)	ChatGPT-5 versus original (*P* value)[a]	Grok-4 versus original (*P* value)	Grok-4 versus ChatGPT-5 (*P* value)
Text structure						
Character count	1117.61 (376.58)	1206.42 (353.00)	1669.52 (234.09)	.17	<.001	<.001
Word count	996.24 (338.14)	1039.33 (297.52)	1393.21 (208.82)	.44	<.001	<.001
Sentence count	36.79 (13.29)	51.30 (12.27)	65.33 (10.15)	<.001	<.001	<.001
Words per sentence	28.05 (6.45)	20.23 (2.29)	21.49 (2.56)	<.001	<.001	.02
Characters per word	1.12 (0.03)	1.16 (0.04)	1.20 (0.04)	<.001	<.001	<.001
Nested sentence ratio	0.32 (0.14)	0.21 (0.06)	0.23 (0.06)	<.001	<.001	.09
Readability						
Lee-Yang Readability Index	28.68 (5.69)	22.39 (2.13)	24.37 (2.32)	<.001	<.001	<.001
Tone friendliness	2.57 (0.29)	2.67 (0.12)	2.67 (0.13)	.04	.02	.76
Content quality						
Existence	11.03 (2.18)	12.72 (1.26)	13.55 (1.24)	<.001	<.001	.004
Specificity	11.55 (3.08)	16.51 (3.38)	18.25 (2.38)	<.001	<.001	.02
Operability	22.75 (5.73)	23.31 (6.12)	23.69 (6.46)	.31	.10	.60
Total	45.33 (8.74)	52.54 (7.92)	55.49 (7.81)	<.001	<.001	.01
Reading comprehension						
Comprehensibility	2.93 (0.45)	3.76 (0.30)	3.69 (0.35)	<.001	<.001	.19
Clarity	2.98 (0.47)	3.78 (0.28)	3.74 (0.28)	<.001	<.001	.32
Trustworthiness	3.30 (0.30)	3.72 (0.26)	3.61 (0.32)	<.001	<.001	.12
Friendliness	3.35 (0.35)	3.75 (0.32)	3.73 (0.27)	<.001	<.001	.79
Professionalism	3.25 (0.25)	3.55 (0.32)	3.64 (0.30)	<.001	<.001	.21
Acceptability	3.20 (0.40)	3.77 (0.37)	3.65 (0.33)	<.001	<.001	.06
Total[b]	19.02 (1.32)	22.33 (0.81)	22.05 (0.90)	<.001	<.001	.11
Dimension means (total[c]/6)	3.17 (0.22)	3.72 (0.13)	3.68 (0.15)	<.001	<.001	.11

[a]*P* values were adjusted for multiple comparisons using the false discovery rate.

[b]Total score of the 3 dimensions in content quality evaluation.

[c]Total score of the 6 dimensions in reading comprehension evaluation.

XSL•FO
RenderX

Figure 1. Text structure and readability indicators across original and large language model–optimized university counseling informed consent forms. Box plots compare 8 structural and readability indicators across original and ChatGPT-5– and Grok-4–rewritten counseling informed consent forms (n=33). Dots represent individual documents. Statistical comparisons were performed using Wilcoxon signed rank tests. Significance thresholds: $P<.05$ (*), $P<.01$ (**), $P<.001$ (***), not significant (ns). (A) Character count, (B) word count, (C) sentence count, (D) words per sentence, (E) characters per word, (F) nested sentence ratio, (G) Lee-Yang Readability Index, and (H) tone friendliness.

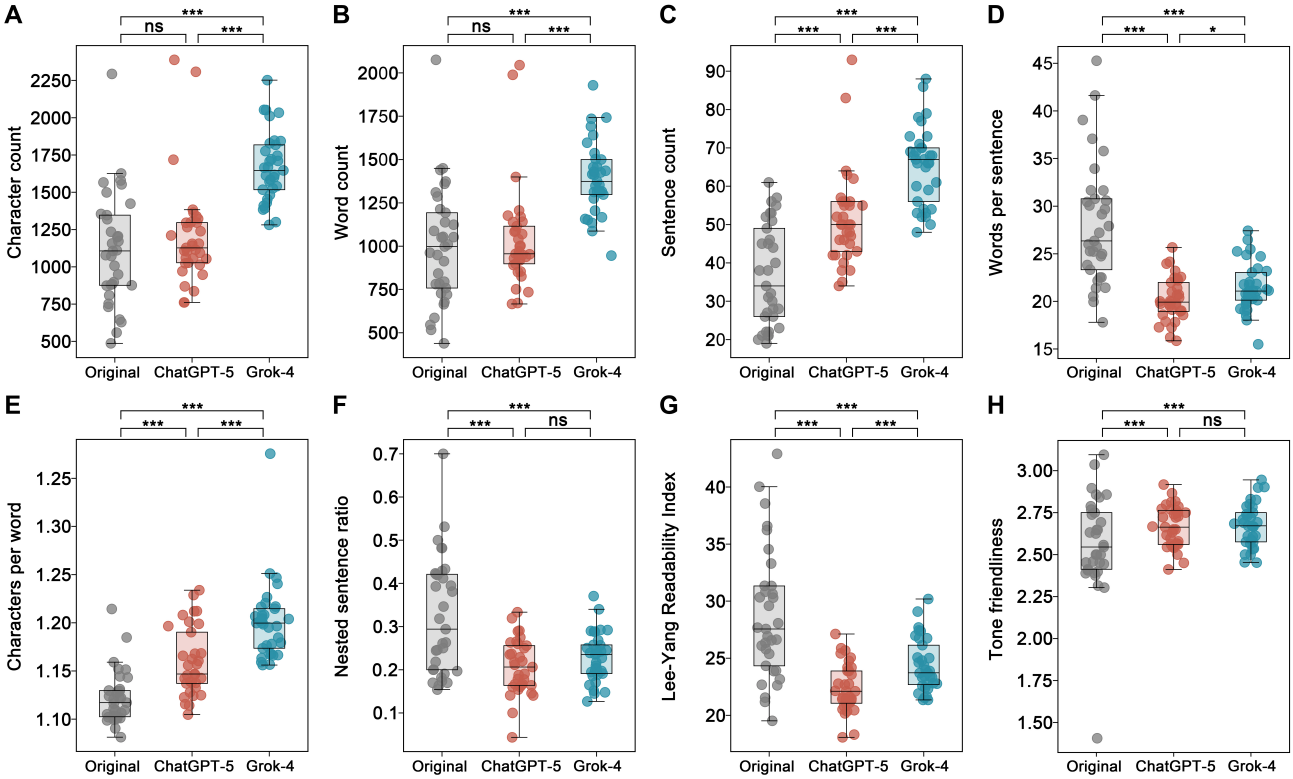

Content Quality

Interrater reliability among 5 expert raters across all documents was high (ICC [2,k]=0.83; Table S5 in Multimedia Appendix 1). Compared with the original texts, ChatGPT (GPT-5)–optimized ICFs showed higher existence (mean 12.72, SD 1.26 vs mean 11.03, SD 2.18; $P<.001$) and specificity scores (mean 16.51, SD 3.38 vs mean 11.55, SD 3.08; $P<.001$), while operability did not differ (mean 23.31, SD 6.12 vs mean 22.75, SD 5.73; $P=.31$). The total score increased significantly (mean 52.54, SD 7.92 vs mean 45.33, SD 8.74; $P<.001$). Grok-4 also showed marked gains, with higher existence (mean 13.55, SD 1.24; $P<.001$) and specificity (mean 18.25, SD 2.38; $P<.001$). Operability increased (mean 23.69, SD 6.46) but was not significant ($P=.09$). The total score rose to a mean of 55.49 (SD 7.81; $P<.001$). Direct comparisons showed Grok-4 scored higher than ChatGPT (GPT-5) in existence ($P=.004$) and specificity ($P=.02$), with no difference in operability ($P=.60$). The overall score was also higher for Grok-4 ($P=.01$) (Table 1). Detailed statistical analysis results are provided in Multimedia Appendix 1, Table S6.

To illustrate performance across the 20 core indicators, a heatmap was generated (Figure 2). The analyses showed that several indicators—confidentiality, exceptions to confidentiality, goals and scope, authorization and revocation, complaints and appeals, disclaimer of boundaries, and counseling modalities—improved significantly under both LLM-optimized versions. These gains were mainly in specificity, with few improvements in existence or operability. Between the 2 LLMs, significant differences appeared only in client rights and language clarity; all other indicators showed no differences. Full statistical results are provided in Tables S7-S11, and rater-level scores in Supplementary Data S1-S5 (Multimedia Appendix 5).

XSL•FO
RenderX

Figure 2. Expert-rated content quality profiles of original and large language models–optimized university counseling informed consent forms. Heatmaps display document-level content quality scores for 33 counseling informed consent forms, evaluated across 20 predefined indicators grouped into three dimensions (existence, specificity, and operability), as well as the total score. Darker colors indicate higher scores. (A) Grok-4–rewritten versions, (B) ChatGPT (GPT-5)–rewritten versions, and (C) original documents. Dimension-level summaries include (i) total score, (ii) existence score, (iii) specificity score, and (iv) operability score.

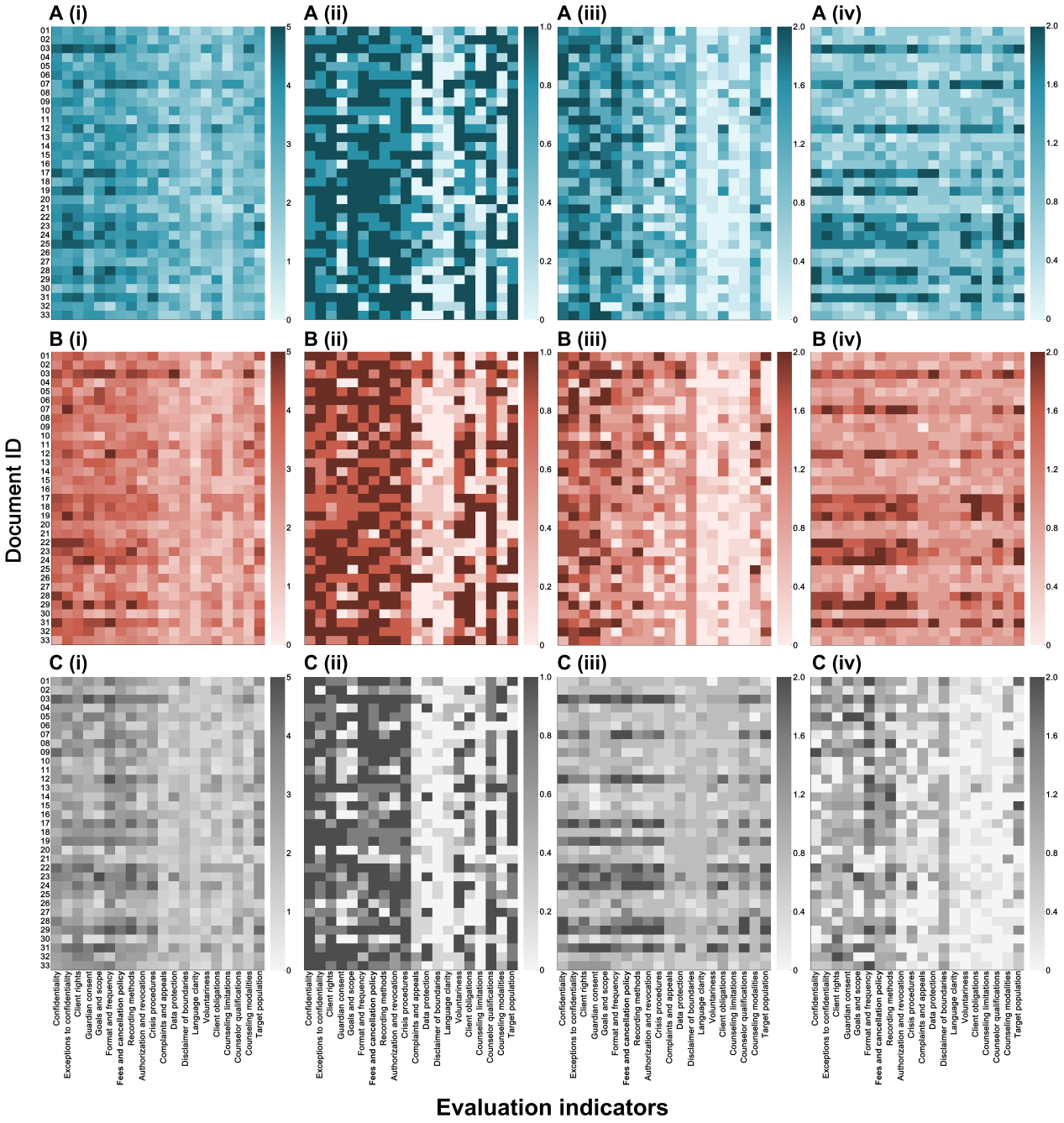

Reading Comprehension

Interrater reliability among 10 readers across all documents was fair (ICC [2,k]=0.59; Table S12 in Multimedia Appendix 1). LLM-optimized ICFs outperformed the originals across all dimensions and in total scores (Table 1). For ChatGPT (GPT-5), scores improved in comprehensibility (mean 3.76, SD 0.30 vs mean 2.93, SD 0.45; *P*<.001), clarity (mean 3.78, SD 0.28 vs mean 2.98, SD 0.47; *P*<.001), trustworthiness (mean 3.72, SD 0.26 vs mean 3.30, SD 0.30; *P*<.001), friendliness (mean 3.75, SD 0.32 vs mean 3.35, SD 0.35; *P*<.001), professionalism (mean 3.55, SD 0.32 vs mean 3.25, SD 0.25; *P*<.001), and acceptability (mean 3.77, SD 0.37 vs mean 3.20, SD 0.40; *P*<.001). The total score also rose (mean 22.33, SD 0.81 vs mean 19.02, SD 1.32; *P*<.001).

Grok-4 showed similar mean gains: comprehensibility 3.69 (SD 0.35; *P*<.001), clarity 3.74 (SD 0.28; *P*<.001), trustworthiness 3.61 (SD 0.32; *P*<.001), friendliness 3.73 (SD 0.27; *P*<.001), professionalism 3.64 (SD 0.30; *P*<.001), and acceptability 3.65 (SD 0.33; *P*<.001), with a total of 22.05 (SD 0.90; *P*<.001).

In direct comparison, ChatGPT (GPT-5) had higher means across most dimensions and the total score, but none reached significance (comprehensibility: *P*=.19; clarity: *P*=.32;

XSL•FO
RenderX

trustworthiness: *P*=.12; friendliness: *P*=.79; professionalism: *P*=.21; and acceptability: *P*=.06). For visualization, dimension means were plotted instead of totals (Figure 3). Detailed statistical analysis results are provided in Table S13 in Multimedia Appendix 1. Complete dimension-level ratings are in Supplementary Data S6-S15 in Multimedia Appendix 5.

Figure 3. Volunteer-rated reading comprehension of original and large language model–optimized university counseling informed consent forms.

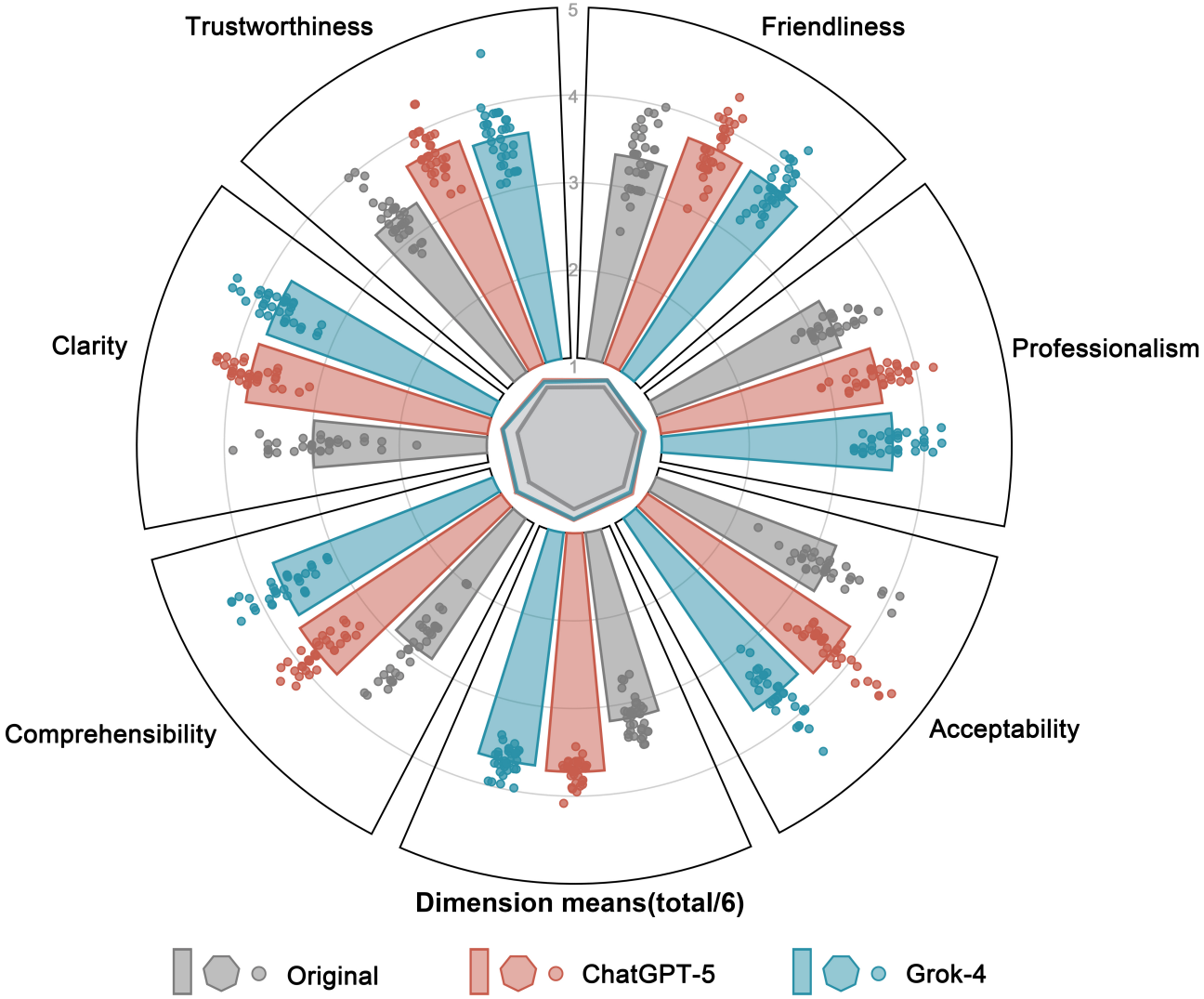

The figure compares reading comprehension scores across original, ChatGPT (GPT-5)–rewritten, and Grok-4–rewritten counseling ICFs (n=33). Mean scores across 6 comprehension dimensions and the total score are shown, with dots indicating document-level distributions. A central radar plot summarizes dimension-level means for comparison.

Linear Mixed-Effects Model

To account for individual differences among raters and texts, LMMs with crossed random effects were constructed, specifying "version" as a fixed effect and including random intercepts for rater and document ID. Separate models were fitted for content quality and reading comprehension (Figure 4). In the content quality model, ChatGPT (GPT-5) yielded higher total scores than the original texts (β=7.20, *P*<.001), with significant gains in existence (β=1.62; *P*<.001) and specificity (β=5.22; *P*<.001), but not operability (β=.36; *P*=.17). Grok-4 also showed significant improvements over the originals in total scores (β=10.15; *P*<.001), existence (β=2.45; *P*<.001), specificity (β=6.96; *P*<.001), and operability (β=.74; *P*=.02). Direct comparison showed Grok-4 scored higher than ChatGPT (GPT-5) in total scores (β=2.95; *P*<.001), existence (β=0.82; *P*<.001), and specificity (β=1.75; *P*<.001), with no difference in operability (β=.38; *P*=.27).

XSL•FO
RenderX

Figure 4. Linear mixed-effects model results for (A) content quality and (B) reading comprehension across original and large language model–optimized university counseling informed consent forms.

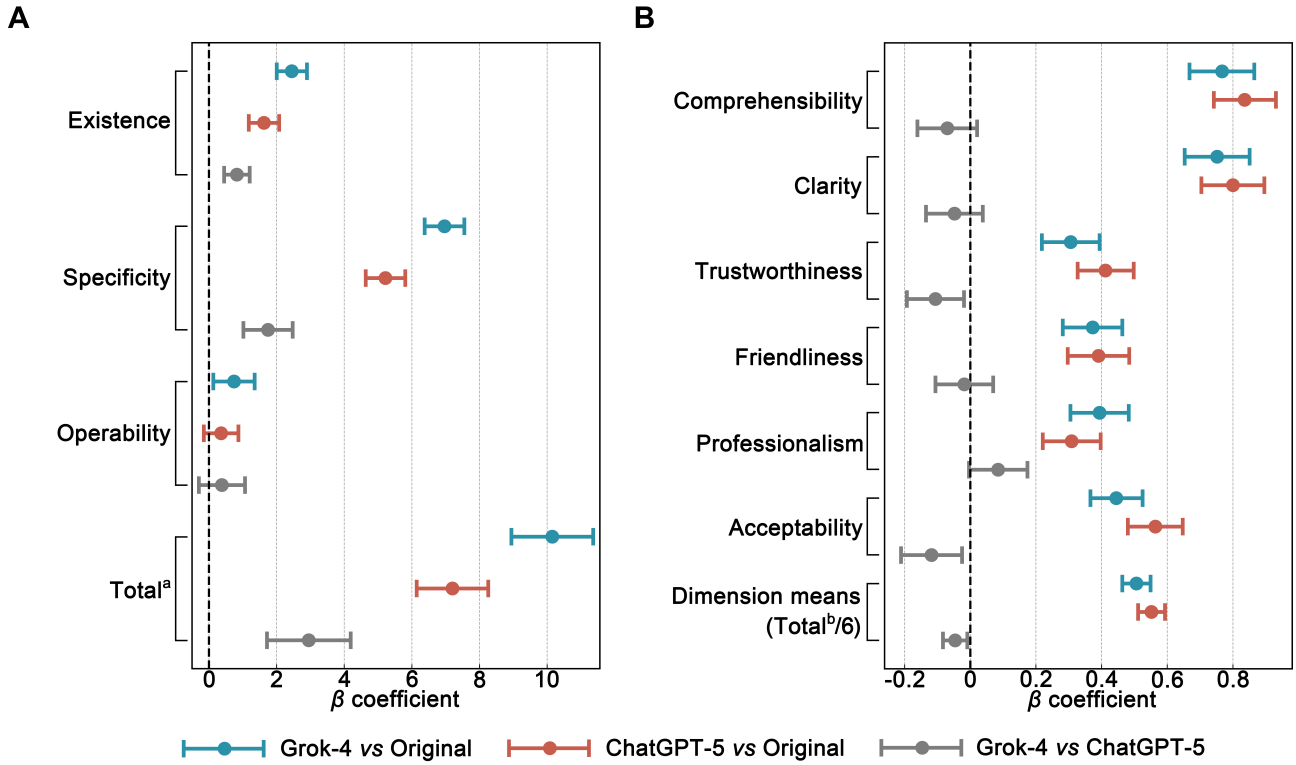

The figure shows LMM estimates comparing original, ChatGPT (GPT-5)–rewritten, and Grok-4–rewritten counseling ICFs (n=33). Models were fitted separately for content quality and reading comprehension outcomes. Positive β coefficients indicate higher scores for the first group in each comparison. The vertical dashed line at β=0 represents no difference between groups. CIs that do not cross the zero line indicate statistically significant differences at $P<.05$.

In the reading comprehension model, ChatGPT (GPT-5) showed significant improvements over the original texts in comprehensibility (β=.84; $P<.001$), clarity (β=.80; $P<.001$), trustworthiness (β=.41; $P<.001$), friendliness (β=.39; $P<.001$), professionalism (β=.31; $P<.001$), acceptability (β=.56; $P<.001$), and dimension means (β=.55; $P<.001$). Grok-4 likewise outperformed the originals, with gains in comprehensibility (β=.77; $P<.001$), clarity (β=.75; $P<.001$), trustworthiness (β=.31; $P<.001$), friendliness (β=.37; $P<.001$), professionalism (β=.39; $P<.001$), acceptability (β=.45; $P<.001$), and dimension means (β=.51; $P<.001$). In direct comparison, Grok-4 scored lower than ChatGPT (GPT-5) in trustworthiness (β=−.11; $P=.02$), acceptability (β=−.12; $P=.01$), and dimension means (β=−.05; $P=.01$). Details are reported in Table 2, with random effects and model fit indices in Multimedia Appendix 1, Tables S14 and S15.

XSL•FO
RenderX

Table . Results of linear mixed-effects model analysis on content quality and reading comprehension.

Dimension and comparison	β (95% CI)	SE	*P* value[a]
Content quality			
Existence			
Grok-4 vs original	2.45 (2.00 to 2.90)	0.23	<.001
ChatGPT-5 vs original	1.62 (1.18 to 2.07)	0.23	<.001
Grok-4 vs ChatGPT-5	0.82 (0.45 to 1.20)	0.19	<.001
Specificity			
Grok-4 vs original	6.96 (6.38 to 7.55)	0.30	<.001
ChatGPT-5 vs original	5.22 (4.63 to 5.80)	0.30	<.001
Grok-4 vs ChatGPT-5	1.74 (1.02 to 2.47)	0.37	<.001
Operability			
Grok-4 vs original	0.74 (0.13 to 1.35)	0.31	.02
ChatGPT-5 vs original	0.36 (–0.15 to 0.87)	0.26	.17
Grok-4 vs ChatGPT-5	0.38 (–0.30 to 1.06)	0.35	.27
Total[b]			
Grok-4 vs original	10.15 (8.94 to 11.36)	0.62	<.001
ChatGPT-5 vs original	7.20 (6.14 to 8.26)	0.54	<.001
Grok-4 vs ChatGPT-5	2.95 (1.71 to 4.19)	0.63	<.001
Reading comprehension			
Comprehensibility			
Grok-4 vs original	0.77 (0.67 to 0.86)	0.05	<.001
ChatGPT-5 vs original	0.84 (0.74 to 0.93)	0.05	<.001
Grok-4 vs ChatGPT-5	–0.07 (–0.16 to 0.02)	0.05	.13
Clarity			
Grok-4 vs original	0.75 (0.65 to 0.85)	0.05	<.001
ChatGPT-5 vs original	0.80 (0.70 to 0.90)	0.05	<.001
Grok-4 vs ChatGPT-5	–0.05 (–0.14 to 0.04)	0.04	.27
Trustworthiness			
Grok-4 vs original	0.31 (0.22 to 0.39)	0.05	<.001
ChatGPT-5 vs original	0.41 (0.33 to 0.50)	0.04	<.001
Grok-4 vs ChatGPT-5	–0.11 (–0.19 to 0.02)	0.04	.02
Friendliness			
Grok-4 vs original	0.37 (0.28 to 0.46)	0.05	<.001
ChatGPT-5 vs original	0.39 (0.30 to 0.48)	0.05	<.001
Grok-4 vs ChatGPT-5	–0.02 (–0.11 to 0.07)	0.05	.69
Professionalism			
Grok-4 vs original	0.39 (0.30 to 0.48)	0.05	<.001
ChatGPT-5 vs original	0.31 (0.22 to 0.40)	0.05	<.001
Grok-4 vs ChatGPT-5	0.08 (0.00 to 0.17)	0.05	.06
Acceptability			
Grok-4 vs original	0.44 (0.37 to 0.53)	0.04	<.001
ChatGPT-5 vs original	0.56 (0.48 to 0.65)	0.04	<.001

XSL•FO
RenderX

Dimension and comparison	β (95% CI)	SE	*P* value[a]
Grok-4 vs ChatGPT-5	–0.12 (–0.21 to 0.03)	0.05	.01
Total[c]			
Grok-4 vs original	3.04 (2.78 to 3.30)	0.13	<.001
ChatGPT-5 vs original	3.31 (3.07 to 3.56)	0.13	<.001
Grok-4 vs ChatGPT-5	–0.28 (–0.50 to 0.06)	0.11	.01
Dimension means (Total[c]/6)			
Grok-4 vs original	0.51 (0.46 to 0.55)	0.02	<.001
ChatGPT-5 vs original	0.55 (0.51 to 0.59)	0.02	<.001
Grok-4 vs ChatGPT-5	–0.05 (–0.08 to 0.01)	0.02	.01

[a]*P* values from the linear mixed-effects models were not adjusted for multiple comparisons.

[b]Total score of the 3 dimensions in content quality evaluation.

[c]Total score of the 6 dimensions in reading comprehension evaluation.

Discussion

Principal Results

This study systematically evaluated the performance of LLMs in rewriting university counseling ICFs. It focused on 2 state-of-the-art closed-source models, ChatGPT (GPT-5) and Grok-4. Both achieved significant improvements over the originals in textual structure and readability, content quality, and reading comprehension. LLM-based rewriting reduced linguistic complexity, enhanced the completeness and specificity of key information, and improved clarity and acceptability, making ICFs easier to understand and apply. Such improvements may help lower barriers to accessing counseling, thereby supporting more timely use of mental health services among young people. LMMs confirmed these advantages across multiple dimensions. This study is also the first to design and apply a multidimensional evaluation framework for university counseling ICFs. The framework deliberately distinguished evaluation perspectives: content quality was assessed by psychology or mental health experts from a counselor's standpoint, while reading comprehension was evaluated by university volunteers simulating clients. This dual-perspective design enabled the assessment of both professional and ethical adequacy and actual acceptability and comprehensibility. Together, the findings underscore the value of LLMs in optimizing counseling ICFs and offer a methodological reference for future research.

Comparison With Prior Work

In terms of textual structure and readability, LLM-based rewriting substantially altered the linguistic features of ICFs. ChatGPT (GPT-5)–generated texts showed shorter words per sentence and lower nested sentence ratios, indicating improved clarity and comprehensibility through sentence splitting and syntactic simplification [47]. The increase in characters per word likely reflects the use of more standardized, professional vocabulary to ensure accuracy while simplifying syntax [4]. The lower Lee-Yang Readability Index and higher tone friendliness suggest reduced overall difficulty with a more approachable style [28]. It should be noted that tone friendliness, as operationalized here, reflects surface-level lexical tendencies and may not fully capture context-dependent or pragmatically appropriate expressions; for example, legally protective wording may serve a positive ethical function despite being classified as strict. Although statistically significant, the absolute increase was small (approximately 0.1 on a 5-point scale), and the linear mapping approach may amplify minor lexical frequency differences. These findings should therefore be interpreted as indicative of modest stylistic shifts rather than substantive changes in empathetic communication. In addition, sentence counts rose significantly, with modest increases in character and word counts, likely reflecting supplementation of missing content [26]. Grok-4 followed a similar strategy but produced longer texts with larger increases in length and sentence count, indicating more active supplementation of absent material, while still improving readability and tone friendliness. However, increased length may also impose practical burdens in counseling settings, highlighting a trade-off between informational completeness and usability. The lack of significant differences between the models in nested sentence ratio and tone friendliness suggests convergence in syntactic and stylistic adjustments [48]. These findings align with prior research. Ali et al [49] reported that GPT-4 simplification of surgical consent forms lowered reading difficulty and enhanced readability, while Beattie et al [50] found LLM-optimized trial consent forms easier to understand. Mirza et al [26] further showed that US trial consent forms are written at overly high reading levels, with each grade-level increase linked to ~16% participant attrition, underscoring the value of simplification. The present study confirms that LLM-based rewriting reduces syntactic complexity and improves readability, with ChatGPT (GPT-5) emphasizing clarity through conciseness and Grok-4 expanding coverage through longer texts, highlighting differences in generation strategies that may shape readers' experiences.

The analysis of content quality indicates that LLM-based rewriting effectively supplemented key information and enhanced the concreteness of items, particularly in the specificity dimension. However, improvements in operability were relatively limited, suggesting that models still face challenges in generating explicit execution pathways and actionable

XSL•FO
RenderX

requirements. Conceptually, operability requires specifying not only what should be done, but also who is responsible, when actions occur, and under what conditions they apply. Such institution-specific procedural details are difficult for LLMs to generate reliably without access to local workflows. Consequently, operability may be less suitable for fully automated optimization and better addressed through human-AI collaboration, with LLMs supporting structural drafting and clinicians refining actionable details. This limited improvement may also reflect the absence of institution-specific "ground truth" information in the source texts and prompts. Because no retrieval-augmented generation or context injection was used, the models lacked access to concrete local procedural details (eg, office hours or emergency contacts), which inherently constrained operability-related gains. Grok-4 showed a more proactive optimization pattern. Taken together, these perspectives were further supported by the findings of the LMM analyses. Previous studies have similarly highlighted the potential of LLMs to improve the quality of ICFs and health-related documents. Shi et al [51] confirmed that LLMs significantly enhanced completeness and specificity in clinical trial consent forms. Oh et al [52] showed that LLMs simplified surgical consent forms while preserving expert-endorsed quality. Rust et al [53] and Romoff et al [54] reported in cardiology discharge summaries and orthopedic educational materials that LLM-based rewriting not only improved readability but also consistently received positive evaluations in content quality. These findings align with the present results, suggesting that LLMs have stable advantages in reinforcing information coverage and refining provisions, while limitations in operability-oriented expressions warrant further investigation.

With regard to practical implementation, operability emerged as a particularly challenging dimension. Indicator-level analyses further revealed that LLM optimization was concentrated on information presentation and provision refinement, with vague or missing clauses supplemented and clarified after rewriting. By contrast, operability and implementation pathways showed only modest gains. While this can be regarded as a limitation at the rewriting level, it also reflects the models' cautious tendency to avoid generating potentially erroneous or misleading operational details in high-risk contexts, thereby reducing the risk of hallucinations. This trend echoes prior research: Alber et al [55] warned of data poisoning risks in medical corpora, underscoring the need for prudence, while Sebastian et al [56] systematically documented hallucinations in LLM-generated medical texts and emphasized risk evaluation mechanisms. Collectively, these observations suggest that the latest generation of models may exhibit enhanced stability in sensitive contexts. Moreover, this pattern leaves scope for improvement through prompt engineering or human-AI collaboration to guide models toward generating more detailed and personalized operational provisions.

It is noteworthy that cases were observed in which existence was scored as 0, while specificity or operability received nonzero scores. This pattern reflects a distinction between structural explicitness and semantic presence incorporated in the evaluation framework. In the present study, the existence dimension was intentionally defined to capture whether information was explicitly listed as an independent item, rather than whether related content appeared anywhere in the document. This design choice was motivated by the intended audience of counseling ICFs, which primarily consists of nonprofessional readers who rely on clear, structured, and itemized presentation to accurately identify key information. Accordingly, information that was mentioned within other sections without explicit listing could still be evaluated for its specificity and operability once identified, but such nonexplicit presentation may reduce retrievability and comprehensibility, as relevant details—even if included—may not be readily recognized without clear structural cues [57]. Although LLM-based optimization improved overall performance in existence and specificity, fully explicit and well-structured presentation was not consistently achieved across all items [49]. These findings underscore the importance of structural clarity in counseling ICFs and suggest that further improvements may benefit from the integration of professional human review and standardized templates to enhance explicitness and ease of information retrieval [26].

With respect to reading comprehension, LLM-based rewriting significantly outperformed the original texts across all 6 dimensions and the total score, indicating that syntactic simplification and structural optimization enhanced clients' understanding of ICFs. The 2 LLMs showed broadly similar patterns, with ChatGPT (GPT-5) slightly higher than Grok-4 on most dimensions. Notably, discrepancies arose between the Wilcoxon test and the LMM in model-to-model comparisons: Wilcoxon signed rank tests suggested no significant differences, whereas LMMs indicated that ChatGPT-5 was significantly superior in trustworthiness and acceptability. This discrepancy may reflect the small sample size (10 readers) and only moderate interrater reliability, which limited the power of nonparametric tests to detect differences. In interpreting these findings, we primarily relied on the LMM results, as they explicitly account for both document- and rater-level variability and are better suited to the hierarchical structure of the data. These observations accord with prior research. Will et al [58] reported that LLM-based rewriting improved the readability of online patient education materials without compromising accuracy, while Ali et al [49] found that optimized surgical consent forms enhanced comprehensibility while preserving professionalism and trustworthiness. Together with the present findings, this evidence suggests that LLMs can refine the structure and language of counseling ICFs to improve client comprehension while maintaining credibility and professional integrity.

Overall, this study is the first to systematically validate the feasibility and value of applying LLMs to optimize ICFs for university mental health counseling. The results demonstrated consistent improvements in text structure, content quality, and comprehensibility, with recognition from both professional evaluators and simulated clients. Notably, the baseline content quality scores of the original documents were relatively low (approximately mid-40s out of 100), and the observed gains therefore represent relative improvements rather than the attainment of an optimal standard. Importantly, the magnitude of these improvements was reflected in the β coefficients derived from the LMMs, which provide interpretable estimates of

XSL•FO
RenderX

practical effect size beyond statistical significance. These improvements suggest that LLMs can function as auxiliary tools in resource-limited higher education settings, enabling the rapid generation of standardized and accessible consent forms to address shortages of trained personnel and professional resources. More importantly, by lowering barriers to comprehension and engagement, LLM-based optimization may facilitate earlier help-seeking, promote more equitable access to counseling services, and strengthen the overall efficiency of youth mental health systems. While rewriting consent forms represents only an initial step, the strong performance observed here underscores the broader potential of LLMs to support standardized [49,51], equitable, and client-centered mental health services for adolescents and young adults worldwide.

In addition, the multidimensional evaluation framework developed in this study offers a methodological reference for future digital mental health research and highlights the value of prompt design and human-AI collaboration in sensitive contexts. Specifically, human-AI collaboration may help address key limitations identified in this study by allowing LLMs to support drafting and structural standardization, while human experts provide contextual judgment, ethical oversight, legal review, and final validation prior to real-world implementation. From a governance perspective, LLM-generated consent materials should therefore be regarded as assistive drafts rather than autonomous outputs, with institutional mechanisms such as mandatory human review and version control playing a central role in mitigating ethical, legal, and accountability risks. In practice, such collaboration may follow a structured workflow, in which LLMs generate an initial draft, counselors or mental health professionals review and refine content for contextual appropriateness and ethical compliance, and institutions provide final oversight and approval prior to deployment.

Limitations

Several limitations should be acknowledged. First, study materials were limited to Chinese-language ICFs from universities in China, which may constrain generalizability. Direct transfer to other settings may be limited by differences in language structure and readability conventions, as well as jurisdiction-specific legal and ethical requirements (eg, confidentiality exceptions, consent norms, and institutional procedures) and broader cultural expectations regarding counseling relationships. Future research should test applicability in English and other languages and adapt the evaluation framework through context-specific calibration of readability metrics and locally relevant ethical and regulatory standards to support cross-cultural use. Second, despite anonymization and randomization, rater blinding may not have been fully preserved. The substantially longer Grok-4–optimized texts may have been visually distinguishable from the originals, potentially introducing unconscious bias. Although efforts were made to reduce this risk, residual expectancy effects cannot be excluded. In addition, potential rater fatigue and sequence bias should be considered. Each expert evaluated 99 documents across 20 indicators, which may have imposed substantial cognitive load. Although the document order was randomized, fatigue-related effects cannot be fully excluded. Third, the reading comprehension assessment included only 10 volunteers, with moderate interrater consistency and no actual clients, underscoring the need for larger, more representative samples, including real audiences. This small sample size also limited statistical power, particularly for detecting subtle differences between LLM models. Future studies should further validate this evaluation framework against real client decision-making processes and counseling outcomes in applied settings. Moreover, university students may not represent individuals with lower health literacy or acute psychological distress, and documents understandable to college volunteers may still pose barriers to more vulnerable populations. Fourth, although our expert ratings captured multidimensional aspects of content quality (eg, completeness, specificity, and operability), the current evaluation did not systematically verify the factual and legal accuracy of LLM-generated content; given the known risk of hallucinations, optimized forms may inadvertently introduce legally incorrect statements, misinterpret jurisdiction-specific regulations (eg, provisions under the Chinese Mental Health Law), or omit critical ethical requirements despite fluent wording. Because no explicit legal texts were provided to the models, such outputs rely entirely on internal training data, which may be outdated or incomplete. This underscores the necessity of mandatory legal and ethical expert review before any real-world implementation. In addition, aggregate mean scores may obscure isolated but clinically significant errors. For example, LLM-generated drafts could omit mandatory clauses or include contextually inaccurate procedural details despite otherwise strong performance. Future studies should incorporate structured qualitative audits to complement quantitative evaluation. Fifth, this study evaluated only 2 closed-source models, ChatGPT (GPT-5) and Grok-4, without considering open-source systems or alternative architectures. In addition, model outputs were generated within a fixed time window, and future updates to these rapidly evolving systems may lead to different results even under identical prompts. As closed-source commercial systems undergo continuous updates, output behavior may shift over time, and future replications may yield different results due to model drift. Moreover, because outputs were generated via consumer-facing web interfaces rather than application programming interfaces with fixed parameters, unrecorded variables may have limited exact replication of the generation environment and further compounded model drift. Future work should therefore broaden model comparisons and incorporate transparent reporting of model versions and access dates to enhance reproducibility. Sixth, although a multidimensional evaluation framework was applied, this study did not conduct an in-depth analysis of specific textual revisions at the case level; future research could incorporate representative before and after examples to complement the quantitative findings. Finally, improvements in operability-related items were limited, suggesting current models remain insufficient in generating explicit execution pathways; this points to directions such as prompt engineering or human-AI collaboration to enhance operability in future work.

XSL•FO
RenderX

Conclusion

This study is the first to systematically evaluate the application of LLMs in optimizing ICFs for university mental health counseling. We found significant improvements in structure, readability, content quality, and comprehension, with consistent recognition from both experts and simulated clients. These findings suggest that LLMs can serve as tools to improve currently deficient counseling consent documentation by enhancing clarity and standardization. However, they may not yet produce fully comprehensive or high-quality consent forms without significant human refinement. For universities or primary institutions with limited professional resources, LLMs may serve as practical auxiliary tools to generate accessible, high-quality materials that facilitate early engagement with counseling. Further research is required to evaluate their safety, accuracy, and practical feasibility in real-world counseling workflows.

Acknowledgments

The authors thank Meiou Wang and Linli Yi for their assistance during the testing phase of this study. We are also grateful to Jiajie Zhang and colleagues for their expert scoring, and to Dongdong Zhu, Yuzhi Zhang, Hongfang Chen, and others for their contributions as readers in the rating process.

Funding

The authors declared no financial support was received for this work.

Data Availability

All data directly supporting the findings of this study are included in Multimedia Appendix 5. Additional underlying data can be made available upon reasonable request to the corresponding author (wwtdoctor02@163.com). Requests from verified academic researchers will be assessed by the Data Access Committee and answered within 1 month, in accordance with institutional and ethical requirements.

Parts of the analysis code used in this study are included in Multimedia Appendix 4. Qualified researchers may request access to the remaining code by contacting the co-corresponding authors (wwtdoctor02@163.com and ljc846702504@163.com). For requests submitted by verified academic researchers, the Data Access Committee will assess the request and grant access within 1 month.

Authors' Contributions

J Luo designed the study; performed code development, statistical analysis, and preliminary testing; and drafted the manuscript. JM coordinated the study and contributed to code development, figure preparation, document collection, preliminary testing, and manuscript writing. DZ, J Li, and YZ collected documents, organized reader and expert ratings, and assisted in manuscript preparation. LZ organized expert evaluations, supervised the study, revised the manuscript, and served as co-corresponding author. WW supervised the study, revised the manuscript, and coordinated submission as corresponding author. J Luo and JM contributed equally as first authors. LZ and WW contributed equally as corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed results of the analyses.
[PDF File, 822 KB - jmir_v28i1e86502_app1.pdf]

Multimedia Appendix 2
Supplementary information including prompt instructions, vocabulary lists, and bilingual (Chinese-English) design details.
[PDF File, 173 KB - jmir_v28i1e86502_app2.pdf]

Multimedia Appendix 3
Complete set of documents in 3 versions (Original, ChatGPT5-generated, and Grok4-generated).
[PDF File, 23240 KB - jmir_v28i1e86502_app3.pdf]

Multimedia Appendix 4
Source code used for data processing and statistical analysis.
[PDF File, 279 KB - jmir_v28i1e86502_app4.pdf]

XSL•FO
RenderX

Multimedia Appendix 5
Supplementary data including all scoring results across documents and versions.
[PDF File, 4142 KB - jmir_v28i1e86502_app5.pdf]

References

1. Dong J, Chen J, Li Y, Huang X, Rong X, Chen L. Relationship between freshmen's psychological health and family economic status in Chinese universities: a latent profile analysis. Psychol Res Behav Manag 2023;16:3489-3502. [doi: 10.2147/PRBM.S424798] [Medline: 37667735]
2. Osborn TG, Li S, Saunders R, Fonagy P. University students' use of mental health services: a systematic review and meta-analysis. Int J Ment Health Syst 2022 Dec 17;16(1):57. [doi: 10.1186/s13033-022-00569-0] [Medline: 36527036]
3. Deng X, Li C. The impact of university freshmen's mental health on academic performance: an empirical study based on M University in Fujian Province, China. Front Public Health 2024;12:1483031. [doi: 10.3389/fpubh.2024.1483031] [Medline: 39735743]
4. Lin GT, Mitchell MB, Hammack-Aviran C, Gao Y, Liu D, Langerman A. Content and readability of US procedure consent forms. JAMA Intern Med 2024 Feb 1;184(2):214-216. [doi: 10.1001/jamainternmed.2023.6431] [Medline: 38079183]
5. García-Álvarez JM, García-Sánchez A. Readability and comprehension of anesthesia informed consent forms in a Spanish county hospital: an observational study. Nurs Rep 2024 May 24;14(2):1338-1352. [doi: 10.3390/nursrep14020101] [Medline: 38921711]
6. Emanuel EJ, Boyle CW. Assessment of length and readability of informed consent documents for COVID-19 vaccine trials. JAMA Netw Open 2021 Apr 1;4(4):e2110843. [doi: 10.1001/jamanetworkopen.2021.10843] [Medline: 33909052]
7. Gerke L, Ladwig S, Pauls F, Trachsel M, Härter M, Nestoriuc Y. Optimized informed consent for psychotherapy: protocol for a randomized controlled trial. JMIR Res Protoc 2022 Sep 30;11(9):e39843. [doi: 10.2196/39843] [Medline: 36178713]
8. Pietrzykowski T, Smilowska K. The reality of informed consent: empirical studies on patient comprehension-systematic review. Trials 2021 Jan 14;22(1):57. [doi: 10.1186/s13063-020-04969-w] [Medline: 33446265]
9. Fischer AE, Venter WDF, Collins S, Carman M, Lalla-Edward ST. The readability of informed consent forms for research studies conducted in South Africa. S Afr Med J 2021 Feb 1;111(2):180-183. [doi: 10.7196/SAMJ.2021.v111i2.14752] [Medline: 33944731]
10. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA 2023 Sep 5;330(9):866-869. [doi: 10.1001/jama.2023.14217] [Medline: 37548965]
11. Wornow M, Xu Y, Thapa R, et al. The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med 2023 Jul 29;6(1):135. [doi: 10.1038/s41746-023-00879-8] [Medline: 37516790]
12. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025 Jan 28;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
13. Liu S, McCoy AB, Wright AP, et al. Leveraging large language models for generating responses to patient messages-a subjective analysis. J Am Med Inform Assoc 2024 May 20;31(6):1367-1379. [doi: 10.1093/jamia/ocae052] [Medline: 38497958]
14. Garcia P, Ma SP, Shah S, et al. Artificial intelligence-generated draft replies to patient inbox messages. JAMA Netw Open 2024 Mar 4;7(3):e243201. [doi: 10.1001/jamanetworkopen.2024.3201] [Medline: 38506805]
15. Schmidt J, Lichy I, Kurz T, et al. ChatGPT as a support tool for informed consent and preoperative patient education prior to penile prosthesis implantation. J Clin Med 2024 Dec 10;13(24):7482. [doi: 10.3390/jcm13247482] [Medline: 39768416]
16. Swisher AR, Wu AW, Liu GC, Lee MK, Carle TR, Tang DM. Enhancing health literacy: evaluating the readability of patient handouts revised by ChatGPT's large language model. Otolaryngol Head Neck Surg 2024 Dec;171(6):1751-1757. [doi: 10.1002/ohn.927] [Medline: 39105460]
17. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
18. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne) 2024;11:1477898. [doi: 10.3389/fmed.2024.1477898] [Medline: 39534227]
19. Chung EM, Zhang SC, Nguyen AT, Atkins KM, Sandler HM, Kamrava M. Feasibility and acceptability of ChatGPT generated radiology report summaries for cancer patients. Digit Health 2023;9:20552076231221620. [doi: 10.1177/20552076231221620] [Medline: 38130802]
20. Informed consent form for university counseling [Report in Chinese]. : The Clinical Psychology Registration Working Committee of the Chinese Psychological Society; 2022 URL: https://cspec.hntou.edu.cn/down/zqtys/202204/P020220419320054681253.pdf [accessed 2026-03-10]
21. ChatGPT. URL: https://chatgpt.com [accessed 2026-03-10]
22. Grok. 2025. URL: https://grok.com [accessed 2026-03-10]
23. Edwards A. OpenAI launches GPT-5 and multilingual performance shows little improvement. Slator. 2025. URL: https://slator.com/openai-launches-gpt5 [accessed 2026-03-10]
24. Grok-4 vs Kimi K2 vs Claude 4: a professional comparison of leading AI models in 2025. iWeaver AI. URL: https://www.iweaver.ai/blog/grok-4-vs-kimi-k2-vs-claude-4-professional-comparison-of-leading-ai-models [accessed 2026-03-10]

XSL•FO
RenderX

25. Karimi AH, Guyler MR, Hecht CJ II, Burkhart RJ, Acuña AJ, Kamath AF. Assessing the readability of clinical trial consent forms for surgical specialties. J Surg Res 2024 Apr;296:711-719. [doi: 10.1016/j.jss.2024.01.045] [Medline: 38367522]
26. Mirza FN, Wu E, Abdulrazeq HF, et al. The literacy barrier in clinical trial consents: a retrospective analysis. EClinicalMedicine 2024 Sep;75:102814. [doi: 10.1016/j.eclinm.2024.102814] [Medline: 39763593]
27. Lakretz Y, Hupkes D, Vergallito A, Marelli M, Baroni M, Dehaene S. Mechanisms for handling nested dependencies in neural-network language models and humans. Cognition 2021 Aug;213:104699. [doi: 10.1016/j.cognition.2021.104699] [Medline: 33941375]
28. Chang LC, Sun CC, Chen TH, Tsai DC, Lin HL, Liao LL. Evaluation of the quality and readability of ChatGPT responses to frequently asked questions about myopia in traditional Chinese language. Digit Health 2024;10:20552076241277021. [doi: 10.1177/20552076241277021] [Medline: 39229462]
29. Harrison RM, Lapteva E, Bibin A. Behavioral nudging with generative AI for content development in SMS health care interventions: case study. JMIR AI 2024 Oct 15;3:e52974. [doi: 10.2196/52974] [Medline: 39405108]
30. Ji M, Bouillon P. Predicting the linguistic accessibility of Chinese health translations: machine learning algorithm development. JMIR Med Inform 2021 Oct 7;9(10):e30588. [doi: 10.2196/30588] [Medline: 34617914]
31. Li Y, Zhou X, Zhou Y, et al. Evaluation of the quality and readability of online information about breast cancer in China. Patient Educ Couns 2021 Apr;104(4):858-864. [doi: 10.1016/j.pec.2020.09.012] [Medline: 32988687]
32. Wang Y, Yang X, Liu J. Navigating sensitive conversations: patient-centered communication and politeness markers in Chinese online medical consultations. Healthcare (Basel) 2024 Dec 6;12(23):2465. [doi: 10.3390/healthcare12232465] [Medline: 39685087]
33. Li C, Fu J, Lai J, et al. Construction of an emotional lexicon of patients with breast cancer: development and sentiment analysis. J Med Internet Res 2023 Sep 12;25:e44897. [doi: 10.2196/44897] [Medline: 37698914]
34. Ethical principles of psychologists and code of conduct. American Psychological Association. 2017. URL: https://www.apa.org/ethics/code [accessed 2026-03-10]
35. Code of ethics and conduct. British Psychological Society. 2018. URL: https://www.bps.org.uk/guideline/code-ethics-and-conduct [accessed 2026-03-10]
36. Chinese Psychological Society. URL: http://www.cpsbeijing.org [accessed 2026-03-10]
37. Comprehensive mental health action plan 2013–2030. : World Health Organization; 2021 Sep 10 URL: https://iris.who.int/server/api/core/bitstreams/69921758-6229-49ba-bd3d-c24736e35829/content [accessed 2026-03-10]
38. Steeb T, Reinhardt L, Harlaß M, Heppt MV, Meier F, Berking C. Assessment of the quality, understandability, and reliability of YouTube videos as a source of information on basal cell carcinoma: web-based analysis. JMIR Cancer 2022 Mar 11;8(1):e29581. [doi: 10.2196/29581] [Medline: 35275067]
39. Glaser J, Nouri S, Fernandez A, et al. Interventions to improve patient comprehension in informed consent for medical and surgical procedures: an updated systematic review. Med Decis Making 2020 Feb;40(2):119-143. [doi: 10.1177/0272989X19896348] [Medline: 31948345]
40. Mani NS, Ottosen T, Fratta M, Yu F. A health literacy analysis of the consumer-oriented COVID-19 information produced by ten state health departments. J Med Libr Assoc 2021 Jul 1;109(3):422-431. [doi: 10.5195/jmla.2021.1165] [Medline: 34629971]
41. Sørensen K, Pelikan JM, Röthlin F, et al. Health literacy in Europe: comparative results of the European Health Literacy Survey (HLS-EU). Eur J Public Health 2015 Dec;25(6):1053-1058. [doi: 10.1093/eurpub/ckv043] [Medline: 25843827]
42. Giordano M, Caccavella VM, Tariciotti L, Della Pepa GM, Olivi A, Polli FM. The reliability and quality of online patient education videos for vestibular schwannoma. Br J Neurosurg 2024 Aug;38(4):811-816. [doi: 10.1080/02688697.2021.1973369] [Medline: 34472385]
43. Zhang X, Zhang R. Impact of physicians' competence and warmth on chronic patients' intention to use online health communities. Healthcare (Basel) 2021 Jul 29;9(8):957. [doi: 10.3390/healthcare9080957] [Medline: 34442093]
44. Bucyibaruta JB, Peu D, Bamford L, van der Wath A. Closing the gaps in defining and conceptualising acceptability of healthcare: a qualitative thematic content analysis. Afr Health Sci 2022 Sep;22(3):703-709. [doi: 10.4314/ahs.v22i3.75] [Medline: 36910403]
45. Fang Y, Liu P, Gao Q. Assessment of knowledge, attitude, and practice toward COVID-19 in China: an online cross-sectional survey. Am J Trop Med Hyg 2021 Feb 19;104(4):1461-1471. [doi: 10.4269/ajtmh.20-0452] [Medline: 33606668]
46. Hall CE, Brooks SK, Potts HWW, Greenberg N, Weston D. Rates of, and factors associated with, common mental disorders in homeworking UK Government response employees' during COVID-19: a cross-sectional survey and secondary data analysis. BMC Psychol 2024 Aug 7;12(1):429. [doi: 10.1186/s40359-024-01921-4] [Medline: 39113130]
47. Leroy G, Kauchak D, Haeger D, Spegman D. Evaluation of an online text simplification editor using manual and automated metrics for perceived and actual text difficulty. JAMIA Open 2022 Jul;5(2):ooac044. [doi: 10.1093/jamiaopen/ooac044] [Medline: 35663117]
48. Leroy G, Kauchak D, Harber P, Pal A, Shukla A. Text and audio simplification: human vs. ChatGPT. AMIA Jt Summits Transl Sci Proc 2024;2024:295-304. [Medline: 38827082]
49. Ali R, Connolly ID, Tang OY, et al. Bridging the literacy gap for surgical consents: an AI-human expert collaborative approach. NPJ Digit Med 2024 Mar 8;7(1):63. [doi: 10.1038/s41746-024-01039-2] [Medline: 38459205]

XSL•FO
RenderX

50. Beattie J, Neufeld S, Yang DX, et al. Using large language models to create patient centered consent forms. Int J Radiat Oncol Biol Phys 2024 Oct;120(2):e612. [doi: 10.1016/j.ijrobp.2024.07.1346]
51. Shi Q, Luzuriaga K, Allison JJ, et al. Transforming informed consent generation using large language models: mixed methods study. JMIR Med Inform 2025 Feb 13;13:e68139. [doi: 10.2196/68139] [Medline: 39946400]
52. Oh N, Kim J, Park S, et al. Large language model-assisted surgical consent forms in non-English language: content analysis and readability evaluation. J Med Internet Res 2025 Jun 19;27:e73222. [doi: 10.2196/73222] [Medline: 40537063]
53. Rust P, Frings J, Meister S, Fehring L. Evaluation of a large language model to simplify discharge summaries and provide cardiological lifestyle recommendations. Commun Med (Lond) 2025 May 29;5(1):208. [doi: 10.1038/s43856-025-00927-2] [Medline: 40442348]
54. Romoff M, Brunette M, Peterson MK, Hashmi SZ, Kim MS. The role of large language models in improving the readability of orthopaedic spine patient educational material. J Orthop Surg Res 2025 May 28;20(1):531. [doi: 10.1186/s13018-025-05955-1] [Medline: 40426209]
55. Alber DA, Yang Z, Alyakin A, et al. Medical large language models are vulnerable to data-poisoning attacks. Nat Med 2025 Feb;31(2):618-626. [doi: 10.1038/s41591-024-03445-1] [Medline: 39779928]
56. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature 2024 Jun;630(8017):625-630. [doi: 10.1038/s41586-024-07421-0] [Medline: 38898292]
57. Gelinas L, Morrell W, Tse T, Glazier A, Zarin DA, Bierer BE. Characterization of key information sections in informed consent forms posted on ClinicalTrials.gov. J Clin Transl Sci 2023;7(1):e185. [doi: 10.1017/cts.2023.605] [Medline: 37745937]
58. Will J, Gupta M, Zaretsky J, Dowlath A, Testa P, Feldman J. Enhancing the readability of online patient education materials using large language models: cross-sectional study. J Med Internet Res 2025 Jun 4;27:e69955. [doi: 10.2196/69955] [Medline: 40465378]

Abbreviations

AI: artificial intelligence
ICC: intraclass correlation coefficient
ICF: informed consent form
LLM: large language model
LMM: linear mixed-effects model

Edited by A Coristine; submitted 25.Oct.2025; peer-reviewed by A gopinath, S Singireddy, Z Liu; accepted 24.Feb.2026; published 01.Apr.2026.

Please cite as:
Luo J, Ma J, Zhan D, Zhou Y, Li J, Zhang L, Wang W
Optimization of University Counseling Consent Forms With Large Language Models: Multidimensional Comparative Evaluation
J Med Internet Res 2026;28:e86502
URL: https://www.jmir.org/2026/1/e86502
doi:10.2196/86502

© Jianchen Luo, Jing Ma, Danni Zhan, Yuhong Zhou, Jiayu Li, Lan Zhang, Wentao Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Well-Being and Cognitive Factors Influencing Health Care Workers' Adherence to Internet-Based Stress Management: Mixed Methods Analysis of a Nonrandomized Controlled Study

Miqi Li[1*], MS; Quan Wang[2*], PhD; Danli Zheng[1], MS; Danni Feng[1], MS; Xiaorong Lang[1], MS

[1]Department of Nursing, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, No.1965 Jiefang Road, Wuhan, Hubei, China

[2]Nursing School, Tong Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China

[*]these authors contributed equally

Corresponding Author:
Danli Zheng, MS
Department of Nursing, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, No.1965 Jiefang Road, Wuhan, Hubei, China

Abstract

Background: High stress levels are common among health care workers (HCWs), threatening their health and workforce stability. Internet-based mobile stress management (MSM) is a promising intervention for reducing work-related stress; however, poor adherence limits effectiveness. Exploring factors influencing HCWs' adherence may thus aid in developing optimal interventions.

Objective: The research aimed to investigate (1) how HCWs' well-being and cognitive factors influenced MSM treatment adherence and (2) what HCWs' specific needs for MSM were.

Methods: This study was a convergent mixed methods secondary analysis of a nonrandomized controlled trial. HCWs who were currently employed, had internet access, had no serious medical problems, and were willing to participate were recruited by convenience sampling through an MSM project in a large Chinese general hospital from August 11, 2021, to January 31, 2022. Those intending to leave the hospital or with insufficient medical condition for follow-up were excluded. Quantitative data were collected from 157 HCWs (n=135, 86% female participants; mean age of 33.7, SD 4.9 y) electronically via Research Electronic Data Capture (REDCap). Measures included sociodemographic characteristics, the Fatigue Assessment Scale, the 14-item Perceived Stress Scale, a user experience questionnaire, an attitudes scale (perceived usefulness, feasibility, and enjoyment), and self-reported practice frequency. Qualitative data were collected via an open-ended question answered by 96 participants. Quantitative data were analyzed using hierarchical regression and structural equation modeling. Qualitative data were analyzed using reflexive-thematic analysis in NVivo (QSR International).

Results: In the quantitative study (n=157), hierarchical regression analyses showed that fatigue was a significant negative predictor of adherence (b=−0.050, 95% CI −0.086 to −0.015; t_{149}=−2.859; 2-tailed P=.005), while user experience (b=0.074, 95% CI 0.042-0.106; t_{148}=4.569; 2-tailed P<.001) and attitudes (b=0.238, 95% CI 0.140-0.336; t_{147}=4.863; 2-tailed P<.001) were positive predictors. Results from the structural equation model demonstrated a significant indirect effect of user experience on adherence through attitudes (indirect effect=0.047, 95% CI 0.021-0.083; P<.001), but no significant direct effect. Separately, qualitative data of responses (n=96) identified four themes about the experience, namely, (1) individualized intervention, (2) effective feedback, (3) reward and constraint mechanisms, and (4) duration of intervention.

Conclusions: This mixed methods study innovatively integrates quantitative and qualitative needs assessment to elucidate the adherence pathway in digital stress management for HCWs. It differentiates itself from prior research by moving beyond merely describing adherence barriers by quantifying the mediating role of attitudes between user experience and treatment adherence. The study indicates that despite fatigue barriers, improving user experience enhances positive attitudes toward MSM, which is vital for ensuring treatment adherence. This contributes to the field by elucidating a key psychological pathway for engagement. For real-world implications, to optimize user experience, future studies should consider focusing on individualized needs, providing effective feedback, developing reward and constraint mechanisms, and designing flexible intervention durations.

Trial Registration: ClinicalTrials.gov NCT05239065; https://clinicaltrials.gov/study/NCT05239065

(J Med Internet Res 2026;28:e73306) doi:10.2196/73306

XSL•FO
RenderX

KEYWORDS

adherence; stress management; internet-based intervention; fatigue; work-related stress; health care workers; user experience

Introduction

Health care workers (HCWs) are often more vulnerable to work-related stress due to multiple stressors that exist in health care contexts, including long working hours, excessive workload, shift work, and inadequate staffing [1-3]. Moreover, the postpandemic era may require HCWs to spend more energy and time on extra work, which places them in a more strained and stressful situation [4,5]. Data released by the World Health Organization in 2025 revealed that one-third of physicians and nurses experience depression or anxiety, with over 10% having experienced suicidal or self-harm ideation [6]. In line with this global pattern, a recent nationwide study among Chinese medical professionals revealed that 28.2% and 24% of participants exhibited major depressive and anxiety symptoms, respectively, with the accumulation of multiple work-related stressors significantly elevating mental health risks [7]. In general, it estimates indicate that the prevalence of stress among HCWs ranges from 29.8% to 63% [8]. The detrimental effects of overwhelming work-related stress on HCWs have been well established, such as mental disorders, sleep disturbance, and physical symptoms [9-11]. Long-term work-related stress may also reduce patient satisfaction and jeopardize patient safety [12,13]. Hence, implementation of stress management interventions is essential in improving the personal health of HCWs and further ensuring the stability of health care human resources.

For now, several stress management techniques, including mindfulness, deep breathing, and regular exercise, have been identified as simple to apply and effective in alleviating both the physical and psychological indices of stress [14-16]. However, adherence to stress management behaviors may not be that easy for HCWs with shift work and irregular working schedules. A recent mixed methods study concluded that cumulative stressors would hinder the application of recommended stress management techniques, resulting in half of them failing to adhere to stress management behaviors [17]. All these findings suggest that it is critical to develop an adaptable and accessible stress management intervention so that HCWs can better use stress management strategies to cope with daily stress.

With the widespread use of digital devices, accumulative research evidence has suggested that mobile stress management (MSM) intervention may produce promising effects on improving stress symptoms and well-being in the workplace [18-20]. However, unlike traditional interventions in which the psychotherapist may supervise the adherence to treatment through face-to-face communications, the adherence of online participants has always been a bottleneck issue in such interventions [21,22]. Recent studies have suggested that personal support might enhance the adherence to internet-based interventions rather than a “pure” self-guided intervention, and the effects varied with the intensity of the intervention [23]. For example, a randomized study found that the adherence rate for stress management in the content-focused guidance sample was 70.5%, higher than both the adherence-focused guidance sample (68.8%) and the administrative guidance sample (42%) [20]. On the one hand, personal guidance may be regarded as an effective approach to improving adherence to internet-based stress management; on the other hand, it would be meaningful to find out factors that influence interindividual differences in nonadherence to such interventions.

Although personal guidance can improve adherence to internet-based stress management, understanding the individual factors underlying nonadherence remains critical. Notably, significant gaps persist regarding how these factors function specifically in MSM interventions targeting HCWs. Previous research identified that well-being might be a critical individual factor that might influence treatment adherence. For example, a qualitative study conducted among general populations identified that high stress levels and fatigue symptoms would hinder individuals from prioritizing and engaging in internet-based interventions [24]. However, another study conducted among HCWs suggested that high workload rather than stress levels may act as a barrier to web-based interventions for modifying their lifestyle, including physical activity [25]. This might be explained by the fact that the appraisal of stress can vary among various people, while some individuals presume stress as a “hindrance,” others would identify it as a “challenge” instead [26]. Consequently, the extent to which high stress levels act as a hindrance for health care professionals is not very definite. It would be fascinating to investigate how individual well-being, as expressed by stress and fatigue complaints, impacts HCWs’ adherence to MSM.

Meanwhile, cognitive elements of therapy adherence should be considered. Users’ perspective on these intervention components or user experience cannot be overlooked. However, research specifically examining how user experience influences adherence to MSM among HCWs remains limited. Because work-related stress can cause a wide range of symptoms, including psychological, physical, and behavioral problems [27-29], those who experience different stress symptoms will require different interventions [30]. According to the Technology Acceptance Model models, individuals will only implement and use an intervention if they believe it is valuable and feasible [31]. To build an effective tailored stress management intervention, it would be required to focus on the user experience and investigate its effects on attitudes and adherence to MSM. This represents a significant gap in the current literature, particularly regarding HCW populations.

To summarize, the current research sought to investigate (1) how HCWs’ well-being and cognitive factors influenced their treatment adherence to MSM and (2) what HCWs’ requests for MSM were. In particular, we suggested the following research hypotheses:

H1: Individual well-being has a significant influence on HCWs’ treatment adherence, and fatigue has a negative effect, although stress levels do not.

XSL•FO
RenderX

H2: The user experience of HCWs would have significant effects on their perceived feasibility and use.

H3: The attitudes would have significant effects on HCWs' treatment adherence.

H4: The attitudes of HCWs would significantly mediate the impact of their user experience on treatment adherence.

Methods

Study Design

This study was a convergent mixed methods secondary analysis of a nonrandomized controlled trial [32,33]. It aimed to investigate the influence of HCWs' well-being and cognitive variables on their adherence to MSM use. This study was reported in compliance with the Good Reporting of a Mixed Methods Study [34]. We acquired secondary data from HCWs who participated in the MSM program at a large general hospital in China from August 11, 2021, to January 31, 2022 [32,33]. Information on the research intervention can be found on the ClinicalTrials.gov website (NCT05239065).

Inclusion and Exclusion

The inclusion criteria were (1) currently employed, (2) self-reported having internet access, (3) did not have any serious medical problems (eg, cancer or stroke), and (4) were willing to participate in this study and sign the consent form. Participants were excluded if they (1) intended to leave the hospital or units for any reason or (2) were in insufficient medical condition to be followed up on.

Participants Characteristics

Sociodemographic characteristics included sex (male or female), age (y), marriage status (married or unmarried), educational background (undergraduate degree or postgraduate degree), position (physician, nurse, or other), working hours (h), income (Chinese Yuan [CNY]), and BMI.

Sampling Procedures

A convenience sample of HCWs was recruited from a large general hospital in China between August 11, 2021, and January 31, 2022. Participants were enrolled from an ongoing MSM program.

Sample Size, Power, and Precision

This study was a secondary analysis of the original MSM program data. In the primary trial, 179 participants were ultimately allocated to and received the intervention. For the purposes of the current analysis, which designated the Fatigue Assessment Scale (FAS) as a key outcome, we applied a strict criterion of complete data: any record with one or more missing items on the 10-item FAS was excluded to ensure the integrity of the primary outcome measure without imputation. According to the inclusion and exclusion criteria, after applying this data completeness rule, 157 participants were ultimately enrolled and completed the quantitative survey. The sample size for the qualitative component was guided by the principle of information saturation, whereby recruitment continued until no substantially new themes emerged from the data. Ultimately, 96 individuals voluntarily agreed to participate.

Measures and Covariates

Well-Being Factors

Fatigue symptoms were measured using the FAS, a regularly used tiredness measure consisting of 10 questions, was used to assess fatigue symptoms [35]. Each item was graded using a 5-point Likert scale ranging from 0 (never) to 4 (always). Those with a higher FAS total score had more severe fatigue symptoms. The FAS has a Cronbach α of 0.883, suggesting that the internal reliability of this scale was satisfactory in this study.

Perceived stress was assessed using the 14-item Perceived Stress Scale (PSS-14) to gage stress levels at the baseline. PSS-14 is an internationally used self-reported 5-point Likert scale, with a greater total score indicating a higher degree of stress [36]. The Cronbach α was 0.835, indicating that the PSS-14 was a credible instrument in this study for measuring felt stress.

Cognitive Factors

User experience was evaluated using an 8-item questionnaire ("Can you tell us about your experience with this research? [in terms of studying organization, quality of staff for communication or explanation, quality of videos, quality of instructional brochure, quality of education regarding the effects of stress on health, quality of education regarding stress management strategies, quality of weekly feedback reports, and portable device and app applications]") with a 5-point Likert scale (1 [extremely terrible] to 5 [extremely excellent]), and the overall user experience score varied from 8 to 40, with a higher score indicating a better user experience. The Cronbach α coefficient for these 8 items was 0.933, indicating that the 8-item questionnaire was trustworthy for assessing HCW user experience.

Attitudes were assessed on three dimensions: perceived use, perceived feasibility, and perceived enjoyment. Perceived usefulness was defined as "the level of one's belief in using technology," perceived feasibility was defined as "a person's belief in the use of a technology that can be easily used," and perceived enjoyment was defined as "the degree of enjoyment of using a technology," as suggested by the previous study. We specifically investigated HCW sentiments using three items: (1) "Did you find the stress management intervention study beneficial in general?" (perceived use), (2) "Did you find the stress management intervention study to be straightforward to complete?" (feasibility), and (3) "Did you find the stress management intervention study interesting overall?" (perceived enjoyment). The 3 items were graded on a 5-point Likert scale, with higher scores suggesting a more favorable attitude to MSM. The Cronbach α coefficient for these 3 items was 0.713, showing that the 3-item questionnaire was consistent enough to assess attitudes against MSM among HCWs.

Treatment Adherence

Treatment adherence was measured by self-reported practice frequency of three stress management techniques: physical activity, deep breathing exercises, and mindfulness, using a 7-point Likert scale ("Did you practice the stress management

XSL•FO
RenderX

interventions?" 1 [never] to 7 [everyday]). A higher overall score suggested more MSM adherence. The Cronbach α coefficient for these 3 items was 0.793, indicating that the 3-item questionnaire was trustworthy in assessing HCWs' adherence to MSM.

Open-Ended Survey

In order to learn about the unique needs of HCWs when using MSM, we requested HCWs to answer an open-ended question (online interview), "Do you have any suggestions for improving the mobile stress management program?"

Data Collection

Research Electronic Data Capture (REDCap), a web-based online survey tool, was used to capture both quantitative and qualitative data. The MSM was a human-assisted intervention that provided weekly feedback and online stress management education. Stress management training was provided by a neuropsychologist in the form of video clips, posters, and messaging. Participants were encouraged to use at least 1 of the 3 recommended stress management approaches on a weekly basis (physical activity, deep breathing, and mindfulness). Participants were asked to keep a weekly journal that documented their practice frequency and mood state, and they were subsequently given feedback based on their replies. Details of the intervention and data collection have been published previously [32,33].

Statistical Analysis

Overview

This study used a convergent mixed methods design. Qualitative and quantitative data were collected concurrently but analyzed separately. The integration occurred during the interpretation phase through triangulation; the qualitative findings (user-identified needs and preferences) were used to explain and contextualize the quantitative results (statistical predictors of adherence). This approach allowed us to understand not only what factors influence adherence but also why and how they matter from the users' perspective.

Quantitative Data Analysis

IBM SPSS (version 26.0; IBM Corp) and IBM SPSS Amos (version 28.0; IBM Corp) were used to analyze quantitative data. For each categorical item, a descriptive analysis of the response frequencies was performed, and the mean (SD) for the continuous variables was calculated. Then, hierarchical regression analyses were run to examine the relative contributions of demographic variables, well-being factors, and cognitive factors in treatment adherence variances. Four sequential models were tested, with R^2 change indicating the additional variance explained at each step and a 2-side α level of .05.

To test the hypothesized mediation model, structural equation modeling with maximum likelihood estimation was used. The mediation effect of attitudes between user experience and adherence was examined using bootstrapping with 5000 resamples. Model fit was assessed using multiple indices including the χ^2/df ratio (3), root mean square error of approximation (<0.08), standardized root mean square residual (<0.08), comparative fit index (>0.90), and Tucker-Lewis index (>0.90) [17]. The indirect effects were examined, and a significant effect was determined if the 95% CI did not include zero.

Qualitative Data Analysis

QSR NVivo 14.0 was used to analyze qualitative data. The open-ended survey responses were analyzed using a pragmatic approach to reflexive-thematic analysis [37]. The analysis was carried out in six steps: (1) familiarization with the data, (2) code generation, (3) theme construction, (4) review of potential themes, (5) theme definition and naming, and (6) report production. In this stage of the study, 2 experienced researchers (ML and QW) independently read the transcript and coded the findings. Both researchers then met to compare their coding and discuss any discrepancies until a consensus was reached on the thematic structure.

Ethical Considerations

This study was approved by the Institutional Review Board (IRB) of the Medical Ethics Committee of Tongji Medical College, Huazhong University of Science and Technology, China (approval number: IRB number 2021S141). All participants provided informed consent electronically prior to participation. For the secondary analysis reported in this manuscript, the original IRB approval and consent procedures permitted the use of deidentified data. All data were collected and stored anonymously using coded identifiers. No personal identifying information was linked to the survey responses, ensuring participant confidentiality. Participants did not receive any financial compensation for their involvement in the study. This manuscript does not contain any personal images, audio or video recordings, or text excerpts that could lead to the identification of an individual participant.

Results

Participant Characteristics

There were 157 participants in total, with 135 (86%) female participants and 22 (14%) male participants. In terms of age, the mean age was 33.7 (SD 4.9) years, 35 (22.3%) participants were under the age of 30 years, 97 (61.8%) were between the ages of 30 and 40 years, and 25 (15.9%) were over the age of 40 years. In terms of marital status, 115 (73.2%) participants were married, while 42 (26.8%) were single. In terms of position, 30 participants (19.1%) were physicians, 104 (66.2%) were nurses, and 23 (14.7%) were other HCWs. When it came to family income, 18 (11.5%) of the participants had less than 10,000 CNY (1 CNY≈0.155 USD). 81 (51.6%) participants had a family income of 10,000 to 15,000 CNY, 35 (22.3%) had a family income of 15,001 to 20,000 CNY, and 23 (14.6%) had a family income of more than 20,000 CNY. When it came to working hours, 22 (14%) participants worked less than 35 hours per week, 74 (47.1%) worked 35 to 44 hours per week, and 61 (38.9%) worked more than 44 hours per week (Table 1).

XSL•FO
RenderX

Table . Baseline characteristics of participants in the quantitative component of a mixed methods study on mobile stress management adherence (n=157)

Characteristics	Participants, n (%)
Sex	
Female	135 (86)
Male	22 (14)
Age group (y)	
<30	35 (22.3)
30 to 40	97 (61.8)
>40	25 (15.9)
Marital status	
Married	115 (73.2)
Unmarried	42 (26.8)
Educational level	
Undergraduate degree	120 (76.4)
Postgraduate degree	37 (23.6)
Occupation	
Physician	30 (19.1)
Nurse	104 (66.2)
Other health care workers	23 (14.7)
Income (Chinese Yuan)[a]	
<10,000	18 (11.5)
10,000 to 15,000	81 (51.6)
15,001 to 20,000	35 (22.3)
>20,000	23 (14.6)
Working (h)	
<35	22 (14)
35 to 44	74 (47.1)
>44	61 (38.9)
BMI	
≤23.9	120 (76.4)
>23.9	37 (23.6)

[a]Income is reported in Chinese Yuan; the average exchange rate was 1 Chinese Yuan≈0.155 USD during the study period of 2021 - 2022.

Descriptive Statistics

Table 2 displays descriptive results for the main study variables. Participants rated perceived usefulness as 2.96 out of 5.0 (SD 0.85), perceived feasibility as 3.71 out of 5.0 (SD 0.72), and perceived enjoyment as 3.59 out of 5.0 (SD 0.69), in that order. The HCWs rated their experience with MSM as 32.03 out of 40.0 (SD 5.01), and the "quality of weekly feedback reports" received the highest mean score of 4.17 (SD 0.73). Participants rated their adherence to MSM for physical activity as 2.91 out of 5.0 (SD 1.15), deep breathing as 3.25 out of 5.0 (SD 1.34), and mindfulness as 2.92 out of 5.0 (SD 1.33), on average.

XSL•FO
RenderX

Table . Descriptive statistics of attitude, use experience, and adherence to mobile stress management in quantitative component (n=157).

Variables	Mean (SD; range)
Experience[a]	32.03 (5.01; 8-40)
Experience regarding study organization	3.74 (0.71; 1-5)
Quality of staff for communication or explanation	4.08 (0.77; 1-5)
Quality of the videos	4.03 (0.80; 1-5)
Quality of education regarding the effects of stress on health	4.09 (0.77; 1-5)
Quality of education regarding stress management strategies	4.06 (0.77; 1-5)
Quality of weekly feedback reports	4.17 (0.73; 1-5)
Applications of portable device and app	4.01 (0.75; 1-5)
Attitude[a]	10.26 (1.82; 3-15)
Perceived usefulness	2.96 (0.85; 1-5)
Perceived feasibility	3.71 (0.72; 1-5)
Perceived interest	3.59 (0.69; 1-5)
Adherence to mobile stress management[a]	9.15 (3.22; 3-21)
Physical activity	2.97 (1.15; 1-7)
Deep breathing	3.25 (1.34; 1-7)
Mindfulness	2.92 (1.33; 1-7)

[a]Composite score of questionnaire.

Hierarchical Regression Analyses

For the hierarchical regression analysis, 4 models were obtained. As shown in Table 3, the demographic variables (sex, age, marital status, position, income, and working hours) were entered into the model in the first step, and no variable was found to be significant in explaining the variance of treatment adherence. When perceived stress and fatigue symptoms were included in step 2, the percentage of treatment adherence explained by the predictors significantly increased from R^2=0.043 in model 1 to R^2=0.093 in model 2, and fatigue symptom was demonstrated to be a significant predictor (b=−0.050; t_{149}=−2.859; and 2-tailed P=.005). Then, after user experience was included as the predictor in step 3, the percentage of treatment adherence explained by the predictors significantly increased from R^2=0.093 in model 2 to R^2=0.205 in model 3, and user experience was demonstrated to be a significant predictor (b=0.074; t_{148}=4.569; and 2-tailed P<.001). In the last step, when attitude was included as the predictor, the percentage of treatment adherence explained by the predictors significantly increased from R^2=0.205 in model 3 to R^2=0.315 in model 4, while attitude was shown to be a significant predictor (b=0.238; t_{147}=4.863; and 2-tailed P<.001), the significance of user experience was shown to be disappearing in model 4 (b=0.031; t_{147}=1.770; and P=.08). All the predictors of model 4 together explained 31.5% of the variance of treatment adherence (R^2=0.315; P<.001).

XSL•FO
RenderX

Table . Hierarchical regression analysis of mobile stress management adherence (n=157).

Predictors	Model 1 (R^2=0.043; F test (5, 151)=1.385)			Model 2 (R^2=0.093; F test (7, 149)=2.182[a])			Model 3 (R^2=0.205; F test (8, 148)=4.774[b])			Model 4 (R^2=0.315; F test (9, 147)=7.520[b])		
	b [c]	β[d]	2-tailed t test (151)	b	β	2-tailed t test (149)	b	β	2-tailed t test (148)	b	β	2-tailed t test (147)
Constant	1.750	—[e]	3.773[b]	2.505	—[e]	3.789[b]	0.076	—[e]	0.093	−0.814	—[e]	−1.040
Sex	0.095	0.031	0.359	0.100	0.032	0.383	0.174	0.056	0.708	0.082	0.027	0.358
Age	0.185	0.084	0.997	0.170	0.078	0.934	0.078	0.036	0.452	−0.068	−0.031	−0.416
Position	0.113	0.131	0.865	0.051	0.035	0.389	0.002	0.001	0.014	0.033	0.023	0.292
Income	−0.185	−0.151	−1.785	−0.173	−0.141	−1.701	−0.119	−0.097	−1.241	−0.052	−0.042	−0.573
Working hours	0.151	0.069	0.794	0.167	0.076	0.883	0.192	0.088	1.084	0.174	0.079	1.054
Perceived stress	—	—	—	0.026	0.152	1.634	0.017	0.100	1.140	0.013	0.076	0.921
Fatigue symptoms	—	—	—	−0.050	−0.264	−2.859[f]	−0.041	−0.214	−2.452[f]	−0.038	−0.201	−2.468[f]
User experience	—	—	—	—	—	—	0.074	0.345	4.569[b]	0.031	0.144	1.770
Attitudes	—	—	—	—	—	—	—	—	—	0.238	0.402	4.863[b]

[a]P<0.05.

[b]P<0.001.

[c]b=unstandardized coefficient.

[d]β=standardized coefficient.

[e]Not applicable.

[f]P<.01.

Mediation Analysis

According to Hu and Bentler's [38] suggested model fit indicators, the model in this paper had a good fitting degree. Table 4 shows that the χ^2/df values were less than 3, the root mean square error of approximation and standardized root mean square residual values were less than 0.80, and the nonnormed fit index, comparative fit index, and adjusted goodness of fit index values were all greater than 0.90, all of which were within the range of acceptable fit indexes. To summarize, the results of confirmatory factor analysis confirm that the empirical model comprising attitude, experience, and adherence has good validity; the collected data are highly reliable; and the research findings are robust.

XSL•FO
RenderX

Table . Goodness of fit indices for the structural equation modeling testing the mediation effect of attitudes.

Index	Goodness-of-fit measure		Model fit
	Good fit	Accept fit	
χ^2/df	$0 \leq \chi^2/df \leq 2$	$2 < \chi^2/df \leq 3$	1.640
P value	$0.05 < P \leq 1.00$	$0.01 < P \leq 0.05$	0.073
RMSEA[a]	0≤ RMSEA ≤0.05	0.05< RMSEA ≤0.08	0.064
SRMR[b]	0≤ SRMR ≤0.05	0.05< SRMR ≤0.08	0.061
NNFI[c]	0.97< NNFI ≤1.00	0.95≤ NNFI ≤0.97	0.951
CFI[d]	0.97< CFI ≤1.00	0.95≤ CFI ≤0.97	0.980
AGFI[e]	0.90< AGFI ≤1.00	0.95≤ AGFI ≤0.97	0.928

[a]RMSEA: root mean square error of approximation.

[b]SRMR: standardized root mean square residual.

[c]NNFI: nonnormed fit index.

[d]CFI: comparative fit index.

[e]AGFI: adjusted goodness of fit index.

Figure 1 depicts the results of attitude mediation between experience and adherence to MSM. The path coefficient from experience to attitude was 0.079 (*P*<.001), and the path coefficient from attitude to adherence was 0.568 (*P*<.001). Meanwhile, the Bootstrapping estimates revealed that the direct influence of experience on adherence was not significant (direct effect=0.002; *P*=.86), implying that experience had a full indirect effect on adherence via the attitude toward MSM (indirect effect=0.047; *P*<.001).

Figure 1. The mediating role of attitudes in the relationship between user experience and adherence to mobile stress management.

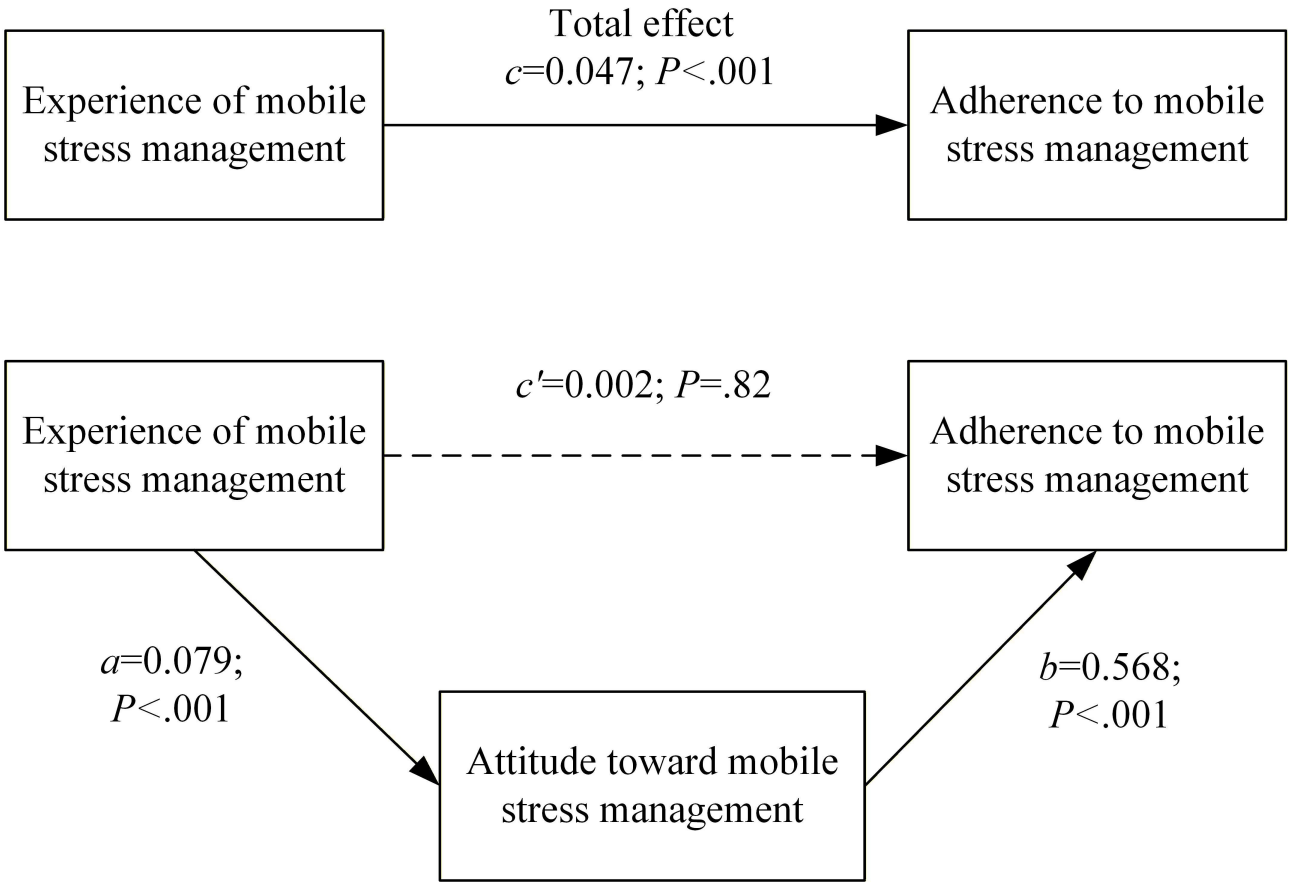

Results of Open-Ended Survey Response

The open-ended survey received 96 responses from 157 participants, which were identified and categorized into four themes: (1) individualized intervention, (2) effective feedback, (3) reward and constraint mechanisms, and (4) duration of intervention. Each theme could have 3 subthemes. Table 5 summarizes the themes, subthemes, and sample quotes.

XSL•FO
RenderX

Table . Themes, subthemes, and sample quotes from the open-ended survey responses regarding needs for improving mobile stress management in qualitative analyses (n=96).

Themes and subthemes	Sample quotes
Individualized intervention	
Diversified modules	• *"Everyone's stress sources are different and require one-on-one counseling from professionals."* (Participant 79) • *"I want to learn more about how to alleviate my anxiety."* (Participant 118) • *"I find it hard to change my sleep quality."* (Participants 83 and 107) • *"I want to know how to reduce negative emotions and deal with doctor-patient conflicts."* (Participant 100)
Adapt to work schedules	• *"I often work too long hours, and I find it difficult to guarantee enough time to practice those techniques."* (Participant 9) • *"I didn't have time to exercise."* (Participant 26)
Staging the intervention	• *"It's hard to get started."* (Participant 137) • *"It's hard to insist on."* (Participant 14, 32, 50, 128) • *"Professionals are required to follow up and guide the practice process."* (Participant 79)
Effective feedback	
Feedback content	• *"The content of the feedback should be more specific to the weekly plan and monthly schedule, which would be more intuitive."* (Participant 68) • *"The weekly questionnaire should be adapted to each individual's situation."* (Participant 41) • *"It would be better if we could get summarized and analyzed feedback data on a weekly or biweekly basis."* (Participant 147)
Feedback frequency	• *"More frequent reminders are needed."* (Participant 143) • *"We need more supervision and feedback."* (Participant 117)
Reward or constraint mechanisms	
Motivations for practice	• *"Giving rewards to stimulate movement would be useful."* (Participant 108) • *"It would work by motivating everyone to exercise through group check-ins."* (Participant 50) • *"It's important to mobilize the wisdom of everyone in the group and share more methods, skills, and experiences together."* (Participant 156)
More restraints	• *"It is completely dependent on self-consciousness and self-restraint."* (Participant 158) • *"It would be best if participants could be supervised and demanded a little more."* (Participant 79) • *"A little more restraint is needed to help us form habits."* (Participant 96)
Duration of intervention	
Short-term intervention	• *"Reduce the experimental cycle."* (Participant 18)
Long-term intervention	• *"I hope the project will continue for a longer time."* (Participant 54, 90)

Discussion

Principal Findings

This research focused on how individual treatment adherence differed depending on their well-being and cognitive factors. By using a convergent mixed methods approach, we specifically examined the direct effects of well-being factors on treatment adherence and the indirect effects of cognitive factors on treatment adherence, as well as learning about the specific requirements of HCWs for MSM interventions.

Overall, adherence to the internet-based MSM was generally modest, with mindfulness and physical activity techniques showing lower adherence compared to deep breathing exercises. The results are consistent with previous studies [20,22], despite

XSL•FO
RenderX

the reduction in dropouts in the internet-based intervention compared to face-to-face, participants often fail to have complete adherence. While physical activity and mindfulness are evidence-based techniques, their lower adherence in our sample may relate to the demanding work schedules of HCWs [8,16,19,39]. A previous study showed that breathing exercises can reduce stress levels even in the short term (mean time of 1.9, SD 1.2 min) [40], which means that breathing exercises may be more available for HCWs who spend most of their time providing health care. HCWs are a special group of people with heavy workloads and limited job autonomy. The time spent setting requirements and the accessibility of management programs were priorities in stress management for this group of people. Therefore, in the future, the stress management of specific interventions needs to be based on the working conditions such as the working hours or environment of different HCWs, taking into account the feasibility of developing a personalized management plan to improve adherence.

On the issue of burnout symptoms, this study revealed that fatigue symptoms were a significant predictor rather than perceived stress. An explanation for this finding is that perceived stress did not reach a statistically significant level in relation to burnout. One possible reason for this could be attributed to insufficient training hours, while another factor may be the nurses' anticipation of encountering stressful work in their future shifts. Additionally, according to Maslow's hierarchy of needs theory, fatigue symptoms represent fundamental and immediate needs, whereas perceived stress can potentially be managed through various methods by HCWs. As professional practitioners, HCWs consciously adapt themselves to reduce stress levels; however, it is important to consider potential biases resulting from psychological effects during the design phase of this study.

Our findings regarding the prominence of fatigue over perceived stress align with broader postpandemic research on psychological well-being, which suggests a shift in the salient components of well-being. Babazadeh [41], using a multidimensional measure of psychological well-being, found that factors like health locus of control and social support were significant predictors, underscoring the importance of targeting specific well-being components and cognitive factors in interventions rather than focusing solely on global stress measures. The significant role of fatigue in our study may reflect its centrality within the somatic symptoms dimension of well-being, which appears highly relevant for HCWs. Furthermore, fatigue and burnout may have intertwined effects, as evidenced by a prospective cohort study reporting that individuals in the "burnout and fatigue" group had a higher proportion of chronic course type compared to those experiencing "pure burnout" or "pure fatigue." Severe burnout and fatigue symptoms become interconnected with each other as well as with other psychological and/or somatic complaints [42]. However, the specific association between fatigue and specific factors remains unclear [43].

The most interesting finding is that the attitude mediates experience and adherence. Our mediation analysis revealed a significant indirect effect of user experience on adherence through attitudes, while the direct effect was not significant. Most existing studies have emphasized the importance of user experience on treatment adherence [8,20]. However, the direct relationship between the experience and adherence is limited to a definite explanation. On the contrary, the discussion of attitude and adherence is a classic issue, and the amount of evidence reports the relationship between attitude and adherence in stress management interventions [8,44,45]. Our research indicates that the direct impact of user experience on treatment adherence disappeared after inputting the variable of attitudes. An implication of the further relationship is the possibility that the improvement of experience with the object of adherence increasing is to make the attitude of the internet-based mobile program or management positive, despite the weak significant association between the experience and attitude. This may provide another perspective on retention in stress management, and a similar relationship in other psychological health of HCWs should be developed in the future.

We also identify the 4 aspects of individualized intervention, effective feedback, reward or constraint mechanisms, and duration of intervention as certain factors to adherence to internet-based MSM. Generally, all the factors highlight the personalized service throughout the whole intervention. The high quality and individual needs of intervention contents are more vital; inadequately supported contents are a hindrance. Furthermore, HCWs, as a special population with a health care education background, face the problem of their personal situations, may not have time to integrate exercises into their daily lives, and may need more personal conversations with the eCoach during the staging and feedback stages. This is in line with previous observations [46-48]. In the reward or constraint mechanisms, HCWs required the increasing demands outwards, which addressed the role of the organization in the interventions. Proving reminders, encouraging each other, and using rewards as the external triggers were taken as effective measures to increase adherence in other evidence [49,50]. To the duration of intervention, extensive studies have pointed out that stress was reduced with the increasing duration of intervention [9,51,52]. Considering that health care often requires demanding physical and psychological efforts, HCWs often have little energy to implement stress management strategies on a daily basis. Therefore, future studies may need to be more focused on less exhausting interventions for those working in health care contexts.

Limitations

Some limitations in the paper also need to be taken into consideration. First, the exploratory study investigated a small sample size with a short-term intervention. Second, the data collection of qualitative research conducted online could result in lacking information. Furthermore, the implication of this paper is that participants in qualitative research may have been positively biased in their support of the intervention because all the interviewees finished the intervention. It would be valuable in future work to extend the sample, focus on those who chose to drop out of the intervention, and explore the reasons for dropping out.

XSL•FO
RenderX

Implications

The findings of this study highlight a central pathway for engagement in digital interventions and point to clear practical directions. By innovatively integrating quantitative modeling with qualitative assessment, this research moves beyond merely describing barriers to adherence. It thereby differentiates itself from prior work by quantitatively establishing the critical mediating role of participant attitudes between user experience and sustained use. This contributes a key insight to the field: successful engagement depends not just on the intervention itself, but on fostering positive attitudes through that experience. Consequently, the real-world implication is straightforward. To optimize adherence, MSM programs for HCWs must prioritize enhancing user experience via personalized content, actionable feedback, incentive mechanisms, and flexible duration, all essential for cultivating the positive attitudes needed for long-term use in high-demand settings.

Conclusion

In summary, by innovatively integrating quantitative modeling with qualitative assessment, this study demonstrates that the value of user experience in stress management adherence to internet-based mobile intervention is fully mediated by participant attitudes, highlighting a key insight that differentiates our work from previous barrier-focused research. For HCWs, especially in high-demand settings, adherence is enhanced by interventions that are perceived as personalized, with actionable feedback, incentive mechanisms, and flexible duration. Future research should incorporate additional mediating variables through larger-scale and longer-term studies to further elucidate these mechanisms.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing. The GAI tool used was DeepSeek-V3.1 Terminus. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors do not bear responsibility for the final outcomes. Declaration submitted by ML on behalf of all the coauthors.

Funding

This work was supported by the National Key R&D Program of China (2021YFC2502200). ML is additionally supported by Scientific Research Fund of Tongji Hospital (TJ2025D05). The funders were not involved in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: ML, QW, DZ
Data curation: ML, QW, DF, XL
Formal analysis: ML, QW
Investigation: ML, QW, DF, XL
Methodology: ML, QW, DF, XL
Project administration: ML, QW, DZ
Supervision: ML, DZ
Validation: ML, QW, DZ
Writing- original draft: ML, QW
Writing-review & editing: ML, QW, DF, XL, DZ

Conflicts of Interest

None declared.

References

1. Joseph B, Joseph M. The health of the healthcare workers. Indian J Occup Environ Med 2016;20(2):71-72. [doi: 10.4103/0019-5278.197518] [Medline: 28194078]
2. Medisauskaite A, Kamau C. Prevalence of oncologists in distress: systematic review and meta-analysis. Psychooncology 2017 Nov;26(11):1732-1740. [doi: 10.1002/pon.4382] [Medline: 28116833]
3. Reith TP. Burnout in United States healthcare professionals: a narrative review. Cureus 2018 Dec 4;10(12):e3681. [doi: 10.7759/cureus.3681] [Medline: 30761233]
4. Caruso CC. Negative impacts of shiftwork and long work hours. Rehabil Nurs 2014;39(1):16-25. [doi: 10.1002/rnj.107] [Medline: 23780784]

XSL•FO
RenderX

5. Rosa D, Terzoni S, Dellafiore F, Destrebecq A. Systematic review of shift work and nurses' health. Occup Med (Lond) 2019 Jun 24;69(4):237-243. [doi: 10.1093/occmed/kqz063] [Medline: 31132107]
6. Healing hands – hurting minds. World Health Organization. 2025. URL: https://www.who.int/europe/news/item/10-10-2025-healing-hands---hurting-minds [accessed 2025-11-17]
7. Chen L, Wu L, Xia J, Cai X, Chen L. Gender differences of the association between work-related stressors and mental health among Chinese medical professionals: a nationwide cross-sectional study. Front Public Health 2025;13:1561588. [doi: 10.3389/fpubh.2025.1561588] [Medline: 40265062]
8. Vizheh M, Qorbani M, Arzaghi SM, Muhidin S, Javanmard Z, Esmaeili M. The mental health of healthcare workers in the COVID-19 pandemic: a systematic review. J Diabetes Metab Disord 2020;19(2):1967-1978. [doi: 10.1007/s40200-020-00643-9] [Medline: 33134211]
9. Bischoff LL, Otto AK, Hold C, Wollesen B. The effect of physical activity interventions on occupational stress for health personnel: a systematic review. Int J Nurs Stud 2019 Sep;97:94-104. [doi: 10.1016/j.ijnurstu.2019.06.002] [Medline: 31234106]
10. George LS, Lais H, Chacko M, Retnakumar C, Krishnapillai V. Motivators and barriers for physical activity among health-care professionals: a qualitative study. Indian J Community Med 2021;46(1):66-69. [doi: 10.4103/ijcm.IJCM_200_20] [Medline: 34035579]
11. Philbrick G, Sheridan NF, McCauley K. An exploration of New Zealand mental health nurses' personal physical activities. Int J Ment Health Nurs 2022 Jun;31(3):625-638. [doi: 10.1111/inm.12981] [Medline: 35166003]
12. Conversano C, Ciacchini R, Orrù G, Di Giuseppe M, Gemignani A, Poli A. Mindfulness, compassion, and self-compassion among health care professionals: what's new? A systematic review. Front Psychol 2020;11:1683. [doi: 10.3389/fpsyg.2020.01683] [Medline: 32849021]
13. Goyal M, Singh S, Sibinga EMS, et al. Meditation programs for psychological stress and well-being: a systematic review and meta-analysis. JAMA Intern Med 2014 Mar;174(3):357-368. [doi: 10.1001/jamainternmed.2013.13018] [Medline: 24395196]
14. Heber E, Ebert DD, Lehr D, et al. The benefit of web- and computer-based interventions for stress: a systematic review and meta-analysis. J Med Internet Res 2017 Feb 17;19(2):e32. [doi: 10.2196/jmir.5774] [Medline: 28213341]
15. Stanic J, Barth J, Danon N, Bondolfi G, Jermann F, Eicher M. Adherence to standardized 8-week mindfulness-based interventions among women with breast or gynecological cancer: a scoping review. J Psychosoc Oncol Res Pract 2021;3(2):e048. [doi: 10.1097/OR9.0000000000000048]
16. Taylor H, Cavanagh K, Field AP, Strauss C. Health care workers' need for headspace: findings from a multisite definitive randomized controlled trial of an unguided digital mindfulness-based self-help app to reduce healthcare worker stress. JMIR mHealth uHealth 2022 Aug 25;10(8):e31744. [doi: 10.2196/31744] [Medline: 36006668]
17. Ebert DD, Kählke F, Buntrock C, et al. A health economic outcome evaluation of an internet-based mobile-supported stress management intervention for employees. Scand J Work Environ Health 2018 Mar 1;44(2):171-182. [doi: 10.5271/sjweh.3691] [Medline: 29144535]
18. Ebert DD, Franke M, Zarski AC, et al. Effectiveness and moderators of an internet-based mobile-supported stress management intervention as a universal prevention approach: randomized controlled trial. J Med Internet Res 2021 Dec 22;23(12):e22107. [doi: 10.2196/22107] [Medline: 34941541]
19. Xia Y, Yang Y. RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: the story they tell depends on the estimation methods. Behav Res Methods 2019 Feb;51(1):409-428. [doi: 10.3758/s13428-018-1055-2] [Medline: 29869222]
20. Zarski AC, Lehr D, Berking M, Riper H, Cuijpers P, Ebert DD. Adherence to internet-based mobile-supported stress management: a pooled analysis of individual participant data from three randomized controlled trials. J Med Internet Res 2016 Jun 29;18(6):e146. [doi: 10.2196/jmir.4493] [Medline: 27357528]
21. Bijelic T, Ekberg E, Willman A, Nilsson IM. Expectations and experiences of internet-based therapy for adolescents with TMD pain. J Oral Facial Pain Headache 2022;36(3-4):237-252. [doi: 10.11607/ofph.3042] [Medline: 36445908]
22. Johansson M, Marcusson-Clavertz D, Gunnarsson C, Olsson I, Kaldo V, Bratt A. Feasibility and preliminary evaluation of internet-based compassion and cognitive-behavioral stress-management courses for health care professionals: a randomized controlled pilot trial. Internet Interv 2022;30:100574. [doi: 10.1016/j.invent.2022.100574] [Medline: 36185345]
23. Todkill D, Powell J. Participant experiences of an internet-based intervention and randomised control trial: interview study. BMC Public Health 2013 Oct 28;13(1):1017. [doi: 10.1186/1471-2458-13-1017] [Medline: 24165325]
24. Zavorotnyy M, Ehrlich F, Nenadic I. Health-related Internet use and treatment adherence: a transdiagnostic comparison of outpatients with major depressive disorder and schizophrenia. Psych J 2020 Apr;9(2):174-184. [doi: 10.1002/pchj.355] [Medline: 32189452]
25. Wu D, Lowry PB, Zhang D, Tao Y. Patient trust in physicians matters-understanding the role of a mobile patient education system and patient-physician communication in improving patient adherence behavior: field study. J Med Internet Res 2022 Dec 20;24(12):e42941. [doi: 10.2196/42941] [Medline: 36538351]

XSL•FO
RenderX

26. Weisel KK, Zarski AC, Berger T, et al. User experience and effects of an individually tailored transdiagnostic internet-based and mobile-supported intervention for anxiety disorders: mixed-methods study. J Med Internet Res 2020 Sep 16;22(9):e16450. [doi: 10.2196/16450] [Medline: 32936085]
27. Cavanagh K. Turn on, tune in and (don't) drop out: engagement, adherence, attrition, and alliance with internet-based interventions oxford guide to low intensity CBT interventions. In: Oxford Guide to Low Intensity CBT Interventions: Oxford University Press; 2010:227-234. [doi: 10.1093/med:psych/9780199590117.003.0021]
28. Coelhoso CC, Tobo PR, Lacerda SS, et al. A new mental health mobile app for well-being and stress reduction in working women: randomized controlled trial. J Med Internet Res 2019 Nov 7;21(11):e14269. [doi: 10.2196/14269] [Medline: 31697244]
29. Li J, Riedel N, Barrech A, et al. Long-term effectiveness of a stress management intervention at work: a 9-year follow-up study based on a randomized wait-list controlled trial in male managers. Biomed Res Int 2017;2017:2853813. [doi: 10.1155/2017/2853813] [Medline: 29181392]
30. Knekt P, Virtala E, Härkänen T, Vaarama M, Lehtonen J, Lindfors O. The outcome of short- and long-term psychotherapy 10 years after start of treatment. Psychol Med 2016 Apr;46(6):1175-1188. [doi: 10.1017/S0033291715002718] [Medline: 26755201]
31. Herr RM, Barrech A, Riedel N, Gündel H, Angerer P, Li J. Long-term effectiveness of stress management at work: effects of the changes in perceived stress reactivity on mental health and sleep problems seven years later. Int J Environ Res Public Health 2018 Feb 3;15(2):255. [doi: 10.3390/ijerph15020255] [Medline: 29401657]
32. Liu W, Wang Q, Zheng D, et al. The effects of a complex interactive multimodal intervention on personalized stress management among health care workers in China: nonrandomized controlled study. J Med Internet Res 2024 Jul 12;26:e45422. [doi: 10.2196/45422] [Medline: 38996333]
33. Wang Q, Collet JP, Mei J, et al. Complex interactive multimodal intervention to improve personalized stress management among healthcare workers in China: a knowledge translation protocol. Digit Health 2023;9:20552076231184052. [doi: 10.1177/20552076231184052] [Medline: 37545629]
34. O'Cathain A, Murphy E, Nicholl J. The quality of mixed methods studies in health services research. J Health Serv Res Policy 2008 Apr;13(2):92-98. [doi: 10.1258/jhsrp.2007.007074] [Medline: 18416914]
35. Mihalits DS, Schiller B, Barrech A, et al. The flipside of work engagement: a qualitative evaluation of a stress management intervention in the workplace. Hu Arenas 2023 Jun;6(2):358-372. [doi: 10.1007/s42087-021-00209-y]
36. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [doi: 10.2307/2136404] [Medline: 6668417]
37. Campbell K, Orr E, Durepos P, et al. Reflexive thematic analysis for applied qualitative health research. Qual Rep 2021;26(6):2011-2028. [doi: 10.46743/2160-3715/2021.5010]
38. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modeling 1999 Jan;6(1):1-55. [doi: 10.1080/10705519909540118]
39. Zhang M, Murphy B, Cabanilla A, Yidi C. Physical relaxation for occupational stress in healthcare workers: a systematic review and network meta-analysis of randomized controlled trials. J Occup Health 2021 Jan;63(1):e12243. [doi: 10.1002/1348-9585.12243] [Medline: 34235817]
40. Kim J, Gray JA, Johnson H. The effect of a web-based deep breathing app on stress of direct care workers: uncontrolled intervention study. J Altern Complement Med 2021 Oct;27(10):876-883. [doi: 10.1089/acm.2020.0541] [Medline: 34227854]
41. Babazadeh T, Ghaffari-Fam S, Shahnavaz-Yoshanluie F, Ranjbaran S. Psychological well-being and factors affecting it after the COVID-19 pandemic. Front Psychol 2024;15:1295774. [doi: 10.3389/fpsyg.2024.1295774] [Medline: 38390417]
42. Leone SS, Huibers MJH, Knottnerus JA, Kant I. A comparison of the course of burnout and prolonged fatigue: a 4-year prospective cohort study. J Psychosom Res 2008 Jul;65(1):31-38. [doi: 10.1016/j.jpsychores.2008.03.018] [Medline: 18582609]
43. Van Dam A, Keijsers G, Verbraak M, Eling P, Becker E. Level and appraisal of fatigue are not specific in burnout. Clin Psychol Psychother 2015;22(2):133-141. [doi: 10.1002/cpp.1869] [Medline: 24022877]
44. De Kock JH, Latham HA, Cowden RG. The mental health of healthcare workers during the COVID-19 pandemic: a narrative review. Curr Opin Psychiatry 2022 Sep 1;35(5):311-316. [doi: 10.1097/YCO.0000000000000805] [Medline: 35855506]
45. Dong F, Liu HL, Yang M, et al. Immediate psychosocial impact on healthcare workers during COVID-19 pandemic in China: a systematic review and meta-analysis. Front Psychol 2021;12:645460. [doi: 10.3389/fpsyg.2021.645460] [Medline: 34122233]
46. De Sio S, Letizia C, Petramala L, et al. Work-related stress and cortisol levels: is there an association? Results of an observational study. Eur Rev Med Pharmacol Sci 2018 Dec;22(24):9012-9017. [doi: 10.26355/eurrev_201812_16672] [Medline: 30575947]
47. Franck E, Haegdorens F, Goossens E, et al. The role of coping behavior in healthcare workers' distress and somatization during the COVID-19 pandemic. Front Psychol 2021;12:684618. [doi: 10.3389/fpsyg.2021.684618] [Medline: 34367005]
48. Pappa S, Sakkas N, Sakka E. A year in review: sleep dysfunction and psychological distress in healthcare workers during the COVID-19 pandemic. Sleep Med 2022 Mar;91:237-245. [doi: 10.1016/j.sleep.2021.07.009] [Medline: 34334303]

XSL•FO
RenderX

49. Panagioti M, Geraghty K, Johnson J, et al. Association between physician burnout and patient safety, professionalism, and patient satisfaction: a systematic review and meta-analysis. JAMA Intern Med 2018 Oct 1;178(10):1317-1331. [doi: 10.1001/jamainternmed.2018.3713] [Medline: 30193239]
50. Montgomery A, van der Doef M, Panagopoulou E, Leiter MP. Connecting Healthcare Worker Well-Being, Patient Safety and Organisational Change The Triple Challenge: Springer Cham; 2020. [doi: 10.1007/978-3-030-60998-6]
51. Bartlett L, Martin A, Neil AL, et al. A systematic review and meta-analysis of workplace mindfulness training randomized controlled trials. J Occup Health Psychol 2019 Feb;24(1):108-126. [doi: 10.1037/ocp0000146] [Medline: 30714811]
52. Zaccaro A, Piarulli A, Laurino M, et al. How breath-control can change your life: a systematic review on psycho-physiological correlates of slow breathing. Front Hum Neurosci 2018;12:353. [doi: 10.3389/fnhum.2018.00353] [Medline: 30245619]

Abbreviations

FAS: Fatigue Assessment Scale
HCW: health care worker
IRB: Institutional Review Board
MSM: mobile stress management
PSS-14: 14-item Perceived Stress Scale
REDCap: Research Electronic Data Capture

Edited by S Brini; submitted 01.Mar.2025; peer-reviewed by J Maciejewski, Q Wang, S Ranjbaran; revised version received 13.Jan.2026; accepted 16.Jan.2026; published 02.Apr.2026.

Please cite as:
Li M, Wang Q, Zheng D, Feng D, Lang X
Well-Being and Cognitive Factors Influencing Health Care Workers' Adherence to Internet-Based Stress Management: Mixed Methods Analysis of a Nonrandomized Controlled Study
J Med Internet Res 2026;28:e73306
URL: https://www.jmir.org/2026/1/e73306
doi:10.2196/73306

© Miqi Li, Quan Wang, Danli Zheng, Danni Feng, Xiaorong Lang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Goal Setting and Anchoring Effects on Meditation Using a Digital Platform: Large-Scale Digital Field Study

Michael Bowen[1,2], MS; Michael Beam[1], PhD

[1]Kent State University, 1800 E Main Street, Kent, OH, United States

[2]Spotify (Sweden), Stockholm, Sweden

Corresponding Author:
Michael Bowen, MS
Kent State University, 1800 E Main Street, Kent, OH, United States

Abstract

Background: Meditation has grown in popularity in recent years, but many people who try meditation often fail to establish a habit. Goal setting has been demonstrated to be an effective technique in behavior change in other health-related contexts but is understudied in the meditation context.

Objective: This study had 2 objectives: (1) to assess the association between goal setting and the number of days people meditated and (2) to evaluate whether anchoring bias in the goal-setting question (via response option order) influences goal selection and subsequent meditation behavior.

Methods: This large-scale quasi-experimental field study included 18,559 Spotify mobile users aged 18 years or older residing in Australia, Canada, New Zealand, the United Kingdom, or the United States who had listened to at least 5 minutes of meditation content from a specified teacher. The in-app experiment consisted of 2 goal-setting test conditions and an active control. In the test conditions, participants selected the number of days they intended to listen to content from the meditation teacher in the next 7 days. The conditions differed only in the order of goal response options (higher goals listed first vs last). The active control rated how much they liked the teacher but did not set a goal. Because responding was optional, selection bias is possible, and the design is quasi-experimental.

Results: The act of setting any goal had a modest positive association with the number of days people meditated in both treatment condition 1 (β=.08, 95% CI 0.01-0.16) and treatment condition 2 (β=.08, 95% CI 0.002-0.15). People who committed to higher goals were also more likely to meditate more than those who committed to lower goals. Additionally, the distribution of goals between the treatment conditions varied (χ^2_2=84.24; P<.001), and the differences in these distributions subsequently yielded differences in the number of days each group meditated, on average ($t_{2744.1}$=−2.34; P=.02; Cohen d=−0.09). Ultimately, placing the highest goal as the first answer choice yielded higher average active days among those who chose a goal, but many more people opted out of answering the question itself.

Conclusions: Simply offering an optional in-app goal-setting prompt, the intention-to-treat estimate from the experiment, did not change meditation engagement at the population level. However, among users who chose to respond and set a goal, goal selection was associated with a modest increase in active days. Response option order (anchoring) shifted both goal selection and opt-out rates. These findings highlight an uptake-engagement trade-off that is relevant for digital behavior change design.

(J Med Internet Res 2026;28:e85801) doi:10.2196/85801

KEYWORDS

goal setting; commitment; precommitment; anchoring; nudging; meditation; mindfulness; app; mobile

Introduction

Background

Meditation has many demonstrated health benefits [1-3], and in recent years, meditation has been increasing in popularity [4]. Many meditators use meditation apps and streaming services to support their practices [5,6]; however, many people who have tried meditation no longer meditate [6], and retention with digital meditation apps and content, in particular, is low [7-9]. Evidence suggests that greater home meditation practice frequency is associated with better outcomes, highlighting the importance of continued practice [10,11]. Developing a new habit, such as meditation, requires repeated behavior that can develop into automaticity [12-15]. One effective strategy to support this process is planning ahead and setting specific goals, typically expressed in terms of when, where, and how much one intends to engage in the behavior. Goal setting has been shown to increase engagement across a variety of domains,

XSL•FO
RenderX

including physical activity, healthy eating, and other health-promoting behaviors [16-19].

Goal setting is an important determinant of goal commitment and pursuit [20], which is a key component of many behavior change theories, both directly in social cognitive theory [21] and through related constructs, such as perceived behavioral control in the theory of planned behavior [22]. For instance, being actively involved in one's own goal setting can yield positive outcomes [20], and stating a goal publicly can enhance commitment to the goal [23], consequently increasing the belief in one's ability to achieve the goals. As such, understanding the influence of goal setting in behavior change is important to test key theories of behavioral regulation and habit formation.

Within the context of meditation, action and coping planning strategies and attitudinal commitments have been positively associated with the number of days practicing meditation [24,25]. However, explicit goal setting and goal-intention effects in meditation are understudied, particularly through rigorous techniques. This knowledge gap is especially noteworthy because of the role that goal setting can play in prominent behavior change theories (eg, social cognitive theory, theory of planned behavior). Additionally, goal setting is associated with behavioral outcomes, and practice frequency is crucial for realizing the mental health benefits of meditation, including benefits from using meditation apps and digital content [26-30].

When setting goals, the measurement interface used to elicit the goal is also an important and understudied component of the experience, particularly in the context of meditation. Cognitive psychology has demonstrated the influence of anchoring. For example, the starting points and positions of choices and the subsequent effects these have on people's judgments and choices [31]. Along these lines, in surveys and survey question design, anchoring biases are a well-known phenomenon that can impact the response distribution [32-34]. In fact, people even rate their success in life differently in surveys when the numeric values on a scale differ [35]. These effects may be explained by dual-process models of judgment, in which people rely on fast, heuristic processing when making quick decisions [31]. In the case of goal commitment within a mobile app, it is plausible that individuals "satisfice" by selecting the first option they come across that seems reasonable, thus making response order a plausible driver of goal commitment. From the perspective of goal-setting theory, anchors that shift potential goals may also change perceived goal difficulty and feasibility, which may influence commitment, self-efficacy, and intention formation [20].

When the objective of asking a question is not explicitly to gather unbiased data for research purposes but rather to encourage people to meditate more, standard survey best practices, such as choice randomization, may be less important than optimizing the goal-setting intervention to outcomes. In a way, these techniques are commonly found in push polls, which are not scientific surveys but rather political persuasion techniques framed as surveys. These kinds of techniques, in which questions have an overt design bias, have been demonstrated to be effective for shifting both memory and opinions [36,37]. These kinds of techniques have also been used in the health behavior goal-setting context. For example, introducing anchoring has demonstrated that people set higher goals when questions are framed as sacrifices (what to forgo) rather than actions (what to undertake) [38]. Consequently, it is important for theorists, applied researchers, and user experience (UX) designers to further understand the effects of anchoring in goal setting, including in the context of meditation. By doing so, these insights can be leveraged for positive impacts on people's mental health.

Research Purpose

With these 2 streams of literature in mind, the purpose of this research was twofold. The first and primary objective was to identify the behavioral association between goal setting and engaging with meditation content through a large-scale digital platform. In doing so, we assessed whether people who preemptively set a goal are more likely to meditate than people who did not select a meditation goal. This objective resulted in the following hypothesis:

- H1: People in the treatment conditions who preemptively state a meditation goal will have more days listening to guided meditation teachings during the 7 days following the intervention, on average, compared to the control group.

The second objective was to explore the effects of anchoring on goal intentions, specifically, whether anchoring different goals in the goal-setting question changes the ways in which people set their goals and subsequently how the anchoring effect impacts meditation behaviors. This objective resulted in the following research hypotheses:

- H2: The effects of anchoring preemptively stated meditation goals in the goal-setting question will result in a different distribution of goal intentions.
- H3: Among people who preemptively state a goal, the effects of anchoring different meditation goals in the goal-setting question will result in varying levels of active days with the meditation content during the 7 days following the intervention.

By exploring these hypotheses, this research can support meditation and mental health app product managers in considering whether to implement goal-setting features in their UX and to provide guidance on how to design the feature. Additionally, the present study can help future researchers who study meditation habit formation, particularly with digital health platforms, understand the role of goal setting within the broader context of behavior change.

Methods

Sampling and Recruitment

This research was conducted among 18,559 users of the Spotify mobile app. To qualify to participate in the study, individuals needed to have listened to at least 5 minutes of content provided by a popular meditation teacher through the Spotify app during the month of June 2025. The content offered by the meditation teacher consisted of both guided meditations and teachings. Additional qualifications included being aged 18 years or older and residing in the United States, United Kingdom, Canada,

XSL•FO
RenderX

Australia, or New Zealand. These countries were chosen because Spotify is available in these countries, the meditation content is in English, and these countries are predominantly English-speaking. Fieldwork was conducted from July 2, 2025, to July 8, 2025. Table 2 in the Results section includes the total sample sizes, country, age, and gender composition of the research sample for each experimental condition. Note that chi-square tests indicated no differences between cells based on country, age, and gender identification.

As the primary experimental intervention, qualified users were assigned to receive 1 of 3 potential one-tap survey questions. Two of the survey questions were treatment conditions, and one was the active control condition. The one-tap survey questions were embedded within a pop-up message in the Spotify mobile app that appeared when a qualified research participant visited the Spotify homepage. Participants could choose to answer the question or dismiss the message without selecting an answer option. Additionally, participants could select a "Prefer not to say" response option if they did not want to select a meditation goal. Upon answering the question, the message was subsequently dismissed automatically, thus ending the intervention.

Because the survey question was optional and, subsequently, goal intention was dependent on a post random assignment event, there was a selection bias in the design when analyzing only people who clicked on the message itself. We detail how this was handled and the implications on the results and interpretation in the Data Analysis Plan section.

Meditation Content

The meditation content evaluated in this study consisted of a large, publicly available audio catalog produced by a single teacher and distributed as a long-running mindfulness podcast. The catalog includes both (1) guided meditation sessions and (2) didactic "teachings" that provide instruction and reflection intended to support practice. Content primarily reflects mindfulness-based approaches and frequently incorporates self-compassion–oriented instruction (eg, attention to breath/body, present-moment awareness, and working with difficult emotions). Episodes vary in length but tend to be between 20 and 90 minutes. The catalog spans more than a decade and contains over 1500 audio items, often with multiple new items released per week.

Experimental Design

Each test condition comprised a single survey question. The 2 treatment conditions were variants of the goal-setting question. Specifically, each goal-setting message requested the participant to select the number of days in the next 7 days they intended to listen to guided meditations and teachings offered by the teacher on Spotify. The variants were nearly identical, with only the question response options scale being flipped, which was designed to test the anchoring effect of having the highest vs the lowest number of days response option at the top of the response battery.

Specifically, treatment 1 asked how many days in the next 7 days the user intended to listen to meditation and teachings offered by the teacher on Spotify. Seven days were chosen because goal-setting theory argues that proximal (short-term) goals can facilitate the performance of complex tasks [20], of which meditation could reasonably be considered. The question response options were listed from top to bottom as "0 days," "1 day," "2-3 days," "4 days or more," and "Prefer not to say." In treatment 2, the question itself was the same, but the options were reversed, so from top to bottom, the options were listed as "4 days or more," "2-3 days," "1 day," "0 days," and "Prefer not to say." Note that selecting "Prefer not to say" was offered as a survey response and was not the same interaction as dismissing the message without selecting any response. However, selecting "Prefer not to say" was not treated as having selected a goal.

The control condition asked participants how they felt about the meditation and teachings offered by the same meditation teacher on Spotify, with the response options listed from top to bottom as "I love them," "I like them," "I have no opinion," "I dislike them," and "Prefer not to say." In using this type of active control, we ensured that the treatment and control conditions had similar interventions, thus equalizing the cognitive effort of responding to a question, controlling for question-behavior effects (where answering any question can influence subsequent behavior), and avoiding inadvertently inducing goal-setting or planning cognitions.

See Table 1 for sample sizes based on participants' interactions with the goal-setting question.

Table . Sample sizes based on message interactions.

	Saw the message (all participants), n (%)	Survey responders[a], n (%)	Selected a goal[b], n (%)
Control	6012 (100)	2428 (40.4)	2249 (37.4)
Treatment 1	6320 (100)	2065 (32.7)	1542 (24.4)
Treatment 2	6227 (100)	2155 (34.6)	1349 (21.7)

[a]Participants who did not dismiss the message and chose an answer to the one-tap survey question.

[b]Participants who did not dismiss the message and did not select "Prefer not to say" as their response.

Following this brief survey-based intervention, we used trace behavioral data from the Spotify app to assess how often users streamed the specific content offered by the respective meditation teacher on the Spotify app.

Active Days Variable

We define active days as the number of days a participant streamed content from the meditation teacher on Spotify. The

XSL•FO
RenderX

postintervention active-days measure (days 0‐7 after the intervention) is the dependent variable for H1 and H3. The preintervention active-days measure (the 30 days preceding the intervention) is also included as an independent covariate. To be considered active on a given day, the participant needed to have streamed content from the teacher for any period of time on the respective day (having pressed play and streamed the content for >0 ms). While this is not necessarily indicative of a meditation session having occurred, it avoids otherwise arbitrary definitions of an active day and is a strong indicator of intent. Summary statistics by cohort are shown in Table 3 in the Results section. In all models, both measures were treated as integer counts.

Data Analysis Plan

The analysis for this study was conducted using Python. The click-through rate of the survey question (the percentage of people who saw the survey question and subsequently chose a response option, as opposed to dismissing it without responding) for each of treatment 1, treatment 2, and the control was 32.7% (2065/6320), 34.6% (2155/6227), and 40.4% (2428/6012), respectively. A chi-square test indicates that these click-through rates are different (χ^2_3=85.71; P<.001). As a result, there is a selection bias introduced into the experimental design. So, while the overall study is a natural field experiment, the analysis of the effects of goal setting is quasi-experimental and should be considered associations.

To test H1, specifically whether people in the treatment conditions who preemptively state a meditation goal will have more days listening to guided meditation teachings during the 7 days following the intervention, on average, compared to the control, a linear regression model was used. The dependent variable in the model was the number of active days on which the participant listened to content from the meditation teacher through the Spotify app during the 7 days after the intervention. While active days is a bounded and nonnormally distributed variable, linear regression is appropriate because it provides a parsimonious and interpretable approximation for mean differences and has been demonstrated to be valid in these conditions when sample sizes are large [39]. This was coded as an integer, with values ranging from 0 to 8. The range is 0 to 8, not 0 to 7, because the day of the intervention, at any point after the question was answered, is included in the count as day 0. Each treatment condition was inserted separately as an independent variable in the model, while the control condition was held out as the reference category. Note that only people who selected a goal were included in the model. To control for past engagement with the meditation content, the number of active days during the 30 days preceding the treatment that the participant listened to meditation content from the meditation teacher through the Spotify app was included as an independent variable. Age was binned into 5 groups and one-hot encoded as 18 to 24, 25 to 34, 35 to 44, 45 to 54, and aged 55 years or older. The 55 years and older group was held out of the model as the reference category. Gender values included female, male, and other and were also one-hot encoded for modeling. Female was held out as the reference category. Country was also inserted into the model, with the United States held out as the reference category.

H2 explored the role of anchoring on the response distributions of goals chosen. To test this hypothesis, we explored the percentage of users who selected each response option in the survey and applied a chi-square test of independence to specifically assess whether these distributions varied.

To test H3, specifically that the anchoring effect between treatment cells would result in different levels of active days between treatment conditions, a 2-sample Welch t test was used to assess differences in postintervention active days.

Ethical Considerations

The study was reviewed and approved as exempt by the Kent State University Institutional Review Board (IRB) under reference number 2145. The consent process was governed by Spotify's Terms of Service and Privacy Policy, as the experiment was conducted on the platform with its members. The research team and the IRB carefully reviewed these policies and the ethical conduct of the study, concluding that consent was obtained through the Terms of Service and Privacy Policy. The IRB also determined that this study qualifies as a benign behavioral intervention. Personally identifiable information about Spotify users was not accessible to the researchers and was safeguarded through encryption. It can only be accessed through a review and approval process by the Spotify Data Protection Office. No compensation for participation was provided.

Results

Table 2 provides the demographic composition of the sample. As previously noted, chi-square tests indicated no differences between cells based on country, age, and gender identification.

Table 3 provides the descriptive statistics for the number of active days before and after the treatment based on the varying levels of engagement with the message.

XSL•FO
RenderX

Table . Demographic composition of experimental conditions.

	Control (N=6012), n (%)	Treatment 1 (N=6320), n (%)	Treatment 2 (N=6227), n (%)
Country			
Australia	631 (10.5)	653 (10.3)	632 (10.1)
Canada	756 (12.6)	767 (12.1)	787 (12.6)
United Kingdom	819 (13.6)	897 (14.2)	917 (14.7)
United States	3684 (61.3)	3867 (61.2)	3733 (59.9)
New Zealand	122 (2)	136 (2.2)	158 (2.5)
Age (years)			
18-24	234 (3.9)	233 (3.7)	244 (3.9)
25-34	1520 (25.3)	1570 (24.8)	1520 (24.4)
35-44	1898 (31.6)	1975 (31.3)	1935 (31.1)
45-54	1223 (20.3)	1272 (20.1)	1275 (20.5)
≥55	1137 (18.9)	1270 (20.1)	1253 (20.1)
Gender identification[a]			
Female	4453 (74.1)	4601 (72.8)	4605 (74)
Male	1257 (20.9)	1394 (22.1)	1344 (21.6)
Other/prefer not to say	59 (1)	72 (1.1)	53 (1)

[a]Gender data is collected at the time of Spotify registration. It is provided to Spotify optionally and is not available for all participants.

Table . Summary statistics for "active days."

Time window	Role in models	Cohort	Mean (SD)
7 days postintervention	Dependent (H1, H3)	All participants	0.95 (1.5)
7 days postintervention	Dependent (H1, H3)	Survey responders[a]	1.11 (1.62)
7 days postintervention	Dependent (H1, H3)	Selected a goal[b]	1.18 (1.66)
30 days preintervention	Independent covariate	All participants	3.93 (4.31)
30 days preintervention	Independent covariate	Survey responders[a]	4.27 (4.52)
30 days preintervention	Independent covariate	Selected a goal[b]	4.49 (4.66)

[a]Participants who did not dismiss the message and chose an answer to the one-tap survey question.

[b]Participants who did not dismiss the message and did not select "Prefer not to say" as their response.

Before testing the specific hypotheses listed previously, we first explored whether simple exposure to the commitment message (ie, participants saw the message) affected the likelihood of participants in the control versus treatment conditions to have more active days following the intervention. This could be considered the intention-to-treat effect from the messages and the results of the true field experiment component of the research. A Welch *t* test showed that the overall mean number of active days between the control condition (mean 0.94, SD 1.48 days) and the treatment conditions combined (mean 0.96, SD 1.51 days) was not different ($t_{12,045.6}$=−0.76; *P*=.45).

In the linear regression used to test H1, participants in both treatment condition 1 (β=.08, 95% CI .01-.16) and treatment condition 2 (β=.08, 95% CI .002-.15) had more average active days with the meditation content compared to the control group. The model controlled for preintervention activity levels, as well as the specific goals and answer choices participants selected, among others, indicating that simply selecting a goal is associated with increased active days. Therefore, we can reject the null hypothesis for H1.

Additionally, people who set higher goals for themselves tended to have more active days on average. The adjusted R^2 for this model was 0.396. See Table 4 for more details on the linear regression model of days active after the intervention.

XSL•FO
RenderX

Table . Coefficient, upper and lower bound CI (95%), and *P* value from linear regression of posttreatment days active.

Variable	Coefficient (95% CI)	*P* value
Constant	0.19 (0.08 to 0.30)	<.001
Preexposure days active	0.21 (0.20 to 0.22)	<.001
Experimental condition		
Control	—[a]	—
Treatment 1	0.08 (0.01 to 0.16)	.03
Treatment 2	0.08 (0.002 to 0.15)	.05
Country		
Australia	0.09 (−0.04 to 0.21)	.17
Canada	0.09 (−0.03 to 0.21)	.13
United Kingdom	0.05 (−0.06 to 0.16)	.34
United States	—	—
New Zealand	0.28 (0.06 to 0.50)	.01
Gender identification		
Female	—	—
Male	0.08 (0 to 0.17)	.06
Age (y)		
18 - 24	−0.36 (−0.54 to −0.19)	<.001
25 - 34	−0.22 (−0.33 to −0.11)	<.001
35 - 44	−0.22 (−0.33 to −0.11)	<.001
45 - 54	−0.07 (−0.19 to 0.04)	.22
55 or older	—	—
Goal intention survey response		
0 days	−0.21 (−0.29 to −0.14)	<.001
1 day	−0.05 (−0.12 to 0.03)	.23
2 - 3 days	0.08 (0 to 0.16)	.04
4 days or more	0.34 (0.23 to 0.44)	<.001
Prefer not to say	—	—
Control survey response		
I dislike them	0.01 (−0.34 to 0.37)	.94
I have no opinion	−0.10 (−0.29 to 0.09)	.30
I like them	−0.02 (−0.15 to 0.10)	.70
I love them	0.14 (0.05 to 0.23)	.003
Prefer not to say	—	—

[a]Not applicable.

To test whether the ordering of the goal response options created an anchoring effect that influenced the distribution of responses in H2, a chi-square test of independence was used. Figure 1 includes the distribution of responses from the survey within the 2 test conditions. The test yielded (χ^2_2=84.24; *P*<.001), indicating significant differences in the distribution of responses across treatment conditions. When the "Prefer not to say" response is removed from the distribution and the chi-square statistic recalculated, the result yielded (χ^2_2=13.03; *P*=.005), indicating that the anchoring effect persists. As a result, we can reject the null hypothesis for H2.

XSL•FO
RenderX

Figure 1. Distribution of goal intention one-tap survey responses based on treatment condition (percentage of users who chose their respective meditation goal by treatment condition).

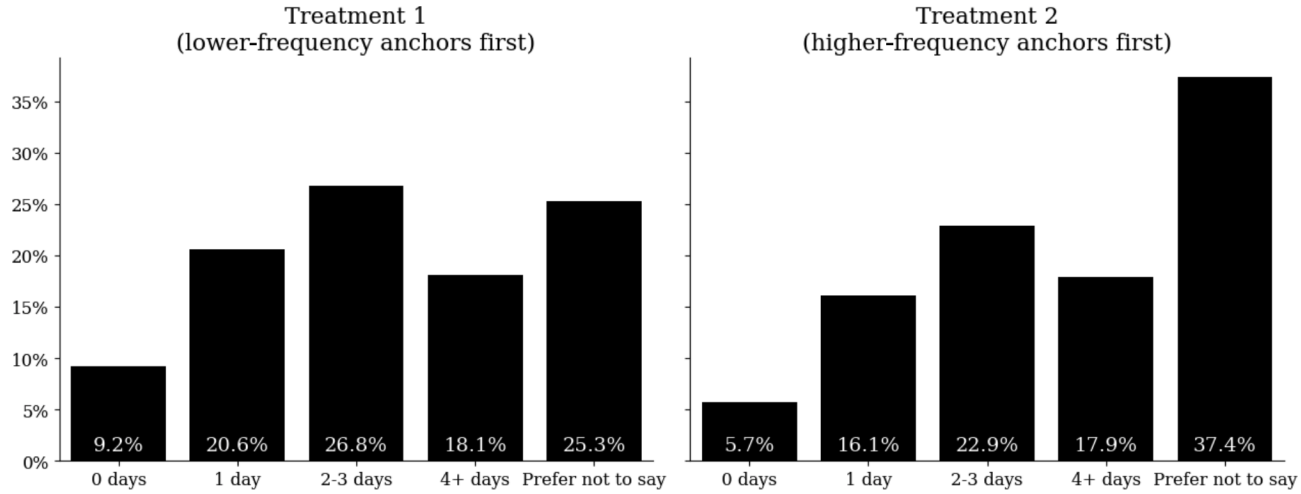

In H3, the Welch *t* test showed that the mean number of active days differed between treatment 1 (mean 1.15, SD 1.62 days) and treatment 2 (mean 1.3, SD 1.78 days; $t_{2744.1}$=−2.34; *P*=.02). A Cohen *d* of −0.09 indicates a modest overall effect of the differences in days active between treatment conditions. As a result, we can reject the null hypothesis for H3. It is important to note, however, that when we included people who did not select a goal (ie, instead they selected "Prefer not to say"), the effect is not significant ($t_{4217.7}$=−0.68; *P*=.50).

Table 5 provides a summary of the hypotheses, test results, and decisions. Additionally, a CONSORT (Consolidated Standards of Reporting Trials) diagram of the full experimental setup is available in Multimedia Appendix 1.

Table . Summary of hypotheses, test results, and decisions.

Hypothesis	Hypothesis (abridged)	Test (*P* value)	Decision
H1	People who preemptively state a meditation goal will have more days listening to guided meditation teachings during the 7 days following the intervention, on average, compared to the control.	Participants who stated a meditation goal were more likely to engage with the meditation content in both Treatment 1 (β=.08, 95% CI 0.01-0.16) and Treatment 2 (β=.08, 95% CI 0.002-0.15), compared to the control.	Reject null hypothesis
H2	The effects of anchoring different meditation goals in the treatment conditions will result in a different distribution of goal intentions.	χ^2 (2) = 84.24 (*P*<.001)	Reject null hypothesis
H3	The effects of anchoring different meditation goals in the goal-setting question will result in varying levels of active days with the meditation content during the 7 days following the intervention.	*t* (2744.1) = −2.34 (*P*=.02)	Reject null hypothesis

Discussion

Principal Results

This study demonstrated that setting short-term goals for guided meditation and teachings through a mobile app was associated with engaging with the meditation content on more days. Specifically, people who preemptively set themselves any goal were more likely to meditate on more days than people in the control condition. Additionally, people who chose higher goals were more likely to engage with meditation content more often, even after controlling for past engagement with the meditation content, as well as age, gender, and country of residence. Conversely, people who selected zero as a goal meditated less often. While the differences were modest, these uptake-conditional results suggest that among participants who elected to respond, goal selection was associated with slightly more active days. The intention-to-treat analysis indicated that simply offering the optional goal-setting prompt (thus including people who chose to ignore the message) did not influence subsequent meditation behavior.

This finding adds to the body of literature that goal setting may be useful in health behavior change in the context of meditation, similar to other health behavior contexts such as physical activity and nutrition [16-19]. Like Epton et al [16] found in their meta-analysis of the effects of goal setting across a range of behaviors, the effect of goal setting from this research was found to be small and should subsequently be considered just one, perhaps small, component of meditation habit formation.

XSL•FO
RenderX

However, while the overall effects may be small, at scale, these kinds of cost-effective nudges can be worthwhile interventions [40]. For meditation content providers and program managers, these findings provide evidence that goal-setting interventions may be useful, if modest, tools to increase engagement with meditation content. Ultimately, for anyone who has a desire to develop a meditation habit, this study provides compelling data that goal setting may be worthwhile in building and maintaining a meditation habit.

This study also provides clear evidence of an anchoring effect based on how the goal-intention question is structured. Specifically, when higher goal response options are anchored at the top of the response battery, the response distributions change substantively. Most notably, many more people selected "Prefer not to say" when the "4+ days per week" option was anchored at the top of the response battery compared to the bottom (37.4% vs 25.3%). This is likely explained by dual-process models of judgment, in which individuals rely on quick, heuristic processing when making low-effort decisions [31]. Within this study environment, individuals likely chose the first plausible goal. Seeing the most ambitious goal at the top may have felt unrealistic or hindered self-efficacy for some people, both of which are important components of goal setting and goal accomplishment [20]. Additionally, people who were in the Treatment 2 condition and selected a meditation goal (ie, did not select "Prefer not to say") subsequently engaged with meditation slightly more than people in Treatment 1 who selected a goal, suggesting that the anchoring effect may not have only affected the response distribution but also had a modest effect on people's propensity to meditate.

These findings indicate that there are trade-offs to anchoring. While it is best practice in survey research to randomize response options or randomly flip ordered scales in survey design [32-34], practitioners and UX designers should consider the outcomes they want to achieve when choosing anchors and whether to flip scales. Push polls are ethically fraught cautionary tales of the impact of asking biased questions [36,37]. However, in the context of health behavior, this technique may lead people to set more ambitious goals [38].

Overall, this study provides theoretical insights about the role of goal setting in meditation habit development and maintenance, adding to the body of literature about goal setting in different contexts. It also offers new insights into the effects of anchoring on goal setting within the meditation context. Finally, the research provides UX designers and product managers practical guidance on the potential impact of goal setting and anchoring on user engagement.

Limitations

There are several limitations to this research that should be considered. First, as described previously, because people could opt out of participating in the one-tap survey, there is a self-selection bias that makes this research a quasi-experimental design and not a full field experiment. Consequently, we cannot confidently conclude that we have established a causal association between goal setting and incremental meditation. While this is an important limitation in drawing conclusions, the nature of the research itself reflects real-world environments and digital platform experiences.

Second, this study was conducted on a single digital platform (Spotify) and includes only a single meditation teacher, which limits the broader generalizability of the findings. Spotify may be unique in that it primarily offers music, podcast, and audiobook content and is not broadly considered a digital mental health services provider. As such, goal-setting exercises on platforms that specialize in digital mental health services may differ. Additionally, the research includes only log behavioral data and does not include important context, such as motivations for meditation and meditation experience. Expanding this research to additional platforms and teachers, to more segments of the meditation population (eg, more experienced meditators who already have strong self-efficacy vs less experienced meditators who need structure), and with additional context (eg, mapping motivations to goal-setting outcomes) could be considered by future researchers.

Additionally, while the single-tap question is a strength of the study because it optimizes response rates and limits nonresponse bias, a single question served as a pop-up has formatting constraints. Consequently, we could not offer more detailed response options and could not probe more deeply into participants' relationship with meditation. Data of this kind may have been helpful to add more context to the insights offered from the study.

Future Directions

The results from this study open up additional avenues for exploration for future researchers. First, future studies may strive to make the research a fully experimental design by making the goal-setting question itself mandatory. This would allow researchers to better understand the full causal effects of goal setting within the meditation technology context. Additionally, as described, it would also be useful to study this phenomenon on other apps and in other contexts, which would allow researchers to generalize the effects of goal setting on meditation more broadly.

This research also looked only at the average effect of the treatment across experimental conditions. Future researchers might further explore how goal setting impacts different cohorts of meditators, such as lighter versus heavier meditators or people with a burgeoning versus an established practice. Doing so would allow theorists to better understand the effects of goal setting throughout the stages of meditation habit formation and for UX designers and product managers to strategically target their goal-setting user growth initiatives.

Conclusion

This research provides insights about the role of goal setting in the meditation context when applied to digital health platforms. Ultimately, goal setting may be a useful and cost-effective tool when applied on a large scale to encourage people to engage with meditation more frequently on digital platforms, although the overall impact may be modest. However, anchoring effects play a significant role in people's willingness to set meditation goals, the goals they set for themselves, and even their incremental meditation engagement. Researchers who study

XSL•FO
RenderX

the theoretical aspects of meditation habit formation can use this research to consider the role of goal setting in their theories and models. Clinical practitioners and UX designers should consider explicit goal setting when encouraging people to practice meditation but think carefully about the effects of anchoring in their designs.

Acknowledgments

This study would not have been possible without the support of Spotify. ChatGPT (GPT-5) was used to help refine the language in some sections of this work in an editorial manner after the original draft was written. The authors reviewed and edited all content and accept full responsibility for the final publication.

Funding

This publication was made possible in part by support from the Kent State University Open Access Publishing Fund and the KSU CCI Research and Creative Fund.

Data Availability

Aggregated outputs used in the manuscript are included in the paper; requests for additional access can be directed to mtbowen@kent.edu. However, approval is unlikely due to contractual and privacy constraints.

Authors' Contributions

Conceptualization: M Bowen
Data curation: M Bowen
Formal analysis: M Bowen
Funding acquisition: M Bowen
Methodology: M Bowen
Project administration: M Bowen
Resources: M Bowen
Software: M Bowen
Supervision: M Beam
Validation: both authors
Visualization: M Bowen
Writing – original draft: M Bowen
Writing – review & editing: both authors

Conflicts of Interest

M Bowen is employed by Spotify, which is a streaming service that primarily offers music, podcasts, and audiobook content. Some content on Spotify is meditation-related. However, the salary and equity of the author are not dependent upon the results of the research. M Beam declares no conflicts of interest.

Multimedia Appendix 1
CONSORT diagram of experimental design and outcomes.
[PNG File, 211 KB - jmir_v28i1e85801_app1.png]

References

1. Galante J, Friedrich C, Dawson AF, et al. Mindfulness-based programmes for mental health promotion in adults in nonclinical settings: a systematic review and meta-analysis of randomised controlled trials. PLoS Med 2021 Jan;18(1):e1003481. [doi: 10.1371/journal.pmed.1003481] [Medline: 33428616]
2. Goldberg SB, Tucker RP, Greene PA, et al. Mindfulness-based interventions for psychiatric disorders: a systematic review and meta-analysis. Clin Psychol Rev 2018 Feb;59:52-60. [doi: 10.1016/j.cpr.2017.10.011] [Medline: 29126747]
3. Luberto CM, Shinday N, Song R, et al. A systematic review and meta-analysis of the effects of meditation on empathy, compassion, and prosocial behaviors. Mindfulness (N Y) 2018 Jun;9(3):708-724. [doi: 10.1007/s12671-017-0841-8] [Medline: 30100929]
4. Davies JN, Faschinger A, Galante J, Van Dam NT. Prevalence and 20-year trends in meditation, yoga, guided imagery and progressive relaxation use among US adults from 2002 to 2022. Sci Rep 2024 Jul 1;14(1):14987. [doi: 10.1038/s41598-024-64562-y] [Medline: 38951149]
5. Lam SU, Xie Q, Goldberg SB. Situating meditation apps within the ecosystem of meditation practice: population-based survey study. JMIR Ment Health 2023 Apr 28;10:e43565. [doi: 10.2196/43565] [Medline: 37115618]

XSL•FO
RenderX

6. Bowen M, Beam M. Mapping mindfulness: assessing the stages of meditation habit formation in the USA using the Sussex Mindfulness Meditation (SuMMed) model. Mindfulness (N Y) 2025 May;16(5):1340-1351. [doi: 10.1007/s12671-025-02555-2]
7. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567. [doi: 10.2196/14567] [Medline: 31573916]
8. Kozlov E, Bantum E, Pagano I, et al. The reach, use, and impact of a free mHealth mindfulness app in the general population: mobile data analysis. JMIR Ment Health 2020 Nov 27;7(11):e23377. [doi: 10.2196/23377] [Medline: 33245289]
9. Bowen M, Beam M, Semb J, Yoo DW. Evaluating theory-driven messaging to overcome the barriers to meditation: large-scale digital field experiment. JMIR Ment Health 2025 Jul 7;12:e71732. [doi: 10.2196/71732] [Medline: 40698564]
10. Parsons CE, Crane C, Parsons LJ, Fjorback LO, Kuyken W. Home practice in mindfulness-based cognitive therapy and mindfulness-based stress reduction: a systematic review and meta-analysis of participants' mindfulness practice and its association with outcomes. Behav Res Ther 2017 Aug;95:29-41. [doi: 10.1016/j.brat.2017.05.004] [Medline: 28527330]
11. Crane C, Crane RS, Eames C, et al. The effects of amount of home meditation practice in mindfulness based cognitive therapy on hazard of relapse to depression in the Staying Well after Depression Trial. Behav Res Ther 2014 Dec;63:17-24. [doi: 10.1016/j.brat.2014.08.015] [Medline: 25261599]
12. Wood W, Neal DT. Healthy through habit: interventions for initiating & maintaining health behavior change. Behav Sci Policy 2016 Apr;2(1):71-83. [doi: 10.1177/237946151600200109]
13. Gardner B. A review and analysis of the use of "habit" in understanding, predicting and influencing health-related behaviour. Health Psychol Rev 2015;9(3):277-295. [doi: 10.1080/17437199.2013.876238] [Medline: 25207647]
14. Gardner B, Lally P. Modelling habit formation and its determinants. In: The Psychology of Habit: Theory, Mechanisms, Change, and Contexts: Springer International Publishing; 2018:207-229. [doi: 10.1007/978-3-319-97529-0_12]
15. Miles E, Matcham F, Strauss C, Cavanagh K. Making mindfulness meditation a healthy habit. Mindfulness (N Y) 2023 Dec;14(12):2988-3005. [doi: 10.1007/s12671-023-02258-6]
16. Epton T, Currie S, Armitage CJ. Unique effects of setting goals on behavior change: systematic review and meta-analysis. J Consult Clin Psychol 2017 Dec;85(12):1182-1198. [doi: 10.1037/ccp0000260] [Medline: 29189034]
17. da Silva MAV, São-João TM, Brizon VC, Franco DH, Mialhe FL. Impact of implementation intentions on physical activity practice in adults: a systematic review and meta-analysis of randomized clinical trials. PLoS One 2018;13(11):e0206294. [doi: 10.1371/journal.pone.0206294] [Medline: 30427874]
18. Adriaanse MA, Vinkers CDW, De Ridder DTD, Hox JJ, De Wit JBF. Do implementation intentions help to eat a healthy diet? A systematic review and meta-analysis of the empirical evidence. Appetite 2011 Feb;56(1):183-193. [doi: 10.1016/j.appet.2010.10.012] [Medline: 21056605]
19. Swann C, Hooper A, Schweickle MJ, et al. Comparing the effects of goal types in a walking session with healthy adults: preliminary evidence for open goals in physical activity. Psychol Sport Exerc 2020 Mar;47:101475. [doi: 10.1016/j.psychsport.2019.01.003]
20. Locke EA, Latham GP. Building a practically useful theory of goal setting and task motivation: a 35-year odyssey. Am Psychol 2002 Sep;57(9):705-717. [doi: 10.1037/0003-066X.57.9.705]
21. Bandura A. Health promotion from the perspective of social cognitive theory. Psychol Health 1998 Jul;13(4):623-649. [doi: 10.1080/08870449808407422]
22. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
23. Hollenbeck JR, Williams CR, Klein HJ. An empirical examination of the antecedents of commitment to difficult goals. J Appl Psychol 1989;74(1):18-23. [doi: 10.1037//0021-9010.74.1.18]
24. Cloughesy JN, Mrazek AJ, Mrazek MD, Schooler JW. Planning to practice: action and coping plans increase days of meditation practiced. Psi Chi J Psychol Res 2020;25(2):203-209. [doi: 10.24839/2325-7342.JN25.2.203]
25. Galla BM, Baelen RN, Duckworth AL, Baime MJ. Mindfulness, meet self-regulation: boosting out-of-class meditation practice with brief action plans. Motiv Sci 2016;2(4):220-237. [doi: 10.1037/mot0000045]
26. Cearns M, Clark SR. The effects of dose, practice habits, and objects of focus on digital meditation effectiveness and adherence: longitudinal study of 280,000 digital meditation sessions across 103 countries. J Med Internet Res 2023 Sep 19;25:e43358. [doi: 10.2196/43358] [Medline: 37725801]
27. Stecher C, Berardi V, Fowers R, Christ J, Chung Y, Huberty J. Identifying app-based meditation habits and the associated mental health benefits: longitudinal observational study. J Med Internet Res 2021 Nov 4;23(11):e27282. [doi: 10.2196/27282] [Medline: 34734826]
28. Zawadzki MJ, Torok ZA, Peña M, Gavrilova L. App-based mindfulness meditation reduces stress in novice meditators: a randomized controlled trial of headspace using ecological momentary assessment. Ann Behav Med 2025 Jan 4;59(1):kaaf025. [doi: 10.1093/abm/kaaf025] [Medline: 40257119]
29. Bostock S, Crosswell AD, Prather AA, Steptoe A. Mindfulness on-the-go: effects of a mindfulness meditation app on work stress and well-being. J Occup Health Psychol 2019 Feb;24(1):127-138. [doi: 10.1037/ocp0000118] [Medline: 29723001]
30. Goleman D, Davidson RJ. Altered Traits: Science Reveals How Meditation Changes Your Mind, Brain, and Body: Avery; 2017.

XSL•FO
RenderX

31. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. Science 1974 Sep 27;185(4157):1124-1131. [doi: 10.1126/science.185.4157.1124] [Medline: 17835457]
32. Krosnick JA, Alwin DF. An evaluation of a cognitive theory of response-order effects in survey measurement. Public Opin Q 1987;51(2):201-219. [doi: 10.1086/269029]
33. Holbrook AL, Krosnick JA, Moore D, Tourangeau R. Response order effects in dichotomous categorical questions presented orally: the impact of question and respondent attributes. Public Opin Q 2007;71(3):325-348. [doi: 10.1093/poq/nfm024]
34. Höhne J, Lenzner T. Investigating response order effects in web surveys using eye tracking. Psihologija 2015;48(4):361-377. [doi: 10.2298/PSI1504361H]
35. Schwarz N, Knauper B, Hippler HJ, Noelle-Neumann E, Clark L. Rating scales: numeric values may change the meaning of scale labels. Public Opin Q 1991;55(4):570-582. [doi: 10.1086/269282]
36. Murphy G, Lynch L, Loftus E, Egan R. Push polls increase false memories for fake news stories. Memory 2021 Jul;29(6):693-707. [doi: 10.1080/09658211.2021.1934033] [Medline: 34080495]
37. Erčulj V, Šulc A. Swayed by leading questions. Qual Quant 2025 Feb;59(1):191-209. [doi: 10.1007/s11135-024-01934-6]
38. Tuk MA, Prokopec S, Van den Bergh B. Do versus don't: the impact of framing on goal-level setting. J Consum Res 2021 Mar 20;47(6):1003-1024. [doi: 10.1093/jcr/ucaa050]
39. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health 2002;23(1):151-169. [doi: 10.1146/annurev.publhealth.23.100901.140546] [Medline: 11910059]
40. Athey S, Grabarz K, Luca M, Wernerfelt N. Digital public health interventions at scale: the impact of social media advertising on beliefs and outcomes related to COVID vaccines. Proc Natl Acad Sci U S A 2023 Jan 31;120(5):e2208110120. [doi: 10.1073/pnas.2208110120] [Medline: 36701366]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
IRB: institutional review board
UX: user experience

Edited by A Stone; submitted 13.Oct.2025; peer-reviewed by O Akhadelor, O Akinpeloye, T Tanimoto; revised version received 09.Mar.2026; accepted 09.Mar.2026; published 09.Apr.2026.

Please cite as:
Bowen M, Beam M
Goal Setting and Anchoring Effects on Meditation Using a Digital Platform: Large-Scale Digital Field Study
J Med Internet Res 2026;28:e85801
URL: https://www.jmir.org/2026/1/e85801
doi:10.2196/85801

© Michael Bowen, Michael Beam. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Bidirectionality Between Perceived Immediate and Long-Term Benefits and Losses and Internet Gaming Disorder Among Chinese Adolescent Gamers: Prospective Longitudinal Study

Siman Li[1], BMed; Jianxin Zhang[2], BMed; Ji-bin Li[3], PhD; Joseph TF Lau[4,5*], PhD; Yanqiu Yu[1*], PhD

[1]Department of Preventive Medicine and Health Education, School of Public Health, Fudan University, Dong'an Road 130, Xuhui District, Shanghai, China

[2]West China School of Public Health, Sichuan University, Chengdu, Sichuan, China

[3]Department of Clinical Research, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangdong, Guangdong, China

[4]Public Mental Health Center, School of Mental Health, Wenzhou Medical University, Wenzhou, China

[5]Zhejiang Provincial Clinical Research Center for Mental Disorders, Affiliated Wenzhou Kangning Hospital of Wenzhou Medical University, Wenzhou, China

[*]these authors contributed equally

Corresponding Author:
Yanqiu Yu, PhD
Department of Preventive Medicine and Health Education, School of Public Health, Fudan University, Dong'an Road 130, Xuhui District, Shanghai, China

Abstract

Background: Adolescents perceive both immediate and long-term benefits and losses related to internet gaming, affecting their risk of internet gaming disorder (IGD). These perceptions could also be shaped and reinforced by IGD, indicating potential bidirectionality.

Objective: This study aimed to investigate the bidirectional relationships between perceived immediate and long-term benefits in 3 domains (mental health, social relationships, and personal achievement) and IGD, and between perceived immediate and long-term losses in 6 domains (mental health, sleep quality, academic performance, family relationships, social relationships, and personal achievement) and IGD.

Methods: A 12-month 2-wave prospective longitudinal study was conducted among junior middle school students who had played internet games in the past 12 months in Guangzhou and Chengdu, China, with a baseline survey (T1, December 2018) and the other identical follow-up survey conducted 1 year later (T2, December 2019). The participating schools were conveniently selected; all Grade 7 and 8 students were invited to self-administer the questionnaires in a classroom setting without the presence of the schoolteachers. The final sample size was 1173 students (mean age 12.5, SD 0.6 y; male: 693/1173, 59.1%). IGD was assessed by using the 9-item *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* IGD checklist.

Results: Cross-lagged panel analysis (adjusting for background factors) showed (1) stronger perceived immediate benefits of mental health (β=.08, 95% CI 0.01-0.15) and personal achievement (β=.10, 95% CI 0.01-0.20) at T1 significantly predicted more IGD symptoms at T2; (2) more IGD symptoms at T1 significantly predicted stronger perceived immediate and long-term benefits of social relationships (immediate: β=.09, 95% CI 0.03-0.15; long-term: β=.11, 95% CI:0.05-0.17) and personal achievement (immediate: β=.12, 95% CI 0.06-0.18; long-term: β=.10, 95% CI 0.04-0.16) at T2; (3) more IGD symptoms at T1 significantly predicted stronger perceived immediate and future losses in mental health (immediate: β=.09, 95% CI 0.03-0.15; long-term: β=.08, 95% CI 0.02-0.14), sleep quality (immediate: β=.10, 95% CI 0.04-0.16; long-term: β=.13, 95% CI 0.07-0.19), academic performance (immediate: β=.09, 95% CI 0.04-0.15; long-term: β=.07, 95% CI 0.01-0.13), and family relationships (immediate: β=.11, 95% CI 0.05-0.17; long-term: β=.10, 95% CI 0.04-0.16) at T2, as well as perceived long-term losses in social relationships at T2 (β=.08, 95% CI 0.02-0.14).

Conclusions: This study was innovative in integrating time perspective into both perceived benefits and losses of internet gaming, a cognitive dimension previously overlooked in literature. The current findings advance the field by revealing the unidimensional predictive effects of IGD on perceived immediate and long-term benefits and losses, with 2 exceptions of perceived immediate and long-term benefits of mental health and personal achievement conversely predicting IGD. These results contribute to the development of effective interventions: the cognitive components should go beyond the general pros and cons of gaming and target the potential temporal bias gamers hold.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e74030) doi:10.2196/74030

KEYWORDS

internet gaming disorder; time perspective; perceived benefits; perceived losses; cross-lagged panel model; adolescents

Introduction

Internet gaming is increasingly popular; the number of internet gamers worldwide reached 3.31 billion in 2023, with a growing rate of 4.3% year-on-year [1]. China had the largest number of internet gamers (668 million as of December 2023) [2]. The internet penetration rate of Chinese minors (aged below 18 y) was 97.2%, and 67.8% of them played internet games regularly [3]. Internet gaming could lead to internet gaming disorder (IGD), a mental disorder included by *the International Classification of Diseases, 11th Revision (ICD-11)* [4], and negative consequences of social (eg, poor parent-child relationships), psychological (eg, depression), and functional (eg, impaired cognitive functions) problems [5-14]. A meta-analysis of 96 studies reported the pooled prevalence of IGD among adolescents ranged from 7.5% to 10% globally [15]; it ranged from 2.4% to 21.5% in China [16,17]. The potentially high adolescent IGD prevalence underscores the need for a better understanding of the precedents and consequences of IGD.

Numerous theoretical and empirical studies highlight the role of gaming-specific cognitions in understanding IGD development and maintenance [18], among which outcome expectancy forms a typical domain. According to social cognitive theory, outcome expectancy refers to tangible or intangible outcomes anticipated as results of a behavior [19]. Outcome expectancy could be manifested in positive (ie, perceived benefits) and negative (eg, perceived losses) domains [19]. Regarding perceived benefits of internet gaming, a recent study summarized 3 key dimensions: mental health (eg, emotional boost), social relationships (eg, social interactions), and personal achievement (eg, improved self-esteem) [20]; all these perceived benefits were positively associated with IGD in multiple cross-sectional studies [21-26]. To our knowledge, these associations have not yet been demonstrated longitudinally. Regarding perceived losses due to internet gaming, internet gaming can entail potential negative consequences [27-30] that constitute losses to gamers. The literature generally identified 6 dimensions of these losses. First, IGD could exacerbate mental problems [5,8]. Second, IGD predicted strongly sleep problems (eg, sleep deprivation [6] and poor sleep quality [31,32]). Third, IGD had negative impacts on family relationships, including worsened parent-child relationships [33] and decreased family functioning [34]. Fourth, although internet gaming provides a platform for online social interactions [35], it could undermine real-life social interactions [36,37] and increase social anxiety or phobia [5,38-40], generating losses in social relationships. Last, internet gaming may lead to declines in academic performance and personal achievement. IGD was longitudinally associated with worsened academic achievement [41,42] and personal achievement [43,44]. Adolescent gamers may be aware of these losses, potentially changing their gaming behaviors and affecting IGD risk. Notably, while existing literature has examined associations between IGD and negative consequences, few studies have directly investigated perceived losses per se, their predictive effects on IGD, or the reverse association.

It is also novel to integrate the time perspective into the concept of perceived benefits and losses of internet gaming. The temporal frames of past, present, and future affect one's decision-making and a variety of addictive behaviors [45-48]. Accordingly, perceived benefits and losses of internet gaming could occur at present (immediate benefits and losses) or in the future (long-term benefits and losses). Our literature review identified only 1 cross-sectional study looking at both perceived immediate and future benefits of internet gaming simultaneously, and it reported positive associations between these 2 perceived benefits and IGD with a comparable effect size [20]. It suggests that perceived future benefits might be as important as perceived immediate benefits and should not be neglected in future studies. In contrast, it was expected that these perceived benefits might be the outcomes of IGD. Our literature search identified a sole study reporting that IGD significantly predicted positive outcome expectancies of using internet gaming as a self-regulatory strategy [49]. In addition, no studies looked at the reciprocity between perceived immediate and long-term losses and IGD.

The reciprocal determinism construct of Social Cognitive Theory postulates that personal cognitive factors are key drivers of health behavior and could be shaped and reinforced by the behavior itself [19], supporting the bidirectional relationships between perceived immediate and long-term benefits and losses and IGD. There is additional theoretical support. The expectancy-value theories postulate that a positive outcome expectancy would increase the likelihood of performing a behavior, while a negative outcome expectancy would decrease the likelihood of performing the behavior [50]. Accordingly, perceived benefits would be expected to increase IGD risk while perceived losses were expected to mitigate it. Conversely, the predictive effect of IGD on perceived benefits and losses could be explained by Cognitive Dissonance Theory [51]. There may be cognitive dissonance when one's behavior contradicts one's well-being or values. To reduce this dissonance without changing behavior, the individual may alter their perceptions, potentially increasing perceived benefits and reducing perceived losses. Conversely, experiencing the negative consequences of IGD could plausibly reduce perceived benefits and heighten perceived losses.

Given the background, this 2-wave prospective longitudinal study aimed to (1) investigate the levels of perceived immediate and long-term benefits of internet gaming in 3 dimensions (mental health, social relationships, and personal achievement) and perceived immediate and long-term losses in 6 dimensions (mental health, sleep quality, academic performance, family relationships, social relationships, and personal achievement) among Chinese adolescents, and (2) examine the bidirectional

XSL•FO
RenderX

relationships between perceived immediate and long-term benefits and losses and IGD. It was hypothesized that (1) perceived immediate and long-term benefits at baseline would be positively associated with IGD at follow-up, (2) perceived immediate and long-term losses at baseline would be negatively associated with IGD at follow-up, (3) IGD at baseline would be significantly associated with perceived immediate and long-term benefits and losses at follow-up, and the directions would be either positive or negative.

Methods

Participants and Data Collection

A 12-month 2-wave prospective longitudinal study was conducted among junior middle school students in Guangzhou and Chengdu, China; baseline and follow-up surveys were performed in December 2018 and 2019, respectively. The inclusion criteria at baseline included full-time grade 7 and grade 8 students, aged 12 - 14 years, who had played internet games in the past 12 months at baseline, and were willing to participate in this survey. The selection of Grade 7 and 8 students was driven by their behavioral risks and environmental shifts specific to early adolescence. Behaviorally, this period involves heightened sensation-seeking and emotional instability, making the immediate gratification of gaming highly attractive [52]. Environmentally, these students are navigating the stressful transition to secondary education and experiencing reduced parental supervision regarding digital device use compared to primary school. Those Grade 9 students were excluded as they would leave school after 1 year and could not be followed up.

With the assistance of local education sectors, 4 and 2 junior middle schools were conveniently selected, and all grade 7 and 8 students of the selected schools were invited for participation; the in-school surveys were approved by school principals. The anonymous survey was self-administered by the students in the classroom setting in the absence of schoolteachers. Before the survey, well-trained field workers prebriefed the students on the objectives, content, logistics, and voluntary nature of this study, and that submitting a completed questionnaire implied informed consent for participation. Written informed consent was not collected to maintain anonymity. Such information was also printed on the cover page of the structured questionnaire. In addition, field workers assisted the students (eg, answering inquiries) and did the quality check during and after the survey, respectively. After 12 months, the students completed the same questionnaire with the same data collection procedures. The reporting of this study follows the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [53], and the checklist is included in the supplementary material.

At baseline, 2463 completed the questionnaire, and a total of 77.4% (1906/2463) questionnaires were matched for both baseline (T1) and follow-up (T2) surveys. The lost to follow-up (530/1906, 21.8%) group was more likely than those followed-up to be older, Chengdu participants, male, not living with both parents, having father's and mother's educational levels of junior middle school or below, and having higher scores of IGD and some domains of perceived immediate and long-term benefits and losses (Table S1 in Multimedia Appendix 1). Among the matched questionnaires, 25.9% (494/1906) were excluded due to not having played internet games in the past 12 months at T1, 5.4% (103/1906) due to missing data in key variables (eg, gaming behaviors and IGD), and 7.1% (136/1906) due to age below12 or above 14 years. The final sample size was 1173 in this study.

Measurements

Background Variables

Background information was collected, including age, sex, city of study, whether living with both parents, father's and mother's educational levels (junior middle school or below, senior middle school or equal, or college or above), and perceived family financial status (very good, good, moderate, poor, or very poor).

IGD

The 9-item *DSM-5* Checklist was used to screen for the presence of IGD symptoms in the past 12 months, including (1) preoccupation, (2) withdrawal symptoms, (3) tolerance, (4) inability to control internet gaming, (5) prioritizing internet games over other activities, (6) continuing to play internet games excessively despite psychosocial problems, (7) deceiving others about gaming time, (8) using internet games for escapism, and (9) impaired functions (eg, social and academic performance) due to internet gaming. All the items were rated with binary response options (0=No and 1=Yes). A total score was calculated by summing the responses, resulting in a range from 0 to 9; higher scores indicate a higher severity of IGD symptoms. The Chinese version of the *DSM-5* Checklist has been validated among adolescents in China with satisfactory reliability and validity indices [54]. In this study, the Cronbach α of this checklist at T1 and T2 was 0.75 and 0.77, respectively.

Perceived Immediate and Long-Term Benefits of Internet Gaming

Perceived immediate and perceived long-term benefits of internet gaming are assessed using the Perceived Overall Immediate Benefits Scale (POIBS) and the Perceived Overall Future Benefits Scale (POFBS; Fudan University), respectively, both of which were developed in Chinese adolescents and comprised 3 dimensions: mental health, social relationships, and personal achievement [20]. Perceived immediate benefits were rated as the extent to which the participants believed that internet gaming had positive impacts AT PRESENT on each of the above 3 dimensions (eg, "How much positive impact do you think internet gaming has AT PRESENT on your mental health?"). Perceived long-term benefits were rated as the extent to which the participants believed that internet gaming had positive impacts IN THE FUTURE on each of the above 3 dimensions (eg, "How much positive impact do you think internet gaming has IN THE FUTURE on your mental health?"). The 6 items were rated by using 5-point Likert scales (0=None to 4=extremely strong); higher scores indicate higher levels of perceived immediate and long-term benefits. The Cronbach α of POIBS and POFBS were 0.76/0.75 and 0.78/0.78 at T1/T2, respectively.

XSL•FO
RenderX

Perceived Immediate and Long-Term Losses Due to Internet Gaming

Overall, 6 dimensions of perceived losses due to internet gaming were considered in this study based on a thorough literature review and expert panel discussion, including (1) mental health, (2) sleep quality, (3) academic performance, (4) family relationships, (5) social relationships, and (6) personal achievement. Following previous publications and taking reference from the above POIBS and POFBS scale [20], perceived immediate losses were rated as the extent to which the participants believed that internet gaming had negative impacts AT PRESENT on each of the above 6 dimensions (eg, "How much negative impacts do you think internet gaming has AT PRESENT on your mental health?"), while perceived long-term losses were rated as the extent to which the participants believed that internet gaming had negative impacts IN THE FUTURE on each of the above 6 dimensions (eg, "How much negative impacts do you think internet gaming has IN THE FUTURE on your mental health?"). The 12 items were rated by using 5-point Likert scales (0=None to 4=extremely strong); higher scores indicate higher levels of perceived immediate and long-term losses. The Cronbach α of perceived immediate and long-term losses scales were 0.87/0.88 and 0.89/0.90 at T1/T2, respectively. Confirmatory factor analyses confirmed the 6-item structure of perceived immediate and long-term losses at both waves, demonstrating satisfactory model fit indices [perceived immediate losses (T1/T2): all factor loadings over 0.67/0.66 (all P<.001), comparative fit index (CFI)=0.97/0.97, Tucker-Lewis Index (TLI)=0.95/0.95, and standardized root mean squared residual (SRMR)=0.03/0.03; perceived long-term losses (T1/T2): all factor loadings over 0.71/0.69 (all P<.001), CFI=0.98/0.97, TLI=0.96/0.95, and SRMR=0.02/0.03].

Data Analysis

Descriptive statistics and normality tests (via Kolmogorov-Smirnov and Shapiro-Wilk) were performed. Little's Missing Completely at Random (MCAR) test was conducted to evaluate the pattern of missing data; Multiple Imputation was used to handle missing data. Paired-sample t tests were used to test for within-individual differences in levels of perceived immediate and long-term benefits and losses at both T1 and T2. Such data analysis was performed by using SPSS software (version 26.0; IBM Corp). A 2-sided P value <.05 was defined as statistically different.

Cross-lagged panel analyses with the maximum likelihood robust estimator were performed using Mplus 8.3; the analyses were conducted across 20 imputed datasets, and parameters were pooled according to Rubin's rules. In total, 9 models were fit to examine the potential bidirectional associations between each pair of perceived immediate versus long-term benefits and losses and IGD. These models had 3 types of correlations: (1) cross-sectional correlations between variables at the same time point, (2) autocorrelations between the level of the same variable at both T1 and T2, and (3) cross-lagged correlations between the variable at T1 and the other variables at T2. To better reveal the cross-lagged correlations (eg, perceived short-term benefits of mental health on IGD), both cross-sectional correlations and autocorrelations, as well as background factors (ie, age, sex, city of study, whether living with both parents, parental educational levels, and perceived family financial status) at T1 were controlled. Satisfactory model fit indices of the models included CFI≥0.90, TLI≥0.90, and SRMR≤0.08 [55].

Ethical Considerations

This study was approved by the Survey and Behavioral Research Ethics Committee of the Chinese University of Hong Kong (number SBRE-18‐430). Participation in this study was entirely voluntary. Prior to data collection, all participants were informed of the study objectives, procedures, and their rights as research participants. Given the anonymous nature of the survey, written informed consent was not obtained. Instead, informed consent was implied by the voluntary completion and submission of the questionnaire. Such information was also printed on the cover page of the structured questionnaire. All data were anonymous and no direct personal identifiers were collected. To enable longitudinal matching while preserving anonymity, instead, the last 4 digits of the father's and mother's phone numbers and the last phonetic letters of the father's and mother's given names were collected for matching purposes. No identifiable information about participants is hence possible in any part of this study or supplementary material. No financial or material incentives were provided to participants for their involvement in this study. In addition, parental approval was sought, and the parental opt-out procedure was exercised.

Results

Descriptive Statistics

Among all participants, the mean age was 12.5 (SD 0.6; range=12-14) years; 72.1% (846/1173) were from Guangzhou and 27.9% (327/1173) from Chengdu; over half (693/1173, 59.1%) were male; 15.6% (183/1173) did not live with both parents; 29.9% (350/1173) perceived that their family financial level was poor or very poor; nearly one-third of the participants' fathers (361/1173, 30.8%) or mothers (335/1173, 28.6%) had received college or above education (Table 1). The mean IGD scores were 2.1 (SD 2.1; range=0-9) and 1.7 (SD 2.1; range=0-9) at T1 and T2, respectively.

XSL•FO
RenderX

Table . Frequencies and proportions of background variables in the 2-wave cross-lagged panel study on bidirectionality between perceived immediate and long-term benefits and losses and internet gaming disorder among adolescent internet gamers in Chengdu and Guangzhou, China (December 2018 to December 2019).

Frequencies and proportions	n (%) (N=1173)
City of study	
Guangzhou	846 (72.1)
Chengdu	327 (27.9)
Sex	
Male	693 (59.1)
Female	480 (40.9)
Living with both parents	
Yes	975 (83.1)
No	183 (15.6)
Missing data	15 (1.3)
Family financial status	
Very good or good	112 (9.5)
Moderate	709 (60.4)
Poor or very poor	350 (29.9)
Missing data	2 (0.2)
Father's education level	
Junior middle school or below	427 (36.4)
Senior middle school or equivalent	316 (26.9)
College or above	361 (30.8)
Missing data	69 (5.9)
Mother's education level	
Junior middle school or below	451 (38.4)
Senior middle school or equivalent	322 (27.5)
College or above	335 (28.6)
Missing data	65 (5.5)

Within-Individual Comparison of Perceived Immediate and Long-Term Benefits and Losses at Both T1 and T2

The results (Table 2) showed (1) among the 3 dimensions of perceived immediate versus long-term benefits, all within-individual differences (except that between perceived immediate and long-term benefits of personal achievement at T2) were statistically significant, but the effect sizes were small (Cohen *d* ranged from 0.06 to 0.12), and (2) among the 6 dimensions of perceived immediate versus long-term losses, all within-individual differences (except those regarding sleep quality and social relationships at T1) were statistically significant with small effect sizes (Cohen *d* ranged from 0.08 to 0.22).

XSL•FO
RenderX

Table . The within-individual comparisons in perceived immediate and long-term benefits and losses in the 2-wave cross-lagged panel study on bidirectionality between perceived immediate and long-term benefits and losses and internet gaming disorder among adolescent internet gamers in Chengdu and Guangzhou, China (December 2018 to December 2019).

Dimension	T1[a], mean (SD)	Cohen *d*	*P* value	T2[b], mean (SD)	Cohen *d*	*P* value
Perceived benefits						
Mental health		−0.08	.004		0.06	.04
Immediate impacts	1.4 (1)			1.5 (1)		
Long-term impacts	1.4 (1.1)			1.4 (1.1)		
Social relationships		0.07	.01		0.09	.002
Immediate impacts	1.5 (1.3)			1.5 (1.2)		
Long-term impacts	1.5 (1.3)			1.4 (1.2)		
Personal achievement		−0.12	<.001		−0.03	.30
Immediate impacts	1.1 (1.2)			1.1 (1.1)		
Long-term impacts	1.2 (1.3)			1.1 (1.2)		
Perceived losses						
Mental health		−0.22	<.001		−0.2	<.001
Immediate impacts	1.1 (1.1)			1.3 (1.1)		
Long-term impacts	1.3 (1.2)			1.4 (1.2)		
Sleep quality		−0.05	.13		−0.08	.01
Immediate impacts	1.1 (1.2)			1.2 (1.2)		
Long-term impacts	1.1 (1.3)			1.3 (1.3)		
Academic performance		−0.08	.007		−0.08	.004
Immediate impacts	1.3 (1.2)			1.3 (1.2)		
Long-term impacts	1.3 (1.3)			1.4 (1.3)		
Family relationships		0.03	.25		−0.1	.001
Immediate impacts	1 (1.2)			1 (1.3)		
Long-term impacts	0.9 (1.2)			1.1 (1.2)		
Social relationships		−0.15	<.001		−0.13	<.001
Immediate impacts	0.7 (1)			0.8 (1)		
Long-term impacts	0.8 (1.1)			0.9 (1.1)		

XSL•FO
RenderX

Dimension	T1[a], mean (SD)	Cohen *d*	*P* value	T2[b], mean (SD)	Cohen *d*	*P* value
Personal achievement		−0.13	<.001		−0.14	<.001
Immediate impacts	0.8 (1.1)			0.9 (1.1)		
Long-term impacts	0.9 (1.2)			1 (1.2)		

[a]T1: assessment at baseline.

[b]T2: assessment at follow-up.

Cross-Lagged Panel Analysis

Bidirectional Relationships Between Perceived Immediate Versus Long-Term Benefits and IGD

Normality tests demonstrated that all key variables (perceived benefits and losses and IGD) at T1 and T2 violated the assumption of normality. Little's MCAR test was statistically significant (χ^2_{1353}=1737.65; *P*<.001), suggesting that data were not missing completely at random. Consequently, the cross-lagged panel models were estimated using a maximum likelihood robust estimator combined with multiple imputation to account for nonnormality and missing patterns. In this section, 3 cross-lagged panel models were fit, each containing a pair of perceived immediate versus long-term benefits and IGD and adjusting for background variables (ie, age, sex, city of study, whether living with both parents, parental educational levels, and perceived family financial status); all models showed satisfactory model fit indices (Figure 1). Higher levels of perceived immediate benefits of mental health and personal achievement (but not social relationships) at T1 predicted more IGD symptoms at T2, while the predictive effects of perceived long-term benefits in the 3 dimensions on IGD were statistically nonsignificant. Conversely, more IGD symptoms at T1 significantly predicted higher levels of perceived immediate and long-term benefits of social relationships and personal achievement (but not mental health) at T2.

Figure 1. Cross-lagged panel model testing the bidirectional relationships between perceived immediate and long-term benefits and internet gaming disorder in the 2-wave cross-lagged panel study among adolescent internet gamers in Chengdu and Guangzhou, China (December 2018 to December 2019). Standardized coefficients were presented. Dashed lines represent nonsignificant paths and solid lines represent significant paths. The models were adjusted for background factors including the city of study, sex, age, father's and mother's educational level, and whether living with both parents. The cross-sectional correlations and autocorrelations were not presented for simplicity. CFI: Comparative Fit Index; IGD: internet gaming disorder; SRMR: standardized root mean square residual; T1: assessment at baseline; T2: assessment at follow-up; TLI: Tucker-Lewis Index.

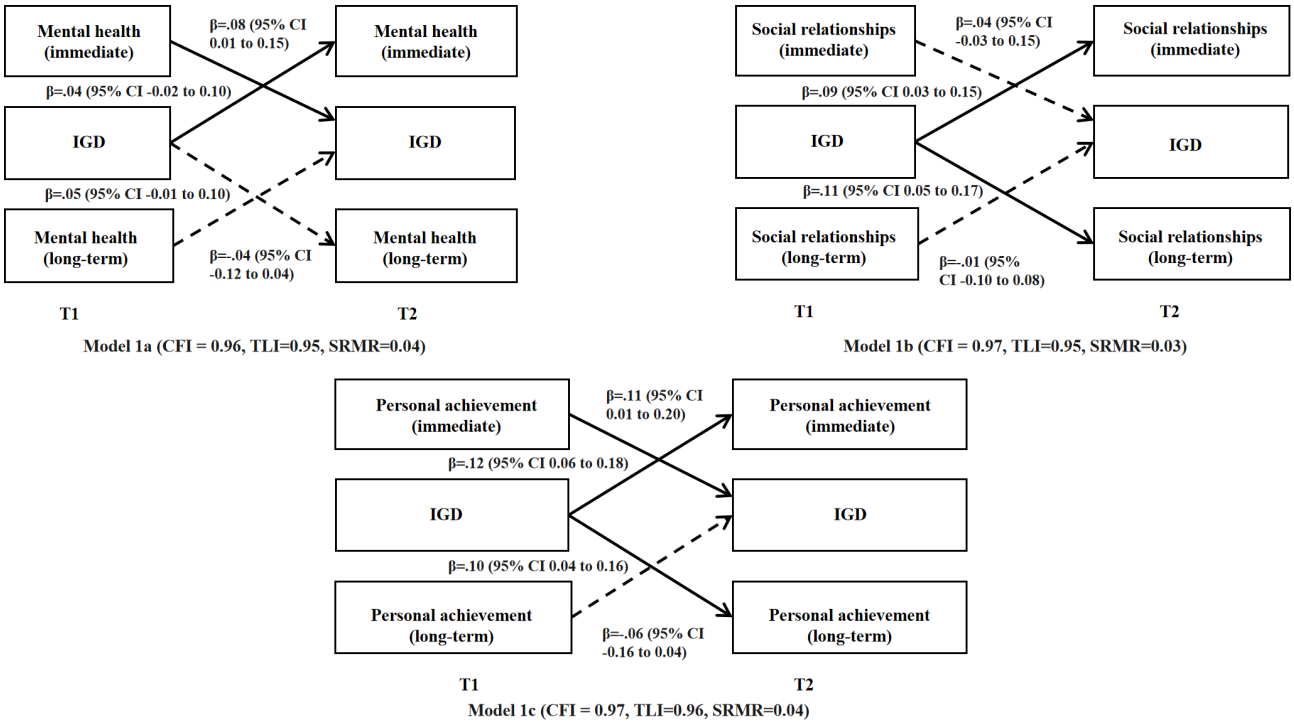

Bidirectional Relationships Between Perceived Immediate Versus Long-Term Losses and IGD

Similarly, 6 cross-lagged panel models were fitted, each containing a pair of perceived immediate versus long-term losses and IGD and adjusting for background variables; all models showed satisfactory model fit indices (Figure 2). All predictive effects of perceived immediate and long-term losses in mental health, sleep quality, academic performance, family relationships, social relationships, and personal achievement at

XSL•FO
RenderX

T1 on IGD at T2 were statistically nonsignificant. Conversely, more IGD symptoms at T1 significantly predicted higher levels of perceived immediate and long-term losses in mental health, sleep quality, academic performance, family relationships, and social relationships (only the long-term losses) at T2; the predictive effect of IGD at T1 on perceived losses in personal achievement was statistically nonsignificant.

Figure 2. Cross-lagged panel model testing the bidirectional relationships between perceived immediate and long-term losses and internet gaming disorder in the 2-wave cross-lagged panel study among adolescent internet gamers in Chengdu and Guangzhou, China (December 2018 to December 2019). Standardized coefficients were presented. Dashed lines represent nonsignificant paths and solid lines represent significant paths. The models were adjusted for background factors including the city of study, sex, age, father's and mother's educational level, and whether living with both parents. The cross-sectional correlations and autocorrelations were not presented for simplicity. CFI: Comparative Fit Index; IGD: internet gaming disorder; SRMR: standardized root mean square residual; T1: assessment at baseline; T2: assessment at follow-up; TLI: Tucker-Lewis Index.

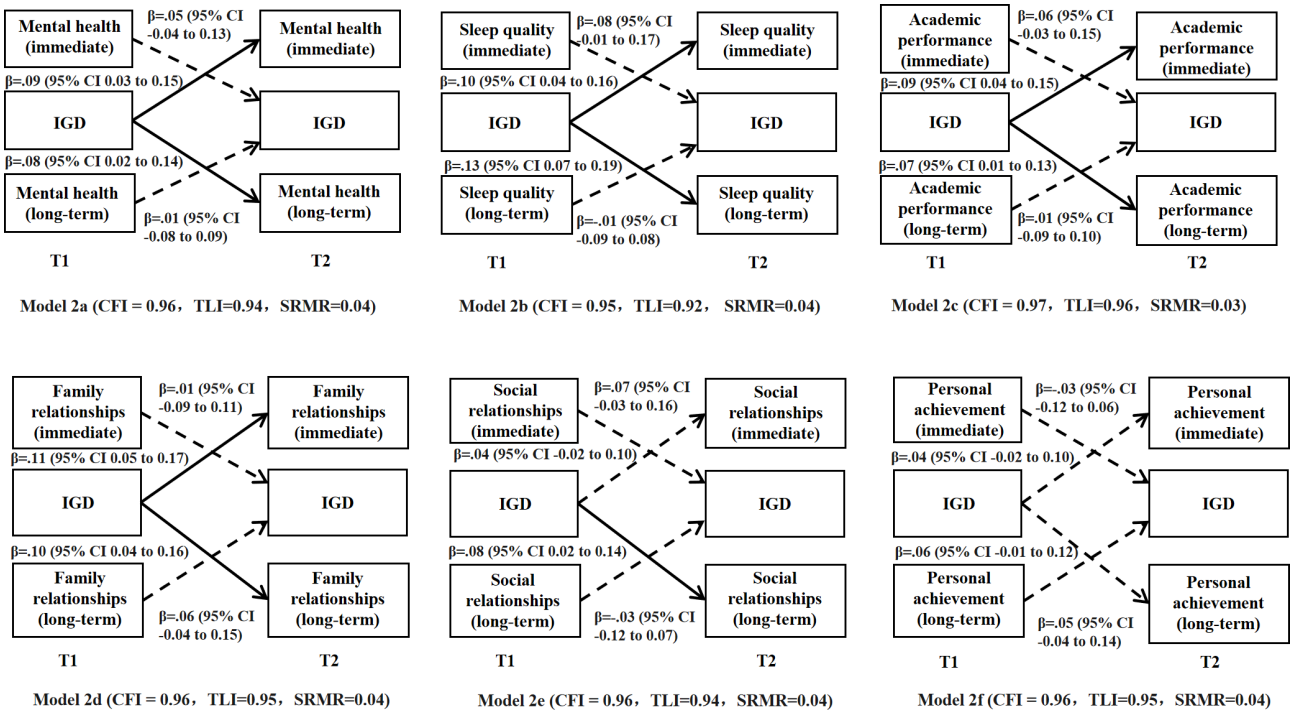

Discussion

Principal Findings

This study was novel to investigate the levels of perceived immediate and long-term benefits in 3 dimensions (mental health, social relationships, and personal achievement) and perceived immediate and long-term losses in 6 dimensions (mental health, sleep quality, academic performance, family relationships, social relationships, and personal achievement). All these levels were comparably low (within a range of 0 - 4); the mean scores for perceived benefits ranged from 1.1 to 1.5, while those for perceived losses ranged from 0.7 to 1.4. These findings reflect the nature of the general adolescent sample, where the majority of participants engage in gaming recreationally without reporting extreme levels of both positive and negative experiences. Nonetheless, the observations confirmed a previous study reporting comparable levels of perceived immediate versus long-term benefits [22] and further underscore the importance of integrating the time perspective into perceived benefits and losses. As the first study to investigate perceived losses, our findings suggest that adolescents may perceive both benefits and losses equally, and perceived losses should be a nonnegligible target in future interventions; future studies in this regard are greatly warranted for confirmation and elaboration.

This longitudinal study was also the first to reveal the interesting bidirectional relationships between perceived immediate and long-term benefits and IGD. Although existing cross-sectional studies reported that perceived benefits of improved mental health, social relationships, and personal achievement were positively associated with IGD [21-26], this longitudinal study only partially supports these findings and revealed the differential predictive effects of these perceptions on IGD. The significant results highlight the importance of perceived immediate mental health and personal achievement benefits. Plausibly, the immediate mental health benefits could offer quick and easily accessible relief from negative emotions for gamers. Furthermore, the instant gratification model postulates that behaviors (eg, internet gaming) providing immediate emotional relief are more likely to be repeated [56], potentially contributing to increased IGD risk. On the other hand, the self-determination theory postulates that satisfaction with the psychological need for competence is powerful in motivating behavior [57]. Accordingly, the sense of accomplishment gained from internet gaming could supply immediate feedback and rewards (ie, immediate personal achievement benefits), motivating repetitive gaming, thereby increasing IGD risk. In addition, the significant predictive effects involving immediate but not long-term benefits corroborate previous publications emphasizing immediate rewards over delayed gratification in the context of IGD [58]. Speculatively, long-term benefits on mental health and personal achievement from internet gaming

XSL•FO
RenderX

might occur gradually and remain less tangible to adolescent gamers, leading to the nonsignificant results. The nonsignificant predictive effects regarding perceived social relationships benefits might be associated with the complexity of social interactions, which can be context-dependent and influenced by external factors outside gaming (eg, peer relationships and behavioral inhibition) [59,60]. It is speculated that social interactions in gaming might not necessarily translate into meaningful or fulfilling relationships, which might reduce their impacts on IGD development.

Conversely, IGD at T1 significantly predicted perceived immediate and long-term benefits of social relationships and personal achievement at T2, but not those of mental health. It suggests that those with more IGD symptoms might amplify perceived benefits related to social relationships and personal achievement. Internet gaming is often designed with structured opportunities for social interactions (eg, multiplayer engagement) and measurable achievements (eg, leveling up) [61]; these aspects might be more pronounced among those with more IGD symptoms, leading to stronger perceptions of these benefits over time. In addition, those with more IGD symptoms may prioritize social and achievement benefits from internet gaming over those from other sources over time [21,62]. In contrast, individuals with more IGD symptoms might experience negative mental health outcomes, such as anxiety and depression [5,8]. These negative psychological conditions may overshadow the mental health benefits and lead to the nonsignificant predictive effect of IGD on perceived mental health benefits.

Unexpectedly, all predictive effects of perceived immediate and long-term losses in 6 domains at T1 on IGD at T2 were statistically nonsignificant. While perceived benefits and losses might be perceived at comparable levels, the psychological impacts of immediate benefits might have stronger influences on developing IGD due to their ability to fulfill psychological needs like competence as aforementioned [20]. This could overshadow the importance of perceived losses in terms of affecting IGD. This speculation was also supported by the concept of reinforcement that positive reinforcement (eg, perceived benefits) tends to have more immediate and stronger effects on motivating behavior in comparison with negative reinforcement or punishment (eg, perceived losses) [63]. Nonetheless, such speculations should be tested in future studies.

Reversely, this study observed differential predictive effects of IGD at T1 on the 6 domains of perceived immediate and long-term losses at T2. First, IGD significantly predicted perceived immediate and long-term losses in mental health, sleep quality, academic performance, and family relationships, suggesting that IGD could lead to noticeable harm in these domains. These findings align with empirical evidence that IGD cases face common negative consequences, such as depression, disturbed sleep patterns, worsened academic performance, and strained family relationships [27,33,41]. Notably, these domains of perceived losses might be more directly observable or harder to rationalize over time and are recommended to be emphasized in IGD intervention programs. Second, IGD at T1 significantly predicted perceived long-term, but not immediate, losses in social relationships, indicating a potential lag effect. It is speculated that, initially, the social benefits of internet gaming might buffer against perceived immediate losses, but, over time, those with more IGD symptoms might experience weakened offline social ties and relationships [64], making the long-term losses more apparent and significant. Last, the predictive effects of IGD at T1 on perceived immediate and long-term losses in personal achievement were statistically nonsignificant, suggesting that individuals with more IGD symptoms might not associate their gaming habits with detriments in personal achievement. Plausibly, as aforementioned, internet gaming may fulfill one's psychological needs of competence [65] and provide satisfactory substitutes for real-life achievements [66], decreasing perceived losses.

The current findings have practical implications, as gaming-specific cognitions are important modifiable constructs in IGD interventions (eg, cognitive-behavioral therapy) [67]. IGD prevention interventions may focus more on the immediate benefits on mental health and personal achievement. For instance, alternatives for immediate rewards of mental health and personal achievements (eg, outdoor physical activities) should be provided for adolescents to reduce their reliance on and sensitivity toward these gaming rewards. It might also be helpful to separate gaming achievement from real-world achievement and then educate adolescents on how to fulfill the latter [68]. In contrast, IGD treatment intervention may underscore the cognitions of potential losses in mental health, sleep quality, academic performance, and family relationships, as well as the long-term losses in social relationships, while perceived rewards of social relationships and personal achievement are recommended to be downplayed.

There are several limitations of this study. First, the attrition rate of this study was about 20%. Attrition analysis revealed that those lost to follow-up demonstrated stronger perceived benefit and losses in some domains, besides specific background differences (eg, age and sex). In addition, Little's MCAR test showed that the missing data were not completely random. It suggests that the study sample may have stronger perceived benefits and losses than the baseline population. Nonetheless, multiple imputation was used in the cross-lagged panel analyses to address the missing data patterns, increasing the robustness of the current findings. Furthermore, the study population was selected from 2 cities in China through convenience sampling, which may lead to selection bias. Cautions are needed when extrapolating the results to other regions or countries. Second, as participants were asked to self-administer the questionnaire, there might be reporting bias, including recall bias and social desirability bias. Third, although this study provides valuable insights on the bidirectional relationships between perceived immediate and long-term benefits and losses and IGD, it comprised only 2 time-points within a 12-month follow-up period. Future longitudinal studies with more time lags and longer study periods are needed to confirm the findings. In addition, while several predictive effects were statistically significant, their associated effect size was modest. It suggests that, despite statistical significance, these associations might have limited practical impacts on individuals. However, it is important to note that small effect sizes could still bear practical implications at the population level. Modest increases or

XSL•FO
RenderX

reductions in each predictor might yield substantial benefits for the population, though future intervention studies are needed for confirmation. Last, there might be other potential losses due to internet gaming that were not included in this study (eg, losses in physical health and the financial domain).

Conclusions

This study was novel in integrating the time perspective into perceived benefits and losses related to internet gaming and investigating their bidirectional relationships with IGD. The results are interesting in that only perceived immediate benefits of mental health and personal achievement predicted IGD, while IGD predicted perceived immediate and long-term benefits on social relationships and personal achievement, as well as perceived immediate and long-term losses in all 6 domains (except personal achievement). These findings enhance the understanding of the relationships between gaming-specific cognitions and IGD and shed new insights for relevant studies, taking into consideration the time perspective. The differential bidirectional relationships are also important for the design of IGD prevention and treatment interventions. Specifically, relevant cognitive components should target the specific temporal bias adolescent gamers possess, rather than focusing merely on the general benefits and losses of internet gaming.

Acknowledgments

We would like to thank the participants for their contribution to the present. The authors declare that no generative artificial intelligence was used in the drafting of this manuscript.

Funding

This study was funded by Shanghai Pujiang Programme (No. 23PJC018) and the National Natural Science Foundation of China (No. 82304258). The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the article.

Data Availability

The data sets generated and/or analysed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization and validation: YY
Methodology and writing-review and editing: SL, YY, and JTFL
Investigation: JZ and JBL
Software, formal analysis, data curation, and writing-original draft: SL
Resources: JTFL
Supervision: YY and JTFL

Conflicts of Interest

None declared.

Multimedia Appendix 1
Attrition analysis in the 2-wave cross-lagged panel study on bidirectionality between perceived immediate and long-term benefits/losses and internet gaming disorder among adolescent internet gamers in Chengdu and Guangzhou, China (December 2018 to December 2019).
[DOCX File, 24 KB - jmir_v28i1e74030_app1.docx]

References

1. Newzoo's global games market report 2023. Newzoo. URL: https://newzoo.com/resources/trend-reports/newzoo-global-games-market-report-2023-free-version [accessed 2026-02-11]
2. China game industry report 2023. CGIGC. URL: https://www.cgigc.com.cn/details.[illegible] [accessed 2026-02-11]
3. (CNNIC). Report on the 5th national survey on internet use by minors. 2023.
4. ICD-11 for mortality and morbidity statistics. URL: https://icd.who.int/browse/2025-01/mms/en [accessed 2026-02-11]
5. Gentile DA, Choo H, Liau A, et al. Pathological video game use among youths: a two-year longitudinal study. Pediatrics 2011 Feb;127(2):e319-e329. [doi: 10.1542/peds.2010-1353] [Medline: 21242221]
6. Kim SY, Kim MS, Park B, Kim JH, Choi HG. Lack of sleep is associated with internet use for leisure. PLoS ONE 2018;13(1):e0191713. [doi: 10.1371/journal.pone.0191713] [Medline: 29360882]
7. Schneider LA, King DL, Delfabbro PH. Family factors in adolescent problematic Internet gaming: a systematic review. J Behav Addict 2017 Sep 1;6(3):321-333. [doi: 10.1556/2006.6.2017.035] [Medline: 28762279]

XSL•FO
RenderX

8. Ahmadi J, Amiri A, Ghanizadeh A, et al. Prevalence of addiction to the internet, computer games, DVD, and video and its relationship to anxiety and depression in a sample of Iranian High School students. Iran J Psychiatry Behav Sci 2014;8(2):75-80. [Medline: 25053960]
9. Henchoz Y, Studer J, Deline S, N'Goran AA, Baggio S, Gmel G. Video gaming disorder and sport and exercise in emerging adulthood: a longitudinal study. Behav Med 2016;42(2):105-111. [doi: 10.1080/08964289.2014.965127] [Medline: 25258243]
10. De Pasquale C, Dinaro C, Sciacca F. Relationship of internet gaming disorder with dissociative experience in Italian university students. Ann Gen Psychiatry 2018;17:28. [doi: 10.1186/s12991-018-0198-y] [Medline: 29983724]
11. Ma HK. Internet addiction and antisocial internet behavior of adolescents. ScientificWorldJournal 2011;11:2187-2196. [doi: 10.1100/2011/308631] [Medline: 22125466]
12. Yamada T, Moshier SJ, Otto MW. No site unseen: predicting the failure to control problematic internet use among young adults. Cogn Behav Ther 2016 Nov;45(6):496-500. [doi: 10.1080/16506073.2016.1205657] [Medline: 27426432]
13. Dong G, Potenza MN. Risk-taking and risky decision-making in Internet gaming disorder: implications regarding online gaming in the setting of negative consequences. J Psychiatr Res 2016 Feb;73:1-8. [doi: 10.1016/j.jpsychires.2015.11.011] [Medline: 26656573]
14. Wang Q, Ren H, Long J, Liu Y, Liu T. Research progress and debates on gaming disorder. Gen Psychiatr 2019;32(3):e100071. [doi: 10.1136/gpsych-2019-100071] [Medline: 31423477]
15. Gao YX, Wang JY, Dong GH. The prevalence and possible risk factors of internet gaming disorder among adolescents and young adults: systematic reviews and meta-analyses. J Psychiatr Res 2022 Oct;154:35-43. [doi: 10.1016/j.jpsychires.2022.06.049] [Medline: 35926424]
16. Paulus FW, Ohmann S, von Gontard A, Popow C. Internet gaming disorder in children and adolescents: a systematic review. Develop Med Child Neuro 2018 Jul;60(7):645-659 [FREE Full text] [doi: 10.1111/dmcn.13754]
17. Liao Z, Huang Q, Huang S, et al. Prevalence of internet gaming disorder and its association with personality traits and gaming characteristics among Chinese adolescent gamers. Front Psychiatry 2020;11:598585. [doi: 10.3389/fpsyt.2020.598585] [Medline: 33312143]
18. Billieux J, Potenza MN, Maurage P, et al. Cognitive Factors Associated with Gaming Disorder: Academic Press; 2020:221-230. [doi: 10.1016/B978-0-12-815298-0.00016-2]
19. Nabavi RT. Bandura's social learning theory & social cognitive learning theory. In: Theory of Developmental Psychology 2012:1-24 URL: https://www.researchgate.net/profile/Mohammad-Bijandi/publication/267750204_Bandura's_Social_Learning_Theory_Social_Cognitive_Learning_Theory/links/67536801b558f41d0fbe2fb1/Banduras-Social-Learning-Theory-Social-Cognitive-Learning-Theory.pdf
20. Yu Y, Fong VWI, Li JB, Zhang J, Lau JTF. The independent impacts of perceived immediate and future benefits of internet gaming on internet gaming disorder among adolescent internet gamers in China: mediation via internet gaming time. Comput Human Behav 2023 Dec;149:107965. [doi: 10.1016/j.chb.2023.107965]
21. Wang HY, Cheng C. The associations between gaming motivation and internet gaming disorder: systematic review and meta-analysis. JMIR Ment Health 2022 Feb 17;9(2):e23700. [doi: 10.2196/23700] [Medline: 35175204]
22. Cabeza-Ramírez LJ, Sánchez-Cañizares SM, Fuentes-García FJ, Santos-Roldán LM. Exploring the connection between playing video games and watching video game streaming: relationships with potential problematic uses. Comput Human Behav 2022 Mar;128:107130. [doi: 10.1016/j.chb.2021.107130]
23. Yildiz Durak H, Haktanir A, Saritepeci M. Examining the predictors of video game addiction according to expertise levels of the players: the role of time spent on video gaming, engagement, positive gaming perception, social support and relational health indices. Int J Ment Health Addict 2023 May 9:1-26. [doi: 10.1007/s11469-023-01073-3] [Medline: 37363766]
24. Cabeza-Ramírez LJ, Rey-Carmona FJ, Del Carmen Cano-Vicente M, Solano-Sánchez M. Analysis of the coexistence of gaming and viewing activities in twitch users and their relationship with pathological gaming: a multilayer perceptron approach. Sci Rep 2022 May 12;12(1):7904. [doi: 10.1038/s41598-022-11985-0] [Medline: 35551493]
25. Zeng W, Wei H, Liu M. Need for distinctiveness leads to pathological internet use? The perspective of cognitive behavioral model. Int J Environ Res Public Health 2023 Jan 16;20(2):1609. [doi: 10.3390/ijerph20021609] [Medline: 36674362]
26. Johannes N, Vuorre M, Przybylski AK. Video game play is positively correlated with well-being. R Soc Open Sci 2021 Feb 17;8(2):202049. [doi: 10.1098/rsos.202049] [Medline: 33972879]
27. Chen YL, Gau SSF. Sleep problems and internet addiction among children and adolescents: a longitudinal study. J Sleep Res 2016 Aug;25(4):458-465. [doi: 10.1111/jsr.12388] [Medline: 26854132]
28. Lemmens JS, Valkenburg PM, Peter J. The effects of pathological gaming on aggressive behavior. J Youth Adolesc 2011 Jan;40(1):38-47. [doi: 10.1007/s10964-010-9558-x] [Medline: 20549320]
29. Gentile D. Pathological video-game use among youth ages 8 to 18: a national study. Psychol Sci 2009 May;20(5):594-602. [doi: 10.1111/j.1467-9280.2009.02340.x] [Medline: 19476590]
30. Yu Y, Yang X, Wang S, et al. Serial multiple mediation of the association between internet gaming disorder and suicidal ideation by insomnia and depression in adolescents in Shanghai, China. BMC Psychiatry 2020 Sep 23;20(1):460. [doi: 10.1186/s12888-020-02870-z] [Medline: 32967648]
31. Lam LT. Internet gaming addiction, problematic use of the internet, and sleep problems: a systematic review. Curr Psychiatry Rep 2014 Apr;16(4):1-9. [doi: 10.1007/s11920-014-0444-1] [Medline: 24619594]

XSL•FO
RenderX

32. Smyth JM. Beyond self-selection in video game play: an experimental examination of the consequences of massively multiplayer online role-playing game play. Cyberpsychol Behav 2007 Oct;10(5):717-721. [doi: 10.1089/cpb.2007.9963]
33. Charlie CWD, HyeKyung C, Khoo A. Role of parental relationships in pathological gaming. Procedia - Social and Behavioral Sciences 2011;30:1230-1236. [doi: 10.1016/j.sbspro.2011.10.238]
34. Ko CH, Wang PW, Liu TL, Yen CF, Chen CS, Yen JY. Bidirectional associations between family factors and internet addiction among adolescents in a prospective investigation. Psychiatry Clin Neurosci 2015 Apr;69(4):192-200. [doi: 10.1111/pcn.12204] [Medline: 24836367]
35. Liu CC, Chang IC. Model of online game addiction: the role of computer-mediated communication motives. Telematics and Informatics 2016 Nov;33(4):904-915. [doi: 10.1016/j.tele.2016.02.002]
36. Kaczmarek LD, Drążkowski D. MMORPG escapism predicts decreased well-being: examination of gaming time, game realism beliefs, and online social support for offline problems. Cyberpsychol Behav Soc Netw 2014 May;17(5):298-302. [doi: 10.1089/cyber.2013.0595] [Medline: 24605951]
37. Chappell D, Eatough V, Davies MNO, Griffiths M. EverQuest—it's just a computer game right? An interpretative phenomenological analysis of online gaming addiction. Int J Ment Health Addiction 2006 Jul;4(3):205-216. [doi: 10.1007/s11469-006-9028-6]
38. Marino C, Canale N, Vieno A, Caselli G, Scacchi L, Spada MM. Social anxiety and internet gaming disorder: the role of motives and metacognitions. J Behav Addict 2020 Oct 12;9(3):617-628. [doi: 10.1556/2006.2020.00044] [Medline: 32750032]
39. Sioni SR, Burleson MH, Bekerian DA. Internet gaming disorder: social phobia and identifying with your virtual self. Comput Human Behav 2017 Jun;71:11-15. [doi: 10.1016/j.chb.2017.01.044]
40. Gioia F, Colella GM, Boursier V. Evidence on problematic online gaming and social anxiety over the past ten years: a systematic literature review. Curr Addict Rep 2022 Mar;9(1):32-47. [doi: 10.1007/s40429-021-00406-3]
41. Brunborg GS, Mentzoni RA, Frøyland LR. Is video gaming, or video game addiction, associated with depression, academic achievement, heavy episodic drinking, or conduct problems? J Behav Addict 2014 Mar;3(1):27-32. [doi: 10.1556/JBA.3.2014.002] [Medline: 25215212]
42. Nie Q, Teng Z, Yang C, Griffiths MD, Guo C. Longitudinal relationships between school climate, academic achievement, and gaming disorder symptoms among chinese adolescents. J Youth Adolescence 2024 Jul;53(7):1646-1665. [doi: 10.1007/s10964-024-01952-5]
43. Teng Z, Pontes HM, Nie Q, Xiang G, Griffiths MD, Guo C. Internet gaming disorder and psychosocial well-being: a longitudinal study of older-aged adolescents and emerging adults. Addict Behav 2020 Nov;110:106530. [doi: 10.1016/j.addbeh.2020.106530] [Medline: 32683173]
44. Nie Q, Griffiths MD, Teng Z. The role of self-esteem in protecting against cyber-victimization and gaming disorder symptoms among adolescents: a temporal dynamics analysis. J Youth Adolesc 2024 Apr;53(4):863-876. [doi: 10.1007/s10964-023-01890-8] [Medline: 37904058]
45. Zimbardo PG, Boyd JN. Putting time in perspective: a valid, reliable individual-differences metric. In: Time Perspective Theory; Review, Research and Application: Essays in Honor of Philip G Zimbardo: Springer; 2014:17-55.
46. Apostolidis T, Fieulaine N, Soulé F. Future time perspective as predictor of cannabis use: exploring the role of substance perception among French adolescents. Addict Behav 2006 Dec;31(12):2339-2343. [doi: 10.1016/j.addbeh.2006.03.008] [Medline: 16626883]
47. Barnett E, Spruijt-Metz D, Unger JB, Rohrbach LA, Sun P, Sussman S. Bidirectional associations between future time perspective and substance use among continuation high-school students. Subst Use Misuse 2013 Jun 18;48(8):574-580. [doi: 10.3109/10826084.2013.787092]
48. Kim J, Hong H, Lee J, Hyun MH. Effects of time perspective and self-control on procrastination and internet addiction. J Behav Addict 2017 Jun 1;6(2):229-236. [doi: 10.1556/2006.6.2017.017] [Medline: 28494615]
49. Dang L, Yang HM, Spada MM, Wu AMS. A three-wave longitudinal study on the underlying metacognitive mechanism between depression and Internet gaming disorder. J Behav Addict 2024 Mar 26;13(1):215-225. [doi: 10.1556/2006.2023.00072] [Medline: 38183452]
50. Oei TP, Baldwin AR. Expectancy theory: a two-process model of alcohol use and abuse. J Stud Alcohol 1994 Sep;55(5):525-534. [doi: 10.15288/jsa.1994.55.525] [Medline: 7990462]
51. Chiou WB, Wan CS. Using cognitive dissonance to induce adolescents' escaping from the claw of online gaming: the roles of personal responsibility and justification of cost. Cyberpsychol Behav 2007 Oct;10(5):663-670. [doi: 10.1089/cpb.2007.9972]
52. Crone EA, Dahl RE. Understanding adolescence as a period of social-affective engagement and goal flexibility. Nat Rev Neurosci 2012 Sep;13(9):636-650. [doi: 10.1038/nrn3313] [Medline: 22903221]
53. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. The Lancet 2007 Oct;370(9596):1453-1457. [doi: 10.1016/S0140-6736(07)61602-X]

XSL•FO
RenderX

54. Yang X, Jiang X, Wu AMS, et al. Validation of the internet gaming disorder symptoms checklist based on the fifth edition of the diagnostic and statistical manual of mental disorders in Chinese adolescents. Child Psychiatry Hum Dev 2023 Feb;54(1):26-33. [doi: 10.1007/s10578-021-01213-7] [Medline: 34338962]
55. Kline RB. Principles and Practice of Structural Equation Modeling: Guilford Publications; 2023. URL: https://books.google.co.in/ [accessed 2026-02-16]
56. Harris C, Laibson D. Instantaneous gratification. Q J Econ 2013 Feb;128(1):205-248. [doi: 10.1093/qje/qjs051]
57. Deci EL, Ryan RM. Self-determination theory. In: Handbook of Theories of Social Psychology: Sage Publications; 2012:416-436. [doi: 10.4135/9781446249215.n21]
58. Weinstein A, Abu HB, Timor A, Mama Y. Delay discounting, risk-taking, and rejection sensitivity among individuals with internet and video gaming disorders. J Behav Addict 2016 Dec;5(4):674-682. [doi: 10.1556/2006.5.2016.081] [Medline: 27958761]
59. Essex MJ, Klein MH, Slattery MJ, Goldsmith HH, Kalin NH. Early risk factors and developmental pathways to chronic high inhibition and social anxiety disorder in adolescence. Am J Psychiatry 2010 Jan;167(1):40-46. [doi: 10.1176/appi.ajp.2009.07010051] [Medline: 19917594]
60. Tillfors M, Persson S, Willén M, Burk WJ. Prospective links between social anxiety and adolescent peer relations. J Adolesc 2012 Oct;35(5):1255-1263. [doi: 10.1016/j.adolescence.2012.04.008] [Medline: 22595770]
61. Wang H, Sun CT. Game reward systems: gaming experiences and social meanings. Presented at: Proceedings of DiGRA 2011 Conference; Sep 14-17, 2011. [doi: 10.26503/dl.v2011i1.594]
62. Collins E, Freeman J. Do problematic and non-problematic video game players differ in extraversion, trait empathy, social capital and prosocial tendencies? Comput Human Behav 2013 Sep;29(5):1933-1940. [doi: 10.1016/j.chb.2013.03.002]
63. Shteingart H, Loewenstein Y. Reinforcement learning and human behavior. Curr Opin Neurobiol 2014 Apr;25:93-98. [doi: 10.1016/j.conb.2013.12.004] [Medline: 24709606]
64. Longman H, O'Connor E, Obst P. The effect of social support derived from world of warcraft on negative psychological symptoms. Cyberpsychol Behav 2009 Oct;12(5):563-566. [doi: 10.1089/cpb.2009.0001]
65. Wan CS, Chiou WB. Why are adolescents addicted to online gaming? An interview study in Taiwan. Cyberpsychol Behav 2006 Dec;9(6):762-766. [doi: 10.1089/cpb.2006.9.762]
66. Snodgrass JG, Dengah HJF 2nd, Lacy MG, Fagan J. A formal anthropological view of motivation models of problematic MMO play: achievement, social, and immersion factors in the context of culture. Transcult Psychiatry 2013 Apr;50(2):235-262. [doi: 10.1177/1363461513487666] [Medline: 23690445]
67. King DL, Delfabbro PH. The cognitive psychopathology of internet gaming disorder in adolescence. J Abnorm Child Psychol 2016 Nov;44(8):1635-1645. [doi: 10.1007/s10802-016-0135-y] [Medline: 26875565]
68. Allen JJ, Anderson CA. Satisfaction and frustration of basic psychological needs in the real world and in video games predict internet gaming disorder scores and well-being. Comput Human Behav 2018 Jul;84:220-229. [doi: 10.1016/j.chb.2018.02.034]

Abbreviations

CFI: comparative fit index
DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
ICD-11: *International Classification of Diseases, 11th Revision*
IGD: internet gaming disorder
MCAR: missing completely at random
POFBS: Perceived Overall Future Benefits Scale
POIBS: Perceived Overall Immediate Benefits Scale
SRMR: standardized root mean squared residual
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TLI: Tucker-Lewis Index

Edited by S Brini; submitted 16.Mar.2025; peer-reviewed by T Baranowski, TY Wang; accepted 26.Jan.2026; published 26.Feb.2026.

Please cite as:
Li S, Zhang J, Li JB, Lau JTF, Yu Y
Bidirectionality Between Perceived Immediate and Long-Term Benefits and Losses and Internet Gaming Disorder Among Chinese Adolescent Gamers: Prospective Longitudinal Study
J Med Internet Res 2026;28:e74030
URL: https://www.jmir.org/2026/1/e74030
doi:10.2196/74030

XSL•FO
RenderX

© Siman Li, Jianxin Zhang, Ji-bin Li, Joseph TF Lau, Yanqiu Yu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Interactions of Technology and Obsessive-Compulsive Disorder Symptomatology in Adults: Qualitative Interview Study

Lucas Occhino-Moede, MD, MPH; Kaitlyn Sulivan-Pascual; Kendall Phelan, MPH; Harrison Wang, BS; Daniel Mokhtar, BS; Elisa Liu; Erica Schug, BS; Megan Mirkis, BS; Thomas Baek, BS; Tamerlane Visher, BA; Ujjwal Pasupulety, MS; Adam Charles Frank, MD, PhD

Department of Psychiatry and Behavioral Sciences, Keck School of Medicine, University of Southern California, 2250 Alcazar St., Suite 200, Los Angeles, CA, United States

Corresponding Author:
Adam Charles Frank, MD, PhD
Department of Psychiatry and Behavioral Sciences, Keck School of Medicine, University of Southern California, 2250 Alcazar St., Suite 200, Los Angeles, CA, United States

Abstract

Background: Obsessive-compulsive disorder (OCD) affects 1% - 3% of the population and is marked by intrusive obsessions and compulsive behaviors that impair daily functioning. As digital technologies have become ubiquitous, their features may interact with OCD symptom dimensions in ways that both exacerbate and alleviate symptoms. While case reports and clinical anecdotes suggest such interactions, systematic investigation of patients' lived experiences with technology remains limited.

Objective: This study aimed to explore how individuals with OCD perceive and navigate their interactions with modern technologies, and to identify how specific features of technology may contribute to, reinforce, or relieve obsessive-compulsive symptom cycles.

Methods: We conducted semistructured interviews (n=24) with adults self-reporting a diagnosis of OCD, recruited through online OCD communities and advocacy networks. Interviews were conducted via the HIPAA (Health Insurance Portability and Accountability Act)-compliant platform Zoom (Zoom Communications) between May and December 2024 (median duration 51, IQR 6.5 minutes). Transcripts were coded in Dedoose (version 9.2.22; SocioCultural Research Consultants) using a constructivist grounded theory approach. Coding proceeded iteratively through open and focused coding, with theoretical saturation reached after 15 interviews. Constant comparison and analytic memoing guided the development of a conceptual framework linking technology features to OCD symptom dimensions.

Results: Participants (median age 26, IQR 12.8, range 20 - 64 years; 67%, 16/24 women, 29%, 7/24 men, and 4%, 1/24 nonbinary) described technology as both a trigger for and a coping tool against OCD symptoms. Analysis produced four central technology-related categories: (1) information-provision platforms (eg, social media, search engines, large language models, etc) that triggered disturbing-thought obsessions and enabled compulsive checking and reassurance-seeking; (2) gamification and quantification features (eg, streaks, progress bars, and tracking metrics) that reinforced "not-just-right" and symmetry-based compulsions; (3) notifications that provoked urges to clear, check, and maintain control, spanning both disturbing-thought and symmetry domains; and (4) user interfaces whose complexity and customizability elicited compulsive ordering, avoidance behaviors, and digital overwhelm.

Conclusions: This study characterizes how interactions between OCD and digital technologies manifest across established symptom domains, most notably disturbing-thought and "not-just-right" categories. Participants overwhelmingly experienced compulsive checking, reassurance-seeking, and ordering behaviors reinforced by features such as information-provision, gamification, notifications, and user interfaces. These findings highlight the clinical relevance of technology-related compulsions and suggest value in their systematic assessment, incorporation into psychoeducation, and consideration in digital design.

(J Med Internet Res 2026;28:e85033) doi:10.2196/85033

KEYWORDS

obsessive-compulsive disorder; OCD; technology; wearables; qualitative; patient perspectives

Introduction

Obsessive-compulsive disorder (OCD) occurs in 1% - 3% of the population [1] and is most commonly characterized by the presence of obsessions and compulsions. Obsessions are intrusive, unwanted thoughts, images, or impulses that are typically experienced as ego-dystonic and distressing. Compulsions are repetitive actions or mental rituals in which an individual engages to relieve anxiety related to obsessions.

XSL•FO
RenderX

Obsessions and compulsions, in the context of OCD, are time consuming and often decrease functioning in social, occupational, and other aspects of life.

OCD is a heterogeneous disorder with variety in the types of obsessions and compulsions experienced, severity of symptoms, and presence of comorbidities. While no two individuals are likely to experience the same pattern of symptoms, factor analytic approaches have been used to parse OCD heterogeneity. These studies have consistently identified the following four broad dimensions: (1) forbidden or taboo thoughts, (2) symmetry and ordering, (3) contamination and cleaning, and (4) hoarding [2,3]. More recent research has used expanded symptom checklists and refined analytic techniques to explore the relationships between these dimensions. The higher-level groupings of incompleteness and disturbing thoughts (distressing images that typically follow themes like contamination, harmful acts, or religion) strongly associate with the other dimensions and thereby are postulated to represent core phenotypes within OCD. Subdimensions within these two categories were also delineated and noted to be structured in a hierarchy. The first order domain of incompleteness was divided into accuracy, mental and perceptual, and “not-just-right” (NJR) behaviors. NJR experiences are typically an internal sense of disturbance related to qualities such as order, placement, or frequency and drive individuals to engage in actions like tapping, counting, or reviewing. Meanwhile, disturbing thoughts encompassed symptoms relating to harm and checking and forbidden thoughts [4].

Excessive or inappropriate use of technology is associated with mental health issues. Children and adolescents are particularly susceptible; higher amounts of screen time and social media use have been linked to a range of mental health symptoms due to disturbances in regular sleep-wake cycles [5]. Technology can also facilitate maladaptive behaviors such as compulsive internet use and gambling disorder [6]. Patients with psychotic disorders have expressed that online platforms can exacerbate their paranoia and delusions through misinformation or targeted harassment [7]. Within OCD, the increasing digitization of society over the past several decades has provided new mediums for the manifestation of symptoms [8]. Recent case reports describe patients with OCD whose symptoms have been exacerbated by technology use. One individual described obsessions about inappropriately reacting to social media posts and would consequently spend hours per day scrolling through her feed to confirm that she had not mistakenly done so. Another patient worried that she had posted shameful content online and would screen record her internet use to later verify appropriate use [9].

However, advances in digital health have also introduced a wide range of technologies for the assessment and management of OCD. Passive monitoring tools, such as actigraphy and wearable devices, have been used to unobtrusively track sleep and activity patterns, often aligning closely with patient reports [10]. Active approaches, including ecological momentary assessment, allow for real-time symptom capture through smartphone prompts, offering insights into daily fluctuations that may be missed by retrospective recall [11]. Beyond monitoring, emerging interventions incorporate smartphone-based apps, cognitive training tools, and even multimodal systems that combine neurophysiological measures with behavioral data [12].

There exists a need to further understand patients’ perspectives toward technology and how its design or features might interact with their OCD. It remains unclear what forms of technology are more likely to trigger obsessions or compulsions, how these symptoms manifest, and whether patients find such interactions distressing or relieving. This qualitative study aims to address these questions by examining the mechanisms through which technology may contribute to or reinforce obsessive-compulsive behaviors.

Methods

Study Design: Constructivist Grounded Theory

We conducted a phenomenological study aimed at understanding the ways in which individuals with lived experience of OCD uniquely interface and interact with technology. Methodology was guided by constructivist grounded theory, which frames analysis as dialogic and iterative, constructing an empirically grounded model of the phenomenon under study [13]. We used the COREQ (Consolidated Criteria for Reporting Qualitative Research) framework, a validated checklist for qualitative reporting, to guide and document important aspects of the research team, methodology, findings, and analysis [14]. The study was approved by the University of Southern California Institutional Review Board (UP-23-01094).

Recruitment and Sample

Participants were recruited through a convenience sampling method between May and December 2024 through online posts in OCD-related communities, including the International OCD Foundation (IOCDF) website, OCD SoCal (a local affiliate of the IOCDF), and existing OCD-focused research participant registries maintained by the laboratory. These online spaces included both moderated forums and institutional web pages aimed at providing information, treatment resources, or research participation opportunities for those with OCD. Recruitment materials included a standardized study announcement containing the study overview, eligibility requirements, and compensation details. Potential participants were routed to contact the study team via email for screening through a REDCap (Research Electronic Data Capture; Vanderbilt University) survey that gathered demographic and self-report data relevant to the inclusion criteria. Study personnel reviewed survey responses in order of submission and contacted participants who met study criteria to schedule an initial virtual meeting to confirm survey responses and obtain informed consent. Following completion of the study, participants were compensated with a US $50 gift card for their time and effort. Inclusion criteria included self-reported age of 18 years and older and a diagnosis of OCD. A total of 27 participants completed the consent process and were interviewed. All interviews were reviewed by study personnel trained in the phenomenology of OCD for content consistent with knowledge of and experience with OCD. A total of 3 interviews were excluded due to credible concerns regarding eligibility (eg, no clear indication of an OCD diagnosis and suspected repeat participation by a single individual). To mitigate this risk, we

XSL•FO
RenderX

subsequently implemented procedures requiring camera-on interviews and a brief discussion of participants' OCD symptoms with the interviewer prior to proceeding. No participants withdrew or declined participation after enrollment. A total of 7 participants had previous contact with laboratory personnel through other unrelated research studies.

Research Team

Data collection and analysis were conducted by a multidisciplinary team, including the principal investigator (ACF; MD, PhD, psychiatrist) and graduate and undergraduate research assistants. Multimedia Appendix 1 provides an overview of team demographics, training, and specific roles. Interviewers disclosed their role in the laboratory to participants at the beginning of each interview. Team members brought varied clinical, academic, and personal perspectives related to OCD and technology and had differing levels of qualitative research experience.

To ensure consistency and competency, all interviewers completed a minimum of 6 weeks of qualitative research training. Standardization efforts included review sessions of the interview guide, observation of senior interviewers, and postinterview debriefings.

Data Collection and Setting

Interviews were conducted via the HIPAA (Health Insurance Portability and Accountability Act)-compliant Zoom software (Zoom Communications) between May 16, 2024, and December 20, 2024. A primary facilitator led each interview, with additional team members present to assist with technical support or occasional follow-up questions. Only study participants and research team members were present during interviews. Interviews ranged from 35 to 65 minutes, with a median duration of 51 minutes.

All interviews were audio- and video-recorded, and transcripts were generated using Zoom's auto-transcription feature. Research assistants verified transcript accuracy and performed deidentification. All audio and video recordings were permanently deleted after transcript verification. No repeat interviews or member checking of transcripts were conducted.

Interview Guide

The semistructured interview guide was developed to explore participants' lived experiences with technologies used for health or health care, particularly in relation to OCD. The guide was informed by phenomenological and user-centered perspectives, emphasizing discovery, motivation, use patterns, benefits, concerns, and perceived interactions with OCD symptoms. While not based on a single theoretical model, the structure reflects elements common to sociotechnical and technology adoption frameworks, with a focus on how individuals make sense of and navigate these tools in daily life [15,16]. Interview questions were piloted within the research team and refined through early interviews as part of our grounded theory approach. Revisions were made iteratively as team discussions and early analysis revealed new areas of interest. A sample of the interview guide is provided in Multimedia Appendix 2.

Data Analysis: Coding

The first 5 transcripts were independently open-coded by pairs of researchers, where open-coding refers to the initial process of breaking data into discrete parts and assigning conceptual labels to segments of interest. Each pair met to discuss initial findings and reach consensus on a shared set of open codes for each transcript. These were then reviewed through group meetings to identify patterns and groupings, generating a set of focused codes informed by emerging theoretical insights that constructed the initial codebook. The coding structure is outlined in Multimedia Appendix 3. Transcripts were uploaded to Dedoose (version 9.2.22; Sociocultural Research Consultants), and the codebook was implemented in the software. The codebook was then applied to each transcript by a pair of researchers, and consensus meetings between coders followed to ensure agreement on code applications.

We operationalized theoretical saturation in our dataset by the inclusion of a "new finding" code in the codebook to capture data relevant to the research question but not represented by existing codes. After all transcripts were coded, the "new finding" excerpts were reviewed collectively to assess whether additional conceptual categories had emerged. New categories were tracked and subsequent emergence of related insights was collapsed, such that "new finding" applications represented only unique and uncaptured theoretical insights [17,18]. We considered interview number 15 the point of theoretical saturation as no subsequent "new finding" applications identified categories or relationships that would meaningfully alter the emerging theory (Multimedia Appendix 4). These discussions led to final adjustments to the codebook. All transcripts were then recoded retrospectively using the updated coding structure.

Data Analysis: Theory Generation

After applying the final codebook, the team engaged in an inductive analytic process to develop a theoretical framework describing a salient emerging topic from the data: understanding interactions between technology and OCD symptoms. Using constant comparison across transcripts and memos, we identified key conceptual categories, examined their interrelationships, and developed an articulation of the processes shaping participants' experiences.

Analysis began with targeted review of excerpts coded with conceptually related codes ("Exacerbating and enabling OCD symptoms with technology" and "Alleviating mental health and OCD symptoms with technology"). These were used to write analytical memos, which guided weekly team discussions and supported theoretical elaboration. Through this iterative process, we developed a conceptual framework that articulates relationships between specific types of technology, the symptom domains they interact with, and distinct stages of the OCD symptom cycle.

Ethical Considerations

The study was approved by the University of Southern California Institutional Review Board (UP-23 - 01094). All study procedures involving human participants adhered to the ethical standards of the institutional review board and the 1964 Declaration of Helsinki and its subsequent amendments. All

XSL•FO
RenderX

subjects received written information regarding study aims, procedures, potential risks, and anticipated benefits, which were reviewed in discussion with study personnel during the onboarding process. All participants provided informed consent before study enrollment. Participation was voluntary, and participants could withdraw at any time without penalty. Participant privacy and confidentiality were strictly protected, and all data were deidentified before analysis. Participants were compensated with a US $50 Tango gift card upon study completion.

Results

Participants

A description of the participants (n=24) is provided in Table 1. No participants refused to participate in the study or withdrew their participation.

Table . Participant demographics (N=24).

Characteristics	Results
Age (years), median (range)	26.2 (20.1 - 63.6)
Gender, n (%)	
Men	7 (29)
Women	16 (67)
Nonbinary	1 (4)
State, n (%)	
California	14 (58)
Pennsylvania	3 (13)
Illinois	2 (8)
Ohio	2 (8)
Florida	1 (4)
Maryland	1 (4)
Virginia	1 (4)

A coding tree with 5 parent codes and 7 child codes was developed during analysis. Among these, the parent codes Exacerbating and enabling OCD symptoms with technology and Alleviating OCD symptoms with technology were most central to addressing the research question and guided the organization of findings. From this coding structure, we identified four pertinent qualities of technology in relation to participants' symptomatic experiences: (1) information provision, (2) gamification and quantification, (3) notifications, and (4) user interfaces. Information-provision technologies had strong interactions with OCD symptoms consistent with the disturbing-thought domain, whereas the 3 other technology characteristics had interactions with both disturbing-thought and symmetry domains. Notably, the application of OCD symptom domains (eg, symmetry and disturbing–thought-based) was not deductively imposed during the interview process or initial codebook development. Rather, these domains were inductively derived from participants' narratives and symptom descriptions. While the interactions between technological characteristics and symptom manifestations emerged organically from the data, the terminology used to describe these domains is drawn from the existing OCD literature to ensure consistency and scientific clarity in reporting [2,3].

Result 1: Information-Provision Technologies and the Disturbing-Thought Domain of OCD

Across interviews, participants described information-providing technologies as implicated in the emergence and reinforcement of symptoms within the disturbing thought-based domain of OCD. These technologies, which range from social media and search engines to health portals and messaging apps, triggered intrusive, ego-dystonic thoughts, while also enabling compulsive behaviors such as checking, reassurance-seeking, and mental review.

Many participants emphasized the role of unsolicited and algorithm-driven content in triggering disturbing-thought symptoms. These triggers were often delivered unexpectedly via video thumbnails, pop-up ads, or headlines, bypassing user intention and presenting emotionally charged or ambiguous material. This dynamic was especially distressing for individuals with harm, health, or moral scrupulosity themes:

> *It just pops up on your page. It can be very triggering….I started getting videos of this one specific food that I really like… This patient that was eating this food a lot developed fatty liver disease and my OCD brain was like, 'oh no'… I've been spending hours every day since seeing that, you know, researching correct dietary choices, researching health.* [P3]

The above quote also highlights the role of information-providing technology in inducing compulsive checking or reviewing to gain certainty or reassurance in response to obsessional distress. In our data, obsessional distress was related to intrusive fears of past wrongdoing, undiagnosed illness, or social harm. Commonly implicated platforms used

XSL•FO
RenderX

for checking or reviewing included search engines, social media timelines, court databases, and health information portals:

> *I will get on my phone…look on social media for like dates and look up people and you know, think that I hurt this person or did I do this? Did I do that?…I'll look my name up on the sex offender registry and I'll be like, oh no, you know, am I on here?…all that information being out there just makes me want to search and search more.* [P2]

Technologies that allowed participants to upload information for others to interact with produced similar patterns in both triggering disturbing-thought obsessions and enabling compulsions. Written communication platforms such as email and messaging apps were particularly salient for participants with social or moral scrupulosity, a subtype of disturbing-thought OCD. Participants reported that the informational permanence of these technologies often led to obsessive worries about their perception by others, including harm, misunderstanding, or being morally inappropriate. Participants then described compulsively reviewing sent messages, analyzing others' tone, and fearing reputational or ethical harm:

> *Yes. I had, I guess I had some OCD or some anxiety anyway about the Teams chat at work that we use for workplace communication…I would go back and check to see…could this be taken [as] inappropriate by this person? And I'll, you know, check, check a few times.* [P1]

Participants who used information-providing technology compulsively often acknowledged that the boundaries between reassurance-seeking, educational inquiry, and compulsive overuse are often opaque:

> *I just compulsively look up health care information…I think initially like all compulsions, when they're carried out, it provides some reassurance, although that's not always healthy.* [P1]

In some cases, participants reported that an obsession-compulsion cycle would subside once they encountered information that felt sufficiently reassuring. However, the threshold for what constituted a "sufficient" or trustworthy source varied across individuals. Both participants P8 and P9 described using search engines to manage health-related obsessions but differed in their sense of reassurance from the information they found:

> *So, I guess just try to, like, put it out of my mind until I can actually talk with someone from the office. So just, like, distract myself, or I'll find something on the internet that will maybe explain it, if I feel like it, is the answer. But usually I'm never like. Usually I'm never satisfied, because, like different websites can say different things. So I mean, that can sometimes work like, if I feel like, okay, I have a reliable website, and you know it says this. Then I'll feel like I found, you know I have the answer or the explanation to it.* [P8]

> *So like, some websites are more trusted than others…But again…I have trust issues deep down. And so the information I'm getting, I don't feel confident in even though I go to town. As soon as I see it, I kind of feel better because, you know, it answers my question. Like I just got a question answered, which is an anxiety, like to not know…Then it's like, I need to keep going and look, look, look, look. And I go to multiple different lengths. So I guess there's really not a point where I feel confident in it. And because the point is that my brain goes into more and more and more possibilities…And then I end up just having a bunch of different things that could be wrong with me. Um, and then I message my doctor and…that's pretty much how that goes.* [P9]

The duality of relief and exacerbation that informational technologies evoke was complex and individualized. The type of clarifying or validating information was also varied and included peer connection, digital monitoring, and access to authoritative sources. Another nuanced experience of participants regarded their deliberate limiting or avoidance of technology platforms perceived as triggering. For some, this involved active efforts to reduce exposure to social media or other content-driven apps:

> *Yeah, social media, like, I have Instagram, kind of only of the social medias, and it's a constant fight. I try to use it as little as I can, down to like, I delete the app every few days…That interacts for sure with my OCD.* [P7]

However, participants also reflected on the complexity of avoidance itself. While reducing exposure to distressing content was sometimes framed as protective, others were careful to distinguish between harmful avoidance (which can reinforce OCD) and what they considered healthy boundary-setting:

> *I guess if we want to talk about media… I try to avoid the amount. And I shouldn't really say 'avoid,' I guess, 'cause that makes it sound like counterproductive avoidance, but more like a self-care kind of avoidance, if you will…I will kind of doom spiral and ruminate and obsess about things like that if I focus on it too much…I just wanted to clarify, not be counterproductive. As we all know, like, avoiding things to heal your OCD doesn't work. That just makes it worse in the long run. So that's why I don't mean like hiding away from it, I mean just having my own healthy boundaries with things.* [P5]

Across both positive and cautionary accounts, participants consistently emphasized how the structure and function of information-providing technologies, whether through unsolicited content or user-driven inquiry, directly shaped the phenomenology of their disturbing-thought–based OCD symptoms.

XSL•FO
RenderX

Result 2: Gamified and Quantified Technologies in Symmetry- and Disturbing-Thought–Based OCD

Several participants described how technologies that quantify behavior or incorporate gamified features—such as streaks, scores, and progress tracking—interacted with their OCD symptoms. These interactions most commonly aligned with symmetry- or NJR-based obsessions. In these cases, discomfort was often triggered by an unbalanced number, an incomplete goal, or an interrupted streak and was accompanied by a vague but compelling internal sense that something was unresolved. Several participants expressed this interaction with health and fitness tracking apps:

> *Sometimes, you are too addicted to getting the perfect [step count] scores…I need to achieve this target, although you might have been drained out and tired throughout the day. But you want to. You want that number to be complete in order to just, like, tick off. It feels like ticking off one of the checkboxes.* [P13]

Participants frequently acknowledged that the distress did not always stem from a clearly articulated fear or consequence. Instead, their behaviors were driven by an ambiguous internal pressure, with goals or numbers taking on a rigid symbolic weight:

> *I would say I'm kind of obsessed with trying to have at least the same or more number of steps every day… I didn't think about that, but that's absolutely true… So I have like, let's say 10,000 or 11,000 steps a day and when I see that it's less, I'm like, oh, no, you know, I need to go for a walk and just go get some groceries just to make it even neater…I check it very often… so that gives me some anxiety probably you know internally…I can't figure out, I mean, the thoughts behind that, like the obsession…* [P15]

The automatic provision of a step count number to the user both initiates distress and enables individuals to perform checking compulsions. This highlights the ambiguity often described by participants in relation to NJR feelings with a fixation on numbers and achievements. One participant powerfully captured how app-based gamification and quantification intersected with OCD perfectionism in a way that borders on overwhelming:

> *Anytime when there's a streak, like you've meditated 9 times in a row, or like the way the apps are kind of gamified. The worst for this is Duolingo…just to give an example where it's like, you have all the cauldrons of stars and bonuses and streaks, and to me, that adds a layer of gamification and stress and numbers and scoring. And related to my OCD perfectionism, that's not as constructive to me…I would say it interacts…when it seems to be a thing I can pass and fail at. Particularly because there are numbers involved and or a kind of like, gamified system…when it feels like that kind of contest…It makes it more complicated.* [P7]

While symmetry-related themes were most common in response to these technologies, some participants described gamified or quantified features as triggering distress related to disturbing-thought–based obsessions. In these cases, the compulsion was not driven by a need for balance or completion, but rather by intrusive fears about consequences if the behavior was not completed:

> *I used the Headspace app to just…calm the thoughts…, but it ironically turned into kind of an OCD pattern where I felt like I couldn't not do the Headspace thing…I remember I had like a 130-day streak, which is great. But I realized that there were some days where I didn't want to do it…[My OCD] kind of left me in a pattern where it was like, you need to do it, or you're gonna fall back, or you're gonna retreat to your old self…The first therapist I met with was the one who kind of helped me be like, it's okay if you miss a day or two—like, the world's not gonna end.* [P14]

This development of intrusive fears and compulsive use of a mental health app underscores both the seemingly innocuous sources of symptom onset and the dual potential of mental health technologies to be either helpful or harmful, depending on how they interact with a user's symptom profile.

Some participants note that there are qualities of quantification technology that exacerbate obsessive-compulsive symptomatology in ways that nondigital alternatives might not:

> *MyFitnessPal was the one where I realized, okay, I need to stop using this…everything I ate, I had to put in. I spent a bunch of time making sure I found as accurate as possible of a submission for what I ate that day…because if you're just journaling, you're not going to see your progress towards this [amount of calories], like if you've gone over by this amount…So it was just very automatic and specific. And I think that's kind of what led to it. Versus if I was journaling, I don't think I would have been as likely to be tracking it so tight.* [P11]

> *I can find myself like, really obsessed with my reading speed and the percentage that I finish a book. So, you can track both your reading speed and the percentage that you finish a book on Kindle versus, you know, paper, and I can find myself feeling stressed or, like, feeling unable to do anything else unless I complete my reading goal. Like, I make my reading goal to reach like 5% of book, for example. And I feel like I cannot stop reading until that's done and that's only on kindle. It's such a quantified thing.* [P4]

Together, these accounts illustrate how features like quantification, gamification, and automated feedback can interact with both symmetry-based and disturbing-thought-based OCD symptoms, transforming everyday technologies into sources of obsessional triggers and compulsive pressure.

Result 3: Notifications and the Urge to Clear, Know, and Control

Participants frequently described notifications as a trigger for compulsive interactions that spanned both NJR- and disturbing-thought–based OCD experiences. Participants described two distinct ways notifications interacted with OCD

XSL•FO
RenderX

symptoms. For some, notifications triggered a need for visual or numerical resolution such as clearing unread counts to zero or removing banners that disrupted the screen, which aligned with NJR experiences. For others, notifications provoked disturbing-thought–related fears about neglecting responsibilities or missing important information, reinforcing compulsive checking behaviors.

For participants whose distress aligned with symmetry-based experiences, notifications disrupted a visual or emotional sense of order:

> *I also have to make sure all the notification numbers on the apps are cleared every time I check it…my wallpaper is a graduation photo with my parents. If a notification pops up and covers my mom, I have to remove it to make sure the important parts of the photo are not covered.* [P6]

> *I get really overwhelmed by my computer and my phone and in relation to my work. It's a space of like, Slack messages that I'm obsessive about, and clearing my inbox to 0 and, you know, checking my email again and again… It becomes like interface [number] 13, that I need to have like cleared and sort of dealt with.* [P7]

For this participant, the presence of notifications created a sense of visual clutter or incompleteness that became intolerable, prompting frequent checking and clearing rituals.

Other participants described notification-related distress that aligned more closely with disturbing-thought–based OCD, particularly around fears of neglecting responsibilities or missing something socially or professionally important. The spontaneity and unpredictability of notifications compounded the sense of being overwhelmed.

> *I just feel a bit burdened about it. Oh, I have a lot of notifications to clear, I need to catch up on, just so that I'm informed. Oh, okay, all of these are there, and I've read through them. So at least I have the knowledge of what was in my notifications.* [P10]

This compulsion of checking and clearing reflects a drive to restore a sense of control by accounting for and resolving the uncertainty that notifications generate.

Result 4: User Interfaces and Experiences of Ordering and Overwhelm

Participants described how the interfaces of personal devices frequently triggered obsessive-compulsive symptoms. Features typically celebrated for enhancing usability, such as customization, integration, and expansive functionality, were often experienced as overwhelming or destabilizing. For individuals with OCD, these features presented countless opportunities to engage in compulsions related to control, ordering, and avoidance, particularly when the interface felt overly stimulating or difficult to contain. These experiences aligned with both symmetry and disturbing-thought–based OCD symptom domains.

For example, some participants described compulsive ordering rituals related to app layout and usage patterns:

> *So, in my OCD, there's a certain way I do things, like there's a certain way I wash my hands. So similarly for social media, I have a certain order in which I use the apps and a certain order and amount of time. On every app there is a routine for it, like, for example, Whenever I wake up, I'll first open Instagram, then LinkedIn, then Handshake, then maybe Snapchat. So, there is a particular order.* [P10]

While certain participants engaged compulsively with interfaces in ways that clearly aligned with symmetry-based symptoms, such as imposing order, others described distress in response to the layered, limitless nature of digital environments. These experiences were characterized by feelings of cognitive overload or a sense of losing control. However, the underlying sources of this distress were more difficult to categorize. For some participants, it appeared to reflect a symmetry-based desire for structure, simplicity, or containment. A participant described this experience while engaging with cloud-based storage systems:

> *If it's my file, it's just abstract. It's weirdly hard for me how everything is virtual, just like not having that order…I work with one organization where it's all through Google Drive. There are hundreds and hundreds of sub folders…Getting lost in these online labyrinths somehow overwhelms my OCD. I struggle when I can't, like, control. And maybe this speaks to apps in terms of simplicity of interface for me. I think that would be more soothing to my OCD, because it wouldn't have that effect like, "this is a thing with like 10 subfolders, and so many different places to explore"…Yeah, if there's too much to it, it can be overwhelming.* [P7]

This sense of digital overload prompted some participants to deliberately restrict the functionality of their devices or to seek out simplified alternatives. The goal was not merely to reduce screen time, but to avoid the sense of overstimulation caused especially by vast integration of technology. A participant reflected on the difference between two wearable devices, seemingly driven by a disturbing-thought fear such as missing important information or falling behind:

> *There's so many different ways to reach me. So that is something that I get very, internally stressed and overwhelmed with…That's one of my main symptoms, is that feeling of always having to catch up and keep up with all the side conversations [on my phone]…The Apple watch reflected my computer, my phone, everything, on my wrist, which was just too much for me. I find the Fitbit to be less invasive because I haven't fully integrated my whole phone into it. I really enjoy the Fitbit for the limited functions that I've got on it right now. I know that it has more capabilities, but I like that I haven't turned those on.* [P12]

Across interviews, participants described how the boundless, integrated nature of digital interfaces transformed their devices into overwhelming, even invasive, objects. Abundance itself became a source of distress as the increasing complexity and

XSL•FO
RenderX

functionality of user interfaces created countless opportunities for engagement that could both facilitate compulsions and trigger anxiety. For some, these technologies reinforced ordering or avoidance rituals; for others, they produced a chaotic sense of overstimulation and loss of control. Several individuals noted limiting or simplifying device functions as a self-protective strategy to reduce compulsive urges and restore a sense of control.

Discussion

Principal Findings

Our findings show that everyday technologies shape the expression, maintenance, and phenomenology of OCD in ways that are both domain-specific and cross-cutting. Information-providing technologies often amplified disturbing-thought symptoms by delivering unexpected or ambiguous content and by enabling compulsive reassurance-seeking, checking, and mental reviewing. Participants described a fluid boundary between appropriate information gathering and compulsive overuse, noting that algorithmic content, the permanence of written communication, and the abundance of online information all contributed to escalating cycles of uncertainty and distress.

Gamified and quantified technologies interacted with symmetry- and NJR-based symptoms by attaching symbolic weight to numbers, streaks, and completion metrics. These features provoked both an internal drive for balance or perfection and, in some cases, intrusive fears about the consequences of breaking a streak. Participants described how automatic feedback loops transformed neutral metrics into triggers for compulsive behavior.

Across symptom domains, notifications acted as potent cues for urges to clear, control, and know. Some participants experienced notifications as visual disorders that required resolution, while others felt compelled to check notifications out of fear they might overlook or miss essential information. These experiences were reflective of both symmetry-based obsessions and disturbing-thought-based obsessions.

Finally, the structure of the user interface itself emerged as a significant source of OCD-related distress. Highly integrated or complex digital environments elicited compulsive ordering, avoidance, or feelings of cognitive overload. Many participants sought simplified devices, simplified digital environments, or limited device functionality to regain a sense of control and reduce compulsions.

Together, these results illustrate how contemporary technologies deliver content and structure interactions in ways that directly shape OCD phenomenology. The design of digital environments, including informational density, automated feedback systems, and levels of integration across platforms, creates conditions in which obsessions can be triggered and compulsions reinforced. These findings point to technology as an active context in which OCD symptoms unfold, rather than a neutral backdrop. This perspective helps clarify why existing research has only begun to capture the breadth and nuance of these experiences, and it provides a foundation for situating our results within the emerging literature.

Previous Research

There is a relative dearth of literature exploring stakeholder perspectives on how OCD interacts with technology, both as a mental health disorder and a lived experience. Previous work has highlighted how symptoms could manifest in the digital realm, including intrusive worries about posting offensive content to the Internet or social media, with associated compulsive checking, screen capture, and reassurance seeking [9]. Expert opinion now suggests assessing for "digital obsessions and compulsions," when evaluating patients with OCD [8]. A number of studies that used technology in the assessment and treatment of OCD also asked participants about their experiences using this technology; however, this information was not collected systematically and remained broadly focused on acceptability and tolerability [19-21]. Finally, one study explored experiences with online peer-support communities in individuals with OCD; thematic analysis revealed factors such as social comparison and misinformation as contributing to negative experiences [22].

This study sought to explore how individuals with OCD perceive and navigate their interactions with modern technology. Notably, neither our interview guide nor codebook was structured to map onto formal OCD dimensions such as those delineated by the Yale-Brown Obsessive Compulsive Scale Symptom Checklist [23]. Yet, clear and consistent patterns emerged that aligned with established symptom clusters, most notably the disturbing-thought (eg, harm, sexual, religious, and moral scrupulosity) and NJR domains. Participants described experiences that mirrored clinical subtypes, underscoring the validity of these dimensions and the extent to which they shape real-world behavior, including digital behavior.

Interestingly, symptoms in the domains of cleanliness, contamination, and excessive or ritualized washing or grooming were less prominent in our data. This absence may reflect that contamination concerns often center on physical contact with contaminants and health-related concerns generally have a focus on the physical body. However, mental contamination is a recognized phenomenon in OCD and shares features with contact contamination symptoms and symmetry and incompleteness symptoms [24]. While this was not present in our data, interactions between technology and mental contamination should be considered for further exploration. Finally, our findings complicate assumptions that certain domains are purely physical. For example, other studies have found that digital clutter can provoke similar emotional distress and difficulty discarding as does physical hoarding [25].

Below, we elaborate on the implications of these findings for 4 key stakeholder groups: patients, clinicians, technology developers, and researchers.

Implications for Patients

Technology can serve as an important source of community, a method for accessing educational and treatment resources, and an approach to monitoring symptoms for individuals with OCD. However, these same tools can exacerbate symptoms and

XSL•FO
RenderX

complicate treatment. Patients may benefit from intentionally monitoring how digital tools affect their emotional state and symptom patterns over time. We recommend that individuals reflect on the distinction between purposeful use of technology and patterns of behavior that drift into compulsive checking, reassurance seeking, or overmonitoring.

In practical terms, this may involve setting time- or context-based limits, pausing engagement with certain platforms during periods of heightened vulnerability, or discussing emerging online behaviors with a therapist to identify early signs of compulsive use. Patients should also be mindful that customization features, achievement systems, and constant access to information can shift from motivating to destabilizing without clear warning. We suggest open and ongoing discussions between patients and their treating clinicians regarding technology in daily life and OCD treatment.

We suggest framing technology not as inherently helpful or harmful, but as a continuum of use that can shift depending on symptom severity, emotional needs, and the design of specific platforms. Approaching digital engagement with this flexible, self-reflective mindset may help patients cultivate healthier, more sustainable relationships with the technologies that shape their daily lives.

Implications for Clinicians

For clinicians, assessment of OCD symptomatology is important, and we concur with previous studies that recommend explicit evaluation for digital behaviors [8]. This can be achieved by adding targeted questions (eg, “Do you feel compelled to check, post, or search online to relieve distress, even when it interferes with your life?”) or by supplementing existing measures with a brief checklist addressing technology-related compulsions, such as repetitive searching, reassurance-seeking, or monitoring notifications. Because current instruments do not systematically probe for these behaviors, clinicians should remain attentive to technology use that functions as a compulsion. It also remains unclear what proportion of OCD symptoms can be attributed uniquely to digital compulsions, and future studies should quantify this domain to inform updates of assessment tools.

With regard to treatment, clinicians should incorporate psychoeducation about how app design, social media features, or online communities can inadvertently reinforce obsessive-compulsive cycles. We suggest clinicians provide concrete examples to patients: streaks and push notifications can induce checking behaviors, social media can reinforce negative stereotypes and foster misinformation, and user interfaces can lead to a feeling of digital overwhelm. These examples can help patients recognize their own experiences and better understand these processes as environmental triggers rather than personal failings. Clinicians might also help patients identify strategies for more adaptive engagement, such as disabling notifications, setting time limits, or designating “offline” periods.

Cognitive-behavioral strategies, particularly exposure and response prevention, can be readily adapted to digital contexts. For instance, therapists might collaborate with patients to create graded exposure hierarchies that include delaying responses to notifications, resisting the urge to refresh a feed, or intentionally interrupting a streak or posting schedule. These exercises can be integrated into therapy sessions or assigned as structured homework.

Finally, clinicians should maintain flexibility as technology evolves. Integrating a brief review of digital behaviors into routine follow-up visits can help identify emerging risks, prevent relapse, and reinforce adaptive digital habits. Tracking these observations in deidentified case notes may also contribute valuable data for refining diagnostic instruments and treatment protocols.

Implications for Technology Developers

Our findings raise important considerations for those who design and develop digital platforms. Many features intended to enhance engagement, such as algorithmic personalization, goal-setting interfaces, streak counters, and push notifications, were described by participants as direct triggers of OCD symptoms.

Rather than removing these features entirely, developers could implement optional design accommodations that give users greater control over their interaction patterns. For example, adjustable interface settings could allow users to disable streak counters, hide quantified feedback (eg, progress bars and daily goals), or consolidate notifications into scheduled digests. Content algorithms might include transparency dashboards where users can view or modify personalization parameters, thereby increasing predictability and reducing uncertainty-driven compulsions. Similarly, introducing a “low-stimulation” or “minimal-feedback” mode could help reduce inadvertent reinforcement of repetitive checking behaviors.

An “OCD-friendly” design, in this context, would prioritize user autonomy, predictability, and control over feedback loops that otherwise promote compulsive engagement. These accommodations could be incorporated without removing core functionality but by expanding the range of interaction options available to all users.

Finally, the mental health technology sector, particularly apps targeting wellness, mindfulness, or productivity, could benefit from evaluating whether their engagement metrics inadvertently reinforce obsessive-compulsive symptom cycles. Collaborating with clinicians, behavioral scientists, and individuals with lived experience of OCD could help establish design standards that promote sustained engagement without exacerbating compulsive behaviors.

Implications for Researchers

This study opens several avenues for future research. There is a clear need to develop and validate new measurement tools that specifically assess the intersection of OCD symptoms and technology use. These tools could help clinicians differentiate between typical digital habits and pathological compulsions, especially in younger or digitally native populations.

Future studies might also explore how different OCD subtypes respond to specific forms of technology in more controlled, quantitative settings. For example, are individuals with NJR symptoms more sensitive to gamified or goal-based interfaces

XSL•FO
RenderX

than those with contamination concerns? Are certain design features (eg, infinite scroll and intermittent rewards) more strongly associated with symptom exacerbation in OCD than in other clinical populations?

Finally, longitudinal studies can address several outstanding questions. First, for individuals with a diagnosis of OCD, what is the time course and progression of technology-related symptoms? Cohort studies of individuals with OCD that include careful assessment for this symptom domain could improve our understanding of the evolution of technology-related obsessions and compulsions. Relatedly, our results demonstrate positive and negative aspects of technology use in adults with OCD; understanding how, and over what period, a neutral or beneficial use of technology transitions into problematic or compulsive use can aid in assessment and evaluation in clinical settings, as well as inform technology design. Finally, tracking technology use in the general population and monitoring for transition to compulsive use could help in identifying specific aspects of the technology and person that may predispose them to these behaviors. Studies in this domain could also assess for conversion to meeting formal OCD diagnostic criteria and evaluate progression of broader OCD symptoms in this population. Overall, understanding these dynamics will be essential as technology continues to evolve and embed itself more deeply into everyday life.

Limitations

This study has several limitations that should be considered when interpreting its findings. First, the demographics of our participants skewed young and were geographically concentrated in California. The median age of participants was 26.2 (range 20.1 - 63.6) years, 14 of the 24 participants resided in California, and 10 participants resided in other states. This geographic concentration likely reflects our recruitment strategies. Seven participants were drawn from other studies conducted at the University of Southern California that required in-person participation in Los Angeles. Additional participants were recruited online through local and national chapters of OCD advocacy organizations. The relatively young age of our sample may, in part, reflect these online recruitment strategies, as previous research has shown that younger adults spend more time online than older adults [26]. Other studies using online recruitment methods have also reported similarly young average participant ages [27-29]. The convenience sampling strategy we used may have contributed to this demographic profile. Regardless of the age profile and geographic distribution of our sample, qualitative research is designed to explore experiences, perspectives, and contexts in depth, rather than to produce findings that are statistically generalizable to a larger population. Its value lies in uncovering rich, detailed insights and generating understanding of complex phenomena, which can then inform theory, practice, or further research.

Second, our participant sample needed a baseline level of technological fluency. All participants were able to access digital recruitment materials, use Zoom for interviews, and speak fluently about their technology use. As a result, the findings may not fully capture the experiences of individuals with lower digital literacy, limited access to technology, or differing generational relationships with digital tools. Individuals for whom technology is so distressing that they cannot consistently use the internet or Zoom would also not be captured in our study.

Additionally, while participants self-identified as having a diagnosis of OCD, we did not conduct formal clinical assessments or structured diagnostic interviews to confirm a diagnosis of OCD or assess specific symptom subtypes. As such, some participant narratives may reflect overlapping symptomatology with related conditions, such as generalized anxiety. However, the use of self-reported diagnosis is an established approach in qualitative mental health research, where the analytic focus is on lived experience, identity, and help-seeking rather than diagnostic validation. Qualitative studies routinely use a self-reported mental health diagnosis as an inclusion criterion for participation, with these diagnoses ranging from OCD to depression to bipolar disorder [30-33]. In line with these precedents, our sampling strategy is appropriate for experiential inquiry, while still requiring cautious interpretation regarding the boundaries between OCD and overlapping or comorbid forms of distress. Additionally, research personnel reviewed transcripts for content consistent with obsessive and compulsive behaviors.

Our analysis was guided by a constructivist grounded theory approach, and results reflect the perspectives and interpretive lens of a research team composed primarily of nonclinicians. While this multidisciplinary team allowed for rich interpretive dialog, the extrapolation of findings was based on analytic interpretation that considered the existing literature on OCD but did not rely on a structured, validated metric. Specifically, aligning participant accounts with disturbing-thought–based or symmetry-based domains was an interpretive process shaped by reflexivity of researchers. Additionally, apparent ambiguity in the data may reflect other barriers to participant disclosure, such as limited insight or the stigmatization of OCD symptoms, rather than the absence of an experience.

Despite these limitations, the study provides novel insights into how individuals with OCD perceive and engage with technology and offers a foundation for future research to more systematically examine these dynamics.

Conclusions

This study highlights how digital technologies can both trigger and sustain OCD symptom patterns. Participants' accounts reveal that the design of these systems often mirrors OCD's own dynamics of uncertainty, reassurance-seeking, and control, blurring the boundary between pathology and platform. Clinically, these findings underscore the importance of assessing technology use as part of symptom formulation and helping patients develop strategies for intentional, rather than compulsive engagement. For designers and digital health developers, we highlight the need for interfaces that minimize reinforcement of compulsive behaviors and allow for user control over triggering features. More broadly, understanding OCD within its technological context reframes it as a disorder increasingly expressed through the architectures of modern attention, emphasizing the shared responsibility of clinicians,

XSL•FO
RenderX

researchers, and technologists in shaping environments that support, rather than exploit, cognitive vulnerability.

Acknowledgments

We would like to thank Dr Rachel Ceasar for invaluable consultation on initial formulations of our interview guide and study protocol. We would like to thank the following research assistants for their effort in data collection, interview transcription, and focused coding: Natalie Bakken, Trevor Bailey, Jenna Kim, Dejan Shakya, and Zoe Elliot. Emma Garland assisted with formatting, copy editing, and manuscript review.

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: literature search and systematization, proofreading and editing, and reformatting. The GenAI tools used were: GPT-4, GPT-4.5, and OpenEvidence. Responsibility for the final manuscript lies entirely with the authors. Gen AI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted with collective responsibility.

Funding

This work was funded by a Brain and Behavior Research Foundation Young Investigator Award to ACF (32015). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: ACF
Data curation: LO, ACF
Formal analysis: LO, KSP, KP, HW, DM, EL, ES, MM, TB, TV, UP, ACF
Funding acquisition: ACF
Investigation: LO, KSP, KP, EL, ES, MM, TB, TV, UP, ACF
Supervision: LO, UP, ACF
Writing – original draft: LO, KSP, KP, HW, DM
Writing – review and editing: ACF

Conflicts of Interest

None declared.

Multimedia Appendix 1
Study team demographics.
[DOCX File, 1811 KB - jmir_v28i1e85033_app1.docx]

Multimedia Appendix 2
Example semistructured interview guide.
[DOCX File, 32 KB - jmir_v28i1e85033_app2.docx]

Multimedia Appendix 3
Sample coding tree.
[DOCX File, 1809 KB - jmir_v28i1e85033_app3.docx]

Multimedia Appendix 4
Table of thematic saturation.
[DOCX File, 1808 KB - jmir_v28i1e85033_app4.docx]

References

1. Ruscio AM, Stein DJ, Chiu WT, Kessler RC. The epidemiology of obsessive-compulsive disorder in the National Comorbidity Survey Replication. Mol Psychiatry 2010 Jan;15(1):53-63. [doi: 10.1038/mp.2008.94] [Medline: 18725912]

XSL•FO
RenderX

2. Bloch MH, Landeros-Weisenberger A, Rosario MC, Pittenger C, Leckman JF. Meta-analysis of the symptom structure of obsessive-compulsive disorder. Am J Psychiatry 2008 Dec;165(12):1532-1542. [doi: 10.1176/appi.ajp.2008.08020320] [Medline: 18923068]
3. Mataix-Cols D, Rosario-Campos MD, Leckman JF. A multidimensional model of obsessive-compulsive disorder. Am J Psychiatry 2005 Feb;162(2):228-238. [doi: 10.1176/appi.ajp.162.2.228] [Medline: 15677583]
4. Cervin M, Miguel EC, Güler AS, et al. Towards a definitive symptom structure of obsessive-compulsive disorder: a factor and network analysis of 87 distinct symptoms in 1366 individuals. Psychol Med 2022 Oct;52(14):3267-3279. [doi: 10.1017/S0033291720005437] [Medline: 33557980]
5. Nagata JM, Lee CM, Hur JO, Baker FC. What we know about screen time and social media in early adolescence: a review of findings from the Adolescent Brain Cognitive Development Study. Curr Opin Pediatr 2025 Aug 1;37(4):357-364. [doi: 10.1097/MOP.0000000000001462] [Medline: 40172268]
6. Monteith S, Glenn T, Geddes JR, et al. Importance of patient online activities. Br J Psychiatry 2025 Oct;227(4):718-720. [doi: 10.1192/bjp.2025.32] [Medline: 40369901]
7. Torous J, Bucci S, Bell IH, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct;20(3):318-335. [doi: 10.1002/wps.20883] [Medline: 34505369]
8. Carmi L, Zohar J, Arush OB, Morein-Zamir S. From checking the door to checking the app: assessment and treatment implications for obsessive-compulsive disorder in the digital era. CNS Spectr 2021 Oct;26(5):457-458. [doi: 10.1017/S1092852920001509] [Medline: 32600488]
9. van Bennekom MJ, de Koning PP, Denys D. Social media and smartphone technology in the symptomatology of OCD. BMJ Case Rep 2018 Aug 28;2018:bcr2017223662. [doi: 10.1136/bcr-2017-223662] [Medline: 30154174]
10. Drummond LM, Wulff K, Rani RS, et al. How should we measure delayed sleep phase shift in severe, refractory obsessive-compulsive disorder? Int J Psychiatry Clin Pract 2012 Oct;16(4):268-276. [doi: 10.3109/13651501.2012.709866] [Medline: 22809128]
11. Gloster AT, Richard DCS, Himle J, et al. Accuracy of retrospective memory and covariation estimation in patients with obsessive-compulsive disorder. Behav Res Ther 2008 May;46(5):642-655. [doi: 10.1016/j.brat.2008.02.010] [Medline: 18417100]
12. Arevian AC, O'Hora J, Rosser J, Mango JD, Miklowitz DJ, Wells KB. Patient and provider cocreation of mobile texting apps to support behavioral health: usability study. JMIR Mhealth Uhealth 2020 Jul 29;8(7):e12655. [doi: 10.2196/12655] [Medline: 32723714]
13. Charmaz K. Constructivist grounded theory. J Posit Psychol 2017 May 4;12(3):299-300. [doi: 10.1080/17439760.2016.1262612]
14. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
15. Mohr DC, Schueller SM, Montague E, Burns MN, Rashidi P. The behavioral intervention technology model: an integrated conceptual and technological framework for eHealth and mHealth interventions. J Med Internet Res 2014 Jun 5;16(6):e146. [doi: 10.2196/jmir.3077] [Medline: 24905070]
16. Rahimi B, Nadri H, Lotfnezhad Afshar H, Timpka T. A systematic review of the technology acceptance model in health informatics. Appl Clin Inform 2018 Jul;9(3):604-634. [doi: 10.1055/s-0038-1668091] [Medline: 30112741]
17. Hennink MM, Kaiser BN, Marconi VC. Code saturation versus meaning saturation: how many interviews are enough? Qual Health Res 2017 Mar;27(4):591-608. [doi: 10.1177/1049732316665344] [Medline: 27670770]
18. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. Soc Sci Med 2022 Jan;292:114523. [doi: 10.1016/j.socscimed.2021.114523] [Medline: 34785096]
19. Rupp C, Falke C, Gühne D, Doebler P, Andor F, Buhlmann U. A study on treatment sensitivity of ecological momentary assessment in obsessive-compulsive disorder. Clin Psychol Psychother 2019 Nov;26(6):695-706. [doi: 10.1002/cpp.2392] [Medline: 31365952]
20. Olbrich H, Stengler K, Olbrich S. Smartphone based Geo-Feedback in obsessive compulsive disorder as facilitatory intervention: a case report. J Obsessive Compuls Relat Disord 2016 Jan;8:75-78. [doi: 10.1016/j.jocrd.2016.01.001]
21. Klein CS, Alt AK, Pascher A, et al. Cognitive behavioral therapy for pediatric obsessive-compulsive disorder delivered via internet videoconferencing: a manualized sensor-assisted feasibility approach. Child Adolesc Psychiatry Ment Health 2024 Dec 4;18(1):154. [doi: 10.1186/s13034-024-00844-7] [Medline: 39633405]
22. Tan YT, Rehm IC, Stevenson JL, De Foe A. Social media peer support groups for obsessive-compulsive and related disorders: understanding the predictors of negative experiences. J Affect Disord 2021 Feb 15;281:661-672. [doi: 10.1016/j.jad.2020.11.094] [Medline: 33234279]
23. Goodman WK, Price LH, Rasmussen SA, et al. The Yale-Brown Obsessive Compulsive Scale. I. Development, use, and reliability. Arch Gen Psychiatry 1989 Nov;46(11):1006-1011. [doi: 10.1001/archpsyc.1989.01810110048007] [Medline: 2684084]

XSL•FO
RenderX

24. Jacoby RJ, Blakey SM, Reuman L, Abramowitz JS. Mental contamination obsessions: an examination across the obsessive-compulsive symptom dimensions. J Obsessive Compuls Relat Disord 2018 Apr;17:9-15. [doi: 10.1016/j.jocrd.2017.08.005]
25. Luxon AM, Hamilton CE, Bates S, Chasson GS. Pinning our possessions: associations between digital hoarding and symptoms of hoarding disorder. J Obsessive Compuls Relat Disord 2019 Apr;21:60-68. [doi: 10.1016/j.jocrd.2018.12.007]
26. Gelles-Watnick R. Americans' use of mobile technology and home broadband. Pew Research Center. 2024 Jan 31. URL: https://www.pewresearch.org/wp-content/uploads/sites/20/2024/01/PI_2024.01.31_Home-Broadband-Mobile-Use_FINAL.pdf [accessed 2025-11-11]
27. Milczarski W, Borkowska A, Białek M. No evidence of risk aversion or foreign language effects in incentivized verbal probability gambles. Judgm decis mak 2025;20:e17. [doi: 10.1017/jdm.2025.3]
28. Gregory SEA. Investigating facilitatory versus inhibitory effects of dynamic social and non-social cues on attention in a realistic space. Psychol Res 2022 Jul;86(5):1578-1590. [doi: 10.1007/s00426-021-01574-7] [Medline: 34374844]
29. Le TP, Bradshaw BT, Pease M, Kuo L. An intersectional investigation of Asian American men's muscularity-oriented disordered eating: associations with gendered racism and masculine norms. Eat Disord 2022;30(5):492-514. [doi: 10.1080/10640266.2021.1924925] [Medline: 33998395]
30. Staiger T, Stiawa M, Mueller-Stierlin AS, et al. Masculinity and help-seeking among men with depression: a qualitative study. Front Psychiatry 2020;11:599039. [doi: 10.3389/fpsyt.2020.599039] [Medline: 33329149]
31. Morton E, Hole R, Murray G, Buzwell S, Michalak E. Experiences of a web-based quality of life self-monitoring tool for individuals with bipolar disorder: a qualitative exploration. JMIR Ment Health 2019 Dec 4;6(12):e16121. [doi: 10.2196/16121] [Medline: 31799936]
32. Arnáez S, Roncero M, López-Santiago J, et al. Fighting against self-stigma in adults with self-reported diagnosis of OCD: a single-arm pilot study using a mobile app-based intervention. Br J Clin Psychol 2025 Sep;64(3):788-805. [doi: 10.1111/bjc.12537] [Medline: 40065191]
33. Wairauch Y. Compulsive rituals in obsessive-compulsive disorder - a qualitative exploration of thoughts, feelings and behavioral patterns. J Behav Ther Exp Psychiatry 2024 Sep;84:101960. [doi: 10.1016/j.jbtep.2024.101960] [Medline: 38513433]

ABBREVIATIONS

COREQ: Consolidated Criteria for Reporting Qualitative Research
HIPAA: Health Insurance Portability and Accountability Act
IOCDF: International OCD Foundation
NJR: not-just-right
OCD: obsessive-compulsive disorder
REDCap: Research Electronic Data Capture

Edited by A Stone; submitted 30.Sep.2025; peer-reviewed by H Maheshwari, R Marshall; revised version received 19.Nov.2025; accepted 17.Dec.2025; published 05.Feb.2026.

Please cite as:
Occhino-Moede L, Sulivan-Pascual K, Phelan K, Wang H, Mokhtar D, Liu E, Schug E, Mirkis M, Baek T, Visher T, Pasupulety U, Frank AC
Interactions of Technology and Obsessive-Compulsive Disorder Symptomatology in Adults: Qualitative Interview Study
J Med Internet Res 2026;28:e85033
URL: https://www.jmir.org/2026/1/e85033
doi:10.2196/85033

© Lucas Occhino-Moede, Kaitlyn Sulivan-Pascual, Kendall Phelan, Harrison Wang, Daniel Mokhtar, Elisa Liu, Erica Schug, Megan Mirkis, Thomas Baek, Tamerlane Visher, Ujjwal Pasupulety, Adam Charles Frank. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessing Usage and Usability of a Narrative-Based Psychoeducational Digital Intervention to Improve Medication Adherence Among Individuals With Schizophrenia in a Stable Phase: Mixed Methods Study

Dian Zhu[1*], PhD; Fangyuan Chang[2*], PhD; Hongyi Yang[2], PhD; Yiwen Wei[2], MEng; Zhao Liu[2], PhD

[1]School of Art Design and Media, East China University of Science and Technology, Shanghai, China

[2]School of Design, Shanghai Jiao Tong University, No.800, Dongchuan RD, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Zhao Liu, PhD
School of Design, Shanghai Jiao Tong University, No.800, Dongchuan RD, Shanghai, China

Abstract

Background: Nonadherence to antipsychotic medication remains one of the most substantial challenges in the management of schizophrenia, contributing to relapse, rehospitalization, and functional decline. Although psychoeducational interventions are a key intervention for relapse prevention, traditional formats often lack interactivity and cultural resonance, thereby limiting engagement and sustained impact. Digital health innovations offer an opportunity to improve both treatment adherence and user experience, but evidence in schizophrenia populations remains limited.

Objective: This study aimed to evaluate the usage patterns, usability, and effectiveness of a narrative-based psychoeducational digital intervention designed to enhance medication adherence among individuals with schizophrenia in the maintenance phase. By employing a mixed methods design, the study integrated quantitative measures of adherence and functioning with qualitative insights into participants' experiences and perceptions.

Methods: A 6-month parallel mixed methods randomized controlled trial was conducted in community mental health settings in Shanghai. Seventy individuals with schizophrenia in a stable phase were randomly assigned (1:1) to the intervention group, which received the digital narrative-based psychoeducation application (Healing Town) in addition to routine community care, or to the control group, which received routine community rehabilitation only. Quantitative evaluation focused on medication adherence, drug attitude, social functioning, and psychiatric symptoms. In parallel, qualitative data were collected through semistructured interviews with patients, caregivers, and clinicians to examine intervention usage, usability, engagement, and perceived impact.

Results: Seventy participants (mean age 44.2, SD 8.057 y; 61% male) were enrolled, and 69 (98.6%) completed the 6-month trial, with one dropout during the intervention period. At 6 months, the intervention group showed significantly higher medication adherence (mean difference 1.27, 95% CI 0.30 - 2.24; *P*=.02) and more positive drug attitudes (mean difference 3.41, 95% CI 1.18 - 5.65; *P*=.002) compared with controls. Improvements in social functioning were significant within the intervention group (*P*=.03) but not between groups. No significant group differences were observed in psychiatric symptoms. Qualitative findings identified three overarching themes: (1) adherence and usability—patients reported enhanced treatment knowledge, confidence, and motivation, though some described challenges with feedback tone and pacing; (2) experiences and attitudes—users valued cultural relevance, immersive narratives, and gamified elements but noted occasional overstimulation; and (3) expectations and recommendations—participants expressed demand for personalized features, reminders, and dynamic content to sustain engagement.

Conclusions: This mixed methods study provides preliminary evidence that a narrative-based digital psychoeducational intervention may enhance medication adherence and attitudes toward medication among individuals with schizophrenia in the maintenance phase, while being perceived as engaging, usable, and culturally relevant. Furthermore, the qualitative findings suggest that supportive feedback, adaptive difficulty, and personalized features may enhance user motivation and optimize future scalability. Overall, this narrative-based digital psychoeducation represents a promising and potentially cost-effective approach to supporting community-based psychiatric rehabilitation, meriting further longitudinal and multisite investigation.

Trial Registration: ClinicalTrials.gov NCT06175559; https://clinicaltrials.gov/study/NCT06175559

(J Med Internet Res 2026;28:e59175) doi:10.2196/59175

XSL•FO
RenderX

KEYWORDS

schizophrenia; psychoeducation; medication adherence; cognitive behavior therapy; mixed methods study

Introduction

Schizophrenia is a chronic and severe mental disorder characterized by pervasive impairments in cognition, emotion, perception, and social functioning. Long-term use of antipsychotic medication is typically required to manage symptoms, prevent relapse, and reduce the risk of hospitalization [1-3]. However, despite advances in pharmacological treatments, approximately 50% of outpatients experience suboptimal medication adherence, which contributes to symptom exacerbation, frequent hospitalizations, and an elevated risk of suicide [4-6]. Improving adherence is not only critical for therapeutic success but also essential for enhancing long-term functional outcomes and quality of life [7].

Patients' subjective understanding of their illness and treatment plays a pivotal role in medication-taking behavior. Misconceptions about drug efficacy or side effects, limited coping strategies, and insufficient illness insight may all diminish motivation to adhere to prescribed regimens. Psychoeducation, a core intervention aimed at enhancing illness awareness and promoting adherence, has been widely implemented in schizophrenia care. It has demonstrated effectiveness in supporting symptom recognition, self-management, and functional recovery. Nevertheless, traditional psychoeducation programs often rely heavily on didactic instruction and lack interactivity, making them less suited for individuals who may experience impairments in attention, memory, and metacognition [8]. Moreover, impaired insight—often associated with prefrontal dysfunction—can lead to resistance toward biomedical explanations, especially when these conflict with patients' personal experiences or cultural beliefs [9,10].

In recent years, digital psychoeducation has shown great promise in enhancing treatment adherence among individuals with mental health conditions. Its advantages lie in offering highly interactive, contextualized, and personalized feedback. By integrating cognitive-behavioral strategies, gamified elements, and multimedia delivery, digital interventions have improved the accessibility of health information while enhancing user engagement and behavioral follow-through [11-13]. A particularly promising development is the emergence of narrative-based digital psychoeducation, which has demonstrated unique potential in reshaping health beliefs, eliciting emotional resonance, and strengthening treatment motivation [14]. Rooted in narrative psychology, this approach emphasizes that when individuals engage emotionally and cognitively in realistic, emotionally rich scenarios and characters, their beliefs about health, attitudes toward illness, and behavioral intentions can be reconstructed through simulated experience and observational learning [15]. Narrative interventions have shown encouraging outcomes in other areas of health education, particularly in sustaining behavior change over time [16].

Although some recent interventions targeting individuals with schizophrenia have begun to incorporate immersive, graphical narratives to enhance engagement and comprehension, this area of research remains in its early stages [17]. As Keats once suggested, narrative analysis should integrate spoken, written, and visual modalities to enable multidimensional expressions of complex psychological experiences [18]. However, many existing digital interventions still emphasize information delivery and sensory presentation, with limited attention to how patients construct meaning, express themselves, or reshape their attitudes toward treatment through narrative. This lack of integration between narrative structures and individual psychological mechanisms may undermine the effectiveness of such interventions in fostering motivation and medication adherence. In the context of long-term schizophrenia management, medication adherence is not only a key predictor of relapse and prognosis but also a reflection of the individual's cognitive, emotional, and attitudinal engagement with treatment. Therefore, conventional educational strategies alone may fall short in activating patients' intrinsic motivation and participatory engagement.

To address this gap, we developed an interactive, narrative-based digital application and conducted a 6-month randomized controlled trial (RCT) combined with qualitative interviews. Using a mixed methods design, we integrated quantitative and qualitative findings to evaluate the usability and user experience of the narrative-based digital psychoeducational application among individuals with schizophrenia in the maintenance phase. In addition, we examined how these user-centered outcomes were related to key 6-month end points, including medication adherence, social functioning, and clinical symptoms.

Methods

Study Design

A mixed methods approach was adopted in this study [19,20] following the ethical guidelines for human research outlined by the American Psychological Association [21].

In January 2023, five experienced psychiatrists contributed to the story script design, while three individuals with schizophrenia evaluated the visual style. After finalizing the script and visuals, the project moved to the application development phase.

Participants were recruited after providing written informed consent and screened for eligibility. Potential participants were identified through the Shanghai Mental Health Information Management System. Community psychiatrists screened eligible individuals and provided study information during routine follow-up visits. Those who expressed interest were subsequently contacted by research staff, who obtained written informed consent. The primary intervention phase was conducted between March 2024 and October 2024, lasting 6 months in total. During this period, participants attended five supervised in-person sessions each week, in which they completed the narrative-based psychoeducation program under therapist guidance (Figure 1). Baseline and postintervention

XSL•FO
RenderX

assessments were conducted. At baseline, participants completed demographic surveys and medication adherence scales. All participants were trained to use the application developed by the research team. The same assessments were repeated at subsequent time points. After the intervention, structured interviews were conducted using standardized procedures. These included think-aloud protocols and semistructured interviews, with responses recorded and later digitized. Audio recordings were made for further analysis.

Figure 1. Study design of the mixed methods study to evaluate a narrative-based psychoeducational digital intervention. QUAL: qualitative; QUANT: quantitative.

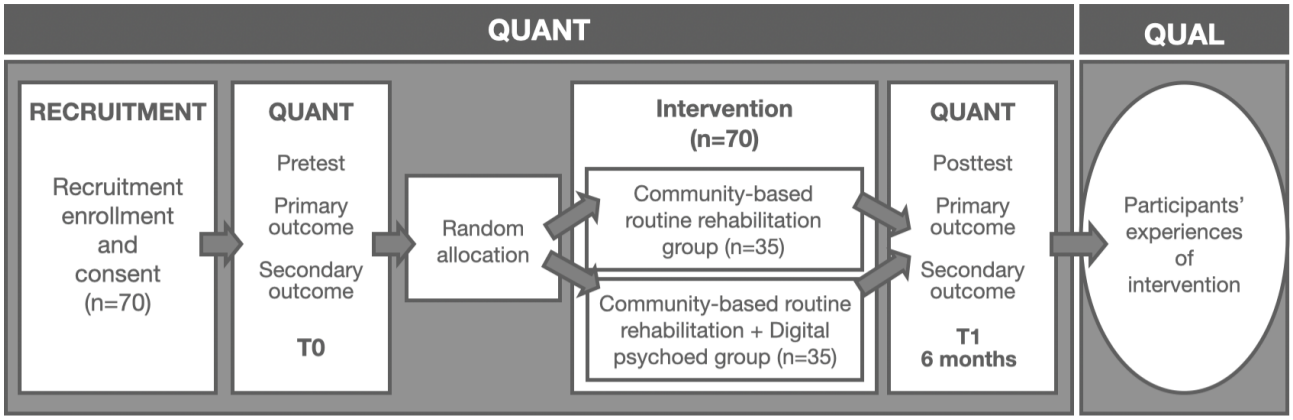

The quantitative component of this study was derived from a previously conducted parallel, two-arm RCT in Shanghai, implemented in accordance with the CONSORT (Consolidated Standards of Reporting Trials; Checklist 1) guidelines [22]. Participants were randomly assigned in a 1:1 ratio to either a community-based usual care group or a digital psychoeducational intervention group, with a 6-month intervention period. The usual care group received standard public health services, including regular home visits, participation in community mental health center rehabilitation programs, and education on maintaining medication adherence. In addition to these services, the intervention group participated in 5 weekly sessions of the narrative-based digital psychoeducation, each lasting 15 to 20 minutes. Each intervention site was staffed with one to two mental health professionals whose role was limited to resolving technical issues with the devices.

Following the 6-month quantitative assessment, a qualitative study was conducted in the intervention group. Semistructured interviews were carried out with all 35 users of the narrative-based digital psychoeducational intervention and 5 mental health professionals (with expertise spanning psychology, public health management, and nursing), ensuring that thematic saturation was achieved. Interviews took place between October 2024 and November 2024, conducted face-to-face in community mental health centers, and lasted an average of 15 minutes. All interviews were audio-recorded with consent, transcribed verbatim, deidentified prior to analysis, and guided by an interview protocol developed based on prior literature and expert consultation. The guide focused on usability, acceptability, user engagement, barriers and facilitators, as well as perceived impacts on medication-related behaviors.

Participant Recruitment

The study recruited individuals registered in Shanghai's community mental health system who had been diagnosed with schizophrenia and were receiving maintenance medication. The trial was conducted across 7 community rehabilitation centers administered by the Shanghai Mental Health Center. Shanghai has been recognized as a national leader in the management of schizophrenia, and digital technologies are increasingly being integrated into community-based services to support patient rehabilitation. Eligible participants were formally enrolled after providing written informed consent.

Stable-phase schizophrenia patients were selected as participants by psychiatrists from the Shanghai Mental Health Center. Trained psychiatrists conducted prospective monthly assessments of schizophrenia patients using the Positive and Negative Syndrome Scale (PANSS), with an evaluation period of at least 1 year. Patients were considered to have a stable condition if their total PANSS score did not change by more than 3 points and if there had been no changes in their medication treatment for at least 6 months prior to the neuropsychological assessment [23]. The inclusion and exclusion criteria are summarized in Table 1.

Participants were further required to be in a clinically stable phase, as determined by their PANSS scores.

XSL•FO
RenderX

Table . Inclusion and exclusion criteria for participant recruitment.

Domain	Inclusion criteria	Exclusion criteria
Registration and diagnosis	• Registered in the Shanghai Mental Health Information Management System • Diagnosed with schizophrenia according to the *International Classification of Diseases, 10th Revision* and determined to be in a clinically stable phase based on PANSS[a] assessment	• Planning to relocate outside of Shanghai • Presence of severe physical illness or organic brain disease • Comorbid with other psychotic disorders
Demographics and capacity	• Aged 18 - 60 y, with at least primary school education • Normal or corrected-to-normal vision and hearing • Ownership of and ability to independently operate a smartphone or other electronic device	• Participation in any intervention or treatment other than medication or basic public health services in the past 6 mo
Treatment status	• Receiving maintenance treatment with second-generation antipsychotics.	—[b]
Consent	• Participant or family member provided informed consent and signed the consent form.	—

[a]PANSS: Positive and Negative Syndrome Scale.

[b]Not available.

Sample Size Calculation

Sample size was calculated using G*Power 3.1 with a 2-tailed independent samples *t* test. Medication adherence scores were designated as the primary outcome measure. Based on a literature review and pilot study results, the expected between-group mean difference with a pooled SD was 0.78 (0.93), yielding an estimated effect size of approximately 0.83. With a significance level of α=.05 and a statistical power of 90%, the required minimum sample size was calculated to be 31 participants per group. Accounting for a 1:1 randomization ratio and a 20% dropout or loss to follow-up rate, the study required at least 76 participants (38 per group). Proportional sampling was conducted across districts based on the number of individuals with schizophrenia in each district.

Randomization, Allocation Concealment, and Blinding

After providing informed consent, participants were randomly assigned in a 1:1 ratio to the community-based routine rehabilitation group or the digital psychoeducation intervention group using a computer-generated random number table. Randomization ensured that baseline characteristics were evenly distributed between groups. These assessments aimed to evaluate the effectiveness of digital psychoeducation interventions in improving medication adherence and other health outcomes.

Intervention

Application Introduction

This study developed a graphic narrative-based digital psychoeducation application that integrates core concepts from narrative psychology and cognitive-behavioral theory. The intervention aims to enhance medication adherence and treatment motivation among individuals living with chronic-phase schizophrenia. Technologically, the application is grounded in interactive graphics, storyline-driven engagement, and behavioral tracking. It consists of three core modules: narrative storylines, cognitive training games, and self-monitoring logs. The intervention is set within fictional yet realistic life scenarios, allowing for high ecological validity. All user interactions—including choice selections, response times, and behavioral paths—are recorded in real time on the backend for subsequent behavioral analysis and feedback generation. The narrative pathways are dynamically open-ended, meaning that different choices lead to varying outcomes, thereby enhancing immersion and personalization (Multimedia Appendix 1).

Narrative Storytelling Psychoeducation

The narrative module is informed by the foundational principle of narrative psychology: individuals construct meaning, integrate experiences, and facilitate behavioral change through storytelling [24]. The application adopts a fictional framework wherein a "nonclinical character supports an individual experiencing schizophrenia," inviting users to engage in role-playing from a third-person perspective. Users observe the story's protagonist and assist them in making treatment-related decisions [25]. Although users are not directly narrating their own experiences, they project aspects of their own reality through simulated interactions, enabling externalization of internal struggles and fostering reappraisal of illness and self-identity. This indirect narrative mechanism encourages emotional involvement and character identification, aligning with the dual persuasive processes of *transportation* and *identification* proposed in narrative communication theory, which are known to support behavior change and motivational development [26,27].

XSL•FO
RenderX

To strengthen the educational impact and structural coherence of the storylines, the narratives were designed based on the classic 3-act structure and were iteratively refined under the guidance of psychiatric rehabilitation specialists and clinical psychiatrists. In total, 30 narrative threads were developed, each targeting specific objectives related to medication education (Multimedia Appendix 2).

Cognitive Games

The cognitive training game is designed to enhance individuals' ability to apply cognitive skills in daily life, thereby supporting functional recovery. Codeveloped by clinical psychiatrists and psychiatric rehabilitation experts, this module includes a series of real-world scenario-based tasks—such as grocery shopping or ordering food at a restaurant—that simulate daily challenges and promote skills such as information integration, planning, and decision-making (Multimedia Appendix 2). The games adopt an adaptive learning mechanism in which task difficulty is dynamically adjusted based on the user's performance. This ensures an optimal balance between challenge and attainability, thereby maintaining engagement and reinforcing learning outcomes.

Beyond strengthening executive function, attention, and problem-solving abilities, this module promotes the transfer of cognitive strategies to real-world environments, offering targeted support for rehabilitation in ecologically relevant contexts.

Self-Management Support

The self-management module enables users to schedule medication intake, access educational content related to medication adherence, and input personalized data—such as the daily dosage and timing of medications. The platform facilitates routine tracking of medication usage and provides accessible information on potential side effects and recommended coping strategies. It also includes educational materials highlighting the risks associated with medication nonadherence, including symptom relapse.

Through structured daily logs, users are encouraged to monitor their behavioral patterns, emotional states, and medication routines, thereby fostering self-awareness and self-regulation. This module supports the development of autonomous health behaviors, which are critical for sustaining long-term treatment engagement and improving functional outcomes.

Outcome Measures

Quantitative Outcomes

Morisky Medication Adherence Scale

The Morisky Medication Adherence Scale (MMAS) [28-31] comprises eight items. The total score ranges from 0 to 8, with higher scores indicating better medication adherence.

Drug Attitude Inventory

The Drug Attitude Inventory (DAI) [32] consists of 10 items, of which 6 are positively scored (correct answers score 1 point and incorrect answers score −1 point), while 4 are reverse-scored (correct answers score −1 point and incorrect answers score 1 point). Higher scores indicate a more favorable attitude toward medication.

Social Disability Screening Schedule

The Social Disability Screening Schedule (SDSS) [33] includes 10 items rated on a 3-point scale (0="no impairment," 2="severe impairment," and 9="not applicable"). The total score ranges from 0 to 20, with higher scores indicating greater social disability. The evaluation will be performed by trained psychiatric professionals at the time of assessment.

Brief Psychiatric Rating Scale

The Brief Psychiatric Rating Scale (BPRS) [34] evaluates 5 factors: anxiety and depression, anergia, thought disturbance, activation, and hostility or suspiciousness. It consists of 18 items scored on a 7-point Likert scale (1="not present" and 7="extremely severe"), with a total score ranging from 18 to 126. Higher scores indicate greater symptom severity. The evaluation will be performed by trained psychiatric professionals at the time of assessment.

Qualitative Outcomes

Participants engaged in semistructured and unstructured interviews accompanied by professional physicians and caregivers. These interviews aimed to explore participants' attitudes toward the application and its impact on medication adherence. Professional health care providers were also interviewed to evaluate the effectiveness of this kind of intervention. At the end of the intervention, patients were surveyed on their satisfaction with the psychoeducation based on several specific questions (Textbox 1).

XSL•FO
RenderX

Textbox 1. Semistructured interview guide used to explore usability and user experience of a narrative-based psychoeducational digital intervention.

Survey on the impact of psychoeducation on medicine adherence (participants)

1. What is the frequency at which the application is utilized?
2. Did you experience any moments during the course of this intervention when you felt the urge to quit up? Why?
3. What motivates or hinders you?
4. Has your medication usage changed as a result of the intervention? How so, if at all. If not, then explain.

Attitudes toward the digital psychoeducation (participants)

1. What is your opinions on such intervention measures?
2. Do you find narratives that are fundamental to be engaging?
3. Which of the chapters that you most impressive?
4. What aspect did you appreciate the most? (scenes, voice-overs, mini-digital psychoeducation interactions, narratives, and characters)

Evaluation of the effectiveness of the digital psychoeducation (mental health professionals)

1. Evaluate the level of patient allure exhibited by the intervention.
2. Kindly assess the efficacy in promoting adherence to medication.
3. In your opinion, which elements are linked to medication adherence? Why?
4. Would you recommend the intervention to your patients?

During each visit, brief semistructured interviews were conducted and audio-recorded to evaluate participants' attitudes toward the digital psychoeducation. During the baseline visit, interview questions focused on general feedback regarding the digital psychoeducation, including its narrative and visual elements. After the intervention, these questions were revisited, along with additional questions addressing specific digital psychoeducation features, such as the self-management check-in system.

Additionally, backend application usage data were collected, including the date and time of app usage and the story paths chosen within the intervention. Researcher YW analyzed each participant's experimental sessions by linking the time of app usage to their final story path selections.

Data Analysis

Data Integration

A sequential explanatory mixed methods design was adopted. Quantitative data from the RCT were first analyzed to assess changes in medication adherence, attitudes, and clinical outcomes. Subsequently, qualitative interviews were conducted to explore participants' experiences and to explain the quantitative results. Integration occurred during the interpretation phase through a triangulation process, in which findings from both data sources were compared and merged to identify areas of convergence, complementarity, and divergence.

Quantitative Analysis

All quantitative survey data were analyzed using SPSS version 26.0 (IBM Corporation). Categorical variables were presented as frequencies and percentages, while continuous variables were expressed as means (SD). Comparisons of categorical variables between groups were conducted using χ^2 tests or Fisher exact tests. For continuous variables, independent sample *t* tests or Mann-Whitney *U* tests were used for between-group comparisons, depending on data distribution. Within-group comparisons were performed using paired *t* tests or Wilcoxon signed rank tests. Baseline characteristics between the intervention and control groups were compared using chi-square tests or independent sample *t* tests. The effects of the intervention were evaluated by comparing baseline and postintervention data using independent sample *t* tests or nonparametric alternatives. All statistical tests were 2-tailed, and a *P* value of <.05 was considered statistically significant. No missing data were observed in this study.

Qualitative Analysis

The transcripts were analyzed in NVivo (version 12; QSR International) through an inductive form of content analysis [35]. At the beginning of the analysis, DZ and FC extracted text segments from 35 participants' transcripts to ensure consistency in coding. They then turned to analyze the rest of the transcripts separately. The records were initially coded to gain insights into the participants' thoughts and behaviors during the intervention process. This involved line-by-line review, extraction of relevant text segments, and identification of codes (n=49), with a focus on participants' attempts at taking action and their choices. Subsequently, subthemes (n=27) related to different attitudes toward adopting digital psychoeducation to enhance compliance were identified through discussions with participants and experts. The authors then collectively discussed the codes and subthemes, further refining them into broader themes (n=3). Examples of data extraction are listed in Multimedia Appendix 2. This approach allowed us to discern participants' varying attitudes toward the digital psychoeducation and their implementation of medication practices and processes. Any discrepancies that emerged during the analysis were discussed with the rest of the authors. To ensure the validity of the codes, the authors employed member checking by reaching out to

XSL•FO
RenderX

participants to validate interpretations of the data [36]. Specifically, comprehensive results were provided to participants to verify the consistency and accuracy of their experiences with the data.

Ethical Considerations

This study was conducted in accordance with the ethical principles of the Declaration of Helsinki and the American Psychological Association's guidelines for research with human participants. The research protocol was reviewed and approved by the Shanghai Jiao Tong University Human Research Ethics Committee (H20230207I). The trial was prospectively registered on ClinicalTrials.gov (NCT06175559). No exemption from ethical review was sought or applied.

All participants, or their legal guardians when appropriate, were provided with detailed verbal and written information about the study procedures, potential risks, and benefits. Written informed consent was obtained prior to enrollment. For participants with limited decision-making capacity, consent was additionally confirmed by a family member or caregiver. The informed consent process covered both primary data collection and the use of deidentified data for secondary analyses without requiring further consent. To protect privacy and confidentiality, all personal identifiers were removed from the study records. The digital psychoeducational application logged only anonymized behavioral data, which were stored in encrypted servers with access restricted to the research team. Interview transcripts were deidentified prior to the analysis. Participants received a gift card valued at approximately US $30 upon completing the study as compensation for their time and effort.

Results

Descriptive Analysis

From March 2024 to October 2024, a total of 82 community-dwelling individuals with schizophrenia undergoing rehabilitation were recruited. Following screening and assessment, 3 participants did not meet the inclusion criteria, and 7 were excluded due to their transition to or planned transition to long-acting injectable medications. The final expected enrollment number was 72 participants (Figure 2). Based on the residential communities of these 72 participants, local intervention sites were identified. Two participants voluntarily withdrew from the study, citing the distance to the intervention site, leaving 70 participants (35 in the community-based rehabilitation group and 35 in the digital psychoeducation intervention group), who completed baseline assessments. During the intervention period, no participants dropped out, and all 70 participants completed the postintervention assessment. The exceptionally low dropout rate can likely be explained by the enforcement of national health policies in China, which require individuals with a confirmed diagnosis of schizophrenia to regularly attend community rehabilitation services as part of their standard care.

Figure 2. CONSORT (Consolidated Standards for Reporting Trials) flow diagram of participant recruitment of a narrative-based psychoeducational digital intervention. BPRS: Brief Psychiatric Rating Scale; DAI: Drug Attitude Inventory; MMAS-8: 8-item Morisky Medication Adherence Scale; SDSS: Social Disability Screening Schedule.

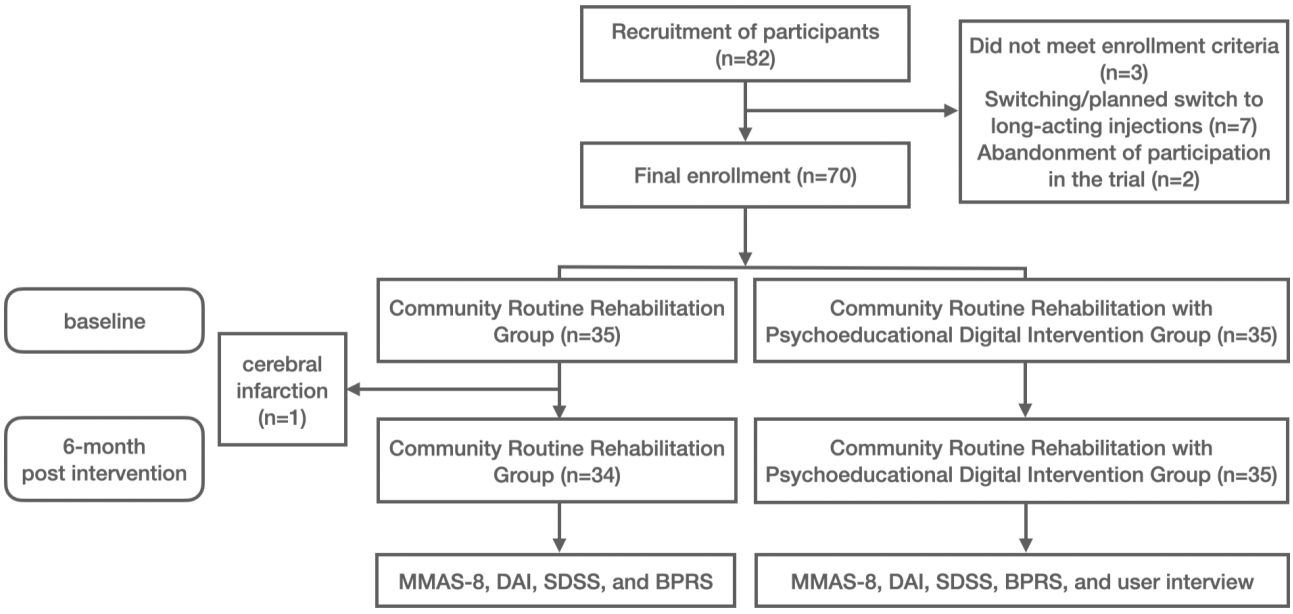

Characteristics of the Study Participants

A total of 70 participants were included in the study, with 43 male patients and 27 female patients. The mean (SD) age of the participants was 44.20 (8.057) years. The distribution of demographic characteristics is shown in Tables 2 and 3. The participants were randomly assigned to either the control group or the intervention group, with 35 individuals in each group. Independent samples *t* test results showed no significant difference in age between the 2 groups ($P>.05$). Chi-square test results indicated no significant differences between the two groups in terms of gender, education level, marital status, or living situation ($P>.05$). Baseline data are provided in Multimedia Appendix 3.

XSL•FO
RenderX

Table . Baseline demographic and clinical characteristics of participants with stable-phase schizophrenia.

Participants	All (N=70)	Control group (n=35)	Experimental group (n=35)	χ^2/*t* test[a] (df)	*P* value
Sex, n				0.54 (1)	.46
Male	43	20	23		
Female	27	15	12		
Age (y), mean (SD)	44.20 (8.06)	44.29 (9.10)	11 (6.99)	0.09 (68)	.93
Education, n				3.29 (2)	.19
Primary school	2	0	2		
Middle school	45	21	24		
College or above	23	14	9		
Marital status, n				0.51 (2)	.77
Married	11	5	6		
Divorced and widowed	10	6	4		
Unmarried	49	24	25		
Residence status, n				0.22 (1)	.64
Living alone	5	3	2		
Living with family	65	32	33		

[a]*t* tests were used for continuous variables (eg, age), and chi-square tests were used for categorical variables (eg, gender, education level, marital status, and residence status).

Table . Demographic characteristics of two participating mental health professionals.

Mental health professionals	Major	Gender	Age (y)	Years of experience
D1	Nursing	Female	39	15
D2	Public health management	Female	32	6
D3	Public health management	Male	41	12
D4	Psychology	Male	30	2
D5	Psychology	Female	42	15

Analysis of Intervention Results

This study evaluated differences between the intervention and control groups across multiple outcome measures, including medication adherence, attitudes toward medication, social functioning, and clinical symptoms (Table 4). The intervention group demonstrated significant improvement in medication adherence. At postintervention, a statistically significant between-group difference was observed, with a mean difference of 1.27 (95% CI 0.30-2.24; *P*=.02). Regarding attitudes toward medication, the intervention group also showed a significant improvement compared to the control group, with a mean difference of 3.41 (95% CI 1.18-5.65; *P*=.002).

XSL•FO
RenderX

Table . Changes in medication adherence, medication attitude, social functioning, and clinical symptoms from baseline to 6 months in a randomized controlled trial.

Characteristics	Intervention group, mean (SD)	Within-group *P* value	Control group, mean (SD)	Within-group *P* value	Mean difference (95% CI)	Between-group *P* value
Medication adherence		.03[a]		.11	1.27 (0.30 to 2.24)	.02[a]
Pre	6.91 (1.66)		7.23 (1.22)			
Post	7.66 (0.49)		6.68 (2.05)			
Medication attitude		<.001[b]		.47	3.41 (1.18 to 5.65)	.002[a]
Pre	4.97 (3.64)		5.83 (3.12)			
Post	8.91 (1.90)		6.35 (3.89)			
Social functioning		.03[a]		.85	−1.91 (−8.55 to 4.73)	.34
Pre	2.20 (2.83)		2.20 (2.79)			
Post	1.06 (1.97)		2.88 (3.68)			
Clinical symptoms		.06		.66	−7.88 (−34.90 to 19.14)	.12
Pre	28.37 (16.04)		25.77 (11.08)			
Post	21.37 (3.49)		26.68 (8.48)			

[a]*P*<.05.

[b]*P*<.001.

Qualitative Feedback and User Experience

Among the secondary outcomes, participants in the intervention group exhibited a notable within-group improvement in social functioning (*P*=.03). However, no statistically significant between-group differences were found in social functioning or clinical symptom severity (*P*>.05).

Application Usage (Back-End Data)

Analysis of back-end data indicated that 34 of 35 (97.14%) participants completed all game content. However, participants exhibited relatively high error rates in decisions related to certain topics, highlighting areas for targeted education in future interventions.

Alcohol Consumption and Medication

Although 25 of 35 (71.43%) participants selected the correct response—“Alcohol should be avoided, as it interacts with antipsychotic medications causing central nervous system depression, and excessive drinking may trigger other severe drug side effects”—a notable proportion (13/35, 37.14%) chose incorrect options, such as “All medicines are toxic, better not take them at all” or “Life is short, it’s fine to skip once in a while.” This suggests persistent misconceptions regarding alcohol–medication interactions.

Medication When Going Out

Some participants (8/35, 22.86%) initially chose unsafe options when deciding whether to take medication while going out, including “Skipping occasionally is fine” or “It’s okay to skip while traveling.” After experiencing negative story consequences, participants corrected their choices to the appropriate response: “Ensure to bring and take medications on time to avoid missed doses.”

Medication Side Effects

Most participants (31/35, 88.57%) correctly responded to in-story side effects by selecting “Consult a professional promptly if experiencing discomfort.” Nevertheless, a subset (7/35, 20%) initially chose the incorrect response, “It’s just lack of exercise, so I can endure it.”

Overall, these back-end data reveal both high engagement with the application and specific knowledge gaps, particularly regarding alcohol use, medication adherence while traveling, and interpretation of side effects. These findings underscore key areas for future educational emphasis.

Result of the Interview

Overview

The quantitative analysis revealed that the intervention significantly improved medication adherence and attitudes toward medication compared with standard community rehabilitation. The qualitative findings complemented these results by illustrating how narrative immersion and interactive gameplay enhanced users’ motivation and engagement, thereby reinforcing the behavioral changes observed in the quantitative phase. The demographic and clinical characteristics of the respondents are shown in Table 2. Thematic saturation was achieved after approximately 25 interviews, when no substantially new codes emerged. The remaining 10 interviews confirmed and enriched existing themes, particularly regarding variations in family support and illness duration. The specifics of the qualitative findings are displayed in Multimedia Appendix 4. Three main higher-order themes emerged from the participants’ and experts’ accounts (Multimedia Appendix 5).

XSL•FO
RenderX

Adherence and Usability

This theme encompassed factors related to medication adherence and the usability of the narrative-based digital intervention. Adherence was strengthened through increased medication knowledge, enhanced confidence in recovery outcomes, and active engagement with interactive narratives. Participants generally reported that the narrative intervention promoted adherence by enhancing knowledge about medication and building awareness of illness consequences, going beyond the functions of traditional reminder tools. This cognitive restructuring process appeared to activate intrinsic motivation. One participant explicitly noted that the intervention content changed their pattern of avoiding medication: "I used to avoid taking my medication due to side effects. After using this app, I realized that failing to reach the prescribed dose could have real consequences. Now I remind myself more consistently to follow the plan." Another participant stated, "I used to reduce doses or skip medication on my own… but now I feel more capable of managing my condition and trusting my treatment plan."

Positive reinforcement, such as the perceived "increase in medication knowledge" and "improvement in health status" after completing each narrative scenario, significantly enhanced participants' self-efficacy and encouraged continued engagement. Additionally, the intervention's self-management interface, which clearly displayed daily progress in medication-taking and task completion, provided ongoing external motivation: "Every time I log in, I feel like I'm working toward my health, which really motivates me." However, feedback tone and clarity were sensitive factors that could influence continued engagement. When participants made incorrect choices, the perceived harshness of outcomes sometimes triggered overstimulation and negative emotions, leading to tension and feelings of being scolded ("A bad ending made me angry, and I didn't understand why I was wrong"), highlighting the importance of strategic and tactful feedback.

Experiences and Attitudes Toward the Digital Psychoeducation

This theme revealed that participants expressed positive attitudes toward the narrative-based intervention, particularly appreciating its novelty and immersive quality. They described this learning approach as "not like a class, more like living inside a story," which significantly enhanced the learning experience. High acceptance was closely related to cultural sensitivity in the intervention, such as familiar home environments, clothing styles, and locally grounded story backgrounds. Participants noted, "The character design feels like people I really know… it is very relatable and comforting." Psychologically, the narrative also provided participants with important emotional connection and a sense of inclusion. Some participants reported feeling socially isolated in real life, but nonplayer characters within the intervention offered companionship and acceptance: "I always feel like people around me don't like me or want to talk to me… only they (in the app) are willing to talk to me."

Despite the overall cultural relevance of the narrative content, limitations remained at the individualized level. Some participants indicated that the narrative did not fully reflect their specific clinical or demographic characteristics, resulting in reduced relevance for certain content. One male participant noted, "Not all themes relate to my personal experiences, and not all side effects happened to me," suggesting that highly themed or specific content may create cognitive distance or misalignment with other user groups.

Expectations for the Digital Psychoeducation

This theme explored participants' suggestions regarding design elements and interactive features of the intervention, aimed at enhancing understanding of medication use. While initial experiences were positive, both participants and therapists emphasized the importance of ongoing content updates and functional completeness for maintaining long-term adherence. One participant remarked, "After completing all the stories, there's nothing new. I hope it could include more everyday situations, like working or traveling with friends." Due to the relatively simple content, one participant completed all narrative scenarios within 6 weeks, resulting in diminished interest later. The rapid consumption of content led some participants to discontinue usage after story updates ceased: "Towards the end, only the daily check-in feature was available, no new storylines. So, I just stopped engaging with the intervention."

From a functional perspective, participants repeatedly emphasized the need for external reminders, not only for new story releases but also for synchronized medication taking, requesting that the app "send notifications like other apps." Therapists highlighted the feasibility of future technological integration, suggesting "It would be great if AI could generate more interactive stories rather than having doctors spend time creating scripts," to ensure a continuous content supply and reduce the burden on clinical staff, thereby enabling sustainable integration of the system. Finally, some participants noted frustration with the lack of a review function: "Sometimes I have to interrupt the narrative due to other things happening, and I forget what happened. When I come back to the story, I realize that progress from before that level wasn't saved…" indicating the need to optimize memory and review mechanisms.

Discussion

Principal Findings

This study evaluated the usability, user experience, and behavioral effects of a narrative-based digital psychoeducational intervention developed for individuals with chronic schizophrenia in community settings. Using a mixed methods design, our findings suggest that, compared with standard community rehabilitation, the intervention significantly improved medication adherence and attitudes toward medication, while no significant changes were observed in clinical symptoms. Complementary qualitative findings further highlighted that participants valued the narrative elements for enhancing knowledge, motivation, and engagement, while also reporting challenges such as negative emotional reactions to corrective feedback. Most participants emphasized that maintaining long-term content engagement remained a concern, underscoring the need for future digital designs to incorporate personalized, accessible, and supportive features.

XSL•FO
RenderX

Specifically, following the narrative-driven digital psychoeducational intervention, participants exhibited notable improvements in medication adherence and medication attitude. Qualitative interviews provided contextual explanations for these quantitative findings, illustrating how the intervention influenced behavior through narrative mechanisms. Immersive storylines and interactive feedback enhanced participants' understanding of medication routines and their sense of control, leading to better adherence and fewer unplanned treatment interruptions. For example, some users mentioned that after completing the story, they "felt responsible for the character's health," reflecting a motivational shift from passive medication-taking to active self-management. This indicates that the digital format not only increased engagement but also fostered emotional resonance and cognitive involvement, supporting the observed behavioral improvements [37,38]. Care providers noted that integrating the intervention into repetitive cognitive rehabilitation sessions encouraged sustained participation in regular health-promoting behaviors. These improvements may be attributed to embedded mechanisms of emotional engagement and cognitive participation within the system [39]. Consistent with prior research, digital technologies such as smartphone applications have been found acceptable and feasible for individuals with psychotic disorders [40]. Existing mobile applications targeting medication adherence in schizophrenia mainly address logistical and memory-related barriers through reminders, dose tracking, and external reinforcement [41,42]. However, for individuals with schizophrenia, key barriers to treatment adherence are often internal, psychological, and cognitive, such as limited insight, misconceptions about side effects, perceived stigma, and reduced intrinsic motivation for long-term therapy [43]. Thus, interventions targeting patients' deeper cognitive representations of illness and treatment offer a promising direction for narrative-based approaches [44]. Immersive narratives provided a judgment-free, low-pressure environment in which participants could temporarily shift from the passive patient role to that of an active, rational decision-maker, engaging in simulated choices about medication and recovery. Many participants perceived the scenarios as familiar and personally relevant, which appeared to reduce resistance toward illness-related content and strengthen their willingness to engage in medication behaviors [41,42]. Although no significant between-group differences were observed in clinical symptoms, within-group improvements were identified. Qualitative narratives illuminated these patterns, as some participants described transient negative emotions during feedback or difficulties sustaining motivation over time, suggesting that behavioral and attitudinal improvements may precede measurable symptom changes [43,44]. Also, given the long-term trajectory of schizophrenia recovery, these findings collectively suggest that increased persistence and motivation could represent intermediate mechanistic changes that may facilitate long-term clinical benefits [45-47].

Regarding user experience and attitudes, most participants expressed positive feedback toward the modular narrative structure, describing the intervention as immersive and engaging. These qualitative findings complemented the quantitative results on adherence and medication attitude, suggesting that enhanced engagement and motivation may have mediated the observed behavioral improvements. By framing participants' self-representation as part of a healthy population, the interactive design reduced perceived stigma and promoted active knowledge acquisition and health behavior. Therapists noted that the intervention embedded medication-related knowledge through a sequence of stories, situational decision points, and real-time feedback, enabling participants to rehearse treatment-related decisions in a safe, low-pressure environment [48]. This narrative-based approach not only facilitated knowledge acquisition but also preserved users' autonomy and sense of realism. The usability and emotional engagement of the intervention appear to underpin the cognitive and motivational mechanisms that supported adherence improvement. As Stephens [49] emphasized, narratives serve not only as a means of organizing experience but also as a tool for identity construction and social positioning. By persuading a virtual character to "take medication on time," participants effectively constructed a socially valued identity—one characterized by cognitive competence and responsibility. This process represents a form of narrative repositioning, in which individuals psychologically shift from passive recipients of care to active agents of self-management, thereby enhancing perceived control and treatment motivation [50].

However, some users reported that certain narratives reflecting specific cultural or gendered contexts limited identification and engagement, highlighting the need for more personalized and culturally adaptive content in future iterations. Additionally, motivational activation was influenced not only by the narrative content but also by functional design elements such as pacing, reward systems, and feedback mechanisms. Participants reported confusion when storylines became overly abstract or fragmented, particularly during periods of cognitive fatigue, emphasizing the importance of clarity and simplicity in interface design for individuals with psychotic disorders. Thus, in addition to ensuring psychological relevance and cultural appropriateness, system design must prioritize simplicity and error tolerance [48]. Future development should continue to involve diverse stakeholders—including therapists, patients, and family members—for iterative discussion and codesign [51]. Mechanistically, participants indicated that the cognitive game components were seamlessly integrated with the narrative elements, enabling repeated practice in recognizing and reframing unhelpful automatic thoughts [42]. The task-based structure and immediate feedback combined behavioral activation with cognitive restructuring, helping participants modify maladaptive thinking patterns and rebuild medication-related cognitive frameworks. This mechanism aligns with the principles of cognitive behavioral therapy, which seeks to enhance adherence by reducing cognitive resistance and emotional conflict while strengthening behavioral consistency and motivation [52,53]. Finally, achievement-based feedback, milestone rewards, and task-driven structures created a sustained motivational environment that supported engagement even when immediate reinforcement was limited [54].

This mixed methods study provides preliminary evidence that the narrative-based digital psychoeducational intervention tested in this study demonstrated good usability and acceptability and may support short-term improvements in medication adherence

XSL•FO
RenderX

and attitudes toward medication among individuals with chronic schizophrenia. Although short-term symptom outcomes did not show significant improvement, the intervention addressed key behavioral and motivational processes critical for long-term rehabilitation. These findings emphasize the potential of this narrative-based, culturally contextualized digital tool to complement traditional community rehabilitation and to inform the development of more sustainable models of psychiatric care. Future research should extend follow-up periods to examine delayed clinical effects, incorporate adaptive and AI-driven narrative content to enhance cultural applicability, and test scalability across different health systems and populations.

Limitations

This study has several limitations. First, the financial incentives provided to participants may have facilitated their continued engagement. However, the primary purpose of these incentives was to compensate participants for their time and travel expenses. The amount of compensation was limited and unlikely to produce long-term changes in medication-taking behavior, suggesting that observed improvements in adherence are more likely attributable to the intervention mechanisms themselves. Nevertheless, we acknowledge that financial incentives may have modestly increased engagement, representing a potential confounding factor. Future studies could consider designing conditions without incentives or employing stratified analyses to control for this potential confounder. Second, all outcome measures relied on self-report, introducing the risk of subjective bias. The involvement of experienced psychiatrists may help mitigate such bias. Third, participants' familiarity and comfort with using smartphones to play the game varied, which could have influenced the effectiveness of the intervention. Additionally, as individuals in the intervention group were aware that they were receiving a digital intervention aimed at improving medication adherence, the absence of participant blinding may have introduced expectancy bias. Finally, while cultural customization likely enhanced the acceptability of the intervention within Chinese community settings, it may limit generalizability to more diverse or international populations. Future iterations of narrative-based psychoeducation should consider incorporating cross-cultural adaptation processes or leveraging AI-driven adaptive narratives to dynamically generate culturally relevant storylines.

Despite these limitations, the findings offer meaningful insights into the potential of narrative-driven digital psychoeducation to enhance medication adherence among individuals with schizophrenia in the long-term chronic phase. The intervention's integration of immersive storytelling, cognitive training, and self-management features presents a promising direction for future psychosocial support tools.

Conclusion

Our mixed methods findings indicate that the narrative-based digital psychoeducational intervention developed in this study demonstrated good usability and acceptability among individuals with chronic schizophrenia and may support short-term improvements in medication adherence and attitudes toward medication. Mobile health interventions of this kind could serve as a valuable complement to traditional community rehabilitation by enhancing behavioral and motivational mechanisms that contribute to long-term recovery. Nevertheless, careful attention should be paid to feedback design and cultural adaptation of narrative content, and tasks should be tailored to users' preferences and cognitive characteristics to ensure sustained usability and long-term effectiveness of the intervention.

Acknowledgments

©MMAS 2006 was used with permission of www.adherence.cc. The Morisky Medication Adherence Scale was used under a paid license granted by Dr. Donald E. Morisky and his colleagues (License Number 4525-8490-4067-7009-3166).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Video of Healing Town.
[MP4 File, 50167 KB - jmir_v28i1e59175_app1.mp4]

Multimedia Appendix 2
Content introduction of Healing Town.
[DOCX File, 3520 KB - jmir_v28i1e59175_app2.docx]

Multimedia Appendix 3
Data at baseline.
[DOCX File, 12 KB - jmir_v28i1e59175_app3.docx]

Multimedia Appendix 4
Themes of the interview.
[DOCX File, 24 KB - jmir_v28i1e59175_app4.docx]

XSL•FO
RenderX

Multimedia Appendix 5
High-order themes emerged from narrations.
[DOCX File, 16 KB - jmir_v28i1e59175_app5.docx]

Checklist 1
CONSORT-EHEALTH checklist (V 1.6.1).
[PDF File, 1222 KB - jmir_v28i1e59175_app6.pdf]

References

1. McCutcheon RA, Reis Marques T, Howes OD. Schizophrenia-an overview. JAMA Psychiatry 2020 Feb 1;77(2):201-210. [doi: 10.1001/jamapsychiatry.2019.3360] [Medline: 31664453]
2. van Os J, Kenis G, Rutten BPF. The environment and schizophrenia. Nature 2010 Nov;468(7321):203-212. [doi: 10.1038/nature09563]
3. Lançon C, Auquier P, Reine G, Bernard D, Addington D. Relationships between depression and psychotic symptoms of schizophrenia during an acute episode and stable period. Schizophr Res 2001 Mar 1;47(2-3):135-140. [doi: 10.1016/s0920-9964(00)00002-5] [Medline: 11278130]
4. Masand PS, Roca M, Turner MS, Kane JM. Partial adherence to antipsychotic medication impacts the course of illness in patients with schizophrenia. Prim Care Companion J Clin Psychiatry 2009;11(4):147-154. [doi: 10.4088/PCC.08r00612] [Medline: 19750066]
5. Cañas F, Alptekin K, Azorin JM, et al. Improving treatment adherence in your patients with schizophrenia: the STAY initiative. Clin Drug Investig 2013 Feb;33(2):97-107. [doi: 10.1007/s40261-012-0047-8] [Medline: 23288695]
6. Higashi K, Medic G, Littlewood KJ, Diez T, Granström O, De Hert M. Medication adherence in schizophrenia: factors influencing adherence and consequences of nonadherence, a systematic literature review. Ther Adv Psychopharmacol 2013 Aug;3(4):200-218. [doi: 10.1177/2045125312474019] [Medline: 24167693]
7. Ben-Zeev D, Scherer EA, Gottlieb JD, et al. mHealth for schizophrenia: patient engagement with a mobile phone intervention following hospital discharge. JMIR Ment Health 2016 Jul 27;3(3):e34. [doi: 10.2196/mental.6348] [Medline: 27465803]
8. Lincoln TM, Wilhelm K, Nestoriuc Y. Effectiveness of psychoeducation for relapse, symptoms, knowledge, adherence and functioning in psychotic disorders: a meta-analysis. Schizophr Res 2007 Nov;96(1-3):232-245. [doi: 10.1016/j.schres.2007.07.022] [Medline: 17826034]
9. Herrera SN, Sarac C, Phili A, et al. Psychoeducation for individuals at clinical high risk for psychosis: a scoping review. Schizophr Res 2023 Feb;252:148-158. [doi: 10.1016/j.schres.2023.01.008] [Medline: 36652831]
10. Çetin N, Aylaz R. The effect of mindfulness-based psychoeducation on insight and medication adherence of schizophrenia patients. Arch Psychiatr Nurs 2018 Oct;32(5):737-744. [doi: 10.1016/j.apnu.2018.04.011] [Medline: 30201202]
11. Al-Shashani A, Abu Sabra MA, Al-Gamal E. The Impact of using digital health interventions and psychoeducation on medication adherence among patients with schizophrenia: a scoping review. Issues Ment Health Nurs 2025 Jul;46(7):735-745. [doi: 10.1080/01612840.2025.2492694] [Medline: 40300193]
12. El-Mallakh P, Findlay J. Strategies to improve medication adherence in patients with schizophrenia: the role of support services. Neuropsychiatr Dis Treat 2015;11:1077-1090. [doi: 10.2147/NDT.S56107] [Medline: 25931823]
13. Cahaya N, Kristina SA, Widayanti AW, Green J. Interventions to improve medication adherence in people with schizophrenia: a systematic review. Patient Prefer Adherence 2022;16:2431-2449. [doi: 10.2147/PPA.S378951] [Medline: 36072918]
14. Li PWC, Yu DSF, Yan BP, Wong CW, Yue SCS, Chan CMC. Effects of a narrative-based psychoeducational intervention to prepare patients for responding to acute myocardial infarction: a randomized clinical trial. JAMA Netw Open 2022 Oct 3;5(10):e2239208. [doi: 10.1001/jamanetworkopen.2022.39208] [Medline: 36306128]
15. Bellack AS, Gold JM, Buchanan RW. Cognitive rehabilitation for schizophrenia: problems, prospects, and strategies. Schizophr Bull 1999;25(2):257-274. [doi: 10.1093/oxfordjournals.schbul.a033377] [Medline: 10416730]
16. Hinyard LJ, Kreuter MW. Using narrative communication as a tool for health behavior change: a conceptual, theoretical, and empirical overview. Health Educ Behav 2007 Oct;34(5):777-792. [doi: 10.1177/1090198106291963] [Medline: 17200094]
17. Landa Y, Mueser KT, Wyka KE, et al. Development of a group and family-based cognitive behavioural therapy program for youth at risk for psychosis. Early Interv Psychiatry 2016 Dec;10(6):511-521. [doi: 10.1111/eip.12204] [Medline: 25585830]
18. Keats PA. Multiple text analysis in narrative research: visual, written, and spoken stories of experience. Qual Res 2009 Apr;9(2):181-195. [doi: 10.1177/1468794108099320]
19. Johnson RB, Onwuegbuzie AJ, Turner LA. Toward a definition of mixed methods research. J Mix Methods Res 2007 Apr;1(2):112-133. [doi: 10.1177/1558689806298224]
20. Creswell JW, Clark VLP. Designing and Conducting Mixed Methods Research: Sage Publications; 2017:434-436.
21. Ethical principles of psychologists and code of conduct. : American Psychological Association; 2016 URL: https://www.apa.org/ethics/code/ethics-code-2017.pdf [accessed 2025-11-25] [doi: 10.1037/14805-030]

XSL•FO
RenderX

22. Liu Z, Zhu D, Zhang Z, et al. Effects of a narrative-based psychoeducational intervention on medication adherence in individuals with schizophrenia: a multicentre, parallel-group randomised controlled trial. EClinicalMedicine 2025 Oct;88:103483. [doi: 10.1016/j.eclinm.2025.103483] [Medline: 40979217]
23. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr Bull 1987;13(2):261-276. [doi: 10.1093/schbul/13.2.261] [Medline: 3616518]
24. Adams DM, Mayer RE, MacNamara A, Koenig A, Wainess R. Narrative games for learning: testing the discovery and narrative hypotheses. J Educ Psychol 2012;104(1):235-249. [doi: 10.1037/a0025595]
25. Cutting JE. Narrative theory and the dynamics of popular movies. Psychon Bull Rev 2016 Dec;23(6):1713-1743. [doi: 10.3758/s13423-016-1051-4] [Medline: 27142769]
26. Zhou C, Occa A, Kim S, Morgan S. A meta-analysis of narrative game-based interventions for promoting healthy behaviors. J Health Commun 2020;25(1):54-65. [doi: 10.1080/10810730.2019.1701586] [Medline: 31829829]
27. Levesque DA, Van Marter DF, Schneider RJ, et al. Randomized trial of a computer-tailored intervention for patients with depression. Am J Health Promot 2011;26(2):77-89. [doi: 10.4278/ajhp.090123-QUAN-27] [Medline: 22040388]
28. Oliveira-Filho AD, Barreto-Filho JA, Neves SJF, Lyra Junior DD. Association between the 8-item Morisky Medication Adherence Scale (MMAS-8) and blood pressure control. Arq Bras Cardiol 2012 Jul;99(1):649-658. [doi: 10.1590/s0066-782x2012005000053] [Medline: 22688844]
29. Moon SJ, Lee WY, Hwang JS, Hong YP, Morisky DE. Accuracy of a screening tool for medication adherence: a systematic review and meta-analysis of the Morisky Medication Adherence Scale-8. PLoS ONE 2017;12(11):e0187139. [doi: 10.1371/journal.pone.0187139] [Medline: 29095870]
30. Wang J, Bian RW, Mo YZ. Validation of the Chinese version of the eight-item Morisky medication adherence scale in patients with type 2 diabetes mellitus. J Clin Gerontol Geriatr 2013 Dec;4(4):119-122. [doi: 10.1016/j.jcgg.2013.06.002]
31. Krousel-Wood M, Islam T, Webber LS, Re RN, Morisky DE, Muntner P. New medication adherence scale versus pharmacy fill rates in seniors with hypertension. Am J Manag Care 2009 Jan;15(1):59-66. [Medline: 19146365]
32. Hogan TP, Awad AG, Eastwood R. A self-report scale predictive of drug compliance in schizophrenics: reliability and discriminative validity. Psychol Med 1983 Feb;13(1):177-183. [doi: 10.1017/s0033291700050182] [Medline: 6133297]
33. Shen YC. Epidemiological study of mental disorders in 12 regions of China: methodology and data analysis. Zhonghua Shen Jing Jing Shen Ke Za Zhi 1986 Apr;19(2):65-69. [Medline: 3743252]
34. Overall JE, Gorham DR. The Brief Psychiatric Rating Scale (BPRS): recent developments in ascertainment and scaling. Psychopharmacol Bull 1988;24(1):97-99. [Medline: 3387516]
35. Graneheim UH, Lindgren BM, Lundman B. Methodological challenges in qualitative content analysis: a discussion paper. Nurse Educ Today 2017 Sep;56:29-34. [doi: 10.1016/j.nedt.2017.06.002] [Medline: 28651100]
36. Carter N, Bryant-Lukosius D, DiCenso A, Blythe J, Neville AJ. The use of triangulation in qualitative research. Oncol Nurs Forum 2014 Sep;41(5):545-547. [doi: 10.1188/14.ONF.545-547] [Medline: 25158659]
37. Grey M, Whittemore R, Jeon S, et al. Internet psycho-education programs improve outcomes in youth with type 1 diabetes. Diabetes Care 2013 Sep;36(9):2475-2482. [doi: 10.2337/dc12-2199] [Medline: 23579179]
38. Alvarez-Jimenez M, Alcazar-Corcoles MA, González-Blanch C, Bendall S, McGorry PD, Gleeson JF. Online, social media and mobile technologies for psychosis treatment: a systematic review on novel user-led interventions. Schizophr Res 2014 Jun;156(1):96-106. [doi: 10.1016/j.schres.2014.03.021] [Medline: 24746468]
39. Killikelly C, He Z, Reeder C, Wykes T. Improving adherence to web-based and mobile technologies for people with psychosis: systematic review of new potential predictors of adherence. JMIR Mhealth Uhealth 2017 Jul 20;5(7):e94. [doi: 10.2196/mhealth.7088] [Medline: 28729235]
40. Lim MH, Penn DL. Using digital technology in the treatment of schizophrenia. Schizophr Bull 2018 Aug 20;44(5):937-938. [doi: 10.1093/schbul/sby081] [Medline: 29878251]
41. Gumley AI, Bradstreet S, Ainsworth J, et al. The EMPOWER blended digital intervention for relapse prevention in schizophrenia: a feasibility cluster randomised controlled trial in Scotland and Australia. Lancet Psychiatry 2022 Jun;9(6):477-486. [doi: 10.1016/S2215-0366(22)00103-1] [Medline: 35569503]
42. Garety P, Ward T, Emsley R, et al. Effects of SlowMo, a blended digital therapy targeting reasoning, on paranoia among people with psychosis: a randomized clinical trial. JAMA Psychiatry 2021 Jul 1;78(7):714-725. [doi: 10.1001/jamapsychiatry.2021.0326] [Medline: 33825827]
43. Maurin KD, Girod C, Consolini JL, et al. Use of a serious game to strengthen medication adherence in euthymic patients with bipolar disorder following a psychoeducational programme: a randomized controlled trial. J Affect Disord 2020 Feb 1;262:182-188. [doi: 10.1016/j.jad.2019.10.008] [Medline: 31668996]
44. Pan MR, Huang F, Zhao MJ, Wang YF, Wang YF, Qian QJ. A comparison of efficacy between cognitive behavioral therapy (CBT) and CBT combined with medication in adults with attention-deficit/hyperactivity disorder (ADHD). Psychiatry Res 2019 Sep;279:23-33. [doi: 10.1016/j.psychres.2019.06.040] [Medline: 31280035]
45. Robinson D, Woerner MG, Alvir JM, et al. Predictors of relapse following response from a first episode of schizophrenia or schizoaffective disorder. Arch Gen Psychiatry 1999 Mar;56(3):241-247. [doi: 10.1001/archpsyc.56.3.241] [Medline: 10078501]

XSL•FO
RenderX

46. Ascher-Svanum H, Faries DE, Zhu B, Ernst FR, Swartz MS, Swanson JW. Medication adherence and long-term functional outcomes in the treatment of schizophrenia in usual care. J Clin Psychiatry 2006 Mar;67(3):453-460. [doi: 10.4088/jcp.v67n0317] [Medline: 16649833]
47. Onitsuka T, Hirano Y, Nakazawa T, et al. Toward recovery in schizophrenia: current concepts, findings, and future research directions. Psychiatry Clin Neurosci 2022 Jul;76(7):282-291. [doi: 10.1111/pcn.13342] [Medline: 35235256]
48. Ma Z, Zytko D. Designing immersive stories for health: choosing character perspective based on the viewer's modality. Int J Hum Comput Interact 2021 Sep 14;37(15):1423-1435. [doi: 10.1080/10447318.2021.1886486]
49. Stephens C. Narrative analysis in health psychology research: personal, dialogical and social stories of health. Health Psychol Rev 2011 Mar;5(1):62-78. [doi: 10.1080/17437199.2010.543385]
50. Bai GN, Wang YF, Yang L, Niu WY. Effectiveness of a focused, brief psychoeducation program for parents of ADHD children: improvement of medication adherence and symptoms. Neuropsychiatr Dis Treat 2015;11:2721-2735. [doi: 10.2147/NDT.S88625] [Medline: 26604761]
51. Howard GS. Culture tales: a narrative approach to thinking, cross-cultural psychology, and psychotherapy. Am Psychol 1991 Mar;46(3):187-197. [doi: 10.1037//0003-066x.46.3.187] [Medline: 2035929]
52. Garety PA, Fowler D, Kuipers E. Cognitive-behavioral therapy for medication-resistant symptoms. Schizophr Bull 2000;26(1):73-86. [doi: 10.1093/oxfordjournals.schbul.a033447] [Medline: 10755670]
53. Westra HA. Comparing the predictive capacity of observed in-session resistance to self-reported motivation in cognitive behavioral therapy. Behav Res Ther 2011 Feb;49(2):106-113. [doi: 10.1016/j.brat.2010.11.007] [Medline: 21159325]
54. Olivet J, Haselden M, Piscitelli S, et al. Results from a pilot study of a computer-based role-playing game for young people with psychosis. Early Interv Psychiatry 2019 Aug;13(4):767-772. [doi: 10.1111/eip.12556] [Medline: 29542863]

Abbreviations

BPRS: Brief Psychiatric Rating Scale
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
DAI: Drug Attitude Inventory
MMAS: Morisky Medication Adherence Scale
PANSS: Positive and Negative Syndrome Scale
RCT: randomized controlled trial
SDSS: Social Disability Screening Schedule

Edited by A Stone; submitted 02.Jul.2025; peer-reviewed by AAL Sawafi, M Tusconi; revised version received 10.Nov.2025; accepted 10.Nov.2025; published 20.Jan.2026.

Please cite as:
Zhu D, Chang F, Yang H, Wei Y, Liu Z
Assessing Usage and Usability of a Narrative-Based Psychoeducational Digital Intervention to Improve Medication Adherence Among Individuals With Schizophrenia in a Stable Phase: Mixed Methods Study
J Med Internet Res 2026;28:e59175
URL: https://www.jmir.org/2026/1/e59175
doi:10.2196/59175

© Dian Zhu, Fangyuan Chang, Hongyi Yang, Yiwen Wei, Zhao Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of and Mechanisms of Change in a Self-Help Web- and App-Based Resilience Intervention on Perceived Stress in the General Working Population: Randomized Controlled Trial

Sandy Hannibal[1], MSc; Dörte Behrendt[1], MSc; Michèle Wessa[2,3,4], PhD; Sarah K Schäfer[5,6], PhD; Nina Dalkner[7], PhD; Dirk Lehr[1], PhD

[1]Department of Health Psychology and Applied Biological Psychology, Institute of Sustainability Psychology, Leuphana University of Lüneburg, Universitätsallee 1, Lüneburg, Germany

[2]Department of Neuropsychology and Psychological Resilience Research, Central Institute of Mental Health, Mannheim, Germany

[3]Research Division Cancer Survivorship and Psychological Resilience, German Cancer Research Center (DKFZ)-Hector Cancer Institute at the University Medical Center Mannheim, Mannheim, Germany

[4]Division Cancer Survivorship and Psychological Resilience, German Cancer Research Center (DKFZ) Heidelberg, Heidelberg, Germany

[5]Department of Clinical Psychology and Psychotherapy for Children and Adolescents, Technische Universität Braunschweig, Braunschweig, Germany

[6]Leibniz Institute for Resilience Research, Mainz, Germany

[7]Division of Psychiatry and Psychotherapeutic Medicine, Medical University of Graz, Graz, Austria

Corresponding Author:
Sandy Hannibal, MSc
Department of Health Psychology and Applied Biological Psychology, Institute of Sustainability Psychology, Leuphana University of Lüneburg, Universitätsallee 1, Lüneburg, Germany

Abstract

Background: Promoting individual resilience—that is, maintaining or regaining mental health despite stressful circumstances—is regarded as an important endeavor to prevent mental illness. However, digital resilience interventions designed to enhance mental health have yielded mixed results. Such heterogeneous effects reflect a variety of unsolved conceptual challenges in interventional resilience research. These range from grounding interventions in resilience frameworks, using theory or targeting etiologically important resilience factors as intervention content, to a lack of knowledge about the mechanisms underlying effects, and using techniques specifically developed to foster psychosocial resources. The web- and app-based resilience intervention RESIST was designed to address these challenges, mainly by using both the Positive Appraisal Style Theory of Resilience as its theoretical foundation and interventional techniques from Strengths-Based Cognitive Behavioral Therapy.

Objective: This study's primary aim was to evaluate the effectiveness of RESIST on perceived stress in a general working population as a means of universal prevention, relative to a waitlist control group. A secondary study aim was to explore the resilience factors of self-efficacy, optimism, self-compassion, and perceived social support, the intervention targets as potential mediators of its effect on stress and self-perceived resilience.

Methods: In total, 352 employees were randomly assigned to either a self-help version of RESIST or a waitlist control group. Data were collected via the web at baseline, postintervention, and at 3- and 6-month (intervention group [IG] only) follow-ups. The primary outcome was perceived stress, measured with the Perceived Stress Scale-10. Secondary outcomes included self-perceived resilience, the resilience factors targeted, and other mental and work-related health outcomes.

Results: The IG reported significantly less stress than controls postintervention (Δ=−3.14; d=−0.54, 95% CI −0.75 to −0.34, and P<.001) and at 3-month follow-up (Δ=−2.79; d=−0.47, 95% CI −0.71 to −0.22, and P=.002). These improvements in the IG were maintained at 6-month follow-up. Favorable between-group differences also were detected for self-perceived resilience and the resilience factors. IG participants completed on average 2.2 (SD 2.3) web-based sessions and used the app's core feature a median of 14 times (IQR 4.00-33.75, range 1-220). The positive effects of the intervention on stress and resilience were primarily mediated by changes in optimism and self-compassion. No evidence was found that self-efficacy and social support also acted as mediators.

Conclusions: In a sample of employees experiencing heightened work-burden levels, RESIST was effective in reducing perceived stress and increasing self-perceived resilience as well as the targeted resilience factors. Mediation analyses suggested that developing a positive future outlook and a self-compassionate attitude toward oneself may be key drivers to enhance resilience. Changing the quality of social relationships and strengthening the belief in one's abilities may require more time, the involvement of others, or personal support from an eCoach to ensure sufficient learning opportunities.

Trial Registration: German Clinical Trials Register DRKS00017605; https://drks.de/search/de/trial/DRKS00017605

XSL•FO
RenderX

(J Med Internet Res 2026;28:e78335) doi:10.2196/78335

KEYWORDS

stress; resilience factor; resilience mechanism; resilience training; digital mental health intervention; internet-based intervention; mobile intervention; occupational eMental health; prevention; RCT; randomized controlled trial; mobile phone

Introduction

Background

Stressors are an integral part of life. By challenging an individual to adapt to new circumstances, they provide opportunities for growth and development. Most people can maintain or quickly regain good mental health despite experiencing stress—a phenomenon referred to as resilience [1-3]. However, when adaptation fails, stressors can contribute to the development of stress-related mental disorders [4]. Given that disorders such as depression and anxiety rank among the top 25 leading causes of global health-related burden [5,6], finding effective strategies to help individuals adapt to stress and prevent mental illness is of high importance. The promotion of individual resilience can, therefore, be an important endeavor to mitigate the potentially harmful effects of stress for individuals and, in turn, societies.

Resilience, however, is a complex phenomenon and the subject of extensive research. Resilience research has not yet reached consensus on how to define or conceptualize it. Although definitions vary, recent conceptual approaches, such as the Positive Appraisal Style Theory of Resilience (PASTOR) [2,3], converge on the idea that resilience reflects a dynamic, active, and multidimensional process of adaptation to stressors. Since this process is potentially modifiable, this opens avenues for resilience-promoting interventions.

Evidence for Resilience-Promoting Interventions

Published evidence on the effectiveness of both nondigital and digital resilience interventions, however, remains mixed, with highly heterogeneous effects observed between primary studies and inconsistent conclusions drawn in systematic reviews and meta-analyses [7-11]. Concerning digital interventions, one meta-analysis identified small-to-moderate effects on self-perceived resilience, but not on stress reduction [7], while another found no effect on self-perceived resilience [8]. Conversely, a recent meta-analysis by Schäfer et al [11] detected small, favorable effects both on perceived stress and self-perceived resilience. Favorable effects were also found for various resilience factors, including self-compassion, optimism, and social support [11]. Overall heterogeneity in the literature may stem from the general lack of consensus on what constitutes a resilience intervention [9,12], resulting in inconsistent inclusion and exclusion decisions between reviews.

Classification of Resilience Interventions

Overview

One approach has been to classify any intervention as a resilience intervention, regardless of more specific characteristics, as long as its primary objective is to modify resilience-related outcomes. These outcomes can include both self-perceived resilience, operationalized as the perceived ability to bounce back and recover from stress, or via related mental health outcomes (eg, stress and well-being).

An alternative approach is to define resilience interventions based on more specific intervention characteristics, including (1) the theory or rationale underlying the intervention; (2) the intervention's content; (3) the interventional techniques applied; and (4) the timing of the intervention relative to stressor exposure. These dimensions contribute to significant variability in the design of resilience interventions and represent a variety of unsolved conceptual challenges in interventional resilience research.

Underlying Theory

Concerning underlying theories and rationales behind interventions, many resilience interventions take a pragmatic, atheoretical approach, often without reference to any specific theoretical foundation. Conversely, some interventions are informed by established theories drawn from related domains, which are then applied to the context of resilience [12]; for example, the Transactional Model of Stress [13]. Notably rare, however, are interventions explicitly grounded in a specific theory of resilience or a genuine resilience framework. As introduced earlier, PASTOR [2,3] provides such a framework. It emphasizes that resilience as an outcome from a dynamic, active, and multicausal process of stressor adaptation is shaped by both internal and external resilience factors. The theory further highlights that these resilience factors might exert their protective effects through common cognitive appraisal processes, and that the core mechanism underlying resilience may be a positive appraisal style of stressors, that is, a general tendency to appraise stressors in a nonnegative, nonthreatening, but positive way.

Intervention Content

Further, an intervention may be considered a resilience intervention based on its content, regardless of the outcomes studied. The most common approach involves targeting intervention content toward building or strengthening certain resilience factors. These factors predominantly consist of a broad range of psychosocial resources, including self-efficacy, locus of control, problem-solving, optimism, cognitive flexibility, mindfulness, self-compassion, and perceived social support [7-9]. Such resources have been shown to be statistically associated with resiliency outcomes and are thought to facilitate adaptive responses to stress [2], thereby helping individuals to bounce back from adversity. Consistent with the idea that access to a broad repertoire of resources facilitates flexible responses to diverse challenging situations [14], resilience interventions should, nonetheless, aim to target multiple resilience factors rather than a single one. For instance, enhancing self-efficacy may have limited value in situations where an individual is unable to influence outcomes through their own actions. However, certain resilience factors are also commonly targeted

XSL•FO
RenderX

in other well-established intervention approaches, including problem-solving approaches [15] and mindfulness-based practices [16]. In contrast, if resilience interventions are categorized as a distinct class of intervention, it is essential to delineate the specific factors that should be targeted to meaningfully differentiate them from other existing approaches. In this regard, the resilience factors addressed within such interventions may be those that reflect more specific and overarching resilience processes and include, for example, factors that may serve as correlates or indicators of higher-order resilience mechanisms, such as a positive appraisal style [2].

The web- and app-based resilience training RESIST is an example of an intervention that is grounded in PASTOR [2], as the intervention's underlying theoretical framework. It aims to foster a set of specific resilience factors, namely self-efficacy, optimism, self-compassion, and social support, associated with a positive appraisal style as a higher-order resilience mechanism [17]. A pilot randomized controlled trial (RCT) has already demonstrated its effectiveness [17], providing preliminary evidence that an intervention based on these principles might be a promising approach to fostering self-perceived resilience.

Mechanisms of Change

Despite certain resilience factors being targeted within given interventions, it nonetheless remains largely unknown whether favorable intervention outcomes can be directly attributed to increases in these resilience factors. In such cases, the targeted resilience factors would function as mediators of the intervention, that is, the intervening variables that statistically account for the relationship between an intervention and its outcome [18]. They may therefore point to mechanisms of change through which the intervention achieves its effects [19]. Notwithstanding the growing number of resilience interventions, empirically based understanding about their mechanisms of change remains scarce [11]; with only a few published studies, often inadequately powered, having explored this [20,21]. Examining the mechanisms underlying resilience interventions could offer valuable insights into core processes promoting resilience. Such insights could help to identify the most appropriate selection of resilience factors for interventions to target, ultimately contributing to their optimization and increased effectiveness.

Interventional Techniques

Concerning the intervention techniques used, which determine how intervention content is delivered, there appears to be a disconnect between the psychosocial resources targeted by resilience interventions and the use of techniques explicitly designed to foster such resources (eg, keeping a diary with a focus on successful experiences to foster resilience by enhancing self-efficacy). Indeed, many resilience interventions adopt a problem-reducing approach [22], rooted in psychotherapeutic approaches via, for example, cognitive restructuring techniques. One contrasting example of an approach explicitly developed to foster individuals' protective resources to build resilience is Strengths-Based Cognitive Behavioral Therapy (Strengths-Based CBT) by Padesky and Mooney [23]. Among the first of its kind, the RESIST intervention applied this approach as the intervention technique used it to target the resilience factors as intervention content [17]. Since RESIST targets employees experiencing workplace stressors, it can be classified, in terms of the timing of the intervention relative to stressor exposure (4), as an intervention for use during stressor exposure [12].

Aims of This Study

Building on the findings of the pilot study assessing RESIST in an indicated preventive setting [17], we conducted a sufficiently powered study in the general working population to compare the effectiveness of RESIST against a waitlist control group (WL) in reducing subjective stress levels, which served as the primary outcome. We conducted further exploratory analyses to examine its effects on various secondary outcomes, including self-perceived resilience and the resilience factors targeted within the intervention. In light of the urgent need to address the gap in knowledge regarding resilience interventions' mechanisms of change, we also examined, as a secondary aim, the mediating effects of the resilience factors emphasized in RESIST on stress and self-perceived resilience.

Methods

Study Design

This study was conducted as a two-arm RCT from November 2019 to August 2021 in Germany. To investigate the intervention effects on the primary outcome of perceived stress, participants were randomly assigned either to the intervention group (IG) with immediate access to the resilience intervention RESIST in the form of a self-help format or to a WL whose members were offered the resilience intervention after 3 months of waiting time.

The study is reported in compliance with the CONSORT-eHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidelines for improving and standardizing the reporting of web-based and mobile health interventions [24] and following the recommendations of Chmitorz et al [12] for reporting trials on resilience interventions. The intervention itself is reported in accordance with the TIDieR (Template for Intervention Description and Replication) checklist [25]; an overview of the checklist, along with information on where to find the corresponding details in the main text, is provided in Multimedia Appendix 1.

Ethical Considerations

This study received ethical approval from the Ethics Committee of Leuphana University of Lüneburg. Before initiating subject enrollment, this study was registered at the German Clinical Trials Registry (DRKS00017605), a primary registry of the World Health Organization. All participants provided written informed consent before taking part in this study. Participants were informed that they could withdraw their consent at any time and request the deletion of their data without experiencing any negative consequences.

All data were deidentified before analysis; personal identifiers were replaced with unique codes, and the coding key was stored securely and separately. Access to identifiable data was restricted to authorized members of the research team.

XSL•FO
RenderX

Participation in this study was entirely voluntary and not financially compensated.

Participant Recruitment

This study was conducted in a universal preventive setting, addressing the general working population. Participants were recruited via (1) articles on resilience containing a link to the study's landing page in two well-known popular German science magazines ("Spektrum der Wissenschaft" and "Spektrum der Psychologie"); (2) newsletters from several statutory German health insurance companies (eg, Daimler BKK) that included a section referring to this study; and (3) social media posts on various platforms (Xing [New Work SE], Instagram [Instagram from Meta], and Facebook [Meta]).

Individuals who entered their email address in a form on this study's landing page or contacted us directly via email received detailed information about this study and were provided access to an online screening questionnaire via email. Inclusion criteria for study participation were (1) being aged at least 18 years, (2) having employment (part-time or full-time), (3) having steady access to the internet, and (4) possessing a smartphone. Exclusion criteria were (1) having an earlier or current diagnosis of a severe psychological disorder (eg, psychosis); (2) undergoing current psychotherapy, including being on a waiting list for or having the intention to receive psychotherapy; (3) current participation in another resilience intervention or stress management program; (4) changes in current medication for a stress disorder, anxiety disorder, or depression within the past four weeks; and (5) current suicidal ideation (agreeing with either the statement "I would like to kill myself" or "I would kill myself if I had the chance" on the related item in the Beck Depression Inventory-II [26]).

Individuals eligible for study participation received a link to the baseline assessment. After completing the baseline assessment and giving informed consent for study participation, participants were randomly assigned to one of the two study groups using a computer-generated randomization list with a ratio of 2:2 and a block size of four. The list was generated by a member of the department of the first and last author, who was not otherwise involved in this study. Randomization was conducted anonymously, without any personal contact between study personnel and participants. Blinding of participants was infeasible because participants were aware of whether they were randomized to the IG or the WL. Participants were informed about their group allocation via email and, depending on the randomization outcome, either were granted immediate access to the web- and app-based intervention (IG) or were promised access after 3 months of waiting time (WL).

Intervention

The feasibility and preliminary effectiveness of the web- and app-based resilience intervention RESIST were positively evaluated in a pilot RCT [17]. As a genuine resilience framework, PASTOR was considered the underlying theoretical foundation of the intervention [2,17]. Overall, the intervention aims to promote a positive appraisal style of stressors and stressful situations to help users adjust to work-related stress. Thus, this study design followed an approach focused on delivering the intervention during ongoing stressor exposure [12]. According to PASTOR, both a positive appraisal style as a resilience mechanism and successful adaptation to stress are facilitated by several resilience factors. Thus, promoting a set of specific resilience factors was at the intervention's core. However, PASTOR is a conceptual framework for this study of resilience, aiming for comprehensiveness in explaining the mechanisms underlying resilient adaptation. Consequently, it does not provide specific guidelines on which factors should be strengthened or how this should be achieved. Thus, the selection of the resilience factors targeted in the training was guided by several considerations: (1) they should be etiologically relevant for fostering adaptive responses to adversity; (2) they should potentially be competencies fundamental to resilience; (3) they should encompass different facets of resilience, including behavioral (eg, self-efficacy), cognitive (eg, optimism), emotional (eg, self-compassion), and interpersonal (eg, social support) dimensions; (4) they should reflect varying temporal orientations (eg, self-efficacy grounded in past achievements and optimism oriented toward the future); (5) they should be distinct from factors commonly targeted in other well-established intervention approaches (eg, mindfulness or active coping, which are typically addressed in the context of mindfulness and stress-management training); and (6) the number of selected factors should remain manageable to enable the design of medium-length interventions, which have been shown to be most effective [27]. Based on these considerations, self-efficacy, optimism, self-compassion, and social support [28-31] were selected as resilience factors targeted in the training.

RESIST was designed as a hybrid intervention to be delivered individually to participants, combining a web-based intervention and a mobile app component. The intervention's format allows participants to complete it anywhere they prefer. The web-based component accessible via a digital platform [32], consisting of longer intervention sessions, is designed for completion on a one-per-week basis. Once one session is completed, the next session is unlocked automatically. The app component that can be downloaded from the Google Play Store and Apple App Store (no longer available) consists of daily exercises. The two components are intentionally interrelated, with daily app exercises informing web-based sessions and web sessions guiding app-based practice. The intervention period for each study participant was 8 weeks. A detailed overview of the interventions' content and exercises can be found in Multimedia Appendix 2. Strengths-Based CBT [23] is applied as the interventional technique addressing *how* the resilience factors are targeted and is comprised of four steps: (1) searching for strengths; (2) building a personal model of resilience (PMRe), that is, a positive self-concept including positive imagery or metaphors; (3) applying the PMRe; and (4) practicing the PMRe.

Using the app component was step 1 (search for strengths) of the Strengths-Based CBT approach (Multimedia Appendix 2). As the app's central feature, users can document experiences of resilience, so-called moments of resilience, throughout the day (see Figure S2A in Multimedia Appendix 3). These moments capture instances when users overcome minor or major obstacles or experience positive events despite ongoing stressors

XSL•FO
RenderX

or challenging life circumstances. During a daily review, these moments are categorized into one of the four resilience factors of self-efficacy, optimism, self-compassion, and social support, based on the user's descriptions (see Figure S2D in Multimedia Appendix 3). Details on other app components are provided in Multimedia Appendix 3.

The web-based component has six sessions. The first session introduces the resilience intervention. Each of the second through fifth sessions addresses step 2 (construct PMRe), step 3, and step 4 (apply+practice PMRe) of the Strengths-Based CBT approach (Multimedia Appendix 2), which are incorporated as follows. The moments of resilience collected with the app are displayed in the web component. Based on step 2 of the Strengths-Based CBT approach, an exercise named resilience self-image was designed to further work with the moments of resilience. Within this exercise, users can select one of the displayed moments of resilience and develop a positive imagery or metaphor describing how they felt in the respective moment (eg, "I felt like a tribal leader, aware of my position and bravely taking responsibility"). This was intended to help participants develop a positive and resilient self-concept. A third exercise, named the resilience project, was developed based on steps 3 and 4 (apply+practice PMRe) of the Strengths-Based CBT approach. Within this exercise, users can plan how to approach and overcome an upcoming stressful situation by making use of their resilience self-image and resilience factors. The sixth session recaps the content of all the previous sessions.

Besides the exercises constructed on the Strengths-Based CBT, sessions include educational elements via text or video, as well as further written or audio exercises on the four resilience factors, each targeted in a single session. The sessions further feature two personas, which represent fictional users who are also completing the resilience training, thereby providing examples of how to complete the exercises. Example screenshots of the web component are available in Multimedia Appendix 3. Each session consists of 8‐10 web pages and requires 45‐60 minutes to complete. Users may complete the intervention at their own pace; however, it is recommended that they finish one session per week. No external reminders were sent to participants to complete the web-based component; however, in-app reminders could be set within the app component. The intervention was not tailored or modified during the course of this study. Access to the intervention can be provided to readers upon request.

Measures

Overview

Web-based data assessment via patient-reported outcome measures took place before randomization for screening purposes (T0), at baseline (T1), immediately postintervention, which was eight weeks after randomization (postintervention assessment [T2]), and three months after randomization (3 mo follow-up [T3]). We further included a 6-month follow-up (T4) assessment for the IG only, guided by existing evidence indicating that between-group intervention effects of similar programs remain largely stable over this period [10,33]. Thus, for ethical reasons, the control group did not continue without intervention beyond the initial 3 months.

Primary Outcome Measure

The primary outcome of this study was subjective stress referring to the last 7 days, measured using the Perceived Stress Scale-10 (PSS-10) [34,35], a well-established self-report measurement that assesses the degree to which people appraise situations in life as stressful—that is, unpredictable, uncontrollable, or overloaded. It consists of 10 items (eg, "In the last month, how often have you felt nervous and 'stressed'?") with each item rated on a scale ranging from 0 to 4. The total score ranges from 0 to 40, with higher scores indicating higher levels of perceived stress. In the present sample, the scale demonstrated high internal consistency, with an α=.89 at baseline.

Secondary Outcome Measures

Overview

Secondary explorative outcomes included resilience-, mental health-, work-, and intervention-related outcomes. Further details and the reliability of all outcome measures are listed in Multimedia Appendix 4.

Resilience and Resilience Factors

Self-perceived resilience—the perceived ability to recover from stress—was measured using the Brief Resilience Scale [36]. The resilience factor self-efficacy was assessed using the General Self-Efficacy Short Scale-3 [37]. As related constructs, internal and external control beliefs were estimated using the Short Scale for the Assessment of Locus of Control [37] for explorative purposes. The resilience factor, optimism, was assessed using the revised version of the Life Orientation Test [38]. The resilience factor, self-compassion, was assessed using the Self-Compassion Scale Short Form [39]. The resilience factor, social support, was measured with the perceived available social support subscale and the support-seeking subscale of the Berlin Social Support Scales [40].

Mental Health Outcomes

Symptoms of depression were measured using the Epidemiologic Studies Depression Scale [41]. Values ≥18 indicate clinically significant levels of depressive symptoms [42]. Psychological distress, including anxiety and psychosomatic symptoms, was measured using the short version of the Brief Symptom Inventory [43].

Consistent with the notion that resilience per se is defined as good mental health despite adversity [44], exposure to different types of stressors was assessed. Critical life events (eg, death of a family member) were measured using an adapted version of the Life History Calendar [45]. Daily hassles were assessed using the Mainz Inventory of Microstressors [46].

Work-Related Outcomes

Adverse circumstances at work were assessed using the short version of the Effort-Reward-Imbalance Questionnaire [47,48], which contains the subscales effort, reward, and overcommitment. For the purpose of describing this study's sample baseline characteristics, the ratio of effort (numerator) over reward (denominator) at T1 was calculated to capture the imbalance between costs and gains experienced at work [49].

XSL•FO
RenderX

According to Lehr et al [49], a ratio >0.715 indicates an adversarial workplace environment.

Concerning productivity in the workplace, days of absenteeism and presenteeism were measured using the respective items of the Trimbos and Institute of Medical Technology Assessment Cost Questionnaire for Psychiatry [50]. Ability to work effectively was measured using the single-item Work Ability Index [51], assessing "current work ability compared with life-time best."

Due to data management errors, data on the secondary outcomes, psychological distress, effort, and rewards are missing for T2 and T4, while data for life events and daily hassles are missing from T2 to T4. These missing data represent minor deviations from this study's protocol.

Intervention Usage and Client Satisfaction

Intervention usage was assessed in two ways: (1) by the number of web-based sessions completed per IG participant (extracted from the online platform); and (2) by the collected number of moments of resilience, the core feature of the app, per participant (as tracked by the native application). Satisfaction with the web- and app-based intervention was assessed postintervention (only for IG) using an adapted version of the Client Satisfaction Questionnaire adjusted to the online format of the intervention [52]. Participants' perception of the app's attractiveness and quality was assessed using the short version of the AttrakDiff [53].

To characterize this study's sample, data on demographic variables such as age, gender, marital status, educational level, employment status, income, and the previous use of health services were collected as part of the screening procedure.

Statistical Analysis

Overview

In accordance with CONSORT-eHEALTH guidelines [24], data were analyzed following an intention-to-treat procedure. Specific details of the analysis are provided below. All statistical analysis was performed using R Studio (R version 4.3.2) [54] with a two-tailed significance level set at $P \leq .05$.

Power Calculation

This study's sample size calculation was based on identifying a realistic and practically meaningful effect size. First, regarding a realistically achievable effect, meta-analytic evidence on unguided web-based interventions for stress reduction was considered [27]. In the meta-analysis by Heber et al [27], an effect size of $d=-0.33$ was reported for unguided stress interventions. Second, by discussion, the research team agreed on 2.5 points as the minimal practically important difference between the IG and WL immediately after the intervention. Assuming an SD of 6.42 points on the PSS-10, as observed in a representative German sample [34], this resulted in an anticipated intervention effect of $d=-0.4$ for the primary outcome in this study. To detect an effect of that size with 80% power and 95% significance, an a priori power-analysis calculation yielded 352 participants required for this study.

Missing Data

All participants provided baseline data. Missing data at T2-T4 were imputed through multivariate imputation by chained equations using the *mice* package [55]. Twenty imputation sets were generated. Variables were included as predictors in the imputation model if they showed a correlation of at least 0.4 with the variable to be imputed and had at least 60% available data, ensuring that only predictors carrying a substantial amount of meaningful information were used [56]. The imputed datasets were analyzed separately, and the parameter estimates and hypothesis tests were ultimately pooled using the rule by Rubin [57].

Intervention Effect

Between-group differences at T2 and T3 were analyzed using analysis of covariance (ANCOVA), as recommended by O'Connell et al [58], to evaluate the interventions' effectiveness in influencing the primary and secondary outcomes. Baseline scores of the respective outcomes were considered covariates. Following recommendations by Harrer et al [56], between-group Cohen *d* and corresponding CIs were derived directly from the covariate-adjusted models used to calculate the magnitude of treatment effect, incorporating the pooled SD (based on Cohen formula) into the models.

Sensitivity analyses for the primary outcome were performed to test the robustness of the intention-to-treat analyses in multiple ways. First, study completers (participants providing data at all measurement points) were analyzed. Second, intervention completers were analyzed, defined as participants having completed at least the first five sessions of the web-based component (incorporating all exercises on the targeted four resilience factors) and collected moments of resilience with the app at least twice per week (≥16) within the training period. Third, the intervention effect was measured using linear mixed models.

Response and Deterioration Rates

Concerning better communicating the results to help-seekers and policy makers, we followed the recommendation by Cuijpers [59] and calculated response and deterioration rates. For all response rates, we calculated the numbers needed to treat to achieve one additional beneficial outcome (number needed to treat for benefit; NNTB) and one additional harmful outcome (numbers needed to treat for harm; NNTH) in the IG, relative to the WL.

We analyzed response rates from various perspectives. First, reliable improvement and deterioration were calculated based on Heber et al [33], who reported a change of ±5.16 points in the PSS-10 following the recommendations of Jacobson and Truax [60]. Second, using anchor-based criteria was advised for determining what constitutes a practically meaningful change [61]. Bauer-Staeb et al [62] highlighted that the magnitude of a practically meaningful change depends on the baseline level of distress and provided reference criteria. Given that the sample was expected to be primarily mildly to moderately distressed, a change of 20% from baseline to postintervention was considered a minimally practical meaningful difference. Third, remission was determined using the preferred method described

XSL•FO
RenderX

by Jacobson and Truax [60] for operationalizing clinically significant change—defined as having a score closer to the mean of the functional than dysfunctional population. The mean for a representative German sample (mean 12.57) served as an estimate for the functional population [34], while the mean score of a highly stressed, help-seeking sample (mean 22.65) was used to define the dysfunctional population [63]. Consequently, we defined remission as scoring 17 points or lower on the PSS-10, excluding participants already considered to have no significant symptoms by this definition at baseline.

Longer-Term Effects

An extended T4 assessment evaluated longer-term interventional effects among participants in the IG. Those effects were analyzed using within-subject comparisons via repeated-measures ANOVA between T1 and T4. Within-group Cohen *d* and respective CIs for the IG were calculated by dividing the change score (T1-T4 difference) by the SD of the change.

Mediation Analyses

To investigate the targeted resilience factors self-efficacy, optimism, perceived support, and self-compassion as potential mediators of the intervention's effect on stress and resilience, multiple single (including one mediator) and parallel mediation analyses (including all mediators at once) were conducted using the *semTools* package [64]. Single mediation analyses were conducted to assess the unique effect of each mediator in isolation, while parallel models provided a more comprehensive view of how multiple mediators operate concurrently and allowed for comparison of their relative contributions. To establish temporal precedence, T2 scores of the potential mediators and T3 scores of stress and resilience were used [18]. Mediator and baseline scores for stress and self-perceived resilience were included as covariates in the model, as recommended by Hayes and Rockwood [65]. A statistically significant effect of the mediation was achieved if the estimated 95% CI of the indirect effect did not cross zero. We conducted all mediation analyses with the study completer sample as additional sensitivity analyses.

Results

Participants and Baseline Characteristics

Figure 1 illustrates the flow of participants. A total of 352 individuals were randomized and assigned to either the IG or WL. Two participants initially allocated to the WL were subsequently excluded from this study due to concurrent involvement in another parallel research project. As a result, the final participant count was 176 participants for the IG and 174 participants for the WL.

XSL•FO
RenderX

Figure 1. Study flow of participants. IG: intervention group; PSS: Perceived Stress Scale; WL: waitlist control group.

XSL•FO
RenderX

Table 1 summarizes the study samples' baseline characteristics. On average, participants were aged 42.8 years; 65.7% (230/350) were female; and 74.9% (262/350) worked full-time. At screening, participants reported an average perceived stress level of 21.58 (SD 6.03). The average depression level among participants at baseline was 11.09 (SD 6.42). A total of 15.7% (55/350) of the participants reported clinically relevant levels of depressive symptoms. Participants reported an average sum of 68.45 (SD 32.72) daily hassles in the preceding 7 days (eg, a conflict or disagreement with a close person), along with 9.24 (SD 5.08) experienced life events (eg, serious illness of self or family member). Almost nine in ten (309/350, 88.3%) reported working under adversarial workplace conditions (Effort Reward Imbalance Scale ratio >0.715) at baseline. Participants' work ability—compared to life-time best indicated as 10 points—was reduced (mean 6.83, SD 1.86).

Table . Baseline characteristics of this study's sample.

	Total	IG[a] (n=176)	WL[b] (n=174)
Sociodemographics			
Age (y), mean (SD)	42.8 (11.2)	41.9 (11.2)	43.8 (11.1)
Female, n (%)	230 (65.7)	123 (69.9)	107 (61.5)
Single, n (%)	116 (33.1)	67 (38.1)	49 (28.2)
Married, n (%)	169 (48.3)	77 (43.8)	92 (52.9)
Relationship, n (%)	40 (11.4)	18 (10.2)	22 (12.6)
Divorced, n (%)	25 (7.1)	14 (8)	11 (6.3)
Education, n (%)			
No university degree	81 (23.1)	41 (23.3)	40 (23)
University degree	269 (76.9)	135 (76.7)	134 (77)
Employment, n (%)			
Full-time	262 (74.9)	143 (81.3)	119 (68.4)
Part-time	88 (25.1)	33 (18.8)	55 (31.6)
Prior experience with a mental health intervention, n (%)			
No	310 (88.6)	152 (86.4)	158 (90.8)
Yes	40 (11.4)	24 (13.6)	16 (9.2)
Experience with psychotherapy, n (%)			
No, never	243 (69.4)	123 (69.9)	120 (69)
Yes, in the past	103 (29.4)	52 (29.6)	51 (29.3)
Perceived stress levels, mean (SD)			
T0 PSS-10[c]	21.6 (6.03)	21.8 (5.7)	21.4 (6.4)
Clinically significant levels of depressive symptoms[d], n (%)			
Yes	55 (15.7)	22 (12.5)	33 (19)
No	295 (84.3)	154 (87.5)	141 (81)
Stressor exposure before intervention			
Adverse workplace situation (ERI-S[e]>0.715), n (%)	309 (88.3)	153 (86.9)	156 (89.7)
Macrostressors[f], mean (SD)	9.24 (5.1)	9.1 (4.9)	9.4 (5.2)
Microstressors[g], mean (SD)	68.5 (32.7)	68.1 (32.2)	68.8 (33.3)

[a]IG: intervention group.

[b]WL: waitlist control group.

[c]PSS-10: Perceived Stress Scale 10.

[d]Scores ≥18 on the Epidemiologic Studies Depression Scale (short form).

[e]ERI-S: Effort-Reward-Imbalance Scale (short form).

[f]Life events (counts).

[g]Daily hassles (counts).

XSL•FO
RenderX

Primary Outcome Measure

Table 2 lists the means and SDs of the primary and all secondary outcomes at all assessment points. At T2, individuals in the IG reported significantly lower levels of perceived stress than individuals in the WL, as demonstrated by ANCOVA ($F_{1,518}$=21.31, P<.001, d=−0.54, 95% CI −0.75 to −0.34, between-group difference: −3.14 points, see Table 3). The between-group effect remained significant and comparable in size at 3-month follow-up ($F_{1,79}$=10.38, P=.002, d=−0.47, 95% CI −0.71 to −0.22, between-group difference: −2.79, see Table 3).

XSL•FO
RenderX

Table . Means and SDs of the intention-to-treat approach's sample (intervention group=176; waitlist control group=174).

Outcome	T1[a]		T2[b]		T3[c]		T4[d]
	IG[e]	WL[f]	IG	WL	IG	WL	IG
Primary outcome, mean (SD)							
Stress	21.35 (6.21)	20.52 (6.71)	16.22 (6.47)	19.36 (6.63)	16.63 (7.33)	19.42 (6.92)	16.11 (6.86)
Secondary outcomes, mean (SD)							
Self-perceived resilience	16.80 (4.70)	16.61 (4.44)	18.84 (3.40)	17.19 (3.31)	19.34 (4.88)	17.86 (4.36)	19.67 (4.48)
Resilience factors, mean (SD)							
Self-efficacy	10.52 (2.13)	10.30 (2.23)	11.25 (2.22)	10.85 (2.02)	11.20 (2.25)	10.62 (2.01)	11.56 (1.98)
Internal control	3.58 (0.76)	3.50 (0.75)	3.81 (0.69)	3.56 (0.75)	3.70 (0.83)	3.46 (0.77)	3.77 (0.79)
External control	2.72 (0.06)	2.65 (0.80)	2.40 (0.80)	2.63 (0.77)	2.33 (0.96)	2.61 (1.05)	2.32 (0.70)
Optimism	14.06 (4.63)	13.90 (4.48)	15.59 (4.62)	14.46 (4.44)	15.89 (4.86)	14.35 (4.72)	16.11 (4.37)
Self-compassion	2.75 (0.72)	2.76 (0.68)	3.27 (0.71)	2.87 (0.69)	3.19 (0.73)	2.87 (0.69)	3.25 (0.74)
Perceived support	26.90 (4.83)	26.16 (5.20)	27.92 (4.38)	26.96 (4.89)	27.63 (4.71)	26.68 (4.90)	27.99 (4.70)
Support seeking	13.48 (3.84)	13.40 (3.76)	14.55 (3.60)	13.62 (3.54)	14.58 (3.70)	13.39 (3.47)	14.75 (3.42)
Mental health, mean (SD)							
Depressive symptoms	10.85 (6.19)	11.33 (6.64)	9.61 (6.16)	11.41 (6.41)	9.61 (6.16)	11.12 (6.61)	9.68 (6.06)
Psychological distress	0.83 (0.53)	0.82 (0.55)	—[g]	—	0.68 (0.52)	0.74 (0.51)	—
Work-related health, mean (SD)							
Work ability	6.80 (1.99)	6.86 (1.73)	7.26 (1.98)	6.75 (1.90)	6.73 (2.27)	6.84 (1.85)	7.17 (2.15)
Effort	8.61 (2.30)	8.83 (2.11)	—	—	7.82 (2.47)	8.59 (2.35)	—
Reward	18.79 (3.75)	18.02 (3.93)	—	—	18.74 (3.92)	17.96 (4.04)	—
Over-commitment	15.99 (3.73)	16.06 (3.44)	14.74 (3.68)	15.61 (3.69)	14.33 (3.55)	15.69 (3.77)	14.28 (3.78)
Absenteeism	7.01 (7.75)	8.05 (11.35)	5.80 (7.88)	8.20 (12.44)	3.69 (4.48)	4.63 (5.75)	7.34 (9.29)
Presenteeism	6.34 (4.90)	7.32 (9.05)	6.53 (6.30)	6.67 (5.67)	6.66 (6.62)	5.26 (4.17)	4.08 (2.89)

[a]T1: baseline.

[b]T2: postintervention (8 weeks after randomization).

[c]T3: 3-month follow-up (3 months after randomization).

[d]T4: 6-month follow-up (6 months after randomization, intervention group only).

[e]IG: intervention group.

[f]WL: waitlist control group.

[g]Missing due to data management errors.

XSL•FO
RenderX

Table . Between-group differences postintervention and at 3-month follow-up for primary and resilience-related secondary outcome measures.

Outcome	Differences between the intervention group and the waitlist control group					
	T2[a]			T3[b]		
	F test (*df*)	*P* value	Cohen *d* (95% CI)	*F* test (*df*)	*P* value	Cohen *d* (95% CI)
Primary outcome						
Stress	21.31 (1, 518)	<.001	−0.54 (−0.75 to −0.34)	10.38 (1, 79)	.002	−0.47 (−0.71 to −0.22)
Secondary outcomes						
Self-perceived resilience	30.46 (1, 253)	<.001	0.47 (0.29 to 0.64)	9.33 (1, 72)	.003	0.29 (0.08 to 0.49)
Resilience factors						
Self-efficacy	4.15 (1, 248)	.04	0.12 (0.06 to 0.30)	4.87 (1, 69)	.03	0.21 (−0.03 to 0.45)
Internal control	11.23 (1, 191)	<.001	0.27 (0.06 to 0.47)	7.82 (1, 181)	.006	0.24 (0.03 to 0.45)
External control	6.82 (1, 132)	.01	−0.35 (−0.57 to −0.12)	4.76 (1, 81)	.03	−0.33 (−0.57 to −0.08)
Optimism	6.35 (1, 49)	.02	0.22 (0.16 to 0.60)	12.38 (1, 89)	.001	0.30 (0.11 to 0.48)
Self-compassion	36.23 (1, 106)	<.001	0.57 (0.39 to 0.76)	11.36 (1, 46)	.002	0.42 (0.18 to 0.66)
Perceived support	5.09 (1, 74)	.03	0.09 (−0.09 to 0.27)	4.18 (1, 78)	.04	0.09 (−0.10 to 0.27)
Support seeking	11.92 (1, 380)	.001	0.25 (−0.09 to 0.40)	12.89 (1, 92)	.001	0.32 (0.14 to 0.49)

[a]T2: postintervention (8 weeks after randomization).

[b]T3: 3-month follow-up (3 months after randomization).

Sensitivity Analyses

Sensitivity analysis based on study completers (T2: d=−0.64, 95% CI −0.89 to −0.39; T3: d=−0.55, 95% CI −0.82 to −0.28) corroborated the results obtained by intention-to-treat analysis. The group difference between intervention completers and the WL at T2, however, was only close to being statistically significant (P=.05, d=−0.42, 95% CI −0.79 to −0.06). In contrast, the group difference at T3 was significant and similar in size to that obtained by the intention-to-treat approach (d=−0.46, 95% CI −0.84 to −0.09). A linear mixed-effects model revealed results similar to those obtained by ANCOVA using imputed data. Equivalently, the interactions between time and intervention at T2 (P<.001, SE=0.75) and T3 (P<.001, SE=0.80) were significant, indicating that the change in stress differed between groups. Specifically, participants in the IG showed a significantly greater reduction in stress from T1 to T2 (β=−4.02, 95% CI −5.48 to −2.56), as well as from T1 to T3 (β=−3.67, 95% CI −5.24 to −2.11), compared to those on the WL.

Response and Deterioration Rates

At T2 and T3, significantly more participants in the IG than WL controls reported reliable and practically meaningful improvement, with NNTB ranging from 3.42 (95% CI 2.56 to 5.16) to 4.35 (95% CI 3.05 to 7.58). The rates of reliable and practically meaningful deterioration were higher in controls than in the IG at both postintervention points, with NNTH values ranging from 7.51 (95% CI 4.80 to 17.21) to 10.81 (95% CI 6.57 to 30.57). However, 4.6% (8/174) to 10.9% (19/174) of participants in the IG also experienced a worsening of stress symptoms. Multimedia Appendix 5 summarizes the results of the response and deterioration analysis.

Secondary Outcome Measures

Table 3 summarizes the results of resilience-related secondary outcome analysis. For self-perceived resilience, there was a significant effect in favor of active treatment at T2 (d=0.47) and T3 (d=0.29). For the resilience factors, there were significantly better outcomes among participants in the IG for all outcomes at T2 and T3. Immediately postintervention, effect sizes were 0.12 for self-efficacy, 0.27 for internal control, −0.35 for external control, 0.22 for optimism, 0.57 for self-compassion, 0.09 for perceived social support, and 0.25 for support seeking.

Multimedia Appendix 6 summarizes the findings of the other mental health- and work-related outcomes assessed. For depression, there was a significant between-group effect at T2 (d=−0.24), but not at T3. Regarding work-related outcomes, significant effects in favor of the IG were found for work ability at T2 (d=0.28), and for overcommitment at T2 (d=−0.22) and T3 (d=−0.36), but not for absenteeism or presenteeism at either of these postintervention measurement points.

XSL•FO
RenderX

Longer-Term Effects

Concerning the intervention's long-term effectiveness (T1-T4), outcomes remained stable, indicating that the beneficial effects were maintained. Within-group effect sizes for the primary and secondary outcomes were all significant, except for absenteeism. The effect size for stress was *d*=−0.80 (95% CI −0.97 to −0.63), for self-perceived resilience *d*=1.19, for self-efficacy *d*=0.80, for optimism *d*=0.87, for self-compassion *d*=1.25, and for perceived social support *d*=0.37. A table summarizing the results of the within-group comparisons obtained by repeated-measures ANOVA is available in Multimedia Appendix 7.

Mediation Analyses

Perceived Stress

Single mediation analysis revealed that indirect effects through optimism (ab=−0.46, 95% CI −0.80 to −0.12) and self-compassion (ab=−0.95, 95% CI −1.49 to −.41) were statistically significant, indicating that these variables mediated the effect of the intervention on perceived stress at T3. In contrast, indirect effects through self-efficacy (ab=−0.11; 95% CI −0.34 to 0.12) and perceived support (ab=−0.04; 95% CI −0.18 to 0.10) were not significant.

As depicted in Figure 2, the results of parallel mediation analysis aligned with the results obtained with single mediation analyses. In a joint model, the indirect effects through optimism (a_2b_2=−0.34, 95% CI −0.63 to −0.06) and self-compassion (a_4b_4=−0.66, 95% CI −1.15 to −0.17) were again statistically significant, while the indirect effects through self-efficacy (a_1b_1=−0.02, 95% CI −.14 to 0.10) and perceived support (a_3b_3=0.04, 95% CI −0.09 to 0.16) were not. The direct effect of the intervention on perceived stress remained significant after incorporating the mediators in the model (*c'*=−2.25, 95% CI −3.79 to −0.70).

Figure 2. Parallel multiple mediation model with 3-month follow-up (T3) stress severity scores as the outcome variable (Y), posttreatment (T2) resilience factors scores as mediators, and baseline values of mediators and outcome as covariates. Intervention (X) is coded 1=IG, 0=WL. Unstandardized beta coefficients are shown with 95% CIs in brackets. IG: intervention group; WL: waitlist control group.

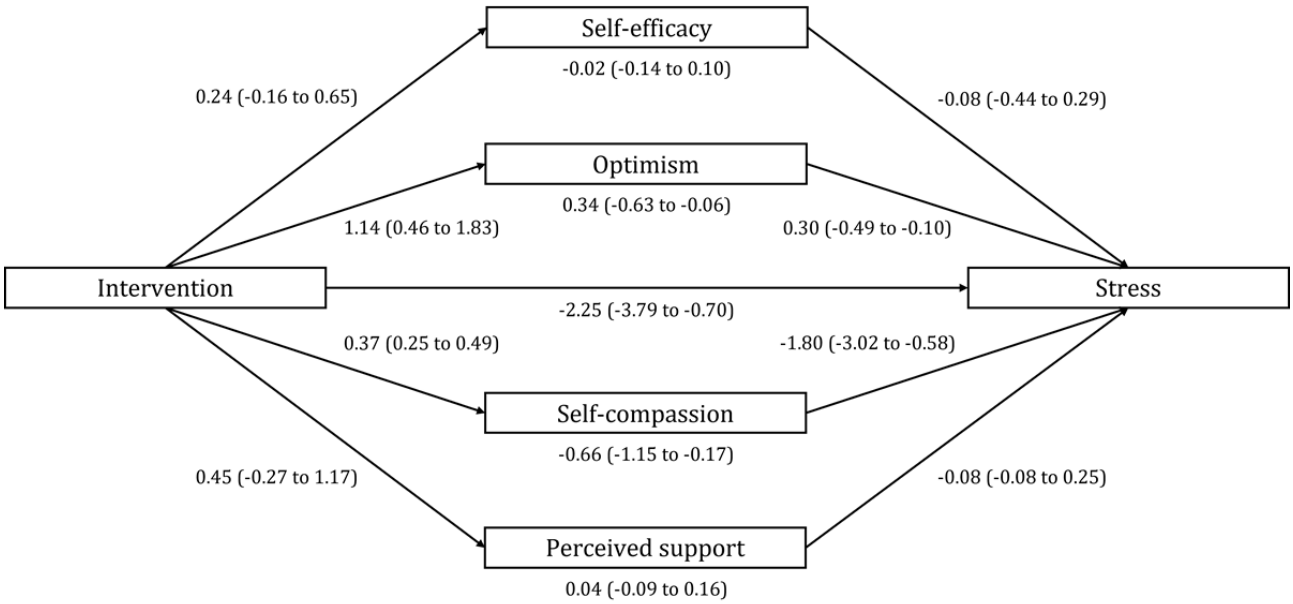

Self-Perceived Resilience

Mediation analyses with self-perceived resilience—that is, the ability to recover from stress—as a dependent outcome revealed a pattern similar to that observed with perceived stress as the outcome. Single mediation analyses showed that the indirect effects through optimism (ab=0.27, 95% CI 0.07 to 0.46) and self-compassion (ab=0.46, 95% CI 0.16 to 0.76) were statistically significant, indicating that these variables mediated the intervention's effect on self-perceived resilience at T3. In contrast, indirect effects through self-efficacy (ab=0.08; 95% CI −0.06 to 0.22) and perceived support (ab=0.01; 95% CI −0.07 to 0.08) were not.

On parallel mediation analysis, the indirect effect through optimism (a_2b_2=0.21, 95% CI 0.06 to 0.37) significantly mediated the intervention's effect on resilience at T3, while the indirect effect through self-compassion (a_4b_4=0.26, 95% CI −0.01 to 0.53), though close, failed to achieve statistical significance. The indirect effects achieved through self-efficacy (a_1b_1=0.04, 95% CI −0.03 to 0.11) and perceived support (a_3b_3=−0.03, 95% CI −0.10 to 0.04) were also not statistically significant. The direct effect of the intervention on self-perceived resilience remained significant after incorporating mediators into the model (*c'*=0.86, 95% CI 0.04 to 1.65). The corresponding mediation model is shown in Figure S1 in Multimedia Appendix 8.

Repeating all single and parallel mediation analyses with this study completer sample as additional sensitivity analyses revealed no differences in the results relative to those obtained with the intention-to-treat approach (see Figures S2 and S3 in Multimedia Appendix 8). The only exception was in the joint model with self-perceived resilience as the dependent variable, in which the indirect effect via self-compassion reached statistical significance (a_4b_4=0.43, 95% CI 0.13 to 0.73).

XSL•FO
RenderX

Intervention Usage and Client Satisfaction

Most participants started with the web component of the intervention (156/176, 88.6%), completing an average of 2.2 (SD 2.3) sessions. Completion rates ranged from 70.5% (110/156) for session 1 to 18.6% (29/156) for all 6 sessions.

Those who used the app component (104/176, 59.1%) collected a median of 14 moments of resilience (range 1 to 220), indicating that half collected these moments almost twice per week over the 8-week intervention period. Based upon participant reports, users collected a total of 3064 moments of resilience, most frequently for self-compassion. Detailed engagement patterns are summarized in Multimedia Appendix 9. Of those participants reporting their intervention usage at T2, 38.1% (32/84) indicated they had not yet finished and expressed an intention to continue.

Participants' satisfaction with the intervention (considering both the web and app components) was high (mean 24.02, SD 6.18). Of those participants providing questionnaire data about their satisfaction with the intervention, 87/99 (87.9%) indicated satisfaction in an "overall, general sense" ("very satisfied" or "mostly satisfied"), and 80/99 (80.5%) indicated satisfaction with the amount of help they received. In addition, 73/99 (73.7%) agreed that the intervention helped them to deal with problems more effectively, and 77/99 (77.8%) said they would recommend the intervention to a friend in need of similar help. The overall user experience of the app was rated as 4.71 (SD 1.04; range 1 to 7; n=99), indicating an overall good level of user experience [53].

Discussion

Principal Results

Overview

For the first time, this study identified a favorable effect of the RESIST intervention, relative to being on a waiting list, on perceived stress in a universal prevention setting. Improvements were evident immediately after the intervention and remained stable 3 and 6 months after randomization. A similar pattern was observed for self-perceived resilience and the resilience factors targeted within the intervention. Effects on other health- and work-related outcomes were mixed.

Concerning the second study aim, mediation analyses suggested that the resilience factors targeted within the training played distinct roles in how the intervention exerted its effects on stress and self-perceived resilience. The strongest evidence for mediation emerged for optimism and self-compassion, with significant indirect effects observed for both variables. In contrast, indirect effects through self-efficacy and social support were not statistically significant.

Effects on Stress and Self-Perceived Resilience

The observed effect on stress reduction immediately postintervention (d=−0.54) slightly exceeded the effect size expected from the a priori sample size calculation (d=−0.40). Building on the findings from the pilot study [17], this study confirmed the intervention's effectiveness in a larger sample within a universal preventive setting, demonstrating that the intervention is effective not only when delivered with guidance, as in the pilot study, but also in a self-help format that requires fewer resources.

When compared with interventions targeting a similar set of resilience factors that may be unique to resilience interventions [21,66], the effect sizes for stress and self-perceived resilience (d=0.47) observed in this study fall between those reported for a web-based resilience intervention for college students [66] (d=−0.34 for stress) and a multicomponent positive psychology intervention for the general population [21] (d=0.67 for resilience).

Concerning intervention techniques, this study's findings align with those of 2 previous studies that also used Strengths-Based CBT [67,68]. In one of these studies [68], distressed college students showed significant postintervention improvements in resilience (d=0.34).

Compared to meta-analytic findings, the stress reduction effect observed for the present intervention postassessment was substantially larger than the nonsignificant effect (g=0.14) reported by Ang et al [7], who used a broad resilience factor approach. In contrast, the findings related to self-perceived resilience closely align with those of Ang et al [7] (g=0.54). The effect of RESIST on self-perceived resilience is also similar in magnitude to that reported for the meta-analysis by Díaz-García et al [8], which adopted an outcome-focused resilience approach, including interventions aimed at modifying resilience or related constructs regardless of the interventions' characteristics. Furthermore, the stress reduction effect of RESIST was slightly larger than the effect size for stress reported in the meta-analysis by Schäfer et al [11] (standard mean difference=0.33, 95% CI −0.41 to −0.24), which applied a broad conceptualization of resilience interventions, also including established intervention formats such as stress-management and mindfulness programs. The effects for self-perceived resilience found in the present study also surpassed those reported by Schäfer et al [11] (standard mean difference=0.22).

Although these comparisons do not establish superiority of interventions that target a specific set of resilience factors that may reflect more overarching resilience processes and may be correlates of higher-order resilience mechanisms, such as a positive appraisal style, the findings suggest that it is valuable to continue developing and investigating interventions based on these principles. This is particularly important if resilience interventions are to be seen as a distinct form of intervention.

The comparisons further underscore the challenges of adequately evaluating and comparing resilience interventions, stemming from the lack of a clear understanding of what constitutes a resilience intervention. The variability in the design of resilience interventions makes it difficult to identify comparable interventions, especially when investigators fail to adequately describe the intervention's characteristics. As mentioned earlier, such conceptual ambiguity is reflected in meta-analyses, where differing definitions of resilience interventions can lead to substantially heterogeneous results and conclusions between meta-analyses. This emphasizes the need for a standardized classification system for resilience interventions. Building on

XSL•FO
RenderX

prior considerations by Chmitorz et al [12], (1) the theory or rationale underlying an intervention, (2) its specific content, (3) the intervention techniques applied, and (4) the timing of the intervention relative to stressor exposure should be considered as dimensions for building a systematic classification system of resilience interventions. This may contribute to identifying the core elements of effective intervention designs.

Response and Deterioration Rates

The NNTB suggests that, on average, three to four individuals must be given access to RESIST for a single individual to experience 20% symptom improvement or reliable improvement in stress symptoms immediately after the intervention. This aligns with the results of other digital mental health interventions such as self-guided stress management training [69]. Corresponding to findings from prior research [70,71], it must be noted, however, that some individuals (12/174, 6.9%) in our IG met the criterion for reliable deterioration. This highlights safety issues within resilience interventions. While resilience interventions labeled positively might seem more appealing than, for example, stress management interventions, this does not guarantee symptom improvement, as some users may experience adverse effects. Potential risks may include heightened frustration or self-blame if participants do not meet perceived expectations of "being resilient" [72]. The positively framed concept of "resilience" may furthermore create implicit pressure to bounce back, potentially reinforcing feelings of inadequacy when setbacks persist despite effort. In the systematic review by Chmitorz et al [12], none of the 43 RCTs included in the analysis assessed adverse effects. Hence, the potential harms of so-called "positive" interventions remain largely overlooked [71], underscoring the need for more conceptual and empirical research in this area.

Effects on Resilience Factors

Besides examining RESIST's effectiveness in reducing stress as a primary outcome and resilience as an important secondary outcome, we also investigated whether the intervention promoted the resilience factors of self-efficacy, optimism, self-compassion, and social support that the intervention was designed to target. The effects observed postintervention were favorable, yet variable in magnitude.

For self-efficacy, we observed a small but statistically significant between-group effect after the intervention (d=0.12). This is numerically similar to the effect reported by Schäfer et al [11] in their meta-analysis (d=0.01). Regarding optimism (ie, the general expectation that one's own outcomes will be positive [73]), our findings indicated a more sizeable intervention effect (d=0.22), in line with results from meta-analyses by Schäfer et al [11] and Malouff and Schutte [74]. The most pronounced effect at T2 was observed for self-compassion (d=0.57; ie, a supportive attitude toward oneself in times of suffering [75]). This finding mirrors results reported by Schäfer et al [11], where self-compassion also showed the largest effect among the resilience factors examined. For perceived social support, we found small between-group differences at T2 (d=0.09), which also aligns with published meta-analyses [31]. Despite the statistically significant findings, the practical significance of these effects is difficult to interpret due to the lack of established criteria for meaningful change in resilience factors. Future research should investigate such criteria, following the methodological recommendations of Cook et al [76].

Mediation Analyses

The mediation analyses further revealed that RESIST not only favorably fostered optimism and self-compassion, but that these factors also contributed to how the intervention exerted its positive effects. With one exception—self-compassion mediated the effect on self-perceived resilience only in this study's completer sample—the mediation results were consistent, irrespective of whether stress or self-perceived resilience was assessed as the outcome, underscoring that the same resilience factors may matter for both positive and negative mental health outcomes. Comparing these mediation findings to prior research results is limited by the scarcity of studies that have investigated the mechanisms of change underlying resilience interventions. Our mediation findings align with isolated prior findings of the previously mentioned multicomponent email-guided positive psychology intervention [21], for which both optimism and self-compassion also emerged as mediators of the intervention's effects.

The findings shed light on the importance of optimism and self-compassion as potential key processes in resilience promotion. In the case of optimism, enhancing the belief that one's own outcomes will be positive may both strengthen someone's confidence that a currently stressful situation can improve over time and foster the perception that such situations are manageable and temporary, thereby reducing perceived stress and strengthening self-perceived resilience. This, in turn, might reflect a positive appraisal style of stressors. This finding aligns with a substantial body of prior research, including both meta-analyses and longitudinal studies, that have revealed the importance of optimism for positive mental health outcomes [30,77-80]. For instance, in a meta-analysis published by Gallagher et al [30], a negative association was identified between optimism and posttraumatic stress disorder. Meanwhile, Romswinkel et al [80] identified optimism both as a predictor of reduced depressive symptoms and as a mediator in the relationship between job stress and depression.

Regarding self-compassion, fostering a kinder and warmer attitude toward oneself during times of stress might have enabled IG participants to better care for themselves and respond with reduced self-criticism, particularly when facing stressors rooted in personal mistakes or feelings of inadequacy [75]. By promoting a more accepting and supportive internal response, the intervention may have helped them to reinterpret such stressors more constructively, thereby establishing a link to a positive appraisal style and ultimately reducing perceived stress. Empirical evidence supports this strong link between self-compassion and mental health [81-85]. For example, one meta-analysis has demonstrated an inverse association between self-compassion and stress [82], while a longitudinal study by Lee et al [81] identified self-compassion as a significant predictor of mental well-being over five years.

In contrast to self-compassion and optimism, no significant mediation effects were observed for self-efficacy and social support. However, it is premature to discount the roles of

XSL•FO
RenderX

self-efficacy and social support in interventions aiming to promote resilience. Differences in the control and ease of implementation, as well as the timescale of observable change, may partly explain the pattern of effects. Self-compassion, for example, can be cultivated through self-directed practices, as evidenced by participants' predominant selection of moments of resilience related to this factor, according to app data. In contrast, fostering self-efficacy could rely more on sufficient learning opportunities and external feedback [86], which may be provided by a mental health professional, such as an eCoach. Increases in perceived social support might depend more on social interactions and environmental factors, potentially requiring more time to become manifest. Including peer support in the intervention could serve as one valuable option to enhance these social interactions.

Strengths and Limitations

This study has several key strengths that help it contribute to the body of research in the field. Its first strength is that it was conducted in accordance with the recommendations for resilience intervention studies outlined by Chmitorz et al [12], including an outcome-oriented resilience definition, and the separate assessment of various mental health outcomes and resilience factors, as well as adverse effects.

Second, to the best of our knowledge, the resilience intervention RESIST examined in this study is among the first to be explicitly grounded in a genuine resilience framework—namely, PASTOR [2], which served as the theoretical foundation of the intervention.

Third, the intervention's content was informed by resilience factors considered etiologically relevant for fostering adaptive responses to adversity [87] and which may reflect resilience-specific competencies.

Fourth, the intervention specifically targeted resilience factors that are typically not addressed in established mental health programs, such as stress management or mindfulness-based interventions. Lastly, by studying mediators, this study addresses the call to study mechanisms of change to optimize intervention development and treatment outcomes from digital resilience interventions [11].

Despite this study's strengths, the results should be interpreted in light of some limitations, with the first two relating to generalizability, the next two to methodological aspects, and the final two to theoretical considerations.

First, the findings may have limited generalizability to other implementation settings; for example, RESIST's effectiveness could differ if the intervention is employer-provided rather than individually initiated. Employees may view employer-led resilience training as contradictory if it ignores structural causes of stress. This concern is supported by a meta-analysis that detected smaller-than-expected effects for organization-implemented e-mental health interventions [88].

Second, this study's sample was predominantly female and highly educated, which may also limit generalizability. However, systematic reviews suggest that sociodemographic factors such as gender or education rarely modify intervention effects [89-91]. The main challenge may therefore be promoting uptake in other groups. Whether this may be achieved, for example, via approaches such as participatory design, tailored recruitment, or interventions addressing mental health indirectly (eg, physical exercise [92]), is not clear yet.

Third, the dropout rate across both intervention arms (82/350, 23.4%) could have negatively affected the results' validity. However, this extent of attrition falls well within the range of similar digital self-help interventions [69,71], and the multiple imputations approach used to address this shortfall is both a robust and widely accepted method for handling missing data [93]. In addition, a range of sensitivity analyses, including mixed model analyses, supported the robustness of the intention-to-treat sample's findings.

Fourth, nearly 40% (32/83) of participants indicated postintervention that they had not yet completed the intervention but intended to continue engaging with it. This may reflect the nature of the self-help format, which typically allows for greater flexibility. It also suggests that the intended pacing, one session and at least 2 app entries per week, may not have been well-matched to participants' preferences or natural usage patterns.

Fifth**,** the applied mediation model implicitly assumes that all mediators change simultaneously over time, thereby limiting its capacity to account for temporal variability in their effects—some mediators may exert their influence more quickly, while others do so more slowly. This limitation is common in studies assessing such models.

Lastly, the interpretability of the mediation analyses' results is limited with regard to drawing conclusions about any potential causal link between specific mediators and individual training components. For instance, it cannot be concluded that the content related to the resilience factors within the training necessarily encompasses the active ingredients driving the intervention's effects. Component studies are needed to further investigate the active ingredients of RESIST.

Theoretical Implications

Based on the study's insights, various theoretical implications and perspectives for future research emerge. First, almost 90% (309/350) of the participants reported experiencing an effort-reward imbalance in their work life and, accordingly, trained resilience in the context of ongoing stressor exposure. A next step would be to more precisely assess the specific stressors experienced by participants and relate them to mental health outcomes, as proposed by Kalisch et al [94]. This would, for instance, enable meaningful comparisons between individuals who show similar improvements in mental health after training, despite facing different levels of adversity.

Second, while the mediation analyses provided initial and valuable insights into the intervention's mechanisms of change, they also highlight several avenues for future research. The precise interrelations, causal directions, and underlying mechanisms linking the resilience factors optimism, self-compassion, and a positive appraisal style as potential resilience mechanisms warrant further investigation to deepen our understanding of their contribution to psychological

XSL•FO
RenderX

adaptation and resilience. Further, for the investigation of RESIST, a logical next step will be to include an assessment of positive appraisal style [95] as an additional construct (which has been published just recently) in future assessments and to examine it as a potential mediator in sequential mediation, onto which the various resilience factors may converge. Additionally, as already discussed, our findings suggest a lack of understanding regarding the varying timeframes required for different resilience factors to develop and change, as well as the degree to which their change can be regulated internally. Regarding the further development of resilience interventions and assuming that resilience factors require more time or external support to develop, so that change becomes better understood, such components could be positioned strategically at the beginning or end of an intervention. For research on mechanisms of change, this underscores the need for a more fine-grained level of investigation of mediators that also respects strict temporal precedence. One promising approach would be to assess the mediating role of resilience factors during the course of the intervention itself. To this end, methods such as session-by-session assessments or ecological momentary assessments in daily life between sessions may be particularly useful [96,97]. An alternative approach could involve using dynamic network models to capture the complexity inherent in change processes within interventions [98].

Third, by targeting multiple resilience factors, our intervention implicitly emphasized the importance of having access to a broad repertoire of resources to draw on various strategies, so individuals are able to respond flexibly to different challenging situations. This idea is reflected in the concept of regulatory flexibility [14], which emphasizes the importance of a good strategy-situation fit, as no single strategy is effective for all stressors. Explicitly integrating elements reflecting such a fit-based perspective into training could be a promising direction for future intervention development, as it could help individuals to learn how to select and apply the most suitable strategies depending on the specific demands of each situation. Examining changes in individuals' repertoire could be a further meaningful outcome for future intervention studies.

Lastly, given that individualized resilience interventions have not yet been explored sufficiently [11], tailoring the content of RESIST based on individual resource profiles could be a promising avenue for future research. For one study, Fassnacht et al [70] adopted a similar approach by integrating an assessment tool into their intervention, which provided participants with a profile of personal strengths and vulnerabilities to support informed decisions about which areas they should focus on. Building on this idea, incorporating a resilience profile assessment to identify participants' existing psychosocial resources—upon which the training might then be constructed—could personalize the approach and potentially enhance its effectiveness.

Practical Implications

The present findings yield several practical implications for implementing digital resilience training, such as RESIST, in occupational settings.

First, adherence to the self-help version of RESIST was low, reflecting a common challenge in digital interventions [99-101]. To enhance adherence, employers should consider recognizing participation in trainings such as RESIST as working time [102]. Low-cost technical measures, such as automated reminders, may also mitigate this issue [103]. More resource-intensive strategies, including personal recruitment, financial incentives, or professional guidance, can further improve engagement and adherence [104,105]. The use of adaptive systems—for example, offering on-demand support during self-help programs or enabling participants in guided formats to opt out of support when preferring independent training—can help limit associated costs. Cost-effectiveness analyses may assist in determining the optimal balance between resource investment to enhance adherence and the resulting health benefits in the population [106].

Second, the observed deterioration rates highlight the need for safety measures, such as providing personal support alongside the intervention or offering clear guidance on where to seek professional external help. Even if the reasons for deterioration remain unclear, as in this study, occupational health care professionals should be aware of potential deterioration and take ethical responsibility into account when offering mental health interventions.

Third, when inviting employees to participate in interventions, occupational health care professionals should communicate potential health benefits appropriately, given the importance of expectations for usage intentions, actual usage, and outcomes [107,108]. The observed response and deterioration rates, together with NNTB and NNTH, provide a solid basis for managing such expectations. Indicators based on standardized mean differences (eg, Cohen *d*) are frequently used but can be difficult for nonresearchers to understand [59] and may not be effective for expectation management.

Fourth, promoting the resilience of individual employees is an important but selective aspect of mental well-being. Interventions such as RESIST should be embedded in a comprehensive occupational health promotion strategy that includes a balanced set of measures addressing both individual and structural, workplace-related aspects [102]. Work-directed interventions—such as improving leadership culture, optimizing workloads, or enhancing effective communication within teams—should complement individual programs. This dual focus helps prevent the impression that employees are solely responsible for their well-being and reinforces the shared responsibility between staff and employers [102].

Conclusions

This study is among the first to evaluate the effectiveness of a self-help digital resilience intervention that was designed to enhance mental health during stressful times by promoting a set of theory- and evidence-informed selection of resilience factors representing trainable competencies. According to the PASTOR framework, they may serve as correlates of a positive appraisal style of stressors as a higher-order resilience mechanism. The interventional techniques used in the intervention, according to Strengths-Based CBT, proved to be feasible and effective, which confirms the results of the pilot

XSL•FO
RenderX

study [17]. Especially when intervention developers find the idea convincing that the methods used to promote resilience should be different from the methods used in psychotherapy aiming at reducing distress, then Strengths-Based CBT appears to be a promising option.

By studying resilience factors as mediators, this trial further addressed the recent call to investigate mechanisms of change within digital resilience interventions [11] to optimize interventional design and outcomes. Mediation analysis suggested that optimism and self-compassion may be important drivers promoting resilient outcomes, while redesigning the intervention seems to be needed to achieve stronger effects for self-efficacy and perceived social support. Future research is needed to investigate the relationships between resilience factors and higher-order resilience mechanisms, the ease of implementation, and the temporal dynamics of how resilience factors exert their protective effects. Finally, a standardized classification system for resilience interventions and a consensus on what constitutes such interventions are urgently needed to enable meaningful comparisons between resilience interventions. This should form the basis of a structured process to improve resilience interventions and make them more effective.

Acknowledgments

Generative artificial intelligence was used to assist in improving spelling and grammar in the initial draft of this paper. The final version was subsequently proofread by a professional editor.

Funding

This publication was funded by the Open Access Publication Fund of Leuphana University Lüneburg. The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of this paper.

Data Availability

The dataset generated and analyzed during this study is available in PubData, the institutional repository of Leuphana University for archiving and publishing datasets [109]. Access is provided by the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: DB (lead), DL (supporting), MW (supporting)
Data curation: SH
Formal analysis: SH
Investigation: DB
Methodology: DB (equal), DL (equal)
Project administration: DB
Supervision: DL
Visualization: SH
Writing – original draft: SH
Writing – review & editing: SH (lead), DB (equal), ND (supporting), SKS (supporting), MW (supporting), DL (equal)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Information on intervention description according to the TIDieR (Template for Intervention Description and Replication) checklist.
[DOCX File, 31 KB - jmir_v28i1e78335_app1.docx]

Multimedia Appendix 2
Intervention content of the web- and app-based resilience training RESIST.
[DOCX File, 17 KB - jmir_v28i1e78335_app2.docx]

Multimedia Appendix 3
Screenshots of the web- and app-based components of RESIST.
[DOCX File, 3645 KB - jmir_v28i1e78335_app3.docx]

Multimedia Appendix 4
Response formats and reliability of patient-reported outcomes measures.
[DOCX File, 15 KB - jmir_v28i1e78335_app4.docx]

XSL•FO
RenderX

Multimedia Appendix 5
Response and deterioration rates for the primary outcome measure of stress.
[DOCX File, 16 KB - jmir_v28i1e78335_app5.docx]

Multimedia Appendix 6
Between-group differences postintervention and at 3-month follow-up for mental health- and work-related secondary outcome measures.
[DOCX File, 18 KB - jmir_v28i1e78335_app6.docx]

Multimedia Appendix 7
Within-subject comparisons for the intervention group from baseline to 6-month follow-up.
[DOCX File, 18 KB - jmir_v28i1e78335_app7.docx]

Multimedia Appendix 8
Additional mediation analyses' results.
[DOCX File, 265 KB - jmir_v28i1e78335_app8.docx]

Multimedia Appendix 9
Engagement with the web and app components of the intervention.
[DOCX File, 14 KB - jmir_v28i1e78335_app9.docx]

Checklist 1
CONSORT-eHEALTH checklist (V1.6).
[PDF File, 1219 KB - jmir_v28i1e78335_app10.pdf]

References

1. Bonanno GA, Westphal M, Mancini AD. Resilience to loss and potential trauma. Annu Rev Clin Psychol 2011;7(1):511-535. [doi: 10.1146/annurev-clinpsy-032210-104526] [Medline: 21091190]
2. Kalisch R, Müller MB, Tüscher O. A conceptual framework for the neurobiological study of resilience. Behav Brain Sci 2015;38:e92. [doi: 10.1017/S0140525X1400082X]
3. Kalisch R, Baker DG, Basten U, et al. The resilience framework as a strategy to combat stress-related disorders. Nat Hum Behav 2017 Nov;1(11):784-790. [doi: 10.1038/s41562-017-0200-8] [Medline: 31024125]
4. Köhler CA, Evangelou E, Stubbs B, et al. Mapping risk factors for depression across the lifespan: an umbrella review of evidence from meta-analyses and Mendelian randomization studies. J Psychiatr Res 2018 Aug;103:189-207. [doi: 10.1016/j.jpsychires.2018.05.020]
5. Vos T, Lim SS, Abbafati C, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct;396(10258):1204-1222. [doi: 10.1016/S0140-6736(20)30925-9]
6. Santomauro DF, Herrera AMM, Shadid J, et al. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 2021 Nov;398(10312):1700-1712. [doi: 10.1016/S0140-6736(21)02143-7]
7. Ang WHD, Chew HSJ, Dong J, Yi H, Mahendren R, Lau Y. Digital training for building resilience: systematic review, meta-analysis, and meta-regression. Stress Health 2022 Dec;38(5):848-869. [doi: 10.1002/smi.3154] [Medline: 35460533]
8. Díaz-García A, Franke M, Herrero R, Ebert DD, Botella C. Theoretical adequacy, methodological quality and efficacy of online interventions targeting resilience: a systematic review and meta-analysis. Eur J Public Health 2021 Jul 7;31(31 Suppl 1):i11-i18. [doi: 10.1093/eurpub/ckaa255] [Medline: 34240159]
9. Ferreira M, Marques A, Gomes PV. Individual resilience interventions: a systematic review in adult population samples over the last decade. Int J Environ Res Public Health 2021 Jul 16;18(14):7564. [doi: 10.3390/ijerph18147564] [Medline: 34300018]
10. Joyce S, Shand F, Tighe J, Laurent SJ, Bryant RA, Harvey SB. Road to resilience: a systematic review and meta-analysis of resilience training programmes and interventions. BMJ Open 2018 Jun 14;8(6):e017858. [doi: 10.1136/bmjopen-2017-017858] [Medline: 29903782]
11. Schäfer SK, von Boros L, Schaubruch LM, et al. Digital interventions to promote psychological resilience: a systematic review and meta-analysis. NPJ Digit Med 2024 Feb 8;7(1):30. [doi: 10.1038/s41746-024-01017-8] [Medline: 38332030]
12. Chmitorz A, Kunzler A, Helmreich I, et al. Intervention studies to foster resilience – a systematic review and proposal for a resilience framework in future intervention studies. Clin Psychol Rev 2018 Feb;59:78-100. [doi: 10.1016/j.cpr.2017.11.002]
13. Folkman S. Stress: Appraisal and Coping: Springer; 1984. [doi: 10.1007/978-1-4419-1005-9_215]

XSL•FO
RenderX

14. Bonanno GA, Chen S, Galatzer-Levy IR. Resilience to potential trauma and adversity through regulatory flexibility. Nat Rev Psychol 2023;2(11):663-675. [doi: 10.1038/s44159-023-00233-5]
15. Harrer M, Adam SH, Fleischmann RJ, et al. Effectiveness of an internet- and app-based intervention for college students with elevated stress: randomized controlled trial. J Med Internet Res 2018;20(4):e136. [doi: 10.2196/jmir.9293]
16. Aikens KA, Astin J, Pelletier KR, et al. Mindfulness goes to work: impact of an online workplace intervention. J Occup Environ Med 2014 Jul;56(7):721-731. [doi: 10.1097/JOM.0000000000000209] [Medline: 24988100]
17. Behrendt D, Boß L, Hannibal S, Kunzler AM, Wessa M, Lehr D. Feasibility and efficacy of a digital resilience training: a pilot study of the strengths-based training RESIST. Internet Interv 2023 Sep;33:100649. [doi: 10.1016/j.invent.2023.100649] [Medline: 37545556]
18. Kazdin AE. Mediators and mechanisms of change in psychotherapy research. Annu Rev Clin Psychol 2007;3(1):1-27. [doi: 10.1146/annurev.clinpsy.3.022806.091432] [Medline: 17716046]
19. Kraemer HC, Wilson GT, Fairburn CG, Agras WS. Mediators and moderators of treatment effects in randomized clinical trials. Arch Gen Psychiatry 2002 Oct;59(10):877-883. [doi: 10.1001/archpsyc.59.10.877] [Medline: 12365874]
20. Cieslak R, Benight CC, Rogala A, et al. Effects of internet-based self-efficacy intervention on secondary traumatic stress and secondary posttraumatic growth among health and human services professionals exposed to indirect trauma. Front Psychol 2016;7:1009. [doi: 10.3389/fpsyg.2016.01009] [Medline: 27458407]
21. Schotanus-Dijkstra M, Pieterse ME, Drossaert CHC, Walburg JA, Bohlmeijer ET. Possible mechanisms in a multicomponent email guided positive psychology intervention to improve mental well-being, anxiety and depression: a multiple mediation model. J Posit Psychol 2019 Mar 4;14(2):141-155. [doi: 10.1080/17439760.2017.1388430]
22. Forbes S, Fikretoglu D. Building resilience: the conceptual basis and research evidence for resilience training programs. Rev Gen Psychol 2018 Dec;22(4):452-468. [doi: 10.1037/gpr0000152]
23. Padesky CA, Mooney KA. Strengths-based cognitive-behavioural therapy: a four-step model to build resilience. Clin Psychol Psychother 2012;19(4):283-290. [doi: 10.1002/cpp.1795] [Medline: 22653834]
24. Eysenbach G, CONSORT-eHEALTH Group. CONSORT-eHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
25. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: Template for Intervention Description and Replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348:g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]
26. Beck AT, Steer RA, Brown G. Beck Depression Inventory–II: American Psychological Association; 1996. [doi: 10.1037/t00742-000]
27. Heber E, Ebert DD, Lehr D, et al. The benefit of web- and computer-based interventions for stress: a systematic review and meta-analysis. J Med Internet Res 2017 Feb 17;19(2):e32. [doi: 10.2196/jmir.5774] [Medline: 28213341]
28. Stewart DE, Yuen T. A systematic review of resilience in the physically ill. Psychosomatics 2011;52(3):199-209. [doi: 10.1016/j.psym.2011.01.036] [Medline: 21565591]
29. Khazanov GK, Ruscio AM. Is low positive emotionality a specific risk factor for depression? A meta-analysis of longitudinal studies. Psychol Bull 2016 Sep;142(9):991-1015. [doi: 10.1037/bul0000059] [Medline: 27416140]
30. Gallagher MW, Long LJ, Phillips CA. Hope, optimism, self-efficacy, and posttraumatic stress disorder: a meta-analytic review of the protective effects of positive expectancies. J Clin Psychol 2020 Mar;76(3):329-355. [doi: 10.1002/jclp.22882] [Medline: 31714617]
31. Wang Y, Chung MC, Wang N, Yu X, Kenardy J. Social support and posttraumatic stress disorder: a meta-analysis of longitudinal studies. Clin Psychol Rev 2021 Apr;85:101998. [doi: 10.1016/j.cpr.2021.101998] [Medline: 33714168]
32. GetOn Platform. URL: https://coach.geton-training.de/ [accessed 2025-12-04]
33. Heber E, Lehr D, Ebert DD, Berking M, Riper H. Web-based and mobile stress management intervention for employees: a randomized controlled trial. J Med Internet Res 2016 Jan 27;18(1):e21. [doi: 10.2196/jmir.5112] [Medline: 26818683]
34. Klein EM, Brähler E, Dreier M, et al. The German version of the Perceived Stress Scale - psychometric characteristics in a representative German community sample. BMC Psychiatry 2016 May 23;16(1):159. [doi: 10.1186/s12888-016-0875-9] [Medline: 27216151]
35. Reis D, Lehr D, Heber E, Ebert DD. The German Version of the Perceived Stress Scale (PSS-10): evaluation of dimensionality, validity, and measurement invariance with exploratory and confirmatory bifactor modeling. Assessment 2019 Oct;26(7):1246-1259. [doi: 10.1177/1073191117715731] [Medline: 28627220]
36. Chmitorz A, Wenzel M, Stieglitz RD, et al. Population-based validation of a German version of the Brief Resilience Scale. PLoS One 2018;13(2):e0192761. [doi: 10.1371/journal.pone.0192761] [Medline: 29438435]
37. Beierlein C, Kovaleva A, Kemper CJ, Rammstedt B. Ein Messinstrument zur Erfassung subjektiver Kompetenzerwartungen: Allgemeine Selbstwirksamkeit Kurzskala (ASKU) [Website in German]. 2012. URL: https://www.ssoar.info/ssoar/bitstream/handle/document/29235/ssoar-2012-beierlein_et_al-ein_messinstrument_zur_erfassung_subjektiver.pdf?sequence=1 [accessed 2025-12-04]
38. Glaesmer H, Hoyer J, Klotsche J, Herzberg PY. Die Deutsche version des Life-Orientation-Tests (LOT-R) zum dispositionellen optimismus und pessimismus [Article in German]. Z Gesundheitspsychologie 2008 Jan;16(1):26-31. [doi: 10.1026/0943-8149.16.1.26]

XSL•FO
RenderX

39. Hupfeld J, Ruffieux N. Validierung einer Deutschen version der Self-Compassion Scale (SCS-D) [Article in German]. Z Klin Psychol Psychother 2011 Apr;40(2):115-123. [doi: 10.1026/1616-3443/a000088]
40. Schulz U, Schwarzer R. Soziale unterstützung bei der krankheitsbewältigung: die Berliner Social Support Skalen (BSSS) [Article in German]. Diagnostica 2003 Apr;49(2):73-82. [doi: 10.1026//0012-1924.49.2.73]
41. Hautzinger M, Bailer M, Hofmeister D. Center for Epidemiological Studies Depression Scale (CES-D; Radloff, LS, 1977)-German adaptation. Psychiatr Prax 2012;39(6):302-304. [doi: 10.1055/s-0032-1326702]
42. Lehr D, Hillert A, Schmitz E, Sosnowsky N. Screening depressiver Störungen mittels Allgemeiner Depressions-Skala (ADS-K) und State-Trait Depressions Scales (STDS-T). Diagnostica 2008 Apr;54(2):61-70 [FREE Full text] [doi: 10.1026/0012-1924.54.2.61]
43. Spitzer C, Hammer S, Löwe B, et al. Die kurzform des Brief Symptom Inventory (BSI -18): erste befunde zu den psychometrischen kennwerten der Deutschen version [Article in German]. Fortschr Neurol Psychiatr 2011 Sep;79(9):517-523. [doi: 10.1055/s-0031-1281602]
44. Mancini AD, Bonanno GA. Predictors and parameters of resilience to loss: toward an individual differences model. J Pers 2009 Dec;77(6):1805-1832. [doi: 10.1111/j.1467-6494.2009.00601.x] [Medline: 19807863]
45. Canli T, Qiu M, Omura K, et al. Neural correlates of epigenesis. Proc Natl Acad Sci U S A 2006 Oct 24;103(43):16033-16038. [doi: 10.1073/pnas.0601674103] [Medline: 17032778]
46. Chmitorz A, Kurth K, Mey LK, et al. Assessment of microstressors in adults: questionnaire development and ecological validation of the Mainz inventory of microstressors. JMIR Ment Health 2020 Feb 24;7(2):e14566. [doi: 10.2196/14566] [Medline: 32130154]
47. Rödel A, Siegrist J, Hessel A, Brähler E. Fragebogen zur messung beruflicher gratifikationskrisen [Article in German]. Z Differentielle Diagnostische Psychol 2004 Jan;25(4):227-238. [doi: 10.1024/0170-1789.25.4.227]
48. Siegrist J, Wege N, Pühlhofer F, Wahrendorf M. A short generic measure of work stress in the era of globalization: effort-reward imbalance. Int Arch Occup Environ Health 2009 Aug;82(8):1005-1013. [doi: 10.1007/s00420-008-0384-3] [Medline: 19018554]
49. Lehr D, Koch S, Hillert A. Where is (im)balance? Necessity and construction of evaluated cut‐off points for effort‐reward imbalance and overcommitment. J Occup O Psychol 2010 Mar;83(1):251-261. [doi: 10.1348/096317909X406772]
50. Bouwmans C, De Jong K, Timman R, et al. Feasibility, reliability and validity of a questionnaire on healthcare consumption and productivity loss in patients with a psychiatric disorder (TiC-P). BMC Health Serv Res 2013 Jun 15;13(1):217. [doi: 10.1186/1472-6963-13-217] [Medline: 23768141]
51. Ahlstrom L, Grimby-Ekman A, Hagberg M, Dellve L. The work ability index and single-item question: associations with sick leave, symptoms, and health – a prospective study of women on long-term sick leave. Scand J Work Environ Health 2010 Sep;36(5):404-412. [doi: 10.5271/sjweh.2917]
52. Boß L, Lehr D, Reis D, et al. Reliability and validity of assessing user satisfaction with web-based health interventions. J Med Internet Res 2016 Aug 31;18(8):e234. [doi: 10.2196/jmir.5952] [Medline: 27582341]
53. Hassenzahl M, Monk A. The inference of perceived usability from beauty. Hum-Comput Interact 2010 Jul;25(3):235-260. [doi: 10.1080/07370024.2010.500139]
54. R Core Team. R: a language and environment for statistical computing. The R Foundation. 2023. URL: https://www.R-project.org/ [accessed 2025-12-04]
55. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw 2011;45:1-67. [doi: 10.18637/jss.v045.i03]
56. Harrer M, Cuijpers P, Schuurmans LKJ, et al. Evaluation of randomized controlled trials: a primer and tutorial for mental health researchers. Trials 2023 Aug 30;24(1):562. [doi: 10.1186/s13063-023-07596-3] [Medline: 37649083]
57. Rubin DB. Multiple Imputation for Nonresponse in Surveys: John Wiley & Sons; 2004. [doi: 10.1002/9780470316696]
58. O'Connell NS, Dai L, Jiang Y, et al. Methods for analysis of pre-post data in clinical research: a comparison of five common methods. J Biom Biostat 2017 Feb 24;8(1):1-8. [doi: 10.4172/2155-6180.1000334] [Medline: 30555734]
59. Cuijpers P. Has the time come to stop using the "standardised mean difference"? Clin Psychol Eur 2021 Sep;3(3):e6835. [doi: 10.32872/cpe.6835] [Medline: 36398102]
60. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 1991;59(1):12-19. [doi: 10.1037/0022-006X.59.1.12]
61. Carrasco-Labra A, Devji T, Qasim A, et al. Minimal important difference estimates for patient-reported outcomes: a systematic survey. J Clin Epidemiol 2021 May;133:61-71. [doi: 10.1016/j.jclinepi.2020.11.024] [Medline: 33321175]
62. Bauer-Staeb C, Kounali DZ, Welton NJ, et al. Effective dose 50 method as the minimal clinically important difference: evidence from depression trials. J Clin Epidemiol 2021 Sep;137:200-208. [doi: 10.1016/j.jclinepi.2021.04.002] [Medline: 33892086]
63. Ebert DD, Franke M, Zarski AC, et al. Effectiveness and moderators of an internet-based mobile-supported stress management intervention as a universal prevention approach: randomized controlled trial. J Med Internet Res 2021 Dec 22;23(12):e22107. [doi: 10.2196/22107] [Medline: 34941541]
64. Jorgensen TD, Pornprasertmanit S, Schoemann AM, et al. SemTools: useful tools for structural equation modeling. The Comprehensive R Archive Network. 2022. URL: https://CRAN.R-project.org/package=semTools [accessed 2025-12-04]

XSL•FO
RenderX

65. Hayes AF, Rockwood NJ. Regression-based statistical mediation and moderation analysis in clinical research: observations, recommendations, and implementation. Behav Res Ther 2017 Nov;98:39-57. [doi: 10.1016/j.brat.2016.11.001] [Medline: 27865431]
66. Roig AE, Mooney O, Salamanca-Sanabria A, Lee CT, Farrell S, Richards D. Assessing the efficacy and acceptability of a web-based intervention for resilience among college students: pilot randomized controlled trial. JMIR Form Res 2020 Nov 11;4(11):e20167. [doi: 10.2196/20167] [Medline: 33174530]
67. Victor P, Krug I, Vehoff C, Lyons N, Willutzki U. Strengths-based CBT: internet-based versus face-to-face therapy in a randomized controlled trial. J Depress Anxiety 2018;07(2):1000301. [doi: 10.4172/2167-1044.1000301]
68. Victor PP, Teismann T, Willutzki U. A pilot evaluation of a strengths-based CBT intervention module with college students. Behav Cogn Psychother 2017 Jul;45(4):427-431. [doi: 10.1017/S1352465816000552] [Medline: 28347377]
69. Ebert DD, Heber E, Berking M, et al. Self-guided internet-based and mobile-based stress management for employees: results of a randomised controlled trial. Occup Environ Med 2016 May;73(5):315-323. [doi: 10.1136/oemed-2015-103269] [Medline: 26884049]
70. Fassnacht DB, Ali K, van Agteren J, et al. A group-facilitated, internet-based intervention to promote mental health and well-being in a vulnerable population of university students: randomized controlled trial of the be well plan program. JMIR Ment Health 2022 May 5;9(5):e37292. [doi: 10.2196/37292] [Medline: 35471196]
71. Lehr D, Freund H, Sieland B, et al. Effectiveness of a guided multicomponent internet and mobile gratitude training program - a pragmatic randomized controlled trial. Internet Interv 2024 Dec;38:100787. [doi: 10.1016/j.invent.2024.100787] [Medline: 39635229]
72. Polivy J, Herman CP. The false-hope syndrome: unfulfilled expectations of self-change. Curr Dir Psychol Sci 2000;9(4):128-131. [doi: 10.1111/1467-8721.00076]
73. Carver CS, Scheier MF. Dispositional optimism. Trends Cogn Sci 2014 Jun;18(6):293-299. [doi: 10.1016/j.tics.2014.02.003] [Medline: 24630971]
74. Malouff JM, Schutte NS. Can psychological interventions increase optimism? A meta-analysis. J Posit Psychol 2017 Nov 2;12(6):594-604. [doi: 10.1080/17439760.2016.1221122]
75. Neff KD. Self-compassion: theory, method, research, and intervention. Annu Rev Psychol 2023 Jan 18;74(1):193-218. [doi: 10.1146/annurev-psych-032420-031047] [Medline: 35961039]
76. Cook JA, Julious SA, Sones W, et al. Practical help for specifying the target difference in sample size calculations for RCTs: the DELTA2 five-stage study, including a workshop. Health Technol Assess 2019 Oct;23(60):1-88. [doi: 10.3310/hta23600] [Medline: 31661431]
77. Alarcon GM, Bowling NA, Khazon S. Great expectations: a meta-analytic examination of optimism and hope. Pers Individ Dif 2013 May;54(7):821-827. [doi: 10.1016/j.paid.2012.12.004]
78. Kaida N, Kaida K. Positive associations of optimism-pessimism orientation with pro-environmental behavior and subjective well-being: a longitudinal study on quality of life and everyday behavior. Qual Life Res 2019 Dec;28(12):3323-3332. [doi: 10.1007/s11136-019-02273-y] [Medline: 31422540]
79. Prati G, Pietrantoni L. Optimism, social support, and coping strategies as factors contributing to posttraumatic growth: a meta-analysis. J Loss Trauma 2009 Aug 26;14(5):364-388. [doi: 10.1080/15325020902724271]
80. Romswinkel EV, König HH, Hajek A. The role of optimism in the relationship between job stress and depressive symptoms. Longitudinal findings from the German Ageing Survey. J Affect Disord 2018 Dec 1;241:249-255. [doi: 10.1016/j.jad.2018.08.005] [Medline: 30138809]
81. Lee EE, Govind T, Ramsey M, et al. Compassion toward others and self-compassion predict mental and physical well-being: a 5-year longitudinal study of 1090 community-dwelling adults across the lifespan. Transl Psychiatry 2021;11(1):1-9. [doi: 10.1038/s41398-021-01491-8]
82. MacBeth A, Gumley A. Exploring compassion: a meta-analysis of the association between self-compassion and psychopathology. Clin Psychol Rev 2012 Aug;32(6):545-552. [doi: 10.1016/j.cpr.2012.06.003]
83. Stutts LA, Leary MR, Zeveney AS, Hufnagle AS. A longitudinal analysis of the relationship between self-compassion and the psychological effects of perceived stress. Self Identity 2018 Nov 2;17(6):609-626. [doi: 10.1080/15298868.2017.1422537]
84. Suh H, Jeong J. Association of self-compassion with suicidal thoughts and behaviors and non-suicidal self injury: a meta-analysis. Front Psychol 2021;12:633482. [doi: 10.3389/fpsyg.2021.633482] [Medline: 34122224]
85. Zessin U, Dickhäuser O, Garbade S. The relationship between self-compassion and well-being: a meta-analysis. Appl Psychol Health Well Being 2015 Nov;7(3):340-364. [doi: 10.1111/aphw.12051] [Medline: 26311196]
86. Bandura A. Self-efficacy: toward a unifying theory of behavioral change. Adv Behav Res Ther 1978 Jan;1(4):139-161. [doi: 10.1016/0146-6402(78)90002-4]
87. Kunzler AM, Helmreich I, Chmitorz A, et al. Psychological interventions to foster resilience in healthcare professionals. Cochrane Database Syst Rev 2020 Jul 5;7(7):CD012527. [doi: 10.1002/14651858.CD012527.pub2] [Medline: 32627860]
88. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on Internet-based mental health interventions — a systematic review. Internet Interv 2014 Oct;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]

XSL•FO
RenderX

89. Haller K, Becker P, Niemeyer H, Boettcher J. Who benefits from guided internet-based interventions? A systematic review of predictors and moderators of treatment outcome. Internet Interv 2023 Sep;33:100635. [doi: 10.1016/j.invent.2023.100635] [Medline: 37449052]
90. Reins JA, Buntrock C, Zimmermann J, et al. Efficacy and moderators of internet-based interventions in adults with subthreshold depression: an individual participant data meta-analysis of randomized controlled trials. Psychother Psychosom 2021;90(2):94-106. [doi: 10.1159/000507819] [Medline: 32544912]
91. Thielecke J, Kuper P, Lehr D, et al. Who benefits from indirect prevention and treatment of depression using an online intervention for insomnia? Results from an individual-participant data meta-analysis. Psychol Med 2024 Jul;54(10):2389-2402. [doi: 10.1017/S0033291724000527] [Medline: 38469832]
92. Kazdin AE. Indirect interventions: lifestyle options to treat mental disorders. Healthcare (Basel) 2025 Feb 26;13(5):505. [doi: 10.3390/healthcare13050505] [Medline: 40077067]
93. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009 Jun 29;338:b2393. [doi: 10.1136/bmj.b2393] [Medline: 19564179]
94. Kalisch R, Köber G, Binder H, et al. The frequent stressor and mental health monitoring-paradigm: a proposal for the operationalization and measurement of resilience and the identification of resilience processes in longitudinal observational studies. Front Psychol 2021;12:710493. [doi: 10.3389/fpsyg.2021.710493] [Medline: 34539510]
95. Petri-Romão P, Engen H, Rupanova A, et al. Self-report assessment of Positive Appraisal Style (PAS): development of a process-focused and a content-focused questionnaire for use in mental health and resilience research. PLoS One 2024;19(2):e0295562. [doi: 10.1371/journal.pone.0295562] [Medline: 38306328]
96. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4(1):1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
97. Trull TJ, Ebner-Priemer U. Ambulatory assessment. Annu Rev Clin Psychol 2013;9(1):151-176. [doi: 10.1146/annurev-clinpsy-050212-185510] [Medline: 23157450]
98. Hofmann SG, Curtiss JE, Hayes SC. Beyond linear mediation: toward a dynamic network approach to study treatment processes. Clin Psychol Rev 2020 Mar;76:101824. [doi: 10.1016/j.cpr.2020.101824] [Medline: 32035297]
99. Boucher EM, Raiker JS. Engagement and retention in digital mental health interventions: a narrative review. BMC Digit Health 2024;2(1):52. [doi: 10.1186/s44247-024-00105-9]
100. Gan DZQ, McGillivray L, Larsen ME, Christensen H, Torok M. Technology-supported strategies for promoting user engagement with digital mental health interventions: a systematic review. Digit Health 2022;8:20552076221098268. [doi: 10.1177/20552076221098268] [Medline: 35677785]
101. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
102. Boß L, Ross J, Reis D, et al. Effectiveness of an integrated platform-based intervention for promoting psychosocial safety climate and mental health in nursing staff: a pragmatic cluster randomised controlled trial. Int J Nurs Stud 2025 Jul;167:105076. [doi: 10.1016/j.ijnurstu.2025.105076] [Medline: 40215602]
103. Alkhaldi G, Hamilton FL, Lau R, Webster R, Michie S, Murray E. The effectiveness of prompts to promote engagement with digital interventions: a systematic review. J Med Internet Res 2016 Jan 8;18(1):e6. [doi: 10.2196/jmir.4790] [Medline: 26747176]
104. Apolinário-Hagen J, Fritsche L, Bierhals C, Salewski C. Improving attitudes toward e-mental health services in the general population via psychoeducational information material: a randomized controlled trial. Internet Interv 2018 Jun;12:141-149. [doi: 10.1016/j.invent.2017.12.002] [Medline: 30135778]
105. Werntz A, Amado S, Jasman M, Ervin A, Rhodes JE. Providing human support for the use of digital mental health interventions: systematic meta-review. J Med Internet Res 2023 Feb 6;25(1):e42864. [doi: 10.2196/42864] [Medline: 36745497]
106. Freund J, Smit F, Lehr D, et al. A universal digital stress management intervention for employees: randomized controlled trial with health-economic evaluation. J Med Internet Res 2024 Oct 22;26:e48481. [doi: 10.2196/48481] [Medline: 39437382]
107. Blut M, Chong AYL, Tsigna Z, et al. Meta-analysis of the Unified Theory of Acceptance and Use of Technology (UTAUT): challenging its validity and charting a research agenda in the red ocean. JAIS 2022;23(1):13-95. [doi: 10.17705/1jais.00719]
108. Thielecke J, Kuper P, Ebert D, et al. Does outcome expectancy predict outcomes in online depression prevention? Secondary analysis of randomised-controlled trials. Health Expect 2024 Feb;27(1):e13951. [doi: 10.1111/hex.13951] [Medline: 39102655]
109. Hannibal S, Behrendt D, Wessa M, Schäfer SK, Dalkner N, Lehr D. RCT Outcome Data RESIST Unguided 2025. [doi: 10.48548/pubdata-2437]

Abbreviations

ANCOVA: analysis of covariance
CBT: cognitive behavioral therapy

XSL•FO
RenderX

CONSORT-eHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
IG: intervention group
NNTB: numbers needed to treat for benefit
NNTH: numbers needed to treat for harm
PASTOR: Positive Appraisal Style Theory of Resilience
PMRe: personal model of resilience
PSS-10: Perceived Stress Scale-10
RCT: randomized controlled trial
TIDieR: Template for Intervention Description and Replication
WL: waitlist control group

Edited by A Mavragani, TDA Cardoso; submitted 31.May.2025; peer-reviewed by RK Kanji, X Liang; revised version received 28.Oct.2025; accepted 29.Oct.2025; published 05.Jan.2026.

Please cite as:
Hannibal S, Behrendt D, Wessa M, Schäfer SK, Dalkner N, Lehr D
Effectiveness of and Mechanisms of Change in a Self-Help Web- and App-Based Resilience Intervention on Perceived Stress in the General Working Population: Randomized Controlled Trial
J Med Internet Res 2026;28:e78335
URL: https://www.jmir.org/2026/1/e78335
doi:10.2196/78335

© Sandy Hannibal, Dörte Behrendt, Michèle Wessa, Sarah K Schäfer, Nina Dalkner, Dirk Lehr. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Predictors of Professional Responses in Nonprofit Mental Health Forums: Interpretable Machine Learning Analysis

Shuang Geng[1], DPhil; Yanghui Li[2], MEng; Jie Wang[2], MEng; Peixuan Chen[3], MSci; Xusheng Wu[4], MEng; Zhiqun Zhang[5], MSci

[1]College of Management, Shenzhen University, Shenzhen, Guangdong, China

[2]College of Artificial Intelligence, Shenzhen University, Shenzhen, Guangdong, China

[3]School of Life Sciences, Central South University, Changsha, Hunan, China

[4]Shenzhen Health Development Research and Data Management Center, Shenzhen, Guangdong, China

[5]Shenzhen Traditional Chinese Medicine Hospital, Shenzhen, Guangdong, China

Corresponding Author:
Xusheng Wu, MEng
Shenzhen Health Development Research and Data Management Center
5 Yuanling Road, Futian
Shenzhen, Guangdong, 518000
China
Phone: 86 19928750177
Email: jysgwuxusheng@163.com

Abstract

Background: Online mental health communities increase access and equity for patients seeking psychological support. User demand and professional contributions are key to their sustainability. While previous research has examined factors influencing physicians' participation in online consultation platforms, limited attention has been given to how post characteristics affect the quantity and length of professional responses in nonprofit mental health communities.

Objective: This study aims to examine how textual (ie, topic, sentiment, title length, and content length) and contextual (ie, page views and posting time) characteristics of inquiries in nonprofit mental health forums influence the quantity and length of responses from mental health professionals, providing insights for enhancing community interactions.

Methods: We collected 18,572 question-and-answer records from a Chinese online mental health platform (August 2024-July 2025). Topic features were extracted using BERTopic, and sentiment features were obtained through a distilled Bidirectional Encoder Representations from Transformers–based sentiment classification model. Additional features were derived from post metadata. We compared 5 machine learning models and identified Light Gradient Boosting Machine as the best performer. We then applied Shapley Additive Explanations (SHAP) analysis to it to evaluate the feature contributions to the prediction of response quantity and length.

Results: In virtual mental health communities, user inquiries fall into 7 topic categories: work, love, depression, boyfriends or girlfriends, school, marriage, and family. Depression-related topics negatively predict response quantity, whereas interpersonal, school, marriage, or family topics are positively correlated. SHAP analysis revealed that page views (SHAP value=0.187) and title length (SHAP value=0.073) are key factors in predicting response quantity, and content length (SHAP value=0.274), sentiment category (SHAP value=0.054), and title length (SHAP value=0.053) are key factors in predicting response length. Posts exhibiting negative emotions are positively related to both the predicted quantity and length of responses, and this effect becomes more pronounced as the degree of emotional intensity increases. Titles with 15-20 characters and content with more than 60 characters are positively correlated with responses, whereas titles with fewer than 7 characters have negative effects. Higher view counts and weekday posts also increase response likelihood.

Conclusions: This study provides insights into how textual and contextual features of patient posts influence professional responses in nonprofit mental health forums. It enhances understanding of voluntary knowledge contribution behaviors in online mental health communities and offers practical guidance for optimizing platform functional design and user posting strategies. Future researchers are encouraged to address the limitations of this study, which focuses solely on response quantity and length, and to explore details of professional responses, such as by developing a comprehensive measure of response quality.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e74359) doi:10.2196/74359

KEYWORDS

influencing factors; online mental health community; predictive analysis; response length; response quantity; sentiment analysis; theme analysis

Introduction

Background

Mental health disorders are critical global health concerns that pose major challenges to both individual well-being and health care systems [1-4]. According to the World Health Organization, approximately 1 billion individuals (more than 12.5% of adults and adolescents) are affected by mental health disorders, which account for approximately 5% of disability-adjusted life years [5]. Individuals living with mental health conditions experience a high burden of illness and face an increased risk of mortality [6]. Therefore, strengthening mental health services has emerged as a notable research concern [7].

In recent years, an increasing number of online mental health communities (OMHCs) have been established, allowing mental health professionals from offline hospitals to offer services on these platforms. These online mental health platforms connect mental health service providers with patient users seeking psychological support and are gaining popularity [8,9]. The benefits of OMHCs are manyfold. For example, OMHCs effectively mitigate social prejudice and stigma associated with mental health disorders. Many individuals delay or avoid seeking professional help because of concerns about being labeled. The anonymity setting of these platforms alleviates such concerns and encourages people with mental health challenges to actively seek psychological support [10,11]. Moreover, OMHCs help address the time and geographical limitations of offline clinical services. This also addresses the uneven distribution of mental health resources and enhances the accessibility and equity of mental health services [12]. Given its importance, this study focuses on OMHCs to better understand the contributing behavior of health professionals on these platforms.

Prior studies on online mental health can be classified into 3 main categories: studies that focus on patient users (eg, user-generated content, user engagement patterns) [13-17], studies that focus on health professionals (eg, contribution behaviors) [18,19], and studies that focus on communities (eg, community quality and value cocreation) [20-22]. Existing studies on the contribution behavior of health professionals suggest that their active participation within the community determines the effectiveness of counseling services [18]. For OMHC managers, understanding the factors that influence the contribution behaviors of health professionals is crucial for improving both the efficiency and quality of community responses, promoting sustained user participation, and enhancing user retention and engagement [23]. From the perspective of patient users, when they seek medical advice and feelings of expression in OMHCs, their posts signal their need for informational or emotional support from professionals. Responses from professional community members may make them feel valued and accepted and enhance their sense of belonging to the community. Moreover, the information shared by professionals benefits not only help-seekers but also lurkers and other community members [24]. Together, the engagement of both patients and professionals helps cultivate a sustainable ecosystem for health knowledge exchange [25]. Therefore, it is imperative to focus on mental health professionals and examine the factors influencing their knowledge contributions in OMHCs.

Previous studies have provided insights into the factors influencing health professionals' knowledge contribution behaviors [20,26-33]. These factors include content-related factors, such as question type, information quality, and readability [26-28]; community-related factors, such as reward mechanisms [26]; and health professional–related factors, such as social relationships, self-efficacy, and reputation needs [30-33]. For example, Srivastava et al [27] analyzed the intent, criticism, readability, and emotion of user posts on Reddit, and found the prominence of "self-criticism" as the most prevalent form of criticism expressed by help-seekers. They also found that individuals who explicitly express their need for help are more likely to receive assistance. These studies explore these factors primarily through the lens of signal theory, motivational theory, self-determination theory, and social exchange theory. For instance, Chen et al [28] reported that emotional and informational language signals increase the likelihood of professionals providing informational and emotional support. Drawing on social exchange theory, Wang et al [29] reported that material and psychological rewards significantly increase the online contributions of health professionals. Additionally, several studies have explored the impact of the intrinsic and extrinsic motivations of professionals' contribution behaviors. Imlawi and Gregg [32] noted that factors such as helping motivator, reputation motivator, and moral obligation motivator influence professionals' contribution continuance intentions. Maheshwari et al [33] found that self-efficacy and reciprocity positively influence the attitude toward knowledge sharing; however, the rewards' moderating effect is not significant.

While there is a wide range of factors that influence professionals' knowledge contribution, post characteristics serve as the most direct medium for patient users to share stories and engage with the community, and they are closely associated with social interaction outcomes [27]. The informational cues (eg, topic) and emotional cues (eg, sentiment) in a post are crucial signals for consultants to understand the patient users' problem and to decide whether to respond. For example, posts with clear problem descriptions may reduce the cognitive effort required by consultants, which is especially important when the professionals are volunteers with limited time. Research in computer-mediated communication suggests that emotional cues are important drivers of social support [34]. Beyond content, posting time is a critical factor that determines a post's visibility. For example, a post published late at night may be quickly buried under a flood of newer posts, making it receive

XSL•FO
RenderX

fewer replies. Investigating the influences of these post features can provide actionable insights beyond those gained from focusing on user or professional factors. Demographic or professional attributes of users and professionals are often static and difficult to change within a platform's context. In contrast, content features are dynamic. Community managers can develop posting guidelines and platform functionalities to help patients craft more effective queries. Therefore, this study aims to address this research gap by investigating diverse post features and providing practical implications that empower users to effectively seek and obtain professional responses. This study conducts a microlevel analysis of post features, which complements the user-level and community-level research in the OMHC landscape. Our work offers granular and actionable explanations for knowledge contribution in OMHCs.

Previous online health community research has usually adopted structural equation modeling, multiple linear regression, and fixed effect modeling to analyze the relationships between focal factors [29,35,36]. The rise of artificial intelligence and machine learning has provided powerful tools for modeling, partitioning, and interpreting the complex relationships between factors. A few recent mental health studies have used machine learning algorithms, such as logistic regression, decision trees, and ensemble models [37-39]. Light Gradient Boosting Machine (LightGBM), an ensemble learning algorithm based on gradient boosting decision trees, is widely used for classification, regression, and ranking tasks because of its efficiency and outstanding performance, particularly in structured and tabular data problems [38]. However, LightGBM, like other ensemble models (eg, Extreme Gradient Boosting [XGBoost]), is often regarded as a "black-box" model and lacks sufficient transparency and interpretability. Explainable machine learning techniques such as Shapley Additive Explanations (SHAP) [39] can visualize the importance of factors in driving model decisions, enabling stakeholders to understand the logic behind the model's outputs and make informed decisions. Therefore, this study combines these 2 techniques to detect important post features that affect the quantity and length of professional responses in OMHCs, aiming to improve community interaction quality and enhance the effectiveness of mental health services.

Objective

The research question for this study is as follows: "In nonprofit OMHCs, how do patient posts' textual features (eg, topic, sentiment, title length, and content length) and contextual features (eg, page view and posting time) influence the response quantity and length of health professionals?" To answer this question, this paper constructs an interpretable machine learning model for analysis. This study aims to deepen the understanding of knowledge contribution in OMHCs and to inform management strategies for building more supportive and efficient mental health communities.

Methods

Overview

The research design for this study is shown in Figure 1. This framework consists of 4 main phases: data collection, thematic and sentiment analysis, predictive model construction, and predictive model interpretation. In the data collection process, we collected a total of 18,572 post-level records using a web crawler. Subsequently, we performed preliminary data cleaning to remove posts with empty titles or empty content. We then analyzed the post topics using BERTopic and removed noisy clusters. A total of 11,154 data entries were obtained after noise (topic cluster index=–1) removal. Sentiment analysis was conducted using distilled Bidirectional Encoder Representations from Transformers (DistilBERT) to determine both the category and intensity of sentiment for each post. During the modeling phase, we trained and compared 5 models (ie, LightGBM, Support Vector Machine [SVM], XGBoost, random forest [RF], ridge regression) to select the best one to develop the final model. These models were trained using textual and contextual features of the posts as predictive variables. The models were designed to predict 2 key outcome variables: response quantity and response length. Finally, we used the SHAP method to interpret the model. It provided global interpretations via summary plots to determine overall feature importance and local interpretations using dependence plots to elucidate how individual features influence the model's predictions.

XSL•FO
RenderX

Figure 1. Study flowchart. Panels 1-2 depict the data collection and processing workflow, and panels 3-4 illustrate the development of the interpretable machine learning model. LightGBM: Light Gradient Boosting Machine; MAE: mean absolute error; MSE: mean squared error; Q&A: question-and-answer; RF: random forest; SHAP: Shapley Additive Explanations; SVM: Support Vector Machine; XGBoost: Extreme Gradient Boosting.

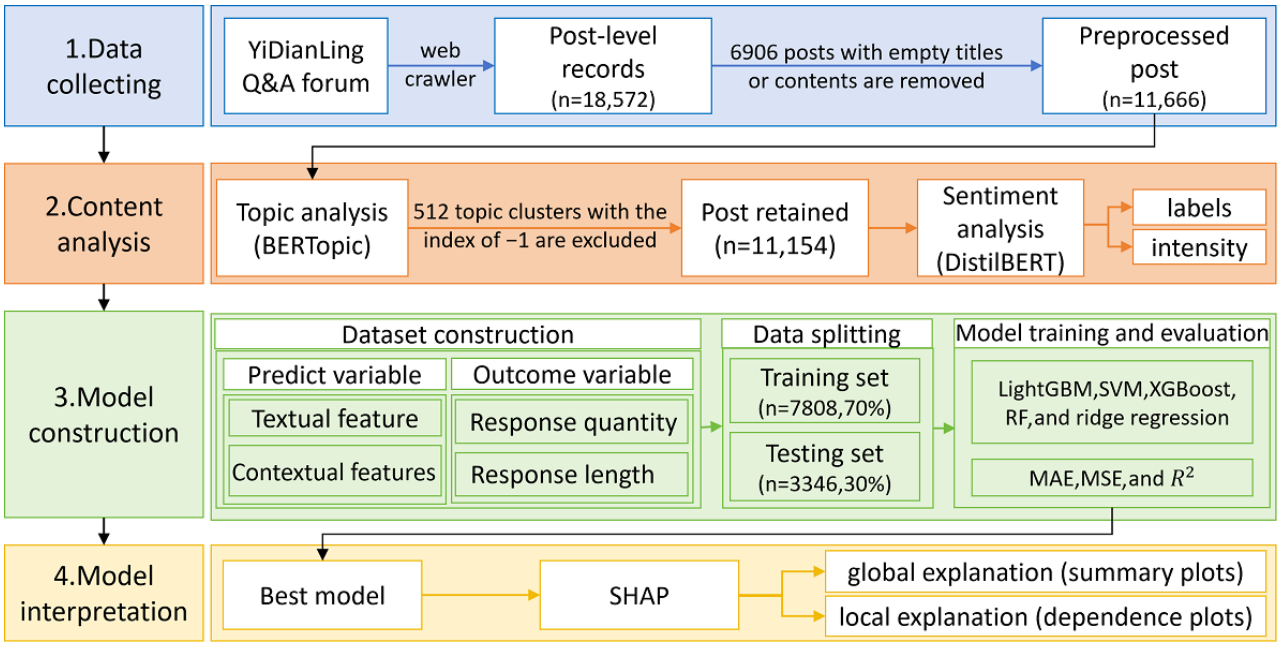

Data Collection

We collected data from the YiDianLing platform [40]. YiDianLing is a leading nonprofit mental health service provider in China, which hosts approximately 50 million registered users and 60,000 professional psychological consultants. The large and diverse user base ensures the representativeness of the data. Moreover, the platform's nationwide coverage provides broad geographic representation. YiDianLing provides a dedicated public question-and-answer forum to facilitate interaction between patient users and certified psychological consultants. In this forum, users can anonymously post their mental health concerns, and psychological consultants can provide free responses. Therefore, the forum generates rich data, including user posts, consultant responses, post view counts, and response volume.

Data from the YiDianLing, which comprised 18,572 entries from August 2024 to July 2025, were collected using a crawler program. An example of the data source page is depicted in Figure 2. Each data sample includes information about the user's post, such as the post title, post content, date of post, number of page views, number of responses, and number of responses provided by psychological consultants. The temporal evolution of the number of posts and the number of responses on the platform is presented in Multimedia Appendix 1. The volume of posts and responses exhibits a high degree of stability, with the number of replies consistently exceeding the number of posts.

XSL•FO
RenderX

Figure 2. Example of extracted features from the post.



Content Analysis

Topic Analysis

BERTopic is a topic modeling technique based on Bidirectional Encoder Representations from Transformers (BERT), which is a pretrained language model based on the transformer architecture that reads text in both directions (left-to-right and right-to-left) to understand the full context [41]. This architecture is based on multilayer neural networks called encoders, and it uses a self-attention mechanism to capture word relationships and context. BERTopic can effectively add interoperability challenges between density-focused clustering and centroid-oriented methods. This study uses the BERTopic model for thematic analysis, as it has demonstrated advantages in various topic modeling benchmark tests [42].

BERTopic facilitates consistent topic identification by leveraging a category-specific version of the term frequency-inverse document frequency (TF-IDF). In this technique, all text in a cluster is considered one entity, and TF-IDF is applied to determine the relevance scores for words within that cluster. By extracting important words in each cluster, descriptions of topics are obtained. This method is known as class-based TF-IDF:



where $f_{x,c}$ represents the frequency of word x in cluster c, f_x denotes the frequency of word x across all clusters, and A signifies the average number of words contained in each cluster.

Sentiment Analysis

In this study, we used the DistilBERT method to classify post sentiment into 5 categories: very negative, negative, neutral, positive, and very positive [43]. This model is a lightweight pretrained language model built on BERT through knowledge distillation techniques [44]. DistilBERT retains BERT's performance in capturing the sentence context and reduces the number of parameters to achieve higher computational efficiency. The used DistilBERT model has been fine-tuned on a multilingual corpus, including Chinese, and has been successfully applied to a wide range of tasks, such as product review classification, social media sentiment analysis, and customer feedback analysis [45,46].

In the sentiment analysis, we first segmented each user post into a set of sentences and applied DistillBERT analysis separately. The predicted sentiment labels were then mapped to numerical scores, with "very negative" as –1, "negative" as –0.5, "neutral" as 0, "positive" as 0.5, and "very positive" as 1. The sentiment scores of the set of sentences were then summed to obtain an overall sentiment score for the post. On the basis of this overall sentiment score, we classified each post into one of three sentiment polarities: posts with sentiment scores less than 0 were classified as negative, posts with a score equal to 0 were classified as neutral, and posts with a score greater than 0 were classified as positive. The sentiment intensity of each post was measured as the logarithm of the absolute value of the overall sentiment score plus 1.

XSL•FO
RenderX

Model Construction

LightGBM is an efficient ensemble method built on gradient boosting decision trees, which construct multiple classifiers and integrate their outputs to obtain the final prediction [47]. It is widely applied to classification, regression, and ranking tasks and can model structured and tabular data [38,48]. Unlike RFs and XGBoost, LightGBM adopts a histogram-based splitting technique that splits data and scans the statistics to determine the best split point, which enables less memory consumption and more efficient training [48].

We compared 5 popular machine learning models: LightGBM, SVM, XGBoost, RF, and ridge regression. We used a 70/30 train-test split, and the data are randomly divided 5 times to reduce the randomness introduced by data splits. The average performance of the 5 trained models is used for final model evaluation using mean absolute error, mean squared error, and R-squared. Based on these evaluations (Multimedia Appendix 2), LightGBM showed the best performance and was selected for subsequent regression prediction and interpretation.

During model training, hyperparameters were tuned on the basis of the official LightGBM documentation [49]. Specifically, the maximum number of leaves per weak learner was set to 40 to mitigate overfitting; the learning rate was set to 0.05 to accelerate convergence and improve prediction accuracy; and feature_fraction was set to 0.8, enabling the model to randomly select a subset of features when constructing each tree, thereby reducing training time. The detailed hyperparameter settings are provided in Multimedia Appendix 3.

Model Interpretation

The black-box nature of traditional machine learning models, such as ensemble methods and neural networks, limits their clinical application in the mental health domain, as stakeholders require transparent and trustworthy decision-making processes [50]. SHAP is a model interpretation method based on cooperative game theory [51]; it provides a unified measure of feature importance by attributing the model's prediction to the marginal contributions of each feature, known as SHAP values. It works as follows.

Consider the ith sample as x_i, where x_{ij} represents the jth feature of the ith sample, and y_i denotes the model's forecast for this sample. The baseline model prediction (often the predicted mean of all samples) is denoted y_{base}. The SHAP value is then derived based on the following formula:



where $f(x_{ij})$ denotes the SHAP value for x_{ij}, indicating the influence of the jth feature of the ith sample on the ultimate prediction y_i. A positive value suggests that the feature enhances the prediction, whereas a negative $f(x_{ij})$ indicates a diminishing effect on the predicted outcome.

We conducted SHAP analysis and visualized the contribution of each feature to the model's prediction using importance ranking summary plots. The dependence plots show the relationship between the changes in a feature's value and its impact on the prediction. The quantified contributions of features, either positive or negative, enhance the transparency of the prediction model.

Ethical Considerations

This study did not involve human participants. The data were publicly accessible information on the YiDianLing platform. All user posts in the question-and-answer forum were published anonymously by patient users, and no identifiable or reidentifiable personal information was collected or processed throughout the research. Therefore, there is no risk to individual privacy or foreseeable harm to users. We conducted our data collection in accordance with the platform's data authorization agreement and ensured that all procedures fully complied with the relevant ethical standards. During the handling of the dataset, we also took steps to maintain data security. This research project received formal approval from the Institutional Review Board of Shenzhen University (approval number PN-202500199).

Results

Topic Analysis and Sentiment Analysis

The results of the topic modeling visualization revealed 8 themes of user posts, as shown in Figure 3. The gray areas in Figure 3 are clusters with a category index of –1 and are considered noise and were excluded from the analysis. Representative keywords for each topic are provided in Multimedia Appendix 4.

XSL•FO
RenderX

Figure 3. Topic clustering results.

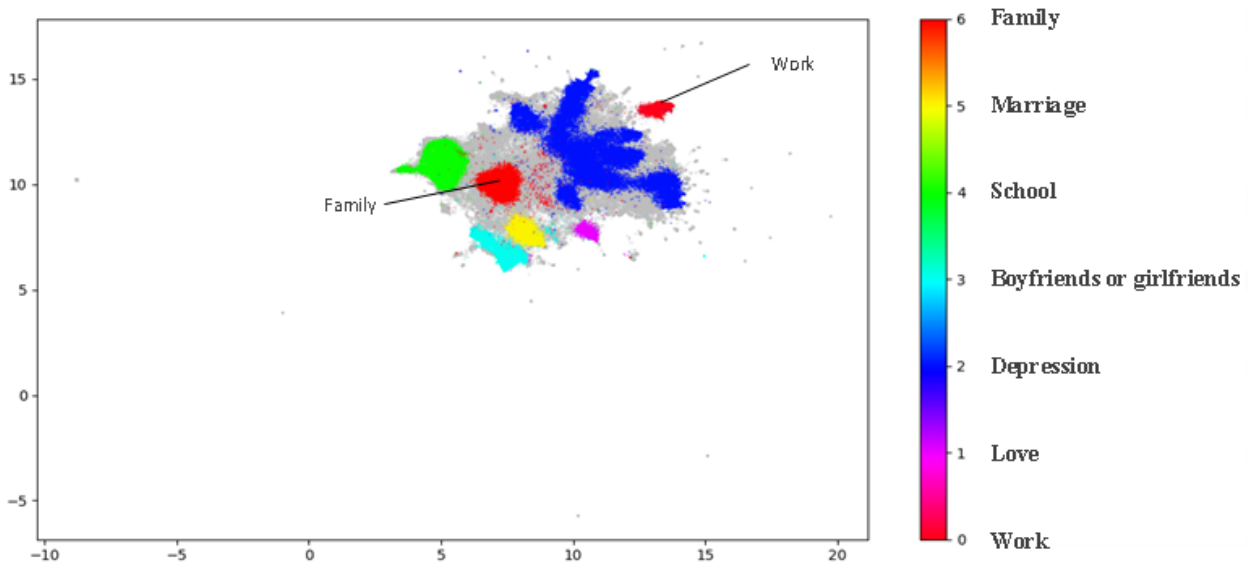

On the basis of the results of thematic clustering, we visualize the frequency of different themes over the observation period in Figure 4. The results indicate that depression consistently remained the predominant theme. Its proportion significantly exceeds that of the other categories, followed by family, school, and work. The overall topic frequency demonstrated stability throughout the period.

Figure 4. Topic frequency evolution.

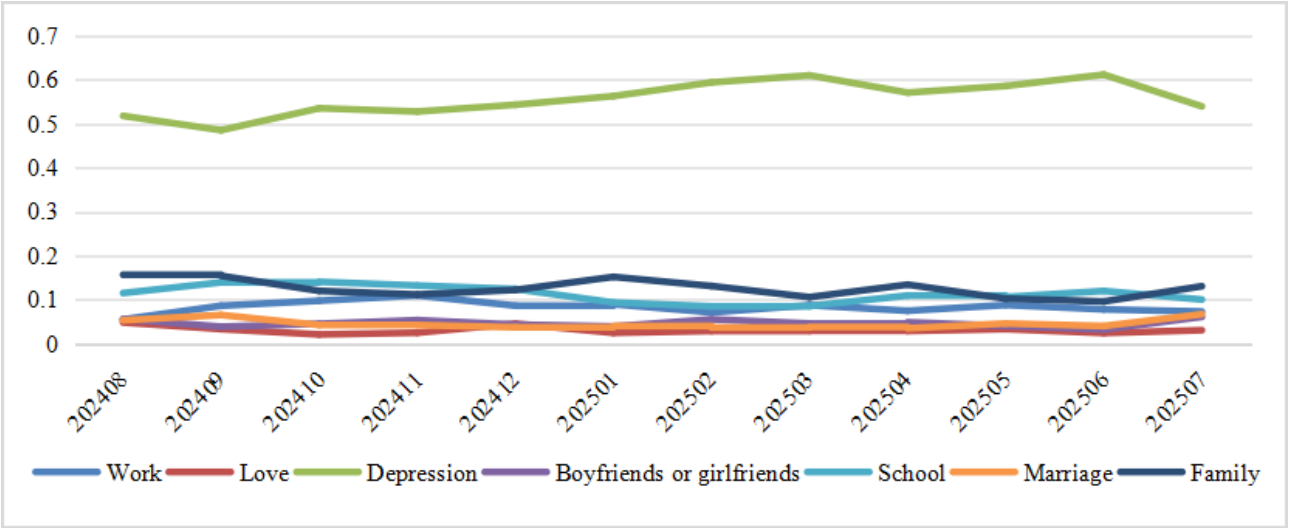

The proportions of sentiment categories at different times are depicted in Figure 5, which shows that the proportion of posts with negative sentiment is greater than those with positive and neutral sentiment. The distribution of sentiment categories remains relatively stable over time, indicating that patient users primarily express negative emotions in the mental health community.

XSL•FO
RenderX

Figure 5. Emotional map for different dates.

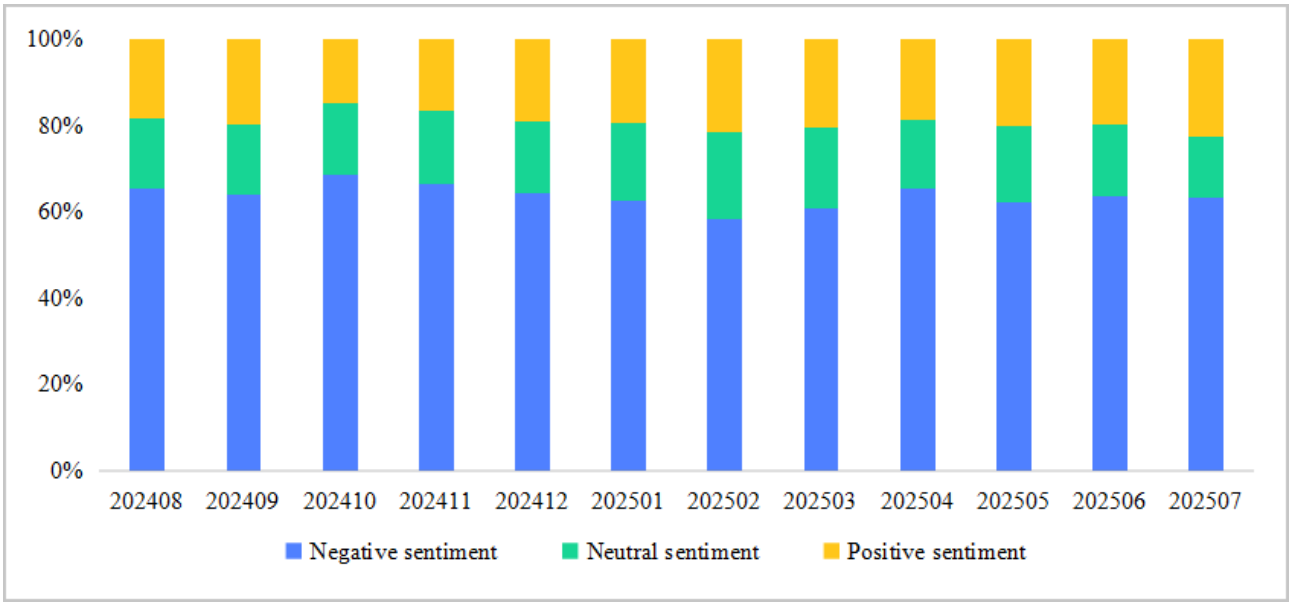

Descriptive Statistics

A total of 11,154 data entries were obtained after topic clustering and noise removal and were used for analysis. The predictor variables included post theme, sentiment category, sentiment intensity, title length, content length, posting year, posting month, posting date, public holiday status, time of day, day of the week, and page view count. The outcome variables comprise the quantity and length of replies, and prediction models are constructed separately for each. We acknowledge that a full picture of community interaction also depends on the quality of responses, a multifaceted construct that encompasses aspects like relevance, empathy, and supportiveness. Capturing this richness quantitatively poses a distinct methodological challenge. We therefore view our work as a critical first step that sets the stage for, and thereby invites, subsequent research to delve into the nuanced quality of professional contributions. A description of the features is provided in Table 1. Descriptive statistics of these selected variables are presented in Table 2, excluding the temporal characteristics of the posts.

XSL•FO
RenderX

Table 1. Description of features.

Feature	Feature description	Variable type
Topic	The topic of the content; 0=work; 1=love; 2=depression; 3=boyfriends or girlfriends; 4=school; 5=marriage; 6=family	Categorical variable
Sentiment category	0=negative; 1=neutral; 2=positive	Categorical variable
Sentiment intensity	The logarithm of the absolute value of the sentiment score plus 1	Continuous variable
Page view	The logarithm of the number of page views posted on a given date	Continuous variable
Year	Year of the post	Categorical variable
Month	Month of the post	Categorical variable
Day	Day of the post	Categorical variable
Title length	The natural logarithm (base e) of the number of Chinese characters in the post title (note: raw character counts are used for result interpretation)	Continuous variable
Content length	The natural logarithm (base e) of the number of Chinese characters in the post content (note: raw character counts are used for result interpretation)	Continuous variable
Hour	0=00:00~00:59, 1=01:00~01:59, 2=02:00~02:59, 3=03:00~03:59, 4=04:00~04:59, 5=05:00~05:59, 6=06:00~06:59, 7=07:00~07:59, 8=08:00~08:59, 9=09:00~09:59, 10=10:00~10:59, 11=11:00~11:59, 12=12:00~12:59, 13=13:00~13:59, 14=14:00~14:59, 15=15:00~15:59, 16=16:00~16:59, 17=17:00~17:59, 18=18:00~18:59, 19=19:00~19:59, 20=20:00~20:59, 21=21:00~21:59, 22=22:00~22:59, 23=23:00~23:59	Categorical variable
Week	0=Monday; 1=Tuesday; 2=Wednesday; 3=Thursday, 4=Friday; 5=Saturday; 6=Sunday	Categorical variable
Holiday	0=no; 1=yes	Categorical variable
Reply quantity	The logarithm of the number of replies to the post	Continuous variable
Reply length	The logarithm of the average reply length of the post	Continuous variable

Table 2. Descriptive statistics.

Features	Value (n=11,154)
Page view, mean (SD)	5.09 (0.67)
Sentiment category, n (%)	
Positive	2117 (19)
Neutral	1904 (17.1)
Negative	7133 (63.9)
Sentiment intensity, mean (SD)	0.69 (0.48)
Topic, n (%)	
Work	917 (8.2)
Love	355 (3.2)
Depression	6139 (55)
Boyfriends or girlfriends	519 (4.7)
School	1269 (11.4)
Marriage	501 (4.5)
Family	1454 (13)
Title length, mean (SD)	2.92 (0.40)
Content length, mean (SD)	4.29 (1.30)
Reply quantity, mean (SD)	0.22 (0.66)
Reply length, mean (SD)	5.25 (0.83)

XSL•FO
RenderX

Comparisons of Model Performance

We evaluated 5 machine learning models: LightGBM, SVM, XGBoost, RF, and ridge regression for predicting response quantity and response length. Tables 3 and 4 present the performance of these models. Results show that LightGBM achieved the lowest mean absolute error (mean 0.2859, SD 0.0072) and mean squared error (mean 0.3100, SD 0.0142), along with the highest R^2 (mean 0.2754, SD 0.0323), statistically outperforming SVM, XGBoost, RF, and ridge regression. Similarly, for response length prediction (Table 4), LightGBM demonstrated superior overall performance. These metrics provide a comprehensive assessment of each model's capabilities. Additionally, calibration curves for the LightGBM model in both prediction tasks are provided in Multimedia Appendix 2, both of which indicate a good model fit.

Table 3. Response quantity prediction performance of the compared models.

Model	MAE[a], mean (SD)	MSE[b], mean (SD)	R^2, mean (SD)
LightGBM[c]	0.2859 (0.0072)	0.3100 (0.0142)	0.2754 (0.0323)
SVM[d]	0.3101 (0.0161)	0.3931 (0.0391)	0.0855 (0.0187)
XGBoost[e]	0.3155 (0.0061)	0.3223 (0.0190)	0.2476 (0.0182)
RF[f]	0.2962 (0.0065)	0.3116 (0.0146)	0.2717 (0.0302)
Ridge regression	0.3693 (0.0049)	0.3516 (0.0195)	0.1789 (0.0221)

Table 4. Response length prediction performance of the compared models.

Model	MAE[a], mean (SD)	MSE[b], mean (SD)	R^2, mean (SD)
LightGBM[c]	0.6988 (0.0079)	0.8367 (0.0155)	0.2766 (0.0221)
SVM[d]	0.6941 (0.0054)	0.8905 (0.0149)	0.2302 (0.0145)
XGBoost[e]	0.7219 (0.0072)	0.8688 (0.0168)	0.2490 (0.0129)
RF[f]	0.7058 (0.0069)	0.8637 (0.0146)	0.2532 (0.0218)
Ridge regression	0.7171 (0.0070)	0.8804 (0.0161)	0.2390 (0.0154)

[a]MAE: mean absolute error.

[b]MSE: mean squared error.

[c]LightGBM: Light Gradient Boosting Machine.

[d]SVM: Support Vector Machine.

[e]XGBoost: Extreme Gradient Boosting.

[f]RF: random forest.

Model Interpretability

Global Interpretability

We applied the SHAP method to the LightGBM model. The global interpretation graphs of LightGBM for predicting the number of replies and the length of replies are shown in Figure 6. The average SHAP value for each feature is detailed in Multimedia Appendix 5. When a SHAP value of 0 is used as the dividing line, the points on the left indicate the features contributing negatively to the prediction, whereas the points on the right indicate positive contributions. The relationship between each feature and the prediction of the number of replies is shown in Figure 6A. This indicates that positively correlated features include page views, content length, and sentiment intensity. Higher values of these features correspond to a greater number of responses received by the posts. The relationship between each feature and the prediction of response length is shown in Figure 6B. These findings indicate that the length of the question content has a positive effect on response length. Most other features are categorical variables, whose effects are not clearly discernible from the figure.

XSL•FO
RenderX

Figure 6. Summary plots of Light Gradient Boosting Machine. (A) Prediction for response quantity. (B) Prediction for response length. SHAP: Shapley Additive Explanations.

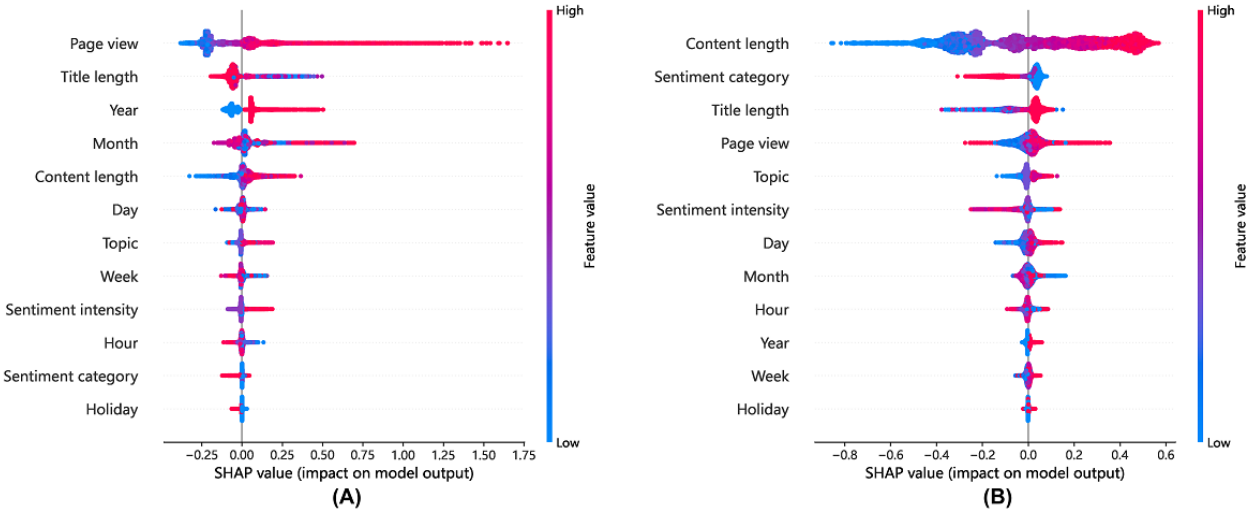

Local Interpretability in Response Quantity Prediction

We constructed feature-SHAP value scatter plots for each feature to analyze its impact on the response quantity (Figure 7). Each dot in the scatter plot represents a single post in our dataset. They illustrate the relationship between feature values (x-axis) and their corresponding SHAP values (y-axis). The SHAP value is a direct measure of how much that specific feature value pushed the model's prediction toward receiving more (positive SHAP value) or fewer (negative SHAP value) replies. These plots answer a critical question: how a specific post characteristic influences a consultant's likelihood to reply, and whether this influence is consistently positive, negative, or more complex?

XSL•FO
RenderX

Figure 7. Shapley Additive Explanations (SHAP) dependence plots of Light Gradient Boosting Machine for predicting response quantity. Panels A-L respectively show the SHAP value distributions for the following features: topic, sentiment category, sentiment intensity, page views, title length, content length, year, month, day, hour, week, and holiday.

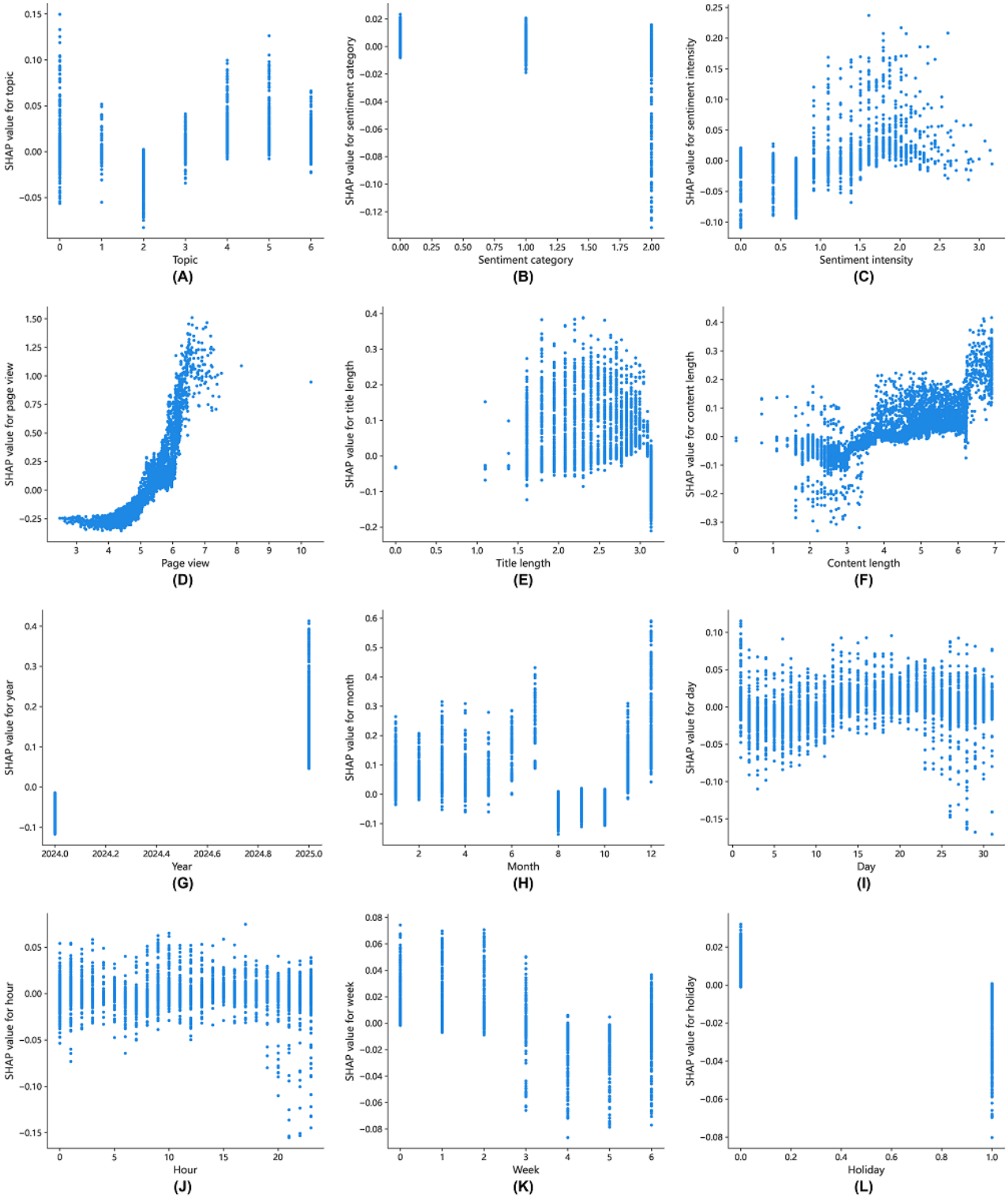

The impact of topic features on response quantity is illustrated in Figure 7A. Topic 2 (depression) has a SHAP value less than 0, whereas topics 3-6 (3: boyfriends or girlfriends; 4: school; 5: marriage; and 6: family) have SHAP values greater than 0. These findings suggest that the topic of depression has a negative effect on response quantity, whereas those related to boyfriends or girlfriends, school, marriage, and family have positive effects. The impact of other topics on response quantity is not clearly defined.

The relationships between post sentiment and response quantity are shown in Figures 7B and 7C. The SHAP values for negative sentiment are greater than 0, indicating a positive contribution to response quantity. In contrast, the SHAP values for positive sentiment are less than 0, indicating a negative contribution. Moreover, when the sentiment intensity exceeds the threshold

XSL•FO
RenderX

of 1.7 (equivalent to a sentiment intensity of 5.5), the post sentiment contributes positively to the response quantity. These findings suggest that mental health professionals are more likely to respond to posts that express negative emotions and provide support to high-risk patient users.

The effects of other features on the prediction of response quantity are presented in Figures 7D-7L. The results indicate that when page views exceed the threshold of 5.5 (approximately 244 views), the title length is between 2.75 and 2.95 (approximately 15 to 20 characters), and the content length exceeds the threshold of 4.85 (approximately 127 characters), the posts are more likely to receive professional responses. With respect to the posting time, when the post is published between December and February, in the middle of the month (days 18 to 21), during the period between Monday and Thursday, and on nonpublic holidays, it is more likely to receive a professional response. Moreover, when the page view is less than 5.5 (approximately 244 views), the title length exceeds 2.95 (approximately 20 characters), or when the post is published from August to October, on weekends, or on public holidays, the SHAP values are less than 0, indicating an inhibitory effect on the prediction of response quantity. Additionally, the posting time (hour of day) does not have a significant effect.

Local Interpretability in Response Length Prediction

We report SHAP dependence plots for each feature to explain the response length predicted by the LightGBM model (Figure 8). According to Figure 8A, the topic of love has a negative effect on the prediction of response length. The results from Figures 8B and 8C show that negative sentiment has a positive effect on predicting response length, whereas positive emotions have an inhibitory effect. When the sentiment intensity is less than 1.5 (equivalent to the original sentiment intensity of 4.5), it negatively affects the prediction of the reply length. As shown in Figure 8E, when the length of the post's title is less than the threshold of 2.0 (approximately 7 characters), the predicted response length decreases. As shown in Figure 8F, the critical value for content length is 4.1 (approximately 60 characters), and posts with lengths exceeding 60 characters are more likely to receive longer responses.

XSL•FO
RenderX

Figure 8. Shapley Additive Explanations (SHAP) dependence plots of Light Gradient Boosting Machine for predicting response length. Panels A-L respectively show the SHAP value distributions for features: topic, sentiment category, sentiment intensity, page views, title length, content length, year, month, day, hour, week, and holiday.

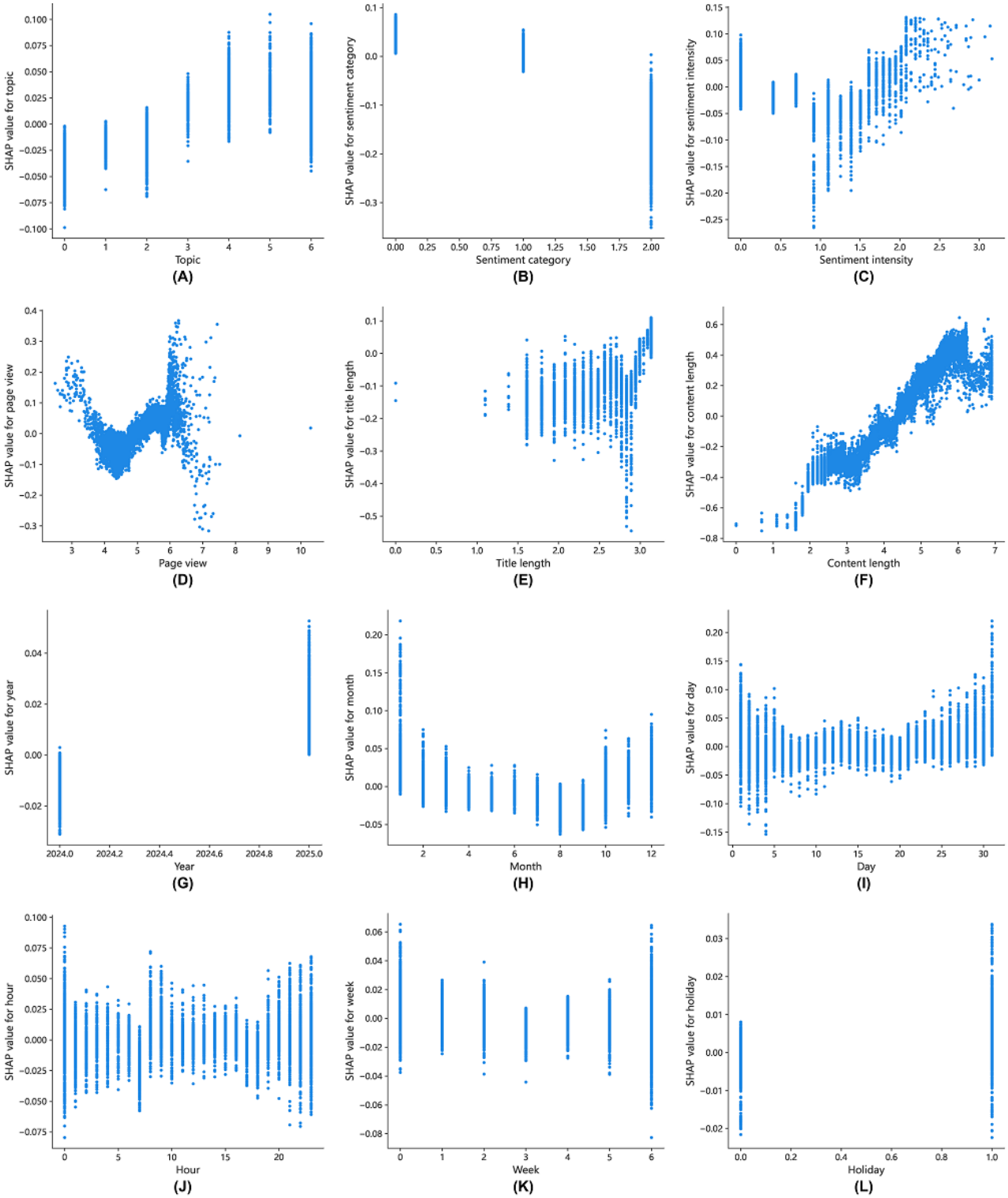

Robustness Analysis

In the robustness analysis, we collected and analyzed forum data from another online mental health platform, YiXinli [52], covering the period from December 2024 to July 2025, with the aim of examining the generalizability of the model interpretations through cross-platform comparison. All the analytical procedures were kept consistent with those applied to the primary dataset. Multimedia Appendix 6 reports the descriptive statistics, topic and sentiment evolution results, model evaluation results, and SHAP-based interpretation outcomes. The findings indicate that the topic distributions, sentiment distributions, and their temporal trends are consistent with those observed on the YiDianLing platform. Moreover, the SHAP interpretations of the LightGBM models for predicting the quantity and length of responses indicate that the

XSL•FO
RenderX

effects of key features are relatively consistent with those observed in the primary dataset. These results provide robust support for our findings.

Discussion

Principal Findings

This study uses interpretable machine learning techniques to analyze question-and-answer posts from a nonprofit OMHC. The findings highlight the pivotal role of various features of the posts in shaping professionals' contribution behaviors. First, patient users' demands for psychological services primarily fall into 7 topic categories: work, love, depression, boyfriends or girlfriends, school, marriage, and family. The majority of user posts are related to depression, which aligns with previous research [14,53]. Users also frequently express concerns related to daily life, work stress, and social relationships [54]. Posts with greater response volumes are often associated with themes such as boyfriends or girlfriends, school, marriage, and family, whereas depression-related posts receive fewer replies. One possible explanation is that depression-related posts are more likely to be posted by diagnosed patients. The forum's professionals are composed mainly of psychological consultants who are not therapists capable of providing clinical treatment. Therefore, they tend to be more cautious when addressing depression-related issues, thus avoiding the risk of inadvertently harming these patient users. This aligns with prior findings that, in the absence of sufficient knowledge about the individual, consultants may avoid overreacting to questions involving illness and emotions [55].

Second, our study revealed that patient posts with stronger negative emotions are more likely to receive social support in nonprofit mental health communities. From the perspective of social support theory, emotional disclosure by patient users in online communities is crucial for fostering social interaction and obtaining support. Expressions of negative emotions may signal users' distress, which may elicit empathy and supportive reactions from psychological counselors [56]. On the other hand, emotional intensity can be viewed as an indicator of patients' level of self-disclosure [57,58]. Intense self-disclosure may enhance users' motivation to engage [59]. Our analysis shows, for example, that when sentiment intensity is relatively high (eg, above 5.5), posts are more likely to attract longer replies. In contrast, when sentiment intensity is relatively low (eg, less than 4.5), posts tend to receive shorter responses. Drawing on these findings, community managers could provide patient users with emotion-related keywords or tags, which can be automatically selected when writing posts. These emotion-related keywords or tags may enhance users' ability to articulate their emotional states.

Third, our findings indicate that the title length and content length of patient users' posts positively influence the quantity and length of professionals' responses. Longer titles and content may contain greater amounts of information, facilitating better understanding by other community members. Therefore, the amount of information provided in posts positively influences the quality of replies [24]. Our analysis revealed that titles with lengths between 15 and 20 characters and contents with lengths of at least 60 characters attracted more and longer responses. In contrast, titles shorter than approximately 7 characters tend to negatively impact the response length. On the basis of these findings, platform designers may provide real-time feedback on the informativeness of titles and content to post writers. Additionally, the platform can offer high-quality example posts for users to enhance the expressiveness of their posts, thus increasing their likelihood of receiving a response.

In our study, post features, including view count, posting time, day of the week, and public holiday status, influence professional responses heterogeneously. Posts with more than 244 views are associated with greater response volumes. This suggests that posts with greater exposure may attract increased professional participation. Posts generated from Monday to Thursday and on nonpublic holidays receive more responses. Because most of the certified counselors on the platform are not full-time clinical physicians working in offline institutions, they tend to be more active and willing to respond online during weekdays. Our results also show that page views and posting time do not affect response length. One possible explanation is that response length may depend on the professionalism, empathy, and motivation of health professionals rather than on the view number or posting time.

Research Implications

Theoretical Implications

This study deepens our understanding of knowledge contribution behaviors in nonprofit OMHCs. By examining how textual and contextual features of user posts influence the quantity and length of responses from mental health professionals, the findings reveal the important factors that shape health professionals' knowledge contributions. In addition, this research introduces interpretable machine learning methods into online mental health. This approach addresses the limitations of traditional regression models and black-box algorithms in explaining the influencing mechanisms. It also provides technical support for a deeper understanding of the factors affecting professional response quantity and length.

Practical Implications

The findings provide practical guidance for community managers. First, managers can categorize post topics to facilitate precise responses from professionals and help other users explore topics of interest. Second, providing predefined emotion-related keywords or tags on the post editing page helps enhance patients' ability to express their emotions and may increase the likelihood of receiving a response. Third, providing feedback on information richness and high-quality post templates may help users improve their expression. Fourth, considering the effects of view counts and posting times, platform operators can optimize content visibility strategies. For example, posts with lower view counts and those published during off-peak hours (eg, late night) can be prioritized. This may balance the exposure across posts of varying popularity and publication times and ensure that these posts receive professional responses. This may also enhance the overall fairness and quality of community interactions. Fifth, platform operators must exercise caution when implementing certain

XSL•FO
RenderX

strategies (eg, real-time feedback on content informativeness), as encouraging longer posts or greater emotional intensity may inadvertently increase the potential psychological burden on patients. To this end, when designing web interfaces, platforms should position these tools as supplementary and optional. It is crucial to ensure users retain control over these functionalities to balance interaction efficiency with psychological safety.

This study, based on Chinese OMHCs, may offer insights for mental health service platforms in other countries. However, we explicitly caution against the direct generalization of our findings to other cultural contexts. The counselor-led, nonprofit forums of the YiDianLing and YiXinLi platforms embed specific sociotechnical norms, such as text-based communication, work culture, and community governance rules. These factors may shape how patient users present their problems and how consultants perceive their role. Therefore, our conclusions should be interpreted as context-specific insights that highlight the need for future research to "unpack" these contextual differences through comparative studies.

In developing our model, we carefully addressed potential sources of bias, such as by using cross-temporal and cross-platform sampling methods to balance both positive and negative scenarios in predictions. When this research is extended to other countries, adjustments need to be made according to local cultural norms and service systems.

Limitations and Future Directions

Although our study offers significant contributions, it also has several limitations. First, the data were obtained primarily from the Chinese mental health platform YiDianLing, with supplementary data from YiXinLi used for robustness checks. Given the potential cultural, platform, and user differences across countries, generalizing our findings would require validation using multinational data from diverse platforms. Second, this study uses publicly available posts and response data from OMHCs. Future research could incorporate multimodal data (eg, images and emojis) to gain a deeper understanding of the interactions between patient users and mental health professionals in nonprofit mental health forums. Third, owing to privacy policies, user-level demographics (eg, sex, age, and membership duration) were unavailable. Subsequent studies or online experiments should examine how such characteristics influence forum participation. Fourth, this study solely explores response quantity and length. However, a complete understanding of community interaction also depends on the quality of responses, a multifaceted construct that encompasses aspects like relevance, empathy, and supportiveness. Future work should prioritize developing validated metrics for response quality to better evaluate professional contribution patterns. Furthermore, this study relies on a LightGBM model interpreted with the SHAP method to analyze post feature importance. Given the limited sample size and post-level features, the predictive accuracy may be limited. Future research should thus incorporate richer predictive features and additional interpretable machine learning techniques (eg, Local Interpretable Model-agnostic Explanations) to validate and extend these insights.

Conclusion

This study uses explainable machine learning methods to investigate the post features that influence response quantity and length in OMHCs. It highlights the importance of the post topic, post title, post length, post sentiment, and posting time. These findings provide insights for platform managers in terms of optimizing functional design and improving the effectiveness of community interactions.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (72471154), the General Project of the Ministry of Education Foundation on Humanities and Social Sciences (24YJC630049), and Shenzhen Science and Technology Innovation Commission (JCYJ20240813143012016).

Data Availability

The data used in this study are available upon reasonable request from the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Temporal dynamics of posts and responses on the YiDianLing platform.
[DOCX File , 17 KB - jmir_v28i1e74359_app1.docx]

Multimedia Appendix 2
Model performance evaluation.
[DOCX File , 248 KB - jmir_v28i1e74359_app2.docx]

Multimedia Appendix 3

XSL•FO
RenderX

The hyperparameter settings for the Shapley Additive Explanations algorithm.
[DOCX File , 12 KB - jmir_v28i1e74359_app3.docx]

Multimedia Appendix 4
Representative keywords for each topic identified by BERTopic.
[DOCX File , 14 KB - jmir_v28i1e74359_app4.docx]

Multimedia Appendix 5
Average absolute Shapley Additive Explanations values of features for predicting response quantity and length.
[DOCX File , 15 KB - jmir_v28i1e74359_app5.docx]

Multimedia Appendix 6
Analysis results for the YiXinli platform.
[DOCX File , 1433 KB - jmir_v28i1e74359_app6.docx]

References

1. De Hert M, Detraux J, Vancampfort D. The intriguing relationship between coronary heart disease and mental disorders. Dialogues Clin Neurosci 2018;20(1):31-40. [doi: 10.31887/DCNS.2018.20.1/mdehert] [Medline: 29946209]
2. Beurel E, Toups M, Nemeroff CB. The bidirectional relationship of depression and inflammation: double trouble. Neuron 2020;107(2):234-256 [FREE Full text] [doi: 10.1016/j.neuron.2020.06.002] [Medline: 32553197]
3. Ohrnberger J, Fichera E, Sutton M. The relationship between physical and mental health: a mediation analysis. Soc Sci Med 2017;195:42-49 [FREE Full text] [doi: 10.1016/j.socscimed.2017.11.008] [Medline: 29132081]
4. Drissi N, Ouhbi S, Janati Idrissi MA, Fernandez-Luque L, Ghogho M. Connected mental health: systematic mapping study. J Med Internet Res 2020;22(8):e19950 [FREE Full text] [doi: 10.2196/19950] [Medline: 32857055]
5. Schuch FB, Vancampfort D. Physical activity, exercise, and mental disorders: it is time to move on. Trends Psychiatry Psychother 2021;43(3):177-184 [FREE Full text] [doi: 10.47626/2237-6089-2021-0237] [Medline: 33890431]
6. Walker ER, McGee RE, Druss BG. Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis. JAMA Psychiatry 2015;72(4):334-341 [FREE Full text] [doi: 10.1001/jamapsychiatry.2014.2502] [Medline: 25671328]
7. Ali K, Farrer L, Gulliver A, Griffiths KM. Online peer-to-peer support for young people with mental health problems: a systematic review. JMIR Ment Health 2015;2(2):e19 [FREE Full text] [doi: 10.2196/mental.4418] [Medline: 26543923]
8. Li Pira G, Aquilini B, Davoli A, Grandi S, Ruini C. The use of virtual reality interventions to promote positive mental health: systematic literature review. JMIR Ment Health 2023;10:e44998 [FREE Full text] [doi: 10.2196/44998] [Medline: 37410520]
9. Tzeng Y, Yin W, Lin K, Wei J, Liou H, Sung H, et al. Factors associated with the utilization of outpatient virtual clinics: retrospective observational study using multilevel analysis. J Med Internet Res 2022;24(8):e40288 [FREE Full text] [doi: 10.2196/40288] [Medline: 35917486]
10. Oexle N, Ajdacic-Gross V, Kilian R, Müller M, Rodgers S, Xu Z, et al. Mental illness stigma, secrecy and suicidal ideation. Epidemiol Psychiatr Sci 2017;26(1):53-60 [FREE Full text] [doi: 10.1017/S2045796015001018] [Medline: 26606884]
11. Banwell E, Hanley T, De Ossorno Garcia S, Mindel C, Kayll T, Sefi A. The helpfulness of web-based mental health and well-being forums for providing peer support for young people: cross-sectional exploration. JMIR Form Res 2022;6(9):e36432 [FREE Full text] [doi: 10.2196/36432] [Medline: 36083629]
12. Liu J, Gao L. Lurking or active? The influence of user participation behavior in online mental health communities on the choice and evaluation of doctors. J Affect Disord 2022;301:454-462. [doi: 10.1016/j.jad.2022.01.074] [Medline: 35066007]
13. Park A, Conway M, Chen AT. Examining thematic similarity, difference, and membership in three online mental health communities from reddit: a text mining and visualization approach. Comput Human Behav 2018;78:98-112 [FREE Full text] [doi: 10.1016/j.chb.2017.09.001] [Medline: 29456286]
14. Feldhege J, Moessner M, Bauer S. Who says what? Content and participation characteristics in an online depression community. J Affect Disord 2020;263:521-527. [doi: 10.1016/j.jad.2019.11.007] [Medline: 31780138]
15. Saha B, Nguyen T, Phung D, Venkatesh S. A framework for classifying online mental health-related communities with an interest in depression. IEEE J Biomed Health Inform 2016;20(4):1008-1015. [doi: 10.1109/JBHI.2016.2543741] [Medline: 27008680]
16. Grub MF. Reddit as a "Safe Space": topic modeling of online mental health communities for depression and anxiety. Weizenbaum J Digit Soc 2025;5(3). [doi: 10.34669/wi.wjds/5.3.3]
17. AbouWarda H, Miscione G. Understanding how discourse themes in an online mental health community on twitter/x drive varied population-specific empowerment processes in alignment with global standards: a qualitative analysis of #bipolarclub. J Med Internet Res 2025;27:e74912 [FREE Full text] [doi: 10.2196/74912] [Medline: 40882197]

XSL•FO
RenderX

18. Zhou J, Zuo M, Ye C. Understanding the factors influencing health professionals' online voluntary behaviors: evidence from YiXinLi, a Chinese online health community for mental health. Int J Med Inform 2019;130:103939. [doi: 10.1016/j.ijmedinf.2019.07.018] [Medline: 31434043]
19. Kim M, Saha K, De Choudhury M, Choi D. Supporters First: understanding online social support on mental health from a supporter perspective. Proc ACM Hum-Comput Interact 2023;7(CSCW1):1-28 [FREE Full text] [doi: 10.1145/3579525]
20. Liu S, Xiao W, Fang C, Zhang X, Lin J. Social support, belongingness, and value co-creation behaviors in online health communities. Telemat Inform 2020;50:101398. [doi: 10.1016/j.tele.2020.101398]
21. Magane KM, Kenney M, Nelson E, Wisk L, Weitzman ER. The quality and safety of online health communities engaging adolescents around depression and substance use: a multisite evaluation. J Adolesc Health 2017;60(2):S77. [doi: 10.1016/j.jadohealth.2016.10.334]
22. Marshall P, Booth M, Coole M, Fothergill L, Glossop Z, Haines J, et al. Understanding the impacts of online mental health peer support forums: realist synthesis. JMIR Ment Health 2024;11:e55750 [FREE Full text] [doi: 10.2196/55750] [Medline: 38722680]
23. Morini V, Sansoni M, Rossetti G, Pedreschi D, Castillo C. Participant behavior and community response in online mental health communities: insights from reddit. Computers in Human Behavior 2025;165:108544. [doi: 10.1016/j.chb.2024.108544]
24. Li J, Liu D, Wan C, Liang Z, Zhu T. Empirical study of factors that influence the perceived usefulness of online mental health community members. Psych J 2023;12(2):307-318. [doi: 10.1002/pchj.629] [Medline: 36726193]
25. Smith-Merry J, Goggin G, Campbell A, McKenzie K, Ridout B, Baylosis C. Social connection and online engagement: insights from interviews with users of a mental health online forum. JMIR Ment Health 2019;6(3):e11084 [FREE Full text] [doi: 10.2196/11084] [Medline: 30912760]
26. Sharma E, De Choudhury CM. Mental health support and its relationship to linguistic accommodation in online communities. 2018 Presented at: CHI '18: CHI Conference on Human Factors in Computing Systems; 2018 April 21-26; Montreal QC Canada p. 1-13. [doi: 10.1145/3173574.3174215]
27. Srivastava A, Gupta T, Cerezo A, Lord SP, Akhtar MS, Chakraborty T. Critical behavioral traits foster peer engagement in online mental health communities. PLoS One 2025;20(1):e0316906 [FREE Full text] [doi: 10.1371/journal.pone.0316906] [Medline: 39804944]
28. Chen L, Baird A, Straub D. A linguistic signaling model of social support exchange in online health communities. Decis Support Syst 2020;130:113233. [doi: 10.1016/j.dss.2019.113233]
29. Wang J, Chiu Y, Yu H, Hsu Y. Understanding a nonlinear causal relationship between rewards and physicians' contributions in online health care communities: longitudinal study. J Med Internet Res 2017;19(12):e427 [FREE Full text] [doi: 10.2196/jmir.9082] [Medline: 29269344]
30. Zhou T. Examining online health community users’ sharing behaviour: a social influence perspective. Inf Dev 2021;38(4):599-608. [doi: 10.1177/02666669211007188]
31. Chen Q, Jin J, Yan X. Understanding physicians' motivations for community participation and content contribution in online health communities. OIR 2022;47(3):604-629. [doi: 10.1108/oir-11-2021-0615]
32. Imlawi J, Gregg D. Understanding the satisfaction and continuance intention of knowledge contribution by health professionals in online health communities. Inform Health Soc Care 2020;45(2):151-167. [doi: 10.1080/17538157.2019.1625053] [Medline: 31328593]
33. Maheshwari B, Sarrion M, Motiani M, O'Sullivan S, Chandwani R. Exploration of factors affecting the use of Web 2.0 for knowledge sharing among healthcare professionals: an Indian perspective. JKM 2020;25(3):545-558 [FREE Full text] [doi: 10.1108/jkm-02-2020-0105]
34. Derks D, Fischer AH, Bos AER. The role of emotion in computer-mediated communication: a review. Comput Hum Behav 2008;24(3):766-785 [FREE Full text] [doi: 10.1016/j.chb.2007.04.004]
35. Zhuo X, Wang W. Why are physicians willing to contribute knowledge? Evidence from online health communities. Comput Hum Behav 2024;152:108095. [doi: 10.1016/j.chb.2023.108095]
36. Feng X, Hu Y, Pfaff H, Liu S, Xie J, Zhang Z. Exploring client preferences for psychological counselors in a chinese online health community: longitudinal study. J Med Internet Res 2024;26:e58428 [FREE Full text] [doi: 10.2196/58428] [Medline: 39388694]
37. Liu H, Zhang L, Wang W, Huang Y, Li S, Ren Z, et al. Prediction of online psychological help-seeking behavior during the COVID-19 pandemic: an interpretable machine learning method. Front Public Health 2022;10:814366 [FREE Full text] [doi: 10.3389/fpubh.2022.814366] [Medline: 35309216]
38. Baba A, Bunji K. Prediction of mental health problem using annual student health survey: machine learning approach. JMIR Ment Health 2023;10:e42420 [FREE Full text] [doi: 10.2196/42420] [Medline: 37163323]
39. Huang Y, Liu H, Chi M, Meng S, Wang W. Digit Health 2025;11:20552076251333480 [FREE Full text] [doi: 10.1177/20552076251333480] [Medline: 40343065]
40. YiDianLing. 2025. URL: https://www.ydl.com/[accessed [accessed 2025-08-19]
41. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational

XSL•FO
RenderX

Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); 2025 December 02; Minneapolis, Minnesota p. 4171-4186. [doi: 10.18653/v1/N19-1423]
42. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv Preprint posted online on March 11, 2022 [FREE Full text] [doi: 10.48550/arXiv.2203.05794]
43. DistilBERT-based multilingual sentiment classification model. Hugging Face. 2025. URL: https://huggingface.co/tabularisai/multilingual-sentiment-analysis [accessed 2025-08-01]
44. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv Preprint posted online on October 2, 2019 [FREE Full text]
45. AlQadi RA, Taie SA, Idrees AM, Elhariri E. Explainable deep learning model for ChatGPT-rephrased fake review detection using DistilBERT. BDCC 2025;9(8):205. [doi: 10.3390/bdcc9080205]
46. Jojoa M, Eftekhar P, Nowrouzi-Kia B, Garcia-Zapirain B. Natural language processing analysis applied to COVID-19 open-text opinions using a distilBERT model for sentiment categorization. AI Soc 2022;39:1-8 [FREE Full text] [doi: 10.1007/s00146-022-01594-w] [Medline: 36439363]
47. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. 2017 Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 December 4 - 9; Long Beach California USA p. 3149-3157.
48. Tai C, Liao T, Chen S, Chung M. Sleep stage classification using light gradient boost machine: exploring feature impact in depressive and healthy participants. Biomed Sigal Process Control 2024;88:105647. [doi: 10.1016/j.bspc.2023.105647]
49. LightGBM parameters tuning. Read the Docs. 2025. URL: https://lightgbm.readthedocs.io/en/latest/Parameters-Tuning.html [accessed 2025-12-04]
50. Le Glaz A, Haralambous Y, Kim-Dufor D, Lenca P, Billot R, Ryan TC, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res 2021;23(5):e15708 [FREE Full text] [doi: 10.2196/15708] [Medline: 33944788]
51. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):56-67 [FREE Full text] [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
52. YiXinLi. URL: https://www.xinli001.com/[accessed [accessed 2025-08-19]
53. González Moreno A, Molero Jurado MDM. Presence of emotions in network discourse on mental health: thematic analysis. Psychiatry Int 2024;5(3):348-359. [doi: 10.3390/psychiatryint5030024]
54. Akar E. Connecting for well-being: a role-based network analysis of online mental health communities. Behav Inf Technol 2025;44(18):4566-4580. [doi: 10.1080/0144929x.2025.2484393]
55. Zhang L, Liu D, Li J, Wan C, Liu X. Exploring linguistic features and user engagement in Chinese online mental health counseling. Heliyon 2024;10(19):e38042. [doi: 10.1016/j.heliyon.2024.e38042] [Medline: 39678785]
56. De Choudhury M, De S. Mental health discourse on reddit: self-disclosure, social support, and anonymity. ICWSM 2014;8(1):71-80. [doi: 10.1609/icwsm.v8i1.14526]
57. Chu TH, Sun M, Crystal Jiang L. Self-disclosure in social media and psychological well-being: A meta-analysis. J Soc Pers Relatsh 2022;40(2):576-599. [doi: 10.1177/02654075221119429]
58. Liu J, Kong J. Why do users of online mental health communities get likes and reposts: a combination of text mining and empirical analysis. Healthcare (Basel) 2021;9(9):1133 [FREE Full text] [doi: 10.3390/healthcare9091133] [Medline: 34574907]
59. Liu J, Liu Y. Exploring the user interaction network in an anxiety disorder online community: an exponential random graph model with topical and emotional effects. Int J Environ Res Public Health 2022;19(11):6354 [FREE Full text] [doi: 10.3390/ijerph19116354] [Medline: 35681939]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
DistilBERT: distilled Bidirectional Encoder Representations from Transformers
LightGBM: Light Gradient Boosting Machine
OMHC: online mental health community
RF: random forest
SHAP: Shapley Additive Explanations
SVM: Support Vector Machine
TF-IDF: term frequency-inverse document frequency
XGBoost: Extreme Gradient Boosting

XSL•FO
RenderX

Edited by J Sarvestan; submitted 23.Mar.2025; peer-reviewed by S Narayan, S Kath; comments to author 07.Apr.2025; accepted 24.Nov.2025; published 05.Jan.2026.

Please cite as:
Geng S, Li Y, Wang J, Chen P, Wu X, Zhang Z
Predictors of Professional Responses in Nonprofit Mental Health Forums: Interpretable Machine Learning Analysis
J Med Internet Res 2026;28:e74359
URL: https://www.jmir.org/2026/1/e74359
doi:10.2196/74359
PMID:

©Shuang Geng, Yanghui Li, Jie Wang, Peixuan Chen, Xusheng Wu, Zhiqun Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effect of AI-Based Natural Language Feedback on Engagement and Clinical Outcomes in Fully Self-Guided Internet-Based Cognitive Behavioral Therapy for Depression: 3-Arm Randomized Controlled Trial

Mirai So[1*], MD, MBA, PhD; Yoichi Sekizawa[2*], MA; Sora Hashimoto[3*], MA; Masami Kashimura[4*], PhD; Hajime Yamakage[5*], MEng; Norio Watanabe[6*], MD, PhD

[1]Department of Psychiatry, Tokyo Dental College, Tokyo, Japan

[2]Research Institute of Economy, Trade and Industry, Tokyo, Japan

[3]United Health Communication Co., Ltd., Tokyo, Japan

[4]Department of Psychology, Faculty of Human Schiences, Tokiwa University, Ibaraki, Japan

[5]Department of Medical Statistics, Satista Co., Ltd., Kyoto, Japan

[6]Department of Psychiatry, Soseikai General Hospital, Kyoto, Japan

[*]all authors contributed equally

Corresponding Author:
Mirai So, MD, MBA, PhD
Department of Psychiatry
Tokyo Dental College
2-9-18 Misakicho, Chiyoda-ku
Tokyo, 101-0061
Japan
Phone: 81 47 322 0151
Fax: 81 47 325 4456
Email: mirai.so@keio.jp

Abstract

Background: Depression remains a major global cause of disability; yet, access to optimal mental health services is limited. Self-guided internet-based cognitive behavioral therapy (iCBT) offers a scalable alternative but is generally less effective than guided programs, showing limited antidepressant effects and incomplete symptomatic and functional recovery. Adherence remains a major barrier. Recent advances in artificial intelligence (AI), particularly natural language processing, enable automated advisory and empathic feedback that may enhance engagement and therapeutic impact. Although previous trials have reported promising effects, most used heterogeneous control conditions, making it difficult to isolate the specific contribution of AI within fully self-guided interventions.

Objective: This randomized controlled trial evaluated whether natural language processing–based AI feedback integrated into a fully self-guided iCBT program improves clinical outcomes and engagement compared with an otherwise identical iCBT program without AI support.

Methods: We recruited 1187 adults aged 20-60 years online and randomly assigned them to AI-augmented iCBT (AI-iCBT; n=396), iCBT without AI (n=397), or a waitlist control (n=394). Both active groups received 6 weekly sessions combining video-based psychoeducation and cognitive restructuring exercises. The AI-iCBT program additionally provided automated empathic and advisory feedback. The primary outcome was depressive symptom severity (Patient Health Questionnaire-9 [PHQ-9]) at week 7 and month 3, analyzed using mixed-effects models for repeated measures under an intention-to-treat framework. Secondary outcomes included a dichotomous PHQ-9 score of ≥10, Quick Inventory of Depressive Symptomatology, Generalized Anxiety Disorder-7, Sheehan Disability Scale, and weekly participation rates. Exploratory analyses assessed the impact of AI functions on engagement and antidepressant effects in the efficacy analysis set (EAS).

Results: In intention-to-treat analyses, no significant between-group differences were observed in mean PHQ-9 scores at week 7 or month 3, whereas engagement analyses showed a significant group × week interaction, with AI-iCBT participants demonstrating consistently higher odds of weekly participation (odds ratio 1.23, 95% CI 1.09-1.39; *P*<.001). Exploratory analyses indicated

XSL•FO
RenderX

that activation of the empathic feedback function strongly predicted adherence (odds ratio 9.99, 95% CI 5.80-17.21; *P*<.001), while advisory feedback was not significant. In EAS analyses, iCBT showed significant short-term improvement versus control at postintervention, whereas at follow-up, only AI-iCBT showed a significantly lower proportion of participants with a PHQ-9 score of ≥10 compared with control (difference –0.15, 95% CI –0.30 to –0.01; *P*=.046). No serious adverse events were reported.

Conclusions: AI support significantly improved adherence to a fully self-administered program. In EAS analyses, AI-iCBT also showed a significantly lower proportion of participants with PHQ-9 score of ≥10 at follow-up compared with control. Empathic feedback emerged as a key mechanism for sustaining engagement, suggesting that AI communication may help maintain participation in scalable digital mental health interventions. Further research is required.

Trial Registration: University Hospital Medical Information Network Clinical Trials Registry (UMIN-CTR) UMIN000019228; https://center6.umin.ac.jp/cgi-open-bin/ctr/ctr_view.cgi?recptno=R000022220

(J Med Internet Res 2026;28:e76902) doi:10.2196/76902

KEYWORDS

Adherence; AI; AI-supported psychotherapy; artificial intelligence; CBT; depression; internet-based CBT; natural language processing; RCT; self-help intervention

Introduction

Depression is a leading global cause of disability [1], substantially impairing quality of life [2,3] and imposing a considerable economic burden, including medical expenses [4] and productivity losses [5]. The rising prevalence of depression, combined with a shortage of health care resources, places a significant strain on health systems and professionals in meeting the growing demand [6-9].

In response, technology-delivered self-help interventions have emerged as promising solutions for managing mental health difficulties. Most of these interventions are based on cognitive behavioral therapy (CBT) and are generally referred to as internet-based CBT (iCBT) [10,11]. iCBT can be delivered either with therapist support (guided) or without therapist support (unguided, self-directed). Compared to traditional face-to-face therapy, iCBT offers major advantages in terms of accessibility, availability, and cost-effectiveness for both patients and providers. Furthermore, its online format provides benefits related to privacy, confidentiality, and anonymity, which can help reduce the stigma often associated with seeking mental health care [12,13].

While iCBT produces clinically meaningful symptom improvements, remission rates tend to be modest (approximately 30%-35%). An individual participant data meta-analysis reported a remission rate of 35.2% and a response rate of 56% [14]. Large-scale individual participant data network meta-analyses have consistently shown that guided iCBT yields higher response and remission rates than unguided formats, reflecting the challenges of engagement and dropout in fully self-administered programs [15-19].

This study specifically examines the unguided, fully self-administered format. Such interventions enable users to manage their symptoms independently and offer potential benefits such as reducing costs, alleviating the burden on health care providers, and improving access to mental health services in regions where such services are difficult to obtain [9,16]. Previous research has demonstrated that sociodemographic factors, such as age and sex, are associated with dropout risk in iCBT [20-22].

Nevertheless, if the effectiveness and engagement of unguided iCBT could be enhanced, the benefits of structured self-help materials would not be limited to fully self-administered interventions but would also extend to guided and blended formats. When patients acquire skills and knowledge through self-help modules, they can participate more effectively in therapist-led sessions, thereby enhancing overall treatment outcomes [23]. Strengthening such “self-help effects” not only amplifies therapeutic gains in guided and blended care but also reduces the time and workload required from therapists. By improving scalability and cost-effectiveness, it further increases the feasibility of implementation in routine practice [24,25]

To address these challenges, natural language processing (NLP), an artificial intelligence (AI) technology that enables the understanding and generation of human language, has increasingly been applied to enhance adherence and engagement through tailored feedback [26-28]. Whereas conventional unguided iCBT typically provides static or generic responses, in this study, we used an NLP-enabled iCBT program with automated advisory and empathic functions. This allows the system to generate advisory and empathic feedback in response to user input, potentially addressing both emotional and procedural barriers simultaneously.

Despite its promise, the specific therapeutic contributions of NLP remain unclear. Previous studies have frequently used heterogeneous control conditions such as waitlists, no intervention, treatment as usual, bibliotherapy, or conversational computer programs [28-32]. This heterogeneity makes it difficult to determine whether NLP provides distinct therapeutic benefits or merely functions as an active placebo by enhancing user expectations.

Against this background, the aim of this study was to conduct a randomized, parallel-group exploratory trial directly comparing 2 unguided, fully self-administered iCBT programs that were identical except for the presence or absence of NLP feedback. This design allowed us to evaluate the therapeutic contribution of NLP within a self-help framework in a blinded comparison of the 2 intervention groups.

XSL•FO
RenderX

Methods

Overview

This study was a 3-arm randomized controlled trial, with double-blinding between the AI-augmented iCBT (AI-iCBT) and iCBT groups, while the waitlist control group was unblinded. The intervention arms consisted of AI-iCBT, an unguided, fully self-administered iCBT program incorporating NLP feedback, and iCBT, an unguided, fully self-administered program without NLP feedback. These 2 arms are hereafter collectively referred to as "unguided iCBT." The waitlist group served as the control condition.

Study Participants

Invitation emails were sent to all monitors registered with Nikkei Research Inc, with the aim of recruiting at least 900 participants. Interested individuals were directed to complete an online screening survey, which included the Patient Health Questionnaire-9 (PHQ-9) [33,34], to determine eligibility.

Eligibility Criteria

Inclusion and exclusion criteria are provided in Textbox 1.

Textbox 1. Eligibility criteria.

Inclusion criteria

- Aged 20-60 years (to target the working-age population and exclude older adults with lower digital literacy)
- Had access to the internet
- Baseline Patient Health Questionnaire-9 score of ≥5 (this threshold was selected to avoid floor effects and ensure adequate symptom levels for change detection) [33,35,36]
- Ability to understand Japanese

Exclusion criteria

- Presence of medical conditions precluding participation as determined by a physician
- Concurrent participation in another cognitive behavioral therapy program
- Diagnosis of schizophrenia
- Severe suicidal ideation
- Diagnosis of dementia
- Substance dependence in the past 12 months (excluding smoking)

Eligible participants were randomly assigned to 1 of the 3 groups. Immediately before the intervention, PHQ-9 scores were reassessed; individuals with scores <5 were excluded from the efficacy analysis set (EAS) but remained in the overall study population. The detailed definition of the EAS is provided in the Analytic Strategy section.

Intervention

The AI-augmented iCBT program, developed by NEC Solution Innovators, Ltd (Tokyo), integrates an NLP module trained on 28,718 prior iCBT records from Japanese users. The entire program was delivered in Japanese, and Figure 1 presents an English-translated version of the original Japanese interface for publication purposes. The program includes a self-guided cognitive restructuring (CR) exercise, where users complete a 7-column thought record to address cognitive distortions. The NLP system processes user inputs, including situation, automatic thoughts, and feelings, referencing a corpus of past responses (Figure 1). It provides 2 types of automated feedback with text and phonation: (1) empathetic messages delivered through an animated character whose expressions, such as smiling or showing sadness, are synchronized with the message content, and (2) advisory messages offering guidance to refine inputs or direct users to appropriate exercises, including suggestions to revise content if the user's input was unclear or misplaced (eg, a feeling given instead of a thought; Figure 2).

In contrast, the non-AI iCBT program retained the same structure but provided only neutral, noncontextual responses, such as generic phrases like "Uh-huh" with neutral facial expressions. Both AI-iCBT and iCBT programs were otherwise identical in content, using the same validated 6-session self-guided iCBT package. This iCBT program has previously demonstrated significant antidepressant effects in a randomized controlled trial among working adults (n=60 per group), compared with a waitlist control, with a medium to large effect size (Cohen d=0.65; P<.005) [37]. This package consisted of 6 weekly sessions, each including a 15-minute video-based psychoeducation module covering standard CBT principles such as behavioral activation and problem-solving, along with a weekly CR exercise in which users applied learned techniques. In this trial, the only difference was the addition of the NLP feedback system. The program was available on both smartphones and PCs. The AI-enhanced features, which were designed in advance to improve user engagement and response accuracy, exhibited high usability, with low dissatisfaction rates reported for both the empathetic (3/32, 9.4%) and advisory (1/24, 4.2%) feedback functions (Figure 3).

XSL•FO
RenderX

Figure 1. Examples of expressions extracted from the natural language processing corpus and categorized into 4 domains: Problem, Trouble, Feeling, and Subjective.

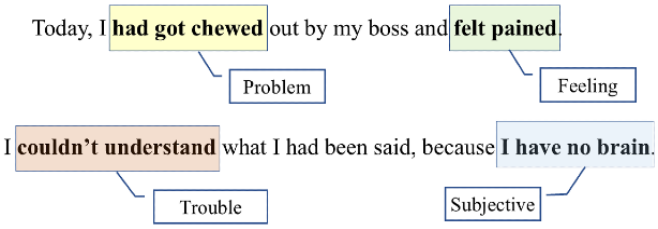

Figure 2. Workflow of artificial intelligence-guided internet-based cognitive behavioral therapy (CBT), showing the structured 7-step cognitive restructuring exercise with automated prompts and feedback. AT: automatic thought; NLP: natural language processing.

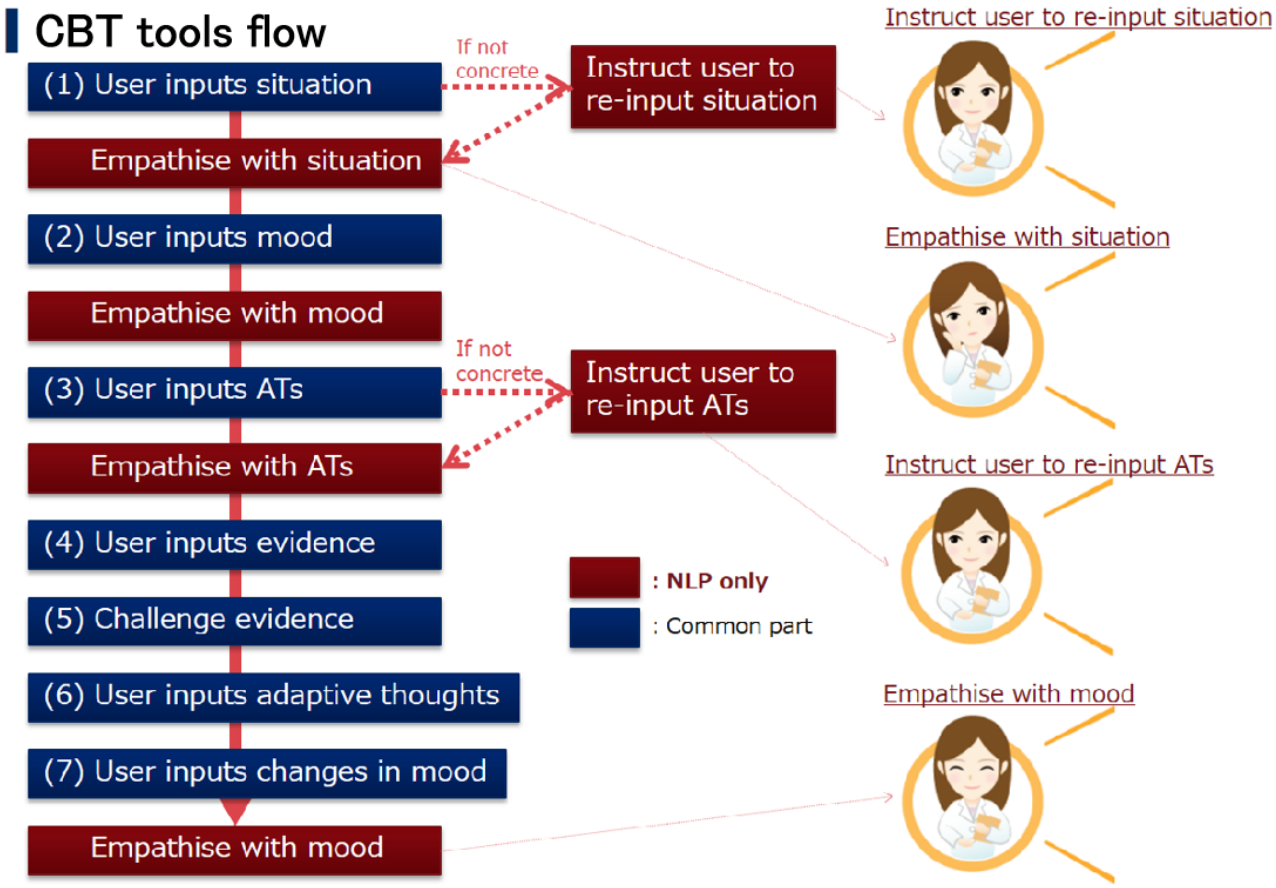

XSL•FO
RenderX

Figure 3. User acceptability ratings of natural language processing–generated feedback for empathy and automatic thought identification. AT: automatic thought.

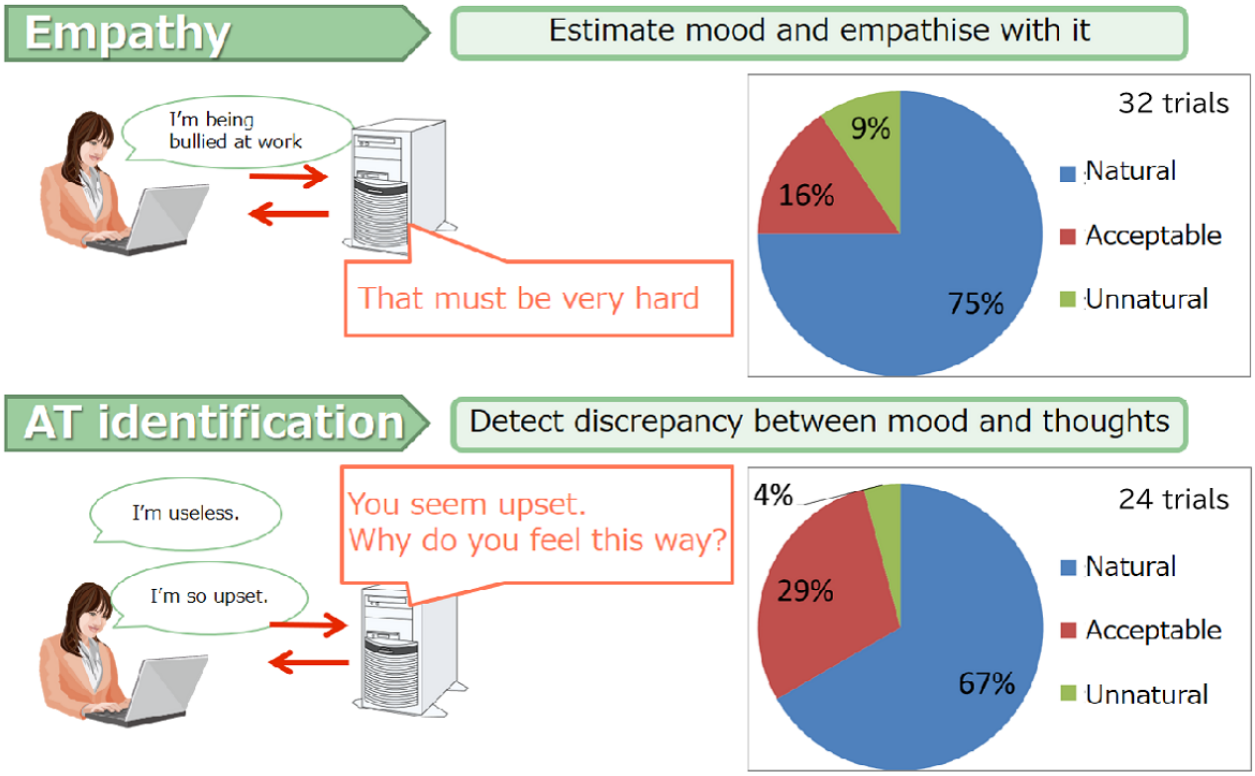

Randomization and Masking

The registered participants were randomly and concurrently assigned to either the AI-iCBT, iCBT, or waitlist groups using a computer-generated random sequence provided by an independent third party. Stratified randomization was applied based on age (≤40 vs >40 years), sex, and baseline PHQ-9 score (≤9 vs >9), as baseline symptom severity has been shown to influence treatment outcomes in self-guided iCBT [23]. Participants in the waitlist group were aware of their allocation and were therefore unblinded, whereas those in the AI-iCBT and iCBT groups were told only that they would participate in iCBT using "the latest technology," without disclosure of their specific group assignment. Accordingly, blinding was implemented between the 2 intervention groups.

Study Procedures

Automated email reminders were sent to participants twice weekly during the 7-week intervention period. Each week, participants in the intervention groups were required to (1) view an online psychoeducational CBT module and (2) perform their allocated (AI-iCBT or iCBT) CR exercise at least once (6 times or more in total). Waitlist participants did not undergo any exercises during this period. All participants were required to complete assessments at baseline, postintervention (week 7), and follow-up (month 3 after postintervention). All intervention and assessment procedures, including attendance and outcome measures, were conducted online.

Outcomes

All primary and secondary outcomes were analyzed based on the intention-to-treat (ITT) population, which included all randomized participants.

Primary Outcome

The primary outcome was the mean PHQ-9 score, assessed at baseline, week 7 (postintervention), and month 3 (follow-up). The PHQ-9 is a widely used self-report measure of depressive symptoms (range 0-27, higher scores indicating greater severity), originally developed by Kroenke et al [33] and validated in Japanese [38].

Secondary Outcomes

Secondary outcomes include (1) proportion of participants with PHQ-9 scores ≥10 (a conventional cutoff for probable major depression) [33,39]. Although not selected as the primary outcome in this study, such binary outcomes are often considered clinically meaningful, as they reflect remission from a diagnostic threshold [40-43]. (2) Quick Inventory of Depressive Symptomatology-Japanese version (QIDS-J) [43,44]—a self-report scale of depressive symptom severity. (3) Generalized Anxiety Disorder-7 (GAD-7) [45,46]—a self-report questionnaire measuring generalized anxiety symptoms. (4) Sheehan Disability Scale (SDS) [47,48], which evaluates functional impairment in work, social, and family life, with SDS ≥10 commonly used as a pragmatic threshold in clinical trials [49].

Engagement outcome included weekly CR exercise attendance rate (defined as attending at least one session per week) in the 2 intervention groups. Program satisfaction at week 7 was

XSL•FO
RenderX

assessed with the Client Satisfaction Questionnaire-8 (CSQ-8) [50,51], for which the Japanese version has demonstrated reliability and validity.

All outcomes except engagement and satisfaction were assessed at baseline, week 7, and month 3.

Analytic Strategy

Overview

In this study, all primary and secondary analyses were conducted in the ITT population, defined as all randomized participants. In addition, 2 exploratory analyses were performed: (1) as an ad hoc exploratory analysis, we examined which AI feedback function (empathy or advisory) contributed more to enhancing engagement, and (2) as an additional exploratory analysis, we assessed continuous and binary PHQ-9 outcomes within the EAS.

Engagement-Enhancing Factors

For this analysis, the 2 intervention groups were combined, and the presence or absence of empathy and advisory feedback during week 1 was examined in relation to engagement from weeks 2 to 6, defined as completing at least 1 exercise per week. The detailed statistical methods are described in the Statistical Analysis section.

Efficacy Analysis Set

The EAS was defined as participants with a baseline PHQ-9 score of ≥5 and completion of at least 3 out of the 6 weekly sessions. Participants with a baseline PHQ-9 score of <5 (minimal symptoms) were excluded, as their inclusion could reduce the power to detect change and dilute the mean effects [52,53]. Furthermore, previous research has demonstrated a dose-response relationship in iCBT, with clinical benefits emerging after completing approximately half of the modules; therefore, the minimum attendance criterion was set at 3 of 6 sessions [54,55].

Statistical Analysis

Overview

The sample size was estimated based on an assumed effect size of 0.10 (Cohen *d*) between the AI-iCBT and iCBT groups, given the absence of directly comparable prior studies. A dropout rate of 50% was anticipated based on patterns observed in similar previous studies. The power was set at 80% with a 2-sided significance level of α=.05. As this was an exploratory study, no adjustment for multiplicity was applied, and nominal *P* values were reported.

The primary and secondary analyses were conducted according to the ITT principle, including all randomized participants. Baseline demographic and clinical characteristics were compared across groups using 1-way ANOVA or chi-square tests.

Continuous outcomes (PHQ-9, QIDS-J, GAD-7, and SDS) were analyzed using a mixed-effects model for repeated measures (MMRM), with intervention, time, and intervention × time interaction as fixed effects, assuming an unstructured covariance structure. Results are presented as least squares means with 95% CIs.

Binary outcomes (PHQ-9 ≥10) were analyzed using generalized linear mixed models (GLMMs) with a binomial distribution and logit link, including intervention, time, and their interaction as fixed effects, and subject as a random effect. Estimated proportions and their 95% CIs were reported. Missing data for the outcomes were handled under the missing at random assumption within the MMRM and GLMMs framework.

CR exercise participation rates (defined as at least 1 completion per week) in the intervention groups were analyzed using GLMMs with a logit link, including intervention, week (as a continuous variable), and intervention × week interaction as fixed effects.

Exploratory Analyses

The following two exploratory analyses were conducted.

Engagement-Enhancing Factors

The dependent variable was defined as achieving at least 1 CR exercise per week across all weeks from week 2 to week 6 (yes/no). Independent variables were the presence or absence of empathy or advisory feedback during week 1. Covariates included age, sex, marital status, education, employment status, history of psychiatric and physical treatment, baseline PHQ-9 score, and intervention group (AI-iCBT vs iCBT), as group differences could confound the association of interest. Analyses were performed using generalized estimating equations logistic regression models to account for within-subject correlation and to estimate population-averaged effects.

Efficacy Analysis Set

In the EAS (participants with a baseline PHQ-9 score of ≥5 and completion of ≥3 sessions), continuous and binary PHQ-9 outcomes were additionally adjusted for age, sex, baseline PHQ-9 score, and medical history as covariates to account for potential group imbalances in the restricted sample.

Sensitivity Analyses

Associate Factors of Low Adherence

To explore potential factors associated with dropout, we compared baseline characteristics between EAS participants who attended ≥3 sessions and those who attended <3 sessions, given the high attrition typically observed in self-guided digital interventions.

Alternative Definition of Caseness

We conducted an exploratory analysis on the binary PHQ-9 outcome in the EAS, applying a stricter definition of depression severity: a PHQ-9 score of ≥10 plus at least 1 core symptom (depressed mood or anhedonia) [56,57], together with an SDS score of ≥10 as an indicator of functional impairment [49,58-60].

All analyses were performed using IBM SPSS Statistics version 26.0.

Reporting Standards

Reporting of this trial followed the CONSORT (Consolidated Standards of Reporting Trials) 2010 statement [61] and the

XSL•FO
RenderX

CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist [62] for internet-based interventions. The completed CONSORT-EHEALTH checklist is submitted as Multimedia Appendix 1.

Ethical Considerations

This study was reviewed and approved by the Hiramatsu Memorial Hospital Ethics Committee (approval number 20150807). All participants provided informed consent electronically prior to enrollment after reading an online information sheet describing the study purpose, procedures, potential risks, and voluntary nature of participation. Participants were informed that they could withdraw at any time without penalty.

The trial was prospectively registered in the University Hospital Medical Information Network Clinical Trials Registry.

All data were anonymized prior to analysis to ensure confidentiality. No personally identifiable information was accessible to the research team. Participants who completed the final assessment received a ¥500 (US $4.5) gift voucher as compensation. No identifiable images or other personal data are presented in this manuscript.

Results

Study Participants

A total of 1187 participants were eligible and randomly allocated to the AI-iCBT (n=396), iCBT (n=397), or waitlist (n=394) groups (ITT population; see Figure 4 for the CONSORT flow diagram). Baseline demographic and clinical characteristics are summarized in Table 1. The mean age was 43.50 (SD 9.85) years, and 699 (58.8%) of 1187 participants were male. Across the 3 groups, demographic and clinical characteristics were well balanced, with no significant differences in depressive symptom severity (PHQ-9: P=.56; QIDS-J: P=.74). No significant baseline differences were found between the AI-iCBT and iCBT groups, confirming the comparability of the 2 active interventions.

Figure 4 shows the flow of participants through the trial, including the numbers assessed for eligibility, randomized, allocated to each study arm (AI-iCBT, iCBT, control), completing follow-up assessments at week 7 and month 3, and included in the ITT analysis.

XSL•FO
RenderX

Figure 4. CONSORT (Consolidated Standards of Reporting Trials) 2010 flow diagram of participant enrollment, allocation, follow-up, and analysis. AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy; iCBT: internet-based cognitive behavioral therapy; ITT: intention-to-treat.

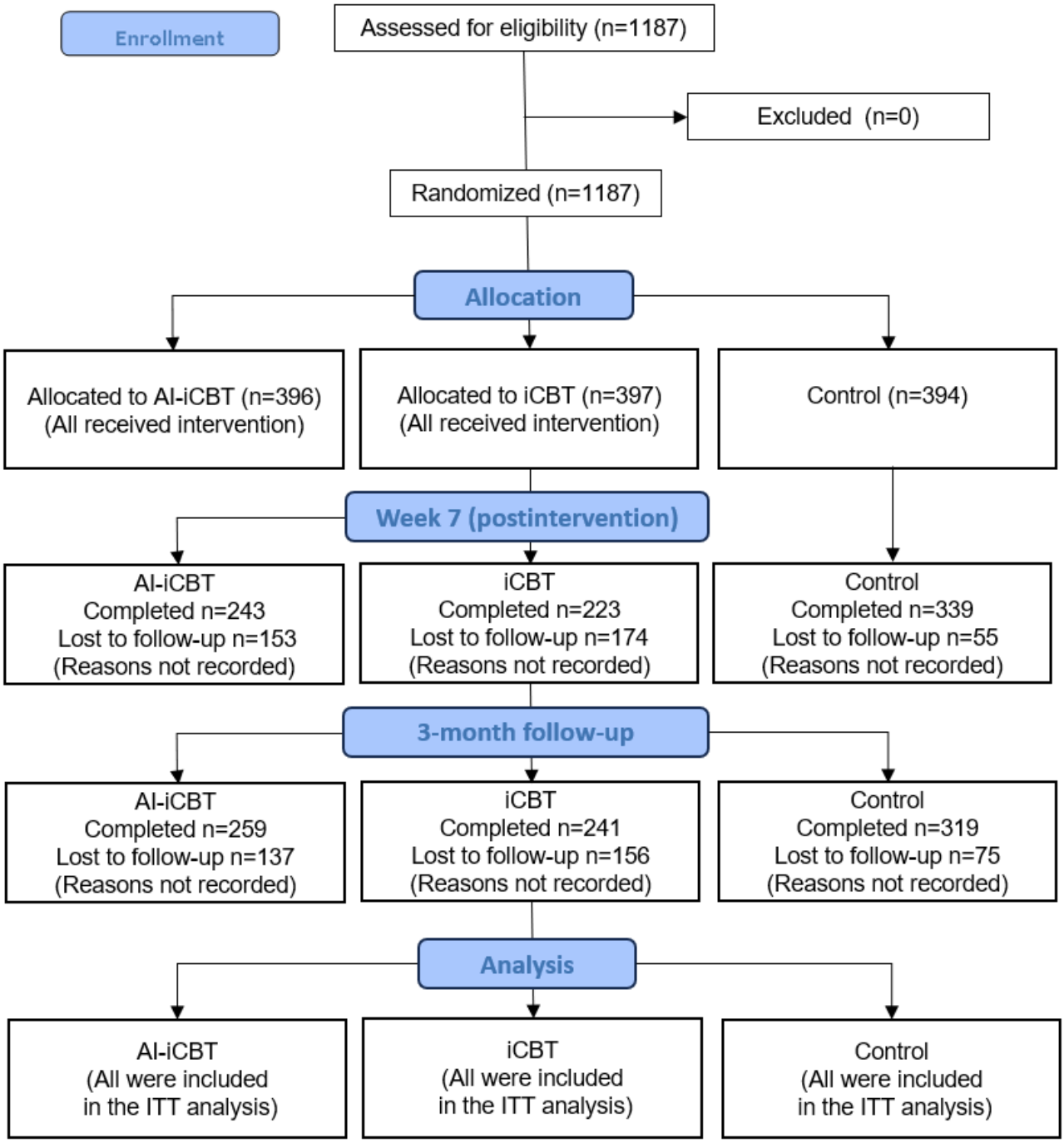

XSL•FO
RenderX

Table 1. Participants' characteristics.

Characteristic	Total (N=1187)	AI-iCBT[a] (n=396)	iCBT[b] (n=397)	Control (n=394)	*P* value (overall)	*P* value (AI-iCBT vs iCBT)[c]
Sex, n (%)					.96[d]	N/A[e]
Male	698 (58.8)	232 (58.6)	232 (58.4)	234 (59.4)		
Female	489 (41.2)	164 (41.4)	165 (41.6)	160 (40.6)		
Age (years)					.81[f]	N/A
Mean (SD)	43.5 (9.9)	43.6 (9.5)	43.2 (9.9)	43.6 (10.1)		
Median (IQR)	44 (36-52)	44 (36.8-51)	44 (36-52)	45 (36-52)		
Minimum-Maximum	20-60	20-60	20-60	20-60		
Marital status, n (%)					.28[d]	N/A
Married	665 (56)	226 (57.1)	219 (55.2)	220 (55.8)		
Divorced	76 (6.4)	29 (7.3)	17 (4.3)	30 (7.6)		
Bereaved	7 (0.6)	2 (0.5)	4 (1)	1 (0.3)		
Single	439 (37)	139 (35.1)	157 (39.5)	143 (36.3)		
Educational background, n (%)					.63[d]	N/A
Junior high school	7 (0.6)	3 (0.8)	1 (0.3)	3 (0.8)		
High school	239 (20.1)	87 (22)	84 (21.2)	68 (17.3)		
Junior college or technical	216 (18.2)	71 (17.9)	73 (18.4)	72 (18.3)		
University or postgraduate	725 (61.1)	235 (59.3)	239 (60.2)	251 (63.7)		
Employment status, n (%)					.78[d]	N/A
Working	972 (81.9)	320 (80.8)	324 (81.6)	328 (83.2)		
Unemployed (seeking)	79 (6.7)	27 (6.8)	30 (7.6)	22 (5.6)		
Unemployed (not seeking)	136 (11.5)	49 (12.4)	43 (10.8)	44 (11.2)		
Medical history, n (%)					.13[d]	N/A
No relevant history	918 (77.3)	316 (79.8)	296 (74.6)	306 (77.7)		
Ambulatory	258 (21.7)	74 (18.7)	97 (24.4)	87 (22.1)		
Hospitalized	11 (0.9)	6 (1.5)	4 (1)	1 (0.3)		
Mental history, n (%)					.93[d]	N/A
In treatment	129 (10.9)	46 (11.6)	44 (11.1)	39 (9.9)		
Treated	184 (15.5)	61 (15.4)	59 (14.9)	64 (16.2)		
No relevant history	874 (73.6)	289 (73)	294 (74.1)	291 (73.9)		
PHQ-9[g] score ≥10	428 (36.1)	142 (35.9)	143 (36)	143 (36.3)	.99[d]	N/A
Baseline scale score, mean (SD)						
PHQ-9	8.7 (5.2)	8.8 (5.2)	8.5 (5.2)	8.9 (5.1)	.56[f]	.41[f]
QIDS-J[h]	8.7 (4.9)	8.8 (4.9)	8.6 (4.9)	8.8 (4.7)	.74[f]	.49[f]
GAD-7[i]	6.0 (4.6)	6.0 (4.4)	5.9 (4.7)	6.2 (4.7)	.59[f]	.93[f]

[a]AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy.

[b]iCBT: internet-based cognitive behavioral therapy.

[c]*P* values represent pairwise comparisons between AI-iCBT and iCBT groups.

[d]*P* value is based on the chi-square test.

[e]N/A: not applicable.

[f]*P* value is based on ANOVA.

XSL•FO
RenderX

[g]PHQ-9: Patient Health Questionnaire-9.

[h]QIDS-J: Quick Inventory of Depressive Symptomatology-Japanese version.

[i]GAD-7: Generalized Anxiety Disorder-7.

Primary and Secondary Outcomes (ITT Population)

The primary outcome, the mean score on the PHQ-9, did not show statistically significant between-group differences compared with the control group at either week 7 or month 3 (AI-iCBT vs control: least squares mean difference –0.47, 95% CI –1.13 to 0.18; *P*=.16; Cohen *d*=–0.10; iCBT vs control: least squares mean difference –0.62, 95% CI –1.28 to 0.04; *P*=.07; Cohen *d*=–0.13; Table 2). No significant differences were observed between the AI-iCBT and iCBT groups. Nevertheless, both intervention groups showed significant within-group reductions from baseline at week 7 and month 3 (all *P*<.001), indicating that depressive symptoms improved over time in both groups.

Table 2. Primary outcome: mean Patient Health Questionnaire-9 scores at baseline, week 7, and month 3 (intention-to-treat population). Values are least squares (LS) means with 95% CIs estimated using a mixed model for repeated measures. Between-group comparisons are shown.

Time point	AI-iCBT[a] group, LS mean (95% CI)	iCBT[b] group, LS mean (95% CI)	Control group, LS mean (95% CI)	AI-iCBT vs control, Δ (95% CI)	iCBT vs control, Δ (95% CI)
Baseline	8.79 (8.47 to 9.11)	8.73 (8.41 to 9.05)	8.81 (8.48 to 9.13)	—[c]	—
Week 7	7.90 (7.49 to 8.31)	7.75 (7.33 to 8.18)	8.28 (7.94 to 8.63)	–0.37 (–1.02 to 0.29); *P*=.27	–0.46 (–1.13 to 0.21); *P*=.18
Month 3	7.37 (6.98 to 7.77)	7.17 (6.76 to 7.58)	7.86 (7.50 to 8.22)	–0.47 (–1.13 to 0.18), *P*=.16	–0.62 (–1.28– to 0.04); *P*=.07

[a]AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy.

[b]iCBT: internet-based cognitive behavioral therapy.

[c]Not available.

For the secondary binary outcome of PHQ-9 ≥10, the overall proportion decreased over time across all groups (Figure 5). At month 3, the proportion was numerically lower in the AI-iCBT group compared with the control group, but between-group differences were not statistically significant in the ITT analysis (Multimedia Appendix 2).

Figure 5. Secondary outcome: proportion of participants with Patient Health Questionnaire-9 (PHQ-9) scores ≥10 at baseline, week 7, and month 3 (intention-to-treat population). Estimated proportions and 95% CIs were derived from generalized linear mixed models with a logit link, including effects for group, time, and their interaction. AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy; iCBT: internet-based cognitive behavioral therapy.

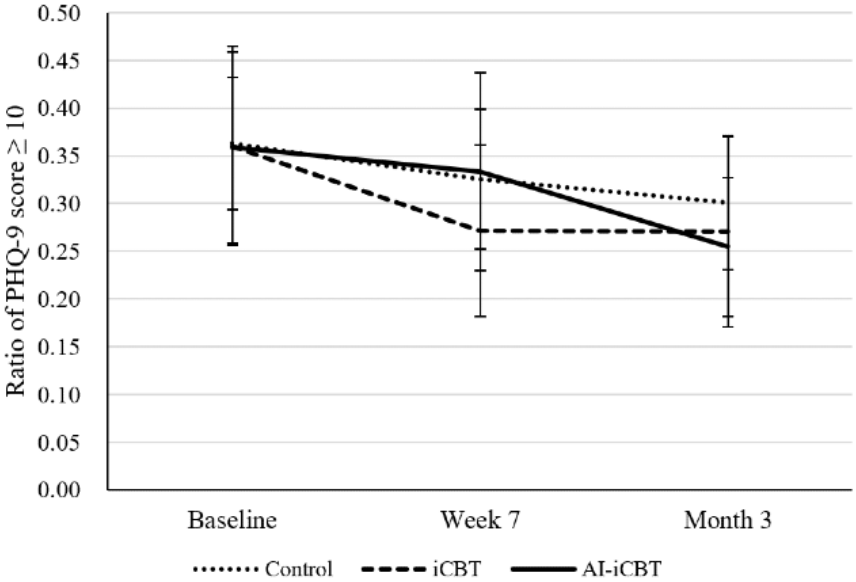

Similar patterns were observed for other secondary measures. QIDS-J and GAD-7 scores improved significantly within both intervention groups but without significant between-group differences. SDS scores showed modest reductions but did not significantly differ from control. Full secondary outcome results are provided in Multimedia Appendix 3.

Engagement and User Satisfaction

Overview

As illustrated in Figure 6, the CR exercise participation rate decreased significantly over time across both intervention groups (odds ratio [OR] 0.751, 95% CI 0.692-0.815; *P*<.001). Participation began at only about half of participants in week

XSL•FO
RenderX

1 and declined further, dropping more steeply in the iCBT group, which fell to around 30% by week 6. In contrast, the AI-iCBT group retained somewhat higher engagement, remaining closer to the low 40% range by week 6, suggesting that AI support helped sustain participation over time. Between-group differences over time were examined using GLMM with a logit link, which showed a significant group × week interaction favoring AI-iCBT (OR 1.23, 95% CI 1.09-1.39; *P*<.001; Table 3). Analyses included all randomized participants in the intervention groups.

Figure 6. Engagement outcome: weekly participation rates in the artificial intelligence–augmented internet-based cognitive behavioral therapy (AI-iCBT) and internet-based cognitive behavioral therapy (iCBT) groups during weeks 1-6 (intention-to-treat population). Participation was defined as completion of at least 1 cognitive restructuring exercise per week. Error bars indicate 95% CIs.

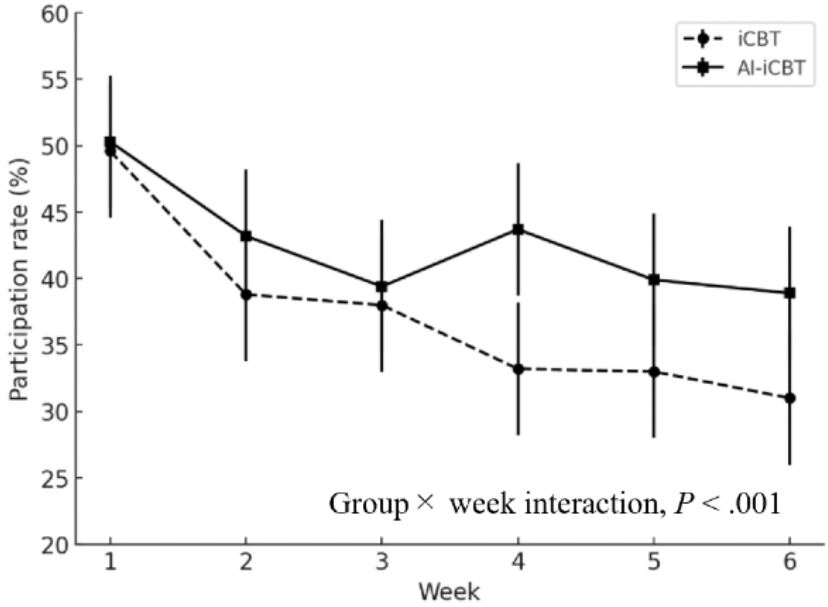

Table 3. Engagement outcome: mixed-effects logistic regression results for weekly cognitive restructuring exercise participation rates (intention-to-treat population). The generalized linear mixed model with a logit link included fixed effects for group, week (continuous and centered), and their interaction (group × week).

Effect	Reference	Odds ratio (95% CI)	*P* value
Group (AI-iCBT[a] vs iCBT[b])	iCBT	0.807 (0.370-1.758)	.59
Week (continuous, centered)	—[c]	0.751 (0.692-0.815)	<.001
Group × week	—	1.229 (1.090-1.386)	<.001

[a]AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy.

[b]iCBT: internet-based cognitive behavioral therapy.

[c]Not available.

User Satisfaction

Assessed at week 7 with the CSQ-8, averaged about 21 out of 32 points in both intervention groups, indicating a moderate to good level of satisfaction. No significant difference was observed between AI-iCBT and iCBT (Multimedia Appendix 4).

Exploratory Analysis of Engagement-Enhancing Factors

Activation of the empathy function was significantly associated with higher participation (OR 9.99, 95% CI 5.80-17.21; *P*<.001). In contrast, activation of the advisory function was not significantly associated with engagement (OR 2.37, 95% CI 0.96-5.83; *P*=.06). Detailed adjusted results are provided in Multimedia Appendix 5.

Exploratory EAS Analysis

In the exploratory EAS analysis, based on the ITT population, 312/396 (78.8%) in the AI-iCBT group, 312/397 (78.6%) in the iCBT group, and 317/394 (80.5%) in the control group had a baseline PHQ-9 score of ≥5. Among the ITT population, the proportion of participants who attended 3 or more sessions was 188/396 (47.5%) in the AI-iCBT group and 158/397 (39.8%) in the iCBT group. Furthermore, the proportion of participants with a baseline PHQ-9 score of ≥5 who attended 3 or more sessions (EAS) was 149/396 (37.6%) in the AI-iCBT group, 134/397 (33.8%) in the iCBT group, and 317/394 (80.5%) in the control group.

Mean PHQ-9 scores decreased significantly from baseline to week 7 in all 3 groups. At week 7, the iCBT group showed a statistically significant improvement compared with the control group (Δ=–1.08, 95% CI –1.98 to –0.18; *P*=.02). However, this

XSL•FO
RenderX

difference was not maintained at month 3. Full numerical results are provided in Multimedia Appendix 6. By contrast, the proportion of participants scoring ≥10 on the PHQ-9 decreased only in the iCBT group compared with the control at week 7 (Figure 7). At month 3, the AI-iCBT group showed a significantly lower proportion compared with control (Δ –0.15, 95% CI –0.30 to –0.01; *P*=.046), while the iCBT group did not differ significantly. The group × time interaction was significant (*P*=.008), indicating that the pattern of improvement differed between intervention groups (see Multimedia Appendix 7 for full numerical results).

Figure 7. Exploratory outcome: proportion of participants with Patient Health Questionnaire-9 (PHQ-9) scores of ≥10 in the efficacy analysis set population at baseline, week 7 (postintervention), and month 3 (follow-up). Error bars indicate 95% CIs. Asterisks represent *P*<.05 versus control. AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy; iCBT: internet-based cognitive behavioral therapy.

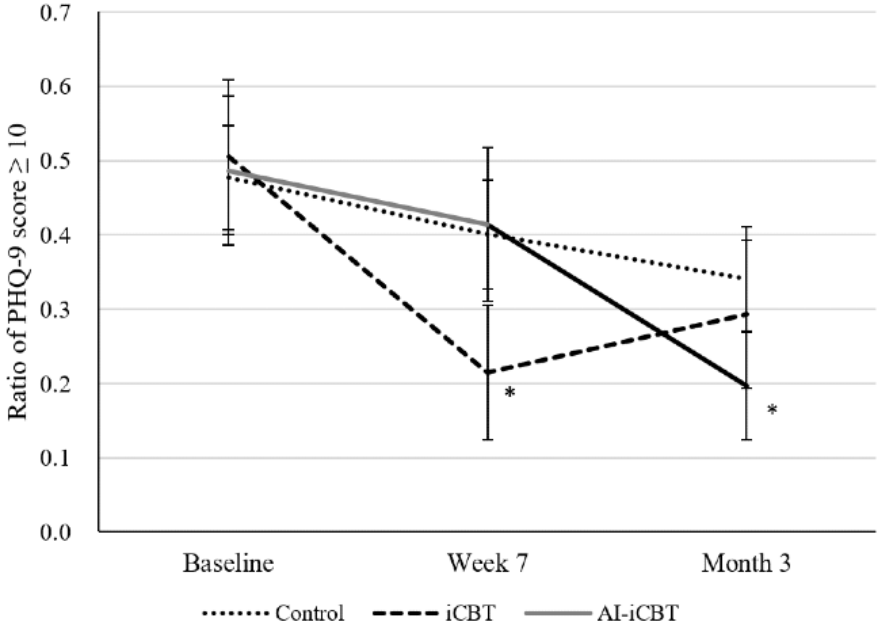

Sensitivity Analyses

In evaluating factors associated with low adherence, participants who attended <3 sessions were more likely to be male (65.1% vs 48.4%; *P*<.001), older (mean age 44.2, SD 9.44 vs mean age 41.7, SD 9.55 years; *P*=.001), and employed (287/341, 84.2% vs 212/283, 74.9%; *P*=.02), compared with those who attended ≥3 sessions. No significant differences were observed for marital status, educational background, medical history, or mental health history (Multimedia Appendix 8).

As a further sensitivity analysis, when applying a stricter definition of depression severity—PHQ-9 score of ≥10 plus at least 1 core symptom and SDS score of ≥10—62.7% (178/284) of participants with a PHQ-9 score of ≥10 at baseline met this definition (AI-iCBT: 44/70, 62.9%; iCBT: 41/71, 57.7%; control: 93/143, 65%), with no significant group differences (Multimedia Appendix 9).

Discussion

Overview

This study has several unique features. First, it directly compared a fully self-administered CR exercise delivered via iCBT, with and without AI-based NLP functionality, under a randomized (partially masked) design. Notably, the addition of AI led to a statistically significant improvement in engagement—an effect, to our knowledge, not previously documented. As AI-based interventions have become increasingly sophisticated and deeply integrated into intervention programs, disentangling the specific contribution of AI has become difficult. In particular, establishing control conditions that differ only in the presence or absence of AI functionality requires substantial resources, and prior studies have therefore often relied on heterogeneous control conditions. By applying a more robust design—feasible in part because the technology was still in a transitional phase—this trial provides new insights into how AI may enhance fully self-administered iCBT and offers a timely perspective for advancing scalable mental health care.

Self-administered interventions are known to be modestly to moderately effective for depression [15,23], but adherence remains a major limitation [15,18,19,21,55,63]. Systematic reviews indicate that approximately one-third to one-half of participants drop out before completing the program [19,64]. In this context, the engagement improvement observed in this study represents a potential step toward overcoming this barrier.

With regard to clinical effectiveness, no additional benefits of NLP feedback were observed in the ITT population. Recent evidence has reported that greater antidepressant effects are associated with lower dropout rates [15,20,65]. In contrast, although no between-group differences in antidepressant effects were found here, the addition of AI feedback was associated with a statistically significant increase in adherence. This suggests a novel engagement-enhancing mechanism, distinct from the traditionally assumed link between larger clinical effects and lower dropout rates.

Exploratory analyses further indicated that the “empathic function” of AI feedback was significantly associated with

XSL•FO
RenderX

improved adherence, whereas the advisory function showed no significant effect. Participants who received empathic responses during their first exercise subsequently demonstrated higher adherence. While self-disclosure was not directly measured, the sense of being supported may have facilitated persistence. These findings align with prior evidence that empathic conversational agents and chatbots support therapeutic alliance and sustained engagement [66-69]. Research in behavioral change has likewise shown that AI-mediated feedback can promote sustained self-management [70], supporting the plausibility of these findings. Such effects of human-AI interaction may have been less visible in prior studies using heterogeneous control conditions but became evident here through the structured randomized design.

Regarding antidepressant effects, no significant between-group differences were observed in the ITT population. In the EAS, results—while requiring cautious interpretation—indicated that at week 7 only the iCBT group showed significant improvement in both the mean PHQ-9 score and the proportion of participants with a PHQ-9 score of ≥10 compared with the control group, whereas the AI-iCBT group did not.

This suggests that the AI function may have attenuated or failed to enhance short-term antidepressant effects. However, this short-term benefit in the iCBT group disappeared at long-term follow-up. For the continuous outcome in particular, the short-term difference was –1.1 points on the PHQ-9, below the minimal clinically important difference (approximately 3 points) [35,71], suggesting limited clinical significance.

By contrast, the short-term dichotomous outcome in the iCBT group represented about a 29% reduction in the proportion of PHQ-9 scores of ≥10 cases relative to the control group. This implies that part of the potential benefit may not have been realized when AI was introduced. At long-term follow-up (month 3), however, only the AI-iCBT group showed a significant reduction of about 15% compared with the control group. These findings highlight the absence of the expected short-term effect in AI-iCBT and the unique long-term effect observed only in AI-iCBT.

The fact that AI-iCBT ultimately demonstrated an effect at long-term follow-up is noteworthy. Although exploratory, this suggests a potential contribution of AI-iCBT in reducing the proportion of participants exceeding a clinically significant severity threshold. One possible explanation is that approaches emphasizing empathy as a core therapeutic skill—such as interpersonal psychotherapy or family therapies—often show slower onset but more enduring gains compared with CBT, lasting well beyond the end of treatment [72-77]. It is possible that the empathy-related function of AI contributed in a similar way, although the underlying mechanisms remain unclear.

The EAS, however, was more restrictive than the ITT population. Among ITT participants with a baseline PHQ-9 score of ≥5, only 149/312 (47.8%) in the AI-iCBT group and 134/312 (42.9%) in the iCBT group attended at least 3 sessions. Furthermore, although a PHQ-9 score of ≥10 is widely recognized as a proxy for “major depression equivalent” in research [33,39], concerns have been raised that it may not be sufficient for diagnostic purposes and may contribute to overdiagnosis [56,78,79]. As a sensitivity analysis, therefore, we used a stricter definition requiring a PHQ-9 score of ≥10 plus at least 1 core symptom (depressed mood or anhedonia) [56,57], together with an SDS score of ≥10 as an indicator of functional impairment [49,58-60]. Results confirmed that only about 62.7% (178/284) of participants who met the PHQ-9 score of ≥10 at baseline also met this stricter definition, highlighting the importance of cautious interpretation (Multimedia Appendix 9).

Taken together, regardless of whether it corresponds to major depression, the finding of a significant reduction in the proportion of participants with clinically meaningful depressive states (PHQ-9 ≥10) compared with the control group may have clinical significance, particularly given the fully self-help nature of the program. From a public health perspective, such a difference could also carry implications for the scalability of self-help programs that do not require therapist involvement.

AI communication, including generative AI, continues to advance rapidly. However, the development of appropriate control programs has often been constrained by logistical and cost-related factors, limiting opportunities to rigorously investigate the antidepressant and anxiolytic effects of AI. Beyond psychiatry, maximizing the effectiveness of self-administered interventions while enhancing engagement remains a critical challenge across health care and welfare domains. This study represents a step toward addressing this challenge.

Limitations

First, most participants were recruited from a research panel with high affinity for digital interventions. Only 10.9% (129/1187) were actual users of mental health services, which is consistent with the national average in Japan, but caution is required in generalizing these findings to broader populations. Second, high dropout rates were observed, with only 37.6% (149/396) of the AI-iCBT group and 33.8% (134/397) of the iCBT group in the ITT population meeting EAS criteria. Additional analyses indicated that low adherence was associated with being male, older, and employed (Multimedia Appendix 8). Consistent with recent studies, time constraints [22], particularly among employed men [21] and older adults [20,21], were confirmed as key barriers to engagement. Third, although the AI-iCBT group consistently showed 5%-10% higher adherence rates than the iCBT group throughout the trial, both groups had already dropped by half to 50% (396/793) at the very first session and continued to decline over time, remaining at only 30%-40%.

Fourth, missing data were substantial, and MMRM and generalized estimating equations were applied to minimize bias. However, these approaches assume data are missing at random. In this study, since attrition occurred according to participant attributes, the possibility of missing not at random cannot be ruled out, and estimates may remain biased. Fifth, a significant group × time interaction was observed in adherence in the AI feedback group, suggesting a potential role of AI in improving engagement. However, this conclusion is based on a single trial and requires replication in different populations and designs, as well as further elucidation of the underlying mechanisms.

XSL•FO
RenderX

Sixth, due to technical issues, session-by-session data on depressive symptoms (Overall Anxiety Severity and Impairment Scale) [80] and anxiety (Overall Depression Severity and Impairment Scale) [81] were lost, precluding more detailed analyses. Future studies should implement automatic data saving and backup systems to prevent such loss. Overall antidepressant effects were small. In addition to limitations of the program itself, this may reflect a ceiling effect due to the predominance of participants with mild depression. Several meta-analyses [23,82,83] report that treatment effects may be less pronounced in cases with mild baseline depression compared to moderate or severe cases. Another limitation is related to our stratified randomization. While stratification by age, sex, and baseline PHQ-9 severity increased internal validity by balancing key prognostic factors, it may also restrict the generalizability of our findings to populations with different distributions of these characteristics.

Despite these challenges, this study provides valuable insights into the potential of AI-enhanced self-help digital interventions, particularly in relation to participant behavior dynamics. Importantly, no major adverse events were reported, underscoring the safety of this innovative approach and its potential to significantly advance mental health care practices.

Acknowledgments

We are grateful to Dr Yutaka Ono (Center for the Development of Cognitive Behavior Therapy Training, Tokyo), Professor Takashi Watanabe (Teikyo Heisei University, Tokyo), Mr Kenichiro Tsumura (T Quest, Chiba), Dr Sosei Yamaguchi (National Center of Neurology and Psychiatry, Tokyo), and all members of the project teams of NEC Solution Innovators, Ltd (Tokyo) for allowing us to unrestrictedly use artificial intelligence–augmented internet-based cognitive behavioral therapy for the study.

Funding

This study is funded by the Research Institute of Economy, Trade, and Industry (a think tank under the Ministry of Economy, Trade, and Industry of the Government of Japan). The funder had no role in study design, data collection, data analysis, data interpretation, or writing of the manuscript.

Data Availability

The data used in this study belong to the Research Institute of Economy, Trade and Industry and can be obtained from the institute upon reasonable request.

Authors' Contributions

MS contributed to conceptualization, investigation, methodology, project administration, resources, software, validation, visualization, and writing—including original draft preparation, review, and editing. YS contributed to conceptualization, funding acquisition, investigation, methodology, project administration, resources, and writing—review and editing. SH contributed to data curation, formal analysis, validation, visualization, and writing—review and editing. MK and HY contributed to secondary analysis and writing—review and editing. NW contributed to conceptualization, investigation, methodology, supervision, validation, and writing—review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist.
[PDF File (Adobe PDF File), 340 KB - jmir_v28i1e76902_app1.pdf]

Multimedia Appendix 2
Secondary outcome: ratio of participants with Patient Health Questionnaire-9 scores ≥10 (intention-to-treat population).
[DOCX File , 17 KB - jmir_v28i1e76902_app2.docx]

Multimedia Appendix 3
Secondary outcomes: mean score of Quick Inventory of Depressive Symptomatology-Japanese version, Generalized Anxiety Disorder-7, and Sheehan Disability Scale (intention-to-treat population).
[DOCX File , 19 KB - jmir_v28i1e76902_app3.docx]

Multimedia Appendix 4

XSL•FO
RenderX

Satisfaction outcome: Client Satisfaction Questionnaire-8 total scores at week 7 (intention-to-treat population, intervention groups only).
[DOCX File , 17 KB - jmir_v28i1e76902_app4.docx]

Multimedia Appendix 5
Associations between artificial intelligence feedback functions and exercise attendance (artificial intelligence–augmented internet-based cognitive behavioral therapy and internet-based cognitive behavioral therapy participants, weeks 2–6).
[DOCX File , 20 KB - jmir_v28i1e76902_app5.docx]

Multimedia Appendix 6
Exploratory outcome: mean Patient Health Questionnaire-9 scores (efficacy analysis set population).
[DOCX File , 18 KB - jmir_v28i1e76902_app6.docx]

Multimedia Appendix 7
Exploratory outcome: ratio of participants with Patient Health Questionnaire-9 scores ≥10 (efficacy analysis set population).
[DOCX File , 18 KB - jmir_v28i1e76902_app7.docx]

Multimedia Appendix 8
Associated factors of low adherence in the efficacy analysis set.
[DOCX File , 20 KB - jmir_v28i1e76902_app8.docx]

Multimedia Appendix 9
Baseline prevalence of major depression proxies in the efficacy analysis set population.
[DOCX File , 18 KB - jmir_v28i1e76902_app9.docx]

References

1. GBD 2017 DiseaseInjury IncidencePrevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the global burden of disease study 2017. Lancet 2018;392(10159):1789-1858 [FREE Full text] [doi: 10.1016/S0140-6736(18)32279-7] [Medline: 30496104]
2. Arias D, Saxena S, Verguet S. Quantifying the global burden of mental disorders and their economic value. EClinicalMedicine 2022;54:101675 [FREE Full text] [doi: 10.1016/j.eclinm.2022.101675] [Medline: 36193171]
3. Hohls JK, König HH, Quirke E, Hajek A. Anxiety, depression and quality of life-a systematic review of evidence from longitudinal observational studies. Int J Environ Res Public Health 2021;18(22):12022 [FREE Full text] [doi: 10.3390/ijerph182212022] [Medline: 34831779]
4. König H, König HH, Konnopka A. The excess costs of depression: a systematic review and meta-analysis. Epidemiol Psychiatr Sci 2019;29:e30 [FREE Full text] [doi: 10.1017/S2045796019000180] [Medline: 30947759]
5. WHO. Mental health at work. World Health Organization. Geneva; 2022. URL: https://www.who.int/news-room/fact-sheets/detail/mental-health-at-work [accessed 2025-12-26]
6. Cunningham PJ. Beyond parity: primary care physicians' perspectives on access to mental health care. Health Aff (Millwood) 2009;28(3):w490-w501. [doi: 10.1377/hlthaff.28.3.w490] [Medline: 19366722]
7. Wang PS, Aguilar-Gaxiola S, Alonso J, Angermeyer MC, Borges G, Bromet EJ, et al. Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. Lancet 2007;370(9590):841-850 [FREE Full text] [doi: 10.1016/S0140-6736(07)61414-7] [Medline: 17826169]
8. Wainberg ML, Lu FG, Riba MB. Global mental health. Acad Psychiatry 2016;40(4):647-649. [doi: 10.1007/s40596-016-0577-0] [Medline: 27259490]
9. Edge D, Watkins ER, Limond J, Mugadza J. The efficacy of self-guided internet and mobile-based interventions for preventing anxiety and depression: a systematic review and meta-analysis. Behav Res Ther 2023;164:104292 [FREE Full text] [doi: 10.1016/j.brat.2023.104292] [Medline: 37003138]
10. Psychological Interventions Implementation Manual: Integrating Evidence-Based Psychological Interventions into Existing Services. Geneva: World Health Organization; 2024.
11. WHO. Consolidated telemedicine implementation guide. World Health Organization. 2022. URL: https://www.who.int/publications/i/item/9789240059184 [accessed 2025-12-26]
12. Binder P, Hjeltnes A. Mindfulness in psychotherapy and society—the need for combining enthusiasm and critical inquiry. Couns and Psychother Res 2021;21(2):247-250 [FREE Full text] [doi: 10.1002/capr.12384]

XSL•FO
RenderX

13. Knowles SE, Lovell K, Bower P, Gilbody S, Littlewood E, Lester H. Patient experience of computerised therapy for depression in primary care. BMJ Open 2015;5(11):e008581 [FREE Full text] [doi: 10.1136/bmjopen-2015-008581] [Medline: 26621513]
14. Andersson G, Carlbring P, Rozental A. Response and remission rates in internet-based cognitive behavior therapy: an individual patient data meta-analysis. Front Psychiatry 2019;10:749. [doi: 10.3389/fpsyt.2019.00749] [Medline: 31708813]
15. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FMG, Furukawa TA, Cuijpers P, Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021;78(4):361-371 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
16. Saad A, Bruno D, Camara B, D'Agostino J, Bolea-Alamanac B. Self-directed technology-based therapeutic methods for adult patients receiving mental health services: systematic review. JMIR Ment Health 2021;8(11):e27404 [FREE Full text] [doi: 10.2196/27404] [Medline: 34842556]
17. Cuijpers P, Noma H, Karyotaki E, Cipriani A, Furukawa TA. Effectiveness and acceptability of cognitive behavior therapy delivery formats in adults with depression: a network meta-analysis. JAMA Psychiatry 2019;76(7):700-707 [FREE Full text] [doi: 10.1001/jamapsychiatry.2019.0268] [Medline: 30994877]
18. Seittu HA, Falk T, Bhatnagar K, Saarni SE. Therapists' role in patient adherence to internet-based cognitive behavioral therapy: qualitative study. J Med Internet Res 2025;27:e71852 [FREE Full text] [doi: 10.2196/71852] [Medline: 40929730]
19. Treanor CJ, Kouvonen A, Lallukka T, Donnelly M. Acceptability of computerized cognitive behavioral therapy for adults: umbrella review. JMIR Ment Health 2021;8(7):e23091 [FREE Full text] [doi: 10.2196/23091] [Medline: 34255714]
20. Fuhr K, Schröder J, Berger T, Moritz S, Meyer B, Lutz W, et al. The association between adherence and outcome in an internet intervention for depression. J Affect Disord 2018;229:443-449. [doi: 10.1016/j.jad.2017.12.028] [Medline: 29331706]
21. Karyotaki E, Kleiboer A, Smit F, Turner DT, Pastor AM, Andersson G, et al. Predictors of treatment dropout in self-guided web-based interventions for depression: an 'individual patient data' meta-analysis. Psychol Med 2015;45(13):2717-2726 [FREE Full text] [doi: 10.1017/S0033291715000665] [Medline: 25881626]
22. Beatty L, Binnion C. A systematic review of predictors of, and reasons for, adherence to online psychological interventions. Int J Behav Med 2016;23(6):776-794. [doi: 10.1007/s12529-016-9556-9] [Medline: 26957109]
23. Karyotaki E, Riper H, Twisk J, Hoogendoorn A, Kleiboer A, Mira A, et al. Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms: a meta-analysis of individual participant data. JAMA Psychiatry 2017;74(4):351-359 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.0044] [Medline: 28241179]
24. Kemmeren L, van Schaik A, Draisma S, Kleiboer A, Riper H, Smit J. Effectiveness of blended cognitive behavioral therapy versus treatment as usual for depression in routine specialized mental healthcare: E-COMPARED trial in the Netherlands. Cogn Ther Res 2023;47(3):386-398. [doi: 10.1007/s10608-023-10363-y]
25. Mathiasen K, Andersen TE, Lichtenstein MB, Ehlers LH, Riper H, Kleiboer A, et al. The clinical effectiveness of blended cognitive behavioral therapy compared with face-to-face cognitive behavioral therapy for adult depression: randomized controlled noninferiority trial. J Med Internet Res 2022;24(9):e36577 [FREE Full text] [doi: 10.2196/36577] [Medline: 36069798]
26. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. Annu Rev Clin Psychol 2018;14:91-118. [doi: 10.1146/annurev-clinpsy-032816-045037] [Medline: 29401044]
27. Nie J, Shao H, Fan Y, Shao Q, You H, Preindl M, et al. LLM-based conversational AI therapist for daily functioning screening and psychotherapeutic intervention via everyday smart devices. arXiv. Preprint posted online on March 16, 2024 2025 [FREE Full text] [doi: 10.1145/3712299]
28. Malgaroli M, Hull TD, Zech JM, Althoff T. Natural language processing for mental health interventions: a systematic review and research framework. Transl Psychiatry 2023;13(1):309 [FREE Full text] [doi: 10.1038/s41398-023-02592-2] [Medline: 37798296]
29. Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res 2021;23(5):e15708. [doi: 10.2196/15708] [Medline: 33944788]
30. Sheehan K, Bhatti PK, Yousuf S, Rosenow W, Roehler DR, Hazekamp C, et al. Natural language processing applied to mental illness detection: a narrative review. NPJ Digit Med 2022;5(1):46 [FREE Full text] [doi: 10.1186/s12889-022-13177-x] [Medline: 35449046]
31. Choudhury MD, Pendse SR, Kumar N. Benefits and harms of large language models in digital mental health. arXiv. Preprint posted online on November 7, 2021 2021 [FREE Full text]
32. Villarreal-Zegarra D, Reategui-Rivera CM, García-Serna J, Quispe-Callo G, Lázaro-Cruz G, Centeno-Terrazas G, et al. Self-administered interventions based on natural language processing models for reducing depressive and anxiety symptoms: systematic review and meta-analysis. JMIR Ment Health 2024;11:e59560 [FREE Full text] [doi: 10.2196/59560] [Medline: 39167795]
33. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]

XSL•FO
RenderX

34. Muramatsu K, Miyaoka H, Kamijima K, Muramatsu Y, Tanaka Y, Hosaka M, et al. Performance of the Japanese version of the Patient Health Questionnaire-9 (J-PHQ-9) for depression in primary care. Gen Hosp Psychiatry 2018;52:64-69. [doi: 10.1016/j.genhosppsych.2018.03.007] [Medline: 29698880]
35. Bauer-Staeb C, Kounali DZ, Welton NJ, Griffith E, Wiles NJ, Lewis G, et al. Effective dose 50 method as the minimal clinically important difference: evidence from depression trials. J Clin Epidemiol 2021;137:200-208 [FREE Full text] [doi: 10.1016/j.jclinepi.2021.04.002] [Medline: 33892086]
36. Kounali D, Button KS, Lewis G, Gilbody S, Kessler D, Araya R, et al. How much change is enough? Evidence from a longitudinal study on depression in UK primary care. Psychol Med 2022;52(10):1875-1882 [FREE Full text] [doi: 10.1017/S0033291720003700] [Medline: 33138872]
37. So M, Sekizawa Y, Yamaguchi Y. A randomised controlled trial investigating the clinical and cost-effectiveness of peer enhanced-computerised cognitive depression. KAKEN. 2015. URL: https://kaken.nii.ac.jp/en/grant/KAKENHI-PROJECT-26350862/ [accessed 2025-12-19]
38. Inagaki M, Ohtsuki T, Yonemoto N, Kawashima Y, Saitoh A, Oikawa Y, et al. Validity of the Patient Health Questionnaire (PHQ)-9 and PHQ-2 in general internal medicine primary care at a Japanese rural hospital: a cross-sectional study. Gen Hosp Psychiatry 2013;35(6):592-597. [doi: 10.1016/j.genhosppsych.2013.08.001] [Medline: 24029431]
39. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. CMAJ 2012;184(3):E191-E196 [FREE Full text] [doi: 10.1503/cmaj.110829] [Medline: 22184363]
40. Cuijpers P, Karyotaki E, Ciharova M, Miguel C, Noma H, Furukawa TA. The effects of psychotherapies for depression on response, remission, reliable change, and deterioration: a meta-analysis. Acta Psychiatr Scand 2021;144(3):288-299. [doi: 10.1111/acps.13335] [Medline: 34107050]
41. Fava M. Depression with physical symptoms: treating to remission. J Clin Psychiatry 2003;64 Suppl 7:24-28. [Medline: 12755649]
42. Keller MB, Lavori PW, Mueller TI, Endicott J, Coryell W, Hirschfeld RM, et al. Time to recovery, chronicity, and levels of psychopathology in major depression. A 5-year prospective follow-up of 431 subjects. Arch Gen Psychiatry 1992;49(10):809-816. [doi: 10.1001/archpsyc.1992.01820100053010] [Medline: 1417434]
43. Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, et al. The 16-item Quick Inventory of Depressive Symptomatology (QIDS), Clinician Rating (QIDS-C), and Self-Report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry 2003;54(5):573-583. [doi: 10.1016/s0006-3223(02)01866-8] [Medline: 12946886]
44. Fujisawa D, Nakagawa A, Tajima M, Ono Y. Development of Japanese version of QIDS-SR (self-report). Jpn J Stress Sci 2010;25(1):43-52 [FREE Full text]
45. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
46. Muramatsu K, Muramatsu Y, Miyaoka H, Fuse K, Yoshimine F, Hosaka M. Validation and utility of a Japanese version of the GAD-7. In: Panminerva Medica. 2009 Presented at: 20th World Congress on Psychosomatic Medicine; August 23-28, 2009; Torino, Italy p. 79.
47. Sheehan KH, Sheehan DV. Assessing treatment effects in clinical trials with the discan metric of the Sheehan Disability Scale. Int Clin Psychopharmacol 2008;23(2):70-83. [doi: 10.1097/YIC.0b013e3282f2b4d6] [Medline: 18301121]
48. Yoshida T, Otsubo T, Tsuchida H, Wada R, Ueshima K, Fukui A. Reliability and validity of the Japanese version of the Sheehan Disability Scale. Jpn J Clin Psychopharmacol 2004;7(10):1645-1653.
49. Soares CN, Zhang M, Boucher M. Categorical improvement in functional impairment in depressed patients treated with desvenlafaxine. CNS Spectr 2019;24(3):322-332 [FREE Full text] [doi: 10.1017/S1092852917000633] [Medline: 29140227]
50. Attkisson CC, Zwick R. The Client Satisfaction Questionnaire. Psychometric properties and correlations with service utilization and psychotherapy outcome. Eval Program Plann 1982;5(3):233-237. [doi: 10.1016/0149-7189(82)90074-x] [Medline: 10259963]
51. Tachimori H, Ito H. Reliability and validity of the Japanese version of the Client Satisfaction Questionnaire. Seishin Igaku (Clin Psychiatry) 1999;41(7):711-717. [doi: 10.11477/mf.1405905056]
52. Wong SYS, Sun YY, Chan ATY, Leung MKW, Chao DVK, Li CCK, et al. Treating subthreshold depression in primary care: a randomized controlled trial of behavioral activation with mindfulness. Ann Fam Med 2018;16(2):111-119 [FREE Full text] [doi: 10.1370/afm.2206] [Medline: 29531101]
53. Harrer M, Sprenger AA, Illing S, Adriaanse MC, Albert SM, Allart E, et al. Psychological intervention in individuals with subthreshold depression: individual participant data meta-analysis of treatment effects and moderators. Br J Psychiatry 2025:1-14 [FREE Full text] [doi: 10.1192/bjp.2025.56] [Medline: 40365980]
54. Donkin L, Hickie IB, Christensen H, Naismith SL, Neal B, Cockayne NL, et al. Rethinking the dose-response relationship between usage and outcome in an online intervention for depression: randomized controlled trial. J Med Internet Res 2013;15(10):e231 [FREE Full text] [doi: 10.2196/jmir.2771] [Medline: 24135213]
55. Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N. A systematic review of the impact of adherence on the effectiveness of e-therapies. J Med Internet Res 2011;13(3):e52 [FREE Full text] [doi: 10.2196/jmir.1772] [Medline: 21821503]

XSL•FO
RenderX

56. Levis B, Benedetti A, Thombs BD, DEPRESSion Screening Data (DEPRESSD) Collaboration. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. BMJ 2019;365:l1476 [FREE Full text] [doi: 10.1136/bmj.l1476] [Medline: 30967483]
57. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatric Annals 2002;32(9):509-515. [doi: 10.3928/0048-5713-20020901-06]
58. Lam RW, Michalak EE, Yatham LN. A new Clinical Rating Scale for work absence and productivity: validation in patients with major depressive disorder. BMC Psychiatry 2009;9:78 [FREE Full text] [doi: 10.1186/1471-244X-9-78] [Medline: 19958540]
59. Luciano JV, Bertsch J, Salvador-Carulla L, Tomás JM, Fernández A, Pinto-Meza A, et al. Factor structure, internal consistency and construct validity of the Sheehan Disability Scale in a Spanish primary care sample. J Eval Clin Pract 2010;16(5):895-901. [doi: 10.1111/j.1365-2753.2009.01211.x] [Medline: 20626541]
60. Sheehan DV, Harnett-Sheehan K, Raj BA. The measurement of disability. Int Clin Psychopharmacol 1996;11 Suppl 3:89-95. [doi: 10.1097/00004850-199606003-00015] [Medline: 8923116]
61. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c332 [FREE Full text] [doi: 10.1136/bmj.c332] [Medline: 20332509]
62. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
63. Eysenbach G. The law of attrition. J Med Internet Res 2005;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
64. Koelen JA, Vonk A, Klein A, de Koning L, Vonk P, de Vet S, et al. Man vs. machine: a meta-analysis on the added value of human support in text-based internet treatments ("e-therapy") for mental disorders. Clin Psychol Rev 2022;96:102179 [FREE Full text] [doi: 10.1016/j.cpr.2022.102179] [Medline: 35763975]
65. Kambeitz-Ilankovic L, Rzayeva U, Völkel L, Wenzel J, Weiske J, Jessen F, et al. A systematic review of digital and face-to-face cognitive behavioral therapy for depression. NPJ Digit Med 2022;5(1):144 [FREE Full text] [doi: 10.1038/s41746-022-00677-8] [Medline: 36109583]
66. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. Front Digit Health 2022;4:847991 [FREE Full text] [doi: 10.3389/fdgth.2022.847991] [Medline: 35480848]
67. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
68. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth 2018;6(11):e12106 [FREE Full text] [doi: 10.2196/12106] [Medline: 30470676]
69. Bickmore T, Gruber A, Picard R. Establishing the computer-patient working alliance in automated health behavior change interventions. Patient Educ Couns 2005;59(1):21-30. [doi: 10.1016/j.pec.2004.09.008] [Medline: 16198215]
70. Prochaska JJ, Vogel EA, Chieng A, Kendra M, Baiocchi M, Pajarito S, et al. A therapeutic relational agent for reducing problematic substance use (woebot): development and usability study. J Med Internet Res 2021;23(3):e24850 [FREE Full text] [doi: 10.2196/24850] [Medline: 33755028]
71. Lynch CP, Cha EDK, Jenkins NW, Parrish JM, Mohan S, Jadczak CN, et al. The minimum clinically important difference for Patient Health Questionnaire-9 in minimally invasive transforaminal interbody fusion. Spine (Phila Pa 1976) 2021;46(9):603-609. [doi: 10.1097/BRS.0000000000003853] [Medline: 33290370]
72. Fairburn CG, Norman PA, Welch SL, O'Connor ME, Doll HA, Peveler RC. A prospective study of outcome in bulimia nervosa and the long-term effects of three psychological treatments. Arch Gen Psychiatry 1995;52(4):304-312. [doi: 10.1001/archpsyc.1995.03950160054010] [Medline: 7702447]
73. Agras WS, Walsh T, Fairburn CG, Wilson GT, Kraemer HC. A multicenter comparison of cognitive-behavioral therapy and interpersonal psychotherapy for bulimia nervosa. Arch Gen Psychiatry 2000;57(5):459-466. [doi: 10.1001/archpsyc.57.5.459] [Medline: 10807486]
74. Carter FA, Jordan J, McIntosh VVW, Luty SE, McKenzie JM, Frampton CMA, et al. The long-term efficacy of three psychotherapies for anorexia nervosa: a randomized, controlled trial. Int J Eat Disord 2011;44(7):647-654. [doi: 10.1002/eat.20879] [Medline: 21997429]
75. Markowitz JC, Petkova E, Neria Y, Van Meter PE, Zhao Y, Hembree E, et al. Is exposure necessary? A randomized clinical trial of interpersonal psychotherapy for PTSD. Am J Psychiatry 2015;172(5):430-440 [FREE Full text] [doi: 10.1176/appi.ajp.2014.14070908] [Medline: 25677355]
76. Bighelli I, Rodolico A, García-Mieres H, Pitschel-Walz G, Hansen WP, Schneider-Thoma J, et al. Psychosocial and psychological interventions for relapse prevention in schizophrenia: a systematic review and network meta-analysis. Lancet Psychiatry 2021;8(11):969-980. [doi: 10.1016/j.envres.2021.112166] [Medline: 34619129]

XSL•FO
RenderX

77. Lemmens LHJM, van Bronswijk SC, Peeters FPML, Arntz A, Roefs A, Hollon SD, et al. Interpersonal psychotherapy versus cognitive therapy for depression: how they work, how long, and for whom-key findings from an RCT. Am J Psychother 2020;73(1):8-14. [doi: 10.1176/appi.psychotherapy.20190030] [Medline: 32122161]
78. Mitchell AJ, Yadegarfar M, Gill J, Stubbs B. Case finding and screening clinical utility of the Patient Health Questionnaire (PHQ-9 and PHQ-2) for depression in primary care: a diagnostic meta-analysis of 40 studies. BJPsych Open 2016;2(2):127-138 [FREE Full text] [doi: 10.1192/bjpo.bp.115.001685] [Medline: 27703765]
79. Levis B, Bhandari PM, Neupane D, Fan S, Sun Y, He C, Depression Screening Data (DEPRESSD) PHQ Group. Data-driven cutoff selection for the Patient Health Questionnaire-9 depression screening tool. JAMA Netw Open 2024;7(11):e2429630 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.29630] [Medline: 39576645]
80. Bentley KH, Gallagher MW, Carl JR, Barlow DH. Development and validation of the Overall Depression Severity and Impairment Scale. Psychol Assess 2014;26(3):815-830. [doi: 10.1037/a0036216] [Medline: 24708078]
81. Norman SB, Cissell SH, Means-Christensen AJ, Stein MB. Development and validation of an Overall Anxiety Severity and Impairment Scale (OASIS). Depress Anxiety 2006;23(4):245-249. [doi: 10.1002/da.20182] [Medline: 16688739]
82. Cuijpers P, Cristea IA, Karyotaki E, Reijnders M, Huibers MJH. How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. World Psychiatry 2016;15(3):245-258 [FREE Full text] [doi: 10.1002/wps.20346] [Medline: 27717254]
83. Mercorio A, Zizolfi B, Barbuto S, Danzi R, Di Spiezio Sardo A, Moawad G, et al. Three-dimensional imaging reconstruction and laparoscopic robotic surgery: a winning combination for a complex case of multiple myomectomy. Fertil Steril 2023;120(1):202-204 [FREE Full text] [doi: 10.1016/j.fertnstert.2023.04.015] [Medline: 37085096]

Abbreviations

AI: artificial intelligence
AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy
CBT: cognitive behavioral therapy
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
CR: cognitive restructuring
CSQ-8: Client Satisfaction Questionnaire-8
EAS: efficacy analysis set
GAD-7: Generalized Anxiety Disorder-7
GLMM: generalized linear mixed models
iCBT: internet-based cognitive behavioral therapy
ITT: intention-to-treat
MMRM: mixed-effects model for repeated measures
NLP: natural language processing
OR: odds ratio
PHQ-9: Patient Health Questionnaire-9
QIDS-J: Quick Inventory of Depressive Symptomatology-Japanese version
SDS: Sheehan Disability Scale

Edited by A Schwartz; submitted 03.May.2025; peer-reviewed by N Titov; comments to author 02.Jun.2025; accepted 21.Nov.2025; published 05.Jan.2026.

Please cite as:
So M, Sekizawa Y, Hashimoto S, Kashimura M, Yamakage H, Watanabe N
Effect of AI-Based Natural Language Feedback on Engagement and Clinical Outcomes in Fully Self-Guided Internet-Based Cognitive Behavioral Therapy for Depression: 3-Arm Randomized Controlled Trial
J Med Internet Res 2026;28:e76902
URL: https://www.jmir.org/2026/1/e76902
doi:10.2196/76902
PMID:

©Mirai So, Yoichi Sekizawa, Sora Hashimoto, Masami Kashimura, Hajime Yamakage, Norio Watanabe. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted

XSL•FO
RenderX

use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Leisure Screen Time, Internet Gaming Disorder, and Mental Health Among Chinese Adolescents: Large-Scale Cross-Sectional Study

Qin Deng[1], MPH; Linna Sha[1], MPH; Jiaojiao Hou[2], MPH; Xunying Zhao[2], MD; Rong Xiang[1], MD; Jiangbo Zhu[1], MPH; Yang Qu[2], MD; Jinyu Zhou[1], MD; Ting Yu[2], MD; Xin Song[1], MD; Sirui Zheng[1], MPH; Tao Han[1], MPH; Bin Yang[1], MD; Mengyu Fan[2*], MD; Xia Jiang[1,2,3*], MD, PhD

[1]Department of Nutrition and Food Hygiene, West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu, China

[2]Department of Epidemiology and Biostatistics, West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu, China

[3]Department of Clinical Neuroscience, Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden

[*]these authors contributed equally

Corresponding Author:
Xia Jiang, MD, PhD
Department of Clinical Neuroscience, Center for Molecular Medicine
Karolinska Institutet
L8:05, CMM Karolinska Universitetssjukhuset Solna
Stockholm, 17176
Sweden
Phone: 86 15528103968
Email: xia.jiang@ki.se

Abstract

Background: Adolescence is a critical period for mental health vulnerability alongside rising digital media exposure. Current evidence often fails to distinguish the distinct roles of leisure screen time (LST) quantity and addictive patterns like internet gaming disorder (IGD) on a comprehensive range of mental health outcomes.

Objective: This study aimed to investigate the independent and joint associations of LST and IGD with multiple mental health conditions among Chinese adolescents.

Methods: We conducted a school-based, cross-sectional survey in Sichuan Province, China. Participants were recruited by random cluster sampling from 20 public schools. The sample comprised 13,240 adolescents (6659/13,240, 50.3% girls) with a mean age of 15.4 (SD 1.6) years. LST was self-reported, and IGD was evaluated using the Internet Gaming Disorder Scale-9 Item Short Form (IGDS9-SF). Mental health outcomes included overall mental health status and 5 specific diseases: psychological distress, depression, paranoia, insomnia, and suicidal ideation, all assessed using validated scales.

Results: The prevalence of excessive LST, IGD, and any mental health disorder was 48.2% (6378/13,240; 95% CI 47.3%-49.0%), 1.4% (188/13,240; 95% CI 1.2%-1.6%), and 55.8% (7387/13,240; 95% CI 54.9%-56.7%), respectively. After adjustment, excessive LST (odds ratio [OR] 1.18, 95% CI 1.09-1.27) and IGD (OR 6.58, 95% CI 5.02-8.62) were independently associated with poor mental health. A dose-response relationship existed for LST quartiles (Q2: OR 1.15, 95% CI 1.04-1.26; Q3: OR 1.24, 95% CI 1.12-1.37; Q4: OR 1.31, 95% CI 1.18-1.46; P_{trend}<.001). Excessive LST was associated with depression (OR 1.16, 95% CIs 1.05-1.29), paranoia (OR 1.22, 95% CI 1.11-1.34), and suicidal ideation (OR 1.15, 95% CI 1.04-1.28), while IGD was associated with all 5 disorders, most notably depression (OR 6.43, 95% CI 4.56-9.06) and paranoia (OR 5.77, 95% CI 4.05-8.21). IGD consistently demonstrated stronger associations than LST: psychological distress (OR 4.40, 95% CI 3.12-6.19 vs OR 1.14, 95% CI 0.98-1.33), depression (OR 6.43, 95% CI 4.56-9.06 vs OR 1.16, 95% CI 1.05-1.29), paranoia (OR 5.77, 95% CI 4.05-8.21 vs OR 1.22, 95% CI 1.11-1.34), insomnia (OR 2.90, 95% CI 2.09-4.05 vs OR 1.12, 95% CI 102-1.22), and suicidal ideation (OR 3.85, 95% CI 2.76-5.37 vs OR 1.15, 95% CI 1.04-1.28). Adolescents with both excessive LST and IGD demonstrated the highest odds of mental health disorders (OR 7.35, 95% CI 5.29-10.22). No significant interaction was found on additive or multiplicative scales.

Conclusions: Both excessive LST and IGD are independently associated with mental health disorders in adolescents, with IGD showing a substantially stronger association. This study is distinct from prior research by simultaneously investigating both screen

XSL•FO
RenderX

time quantity and addictive usage patterns, and by comprehensively assessing 5 distinct mental health outcomes. Longitudinal studies are needed to better understand the long-term effects.

(J Med Internet Res 2026;28:e80737) doi:10.2196/80737

KEYWORDS

adolescent; cross-sectional study; internet gaming disorder; leisure screen time; mental health; screen media activity

Introduction

Adolescence represents a critical period for psychological development, as up to half of all mental health conditions start before age 14 years [1,2]. Globally, 1 in 7 adolescents experience at least one mental health disorder, contributing to 15% of the overall disease burden in this age group [3]. Depression and anxiety are among the leading causes of illness and disability, with suicide being the fourth leading cause of death for those aged 15-19 years [3]. Additionally, psychological distress [4], paranoia [5], and insomnia [6] are also commonly prevalent. Poor mental health at a relatively early stage poses significant adverse effects that can span from the short to the long term [7], including school disengagement, reduced quality of life, and increased mortality. The widespread underrecognition and treatment delays—where only 15.4% of adolescents seek prompt professional help after their initial request—further exacerbate these issues [8].

Alongside the rise in mental health disorders over the years, time spent on screen-based activities continues to increase by as much as 2 hours per day since 2010, particularly among adolescents [9,10]. Data from the Adolescent Brain Cognitive Development Study reveal that adolescents now spend an average of over 5.5 hours daily on noneducational screen media, underscoring a global trend of pervasive digital engagement [11]. While moderate screen use may offer cognitive benefits such as enhanced multitasking [12], excessive use is linked to sedentary behaviors, impaired parent–child interactions, and behavioral addictions [13,14]. Each additional hour of daily screen time in late childhood predicted increased depressive symptoms in early adolescence, an effect mediated by shorter sleep duration and changes in white matter organization in brain regions responsible for emotion regulation [15]. Current research often conflates educational and leisure screen time (LST), obscuring the unique risk of LST [16,17]. A dose-response study of Chinese adolescents revealed that LST exceeding 1 hour/day linearly increases mental health problems, with video-based and gaming content showing the strongest negative associations [18]. Nevertheless, existing studies have mainly investigated depression and anxiety, the 2 most prevalent mental health issues, while largely disregarding other important disorders such as paranoia and insomnia that affect a nontrivial proportion (20%-30%) of adolescents [5,17].

The amount of leisure time spent using screen-based media does not necessarily represent the extent of harm. Crucially, recent research underscores that addictive use patterns—characterized by compulsive use, loss of control, and continued use despite negative consequences—may represent a far greater risk than screen time duration alone [14]. Studies found that a small number of individuals, despite showing shorter screen time, exhibited characteristics of addictive behaviors [19]. A primary concern for adolescents is gaming addiction, particularly the internet gaming disorder (IGD), defined by *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]) as the persistent and recurrent use of internet to play games, often with others. IGD leads to clinically significant impairment or distress and has been listed as a tentative disorder in need of further study [20]. IGD has been linked to distinct neurobiological changes and impairments in executive function, underscoring its status as a behavioral addiction beyond mere excessive use [21]. Interestingly, a substantial proportion of individuals with long screen time do not meet the symptomatic criteria of IGD, while those who meet criteria do not necessarily show long screen time [22,23]. This dissociation underscores the need to examine both behaviors simultaneously.

The interplay between LST and IGD in relation to a broad spectrum of mental health outcomes remains inadequately explored. While longitudinal evidence suggests that internalizing symptoms (eg, depression, anxiety) fully mediate the link between problematic internet use and subsequent self-harm behaviors, this model did not account for the joint influence of general screen time [24]. Furthermore, seminal research has demonstrated that trajectories of addictive digital media use are independently associated with heightened risks of suicidality, regardless of baseline screen time [14]. Moreover, while high LST and IGD independently poses negative effects on mental health, it is likely that those who exhibit both behaviors are at the highest risk. However, few studies have analyzed leisure and addictive screen-based media use by taking into consideration their independent and joint effects on mental health outcomes.

This cross-sectional study aims to comprehensively investigate the independent and joint associations of LST and IGD with an array of mental health among Chinese adolescents, including overall mental health and its 5 major diseases, namely, depression, psychological distress, paranoia, insomnia, and suicidal ideation.

Methods

Ethical Considerations

The research was conducted in accordance with the Declaration of Helsinki and received approval from the Ethics Committee of West China Fourth Hospital and West China School of Public Health, Sichuan University (Gw112023133; June 13, 2023). Written informed consent was obtained from parents or guardians after they received a detailed information sheet outlining the study purpose, procedures, and voluntary nature of participation. Student assent was also secured at the time of the survey. No compensation was offered to participants for

XSL•FO
RenderX

their involvement in this study. Beyond a lack of financial compensation, the study also prioritized participant anonymity by eliminating all personally identifiable information from the collected data.

Additionally, this study did not collect any personally identifiable information, such as facial images, voice recordings, or other biometric data. All data presented in this manuscript and supplementary materials are fully anonymized, ensuring that no individual participant can be identified.

Study Population

This cross-sectional study analyzed data from a survey project on the mental health of children and adolescents in Pidu District, Chengdu, conducted from June to December 2023 by the West China School of Public Health and the West China Fourth Hospital, Sichuan University. We adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines in reporting cross-sectional studies [25].

The survey aimed to gain a comprehensive understanding of the mental health status of children and adolescents residing in Pidu District and to effectively address the mental health issues affecting this population. To ensure the representativeness of study samples, we used a random cluster sampling design based on the number and proportion of primary and secondary students in Pidu District, considering types of school and their urban–rural distribution. A total of 20 schools were selected, comprising 17,973 registered students for the questionnaire survey.

Data collection was conducted in classroom settings during regular school hours. Trained research staff administered structured, paper-based questionnaires to all participants. To ensure standardized procedures, staff delivered standardized verbal instructions to students before distributing the questionnaires, detailing the study's purpose, its voluntary nature, confidentiality protocols, and questionnaire completion guidelines. Participants then completed the surveys independently. Research staff supervised the process and were available to provide neutral clarification on questions but refrained from influencing participant responses. Of the 16,482 participants invited, a total of 16,325 were present on the survey day and completed the questionnaire, resulting in a response rate of 98.99%. Participant eligibility was restricted to Han Chinese adolescents aged 12-18 years. Ethnicity was self-identified in the questionnaire, and age was calculated from the self-reported date of birth and ensured to be within the specified range at the time of data collection. After applying the eligibility criteria, a total of 13,240 participants were included in the final analytical sample. The high response rate minimized concerns about selection bias.

Assessment of LST and IGD

LST was measured using a self-reported questionnaire. Participants were asked to recall and report the average time they spent on leisure screen activities (including entertainment-based use of smartphones, computers, tablets, and televisions) on a typical day. The average daily LST was calculated using the formula: ([Weekday LST × 5] + [Weekend LST × 2])/7. In this study, LST was categorized into 2 groups according to World Health Organization standards, low (<2 h/day) and high (≥2 h/day) [26], and further into 4 groups based on quartiles, Q1 (≤0.93 h/day), Q2 (0.93 <LST ≤1.86 h/day), Q3 (1.86 <LST ≤4.19 h/day), and Q4 (>4.19 h/day).

IGD was assessed using the Internet Gaming Disorder Scale-9 Item Short Form (IGDS9-SF) [27], consisting of 9 items reflecting the diagnostic criteria for IGD as defined by the American Psychological Association. Participants responded to each item through a Likert scale ranging from 1 ("never") to 5 ("very often"). Total scores, ranging from 9 to 45, were calculated by summing across the responses. This threshold was identified as clinically optimal for the Chinese population in a previous validation study [28], which, although conducted in a sample with a mean age of 20 years, remains one of the most established cutoffs for the IGDS9-SF in the absence of a universally validated threshold for early- to mid-adolescents. We selected this cutoff to ensure consistency with prior research and because of its high specificity, which minimizes false positives. Furthermore, the scale demonstrated adequate internal consistency in our adolescent sample (Cronbach α=0.88), supporting its reliability for this age group.

To identify the joint association of the 2 types of screen-based behaviors, we defined a combined pattern based on the level of LST (low vs high) and status of IGD (absence vs presence). We designated the group with LST of less than 2 hours and without IGD as the reference group, referred to as "Low No." The other 3 comparison groups were "High No," "Low Yes," and "High Yes."

Assessment of Mental Health Disorders

We applied well-established validated scales to investigate the 5 common mental health disorders among adolescents. Specifically, psychological distress was measured using the Kessler 6-item (K6) scale, which assesses the frequency of psychopathological symptoms or behaviors. A K6 score of ≥13 indicates psychological distress [29,30]. Depression was evaluated using the Kutcher Adolescent Depression 6-item (KADS-6) scale, specifically designed for children and adolescents to effectively identify major depressive episodes. A KADS-6 score of ≥6 indicates depression [31]. Paranoia was assessed using the psychoticism and paranoid ideation subscales of the Symptom Checklist-90 (SCL-90), with a score of ≥12 indicating paranoid ideation [32]. Insomnia was measured using the Insomnia Severity Index (ISI) scale, with a score of ≥8 indicating sleep disturbances [33]. Suicidal ideation was assessed using the Suicide Behaviors Questionnaire-Revised (SBQ-R) scale, with a score of ≥8 reflecting risk of suicidal behaviors [34]. Adolescents exhibiting any (or more) of these 5 symptoms were considered to have mental health disorders. The internal consistency of all scales used in this study appeared to be robust, with Cronbach α values being 0.88 (K6), 0.85 (KADS-6), 0.81 (SCL-90), 0.88 (ISI), and 0.84 (SBQ-R), respectively.

Assessment of Covariates

This analysis included a comprehensive set of variables that may potentially confound the association between screen-based

XSL•FO
RenderX

behaviors and adolescent mental health (Table S1 in Multimedia Appendix 1). Demographic variables included age, sex, area of residence, economic status, single-child status, type of caretakers, and parental educational level. The area of residence was classified as "urban" or "rural." Economic status was categorized into "high-income" or "low- to middle-income." Single-child status was determined based on the presence of siblings ("yes" or "no"). Caretaker type was classified as "both parents and others" or "one parent." Parental educational level was defined by the highest education attained: "junior school and below," "high school or vocational," and "college and above." Anthropometric measures included BMI (kg/m^2), calculated from height and weight. Lifestyle factors included smoking status, drinking status, academic performance, physical health status, dietary habits, and physical activity (PA) level. Smoking and drinking were self-reported as "yes" or "no." Academic performance and physical health status were categorized as "low," "moderate," or "high." Dietary habits—including breakfast, vegetables and fruits, protein consumption, sugary beverages, desserts, and fried food—were divided into "never eat," "1-2 times per week," "3-4 times per week," "5-6 times per week," and "eat every day." PA levels were categorized based on daily durations into 4 quartiles: Q1 (≤1.80 hours), Q2 (1.81-4.80 hours), Q3 (4.81-9.50 hours), and Q4 (>9.50 hours).

Statistical Analysis

The missing data pattern was evaluated using the Little's Missing Completely at Random test. A nonsignificant result (χ^2_{8836}=4638.4; P>.05) confirmed that the data were missing completely at random, thereby justifying the application of multiple imputation. Consequently, the Multiple Imputation by Chained Equations approach was used to handle the missing values, given the low overall missing rate (maximum: 5.7%) [35]. We generated m=5 imputed datasets using predictive mean matching with a maximum of maxit=50 iterations, considered sufficient to achieve stability for datasets with low missing rates. Convergence was assessed by ensuring the algorithm completed the specified iterations without issues. To validate the imputation, we conducted analyses using both raw and imputed datasets, considering results statistically significant only if they were consistent across both methods. Descriptive summaries were presented as mean (SD) for continuous variables and as quantity (proportion) for categorical variables. Differences in continuous and categorical variables between groups (individuals with or without mental health disorders) were assessed using *t* tests (2-tailed) and chi-square tests, as appropriate.

To assess multicollinearity between LST and IGD, we used Pearson correlation coefficients (*r*) with values greater than 0.5 and variance inflation factor values exceeding 10 as diagnostic tests. The variance inflation factor value was found to be less than 1.02, indicating that multicollinearity was of minimal concern.

Multivariable logistic regression was used to examine the associations between screen-based behaviors and mental health among adolescents, controlling for confounders. First, we assessed the independent association of LST (high vs low) and IGD (presence vs absence) with overall mental health condition, making mutual adjustment for each other. Meanwhile, we also categorized LST into quartiles based on its duration (Q1-Q4) and used the Cochran-Armitage test to evaluate its dose-response effect [36]. Second, we investigated the joint effects and interactions of LST and IGD on overall mental health condition, adopting the Cochran-Armitage test to evaluate a linear trend. Finally, subgroup analyses were conducted by sex, area, economic status, single-child status, and type of caretakers. Interaction between screen-based behaviors and stratification factors was performed using the likelihood ratio test comparing models with and without a cross-product term [37]. Such a comprehensive analytical framework was further applied in parallel to each specific mental health disorder. In addition, to account for the cluster sampling design (participants nested within 20 schools), we used multilevel mixed-effects models with random intercepts for schools as sensitivity analyses.

Statistical analysis were performed using R (version 4.3.3; The R Core Team). Association estimates were presented in the form of odds ratios (ORs) and their 95% CIs. To minimize type I error, Bonferroni correction was applied, with a 2-sided *P* value <.05/5 considered statistically significant for the 5 mental health disorders. Trend tests and interaction *P* values were defined as statistically significant at *P*<.05.

Results

Participant Characteristics

The main characteristics of participants according to the status of mental health disorders are presented in Table S1 in Multimedia Appendix 1. This study included 13,240 adolescents (6659/13,240, 50.3% girls) with a mean age of 15.4 (SD 1.6) years. Among these adolescents, 60% (7947/13,240; 95% CI 59.1%-60.9%) lived in the rural areas, and a majority (12,264/13,240, 92.5%; 95% CI 92.0%-93.0%) were from low- to middle-income families. More than half of the participants (8543/13,240, 64%; 95% CI 63.2%-64.8%) had no siblings, and a significant proportion of parents (55.3%-58.9%) received low levels of education.

Regarding the outcomes, a total of 7391 adolescents (55.8%; 95% CI 54.9%-56.7%) were identified as having at least 1 of the 5 mental health disorders. Insomnia was the most prevalent, affecting 35.3% (4677/13,240; 95% CI 34.5%-36.1%) of participants, followed by paranoia (3670/13,240, 28.4%; 95% CI 27.6%-29.2%), suicidal ideation (3155/13,240, 23.8%; 95% CI 23.0%-24.6%), depression (2733/13,240, 20.6%; 95% CI 19.9%-21.3%), and psychological distress (1091/13,240, 8.2%; 95% CI 7.7%-8.7%).

Regarding the exposures, a total of 6378 adolescents (48.2%; 95% CI 47.3%-49.0%) had LST exceeding 2 hours per day, and 188 individuals (1.4%; 95% CI 1.2%-1.6%) met the criteria for IGD. Notably, among those with IGD (n=188), a total of 58 individuals (30.9%; 95% CI 24.3%-38.1%) had LST of 2 hours or less.

The prevalence of any mental health condition was higher among females, those with single parents or other caretakers, those with low academic performance, smokers, drinkers, those

XSL•FO
RenderX

with poor physical health, those who rarely consumed breakfast, fruits, and vegetables or protein, as well as those who frequently consumed unhealthy foods (sugary beverages, sweets, and fried food) (P<.05 for all).

Independent Association of LST and IGD on Mental Health Conditions

Results on the associations of each specific screen-based behavior with mental health disorders are shown in Table 1. After adjusting for potential confounders, LST >2 hours/day (OR 1.18, 95% CI 1.09-1.27) and the presence of IGD (OR 6.58, 95% CI 5.02-8.62) were significantly and independently associated with poorer mental health status. In addition, the odds of overall mental health issues showed a marked increase as LST increased (Q2: OR 1.15, 95% CI 1.04-1.26; Q3: OR 1.24, 95% CI 1.12-1.37; Q4: OR 1.31, 95% CI 1.18-1.46; P_{trend}<.001) (Table S2 in Multimedia Appendix 1).

Table 1. Independent associations of leisure screen time and internet gaming disorder with mental health disorders among Chinese adolescents (P<.05/5).

Mental health outcome	Leisure screen time (h/day)			Internet gaming disorder		
	<2	≥2		No	Yes	
	OR[a,b]	OR (95% CI)	*P* value	OR[c]	OR (95% CI)	*P* value
Overall mental health						
Model 1[d]	Ref	1.56 (1.45-1.67)	<.001	Ref	11.24 (8.69-14.55)	<.001
Model 2[e]	Ref	1.19 (1.11-1.28)	<.001	Ref	6.68 (5.10-8.75)	<.001
Model 3[f]	Ref	1.18 (1.09-1.27)	<.001	Ref	6.58 (5.02-8.62)	<.001
Psychological distress						
Model 1	Ref	1.61 (1.40-1.85)	<.001	Ref	7.86 (5.77-10.71)	<.001
Model 2	Ref	1.17 (1.01-1.37)	.04	Ref	4.47 (3.18-6.29)	<.001
Model 3	Ref	1.14 (0.98-1.33)	.09	Ref	4.40 (3.12-6.19)	<.001
Depression						
Model 1	Ref	1.57 (1.43-1.73)	<.001	Ref	9.98 (7.29-13.66)	<.001
Model 2	Ref	1.19 (1.07-1.32)	.002	Ref	6.52 (4.63-9.18)	<.001
Model 3	Ref	1.16 (1.05-1.29)	.005	Ref	6.43 (4.56-9.06)	<.001
Paranoia						
Model 1	Ref	1.52 (1.39-1.65)	<.001	Ref	8.38 (5.99-11.72)	<.001
Model 2	Ref	1.24 (1.13-1.36)	<.001	Ref	5.88 (4.13-8.36)	<.001
Model 3	Ref	1.22 (1.11-1.34)	<.001	Ref	5.77 (4.05-8.21)	<.001
Insomnia						
Model 1	Ref	1.41 (1.30-1.53)	<.001	Ref	4.34 (3.18-5.93)	<.001
Model 2	Ref	1.13 (1.03-1.23)	.01	Ref	2.94 (2.11-4.09)	<.001
Model 3	Ref	1.12 (1.02-1.22)	.01	Ref	2.90 (2.09-4.05)	<.001
Suicidal ideation						
Model 1	Ref	1.52 (1.39-1.67)	<.001	Ref	6.33 (4.67-8.58)	<.001
Model 2	Ref	1.17 (1.06-1.29)	.003	Ref	3.90 (2.80-5.45)	<.001
Model 3	Ref	1.15 (1.04-1.28)	.01	Ref	3.85 (2.76-5.37)	<.001

[a]OR: odds ratio.

[b]The reference group for leisure screen time was <2 h/day.

[c]The reference group for internet gaming disorder was "No."

[d]Model 1 (partially adjusted model) was adjusted for age, sex, area, and economic status.

[e]Model 2 (fully adjusted model) was adjusted for age, sex, area, economic status, single child, caretaker, father's educational level, mother's educational level, BMI, smoking, drinking, academic performance, health status, dietary factors, and physical activity level.

[f]Model 3 adjusted for the covariates included in Model 2 and additionally included mutual adjustment for leisure screen time and internet gaming disorder.

For specific disorders, excessive LST was associated with significantly higher odds of depression (OR 1.16, 95% CI 1.05-1.29), paranoia (OR 1.22, 95% CI 1.11-1.34), and suicidal ideation (OR 1.15, 95% CI 1.04-1.28). Similarly, IGD was

XSL•FO
RenderX

significantly associated with psychological distress (OR 4.40, 95% CI 3.12-6.19), depression (OR 6.43, 95% CI 4.56-9.06), paranoia (OR 5.77, 95% CI 4.05-8.21), insomnia (OR 2.90, 95% CI 2.09-4.05), and suicidal ideation (OR 3.85, 95% CI 2.76-5.37).

In subgroup analysis, no factor was found to significantly modify the associations between LST and any mental health condition, nor between IGD and any mental health condition (all $P_{interaction}$>.05) (Figure 1 and Tables S7 and S8 in Multimedia Appendix 1).

Figure 1. Subgroup analysis of the associations of screen-based behaviors with mental health disorders in Chinese adolescents. (A) Association between leisure screen time (LST; h/day) and mental health; (B) association between internet gaming disorder (IGD) and mental health.

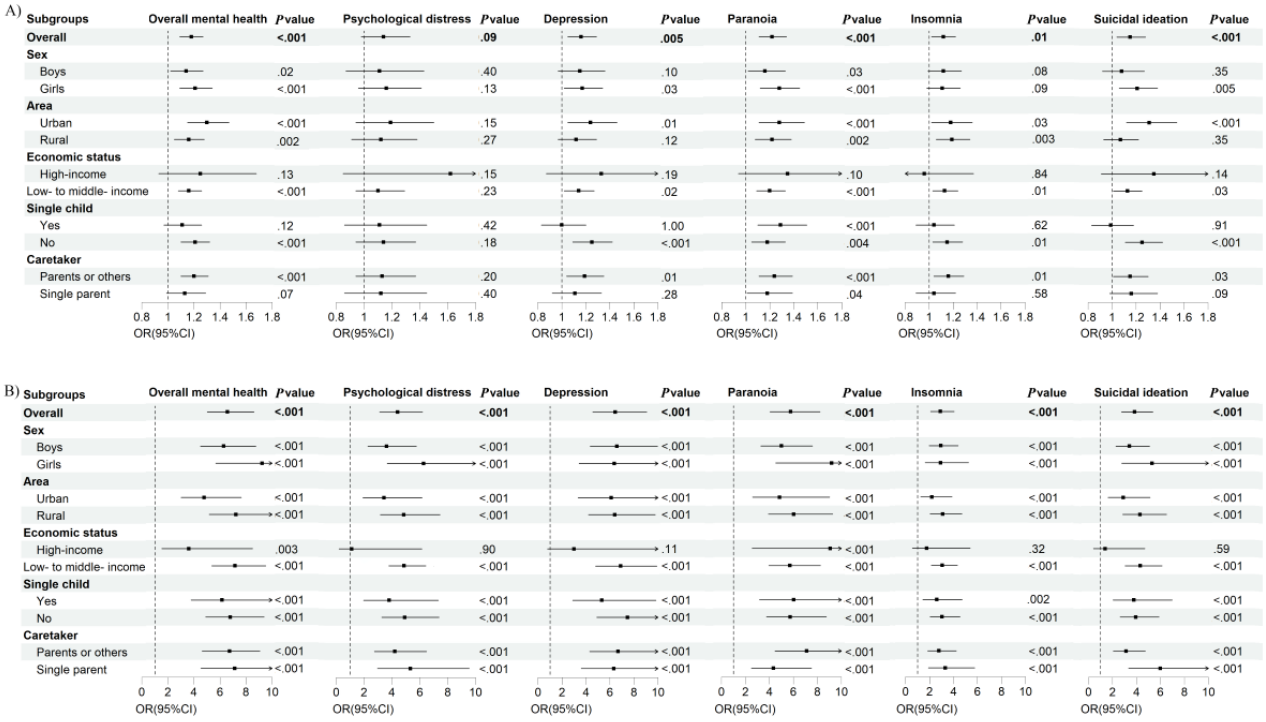

Joint Associations of LST and IGD on Mental Health Conditions

We further examined the joint associations of LST and IGD with the outcomes, as presented in Table 2 and Figure 2. The odds of overall mental health conditions were significantly higher among adolescents who exhibited at least one type of screen-based behavior (High No: OR 1.18, 95% CI 1.09-1.27; Low Yes: OR 7.34, 95% CI 4.59-11.74). Notably, the highest odds were observed when both excessive LST and symptoms of IGD were present simultaneously (High Yes: OR 7.35, 95% CI 5.29-10.22). However, no significant interaction between LST and IGD, either on the additive scale (relative excess risk due to interaction=–0.77, 95% CI –22.05 to 17.02; attributable proportion=–0.01, 95% CI –2.86 to 0.69; synergy index=0.99, 95% CI 0.23-4.26) or on the multiplicative scale (OR 0.88, 95% CI 0.23-3.37), was observed (Table 3). A similar pattern of results was observed with each of the 5 individual mental health outcomes.

XSL•FO
RenderX

Table 2. Combined effect of leisure screen time (LST) and internet gaming disorder (IGD) on mental health (*P*<.05/5). Odds ratios (ORs) of the joint associations were not directly comparable to those of the independent associations as shown in as they were derived from different models with different reference groups and adjustment strategies.

Mental health outcome	LST_IGD						
	Low No	High No		Low Yes		High Yes	
	OR[a]	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Overall mental health	Ref	1.18 (1.09-1.27)	<.001	7.34 (4.59-11.74)	<.001	7.35 (5.29-10.22)	<.001
Psychological distress	Ref	1.14 (0.98-1.34)	.09	4.94 (2.63-9.28)	<.001	4.81 (3.19-7.27)	<.001
Depression	Ref	1.17 (1.05-1.30)	.004	7.78 (4.32-14.00)	<.001	6.80 (4.46-10.37)	<.001
Paranoia	Ref	1.23 (1.12-1.35)	<.001	8.96 (4.70-17.11)	<.001	5.74 (3.77-8.75)	<.001
Insomnia	Ref	1.12 (1.02-1.22)	.01	2.71 (1.53-4.77)	<.001	3.36 (2.23-5.07)	<.001
Suicidal ideation	Ref	1.15 (1.04-1.28)	.006	3.81 (2.14-6.79)	<.001	4.46 (2.96-6.73)	<.001

[a]Adjusted for age, sex, area, economic status, single child, caretaker, father's educational level, mother's educational level, BMI, smoking, drinking, academic performance, health status, dietary factor and physical activity level.

XSL•FO
RenderX

Figure 2. Combined effect of leisure screen time (LST) and internet gaming disorder (IGD) on mental health disorders among Chinese adolescents.

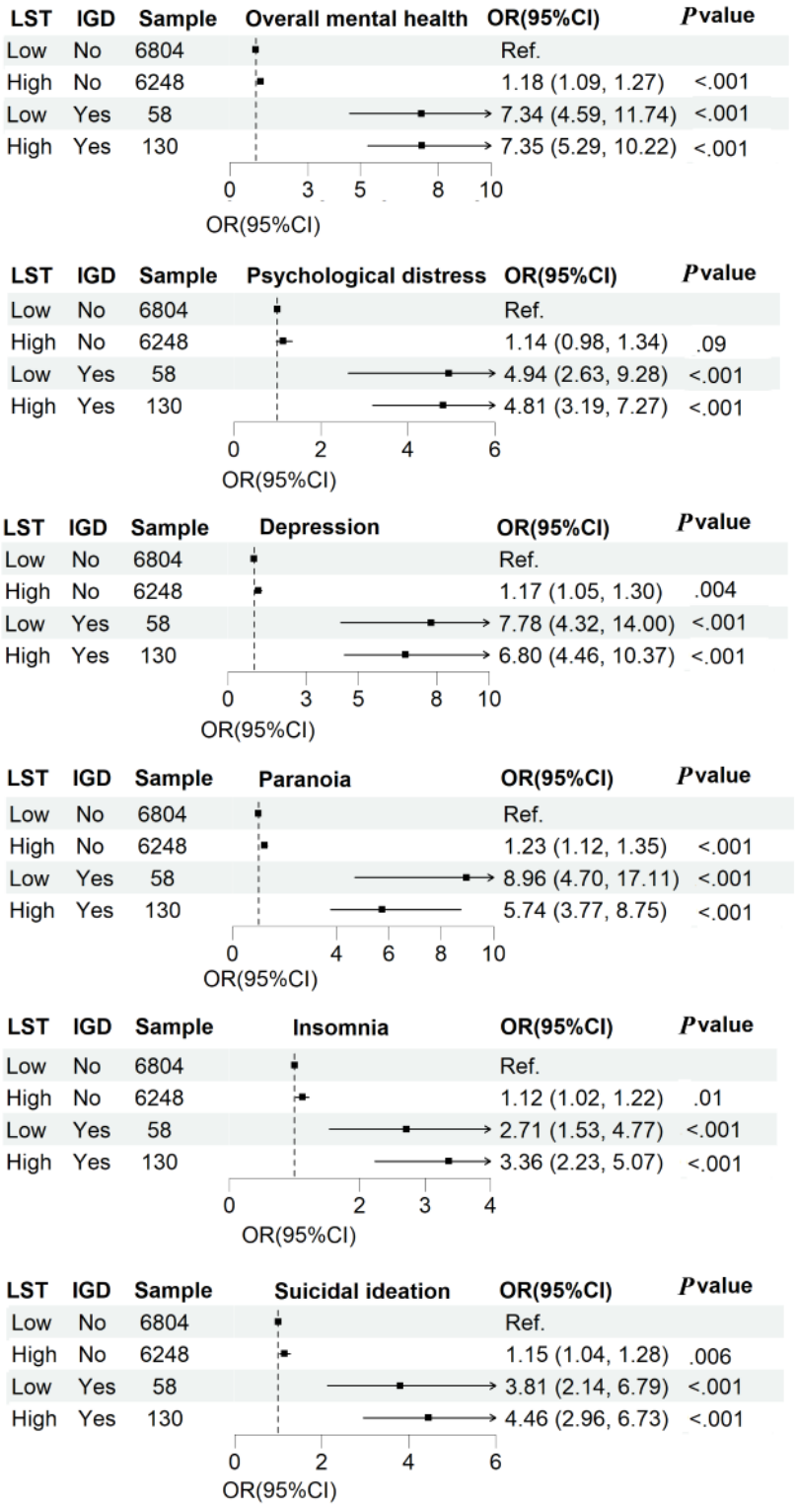

XSL•FO
RenderX

Table 3. Interaction analysis between leisure screen time and internet gaming disorder on mental health disorders among Chinese adolescents.

Mental health outcome	Additive interaction			Multiplicative interaction
	RERI[a] (95% CI)	AP[b] (95% CI)	S[c] (95% CI)	OR[d,e] (95% CI)
Overall mental health	–0.77 (–22.05 to 17.02)	–0.01 (–2.86 to 0.69)	0.99 (0.23 to 4.26)	0.88 (0.23 to 3.37)
Psychological distress	–0.27 (–4.83 to 3.00)	–0.06 (–1.20 to 0.47)	0.93 (0.38 to 2.30)	0.85 (0.40 to 1.79)
Depression	–1.14 (–7.73 to 3.75)	–0.17 (–1.38 to 0.41)	0.84 (0.37 to 1.88)	0.75 (0.36 to 1.54)
Paranoia	–3.45 (–11.80 to 1.73)	–0.60 (–2.43 to 0.22)	0.58 (0.24 to 1.37)	0.52 (0.24 to 1.12)
Insomnia	0.54 (–1.80 to 2.58)	0.16 (–0.69 to 0.54)	1.30 (0.47 to 3.58)	1.11 (0.55 to 2.23)
Suicidal ideation	0.50 (–2.81 to 3.28)	0.11 (–0.79 to 0.53)	1.17 (0.47 to 2.88)	1.17 (0.47 to 2.88)

[a]RERI: relative excess risk due to interaction.

[b]AP: attributable proportion due to interaction.

[c]S: synergy index.

[d]OR: odds ratio.

[e]Adjusted for age, sex, area, economic status, single child, caretaker, father's educational level, mother's educational level, BMI, smoking, drinking, academic performance, health status, dietary factor and physical activity level.

Repeating all analyses using raw data, we obtained consistent findings (Tables S4-S6 and S9-S12 in Multimedia Appendix 1). Sensitivity analyses using multilevel models accounting for school clustering yielded results consistent with primary findings, with all significant associations remaining unchanged. Intraclass correlation coefficients (0.76%-2.08%) confirmed that most variance (>98%) occurred at the individual level, supporting the robustness of results (Tables S18 and S19 in Multimedia Appendix 1).

Discussion

Principal Findings

In this large-scale cross-sectional study of more than 10,000 Chinese adolescents, 48.2% (6378/13,240) of the study population reported excessive LST, and 1.4% (188/13,240) met the criteria for IGD. Both excessive LST (>2 h/day) and IGD were independently associated with increased odds of overall mental health disorders and its 5 major conditions: psychological distress, depression, paranoia, insomnia, and suicidal ideation. The association was markedly stronger for IGD than for LST. The observed dose-response relationship across LST quartiles, with progressively higher ORs across LST quartiles, provides compelling evidence for a graded association in which increasing screen time exposure corresponds to incrementally worse mental health outcomes. Furthermore, while a significant joint association was observed—where adolescents exhibiting both behaviors demonstrated a 7-fold rise in the probability of mental health disorders—no statistically significant interaction was found, indicating an additive rather than synergistic effect. The precision of the main effect estimates further strengthens the robustness of our findings.

These findings necessitate an expansion of current public health guidelines, which predominantly target screen time duration, to include specific strategies for early identification and management of IGD. This study makes a critical contribution by clearly establishing IGD as a severe and distinct factor, independent of mere usage quantity. Methodologically, the innovative joint-association analysis revealing their additive relationship provides a more sophisticated framework for assessing digital media behaviors on mental health.

Prevalence of LST and IGD Among Adolescents

Approximately 48.2% (6378/13,240) of adolescents in this study reported LST >2 hours/day, surpassing the rate reported in a recent study conducted in Guangzhou, China (20.7%) [38]. The prevalence of IGD in this study was only 1.4% (188/13,240), much lower than the global rate of 8.8% [39]. Such discrepancy may be attributed to several factors. First, it likely reflects our use of a conservative, high-specificity cutoff score (≥32 on IGDS9-SF) to minimize false positives, whereas many epidemiological studies use lower, more sensitive thresholds that yield higher prevalence rates. Second, our sample was drawn from a general school-based population in a specific sociocultural context (Pidu District, Chengdu), which may exhibit different characteristics compared to the more diverse and potentially higher-risk samples included in the global meta-analyses.

Despite the lower prevalence, the large overall sample size (N=13,240) ensured that this study was sufficiently powered to detect strong and statistically significant associations, as evidenced by the strength and consistency of the association between IGD and poorer mental health. It is also important to note that the large population size of China means that even a 1.4% prevalence translates to a substantial absolute number of adolescents affected by IGD, underscoring the importance for policymakers and guardians to closely monitor screen-based behaviors among adolescents. Of note, we found that a large proportion of adolescents meeting the symptomatic criteria for IGD (58/188, 30.9%) reported LST of less than 2 hours/day; conversely, a substantial proportion of participants who did not meet the criteria for IGD (6248/13,052, 47.9%) reported excessive LST. These findings further illustrate that spending a large amount of time on gaming or using media does not necessarily equal to addictive use, and vice versa.

Adolescents living with a single parent or without siblings were more likely to engage in excessive LST. This may stem from reduced supervision in single-parent households or absence of

XSL•FO
RenderX

siblings that limit the opportunities for outdoor or other non–screen-based activities, thereby contributing to excessive LST. In addition, boys were more likely than girls to exhibit symptoms of IGD. Boys typically prefer to participate in active and combative activities, whereas girls tend to engage more in social and conversational activities. These preferences may explain why boys are more often drawn to the virtual worlds of online games [40].

The Association of LST and IGD With Mental Health Conditions

Our findings indicated that excessive LST and IGD, both independently and jointly, were associated with mental health disorders. Prior studies showed that greater time spent on screen media was correlated with a higher prevalence of multiple mental health disorders, such as depression and psychological distress [41-43]. This study extends the existing body of work by providing new evidence for additional mental health outcomes, including paranoia and suicidal ideation in adolescents.

One possible explanation is that screen-based activities displace time for other beneficial activities, such as PA [44], a well-established factor protecting against mental health disorders [45,46]. Indeed, previous studies have demonstrated that PA can help reduce excessive screen-based behaviors and improve mental health. Since PA and screen-based behaviors often co-develop during adolescence [47], fostering healthy lifestyles is crucial for both the current and the future well-being of this population. Another key factor is the inherent harm posed by screen-based behaviors. Beyond educational content, the internet and media are flooded with violent, graphic, and pornographic materials, which further exacerbate mental health [48].

Adolescents with both high levels of LST and symptoms of IGD had significantly higher odds of mental health issues compared to those with excessive LST or IGD alone. These behaviors may reinforce each other: high LST could increase the likelihood of developing IGD, while IGD, in turn, may encourage further screen time, creating a feedback loop that exacerbates mental health. Despite such co-occurrence, the absence of significant interactions, neither additive nor multiplicative, suggests that the associations of LST and IGD with mental health are independent rather than synergistic.

This conclusion is supported by the nonsignificant interaction, indicating a lack of synergy where the joint presence does not create a disproportionately greater association. In other words, while the combination of high LST and IGD was associated with a higher likelihood of mental health problems, their interaction did not statistically amplify or mitigate the effect of each behavior. Notably, the group with IGD but low LST (“Low Yes”) exhibited substantially higher odds of mental health disorders than the group with high LST but no IGD (“High No”). This indicates that while both behaviors warrant attention, IGD constitutes a disproportionately strong associated factor. Therefore, future research and public health interventions should address both factors, prioritizing the identification and management of IGD within a framework that includes multidimensional assessments of both screen time and gaming disorder symptoms.

Although no significant modifier was found to influence the associations between LST and mental health disorders, our findings suggested that the association was somewhat more pronounced in girls. Consistent with previous studies, girls tend to spend more time on screen-based media than boys [49,50] and report poorer mental health outcomes [51]. Subgroup analyses revealed distinct patterns of vulnerability: urban residents, adolescents from higher-income families, nonsingle children, and those living with parents or caregivers showed stronger associations between excessive LST and mental health problems. In contrast, the link between IGD and mental health issues was more pronounced among girls, nonsingle children, rural adolescents, individuals from low- to middle-income families, and those in single-parent households. According to a report released by the China Internet Network Information Center, while internet penetration rates among Chinese adolescents are comparable across urban and rural areas, differences emerge in the use of specific online applications [52,53]. Moreover, previous research has shown that parenting style and family function play a pivotal role in the development of screen-based media dependency [54].

Evidence also indicates that parental absence can contribute to depression in children [55], potentially due to disruptions in child–parent relationships resulting from parental migration [56] and reduced communication [57]. These factors may influence the impact of LST and IGD on mental health, as variations in family dynamics and parenting styles affect how screen-based behaviors influence emotional and psychological outcomes. For instance, supportive family environments may mitigate the negative effects of excessive screen time while dysfunctional family interactions could exacerbate the associated negative outcomes.

Limitations

Regarding the limitations, first, due to the cross-sectional design of this study, it is not possible to establish causal relationships between exposure and outcomes. Second, although random cluster sampling was used, our participants were exclusively Han Chinese adolescents from Pidu District, Chengdu. While this sample captured urban–rural diversity within the district, it may not fully represent the broader socioeconomic, cultural, and regional heterogeneity across China. Therefore, caution is warranted when extrapolating our findings to inform national-level policies, and they should be interpreted as preliminary evidence primarily relevant to similar regional contexts. Third, a key limitation is that all data, including parameters for LST, IGD, and mental health symptoms, were based on self-report. This may introduce reporting bias. To address these limitations, future research could consider using longitudinal and interventional studies to validate the detrimental effects of screen-based behaviors on adolescent mental health. Expanding the research scope to include adolescents from multiple regions might enhance the representativeness and generalizability of findings. Additionally, using objective measurement tools, such as wearable devices or monitoring instruments, for screen time, alongside clinical interviews for diagnosing IGD and mental health disorders, would greatly improve the objectivity and accuracy of the assessments.

XSL•FO
RenderX

Conclusions

This large-scale population-based study provides novel evidence that LST and IGD represent distinct dimensions of digital engagement with independent and additive associations with mental health problems in Chinese adolescents. Methodologically, this study advances the field by simultaneously examining both screen time quantity and addictive usage patterns, using a comprehensive assessment of 5 mental health outcomes often overlooked in previous research. These findings highlight the need to expand public health strategies beyond screen time limits to include early IGD screening, emphasizing that both the quantity and quality of digital engagement require consideration. Future longitudinal studies should confirm these relationships and inform targeted interventions.

Acknowledgments

The authors would like to express their sincere gratitude to all the participants and their families for their valuable contribution to the cross-sectional survey on the mental health of children and adolescents in Pidu District, Chengdu.

Funding

This work was supported by the National Natural Science Foundation of China (82204170), the Science Fund for Creative Research Groups of Science and Technology Bureau of Sichuan Province (2024NSFTD0030), the Recruitment Program for Young Professionals of China, and other projects from West China School of Public Health and West China Fourth Hospital, Sichuan University.

Data Availability

The datasets generated and/or analyzed during this study are not publicly available, but can be obtained from the corresponding author upon reasonable request.

Authors' Contributions

XJ, MF, and QD conceived and designed the analysis. QD, LS, JH, XZ, RX, J Zhu, YQ, J Zhou, TY, XS, TH, SZ, and BY analyzed the data and interpreted the results. QD drafted the manuscript. MF and XJ revised the manuscript. XJ and MF contributed equally as co-corresponding authors. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables.
[XLSX File (Microsoft Excel File), 93 KB - jmir_v28i1e80737_app1.xlsx]

References

1. Khalaf AM, Alubied AA, Khalaf AM, Rifaey AA. The impact of social media on the mental health of adolescents and young adults: a systematic review. Cureus 2023;15(8):e42990 [FREE Full text] [doi: 10.7759/cureus.42990] [Medline: 37671234]
2. Pfeifer JH, Allen NB. Puberty initiates cascading relationships between neurodevelopmental, social, and internalizing processes across adolescence. Biol Psychiatry 2021;89(2):99-108 [FREE Full text] [doi: 10.1016/j.biopsych.2020.09.002] [Medline: 33334434]
3. Adolescent health. World Health Organization. 2025. URL: https://www.who.int/health-topics/adolescent-health/ [accessed 2025-12-16]
4. Malmir H, Mahdavi FS, Ejtahed H, Kazemian E, Chaharrahi A, Mohammadian Khonsari N, et al. Junk food consumption and psychological distress in children and adolescents: a systematic review and meta-analysis. Nutr Neurosci 2023;26(9):807-827. [doi: 10.1080/1028415X.2022.2094856] [Medline: 35816403]
5. Nunes ML, Bruni O. Insomnia in childhood and adolescence: clinical aspects, diagnosis, and therapeutic approach. J Pediatr (Rio J) 2015;91(6 Suppl 1):S26-S35 [FREE Full text] [doi: 10.1016/j.jped.2015.08.006] [Medline: 26392218]
6. de Zambotti M, Goldstone A, Colrain IM, Baker FC. Insomnia disorder in adolescence: diagnosis, impact, and treatment. Sleep Med Rev 2018;39:12-24 [FREE Full text] [doi: 10.1016/j.smrv.2017.06.009] [Medline: 28974427]
7. Kieling C, Baker-Henningham H, Belfer M, Conti G, Ertem I, Omigbodun O, et al. Child and adolescent mental health worldwide: evidence for action. Lancet 2011;378(9801):1515-1525. [doi: 10.1016/S0140-6736(11)60827-1] [Medline: 22008427]

XSL•FO
RenderX

8. Lu H, Yu Y, Wang DB, Wu AMS, Chen JH, Zhang G, et al. Association between interpersonal resources and mental health professional help-seeking among Chinese adolescents with probable depression: mediations via personal resources and active coping. BMC Psychiatry 2024;24(1):840 [FREE Full text] [doi: 10.1186/s12888-024-06271-4] [Medline: 39574049]
9. Chen S, Liu Y, Tremblay MS, Hong J, Tang Y, Cao Z, et al. Meeting 24-h movement guidelines: prevalence, correlates, and the relationships with overweight and obesity among Chinese children and adolescents. J Sport Health Sci 2021;10(3):349-359 [FREE Full text] [doi: 10.1016/j.jshs.2020.07.002] [Medline: 32679341]
10. Schønning V, Hjetland GJ, Aarø LE, Skogen JC. Social media use and mental health and well-being among adolescents - a scoping review. Front Psychol 2020;11:1949 [FREE Full text] [doi: 10.3389/fpsyg.2020.01949] [Medline: 32922333]
11. Nagata JM, Wong JH, Kim KE, Richardson RA, Nayak S, Potes C, et al. Social media use trajectories and cognitive performance in adolescents. JAMA 2025;334(21):1948-1950. [doi: 10.1001/jama.2025.16613] [Medline: 41082212]
12. Eales L, Wiglesworth A, Cullen KR, Klimes-Dougan B. Screen time, problematic media use, and clinical concerns in the ABCD study: differences by sex and race/ethnicity. Dev Psychopathol 2025:1-14. [doi: 10.1017/S0954579425100655] [Medline: 40970449]
13. Castro O, Bennie J, Vergeer I, Bosselut G, Biddle SJH. How sedentary are university students? A systematic review and meta-analysis. Prev Sci 2020;21(3):332-343. [doi: 10.1007/s11121-020-01093-8] [Medline: 31975312]
14. Xiao Y, Meng Y, Brown TT, Keyes KM, Mann JJ. Addictive screen use trajectories and suicidal behaviors, suicidal ideation, and mental health in US youths. JAMA 2025;334(3):219-228. [doi: 10.1001/jama.2025.7829] [Medline: 40531519]
15. Alfano CA, Moreno JP. Screen use in late childhood and early adolescence-a search for balance. JAMA Pediatr 2025;179(9):950-951. [doi: 10.1001/jamapediatrics.2025.1726] [Medline: 40549408]
16. Drescher AA, Goodwin JL, Silva GE, Quan SF. Caffeine and screen time in adolescence: associations with short sleep and obesity. J Clin Sleep Med 2011;7(4):337-342 [FREE Full text] [doi: 10.5664/JCSM.1182] [Medline: 21897768]
17. Wu X, Tao S, Zhang Y, Zhang S, Tao F. Low physical activity and high screen time can increase the risks of mental health problems and poor sleep quality among Chinese college students. PLoS One 2015;10(3):e0119607 [FREE Full text] [doi: 10.1371/journal.pone.0119607] [Medline: 25786030]
18. Zhou LH, Yu B, Xiao CC, Chen J, Zhu YZ, Yu QY, et al. [Dose-dependent associations between screen time, contents and adolescents' mental health]. Zhonghua Liu Xing Bing Xue Za Zhi 2025;46(6):1030-1035. [doi: 10.3760/cma.j.cn112338-20241014-00632] [Medline: 40518398]
19. Yin M, Huang S, Yu C. Depression and internet gaming disorder among chinese adolescents: a longitudinal moderated mediation model. Int J Environ Res Public Health 2023;20(4):3633 [FREE Full text] [doi: 10.3390/ijerph20043633] [Medline: 36834332]
20. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). Washington: American Psychiatric Association; 2013.
21. Wei C, Xu Q. The longitudinal relationship between school climate and internet gaming addiction among Chinese children: a moderated mediation model. Front Psychiatry 2025;16:1662888 [FREE Full text] [doi: 10.3389/fpsyt.2025.1662888] [Medline: 40995072]
22. Infanti A, Valls-Serrano C, Perales JC, Vögele C, Billieux J. Gaming passion contributes to the definition and identification of problematic gaming. Addict Behav 2023;147:107805 [FREE Full text] [doi: 10.1016/j.addbeh.2023.107805] [Medline: 37523871]
23. Kardefelt-Winther D, Heeren A, Schimmenti A, van Rooij A, Maurage P, Carras M, et al. How can we conceptualize behavioural addiction without pathologizing common behaviours? Addiction 2017;112(10):1709-1715 [FREE Full text] [doi: 10.1111/add.13763] [Medline: 28198052]
24. Gong X, Zhou J, Hao S. Longitudinal bidirectional relations between problematic internet game use and nonsuicidal self-injury among early adolescents: the mediating role of internalizing symptoms. Computers in Human Behavior 2025;165:108564 [FREE Full text] [doi: 10.1016/j.chb.2025.108564]
25. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
26. Council on Communications and Media. Children, adolescents, and the media. Pediatrics 2013;132(5):958-961. [doi: 10.1542/peds.2013-2656] [Medline: 28448255]
27. Pontes H, Griffiths M. Measuring DSM-5 internet gaming disorder: development and validation of a short psychometric scale. Computers in Human Behavior 2015;45:137-143 [FREE Full text] [doi: 10.1016/j.chb.2014.12.006]
28. Qin L, Cheng L, Hu M, Liu Q, Tong J, Hao W, et al. Clarification of the cut-off score for nine-item internet gaming disorder scale-short form (IGDS9-SF) in a Chinese context. Front Psychiatry 2020;11:470 [FREE Full text] [doi: 10.3389/fpsyt.2020.00470] [Medline: 32528331]
29. Furukawa TA, Kawakami N, Saitoh M, Ono Y, Nakane Y, Nakamura Y, et al. The performance of the Japanese version of the K6 and K10 in the world mental health survey Japan. Int J Methods Psychiatr Res 2008;17(3):152-158 [FREE Full text] [doi: 10.1002/mpr.257] [Medline: 18763695]
30. Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SLT, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002;32(6):959-976. [doi: 10.1017/s0033291702006074] [Medline: 12214795]

XSL•FO
RenderX

31. Romero-Acosta K, Lipps GE, Lowe GA, Gibson R, Ramirez-Giraldo A. The validation of the Kutcher Adolescent Depression 6-item scale in a sample of colombian preadolescents and adolescents. Eval Health Prof 2024;47(1):27-31. [doi: 10.1177/01632787231175931] [Medline: 37186570]
32. Dang W, Xu Y, Ji J, Wang K, Zhao S, Yu B, et al. Study of the SCL-90 scale and changes in the Chinese norms. Front Psychiatry 2020;11:524395 [FREE Full text] [doi: 10.3389/fpsyt.2020.524395] [Medline: 33584353]
33. First MB. Diagnostic and statistical manual of mental disorders, 5th edition, and clinical utility. J Nerv Ment Dis 2013;201(9):727-729. [doi: 10.1097/NMD.0b013e3182a2168a] [Medline: 23995026]
34. Osman A, Bagge CL, Gutierrez PM, Konick LC, Kopper BA, Barrios FX. The suicidal behaviors questionnaire-revised (SBQ-R): validation with clinical and nonclinical samples. Assessment 2001;8(4):443-454. [doi: 10.1177/107319110100800409] [Medline: 11785588]
35. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011;20(1):40-49 [FREE Full text] [doi: 10.1002/mpr.329] [Medline: 21499542]
36. Fu TS, Lee C, Gunnell D, Lee W, Cheng AT. Changing trends in the prevalence of common mental disorders in Taiwan: a 20-year repeated cross-sectional survey. Lancet 2013;381(9862):235-241. [doi: 10.1016/S0140-6736(12)61264-1] [Medline: 23151370]
37. Wang Y, Mínguez-Alarcón L, Gaskins AJ, Missmer SA, Rich-Edwards JW, Manson JE, et al. Association of spontaneous abortion with all cause and cause specific premature mortality: prospective cohort study. BMJ 2021;372:n530 [FREE Full text] [doi: 10.1136/bmj.n530] [Medline: 33762255]
38. Wang W, Du X, Guo Y, Li W, Zhang S, Zhang W, et al. Associations among screen time, sleep duration and depressive symptoms among Chinese adolescents. J Affect Disord 2021;284:69-74. [doi: 10.1016/j.jad.2021.01.082] [Medline: 33582434]
39. Gao Y, Wang J, Dong G. The prevalence and possible risk factors of internet gaming disorder among adolescents and young adults: systematic reviews and meta-analyses. J Psychiatr Res 2022;154:35-43. [doi: 10.1016/j.jpsychires.2022.06.049] [Medline: 35926424]
40. Wang J, Sheng J, Wang H. The association between mobile game addiction and depression, social anxiety, and loneliness. Front Public Health 2019;7:247. [doi: 10.3389/fpubh.2019.00247] [Medline: 31552213]
41. Li L, Abbey C, Wang H, Zhu A, Shao T, Dai D, et al. The association between video game time and adolescent mental health: evidence from rural China. Int J Environ Res Public Health 2022;19(22):14815 [FREE Full text] [doi: 10.3390/ijerph192214815] [Medline: 36429534]
42. Nagata JM, Al-Shoaibi AAA, Leong AW, Zamora G, Testa A, Ganson KT, et al. Screen time and mental health: a prospective analysis of the adolescent brain cognitive development (ABCD) study. BMC Public Health 2024;24(1):2686 [FREE Full text] [doi: 10.1186/s12889-024-20102-x] [Medline: 39370520]
43. Hoare E, Milton K, Foster C, Allender S. The associations between sedentary behaviour and mental health among adolescents: a systematic review. Int J Behav Nutr Phys Act 2016;13(1):108 [FREE Full text] [doi: 10.1186/s12966-016-0432-4] [Medline: 27717387]
44. Khan A, Burton N. Is physical inactivity associated with depressive symptoms among adolescents with high screen time? Evidence from a developing country. Mental Health and Physical Activity 2017;12:94-99 [FREE Full text] [doi: 10.1016/j.mhpa.2017.03.001]
45. Hrafnkelsdottir SM, Brychta RJ, Rognvaldsdottir V, Gestsdottir S, Chen KY, Johannsson E, et al. Less screen time and more frequent vigorous physical activity is associated with lower risk of reporting negative mental health symptoms among Icelandic adolescents. PLoS One 2018;13(4):e0196286 [FREE Full text] [doi: 10.1371/journal.pone.0196286] [Medline: 29698499]
46. Zink J, Belcher BR, Imm K, Leventhal AM. The relationship between screen-based sedentary behaviors and symptoms of depression and anxiety in youth: a systematic review of moderating variables. BMC Public Health 2020;20(1):472 [FREE Full text] [doi: 10.1186/s12889-020-08572-1] [Medline: 32272906]
47. Hills AP, King NA, Armstrong TP. The contribution of physical activity and sedentary behaviours to the growth and development of children and adolescents: implications for overweight and obesity. Sports Med 2007;37(6):533-545. [doi: 10.2165/00007256-200737060-00006] [Medline: 17503878]
48. Cuong VM, Assanangkornchai S, Wichaidit W, Minh Hanh VT, My Hanh HT. Associations between gaming disorder, parent-child relationship, parental supervision, and discipline styles: findings from a school-based survey during the COVID-19 pandemic in vietnam. J Behav Addict 2021;10(3):722-730 [FREE Full text] [doi: 10.1556/2006.2021.00064] [Medline: 34564065]
49. Monica A, Jiang JJ. Teens' Social Media Habits and Experiences. 2018. URL: http://tony-silva.com/eslefl/miscstudent/downloadpagearticles/teensandsocialmedia-pew.pdf [accessed 2025-12-16]
50. Kapidzic SCHS. Teens, gender, and self-presentation in social media. International Encyclopedia of Social and Behavioral Sciences. 2015. URL: https://homes.luddy.indiana.edu/herring/teens.gender.pdf [accessed 2025-12-16]
51. Hysing M, Pallesen S, Stormark KM, Lundervold AJ, Sivertsen B. Sleep patterns and insomnia among adolescents: a population-based study. J Sleep Res 2013;22(5):549-556. [doi: 10.1111/jsr.12055] [Medline: 23611716]

XSL•FO
RenderX

52. Pontes HM. Investigating the differential effects of social networking site addiction and internet gaming disorder on psychological health. J Behav Addict 2017;6(4):601-610 [FREE Full text] [doi: 10.1556/2006.6.2017.075] [Medline: 29130329]
53. Gabarron E, Denecke K, Lopez-Campos G. Evaluating the evidence: a systematic review of reviews of the effectiveness and safety of digital interventions for ADHD. BMC Psychiatry 2025;25(1):414 [FREE Full text] [doi: 10.1186/s12888-025-06825-0] [Medline: 40264083]
54. Xiuqin H, Huimin Z, Mengchen L, Jinan W, Ying Z, Ran T. Mental health, personality, and parental rearing styles of adolescents with internet addiction disorder. Cyberpsychol Behav Soc Netw 2010;13(4):401-406. [doi: 10.1089/cyber.2009.0222] [Medline: 20712498]
55. Zhou M, Sun X, Huang L, Zhang G, Kenny K, Xue H, et al. Parental migration and left-behind children's depressive symptoms: estimation based on a nationally-representative panel dataset. Int J Environ Res Public Health 2018;15(6) [FREE Full text] [doi: 10.3390/ijerph15061069] [Medline: 29795049]
56. He B, Fan J, Liu N, Li H, Wang Y, Williams J, et al. Depression risk of 'left-behind children' in rural China. Psychiatry Res 2012;200(2-3):306-312. [doi: 10.1016/j.psychres.2012.04.001] [Medline: 22572158]
57. Wu Q, Lu D, Kang M. Social capital and the mental health of children in rural China with different experiences of parental migration. Soc Sci Med 2015;132:270-277. [doi: 10.1016/j.socscimed.2014.10.050] [Medline: 25465499]

Abbreviations

DSM-5: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition)
IGD: internet gaming disorder
IGDS9-SF: Internet Gaming Disorder Scale-9 Item Short Form
ISI: Insomnia Severity Index
K6: Kessler 6-item
KADS-6: Kutcher Adolescent Depression 6-item
LST: leisure screen time
OR: odds ratio
PA: physical activity
SBQ-R: Suicide Behaviors Questionnaire-Revised
SCL-90: Symptom Checklist-90
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 15.Jul.2025; peer-reviewed by H Wan, X Liang, A Shivanna; comments to author 14.Sep.2025; accepted 07.Nov.2025; published 15.Jan.2026.

Please cite as:
Deng Q, Sha L, Hou J, Zhao X, Xiang R, Zhu J, Qu Y, Zhou J, Yu T, Song X, Zheng S, Han T, Yang B, Fan M, Jiang X
Leisure Screen Time, Internet Gaming Disorder, and Mental Health Among Chinese Adolescents: Large-Scale Cross-Sectional Study
J Med Internet Res 2026;28:e80737
URL: https://www.jmir.org/2026/1/e80737
doi:10.2196/80737
PMID:

©Qin Deng, Linna Sha, Jiaojiao Hou, Xunying Zhao, Rong Xiang, Jiangbo Zhu, Yang Qu, Jinyu Zhou, Ting Yu, Xin Song, Sirui Zheng, Tao Han, Bin Yang, Mengyu Fan, Xia Jiang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Longitudinal Between- and Within-Person Associations Among Screen Time, Bedtime, and Daytime Sleepiness Among Adolescents: Three-Wave Prospective Panel Study

Michał Tkaczyk[1], PhD; Albert J Ksinan[2], PhD; David Smahel[1], Prof Dr

[1]Interdisciplinary Research Team on Internet and Society, Faculty of Social Studies, Masaryk University, Brno, Czech Republic
[2]RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic

Corresponding Author:
Michał Tkaczyk, PhD
Interdisciplinary Research Team on Internet and Society
Faculty of Social Studies
Masaryk University
Joštova 218/10
Brno, 602 00
Czech Republic
Phone: 420 549 49 7289
Email: tkaczyk@fss.muni.cz

Abstract

Background: Daytime sleepiness is prevalent among adolescents and linked to multiple health and functional impairments. Prior research has identified digital media use and insufficient sleep as key predictors, yet the reciprocal longitudinal associations among screen time, sleep, and daytime sleepiness remain understudied.

Objective: This study examined the between- and within-person reciprocal longitudinal associations among adolescents' screen time, bedtime, and daytime sleepiness. It also tested whether potential adverse effects of screen time were less pronounced over time among adolescents who limited their screen exposure before sleep at baseline.

Methods: We conducted a prospective 3-wave panel study at 6-month intervals in a quota-based sample of 2500 Czech adolescents (mean age 13.43, SD 1.70 years; 1250/2500, 50% girls). Data were collected through an online survey. Screen time was assessed with 3 items covering total use of computers, smartphones, tablets, and television on a typical school day; bedtime with a single item on usual bedtime before school days; and daytime sleepiness with 4 items from the Pediatric Daytime Sleepiness Scale. Screen time restriction within 1 hour before sleep at baseline was examined as a moderator. Age and sex were included as covariates.

Results: Between- and within-person associations were estimated using random intercept cross-lagged panel models. Adolescents with higher screen time reported later bedtimes (r=0.23, 95% CI 0.15-0.31; P<.001) and greater daytime sleepiness (r=0.25, 95% CI 0.16-0.34; P<.001). No direct within-person effects emerged between screen time and daytime sleepiness (W1→W2: β=.02, 95% CI –0.11 to 0.16; P=.71; W2→W3: β=.02, 95% CI –0.10 to 0.14; P=.66). However, increases in screen time at Wave 1 predicted later bedtime at Wave 2 (β=.14, 95% CI 0.01-0.27; P<.05), which in turn predicted higher screen time at Wave 3 (β=.24, 95% CI 0.11-0.36; P<.001). Temporary within-person spikes in screen time coincided with delayed bedtimes (W1: r=0.16, 95% CI 0.04-0.27; P<.01; W2: r=0.23, 95% CI 0.09-0.36; P<.001; W3: r=0.09, 95% CI 0.00-0.20; P=.049). Baseline screen time restriction did not moderate within-person effects ($\Delta\chi^2_6$=5.3; P=.51).

Conclusions: This study is the first to test reciprocal longitudinal associations among adolescents' screen time, bedtime, and daytime sleepiness while separating between- and within-person associations, thereby addressing potential bias common in prior cross-lagged panel studies. The findings refine theoretical understanding by indicating a complex, bidirectional, and mutually reinforcing interplay between screen time and bedtime over time—even when individual differences are accounted for—whereas daytime sleepiness does not appear to be affected by this dynamic. Additionally, negatively correlated, within-person fluctuations in screen time and bedtime indicate that these behaviors are partly mutually exclusive and likely shaped by shared contextual influences. Screen time restriction before sleep did not mitigate within-person effects, indicating that interventions should prioritize consistent sleep schedules rather than focusing solely on reducing screen use.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e78972) doi:10.2196/78972

KEYWORDS

adolescents; bedtime; daytime sleepiness; random intercept cross-lagged panel model; RI-CLPM; screen restriction; screen time

Introduction

Background

Daytime sleepiness is an important, yet understudied, dimension of adolescents' sleep health [1]. Its prevalence varies widely across countries, ranging from 7.8% to 55.8% [2], and it is notably higher in adolescence than in adulthood [3]. Importantly, daytime sleepiness plays a central role in mediating the adverse effects of sleep impairment on adolescent health and well-being [4]. Studies have linked it to lower health-related quality of life [5], depressive symptoms, anxiety [6], heightened risk of mood disorders [7], and lower educational achievement [8]. Given its central role in linking sleep impairment to adverse outcomes, understanding the factors and processes contributing to daytime sleepiness in adolescence warrants greater scholarly attention.

Daytime sleepiness arises from an interplay of intrinsic (eg, brain maturation and sleep disorders) and extrinsic (eg, early school start times and poor sleep hygiene) factors [4]. Among these, insufficient sleep and late bedtimes on schooldays have been identified as the most direct contributors to daytime sleepiness among adolescents [9]. Digital media use is an important extrinsic factor known to affect sleep duration and bedtime timing; yet, most research on this association has been cross-sectional, limiting causal interpretations and leaving the direction of effects unclear [10,11].

Although the number of longitudinal studies is increasing [12], the vast majority do not distinguish between-person from within-person associations, which can lead to misleading conclusions about causal effects [13,14]. Studies that do separate these effects typically focus on short-term dynamics, such as day-to-day changes [15-18], often in small convenience samples, which limits their relevance for understanding longer-term processes.

To address these gaps in prior research, this study is the first to examine the longitudinal, reciprocal associations among screen time, bedtime, and daytime sleepiness, accounting for both stable between-person differences and within-person processes.

The study also tests whether restricting screen time before sleep moderates these associations. Clarifying whether daytime sleepiness emerges primarily from stable between-person differences, dynamic within-person processes, or both can advance theoretical understanding of how digital media use and adolescent sleep health influence each other. It may also help determine whether interventions should target stable behavioral patterns—such as sleep-related lifestyle habits, household routines, or family norms around screen use—or instead focus on longer-term individual trajectories—such as gradual increases in screen time or seasonal shifts in bedtime habits—or integrate both approaches.

Prior Work

Associations Among Screentime, Bedtime, and Daytime Sleepiness

Cross-sectional research consistently demonstrates positive associations between various screen-based activities—such as television watching, internet use, video gaming, and phone use—and both delayed bedtimes and increased daytime sleepiness [10,11]. Whereas this evidence cannot be a basis for causal interpretations, it suggests that, for some adolescents, higher screen time, later bedtimes, and greater daytime sleepiness tend to co-occur. This pattern likely reflects stable between-person differences that may be linked to external factors such as individual traits (eg, social anxiety), lifestyle demands (eg, extracurricular commitments), and family environment characteristics (eg, parenting style and household rules) [19-22].

The recent synthesis of evidence suggests that the causal link between screen time and sleep health is bidirectional, involving 2 potential pathways [23]. The screen-time-affecting-sleep pathway posits that media use, in particular before or after bedtime, contributes to shorter sleep duration and poorer sleep quality. Four explanatory mechanisms have been proposed: melatonin suppression due to blue light exposure, psychological arousal, displacement of sleep time, and sleep interruptions [24-26]. Of these, only displacement—that is, delayed bedtime due to screen time—and nighttime interruptions from notifications appear to have a substantial impact on sleep [23].

Conversely, the impaired-sleep-affecting-screen-time pathway posits that changes in sleep can contribute to increased media use. Three mechanisms explain this effect. Circadian phase shifts in puberty result in extended evening free time for media use [27,28]. Adolescents may use digital media to cope with sleep difficulties [23,29]. Daytime sleepiness is associated with more sedentary behavior, including prolonged screen time [30].

Longitudinal evidence supporting the 2 pathways is mixed. Some adolescent studies support the screen-time-affecting-sleep pathway (eg, meta-analysis by Pagano et al [12]), others report reciprocal associations [31,32], and some find minimal or no effects [33-35]. Evidence for the sleep-impairment-affecting-screen-time pathway exists, but in young adult samples [36]. There are also some longitudinal studies that found little or no support for either pathway or only marginal effects [33-35]. Such mixed findings may partly stem from conflating between-person and within-person associations in prior longitudinal studies.

To date, only 2 longitudinal studies have investigated the within-person associations between electronic media use and sleep-related outcomes—one focusing on daytime sleepiness [34] and the other on bedtime [33]. The former did not find significant within-person associations between the frequency of social media use and daytime sleepiness in Dutch adolescents

XSL•FO
RenderX

(aged 11-15 years), but did find between-person associations [34]. The latter, a 5-wave study of Finnish adolescents (aged 13-14 years at baseline), found no lagged effects and limited evidence of concurrent within-person associations—higher-than-usual social media use coincided with later-than-usual bedtime, but only in wave 1 [33]. These sparse findings suggest that the link between media use and sleepiness may arise from stable individual differences rather than changes over time.

A Moderating Role of Screen Time Restriction Before Sleep

Not all screen use is equally adverse for sleep health. In particular, evening screen time is considered detrimental to adolescent sleep [23], and restricting it is a common sleep hygiene recommendation [37]. Among adolescents, presleep screen restriction often results from parent-set technology rules, which cross-sectional studies have linked to less screen use, an earlier bedtime, and longer sleep duration [23]. While many adolescents do not follow their parents' technology rules and recommendations [38], research synthesis suggests that interventions aimed at reducing prebedtime screen use lead to modest improvements in bedtime and sleep duration [39]. Although this evidence suggests a potentially protective effect of reducing evening screen use, evidence on whether presleep screen restrictions moderate the longitudinal relationship between adolescents' screen time and sleep health is largely missing.

Covariates

Both screen use and sleep health vary by age and sex and therefore are important to consider when interpreting associations among adolescents' screen time and sleep health. In particular, older adolescents sleep less, go to bed later, and spend more time on screens, and younger adolescents are more likely to limit evening screen use [40-42]. Findings on sex differences in sleep are mixed. Some studies report no substantial differences [40], while others show girls sleep more than boys [43], or the reverse [44]. Daytime sleepiness findings are also inconsistent [45]. Sex differences in screen time are clearer. Boys exceed screen time limits more often [46], and sleep-disrupting screen activities differ—girls' sleep is more affected by social media, while boys' is impacted by video games [42]. Together, these patterns indicate that age and sex are important individual factors in understanding variation in adolescents' screen use and sleep health.

This Study

Prior longitudinal studies have rarely distinguished between stable between-person differences and within-person fluctuations in digital media use and sleep, leaving uncertainty about whether observed associations reflect enduring individual characteristics or dynamic changes over time [19]. The few adolescent studies that applied this distinction produced inconclusive results, with limited evidence for lagged or concurrent within-person effects [33,34]. To address this gap, this study extends prior work by examining the reciprocal longitudinal associations among adolescents' screen time, bedtime, and daytime sleepiness while separating between- and within-person processes. This allows us to clarify whether screen time and sleep co-vary because they influence each other over time or because of stable individual differences among adolescents. Furthermore, testing the moderating role of screen time restriction before sleep provides evidence on whether this common sleep hygiene recommendation mitigates longer-term effects of screen use on sleep health.

Specifically, we hypothesize that adolescents with higher overall screen time go to bed later and experience greater daytime sleepiness (Hypothesis 1); that increases in screen time are associated with a corresponding delay in bedtime and an increase in daytime sleepiness at the subsequent wave (Hypothesis 2), as well as within the same wave (Hypothesis 3); that delayed bedtime and increased daytime sleepiness are each associated with a subsequent increase in screen time (Hypothesis 4). The within-person effects expected in Hypotheses 2-4 reflect changes relative to a person's typical patterns. Finally, we hypothesize that within-person associations are weaker among adolescents who restrict their screen use before sleep (Hypothesis 5).

Methods

Ethical Considerations

The study was approved by the Research Ethics Committee at Masaryk University (EKV-2018-068). Before participation, respondents were informed about the nature and purpose of the survey, their right to decline involvement, and their ability to skip any questions by selecting the "I prefer not to say" option available for all items. Informed consent was obtained from both adolescents and parents. Parents were instructed not to be present during the adolescent survey to protect privacy. Adolescents were asked to indicate if an adult had observed or intervened. Although most caregivers appeared to comply, this could not be independently verified. All data were fully deidentified prior to analysis, and no identifying information was collected or stored. No identification of individual participants in any images of the manuscript or supplementary material is possible.

Participants received reward points equivalent to approximately US $4, added to the panelist's account and redeemable as cash or for charity donations.

Study Design and Setting

A longitudinal observational design was used. This 3-wave prospective panel study was a part of a larger multifocal study examining various aspects of adolescents' use of information and computer technologies and their impact on well-being. The first wave of data collection took place in June 2021, the second in November and December 2021, and the third in May and June 2022, with approximately 6 months between each wave. This study adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guidelines [47]; the completed STROBE checklist is provided in Multimedia Appendix 1 [48].

Participants

This study was conducted on a sample of 2500 Czech adolescents aged 11-16 years (mean age 13.43, SD 1.70 years;

XSL•FO
RenderX

1250/2500, 50% girls). Data were collected in the Czech Republic by an external research agency that recruited participants from existing online panels using face-to-face interviews, computer-assisted telephone interviewing, and online methods. Eligible participants were Czech households with at least 1 adolescent aged 11-16 years and a caregiver, enabling data collection from adolescent-parent dyads within the same household. Quota sampling was used to ensure equal representation of gender, age, and their combination and to ensure that the sample reflected the distribution of Czech households with children based on households' socioeconomic status (head of the household's education level) and place of residence (Nomenclature of Territorial Units for Statistics, level 3, municipality size, European Commission, 2020). Out of 2500 participants initially recruited at Wave 1, a total of 1654 completed Wave 2, corresponding to an attrition rate of 33.8% (846/2500). At Wave 3, a total of 1102 participants remained in the study. The overall attrition rate from Wave 1 to Wave 3 was 44.1% (1102/2500), with an incremental attrition rate of 33.4% (552/1654) between Wave 2 and Wave 3.

Measures

Screen Time

Screen time was assessed with 3 items, each starting with the question: "How much time (hours and minutes) do you spend doing the following activities during a typical school day?" The three items were: (1) "using a computer (PC or notebook)," (2) "using a cell phone or tablet," and (3) "watching TV, including various videos on TV (eg, DVD, Netflix)." In response to these items, respondents picked hours and minutes using a time spinner. The screen time score was then computed by adding up the scores of each item.

Bedtime

Bedtime was measured with 1 item: "When do you usually go to bed before school days?" In response to this item, respondents picked hours and minutes using a time spinner.

Daytime Sleepiness

Daytime sleepiness was measured using 4 items from the Pediatric Daytime Sleepiness Scale, which contains 8 items assessing the frequency of specific daytime sleepiness symptoms [49]. The 4 items were "You get sleepy or drowsy while doing your homework," "You have trouble getting out of bed in the morning," "You tell yourself that you need more sleep," and "You are tired and grumpy during the day." The items were rated on a 5-point scale: (1) "never,"(2) "rarely," (3) "sometimes," (4) "often," and (5) "very often." For each measurement occasion, a composite score was calculated as the mean of the items measuring the construct. A higher score indicates higher daytime sleepiness. Cronbach α was computed to assess the reliability of the scale across 3 waves. Reliability estimates were: α=0.77 for Wave 1, α=0.81 for Wave 2, and α=0.82 for Wave 3. These results indicate that the scale has acceptable internal consistency over time. Mean scores of the observed items were used for daytime sleepiness in analyses due to convergence issues when the latent variable was incorporated into the trivariate random intercept cross-lagged panel model (RI-CLPM).

Screen Time Restriction Within 1 Hour Before Sleep

Screen time restriction within 1 hour before sleep was measured at Wave 1. First, respondents were asked: "How long before going to sleep do you usually stop using all devices with a screen, ie, phone, tablet, computer, television?" Respondents picked hours and minutes using a time spinner in response to this item. Then, these data were transformed into a binary variable with values of 0 for adolescents who reported less than 60 minutes and 1 for adolescents who reported 60 minutes or more.

Covariates

Sex and age at baseline were self-reported at Wave 1 and were both included as time-invariant covariates in the analysis. Sex was coded as 0 for girls and 1 for boys, and age was grouped into 11-13 years (0) and 14-16 years (1).

Statistical Analysis

To examine the associations between screen time, bedtime, and daytime sleepiness over time while accounting for both between- and within-person sources of variance, we used RI-CLPMs fitted in *lavaan* (version 0.6-18) in R (version 4.4.1; R Core Team), allowing unbiased estimation of within-person effects net of stable individual differences [14]. The robust maximum likelihood estimator (MLR) was used, as it adjusts standard errors and chi-square statistics to accommodate nonnormal data (Section 3: "Testing normality assumptions" in supplementary materials provided by Tkaczyk et al [48]), yielding more accurate parameter estimates [50]. The proportion of missing data for the key time-varying variables ranged from 0.0% to 7.3% across waves. Little's Missing Completely at Random (MCAR) test indicated that the data were not completely missing at random (χ^2_{377}=787.8; P<.001; normed χ^2_{377}=2.1), suggesting a small to moderate deviation from MCAR. Given the low proportion of missing data (<8% per variable), full information maximum likelihood (FIML) estimation was used to handle missing values. For a detailed breakdown of percentages of missingness for each variable and wave, and results of logistic regressions testing the relationship between key analytical variables, demographics, and dropouts are provided in Section 1: "Attrition analysis" in supplementary materials provided by Tkaczyk et al [48].

To obtain more robust estimates, nonparametric bootstrapping with 2000 resamples was used to estimate 95% CIs for both unstandardized and standardized effects. Standardized coefficients represent the SD change in outcomes per 1 SD change in exposure. Chi-square difference tests were used to compare the fit of a nested model with constraints to the fit of the unconstrained model unless otherwise specified. The modeling approach was adapted from Mulder and Hamaker [51]. In the first step, the unconstrained RI-CLPM was compared to a model where all random intercept variances and covariances were set to zero (statistically equivalent to cross-lagged panel model [CLPM]) to test for stable between-unit differences, using the chi-bar-square test [52]. The comparison showed that the RI-CLPM fit the data better ($\Delta\chi^2_6$=286.8; P<.001). In addition, random intercepts of all 3 constructs had significant variance, indicating that there were some stable between-person

XSL•FO
RenderX

differences in screen time, bedtime, and daytime sleepiness over time. Second, to assess population-level changes in observed variables, we fixed grand means over time and compared this model to the unconstrained version. The comparison showed that the model without the constraints fit data better ($\Delta\chi^2_6$=74.1; *P*<.001), which implies that, on average, there was some change over time in all 3 variables. Third, to test whether the associations between screen time, bedtime, and daytime sleepiness were time-invariant, we constrained the autoregressive and cross-lagged paths, as well as the residual covariances. The model-building procedure indicated the fully unconstrained model as the best-fitting model Table 1. At this point, covariates (age and sex) were added to the model. The final model showed an adequate fit (χ^2_{15}=46.7; *P*<.001; Comparative Fit Index [CFI]=0.994; Tucker-Lewis Index [TLI]=0.977; root-mean-square error of approximation [RMSEA]=0.029, 90% CI 0.020-0.039; standardized root-mean-square residual [SRMR]=0.020). Fourth, moderation by screen time restriction before bed was tested using a multiple-group extension to RI-CLPM [51].

Table 1. Model fit indices for random intercept cross-lagged panel models (RI-CLPMs) examining longitudinal associations between screen time, bedtime, and daytime sleepiness across 3 waves in a longitudinal study of adolescents (aged 11-16 years). Data were collected in the Czech Republic between June 2021 and June 2022.

Model	χ^2(*df*)	CFI[a]	SRMR[b]	RMSEA[c]	TLI[d]	AIC[e]	BIC[f]
M0[g]	7.3 (3)	0.999	0.010	0.024	0.989	45591.676	45888.702
M1[h]	294.0 (9)	0.938	0.044	0.113	0.750	45866.448	46128.530
M2[i]	81.4 (9)	0.984	0.026	0.057	0.937	45653.804	45915.886
M3[j]	42.7 (18)	0.995	0.023	0.029	0.989	45597.131	45806.796
M3 + Covs[k]	46.7 (15)	0.994	0.020	0.029	0.977	45301.270	45633.241

[a]CFI: Comparative Fit Index.

[b]SRMR: standardized root-mean-square residual.

[c]RMSEA: root-mean-square error of approximation.

[d]TLI: Tucker-Lewis Index.

[e]AIC: Akaike information criterion.

[f]BIC: Bayesian information criterion.

[g]M0: fully unconstrained RI-CLPM.

[h]M1: cross-lagged panel model [CLPM].

[i]M2: RI-CLPM with grand means constrained over time.

[j]M3: RI-CLPM with constraint over time imposed on auto-regressive paths, cross-lagged paths, and residual (co)variances.

[k]M3 + Covs: M0 with covariates (age and sex).

Results

Descriptive Analysis

Table 2 displays pairwise correlations for time-varying variables across waves, along with their descriptive statistics, skewness, and kurtosis. The means of daytime sleepiness are close to "sometimes" (Wave 1: 2.81, SD 0.80; Wave 2: 2.82, SD 0.82; Wave 3: 2.84, SD 0.82). At Wave 1, approximately every third (788/2500, 32%) participant reported having trouble getting out of bed in the morning often or very often. Getting sleepy or drowsy while doing homework was the least frequent symptom—at Wave 1, approximately every sixth (400/2494, 16%) participant reported experiencing it often or very often.

XSL•FO
RenderX

Table 2. Pearson correlations and descriptive statistics for screen time, bedtime, and daytime sleepiness across 3 waves in a longitudinal study of adolescents (aged 11-16 years). Data were collected in the Czech Republic between June 2021 and June 2022. All correlation coefficients (r) are significant at $P<.001$.

Variable	ST[a] (W1[b])	ST (W2[c])	ST (W3[d])	BT[e] (W1)	BT (W2)	BT (W3)	DS[f] (W1)	DS (W2)	DS (W3)
ST (W1), *r*	1.00								
ST (W2), *r*	0.63	1.00							
ST (W3), *r*	0.59	0.62	1.00						
BT (W1), *r*	0.23	0.16	0.14	1.00					
BT (W2), *r*	0.17	0.20	0.21	0.61	1.00				
BT (W3), *r*	0.13	0.11	0.18	0.56	0.63	1.00			
DS (W1), *r*	0.13	0.09	0.09	0.22	0.17	0.13	1.00		
DS (W2), *r*	0.15	0.15	0.13	0.20	0.23	0.18	0.57	1.00	
DS (W3), *r*	0.14	0.13	0.16	0.21	0.19	0.20	0.55	0.64	1.00
Mean (SD), hh:mm or scale	06:23 (02:40)	06:11 (02:36)	06:02 (02:37)	09:48 00:56	09:47 (00:56)	09:57 (00:58)	2.81 (0.80)	2.82 (0.82)	2.84 (0.82)
Skewness	0.38	0.51	0.59	0.04	0.31	0.26	0.15	0.13	0.10
Kurtosis	−0.54	−0.29	−0.23	0.28	0.83	0.58	0.14	0.11	0.09

[a]ST: screen time (hh:mm).

[b]W1: Wave 1.

[c]W2: Wave 2.

[d]W3: Wave 3.

[e]BT: bedtime (hh:mm PM).

[f]DS: daytime sleepiness (1-5).

Average bedtimes at each wave were before 10 PM (Wave 1: 9:48, SD 00:56; Wave 2: 9:47, SD 00:56; Wave 3: 9:57, SD 00:58). At Wave 1, a total of 14% (353/2465) of participants reported bedtime at 11:00 PM or later (213/1621, 13% at Wave 2 and 190/1093, 17% at Wave 3). Average total daily screen times were close to 6 hours at each wave and showed a decreasing tendency across time (Wave 1: 06:23, SD 02:40; Wave 2: 06:11, SD 02:36; Wave 3: 06:02, SD 02:37).

Intraclass correlation coefficients (ICCs) revealed that between-person differences accounted for approximately 64% of the variance in screen time, 60% in bedtime, and 58% in daytime sleepiness, indicating a smaller but substantial proportion of variance due to within-person changes over time. All variables showed statistically significant and positive correlations both within and across waves.

Between-Person Associations Among Screen Time, Bedtime, and Daytime Sleepiness

Standardized path coefficients of the final RI-CLPM are presented in Figure 1.

The analysis revealed significant positive associations between the random intercepts of screen time and bedtime (r=0.23, 95% CI 0.15-0.31; P<.001), screen time and daytime sleepiness (r=0.25, 95% CI 0.16-0.34; P<.001), and bedtime and daytime sleepiness (r=0.31, 95% CI 0.22-0.41; P<.001). Consistent with Hypothesis 1, these correlations indicate that adolescents who typically use screens more also tend to go to bed later and experience higher daytime sleepiness. Additionally, those with later bedtimes tend to experience higher daytime sleepiness.

XSL•FO
RenderX

Figure 1. Standardized path coefficients of the final random intercept cross-lagged panel model testing between- and within-person associations among screen time, bedtime, and daily sleepiness across 3 measurement waves in a longitudinal study of adolescents (aged 11-16 years) conducted in the Czech Republic between June 2021 and June 2022. The model controls for the effects of age (at Wave 1 [W1]) and sex on the random intercepts of the time-varying variables. Solid black lines represent significant paths, and dashed lines represent nonsignificant paths. Solid gray paths were fixed to 1. **P*<.05; ***P*<.01; ****P*<.001.

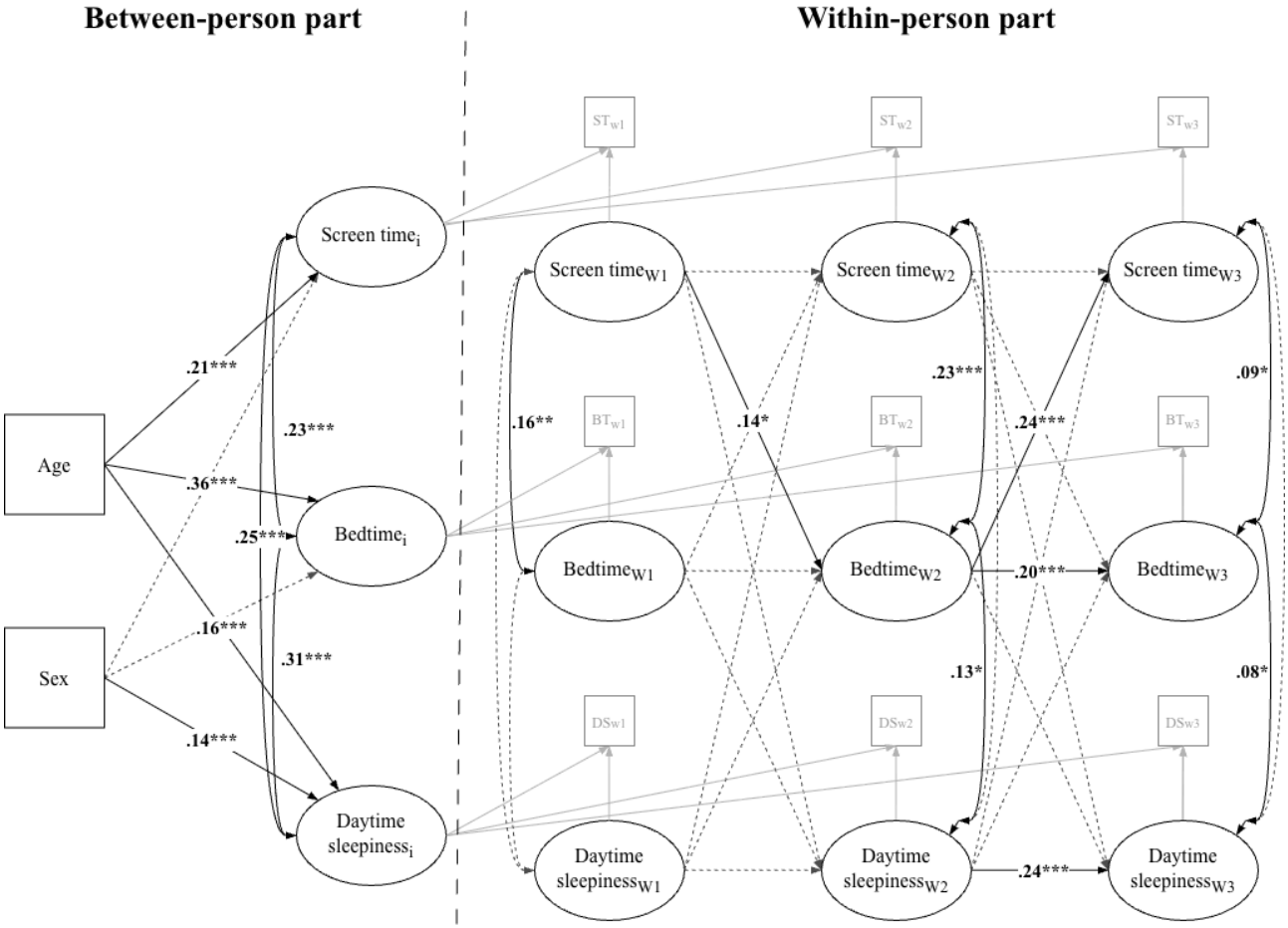

Within-Person Associations Among Screen Time, Bedtime, and Daytime Sleepiness

The analysis identified 2 significant cross-lagged effects. Consistent with Hypothesis 2, elevated screen time at Wave 1 relative to a person's usual patterns, was associated with elevated bedtime at Wave 2 (β=.14, 95% CI 0.01-0.27; *P*=.02). Similarly, in line with Hypothesis 4, elevated bedtime at Wave 2, relative to a person's usual patterns, was associated with increased screen time at Wave 3 (β=.24, 95% CI 0.11-0.36; *P*<.001). No evidence was found for the remaining cross-lagged paths hypothesized in Hypothesis 2 or Hypothesis 4.

Consistent with Hypothesis 3, the analysis revealed consistent concurrent associations between the within-person components of screen time and bedtime (Wave 1: β=.16, 95% CI 0.04-0.27; *P*=.007; Wave 2: β=.23, 95% CI 0.010-0.36; *P*<.001; Wave 3: β=.09, 95% CI 0.01-0.19; *P*=.049), indicating that an increase in screen time—relative to a person's usual patterns—was associated with a corresponding delay in bedtime with the same wave. No evidence was found to support Hypothesis 3. Additionally, a significant concurrent association between bedtime and daytime sleepiness was observed at Wave 2 (β=.13, 95% CI 0.01-0.26; *P*=.045) and Wave 3 (β=.08, 95% CI 0.00-0.17; *P*=.04). This indicates that, within these waves, a delay in bedtime was associated with elevated daytime sleepiness relative to a person's usual level of sleepiness.

The analysis also revealed autoregressive effects. Elevated bedtime at Wave 2, relative to a person's usual patterns, was associated with elevated bedtime at Wave 3 (β=.20, 95% CI 0.05-0.36; *P*<.001), indicating that a delay in bedtime—relative to a person's usual patterns—tends to carry over time. A similar autoregressive effect was observed for daytime sleepiness, with elevated sleepiness at Wave 2 associated with elevated sleepiness at Wave 3 (β=.24, 95% CI 0.12-0.37; *P*<.001).

The Role of Covariates

Age significantly predicted the intercepts of screen time (β=.21, 95% CI 0.16-0.25; *P*<.001), bedtime (β=.36, 95% CI 0.32-0.41; *P*<.001), and daytime sleepiness (β=.16, 95% CI 0.11-0.21; *P*<.001), indicating that older adolescents (aged 14-16 years) typically spent more time using screen media, have later bedtimes, and experience higher daytime sleepiness compared with younger adolescents (aged 11-13 years). Sex (boy=1) significantly predicted the intercept of daytime sleepiness (β=.14, 95% CI 0.10-0.19; *P*<.001), indicating that typical levels of daytime sleepiness are higher for boys than for girls (Table 3).

XSL•FO
RenderX

Table 3. Estimated parameters of the random intercept cross-lagged panel model (RI-CLPM) testing between- and within-person associations of screen time, bedtime, and daytime sleepiness across 3 measurement waves in a longitudinal study of adolescents (aged 11-16 years) conducted in the Czech Republic between June 2021 and June 2022. The model controls for the effects of age (at Wave 1) and sex on the random intercepts of the time-varying variables.

Parameter	*B*	SE	95% CI	*P* value	β
Between-person associations					
Correlations					
ST_i[a] ↔ BT_i[b]	0.314	0.062	0.184 to 0.436	<.001	.229
ST_i ↔ DS_i[c]	0.296	0.050	0.177 to 0.407	<.001	.250
BT_i ↔ DS_i	0.123	0.019	0.084 to 0.159	<.001	.312
Covariates					
Age → ST_i	0.859	0.100	0.658 to 1.05	<.001	.206
Sex → ST_i	0.050	0.099	−0.138 to 0.247	.62	.012
Age → BT_i	0.523	0.033	0.457 to 0.591	<.001	.362
Sex → BT_i	0.054	0.033	−0.012 to 0.112	.08	.038
Age → DS_i	0.191	0.029	0.135 to 0.248	<.001	.160
Sex → DS_i	0.172	0.028	0.113 to 0.228	<.001	.144
Within-person associations					
Autoregressive paths					
ST[d] (W1[e]) → ST (W2[f])	0.100	0.070	−0.049 to 0.235	.15	.100
BT[g] (W1) → BT (W2)	0.047	0.081	−0.145 to 0.233	.56	.045
DS[h] (W1) → DS (W2)	0.027	0.074	−0.147 to 0.183	.72	.025
ST (W2) → ST (W3[i])	0.056	0.068	−0.103 to 0.201	.41	.056
BT (W2) → BT (W3)	0.224	0.059	0.04 to 0.387	<.001	.200
DS (W2) → DS (W3)	0.245	0.051	0.112 to 0.362	<.001	.242
Cross-lagged paths					
ST (W1) → BT (W2)	0.051	0.023	0.002 to 0.099	.03	.139
ST (W1) → DS (W2)	0.008	0.020	−0.037 to 0.051	.71	.023
BT (W1) → ST (W2)	0.131	0.171	−0.238 to 0.463	.44	.046
BT (W1) → DS (W2)	0.006	0.057	−0.124 to 0.133	.92	.006
DS (W1) → ST (W2)	0.001	0.186	−0.411 to 0.398	>.99	.000
DS (W1) → BT (W2)	-0.018	0.072	−0.178 to 0.125	.81	−.015
ST (W2) → BT (W3)	-0.009	0.022	−0.057 to 0.037	.69	−.021
ST (W2) → DS (W3)	0.008	0.018	−0.037 to 0.051	.66	.024
BT (W2) → ST (W3)	0.635	0.162	0.309 to 0.976	<.001	.235
BT (W2) → DS (W3)	-0.010	0.044	−0.113 to 0.084	.83	−.011
DS (W2) → ST (W3)	0.116	0.173	−0.25 to 0.481	.50	.038
DS (W2) → BT (W3)	0.102	0.061	−0.025 to 0.23	.09	.080
Residual covariances					
ST (W1) ↔ BT (W1)	0.156	0.058	0.04 to 0.277	.007	.158
ST (W1)↔ DS (W1)	0.043	0.047	−0.059 to 0.156	.05	.050
BT (W1) ↔ DS (W1)	0.025	0.018	−0.011 to 0.063	.15	.084
ST (W2) ↔ BT (W2)	0.229	0.066	0.087 to 0.364	<.001	.225

XSL•FO
RenderX

Parameter	*B*	SE	95% CI	*P* value	β
ST (W2) ↔ DS (W2)	0.098	0.057	−0.021 to 0.213	.09	.108
BT (W2) ↔ DS (W2)	0.042	0.021	−0.005 to 0.088	.045	.127
ST (W3) ↔ BT (W3)	0.099	0.050	−0.008 to 0.209	.049	.091
ST (W3) ↔ DS (W3)	0.077	0.040	−0.006 to 0.159	.06	.090
BT (W3) ↔ DS (W3)	0.030	0.014	−0.001 to 0.061	.04	.083

[a]ST_i: screen time latent intercept.

[b]BT_i: bedtime latent intercept.

[c]DS_i: daytime sleepiness latent intercept.

[d]ST: screen time.

[e]W1: Wave 1.

[f]W2: Wave 2.

[g]BT: bedtime.

[h]DS: daytime sleepiness.

[i]W3: Wave 3.

The Moderating Role of Screen Time Restriction Before Bed

Against Hypothesis 5, comparisons of multiple group RI-CLPMs with and without constraints across groups showed no differences in correlations between random intercepts ($\Delta\chi^2_3$=6.0; *P*=.11), residual covariances ($\Delta\chi^2_6$=3.5; *P*=.74), or cross-lagged associations ($\Delta\chi^2_6$=5.3; *P*=.51) across adolescents who restricted screen time 1 hour before bed at Wave 1 and those who did not. However, some significant differences between those groups were found ($\Delta\chi^2_2$=32.0; *P*<.001). Adolescents who restricted their screen time before bed reported, on average, shorter screen time (by 27 minutes and 28 seconds), earlier bedtime (22 minutes and 12 seconds), and lower daytime sleepiness (Δ=0.159; Table 3).

Discussion

Principal Results

This 3-wave prospective panel study examined bidirectional relationships between screen time, bedtime, and daytime sleepiness in a large representative sample of early to midadolescents in the Czech Republic. Findings at the between-person level showed that higher screen time, later bedtimes, and increased daytime sleepiness tend to co-occur among adolescents. At the within-person level, results revealed a bidirectional, transactional association between screen time and bedtime, suggesting mutual reinforcement over time. Additionally, temporary, wave-specific deviations in screen time and bedtime—relative to a person's usual patterns—were positively correlated, suggesting that increases in screen time and delays in bedtime tend to co-occur within individuals at the same wave. Finally, while restricting screen time before sleep did not modify these associations, adolescents who restricted screen time had lower typical screen time, earlier bedtimes, and less daytime sleepiness on average.

Between-Person Associations Among Screen Time, Bedtime, and Daytime Sleepiness

Consistent with Hypothesis 1, the analysis revealed small to medium positive correlations between screen time, bedtime, and daytime sleepiness at the between-person level, aligning with findings from cross-sectional studies [4,10,11,53]. However, previous RI-CLPM studies reported mixed correlation patterns. For instance, Maksniemi et al [33] found no significant between-person correlations between active social media use and bedtime, whereas 2 other studies reported medium positive correlations between social media use and daytime sleepiness [34] and between media multitasking and sleep problems [35]. Such inconsistencies may reflect differences across studies in how media use was conceptualized and defined (eg, active vs general social media use).

The between-person associations observed in this study indicate that higher screen time and poorer sleep co - occur as relatively stable individual tendencies, likely shaped by other stable factors. For example, late chronotype may predispose some adolescents to later bedtimes and heavier evening media use [54]. Prior work has shown that modifiable factors—such as parenting style [55], parental sleep [21], media habits [56], and household rules [57,58]—also influence both adolescent media habits and sleep. To guide better-targeted interventions, future longitudinal RI-CLPM studies should investigate how various modifiable family and lifestyle factors influence the media–sleep association over time.

Within-Person Associations Among Screen Time, Bedtime, and Daytime Sleepiness Over Time

Consistent with Hypothesis 2, increased screen time was associated with delayed bedtime 6 months later, but only between Waves 1 and 2. According to the interpretation guidelines proposed by Orth et al [59], the effect is considered large. The association, although not consistent across all waves, aligns with prior longitudinal research, including a 6-wave study based on data from the ABCD study among adolescents aged 11-14 years [60] and a 2-wave study among adolescents aged 13-14 years [31], both of which link media use to later bedtimes

XSL•FO
RenderX

over time. The present findings extend the prior evidence by demonstrating the association even when controlling for stable between‐person differences. Other RI-CLPM studies did not find cross-lagged effects; for instance, Maksniemi et al [33] found no association between active social media use and bedtime. Such discrepancies may reflect differences in conceptualizing media use—overall screen time versus active social media use—which involve distinct pathways linking media to sleep. Whereas active social media use mainly disrupts sleep through presleep arousal [61], total screen time is more closely related to blue light exposure and sleep displacement, the latter showing stronger and substantial associations with reduced sleep duration [23].

Contrary to Hypothesis 2, this study found no evidence of a direct within-person association between screen time and daytime sleepiness in the long term. This result is consistent with 2 previous RI-CLPM studies. Van Der Schuur et al [34] also found no evidence of a long-term within-person association between social media use and daytime sleepiness, aside from a small effect of social media stress among girls. Van der Schuur et al [35] found no direct path from media multitasking to sleep problems (including daytime sleepiness), except for a marginally significant effect of media multitasking among girls. Although direct effects were absent, indirect pathways remain plausible. Daytime sleepiness may occur when screen use results in later bedtimes [24]. Although bedtime was not formally tested as a mediator in this study, which should be considered a limitation, future longitudinal studies might examine bedtime delay as a pathway linking screen time to daytime sleepiness.

Consistent with Hypothesis 3, temporary increases in screen time coincided with temporary delays in bedtime across all 3 waves, indicating concurrent within-person associations between the two. Similar results were found by Maksniemi et al [33] in a single wave, whereas other RI-CLPM studies did not examine concurrent associations [34,35]. This pattern likely reflects the mutually exclusive nature of screen use and sleep within daily time allocation [62]; yet, the association manifests itself in period-specific, typical patterns of behavior—during periods when bedtime is delayed, adolescents have more opportunities for screen use, and conversely, during periods with greater screen use, they have less time available for sleep. Findings further indicate that bedtime remains sensitive to short-term, period-specific changes in screen time (and vice versa) and that both may share common contextual drivers.

Against Hypothesis 3, this study found no evidence of a correlated change between screen time and daytime sleepiness, suggesting that short-term increases in screen time do not directly coincide with increased sleepiness. Similarly, a diary study on smartphone use and next-day sleepiness found no such effects [18]. Delayed bedtimes in Waves 2 and 3 were concurrently linked to increased daytime sleepiness, likely due to shorter sleep duration [63]. Overall, the pattern of longitudinal associations found in this study suggests that while screen time and daytime sleepiness are not directly linked at the within-person level, an indirect path is possible, whereby delayed bedtime may mediate the association between technology use and daytime sleepiness.

Consistent with Hypothesis 4, this study found a within-person cross-lagged effect of bedtime on screen time in the subsequent wave: a later-than-usual bedtime predicted increased screen time 6 months later, but only between Waves 2 and 3. This finding aligns with prior longitudinal research showing reciprocal effects between poorer sleep and greater media use [31,36]. The RI-CLPM study by Van der Schuur et al [34] provided partial evidence for the opposite direction, with increased daytime sleepiness predicting decreased social media use over time among boys. Unlike earlier RI-CLPM studies, this study supports the sleep-impairment-affecting-screen-time pathway, demonstrating a substantial effect even after accounting for stable between-person differences. Although this effect was not consistent across all waves, it suggests that adolescents may extend screen use to fill additional evening hours, which likely arises from circadian shifts or related factors [28].

Overall, discrepancies in cross-lagged effects across RI-CLPM studies may partly reflect differences in the time intervals between measurements. This study used a 6-month interval, whereas Van der Schuur et al [34,35] used a 3- to 4-month interval, and Maksniemi et al [33] used a 1-year interval. The absence of cross-lagged effects in some cases suggests that these intervals may not have been optimal for capturing the underlying dynamics [33]. Future research could benefit from greater use of different temporal designs, such as shortitudinals, to identify optimal temporal windows for detecting within-person effects and the temporal dynamics through which media use influences sleep across adolescence.

Taken together, the cross-lagged pattern (screen time → bedtime between Waves 1 and 2; bedtime → screen time between Waves 2 and 3) suggests a reinforcing cycle between increased screen time and delayed bedtime over time. While previous research identified bidirectional links between screen time and sleep [31], this study extends prior work by being the first to demonstrate this reinforcing pattern longitudinally using an RI-CLPM that accounts for stable between-person differences. The autoregressive effects further indicate that delayed bedtimes tend to carry over across waves—a finding also reported in other RI-CLPM studies [33], which may reflect adolescent circadian shifts or habitual delays associated with greater autonomy or increased school demands [9]. Considering that delayed bedtimes were concurrently linked to greater daytime sleepiness and prospectively to higher screen time, interventions that promote earlier and more consistent sleep schedules, rather than solely limiting screen use, may be more effective for improving adolescent sleep health.

Effects of Screen Time Restriction Before Sleep

Contrary to Hypothesis 5, this study found no evidence that restricting screen time before sleep affected within-person associations between screen time and sleep, particularly regarding the development of sleep displacement over time. Prior findings are mixed—while experimental studies have shown improvements in sleep outcomes [64,65], observational studies often report no adverse effects of prebedtime smartphone use [17,18], with inconsistent adherence to parent-set rules frequently cited as a limiting factor [61]. These discrepancies

XSL•FO
RenderX

likely reflect differences in study design, sampling strategies, and time frames (eg, short- vs long-term). It should also be noted that the comparison groups were defined based on screen time restriction assessed at Wave 1 only. However, this behavior was not stable over time—among those who reported limiting their screen use at Wave 1, only 40% (284/710) did so across all 3 waves, and 65% (460/710) did so at least once thereafter. Future research should account for this temporal variability when examining the long-term effects of screen time restriction.

Adolescents who reported restricting screen use before sleep also tended to report lower overall screen exposure, earlier bedtimes, and less daytime sleepiness than their peers. Although these between-person differences may indicate a protective role of screen time restriction, they could also reflect other stable characteristics such as family environment (eg, parenting style), chronotype, or self-regulation. Prior research has linked adverse parenting styles to poorer sleep quality and greater daytime sleepiness [55], and greater sleepiness to lower self-regulation and eveningness chronotype [66]. Future longitudinal studies should account for these factors and examine their potential moderating roles in the relationship between screen use and sleep outcomes.

Limitations

Several limitations should be considered when interpreting these findings. First, the study relied on self-reported measures of screen time and sleep, which may be prone to inaccuracy [67,68]. Because overall screen time was calculated by summing reported use across multiple devices that could have been used simultaneously, average values may overestimate actual exposure. Future research should integrate digital trace data [69] and wrist-worn accelerometers data [70] for more accurate measurements.

Second, measurement simplifications—using total screen time and an abbreviated version of the Pediatric Daytime Sleepiness Scale [49]—may have reduced precision and obscured associations with sleep [71]. Future studies should use more detailed measures that account for media functions, content, and context of use [72,73].

Third, with only three waves spaced 6 months apart, the design was insufficient for modeling longer-term developmental trajectories [53,74] or accounting for seasonal variability in screen time and sleep [75,76]. Longer follow-up and more frequent measurement occasions would allow finer modeling of these changes.

Fourth, attrition was higher than in comparable school-based studies [33-35], likely because data were collected through an online panel and required the agreement of both adolescent and parent or caregiver. Online panels typically exhibit higher attrition rates due to the sustained participant burden and email-based recontact [77,78], and similar rates have been reported in other adolescent panel studies [79]. Notably, attrition remained high despite offering substantially increased incentives (160% in Waves 1 and 2; 280% in Wave 3). Dropouts reported slightly higher baseline screen time (Tables S1 and S2 in supplementary materials provided by Tkaczyk et al [48]), which may limit generalizability to heavy screen users.

Finally, data collection partially overlapped with COVID-19 social distancing measures, which were associated with increased screen time and later bedtimes among adolescents [60,80]. The stringency of restrictions varied across waves: Wave 1 (June 2021) coincided with the strictest measures, Wave 2 (November-December 2021) with moderate restrictions, and Wave 3 (May-June 2022) after their removal [81]. This variation may partly explain the observed decrease in screen time and the stability of bedtime between Waves 1 and 2.

Conclusion

This study is the first to test reciprocal longitudinal associations among adolescents' screen time, bedtime, and daytime sleepiness while separating between- and within-person processes, thereby addressing bias common in prior cross-lagged panel studies. The findings refine theoretical understanding by showing a complex, bidirectional, and mutually reinforcing interplay between screen time and bedtime over time, even after accounting for stable individual differences. Between-person associations revealed that adolescents with higher screen use had poorer sleep, likely reflecting the influence of relatively stable individual and environmental factors. Although specific cross-lagged effects varied across waves, the overall pattern supports both the screen-time-affecting-sleep and sleep-impairment-affecting-screen-time pathways, whereas daytime sleepiness was not affected by this dynamic. Negatively correlated within-person fluctuations further indicate that screen time and bedtime are partly mutually exclusive and may share contextual drivers.

Screen time restriction before sleep did not moderate within-person effects. However, at the between-person level, adolescents who practiced it reported lower screen use, earlier bedtimes, and less daytime sleepiness. Taken together, these findings suggest that interventions emphasizing consistent sleep schedules and supportive family routines—rather than focusing solely on limiting screen use—may be most effective for promoting adolescent sleep health. Future research should incorporate objective measurements on multiple time scales and relevant moderators.

Acknowledgments

The authors thank Dr David Lacko, Dr Vojtěch Mýlek, and Martin Tancoš for their thoughtful consultations during the preparation of this manuscript. During the preparation of this work, we used the generative artificial intelligence (AI) tool ChatGPT by OpenAI [82] to improve language clarity and readability. After using this service, we reviewed and edited the content as needed and take full responsibility for the publication's content.

XSL•FO
RenderX

Funding

This work has been funded by a grant from the Programme Johannes Amos Comenius under the Ministry of Education, Youth and Sports of the Czech Republic from the project "Research of Excellence on Digital Technologies and Wellbeing CZ.02.01.01/00/22_008/0004583" which is co-financed by the European Union. The work of AJK was supported from Operational Programme Johannes Amos Comenius—Project MSCAfellow5_MUNI (No. CZ.02.01.01/00/22_010/0003229). The funding sources were not involved in any research decisions.

Data Availability

The data used in this study are openly available on the Open Science Framework (OSF) [48].

Authors' Contributions

Conceptualization: MT (lead), AJK (supporting), DS (supporting)
Formal analysis: MT (lead), AJK (supporting)
Funding acquisition: DS
Methodology: MT (lead), AJK (supporting)
Project administration: DS
Supervision: DS
Validation: MT (lead), AJK (supporting)
Visualization: MT
Writing – original draft: MT
Writing – review & editing: MT (lead), AJK (supporting), DS (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
STROBE checklist.
[PDF File (Adobe PDF File), 183 KB - jmir_v28i1e78972_app1.pdf]

References

1. Buysse DJ. Sleep health: can we define it? Does it matter? Sleep 2014;37(1):9-17 [FREE Full text] [doi: 10.5665/sleep.3298] [Medline: 24470692]
2. Pereira É, Teixeira CS, Louzada FM. Sonolência diurna excessiva em adolescentes: prevalência e fatores associados. Rev Paul Pediatr 2010;28(1):98-103. [doi: 10.1590/s0103-05822010000100015]
3. Roehrs T, Carskadon MA, Dement WC, Roth T. Daytime sleepiness and alertness. In: Principles and Practice of Sleep Medicine. United States: Elsevier; 2005:39-50.
4. Moore M, Meltzer LJ. The sleepy adolescent: causes and consequences of sleepiness in teens. Paediatr Respir Rev 2008;9(2):114-121. [doi: 10.1016/j.prrv.2008.01.001] [Medline: 18513671]
5. Gustafsson M, Laaksonen C, Aromaa M, Asanti R, Heinonen OJ, Koski P, et al. Association between amount of sleep, daytime sleepiness and health-related quality of life in schoolchildren. J Adv Nurs 2016;72(6):1263-1272. [doi: 10.1111/jan.12911] [Medline: 26899487]
6. Moore M, Kirchner HL, Drotar D, Johnson N, Rosen C, Ancoli-Israel S, et al. Relationships among sleepiness, sleep time, and psychological functioning in adolescents. J Pediatr Psychol 2009;34(10):1175-1183 [FREE Full text] [doi: 10.1093/jpepsy/jsp039] [Medline: 19494088]
7. Lofthouse N, Gilchrist R, Splaingard M. Mood-related sleep problems in children and adolescents. Child Adolesc Psychiatr Clin N Am 2009;18(4):893-916. [doi: 10.1016/j.chc.2009.04.007] [Medline: 19836695]
8. Dewald JF, Meijer AM, Oort FJ, Kerkhof GA, Bögels SM. The influence of sleep quality, sleep duration and sleepiness on school performance in children and adolescents: a meta-analytic review. Sleep Med Rev 2010;14(3):179-189. [doi: 10.1016/j.smrv.2009.10.004] [Medline: 20093054]
9. Carskadon MA. Sleep in adolescents: the perfect storm. Pediatr Clin North Am 2011;58(3):637-647 [FREE Full text] [doi: 10.1016/j.pcl.2011.03.003] [Medline: 21600346]
10. Hale L, Guan S. Screen time and sleep among school-aged children and adolescents: a systematic literature review. Sleep Med Rev 2015;21:50-58 [FREE Full text] [doi: 10.1016/j.smrv.2014.07.007] [Medline: 25193149]
11. Lund L, Sølvhøj IN, Danielsen D, Andersen S. Electronic media use and sleep in children and adolescents in western countries: a systematic review. BMC Public Health 2021;21(1):1598 [FREE Full text] [doi: 10.1186/s12889-021-11640-9] [Medline: 34587944]

XSL•FO
RenderX

12. Pagano M, Bacaro V, Crocetti E. “Using digital media or sleeping … that is the question”. A meta-analysis on digital media use and unhealthy sleep in adolescence. Comput Hum Behav 2023;146:107813. [doi: 10.1016/j.chb.2023.107813]
13. Berry D, Willoughby MT. On the practical interpretability of cross-lagged panel models: rethinking a developmental workhorse. Child Dev 2017;88(4):1186-1206. [doi: 10.1111/cdev.12660] [Medline: 27878996]
14. Hamaker EL, Kuiper RM, Grasman RPPP. A critique of the cross-lagged panel model. Psychol Methods 2015;20(1):102-116. [doi: 10.1037/a0038889] [Medline: 25822208]
15. Lee PH, Tse ACY, Wu CST, Mak YW, Lee U. Temporal association between objectively measured smartphone usage, sleep quality and physical activity among Chinese adolescents and young adults. J Sleep Res 2021;30(4):e13213. [doi: 10.1111/jsr.13213] [Medline: 33049798]
16. Brosnan B, Haszard JJ, Meredith-Jones KA, Wickham S, Galland BC, Taylor RW. Screen use at bedtime and sleep duration and quality among youths. JAMA Pediatr 2024;178(11):1147-1154. [doi: 10.1001/jamapediatrics.2024.2914] [Medline: 39226046]
17. Siebers T, Beyens I, Baumgartner SE, Valkenburg PM. Adolescents’ digital nightlife: the comparative effects of day- and nighttime smartphone use on sleep quality. Commun Res 2024:00936502241276793. [doi: 10.1177/00936502241276793]
18. Tkaczyk M, Lacko D, Elavsky S, Tancoš M, Smahel D. Are smartphones detrimental to adolescent sleep? An electronic diary study of evening smartphone use and sleep. Comput Hum Behav 2023;149:107946. [doi: 10.1016/j.chb.2023.107946]
19. Bartel KA, Gradisar M, Williamson P. Protective and risk factors for adolescent sleep: a meta-analytic review. Sleep Med Rev 2015;21:72-85. [doi: 10.1016/j.smrv.2014.08.002] [Medline: 25444442]
20. Gunnell KE, Flament MF, Buchholz A, Henderson KA, Obeid N, Schubert N, et al. Examining the bidirectional relationship between physical activity, screen time, and symptoms of anxiety and depression over time during adolescence. Prev Med 2016;88:147-152. [doi: 10.1016/j.ypmed.2016.04.002] [Medline: 27090920]
21. Khor SPH, McClure A, Aldridge G, Bei B, Yap MB. Modifiable parental factors in adolescent sleep: a systematic review and meta-analysis. Sleep Med Rev 2021;56:101408. [doi: 10.1016/j.smrv.2020.101408] [Medline: 33326915]
22. Short MA, Gradisar M, Lack LC, Wright HR, Dewald JF, Wolfson AR, et al. A cross-cultural comparison of sleep duration between US and Australian adolescents: the effect of school start time, parent-set bedtimes, and extracurricular load. Health Educ Behav 2013;40(3):323-330 [FREE Full text] [doi: 10.1177/1090198112451266] [Medline: 22984209]
23. Bauducco S, Pillion M, Bartel K, Reynolds C, Kahn M, Gradisar M. A bidirectional model of sleep and technology use: a theoretical review of how much, for whom, and which mechanisms. Sleep Med Rev 2024;76:101933 [FREE Full text] [doi: 10.1016/j.smrv.2024.101933] [Medline: 38657359]
24. Cain N, Gradisar M. Electronic media use and sleep in school-aged children and adolescents: a review. Sleep Med 2010;11(8):735-742. [doi: 10.1016/j.sleep.2010.02.006] [Medline: 20673649]
25. Lissak G. Adverse physiological and psychological effects of screen time on children and adolescents: literature review and case study. Environ Res 2018;164:149-157. [doi: 10.1016/j.envres.2018.01.015] [Medline: 29499467]
26. Van den Bulck J. Text messaging as a cause of sleep interruption in adolescents, evidence from a cross-sectional study. J Sleep Res 2003;12(3):263. [doi: 10.1046/j.1365-2869.2003.00362.x] [Medline: 12941066]
27. Hansen SL, Capener D, Daly C. Adolescent sleepiness: causes and consequences. Pediatr Ann 2017;46(9):e340-e344. [doi: 10.3928/19382359-20170816-01] [Medline: 28892550]
28. Crowley SJ, Acebo C, Carskadon MA. Sleep, circadian rhythms, and delayed phase in adolescence. Sleep Med 2007;8(6):602-612. [doi: 10.1016/j.sleep.2006.12.002] [Medline: 17383934]
29. Eggermont S, Van den Bulck J. Nodding off or switching off? The use of popular media as a sleep aid in secondary-school children. J Paediatr Child Health 2006;42(7-8):428-433. [doi: 10.1111/j.1440-1754.2006.00892.x] [Medline: 16898880]
30. Malheiros LEA, da Costa BG, Lopes MV, Chaput J, Silva KS. Association between physical activity, screen time activities, diet patterns and daytime sleepiness in a sample of Brazilian adolescents. Sleep Med 2021;78:1-6. [doi: 10.1016/j.sleep.2020.12.004] [Medline: 33370617]
31. Mazzer K, Bauducco S, Linton SJ, Boersma K. Longitudinal associations between time spent using technology and sleep duration among adolescents. J Adolesc 2018;66:112-119. [doi: 10.1016/j.adolescence.2018.05.004] [Medline: 29842997]
32. Poulain T, Vogel M, Buzek T, Genuneit J, Hiemisch A, Kiess W. Reciprocal longitudinal associations between adolescents' media consumption and sleep. Behav Sleep Med 2019;17(6):763-777. [doi: 10.1080/15402002.2018.1491851] [Medline: 30040503]
33. Maksniemi E, Hietajärvi L, Ketonen EE, Lonka K, Puukko K, Salmela-Aro K. Intraindividual associations between active social media use, exhaustion, and bedtime vary according to age-a longitudinal study across adolescence. J Adolesc 2022;94(3):401-414. [doi: 10.1002/jad.12033] [Medline: 35390194]
34. van der Schuur WA, Baumgartner SE, Sumter SR. Social media use, social media stress, and sleep: examining cross-sectional and longitudinal relationships in adolescents. Health Commun 2019;34(5):552-559 [FREE Full text] [doi: 10.1080/10410236.2017.1422101] [Medline: 29313723]
35. van der Schuur WA, Baumgartner SE, Sumter SR, Valkenburg PM. Media multitasking and sleep problems: a longitudinal study among adolescents. Comput Hum Behav 2018;81:316-324. [doi: 10.1016/j.chb.2017.12.024]
36. Tavernier R, Willoughby T. Sleep problems: predictor or outcome of media use among emerging adults at university? J Sleep Res 2014;23(4):389-396. [doi: 10.1111/jsr.12132] [Medline: 24552437]

XSL•FO
RenderX

37. Hale L, Kirschen GW, LeBourgeois MK, Gradisar M, Garrison MM, Montgomery-Downs H, et al. Youth screen media habits and sleep: sleep-friendly screen behavior recommendations for clinicians, educators, and parents. Child Adolesc Psychiatr Clin N Am 2018;27(2):229-245 [FREE Full text] [doi: 10.1016/j.chc.2017.11.014] [Medline: 29502749]
38. Grasaas E, Ostojic S, Jahre H. Adherence to sleep recommendations is associated with higher satisfaction with life among Norwegian adolescents. BMC Public Health 2024;24(1):1288 [FREE Full text] [doi: 10.1186/s12889-024-18725-1] [Medline: 38730403]
39. Martin KB, Bednarz JM, Aromataris EC. Interventions to control children's screen use and their effect on sleep: a systematic review and meta-analysis. J Sleep Res 2021;30(3):e13130. [doi: 10.1111/jsr.13130] [Medline: 32567219]
40. Gariepy G, Danna S, Gobiņa I, Rasmussen M, Gaspar de Matos M, Tynjälä J, et al. How are adolescents sleeping? Adolescent sleep patterns and sociodemographic differences in 24 European and North American countries. J Adolesc Health 2020;66(6S):S81-S88 [FREE Full text] [doi: 10.1016/j.jadohealth.2020.03.013] [Medline: 32446613]
41. Smahel D, Machackova H, Mascheroni G, Dedkova L, Staksrud E, Ólafsson K, EU Kids Online network. EU Kids Online 2020: survey results from 19 countries. EU Kids Online, The London School of Economics and Political Science 2020. [doi: 10.21953/lse.47fdeqj01ofo]
42. Smith C, de Wilde T, Taylor RW, Galland BC. Prebedtime screen use in adolescents: a survey of habits, barriers, and perceived acceptability of potential interventions. J Adolesc Health 2020;66(6):725-732. [doi: 10.1016/j.jadohealth.2019.12.007] [Medline: 32044232]
43. Moore M, Slane J, Mindell JA, Burt SA, Klump KL. Sleep problems and temperament in adolescents. Child Care Health Dev 2011;37(4):559-562 [FREE Full text] [doi: 10.1111/j.1365-2214.2010.01157.x] [Medline: 21083682]
44. Maslowsky J, Ozer EJ. Developmental trends in sleep duration in adolescence and young adulthood: evidence from a national United States sample. J Adolesc Health 2014;54(6):691-697 [FREE Full text] [doi: 10.1016/j.jadohealth.2013.10.201] [Medline: 24361237]
45. Gibson ES, Powles AP, Thabane L, O'Brien S, Molnar DS, Trajanovic N, et al. "Sleepiness" is serious in adolescence: two surveys of 3235 Canadian students. BMC Public Health 2006;6(1):116 [FREE Full text] [doi: 10.1186/1471-2458-6-116] [Medline: 16670019]
46. Atkin AJ, Sharp SJ, Corder K, Van Sluijs EMF. Prevalence and correlates of screen time in youth: an international perspective. Am J Prev Med 2014;47(6):803-807. [doi: 10.1016/j.amepre.2014.07.043] [Medline: 25241193]
47. von Elm E, Altman DG, Egger M. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Epidemiology 2007;18(6):800-804. [doi: 10.1097/ede.0b013e3181577654]
48. Tkaczyk M, Ksinan AJ, Smahel D. Supplementary Materials and Analyses for “Longitudinal Between- and Within-Person Associations of Screen Time, Bedtime, and Daytime Sleepiness Among Adolescents: A Three-Wave Prospective Panel Study.”. 2025. URL: https://osf.io/hsjp7/ [accessed 2025-12-17]
49. Drake C, Nickel C, Burduvali E, Roth T, Jefferson C, Badia P. Pediatric Daytime Sleepiness Scale. 2011. URL: https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft02761-000 [accessed 2025-12-17]
50. Yuan KH, Bentler PM. Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. Sociological Methodology 2000;30(1):165-200. [doi: 10.1111/0081-1750.00078]
51. Mulder JD, Hamaker EL. Three extensions of the random intercept cross-lagged panel model. Struct Equ Modeling 2020;28(4):638-648. [doi: 10.1080/10705511.2020.1784738]
52. Stoel RD, Garre FG, Dolan C, van den Wittenboer G. On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. Psychol Methods 2006;11(4):439-455. [doi: 10.1037/1082-989x.11.4.439]
53. Campbell IG, Burright CS, Kraus AM, Grimm KJ, Feinberg I. Daytime sleepiness increases with age in early adolescence: a sleep restriction dose-response study. Sleep 2017;40(5):zsx046 [FREE Full text] [doi: 10.1093/sleep/zsx046] [Medline: 28419388]
54. Vollmer C, Jankowski KS, Díaz-Morales JF, Itzek-Greulich H, Wüst-Ackermann P, Randler C. Morningness-eveningness correlates with sleep time, quality, and hygiene in secondary school students: a multilevel analysis. Sleep Med 2017;30:151-159. [doi: 10.1016/j.sleep.2016.09.022] [Medline: 28215240]
55. Brand S, Hatzinger M, Beck J, Holsboer-Trachsler E. Perceived parenting styles, personality traits and sleep patterns in adolescents. J Adolesc 2009;32(5):1189-1207. [doi: 10.1016/j.adolescence.2009.01.010] [Medline: 19237190]
56. Lee HE, Kim JY, Kim C. The influence of parent media use, parent attitude on media, and parenting style on children's media use. Children (Basel) 2022;9(1):37 [FREE Full text] [doi: 10.3390/children9010037] [Medline: 35053662]
57. Buxton OM, Chang AM, Spilsbury JC, Bos T, Emsellem H, Knutson KL. Sleep in the modern family: protective family routines for child and adolescent sleep. Sleep Health 2015;1(1):15-27 [FREE Full text] [doi: 10.1016/j.sleh.2014.12.002] [Medline: 26779564]
58. Pedersen J, Rasmussen MG, Olesen LG, Klakk H, Kristensen PL, Grøntved A. Recreational screen media use in Danish school-aged children and the role of parental education, family structures, and household screen media rules. Prev Med 2022;155:106908 [FREE Full text] [doi: 10.1016/j.ypmed.2021.106908] [Medline: 34915040]
59. Orth U, Meier LL, Bühler JL, Dapp LC, Krauss S, Messerli D, et al. Effect size guidelines for cross-lagged effects. Psychol Methods 2024;29(2):421-433 [FREE Full text] [doi: 10.1037/met0000499] [Medline: 35737548]

XSL•FO
RenderX

60. Kiss O, Nagata JM, de Zambotti M, Dick AS, Marshall AT, Sowell ER, et al. Effects of the COVID-19 pandemic on screen time and sleep in early adolescents. Health Psychol 2023;42(12):894-903 [FREE Full text] [doi: 10.1037/hea0001251] [Medline: 36972087]
61. Bauducco SV, Flink IK, Jansson-Fröjmark M, Linton SJ. Sleep duration and patterns in adolescents: correlates and the role of daily stressors. Sleep Health 2016;2(3):211-218. [doi: 10.1016/j.sleh.2016.05.006] [Medline: 29073425]
62. Gába A, Dygrýn J, Štefelová N, Rubín L, Hron K, Jakubec L, et al. How do short sleepers use extra waking hours? A compositional analysis of 24-h time-use patterns among children and adolescents. Int J Behav Nutr Phys Act 2020;17(1):104 [FREE Full text] [doi: 10.1186/s12966-020-01004-8] [Medline: 32795287]
63. Asarnow LD, McGlinchey E, Harvey AG. The effects of bedtime and sleep duration on academic and emotional outcomes in a nationally representative sample of adolescents. J Adolesc Health 2014;54(3):350-356 [FREE Full text] [doi: 10.1016/j.jadohealth.2013.09.004] [Medline: 24225447]
64. Bartel K, Scheeren R, Gradisar M. Altering adolescents' pre-bedtime phone use to achieve better sleep health. Health Commun 2019;34(4):456-462. [doi: 10.1080/10410236.2017.1422099] [Medline: 29313721]
65. Duraccio KM, Zaugg KK, Blackburn RC, Jensen CD. Does iPhone night shift mitigate negative effects of smartphone use on sleep outcomes in emerging adults? Sleep Health 2021;7(4):478-484. [doi: 10.1016/j.sleh.2021.03.005] [Medline: 33867308]
66. Owens JA, Dearth-Wesley T, Lewin D, Gioia G, Whitaker RC. Self-regulation and sleep duration, sleepiness, and chronotype in adolescents. Pediatrics 2016;138(6):e20161406. [doi: 10.1542/peds.2016-1406] [Medline: 27940688]
67. Parry DA, Davidson BI, Sewall CJR, Fisher JT, Mieczkowski H, Quintana DS. A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. Nat Hum Behav 2021;5(11):1535-1547. [doi: 10.1038/s41562-021-01117-5] [Medline: 34002052]
68. Lauderdale DS, Knutson KL, Yan LL, Liu K, Rathouz PJ. Self-reported and measured sleep duration: how similar are they? Epidemiology 2008;19(6):838-845. [doi: 10.1097/ede.0b013e318187a7b0]
69. Stier S, Breuer J, Siegers P, Thorson K. Integrating survey data and digital trace data: key issues in developing an emerging field. Soc Sci Comput Rev 2019;38(5):503-516. [doi: 10.1177/0894439319843669]
70. Doherty A, Jackson D, Hammerla N, Plötz T, Olivier P, Granat MH, et al. Large scale population assessment of physical activity using wrist worn accelerometers: the UK Biobank Study. PLoS One 2017;12(2):e0169649 [FREE Full text] [doi: 10.1371/journal.pone.0169649] [Medline: 28146576]
71. Kaye KL, Orben A, Ellis AD, Hunter CS, Houghton S. The conceptual and methodological mayhem of "Screen Time". Int J Environ Res Public Health 2020;17(10):3661 [FREE Full text] [doi: 10.3390/ijerph17103661] [Medline: 32456054]
72. Sumter SR, Baumgartner SE, Wiradhany W. Beyond screentime: a 7-day mobile tracking study among college students to disentangle smartphone screentime and content effects on sleep. Behav Inf Technol 2024;44(6):1260-1276. [doi: 10.1080/0144929x.2024.2350663]
73. Livingstone S, Pothong K. Beyond screen time: Rethinking children's play in a digital world. J Health Visiting 2022;10(1):32-38. [doi: 10.12968/johv.2022.10.1.32]
74. Orth U, Clark DA, Donnellan MB, Robins RW. Testing prospective effects in longitudinal research: comparing seven competing cross-lagged models. J Pers Soc Psychol 2021;120(4):1013-1034 [FREE Full text] [doi: 10.1037/pspp0000358] [Medline: 32730068]
75. Devís-Devís J, Peiró-Velert C, Beltrán-Carrillo VJ, Tomás JM. Screen media time usage of 12-16 year-old Spanish school adolescents: effects of personal and socioeconomic factors, season and type of day. J Adolesc 2009;32(2):213-231. [doi: 10.1016/j.adolescence.2008.04.004] [Medline: 18694592]
76. Quante M, Wang R, Weng J, Kaplan ER, Rueschman M, Taveras EM, et al. Seasonal and weather variation of sleep and physical activity in 12-14-year-old children. Behav Sleep Med 2019;17(4):398-410 [FREE Full text] [doi: 10.1080/15402002.2017.1376206] [Medline: 28922020]
77. Kocar S, Biddle N. The power of online panel paradata to predict unit nonresponse and voluntary attrition in a longitudinal design. Qual Quant 2023;57(2):1055-1078 [FREE Full text] [doi: 10.1007/s11135-022-01385-x] [Medline: 35493336]
78. Lugtig P, Das M, Scherpenzeel A. In: Callegaro M, Baker R, Bethlehem J, Göritz AS, Krosnick JA, Lavrakas PJ, editors. Nonresponse and Attrition in A Probability‐Based Online Panel for the General Population. USA: Wiley; 2014:135-153.
79. Matthes J, Koban K, Neureiter A, Stevic A. Longitudinal relationships among fear of COVID-19, smartphone online self-disclosure, happiness, and psychological well-being: Survey study. J Med Internet Res 2021;23(9):e28700 [FREE Full text] [doi: 10.2196/28700] [Medline: 34519657]
80. Madigan S, Eirich R, Pador P, McArthur BA, Neville RD. Assessment of changes in child and adolescent screen time during the COVID-19 pandemic: a systematic review and meta-analysis. JAMA Pediatr 2022;176(12):1188-1198 [FREE Full text] [doi: 10.1001/jamapediatrics.2022.4116] [Medline: 36342702]
81. Measures Adopted by the Czech Government Against the Coronavirus.: Government of the Czech Republic URL: https://vlada.gov.cz/en/media-centrum/aktualne/measures-adopted-by-the-czech-government-against-coronavirus-180545/ [accessed 2025-09-22]
82. ChatGPT Large Language Model.: OpenAI; 2025. URL: https://chat.openai.com/chat [accessed 2025-12-17]

XSL•FO
RenderX

Abbreviations

CFI: Comparative Fit Index
CLPM: cross-lagged panel model
FIML: full information maximum likelihood
ICC: intraclass correlation coefficient
MCAR: Missing Completely at Random
MLR: maximum likelihood estimator
RI-CLPM: random intercept cross-lagged panel model
RMSEA: root-mean-square error of approximation
SRMR: standardized root-mean-square residual
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TLI: Tucker-Lewis Index

Edited by S Brini; submitted 13.Jun.2025; peer-reviewed by R Taylor, T Poulain; comments to author 08.Sep.2025; revised version received 01.Dec.2025; accepted 02.Dec.2025; published 21.Jan.2026.

Please cite as:
Tkaczyk M, Ksinan AJ, Smahel D
Longitudinal Between- and Within-Person Associations Among Screen Time, Bedtime, and Daytime Sleepiness Among Adolescents: Three-Wave Prospective Panel Study
J Med Internet Res 2026;28:e78972
URL: https://www.jmir.org/2026/1/e78972
doi:10.2196/78972
PMID:

©Michał Tkaczyk, Albert J Ksinan, David Smahel. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Tailored Internet-Delivered Mindfulness-Based Interventions for Patients With Hepatocellular Carcinoma After Transarterial Chemoembolization: Qualitative Study

Zengxia Liu[1,2,3], PhD; Min Li[4], MS; Yong Jia[3], PhD; Li Chen[3], Prof Dr

[1]School of Nursing, Anhui Medical University, Hefei, China

[2]Department of Nursing, First Affiliated Hospital of Anhui Medical University, Hefei, China

[3]School of Nursing, Jilin University, Changchun, Jilin, China

[4]Department of Intervention, First Hospital of Jilin University, Changchun, China

Corresponding Author:
Li Chen, Prof Dr
School of Nursing
Jilin University
965 Xinjiang Street
Changchun, 130021
China
Phone: 1 15943033929
Fax: 1 043185619366
Email: chen_care@126.com

Abstract

Background: Patients with hepatocellular carcinoma (HCC) undergoing transarterial chemoembolization (TACE) experience significant psychological distress, impacting outcomes. While mindfulness-based interventions (MBIs) are beneficial, access is limited. Internet-delivered MBIs (iMBIs) offer an accessible alternative; yet, qualitative understanding of patient experiences with tailored iMBIs for this specific population is lacking.

Objective: This study aimed to explore the facilitators and barriers of patients with HCC post TACE and participated in tailored iMBIs.

Methods: From November 2020 to December 2022, 11 patients with HCC post TACE who had taken part in tailored iMBIs were purposively recruited from a tertiary hospital in Jilin Province. Data were collected through semistructured interviews lasting 30-60 minutes. The interviews were analyzed using conventional content analysis.

Results: Five main categories emerged from the analysis: (1) mindfulness mindset, including acceptance, calmness, and mood improvement; (2) improvement of physical discomfort, such as better sleep, pain relief, reduced gastrointestinal symptoms, and increased activity levels; (3) resistance to mindfulness practice, including perceived lack of effectiveness, unsuitable conditions, equipment limitations, and difficulty concentrating; (4) support and encouragement, involving social support, supervision, and professional guidance; and (5) accessibility and convenience characterized by restoration of life balance and user-friendly features of the practice. Each category encompassed several subcategories reflecting the diverse experiences of participants.

Conclusions: While iMBIs were generally perceived as convenient and accessible, challenges such as equipment limitations were noted. Future implementation should focus on enhancing supportive factors to improve adherence, minimizing barriers, and refining the design and delivery of iMBI programs.

Trial Registration: Chinese Clinical Trial Registry ChiCTR1900027976; https://www.chictr.org.cn/showproj.html?proj=46657

(J Med Internet Res 2026;28:e78337) doi:10.2196/78337

KEYWORDS

internet; mindfulness; hepatocellular carcinoma; cancer; qualitative study

XSL•FO
RenderX

Introduction

Primary liver cancer is the third leading cause of cancer-related death worldwide and one of the most prevalent malignancies. China accounts for 45.27% of global cases and 47.12% of liver cancer–related deaths [1]. In 2022, there were approximately 368,000 new cases and 317,000 deaths from liver cancer in China, ranking fourth in cancer incidence and second in cancer-related mortality [2]. Hepatocellular carcinoma (HCC) is the most common histological subtype of primary liver cancer, comprising 75%-85% of all cases [3]. Transarterial chemoembolization (TACE) is a widely accepted local treatment for HCC and is the preferred option for patients with intermediate-stage HCC who are not suitable candidates for surgery [4,5]. However, patients often require multiple TACE sessions to manage disease progression. Despite treatment, they continue to face high risks of recurrence, metastasis, and various complications [6]. Moreover, the effectiveness of TACE diminishes with repeated use, and frequent treatments can result in progressive liver damage [7].

Due to the high mortality rate of HCC, research showed that the 5-year survival rate of patients with HCC after diagnosis was only 19.5% [8].Therefore, most patients with HCC face a serious threat of death, and their psychological distress is generally higher than that of other patients with cancer. During TACE treatment, some adverse reactions may occur, such as nausea, vomiting, fever, and pain. TACE patients have to face not only the physical pain brought by the disease and treatment but also the fear of disease recurrence, the high medical costs of repeated TACE treatments, and the pressure of being unable to return to society. This causes them to bear tremendous mental stress, resulting in widespread and severe psychological distress among patients with HCC post TACE [9,10]. This psychological burden can weaken patients' immune response, reduce treatment compliance, and adversely affect treatment outcomes, ultimately increasing the likelihood of tumor recurrence, deterioration, and metastasis [11].

Psychological distress in patients with cancer often arises not from a single external stressor that can be resolved or avoided but from the persistent fear of recurrence. Coping effectively, therefore, requires adapting to this ongoing internal experience. Mindfulness-based interventions (MBIs), which are designed to help individuals manage distressing thoughts and emotions, are particularly well suited for this purpose. The effectiveness of MBIs in alleviating psychological distress among patients with cancer has been well documented [12,13]. However, patients with HCC post TACE face unique physiological and psychological challenges. The uncertainty surrounding treatment outcomes, fear of recurrence, financial burden of repeated treatments, and treatment-related side effects collectively inflict significant physical and emotional strain. These factors highlight the urgent need to develop tailored MBIs specifically for this population—interventions that take into account not only the psychological profile and coping mechanisms of patients with HCC but also their physical health status. Despite the proven benefits of MBIs, several barriers limit access, including a shortage of trained therapists, high costs, and practical constraints, such as limited mobility and time availability among patients with cancer [14]. Addressing these challenges is essential to ensure that MBIs are accessible and effective for those who need them most.

Internet-delivered mindfulness-based interventions (iMBIs) offer a promising solution to many of the barriers associated with accessing psychosocial care in cancer treatment settings. Compared with traditional face-to-face interventions, iMBIs have several advantages. They can be accessed online at any time, allowing participants to engage with the program at their own convenience. This flexibility is particularly beneficial for patients with HCC and helps overcome practical challenges such as geographical distance, transportation difficulties, cancer-related fatigue, and mobility issues. This flexibility is particularly beneficial for patients with HCC post TACE, as most of them have a heavy burden of physical symptoms. This method helps overcome practical challenges such as geographic distance, transportation difficulties, cancer-related fatigue, and limited mobility. While iMBIs are associated with a higher rate of participant dropout, studies have nonetheless demonstrated their effectiveness in reducing psychological distress among patients with cancer [15-17].

Most studies investigating iMBIs in patients with cancer rely on standardized questionnaires to assess outcomes [18-20]. However, such quantitative approaches fail to capture the personal experiences and perceptions of participants. These subjective experiences are difficult to evaluate through usage data or survey metrics alone. Therefore, the primary aim of this qualitative study was to explore participants' experiences and perceived effects of tailored iMBIs. This approach allows for a deeper understanding of their thoughts and feelings, a comprehensive examination of both the benefits and the challenges of the intervention, and an analysis of factors that influence participant adherence and engagement.

Methods

Research Design

To gain a deep understanding of participants' experiences with iMBIs, as well as the interventions' acceptability and applicability, a qualitative descriptive design using semistructured personal interviews was used. The study followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) [21].

Participants

From November 2020 to December 2022, the study was conducted at the First Hospital of Jilin University. Potential participants were identified by evaluating and screening patient medical records to determine eligibility based on the inclusion criteria. The inclusion criteria for participants were as follows: (1) aged 18 years or older; (2) diagnosed with HCC based on the diagnostic criteria of the European Association for the Study of the Liver [22]; (3) currently received TACE; (4) able to operate a smartphone and use WeChat frequently (at least 5 times per week); (5) able to read, write, and speak Chinese; and (6) provided informed consent and voluntarily agreed to participate in the study. The exclusion criteria for participants were as follows: (1) diagnosed with other types of tumors or

XSL•FO
RenderX

mental stress disorders, (2) taking psychotropic medications during the study period, (3) received mindfulness intervention, and (4) experiencing poor health that interferes with normal communication. Those who met the criteria were invited to participate in a mindfulness intervention. Interviewees were then selected from among the participants who completed the iMBIs intervention.

In accordance with the principle of maximum variation sampling, participants were chosen to reflect a wide range of differences in age, gender, education level, and frequency of TACE treatments. The sample size was determined according to the concept of data saturation, meaning that collection ceased once no new codes emerged and the data appeared to have stabilized. After 11 patients were interviewed and the data were analyzed, the research team agreed that data saturation had been reached, and data collection was stopped.

Interventions

Overview

A cross-sectional survey of patients post TACE revealed that trait mindfulness, perceived stress, and experiential avoidance are potential psychological mechanisms influencing anxiety and depression in patients with HCC [10]. Based on this finding, the research team developed a tailored MBI for patients with HCC post TACE. The intervention was informed by the core principles of Mindfulness-Based Stress Reduction (MBSR) [23], Acceptance and Commitment Therapy [24,25], and Mindfulness-Based Cancer Recovery, as well as insights from a systematic review and an assessment of the characteristics and internet usage habits of this patient population. An initial draft of the tailored MBI was created and then refined through expert consultation and revisions.

A self-guided intervention was delivered through WeChat groups and an official WeChat account, which was used to distribute weekly mindfulness content. This content was presented in various formats, including text, audio, video, and images, via the official WeChat official account. The interface of the WeChat official account for the intervention content is shown in Figure 1. The intervention lasted for 6 weeks and included an initial in-hospital face-to-face session during the first week, followed by web-based sessions from the second to the sixth week.

Figure 1. The interface of the WeChat official account for the intervention content.

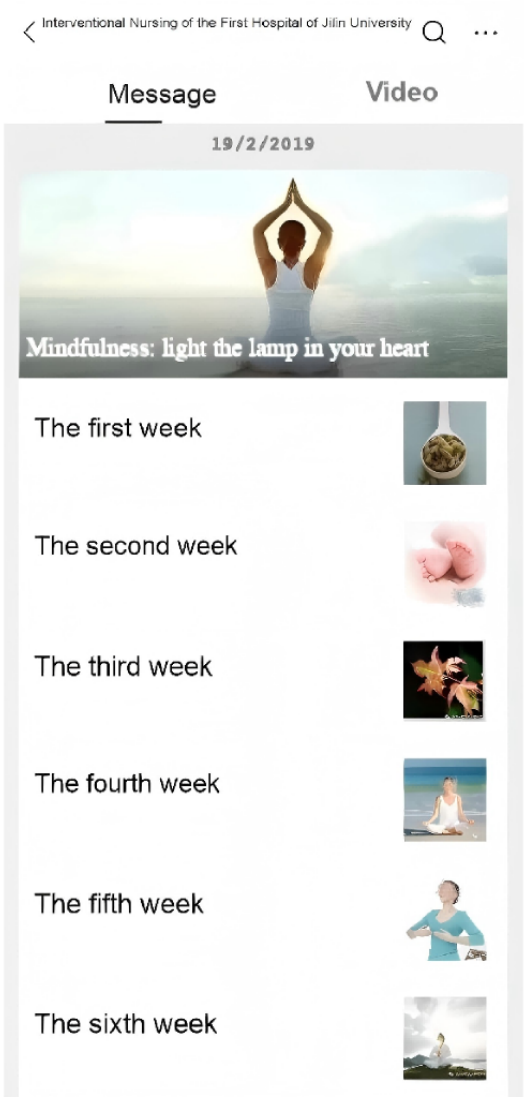

XSL•FO
RenderX

In-Hospital Face-to-Face Intervention (Week 1)

The face-to-face sessions were conducted in the interventional department classroom, with 5-8 participants per group. Each session lasted 1-1.5 hours. During this session, the research team introduced the study's purpose and content using presentation slides, provided an overview of mindfulness and its benefits, and helped participants understand the attitudes and principles underlying mindfulness practice. The session also familiarized participants with the structure and content of the upcoming web-based intervention, ensuring that they were prepared for the self-guided practice in the following weeks.

Out-of-Hospital Web-Based Intervention (Weeks 2-6)

For the remainder of the intervention, content was delivered weekly through WeChat in the form of audio, video, or image-based materials. Participants were encouraged to engage in daily mindfulness practice for 10-20 minutes, 5 days per week. The research team provided regular reminders and encouragement via WeChat and phone calls to ensure adherence. Participants were also encouraged to "check in" daily in the WeChat group by sharing their experiences and reflections on mindfulness practice. In addition, the research team shared educational materials through the WeChat group, covering topics such as mindfulness, HCC, and TACE treatment. They also responded to participants' questions regarding the disease, mindfulness practices, and other related concerns. The specific intervention content is detailed in Table 1.

Before interventions, 5 patients were invited to conduct a pre-experiment. After the first 2 weeks of intervention, adjustments were made according to the participants' feedback. Participants reported that the text on the official account interface was too small, which was not conducive to reading. Therefore, the researchers reorganized the layout to make the interface clearer and more esthetically pleasing. Participants also mentioned that there was too much text content, which was not conducive to reading and practice. As a result, the researchers converted some of the text content into audio to facilitate participants' learning and practice. Participants suggested that more pictures and videos could be added to make the content more vivid. The researchers added some pictures and video content to make the content more vivid and practical.

XSL•FO
RenderX

Table 1. The tailored internet-delivered mindfulness-based interventions for patients with HCC[a] post TACE[b].

Week	Theme of intervention	Summary of contents	Homework	Intervention mode
Week 1	Mindfulness and HCC and self-acceptance	1. Knowledge introduction: Introducing content related to mindfulness, HCC and TACE treatment, and the relationship between mindfulness and cancer. 2. Mindful breathing: Guide participants to practice mindful breathing, exhaling and inhaling easily, experiencing every detail of breathing, including the process, changes and pauses of breathing, accepting their breathing state, and being aware of the changes in their thoughts and attention. 3. "The Missing Piece": Watch the video "The Missing Corner" to prompt participants to think about "unchangeable imperfections," learn to accept negative events or emotions in life, and encourage participants to share their feelings after watching.	Mindful breathing	Face-to-face intervention in the hospital, 5-8 participants per group, 1-1.5 hours per session, introducing the purpose and content of the study, and guiding the completion of the first session.
Week 2	Cancer and stress and cognitive dissociation	1. Knowledge introduction: Describe knowledge related to cancer and stress, and typical stress reactions to cancer stress events. 2. Body scan: Introduce the connotation and requirements of body scan. Guide participants to feel and experience the sensation of each part of the body in order from feet to head. Pay attention to the feelings of each part, whether it is comfortable or not. Do not judge. Emphasize full commitment to the body-scanning practice. 3. "I have such an idea now": Guide participants to practice "I have such an idea now." Ask the participants to state a thought that makes them depressed, such as "Why did I get liver cancer?" Repeat this several times. After 3 minutes, add "I have an idea. Why did I get liver cancer?" before this thought. Repeat this several times. After 3 minutes, add a sentence before this idea, "I noticed that I have an idea. Why did I get liver cancer?" The main purpose of this exercise is to distinguish the real self from the imagined self, achieve cognitive dissociation, and prevent these thoughts from affecting an individual's life.	Body scan	Out-of-hospital online intervention, 20 minutes per day, at least 5 days per week, learn the intervention content via official account and WeChat, and 10 minutes of homework practice.
Week 3	Identifying avoidance reactions and self-awareness	1. Meditation practice: Guide participants to practice mindful meditation. Be aware of the thoughts and emotions that enter your mind. Pay attention to these thoughts and emotions but do not focus too much on them. Try to let go and shift your attention to your breathing. Then slowly shift your attention to your body, voice, thoughts, and emotions, and return to your breathing. In this repetitive process, gradually become aware of and feel your thoughts and emotions, and gradually and spontaneously coexist with them. 2. "Chessboard metaphor": Guide participants to imagine a "chess game." No longer caring about winning or losing, just providing a venue for the game of chess, as an observer, to observe the competition. Through the "chessboard metaphor," participants are guided to stop getting entangled in their inner struggles, become observers of their own psychology, and achieve self-awareness.	Meditation practice	Out-of-hospital online intervention, 20 minutes per day, at least 5 days per week, learn the intervention content via official account and WeChat, and 10 minutes of homework practice.
Week 4	Practice mindfulness in life, focusing on the present	1. Mindfulness walking: Guide patients to practice mindfulness walking and feel the current movements while practicing. From the lifting of the foot to the forward movement and the landing process, the sensation of each part of the foot, when practicing, feel the current movement and the current feeling, and focus your attention on the action of mindful walking. 2. Mindfulness eating: Guide patients to practice mindfulness eating. Slow down your eating speed, focus on the movements and sensations of your tongue and mouth while eating, experience the changes and flavors of food in your mouth, and pay attention to observing your inner feelings and emotions during the practice.	Mindful walking	Out-of-hospital online intervention, 20 minutes per day, at least 5 days per week, learn the intervention content via official account and WeChat, and 10 minutes of homework practice.

XSL•FO
RenderX

Week	Theme of intervention	Summary of contents	Homework	Intervention mode
Week 5	Nonselective awareness, clear value	1. Awareness practice: Explain the key points of awareness practice to patients and be aware of everything around them. Perceive the things around you and the sensations of your body with an accepting attitude, including pain, fatigue, and so on, and embrace everything. In practice, no matter what you perceive, trust and accept yourself. 2. "Compass" metaphor: Guide participants to think about the guiding role of the "compass" in the journey. By using this metaphor for goals and values, it guides participants to think about the direction and goals they want to move forward in, and based on their own values, they choose to set a small goal and take action.	Awareness practice	Out-of-hospital online intervention, 20 minutes per day, at least 5 days per week, learn the intervention content via official account and WeChat, and 10 minutes of homework practice.
Week 6	Grateful life, committed to action	1. Love meditation, good wishes: Cultivate the mindfulness practice of kindness and bestow your blessings upon others during the practice. You can choose different blessings in sequence to send to different people, including yourself, your family, friends, or even someone you dislike. Fill the participants' hearts with tolerance and love. 2. "Passengers on the Bus": Through the short film "Passenger Practice on the Bus," participants are guided to compare their own lives to buses and think about how to keep moving forward on the path of happiness when encountering different passengers, such as "stress," "anxiety," "happiness," and so forth, thereby contemplating the significance of committing to action.	Love meditation	Out-of-hospital online intervention, 20 minutes per day, at least 5 days per week, learn the intervention content via official account and WeChat, and 10 minutes of homework practice.

[a]HCC: hepatocellular carcinoma.

[b]TACE: transarterial chemoembolization.

Data Collection

Based on the disease characteristics and treatment experiences of patients with HCC following TACE, a preliminary interview guide was developed through a literature review and discussions among the research team. Prior to the formal interviews, 2 participants were selected for pilot interviews. The final interview guide was revised based on the issues identified during these pilot sessions and is shown in Table 2. Demographic information was collected before the interviews.

Two researchers (ZL and YJ), graduate students pursuing a Doctor of Nursing degree, approached potential participants during conducting the research. Once the patients had completed their consultations and treatments, we invited them to participate in this study. If they agreed to participate, we clearly explained the study's purpose and procedures to them. The researchers coordinated with participants in advance to schedule interviews and conducted them in a private room at the hospital to ensure confidentiality. At the beginning of each session, the researchers introduced themselves and explained the purpose, methods, and content of the interview. Participants were informed that the interview would be audio-recorded to ensure the completeness and accuracy of the data. Informed consent was obtained prior to starting the interview. During the interviews, the researchers maintained a neutral and nondirective attitude, using appropriate and nonleading language to avoid influencing participants' responses. Data collection and analysis were carried out concurrently to support iterative refinement of the findings.

During the interview, we used the interview tools proposed by Robinson [26], which include descriptive, individualized memory, exploratory, and clarifying inquiries. The individualized memory inquiry was used to guide the participants in recalling specific periods to obtain detailed information, such as the duration of their participation in iMBIs. The descriptive inquiry was used to explore the participants' feelings of participation. The interpretive inquiry was used to reveal the participants' views on iMBIs. The clarifying inquiry was used to clarify keywords and the implicit meanings of expressions. Based on the participants' responses, we modified or omitted the interview questions, or introduced new questions to explore new emerging topics, such as the influence of family on the participants' participation in iMBIs.

Table 2. Interview outline.

Number	Question
1	Why did you participate in this network-based mindfulness training?
2	What are your experiences while participating in this training?
3	What effects does network-based mindfulness training have on your body?
4	How does the network-based mindfulness training affect you psychology?

XSL•FO
RenderX

Ethical Considerations

The study was executed in accordance with the principles of the Declaration of Helsinki. The research protocol was approved by the ethics committee of the School of Nursing at Jilin University (approval no. 2019112001) and was registered with the Chinese Clinical Trial Registration Center (ChiCTR1900027976) on December 7, 2019. Prior to the interviews, participants were informed about the purpose and content of the study. They were assured that all data would be used solely for research and publication purposes, and that their personal information would remain confidential. Participants were also informed of their right to withdraw from the study at any time without any consequences. Written informed consent was obtained from all participants before the study commenced.

Data Analysis

One researcher (LZ) participating in the interview listened to the recordings repeatedly within 24 hours after each interview and transcribed them into text. Another researcher (LM) participating in the interview checked the transcribed text and imported it into NVivo 11 software [27] for analysis. Conventional content analysis [28,29] was used to analyze the data. Three researchers (LZ, JY, and CL) independently analyzed and transcribed the data.

Methodological Rigor and Trustworthiness of Data

Members of the research team include liver cancer clinical experts, nursing specialists, and psychologists. Interviews were conducted by 2 researchers (LZ and LM) to avoid information omission. To establish credibility, we used both analyst triangulation and data triangulation. Analyst triangulation was used to ensure intersubjective stability of the results and involved the independent analysis of the data by 3 researchers (LZ, JY, and CL), followed by a comparison of the analysis. The data were collated and analyzed by experienced researchers in qualitative research, the speech information of the interviewees was comprehensively retained, and complete and detailed textual data were established. The 2 researchers (LZ and JY) organized, analyzed, and transcribed the data independently. If there were differences, they were resolved through discussion or consultation with the third researcher (CL) until the results were reached.

Results

Characteristics of Participants

In this study, interviews were conducted with 9 participants, during which no new information emerged. To confirm data saturation, interviews with 2 additional participants were carried out, and again, no new themes were identified. Thus, a total of 11 participants were included in the final analysis. Their demographic and clinical characteristics are shown in Table 3.

Table 3. Characteristics of semidepth interview participants (N=11).

Number	Age (years)	Sex	Education level	Number of TACEs[a]
1	38	Male	Senior high school	1
2	50	Male	College school	8
3	53	Female	Senior high school	1
4	42	Female	College school	4
5	61	Male	Primary school	5
6	57	Female	Senior high school	1
7	49	Female	College school	7
8	62	Male	Senior high school	3
9	63	Female	Senior high school	3
10	48	Male	College school	7
11	40	Male	Primary school	4

[a]TACE: transarterial chemoembolization.

Categories

Overview

Through repeated analysis and organization of the interview data, 5 main categories reflecting the participants' experiences were identified: mindfulness mindset, improvement of physical discomfort, resistance to mindfulness practice, support and encouragement, and accessibility and convenience. Each main category is further divided into several subcategories, as shown in Textbox 1.

XSL•FO
RenderX

Textbox 1. Relevant categories corresponding to experiences of participants.

1. **Mindful mindset**
   - 1.1 Acceptance
   - 1.2 Calm
   - 1.3 Mood improvement
     - 1.3.1 Relaxation
     - 1.3.2 Gratitude
2. **Improvement of physical discomfort**
   - 2.1 Improve sleep
   - 2.2 Pain relief
   - 2.3 Reduce gastrointestinal symptoms
   - 2.4 Increase activities
3. **Mindfulness practice resistance**
   - 3.1 Lack of effect
   - 3.2 Condition
     - 3.2.1 Fatigue
     - 3.2.2 Pain
   - 3.3 Device usage restrictions
   - 3.4 Difficulty focusing
   - 3.5 Lack of motivation to practice
4. **Support and encourage**
   - 4.1 Social support
   - 4.2 Supervision and guidance
5. **Accessibility and convenience**
   - 5.1 Restore a sense of balance in your life
   - 5.2 Practice features

Main Category 1: Mindfulness Mindset

The mindfulness mindset refers to the ability of patients with HCC following TACE to face their condition with a nonjudgmental attitude, while maintaining a curious, open, and accepting approach toward their present thoughts and emotions.

General Category 1.1: Acceptance

Six participants reported that practicing mindfulness helped them adopt a more accepting attitude toward their illness and the challenges they faced in life.

> *My mother and I both have hepatitis, and my mother died of HCC, so I've always been worried about my health. I get scared every time I go for a check-up. But after doing the mindfulness exercises, my mindset changed. I gradually realized that I shouldn't overthink things every day. Since I have this disease, I just need to follow the doctor's advice and receive treatment. Many treatments are effective, but overthinking isn't good for my health.* [Interview ig4]

General Category 1.2: Calm

Seven participants reported that after practicing mindfulness, they experienced a greater sense of inner calm and became more attuned to the beauty in their surroundings.

> *My mindset has improved a lot. I can now perceive things around me more calmly and appreciate the little moments in life. I'm more willing to take walks, enjoy the scenery, and notice the beauty in simple things, like a tree or a flower by the roadside.* [Interview ig9]

General Category 1.3: Mood Improvement

Subcategory 1.3.1: Relaxation

Five participants shared that mindfulness practice helped them better regulate their emotions, reduce anxiety, and achieve a more relaxed state of mind.

XSL•FO
RenderX

> *Every time I went for a check-up, I would get really nervous if there was any change in my test results. Even small changes would worry me for a long time. Even when the doctor reassured me that everything was fine, I would still look up all kinds of information online. Now, when I notice myself feeling anxious, I try to adjust my emotions, relax, and avoid overthinking things that haven't even happened.* [Interview ig4]

Subcategory 1.3.2: Gratitude

Two participants expressed gratitude. After practicing mindfulness, they were able to approach people and situations around them with a more thankful attitude.

> *I've always had a bad temper and often got angry. My family would give in to me because of my poor health, which made me even angrier. After listening to the exercises, my mindset gradually improved. Now, when I encounter unpleasant things, I can stay calm. I also recognize how much my family has sacrificed for me, and my attitude has changed a lot. The atmosphere at home is much better.* [Interview ig3]

Main Category 2: Improvement of Physical Discomfort

Improvement of physical discomfort refers to the reduction of physical symptoms and alleviation of discomfort experienced by patients with HCC after TACE.

General Category 2.1: Improve Sleep

Ten participants reported significant improvements in sleep quality after mindfulness practice, with less presleep rumination and easier time falling asleep.

> *I've always had poor sleep, often lying in bed for a long time without falling asleep. The worse my sleep, the more I get lost in my thoughts. But since I started these exercises, I fall asleep shortly after listening. It helps stop my random thoughts.* [Interview ig3]

General Category 2.2: Pain Relief

Three participants said that mindfulness helped relieve pain and reduce physical discomfort.

> *I used to feel discomfort all over my body and pain everywhere. During the exercises, I take deep breaths, relax, follow the teacher, and slowly let go of tension. These uncomfortable feelings fade away, and I feel much better.* [Interview ig6]

General Category 2.3: Alleviation of Gastrointestinal Symptoms

Two participants reported that mindfulness practice helped ease gastrointestinal symptoms, improved their appetite, and reduced bloating.

> *I often had no appetite, felt uncomfortable after eating, and experienced bloating. Now I do these exercises daily, and these discomforts have improved. Sitting and relaxing each day reduces my irritability and unease.* [Interview ig1]

General Category 2.4: Increase Activities

Two participants noted that mindfulness helped reduce fatigue and encouraged them to be more active.

> *My whole body felt weak. I would just sit or lie down, unwilling to do anything. Sometimes I felt exhausted for long periods. After starting these exercises and practicing regularly, I feel much better. I'm more willing to move and have things to do.* [Interview ig4]

Main Category 3: Resistance to Mindfulness Practice

Resistance refers to factors that hinder patients' ability to consistently participate in mindfulness practice during the intervention.

General Category 3.1: Lack of Effectiveness

Two participants felt that the intervention had little effect on their psychological or physical state.

> *With this disease, it feels like a death sentence. I wanted to give up, but my family insisted I continue treatment. It feels like all efforts are just a loss of life and money, and this practice doesn't help.* [Interview ig1]

General Category 3.2: Conditions

Subcategory 3.2.1: Fatigue

Three participants reported that fatigue and physical discomfort made it difficult to practice mindfulness regularly.

> *I have poor appetite and low energy every day. I feel uncomfortable all over, the treatment isn't working, and I just want to lie down all day without doing anything. It's hard to practice.* [Interview ig4]

Subcategory 3.2.2: Pain

Two participants said that pain interfered with their ability to engage in mindfulness.

> *This pain is torturous. I'm not in the mood for anything, and it hasn't gotten better. When will I get better? This disease is unbearable.* [Interview ig6]

General Category 3.3: Device Usage Restrictions

Two participants said that distractions from their phones prevented them from completing mindfulness exercises consistently.

> *My phone keeps buzzing with messages, interrupting my practice. It's hard to stay quiet and focus for even a short time. There are just too many distractions.* [Interview ig10]

General Category 3.4: Difficulty in Focusing

Two participants found it hard to concentrate during mindfulness practice, leading to distraction and difficulty persisting.

> *When practicing at home, I often got lost in my thoughts. For example, when asked to focus on my abdomen, I'd start worrying about my illness and other messy things. It was hard to fully follow and stick with it.* [Interview ig3]

XSL•FO
RenderX

General Category 3.5: Lack of Motivation to Practice

Two participants reported difficulty maintaining regular practice due to low motivation.

> *I feel irritable and tired every day. I kept up with the exercises for a few days, but then felt it didn't work. Many times I was just too lazy to do the exercises.* [Interview ig1]

Main Category 4: Support and Encourage

Support and encouragement refer to the social and professional support received by patients during the online mindfulness intervention.

General Category 4.1: Social Support

Two participants shared that family involvement and communication through WeChat groups helped them persevere.

> *My wife started practicing with me in the hospital and continued after we got home. We share our experiences and feelings. Sometimes I don't want to practice, but she encourages me. Practicing together helped me stick with it.* [Interview ig5]

General Category 4.2: Supervision and Guidance

Three participants found that daily check-ins and reminders in the WeChat group, along with exchanges with other patients and medical staff, motivated their persistence.

> *Everyone checked in daily and shared their feelings in the WeChat group. Seeing fellow patients from my ward stick with it and ask questions encouraged me. After more communication, I kept going.* [Interview ig2]

Main Category 5: Accessibility and Convenience

Accessibility and convenience refer to patients' perceptions of how easy and convenient the online mindfulness intervention was to use.

General Category 5.1: Restore a Sense of Balance in Your Life

Three participants said that practicing mindfulness at the same time and place daily helped regulate their lives and restore balance and encouraged ongoing practice.

> *Since my diagnosis, my family has taken special care of me, not letting me do housework. I just lay around and felt useless. After practicing daily, I feel like I've accomplished something and have more energy.* [Interview ig4]

General Category 5.2: Practice Features

Three participants appreciated the convenience of the online format, finding it time-saving, flexible, and easy to fit into their daily routines.

> *I live in a rural area, and it's always hard to come to the hospital, especially without family to accompany me. Now I can practice anytime online after finding the audio. The WeChat method is convenient. When I feel down or have free time, I just practice—it doesn't have to be at a fixed time, so it's more flexible.* [Interview ig11]

Discussion

Principal Findings

This study found that participants experienced both physical and psychological benefits from the intervention. The physical improvements were primarily related to the alleviation of physical symptoms, consistent with findings from Nissen et al [20], who reported that iMBIs can enhance sleep and increase activity levels in patients with cancer. From a psychological perspective, studies by Eyles et al [30] and Weitz et al [31] demonstrated that an 8-week MBSR program can improve the mental health of patients with breast cancer, helping them live more fully in the present. Living in the present is associated with letting go of anxiety and rediscovering happiness [24]. Mindfulness interventions may also lead to changes in neurophysiological activity, brain structure, and function, which contribute to the positive psychological experiences reported by patients with cancer [32]. The intervention program of this study was based on the relevant contents of MBSR, Acceptance and Commitment Therapy, and Mindfulness-Based Cancer Recovery. Participants can have a better understanding of their own diseases and reduce the worries caused by the lack of disease knowledge. The intervention program incorporated elements of acceptance and gratitude. After the intervention, participants reduced their evasive attitude toward some negative emotions and events, faced negative thoughts and feelings with an accepting attitude, improved psychological flexibility, and thereby reduced anxiety and depression. Participants can better maintain a nonjudgmental mindset to continuously perceive and focus on their current experiences, avoiding their consciousness from diverging and wandering in the virtual world of thought, achieving the goal of stopping distractions, concentrating on real things, and thus attaining mental liberation, thereby enhancing trait mindfulness. These outcomes align with the mindfulness-to-meaning theory proposed by Garland et al [33], which suggests that MBIs enhance trait mindfulness and reduce psychological distress.

Social support and guidance were significant facilitators of mindfulness practice. The WeChat group used in this study provided a platform for participants to share experiences and feelings about mindfulness practice and their illness. The involvement of medical staff enabled timely professional guidance, enhancing participants' confidence and adherence. The group check-ins and communication encouraged mutual supervision, motivating participants to practice consistently. During the initial in-hospital phase, mindfulness was introduced to both patients and their families, encouraging joint participation, which helped mitigate the limitations of a solely online intervention. Research by Zulman et al [34] supports this approach, showing that engaging both patients with cancer and their caregivers can foster better communication and mutual support.

However, participants also encountered several obstacles, including perceived ineffectiveness, illness-related fatigue, equipment limitations, difficulty concentrating, lack of

XSL•FO
RenderX

motivation, and challenges completing practices on time. These barriers may stem from the advanced disease stage and poor health status of patients with HCC [35], which can hinder the ability to engage in even simple exercises. Future intervention designs should consider the patient's physical condition and create more tailored programs. A lack of professional mindfulness guidance also contributed to reduced compliance. Although this study used health care staff as facilitators, guidance provided through WeChat lacked real-time responsiveness and personalization, which likely weakened adherence. Previous research has shown that iMBIs with interactive guidance have significantly stronger effects on mindfulness and stress reduction than unguided interventions [36,37].

Participant inclusion criteria also influenced outcomes. High or low baseline levels of negative emotions can reduce intervention effectiveness. For example, Compen et al [15] used a Hospital Anxiety and Depression Scale cutoff of ≥11 and reported significantly better results. If patients are already in a stable psychological state, mindfulness interventions may unintentionally reinforce their identity as "patients," potentially reducing effectiveness. Therefore, selecting participants with appropriate levels of psychological distress is crucial to avoid ceiling or floor effects and to enhance the efficacy of the intervention.

Implications for Practice and Research

The effectiveness of iMBIs is closely tied to the platform used for delivery. Future iMBI platforms should be designed to meet users' individual and shared needs, enhancing usability and comfort. Comprehensive mobile health apps tailored for patients post TACE could be further developed. For instance, Subnis et al [38] created a mindfulness app that offers guided meditations, audio lectures, timers, logs, and stress assessments using facial biosignals. Such tools can reduce medical costs and address patients' psychological needs, providing a sustainable platform for continuous support.

Strengths and Limitations

This study tailored the iMBI to the physiological and psychological characteristics of participants and delivered it via WeChat, in line with their internet usage habits. Participants appreciated the convenience and flexibility of being able to practice at home, which reduced travel-related burdens. The study also emphasized family involvement, which is particularly relevant in the Chinese cultural context where family support plays a central role.

However, limitations remain. Compared with face-to-face mindfulness interventions, iMBIs lack real-time group interaction and peer support. Although WeChat enabled communication, it was still somewhat limited. Using health care providers instead of trained mindfulness instructors meant that the guidance lacked expertise and real-time feedback. Moreover, the WeChat platform itself had functional limitations, lacking features such as real-time monitoring, personalized feedback, and automatic recording of practice sessions. Another limitation is that interviews were conducted postintervention, relying on participants' recollections, which may introduce recall bias. This method also did not allow for tracking participants' evolving experiences over the course of the intervention.

Conclusions

This study used semistructured interviews to explore participants' experiences with iMBIs. Participants reported a variety of benefits, particularly psychological ones. Nevertheless, they also conveyed negative experiences, such as ineffectiveness, emotional resistance, difficulty concentrating, and low motivation. Factors such as social support, supervision, and restored life balance promoted adherence, while device-related distractions and the emotional reminder of illness hindered it. Although iMBIs offer convenience and accessibility, there are still important issues to address. Future efforts should explore participants' individualized needs more deeply and develop comprehensive, user-friendly apps to enhance both the comfort and the effectiveness of mindfulness interventions.

Acknowledgments

The authors would like to thank all participants and researchers for their cooperation.

Funding

Funding for this study was provided by University Natural Science Research Project of Anhui Provincial (grant 2023AH050583), Joint Special Project on Nursing of Anhui Institute of Translational Medicine (grant 2024zhyx-hl-B03), and General Project of Quality Engineering Teaching and Research in Anhui Province (grant 2023jyxm0252).

Data Availability

The data are not publicly available because they contain information that could compromise the privacy of study participants but are available from the corresponding author on reasonable request.

Authors' Contributions

ZL contributed to conceptualization, methodology, formal analysis, data curation, writing—original draft, and writing—review and editing. ML participated in methodology and data curation. YJ participated in data curation and writing—review and editing. LC participated in funding acquisition, project administration, and supervision.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ checklist.
[PDF File (Adobe PDF File), 92 KB - jmir_v28i1e78337_app1.pdf]

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263 [FREE Full text] [doi: 10.3322/caac.21834] [Medline: 38572751]
2. Han B, Zheng R, Zeng H, Wang S, Sun K, Chen R, et al. Cancer incidence and mortality in China, 2022. J Natl Cancer Cent 2024;4(1):47-53 [FREE Full text] [doi: 10.1016/j.jncc.2024.01.006] [Medline: 39036382]
3. NA. Hepatocellular carcinoma. Nat Rev Dis Primers 2021 Jan 21;7(1):7. [doi: 10.1038/s41572-021-00245-6] [Medline: 33479233]
4. Johnson BW, Wright GP. Regional therapies for the treatment of primary and metastatic hepatic tumors: a disease-based review of techniques and critical appraisal of current evidence. Am J Surg 2019;217(3):541-545. [doi: 10.1016/j.amjsurg.2018.10.018] [Medline: 30782316]
5. Tang Z, Bai T, Wei T, Wang X, Chen J, Ye J, et al. TACE combined lenvatinib plus camrelizumab versus TACE alone in efficacy and safety for unresectable hepatocellular carcinoma: a propensity score-matching study. BMC Cancer 2024;24(1):717 [FREE Full text] [doi: 10.1186/s12885-024-12484-3] [Medline: 38862932]
6. Kim MS, Kang M, Park J, Ryu JM. Nurses' comfort care of transarterial chemoembolization patients based on their perceptions around postembolization syndrome and symptom interference. Nurs Open 2023;10(5):2877-2885 [FREE Full text] [doi: 10.1002/nop2.1529] [Medline: 36565057]
7. Zhang L, Zhang X, Li Q, Makamure J, Liu Z, Zhao D, et al. Transarterial chemoembolization failure in patients with hepatocellular carcinoma: incidence, manifestation and risk factors. Clin Res Hepatol Gastroenterol 2023;47(2):102071. [doi: 10.1016/j.clinre.2022.102071] [Medline: 36539181]
8. Liu M, Liu X, Liu S, Xiao F, Guo E, Qin X, et al. Big data-based identification of multi-gene prognostic signatures in liver cancer. Front Oncol 2020;10:847 [FREE Full text] [doi: 10.3389/fonc.2020.00847] [Medline: 32547951]
9. Lan SC, Lin YE, Chen SC, Lin YF, Wang YJ. Effects of acupressure on fatigue and depression in hepatocellular carcinoma patients treated with transcatheter arterial chemoembolization: a quasi-experimental study. Evid Based Complement Alternat Med 2015;2015:496485 [FREE Full text] [doi: 10.1155/2015/496485] [Medline: 25802540]
10. Liu Z, Li M, Jia Y, Zheng L, Chen L. Effect of perceived stress on psychological distress in hepatocellular carcinoma patients undergoing TACE: the mediating role of experiential avoidance and the moderating role of trait mindfulness. BMC Cancer 2025;25(1):254 [FREE Full text] [doi: 10.1186/s12885-025-13679-y] [Medline: 39948489]
11. Cao W, Li J, Hu C, Shen J, Liu X, Xu Y, et al. Symptom clusters and symptom interference of HCC patients undergoing TACE: a cross-sectional study in China. Support Care Cancer 2013;21(2):475-483. [doi: 10.1007/s00520-012-1541-5] [Medline: 23010958]
12. Kubo A, Kurtovich E, McGinnis M, Aghaee S, Altschuler A, Quesenberry C, et al. A randomized controlled trial of mHealth mindfulness intervention for cancer patients and informal cancer caregivers: a feasibility study within an integrated health care delivery system. Integr Cancer Ther 2019;18:1534735419850634 [FREE Full text] [doi: 10.1177/1534735419850634] [Medline: 31092044]
13. Goldin PR, Thurston M, Allende S, Moodie C, Dixon ML, Heimberg RG, et al. Evaluation of cognitive behavioral therapy vs mindfulness meditation in brain changes during reappraisal and acceptance among patients with social anxiety disorder: a randomized clinical trial. JAMA Psychiatry 2021;78(10):1134-1142 [FREE Full text] [doi: 10.1001/jamapsychiatry.2021.1862] [Medline: 34287622]
14. Martín J, García S, Anton-Ladislao A, Ferreiro J, Martín M, Padierna A, CAMISS-prospective group. Variables related to health-related quality of life among breast cancer survivors after participation in an interdisciplinary treatment combining mindfulness and physiotherapy. Cancer Med 2023;12(12):13834-13845 [FREE Full text] [doi: 10.1002/cam4.6035] [Medline: 37165927]
15. Compen F, Bisseling E, Schellekens M, Donders R, Carlson L, van der Lee M, et al. Face-to-face and internet-based mindfulness-based cognitive therapy compared with treatment as usual in reducing psychological distress in patients with Cancer: a multicenter randomized controlled trial. J Clin Oncol 2018;36(23):2413-2421. [doi: 10.1200/jco.2017.76.5669]
16. Liu Z, Li M, Jia Y, Wang S, Zheng L, Wang C, et al. A randomized clinical trial of guided self-help intervention based on mindfulness for patients with hepatocellular carcinoma: effects and mechanisms. Jpn J Clin Oncol 2022;52(3):227-236. [doi: 10.1093/jjco/hyab198] [Medline: 35088079]

XSL•FO
RenderX

17. Shao D, Zhang H, Cui N, Sun J, Li J, Cao F. The efficacy and mechanisms of a guided self-help intervention based on mindfulness in patients with breast cancer: a randomized controlled trial. Cancer 2021;127(9):1377-1386 [FREE Full text] [doi: 10.1002/cncr.33381] [Medline: 33332582]
18. Ireland MJ, Clough B, Gill K, Langan F, O'Connor A, Spencer L. A randomized controlled trial of mindfulness to reduce stress and burnout among intern medical practitioners. Med Teach 2017;39(4):409-414. [doi: 10.1080/0142159x.2017.1294749]
19. Sommers-Spijkerman M, Austin J, Bohlmeijer E, Pots W. New evidence in the booming field of online mindfulness: an updated meta-analysis of randomized controlled trials. JMIR Ment Health 2021;8(7):e28168 [FREE Full text] [doi: 10.2196/28168] [Medline: 34279240]
20. Nissen ER, O'Connor M, Kaldo V, Højris I, Borre M, Zachariae R, et al. Internet-delivered mindfulness-based cognitive therapy for anxiety and depression in cancer survivors: a randomized controlled trial. Psychooncology 2020;29(1):68-75 [FREE Full text] [doi: 10.1002/pon.5237] [Medline: 31600414]
21. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
22. Bruix J, Sherman M, Llovet JM, Beaugrand M, Lencioni R, Burroughs AK, EASL Panel of Experts on HCC. Clinical management of hepatocellular carcinoma. Conclusions of the Barcelona-2000 EASL conference. European Association for the Study of the Liver. J Hepatol 2001;35(3):421-430. [doi: 10.1016/s0168-8278(01)00130-1] [Medline: 11592607]
23. Shapiro SL, Brown KW, Thoresen C, Plante TG. The moderation of mindfulness-based stress reduction effects by trait mindfulness: results from a randomized controlled trial. J Clin Psychol 2011;67(3):267-277. [doi: 10.1002/jclp.20761] [Medline: 21254055]
24. Finucane A, Mercer SW. An exploratory mixed methods study of the acceptability and effectiveness of mindfulness -based cognitive therapy for patients with active depression and anxiety in primary care. BMC Psychiatry 2006;6(1). [doi: 10.1186/1471-244x-6-14]
25. Fledderus M, Oude Voshaar MAH, Ten Klooster PM, Bohlmeijer ET. Further evaluation of the psychometric properties of the acceptance and action questionnaire-II. Psychol Assess 2012;24(4):925-936. [doi: 10.1037/a0028200] [Medline: 22545700]
26. Robinson OC. Probing in qualitative research interviews: Theory and practice. Qual Res Psychol 2023;20(3):382-397. [doi: 10.1080/14780887.2023.2238625]
27. QIP L. NVivo qualitative data analysis software (version 11). 2014. URL: https://lumivero.com/products/nvivo/ [accessed 2026-01-08]
28. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
29. Lindgren BM, Lundman B, Graneheim UH. Abstraction and interpretation during the qualitative content analysis process. Int J Nurs Stud 2020;108:103632. [doi: 10.1016/j.ijnurstu.2020.103632]
30. Eyles C, Leydon GM, Hoffman CJ, Copson ER, Prescott P, Chorozoglou M, et al. Mindfulness for the self-management of fatigue, anxiety, and depression in women with metastatic breast cancer: a mixed methods feasibility study. Integr Cancer Ther 2015;14(1):42-56 [FREE Full text] [doi: 10.1177/1534735414546567] [Medline: 25161198]
31. Weitz MV, Fisher K, Lachman VD. The journey of women with breast cancer who engage in mindfulness-based stress reduction: a qualitative exploration. Holist Nurs Pract 2012;26(1):22-29. [doi: 10.1097/HNP.0b013e31823c008b] [Medline: 22157506]
32. Hölzel BK, Carmody J, Vangel M, Congleton C, Yerramsetti SM, Gard T, et al. Mindfulness practice leads to increases in regional brain gray matter density. Psychiatry Res 2011;191(1):36-43 [FREE Full text] [doi: 10.1016/j.pscychresns.2010.08.006] [Medline: 21071182]
33. Garland EL, Fredrickson B, Kring AM, Johnson DP, Meyer PS, Penn DL. Upward spirals of positive emotions counter downward spirals of negativity: insights from the broaden-and-build theory and affective neuroscience on the treatment of emotion dysfunctions and deficits in psychopathology. Clin Psychol Rev 2010;30(7):849-864 [FREE Full text] [doi: 10.1016/j.cpr.2010.03.002] [Medline: 20363063]
34. Zulman DM, Schafenacker A, Barr KLC, Moore IT, Fisher J, McCurdy K, et al. Adapting an in-person patient-caregiver communication intervention to a tailored web-based format. Psychooncology 2012;21(3):336-341 [FREE Full text] [doi: 10.1002/pon.1900] [Medline: 21830255]
35. van den Hurk DGM, Schellekens MPJ, Molema J, Speckens AEM, van der Drift MA. Mindfulness-based stress reduction for lung cancer patients and their partners: results of a mixed methods pilot study. Palliat Med 2015;29(7):652-660 [FREE Full text] [doi: 10.1177/0269216315572720] [Medline: 25701663]
36. Arvidsson B, Allard E, Sjögren E, Lennernäs H, Sjöberg PJR, Bergquist J. Online capillary solid phase extraction and liquid chromatographic separation with quantitative tandem mass spectrometric detection (SPE-LC-MS/MS) of ximelagatran and its metabolites in a complex matrix. J Chromatogr B Analyt Technol Biomed Life Sci 2009;877(3):291-297. [doi: 10.1016/j.jchromb.2008.12.017] [Medline: 19117807]

XSL•FO
RenderX

37. Spijkerman MPJ, Pots WTM, Bohlmeijer ET. Effectiveness of online mindfulness-based interventions in improving mental health: a review and meta-analysis of randomised controlled trials. Clin Psychol Rev 2016;45:102-114 [FREE Full text] [doi: 10.1016/j.cpr.2016.03.009] [Medline: 27111302]
38. Subnis UB, Farb NA, Piedalue KL, Speca M, Lupichuk S, Tang PA, et al. A smartphone app-based mindfulness intervention for cancer Survivors: protocol for a randomized controlled trial. JMIR Res Protoc 2020;9(5):e15178 [FREE Full text] [doi: 10.2196/15178] [Medline: 32390591]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
HCC: hepatocellular carcinoma
iMBIs: internet-delivered mindfulness-based interventions
MBI: mindfulness-based intervention
MBSR: Mindfulness-Based Stress Reduction
TACE: transarterial chemoembolization

Edited by A Stone; submitted 31.May.2025; peer-reviewed by C Sun, W Zhou; comments to author 21.Oct.2025; revised version received 28.Nov.2025; accepted 31.Dec.2025; published 29.Jan.2026.

Please cite as:
Liu Z, Li M, Jia Y, Chen L
Tailored Internet-Delivered Mindfulness-Based Interventions for Patients With Hepatocellular Carcinoma After Transarterial Chemoembolization: Qualitative Study
J Med Internet Res 2026;28:e78337
URL: https://www.jmir.org/2026/1/e78337
doi:10.2196/78337
PMID:

©Zengxia Liu, Min Li, Yong Jia, Li Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Associations Among Cyberbullying Victimization, Inhibitory Control, Neural Activation of Error Processing, and Mental Health Problems in Adolescents: Neuroimaging, Retrospective Longitudinal Cohort Study Using the Adolescent Brain Cognitive Development Data

Xuanyu Zhang[1], BS; Chengyan Xie[2], BS; Yu Chen[3], PhD; Boyu Qiu[1], PhD

[1]School of Health Management, Guangzhou Medical University, Guangzhou, Guangdong, China

[2]The Third School of Clinical Medicine, Guangzhou Medical University, Guangzhou, China

[3]Department of Psychiatry, Yale University School of Medicine, New Haven, CT, United States

Corresponding Author:
Boyu Qiu, PhD
School of Health Management, Guangzhou Medical University
No.1, Xinzao Road
Xinzao Town, Panyu District
Guangzhou, Guangdong, 511436
China
Phone: 86 37103516 ext 37103516
Email: boyuqiu@gzhmu.edu.cn

Abstract

Background: Cyberbullying victimization is prevalent and closely linked to mental health problems. However, existing research, often limited by cross-sectional designs and a focus on direct relationships, has yielded inconsistent results. Furthermore, the biological mechanisms underlying the relationship between cyberbullying victimization and psychopathological outcomes remain largely unclear at present.

Objective: This retrospective cohort study aimed to explore the longitudinal associations among cyberbullying victimization, inhibitory control, brain activation during error processing, and mental health problems among adolescents.

Methods: We curated the clinical, behavioral, and neuroimaging data (551/1186, 46.5% girls; 9-10 years at baseline) from the Adolescent Brain Cognitive Development study, a nationally representative cohort established through school-based probability sampling (selected factors included gender, race/ethnicity, socioeconomic status, and urbanicity). Participants were assessed by the cyberbullying question, the functional magnetic resonance imaging stop signal task for inhibitory control and error processing, and the Child Behavioral Checklist for externalizing and internalizing problems at 2-year (T1) and 4-year follow-up (T2). Linear mixed models were used to examine the retrospective longitudinal associations between these clinical, behavioral, and neuroimaging factors.

Results: Linear mixed models showed that victims of cyberbullying at T1 exhibited significantly greater externalizing problems at T2 (β=0.25, 95% CI 0.06-0.45, P_{FDR}=.02), but not for internalizing problems (β=–0.01, 95% CI –0.20 to 0.19, P_{FDR}=.99) or deficits in inhibitory control (Correct Stop Rate: β=–0.02, 95% CI –0.26 to 0.21, P_{FDR}=.85; Stop Signal Reaction Time: β=–0.07, 95% CI –0.27 to 0.13, P_{FDR}=.85). Furthermore, cyberbullying victimization at T1 contributed to higher activation in the bilateral superior parietal gyri (left: β=0.36, 95% CI 0.10-0.61, P_{FDR}=.04; right: β=0.34, 95% CI 0.08-0.59, P_{FDR}=.04), right inferior parietal gyrus (β=0.32, 95% CI 0.07-0.57, P_{FDR}=.04), and right posterior cingulate cortex (β=0.34, 95% CI 0.09-0.60, P_{FDR}=.04) during error processing at T2. However, these neural alterations did not significantly mediate between cyberbullying victimization at T1 and externalizing problems at T2.

Conclusions: This longitudinal functional magnetic resonance imaging study investigates neural correlates of cyberbullying victimization in adolescents. By extending prior research that has relied primarily on cross-sectional or behavioral data, this research demonstrates that this form of victimization is associated with altered neural activation during error processing in later development. The pattern of nonsignificant impairment in inhibitory control and mediation to externalizing problems suggests that these neural impacts may be better characterized by a state of heightened sensitivity and compensatory engagement than by

XSL•FO
RenderX

direct damage. Overall, this study points to the error-processing network as a potential target for cognitive interventions and establishes a foundation for further exploration of other neural mechanisms between cyberbullying victimization and mental health outcomes.

(J Med Internet Res 2026;28:e75126) doi:10.2196/75126

KEYWORDS

cyberbullying victimization; mental health problems; error processing; functional magnetic resonance imaging; stop signal task

Introduction

Literature Review

Bullying is a prevalent risk factor for mental health disorders among adolescents [1,2]. In an era characterized by web-based relationships and interactions, cyberbullying has become increasingly common [3,4], which is defined as the repeated and deliberate harm by a perpetrator to a victim through mobile phones, computers, and other electronic devices [5]. A growing number of studies have linked cyberbullying victimization to mental health problems in adolescents, including more internalizing problems [6-8] such as depression [9,10] and anxiety [11,12] as well as more externalizing problems such as aggressive [13,14] and rule-breaking behaviors [15,16].

However, there exists a subset of studies that failed to find the significant effects of cyberbullying victimization on adolescents' mental health problems [17-20]. Such heterogeneity on this topic may be attributed to differences in experimental design (eg, cross-sectional vs longitudinal study) and variable measurement (eg, different questionnaire assessments) across studies, particularly the tendency to focus solely on direct observational relationships without sufficiently investigating the underlying psychological and physiological mechanisms that may explain how and for whom these effects occur [21]. Therefore, there is a current need for more longitudinal studies that use consistent, well-validated measures and incorporate biological perspectives to clarify the pathways linking cyberbullying victimization to mental health.

Grounded in the stress process model for internalizing problems [22] and the general aggression model for externalizing problems [23], the association between cyberbullying victimization and these psychopathological outcomes may be mediated by inhibitory control and its associated neural mechanisms. Inhibitory control, the ability to suppress inappropriate responses to stimuli [24], is negatively affected by various adverse childhood experiences (ACEs) [25], such as threat [26,27], deprivation [28,29], and abuse [30,31]. Inhibitory control deficits disrupt an individual's emotional and behavioral regulation, further contributing to internalizing and externalizing problems [32-34]. Given that the cyberbullying victimization in adolescents shares key characteristics with ACEs (ie, exposure to adversity, trauma, stress, or threats) and is closely linked to mental health problems, it is pertinent to explore whether diminished inhibitory control mediates the association between cyberbullying victimization and psychopathological outcomes.

According to the error detection theory [35,36], performance on inhibitory control is closely linked to the brain activation during error processing. Specifically, when individuals fail to inhibit a response, the cingulate cortex typically shows increased activation, as it reflects the perception of errors (ie, failed inhibition) [37,38] and supports subsequent cognitive and behavioral adjustments [39,40]. Moreover, previous studies have shown that ACEs impede the error monitoring circuitry, leading to blunted activation in cingulate cortices during failed inhibition [41,42]. Therefore, we focused on the cingulate cortices as regions of interest (ROIs) to investigate whether or how cyberbullying victimization decreases neural responses and contributes to inhibitory control deficits. The cingulate cortex further interacts with frontal [43,44] and parietal [45,46] cortices during error processing to maintain inhibitory control, which are key regions of the cognitive control network closely linked to information integration, attention regulation, and adaptive behaviors [47,48]. Given that ACEs have been suggested to reduce activation and connectivity within this network [49,50], we also included frontoparietal regions as ROIs to examine potential alterations of brain activation in cyberbullying victims.

Notably, several studies have indicated that adolescents with ACEs may exhibit excessive, rather than insufficient, activation in cingulate, frontal, and parietal cortices during response inhibition and error processing [51-53]. Such brain over-activation within the cognitive control network may reflect heightened sensitivity to errors and hypervigilance to adversity following trauma and stress caused by ACEs [54] and predicts the development of mental health problems such as depression and anxiety [55-57]. Since cyberbullying victimization is also traumatic [5] and stressful [58] for adolescents, it is plausible that a mechanism exists whereby cyberbullying victimization directly mediates excessive activation of cingulate, frontal, and parietal cortices, contributing to psychopathological outcomes.

McLoughlin et al [59] provided initial functional magnetic resonance imaging (fMRI) evidence supporting this potential mechanism. Their findings showed that viewing cyberbullying-related stimuli activated responses in many brain regions (eg, frontal, cingulate, and parietal cortices), closely associated with cognitive and emotional processing [60]. Despite the negative impacts of viewing cyberbullying content, adolescents may also be victims of cyberbullying rather than bystanders. Exploring the specific neural harms of cyberbullying victimization can contribute to a comprehensive understanding of the negative impacts of cyberbullying as well as the development of effective postvictimization interventions. However, based on their review [61] and to our knowledge, no longitudinal studies have investigated the neural alterations after adolescents experienced cyberbullying.

XSL•FO
RenderX

The abilities of inhibitory control and error monitoring improve progressively throughout adolescence, which are essential for maturing adaptive behaviors and reducing the risk of psychopathological outcomes [62,63]. This progression is supported by steadily increasing activation in related brain regions during neurodevelopment [64,65]. Notably, early adolescence is a critical period during which the neurodevelopment is particularly vulnerable to bullying [66] and highly sensitive to acceptance and rejection, especially through social media [67]. Therefore, it is important to examine the specific impacts of cyberbullying victimization on the adolescent brain using a longitudinal design. Building on the longitudinal associations, the temporal mechanisms of cyberbullying victimization and mental health problems, as well as the mediating role of biological factors between them, can be more clearly elucidated.

This Study

Using data from a large-scale longitudinal study in the United States—the Adolescent Brain Cognitive Development (ABCD) study [68]—this study aimed to explore the longitudinal associations among cyberbullying victimization, inhibitory control, brain activation during error processing, and mental health problems among adolescents. First, we explored how inhibitory control and ROIs' activation during error processing at T2, both measured using the stop signal task (SST), were influenced by cyberbullying victimization at T1. We broadly hypothesized that cyberbullying victimization at T1 predicted impaired inhibitory control and altered brain activation in the error-monitoring regions at T2. Second, we examined whether the cyberbullying victimization at T1 significantly predicted the mental health problems at T2. Based on the existing evidence, we hypothesized that individuals who were victims of cyberbullying at T1 would exhibit more internalizing and externalizing problems at T2 as compared with nonvictims at T1. Finally, we hypothesized that alterations in brain activity during error processing might mediate the relationship between cyberbullying victimization and mental health problems.

Methods

Study Design and Participants

This retrospective cohort study uses data from the ABCD study, which is an ongoing and 10-year longitudinal study of neurocognitive development in adolescents. To form a nationally representative sample, the ABCD study used school-based probability sampling to recruit its baseline cohort between 2016 and 2018. With school selection informed by gender, race/ethnicity, socioeconomic status, and urbanicity, it enrolled adolescents aged 9-10 years from 21 sites across the United States to mitigate selection bias [69]. As data pertaining to cyberbullying are available from the 2-year follow-up and imaging data are collected every other year, the data used in this study were obtained from the 2-year follow-up (T1) and 4-year follow-up (T2) of the ABCD Study Data Release 5.1. Participants were excluded if (1) they had missing data pertaining to cyberbullying and mental health problems, (2) they had missing data pertaining to fMRI of SST, (3) their imaging data were recommended for exclusion by the ABCD team, or (4) their behavioral performance of SST was recommended for exclusion by the ABCD team. For more details regarding the exclusion of imaging data and behavioral performance of SST, please refer to the ABCD Human Subjects Study (2024) [70].

The flowchart of participant exclusion is detailed in Figure S1 in Multimedia Appendix 1 [71-75], with most of the missing data due to the loss of mental health and brain imaging data. The Little's MCAR test suggested that the data were missing at random (χ^2_{24}= 31.07, *P*=.15) [76]. Therefore, participants who did not meet the criteria were deleted from the present analysis, and the final sample consisted of 1186 participants (mean age of 11.96 years; 551/1186, 46.5% female; Table 1 provides demographic information, and Table S1 in Multimedia Appendix 1 [71-75] provides comparison of included and excluded participants' characteristics). The adequacy of this sample size was evaluated using Monte Carlo simulations [77]: for the linear mixed models (LMMs) examining longitudinal associations between cyberbullying victimization, inhibitory control, brain activation during error processing, and mental health problems, these models provided 99.26% (95% CI 99.00-100.00) statistical power to detect effect size β=0.2 at α of .05. This study was reported in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement [78], and the completed checklist is provided in Multimedia Appendix 2 [71-75].

XSL•FO
RenderX

Table 1. Descriptive statistics for demographic characteristics at T1 in the Adolescent Brain Cognitive Development study.

Characteristics	Cyberbullying victims (*n*=61)	No cyberbullying experience (*n*=1125)	*t* test (*df*)	Chi-square (*df*)	*P* value
Age (years), mean (SD)	11.96 (0.58)	11.96 (0.64)	0.07 (1184)	—[a]	.95
Sex, n (%)			—	0.13 (1)	.72
Female	27 (44.3)	524 (46.6)			
Male	34 (55.7)	601 (53.4)			
Race and ethnicity, n (%)			—	2.94 (4)	.57
White	41 (67.2)	661 (58.8)			
Black	4 (6.6)	100 (8.9)			
Hispanic	9 (14.8)	219 (19.5)			
Asian	0 (0.0)	22 (2.0)			
Other	7 (11.5)	123 (10.9)			
Family income (US $), n (%)			—	0.48 (3)	.92
Less than 50,000	13 (21.3)	221 (19.6)			
50,000 through 99,999	19 (31.1)	324 (28.8)			
100,000 and greater	25 (41.0)	512 (45.5)			
Refused to answer/do not know	4 (6.6)	68 (6.0)			
Parents' highest education, n (%)			—	1.36 (3)	.72
High school education or less	8 (13.1)	143 (12.7)			
Some college	9 (14.8)	184 (16.4)			
Associate or Bachelor Degree	31 (50.8)	496 (44.1)			
Post-Graduate Degree	13 (21.3)	302 (26.8)			
Offline victimization, mean (SD)	12.45 (8.81)	12.40 (4.00)	0.10 (1184)	—	.92
Adverse childhood experiences, mean (SD)	1.47 (1.01)	1.44 (1.15)	0.21 (1184)	—	.83
Family conflict, mean (SD)	1.91 (1.80)	1.89 (1.20)	0.09 (1184)	—	.93
Social media use, hours, mean (SD)	0.47 (0.79)	0.51 (1.25)	0.27 (1184)	—	.79

[a]Not applicable.

Ethical Considerations

In accordance with the Declaration of Helsinki, the ABCD study's procedures were reviewed, exempted, and approved by the Centralized Institutional Review Board (IRB) at the University of California, San Diego (IRB# 160091), as well as each individual site's IRB. Informed consent was obtained from all participants and their legal guardians for both primary data collection and secondary analyses. The IRB confirmed that no additional consent was required for the secondary use of deidentified data. To ensure privacy and confidentiality, all data were deidentified and anonymized before analysis, with no personally identifiable information retained or disclosed. Participants and their families received financial compensation for their time, typically including US $200 for the parent or guardian, US $100 in gifts and gift cards for the child, and coverage for travel costs and sibling childcare. Images and Multimedia Appendix 1 [71-75] ensure participant anonymity, with no identifiable information included.

Questionnaire Measures

Cyberbullying Victimization

A modified version of the cyberbullying scale [79] was used to investigate whether adolescents were victims of cyberbullying. To report cyberbullying victimization, participants responded to the question, "In the past 12 months, have you been cyberbullied where someone intentionally tried to hurt you or be mean to you online, in texts or group texts, or on social media (like Instagram or Snapchat)?". Participants answered either "yes" or "no," and those who responded "yes" were coded as experiencing cyberbullying victimization [80,81].

Mental Health Problems

The mental health problems of adolescents were assessed using the parent-reported Child Behavior Checklist [82]. On a 3-point scale (0=not true to 2=very true or very often), parents reported on the adolescent's internalizing (T1: Cronbach α=0.88; T2: Cronbach α=0.87) and externalizing problems (Cronbach α=0.90; T2: Cronbach α=0.88). The internalizing score is the

XSL•FO
RenderX

sum of all items related to “anxious depressive symptoms (12 items; eg, Fears mistakes)” and “withdrawn depressive symptoms (8 items; eg, Lacks energy),” while the externalizing score is the sum of all items related to “rule-breaking behaviors (13 items; eg, Breaks rules)” and “aggressive behaviors (17 items; eg, Gets in fights).” The raw total score on each problem dimension was converted to norm-referenced *T*-score (range with a mean of 50 and SD of 10), with a higher number indicating greater symptom severity.

SST

The SST [83] assesses participants’ behavioral response inhibition through 300 “Go” trials and 60 “Stop” trials. During frequent “Go” trials, participants are required to press a button promptly and correctly upon seeing an arrow indicating left or right. In contrast, on “Stop” trials, participants are required to refrain from pressing the button upon seeing a vertical arrow. Both the correct “Stop” rate and stop signal reaction time (a specific calculation method is provided in Multimedia Appendix 1 [71-75]) are indicators of inhibitory control ability. Furthermore, the stop signal delay, defined as the interval between Go and Stop trials, is calibrated to maintain response inhibition accuracy at approximately 50%. If a trial is successfully inhibited, the stop signal delay is increased by 50 milliseconds, whereas it is decreased by 50 milliseconds following unsuccessful inhibition. As SST ensures nearly 50% inhibition of failure, it is optimally suited for testing brain activation during error detection. For more details regarding SST design, please refer to Casey et al [68].

fMRI Data Acquisition and Preprocessing

The ABCD study used a harmonized neuroimaging protocol across 21 sites, using three 3T scanner platforms: Siemens Prisma (Siemens Healthineers), GE 750 (GE Healthcare), and Philips (Philips Healthcare). The ABCD’s Data Analysis, Informatics, and Resource Core was responsible for the processing of all scan data, ensuring both high quality and uniform standards across participating sites. Corrections were made for head motion, B0, nonlinearity distortions, and between-scan motion for each participant [84-86]. The ROIs’ values were derived from the average time courses of cortical regions, which were defined by FreeSurfer’s anatomical parcellations and Desikan-Killiany Atlas [87]. General linear models in AFNI’s 3dDeconvolve (Cox, 1996) were used to model task-related activation strength. Mean beta coefficients and standard errors for each ROI’s time series were calculated. For more details regarding fMRI preprocessing techniques and the analysis of task-based fMRI data, please refer to Casey et al [68] and Hagler et al [88].

In this study, mean beta weights for failed stop contrasts (ie, incorrect stop contrasted with correct go and incorrect stop contrasted with correct stop) were used to measure brain activation during error processing [68]. Based on previous studies, we selected the cortical ROIs (Figure 1) most likely recruited by inhibitory control and error processing [43,44,53]: medial frontal, rostral middle frontal, rostral anterior cingulate, and posterior cingulate cortices. The superior and inferior parietal regions, which are closely related to attention, sensation, and adaptive behaviors [45,89], were also examined. The beta weights for each ROI in the right and left hemispheres were used in subsequent analyses.

Figure 1. Regions of interest for failed inhibition analysis in the Adolescent Brain Cognitive Development study. The regions of interest included medial frontal, rostral middle frontal, rostral anterior cingulate, and posterior cingulate cortices. The brain activation during error processing was reflected by the mean beta weights for failed stop contrasts (ie, incorrect stop contrasted with correct go and incorrect stop contrasted with correct stop). All *P* values were adjusted for multiple comparisons using the false discovery rate method, and $P<.05$ indicated significance.

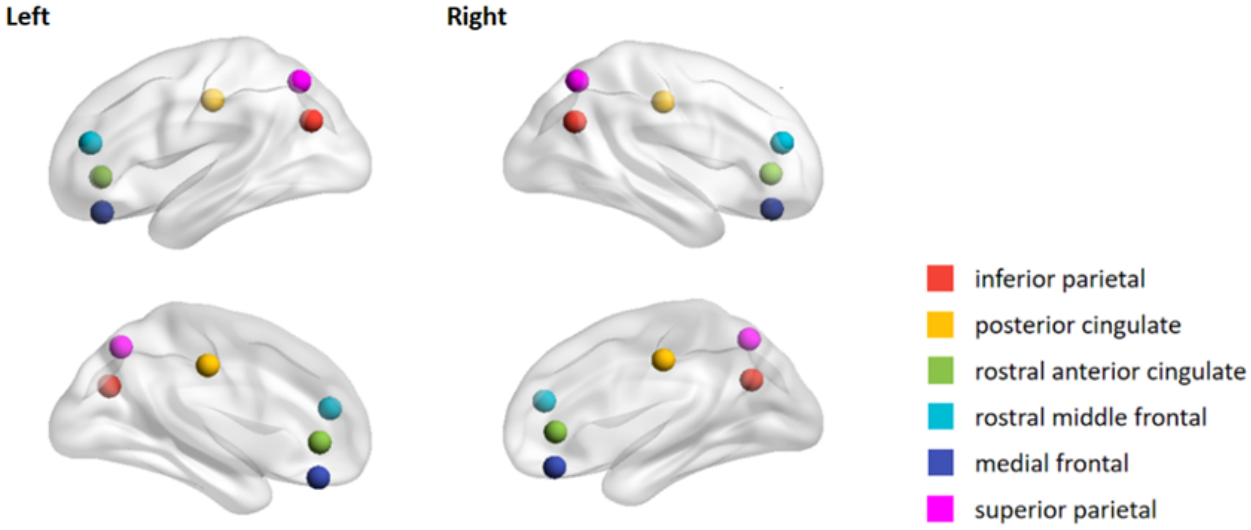

Covariates

Potential confounders include adolescents’ age, biological sex, race/ethnicity, family income, parents’ highest educational attainment, offline victimization, ACEs, family conflict, and social media use at T1. Adolescents’ biological sex was coded into 1=male and 2=female. Race was coded into 5 categorical variables, including 1=White, 2=Black, 3=Hispanic, 4=Asian, and 5=other races. Family income was categorized into 3 brackets, ranging from 1 (less than US $50,000) to 3 (US

XSL•FO
RenderX

$100,000 and greater). Parents' educational attainment was the highest educational degree in the family, ranging from 1 (less than a high school diploma) to 4 (postgraduate degree). Other covariates remained in the continuous format to be controlled (scoring details are provided in Multimedia Appendix 1 [71-75]).

Statistical Analyses

All data analyses were conducted on R (version 4.4.1; R Development R Core Team, 2024). LMMs in the lme4 package [90] were used to explore the retrospective longitudinal association between cyberbullying victimization, inhibitory control, brain activation during error processing, and mental health problems. First, mental health outcomes (ie, internalizing and externalizing problems) at T2 served as the dependent variables, with whether cyberbullying victimization at T1 as the independent variable, which was included as a fixed effect in the model. Research sites and family id were treated as random effects to reflect the nested structure of the data. Second, the performance on inhibitory control (ie, correct "Stop" rate and stop signal reaction time from the SST) and the ROIs' activation during error processing (ie, beta weight of failed stop from the SST) at T2 were used as dependent variables, while cyberbullying victimization at T1 remained the independent variable. MRI scanner type was added as a random effect in these brain measurement models.

Further analyses followed up any significant findings by examining whether cyberbullying victimization at T1 was associated with mental health outcomes at T2 via altered brain activation patterns during error processing. According to Cole and Maxwell's recommendation [91] regarding 2 waves of longitudinal studies, a half-longitudinal mediation model was established to explore the mediating role of brain activation during error processing (Figure 2). Specifically, we calculated the regression coefficient from T1 cyberbullying victimization to T2 brain measures (path a) and the regression coefficient from T1 brain measures to T2 mental health outcomes (path b). The significance of the mediating effect (path a × b) was calculated using bootstrap analysis with 5000 resamples to generate bias-corrected CI. When the 95% CI does not contain 0, then the mediating effect of brain activation is statistically significant.

Adolescents' biological sex, age, race/ethnicity, family income, and highest parental education attainment were included as demographic covariates in all analyses. In the analyses predicting brain activation of SST, we additionally covaried mean framewise displacement and its quadratic effects during the fMRI scan [92]. Furthermore, to control for the developmental effect of the dependent variable in each LMM, we separately covaried T1 level of mental health, inhibitory control, and activation of ROIs during error processing in the corresponding model [71,93]. For example, the T1 level of externalizing problems was included as a covariate in the LMM exploring the link between whether cyberbullying victimization at T1 and T2 levels of externalizing problems. Finally, we used the inverse probability weighting method [94] in the survey package [95] to test the robustness of the results from all LLMs and mediation models. False discovery rate (FDR) was applied to correct for multiple comparisons, and P_{FDR}<.05 was considered to reach the significant level.

Figure 2. Conceptual half-longitudinal mediation model examining neural mechanisms linking cyberbullying victimization to mental health outcomes in US adolescents. For all analyses, participants' age, biological sex, race/ethnicity, family income, parents' highest educational attainment, offline victimization, adverse childhood experiences, family conflict, and social media use were included as covariates. Path a (T1 victimization → T2 brain measures) additionally covaried mean framewise displacement during the fMRI scan and the T1 activation level of regions of interest, while path b (T1 brain measures → T2 mental health) added T1 level of mental health to the control.

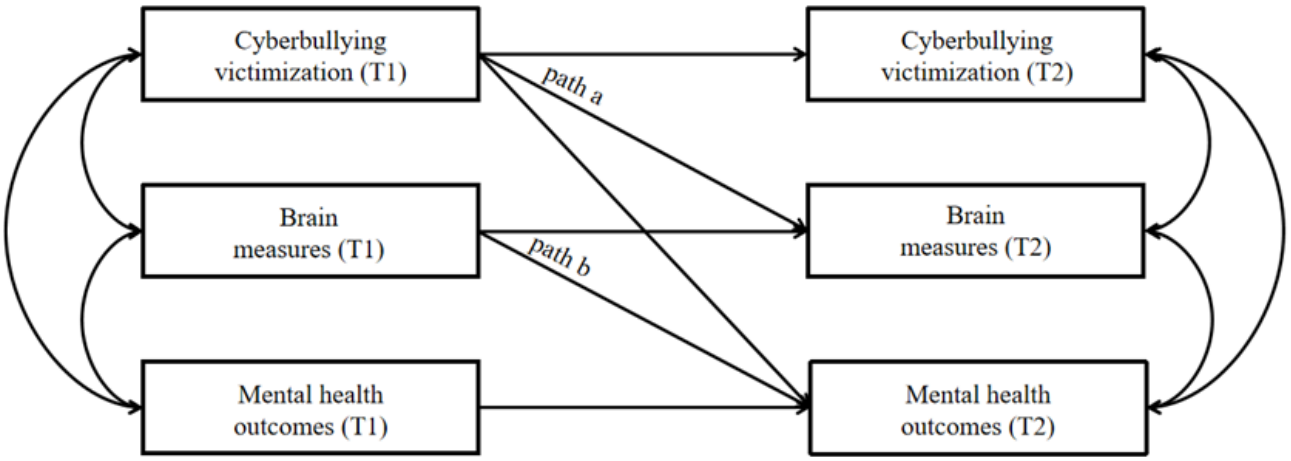

Results

Descriptive Statistics for Demographic Characteristics

Descriptive statistics for demographic characteristics at T1 are presented in Table 1. Independent samples *t* tests and chi-square test indicated that no significant difference was found between the cyberbullying victims and participants with no cyberbullying experiences in age, offline victimization, ACEs, family conflict, and social media use, t_{1184}< 0.27, *P*>.79, sex, race/ethnicity, family income, or parents' highest educational attainment, χ^2_4< 2.94, *P*>.57.

XSL•FO
RenderX

Associations Between Cyberbullying at T1 and ROIs' Activation During Error Processing at T2

The associations between cyberbullying victimization at T1 and ROIs' activation during error processing at T2 are shown in Table 2. For the contrast of incorrect stop versus correct go, cyberbullying victimization at T1 was significantly associated with higher activation of bilateral superior parietal gyri (left: β=0.36, 95% CI 0.10-0.61, SE 0.01, P_{FDR}=.04; right: β=0.34, 95% CI 0.08-0.59, SE 0.01, P_{FDR}=.04), right inferior parietal gyrus (β=0.32, 95% CI 0.07-0.57, SE 0.01, P_{FDR}=.04), and right posterior cingulate cortex (β=0.34, 95% CI 0.09-0.60, SE 0.01, P_{FDR}=.04) at T2. For the contrast of incorrect stop versus correct stop, significant heightened activation was also observed in the bilateral superior parietal gyri (left: β=0.32, 95% CI 0.08-0.52, SE 0.01, P_{FDR}=.04; right: β=0.30, 95% CI 0.05-0.48, SE 0.01, P_{FDR}=.04), bilateral inferior parietal gyri (left: β=0.30, 95% CI 0.06-0.50, SE 0.01, P_{FDR}=.04; right: β=0.30, 95% CI 0.04-0.47, SE 0.01, P_{FDR}=.04), and right posterior cingulate cortex (β=0.34, 95% CI 0.10-0.60, SE 0.01, P_{FDR}=.04). No other significant ROIs' activated alteration during error processing at T2 were predicted by cyberbullying victimization at T1 after correction for multiple comparisons.

Table 2. Associations between cyberbullying victimization at T1 and cortical regions of interest activation during error processing at T2 in the Adolescent Brain Cognitive Development study.

	Left hemisphere					Right hemisphere				
	β[a]	95% CI	SE[b]	*P* value[c]	P_{FDR} value[c]	β	95% CI	SE	*P* value[c]	P_{FDR} value[c]
Incorrect stop contrasted with correct go										
Superior parietal	0.36	0.10 to 0.61	0.01	.005[d]	.04[d]	0.34	0.08 to 0.59	0.01	.009[d]	.04[d]
Inferior parietal	0.26	0.00 to 0.51	0.01	.047[d]	.09	0.32	0.07 to 0.57	0.01	.01[d]	.04[d]
Medial frontal	0.22	–0.03 to 0.48	0.03	.08	.14	0.20	–0.06 to 0.45	0.03	.13	.20
Rostral middle frontal	0.15	–0.10 to 0.41	0.02	.24	.28	0.17	–0.09 to 0.42	0.02	.20	.27
Rostral anterior cingulate	–0.01	–0.27 to 0.24	0.02	.91	.91	–0.03	–0.29 to 0.22	0.02	.81	.88
Posterior cingulate	0.27	0.01 to 0.53	0.01	.04[d]	.09	0.34	0.09 to 0.60	0.01	.008[d]	.04[d]
Incorrect stop contrasted with correct stop										
Superior parietal	0.32	0.08 to 0.52	0.01	.006[d]	.04[d]	0.30	0.05 to 0.48	0.01	.01[d]	.04[d]
Inferior parietal	0.30	0.06 to 0.50	0.01	.01[d]	.04[d]	0.30	0.04 to 0.47	0.01	.02[d]	.04[d]
Medial frontal	0.17	–0.08 to 0.38	0.04	.10	.17	0.22	–0.06 to 0.45	0.03	.11	.17
Rostral middle frontal	–0.03	–0.28 to 0.23	0.02	.83	.83	0.05	–0.21 to 0.31	0.02	.70	.76
Rostral anterior cingulate	–0.06	–0.31 to 0.20	0.02	.66	.76	–0.02	–0.27 to 0.21	0.02	.63	.76
Posterior cingulate	0.28	0.04 to 0.58	0.01	.04[d]	.08	0.34	0.10 to 0.60	0.01	.006[d]	.04[d]

[a]β: standardized coefficient.

[b]SE of β.

[c]All *P* values were adjusted for multiple comparisons using the false discovery rate (FDR) method.

[d]*P*<.05.

Associations Between Cyberbullying at T1 and Mental Health Outcomes/Inhibitory Control at T2

Descriptive statistics for mental health outcomes and inhibitory control at T1 and T2 are presented in Table 3. The LMMs revealed that cyberbullying victimization at T1 significantly predicted increased externalizing problems at T2 (β=0.25, 95% CI 0.06-0.45, SE 0.87, P_{FDR}=.02). In contrast, internalizing problems and inhibitory control performance at T2 were not significantly associated with cyberbullying victimization at T1 (Table 4).

XSL•FO
RenderX

Table 3. Descriptive statistics for mental health outcomes and stop signal task (SST) behavioral performance at T1 and T2 in the Adolescent Brain Cognitive Development study.

	T1		T2	
	Cyberbullying victims (n=61)	No cyberbullying experience (n=1125)	Cyberbullying victims (n=61)	No cyberbullying experience (n=1125)
Mental health outcomes, mean (SD)				
Internalizing problems	49.39 (9.33)	47.65 (10.34)	48.85 (8.76)	47.85 (10.72)
Externalizing problems	46.69 (8.92)	44.17 (9.10)	46.97 (8.90)	43.08 (8.77)
SST behavioral performance, mean (SD)				
Correct stop rate (%)	50.10 (6.60)	50.90 (5.40)	48.80 (6.90)	49.20 (6.30)
Stop signal reaction time (ms)	284.41 (58.00)	276.84 (57.24)	264.49 (56.36)	261.73 (50.77)

Table 4. Associations between cyberbullying victimization at T1 and mental health outcomes/inhibitory control at T2 in the Adolescent Brain Cognitive Development study.

	β[a]	95% CI	SE[b]	*P* value[c]	P_{FDR} value[c]
Mental health outcomes					
Internalizing problems	–0.01	–0.20 to 0.19	1.06	.99	.99
Externalizing problems	0.25	0.06 to 0.45	0.87	.01	.02[d]
SST[e] behavioral performance					
Correct stop rate (%)	–0.02	–0.26 to 0.21	0.01	.85	.85
Stop signal reaction time	–0.07	–0.27 to 0.13	5.14	.49	.85

[a]β: standardized coefficient.

[b]SE of β.

[c]All *P* values were adjusted for multiple comparisons using the false discovery rate (FDR) method.

[d]$P<.05$.

[e]SST: stop signal task.

Mediation Analyses

As shown in Table 5, the mediation effects of parietal gyri or posterior cingulate cortex activated alteration during error processing between cyberbullying victimization at T1 and externalizing problems at T2 did not reach a significant level after correction for multiple comparisons (P_{FDR}s>.74; specific path coefficients are presented in Table S2 in Multimedia Appendix 1 [71-75]).

Table 5. Mediation effects of brain activation during error processing between cyberbullying victimization at T1 and externalizing problems at T2 in the Adolescent Brain Cognitive Development study.

	Left hemisphere				Right hemisphere			
	Indirect effect	95% CI	*P* value[a]	P_{FDR} value[a]	Indirect effect	95% CI	*P* value[a]	P_{FDR} value[a]
Incorrect stop contrasted with correct go								
Superior parietal	0.004	–0.003 to 0.006	.66	>.99	0.001	–0.005 to 0.005	>.99	>.99
Inferior parietal	0.000	–0.003 to 0.004	.87	>.99	–0.004	–0.007 to 0.002	.47	>.99
Posterior cingulate	–0.005	–0.006 to 0.004	.62	>.99	–0.006	–0.007 to 0.005	.65	>.99
Incorrect stop contrasted with correct stop								
Superior parietal	0.003	–0.002 to 0.004	.74	.82	0.003	–0.001 to 0.005	.64	.82
Inferior parietal	0.003	–0.002 to 0.004	.82	.82	0.006	–0.001 to 0.008	.74	.82
Posterior cingulate	0.005	–0.004 to 0.006	.54	.82	0.006	–0.001 to 0.008	.44	.82

[a]All *P* values were adjusted for multiple comparisons using the false discovery rate (FDR) method.

XSL•FO
RenderX

Inverse Probability Weighting Analysis

After inverse probability weighting, the results remained consistent with the primary analyses (Tables S3-S6 in Multimedia Appendix 1 [71-75]). Specifically, cyberbullying victimization at T1 was significantly associated with higher activation in the bilateral superior parietal gyri, inferior parietal gyri, and right posterior cingulate cortex at T2 (Table S3 in Multimedia Appendix 1 [71-75]). Behaviorally, while it significantly predicted increased externalizing problems, its associations with internalizing problems and inhibitory control performance still did not reach a significant level (Table S4 in Multimedia Appendix 1 [71-75]). Moreover, the mediation effects of activation in the parietal gyri or posterior cingulate cortex during error processing between cyberbullying victimization and externalizing problems were also nonsignificant (Tables S5 and S6 in Multimedia Appendix 1 [71-75]).

Discussion

Principal Findings

This retrospective cohort study explored the longitudinal association between cyberbullying victimization, inhibitory control, brain activation during error processing, and mental health problems among American adolescents. Adolescents who experienced cyberbullying victimization demonstrated heightened activation in the parietal gyri and posterior cingulate cortices during error processing. Additionally, these victims showed more subsequent externalizing rather than internalizing problems. However, the mediating effects of brain-activated alteration on the link between cyberbullying victimization and mental health outcomes did not reach significant levels.

Brain Activation During Error Processing

Our study showed that cyberbullying victimization significantly contributed to higher activation in parietal regions and posterior cingulate cortices during error processing, which may reflect their high sensitivity and vigilance to adversity, thus avoiding potential injury [54]. The parietal cortex plays a critical role in an individual's attention regulation, goal orientation, and motor coordination [45,88]. Increased activation of the parietal cortex may not only make cyberbullying victims more vigilant and sensitive to adversity, but also underpin their subsequent behavioral adjustments to distance themselves from potential injury. Victims of cyberbullying experienced web-based, anonymous, and spatially/temporally unlimited harm (eg, verbal scold and satirize) in a web-based environment [10,96], potentially leading them to distance themselves from various harm in any real-life situations. This generalized avoidance tendency could be supported by the parietal gyri, given their involvement in relevant cognitive functions.

The posterior cingulate cortex, a core structure in the default mode network, is closely associated with an individual's self-reflection, attention shift, and emotion regulation [97,98]. However, its interpretation in our findings requires caution, as this region typically deactivates during goal-directed tasks in typical populations [99]. Therefore, the observed higher activation among victims may not reflect an absolute increase above a resting baseline, but rather a reduction in the deactivation. This pattern could signify either a failure to suppress self-referential processing, potentially due to intrusive thoughts related to victimization experiences, or the active recruitment of cognitive resources for heightened internal monitoring and emotion regulation to stabilize the internal state during adversity [100]. It is also worth noting that the significant change in activation of the inferior parietal and posterior cingulate cortices was only observed in the right hemisphere. Combined with the fact that no significant difference in frontal and anterior cingulate activation was found between cyberbully victims and nonvictims, it seems that the effects of cyberbullying partially lead to an increased cognitive load, but do not widely damage the error-monitoring circuits and cognitive control network.

From the perspective of latent vulnerability theory [101,102], a nonsignificant change in the anterior cingulate cortex but significantly higher activation in posterior cingulate and parietal cortices could be viewed as adaptive responses and adjustments, rather than mere damage after cyberbullying victimization. Furthermore, these changes could serve as generalized neural impacts of this experience, as this study only used category variables to define cyberbullying but did not consider its intensity. However, given that this form of victimization has been shown to elicit higher cortisol secretion levels and greater perceived stress in adolescents [58], which could disrupt the hypothalamic-pituitary-adrenal axis feedback [103], further studies are warranted to further explore the relationship between characteristic neural alterations and the intensity of cyberbullying victimization.

Externalizing Problems

The relationship between cyberbullying victimization and externalizing problems was consistent with previous studies [13-15]. Yet the mediating effects of brain over-activation during error processing between cyberbullying victimization and externalizing problems did not reach a significant level. As mentioned above, the higher activation of parietal and posterior cingulate cortices in SST may serve as adaptive adjustments, rather than severe neural circuit damage. Moreover, dealing with real-life "errors" may generally require more cognitive resources and neural load than processing failed inhibition in SST. In other words, the standardized stimuli and the monotonous failed inhibition in SST may not adequately elicit the actual neural changes of cyberbullying victimization. Consequently, despite certain differences in brain activation measured by the SST between cyberbullying victims and nonvictims, such differences could not sufficiently mediate the relationship between cyberbullying victimization and externalizing problems.

The sustained performance of inhibitory control in cyberbullying victims also provides a potential explanation for the insufficient mediated effect of brain-activated alteration. Previous studies suggested that poor inhibitory control is a predictor of externalizing problems [104,105]. However, no significant predictive difference was found in inhibitory control performance between cyberbullying victims and nonvictims of this study, which aligned with Lim et al [53] and Carrion et al

XSL•FO
RenderX

[106]. Therefore, we speculate that the higher neurological load in the parietal gyri and posterior cingulate cortices did not lead to more failed inhibition in SST (ie, decompensatory effect), but rather maintained the performance of inhibitory control (ie, compensatory effect) with heightened attention and self-reflection, resulting in no evidence for a significant mediation effect for externalizing problems. In general, both the potential lack of representation and the protective bias role of altered brain activation measured by SST may contribute to this study's failure to indicate the significant mediated brain effects.

Based on existing studies, the link between cyberbullying victimization and externalizing problems may be effectively explained by other mechanisms. One potential mechanism is that the experience of cyberbullying elicited the victim's aggressive tendencies (eg, hostility and retaliatory motivation) [14,107,108]. For example, Tong et al [14] engaged 464 rural adolescents to examine the association between cyberbullying victimization and malevolent creativity. The results revealed that victims mediated increased hostile attribution leading to serious malevolent creativity. Another potential mechanism is that cyberbullying victimization led to the adoption of externalizing behaviors through imitation [107,109]. Specifically, according to the general learning model [110], we speculate that cyberbullying is a situational factor that influences victims' observational learning processes. These influences manifested in internal states, including perception and emotion, and subsequently contributed to externalizing problems. For instance, victims of cyberbullying perceived the characteristics of the bullying conducted in the web-based environment (eg, anonymity and no strength differential). They may develop a positive attitude towards such characteristics, thereby engaging in cyberbullying [107,109].

Overall, this study revalidated the longitudinal association between cyberbullying and externalizing problems but failed to indicate the significant neural mediation measured by SST. More studies are needed in the future to comprehensively explore other potential mediated mechanisms.

Internalizing Problems

Findings in previous studies regarding whether the cyberbullying victimization could predict internalizing problems (eg, depression and anxiety) are controversial [9,10,17,18]. Similarly with Chu et al [17] and Frison et al [18], this study found that cyberbullying victimization did not significantly predict internalizing problems. One potential explanation is that individuals have the ability to regulate their emotions, which might mitigate the internalized impacts of cyberbullying [111,112]. Since the emotional regulation is closely linked to the activation of the posterior cingulate cortex and the performance of inhibitory control [97,100], we speculate that heightened activation of the posterior cingulate cortex during error processing may reflect cyberbullying victims' increased load of emotional regulation to cope with adversity. Such increased load may also serve as the neural compensation and protection, through which cyberbullying victimization could not directly and severely disrupt the individual's internal state.

Another potential explanation is the different measurement methods used for cyberbullying and emotions. Specifically, cyberbullying is a trait measure (ie, assessing the frequency of cyberbullying within a certain time frame by self-report) [113], and emotions are conceptualized as state processes (ie, measurement dependent on current arousal) [114]. The mismatch between these measurement methods leads to an inability to sensitively capture the immediate impact of cyberbullying on emotions. Therefore, we speculate that the significant longitudinal effect of cyberbullying on internalizing problems only emerges when there is sufficient emotional destructiveness and prolonged duration of victimization, which aligns with the hypothesis proposed by Frison et al [18]. Given that this study used a categorical variable to define cyberbullying, which may be more reflective of general victimization, future studies should explore whether there is a longitudinal association between the intensity and frequency of cyberbullying and the severity of neural damage, as well as internalizing problems [21].

Theoretical and Practical Implications

For theoretical implications, our findings promote the integration of the error detection theory [35,36] and psychopathological outcomes. The potential neurocompensatory response observed in the posterior cingulate and parietal cortices during failed inhibition could provide a novel account for the stress process model [22], illustrating how chronic social stress may manifest as heightened neural sensitivity to errors while maintaining behavioral performance. Similarly, these neural findings offer a mechanistic elaboration of the general aggression model [23,115] for externalizing problems, specifying how victimization might strain cognitive control resources without immediate behavioral manifestations. For practical implications, the significant longitudinal association with externalizing problems highlights the importance of incorporating cyberbullying victimization screening into mental health assessments at an early stage. Furthermore, the neural signature identified here might inform the development of targeted interventions, such as cognitive training protocols designed to reinforce compensatory processes, which hold potential for enhancing emotion regulation and mitigating the risk of externalizing behaviors in affected youth [116,117].

Limitations and Future Directions

This study has a few limitations. The first limitation stems from data availability in the ABCD Study Release 5.1. The analytical sample size was reduced primarily due to the partial availability of the 4-year follow-up (T2) data. This attrition likely contributed to the relatively low prevalence of cyberbullying victims (≈5%), which may limit statistical power to detect smaller effects. Furthermore, although our analysis did not reveal systematic differences in baseline characteristics between included and excluded participants, the potential selection bias (eg, the unexpected similarity in offline victimization levels between victims and nonvictims) should be considered when interpreting the results and their generalizability.

Second, the analysis was constrained to only 2 time points (T1 and T2) available in this data release. This limits the representation of developmental trajectories and the more rigorous longitudinal mediation analysis (eg, a T1→T2→T3

XSL•FO
RenderX

pathway). Future studies should incorporate larger samples of affected individuals and more waves of neuroimaging data across development to examine how cyberbullying victimization influences mental health via neural changes over time.

Third, cyberbullying was analyzed as a categorical variable (victims vs nonvictims) in this study. This approach was adopted because the item measuring frequency/intensity had a high rate of missing data in our analytical sample, which precludes examining the impact of victimization severity on mental health problems, especially internalizing problems. Moreover, this study only focused on externalizing problems, internalizing problems, and brain activation during error processing. Cyberbullying may be significantly associated with other mental/physical health problems (eg, eating disorder, suicide, and sleep disturbance) [79,118,119] as well as social media use [67,71], it is necessary to comprehensively examine the longitudinal impacts of cyberbullying on a variety of health problems and brain states (eg, morphology, resting-state connectivity, and other dynamic states).

Conclusions

Drawing on the ABCD study, this fMRI investigation advances prior cross-sectional or behavioral research by adopting a longitudinal neurobiological perspective. It reveals that cyberbullying victimization predicts significantly heightened activation in parietal and posterior cingulate regions during error processing, alongside increased externalizing problems over time. The absence of significant deficits in inhibitory control and mediation effects suggests that these neural impacts may be better characterized by a state of heightened sensitivity and compensatory engagement after victimization rather than direct impairment. In general, these findings identify the error-processing network as a candidate neural system for future study and motivate continued exploration of other neural pathways linking cyberbullying victimization to mental health, with the ultimate aim of mitigating negative consequences.

Acknowledgments

This study was supported by the GuangDong Basic and Applied Basic Research Foundation (#2023A1515110394) and the Guangzhou Basic and Applied Basic Research Foundation (#2024A04J5080). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript. The authors acknowledge that generative artificial intelligence (AI) technology was used to identify and edit word errors. Data used in the preparation of this article were obtained with permission from the Adolescent Brain Cognitive Development (ABCD) Study, held in the NIMH Data Archive. This is a multisite, longitudinal study designed to recruit more than 10,000 children aged 9-10 years and follow them over 10 years into early adulthood. The ABCD Study is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041048, U01DA050989, U01DA051016, U01DA041022, U01DA051018, U01DA051037, U01DA050987, U01DA041174, U01DA041106, U01DA041117, U01DA041028, U01DA041134, U01DA050988, U01DA051039, U01DA041156, U01DA041025, U01DA041120, U01DA051038, U01DA041148, U01DA041093, U01DA041089, U24DA041123, and U24DA041147. ABCD consortium investigators designed and implemented the study, and provided data, but did not necessarily participate in the analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the National Institutes of Health or ABCD consortium investigators. The ABCD data repository grows and changes over time.

Data Availability

Data were obtained from the 2-year follow-up and 4-year follow-up of the Adolescent Brain Cognitive Development (ABCD) Study (data release 5.1). The data that support the findings of this study are available at the NIMH Data Archive upon data access request.

Authors' Contributions

XZ, CX, and BQ contributed to the conceptualization of the study. The methodology was developed by XZ, YC, and BQ. XZ, YC, and BQ prepared the original draft of the manuscript, while XZ, CX, CZ, YC, and BQ contributed to the review and editing of the manuscript. Supervision was provided by YC and BQ, and funding acquisition was managed by BQ. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional material.
[DOCX File , 159 KB - jmir_v28i1e75126_app1.docx]

Multimedia Appendix 2
STROBE checklist.

XSL•FO
RenderX

[DOCX File , 22 KB - jmir_v28i1e75126_app2.docx]

References

1. Brunstein Klomek A, Marrocco F, Kleinman M, Schonfeld IS, Gould MS. Bullying, depression, and suicidality in adolescents. J Am Acad Child Adolesc Psychiatry 2007;46(1):40-49. [doi: 10.1097/01.chi.0000242237.84925.18] [Medline: 17195728]
2. Fraguas D, Díaz-Caneja CM, Ayora M, Durán-Cutilla M, Abregú-Crespo R, Ezquiaga-Bravo I, et al. Assessment of school anti-bullying interventions: a meta-analysis of randomized clinical trials. JAMA Pediatr 2021;175(1):44-55 [FREE Full text] [doi: 10.1001/jamapediatrics.2020.3541] [Medline: 33136156]
3. Kennedy RS. Bullying trends in the United States: a meta-regression. Trauma Violence Abuse 2021;22(4):914-927. [doi: 10.1177/1524838019888555] [Medline: 31789118]
4. Patchin J, Hinduja S. Cyberbullying among tweens in the United States: prevalence, impact, and helping behaviors. J Early Adolesc 2021;42(3):414-430. [doi: 10.1177/02724316211036740]
5. Englander E, Donnerstein E, Kowalski R, Lin C, Parti K. Defining cyberbullying. Pediatrics 2017;140(Suppl 2):S148-S151. [doi: 10.1542/peds.2016-1758U] [Medline: 29093051]
6. Aboujaoude E, Savage MW, Starcevic V, Salame WO. Cyberbullying: review of an old problem gone viral. J Adolesc Health 2015;57(1):10-18. [doi: 10.1016/j.jadohealth.2015.04.011] [Medline: 26095405]
7. Baldry AC, Farrington DP, Sorrentino A. “Am I at risk of cyberbullying”? a narrative review and conceptual framework for research on risk of cyberbullying and cybervictimization: the risk and needs assessment approach. Aggress Violent Behav 2015;23:36-51. [doi: 10.1016/j.avb.2015.05.014]
8. Kowalski RM, Giumetti GW, Schroeder AN, Lattanner MR. Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth. Psychol Bull 2014;140(4):1073-1137. [doi: 10.1037/a0035618] [Medline: 24512111]
9. Gámez-Guadix M, Orue I, Smith PK, Calvete E. Longitudinal and reciprocal relations of cyberbullying with depression, substance use, and problematic internet use among adolescents. J Adolesc Health 2013;53(4):446-452. [doi: 10.1016/j.jadohealth.2013.03.030] [Medline: 23721758]
10. Li L, Jing R, Jin G, Song Y. Longitudinal associations between traditional and cyberbullying victimization and depressive symptoms among young Chinese: a mediation analysis. Child Abuse Negl 2023;140:106141. [doi: 10.1016/j.chiabu.2023.106141] [Medline: 36958096]
11. Fahy AE, Stansfeld SA, Smuk M, Smith NR, Cummins S, Clark C. Longitudinal associations between cyberbullying involvement and adolescent mental health. J Adolesc Health 2016;59(5):502-509. [doi: 10.1016/j.jadohealth.2016.06.006] [Medline: 27528471]
12. Rose CA, Tynes BM. Longitudinal associations between cybervictimization and mental health among U.S. adolescents. J Adolesc Health 2015;57(3):305-312 [FREE Full text] [doi: 10.1016/j.jadohealth.2015.05.002] [Medline: 26115909]
13. Lu Y, Avellaneda F, Torres ED, Rothman EF, Temple JR. Adolescent cyberbullying and weapon carrying: cross-sectional and longitudinal associations. Cyberpsychol Behav Soc Netw 2019;22(3):173-179 [FREE Full text] [doi: 10.1089/cyber.2018.0463] [Medline: 30614736]
14. Tong D, Shi Y, Gu X, Lu P. Bullying victimization and malevolent creativity in rural adolescents: the longitudinal mediational role of hostile attribution. Cyberpsychol Behav Soc Netw 2024;27(6):420-425. [doi: 10.1089/cyber.2023.0499] [Medline: 38511278]
15. Gong X, Huebner ES, Tian L. Bullying victimization and developmental trajectories of internalizing and externalizing problems: the moderating role of locus of control among children. Res Child Adolesc Psychopathol 2021;49(3):351-366. [doi: 10.1007/s10802-020-00752-2] [Medline: 33404945]
16. Jung Y, Leventhal B, Kim YS, Park TW, Lee S, Lee M, et al. Cyberbullying, problematic internet use, and psychopathologic symptoms among Korean youth. Yonsei Med J 2014;55(3):826-830 [FREE Full text] [doi: 10.3349/ymj.2014.55.3.826] [Medline: 24719154]
17. Chu X, Fan C, Lian S, Zhou Z. Does bullying victimization really influence adolescents' psychosocial problems? a three-wave longitudinal study in China. J Affect Disord 2019;246:603-610. [doi: 10.1016/j.jad.2018.12.103] [Medline: 30605879]
18. Frison E, Subrahmanyam K, Eggermont S. The short-term longitudinal and reciprocal relations between peer victimization on facebook and adolescents' well-being. J Youth Adolesc 2016;45(9):1755-1771. [doi: 10.1007/s10964-016-0436-z] [Medline: 26880284]
19. Pabian S, Vandebosch H. An investigation of short-term longitudinal associations between social anxiety and victimization and perpetration of traditional bullying and cyberbullying. J Youth Adolesc 2016;45(2):328-339. [doi: 10.1007/s10964-015-0259-3] [Medline: 25687265]
20. van den Eijnden R, Vermulst A, van Rooij AJ, Scholte R, van de Mheen D. The bidirectional relationships between online victimization and psychosocial problems in adolescents: a comparison with real-life victimization. J Youth Adolesc 2014;43(5):790-802 [FREE Full text] [doi: 10.1007/s10964-013-0003-9] [Medline: 23979296]
21. Kwan I, Dickson K, Richardson M, MacDowall W, Burchett H, Stansfield C, et al. Cyberbullying and children and young people's mental health: a systematic map of systematic reviews. Cyberpsychol Behav Soc Netw 2020;23(2):72-82 [FREE Full text] [doi: 10.1089/cyber.2019.0370] [Medline: 31977251]

XSL•FO
RenderX

22. Wang L. The effects of cyberbullying victimization and personality characteristics on adolescent mental health: an application of the stress process model. Youth Soc 2021;54(6):935-956. [doi: 10.1177/0044118x211008927]
23. Zhao Y, An W, Zhang Y, Yang C, Li W, Li J, et al. Cyberbullying victimization and perpetration in adolescents: a longitudinal moderated mediation model. J Youth Adolesc 2025;54(4):890-902. [doi: 10.1007/s10964-024-02105-4] [Medline: 39466550]
24. Aron AR. The neural basis of inhibition in cognitive control. Neuroscientist 2007;13(3):214-228. [doi: 10.1177/1073858407299288] [Medline: 17519365]
25. Rahapsari S, Levita L. The impact of adverse childhood experiences on cognitive control across the lifespan: a systematic review and meta-analysis of prospective studies. Trauma Violence Abuse 2025;26(4):712-733 [FREE Full text] [doi: 10.1177/15248380241286812] [Medline: 39396188]
26. Westermann N, Busching R, Klein AM, Warschburger P. The longitudinal interplay between adverse peer experiences and self-regulation facets: a prospective analysis during middle childhood. Res Child Adolesc Psychopathol 2024;52(2):293-308 [FREE Full text] [doi: 10.1007/s10802-023-01117-1] [Medline: 37782355]
27. Savopoulos P, Brown S, Anderson PJ, Gartland D, Bryant C, Giallo R. Intimate partner violence during infancy and cognitive outcomes in middle childhood: Results from an Australian community-based mother and child cohort study. Child Dev 2022;93(4):e396-e411. [doi: 10.1111/cdev.13736] [Medline: 35137950]
28. Frenkel TI, Donzella B, Frenn KA, Rousseau S, Fox NA, Gunnar MR. Moderating the risk for attention deficits in children with pre-adoptive adversity: the protective role of shorter duration of out of home placement and children's enhanced error monitoring. J Abnorm Child Psychol 2020;48(9):1115-1128 [FREE Full text] [doi: 10.1007/s10802-020-00671-2] [Medline: 32607754]
29. Golm D, Sarkar S, Mackes NK, Fairchild G, Mehta MA, Rutter M, ERA Young Adult Follow Up study team, et al. The impact of childhood deprivation on adult neuropsychological functioning is associated with ADHD symptom persistence. Psychol Med 2021;51(15):2675-2684. [doi: 10.1017/S0033291720001294] [Medline: 32419675]
30. Amédée LM, Cyr C, Jean-Thorn A, Hébert M. Executive functioning in child victims of sexual abuse: a multi-informant comparative study. Child Abuse Negl 2024;152:106737 [FREE Full text] [doi: 10.1016/j.chiabu.2024.106737] [Medline: 38564916]
31. Wei X, Lü W. Childhood abuse and depressive symptoms in adolescents: affective inhibitory control as a mediator and sensory processing sensitivity as a moderator. Child Abuse Negl 2024;154:106957. [doi: 10.1016/j.chiabu.2024.106957] [Medline: 39053224]
32. Kahle S, Utendale WT, Widaman KF, Hastings PD. Parasympathetic regulation and inhibitory control predict the development of externalizing problems in early childhood. J Abnorm Child Psychol 2018;46(2):237-249. [doi: 10.1007/s10802-017-0305-6] [Medline: 28493111]
33. Poland SE, Monks CP, Tsermentseli S. Cool and hot executive function as predictors of aggression in early childhood: differentiating between the function and form of aggression. Br J Dev Psychol 2016;34(2):181-197. [doi: 10.1111/bjdp.12122] [Medline: 26615980]
34. Sætren SS, Augusti E, Hafstad GS. Affective inhibitory control and risk for internalizing problems in adolescents exposed to child maltreatment: a population-based study. J Abnorm Psychol 2021;130(2):113-125. [doi: 10.1037/abn0000582] [Medline: 33315413]
35. Coles MG, Scheffers MK, Holroyd CB. Why is there an ERN/Ne on correct trials? response representations, stimulus-related components, and the theory of error-processing. Biol Psychol 2001;56(3):173-189. [doi: 10.1016/s0301-0511(01)00076-x] [Medline: 11399349]
36. Falkenstein M, Hoormann J, Christ S, Hohnsbein J. ERP components on reaction errors and their functional significance: a tutorial. Biol Psychol 2000;51(2-3):87-107. [doi: 10.1016/s0301-0511(99)00031-9] [Medline: 10686361]
37. Botvinick MM, Cohen JD, Carter CS. Conflict monitoring and anterior cingulate cortex: an update. Trends Cogn Sci 2004;8(12):539-546. [doi: 10.1016/j.tics.2004.10.003] [Medline: 15556023]
38. Yeung N, Botvinick MM, Cohen JD. The neural basis of error detection: conflict monitoring and the error-related negativity. Psychol Rev 2004;111(4):931-959. [doi: 10.1037/0033-295x.111.4.939] [Medline: 15482068]
39. Ridderinkhof KR, Ullsperger M, Crone EA, Nieuwenhuis S. The role of the medial frontal cortex in cognitive control. Science 2004;306(5695):443-447. [doi: 10.1126/science.1100301] [Medline: 15486290]
40. Stevens M, Kiehl K, Pearlson G, Calhoun V. Brain network dynamics during error commission. Hum Brain Mapp 2009;30(1):24-37 [FREE Full text] [doi: 10.1002/hbm.20478] [Medline: 17979124]
41. Fava NM, Trucco EM, Martz ME, Cope LM, Jester JM, Zucker RA, et al. Childhood adversity, externalizing behavior, and substance use in adolescence: mediating effects of anterior cingulate cortex activation during inhibitory errors. Dev Psychopathol 2019;31(4):1439-1450 [FREE Full text] [doi: 10.1017/S0954579418001025] [Medline: 30585564]
42. Stevens JS, Ely TD, Sawamura T, Guzman D, Bradley B, Ressler KJ, et al. Childhood maltreatment predicts reduced inhibition-related activity in the rostral anterior cingulate in PTSD, but not trauma-exposed controls. Depress Anxiety 2016;33(7):614-622 [FREE Full text] [doi: 10.1002/da.22506] [Medline: 27062552]
43. Iannaccone R, Hauser TU, Staempfli P, Walitza S, Brandeis D, Brem S. Conflict monitoring and error processing: new insights from simultaneous EEG-fMRI. Neuroimage 2015;105:395-407. [doi: 10.1016/j.neuroimage.2014.10.028] [Medline: 25462691]

XSL•FO
RenderX

44. Wessel JR. An adaptive orienting theory of error processing. Psychophysiology 2018;55(3). [doi: 10.1111/psyp.13041] [Medline: 29226960]
45. Corbetta M, Shulman GL. Control of goal-directed and stimulus-driven attention in the brain. Nat Rev Neurosci 2002;3(3):201-215. [doi: 10.1038/nrn755] [Medline: 11994752]
46. Dosenbach NUF, Fair DA, Miezin FM, Cohen AL, Wenger KK, Dosenbach RAT, et al. Distinct brain networks for adaptive and stable task control in humans. Proc Natl Acad Sci U S A 2007;104(26):11073-11078 [FREE Full text] [doi: 10.1073/pnas.0704320104] [Medline: 17576922]
47. Breukelaar IA, Antees C, Grieve SM, Foster SL, Gomes L, Williams LM, et al. Cognitive control network anatomy correlates with neurocognitive behavior: a longitudinal study. Hum Brain Mapp 2017;38(2):631-643 [FREE Full text] [doi: 10.1002/hbm.23401] [Medline: 27623046]
48. Niendam TA, Laird AR, Ray KL, Dean YM, Glahn DC, Carter CS. Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions. Cogn Affect Behav Neurosci 2012;12(2):241-268 [FREE Full text] [doi: 10.3758/s13415-011-0083-5] [Medline: 22282036]
49. McLaughlin KA, Weissman D, Bitrán D. Childhood adversity and neural development: a systematic review. Annu Rev Dev Psychol 2019;1:277-312 [FREE Full text] [doi: 10.1146/annurev-devpsych-121318-084950] [Medline: 32455344]
50. Teicher MH, Samson JA, Anderson CM, Ohashi K. The effects of childhood maltreatment on brain structure, function and connectivity. Nat Rev Neurosci 2016;17(10):652-666. [doi: 10.1038/nrn.2016.111] [Medline: 27640984]
51. Keator DB, Salgado F, Madigan C, Murray S, Norris S, Amen D. Adverse childhood experiences, brain function, and psychiatric diagnoses in a large adult clinical cohort. Front Psychiatry 2024;15:1401745 [FREE Full text] [doi: 10.3389/fpsyt.2024.1401745] [Medline: 39469474]
52. Lackner CL, Santesso DL, Dywan J, O'Leary DD, Wade TJ, Segalowitz SJ. Adverse childhood experiences are associated with self-regulation and the magnitude of the error-related negativity difference. Biol Psychol 2018;132:244-251. [doi: 10.1016/j.biopsycho.2018.01.006] [Medline: 29309827]
53. Lim L, Hart H, Mehta M, Simmons A, Mirza K, Rubia K. Neural correlates of error processing in young people with a history of severe childhood abuse: an fMRI study. Am J Psychiatry 2015;172(9):892-900. [doi: 10.1176/appi.ajp.2015.14081042] [Medline: 25882324]
54. Santesso DL, Dzyundzyak A, Segalowitz SJ. Age, sex and individual differences in punishment sensitivity: factors influencing the feedback-related negativity. Psychophysiology 2011;48(11):1481-1489. [doi: 10.1111/j.1469-8986.2011.01229.x] [Medline: 21711354]
55. Abi-Dargham A, Moeller SJ, Ali F, DeLorenzo C, Domschke K, Horga G, et al. Candidate biomarkers in psychiatric disorders: state of the field. World Psychiatry 2023;22(2):236-262 [FREE Full text] [doi: 10.1002/wps.21078] [Medline: 37159365]
56. Hajcak G, Klawohn J, Meyer A. The utility of event-related potentials in clinical psychology. Annu Rev Clin Psychol 2019;15:71-95. [doi: 10.1146/annurev-clinpsy-050718-095457] [Medline: 31067414]
57. Olvet DM, Hajcak G. The error-related negativity (ERN) and psychopathology: toward an endophenotype. Clin Psychol Rev 2008;28(8):1343-1354 [FREE Full text] [doi: 10.1016/j.cpr.2008.07.003] [Medline: 18694617]
58. González-Cabrera J, Calvete E, León-Mejía A, Pérez-Sancho C, Peinado J. Relationship between cyberbullying roles, cortisol secretion and psychological stress. Comput Human Behav 2017;70:153-160. [doi: 10.1016/j.chb.2016.12.054]
59. McLoughlin LT, Shan Z, Broadhouse KM, Lagopoulos J, Winks N, Hermens DF. Elucidating the neurobiology of cyberbullying using functional Magnetic Resonance Imaging (fMRI): a hypothesis. Aggress Violent Behav 2020;50:101360. [doi: 10.1016/j.avb.2019.101360]
60. McLoughlin LT, Shan Z, Broadhouse KM, Winks N, Simcock G, Lagopoulos J, et al. Neurobiological underpinnings of cyberbullying: a pilot functional magnetic resonance imaging study. Hum Brain Mapp 2020;41(6):1495-1504 [FREE Full text] [doi: 10.1002/hbm.24890] [Medline: 31797458]
61. McLoughlin LT, Lagopoulos J, Hermens DF. Cyberbullying and adolescent neurobiology. Front Psychol 2020;11:1511 [FREE Full text] [doi: 10.3389/fpsyg.2020.01511] [Medline: 32670171]
62. Crone EA, Dahl RE. Understanding adolescence as a period of social-affective engagement and goal flexibility. Nat Rev Neurosci 2012;13(9):636-650. [doi: 10.1038/nrn3313] [Medline: 22903221]
63. Steinberg L. Cognitive and affective development in adolescence. Trends Cogn Sci 2005;9(2):69-74. [doi: 10.1016/j.tics.2004.12.005] [Medline: 15668099]
64. Rubia K. Functional brain imaging across development. Eur Child Adolesc Psychiatry 2013;22(12):719-731 [FREE Full text] [doi: 10.1007/s00787-012-0291-8] [Medline: 22729957]
65. Tamnes CK, Ostby Y, Fjell AM, Westlye LT, Due-Tønnessen P, Walhovd KB. Brain maturation in adolescence and young adulthood: regional age-related changes in cortical thickness and white matter volume and microstructure. Cereb Cortex 2010;20(3):534-548. [doi: 10.1093/cercor/bhp118] [Medline: 19520764]
66. Cooke EM, Connolly EJ, Boisvert DL, Hayes BE. A systematic review of the biological correlates and consequences of childhood maltreatment and adverse childhood experiences. Trauma Violence Abuse 2023;24(1):156-173. [doi: 10.1177/15248380211021613] [Medline: 34105421]

XSL•FO
RenderX

67. Crone EA, Konijn EA. Media use and brain development during adolescence. Nat Commun 2018;9(1):588 [FREE Full text] [doi: 10.1038/s41467-018-03126-x] [Medline: 29467362]
68. Casey B, Cannonier T, Conley MI, Cohen AO, Barch DM, Heitzeg MM, ABCD Imaging Acquisition Workgroup. The adolescent brain cognitive development (ABCD) study: imaging acquisition across 21 sites. Dev Cogn Neurosci 2018;32:43-54 [FREE Full text] [doi: 10.1016/j.dcn.2018.03.001] [Medline: 29567376]
69. Garavan H, Bartsch H, Conway K, Decastro A, Goldstein RZ, Heeringa S, et al. Recruiting the ABCD sample: design considerations and procedures. Dev Cogn Neurosci 2018;32:16-22 [FREE Full text] [doi: 10.1016/j.dcn.2018.04.004] [Medline: 29703560]
70. ABCD Human Subjects Study. National Institute for Health. 2024. URL: https://nda.nih.gov/abcd [accessed 2026-01-20]
71. He X, Hu J, Yin M, Zhang W, Qiu B. Screen media use affects subcortical structures, resting-state functional connectivity, and mental health problems in early adolescence. Brain Sci 2023;13(10):1452 [FREE Full text] [doi: 10.3390/brainsci13101452] [Medline: 37891820]
72. Verbruggen F, Aron AR, Band GP, Beste C, Bissett PG, Brockett AT, et al. A consensus guide to capturing the ability to inhibit actions and impulsive behaviors in the stop-signal task. Elife 2019 Apr 29;8:e46323 [FREE Full text] [doi: 10.7554/eLife.46323] [Medline: 31033438]
73. Prinstein MJ, Boergers J, Vernberg EM. Overt and relational aggression in adolescents: social-psychological adjustment of aggressors and victims. J Clin Child Psychol 2001 Dec;30(4):479-491. [doi: 10.1207/S15374424JCCP3004_05] [Medline: 11708236]
74. Felitti VJ, Anda RF, Nordenberg D, Williamson DF, Spitz AM, Edwards V, et al. Reprint of: relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults: the adverse childhood experiences (ACE) study. Am J Prev Med 2019 Jun;56(6):774-786. [doi: 10.1016/j.amepre.2019.04.001] [Medline: 31104722]
75. Lanz M, Maino E. Family environment scale. In: Encyclopedia of Quality of Life and Well-Being Research. Dordrecht: Springer; 2014:2170-2173.
76. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988;83(404):1198-1202. [doi: 10.1080/01621459.1988.10478722]
77. Green P, MacLeod CJ. SIMR?: an R package for power analysis of generalized linear mixed models by simulation. Methods Ecol Evol 2016;7(4):493-498. [doi: 10.1111/2041-210x.12504]
78. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007;370(9596):1453-1457 [FREE Full text] [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
79. Stewart RW, Drescher CF, Maack DJ, Ebesutani C, Young J. The development and psychometric investigation of the cyberbullying scale. J Interpers Violence 2014;29(12):2218-2238. [doi: 10.1177/0886260513517552] [Medline: 24424252]
80. Arnon S, Brunstein Klomek A, Visoki E, Moore TM, Argabright ST, DiDomenico GE, et al. Association of cyberbullying experiences and perpetration with suicidality in early adolescence. JAMA Netw Open 2022;5(6):e2218746 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.18746] [Medline: 35759263]
81. Trompeter N, Testa A, Raney JH, Jackson DB, Al-Shoaibi AAA, Ganson KT, et al. The association between adverse childhood experiences (ACEs), bullying victimization, and internalizing and externalizing problems among early adolescents: examining cumulative and interactive associations. J Youth Adolesc 2024;53(3):744-752 [FREE Full text] [doi: 10.1007/s10964-023-01907-2] [Medline: 38066316]
82. Achenbach TM, Ruffle TM. The child behavior checklist and related forms for assessing behavioral/emotional problems and competencies. Pediatr Rev 2000;21(8):265-271. [doi: 10.1542/pir.21-8-265] [Medline: 10922023]
83. Logan G, Cowan W, Davis K. On the ability to inhibit simple and choice reaction time responses: a model and a method. J Exp Psychol Hum Percept Perform 1984;10(2):276-291. [doi: 10.1037//0096-1523.10.2.276] [Medline: 6232345]
84. Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 1996;29(3):162-173. [doi: 10.1006/cbmr.1996.0014] [Medline: 8812068]
85. Holland D, Kuperman JM, Dale AM. Efficient correction of inhomogeneous static magnetic field-induced distortion in echo planar imaging. Neuroimage 2010;50(1):175-183 [FREE Full text] [doi: 10.1016/j.neuroimage.2009.11.044] [Medline: 19944768]
86. Jovicich J, Czanner S, Greve D, Haley E, van der Kouwe A, Gollub R, et al. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. Neuroimage 2006;30(2):436-443. [doi: 10.1016/j.neuroimage.2005.09.046] [Medline: 16300968]
87. Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 2006;31(3):968-980. [doi: 10.1016/j.neuroimage.2006.01.021] [Medline: 16530430]
88. Hagler DJ, Hatton S, Cornejo MD, Makowski C, Fair DA, Dick AS, et al. Image processing and analysis methods for the adolescent brain cognitive development study. Neuroimage 2019;202:116091 [FREE Full text] [doi: 10.1016/j.neuroimage.2019.116091] [Medline: 31415884]
89. Andersen RA, Buneo CA. Intentional maps in posterior parietal cortex. Annu Rev Neurosci 2002;25:189-220. [doi: 10.1146/annurev.neuro.25.112701.142922] [Medline: 12052908]

XSL•FO
RenderX

90. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Soft 2015;67(1). [doi: 10.18637/jss.v067.i01]
91. Cole DA, Maxwell SE. Testing mediational models with longitudinal data: questions and tips in the use of structural equation modeling. J Abnorm Psychol 2003;112(4):558-577. [doi: 10.1037/0021-843X.112.4.558] [Medline: 14674869]
92. Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage 2012;59(3):2142-2154 [FREE Full text] [doi: 10.1016/j.neuroimage.2011.10.018] [Medline: 22019881]
93. Yang FN, Liu TT, Wang Z. Corticostriatal connectivity mediates the reciprocal relationship between parent-reported sleep duration and impulsivity in early adolescents. J Child Psychol Psychiatry 2023;64(11):1545-1554. [doi: 10.1111/jcpp.13843] [Medline: 37248201]
94. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res 2013;22(3):278-295. [doi: 10.1177/0962280210395740] [Medline: 21220355]
95. Lumley T. Analysis of complex survey samples. J Stat Soft 2004;9(8). [doi: 10.18637/jss.v009.i08]
96. Baiden P, Tadeo SK. Investigating the association between bullying victimization and suicidal ideation among adolescents: evidence from the 2017 youth risk behavior survey. Child Abuse Negl 2020;102:104417. [doi: 10.1016/j.chiabu.2020.104417] [Medline: 32113078]
97. Leech R, Kamourieh S, Beckmann CF, Sharp DJ. Fractionating the default mode network: distinct contributions of the ventral and dorsal posterior cingulate cortex to cognitive control. J Neurosci 2011;31(9):3217-3224 [FREE Full text] [doi: 10.1523/JNEUROSCI.5626-10.2011] [Medline: 21368033]
98. Leech R, Sharp DJ. The role of the posterior cingulate cortex in cognition and disease. Brain 2014;137(Pt 1):12-32 [FREE Full text] [doi: 10.1093/brain/awt162] [Medline: 23869106]
99. Chaarani B, Hahn S, Allgaier N, Adise S, Owens MM, Juliano AC, ABCD Consortium. Baseline brain function in the preadolescents of the ABCD study. Nat Neurosci 2021;24(8):1176-1186 [FREE Full text] [doi: 10.1038/s41593-021-00867-9] [Medline: 34099922]
100. Pessoa L. How do emotion and motivation direct executive control? Trends Cogn Sci 2009;13(4):160-166 [FREE Full text] [doi: 10.1016/j.tics.2009.01.006] [Medline: 19285913]
101. McCrory EJ, Viding E. The theory of latent vulnerability: reconceptualizing the link between childhood maltreatment and psychiatric disorder. Dev Psychopathol 2015;27(2):493-505. [doi: 10.1017/S0954579415000115] [Medline: 25997767]
102. Puetz VB, McCrory E. Exploring the relationship between childhood maltreatment and addiction: a review of the neurocognitive evidence. Curr Addict Rep 2015;2(4):318-325 [FREE Full text] [doi: 10.1007/s40429-015-0073-8] [Medline: 26550550]
103. O'Connor DB, Thayer JF, Vedhara K. Stress and health: a review of psychobiological processes. Annu Rev Psychol 2021;72(1):663-688. [doi: 10.1146/annurev-psych-062520-122331] [Medline: 32886587]
104. Chaku N, Yan R, Kelly DP, Zhang Z, Lopez-Duran N, Weigard AS, et al. 100 days of adolescence: elucidating externalizing behaviors through the daily assessment of inhibitory control. Res Child Adolesc Psychopathol 2024;52(1):93-110 [FREE Full text] [doi: 10.1007/s10802-023-01071-y] [Medline: 37405589]
105. Young SE, Friedman NP, Miyake A, Willcutt EG, Corley RP, Haberstick BC, et al. Behavioral disinhibition: liability for externalizing spectrum disorders and its genetic and environmental relation to response inhibition across adolescence. J Abnorm Psychol 2009;118(1):117-130 [FREE Full text] [doi: 10.1037/a0014657] [Medline: 19222319]
106. Carrion VG, Garrett A, Menon V, Weems CF, Reiss AL. Posttraumatic stress symptoms and brain function during a response-inhibition task: an fMRI study in youth. Depress Anxiety 2008;25(6):514-526. [doi: 10.1002/da.20346] [Medline: 17598145]
107. Barlett C, Gentile D. Attacking others online: the formation of cyberbullying in late adolescence. Psychol Pop Media Cult 2012;1(2):123-135. [doi: 10.1037/a0028113]
108. Wang C, Li X, Xia L. Long-term effect of cybervictimization on displaced aggressive behavior across two years: mutually predicting mediators of hostile emotion and moral disengagement. Comput Hum Behav 2023;141:107611. [doi: 10.1016/j.chb.2022.107611]
109. Barlett CP. Cyberbullying as a learned behavior: theoretical and applied implications. Children (Basel) 2023;10(2):325 [FREE Full text] [doi: 10.3390/children10020325] [Medline: 36832455]
110. Gentile DA, Groves CL, Gentile JR. The general learning model. In: Learning by Playing. England: Oxford University Press; 2014:121-142.
111. Kassis W, Artz S, Scambor C, Scambor E, Moldenhauer S. Finding the way out: a non-dichotomous understanding of violence and depression resilience of adolescents who are exposed to family violence. Child Abuse Negl 2013;37(2-3):181-199. [doi: 10.1016/j.chiabu.2012.11.001] [Medline: 23260117]
112. Lin L, Chien Y, Chen Y, Wu C, Chiou H. Bullying experiences, depression, and the moderating role of resilience among adolescents. Front Public Health 2022;10:872100 [FREE Full text] [doi: 10.3389/fpubh.2022.872100] [Medline: 35692326]
113. Chun J, Lee J, Kim J, Lee S. An international systematic review of cyberbullying measurements. Comput Hum Behav 2020;113:106485. [doi: 10.1016/j.chb.2020.106485]

XSL•FO
RenderX

114. Mulligan K, Scherer KR. Toward a working definition of emotion. Emot Rev 2012;4(4):345-357. [doi: 10.1177/1754073912445818]
115. Teng H, Zhu L, Zhang X, Qiu B. The relationship between cognitive resource consumption during gameplay and postgame aggressive behaviors: between-subjects experiment. JMIR Serious Games 2023;11:e48317 [FREE Full text] [doi: 10.2196/48317] [Medline: 37990585]
116. Kasturiratna KTAS, Hartanto A, Chen CHY, Tong EMW, Majeed NM. Umbrella review of meta-analyses on the risk factors, protective factors, consequences and interventions of cyberbullying victimization. Nat Hum Behav 2025;9(1):101-132. [doi: 10.1038/s41562-024-02011-6] [Medline: 39516404]
117. Zhang X, Teng H, Zhu L, Qiu B. Short-term exposure to aggressive card game: releasing emotion without escalating post-game aggression. Front Psychol 2025;16:1505360 [FREE Full text] [doi: 10.3389/fpsyg.2025.1505360] [Medline: 40236967]
118. Cheng CM, Chu J, Ganson KT, Trompeter N, Testa A, Jackson DB, et al. Cyberbullying and eating disorder symptoms in US early adolescents. Int J Eat Disord 2023;56(12):2336-2342. [doi: 10.1002/eat.24034] [Medline: 37671456]
119. Nagata JM, Yang JH, Singh G, Kiss O, Ganson KT, Testa A, et al. Cyberbullying and sleep disturbance among early adolescents in the U.S. Acad Pediatr 2023;23(6):1220-1225 [FREE Full text] [doi: 10.1016/j.acap.2022.12.007] [Medline: 36581100]

Abbreviations

ABCD: Adolescent Brain Cognitive Development
ACE: adverse childhood experience
FDR: false discovery rate
fMRI: functional magnetic resonance imaging
IRB: Institutional Review Board
LMM: linear mixed model
ROI: regions of interest
SST: stop signal task
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani, S Brini; submitted 30.Mar.2025; peer-reviewed by DA Lopez, M Yin, S Chen; comments to author 12.Sep.2025; accepted 08.Jan.2026; published 18.Feb.2026.

Please cite as:
Zhang X, Xie C, Chen Y, Qiu B
Associations Among Cyberbullying Victimization, Inhibitory Control, Neural Activation of Error Processing, and Mental Health Problems in Adolescents: Neuroimaging, Retrospective Longitudinal Cohort Study Using the Adolescent Brain Cognitive Development Data
J Med Internet Res 2026;28:e75126
URL: https://www.jmir.org/2026/1/e75126
doi:10.2196/75126
PMID:

©Xuanyu Zhang, Chengyan Xie, Yu Chen, Boyu Qiu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Adherence, Acceptability, and Sexual Health Outcomes of the Odeya App–Based Intervention for Sexual Distress in Women With Endometriosis: Randomized Controlled Mixed Methods Trial

Selina Marie Kronthaler[1], MSc; Eden Kosman[1], MSc; Tatjana Tissen-Diabaté[2], MSc; Elena Mühle[1]; Luzia Weber-Schallauer[1], MSc; Therese Pross[3], MD, MBA; Maria Margarete Karsten[3], MD; Klaus Michael Beier[1], MD, PhD; Laura Hatzler[1,4], MD

[1]Institute of Sexology and Sexual Medicine, Charité - Universitätsmedizin Berlin, Berlin, Germany

[2]Institute of Social Medicine, Epidemiology and Health Economics, Charité – Universitätsmedizin Berlin, Berlin, Germany

[3]Department of Gynecology and Breast Center, Charité – Universitätsmedizin Berlin, Berlin, Germany

[4]Department of Urology, Charité - Universitätsmedizin Berlin, Berlin, Germany

Corresponding Author:
Laura Hatzler, MD
Institute of Sexology and Sexual Medicine
Charité - Universitätsmedizin Berlin
Charité Platz 1
Berlin, 10117
Germany
Phone: 49 30 450 617 139
Fax: 49 30 450 529 992
Email: laura.hatzler@charite.de

Abstract

Background: Evidence-based interventions effectively treat sexual dysfunctions. Up to 13.5% of women with gynecological conditions are affected, yet access to therapy is limited. Self-guided digital interventions may offer scalable, accessible first-line support.

Objective: This randomized controlled mixed methods pilot trial evaluated adherence, acceptance, and safety of the Odeya app and changes in sexual and health outcomes among women with sexual dysfunctions and endometriosis.

Methods: Following online and flyer-based recruitment, participants completed an online screening and were randomized to either an intervention group (IG) receiving 8 self-guided app modules targeting biopsychosocial aspects of sexuality or to a control group (CG) receiving routine care. Self-administered online questionnaires were completed at baseline (T0), midintervention (T1), postintervention (T2), and 6-month follow-up (T3). Standardized instruments assessed acceptance (Client Satisfaction Questionnaire-Internet [CSQ-I] and German mHealth App Usability Questionnaire [G-MAUQ]), safety (Inventory for the balanced assessment of Negative Effects of Psychotherapy-Online Intervention), sexual health (Female Sexual Distress Scale-Desire/Arousal/Orgasm [FSDS-DAO], Female Sexual Function Index-German version [FSFI-d], and Partnership Questionnaire), and overall health (Patient-Reported Outcome Measurement Information System-29-Item Profile, Beck Depression Inventory-II, and Generalized Anxiety Disorder-7). Adherence indicators included module completion, dropout rates, and symptom tracker use. Group differences were examined descriptively and using Cohen *d*. Qualitative data were collected through free-list questionnaires from dropouts (n=11) and interviews with completers (IG: n=3; CG: n=2).

Results: A total of 60 women (mean age 31.12, SD 6.67 years) with confirmed or suspected endometriosis and sexual distress (FSDS-DAO score >18) were randomized to the IG (n=29) or CG (n=31). IG participants completed on average 61.2% (4.9/8) of modules; the dropout rate was 65.5% (19/29). Emotional strain, time demands, and technical issues were key barriers causing dropout, while persona-based stories facilitated engagement. Participants wished for more professional interaction. IG completers (n=10, 34.5%) showed lower baseline depression and anxiety but higher sexual distress. Satisfaction was high (CSQ-I=26.60; G-MAUQ=5.38). Although some adverse health changes were reported, findings indicate safety. FSDS-DAO scores decreased in both groups, with mean reductions from baseline of −10.39, −12.61, and −14.98 in the IG and −3.68, −14.83, and −6.92 in the CG from T1 to T3, respectively. Moderate to large between-group effects favoring the IG were observed at T1 (*d*=−0.66) and T3 (*d*=−0.79). Sexual function (FSFI-d) improved only in the IG (T1-T3: *d*=0.16-1.00). Qualitative findings highlighted rediscovery

XSL•FO
RenderX

of positive sexual experiences, improved communication, and increased openness. Both groups reported improvements in anxiety, depression, and physical functioning, with additional gains in emotion regulation, distress reduction, and body awareness reported in the IG. Women emphasized symptom complexity and a need for more professional guidance.

Conclusions: The self-guided intervention was well accepted and showed preliminary improvements among completers. Adherence and sustained engagement seemed shaped by baseline psychosocial health, pointing to a need for tailored adaptations and larger confirmatory trials.

Trial Registration: German Clinical Trials Register DRKS00034351; https://drks.de/search/en/trial/DRKS00034351

(J Med Internet Res 2026;28:e86042) doi:10.2196/86042

KEYWORDS

app-based intervention; digital health intervention; endometriosis; randomized controlled trial; sexual distress; sexual dysfunction

Introduction

Overview

Sexual health is a fundamental aspect of overall well-being and quality of life [1-3]. Sexual dysfunctions, such as problems with sexual desire, arousal, orgasm, or pain, accompanied by clinically relevant distress, as defined by the *ICD-11* (*International Classification of Diseases, 11th Revision*) [4], were reported by 17.5% of women in a representative German study [5]. The biopsychosocial risk factors include relationship-related difficulties, poor mental health, chronic conditions, cultural factors, and lack of knowledge and experience [6-10].

Among individuals with chronic diseases, including gynecological conditions such as endometriosis, sexual health issues are particularly pronounced [11]. In a recent representative German study, 75.2% of those with chronic conditions reported problems in sexual function, and 19.3% met *ICD-11* criteria for sexual dysfunction, with 2.56-fold higher odds compared to individuals without chronic conditions [12]. Among all chronic condition groups, women with gynecological conditions showed the highest prevalence of problems in sexual function (84.3%), and 13.5% of these women met criteria for sexual dysfunction with distress [12]. Chronic diseases frequently contribute to sexual dysfunction through physical, hormonal, psychological, and treatment-related factors [13-15], with downstream impacts on mental health, relationship satisfaction, and health care costs [16-19].

Endometriosis affects 6.8% of women worldwide [20] and substantially impairs sexual health and quality of life [21,22]. The association between endometriosis and sexual dysfunction is well-documented, with many reporting sexual pain, decreased sexual satisfaction, and overall reduced sexual functioning [11,23]. Surgical interventions may further worsen outcomes [24]. Care access is hindered by shame and stigma, insufficient awareness, high costs, and gaps in provider training [12,25-30].

Effective treatment for sexual dysfunction with distress requires a personalized, multimodal, interdisciplinary approach addressing the individual's set of biopsychosocial etiological factors [10,31-36]. Recommended strategies combine somatic interventions (eg, pelvic floor therapy and hormonal and medical treatments) [37-40] with sex and couples therapy (eg, sensate focus and communication training), educational components (eg, psychoeducation on anatomy and physiology) [33,41], and lifestyle-based strategies such as adapted physical activity [42,43]. Furthermore, evidence supports strengthening the mind-body integration through exercises on body perception, mindfulness, and reflective techniques [41,44-46]. Underlying mental health and somatic conditions should always be addressed in interdisciplinary approaches, ideally by a multidisciplinary team trained in sexual medicine [35,47].

Despite high prevalence, sexual dysfunctions remain underreported and undertreated in Germany [12]. Persistent access barriers—including limited specialist availability, long waiting lists, and regional disparities—contrast with strong interest in digital interventions such as app- or web-based programs with exercises and educational content on sexual health [12,29,48].

Digital self-help interventions can help overcome these barriers [16], offering accessibility, anonymity, and cost-effectiveness [49-51]. Online treatments—especially cognitive-behavioral and mindfulness-based programs—have shown moderate to large effects on sexual function and satisfaction [50,52-56]. In 2019, the German Digital Healthcare Act introduced a regulatory framework for the approval and reimbursement of software as a medical device, referred to as digital health applications (DiGAs) [57]. Currently, 2 sexual health-related DiGAs are approved and permanently listed in the registry of the German competent authority (Bundesinstitut für Arzneimittel und Medizinprodukte [BfArM]): HelloBetter Vaginismus for vaginismus [50,58] and Kranus Edera for erectile dysfunction [59]. Additional DiGAs targeting urogenital health include Endo App for endometriosis symptom management [60,61] and Kranus Lutera for lower tract symptoms [62]. Negotiated 90-day prices averaged €221.09 (US $258.70) [63], with full statutory reimbursement.

Patient-centered digital intervention development [64] has demonstrated high user satisfaction [65,66] and positive outcomes across diverse populations, including cancer [67-69], mental health conditions such as depression and anxiety [70-72], and chronic pain [73-75]. Yet adherence and engagement remain challenging—especially in self-guided formats [76-82]—and effects on engagement are mixed [73,83-89]. A major gap persists in digital interventions addressing sexual distress in women with gynecological conditions such as endometriosis, underscoring the need for tailored, patient-centered solutions [90]. Existing digital tools rarely address sexuality as a relational

XSL•FO
RenderX

resource; insights from syndyastic sex therapy may offer a useful framework for future digital models [91].

Objective

This study evaluated the pilot implementation of a self-guided smartphone app intervention (Odeya) in women with sexual distress and diagnosed or suspected endometriosis. The Odeya app was developed using evidence-based content for patients with sexual dysfunctions in conjunction with clinically relevant distress and was tailored, within a patient-centered framework, to the specific needs of women with endometriosis. The primary objectives were to assess (1) adherence, (2) acceptance, and (3) safety. The secondary objective was to exploratorily examine effects on sexual and overall health-related outcomes using a mixed methods design.

Methods

Trial Design

A preceding patient-centered, iterative development phase informed the intervention and functioned as a preuse acceptability assessment of burden, usability, relevance, and coherence (E Kosman, MSc, et al, unpublished data, 2026). Building on this foundation, this open-label, 2-arm pilot randomized controlled trial (July 2024 to July 2025) used an expansion-type mixed methods design, collecting quantitative and qualitative data to enrich and explain emerging findings (see Multimedia Appendix 1 [6-8,16,33,41,45,47,64,65,92-127] for design rationale and prestudy procedures) [128]. Longitudinal self-reported data were collected in the intervention group (IG) and control group (CG) at baseline (T0) after randomization, peritreatment (T1; week 5 in the CG/after completing module 5 in the IG), posttreatment (T2; week 8 in the CG/after completing module 8 in the IG), and at 6-month follow-up (T3). The IG received Odeya app access; the CG was allowed to use existing treatment options within the health care system (treatment as usual) and offered later access. The trial was preregistered (German Clinical Trials Register: DRKS00034351) [129] and adhered to Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online TeleHealth checklist (Multimedia Appendix 2) [130], and National Institutes of Health best practice guidelines for mixed methods research in the health sciences [131].

Participants

Participants had to meet the following criteria to be eligible: (1) sufficient understanding of the German language, (2) being at least 18 years of age, (3) a physician-suspected or confirmed diagnosis of endometriosis or adenomyosis, (4) clinically relevant sexual distress (Female Sexual Distress Scale-Desire/Arousal/Orgasm [FSDS-DAO] >18), and (5) owning a smartphone (iOS or Android). The exclusion criteria were (1) current severe depression (Beck Depression Inventory-II ≥29) [92], (2) current severe anxiety (Generalized Anxiety Disorder 7-item ≥15) [93], (3) suicidal tendencies in the last 5 years, (4) current symptoms of posttraumatic stress disorder (PTSD), (5) substance dependence in the last 2 years, (6) current psychosis or dissociative symptoms, and (7) current pregnancy. If the online screening responses indicated the presence of potential PTSD or substance dependence, the participants were invited via email to take part in an additional telephone screening interview. The International Trauma Questionnaire [132] was used to assess PTSD, and the relevant section of the Structured Clinical Interview for the *DSM-IV* (*Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*) [133] was used to assess substance use.

Intervention

The Odeya intervention was a self-guided smartphone app intervention developed to address sexual distress in women with endometriosis, for both single and partnered users.

It was developed at the Institute of Sexology and Sexual Medicine, Charité–Universitätsmedizin Berlin, within the Berlin Institute of Health Digital Health Accelerator program, in collaboration with Hybrid Heroes GmbH, using a patient-centered, iterative process (E Kosman, MSc, et al, unpublished data, 2026). The intervention comprises 8 self-guided modules and a symptom-tracking tool, intended for completion over 12 weeks with a 4-week buffer. Modules unlock weekly, each taking ~60 minutes plus 15-60 minutes of exercises. Users can pause and resume at any time. See Table 1 and Figure 1 for a list of module topics and examples of user interface screens. Delivery was multimodal (text, video, audio, graphics, or interactive tasks) with exercises such as pelvic floor training, guided body-based activities, mindfulness, and sensate focus for solo or partnered practice. Participants in relationships were encouraged to involve their partners, while equivalent solo options were provided. The symptom-tracking tool monitors 4 domains (body, mind, social, and sexuality) using 1-10 ratings for pain, stress, self-care, and sexual satisfaction, plus standardized and personalized symptoms. See the Methods section in Multimedia Appendix 1 for further details on development, personas, technical notes, and symptom tracking.

XSL•FO
RenderX

Table 1. Overview of the Odeya intervention modules.

Module	Title	Purpose and key topics	Relational focus
1	You are not alone	Introduces sexual dysfunction and its interplay with endometriosis; fosters health literacy and support.	Normalizes relational strain within the biopsychosocial framework of sexual distress.
2	My body	Builds body awareness and sexual anatomy knowledge; covers stimulation techniques for solo and partnered activity.	Encourages exploration of bodily responses during solo and partnered touch.
3	My pain is real	Explains pain mechanisms and central sensitization; provides strategies for sexual pain management.	Promotes communication about pain, joint coping, and coregulation during intimacy.
4	My sexual self	Explores personal needs, preferences, boundaries, and the role of fantasies in sexual agency.	Encourages sharing preferences and boundaries in communication with partners.
5	My emotional network	Addresses dysfunctional beliefs using the Fear-Avoidance Model and cognitive restructuring.	Reflects on how emotional and cognitive patterns affect relationship dynamics and intimacy.
6	My sexual communication	Enhances communication skills and intimacy through couples' exercises (eg, sensate focus).	The full module has a relationship focus. Emphasizes sexuality as relational communication and includes guided communication strategies and sensate focus exercises for couples.
7	My sexual response	Explains the sexual response cycle in a partnered context, contextual influences, and effects of stress on arousal.	The full module has a relationship focus. Addresses dyadic factors influencing arousal, responsiveness, and shared satisfaction.
8	My resources	Summarizes learnings; develops personalized routines and booster strategies for sustained sexual health.	Encourages maintaining intimacy and open communication with partners.

Figure 1. Interface of the Odeya intervention. Five smartphone screens are shown from left to right: (A) symptom-tracking overview displaying symptom levels (upper section) and present symptoms for a single entry (lower section); (B) overview of intervention modules including modules 1 and 2; (C) example of module content with graphics and an audio exercise; (D) example of psychoeducational content presented as text; and (E) example of psychoeducational content presented as video.

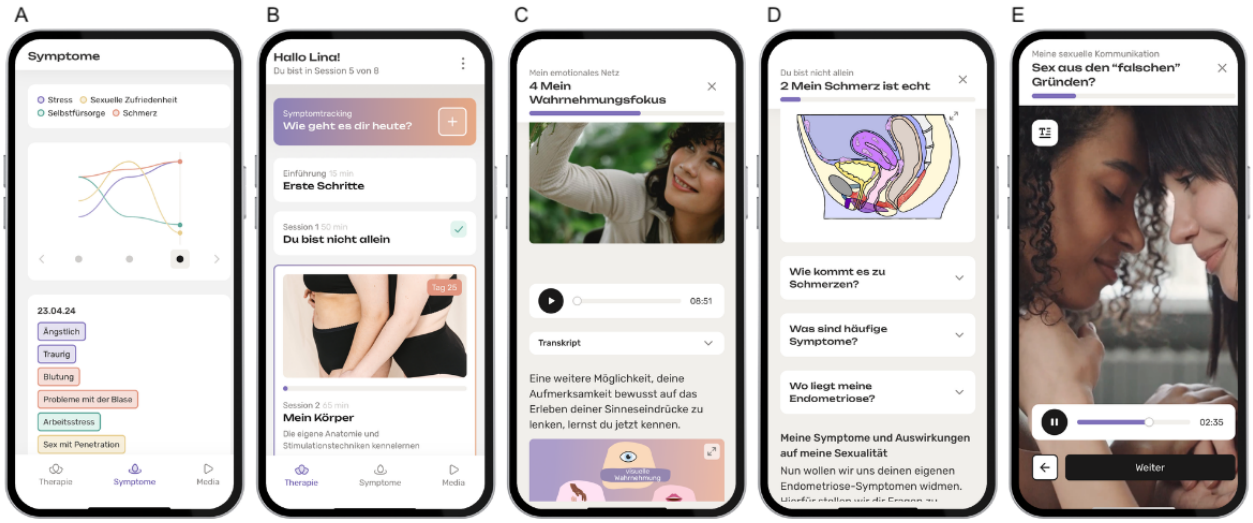

Procedure

Recruitment took place from May to October 2024. Participants were recruited online through study announcements posted by endometriosis associations on their websites, Instagram accounts, and in endometriosis-related Facebook groups in Germany, Austria, and Switzerland. In-person recruitment took place via the Department of Gynecology at Charité–Universitätsmedizin Berlin and through a flyer campaign in outpatient gynecological practices across Germany (see Methods section in Multimedia Appendix 1 for selection procedures). Interested individuals emailed the study team, provided medical documentation of suspected or confirmed endometriosis or adenomyosis, and received study information. A pseudonym and masked email address were created for each participant [134] using the tool AnonAddy to protect participant

XSL•FO
RenderX

identity. For screening and follow-up, they received a link to the data management platform REDCap (Research Electronic Data Capture; Vanderbilt University) [135,136] via email hosted at Charité–Universitätsmedizin Berlin. Participants were invited to a telephone interview with a clinical psychologist if further clarification was required. After randomization, the IG received access to the Odeya app. Qualitative interviews were offered to three groups: (1) IG dropouts, (2) IG completers, and (3) CG completers.

Randomization

Balanced block randomization with 4 blocks was generated in R software (version 4.5.1; R Foundation for Statistical Computing) and implemented in REDCap (version 15.5.29) by the research team to ensure a 1:1 allocation ratio between the IG and CG. Randomization was stratified by relationship status (single vs in a relationship) and age (<30 vs ≥30 years). Participants were informed of their allocation.

Quantitative Measures

This study assessed (1) adherence through module completion, app use duration, dropout status (IG: >4 weeks inactivity with module-locked assessments; CG: nonresponse to scheduled questionnaires), and symptom tracking activity; (2) acceptability, measured with the Client Satisfaction Questionnaire-Internet (CSQ-I), the German mHealth App Usability Questionnaire (G-MAUQ), and a single Visual Analog Scale for Client Satisfaction item on overall satisfaction; (3) safety, evaluated with the Inventory for the balanced assessment of Negative Effects of Psychotherapy-Online Intervention (INEP-ON) and self-reported changes in health status; and (4) sexual and health-related outcomes, including the FSDS-DAO, Female Sexual Function Index-German version (FSFI-d), Fear of Sexuality Questionnaire (FSQ), Vaginal Penetration Cognition Questionnaire (VPCQ), Central Sensitization Inventory-German version (CSI-GE), Partnership Questionnaire (PFB), and the Patient-Reported Outcome Measurement Information System-29-Item Profile (PROMIS-29). Detailed descriptions of assessment time points, selected measures, and additional questionnaires implemented in the broader study framework are provided in Multimedia Appendix 1 (Methods section and Table S1 in Multimedia Appendix 1).

Statistical Analysis

Descriptive statistics (means, SDs, and medians with IQR) were computed for all sexual and health-related continuous variables at baseline (T0), T1, T2, and 6-month follow-up (T3). Change scores (Δ=value$_t$ – value$_0$) were analyzed with linear models for continuous outcomes. Models included group, time, and their interaction, adjusted for the respective baseline value. Adjusted mean changes and 95% CIs were obtained via the emmeans package in R software (version 4.5.1) and visualized. Adjusted between-group differences (IG-CG) are presented as mean differences with 95% CIs; *P* values were not shown, given the exploratory design. Standardized effect sizes (adjusted Cohen d; small: ≥0.2; medium: ≥0.5; large: ≥0.8) [137] were computed for the comparison of change to baseline means. For descriptive reporting, pooled SDs of the individual change scores per time point were added. As all questionnaires used mandatory fields, no item-level missing data occurred. Missingness was limited to uncompleted measurement time points and handled by available-case analysis without imputation. A predefined target sample size of n=60 was based on feasibility considerations within the mixed methods pilot design.

Qualitative Interviews

Overview

The qualitative substudy included 16 participants sampled from IG dropouts (n=11; mean age 29.36, SD 3.91 years), IG completers (n=3; mean age 37.5, SD 13.44 years), and CG completers (n=2; mean 33.67, SD 5.69 years; see Table S2 in Multimedia Appendix 1 for participant characteristics). While all IG participants were invited, CG completers were recruited stepwise using criterion sampling to ensure sociodemographic diversity (age and relationship status) due to feasibility constraints. Given the small, uneven subgroup sizes, the qualitative data were used to capture a breadth of perspectives and to identify barriers and facilitators relevant to feasibility and implementation. Interviews and analysis proceeded iteratively; recruitment ended when no additional overarching feasibility-relevant issues were identified in the final interviews (further details in Methods section in Multimedia Appendix 1). Three distinct semistructured interview guides (Multimedia Appendix 3) were developed for each participant group to explore usability, engagement barriers, perceived helpfulness, and contextual factors.

Qualitative Data Analysis

IG dropout data were analyzed using a free-listing approach, a qualitative elicitation method that generates structured and quantifiable data by asking participants to spontaneously enumerate all responses relevant to a prompt. This method captures how individuals naturally conceptualize and prioritize health-related experiences, making it particularly useful for identifying perceived barriers, needs, and salient usability issues from participants' own language [64,94]. Lists were subsequently cleaned, consolidated, and organized into categories through iterative coding and consensus discussions [95]. Thematic salience was derived by examining the frequency, following established guidelines [96]. Interviews with IG and CG completers were transcribed using *f4transkript* [97] and analyzed with a qualitative content analysis following Schreier's toolbox model [98]. Coding was conducted inductively: an initial coding frame was developed from the data through pilot coding, iteratively refined as further transcripts were coded, and subsequently applied to the full dataset in MAXQDA (version 24.3; VERBI Software GmbH). Analytical rigor was supported through memo-writing and regular peer debriefing within the research team. Further details are provided in Multimedia Appendix 1 (Methods).

Triangulation of Qualitative and Quantitative Data

In line with the expansion-type mixed methods design, quantitative and qualitative data were analyzed separately by researchers with complementary methodological expertise. Integration occurred at the interpretation stage through triangulation, whereby findings from both data strands were systematically compared and synthesized in relation to study

XSL•FO
RenderX

end points through iterative discussion within the research team. Integrated findings are presented in a joint display table.

Ethical Considerations

The human participant study was approved by the Ethics Committee of Charité–Universitätsmedizin Berlin (EA4/217/23). Participants received no compensation and could contact the research team for technical support during the study period. Data were collected and managed in REDCap hosted at Charité. Written informed consent was obtained via postal mail.

Results

Characteristics of Study Population

Between July and November 2024, 187 individuals were screened. Of those, 132 completed the screening process, and 72 were excluded due to eligibility screening (Figure 2). After randomization, 29 women were assigned to the IG, and 31 were assigned to the CG. The mean age of the 60 included participants was 31.12 (SD 6.67) years, with ages ranging from 21 to 59 years (Table 2). Most participants (51/60, 85%) were in relationships and reported high relationship satisfaction (median 8.0, IQR 7.0-9.0). The overall level of education was high, with 81.7% (49/60) having 12 or more years of education. Only 1.7% (1/60) had previously accessed sex and couples therapy. However, half of the participants reported previous experience with psychotherapy. For information on motivation, participation expectations, and FSDS-DAO scores before and after randomization, see the Methods section in Multimedia Appendix 1.

Figure 2. Flowchart of participants. CG: control group; FSDS-DAO: Female Sexual Distress Scale-Desire/Arousal/Orgasm; IG: intervention group.

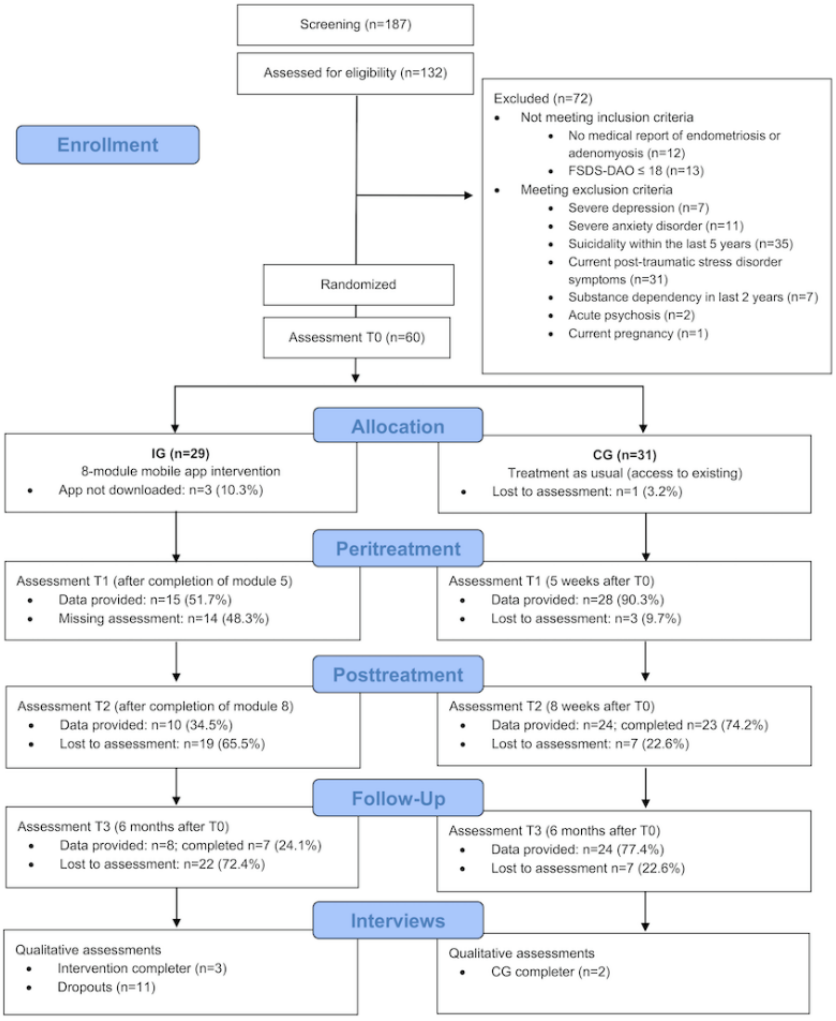

XSL•FO
RenderX

Table 2. Characteristics of the study population at baseline (T0) stratified by intervention group (IG) and control group (CG). Percentages refer to column totals.

Characteristic	Total (N=60)	IG (n=29)	CG (n=31)
Age (years), mean (SD)	31.12 (6.67)	31.1 (5.29)	31.13 (7.83)
In a relationship, n (%)	51 (85)	26 (89.7)	25 (80.6)
Relationship duration[a], mean (SD)	75.80 (67.64)	69.52 (56.86)	82.33 (78.01)
Relationship satisfaction (0-10)[b], median (IQR)	8.0 (7.0-9.0)	8.0 (7.0-9.0)	8.0 (7.0-9.0)
Education ≥12 years, n (%)	49 (81.7)	23 (79.3)	26 (83.9)
Urban residence, n (%)	33 (55)	16 (55.2)	17 (54.8)
Heterosexual, n (%)	51 (85)	23 (79.3)	28 (90.3)
Partnered intimacy[c,d], n (%)	47 (78.3)	24 (82.8)	23 (74.2)
Masturbation[c], n (%)	20 (33.3)	7 (24.1)	13 (41.9)
Religious, n (%)	23 (38.3)	10 (34.5)	13 (41.9)
Histologically confirmed endometriosis, n (%)	52 (86.7)	26 (89.7)	26 (83.9)
Previous operation, n (%)	53 (88.3)	26 (89.7)	27 (87.1)
Hormonal medication, n (%)			
Progestin-only contraceptive pill	17 (28.3)	10 (34.5)	7 (22.6)
Combined oral contraceptive pill	6 (10)	2 (6.9)	4 (12.9)
Sex therapy or couples therapy, n (%)	1 (1.7)	0 (0)	1 (3.2)
Psychotherapy, n (%)	31 (51.7)	17 (58.6)	14 (45.2)
Other diagnoses, n (%)			
PCOS[e]	0 (0)	0 (0)	0 (0)
PMS[f]	13 (21.7)	8 (27.6)	5 (16.1)
Uterus myomatosus	3 (5)	3 (10.3)	0 (0)
Uterus prolapse	1 (1.7)	0 (0)	1 (3.2)
Incontinence	3 (5)	1 (3.4)	2 (6.5)
Infertility	1 (1.7)	1 (3.4)	0 (0)
Vulvodynia	1 (1.7)	0 (0)	1 (3.2)
Lichen sclerosus	0 (0)	0 (0)	0 (0)
Cancer	2 (3.3)	2 (6.9)	0 (0)
Lifestyle, n (%)			
Physical activity	41 (68.3)	22 (75.9)	19 (61.3)
Healthy diet	53 (88.3)	25 (86.2)	28 (90.3)
Smoking	5 (8.3)	2 (6.9)	3 (9.7)
Alcohol consumption	8 (13.3)	3 (10.3)	5 (16.1)
BDI-II[g] (0-63), mean (SD)	13.3 (7.07)	13.45 (6.99)	13.16 (7.26)
GAD-7[h] (0-21), mean (SD)	5.63 (3.4)	6.55 (3.72)	4.77 (2.88)
SSP-F[i], n (%)	54 (90)	25 (86.2)	29 (93.5)
BSI GSI[j], mean (SD)	54.07 (11.88)	54.86 (12.74)	53.32 (11.18)
BSI PSDI[k], mean (SD)	53.53 (10.14)	54.52 (11.33)	52.61 (8.98)
BSI PST[l], mean (SD)	55.05 (12.73)	55.45 (12.96)	54.68 (12.72)
CTQ[m]: Sexual Abuse[n], n (%)	7 (11.7)	3 (10.3)	4 (12.9)

XSL•FO
RenderX

Characteristic	Total (N=60)	IG (n=29)	CG (n=31)
CTQ: Any Trauma[o], n (%)	22 (36.7)	9 (31)	13 (41.9)

[a]In months.

[b]Higher values indicate better outcomes.

[c]More than once per week.

[d]Partnered intimacy referred to activities such as cuddling and kissing.

[e]PCOS: polycystic ovary syndrome.

[f]PMS: premenstrual syndrome.

[g]BDI-II: Beck-Depression-Inventory-II.

[h]GAD-7: Generalized Anxiety Disorder Scale-7.

[i]SSP-F: Screening for Sexual Problems.

[j]BSI GSI: Brief Symptom Inventory Global Severity Index.

[k]BSI PSDI: Brief Symptom Inventory Positive Symptom Distress Index.

[l]BSI PST: Brief Symptom Inventory Positive Symptom Total.

[m]CTQ: Childhood Trauma Questionnaire.

[n]Sexual trauma was defined using the Childhood Trauma Questionnaire Sexual Abuse Subscale, with a cutoff score of 8 [138].

[o]Any trauma was defined as meeting the Childhood Trauma Questionnaire cutoff for at least moderate severity [138].

Main Findings

Overview

The following section reports adherence and user behavior, including module completion, dropout rates, time spent in the app, reasons for discontinuation, and use of the symptom tracker, as well as acceptance and user satisfaction assessed with validated questionnaires. Qualitative facilitators and barriers to adherence and acceptance follow in subsequent sections. Safety outcomes comprised balanced effects, self-reported health changes, and stressful life events.

Adherence and User Behavior

Dropout Rates, Time, and Reasons

Participants in the IG completed a median of 6 (IQR 2-8) modules (mean 4.9, SD 2.97; 4.9/8, 61.2% of the total content). All 8 modules were completed by 34.5% (10/29) of the IG, with a dropout rate of 65.5% (19/29), which is higher than in the CG (7/31, 22.6%). Mean app usage duration, defined as the time from first log-in until either completion of module 8 or the last recorded in-app activity (therapeutic content or symptom tracking), was 15 weeks (range 0-30) in the IG. Among completers, the mean duration was 18 (range 10-30) weeks, whereas noncompleters used the app for an average of 13 (range 0-30) weeks. IG dropouts were distributed throughout the course of the intervention: before starting the app (n=3), after module 1 (n=3), module 2 (n=2), module 3 (n=1), module 4 (n=4), module 5 (n=1), module 6 (n=4), and module 7 (n=1). Baseline characteristics (Table S3) and baseline values of outcomes (Table S4) of IG completers and IG dropouts are presented in Multimedia Appendix 1. Health care behaviors showed little change; only isolated therapy initiations occurred (Results section in Multimedia Appendix 1). Reasons for dropout included time constraints, technical difficulties, life changes, and perceived length of app units (Tables S5-S7 in Multimedia Appendix 1).

Adherence Facilitators and Barriers

High initial motivation and the persona-based progress stories supported adherence:

> *You also got to witness the progress of the three women, and that was motivating and encouraging.* [IG1]

Several participants felt acknowledged by the intervention, contrasting it with previous clinical encounters where their concerns had not been taken seriously.

Participants identified unexpectedly high time demands and emotional strain of prolonged self-reflection as key barriers:

> *It kept getting longer and more extensive.... I think it was supposed to be 50 minutes per module..., but I sometimes needed almost three times as long.... It was really exhausting to constantly engage with myself.* [IG1]

Some wished for more interaction with health care professionals or peers and reported difficulties applying insights to daily life and relationships.

> *When it comes to communication in the relationship that was also a very big topic in the app and I've always struggled with that. There were helpful impulses, but I couldn't really implement them at that point.* [IG2]

Technical issues—videos stopping without replay options, restrictive response formats, and static text fields limiting review of longer entries—further affected usability and adherence.

Use of Symptom Tracker: Frequency and Evaluation

Symptom tracking usage varied widely across participants. The average number of tracking entries during intervention was 18.34 (SD 23.77; range 0-99), with completers averaging 23.20 (SD 34.03; range 4-99) and dropouts 15.79 (SD 16.73; range 0-46). The most frequently self-selected tracked symptoms included, for emotional states: feeling relaxed, sad, or tired; for physical symptoms: bowel problems, intake of pain medication

XSL•FO
RenderX

or hormones, and bladder problems; for social aspects: participation in social activities, sports, or work-related stress; and for sexuality: no sexual activity, masturbation, or sexual intercourse. Figure 3 shows the mean tracked levels of sexual satisfaction, self-care, pain, and stress.

Qualitative feedback on tracking reminders was mixed; some participants perceived them as redundant or burdensome, while one reported them as helpful.

Figure 3. Rating of intensity levels of sexual satisfaction, self-care, pain, and stress in the IG, stratified for completers (n=10) and dropouts (n=15), showing median rating averaged across the usage weeks.

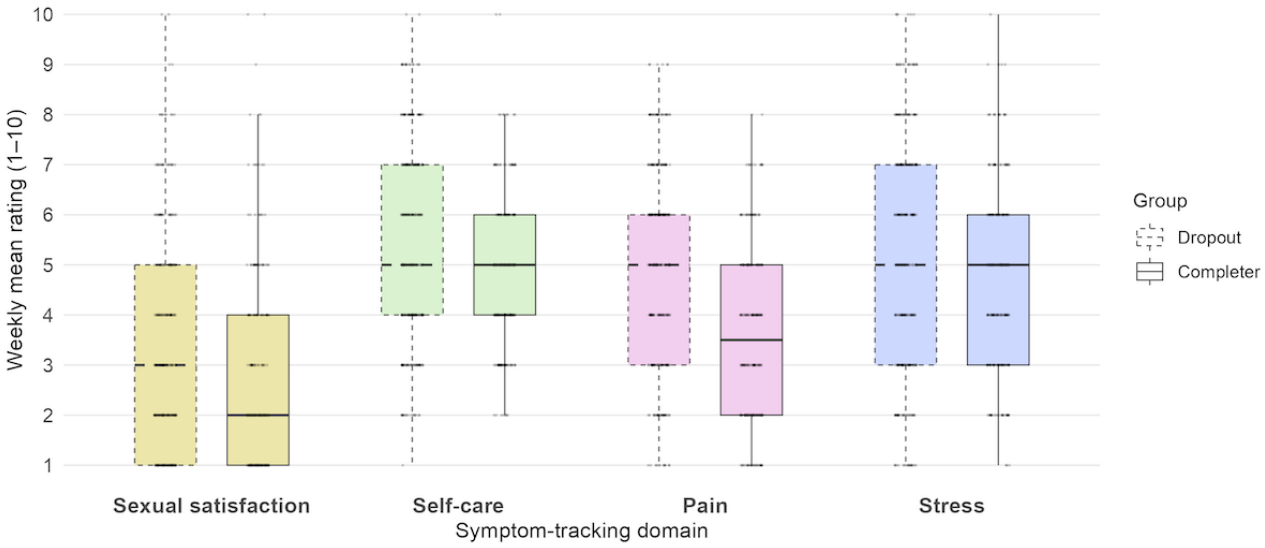

Acceptance

User Satisfaction

At midintervention, the median satisfaction score was 8.00 (IQR 7.0-8.0; mean 7.53, SD 1.36; n=15), and at postintervention (T2), 7.50 (IQR 6.0-8.0; mean 7.00, SD 1.83; n=10) on a 10-point scale (Visual Analog Scale for Client Satisfaction). At T2, CSQ-I ratings had a mean of 26.60 (SD 4.12) of 32. The CSQ-I sum of items assessing general satisfaction (items 1, 2, 3, 5, and 7) averaged 16.30 (SD 2.45), item 4 assessing recommendation 3.50 (SD 0.71), item 6 assessing helpfulness 3.10 (SD 0.88), and item 8 assessing likelihood of reuse 3.70 (SD 0.67). Based on these ratings, 80% (8/10) of participants agreed or strongly agreed with CSQ-I total items, 90% (9/10) reported that they would recommend or reuse the intervention, and 70% (7/10) reported that the intervention was helpful and fulfilled their general satisfaction. Usability, assessed with the G-MAUQ, showed a total score of 5.38 (SD 0.74), with subscales of ease of use (mean 6.46, SD 0.64), interface satisfaction (mean 5.28, SD 0.78), and usefulness (mean 4.46, SD 1.69). For detailed ratings, including median, minimum, and maximum values, see Table S8 in Multimedia Appendix 1.

These findings were closely mirrored in participant interviews, where usability, app design, and comprehensive content were highlighted as facilitating factors.

> *First of all, in terms of the technical aspects and structure, I found the app very clear and easy to use.* [IG1]

Qualitatively, some users expressed discomfort or ambivalence about specific exercises, such as guided masturbation or vulva self-exploration.

Facilitators and Barriers to Acceptance

Multimedia elements and the option to proceed at one's own pace were highly valued. Participants appreciated the intervention's diverse components, including normalization through personas ("not feeling alone"), psychoeducation, practical and communication exercises, structured reflection, partners' involvement, and gradual exposure to sensitive topics, especially the sensate focus exercises.

However, several barriers were identified. Technical issues and content overload inhibited satisfaction for some. These frustrations might explain lower satisfaction scores or disengagement in some cases. Weekly exercises were challenging, leading to self-doubt and frustration, although these feelings were partly mitigated by greater awareness of social influences, as one participant described:

> *Sometimes I felt very frustrated, doubting if it was my fault.... I even felt guilty after reading or doing certain exercises.... However, I was able to balance this by reflecting on what I'd learned from my parents' attitudes toward sexuality...phrases from friends or relationships. These insights gave me "aha" moments, helping me feel less desperate and understand that other factors might be involved.* [IG3]

Safety

Balanced Effects

The most frequently endorsed positive effects in the INEP-ON outcome were the helpfulness of new ways of thinking (T2: 10/10, 100%; T3: 6/7, 85.7%) and moderator support (T2: 9/10, 90%; T3: 4/6, 57.1%; Table S9 in Multimedia Appendix 1). Increased motivation for psychotherapy was reported by 60% (6/10) at T2 and 42.9% (3/7) at T3. Improved overall well-being

XSL•FO
RenderX

was noted by 90% (9/10) postintervention, but only 28.6% (2/7) at follow-up, while 57.1% (4/7) reported deterioration. Negative responses on balanced items were otherwise rare (≤10% at T2 and <30% at T3; see Results section in Multimedia Appendix 1). Items addressing exclusively negative effects (Table S10 in Multimedia Appendix 1) showed longer periods of not feeling well in 50% (5/10) at T2 and 71.4% (5/7) at T3 (see Results section in Multimedia Appendix 1 for details).

Self-Reported Health Changes and Stressful Life Events

A total of 6 women in the IG and 8 in the CG reported events, including new medical diagnoses (eg, cardiac arrhythmia, asthma, adenomyosis, suspected lipedema, and migraine), hospitalizations, bereavement, or relationship breakups. In the IG, only 1 affected participant dropped out; the others completed the study.

Secondary Outcomes

Overview

Secondary outcomes included exploratory analyses of changes in sexual health (FSDS-DAO, FSFI-d, FSQ, VPCQ, CSI-GE), relationship (PFB), and overall health (PROMIS-29). Table S11 in Multimedia Appendix 1 reports mean scores and Table S12 baseline-adjusted changes with effect sizes for all outcomes. Figure 4 illustrates sexual health outcomes based on total scores.

Figure 4. Sexual health outcomes over time as changes from baseline (Δ) for intervention group (IG) and control group (CG) across T0-T3. Points and lines depict estimated marginal means of the change scores; error bars indicate 95% CIs. Above the x-axis, the between-group effect size (d) from the linear model (Δ ~ Group × Time + Baseline) is shown for each time point. Shaded backgrounds denote study phases (on treatment: T0-T2; follow-up: T3). Panels A-D correspond to outcomes. CSI-GE: Central Sensitization Inventory-German version; FSDS-DAO: Female Sexual Distress Scale-Desire/Arousal/Orgasm; FSFI-d: Female Sexual Function Index-German version; PFB: Partnership Questionnaire; T0: baseline; T1: after module 5/5 weeks after baseline; T2: after module 8/8 weeks after baseline; T3: 6-month follow-up after baseline.

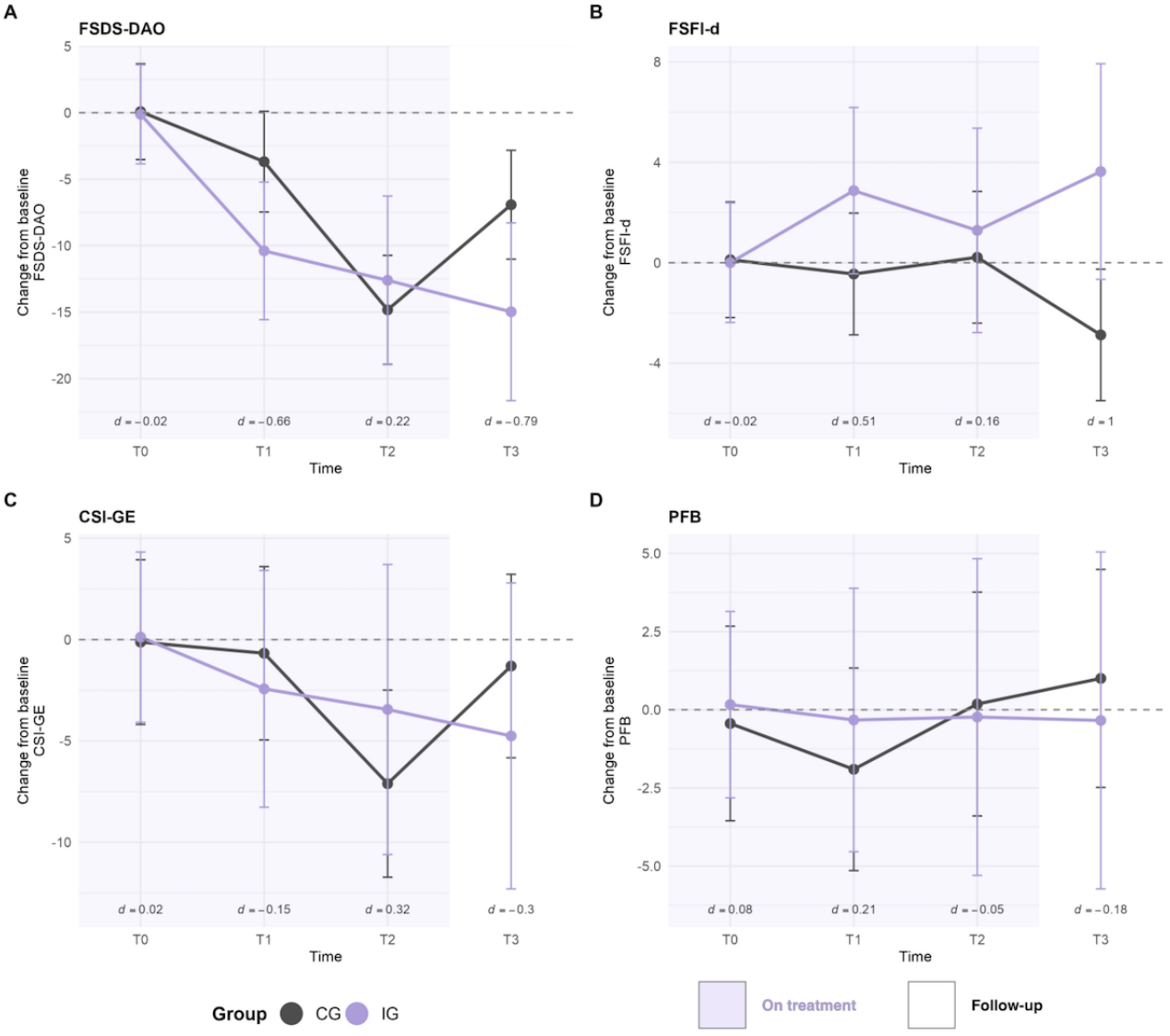

XSL•FO
RenderX

Sexual Health and Relationship-Related Outcomes

The IG showed a stronger early reduction in sexual distress (FSDS-DAO) (T1: IG Δ=–10.39 vs CG Δ=–3.68; between groups Δ=–6.71, 95% CI –13.13 to –0.29; *d*=–0.66) while both groups improved by T2 (Figure 4 and Table S12 in Multimedia Appendix 1). At T3 the IG again showed a stronger reduction (T3: Δ=–8.05, 95% CI –15.89 to –0.22; *d*=–0.79).

Sexual function (FSFI-d) improved continuously in the IG (T1-T3 Δ=1.29-3.63), with the largest between-group difference at T3 (Δ=6.51, 95% CI 1.48-11.55; *d*=1.00).

Fear of coitus (FSQ) showed small between-group effects, with the IG showing greater improvement at T1 (*d*=−0.21) and T3 (*d*=−0.14), but the CG at T2 (*d*=0.17). Fear of noncoital activity decreased continuously in the IG by about 0.6 points at T2 (scale ranging from 5-25), with small to large effects between groups (*d*=−0.14 to −0.84; Table S10 in Multimedia Appendix 1).

Regarding vaginal penetration cognition (VPCQ), results were mixed, with small to moderate improvements in the IG, primarily at T1 and T3 (eg, between-group effects: at T3, control cognitions d=0.48; catastrophic and pain cognitions d=−0.17). The only scale showing continuous improvement favoring the IG was incompatibility cognitions, with between-group effect sizes ranging from d=−0.34 to −0.89.

Central sensitization (CSI-GE) improved in both groups. Small effects favored the IG at T1 (*d*=−0.15) and T3 (*d*=−0.30), while at T2 a small effect favored the CG (*d*=0.32).

Partnership quality (PFB) demonstrated only small changes over time (T1: Δ=1.58; d=0.21; T2: Δ=−0.42; d=−0.05; T3: Δ=−1.34; d=−0.18). Subscale effects were likewise small: disruptive behavior worsened slightly in the IG at T3 (d=0.35), whereas communication improved in the IG at T1 (d=0.26) and T3 (d=0.21), with no change at T2 (d=−0.01).

Qualitative interviews indicated that IG participants frequently reported new positive sexual experiences:

> *Using the app helped me to rediscover how to have positive experiences with sexuality—across the entire spectrum of what sexuality can be.* [IG1]

Several reported that their perception and communication of sexual pain improved, although complete pain relief was rarely achieved:

> *From the communication perspective, I felt that I was often able to say, "Okay, I have pain now," and I could localize it, to give him that feedback.* [IG3]

Further, IG participants described gaining more courage and reduced fear and avoidance of sexual activities, reporting increased confidence in approaching both coital and noncoital intimacy:

> *To find the courage to start again and actually do something. And also, those small steps—that was exactly what I needed. I can imagine many others feel the same, because there's always that fear of penetrative sex.* [IG1]

Qualitative analyses revealed that participants in the IG experienced increased openness and more frequent communication with their partners regarding sexuality and recognized the issue as a shared concern affecting both partners, rather than an individual problem. One participant noted:

> *In the beginning, when I started using the app, I really tried to communicate more with my partner.... It's usually such an uncomfortable topic for me—I've never talked about it openly. But I tried to tell him about the app, showed him parts of it.... I realized, it's not just my issue, it's something that affects the relationship too, and we're both suffering in some way.* [IG3]

Additionally, several participants described a shift toward a more constructive and collaborative dynamic, moving away from blame:

> *We talked about that very openly—that it's not about anyone being at fault or him hurting me or anything like that, but rather that we go through the process together and figure out, together, what feels good and what doesn't.* [IG3]

Although CG participants had no access, qualitative interviews suggested that study participation itself fostered self-reflection prompted by study questionnaires.

> *I actually liked that it forced me to keep reflecting...on what's going on in that part of my life, my sexuality or lack of libido. [It was really] helpful].* [CG1]

CG participants also reported positively impacting life events over the course of the study (eg, hormonal changes and changes in sexual life circumstances) and expressed continued interest in the intervention after the study.

Health-Related Quality of Life

Results for the PROMIS-29 domains were mixed overall. Anxiety, depression, fatigue, and physical function improved in both groups at T2, although between-group comparisons favored the CG at several time points. In contrast, social participation, pain interference, and pain intensity improved across all time points in the IG, with between-group effects favoring the IG at T2 and T3 and showing moderate to large effect sizes.

Qualitatively, participants in the IG attributed enhanced emotional regulation and reduced distress to the intervention, linking these psychological changes to better physical health management:

> *That was one of my goals—to somehow get out of this negative emotional state. And it actually worked. Of course, there are still phases or moments when you think, "Hmm..." But then things come to mind—things you can do, right? And you realize, okay, even if it's not working right now, that's okay too—it's not the end of the world.* [IG1]

Self-reflection and body awareness were recurring qualitative themes reported by both groups, further elucidating why both showed improvement in those domains.

XSL•FO
RenderX

Integration of Quantitative and Qualitative Findings

The integration of quantitative outcomes and qualitative interviews (Table 3) provides a nuanced picture of the intervention's effects on sexual health, relationships, well-being, and user engagement. Quantitative findings contextualized quantitative improvements and highlighted adherence challenges and the complexity of symptom management.

XSL•FO
RenderX

Table 3. Synthesis: integration of quantitative and qualitative results. Intervention group (IG) > control group (CG) indicates between-group differences in favor of the intervention group, whereas IG < CG indicates differences in favor of the control group.

Domain	Quantitative findings	Qualitative themes	Synthesis
Adherence	• IG completion 34.5% (10/29), dropout 65.5% (19/29) vs CG 22.6% (7/31) • Symptom tracking more frequent in completers • Frequent use of individual symptom feature • Outside app: start of psychotherapy, no change in costs	• Facilitator: persona stories • Barriers: emotional strain, time demands • Reminders: helpful or confusing • Dropout: time, technical issues, life changes • Need: more professional/patient interaction	High dropout aligns with time/emotional strain; progress stories may support adherence. Mixed feedback on tracking suggests the need for customizable reminders. Behavior outside app indicates broader health-seeking changes.
Acceptance	• CSQ-I[a] mean 26.6 (SD 4.1) of 32 • G-MAUQ[b] mean 5.4 (SD 0.7) of 7; "Ease of use" mean 6.5 (SD 0.6) of 7 • Mean price €83 (range €0-€300; €1=US $1.17); 30% (3/10) no self-pay	• Facilitators: clarity, multimedia, flexibility, reimbursement • Barriers: technical issues, content overload	High satisfaction matches positive reports; barriers may explain disengagement. Payment variability underscores the need for reimbursement.
Safety	• INEP-ON[c]: learned strategies, moderator support • "Negative events" rare (≤10% in T2, <30% in T3); mainly periods of not feeling well	• Reports of health changes, stressful life events in both groups	Findings suggest overall safety; negative responses linked more to external stressors than intervention.
Sexual health	• FSDS-DAO[d]: improved in both groups, IG > CG at T1, T3 • FSFI-d[e]: IG > CG • FSQ[f] "coital": improved in both, IG vs CG mixed; "noncoital": IG > CG • VPCQ[g] "control," "catastrophic": improved in both groups, IG vs CG mixed; "Incompatibility": IG > CG • PFB[h]: no change at T2; IG vs CG mixed; "Communication": IG > CG at T1, T3 • CSI-GE[i]: improved in both groups, IG > CG at T1; T3, IG < CG at T2	• Facilitators: rediscovery of positive experiences, improved pain communication, openness, collaboration • Barriers: ambivalence toward some exercises • CG: self-reflection via questionnaires and self-applied tools/books	Quantitative gains align with reports of reduced distress and better sex-related communication. Although some changes were not strongly reflected in quantitative scores, participants described meaningful qualitative improvements. CG improvements may be attributed to study participation and self-reflection.
Overall Health	• PROMIS-29[j]: "anxiety," "depression," "fatigue," "physical function" improved at T2 in both groups, mostly IG < CG; "Social participation": IG > CG at T3; "Pain interference", "pain intensity": IG > CG at T2, T3.	• IG: improved emotional regulation, distress reduction • Both: more self-reflection, body awareness	Both groups, especially the CG, showed broad improvements in psychological well-being and physical functioning, while the IG exhibited reductions in pain interference and intensity. Qualitative reports of greater emotional regulation, open partner communication, and body awareness in the IG support these findings, indicating that self-regulatory and relational processes may reduce pain-related avoidance and attentional capture by pain.

[a]CSQ-I: Client Satisfaction Questionnaire-Internet Version.

[b]G-MAUQ: German mHealth App Usability Questionnaire.

[c]INEP-ON: Inventory for the balanced assessment of Negative Effects of Psychotherapy.

[d]FSDS-DAO: Female Sexual Distress Scale-Desire/Arousal/Orgasm.

[e]FSFI-d: Female Sexual Function Index-German version.

[f]FSQ: Fear of Sexuality Questionnaire.

[g]VPCQ: Vaginal Penetration Cognition Questionnaire.

[h]PFB: Partnership Questionnaire.

[i]CSI-GE: Central Sensitization Inventory-German Version.

[j]PROMIS-29: Patient-Reported Outcome Measurement Information System-29-Item Profile.

XSL•FO
RenderX

Discussion

Principal Findings

This pilot implementation study examined the adherence, acceptability, safety, and exploratory effects of the self-guided Odeya app for women with sexual distress and diagnosed or suspected endometriosis. Adherence was moderate, with most completers working through several modules over an average duration of 18 weeks, while dropout was high. Overall satisfaction with the app was strong, and no negative life and health changes were attributed to the intervention. Quantitative outcomes showed reductions in sexual distress in both groups, with some advantages for the IG in sexual function, penetration-related fears, and partner communication. Broader health and mental health outcomes showed mixed but generally positive changes across groups. Qualitative feedback supported these trends, describing more positive sexual experiences, improved partner communication, greater pain awareness, and reduced fear of penetration.

Adherence

Overview

The intervention dropout rates in our study were notably high, a phenomenon typically reported in digital interventions for chronic conditions. A meta-analysis found a pooled dropout rate of 43% across app-based interventions for chronic diseases [76]. In the HelloBetter Vaginismus DiGA effectiveness trial, Zarski et al [50] reported dropout of 22% postintervention among women with dyspareunia, with participants completing on average 6 of 8 modules. In an early study, the Endo App retained 64.4% (29/45) of endometriosis patients in week 12 [61]. The German DiGA registry (BfArM) reports low attrition in IGs for several DiGAs, including the Endo App (3.75% dropout; NCT04883073) [60], Kranus Edera for erectile dysfunction (4.1%) [59], and Kranus Lutera for incontinence (5.4%) [62]. In comparison with these trials, dropout in our study was higher in the IG than in the CG [50,59,61,62]. Several factors may explain the comparatively high dropout in our sample.

Health Status

First, our study population comprised individuals with endometriosis and sexual distress, a particularly complex clinical profile that may require more intensive support than fully self-guided interventions can provide [69]. Unlike trials with stricter exclusion criteria (eg, excluding chronic pain or moderate depression [49,50,58], recent medication or surgery changes [60], chronic infections [62], or postsurgical/cardiovascular risk) [59], our study was deliberately inclusive, excluding only severe depressive or anxiety symptoms. This approach enhanced external validity but may have increased attrition among participants facing multiple health burdens. At baseline, dropouts reported lower quality of life and greater symptom burden (anxiety, depression, stress, fatigue, and pain) yet less sexual distress and better sexual function than completers. Visual inspection of the symptom-tracking data suggested slightly higher median values among dropouts for sexual satisfaction and pain compared to IG completers. This health status pattern may also explain the continuous dropout observed in our study, as disengagement likely reflects participants' health burden rather than low initial motivation or lack of interest in the intervention—contrasting with the early-stage dropout typically reported in digital mental health apps [139,140]. Some studies found no link between depression or pain and attrition [79,81]. However, a systematic review and a meta-analysis identified baseline depression and comorbid anxiety as predictors of dropout [77,80]. Further, most participants were in relationships. Being partnered has been identified as a risk factor for sexual dysfunction with distress [12], whereas being single has already been associated with higher dropout [81].

App Engagement

Second, app engagement and adherence may operate in both directions. While active tracking could facilitate greater adherence, it is also plausible that individuals with better health or mental well-being were more likely to engage with the app. Dropouts tracked symptoms less frequently, indicating lower initial engagement with self-monitoring features. Early engagement predicts adherence in digital health programs [78,81,82] and may mark a critical window for retention strategies. Evidence on the link between tracking and adherence in digital health interventions is mixed: some studies report benefits of consistent symptom tracking, while others show only transient effects [68,86,89]. In our study, reduced tracking and higher dropout among participants with elevated anxiety and depression suggest that mental health symptoms hinder sustained engagement. This aligns with meta-analytic evidence showing that engagement modestly predicts improved outcomes, with specific indicators such as module completion being most predictive [83].

Program Design

Third, compensation structures and intervention duration may have influenced adherence. Studies with better adherence rates often provided participant compensation or had shorter intervention periods [84], factors that can significantly impact engagement and completion rates. Although a 12-week duration is clinically appropriate and common [50,58,59,61,62], it may be demanding for participants with chronic pain and its psychological burden, as noted in prior studies [141]; furthermore, modules were reported to exceed the displayed module duration of 60 minutes.

Acceptance and User Experience

Overview

Quantitative results showed high user satisfaction, aligning with findings in women with genito-pelvic pain/penetration disorder (CSQ-I mean 28.03, SD 3.96; 67/72, 93% satisfaction) [50] and supported by qualitative reports of positive experiences. Similar satisfaction levels have been reported in other digital interventions across different patient populations using the CSQ-I [65,72]. Usability (G-MAUQ) ratings were particularly high for ease of use, likely reflecting the patient-centered development process (E Kosman, MSc, et al, unpublished data, 2026) and the app's reduced functions, with scores comparable to other digital interventions [66,85]. Perceived usefulness was

XSL•FO
RenderX

slightly lower in our study, which may have contributed to attrition.

Content Overload and Usability as Barriers

Content overload, reflected in the high time demands of the modules and the emotional strain of sustained self-reflection [142], highlighted the challenge of balancing comprehensive content with practical feasibility for participants. Similar findings show that time-intensive digital interventions hinder engagement in this population [74]. Moreover, even seemingly minor usability issues can significantly affect user experience and sustained engagement [73,75,87].

Guidance

Interview participants frequently requested more clinician and peer contact—consistent with evidence that human support drives engagement [88]—and reported difficulties applying content and feeling insufficiently guided. This underscores the potential value of integrating human support into digital tools for complex chronic conditions. Interventions including personal guidance achieved better adherence [56,76] compared with our fully self-directed approach. Blended care models may enhance both engagement and acceptance by providing structured support alongside digital self-management [143].

Willingness to Pay

Despite high user satisfaction, willingness to pay for a comparable DiGA was limited. Participants expressed generally low willingness to pay, substantially lower than prices of comparable DiGAs in Germany [57,63], and many felt that such tools should be financed by the health care system rather than by users themselves. Similar patterns have been reported, linking reluctance to pay to expectations of public coverage, skepticism toward digital tools, and limited awareness of benefits [144].

Safety

None of the health and life changes was attributable to the intervention. Comparisons with other digital interventions are limited, as adverse events are rarely assessed or reported in German DiGAs [71]. Our findings accord with a recent meta-analysis indicating that mental health apps rarely cause harm and do not increase adverse events versus controls [70].

Changes in Sexual and Overall Health

Although not powered for clinical end points, the trial provided exploratory indications of improvements, with small to large effects based on between-group mean differences. Our findings are consistent with evidence that both face-to-face and online interventions can reduce sexual distress and improve sexual function and satisfaction [16,41,44], that psychoeducation increases knowledge and reduces performance anxiety [31], and that mindfulness improves desire, arousal, and satisfaction while mitigating sexual distress [45,46]. An online self-help trial for women with dyspareunia reported medium to large improvements in genital pain and penetration-related cognitions and small to medium improvements in sexual function, anxiety, and well-being [50]. A psychoeducational web-based program for cancer survivors improved sexual communication but showed limited impact on relationship outcomes [67]. Consistent with this pattern [50,51,53,54], our trial also did not affect relationship satisfaction. Qualitative data aligned with established mechanisms [16,55]: participants described sensate focus and self-stimulation exercises as helpful for improving desire, and cognitive restructuring for reducing maladaptive sexual beliefs. These processes were reflected quantitatively by reduced incompatibility cognitions in the IG. However, it should be noted that some between-group differences may reflect random error rather than genuine effects.

Consistent with prior research [50], control participants also improved in health (eg, depression, anxiety, and fatigue) and sexual distress. However, at follow-up, reductions in sexual distress persisted strongly only in the IG. Qualitative data suggest that repeated self-assessment may have acted as a catalyst for self-reflection and proactive coping, a Hawthorne-like effect [145]. Additional mechanisms include mere-measurement effects [146], hope and expectancy [147], natural recovery [50,148], and enhanced self-monitoring [149]. Despite these effects, CG participants continued to desire intervention access, indicating ongoing unmet needs.

Implications for Future Digital Interventions

Despite the high dropout rate, satisfaction metrics in the IG indicated that engaged participants found the program helpful. Thus, content appears valuable, whereas delivery and support structures may require optimization. Following qualitative feedback indicating a desire for more guidance and professional contact, we propose key design considerations for future digital interventions (Textbox 1).

XSL•FO
RenderX

Textbox 1. Key design considerations to enhance engagement, adherence, and user experience in future digital sexual health interventions targeting individuals with endometriosis

1. Hybrid approaches: incorporating human support elements—and, where appropriate, AI-driven assistance—such as periodic check-ins with health care providers, trained coaches, or moderated peer communities, may enhance engagement, adherence, and outcomes [150-152].
2. Tailored content delivery: reducing the time burden and emotional intensity of modules while maintaining therapeutic effectiveness could improve completion rates. Flexible options—such as individually selectable modules, exercises, or notifications and supplementary materials (eg, workbooks)—may better address diverse user needs [153,154].
3. Technical optimization: addressing usability issues and ensuring a seamless user experience—potentially through the integration of gamification features—is fundamental to sustaining user engagement [155].
4. Screening for intervention readiness: given the relationship between baseline psychological distress and dropout, screening for intervention readiness and providing additional support for participants at high risk of dropout may improve outcomes [156].
5. Patient involvement: beyond developing the intervention through a patient-centered approach, we recommend involving patients or patient influencers in the go-to-market phase to ensure effective reach and trust [64,157-160].

Integrating the Relational Dimension

Beyond the structural and delivery-related design considerations outlined above, the relational dimension constitutes a further critical aspect of future digital sexual health interventions. Our findings address a recently identified structural limitation in current DiGAs for sexual dysfunctions: the insufficient consideration of the relationship dimension. While existing DiGAs primarily address individual functional parameters and measure success through standardized scores such as the International Index of Erectile Function - 5, sexual dysfunctions are frequently embedded in relational components that require dyadic intervention approaches [91].

Future Research

Future research should test efficacy for core sexual health outcomes, particularly sexual distress, in adequately powered randomized controlled trials and, in larger samples, identify subgroups most likely to benefit. Incorporating participants' recommendations—such as integrating the program into blended-care models with chat support or video consultations [12]—may enhance value and warrant evaluation. Longer, more flexible trials should also examine motivational drivers and dropout trajectories, as a 4-week inactivity threshold may be overly restrictive (eg, for women recovering from endometriosis surgery). Moreover, future iterations could include both optional and mandatory partner components addressing partners directly (eg, psychoeducation modules on endometriosis and its relational impact).

Limitations

This study has several limitations. First, several methodological aspects limit the interpretability of findings. The trial was not powered to detect efficacy or between-group differences, constraining comparative analyses. Additionally, although aspects of preuse acceptability were addressed during development, they were not systematically assessed in this study. Without a dedicated prepilot acceptability study, some usability issues (eg, module length and navigation) may not have been identified beforehand. Finally, combining assessments of acceptability and preliminary effectiveness within a single early-phase trial may have constrained sampling strategies and methodological specificity. Second, dropout rates were high—particularly in the IG—resulting in substantially reduced sample sizes for both quantitative and qualitative components. This further restricts within-group analyses and impedes meaningful comparisons between the IG and CG. Participants who discontinued showed higher baseline symptom severity (eg, anxiety and depression), potentially biasing findings toward more favorable outcomes. Third, satisfaction outcomes may be influenced by recruitment and measurement constraints. Because the recruitment strategy relied primarily on online channels, the sample may overrepresent women with higher digital literacy or a stronger preference for online support formats. Satisfaction was also only assessed among women who completed the intervention due to predefined measurement timing, likely inflating acceptability ratings. Fourth, sample characteristics and condition-specific focus limit generalizability. The sample covered a broad age range (up to 59 years), representing heterogeneous life stages (eg, perimenopause and menopause) with potentially distinct biopsychosocial profiles. Participants were also highly educated and predominantly partnered, which is not representative of the wider population. Moreover, although the intervention was developed through a patient-centered approach with women affected by different gynecological conditions (E Kosman, MSc, et al, unpublished data, 2026), it was only tested in women with endometriosis or adenomyosis, further limiting generalizability. Fifth, while the intervention included optional partner-focused components, systematic involvement of partners was limited, and dyadic processes were not comprehensively addressed. This represents a conceptual limitation, given that sexual health in the context of chronic conditions such as endometriosis is inherently relational. Finally, a coding error prevented proper attribution of INEP-ON items, obscuring whether reported positive or negative changes were related to life circumstances or the intervention itself.

Conclusions

The digital intervention Odeya appeared acceptable and safe for women who engaged with the app, with initial indications of improvements in sexual health–related outcomes among completers. Sustained engagement in fully self-guided formats, however, appears to vary by baseline psychosocial status, with individuals reporting better psychosocial health showing higher adherence. Adequately powered trials should establish efficacy, identify moderators of benefit, and determine the support intensity needed to maximize impact. Self-guided digital interventions may present one accessible and scalable

XSL•FO
RenderX

component within stepped-care models for sexual health, although some users are likely to require additional guidance and personalization.

Acknowledgments

We sincerely thank all patients who participated in the study and the qualitative development process. We also acknowledge the support of the Digital Health Accelerator Program of the Berlin Institute of Health at Charité–Universitätsmedizin Berlin, which enabled the collaboration with Hybrid Heroes for the development of the app-based intervention.

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: Literature search and systematization; Code generation; Code optimization;Proofreading and editing; Summarizing text; Adapting and adjusting emotional tone; Translation; Reformatting; Quality assessment; Identification of limitations; Recommendations; and Publication support.

The GenAI tools used were ChatGPT (GPT-5.1), Claude (Claude Sonnet 4.5), and Perplexity AI. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was supported by the Digital Health Accelerator Program of the Berlin Institute of Health, which also enabled the development of the app-based intervention in collaboration with Hybrid Heroes. The funder had no role in the study design, data collection, analysis, interpretation, or the decision to submit the manuscript for publication.

Data Availability

The datasets analyzed during this study are not publicly available due to restrictions imposed by the ethics approval.

Authors' Contributions

Conceptualization: LH, SMK, EK
Data curation: SMK, TT-D, LWS, EM
Formal analysis: SMK, EK, TT-D
Funding acquisition: LH, SMK
Methodology: LH, SMK, EK
Project administration: SMK, LH
Resources: KMB, LH
Supervision: LH, KMB, MMK, TP
Validation: LH
Visualization: SMK, EK
Writing – original draft: SMK, EK, LH, LW-S
Writing – review & editing: SMK, EK, LH, KMB, MMK, TP, EM, LW-S

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary methods, results, and tables.
[PDF File (Adobe PDF File), 850 KB - jmir_v28i1e86042_app1.pdf]

Multimedia Appendix 2
CONSORT‐eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1055 KB - jmir_v28i1e86042_app2.pdf]

Multimedia Appendix 3
Interview protocols.
[PDF File (Adobe PDF File), 176 KB - jmir_v28i1e86042_app3.pdf]

References

1. Beier KM, Loewit K. Praxisleitfaden Sexualmedizin. Heidelberg: Springer Berlin; 2011.

XSL•FO
RenderX

2. Lee DM, Vanhoutte B, Nazroo J, Pendleton N. Sexual health and positive subjective well-being in partnered older men and women. J Gerontol B Psychol Sci Soc Sci 2016;71(4):698-710 [FREE Full text] [doi: 10.1093/geronb/gbw018] [Medline: 26993519]
3. Mitchell KR, Lewis R, O'Sullivan LF, Fortenberry JD. What is sexual wellbeing and why does it matter for public health? Lancet Public Health 2021;6(8):e608-e613 [FREE Full text] [doi: 10.1016/S2468-2667(21)00099-2] [Medline: 34166629]
4. International Classification of Diseases (ICD-11). Geneva: World Health Organization; 2019.
5. Briken P, Matthiesen S, Pietras L, Wiessner C, Klein V, Reed GM, et al. Estimating the prevalence of sexual dysfunction using the new ICD-11 guidelines. Dtsch Arztebl Int 2020;117(39):653-658 [FREE Full text] [doi: 10.3238/arztebl.2020.0653] [Medline: 33357346]
6. McCabe MP, Sharlip ID, Lewis R, Atalla E, Balon R, Fisher AD, et al. Risk factors for sexual dysfunction among women and men: a consensus statement from the fourth international consultation on sexual medicine 2015. J Sex Med 2016;13(2):153-167. [doi: 10.1016/j.jsxm.2015.12.015] [Medline: 26953830]
7. McCool-Myers M, Theurich M, Zuelke A, Knuettel H, Apfelbacher C. Predictors of female sexual dysfunction: a systematic review and qualitative analysis through gender inequality paradigms. BMC Womens Health 2018;18(1):108 [FREE Full text] [doi: 10.1186/s12905-018-0602-4] [Medline: 29929499]
8. Mitchell KR, Mercer CH, Ploubidis GB, Jones KG, Datta J, Field N, et al. Sexual function in Britain: findings from the third national survey of sexual attitudes and lifestyles (Natsal-3). Lancet 2013;382(9907):1817-1829. [doi: 10.1016/s0140-6736(13)62366-1]
9. Grosse-Rueschkamp MS, Kronthaler SM, Legeland LT, Schulze D, Beier KM, Hatzler L. What knowledge matters for women's sexual wellbeing? Validating a questionnaire measuring sexual knowledge using public and patient involvement. J Sex Med 2025;22(11):2104-2113. [doi: 10.1093/jsxmed/qdaf209] [Medline: 40974279]
10. Perelman MA. The sexual tipping point: a mind/body model for sexual medicine. J Sex Med 2009;6(3):629-632. [doi: 10.1111/j.1743-6109.2008.01177.x] [Medline: 19210711]
11. González-Mesa E, Moya-Bejarano D, Butrón-Hinojo CA, Marín-Sánchez P, Blasco-Alonso M, Jimenez-López JS, et al. Correlates of sexual function in a sample of Spanish women with endometriosis. J Clin Med 2021;10(21):4957 [FREE Full text] [doi: 10.3390/jcm10214957] [Medline: 34768476]
12. Kronthaler SM, Tissen-Diabaté T, Karsten MM, Blohmer J, Beier KM, Hatzler L. Assessment of mental and chronic health conditions as determinants of health care needs and digital innovations for women with sexual dysfunction: cross-sectional population-based survey study in Germany. J Particip Med 2025;17:e71301 [FREE Full text] [doi: 10.2196/71301] [Medline: 40865091]
13. Brenk-Franz K, Häuser W, Brähler E, Schneider N, Hoy M, Strauß B. Sexual dysfunctions and health-related impairment in patients with chronic disease. Dtsch Arztebl Int 2024;121(2):66-67 [FREE Full text] [doi: 10.3238/arztebl.m2023.0218] [Medline: 38427942]
14. Sobel T, David P. Impact of chronic medical disease on sexual function and other conditions. Obstet Gynecol Clin North Am 2024;51(2):323-340. [doi: 10.1016/j.ogc.2024.02.006] [Medline: 38777487]
15. Valeiro C, Matos C, Scholl J, van Hunsel F. Drug-induced sexual dysfunction: an analysis of reports to a national pharmacovigilance database. Drug Saf 2022;45(6):639-650. [doi: 10.1007/s40264-022-01174-3] [Medline: 35386045]
16. Zarski AC, Velten J, Knauer J, Berking M, Ebert DD. Internet- and mobile-based psychological interventions for sexual dysfunctions: a systematic review and meta-analysis. NPJ Digit Med 2022;5(1):139 [FREE Full text] [doi: 10.1038/s41746-022-00670-1] [Medline: 36085306]
17. Atlantis E, Sullivan T. Bidirectional association between depression and sexual dysfunction: a systematic review and meta-analysis. J Sex Med 2012;9(6):1497-1507. [doi: 10.1111/j.1743-6109.2012.02709.x] [Medline: 22462756]
18. Xie Y, Shi L, Xiong X, Wu E, Veasley C, Dade C. Economic burden and quality of life of vulvodynia in the United States. Curr Med Res Opin 2012;28(4):601-608. [doi: 10.1185/03007995.2012.666963] [Medline: 22356119]
19. Beier KML, Kurt K. Sexual Medicine in Clinical Practice. New York, NY: Springer; 2013.
20. Göhring J, Drewes M, Kalder M, Kostev K. Germany endometriosis pattern changes; prevalence and therapy over 2010 and 2019 years: a retrospective cross-sectional study. Int J Fertil Steril 2022;16(2):85-89 [FREE Full text] [doi: 10.22074/IJFS.2021.528397.1113] [Medline: 35639651]
21. Buggio L, Barbara G, Facchin F, Frattaruolo MP, Aimi G, Berlanda N. Self-management and psychological-sexological interventions in patients with endometriosis: strategies, outcomes, and integration into clinical care. Int J Womens Health 2017;9:281-293. [doi: 10.2147/ijwh.s119724]
22. Zondervan KT, Becker CM, Missmer SA. Endometriosis. N Engl J Med 2020;382(13):1244-1256. [doi: 10.1056/NEJMra1810764] [Medline: 32212520]
23. Fairbanks F, Abdo CH, Baracat EC, Podgaec S. Endometriosis doubles the risk of sexual dysfunction: a cross-sectional study in a large amount of patients. Gynecol Endocrinol 2017;33(7):544-547. [doi: 10.1080/09513590.2017.1302421] [Medline: 28347195]
24. Edgley K, Saunders PTK, Whitaker LHR, Horne AW, Tsanas A. Insights into endometriosis symptom trajectories and assessment of surgical intervention outcomes using longitudinal actigraphy. NPJ Digit Med 2025;8(1):236 [FREE Full text] [doi: 10.1038/s41746-025-01629-8] [Medline: 40316659]

XSL•FO
RenderX

25. Velten J, Pantazidis P, Benecke A, Bräscher A, Fehm L, Fladung A, et al. Wie häufig werden diagnosen aus dem bereich der sexuellen Funktionsstörungen an deutschen hochschulambulanzen für psychotherapie an psychologischen Instituten vergeben? Z Sex Forsch 2021;34(01):5-14. [doi: 10.1055/a-1362-2243]
26. Barnhoorn PC, Prins IC, Zuurveen HR, Oudsten BLD, Ouden MEMD, Numans ME, et al. Let's talk about sex: exploring factors influencing the discussion of sexual health among chronically Ill patients in general practice. BMC Prim Care 2022;23(1):49 [FREE Full text] [doi: 10.1186/s12875-022-01660-8] [Medline: 35305576]
27. As-Sanie S, Black R, Giudice LC, Gray Valbrun T, Gupta J, Jones B, et al. Assessing research gaps and unmet needs in endometriosis. Am J Obstet Gynecol 2019;221(2):86-94. [doi: 10.1016/j.ajog.2019.02.033] [Medline: 30790565]
28. Illum LRH, Forman A, Melgaard A, Hansen KE, Hansen SN, Nyegaard M, et al. Temporal and regional differences in the incidence of hospital-diagnosed endometriosis: a Danish population-based study. Acta Obstet Gynecol Scand 2022;101(7):737-746 [FREE Full text] [doi: 10.1111/aogs.14364] [Medline: 35434780]
29. Velten J, Margraf J. Exploring barriers and facilitators to women's intention and behavior to seek treatment for distressing sexual problems. PLoS One 2023;18(7):e0288205 [FREE Full text] [doi: 10.1371/journal.pone.0288205] [Medline: 37463132]
30. Kocas HD, Rubin LR, Lobel M. Stigma and mental health in endometriosis. Eur J Obstet Gynecol Reprod Biol X 2023;19:100228 [FREE Full text] [doi: 10.1016/j.eurox.2023.100228] [Medline: 37654520]
31. Avasthi A, Grover S, Sathyanarayana Rao TS. Clinical practice guidelines for management of sexual dysfunction. Indian J Psychiatry 2017;59(Suppl 1):S91-S115 [FREE Full text] [doi: 10.4103/0019-5545.196977] [Medline: 28216788]
32. Parish SJ, Hahn SR, Goldstein SW, Giraldi A, Kingsberg SA, Larkin L, et al. The International Society for the Study of Women's Sexual Health process of care for the identification of sexual concerns and problems in women. Mayo Clin Proc 2019;94(5):842-856 [FREE Full text] [doi: 10.1016/j.mayocp.2019.01.009] [Medline: 30954288]
33. Pereira VM, Arias-Carrión O, Machado S, Nardi AE, Silva AC. Sex therapy for female sexual dysfunction. Int Arch Med 2013;6(1):37 [FREE Full text] [doi: 10.1186/1755-7682-6-37] [Medline: 24066697]
34. Goldstein I, Kim NN, Clayton AH, DeRogatis LR, Giraldi A, Parish SJ, et al. Hypoactive sexual desire disorder: International Society for the Study of Women's Sexual Health (ISSWSH) expert consensus panel review. Mayo Clin Proc 2017;92(1):114-128 [FREE Full text] [doi: 10.1016/j.mayocp.2016.09.018] [Medline: 27916394]
35. Goldstein I, Komisaruk BR, Pukall CF, Kim NN, Goldstein AT, Goldstein SW, et al. International Society for the Study of Women's Sexual Health (ISSWSH) review of epidemiology and pathophysiology, and a consensus nomenclature and process of care for the management of persistent genital arousal disorder/Genito-pelvic dysesthesia (PGAD/GPD). J Sex Med 2021;18(4):665-697 [FREE Full text] [doi: 10.1016/j.jsxm.2021.01.172] [Medline: 33612417]
36. Bornstein J, Preti M, Radici G, Stockdale CK, Vieira-Baptista P. Vulvodynia: a neglected chronic pain diagnosis. Pain 2019;160(7):1680-1681. [doi: 10.1097/j.pain.0000000000001559] [Medline: 31219952]
37. Jorge CH, Bø K, Chiazuto Catai C, Oliveira Brito LG, Driusso P, Kolberg Tennfjord M. Pelvic floor muscle training as treatment for female sexual dysfunction: a systematic review and meta-analysis. Am J Obstet Gynecol 2024;231(1):51-66.e1. [doi: 10.1016/j.ajog.2024.01.001] [Medline: 38191016]
38. Sahin E, Brand A, Cetindag EN, Messelink B, Yosmaoglu HB. Pelvic physical therapy for male sexual disorders: a narrative review. Int J Impot Res 2025;37(12):941-949. [doi: 10.1038/s41443-025-01034-5] [Medline: 40016382]
39. van Reijn-Baggen DA, Han-Geurts IJM, Voorham-van der Zalm PJ, Pelger RCM, Hagenaars-van Miert CHAC, Laan ETM. Pelvic floor physical therapy for pelvic floor hypertonicity: a systematic review of treatment efficacy. Sex Med Rev 2022;10(2):209-230 [FREE Full text] [doi: 10.1016/j.sxmr.2021.03.002] [Medline: 34127429]
40. Parish SJ, Simon JA, Davis SR, Giraldi A, Goldstein I, Goldstein SW, et al. International society for the study of women's sexual health clinical practice guideline for the use of systemic testosterone for hypoactive sexual desire disorder in women. J Sex Med 2021;18(5):849-867 [FREE Full text] [doi: 10.1016/j.jsxm.2020.10.009] [Medline: 33814355]
41. Frühauf S, Gerger H, Schmidt HM, Munder T, Barth J. Efficacy of psychological interventions for sexual dysfunction: a systematic review and meta-analysis. Arch Sex Behav 2013;42(6):915-933 [FREE Full text] [doi: 10.1007/s10508-012-0062-0] [Medline: 23559141]
42. Almuqahwi A, Alabdrabulridha H, Aljumaiah RM, Alfaifi AJ, Alnaim MF, Alfaifi IA, et al. A systematic review on the relationship between physical activity and sexual function in adults. Cureus 2023;15(12):e51307 [FREE Full text] [doi: 10.7759/cureus.51307] [Medline: 38288234]
43. Maseroli E, Rastrelli G, Di Stasi V, Cipriani S, Scavello I, Todisco T, et al. Physical activity and female sexual dysfunction: a lot helps, but not too much. J Sex Med 2021;18(7):1217-1229. [doi: 10.1016/j.jsxm.2021.04.004] [Medline: 34099414]
44. Kharaghani R, Khani ME, Mahmoodi Dangesaraki M, Damghanian M. The effects of psychological interventions on sexual function of women: a systematic review and meta-analysis. Nursing Practice Today 2020;7(3). [doi: 10.18502/npt.v7i3.3345]
45. Jaderek I, Lew-Starowicz M. A systematic review on mindfulness meditation-based interventions for sexual dysfunctions. J Sex Med 2019;16(10):1581-1596. [doi: 10.1016/j.jsxm.2019.07.019] [Medline: 31570137]
46. Brotto LA, Stephenson KR, Marshall N, Balvan M, Okara Y, Mahar EA. Evaluating a digital health tool designed to improve low sexual desire in women: mixed-methods implementation science study. J Med Internet Res 2025;27:e69828 [FREE Full text] [doi: 10.2196/69828] [Medline: 40132194]

XSL•FO
RenderX

47. Basson R, Rees P, Wang R, Montejo AL, Incrocci L. Sexual function in chronic illness. J Sex Med 2010;7(1 Pt 2):374-388. [doi: 10.1111/j.1743-6109.2009.01621.x] [Medline: 20092445]
48. Hobbs LJ, Mitchell KR, Graham CA, Trifonova V, Bailey J, Murray E, et al. Help-seeking for sexual difficulties and the potential role of interactive digital interventions: findings from the third British national survey of sexual attitudes and lifestyles. J Sex Res 2019;56(7):937-946 [FREE Full text] [doi: 10.1080/00224499.2019.1586820] [Medline: 30907691]
49. Zarski AC, Berking M, Fackiner C, Rosenau C, Ebert DD. Internet-based guided self-help for vaginal penetration difficulties: results of a randomized controlled pilot trial. J Sex Med 2017;14(2):238-254. [doi: 10.1016/j.jsxm.2016.12.232] [Medline: 28161080]
50. Zarski AC, Berking M, Ebert DD. Efficacy of internet-based treatment for genito-pelvic pain/penetration disorder: results of a randomized controlled trial. J Consult Clin Psychol 2021;89(11):909-924. [doi: 10.1037/ccp0000665] [Medline: 34843312]
51. van Lankveld J. Internet-based interventions for women's sexual dysfunction. Curr Sex Health Rep 2016;8(3):136-143 [FREE Full text] [doi: 10.1007/s11930-016-0087-9] [Medline: 27512346]
52. Brotto LA, Goldmeier D. Mindfulness interventions for treating sexual dysfunctions: the gentle science of finding focus in a multitask world. J Sex Med 2015;12(8):1687-1689. [doi: 10.1111/jsm.12941] [Medline: 26289539]
53. Hucker A, McCabe MP. An online, mindfulness-based, cognitive-behavioral therapy for female sexual difficulties: impact on relationship functioning. J Sex Marital Ther 2014;40(6):561-576. [doi: 10.1080/0092623X.2013.796578] [Medline: 24308322]
54. Jones LM, McCabe MP. The effectiveness of an internet-based psychological treatment program for female sexual dysfunction. J Sex Med 2011;8(10):2781-2792. [doi: 10.1111/j.1743-6109.2011.02381.x] [Medline: 21771279]
55. Meyers M, Margraf J, Velten J. A qualitative study of women's experiences with cognitive-behavioral and mindfulness-based online interventions for low sexual desire. J Sex Res 2022;59(9):1082-1091. [doi: 10.1080/00224499.2022.2056565] [Medline: 35435791]
56. Mahar EA, Stephenson KR, Brotto LA. A randomized controlled trial of online mindfulness and cognitive-behavioral interventions for sexual interest/arousal disorder in women: eSense. Behav Res Ther 2025;188:104732 [FREE Full text] [doi: 10.1016/j.brat.2025.104732] [Medline: 40147247]
57. Goeldner M, Gehder S. Digital health applications (DiGAs) on a fast track: insights from a data-driven analysis of prescribable digital therapeutics in Germany from 2020 to mid-2024. J Med Internet Res 2024;26:e59013 [FREE Full text] [doi: 10.2196/59013] [Medline: 39208415]
58. Zarski AC, Berking M, Ebert DD. Efficacy of internet-based guided treatment for genito-pelvic pain/penetration disorder: rationale, treatment protocol, and design of a randomized controlled trial. Front Psychiatry 2017;8:260 [FREE Full text] [doi: 10.3389/fpsyt.2017.00260] [Medline: 29403395]
59. Kliesch S, Cremers J, Krallmann C, Epplen R, Scheffer B, Schubert T, et al. App-based therapy of erectile dysfunction using a digital health application (EDDIG Study): a randomized, single-blind, controlled trial. Eur Urol Focus 2024;10(6):1003-1010 [FREE Full text] [doi: 10.1016/j.euf.2024.05.020] [Medline: 38853028]
60. Impact on quality of life by additional use of the digital health application "Endo-App" to standard therapy: monocentric, prospective, randomised trial. ClinicalTrials.gov. 2021. URL: https://clinicaltrials.gov/study/NCT04883073 [accessed 2026-01-23]
61. Rohloff N, Götz T, Kortekamp SS, Heinze NR, Weber C, Schäfer SD. Influence of app-based self-management on the quality of life of women with endometriosis. Cureus 2024;16(8):e67655. [doi: 10.7759/cureus.67655] [Medline: 39314601]
62. Gratzke C, Schönburg S, Eger S, Raude K, Grabbert M, Astheimer S, et al. A randomized trial of an app-based therapeutic for lower urinary tract symptoms. NEJM Evid 2025;4(4):EVIDoa2400290. [doi: 10.1056/EVIDoa2400290] [Medline: 40130971]
63. Rödiger H, Busse R, Henschke C. The introduction of maximum reimbursement prices for digital health applications in Germany in 2022: current developments. Health Policy 2025;155:105284 [FREE Full text] [doi: 10.1016/j.healthpol.2025.105284] [Medline: 40056714]
64. Burchert S, Alkneme MS, Bird M, Carswell K, Cuijpers P, Hansen P, et al. User-centered app adaptation of a low-intensity E-mental health intervention for Syrian refugees. Front Psychiatry 2018;9:663 [FREE Full text] [doi: 10.3389/fpsyt.2018.00663] [Medline: 30740065]
65. Boß L, Lehr D, Reis D, Vis C, Riper H, Berking M, et al. Reliability and validity of assessing user satisfaction with web-based health interventions. J Med Internet Res 2016;18(8):e234 [FREE Full text] [doi: 10.2196/jmir.5952] [Medline: 27582341]
66. Moorthy P, Weinert L, Harms BC, Anders C, Siegel F. German version of the mHealth app usability questionnaire in a cohort of patients with cancer: translation and validation study. JMIR Hum Factors 2023;10:e51090 [FREE Full text] [doi: 10.2196/51090] [Medline: 37910144]
67. Bergström C, Skog R, Eriksson LE, Lampic C, Wettergren L. Efficacy of a web-based psychoeducational intervention targeting young adults with sexual problems 1.5 years after cancer diagnosis-results from a randomized controlled trial. Digit Health 2024;10:20552076241310037 [FREE Full text] [doi: 10.1177/20552076241310037] [Medline: 39741983]

XSL•FO
RenderX

68. Kim SH, Sung JH, Yoo SH, Kim S, Lee K, Oh EG, et al. Effects of digital self-management symptom interventions on symptom outcomes in adult cancer patients: a systematic review and meta-analysis. Eur J Oncol Nurs 2023;66:102404. [doi: 10.1016/j.ejon.2023.102404] [Medline: 37517339]
69. Schover LR, Yuan Y, Fellman BM, Odensky E, Lewis PE, Martinetti P. Efficacy trial of an Internet-based intervention for cancer-related female sexual dysfunction. J Natl Compr Canc Netw 2013;11(11):1389-1397. [doi: 10.6004/jnccn.2013.0162] [Medline: 24225972]
70. Linardon J, Torous J, Firth J, Cuijpers P, Messer M, Fuller-Tyszkiewicz M. Current evidence on the efficacy of mental health smartphone apps for symptoms of depression and anxiety. A meta-analysis of 176 randomized controlled trials. World Psychiatry 2024;23(1):139-149. [doi: 10.1002/wps.21183] [Medline: 38214614]
71. Schreiter S, Mascarell-Maricic L, Rakitzis O, Volkmann C, Kaminski J, Daniels MA. Digital health applications in the area of mental health. Dtsch Arztebl Int 2023;120(47):797-803 [FREE Full text] [doi: 10.3238/arztebl.m2023.0208] [Medline: 37732500]
72. Guo W, Nazari N, Sadeghi M. Cognitive-behavioral treatment for insomnia and mindfulness-based stress reduction in nurses with insomnia: a non-inferiority internet delivered randomized controlled trial. Peer J 2024;12:e17491. [doi: 10.7717/peerj.17491] [Medline: 39071123]
73. Ross EL, Jamison RN, Nicholls L, Perry BM, Nolen KD. Clinical integration of a smartphone app for patients with chronic pain: retrospective analysis of predictors of benefits and patient engagement between clinic visits. J Med Internet Res 2020;22(4):e16939 [FREE Full text] [doi: 10.2196/16939] [Medline: 32297871]
74. Strain JDR, Welch L, Sadler E. Systematic review and narrative synthesis of the experiences of individuals with chronic pain participating in digital pain management interventions. PLoS One 2024;19(7):e0306455 [FREE Full text] [doi: 10.1371/journal.pone.0306455] [Medline: 38976713]
75. Vad VB, Madrazo-Ibarra A, Estrin D, Pollak JP, Carroll KM, Vojta D, et al. "Back Rx, a personalized mobile phone application for discogenic chronic low back pain: a prospective pilot study". BMC Musculoskelet Disord 2022;23(1):923 [FREE Full text] [doi: 10.1186/s12891-022-05883-9] [Medline: 36261825]
76. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020;22(9):e20283 [FREE Full text] [doi: 10.2196/20283] [Medline: 32990635]
77. Christensen H, Griffiths KM, Farrer L. Adherence in internet interventions for anxiety and depression. J Med Internet Res 2009;11(2):e13 [FREE Full text] [doi: 10.2196/jmir.1194] [Medline: 19403466]
78. Bremer V, Chow PI, Funk B, Thorndike FP, Ritterband LM. Developing a process for the analysis of user journeys and the prediction of dropout in digital health interventions: machine learning approach. J Med Internet Res 2020;22(10):e17738 [FREE Full text] [doi: 10.2196/17738] [Medline: 33112241]
79. Cohen KA, Schleider JL. Adolescent dropout from brief digital mental health interventions within and beyond randomized trials. Internet Interv 2022;27:100496 [FREE Full text] [doi: 10.1016/j.invent.2022.100496] [Medline: 35257001]
80. Karyotaki E, Kleiboer A, Smit F, Turner DT, Pastor AM, Andersson G, et al. Predictors of treatment dropout in self-guided web-based interventions for depression: an 'individual patient data' meta-analysis. Psychol Med 2015;45(13):2717-2726 [FREE Full text] [doi: 10.1017/S0033291715000665] [Medline: 25881626]
81. Moshe I, Terhorst Y, Paganini S, Schlicker S, Pulkki-Råback L, Baumeister H, et al. Predictors of dropout in a digital intervention for the prevention and treatment of depression in patients with chronic back pain: secondary analysis of two randomized controlled trials. J Med Internet Res 2022;24(8):e38261 [FREE Full text] [doi: 10.2196/38261] [Medline: 36040780]
82. Pedersen DH, Mansourvar M, Sortsø C, Schmidt T. Predicting dropouts from an electronic health platform for lifestyle interventions: analysis of methods and predictors. J Med Internet Res 2019;21(9):e13617 [FREE Full text] [doi: 10.2196/13617] [Medline: 31486409]
83. Gan DZQ, McGillivray L, Han J, Christensen H, Torok M. Effect of engagement with digital interventions on mental health outcomes: a systematic review and meta-analysis. Front Digit Health 2021;3:764079 [FREE Full text] [doi: 10.3389/fdgth.2021.764079] [Medline: 34806079]
84. Goodday SM, Karlin E, Brooks A, Chapman C, Harry C, Lugo N, et al. Value of engagement in digital health technology research: evidence across 6 unique cohort studies. J Med Internet Res 2024;26:e57827 [FREE Full text] [doi: 10.2196/57827] [Medline: 39226552]
85. Gudmundsdóttir SL, Ballarini T, Ámundadóttir ML, Mészáros J, Eysteinsdottir JH, Thorleifsdottir RH, et al. Engagement, retention, and acceptability in a digital health program for atopic dermatitis: prospective interventional study. JMIR Form Res 2023;7:e41227 [FREE Full text] [doi: 10.2196/41227] [Medline: 36975050]
86. Huberty J, Green J, Puzia M, Stecher C. Evaluation of mood check-in feature for participation in meditation mobile app users: retrospective longitudinal analysis. JMIR Mhealth Uhealth 2021;9(4):e27106 [FREE Full text] [doi: 10.2196/27106] [Medline: 33890865]
87. Jamison RN, Jurcik DC, Edwards RR, Huang CC, Ross EL. A pilot comparison of a smartphone app with or without 2-way messaging among chronic pain patients: who benefits from a pain app? Clin J Pain 2017;33(8):676-686 [FREE Full text] [doi: 10.1097/AJP.0000000000000455] [Medline: 27898460]

XSL•FO
RenderX

88. Lipschitz JM, Pike CK, Hogan TP, Murphy SA, Burdick KE. The engagement problem: a review of engagement with digital mental health interventions and recommendations for a path forward. Curr Treat Options Psychiatry 2023;10(3):119-135 [FREE Full text] [doi: 10.1007/s40501-023-00297-3] [Medline: 38390026]
89. Wang T, Ho MH, Tong MCF, Chow JC, Voss JG, Lin CC. Effects of patient-reported outcome tracking and health information provision via remote patient monitoring software on patient outcomes in oncology care: a systematic review and meta-analysis. Semin Oncol Nurs 2023;39(5):151473. [doi: 10.1016/j.soncn.2023.151473] [Medline: 37516624]
90. Pettersson A, Berterö CM. How women with endometriosis experience health care encounters. Womens Health Rep (New Rochelle) 2020;1(1):529-542 [FREE Full text] [doi: 10.1089/whr.2020.0099] [Medline: 33786519]
91. Beier KM, von Heyden M. Sexuelle funktionsstörungen im digitalen zeitalter – therapeutische potenziale und entwicklungsperspektiven digitaler gesundheitsanwendungen. Gynäkologische Endokrinologie 2025;23(4):247-252. [doi: 10.1007/s10304-025-00645-x]
92. Beck AT, Steer RA, Brown G. Beck Depression Inventory (BDI-II). 2nd ed. San Antonio, TX: Pearson; 1996.
93. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
94. Decker-Palmer M, Klodowski D, Thompson T, Lanoue M, Messina A, Schroeder D, et al. Freelisting: a technique for enhancing the community health needs assessment. Community Health Equity Res Policy 2024;44(2):201-208. [doi: 10.1177/2752535X221146232] [Medline: 36513364]
95. Keddem S, Barg FK, Frasso R. Practical guidance for studies using freelisting interviews. Prev Chronic Dis 2021;18:E04 [FREE Full text] [doi: 10.5888/pcd17.200355] [Medline: 33444525]
96. Applied Mental Health Research Group. Module 1: qualitative assessment (adult version). In: Design, Implementation, Monitoring, and Evaluation of Mental Health and Psychosocial Assistance Programs for Trauma Survivors in Low Resource Countries: A User's Manual for Researchers and Program Implementers. Module 1: Qualitative Assessment (Adult Version). Baltimore, MD: Johns Hopkins Bloomberg School of Public Health; 2013.
97. f4transkript. Audiotranskription. 2025. URL: https://www.audiotranskription.de/f4transkript [accessed 2026-02-06]
98. Schreier M. Ways of doing qualitative content analysis: disentangling terms and terminologies. Forum Qual Soc Res 2014;15(1). [doi: 10.17169/fqs-15.1.2043]
99. Richards DA, Bazeley P, Borglin G, Craig P, Emsley R, Frost J, et al. Integrating quantitative and qualitative data and findings when undertaking randomised controlled trials. BMJ Open 2019;9(11):e032081 [FREE Full text] [doi: 10.1136/bmjopen-2019-032081] [Medline: 31772096]
100. Hwang Y, Knobf MT, Sadler LS. Integration in mixed-methods research with an exemplar explanatory sequential study. Nurs Res 2025;74(2):144-149. [doi: 10.1097/NNR.0000000000000796] [Medline: 39616427]
101. Basson R. Sexual function of women with chronic illness and cancer. Womens Health (Lond) 2010;6(3):407-429 [FREE Full text] [doi: 10.2217/whe.10.23] [Medline: 20426607]
102. Attkisson CC, Zwick R. The client satisfaction questionnaire. Psychometric properties and correlations with service utilization and psychotherapy outcome. Eval Program Plann 1982;5(3):233-237. [doi: 10.1016/0149-7189(82)90074-x] [Medline: 10259963]
103. Kopka M, Slagman A, Schorr C, Krampe H, Altendorf M, Balzer F, et al. German mHealth App Usability Questionnaire (G-MAUQ) and short version (G-MAUQ-S): translation and validation study. Smart Health 2024;34:100517. [doi: 10.1016/j.smhl.2024.100517]
104. Hayes MHS, Paterson DG. Experimental development of the graphic rating method. Psychol Bull 1921;18(2):98-99 [FREE Full text]
105. Ladwig I, Rief W, Nestoriuc Y. Welche risiken und nebenwirkungen hat psychotherapie? - Entwicklung des inventars zur erfassung negativer effekte von psychotherapie (INEP). Verhaltenstherapie 2014;24(4):252-263. [doi: 10.1159/000367928]
106. Derogatis LR, Revicki DA, Rosen RC, Jordan R, Lucas J, Spana C. Psychometric validation of the Female Sexual Distress Scale-Desire/Arousal/Orgasm. J Patient Rep Outcomes 2021;5(1):100 [FREE Full text] [doi: 10.1186/s41687-021-00359-1] [Medline: 34559353]
107. Velten J, Zarski AC. Therapie-Tools Sexuelle Funktionsstörungen: Mit E-Book inside und Arbeitsmaterial. 1st ed. Weinheim: Beltz; 2022.
108. Berner MM, Kriston L, Zahradnik HP, Härter M, Rohde A. Überprüfung der Gültigkeit und Zuverlässigkeit des deutschen Female Sexual Function Index (FSFI-d). Geburtsh Frauenheilk 2004;64(3):293-303. [doi: 10.1055/s-2004-815815]
109. Bornefeld-Ettmann P, Steil R, Höfling V, Weßlau C, Lieberz K, Rausch S, et al. Validation of the German version of the Sexual Self-Esteem Inventory for Women and its application in a sample of sexually and physically abused women. Sex Roles 2017;79(1-2):109-122. [doi: 10.1007/s11199-017-0849-5]
110. Ter Kuile MM, Melles R, de Groot HE, Tuijnman-Raasveld CC, van Lankveld JJDM. Therapist-aided exposure for women with lifelong vaginismus: a randomized waiting-list control trial of efficacy. J Consult Clin Psychol 2013;81(6):1127-1136. [doi: 10.1037/a0034292] [Medline: 24060195]
111. Klaassen M, Ter Kuile MM. Development and initial validation of the Vaginal Penetration Cognition Questionnaire (VPCQ) in a sample of women with vaginismus and dyspareunia. J Sex Med 2009;6(6):1617-1627. [doi: 10.1111/j.1743-6109.2009.01217.x] [Medline: 19538428]

XSL•FO
RenderX

112. Hahlweg K. Fragebogen zur Partnerschaftsdiagnostik (PFB) [Partnership Questionnaire]. Göttingen: Hogrefe; 1996. URL: https://www.testzentrale.de/shop/fragebogen-zur-partnerschaftsdiagnostik.html [accessed 2026-01-26]
113. Hinz A, Stöbel-Richter Y, Brähler E. Der Partnerschaftsfragebogen (PFB). Diagnostica 2001;47(3):132-141. [doi: 10.1026//0012-1924.47.3.132]
114. Klingler OJ, Loewit KK. Der Fragebogen „Ressourcen in Sexualität und Partnerschaft“ (RSP) – konzeption und erste ergebnisse zur validität. Z Differ Diagnost Psychol 1996;17(4):268-275 [FREE Full text]
115. Derogatis LR. Brief Symptom Inventory (BSI): Administration, scoring, and procedures manual. 3rd ed. Minneapolis: National Computer Systems; 1993.
116. Geisheim C, Hahlweg K, Fiegenbaum W, Frank M, Schröder B, von Witzleben I. Das Brief Symptom Inventory (BSI) als instrument zur qualitätssicherung in der psychotherapie. Diagnostica 2002;48(1):28-36. [doi: 10.1026//0012-1924.48.1.28]
117. Klein EM, Brähler E, Dreier M, Reinecke L, Müller KW, Schmutzer G, et al. The German version of the Perceived Stress Scale - psychometric characteristics in a representative German community sample. BMC Psychiatry 2016;16:159 [FREE Full text] [doi: 10.1186/s12888-016-0875-9] [Medline: 27216151]
118. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, PROMIS Cooperative Group. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH roadmap cooperative group during its first two years. Med Care 2007;45(5 Suppl 1):S3-S11 [FREE Full text] [doi: 10.1097/01.mlr.0000258615.42478.55] [Medline: 17443116]
119. Schmidt S, Mühlan H, Power M. The EUROHIS-QOL 8-item index: psychometric results of a cross-cultural field study. Eur J Public Health 2006;16(4):420-428. [doi: 10.1093/eurpub/cki155] [Medline: 16141303]
120. Klute M, Laekeman M, Kuss K, Petzke F, Dieterich A, Leha A, et al. Cross-cultural adaptation and validation of the German Central Sensitization Inventory (CSI-GE). BMC Musculoskelet Disord 2021;22(1):708 [FREE Full text] [doi: 10.1186/s12891-021-04481-5] [Medline: 34407773]
121. Brookes ST, Donovan JL, Wright M, Jackson S, Abrams P. A scored form of the Bristol Female Lower Urinary Tract Symptoms Questionnaire: data from a randomized controlled trial of surgery for women with stress incontinence. Am J Obstet Gynecol 2004;191(1):73-82. [doi: 10.1016/j.ajog.2003.12.027] [Medline: 15295345]
122. Rybak A. Survey mode and nonresponse bias: A meta-analysis based on the data from the international social survey programme waves 1996-2018 and the European social survey rounds 1 to 9. PLoS One 2023;18(3):e0283092 [FREE Full text] [doi: 10.1371/journal.pone.0283092] [Medline: 36928697]
123. Patton MQ. Qualitative Research & Evaluation Methods: Integrating Theory and Practice. 4th ed. Thousand Oaks: SAGE Publications; 2015.
124. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: Medical Research Council guidance. BMJ 2015;350:h1258 [FREE Full text] [doi: 10.1136/bmj.h1258] [Medline: 25791983]
125. Miro. URL: https://miro.com/ [accessed 2026-01-26]
126. Blair C, Best P, Burns P, Campbell A, Davidson G, Duffy J, et al. 'Getting involved in research': a co-created, co-delivered and co-analysed course for those with lived experience of health and social care services. Res Involv Engagem 2022;8(1):20 [FREE Full text] [doi: 10.1186/s40900-022-00353-x] [Medline: 35578275]
127. Wollin-Giering S, Hoffmann M, Höfting J, Ventzke C. Automatic transcription of English and German qualitative interviews. Forum Qual Sozialforschung 2024;25(1). [doi: 10.17169/fqs-25.1.4129]
128. Plano Clark VL, Schumacher K, West C, Edrington J, Dunn LB, Harzstark A, et al. Practices for embedding an interpretive qualitative approach within a randomized clinical trial. J Mix Methods Res 2013;7(3):219-242. [doi: 10.1177/1558689812474372]
129. Laura H. Zweiarmige mixed-methods machbarkeitsstudie zur anwendung einer mobilen applikation für betroffene von endometriose und sexuellen funktionsstörungen. Deutsches Register Klinischer Studien. 2024. URL: https://drks.de/ [accessed 2026-02-01]
130. Eysenbach G. CONSORT-EHEALTH: implementation of a checklist for authors and editors to improve reporting of web-based and mobile randomized controlled trials. Stud Health Technol Inform 2013;192:657-661. [Medline: 23920638]
131. Creswell JW, Klassen AC, Clark VLP, Smith KC. Best Practices for Mixed Methods Research in the Health Sciences. Bethesda, MD: Office of Behavioral and Social Sciences Research, National Institutes of Health; 2011.
132. Cloitre M, Shevlin M, Brewin CR, Bisson JI, Roberts NP, Maercker A, et al. The international trauma questionnaire: development of a self-report measure of ICD-11 PTSD and complex PTSD. Acta Psychiatr Scand 2018;138(6):536-546. [doi: 10.1111/acps.12956] [Medline: 30178492]
133. Kübler U. Structured clinical interview for DSM-IV (SCID). In: Gellman MD, Turner JR, editors. Encyclopedia of Behavioral Medicine. New York, NY: Springer New York; 2013:1919-1920.
134. Rodriguez A, Tuck C, Dozier MF, Lewis SC, Eldridge S, Jackson T, et al. Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: a scoping review. Clin Trials 2022;19(4):452-463 [FREE Full text] [doi: 10.1177/17407745221087469] [Medline: 35730910]
135. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]

XSL•FO
RenderX

136. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, REDCap Consortium. The REDCap Consortium: building an international community of software platform partners. J Biomed Inform 2019;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
137. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
138. Klinitzke G, Romppel M, Häuser W, Brähler E, Glaesmer H. [The German version of the Childhood Trauma Questionnaire (CTQ): psychometric characteristics in a representative sample of the general population]. Psychother Psychosom Med Psychol 2012;62(2):47-51. [doi: 10.1055/s-0031-1295495] [Medline: 22203470]
139. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health Apps: systematic search and panel-based usage analysis. J Med Internet Res 2019;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
140. Boucher EM, Raiker JS. Engagement and retention in digital mental health interventions: a narrative review. BMC Digit Health 2024;2(1):52. [doi: 10.1186/s44247-024-00105-9]
141. Diaz-Mohedo E, Hita-Contreras F, Castro-Martin E, Pilat A, Perez-Dominguez B, Valenza-Peña G. Using myofascial therapy to improve psychological outcomes, quality of life, and sexual function in women with chronic pelvic pain-a case series. Healthcare (Basel) 2024;12(3):304 [FREE Full text] [doi: 10.3390/healthcare12030304] [Medline: 38338190]
142. Livieri G, Mangina E, Protopapadakis ED, Panayiotou AG. The gaps and challenges in digital health technology use as perceived by patients: a scoping review and narrative meta-synthesis. Front Digit Health 2025;7:1474956 [FREE Full text] [doi: 10.3389/fdgth.2025.1474956] [Medline: 40212901]
143. Erbe D, Eichert H, Riper H, Ebert DD. Blending face-to-face and internet-based interventions for the treatment of mental disorders in adults: systematic review. J Med Internet Res 2017;19(9):e306 [FREE Full text] [doi: 10.2196/jmir.6588] [Medline: 28916506]
144. Liu H, Xie Z, Or C. Willingness to pay for health apps, its sociodemographic correlates, and reasons for being unwilling to pay. Digit Health 2024;10:20552076241248925 [FREE Full text] [doi: 10.1177/20552076241248925] [Medline: 38698831]
145. Berkhout C, Berbra O, Favre J, Collins C, Calafiore M, Peremans L, et al. Defining and evaluating the Hawthorne effect in primary care, a systematic review and meta-analysis. Front Med (Lausanne) 2022;9:1033486 [FREE Full text] [doi: 10.3389/fmed.2022.1033486] [Medline: 36425097]
146. Long PA, Huberts AS, di Torrero AN, Otto LR, Rogge AA, Ritschl V, et al. The mere-measurement effect of patient-reported outcomes: a systematic review and meta-analysis. Qual Life Res 2025;34(5):1211-1220. [doi: 10.1007/s11136-025-03909-y] [Medline: 39907942]
147. Gallagher MW, Long LJ, Richardson A, D'Souza J, Boswell JF, Farchione TJ, et al. Examining hope as a transdiagnostic mechanism of change across anxiety disorders and CBT treatment protocols. Behav Ther 2020;51(1):190-202 [FREE Full text] [doi: 10.1016/j.beth.2019.06.001] [Medline: 32005336]
148. Bradford A, Meston CM. Placebo response in the treatment of women's sexual dysfunctions: a review and commentary. J Sex Marital Ther 2009;35(3):164-181. [doi: 10.1080/00926230802716302] [Medline: 19360517]
149. Song MK, Ward SE. Assessment effects in educational and psychosocial intervention trials: an important but often-overlooked problem. Res Nurs Health 2015;38(3):241-247 [FREE Full text] [doi: 10.1002/nur.21651] [Medline: 25728502]
150. Chen K, Huang JJ, Torous J. Hybrid care in mental health: a framework for understanding care, research, and future opportunities. NPP Digit Psychiatry Neurosci 2024;2(1):16. [doi: 10.1038/s44277-024-00016-7] [Medline: 41360926]
151. Loughnane C, Laiti J, O'Donovan R, Dunne PJ. Systematic review exploring human, AI, and hybrid health coaching in digital health interventions: trends, engagement, and lifestyle outcomes. Front Digit Health 2025;7:1536416 [FREE Full text] [doi: 10.3389/fdgth.2025.1536416] [Medline: 40343215]
152. Saleem M, Kühne L, De Santis KK, Christianson L, Brand T, Busse H. Understanding engagement strategies in digital interventions for mental health promotion: scoping review. JMIR Ment Health 2021;8(12):e30000 [FREE Full text] [doi: 10.2196/30000] [Medline: 34931995]
153. Moe-Byrne T, Shepherd J, Merecz-Kot D, Sinokki M, Naumanen P, Hakkaart-van Roijen L, et al. Effectiveness of tailored digital health interventions for mental health at the workplace: a systematic review of randomised controlled trials. PLOS Digit Health 2022;1(10):e0000123 [FREE Full text] [doi: 10.1371/journal.pdig.0000123] [Medline: 36812547]
154. Bidargaddi N, Almirall D, Murphy S, Nahum-Shani I, Kovalcik M, Pituch T, et al. To prompt or not to prompt? A microrandomized trial of time-varying push notifications to increase proximal engagement with a mobile health app. JMIR Mhealth Uhealth 2018;6(11):e10123 [FREE Full text] [doi: 10.2196/10123] [Medline: 30497999]
155. Bitrián P, Buil I, Catalán S. Enhancing user engagement: The role of gamification in mobile apps. J Bus Res 2021;132:170-185 [FREE Full text] [doi: 10.1016/j.jbusres.2021.04.028]
156. Twomey C, O'Reilly G, Meyer B. Effectiveness of an individually-tailored computerised CBT programme (Deprexis) for depression: a meta-analysis. Psychiatry Res 2017;256:371-377. [doi: 10.1016/j.psychres.2017.06.081] [Medline: 28686935]
157. Molina-Recio G, Molina-Luque R, Jiménez-García AM, Ventura-Puertos PE, Hernández-Reyes A, Romero-Saldaña M. Proposal for the user-centered design approach for health apps based on successful experiences: integrative review. JMIR Mhealth Uhealth 2020;8(4):e14376 [FREE Full text] [doi: 10.2196/14376] [Medline: 32319965]

XSL•FO
RenderX

158. Saparamadu AADNS, Fernando P, Zeng P, Teo H, Goh A, Lee JMY, et al. User-centered design process of an mHealth app for health professionals: case study. JMIR Mhealth Uhealth 2021;9(3):e18079 [FREE Full text] [doi: 10.2196/18079] [Medline: 33769297]
159. de Paula D, Jühling D, Uebernickel F. Designing digital health innovation for stigma-affected health conditions: a systems-based view to managing chronic pelvic pain. Int J Hum Comput Interact 2025:1-15. [doi: 10.1080/10447318.2025.2485494]
160. Willis E, Friedel K, Heisten M, Pickett M, Bhowmick A. Communicating health literacy on prescription medications on social media: in-depth interviews with "patient influencers". J Med Internet Res 2023;25:e41867 [FREE Full text] [doi: 10.2196/41867] [Medline: 36912881]

Abbreviations

BfArM: German Federal Institute for Drugs and Medical Devices (Bundesinstitut für Arzneimittel und Medizinprodukte)
CG: control group
CSI-GE: Central Sensitization Inventory-German Version
CSQ-I: Client Satisfaction Questionnaire- Internet Version
DiGA: digital health application
DSM-IV: Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition
FSDS-DAO: Female Sexual Distress Scale-Desire/Arousal/Orgasm
FSFI-d: Female Sexual Function Index-German version
FSQ: Fear of Sexuality Questionnaire
G-MAUQ: German mHealth App Usability Questionnaire
ICD-11: International Classification of Diseases, 11th Revision
IG: intervention group
INEP-ON: Inventory for the balanced assessment of Negative Effects of Psychotherapy
PFB: Partnership Questionnaire
PROMIS-29: Patient-Reported Outcome Measurement Information System-29-Item Profile
PTSD: posttraumatic stress disorder
REDCap: Research Electronic Data Capture
T0: baseline
T1: after module 5 in intervention group/5 weeks after baseline in control group
T2: after module 8 in intervention group/8 weeks after baseline in control group
T3: 6-month follow-up after baseline
VPCQ: Vaginal Penetration Cognition Questionnaire

Edited by A Stone; submitted 26.Oct.2025; peer-reviewed by G Cabagno; comments to author 20.Nov.2025; revised version received 27.Dec.2025; accepted 31.Dec.2025; published 19.Feb.2026.

Please cite as:
Kronthaler SM, Kosman E, Tissen-Diabaté T, Mühle E, Weber- Schallauer L, Pross T, Karsten MM, Beier KM, Hatzler L
Adherence, Acceptability, and Sexual Health Outcomes of the Odeya App–Based Intervention for Sexual Distress in Women With Endometriosis: Randomized Controlled Mixed Methods Trial
J Med Internet Res 2026;28:e86042
URL: https://www.jmir.org/2026/1/e86042
doi:10.2196/86042
PMID:

©Selina Marie Kronthaler, Eden Kosman, Tatjana Tissen-Diabaté, Elena Mühle, Luzia Weber- Schallauer, Therese Pross, Maria Margarete Karsten, Klaus Michael Beier, Laura Hatzler. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Mental Health Professionals' Perceptions of Benefits and Disadvantages of Telehealth: International Mixed Methods Study

Madeline I Montoya[1], BA; Cary S Kogan[1], PhD; Brigitte Khoury[2], PhD; José A García-Pacheco[3], MA; Rebeca Robles[4], PhD; Tahilia J Rebello[5], PhD; Maya Kulygina[6], PhD; Geoffrey M Reed[7], PhD

[1]School of Psychology, University of Ottawa, Ottawa, ON, Canada

[2]Department of Psychiatry, American University of Beirut Medical Center, Beirut, Lebanon

[3]Centro de Investigación y Docencia Económicas, Mexico City, Mexico

[4]Center for Global Mental Health Research, Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz, Mexico City, Mexico

[5]Columbia University Vagelos College of Physicians and Surgeons, New York, NY, United States

[6]Training and Research Centre, Mental-Health Clinic No.1 Named After N.A. Alekseev, Moscow, Russian Federation

[7]Department of Psychiatry, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, United States

Corresponding Author:
Cary S Kogan, PhD
School of Psychology
University of Ottawa
Vanier Building
136 Jean-Jacques Lussier Pvt
Ottawa, ON, K1N 6N5
Canada
Phone: 1 6135625800 ext 4804
Email: ckogan@uottawa.ca

Abstract

Background: Telehealth has become an integral component of mental health care delivery worldwide. Understanding provider perceptions is essential to guiding its continued implementation.

Objective: This international study used quantitative and qualitative methodologies to examine and broaden our understanding of the benefits and concerns related to telehealth for mental health care.

Methods: An internet-based survey was conducted during the COVID-19 pandemic between November 11 and December 18, 2020, among mental health professionals, primarily psychiatrists and psychologists, registered with the World Health Organization's Global Clinical Practice Network. Clinicians completed the survey in 1 of 6 languages (Chinese, English, French, Japanese, Russian, or Spanish). Descriptive statistics and logistic regressions were used to analyze quantitative survey data on concerns and implementation of telehealth. Responses to an open-ended question about providers' perspectives on the benefits of telehealth were analyzed qualitatively.

Results: In total, 847 participants completed the telehealth section of the survey, and 496 provided a response to the open-ended question. Quantitative data on telehealth use and concerns revealed that clinicians' primary concerns focused on technical issues and clinical effectiveness relative to in-person services, specifically, loss of clinical information (eg, nonverbal behavior) and challenges with establishing a therapeutic alliance. Findings varied by profession, World Health Organization region, and telehealth training and experience. Qualitative data examining benefits fell into 3 major areas: accessibility and reach of mental health services, efficiency and flexibility for clinicians, and enhancement of clinical processes and outcomes. Taken together, findings revealed a trade-off between telehealth benefits and disadvantages.

Conclusions: From the perspective of mental health professionals, telehealth practice comes with key challenges and valuable benefits. Findings offer important considerations for the implementation of telehealth systems, including the importance of training and education and balancing trade-offs to optimize care.

(J Med Internet Res 2026;28:e75905) doi:10.2196/75905

KEYWORDS

global mental health; telehealth; mental health professionals; COVID-19; implementation science

XSL•FO
RenderX

Introduction

Following the onset of the COVID-19 pandemic and subsequent disruption of mental health services, clinicians and health care systems around the world experienced an abrupt shift to remote delivery of mental health services [1,2]. The term telehealth will be used, which is defined by the World Health Organization (WHO) as "the delivery of health care services, where patients and providers are separated by distance" [3]. Telehealth encompasses the use of information communication technology for the diagnosis and treatment of diseases and injuries, research, and continuing education of health professionals. This broad definition will include services provided by all health professionals working in mental health (eg, psychiatry and psychology) across various modalities, both synchronous (eg, videoconferencing or telephone) and asynchronous (eg, SMS text message or email).

Telehealth became central for continuity of mental health care throughout the pandemic and as a solution to meeting increased demand [4,5]. Governments and systems launched digital platforms [6,7] and eased regulatory, policy, and billing restrictions [8], while professional bodies produced guidelines to assist clinicians [9,10]. These combined efforts led to a substantial global surge in telehealth use for mental health care, which has remained robust in the postpandemic era [11,12].

Numerous studies conducted prior to and during the COVID-19 pandemic have shown favorable perceptions about telehealth including high client satisfaction and provider acceptability [1,13]. The evidence has demonstrated the feasibility and effectiveness of telehealth compared to in-person services for the assessment and treatment of various mental disorders and patient populations [14-16]. Telehealth can improve access to care in underserved areas and for individuals who typically experience barriers to accessing in-person services [1]. Telehealth can also be convenient (eg, reducing burdensome travel [17], cost-effective, and time-efficient for patients and providers [18]). Telepsychiatry research suggests additional advantages such as reduced psychiatric admissions [19,20] and greater scheduling flexibility [21]. However, the literature on the benefits of telehealth is fragmented, lacking a cohesive perspective across mental health providers and geographic regions.

Despite largely positive accounts of the potential of telehealth, numerous challenges still exist. Research conducted prior to and during the COVID-19 pandemic found that mental health professionals from various countries remain concerned with clinical elements of telehealth such as quality or effectiveness of care compared to in-person services [2,18] and assessing patients remotely. Technological challenges remain a barrier to telehealth use, particularly technical issues (eg, poor connectivity) and client access to technology [1,22,23]. Finally, although professional guidelines and easing of restrictions addressed many ethical, regulatory, and legal barriers to telehealth, some studies show that providers during the pandemic still reported concerns about patient privacy and confidentiality [2]. Therefore, it is important to comprehensively examine which clinical, technological, and ethical concerns are most challenging for mental health providers. For example, the specific sources that contribute to providers' perceptions of telehealth as less effective relative to in-person services remain unclear. Several studies suggest that mental health professionals report difficulty establishing a therapeutic alliance with their clients via telehealth [19,24], although findings are mixed [25]. Others report difficulty with interpreting patients' nonverbal communication [26]. Still other research suggests that difficulties with therapeutic skills (eg, showing empathy) is also a factor [27].

The prevailing view is that telehealth is a largely effective and more equitable means of service provision that is expected to remain a significant component of service provision. However, it may not be appropriate or effective for everyone. More evidence is needed to understand if telehealth is feasible and effective for certain patient populations and types of services [1,15]. Therefore, mental health professionals must assess the advantages and disadvantages of offering telehealth services in their particular contexts with specific patients. Moreover, given that clinicians are considered gatekeepers to implementation and sustainability of telehealth for mental health care [23], understanding provider perceptions is key to informing continued use.

The COVID-19 pandemic created a unique opportunity to assess the perceptions of telehealth providers not only from numerous clinical settings and geographical locations but also from professionals with greater generalizability than the subset of participants in prepandemic telehealth studies who may have been influenced by selection bias (ie, interest in telehealth) and existing infrastructure for virtual care. Therefore, research conducted during the pandemic contains informative perspectives from both groups of mental health providers: those forced to transition to telehealth and those familiar with and already implementing telehealth.

This mixed methods study used distinct but complementary qualitative and quantitative research questions to obtain more comprehensive information on the experiences, concerns, and benefits of telehealth from the perspective of international mental health providers during the COVID-19 pandemic [28,29]. Participants were drawn from the WHO's Global Clinical Practice Network (GCPN), a multilingual and international network of psychologists, psychiatrists, and other mental health professionals from 160 countries representing all WHO global regions [30]. It was hypothesized that the perspectives of mental health professionals would mirror previous findings on the advantages and disadvantages of telehealth as well as illustrate differential experiences across contexts (eg, world region) and professional demographics (eg, discipline, telehealth training, and experience [1,2]). Telehealth research often examines psychologists and psychiatrists either separately (ie, telepsychiatry) or treats them as a single group despite meaningful differences in the scope of practice (eg, medication management and physical health assessments). This study examines differences between the 2 professions, as discipline was hypothesized to shape telehealth use and perceptions. The term clinician is used to refer collectively to psychologists, psychiatrists, and other mental health professionals registered with the GCPN.

XSL•FO
RenderX

Methods

Study Design, Participants, and Procedure

This study was part of the "Longitudinal COVID-19 Survey of Mental Health Professionals Project," a 3-wave longitudinal internet-based survey study examining the effects of the COVID-19 pandemic on global mental health professionals' practice including telehealth [2] and occupational well-being [31]. This study focuses on the survey questions related to telehealth included in the second wave of data collection, which occurred between November 11 and December 18, 2020 (for a total data collection period of 5 weeks). A convergent mixed methods design was used, in which quantitative and qualitative data were collected in parallel (ie, in the same survey), analyzed separately with equal priority, and then integrated using a joint display of the 2 forms of data to provide additional insights [32]. This study used the GRAMMS (Good Reporting of a Mixed Methods Study) criteria [33] to guide reporting (Multimedia Appendix 1).

The Longitudinal COVID-19 Survey of Mental Health Professionals Project procedures and recruitment for the telehealth section of the survey have been described elsewhere in more detail [2]. In brief, participants were recruited from the GCPN. To be eligible for registration with the GCPN, providers must have completed their training and be qualified to provide mental health services in their countries. Initially, GCPN members registered because of interest in participating in field studies related to the *International Classification of Diseases, 11th Revision* (*ICD-11*) being conducted by the WHO [30].

The online survey was administered using Qualtrics in Chinese, English, French, Japanese, Russian, and Spanish. Content was initially developed in English and then translated by fluent experts affiliated with the GCPN's International Advisory Group.

GCPN members proficient in the corresponding survey languages were invited to participate via email containing an individualized survey. Upon accessing the link, participants were presented with a description of the study. Reminder emails were sent 7 and 14 days after the initial email. Participants who indicated that they had provided direct clinical services in the past 2 weeks and had used telehealth for either new or continuing patients were directed to the telehealth survey questions.

Ethical Considerations

The study was approved by the institutional review board at Columbia University/New York State Psychiatric Institute (registration: #6886) and the University of Ottawa (registration: H-06-20-5973). Participants, who were practicing mental health professionals and members of the GCPN, were asked to provide consent for participation by selecting "Yes" to proceed with the online survey after reviewing a description of the study. Participants were not compensated for participation. Study data were deidentified to maintain participant privacy and confidentiality.

Survey

Quantitative data were collected from closed-ended items about telehealth concerns and implementation, and qualitative data were collected from an open-ended question about clinicians' perspectives on the benefits of telehealth. Telehealth survey items were developed from a review of the existing literature at the time (including data from the first-wave telehealth survey published in Montoya et al [2]), together with group discussion and expert review by the authors and the GCPN's International Advisory Group (Multimedia Appendix 2).

Upon starting the survey, participants were informed that they would be asked about their recent work activities and to have their work calendars accessible when completing the survey. Participants were about years of telehealth experience, if they had received training on technological, ethical, legal, and clinical aspects of telehealth service delivery or no training at all, and current (ie, past 2 weeks) telehealth use. Demographic and telehealth use survey items included closed-ended and multiple response (or check-all-that-apply) questions. Likert-type questions about perceived effectiveness of telehealth for specific age groups and treatment modality (eg, individuals and couples) were also included.

To identify specific challenges, 4 separate checklists were presented, which focused on clinical, technical, ethical, legal, and professional, and finally, administrative elements of providing telehealth services. If participants endorsed concerns about "clinical effectiveness relative to in-person services" in the checklist of clinical elements, they were directed to an additional checklist containing more detailed concerns related to clinical effectiveness (eg, building and maintaining a therapeutic alliance and showing empathy). These options for specific concerns were based on the telehealth literature [13] as well as basic clinical skills that are considered important for therapeutic change [34,35]. Participants could select multiple items for each checklist and use an open-text field to specify an "other" response.

The following open-ended question was included to capture participants' perspectives on the benefits of telehealth: "In addition to making it possible to consult with patients remotely, have there been any benefits of using telehealth services?"

Quantitative Data Analysis

Of a total of 15,179 GCPN members who were invited to participate in the larger study, 2124 (13.9%) agreed to participate in the survey, and 2088 (98.3%) of those participants provided sufficient data for analysis (ie, <90% missing survey data). Of those participants, 847 completed the telehealth section. Descriptive and chi-square statistics were used to summarize and compare quantitative survey data. Logistic regressions were used to examine predictors of clinicians' common concerns. Statistical calculations were performed using SPSS (version 29; IBM Corp).

Qualitative Data Analysis

The purpose of the qualitative data analysis was to examine and summarize the content of clinicians' responses. The analysis was guided by recommendations from Miles et al [29] and

XSL•FO
RenderX

consisted of first- and second-cycle coding methods. First cycle coding is a way of assigning initial descriptive codes to the data, whereas second cycle coding is the reorganization of these first cycle codes into meaningful themes, categories, or concepts [36].

Of the 847 clinicians who completed the telehealth section survey, 496 (58.6%) provided free-text responses to the open-ended question. Responses ranged from 1-word phrases to a paragraph; however, most were a few sentences in length. In general, characteristics of those who completed the open-ended question were similar to those who did not. Significant differences were found for telehealth training, with a greater proportion of participants with telehealth training answering the open-ended question (χ^2_1=12.1; P<.001). Moreover, those who responded to the open-ended question were older (t_{845}=4.6; P<.001) and had slightly more years of clinical experience (t_{845}=2.8; P=.003). All 496 responses were imported to and analyzed using NVivo (version 14; Lumivero). Two of the authors (MIM and CSK) coded the data and consulted with a researcher with experience in approaches relevant to this study. Coders were either actively or passively fluent in the 3 most common survey languages (English, Spanish, and French); thus, 421 (84.5%) responses were kept in their original language until the end of data analysis. MIM, the primary coder, was fluent in the former 3 languages. CSK was fluent in English and French and could read and understand Spanish. Chinese, Japanese, and Russian responses were translated to English prior to coding using DeepL Translator, a machine translation tool known for its accuracy and natural, high-quality translations. DeepL Translator is owned by DeepL SE, a German-based company that uses artificial intelligence to offer various translation supports.

Following an inductive approach, MIM and CSK used initial coding (an open-ended, data-driven coding process [36]) and in vivo coding (a coding technique that uses participants' exact words or phrases to extract ideas [36]) to independently code 150 responses (33% of the data). This initial sample was selected by sorting the full sample of responses by WHO region and coding a proportional number of responses based on WHO region and language. Responses were intentionally skipped as needed to ensure proportional representation. For example, 10% (15/150) of responses were from Western Pacific-Asia; thus, 15 responses from this region were coded as part of the initial sample of 150 responses. Next, coders shared their initial codes to collaboratively develop a preliminary codebook.

Discrepancies in code definitions, labels, and interpretations were resolved through discussion until consensus was reached. Using this initial codebook, a process of intercoder reliability was conducted. MIM and CSK independently coded a sample of 100 (20%) responses different from the initial sample that was used to develop initial codes. κ statistics (overall κ=0.65) and percentage agreement (ranging from 65% to 100%) were used to identify codes with lower agreement. The coders discussed inconsistencies until consensus was reached. Codes were merged, dropped, and refined until a final codebook was generated. MIM then coded all 496 responses.

Two second-cycle coding methods were used: pattern and axial coding. MIM conducted a process of pattern coding to identify emergent themes, configurations, or explanations in the data [29]. Axial coding was used to distinguish properties and dimensions of emergent categories [36] such as benefits specific to patients versus clinicians. Experiential differences and similarities across the WHO region and profession were also examined. MIM and CSK discussed and developed emergent themes and overall conceptualization. The relative frequency with which codes were applied was used to distinguish between topics that arose more often than others to better characterize the diversity of perceptions reported by clinicians [37].

Efforts were taken to ensure the quality of the data analysis. Methodological and analytic documentation (eg, memo writing) was used to keep track of the coding process and related decisions [36]. During meetings, MIM and CSK attempted to maintain reflexivity and note sources of bias during analysis (eg, the impact of one's own sociocultural background on data interpretation). Furthermore, coders were blinded to all demographic information when analyzing codable text except for language and profession. The rationale was that access to this information could help contextualize responses (eg, assessment could include physical examination in psychiatry). Coders searched for disconfirming evidence (ie, counterexamples to recurring patterns), and that no substantively new themes or patterns emerged in the final stages of coding the data sample [29].

Integrating Quantitative and Qualitative Data

After separate analyses, frequently endorsed concerns (quantitative data) and identified benefits (qualitative data) were brought together in a table for elaboration [32]. MIM and CSK met to conduct a higher-level interpretation of findings to supplement and enhance understanding of the advantages and disadvantages of telehealth for mental health care [28].

Results

Quantitative: Telehealth Use and Concerns

Overview

Clinician demographic and professional characteristics are provided in Table 1. Participants represented all WHO regions and 73 countries. Most worked in mental health settings (754/847, 89%), followed by substance abuse specialty programs (200/847, 23.6%), university or other educational institutions (182/847, 21.5%), other general medical settings (148/847, 17.5%), primary care settings (67/847, 7.9%), and finally, hospital emergency departments (15/847, 1.8%). Approximately half of the participants (446/847, 52.7%) reported working in multiple settings. Most participants (776/847, 91.6%) reported that they had provided mental health services (such as diagnostic assessment, psychological assessment, or psychotherapy) during the past 2 weeks. Additionally, 84.8% (263/310) of psychiatrists had also provided psychopharmacotherapy, and 29% (90/310) had provided other medical services (including general medical services and management of comorbid medical conditions).

XSL•FO
RenderX

Participants reported using various modalities for direct clinical services (assessment, treatment, or monitoring of patients) including in person (607/847, 71.7%), videoconferencing (632/847, 74.6%), telephone (654/847, 77.2%), chat (353/847, 41.7%), and email (359/847, 42.4%). Psychiatrists were more likely to see patients in person compared to psychologists (χ^2_1=28.5; *P*<.001). Table 2 provides the number of hours spent and patients seen using telehealth for psychiatrists and psychologists in the past 2 weeks.

Table 3 shows providers' perceived effectiveness of telehealth by treatment modality (individual, group, families, and couples) and patient age groups (older adults, adults, adolescents, and children).

XSL•FO
RenderX

Table 1. Clinician demographics and occupational characteristics.

	Sample of telehealth participants (N=847)
Gender, n (%)	
Men	372 (43.9)
Women	472 (55.7)
Other self-identified	3 (0.4)
Language, n (%)	
English	454 (53.6)
Spanish	174 (20.5)
Japanese	67 (7.9)
Chinese	29 (3.4)
French	53 (6.3)
Russian	70 (8.3)
WHO[a] region[b], n (%)	
Africa	22 (2.6)
Americas-South	135 (15.9)
Americas-North	156 (18.4)
Eastern Mediterranean	13 (1.5)
Europe	343 (40.5)
South-East Asia	48 (5.7)
Western Pacific-Asia	106 (12.5)
Western Pacific-Oceania	24 (2.8)
Country income level, n (%)	
Low	4 (0.5)
Lower-middle	87 (10.3)
Upper-middle	219 (25.9)
High	537 (63.4)
Profession, n (%)	
Psychiatry	310 (36.6)
Psychology	384 (45.3)
Other[c]	153 (18.1)
Age (years)	
Mean (SD)	52.3 (11.3)
Range	25-89
Median (IQR)	52 (43-60)
Years of experience	
Mean (SD)	20.9 (10.2)
Range	0-57
Median (IQR)	20 (13-29)
Telehealth training, n (%)	
Yes	475 (56.1)
No training	372 (43.9)
Telehealth experience, n (%)	

XSL•FO
RenderX

	Sample of telehealth participants (N=847)
Less than 1 year	419 (49.5)
More than 1 year	428 (50.5)

[a]WHO: World Health Organization.

[b]Participants were predominantly from the following countries in each WHO region—Africa: Nigeria, South Africa; Americas-South: Mexico, Argentina, Brazil, Columbia, Peru; Americas-North: the United States, Canada; Eastern Mediterranean: United Arab Emirates, Egypt; Europe: Spain, Russia, Germany, the United Kingdom, France, Denmark, Switzerland; South-East Asia; India; Western Pacific-Asia: Japan, China; and Western Pacific-Oceania: Australia, New Zealand.

[c]Other professionals included predominantly social work, nursing, counseling, occupational therapy, and sex therapy.

Table 2. Number of patients seen and hours spent using telehealth in the past 2 weeks.

	Values, n	Psychologist, mean (SD)	Values, n	Psychiatrist, mean (SD)
Telephone	263		245	
Patients		7.79 (13.34)		8.48 (9.27)
Hours		7.40[a] (10.06)		5.79[a] (7.09)
Videoconferencing	332		201	
Patients		10.97 (13.36)		9.83 (11.77)
Hours		14.96[a] (17.12)		10.66[a] (12.29)
Instant messaging	126		141	
Patients		5.82 (10.08)		7.77 (10.26)
Hours		6.53 (14.41)		5.35 (10.04)
Email	137		114	
Patients		3.58 (5.21)		4.81 (4.88)
Hours		2.52 (3.59)		2.31 (2.76)

[a]Significantly different at $P<.05$.

Table 3. Clinicians who reported that the telehealth services they provided for specific patient populations were about the same or more effective than in-person by treatment modality and patient age groups.

	Value, n/N (%)
Patient age groups	
Adults (65+ years)	155/406 (38.2)
Adults (19-64 years)	468/751 (62.3)
Adolescents (13-18 years)	192/359 (53.5)
Children (0-12 years)	42/188 (22.3)
Treatment modality	
Groups	68/200 (34)
Families	88/271 (32.5)
Couples	106/270 (39.3)
Individual assessment and treatment	425/774 (54.9)

Table 4 lists the most common concerns (endorsed by 40% or more) regarding clinical, technical, ethical, legal, professional, and administrative aspects of telehealth provision. Psychologists were significantly more concerned about a lack of private space (χ^2_1=7.6; P=.006) and technical issues (χ^2_1=4.8; P=.03) when compared to psychiatrists. Conversely, psychiatrists were significantly more concerned with remote assessment of patients (χ^2_1=32.5; P<.001), assessing high-risk patients and managing emergencies (χ^2_1=12.4; P<.001), showing empathy (χ^2_1=5.6; P=.02), and licensing or credentialing requirements (χ^2_1=36.1; P<.001) when compared to psychologists.

XSL•FO
RenderX

Table 4. List of most frequent telehealth concerns endorsed by at least 40% of psychiatrists, psychologists, and/or other professionals (mainly social work, nursing, counseling, occupational therapy, and sex therapy).

	Psychiatrist, n (%)	Psychologist, n (%)	Other, n (%)	Total, n (%)
Clinical concerns				
Clinical effectiveness relative to in-person services	196 (68.5)	224 (60.7)	86 (57)	506 (62.8)
Loss of clinical information	158 (81.4)	175 (78.8)	63 (75)	396 (79.2)
Building and maintaining a therapeutic alliance	141 (72.7)	150 (67.6)	60 (71.4)	351 (70.2)
Showing empathy	103 (53.1)	92 (41.4)	42 (50)	237 (47.4)
Implementing therapeutic interventions	68 (35.1)	97 (43.7)	43 (51.2)	208 (41.6)
Assessment of high-risk patients and managing emergencies remotely	194 (67.8)	189 (51.2)	77 (51)	460 (57.1)
Lack of private space for patients during telehealth sessions	128 (44.8)	199 (53.9)	60 (39.7)	387 (48)
Remote assessment of patients (eg, diagnostic assessments, self-report measures, lack of opportunity for physical examinations, and medication blood level monitoring)	171 (59.8)	129 (35)	58 (38.4)	358 (44.4)
Decrease in patient engagement compared to in-person services	119 (41.6)	151 (40.9)	70 (46.4)	340 (42.2)
Technological concerns				
Technical issues (eg, audio or video quality and stability of internet connection)	211 (71)	294 (78.4)	116 (78.9)	621 (75.8)
Patients' access to equipment	177 (59.6)	230 (61.3)	93 (63.3)	500 (61.1)
Patient familiarity with connecting	137 (46.1)	166 (44.3)	77 (52.4)	380 (46.4)
Ethical, legal, and professional concerns				
Patient consent, privacy, security, and confidentiality	112 (42.6)	128 (38.6)	61 (42.4)	301 (40.7)
Licensure or credentialing requirements	116 (44.1)	70 (21.1)	34 (23.6)	220 (29.8)
Administrative concerns				
Disruption of routine and workflow	121 (40.7)	131 (34.9)	58 (39.5)	310 (37.9)

Logistic Regressions

When chi-square tests showed associations between at least 2 demographic variables and the most common concerns, binomial logistic regressions were used to determine whether demographic and professional variables affected the likelihood that clinicians endorsed a common concern. Predictors included age, years of clinical experience, profession (psychiatrists, psychologists, and other professionals), telehealth experience (<1 year vs >1 year), telehealth training received (no vs yes), and WHO region. Age and years of clinical experience did not significantly contribute to overall model fit and were thus excluded from the final models (Table 5).

XSL•FO
RenderX

Table 5. Logistic regression of predictors of common telehealth concerns.

Predictors	OR[a] (95% CI)	*P* value
Clinical effectiveness relative to in-person services		
Profession[b]		
Psychiatrist	1.336 (0.847-2.106)	.21
Psychologist	1.045 (0.697-1.566)	.83
WHO[c] region[d]		
Africa	1.671 (0.631-4.428)	.30
Americas-South	1.301 (0.778-2.175)	.32
Eastern Mediterranean	3.691 (0.771-17.662)	.10
Europe	1.314 (0.853-2.024)	.22
South-East Asia	1.394 (0.695-2.795)	.35
Western Pacific-Asia	1.408 (0.779-2.545)	.26
Western Pacific-Oceania	1.179 (0.482-2.886)	.72
Telehealth training[e] (none)	1.378 (1.011-1.878)	.04
Telehealth experience[f] (<1 year)	1.348 (1.004-1.812)	.047
Remote assessment of patients		
Profession[b]		
Psychiatrist	2.079 (1.319-3.276)	.002
Psychologist	0.901 (0.595-1.363)	.62
WHO region[d]		
Africa	2.077 (0.791-5.453)	.14
Americas-South	0.920 (0.545-1.551)	.75
Eastern Mediterranean	1.167 (0.354-3.855)	.80
Europe	0.753 (0.482-1.175)	.21
South-East Asia	2.047 (1.016-4.124)	.045
Western Pacific-Asia	1.877 (1.036-3.401)	.04
Western Pacific-Oceania	0.720 (0.284-1.823)	.49
Telehealth training[e] (none)	1.429 (1.050-1.944)	.02
Telehealth experience[f] (<1 year)	1.010 (0.751-1.358)	.95
Lack of private space for patients during telehealth sessions		
Profession[b]		
Psychiatrist	1.447 (0.922-2.271)	.11
Psychologist	1.837 (1.224-2.759)	.003
WHO region[d]		
Africa	0.758 (0.301-1.908)	.56
Americas-South	1.372 (0.822-2.290)	.23
Eastern Mediterranean	0.500 (0.152-1.641)	.25
Europe	0.658 (0.428-1.012)	.06
South-East Asia	1.048 (0.532-2.064)	.89
Western Pacific-Asia	0.521 (0.292-0.930)	.03
Western Pacific-Oceania	0.819 (0.337-1.990)	.66

XSL•FO
RenderX

Predictors	OR[a] (95% CI)	*P* value
Telehealth training[e] (none)	1.242 (0.920-1.677)	.16
Telehealth experience[f] (<1 year)	1.043 (0.783-1.389)	.77
Technical difficulties		
Profession[b]		
Psychiatrist	0.923 (0.547-1.558)	.77
Psychologist	1.042 (0.635-1.707)	.87
WHO region[d]		
Africa	0.739 (0.242-2.262)	.60
Americas-South	0.875 (0.461-1.659)	.68
Eastern Mediterranean	0.423 (0.115-1.564)	.20
Europe	0.695 (0.403-1.197)	.19
South-East Asia	0.674 (0.294-1.545)	.35
Western Pacific-Asia	0.362 (0.189-0.696)	.002
Western Pacific-Oceania	0.478 (0.181-1.265)	.14
Telehealth training[e] (none)	0.920 (0.653-1.297)	.64
Telehealth experience[f] (<1 year)	1.556 (1.115-2.173)	.009
Patient consent, privacy, security, and confidentiality		
Profession		
Psychiatrist	0.720 (0.399-1.301)	.28
Psychologist	0.742 (0.430-1.282)	.29
WHO region		
Africa	1.504 (0.446-5.069)	.51
Americas-South	0.199 (0.095-0.415)	<.001
Eastern Mediterranean	0.926 (0.240-3.577)	.91
Europe	0.700 (0.400-1.226)	.21
South-East Asia	3.055 (1.095-8.521)	.03
Western Pacific-Asia	0.681 (0.327-1.417)	.30
Western Pacific-Oceania	0.535 (0.165-1.739)	.30
Telehealth training (none)	0.802 (0.548-1.175)	.26
Telehealth experience (<1 year)	0.745 (0.511-1.085)	.13

[a]OR: odds ratio.

[b]Other professionals as reference.

[c]WHO: World Health Organization.

[d]Americas-North as reference.

[e]Received telehealth training as reference.

[f]More than 1 year of telehealth experience as reference.

For clinical concerns of telehealth, the model investigating endorsement of clinical effectiveness relative to in-person services was statistically significant (χ^2_{11}=21.4; Nagelkerke R^2=0.04; *P*=.03). A lack of telehealth training and less than 1 year of telehealth experience were significant predictors of the likelihood of clinicians reporting clinical effectiveness of telehealth as a concern. The model predicting concern with remote assessment of patients was statistically significant (χ^2_{11}=71.0; Nagelkerke R^2=0.14; *P*<.001). The model suggested that psychiatrists, clinicians from South-East Asia and Western-Pacific Asia (compared to clinicians from Americas-North), and clinicians with no telehealth training were more likely to report remote assessment as a telehealth concern. The model predicting concern with lack of private space for

XSL•FO
RenderX

patients was also statistically significant (χ^2_{11}=30.0; Nagelkerke R^2=0.05; P=.002) with psychologists, clinicians from Americas-North (compared to clinicians from Western Pacific-Asia) being more likely to endorse this concern.

Regarding technological concerns, the model predicting endorsement of technical difficulties (eg, audio or video quality and stability of internet connection) was statistically significant (χ^2_{11}=30.0; Nagelkerke R^2=0.05; P=.007). Clinicians with less than 1 year of telehealth experience were more likely to report technical difficulty as a concern. Conversely, compared to participants living in the Americas-North, clinicians in Western-Pacific Asia were less likely to endorse technical difficulties. Finally, the model predicting concern with patient consent, privacy, security, and confidentiality was statistically significant (χ^2_{11}=48.5; Nagelkerke R^2=0.12; P<.001). The WHO region was the only significant predictor. Compared to clinicians living in Americas-North, clinicians in Americas-South were less likely to endorse concern with patient consent, privacy, security, and confidentiality, whereas clinicians in South-East Asia were conversely more likely to report this concern.

Qualitative Results: Benefits of Telehealth

Overview

Three key themes related to the benefits of telehealth were identified: accessibility and reach of mental health services, efficiency and flexibility for clinicians, and enhancement of clinical process and outcomes. In general, clinicians' responses highlighted multiple benefits of telehealth. Codes related to flexibility and convenience, increased access, and time savings were most commonly and consistently cited across the WHO region and profession. However, the context and conditions in which respondents described these codes differed and thus gave rise to several subthemes.

Themes and their respective subthemes are presented in the section below, with representative quotes. Representative responses written in any of the 5 other languages other than English are shown in their original form alongside an English translation. Where descriptive differences by WHO region or profession were observed, these are discussed within corresponding themes. Salient or evocative codes that did not fit the main themes are discussed further below.

Theme 1: Accessibility and Reach of Mental Health Services

This theme was developed primarily from a review of responses where codes related to flexibility and convenience for patients, time savings, and greater access to services were applied. The following 3 subthemes were included:

Reduces Barriers to Care

Clinician responses suggested that telehealth reduces burdens and barriers associated with accessing and receiving in-person mental health services. Many providers indicated that telehealth improved access by reducing geographical and logistical barriers. For example: "Reaching easily patients that live far or in islands where there are no mental health services" (Psychiatrist, Greece).

Others described a reduction in costs and stress that come with traveling to in-person appointments for patients: "Convenience for families not needing to drive/find parking/childcare for siblings etc." (Psychologist, Australia).

Several clinicians described how telehealth increased access for specific populations that typically experience challenges with attending in-person services including older people, individuals with physical and mental comorbidities, and those with mobility difficulties: "It has meant much greater access for people with disabilities and people with more full schedules, which is fantastic" (Psychologist, the United States).

Expands Reach and Cost-Effectiveness of Mental Health Service Delivery

Many clinicians described being able to contribute to the creation of greater access through telehealth. This was indicated on both a provider and a systemic level. For instance: "It has meant I can work with people throughout the state(s) I am licensed in, which creates much more access" (Psychologist, the United States).

Some clinicians reported being able to "see" more patients and having "mas cobertura" ("greater coverage"). Whereas others, such as a psychiatrist in India, cited greater "availability of services in distress, emergency and follow up."

Some also indicated a reduction in overhead expenses. For example, a psychologist in China reported feeling "... less financially burdened for office space." Similarly, a psychologist in Mexico reported, "disminución en costos de alquiler de consultorio" ("Decrease in office rental costs").

Safety of Services Provided

This subtheme came from a code specific to the context of the COVID-19 pandemic, which focused on reduced risk to health. For instance, a psychiatrist in India noted that telehealth, "... reduces risk of exposure to COVID-19 both for the doctor and patient and his/her family."

Theme 2: Efficiency and Flexibility for Clinicians

This theme was developed from codes related to flexibility and convenience for mental health professionals, time savings, provider self-care, and access to other professionals. A review of these texts illustrated the benefits of telehealth that were provider-specific in either a personal or professional capacity. The following 2 subthemes emerged:

Enables Clinicians to Use Their Work Time More Effectively and Flexibly

Many clinicians discussed better time management as a benefit of telehealth. For example, a psychiatrist in Switzerland indicated: "Organisation plus facile de l'agenda." ("Easier to manage my schedule."). Moreover, clinicians described how this flexibility allowed them more time to pursue personal, domestic, or professional activities:

> *He podido aprovechar mejor los espacios intermedios, prácticamente no tengo horarios muertos. (I have been able to take better advantage*

XSL•FO
RenderX

of the time between appointments. I have practically no dead time.) [Psychologist, Mexico]

I can work from home and be available to my partner in a more immediate way, and I can utilize the resources in home more flexibly. I save 6-8 hours a week in commuting time. [Psychologist, the United States]

Seguir trabajando en otros ámbitos. (Continue working in other areas.) [Psychologist, Argentina]

Promotes Provider's Well-Being and Access to Other Professionals

Some providers reported that working from home promoted their capacity for self-care and well-being. For example:

[M]oins d'énergie dépensée en préparatifs et déplacements = moins fatigué. ([L]ess energy spent on preparation and travel = less tired.) [Psychologist, Canada]

Me siento más relajada al poder acceder fácilmente a tomar algún alimento o descansar entre consulta y consulta en un espacio diferente, otro sillón, la cama, el jardín, etc. (I feel more relaxed as I can easily get something to eat or rest between sessions in a different space, another couch, the bed, the garden etc.) [Psychologist, Mexico]

Some clinicians also reported easier access to professional development opportunities (eg, "docencias (teachings)" and "capacitaciones (trainings)"), colleagues, and clinical meetings:

Attending multiagency meetings, which are sometimes many miles from the previous appointment, are now attendable via MS Teams, just a couple of clicks away. [Psychiatrist, the United Kingdom]

Les échanges avec des collègues. (Discussions with colleagues) [Psychiatrist, Senegal]

Theme 3: Enhancement of Clinical Processes and Outcomes

This theme was developed from a review of text where codes related to clinical activities and patient experience (eg, comfort, attendance, and anonymity) were applied. The following 2 subthemes were included:

Improves Clinical and Therapeutic Activities

Some clinicians reported that various clinical processes were improved with telehealth. For example, a psychologist in the United States described "easier access to internet search while in therapy to support therapy." This also included progress tracking, maintaining the alliance, psychoeducation, and follow-ups. Other clinicians described the benefits of having access to patients' homes and family life along with greater family or caregiver involvement:

Useful to see in client's homes and the conditions in which they live. [Social worker, Canada]

Consultation avec des parents séparés qui ne seraient pas venus ensemble en presence. (Consultations with separated parents who would have never attended together in person.) [Psychiatrist, France]

Others reported treatment benefits for specific diagnoses. For example:

Interventions with patients who have Obsessive Compulsive Disorder are far more effective, because patient is in their own home and can practice behavioural exposure during the session. [Psychologist, Canada]

(Easy for autistic patients to understand information and share PC screen.) [Psychiatrist, Japan]

Improves Patient Comfort, Attendance, and Engagement

Several clinicians from nearly all WHO regions indicated that telehealth improved patient attendance. For example, a psychologist in New Zealand noted, "better adherence and fewer did not attend!" Some differences among WHO regions were noted. Many of the responses around patient comfort came from providers in the Americas-North region. Clinicians discussed patient comfort in various ways:

Some patients prefer it, and are more comfortable/relaxed their own space than they would be in a therapy room—so they engage better. [Psychologist, Australia]

Some clients have felt more comfortable disclosing vulnerable material. [Psychologist, the United States]

Nearly all of the responses related to the benefit of patient privacy and anonymity were provided by clinicians in the South-East Asia and Western Pacific-Asia regions and described as an advantage of telehealth:

Some subjects did express that they are happy to have online telehealth services as they do not need to feel noticed by their peers and general public visiting a mental health professional. [Counselor, India]

People can be anonymous for those who wants to keep their identity anonymous. [Psychologist, Philippines]

Other Notable Codes

Disadvantages or No Additional Benefits

Although the framing of the short-answer questions was focused on the benefits of telehealth, several providers (58/496, 11.9% responses) indicated that there were no additional benefits aside from remote consultations or exclusively reported disadvantages of telehealth. The vast majority of responses to which this code was applied came from clinicians in Europe, Americas-South, and Western Pacific-Asia. Most of these responses were brief (eg, "no" and "none"), while only a few responses were more descriptive:

No. The Covid experience has confirmed to me that I don't think it's a useful longterm alternative to in person therapy. [Psychologist, New Zealand]

The only benefit is that the service can be provided despite COVID, but I believe we lose out on many of

XSL•FO
RenderX

the small, but important nuances of in-person therapy when limited to on-line intervention. [Psychologist, Canada]

Positive Disruption of Mental Health Profession

A few clinicians, particularly providers in the Americas-North region, described a systemic level benefit, suggesting that the telehealth shift had challenged and advanced the mental health field and profession by altering service delivery, lifting licensing restrictions, and "kickstarting" new areas:

Offre aux patients un modèle de souplesse et de flexibilité; permet d'approfondir les enjeux autour des limites et des frontière.(Offers patients a model of flexibility; allows for the exploration of issues around limits and boundaries.) [Psychologist, Canada]

In South Africa, remote consulting was not permitted by the medical regulatory body. We hope that this will change in future, so that psychiatric services can be made accessible to underserved areas where there are no psychiatrists. This would benefit particularly poor and disadvantaged areas. This can also be a way of upskilling primary care providers in these areas. [Psychiatrist, South Africa]

Comparing Quantitative and Qualitative Results

A joint display of the 2 forms of data offered a complimentary perspective on the concerns and benefits of telehealth according to clinicians [32]. Table 6 presents these contradictory findings as thematic intersections.

XSL•FO
RenderX

Table 6. Integrating qualitative and quantitative findings from mental health professionals to generate complementary insights.

	Benefits of telehealth (qualitative)	Concerns with telehealth (quantitative)	Provider- and system-level considerations when evaluating the trade-off
Thematic intersection 1	• Improves clinical and therapeutic activities (eg, follow-ups, greater family involvement, more insight into patients' home and family life, easier use of online tools to support therapy, and diagnosis-specific treatment benefits; 52/496, 10.5% responses)	• Clinical effectiveness compared to in-person (506/806, 62.8%), especially, • Loss of clinical information (396/500, 79.2%) • Building and maintaining the therapeutic alliance (351/500, 70.2%) • Showing empathy (237/500, 47.4%) • Implementing therapeutic interventions (208/500, 41.6%)	• Patient characteristics including disorder or diagnosis, age group, and treatment modality (individual vs group or family or couple) • Provider characteristics including telehealth training and experience; telehealth strategies tailored to meet discipline-specific clinical needs • Evaluating the loss versus the gain of clinical information • Would telehealth enhance or impede therapeutic interventions? • Role of flexible blended models that incorporate telehealth and in-person
Thematic intersection 2	• Expands reach of mental health services (64/496, 12.9% responses) • Timely access to emergency services (6/496, 1.2% responses)	• Assessing high-risk patients and managing emergencies remotely (460/806, 57.1%) • Remote assessment of patients (358/806, 44.4%)	• Patient characteristics including reliable access to technology and internet; technological literacy • Provider characteristics including telehealth training and experience • Regional characteristics: is this an area or population with no access to mental health services? • Are there discipline-specific clinical needs such as infrastructure for physical examination? • Opportunity for telehealth models to work closely with providers from remote regions for in-person care and capacity building?
Thematic intersection 3	• Improves patient attendance (40/496, 8.1% responses) • Improves patient comfort (eg, more at ease in their own homes and more disclosures; 25/496, 5% responses)	• Decreased patient engagement compared to in-person (340/806, 42.2%) • Lack of private space for patients during telehealth sessions (387/806, 48%)	• Patient characteristics including disorder or diagnosis, age group, treatment modality, preferences, and living circumstances • Will the lack of privacy in the home environment be an issue? • Regional characteristics: could the privacy or anonymity of telehealth help reduce stigma challenges about mental illness and seeking care?

Discussion

Overview

This study used quantitative and qualitative methodologies to comprehensively examine and expand upon our understanding of the perspectives, benefits, and concerns related to telehealth service provision among a large, international, and multilingual sample of mental health professionals during the COVID-19 pandemic (November-December 2020). Participants were mainly psychologists and psychiatrists working in mental health settings and providing mental health services such as diagnostic

XSL•FO
RenderX

assessment or psychotherapy via telehealth. Most psychiatrists had also provided psychopharmacotherapy services. Half of participating clinicians had less than a year of telehealth experience, suggesting that they likely adopted telehealth because of the pandemic. A little over half of the participants had received some form of telehealth training. Participants were mostly using videoconferencing and/or telephone. This study offers a global snapshot of telehealth implementation from the perspective of clinicians with both previous telehealth experience and new adopters. Findings suggest several important conclusions that can inform recommendations for training and practice for telehealth in mental health care.

Clinicians Are Primarily Concerned About Clinical Effectiveness and Technological Challenges

Consistent with previous research, clinical effectiveness relative to in-person services was a top concern among clinicians [2,18,38]. However, this study further highlights that clinicians' primary concerns about clinical effectiveness were regarding the loss of clinical information (eg, nonverbal behavior) and challenges in building and maintaining a therapeutic alliance. Assessment of high-risk patients and managing emergencies remotely was another common clinical concern. These findings support previous research that indicates that major limitations of telehealth practice are reduced availability of clinical data to support assessment or diagnosis [26] and greater difficulty picking up on nonverbal cues [1]. Diminished sense of connection and therapeutic alliance [18,19] has also been raised as a notable challenge associated with telehealth practice, despite this not being fully reflected among findings in the telehealth efficacy literature [39].

With respect to technological concerns, technical issues (eg, audio or video quality and stability of internet connection) were the top concerns followed by patient-related factors such as patients' access to equipment and familiarity with connecting. The COVID-19 pandemic compelled clinicians to adapt to technology and the technical issues inherent to telehealth. Current best-practice guidelines suggest that proactively establishing backup means of communication with clients can help mitigate these challenges [10]. However, the pandemic has also exposed disparities in access to mental health care among vulnerable populations across low-income and middle-income countries (LMICs) as well as within high-income countries [6]. Hence, client-specific challenges such as technology access, reliability, and competency remain significant barriers to telehealth.

Overall, these top concerns are broadly consistent with the literature [1,18]. However, the current findings build on previous research by examining how demographic and professional characteristics influence the endorsement of common concerns and perceptions, offering further context and insight. For instance, a lack of telehealth training was associated with increased endorsement of common concerns such as clinical effectiveness of telehealth relative to in-person services and remote assessment of patients. Likewise, those with minimal experience using telehealth (ie, less than 1 year) were more likely to report technical difficulties and clinical effectiveness of telehealth relative to in-person services as a concern. These findings support the literature, suggesting that experience and training are related to provider perceptions of telehealth [2,13,40].

Telehealth Offers Valuable Benefits for Patients and Clinicians That Are Consistent Across Global Contexts

Qualitative data revealed numerous advantages to virtual mental health care. These benefits fell into 3 major areas: accessibility and reach of mental health services, efficiency and flexibility for clinicians, and enhancement of clinical process and outcomes. These main benefits broadly align with existing studies conducted across LMICs and higher-income countries [1,17,18,21,26,41]. However, this international sample extends current research by highlighting the significant advantages of telehealth cited by clinicians from nearly all WHO regions. Specifically, remote mental health care is convenient and flexible for patients because it reduces burdens and barriers to accessing care. Clinicians' responses illustrated the various economical, geographical, and logistical barriers that clients regularly encounter when accessing in-person mental health care. Moreover, telehealth improved patient access by enabling clinicians to see more patients and extend specialized services to a broader population. Thus, continued implementation of telehealth services alongside in-person care can help ensure equitable mental health access.

Based on participants' perceptions, telehealth also offers convenience and flexibility for mental health professionals by giving them greater control over their time and clinical outreach. Previous studies have spoken about the personal benefits of telehealth for mental health professionals including more efficient use of time [18] and improved work-life balance [22]. This study further illustrates how telehealth can support clinician efficiency (eg, no commuting), well-being and quality of life (eg, working in a relaxing home environment), and professional development. As such, telehealth offers valuable benefits for mental health professionals, particularly for clinicians at life stages where flexibility and work-life balance are important. This is likewise relevant to occupational stress and burnout, as the COVID-19 pandemic underscored the need to safeguard the capacity of the global mental health workforce [31]. Health care systems and clinicians should leverage the above benefits to promote provider efficiency and health, such as policies supporting hybrid work models, while maintaining quality care.

Increased safety from COVID-19 and improved patient attendance were among the most prominently reported benefits among participants. The latter is consistent with the literature [21,22]. The reduction of barriers to access likely facilitates client attendance. Altogether, qualitative findings underscore why telehealth has become a routine practice for many clinicians in a postpandemic era.

Psychologists and Psychiatrists Engage Differently With Telehealth

Our quantitative data also revealed informative data regarding professional differences between psychologists' and psychiatrists' use (Table 2) and perceptions of telehealth. Findings indicated that psychologists spent significantly more hours delivering clinical services via videoconferencing and

XSL•FO
RenderX

telephone than psychiatrists. Conversely, there were no significant differences in the number of patients seen between the 2 disciplines across telehealth modalities. With respect to endorsement of telehealth concerns, psychiatrists were more likely to report challenges with remote assessment of patients, assessing high-risk patients and managing emergencies, showing empathy, and licensing or credentialing requirements compared to psychologists. Conversely, psychologists were more likely to be concerned about a lack of private space for patients during telehealth sessions compared to psychiatrists.

These findings likely reflect differences in the clinical services each discipline provides. Psychiatrists' concerns emphasize their need for additional infrastructural support for adapting in-person interventions like physical examinations or blood pressure monitoring [26]. Moreover, participating psychiatrists provided other clinical activities such as medication management or general medical services that are briefer in nature compared to assessment or psychotherapy services. Likewise, psychologists, who spend more direct time with patients (eg, providing psychotherapy), may prioritize ensuring that patients have a private space conducive to treatment. Altogether, these findings stress the need for telehealth strategies that are tailored to the needs of psychologists and psychiatrists. This could entail specialized telehealth training, telehealth models that integrate in-person and remote care, and improved regulatory frameworks to support remote psychiatric care. Likewise, platforms that optimize usability to accommodate different clinical needs and policies that focus on improving the standards and accessibility of telehealth may be beneficial.

Telehealth Service Delivery Involves Balancing Trade-Offs

An integration of quantitative and qualitative findings (Table 6) illustrated that there was much overlap among concerns and benefits of telehealth. For instance, some clinicians noted that telehealth provided valuable clinical insights into patients' home and family environments, while others endorsed concerns about a loss of clinical information, such as nonverbal cues. Likewise, providers observed improved attendance, engagement, and continuity of care. However, clinicians also worried about the therapeutic alliance and patient engagement. This overlap in perspectives has been reported in the literature [22,25,40]. Data from the international sample of mental health professionals presented here offer a nuanced perspective: there are inherent trade-offs between telehealth's benefits and disadvantages.

It is important for health care systems and stakeholders to be aware of the potential advantages and disadvantages of telehealth practice when planning mental health service. Flexible telehealth models can harness advantages while addressing concerns about clinical effectiveness and technology through hybrid or capacity-building initiatives that integrate available in-person support within patient communities (eg, medical personnel for physical examinations, psychometrist for cognitive assessments, and staff for crisis management).

At the provider level, clinicians must evaluate telehealth's trade-offs to determine its suitability for each patient. Will telehealth enhance or limit therapeutic outcomes? The available literature recommends evaluating provider competencies, patient needs and preferences, and organizational resources when determining the appropriateness of telehealth among other considerations [42]. For instance, this study found that perceived effectiveness of telehealth was higher for individual assessment and treatment as well as adult-aged patients. Likewise, service modality, particularly family or group therapy, or initial assessment may impact rapport building [39]. Research also suggests that having the initial patient interaction be in-person can help establish rapport prior to telehealth services [25,43]. Finally, adapting clinician communication style to include greater use of vocal tone and facial expression can help convey empathy [43].

Differences Among WHO Regions

Quantitative findings identified that participants from South-East Asia and Western Pacific-Asia were more likely to endorse concerns related to remote assessment and patient privacy, security, and confidentiality when compared to participants in the Americas-North. Conversely, participants from the Americas-South and Western Pacific-Asia were less likely to report concerns about patient privacy, security, and confidentiality, and technical difficulties, respectively. Qualitative findings revealed that clinicians in Europe, Americas-South, and Western-Pacific Asia reported disadvantages to telehealth (eg, inferiority to in-person care and difficulty obtaining payment for services) or simply indicated no additional benefits aside from remote contact. Reasons for variations in participant impressions across WHO regions are likely multifaceted; however, a few factors can be considered. First, the pandemic necessitated an immediate, global shift to telehealth with almost complete virtualization of many services at the time of data collection [7], meaning that many clinicians were faced with adopting unfamiliar telehealth platforms. As such, WHO regions varied in readiness for such changes in practice and training. For instance, in a survey study with medical students, trainee and early career psychiatrists from South-East Asia and Western-Pacific Asia regions conducted in early 2021 found limited opportunities for training and knowledge of telehealth among young professionals [41]. Some regions also faced key barriers to uptake including reservations about telehealth among clinicians, inadequate communication infrastructure (eg, in some parts of Latin America), and lack of regulatory frameworks (eg, professional licensing) in some parts of Europe [44-46].

Qualitative data also found that privacy and anonymity as benefits of telehealth were primarily noted by participants in the South-East Asia and Western Pacific-Asia regions—advantages that have been less addressed in the literature. Challenges related to stigma around mental illness in these regions may explain this heightened benefit among its mental health professionals [47]. A few participants, particularly those in the Americas-North, believed that the abrupt shift to telehealth had advanced the field of mental health. Quantitative data identified that clinicians in the Americas-North were more likely to be concerned about a lack of private space for patients during telehealth sessions compared to clinicians from Western Pacific-Asia. These findings may reflect how certain regions, such as North America—where much telehealth research has

XSL•FO
RenderX

been conducted [23,26], had more familiarity with the implementation of virtual models.

Countries and regions must now reflect on the future direction of telehealth. This study suggests that to capitalize on telehealth's broad advantages, investing in telehealth training and education and the dissemination of resources for mental health professionals are crucial. Countries should also address disparities in access to digital technology, known as the "digital divide," that affects LMICs and high-income countries. Regional telehealth models could promote digital literacy among specific patient groups, assess the use of tools that require less bandwidth (eg, telephone and instant messaging), or offer stability and privacy for telehealth services through local health centers or schools.

Limitations

This study has some limitations. Study findings were based on provider perspectives rather than objective measures of quality of mental health care or ethical practices. It will be important to further examine the patient perspective and assess the quality of services delivered. Given that data were gathered during the early stages of the pandemic, clinicians' perceptions and responses were likely impacted by their evolving contexts at the time. Furthermore, the quality of telehealth training indicated by clinicians was unknown. Since participant perceptions were based on multiple telehealth modalities (synchronous and asynchronous) and clinical contexts, direct comparisons (eg, telephone vs videoconferencing) were not possible. As such, present findings remain general and based on clinicians' own combination of telehealth service modalities provided in a public health emergency. Although this study had representation from all WHO regions and country income levels, most participating clinicians were from high- and upper-middle–income countries. Finally, given the cross-language qualitative data, it is possible that translation decisions and coders being outside of the sociocultural environments of many participant clinicians (coders were based in North America) impacted interpretation. Although the number of short-answer responses in Japanese or Chinese was small (roughly 44/496, 9%), the DeepL translations were not reviewed by fluent speakers, which may limit interpretive accuracy. Future research should explore implementation and patient outcomes within specific telehealth modalities, clinical settings, and geographical contexts, particularly its potential in LMICs.

Conclusions

The results of this study suggest that telehealth for mental health services presents notable clinical and technological challenges that are common across global contexts. Certain competencies such as telehealth training and experience may help mitigate these concerns. Moreover, given that psychologists and psychiatrists seem to engage with telehealth differently, specialized training and tailored approaches may also help implementation. Nonetheless, results also suggest that telehealth offers valuable benefits for patients and providers. Blended models that integrate virtual and in-person care or capacity-building platforms for underserved areas are worth exploring. In a postpandemic era with more deliberate implementation of telehealth services and a new generation of digitally trained clinicians, we can balance its benefits and limitations to optimize mental health care and clinical outcomes globally.

Acknowledgments

The authors would like to thank Konrad Czechowski from the School of Psychology at the University of Ottawa for support with qualitative data analyses and reporting.

Funding

This study was funded by the Canadian Institutes of Health Research (grant VR4-172750).

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the nature of our original ethics approval. Anonymized, group-level data are available from the corresponding author on reasonable request.

Authors' Contributions

CSK acquired funding. All authors contributed to conceptualization and methodology including study design and survey. GMR and TJR contributed to project administration. CSK contributed to supervision. JAG-P contributed to initial data curation. MIM conducted formal analysis and analyzed quantitative data. MIM and CSK analyzed the qualitative data. MIM wrote the first draft with critical input from CSK. BK, GMR, JAG-P, MK, RR, and TJR helped with reviewing and editing the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
GRAMMS checklist.
[DOCX File , 19 KB - jmir_v28i1e75905_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Survey questions.
[DOCX File , 24 KB - jmir_v28i1e75905_app2.docx]

References

1. Appleton R, Williams J, Vera San Juan N, Needle JJ, Schlief M, Jordan H, et al. Implementation, adoption, and perceptions of telemental health during the COVID-19 pandemic: systematic review. J Med Internet Res 2021;23(12):e31746. [doi: 10.2196/31746] [Medline: 34709179]
2. Montoya MI, Kogan CS, Rebello TJ, Sadowska K, Garcia-Pacheco JA, Khoury B, et al. An international survey examining the impact of the COVID-19 pandemic on telehealth use among mental health professionals. J Psychiatr Res 2022;148:188-196 [FREE Full text] [doi: 10.1016/j.jpsychires.2022.01.050] [Medline: 35131587]
3. Global diffusion of eHealth: making universal health coverage achievable: report of the third global survey on eHealth. World Health Organization. 2016. URL: https://www.who.int/publications/i/item/9789241511780 [accessed 2025-02-23]
4. Talevi D, Socci V, Carai M, Carnaghi G, Faleri S, Trebbi E, et al. Mental health outcomes of the CoViD-19 pandemic. Riv Psichiatr 2020;55(3):137-144. [doi: 10.1708/3382.33569] [Medline: 32489190]
5. Xiang YT, Yang Y, Li W, Zhang L, Zhang Q, Cheung T, et al. Timely mental health care for the 2019 novel coronavirus outbreak is urgently needed. Lancet Psychiatry 2020;7(3):228-229 [FREE Full text] [doi: 10.1016/S2215-0366(20)30046-8] [Medline: 32032543]
6. Kola L, Kohrt BA, Hanlon C, Naslund JA, Sikander S, Balaji M, et al. COVID-19 mental health impact and responses in low-income and middle-income countries: reimagining global mental health. Lancet Psychiatry 2021;8(6):535-550 [FREE Full text] [doi: 10.1016/S2215-0366(21)00025-0] [Medline: 33639109]
7. Zangani C, Ostinelli EG, Smith KA, Hong JSW, Macdonald O, Reen G, et al. Impact of the COVID-19 pandemic on the global delivery of mental health services and telemental health: systematic review. JMIR Ment Health 2022;9(8):e38600 [FREE Full text] [doi: 10.2196/38600] [Medline: 35994310]
8. Pandya A, Waller M, Portnoy JM. The regulatory environment of telemedicine after COVID-19. J Allergy Clin Immunol Pract 2022;10(10):2500-2505 [FREE Full text] [doi: 10.1016/j.jaip.2022.07.001] [Medline: 35842133]
9. Interim ethical guidelines for psychologists providing psychological services via electronic media. Canadian Psychological Association. 2020. URL: https://cpa.ca/docs/File/Ethics/CPAe-therapyGuidelinesUpdate2020.pdf [accessed 2026-02-05]
10. CPA guidelines on telepsychology. Canadian Psychological Association. 2023. URL: https://cpa.ca/docs/File/Practice/Telepsychology_Guidelines-September_2023-EN_FINAL.pdf [accessed 2026-02-05]
11. Cantor JH, McBain RK, Ho P, Bravata DM, Whaley C. Telehealth and in-person mental health service utilization and spending, 2019 to 2022. JAMA Health Forum 2023;4(8):e232645 [FREE Full text] [doi: 10.1001/jamahealthforum.2023.2645] [Medline: 37624614]
12. Page C, Stamm K. Telehealth and hybrid practice are here to stay. American Psychological Association. 2024. URL: https://www.apa.org/monitor/2024/09/telehealth-hybrid-practice [accessed 2025-11-17]
13. Connolly SL, Miller CJ, Lindsay JA, Bauer MS. A systematic review of providers' attitudes toward telemental health via videoconferencing. Clin Psychol (New York) 2020;27(2):10.1111/cpsp.12311 [FREE Full text] [doi: 10.1111/cpsp.12311] [Medline: 35966216]
14. Hubley S, Lynch SB, Schneck C, Thomas M, Shore J. Review of key telepsychiatry outcomes. World J Psychiatry 2016;6(2):269-282 [FREE Full text] [doi: 10.5498/wjp.v6.i2.269] [Medline: 27354970]
15. Sugarman DE, Busch AB. Telemental health for clinical assessment and treatment. BMJ 2023;380:e072398. [doi: 10.1136/bmj-2022-072398] [Medline: 36646462]
16. Varker T, Brand RM, Ward J, Terhaag S, Phelps A. Efficacy of synchronous telepsychology interventions for people with anxiety, depression, posttraumatic stress disorder, and adjustment disorder: a rapid evidence assessment. Psychol Serv 2019;16(4):621-635. [doi: 10.1037/ser0000239] [Medline: 29809025]
17. Liberati E, Richards N, Parker J, Willars J, Scott D, Boydell N, et al. Remote care for mental health: qualitative study with service users, carers and staff during the COVID-19 pandemic. BMJ Open 2021;11(4):e049210 [FREE Full text] [doi: 10.1136/bmjopen-2021-049210] [Medline: 33888531]
18. Watson JD, Pierce BS, Tyler CM, Donovan EK, Merced K, Mallon M, et al. Barriers and facilitators to psychologists' telepsychology uptake during the beginning of the COVID-19 pandemic. Int J Environ Res Public Health 2023;20(8):5467 [FREE Full text] [doi: 10.3390/ijerph20085467] [Medline: 37107748]
19. Jaywant A, Aulitzky W, Avari J, Buchheim A, Dubin M, Galffy M, et al. Multinational perspectives on changes to psychiatric care during the COVID-19 pandemic: a survey of practicing psychiatrists. Neuropsychiatr 2023;37(3):115-121. [doi: 10.1007/s40211-022-00452-x] [Medline: 36600105]
20. Shalev L, Eitan R, Rose AJ. The use of telepsychiatry services in emergency settings: scoping review. J Med Internet Res 2024;26:e51814 [FREE Full text] [doi: 10.2196/51814] [Medline: 39008831]
21. Guinart D, Marcy P, Hauser M, Dwyer M, Kane JM. Mental health care providers' attitudes toward telepsychiatry: a systemwide, multisite survey during the COVID-19 pandemic. Psychiatr Serv 2021;72(6):704-707. [doi: 10.1176/appi.ps.202000441] [Medline: 33593104]

XSL•FO
RenderX

22. Lipschitz JM, Connolly SL, Van Boxtel R, Potter JR, Nixon N, Bidargaddi N. Provider perspectives on telemental health implementation: lessons learned during the COVID-19 pandemic and paths forward. Psychol Serv 2023;20(Suppl 2):11-19 [FREE Full text] [doi: 10.1037/ser0000625] [Medline: 35201809]
23. Cowan KE, McKean AJ, Gentry MT, Hilty DM. Barriers to use of telepsychiatry: clinicians as gatekeepers. Mayo Clin Proc 2019;94(12):2510-2523. [doi: 10.1016/j.mayocp.2019.04.018] [Medline: 31806104]
24. Farrer LM, Clough B, Bekker MJ, Calear AL, Werner-Seidler A, Newby JM, et al. Telehealth use by mental health professionals during COVID-19. Aust N Z J Psychiatry 2023;57(2):230-240 [FREE Full text] [doi: 10.1177/00048674221089229] [Medline: 35360958]
25. Galvin E, Desselle S, Gavin B, Quigley E, Flear M, Kilbride K, et al. Patient and provider perspectives of the implementation of remote consultations for community-dwelling people with mental health conditions: a systematic mixed studies review. J Psychiatr Res 2022;156:668-678 [FREE Full text] [doi: 10.1016/j.jpsychires.2022.10.051] [Medline: 36399859]
26. Uscher-Pines L, Sousa J, Raja P, Mehrotra A, Barnett ML, Huskamp HA. Suddenly becoming a "Virtual Doctor": experiences of psychiatrists transitioning to telemedicine during the COVID-19 pandemic. Psychiatr Serv 2020;71(11):1143-1150 [FREE Full text] [doi: 10.1176/appi.ps.202000250] [Medline: 32933411]
27. Lin T, Stone SJ, Heckman TG, Anderson T. Zoom-in to zone-out: Therapists report less therapeutic skill in telepsychology versus face-to-face therapy during the COVID-19 pandemic. Psychotherapy (Chic) 2021;58(4):449-459. [doi: 10.1037/pst0000398] [Medline: 34881922]
28. Schoonenboom J, Johnson RB. How to construct a mixed methods research design. Kolner Z Soz Sozpsychol 2017;69(Suppl 2):107-131 [FREE Full text] [doi: 10.1007/s11577-017-0454-1] [Medline: 28989188]
29. Miles MB, Huberman AM, Saldana J. Qualitative Data Analysis: A Methods Sourcebook. 3rd Edition. Thousand Oaks, CA: SAGE Publications; 2014.
30. Reed GM, Rebello TJ, Pike KM, Medina-Mora ME, Gureje O, Zhao M, et al. WHO's Global Clinical Practice Network for mental health. Lancet Psychiatry 2015;2(5):379-380. [doi: 10.1016/S2215-0366(15)00183-2] [Medline: 26360271]
31. Kogan CS, Garcia-Pacheco JA, Rebello TJ, Montoya MI, Robles R, Khoury B, et al. Longitudinal impact of the COVID-19 pandemic on stress and occupational well-being of mental health professionals: an international study. Int J Neuropsychopharmacol 2023;26(10):747-760 [FREE Full text] [doi: 10.1093/ijnp/pyad046] [Medline: 37531283]
32. Creswell JW. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. 4th Edition. Los Angeles, CA: SAGE; 2014.
33. O'Cathain A, Murphy E, Nicholl J. The quality of mixed methods studies in health services research. J Health Serv Res Policy 2008;13(2):92-98. [doi: 10.1258/jhsrp.2007.007074] [Medline: 18416914]
34. Elliott R, Bohart AC, Watson JC, Murphy D. Therapist empathy and client outcome: an updated meta-analysis. Psychotherapy (Chic) 2018;55(4):399-410 [FREE Full text] [doi: 10.1037/pst0000175] [Medline: 30335453]
35. Wampold BE. How important are the common factors in psychotherapy? An update. World Psychiatry 2015;14(3):270-277 [FREE Full text] [doi: 10.1002/wps.20238] [Medline: 26407772]
36. Saldaña J. The Coding Manual for Qualitative Researchers. Thousand Oaks, CA: Sage Publications Ltd; 2009:xi-223.
37. Maxwell JA. Using numbers in qualitative research. Qual Inq 2010;16(6):475-482. [doi: 10.1177/1077800410364740]
38. Perle JG, Burt J, Higgins WJ. Psychologist and physician interest in telehealth training and referral for mental health services: an exploratory study. J Technol Hum Serv 2014;32(3):158-185. [doi: 10.1080/15228835.2014.894488]
39. Sugarman DE, Horvitz LE, Greenfield SF, Busch AB. Clinicians' perceptions of rapid scale-up of telehealth services in outpatient mental health treatment. Telemed J E Health 2021;27(12):1399-1408 [FREE Full text] [doi: 10.1089/tmj.2020.0481] [Medline: 33600272]
40. Lawson JL, Doran JM, O'Shea M, Abel EA. The good, the bad, the uncertain: diverse provider experiences with telemental health during COVID-19. Psychiatr Q 2022;93(3):753-774 [FREE Full text] [doi: 10.1007/s11126-022-09990-7] [Medline: 35689752]
41. Orsolini L, Jatchavala C, Noor IM, Ransing R, Satake Y, Shoib S, et al. Training and education in digital psychiatry: a perspective from Asia-Pacific region. Asia Pac Psychiatry 2021;13(4):e12501 [FREE Full text] [doi: 10.1111/appy.12501] [Medline: 34873845]
42. Luxton DD, Nelson EL, Maheu M. A Practitioner's Guide to Telemental Health: How to Conduct Legal, Ethical, and Evidence-Based Telepractice. 1st Edition. Washington, DC: American Psychological Association; 2023.
43. Powell M. Empathy in digital healthcare [doctoral dissertation]. University of Central Florida. 2024. URL: https://stars.library.ucf.edu/etd2024/60 [accessed 2026-02-05]
44. Castaño Pérez GA, Rojas Bernal LA, Maldonado MA. Telemedicine for mental health in Latin America. In: Mental Health & Addiction: An Integrated Care in Latin America. Canada: CDS Press; 2023:26-56.
45. Haddouk L, Schneider B, Daele TV, Lang F. Telepsychology in France following the COVID-19 crisis. Eur J Psychol Open 2025;84(1):13-19 [FREE Full text] [doi: 10.1024/2673-8627/a000063]
46. Omboni S, Padwal RS, Alessa T, Benczúr B, Green BB, Hubbard I, et al. The worldwide impact of telemedicine during COVID-19: current evidence and recommendations for the future. Connect Health 2022;1:7-35 [FREE Full text] [doi: 10.20517/ch.2021.03] [Medline: 35233563]

XSL•FO
RenderX

47. Singh S, Sagar R. Online psychotherapy during the COVID-19 pandemic: the good, the bad, and the ugly. Indian J Psychol Med 2022;44(2):177-180 [FREE Full text] [doi: 10.1177/02537176211070427] [Medline: 35655984]

Abbreviations

GCPN: Global Clinical Practice Network
GRAMMS: Good Reporting of a Mixed Methods Study
ICD-11: International Classification of Diseases, 11th Revision
LMIC: low-income and middle-income country
WHO: World Health Organization

Edited by A Stone; submitted 14.Apr.2025; peer-reviewed by NG Badr, C Babbage; comments to author 06.Nov.2025; revised version received 08.Jan.2026; accepted 30.Jan.2026; published 04.Mar.2026.

Please cite as:
Montoya MI, Kogan CS, Khoury B, García-Pacheco JA, Robles R, Rebello TJ, Kulygina M, Reed GM
Mental Health Professionals' Perceptions of Benefits and Disadvantages of Telehealth: International Mixed Methods Study
J Med Internet Res 2026;28:e75905
URL: https://www.jmir.org/2026/1/e75905
doi:10.2196/75905
PMID:

©Madeline I Montoya, Cary S Kogan, Brigitte Khoury, José A García-Pacheco, Rebeca Robles, Tahilia J Rebello, Maya Kulygina, Geoffrey M Reed. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Misinformation in Social Media Narratives on Highly Pathogenic Avian Influenza: Systematic Content Analysis of Facebook and Instagram Posts

Ahmed Al-Rawi[1*], PhD; Betty Ackah[1*], PhD; Abdelrahman Fakida[1*], MA; Kelley Lee[2*], DPhil

[1]School of Communication, Simon Fraser University, Schrum Science Centre-K 9653, Burnaby, BC, Canada

[2]Faculty of Health Sciences, Simon Fraser University, Burnaby, BC, Canada

[*]all authors contributed equally

Corresponding Author:
Ahmed Al-Rawi, PhD
School of Communication, Simon Fraser University, Schrum Science Centre-K 9653, Burnaby, BC, Canada

Abstract

Background: Recurrent outbreaks of the highly pathogenic avian influenza (HPAI) A (H5N1) virus in farmed poultry, and reports of infections in dairy cattle herds in the United States since March 2024, have triggered concerns about the spillover threat to human populations and a subsequent influenza pandemic. The increasing threat that H5N1 poses to human health has led to more vigilant public health monitoring of these developments. In addition to intensifying surveillance, preventative strategies—like vaccinating those at higher risk—are being evaluated to help minimize infection and spread.

Objective: Efforts to mitigate and respond to such an event will entail broad public health interventions including vaccination. However, analysis of the COVID-19 pandemic suggests that information quality can significantly impact the effectiveness of such measures by influencing public understanding and trust. Misinformation about H5N1 and other viruses circulating online often includes inaccurate information about transmission, prevention, and the severity of the viruses. By systematically analyzing these false narratives, public health authorities can better tailor their pandemic prevention, preparedness, and response strategies.

Methods: In light of the emerging threat of H5N1, we analyzed the content of social media posts from Facebook (approximately 350,000) and Instagram (n=69,551) related to HPAI. Using 40 keywords associated with misinformation, we identified over 500 posts explicitly mentioning H5N1 and related terms for further systematic analysis. Posts were coded to identify targets and topics in the social media narratives. The “target” refers to the organization or person mentioned in the post, while the “topic” refers to the primary issue or subject being addressed.

Results: Our content analysis identifies 7 main targets of misinformation, including government (149/544, 27%), health authorities (108/544, 20%), and international organizations (74/544, 14%). Also, from the 6 topics that have been identified, we found that the most widespread one was that authority figures purposefully engineer pandemics to achieve multiple political, economic, and other objectives (362/544, 67%) followed by societal destruction (121/544, 22%), and anti-vaccination (84/544, 15%). Other themes include societal destruction and religious allusions and prophecies.

Conclusions: Our analysis of online content showed that H5N1 misinformation was primarily aimed at individuals or groups with differing degrees of political or institutional authority, such as government leaders and public health officials. These figures were often the focus due to their involvement in making health policy decisions and implementing public health measures. Decision-making entities and individuals were the target of various misinformation narratives. Results demonstrate the ongoing need for monitoring health misinformation to inform evolving public health responses to HPAI.

(J Med Internet Res 2026;28:e73275) doi:10.2196/73275

KEYWORDS

highly pathogenic avian influenza; H5N1; social media; misinformation; health communication

Introduction

The highly pathogenic avian influenza (HPAI) H5N1 virus was first identified in poultry flocks in 1959, with the first human infection detected in 1997 [1-3]. The virus has since circulated worldwide, causing serious outbreaks in wild and domesticated birds, with periodic human cases reported [4]. Since 2021, H5N1 has been identified in a growing number of other wild, domestic, and zoo mammals (48 species by August 2024), in some cases causing major die-offs [5,6]. In March 2022, infections were first detected in dairy cattle in the United States, with more than 700 herds confirmed as infected by January 2025 [7]. Around 40 human infections linked to dairy cattle by January 2025 have

XSL•FO
RenderX

raised concerns about the possibility of an influenza pandemic in human populations [8]. While the current assessed risk of human infection ranges in severity from low to medium symptoms compared to earlier assessments [4,9], the further spread of H5N1 to other mammals increases the potential for genetic reassortment that promotes further spillover events and sustained human-to-human transmission.

The growing risk to human health from H5N1 has prompted stronger public health monitoring to these unfolding events. Alongside heightened surveillance efforts, preventative measures such as vaccination of high-risk populations are being considered to reduce infection and transmission. However, substantial evidence suggests that the ability to implement such measures could be limited in the wake of the COVID-19 pandemic [10]. Public trust was a key factor in achieving effective public health outcomes during COVID-19 specifically to achieve public compliance with recommended measures [11]. Online misinformation has been found to erode public trust, which has, in turn, facilitated vaccine hesitancy and refusal [12]. However, levels of public trust in the United States in key health agencies declined between June 2023 and December 2024, with 53% (down from 64%) trusting the Food and Drug Administration to make the right recommendations. Sixty-one percent (down from 66%) trust the recommendations of the US Centers for Disease Control and Prevention (CDC) [13]. While there are multiple factors contributing to increased mistrust, a decline in quality of the health information environment is a key issue.

This study seeks to inform emerging public health responses to H5N1 by understanding the content of misinformation on selected social media platforms related to H5N1. To date, there has been limited study of this subject. In part, this is due to increasing controls put into place by social media companies regarding access to data including for scholarly purposes [14]. For instance, X (formerly Twitter) discontinued its free Academic API (application programming interface) in February 2023. Meta's CrowdTangle, which allowed researchers free but limited access to data from Facebook and Instagram, is no longer available from August 14, 2024 [15,16]. This reduced access poses new challenges for understanding trends in misinformation. Using CrowdTangle and another data collection tool, this study navigates these barriers to understand the targets and topics of misinformation related to H5N1 on Facebook and Instagram. The findings show that such data can offer useful insights when monitoring social and behavioral factors that may impact public health efforts to mitigate or control a public health emergency [17].

Studies show that the ubiquity of internet connectivity in many societies has now made social media platforms primary information sources and means of communication [18]. Social media serve as virtual spaces for interactive engagements, particularly during times of uncertainty such as public health emergencies [19]. While these technologies enable connectivity in a globalized world, they can also be the source and rapid spreader of low-quality information [20]. This includes misinformation defined as misleading, inaccurate, or false information, although in some cases, there might be a facet of truth. Misinformation is distinguished from disinformation primarily by intent. Disinformation refers to the deliberate spread of false content with the intent to mislead or misinform [21,22]. However, due to the challenges of deducing the intent behind false information, we use the term misinformation to collectively describe all content in our dataset that is associated with falsehood or misleading claims. This approach is aligned with the trend by researchers toward understanding information quality separate from intent [23]. There is substantial evidence that both are part of what has been termed the post-truth world where people reject objective facts in favor of information aligned with ideology, identity, and other biases over verified facts [24,25]. This resulted in a proliferation of health misinformation across social media platforms during the COVID-19 pandemic. Moreover, this trend is likely to continue, and even worsen, as social media platforms such as X and Meta discard commitments to fact-checking in favor of leaving online communities "to decide when posts are potentially misleading and need more context" [26].

There remains limited but growing scholarly study on the content of misinformation related to H5N1. Studies show that misinformation concerning infectious diseases is more substantial and spreads more widely and quickly on social media than verifiable information [27-29]. Previous studies to date have analyzed the relationship between misinformation and a broad range of infectious diseases such as Ebola, H1N1, and Middle East Respiratory Syndrome [25,30,31]. Other studies have investigated social media misinformation against vaccination [27].

The structural features of social media platforms further the spread of misinformation as they often prioritize sensational content. Vosoughi et al [28] examined approximately 126,000 stories tweeted by about 3 million users on X to reveal that misinformation diffused at a significantly accelerated rate than verifiable information. Characteristics such as their novelty and emotional intensity facilitate their rapid diffusion. This phenomenon is particularly evident during health crises like avian influenza outbreaks [32].

For example, Yan et al [33] explore the relationship between social networking status, that is, the modes of information diffusion, relationship building, specific algorithmic exigencies, and influenza susceptibility. Gori and colleagues [34] examined social media debates on vaccines during the early part of the COVID-19 pandemic to understand the dynamics of vaccine readiness in Italy. The study highlighted how online debates using anti-vaccination rhetoric can influence the behavior of the vaccine-hesitant compared to pro-vaccination arguments.

Other studies posit the salience of factors such as people's reliance on these platforms as primary sources of news and individual-level predispositions to a conspiratorial mindset [35,36]. As well, public concerns about the impacts of infectious disease outbreaks on social media have intensified in the wake of the COVID-19 pandemic. These underlying factors have amplified the rapid spread of misinformation, contributing to the rise of infodemics during epidemics and pandemics. They present significant dangers, leading to public panic, reducing trust in health authorities, and resulting in uninformed health decisions that inhibit optimal uptake of public health

XSL•FO
RenderX

interventions. People who adopt inaccurate information about disease events are more likely to reject factual information about their severity, thus inhibiting their willingness to comply with appropriate measures [37].

Evidence suggests that the public's interpretation of information on infectious diseases on social media platforms can influence behaviors in response to relevant interventions, which, in turn, can impact the trajectory of a health crisis [38]. This is particularly important for early interventions during an emerging crisis that have the potential to become pandemics [39]. Health crises tend to heighten levels of public uncertainty, anxiety, and fear. Studies show that online spaces offer opportunities for people to engage with social networks and information sources to allay their concerns [40,41]. However, as Wang et al [29] found, elevated levels of anxiety and fear can make people more susceptible to poor quality information.

The above research has identified best practices for health communication to attend to public concerns, facilitate appropriate interpretation of the information, and support trust in public health interventions [17]. For instance, the enduring impact of COVID-19 has affected people's trust toward organizations like the US CDC and the government, in some cases augmenting an aversion to vaccination in general. Knowing these factors is key for appropriate communication framing [17]. Indeed, the COVID-19 pandemic showed how misinformation can significantly hamper public health interventions during health emergencies.

In this context, a better understanding of the content of misinformation on social media can inform effective public health responses to the evolving threat from H5N1 [42]. Misinformation on H5N1 may thus hinder public health efforts to identify, mitigate, or respond to the spread of the virus in animal and human populations by fostering further mistrust in public health responses [43]. Misinformation also has repercussions on individual and societal health outcomes by influencing health-seeking behaviors and compliance with public health interventions [17,29,44]. Although there are different subtypes of highly pathogenic avian influenza, such as H5N6 and H7N9 that have raised concern as well and remained important to analyze, our scope focuses on H5N1 due to its role as a primary subtype on known global outbreaks and designated by health authorities as a virus with pandemic potential. Moreover, H5N1 received wide media and public attention, making it particularly relevant for a study on health misinformation [45]. In sum, research on health misinformation related to H5N1 is limited, which this study seeks to address amid the ongoing spread of the virus. Thus, this study aims to answer the following research question: What is the content of misinformation about H5N1 on social media and how can this inform public health efforts at pandemic prevention, preparedness, and response?

Methods

Overview

Using scientific and colloquial terms associated with HPAI and specifically H5N1, such as bird flu, avian influenza, and avian flu, we collected approximately 350,000 Facebook and Instagram posts referencing H5N1 and associated terms using both CrowdTangle and a webometric tool called Zeeschuimer for more Instagram data. The data included both textual and multimedia content. We focused on these platforms because they are among the 3 most accessed social media platforms globally. Facebook is the market leader with an estimated active user population of over 3 billion [46]. Although YouTube is ahead of Instagram as the second most popular platform, its content being primarily video-based, with minimal text, excluded it from this study. We also selected Facebook and Instagram to expand the knowledge base on health misinformation beyond the platform X, especially regarding the affordances of social media platforms.

The CrowdTangle data were collected from June 5, 2014, to June 4, 2024, the posts are not limited to geographical demarcations. We collected the additional Instagram data with Zeeschuimer in July 2024. The total Instagram posts were (n=69,551). For both the Facebook and Instagram datasets, we employed a Python script to identify misinformation-related posts using terms such as "plandemic," "medicalmurder," and "LuciferShot" [47,48]. We compiled this list of terms based on research on health misinformation from academic literature, [49,50], online exploratory observations, and notes from the research team. Keywords were specifically focused on pandemic and vaccine conspiracy–related terms, demonstrating general connections for misinformation narratives between COVID-19 and vaccines. The study period from June 5, 2014, to June 5, 2024, was selected to ensure a proper review of discussions on the virus and its continued activity and outbreaks. A specific instance occurred in March 2024 when H5N1 infections were reported in dairy cattle in the United States for the first time, highlighting its continued relevance for public discussions [51].

Our final dataset contains 544 posts after removing duplicates. Though the volume of posts flagged as misinformation and misleading appears to be limited, we note several limitations. First, and in searching for specific keywords, it is difficult to confirm the identification of all the available misleading or misinformation-filled posts in the dataset, but the above approach has been effective in identifying a relevant subset of data that we can further explore. In addition, since the two data collection tools are not comprehensive in getting data from Meta, we cannot make any claims about the volume of misinformation on Facebook and Instagram. Moreover, as the initial number of gathered posts (350,000) is far larger than the remaining posts we analyzed, the difference in numbers is due to our inclusion criteria that excluded any duplicates, irrelevant use of keywords, and posts beyond our scope. This reduction also represents the inclusion criteria to ensure relevance.

Employing a general inductive approach [52] as part of systematic content analysis approach, 3 coders carried out a close reading of a 10% sample of the dataset to outline the emerging and dominant arguments. The analysis was multimodal, which involves the text and any associated multimedia content. Following a process of several rounds of phased discussions, the researchers organized the broad themes into 2 categories (targets and topics). These categories were used to link the analysis to the research objectives of

XSL•FO
RenderX

understanding the prevailing misinformative concepts on HPAI to inform effective public health agenda. The target is the organization or individual discussed in the post. The topic concerns the main subject or issue in the post, ensuring that subcategories were mutually exclusive and that the themes of the various subcategories are unique to each category, while the central messages do not overlap. Two researchers then used this initial codebook on a sample of the first 100 posts to test the codebook and test intercoder reliability. After the first coding round, the researchers adjusted the target section of the codebook, adding corporations as a category to account for multinational and other commercial entities, along with a category for irrelevant posts. The codebook has nine categories under target: (1) government had the highest occurring target references (n=149), (2) public health authorities had 108 references, (3) international organizations (n=74), (4) elites (n=64), (5) pharmaceutical industries (n=52), (6) mainstream media (n=43), and (7) corporations (n=23). The final categories for targets were (8) other/unclear (n=113) and (9) irrelevant (n=53). We distinguished between these 2 categories because while some posts were focused on H5N1 but did not address a target (other), others did have targets but had topics that were unrelated to avian influenza (irrelevant).

Researchers also adjusted the topics section following the first round of coding to address vagueness and overlapping categories. After the second round of intercoder reliability, further adjustments were made, including renaming a category to religious allusions and prophecies to include a broader range of religious content. Moreover, we removed an alternative medicine category as there were only 2 posts, and they were better classified under other. The adjusted seven categories are (1) pandemic engineering, which accounted for 362 topics, (2) societal destruction (n=121, 22%), (3) anti-vaccination (n=84, 15%), (4) elections/democracy (n=66, 12%), (5) unjustified food insecurity (n=59, 11%), (6) religious allusions and prophecies (n=26, 5%), and (7) other/irrelevant (n=77, 14%). The final category contained posts that were not focused on H5N1 but on other pathogens and diseases such as SARS-CoV-2/COVID-19. Others did not have a clear message and sometimes also used contradictory hashtags [49]. There were also 2 posts advancing misinformative pseudoscientific remedies to H5N1 infections by advertising plant-based vaccines. After 3 attempts involving the modifications described above, intercoder reliability was statistically acceptable at 0.887 for targets and 0.899 for topics using Krippendorff Alpha [53]. Though limited contextual insight was identified due to lack of context, we also examined the nature of the social media posters by reviewing their profiles and previous posts. Since the evidence we collected on users' profiles was only anecdotal, we did not report these findings in this study. Finally, as we analyzed publicly available social media data, all posts used have been deidentified and paraphrased to prevent re-identification and to protect user privacy.

Ethical Considerations

This project received an exemption from the requirements of Research Ethics Review at Simon Fraser University (30002857). All analysis is based on publicly available and deidentified data. All publicly available social media content used in this study was deidentified to prevent reidentification of posts or individuals.

Results

Targets

Types of Targets

The identified targets (see Figure 1) were individuals or entities that possess varying levels of economic, political, or decision-making power. An underlying theme in the claims made against them centered on implying that they intentionally generated public concern.

XSL•FO
RenderX

Figure 1. Percentage distribution of targets of misinformation on H5N1.

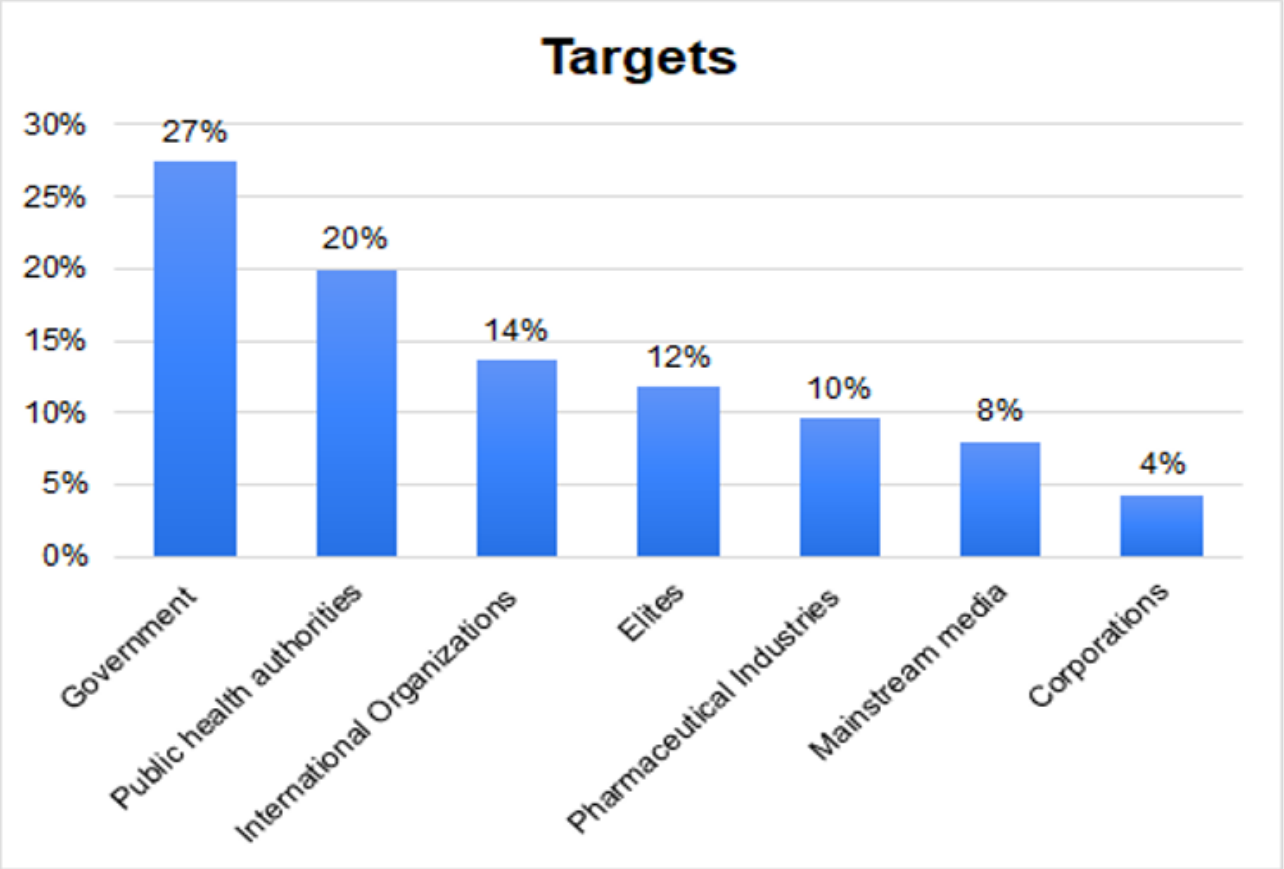

Government

This category constituted 27% (149/544) of targets, with figures being rounded off to the nearest whole number. The posts addressed the roles and responsibilities of governments and governmental entities in relation to pandemics, including heads of states, ministries, officials, and agencies in different countries, and the United States Food and Drug Administration. The category included posts claiming that the government and its "deep state" actors created pandemics for their own benefit such as to influence elections or to benefit other corrupt actors and individuals. Posts often highlighted the COVID-19 pandemic measures that users claimed resulted in societal destruction, suggesting that warnings of avian influenza can lead governments to reimpose similar restrictions. The following paraphrased post illustrates this point:

> *Don't worry about #covid_19 the government is scaring you and lying to you to start a new digital money system It's the same pattern they used before to [expletive] us and have reasons to push another PLANDEMIC. [expletive] them!*

In this example, wherein posts may contain typos as they are reproduced verbatim, the user points out different health crises such as COVID-19, Ebola, and avian influenza, suggesting that these crises were falsely created by the government as part of a larger conspiracy to impose certain financial measures on citizens. These are similar to other posts in the dataset in which users connect different health crises to argue that governments created or pushed these health crises to sow fear among citizens, thus creating conditions that allow the imposition of corrupt measures.

Public Health Authorities

Specific public health officials were targeted in posts, notably well-known figures such as Anthony Fauci, then director of the National Institute of Allergy and Infectious Diseases, and Robert Redfield, then director of the CDC. For instance, users frequently highlighted Redfield's past warnings about the possible threat of an avian influenza pandemic as evidence of planned manipulation and creation of a "plandemic." These posts also referred to public health officials as part of a broader network of parties that benefit from spreading fear about pandemics to advance individual interests. Public health officials comprised 20% (108/544) of the targets.

International Organizations

Fourteen percent (74/544) of the social media posts identified linked international organizations, especially those involved in pandemic responses such as the World Health Organization (WHO) and the World Economic Forum, with efforts to push political agendas that the posts claimed serve or align with their interests. This post, for example, accused the WHO of spreading false information regarding an avian influenza death to create a health crisis that can help control the public and push for certain medical interventions. The paraphrased excerpt below partly explains this claim:

> *The terrorist global health care organization is making up stories about someone dying after being infected with #avian influenza… because those in*

XSL•FO
RenderX

> *power want to turn us into mindless people… to push their medicines and spread panic and keep us under their control. Don't trust them; they are manipulating you. Make sure to focus on your healthy habits and keep learning and praying.*

Elites

Elites are individuals perceived by users to hold power and influence in the health sector and pandemic response such as Bill Gates, American businessman and philanthropist, and Klaus Schwab, the founder of the World Economic Forum. Constituting 12% (64/544) of the targets, the posts connected them with larger conspiracy theories, suggesting they had ambitions to reduce meat consumption, which was linked to perceived food shortages, and advanced the "Great Reset," a theory claiming that the COVID-19 pandemic measures were orchestrated by global elites allegedly to achieve their own goals, such as the following paraphrased quote:

> *It seems that the next Plandemic could be happening soon. This could be airborne and will be blamed on the new bird flu strain. Be careful, there is always another plan in the works. Billionaires already told us that they will go for our attention during the new pandemic, and the global health organizations and those globalists planned this new disease and created a Global Pandemic plan because they want to control our nations through a "Health" emergency.*

Pharmaceutical Industries

Posts focused on pharmaceutical companies accounted for 9.57% (52/544) of posts. Several users who promoted skeptical messages about avian influenza and pandemics highlighted the role they claimed these pharmaceutical companies played with other discussed entities such as the government and international organizations to amplify perceived medical threats and eventually sell the vaccines these companies produced. For instance, one post commented on avian influenza-related news saying that "psychic vaccine makers" prepared a vaccine for avian influenza a year in advance, suggesting that this is a "plandemic." Such posts reference vaccine developments as evidence that pandemics are intentionally created to help vaccine manufacturers.

Mainstream Media

Posts targeting mainstream media amounted to approximately 8% (43/544). Users who criticized mainstream media outlets portrayed them as tools used to spread fear about pandemics and diseases, and sometimes to encourage people to take vaccines, by claiming that they are manipulating and exaggerating to eventually support corruption that allows governments and pharmaceutical companies to benefit from pandemics.

Other Private Companies

Four percent (23/544) of the posts identified targeted nonpharmaceutical companies. Among the allegations were claims that various companies influence governments and organizations to declare pandemics, given that these companies were described as profiting from health crises. Some posts alleged that food companies, for example, pushed governments to use avian influenza to reduce the number of poultry in small farms or to mandate vaccines that could harm chickens. Such claims argue that these corrupt measures create food shortages that eventually corporations that control poultry supply will benefit from.

Topics

Types of Topics

The dataset included an extensive range of topics that were categorized according to the most prevalent narrative arguments (see Figure 2). The proportion of posts referring to the engineering of infectious disease pandemics was significant. The topics largely attributed major infectious disease events to human actions rather than any roles that pathogens play.

XSL•FO
RenderX

Figure 2. Percentage distribution of H5N1 misinformation topics.

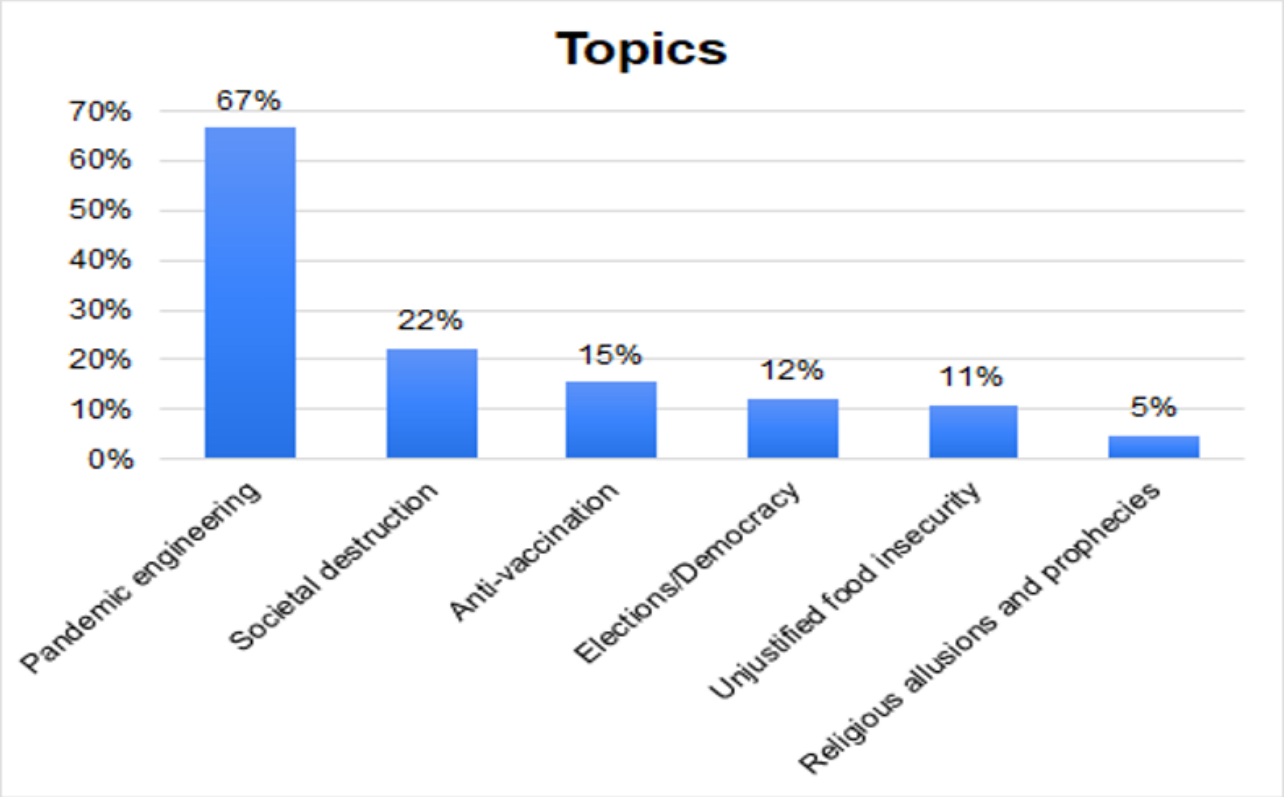

Pandemic Engineering

These posts (362/544, 67%) claim that the H5N1 outbreaks were allegedly engineered. They claimed that viruses causing widespread outbreaks are fabricated bioweapons aimed at stimulating economic gain for their creators or enshrining political power for governments or international organizations. Other targets of pandemic engineering claims are pharmaceutical companies and elites like Bill Gates. Popular hashtags associated with this theme include "#scamdemic" and "#plandemic." This post illustrates narratives about antiviral medications being specifically designed as bioweapons to manipulate people's beliefs such as the following paraphrased claim:

> *That is strange, they are already creating a new medication for the new flu. This drug and its company already know all the potential side effects …including sensory symptoms and continued false beliefs. The company that created the drug could just be saying that side effects could seem like a zombie apocalypse…They are clearly experimenting these medications on us to change our thoughts and memories!*

Societal Destruction

Amounting to 22% (121/544) of posts, this category is made up of posts that described widespread negative consequences from H5N1 outbreaks and public health measures. The consequences outlined were either projections of future occurrences or outcomes that are described as already happening, such as suicides and economic crashes. In the example below, the user indicated that pandemics of any kind could result in economic crashes and wars. Thus, alerts of avian influenza and swine flu infections were particularly disturbing in the wake of the COVID-19 pandemic and its ensuing effect. The paraphrased quote below further explains this point:

> *Oh, no! Now there is pig flu in addition to covid? They are saying that influenza this year, will be bad. Pandemics can start economic disasters and economic issues can lead to conflict. The elites and globalists and others are creating a perfect storm." It's the same people who created the New World Order and are now using China to bring that "perfect storm" on Chinese people and affecting all of us.*

This post encapsulates multiple theories that proponents of misinformation disseminate. One is the idea that pandemics are created to intentionally destabilize the political and economic systems of the world. Amid the ensuing chaos, global hegemonic figures are portrayed as aiming to establish a new hierarchy, a New World Order. The transformed global landscape is generally framed in economic terms; dominant forces will gain tremendous wealth and concretize their power. In this instance, the misinformation posts deviate from systematic examinations of pandemic impacts by overwhelmingly focusing on the negative impacts of pandemics and a purported new world order, while rejecting the public health measures (such as vaccinations), which could help mitigate disease spread and therefore minimize the impact.

Anti-Vaccination

Anti-vaccination sentiments were the third most dominant topic at 15% (84/544). The example below demonstrates the majority of the reasons that users gave for rejecting vaccination, including the profiteering motivations driving pharmaceutical companies to manufacture vaccines and vaccines being allegedly used to

XSL•FO
RenderX

implement a depopulation agenda. Others asserted that the vaccine manufacturing process is largely experimental and that not enough tests are run before they are administered to humans. Another conspiracy theory was that vaccines are used to insert digital biometric identity mechanisms into people, which would then be used to enforce a digital economy.

> *This is not good! They will do anything for money. These companies and health care providers will specially target poor neighborhoods. Vaccines are hurting our children, they're toxic and made of chicken embryos and from bird flu, pigeon flu, avian influenza, H1N1 and other viruses. Stop using our kids for testing drugs! Wake up and stop it!*

A notable aspect of misinformative posts against vaccines was how proponents seek to legitimize their messages by adapting "seeds" of facts. For instance, some posts refer to the history of cell strains originally derived from human fetal tissue in the development of some vaccines (from text in an image): "What's in your needle?… measles virus…monkey kidneys…aborted fetal tissues…chicken embryo."

Elections/Democracy

Posts discussing the connection between H5N1 and topics related to elections or democracy made up 12% (66/544) of the subset. Some users emphasized the political ramifications of an avian influenza outbreak. They include a perceived attempt by liberal governments to use outbreaks to curtail people's rights and freedoms and a socialist takeover by China supported by the United States and other governments. These allegations were frequently presented in tandem with other conspiracy theories. Many posts thus contended that the attacks on democratic freedoms would be facilitated by the modified vaccines being used to digitally tag everyone. The perceived attacks on rights and freedoms are usually linked to people's rejection of social distancing, vaccination, and other mandates, which were instituted in many countries during the COVID-19 outbreak. Posts frequently included multiple conspiracy-related narratives connected to reinforce the claim. In some cases, even contradictory assertions are believed to be true at the same time. This example from our dataset revealed the convergence of an engineered pandemic, fake climate change, and attempts to sabotage elections, all of which are being allegedly implemented by globalists.

> *I'm actually worried about those who turned against their neighbors during the pandemic (while those who doubted all this were actually right); this is similar to how people snitched on their neighbors during wars. The same could happen when the next fake pandemic is created bird influenza? Climate change? The next political scheme and elections?*

Misinformation About the Food System

Among the actions that health authorities in Canada and internationally implemented to control H5N1 infections in farmed flocks are quarantines, restrictions on imports and exports, culling of birds, and disinfection of objects and premises. Another concern that was expressed in the dataset involved food security in relation to HPAI interventions such as depopulation of poultry farms. The posts below illustrate assertions about authorities' intentional use of avian influenza to cause food insecurity by creating insufficiency in meat, eggs, and other food supplies.

> *They are using fake results in PCR tests to kill animals and prepare for an Avian influenza. Plandemic and land misuse reduce the amount of food available…this is all to create a crisis and push people towards digital IDs and force them to take vaccinations and what's worst is that they will replace meat with fake products and insects. We have to fight; this doesn't support the corrupt system.*

The example demonstrates the above-discussed confluence of conspiracy theories, where proclamations of fabricated food crises feed into fears about digital identification, fossil fuels, and a new economic world order. Another post linking avian influenza to an agenda to advance insect and lab-grown meat diets included the following hashtags touting multiple theories since hashtags produce hypertexts [54]: #foodshortage #thegreatreset #depopulationagenda #newworldorderagenda #globalists #elites #climatechangehoax #clotshot #birdflu #populationcontrol #labgrownmeat #waronmeat #warmongers #billgatesisevil.

A final noteworthy theme under the food system is allegations that the Chinese government is gaining control over food supplies by unleashing H5N1 as a weapon of dominance. Misinformation on the food system was 11% (59/544) of the subset.

Religious Allusions and Prophecies

This category comprised approximately 5% (26/544). They all espoused Christian or Biblical allusions, including repentance from sin as an infection prevention method, apocalyptic prophecies on various catastrophes, which human societies will experience as the world nears the end of its lifespan, and vaccines being used to implant the mark of the beast. Chapter 13 of the Biblical Book of Revelation notes that during Earth's final days, a beast will assume global leadership and require everyone to receive an identifying mark that will enable them to engage in societal transactions. The marks will be placed either on foreheads or on right hands.

> *The prediction is about to happen; don't accept these tech implants that they want to put in your hand and head. They want to add online money to your body. This bird flu is part of the next fake pandemic to spread digital identification and currency.*

The above illustrates misinformative claims that vaccines contained materials such as nanochips that marked immunized people with the mark of the beast. A recurring aspect of the messages in this category was the correlations made between announcements forecasting disease events and Biblical prophecies. Over time, and especially with reports of infections in mammals such as cows, various health experts have sounded warnings about imminent avian flu pandemics. News stories on such statements have bolstered religious myths, with people claiming that they were fulfillments of prophetic proclamations as shown in the example below.

XSL•FO
RenderX

> *Remember that former public health official and what he said in this video. Remember when he mentioned that the next likely pandemic could be coming from bird flu?! He didn't confirm that it will be for sure, but let's keep an eye on this. We might see a biblical prophecy about wars and famines; these are the signs of the end times.* [#Matthew 24:7 KJV.]

Discussion

Principal Findings

The findings from this analysis of misinformation on 2 major social media platforms raise implications for public health policy and practice on the evolving H5N1 situation. Some evidence suggests that efforts to counter misinformation, such as debunking or fact-checking, may unintentionally amplify misinformation by elevating its perceived significance [32]. As the sample of posts analyzed represents a small portion of the initial dataset, our study does not aim to show prevalence on various platforms but to describe misinformation narratives and their types. The qualitative analysis of these narratives can demonstrate the construction and development of such narratives on this topic.

Importantly, findings suggest the need for the public health community to carefully consider such misinformation. Evidence shows that misinformation on social media has a differential diffusion pattern compared to traditional media [23]. Moreover, studies show that misinformation garners more engagement than facts. As Vosoughi et al [28] found, "falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information." The top 1% of the most popular false news was disseminated among 1000 to 100,000 people. Verifiable information, however, barely reached more than 1000 people. Misinformation had a 70% higher chance of being retweeted compared to facts. By contrast, the truth took about 6 times as long as false information to reach 1500 people. This more rapid and far-reaching diffusion pattern of misinformation is also reported by studies of medical large language models, which have been found to be vulnerable to "data poisoning" attacks [55]. Large language models, trained on web-scraped data, are increasingly being explored to support health care [56]. However, as Alber et al [47] wrote, "replacement of just 0.001% of training tokens with medical misinformation results in harmful models more likely to propagate medical errors". Thus, a deeper understanding is needed over time of their diffusion pattern and potential role in the spread of misinformation.

Our findings provide a baseline from which ongoing monitoring for trends in patterns of spread can be undertaken and lay the foundation for future work exploring misinformation on the HPAI virus in online spaces. For example, findings in this study about the use of scientific terminology in anti-vaccination posts showed similar points to the investigation by Faasse et al [57] of the dynamics of language use in pro- and anti-vaccination commentary on Facebook. They found that messages against vaccination used linguistic markers that resembled scientific language. Under the elections/democracy topic, posts analyzed reflected patterns noted in other studies regarding the belief in multiple conspiracy theories at the same time [36]. For instance, respondents in Miller's study asserted that COVID-19 was a bioweapon manufactured by China, and the United States accidentally released it into society. In the study by Hellsten et al [58] on X communication networks during avian influenza outbreaks in the Netherlands from 2015 to 2017, they found frequent instances of concurring themes such as the rejection of health regulations #kippendans (chicken dance), #vuurwerk (firework), and #jacht (hunting). Thus, a principal recommendation is that public health systems need to use early warning data of social and behavioral factors, including social media monitoring of misinformation, to inform the development of public health messaging.

Second, our systematic analysis of the content of the posts deemed misinformation suggests that claims about H5N1 are an extension of dominant narratives, by topics and targets, put forth during the COVID-19 pandemic and other issue areas such as climate change. This could be seen, for example, in patterns of blaming a foreign entity or minority communities, such as blaming the Chinese government for H5N1, for a major disease outbreak, which has occurred regularly in history. Examples found included the massacre of Jews during the Black Death of the 15th century and during The Great Plague and the Spanish Flu of 1918‐1919; the lesbian, gay, bisexual, transgender/transsexual, queer, and other minority sexual orientations and gender identities community during the global AIDS pandemic; and Asian people during the COVID-19 pandemic [59].

An overarching theme as well was mistrust of the motivations for public health messaging and interventions. Our findings suggest that views sustaining mistrust persist, framed around a broader narrative of control by powerful elites to achieve a new world order favorable to their interests [60]. H5N1 has become subsumed within this world view, with public health concerns about the virus attributed to self-interested "pandemic engineering." [61] Misinformation about the food system also demonstrated patterns of such mistrust in authorities following measures to control H5N1 infections such as bird depopulation. Bird depopulation can appear drastic because it involves both healthy and infected birds within a given perimeter. It has, however, shown to be the most effective method for stemming infection spread [8,62,63]. Misinformation against this measure typically has ominous labels such as murder and mass slaughtering. Additionally, while health authorities and governments have announced that H5N1 does not currently present food safety risks if animal products are appropriately processed or cooked [55,64], some misinformative posts called the safety of dairy products into question.

Another pattern of mistrust found was under religious allusions and prophecies topics, as users connected prominent public health communications, such as one made by the WHO Director-General [65], with religious narratives seen as the realization of prophetic predictions. This topic, in general, offers insights among certain religious groups that may resist mainstream information. Christian beliefs and those of other religious and insular communities are primarily grounded in faith rather than empirical evidence [66]. Such groups, particularly the fundamentalist ones, reject mainstream and

XSL•FO
RenderX

external information, reinforcing their worldviews and possibly a greater risk of accepting misinformation [67-69]. Therefore, the credibility of such religious declarations relies on people's acceptance of the parallels with Bible verses, which could reduce the need for expositions based on objective reality. Douglas further contends that the growing relationship between misinformation and religion is due in part to religious communities being principal targets [70]. This relates to the historical foundations of misinformation that was facilitated by Christian fundamentalists seeking alternative news sources as they rejected mainstream media's inclination toward what they deem as biased expert knowledge, which usually deviates from religious dogma [6].

Given this broader framing, rebuilding and protecting public trust in the state, market, and civil society may require efforts beyond the public health community. Instead, strategies to sustain public trust in health practitioners and scientists, which remains high in most countries [71], as sources of quality information can be supported. For example, "localizing" public health efforts by working in meaningful partnership with community leaders can "humanize" and thus increase trust in public health messaging [72].

Third, our findings demonstrate the importance of protecting information quality as the H5N1 situation evolves while balancing the need to not draw undue attention to misinformation. Unlike COVID-19, the public health community has been managing the risks from H5N1 since the late 1990s, so much is already known about the virus [73]. Since 2022, the spread to a growing number of mammal species including dairy farms has increased media attention and public awareness [2]. Evidence suggests that there is thus an opportunity for public health messaging to facilitate public understanding of this issue while it is still emerging. For example, Vos and Buckner [74] found that sensemaking plays a key role in the success of public health interventions during health crises. Sensemaking involves the way the public interprets and frames health information they receive about emerging health events. The misinformation identified in this paper suggests H5N1 is already interpreted using existing conspiracy-related worldviews (eg, New World Order). To counter misinformed sensemaking, there is a need to avoid information vacuums. Moreover, public health messaging must not only be factually accurate but in a form that facilitates sensemaking. Accordingly, there is a need to provide appropriate information for the public to understand, and not just be informed, about H5N1. The insights into online health misinformation on HPAI that this study provides are instrumental for supplementing surveillance structures, guiding communication mechanisms with an understanding of the misgivings by some, which could be fueling hesitation toward and rejection of interventions and ultimately bolstering pandemic readiness measures.

Importantly, the ability to monitor the content and volume of misinformation on social media, and proactively message about the evolving H5N1 situation, is likely to be more important yet more challenging amid changing practices by social media platforms to limit access to data and to protect information quality. One of the strengths of this study is the novel expansion of online health misinformation analysis beyond the textual content of X platform. Our multimodal analysis of textual and multimedia data from Instagram and Facebook broadens the scope of knowledge of online public narratives, particularly about infectious diseases. Despite Meta's claims about the strict moderation of posts,75 the majority of posts identified as misinformation were not flagged by both Facebook and Instagram for fact-checking. There were only 5 posts in our dataset that were flagged as inaccurate (less than 1%). The generic notification stated they were not completely factual, and they contained information similar to other posts that third-party fact checkers had reviewed. In 2018, CEO Mark Zuckerberg stated that Meta's research showed, "that no matter where we draw the lines for what is allowed, as a piece of content gets close to that line (demarcating prohibited content which contains misinformation and harmful material), people will engage with it more on average—even when they tell us afterwards they don't like the content" [76]. The announcement in January 2025 that Meta would end third-party fact-checking, in favor of allowing people to "express themselves freely," could have implications for protecting information quality [77]. Given the ubiquity of social media and its centrality in sustaining social connection, especially during times of uncertainty such as public health emergencies, the public health community can benefit from understanding and navigating this evolving landscape as a core component of future pandemic preparedness and response.

Limitations

A key limitation was the inability to include an analysis of the volume of misinformation on Meta platforms due to the lack of data collection tools. Moreover, we were unable to analyze the ideological leanings of users to further contextualize the narratives. We conducted a brief review of the profiles of users in our dataset to explore their potential ideological or political affiliations. Though some users' profiles suggested specific political ideologies, this analysis was not included in the study, as confirming such affiliations would require an ethnographic approach, which falls outside the scope of this paper.

Additionally, our paper was limited to English-Language and region-specific search terms that mostly covered North America, along with a focus on some US public figures such as Anthony Fauci. This could have excluded similar discussions from other regions and countries such as the United Kingdom and Germany, where different terms were used. Future research could address this by including additional relevant keywords from different regions such as "flockdown" and "raw milk." Another limitation was the prioritization of pandemic and vaccine-related terms in our dataset, which meant that other misinformation narratives about transmission and treatment, for example, may not have been captured. This limitation should be considered when examining our findings, as the topic categories identified were a result of our keyword search focus. Thus, findings reflect a specific focus of H5N1 misinformation.

Conclusions

Misinformation on HPAI on social media has implications for public health as recurrent avian influenza outbreaks in animal populations have contributed to fears about the pandemic potential of the disease for humans. Notably, H5N1 in human

XSL•FO
RenderX

populations is associated with several clinical symptoms; thus, decisive and swift actions are needed to stem any further infections. This finding underscores the importance of reviewing contextual factors like online misinformation, which could interfere with health interventions.

Our online investigation revealed that the targets of H5N1 misinformation narratives were largely people or entities who wield varying levels of political or other authority such as government and public health officials. These individuals received attention for their varying roles in health-related decision-making and in facilitating public health interventions.

We also found that the most prevalent misinformation themes were accusations of pandemic engineering, societal destruction, anti-vaccination sentiments, impact on elections or democracy, causing food insecurity, and religious allusions and prophecies. These assertions took various forms, including accusations that the virus itself did not occur naturally but claimed that the virus was artificially created as a bioweapon. The messages that contained anti-vaccination sentiments concluded that health interventions for avian influenza would include vaccines that would insert digital biometric identification into people, which could be used to either instigate population control or enforce a digital economy.

Acknowledgments

The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript.

Funding

This work was supported by the Bridge Research Consortium (BRC), part of Canada's Immuno-Engineering and Biomanufacturing Hub. The BRC is funded by the Canada Biomedical Research Fund and Biomedical Research Infrastructure Fund (grant: CBRF-2023-00122) and BC Knowledge Development Fund.

Data Availability

The dataset used in this manuscript can be shared with any interested scholar.

Conflicts of Interest

None declared.

References

1. Banks J, Speidel EC, Alexander DJ. Characterisation of an avian influenza A virus isolated from a human--is an intermediate host necessary for the emergence of pandemic influenza viruses? Arch Virol 1998;143(4):781-787. [doi: 10.1007/s007050050329] [Medline: 9638147]
2. Pereira HG, Tůmová B, Law VG. Avian influenza A viruses. Bull World Health Organ 1965;32(6):855-860. [Medline: 5294310]
3. Snowden FM. Epidemics and Society: From the Black Death to the Present: Yale University Press; 2019. [doi: 10.12987/9780300249149]
4. Avian influenza a(H5N1) virus. World Health Organization. 2024. URL: https://www.who.int/teams/global-influenza-programme/avian-influenza/avian-a-h5n1-virus [accessed 2026-02-20]
5. 2020-2024 highlights in the history of avian influenza (bird flu) timeline. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/bird-flu/avian-timeline/2020s.html [accessed 2026-02-20]
6. Ridgely SB. Conservative Christianity and the creation of alternative news: an analysis of Focus on the Family's multimedia empire. Relig Am Cult 2020;30(1):1-25. [doi: 10.1017/rac.2020.1]
7. Rust S. USDA says it will start testing milk for H5N1 bird flu virus nationwide. Los Angeles Times. 2024 Dec 6. URL: https://www.latimes.com/environment/story/2024-12-06/usda-says-it-will-start-testing-milk-for-h5n1-nationwide [accessed 2026-02-20]
8. Minister bibeau meets with stakeholders from the agricultural sector to discuss the impacts of highly pathogenic avian influenza. Canada.ca. 2022 Apr 29. URL: https://www.canada.ca/en/food-inspection-agency/news/2022/04/readout-from-minister-bibeau-minister-of-agriculture-and-agri-food-on-the-impacts-of-highly-pathogenic-avian-influenza.html [accessed 2026-02-20]
9. Tian H, Xu B. Persistence and transmission of avian influenza A (H5N1): virus movement, risk factors and pandemic potential. Ann GIS 2015 Jan 2;21(1):55-68. [doi: 10.1080/19475683.2014.992368]
10. Clark EC, Neumann S, Hopkins S, Kostopoulos A, Hagerman L, Dobbins M. Changes to public health surveillance methods due to the COVID-19 pandemic: scoping review. JMIR Public Health Surveill 2024 Jan 19;10:e49185. [doi: 10.2196/49185] [Medline: 38241067]

XSL•FO
RenderX

11. Lewandowsky S. The 'post-truth' world, misinformation, and information literacy: a perspective from cognitive science. In: Informed Societies why Information Literacy Matters for Citizenship, Participation and Democracy: Facet Publishing; 2020:69-87. [doi: 10.29085/9781783303922]
12. Shahbazi M, Bunker D. Social media trust: fighting misinformation in the time of crisis. Int J Inf Manage 2024 Aug;77:102780. [doi: 10.1016/j.ijinfomgt.2024.102780]
13. KFF tracking poll on health information and trust: January 2025. KFF. 2025 Jan 28. URL: https://www.kff.org/health-information-trust/kff-tracking-poll-on-health-information-and-trust-january-2025/ [accessed 2026-02-20]
14. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021 Jan 20;23(1):e17187. [doi: 10.2196/17187] [Medline: 33470931]
15. Calma J. Twitter just closed the book on academic research. The Verge. 2023 May 31. URL: https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research [accessed 2026-02-20]
16. Ortutay B. Meta kills off misinformation tracking tool crowdtangle despite pleas from researchers, journalists. Associated Press News. 2024 Aug 14. URL: https://apnews.com/article/meta-crowdtangle-research-misinformation-shutdown-facebook-977ece074b99adddb4887bf719f2112a [accessed 2026-02-20]
17. Liu J, Kassas B, Lai J, Kropp J, Gao Z. Understanding the role of risk preferences and perceptions in vaccination decisions and post-vaccination behaviors among U.S. households. Sci Rep 2024;14(1):3190. [doi: 10.1038/s41598-024-52408-6]
18. Westerman D, Spence PR, Van Der Heide B. Social media as information source: recency of updates and credibility of information. J Comput-Mediat Comm 2014 Jan;19(2):171-183. [doi: 10.1111/jcc4.12041]
19. Yang Y, Zhang Y, Xiang A. Information interaction and social support: exploring help-seeking in online communities during public health emergencies. BMC Public Health 2023 Jun 27;23(1):1250. [doi: 10.1186/s12889-023-16151-3] [Medline: 37370074]
20. Disease outbreak news (dons). World Health Organization. URL: https://www.who.int/emergencies/disease-outbreak-news [accessed 2026-02-20]
21. Al-Rawi A. Gatekeeping fake news discourses on mainstream media versus social media. Soc Sci Comput Rev 2019 Dec;37(6):687-704. [doi: 10.1177/0894439318795849]
22. A multi-dimensional approach to disinformation: report of the independent high level group on fake news and online disinformation (3). : European Commission; 2018 URL: https://www.ami.info/wp-content/uploads/2018/03/HLEGReportonFakeNewsandOnlineDisinformation.pdf [accessed 2026-02-20]
23. Chen S, Xiao L, Kumar A. Spread of misinformation on social media: what contributes to it and how to combat it. Comput Human Behav 2023 Apr;141:107643. [doi: 10.1016/j.chb.2022.107643]
24. Lewandowsky S, Cook J, Schmid P, et al. The COVID-19 Vaccine Communication Handbook: A Practical Guide for Improving Vaccine Communication and Fighting Misinformation: SciBeh; 2021. URL: https://www.immunize.ca/sites/default/files/Resource%20and%20Product%20Uploads%20(PDFs)/COVID-19/final-COVID19-vaccinehandbook-4-production.pdf [accessed 2026-02-20]
25. Song J, Song TM, Seo DC, Jin DL, Kim JS. Social big data analysis of information spread and perceived infection risk during the 2015 Middle East respiratory syndrome outbreak in South Korea. Cyberpsychol Behav Soc Netw 2017 Jan;20(1):22-29. [doi: 10.1089/cyber.2016.0126] [Medline: 28051336]
26. Gibson K. Meta to end fact-checking, replacing it with community-driven system akin to elon musk's x. CBS News. 2025 Jan 8. URL: https://www.cbsnews.com/news/meta-facebook-instagram-fact-checking-mark-zuckerberg/ [accessed 2026-02-20]
27. Broniatowski DA, Jamison AM, Qi S, et al. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. Am J Public Health 2018 Oct;108(10):1378-1384. [doi: 10.2105/AJPH.2018.304567] [Medline: 30138075]
28. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science 2018 Mar 9;359(6380):1146-1151. [doi: 10.1126/science.aap9559] [Medline: 29590045]
29. Wang Y, McKee M, Torbica A, Stuckler D. Systematic literature review on the spread of health-related misinformation on social media. Soc Sci Med 2019 Nov;240:112552. [doi: 10.1016/j.socscimed.2019.112552] [Medline: 31561111]
30. Adebimpe WO, Adeyemi DH, Faremi A, Ojo JO, Efuntoye AE. The relevance of the social networking media in Ebola virus disease prevention and control in Southwestern Nigeria. Pan Afr Med J 2015;22(Supp 1):7. [doi: 10.11604/pamj.supp.2015.22.1.6165]
31. Smith TO. A little birdie told me: H1N1 information and misinformation exchange on twitter. Social Science Research Network. 2010. URL: https://papers.ssrn.com/abstract=1739306 [accessed 2026-02-20]
32. Chowdhury N, Khalid A, Turin TC. Understanding misinformation infodemic during public health emergencies due to large-scale disease outbreaks: a rapid review. Z Gesundh Wiss 2023 Apr;31(4):553-573. [doi: 10.1007/s10389-021-01565-3]
33. Yan Q, Shan S, Zhang B, et al. Monitoring the relationship between social network status and influenza based on social media data. Disaster Med Public Health Prep 2023 Sep 18;17:e490. [doi: 10.1017/dmp.2023.117] [Medline: 37721020]
34. Gori D, Reno C, Remondini D, Durazzi F, Fantini MP. Are we ready for the arrival of the new COVID-19 vaccinations? Great promises and unknown challenges still to come. Vaccines (Basel) 2021 Feb 18;9(2):173. [doi: 10.3390/vaccines9020173] [Medline: 33670697]

XSL•FO
RenderX

35. Enders AM, Uscinski JE, Seelig MI, et al. The relationship between social media use and beliefs in conspiracy theories and misinformation. Polit Behav 2023;45(2):781-804. [doi: 10.1007/s11109-021-09734-6] [Medline: 34248238]
36. Miller JM. Do COVID-19 conspiracy theory beliefs form a monological belief system? Can J Pol Sci 2020 Jun;53(2):319-326. [doi: 10.1017/S0008423920000517]
37. Motta M, Stecula D, Farhart C. How right-leaning media coverage of COVID-19 facilitated the spread of misinformation in the early stages of the pandemic in the US. Can J Pol Sci 2020 Jun;53(2):335-342. [doi: 10.1017/S0008423920000396]
38. Choi DH, Yoo W, Noh GY, Park K. The impact of social media on risk perceptions during the MERS outbreak in South Korea. Comput Human Behav 2017 Jul;72:422-431. [doi: 10.1016/j.chb.2017.03.004] [Medline: 32288176]
39. Veil S, Reynolds B, Sellnow TL, Seeger MW. CERC as a theoretical framework for research and practice. Health Promot Pract 2008 Oct;9(4 Suppl):26S-34S. [doi: 10.1177/1524839908322113] [Medline: 18936257]
40. Ahmed W, Bath PA, Sbaffi L, Demartini G. Novel insights into views towards H1N1 during the 2009 Pandemic: a thematic analysis of Twitter data. Health Info Libr J 2019 Mar;36(1):60-72. [doi: 10.1111/hir.12247] [Medline: 30663232]
41. Qin Z, Ronchieri E. Exploring pandemics events on Twitter by using sentiment analysis and topic modelling. Appl Sci (Basel) 2022;12(23):11924. [doi: 10.3390/app122311924]
42. Soltani M, Sharif S, Dara R. Analyzing social media discourse of avian influenza outbreaks. Nat Lang Process J Sep 2025 Sep;12(1):100176. [doi: 10.1016/j.nlp.2025.100176]
43. Brainard J, Hunter PR. Misinformation making a disease outbreak worse: outcomes compared for influenza, monkeypox, and norovirus. Simulation 2019 Apr;96(4):365-374. [doi: 10.1177/0037549719885021] [Medline: 34285423]
44. Fault lines: the expert panel on the socioeconomic impacts of science and health misinformation. : Council of Canadian Academies; 2023 URL: https://cca-reports.ca/wp-content/uploads/2023/02/Report-Fault-Lines-digital.pdf [accessed 2026-02-20]
45. Online discourse about bird flu spikes in the US and globally. Public Health Communications Collaborative. 2024 Apr 30. URL: https://publichealthcollaborative.org/alerts/online-discourse-about-bird-flu-spikes-in-the-u-s-and-globally/ [accessed 2026-02-20]
46. Dixon SJ. Most popular social networks worldwide as of february 2025, ranked by number of monthly active users (in millions). Statista. 2024 Jul 10. URL: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ [accessed 2026-02-20]
47. Alber DA, Yang Z, Alyakin A, et al. Medical large language models are vulnerable to data-poisoning attacks. Nat Med 2025 Feb;31(2):618-626. [doi: 10.1038/s41591-024-03445-1] [Medline: 39779928]
48. Al-Rawi A. Script for extracting most frequent words, phrases, and sentiments. GitHub. 2025 Sep 23. URL: https://gist.github.com/ahmedalrawi/c5879cfc52342d3ed778c35c755a42ad [accessed 2026-02-20]
49. Al-Rawi A, Jamieson K. Characterizing the gendered Twitter discussion of COVID-19 hoax. Health Commun 2023 Dec;38(14):3366-3375. [doi: 10.1080/10410236.2022.2149112] [Medline: 36411526]
50. Lanier HD, Diaz MI, Saleh SN, Lehmann CU, Medford RJ. Analyzing COVID-19 disinformation on Twitter using the hashtags #scamdemic and #plandemic: Retrospective study. PLOS ONE 2022;17(6):e0268409. [doi: 10.1371/journal.pone.0268409] [Medline: 35731785]
51. Current situation: bird flu in dairy cows. Centers for Disease Control and Prevention. 2025 Jul 7. URL: https://www.cdc.gov/bird-flu/situation-summary/mammals.html [accessed 2026-02-20]
52. Thomas DR. A general inductive approach for analyzing qualitative evaluation data. Am J Eval 2006 Jun;27(2):237-246. [doi: 10.1177/1098214005283748]
53. Krippendorff K. Reliability in content analysis: some common misconceptions and recommendations. Hum Commun Res 2004 Jul 1;30(3):411-433. [doi: 10.1111/j.1468-2958.2004.tb00738.x]
54. Albu OB, Etter M. Hypertextuality and social media: a study of the constitutive and paradoxical implications of organizational Twitter use. Manag Commun Q 2016;30(1):5-31. [doi: 10.1177/0893318915601161]
55. Avian influenza. European Food Safety Authority. URL: https://www.efsa.europa.eu/en/topics/topic/avian-influenza [accessed 2026-02-20]
56. Nazi ZA, Peng W. Large language models in healthcare and medical domain: a review. Informatics (MDPI) 2024;11(3):57. [doi: 10.3390/informatics11030057]
57. Faasse K, Chatman CJ, Martin LR. A comparison of language use in pro- and anti-vaccination comments in response to a high profile Facebook post. Vaccine 2016 Nov 11;34(47):5808-5814. [doi: 10.1016/j.vaccine.2016.09.029] [Medline: 27707558]
58. Hellsten I, Jacobs S, Wonneberger A. Active and passive stakeholders in issue arenas: a communication network approach to the bird flu debate on Twitter. Public Relat Rev 2019 Mar;45(1):35-48. [doi: 10.1016/j.pubrev.2018.12.009]
59. Davis M. The Monster Enters: COVID-19, Avian Flu, and the Plagues of Capitalism: Verso Books; 2022. URL: https://www.versobooks.com/en-gb/products/2753-the-monster-enters?srsltid=AfmBOopdZFr6YRDmmIZ5n5n_l0dqSjtb1xi6kdqQGA4CpnINQ5y6bNKV [accessed 2026-02-20]
60. Gallagher A, O'Connor C. The 'great reset'. Institute for Strategic Dialogue. 2023 Mar 14. URL: https://www.isdglobal.org/isd-explainer/the-great-reset/ [accessed 2026-02-20]

XSL•FO
RenderX

61. Kilander G. Former director of the CDC predicts the next pandemic will be from bird flu. The Independent. 2024 Jun 17. URL: https://www.independent.co.uk/news/world/americas/bird-flu-cdc-covid-pandemic- b2563711.html [accessed 2026-02-20]
62. Backer JA, van Roermund HJW, Fischer EAJ, van Asseldonk M, Bergevoet RHM. Controlling highly pathogenic avian influenza outbreaks: an epidemiological and economic model analysis. Prev Vet Med 2015 Sep 1;121(1-2):142-150. [doi: 10.1016/j.prevetmed.2015.06.006] [Medline: 26087887]
63. Martcheva M. Avian flu: modeling and implications for control. J Biol Syst 2014 Mar;22(1):151-175. [doi: 10.1142/S0218339014500090]
64. Status of ongoing avian influenza response by province. Canadian Food Inspection Agency. 2024 Jun 26. URL: http://inspection.canada.ca/en/animal-health/terrestrial-animals/diseases/reportable/avian-influenza/latest-bird-flu-situation/status-ongoing-avian-influenza-response [accessed 2026-02-20]
65. Disinformation and public health. World Health Organization. 2024 Feb 6. URL: https://www.who.int/news-room/questions-and-answers/item/disinformation-and-public-health [accessed 2026-02-20]
66. Nieminen P, Loikkanen J, Ryökäs E, Mustonen AM. Nature of evidence in religion and natural science. Theol Sci 2020 Jul 2;18(3):448-474. [doi: 10.1080/14746700.2020.1786221]
67. Chauchard S, Badrinathan S. The Religious Roots of Belief in Misinformation: Experimental Evidence from India: Cambridge University Press; 2025. [doi: 10.1017/S0007123425100616]
68. Jedinger A, Siegers P. Religion, Spirituality, and Susceptibility to Conspiracy Theories: Examining the Role of Analytic Thinking and Post-Critical Beliefs: Cambridge University Press; 2024. [doi: 10.1017/S1755048324000130]
69. Łowicki P, Marchlewska M, Molenda Z, Karakula A, Szczepańska D. Does religion predict coronavirus conspiracy beliefs? Centrality of religiosity, religious fundamentalism, and COVID-19 conspiracy beliefs. Pers Individ Dif 2022 Mar;187:111413. [doi: 10.1016/j.paid.2021.111413] [Medline: 34866726]
70. Douglas C. Religion and fake news: faith-based alternative information ecosystems in the US and Europe. Rev Faith Int. Aff 2018 Jan 2;16(1):61-73. [doi: 10.1080/15570274.2018.1433522]
71. 2025 edelman trust barometer special report: trust and health. : Edelman Trust Institute; 2025 URL: https://www.edelman.com/sites/g/files/aatuss191/files/2025-04/2025%20Edelman%20Trust%20Barometer%20Trust%20and%20Health%20FINAL.pdf [accessed 2026-02-20]
72. Sundelson AE, Jamison AM, Huhn N, Pasquino SL, Sell TK. Fighting the infodemic: the 4 i framework for advancing communication and trust. BMC Public Health 2023 Aug 30;23(1):1662. [doi: 10.1186/s12889-023-16612-9] [Medline: 37644563]
73. Alexander DJ. A review of avian influenza in different bird species. Vet Microbiol 2006;74(1–2):3-13. [doi: 10.1016/S0378-1135(00)00160-7]
74. Vos SC, Buckner MM. Social media messages in an emerging health crisis: tweeting bird flu. J Health Commun 2016;21(3):301-308. [doi: 10.1080/10810730.2015.1064495] [Medline: 26192209]
75. Misinformation. Meta. 2025. URL: https://transparency.fb.com/policies/community-standards/misinformation [accessed 2026-02-20]
76. Zuckerberg M. A blueprint for content governance and enforcement. Facebook.com. 2018. URL: https://opencasebook.org/casebooks/270-governing-digital-technology-spring-2020/resources/4.1.1.1-a-blueprint-for-content-governance-and-enforcement-by-mark-zuckerberg-facebook-2018/ [accessed 2026-02-20]
77. Kaplan J. More speech, fewer mistakes. Meta. 2025 Jan 7. URL: https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/ [accessed 2026-02-20]

Abbreviations

CDC: Centers for Disease Control and Prevention
HPAI: highly pathogenic avian influenza
WHO: World Health Organization

Edited by A Mavragani; submitted 28.Feb.2025; peer-reviewed by E Lau, R Tarlinton; revised version received 06.Jan.2026; accepted 28.Jan.2026; published 09.Apr.2026.

Please cite as:
Al-Rawi A, Ackah B, Fakida A, Lee K
Misinformation in Social Media Narratives on Highly Pathogenic Avian Influenza: Systematic Content Analysis of Facebook and Instagram Posts
J Med Internet Res 2026;28:e73275
URL: https://www.jmir.org/2026/1/e73275
doi:10.2196/73275

XSL•FO
RenderX

© Ahmed Al-Rawi, Betty Ackah, Abdelrahman Fakida, Kelley Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patient Perceptions of Ozempic (Semaglutide) for Weight Loss: Mixed Methods Analysis of Online Medication Reviews

Abanoub J Armanious[1,2], BS; Rachel-Mae Hunter[1,2], BS; Kristi R Griffiths[3], PhD; Hannah E Bowrey[4,5], PhD; Robyn M Brown[6], PhD; Morgan H James[1,2,4,5], PhD

[1]Department of Psychiatry, Rutgers Robert Wood Johnson Medical School, Piscataway, NJ, United States

[2]Rutgers Addiction Research Center, Brain Health Institute, Rutgers, The State University of New Jersey, Piscataway, NJ, United States

[3]InsideOut Institute, Faculty of Medicine and Health, The University of Sydney and Sydney Local Health District, Sydney, Australia

[4]School of Psychology, Faculty of Science, The University of Sydney, Camperdown, Australia

[5]Brain and Mind Centre, The University of Sydney, Camperdown, Australia

[6]Department of Biochemistry and Pharmacology, The University of Melbourne, Parkville, Australia

Corresponding Author:
Morgan H James, PhD
School of Psychology, Faculty of Science
The University of Sydney
Camperdown
Australia
Phone: 61 293510774
Email: morgan.james@sydney.edu.au

Abstract

Background: Ozempic (semaglutide) has received widespread attention for its appetite-suppressing effects, leading to extensive off-label use for weight loss. Although gastrointestinal side effects are well documented, less is known about how users assess the trade-off between perceived benefits and adverse effects, or how these assessments influence treatment discontinuation. Importantly, existing insights are often limited to clinical trial populations and may not fully reflect real-world experiences.

Objective: This study applies a novel infoveillance approach to examine patient-reported experiences with off-label Ozempic use for weight loss and to identify the factors most strongly associated with user satisfaction and treatment discontinuation.

Methods: We analyzed 60 publicly available, self-selected, anonymous user reviews of Ozempic from Drugs.com. Reviews were initially examined using thematic analysis to identify key themes describing patients' lived experiences with treatment. These qualitative themes were then linked to user-provided ratings of perceived drug efficacy (1-10 scale) and statements regarding intent to continue or discontinue treatment. This mixed methods approach enabled the integration of qualitative depth with quantitative patterns within naturally occurring, deidentified online data.

Results: Three major themes emerged from the thematic analysis: (1) change in body weight and appetite, (2) nonweight-related symptoms and side effects, and (3) plans for ongoing use versus discontinuation. Two-thirds of respondents reported reduced appetite, food cravings, or body weight. Gastrointestinal complaints were common (reported by 37 of 60, 62%, reviewers) but did not significantly (P=.39) influence satisfaction ratings or decisions to continue treatment. Instead, minimal/no weight loss and the emergence of nongastrointestinal side effects were more frequently associated with low overall satisfaction and discontinuation. Effective weight loss, even when accompanied by gastrointestinal side effects, was associated with a greater willingness to continue Ozempic treatment.

Conclusions: This study presents a novel application of infoveillance methods to characterize real-world patient attitudes toward off-label Ozempic use. Satisfaction was driven primarily by perceived effectiveness rather than tolerability. Key limitations are the self-selected nature of the sample, reliance on anonymous, self-reported data, and the lack of demographic, dosing, or treatment-duration information. Nonetheless, these findings underscore the value of online health forums as a rich and underutilized source of patient-centered insights to inform obesity treatment strategies, adherence interventions, and public health communication.

(J Med Internet Res 2026;28:e78391) doi:10.2196/78391

KEYWORDS

Thematic analysis; discussion board; pharmacotherapy; glucagon like peptide 1; GLP-1; incretin mimetic; craving.

XSL•FO
RenderX

Introduction

Background

The prevalence of obesity has more than doubled since 1990, contributing to a rise in chronic diseases associated with higher body weight, including type 2 diabetes (T2D) and cardiovascular diseases [1]. Consequently, the need for effective interventions to address obesity is urgent. Although lifestyle interventions (diet and exercise) are considered first-line strategies for weight management, they are often ineffective in the long term [2,3]. The few pharmacotherapies approved for treating overweight and obesity have historically produced only modest results, typically achieving 5%-10% weight loss [4]. Moreover, although bariatric surgery is effective for many patients, it carries significant morbidity and mortality risks [5,6]. Accordingly, the advent of glucagon-like peptide-1 receptor (GLP-1R) agonists, which are highly effective in promoting substantial weight loss and improving related health outcomes, has dramatically redefined the treatment landscape for these conditions.

Semaglutide, an incretin mimetic, is a GLP-1R agonist that stabilizes blood glucose by stimulating insulin secretion and inhibiting glucagon production [7]. It delays gastric emptying and influences appetite-regulating neural pathways, increasing satiety and reducing food intake in some individuals [8]. The first FDA-approved formulation of semaglutide was Ozempic, a once-weekly subcutaneous injection for the management of T2D [9]. This regimen offered the convenience of less frequent dosing compared with the previously approved incretin mimetic liraglutide, which required daily injections [7,10,11]. A substantial body of clinical trial data now indicates that, compared with placebo, Ozempic significantly reduces body weight (7.9%-17.3%) [12-16], lowers HbA_{1c} (glycated hemoglobin) levels, waist circumference, and systolic blood pressure, and improves overall physical functioning [13,14,16]. In clinical trials, improvements in these outcomes are generally observed within 3 months of initiating treatment [14,16-18]. In light of these findings, there has been intense interest in the off-label use of Ozempic for cosmetic weight loss, fueled in part by its popularization in mainstream and social media [19], despite other semaglutide formulations, such as Wegovy, being specifically approved for the treatment of obesity [20].

Despite their efficacy in promoting weight loss, semaglutide treatments are associated with several adverse events that vary in severity. Gastrointestinal complaints are the most common, with prevalence ranging from 41.9% to 82.8% (more common at higher doses), and include symptoms such as nausea, vomiting, constipation, and diarrhea [12-16]. Less frequent adverse events include headache, allergic reactions, and gallbladder-related disorders [15]. Adverse events are a major contributor to nonadherence to weight loss medications, including GLP-1R agonists [21-23]. In the SUSTAIN-6 clinical trial of injectable semaglutide, 22.6% of patients discontinued treatment prematurely during the 24-month study period. However, “real-world” discontinuation rates appear to be much higher. One study reported a 12-month discontinuation rate of 33% following initiation of once-weekly Ozempic therapy [24], while another study, which did not differentiate between different forms of injectable GLP-1R agonists, reported that 70.1% of patients discontinued treatment within 24 months [25].

Thus, among individuals using Ozempic for weight loss, there may be a conflict between experiencing the desired effects of treatment and managing its associated side effects. Understanding how patients navigate this trade-off is essential for optimizing adherence and maximizing therapeutic outcomes.

Aims

Here, we employed a mixed methods infoveillance approach to examine attitudes toward Ozempic among individuals with lived experience of off-label use. Specifically, we conducted a thematic analysis of user-generated reviews posted on Drugs.com [26], followed by quantitative modeling to assess how emergent themes were associated with perceived efficacy and treatment discontinuation. A major advantage of this approach is that it is not constrained by the a priori hypotheses typical of traditional quantitative designs, allowing for the emergence of unsolicited insights that might otherwise be overlooked [27]. As a form of infodemiological research, this study illustrates how publicly available, user-generated data can serve as a powerful resource for capturing patient-centered perspectives on medication effectiveness, tolerability, and real-world barriers to adherence. Despite the inherent limitations of using self-selected, anonymous online data, these findings provide unique and timely insight into how individuals evaluate the benefits and drawbacks of off-label Ozempic use for weight loss.

Methods

Data Collection

Data consisted of reviews of Ozempic submitted to Drugs.com, a website that provides peer-reviewed and independent information on more than 24,000 prescription drugs, over-the-counter medicines, and natural products. A unique feature of Drugs.com is its platform that allows members of the public to submit open-ended reviews and quantitative ratings of specific medications, enabling analysis of how these outcomes are related. We extracted data exclusively from respondents who selected “weight loss” as the condition for which they were using Ozempic, despite this not being an approved indication. Notably, Drugs.com has since removed “weight loss” as an option (current options now include T2D, cardiovascular risk reduction, and chronic kidney disease), making these data particularly valuable for capturing experiences of individuals using Ozempic specifically for weight loss. Using a display name, respondents are prompted to “comment on your experience with Ozempic” and are encouraged to “describe how the medication helped (or why it didn’t work); the benefits, adverse events, dosage, ease of use” in a single textbox. No demographic data are collected. Respondents can also provide a quantitative rating of the drug on a scale from 1 (not effective) to 10 (most effective) and indicate the duration of medication use.

Data were downloaded in June 2023. No retrospective time limit on reviews was imposed; the oldest review dated from

XSL•FO
RenderX

February 2023, and the most recent from June 2023. User reviews of Ozempic in which weight loss was listed as the primary indication were extracted, yielding a total of 78 reviews. As described below, a total of 60 reviews were analyzed before reaching thematic saturation.

This study employed a sequential mixed methods design, integrating qualitative and quantitative analyses of user-generated data. In the first phase, an inductive thematic analysis was conducted to identify patterns and themes within users' open-ended narratives describing their experiences with Ozempic. In the second phase, these emergent themes were quantitatively examined in relation to users' numerical satisfaction ratings. This design was selected to provide complementary insights, capturing the contextual richness of lived experience while enabling empirical assessment of the factors most strongly associated with user satisfaction and treatment discontinuation. This approach aligns with our previously published work [27]. We acknowledge that qualitative interpretation is inherently influenced by coder perspectives; measures taken to minimize and reflect on this potential bias are detailed in the "Thematic Data Analysis" section.

Ethical Considerations

This study analyzed secondary, publicly available, deidentified user reviews from Drugs.com. No interaction with users occurred, and no direct or indirect identifiers were collected or stored. Data-minimization procedures were applied: only text necessary for analysis was retained, quotes were screened to exclude potentially identifying details, and results are reported in aggregate wherever possible. In accordance with institutional guidance for research using publicly accessible data, this project was not subject to human participant review.

We acknowledge that, despite deidentification and public availability, online health narratives may still pose residual privacy risks (eg, potential reidentification via unique combinations of details) and may reflect audience expectations. Our use of these data was limited to analytic purposes, with careful curation of verbatim quotations and deliberate avoidance of stigmatizing language.

Thematic Data Analysis

Qualitative data analysis was conducted using NVivo 14 software (Lumivero, LLC). Data were analyzed using a thematic analysis approach, as initially outlined by Braun and Clarke [28], following a procedure we have described previously [27] and similar to those used in studies employing alternative online data sources [29,30]. Briefly, themes were generated through an iterative process of reading through each review, suggesting themes, re-reading, and comparing categories across multiple cycles of analysis (Figure 1). To facilitate this process, the dataset was randomly divided into batches of 15 reviews. Each batch was independently reviewed by 2 coders (AJA and RMH), and excerpts relevant to the research question were coded according to a data-driven, "bottom-up" principle. This approach minimized the influence of any preconceived ideas the reviewers might have had about respondent perceptions of Ozempic. After each set of 15 reviews, both coders met with a third-party noncoder (MHJ) to compare identified codes against the original data and with each other, ensuring that the codes were coherent, consistent, and distinctive. Thematic saturation was reached when 2 consecutive batches of 15 reviews yielded no new codes or subthemes. Saturation was confirmed by consensus among the 2 coders (AJA and RMH) and the independent reviewer (MHJ), consistent with the reflexive and inductive approach to thematic analysis described previously [27,28]. The initial analysis yielded 34 distinct coding categories, which were subsequently grouped and refined into 3 overarching themes. There were no predefined criteria for determining what constituted a separate theme; rather, meaningful clusters of codes were identified, reviewed, and iteratively refined.

It is important to acknowledge that thematic coding may have been influenced by the positions and potential biases of the authors. At the time of coding, AJA (male) was a post-baccalaureate researcher (BSc with concentrations in Cell Biology and Neuroscience, Public Health, and Religion), and RMH (female) was an undergraduate student majoring in Biological Sciences on the predental medicine track. Both were conducting laboratory research on the neurobiological basis of eating disorders. MHJ (male) was a researcher with expertise in the neurobiology of motivated behaviors, including feeding. To minimize potential coder bias, the coding team underwent structured training and employed standardized procedures for codebook development. Coders met regularly to review emerging codes, reconcile discrepancies, and discuss interpretations with the independent reviewer (MHJ) throughout the analytic process.

XSL•FO
RenderX

Figure 1. Flowchart outlining the process adopted to carry out the qualitative analysis portion of the study. First, reviews of Ozempic were extracted from Drugs.com and then randomly batched into groups of 15, with the first batch undergoing thorough familiarization via reading and rereading by the coders. Next, initial coding was conducted using NVivo 14 software, with subsequent validation by a noncoder. Similar analyses were carried out on a second batch of reviews; this process was repeated until the coders and non-coder agree that analysis of an additional batch of 15 reviews was unlikely to result in the identification of additional unique codes (i.e. thematic saturation). Codes were then organized into themes and sub-themes, which underwent iterative review and refinement. Finally, representative quotes were selected to illustrate each of the subthemes.

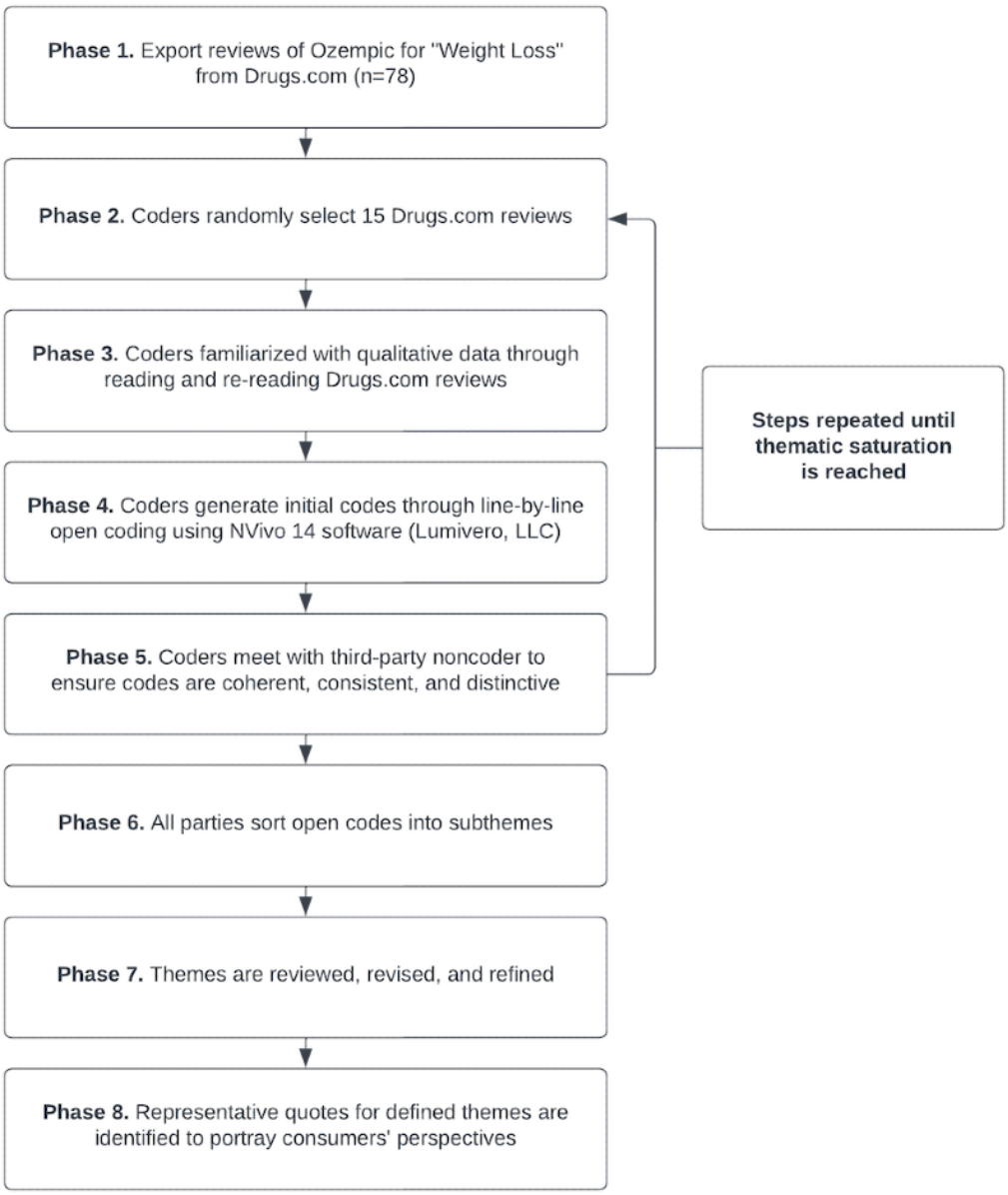

Quantitative Analysis of User Reviews

We also sought to examine how the qualitative themes identified through thematic analysis related to the quantitative rating scores (1-10 scale) provided by respondents. Of the 60 reviews analyzed, 54 included an associated quantitative rating. We calculated the median user rating for participants whose reviews contributed to each subtheme. Based on a frequency histogram of rating scores, the data were divided into 2 groups on either side of the median score (7.5), effectively creating clusters of “higher ratings” (n=27) and “lower ratings” (n=27; see Figure 2A). A median split was used because user ratings were bimodally distributed, with most responses clustering at the extreme values (1 and 10). In this context, the median provided a more robust and meaningful threshold than the mean, enabling clearer categorical comparisons between users who expressed generally positive versus negative appraisals of Ozempic. We plotted the proportions of reviews mentioning each subtheme in the higher- and lower-rating groups, reflecting their relative frequencies within each group. For visualization purposes, each subtheme was classified according to the predominant sentiment expressed by respondents: “positive” (eg, treatment-associated weight loss), “neutral” (eg, injection process), or “negative” (eg, nausea and gastrointestinal complaints; see Figure 2B). Separate chi-square tests were conducted to compare the frequency of each subtheme’s representation between higher- and lower-rating groups. A 2-sided test with a type I error rate of 0.05 was used for all analyses.

XSL•FO
RenderX

Figure 2. A) Histogram depicting the frequency of respondents' quantitative ratings of Ozempic on a 1-10 scale. Data were bimodal, with the most frequent scores being 1 and 10. For subsequent analyses, we divided respondents into those who provided ratings above or below the median score of 7.5 ('higher' vs. 'lower' ratings). B) Respondents who provided higher quantitative ratings of Ozempic (8-10 out of 10) were more likely to have also contributed to subthemes associated with positive sentiment (as described by the respondents), including 'weight loss' and 'appetite suppression'. Respondents who provided lower ratings (1-7 out of 10) were more likely to have contributed to several more negatively valanced subthemes, including 'minimal/no weight loss' 'other physiological (nongastrointestinal) symptoms,' and 'plans to discontinue treatment'. Comparisons between higher vs. lower ratings made using χ^2 analyses. *P<0.05, **P<0.01. Numbers/letters in parentheses reflect the subthemes described in the Results section.

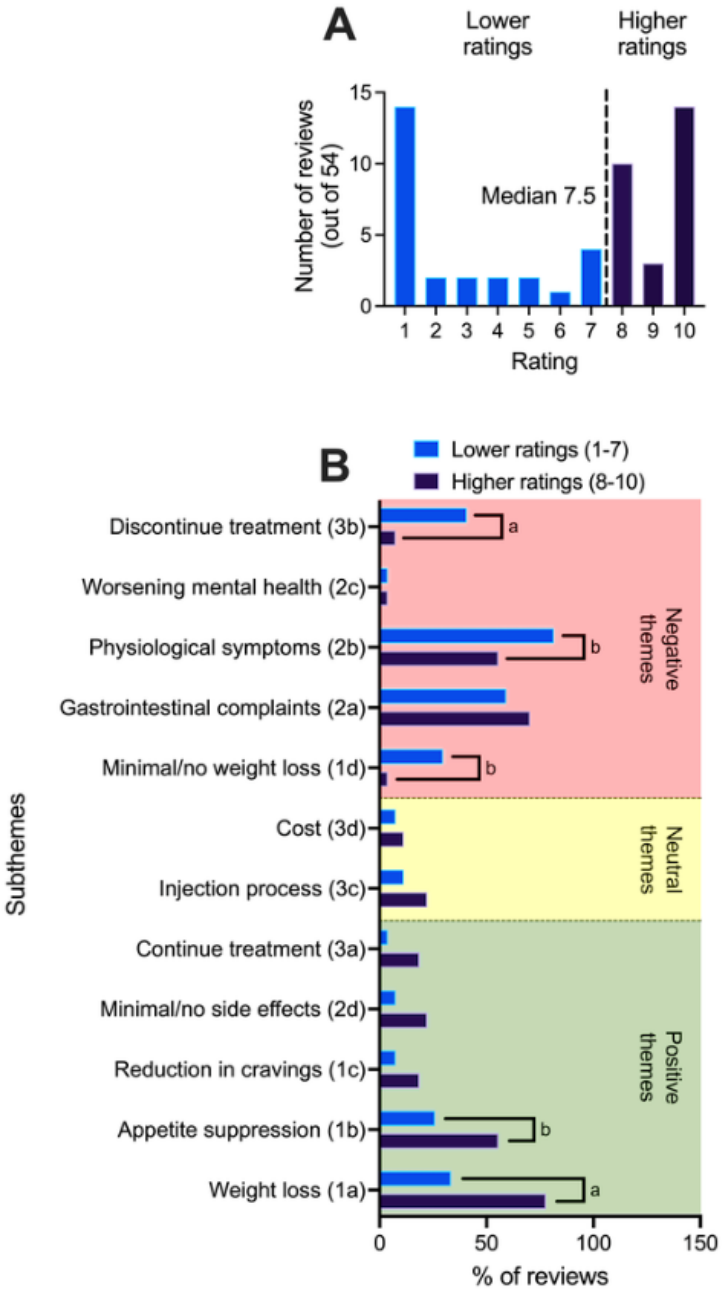

Reporting Standards

This study was designed and reported in accordance with the Good Reporting of a Mixed Methods Study (GRAMMS; Multimedia Appendix 1) and Standards for Reporting Qualitative Research (SRQR; Multimedia Appendix 2) guidelines to ensure transparency and reproducibility.

Results

Thematic Analysis

Overview

Our thematic analysis reached saturation after analyzing 60 responses. Three major themes emerged from these analyses (outlined in Tables 1-3). Each overarching theme comprised several related subthemes that reflected a spectrum of responses (eg, weight loss vs no weight loss). Some respondents contributed to multiple related subthemes (eg, individuals who

XSL•FO
RenderX

initially lost weight with Ozempic but later regained weight were included in both "Weight loss and related outcomes" and "No/minimal weight loss or weight rebound"). Below, we describe each theme and provide representative verbatim examples. For each subtheme, we also report the median quantitative rating (1-10 scale) among all participants who contributed to that subtheme, along with the respective median absolute deviation (MAD) from the median.

Theme 1: Change in Body Weight/Appetite

Approximately half of respondents (33/60, 55%) indicated that they experienced weight loss at some point during Ozempic treatment (Table 1; subtheme 1a). Some users also reported improvements in weight-related outcomes, including reductions in cholesterol levels. Quantitative ratings from respondents contributing to this subtheme (median 8.5, MAD 1.5) were higher than the overall group median (median 7.5, MAD 2.5). A substantial proportion of respondents (22/60, 37%) reported that Ozempic treatment was associated with appetite suppression (Table 1; subtheme 1b). The median quantitative ratings for this group were similar to those for subtheme 1a (median 8.0, MAD 1.5), reflecting the considerable overlap in respondents (n=15) contributing to both subthemes. Additionally, a subtheme emerged regarding a reduction in food cravings (Table 1; subtheme 1c), with 8 of 60 (13%) respondents reporting that Ozempic treatment was associated with decreased cravings, particularly for sugary and greasy foods (median 10.0, MAD 0.0). A total of 40 unique respondents were coded under subthemes 1a, 1b, or 1c, indicating that 40 out of 60 (67%) respondents reported reductions in weight, appetite, or cravings. By contrast, 11 out of 60 (18%) respondents expressed no/minimal weight loss or weight rebound (Table 1; subtheme 1d; median 1.0, MAD 0.0). Among these, some noted they had not lost any weight, while others reported weight loss occurring slowly. Four respondents in subtheme 1d also appeared in subthemes 1a, 1b, or both, suggesting that although they initially experienced reductions in weight or appetite, these effects were not sustained over time (ie, weight loss plateaued or reversed).

Table 1. Representative quotes for theme 1: "Change in Body Weight/Appetite."

Subtheme	n	Median (median absolute deviation) drug rating (0-10)	Examples of review comments (and drug rating associated with comment)
1a. Weight loss and related outcomes	33	8.5 (1.5)	• I've been taking Ozempic for almost 1 year and I have lost 55 lbs. [Rating N/A[a]] • For the first time in 30 years I don't go to bed kicking myself for what I've eaten or making promises to myself to make amends for overeating. [Rating 7] • I started at 192 lbs Nov 15th and as of May 1st I weigh 152 lbs. [Rating 10] • I am more than pleased that my cholesterol is now 180 from 270 on medication. I have never been below 200 total cholesterol in my life! [Rating 10] • In the first month, I lost 15 pounds on the lowest dose. [Rating 10]
1b. Appetite suppression	22	8.0 (1.5)	• I don't have much an appetite and I feel fuller faster and longer. [Rating 10] • My appetite was reduced by 90%. I used to overeat, but now I can only manage two small meals a day. [Rating 9] • It curbed my appetite from the moment I took the 1st dose. I didn't feel any hunger despite being on a low-calorie diet and exercising five times a week. [Rating 10] • Yes, it makes you eat way less, I was never hungry but made myself eat because I had to. [Rating 2]
1c. Reduction in food cravings	8	10.0 (0.0)	• I found my sugar cravings disappeared once I started taking 1 mg. Up until then I still craved sugary foods. I lost all interest in greasy food (fries, anything deep fried etc) from .5 mg and up. [Rating 10] • Ozempic helped me cure my sugar addiction and greediness. [Rating 10] • I don't crave junk food and only eat 1/3 of what I used to since I stay full longer. It's nice [Rating 8]
1d. No/minimal weight loss or weight rebound	11	1.0 (0.0)	• I am loosing [sic] weight but it is very slow (.25/week if I am lucky) [Rating 5] • It has done absolutely NOTHING for me for weight loss. [Rating 1] • it seems to have plateaued as I haven't lost any weight since Christmas and it's now March. [Rating N/A] • To date, I am not losing anything and most weeks I have gained the weight back. [Rating 1]

[a]N/A: not applicable.

XSL•FO
RenderX

Theme 2: Nonweight-Related Symptoms and Side Effects

The majority of respondents (48/60, 80%) indicated that Ozempic treatment was associated with nonweight-related symptoms and side effects that varied in nature and severity. Nausea and gastrointestinal complaints were the most frequently reported (Table 2; subtheme 2a; 37/60, 62%), including general nausea, vomiting, burping, and severe constipation. Interestingly, the quantitative ratings associated with this subtheme (median 8.0, MAD 2.0) were similar to the overall median across all respondents (median 7.5, MAD 2.5), suggesting that the presence or absence of these symptoms was not a decisive factor in participants' overall appraisal of the medication (discussed further below). Respondents also described a range of other physiological (nongastrointestinal) symptoms, including headaches, gallbladder complications, severe dehydration, blood loss, and anemia (Table 2; subtheme 2b; 40/60; 67%; median 7.0, MAD 3.0). Two respondents reported negative impacts on mental health, specifically the onset or worsening of depressive symptoms (Table 2; subtheme 2c; median 4.5, MAD 3.5). Finally, a minority of respondents (8/60, 13%) explicitly indicated that they did not experience any distressing adverse events (Table 2; subtheme 2d; median 10.0, MAD 0.0).

Table 2. Representative quotes for theme 2: "Nonweight-Related Symptoms and Side Effects."

Subtheme	n	Median (median absolute deviation) drug rating (0-10)	Examples of review comments (associated quantitative rating of Ozempic efficacy)
2a. Nausea and gastrointestinal complaints	37	8.0 (2.0)	• It started with huge belches and nausea. That night I vomited and was so lethargic and nauseous that I didn't get out of bed for 3 days. [Rating 5] • The most consistent symptom throughout my 5 months on Ozempic has been severe constipation. The inability to down lots of water like I used to has only added to the constipation. Nausea has also been prevalent from early on but reached the point of unbearable after a few weeks on 2 mg. Started throwing up daily around that time as well which was when the costs started outweighing the benefits. [Rating 7] • I have had some very bad nausea, vomiting and diarrhea. Also lots of burping and it smells terrible. When I have vomited - it is sooo much volume. More than anytime in my life. [Rating 7] • I have had massively bad headaches, nausea, vomiting, and stomach pain. After the first injection, I ended up in the ER because of my stomach pain, and then again 5 days later. I haven't been able to keep anything down. I can barely keep 2 sips of water down. I can't even take any of my prescription medications because I am constantly throwing them back up. [Rating 3]
2b. Other physiological (nongastrointestinal) symptoms	40	7.0 (3.0)	• I waited for three weeks and then tried again. However, after two injections, I became severely dehydrated and ended up in the ICU at the Heart Center. I had collapsed at a baby shower due to dehydration and was experiencing blood loss in my stools, anemia, and electrolyte imbalances. I could have died, but thankfully, I am still here to share my story. [Rating 1] • I've gotten a few headaches [Rating 10] • The sad news is I got gallbladder problems from ozempic and my gallbladder has to be removed. [Rating 1] • However, after a couple of months, I started getting abdominal pains in the upper right quadrant that extended to my back, between my shoulders. The pains would manifest a day or two after the shots and last several hours. It was worse during standing or walking and not a problem when sitting. [Rating 8] • I do get a case of terrible heartburn after each injection. [Rating 1] • ...swelling of the throat, and difficulty swallowing which has not stopped since the first and last dose [Rating 1] • I have constant dizziness [Rating 4] • I had weird sores on my tongue. [Rating 5]
2c. Detrimental mental health outcomes	2	4.5 (3.5)	• The worst side effect for me was depression. I would have mild anxiety and this drug made it a lot worse and made my mood very low...I just couldn't stick feeling so low in my mood. [Rating 8] • Depression - very strange feeling, almost like out of body experience. For first time found myself HATING body...Difficulty focusing. [Rating 1]
2d. Minimal or no experience with side effects	8	10.0 (0.0)	• Obviously some people have terrible side effects, but I've had none (not even a headache!). [Rating 10] • I haven't had any side effects, only positive ones! [Rating 10] • No side effects whatsoever but saying that I was (and still am) eating healthy (no greasy/fried food, no sugar, no alcohol, low fat). [Rating 10]

XSL•FO
RenderX

Theme 3: Plans for Ongoing Use Versus Discontinuation

Some respondents (n=20) discussed their intentions regarding continuation or discontinuation of Ozempic treatment, as well as practical considerations related to ongoing use. A small subset (n=6) explicitly reported plans to continue treatment (Table 3; subtheme 3a; median 9.0, MAD 1.0), often despite experiencing side effects. By contrast, a larger number (n=14) explicitly indicated plans to discontinue Ozempic (Table 3; subtheme 3b), and their quantitative ratings were correspondingly lower than the overall group median (median 3.0, MAD 2.0). Across the entire sample, 10 out of 60 (17%) respondents discussed the injection process associated with Ozempic administration (Table 3; subtheme 3c; median 8.0, MAD 1.0). Most users described the process as straightforward and convenient, with only 3 reporting difficulties. A small number (n=5) mentioned the cost of Ozempic (Table 3; subtheme 3d; median 8.0, MAD 1.0); while some considered it manageable when covered by insurance or medical cards, others identified cost as a barrier to continued use.

Table 3. Representative quotes for theme 3: "Plans for Ongoing Use Versus Discontinuation."

Subtheme	n	Median (median absolute deviation) drug rating (0-10)	Examples of review comments (associated quantitative rating of Ozempic efficacy)
3a. Plans to continue with treatment	6	9.0 (1.0)	• I have had all 5 of the main side effects, nausea, stomach pain, vomiting, diarrhea, and constipation. I am happy with the weight loss, so am learning to manage these. [Rating 8] • The side effects at the beginning were worth it for me but from the sounds of it, mine weren't that bad. I have never once thrown up. [Rating 10] • What it has done is to force me to give up bad habits because I do not like staying nauseous. I do believe this negative reinforcement will make me sustain weight loss after I have finished my rounds. [Rating 10]
3b. Plans to discontinue treatment	14	3.0 (2.0)	• I do not plan to continue. No pain, no gain. I'll get my loss the old-fashioned way with proper diet and exercise. Never again want to feel this way on Ozempic. [Rating 1] • ...while watching TV, I heard an Ozempic commercial that included warnings about gallbladder problems and pancreatitis. I immediately stopped using it and went to my doctor. [Rating 1] • I have had pretty much all the side effects possible. I have missed four days of work because I can't leave my bathroom. I can't keep anything in my stomach with it hurting and the diarrhea will not stop. I'm getting dehydrated but If I drink it goes right through me. I have a family to look after and I can't. All I want to do is sleep. I don't think I will be taking my second injection. [Rating N/A[a]] • I've been on Ozempic for 4 months. Recently raised my dose to .75 mg but I am stopping this medication. Yes, I lost 15 lbs but suffered with the worse sulfur gas, was sick in bed at times, missing appointments because I felt that crap. [Rating 2] • This is my third week and I feel absolutely horrid!! I have constant dizziness, extreme fatigue, and generally feel like crap. I don't think I will continue. It's not worth it to lose maybe 20 lbs but can't get out of bed and function. I am tired of feeling like death warmed over. [Rating 4] • Got up to 2mg after 2 years at 1mg. Stopped because I began vomiting and feeling nauseous [Rating 7]
3c. Injection process	10	8.0 (1.0)	• I had to go to my doctor's office to have them me show how to use the pen from priming it to injecting the pen. It was easy. Even the directions from the website is easier. [Rating 10] • It is very convenient to use (inject once a week), and the needle is so thin (less than a hair) you don't even feel it. [Rating 10] • Dosing 2mg is difficult as a needle change is required. [Rating 7] • Incorrect filling of the pen, it never has enough according to dosage. [Rating N/A]
3d. Cost	5	8.0 (1.0)	• I'm in Ireland, and it costs €150 a month for four injections, which are covered by my medical card. [Rating 9] • Unfortunately, it did take a toll on my wallet, and I eventually had to switch from my hospital/GP to getting it online by telehealth from semalean. My son has also started taking it, and I would recommend it to anyone who is curious, but only if they are willing to see if the side effects apply to them. If they do not, and the treatment is affordable or covered by insurance, it can be truly amazing. [Rating 8] • Don't waste your money on this stuff. [Rating 1]

[a]N/A: not applicable.

XSL•FO
RenderX

Identification of Themes Contributing to Higher Versus Lower Quantitative Ratings of Ozempic Efficacy

Across the 54 participants who provided a quantitative rating of Ozempic's efficacy on the 1-10 scale, the average rating was 5.98 (SD 3.65), and the median was 7.5 (MAD 2.5; Figure 2A). For comparison, the average rating among all available user reviews of Ozempic for weight loss on Drugs.com at the time of data collection (n=78) was 6.04, with a median of 7.0 (MAD 3.0). Statistical analysis indicated no significant difference between these 2 distributions (Mann-Whitney *U* test, *P*=.996), indicating that our analytic subsample was representative of the broader dataset. The most frequent scores were 1 and 10 (n=14 each), indicating that more than half of respondents (28/54, 52%) rated Ozempic as either the highest or lowest possible score (see Figure 2A). Based on the distribution of user ratings, we divided the data into 2 groups using a median split: those with higher ratings (scores of 8-10; n=27) and those with lower ratings (scores of 1-7; n=27; Figure 2A). To examine which qualitative subthemes were associated with more positive versus negative evaluations of Ozempic, we conducted 2 × 2 chi-square tests comparing the frequency of each subtheme across the higher- and lower-rating groups (Figure 2B).

As expected, several subthemes reflecting positive sentiment were more frequently observed among respondents who provided higher ratings of perceived efficacy than among those who provided lower ratings. These included weight loss (subtheme 1a; 21/27, 78%, vs 9/27, 33%; χ^2_1[n=54]=10.80, *P*=.001) and appetite suppression (subtheme 1b; 15/27, 56%, vs 7/27, 26%; χ^2_1[n=54]=4.91, *P*=.03). Although reduction in food cravings was mentioned more often among higher-rating respondents (subtheme 1c; 5/27, 19%, vs 2/27, 7%), this difference did not reach statistical significance (*P*=.22). Three subthemes reflecting negative sentiment were observed at significantly higher frequencies among respondents providing lower ratings. These included no/minimal weight loss or weight rebound (subtheme 1d; 8/27, 30%, vs 1/27, 4%; χ^2_1[n=54]=6.53, *P*=.01), physiological (non-gastrointestinal) symptoms (subtheme 2b; 22/27, 81%, vs 15/27, 56%; χ^2_1[n=54]=4.21, *P*=.04), and plans to discontinue treatment (subtheme 3b; 11/27, 41%, vs 2/27, 7%; χ^2_1[n=54]=8.21, *P*=.004). Interestingly, the frequency of nausea/gastrointestinal complaints (subtheme 2a; 19/27, 70%, vs 16/27, 59%; χ^2_1[n=54]=0.73, *P*=.39) did not differ significantly between the higher- and lower-ratings groups.

Finally, we conducted exploratory analyses to better understand the profile of the relatively small number of respondents who explicitly indicated their intention to continue (n=6; Figure 3A) versus discontinue (n=14; Figure 3B) Ozempic treatment. All respondents (6/6, 100%) who intended to continue treatment reported ongoing weight loss, appetite suppression, or reduction in food cravings (ie, subthemes 1a, 1b, and 1c, but not 1d); this proportion differed significantly from those who intended to discontinue treatment (5/14, 36%; χ^2_1[n=20]=7.01, *P*=.008). By contrast, both groups reported a high frequency of adverse events (subthemes 2a, 2b, or 2c; 6/6, 100%, vs 13/14, 93%; *P*=.50). All 6 continuers reported nausea and gastrointestinal complaints (subtheme 2a), and 4 experienced other physiological symptoms (subtheme 2b). Similarly, 9 of the 14 discontinuers experienced nausea, and 12 reported other physiological symptoms. Together, these data suggest that the intention to discontinue Ozempic may be driven primarily by a lack of perceived efficacy (failure to lose weight), rather than by side effects or adverse events.

XSL•FO
RenderX

Figure 3. Weight loss and side effect outcomes among respondents who explicitly indicated an intention to continue (n=6) or discontinue (n=14) Ozempic treatment in the long term. Among those intending to continue (A), 100% reported ongoing weight loss, appetite suppression, and/or reduction in cravings (i.e., subthemes 1a, b, c, but not d), as well as experiencing some type of side effect (subtheme 2a, b, and/or c). Users who intended to discontinue treatment (B) were significantly less likely to report ongoing weight loss and/or appetite suppression (36%) and had a high frequency of adverse events (93%). **p<0.01, $\chi 2$ analyses).

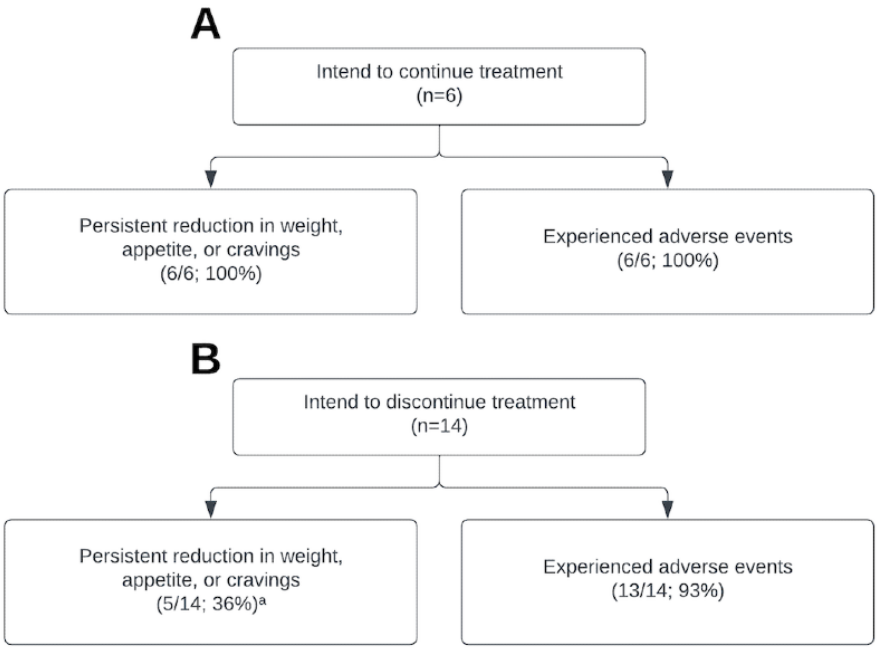

Discussion

Principal Findings and Comparison to Prior Work

This study provides new insight into how individuals perceive and evaluate the off-label use of Ozempic for weight loss, based on unsolicited, real-world data from an online medication review platform. Using a mixed methods approach, we found that user satisfaction was driven primarily by perceived effectiveness in promoting weight loss and appetite suppression, whereas gastrointestinal side effects were common but exerted limited influence on overall evaluations or decisions to continue treatment. Rather, discontinuation was most strongly associated with no/minimal weight loss or the occurrence of other, nongastrointestinal side effects. These findings highlight that, for many users, perceived efficacy outweighed tolerability concerns—a perspective that may be underrepresented in traditional clinical trials—and demonstrate the potential of infoveillance methods to capture patient-centered attitudes that shape treatment adherence.

Of the 60 respondents, 40 (67%) reported reduced weight, appetite, or cravings as a result of Ozempic treatment. This finding aligns with clinical trial data demonstrating the broad efficacy of semaglutide in promoting weight loss, with reductions of up to 17.3% observed after approximately 1 year of treatment, depending on dose and patient population [12,14,15]. Across these studies, approximately 13.5% of participants failed to achieve ≥5% weight loss with semaglutide 2.4 mg, comparable to the 18% (11/60) of participants in our sample who reported minimal or no overall weight loss. Consistent with these findings, subthemes related to weight outcomes were major contributors to respondents' overall quantitative ratings of Ozempic: subthemes 1a and 1b ("weight loss and related outcomes" and "appetite suppression") were associated with higher overall ratings, whereas subtheme 1d (no/minimal weight loss or weight rebound) occurred more frequently among lower ratings. Moreover, reductions in weight, appetite, and cravings (subthemes 1a, 1b, and 1c) were strongly associated with respondents' intention to continue versus discontinue treatment, underscoring these outcomes as being closely associated with long-term medication adherence. Notably, a retrospective study of electronic health records in the United States reported that semaglutide treatment was associated with higher persistence rates at 1 year (40%) compared with other weight loss medications, including liraglutide (17%), phentermine-topiramate (13%), and naltrexone-bupropion (10%) [31], likely reflecting its superior efficacy in promoting weight loss.

Several respondents indicated that although Ozempic treatment initially led to weight loss, this effect had plateaued or even reversed with continued use. This observation aligns with evidence from clinical studies showing that weight loss tends to plateau after approximately 1 year of semaglutide treatment, with weight regain often emerging during the second year in trial extension cohorts [32]. Such plateaus are consistent with those observed following other weight loss interventions and are thought to reflect metabolic adaptations, including reductions in resting and nonresting energy expenditure, accompanied by compensatory changes in appetite-regulating hormones [2,33-35]. In our dataset, very few respondents reported their treatment duration; therefore, we were unable to determine whether this variable mediated positive versus negative appraisals of Ozempic's efficacy. This should be a focus of future research, as it is conceivable that patient attitudes toward Ozempic become increasingly negative as weight loss plateaus or reverses over the longer term. Such dynamics likely have

XSL•FO
RenderX

important implications for long-term medication adherence, including for semaglutide formulations specifically approved for obesity and overweight [36].

A high proportion of respondents reported experiencing gastrointestinal complaints, including nausea, diarrhea, and vomiting. This aligns with clinical trial data identifying these as the most common adverse events associated with semaglutide treatment, with prevalence ranging from 41.9% to 82.8% across studies [12-16], as well as with preclinical evidence that GLP-1R agonists act on hindbrain regions involved in emesis control [37]. In our sample, these adverse events occurred with similar frequency among respondents who provided higher versus lower quantitative ratings, suggesting that gastrointestinal symptoms did not substantially influence overall attitudes toward Ozempic as a weight loss medication. This aligns with data from a previous study showing that 99.5% of gastrointestinal adverse events were nonserious, transient, and occurred most frequently during or shortly after dose escalation [38]. Moreover, across several clinical trials, treatment discontinuation due to gastrointestinal complaints was relatively uncommon, affecting only 3.4%-4.2% of participants [14,16]. Together, these findings indicate that gastrointestinal side effects are generally well tolerated and often regarded as "acceptable," particularly among individuals who experience meaningful weight loss (as described by 1 respondent: "I have had all 5 of the main side effects, nausea, stomach pain, vomiting, diarrhea, and constipation. I am happy with the weight loss, so am learning to manage these"). By contrast, users who reported other physiological (nongastrointestinal) symptoms tended to give lower quantitative ratings. This may reflect the greater severity of some of these adverse events, with several users indicating hospitalization due to complications such as severe dehydration or gallbladder removal. Although we cannot confirm that these outcomes were directly attributable to Ozempic treatment, serious treatment-associated adverse events have been reported in approximately 10% of participants in large-scale studies [15,16], including gallbladder disorders such as cholelithiasis and cholecystitis, which have led to treatment discontinuation in some cases [12].

Our data also included 2 instances in which respondents reported experiencing depression symptoms that they attributed to Ozempic treatment. Recent discussions have raised concerns about a possible association between semaglutide use and adverse mental health outcomes, particularly suicidal ideation [39]. This aligns with a recent FDA submission noting a disproportionate number of reports of "depression/suicidal" and suicidal ideation among individuals treated with semaglutide, although no causal relationship was established [39]. However, other studies, including a recent meta-analysis of 25 clinical trials, have found no association between GLP-1R agonists and suicidal or self-injurious behaviors [40,41], and some evidence even suggests a lower risk of these outcomes compared with other medications for obesity and T2D [42]. These mixed findings mirror patterns observed among bariatric surgery patients, where treatment has been associated not only with improvements in depression and anxiety but also with an increased risk of suicidality and self-injurious behavior [43]. Collectively, these data highlight the need for further research, including prospective studies and controlled clinical trials, to clarify the potential mental health risks associated with semaglutide use. Complementary preclinical investigations may also be necessary to identify shared neurobiological pathways underlying the regulation of appetite and mood.

Limitations and Future Directions

We acknowledge several important limitations of our study. First, our data were opportunistic and derived exclusively from a single website (Drugs.com), which does not publish demographic information about its users. As such, it is unclear to what extent our findings are representative of the broader population of Ozempic users. Online reviewers may also differ systematically from the general treatment population (eg, in health literacy, socioeconomic status, or engagement with digital health platforms), potentially biasing the types of experiences shared. Furthermore, self-selection biases may amplify extreme positive or negative perspectives, leading to an overrepresentation of polarized views [44]. Although we mitigated this by including all eligible reviews and reporting aggregate rather than individual data, future research should extend these findings through prospective, consented studies that collect demographic and clinical information, enabling the hypotheses generated here to be tested in more representative and generalizable samples.

Second, cumulative evidence indicates that weight loss in response to GLP-1R agonists may be more pronounced among women [45-47]; future research should therefore examine whether perceptions of Ozempic's clinical benefits and side effect profile differ by sex. Relatedly, because the data were self-reported, we could not independently verify clinical outcomes, adverse events, or the reasons for Ozempic use. Although this limitation may introduce some inaccuracy, the strong consistency of themes across respondents provides reassurance regarding the reliability of the data. Third, weight loss outcomes and adverse events are likely influenced by both the dose of Ozempic and the duration of treatment. These variables were not available in the present dataset and therefore, could not be analyzed. Finally, online reviews of consumer products, including medications, may be shaped by contextual factors such as prior reviews or platform norms, which could "prime" respondents to emphasize certain outcomes over others. This limitation underscores the need to triangulate infoveillance data with controlled, prospective designs to validate and extend these findings.

Clinical Implications

These findings offer practical insights for both clinicians and their patients. For clinicians, acknowledging that side effects are common, and occasionally serious, can help guide expectation-setting, safety monitoring, and decisions about when to consider alternative treatments. For patients, understanding that side effects vary in severity and that weight loss may plateau over time can support more realistic expectations and informed discussions with health care providers about whether to continue or adjust therapy. Together, these insights can foster clearer communication and more patient-centered decision-making.

XSL•FO
RenderX

Ethical Implications of Using Deidentified Online Data

Although we analyzed public, deidentified posts, we acknowledge that ethical considerations remain, including (1) the potential for reidentification through rare combinations of clinical details; (2) users' possible expectation that posts were intended for peer support rather than research; and (3) the risk of unintended harm, such as reinforcing stigma around the use of medications for weight loss. We mitigated these risks by limiting data collection to information necessary for analysis, screening quotations to remove potentially identifying details, reporting results in aggregate, and using neutral, nonsensational language. Future research should build on the insights from this study through prospective designs with informed consent, enabling hypotheses generated here to be tested in more representative and ethically robust samples that include limited demographic and clinical data.

Conclusion

Attitudes toward Ozempic were shaped primarily by its perceived effectiveness in promoting weight loss, with gastrointestinal side effects exerting minimal influence on overall satisfaction. For many users, the benefits of appetite suppression and weight reduction outweighed treatment-related discomfort. By leveraging an infoveillance approach, this study identified key patient-reported factors driving satisfaction and discontinuation that may be underrepresented in traditional research. These findings provide a foundation for future structured studies aimed at improving adherence and optimizing treatment strategies for individuals with overweight and obesity.

Acknowledgments

AJA gratefully acknowledges support through a Post-Baccalaureate Research Experience for LSAMP Students (PRELS) stipend from the National Science Foundation via the Garden State-LSAMP (NSF Award HRD-1909824). RTH gratefully acknowledges support through an LSAMP Undergraduate Summer Research stipend from the National Science Foundation via the Garden State-LSAMP (NSF Award HRD-1909824). This work was funded by grants to MHJ from the University of Sydney (Horizon Fellowship); the National Institute on Drug Abuse (R00 DA04765 and R37 DA061303); the National Heart, Lung, and Blood Institute (U01HL150852); Rutgers Optimizes Innovation; the New Jersey Health Foundation (PC144-23 and PC98-22); and an International Collaborative Research Grant from Rutgers Global. Artificial intelligence (ChatGPT; OpenAI) was used for minor language refinement and editing; all outputs were reviewed and modified, and the final content is the authors' responsibility. Figures 1 and 3 were created using Lucidchart (Lucid Software Inc). Special thanks to the individuals who shared their experiences anonymously on Drugs.com; we hope this research helps improve care for such individuals.

Data Availability

The data that support this study are available upon reasonable request from the corresponding author.

Authors' Contributions

AJA and MHJ conceptualized the study. AJA and RTH collected and coded the data. AJA, RTH, RMB, and MHJ reviewed the data and discussed potential themes. AJA and MHJ drafted the manuscript, and AJA, RTH, KRG, HEB, RMB, and MHJ edited and approved the final version.

Conflicts of Interest

MHJ is an inventor on patent PCT/US23/27918 titled "Therapeutic combinations and methods," which describes novel approaches for reducing overeating. All other authors declare no competing interests.

Multimedia Appendix 1
GRAMMS (Good Reporting of a Mixed Methods Study) checklist.
[PDF File (Adobe PDF File), 62 KB - jmir_v28i1e78391_app1.pdf]

Multimedia Appendix 2
SRQR (Standards for Reporting Qualitative Research) checklist.
[PDF File (Adobe PDF File), 76 KB - jmir_v28i1e78391_app2.pdf]

References

1. WHO (World Health Organization). Obesity and overweight. WHO. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2025-11-11]
2. Hall KD, Kahan S. Maintenance of lost weight and long-term management of obesity. Med Clin North Am 2018 Jan;102(1):183-197 [FREE Full text] [doi: 10.1016/j.mcna.2017.08.012] [Medline: 29156185]
3. Wadden TA, Tronieri JS, Butryn ML. Lifestyle modification approaches for the treatment of obesity in adults. Am Psychol 2020;75(2):235-251 [FREE Full text] [doi: 10.1037/amp0000517] [Medline: 32052997]

XSL•FO
RenderX

4. Coutinho W, Halpern B. Pharmacotherapy for obesity: moving towards efficacy improvement. Diabetol Metab Syndr 2024 Jan 03;16(1):6 [FREE Full text] [doi: 10.1186/s13098-023-01233-4] [Medline: 38172940]
5. Arterburn DE, Telem DA, Kushner RF, Courcoulas AP. Benefits and risks of bariatric surgery in adults. JAMA 2020 Sep 01;324(9):879. [doi: 10.1001/jama.2020.12567]
6. Haddad A, Suter M, Greve JW, Shikora S, Prager G, Dayyeh BA, et al. Therapeutic options for recurrence of weight and obesity related complications after metabolic and bariatric surgery: an IFSO position statement. Obes Surg 2024 Nov;34(11):3944-3962. [doi: 10.1007/s11695-024-07489-7] [Medline: 39400870]
7. Mahapatra MK, Karuppasamy M, Sahoo BM. Semaglutide, a glucagon like peptide-1 receptor agonist with cardiovascular benefits for management of type 2 diabetes. Rev Endocr Metab Disord 2022 Jun;23(3):521-539 [FREE Full text] [doi: 10.1007/s11154-021-09699-1] [Medline: 34993760]
8. Ard J, Fitch A, Fruh S, Herman L. Weight loss and maintenance related to the mechanism of action of glucagon-like peptide 1 receptor agonists. Adv Ther 2021 Jun;38(6):2821-2839 [FREE Full text] [doi: 10.1007/s12325-021-01710-0] [Medline: 33977495]
9. Novo Nordisk A/S. Ozempic® prescribing information. Ozempic. 2025 10. URL: https://www.ozempic.com/prescribing-information.html [accessed 2025-12-23]
10. Jain AB, Ali A, Gorgojo Martínez JJ, Hramiak I, Kavia K, Madsbad S, et al. Switching between GLP-1 receptor agonists in clinical practice: expert consensus and practical guidance. Int J Clin Pract 2021 Feb;75(2):e13731 [FREE Full text] [doi: 10.1111/ijcp.13731] [Medline: 32975890]
11. Mahapatra MK, Karuppasamy M, Sahoo BM. Therapeutic potential of semaglutide, a newer GLP-1 receptor agonist, in abating obesity, non-alcoholic steatohepatitis and neurodegenerative diseases: a narrative review. Pharm Res 2022 Jun;39(6):1233-1248 [FREE Full text] [doi: 10.1007/s11095-022-03302-1] [Medline: 35650449]
12. O'Neil PM, Birkenfeld AL, McGowan B, Mosenzon O, Pedersen SD, Wharton S, et al. Efficacy and safety of semaglutide compared with liraglutide and placebo for weight loss in patients with obesity: a randomised, double-blind, placebo and active controlled, dose-ranging, phase 2 trial. Lancet 2018 Aug 25;392(10148):637-649. [doi: 10.1016/S0140-6736(18)31773-2] [Medline: 30122305]
13. Rubino D, Abrahamsson N, Davies M, Hesse D, Greenway FL, Jensen C, STEP 4 Investigators. Effect of continued weekly subcutaneous semaglutide vs placebo on weight loss maintenance in adults with overweight or obesity: the STEP 4 randomized clinical trial. JAMA 2021 Apr 13;325(14):1414-1425 [FREE Full text] [doi: 10.1001/jama.2021.3224] [Medline: 33755728]
14. Wadden TA, Bailey TS, Billings LK, Davies M, Frias JP, Koroleva A, et al. Effect of subcutaneous semaglutide vs placebo as an adjunct to intensive behavioral therapy on body weight in adults with overweight or obesity. JAMA 2021 Apr 13;325(14):1403. [doi: 10.1001/jama.2021.1831]
15. Wilding JP, Batterham RL, Calanna S, Davies M, Van Gaal LF, Lingvay I, et al. Once-weekly semaglutide in adults with overweight or obesity. N Engl J Med 2021 Mar 18;384(11):989-1002. [doi: 10.1056/nejmoa2032183]
16. Davies M, Færch L, Jeppesen OK, Pakseresht A, Pedersen SD, Perreault L, et al. Semaglutide 2.4 mg once a week in adults with overweight or obesity, and type 2 diabetes (STEP 2): a randomised, double-blind, double-dummy, placebo-controlled, phase 3 trial. The Lancet 2021 Mar;397(10278):971-984. [doi: 10.1016/s0140-6736(21)00213-0]
17. Ghusn W, De la Rosa A, Sacoto D, Cifuentes L, Campos A, Feris F, et al. Weight loss outcomes associated with semaglutide treatment for patients with overweight or obesity. JAMA Netw Open 2022 Sep 01;5(9):e2231982 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.31982] [Medline: 36121652]
18. Jordan G, Young S, Alemán JO. Weight loss pharmacotherapy: current and future therapies. Gastrointest Endosc Clin N Am 2024 Oct;34(4):591-608. [doi: 10.1016/j.giec.2024.06.006] [Medline: 39277293]
19. Basch C, Narayanan S, Tang H, Fera J, Basch C. Descriptive analysis of TikTok videos posted under the hashtag #Ozempic. Journal of Medicine, Surgery, and Public Health 2023 Sep 27:100013 [FREE Full text] [doi: 10.1016/j.glmedi.2023.100013]
20. Madsbad S, Holst JJ. The promise of glucagon-like peptide 1 receptor agonists (GLP-1RA) for the treatment of obesity: a look at phase 2 and 3 pipelines. Expert Opin Investig Drugs 2025 Mar;34(3):197-215 [FREE Full text] [doi: 10.1080/13543784.2025.2472408] [Medline: 40022548]
21. Garber AJ. Long-acting glucagon-like peptide 1 receptor agonists: a review of their efficacy and tolerability. Diabetes Care 2011 May;34 Suppl 2(Suppl 2):S279-S284 [FREE Full text] [doi: 10.2337/dc11-s231] [Medline: 21525469]
22. McGovern A, Tippu Z, Hinton W, Munro N, Whyte M, de Lusignan S. Comparison of medication adherence and persistence in type 2 diabetes: a systematic review and meta‐analysis. Diabetes Obesity Metabolism 2017 Dec 12;20(4):1040-1043. [doi: 10.1111/dom.13160]
23. Guerci B, Charbonnel B, Gourdy P, Hadjadj S, Hanaire H, Marre M, et al. Efficacy and adherence of glucagon-like peptide-1 receptor agonist treatment in patients with type 2 diabetes mellitus in real-life settings. Diabetes Metab 2019 Dec;45(6):528-535. [doi: 10.1016/j.diabet.2019.01.006] [Medline: 30677504]
24. Uzoigwe C, Liang Y, Whitmire S, Paprocki Y. Semaglutide once-weekly persistence and adherence versus other GLP-1 RAs in patients with type 2 diabetes in a US real-world setting. Diabetes Ther 2021 May;12(5):1475-1489 [FREE Full text] [doi: 10.1007/s13300-021-01053-7] [Medline: 33837922]

XSL•FO
RenderX

25. Weiss T, Carr RD, Pal S, Yang L, Sawhney B, Boggs R, et al. Real-world adherence and discontinuation of glucagon-like peptide-1 receptor agonists therapy in type 2 diabetes mellitus patients in the United States. Patient Prefer Adherence 2020;14:2337-2345 [FREE Full text] [doi: 10.2147/PPA.S277676] [Medline: 33273810]
26. Drugs.com. URL: https://www.drugs.com/ [accessed 2023-06-26]
27. Armanious A, Asare A, Mitchison D, James M. Patient perceptions of lisdexamfetamine as a treatment for binge eating disorder: an exploratory qualitative and quantitative analysis. Psychiatry Res Commun 2024 Dec;4(4):100195 [FREE Full text] [doi: 10.1016/j.psycom.2024.100195] [Medline: 39664649]
28. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
29. Mayor E, Bietti LM. A social media study of portrayals of bipolar disorders on YouTube: content and thematic analyses. J Med Internet Res 2025 Apr 25;27:e67129 [FREE Full text] [doi: 10.2196/67129] [Medline: 40279634]
30. Gilbert MK, Daughton AR, Chilcoat HD, Laffont CM, Strafford S, DeVeaugh-Geiss AM. Social listening for patient experiences with stopping extended-release buprenorphine: content analysis of reddit messages. J Med Internet Res 2025 Apr 25;27:e71245. [doi: 10.2196/71245]
31. Gasoyan H, Pfoh E, Schulte R, Le P, Rothberg M. Early- and later-stage persistence with antiobesity medications: a retrospective cohort study. Obesity (Silver Spring) 2024 Mar;32(3):486-493. [doi: 10.1002/oby.23952] [Medline: 38053443]
32. Wilding JPH, Batterham RL, Davies M, Van Gaal LF, Kandler K, Konakli K, STEP 1 Study Group. Weight regain and cardiometabolic effects after withdrawal of semaglutide: the STEP 1 trial extension. Diabetes Obes Metab 2022 Aug;24(8):1553-1564 [FREE Full text] [doi: 10.1111/dom.14725] [Medline: 35441470]
33. Sumithran P, Prendergast LA, Delbridge E, Purcell K, Shulkes A, Kriketos A, et al. Long-term persistence of hormonal adaptations to weight loss. N Engl J Med 2011 Oct 27;365(17):1597-1604. [doi: 10.1056/NEJMoa1105816] [Medline: 22029981]
34. Heckman BW, Mathew AR, Carpenter MJ. Treatment burden and treatment fatigue as barriers to health. Curr Opin Psychol 2015 Oct 01;5:31-36 [FREE Full text] [doi: 10.1016/j.copsyc.2015.03.004] [Medline: 26086031]
35. Fothergill E, Guo J, Howard L, Kerns JC, Knuth ND, Brychta R, et al. Persistent metabolic adaptation 6 years after "The Biggest Loser" competition. Obesity (Silver Spring) 2016 Aug;24(8):1612-1619 [FREE Full text] [doi: 10.1002/oby.21538] [Medline: 27136388]
36. Singh G, Krauthamer M, Bjalme-Evans M. Wegovy (semaglutide): a new weight loss drug for chronic weight management. Journal of Investigative Medicine 2023 May 25;70(1):5-13. [doi: 10.1136/jim-2021-001952]
37. Borner T, Geisler CE, Fortin SM, Cosgrove R, Alsina-Fernandez J, Dogra M, et al. GIP receptor agonism attenuates GLP-1 receptor agonist-induced nausea and emesis in preclinical models. Diabetes 2021 Nov;70(11):2545-2553 [FREE Full text] [doi: 10.2337/db21-0459] [Medline: 34380697]
38. Wharton S, Calanna S, Davies M, Dicker D, Goldman B, Lingvay I, et al. Gastrointestinal tolerability of once-weekly semaglutide 2.4 mg in adults with overweight or obesity, and the relationship between gastrointestinal adverse events and weight loss. Diabetes Obes Metab 2022 Jan;24(1):94-105 [FREE Full text] [doi: 10.1111/dom.14551] [Medline: 34514682]
39. McIntyre RS, Mansur RB, Rosenblat JD, Kwan ATH. The association between glucagon-like peptide-1 receptor agonists (GLP-1 RAs) and suicidality: reports to the Food and Drug Administration Adverse Event Reporting System (FAERS). Expert Opin Drug Saf 2024 Jan;23(1):47-55. [doi: 10.1080/14740338.2023.2295397] [Medline: 38087976]
40. Zhou J, Zheng Y, Xu B, Long S, Zhu L, Liu Y, et al. Exploration of the potential association between GLP-1 receptor agonists and suicidal or self-injurious behaviors: a pharmacovigilance study based on the FDA Adverse Event Reporting System database. BMC Med 2024 Feb 14;22(1):65 [FREE Full text] [doi: 10.1186/s12916-024-03274-6] [Medline: 38355513]
41. Chen J, Zhang Q, Wu Q, Zhang X, Xiang Z, Zhu S, et al. Impact of GLP-1 receptor agonists on suicide behavior: a meta-analysis based on randomized controlled trials. J Diabetes 2025 Sep;17(9):e70151. [doi: 10.1111/1753-0407.70151] [Medline: 40887719]
42. Wang W, Volkow ND, Berger NA, Davis PB, Kaelber DC, Xu R. Association of semaglutide with risk of suicidal ideation in a real-world cohort. Nat Med 2024 Jan;30(1):168-176 [FREE Full text] [doi: 10.1038/s41591-023-02672-2] [Medline: 38182782]
43. Law S, Dong S, Zhou F, Zheng D, Wang C, Dong Z. Bariatric surgery and mental health outcomes: an umbrella review. Front Endocrinol (Lausanne) 2023;14:1283621 [FREE Full text] [doi: 10.3389/fendo.2023.1283621] [Medline: 38027159]
44. Bhole B, Hanna B. The effectiveness of online reviews in the presence of self-selection bias. Simulation Modelling Practice and Theory 2017 Sep;77:108-123. [doi: 10.1016/j.simpat.2017.05.005]
45. Buysschaert M, Preumont V, Oriot PR, Paris I, Ponchon M, Scarnière D, UCL Study Group for Exenatide. One-year metabolic outcomes in patients with type 2 diabetes treated with exenatide in routine practice. Diabetes Metab 2010 Nov;36(5):381-388. [doi: 10.1016/j.diabet.2010.03.009] [Medline: 20598606]
46. Anichini R, Cosimi S, Di Carlo A, Orsini P, De Bellis A, Seghieri G, et al. Gender difference in response predictors after 1-year exenatide therapy twice daily in type 2 diabetic patients: a real world experience. Diabetes Metab Syndr Obes 2013;6:123-129 [FREE Full text] [doi: 10.2147/DMSO.S42729] [Medline: 23630427]

XSL•FO
RenderX

47. Quan H, Zhang H, Wei W, Fang T. Gender-related different effects of a combined therapy of exenatide and metformin on overweight or obesity patients with type 2 diabetes mellitus. J Diabetes Complications 2016;30(4):686-692. [doi: 10.1016/j.jdiacomp.2016.01.013] [Medline: 26873871]

Abbreviations

GLP-1R: glucagon-like peptide-1 receptor
GRAMMS: Good Reporting of a Mixed Methods Study
HbA1c: glycated hemoglobin
MAD: median absolute deviation (from median)
SRQR: Standards for Reporting Qualitative Research
T2D: type 2 diabetes

Edited by A Mavragani, N Cahill; submitted 02.Jun.2025; peer-reviewed by B Khandakar, C Annan, BT Oladun, M Almashmoum; comments to author 18.Aug.2025; accepted 04.Nov.2025; published 09.Jan.2026.

Please cite as:
Armanious AJ, Hunter RM, Griffiths KR, Bowrey HE, Brown RM, James MH
Patient Perceptions of Ozempic (Semaglutide) for Weight Loss: Mixed Methods Analysis of Online Medication Reviews
J Med Internet Res 2026;28:e78391
URL: https://www.jmir.org/2026/1/e78391
doi:10.2196/78391
PMID:

©Abanoub J Armanious, Rachel-Mae Hunter, Kristi R Griffiths, Hannah E Bowrey, Robyn M Brown, Morgan H James. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Public Emotional and Thematic Responses to Major Emergencies on Social Media, 2024-2025: Cross-Sectional Convergent Mixed Methods Study

Xingrong Guo[1], PhD; Yiqian Fan[1], MA; Yiming Guo[2], PhD

[1]College of Foreign Languages, Shanghai Maritime University, Shanghai, China

[2]School of Economics and Management, Shanghai Maritime University, Shanghai, China

Corresponding Author:
Xingrong Guo, PhD
College of Foreign Languages
Shanghai Maritime University
1150 Haigang Ave
Shanghai, 201306
China
Phone: 86 02138282732
Email: xmguo@shmtu.edu.cn

Abstract

Background: During 2024-2025, global emergencies triggered intense online discourse, presenting a unique opportunity to examine how cultural factors shape emotional expression and knowledge dissemination. Understanding these dynamic mechanisms is crucial for enhancing the effectiveness of digital health communication and optimizing crisis response strategies.

Objective: We analyzed how cultural and linguistic contexts influence emotional expression and thematic framing in social media comments during major emergencies in 2024-2025. We uncovered cross-cultural differences in collective emotions and narrative focuses, explaining how affective stance and discourse framing jointly shape the public construction of crisis meaning.

Methods: We used a cross-sectional, convergent mixed methods design. Data were collected retrospectively from X (formerly Twitter; X Corp) and Weibo (Sina Weibo) between January 1 and December 31, 2024. Using purposive sampling, we selected 5-6 representative emergency events per month based on online visibility (capped at 600 comments/event). The dataset included 19,813 comments from X and 6536 comments from Weibo. Emotions were identified using a Cross-lingual Language Model-Robustly optimized Bidirectional Encoder Representations from Transformers approach, and thematic patterns were extracted with Bidirectional Encoder Representations from Transformers Topic. Integrated Gradients was used to interpret model outputs, while clustering and network analysis were applied to visualize cross-cultural patterns. Hofstede's cultural dimensions theory helped interpret cultural influences on discourse. This mixed computational approach enabled a detailed comparison of emotional structures and thematic discourse across linguistic communities.

Results: Significant cross-platform differences were observed in emotional distribution (χ^2_8=8025.60; P<.001). Compared to X users, Weibo users, representing a collectivist culture, expressed concentrated negative emotions (20.37%; odds ratio [OR] 15.76, 95% CI 13.90-17.85), surprise (19.70%; OR 2.53, 95% CI 2.32-2.73), and fear (16.68%; OR 1.72, 95% CI 1.58-1.86), reflecting group-oriented anxiety and emotional contagion. In contrast, X (formerly Twitter) users in individualist contexts displayed dispersed sarcasm (43.49%; OR 55.19, 95% CI 43.95-69.21) and worry (15.30%; OR 55.27, 95% CI 34.74-87.88), indicating personalized and critical emotional styles. Topic modeling revealed dense clusters around "safety," "pray," and "resettlement" on Weibo, whereas X (formerly Twitter) comments emphasized decentralized themes of critique and responsibility. Semantic network analysis revealed a cohesive fear-prayer-rescue chain on Weibo and fragmented, debate-oriented interactions on X (formerly Twitter).

Conclusions: Emergency discourse is not neutral but is systematically structured by cultural values that shape emotions and themes. Integrating multilingual computational and qualitative methods, we offer a replicable framework using large-scale data, moving crisis and infodemiological research beyond single-platform or survey-based approaches. Our findings advance theory-informed understanding of how cultural meaning systems translate into observable digital discourse under conditions of risk and uncertainty. They also offer practical implications for governments, public health agencies, international organizations, and digital platforms by informing culturally adaptive, platform-specific risk communication, community moderation, and crisis

XSL•FO
RenderX

engagement strategies that can strengthen public trust, improve compliance with protective behaviors, and mitigate infodemic-related harms.

(J Med Internet Res 2026;28:e84648) doi:10.2196/84648

KEYWORDS

emergencies; cross-culture; XLM-RoBERTa; BERTopic; emotion; topic; cross-lingual language model–robustly optimized BERT approach; Bidirectional Encoder Representations from Transformers Topic

Introduction

Background

This study investigates the emotional and thematic patterns exhibited in social media comments during key global emergencies spanning 2024-2025, focusing on how cultural factors mold emotional expression and information dissemination. Understanding these dynamic mechanisms is essential for improving digital health communication and crisis response strategies. Global natural disasters, public health crises, and human-made accidents have intensified in recent years, placing unprecedented demands on emergency communication and mental health support. Social media platforms now serve as real-time spaces for emotional expression and public discourse during such events [1]. For example, on May 1, 2024, a highway in Meizhou, Guangdong, collapsed after prolonged heavy rain and unstable geological conditions. The disaster caused 52 deaths and injured 30 people. Another tragedy occurred on December 29, 2024, when a plane crashed in South Korea. Only 2 passengers survived, while 179 passengers lost their lives. Both incidents generated millions of online reactions within hours.

Social media has increasingly been acknowledged as a vital tool in emergency management. It allows authorities to monitor public sentiment in real time, evaluate opinion-related risks, and enhance the quality of risk communication [2-4]. Surveys show that 75% of people use or plan to use social media during emergencies, and 77% believe it delivers information faster than traditional channels [5]. Weibo (Sina Weibo) and X (formerly Twitter; X Corp) spread emergency updates more quickly than conventional media and provide spaces for emotional exchange and debate [5].

Limited understanding exists of how different cultural groups emotionally respond to the same emergency through online comments [6]. Although many studies have explored the role of tweets or posts during emergencies [7,8], such as information dissemination, public emotion analysis of posts, and crisis communication strategies. User comments are often overlooked because they are more spontaneous and interaction-driven [9]. While a substantial portion of research has focused on single-language or single-region datasets, some cross-lingual studies exist [10,11]. However, these studies often do not fully explore cross-cultural differences in public responses, particularly in the context of emergencies across multiple social media platforms. Therefore, there remains a need for research that systematically examines cross-cultural emotional and thematic patterns in multilingual online discourse. To bridge this gap, the study draws on comments posted during major emergencies in 2024-2025 on Weibo and X (formerly Twitter), comparing how cultures differ in emotional expression and discourse patterns. Multilingual emotion recognition is performed with the Cross-lingual Language Model-Robustly optimized Bidirectional Encoder Representations from Transformers (BERT) approach (XLM-RoBERTa), while BERTopic modeling is used to cluster topics and trace emotions such as fear, anger, and sarcasm across cultures. By revealing how emotions are shaped in online discourse, the research explains the psychological factors behind users' emergency responses and offers practical guidance for emergency communication that respects cultural differences.

The paper is structured as follows. Section 1 details data collection and the methods for preprocessing, emotion labeling, and topic modeling. Section 2 reports results on emotion patterns, topic clusters, and cross-cultural comparisons. Section 3 interprets these findings through cultural theories and discusses practical implications, concluding with limitations and directions for future work.

Literature Review

Social Media as a Digital Health Communication Tool in Emergencies

Digital health, or the use of digital technologies for health, has become an important field that applies routine and new forms of information and communications technology to meet health needs [12]. Social media platforms such as Weibo and X (formerly Twitter) enable fast and interactive communication. People can share personal experiences, questions, and feelings in real time, adding to official announcements. This immediate feedback builds a shared understanding of a situation. It also helps officials respond to misinformation and ease public anxiety, making social media a key setting for digital health efforts [13-15].

Existing research has confirmed that during crises such as the COVID-19 pandemic, digital health communication can lower uncertainty and encourage protective behaviors [16]. Other studies show that effective online risk communication can reduce fear, prevent rumors, and improve a community's ability to handle crises [17]. However, most of this research focuses on original posts or organizational messages. Less attention has been paid to the active and multilingual comment sections. In reality, comment sections often contain spontaneous, emotional conversations and reveal how people from different cultural backgrounds interpret health information. This content is crucial for understanding how people process risk information and decide how to respond. Despite the global reach of social media, there is still little cross-lingual and cross-cultural analysis of digital health communication.

XSL•FO
RenderX

The Unique Value of Comment Discourse in Emergency Communication

With the growth of social media, Weibo and X (formerly Twitter) have become major spaces for public emotion and opinion during crises. Unlike original posts that mainly broadcast information, comment threads allow real-time dialogue and shared narratives [18,19]. This bottom-up interaction captures immediate, authentic public sentiment [20] and reveals insights often missing from the posts themselves [19].

Comments play 2 key roles in society. They show natural emotional reactions and shape public agendas, influencing social norms and even public policy. Since discourse reflects cultural norms, emotional tones, and interaction styles vary across contexts [21]. For example, collectivist cultures stress emotional connection and group harmony, while individualist cultures focus more on self-expression [22]. Thus, comment styles on Weibo and X (formerly Twitter) during emergencies can differ, highlighting the need for cross-cultural comparison [23].

In social media research, comment sections have been less studied. One key reason is that these texts are often unstructured, fragmented, and full of everyday wording [24]. Nonstandard grammar, mixed languages, local slang, and subtle emotions further challenge traditional text analysis models in processing them [25,26]. These features make traditional models less effective, so studies on emergency communication have often left comments aside [27].

Emotional Characteristics and Multiemotion Analysis in Comments

Emotional expression is shaped by language and culture, as Dewaele and Pavlenko [28] emphasized in their cross-linguistic perspective on emotions, showing that different languages provide distinct repertoires for conveying feelings. In traditional sentiment analysis, Gul et al [2] note that it typically categorizes emotions into 3 broad types: positive, neutral, and negative. However, Kant et al [29] argue that this approach fails to capture the complexity of specific feelings such as fear, anger, worry, shock, and sarcasm, leading to an oversimplified understanding. Recent studies have shown that classifying emotions into 2 or 3 types is insufficient for crisis communication [4,30]. Regarding this issue, multiemotion labeling systems with categories such as fear, anger, and sarcasm offer a more accurate picture of public reactions and their potential policy implications [11,31].

Cross-lingual and cross-cultural sentiment analysis is still relatively understudied. Although social media is global, most research still relies on monolingual datasets and ignores culturally mixed communication [32]. Key linguistic characteristics, including the use of metaphors, styles of emotional expression, and cultural norms, differ greatly across languages [26]. These differences present challenges to current analytical models. Multilingual deep learning models such as XLM-RoBERTa can capture cross-linguistic semantic meanings [33,34]. Transformer-based topic modeling tools, such as BERTopic, demonstrate stronger performance in extracting coherent discussion themes from unstructured comment data infused with emotional content [35,36].

Existing Approaches in Social Media Emergency Research

Emergency communication research on social media mainly uses 3 methods: network analysis, content analysis, and sentiment analysis.

Network analysis studies how information spreads structurally during emergencies. For example, Zhang et al [37] showed that network structures affect how information spreads over time. Han et al [38] proposed a convolutional neural network with an extreme learning machine model-based algorithm to analyze the emotional influence of Weibo users. It measures how emotions spread among users. Singh and Singh [39] used text and graph multiview learning for tweet sentiment analysis, revealing structural and semantic connections. However, these kinds of studies place less emphasis on the content and emotions in messages.

Content analysis focuses on message features and communication strategies. Kada et al [40] analyzed government social media posts during COVID-19, and Chen and Ping [41] used the Wuli-Shili-Renli method for natural disasters. Nevertheless, content analysis has limitations in automating and scaling up when dealing with large amounts of user comments.

Sentiment analysis is widely used to gauge public emotions. Ou et al [30] explored the evolution of public sentiments, filling the gap in multiemotion classification studies. Studies [2,29] have found that negative feelings often dominate, while Halse et al [4] underscored their role in detecting trust. However, most sentiment studies only assess basic polarity (positive, negative, and neutral). Multiemotion classification remains scarce, especially in cross-lingual settings [32,42].

Recently, researchers have combined topic modeling methods with sentiment analysis to better capture themes and emotional tones. Babalola et al [43] reported that BERTopic is more effective than traditional models. Its usefulness in health-related social media studies was also confirmed by Khodeir and Elghannam [36] and Ma et al [44].

Hofstede's Cultural Dimensions Framework

To analyze the discourse of cross-cultural comments, Hofstede's [45] cultural dimensions theory is used in this study. It is a framework that not only compares the patterns of national cultures but also explains how culture influences comment style. Hofstede identifies 6 dimensions: power distance, individualism and collectivism, masculinity and femininity, uncertainty avoidance, long- and short-term orientation, and indulgence and restraint. These dimensions link cultural values to emotional and discursive behaviors. By applying Hofstede's framework to the comparative analysis, this study aims to interpret emotional and thematic differences within broader cultural contexts.

Hofstede's framework is applied in many areas beyond theory. It appears in studies on technology adoption [46], educational behavior [47], and online communication. Research [48] shows

XSL•FO
RenderX

cultural dimensions such as individualism and power distance strongly affect behavioral intention and emotional expression across nations.

Research Gap and Study Contribution

Although research on emergency communication through social media has grown, 3 key gaps remain. First, comment sections have not been fully used. Their unstructured format and the complexity of analysis limit their practical application [24,25]. Second, most studies rely on monolingual datasets and seldom examine cross-cultural comment data [32,49]. Third, research on multiemotion sentiment analysis in cross-cultural emergencies is still scarce [11,30].

To fill these gaps, this study proposes an integrated framework. The framework combines BERTopic and XLM-RoBERTa to perform cross-cultural, multiemotion analysis of comments from Weibo and X (formerly Twitter) during emergencies. Its purpose is to expand research methods in emergency communication and to broaden the scope of discourse analysis in digital public spaces. It emphasizes both multilingual coverage and emotional sensitivity.

Research Questions

This study examines user comments on Weibo and X (formerly Twitter) from multiple countries during 2024-2025 emergency events, focusing on emotional structures and thematic discourse. The analysis seeks to identify key topics and emotional patterns within social media comments and to explore how cultural factors shape styles of expression. To address these aims, XLM-RoBERTa is applied for sentiment analysis and BERTopic for topic modeling, providing data that enable detailed cross-cultural comparison.

This study seeks to explore the following research questions:

- RQ1: How do social media comments on emergency events express collective emotions and construct shared meanings?
- RQ2: In what ways do emotional expressions and narrative focuses differ across cultural and linguistic communities?
- RQ3: How are emotional valence and narrative focus interrelated in shaping the cross-cultural representation of emergencies?

Guided by these questions, the study proposes that the distribution and valence of emotional expressions differ significantly across cultural groups, reflecting distinct affective orientations and underlying value systems. Linguistic communities are also expected to demonstrate culturally specific narrative focuses when interpreting emergency events, revealing divergent framing patterns in emotional discourse. Moreover, emotional valence and narrative focus are assumed to be interrelated across cultures, suggesting that affective stance and discourse framing jointly contribute to the construction of crisis meaning.

Methods

Mixed Methods Design Overview

This study used a convergent mixed methods design, integrating quantitative computational analyses with qualitative content and discourse analysis to examine cross-cultural emotional and thematic patterns in social media responses to emergencies. The quantitative phase included multilingual emotion classification, frequency statistics, topic modelling, semantic co-occurrence analysis, and statistical testing. The qualitative phase involved manual coding by trained human coders, interpretive examination of representative comments, and theory-driven discourse analysis based on Hofstede's cultural dimensions. Both strands were conducted in parallel, and findings were integrated during interpretation to triangulate results and enhance validity. In adherence with best practices for observational studies, this paper was drafted using the JARS (Journal Article Reporting Standards) guideline [50] and was edited according to the JARS reporting checklist [51], which is included in Multimedia Appendix 1 Checklist 1.

Quantitative Component

Study Design and Data Collection

To explore public responses to emergencies, this study used a cross-sectional observational design and systematically collected social media comments from Weibo and X (formerly Twitter) between January 1 and December 31, 2024. At the end of each month, events from the previous month were systematically reviewed within a 1-week window. Continuous or nonbreaking events were excluded to ensure that only discrete emergency events were included in the dataset. Furthermore, the total number of comments across all major relevant hashtags for that event on Weibo and X (formerly Twitter) did not exceed 600 during the selected evaluation period. Given the potentially large data volume and uneven discussion intensity across events, a purposive sampling method was used. Each month, 5-6 representative events were selected based on topic relevance, comment volume, and overall online visibility, measured by comment counts, repost numbers, and trending hashtag rankings. Data were collected retrospectively after each event. Within the archived datasets, iterative sampling was conducted in successive batches until no new emotional categories or thematic patterns emerged, indicating the achievement of analytical saturation. This approach ensured that the dataset captured events that generated substantial public interaction and emotional expression on both platforms. For example, regarding the January 2024 Japan earthquake, specific trending hashtags were identified, including "# 日本地震 (Earthquake in Japan)" on Weibo and "#JapanEarthquake2024" on X (formerly Twitter). Posts related to these events were then systematically collected, with data acquisition carried out via custom Python scripts (Python Software Foundation), strictly adhering to the platforms' developer agreement and used solely for noncommercial, academic research purposes.

In total, 26,349 valid comments were gathered through this process, with 19,813 comments from X (formerly Twitter) and 6536 from Weibo. The combination of random sampling, multistage cleaning, and independent coding ensured that the final dataset remained representative, reproducible, and reliable for analysis.

XSL•FO
RenderX

Measures, Predictors, and Confounders

This study focuses on emotional patterns and thematic content reflected in cross-cultural public comments. The XLM-RoBERTa model was used, achieving an accuracy of 78.37%, outperforming traditional models, including the support vector machine (64.2% accuracy) and XGBoost (62% accuracy) [52]. To better capture emotions, the quantitative results went beyond basic labels ("negative," "positive," and "neutral") and included additional categories such as "worried," "fear," "angry," "sarcasm," "shock," and "sad." They also captured thematic structures and semantic network forms identified by computational models. In comparative analysis, platform type (Weibo and X, formerly Twitter) and event category (natural disasters and public crises) were regarded as key influencing factors. The study aimed to examine their differentiated effects on online emotional expression and discourse construction through analyzing these factors. In addition, several background variables were considered, covering event scale, total number of comments, posting time, and geographic origin. Quantitative processing of these variables was designed to ensure the statistical robustness of the main findings.

Data Processing and Sampling Procedures

As shown in Figure 1, the analysis proceeded in 5 major stages: pretraining, information collection, data processing, model preparation, and visual analysis. A multistage cleaning process was adopted, combining automated filtering and manual checks to eliminate spam, blank comments, advertisements, and duplicates. The collected data mainly included user IDs, posting times, usernames, comment content, repost counts, comment counts, such as counts and geographic locations. Due to the exclusion of invalid and incomplete entries during preprocessing (such as blank text content or missing timestamps), the final analytical dataset (N=26,349) consisted of complete cases, making multiple imputation unnecessary. The data were cleaned and stop words removed using different natural language processing libraries depending on language (Jieba for Chinese, Natural Language Toolkit for European languages, and stopwords-ISO [International Organization for Standardization] for others). The multistage cleaning process also involved eliminating irrelevant information, HTML tags, meaningless symbols, and duplicate comments. All personally identifiable information was destroyed immediately after collection, and only valid data was retained. For events exceeding 600 comments, Python random sampling was used to cap the dataset per event to maintain representativeness and reduce class imbalance.

XSL•FO
RenderX

Figure 1. Data analysis procedure. The data analysis included 5 steps: pretraining data, collecting comments via Python, processing comments, model preparation, and building models for visual analysis. BERTopic: Bidirectional Encoder Representations from Transformers Topic; HTML: HyperText Markup Language; ISO: International Organization for Standardization; NLTK: Natural Language Toolkit; XLM-ROBERTa: Cross-lingual Language Model–Robustly Optimized BERT Approach.

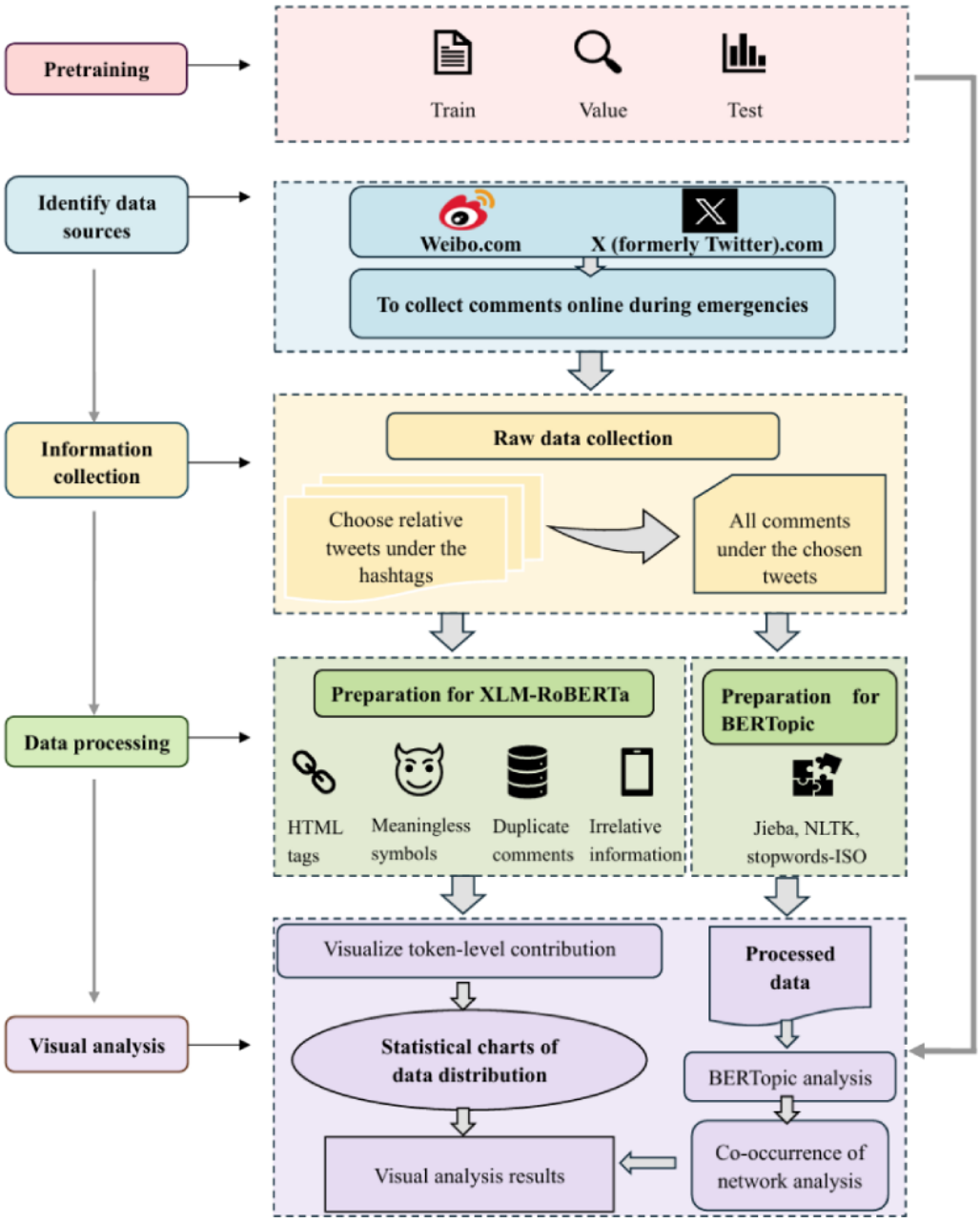

Computational Modeling for Sentiment Classification and Topic Clustering

Following the multilingual fine-tuning approach outlined by Rasool et al [53], the study used the XLM-RoBERTa-base model using a learning rate of 2×10^{-5}, batch size of 16, and 15 training epochs. The optimizer was AdamW with fused precision optimization adamw_torch_fused and a linear learning rate scheduler. The maximum sequence length was set to 128 tokens. Mixed-precision training (FP16) was enabled to optimize memory usage and training efficiency. To address class imbalance across emotion labels, a weighted binary cross-entropy loss function was used, with class weights inversely proportional to their frequency in the training corpus. Continuous emotion probability scores were binned into discrete emotional labels using predefined thresholds (neutral=0.12, surprise=0.10, positive=0.25, negative=0.30, sarcasm=0.25, fear=0.20, sad=0.25, worried=0.25, and anger=0.28). A total of 11,933 sentences were used for pretraining [54] to help the model better handle specific emotions, where meanings differ from literal words [55].

To mitigate potential semantic misalignment for low-resource languages (eg, Hindi and Indonesian), the research adopted a back-translation data augmentation strategy using the Many-to-Many 100 multilingual translation model. It has been shown to improve model quality in low-resource languages [56]. Each comment was translated from its source language to English and then back to the original language, enriching contextual diversity and improving cross-lingual embedding alignment before fine-tuning.

XSL•FO
RenderX

XLM-RoBERTa was applied to real data only after it showed adequate performance. As shown in Figure 2, the correlation checks between labels (ranging from –0.39 to 0.15) confirmed low overlap, indicating that the model has sufficient ability to process text. In addition, a manual verification step was conducted to ensure classification accuracy and validate the model's reliability. Two trained coders independently analyzed a random sample of 300 comments (150 from Weibo and 150 from X) to check the alignment between automated labels and human interpretation. After 2 coding rounds, the inter-rater reliability (Cohen κ) reached 0.88, confirming the model's high precision in capturing emotional nuances across both platforms. Nevertheless, because the model relies primarily on linguistic and semantic patterns without incorporating broader contextual cues, it may not fully capture context-dependent expressions of sarcasm or other nuanced emotional tones. These methodological constraints were taken into account when interpreting the results.

For topic clustering, the study adopted the BERTopic framework, which outperforms traditional models such as latent Dirichlet allocation and nonnegative matrix factorization when handling short texts [57]. Grootendorst [58] upgraded BERTopic, and this study adopted his improved version. As shown in Figure 3, BERTopic combines BERT embeddings, UMAP for dimensionality reduction, and HDBSCAN clustering to extract key topics from comments, followed by class-based term frequency-inverse document frequency (c-TF-IDF) to generate keywords for each topic.

Figure 2. Correlation matrix of the 9 identified emotion labels. The heatmap displays the Pearson correlation coefficients (r) between all emotion pairs, demonstrating generally low linear correlation across the labels.

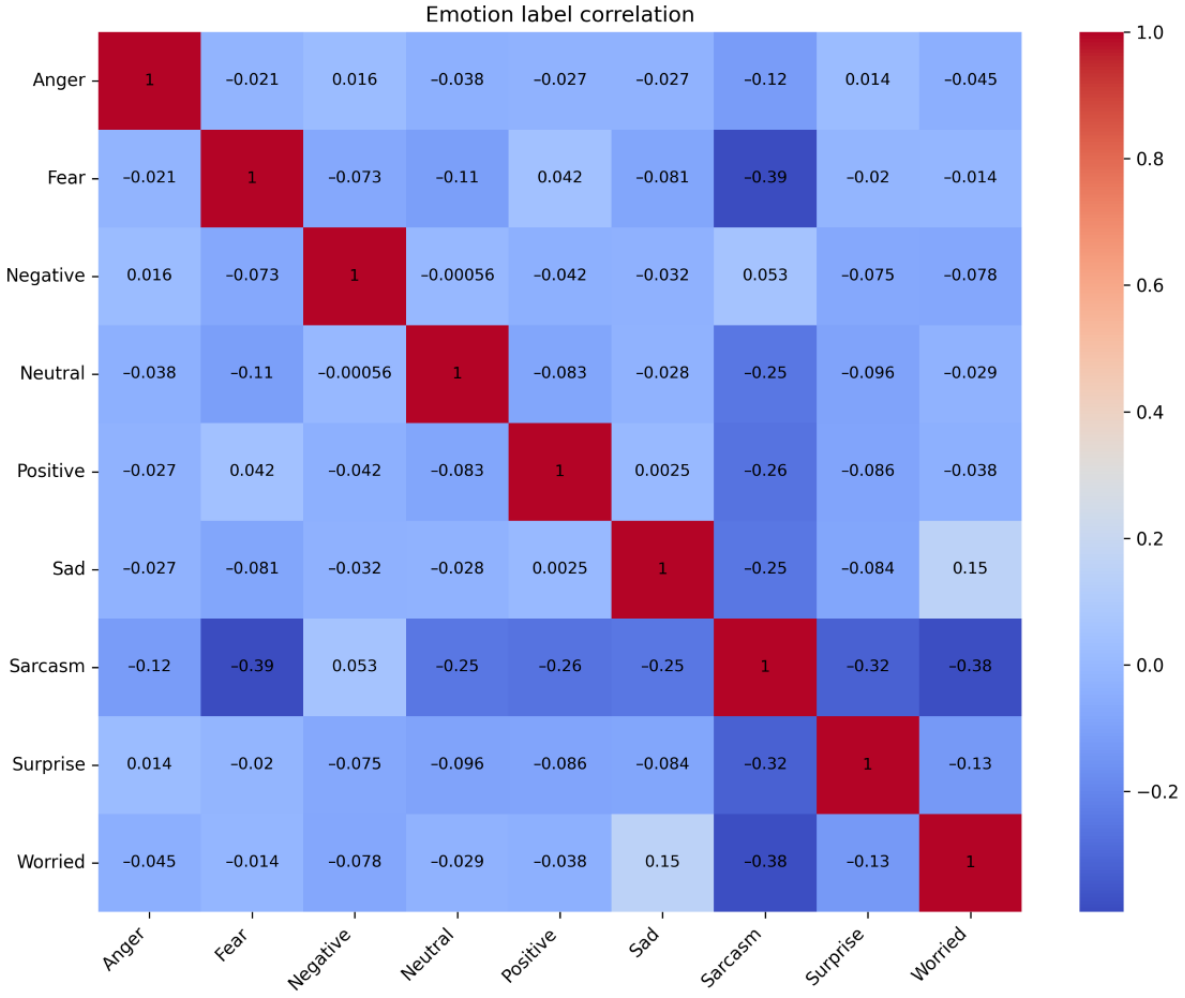

XSL•FO
RenderX

Figure 3. Bidirectional Encoder Representations from Transformers Topic (BERTopic) modeling workflow for extracting topics from social media comments. The process consists of three main steps: (1) document embedding using Bidirectional Encoder Representations from Transformers (BERT) and Transformers, (2) document clustering using Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), and (3) topic creation via class-based term frequency-inverse document frequency (c-TF-IDF) and Maximal Marginal Relevance to generate topic representations and keywords.

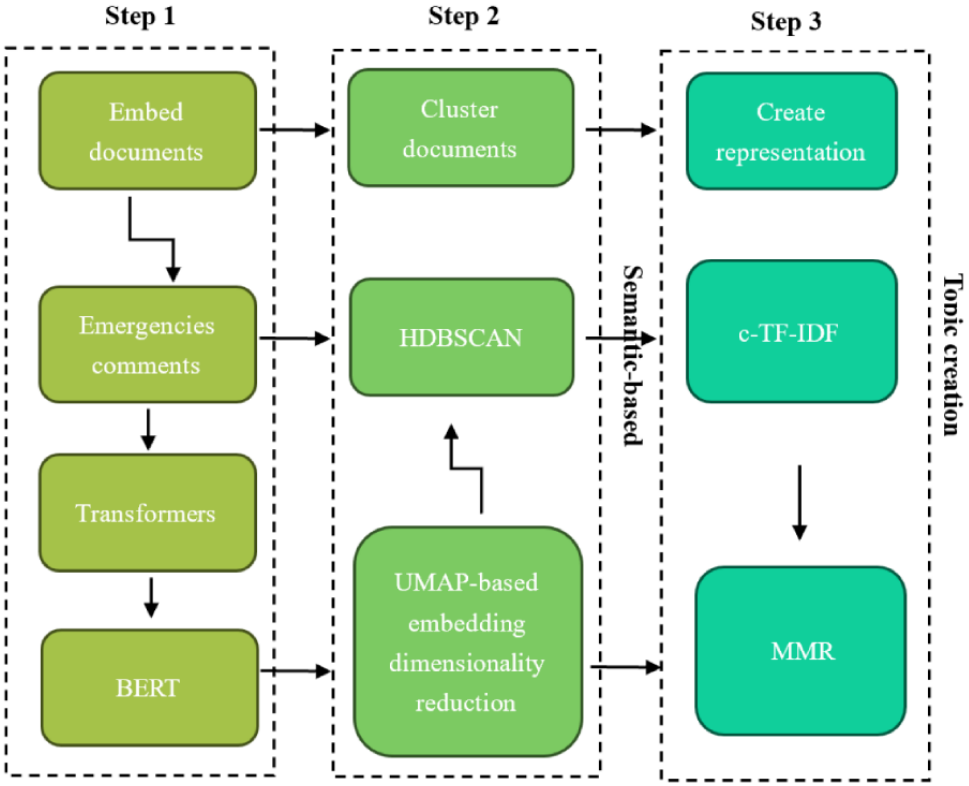

Statistical Analysis

Descriptive statistics were used to summarize distributions of comments and emotion categories. Proportions were reported with their 95% CIs, which served to quantify the precision of the parameter estimates for the analyzed dataset. Chi-square tests were conducted to assess cross-platform differences. Correlation matrices were generated to evaluate overlap among emotion labels. Statistical charts were created to visualize data distributions.

Model Interpretability Analysis

To address the inherent black-box nature of the fine-tuned XLM-RoBERTa model, the study used the Integrated Gradients (IG) attribution method [59]. IG attributes the model's prediction to input features by calculating the path integral of the gradient from a baseline (zero embedding) to the actual input. It satisfies the axioms of sensitivity (changes to an essential input feature must lead to a change in attribution) and implementation invariance (attributions must be independent of the specific model implementation) [60].

The IG attribution score for an input feature is computed as follows [61]:



Where F(x) is the model's prediction function (the target logit), is the input embedding, and χ' is the baseline embedding.

For the XLM-RoBERTa implementation, the attribution target was set to the logit output corresponding to the predicted emotion category. The baseline input "χ'" was defined as the zero embedding vector. This vector represents the absence of textual information and is a standard practice in natural language processing interpretation. The path integral required by IG was numerically approximated using the Riemann sum with 50 steps. A balance between computational efficiency and convergence accuracy was achieved with this setting. Since XLM-RoBERTa uses high-dimensional token embeddings, the raw IG output is a vector for each subword token. The final scalar token-level attribution score was derived by averaging the contribution across all dimensions of the embedding vector. This process yields a single value, which clearly indicates whether a token supports (positive score) or suppresses (negative score) the classification decision.

Qualitative Component

Data Sources and Sampling

The qualitative component was based on the same corpus of publicly available social media comments analyzed in the quantitative phase. Rather than generating new qualitative data or recruiting participants, this study examined naturally occurring textual data posted on Weibo and X (formerly Twitter) in response to emergency events during 2024-2025. These comments represent unsolicited public expressions produced in real-world digital environments and were therefore treated

XSL•FO
RenderX

as textual data sources rather than participant-generated responses.

Qualitative analysis focused on interpreting how emotions and meanings were discursively constructed within the broader patterns identified by computational analysis. Representative excerpts were examined to contextualize emotion categories, topic structures, and semantic relationships observed at the aggregate level. This approach allowed qualitative interpretation to be embedded within the full dataset, supporting explanation and triangulation without introducing a separate qualitative sample.

Manual Coding Procedures

To ensure the methodological integrity of the qualitative discourse analysis, the study conducted a separate theoretical verification. The verification was meant to systematically link emotional themes to Hofstede's cultural dimensions through the emotion-culture mapping framework. Two coders independently analyzed representative comments for theory-driven mapping. The intercoder reliability for this thematic categorization reached a Cohen κ of 0.82. Coding focused on the dominant semantic meaning and pragmatic context, such as sarcasm. This approach helped capture cultural nuance. Any discrepancies were resolved through discussion, and full consensus was reached to provide an empirically supported foundation for cross-cultural interpretations.

Discourse and Pragmatic Analysis

Linguistic differences help reveal cultural disparities. To explain how emotions and comment styles differ across cultures, Hofstede's cultural dimensions theory [45] is applied in this study. For instance, expressions such as "congratulations" in an emergency context were interpreted not as positive but as having a "sarcastic" connotation. Representative comments were analyzed at the semantic level to explain how emotions and comment styles differ between the collectivist context (Weibo) and the individualist context (X, formerly Twitter).

Integration of Quantitative and Qualitative Strands

The study combined findings by linking emotional tone (quantitative) with narrative focus (topic modeling). This mixed methods approach showed how affective orientation and discourse framing together shaped cross-cultural meaning construction. Semantic co-occurrence network analysis was combined with statistical distributions. The research emphasized the interaction between emotion and narrative structure and demonstrated that distinct emotional tendencies matched particular thematic patterns in multilingual emergency communication.

Reflexivity and Research Stance

The study acknowledges that, while the model relies on linguistic patterns, it may not fully capture all context-dependent expressions. Therefore, methodological constraints were taken into account when interpreting the results. The interpretation of "cultural disparities" is framed through the specific lens of Hofstede's theory, acknowledging the distinction between the Chinese cultural context and the Western-dominated context of X (formerly Twitter).

Ethical Considerations

This study drew on publicly accessible comments about 2024-2025 emergency events from X (formerly Twitter) and Weibo. All data were collected with Python scripts, following the platforms' terms of service. The study was deemed exempt from formal human research ethics approval by institutional guidelines, as it involved no direct interaction with individuals and used only publicly available data. Informed consent was not required since the data were public, and platform policies already notify users of potential academic use. To ensure privacy, user identifiers (such as usernames and IP addresses) were collected but immediately discarded, and all textual content was thoroughly anonymized by removing indirect identifiers before analysis. In addition, no images, figures, examples, or supplementary materials included in this paper contain information that could lead to the identification of individual users. All illustrative excerpts and visualizations are fully anonymized and presented in aggregate form. Therefore, no identifiable personal data are disclosed, and additional individual consent was not required. The research strictly followed the ethical principles of the Declaration of Helsinki and complied with major data protection regulations, including the General Data Protection Regulation and the Chinese Cybersecurity Law.

Results

Interpretability Analysis

A comprehensive interpretability analysis was conducted for the 9 emotion labels. The analysis consistently demonstrated that the XLM-RoBERTa model's predictions were highly dependent on tokens with explicit emotional valence, matching established principles of affective linguistics. As examples, local attribution results for the "worried" and "anger" emotion classifications are shown in Figures 4 and 5. For predicting "worried," the model relies heavily on specific entities or situations that may pose harm. Tokens with the highest positive contributions, such as "dest" (destination), "_har" (hardship or harm), and "locat," point to entities or difficulties that clearly strengthen the model's judgment of worry.

For predicting "anger," decisions by the model depend strongly on tokens representing highly critical or aggressive language. The tokens with the highest attribution scores, including _h, isha, and _han, correspond to strongly negative words and serve as key evidence for XLM-RoBERTa to identify anger.

In conclusion, the IG interpretability analysis provides direct evidence validating that XLM-RoBERTa's decision-making mechanism is trustworthy and intuitive. This subword-level analysis reveals that the model effectively distinguishes between relying on specific, contextual tokens and relying on strong, affective tokens, significantly enhancing transparency and trust in the model's operations.

XSL•FO
RenderX

Figure 4. Integrated Gradients (IG) token contribution scores for the "Anger" emotion, validating the Cross-lingual Language Model–Robustly optimized BERT approach (XLM-RoBERTa) model. The bar chart displays tokens that positively (green) or negatively (red) contribute to the model's prediction of "anger".

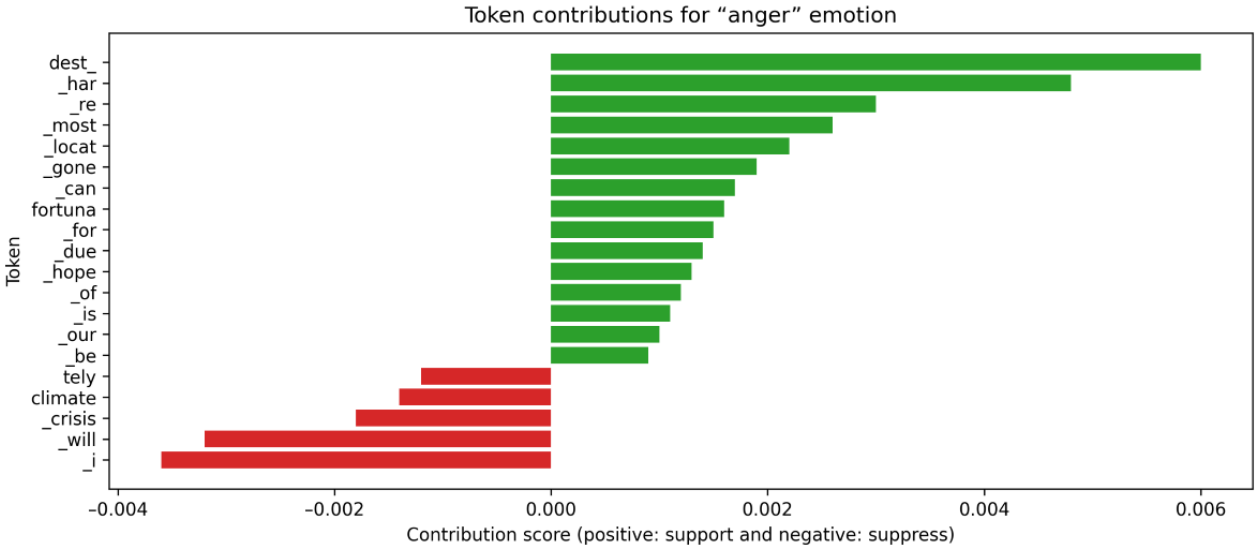

Figure 5. Integrated Gradients (IG) token contribution scores for the "Worried" emotion, validating the Cross-lingual Language Model–Robustly optimized BERT approach (XLM-RoBERTa) model. The bar chart shows the positive (green) and negative (red) contribution scores of key tokens.

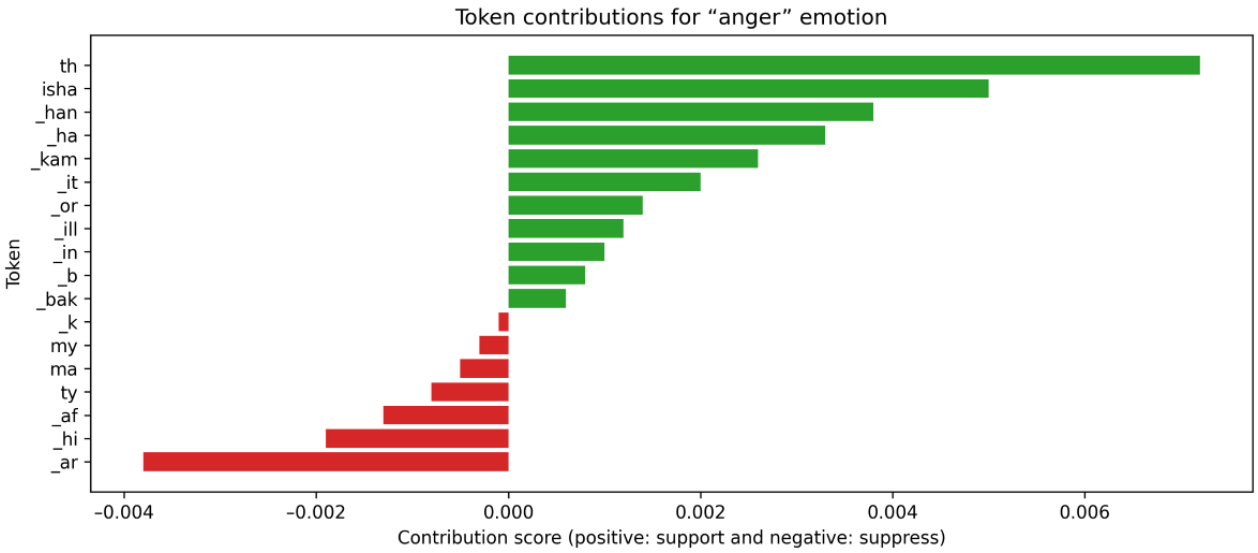

Emotional Distribution and Cultural Difference

Understanding how people feel about emergencies helps us understand how the public perceives risks and their attitudes toward them. This helps policymakers make timely decisions that consider culture. These decisions improve social stability. To test whether there is a significant difference, $P<.001$, in the distribution of emotional labels between 2 platforms, a chi-square test of independence was conducted. As shown in Table 1, the results show that there is a highly significant association between the platforms and the distribution of emotional labels. This indicates that there are obvious statistical differences in the proportions of different emotions expressed by users on Weibo and X (formerly Twitter).

XSL•FO
RenderX

Table 1. Statistical differences in the proportion of emotional labels between Weibo and X (formerly Twitter). The results ($\chi^2$8=8025.60; *P*<.001) indicate a highly significant association between the platform used and the type of emotional labels expressed by users. All expected cell counts exceeded the minimum requirement for chi-square analysis (181.76).

Test statistic	Value (*df*)	Asymptotic significance (2-sided)
Pearson chi-square	8025.598 (8)	<.001
Likelihood ratio	8551.340 (8)	<.001
Linear-by-linear association	3008.296 (1)	<.001
N (valid cases)	29040	—

To visually illustrate the emotional distribution patterns of user comments, Figures 6 and 7 illustrate the monthly emotional distribution patterns for both platforms from 2024 to 2025. Platform X (formerly Twitter) exhibited a significantly higher comment volume; from an initial collection of 19,813 entries, the final analyzed sample for Platform X consisted of 21,600 emotion labels (mean 1800, SD 601.78). Similarly, the Weibo dataset was refined from an original 6536 entries to a final sample of 5527 emotion labels (mean 460, SD 223.13). The distribution of emotional labels on each platform showed certain differences (as shown in Table 2). “Fear” (16.68%) and “Negative” (20.37%) emotions were most prominent on Weibo, while “Sarcasm” (43.49%) dominated on X (formerly Twitter). As the emotion classification was derived from linguistic patterns without full contextual interpretation, the results should be viewed as indicative rather than absolute, particularly for context-dependent emotions such as sarcasm. Overall, the distribution reveals clear cross-platform differences, suggesting culturally distinct emotional response patterns during crises.

On Weibo (Figure 6), negative emotions such as “Fear” and “Worried” rise sharply in crisis months such as July and November. This pattern matches China’s high uncertainty avoidance tendency. Data from X (formerly Twitter; Figure 7) show a more stable distribution of these emotions (coefficient of variation [CV]=0.39), reflecting higher uncertainty tolerance in individualistic cultures with low avoidance traits. Emotional responses on Weibo often evolve into collective worry; X (formerly Twitter) users, however, express emotions more individually.

“Positive” emotions (such as prayers and gratitude) on Weibo increase sharply during disasters. For example, there were 192 labels in May and 105 in December. This pattern matches collectivist cultures’ focus on unity. On platform X (formerly Twitter), “Positive” emotions show a more even distribution (CV=0.51), suggesting higher emotional self-regulation and less need for group comfort.

“Sarcasm” shows the greatest difference. On platform X (formerly Twitter), sarcasm occurs often (n=9393, 43.49%; 95% CI 42.8%-44.2%). It spreads evenly across the year (CV=0.39) and reaches the highest level during emergencies, usually targeting institutional issues. This fits cultures with low power distance and strong free expression norms. On Weibo, sarcasm is rare (n=76), mainly appearing when disaster response is clearly poor. This reflects the caution typical of high power distance and collectivist environments.

“Surprise” appears on both platforms (Weibo: n=1089 and X [formerly Twitter]: n=1913), but the tone differs. On X (formerly Twitter), people explicitly show surprise at event scale or shock, matching direct cultural styles. On Weibo, surprise is brief, quickly turning into emotions such as sympathy or worry. This fits the pattern of emotional control in high-uncertainty-avoidance cultures. These emotional patterns across platforms highlight how culture influences not only the frequency but also the style and timing of public emotional expression in response to emergencies.

XSL•FO
RenderX

Figure 6. Monthly distribution of 9 emotional labels on the Weibo Platform (2024-2025). This stacked bar chart presents the absolute monthly frequency and proportional distribution of the 9 classified emotional tags.

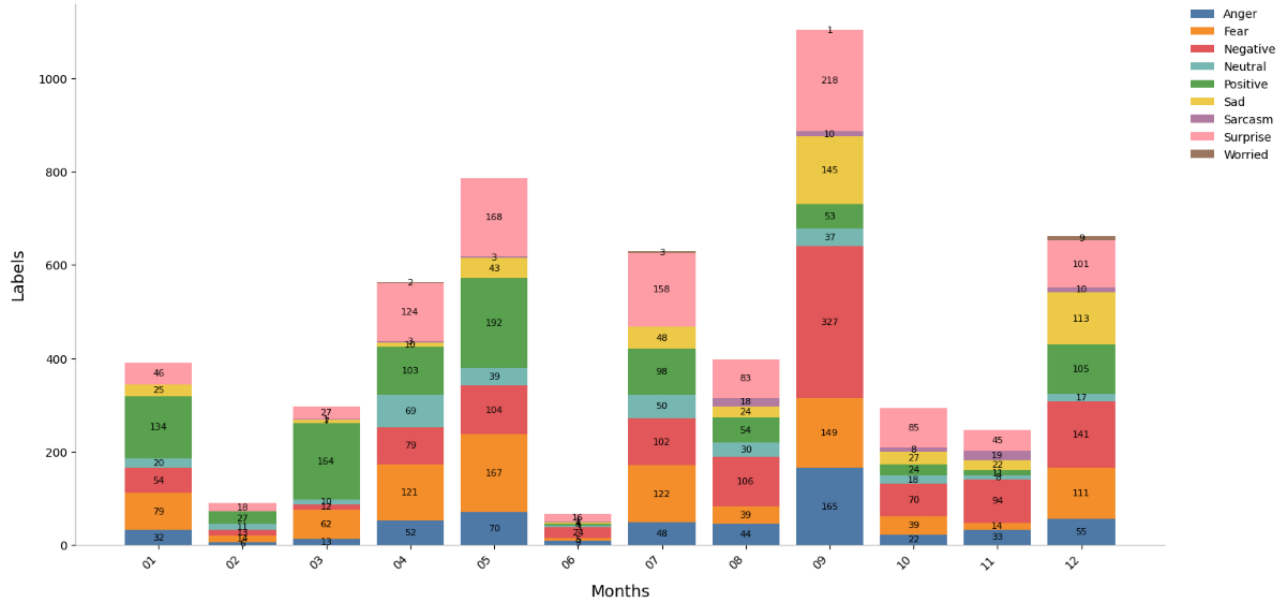

Figure 7. Monthly distribution of 9 emotional labels on the X (formerly Twitter) platform (2024-2025). This stacked bar chart displays the absolute monthly frequency and proportional distribution of the 9 classified emotional tags.

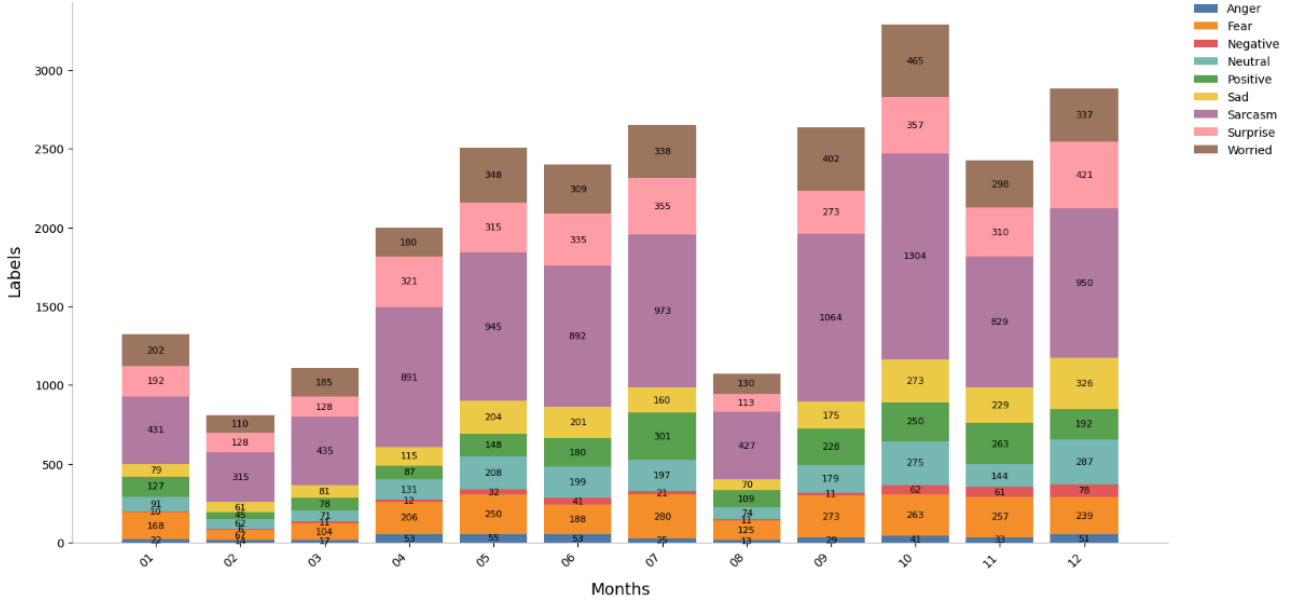

XSL•FO
RenderX

Table 2. The overall emotional distribution comparison between Weibo and X (formerly Twitter).

Tags	Weibo count	Value, % (95% CI)	X (formerly Twitter) count	Value, % (95% CI)	Total count	Value, % (95% CI)	OR[a] (95% CI)
Anger	549	9.93 (9.1-10.7)	406	1.88 (1.7-2.1)	955	3.52 (3.3-3.7)	5.76 (5.04-6.57)
Fear	922	16.68 (15.7-17.7)	2252	10.43 (10.0-10.8)	3174	11.7 (11.3-12.1)	1.72 (1.58-1.87)
Negative	1126	20.37 (19.3-21.4)	345	1.60 (1.4-1.8)	1471	5.42 (5.2-5.7)	15.76 (13.91-17.85)
Neutral	313	5.66 (5.1-6.3)	1918	8.88 (8.5-9.3)	2231	8.22 (7.9-8.6)	0.62 (0.54-0.70)
Positive	967	17.5 (16.5-18.5)	2008	9.30 (8.9-9.7)	2975	10.97 (10.6-11.3)	2.07 (1.90-2.25)
Sad	467	8.45 (7.7-9.2)	1974	9.14 (8.8-9.5)	2441	9.00 (8.7-9.3)	0.92 (0.83–1.02)
Sarcasm	76	1.37 (1.1-1.7)	9393	43.49 (42.8-44.2)	9469	34.91 (34.3-35.3)	0.02 (0.01-0.02)
Surprise	1089	19.7 (18.7-20.8)	1913	8.86 (8.5-9.2)	3002	11.07 (10.7-11.4)	2.53 (2.33-2.74)
Worried	18	0.33 (0.2-0.5)	3304	15.3 (14.8-15.8)	3322	12.25 (11.9-12.6)	0.02 (0.01-0.03)
Total	5527	—	21600	—	27127	—	—

[a]R: odds ratio; calculated with X (formerly Twitter) as the reference group. All OR values include leading zeros for values <1 per journal requirements.

Discovery of Online Public Opinion Topics

In the digital age, understanding the main topics in online public discourse during emergencies is crucial for understanding public attention, concerns, and information dissemination patterns. This section explores the dominant themes that appear on the 2 social media platforms in relation to emergency events. It uses c-TF-IDF to identify and rank key topics.

Figures 8 and 9 display the top 8 themes on Weibo and X (formerly Twitter), respectively. A total of 1071 topics were calculated for X (formerly Twitter), and 345 for Weibo. It can be observed that on both platforms, keywords directly related to emergency events (such as “rainstorm,” “hurricane,” “fire,” “airplane,” “death,” and so on) occupy significant positions. Regardless of cultural background, the public generally shows instinctive concern for disaster impacts when faced with real-world problems such as threats to life and property loss. Both platforms focus heavily on discussions about natural disasters (earthquakes, floods, typhoons, and so on). Characteristic words such as “earthquake,” “rainstorm” (on Weibo),” “hurricane,” “storm,” and “natural disasters” (on X [formerly Twitter]) are all high-frequency themes. This reflects that emergency events attract widespread attention across cultures, consistent with humans’ instinctive responses to survival and safety.

For comparative analysis, this study focused on the 7 most frequent topics within a subset of clusters from Weibo and X (formerly Twitter). For each of these topics, the 10 most frequent topic words were extracted (refer to Tables 3 and 4). The analysis identified representative themes, including “Rainfall impact” and “Fire crisis.”

These were further examined using intertopic distance maps (Figures 10 and 11), which visualize topic prevalence and semantic similarity, revealing that discussions on Weibo were more concentrated, while those on X were more dispersed.

Figure 8. Dominant themes and ranked keywords on the Weibo platform, identified by class-based term frequency-inverse document frequency (c-TF-IDF) (Topics A-H). This figure displays the top 8 (of 345 total) extracted themes, highlighting the most representative Chinese tokens and their c-TF-IDF scores.

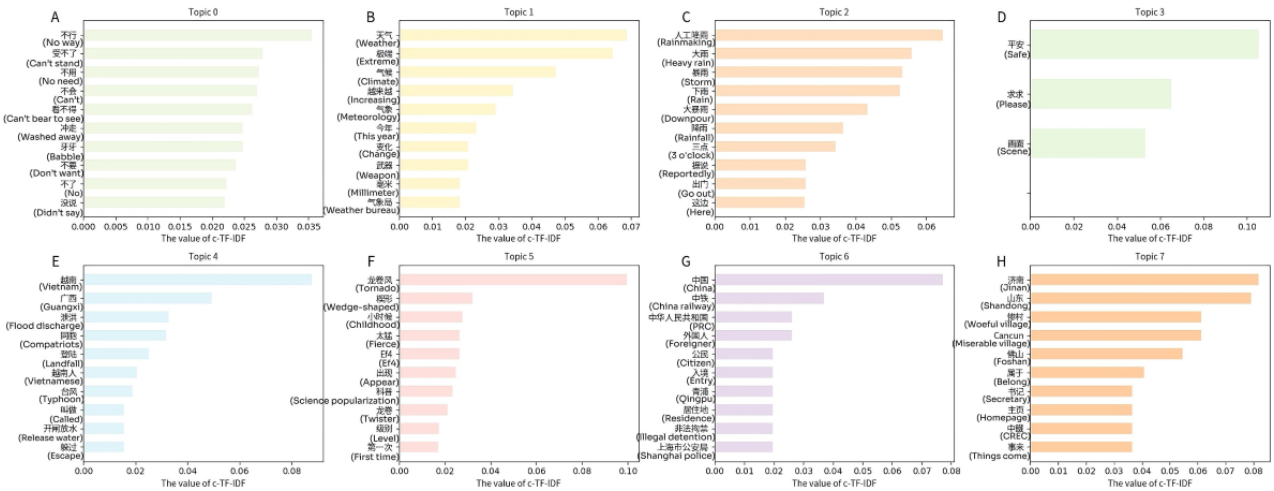

XSL•FO
RenderX

Figure 9. Dominant themes and ranked keywords on the X (formerly Twitter) platform, identified by class-based term frequency-inverse document frequency (c-TF-IDF) (Topics A-H). This figure presents the top 8 (of 1071 total) extracted themes, showing the highest-ranking tokens and their c-TF-IDF scores.

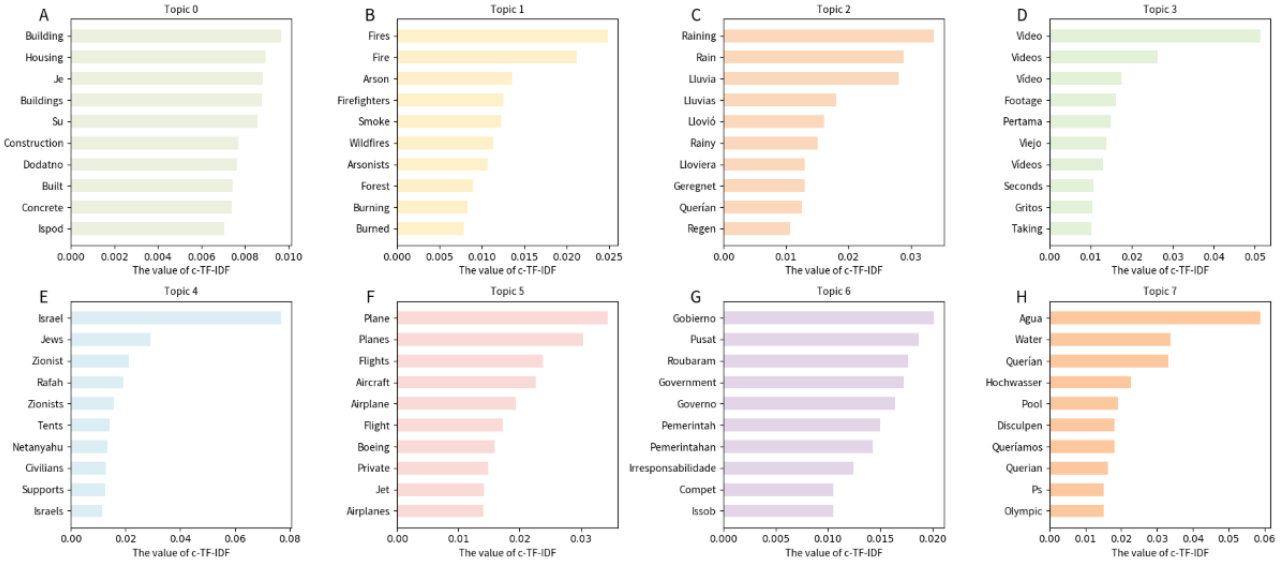

XSL•FO
RenderX

Table 3. The 7 most frequent topics and top keywords within selected clusters on the Weibo platform. This table presents the top 7 themes and their 10 highest-weighted keywords (based on class-based term frequency-inverse document frequency [c-TF-IDF] scores) that dominate public discourse on Weibo regarding emergency events.

Keywords	Weights
Topic 0	
城市(City)	0.03
危险(Danger)	0.07
气象(Meteorological)	0.02
我们(We)	0.01
越南(Vietnam)	0.11
猴子(Monkey)	0.02
他们(They)	0.05
打开(Open)	0.03
三浦(Sanpu)	0.04
这样(This way)	0.03
Topic 1	
下雨(Raining)	0.06
三点(Three o’clock)	0.05
暴雨(Heavy rain)	0.04
出门(Go out)	0.07
这边(Here)	0.03
大雨(Heavy rain)	0.06
凌晨(Early morning)	0.4
下雨天(Rainy day)	0.3
平安(Safety)	0.22
今晚(Tonight)	0.12
Topic 2	
人工降雨(Artificial rain)	0.04
大雨(Heavy rain)	0.06
大暴雨(Torrential rain)	0.04
降雨(Rainfall)	0.02
太多了(Too much)	0.02
强暴雨(Severe rainstorm)	0.46
吓人(Frightening)	0.85
平安(Safety)	0.22
顺利(Smoothly)	0.11
雨势(Rain intensity)	0.12
Topic 3	
洪灾(Floods)	0.03
遭遇(Encounter)	0.03
中东地区(Middle East)	0.02
东非(East Africa)	0.03
上古(Ancient times)	0.01
旱灾(Drought)	0.31
防灾(Disaster prevention)	0.03

XSL•FO
RenderX

Keywords	Weights
失踪(Missing)	0.08
甘蔗(Sugarcane)	0.02
极端(Extreme)	0.09
Topic 4	
火之歌(A Song of Ice and Fire)	0.07
火烧(Burning)	0.04
起火(Catch fire)	0.04
大火(Big fire)	0.04
火灾(Fire disaster)	0.04
纵火(Arson)	0.03
烟头(Cigarette butt)	0.02
烟火(Fireworks)	0.02
绝好(Excellent)	0.02
素材(Material)	0.01
Table 5	
山火(Forest fire)	0.08
扑灭(Extinguish)	0.05
肆虐(Raging)	0.04
尽快(As soon as possible)	0.04
火灾(Fire disaster)	0.04
救火(Firefighting)	0.03
见证(Witness)	0.02
菲政府(Philippine government)	0.01
关爱(Care)	0.03
减轻(Alleviate)	0.02
Topic 6	
祈祷(Pray)	0.32
人们(People)	0.05
保守(Conservative)	0.04
无人(No one)	0.8
恳求(Plead)	0.75
祈祷平安(Pray for safety)	0.17
阿弥陀佛祈祷(Amitabha prayer)	0.8
生活(Life)	0.04
活着(Alive)	0.04
大过年(During Chinese New Year)	0.11

XSL•FO
RenderX

Table 4. The 7 most frequent topics and top keywords within selected clusters on the X (formerly Twitter) platform. This table outlines the top 7 themes and their 10 highest-weighted keywords (based on class-based term frequency-inverse document frequency [c-TF-IDF] scores) identified in the X discourse on emergency events.

Keywords	Weights
Topic 0	
Fires	0.08
Fire	0.07
Arson	0.04
Smoke	0.3
Firefighters	0.03
Arsonists	0.04
Burning	0.02
Burned	0.01
(Fire)	0.04
Flames	0.19
Topic 1	
Gas	0.11
Fuel	0.3
Residential	0.14
Tanker	0.28
Cylinder	0.04
Station	0.03
Lorry	0.05
Embakasi	0.07
Plant	0.03
Cylinders	0.05
Topic 2	
Portugal	0.14
Wildfire	0.05
Fires	0.05
Wildfires	0.04
Spanish	0.04
Portugals	0.04
Portuguese	0.04
Declaran	0.02
Alguém	0.03
Arson	0.04
Topic 3	
Lava	0.07
Lewotobi	0.08
Lakilaki	0.05
Mount	0.09
Ash	0.07
Indonesia	0.15

XSL•FO
RenderX

Keywords	Weights
Evacuations	0.07
Erupted	0.07
Eruptions	0.06
16000	0.04
Topic 4	
Bus	0.17
Thailand	0.11
44	0.77
Children	0.13
Returning	0.08
Siswa	0.05
Guru	0.03
Bangkok	0.07
Teachers	0.11
Thani	0.02
Topic 5	
Gobierno	0.25
Government	0.07
Gobierna	0.12
Governo	0.08
Pemerintah	0.09
Federal	0.03
Pusat	0.06
Roubaram	0.05
Años	0.05
Provincias	0.08
Topic 6	
Rescue	0.16
Underway	0.03
Rescued	0.13
Certain	0.05
Humourrescue	0.03
Baseball	0.02
Operationswhat	0.02
Isvbeing	0.03
Resgate	0.05
Missions	0.1

XSL•FO
RenderX

Figure 10. Bidirectional Encoder Representations from Transformers Topic (BERTopic) intertopic distance map visualizing the semantic distribution and prevalence of topics on Weibo. This map uses UMAP-based dimensionality reduction to plot topics, where the size of each circle corresponds to the topic frequency. The proximity of the circles illustrates semantic similarity. The overall visualization confirms that public discourse on Weibo is thematically concentrated.

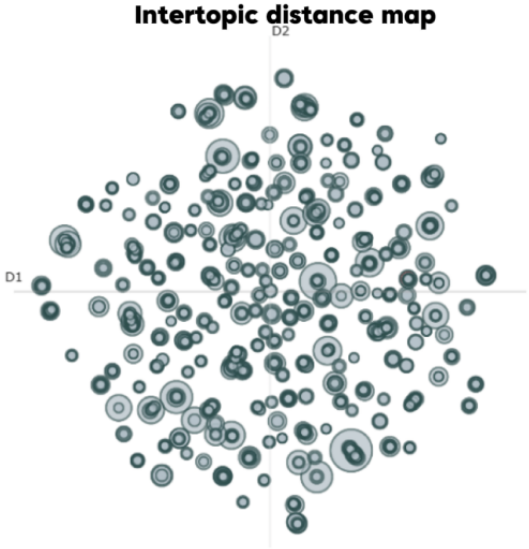

The intertopic distance map for X (formerly Twitter; Figure 11) reveals a relatively dispersed topic structure, suggesting that emergency event discussions cover a wide range of issues. A key trait of its discourse is the coexistence of several distinct clusters alongside some relatively isolated topics. For example, one prominent cluster centered on the immediate impact and characteristics of specific disasters, particularly those related to weather events. As shown in Table 4, this cluster covers Topic 0 (“fire, arson, smoke...”), with a count of 198; Topic 1 (“gas, fuel, residential, tanker”), with a count of 99; and Topic 2 (“portugal, wildfire, fires, wildfires, alguém”), with a count of 30, as evidenced by their proximity in the map and shared keywords. These topics highlight discussions about sudden natural disaster events and the dangers they pose to affected urban areas. Another cluster appears to revolve around the broader disaster context, including Topic 3 (“lava, lewotobi, lakilaki, mount, ash...”), with a count of 23, and Topic 4 (“bus, Thailand, children, returning, siswa...”), with a count of 17. While these topics are distinct, they all focus on sudden fire-related events. Simultaneously, a corresponding yet somewhat independent cluster emerges. It includes Topic 5 (“gobierno, government, gobierna, governo, pemerintah, federal, pusat, roubaram, años, provincias”), with a count of 167. Topic 6 (“rescue, underway, rescued, baseball, operations, isvbeing, resgate, missions”), with a count of 11, stands out as a major theme, indicating a strong emphasis on rescue efforts and on-the-scene situations during emergencies. The topic including “earthquake, sismo, earthquakes, terremoto...” is also relatively large, with a count of 63, suggesting substantial discussion surrounding seismic events. Notably, distinct topics focusing on long-term recovery or preventive measures are relatively scarce, though these elements appear to some extent in earthquake and flood themes.

In contrast, as shown in Figure 10, the topic structure on Weibo is more concentrated, suggesting a stronger connection among themes. Similar to X (formerly Twitter), a significant cluster on Weibo pertains to natural disasters, but with a greater emphasis on direct impacts and local contexts. It is evident within Cluster 1 that Topic 0 from Table 3 (count=51), including “城市 (city),” “危险 (danger),” “气象 (meteorological),” “我们 (we),” “越南 (Vietnam),” “猴子 (monkey),” “他们 (they),” “打开 (open),” “三浦 (Sanpu),” and “这样 (this)”; Topic 1 (count=30), with terms such as “下雨 (raining),” “三点 (3 o’clock),” “暴雨 (heavy rain),” “出门 (go out),” “这边 (here),” “大雨 (heavy rain),” “凌晨 (early morning),” and “下雨天 (rainy day)”; Topic 2 (count=20), including “人工降雨 (artificial rain),” “大雨 (heavy rain),” “大暴雨 (torrential rain),” “降雨 (rainfall),” “太多了 (too much),” “强暴雨 (severe rainstorm),” “吓人 (frightening),” “平安 (peace),” and “顺利 (smoothly)”; and Topic 3 (count=7), with “洪灾 (floods),” “遭遇 (encounter),” “中东地区 (Middle East),” “东非 (East Africa),” “上古 (ancient times),” “旱灾 (drought),” and “边境县 (border county),” are closely located on the map and share keywords related to weather events and urban impacts. Another cluster encompasses similar natural disaster themes, such as Topic 4, including “火之歌 (A Song of Ice and Fire),” “火烧 (burning),” “起火 (catch fire),” “大火 (big fire),” “火灾 (fire disaster),” “纵火 (arson),” “烟头 (cigarette butt),” and “烟火 (fireworks)”; and Topic 5, including “山火 (forest fire),” “扑灭 (extinguish),” “肆虐 (raging),” “尽快 (as soon as possible),” “火灾 (fire disaster),” “救火 (firefighting),” “见证 (witness),” “政府 (government),” “关爱 (care),” and “减轻 (alleviate).” Although different from the rainfall themes, these topics still fall under the broader category of natural disasters. Similarly, themes such as Topic 6 (count=24) are more mixed, including “祈祷 (pray),” “人们 (people),” “保守 (conservative),” “无人 (no one),” “恳求 (plead),” “祈祷平安 (pray for peace),” “阿弥陀佛祈祷 (Amitabha prayer),” and “生活 (life).”

XSL•FO
RenderX

Figure 11. Bidirectional Encoder Representations from Transformers Topic (BERTopic) intertopic distance map visualizing the semantic distribution and prevalence of topics on X (formerly Twitter). This map displays the relative distance and size (topic prevalence) of topics derived from the BERTopic model. The broad spread of the circles across the map confirms that public discourse on X is thematically dispersed and covers a wider range of distinct issues than Weibo.

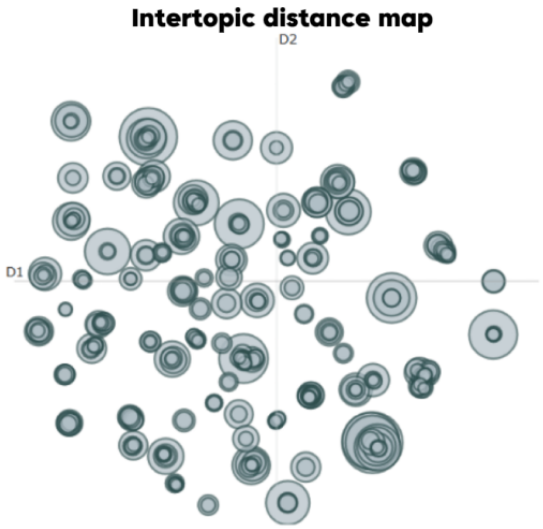

Topic analysis uncovers a common trend. Both platforms place strong emphasis on the direct impact of natural disasters. This focus is supported by keyword clusters related to heavy rainfall, floods, and earthquakes. This shared focus reflects basic human reactions to the dangers posed by these events. Weibo users tend to focus more on the natural phenomena themselves, such as “暴雨 (heavy rain),” “气候变暖 (climate warming),” and “极端天气 (extreme weather).” Their comments primarily describe disaster severity and meteorological anomalies, with emotions dominated by “担忧 (worry)” and “无奈 (helplessness).” In contrast, X (formerly Twitter) users often connect “rainfall” or “storm” to climate change or government action, expressing more explicit criticism. This difference corresponds to Hofstede’s cultural dimensions theory. In highly collectivist societies (Weibo), discourse emphasizes shared communal emotions; in low collectivist contexts (X [formerly Twitter]), it prioritizes individual critique and responsibility attribution.

Under the “Fire crisis” theme, Weibo comments frequently feature keywords such as “消防员 (firefighters),” “肆虐 (raging),” and “扑灭 (extinguish),” expressing respect for firefighters and concern about the spread of disasters. Emotions here are concentrated in “fear” and “worried.” On X (formerly Twitter), fire-related discussions include sympathy for victims, along with more anger toward arsonists and questions about the adequacy of firefighting resources. This phenomenon illustrates how people in high uncertainty avoidance cultures tend to simplify perceived causes of disasters. It also shows how the public in low power distance cultures more directly criticizes institutional failures.

For the “Air crash” theme, Weibo users primarily express “震惊 (surprise)” and “悲痛 (sadness),” using keywords such as “飞机 (airplane),” “故障 (malfunction),” “失控 (out of control),” and “逝者 (the deceased).” This reflects emotional responses and a tendency toward collective mourning when facing uncontrollable disasters. X (formerly Twitter) comments, by contrast, more frequently focus on flight safety standards and airline management, adopting a calmer and more rational tone. This pattern underscores the emphasis on factual accuracy and responsibility in highly individualistic cultural contexts. The theme also includes some sarcastic remarks on X (formerly Twitter), such as comments mocking the slowness of accident investigations, which are barely present on Weibo.

This study shows how emotional expressions and narrative focus differ across cultures. It introduces an “emotion-culture mapping table” (Table 5). The table compares user responses to emergency topics on Weibo and X (formerly Twitter). It also links these response patterns to Hofstede’s cultural dimensions. In addressing possible subjective interpretation, especially when linking emotions and themes to cultural dimensions in Table 5, a verification step was added. A subset of comments was independently analyzed by 2 human coders to confirm the categorization principles in the emotion-culture mapping table. The intercoder reliability test, measured by Cohen κ, showed substantial agreement (κ=0.82). This validation step ensures that the interpretations in Table 5 are not isolated but are systematic and empirically supported by a reliable coding method.

The table outlines 6 cultural dimensions derived from 4 core theories (such as high or low power distance). For each dimension, it details the corresponding “Narrative Focus Examples” and provides relevant “Dimension explanation and link,” thus clearly constructing the mapping relationship between culture and public narratives.

XSL•FO
RenderX

Table 5. The emotion-culture mapping table provides for a clearer understanding of emotions within the framework of Hofstede's cultural dimension theory. This table systematically compares the observed dominant emotions and narrative focus on the 2 platforms and provides a theoretical explanation for observed cross-cultural variations in online discourse regarding emergency events.

Hofstede dimension	Cultural context	Platform	Dominant emotions	Narrative focus examples	Dimension explanation and link
Power distance	Low power distance	X (formerly Twitter)	Anger and sarcasm	Direct criticism of institutional failures.	Low power distance cultures tend to question authority and power structures, and the public is more likely to directly criticize institutional failures.
Power distance	High power distance	Weibo	Fear and worry	Respect for firefighters and concern about the disaster spread.	High power distance cultures tend to respect authority, focus more on collective emotions, and exhibit less criticism of institutions.
Uncertainty avoidance	High uncertainty avoidance	Weibo	Fear and worry	Concern about the disaster spread.	High uncertainty avoidance cultures tend to seek certainty, and explanations of disaster causes may be simplified, with a focus on immediate threats.
Uncertainty avoidance	Low uncertainty avoidance	X (formerly Twitter)	Anger and sarcasm	Anger toward arson and questioning the adequacy of firefighting resources.	Low uncertainty avoidance cultures are more tolerant of uncertainty and tend to explore complex disaster causes, including human and institutional factors.
Individualism	High individualism	X (formerly Twitter)	Anger, sarcasm, and sad	Discussion of flight safety standards, airline management, a strong focus on factual truth, and responsibility attribution.	High individualistic cultures emphasize personal responsibility and autonomy, focusing on factual truth, responsibility attribution, and rational analysis.
Individualism	Low individualism	Weibo	Fear, surprise, and positive (pray and wish)	Emotional outburst and collective mourning.	Low individualistic cultures emphasize group cohesion and collective emotions, and tend to express emotions communally, such as through collective mourning.
Collectivism	High collectivism	Weibo	Worried, fear, and positive	Description of disaster intensity, meteorological anomalies, and sharing of communal emotions.	High collectivist cultures emphasize group connections and shared feelings, and tend to share common worries and helplessness, with a focus on natural phenomena.
Collectivism	Low collectivism	X (formerly Twitter)	Sarcasm	Linking rainfall to climate change and government responses, individual criticism, and responsibility attribution.	Low collectivist cultures tend toward independent thinking and critical analysis, focusing on the underlying causes of events and responsibility.

Semantic co-occurrence network graphs (Figures 12 and 13) were constructed based on word adjacency data (the top 155 terms). In the graphs, node size indicated word frequency, edge thickness represented co-occurrence strength, and colors denoted semantic communities. This approach helps to reveal semantic cores, thematic connections between clusters, and cultural differences in public discourse across platforms. Overall, the Weibo network displayed cohesive and centralized emotional narratives, whereas the X (formerly Twitter) network showed fragmented and decentralized discourse patterns. Unlike topic clustering (Figures 10 and 11), this method did not group texts by overall theme similarity; instead, it focused on word connections at a lexical level, providing a bottom-up perspective of how meaning clusters naturally formed in the discourse.

XSL•FO
RenderX

Figure 12. Semantic co-occurrence network graph of top terms on the Weibo platform. The network visualizes the thematic connections and semantic cores in the public discourse, constructed from the top 155 most frequent terms. Node size represents word frequency, edge thickness indicates co-occurrence strength, and colors delineate distinct thematic clusters.

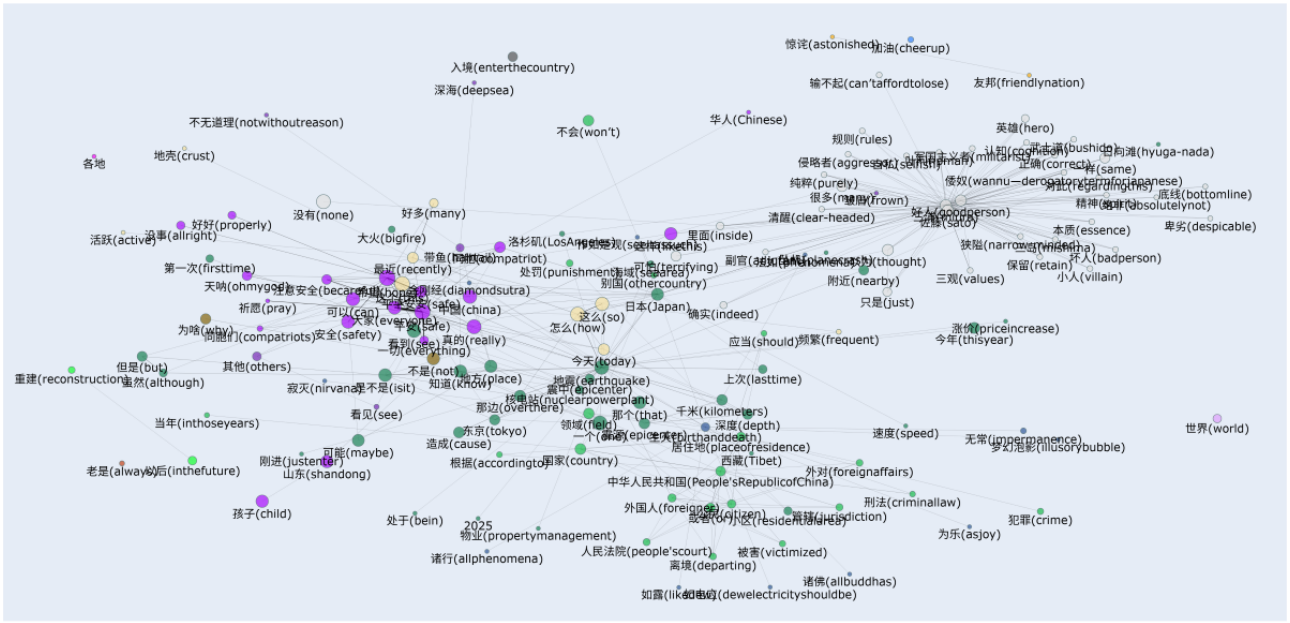

Figure 13. Semantic co-occurrence network graph of top terms on the X (formerly Twitter) platform. The network visualizes the thematic connections and semantic cores in the public discourse, constructed from the top 155 most frequent terms. Node size represents word frequency, the thickness of the edge indicates the co-occurrence strength, and colors denote thematic clusters.

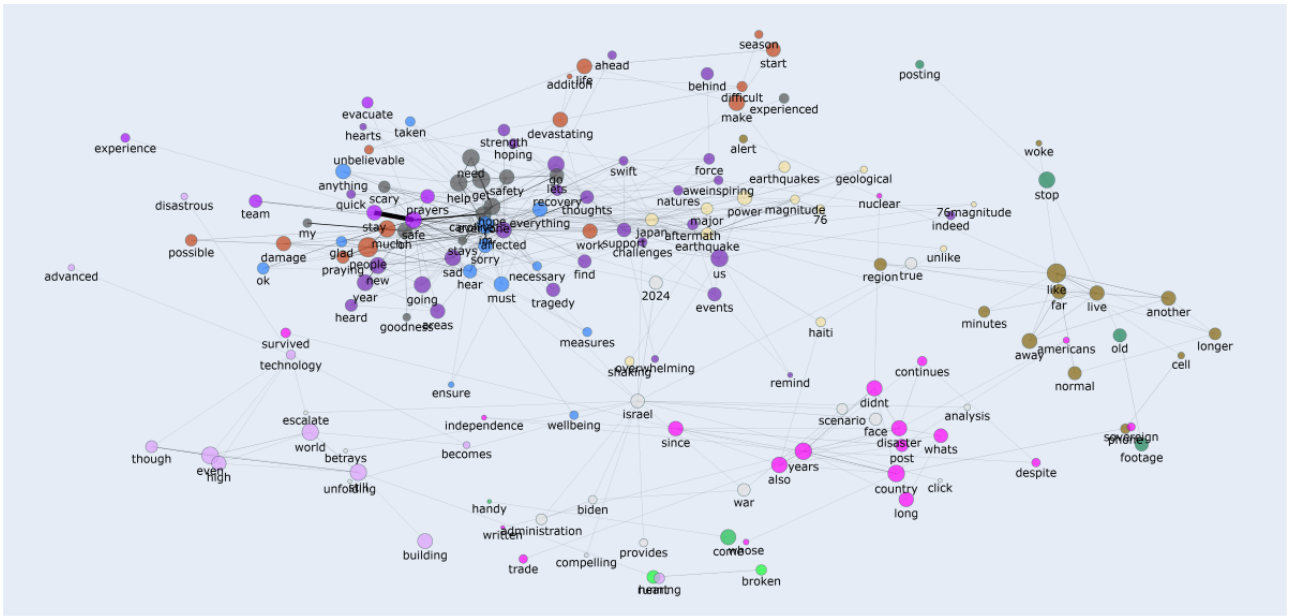

Take Weibo as an example. On this social platform, words such as “平安 (safety),” “祈祷 (pray),” “安置 (resettlement),” and “地方 (local)” created tight clusters. This finding suggests a strong narrative connection among emotional support, local identity, and institutional action. Conversely, semantic clusters on X (formerly Twitter; eg, “pray,” “rescue,” “fire,” and “government”) were more scattered. This pattern shows narratives that were emotional yet fragmented; instead of shaping a single shared story, multiple narrative threads exist side by side. High-frequency words found in topic clustering (eg, “fire” and “floods”) spread across different parts of the network. These words also appear together with various other keywords, highlighting emotional differences that topic models alone failed to capture. One specific case is “fire” on X (formerly Twitter). In 1 cluster, it appears alongside “arson” and “anger,” while in another cluster, it is linked to “rescue” and “prayers.” This example proves there were emotional complexities that a single topic label could not fully represent.

These patterns are consistent with the cultural differences discussed previously. The semantic cohesion observed in the Weibo graph suggests that collectivist cultures tend to build discourse around shared symbols and centralized values. In contrast, the scattered distribution of words on the X (formerly Twitter) platform implies a more decentralized response style, which is typical of individualistic cultures. Therefore, the semantic co-occurrence network analysis does not merely repeat

XSL•FO
RenderX

the conclusions of topic clustering. Instead, it contributes to a deeper understanding of how users interconnected emotions, facts, and cultural expressions during emergencies.

Discussion

Principal Findings

This study provides empirical evidence that cultural values influence emotional expression and discourse during global emergencies, directly relating to the research goals stated in the introduction. The study used a large-scale dataset of comments from Weibo and X (formerly Twitter) (2024-2025) and identified significant cross-platform differences in both emotional distribution and topic structures. On Weibo, negative emotions such as fear and worry are more common, reflecting China's high uncertainty avoidance. Collective expressions, including prayers and wishes for safety, are also frequent, reflecting collectivist orientations. By contrast, X (formerly Twitter) comments contain more sarcasm and criticism, reflecting individualistic values and lower power distance.Topic modeling supported these patterns: Weibo discourse focuses more on disaster severity and communal support, whereas X (formerly Twitter) discourse emphasizes accountability, institutional response, and climate-related narratives. These findings confirm that Hofstede's cultural dimensions—including individualism, collectivism, power distance, and uncertainty avoidance—are useful for explaining cultural differences in online crisis communication.

Comparison With Previous Work

The study supports earlier findings that cultural background strongly shapes emotional expression in crises. On Chinese platforms, collectivist traits dominate, as noted by Zhang [11], where prayers and calls for peace act as clear signs of group unity. This corresponds closely to the Weibo data analyzed in the research, where high-frequency expressions include keywords such as "pray" and "wish for safety." These expressions demonstrate the tendency of collective emotions. From a Western context perspective, the sarcastic and individualized responses described by Imran et al [62] and Maynard and Greenwood [63] match the study's observations. Sarcasm appeared widespread on X (formerly Twitter), often targeting institutional failures. Wang et al [7] also emphasized that cultural orientation affects how people attribute responsibility and process crisis-related information. The study extends this perspective: X (formerly Twitter), which has lower power distance, shows users openly criticizing authorities. On Weibo, however, higher power distance prevails, and emotions lean more toward worry and collective solidarity. The study by Matsumoto et al [64] emphasized the impact of cultural differences on emotional expression, yet it relied heavily on experimental and questionnaire data, with no inclusion of natural language evidence in the context of social media. Building on this foundation, this study helps fill the gap in empirical natural language research within this field, providing a new supporting dimension for the association between cultural differences and emotional expression.

The research also matches earlier studies when examining discourse structure and interaction styles. Han and Wang [65] and Chen and Yik [66] pointed out that Chinese users' discourse tends to emphasize group harmony and shared emotions. This was confirmed by the topic and semantic network analysis in this study, which revealed cohesive clusters around words such as "pray" and "safety." Discourse on X (formerly Twitter), by contrast, appeared more fragmented. Loosely connected clusters formed around terms such as "damage," "government," and "help," a pattern consistent with Gu et al [67], who stressed the openness and diversity of Western crisis communication. In addition, previous research has shown that in low power distance cultures, Western publics are more likely to hold institutions directly accountable [6,68]. The study contributes to this body of work by demonstrating that sarcasm, specifically, acts as a discursive marker of individualistic cultural values. In addition, as noted in the study by Du et al [69], although Chinese participants scored slightly lower than American participants in understanding sarcasm, they pointed out that "collectivism" is instead associated with better sarcasm comprehension. However, understanding sarcasm does not equate to using sarcasm. This is influenced by cultural contexts and platform rules. Precisely due to this discrepancy, only a minimal amount of sarcasm was observed in the Weibo corpus analyzed in this study. This finding demonstrates the importance of this study's focus on cultural dimensions. When these observations are taken together, the research both confirms and advances previous scholarship. It integrates Hofstede's cultural dimensions with a multimethod analysis, and in doing so, provides a more comprehensive account of the culturally distinctive online responses observed.

Limitations and Future Directions

This study has several limitations. First, the dataset was limited to 2 platforms, which, while influential, cannot fully represent the diversity of cultural and digital contexts. Future research should include additional platforms such as Reddit, YouTube, or regional networks to enhance representativeness. Second, cultural differences were inferred from platform-level data, which may risk oversimplifying individual identities and cross-cultural hybridity. Third, although XLM-RoBERTa and BERTopic provided robust classification, automated labeling relies on linguistic and semantic cues and does not fully incorporate contextual validation. Because of this, some detailed expressions, such as sarcasm, might be undercounted or labeled wrong. Manual validation and mixed methods approaches could strengthen accuracy in future studies. Finally, while the 2024-2025 time frame captured multiple emergencies, it did not account for the effects of event type or severity. Longer longitudinal studies across multiple regions would help clarify how cultural and contextual factors interact in shaping online discourse.

Although event scale, comment volume, and posting time were not the main focus, they could still indirectly influence the observed cross-cultural differences in emotion and discourse patterns.

Conclusions

This study shows that digital emergency discourse is structured by culturally embedded values rather than operating as a neutral information space. Cross-platform analysis demonstrates

XSL•FO
RenderX

systematic associations between cultural dimensions and patterns of emotional expression, narrative focus, and responsibility attribution. By operationalizing cultural dimensions through a mixed methods analysis of large-scale, multilingual, naturally occurring social media data, this study advances infodemiology scholarship from descriptive mapping of online content to theory-informed, mechanism-oriented analysis, beyond survey-based and single-platform approaches. It provides a replicable comparative framework for examining how cultural meaning systems are translated into observable digital traces under conditions of uncertainty, risk, and collective sense-making.

From an applied perspective, the findings offer concrete guidance for multiple stakeholders. For governments and public health authorities, the results highlight the need to tailor risk communication to cultural expectations by emphasizing reassurance, unity, and structured guidance in high–uncertainty-avoidance contexts, and transparency, responsiveness, and accountability in low–power-distance contexts. For international organizations, the findings suggest that global crisis messaging should prioritize culturally neutral framing, shared risk narratives, and locally adaptable templates to minimize misunderstanding across cultural boundaries. For digital platforms, culturally sensitive governance is essential, including rapid rumor suppression in high-anxiety environments and careful differentiation between legitimate public critique and harmful misinformation in low–power-distance contexts. Together, these implications support the design of more equitable, trustworthy, and context-aware infodemic management strategies for future public health and safety emergencies.

Acknowledgments

The authors declare that generative artificial intelligence (ChatGPT 5; OpenAI) was used under complete human oversight to assist with grammatical and lexical issues while writing the paper. All changes were human-verified and further edited by the authors, who accept full responsibility for the final content.

Funding

No external financial support or grants were received for this work.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
JARS (Journal Article Reporting Standards) reporting checklist.
[DOCX File , 30 KB - jmir_v28i1e84648_app1.docx]

References

1. Dubovi I, Tabak I. Interactions between emotional and cognitive engagement with science on YouTube. Public Underst Sci 2021;30(6):759-776 [FREE Full text] [doi: 10.1177/0963662521990848] [Medline: 33546572]
2. Gul S, Shah TA, Ahad M, Mubashir M, Ahmad S, Gul M, et al. Twitter sentiments related to natural calamities. EL 2018;36(1):38-54. [doi: 10.1108/el-12-2015-0244]
3. Li S, Sun X. Application of public emotion feature extraction algorithm based on social media communication in public opinion analysis of natural disasters. PeerJ Comput Sci 2023;9:e1417 [FREE Full text] [doi: 10.7717/peerj-cs.1417] [Medline: 37346715]
4. Halse SE, Tapia A, Squicciarini A, Caragea C. An emotional step toward automated trust detection in crisis social media. Inf Commun Soc 2017;21(2):288-305. [doi: 10.1080/1369118x.2016.1272618]
5. Reuter C, Spielhofer T. Towards social resilience: a quantitative and qualitative survey on citizens' perception of social media in emergencies in Europe. Technol Forecast Soc Change 2017;121:168-180. [doi: 10.1016/j.techfore.2016.07.038]
6. Rahmani M, Muzwagi A, Pumariega AJ. Cultural factors in disaster response among diverse children and youth around the world. Curr Psychiatry Rep 2022;24(10):481-491 [FREE Full text] [doi: 10.1007/s11920-022-01356-x] [Medline: 35953637]
7. Wang B, Liu B, Zhang Q. An empirical study on twitter's use and crisis retweeting dynamics amid Covid-19. Nat Hazards (Dordr) 2021;107(3):2319-2336 [FREE Full text] [doi: 10.1007/s11069-020-04497-5] [Medline: 33469243]
8. Cheng B. Investigating Chinese Microblogging Through a Citizen Journalism Perspective. Sydney: University of Technology Sydney; 2020. URL: https://www.proquest.com/openview/7c45b8844cf6ac9e5d46cbaae1ee0741/1.pdf?pq-origsite=gscholar&cbl=2026366&diss=y

XSL•FO
RenderX

9. Hyvärinen H, Beck R. Emotions trump facts: the role of emotions on social media: a literature review. Published online 2018 Jan 3:1797-1806 [FREE Full text] [doi: 10.24251/HICSS.2018.226]
10. Java A, Song X, Finin T, Tseng B. Why we Twitter: understanding microblogging usage and communities. 2007 Presented at: WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis; August 12, 2007; San Jose, California p. 56-65. [doi: 10.1145/1348549.1348556]
11. Zhang Q. Research on the emotional sentiment of TikTok public opinion in response to sudden natural disasters? A case study of the Zhengzhou severe rainstorm disaster. In: Master's Thesis. China, Henan, Zhengzhou: Zhengzhou University of Aeronautics; 2024.
12. Recommendations on digital interventions for health system strengthening. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2025-09-20]
13. Al-Dmour H, Masa'deh R, Salman A, Abuhashesh M, Al-Dmour R. Influence of social media platforms on public health protection against the COVID-19 pandemic via the mediating effects of public health awareness and behavioral changes: integrated model. J Med Internet Res 2020;22(8):e19996 [FREE Full text] [doi: 10.2196/19996] [Medline: 32750004]
14. Welch V, Petkovic J, Simeon R, Presseau J, Gagnon D, Hossain A, et al. PROTOCOL: Interactive social media interventions for health behaviour change, health outcomes, and health equity in the adult population. Campbell Syst Rev 2018;14(1):1-38 [FREE Full text] [doi: 10.1002/CL2.213] [Medline: 37131397]
15. Al-Dmour H, Masa'deh R, Salman A, Al-Dmour R, Abuhashesh M. The role of mass media interventions on promoting public health knowledge and behavioral social change against COVID-19 pandemic in Jordan. Sage Open 2022;12(1):21582440221082125. [doi: 10.1177/21582440221082125]
16. Ju I, Ohs J, Park T, Hinsley A. Interpersonal communication influence on health-protective behaviors amid the COVID-19 crisis. Health Commun 2023;38(3):468-479. [doi: 10.1080/10410236.2021.1956038] [Medline: 34313168]
17. Mheidly N, Fares J. Leveraging media and health communication strategies to overcome the COVID-19 infodemic. J Public Health Policy 2020;41(4):410-420 [FREE Full text] [doi: 10.1057/s41271-020-00247-w] [Medline: 32826935]
18. Siddiqui S, Singh T. Social media its impact with positive and negative aspects. IJCATR 2016;5(2):71-75. [doi: 10.7753/ijcatr0502.1006]
19. Jakob J, Chan C, Dobbrick T, Wessler H. Discourse integration in positional online news reader comments: patterns of responsiveness across types of democracy, digital platforms, and perspective camps. New Media Soc 2023;26(11):6796-6814. [doi: 10.1177/14614448231183704]
20. Knuth D, Szymczak H, Kuecuekbalaban P, Schmidt S. Social media in emergencies: How useful can they be. 2016 Presented at: 3rd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM); December 13-15, 2016; Vienna, Austria p. 1-7. [doi: 10.1109/ict-dm.2016.7857226]
21. Portwood JD. Book review: international and comparative industrial relations: culture's consequences: international differences in work-related values. ILR Review 1982;36(1):129-130. [doi: 10.1177/001979398203600113]
22. Alsaleh DA, Elliott MT, Fu FQ, Thakur R. Cross-cultural differences in the adoption of social media. JRIM 2019;13(1):119-140. [doi: 10.1108/jrim-10-2017-0092]
23. Wood A, Kleinbaum AM, Wheatley T. Cultural diversity broadens social networks. J Pers Soc Psychol 2023;124(1):109-122. [doi: 10.1037/pspi0000395] [Medline: 35266781]
24. Naab TK, Küchler C. Content Analysis in the Research Field of Online User Comments. In: Oehmer-Pedrazzi F, Kessler SH, Humprecht E, Sommer K, Castro L, editors. Standardisierte Inhaltsanalyse in der Kommunikationswissenschaft – Standardized Content Analysis in Communication Research. Wiesbaden, Germany: Springer Fachmedien Wiesbaden; 2023:441-450.
25. Manuel K, Indukuri K, Krishna P. Analyzing internet slang for sentiment mining. 2010 Presented at: Second Vaagdevi International Conference on Information Technology for Real World Problems; December 09-11, 2010; Warangal, India p. 9-11. [doi: 10.1109/vcon.2010.9]
26. Bi Y. The influence of differences between Chinese and English thinking patterns on Chinese-English translation. Int J Soc Sci Educ Res 2021;4(5):284-289 [FREE Full text] [doi: 10.6918/IJOSSER.202105_4(5).0038]
27. Sener B, Akpinar E, Ataman MB. Unveiling the dynamics of emotions in society through an analysis of online social network conversations. Sci Rep 2023;13(1):14997 [FREE Full text] [doi: 10.1038/s41598-023-41573-9] [Medline: 37696868]
28. Dewaele J, Pavlenko A. Languages and emotions: a crosslinguistic perspective. J Multiling Multicult Dev 2004;25(2-3):93-93. [doi: 10.1080/01434630408666522]
29. Kant N, Puri R, Yakovenko N, Catanzaro B. Practical text classification with large pre-trained language models. arXiv 2018 [FREE Full text] [doi: 10.48550/arXiv.1812.01207]
30. Ou Y, de Bruijn G, Schulz PJ. Social media as an emotional barometer: bidirectional encoder representations from transformers-long short-term memory sentiment analysis on the evolution of public sentiments during influenza A on Sina Weibo. J Med Internet Res 2025;27:e68205 [FREE Full text] [doi: 10.2196/68205] [Medline: 40900625]
31. Multi-class emotion classification for short texts. GitHub. 2020. URL: https://tlkh.github.io/text-emotion-classification/ [accessed 2025-07-16]
32. Anno S, Kimura Y, Sugita S. Using transformer-based models and social media posts for heat stroke detection. Sci Rep 2025;15(1):742 [FREE Full text] [doi: 10.1038/s41598-024-84992-y] [Medline: 39753702]

XSL•FO
RenderX

33. Leburu-Dingalo T, Ntwaagae K, Motlogelwa N, Thuma E, Mudongo M. Application of XLM-RoBERTa for multi-class classification of conversational hate speech. 2022 Presented at: Forum for Information Retrieval Evaluation (FIRE) 2021; December 13-17, 2021; Virtual, India p. 590-595 URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85160755477&partnerID=40&md5=e0fdc9ca66e64b5de9146ab6521d980e
34. Barbieri F, Anke LE, Camacho-Collados J. XLM-T: multilingual language models in Twitter for sentiment analysis and beyond. 2022 Presented at: Proceedings of the Thirteenth Language Resources and Evaluation Conference; 20-25 June, 2022; Marseille, France p. 258-266.
35. Xu WW, Tshimula JM, Dubé È, Graham JE, Greyson D, MacDonald NE, et al. Unmasking the twitter discourses on masks during the COVID-19 pandemic: user cluster-based BERT topic modeling approach. JMIR Infodemiology 2022;2(2):e41198 [FREE Full text] [doi: 10.2196/41198] [Medline: 36536763]
36. Khodeir N, Elghannam F. Efficient topic identification for urgent MOOC Forum posts using BERTopic and traditional topic modeling techniques. Educ Inf Technol 2024;30(5):5501-5527. [doi: 10.1007/s10639-024-13003-4]
37. Zhang Z, Cui L, Fang A, Pan Z, Zhang Z, Zhang H. Information dissemination analysis using a time-weight null model: a case study of Sina micro-blog. IEEE Access 2018;6:71181-71193. [doi: 10.1109/access.2018.2881514]
38. Han D, Wei F, Bai L, Tang X, Zhu T, Wang G. An algorithm of Sina microblog user's sentimental influence analysis based on CNN+ELM model. 2018 Presented at: Proceedings of ELM 2018; November 21-23, 2018; Singapore, Singapore p. 86-97. [doi: 10.1007/978-3-030-23307-5_10]
39. Singh LG, Singh SR. Sentiment analysis of tweets using text and graph multi-views learning. Knowl Inf Syst 2024;66(5):2965-2985. [doi: 10.1007/s10115-023-02053-8]
40. Kada A, Chouikh A, Mellouli S, Prashad AJ, Straus SE, Fahim C. An exploration of Canadian government officials' COVID-19 messages and the public's reaction using social media data. PLoS One 2022;17(9):e0273153 [FREE Full text] [doi: 10.1371/journal.pone.0273153] [Medline: 36054094]
41. Chen Y, Ping S. Effectiveness of emergency information for sudden natural disaster events in the new media era: grounded analysis based on WSR methodology. J Kunming Univ Sci Technol (Nat Sci Ed) 2024;49(02):182-193 [FREE Full text] [doi: 10.16112/j.cnki.53-1223/n.2024.02.292]
42. Hasan M, Islam L, Jahan I, Meem SM, Rahman RM. Natural language processing and sentiment analysis on Bangla social media comments on Russia–Ukraine war using transformers. Vietnam J Comp Sci 2023;10(03):329-356. [doi: 10.1142/s2196888823500021]
43. Babalola O, Ojokoh B, Boyinbode O. Comprehensive evaluation of LDA, NMF, and BERTopic's performance on news headline topic modeling. J. Comput. Theor. Appl 2024;2(2):268-289. [doi: 10.62411/jcta.11635]
44. Ma L, Chen R, Ge W, Rogers P, Lyn-Cook B, Hong H, et al. AI-powered topic modeling: comparing LDA and BERTopic in analyzing opioid-related cardiovascular risks in women. Exp Biol Med (Maywood) 2025;250:10389 [FREE Full text] [doi: 10.3389/ebm.2025.10389] [Medline: 40093658]
45. Hofstede G. Dimensionalizing cultures: the Hofstede model in context. ORPC 2011;2(1):1-26. [doi: 10.9707/2307-0919.1014]
46. Masoud R, Liu Z, Ferianc M, Treleaven P, Rodrigues M. Cultural alignment in large language models: an explanatory analysis based on Hofstede's cultural dimensions. arXiv. 2023. URL: https://arxiv.org/abs/2309.12342 [accessed 2026-01-11]
47. Alqarni A. Hofstede's cultural dimensions in relation to learning behaviours and learning styles: a critical analysis of studies under different cultural and language learning environments. J Lang Linguist Stud 2022;18(2022):721-739 [FREE Full text]
48. Jan J, Alshare KA, Lane PL. Hofstede's cultural dimensions in technology acceptance models: a meta-analysis. Univ Access Inf Soc 2022;23(2):717-741. [doi: 10.1007/s10209-022-00930-7]
49. Köksal A, Özgür A. Twitter dataset and evaluation of transformers for Turkish sentiment analysis. 2021 Presented at: 29th Signal Processing and Communications Applications Conference (SIU); June 09-11, 2021; Istanbul, Turkey p. 1-4. [doi: 10.1109/siu53274.2021.9477814]
50. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
51. von EE, Altman D, Egger M. The STROBE reporting checklist. EQUATOR Network. 2025. URL: https://resources.equator-network.org/reporting-guidelines/strobe/strobe-checklist.docx [accessed 2025-10-28]
52. Gaikwad A, Belhekar P, Kottawar V. Advancing multilingual sentiment understanding with XGBoost, SVM, and XLM-RoBERTa. 2025 Presented at: Proceedings of the 5th International Conference on Data Science, Machine Learning and Applications; Volume 2; December 15-16, 2023; Hyderabad, India p. 990-1000. [doi: 10.1007/978-981-97-8043-3_155]
53. Rasool A, Aslam S, Hussain N, Imtiaz S, Riaz W. nBERT: harnessing NLP for emotion recognition in psychotherapy to transform mental health care. Information 2025;16(4):301. [doi: 10.3390/info16040301]
54. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: Proceedings of NAACL-HLT 201; June 2-7, 2019; Minneapolis, Minnesota p. 4171-4186. [doi: 10.18653/v1/2024.naacl-short]
55. Grice HP. Logic and conversation. In: Speech Acts. Leiden, The Netherlands: Brill; 1975:41-58.

XSL•FO
RenderX

56. McNamee P, Duh K. An extensive exploration of back-translation in 60 languages. 2023 Presented at: Findings of the Association for Computational Linguistics: ACL 2023; July, 2025; Toronto, Canada p. 8166-8183.
57. de Groot M, Aliannejadi M, R. Haas M. Experiments on generalizability of BERTopic on multi-domain short text. arXiv 2022 [FREE Full text] [doi: 10.48550/arXiv.2212.08459]
58. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794 2022 [FREE Full text] [doi: 10.48550/ARXIV.2203.05794]
59. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. 2017 Presented at: ICML'17: Proceedings of the 34th International Conference on Machine Learning - Volume 70; August 6-11, 2017; Sydney, Australia p. 3319-3328 URL: https://proceedings.mlr.press/v70/sundararajan17a.html
60. Sanyal S, Ren X. Discretized integrated gradients for explaining language models. 2021 Presented at: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; November 7-11, 2021; Punta Cana, Dominican Republic p. 10285-10299. [doi: 10.18653/v1/2021.emnlp-main.805]
61. Ortigossa ES, Dias FF, Barr B, Silva CT, Nonato LG, Ortigossa ES. T-explainer: a model-agnostic explainability framework based on gradients. IEEE Intell Syst 2025;40(5):34-44. [doi: 10.1109/mis.2025.3564330]
62. Imran AS, Daudpota SM, Kastrati Z, Batra R. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets. IEEE Access 2020;8:181074-181090. [doi: 10.1109/access.2020.3027350]
63. Maynard D, Greenwood M. Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. 2014 Presented at: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); December 29, 2025; Reykjavik, Iceland p. 4238-4243 URL: https://eprints.whiterose.ac.uk/130763/
64. Matsumoto D, Seung Hee Yoo, Fontaine J. Mapping expressive differences around the world. J Cross-Cult Psychol 2008;39(1):55-74. [doi: 10.1177/0022022107311854]
65. Han X, Wang J. Using social media to mine and analyze public sentiment during a disaster: a case study of the 2018 Shouguang city flood in China. IJGI 2019;8(4):185. [doi: 10.3390/ijgi8040185]
66. Chen X, Yik M. The emotional anatomy of the Wuhan lockdown: sentiment analysis using Weibo data. JMIR Form Res 2022;6(11):e37698 [FREE Full text] [doi: 10.2196/37698] [Medline: 36166650]
67. Gu M, Guo H, Zhuang J. Social media behavior and emotional evolution during emergency events. Healthcare (Basel) 2021;9(9):1109 [FREE Full text] [doi: 10.3390/healthcare9091109] [Medline: 34574883]
68. Oz T, Havens R, Bisgin H. Assessment of blame and responsibility through social media in disaster recovery in the case of #flintwatercrisis. Front Commun 2018;3:45. [doi: 10.3389/fcomm.2018.00045]
69. Du Y, He H, Chu Z. Cross-cultural nuances in sarcasm comprehension: a comparative study of Chinese and American perspectives. Front Psychol 2024;15:1349002 [FREE Full text] [doi: 10.3389/fpsyg.2024.1349002] [Medline: 38445055]

Abbreviations

BERTopic: Bidirectional Encoder Representations from Transformers Topic
c-TF-IDF: class-based term frequency-inverse document frequency
CV: coefficient of variation
IG: Integrated Gradients
ISO: International Organization for Standardization
JARS: Journal Article Reporting Standards
XLM-RoBERTa: Cross-lingual Language Model–Robustly optimized BERT approach

Edited by S Brini; submitted 23.Sep.2025; peer-reviewed by P Babatuyi, A Rasool, BAF Assunção; comments to author 14.Oct.2025; accepted 24.Dec.2025; published 20.Jan.2026.

Please cite as:
Guo X, Fan Y, Guo Y
Public Emotional and Thematic Responses to Major Emergencies on Social Media, 2024-2025: Cross-Sectional Convergent Mixed Methods Study
J Med Internet Res 2026;28:e84648
URL: https://www.jmir.org/2026/1/e84648
doi:10.2196/84648
PMID:

©Xingrong Guo, Yiqian Fan, Yiming Guo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Cocreating Principles for Digital Health Equity: Cross-Sectional, Qualitative Study for Participatory Human-Centered Design in Catalonia

Jordi Piera-Jiménez[1,2], MBA, PhD; Núria Vilarasau Creus[3], MSc; Ada Maymó Costa[3], MA; Xabier Michelena[1,2,4], MD, PhD; Andrea Climent Fageda[2], MSc; Alèxia Farré[3], BSc; László Herczeg[3], MA; Lekshmy Parameswaran[3], MEng, MA; Gerard Carot-Sans[1,2], PhD; Luis Valle[2], MBA, PhD

[1]Digitalization for the Sustainability of the Healthcare System (DS3) research group, Barcelona, Spain

[2]Information Sysems Directorate, Catalan Health Service, Gran Via de les Corts Catalanes 587, Barcelona, Spain

[3]The Care Lab, Barcelona, Spain

[4]Rheumatology Research Group, Vall d'Hebron Research Institute, Barcelona, Spain

Corresponding Author:
Jordi Piera-Jiménez, MBA, PhD
Digitalization for the Sustainability of the Healthcare System (DS3) research group, Barcelona, Spain

Abstract

Background: Digital health technologies promise to democratize health care access yet often exacerbate existing inequalities when developed through traditional top-down approaches that prioritize technology implementation and exclude end users from design processes. The COVID-19 pandemic accelerated digital transformation while simultaneously exposing how technology can both bridge and widen gaps in health care access. Understanding how to systematically integrate equity considerations into digital health transformation across entire health systems has become increasingly urgent.

Objective: This study aims to cocreate actionable design principles for equitable digital health transformation through a large-scale participatory human-centered design (PHCD) process involving diverse stakeholders across Catalonia's health care ecosystem (northeast Spain), with the aim of establishing guidelines for information systems that support a person-centered, integrated, and longitudinal care delivery model.

Methods: We conducted a qualitative PHCD research study involving 265 participants representing diverse stakeholder groups: citizens and informal caregivers (n=106), health care professionals (n=83), health care managers and leaders (n=50), and experts representing various domains of digital health innovation (n=26). Through two sequential rounds of participation between June 2024 and April 2025, we used design thinking methodologies and cocreation tools in 24 sessions across Catalan geography and 7 topic-specific expert sessions. Data collection used innovative visual tools, including journey mapping, care model animations, future scenario storyboarding, and facilitated ideation techniques. Analysis followed an inductive-deductive approach combining affinity mapping, thematic synthesis, and participant validation to transform stakeholder proposals into actionable design principles.

Results: Participants identified critical barriers to digital health equity, including digital literacy gaps, fragmented information systems, a lack of user involvement in design, and insufficient consideration of vulnerable populations' needs. The cocreation process yielded 10 fundamental principles: (1) the person and their care circle at the center, (2) health for everyone, everywhere, (3) tools for more compassionate care, (4) a better professional experience, (5) an active role of the population, (6) personalized and proactive care, (7) systematic use of data for decision-making, (8) integrated quality data working for health, (9) an information system that builds trust, and (10) collaboration as a driver of innovation.

Conclusions: This study shows how large-scale, rigorously conducted PHCD can uncover and address equity barriers in health information systems. Beyond producing 10 actionable design principles, it highlights how engaging diverse stakeholders can turn complex inequities into practical guidance for equitable digital transformation. The resulting principles provide a framework for creating person-centered systems that are robust, inclusive, and accessible to all, while underscoring the need for enduring partnerships among public institutions, researchers, design experts, and communities as a foundation for sustainable and equitable digital health innovation.

(J Med Internet Res 2026;28:e84129) doi:10.2196/84129

KEYWORDS

digital health; equity; human-centred design; participatory design; digital health strategy; health information systems; co-creation; user experience

XSL•FO
RenderX

Introduction

The promise of digital health technologies to transform health care delivery has never been more apparent than in the wake of the COVID-19 pandemic. While the rapid adoption of telemedicine and mobile health apps demonstrated the potential of technology to maintain health care access during unprecedented disruption, it simultaneously revealed a troubling paradox: the very innovations designed to improve health care access often exacerbated existing inequities [1,2]. Vulnerable populations, including older adults, those with limited digital literacy, people with disabilities, and communities facing language barriers, frequently found themselves excluded from these digital advances, transforming potential bridges to care into barriers [3,4].

This paradox stems from a fundamental flaw in how health technologies are traditionally developed. Despite decades of evidence supporting user-centered approaches, health information systems continue to be designed primarily through top-down processes that prioritize technical requirements and organizational workflows over the needs, contexts, and capabilities of end users [5-7]. Health care professionals report spending up to 50% of their time on administrative tasks related to electronic health records (EHRs), reducing time available for direct patient care [8], while at the same time, patients struggle to navigate increasingly complex digital interfaces that assume levels of digital literacy many do not possess [9]. These challenges are particularly pronounced for vulnerable populations whose voices are rarely heard in technology design processes [10].

Participatory human-centered design (PHCD) offers a radically different approach to health technology development. Rooted in the principle that those who will use a system should be central to its design, PHCD engages users as partners throughout the development process, from initial problem definition through implementation and evaluation [11-14]. In health care contexts, this means involving patients, caregivers, health care professionals, and community members not merely as sources of requirements but as cocreators of solutions [15,16]. This orientation resonates with the broader global health literature describing human-centered design as a disciplined yet flexible practice that bridges design and implementation by iteratively engaging stakeholders and aligning innovation with human values and equity goals [17,18]. This approach is particularly crucial for addressing health equity, as it ensures that the voices of those most at risk of digital exclusion, often the same populations experiencing the greatest health disparities, are not just heard but actively shape the solutions developed [19,20]. Growing evidence demonstrates that health information systems developed through genuine participatory processes achieve higher adoption rates, better user satisfaction, and more equitable outcomes across diverse populations [21]. Yet, recent reviews highlight that while human-centered approaches can enhance equity for participants directly involved in cocreation, their broader, system-level impact often remains untested due to limited scaling and evaluation [22].

The concept of digital health equity extends far beyond mere access to technology. As defined by the World Health Organization (WHO), it encompasses the ability of all individuals to benefit from digital health solutions regardless of their socioeconomic status, geographic location, age, disability, or cultural background [23]. Operationalizing this vision requires addressing what recent collaborative frameworks identify as digital determinants of health (i.e., the complex, interconnected factors that shape whether digital health interventions produce equitable outcomes across populations) [24-27]. The adaptation of Richardson et al [26] of the National Institute of Minority Health and Health Disparities Research Framework [28] suggests that digital health equity operates across 4 different levels of influence (individual, interpersonal, community, and societal), each containing specific determinants that must be simultaneously addressed [26]. Similarly, recent work on equitable digital health design further proposes operational frameworks (eg, the Double Diamond and IDEAS models) that guide practitioners in deliberately embedding equity considerations throughout all design phases and in fostering structured collaboration with underserved groups [29].

Like many other countries worldwide, the COVID-19 crisis accelerated the digital transformation of health care in Catalonia (northeast Spain), with digital consultations increasing from 2% to more than 25% of all primary care visits between 2019 and 2021 [30,31]. However, this rapid digitalization revealed significant disparities across the region's 8 million inhabitants of the region. Rural areas struggled with inadequate internet infrastructure, older adult populations faced overwhelming digital interfaces, immigrant communities encountered language barriers in digital platforms, and health care professionals grappled with fragmented systems that hindered rather than helped care coordination [32]. These challenges acquired particular significance as Catalonia prepared to implement its ambitious Digital Health Strategy (2026 - 2031), a transformative initiative designed to support the future needs of health care delivery. Before embarking on this large-scale transformation, policymakers determined that equity principles must be embedded from the outset to avoid amplifying existing disparities.

Recognizing both the urgent need to address these digital equity challenges and the imperative to build equity into the upcoming Digital Health Strategy from its foundation, the Catalan Department of Health launched on a major participatory initiative to reimagine health information systems for the future. Rather than starting with predetermined technological solutions, the Department adopted a PHCD approach that would first identify gaps in the current health care delivery model and envision how future care should be organized, then determine how health information technologies could support this transformation. Following this approach, the idea was to reflect the growing recognition that creating truly equitable health information systems requires fundamentally rethinking not just what we build, but how we build it and who is involved in the building process [33].

This paper presents the results of the aforementioned large-scale PHCD process conducted across Catalonia, aimed at providing guidance to the design of a new and equitable health information

XSL•FO
RenderX

infrastructure that places the citizen at the center. The study addressed 2 main questions: (1) What are the lived experiences and challenges of diverse stakeholders navigating the Catalan health care system, particularly regarding barriers to equitable access and digital health engagement? (2) What fundamental principles and guidelines should inform the design of health information systems to ensure they are equitable, accessible, and responsive to the needs of all populations, especially those traditionally excluded from design processes?

Beyond these specific research questions, the study constitutes the first large-scale use case of user involvement in human-centered design for health care system transformation, providing valuable insights for replicating and scaling participatory design approaches across complex health systems.

Methods

Study Design and Conceptual Framework

We conducted a qualitative participatory design research study to develop a collective vision for Catalonia's future health care model and actionable design principles for digital health information systems. Our framework combined 2 complementary approaches. Human-centered design placed people at the core of the process, grounding design in their needs and contexts [34,35]. Participatory design ensured that diverse stakeholders were active cocreators rather than passive consultees [36]. This combined approach (ie, PHCD) was chosen because it directly addresses a fundamental cause of digital health inequity: traditional top-down development processes that systematically exclude the voices of those most affected by health disparities, particularly vulnerable populations who experience the greatest barriers to digital health access. PHCD engages stakeholders at all levels, from citizens and caregivers to frontline professionals, managers, and technical experts, capturing how barriers manifest and interact across individual, interpersonal, organizational, and systemic dimensions. This approach ensures a deep understanding of user needs before developing solutions, avoiding the premature solutioning that often leads to health technology failure [11,37].

The study design was structured around 2 sequential phases of participation, each addressing distinct but interconnected research questions. Both rounds used consistent participatory design methods, including journey mapping, system visualization, facilitated ideation, and collaborative prioritization, with Round 1 tools focusing on current care experiences and future care model visioning, and Round 2 tools adapted to translate identified needs into specific design principles for health information systems. Round 1 was deliberately designed to focus on health care delivery experiences and needs without introducing technology considerations, thereby avoiding the risk of biasing discussions toward predetermined technical solutions or constraining participant imagination to existing digital tools. This approach ensured that identified barriers and desired futures emerged authentically from lived experiences rather than being shaped by assumptions about technological feasibility or current system constraints. Only after establishing this grounded understanding of real citizen and professional needs did Round 2 shift focus to cocreating specific design principles for health information systems, ensuring that technology would serve human needs rather than drive the transformation agenda. The 2-phase approach enabled progressive deepening of understanding, moving from broad exploration of health care system experiences to focused cocreation of specific design principles for future health information systems. Finally, we deliberately adopted a multilevel scope encompassing technological design, organizational implementation, and broader system considerations. This approach was grounded on the fact that digital health equity failures occur across all these dimensions [26,38-40], and our PHCD methodology required understanding the full health care delivery context before narrowing to specific IT requirements.

The manuscript has been prepared according to the Standards for Reporting Qualitative Research (SRQR) guidelines [41].

Project Team Organization

The study was managed through a collaborative governance structure designed to ensure scientific rigor, meaningful stakeholder engagement, and operational excellence. A project steering board comprised representatives from the Catalan Health Service (JP-J, XM, LLV, and AC). The steering board was responsible for defining project objectives, monitoring progress against milestones, ensuring alignment with regional health transformation goals, making decisions about methodological adaptations based on emerging insights, and engaging with territorial health coordinators to recruit participants.

The operational implementation was carried out by a specialized design practice in multistakeholder engagement, human-centered design, and participatory design methodologies for enabling health and care system transformation. This team included supervisors (LP and LH) who provided methodological guidance and oversaw the overall design process to ensure consistency and quality. A dedicated tandem project management support structure coordinated logistics, including scheduling meetings across multiple territories, managing participant communications, providing technical support, and coordinating the overall design methodology and facilitation team (NVC). A team of 4 trained service-system designers who designed, facilitated, and synthesized the sessions, 2 lead facilitators who guided overall discussion flow and ensured all voices were heard (NVC and AMC), and 2 support facilitators who helped to document information during sessions, assisted participants with cocreation activities, and provided real-time analysis support (AF and AC). This team structure ensured comprehensive documentation while enabling smooth session flow and participant engagement.

The steering board and operational team established a coordinated governance structure with weekly alignment sessions to ensure methodological consistency, address implementation challenges, and maintain project objectives throughout the process.

Participants and Recruitment

Participants' recruitment and involvement adhered to the 4 core criteria outlined in the Catalan Department of Health's

XSL•FO
RenderX

Framework for Citizen Participation in Health, ensuring adequate methodological definition, maintaining scientific evidence and ethical standards, protecting vulnerable and minority populations, and preserving public health system sustainability [42].

Recruitment followed a purposive sampling strategy designed to maximize diversity and ensure representation of voices often excluded from technology design processes [43]. Sampling was operationalized through the health coordinators of each Catalan health region. These coordinators, who are managers within the Catalan Health Service, are responsible for planning and translating system-level health care policies into practice across the 10 health care regions of Catalonia. Once the study protocol was approved, and before fieldwork started, all 10 health coordinators attended a kick-off meeting in which the research team (ie, the authors) presented the study aims and procedures and provided detailed instructions for participant identification and recruitment. Comprehensive descriptions of the selection criteria and recruitment process are provided in the Multimedia Appendix 1. The health coordinators identified potential participants based on explicit diversity criteria, including gender and age distribution, organization types (public, private, and nonprofit sectors), health and social care service levels, and geographic representation across urban, semiurban, and rural territories. Health coordinators were explicitly asked to include, whenever possible and while meeting eligibility requirements, participants from potentially vulnerable groups such as older adults, immigrants, individuals with disabilities, and those with lower socioeconomic status. The project manager conducted weekly meetings with health coordinators for monitoring the recruitment process.

The study targeted four key stakeholder groups, each bringing essential perspectives to the cocreation process: (1) citizens and informal caregivers with diverse health care experiences (including expert patients and representatives of patient associations), who provide lived experience of navigating the health care system and represent the end user perspective; (2) health and social care professionals across different disciplines and care settings, who offer frontline insights into system functionality and daily operational challenges; (3) health and social care managers and leaders from various organizational levels, who contribute strategic and organizational perspectives on system implementation and sustainability; and (4) experts representing various domains of strategy, innovation, and technology, who provide technical expertise and broader sectoral knowledge to inform feasible and evidence-based solutions.

Territorial and Contextual Considerations

We recognized that territorial context significantly influences health care experiences and digital health needs [44,45]. Our sampling strategy deliberately included representation across urban, semiurban, and rural territories to ensure diverse contextual factors would be captured [3]. We also considered intersectional factors in our analysis framework, recognizing that age, gender, socioeconomic status, immigration status, and disability interact with geographic location to create unique barriers and opportunities [10]. We conducted an informed selection of the locations in consultation with the respective health regions to ensure comprehensive coverage of key regions and strategic areas across the Catalan territory.

Timeline and Methods

The study spanned from June 2024 to April 2025, beginning with preparation activities conducted from June to September 2024 (Figure 1). Data collection occurred through 2 sequential rounds of participation and analysis from October 2024 to March 2025, followed by consolidation of results and synthesis conducted from March to April 2025. Each round addressed distinct but interconnected research objectives, building progressively from understanding current health care experiences to cocreating specific design principles for future digital health information systems.

Figure 1. Project timeline, with activities (in blue) and outputs (in gray). R1: round 1; R2: round 2.

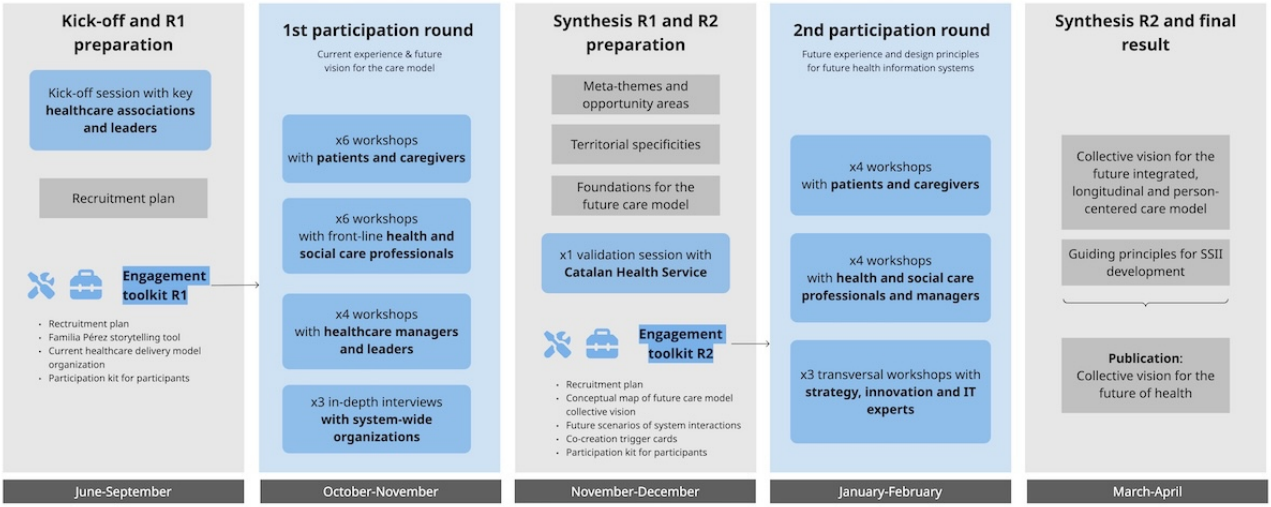

Round 1, conducted from October to November 2024, focused on collective reflection about the care model. This round aimed to understand current lived experiences with the health care system, identify what works well and what needs improvement, and envision desired futures. The central guiding questions included, “What do we understand by a person-centered, longitudinal, and integrated care model?” “What needs does this model address?” and “What must change in the current

XSL•FO
RenderX

system to achieve this vision?" Data were collected throughout 16 cocreation sessions conducted face-to-face and strategically selected to represent Catalonia's geographic and demographic diversity. Sessions were distributed across urban territories (Badalona, Barcelona, Girona, Lleida, and Tarragona), semiurban territories (Igualada, Manresa, Sant Cugat, Sant Feliu de Llobregat, Tortosa, and Vic), and rural territories (Amposta, Navàs, Palamós, Tremp, and Sort). Each 2-hour session included between 8 and 12 participants and followed a carefully structured format designed to maximize participation and insight generation. Additionally, 3 online one-hour group interviews with 2 participants each complemented the sessions across the territory, focusing on macrolevel perspectives and strategic insights into system-level challenges and opportunities. The organizations participating in these interviews were: the Agency for Quality and Assessment in Healthcare, the TICSalut and Social Foundation, and the Catalan Integrated Care Agency.

Round 2, conducted between January and February 2025, was devoted to cocreating specific design principles for digital health information systems. Building on insights from Round 1, this round transformed needs and barriers into actionable principles. The central question guiding this round was, "What design principles should guide the future development of health information systems to ensure they serve all citizens equitably?" This round started with a validation session with 7 experts from the Catalan Health Service, who reviewed Round 1 learnings and bridged into defining the focus for Round 2. It used 8 additional face-to-face sessions, maintaining geographic representation while focusing on the development of design principles. These sessions were distributed across urban territories (Barcelona), semiurban territories (Olot, Puigcerdà, Reus, Tàrrega, and Vilanova i la Geltrú), and rural territories (Mora d'Ebre, Solsona). Each session lasted 2 hours and included between 8 and 12 participants. Three topic-specific expert sessions brought together experts in specific domains identified from Round 1 learnings. Each session took 2 and a half hours and included 6 expert participants focusing on particular aspects of digital health transformation, such as strategy, innovation, and technology, providing a more macro perspective on the central question.

The locations of the 2 rounds across the Catalan geography are depicted in Figure 2.

XSL•FO
RenderX

Figure 2. Territorial distribution of the face-to-face participatory sessions across the 2 rounds.

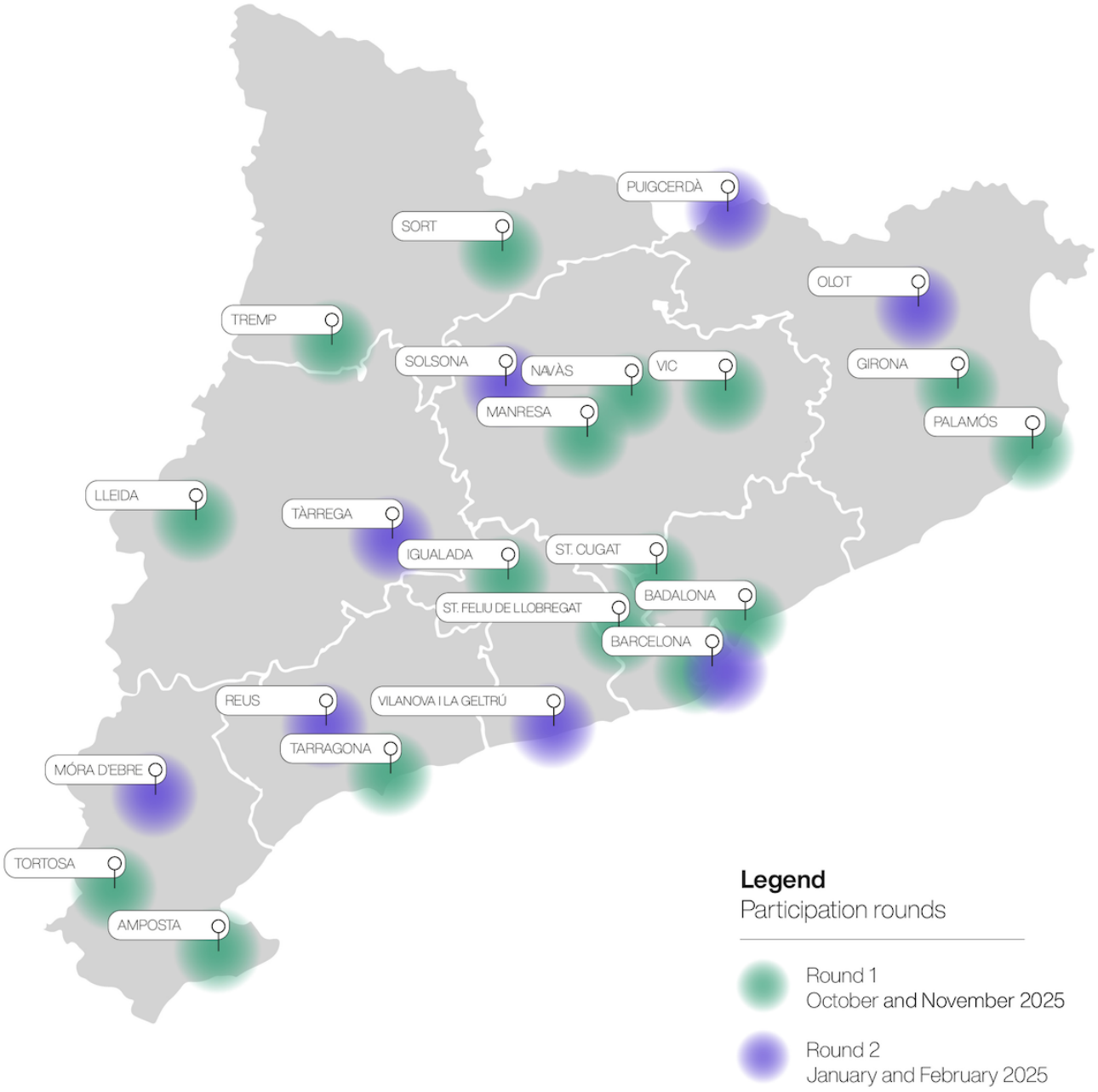

All sessions followed a consistent structure designed to create safe, inclusive spaces for participation. Sessions began with a 15-minute opening that established context, introduced facilitators, and established participation ground rules to ensure safe sharing of diverse perspectives. Core activities lasting 90 - 105 minutes varied by round and session type but consistently used visual facilitation tools, including journey mapping, experience sharing, ideation techniques, and collaborative prioritization exercises. Sessions concluded with 10 - 15 minutes for summarizing key insights, outlining next steps, and acknowledging participant contributions.

Sessions used innovative visual tools including journey mapping and storytelling to understand current health care experiences, scenario building for future visioning, and facilitated ideation techniques for developing design principles. Large-format posters, sticky notes, and collaborative boards enabled participants to contribute ideas both verbally and visually (reflecting individually and sharing collectively). Session logistics were carefully adapted to ensure inclusive participation across diverse contexts. Rural sessions addressed infrastructure challenges through flexible scheduling and accessible locations, while urban sessions accommodated participants' complex schedules while managing larger group dynamics. All venues provided step-free access, appropriate facilities for participants with disabilities, and professional translation services when needed. Transport support was provided where needed, particularly for participants from vulnerable populations or remote areas.

At the end of each session, a satisfaction questionnaire was provided to each participant to assess their participation experience and the perceived value provided by the workshop. The underpinning idea behind this feedback collection was to improve the forthcoming workshop dynamics in case any issue was identified (Multimedia Appendix 1).

XSL•FO
RenderX

Cocreation Tools

Central to the methodology was the development and use of innovative cocreation tools designed to make participation accessible and engaging for all stakeholders. A total of 4 tools were developed, 2 storytelling visual narratives were designed for the sessions with citizens, caregivers, and professionals. And 2 system maps for the sessions with health care managers and experts. The cocreation materials design drew from established journey mapping [46-48] and system visualization methodologies [49,50] in health care service design. The research team's approach was informed by their prior development of the WhocCares publication and the WeCare Toolkit (National Council of Social Service, Singapore and fuelfor) for caregiver support in Singapore [51], which demonstrated effective use of visual tools for engaging diverse stakeholders in the cocreation of health and social care service innovations. The full set of tools can be found in the Multimedia Appendix 1.

The "Pérez Family" journey mapping and storytelling tool consisted of a visual narrative depicting 8 health care interaction scenarios across the life course, from prenatal care through end-of-life support (Figure 3). This tool was designed with health care experts from the Catalan Health Service to ensure scenarios were realistic, relatable, and comprehensive, translating complex processes and multiple care pathways into accessible storylines to trigger exploration and discussion. The full poster of the "Pérez Family" tool is provided as supplementary material in Multimedia Appendix 1.

Figure 3. Translated sample of the "Pérez Family" journey mapping and storytelling tool used in face-to-face sessions with patients, caregivers, and professionals.



The system maps for professional and manager sessions included behind-the-scenes system processes and pain points (Figure 4).

XSL•FO
RenderX

Figure 4. Current health care delivery model organization used in face-to-face sessions with health care leaders and managers. IS3 (*Integrador de serveis de salut*): care process management; HC3 (*Història clínica compartida de catalunya*): shared care record*; eCAP:* primary care electronic health record; HIS: hospital information system; HS (*Historial social*): social care record; SITREM (*Sistema de tractament d'Emergències*): emergency treatment;

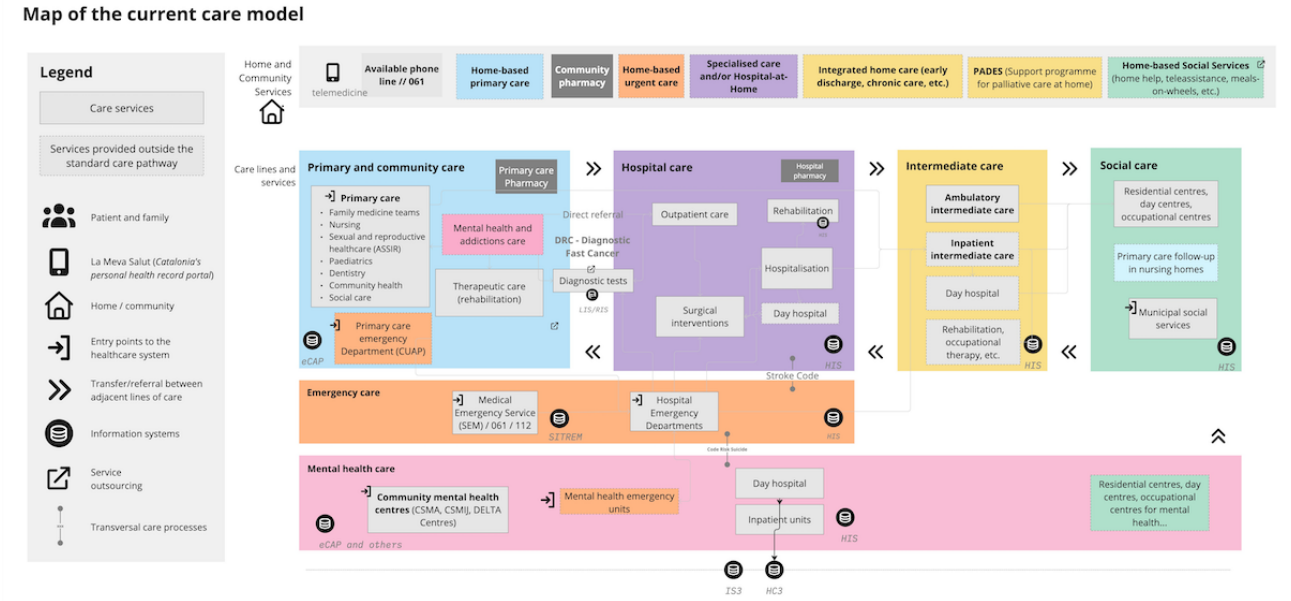

For Round 2, future scenarios were designed to portray a visual narrative of future interactions within 6 key themes emergent from Round 1 analysis: proximity care, empathic and quality interactions, collaborative and continuity care, personalized and proactive health prevention and promotion, active citizen participation in health, and holistic view of health. A conceptual map of the future care model vision resulting from Round 1 was another key visual used in the expert sessions.

Visual facilitation materials were carefully designed to support different communication styles and literacy levels. A facilitation toolkit was designed to enable issues to be seen at different levels of zoom or resolution, from lived experience to system overview, through technical backend, and overall delivery model. Large-format posters enabled collective visioning, while color-coded notes provided a simple yet effective method for categorizing inputs (Figure 5). Participants were given a participation kit with materials printed individually for better readability (Multimedia Appendix 1). All materials were tested for accessibility and refined based on feedback.

XSL•FO
RenderX

Figure 5. Example of a large-format poster for collective visioning with participants adding color-coded notes for item categorization.



Data Collection

Data collection was designed to capture both the content and context of cocreation activities through multiple complementary approaches. The facilitation team used a structured approach to data capture, with each member having specific responsibilities.

The lead facilitator maintained a neutral stance while using follow-up questions to uncover the deeper reasoning and meaning behind participants' statements. The support facilitator observed and documented group dynamics, body language, nonverbal cues, power relations, and patterns of interaction. Dedicated notetakers captured verbatim statements, reasoning provided by participants, and contextual explanations. One notetaker focused exclusively on capturing verbatim quotes to preserve original language. Observer roles photographed and video-recorded key moments, interactions (such as voting, debating, and consensus-building), and visual outputs to document the atmosphere and participation patterns.

Multiple types of data were systematically collected from each session (1) physical artifacts including color-coded sticky notes, large-format posters, and flipcharts with participant handwriting and drawings; (2) documented narratives of participants' lived experiences and system interactions; (3) structured observational notes on group dynamics, including dominant voices, facilitation influence, and the sequence in which ideas emerged and built upon each other; (4) photographic and video documentation; and (5) verbatim capture of participants' oral explanations and reasoning. Table 1 summarizes the data collection methods, mapped to objectives and methods, according to project phases.

XSL•FO
RenderX

Table . Overview of methodological approaches and data collection tools by study phase.

Study phase	Timeline	Research objectives	Methods	Data collection tooling
Kick-off and Round 1 (R1) preparation	June-September 2024	• Develop cocreation tools. • Establish governance structure. • Recruit participants. • Design session protocols.	• Co-design of visual tools with health experts. • Tool testing and refinement.	• Recruitment tracking forms. • Tool design documentation. • Pilot session feedback.
Round 1: current experience and future vision for the care model	October-November 2024	• Understand current health care experiences. • Identify barriers and facilitators. • Envision desired future care model.	• Journey mapping. • System visualization. • Experience sharing. • Future visioning. • Collaborative discussion.	• "Pérez Family" journey tool. • Current care model system maps. • Color-coded sticky notes. • Large-format posters. • Verbatim capture. • Observational notes. • Photography and video. • Postsession debriefs. • Satisfaction questionnaires.
Synthesis R1 and Round 2 (R2) preparation	November-December 2024	• Validate Round 1 findings. • Define focus areas for Round 2.	• Expert validation session. • Collaborative synthesis.	• Summary presentations. • Validation discussion notes.
Round 2: future experience and design principles for future health information systems	January-February 2025	• Transform needs into actionable principles. • Define guidelines for health information systems. • Validate and refine principles.	• Scenario building. • Facilitated ideation. • Collaborative prioritization. • Principle refinement.	• Future scenario storyboards. • Future care model conceptual maps. • Color-coded sticky notes. • Large-format posters. • Verbatim capture. • Observational notes. • Photography and video. • Postsession debriefs. • Satisfaction questionnaires.
Synthesis R2 and final result	March-April 2025	• Synthesize findings across rounds. • Finalize design principles. • Prepare dissemination materials.	• Design synthesis. • Thematic analysis. • Principle validation.	• Digital collaboration boards (Miro). • Synthesis workshop outputs. • Final principle documentation. • Internal synthesis sessions.

Immediately following each session, the facilitation team conducted structured debriefs that captured observations on group dynamics (including identification of dominant participants, power relations, facilitation effects, and the order in which contributions emerged), surprising insights that emerged during sessions, patterns of agreement or tension, and participants' clarifications of their contributions. These immediate debriefs served to contextualize the visual outputs while observations were still fresh, ensuring that team observations about the process of cocreation informed the interpretation of the products (visual results). The debriefs used a standardized template covering key themes that emerged, surprising insights or unexpected perspectives, group dynamics observations (including any imbalances in participation or influence), emerging patterns across different voices, and new questions arising for subsequent sessions.

Data collection was extended at least until saturation, assessed through continuous monitoring of topics. As sessions progressed, we maintained a systematic list of topics and themes being discussed across workshops. The iterative download and synthesis process allowed clustering ideas and comments around

XSL•FO
RenderX

different themes for subsequent identification of both recurrent and unexplored topics. Saturation was reached when no new ideas or experiences were being captured.

Data Analysis

We used a systematic multistage analysis process combining inductive and deductive approaches, consistent with best practices in qualitative research [52]. Analysis began immediately after each session with the structured team debriefs (steering board and specialized design practice). These direct downloads enabled key takeaways to be captured whilst fresh and vivid, allowing the methodology to be tuned between sessions to optimize the approach and outcomes and boost learnings and insights. Our front-stage and backstage teams processed detailed data in agile loops as fieldwork progressed, which avoided a build-up of data and ensured learnings were taken quickly to deepen insights.

All physical materials, including colored notes, posters, and flipchart outputs, were photographed and systematically transcribed into digital collaboration boards using Miro software (Figure 6). This digitization process maintained original participant language and verbatim, preserved visual relationships between insights and ideas, and included contextual notes about discussion dynamics and nonverbal communications.

Figure 6. Example of digital boards (Miro software) used to analyze workshop outputs through axial coding. Contributions are mentioned in the Catalan language.

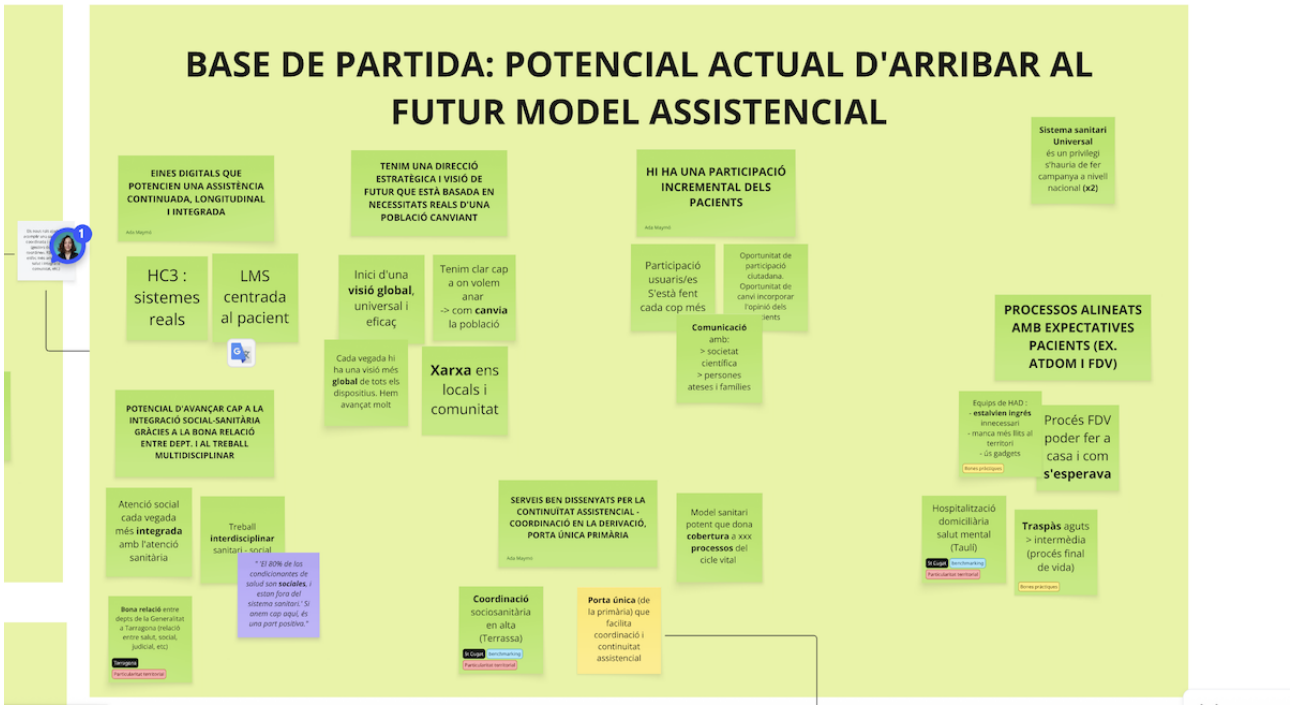

Multiple units of analysis were used. Visual outputs (colored notes and flipchart content) captured discrete participant insights and ideas. One researcher specifically documented verbatim quotes to preserve participant language. Debrief notes synthesized contextual observations about group dynamics and emergent patterns. These complementary sources enabled triangulation across data types.

Each data unit was tagged with structured metadata: geographic territory (urban, semiurban, and rural), stakeholder type (citizen, professional, manager, and expert), and workshop identifier. A systematic topic list tracked emerging themes across sessions to monitor saturation.

Analysis combined affinity mapping [53] with thematic synthesis [52]. Multiple analytical methods were integrated to ensure rigorous interpretation [54]. Initial affinity mapping clustered related insights on digital boards, grouping similar participant contributions whilst preserving their source metadata. This visual organization revealed natural patterns and relationships within the data [55]. Three researchers with complementary analytical perspectives—lived experience (citizen perspectives), operational processes (professional practice), and system architecture (organizational design)—then conducted thematic analysis. Open coding identified discrete concepts, followed by axial coding to explore relationships and create preliminary categories. Iterative team analysis sessions consolidated categories into broader themes, identifying priority challenges and opportunities.

Analytical consistency was ensured through regular synthesis sessions amongst the coding team and validation discussions with project steering board members (health information technology and clinical experts). These sessions cross-checked interpretations, identified analytical blind spots, and confirmed no significant gaps in knowledge capture.

From the Round 1 themes, we extracted foundational elements through design synthesis methods [56]. These foundations were translated into draft design principles through internal team workshops. Round 2 sessions then validated, refined, and finalized these principles through participant feedback.

XSL•FO
RenderX

Ethical Considerations

All participants provided written informed consent, and the study received approval from the Ethics Committee of Hospital Universitari de Bellvitge (reference PR123/24). Participants were informed that their participation was voluntary and could be withdrawn at any time without consequence. Data confidentiality was maintained through anonymization procedures, and all data were stored securely with access limited to research team members. Participants provided informed consent for the recording of verbatim statements and photographs during the sessions; no images were captured for participants who rejected consent. To ensure inclusive participation, transport and professional translation services were offered when necessary. In line with the PHCD approach, we organized give-back online sessions to share the findings with participants and inform them of subsequent steps, ensuring the research delivered value to all involved.

Results

Overview

A total of 265 participants across the Catalan health care ecosystem participated in a 2-round PHCD process that explored current health care delivery challenges and cocreated principles for equitable digital health transformation. The findings reveal both systemic barriers and collaborative solutions emerging from systematic engagement with diverse stakeholders across Catalonia's health care territories.

Participant Characteristics and Territorial Distribution

Participant demographics and territorial distribution are presented in Table 2, reflecting robust representation across stakeholder groups and geographic contexts throughout the study period. The sample included citizens and informal caregivers (n=106), health care professionals (n=83), health care managers and leaders (n=50), and experts in various domains of digital health innovation (n=26). Territorial representation spanned urban centers, rural areas, and intermediate zones, ensuring diverse infrastructure and resource contexts informed the findings.

XSL•FO
RenderX

Table . Demographic characteristics of participants.

Participant characteristics	Values, n (%)
Patients and caregivers (n=106)	
Gender	
Men	39 (36.8)
Women	67 (63.2)
Age (years; missing=9; n=97)	
18 - 35	2 (2.1)
36 - 50	16 (16.5)
51 - 65	38 (39.2)
>65	41 (42.3)
Profile (n=106)	
Patient	68 (64.2)
Caregiver	38 (35.8)
Territorial context	
Urban	20 (18.9)
Semiurban	49 (46.2)
Rural	37 (34.9)
Health and social professionals, experts, and leaders (n=159)	
Gender	
Male	57 (35.8)
Female	102 (64.2)
Age (years; n=159)	
18 - 35	7 (4.4)
36 - 50	100 (62.9)
51 - 65	52 (32.7)
>65	0 (0.0)
Profile (n=159)	
Family physician	4 (2.5)
Hospital and intermediate care physician	16 (10.1)
Primary care nurse	11 (6.9)
Hospital and intermediate care nurse	9 (5.7)
Psychology	4 (2.5)
Social worker	7 (4.4)
Health administration	7 (4.4)
New roles (community well-being, physiotherapist, and nutritionist)	7 (4.4)
Others (emergency service, pharmacy and care homes)	13 (8.2)
Health care managers and leaders	50 (31.4)
Experts in digital health innovation	26 (16.4)
Territorial context	
Urban	91 (57.2)
Semiurban	48 (30.2)
Rural	20 (12.6)

XSL•FO
RenderX

Through journey mapping and storytelling activities, Round 1 participants articulated 13 priority barriers to digital health equity that span technological, organizational, and sociocultural dimensions. These barriers emerged consistently across urban, semiurban, and rural contexts, though with varying emphases reflecting territorial specificities (see Figure 7 for territorial variations).

Figure 7. Summary of barriers to digital health equity identified during Round 1, contextualized per geographic area.

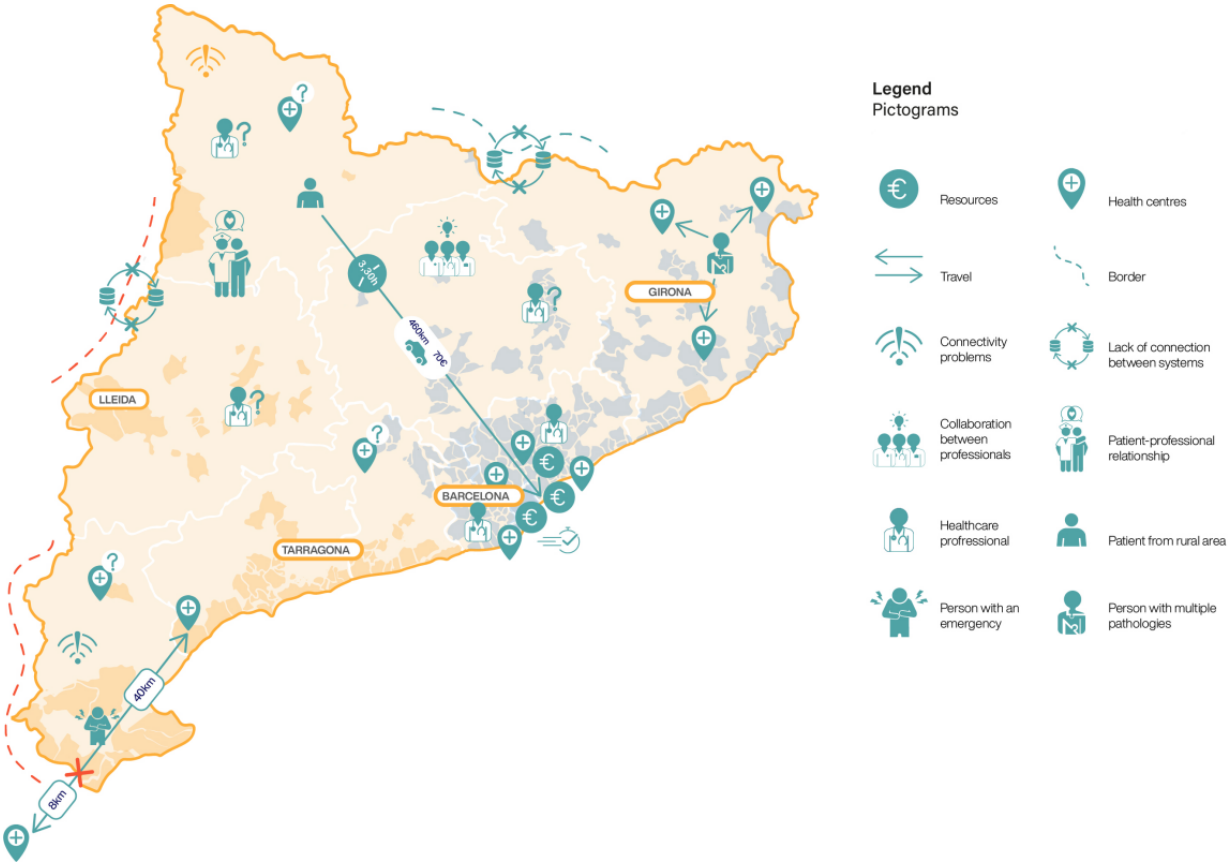

While these barriers emerged across all territories, their manifestation and intensity varied by geographic context. Rural areas emphasized connectivity and resource access challenges; urban contexts highlighted diversity and complexity issues; border regions contributed unique perspectives on administrative complexity and cross-jurisdictional care. Table 3 presents the 13 priority barriers with illustrative participant quotes demonstrating lived experiences across these diverse contexts. Detailed descriptions of all 13 barriers with extended context and additional participant quotes are available in Multimedia Appendix 2.

XSL•FO
RenderX

Table . Barriers to digital health equity identified in round 1, with illustrative participant quotes.

Barrier[a]	Description	Participant quote[b]
Fragmented health and social care integration	Services function in a fragmented manner due to hyperspecialization, generating multiple referrals that break care continuity. Primary care, under pressure, cannot assume its connecting role.	"Everything to do with the social system is a black box. I don't know what happens there. And we should know! 80% of health determinants are social." [Health care manager]
Insufficient emotional well-being support	Growing need for emotional support with a preventive, holistic, and community approach. Current system designed primarily for emergencies and physical pathologies.	"Community networks and neighbors are great prescribers of health and could be protective and informative elements." [Health care manager]
New citizen expectations are creating system pressure	Citizens have growing expectations about immediacy, flexibility, and personalization influenced by digital society, generating pressure on a system lacking resources to respond.	"We need to move towards a one-stop-shop system. It will make life easier for both the user and us." [Health care manager]
Lack of health education and prevention	Many people lack knowledge about self-care, prevention, and appropriate use of health services, leading to system misuse (eg, A&E saturation for nonurgent cases).	"Health literacy is fundamental. If people don't understand their health, they can't manage it properly." [Health care professional]
Social, language, and digital barriers	Digital gaps, medical terminology, and language barriers prevent many from accessing and navigating the system, especially older adults and migrant populations. Overall, 48% of the foreign population is at risk of poverty or social exclusion.	"Many times I don't understand the doctors' calls. If they give me a biopsy result I'm none the wiser, or I have to say, please, tell me in words I can understand. I prefer to go to the place rather than receive a call." [Service user]
Insufficient resources for aging and complexity	Progressive aging brings increased care needs for people with frailty and complexity. The current system lacks sufficient human and material resources for continuous, specialized, often home-based care.	"Resources should be moved towards the patient, rather than moving the patient towards resources. For example, prioritizing home care over residential facilities." [Health care manager]
Waiting lists with human consequences	Long waiting lists carry health risks and a strong emotional impact. Overall, 64% of patients wait >5 days to access a family physician; 150 days average for surgical intervention. Many are forced to turn to private health care, generating access inequalities.	"I'm an electrician, I'm self-employed, and they were giving me a physio appointment in 8 months. I can't wait, what will I eat? If you don't have private healthcare, you're done for." [Service user]
Resource planning based on quantitative criteria	Planning and resource allocation based on quantitative criteria (number of appointments, procedures) rather than health outcome indicators, leading to inadequate incentives.	"As a professional, you make the diagnosis, but when it comes to implementing truly, it's not possible because a specialist is missing. Something behind there should be a map of specialists that lets us zoom in and see, for example, the difference between supply and demand in areas like allergies and dermatology." [Health care professional]
Primary care unable to assume connecting role	Primary care should coordinate and ensure continuity, but pressure, lack of resources, and system fragmentation prevent it from exercising this role, generating breaks in care continuity.	"Primary Care must be the real entry and staying point for care. It shouldn't be a referral center. We need to do good work and strengthen Primary Care more." [Health care professional]
Inadequate protocols for chronicity and mental health	System excels at emergencies but lacks reinforcement in continuous preventive care and adequate support for daily management of chronic diseases. Overall, 37.8% of the adult population has chronic illness; 9.84% increase in chronic mental health patients since 2017.	"Mental health is left to God, more resources should be allocated to it. [...] Pure mental pathology is 'the great forgotten.' It works well in critical moments and acute hospitalization, but it doesn't have enough resources for chronic, continuous care within the community." [Service user and Health care professional]
High professional turnover	High turnover due to work culture changes and stressful conditions significantly affects care quality and continuity. Rotation prevents creating trust bonds essential for effective care, particularly impacting older patients and those with mental health needs.	"My psychologist has changed every other time. Each time you go you have to expose yourself again and explain what's happening to you." — Service user"I've even thought about leaving medicine and changing jobs. I work loads of hours, and you have no help from anywhere. I'm taking diazepam for anxiety." [Health care professional]

XSL•FO
RenderX

Barrier[a]	Description	Participant quote[b]
Administrative burden compromising care	Health care professionals dedicate significant time to administrative and bureaucratic tasks, reducing time for direct patient care, compromising care quality, and contributing to staff burnout.	”There should be an automatic process that allows you to spend more time with the patient. Here some technology like AI or a different way of entering information could play an important role.” [Health care manager]
Multiple noninteroperable information systems	Multiple noninteroperable systems hinder clinical information access and sharing between professionals from different care settings, generating inefficiencies, duplications, risk of errors, and coordination difficulties.	”Information should be accessible to all professionals because the information belongs to the patient, not to the professionals.” [Health care professional]

[a]Barriers synthesized from 189 participants across 19 round 1 sessions (16 territorial cocreation sessions +3 expert interviews).

[b]Quotes translated from Catalan/Spanish.

Cocreated Design Principles for Equitable Health Information Systems

Round 2 involved 8 territorial cocreation sessions, one validation session with Catalan Health Service experts, and 3 topic-specific sessions. Building on barriers identified in Round 1, participants cocreated 10 fundamental principles for equitable health information systems that address technological, organizational, and social dimensions of digital health equity (Figure 8).

Figure 8. Cocreated fundamental principles for equitable digital health information systems.

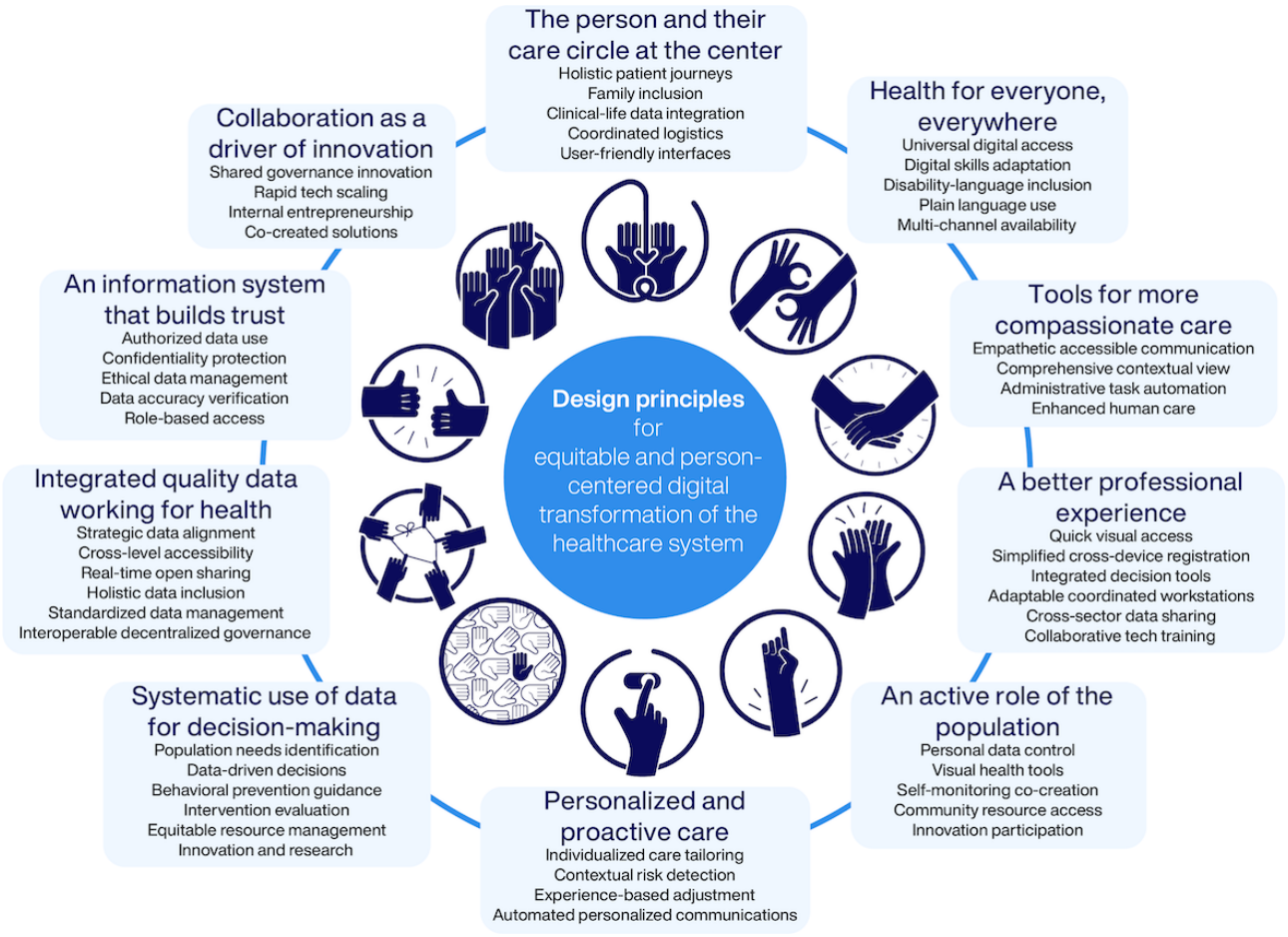

Territorial Context and Principle Development

Territorial analysis revealed how geographic and infrastructure contexts shaped both barrier experiences and the development of information system design principles. Rural areas emphasized connectivity and resource access issues, leading to a stronger emphasis on principles addressing “Health for Everyone, Everywhere” and “Collaboration as a Driver of Innovation.” Urban contexts highlighted diversity and complexity challenges, informing principles focused on cultural responsiveness and care coordination. Border regions contributed unique perspectives on administrative complexity and cross-jurisdictional care access, strengthening principles addressing system integration and user navigation support. These territorial differences informed the development of principles that explicitly address contextual variation rather than assuming standardized solutions.

XSL•FO
RenderX

Stakeholder Perspective Alignment and Divergence

While remarkable alignment emerged across stakeholder groups on fundamental principles, meaningful divergences in emphasis provided valuable insights for implementation. Health care professionals prioritized workflow integration and administrative burden reduction more strongly than other groups, directly informing the principle of "A Better Professional Experience." Patients and caregivers emphasized ease of use and personal control over health information, strengthening principles focused on user empowerment and data governance. Health care managers and leaders focused on sustainability and implementation feasibility, contributing to principles addressing systematic data use and collaborative innovation. Experts in digital health innovation highlighted technical interoperability and innovation potential, enriching principles related to data integration and system connectivity.

Although rare, certain divergences revealed conflicting views on the same topic. For instance, regarding mental health data sharing, parents expressed concerns about schools accessing their children's information, while professionals emphasized the need for cross-team coordination to optimize care. Such tensions were addressed by exploring underlying reasoning during facilitation and using divergences generatively during synthesis. Rather than forcing consensus, conflicting perspectives directly informed principles that could respond equitably to multiple stakeholder needs, such as Principle 9 on data governance, which emphasizes individual control over context-specific information sharing.

Implementation Considerations

Participants not only cocreated principles but also provided detailed insights into implementation considerations. They emphasized the need for phased implementation approaches that build confidence and capability gradually, strong change management and training programs for both citizens and health and social care professionals, and maintaining human connection even as systems become more automated. The cocreation process itself was identified as a valuable model for ongoing stakeholder engagement throughout implementation, with participants requesting continued involvement in system development and evaluation.

Participation Experience and Engagement

Participants demonstrated high levels of engagement throughout the cocreation process. Postsession evaluation surveys indicated strong satisfaction with the participatory experience, with participants reporting that they felt heard, respected, and valued in the process. The visual facilitation tools, particularly the "Pérez Family" journey mapping and storytelling, proved effective in enabling participants with varying communication styles and literacy levels to contribute meaningfully to discussions while providing a common visual framework upon which diverse experiences, ideas, and perspectives could be consolidated to create a shared mental model. The diversity of profiles facilitated rich exchanges between different perspectives, with participants frequently noting the value of hearing experiences from other stakeholder groups.

Discussion

Principal Findings

In this study, PHCD allowed us to generate a deep understanding of lived experiences and challenges faced by diverse stakeholders navigating the health care ecosystem in Catalonia, while simultaneously cocreating principles for guiding equitable digital transformation. Through engaging citizens, caregivers, professionals, managers, and different expert profiles in collective reflection and design, the process not only uncovered structural and experiential barriers but also translated these insights into actionable principles for future health information systems. Finally, the implemented and documented large-scale, iterative co-design approach provides a replicable methodology that other health systems can adapt to embed equitable care supported by health information technology as a central value in their transformation processes.

Our participatory approach yielded ten design principles that address critical gaps in current digital health development. These principles emerge from a comprehensive understanding of real-world challenges faced by patients, caregivers, health care professionals, and system leaders across diverse territorial contexts. The cocreation process revealed that digital health equity requires attention to multiple interconnected dimensions: technological accessibility, organizational readiness, professional workflow integration, and citizen empowerment in data governance.

The barriers identified (eg, digital literacy gaps, fragmented information systems, and insufficient consideration of vulnerable populations) align with the Digital Health Care Equity Framework in the United States [27], which similarly emphasizes addressing digital determinants of health through participatory design and community engagement. Our findings extend this work by showcasing how large-scale participatory processes can systematically translate these equity considerations into actionable design principles for driving digital transformations in the health care system.

The 10 principles address long-standing challenges in health IT development, particularly the gap between technological capabilities and real-world usability. Recent research continues to document usability problems, with over one-third of medication safety events in pediatric hospitals related to EHR usability issues[57].

The principle of "The Person and their Circle at the Center" reflects growing recognition that meaningful stakeholder involvement throughout design is essential [58]. This aligns with evidence from an umbrella review identifying participatory design and community collaboration as crucial for creating digital health tools serving diverse user needs [59]. Our study extends this evidence by scaling the community collaboration across an entire regional health system and applying it to define core design principles.

The emphasis on data control and citizen agency in the principle "An Information System that Builds Trust" resonates with recent research on patient preferences for health information exchange design, revealing complex attitudes toward health information

XSL•FO
RenderX

control and supporting our finding that genuine patient engagement is essential for addressing privacy and trust concerns [60].

Knowledge Contributions Beyond Design Artifacts

While producing 10 specific design principles, our research contributes to understanding digital health equity beyond these outputs through three key areas.

First, our findings illustrate how digital health equity barriers operate across multiple interconnected levels in practice. Previous frameworks have conceptualized equity barriers at individual, interpersonal, community, and societal levels [26,28]; implementation guidance has often remained siloed [22,29]. Our participatory process documented how stakeholders experience these barriers as deeply interconnected: technically accessible interfaces are undermined when organizational training support is missing; well-designed systems fail when procurement policies exclude user input; individual digital literacy interventions have limited impact when infrastructure gaps persist. In this regard, our findings suggest that equity-centered transformation may require coordinated technical, organizational, and systemic change because barriers appear to operate synergistically across these dimensions.

Second, our study documents how participatory design processes can serve as more than requirements-gathering exercises. Traditional health IT development treats stakeholder engagement as a requirements-gathering phase that feeds into expert-driven design [5,6]. Our experience indicates that participatory processes may generate equity-relevant outcomes through several mechanisms: surfacing barriers invisible to system designers [10], legitimizing experiential knowledge alongside technical expertise [18], and building stakeholder investment in transformation [59].

Finally, our research contributes to growing evidence highlighting the challenge of retrofitting equity considerations, which instead need to be embedded from the outset. The barriers identified appeared to stem not from isolated technical shortcomings but from design processes that systematically excluded diverse users [61]. Therefore, increasing equity may require restructuring design processes themselves, shifting from technology-driven implementation, where solutions are predetermined, to genuinely participatory cocreation, where diverse voices shape problem definition, solution ideation, and implementation priorities [14,15,21]. This concept aligns with that of other authors claiming that design processes themselves can either reproduce or mitigate existing health inequities [1,2,10].

Limitations and Transferability

While our study provides valuable insights into participatory design for digital health equity, the external validity of our principles and conclusions may be influenced by intrinsic features of the study design.

The geographic focus on a single area under the same health care system (ie, Catalonia) may have led to some insights not applicable to other countries and thus limiting transferability. However, we consider most of our findings applicable and valuable beyond regional boundaries through several dimensions. First, many of the identified barriers represent universal systemic challenges (workforce shortages, resource constraints, service fragmentation, and navigation complexity) documented across diverse health care contexts. Second, the methodological framework itself is transferable due to the inclusion of territorial segmentation (rural vs urban contexts) and multilevel stakeholder engagement (citizens to policymakers and health care to social care), which are categories relevant to most health systems. Third, specific findings regarding data duplication between public and private providers are particularly relevant and common in mixed-model health care systems, which are increasingly prevalent globally. Finally, the governance of the Catalan health system is regionally commissioned, like other environments, such as England's Integrated Care Boards, enhancing organizational-level transferability. Future research should examine how these principles perform across different health care contexts, particularly in low-resource settings where digital equity challenges may be more pronounced.

Our study captured stakeholder perspectives at a specific point during rapid digital transformation. The long-term sustainability and relevance of identified principles will require ongoing evaluation as digital health technologies and user needs evolve. However, we attempted to ensure future-proofing through grounding principles in fundamental human-centered design values, focusing on equity considerations that persist regardless of specific technologies, and validating principles through diverse stakeholder perspectives representing enduring health care needs.

Another feature that may influence generalizability is the participant profile involved in the sessions. Despite our effort in recruiting a diverse profile regarding roles, geographic location, and demographic characteristics, the voluntary nature of participation may have introduced selection bias towards individuals with a stronger interest in digital health transformation. This was particularly relevant for migrant and underserved communities, which were underrepresented. These populations often face the most significant barriers to digital health access. The exclusion of these voices represents a significant limitation, as these communities are frequently most affected by digital health inequities.

Future research could explore mechanisms for engaging harder-to-reach populations, with particular attention to developing culturally appropriate recruitment strategies for migrant communities and other underserved populations.

Implications for Practice and Policy

Our findings have important implications for health care organizations, technology vendors, and policymakers. For health care organizations, our study highlights the value and feasibility of investing in meaningful participatory processes before, during, and after health information system implementations. The principles provide a framework for evaluating existing systems and guiding future technology decisions with explicit attention to equity considerations.

XSL•FO
RenderX

A critical insight is the importance of innovative partnership approaches when navigating system complexity to drive large-scale engagement and co-design processes. In our experience, successful system-level transformation requires sophisticated collaboration models bringing together diverse expertise and perspectives. The partnership between public health authorities, academic institutions, and specialized design practitioners created a unique configuration enabling navigation of complex health care system dynamics whilst maintaining methodological rigor and stakeholder trust. This proved essential for accessing diverse participant networks, managing territorial complexities, and ensuring that cocreated principles would be both evidence-based and implementable within existing structures.

For technology vendors, our research highlights the importance of embedding human-centered design processes throughout product development, not merely as compliance exercises. Persistent usability problems documented in EHR research suggest current vendor approaches to stakeholder engagement are insufficient [52]. Our study provides evidence that more comprehensive participatory approaches can yield insights leading to more effective, equitable, and sustainable system designs.

For policymakers, our findings support the need for regulatory frameworks that incentivize or require meaningful stakeholder engagement in health technology development. Our study demonstrates the feasibility and value of comprehensive participatory approaches that could inform future policy development. Additionally, policymakers should consider how to support and incentivize the development of collaborative partnerships and co-design capabilities within health care systems, recognizing that these represent critical infrastructure for equitable digital transformation.

Future Directions

The 10 principles provide a foundation for future research examining their implementation and effectiveness in practice. Critical areas include extending these principles to social and community care settings, examining how they apply to artificial intelligence–enabled health technologies [62], and exploring how they can inform the development of community health information systems. Implementation research examining how organizations can operationalize these principles, including necessary capabilities, resource requirements, and implementation strategies, would provide valuable guidance for translating insights into improved health information systems.

Conclusions

This study showcases how PHCD, conducted at a large scale and grounded in rigorous methodology while engaging all relevant stakeholders, can serve as a powerful tool to identify equity barriers in health information systems and inform relevant and specific principles driving digital transformation of health care systems. The 10 principles emerging from this study provide a comprehensive framework for developing person-centered health information systems that are not only technically robust but also genuinely responsive to user needs, culturally appropriate, and accessible to all populations.

This study makes three distinctive contributions that differentiate it from existing research. First, it is the first successful application of PHCD at a regional health system scale, substantially exceeding the typical scale of participatory design studies in health care. Second, it provides empirical evidence of how equity barriers operate synergistically across technical, organizational, and systemic levels in practice, challenging the siloed implementation approaches suggested by previous conceptual frameworks. Third, it exemplifies how participatory processes can serve as transformative mechanisms for embedding equity from the outset, rather than the requirements-gathering exercises that characterize traditional health IT development. These contributions provide both a methodological blueprint for scaling participatory approaches and evidence that restructuring design processes themselves is essential for achieving digital health equity.

Acknowledgments

We wish to acknowledge the valuable contributions of those individuals who participated in the study, including patients, caregivers, health care professionals, and managers who contributed to the cocreation sessions. We are grateful to the territorial health coordinators across Catalonia who facilitated participant recruitment and supported session logistics in their respective regions. We acknowledge the invaluable contributions of The Care Lab team members who made this research possible: Andreu Coy, Andrea Barbiero, and Mercè Gamell, whose expertise in project management, facilitation, documentation, and methodological support was essential to the success of this participatory process. The authors disclose the use of generative artificial intelligence software (Grammarly) for grammar accuracy and writing consistency. No artificial intelligence bots were used to generate contents.

Funding

This study was funded by the European Union – NextGenerationEU, under the Recovery, Transformation, and Resilience Plan (PRTR) of the Spanish Government. The views expressed are those of the authors and not necessarily those of the Spanish Ministry of Health or the Funding bodies, which did not contribute to study design or results interpretation.

Data Availability

The transcription of discussions, pictures of the sticky notes as part of the participatory process, and clustering of themes (in Catalan and Spanish) are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

JP-J conceived and designed the study, developed the methodology, acquired funding, supervised the research, and wrote the original draft of the manuscript. XM, LLV, AC, LP, LH, NV, and AM contributed to study design and methodology development. JP-J, XM, LLV, AC, LP, LH, AB, AF, NV, and AM conducted data analysis and interpretation. AF, NV, and AM managed data collection and curation. XM, LP, NV, AM, and GCS contributed to manuscript writing and revision. AC provided project administration and coordination. NV, AF, and AM created visualizations and figures. All authors contributed to the interpretation of results, critically reviewed the manuscript for important intellectual content, and approved the final version for publication.

Conflicts of Interest

JP-J is Associate Editor of the *Journal of Medical Internet Research.*

Multimedia Appendix 1
Participant selection and recruitment strategy, co-design materials, and participation surveys.
[PDF File, 12001 KB - jmir_v28i1e84129_app1.pdf]

Multimedia Appendix 2
Round 1 outcomes (extended version).
[PDF File, 7802 KB - jmir_v28i1e84129_app2.pdf]

References

1. Crawford A, Serhal E. Digital health equity and COVID-19: the innovation curve cannot reinforce the social gradient of health. J Med Internet Res 2020 Jun 2;22(6):e19361. [doi: 10.2196/19361] [Medline: 32452816]
2. Rodriguez JA, Clark CR, Bates DW. Digital health equity as a necessity in the 21st century cures act era. JAMA 2020 Jun 16;323(23):2381-2382. [doi: 10.1001/jama.2020.7858] [Medline: 32463421]
3. Lyles CR, Sharma AE, Fields JD, Getachew Y, Sarkar U, Zephyrin L. Centering health equity in telemedicine. Ann Fam Med 2022;20(4):362-367. [doi: 10.1370/afm.2823] [Medline: 35879077]
4. Ramsetty A, Adams C. Impact of the digital divide in the age of COVID-19. J Am Med Inform Assoc 2020 Jul 1;27(7):1147-1148. [doi: 10.1093/jamia/ocaa078] [Medline: 32343813]
5. Martikainen S, Korpela M, Tiihonen T. User participation in healthcare IT development: a developers' viewpoint in Finland. Int J Med Inform 2014 Mar;83(3):189-200. [doi: 10.1016/j.ijmedinf.2013.12.003] [Medline: 24382475]
6. Kushniruk A, Nøhr C. Participatory design, user involvement and health IT evaluation. Stud Health Technol Inform 2016;222(139–151):139-151. [Medline: 27198099]
7. Clemensen J, Rothmann MJ, Smith AC, Caffery LJ, Danbjorg DB. Participatory design methods in telemedicine research. J Telemed Telecare 2017 Oct;23(9):780-785. [doi: 10.1177/1357633X16686747] [Medline: 28027678]
8. Arndt BG, Beasley JW, Watkinson MD, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med 2017 Sep;15(5):419-426. [doi: 10.1370/afm.2121] [Medline: 28893811]
9. Irizarry T, DeVito Dabbs A, Curran CR. Patient portals and patient engagement: a state of the science review. J Med Internet Res 2015 Jun 23;17(6):e148. [doi: 10.2196/jmir.4255] [Medline: 26104044]
10. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018 Aug 1;25(8):1080-1088. [doi: 10.1093/jamia/ocy052] [Medline: 29788380]
11. Boy GA, Riedel N. Participatory human-centered design: user involvement and design cross-fertilization. In: Kurosu M, editor. Human Centered Design: Springer; 2009:835-843. [doi: 10.1007/978-3-642-02806-9_96]
12. Giacomin J. What is human centred design? Des J 2014 Dec;17(4):606-623. [doi: 10.2752/175630614X14056185480186]
13. Bazzano AN, Martin J, Hicks E, Faughnan M, Murphy L. Human-centred design in global health: a scoping review of applications and contexts. In: Virgili G, editor. PLoS ONE 2017;12(11):e0186744. [doi: 10.1371/journal.pone.0186744] [Medline: 29091935]
14. Chen E, Leos C, Kowitt SD, Moracco KE. Enhancing community-based participatory research through human-centered design strategies. Health Promot Pract 2020 Jan;21(1):37-48. [doi: 10.1177/1524839919850557] [Medline: 31131633]
15. Göttgens I, Oertelt-Prigione S. The application of human-centered design aproaches in health research and innovation: a narrative review of current practices. JMIR Mhealth Uhealth 2021 Dec 6;9(12):e28102. [doi: 10.2196/28102] [Medline: 34874893]
16. Altman M, Huang TTK, Breland JY. Design thinking in health care. Prev Chronic Dis 2018 Sep 27;15:E117. [doi: 10.5888/pcd15.180128] [Medline: 30264690]
17. Holeman I, Kane D. Human-centered design for global health equity. Inf Technol Dev 2020 Jul 2;26(3):477-505. [doi: 10.1080/02681102.2019.1667289]

XSL•FO
RenderX

18. Bratteteig T, Wagner I. Disentangling power and decision-making in participatory design. 2012 Aug 12 Presented at: PDC '12; Aug 12, 2012; Roskilde, Denmark p. 41-50 URL: https://dl.acm.org/doi/proceedings/10.1145/2347635 [accessed 2025-12-18] [doi: 10.1145/2347635.2347642]
19. Merkel S, Kucharski A. Participatory design in gerontechnology: a systematic literature review. Gerontologist 2019 Jan 9;59(1):e16-e25. [doi: 10.1093/geront/gny034] [Medline: 29788319]
20. Duque E, Fonseca G, Vieira H, Gontijo G, Ishitani L. A systematic literature review on user centered design and participatory design with older people. 2019 Oct 22 Presented at: Proceedings of the 18th Brazilian Symposium on Human Factors in Computing Systems; Oct 22, 2019; Vitória Espírito Santo, Brazil p. 1-11 URL: https://dl.acm.org/doi/proceedings/10.1145/3357155 [accessed 2025-12-18] [doi: 10.1145/3357155.3358471]
21. Clemensen J, Larsen SB, Kyng M, Kirkevold M. Participatory design in health sciences: using cooperative experimental methods in developing health services and computer technology. Qual Health Res 2007 Jan;17(1):122-130. [doi: 10.1177/1049732306293664] [Medline: 17170250]
22. Evans L, Evans J, Pagliari C, Källander K. Scoping review: exploring the equity impact of current digital health design practices. Oxford Open Digital Health 2023 Jan 1;1. [doi: 10.1093/oodh/oqad006]
23. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2025-12-03]
24. Sanders EBN, Stappers PJ. Co-creation and the new landscapes of design. CoDesign 2008 Mar;4(1):5-18. [doi: 10.1080/15710880701875068]
25. Wadley G, Lederman R, Gleeson J, Alvarez-Jimenez M. Participatory design of an online therapy for youth mental health. 2013 Nov 25 Presented at: OzCHI '13; Nov 25, 2013; Adelaide Australia p. 517-526 URL: https://dl.acm.org/doi/proceedings/10.1145/2541016 [accessed 2025-12-18] [doi: 10.1145/2541016.2541030]
26. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119. [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
27. Hatef E, Hudson Scholle S, Buckley B, Weiner JP, Austin JM. Development of an evidence- and consensus-based Digital Healthcare Equity Framework. JAMIA Open 2024 Dec;7(4):ooae136. [doi: 10.1093/jamiaopen/ooae136] [Medline: 39553827]
28. Alvidrez J, Castille D, Laude-Sharp M, Rosario A, Tabor D. The National Institute on Minority Health and Health Disparities Research Framework. Am J Public Health 2019 Jan;109(S1):S16-S20. [doi: 10.2105/AJPH.2018.304883] [Medline: 30699025]
29. Bucher A, Chaudhry BM, Davis JW, et al. How to design equitable digital health tools: a narrative review of design tactics, case studies, and opportunities. In: Kuo PC, editor. PLOS Digit Health 2024 Aug;3(8):e0000591. [doi: 10.1371/journal.pdig.0000591] [Medline: 39172776]
30. Pérez Sust P, Solans O, Fajardo JC, et al. Turning the crisis into an opportunity: digital health strategies deployed during the COVID-19 outbreak. JMIR Public Health Surveill 2020 May 4;6(2):e19106. [doi: 10.2196/19106] [Medline: 32339998]
31. Solans O, Vidal-Alaball J, Roig Cabo P, et al. Characteristics of citizens and their use of teleconsultations in primary care in the Catalan public health system before and during the COVID-19 pandemic: retrospective descriptive cross-sectional study. J Med Internet Res 2021 May 27;23(5):e28629. [doi: 10.2196/28629] [Medline: 33970867]
32. Vidal-Alaball J, López Seguí F, Garcia Domingo JL, et al. Primary care professionals' acceptance of medical record-based, store and forward provider-to-provider telemedicine in Catalonia: results of a web-based survey. Int J Environ Res Public Health 2020 Jun 8;17(11):4092. [doi: 10.3390/ijerph17114092] [Medline: 32521740]
33. Moll J, Rexhepi H, Cajander Å, et al. Patients' experiences of accessing their electronic health records: national patient survey in Sweden. J Med Internet Res 2018 Nov 1;20(11):e278. [doi: 10.2196/jmir.9492] [Medline: 30389647]
34. Norman DA, Draper SW. User Centered System Design: New Perspectives on Human-Computer Interaction 1986. URL: https://www.taylorfrancis.com/books/9781482229639 [accessed 2025-12-03]
35. IDEO. Human Centered Design: Toolkit, 2nd edition: IDEO; 2011.
36. Steen M. Tensions in human-centred design. CoDesign 2011 Mar;7(1):45-60. [doi: 10.1080/15710882.2011.563314]
37. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
38. Trisha G, Seye A. The NASSS Framework – a Synthesis of Multiple Theories of Technology Implementation: Studies in Health Technology and Informatics IOS Press; 2019. [doi: 10.3233/SHTI190123]
39. Marcus HJ, Ramirez PT, Khan DZ, et al. The IDEAL framework for surgical robotics: development, comparative evaluation and long-term monitoring. Nat Med 2024 Jan;30(1):61-75. [doi: 10.1038/s41591-023-02732-7] [Medline: 38242979]
40. Rangachari P, Al Arkoubi K, Shindi R. A multi-level framework for advancing digital health equity in learning health systems: aligning practice and theory with the Quintuple Aim. Int J Equity Health 2025 Oct 7;24(1):253. [doi: 10.1186/s12939-025-02663-4] [Medline: 41057836]
41. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]

XSL•FO
RenderX

42. Marc de la participació ciutadana en salut: facilitem i potenciem la participació ciutadana; fem-ho entre tots [Article in Catalan]. Secretaria d'Atenció Sanitària i Participació. 2017. URL: https://scientiasalut.gencat.cat/bitstream/handle/11351/3503/marc_participacio_ciutadana_salut_2017.pdf?sequence=1&isAllowed=y [accessed 2025-08-11]
43. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015 Sep;42(5):533-544. [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
44. Baum FE, Ziersch AM, Zhang G, Osborne K. Do perceived neighbourhood cohesion and safety contribute to neighbourhood differences in health? Health Place 2009 Dec;15(4):925-934. [doi: 10.1016/j.healthplace.2009.02.013] [Medline: 19403326]
45. Hirko KA, Kerver JM, Ford S, et al. Telehealth in response to the COVID-19 pandemic: implications for rural health disparities. J Am Med Inform Assoc 2020 Nov 1;27(11):1816-1818. [doi: 10.1093/jamia/ocaa156] [Medline: 32589735]
46. Trebble TM, Hansi N, Hydes T, Smith MA, Baker M. Process mapping the patient journey: an introduction. BMJ 2010 Aug 13;341(aug13 1):c4078. [doi: 10.1136/bmj.c4078] [Medline: 20709715]
47. McCarthy S, O'Raghallaigh P, Woodworth S, Lim YL, Kenny LC, Adam F. An integrated patient journey mapping tool for embedding quality in healthcare service reform. Journal of Decision Systems 2016 Jun 10;25(sup1):354-368. [doi: 10.1080/12460125.2016.1187394]
48. Simonse L, Albayrak A, Starre S. Patient journey method for integrated service design. Design for Health 2019 Jan 2;3(1):82-97. [doi: 10.1080/24735132.2019.1582741]
49. Best A, Greenhalgh T, Lewis S, Saul JE, Carroll S, Bitz J. Large-system transformation in health care: a realist review. Milbank Q 2012 Sep;90(3):421-456. [doi: 10.1111/j.1468-0009.2012.00670.x] [Medline: 22985277]
50. Joseph AL, Kushniruk AW, Borycki EM. Patient journey mapping: Current practices, challenges and future opportunities in healthcare. Knowl Manag E-Learn 2020 Dec 26:387-404. [doi: 10.34105/j.kmel.2020.12.021]
51. WeCare Toolkit. National Council of Social Service. 2016. URL: https://isomer-user-content.by.gov.sg/24/c875b227-ad66-4127-b982-34412636febd/We-Care-Toolkit.pdf [accessed 2025-10-26]
52. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
53. Beyer H, Holtzblatt K. Contextual Design: Defining Customer-Centered Systems: Morgan Kaufmann; 2009.
54. Miles MB, Huberman AM, Saldaña J. Qualitative Data Analysis: A Methods Sourcebook: SAGE; 2020.
55. Martin B. Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions: Rockport Publishers; 2012.
56. Kolko J. Exposing the Magic of Design: A Practitioner's Guide to the Methods and Theory of Synthesis First Issued as an Oxford University Press Paperback: Oxford University Press; 2015.
57. Ratwani RM, Savage E, Will A, et al. Identifying Electronic Health Record Usability And Safety Challenges In Pediatric Settings. Health Aff (Millwood) 2018 Nov;37(11):1752-1759. [doi: 10.1377/hlthaff.2018.0699]
58. Barker W, Chang W, Everson J, et al. The evolution of health information technology for enhanced patient-centric care in the United States: data-driven descriptive study. J Med Internet Res 2024 Oct 28;26:e59791. [doi: 10.2196/59791] [Medline: 39466303]
59. Kilfoy A, Hsu TCC, Stockton-Powdrell C, Whelan P, Chu CH, Jibb L. An umbrella review on how digital health intervention co-design is conducted and described. NPJ Digit Med 2024 Dec 23;7(1):374. [doi: 10.1038/s41746-024-01385-1] [Medline: 39715947]
60. Jabour AM. Putting patients at the center of health information exchange design: an exploration of patient preferences for data sharing. Health Informatics J 2024;30(3):14604582241277029. [doi: 10.1177/14604582241277029] [Medline: 39142341]
61. van Velsen L, Ludden G, Grünloh C. The limitations of user-and human-centered design in an eHealth context and how to move beyond them. J Med Internet Res 2022 Oct 5;24(10):e37341. [doi: 10.2196/37341] [Medline: 36197718]
62. Wang T, Emami E, Jafarpour D, Tolentino R, Gore G, Rahimi SA. Integrating equity, diversity, and inclusion throughout the lifecycle of artificial intelligence for healthcare: a scoping review. In: Kuo PC, editor. PLOS Digit Health 2025 Jul;4(7):e0000941. [doi: 10.1371/journal.pdig.0000941] [Medline: 40658719]

Abbreviations

EHR: electronic health record
PHCD: participatory human-centered design
SRQR: Standards for Reporting Qualitative Research
WHO: World Health Organization

XSL•FO
RenderX

Edited by S Brini; submitted 15.Sep.2025; peer-reviewed by C Xie, C Kunze; revised version received 19.Nov.2025; accepted 20.Nov.2025; published 06.Jan.2026.

Please cite as:
Piera-Jiménez J, Vilarasau Creus N, Maymó Costa A, Michelena X, Climent Fageda A, Farré A, Herczeg L, Parameswaran L, Carot-Sans G, Valle L
Cocreating Principles for Digital Health Equity: Cross-Sectional, Qualitative Study for Participatory Human-Centered Design in Catalonia
J Med Internet Res 2026;28:e84129
URL: https://www.jmir.org/2026/1/e84129
doi:10.2196/84129

© Jordi Piera-Jiménez, Núria Vilarasau Creus, Ada Maymó Costa, Xabier Michelena, Andrea Climent Fageda, Alèxia Farré, László Herczeg, Lekshmy Parameswaran, Gerard Carot-Sans, Luis Valle. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Telemedicine Adoption for Managing Chronic and Rare Diseases in Indonesia During and Beyond the COVID-19 Era: Qualitative Study

Christine Gracia Pratama[1], MBA; Rima Nurlianti[1], MSc; Cornelius Herstatt[2,3], Prof Dr oec publ; Moritz Goeldner[1,4], Prof Dr

[1]Technology and Innovation Management, Hamburg University of Technology (TUHH), Am Schwarzenberg-Campus 4, Hamburg, Germany

[2]Leuphana University of Lüneburg, Lüneburg, Germany

[3]University of Auckland, Auckland, New Zealand

[4]Working Group for Data-Driven Innovation, Hamburg University of Technology, Hamburg, Germany

Corresponding Author:
Christine Gracia Pratama, MBA
Technology and Innovation Management, Hamburg University of Technology (TUHH), Am Schwarzenberg-Campus 4, Hamburg, Germany

Abstract

Background: Telemedicine has emerged as a valuable tool for improving health care delivery, especially in low-resource and geographically isolated regions. In Indonesia, the COVID-19 pandemic highlighted the need for digital health solutions to manage chronic diseases and improve access to specialists.

Objective: This study aimed to examine the rapid adoption of telemedicine in Indonesia, focusing on its role in managing chronic and rare diseases and highlighting both its benefits and challenges to long-term viability during and post pandemic.

Methods: A qualitative study was used, following the Consolidated Criteria for Reporting Qualitative Research 32-item checklist and guided by established theories and frameworks. Recruitment followed a systematic multistage procedure suited to the context of Indonesia, an upper-middle-income country and Group of Twenty member characterized by wide regional and socioeconomic disparities. Data were collected through 15 semistructured interviews conducted between October and December 2024. The interviews were divided into 2 phases: first with physicians from various medical specialties and 6 out of 7 regions within Indonesia, and then with patients diagnosed with autoimmune diseases. All interviews were audio-recorded, transcribed verbatim, and analyzed inductively to identify recurrent themes and subthemes.

Results: Before the COVID-19 pandemic, telemedicine adoption in Indonesia was limited, with many physicians recalling its nonexistence before 2015. The pandemic accelerated the adoption of virtual platforms, which became essential for patient consultation, follow-ups, and medication management. Though its use decreased after the pandemic, telemedicine remains valuable, particularly for chronic and rare disease management in remote areas with limited access to specialists. At least three overarching themes were identified: (1) acceleration and normalization of telemedicine (the pandemic catalyzed the integration of virtual consultations into routine care, transforming initial skepticism into broad acceptance); (2) accessibility and cost efficiency (telemedicine reduced financial and travel burdens, especially for patients in remote areas, with several reporting cost savings from hybrid consultation models); and (3) limitations of virtual care (physicians highlighted constraints related to the absence of physical examination).

Conclusions: The findings reveal that telemedicine has become a valuable tool in Indonesia for managing chronic and rare diseases, particularly for follow-up care and medical specialist access across diverse geographical areas. While telemedicine has improved health care accessibility and demonstrated significant cost benefits by reducing transportation and consultation costs, challenges such as limited infrastructure and physician burnout remain. Long-term success will depend on the development of sustainable regulatory frameworks, continued investment in digital infrastructure, and a focus on optimizing cost-effectiveness in health care delivery.

(J Med Internet Res 2026;28:e83462) doi:10.2196/83462

KEYWORDS

digital health; telemedicine; chronic disease; rare disease; COVID-19; Indonesia

XSL•FO
RenderX

Introduction

Background

Digital technologies offer new ways to improve health care coverage and quality, particularly in settings with limited resources and widespread geographic dispersion. Integrating digital technologies has made care more efficient, accessible, and patient-centered. The global adoption of digital health care, such as telemedicine, was accelerated by the COVID-19 pandemic [1-3]. Telemedicine has gained prominence as a crucial component of health care delivery. It was strongly advocated during the pandemic for its potential to address immediate health care needs while reducing the risk associated with in-person visits [4-6]. This shift was not just a temporary solution; it marked the beginning of a permanent transformation in health care delivery models in both the Global North and the Global South. Importantly, telemedicine also holds potential to support progress toward sustainable development goal (SDG) 3 (good health and well-being) and goal 10 (reducing inequalities) by enabling more equitable access to health care services across diverse socioeconomic and geographic contexts.

With a population exceeding 270 million and more than 17,500 islands, Indonesia faces significant disparities in health care access due to its fast-growing population, urbanization, and huge distances between health care providers and patients. This intensified during the pandemic, particularly among individuals with chronic, noncommunicable diseases (NCDs) [7]. Therefore, digital health solutions have the potential to ensure equitable and continuous care in this region. Early evidence suggests that telemedicine could reduce transportation costs by 4% and shorten patient travel times without compromising the level of care [8]. However, amidst the effectiveness of telemedicine, barriers such as digital literacy, educational level, age, rural residence, and internet access continue to affect its adoption [9-11].

Evolution, Definitions, and Current Relevance

The digital transformation of health care systems has the potential to greatly improve quality while reducing costs [12,13]. Telemedicine, which originated in the early 20th century with the use of telephone lines to transmit medical data, has evolved significantly and become a vital component of modern health care delivery. Early applications, such as remotely transmitting electrocardiograms and X-rays, laid the groundwork for telemedicine's role in improving health care access, particularly in emergencies and rural areas [14].

The World Health Organization defines telemedicine as the use of information and communication technology to provide health care remotely, particularly in regions with limited access to medical services [12]. Telemedicine is highly patient-centered because it is convenient and caters to patients' digital preferences. This makes telemedicine an essential component of patient care management in areas with limited traditional options [15]. Telemedicine enables communication between health care providers and patients through technology [16]. It encompasses both synchronous and asynchronous methods of delivering health care services remotely and is provided through various digital platforms, including secure messaging, mobile apps, and video conferencing [16-19]. Telemedicine is a broad health care service that uses telecommunication and electronic technologies to support remote consultations, monitoring, nursing, and rehabilitation between patients and health care providers [4,20].

Telemedicine Adoption in Asia-Pacific and Organization for Economic Cooperation and Development Countries

In the Asia-Pacific region, telemedicine experienced substantial growth during the COVID-19 pandemic, nearly doubling between 2019 and 2021. Australia recorded a ninefold increase in usage during this period [20,21]. However, postpandemic trends show a stabilization or decline in some countries. For instance, between 2021 and 2023, Singapore and Australia each saw a 9 percentage points rise in telemedicine adoption (from 34% to 43% and 45% to 54%, respectively), while Indonesia's adoption increased by 7 percentage points (from 51% to 58%) [22,23]. Conversely, from the same study, India experienced a 4 percentage points decline (from 59% to 55%), reflecting regional disparities in digital infrastructure, regulatory support, and user engagement.

In Organization for Economic Cooperation and Development countries, telemedicine use was limited before the pandemic, with an average of only 0.6 teleconsultations per patient annually in 2019 [24]. Sweden led in adoption, with 47% of Swedish adults reporting in early 2021 that they had at least 1 teleconsultation since the start of the pandemic [25,26]. France also saw a dramatic spike, from 136,882 consultations in 2019 to 4.5 million in April 2020 alone, driven by policy changes such as full reimbursement [27]. Germany's growth was short-lived; after a surge from 3000 consultations in 2019 to 2.7 million in 2020, usage declined by 40% in 2022 [28,29]. Physician adoption mirrored this trend: in Sweden, 50% of physicians conducted 25% - 75% of consultations online weekly, compared to just 6% in France and 14 % in Germany [30]. These patterns underscore the need for adaptive digital health policies tailored to each country's health care context [31].

Digital Health Adoption and COVID-19

The COVID-19 pandemic catalyzed an unprecedented acceleration in digital adoption across health care systems globally, particularly in the area of telemedicine. Mobile apps offering teleconsultation, contact tracing, symptom monitoring, and remote patient management became essential tools for maintaining health care continuity during lockdowns. Regional distinctions shaped the design and use of these platforms: for example, Asian countries emphasized user accessibility and functional simplicity to promote widespread adoption, while North American and European apps primarily focused on disseminating health information with limited interactivity [32]. Additionally, global differences in regulatory environments, such as Germany's stringent data privacy laws compared to India's more rapid, utilitarian deployment of contact-tracing apps, reflect diverse policy priorities during the pandemic [33]. These variances underscore the importance of localized strategies for telehealth implementation and regulation, particularly as countries transition into the postpandemic phase.

XSL•FO
RenderX

Post-COVID-19 telemedicine is increasingly recognized as a foundational element of modern health care, extending far beyond its initial role as an emergency response. The World Health Organization emphasized its potential not only for acute care but also as a sustainable solution for managing long-term health conditions. Chronic diseases, which demand ongoing monitoring and personalized care, have emerged as a central focus for telehealth apps [5,6]. This is particularly relevant for rare diseases, which often require specialist consultations and coordinated care that may not be readily available in all geographic regions. The decentralized model enabled by telemedicine can mitigate geographic and systemic barriers, facilitating access to specialist expertise for rare disease patients and improving outcomes through timely interventions and continuity of care [34,35].

Indonesia exemplifies the strategic integration of telemedicine into national health planning in the postpandemic context. With a population exceeding 270 million distributed across thousands of islands, the government has prioritized digital health to overcome logistical challenges and expand health care accessibility [36]. The Ministry of Health's "Blueprint of Digital Transformation Strategy 2024" developed in collaboration with the United Nations Development Program outlines a comprehensive framework for digital health system development [37,38]. This includes infrastructure to support the management of chronic and rare diseases, particularly in view of the increasing prevalence of NCDs in the country [7,39]. The blueprint reflects a broader regional trend toward embedding telehealth within long-term health care policy, signaling a shift from reactive digital health adoption to proactive, system-wide transformation. Such strategies are vital to ensuring equitable care for patients with chronic and rare diseases in both urban and remote settings.

Use of Telemedicine for Chronic and Rare Disease Management

Chronic and rare disease management requires a continuous, coordinated, and patient-centered approach, focusing on monitoring, medication adherence, education, and preventive care to improve outcomes and quality of life [40,41]. Chronic NCD diseases such as diabetes, cardiovascular conditions, and autoimmune disorders—including rare diseases like systemic lupus erythematosus and Sjögren's disease—demand sustained health care interventions due to their complexity and long-term nature [42,43]. Telemedicine has emerged as a critical enabler of chronic care, particularly during and after the COVID-19 pandemic, by bridging geographic gaps, reducing travel costs, and supporting remote management for patients in underserved areas [32,44,45]. In Indonesia, for example, the Program Pengelolaan Penyakit Kronis under Badan Pengelola Jaminan Sosial Kesehatan (BPJS; National Health Insurance Agency) illustrates the integration of telemedicine in chronic disease management, offering education, routine monitoring, and preventive care for diabetes and hypertension [24,46].

The post-pandemic era has highlighted telemedicine's broader clinical use in managing both chronic and rare diseases. Studies demonstrate its efficacy in improving medication adherence and reducing psychological distress among patients with conditions such as rheumatoid arthritis, while also enhancing satisfaction in postoperative and psychiatric care when delivered via video consultations [47-50]. Although many autoimmune patients still prefer initial face-to-face consultations, hybrid models—where video is used for follow-ups—have been well received, with 84% reporting video consultations as effective when paired with subsequent in-person visits [43]. Beyond synchronous video, asynchronous tools such as SMS have proven particularly effective among younger patients, promoting continuous engagement through reminders and 2-way communication [51]. To ensure quality and patient safety, frameworks like the Asia-Pacific League of Associations for Rheumatology telemedicine guidelines recommend time-limited follow-ups, prioritization through preconsultation screening, and strict adherence to data privacy standards [52]. Collectively, these strategies underscore telemedicine's critical role in enhancing access, continuity, and personalization of care for patients with complex and chronic conditions.

Cost-Benefit Analysis

Telemedicine has emerged as a cost-effective solution for improving health care access and quality, particularly in regions facing shortages of medical professionals. In Association of Southeast Asian Nation countries, including Indonesia, the scarcity of physicians in rural areas exacerbates health care challenges. With only 7.5 medical practitioners per 10,000 population, Indonesia lags behind other nations like Malaysia and Singapore, where the availability of physicians is much higher. Telemedicine closes this gap by offering remote consultations, which reduces the need for patients to travel and minimizes health care costs for patients and providers alike [49]. The increasing adoption of telemedicine has proven to reduce consultation, travel, and time-related expenses, while also enhancing the overall quality of care.

Studies from Europe also highlight the cost-effectiveness of telemedicine. A comprehensive analysis found that telemedicine was cost-effective in 73.3% of cases, with reduced consultation costs and travel expenses being significant contributors to savings [53]. In Sweden, for example, digital consultations cost 41.5% less than in-person visits [54]. Similarly, a remote monitoring program for diabetic patients in France demonstrated cost savings, with the remote group spending €1334 (US $1513, December 2021 exchange rate) less compared to the control group over 1 year [55]. These findings underscore the potential of telemedicine to reduce both health care costs and improve patient outcomes, particularly in systems where teleconsultations are integrated into public health systems with reimbursement options.

Telemedicine also shows considerable promise in reducing nonmedical expenses, such as transportation costs and lost time from work or school. Studies from Bangladesh and pediatric rheumatology research highlight the substantial savings telemedicine provides by cutting travel time by 56% and associated costs by 94% [43,56]. In Indonesia, the cost of teleconsultations varies depending on the platform, with independent companies offering more affordable options compared to hospital-integrated services. However, telemedicine remains largely inaccessible through public health insurance,

XSL•FO
RenderX

as it is only reimbursed via private insurance or self-payment [57]. Despite these limitations, text messaging and other telemedicine tools continue to be cost-effective solutions for chronic disease management, particularly in underserved areas with limited health care access [58,59]. This qualitative study aims to examine how access to telemedicine supports care for managing chronic and rare diseases in Indonesia, drawing on perspectives from both patients and health care professionals in the post-pandemic context.

Methods

Ethical Considerations

Informed consent was obtained from all participants prior to data collection. Data were collected, anonymized, and stored securely in line with international ethical standards. No personal or social information was collected, and all data was fully anonymized to ensure privacy and confidentiality. Participants were not offered any compensation for taking part in the study.

This study was conducted in accordance with the principles of the Declaration of Helsinki. Ethical approval was obtained from the German Association for Experimental Economic Research Institutional Review Board (certificate number syBSVy9b). No identifying images or personal or clinical details of participants are included.

Research Design

This study adopts a qualitative research design following the Consolidated Criteria for Reporting Qualitative Research (COREQ) 32-item checklist as provided as Checklist 1 [60,61]. The topic guides were developed by RN and CGP in accordance with the COREQ 32-item checklist to ensure comprehensive coverage of key methodological domains. Guide development was informed by prior literature on telemedicine adoption and qualitative interviewing and subsequently reviewed by MG to establish content validity. A deductive analytic approach was applied, drawing on established theories and frameworks to inform the research questions and interview guide development.

The study comprises two primary participant groups: physicians and patients. Semistructured interviews were chosen as the main data collection method due to their suitability for eliciting in-depth responses on telemedicine experiences and challenges [62,63]. The interviews were conducted in 2 phases. First, physicians from different regions and medical specialties were interviewed. Next, case studies of patients diagnosed with autoimmune diseases were examined.

Data Collection

Recruitment

Recruitment followed a systematic multistage procedure suited to the context of Indonesia, an upper-middle-income country and Group of Twenty member characterized by wide regional and socioeconomic disparities [64]. For physicians, the inclusion criteria required active medical practice in Indonesia, prior experience using or familiarity with telemedicine in managing chronic or rare diseases, and willingness to participate in an online interview. For patients, inclusion criteria included having a clinically confirmed chronic or rare autoimmune condition and experience with telemedicine services. A social media-based strategy was used, as it has proven effective in engaging participants and achieving broader outreach across socioeconomic categories [65-68]. Physicians were approached through professional platforms (LinkedIn and ResearchGate), email, and Instagram, while patients were recruited through content-sharing platforms (Instagram and TikTok) and WhatsApp community groups. This multiplatform strategy enabled access to both health care professionals and hard-to-reach patient populations [66,68,69]. Previous studies underscored notable gender disparities, revealing a predominance of female patients, which corresponds with the higher incidence of autoimmune diseases observed in women [69].

Furthermore, standardized invitations outlining the study, eligibility criteria, and confidentiality were shared, thus participation occurred according to respondents' availability. Digital consent was obtained prior to interviews. Although social media enabled broad outreach, engagement from marginalized groups remained uneven [65].

For this study, Indonesia was categorized into Western and Eastern regions. Western Indonesia comprises Sumatra, Java, and Kalimantan, while Eastern Indonesia includes Nusa Tenggara, Sulawesi, Maluku, and Papua. Participants were drawn from 6 of Indonesia's 7 major regions, including Sumatra, Java, Kalimantan, Nusa Tenggara, Sulawesi, and Papua. Recruitment from the Maluku Islands proved challenging due to the absence of referral connections with physicians or patients with autoimmune conditions, as well as the region's small population.

National telecommunications statistics indicate substantial regional disparities in digital connectivity. Western Indonesia generally demonstrates higher internet access, with an average of 80% of the population, compared to around 40% in the Eastern regions [70]. The lowest internet penetration was observed in parts of Papua, where access declined to approximately 6% in 2024 [70]. Similar results were found for cellular signal coverage across 7 Indonesian regions [70].

Interviews With Physicians and Patients

A total of 35 physicians were approached, resulting in 15 semistructured interviews conducted between August and November 2024. The participants, representing 6 medical specialties, were chosen for their experience with telemedicine, research involvement, or public health advocacy. For physicians, the inclusion criteria required active medical practice in Indonesia, prior experience using or familiarity with telemedicine in managing chronic or rare diseases, and willingness to participate in an online interview. For patients, inclusion criteria included having a clinically confirmed chronic or rare autoimmune condition and experience with telemedicine services. Each interview, conducted in Bahasa Indonesia via Zoom (Zoom Communications, Inc) or WhatsApp (Meta Platforms, Inc), lasted 25 to 63 minutes (mean 48, SD 13.4), following a 5-section interview guide with 27 questions as shown in Appendix S1 (Multimedia Appendix 1). In total, the interviews yielded 743 minutes of recordings and 163 pages of transcripts. The interviews were primarily conducted by RN,

XSL•FO
RenderX

with several sessions co-led or supported by CGP. Physicians were recruited through social media, professional platforms, and referrals, with 1 session including a passive observer (a community leader).

In the second phase, 15 patients diagnosed with autoimmune diseases were contacted, leading to 9 interviews conducted between November 15 and December 15, 2024. All interviews were conducted by RN. Recruitment took place through patient communities on Instagram (Meta Platforms, Inc) and WhatsApp (Meta Platforms, Inc), with consent from community leaders. Each interview had a typical duration of 26 to 70 minutes (mean 54, SD 16.1), guided by a 6-section, 34-question format, and transcribed verbatim. Appendix S3 (Multimedia Appendix 1) contains the interview guide, and a detailed interview summary is attached in Appendix S4 (Multimedia Appendix 1). The total output was 539 minutes of recordings and 197 transcript pages, with 1 session also involving a nonparticipant community leader. Table 1 summarizes key details of both phases.

Table . Overview of interview characteristics and participant details.

Category	Physicians	Patients
Interview period	October 10-November 10, 2024	November 15-December 15, 2024
Participants reached	35 physicians	15 patients
Number of participants	15 physicians	9 patients
Recruitment method	Direct outreach via ResearchGate, LinkedIn, Instagram, and referral through mutual connections	Community posts and broadcasts on Instagram and WhatsApp
Language	Bahasa Indonesia	Bahasa Indonesia
Interview duration	25-63 min (mean 48, SD 13.4)	26-70 min (mean 54, SD 16.1)
Interview format	Semistructured (5 sections and 27 questions)	Semistructured (6 sections and 34 questions)
Total recorded time	743 min	539 min
Transcript length	163 pages	197 pages

Data Analysis

Thematic analysis is used to systematically identify, analyze, and interpret patterns within qualitative data [71]. Transcripts were initially reviewed for accuracy and completeness. Zoom (Zoom Communications, Inc) interviews were transcribed using built-in features, while WhatsApp (Meta Platform) recordings were processed using QuickTime Player (Apple Inc) and TurboScribe (Leif Erikson Ventures, LLC). All transcripts were manually checked and then translated into English using ChatGPT 4.0 (OpenAI) and DeepL (DeepL SE) to ensure consistency and accessibility for non-Indonesian—speaking readers. All translated transcripts were then reviewed again manually by 2 of the authors, both of whom are native speakers and 1 of whom has medical knowledge. To maintain confidentiality and reduce participant burden, transcripts were not returned to participants for verification. Identifying details were anonymized throughout. The data analysis process is displayed in Figure 1.

Figure 1. Steps involved in the qualitative data analysis process.

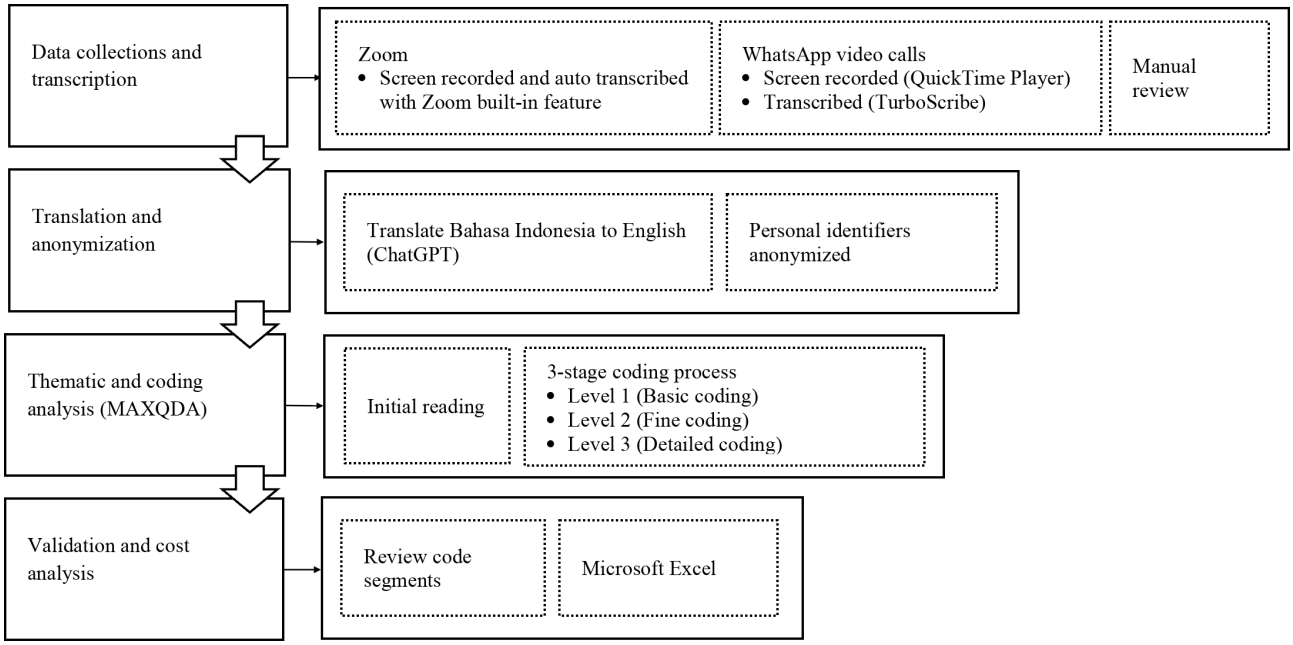

XSL•FO
RenderX

Data were analyzed using MAXQDA (VERBI Software GmbH), following a structured, 3-level coding process [72]. First-level coding identified broad themes that served as the foundation for analysis. Second-level coding, or fine coding, further refined these categories into specific subcategories for deeper analysis of telemedicine experiences. Third-level coding introduced additional granularity, breaking down subcategories into more specific classifications to capture nuanced patterns within the data. Coding was primarily conducted by RN, while CGP and MG independently reviewed portions of the transcripts to validate emerging codes and themes. Theme validation extended beyond partial co-coding through iterative comparison of the final themes with the original transcripts to ensure accurate representation of participants' perspectives. Interpretive discrepancies were discussed among the authors and resolved through consensus. In addition, the researchers also held regular discussions throughout the coding process and participated in debriefing sessions with MG to further strengthen analytical rigor. Data saturation was determined by RN during the coding process and subsequently confirmed through joint discussions with CGP and MG. Saturation was considered achieved once participants from all major Indonesian regions were represented and recurring patterns consistently emerged across interviews. This determination was further supported by verification through field notes and preliminary coding reviews, with no additional first-order concepts or second-order themes identified [73].

In addition to thematic analysis, cost-related data mentioned in patient interviews (eg, transportation, consultation, and medication costs) were extracted and compiled using Microsoft Excel (Microsoft Corporation) to support an exploratory cost-benefit dimension.

Interviews with physicians were thematically coded across 6 domains: general experience, telemedicine usage, perceived benefits, challenges, health care accessibility, and future outlook. Physicians highlighted telemedicine's effectiveness in monitoring stable patients, especially in remote areas, citing improved accessibility and cost-efficiency. However, they also noted challenges such as diagnostic limitations, poor internet connectivity, legal concerns, and disparities in patient digital literacy. Telemedicine adoption varied by region, case complexity, and health care setting, with usage patterns influenced by the COVID-19 pandemic and shifting postpandemic practices. The coding structure is visualized in Appendix S5 (Multimedia Appendix 1).

Patient interviews were classified according to demographics, medical history, telemedicine usage, health care access and costs, challenges, and overall perceptions. Participants reported diverse experiences shaped by illness severity, hospital class, payment methods (eg, BPJS and private insurance), and regional access to specialists. Many appreciated the reduced travel and financial burden telemedicine offered, while others preferred in-person care for complex diagnostics. Barriers included infrastructure limitations, insurance gaps, and varying levels of digital literacy. Overall, patients valued the flexibility of telemedicine but saw it as a complementary tool rather than a full replacement for face-to-face consultations. Cost-related findings were supported by comparative analysis using MAXQDA (VERBI Software GmbH) and Microsoft Excel (Microsoft Corporation) to highlight reductions in nonmedical expenses. Appendix S6 (Multimedia Appendix 1) presents the coding analysis level of the patient.

Results

Geographic Distribution of Participants

Figure 2 illustrates the geographic distribution of study participants across Indonesia. Physicians and patients are represented according to their respective regions. This highlights the coverage across 6 major regions and shows regional representation.

Figure 2. Distribution of physician and patient interviews across Indonesia.

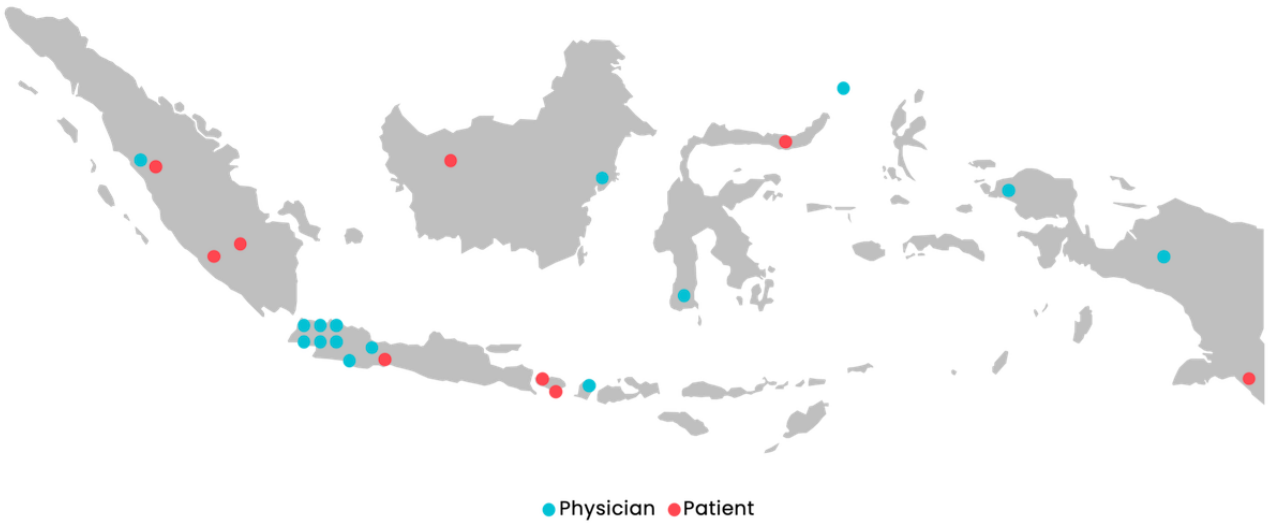

XSL•FO
RenderX

Physicians' Perspectives

Overview

The study involved 15 physicians from 6 regions in Indonesia, as illustrated in Figure 2, representing a range of specialties: general practice (n=5), immunology (n=4), rheumatology (n=3), plastic surgery (n=1), pediatrics (n=1), and ophthalmology (n=1). Participants' clinical experience varied, with 2 participants having over 20 years of practice, 6 having 16 - 20 years, 2 having 6 - 10 years, and 5 having between 1 - 5 years of experience.

Additionally, the physicians practiced across all hospital classifications, as defined in Article 340/MENKES/PER/III/2010, which categorizes hospitals based on service capabilities, infrastructure, and available resources [74-76]. The classification system categorizes hospitals into four types: class A, B, C, and D. Class A hospitals, primarily located in major cities, serve as top referral centers with extensive facilities and a wide range of medical and subspecialty services. Class B hospitals, mostly found in regional areas, provide at least 11 specialist and limited subspecialist services. Class C hospitals focus on essential specialist care and are equipped with a minimum number of medical specialties. Class D hospitals, commonly referred to as Community Health Centers (Pusat Kesehatan Masyarakat), offer basic medical services with limited specialized care.

Evolution of Telemedicine Adoption

Before the COVID-19 pandemic, telemedicine adoption in Indonesia was sporadic and limited. Among 15 physicians surveyed, 12 reported using telemedicine. Additionally, 1 physician recalled that "telemedicine didn't exist yet" around 2015, reflecting the nascent state of digital health at the time.

The pandemic became a turning point. A total of 11 out of 15 physicians began using virtual platforms such as WhatsApp (Meta Platforms, Inc), Zoom (Zoom Communications, Inc), standalone health tech apps like HaloDoc (PT Media Dokter Investama) and Alodokter (PT Alodokter Teknologi Indonesia), as well as hospital-based telemedicine systems [65-68]. During this period, telemedicine was particularly valuable for triaging patients, conducting follow-ups, and managing medications. The need to reduce physical contact due to infection risks made remote consultations not just practical but essential. Many hospitals and clinics swiftly adopted telemedicine services, though often with basic tools or improvised infrastructure.

As COVID-19 cases declined, physicians observed a reduction in telemedicine usage as patients gradually returned to in-person visits. Still, telemedicine continues to serve an important role, especially for follow-up care and for managing chronic or rare conditions. This transition is reflected in 1 physician's experience:

> *At one point, I was handling up to 23 patients a day, but during a single Zoom call, there could be five family members present. Afterward, I would continue treating the rest of my patients. That was mostly in 2019 - 2020, but then the numbers started declining. Post-COVID, it's been fewer, I primarily manage autoimmune patients from remote areas.* [D13]

While full replacement of in-person care remains unlikely, the pandemic has established telemedicine as a lasting complement within hybrid care models. Half of the physicians viewed it as unsuitable for initial diagnoses, but they now use it for monitoring, education, second opinions, and psychological support. This approach is particularly effective for extending care and reducing financial burdens to patients in remote locations. In this context, pediatric cardiologists in Jakarta conduct monthly follow-ups via video consultation with patients in Aceh, located in the westernmost part of Indonesia. Given the long waiting times and the need for continuous monitoring, a process known as the surgical conference is used to assess and prioritize patient conditions. One physician shared:

> *They are usually asked to have monthly follow-ups via the online clinic. They also check the status of the surgical conference queue. So, it just keeps moving along.* [D4]

Challenges

In total, 14 out of 15 physicians identified the lack of physical examination as a key limitation in telemedicine, as noted:

> *However, if I've never met a patient in person, never done a physical examination, or never had direct contact, it's challenging to diagnose complex diseases [...] for diagnoses that require physical examinations, telemedicine can't fully replace in-person visits.* [D7]

Telemedicine has also increased workloads. Two physicians reported burnout due to additional consultation demands. A physician explained:

> *Right now, telemedicine is just a replacement for face-to-face consultations. It does not reduce the workload; it adds to it. [...] So, telemedicine adds extra time slots. To reduce burnout, we need more doctors and better networking.* [D7]

Legal uncertainty adds another layer of concern. Five physicians questioned data ownership when using third-party apps, emphasizing "[a]nother issue is significant patient privacy concerns regarding data security and encryption measures" (D9).

Future Directions

Telemedicine is advancing toward greater accessibility, efficiency, and diagnostic accuracy. A key innovation is the creation of specialized hubs linking medical teams across hospitals. As explained by a physician, this hub-and-spoke model is underway at his or her hospital, beginning with neurosurgery:

> *For example, at our hospital, what we plan for the future is to have this. Currently, our neurosurgery team is already working on it. The main hub for the neurosurgery team is at our hospital. But then, in every hospital branch, there is a hub that becomes part of the neurosurgery network.* [D7]

XSL•FO
RenderX

Another promising development is artificial intelligence (AI)–assisted diagnostics. A physician highlighted a project targeting diabetic retinopathy:

> *We are currently developing AI tools for diabetic retinopathy [...] The AI analyzes the photos to categorize them automatically [...] In remote regions, access to equipment is still a challenge.* [D8]

Additionally, a physician cited virtual reality (VR) integration in Japan to support remote assessments:

> *As currently being implemented at the University of Nagasaki, Japan [...] Patients visit a clinic on a remote island, and physicians remotely assess their condition using VR.* [D15]

Patients' Perspectives

Participant Characteristics

The data is based on interviews with 9 patients diagnosed with various autoimmune conditions, including rheumatoid arthritis, systemic lupus erythematosus, Sjögren's disease, psoriasis, antiphospholipid syndrome, Graves' disease, and scleroderma, with time since diagnosis ranging from 2 to 17 years. Participants, aged between 26 and 59 years, came from diverse professional backgrounds and lived in both urban and rural areas. Seven were covered by Indonesia's national health insurance (BPJS), while 2 financed their health care independently.

Telemedicine Adoption Shaped by Health Care Access

Patients with autoimmune diseases in Indonesia face significant disparities in health care access. Accurate diagnosis typically requires referral to class A hospitals located in major cities, due to the scarcity of subspecialists in rural or outer island regions. The average diagnosis delay exceeded 1 year. One patient was diagnosed in Malaysia due to the lack of domestic expertise during the early 2000s, as local specialists were unavailable at the time.

While some patients now receive initial evaluations at class B hospitals, advanced care still necessitates travel to central facilities. A patient from eastern Indonesia reported a 400 km journey, taking 10 hours by bus through mountainous terrain, costing up to IDR 10 million (€580) per visit.

> *The eight to ten hours by land is really far, especially with the poor infrastructure with mountain roads and winding paths [...], so my mother called instead. There is no consultation fee; only the medicine has a cost, but it is expensive.* [P6]

Due to the physical and financial burden, the patient was later transferred to a nearer class B hospital, albeit with limited specialist support.

Telemedicine has emerged as a critical solution for bridging these access gaps. Among the 9 patients interviewed, those living in urban or semiurban areas within 5 km of a hospital preferred in-person care, citing convenience and minimal travel time.

> *The distance is probably one kilometer. You can walk or ride a motorbike. It usually takes about 5-10 minutes by motorbike.* [P9]

In contrast, 6 patients residing over 50 km from the nearest hospital relied heavily on telemedicine for follow-ups and flare-ups. Poor infrastructure meant that a 50 km trip could take up to 3 hours, pushing patients to communicate directly with their physicians via WhatsApp text and video calls. This is an informal, yet reliable channel in the absence of more robust telehealth systems.

> *I once had a check-up, and everything was fine, with no complaints for a month. Then, just a week after the check-up, I suddenly experienced pain in my hand, swelling, and an unusual toothache. When I consulted a dentist, they found no inflammation, but the pain persisted. [...] I messaged the doctor directly and sent a photo of the dental examination results.* [P2]

Notably, 3 patients with multiple chronic conditions reported using WhatsApp video calls during emergencies to replace in-person visits. They maintained direct contact with their physicians with whom they had been in care for more than 5 years.

> *I was able to video call my personal doctor directly via WhatsApp while I was in isolation due to COVID-19 to discuss my symptoms along with my autoimmune conditions. The consultation lasted 10 - 15 minutes and was free of charge.* [P5]

Telemedicine was also used for prescription renewals and noncritical follow-ups. Two patients receiving care at specialized autoimmune clinics reported being notified via WhatsApp when their condition was stable, skipping the usual in-person appointment. Interestingly, none of the patients used commercial health-tech apps (eg, Halodoc), preferring direct contact with trusted physicians. As 1 patient explained:

> *I already know the doctor, so I just message them directly.* [P2]

In remote regions where specialist access remains sparse and health literacy is rising, patients have begun turning to social media for medical advice. This noted the role of social media in improving access to medical information and direct physician interaction [66,67].

> *I asked him on Instagram or TikTok, "Doctor, is there an alternative to Myfortic because it's hard to find?"* [P1]

Together, these findings highlight how geography, infrastructure, and personal networks shape health care access and telemedicine use. Patients in remote areas, particularly those with long-term physician relationships, increasingly rely on digital tools to overcome systemic barriers to consistent and affordable care.

Cost-Effectiveness

Telemedicine significantly reduced financial burdens for patients requiring frequent monitoring. Of the 9 participants, 6 reported measurable cost savings after switching from monthly in-person visits to a hybrid model with 1 in-person visit and 5 virtual consultations over 6 months.

XSL•FO
RenderX

At the time of the study, the exchange rate was approximately €1=17,400 IDR (US $1=IDR 15,500 [2024 average]) [77,78]. The average cost per in-person visit was IDR 2 million (€115), with transportation and accommodation expenses increasing the total cost substantially for remote patients. In comparison, teleconsultations conducted via WhatsApp were free. Table 2 below summarizes the financial impact, as reported by multiple participants in this study.

These results reflect a cost reduction of 65%, from IDR 42.0 million (€2484) to IDR 16.5 million (€869), corresponding to an absolute savings of IDR 25.5 million (€1615), without compromising care quality for stable patients. However, public insurance (BPJS) does not fully cover all lab tests or medications. Out-of-pocket costs varied from IDR 200,000 (€12) to IDR 7 million (€405) per month, depending on the disease progression. One patient noted paying IDR 7 million monthly during the first year post diagnosis due to uncovered medications.

Hence, telemedicine offers significant financial and logistical relief for patients in eastern Indonesia, reducing the need for costly and exhausting multileg hospital trips. Thus, telemedicine is highly valued, not only for its economic advantages but also for its emotional and physical relief.

Table . Exemplary comparison of cost savings based on patient interviews.

Category	In-person visits	Hybrid model	Cost savings
Visit frequency (6 mo)	6 in-person visits (monthly)	1 in-person visit+5 telemedicine consultations	—[a]
Cost per in-person visit (IDR)[b]	2,000,000	2,000,000	0
Total in-person visit cost (IDR)	12,000,000 (IDR 2,000,000x6)	2,000,000 (1 visit)	10,000,000
Telemedicine consultation cost (IDR)	0	0 (via WhatsApp)[c]	0
Medication costs (IDR)	9,000,000 (IDR 1,500,000 x 6)	9,000,000 (IDR 1,500,000 x 6)	0
Transportation costs (IDR)	12,000,000 (round-trip for all visits)	2,000,000 (1 trip)	10,000,000
Additional costs (accommodation, meals) (IDR)	9,000,000 (overnight stays, meals)	1,500,000	7,500,000
Total cost (IDR)	42,000,000	16,500,000	25,500,000 saved

[a]Not applicable.

[b]All costs are reported in Indonesian Rupiah (IDR). Conversion to US dollars (US $) was performed using an exchange rate of US $1=IDR 15,500 (2024 average).

[c]Free, under specific conditions (eg, a close physician-patient relationship, typically established for at least 5 years since the initial diagnosis).

Discussion

Overview

This study examines the impact of telemedicine adoption on the management of chronic and rare diseases in Indonesia. Telemedicine has demonstrated significant potential in enhancing chronic illness care and has been recognized as a critical component of Indonesia's health care response during the COVID-19 pandemic [79-81]. By exploring the perspectives of both patients and health care professionals, this research offers valuable insights into the effectiveness, challenges, and opportunities of telemedicine for managing chronic and rare diseases in the postpandemic era.

Principal Findings and Alignment With Prior Literature

The results confirm existing theoretical insights into telemedicine adoption, particularly in Indonesia. The findings align with previous studies, which highlight both synchronous (real-time consultations) and asynchronous (store-and-forward) telemedicine models [16,18]. Physicians' reliance on WhatsApp, Zoom, and hospital-based platforms demonstrates the transition from minimal telemedicine adoption prepandemic to a more structured implementation during and post–COVID-19.

However, the results reveal that despite growing adoption, telemedicine is still not evenly distributed. This reflects reports that, although telemedicine adoption increased during the pandemic, its growth has since plateaued or declined in some regions [22,23]. The Indonesian data aligns particularly well with the trend observed in India, where slow adoption is due to remote infrastructure challenges.

From the perspectives of both patients and health care providers, telemedicine was associated with improved access to care, reduced travel time and costs, and was particularly useful for follow-ups and managing chronic diseases. However, shared concerns included clinical limitations, unreliable internet connections, and legal uncertainties. Referring to Figure 3, health care professionals benefit from expanded patient reach and improved remote monitoring but face limitations such as the inability to conduct physical examinations and increased administrative burdens. On the patient side, benefits like improved access to specialists and reduced travel are offset by challenges such as low digital literacy and poor internet access.

However, this analysis may not fully capture variations across rural Indonesia. Some areas, particularly those with digitally literate younger chronic patients and stable internet access, face physician shortages rather than technological barriers. In such cases, patients often turn to social media messaging platforms

XSL•FO
RenderX

to seek timely medical guidance from appropriate specialists, reflecting the evolving role of asynchronous telemedicine in bridging health care gaps [49,51,66]. Telemedicine is most effective and feasible for patients with high digital literacy, stable internet access, and low disease activity [43]. This further supports the trend in Indonesia, where telemedicine serves as a viable alternative for patients who meet these criteria but lack immediate access to physicians. Our study contributes that access to medical specialists is very limited outside the capital, and that telemedicine could significantly improve care across the country by improving access and lowering cost. Regarding chronic and rare disease management, the physicians' insights reflect the Asia-Pacific League of Associations for Rheumatology recommendations that telemedicine is effective for follow-ups but should not replace face-to-face visits for initial diagnoses [52]. This aligns with findings that emphasize the importance of physical examinations in clinical decision-making [49]. Moreover, the results echo research showing that over half of autoimmune patients prefer in-person consultations, particularly for first-time visits [43]. However, once a patient-physician relationship is established, they are more open to telemedicine follow-ups, reflecting evidence that patients prefer continuity with familiar physicians [15].

Figure 3. Summary of the benefits and limitations of telemedicine in Indonesia.

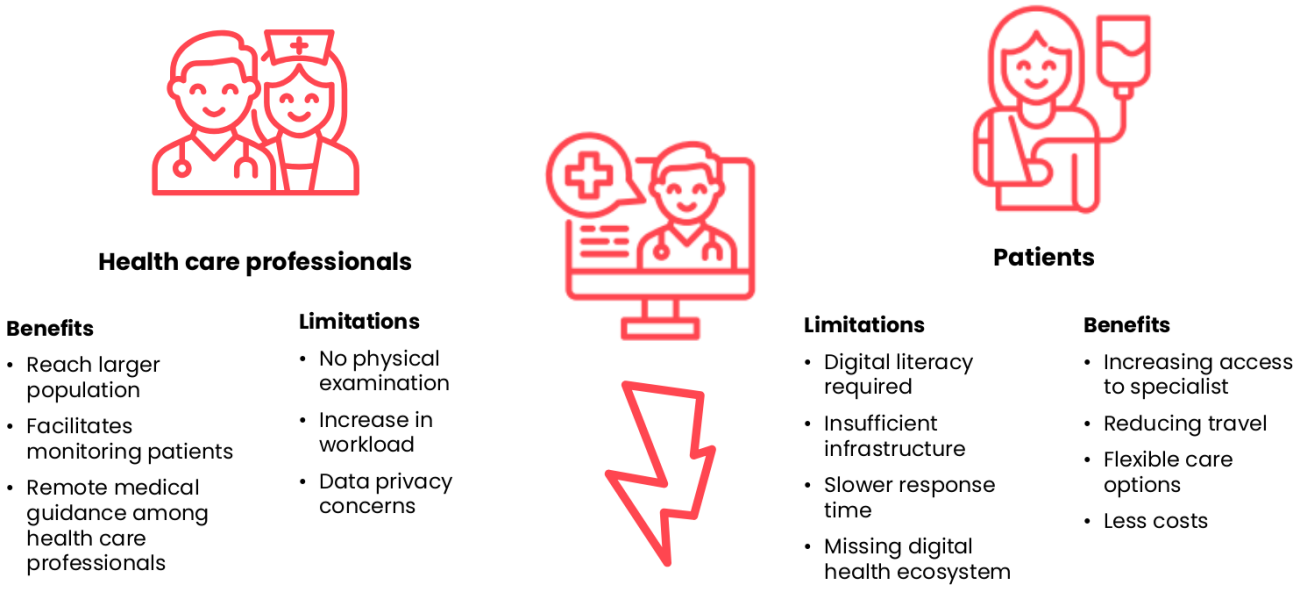

The cost-effectiveness of telemedicine in chronic disease management is also validated. Telemedicine has been shown to reduce travel costs and improve medication adherence, a pattern echoed in physicians' reports of its benefits for remote patients in Kalimantan and Sulawesi [45,47]. These results reinforce previous findings that telemedicine reduces costs for patients and providers, reducing inequalities in health care delivery and thus advancing SDGs 3 and 10. A systematic review of studies conducted in high-income countries further indicates that patients with chronic diseases show a measurable willingness to pay for telemedicine, ranging between 19% and 70% across studies, with higher willingness linked to greater distance from health facilities, while older patients were less willing to pay [82]. In Sweden, digital consultations led to a 41.5% reduction in health care expenses [54], a trend mirrored in Indonesia, where chat-based teleconsultations are considerably more affordable than face-to-face visits [83,84]. However, the study also highlights barriers to affordability. Unlike Sweden and France, where telemedicine consultations are reimbursed by statutory health insurances, Indonesia's system still relies on private insurance or out-of-pocket payments [55,85]. This limits access for low-income populations and supports concerns that telemedicine may exacerbate health care inequalities if not implemented equitably [49,86].

Beyond cost savings, telemedicine improves access to specialists for patients in remote areas who require consultations with rare disease specialists without geographical constraints [87,88]. This is evident in Indonesia's surgical conference model, where pediatric cardiologists conduct video consultations to assess and monitor children awaiting congenital heart disease surgery. Therefore, this study confirms existing challenges from the literature.

Practical Implications: Potential for Future Telemedicine Advancement

While telemedicine presents substantial benefits, physicians continue to recognize its limitations, particularly the necessity of physical examinations for certain medical fields such as rheumatology, neurology, and surgery [89]. Although effective for initial screenings and follow-ups, these specialties often require in-person assessments. Additionally, concerns have been raised about physician burnout, as virtual consultations expand patient reach beyond local boundaries and increase workload [59,90]. While the phenomenon of an increased workload and the resulting risk of burnout is not new to health care professionals, it certainly needs more attention in the context of telemedicine, from scholars and regulators alike. Similarly, concerns have been raised in the Association of Southeast Asian Nation-Japan Center report, where Indonesian health care providers express uncertainty over data security and ownership when using independent telemedicine platforms [91]. While unregulated platforms such as WhatsApp and Zoom are widely used at the moment (also in other health-related contexts such

XSL•FO
RenderX

as nutrition monitoring), it remains to be seen whether regulators will enforce other platforms in the future to increase privacy and security, while maintaining low entry barriers and an easy-to-use system [92]. Furthermore, infrastructural challenges—specifically unreliable electricity and limited internet connectivity in remote areas—continue to hinder telemedicine implementation in rural Indonesia [6,33].

Despite these challenges, the findings point to significant opportunities for advancing telemedicine. Physicians' vision for the development of telemedicine hubs across Indonesia aligns with Sweden's integrated telehealth model, which connects regional hospitals and primary care centers with specialized institutions to optimize care delivery [26]. Technological innovations such as VR can enhance remote diagnostics and treatment by offering immersive patient experiences, particularly beneficial in underserved regions [16,93]. Moreover, AI-assisted tools, including automated retinal screening for diabetic retinopathy, show promise in increasing diagnostic accuracy and early intervention. Integrating such tools into national health programs like Program Pengelolaan Penyakit Kronis could significantly strengthen chronic disease management, especially for diabetes, by facilitating proactive care and reducing complications [30].

Study Limitations and Directions for Future Research

While this study provides valuable insight into telemedicine implementation in a low-resource setting such as Indonesia, it is not without limitations. First, the sample size and scope were limited to a specific group of physicians and patients, which may not fully capture the diversity of Indonesia's health care landscape. Although our qualitative analysis did not reveal notable differences in telemedicine use based on gender or socioeconomic status within the small patient sample, we acknowledge that the composition of the sample was limited and skewed. Most of the participants were female and were predominantly insured by BPJS. Second, the study did not sufficiently account for regional disparities. Although adoption trends were discussed, variations in infrastructure quality, digital literacy, and health care accessibility between urban and rural areas were not comprehensively addressed. These disparities significantly influence the rate of telemedicine uptake and its overall effectiveness. Third, the study adopts a primarily short-term perspective, focusing on current patterns of telemedicine use without evaluating its long-term impact on patient outcomes, health care systems, or cost-efficiency. Longitudinal studies are needed to assess whether telemedicine can sustainably support health care delivery in Indonesia over time. At last, while the study acknowledges certain challenges related to technology and policy, it does not provide an in-depth analysis of regulatory environments, reimbursement structures, or technological readiness. These factors are crucial for understanding the scalability and sustainability of telemedicine services.

Future research should aim to expand sample diversity by including a wider range of health care professionals, patient populations, facility types, and geographic regions. In particular, studies with larger and more demographically diverse participant groups are needed to explore potential variations in telemedicine access, use, and user experience across gender, income, and insurance categories. This broader scope would enable a more representative and comprehensive understanding of telemedicine adoption across Indonesia. Additionally, long-term studies are necessary to evaluate the effects of telemedicine on patient health outcomes, cost-effectiveness, and physician workload. Exploring alternatives to in-person examinations, such as VR-based assessments, as trialed in Japan, may offer innovative solutions to address diagnostic challenges in remote settings. Furthermore, a detailed analysis of policy frameworks, legal regulations, data security protocols, and funding mechanisms is essential to assess their influence on telemedicine integration. Comparative studies with countries facing similar health care challenges could also provide valuable benchmarks and help identify best practices for more effective and equitable telemedicine implementation in Indonesia.

Conclusions

The findings of this study align with existing literature while providing additional insights into the regional disparities and infrastructure challenges related to telemedicine adoption in a major South-East Asian country. A noticeable shift in telemedicine usage occurred before, during, and after the COVID-19 pandemic. Prior to the pandemic, telemedicine adoption was minimal, with limited engagement from both physicians and patients. During the pandemic, there was a significant increase in the use of telemedicine platforms, as remote consultations became essential for managing patient volumes through tools like WhatsApp, Zoom, independent health technology applications, and hospital-based systems. Therefore, targeted investment in digital infrastructure and internet connectivity, particularly in underserved regions, would help reduce geographic inequities.

Postpandemic, the use of telemedicine for general consultations has decreased, though it continues to be important for follow-up care, the management of chronic and rare diseases, and improving access to specialists in remote areas, thereby supporting progress toward SDGs 3 and 10. While telemedicine has contributed to increased health care accessibility and cost-effectiveness, challenges remain, such as inadequate infrastructure, physician burnout, legal uncertainties, and the inability to conduct physical examinations remotely, which can hinder its impact on health equity. Hence, implementing support mechanisms for health care professionals, such as workload management systems and mental health progress, is crucial for long-term adoption.

Emerging models, such as telemedicine hubs, VR-based diagnostics, and AI-assisted screening, offer potential solutions to bridge these gaps and further promote equitable health care delivery in line with SDG 3 and SDG 10. However, the long-term success and broader adoption of telemedicine in Indonesia require sustainable regulatory frameworks, adequate funding models, consideration of the well-being of health care professionals, and continued investment in digital infrastructure. Addressing these challenges will be essential to maximize telemedicine's contribution to more inclusive and resilient health care systems that advance global commitments to SDG 3 and 10.

XSL•FO
RenderX

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) tools in the research and manuscript preparation process. In accordance with the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and language editing and translation of interview data. The GAI tools used were ChatGPT (versions 4o and 5; OpenAI) and DeepL (DeepL SE). The authors take full responsibility for the integrity, accuracy, and final content of the manuscript. GAI tools are not listed as authors and do not bear responsibility for the study's findings or conclusions. Declaration submitted by: CGP, RN, CH, and MG.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The datasets generated and/or analyzed during the current study are not publicly available due to privacy and confidentiality agreements with participants, but deidentified data are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: CGP, RN, CH, MG
Formal analysis: CGP, RN, MG
Investigation: CGP, RN
Methodology: CGP, RN, MG
Supervision: CH, MG
Validation: CGP, RN, MG
Writing – original draft: CGP, RN, MG
Writing – review & editing: CGP, RN, CH, MG

Conflicts of Interest

None declared.

Multimedia Appendix 1
Qualitative research design.
[DOCX File, 71 KB - jmir_v28i1e83462_app1.docx]

Checklist 1
COREQ checklist.
[DOCX File, 20 KB - jmir_v28i1e83462_app2.docx]

References

1. Amagai S, Vonesh E, Adams J, Luo Y. Closing the gap: addressing telehealth disparities across specialties in the sustained pandemic era. NPJ Digit Med 2024 Aug 21;7(1):217. [doi: 10.1038/s41746-024-01201-w] [Medline: 39164391]
2. Hodgkins M, Barron M, Jevaji S, Lloyd S. Physician requirements for adoption of telehealth following the SARS-CoV-2 pandemic. NPJ Digit Med 2021 Feb 9;4(1):19. [doi: 10.1038/s41746-021-00390-y] [Medline: 33564102]
3. Lytras MD, Housawi AA, Alsaywid BS, editors. Digital Transformation in Healthcare in Post-Covid-19 Times: Next Generation Technology Driven Personalized Medicine and Smart Healthcare: Academic Press; 2023. URL: https://www.sciencedirect.com/book/edited-volume/9780323983532/digital-transformation-in-healthcare-in-post-covid-19-times [accessed 2026-02-18]
4. Haleem A, Javaid M, Singh RP, Suman R. Telemedicine for healthcare: capabilities, features, barriers, and applications. Sens Int 2021;2:100117. [doi: 10.1016/j.sintl.2021.100117] [Medline: 34806053]
5. Implementing telemedicine services during COVID-19: guiding principles and considerations for a stepwise approach. WHO. 2020. URL: https://www.who.int/publications/i/item/WPR-DSE-2020-032 [accessed 2024-08-09]
6. WHO issues consolidated guide to running effective telemedicine services: when delivered effectively, telemedicine can broaden healthcare access. WHO. 2022. URL: https://www.who.int/news/item/10-11-2022-who-issues-new-guide-to-running-effective-telemedicine-services [accessed 2024-08-09]
7. Indonesia. IHME. 2024. URL: https://www.healthdata.org/research-analysis/health-by-location/profiles/indonesia [accessed 2024-08-10]

XSL•FO
RenderX

8. Digital health transformation and telehealth regulation in improving health access in indonesia. UNDP Indonesia YouTube page. 2022 Jan 9. URL: https://www.youtube.com/watch?v=fbNezaOdWLY [accessed 2026-02-28]
9. Bakibinga P, Kamande E, Kisia L, Omuya M, Matanda DJ, Kyobutungi C. Challenges and prospects for implementation of community health volunteers' digital health solutions in Kenya: a qualitative study. BMC Health Serv Res 2020 Sep 21;20(1):888. [doi: 10.1186/s12913-020-05711-7] [Medline: 32957958]
10. Danila MI, Gavigan K, Rivera E, et al. Patient perceptions and preferences regarding telemedicine for autoimmune rheumatic diseases care during the COVID‐19 pandemic. Arthritis Care Res (Hoboken) 2022 Jul;74(7):1049-1057. [doi: 10.1002/acr.24860]
11. Kaihlanen AM, Virtanen L, Buchert U, et al. Towards digital health equity - a qualitative study of the challenges experienced by vulnerable groups in using digital health services in the COVID-19 era. BMC Health Serv Res 2022 Feb 12;22(1):188. [doi: 10.1186/s12913-022-07584-4] [Medline: 35151302]
12. Consolidated telemedicine implementation guide. : World Health Organization; 2022 URL: https://iris.who.int/server/api/core/bitstreams/d119a69e-e56d-4c78-a2b9-52572e14217c/content [accessed 2026-02-19]
13. Goeldner M, Gehder S. Digital health applications (DiGAs) on a fast track: insights from a data-driven analysis of prescribable digital therapeutics in Germany from 2020 to mid-2024. J Med Internet Res 2024 Aug 29;26:e59013. [doi: 10.2196/59013] [Medline: 39208415]
14. Higgins C, Dunn E, Conrath D. Telemedicine: an historical perspective. Telecomm Pol 1984 Dec;8(4):307-313. [doi: 10.1016/0308-5961(84)90044-2]
15. Kruse CS, Molina-Nava A, Kapoor Y, Anerobi C, Maddukuri H. Analyzing the effect of telemedicine on domains of quality through facilitators and barriers to adoption: systematic review. J Med Internet Res 2023 Jan 5;25:e43601. [doi: 10.2196/43601] [Medline: 36602844]
16. Sugawara Y, Hirakawa Y, Iwagami M, et al. Issues in the adoption of online medical care: cross-sectional questionnaire survey. J Med Internet Res 2024 Nov 27;26:e68439. [doi: 10.2196/68439] [Medline: 39602799]
17. Ronquillo Y, Meyers A, Korvek SJ. Digital health. In: StatPearls: StatPearls Publishing; 2025. URL: https://www.ncbi.nlm.nih.gov/books/NBK470260/ [accessed 2026-02-19]
18. Eswaran H, Dawson L, Hernandez M, et al. Telehealth: current definitions and future trends. : The Rural Telehealth Evaluation Center; 2022 URL: https://idhi.uams.edu/rtec/wp-content/uploads/sites/4/2022/05/Telehealth-Definitions-Paper-06MAY2022-1.pdf [accessed 2026-02-19]
19. Haimi M, Gesser-Edelsburg A. Application and implementation of telehealth services designed for the elderly population during the COVID-19 pandemic: a systematic review. Health Informatics J 2022;28(1):14604582221075561. [doi: 10.1177/14604582221075561] [Medline: 35175881]
20. Online doctor consultations. Statista. 2024. URL: https://www.statista.com/outlook/hmo/digital-health/online-doctor-consultations/indonesia [accessed 2026-02-19]
21. Osman S, Churruca K, Ellis LA, Luo D, Braithwaite J. The unintended consequences of telehealth in Australia: critical interpretive synthesis. J Med Internet Res 2024 Aug 27;26:e57848. [doi: 10.2196/57848] [Medline: 39190446]
22. Telehealth adoption rate in the Asia-Pacific region in 2019 and 2021, with a forecast for 2024, by selected country. Statista. 2024. URL: https://www.statista.com/statistics/1294089/apac-telehealth-adoption-rate-by-country/ [accessed 2026-02-19]
23. Kapur V, Boulton A, Angus D, Sukhrani D. Asia-pacific telemedicine is here to stay. Bain & Company. 2024. URL: https://www.bain.com/insights/asia-pacific-telemedicine-is-here-to-stay-snap-chart [accessed 2026-02-19]
24. Keelara R, Sutherland E, Almyranti M. Beyond the pandemic: leading practices for the future of telemedicine. : OECD; 2025 URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/01/beyond-the-pandemic_0b39f409/98863928-en.pdf [accessed 2026-02-19]
25. Sweden: country health profile 2023. : OECD Publishing; 2023 URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/12/sweden-country-health-profile-2023_1f0c8967/ec938a6d-en.pdf [accessed 2026-02-19]
26. The digital transformation of healthcare. Health by Sweden. 2025. URL: https://www.healthbysweden.com/the-digital-transformation-of-healthcare/ [accessed 2025-02-05]
27. Chittim G, Pappas A, Bomba J. The changing fortunes of telemedicine in Europe – past, present, and future beyond COVID-19. Health Advances. 2020. URL: https://healthadvances.com/insights/blog/the-changing-fortunes-of-telemedicine-in-europe [accessed 2025-02-15]
28. Messal H, Müller T, Richter L, Silberzahn T. Germany's eHealth transformation makes good but uneven progress. McKinsey & Company. 2022. URL: https://www.mckinsey.com/industries/life-sciences/our-insights/germanys-ehealth-transformation-makes-good-but-uneven-progress [accessed 2025-02-01]
29. Beratung am Bildschirm: videosprechstunden werden wieder deutlich weniger genutzt [Web page in German]. Welt. 2024. URL: https://www.welt.de/regionales/hamburg/article252891742/Beratung-am-Bildschirm-Videosprechstunden-werden-wieder-deutlich-weniger-genutzt.html [accessed 2025-02-05]
30. The COVID-19 pandemic and the future of telemedicine. : OECD; 2023 URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/01/the-covid-19-pandemic-and-the-future-of-telemedicine_1c878192/ac8b0a27-en.pdf [accessed 2026-02-19]

XSL•FO
RenderX

31. Holl F, Kircher J, Hertelendy AJ, Sukums F, Swoboda W. Tanzania's and Germany's digital health strategies and their consistency with the World Health Organization's global strategy on digital health 2020-2025: comparative policy analysis. J Med Internet Res 2024 Mar 18;26:e52150. [doi: 10.2196/52150] [Medline: 38498021]
32. Omboni S, Padwal RS, Alessa T, et al. The worldwide impact of telemedicine during COVID-19: current evidence and recommendations for the future. Connect Health 2022 Jan 4;1:7-35. [doi: 10.20517/ch.2021.03] [Medline: 35233563]
33. Davalbhakta S, Advani S, Kumar S, et al. A systematic review of smartphone applications available for coronavirus disease 2019 (COVID-19) and the assessment of their quality using the Mobile Application Rating Scale (MARS). J Med Syst 2020 Aug 10;44(9):164. [doi: 10.1007/s10916-020-01633-3] [Medline: 32779002]
34. Dorsey ER, Topol EJ. Telemedicine 2020 and the next decade. Lancet 2020 Mar 14;395(10227):859. [doi: 10.1016/S0140-6736(20)30424-4] [Medline: 32171399]
35. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nat Biomed Eng 2021 Jun;5(6):493-497. [doi: 10.1038/s41551-021-00751-8] [Medline: 34131324]
36. Indonesia population. World Population Review. 2024. URL: https://worldpopulationreview.com/countries/indonesia-population [accessed 2024-07-14]
37. Indonesia launches a blueprint on digital health to expand inclusive health care coverage. UNDP Indonesia. 2021. URL: https://www.undp.org/indonesia/press-releases/indonesia-launches-blueprint-digital-health-expand-inclusive-health-care-coverage [accessed 2026-02-19]
38. Digitalizing Indonesia's health sector, a critical step towards SDGs achievement. UNDP Indonesia. 2021. URL: https://www.undp.org/indonesia/news/digitalizing-indonesias-health-sector-critical-step-towards-sdgs-achievement#:~:text=Indonesia's%20Ministry%20of%20Health%20mustf%2Dneeded%20herd%20immunity%20level [accessed 2026-02-19]
39. Wijaya JH, Octavius GS, Hwei LRY. A literature review of telemedicine in Indonesia: past, present, and future prospective. J Administr Kesehat Indones 2022;10(2):261-272. [doi: 10.20473/jaki.v10i2.2022.261-272]
40. Bodenheimer T, Wagner EH, Grumbach K. Improving primary care for patients with chronic illness. JAMA 2002 Oct 9;288(14):1775-1779. [doi: 10.1001/jama.288.14.1775] [Medline: 12365965]
41. Bashshur RL, Shannon GW, Smith BR, et al. The empirical foundations of telemedicine interventions for chronic disease management. Telemed J E Health 2014 Sep;20(9):769-800. [doi: 10.1089/tmj.2014.9981] [Medline: 24968105]
42. Classification of rare diseases. Orphanet. URL: https://www.orpha.net/en/disease/classification [accessed 2025-05-20]
43. Jackson LE, Edgil TA, Hill B, et al. Telemedicine in rheumatology care: a systematic review. Semin Arthritis Rheum 2022 Oct;56:152045. [doi: 10.1016/j.semarthrit.2022.152045] [Medline: 35843158]
44. Riaz MM, Mahmood SBZ, Nasir N. Telemedicine for rheumatology patients in COVID-19 pandemic: perspectives of patients. J Patient Exp 2022;9:23743735221092635. [doi: 10.1177/23743735221092635] [Medline: 35441090]
45. Patel KB, Turner K, Alishahi Tabriz A, et al. Estimated indirect cost savings of using telehealth among nonelderly patients with cancer. JAMA Netw Open 2023 Jan 3;6(1):e2250211. [doi: 10.1001/jamanetworkopen.2022.50211] [Medline: 36626174]
46. Nappoe SA, Djasri H, Kurniawan MF. Chronic Disease Management Programme (PROLANIS) in Indonesia. : World Health Organization; 2023 URL: https://iris.who.int/server/api/core/bitstreams/6546593d-9cd6-4dc5-b47b-4c6031b28c3f/content [accessed 2026-02-19]
47. Ma Y, Zhao C, Zhao Y, et al. Telemedicine application in patients with chronic disease: a systematic review and meta-analysis. BMC Med Inform Decis Mak 2022 Apr 19;22(1):105. [doi: 10.1186/s12911-022-01845-2] [Medline: 35440082]
48. Sharareh B, Schwarzkopf R. Effectiveness of telemedical applications in postoperative follow-up after total joint arthroplasty. J Arthroplasty 2014 May;29(5):918-922.e1. [doi: 10.1016/j.arth.2013.09.019] [Medline: 24342278]
49. Greenhalgh T, Vijayaraghavan S, Wherton J, et al. Virtual online consultations: advantages and limitations (VOCAL) study. BMJ Open 2016 Jan 29;6(1):e009388. [doi: 10.1136/bmjopen-2015-009388] [Medline: 26826147]
50. Choi NG, Hegel MT, Marti N, Marinucci ML, Sirrianni L, Bruce ML. Telehealth problem-solving therapy for depressed low-income homebound older adults. Am J Geriatr Psychiatry 2014 Mar;22(3):263-271. [doi: 10.1097/JGP.0b013e318266b356] [Medline: 23567376]
51. Nelson LA, Spieker A, Greevy R, LeStourgeon LM, Wallston KA, Mayberry LS. User engagement among diverse adults in a 12-month text message–delivered diabetes support intervention: results from a randomized controlled trial. JMIR Mhealth Uhealth 2020 Jul 21;8(7):e17534. [doi: 10.2196/17534] [Medline: 32706738]
52. Ahmed S, Grainger R, Santosa A, et al. APLAR recommendations on the practice of telemedicine in rheumatology. Int J Rheum Dis 2022 Mar;25(3):247-258. [doi: 10.1111/1756-185X.14286] [Medline: 35043576]
53. Market study on telemedicine. : European Commission; 2018 URL: https://health.ec.europa.eu/system/files/2019-08/2018_provision_marketstudy_telemedicine_en_0.pdf [accessed 2026-02-19]
54. Ekman B. Cost analysis of a digital health care model in Sweden. Pharmacoecon Open 2018 Sep;2(3):347-354. [doi: 10.1007/s41669-017-0059-7] [Medline: 29623633]
55. Mounié M, Costa N, Gourdy P, et al. Cost-effectiveness evaluation of a remote monitoring programme including lifestyle education software in type 2 diabetes: results of the Educ@dom study. Diabetes Ther 2022 Apr;13(4):693-708. [doi: 10.1007/s13300-022-01207-1] [Medline: 35133640]

XSL•FO
RenderX

56. Sorwar G, Rahamn MM, Uddin R, Hoque MR. Cost and time effectiveness analysis of a telemedicine service in Bangladesh. Stud Health Technol Inform 2016;231:127-134. [Medline: 27782024]
57. Key Contacts in Indonesia. DLA Piper. 2021. URL: https://www.dlapiperintelligence.com/telehealth/countries/index.html?c=ID&t=contacts-section [accessed 2025-02-05]
58. Hall AK, Cole-Lewis H, Bernhardt JM. Mobile text messaging for health: a systematic review of reviews. Annu Rev Public Health 2015 Mar 18;36:393-415. [doi: 10.1146/annurev-publhealth-031914-122855] [Medline: 25785892]
59. Kruse CS, Krowski N, Rodriguez B, Tran L, Vela J, Brooks M. Telehealth and patient satisfaction: a systematic review and narrative analysis. BMJ Open 2017 Aug 3;7(8):e016242. [doi: 10.1136/bmjopen-2017-016242] [Medline: 28775188]
60. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
61. Donnelly C, Janssen A, Shah K, Harnett P, Vinod S, Shaw TJ. Qualitative study of international key informants' perspectives on the current and future state of healthcare quality measurement and feedback. BMJ Open 2023 Jun 7;13(6):e073697. [doi: 10.1136/bmjopen-2023-073697] [Medline: 37286326]
62. Kallio H, Pietilä AM, Johnson M, Kangasniemi M. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. J Adv Nurs 2016 Dec;72(12):2954-2965. [doi: 10.1111/jan.13031] [Medline: 27221824]
63. Khanassov V, Ilali M, Ruiz AS, Rojas-Rozo L, Sourial R. Telemedicine in primary care of older adults: a qualitative study. BMC Prim Care 2024 Jul 17;25(1):259. [doi: 10.1186/s12875-024-02518-x] [Medline: 39020277]
64. Harapan BN, Harapan T, Theodora L, Anantama NA. From archipelago to pandemic battleground: unveiling Indonesia's COVID-19 crisis. J Epidemiol Glob Health 2023 Dec;13(4):591-603. [doi: 10.1007/s44197-023-00148-7] [Medline: 37707715]
65. Pathak L, Hernandez-Ramos R, Rosales K, et al. Using social media to recruit a diverse sample of participants for a mobile health (mHealth) intervention to increase physical activity: exploratory study. JMIR Mhealth Uhealth 2025 Apr 28;13:e56329. [doi: 10.2196/56329] [Medline: 40294424]
66. Darmawan I, Bakker C, Brockman TA, Patten CA, Eder M. The role of social media in enhancing clinical trial recruitment: scoping review. J Med Internet Res 2020 Oct 26;22(10):e22810. [doi: 10.2196/22810] [Medline: 33104015]
67. Uy L, Safwan N, Ahmad R, et al. Research recruitment strategies in the digital era: opportunities and challenges of social media. Contemp Clin Trials 2025 Sep;156:108033. [doi: 10.1016/j.cct.2025.108033] [Medline: 40738218]
68. Topolovec-Vranic J, Natarajan K. The use of social media in recruitment for medical research studies: a scoping review. J Med Internet Res 2016 Nov 7;18(11):e286. [doi: 10.2196/jmir.5698] [Medline: 27821383]
69. Angum F, Khan T, Kaler J, Siddiqui L, Hussain A. The prevalence of autoimmune disorders in women: a narrative review. Cureus 2020 May 13;12(5):e8094. [doi: 10.7759/cureus.8094] [Medline: 32542149]
70. Statistik Telekomunikasi Indonesia 2024 [Article in Indonesian]. : Badan Pusat Statistik Indonesia; 2025 URL: https://tinyurl.com/mrxedn4j [accessed 2026-02-19]
71. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
72. Rädiker S, Kuckartz U. Focused Analysis of Qualitative Interviews with MAXQDA: Step by Step: Maxqda Press; 2020. [doi: 10.36192/978-3-948768072]
73. Gehder S, Goeldner M. Unpacking performance factors of innovation systems and studying Germany's attempt to foster the role of the patient through a market access pathway for digital health applications (DiGAs): exploratory mixed methods study. J Med Internet Res 2025 Jan 6;27:e66356. [doi: 10.2196/66356] [Medline: 39761562]
74. Peraturan menteri kesehatan nomor 340 tahun 2020 tentang klasifikasi rumah sakit [Article in Indonesian]. : Kementerian Kesehatan Republik Indonesia; 2010 URL: https://pelayanan.jakarta.go.id/download/regulasi/peraturan-menteri-kesehatan-nomor-340-tentang-klasifikasi-rumah-sakit.pdf [accessed 2026-02-19]
75. The republic of Indonesia health system review. WHO. 2017. URL: https://iris.who.int/bitstream/handle/10665/254716/9789290225164-eng [accessed 2025-03-01]
76. Sungkono KR, Sarno R. Determine types of Indonesian hospital by criteria-based proses model, k-means cluster, and hierarchical average linkage. Presented at: 2019 International Seminar on Application for Technology of Information and Communication (iSemantic); Sep 21-22, 2019. [doi: 10.1109/ISEMANTIC.2019.8884299]
77. JISDOR. Bank Indonesia. URL: https://www.bi.go.id/id/statistik/informasi-kurs/jisdor/Default.aspx [accessed 2024-12-31]
78. Euro foreign exchange reference rates. European Central Bank. 2024. URL: https://www.ecb.europa.eu/stats/policy_and_exchange_rates/euro_reference_exchange_rates/html/index.en.html [accessed 2024-12-31]
79. Indonesia's telemedicine industry one of fastest-growing: minister. Antara News. URL: https://en.antaranews.com/news/291951/indonesias-telemedicine-industry-one-of-fastest-growing-minister [accessed 2024-08-11]
80. Corbett JA, Opladen JM, Bisognano JD. Telemedicine can revolutionize the treatment of chronic disease. Int J Cardiol Hypertens 2020 Dec;7:100051. [doi: 10.1016/j.ijchy.2020.100051] [Medline: 33330846]
81. Shaqeel M. President Jokowi gives appreciation for the contribution of various parties in handling the COVID-19 pandemic. Industry.co.id. 2023. URL: https://en.industry.co.id/read/8067/

XSL•FO
RenderX

president-jokowi-gives-appreciation-for-the-contribution-of-various-parties-in-handling-the-covid-19-pandemic#google_vignette [accessed 2024-08-11]
82. Chua V, Koh JH, Koh CHG, Tyagi S. The willingness to pay for telemedicine among patients with chronic diseases: systematic review. J Med Internet Res 2022 Apr 13;24(4):e33372. [doi: 10.2196/33372] [Medline: 35416779]
83. Dokter umum - konsultasi online terpercaya 24 jam [Article in Indonesian]. Halodoc. 2020. URL: https://www.halodoc.com/tanya-dokter/kategori/dokter-umum [accessed 2025-02-23]
84. MySiloam. Siloam Hospitals. 2025. URL: https://www.siloamhospitals.com/en/mysiloam [accessed 2025-02-23]
85. Key Contacts in Sweden. DLA Piper. 2021. URL: https://www.dlapiperintelligence.com/telehealth/countries/index.html?c=SE&t=contacts-section [accessed 2025-02-25]
86. Yao R, Zhang W, Evans R, Cao G, Rui T, Shen L. Inequities in health care services caused by the adoption of digital health technologies: scoping review. J Med Internet Res 2022 Mar 21;24(3):e34144. [doi: 10.2196/34144] [Medline: 35311682]
87. Rahman FF. Indonesia's healthcare landscape: embracing innovation in the new health regime. Curr Med Res Opin 2024 May 6;2024:1-5. [doi: 10.1080/03007995.2024.2349732] [Medline: 38683759]
88. Bin KJ, Santana Alves PG, Costa R, et al. User experience regarding digital primary health care in Santarém, Amazon: evaluation of patient satisfaction and Doctor's feedback. JMIR Form Res 2023 Jan 11;7:e39034. [doi: 10.2196/39034] [Medline: 36630164]
89. Kaissar S, Zara S, Fabrice F, et al. Physicians' satisfaction with the use of teleconsultation in France. J Hosp Manag Health Policy 2023;7:1. [doi: 10.21037/jhmhp-22-76]
90. Mandal S, Wiesenfeld BM, Mann DM, Szerencsy AC, Iturrate E, Nov O. Quantifying the impact of telemedicine and patient medical advice request messages on physicians' work-outside-work. NPJ Digit Med 2024 Feb 14;7(1):35. [doi: 10.1038/s41746-024-01001-2] [Medline: 38355913]
91. Navallo K, Detros K. Assessing digital health adoption in ASEAN. : ASEAN-Japan Centre; 2024 URL: https://www.asean.or.jp/main-site/wp-content/uploads/2024/07/Assessing-Digital-Health-Adoption-in-ASEAN.pdf [accessed 2026-02-19]
92. Aisyah DN, Mayadewi CA, Utami A, et al. Using WhatsApp for nutrition surveillance among children under 5 years in West Java, Indonesia: cross-sectional survey and feasibility study. JMIR Pediatr Parent 2025 May 15;8:e58752. [doi: 10.2196/58752] [Medline: 40372943]
93. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. NPJ Digit Med 2023 Oct 9;6(1):186. [doi: 10.1038/s41746-023-00927-3] [Medline: 37813960]

Abbreviations

AI: artificial intelligence
BPJS: Badan Pengelola Jaminan Sosial Kesehatan (National Health Insurance Agency)
COREQ: Consolidated Criteria for Reporting Qualitative Research
GAI: generative artificial intelligence
NCD: noncommunicable disease
SDG: sustainable development goal
VR: virtual reality

Edited by A Stone; submitted 03.Sep.2025; peer-reviewed by A Rasool, IAP Tarquino; accepted 02.Feb.2026; published 19.Mar.2026.

Please cite as:
Pratama CG, Nurlianti R, Herstatt C, Goeldner M
Telemedicine Adoption for Managing Chronic and Rare Diseases in Indonesia During and Beyond the COVID-19 Era: Qualitative Study
J Med Internet Res 2026;28:e83462
URL: https://www.jmir.org/2026/1/e83462
doi:10.2196/83462

© Christine Gracia Pratama, Rima Nurlianti, Cornelius Herstatt, Moritz Goeldner. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Telehealth Intervention to Reduce Sedentary Behavior in Older Adults With Type 2 Diabetes: Development and Feasibility Study

Xiaoyan Zhang[1,2*], PhD; Dan Yang[3*], PhD; Sihan Chen[1], BSN; Lei Wang[4], BSN; Xin Li[1], BSN; Ziyan Wang[1], BSN; Meiqi Meng[1], MSN; Xuejing Li[1], PhD; Hongzhan Jiang[1], MSN; Yufang Hao[1], PhD; Jiayin Luo[2], MSN

[1]School of Nursing, Beijing University of Chinese Medicine, Liangxiang High Education Park, Fangshan District, Beijing, China

[2]Department of Vascular Surgery, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing, China

[3]Department of Nursing Administration, Fangshan Hospital Beijing University of Chinese Medicine, Beijing, China

[4]Department of Nursing, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Yufang Hao, PhD
School of Nursing, Beijing University of Chinese Medicine, Liangxiang High Education Park, Fangshan District, Beijing, China

Abstract

Background: Sedentary behavior (SB) is a modifiable risk factor for complications in older adults with type 2 diabetes mellitus (T2DM). Despite widespread adoption of digital health platforms, theory-driven telehealth interventions specifically targeting SB reduction remain limited, particularly those incorporating cultural adaptation and behavioral change frameworks.

Objective: This study aims to develop and evaluate the feasibility of a theory-based personalized telehealth intervention to reduce SB in older adults with T2DM in China.

Methods: The intervention was developed over 14 months (January 2022-February 2023) following the intervention mapping and Behavior Change Wheel frameworks. A panel of 19 multidisciplinary experts (90.5% response rate) refined the program through a systematic iterative process. Subsequently, a 7-week quasi-experimental study (pre-post self-controlled design) was conducted to assess feasibility. We recruited 30 community-dwelling older adults with T2DM via WeChat-based convenience sampling. The primary outcome was SB measured by the Measure of Older Adults' Sedentary Time for Type 2 Diabetes Mellitus questionnaire. Secondary outcomes included cardiovascular risk (blood pressure), glycemic control (fasting blood glucose), Diabetes-Specific Quality of Life, social isolation, BMI, and fall incidence. Pre-post changes from baseline to 7 weeks were statistically evaluated to assess the intervention's feasibility and preliminary impact.

Results: The intervention comprises 5 components: an eHealth education manual, a motion graphics library, an SMS text messaging library, a WeChat Q&A group, and a material incentive package. These components address "knowledge," "social support," and "intention" determinants through "education," "enablement," and "incentivisation" functions, respectively. All components used the "service provision" policy and various behavior change techniques. Preliminary feasibility testing (n=31) showed reduced sedentary time by 1.12 hours/day (P<.001) and improved social connectivity scores (P=.001).

Conclusions: This study demonstrates the feasibility and potential impact of a systematically developed telehealth intervention for reducing SB in older adults with T2DM in China. The integration of intervention mapping with the Behavior Change Wheel provides a replicable framework for developing theory-driven digital health interventions. With significant reductions in sedentary time and improved social connectivity, this culturally adapted approach offers a scalable model for chronic disease self-management in aging populations. The systematic methodology and positive preliminary outcomes support further large-scale evaluation of evidence-based telehealth solutions for behavioral modification in diabetes care.

(J Med Internet Res 2026;28:e80827) doi:10.2196/80827

KEYWORDS

intervention mapping; behavior change intervention; sedentary behavior; older adults; mixed methods; telehealth interventions

Introduction

Type 2 diabetes mellitus (T2DM) represents a global health challenge characterized by insufficient insulin secretion and/or dysfunction [1]. The International Diabetes Federation reported that diabetes prevalence among older adults (aged ≥65 y) reached 19.3% globally in 2020, with projections indicating an increase to 27.6% by 2045 [2]. This trend is particularly

XSL•FO
RenderX

pronounced in upper-middle-income countries (UMICs), where age-standardized mortality rates remain substantially elevated [3]. China exemplifies this challenge with a T2DM prevalence of 30.2% among its older population [4]. Given the profound impact, lifestyle modifications offer a potentially cost-effective and sustainable method for improving T2DM outcomes [5], with sedentary behavior (SB) emerging as a critical modifiable risk factor.

SB is defined as any waking behavior characterized by an energy expenditure of 1.5 metabolic equivalents or less while sitting, reclining, or lying [6]. Distinct from physical inactivity, adverse health effects may persist even when recommended activity levels are met [7]. Globally, 67% of adults aged 60 years and older spend over 8.5 hours daily in SB [8], with 69.8% of Chinese older adults with diabetes classified as sedentary [9]. Extensive research confirms SB's adverse impacts, including increased all-cause mortality [10], cardiovascular disease [11], cancer [12], and metabolic disorders [13-16]. SB represents a significant modifiable risk factor requiring targeted intervention in T2DM management.

While the World Health Organization (WHO) and various countries' guidelines recommend reducing sedentary time [17-19], promoting physical activity alone proves insufficient [20,21]. Current SB interventions face challenges: moderate activity often replaces light activity rather than sedentary time [22], and recommended activity levels may paradoxically increase compensatory SB [23]. Notably, a 2021 Cochrane systematic review revealed no statistically significant effects of current SB interventions for community-dwelling older adults [24]. As the global chronic disease burden increases, telemedicine emerges as a crucial tool for providing accessible care [25]. While small-scale studies suggest potential efficacy [26], patient-oriented behavioral interventions remain fragmented—a challenge prevalent globally in digitally advancing health care systems [27,28]. There is a pressing need for theory-based, culturally adapted approaches that systematically integrate theoretical frameworks into remote SB intervention design.

Traditional implementation strategy development often relies on the "It Seemed Like a Good Idea at the Time" principle, targeting barriers without theoretical foundation, thereby compromising outcome interpretation and failing to elucidate mechanisms of action [29]. The UK Medical Research Council recommends systematic, theoretically grounded approaches [29]. Implementation science has emerged as a discipline exploring how to systematically integrate evidence-based interventions into routine practice [30-32]. To operationalize these principles, the intervention mapping (IM) framework [33,34] provides a structured, theory-driven approach that constructs logical problem models, identifies behavioral determinants, and matches them with behavioral objectives. Successfully applied in SB reduction interventions across diverse populations [35-38], IM ensures systematic development of theoretically grounded strategies.

Our foundational work involved completing step 1 of the IM framework through previously published research [39]. This foundational phase used a 2-phase sequential explanatory mixed methods design based on the Behavior Change Wheel (BCW). Initially, we used an established conceptual framework (Figure 1) to conduct a cross-sectional survey, which validated key SB patterns and determinants [39]. Subsequently, qualitative interviews were conducted to enrich the model with patient perspectives and identify contextual influencing factors (X Zhang et al, PhD, unpublished data, August 2025). Through the integration of these findings, we finalized the SB logic model (Figure 2), which systematically delineates 20 modifiable determinants across capability, motivation, and opportunity dimensions, as well as a 3-stage behavioral trajectory: multipoint triggering, immersive maintenance, and transitional breaking contexts. This logic model (Figure 2) serves as the theoretical and empirical foundation for the systematic development and testing of the "Double-S" intervention in this study.

XSL•FO
RenderX

Figure 1. Preliminary conceptual framework of sedentary behavior for older adults with type 2 diabetes mellitus (T2DM). This framework was developed prior to empirical data collection by integrating the capability, opportunity, motivation, and behavior model, the theoretical domains framework, and existing literature. The model depicts a logical flow from potential determinants mapped onto COM-B dimensions on the left, through various behavioral contexts in the center, to potential health outcomes on the right. This initial framework was subsequently validated and refined through 2-phase empirical studies ([39]; X Zhang et al, PhD, unpublished data, August 2025), leading to the finalized logic model presented in Figure 2.

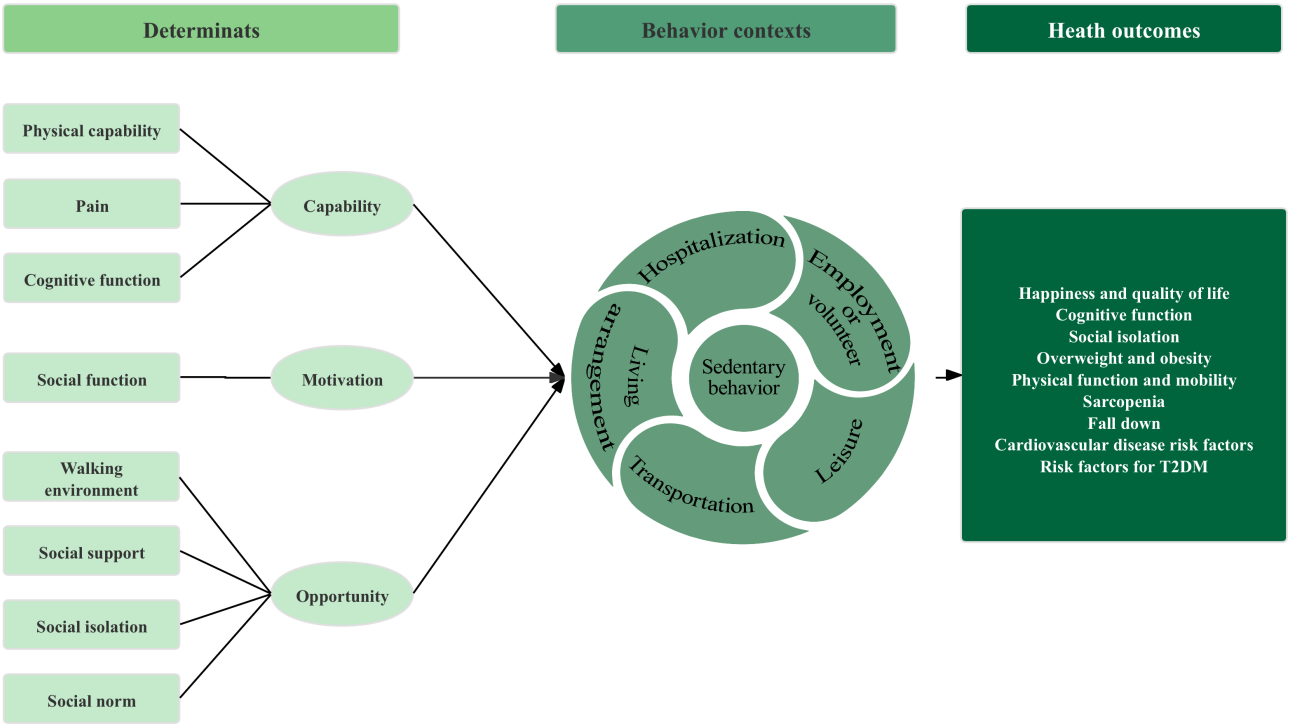

Figure 2. Logical model of sedentary behavior in older adults with type 2 diabetes mellitus (T2DM).

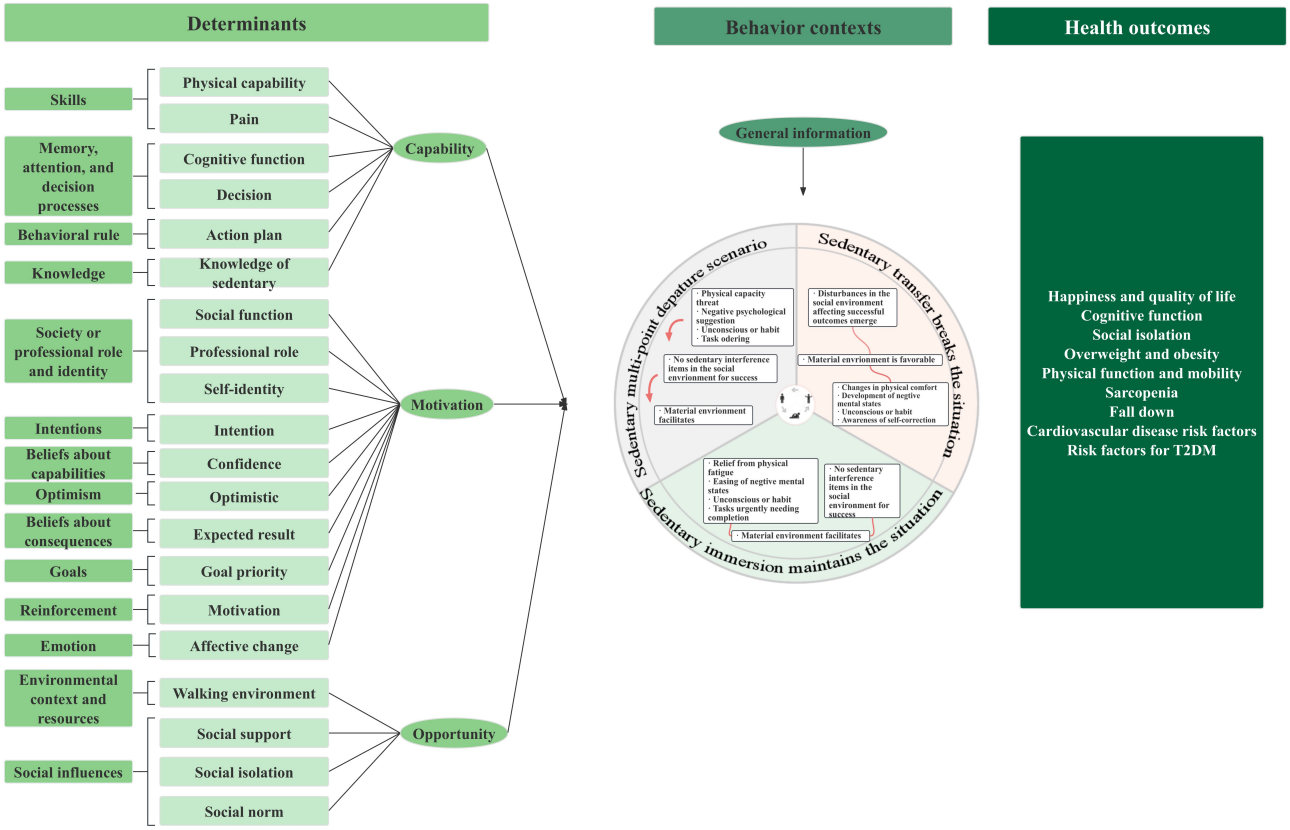

Building on these foundational insights, this study aims to translate identified behavioral determinants into a practical, technology-enabled application. We introduced the theory-driven remote intervention program “Double-S” (stop SB, stand up), designed to specifically address the identified determinants and reduce SB among community-dwelling older

XSL•FO
RenderX

adults with T2DM. This research sought to bridge the gap between theoretical frameworks and practical telehealth delivery by addressing the following specific research questions: (1) How can a multicomponent telehealth intervention be systematically developed for older Chinese adults with T2DM using the IM [33,34,40] and BCW [41] frameworks? (2) Is the resulting "Double-S" program feasible and acceptable for this population, and what are its preliminary impacts on behavioral and health-related outcomes?

Methods

Study Design

This study used the IM framework to develop and evaluate a theory-based behavior change intervention. The process focused on IM steps 2 through 6, conducted between January 2022 and February 2023, spanning 14 months (Figure 3). Building on the logic model established in step 1 (Figure 2), which was previously completed and published [39], this study focuses on IM steps 2 through 6. A detailed summary of the sampling strategies, recruitment channels, and participant involvement (and the independence of participants) across all stages is provided in Multimedia Appendix 1. IM consists of six iterative steps: (1) completed previously [39], (2) formulating program outcomes and objectives, (3) designing the program, (4) producing program components and materials, (5) planning program implementation, and (6) planning for evaluation. Each step used core processes to identify key literature, apply theory, and collect additional data [40]. Experts were purposively recruited through professional networks and institutional recommendations between January 2022 and February 2022. The research team identified potential participants based on their expertise and extended personal invitations via email, accompanied by detailed program information. The panel comprised experts from various fields: implementation science, evidence-based practice, education, psychology, geriatric nursing, clinical nursing, and diabetes specialist nursing. Eligibility criteria required professionals to hold a relevant postgraduate degree (master's or higher) or possess over 10 years of specialized work experience. Of the 21 professionals identified and invited, 19 accepted the invitation (90.5% response rate). The recruitment was conducted through the research team's professional database and academic societies related to geriatric nursing in Beijing. No formal interviews were conducted; selection was based solely on professional credentials and alignment with the program's objectives.

Figure 3. Intervention mapping (IM) process and study timeline for the "Double-S" program. APEASE: Acceptability, Practicability, Effectiveness, Affordability, Safety, Equity; BCW: Behavior Change Wheel; SB: sedentary behavior.

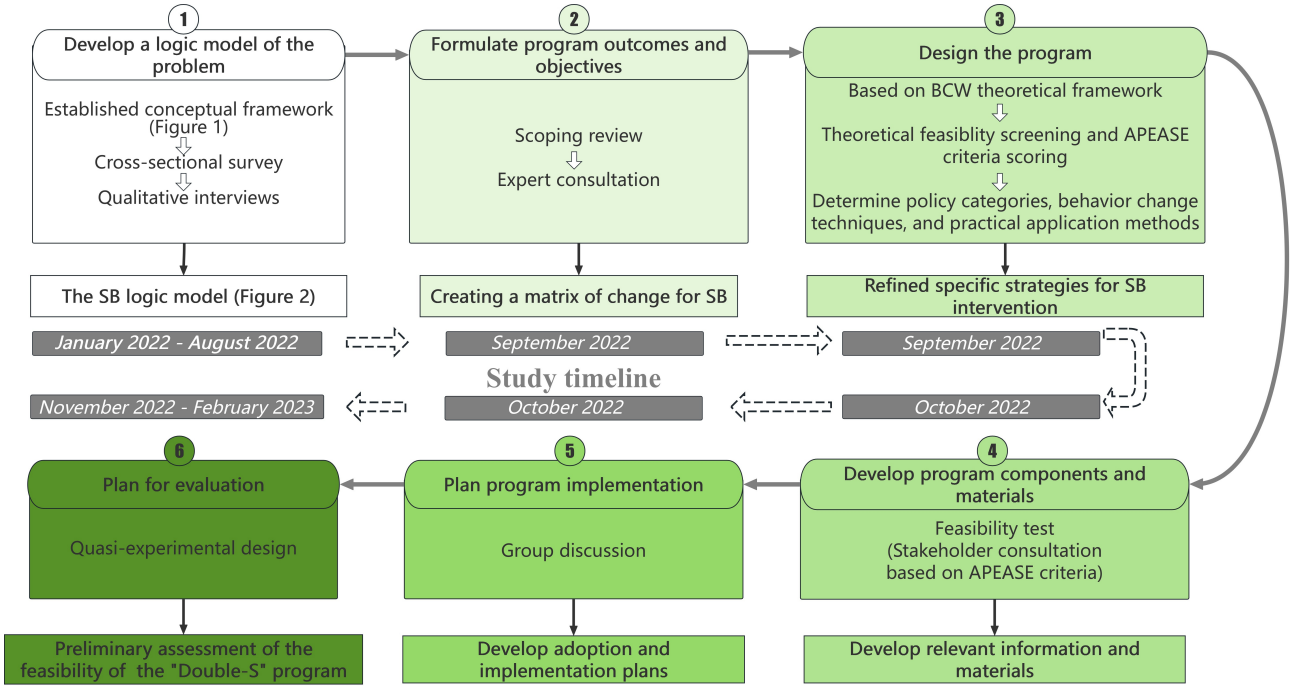

Ethical Considerations

The study protocol was approved by the ethics committee of Beijing University of Chinese Medicine (2022BZYLL0505). Digital or written informed consent was obtained from all participants prior to their enrollment. To protect participant privacy, all data were anonymized and deidentified during analysis. All participants (experts, stakeholders, and older adults) participated voluntarily and did not receive financial compensation for their involvement.

Determining Program Outcomes and Change Objectives

This phase aimed to establish behavioral outcomes for the SB intervention and articulate performance objectives (POs) for each outcome, specifying what participants must learn, do, or change to achieve particular results. The process encompassed three key steps. First, based on preliminary literature research, we defined expected behavioral change outcomes for SB intervention strategies by synthesizing evidence from multiple sources: (1) the Cochrane systematic review [24] examined the effectiveness of interventions to reduce sedentary time in adults

XSL•FO
RenderX

with type 2 diabetes, providing evidence on the feasibility and health benefits of reducing daily sedentary time; (2) our scoping review [42] systematically mapped existing evidence on SB interventions in older adults, complementing the disease-specific evidence from the Cochrane review with age-related considerations; and (3) the clinical practice guideline for exercise therapy of T2DM in China (2024 edition) [43] provided specific recommendations on the maximum duration of single sedentary bouts for patients with T2DM. These reviews complemented each other by providing both disease-specific evidence (population with T2DM) and age-specific considerations (older adults), enabling us to establish appropriate behavioral targets for our specific population of older adults with T2DM. For these expected outcomes, we set POs based on the previously developed 3-stage SB trajectory model (specifically, sub-behaviors of the target behavior). These POs specified what the study participants needed to learn, implement, or change to achieve the specific outcomes. Second, selecting important and modifiable behavioral determinants. Guided by the SB logic model, we developed an expert consultation questionnaire to systematically evaluate each determinant. This evaluation was conducted by the intervention strategy development group, consisting of a multidisciplinary panel of 19 experts with a mean professional or research experience of 7.39 (SD 6.94) years. The panel's expertise spanned various fields, including implementation science, evidence-based practice, education, psychology, geriatric nursing, community nursing, and specialized diabetes nursing. Members of the intervention strategy development group (n=19) independently rated each determinant on 2 dimensions using a 5-point Likert scale (1="very low" to 5="very high"): (1) importance—the extent to which the determinant influences SB in older adults, and (2) modifiability—the extent to which the determinant can be changed through intervention strategies. Determinants scoring 3 or more (indicating moderate to high levels) on both dimensions were selected as priority targets for intervention development. This scoring and threshold system was established to ensure that the resulting intervention would focus on determinants with the highest potential for meaningful behavior change and practical feasibility. Third, creating a matrix of change for SB. The intervention strategy group constructed a change matrix based on the screened determinants and established POs. This step aimed to operationalize objectives, enhancing their practicality and feasibility.

Program Design

Framework

This phase aimed to translate change objectives into practical intervention strategies using a theory-driven approach. We adopted the BCW [41] as our overarching framework, as it systematically integrates the Capability, Opportunity, Motivation, and Behavior (COM-B) model, the theoretical domains framework (TDF), and behavior change techniques (BCTs). The COM-B model explains behavior change mechanisms by analyzing the interplay of 3 factors: capability, opportunity, and motivation. Building on COM-B, the TDF provides a more detailed classification of behavioral determinants (Multimedia Appendix 2). BCTs are specific intervention measures designed to modulate behavioral determinants [44]. Based on this theoretical framework and its components, intervention strategies were developed through the following steps.

Identifying Intervention Functions

Based on the relationship between determinants identified by the TDF and intervention functions recommended by the BCW (Multimedia Appendix 3), we screened theoretically feasible intervention functions. Subsequently, the intervention development group applied the APEASE (Affordability, Practicability, Effectiveness/Cost-effectiveness, Acceptability, Side-Effects/Safety, Equity) criteria [45] to score the prescreened intervention functions. Functions with the highest scores were selected for strategy design. The APEASE scoring system ranged from 1 to 3, with higher scores indicating greater development potential for the intervention function (Multimedia Appendix 4).

Determining Policy Categories, BCTs, and Practical Application Methods

First, we identified theoretically applicable policy categories to support the selected intervention functions, primarily based on the BCW-recommended matching between intervention functions and policy categories (Multimedia Appendix 3). Subsequently, we referred to the results of our previous literature review [42] on SB interventions for older adults and the BCTs recommended in the BCW guide that correspond to specific intervention functions. Group members then determined the BCTs by mapping the selected intervention functions to the specific needs of older adults with T2DM, guided by both the BCW manual and findings from our previous literature review [42]. Finally, considering the identified intervention functions, policy categories, and BCTs, as well as the context and environment of intervention implementation, the intervention development group further refined specific strategies for SB intervention in community-dwelling older adults with T2DM.

Program Production

This phase aimed to refine the structure of intervention strategies and to draft, test for feasibility, and produce relevant information and materials. The intervention strategy group developed an initial draft of SB intervention strategies for community-dwelling older adults with T2DM, based on previously identified functions, policy categories, and BCTs. Throughout the iterative process, POs, change objectives, and practical strategies were continuously optimized.

To test the feasibility of strategies, a stakeholder consultation method was used. A panel of 9 stakeholders was recruited through purposive sampling, including experts in evidence-based medicine and nursing, specialists in education, clinicians in diabetes care, and patient representatives. These 9 stakeholders were independent of the 19 experts involved in step 2 to ensure an objective external review, although they shared similar eligibility criteria. The consultation was conducted through (eg, individual email-based or face-to-face sessions). Each stakeholder was provided with a detailed description of the proposed intervention and used a custom-designed APEASE scoring sheet to evaluate intervention strategies across 6 dimensions: acceptability, practicability, effectiveness or

XSL•FO
RenderX

cost-effectiveness, affordability, side effects or safety, and equity. Each dimension was rated on a 3-point Likert scale. Data were analyzed using descriptive statistical methods.

Program Implementation Plan

This phase focused on developing adoption and implementation plans in consultation with the intervention development group. Key stakeholders from evidence-based medicine or nursing, education, and diabetes care were engaged to discuss the integration of intervention strategies into existing services, as well as the logistics of delivery (how, when, and by whom). Training guidelines and curriculum plans were developed to equip community older people's health and diabetes management professionals with the tools to address the primary barriers contributing to SB in older adults with T2DM. In designing remote interventions, we specifically considered the resource constraints and technological infrastructure challenges common in UMICs. We opted for low-cost, easily implementable remote intervention solutions that do not require expensive equipment or complex technical support.

Evaluation Plan

This phase aimed to preliminarily assess the feasibility of the "Double-S" program for community-dwelling older adults with T2DM. A quasi-experimental design (pre-post self-controlled study) was used (registration: ER20220111).

The study population comprised community-dwelling older adults with T2DM. A convenience sampling method was used to recruit participants via WeChat in November 2022. Recruitment flyers were posted in 3 purposively selected WeChat groups of community-dwelling older adults and shared through the "Moments" functions of community health workers. Interested individuals contacted the research team directly via WeChat or phone. The researchers then screened these potential participants against the inclusion criteria and explained the study's purpose. Eligible candidates who provided informed consent were then purposively selected to ensure diversity in age and health status and formally enrolled for interviews. Inclusion criteria were: aged 60 years and older, diagnosed with T2DM, permanent community residents, ability to use WeChat on smartphones, and voluntary participation. Exclusion criteria included communication disorders, mental disorders, physical disabilities preventing standing, and comorbid severe illnesses. No formal sample size calculation was conducted due to the exploratory nature of this feasibility study; a target sample of 30 patients was set.

The "Double-S" program was delivered over a 7-week period using a WeChat-based telehealth approach. The intervention was facilitated by a trained research assistant and consisted of four primary components as follows:

- Digital education: an eHealth education manual, designed as a cloud-based document, was provided to participants for real-time access.
- Behavioral prompts: standardized reminders were selected from a predeveloped prompt library and sent to the WeChat group to encourage participants to reduce and break sedentary bouts.
- Visual learning: motion graphics (MG) animations were distributed through the WeChat group and a synchronized public video channel to demonstrate sedentary-breaking exercises.
- Interactive support: real-time Q&A sessions were conducted within the WeChat group to address participant concerns and provide peer support.

Participants underwent structured assessments at 2 specific time points: baseline (T0; week 0) and postintervention (T1; immediately following the completion of week 7). To ensure data accuracy and efficiency, all assessments were conducted online via the "Wenjuanxing" professional survey platform (similar to SurveyMonkey). The research team provided remote guidance to participants who required assistance in navigating the electronic questionnaires.

The intervention implementation adhered to the plan formulated in step 5, with assessment focusing on primary and secondary outcome measures. The primary outcome, SB, was measured using the Measure of Older Adults' Sedentary Time for T2DM Patients questionnaire [39]. Secondary outcomes encompassed a comprehensive range of health and psychosocial factors: cardiovascular disease risk (assessed by the most recent morning blood pressure readings), T2DM risk factors (evaluated using the most recent fasting blood glucose levels), quality of life (measured with the Diabetes-Specific Quality of Life scale) [46], social isolation (assessed using the Chinese version of the Lubben Social Network Scale-6 [LSNS-6]) [47,48], overweight and obesity (determined by BMI), and fall incidence (recorded as the number of individuals experiencing falls and their consequences within the past week).

Data were analyzed by SPSS Statistics v25.0 (IBM Corp). Descriptive statistics were used for continuous variables. Paired 2-tailed *t* tests or Wilcoxon signed-rank tests were used to compare preintervention and postintervention differences, depending on data distribution. The McNemar test was used for paired categorical variables. The significance level for all statistical tests was set at .05.

Results

Program Outcomes and Change Objectives

Behavioral Change Outcomes and POs

Based on the previously established logic model ([39]; X Zhang et al, PhD, unpublished data, August 2025), the intervention strategy group developed the following behavioral change outcome and POs. First, older adults with T2DM should reduce their average sedentary time by 1 hour per day (based on evidence synthesis from Cochrane systematic review [24] and our scoping review [42]), with a recommendation that single bouts of SB not exceed 30 minutes (based on the *Guideline for Exercise Therapy of T2DM in China, 2024 Edition* [43]). Second, to achieve the expected behavioral outcome, 4 POs were established based on the 3-stage trajectory of SB (Multimedia Appendix 5) as follows:

- PO1: patients should comprehend the risks and consequences of SB (prebehavior).

XSL•FO
RenderX

- PO2: patients select appropriate strategies to address SB (during behavior).
- PO3: patients implement reminders and interruptions during sedentary periods (during behavior).
- PO4: patients engage in appropriate physical activity to alleviate discomfort from SB (postbehavior).

Selection of Important and Modifiable Determinants of SB

Following the evaluation by the multidisciplinary panel described in the "Program Production" section, the assessment results (Table 1) identified knowledge, social influences, and intention as the priority TDF domains meeting both importance and modifiability criteria. These identified determinants were mapped directly to the TDF framework during the analysis. Further analysis within the social influences domain revealed that it encompassed social support, social isolation, and social norms. Considering feasibility constraints, social support was selected as the representative factor for social influence.

Table . Degree of support, importance, and modifiability scores for the determinants of sedentary behavior.

TDF[a] domains (behavioral determinants)	Literature review [49]	Cross-sectional study [39]	Qualitative research[b]	Importance, mean (SD)	Modifiability, mean (SD)
Knowledge	N[c]	N	Y[d]	4.00 (0.79)	4.37 (0.81)
Memory, attention, and decision processes	Y	Y	Y	4.74 (0.55)	2.68 (0.92)
Behavioral regulation	N	N	Y	4.58 (0.75)	2.84 (0.74)
Skills	N	Y	Y	4.00 (0.79)	2.79 (0.77)
Environmental context and resources	N	Y	Y	3.95 (1.05)	2.53 (1.09)
Social influences	N	Y	Y	3.63 (0.98)	3.47 (0.60)
Society or professional roles and identity	N	Y	Y	3.84 (0.87)	2.74 (0.85)
Intentions	N	N	Y	4.26 (0.91)	3.16 (0.49)
Beliefs about capabilities	N	N	Y	4.16 (0.67)	2.95 (0.76)
Optimism	N	N	Y	3.58 (1.09)	2.79 (0.83)
Beliefs about consequences	N	N	Y	3.47 (1.14)	2.74 (0.64)
Goals	N	N	Y	3.58 (1.04)	2.32 (0.46)
Reinforcement	N	N	Y	3.47 (0.60)	2.89 (0.64)
Emotion	N	N	Y	3.21 (0.77)	2.47 (0.60)

[a]TDF: theoretical domains framework.

[b]X Zhang et al, PhD, unpublished data, August 2025.

[c]N: not yet discovered.

[d]Y: support.

Change Matrix for SB

Based on the established POs and the selected determinants, specific change objectives were formulated to create a comprehensive change matrix, as shown in Table 2.

XSL•FO
RenderX

Table . Matrix of change objectives for sedentary behavior reduction: linking performance objectives (POs) with theoretical determinants.

PO	Determinants		
	Knowledge	Social support	Intentions
PO1: the patient should comprehend the risks and consequences of SB (before behavior).	The patient should comprehend the hazards of SB[a] (current situation, consequences, and its relationship with T2DM[b]).	Researchers or family members provide social support to the patient regarding PO1.	The patient expresses an intention to reduce SB.
PO2: the patient chooses appropriate strategies to cope with SB (during behavior).	The patient should comprehend strategies to counteract SB, such as reminders and behavior reversal.	Researchers or family members provide social support to the patient regarding PO2.	The patient expresses an intention to adopt appropriate coping strategies.
PO3: The patient is reminded and interrupted during SB (during behavior).	Family members should comprehend the hazards of SB (current situation, consequences, and its relationship with T2DM).	Researchers or family members provide social support to the patient regarding PO3.	Family members express an intention to monitor and remind the patient.
PO4：the patient engages in appropriate physical activities to alleviate discomfort from SB (after behavior).	The patient should comprehend how to engage in physical activity.	Researchers or family members provide social support to the patient regarding PO4.	The patient expresses an intention to engage in physical activity.

[a]SB: sedentary behavior.

[b]T2DM: type 2 diabetes mellitus.

Program Design

Selection of Intervention Functions

Intervention functions were evaluated using the APEASE criteria, with results presented in Multimedia Appendix 6. The BCW framework was used to align intervention functions with determinants: (1) knowledge: education (mean total score of 16.89, SD 1.10), (2) intentions: incentivization (mean highest score of 15.21, SD 1.08), and (3) social support: enablement (mean highest score of 15.10, SD 1.33).

Determination of Policy Categories, BCTs, and Practical Applications

The theoretical constructs were systematically translated into practical intervention strategies through integration of the following components: behavioral change outcomes, POs, determinants, change matrix, intervention functions, and policy categories. The resulting practical applications are comprehensively outlined in Multimedia Appendix 7.

Intervention Production

Draft Intervention Strategies

Through 2 rounds of expert panel discussions (n=19), the group determined that simple SMS text messaging reminders and social media–based support groups were particularly effective in reducing SB. These low-cost, highly acceptable intervention methods may be applicable in other resource-limited settings. Following two rounds of expert panel discussions involving 19 intervention group specialists, a comprehensive intervention strategy for SB in the community-dwelling older adults with T2DM was formulated. The strategy comprised 5 key components: eHealth education manual, MG animation library, SMS text messaging library, WeChat Q&A group, and material incentive package. The eHealth education manual, MG animation library, and SMS text messaging library targeted the "knowledge" determinant of SB. The WeChat Q&A group primarily addressed the "social support" determinant, while the material incentive package focused on the "intentions" determinant. Full descriptions of these intervention strategies are provided in Multimedia Appendix 8.

Feasibility Assessment of Intervention Strategies

Eight multidomain stakeholders (mean age of 44.75, SD 15.56 y) participated in the feasibility assessment, including 2 implementation science and evidence-based nursing or medicine stakeholders, 2 education stakeholders, 2 diabetes clinical nursing stakeholders, and 2 patient representatives. APEASE scoring results are presented in Multimedia Appendix 9. The eHealth education manual, MG animation library, SMS text messaging library, and material incentive package all achieved mean scores of 3.00 (SD 0.00), while the WeChat Q&A group scored 2.88 (SD 0.35).

Program Implementation Plan

Specific POs for intervention implementation are outlined in Table 3. These objectives were established by the research team based on the behavioral determinants identified in the logic model (Table 2). To achieve the adoption and implementation change objectives, the intervention design group selected a series of BCTs grounded in theory and integrated them with the 5 intervention components (Multimedia Appendix 10): eHealth education manual, MG animation library, SMS text messaging library, WeChat Q&A group, and material incentive package.

XSL•FO
RenderX

Table . Adoption and implementation outcomes and performance objectives (POs) for the intervention program.

PO	Context
PO1	The patient should comprehend the hazards of SB[a] (before behavior).
PO2	The patient should comprehend strategies to counteract SB, such as reminders and behavior reversal.
PO3	Family members should comprehend the hazards of SB (current situation, consequences, and its relationship with T2DM[b]).
PO4	The patient should comprehend how to engage in physical activity.

[a]SB: sedentary behavior.

[b]T2DM: type 2 diabetes mellitus.

Evaluation Results

A total of 31 community-dwelling older adults with T2DM (mean age 66.35, SD 3.94 years) completed the 7-week intervention. All 31 participants successfully completed both the baseline (T0) and postintervention (T1) assessments. Detailed participant characteristics are provided in Multimedia Appendix 11.

Following the implementation of the intervention, significant reductions in SB were observed among participants. Total sedentary time decreased by 1.12 hours per day compared to preintervention levels (P<.001). Notably, screen-based sedentary time showed a significant reduction of 0.33 hours per day (P=.01). However, no statistically significant changes were observed in sedentary time related to reading (P=.71), socializing (P=.87), transportation (P=.32), hobbies (P=.85), or other activities (P=.56).

The intervention also yielded improvements in social network scores as measured by the LSNS-6. Post-intervention LSNS-6 scores showed significant improvement (P<.001), with the family dimension demonstrating marked enhancement (P<.001). The friend dimension, however, did not show statistically significant changes (P=.46).

For participants with T2DM, no statistically significant changes were observed in morning blood pressure, fasting blood glucose, Diabetes-Specific Quality of Life scores, or BMI. No fall incidents were reported either before or during the intervention period.

Discussion

Principal Findings

This study aimed to develop and evaluate a theory-driven remote intervention strategy, “Double-S,” to reduce SB among older adults with T2DM in China. Our findings suggest that the systematic application of the IM framework is a feasible approach for developing targeted behavioral interventions. To our knowledge, this research represents one of the first theory-driven and evidence-based remote intervention strategies specifically targeting SB in older Chinese adults with T2DM in community settings. By integrating theoretical frameworks and addressing the specific needs of a rapidly aging population in a digitally evolving environment, this work contributes to the growing body of knowledge on culturally adapted, evidence-based interventions for chronic disease management in diverse global settings. Specifically, the intervention led to a promising reduction in total sedentary time by 1.12 hours per day and screen-based sedentary time by 0.33 hours per day. Furthermore, it significantly enhanced social network scores within the family dimension. However, these results should be interpreted as preliminary indicators of the program’s potential rather than definitive proof of efficacy, given the feasibility nature and limited sample size of this pilot evaluation.

Comparison With Prior Work and Implications

The core advantage of IM lies in its ability to explicitly link intervention strategies with theory, supporting hypothesis testing and deepening our understanding of causal mechanisms [33,50]. By addressing the gap in theory-grounded, culturally adapted digital health tools, our study offers a novel design contribution to the chronic disease management market in China and other similar settings. In diverse health care settings, IM is crucial for developing systematic approaches to use available resources and translate research findings into routine practice. Building on the BCW theory, this study constructed a problem logic model for SB among older adults with T2DM. The model integrates determinants (capability, motivation, and opportunity), the Chinese community context, the 3-stage SB trajectory, and health outcomes. By addressing this gap, our study offers new perspectives on improving behavior change interventions in diverse health care environments.

While IM has been widely applied in Western health care contexts [51-53], our research explores its applicability in Asian settings. Previous reviews [42] indicated that Western interventions predominantly target individual psychological aspects through face-to-face counseling or wearable devices.

However, SB is a habitual, unconscious behavior deeply influenced by cultural and environmental factors. The perception of SB and social norms varies across cultural contexts, affecting not only individual behavior but also health system design. Consequently, Western strategies need to be re-examined and adapted for Asian health care settings to ensure their effectiveness.

In community-based settings like China, high-cost individualized interventions pose challenges due to population aging and diabetes pressures [54]. Thus, our research shifted focus towards social and policy-level intervention strategies. Initial findings suggest that older adults in China often obtain health information

XSL•FO
RenderX

through public media, aligning with current science popularization policies. The high feasibility and acceptability scores from both multidisciplinary stakeholders and older users highlight significant opportunities for scaling up the "Double-S" program.

The widespread use of smartphones offers new channels for rapid health information dissemination. Remote interventions combined with digital tools provide advantages such as broad coverage and cost-effectiveness. Given the deep integration of digital devices into daily life, leveraging digital means for diabetes management has become increasingly necessary [55].

Beyond testing under controlled conditions, the preliminary evaluation results from both stakeholder and user perspectives provide a strong foundation for future implementation testing. This includes exploring how the "Double-S" program can be integrated into routine community health care systems for large-scale application. However, while the "Double-S" proved feasible under controlled conditions, real-world implementation may be constrained by structural barriers. This finding highlights potential challenges in transitioning from experimental settings to practical applications.

Digital infrastructure in rural areas and technological adaptability among older populations generally lag behind urban areas [56]. This can subject beneficiaries to "double exclusion" regarding technology access and service availability. Policymakers should prioritize bridging the gap between "technology popularization" and "institutional adaptation" rather than merely hardware upgrades.

Regarding effectiveness, the results showed significant improvements in social isolation but limited impact on physiological indicators like BMI or blood glucose. This aligns with findings from Ferguson et al [57] and Chastin et al [24], who indicated that metabolic improvements from SB reduction are typically modest and nonsignificant in the short term.

These findings reveal the dual nature of remote interventions: while cost-effective in enhancing social support networks, they may face challenges in reversing long-term metabolic disorders quickly. Heterogeneity in intervention intensity and duration likely compromises cross-study comparability, suggesting that determining core parameters within specific cultural contexts is crucial.

Limitations

This study has several limitations that should be acknowledged. First, the feasibility study used a convenience sampling method with a small sample size (N=31), which may limit the generalizability of the findings to the broader population with T2DM in China. Importantly, as this was a feasibility study, it lacked the statistical power to definitively assess clinical effectiveness; thus, the observed reductions in sedentary time should be viewed as preliminary. Second, the 7-week intervention period was relatively short, which likely explains the lack of significant changes in physiological and metabolic outcomes. Third, data collection relied on self-reported questionnaires, which are subject to recall bias. Future research should use randomized controlled trials with longer follow-up periods and objective measures of SB to further validate these findings.

Conclusions

The "Double-S" program demonstrates that theory-driven telehealth interventions can be successfully developed and adapted to specific cultural and technological contexts to address the public health challenge of SB. Beyond the immediate reduction in sedentary time, this research highlights the critical role of social media ecosystems and family-based social support in managing chronic diseases among older adults. Rather than providing definitive evidence of efficacy, this study offers a scalable design model and a validated methodological foundation for health care systems in UMICs. Successfully scaling such digital health solutions requires moving beyond hardware deployment toward institutional adaptation and the development of localized, culturally sensitive outcome sets. The positive outcomes from this pilot phase support the progression toward large-scale implementation trials to further validate the program's impact.

Acknowledgments

The authors would like to thank the health care professionals and patients who participated in the stakeholder consultations and pilot study.

Funding

This study was funded by the National High Level Hospital Clinical Research Funding (BJ-2024-198), the Fundamental Research Funds for the Central Universities (3332024098), and the Research Funding of the Chinese Nursing Association (ZHKYQ202416). The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Data Availability

All of the data are available upon reasonable request to the corresponding author.

Data Availability

All of the data are available on reasonable request to the corresponding author.

XSL•FO
RenderX

Authors' Contributions

XZ and DY conceptualized and designed the study, and are co-first authors of this paper. LW, XL, and JL participated in the investigation and data curation. The formal data analysis was conducted by XZ, DY, SC, ZW, MM, and HJ. Throughout the study, YH and JL were responsible for program administration and provided methodological guidance. XZ and DY wrote the original draft of the manuscript. SC and ZW reviewed and edited the manuscript. LW, XL, ZW, MM, XL, HJ, YH, and JL reviewed the manuscript and provided valuable revisions. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed recruitment and participant flow.
[DOCX File, 13 KB - jmir_v28i1e80827_app1.docx]

Multimedia Appendix 2
The matching relationship between the determinants identified by theoretical domains framework and the capability, opportunity, motivation, and behavior components.
[DOCX File, 19 KB - jmir_v28i1e80827_app2.docx]

Multimedia Appendix 3
Mapping relationships between behavioral determinants, intervention functions, and policy categories.
[DOCX File, 24 KB - jmir_v28i1e80827_app3.docx]

Multimedia Appendix 4
APEASE (Affordability, Practicability, Effectiveness/Cost-effectiveness, Acceptability, Side-Effects/Safety, Equity) criteria.
[DOCX File, 13 KB - jmir_v28i1e80827_app4.docx]

Multimedia Appendix 5
Three-stage trajectory of sedentary behavior.
[DOCX File, 439 KB - jmir_v28i1e80827_app5.docx]

Multimedia Appendix 6
Intervention function APEASE (Affordability, Practicability, Effectiveness/Cost-effectiveness, Acceptability, Side-Effects/Safety, Equity) score.
[DOCX File, 17 KB - jmir_v28i1e80827_app6.docx]

Multimedia Appendix 7
The practical application of intervention strategies converts results.
[DOCX File, 29 KB - jmir_v28i1e80827_app7.docx]

Multimedia Appendix 8
“Double-S” intervention strategy for older adult patients with type 2 diabetes.
[DOCX File, 18 KB - jmir_v28i1e80827_app8.docx]

Multimedia Appendix 9
“Double-S” program APEASE (Affordability, Practicability, Effectiveness/Cost-effectiveness, Acceptability, Side-Effects/Safety, Equity) score results.
[DOCX File, 15 KB - jmir_v28i1e80827_app9.docx]

Multimedia Appendix 10
The 5 components of the “Double-S” Plan.
[DOCX File, 1216 KB - jmir_v28i1e80827_app10.docx]

Multimedia Appendix 11

XSL•FO
RenderX

General patient information (N=31).
[DOCX File, 16 KB - jmir_v28i1e80827_app11.docx]

References

1. Alberti K, Zimmet PZ. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. Diabet Med 1998 Jul;15(7):539-553. [doi: 10.1002/(SICI)1096-9136(199807)15:7<539::AID-DIA668>3.0.CO;2-S] [Medline: 9686693]
2. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in diabetes prevalence and treatment from 1990 to 2022: a pooled analysis of 1108 population-representative studies with 141 million participants. Lancet 2024 Nov 23;404(10467):2077-2093. [doi: 10.1016/S0140-6736(24)02317-1] [Medline: 39549716]
3. Liu J, Bai R, Chai Z, Cooper ME, Zimmet PZ, Zhang L. Low- and middle-income countries demonstrate rapid growth of type 2 diabetes: an analysis based on Global Burden of Disease 1990-2019 data. Diabetologia 2022 Aug;65(8):1339-1352. [doi: 10.1007/s00125-022-05713-6] [Medline: 35587275]
4. Zhao W. Interpretation on clinical guidelines for prevention and treatment of type 2 diabetes mellitus in the elderly in China (2022 edition). Med J Peking Union Med Coll Hosp 2022;13(4):574-580. [doi: 10.12290/xhyxzz.2022-0199]
5. Ning X, Chen X. The new role of lifestyle interventions in the clinical prevention and treatment of type 2 diabetes. J Pract Med 2025;41(17):2611-2616. [doi: 10.3969/j.issn.1006-5725.2025.17.001]
6. Tremblay MS, Aubert S, Barnes JD, et al. Sedentary behavior research network (SBRN)–terminology consensus project process and outcome. Int J Behav Nutr Phys Act 2017 Jun 10;14(1):75. [doi: 10.1186/s12966-017-0525-8] [Medline: 28599680]
7. Lee IM, Shiroma EJ, Lobelo F, Puska P, Blair SN, Katzmarzyk PT. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. The Lancet 2012 Jul;380(9838):219-229. [doi: 10.1016/S0140-6736(12)61031-9]
8. Harvey JA, Chastin SFM, Skelton DA. Prevalence of sedentary behavior in older adults: a systematic review. Int J Environ Res Public Health 2013 Dec 2;10(12):6645-6661. [doi: 10.3390/ijerph10126645] [Medline: 24317382]
9. Li H, Zhang J, Gu R, Li H, Li L, Jin Y. Status and influencing factors of sedentary behavior in community-dwelling elderly diabetic patients. Chin Nurs Manage 2023;23(5):698-703. [doi: 10.3969/j.issn.1672-1756.2023.05.012]
10. Xu C, Furuya-Kanamori L, Liu Y, et al. Sedentary behavior, physical activity, and all-cause mortality: dose-response and intensity weighted time-use meta-analysis. J Am Med Dir Assoc 2019 Oct;20(10):1206-1212.e3. [doi: 10.1016/j.jamda.2019.05.001] [Medline: 31272857]
11. Cabanas-Sánchez V, Guallar-Castillón P, Higueras-Fresnillo S, Rodríguez-Artalejo F, Martínez-Gómez D. Changes in sitting time and cardiovascular mortality in older adults. Am J Prev Med 2018 Mar;54(3):419-422. [doi: 10.1016/j.amepre.2017.10.010] [Medline: 29305068]
12. Schmid D, Leitzmann MF. Television viewing and time spent sedentary in relation to cancer risk: a meta-analysis. J Natl Cancer Inst 2014 Jul;106(7):dju098. [doi: 10.1093/jnci/dju098] [Medline: 24935969]
13. Lee SH, Son C, Yeo S, Ha IH. Cross-sectional analysis of self-reported sedentary behaviors and chronic knee pain among South Korean adults over 50 years of age in KNHANES 2013-2015. BMC Public Health 2019 Oct 26;19(1):1375. [doi: 10.1186/s12889-019-7653-9] [Medline: 31655569]
14. Chastin SFM, Mandrichenko O, Helbostadt JL, Skelton DA. Associations between objectively-measured sedentary behaviour and physical activity with bone mineral density in adults and older adults, the NHANES study. Bone 2014 Jul;64:254-262. [doi: 10.1016/j.bone.2014.04.009] [Medline: 24735973]
15. Beunza JJ, Martínez-González MA, Ebrahim S, et al. Sedentary behaviors and the risk of incident hypertension: the SUN Cohort. Am J Hypertens 2007 Nov;20(11):1156-1162. [doi: 10.1016/j.amjhyper.2007.06.007] [Medline: 17954361]
16. Ohlsson C, Gidestrand E, Bellman J, et al. Increased weight loading reduces body weight and body fat in obese subjects—a proof of concept randomized clinical trial. EClinicalMedicine 2020 May;22:100338. [doi: 10.1016/j.eclinm.2020.100338] [Medline: 32510046]
17. Izquierdo M, Merchant RA, Morley JE, et al. International exercise recommendations in older adults (ICFSR): expert consensus guidelines. J Nutr Health Aging 2021;25(7):824-853. [doi: 10.1007/s12603-021-1665-8] [Medline: 34409961]
18. Katzmarzyk PT, Powell KE, Jakicic JM, et al. Sedentary behavior and health: update from the 2018 Physical Activity Guidelines Advisory Committee. Med Sci Sports Exerc 2019 Jun;51(6):1227-1241. [doi: 10.1249/MSS.0000000000001935] [Medline: 31095080]
19. Bull FC, Al-Ansari SS, Biddle S, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br J Sports Med 2020 Dec;54(24):1451-1462. [doi: 10.1136/bjsports-2020-102955] [Medline: 33239350]
20. Prince SA, Saunders TJ, Gresty K, Reid RD. A comparison of the effectiveness of physical activity and sedentary behaviour interventions in reducing sedentary time in adults: a systematic review and meta-analysis of controlled trials. Obes Rev 2014 Nov;15(11):905-919. [doi: 10.1111/obr.12215] [Medline: 25112481]
21. Martin A, Fitzsimons C, Jepson R, et al. Interventions with potential to reduce sedentary time in adults: systematic review and meta-analysis. Br J Sports Med 2015 Aug;49(16):1056-1063. [doi: 10.1136/bjsports-2014-094524] [Medline: 25907181]

XSL•FO
RenderX

22. Gomersall S, Maher C, English C, Rowlands A, Olds T. Time regained: when people stop a physical activity program, how does their time use change? A randomised controlled trial. PLoS ONE 2015;10(5):e0126665. [doi: 10.1371/journal.pone.0126665] [Medline: 26023914]
23. Melanson EL. The effect of exercise on non-exercise physical activity and sedentary behavior in adults. Obes Rev 2017 Feb;18(Suppl 1):40-49. [doi: 10.1111/obr.12507] [Medline: 28164451]
24. Chastin S, Gardiner PA, Harvey JA, et al. Interventions for reducing sedentary behaviour in community-dwelling older adults. Cochrane Database Syst Rev 2021 Jun 25;6(6):CD012784. [doi: 10.1002/14651858.CD012784.pub2] [Medline: 34169503]
25. Gentili A, Failla G, Melnyk A, et al. The cost-effectiveness of digital health interventions: a systematic review of the literature. Front Public Health 2022;10:787135. [doi: 10.3389/fpubh.2022.787135] [Medline: 36033812]
26. Howland C, Wakefield B. Assessing telehealth interventions for physical activity and sedentary behavior self-management in adults with type 2 diabetes mellitus: an integrative review. Res Nurs Health 2021 Feb;44(1):92-110. [doi: 10.1002/nur.22077] [Medline: 33091168]
27. China's digital health development report was released, and digital health community leads the new trend of the industry. Global Times. 2020 Jun 24. URL: https://www.huanqiu.com/article/3ymeOncYqfv [accessed 2025-12-30]
28. Cui F, Ma Q, He X, et al. Implementation and application of telemedicine in China: cross-sectional study. JMIR Mhealth Uhealth 2020 Oct 23;8(10):e18426. [doi: 10.2196/18426] [Medline: 33095175]
29. Skivington K, Matthews L, Simpson SA, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. Int J Nurs Stud 2024 Jun;154:104705. [doi: 10.1016/j.ijnurstu.2024.104705] [Medline: 38564982]
30. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implementation Sci 2009 Dec;4(1):50. [doi: 10.1186/1748-5908-4-50]
31. Proctor EK, Landsverk J, Aarons G, Chambers D, Glisson C, Mittman B. Implementation research in mental health services: an emerging science with conceptual, methodological, and training challenges. Adm Policy Ment Health 2009 Jan;36(1):24-34. [doi: 10.1007/s10488-008-0197-4] [Medline: 19104929]
32. Proctor EK, Powell BJ, McMillen JC. Implementation strategies: recommendations for specifying and reporting. Implement Sci 2013 Dec 1;8:139. [doi: 10.1186/1748-5908-8-139] [Medline: 24289295]
33. Bartholomew LK, Parcel GS, Kok G. Intervention mapping: a process for developing theory- and evidence-based health education programs. Health Educ Behav 1998 Oct;25(5):545-563. [doi: 10.1177/109019819802500502] [Medline: 9768376]
34. Fernandez ME, Ruiter RAC, Markham CM, Kok G. Intervention mapping: theory-and evidence-based health promotion program planning: perspective and examples. Front Public Health 2019;7:209. [doi: 10.3389/fpubh.2019.00209] [Medline: 31475126]
35. De Decker E, De Craemer M, De Bourdeaudhuij I, et al. Using the intervention mapping protocol to reduce European preschoolers' sedentary behavior, an application to the ToyBox-Study. Int J Behav Nutr Phys Act 2014 Feb 19;11:19. [doi: 10.1186/1479-5868-11-19] [Medline: 24552138]
36. Huang Y, Benford S, Price D, et al. Using Internet of Things to reduce office workers' sedentary behavior: intervention development applying the behavior change wheel and human-centered design approach. JMIR Mhealth Uhealth 2020 Jul 29;8(7):e17914. [doi: 10.2196/17914] [Medline: 32723716]
37. Ezeugwu VE, Manns PJ. Using intervention mapping to develop and implement a home-based sedentary behavior change intervention after stroke. Transl Behav Med 2020 Feb 3;10(1):87-95. [doi: 10.1093/tbm/iby128] [Medline: 30566661]
38. Direito A, Walsh D, Hinbarji M, et al. Using the intervention mapping and behavioral intervention technology frameworks: development of an mHealth intervention for physical activity and sedentary behavior change. Health Educ Behav 2018 Jun;45(3):331-348. [doi: 10.1177/1090198117742438] [Medline: 29216765]
39. Zhang X, Yang D, Luo J, et al. Determinants of sedentary behavior in community-dwelling older adults with type 2 diabetes based on the behavioral change wheel: a path analysis. BMC Geriatr 2024 Jun 6;24(1):502. [doi: 10.1186/s12877-024-05076-0]
40. Kok G, Peters LWH, Ruiter RAC. Planning theory- and evidence-based behavior change interventions: a conceptual review of the intervention mapping protocol. Psicol Reflex Crit 2017 Oct 18;30(1):19. [doi: 10.1186/s41155-017-0072-x] [Medline: 32026109]
41. Michie S, van Stralen MM, West R. The Behaviour Change Wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
42. Zhang X, Li X, Yang D, Meng M, Hao Y. Intervention strategies for reducing sedentary behavior among older adults:a scoping review. Chin J Nurs 2024;59(8):1012-1020. [doi: 10.3761/j.issn.0254-1769.2024.08.017]
43. National Center of Gerontology, Chinese Diabetes Society, China Sport Science Society. Guideline for exercise therapy of type 2 diabetes mellitus in China. Chin J Sports Med 2024;43(6):419-452. [doi: 10.12114/j.issn.1007-9572.2024.A0019]
44. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]

XSL•FO
RenderX

45. Global strategy on digital health 2020-2025. : World Health Organization; 2021 URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2026-02-27]
46. Ye H, Zhu X, Zhu X, Lin Y, Zhuang J. Evaluation of reliability, validity and reactivity of Diabetes Specific Quality of Life scale in rural-urban fringe diabetic patients. J Nurs 2009;16(24):60-61. [doi: 10.16460/j.issn1008-9969.2009.24.026]
47. Chang Q, Sha F, Chan CH, Yip PSF. Validation of an abbreviated version of the Lubben Social Network Scale ("LSNS-6") and its associations with suicidality among older adults in China. PLoS ONE 2018;13(8):e0201612. [doi: 10.1371/journal.pone.0201612] [Medline: 30071067]
48. Yang K. Status of social network and its influencing factors among elderly patients with type 2 diabetes in Beijing. J Nurs Sci 2020;35(10):94-97. [doi: 10.3870/j.issn.1001-4152.2020.10.094]
49. Zhang X. Sedentary behavior mechanism and intervention strategies in community-dwelling elderly patients with T2DM [PhD Thesis]. : Beijing University of Chinese Medicine; 2018.
50. Lewis CC, Klasnja P, Powell BJ, et al. From classification to causality: advancing understanding of mechanisms of change in implementation science. Front Public Health 2018;6:136. [doi: 10.3389/fpubh.2018.00136] [Medline: 29868544]
51. Fernandez ME, Ten Hoor GA, van Lieshout S, et al. Implementation mapping: using intervention mapping to develop implementation strategies. Front Public Health 2019;7:158. [doi: 10.3389/fpubh.2019.00158] [Medline: 31275915]
52. Hurley DA, Murphy LC, Hayes D, et al. Using intervention mapping to develop a theory-driven, group-based complex intervention to support self-management of osteoarthritis and low back pain (SOLAS). Implementation Sci 2016 Dec;11(1):56. [doi: 10.1186/s13012-016-0418-2]
53. Mirzaei-Alavijeh M, Jouybari TA, Jalilian F, Motlagh ME, Jalilian F. Using intervention mapping approach to finding socio-cognitive determinants of diabetes preventive behaviors. J Prev Med Hyg 2019 Sep;60(3):E237-E242. [doi: 10.15167/2421-4248/jpmh2019.60.3.1159] [Medline: 31650060]
54. Yan Y, Wu T, Zhang M, Li C, Liu Q, Li F. Prevalence, awareness and control of type 2 diabetes mellitus and risk factors in Chinese elderly population. BMC Public Health 2022 Jul 19;22(1):1382. [doi: 10.1186/s12889-022-13759-9] [Medline: 35854279]
55. Sriram RD, Reddy SSK. Artificial intelligence and digital tools: future of diabetes care. Clin Geriatr Med 2020 Aug;36(3):513-525. [doi: 10.1016/j.cger.2020.04.009] [Medline: 32586478]
56. Ming Y, Li Y, Liu Y. Digital technologies as solutions to China's aging population: a systematic review of their opportunities and challenges in rural development. Front Public Health 2024;12:1416968. [doi: 10.3389/fpubh.2024.1416968] [Medline: 39882121]
57. Ferguson T, Olds T, Curtis R, et al. Effectiveness of wearable activity trackers to increase physical activity and improve health: a systematic review of systematic reviews and meta-analyses. Lancet Digit Health 2022 Aug;4(8):e615-e626. [doi: 10.1016/S2589-7500(22)00111-X] [Medline: 35868813]

Abbreviations

APEASE: Affordability, Practicability, Effectiveness/Cost-effectiveness, Acceptability, Side Effects/Safety, Equity
BCT: behavior change technique
BCW: Behavior Change Wheel
COM-B: capability, opportunity, motivation, and behavior
IM: intervention mapping
LSNS-6: Lubben Social Network Scale-6
MG: motion graphics
PO: performance objective
SB: sedentary behavior
T2DM: type 2 diabetes mellitus
TDF: theoretical domains framework
UMIC: upper-middle-income country
WHO: World Health Organization

Edited by A Stone; submitted 17.Jul.2025; peer-reviewed by IB Felix, LF Belo; revised version received 12.Feb.2026; accepted 18.Feb.2026; published 26.Mar.2026.

Please cite as:
Zhang X, Yang D, Chen S, Wang L, Li X, Wang Z, Meng M, Li X, Jiang H, Hao Y, Luo J
Telehealth Intervention to Reduce Sedentary Behavior in Older Adults With Type 2 Diabetes: Development and Feasibility Study
J Med Internet Res 2026;28:e80827
URL: https://www.jmir.org/2026/1/e80827
doi:10.2196/80827

XSL•FO
RenderX

© Xiaoyan Zhang, Dan Yang, Sihan Chen, Lei Wang, Xin Li, Ziyan Wang, Meiqi Meng, Xuejing Li, Hongzhan Jiang, Yufang Hao, Jiayin Luo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Association Between Physician Communication Features and Patient Outcomes in Telemedicine: Retrospective Cross-Sectional Observational Study

Yipei Wang[1,2*], MBA, PhD; Shu Wang[3*], BA; Ke Zhang[1,4*], MD; Zhijie Liu[5]; Qingbian Ma[6], MD; Hong Ji[7], MS; Zheng Hou[1,8], MD; Tracy Xiao Liu[3], PhD; Xuedong Xu[9], MBA; Xinxia Wu[1], MS; Changxiao Jin[2], MS

[1]Department of Medical Affairs, Office of Internet Hospital, Peking University Third Hospital, Beijing, China

[2]Institute of Hospital Management, Peking University Third Hospital, 49 North Huayuan Road, Haidian District, Beijing, China

[3]Department of Economics, School of Economics and Management, Tsinghua University, Beijing, China

[4]Department of Otolaryngology, Peking University Third Hospital, Beijing, China

[5]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China

[6]Department of Emergency Medicine, Peking University Third Hospital, Beijing, China

[7]Information Management and Big Data Center, Peking University Third Hospital, Beijing, China

[8]Department of Obstetrics and Gynecology, Peking University Third Hospital, Beijing, China

[9]Peking University People's Hospital, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Changxiao Jin, MS
Institute of Hospital Management, Peking University Third Hospital, 49 North Huayuan Road, Haidian District, Beijing, China

Abstract

Background: Asynchronous telemedicine is a crucial component of multichannel health care, where effective communication drives satisfaction. However, the effectiveness of communication features remains poorly understood. Prior research relied on subjective surveys or small-scale simulations, failing to link features to objective outcomes. Understanding these features is critical for optimizing physician engagement and establishing quality indicators to enhance the patient experience.

Objective: This study aimed to bridge this gap by leveraging a large-scale real-world dataset to quantify the association between physicians' communication features—including response modalities, length, and sequence—and patient repurchase behavior, as well as review scores, within a high-autonomy health care setting.

Methods: This retrospective cross-sectional study analyzed 304,337 paid, patient-initiated virtual visits from a Chinese academic medical center (2021 - 2023), which included 823,135 physician responses. The sample was selected after applying a series of exclusion criteria, such as free consultations, team-based visits, and outlier data. The key exposures were the modality of physician responses, response length, and response sequence. Outcomes included patient loyalty and satisfaction. Loyalty was operationalized as follow-up visits within 6 months, with a 30-day exclusion period applied to same-physician (*fv1*) and same-department (*fv2*) revisits to filter out clinical necessity, but not to hospital-wide revisits (*fv3*). Satisfaction was measured by the review scores. We used probit and ordinary least squares regressions to examine the relationships between communication features and patient outcomes.

Results: Regarding loyalty, audio-only visits were associated with the lowest *fv1*, with an average marginal effect (AME) of −0.030 (95% CI −0.043 to −0.016, *P*<.001), translating to a 30.9% (0.030/0.097) reduction compared to text-only visits. Regarding satisfaction, audio messages were associated with a significantly increased likelihood of patients providing reviews, with an AME of 0.041 (95% CI 0.006 - 0.076, *P*=.02), but they did not affect review scores after adjusting for inverse Mills ratios. Increased numbers of text and audio replies were (marginally) associated with improved *fv1*, with AMEs of 0.009 (95% CI 0.006 - 0.011, *P*<.001) and 0.007 (95% CI −0.000 to 0.016, *P*=.06), respectively. Visits beginning with a sub-5-second audio response and ending with text had significantly higher *fv1* than text-only visits, with an AME of 0.069 (95% CI 0.018 - 0.120, *P*=.008). The same patterns hold for *fv2* and *fv3*. Based on the Bonferroni test, coefficients with a *P* value smaller than α=.050/3=.017 or α=.50/2=.025 were regarded as significant when evaluating the association with patient loyalty or satisfaction, respectively.

Conclusions: Physician communication practices were significantly associated with patient loyalty and satisfaction. This study is innovative in leveraging large-scale real-world data to systematically examine physician communication. It differs from existing studies by transcending prior survey-based research limitations. It introduces an effective hybrid approach, balancing human

XSL•FO
RenderX

connection with text clarity in the field. Its implication in the real world is providing data-driven evidence to guide clinicians and policymakers in designing high-quality telemedicine services.

(J Med Internet Res 2026;28:e86977) doi:10.2196/86977

KEYWORDS

telemedicine; asynchronous visits; physician-patient communication; patient loyalty; patient satisfaction; behavioral pattern; quality of virtual care; health-seeking behavior

Introduction

Telemedicine has emerged as a crucial component of the multichannel service delivery [1-3], significantly expanding access to care beyond the COVID-19 pandemic [4-6]. While granting increased convenience [7,8], enhanced accessibility [9,10], and improved clinical outcomes [11,12], the shift to virtual settings strips away physical interactions, making the communication skills of physicians essential, as they positively influence patient satisfaction and engagement [13-16]. In the specific context of asynchronous telemedicine, where visual cues are lacking, physicians' choices of text and/or audio messages become critical. Physicians face a dilemma: text messages offer clarity and ease of review, whereas audio communication enables faster expression and conveys a sense of human presence [17-19]. Striking the right balance between these modalities is essential for establishing a robust physician-patient connection [20,21].

Nonetheless, few studies have comprehensively investigated which features of physician response behaviors most effectively enhance the quality of virtual care. Prior research on this topic predominantly focused on in-person visit settings, examining the influence of computer-based electronic medical record writing and patient demographic factors on physician-patient interactions [22,23]. In the context of telemedicine, studies have explored the effect of audio use on satisfaction [24], yielding inconsistent results [18]. Some studies have quantified the length of physician responses and examined its impact on diagnostic accuracy and patient satisfaction using standardized patients [25,26]. Although these studies have value, they exhibit several shortcomings, including the subjectivity of patient satisfaction as an indicator, the artificiality of the standardized patient approach, and constraints on disease categories and sample sizes. Taken together, there is a scarcity of large-scale real-world evidence linking physician communication features to objective behavioral outcomes such as patient loyalty.

The Chinese health care context offers a unique setting to examine patient loyalty due to its high degree of consumer autonomy. Unlike systems with strict gatekeeping, China lacks a mandatory primary care referral system, granting patients direct access to specialists [27-29]. Furthermore, the government-led universal health insurance eliminates restrictive provider networks, allowing patients to seek care from any public institution without facing out-of-network penalties [30]. This autonomy is particularly pronounced in the asynchronous telemedicine examined in this study. In contrast to synchronous virtual visits, which are typically scheduled by physicians for postdischarge or routine follow-up visits based on clinical necessity, asynchronous consultations are patient-initiated and on-demand. Patients voluntarily seek advice, and physicians respond during flexible intervals. Therefore, unconstrained by either insurance policies or provider-mandated appointment requirements, the asynchronous setting effectively isolates patient behavior from structural and clinical constraints. Consequently, within the context of asynchronous consultations in China, patient repurchase acts as a revealed preference, driven predominantly by the overall care experience rather than by insurance constraints or referral mandates.

Our study leverages this unique setting to address the research gap by systematically investigating the association between physician communication features and patient outcomes. The primary objective of this cross-sectional study is to examine how response modality, length, and sequence relate to objective patient loyalty and subjective patient satisfaction. By using large-scale real-world data from 304,337 asynchronous visits at an academic medical center (AMC) in China, we aim to elucidate the distinct utility of these communication features and identify effective engagement patterns. Ultimately, this work is expected to establish an empirical basis for defining quality indicators and improving the patient experience of telemedicine.

Methods

Inclusion and Exclusion Criteria

This cross-sectional study was designed and reported in accordance with the Journal Article Reporting Standards (JARS) [31]. Our study used all administrative records (368,317 visits) of telemedicine visits from Peking University Third Hospital, a large AMC in China, covering the period from January 2021 to December 2023. These records were directly exported from the hospital information system (HIS) of the AMC. In this study, a "visit" refers to a complete, billable asynchronous consultation episode, distinct from a single message or reminder. To reflect the dynamics of individual physicians responding to patients' requests on a daily basis, the following criteria were applied (Multimedia Appendix 1): (1) free telemedicine visits provided by physicians on a voluntary basis were excluded, as physicians' behavior might change (n=35,804); (2) visits responded to by a medical team, such as multidisciplinary treatments, and those by physicians in diagnostic and therapeutic service departments, were excluded (n=2217); (3) visits initiated by physicians were excluded (n=97); (4) visits where physicians only prescribed medication without replying with a response were excluded (n=7075); (5) to ensure that our results were not skewed by specific physicians, we further excluded visits of "outlier physicians" from the analysis. Specifically, we focused on high-volume physicians, defined as those whose visit counts exceeded the median. We then analyzed the distributions of

XSL•FO
RenderX

response volume (number of messages) and duration (length of text/audio). Upon observing structural discontinuities at the extreme tails of these distributions, we applied a gap-based exclusion criterion: the top-ranked physician was removed if their mean value was more than twice that of the second-ranked physician. This conservative threshold was chosen to eliminate extreme anomalies likely driven by nonstandard usage patterns (eg, systematic templates) that could act as high leverage points and skew the regression estimates, while retaining valid high-performing physicians (n=12,086); and (6) given our research interest, we excluded visits containing image responses sent by physicians (n=6701). Altogether, the final dataset contained 304,337 (82.63% of 368,317 typical visits) telemedicine visits provided by 781 unique physicians, requested by 140,466 unique patients.

Participant Characteristics and Dataset Description

The 781 unique physicians exhibited a balanced gender distribution, with the average age in the mid-forties. Around 85% of individuals were from the fields of internal medicine, surgery, and obstetrics & gynecology. The proportions of different professional titles, such as chief, deputy chief, and attending physicians, were approximately equal (Table 1). Over the 3-year period, the average number of asynchronous visits per patient was 2.069, 1.810, and 1.786 per year. Over 60% of patients attended only once each year (Table 1).

XSL•FO
RenderX

Table . Baseline characteristics at the visit level (N=304,337), physician level (N=781), and patient level (N=140,466) at Peking University Third Hospital, Beijing, China (2021 - 2023).

	2021	2022	2023
Physician level			
Total, n (%)	582 (100)	698 (100)	706 (100)
Field, n (%)			
Obstetrics and gynecology	70 (12)	90 (12.9)	94 (13.3)
Internal medicine	229 (39.3)	280 (40.1)	279 (39.5)
Surgery	185 (31.8)	214 (30.7)	226 (32)
Pediatrics	27 (4.6)	30 (4.3)	28 (4)
Ophthalmology and otorhinolaryngology	71 (12.2)	84 (12)	79 (11.2)
Position at the year, n (%)			
Attending physician	149 (25.6)	224 (32.1)	226 (32)
Deputy chief physician	239 (41.1)	250 (35.8)	252 (35.7)
Chief physician	194 (33.3)	224 (32.1)	228 (32.3)
Gender, n (%)			
Female	303 (52.1)	353 (50.6)	358 (50.7)
Male	279 (47.9)	345 (49.4)	348 (49.3)
Age at the year, mean (SD)	44,735 (8.235)	44,838 (8.381)	45,072 (8.573)
Number of visits per physician, mean (SD)	145,847 (495.547)	145,479 (413.860)	167,011 (503.085)
Patient level			
Total, n (%)	41,033 (100)	56,093 (100)	66,013 (100)
Number of visits, n (%)			
1	24,786 (60.4)	37,733 (67.3)	44,669 (67.7)
2	7268 (17.7)	9011 (16.1)	10,643 (16.1)
3	3416 (8.3)	3815 (6.8)	4443 (6.7)
4	1885 (4.6)	1951 (3.5)	2284 (3.5)
5	1191 (2.9)	1175 (2.1)	1336 (2)
6	757 (1.8)	781 (1.4)	815 (1.2)
≥7	1730 (4.2)	1627 (2.9)	1823 (2.8)
Visit level			
Total, n (%)	84,883 (100)	101,544 (100)	117,910 (100)
Field, n (%)			
Obstetrics and gynecology	47,863 (56.4)	47,980 (47.3)	53,948 (45.8)
Internal medicine	18,622 (21.9)	27,356 (26.9)	28,554 (24.2)
Surgery	13,080 (15.4)	16,733 (16.5)	27,185 (23.1)
Pediatrics	3428 (4)	6192 (6.1)	4417 (3.7)
Ophthalmology and otorhinolaryngology	1890 (2.2)	3283 (3.2)	3806 (3.2)
Position at the year, n (%)			
Attending physician	11,041 (13)	12,685 (12.5)	15,417 (13.1)
Deputy chief physician	26,011 (30.6)	30,704 (30.2)	38,440 (32.6)
Chief physician	47,831 (56.3)	58,155 (57.3)	64,053 (54.3)

XSL•FO
RenderX

	2021	2022	2023
Total text length, mean (SD)	70.564 (95.751)	69.445 (92.141)	69.876 (95.651)
Total audio length, mean (SD)	24.261 (41.606)	24.411 (43.024)	22.078 (40.886)

Around half of the 304,337 visits pertained to obstetrics and gynecology, whereas roughly a quarter were related to internal medicine. The Department of Reproductive Medicine in the field of obstetrics and gynecology had the highest number of asynchronous visits, making up around 40% of the total visits (Figure 1A and Multimedia Appendix 2).

Figure 1. Study overview and distribution of 823,135 physician responses across 304,337 retrospective asynchronous telemedicine visits at Peking University Third Hospital, Beijing, China (2021 - 2023).

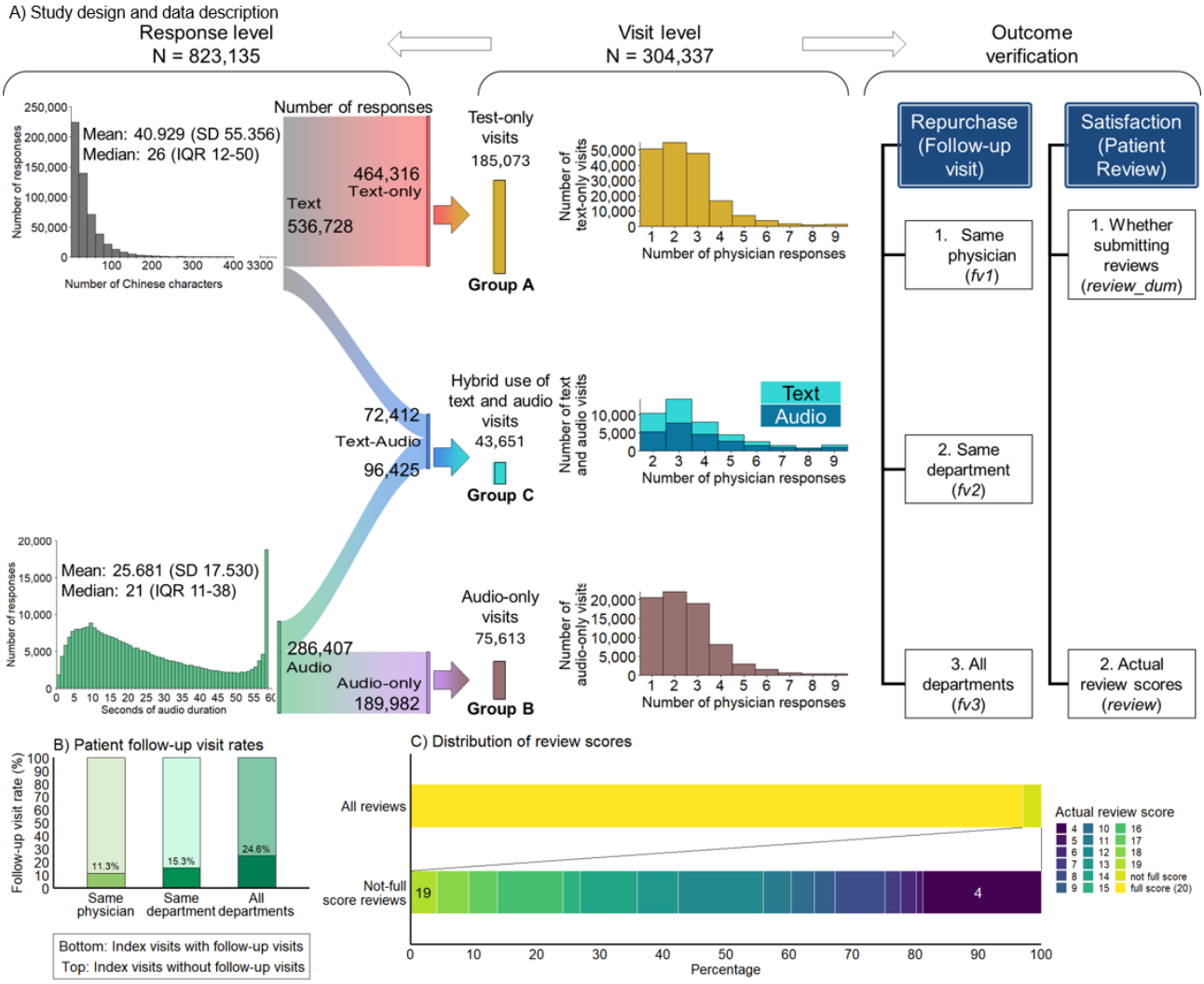

Between 2021 and 2023, the number of visits exhibited a considerable compound annual growth rate of 17.9%. A modest rise in the number of physicians offering asynchronous treatment was observed, with a notable surge in 2022. In 2023, there was a 14.8% (21,533/145,478) rise in the average workload per physician, with significant heterogeneity noticed among individuals.

These 304,337 visits included 823,135 physician responses, translating into 2.705 responses per visit. Of all 823,135 responses, 65.2% (n=536,728) were delivered in text modality and 34.8% (n=286,407) in audio modality. Text responses had no word restriction; however, each audio response was restricted to a maximum duration of 59 seconds, similar to the widely used instant communication application WeChat in China. On average, each text response contained 40.929 Chinese characters, whereas each audio response lasted 25.681 seconds (Figure 1A).

A total of 464,316 text responses formed 185,073 text-only visits, 189,982 audio responses formed 75,613 audio-only visits, and the additional 43,651 visits included both text and audio information. Regardless of the modalities, most visits consisted of no more than 3 physician responses (Figure 1A). Among all index visits, the follow-up visit rates within 6 months were 11.3% (13,937/123,504), 15.3% (15,203/99,126), and 24.8% (19,791/80,557) when considering the scope of the same physician (*fv1*), same department (*fv2*), and all departments (*fv3*) (Figure 1B). Additionally, the percentage of visits reviewed by patients (*review_dum*) was 5% (15,268/304,337), with an

XSL•FO
RenderX

average score (*review*) of 19.299 (SD 2.981) out of 20 (Figure 1C).

Sampling Procedures

This study used a retrospective cross-sectional observational design. All visits that met the predefined eligibility criteria during the study period were included. Therefore, no sampling was performed; instead, the study used a census of all eligible visit records. Accordingly, there were no self-selection concerns in our sample. The dataset was obtained directly from the hospital in Beijing in April 2024 and was analyzed subsequently.

Sample Size, Power, and Precision

As this was a retrospective study using existing data from HIS, a formal a priori power analysis was not conducted. The sample size was determined by the data in the Peking University Third Hospital within the specified period (ie, January 2021 to December 2023). All eligible records within this timeframe were included to maximize statistical power. Thus, the intended sample size was 368,317 visits. Following the inclusion and exclusion criteria, we achieved an actual sample size of 304,337 visits. The difference resulted from the exclusion process (steps 1 - 6 described in Multimedia Appendix 1).

With this substantial sample size, the study is adequately powered to conduct statistical analyses, ensuring high precision in parameter estimates.

Measures and Covariates

Our study focused on 3 features of physician responses, including the modality of physician responses (text-only, audio-only, or hybrid), response length, and response sequence. Two outcomes (patient loyalty and patient satisfaction) were used to examine the consequences of these physician characteristics.

Outcome: Patient Loyalty

Revisit intention is commonly used to measure consumers' recognition of a product or service and has been studied across various fields, such as tourism, catering, and health care [32-34]. However, revisit intention is a subjective measure that may lack accuracy. Thus, we focused on actual revisits, which provide an objective description of behavior. To distinguish behavioral loyalty from clinical dependency and avoid confusion with the negative quality indicator "revisit," we specifically used "follow-up visit" as a proxy for patient loyalty.

This measurement strategy is grounded in 2 key premises. First, revisits within the acute phase (eg, 14, 21, or 30 d) often serve as negative indicators in health care, reflecting unresolved medical issues or unmet expectations requiring additional treatment in a relatively short period of time [35-41]. Second, in this setting where patients have sufficient autonomy to choose providers for ongoing care, a voluntary return after the acute phase reflects a revealed preference rather than a lack of alternatives [42]. In this context, a follow-up visit is defined as the patient reusing the asynchronous service after a specific exclusion period following the index visit. Based on this rationale, a subsample of index visits was constructed to accurately capture this long-term behavioral loyalty.

To do so, we defined 2 dimensions—period and scope—to help formulate the concept of an index visit and a follow-up visit. The period was considered to be 6 months (182 d), which was the typical timeframe that the Clinician & Group Survey used to assess the lasting effects of a visit [43,44]. Meanwhile, the scope varied across different measurements of visit loyalty. Specifically, the scopes of (1) the same physician, (2) the same department, and (3) all departments applied to the 3 measures: follow-up visit 1 (*fv1*), follow-up visit 2 (*fv2*), and follow-up visit 3 (*fv3*), respectively. The 3 subsamples of index visits were thus defined as the visits without any visit with the same physician, in the same department, or in all departments in the 6 months preceding them (Figure 2).

XSL•FO
RenderX

Figure 2. Schematic illustration of index and follow-up visits used to measure patient loyalty across 3 hierarchical scopes: same physician (*fv1*, light blue circle), same department (*fv2*, dark blue circle), and same hospital (*fv3*, gold circle).

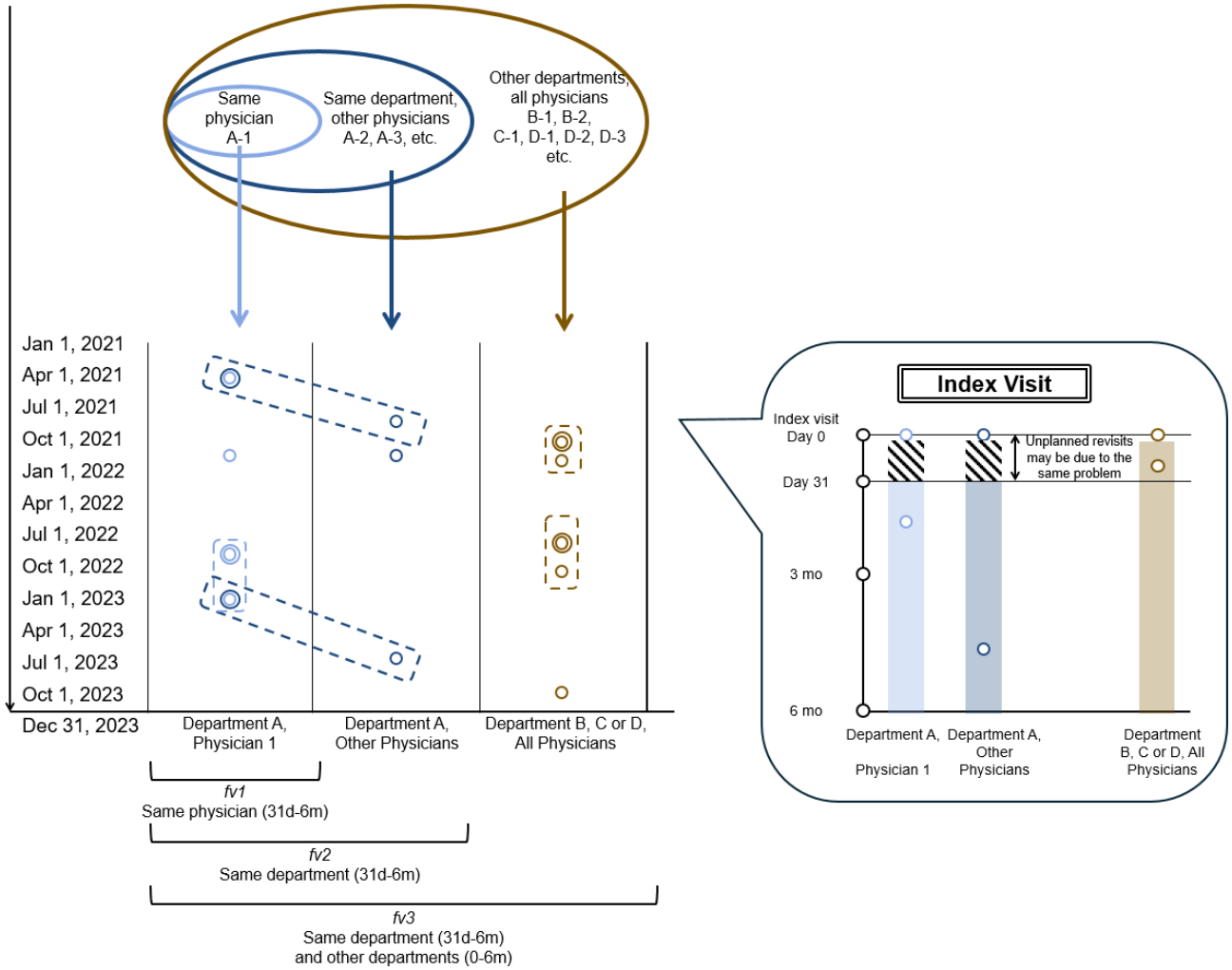

If there was at least one follow-up visit with the same physician, within the same department, or across all departments within 6 months after the corresponding index visit, then *fv1*, *fv2*, or *fv3* equaled 1; otherwise, it was equal to 0. Meanwhile, index visits followed by a visit to the same physician within the subsequent 30 days were excluded for *fv1*, whereas those followed by a visit within the same department were excluded for *fv2* and *fv3*, as immediate follow-up visits were often regarded as a signal of poor quality [35-41]. Note that when the scope was extended to all departments (ie, *fv3*), only follow-up visits within the same department within 30 days were excluded, whereas cross-department visits were retained, as these cross-department follow-up visits were unlikely to reflect poor quality of care and were more likely to indicate new demand. Finally, to ensure we could observe the entire 6-month period before and after the index visits, we excluded 2 types of index visits for all 3 measurements: (1) those in the last 6 months of our sample period and (2) those in the first 6 months of our sample period. In this way, we obtained 123,504, 99,126, and 80,556 index visits for *fv1*, *fv2*, and *fv3*, respectively.

Outcome: Patient Satisfaction

Patient satisfaction is commonly used as an evaluation metric for services in health care settings [45]. Listed as 1 of the 5 domains of the Quintuple Aim, patient satisfaction is strongly related to quality [46-48]. Furthermore, research has demonstrated that it is positively associated with patients' intention to revisit [34,45,49].

In our study, patients could review their experience on a voluntary basis after having asynchronous visits. The review asked the patients to rate their experience on a scale ranging from a minimum of 4 points to a maximum of 20 points. Based on this, 2 measurements were constructed: a dummy variable capturing whether the patient submitted a review (*review_dum*) and the corresponding review score conditional on the submission (*review*).

Control Variables

We defined some other variables to describe the characteristics of physicians and added them as control variables in the regression analyses. Based on the classification of departments in the sample hospital and different disease types, all departments were divided into internal medicine, surgery, obstetrics and gynecology, pediatrics, and ophthalmology and otorhinolaryngology (including ophthalmology, otolaryngology [ENT], and stomatology). Each field was defined as a dummy variable. Furthermore, 3 dummy variables were defined for attending physicians, deputy chief physicians, and chief physicians, respectively. Additionally, we counted the number

XSL•FO
RenderX

of Chinese characters and the seconds of audio duration by physicians per visit as control variables.

Data Collection

Data were directly extracted from the HIS of Peking University Third Hospital. The dataset integrates administrative records (eg, timestamps and payment status) and clinical contents (eg, message modality and response length).

Quality of Measurements

We used all data exactly as recorded in the system. This direct-output process did not involve human data collectors, thus eliminating any potential bias or intercollector variability associated with manual data collection.

Instrumentation

In this study, we developed a set of instruments to comprehensively assess patient loyalty and satisfaction. Loyalty was evaluated through follow-up visits, defined within different periods and scopes. Satisfaction was measured based on the raw review score provided by the patient, including whether a review was submitted and the review score itself.

Masking

Because this was a retrospective study rather than a randomized controlled trial, blinding was not applicable, as data collection occurred after clinical interactions had been completed.

Conditions and Design

This study used a nonexperimental design (ie, no experimental manipulation). As an observational design, it used retrospective data.

Data Diagnostics

There were no missing data in the dataset, ensuring a complete set of records for analysis. All visit-level continuous variables were censored at the top 1% to alleviate concerns related to outliers.

Analytic Strategy

All statistical analyses and visualizations were performed using Stata (MP 18.0), R (R/2024.04.2+764), and Origin (2024bSr0H).

To examine the role of the response modality, we split the whole sample into 3 subsamples: text-only visits (Group A), audio-only visits (Group B), and visits with both modalities (Group C) (Figure 1A). We also created 3 dummy variables to indicate each subsample, respectively. In the regression, we included 2 dummies for Groups B and C, as well as control variables for physician title (a dummy for deputy chief physicians and another for chief physicians). Moreover, the day-of-week fixed effect was added to capture features specific to the day of the week, such as weekdays and weekends. The year-month fixed effect was added to capture month-specific shocks common to all physicians, such as seasonal fluctuations in workload. The physician fixed effect was also included to capture individual behavioral habits. Standard errors were clustered at the physician level. Probit regressions were applied when the dependent variables were *fv1*, *fv2*, *fv3*, and *review_dum* using the Stata command *probit*, and the average marginal effects (AMEs) were reported. Ordinary least squares regressions were applied when the dependent variable was *review* using the Stata command *reghdfe*. Notably, only a subsample of visits contained reviews, which might lead to sample selection concerns. To account for such bias, we followed the standard procedure of the Heckman two-step method [50] by calculating the inverse Mills ratio and adding it into the regression in the second step. To further examine the differences between visits with both text and audio responses and those with only audio responses, we used the Wald test with the Stata command *test* after the regression to examine whether the coefficients of being in Group B and Group C were equivalent.

In a similar vein, we conducted regression analysis using the same models and dependent variables to evaluate the extent to which the length of responses affects patient loyalty and patient satisfaction. Specifically, we included the number of text responses (*text_num*) and the number of audio responses (*audio_num*) as the main independent variables and controlled for the total Chinese characters and total audio duration. Additionally, we added the same set of control variables and fixed effects as in the analysis regarding the response modality. Moreover, as a companion exercise, we further considered the extreme cases of visits with only audio responses, where the total audio duration falls within the range of 55 - 59 seconds (Group B1). We investigated whether splitting long audio responses mattered by including *audio_num* while controlling for the total audio duration.

We next considered the role of the sequence of audio responses. Specifically, we constructed a subsample that included 2 types of visits, both of which included only one audio response. The first started with the audio response (Group C1), whereas the second ended with the audio response (Group C2). We constructed a dummy variable to indicate the second type of visit. By incorporating the dummy variable, as well as controlling for *text_num*, *audio_num*, total Chinese characters, total audio duration, the physician title, and 3 sets of fixed effects, the coefficient of the dummy represented the difference in patient loyalty and satisfaction associated with moving the audio response from the opening to the ending.

The 3 sets of examinations corresponded to our prespecified hypotheses regarding response modalities, response length, and response sequence, each comprising 5 tests. To control the family-wise Type I error rate arising from multiple comparisons, Bonferroni correction was applied. The 3 outcomes related to patient loyalty (*fv1*, *fv2*, and *fv3*) and the 2 outcomes related to patient satisfaction (whether a review was provided and the review score) were each treated as a separate family. Accordingly, *P* values smaller than α=.050/3=.017 for loyalty outcomes and α=.050/2=.025 for satisfaction outcomes were considered statistically significant under the Bonferroni correction.

Then, as an exploratory analysis, the length and position of audio responses were jointly considered using additionally constructed subsamples that also contained 2 types of visits. The first type required the visit to start with the audio response, while the length was required to be shorter than the threshold *t*, ranging from 2 to 60 seconds. The second type included all text-only visits (Group A). After including the dummy variable

XSL•FO
RenderX

for Group A as an independent variable, we repeated the above regression analysis using different subsamples with different thresholds *t* combined with Group A, thus obtaining 59 regression results.

Beyond the dimension of audio length, the variation in position was also considered. We constructed a subsample containing visits with one audio response shorter than 5 seconds while not in the first position, as well as text-only visits (Group A). All visits were required to have at least 2 physician responses to facilitate the comparison. Using the subsample and constructing a dummy for the visits with audio responses, we could examine the difference between the above 2 types of visits regarding outcomes.

Ethical Considerations

This study protocol was approved by the Institutional Review Board (IRB) of Peking University Third Hospital (approval number IRB00006761-M2024701). The requirement for informed consent was waived by the IRB due to the retrospective nature of the study and the use of deidentified data. All data were strictly anonymized and deidentified prior to analysis to protect participant privacy and confidentiality. No compensation was provided to participants for this study. No identifiable images or personal details of participants were included in this manuscript.

Results

Response Modalities

As for the follow-up visit rates, the AME for pure audio response visits was −0.030 (95% CI −0.043 to −0.016, *P*<.001), −0.030 (95% CI −0.045 to −0.016, *P*<.001), and −0.037 (95% CI −0.054 to −0.019, *P*<.001) for *fv1*, *fv2*, and *fv3*, respectively, indicating that the follow-up rates for pure audio visits were 3 - 4 percentage points lower than those for pure text visits (Figure 3). While the combination of text and audio surpassed pure audio in eliciting follow-up visits, its coefficients remained negative in comparison to pure text. In summary, text-only visits (Group A) were associated with the highest patient loyalty, followed by combinations of text and audio (Group C), while audio-only visits (Group B) performed the worst, obtaining 16.6% - 30.9% lower follow-up visit rates than those achieved by text-only visits.

Figure 3. Association between response modality and outcomes (N=304,337 visits) at Peking University Third Hospital, Beijing, China (2021 - 2023). Each subfigure reports the results from one regression. The blue dots represent average marginal effects on *fv1*, *fv2*, *fv3*, and *review_dum* in Probit regressions (A-D) and coefficients for *review* in the ordinary least squares regression (E). Light blue vertical lines with caps represent 95% CIs. Wald tests were used to examine whether the coefficients of the 2 dummies were equal.

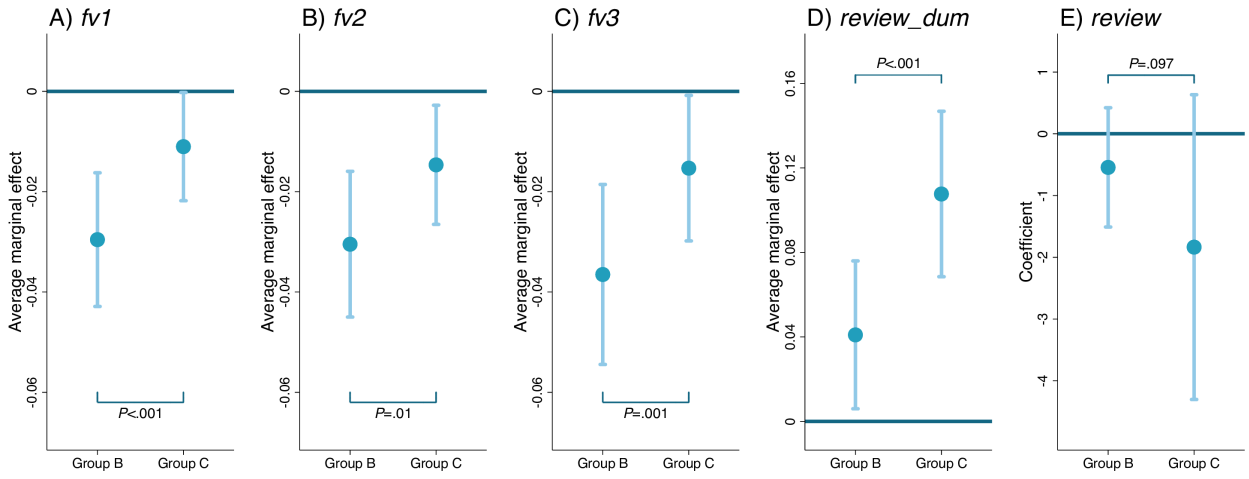

As for the review, whether the use of audio would promote the likelihood of receiving reviews was examined first. The AME of pure audio visits was 0.041 (95% CI 0.006 - 0.076, *P*=.02), indicating that physicians were more likely to receive patient reviews following the provision of medical services in the pure audio modality. The AME for the combination of text and audio was 0.108 (95% CI 0.068 - 0.147, *P*<.001), suggesting that visits in mixed information modalities were most evaluated by patients, followed by those in pure audio, while those in pure text were the least. Subsequently, the linkage between various response modalities and review scores was investigated. Upon adjusting for the inverse Mills ratio, it was determined that there were no statistically significant differences in the review scores among Groups A, B, and C. Note that Group B could achieve a higher level of satisfaction than Group A when the inverse Mills ratio was not controlled (Multimedia Appendix 3).

Response Length

Alongside the modality of response, length is a critical measure in telemedicine. Therefore, the association between response length and outcomes was further investigated. The number of responses was used as a metric to measure visit length when the total information load was fixed.

As for patient loyalty, results demonstrated that the AMEs of *text_num* were 0.009 (95% CI 0.006 - 0.011, *P*<.001), 0.008 (95% CI 0.005 - 0.011, *P*<.001), and 0.008 (95% CI 0.004−0.012 *P*<.001) for *fv1*, *fv2*, and *fv3*, respectively. The AMEs of *audio_num* were 0.007 (95% CI −0.000 to 0.016, *P*=.06), 0.011 (95% CI 0.005−0.018, *P*<.001), and 0.013 (95% CI 0.004 - 0.022, *P*=.006), respectively (whole sample, Figure 4). The findings indicated that irrespective of the modality—text or audio—used for telemedicine services, when the information

XSL•FO
RenderX

load was fixed, an increased number of physician responses correlated with enhanced patient loyalty.

Figure 4. Association between response length and outcomes at Peking University Third Hospital, Beijing, China (2021 - 2023). Each subfigure reports the results of 2 regressions, using the whole sample (N=304,337 visits) and Group B1 (n=3674 visits), respectively. Blue dots represent average marginal effects on *fv1*, *fv2*, *fv3*, and *review_dum* in probit regressions (A-D) and coefficients for *review* in ordinary least squares regressions (E). Light blue vertical lines with caps represent 95% CIs.

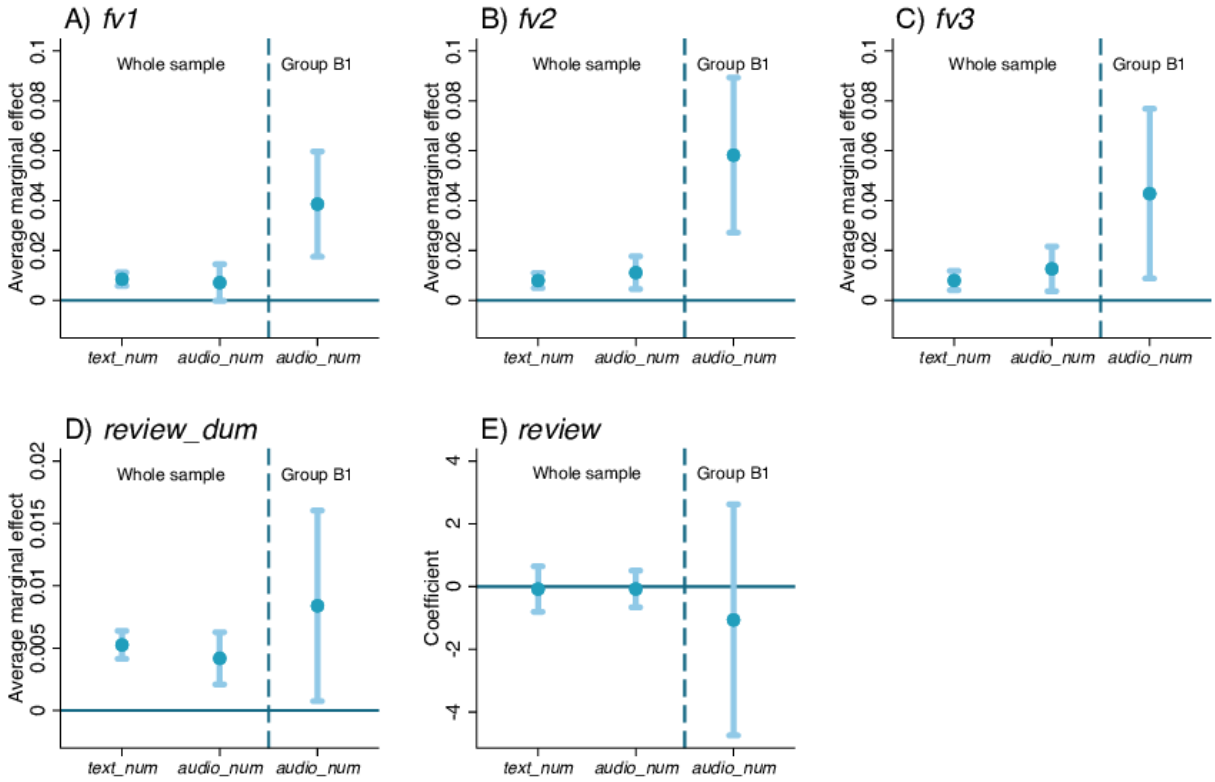

To further validate the above conclusions, we leveraged the design of the telemedicine platform, where a single audio response was limited to a maximum of 59 seconds, to create a more extreme scenario. Specifically, cases where the total audio length of physician responses in a visit fell between 55 and 59 seconds were selected, and we investigated whether there would be differences in the outcome measures as the number of audio responses increased (Group B1, Figure 4).

The results showed that the AMEs of *audio_num* were 0.039 (95% CI 0.017 - 0.060, *P*<.001), 0.058 (95% CI 0.027 - 0.089, *P*<.001), and 0.043 (95% CI 0.009 - 0.077, *P*=.014) for *fv1*, *fv2*, and *fv3*, respectively (Figure 4). This positive association further confirmed that patient loyalty improved when lengthy audio information was divided into several shorter segments.

As for the review, each additional text and audio response was linked to a 0.5% (95% CI 0.4% - 0.6%, *P*<.001) and 0.4% (95% CI 0.2% - 0.6%, *P*<.001) increase in the probabilities of patients providing reviews, respectively. Yet, both additional text and audio responses did not significantly alter the actual review score, indicating that segmenting text or audio into shorter responses was not associated with improved patient satisfaction (Figure 4).

Sequence

In addition to response length, the sequence in which text and audio messages were presented may also be connected with outcomes in visits containing both modalities. We restricted the sample to Group C to examine the association of audio responses positioned at various rounds and the outcomes. Specifically, 2 groups were constructed from the sample. The first group comprised visits in which the first response was in the audio modality and the last response was not (audio-opening visits, Group C1). The second group encompassed visits that ended with an audio response and did not begin with an audio response (audio-ending visits, Group C2).

Firstly, the follow-up visit rates associated with these 2 groups of visits were evaluated. The findings indicated that the AMEs for being in Group C2 were −0.019 (95% CI −0.033 to −0.005, *P*=.007), −0.022 (95% CI −0.042 to −0.001, *P*=.04), and −0.023 (95% CI −0.045 to −0.000, *P*=.046) on *fv1*, *fv2*, and *fv3*, respectively, compared to Group C1 (Table 2). This demonstrated that patients exhibited increased loyalty to the same physician when physicians used audio responses at the beginning of a telemedicine session instead of at the end. Note that the associations for *fv2* and *fv3* were no longer significant after Bonferroni correction.

XSL•FO
RenderX

Table . Association between the position and sequence of the audio responses and outcomes at Peking University Third Hospital, Beijing, China (2021 - 2023).[a]

	fv1	*fv2*	*fv3*	*review_dum*	*review*
Comparing the audio-opening visits (Group C1) and audio-ending visits (Group C2)					
Group C2 (95% CI, *P* value)	−0.019 (−0.033 to −0.005, *P*=.007)	−0.022 (−0.042 to −0.001, *P*=.04)	−0.023 (−0.045 to −0.000, *P*=.046)	0.005 (−0.001 to 0.011, *P*=.098)	0.170 (−1.017 to 1.358, *P*=.78)
Comparing the visits with nonbeginning short audio group and text-only visits (Group A)					
Visits with nonbeginning short audio (95% CI, *P* value)	0.152 (−0.032 to 0.337, *P*=.11)	0.189 (−0.045 to 0.422, *P*=.11)	0.212 (−0.150 to 0.575, *P*=.25)	−0.055 (−0.188 to 0.077, *P*=.41)	−1.960 (−15.598 to 11.677, *P*=.78)
Controls	Y	Y	Y	Y	Y
Physician FE[b]	Y	Y	Y	Y	Y
Day-of-week FE	Y	Y	Y	Y	Y
Year-month FE	Y	Y	Y	Y	Y

[a]The coefficients in the first row were estimated using the subsample Group C1 and Group C2 (n=33,758 visits), and those in the second row were estimated using the subsample Group A and visits with only one nonbeginning short audio response of less than 5 seconds in Group C, both restricted to visits with at least 2 physician responses (n=134,698 visits). Average marginal effects for *fv1*, *fv2*, *fv3*, and *review_dum* from probit regressions are reported in the first 4 columns, and coefficients for *review* from ordinary least squares regressions are reported in the last column.

[b]FE: fixed effect ____.

Secondly, the association between audio position and reviews was examined. The findings indicated no significant association between audio position and the probability of patients submitting reviews or the actual review scores.

Role of Audio

The previous results suggested that, while the audio modality did not facilitate the follow-up visit rate, as reflected by a diminished patient loyalty for Group C compared to Group A, brevity and being in the opening position of audio responses were individually associated with elevated follow-up visit rates. Therefore, the follow-up visit rates of Group C1 and Group A were compared to examine if the decrease in follow-up visit rates due to audio use could be mitigated when the audio was short and positioned at the beginning.

Results showed that when the length of the introductory audio was less than 5 seconds, the follow-up visit rates in this group were significantly higher than those of visits with only text, with the AMEs of 0.069 (95% CI 0.018 - 0.120, *P*=.008), 0.077 (95% CI 0.018 - 0.136, *P*=.011), and 0.099 (95% CI 0.001 - 0.197, *P*=.049) on *fv1*, *fv2*, and *fv3*, respectively (Figure 5). As for reviews, no matter how the length of the introductory audio response changed, there were no significant differences between Group C1 and the text-only visits in terms of the probability of getting reviews or the actual review scores.

XSL•FO
RenderX

Figure 5. Association between the length of the first audio response and outcomes at Peking University Third Hospital, Beijing, China (2021 - 2023). Each subfigure contains results from 59 regressions, each comparing visits with only 1 audio response placed at the first position and shorter than *t* seconds (2≤t≤60) with visits consisting solely of text. The blue curves represent average marginal effect (AMEs) of being the visit with an introductory audio response on *fv1*, *fv2*, *fv3*, and *review_dum* in probit regressions in (A-D), and its coefficients on *review* in ordinary least squares regressions (E). The light blue shades represent 95% CIs. The blue dots and vertical lines with caps represent the AMEs or coefficient and 95% CIs, respectively, when *t*=5 (meaning subsample with pure-text visits and visits with an introductory audio response that are shorter than 5 s).

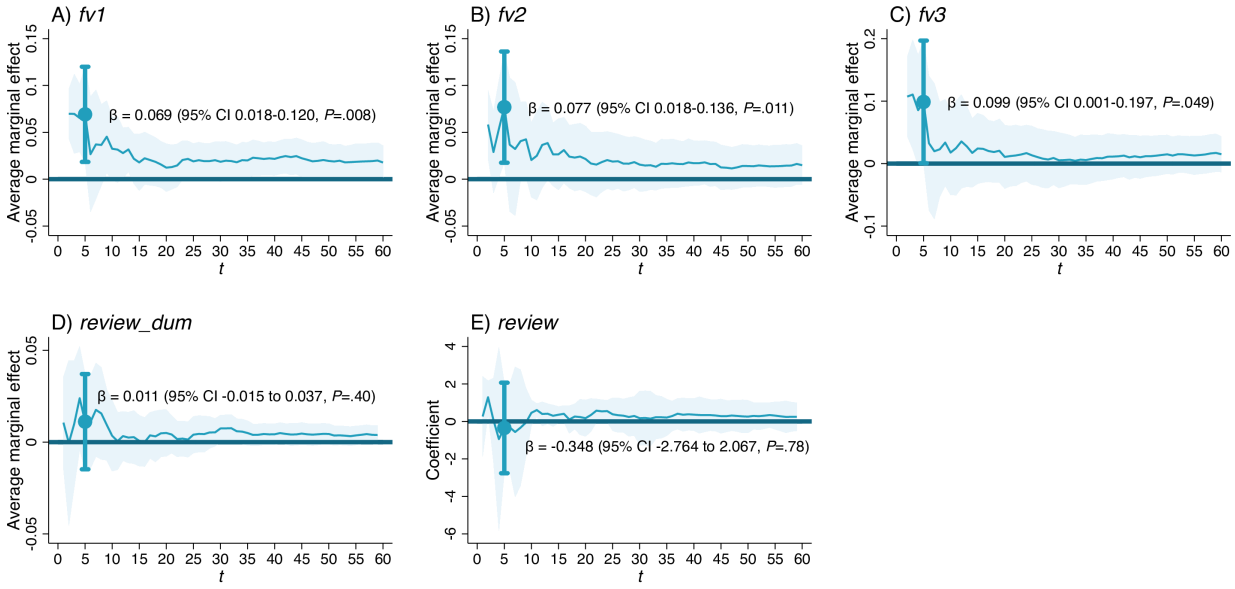

Therefore, visits featuring solely brief introductory audio responses yielded a higher follow-up visit rate than those with only text responses. In addition, this effect diminished when the audio length reached 5 seconds.

Subsequently, short duration (less than 5 s) was maintained as a favorable condition, and whether changing the position of short audio responses would be associated with elevated or diminished follow-up visit rates compared to Group A was assessed. We identified visits in Group C that contained only one short audio response, which was not positioned in the first place, as the nonbeginning audio response group, then compared this group to text-only visits with at least 2 physician responses. No significant effects were observed between the 2 groups of visits (Table 2).

Discussion

Principal Results

This study is the first to systematically examine the effects of physicians' communication modality, length, and sequence on patient behavioral outcomes in asynchronous telemedicine. Our primary findings indicated that the repurchase rate was highest in text-only visits, followed by text-audio hybrid visits, and lowest in audio-only visits. The reduced average length of both text and audio responses was correlated with higher patient loyalty. We also identified a response pattern in which virtual visits begin with a brief introductory audio message followed by text, which was associated with the highest patient loyalty. Additionally, audio responses were connected to a higher likelihood of patients providing reviews without changing the review scores. Our research illustrates specific behavioral patterns of physicians in telemedicine, elucidates the association between various communication features and outcomes, and serves as a reference for delivering high-quality and efficient telemedicine services.

This study improves our current comprehension of physician behaviors in telemedicine by leveraging a large-scale dataset of real-world interactions. A key methodological strength lies in our dual-outcome measurement, which integrates objective longitudinal repurchase behavior with subjective satisfaction scores. While satisfaction may accurately represent the immediate consumption experience, it remains a subjective assessment prone to selection bias, often suffering from a low response rate, which is estimated to be only 5% - 10% in natural settings [51]. In contrast, repurchase behavior serves as an objective, long-term metric free from selection bias [52,53]. By combining these 2 metrics, our study overcomes the limitations of relying solely on self-reported data [34,54,55], offering a more comprehensive evaluation of patient outcomes. In addition, our study thoroughly characterized various dimensions of physician responses, including the modality, length, and sequence. This multidimensional analysis approach establishes a robust empirical foundation for understanding physician-patient communication behaviors.

Central to the validity of this multidimensional framework is the rigorous construction process of the patient loyalty proxy, which was carefully designed to distinguish behavioral preference from clinical necessity or dependency. Unlike in general consumer services, a repeat visit in health care does not inherently signal satisfaction; it may instead reflect unresolved medical needs [56,57]. To address this problem, we imposed specific temporal and scope constraints to isolate true loyalty.

XSL•FO
RenderX

Regarding the temporal dimension, we established a 1‐ to 6-month window to balance the exclusion of clinical necessity against the attribution of care quality. On the lower bound, we excluded revisits occurring within 30 days [38]. In clinical contexts, immediate returns often serve as proxies for unplanned revisits, signaling ineffective initial communication or unresolved symptoms rather than satisfaction [35-41]. On the upper bound, we restricted the observation to 6 months. This cutoff is empirically grounded in the temporal decay of visit influence; for instance, prior research indicates that while initial patient experience significantly predicts retention at 6 months, this association fades and becomes nonsignificant by 12 months [43,44]. Thus, limiting the window ensures that the measured follow-up behavior remains attributable to the impact of the index visit rather than long-term habituation.

Regarding the scope of loyalty, we stratified follow-up visits into 3 levels—the same physician, the same department, and all departments—to capture the multilayered nature of patient trust [58]. While returning to the same physician (*fv1*) serves as the most direct measure of interpersonal loyalty, patients in high-demand academic centers often manifest institutional trust by transferring their confidence to a physician's colleagues [59,60]. By broadening the scope to the same department (*fv2*) and all departments (*fv3*), we capture the spillover effect of effective communication, where a positive interaction with an individual provider enhances the perceived reliability of the entire medical team [59,60].

Crucially, our exclusion criteria incorporated a nuanced interaction between time and scope to account for disease relatedness. We applied the 30-day exclusion period strictly to revisits involving the same physician or department, as these were highly likely to address the same unresolved medical needs. In contrast, visits to different departments within a month were not subject to this exclusion constraint. This distinction is pivotal, as an immediate return to the same specialty suggests a failure to resolve the initial problem, whereas a visit to a different specialty typically indicates a new, unrelated health need [61]. Thus, retaining these cross-specialty visits allows us to capture a broader form of hospital-wide loyalty without confounding it with the negative quality signal of clinical redundancy. Yet, given that returning to the same individual physician offers the most direct and unmediated reflection of the specific interaction's quality, we designated *fv1* as our primary outcome measure to maximize internal validity.

Our focus on repurchase behavior complements the broader literature on patient nonattendance [16]. While nonattendance metrics capture passive adherence to scheduled appointments, the asynchronous workflow in our study operates on a patient-initiated, on-demand basis where prebooking is absent [62]. In this context, engagement is not measured by the absence of a no-show but by the active decision to repurchase. Thus, our loyalty metric represents a deliberate reselection of the provider, serving as a distinct, proactive indicator of the patient-provider connection.

The specific aspects that these 2 outcome indicators highlight align with the different factors that patients encounter during a telemedicine visit. In a virtual visit, a patient's holistic experience is derived from 2 distinct aspects: the comprehension of medical information, which constitutes an evaluation of factual matters, and the perception of physicians' attitudes, which represents an evaluation of interpersonal interactions. We shall examine the importance of the findings from these 2 aspects [63].

Research indicated that comprehension of information was a significant determinant in fostering repurchase behavior [64]. Textual information surpasses auditory in improving understanding of lengthy and complicated material [65]. Meanwhile, medicine is considered a complex field that most individuals have limited knowledge of, resulting in significant information asymmetry [66]. These could explain our findings that text-only visits resulted in more repurchasing behaviors. Because text responses may improve patients' comprehension of physicians' information and provide convenient access for future reference, thereby fostering recognition and promoting repurchases, the outcome metric from an objective and long-term perspective.

Another crucial dimension of outcomes is patient satisfaction, which primarily reflects patients' feelings and attitudes toward physicians. Prior research has indicated that audio communication is more effective than text in bridging gaps among those unable to meet in person [42], and it correlated with increased satisfaction [24]. However, in this study, the use of audio responses resulted in an enhancement of the probability of providing reviews, rather than review scores after adjusting for the inverse Mills ratio. Yet, in the absence of controlling for selection bias, the review scores were higher in the audio-only visits (P=.04), leading us to hypothesize that the increase in satisfaction levels attributed to audio might be a result of selection bias.

Beyond the response modality, a notable discovery is that shorter responses had better outcomes not only for audio responses but also for text responses. When the load of information from the physician was consistent, dividing the content into shorter responses correlated with a higher probability of follow-up visits and an enhanced inclination to provide feedback. This finding aligns with previous empirical research conducted in online health communities, which demonstrated that although physicians' knowledge-sharing behaviors could effectively enhance patient engagement, overly lengthy medical information might inadvertently induce patient anxiety [67], suggesting the importance of response segmentation in physician-patient communication.

Our findings indicated that audio responses were adversely associated with patients' repurchase behavior. The repurchase rate for audio-only or hybrid visits was inferior to that of text alone. Does this imply that audio should not be used at all? Not particularly. We identified that visits featuring a brief introductory audio message (<5 s) followed by text yielded the highest patient loyalty, highlighting it as a particularly effective communication pattern within our sample. This served as evidence that audio itself was not inherently flawed. Given the inability to access substantive content based on communication norms, we hypothesize that these initial, sub–5-second audio clips likely function as personalized greetings or empathetic

XSL•FO
RenderX

openings rather than complex medical explanations. This initial human touch may help establish an emotional connection, while the subsequent text ensures the clarity and reviewability of the medical advice.

Limitations

This study used physician-patient interaction records in the real world and thoroughly examined the roles of text and audio usage, offering a reference for physicians in telemedicine practice; nevertheless, it also had specific limitations.

First, this study focused on observable communication features and did not examine the substantive content of asynchronous care. As a result, the informational richness and clinical appropriateness of medical advice, as well as their roles in shaping communication practices, could not be directly assessed. Future research may extend this work by incorporating content-based or semantic analysis with data on medical specialization and case characteristics to provide a more comprehensive evaluation of how communication quality and clinical complexity jointly affect telemedicine outcomes.

Second, while our study documents associations between physicians' response features and patient behavioral outcomes, it does not allow us to infer causal relationships. Retrospective investigations do indeed possess such limitations. To alleviate the concern, we prioritized controlling for a series of covariates and added abundant fixed effects at the physician level, day-of-week level, and year-month level. Therefore, our results are unlikely to be driven by differences across physicians or time periods. While we cannot fully eliminate endogeneity in a retrospective study, we have taken extensive steps to mitigate such concerns in our analysis.

Third, the external validity of our findings may be influenced by institutional, platform, and cultural factors. The data were drawn from a single large AMC in China, where high patient volume and strong professional norms may shape communication patterns differently from those in primary or community care settings. In addition, platform-specific design features, such as the 59-second audio length limit, may act as technological nudges encouraging concise responses, which is a common constraint across major Chinese platforms. These factors could limit generalizability to other health care systems or international contexts with different technical parameters and cultural expectations around physician-patient interactions. Future cross-institutional, cross-platform, and cross-cultural studies are needed to examine how these elements interact to shape effective communication practices.

Conclusions

This study is innovative in using a large-scale dataset to thoroughly examine the distinct links between specific communication features and patient loyalty outcomes in telemedicine. It differs from existing studies by leveraging a large-scale real-world dataset and objective indicators, diverging from prior research that relied on small-scale simulations or subjective surveys. It provides critical empirical evidence to the field, establishing that while text-only visits facilitate content comprehension and audio encourages feedback, a hybrid pattern with a brief audio introduction followed by text is associated with the highest patient loyalty. Its implication in the real world is offering critical knowledge on how physicians can strategically select engagement tactics to improve service quality and patient retention.

Acknowledgments

Generative artificial intelligence played no role in the writing of this manuscript.

Funding

This work was supported by the Capital Foundation of Medical Development to KZ (2024-2-40915) and the National Natural Science Foundation of China to YW (72204005) and TXL (72342032). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets used in this study are not publicly available due to their proprietary nature but may be made accessible upon reasonable request and with the permission of the corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Flowchart of virtual visits after applying exclusion and inclusion criteria.
[DOCX File, 188 KB - jmir_v28i1e86977_app1.docx]

Multimedia Appendix 2
The number of visits (N=304,337 visits) and physicians (N=781) across years and departments at Peking University Third Hospital, Beijing, China (2021-2023). Departments with visit volume of less than 1% each year are consolidated into "Others."
[DOCX File, 21 KB - jmir_v28i1e86977_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
The impact of response modality on satisfaction without controlling inverse Mills ratio (N=304,337 visits) at Peking University Third Hospital, Beijing, China (2021-2023).
[DOCX File, 17 KB - jmir_v28i1e86977_app3.docx]

References

1. Chang E, Penfold RB, Berkman ND. Patient characteristics and telemedicine use in the US, 2022. JAMA Netw Open 2024 Mar 4;7(3):e243354. [doi: 10.1001/jamanetworkopen.2024.3354] [Medline: 38517438]
2. Dorn SD. Backslide or forward progress? Virtual care at U.S. healthcare systems beyond the COVID-19 pandemic. NPJ Digit Med 2021 Jan 8;4(1):6. [doi: 10.1038/s41746-020-00379-z] [Medline: 33420420]
3. Dorsey ER. The new platforms of health care. NPJ Digit Med 2021 Jul 15;4(1):112. [doi: 10.1038/s41746-021-00478-5] [Medline: 34267312]
4. Shaver J. The state of telehealth before and after the COVID-19 pandemic. Prim Care 2022 Dec;49(4):517-530. [doi: 10.1016/j.pop.2022.04.002] [Medline: 36357058]
5. Appointments in general practice. : NHS England; 2024 URL: https://digital.nhs.uk/data-and-information/publications/statistical/appointments-in-general-practice/june-2024 [accessed 2026-02-28]
6. Amagai S, Vonesh E, Adams J, Luo Y. Closing the gap: addressing telehealth disparities across specialties in the sustained pandemic era. NPJ Digit Med 2024 Aug 21;7(1):217. [doi: 10.1038/s41746-024-01201-w] [Medline: 39164391]
7. Roth M. Telehealth: how asynchronous communication creates provider efficiencies. Accreditation & Quality Compliance Center. URL: https://www.accreditationqualitycenter.com/articles/telehealth-how-asynchronous-communication-creates-provider-efficiencies [accessed 2026-02-28]
8. Rzewuska Díaz M, Locock L, Keen A, Melvin M, Myhill A, Ramsay C. Implementation of a web-based outpatient asynchronous consultation service: mixed methods study. J Med Internet Res 2024 Jun 4;26:e48092. [doi: 10.2196/48092] [Medline: 38833695]
9. Wang Y, Zhang P, Xing Y, et al. Telemedicine integrated care versus in-person care mode for patients with short stature: comprehensive comparison of a retrospective cohort study. J Med Internet Res 2024 Nov 19;26:e57814. [doi: 10.2196/57814] [Medline: 39560983]
10. Barbosa W, Zhou K, Waddell E, Myers T, Dorsey ER. Improving access to care: telemedicine across medical domains. Annu Rev Public Health 2021 Apr 1;42:463-481. [doi: 10.1146/annurev-publhealth-090519-093711] [Medline: 33798406]
11. Nguyen OT, Alishahi Tabriz A, Huo J, Hanna K, Shea CM, Turner K. Impact of asynchronous electronic communication-based visits on clinical outcomes and health care delivery: systematic review. J Med Internet Res 2021 May 5;23(5):e27531. [doi: 10.2196/27531] [Medline: 33843592]
12. Leighton C, Cooper A, Porter A, Edwards A, Joseph-Williams N. Effectiveness and safety of asynchronous telemedicine consultations in general practice: a systematic review. BJGP Open 2024 Apr;8(1). [doi: 10.3399/BJGPO.2023.0177] [Medline: 37783479]
13. Levinson W, Lesser CS, Epstein RM. Developing physician communication skills for patient-centered care. Health Aff (Millwood) 2010 Jul;29(7):1310-1318. [doi: 10.1377/hlthaff.2009.0450] [Medline: 20606179]
14. Drossman DA, Chang L, Deutsch JK, et al. A review of the evidence and recommendations on communication skills and the patient-provider relationship: a Rome foundation working team report. Gastroenterology 2021 Nov;161(5):1670-1688. [doi: 10.1053/j.gastro.2021.07.037] [Medline: 34331912]
15. Mason AN. The most important telemedicine patient satisfaction dimension: patient-centered care. Telemed J E Health 2022 Aug;28(8):1206-1214. [doi: 10.1089/tmj.2021.0322] [Medline: 34882032]
16. Greenup EP, Best D. Systematic review and meta-analysis of no show or non-attendance rates among telehealth and in-person models of care. BMC Health Serv Res 2025 May 9;25(1):663. [doi: 10.1186/s12913-025-12826-2] [Medline: 40340966]
17. He X, Wu D. Does additional audio really work? A study on users' cognitive behavior with audio-visual dual-channel in panoramic digital museum. Inf Manag 2023 Jun;60(4):103791. [doi: 10.1016/j.im.2023.103791]
18. Xia A, Perera SC, Ahmed MU, Tang J, Wang JJ. Voice or text? The role of physician media choice on patient experience in online medical communities. Decis Sci 2024 Dec;55(6):620-638. [doi: 10.1111/deci.12654]
19. Valenzuela-Gálvez ES, González-Benito Ó, Garrido-Morgado Á. Sound in mode on: how sound stimuli influence customer engagement? Electron Commer Res 2026 Feb;26(1):783-804. [doi: 10.1007/s10660-024-09902-w]
20. Culmer N, Smith TB, Stager C, et al. Asynchronous telemedicine: a systematic literature review. Telemed Rep 2023;4(1):366-386. [doi: 10.1089/tmr.2023.0052] [Medline: 38143795]
21. Yan M, Tan H, Jia L, Akram U. The antecedents of poor doctor-patient relationship in mobile consultation: a perspective from computer-mediated communication. Int J Environ Res Public Health 2020 Apr 9;17(7):2579. [doi: 10.3390/ijerph17072579] [Medline: 32283741]
22. Street RL, Gordon H, Haidet P. Physicians' communication and perceptions of patients: is it how they look, how they talk, or is it just the doctor? Soc Sci Med 2007 Aug;65(3):586-598. [doi: 10.1016/j.socscimed.2007.03.036] [Medline: 17462801]

XSL•FO
RenderX

23. Alkureishi MA, Lee WW, Lyons M, et al. Impact of electronic medical record use on the patient-doctor relationship and communication: a systematic review. J Gen Intern Med 2016 May;31(5):548-560. [doi: 10.1007/s11606-015-3582-1] [Medline: 26786877]
24. Liu Y, Zhang X, Liu L, Lai KH. Does voice matter? Investigating patient satisfaction on mobile health consultation. Inf Process Manag 2023 Jul;60(4):103362. [doi: 10.1016/j.ipm.2023.103362]
25. Zeng Z, Xu DR, Cai Y, Gong W. Assessing quality of direct-to-consumer telemedicine in China: a cross-sectional study using unannounced standardised patients. BMJ Qual Saf 2024 Aug 16;33(9):609-612. [doi: 10.1136/bmjqs-2024-017072] [Medline: 38688711]
26. Gong X, Hou M, Guo R, Feng XL. Investigating the relationship between consultation length and quality of tele-dermatology E-consults in China: a cross-sectional standardized patient study. BMC Health Serv Res 2022 Sep 22;22(1):1187. [doi: 10.1186/s12913-022-08566-2] [Medline: 36138410]
27. Peeters KMM, Reichel LAM, Muris DMJ, Cals JWL. Family physician-to-hospital specialist electronic consultation and access to hospital care: a systematic review. JAMA Netw Open 2024 Jan 2;7(1):e2351623. [doi: 10.1001/jamanetworkopen.2023.51623] [Medline: 38214930]
28. Cheng TC, Yip W. Policies, progress, and prospects for internet telemedicine in China. Health Syst Reform 2024 Dec 17;10(2):2389570. [doi: 10.1080/23288604.2024.2389570] [Medline: 39437237]
29. Geng J, Li R, Wang X, et al. Exploring the lack of continuity of care in older cancer patients under China's "integrated health system" reform. Age Ageing 2024 Oct 1;53(10):afae213. [doi: 10.1093/ageing/afae213] [Medline: 39373574]
30. Yip W, Fu H, Jian W, et al. Universal health coverage in China part 1: progress and gaps. Lancet Public Health 2023 Dec;8(12):e1025-e1034. [doi: 10.1016/S2468-2667(23)00254-2] [Medline: 38000882]
31. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
32. Yan X, Wang J, Chau M. Customer revisit intention to restaurants: evidence from online reviews. Inf Syst Front 2015 Jun;17(3):645-657. [doi: 10.1007/s10796-013-9446-5]
33. Alegre J, Cladera M. Analysing the effect of satisfaction and previous visits on tourist intentions to return. Eur J Mark 2009 May 29;43(5/6):670-685. [doi: 10.1108/03090560910946990]
34. Woo S, Choi M. Medical service quality, patient satisfaction and intent to revisit: case study of public hub hospitals in the Republic of Korea. PLoS ONE 2021;16(6):e0252241. [doi: 10.1371/journal.pone.0252241] [Medline: 34181662]
35. Glock H, Jakobsson U, Borgström Bolmsjö B, Milos Nymberg V, Wolff M, Calling S. eVisits to primary care and subsequent health care contacts: a register-based study. BMC Prim Care 2024 Aug 12;25(1):297. [doi: 10.1186/s12875-024-02541-y] [Medline: 39135150]
36. Zeltzer D, Einav L, Rashba J, Balicer RD. The impact of increased access to telemedicine. J Eur Econ Assoc 2024 Apr 3;22(2):712-750. [doi: 10.1093/jeea/jvad035]
37. Uscher-Pines L, Mehrotra A. Analysis of Teladoc use seems to indicate expanded access to care for patients without prior connection to a provider. Health Aff (Millwood) 2014 Feb;33(2):258-264. [doi: 10.1377/hlthaff.2013.0989] [Medline: 24493769]
38. Duseja R, Bardach NS, Lin GA, et al. Revisit rates and associated costs after an emergency department encounter: a multistate analysis. Ann Intern Med 2015 Jun 2;162(11):750-756. [doi: 10.7326/M14-1616] [Medline: 26030633]
39. Ye H, Yi J. Patient-physician race concordance, physician decisions, and patient outcomes. Rev Econ Stat 2023 Jul 11;105(4):766-779. [doi: 10.1162/rest_a_01236]
40. Chan DC. Teamwork and moral hazard: evidence from the emergency department. J Pol Econ 2016 Jun;124(3):734-770. [doi: 10.1086/685910]
41. Chan DC. The efficiency of slacking off: evidence from the emergency department. Econometrica 2018;86(3):997-1030. [doi: 10.3982/ECTA13565]
42. Sprecher S. Initial interactions online-text, online-audio, online-video, or face-to-face: effects of modality on liking, closeness, and other interpersonal outcomes. Comput Human Behav 2014 Feb;31:190-197. [doi: 10.1016/j.chb.2013.10.029]
43. Examples of physician quality measures for consumers. Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/talkingquality/measures/setting/physician/examples.html [accessed 2026-02-28]
44. Quigley DD, Reynolds K, Dellva S, Price RA. Examining the business case for patient experience: a systematic review. J Healthc Manag 2021;66(3):200-224. [doi: 10.1097/JHM-D-20-00207] [Medline: 33960966]
45. Saini GK, Matta JK. Service quality and re-visit intention: a comparative study of charitable and corporate hospitals. J Health Manag 2014 Dec;16(4):595-612. [doi: 10.1177/0972063414548562]
46. Juwaheer TD, Kassean H. Exploring quality perceptions of health care operations: a study of public hospitals of Mauritius. J Hosp Mark Public Relations 2006;16(1-2):89-111. [doi: 10.1300/J375v16n01_07] [Medline: 17194688]
47. Nundy S, Cooper LA, Mate KS. The quintuple aim for health care improvement: a new imperative to advance health equity. JAMA 2022 Feb 8;327(6):521-522. [doi: 10.1001/jama.2021.25181] [Medline: 35061006]
48. Kruse CS, Krowski N, Rodriguez B, Tran L, Vela J, Brooks M. Telehealth and patient satisfaction: a systematic review and narrative analysis. BMJ Open 2017 Aug 3;7(8):e016242. [doi: 10.1136/bmjopen-2017-016242] [Medline: 28775188]

XSL•FO
RenderX

49. Olsen SO. Comparative evaluation and the relationship between quality, satisfaction, and repurchase loyalty. J Acad Mark Sci 2002 Jan 1;30(3):240-249. [doi: 10.1177/0092070302303005]
50. Heckman JJ. Sample selection bias as a specification error. Econometrica 1979 Jan;47(1):153-161. [doi: 10.2307/1912352]
51. Online review statistics: the definitive list. Luisa Zhou. URL: https://luisazhou.com/blog/online-review-statistics [accessed 2024-12-30]
52. Mittal V, Kamakura WA. Satisfaction, repurchase intent, and repurchase behavior: Investigating the moderating effect of customer characteristics. J Mark Res 2001 Feb;38(1):131-142. [doi: 10.1509/jmkr.38.1.131.18832]
53. Curtis T, Abratt R, Rhoades D, Dion P. Customer loyalty, repurchase and satisfaction: a meta-analytical review. J Consum Satisfaction, Dissatisfaction Compl Behav 2011;24:1-26 [FREE Full text]
54. Petrie S, Laur C, Rios P, et al. Quality measures of virtual care in ambulatory healthcare environments: a scoping review. BMJ Open 2024 Apr 2;14(4):e078214. [doi: 10.1136/bmjopen-2023-078214] [Medline: 38569716]
55. Wu DC, Zhao X, Wu J. Online physician-patient interaction and patient satisfaction: empirical study of the internet hospital service. J Med Internet Res 2023 Aug 24;25:e39089. [doi: 10.2196/39089] [Medline: 37616031]
56. Namgung M, Lee DH, Bae SJ, et al. A comparison of emergency department revisit rates of pediatric patients between pre-COVID-19 and COVID-19 periods. Children (Basel) 2022 Jul 4;9(7):1003. [doi: 10.3390/children9071003] [Medline: 35883987]
57. Juang WC, Cai ZX, Chen CM, You ZH. Assessing revisit risk in emergency department patients: machine learning approach. JMIR AI 2025 Aug 7;4:e74053. [doi: 10.2196/74053] [Medline: 40773678]
58. Meyer SB, Brown P, Calnan M, et al. Development and validation of the trust in multidimensional healthcare systems scale (TIMHSS). Int J Equity Health 2024 May 8;23(1):94. [doi: 10.1186/s12939-024-02162-y] [Medline: 38720303]
59. Greene J, Samuel-Jakubos H. Building patient trust in hospitals: a combination of hospital-related factors and health care clinician behaviors. Jt Comm J Qual Patient Saf 2021 Dec;47(12):768-774. [doi: 10.1016/j.jcjq.2021.09.003] [Medline: 34654668]
60. Taylor LA, Nong P, Platt J. Fifty years of trust research in health care: a synthetic review. Milbank Q 2023 Mar;101(1):126-178. [doi: 10.1111/1468-0009.12598] [Medline: 36689251]
61. Hutchinson CL, Curtis K, McCloughen A. Characteristics of patients who return unplanned to the ED, and factors that contribute to their decision to return: integrated results from an explanatory sequential mixed methods inquiry. Australas Emerg Care 2024 Mar;27(1):71-77. [doi: 10.1016/j.auec.2023.08.004] [Medline: 37741746]
62. AlOsaimi RB, Aljehani NM, Aljuhani T. Examining the reasons for patient no-show in an outpatient adult rehabilitation department in King Abdulaziz Medical City, Riyadh: a cross-sectional study. Health Sci Rep 2025 May;8(5):e70783. [doi: 10.1002/hsr2.70783] [Medline: 40309620]
63. Yang Z, Velasco Vizcaíno F, Jin M. Enhancing patient wellbeing through telemedicine services: the impact of cognitive and affective service quality ratings and physicians' long-term orientation. Int J Inf Manage 2025 Jun;82:102873. [doi: 10.1016/j.ijinfomgt.2025.102873]
64. Hollebeek LD, Macky K. Digital content marketing's role in fostering consumer engagement, trust, and value: framework, fundamental propositions, and implications. J Interact Mark 2019 Feb;45(1):27-41. [doi: 10.1016/j.intmar.2018.07.003]
65. Leahy W, Sweller J. Cognitive load theory and the effects of transient information on the modality effect. Instr Sci 2016 Feb;44(1):107-123. [doi: 10.1007/s11251-015-9362-9]
66. Arrow KJ. Uncertainty and the welfare economics of medical care. Am Econ Rev 1963;53(5):941-973 [FREE Full text]
67. Meng F, Liu Y, Zhang X, Liu L. General knowledge-sharing and patient engagement in online health communities: an inverted U-shaped relationship. J Knowl Manag 2024 Mar 18;28(3):763-788. [doi: 10.1108/JKM-12-2022-0986]

Abbreviations

AMC: academic medical center
AME: average marginal effect
HIS: hospital information system
IRB: Institutional Review Board
JARS: Journal Article Reporting Standards

XSL•FO
RenderX

Edited by S Brini; submitted 03.Nov.2025; peer-reviewed by EP Greenup, R Zhang; revised version received 03.Feb.2026; accepted 18.Feb.2026; published 26.Mar.2026.

Please cite as:
Wang Y, Wang S, Zhang K, Liu Z, Ma Q, Ji H, Hou Z, Liu TX, Xu X, Wu X, Jin C
Association Between Physician Communication Features and Patient Outcomes in Telemedicine: Retrospective Cross-Sectional Observational Study
J Med Internet Res 2026;28:e86977
URL: https://www.jmir.org/2026/1/e86977
doi:10.2196/86977

© Yipei Wang, Shu Wang, Ke Zhang, Zhijie Liu, Qingbian Ma, Hong Ji, Zheng Hou, Tracy Xiao Liu, Xuedong Xu, Xinxia Wu, Changxiao Jin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Primary Health in Rwanda: Qualitative Study of User Experiences and Implementation Lessons From Babyl's Telemedicine Platform

Sabine Musange Furere[1], MD; Gashaija Absolomon[1], MPH; Nathalie Umutoni[1], MD; Stella Matutina Umuhoza[2,3], MPH; Felix K Rubuga[1,2], PhD, MPH; Yvonne Delphine Nsaba-Uwera[4], MSN (MWH); Patrick Singa Muhoza[5], MPH; Regis Hitimana[2,6], PhD; Jeanine Condo[1,2,7], PhD, MD; James Humuza[2], PhD, MD

[1]Centre for Impact, Innovation and Capacity Building for Health Information Systems and Nutrition (CIIC-HIN), Kigali, Rwanda

[2]School of Public Health, University of Rwanda, KK 19 Ave, Kigali, Rwanda

[3]Partners In Health - Inshuti Mu Buzima, Kigali, Rwanda

[4]School of Nursing and Midwifery, University of Rwanda, Kigali, Rwanda

[5]Irembo, Kigali, Rwanda

[6]Rwanda Social Security Board (RSSB), Kigali, Rwanda

[7]Tulane School of Public Health and Tropical Medicine, New Orleans, LA, United States

Corresponding Author:
James Humuza, PhD, MD
School of Public Health, University of Rwanda, KK 19 Ave, Kigali, Rwanda

Abstract

Background: Digital health innovations address health care accessibility challenges in low- and middle-income countries. Babyl, Rwanda's largest telemedicine platform, reached 450 of 510 health facilities and enrolled 2 million patients before halting in September 2023 for system redesign. Limited research has explored implementation experiences and user perspectives that influenced its sustainability.

Objective: This study aims to explore user experiences and implementation lessons from Babyl's digital health platform, examining drivers that supported or hindered adoption and scale-up. This qualitative study uniquely examines the lived experiences of diverse stakeholders, active users, lapsed users, nonusers, health care providers, and Babyl agents to understand implementation challenges that contributed to the platform's halt.

Methods: A qualitative, cross-sectional study used 20 focus group discussions (FGDs) and 32 key informant interviews (KIIs) across 12 health centers in ten districts with diverse utilization rates, geographic locations, and Babyl agent availability. FGDs captured collective community perspectives while KIIs provided in-depth individual experiences, enabling data triangulation. FGDs included active users, lapsed users, registered nonusers, and eligible nonregistrants. KIIs involved health center heads, health care providers, and Babyl agents. Data were analyzed using thematic analysis following Braun and Clarke's framework. Data saturation was achieved when no new themes emerged from the last 3 FGDs and 5 KIIs. All transcripts were validated through member checking with a subset of participants, and intercoder reliability was established with a Cohen kappa of 0.82 across 2 independent coders.

Results: Five themes emerged: (1) knowledge and perceptions of digital health, (2) enablers and barriers to utilization, (3) experience and satisfaction, (4) benefits, and (5) improvement suggestions. Participants held positive perceptions of digital health for improving access and reducing wait times. Key enablers included qualified providers, convenience, privacy, and Babyl agents. Major barriers included negative perceptions of remote care quality, service delays, limited digital literacy, device access challenges, and inadequate health facility integration. Users reported high satisfaction with consultations but experienced process confusion. Patient and provider perspectives diverged: patients emphasized convenience, while providers expressed concerns about diagnostic limitations without physical examination. Digital literacy and smartphone access were pronounced barriers among rural and older participants. Recommendations included community mobilization, universal agent deployment, expanded coverage, and sustainable financing.

Conclusions: Multiple implementation challenges at individual, community, health system, and policy levels contributed to Babyl's discontinuation. Critical lessons include the importance of genuine health system integration, sustainable financing, stakeholder engagement, and gradual scaling. Findings provide insights for Rwanda's health sector digitalization and other African nations investing in telemedicine platforms.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e84832) doi:10.2196/84832

KEYWORDS

digital health; telemedicine; implementation science; qualitative research; Rwanda; low- and middle-income countries; health system integration

Introduction

Digital health innovations have emerged as transformative solutions to address health care accessibility challenges, particularly in low- and middle-income countries (LMICs) where geographic barriers, provider shortages, and infrastructure limitations impede effective health care delivery [1]. The World Health Assembly of the World Health Organization (WHO) unanimously adopted the use of telehealth in 2005 as a means of improving access to specialized health care services and contributing towards achieving Universal Health Coverage goals [2,3]. The expansion of mobile phone usage in emerging markets has created unprecedented opportunities for telemedicine platforms to revolutionize health care access and convenience, offering remote consultation services that can bridge the gap between patients and health care providers [4,5].

Rwanda's health care landscape exemplifies both the potential and challenges of implementing digital health solutions at scale. The country's well-structured health care system operates through a hierarchical model, with 499 health centers (HCs) serving as the backbone of primary health care delivery, addressing approximately 80% of the disease burden at the peripheral level [6,7]. Each HC provides a package of services, including curative, preventive, promotional, and rehabilitation services, while supervising health posts and community health workers in their catchment areas. This systematic approach to health care delivery created a conducive environment for digital health integration.

Babyl, an innovative digital health care platform, was introduced in Rwanda with the mission to transform health care service delivery through digital consultations facilitated by experienced health care providers. Leveraging Rwanda's high mobile phone penetration rate, Babyl's platform connected citizens with health care professionals through artificial intelligence–powered chatbots that triaged patients' medical concerns, offered treatment recommendations, and facilitated appointments with remote physicians based on symptom severity [7,8]. The platform's integration with Rwanda's existing health care infrastructure represented a significant commitment to leveraging technology for improved health outcomes, underscored by the Government of Rwanda's initial 10-year partnership agreement with the platform [8].

The scale of Babyl's implementation was remarkable, reaching 450 out of 510 primary health facilities across Rwanda, representing over 80% of the primary health care market coverage before it halted in September 2023 as the system is being redesigned. By August 2020, the platform had enrolled 2 million patients, with approximately half having actively used the service to seek care. The service primarily catered to Community-Based Health Insurance (CBHI) members, who comprised over 80% of Babyl's client base, alongside Rwandaise d'Assurance Maladie members and private clients. Rwandaise d'Assurance Maladie is the public medical insurance scheme that primarily covers employees in the public sector and also offers coverage to private sector employees and pensioners. It is coordinated by the Rwanda Social Security Board [9].

Despite the promising scale of implementation, the platform's halt in September 2023 for system redesign highlighted critical gaps in understanding the complex factors that influence the adoption, utilization, and sustainability of large-scale digital health interventions. While previous studies, including Dalberg's initial impact assessment, examined effects on access, equity, patient satisfaction, and cost-effectiveness compared with traditional care [7,10-12]. These evaluations did not provide insights into the deeper implementation challenges and user experiences that ultimately determine the success or failure of such platforms.

Notably, this study distinguishes itself from prior evaluations by examining the lived experiences and perspectives of diverse stakeholder groups, including active users, lapsed users, nonusers, health care providers, and Babyl agents, to understand the multifaceted barriers that contributed to the platform's discontinuation. Unlike previous quantitative assessments focused primarily on utilization metrics and clinical outcomes, this qualitative investigation explores the contextual, social, and systemic factors that shaped user experiences and implementation outcomes. This study contributes to the literature through comprehensive stakeholder inclusion, implementation-focused analysis, and multi-level examination of factors influencing digital health platform adoption and sustainability.

The complexity of implementing digital health solutions extends beyond technical capabilities to encompass individual, community, and health system factors that can either facilitate or hinder adoption. Health care providers must adapt to new diagnostic approaches without physical examination capabilities, potentially altering clinical decision-making processes and provider-patient interactions [13,14]. Patients may face challenges in communicating illness concerns and treatment preferences through digital platforms, while also navigating new technologies that require different levels of digital literacy and access to reliable communication infrastructure.

Understanding these multifaceted implementation experiences becomes crucial as countries across sub-Saharan Africa and other low- and middle-income settings continue to explore and invest in digital health solutions [15,16]. The WHO Digital Health Guidelines emphasize the importance of understanding contextual factors and user experiences in digital health implementation [17]. The COVID-19 pandemic accelerated the global adoption of telemedicine and digital care models, making

XSL•FO
RenderX

it imperative to learn from large-scale implementations like Babyl to inform future digital health initiatives [18].

This study uniquely addresses critical knowledge gaps by providing the first comprehensive, multi-stakeholder qualitative evaluation of Rwanda's Babyl telemedicine platform, explicitly contrasting perspectives across patients (active users, lapsed users, registered nonusers, and eligible nonregistrants), health care providers, HC administrators, and Babyl agents. Unlike previous evaluations focused primarily on utilization metrics or single stakeholder groups, this study systematically examines the complex interplay between individual, community, health system, and technological factors shaping implementation outcomes, while explicitly linking demographic characteristics, including age, education, geographic location, and digital literacy, to adoption patterns and user experiences. These insights extend beyond Rwanda's context to inform digital health implementation strategies across sub-Saharan Africa and other LMICs pursuing universal health coverage through technological innovations.

Methods

Study Design and Setting

This qualitative, cross-sectional study was conducted across Rwanda's health care system to explore key drivers that supported or hindered the adoption and scale-up of Babyl's digital health services in Rwanda. The study used focus group discussions (FGDs) and key informant interviews (KIIs) to capture perspectives from multiple stakeholder groups involved in or affected by the digital health platform implementation. The combination of FGDs and KIIs was strategically chosen to capture both collective community perspectives and in-depth individual experiences. FGDs enabled exploration of shared experiences, social norms, and community-level barriers through group interactions, while KIIs provided detailed accounts from health care providers and program implementers who could speak to operational challenges and system-level implementation issues. This methodological triangulation strengthened the validity and comprehensiveness of findings by allowing cross-verification of themes across different data sources and participant perspectives.

Sampling Strategy and Participant Selection

A purposive, maximum variation sampling strategy was used to ensure representation of diverse experiences and perspectives across different contexts. The study used a multi-stage sampling approach. First, HCs were purposively selected based on three criteria: (1) Babyl service utilization rates (high, medium, and low), (2) geographic distribution across Rwanda's provinces, and (3) presence or absence of Babyl agents at the facility level. This stratification ensured representation of diverse implementation contexts and user experiences.

Twelve HCs were selected across 10 districts, distributed as follows: two from the City of Kigali (urban), four from the Eastern Province (mixed urban-rural), three from the Southern Province (predominantly rural), two from the Western Province (rural), and one from the Northern Province (rural). This geographic distribution captured variations in infrastructure, digital connectivity, and socioeconomic contexts that might influence digital health service adoption.

Within each selected HC catchment area, recruitment occurred through multiple channels. HC staff identified potential participants from patient registers and community health worker networks. Community meetings were held at centrally located venues to explain the study's purpose and answer questions. Interested individuals were screened for eligibility using a structured screening form. Recruitment continued until data saturation was achieved and the target sample size was reached. For FGDs, four distinct user categories were identified:

1. Active users: CBHI members who had registered for and actively used Babyl services within the past 6 months.
2. Lapsed users: CBHI members who had registered but stopped using Babyl services.
3. Registered nonusers: CBHI members who registered for Babyl but never used the services.
4. Eligible nonregistrants: CBHI members aware of Babyl services but never registered.

Participants were eligible if they were 18 years or older, CBHI members in the selected catchment areas, and able to provide informed consent. For KII participants, eligibility criteria included being used at the selected HC or serving as a Babyl agent for at least 6 months, ensuring adequate experience with the platform implementation. For KIIs, three categories of key informants were purposively selected:

1. HC heads: Administrative leaders overseeing HC operations.
2. Health care providers: Nurses, clinical officers, and physicians who interacted with Babyl-referred patients.
3. Babyl agents: Community-based facilitators who assisted users with platform registration and navigation.

The final sample size of 192 participants (160 in FGDs and 32 in KIIs) was determined based on several criteria: (1) Data saturation confirmed when no new themes emerged from the final 3 FGDs and 5 KIIs; (2) Representation of all four user categories across diverse geographic settings; (3) Multi-site coverage, including stakeholders from all 12 HCs; (4) Methodological guidelines suggesting 15‐20 FGDs for thematic saturation in multi-site studies; and (5) Distribution of 4‐5 FGDs per user category. The sample size exceeded the minimum recommendations.

Data Collection and Procedures

Data collection was conducted between October 2022 and January 2023 by a team of 6 trained research assistants with prior qualitative research experience. All data collectors underwent a 3-day training covering qualitative interviewing techniques, research ethics, study objectives, and the semi-structured interview guides. Mock interviews were conducted to standardize data collection approaches across team members.

The research team comprised public health researchers with backgrounds in health systems strengthening, digital health, and implementation science. The principal investigator is a PhD holder with 8 years of experience in health policy research. The

XSL•FO
RenderX

study coordinator is a female MPH graduate with expertise in qualitative methods. Research assistants included 4 females and 2 males, all with bachelor's degrees in public health. No team members had prior professional relationships with Babyl or the participants.

Semi-structured interview guides were developed based on a literature review, consultation with digital health experts, and the research team's expertise in health systems and implementation science. The interview guides were adapted from WHO's framework for digital health implementation research [19] and incorporated constructs from the Consolidated Framework for Implementation Research [20]. The guides were pretested with 5 users and 2 health care providers from a nonstudy HC. Based on feedback, questions were refined for clarity, cultural appropriateness, and to reduce technical jargon.

The guides covered key domains including: (1) awareness and perceptions of digital health services, (2) facilitators and barriers to adoption, (3) user experiences with service delivery processes, (4) satisfaction with care quality, (5) perceived benefits and challenges, and (6) recommendations for improvement. Content validity was established through expert review by 3 implementation science researchers and 2 clinicians with telemedicine experience. Interview guides are provided as Multimedia Appendices 1-4.

Each FGD included 6 - 10 participants of the same user category to promote open discussion and ensure homogeneity within groups. FGDs were conducted in Kinyarwanda (local language) at community meeting spaces near the HCs, lasted 60 - 90 minutes, and were audio-recorded with participant consent. Two team members facilitated each FGD: one as moderator and one as note-taker to ensure comprehensive documentation of verbal and nonverbal data.

KIIs were conducted individually in private rooms at HCs or Babyl agent offices, lasted 45 - 60 minutes, and were also audio-recorded. Interviews were scheduled at times convenient for participants to minimize disruption to their work responsibilities.

Field notes were taken during all interviews to capture contextual information, nonverbal cues, and preliminary observations. A reflexive journal was maintained by the research team lead to document potential biases, decision-making processes, and emerging insights throughout data collection.

In addition to qualitative data, demographic and contextual information were collected from all participants using structured questionnaires administered prior to FGDs and KIIs. For FGD participants, the demographic questionnaire captured age, gender, education level, geographic location (urban/rural), smartphone ownership status, digital literacy self-assessment, health insurance type, and Babyl service utilization history. For KII participants, the questionnaire captured professional role, years of professional experience, duration of Babyl involvement, health facility location, and service coverage characteristics.

Data Analysis

All audio recordings were transcribed verbatim in Kinyarwanda within 48 hours of data collection by professional transcribers trained in qualitative transcription standards. Transcripts were then translated into English by bilingual team members. A back-translation process was used for 20% of transcripts to ensure translation accuracy and conceptual equivalence. Transcripts were anonymized, with all identifying information removed and replaced with participant codes.

Transcript quality was verified through a validation process where the lead researchers reviewed 30% of transcripts against audio recordings. Additionally, summary transcripts were shared with a subset of 15 participants for member checking to ensure accurate representation of their views.

Data were analyzed using iterative, inductive thematic analysis following Braun and Clarke's 6-phase framework [18] : (1) familiarization with data, (2) generating initial codes, (3) searching for themes, (4) reviewing themes, (5) defining and naming themes, and (6) producing the report.

Two researchers independently coded the first 4 transcripts to develop an initial codebook. Codes were compared and discussed to achieve consensus, resulting in a refined codebook with clear code definitions. This codebook was then applied to all transcripts using NVivo (version 12; Lumivero) qualitative analysis software. Intercoder reliability was assessed on 20% of transcripts coded independently by 2 researchers, achieving a Cohen kappa of 0.82, indicating strong agreement.

The coding process began with open coding, where segments of text were assigned descriptive codes. A total of 127 initial codes were identified during open coding. These codes were then grouped into broader categories through axial coding, examining relationships and patterns. Finally, selective coding identified overarching themes that captured the essence of participants' experiences and perspectives.

The analysis was conducted iteratively, with constant comparison between emerging themes and raw data. Deviant cases were actively sought and analyzed to ensure a comprehensive understanding. Regular team meetings were held to discuss emerging themes, resolve discrepancies, and refine the thematic structure.

Thematic saturation was formally assessed by monitoring theme emergence across sequential data collection phases. Data collection occurred in 3 waves. When analysis of the final 3 FGDs and 5 KIIs revealed no new themes, data collection was concluded.

Data credibility was enhanced through (1) methodological triangulation using both FGDs and KIIs, (2) investigator triangulation with multiple researchers involved in coding and analysis, (3) participant validation through member checking, (4) peer debriefing sessions with external qualitative research experts, and (5) maintaining a detailed audit trail documenting all analytical decisions.

Quantitative data were analyzed using descriptive statistics and chi-square tests to examine associations between demographic characteristics and technology access patterns. Statistical significance was set at $P<.05$. These quantitative findings are presented alongside qualitative data to provide a comprehensive understanding of barriers.

XSL•FO
RenderX

Ethical Considerations

Ethical approval for this study was obtained from the University of Rwanda, College of Medicine and Health Sciences Institutional Review Board (approval number 335/CMHS IRB/2022). Additional administrative clearance was granted by the Rwanda Ministry of Health and participating district health offices. The study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki.

Written informed consent was obtained from all participants after providing detailed information about the study purpose, procedures, risks, benefits, and their right to withdraw at any time without consequences. For participants with limited literacy, the consent form was administered orally in the presence of a witness, and verbal consent was audio-recorded.

FGD participants received 5000 Rwandan francs (approximately US $5) as compensation for transportation and time. KII participants (health care providers, administrators, and Babyl agents) were interviewed during working hours with supervisor approval and did not receive monetary compensation. Refreshments were provided during all FGDs and interviews.

Participant privacy and confidentiality were ensured by conducting interviews in private settings, anonymizing transcripts, and storing audio recordings and transcripts on password-protected computers accessible only to the research team. Data will be securely stored for 5 years before permanent deletion. No adverse events were reported during the study.

Results

Overview

The analysis of 20 FGDs and 32 KIIs revealed five major themes that emerged from combining client and health facility personnel responses: (1) knowledge and perceptions of digital health services, (2) enablers and barriers to using digital services, (3) experience and satisfaction using Babyl digital services, (4) benefits of using digital services, and (5) suggestions to improve uptake of digital services. The KIIs also provided opportunities to better understand Babyl's operations related to agent recruitment processes, client facilitation experiences, and engagement with health facilities.

Participant Characteristics

The demographic distribution of FGD and KII participants is summarized in Tables 1 and 2, respectively.

Table . Demographic characteristics of focus group discussion participants (n=160).

Characteristic	Active users (n=40)	Lapsed users (n=40)	Registered nonusers (n=40)	Eligible nonregistrants (n=40)
Sex, n (%)				
Female	25 (62.5)	27 (67.5)	28 (70.0)	21 (52.5)
Male	15 (37.5)	13 (32.5)	12 (30.0)	19 (47.5)
Age group (years), n (%)				
18 - 30	12 (30.0)	8 (20.0)	6 (15.0)	10 (25.0)
31 - 50	26 (65.0)	28 (70.0)	30 (75.0)	28 (70.0)
>50	2 (5.0)	4 (10.0)	4 (10.0)	2 (5.0)
Age (years), mean (SD)	35.2 (8.4)	36.8 (9.1)	38.4 (8.7)	42.1 (10.2)
Education level, n (%)				
Primary or less	10 (25.0)	15 (37.5)	25 (62.5)	28 (70.0)
Secondary	24 (60.0)	20 (50.0)	13 (32.5)	10 (25.0)
Tertiary	6 (15.0)	5 (12.5)	2 (5.0)	2 (5.0)
Location, n (%)				
Urban	11 (27.5)	9 (22.5)	10 (25.0)	6 (15.0)
Rural	29 (72.5)	31 (77.5)	30 (75.0)	34 (85.0)
Phone ownership, n (%)				
Own phone	38 (95.0)	36 (90.0)	33 (82.5)	27 (67.5)
Shared/no phone	2 (5.0)	4 (10.0)	7 (17.5)	13 (32.5)
Insurance type, n (%)				
CBHI[a]	35 (87.5)	32 (80.0)	33 (82.5)	34 (85.0)
Other/None	5 (12.5)	8 (20.0)	7 (17.5)	6 (15.0)

[a]CBHI: Community-Based Health Insurance.

XSL•FO
RenderX

Table . Characteristics of key informant interview participants (n=32).

Characteristic	Health center heads (n=6)	Health care providers (n=18)	Babyl agents (n=8)
Sex, n (%)			
Female	2 (33.3)	10 (55.6)	5 (62.5)
Male	4 (66.7)	8 (44.4)	3 (37.5)
Age group (years), n (%)			
25 - 35	1 (16.7)	6 (33.3)	4 (50.0)
36 - 45	3 (50.0)	9 (50.0)	3 (37.5)
>45	2 (33.3)	3 (16.7)	1 (12.5)
Age (years), mean (SD)	42.5 (7.2)	38.4 (6.8)	34.2 (5.9)
Professional role, n (%)			
Nurses	—[a]	12 (66.7)	—
Lab technicians	—	4 (22.2)	—
Pharmacists	—	2 (11.1)	—
Facility-based	6 (100)	—	8 (100)
Years of experience, n (%)			
<2	0 (0)	2 (11.1)	3 (37.5)
2 - 5	1 (16.7)	8 (44.4)	5 (62.5)
6 - 10	3 (50.0)	6 (33.3)	0 (0)
>10	2 (33.3)	2 (11.1)	0 (0)
Years of experience, mean (SD)	10.2 (4.8)	6.3 (4.2)	2.8 (1.5)
Health center type, n (%)			
Urban	2 (33.3)	7 (38.9)	3 (37.5)
Rural	4 (66.7)	11 (61.1)	5 (62.5)
Babyl experience (years), n (%)			
<1	1 (16.7)	4 (22.2)	1 (12.5)
1 - 2	2 (33.3)	8 (44.4)	3 (37.5)
>2	3 (50.0)	6 (33.3)	4 (50.0)
Babyl experience (years), mean (SD)	2.4 (1.1)	1.8 (0.9)	2.1 (0.8)

[a]Not applicable.

A total of 192 participants were enrolled in this study, comprising 160 individuals in 20 FGDs and 32 key informants in individual interviews (KIIs). Among FGD participants, 98 (61.3%) were female, and 62 (38.7%) were male. The majority (112/160, 70.0%) were aged 31 - 50 years, with 36 (22.5%) participants aged 18 - 30 years, and 12 (7.5%) participants older than 50 years.

Demographic characteristics significantly influenced digital health adoption patterns. Twenty-eight younger participants (18 - 30 y) demonstrated greater digital literacy and smartphone ownership (76.5%) compared with 4 participants older than 50 years (34.8%), facilitating easier platform adoption. Among the 12 participants older than 50 years, 11 (91.7%) participants reported difficulty navigating the digital platform. Female participants more frequently emphasized privacy as a key benefit, particularly for reproductive health consultations. Male participants more commonly highlighted time efficiency and workplace convenience. Rural participants (92/160, 57.5% of FGDs) faced compounded barriers: lower smartphone ownership (34.8% vs 76.5% in urban areas), poorer network connectivity (78.3% reported “poor” or “very poor” connection vs 23.5% urban), and limited Babyl agent presence (available in only 2 of 7 rural HCs vs all 5 urban centers).

Among active users (n=40), most were CBHI members (35/40, 87.5%), aged 25 - 45 years, predominantly female (25/40, 62.5%), rural (29/40, 72.5%), and phone owners (38/40, 95%), with higher digital literacy and positive digital health experiences. Lapsed users (n=40) had similar demographic profiles but reported technology challenges and service dissatisfaction. Registered nonusers (n=40) were mainly female (28/40, 70%), older (mean age of 38 y), with lower digital literacy (25/40, 62.5% had only primary education), and largely

XSL•FO
RenderX

rural (30/40, 75%). Eligible nonregistrants (n=40) were the oldest group (mean age of 42 y) with the lowest digital literacy and phone ownership (27/40, 67.5%), often relying on shared devices.

Table 2 shows that key informants had substantial professional experience and adequate exposure to Babyl services to provide informed perspectives. HC heads had the longest professional experience (mean 10.2 y), providing institutional memory and leadership perspective on digital health integration. Health care providers, predominantly nurses (12/18, 66.7%), had moderate experience (mean 6.3 y) and represented frontline staff managing the interface between digital and traditional services. Babyl agents were younger (mean age 34.2 y), had shorter professional tenure (mean 2.8 y), but substantial Babyl-specific experience (mean 2.1 y), positioning them to understand both technical platform issues and community-level barriers. The predominance of rural HCs (61.1% - 66.7% across groups) ensured representation of contexts where implementation challenges were most pronounced. Variation in Babyl experience duration (1 - 2+ years across groups) captured perspectives from both early implementation and more mature operational phases.

Table 3 presents key findings organized by stakeholder group, revealing both convergent and divergent perspectives across patients/users, health care providers, HC heads, and Babyl agents. While all stakeholders recognized Babyl's potential to improve health care access, critical disconnects emerged: patients emphasized front-end barriers, including registration complexity and agent availability; health care providers focused on back-end challenges including workflow integration and lack of access to consultation records; HC heads highlighted coordination difficulties and unclear protocols, and Babyl agents identified gaps in training, equipment, and support systems. These divergent priorities reflect each group's position within the implementation ecosystem and underscore the importance of multi-stakeholder engagement in digital health design and implementation.

XSL•FO
RenderX

Table . Key findings by stakeholder group.

Theme	Patients/Users (n=160)	Health care providers (n=12)	Health center heads (n=6)	Babyl agents (n=14)
Knowledge and Perceptions	Mixed awareness levels; valued convenience but had an incomplete understanding of coverage and eligibility; trust concerns about remote diagnosis quality	Skepticism about diagnostic accuracy without physical examination; concerns about consultation quality and clinical decision-making limitations	Recognized potential for expanded access but concerned about facility burden, resource allocation, and unclear role in digital health integration	Strong belief in service value and community health impact; frustration when unable to assist users due to system limitations or lack of support
Enablers	Convenience and time savings; privacy for sensitive health issues; cost savings from avoided travel; access to qualified doctors; medication prescription and dispensing	Potential for reduced outpatient department congestion; access to physician consultation for complex cases; opportunity to refer patients requiring specialist input	Expanded health care access for the catchment population; potential to reduce referrals to district hospitals; alignment with national health coverage goals	Direct patient interaction and community engagement; satisfaction from helping users access health care; visible community health impact when services worked well
Barriers	Digital literacy gaps and limited smartphone access; agent absence at health centers; registration process complexity; poor network connectivity (especially rural); service coverage exclusions (chronic diseases, pediatrics)	No access to Babyl consultation records; medication stock-outs disrupting prescribed treatments; workflow disruptions from parallel service delivery; lack of feedback mechanisms to learn from consultations; diagnostic uncertainty without physical examination	Unclear integration protocols and coordination mechanisms; additional workload without corresponding resources; medication supply chain issues; ambiguous accountability when problems arose	Inadequate technical and patient education training; equipment failures and resource constraints; inconsistent presence at facilities; blamed for system issues beyond their control; limited escalation support
Experience and Satisfaction	High satisfaction when services functioned properly; frustration with technical issues, delays, and process complexity; pronounced age-related challenges (older adults required agent assistance 100%); lapsed users cited process friction exceeding perceived benefit	Frustration with incomplete patient information arriving at facilities; inability to provide continuity of care without consultation histories; diagnostic uncertainty affecting clinical confidence; felt Babyl operated as a parallel rather than an integrated system	Difficulty coordinating between Babyl and facility-based services; unclear accountability and escalation pathways when problems arose; concerned about sustainability and long-term resource implications	Satisfaction when helping users successfully; demotivation from facing blame for system failures; limited technical support for troubleshooting; felt undervalued despite frontline role
Recommendations	Universal agent availability at all health centers; simplified registration process; expanded service coverage (chronic diseases, pediatrics); improved rural network infrastructure; clearer information about eligibility and services	Full health management information system integration; access to consultation histories and prescribed treatments; regular coordination meetings with Babyl; clear referral and feedback protocols; assurance of medication availability	Clear operational guidelines and integration protocols; regular stakeholder coordination meetings; dedicated resources for digital health integration; sustainable financing clarity and commitment	Enhanced training (technical troubleshooting + patient education); reliable equipment and backup systems; consistent facility presence and defined roles; clear escalation procedures; recognition and support for frontline work

Theme 1: Knowledge and Perceptions of Digital Health Services

Overview

Three common subthemes emerged: (1) sources of awareness about Babyl services, (2) perceptions of service quality and legitimacy, and (3) understanding of eligibility and coverage. These subthemes reveal how information dissemination and initial perceptions shaped enrollment decisions.

Sources of Awareness

Awareness of Babyl services came primarily through 3 channels. Health facility visits were the most common source, where patients learned about services during routine consultations or registrations. Radio advertisements were mentioned in 7 of 8 urban FGDs but only 3 of 8 rural FGDs, indicating geographic disparities in media reach. One participant stated,

> *I heard about it from the radio. They said we can talk to doctors without going to the health center.* [active user, FGD participant #3]

Word-of-mouth through community networks was particularly important in rural areas, where 6 of 8 FGDs cited neighbors or community health workers as information sources. Registered nonusers and eligible nonregistrants frequently mentioned incomplete or unclear information as a reason for nonengagement. As one participant explained,

> *I heard about it, but I wasn't sure if it was free with my insurance.* [eligible nonregistrant, FGD participant #8]

Perceptions of Quality and Legitimacy

Perceptions of service quality varied substantially across user categories. Active users expressed confidence in the

XSL•FO
RenderX

qualifications of Babyl doctors, emphasizing their ability to diagnose and prescribe appropriately. A typical comment was,

> *They have doctors, real doctors who can prescribe medicine.* [registered nonuser, FGD participant #2]

However, skepticism was common among nonregistrants and registered nonusers, who questioned whether remote consultations could match in-person care. This skepticism was more pronounced in rural areas (6 of 8 rural FGDs) compared with urban areas (3 of 8 urban FGDs) and among older participants (>50 y).

While 89% of CBHI members had heard of Babyl services, only 34% could accurately describe the service scope and eligibility. This awareness-understanding gap created adoption barriers across user groups. Active users learned through trial and error with agent assistance, while registered nonusers held persistent misunderstandings (eg, believing services required payment). Provider skepticism about remote diagnosis reinforced community doubts, creating a negative cycle where patients avoided services that providers did not endorse. This demonstrates that awareness campaigns alone are insufficient; comprehensive education addressing service scope, appropriate use cases, and complementarity with facility-based care is essential.

Theme 2: Enablers and Barriers to Digital Service Utilization

Overview

Four subthemes emerged operating at different levels: (1) individual-level factors, including technology access and digital literacy, (2) community-level factors such as social norms and trust, (3) health system factors, including integration with existing services, and (4) technological infrastructure. Findings are organized by stakeholder perspectives to highlight divergent experiences.

Patient Perspectives on Enablers

Patients, particularly active users, identified several key enablers. Convenience and time savings were universally emphasized, with one participant noting,

> *I don't have to travel to the health center or wait in line.* [active user, FGD participant #5]

Privacy emerged as a significant enabler, especially for female participants (mentioned in 68% of female vs 42% of male participant comments) and for sensitive health concerns. As one woman explained,

> *I can get advice without anyone knowing my problem.* [active user, FGD participant #9]

For working-age adults (31 - 50 y), the ability to consult without missing work was critical.

> *I don't have to miss work to see a doctor.* [active user, FGD participant #10]

Access to qualified doctors and prescribed medications was an additional enabler, particularly valued in rural areas where specialist access is limited.

Patient Perspectives on Barriers

Demographic and digital literacy data were collected using a structured questionnaire administered to all FGD participants (n=160). The questionnaire captured age, gender, education level, smartphone ownership, phone sharing patterns, and self-reported digital literacy. Quantitative data were analyzed using descriptive statistics and chi-square tests to examine associations between demographic characteristics and technology access patterns. Statistical significance was set at $P<.05$. These quantitative findings are presented alongside qualitative data to provide a comprehensive understanding of barriers.

Individual-level barriers were substantial and strongly correlated with demographic characteristics. Technology access challenges included a lack of smartphone ownership (affecting 65.2% of adults older than 50 y old vs 23.5% of 18 - 30 y olds), reliance on shared devices (creating privacy concerns), and device limitations. One lapsed user described,

> *My phone is very old and slow. The app takes too long to open.* [lapsed user, FGD participant #2]

Smartphone ownership ranged from 76.5% (18 - 30 y) to 34.8% (>50 y), $P<.001$. Rural participants older than 45 years had lower ownership (23.8%) than urban counterparts (71.4%, $P<.001$). Educational attainment predicted ownership: 85.7% secondary-educated versus 38.9% primary-educated owned smartphones ($P<.001$). Female participants were 2.3× more likely to share phones (62.5% vs 27.4% male, $P<.001$), creating privacy barriers for reproductive health discussions.

Digital literacy barriers were pronounced, with 91.7% of participants older than 50 years reporting difficulty using smartphone apps compared with 22.2% of 18 - 30-year-olds. An eligible nonregistrant stated,

> *I don't know how to use these apps. I need someone to help me.* [eligible nonregistrant, FGD participant #8]

Digital literacy showed a 4-fold age gradient: 91.7% of those older than 50 years needed continuous assistance versus 22.2% of 18 - 30 years ($P<.001$). Educational patterns were equally pronounced: 89.7% primary-educated versus 23.4% secondary-educated required assistance ($P<.001$). All 8 participants, both older than 50 years and primary-educated, required help for every interaction, averaging 4.2 registration attempts.

Process complexity deterred many potential users, particularly the registration process, which required multiple steps, including phone verification, insurance validation, and profile setup.

> *I tried to register but got confused with all the steps.* [registered nonuser, FGD participant #6]

Trust concerns centered on data privacy, consultation quality without physical examination, and medication dispensing procedures. Service coverage limitations, particularly the exclusion of chronic diseases (diabetes, hypertension) and pediatric care, created significant barriers. Network connectivity issues disproportionately affected rural users, with 78.3%

XSL•FO
RenderX

reporting poor or intermittent connections compared with 23.5% of urban users.

> *The network doesn't work well in my village.* [eligible nonregistrant, FGD participant #3]

Health Care Provider Perspectives

Health care providers identified both enablers and barriers from their operational standpoint. Key enablers included reduced outpatient department (OPD) congestion for simple cases and access to physician expertise for complex cases. As one provider explained,

> *For patients who only need simple consultations, this reduces our workload.* [health care provider, KII]

However, providers raised concerns about diagnostic limitations, noting,

> *Without physical examination, it's difficult to make accurate diagnoses.* [health center head, KII]

Workflow disruptions occurred when Babyl patients arrived at HCs without proper documentation or when consultation records were not accessible to facility staff.

> *We don't know what the doctor prescribed or why patients were referred.* [health care provider, KII]

Medication stock-outs created frustration for both providers and patients.

> *Sometimes the medication is out of stock when patients come to collect it.* [health care provider, KII]

The lack of feedback mechanisms meant providers could not learn from Babyl consultations or understand clinical outcomes.

Babyl Agent Perspectives

Babyl agents identified community relationships and collaborative HC staff as key enablers, but faced significant operational barriers. Equipment inadequacy was a persistent challenge.

> *Sometimes the equipment doesn't work, and I have to use my own phone.* [Babyl agent, KII]

Agents experienced blame for system failures beyond their control, affecting their motivation and community relationships.

> *When the system fails, or verification takes long, patients blame me.* [Babyl agent, KII]

Insufficient training on both technical troubleshooting and patient education limited their effectiveness. The absence of dedicated agents at many HCs (particularly rural facilities) was identified as a critical gap, with patients having no one to assist with registration or resolve issues.

Barriers compound multiplicatively for vulnerable populations. Older rural women with primary education faced simultaneously: limited smartphone ownership, agent absence, poor connectivity, registration complexity, and service coverage gaps. These disadvantages are concentrated among populations with the greatest potential benefit. While patients emphasized front-end barriers (registration, digital literacy, agent availability), providers focused on back-end challenges (workflow disruption, information access), indicating multilevel intervention needs. That 100% of participants older than 50 years required agent assistance, yet agents were absent from 41.7% of HCs, exemplifies how self-service assumptions created insurmountable barriers for vulnerable groups.

Theme 3: Experience and Satisfaction Using Babyl Services

Overview

Three subthemes characterized user experiences: (1) registration and enrollment processes, (2) consultation quality and clinical interactions, and (3) factors influencing continued use or discontinuation. Experiences varied significantly by user category and demographic characteristics.

Patient Registration Experiences

Age was the strongest predictor of registration experience. Younger participants (18‐30 y) predominantly enrolled independently (89.5%), using smartphones and navigating the app without assistance.

> *I was able to register easily on my own using the app.* [active user, FGD participant #2]

In contrast, 100% of participants older than 50 years required agent assistance for registration.

> *The agent helped me step by step to register and call.* [active user, FGD participant #6]

This agent dependency created a critical vulnerability when agents were absent, older adults could not access services.

Registration success was age-stratified: 89.5% of 18‐30 years registered independently versus 0% of older than 50 years (100% required assistance). Educational attainment was equally predictive: 74.3% secondary-educated versus 15.6% primary-educated registered independently. Rural participants required 2.8 versus 1.3 registration attempts (urban, $P<.001$). Independent registrants were 3.7× more likely to initiate consultations (67.2% vs 18.1%), suggesting registration difficulty created psychological barriers to utilization.

Consultation Experiences and Discontinuation

Among lapsed users, the most common pattern was initial satisfaction with clinical outcomes but frustration with process complexity, leading to discontinuation.

> *The medicine worked, but the process was too confusing to use again.* [lapsed user, FGD participant #9]

Circumstantial discontinuation occurred when users' health needs fell outside service coverage, particularly for chronic disease management.

> *I tried it once for a cough, but now I just go to the health center.* [lapsed user, FGD participant #2]

Registered nonusers typically completed enrollment but never initiated consultations due to persistent trust concerns or lack of immediate need.

XSL•FO
RenderX

Provider and Agent Experiences

Health care providers reported operational challenges, including the inability to access consultation records, medication stock-outs disrupting prescribed treatments, and a lack of clinical feedback limiting learning opportunities. Babyl agents experienced satisfaction when assisting successful users but felt blamed for system failures beyond their control, including verification delays and network connectivity problems.

User experiences revealed distinct trajectories: active users progressed through initial challenges (2‐3 attempts) to routine utilization; lapsed users reported satisfaction but discontinued due to process friction; registered nonusers completed enrollment but never initiated consultations. Age predicted experience quality through process navigation capability; older adults required sustained agent assistance. Without agent availability, consultation initiation was impossible regardless of need severity. Provider experiences paralleled patient frustrations: initial enthusiasm diminished when medication stock-outs and inability to access consultation records created operational challenges.

Theme 4: Benefits of Digital Health Services

Overview

Four subthemes emerged regarding benefits: (1) access to specialized care and qualified providers, (2) time and cost savings, (3) health system efficiency improvements, and (4) innovation and learning opportunities. Benefits were emphasized differently by different stakeholder groups.

Patient-Identified Benefits

Patients most frequently emphasized access to doctors and specialists.

> *I talked to a real doctor from Kigali who understood my problem.* [active user, FGD participant #4]

This access benefit was particularly valued in rural areas where specialist consultations require traveling to district or referral hospitals. Time savings were mentioned more frequently by male participants (72%) and working-age adults than by female or older participants. Cost savings, including transport, lost work time, and informal payments, were emphasized more in rural FGDs (7 of 8) compared with urban FGDs (3 of 8).

> *I saved money on transport and time away from my business.* [active user, FGD participant #9]

Privacy and convenience benefits allowed participants to seek care for sensitive health issues and at convenient times.

Benefit valuations varied by demographics. Male participants emphasized time savings more than females (72.6% vs 48.0%), while females emphasized privacy (68.4% vs 41.9%). Age influenced perceptions: younger participants valued convenience, middle-aged participants emphasized time/cost, and older participants valued doctor access. Rural participants saved 2000‐3000 RWF (approximately US $1.90-2.86) versus urban 500‐1000 RWF (approximately US $0.48-0.95) per avoided visit, reflecting greater travel distances.

Provider and Agent-Identified Benefits

Health care providers valued the ability to refer complex cases to physicians.

> *Patients can consult doctors from Kigali for complex cases.* [health center head, KII]

Reduced OPD congestion for routine consultations allowed providers to focus on patients requiring physical examination.

> *We can focus on patients who really need physical examination.* [health care provider, KII]

However, providers noted that the lack of feedback limited their learning from consultations. Babyl agents identified community health impact and the satisfaction of facilitating access as key benefits, noting that successful users often became service advocates.

Stakeholder benefit perceptions diverged critically. Patients emphasized individual benefits (time savings, cost reduction, and privacy); providers emphasized system-level advantages (reduced congestion and specialist access). However, providers' anticipated benefits remained unrealized due to implementation gaps. Babyl operated in parallel to facility services; patients still presented for medication collection. Inability to access consultation records created fragmented care. This disconnect created sustainability challenges: patients experienced sufficient benefit to continue utilization, but providers experienced more burdens than benefits, threatening institutional buy-in essential for scaling.

Theme 5: Suggestions for Service Improvement

Overview

Four subthemes emerged regarding improvement recommendations: (1) community-level awareness and engagement strategies, (2) health facility-level infrastructure and staffing, (3) service delivery modifications, and (4) policy and sustainability considerations. Recommendations varied by stakeholder role and context.

Community-Level Recommendations

Expanded awareness campaigns using multiple channels (radio, community meetings, health facility outreach) were universally recommended.

> *They should announce it at community meetings in villages.* [eligible nonregistrant, FGD participant #1]

Participants suggested deploying community health worker mobilizers in cells and villages, particularly in rural areas. Clear, simplified information about eligibility, covered services, costs, and registration steps was emphasized across all user categories.

Health Facility-Level Recommendations

The most critical recommendation was ensuring dedicated Babyl agents at all HCs, especially rural facilities.

> *There should be an agent at every health center to help people.* [lapsed user, FGD participant #4]

Additional needs included adequate equipment (smartphones, tablets, backup power), improved network infrastructure, comprehensive agent training covering technical troubleshooting

XSL•FO
RenderX

and patient education, and better integration of Babyl consultation records with health facility systems.

Service Delivery Recommendations

Service expansion to include pediatric consultations (for caregivers) was frequently requested. Participants suggested accommodating phone sharing through mechanisms like designated family members or community phone points. Partnerships with health posts for last-mile service delivery in remote areas were proposed. Chronic disease coverage was the most common service expansion request.

> *It should cover chronic diseases like diabetes and hypertension.* [registered nonuser, FGD participant #1]

Rural network infrastructure improvement was critical.

> *The network infrastructure in rural areas needs improvement.* [Babyl agent, KII]

Simplified registration processes, medication stock management improvements, and patient feedback mechanisms were additional operational recommendations.

Policy and Sustainability Considerations

Health care providers and HC heads emphasized sustainable financing mechanisms, clear governance and accountability structures, and formal integration into national health policies and strategies. These systemic recommendations reflected concerns about program continuity and scale-up.

Demographic Patterns in Digital Health Adoption

Demographics functioned as intersecting determinants, creating compounded barriers. Older rural women with primary education faced simultaneously: smartphone nonownership (65.2%), digital illiteracy (91.7% needed assistance), poor connectivity (78.3% rural), agent absence (5 of 7 rural facilities), and service coverage gaps. Successful adopters were advantaged: 74.6% secondary-educated (vs 29.7% nonregistrants, $P<.001$), 87.5% urban, 95.0% smartphone owners, mean age 33.2 years (vs 42.1, $P<.001$). Each additional disadvantage multiplicatively reduced adoption: 87.5% utilization (zero disadvantages) versus 12.3% (3‐4 disadvantages). Digital health primarily reached already-advantaged populations, potentially exacerbating inequities.

Thematic synthesis reveals critical interdependencies among findings that explain implementation outcomes and point toward multilevel intervention requirements. Knowledge gaps identified in theme 1 directly amplified barrier impacts documented in theme 2: participants unaware that Babyl agents could provide registration assistance perceived enrollment as impossible and never attempted it, while informed participants with equivalent digital literacy limitations successfully navigated identical technical challenges with agent support. Service experiences captured in theme 3 created feedback loops affecting subsequent adoption: positive experiences generated word-of-mouth promotion that encouraged community members' service uptake ("My neighbor used it successfully, so I tried it too," FGD #7), while negative experiences, particularly older adults' repeated inability to access services without agents, circulated through social networks actively deterring potential users from attempting enrollment. Perceived benefits in theme 4 remained theoretical rather than experienced for most participants because theme 2 barriers prevented service access; the value proposition was sound, but inaccessibility meant benefits could not be realized. Provider integration challenges documented in theme 2 directly undermined patient-perceived benefits highlighted in theme 4: medication unavailability at HCs negated Babyl consultations' convenience advantages when patients still required facility visits, while lack of clinical follow-up information prevented the care continuity that would optimize health outcomes. Participants' improvement recommendations in theme 5 implicitly recognized these interdependencies: universal agent deployment simultaneously addresses theme 1 knowledge gaps and theme 2 barrier navigation challenges; health management information system integration resolves theme 2 provider workflow challenges while enabling theme 3 positive experiences through care continuity; expanded service coverage addresses theme 2 coverage gaps while enhancing theme 4 benefit realization. The finding that no single intervention type dominated stakeholder recommendations participants proposed community-level awareness strategies, facility-level agent deployment and equipment improvements, service-level coverage expansion and process simplification, and policy-level financing and governance solutions—indicates that stakeholders intuitively understood that multilevel implementation challenges require correspondingly multilevel solutions. This systems perspective, emerging organically from participants' lived experiences rather than being imposed by implementation science frameworks, suggests that user-centered digital health design must attend to complex interactions among individual capabilities, community resources, health system structures, technological infrastructure, and policy environments.

Discussion

Principal Findings

This qualitative study provides insights into the user experiences and implementation lessons from Babyl's large-scale digital health platform in Rwanda, representing one of the most extensive telemedicine implementations in sub-Saharan Africa. Despite Babyl's impressive reach, covering 450 of 510 primary health facilities and enrolling 2 million patients, the eventual closure in September 2023 underscores the critical importance of understanding implementation challenges beyond initial uptake metrics [7]. Importantly, the findings from this study are now informing Rwanda's current redesign of its digital health strategy, providing evidence-based insights for developing more sustainable and effective digital health interventions. The findings reveal a complex interplay of individual, community, health system, and technological factors influencing digital health adoption, utilization, and sustainability, contributing valuable lessons for future initiatives in similar contexts and supporting Rwanda's broader health system strengthening goals [21,22].

This study advances digital health implementation science through three distinct contributions to the literature. First, the

XSL•FO
RenderX

comprehensive stakeholder inclusion approach sampled diverse groups, including active users, lapsed users (who discontinued services), registered nonusers (who never used despite registering), eligible nonregistrants (who were aware but never registered), health care providers, HC heads, and Babyl agents. Unlike previous evaluations that focused primarily on active users or clinical outcomes [7,10-12], this comprehensive sampling reveals the full spectrum of adoption patterns and identifies barriers across the complete user journey from awareness through sustained utilization. This methodological approach demonstrates that evaluating only successful adopters creates selection bias that obscures critical implementation barriers.

Second, the implementation-focused analytical framework examined why and how implementation succeeded or failed across different contexts, rather than solely measuring what outcomes were achieved. By examining facilitators, barriers, and contextual factors through an implementation science lens informed by WHO digital health guidelines [19] and the Consolidated Framework for Implementation Research [20], the study provides actionable insights for future digital health initiatives. This focus on implementation processes complements outcome-focused evaluations by revealing the mechanisms underlying utilization patterns.

Third, the multi-level analysis examined implementation challenges at individual levels (user characteristics, digital literacy), community levels (social norms, agent availability), health system levels (provider readiness, workflow integration), and policy levels (financing, governance structures). This holistic framework demonstrates that digital health implementation challenges cannot be addressed through single-level interventions. The finding that 100% of participants older than 50 years required agent assistance, yet 41.7% of HCs lacked agents, exemplifies how individual-level needs and system-level resource allocation interact to determine access outcomes. This multi-level perspective provides a comprehensive understanding of the complex, interacting factors that influenced Babyl's trajectory and ultimate discontinuation.

The findings reveal profound equity considerations that must inform future digital health implementations. The demographic analysis demonstrates that digital health adoption is not merely a matter of individual choice but is fundamentally shaped by structural inequities in age, education, gender, and geography. Older adults faced substantial barriers, including limited smartphone ownership and digital literacy challenges. Educational attainment strongly predicted utilization patterns, with active users more likely to have secondary or higher education. Geographic disparities were pronounced, with rural participants experiencing lower smartphone ownership and substantially worse network connectivity. Gender differences emerged particularly in privacy concerns, with female participants emphasizing confidential consultations more frequently than male participants. These overlapping disadvantages create compounded barriers for the most vulnerable populations; older, less educated, rural, female participants face multiple simultaneous obstacles to digital health access. Future digital health interventions must proactively address these equity considerations through targeted strategies, including age-appropriate interfaces, low-literacy design, substantial rural infrastructure investments, and gender-sensitive implementation approaches.

Participants' experiences highlight the critical importance of digital literacy support and trusted community channels for adoption. The finding that older participants consistently required agent assistance for registration, while younger participants more often enrolled independently, demonstrates that digital health platforms cannot rely on user self-service alone.

The absence of dedicated Babyl agents at many HCs, identified as the most critical implementation gap across all stakeholder groups, meant that older and less digitally literate potential users had no pathway to access services. Participants' emphasis on word-of-mouth promotion and community health worker involvement reflects the importance of trusted local channels for health information dissemination [8]. Reports of users deleting verification codes due to cyberfraud concerns illustrate how broader digital security anxieties can undermine adoption, suggesting that digital health implementations require accompanying digital literacy and security education components [23].

Health care provider perspectives revealed critical health system integration gaps that limited clinical effectiveness and provider buy-in. Providers reported an inability to access Babyl consultation records, creating continuity of care challenges and workflow disruptions when patients arrived at facilities. The frequent medication stock-outs for prescribed treatments undermined patient confidence and created provider frustration. Lack of feedback mechanisms meant providers could not learn from Babyl consultations or understand clinical outcomes, limiting professional development opportunities. While providers valued reduced OPD congestion for simple cases and access to physician expertise for complex cases, concerns about diagnostic limitations without physical examination and parallel rather than integrated service delivery prevented full provider engagement. These findings demonstrate that technical platform functionality is insufficient; digital health implementations require genuine integration with clinical workflows, information systems, and pharmaceutical supply chains [5,17].

Digital health implementation in LMICs faces distinct challenges, including health care provider shortages, limited infrastructure access, inadequate education opportunities, greater disease burden, and insufficient data management capabilities compared with high-income countries [24,25]. Babyl's experience can be contextualized within broader LMIC digital health implementations. India's e-Sanjeevani platform delivered over 10 million teleconsultations through government-led integration within the national digital health strategy, while Kenya's M-TIBA platform connected 4 million users with private-sector financing models integrated into existing health insurance schemes. Brazil's Portal Telemedicina demonstrates infrastructure integration connecting remote primary care providers with specialists nationwide [24,25]. Unlike these government-integrated approaches, Babyl operated as a parallel system with limited health management information system integration and donor-dependent financing. Participants'

XSL•FO
RenderX

emphasis on national health system integration and expanded insurance coverage reflects awareness of structural requirements for digital health sustainability.

The findings highlight critical service design considerations for future implementations. The exclusion of chronic disease management (diabetes, hypertension) and pediatric consultations created significant coverage gaps, with many lapsed users discontinuing after initial acute illness treatment when ongoing health needs fell outside service scope. Participants' recommendations for telephone sharing mechanisms and health post partnerships for remote areas reflect practical solutions to device access barriers. The registration process complexity, requiring multiple steps including phone verification, insurance validation, and profile setup, deterred potential users, particularly older adults and those with lower digital literacy. Successful registration required either independent technological capability or access to agent support; the absence of agents at many facilities meant the registration barrier was insurmountable for substantial population segments. These implementation details, often overlooked in digital health planning, proved critical to adoption and sustained utilization.

Implementation Lessons and Policy Implications

The findings yield several implementation lessons for future digital health initiatives in similar contexts. First, sustainable financing mechanisms integrated within health system structures are essential. Babyl's closure despite technical success and user satisfaction demonstrates that donor-dependent models face inherent sustainability challenges. Second, comprehensive health system integration from implementation outset is critical; parallel systems create workflow disruptions, continuity of care gaps, and provider resistance that undermine effectiveness regardless of platform technical capabilities. Third, differentiated implementation approaches are required for diverse population segments; one-size-fits-all strategies fail to address the distinct needs of older versus younger, rural versus urban, and high versus low digital literacy users. Fourth, adequate human resource allocation for user support is nonnegotiable. The absence of agents at HCs created insurmountable barriers for populations most in need of support. Finally, explicit governance frameworks clarifying digital health service accountability within existing health management structures are essential to prevent implementation confusion and ensure sustainability [2,3,12].

Strengths and Limitations

This study has several strengths. The purposive sampling strategy across 4 distinct user categories (active users, lapsed users, registered nonusers, and eligible nonregistrants) enabled examination of the full adoption spectrum rather than only successful users, revealing critical barriers typically missed in implementation evaluations. Multi-stakeholder inclusion of patients, health care providers, HC heads, and Babyl agents provided triangulated perspectives on implementation facilitators and challenges from different vantage points within the health system. Geographic diversity across 6 districts with both urban and rural representation ensured findings captured varied contexts. Data collection timing during active implementation enabled exploration of real-time experiences rather than retrospective recall. The use of established qualitative methods with rigorous analysis procedures, including independent coding, regular team discussions, member checking, and audit trails, enhanced analytical validity [18].

Several limitations should be considered when interpreting findings. (1) The study relied on self-reported experiences and perceptions, which may be subject to recall bias and social desirability response patterns, particularly when discussing reasons for non-adoption or discontinuation. (2) Data collection occurred during Babyl's active implementation period before closure was announced; participants' perceptions may have differed had they known about impending service discontinuation. (3) Eligible nonregistrants who were unaware of Babyl services were not included, meaning perspectives from the least-reached population segments are absent. (4) The study did not directly assess clinical outcomes, cost-effectiveness, or health impact, limiting conclusions to user experiences and implementation processes rather than clinical or economic outcomes. (5) Although data saturation was documented through iterative data collection and analysis, the possibility remains that additional perspectives may exist within the broader population. (6) The specific context of Rwanda's health system, insurance coverage structures, and telecommunications infrastructure may limit direct generalizability to other LMICs, though conceptual transferability to similar contexts is likely. (7) Language translation from Kinyarwanda to English may have introduced subtle meaning shifts despite back-translation validation procedures. (8) Selection bias may have occurred despite purposive sampling efforts, as participants willing to discuss digital health services may differ from those who declined participation. (9) The cross-sectional design captured experiences at a single time point; longitudinal follow-up would have enabled examination of evolving perceptions and utilization patterns over time.

Conclusions

This qualitative study explored user experiences and implementation challenges of Babyl's telemedicine platform in Rwanda, which operated from 2016 until its closure in September 2023. Our findings reveal that while participants recognized specific benefits of digital consultation, particularly convenience, reduced travel costs, and privacy, multiple interconnected barriers at individual, community, health system, and policy levels impeded widespread adoption and sustained utilization. Through analysis of experiences from active users, lapsed users, registered nonusers, and eligible nonregistrants, alongside perspectives from health care providers, HC heads, and Babyl agents, the study identified 5 major implementation themes that collectively explain the platform's challenges.

Key insights from participants emphasize that successful digital health implementation extends beyond technological functionality to encompass critical human and structural elements. First, participants across all categories emphasized that digital literacy support and human assistance are nonnegotiable requirements, not optional enhancements. All participants (100%) older than 50 years required agent assistance, yet agent absence at HCs was identified as the single most critical implementation gap. Second, health care providers

XSL•FO
RenderX

consistently reported that the platform operated as a parallel system rather than an integrated component of health service delivery, creating workflow disruptions, continuity of care gaps, and medication stock-out frustrations that undermined both provider satisfaction and patient confidence. Third, participants highlighted significant equity implications: older adults, individuals with lower educational attainment, rural residents, and those lacking smartphones faced compounded barriers that the service design failed to adequately address.

These findings carry practical implications for Rwanda's current digital health redesign process and for similar initiatives in comparable contexts. Participants' recommendations, validated across stakeholder groups, converge on five critical requirements: (1) genuine integration with existing health information systems and clinical workflows rather than parallel operation; (2) sustainable financing models embedded within national health system structures rather than donor-dependent arrangements; (3) universal availability of trained agents providing enrollment and troubleshooting support; (4) differentiated implementation strategies that accommodate varying levels of digital literacy, device access, and infrastructure; and (5) explicit governance frameworks clarifying roles, responsibilities, and accountability within health management structures. While this study focused specifically on Babyl's experience in Rwanda, these implementation lessons may inform digital health initiatives in settings with similar health system structures, resource constraints, and digital infrastructure development stages.

Acknowledgments

We acknowledge the contributions of all study participants who shared their experiences and insights. We thank the Rwanda Ministry of Health, University of Rwanda School of Public Health, and CIIC-HIN for their support in conducting this research. We also acknowledge the Bill and Melinda Gates Foundation for funding the broader evaluation project of which this qualitative study was a component. All authors declared that they had insufficient or no funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article. No artificial intelligence tools or large language models were used in the design, conduct, analysis, interpretation, or writing of this study. All data collection, thematic analysis, manuscript writing, and editing were performed exclusively by the research team members listed as authors.

Funding

This research was supported by the Centre for Impact, Innovation and Capacity Building for Health Information Systems and Nutrition (CIIC-HIN) with funding from the Bill & Melinda Gates Foundation (Grant Number: INV-030342). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The views expressed are those of the authors and do not necessarily reflect the official policy or position of the funding organizations.

Data Availability

The deidentified data that support the findings of this study are included in the manuscript in the form of anonymized quotes and thematic analyses. Full interview transcripts cannot be shared publicly due to privacy concerns and ethical considerations regarding participant confidentiality. However, the interview guide and coding framework are available from the corresponding author upon reasonable request, subject to approval by the ethical review board.

Authors' Contributions

The original draft was written by SMF, JH, JC, and GA. The methodology and formal analysis were developed and performed by SMF, JH, JC, and GA. The manuscript underwent review and editing by MSF, GA, UN, USM, FRK, YDN, PS, RH, JC, and JH. All authors contributed to the conceptualization and design of the study. Data collection was supervised by JH and JC. Critical revisions and intellectual content were provided by all co-authors. All authors reviewed and approved the final manuscript for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Focus Group Discussion Guide for Babyl registered clients who never used digital services and nonregistered eligible CBHI (Community-Based Health Insurance) members who are aware of Babyl services.
[DOCX File, 22 KB - jmir_v28i1e84832_app1.docx]

Multimedia Appendix 2
Key Informant interview guide for health care providers.
[DOCX File, 24 KB - jmir_v28i1e84832_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Key Informant interview guide for Babyl agents.
[DOCX File, 68 KB - jmir_v28i1e84832_app3.docx]

Multimedia Appendix 4
Key Informant interview guide with health center staff.
[DOCX File, 23 KB - jmir_v28i1e84832_app4.docx]

References

1. Ahmed MM, Okesanya OJ, Olaleke NO, et al. Integrating digital health innovations to achieve universal health coverage: promoting health outcomes and quality through global public health equity. Healthcare (Basel) 2025 May 5;13(9):1060. [doi: 10.3390/healthcare13091060] [Medline: 40361838]
2. Fifty-eighth world health assembly. : World Health Organization (WHO); 2005 URL: https://apps.who.int/gb/ebwha/pdf_files/wha58-rec1/english/a58_2005_rec1-en.pdf [accessed 2026-03-30]
3. Al-Shorbaji N. The World Health Assembly resolutions on eHealth: eHealth in support of universal health coverage. Methods Inf Med 2013;52(6):463-466. [doi: 10.1055/s-0038-1627062] [Medline: 24310395]
4. Hollander JE, Carr BG. Virtually perfect? Telemedicine for Covid-19. N Engl J Med 2020 Apr 30;382(18):1679-1681. [doi: 10.1056/NEJMp2003539] [Medline: 32160451]
5. Ibrahim MS, Mohamed Yusoff H, Abu Bakar YI, Thwe Aung MM, Abas MI, Ramli RA. Digital health for quality healthcare: a systematic mapping of review studies. Digit Health 2022;8:20552076221085810. [doi: 10.1177/20552076221085810] [Medline: 35340904]
6. Primary health care systems (primasys). Nigeria Governors' Forum. 2017. URL: https://ngfrepository.org.ng:8443/jspui/bitstream/123456789/3163/1/PRIMARY HEALTH CARE SYSTEMS-CASE STUDY FROM NIGERIA.pdf [accessed 2025-08-15]
7. Telehealth innovation in Rwanda: lessons and insights from the Babyl experience. Percept. 2021. URL: https://percept.co.za/wp-content/uploads/2021/05/Brief-3-babyl.pdf [accessed 2025-08-28]
8. Mitgang EA, Blaya JA, Chopra M. Digital health in response to COVID-19 in low- and middle-income countries: opportunities and challenges. Glob Policy 2021 Jul;12(Suppl 6):107-109. [doi: 10.1111/1758-5899.12880] [Medline: 34230840]
9. Nyakabau MR. Digital-first integrated care: Rwanda's innovative digital health care service. : Transform Health Coalition; 2021 Nov URL: https://transformhealthcoalition.org/wp-content/uploads/2021/11/Rwanda.pdf [accessed 2026-03-18]
10. Totten AM, McDonagh MS, Wagner JH. The Evidence Base for Telehealth: Reassurance in the Face of Rapid Expansion during the COVID-19 Pandemic: Rockville (MD): Agency for Healthcare Research and Quality (US); 2020. [Medline: 32479040]
11. Jiang X, Ming WK, You JHS. The cost-effectiveness of digital health interventions on the management of cardiovascular diseases: systematic review. J Med Internet Res 2019 Jun 17;21(6):e13166. [doi: 10.2196/13166] [Medline: 31210136]
12. Digital-in-health: unlocking the value for everyone digit unlocking value everyone. Washington (DC): World Bank. 2023. URL: https://www.worldbank.org/en/topic/health/publication/digital-in-health-unlocking-the-value-for-everyone [accessed 2026-03-18]
13. Ashwood JS, Mehrotra A, Cowling D, Uscher-Pines L. Direct-To-consumer telehealth may increase access to care but does not decrease spending. Health Aff (Millwood) 2017 Mar;36(3):485-491. [doi: 10.1377/hlthaff.2016.1130]
14. Singh H. Diagnostic errors: moving beyond "no respect" and getting ready for prime time. BMJ Qual Saf 2013 Oct;22(10):789-792. [doi: 10.1136/bmjqs-2013-002387] [Medline: 24048615]
15. Madhivanan A, Venugopal V, Dongre AR. Physical violence against doctors: a content analysis from online Indian newspapers. Indian J Community Med 2020;45(1):108-109. [doi: 10.4103/ijcm.IJCM_215_19] [Medline: 32029997]
16. Cory N, Stevens P. Building a global framework for digital health services in the era of COVID-19. Inf Technol Innov Found 2020 [FREE Full text]
17. Labrique A, Agarwal S, Tamrat T, Mehl G. WHO digital health guidelines: a milestone for global health. NPJ Digit Med 2020;3(1):120. [doi: 10.1038/s41746-020-00330-2] [Medline: 33015373]
18. Eslami Jahromi M, Ayatollahi H. Utilization of telehealth to manage the Covid-19 pandemic in low- and middle-income countries: a scoping review. J Am Med Inform Assoc 2023 Mar 16;30(4):738-751. [doi: 10.1093/jamia/ocac250]
19. World Health Organization. Monitoring and Evaluating Digital Health Interventions A Practical Guide to Conducting Research and Assessment: World Health Organization; 2016:1-144 URL: https://www.who.int/publications/i/item/9789241511766 [accessed 2026-03-27]
20. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implementation Sci 2009 Dec;4(1):50. [doi: 10.1186/1748-5908-4-50]

XSL•FO
RenderX

21. 4X4 reform: executive summary towards quadrupling the current workforce in 4 years (4X4); uplifting quantity and quality of care. Ministry of Health. 2024. URL: https://www.moh.gov.rw/index.php?eID=dumpFile&t=f&f=92859&token=3c26da47182b639efd749f9c6829633dc0fcada4 [accessed 2025-10-25]
22. Health sector strategic plan V. MOH-Rwanda. URL: https://www.moh.gov.rw/index.php?eID=dumpFile&t=f&f=117507&token=f2a527e089201f73bec29f8ce6d6c4dc55ed241e [accessed 2025-11-05]
23. Global Strategy on Digital Health 2020-2025: World Health Organization; 2021. URL: https://iris.who.int/ [accessed 2026-03-18]
24. McKinsey & Company. Unlocking digital healthcare in lower- and middle-income countries. McKinsey. 2021 Sep 22. URL: https://www.mckinsey.com/industries/healthcare/our-insights/unlocking-digital-healthcare-in-lower-and-middle-income-countries [accessed 2026-11-18]
25. Bhushan I, Anandampillai K, Agarwal S. Navigating complexities: agile digital health initiatives in developing countries. Oxf Open Digit Health 2024;2:oqae032. [doi: 10.1093/oodh/oqae032] [Medline: 40230973]

Abbreviations

CBHI: Community-Based Health Insurance
FGD: focus group discussion
HC: health center
KII: key informant interview
LMIC: low- and middle-income country
OPD: outpatient department
RAMA: Rwandaise d'Assurance Maladie
WHO: World Health Organization

Edited by A Stone; submitted 25.Sep.2025; peer-reviewed by KL Mauco, R Ranjan; revised version received 10.Jan.2026; accepted 28.Jan.2026; published 01.Apr.2026.

Please cite as:
Musange Furere S, Absolomon G, Umutoni N, Matutina Umuhoza S, K Rubuga F, Nsaba-Uwera YD, Singa Muhoza P, Hitimana R, Condo J, Humuza J
Digital Primary Health in Rwanda: Qualitative Study of User Experiences and Implementation Lessons From Babyl's Telemedicine Platform
J Med Internet Res 2026;28:e84832
URL: https://www.jmir.org/2026/1/e84832
doi:10.2196/84832

© Sabine Musange Furere, Gashaija Absolomon, Nathalie Umutoni, Stella Matutina Umuhoza, Felix K Rubuga, Yvonne Delphine Nsaba-Uwera, Patrick Singa Muhoza, Regis Hitimana, Jeanine Condo, James Humuza. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Telehealth Use and Modality Choice Among US Adults: Shorrocks-Shapley Decomposition of a 2022 Cross-Sectional National Survey

Corneliu Bolbocean[1], BA, MA, MPPM, PhD; Corey Hayes[2], MPH, PhD; Cari Bogulski[3], PhD

[1]Nuffield Department of Primary Care Health Sciences, University of Oxford, Radcliffe Primary Care Building, Radcliffe Observatory Quarter, Woodstock Rd, Oxford, United Kingdom

[2]Department of Pharmacy Practice, University of Arkansas for Medical Sciences, Little Rock, AR, United States

[3]Department of Biomedical Informatics in the College of Medicine, UAMS, Little Rock, AR, United States

Corresponding Author:
Corneliu Bolbocean, BA, MA, MPPM, PhD
Nuffield Department of Primary Care Health Sciences, University of Oxford, Radcliffe Primary Care Building, Radcliffe Observatory Quarter, Woodstock Rd, Oxford, United Kingdom

Abstract

Background: Telehealth use surged during the COVID-19 pandemic and has stabilized at levels substantially above prepandemic baselines. However, concerns persist that the digital divide may reproduce or widen disparities in access. Understanding the determinants of telehealth use—and particularly modality choice between video and audio—is essential for designing policies that promote equitable access in the post–public health emergency era.

Objective: This study aims to identify determinants of telehealth use and modality among US adults in 2022 and quantify the relative contributions of digital, geographic, clinical, and socioeconomic domains.

Methods: We conducted a cross-sectional secondary analysis of the sixth cycle of the Health Information National Trends Survey, administered in 2022 by the National Cancer Institute, a nationally representative, 2-stage stratified random probability survey of civilian, noninstitutionalized US adults aged 18 years or older. Sampled households were recruited via mailed invitations, and 1 adult per household was randomly selected using the next birthday method and invited to complete a self-administered questionnaire between February 2022 and November 2022 (N=6252). The primary analytic sample included respondents with nonmissing telehealth modality responses (n=6046, 59.4% female; mean age of 55.1 y). Individual-level data were linked to county-level American Community Survey socioeconomic indicators and broadband availability measures. The primary outcome was telehealth use, categorized as video (n=1641, 27.2%; 95% CI 25.5%-29.1%), audio-only (n=876, 12.1%; 95% CI 10.9%-13.4%), or none (n=3529, 60.7%; 95% CI 58.6%-62.7%). We estimated 4 binary contrasts using survey-weighted linear probability models with jackknife variance estimation, reporting absolute risk differences in percentage points (pp) with 95% CIs. We applied Shorrocks-Shapley decomposition to quantify each predictor domain's contribution to explained variance.

Results: Nationally, 39.3% (n=2517; 95% CI 37.3% - 41.4%) reported any telehealth use in the past 12 months. In survey-weighted linear probability models (α=.05), significant predictors of any telehealth vs none included: male sex (−9.7 pp, 95% CI −14.0 to −5.4), disability status (+22.5 pp, 95% CI 16.1-28.8), and health app use (+18.4 pp, 95% CI 12.0-24.8). For video vs audio-only telehealth, insurance coverage increased video use (+21.2 pp, 95% CI 13.0-29.3), while basic cell phone only (vs smartphone) decreased video use (−20.1 pp, 95% CI −33.5 to −6.8). Shorrocks-Shapley decomposition revealed that digital access and eHealth behaviors explained 40.4% of variance in video vs audio choice and 33.4% of video vs none; geography explained 40.5% of audio vs none; digital factors (25.7%), geography (19.7%), and health status and needs (15.5%) all contributed substantially to any vs none.

Conclusions: Digital access and eHealth behaviors collectively explain more variance in modality choice than traditional sociodemographic factors. Telehealth uptake reflects a combination of digital factors, geography, and clinical need, whereas video modality specifically hinges on digital readiness. Interventions pairing sustained insurance coverage with targeted investments in device access, affordable high-speed connectivity, and digital literacy training are most likely to narrow persistent telehealth gaps.

(J Med Internet Res 2026;28:e81879) doi:10.2196/81879

KEYWORDS

telehealth; digital divide; eHealth literacy; Shorrocks-Shapley decomposition; Health Information National Trends Survey; HINTS

XSL•FO
RenderX

Introduction

The landscape of health care delivery in the United States has been reshaped by the expansion of telehealth—the use of electronic information and telecommunications technologies to support clinical care, education, and administration [1-3]. Services delivered by telephone, video, remote monitoring, or asynchronous platforms can lower geographic and logistical barriers to care [3]. However, as telehealth has scaled, concerns persist that the "digital divide" may reproduce or even widen disparities in access and outcomes [4-9].

Evidence documents persistent differences in telehealth access across sociodemographic groups. Older adults, racial and ethnic minorities, individuals with lower educational attainment, and those residing in rural areas consistently demonstrate lower telehealth use rates [10]. A systematic review found that telehealth adoption during the pandemic was significantly lower among Black and Hispanic patients compared to White patients after accounting for clinical and geographic factors [11]. Recent analyses of post-pandemic telehealth patterns reveal that while overall use has stabilized, disparities by age, income, and digital literacy persist [12]. Critically, modality choice—whether patients use video or audio-only visits—varies substantially by socioeconomic status (SES) and digital access, with audio-only visits more common among older, lower-income, and rural populations [13].

Use of telehealth surged during the initial phase of the COVID-19 public health emergency (PHE) and then declined from its peak, but it has stabilized at levels substantially higher than before the pandemic. For example, among Medicare fee-for-service beneficiaries, telehealth use in 2021 was approximately 38-fold higher than prepandemic levels (from 0.1% in 2019 to 3.8% of visits in 2021) [14]. By late 2022, about 15% of Medicare beneficiaries received telehealth in a given quarter, and in the last quarter of 2023, more than 1 in 10 beneficiaries (ie, 13%) used telehealth at least once, well above the pre-2020 baseline [15,16]. Also, in population-based surveys, 37% of US adults reported telemedicine use in 2021 and 43% of adults with a health care visit reported telemedicine use in 2022 (70% video; 30% audio) [17,18]. Taken together, these data indicate that telehealth has become a substantial component of routine care, with meaningful variation by modality (video vs audio).

The PHE also saw major federal investments in household internet affordability that intersect directly with telehealth access. The Affordable Connectivity Program (ACP) launched at the end of 2021 expanded subsidized broadband, which increased adoption among lower-income households and provided a discount for eligible households and was active throughout 2022 [19,20]. By the end of 2022, roughly 15.4 million households had enrolled (around 28% of eligible households), growing to more than 23 million by the February 2024 enrollment freeze [21,22]. The ACP ended in June 2024 when Congress did not appropriate additional funding, potentially affecting an estimated 23 million households [23]. Evidence after the ACP wind-down suggests that broadband affordability shocks can translate into foregone telehealth among subsidy-eligible households: in a post-ACP consumer survey of over 68,000 lifeline and/or former ACP participants fielded in November 2024, 36.1% of respondents who were previously enrolled in ACP reported being unable to continue with telehealth, rehabilitation, or remote medical monitoring without ACP [24]. These policy changes have altered the digital access landscape in ways that may affect who can use—and how they use—telehealth.

The digital divide is multidimensional, encompassing access to devices and connectivity, the quality and affordability of home internet, and the skills and engagement required to use digital tools effectively [25]. These dimensions map directly onto telehealth modality: video visits typically require higher-bandwidth connectivity, compatible devices, and greater comfort with app-based workflows, whereas audio-only visits may remain feasible when bandwidth, device capability, or digital skills are constrained [26,27]. However, much of the existing literature emphasizes individual predictors without quantifying the relative explanatory importance of digital, geographic, clinical, and socioeconomic domains. A domain-level approach can help distinguish which levers are most policy-relevant for improving overall uptake vs enabling equal access to video-based care.

Despite growing research on telehealth disparities, significant gaps remain in understanding the relative importance of different determinant domains. Most existing studies examine individual predictors in isolation rather than quantifying how much each domain—demographics, SES, health needs, digital access, and geography—contributes to explaining telehealth use patterns. Furthermore, few studies have simultaneously examined predictors of both any telehealth use and modality choice (video vs audio-only) in the post-PHE period, when several temporary telehealth flexibilities required repeated legislative extensions and telehealth patterns stabilized into new norms.

We use nationally representative data from the sixth cycle of the Health Information National Trends Survey (HINTS 6), administered in 2022 by the National Cancer Institute, to describe contemporary patterns of telehealth use (phone or video) and to evaluate how sociodemographic, clinical, area-level and geographic measures, psychosocial, and digital access factors relate to (1) any telehealth use and (2) modality choice. Using the Shorrocks-Shapley decomposition models, we quantify the relative contribution of each explanatory domain to explained variance in telehealth outcomes [28].

Our objectives were to (1) estimate national prevalence of telehealth use by modality (video, audio-only, and none) among US adults in 2022; (2) identify individual and contextual determinants of telehealth use and modality choice using survey-weighted regression models; and (3) quantify the relative contribution of digital access or eHealth behaviors, geography, clinical need, and socioeconomic factors to explained variance using the Shorrocks-Shapley decomposition.

XSL•FO
RenderX

Methods

Research Design

This study used a cross-sectional observational design using secondary analysis of nationally representative survey data linked to area-level contextual measures. The primary aim was to identify individual and contextual determinants of telehealth use and modality choice among US adults in 2022. We report this cross-sectional observational study in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for cross-sectional studies [29].

Participants

Inclusion and Exclusion Criteria

The study population included civilian, noninstitutionalized adults aged 18 years and older residing in the United States. Eligibility was determined by the HINTS sampling frame, which excluded individuals living in institutionalized settings (eg, prisons and nursing homes), those without a valid mailing address, and residents of US territories. For this analysis, we included all respondents with valid responses to the telehealth use question.

Sampling Procedures

We analyzed the HINTS 6, a nationally representative survey of civilian US adults administered by the National Cancer Institute from February 2022 through November 2022 [30,31]. HINTS uses a complex 2-stage stratified random probability sampling design: first, a stratified sample of addresses was selected from a comprehensive address file, then 1 adult per household was randomly selected to complete the survey. The survey provides person-level sampling weights and 50 jackknife replicate weights to allow reliable national inference. The restricted-use version of HINTS (available to researchers through a data use agreement with the National Cancer Institute) contains geocoding variables (county Federal Information Processing Series [FIPS] codes) not included in the public-use file, enabling linkage with area-level datasets.

Sample Size

The HINTS 6 cycle included 6252 respondents. The primary analytic sample for telehealth modality outcomes included respondents with nonmissing telehealth modality responses (n=6046). This sample size provides adequate statistical power to detect small-to-medium effect sizes (Cohen $d \geq 0.10$) with 80% power at $\alpha.05$ for the primary regression analyses, accounting for the design effect of the complex survey design.

Data Linkage

Consistent with prior literature, we appended area-level measures to each respondent via county FIPS codes. County socioeconomic context for 2022 (total population, poverty rate, unemployment rate, percent without a high school diploma, and percent aged ≥65 y) was derived from the American Community Survey [32]. County-level broadband availability was obtained from the National Institute on Minority Health and Health Disparities (NIMHD) health disparities (HD) Pulse portal as the percent of residents with broadband access [33]. We additionally obtained the Centers for Disease Control and Prevention/Agency for Toxic Substances and Disease Registry (CDC/ATSDR) Social Vulnerability Index (SVI) to characterize local vulnerability [34].

Ethical Considerations

This study analyzed the 2022 HINTS 6 restricted-use deidentified dataset. The University of Arkansas for Medical Sciences Institutional Review Board determined this study did not constitute human participants research (protocol 297801). The original HINTS data collection was approved by the Westat Institutional Review Board, and all survey respondents provided implied informed consent by completing and returning the questionnaire. The restricted-use dataset contained no direct identifiers; geographic information was limited to county-level codes. Respondents received no compensation for the secondary analysis; compensation for the original HINTS survey included a US $2 preincentive and an opportunity for a US $25 postsurvey incentive. No images containing identifiable individuals were included in this study.

Measures

Outcome Variable

Telehealth exposure in the 12 months preceding the survey was categorized into 3 mutually exclusive groups—video, audio-only, and no telehealth—yielding 1641 (27.2%, 95% CI 25.5% - 29.1%), 876 (12.1%, 95% CI 10.9% - 13.4%), and 3529 (60.7%, 95% CI 58.6% - 62.7%) respondents, respectively (Table 1). Respondents reporting any video visit were assigned to the video group (even if they also had phone visits); audio-only indicates telephone visits with no video exclusively. We used “telehealth” to denote synchronous video or phone visits. Other digital health activities (apps, portals, and information seeking) were considered model predictors, not modalities of telehealth delivery.

XSL•FO
RenderX

Table . Descriptive characteristics by telehealth modality (video, audio only, and none) among US adults[a].

Characteristics	Video	Audio only	No telehealth
Participants, n (%)	1641 (27.2)	876 (12.1)	3529 (60.7)
Age (y), mean (SD)	52.84 (16.46)	58.57 (17.55)	56.26 (17.58)
Female, n (%)	1071 (65.3)	566 (64.6)	2012 (57)
High school graduation, n (%)	1499 (96.7)	762 (92.1)	3120 (92.5)
Race and ethnicity, n (%)			
Asian	51 (3.1)	29 (3.3)	116 (3.3)
Black	241 (14.7)	135 (15.4)	551 (15.6)
Hispanic	262 (16)	177 (20.2)	542 (15.4)
White	1155 (76.5)	566 (71.7)	2353 (72.6)
Other	169 (10.3)	119 (13.6)	422 (12)
Metropolitan status, n (%)			
Large central metro	684 (41.7)	355 (40.5)	1220 (34.6)
Large fringe metro (suburban)	393 (23.9)	176 (20.1)	769 (21.8)
Medium or small metro	401 (24.4)	250 (28.5)	1016 (28.8)
Micropolitan	96 (5.9)	55 (6.3)	307 (8.7)
Noncore (rural)	66 (4)	40 (4.6)	216 (6.1)
Income group (US $, n (%)			
Low ($20,000)	213 (14.4)	155 (20)	555 (17.5)
Middle (20,000-75,000)	557 (37.6)	351 (45.2)	1457 (45.9)
High (≥75,000)	788 (48)	270 (34.8)	1292 (36.6)
Employed, n (%)	912 (55.6)	349 (41.9)	1807 (51.2)
Cancer history, n (%)	258 (15.7)	134 (15.3)	487 (13.8)
Hypertension, n (%)	731 (44.5)	452 (51.6)	1477 (41.9)
Disabled, n (%)	172 (11)	96 (11.5)	211 (6.3)
PROMIS[b] Isolation T-score, mean (SD)	47.11 (9.97)	46.36 (9.46)	45.40 (9.30)
PROMIS Meaning T-score, mean (SD)	54.16 (10.29)	55.10 (9.75)	55.61 (10.08)
Health Insurance, n (%)	1559 (95.1)	813 (92.8)	3165 (89.7)
Health numeracy, n (%)			
Easy or very easy	1241 (75.6)	624 (71.2)	2495 (70.7)
Hard or very hard	369 (22.5)	231 (26.4)	951 (26.9)
Use of health or wellness apps, n (%)			
Used app	1103 (67.2)	414 (47.3)	1441 (40.8)
Did not use or no apps or no smart device	498 (30.3)	385 (43.9)	1871 (53)
Sought health information online, n (%)			
No	128 (7.8)	95 (10.8)	506 (14.3)
Yes	1367 (83.3)	586 (66.9)	2294 (65)
Not internet user	120 (7.3)	153 (17.5)	586 (16.6)
Satisfaction with home internet for health needs, n (%)			
Satisfied (extremely or very)	1075 (65.5)	439 (50.1)	1789 (50.7)
Somewhat	354 (21.6)	194 (22.1)	799 (22.6)

XSL•FO
RenderX

Characteristics	Video	Audio only	No telehealth
Not satisfied or nonuser	185 (11.3)	207 (23.6)	802 (22.7)
Device ownership, n (%)			
Any smart device (tablet or smartphone or multiple)	1582 (96.4)	750 (85.6)	3091 (87.6)
No smart device (basic cell only or none)	54 (3.3)	120 (13.7)	417 (11.8)
Census division, n (%)			
New England	80 (4.9)	39 (4.5)	108 (3.1)
Middle Atlantic	192 (11.7)	83 (9.5)	399 (11.3)
East North Central	170 (10.4)	98 (11.2)	481 (13.6)
West North Central	51 (3.1)	17 (1.9)	210 (6)
South Atlantic	444 (27.1)	204 (23.3)	878 (24.9)
East South Central	51 (3.1)	39 (4.5)	202 (5.7)
West South Central	232 (14.1)	117 (13.4)	567 (16.1)
Mountain	126 (7.7)	58 (6.6)	255 (7.2)
Pacific	295 (18)	221 (25.2)	429 (12.2)

[a]n is unweighted and percentage is survey-weighted.

[b]PROMIS: Patient-Reported Outcomes Measurement Information System.

Explanatory Variables

A comprehensive set of explanatory variables to explain telehealth use was constructed based on prior literature on the digital divide and health care access [7,30,31,35]. Variable definitions and coding for individual-level demographics, SES, health status, digital access and literacy (device ownership, satisfaction with home internet for health needs, health numeracy, health or wellness app use, and online health information seeking), and geographic indicators (US Census division and urbanicity) are summarized in Appendix A in Multimedia Appendix 1.

Digital Readiness Variables

We modeled digital readiness using separate indicators for device ownership, internet satisfaction, health app use, and online health information seeking rather than combining them into a single composite score via factor analysis. This approach was chosen for 3 reasons. First, these constructs are theoretically distinct: device ownership reflects material access, internet satisfaction captures infrastructure quality, and app use or information seeking represents behavioral engagement and digital skills [36]. Second, separate coefficients provide policy information—interventions targeting device distribution differ from those promoting digital literacy or improving broadband infrastructure [37]. Finally, the Shorrocks-Shapley decomposition explicitly partitions explained variance across covariate domains, allowing us to quantify the relative importance of “digital access” vs “health literacy and eHealth” without collapsing distinct mechanisms into a single latent factor.

Psychosocial Variables

HINTS 6 includes validated Patient-Reported Outcomes Measurement Information System (PROMIS) short forms measuring psychosocial constructs [38,39]. The PROMIS Social Isolation 4-item short form assesses perceptions of being excluded, disconnected, or isolated from others; scores are reported as T-scores (population: mean 50, SD 10), with higher scores indicating greater isolation. The PROMIS Meaning and Purpose 4-item short form assesses sense of meaning, purpose, and engagement in life, with higher T-scores indicating greater sense of purpose.

PROMIS measures were included as exploratory covariates to assess whether psychosocial context contributed meaningfully beyond digital, geographic, clinical, and socioeconomic domains. Social isolation has been associated with reduced health care use, lower technology adoption, and barriers to accessing care [40,41]. Conversely, individuals with a greater sense of meaning and purpose may be more motivated to engage with health services and adopt health-promoting technologies [42]. These constructs may therefore influence both the decision to use telehealth and the choice of modality, independent of traditional sociodemographic and health-related factors.

Analytic Strategy

Primary Analyses

We report descriptive statistics and estimate multivariable associations using survey-weighted linear probability models (LPMs) for four contrasts: (1) video vs audio-only, (2) video vs no telehealth, (3) audio-only vs no telehealth, and (4) any telehealth (video or audio) vs no telehealth. The LPM directly models the probability of the outcome as a linear function of covariates and with survey weights provides easily interpretable absolute risk differences in percentage points [43,44].

XSL•FO
RenderX

Coefficients are interpreted as the percentage-point (pp) change in the probability of the outcome for a 1-unit change in the predictor, holding other variables constant.

Survey Design Adjustments

All adjusted analyses incorporated the complex survey design of HINTS 6. We used the final person-level weight and the 50 jackknife replicate weights to obtain design-consistent point estimates and robust standard errors, following HINTS technical guidance [45]. This approach yields nationally representative estimates with SEs and *P* values adjusted for stratification and clustering. Replicate-weight variance estimation used the jackknife mean squared error (MSE).

Decomposition Analysis

To quantify the explanatory contribution of prespecified covariate domains, we applied a Shorrocks-Shapley decomposition of the model R^2, partitioning the total R^2 into additive contributions from groups of predictors (demographics, SES, health status and needs, health system access, geography, urbanicity, psychosocial factors, digital access, health literacy and eHealth, and area context) [28,46]. Decompositions were implemented for each survey-weighted LPM using the *shapley2* module in Stata (StataCorp LLC) [47]. The Shorrocks-Shapley approach does not imply causal ordering or construct hierarchy; it quantifies predictive importance.

Missing Data Analysis

We evaluated item-level missingness across all study variables prior to model estimation. Missingness was assessed for continuous measures (including PROMIS T-scores and self-rated health) and for sociodemographic covariates (including employment status, disability status, race or ethnicity, and household income). To minimize loss of observations in adjusted models and to preserve the survey sample structure, selected categorical covariates with modest item nonresponse and/or structurally nonapplicable responses (indicators related to digital access and online behaviors) were coded with an explicit "missing/unknown" category. These missingness categories were included as indicator levels in the regression models to avoid listwise deletion for those variables, and coefficients for missingness indicators were not interpreted substantively.

To evaluate the plausibility of common missing-data mechanisms, we first applied the Little Missing Completely at Random (MCAR) test [48]. In this way, we assessed whether observed values differed systematically across missing data. Second, because income exhibited the largest item nonresponse among key sociodemographic measures, we performed a supplementary missingness model in which an indicator for income nonresponse was regressed with minimal missingness (age, census division, cancer history, and insurance status). These diagnostics were used to guide the selection of an appropriate missing-data handling strategy for sensitivity analyses.

As a sensitivity analysis, we conducted multiple imputation by chained equations under a missing at random (MAR) framework conditional on observed data. We generated m=20 imputed datasets. Variables imputed included household income, race or ethnicity, employment status, disability status, self-rated health, sex, marital status, education, and PROMIS psychosocial measures. The imputation models incorporated all variables used in the primary analyses, along with auxiliary predictors expected to be associated with missingness and/or the imputed values (including age, census division, urbanicity, cancer history, high blood pressure, and insurance status). After imputation, we re-estimated the primary survey-weighted regression models in each imputed dataset and pooled point estimates and SEs across imputations using standard combining rules. We then compared pooled estimates from the imputed analyses with the complete-case results to evaluate robustness; full imputation model specifications, diagnostics, and pooled regression outputs are provided in Multimedia Appendix 1.

Software: Statistical Threshold

All analyses were conducted in Stata/MP 18 (StataCorp LLC) using survey procedures (svy: prefix) with the provided jackknife replicate weights and MSE option, consistent with HINTS methodological guidelines. Tests were 2-sided with an α of .05.

Results

Sample Characteristics by Telehealth Modality

Table 1 summarizes characteristics by telehealth modality. Among 6252 respondents in HINTS 6, 1641 reported any video telehealth, 876 reported exclusively audio-only telehealth, and 3529 reported no telehealth in the past 12 months. These correspond to survey-weighted prevalence of 27.2% (95% CI 25.5% - 29.1%), 12.1% (95% CI 10.9% - 13.4%), and 60.7% (95% CI 58.6% - 62.7%), respectively. Video users were younger on average (52.8, SD 16.2 y) than those with no telehealth (56.3, SD 17.1 y) and audio-only users (58.6, SD 16.8 y). Women comprised a larger share of video (1071/1641, 65.3%) and audio-only (566/876, 64.6%) users than nonusers (2012/3529, 57%). Compared with nonusers, video users more frequently reported higher income (788/1641, 48% vs 1292/3529, 36.6% at ≥US $75,000), current employment (912/1641, 55.6% vs 1807/3529, 51.2%), and insurance coverage (1559/1641, 95.1% vs 3165/3529, 90%). Digital readiness was notably higher among video users: any smart device ownership (1582/1641, 96.4% vs 3091/3529, 87.6%), health or wellness app use (1103/1641, 67.2% vs 1441/3529, 40.8%), and online health information seeking (1367/1641, 83.3% vs 2294/3529, 65%). Satisfaction with home internet for health needs was also higher in the video group (1075/1641, 65.5% extremely or very satisfied vs 1789/3529, 50.7% among nonusers). Geographic composition differed across modalities, with video visits more common among residents of large central metros (684/1641, 41.7%) and large fringe metro or suburban areas (393/1641, 23.9%) compared with nonusers (1220/3529, 34.6% and 769/3529, 21.8%, respectively). White respondents made up a similar share across video, audio-only, and no telehealth groups, while Hispanic respondents were somewhat more represented among audio-only users than among video users or those who did not use telehealth.

XSL•FO
RenderX

Regression Estimates

Key Predictors

Table 2 reports key predictors from survey-weighted LPMs for all 4 contrasts, presenting marginal effects as pp differences. Below we summarize the key patterns; full coefficient estimates and CIs for all covariates appear in Tables S5-S8 in Multimedia Appendix 1.

Table . Key predictors from survey-weighted linear probability models of telehealth outcomes, Health Information National Trends Survey (HINTS 6)[a].

Variables	Video vs audio[b] (n=2069)	Video vs none[c] (n=4317)	Audio vs none[d] (n=3632)	Any vs none[e] (n=5009)
Demographics and socioeconomics				
Male	−0.2 (−5.5 to 5.2; *P*=.94)	−8.5 (−12.4 to −4.6; *P*<.001)	−5.5 (−8.1 to −2.8; *P*<.001)	−9.7 (−14.0 to −5.4; *P*<.001)
White	2.4 (−3.4 to 8.2; *P*=.42)	6.4 (2.3 to 10.6; *P=0.001*)	3.7 (0.7 to 6.7; *P*=.02)	7.1 (2.2 to 12.0; *P<0.001*)
Employed	9.1 (2.6 to 15.5; *P=0.001*	3.8 (−1.6 to 9.2; *P*=.17)	−3.3 (−6.8 to 0.2; *P*=.07)	0.9 (−3.3 to 5.1; *P*=.67)
Health status and access				
Disabled	9.6 (1.6 to 17.6; *P*=.02)	23.7 (17.8 to 29.6; *P*<.001)	11.9 (5.1 to 18.7; *P*<.001)	22.5 (16.1 to 28.8; *P*<.001)
High blood pressure	−1.7 (−7.8 to 4.3; *P*=.58)	5.8 (1.8 to 9.8; *P=0.001*)	5.8 (2.6 to 8.9; *P*<.001)	8.0 (4.8 to 11.2; *P*<.001)
Health insurance	21.2 (13.0 to 29.3; *P*<.001)	13.3 (9.2 to 17.3; *P*<.001)	1.1 (−4.0 to 6.3; *P*=.67)	11.5 (4.6 to 18.4; *P*=.001)
Digital access				
Basic cell only (reference: landline or none)	−20.1 (−33.5 to −6.8; *P=0.001*)	−7.4 (−16.4 to 1.7; *P*=.11)	2.6 (−3.9 to 9.1; *P*=.44)	−4.0 (−10.6 to 2.6; *P*=.23)
Internet: extremely satisfied	8.6 (1.3 to 15.9; *P*=.02)	7.7 (1.1 to 14.4; *P*=.02)	0.1 (−5.4 to 5.6; *P*=.97)	5.1 (−1.3 to 11.5; *P*=.12)
eHealth behaviors				
Used health apps	5.5 (−3.4 to 14.5; *P*=.23)	17.5 (12.5 to 22.6; *P*<.001)	9.1 (5.5 to 12.7; *P*<.001)	18.4 (12.0 to 24.8; *P*<.001)
Seeks health information online	6.6 (−5.5 to 18.7; *P*=.28)	7.2 (1.0 to 13.4; *P*=.02)	1.7 (−1.0 to 4.4; *P*=.22)	7.2 (2.4 to 12.0; *P=0.001*)

[a]Coefficients represent percentage-point differences in probability of outcome (95% CI; *P* value). Models additionally controlled for: age, marital status, income, education, self-rated health, cancer history, census division, urbanicity, Patient-Reported Outcomes Measurement Information System (PROMIS) isolation and meaning scores, other device categories, internet satisfaction levels, health numeracy, and county-level characteristics.

[b]R^2 for “video vs audio” is 0.08.

[c]R^2 for “video vs none” is 0.13.

[d]R^2 for “audio vs none” is 0.07.

[e]R^2 for “any vs none” is 0.12.

Demographics and Socioeconomic Factors

Gender differences were consistent: men were 8 to 10 pp less likely than women to use telehealth across all uptake contrasts (*P*<.001 for video and any telehealth vs none). White respondents had 4 to 7 pp higher use than non-White respondents, but race or ethnicity was not significantly associated with video vs audio modality choice. Adults aged 35 to 49 years showed elevated telehealth use (approximately 6 - 7 pp) compared with older age groups. Among telehealth users, employment and insurance coverage were strongly associated with choosing video over audio (9.1 pp and 21.2 pp, respectively).

Health Status and Needs

Health-related factors showed the largest effect sizes. Disability was strongly associated with telehealth use: 22 to 24 pp for video and any telehealth vs none, and 12 pp for audio-only vs none. Among telehealth users, disability was also associated with choosing video over audio (9.6 pp). Cancer history (6 - 7 pp) and hypertension (6 - 8 pp) were positively associated with all forms of telehealth uptake, but were not significant predictors of choosing video vs audio. Insurance coverage was consistently positive (11 - 13 pp for uptake; 21 pp for video vs audio choice).

Digital readiness

Digital factors differentiated video from audio use. Basic cell phone only (vs smartphone or tablet) was associated with 20 pp lower video use among telehealth users. Internet satisfaction strongly favored video: “extremely” or “very satisfied” users were 8 to 9 pp more likely to choose video over audio. Health app use showed the largest digital association with uptake: 17 to 18 pp for video and any telehealth vs none. Online health information seeking was positively associated with video (7 pp) but not audio-only uptake.

Geographic Variation

Geography was especially salient for audio-only use. Relative to New England, audio-only use was 9 to 19 pp lower across multiple census divisions (East North Central, West North

XSL•FO
RenderX

Central, South Atlantic, and West South Central). For any telehealth vs none, the West North Central showed the largest deficit (−21 pp). Geographic variation for video use was more modest, with only West North Central showing a significant deficit (−13 pp vs New England).

Psychosocial Factors

PROMIS measures showed small associations. Social isolation was weakly associated with any telehealth use (0.2 pp per T-score point). Among telehealth users, greater meaning in life was associated with lower video use (−0.3 pp per point), though this effect was small in magnitude.

Shorrocks-Shapley Decomposition

Explained Variation and Shapley Decomposition

Model fit, expressed as Shapley-based R^2, was modest and varied by contrast: 0.08 (video vs audio), 0.13 (video vs none), 0.07 (audio vs none), and 0.12 (any vs none; Table 3). Decomposition of R^2 across prespecified domains highlighted distinct drivers for modality choice vs telehealth uptake. For video vs audio, digital access was the leading contributor (25.1% of explained variance), followed by SES (23.5%), health literacy and eHealth (15.3%), and geography (13.8%). Together, digital access+health literacy and eHealth accounted for 40.4% of explained variance, underscoring that device ownership, internet satisfaction, and online engagement are central to video over audio.

Table . Shapley decomposition of survey-weighted linear probability models (LPMs)[a].

Variable group	Video vs audio-only, R^2 (%)	Video vs no telehealth, R^2 (%)	Audio-only vs no telehealth, R^2 (%)	Any telehealth vs no telehealth, R^2 (%)
Demographics	0.0029 (3.65)	0.0164 (12.61)	0.0036 (5.13)	0.0142 (11.87)
Socioeconomic status	0.0188 (23.49)	0.0154 (11.83)	0.0049 (7.03)	0.0095 (7.91)
Health status and needs	0.0057 (7.12)	0.0153 (11.79)	0.0148 (21.08)	0.0186 (15.53)
Health system access	0.0030 (3.80)	0.0065 (5.00)	0.0013 (1.83)	0.0057 (4.76)
Geography	0.0110 (13.78)	0.0158 (12.15)	0.0283 (40.49)	0.0237 (19.71)
Urbanicity	0.0011 (1.34)	0.0042 (3.25)	0.0013 (1.91)	0.0038 (3.19)
Psychosocial factors	0.0016 (1.95)	0.0104 (7.97)	0.0011 (1.53)	0.0088 (7.34)
Digital access	0.0201 (25.07)	0.0186 (14.31)	0.0010 (1.36)	0.0099 (8.23)
Health literacy and eHealth	0.0122 (15.27)	0.0248 (19.06)	0.0062 (8.79)	0.0209 (17.45)
Area context	0.0036 (4.52)	0.0026 (2.03)	0.0076 (10.85)	0.0048 (4.03)
Total	0.08 (100)	0.13 (100)	0.07 (100)	0.12 (100)

[a]Shapley decomposition apportions the total R^2 of each model into the marginal contributions from distinct groups of predictor variables. All models were survey-weighted. Variable groups defined as (1) demographics: gender, marital status, age group, and race or ethnicity (White vs non-White); (2) socioeconomic status: household income, education level, and employment status; (3) health status and needs: disability status, self-rated health, cancer history, and high blood pressure history; (4) health system access: insurance status; (5) geography: US Census division; (6) urbanicity: urban, suburban, and rural; (7) psychosocial factors: Patient-Reported Outcomes Measurement Information System (PROMIS)—"Social Isolation" and "Meaning and Purpose" scores; (8) digital access: device ownership, satisfaction with home internet for health needs, and county-level percent broadband access; (9) health literacy and eHealth: self-reported health numeracy, use of health or wellness apps, and online health information-seeking behavior; and (10) area context: county-level characteristics including total population, poverty rate, unemployment rate, percent without a high school diploma, and percent of population aged ≥65 years.

For video vs no telehealth, the pattern remained digitally oriented but more diffuse: health literacy and eHealth (19.1%) and digital access (14.3%) ranked highly, as did demographics (12.6%), geography (12.2%), SES (11.8%), and health status and needs (11.8%). The digital stack contributed 33.4%, while demographics and SES jointly explained 24.4%, and geography, urbanicity, and area context explained 17.4%.

For audio-only vs no telehealth, the leading domain shifted decisively to place: geography alone explained 40.5%, followed by health status and needs (21.1%) and area context (10.9%). The digital stack was comparatively small (10.2%), while the aggregate place–based share reached 53.2%, indicating that where people live (division and local socioeconomic context) is especially salient for audio uptake. Finally, for any telehealth vs none, contributions were balanced across domains: geography (19.7%) and health literacy and eHealth (17.5%) led, followed by health status and needs (15.5%) and demographics (11.9%). The digital stack accounted for 25.7%, place-based factors for 26.9%, and demographics+SES for 19.8%. Across results, insurance and urbanicity contributed modest shares, whereas psychosocial factors contributed minimally (2%) for video vs audio but were more substantial for video vs none (8%) and

XSL•FO
RenderX

any vs none (7.3%), suggesting a limited role in modality choice but some role in overall uptake.

Missing Data Patterns and Multiple Imputation Sensitivity Analysis

Item-level missingness varied across variables. Key continuous variables had moderate missingness: PROMIS Isolation T-scores (n=334, 5.3%), PROMIS Meaning T-scores (n=314, 5%), and self-rated health (n=234, 3.7%). Sociodemographic variables with missingness included: employment status (n=390, 6.2%), disability status (n=390, 6.2%), race or ethnicity (n=619, 9.9%), and income (n=732, 11.7%). For categorical variables with modest missingness, we used a "missing-as-its-own-category" approach: cancer history (5.9% missing or unknown), digital access indicators (0.6% - 14.8% missing or not applicable), and health app use or online health information seeking (3.7% - 5.8% missing).

The null hypothesis of the MCAR test [48] was rejected (χ^2_{530}=982.1; P<.001), indicating that data were not MCAR. As a supplementary analysis, we conducted logistic regression examining whether missingness on income (the variable with the highest missingness) was predicted by fully observed covariates; this analysis also indicated systematic missingness (χ^2_1=1354.6, P<.001), with missingness associated with age, census division, cancer history, and insurance status. Together, these results indicate that missingness is related to observed covariates. Accordingly, we conducted multiple imputation by chained equations under a MAR assumption as a recommended sensitivity approach for item nonresponse in observational studies [49,50].

Multiple Imputation Sensitivity Analysis

Given the MAR finding, we conducted a sensitivity analysis using multiple imputation by chained equations with m=20 imputations. Imputed variables included income, race, employment, disability, self-rated health, gender, marital status, education, and PROMIS scores. Imputation models included all analysis variables plus auxiliary predictors (age, census division, urbanicity, cancer history, high blood pressure, and insurance status). Results from multiple imputations analyses were consistent with complete-case estimates, supporting the robustness of our primary findings (see Table S1 in Multimedia Appendix 1). Complete multiple imputations regression results with 95% CIs are provided in Tables S2–S4 and accompanying data files in Multimedia Appendix 1.

Discussion

Principal Findings

In a nationally representative survey of US adults in 2022, we found that overall telehealth use (any vs none) showed large regression effect sizes for health needs (eg, disability, +22.5 pp) and insurance coverage (+11.5 pp), whereas video vs audio-only modality choice among telehealth users was most strongly differentiated by digital readiness (device access and satisfaction with home internet) and eHealth behaviors. Shorrocks-Shapley decomposition revealed distinct variance patterns: for any telehealth vs none, geography (19.7%), digital factors (25.7%), and health status (15.5%) all contributed substantially; digital domains explained the largest share of variation in video vs audio choice (40.4%), while geography and area context contributed most to explaining audio-only uptake (53.2% combined for audio vs none). Men were less likely than women to use telehealth, consistent with gender differences in health care utilization [51], and telehealth users with limited device capability were more likely to rely on audio-only visits. Hypertension was positively associated with telehealth uptake but not with video vs audio choice, potentially reflecting that patients with chronic conditions requiring frequent visits benefit from any telehealth modality without a strong preference for video.

This study advances prior work in 3 ways. First, we distinguish modality choice (video vs audio) from uptake (any telehealth vs none) in a unified, survey-weighted framework, showing that digital access+health literacy and eHealth domains together explain over 40% of the variance in video vs audio choice (40.4%) and one-third in video vs none (33.4%). Second, we explicitly separate individual digital behaviors (app use and health information seeking) from device ownership and satisfaction, demonstrating that behavioral engagement is strongly linked to uptake above and beyond device availability. Third, we link person-level data with area context (division, urbanicity, and county characteristics) and apportion explained variance across domains using a Shorrocks-Shapley decomposition adapted to complex surveys—offering decision makers or researchers a transparent view of the relative leverage points for potential interventions.

A key strength of this study relative to the telehealth disparities literature is the granularity of the person-level technology data available in HINTS 6, which allows us to operationalize digital readiness beyond proxies such as binary internet access or area-level broadband coverage alone. Specifically, HINTS captures multiple, policy-relevant dimensions of digital readiness—device access and ownership, perceived adequacy of home internet for meeting health-related needs, and behavioral engagement with digital health. We leveraged this richness in a two-level way to avoid inconsistent aggregation: (1) we modeled these measures as distinct indicators in the regression analyses to preserve interpretability and map directly onto different intervention levers, and (2) we grouped them into prespecified domains in the Shorrocks-Shapley decomposition to provide a coherent summary of the relative explanatory contribution of the broader digital readiness "stack" compared with sociodemographic, clinical, and place-based domains. Future work could apply psychometric methods to develop and validate a composite digital readiness scale. In contrast, our indicator plus–domain approach prioritizes actionable interpretability for policy or practice [52,53].

The decomposition analysis shows 3 patterns. First, the digital access+health literacy and eHealth consistently explain a large fraction of variation, but its weight is contrast-specific: it is largest for video vs audio and video vs none, smaller for any vs none, and smallest for audio vs none. Second, geography+urbanicity+area context is most influential for audio vs none, moderate for any vs none and video vs none, and least for video vs audio. Third, the combined health status and

XSL•FO
RenderX

needs+health system access contribution is largest when the contrast is framed as use vs nonuse (audio vs none and any vs none), and smaller when the choice is video vs audio. Finally, because the effect size for social isolation was small and imprecise across modality contrasts, we interpret this as suggestive at best. It may reflect unmeasured factors (eg, mobility or transport barriers or care preferences) correlated with isolation rather than a direct effect.

Taken together, these patterns imply that (1) modality is largely a function of digital readiness and skills, while (2) adoption (any use vs none) is a joint function of clinical need, geography or place, and digital readiness. Despite modest R^2—typical for behavioral use models—the decomposition points to clear, policy-addressable levers: device access and digital engagement for video-based care; and place-based strategies (alongside coverage and need) to broaden overall telehealth uptake.

A prior nationwide study in older adults shows that living in a neighborhood at the lowest tertile of broadband internet subscription is associated with a significantly lower likelihood of using telehealth [54]. We similarly observe strong associations between digital readiness and video use. Lacking a smart device is strongly associated with choosing audio over video but was not significantly associated with any telehealth vs none in adjusted models. In uptake models, app use and online health information seeking are positive, whereas device category is not significantly associated with any telehealth vs none. However, our adjusted county-level percent broadband access shows small associations with any and audio telehealth. This likely reflects differences between availability vs quality or speed and aggregation by counties, which might dilute neighborhood heterogeneity. Together, the findings underscore that closing gaps requires more than network presence; adoption and affordability are important. Our modality-specific results align: device richness and internet satisfaction push use toward video over audio, while those with fewer devices or lower satisfaction are more likely to rely on audio. This is consistent with a study that emphasized the importance of optimal broadband connectivity for improved mental health access and raised quality considerations as visual cues may provide useful clinical information and improve patient-clinician rapport/access to video-based care [55].

Furthermore, a recent study by Tilhou et al [56] reported that telehealth expansion exposed differences in the use of telehealth services that persisted even among beneficiaries with high-speed internet and that expansion of telehealth service or access to high-speed internet might not close gaps in use. Despite strong associations for digital access and eHealth behaviors, geography and social context still explain sizable variance. Thus, high-quality connectivity is likely important but not sufficient without attention to coverage, device access, and digital skills or engagement. Our domain-level results empirically support that digital inclusion is considered a “super” social determinant of health [37] as the device access and skills or engagement carry a substantial explanatory weight, while area context signals persistent place-based differences. The decomposition attributes important shares to geography and smaller but nontrivial shares to area context even after detailed individual controls, reinforcing that place effects remain operative for digital health adoption, consistent with a study conducted in Chicago which found that segregation and concentrated poverty influence why residents do not have home internet access [57].

Policy and Practice Implications

Our findings point to complementary levers. First, sustaining insurance coverage and targeting populations with elevated needs (eg, disability and chronic conditions) may yield large absolute gains in telehealth use. Second, programs that upgrade device access and digital engagement (eg, app enablement and coaching in online health information seeking) should accompany broadband deployment; these address the “devices” and “skills or training” pillars alongside connectivity. Third, where home connectivity is constrained, offering high-bandwidth alternatives (community clinics and libraries) and maintaining audio-only options preserves access while longer-term infrastructure and affordability gaps are addressed. Finally, because geography and area context explain substantial variance, especially for audio telehealth–place-based investments (including affordability supports) are warranted. These implications are directly supported by our Shorrocks-Shapley decomposition, which quantifies that digital readiness is the dominant explanatory domain for video vs audio modality choice; whereas place-based factors play a larger role for audio-only uptake.

The overarching goal of telehealth is to democratize high-value care by overcoming geographic and logistical barriers. However, our findings indicate that this promise might remain unfulfilled. The significant differences identified in our nationally representative analysis reveal that telehealth, in its current form, may risk exacerbating the very differences it is intended to solve [58]. These results carry direct implications for policymakers, health care systems, and technology developers aiming to ensure the expansion of virtual care is available to all individuals.

In our data, “geography” primarily captures regional policy environments, market or provider supply for virtual care, as well as digital infrastructure and adoption. “Underserved” in this context refers to areas with lower broadband adoption or affordability and speeds, higher area deprivation, rural or micropolitan location, and limited provider supply for video-capable care. For audio only vs no telehealth, geography explained the largest share of the model’s variance—substantially more than in other contrasts—underscoring that a one-size-fits-all strategy is insufficient. The significant reductions in the probability of telehealth use by geography, for example, for residents of the Midwest and South compared to the Northeast point to potential structural issues. This aligns with research highlighting how state-level policies, particularly those governing reimbursement and physician licensure, create a fragmented regulatory landscape that hinders uniform adoption [59]. To address this, federal and state policies should prioritize targeted investments in underserved regions [60,61]. This includes not only expanding community infrastructure like broadband internet, a known barrier [60-64], but also incentivizing provider adoption through sustainable reimbursement models and advancing interstate

XSL•FO
RenderX

licensure compacts to improve the supply of virtual care in these areas.

Overall, since affordability is a binding constraint for adoption, recent federal policy changes—including 2025 adjustments to Broadband Equity, Access, and Deployment program affordability requirements and the termination of Digital Equity Act grant programs—threaten to undermine gains in telehealth use, particularly in high-need regions where subsidized connectivity and digital inclusion supports have been critical [65,66].

Second, our analysis provides robust empirical weight to the emerging concept of digital determinants of health [63,64]. We found that the combined domains of health literacy and digital access are as important as geography in explaining telehealth use. Critically, this difference is not merely about access to a device but about the capability to use it effectively for health. We did not find consistent negative associations for low health numeracy after adjustment. Instead, app use and online information seeking were the behavioral factors most strongly linked to higher telehealth uptake. One interpretation is that these behaviors proxy digital readiness—familiarity with app-based health workflows similar to those required for scheduling, check-in, and joining video visits. Patient studies of video visits report that technical and setup challenges (eg, connecting to the visit, audio or video problems, and confusing initial setup) remain common and systematic reviews identify limited technical literacy and technology barriers as recurring obstacles to telehealth use in older adults. Accordingly, this pattern is hypothesis-generating and suggests that the usability burden of telehealth platforms could disproportionately deter patients with less prior digital health experience [5,8,9,67].

Overall, initiatives should address both digital access and digital literacy [68]. This requires a 2-pronged approach: funding for community-based digital health literacy programs and a push for user-centered design principles in the development of patient-facing technologies to reduce overall burden on users.

Limitations

Cross-sectional data preclude causal inference and telehealth measures were self-reported and lacked encounter frequency or quality. County-level broadband access may mis-measure adoption, affordability, and speed (and operates at a coarser geography); collinearity with urbanicity and digital behaviors may attenuate or reverse associations for the county broadband variable. Model R^2 values are modest, indicating unmeasured determinants (eg, provider supply). Future research should examine finer gradations of educational measure to distinguish groups that may face distinct barriers to telehealth adoption. Nevertheless, the survey-weighted national scope, explicit modality contrasts, and Shorrocks-Shapley decomposition across domains provide a comprehensive, policy-relevant portrait of telehealth determinants in 2022. Furthermore, understanding intersectional perspectives across race, income, education, and broadband availability might be an area of fruitful research.

Conclusions

Video telehealth uptake in the United States reflects intersecting determinants with modality choice driven most strongly by digital readiness and overall uptake shaped by digital factors, place-based characteristics, and health needs. Policies that simultaneously advance connectivity, affordability, device access, and digital skills support are likely to narrow persistent gaps in telehealth use across modalities. Methodologically, this study advances the telehealth disparities literature by applying Shorrocks-Shapley decomposition to survey-weighted national data to quantify the relative explanatory contribution of determinant domains rather than examining predictors in isolation. Our approach provides a domain-level ranking of where disparities are most strongly concentrated, distinguishing digital readiness mechanisms from place-based constraints. Leveraging HINTS' rich person-level technology measures, we show that digital access and eHealth behaviors account for a larger share of explained variation in video-vs-audio modality choice than traditional sociodemographic characteristics. These findings support a dual strategy: maintain reimbursable audio-only care to preserve access while digital constraints persist and pair sustained coverage with targeted investments in devices, affordable high-speed connectivity, and digital skills supports to enable more equitable access to video-based telehealth.

Acknowledgments

No generative artificial intelligence tools were used to write or edit the scientific content of this manuscript. Any grammar or spell checking tools did not alter the authors' interpretation of findings.

Funding

This research has been funded by Health Resources and Services Administration (grant U3GRH40001)

Data Availability

This study used the Health Information National Trends Survey restricted-use dataset containing county identifiers; access requires approval through the National Cancer Institute restricted data request process and a data use agreement, and the authors are not permitted to redistribute the restricted-use file. County-level contextual indicators were obtained from the American Community Survey, NIMHD HD Pulse broadband indicators, and the CDC/ATSDR Social Vulnerability Index. Supplementary data files are provided in Appendix D in Multimedia Appendix 1, including complete LPM regression results with 95% CIs for all four contrasts (LPM Video vs Audio Complete.csv, LPM Video vs NoTelehealth Complete.csv, LPM Audio vs NoTelehealth Complete.csv, LPM AnyTelehealth vs NoTelehealth Complete.csv); multiple imputation results (MI regression results CI.csv, MI vs CompleteCase

XSL•FO
RenderX

comparison.csv, MI regression results.csv); and missingness diagnostics (littles mcar test results.txt). Analysis code will be made available in a public repository upon acceptance; during peer review, code is available from the corresponding author upon reasonable request.

Authors' Contributions

Concept and design: C Bolbocean.
Data acquisition of data: C Bolbocean, C Bogulski.
Drafting of the manuscript: C Bolbocean.
Statistical analysis: C Bolbocean.
Interpretation of data: C Bolbocean, CH, C Bogulski.
Critical revision of the paper for important intellectual content: C Bolbocean, CH, C Bogulski.
Obtaining funding: C Bolbocean, CH, C Bogulski.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials including definition of variables, complete linear probability model results, multiple imputation sensitivity analysis, and supplementary data files.
[PDF File, 265 KB - jmir_v28i1e81879_app1.pdf]

References

1. What is telehealth? Health Resources and Services Administration, US Department of Health and Human Services. 2024. URL: https://www.hrsa.gov/telehealth/what-is-telehealth [accessed 2025-06-10]
2. Kichloo A, Albosta M, Dettloff K, et al. Telemedicine, the current COVID-19 pandemic and the future: a narrative review and perspectives moving forward in the USA. Fam Med Community Health 2020 Aug;8(3):e000530. [doi: 10.1136/fmch-2020-000530] [Medline: 32816942]
3. Bashshur RL, Howell JD, Krupinski EA, Harms KM, Bashshur N, Doarn CR. The empirical foundations of telemedicine interventions in primary care. Telemed J E Health 2016 May;22(5):342-375. [doi: 10.1089/tmj.2016.0045] [Medline: 27128779]
4. Lyles CR, Schillinger D, Sarkar U. Connecting the dots: health information technology expansion and health disparities. PLoS Med 2015 Jul;12(7):e1001852. [doi: 10.1371/journal.pmed.1001852] [Medline: 26172977]
5. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018 Aug 1;25(8):1080-1088. [doi: 10.1093/jamia/ocy052] [Medline: 29788380]
6. Kates J, Ranji U, Beamesderfer A, Salganicoff A, Dawson L. Health and access to care and coverage for lesbian, gay, bisexual, and transgender (LGBT) individuals in the US. : Kaiser Family Foundation; 2018 URL: https://collections.nlm.nih.gov/master/borndig/101740255/Issue-Brief-Health-and-Access-to-Care-and-Coverage-for-LGBT-Individuals-in-the-US.pdf [accessed 2026-02-14]
7. Nouri S, Khoong EC, Lyles CR, Karliner L. Addressing equity in telemedicine for chronic disease management during the COVID-19 pandemic. NEJM Catalyst Innov Care Deliv 2020;1(3). [doi: 10.1056/CAT.20.0123]
8. Ternes S, Lavin L, Vakkalanka JP, et al. The role of increasing synchronous telehealth use during the COVID-19 pandemic on disparities in access to healthcare: a systematic review. J Telemed Telecare 2025 Sep;31(8):1091-1113. [doi: 10.1177/1357633X241245459] [Medline: 38646804]
9. Chang JE, Lai AY, Gupta A, Nguyen AM, Berry CA, Shelley DR. Rapid transition to telehealth and the digital divide: implications for primary care access and equity in a post‐COVID era. Milbank Q 2021 Jun;99(2):340-368. [doi: 10.1111/1468-0009.12509] [Medline: 34075622]
10. Garfan S, Alamoodi AH, Zaidan BB, et al. Telehealth utilization during the COVID-19 pandemic: a systematic review. Comput Biol Med 2021 Nov;138:104878. [doi: 10.1016/j.compbiomed.2021.104878] [Medline: 34592585]
11. Vakkalanka JP, Gadag K, Lavin L, et al. Telehealth use and health equity for mental health and substance use disorder during the COVID-19 pandemic: a systematic review. Telemed J E Health 2024 May;30(5):1205-1220. [doi: 10.1089/tmj.2023.0588] [Medline: 38227387]
12. Eberly LA, Khatana SAM, Nathan AS, et al. Telemedicine outpatient cardiovascular care during the COVID-19 pandemic: bridging or opening the digital divide? Circulation 2020 Aug 4;142(5):510-512. [doi: 10.1161/CIRCULATIONAHA.120.048185] [Medline: 32510987]
13. Forcino RC, Johnson JK, Elwyn G. Telehealth modality preferences for video and voice-only visits among US clinicians and the public: cross-sectional, web-based survey study. J Med Internet Res 2025 Jun 13;27:e72276. [doi: 10.2196/72276] [Medline: 40513030]

XSL•FO
RenderX

14. Samson LW, Couture SJ, Creedon TB, Jacobus-Kantor L, Sheingold S. Updated Medicare FFS telehealth trends by beneficiary characteristics, visit specialty, and state, 2019-2021. : Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services; 2023 Mar URL: https://aspe.hhs.gov/sites/default/files/documents/cb83f6f25c25c3a3529807f23cd2327d/medicare-telehealth-updated-trends-report.pdf [accessed 2026-03-06]
15. Saharkhiz M, Rao T, Parker-Lue S, Borelli S, Johnson K, Cataife G. Telehealth expansion and Medicare beneficiaries' care quality and access. JAMA Netw Open 2024 May 1;7(5):e2411006. [doi: 10.1001/jamanetworkopen.2024.11006] [Medline: 38739388]
16. What to know about medicare coverage of telehealth. Kaiser Family Foundation. 2024. URL: https://www.kff.org/medicare/what-to-know-about-medicare-coverage-of-telehealth/ [accessed 2025-10-14]
17. Lucas JW, Villarroel MA. Telemedicine use among adults: United States, 2021. NCHS Data Brief 2022 Oct(445):1-8. [Medline: 36255940]
18. Chang E, Penfold RB, Berkman ND. Patient characteristics and telemedicine use in the US, 2022. JAMA Netw Open 2024 Mar 4;7(3):e243354. [doi: 10.1001/jamanetworkopen.2024.3354] [Medline: 38517438]
19. FCC launches Affordable Connectivity Program. Federal Communications Commission. 2021 Dec. URL: https://www.fcc.gov/fcc-launches-affordable-connectivity-program [accessed 2026-02-14]
20. Affordable Connectivity Program. Federal Communications Commission. 2024. URL: https://www.fcc.gov/acp [accessed 2026-02-14]
21. Half of ACP-eligible households still unaware of the program. Benton Institute for Broadband & Society. 2023 Mar. URL: https://www.benton.org/blog/half-acp-eligible-households-still-unaware-program [accessed 2026-02-14]
22. ACP enrollment and claims tracker. Universal Service Administrative Company. 2024. URL: https://www.usac.org/about/affordable-connectivity-program/acp-enrollment-and-claims-tracker [accessed 2026-02-14]
23. Affordable connectivity program wind down. Federal Communications Commission. 2024. URL: https://www.fcc.gov/acp-wind-down [accessed 2026-03-06]
24. Annual consumer survey fielded november 1–20, 2024; 68,000+ respondents. National Lifeline Association. 2025. URL: https://eqm2782zctq.exactdn.com/wp-content/uploads/2025/01/NaLA-Annual-Consumer-Survey-1.22.25-FINAL.pdf [accessed 2026-03-06]
25. van Deursen AJ, van Dijk JA. The first-level digital divide shifts from inequalities in physical access to inequalities in material access. New Media Soc 2019 Feb;21(2):354-375. [doi: 10.1177/1461444818797082]
26. Rodriguez JA, Betancourt JR, Sequist TD, Ganguli I. Differences in the use of telephone and video telemedicine visits during the COVID-19 pandemic. Am J Manag Care 2021 Jan;27(1):21-26. [doi: 10.37765/ajmc.2021.88573] [Medline: 33471458]
27. Sachs JW, Graven P, Gold JA, Kassakian SZ. Disparities in telephone and video telehealth engagement during the COVID-19 pandemic. JAMIA Open 2021 Jul;4(3):ooab056. [doi: 10.1093/jamiaopen/ooab056] [Medline: 34632322]
28. Shorrocks AF. Decomposition procedures for distributional analysis: a unified framework based on the Shapley value. J Econ Inequal 2013 Mar;11(1):99-126. [doi: 10.1007/s10888-011-9214-z]
29. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. The Lancet 2007 Oct;370(9596):1453-1457. [doi: 10.1016/S0140-6736(07)61602-X]
30. Blake KD, Moser RP, D'Angelo H, Gaysynsky A, Vanderpool RC. The evolution of NCI's Health Information National Trends Survey: methods, data, and future directions. J Natl Cancer Inst 2025 Jun 1;117(6):1110-1116. [doi: 10.1093/jnci/djae317] [Medline: 39657961]
31. Finney Rutten LJ, Blake KD, Skolnick VG, Davis T, Moser RP, Hesse BW. Data resource profile: the national cancer institute's health information national trends survey (HINTS). Int J Epidemiol 2020 Feb 1;49(1):17-17j. [doi: 10.1093/ije/dyz083] [Medline: 31038687]
32. American Community Survey data. US Census Bureau. URL: https://www.census.gov/programs-surveys/acs/data.html [accessed 2026-02-14]
33. Broadband access for United States by county: all races (includes Hispanic/Latino), both sexes, all ages, 2019-2023 (sorted by percent). National Institute on Minority Health and Health Disparities. URL: https://hdpulse.nimhd.nih.gov/data-portal/physical/table?age=001&age_options=ageall_1&demo=235&demo_options=broadband_1&physicaltopic=006 [accessed 2026-02-14]
34. SVI data & documentation download. Centers for Disease Control and Prevention (CDC) and Agency for Toxic Substances and Disease Registry (ATSDR). URL: https://www.atsdr.cdc.gov/place-health/php/svi/svi-data-documentation-download.html [accessed 2026-03-06]
35. Lam K, Lu AD, Shi Y, Covinsky KE. Assessing telemedicine unreadiness among older adults in the United States during the COVID-19 pandemic. JAMA Intern Med 2020 Oct 1;180(10):1389-1391. [doi: 10.1001/jamainternmed.2020.2671] [Medline: 32744593]
36. van Dijk J. Digital divide research, achievements and shortcomings. Poetics 2006 Aug;34(4-5):221-235. [doi: 10.1016/j.poetic.2006.05.004]

XSL•FO
RenderX

37. Sheon A, Khoong EC. Pillars, policies, and plausible pathways linking digital inclusion and health equity. Health Aff 2024. [doi: 10.1377/hpb20240327.505]
38. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH roadmap cooperative group during its first two years. Med Care 2007 May;45(5 Suppl 1):S3-S11. [doi: 10.1097/01.mlr.0000258615.42478.55] [Medline: 17443116]
39. Hays RD, Bjorner JB, Revicki DA, Spritzer KL, Cella D. Development of physical and mental health summary scores from the Patient-Reported Outcomes Measurement Information System (PROMIS) global items. Qual Life Res 2009 Sep;18(7):873-880. [doi: 10.1007/s11136-009-9496-9] [Medline: 19543809]
40. Nicholson NR. Social isolation in older adults: an evolutionary concept analysis. J Adv Nurs 2009 Jun;65(6):1342-1352. [doi: 10.1111/j.1365-2648.2008.04959.x] [Medline: 23163719]
41. Hawkley LC, Cacioppo JT. Loneliness matters: a theoretical and empirical review of consequences and mechanisms. Ann Behav Med 2010 Oct;40(2):218-227. [doi: 10.1007/s12160-010-9210-8] [Medline: 20652462]
42. Roepke AM, Jayawickreme E, Riffle OM. Meaning and health: a systematic review. Appl Res Qual Life 2014 Dec;9(4):1055-1079. [doi: 10.1007/s11482-013-9288-9]
43. Lee MJ, Lee G, Choi JY. Linear probability model revisited: why it works and how it should be specified. Sociol Methods Res 2025;54(1):173-186. [doi: 10.1177/00491241231176851]
44. Wooldridge JM. Econometric Analysis of Cross Section and Panel Data, 2nd edition: MIT press; 2002.
45. Moser RP, Naveed S, Cantor D, et al. Health Information National Trends Survey: Integrative analytic methods using population-level cross-sectional data. : National Cancer Institute; 2013 URL: https://hints.cancer.gov/docs/hints_ida_report.pdf [accessed 2026-02-14]
46. Suo Z, Shao L, Lang Y. A study on the factors influencing the utilization of public health services by China's migrant population based on the Shapley value method. BMC Public Health 2023 Nov 24;23(1):2328. [doi: 10.1186/s12889-023-17193-3] [Medline: 38001411]
47. Juarez FWC. Shapley2: Stata module to compute additive decomposition of estimation statistics by regressors or groups of regressors. Statistical Software Components. 2015. URL: https://ideas.repec.org/c/boc/bocode/s457543.html [accessed 2026-02-14]
48. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988 Dec;83(404):1198-1202. [doi: 10.1080/01621459.1988.10478722]
49. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009 Jun 29;338:b2393. [doi: 10.1136/bmj.b2393] [Medline: 19564179]
50. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med 2011 Feb 20;30(4):377-399. [doi: 10.1002/sim.4067] [Medline: 21225900]
51. Bertakis KD, Azari R, Helms LJ, Callahan EJ, Robbins JA. Gender differences in the utilization of health care services. J Fam Pract 2000 Feb;49(2):147-152. [Medline: 10718692]
52. Greco S, Ishizaka A, Tasiou M, Torrisi G. On the methodological framework of composite indices: a review of the issues of weighting, aggregation, and robustness. Soc Indic Res 2019 Jan;141(1):61-94. [doi: 10.1007/s11205-017-1832-9]
53. Nardo M, Saisana M, Saltelli A, Tarantola S, Hoffmann A, Giovannini E. Handbook on Constructing Composite Indicators: Methodology and User Guide: OECD Publishing; 2008. [doi: 10.1787/9789264043466-en]
54. Okoye SM, Mulcahy JF, Fabius CD, Burgdorf JG, Wolff JL. Neighborhood broadband and use of telehealth among older adults: cross-sectional study of national survey data linked with census data. J Med Internet Res 2021 Jun 14;23(6):e26242. [doi: 10.2196/26242] [Medline: 34125071]
55. O'Shea AMJ, Howren MB, Mulligan K, Haraldsson B, Shahnazi A, Kaboli PJ. Quantifying the digital divide: associations of broadband internet with tele-mental health access before and during the COVID-19 pandemic. J Gen Intern Med 2023 Jul;38(Suppl 3):832-840. [doi: 10.1007/s11606-023-08120-8] [Medline: 37340258]
56. Tilhou AS, Jain A, DeLeire T. Telehealth expansion, internet speed, and primary care access before and during COVID-19. JAMA Netw Open 2024 Jan 2;7(1):e2347686. [doi: 10.1001/jamanetworkopen.2023.47686] [Medline: 38180762]
57. Mossberger K, Tolbert CJ, Bowen D, Jimenez B. Unraveling different barriers to internet use: urban residents and neighborhood effects. Urban Aff Rev 2012;48(6):771-810. [doi: 10.1177/1078087412453713]
58. SteelFisher GK, McMurtry CL, Caporello H, et al. Video telemedicine experiences in COVID-19 were positive, but physicians and patients prefer in-person care for the future: study examines patient and physician opinion of telemedicine experiences during COVID-19. Health Aff 2023;42(4):575-584. [doi: 10.1377/hlthaff.2022.01027]
59. Turner Lee N, Karsten J, Roberts J. Removing regulatory barriers to telehealth before and after COVID-19. : Brookings Institution; 2020 URL: https://www.brookings.edu/wp-content/uploads/2020/05/Removing-barriers-to-telehealth-before-and-after-COVID-19_PDF.pdf [accessed 2026-02-14]
60. Sheerah HA, AlSalamah S, Alsalamah SA, et al. The rise of virtual health care: transforming the health care landscape in the Kingdom of Saudi Arabia: a review article. Telemed J E Health 2024 Oct;30(10):2545-2554. [doi: 10.1089/tmj.2024.0114] [Medline: 38984415]

XSL•FO
RenderX

61. Epstein JA, Lkhagvajav Z, Young T, Bertram A, Yeh HC, Taylor CO. Will the doctor “see” you now? The development and implementation of a targeted telemedicine system for primary care. ACI open 2023 Jul;7(2):e71-e78. [doi: 10.1055/s-0043-1776038] [Medline: 37900978]
62. Kane CK, Gillis K. The use of telemedicine by physicians: still the exception rather than the rule. Health Aff (Millwood) 2018 Dec;37(12):1923-1930. [doi: 10.1377/hlthaff.2018.05077] [Medline: 30633670]
63. Jahnel T, Dassow HH, Gerhardus A, Schüz B. The digital rainbow: digital determinants of health inequities. Digit Health 2022;8:20552076221129093. [doi: 10.1177/20552076221129093] [Medline: 36204706]
64. Chidambaram S, Jain B, Jain U, et al. An introduction to digital determinants of health. PLOS Digit Health 2024 Jan;3(1):e0000346. [doi: 10.1371/journal.pdig.0000346] [Medline: 38175828]
65. Broadband equity, access, and deployment (BEAD) program: BEAD restructuring policy notice. : National Telecommunications and Information Administration; 2025 URL: https://www.ntia.gov/sites/default/files/2025-06/bead-restructuring-policy-notice.pdf [accessed 2026-02-14]
66. NTIA broadband programs: semiannual status report. : US Department of Commerce, Office of Inspector General; 2025 URL: https://www.oig.doc.gov/wp-content/OIGPublications/OIG-25-031-I_FinalReport-SECURED.pdf [accessed 2026-02-14]
67. Kruse C, Fohn J, Wilson N, Nunez Patlan E, Zipp S, Mileski M. Utilization barriers and medical outcomes commensurate with the use of telehealth among older adults: systematic review. JMIR Med Inform 2020 Aug 12;8(8):e20359. [doi: 10.2196/20359] [Medline: 32784177]
68. Roberts ET, Mehrotra A. Assessment of disparities in digital access among Medicare beneficiaries and implications for telemedicine. JAMA Intern Med 2020 Oct 1;180(10):1386-1389. [doi: 10.1001/jamainternmed.2020.2666] [Medline: 32744601]

Abbreviations

ACP: Affordable Connectivity Program
FIPS: Federal Information Processing Series
HINTS: Health Information National Trends Survey
HINTS: Health Information National Trends Survey
LPM: linear probability model
MAR: missing at random
MCAR: missing completely at random
PHE: public health emergency
pp: percentage point
PROMIS: Patient-Reported Outcomes Measurement Information System
SES: socioeconomic status
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 05.Aug.2025; peer-reviewed by A Sheon, M Okonkwo, O Okunbor; accepted 30.Jan.2026; published 18.Mar.2026.

Please cite as:
Bolbocean C, Hayes C, Bogulski C
Telehealth Use and Modality Choice Among US Adults: Shorrocks-Shapley Decomposition of a 2022 Cross-Sectional National Survey
J Med Internet Res 2026;28:e81879
URL: https://www.jmir.org/2026/1/e81879
doi:10.2196/81879

© Corneliu Bolbocean, Corey Hayes, Cari Bogulski. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Attitudes Toward Video Consultations From the Perspective of Physicians and Psychotherapists in German Outpatient Care After the COVID-19 Pandemic: Survey Study

Lara Kleinschmidt[1], MSc; Juergen Wasem[1], Prof Dr; Nikola Blase[1], Dr med; Beatrice Nauendorf[2], MSc; Juliane Malsch[2], MSc; Matthias Brittner[3], MA; Paul Brandenburg[4], MA; André Aeustergerling[5], MA; Theresa Hüer[1], Dr rer pol

[1]Institute for Health Care Management and Research, University of Duisburg-Essen, Thea-Leymann-Straße 9, Essen, Germany

[2]Kassenärztliche Vereinigung Berlin, Berlin, Germany

[3]Kassenärztliche Vereinigung Westfalen-Lippe, Dortmund, Germany

[4]Kassenärztliche Vereinigung Schleswig-Holstein, Bad Segeberg, Germany

[5]Kassenärztliche Vereinigung Mecklenburg-Vorpommern, Schwerin, Germany

Corresponding Author:
Lara Kleinschmidt, MSc
Institute for Health Care Management and Research, University of Duisburg-Essen, Thea-Leymann-Straße 9, Essen, Germany

Abstract

Background: Although video consultations (VCs) are permitted in German outpatient care and have seen a notable rise during the COVID-19 pandemic, their use still does not seem to have become established in Germany.

Objective: This survey aims to evaluate the attitudes of physicians and psychotherapists with regard to the use of VC after the COVID-19 pandemic, in particular in the context of types of treatment and suitable medical fields.

Methods: A standardized questionnaire was sent out to all 34,095 physicians and psychotherapists in 4 German regions. The analysis consisted of both descriptive and inferential statistics. Subgroup analysis included gender, age groups, community size of practice location, VC experience, type and ownership of practice, and area of medical care. Binary logistic regression was conducted to determine whether physicians' and psychotherapists' individual factors, organizational factors, or area of medical care were associated with at least monthly VC provision or interest in VC provision.

Results: The response rate was 17.9%, including a total of 5930 participants in the analysis. About 40% (2216/5863) of the physicians and psychotherapists surveyed stated that they offer VC at least once a month. In the area of medical care, the odds ratio (OR) of at least monthly VC provision in psychotherapeutic care was about 8.2 (95% CI 7.4 - 1.64; *P*<.001) compared to primary care, whereas in specialist care, the odds for monthly VC provision were approximately 50% lower than in primary care (OR 0.5, 95% CI 0.43-0.59; *P*<.001). Further, female participants have higher odds to provide VC at least once a month (OR 1.163, 95% CI 1.01 - 1.34; *P*=.03). The odds for monthly VC provision in older age groups are approximately 60% higher than in the <40 years old age group (OR 0.41, 95% CI 0.32-0.52; *P*<.001). Around 80% (4347/5442) of the participants expressed interest in VC use. The most common occasions for which treatment by VC was reported to be suitable were discussing test results (1422/1896, 75.0%), taking the patient's medical history (1195/2147, 55.7%), issuing prescriptions for drugs and remedies (793/1204, 65.9%), and the issuing of incapacity certificates for work (677/1042, 65.0%).

Conclusions: There has been an increase in the self-reported uptake of VC among physicians and psychotherapists compared to prepandemic levels, although this remains at a relatively low level in primary and specialist care. A significant proportion of doctors and psychotherapists have expressed an interest in using VC after the pandemic period. However, this self-reported use is not yet reflected in actual usage data, suggesting the need for further investigation into the underlying factors influencing the gap and identifying potential enablers. Further, these self-assessments by service providers on suitable types of treatment and suitable medical fields can inform political decision-making.

(J Med Internet Res 2026;28:e73757) doi:10.2196/73757

KEYWORDS

eHealth; telehealth; telemedicine; remote care; mental health services; provider views; virtual medicine; attitude of health personnel; COVID-19; video consultation; survey; suitable types of treatment; medical fields; German outpatient medical care; physician and psychotherapist perspective; provider perspective

XSL•FO
RenderX

Introduction

The utilization of telemedicine, in particular the implementation of video consultation (VC), has become increasingly popular since the onset of the global COVID-19 pandemic in 2020 [1-3]. In Germany, VC could be billed for certain diagnoses or check-ups in outpatient care since 2017. In 2019, the legislative initiative of the Digital Healthcare Act was implemented with the objective of accelerating the digital transformation of the nation's health care infrastructure. Thereafter, VCs have been eligible for billing in almost all medical specialties, as physicians and psychotherapists have been granted permission to use them at their discretion, without the prior restrictions [4].

However, uptake in its use was only reported during the COVID-19 pandemic as it was a relevant option for preventing infections [4-6]. Toward the end of the pandemic, its use has declined again and does still not appear to have become established into everyday medical practice [7]. A representative survey of 2800 medical practices in Germany by Albrecht et al [8] revealed that in 2021, 20% of practices offered VC, compared to 25% during the beginning of the COVID-19 pandemic in 2020. Among psychotherapy practices, VCs are used more frequently, with about three quarters of practices offering the service in 2020 and 2021 [8]. Claims data analysis by the statutory health insurance fund BARMER reported that in 2021 around 80% of the VCs have been provided by psychiatrists and psychotherapists, followed by general practitioners (GPs) who provided 14% of all VCs. Less prevalent are the medical specialties gynecologists (1.8%) and pediatricians (1.5%) [7]. According to the latest figures reported by the Zentralinstitut kassenärztliche Versorgung, the use of VC in Germany did not demonstrate an upward trend from 2022 to 2024, with approximately 2.7 million VCs recorded in both years [9].

For the period before and during the COVID-19 pandemic, there are study results that have identified a number of factors that influence the provision of VC services. Besides medical specialty, these include the age of the service provider, the community size of practice location, and the size of the service provider's practice. A greater uptake in VC use during the COVID-19 pandemic was observed among younger physicians and psychotherapists [6,8] and among providers in an urban practice [6,10]. In terms of practice type, Kane et al [11] and Albrecht et al [8] reported higher use in larger practices, multispecialty practices, or practices with a nonphysician ownership structure. Conversely, Knörr et al [10] found no associations for use but did find associations for perceived benefit with group practices rather than solo practices. At that time, VCs were most frequently employed for the purposes of discussing examination results and taking a patient's medical history. VCs have previously been reported as being inadequate for the purposes of diagnosis and determining medical indication [8].

Moreover, VCs have been most commonly used as an adjunct to in-person treatment. However, in rural regions with a lower density of physicians, VCs have been increasingly provided without additional personal contact in a billing quarter, that is, in 3 months [6]. This indicates that VC can also be a suitable method in the context of a shortage of medical care. Internationally, service providers have reported positive experiences with chronic disease management, mental health support, and medication management during the pandemic [12]. Despite the fact that experiences with VC during the pandemic have frequently been characterized as positive overall [13], health care providers appear to be rather hesitant about implementing them into routine practice [14].

The sudden and unanticipated need for VC adoption during the course of the pandemic was widely regarded as an opportunity for their long-term integration into routine treatment. At the same time, it accelerated ongoing political initiatives in Northern Europe to digitize health care [15-17]. However, in Germany, measures to promote the use of VC do not seem to have been effectively implemented, which has created uncertainty about the sustainability of VC. This underscores the need for research on how VC use evolved. Although there is plenty of research on providers' experiences with VC during the pandemic, given the potential shifts in attitudes toward such services resulting thereof, it is crucial to examine the subject from a provider's perspective and their attitudes toward telemedicine services in a postpandemic context [14]. Hence, the objective of this study is to examine the attitudes of physicians and psychotherapists with regard to the use of VC after the COVID-19 pandemic in German outpatient care. Specifically, the study examines factors associated with VC provision and interest in VC provision. Further, the study investigates which types of treatments and medical fields are deemed suitable by the medical and psychotherapeutic service providers in nonpandemic conditions. This survey of service providers' perspectives can assist in ascertaining how VC can be used sensibly. For many, their familiarity with VC was built in the context of the pandemic, when it was quickly implemented. There is now a need for evaluation to establish an evidence base for its optimal use [16,18].

Methods

Study Design

This survey was conducted as part of the German study "Preference-based use of video consultations in urban and rural regions," which was funded by the Innovation Fund of the German Federal Joint Committee (funding no. 01VSF20011). A study protocol was published previously [19]. Based on groundwork research within the study—a literature review and focus group discussions—a standardized questionnaire was developed.

The questionnaire was divided into three segments: (1) attitudes and experiences regarding VC provision, (2) preference survey using discrete choice experiments, and (3) sociodemographic information as well as information about the medical profession.

First, the survey asked whether VC was offered and, if so, how frequently. A follow-up question enquired how long VC had been part of their medical treatment. The first section also asked about suitable types of treatment and medical fields for which VC could be (at least partially) suitable. There were also

XSL•FO
RenderX

questions about the exclusive use of VC (eg, no additional face-to-face contact in a quarter for the treatment case), the use of VC in treating children and adolescents, and potential future VC provision. All questions had categorical responses except the questions about suitable types of treatment and suitable medical fields for treatment via VC, which were 5-scale Likert questions (highly unsuitable, unsuitable, to some extent, suitable, and very suitable). All Likert scales had an "I don't know" option, which was coded as missing during the analysis. Likert scales also included open-text response options to allow for further explanation of the participants' experiences.

The attitude of inhibiting and promoting factors for the provision of VC and the results of the second part of the survey (preference elicitation) are published separately. Thus, no further details are given here.

The third section included sociodemographic questions and information about the participants' medical profession. These included membership at one of the four participating regional Associations of Statutory Health Insurance Physicians (ASHIPs), community size of practice location, employment status, area of medical care, and medical specialty or psychotherapy practice.

Prior to conducting the survey, a pretest with think-aloud and probing methods was done [20]. The pretest included 6 physicians with different medical specialties.

Ethical Considerations

At the beginning of the survey, all participants received information about the objective of the study and how their data would be dealt with. Participation was anonymous and voluntary, and participants could opt out at anytime. Entries could be made on paper or online using a QR code using QuestionPro online survey software (QuestionPro, Inc). The incentive to participate in the survey was amplified by the prospect of winning a tablet in a raffle. The local ethics committee of the Medical Faculty of the University of Duisburg-Essen approved the study (reference: 21 - 10283-BO). Completion time of the survey was around 20 minutes. An excerpt of the survey can be found in Multimedia Appendix 1.

Recruitment

The 4 participating ASHIPs in the regions Berlin, Westphalia-Lippe, Mecklenburg-Western Pomerania, and Schleswig-Holstein sent the survey via postal delivery to all their members (physicians and psychotherapists) who met the inclusion criterion. Inclusion criterion was being eligible to provide VC. In Germany, outpatient medical care in the field of psychotherapeutic/psychiatric disorders is provided by medical care providers (eg, medical psychotherapists, psychiatrists, neurologists, family physicians) and nonphysician care providers called psychological psychotherapists. As both groups are allowed to provide VC and are reimbursed under the same remuneration system, psychological psychotherapists were also included in the study. The ASHIPs cover rural as well as urban regions.

In order to gain insight into the postpandemic perspective, the survey was conducted between November 2022 and March 2023. No reminder was sent out, as the target of 10% response rate was reached early.

Statistical Analysis

Descriptive data are presented as frequencies (n) and percentages (%). All 49 medical specialty categories were aggregated to 12 final medical specialist groups to simplify the evaluation. For the gender, participants who answered "diverse" were coded as missing in the gender variable for the purposes of subgroup analysis. This decision was taken to safeguard the privacy of those involved because of the relatively limited number of participants with this response. As not all questions were mandatory, missing answers were possible, and no further conclusions can be drawn about the number of "diverse" cases.

Subgroup analyses included gender (male or female), age groups (<40 y, 40 - 50 y, 51 - 60 y, and >60 y), community size in reference to the classification "Stadt-und Gemeindetyp" as rural or urban by the Federal Institute for Research on Building, Urban Affairs, and Spatial Development (rural community, small town, middle town, or large city) [21], VC experience (yes or no), type of practice (individual practice, joint practice, group practice, or medical care unit), ownership of practice (self-employed and employed), and area of medical care (primary care, specialist care, or psychotherapeutic care). Physicians and psychotherapists were categorized as having VC experience if they stated they offered VC at least once a month. In Germany, the distinction between joint practice and group practice is primarily evident in the manner of billing. In joint practice, a joint billing system is employed, whereas in group practice, the physicians share the facilities but operate independently of one another. Medical care units are outpatient facilities in which physicians are employed exclusively and which can be operated by physicians, hospitals, or other licensed providers. With regard to the variable type of practice, multiple responses were permitted in the questionnaire; however, cases with multiple answers were excluded for the subgroup analysis. Regarding areas of medical care, in primary care, all family physicians, including GPs, general internists, and pediatricians, are displayed. Specialist care, on the other hand, comprised physicians who have indicated that they work in specialist care but not in the areas of family medicine and psychiatry or psychotherapy. With regard to psychotherapeutic care, all physicians or nonphysician service providers (psychological psychotherapists) who have reported working in this field were included. All participants working in psychotherapy, psychiatry, or neurology care were referred to in the study as "psychotherapists," unless otherwise stated.

First, a descriptive analysis was conducted, followed by an association analysis as a second step and binary logistic regression as a third step. Correlation and association coefficients were calculated using the Stuart-Kendall Tau-c for ordinal variables and Pearson χ^2 categorical variables. Effect sizes were indicated by Stuart-Kendall Tau-c for ordinal variables and Cramer *V* for categorical variables [22,23]. In accordance with the criteria established by Cohen (*d*), effect sizes that fall below the threshold of 0 are considered inconsequential and thus excluded from the reporting. Effect sizes greater than 0.1 and up to 0.3 are considered weak, while

XSL•FO
RenderX

those greater than 0.3 are considered moderate, and those greater than 0.5 are considered strong [24].

Binary logistic regression was conducted to determine whether physicians' and psychotherapists' individual factors (gender, age, and region), organizational factors (ownership or type of practice), or area of medical care are associated with at least monthly VC provision (yes or no). Similarly, it was then examined whether these factors are related to the general interest in VC provision (yes or no). The aim is to test whether the predictors influence the probability of VC provision or interest in VC and to determine their odds. Multicollinearity diagnostics indicated no concerns for the predictor variables (all tolerance values >0.20; all variance inflation factor values <5). All independent variables were included in the analysis simultaneously in order to examine their contribution to predicting each dependent variable. The level of statistical significance was set at α=.05, with a *P* value of ≤.05 deemed significant. Statistical analyses were performed using the statistical software IBM SPSS Statistics version 28 (IBM Corp).

Results

Response Rate and Participants' Characteristics

The study questionnaire was sent to 34,095 physicians and psychotherapists, of whom 33,994 participants could be contacted. The remaining 101 questionnaires could not be delivered. The total number of returned questionnaires was 6112, resulting in a response rate of 17.9%. About 60.5% (3700/6112) of participants responded online and 39.5% (2412) responded by paper. Following the exclusion of 182 questionnaires that were either implausible or blank, a total of 5930 respondents were included in the analysis. They represent around 17.4% (5930/34,095) of physicians and psychotherapists in the four regions surveyed.

The study sample comprised a slight overrepresentation of women (3140 to 2378; 56.9% to 32.1%) and a slight underrepresentation of medical service providers in the older age groups (>60 y: 1359/5539, 24.5% to 10,261/34,095, 30.1%) compared to the basic population. Table 1 summarizes the sociodemographic characteristics and information on the medical profession.

XSL•FO
RenderX

Table . Characteristics and medical profession of survey participants and basic population.

	Participants, n (%)	Physicians and psychotherapists in selected regions, n (%)[a]
Total	5930 (100)	34,095 (100)
Gender (n=5518)		
Female	3140 (56.9)	17,452 (51.2)
Male	2378 (43.1)	16,643 (48.8)
Age groups (aggregated) (n=5539)		
<40 years	749 (13.5)	3538 (10.4)
40 - 50 years	1478 (26.7)	8317 (24.4)
51 - 60 years	1952 (35.2)	11,979 (35.1)
>60 years	1359 (24.5)	10,261 (30.1)
Community size of practice location (n=5525)		—[b]
Rural community	372 (6.7)	
Small town	965 (17.5)	
Middle town	1411 (25.5)	
Large city	2777 (50.3)	
Employment status[c] (n=5542)		—
Self-employed	4421 (79.8)	
Employed	1161 (20.9)	
Type of practice[c] (n=5542)		—
Individual practice	2864 (50.1),	
Joint practice[d]	1526 (26.7)	
Group practice[d]	807 (14.1)	
Medical care unit[d]	521 (9.1)	
Area of medical care (n=5531)		—
Primary care	1591 (28.8)	
Specialist care	1747 (31.6)	
Psychotherapeutic care	2193 (39.6)	

[a]Data of all 34,095 physicians and psychotherapists from the Associations of Statutory Health Insurance Physicians (ASHIPs) in the regions Berlin, Westphalia-Lippe, Mecklenburg-Western Pomerania, and Schleswig-Holstein. All specialist groups that are theoretically authorized to provide video consultation in Q4 2023 are included. Source: Data provided by the participating ASHIPs.

[b]Not available.

[c]Multiple answers possible.

[d]German terms: joint practice: Berufsausübungsgemeinschaft; group practice: Praxisgemeinschaft; medical care unit: Medizinisches Versorgungszentrum.

About half (2777/5525) of the participants have their practice location in a large city, 25.5% (1411/5525) run their practice in a medium-sized town, 17.5% (965/5525) in a small town, and 6.7% (372/5525) in the countryside. Most of the participants (4549/5542, 82.1%) are self-employed. Around half of the participants (2873/5542, 51.8%) run an individual practice, 27.6% (1530/5542) a joint practice, and 14.6% (811/5542) a group practice. Only 9.5% (529/5542) of the participants state that they are employed in a medical care unit. The participants also were asked to indicate their area of medical care. About 28.8% (1591/5531) of the participants practice in primary care, while a comparable proportion (1747/5531, 31.6%) works in specialist care. A larger proportion of participants (2193/5531, 39.6%) is engaged in psychotherapeutic/psychiatric care.

Most common medical care areas are primary care (1438/5484, 26.1%), nonphysician care providers in psychological psychotherapy (1293/5484, 23.6%), physicians in the field of psychotherapy, psychiatry, or neurology (464/5484, 8.5%), and psychotherapists for children and adolescents (458/5484, 8.4%). Other, less frequent areas of medical care in which the participants are trained in include gynecology, gynecological endocrinology, and reproductive medicine (315/5484, 5.7%), orthopedics and trauma surgery (182/5484, 3.3%), pediatrics

XSL•FO
RenderX

and adolescent medicine in primary care (153/5484, 2.8%), otorhinolaryngology (130/5484, 2.4%), and venereal diseases and dermatology (113/5484, 2.1%). Table 2 presents a summary of the (aggregated) medical specialties displayed in this survey in comparison to the basic population.

Table . Medical specialty (aggregated).

	Participants, n (%)	Physicians and psychotherapists in selected regions, n (%)[a]
Total	5484 (100)	34,104[b] (100)
Primary care (excluding pediatrics in primary care)	1438 (26.1)	10,736 (31.5)
Psychological psychotherapy	1293 (23.6)	5446 (16.0)
Psychotherapy/psychiatry/neurology (excluding psychological psychotherapists and psychotherapists for children/adolescents)	464 (8.5)	2389 (7.0)
Psychotherapists for children and adolescents	458 (8.4)	1590 (4.7)
Gynecology, gynecological endocrinology, and reproductive medicine	315 (5.7)	2394 (7.0)
Orthopedics and trauma surgery	182 (3.3)	1702 (5.0)
Pediatrics (in primary care)	153 (2.8)	1350 (4.0)
Otorhinolaryngology	130 (2.4)	898 (2.6)
Surgery (excluding orthopedics and trauma surgery)	128 (2.3)	863 (2.6)
Venereal diseases and dermatology	113 (2.0)	793 (2.3)
Pediatrics (specialist care)	44 (0.8)	144 (0.4)
Other	766 (13.2)	5799 (17.0)

[a]Data of all 34,095 physicians and psychotherapists from the Associations of Statutory Health Insurance Physicians (ASHIPs) in the regions Berlin, Westphalia-Lippe, Mecklenburg-Western Pomerania, and Schleswig-Holstein. All specialist groups that are theoretically authorized to provide video consultation in Q4 2023 are included. Source: Data provided by the participating ASHIPs.

[b]Since in a few cases more than one specialist group was assigned to a physician or psychotherapist in an ASHIP, the number in this table (34,104) differs slightly from the total number of respondents (34,095).

Provision of Video Consultations

Provision of VC

Among the physicians and psychotherapists who have already used VC, 78.3% (1722/2199) have offered them since the beginning of the COVID-19 pandemic; 11.2% (246/2199) started using VC later on during the pandemic (from 2021 onwards). However, only 10.2% (231/2199) have been providing VC since before the COVID-19 pandemic.

Of the physicians and psychotherapists, 37.8% (2216/5863) reported using VC at least once a month, with 36.6% using VC several times a week, 33.6% at least once a week, and 29.5% at least once a month. Meanwhile, 62.2% of participants (3647/5863) report that they have rarely or never offered VC. Significant associations with gender, age, community size of practice location, type of practice, ownership of practice, and area of medical care were identified with VC provision (see Multimedia Appendix 2).

The findings of the logistic regression analysis demonstrate that at least monthly VC provision is associated with each included predictor except community size of practice location. For the area of medical care, the odds ratio (OR) of at least monthly VC provision in psychotherapeutic care was about 8.2 compared to primary care (95% CI 7.4 - 1.64; *P*<.001), whereas in specialist care, the odds for monthly VC provision were approximately 50% lower than in primary care (OR 0.5, 95% CI 0.43-0.59; *P*<.001). Regarding medical specialty, within psychotherapeutic care, medical professionals who offer VC most frequently are psychological psychotherapists (985/1282, 76.8%), psychotherapists for children and adolescents (333/456, 73%), and psychotherapists, psychiatrists, and neurologists (245/457, 53.6%). At a considerable distance from the frequency observed in psychotherapeutic care, 22.7% (10/44) of pediatric specialists, 20% (285/1422) of physicians in primary care, and 19.6% (22/112) of specialists for venereal diseases and dermatology have stated to have VC experience.

Individual factors such as gender and age were also significant predictors for monthly VC provision. Female participants have higher odds to provide VC at least once a month (OR 1.163, 95% CI 1.01 - 1.34; *P*=.03). The odds for monthly VC provision in older age groups are approximately 60% higher than in the <40 years old age group (OR 0.41, 95% CI 0.32-0.52; *P*<.001). With regard to organizational factors, self-employed participants had about 50% higher odds of providing VC than employed participants (OR 0.480, 95% CI 0.39-0.59; *P*<.001). Regarding the type of practice, participants working in group practice (OR 1.33; 95% CI 1.07 - 1.66; *P*=.01) and medical care units (OR

XSL•FO
RenderX

1.39, 95% CI 1.04 - 1.85; *P*=.03) had significantly higher odds to provide VC at least once a month compared to those in individual practice. The model was statistically significant, χ^2_{13}=1937.52, *P*<.001, and explained approximately 41% according to Nagelkerke pseudo-R^2. Further, the Hosmer–Lemeshow test was nonsignificant, χ^2_8=12.32, *P*=.14, indicating good model fit.

Interest in VC Provision

The majority of the physicians and psychotherapists (4347/5442, 79.9%) express interest in potentially providing VC. There appears to be a significant association between interest in use by gender, age group, VC experience, and area of medical care (see Multimedia Appendix 3).

The patterns evident in current at least monthly VC provision are also reflected in the results of the logistic regression analysis of general interest in VC provision. Interest in providing VC is significantly associated with gender, age group, practice location, type of practice, and area of medical care. For instance, logistic regression indicated a significant association between age group and the likelihood of interest in VC provision (*P*<.001). Compared to the reference group of physicians and psychotherapists under 40 years, those aged 40 - 50 years had 51% lower odds of the outcome (OR 0.49, 95% CI 0.34-0.69; *P*<.001). The 50 - 60 years age group also exhibited lower odds compared to the reference group (OR 0.3, 95% CI 0.21-0.42; *P*<.001). The strongest effect was observed among individuals aged over 60 years, who had approximately 80% lower likelihood of interest in VC provision (OR 0.2, 95% CI 0.15-0.28; *P*<.001). The overall model for interest in VC provision was statistically significant, χ^2_{13}=574.23, *P*<.001, indicating that the set of predictors reliably distinguished interest in use of VC provision. The model explained approximately 17% according to Nagelkerke pseudo-R^2. The Hosmer–Lemeshow goodness-of-fit test was nonsignificant, χ^2_8=13.77, *P*=.09, suggesting acceptable model fit. Table 3 presents details on the results of the logistic regression models.

XSL•FO
RenderX

Table . Binary logistic regression for at least monthly VC[a] provision and interest in VC provision.

Predictor	At least monthly VC provision[b] (n=5232)		Interest in VC provision [c] (n=5173)	
	P value	OR (95% CI)	*P* value	OR (95% CI)
Gender (reference: male)				
Female	.03	1.163 (1.01-1.34)	.03	1.315 (1.13-1.53)
Age group (reference: <40 y)				
40 - 50 y	<.005	0.711 (0.56-0.90)	<.001	0.486 (0.34-0.69)
51 - 60 y	<.008	0.729 (0.58-0.9)	<.001	0.300 (0.21-0.42)
>60 y	<.001	0.409 (0.32-0.52)	<.001	0.204 (0.15-0.28)
Community size of practice location (reference: rural community)				
Small town	.10	0.783 (0.59-1.05)	.79	0.958 (0.7-1.31)
Middle town	.20	0.833 (0.63-1.10)	.96	0.993 (0.73-1.35)
Large city	.38	1.128 (0.864-1.47)	.21	1.210 (0.9-1.63)
Ownership of practice (reference: self-employed)				
Employed	<.001	0.480 (0.39-0.59)	.30	0.893 (0.72-1.11)
Type of practice (reference: individual practice)				
Joint practice	.05	1.175 (0.999-1.38)	<.001	1.560 (1.31-1.85)
Group practice	.01	1.331 (1.067-1.66)	<.001	1.385 (1.08-1.78)
Medical care unit	.03	1.387 (1.039-1.85)	.01	2.713 (1.96-3.76)
Area of medical care (reference: primary care)				
Specialist care	<.001	0.503 (0.425-0.59)	.01	0.807 (0.68-0.96)
Psychotherapeutic care	<.001	8.917 (7.472-10.64)	<.001	3.850 (3.11-4.77)
Constant	0.33	0.845	<.001	5.606

[a]VC: video consultation.

[b]At least monthly VC provision: pseudo-R^2=0.41.

[c]Interest in VC provision: pseudo-R^2=0.17.

Exclusive Use of VC

Almost half (2702/5651, 47.8%) of the physicians and psychotherapists surveyed can imagine treating patients exclusively via VC without further face-to-face appointments for the treatment case in the same quarter. There are differing attitudes toward the exclusive use of VC, depending on whether the participants have VC experience (χ^2_1=320.997, *P*<.001; Cramer *V*=0.239), their age group (χ^2_3=126.589, *P*<.001; Cramer *V*=0.152), or their medical specialty (χ^2_{11}=146.092, *P*<.001; Cramer *V*=0.164). Physicians and psychotherapists with VC experience (1345/2676, 63%) are more likely to imagine treating patients in a treatment case exclusively via VC than participants who have not or rarely provided VC before. Further, medical specialists with specialist training in venereal diseases and dermatology (80/113, 70.8%), GPs (720/1423, 50.6%), nonphysician care providers in psychological psychotherapy (721/1285, 56.1%), and physicians who specialized in psychotherapy, psychiatry, and neurology (232/459, 50.5%) are more likely to consider treating patients in a treatment case exclusively via VC.

Suitable Medical Fields

Regarding suitable medical fields, the participants were asked if parts of their treatment could be carried out via VC for the enquired medical fields. Results only include participants with

XSL•FO
RenderX

previous VC experience, as it is assumed that a more precise assessment can be made. The participants indicated a potential for using VC, in particular, in the area of mental and behavioral disorders. In the field of mental disorders, nonorganic sleep disorders (very suitable or suitable: 1296/1995, 65.0%), affective disorders (very suitable or suitable: 1123/1991, 56.4%), anxiety disorders (very suitable or suitable: 1110/2014, 55.1%), sexual dysfunctions (very suitable or suitable: 952/1836, 51.3%), or obsessive-compulsive disorders (very suitable or suitable: 952/1974, 48.2%) are considered to be very suitable or suitable most often. An exception is cases of schizophrenia (highly unsuitable or unsuitable: 1047/1764, 59.4%) as well as schizotypal and delusional disorders caused by psychotropic substances (highly unsuitable or unsuitable: 824/1837, 44.9%), which are rather reported as highly unsuitable or suitable for VC.

Among chronic illnesses, the treatment of chronic pain (very suitable or suitable: 936/1517, 61.7%), metabolic disorders (very suitable or suitable: 464/1009, 46.0%), and allergies (very suitable or suitable: 413/1020, 40.5%) as well as dermatological diseases (very suitable or suitable: 368/989, 37.2%) are considered very suitable or suitable for VC treatment.

Acute illnesses are rather considered highly unsuitable or unsuitable, with the exception of acute headaches (to some extent suitable: 481/1238, 38.9%) or dermatoses (to some extent suitable: 337/922, 36.6%), which are considered suitable to some extent. Further details on the attitude of suitable medical fields from the perspective of physicians and psychiatrists with VC experience can be found in Multimedia Appendix 4.

Types of Treatment

To assess potential suitable types of treatment, only findings of participants with VC experience are displayed as a better understanding is assumed. Significant associations have been identified by participants who have experience in VC provision for all queried types of treatment. They are more likely to consider the types of treatment as suitable for VC provision than participants with little or no experience; for details on their associations and effect sizes, see Multimedia Appendix 5.

According to over 60% of these physicians and psychotherapists, discussing test results (1422/1896, 75%), issuing of prescriptions for drugs and remedies (793/1204, 65.9%), and issuing of incapacity certificates (677/1042, 65.0%) and treatment planning (1379/2131, 64.7%) are very suitable or suitable types of treatment via VC (see Table 4). Follow-up appointments (771/1350, 57.1%) and taking a patient's medical history (1195/2147, 55.7%) are considered very suitable or suitable by over half of the participants. Diagnostic procedures appear to be the least suitable form of treatment for VC, with only 24.6% of participants (very suitable or suitable: 523/2126), indicating agreement with these types of treatment.

Table . Attitudes of physicians and psychotherapists with VC[a] experience towards the suitability of types of treatment for VC.

	Highly unsuitable, n (%)	Unsuitable, n (%)	To some extent, n (%)	Suitable, n (%)	Very suitable, n (%)
Discussion of test results (n=1896)	29 (1.5)	101 (5.3)	344 (18.1)	706 (37.2)	716 (37.8)
Medical history (n=2147)	85 (4.0)	343 (16.0)	524 (24.4)	633 (29.5)	562 (26.2)
Issuing prescriptions for drugs and remedies (n=1204)	49 (4.1)	136 (11.3)	226 (18.8)	429 (35.6)	364 (30.2)
Issuing incapacity certificate (n=1042)	44 (4.2)	100 (9.6)	221 (21.2)	366 (35.1)	311 (29.8)
Therapy treatment planning (n=2131)	35 (1.6)	172 (8.1)	545 (25.6)	919 (43.1)	460 (21.6)
Individual psychiatric/psychotherapeutic consultations (n=2017)	27 (1.3)	122 (6.0)	546 (27.1)	743 (36.8)	579 (28.7)
Follow-up checks (eg, wound healing, medication) (n=1350)	34 (2.5)	130 (9.6)	415 (30.7)	485 (35.9)	286 (21.2)
Group sessions (eg, in psychotherapy) (n=977)	178 (18.2)	381 (39.0)	256 (26.2)	105 (10.7)	57 (5.8)
(Further) diagnostic work-up (n=2126)	151 (7.1)	669 (31.5)	783 (36.8)	375 (17.6)	148 (7.0)

[a]VC: video consultation.

Regarding area of medical care of participants with VC experience, significant associations have been identified for taking a patient's medical history (χ^2_8=21.266, P<.001; Cramer V=0.229), further diagnostic work-up (χ^2_8=85.333, P<.001; Cramer V=0.145), issuing an incapacity certificate (χ^2_8=36.051, P<.001; Cramer V=0.135), the discussion of test results

XSL•FO
RenderX

(χ^2_8=164.154, *P*<.001; Cramer *V*=0.213), follow-up checks (χ^2_8=71.743, *P*<.001; Cramer *V*=0.167), and individual psychiatric or psychotherapeutic consultations (χ^2_8=127.499, *P*<.001; Cramer *V*=0.182). Taking a patient's medical history via VC is deemed very suitable or suitable rather by providers in primary care (very suitable or suitable: 259/308, 84.1%) and specialist care (very suitable or suitable: 141/198, 71.2%), less than by providers in psychotherapeutic care (very suitable or suitable: 739/1544, 47.9%). It is similar in the case of discussing test results, 91.2% (281/308) of GPs and 90.2% (175/194) of specialists report them to be a very suitable or suitable option via VC, while 68.6% (895/1305) of psychotherapists consider them very suitable or suitable. Conversely, the attitudes of the suitability of (further) diagnostics are the opposite. Although (further) diagnostics are deemed highly unsuitable or suitable, health care professionals engaged in the delivery of psychotherapeutic care (very suitable or suitable: 441/1526, 28.9%) are more likely to view them as suitable than GPs (very suitable or suitable: 37/303, 12.2%) and specialists (very suitable or suitable: 19/201, 9.5%). For follow-up checks, 64.5% (517/802) of psychotherapists and about half of specialists (98/184, 53.3%) consider these very suitable or suitable for VC, whereas only 38.3% (116/303) of GPs agree on their suitability. In particular, in the context of individual psychotherapeutic consultations, psychotherapists (very suitable or suitable: 1079/1569, 68.8%) consider these to be suitable for VC, and also approximately 50% of GPs (very suitable or suitable: 122/241, 50.6%) and specialist care (very suitable or suitable: 55/111, 49.5%) hold this view. For further associations or correlations and effect sizes, see Multimedia Appendix 6.

Discussion

Principal Findings

This study examined the attitude toward the use of VC in German outpatient care after the COVID-19 pandemic from a provider perspective. For the majority of the surveyed physicians and psychotherapists, the advent of the COVID-19 pandemic served as a catalyst, with approximately 80% commencing VC use during the initial phase of the pandemic since 2020, particularly within the area of psychotherapeutic care. VC provision is reported more often in psychotherapeutic care (73%) and still less often in primary (20%) and specialist care (12%). Despite over half of service providers reporting little or no VC provision, with approximately 40% stating otherwise, this represents a notable increase in providers who have used VC so far compared to the prepandemic period. However, the survey results do not reveal whether they have incorporated VC into their daily routine or whether it is merely a sporadic occurrence necessitated by the circumstances during the pandemic. Current claims data analyses indicate that VC provision is not regular, as use has experienced a decline once again following the conclusion of the COVID-19 pandemic [7].

According to our findings, the odds of VC provision are higher among female physicians and psychotherapists, younger age groups, and those living in urban areas. These results are coherent with previously published research on German claims data done earlier in the study by Hüer et al [6] covering the time period during the COVID-19 pandemic. Therefore, there seem to be no changes in the analyzed factors associated with VC provision during and after the COVID-19 pandemic. In terms of organizational factors, practice types with shared facilities demonstrated higher odds for VC provision than participants in an individual practice setting. During the pandemic, Hüer et al [6] and Knörr et al [10] did not observe a higher frequency of use by group practices. However, Knörr et al [10] noted that the perceived benefit is greater in group practices. Recent systematic reviews emphasized the importance of perceived usefulness and ease of use in the adoption of telemedicine [25,26]. The findings of this study suggest that shared facilities may result in a reduction of organizational barriers to VC use and offer enhanced practical benefits or ease of use.

Conversely, self-employed physicians also demonstrated higher odds for VC provision. The current literature suggests that the barriers are particularly related to the expected usefulness within their professional environment, making the indications inconclusive.

Globally, many countries have integrated telehealth into routine care to some extent, although the degree varies widely [27]. The 2022 International Health Policy Survey of General Practitioners, conducted by the Commonwealth Fund in 10 high-income countries, examined opinions on the user-friendliness and effectiveness of telemedicine in the period following the peak of the pandemic. In the majority of countries, physicians reported a number of advantages, including financial compensation, improved timeliness of care, and enhanced ability for mental health care. German physicians demonstrated the lowest levels of satisfaction across the majority of stated metrics, while medical professionals in the United Kingdom, Australia, Canada, and the United States reported higher levels of satisfaction and ease of implementation of the telehealth platform. In Germany, telehealth expanded more slowly postpandemic, with only about 30% - 40% of German primary care doctors stating telehealth to be easy to implement, citing cost of the platform as a barrier [28]. This indicates that besides readiness of users and providers, national policies and technical infrastructure have an influence on VC adoption [16,28]. Nevertheless, the results suggest that there is potential to establish use among medical service providers, as approximately 80% expressed interest in potentially using VC. With younger providers in particular using and being interested in using VC, it is likely that an age-related dip in VC provision will diminish in the future. In addition, these results can also help identify subgroups that might perceive more barriers in VC adoption. Elder and male physicians and psychotherapists showed less odds of VC adoption. Research suggests that those groups show less readiness and might lack familiarity or trust in new technologies [26,29]. For the time being, it is recommended that efforts are made to facilitate use [26], with a particular focus on engaging older age groups, as even around 70% of physicians and psychotherapists with little or no VC experience reported interest in using VC. This objective could be realized through the implementation of knowledge pertaining to the advantages of VC in the obligatory Continuing Medical Education administered by the German Medical Associations.

XSL•FO
RenderX

Although awareness of VC has been heightened by the COVID-19 pandemic and there is considerable interest in its use, VC still appears to be used irregularly at present [7], suggesting that there may currently be inhibiting factors inhibiting the expansion of VC provision (eg, concerns about a deterioration in the physician–patient relationship [13] or factors relating to the health care system [16]). This issue merits a need for further research to investigate the reasons for the current reluctance to use VC. In order to explore why the uptake of VC has not been sustainable and to develop strategies that can be used to ensure its long-term integration, a look at hindering factors, particularly in the German health care system, or qualitative research on sustainable implementation strategies would also be an appropriate approach.

Globally, many countries have integrated telehealth into routine care to some extent, although the degree varies widely [27]. The 2022 International Health Policy Survey of General Practitioners, conducted by the Commonwealth Fund in 10 high-income countries, examined opinions on the user-friendliness and effectiveness of telemedicine in the period following the peak of the pandemic. In the majority of countries, physicians reported a number of advantages, including financial compensation, improved timeliness of care, and enhanced ability for mental health care. German physicians demonstrated the lowest levels of satisfaction across the majority of stated metrics, while medical professionals in the United Kingdom, Australia, Canada, and the United States reported higher levels of satisfaction and ease of implementation of the telehealth platform. In Germany, telehealth expanded more slowly postpandemic, with only about 30% - 40% of German primary care doctors stating telehealth to be easy to implement, citing cost of the platform as a barrier [28]. This indicates that besides readiness of users and providers, national policies and technical infrastructure have an influence on VC adoption [16,28]. Nevertheless, the results suggest that there is potential to establish use among medical service providers, as approximately 80% expressed interest in potentially using VC. With younger providers in particular using and being interested in using VC, it is likely that an age-related dip in VC provision will diminish in the future. In addition, these results can also help identify subgroups that might perceive more barriers in VC adoption. Elder and male physicians and psychotherapists showed less odds of VC adoption. Research suggests that those groups show less readiness and might lack familiarity or trust in new technologies [26,29]. For the time being, it is recommended that efforts are made to facilitate use [26], with a particular focus on engaging older age groups, as even around 70% of physicians and psychotherapists with little or no VC experience reported interest in using VC. This objective could be realized through the implementation of knowledge pertaining to the advantages of VC in the obligatory Continuing Medical Education administered by the German Medical Associations. In light of the heightened awareness of VC brought on by the COVID-19 pandemic and given the great interest in the use of VC, VCs still seem to be used irregularly [7], which suggests that there might currently be inhibiting factors that impede the expansion of VC provision (eg, fear of deterioration in the physician–patient relationship [13] or health system factors [16]). This issue merits a need for further research to investigate the reasons for the current reluctance to use VC. In order to explore why the uptake of VC has not been sustainable and to develop strategies that can be used to ensure its long-term integration, a look at hindering factors, particularly in the German health care system, or qualitative research on sustainable implementation strategies would also be an appropriate approach.

Regarding VC as an add-on service or as a service without face-to-face patient contact in the same quarter for the treatment case, there is no clear agreement or rejection in favor of its exclusive use. According to the findings of the study, a slightly below-average number of respondents were in favor of exclusive use. However, it should be noted that in Germany, discounts in honorarium are applied when VC is used exclusively during a given billing quarter [4]. Hence, the attitude of suitability of exclusive VC use may be biased due to the German billing regulations. Nevertheless, the findings suggest that exclusive use may be more appropriate for certain medical specialties. Physicians in venereal diseases and dermatology, as well as nonphysician care providers in psychological psychotherapy, demonstrated a greater tendency to consider the exclusive use of VC.

This research suggests that VC may be particularly suited to the care of patients with mental and behavioral health problems [30]. However, there are limitations for certain diagnoses, such as schizophrenia, schizotypal, and delusional disorders caused by psychotropic substances. This assessment has been widely acknowledged in relevant literature and is accompanied by an increase in the use of VC and perceived suitability by providers of psychotherapeutic care [31,32]. Furthermore, the results suggest that VCs appear to be suitable for selected chronic illnesses. Current research also validates this for selected indications, for example, for diabetes [33], obstructive sleep apnea [34], or dermatological diseases such as atopic dermatitis [35].

Regarding the findings on types of treatment, there are indications that physicians and psychotherapists consider VC to be particularly suitable to those types of treatment which are mainly based on conversations (eg, discussing test results or taking the patient's medical history, therapy treatment planning) or which require a higher degree of administrative work (eg, issuing of prescriptions for drugs and remedies and issuing of incapacity certificates). This may particularly apply to well-known patients and those with long-term conditions [12,13]. The results of our study, as well as existing research, indicate that diagnostics may not be the optimal fit for this particular modality of treatment in primary and specialist care. For psychotherapeutic care, further diagnostic workup has been rated more positively, which is supported by another German survey [8]. During the COVID-19 pandemic, follow-up checks were reported to be a well-accepted type of treatment [12,36,37]. According to this study, after the COVID-19 pandemic, this was not fully confirmed, with only 33% finding it suitable and 36% suitable to some extent. The variability of the suitability of follow-up checks seems to be dependent on different treatments. In the area of psychotherapeutic care, follow-up visits are more likely to be considered appropriate than in primary care or specialist care. Moreover, differences are

XSL•FO
RenderX

observed with regard to the presence or absence of VC experience. This suggests that as a result of the adoption of VC by physicians and psychotherapists, attitudes have shifted toward a more positive outlook. It can be reasonably concluded that training programs may assist in reducing the current deficit of knowledge and in fostering greater confidence in the provision of VC.

Limitations

The study achieves a satisfactory response rate of around 18% and thus results in a large sample size of 5930, which is typically challenging to obtain when studying physicians or psychotherapists. This was made possible due to the involvement of ASHIPs in contacting all of the physicians in their respective regions. However, it is important to note that the participants do not fully represent the original population. The sample comprises a slightly higher proportion of female participants and a slightly higher proportion of individuals under the age of 40 than is observed in the basic population. The mode of data collection did not influence the variables related to the main findings. However, differences were observed in demographic characteristics such as the age subgroup between participants recruited online and those surveyed via paper-pencil. The option to participate on paper next to online via a QR code was incorporated with the objective of mitigating the exclusion of offline participants, a demographic which is predominantly composed of older individuals. Moreover, it should be acknowledged that there is a considerably greater number of physicians and medical service providers from the psychotherapeutic care area than represented in the surveyed basic population. This could potentially introduce a selection bias, as psychotherapists are more likely to use VC and may therefore have responded differently from the average due to their vested interest in VC. Psychotherapists may perceive VC as being congruent with their professional practice, particularly with regard to continuity of care and emergency situations [1,2,30,38]. Conversely, some medical practitioners, particularly those specializing in physical disciplines, may regard VC as being incongruent with the necessity for physical examinations [39]. Subsequently, the results may lead to more positive attitudes toward VC, which could restrict the generalizability of the findings to other outpatient providers. Subgroup analysis differentiated according to the areas of primary care, and specialist and psychotherapeutic cares was done in order to be able to make a more precise distinction.

While the rollout of VCs has been comprehensively researched in many countries since the start of the pandemic, there has been limited investigation into this for the German outpatient sector up to now. This study, therefore, adds value by taking a closer and more detailed look at the situation in outpatient care in Germany. Yet, physicians and psychiatrists, as well as psychotherapists working exclusively in the private sector, were not included. However, as approximately 90% of insured persons in Germany have statutory health insurance and only members of ASHIPs are permitted to bill patients with statutory health insurance, the relevance of exclusively private practitioners is negligible. Nevertheless, since only German physicians and psychotherapists were surveyed, the transferability of the results to other countries with different regulatory systems may be limited. Further, studies could explore the use of VC in the inpatient or care settings where there may also be potential [40,41]. Additionally, all specific medical specialties in the field of psychiatric/neurological and psychotherapeutic care are referred to as “psychotherapists” in this study. It is important to note that there are differences in practicing depending on specific medical specialty. Further research could go into depth according to specific medical specialties.

While the logistic regression models provided useful insights, several limitations should be noted. First, as this was a cross-sectional study, the associations identified cannot be interpreted causally. The logistic regression model was employed for the purpose of exploratory analysis, with the objective of ascertaining which variables exert a significant effect when controlling for the other variables. Second, although the first model showed substantial explanatory power (Nagelkerke R^2=0.41) and the second model moderate power (Nagelkerke R^2=0.17), a certain proportion of variance remains unexplained, suggesting that additional factors not included in the analysis may influence the outcome. These might include digital maturity, technical infrastructure, and remuneration terms [42-44].

As this was an exploratory cross-sectional study, a self-constructed questionnaire was used. Although no full psychometric validation was conducted, the instrument was informed by prior qualitative analyses and pretesting to ensure content relevance and comprehensibility. This limits the extent to which the results can be generalized, and the approach is appropriate for an exploratory design and provides valuable initial insights to guide future research. Furthermore, systematic errors cannot be completely excluded. In this study, the use of batteries of items (eg, Likert scales) may have led to a tendency for the questions to be averaged or not to be differentiated. Concerning measures of association or correlation, the calculated effect sizes are according to Cohen value categorized as weak and a few as moderate.

Conclusion

Although the actual use of VC in general practice and specialist care is still relatively low, most physicians and psychotherapists are in favor of their use after the COVID-19 pandemic. A wide range of medical indications is considered to be suitable, particularly in the area of psychotherapeutic care and use for chronic illnesses. The findings indicate that German service providers do not appear to be averse to the use of VC as they show high interest in their provision. Potential barriers to the exclusive use of VC may include factors such as remuneration or the medical specialty of the practitioner. Furthermore, in the case of certain medical specialties, treatment, particularly diagnosis, cannot be provided without a physical examination. It is advisable to incorporate the providers’ perspectives into the ongoing refinement of VC policies and practices, as these insights are important for ensuring the successful implementation of VC as a therapeutic modality. This research on physicians’ and psychotherapists’ attitudes adds to a baseline knowledge of different attitudes by providers and can provide grounds to create a shared vision, which has shown to be a driver

XSL•FO
RenderX

in the adoption process of digital technology [45]. Besides, the attitude and conduct of medical professionals play a fundamental role in ensuring that patients experience a sufficient level of comfort during the course of their treatment [46]. Given that physicians and psychotherapists are more likely to endorse video-based treatment modalities when they have prior experience with them, expanding training programs may help reduce knowledge gaps, alleviate uncertainty, and foster greater confidence in their use in outpatient care. The expansion of the digital infrastructure in the future, which facilitates the dissemination of information, has the potential to diminish potential barriers. Nevertheless, the results do not provide insights into the hurdles that exist in daily use and how exactly their use can be promoted. Further research is necessary to ascertain the full implications.

Funding

This study was entirely funded by the Innovation Fund of the German Federal Joint Committee (number 01VSF20011). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

Primary data of the survey are not publicly available due to data protection reasons.

Authors' Contributions

All authors contributed to the study organization and conception of the surveys. LK was responsible for the survey analysis and was the major contributor in writing this manuscript. The analysis and the manuscript draft were critically revised by TH. All authors contributed to the manuscript in different stages and also read and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Excerpt of the survey.
[DOCX File, 63 KB - jmir_v28i1e73757_app1.docx]

Multimedia Appendix 2
Subgroup analysis for provision of video consultations.
[DOCX File, 17 KB - jmir_v28i1e73757_app2.docx]

Multimedia Appendix 3
Subgroup analysis for interest in video consultation provision.
[DOCX File, 16 KB - jmir_v28i1e73757_app3.docx]

Multimedia Appendix 4
Attitudes on suitable medical fields (only for participants with video consultation experience).
[DOCX File, 22 KB - jmir_v28i1e73757_app4.docx]

Multimedia Appendix 5
Association and effect size of suitable types of treatment for participants with video consultation experience.
[DOCX File, 15 KB - jmir_v28i1e73757_app5.docx]

Multimedia Appendix 6
Association/correlation and effect size of suitable types of treatment (only for participants with video consultation experience).
[DOCX File, 19 KB - jmir_v28i1e73757_app6.docx]

References

1. Zangani C, Ostinelli EG, Smith KA, et al. Impact of the COVID-19 pandemic on the global delivery of mental health services and telemental health: systematic review. JMIR Ment Health 2022 Aug 22;9(8):e38600. [doi: 10.2196/38600] [Medline: 35994310]
2. Mann DM, Chen J, Chunara R, Testa PA, Nov O. COVID-19 transforms health care through telemedicine: evidence from the field. J Am Med Inform Assoc 2020 Jul 1;27(7):1132-1135. [doi: 10.1093/jamia/ocaa072] [Medline: 32324855]

XSL•FO
RenderX

3. Stein DJ, Naslund JA, Bantjes J. COVID-19 and the global acceleration of digital psychiatry. Lancet Psychiatry 2022 Jan;9(1):8-9. [doi: 10.1016/S2215-0366(21)00474-0] [Medline: 34921798]
4. Bericht des Bewertungsausschusses und des ergänzenden Bewertungsausschusses zur telemedizinischen Leistungserbringung im Einheitlichen Bewertungsmaßstab [Report in German]. : Deutscher Bundestag; 2022 URL: https://dserver.bundestag.de/btd/20/049/2004982.pdf [accessed 2025-11-28]
5. Smith AC, Thomas E, Snoswell CL, et al. Telehealth for global emergencies: implications for coronavirus disease 2019 (COVID-19). J Telemed Telecare 2020 Jun;26(5):309-313. [doi: 10.1177/1357633X20916567] [Medline: 32196391]
6. Hüer T, Walendzik A, Kleinschmidt L, et al. Use of video consultation between 2017 and 2020 in outpatient medical care in Germany and characteristics of their user groups: analysis of claims data. JMIR Form Res 2025 Mar 14;9:e60170. [doi: 10.2196/60170] [Medline: 40085136]
7. Mangiapane S. Fünf jahre videosprechstunde in der vertragsärztlichen versorgung: hat die pandemie zum durchbruch geführt [Book in German]. In: Repschläger U, Schulte C, Osterkamp N, editors. Gesundheitswesen Aktuell 2023 Beiträge Und Analysen: Barmer Institut fur Gesundheitssystemforschung; 2023:82-99 URL: https://www.bifg.de/media/dl/gesundheitswesen-aktuell/2023/gwa-2023_mangiapane.pdf [accessed 2025-11-28]
8. Albrecht M, Otten M, Bernhard J. PraxisBarometer digitalisierung 2022: befragung von vertragsärztinnen/-ärzten und vertrags [Report in German]. : iGES; 2022 URL: https://www.iges.com/sites/igesgroup/iges.de/myzms/content/e6/e1621/e10211/e27603/e27639/e27640/e27642/attr_objs27737/IGES_PraxisBaroDigit2021_Kurzbericht_ger.pdf [accessed 2025-11-28]
9. Mangiapane S, Kretschmann J, Czihal T, Stillfired DV. Zi-trendreport zur vertragsärztlichen versorgung: bundesweiter tabellarischer report vom 1 [Report in German]. : Zentralinstitut kassenarztliche Versorgung; 2025 URL: https://www.zi.de/fileadmin/Downloads/Service/Publikationen/Zi-Trendreport_2023-Q2.pdf [accessed 2024-07-22]
10. Knörr V, Dini L, Gunkel S, et al. Use of telemedicine in the outpatient sector during the COVID-19 pandemic: a cross-sectional survey of German physicians. BMC Prim Care 2022 Apr 23;23(1):92. [doi: 10.1186/s12875-022-01699-7] [Medline: 35461212]
11. Kane CK, Gillis K. The use of telemedicine by physicians: still the exception rather than the rule. Health Aff 2018 Dec;37(12):1923-1930. [doi: 10.1377/hlthaff.2018.05077] [Medline: 30633670]
12. Ward K, Vagholkar S, Sakur F, Khatri NN, Lau AYS. Visit types in primary care with telehealth use during the COVID-19 pandemic: systematic review. JMIR Med Inform 2022 Nov 28;10(11):e40469. [doi: 10.2196/40469] [Medline: 36265039]
13. Wanderås MR, Abildsnes E, Thygesen E, Martinez SG. Video consultation in general practice: a scoping review on use, experiences, and clinical decisions. BMC Health Serv Res 2023 Mar 30;23(1):316. [doi: 10.1186/s12913-023-09309-7] [Medline: 36997997]
14. Saeed S, Singhal M, Kaur KN, et al. Acceptability and satisfaction of patients and providers with telemedicine during the COVID-19 pandemic: a systematic review. Cureus 2024 Mar;16(3):e56308. [doi: 10.7759/cureus.56308] [Medline: 38628988]
15. Hoff T, Lee DR. Physician satisfaction with telehealth: a systematic review and agenda for future research. Qual Manag Health Care 2022;31(3):160-169. [doi: 10.1097/QMH.0000000000000359] [Medline: 35132008]
16. Assing Hvidt E, Atherton H, Keuper J, et al. Low adoption of video consultations in post-COVID-19 general practice in Northern Europe: barriers to use and potential action points. J Med Internet Res 2023 May 22;25:e47173. [doi: 10.2196/47173] [Medline: 37213196]
17. Tyre MJ, Orlikowski WJ. Windows of opportunity: temporal patterns of technological adaptation in organizations. Organ Sci 1994 Feb;5(1):98-118. [doi: 10.1287/orsc.5.1.98]
18. Fahy N, Williams GA, Habicht T, et al. Use of Digital Health Tools in Europe: Before, During and After COVID-19: European Observatory on Health Systems and Policies; 2021. [Medline: 35099866]
19. Kleinschmidt L, Walendzik A, Wasem J, et al. Preference-based implementation of video consultations in urban and rural regions in outpatient care in Germany: protocol for a mixed methods study. JMIR Res Protoc 2024 Apr 11;13:e50932. [doi: 10.2196/50932] [Medline: 38602749]
20. Häder M. Empirische Sozialforschung [Book in German]: Springer Fachmedien Wiesbaden; 2019.
21. Milbert A, editor. Raumabgrenzungen und Raumtypen des BBSR [Book in German]: Bonn; 2012:111.
22. Stuart A. The estimation and comparison of strengths of association in contingency tables. Biometrika 1953 Jun;40(1/2):105. [doi: 10.2307/2333101]
23. Cramér H. Mathematical Methods of Statistics (PMS-9): Princeton University Press; 2016:1593.
24. Cohen J. In: Hillsdale NJ, editor. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Erlbaum; 1988:567. [doi: 10.4324/9780203771587]
25. Garavand A, Aslani N, Nadri H, Abedini S, Dehghan S. Acceptance of telemedicine technology among physicians: a systematic review. Inform Med Unlocked 2022;30:100943. [doi: 10.1016/j.imu.2022.100943]
26. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023 Sep 18;6(1):161. [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
27. Ndwabe H, Basu A, Mohammed J. Post pandemic analysis on comprehensive utilization of telehealth and telemedicine. Clinical eHealth 2024 Dec;7:5-14. [doi: 10.1016/j.ceh.2023.12.002]

XSL•FO
RenderX

28. Gunja M, Gumas ED, Williams R. How primary care physicians experience telehealth: an international comparison—findings from the 2022 international health policy survey of primary care physicians. : The Commonwealth Fund; 2023 URL: https://tinyurl.com/3pxu6chz [accessed 2025-11-27]
29. Weik L, Fehring L, Mortsiefer A, Meister S. Understanding inherent influencing factors to digital health adoption in general practices through a mixed-methods analysis. NPJ Digit Med 2024 Feb 27;7(1):47. [doi: 10.1038/s41746-024-01049-0] [Medline: 38413767]
30. Appleton R, Williams J, Vera San Juan N, et al. Implementation, adoption, and perceptions of telemental health during the COVID-19 pandemic: systematic review. J Med Internet Res 2021 Dec 9;23(12):e31746. [doi: 10.2196/31746] [Medline: 34709179]
31. Shaker AA, Simonsen E, Tarp K, et al. Capturing patients' and clinicians' experiences of using video consultations in mental health outpatient services: qualitative thematic analysis. JMIR Form Res 2024 Aug 21;8:e50580. [doi: 10.2196/50580] [Medline: 39167796]
32. Shaker AA, Austin SF, Storebø OJ, et al. Psychiatric treatment conducted via telemedicine versus in-person modality in posttraumatic stress disorder, mood disorders, and anxiety disorders: systematic review and meta-analysis. JMIR Ment Health 2023 Jul 5;10:e44790. [doi: 10.2196/44790] [Medline: 37277113]
33. Sotomayor F, Hernandez R, Malek R, Parimi N, Spanakis EK. The effect of telemedicine in glycemic control in adult patients with diabetes during the COVID-19 era - a systematic review. J Clin Med 2023 Aug 31;12(17):5673. [doi: 10.3390/jcm12175673] [Medline: 37685740]
34. Alsaif SS, Kelly JL, Little S, et al. Virtual consultations for patients with obstructive sleep apnoea: a systematic review and meta-analysis. Eur Respir Rev 2022 Dec 31;31(166):220180. [doi: 10.1183/16000617.0180-2022] [Medline: 36517048]
35. Verma L, Turk T, Dennett L, Dytoc M. Teledermatology in atopic dermatitis: a systematic review. J Cutan Med Surg 2024;28(2):153-157. [doi: 10.1177/12034754231223694] [Medline: 38205736]
36. Richter JG, Chehab G, Reiter J, et al. Evaluation of the use of video consultation in German rheumatology care before and during the COVID-19 pandemic. Front Med (Lausanne) 2022;9:1052055. [doi: 10.3389/fmed.2022.1052055] [Medline: 36507506]
37. Johnsen TM, Norberg BL, Kristiansen E, et al. Suitability of video consultations during the COVID-19 pandemic lockdown: cross-sectional survey among Norwegian general practitioners. J Med Internet Res 2021 Feb 8;23(2):e26433. [doi: 10.2196/26433] [Medline: 33465037]
38. Galvin E, Desselle S, Gavin B, et al. Stakeholder perspectives and experiences of the implementation of remote mental health consultations during the COVID-19 pandemic: a qualitative study. BMC Health Serv Res 2023 Jun 13;23(1):623. [doi: 10.1186/s12913-023-09529-x] [Medline: 37312119]
39. Kulkarni AJ, Thiagarajan AB, Skolarus TA, Krein SL, Ellimoottil C. Attitudes and barriers toward video visits in surgical care: insights from a nationwide survey among surgeons. Surgery 2024 Jul;176(1):115-123. [doi: 10.1016/j.surg.2024.03.033] [Medline: 38734503]
40. Asiri A, AlBishi S, AlMadani W, ElMetwally A, Househ M. The use of telemedicine in surgical care: a systematic review. Acta Inform Med 2018 Oct;26(3):201-206. [doi: 10.5455/aim.2018.26.201-206] [Medline: 30515013]
41. May S, Jonas K, Fehler GV, Zahn T, Heinze M, Muehlensiepen F. Challenges in current nursing home care in rural Germany and how they can be reduced by telehealth - an exploratory qualitative pre-post study. BMC Health Serv Res 2021 Sep 6;21(1):925. [doi: 10.1186/s12913-021-06950-y] [Medline: 34488746]
42. Kerr G, Greenfield G, Li E, et al. Factors associated with the availability of virtual consultations in primary care across 20 countries: cross-sectional study. J Med Internet Res 2025 Mar 19;27:e65147. [doi: 10.2196/65147] [Medline: 40105882]
43. Mold F, Cooke D, Ip A, Roy P, Denton S, Armes J. COVID-19 and beyond: virtual consultations in primary care-reflecting on the evidence base for implementation and ensuring reach: commentary article. BMJ Health Care Inform 2021 Jan;28(1):e100256. [doi: 10.1136/bmjhci-2020-100256] [Medline: 33436372]
44. Donaghy E, Atherton H, Hammersley V, et al. Acceptability, benefits, and challenges of video consulting: a qualitative study in primary care. Br J Gen Pract 2019 Sep;69(686):e586-e594. [doi: 10.3399/bjgp19X704141] [Medline: 31160368]
45. Kateb S, Ruehle RC, Kroon DP, van Burg E, Huber M. Innovating under pressure: adopting digital technologies in social care organizations during the COVID-19 crisis. Technovation 2022 Jul;115:102536. [doi: 10.1016/j.technovation.2022.102536]
46. Orrange S, Patel A, Mack WJ, Cassetta J. Patient satisfaction and trust in telemedicine during the COVID-19 pandemic: retrospective observational study. JMIR Hum Factors 2021 Apr 22;8(2):e28589. [doi: 10.2196/28589] [Medline: 33822736]

Abbreviations

ASHIP: Association of Statutory Health Insurance Physicians
GP: general practitioner
OR: odds ratio
VC: video consultation

XSL•FO
RenderX

Edited by J Sarvestan; submitted 12.Mar.2025; peer-reviewed by C Doarn, E Galvin; revised version received 15.Nov.2025; accepted 17.Nov.2025; published 06.Jan.2026.

Please cite as:
Kleinschmidt L, Wasem J, Blase N, Nauendorf B, Malsch J, Brittner M, Brandenburg P, Aeustergerling A, Hüer T
Attitudes Toward Video Consultations From the Perspective of Physicians and Psychotherapists in German Outpatient Care After the COVID-19 Pandemic: Survey Study
J Med Internet Res 2026;28:e73757
URL: https://www.jmir.org/2026/1/e73757
doi:10.2196/73757

© Lara Kleinschmidt, Juergen Wasem, Nikola Blase, Beatrice Nauendorf, Juliane Malsch, Matthias Brittner, Paul Brandenburg, André Aeustergerling, Theresa Hüer. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Communication Challenges and Mitigation Strategies in Primary Care Virtual Consultations: Qualitative Study

Ahmed Alboksmaty[1], MBBS, MPH; Tetiana Lunova[1], PhD; Ara Darzi[1], OM, KBE, PC; Ana-Luisa Neves[1,2], PhD

[1]Department of Surgery and Cancer, Institute of Global Health Innovation, Imperial College London, St Mary's Campus, Norfolk Place, London, United Kingdom

[2]Department of Primary Care and Public Health, Imperial College London, London, United Kingdom

Corresponding Author:
Ahmed Alboksmaty, MBBS, MPH
Department of Surgery and Cancer, Institute of Global Health Innovation, Imperial College London, St Mary's Campus, Norfolk Place, London, United Kingdom

Abstract

Background: The growing reliance on virtual consultations in primary care has reshaped traditional general practitioner (GP)–patient communication dynamics, presenting new challenges that affect care quality and safety.

Objective: This study explores communication challenges and gaps, particularly relevant to virtual consultations compared with face-to-face interactions, as well as identifying mitigation strategies from both GPs' and patients' perspectives.

Methods: This qualitative study employed 4 online focus group discussions with a purposive sample of UK-based GPs and patients. Data were analyzed using a deductive-inductive thematic approach with NVivo software. The extended Shannon-Weaver communication model and the Capability, Opportunity, Motivation and Behavior model guided the analysis of communication challenges and mitigation strategies, respectively. The Consolidated Criteria for Reporting Qualitative Research were followed to ensure rigorous reporting.

Results: A total of 21 participants (12 patients and 9 GPs) took part in 4 online focus group discussions, 2 for patients and 2 for GPs. Six key themes on communication challenges emerged: 5 aligned with the extended Shannon-Weaver communication model (related to the sender-encoder, message, channel, receiver-decoder-feedback, and context), and a new one was inductively identified (patient autonomy and inclusivity). GPs, as senders, highlighted missing visual cues, affecting message clarity in remote communication channels. Patients, as receivers, reported difficulties explaining symptoms remotely, reduced emotional connection, and perceived empathy, linked to contextual challenges and the need for inclusive communication. Mitigation strategies were mapped to the Capability, Opportunity, Motivation and Behavior model: capability (training/resources), opportunity (triage/tools), and motivation (patient engagement/system adaptability), with participants emphasizing tailored training, standardized approaches, and flexible models to support effective and inclusive virtual communication.

Conclusions: This study highlights communication gaps in virtual consultations and proposes actionable mitigation strategies. Tailored use of virtual modalities, supported by structured training and policy efforts, is essential to ensure effective and safe remote communication.

(J Med Internet Res 2026;28:e79399) doi:10.2196/79399

KEYWORDS

communication; remote care; patient-physician relationship; virtual care; telehealth

Introduction

Primary care has become increasingly complex and demanding, driven by the variety of patients' conditions, the multifaceted nature of health care systems, the introduction of new health technologies, and the changing needs and preferences of patients and communities [1,2]. These factors have proposed a shift toward virtual consultations, emerging a transformation in primary care delivery [3]. This trend gained considerable momentum during the COVID-19 pandemic, as telephone and video consultations became essential for maintaining health care access amid disruptions to traditional face-to-face services [3,4]. While these virtual modalities have enhanced the accessibility and sustainability of services, they also necessitate a much-needed reevaluation of the core components of communication in health care to uphold quality and safety standards [3,4].

Effective communication is fundamental to high-quality primary care, acting as the foundation for building trust, fostering mutual understanding, and delivering patient-centered care [5]. The Shannon-Weaver communication model (SWCM), a foundational framework in communication theory, offers a

XSL•FO
RenderX

valuable perspective for critically analyzing GP-patient interactions across different consultation modalities (virtual and face-to-face) [6]. The extended SWCM comprises 9 components, ie, sender, encoder, message, channel, noise, decoder, receiver, feedback, and context, which collectively outline the classic communication process and can be adapted to various contexts [6].

Traditionally, the primary channel for communication between general practitioners (GPs) and patients has been face-to-face consultations. While this method facilitates direct engagement, it is not without challenges, such as time constraints, language and cultural barriers, and environmental distractions, factors that both GPs and patients have learned to navigate and challenge [7]. However, the shift to virtual consultation modalities disrupts these established dynamics, introducing new communication barriers that require tailored mitigation strategies to ensure effective and equitable care delivery [3].

Virtual consultations, whether conducted via phone or video calls, present an additional range of challenges, including technical difficulties, communication barriers, privacy and confidentiality concerns, limitations in clinical assessment, and adaptation challenges for both GPs and patients [3,8,9]. While issues with mutual understanding during in-person consultations might arise from the use of jargon or complex medical terminology, virtual consultations introduce the added difficulty of missing nonverbal cues and expressions, which can exacerbate communication gaps [10]. These distinctive challenges in remote interactions can impact the quality and safety of primary care, indicating the importance of assessing these issues in practice from the perspectives of both GPs and patients [3].

Extensive research has explored communication challenges, gaps, required skills, and improvement strategies in traditional face-to-face medical consultations within primary care [5,11,12]. However, there is limited evidence on the unique challenges and implications introduced by virtual consultation modalities. This study aims to address this gap by exploring the communication barriers and gaps specific to virtual consultations compared to traditional in-person consultations in primary care. It further seeks to identify potential strategies for mitigating these challenges based on the COM-B (capability, opportunity, motivation, and behavior) model [13], which offers a framework for identifying and presenting strategies for improvement. The assessment incorporates perspectives from both GPs and patients, who are the primary participants in the consultation process.

Methods

Study Design

This study adopted a qualitative approach, employing focus group discussions. The study adhered to the Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist to ensure comprehensive reporting of the research processes and outcomes [14].

Study Participants and Recruitment

Participants included patients and GPs with experience of being involved in primary care consultations, both face-to-face and remote interactions within the United Kingdom. Adult patients aged 18 years or older who could communicate in English and had lived experience of virtual consultations (video or audio) with GPs were targeted to facilitate meaningful engagement and interaction during the discussions. Patients were recruited through Valuing Our Intellectual Capital and Experience, an online platform for community engagement and involvement in research [15].

GPs were recruited through a combination of convenience and purposive sampling via the research team's networks, followed by snowball sampling to ensure diversity among participants [16]. GP participants were required to have professional experience conducting virtual consultations (video or audio) in primary care within the UK National Health Service to ensure shared familiarity with the primary care context, including relevant policies and regulations.

All participants received, via email, a detailed participant information leaflet describing the study background, proposed methodology, and objectives. Participants were encouraged to ask questions for clarification before providing informed consent and participating in the focus groups.

Data Collection

Four online focus groups were conducted via Microsoft Teams between June and August 2024: 2 with GPs and 2 with patients. Each group participated independently to capture distinct perspectives. The focus groups were conducted as part of a larger study aimed at identifying the safety implications of virtual consultations in primary care. A researcher (TL) with a clinical background, PhD qualification, and expertise in patient safety and digital health in primary care moderated all 4 focus groups. A senior researcher (ALN) comoderated the first GP focus group to ensure alignment of the discussions with the study's objectives.

A semistructured topic guide was developed to facilitate interactive discussions during the focus groups, covering key aspects of communication challenges and potential solutions. The guide was piloted within the research team to ensure clarity and a logical flow. The questions in the topic guide aimed at exploring communications issues and strategies are presented in Textbox 1.

XSL•FO
RenderX

Textbox 1. Questions relevant to discussing communication issues in the topic guide.

- Communication problems in virtual consultations (this section aimed to explore the communication issues experienced by participants [GPs and patients] during virtual consultations).
  - What communication issues do you think can arise in virtual consultations?
  - Have you experienced any communication problems when having a virtual consultation?
  - How do you think these issues can be solved?

Data Analysis

Microsoft Teams' transcription feature was used to generate automatic transcripts of the discussions, which were later reviewed against the audio recordings by the focus group moderator to ensure accuracy before finalization and analysis. An interpretivist approach was adopted for this qualitative research [17], considering that participants construct meaning through their experiences and regular involvement in consultation dynamics. The analytical focus was therefore on understanding how participants interpreted and made sense of communication within virtual consultations. Transcripts were not returned to participants for further review; instead, the transcripts were directly analyzed by the research team to identify recurring themes and insights into participants' experiences and perspectives.

A deductive-inductive approach was employed for data analysis to ensure a comprehensive assessment while allowing for new themes to emerge [18]. The deductive analysis was guided by the components of the extended SWCM model [6], which were set as proposed themes for communication challenges and gaps. The extended SWCM originally included 9 components, which were reviewed and refined by the research team into 7 key components (Table 1), representing the major deductive themes for this analysis. For the analysis of suggested mitigation strategies, the COM-B model was adopted as the framework for deductive themes, as illustrated in Figure 1.

Figure 1. Redefined COM-B model's components, adapted to be used for the deductive analysis of the mitigation strategies to improve communication during virtual consultations. COM-B: Capability, Opportunity, Motivation and Behavior; GP: general practitioner.

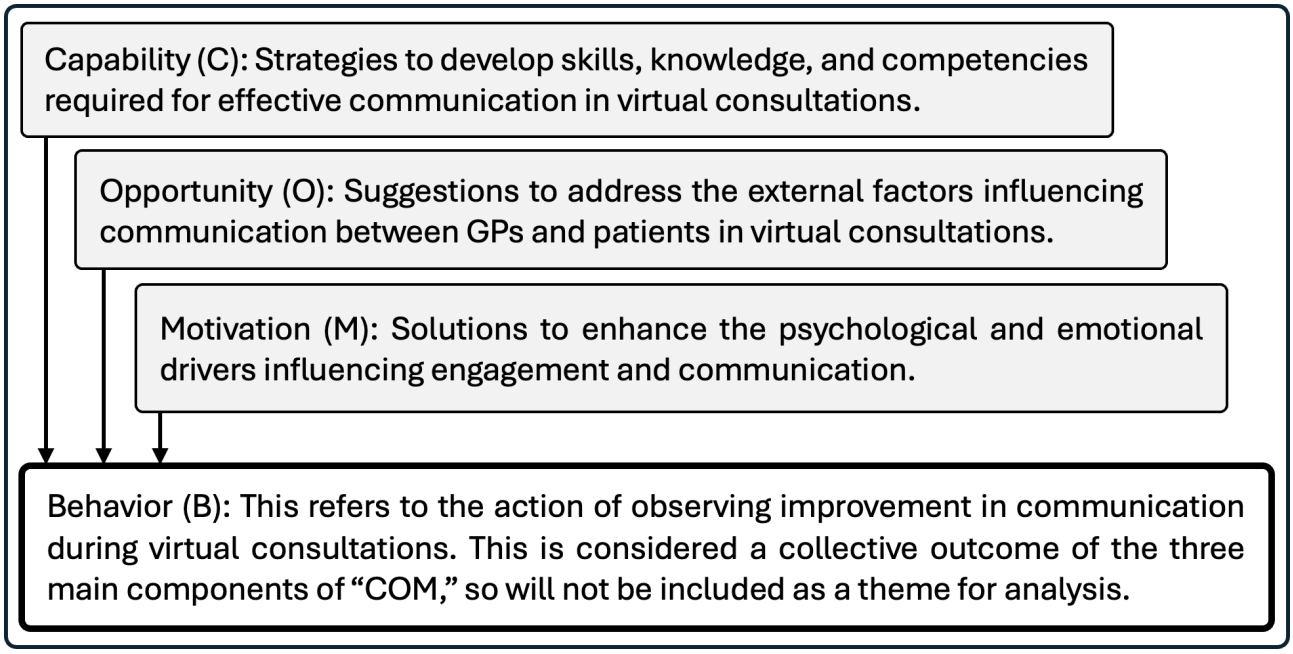

XSL•FO
RenderX

Table . Adapting the SWCM's[a] components for the deductive analysis of communication challenges and gaps.

SWCM components	Adapted definitions
Sender-Encoder	• Merged definition: Combines the roles of the sender (GP[b]) and the encoder, as they are closely intertwined. The sender is the person who initiates the communication, and the encoder refers to the process of transforming thoughts into verbal and nonverbal expressions and talks. • Contextualization to our research: How GPs articulate and communicate their questions, advice, and thoughts with their patients during virtual consultations.
Message	• Definition: The content of the communication, including clinical advice, questions, and emotional cues. • Contextualization to our research: How the clinical discussion is constructed, its clarity, and its comprehensiveness through different modalities.
Channel	• Definition: The medium through which the communication is delivered, virtual (eg, video call, telephone) or face-to-face. • Contextualization to our research: The limitations and merits of each modality in conveying messages effectively, particularly regarding technological or physical limitations.
Noise	• Definition: Any factors that disrupt or distort the communication process. • Contextualization to our research: This would include, for example, poor connectivity, background distractions, emotional and cognitive barriers.
Decoder-Receiver	• Merged definition: Combines the roles of the decoder (interpreting the message) and the receiver (the person/patient for whom the message is intended). • Contextualization to our research: How effectively patients understand the information, particularly the challenges posed by reduced nonverbal cues in phone consultations.
Feedback	• Definition: The responses or reactions from the receiver (patient) that indicate whether the message was understood as intended. • Contextualization to our research: How feedback mechanisms differ between virtual and face-to-face settings.
Context	• Definition: The situational factors influencing the communication process, such as the patient-clinician relationship, and the differences in medical conditions. • Contextualization to our research: How different contexts impact the communication dynamics, including the adaptability of both patients and clinicians to virtual or face-to-face settings.

[a]SWCM: Shannon-Weaver communication model.

[b]GP: general practitioner.

The initial analysis was conducted by a researcher (AA) with expertise in qualitative data analysis and the study topic. Analyses of the patients' and GPs' focus group datasets were first conducted independently, followed by triangulation across both sources to support a comprehensive and coherent interpretation of the findings [19,20]. This process combined a deductive approach, using the predetermined models, with an inductive approach to identify any newly emerging themes. NVivo software was used to facilitate the analysis [21].

Data adequacy was supported by the concept of information power [22], given the focused discussions, the relevance of participants recruited through purposive sampling, and the strong alignment between the data and the study objectives. During analysis, the research team observed thematic stability across the focus groups, indicating that additional data collection was unlikely to generate new insights and suggesting a level of data saturation appropriate for the study aims [23]. Preliminary themes were further refined through iterative analysis and discussions among all members of the research team to reach a consensus on the final themes.

Ethical Considerations

Ethics approval for this study was obtained from the Imperial College Research Ethics Committee (ICREC 6833567) before commencing the recruitment and study processes. All

XSL•FO
RenderX

participants provided informed consent prior to taking part in the focus group discussions. This consent was obtained after participants had reviewed comprehensive participant information leaflets detailing the study's aims and procedures, and after they had the opportunity to ask questions and receive clarification on any aspect of the study. To protect privacy and confidentiality, data were pseudoanonymized; identifiable details were removed after verification of transcription accuracy, and only deidentified qualitative data were used in the analysis. Participants received a £30 (US $40.36) voucher as a token of appreciation for their time and contribution.

Results

Participant Characteristics

A total of 21 participants (12 patients and 9 GPs) were involved in 4 focus group discussions; 2 patient groups, each with 6 participants, and 2 GP groups with 5 and 6 participants, respectively. None of the invited participants refused to take part. Each online focus group lasted for an average duration of 90 minutes. All participating GPs (n=9) were based in England, along with the majority of patients (10/12). The participating GPs had a mean of 5.33 (SD 4.06) years of National Health Service experience, with a range spanning from 2 to 16 years. Table 2 provides an overview of the participants' characteristics.

XSL•FO
RenderX

Table . Self-reported participant characteristics.

Characteristic	Patients	GPs[a]
Total number of participants, n (%)	12 (57.14)	9 (42.86)
Age bands, n (%)		
18- - 29 years (n=1, 4.76%)	1 (4.76)	0 (0)
30- - 39 years (n=10, 47.62%)	3 (14.29)	7 (33.33)
40- - 49 years (n=4, 19.05%)	2 (9.52)	2 (9.52)
50- - 59 years (n=2, 9.52%)	2 (9.52)	0 (0)
≥60 years (n=4, 19.05%)	4 (19.05)	0 (0)
Missing (n=0, 0%)	0 (0)	0 (0)
Sex, n (%)		
Female (n=13, 61.90%)	8 (38.10)	5 (23.81)
Male (n=8, 38.10%)	4 (19.05)	4 (19.05)
Missing (n=0, 0%)	0 (0)	0 (0)
Geographic location, n (%)		
Bournemouth (n=1, 4.76%)	1 (4.76)	0 (0)
Bristol (n=1, 4.76%)	1 (4.76)	0 (0)
Cambridge (n=2, 9.52%)	0 (0)	2 (9.52)
Edinburgh (n=1, 4.76%)	1 (4.76)	0 (0)
Leicester (n=1, 4.76%)	0 (0)	1 (4.76)
Lisburn (n=1, 4.76%)	1 (4.76)	0 (0)
London (n=9, 42.86%)	3 (14.29)	6 (28.57)
Newcastle Upon Tyne (n=4, 19.05%)	4 (19.05)	0 (0)
Walsingham (n=1, 4.76%)	1 (4.76)	0 (0)
Missing (n=0, 0%)	0 (0)	0 (0)
Ethnic group, n (%)		
Arab (n=2, 9.52%)	1 (4.76)	1 (4.76)
Asian or Asian British (n=6, 28.57%)	2 (9.52)	4 (19.05)
Black or Black British (n=2, 9.52%)	2 (9.52)	0 (0)
Mixed (n=1, 4.76%)	1 (4.76)	0 (0)
Other (n=1, 4.76%)	0 (0)	1 (4.76)
White (n=9, 42.87%)	6 (28.57)	3 (14.29)
Missing (n=0, 0%)	0 (0)	0 (0)

[a]GP: general practitioner.

Communication Challenges and Gaps

Overview

Suggestions from GPs and patients were overall aligned, reflecting shared struggles in communication during virtual consultations and a common desire for improvement. Six themes were identified through the data analysis of both patient and GP focus groups, comprising 5 deductive themes derived from merging some of the SWCM components and 1 inductive theme [6]. The 5 deductive themes included communication challenges related to (1) sender-coder, (2) message, (3) channel, (4) receiver-decoder-feedback, and (5) context with modality appropriateness. A newly identified inductive theme was labeled patient autonomy and inclusivity. A description of each theme and its subthemes is provided in Table 3.

XSL•FO
RenderX

Table . Major themes, description, and subthemes identified through the data analysis.

Major theme	Description	Subthemes
Theme 1: Sender–Encoder	How GPs[a] articulate and communicate their questions, advice, and thoughts with their patients during virtual consultations.	• Issues in continuity and coordination of care • Misunderstanding, bias, and missing clinically important points • Lack of empathy and emotional engagement
Theme 2: Message	How the clinical discussion is constructed, its clarity, and its comprehensiveness through different modalities.	• Less clarity and comprehensiveness in virtual consultations • Limited diagnostic visualization during phone consultations
Theme 3: Channel	The limitations and merits of each consultation modality in conveying messages effectively, particularly regarding technological or physical limitations.	• Technical and organizational barriers • Physical barriers in virtual communication • Accessibility and reliability issues for stable connection in virtual consultations
Theme 4: Receiver–Decoder–Feedback	How patients interpret and understand clinical information and discussions and how their responses and interactions indicate understanding and engagement.	• Lack of trust and confidence in virtual consultations • Communication issues that affect understanding, including language barriers • Resistance by patients to engage in virtual consultations • Unclear understanding and missing engagement cues • Missing natural interactions and non-verbal cues in phone consultations
Theme 5: Context and modality appropriateness	The situational factors and suitability of communication methods that influence the effectiveness and engagement of interactions between GPs and patients.	• System and structural limitations • Influence of cultural and social norms • Questionable patient expectations and trust • No standardized system to decide on the consultation type
Theme 6: Patient autonomy and inclusivity	How patient autonomy and inclusivity influence communication between GPs and patients, focusing on the extent to which patients can control their consultation choices and the barriers to full participation in virtual care.	• Barriers to choosing consultation modality • Technical challenges faced by some patient groups in engaging with virtual consultations • Patient preferences for in-person consultations • Inequality in access to technology

[a]GP: general practitioner.

Theme 1: Sender–Encoder (GPs)

Virtual consultations occasionally disrupted clinicians' and patients' ability to interpret emotional and clinical cues, leading to a mutual sense of disconnection that affects relational quality. GPs expressed concerns about missing nonverbal and visual clinical signs during phone consultations, negatively affecting diagnostic decision-making. They also highlighted the extra efforts needed for maintaining engaging conversations with patients remotely, a difficulty further exacerbated by increasing workload pressures. Patients, in turn, reported feeling a lack of empathy and emotional connection during virtual consultations, highlighting a perceived disconnect between themselves and their GPs in virtual settings, particularly in phone consultations.

> *Missing non-verbal communication, the stuff we wouldn't get in a virtual [phone] consultation, you can't see them [patients] walk into the room, can't see how they're sitting, or notice the little things that might make you think differently.* [GP 3.3]

> *Many of the emotions and empathy, at least towards my case, are almost completely gone when it comes to phone and video consultations.* [Patient 2.1]

Theme 2: Message (Content of the Medical Consultation)

The lack of direct interaction in virtual consultations complicated symptom description and clinical explanation, making it harder for both patients and GPs to achieve clarity, especially when addressing complex or sensitive issues. Patients raised significant concerns about the difficulty of clearly describing symptoms and medical complaints over the phone. Likewise, GPs reported challenges in effectively explaining diagnostic reasoning and management plans to patients and ensuring their understanding during virtual consultations, which is of more concern during phone calls. These issues became particularly challenging when discussing sensitive or complex medical conditions.

XSL•FO
RenderX

> *…Because the doctor can't pick up on non-verbal cues that might explain a complaint or tell if the patient wants to say something else.* [Patient 3.2]

Theme 3: Channel

Participants reflected that the risk of technical unreliability in remote communication tools essentially reduces the perceived quality of virtual consultations. Patients expressed increased concerns about the timing of their appointments, often disrupted by delays in calls, the reliability of their internet or phone connections, and the challenge of effectively describing symptoms over the phone, particularly for non-native speakers. Similarly, GPs reported concerns about dropped calls or lost connections during consultations, as well as issues with the quality of sound and images shared online, which could hinder effective communication.

> *The number of times I've called patients—too many—they've had a really bad signal, no alternative number to call, or they're out in a supermarket or somewhere else, making me just can't understand a word they're saying.* [GP 4.3]

Theme 4: Receiver–Decoder–Feedback

Persistent challenges in trust, confidence, and equitable access to care influenced patient engagement with virtual consultations, with the absence of nonverbal interactions over the phone often weakening the mutual understanding essential for safe and reassuring care. Despite the normalization of virtual consultations for GP appointments in the United Kingdom, particularly since the COVID-19 pandemic, patients continued to express a lack of confidence and trust in this method for addressing serious medical conditions. This has been perceived by GPs as a form of resistance, especially among elderly patients, toward accepting virtual consultations as the only method of communication, even when deemed clinically safe.

> *There's a lot of reluctance towards virtual consultations across the country… some people say there's no place for virtual consultations because they've been brought up with the face-to-face discussion model.* [GP 1.4]

Patients further expressed concerns about equity in accessing and benefiting from remote care, particularly among individuals with special needs, such as those with hearing impairments, mental health issues, in need of an interpreter, or learning difficulties. Both GPs and patients reported ongoing challenges with interactive feedback during virtual consultations, largely due to the reliance on verbal cues alone, with potentially facial reactions in video consultations, to convey messages and management plans on most occasions.

Theme 5: Context and Modality Appropriateness

Both GPs and patients emphasized that effective communication depends on aligning consultation modality with clinical needs and contextual expectations. GPs highlighted the need for a robust triage system and effective coordination policies to ensure patients receive the most appropriate and convenient consultation modality tailored to their individual needs. Patients echoed this sentiment, emphasizing that communication during medical appointments extends beyond the consultation itself and is influenced by health care system structure, social norms, and community culture. Bridging these perspectives, both groups stressed that aligning the consultation modality with patient expectations and needs is crucial for building trust and ensuring effective communication.

> *I needed to be present with the doctor. I needed to be there, face-to-face, when talking about sensitive and difficult things. I think sometimes GPs need to acknowledge that a quick call or video call isn't the right way to talk to a patient at certain times.* [Patient 1.1]

Theme 6: Patient Autonomy and Inclusivity

Patients shared a wish to have their voices heard beyond their clinical symptoms, particularly in deciding the type of consultation they receive with their GP. They further explained that some patients struggle to engage, express their emotions, or maintain focus during remote conversations, making in-person appointments more suitable for their personal traits, regardless of their symptoms. Conversely, patients with demanding jobs or caregiving responsibilities may prefer communicating remotely for their convenience. GPs acknowledged this perspective, emphasizing that person-centered and equitable care requires understanding each patient's preferences and capabilities while also assessing their medical needs.

> *I would like a bit more patient choice over which consultations you may have. There may be some conditions where you're quite happy to talk about them over the phone, while with others, you know you'd be better off having an in-person or video consultation.* [Patient 3.2]

Mitigation Strategies to Improve Communication

The proposed mitigation strategies targeted the identified challenges and gaps, which were categorized into 3 deductive themes (capability, opportunity, and motivation) based on the COM-B model [13]. Table 4 outlines these major themes and the associated subthemes based on the data analysis.

XSL•FO
RenderX

Table . Theme of the mitigation strategies based on the capability, opportunity, motivation, and behavior model.

Major theme	Subthemes
Theme 1: Capability	• Training and skill development for virtual consultations • Resource provision and standardization • Enhancing patient understanding
Theme 2: Opportunity	• Triage and risk mitigation • Enhancing communication through supplementary materials • Addressing barriers to effective communication
Theme 3: Motivation	• Patient and doctor preferences in consultation modalities • Supporting patient preparedness and engagement • Monitoring and adapting models of care

Theme 1: Capability

Training and skill development for virtual communication was highlighted by both patients and GPs as essential for improving interactions during virtual consultations. It was noted that strong communication skills in face-to-face settings do not necessarily translate to effective communication over the phone, and vice versa. Therefore, it was suggested that medical students and doctors in training should gain practical experience in conducting virtual consultations and leading virtual clinical appointments. Additionally, there was a shared call for equipping GPs with modern tools and reliable infrastructure to enable seamless and effective communication with patients remotely.

> *I think we need to train our GP trainees using a different model of communication.* [GP 1.4]
>
> *…training [of GPs] on using specific platforms and building rapport over the phone to aid patient-doctor communication, I think, is important, how you do that over the phone to achieve the maximal effect.* [Patient 3.2]

Theme 2: Opportunity

The primary recommendation in this theme was to set a standardized triage system and clear protocols to determine the most suitable consultation modality based on each patient's individual needs, preferences, and characteristics. Offering video consultations as a standard option, in addition to phone appointments, was also suggested as a way to enhance engagement and build stronger professional doctor-patient relationships. Another proposal involved maintaining a registry of patients who face barriers to virtual consultations, such as those living in areas with poor signal, experiencing language barriers, suffering from hearing impairments, or having complex medical conditions.

> *In our practices, there are two completely separate approaches. One practice has a policy that if patients want virtual, they get virtual, and if they want face-to-face, they get face-to-face. In the other practice, the receptionist decides.* [GP 3.4]

Theme 3: Motivation

A GP's comfort and familiarity with different consultation modalities were identified as important factors influencing communication quality, whether virtually or face-to-face. Some participants proposed a system that allocates patients to GPs based on their preferred consultation modality, for instance, assigning virtual consultations to GPs who favor them, while those who prefer direct communication handle face-to-face appointments. However, concerns were raised about the potential impact of this approach on continuity of care.

> *One approach could be to ask patients whether they would prefer a virtual consultation with a new doctor or one they are familiar with. Personally, I would feel more comfortable talking to a doctor I know in person.* [GP 2.4]

Patients also suggested offering incentives to encourage engagement in virtual consultations, such as implementing a fast-track appointment booking system for those opting for virtual appointments. However, some participants noted potential inequality issues with such incentives, as certain patients may be unable to choose virtual consultations due to personal or medical needs. Additionally, a system to promptly update patients about any changes to their appointment times was recommended to help them prepare accordingly and feel more confident in the system.

Discussion

Summary of Results

This study adopted a qualitative approach using focus groups with GPs and patients to identify communication challenges in primary care virtual consultations compared to traditional face-to-face appointments, as well as exploring mitigation strategies. Four online focus groups were conducted, 2 with GPs (n=9) and 2 with patients (n=12).

Six major themes emerged regarding the challenges and gaps in virtual consultations, 5 of which align with the SWCM model: sender-encoder, message, channel, receiver-decoder-feedback, and context with modality appropriateness. An additional theme, patient autonomy and inclusivity, was identified inductively. For GPs as sender-encoders, misinterpretations and missed clinically significant cues were major concerns. The "message" was often compromised by limited clarity in dialog and difficulties in communicating diagnostic information virtually. Technical issues, poor connectivity, and accessibility barriers complicated the remote "channels" of communication. Patients, as receiver-decoders, reported a lack of trust in discussing

XSL•FO
RenderX

complex medical needs remotely and perceived reduced empathy and emotional engagement from GPs in virtual consultations. Social norms and cultural factors, especially among elderly patients, shaped preferences and communication styles, highlighting the importance of "context and modality appropriateness." Furthermore, patient autonomy and inclusivity in choosing virtual consultations influenced their engagement and willingness to communicate effectively during virtual primary care appointments.

Proposed mitigation strategies were mapped to the COM-B model: capability (training and resource provision), opportunity (triage systems and supplementary communication tools), and motivation (patient engagement and system adaptability). Participants highlighted the need for tailored capacity building through virtual-communication-focused training and skills development, supported by resource provision and standardized approaches. Recommendations to enhance the opportunity for better communication in virtual consultations included implementing triage and risk mitigation strategies, using supplementary materials to support remote interactions, and continuously identifying and addressing communication barriers. To strengthen motivation among both GPs and patients, participants emphasized the importance of respecting individual preferences for consultation modality, offering preparatory guidance and materials to support patient engagement, and ensuring flexibility and adaptability in care models.

Findings in Context of Existing Literature

Consistent with our findings, previous studies have identified common challenges in virtual consultations. Patients often struggle to articulate complex symptoms over the phone, while GPs face difficulties ensuring patient understanding of medical information and management plans [24]. These challenges are particularly pronounced in sensitive discussions, cases involving complex health needs, conditions requiring physical examination [25], and with new patients unfamiliar to the GP [26,27]. Some conditions, such as mental health concerns, raise questions about the appropriateness of virtual consultations on some occasions [28]. Our participants described a perceived lack of empathy in virtual appointments, contributing to dissatisfaction in mental health care. However, previous studies reported contrasting evidence, suggesting that offering timely virtual appointments may enhance mental health care by enhancing accessibility to medical advice whenever needed [3,28,29]. This inconsistency in evidence for some medical conditions highlights the importance of contextualizing and adapting care and communication models, including the choice of consultation modality, to align with individual patient needs and clinical circumstances [11]. It is also important to note that patient preferences for communicating remotely versus face-to-face are influenced not only by their health needs but also by personal characteristics [30]. Evidence shows that younger individuals, those with busy schedules, and highly educated patients tend to favor virtual appointments for convenience and accessibility [3,31].

Existing literature has tended to identify similar communication challenges in primary care consultations, but these are often presented in a fragmented manner [26,29,32]. Our study synthesizes these challenges within a structured, theoretically informed framework [6], with particular emphasis on virtual consultations as an emerging norm in primary care [3]. By framing communication issues as a shared responsibility between patients and GPs [33], we offer a more nuanced interpretation that supports practical approaches to mitigation. Through the application of the adapted SWCM framework [6], we further highlight how patient, clinician, technological, and contextual factors intersect, enabling us to propose targeted interventions that may lead to meaningful and sustained improvements in communication practices.

Importantly, communication challenges in virtual consultations extend beyond the consultation itself, involving both pre- and post-encounter interactions [32,34,35]. A qualitative study conducted in Australia highlighted pre-consultation patient-related factors that influence GP-patient communication, particularly health literacy and familiarity with digital platforms [26]. Many patients struggle with completing pre-consultation forms required for describing their symptoms and concerns for booking virtual appointments, especially when selecting appropriate terms to describe their symptoms [32]. From a GP's perspective, familiarity with the patient, as well as access to and time for reviewing medical records, plays a crucial role in shaping communication style and ensuring a high-quality virtual consultation [27,36]. Existing tools also provide structured checklists and guidance to support high-quality virtual consultations, with communication highlighted as a central component [37]. For example, the Telehealth Etiquette Competency Checklist (TECC) emphasizes communication alongside technological, environmental, and confidentiality considerations [37].

Effective communication remains essential even after the consultation itself, particularly in follow-up messaging, referrals, and sharing components of the agreed-upon management plan [34]. The widespread adoption of virtual consultations has raised concerns regarding continuity of care, as it may challenge the foundation of building a strong professional relationship between GPs and their patients, which is an essential factor in enhancing the consultation experience and overall quality of care [38]. Ensuring clarity in follow-up plans after virtual consultations is crucial, particularly in guiding patients on how to seek further care if their concerns persist or if the initial virtual visit does not fully address their health needs [8,35]. Clear and efficient remote communication can help reduce unnecessary follow-up visits, whether in primary care or to hospital and ambulatory care [3].

To mitigate the identified communication barriers and enhance interactions during virtual consultations, GPs often use verbal affirmations, such as "yes" and "I see," to prove active listening [33]. They may also verbally narrate any concurrent tasks, such as reviewing investigation results or previous reports, to maintain transparency with patients for reassurance. Such capacity-building requirements for these supporting skills have been proposed by the participants in our study to improve communication.

Evidence suggests that, when feasible, video consultations may be preferable for initial appointments to establish rapport, for

XSL•FO
RenderX

assessments requiring observation of physical signs [25], and for counseling appointments. Existing literature also offers evidence-based guidelines, emphasizing aspects regarding the knowledge, skills, attitudes, and teaching strategies required for high-quality videoconferencing, which primary care teams can adapt to their local context [39]. However, no clear advantage or superiority of video consultations over face-to-face interactions has been established [40]. The collective evidence from broader literature indicates that both GPs and patients generally prefer communication through face-to-face consultations, followed by video calls when feasible, with phone consultations being the least preferred [41]. However, this hierarchy may vary depending on the context; for instance, when the consultation is for routine medication renewal, or patients live far away from their GP surgeries, phone or video consultations could be more practical [42].

Strengths and Limitations

This study focuses on the core aspect of virtual consultation -communication- as a critical determinant of both satisfaction and patient preference in clinical practice [5]. Focus groups were employed due to their ability to foster an in-depth dialog among participants and to identify shared concerns effectively [43,44]. However, focus groups could potentially allow vocal participants to dominate the discussion and the possibility that some individuals may share less in-depth reflections in a group setting [45]. To mitigate these risks, the moderator had substantial experience in conducting similar research, ensuring balanced participation and providing sufficient time for individuals to share their views [45]. Additionally, in terms of methodology, future research using direct observation or other objective methods could reveal communication nuances that focus group discussions may miss [46]. This would enable a clearer assessment of communication aspects in different consultation modalities.

The analysis was strengthened by the use of 2 theoretical frameworks to structure the data analysis and guide its interpretation: the SWCM framework [6], which addresses the challenges and gaps, and the COM-B model [13], which informs the development of mitigation strategies. Our analysis integrated insights from both GP and patient transcripts, emphasizing that effective communication is inherently reciprocal and requires active engagement from both parties.

While we aimed to recruit a diverse patient population, further research is needed to explore communication challenges among particular groups with potential distinct needs, such as patients requiring interpreters or those with mental health conditions and complex chronic illnesses like cancer. Additionally, all participating GPs were based in England, with an average of 5.4 years of experience. Future studies should include GPs at varying career stages and across different practice settings to identify role-specific communication challenges and potential training needs. Moreover, our study categorized both video and telephone consultations under the broad term of virtual consultations; however, each modality presents unique communication dynamics that warrant further investigation.

Implications for Research, Policy, and Practice

Virtual consultations have become an integral part of the “norm” in primary care, whether in the UK or globally. Therefore, it is essential to optimize remote communication to uphold patient safety and care quality [3]. The communication challenges and mitigation strategies identified in this study, structured using theoretical frameworks, provide actionable themes to be considered by policymakers and practitioners for improvement plans. For example, targeted communication training for GPs, ensuring that GP practices are equipped with stable and reliable technology for virtual consultations, and implementing guidelines to allocate virtual appointments to patients who are both medically and personally suited for them are all key areas for intervention.

Further research should focus specifically on GP-patient communication during virtual consultations, examining its impact on patient satisfaction and clinical outcomes, including safety. Policy efforts should extend beyond the provision of technology and infrastructure to include structured training programs that equip GPs and other clinicians with the skills necessary for effective remote communication. The training efforts should target not only qualified GPs but also undergraduate medical students and other healthcare professionals who are involved in patient care provision, potentially through virtual appointments [12]. In practice, the allocation of virtual consultations should consider both the GP’s level of experience in remote communication and the patient’s health condition and background. Communication in healthcare should be viewed holistically, not merely as spoken language, but as the means through which care is delivered, tailored to the needs of each patient.

Conclusion

This study highlights the pivotal role of communication in virtual consultations, highlighting existing gaps and potential strategies for improvement in primary care. Through focus groups with GPs and patients, we identified key communication challenges and potential mitigation strategies. While virtual consultations offer convenience, their use should be tailored to individual patient needs and clinical contexts. Future research should explore the distinct communication dynamics of different virtual modalities in primary care, informing policymakers and practitioners with avenues for improvement to ensure equitable, high-quality, and patient-centered primary care delivery.

Funding

This study was supported by the National Institute for Health and Care Research (NIHR) North-West London Patient Safety Research Collaboration (NIHR NWL PSRC, Ref. NIHR204292), with infrastructure support from the NIHR Imperial Biomedical Research Centre. ALN is also supported by the NIHR Applied Research Collaboration North-West London. The views expressed

XSL•FO
RenderX

in this publication are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. The funders had no role in study design, data collection and analysis, decision to publish, or writing the manuscript.

Authors' Contributions

AD, ALN, and TL contributed to the conceptualization of the study. TL (she/her), a postdoctorate research associate, supported by ALN, a senior clinical lecturer in digital health, led the conduct and transcription of the interviews. AA and ALN led the conceptualization of the data analysis approach and the selection of the theoretical frameworks guiding the analysis. AA conducted the initial thematic data analysis. TL reviewed and discussed the initial analysis outcomes with AA, while ALN reviewed and confirmed the final agreed analysis outcomes. All authors contributed to and confirmed the interpretation of the results. AA drafted the initial full manuscript. AD, ALN, and TL reviewed and provided feedback on the manuscript. All authors read and approved the final version of the manuscript. The corresponding author confirms that all listed authors meet the authorship criteria and that no individuals meeting the criteria have been omitted.

Conflicts of Interest

None declared.

References

1. Mount JK, Massanari RM, Teachman J. Patient care complexity as perceived by primary care physicians. Fam Syst Health 2015 Jun;33(2):137-145. [doi: 10.1037/fsh0000122] [Medline: 25893538]
2. Neve G, Fyfe M, Hayhoe B, Kumar S. Digital health in primary care: risks and recommendations. Br J Gen Pract 2020 Dec;70(701):609-610. [doi: 10.3399/bjgp20X713837] [Medline: 33243917]
3. Campbell K, Greenfield G, Li E, et al. The impact of virtual consultations on the quality of primary care: systematic review. J Med Internet Res 2023 Aug 30;25:e48920. [doi: 10.2196/48920] [Medline: 37647117]
4. Ezeamii VC, Okobi OE, Wambai-Sani H, et al. Revolutionizing healthcare: how telemedicine is improving patient outcomes and expanding access to care. Cureus 2024 Jul;16(7):e63881. [doi: 10.7759/cureus.63881] [Medline: 39099901]
5. King A, Hoppe RB. “Best practice” for patient-centered communication: a narrative review. J Grad Med Educ 2013 Sep;5(3):385-393. [doi: 10.4300/JGME-D-13-00072.1] [Medline: 24404300]
6. Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948;27(3):379-423. [doi: 10.1002/j.1538-7305.1948.tb01338.x]
7. Payne R, Dakin F, MacIver E, et al. Challenges to quality in contemporary, hybrid general practice a multi-site longitudinal case study. Br J Gen Pract 2024 Jan;75(750):e1-e11. [doi: 10.3399/BJGP.2024.0184] [Medline: 39117426]
8. Turner A, Morris R, Rakhra D, et al. Unintended consequences of online consultations: a qualitative study in UK primary care. Br J Gen Pract 2022 Feb;72(715):e128-e137. [doi: 10.3399/BJGP.2021.0426] [Medline: 34903520]
9. Williams S, Barnard A, Collis P, et al. Remote consultations in primary care across low-, middle- and high-income countries: implications for policy and care delivery. J Health Serv Res Policy 2023 Jul;28(3):181-189. [doi: 10.1177/13558196221140318] [Medline: 36484225]
10. Murphy M, Scott LJ, Salisbury C, et al. Implementation of remote consulting in UK primary care following the COVID-19 pandemic: a mixed-methods longitudinal study. Br J Gen Pract 2021;71(704):e166-e177. [doi: 10.3399/BJGP.2020.0948] [Medline: 33558332]
11. Moulaei K, Sheikhtaheri A, Fatehi F, Shanbehzadeh M, Bahaadinbeigy K. Patients’ perspectives and preferences toward telemedicine versus in-person visits: a mixed-methods study on 1226 patients. BMC Med Inform Decis Mak 2023 Nov 15;23(1):261. [doi: 10.1186/s12911-023-02348-4] [Medline: 37968639]
12. Thuraisingham C, Abd Razak SS, Nadarajah VD, Mamat NH. Communication skills in primary care settings: aligning student and patient voices. Educ Prim Care 2023 May;34(3):123-130. [doi: 10.1080/14739879.2023.2210097] [Medline: 37194600]
13. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
14. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
15. Newcastle University. Voice. URL: https://voice-global.org [accessed 2026-01-10]
16. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015 Sep;42(5):533-544. [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
17. Schwandt TA. Constructivist, interpretivist approaches to human inquiry. In: Denzin NK, Lincoln YS, editors. Handbook of Qualitative Research: Sage; 1994:118-137 URL: https://www.scirp.org/reference/referencespapers?referenceid=435054 [accessed 2026-01-10]

XSL•FO
RenderX

18. Bonner C, Tuckerman J, Kaufman J, et al. Comparing inductive and deductive analysis techniques to understand health service implementation problems: a case study of childhood vaccination barriers. Implement Sci Commun 2021 Sep 15;2(1):100. [doi: 10.1186/s43058-021-00202-0] [Medline: 34526141]
19. Ramani S, Mann K. Introducing medical educators to qualitative study design: twelve tips from inception to completion. Med Teach 2016 May;38(5):456-463. [doi: 10.3109/0142159X.2015.1035244] [Medline: 25897710]
20. Carter N, Bryant-Lukosius D, DiCenso A, Blythe J, Neville AJ. The use of triangulation in qualitative research. Oncol Nurs Forum 2014 Sep;41(5):545-547. [doi: 10.1188/14.ONF.545-547] [Medline: 25158659]
21. NVivo 15 - the most trusted qualitative analysis software (QDA) is even better. Lumivero. 2020. URL: https://lumivero.com/products/nvivo [accessed 2026-01-10]
22. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
23. Saunders B, Sim J, Kingstone T, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907. [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
24. Khanji MY, Gallagher AM, Rehill N, Archbold RA. Remote consultations: review of guiding themes for equitable and effective delivery. Curr Probl Cardiol 2023 Aug;48(8):101736. [doi: 10.1016/j.cpcardiol.2023.101736] [Medline: 37075908]
25. Honarbakhsh S, Sporton S, Monkhouse C, Lowe M, Earley MJ, Hunter RJ. Remote clinics and investigations in arrhythmia services: what have we learnt during Coronavirus disease 2019? Arrhythm Electrophysiol Rev 2021 Jul;10(2):120-124. [doi: 10.15420/aer.2020.37] [Medline: 34401185]
26. Nguyen AD, White SJ, Tse T, et al. Communication during telemedicine consultations in general practice: perspectives from general practitioners and their patients. BMC Prim Care 2024 Sep 4;25(1):324. [doi: 10.1186/s12875-024-02576-1] [Medline: 39232645]
27. Schers H, van den Hoogen H, Bor H, Grol R, van den Bosch W. Familiarity with a GP and patients' evaluations of care. A cross-sectional study. Fam Pract 2005 Feb;22(1):15-19. [doi: 10.1093/fampra/cmh721] [Medline: 15640289]
28. Bulkes NZ, Davis K, Kay B, Riemann BC. Comparing efficacy of telehealth to in-person mental health care in intensive-treatment-seeking adults. J Psychiatr Res 2022 Jan;145:347-352. [doi: 10.1016/j.jpsychires.2021.11.003] [Medline: 34799124]
29. Antonio S, Joseph D, Parsons J, Atherton H. Experiences of remote consultation in UK primary care for patients with mental health conditions: a systematic review. Digit Health 2024;10:20552076241233969. [doi: 10.1177/20552076241233969] [Medline: 38465292]
30. Lunova T, Hurndall KH, Crespo R, et al. Impact of the cost-of-living crisis on patient preferences towards virtual consultations. J Telemed Telecare 2025 Sep;31(8):1175-1185. [doi: 10.1177/1357633X241255411] [Medline: 38767152]
31. Crook RL, Iftikhar H, Moore S, Lowdon P, Modarres P, Message S. A comparison of in-person versus telephone consultations for outpatient hospital care. Future Healthc J 2022 Jul;9(2):154-160. [doi: 10.7861/fhj.2022-0006] [Medline: 35928204]
32. Atherton H, Eccles A, Poltawski L, Dale J, Campbell J, Abel G. Investigating patient use and experience of online appointment booking in primary care: mixed methods study. J Med Internet Res 2024 Jul 8;26:e51931. [doi: 10.2196/51931] [Medline: 38976870]
33. White SJ, Nguyen AD, Roger P, et al. Tailoring communication practices to support effective delivery of telehealth in general practice. BMC Prim Care 2024 Jun 27;25(1):232. [doi: 10.1186/s12875-024-02441-1] [Medline: 38937674]
34. Lyles CR, Gupta R, Tieu L, Fernandez A. After-visit summaries in primary care: mixed methods results from a literature review and stakeholder interviews. Fam Pract 2019 Mar 20;36(2):206-213. [doi: 10.1093/fampra/cmy045] [Medline: 29846584]
35. Reed M, Huang J, Graetz I, Muelly E, Millman A, Lee C. Treatment and follow-up care associated with patient-scheduled primary care telemedicine and in-person visits in a large integrated health system. JAMA Netw Open 2021 Nov 1;4(11):e2132793. [doi: 10.1001/jamanetworkopen.2021.32793] [Medline: 34783828]
36. Jeffers H, Baker M. Continuity of care: still important in modern-day general practice. Br J Gen Pract 2016 Aug;66(649):396-397. [doi: 10.3399/bjgp16X686185] [Medline: 27481958]
37. Pittmann R, Danaher-Garcia N, Adair White BA, Thompson A. Development and validation of the Telehealth Etiquette Competency Checklist: a Delphi study. J Telemed Telecare 2025 Oct;31(9):1308-1316. [doi: 10.1177/1357633X241279494] [Medline: 39311041]
38. de Visser RO, Nwamba C, Brearley E, Shafiei V, Hart L. Remote consultations in primary care: patient experiences and suggestions for improvement. J Health Psychol 2024 Oct;29(12):1321-1335. [doi: 10.1177/13591053241240383] [Medline: 38581309]
39. Koppel PD, De Gagne JC, Webb M, et al. Guidelines for rapport-building in telehealth videoconferencing: interprofessional e-Delphi study. JMIR Med Educ 2025 Aug 7;11:e76260. [doi: 10.2196/76260] [Medline: 40773683]
40. Thiyagarajan A, Grant C, Griffiths F, Atherton H. Exploring patients' and clinicians' experiences of video consultations in primary care: a systematic scoping review. BJGP Open 2020;4(1):bjgpopen20X101020. [doi: 10.3399/bjgpopen20X101020] [Medline: 32184212]
41. Carrillo de Albornoz S, Sia KL, Harris A. The effectiveness of teleconsultations in primary care: systematic review. Fam Pract 2022 Jan 19;39(1):168-182. [doi: 10.1093/fampra/cmab077] [Medline: 34278421]

XSL•FO
RenderX

42. Powell RE, Henstenburg JM, Cooper G, Hollander JE, Rising KL. Patient perceptions of telehealth primary care video visits. Ann Fam Med 2017 May;15(3):225-229. [doi: 10.1370/afm.2095] [Medline: 28483887]
43. Hamilton AB, Finley EP. Reprint of: Qualitative methods in implementation research: an introduction. Psychiatry Res 2020 Jan;283:112629. [doi: 10.1016/j.psychres.2019.112629] [Medline: 31735374]
44. Kitzinger J. Qualitative research. Introducing focus groups. BMJ 1995 Jul 29;311(7000):299-302. [doi: 10.1136/bmj.311.7000.299] [Medline: 7633241]
45. Guest G, Namey E, O'Regan A, Godwin C, Taylor J. Comparing interview and focus group data collected in person and online. : Patient-Centered Outcomes Research Institute (PCORI); 2020. [doi: 10.25302/05.2020.ME.1403117064]
46. Amelung D, Whitaker KL, Lennard D, et al. Influence of doctor-patient conversations on behaviours of patients presenting to primary care with new or persistent symptoms: a video observation study. BMJ Qual Saf 2020 Mar;29(3):198-208. [doi: 10.1136/bmjqs-2019-009485] [Medline: 31326946]

Abbreviations

COM-B: Capability, Opportunity, Motivation and Behavior model
COREQ: Consolidated Criteria for Reporting Qualitative Research
GP: general practitioner
SWCM: Shannon-Weaver communication model

Edited by A Stone; submitted 20.Jun.2025; peer-reviewed by C Gray, R Pittmann; revised version received 29.Nov.2025; accepted 31.Dec.2025; published 20.Jan.2026.

Please cite as:
Alboksmaty A, Lunova T, Darzi A, Neves AL
Communication Challenges and Mitigation Strategies in Primary Care Virtual Consultations: Qualitative Study
J Med Internet Res 2026;28:e79399
URL: https://www.jmir.org/2026/1/e79399
doi:10.2196/79399

© Ahmed Alboksmaty, Tetiana Lunova, Ara Darzi, Ana-Luisa Neves. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Gender Concordance and Patient Outcomes in Indian Telemedicine: Retrospective Cross-Sectional Quantitative Study of 286,000 Consultations

Nafisa Vaz[1], PhD; Vishalkumar Jani[2], PhD

[1]Healthcare Management Department, Goa Institute of Management, Sattari, Sanquelim, Goa, India

[2]Practo Technologies Pvt Ltd., Bangalore, Karnataka, India

Corresponding Author:
Nafisa Vaz, PhD
Healthcare Management Department, Goa Institute of Management, Sattari, Sanquelim, Goa, India

Abstract

Background: Gender concordance (GC) between patients and physicians has been linked to trust and satisfaction in traditional health care. However, its role in telemedicine, especially in culturally complex settings like India, is underexplored. In India's culturally diverse and gender-sensitive context, understanding GC becomes particularly relevant for specialties such as gynecology, dermatology, psychiatry, and urology, where discussions often involve intimate or stigmatized concerns. Despite rapid telemedicine expansion, little empirical evidence exists on whether GC affects patient-reported outcomes in this context.

Objective: This study examined whether GC significantly influences patient satisfaction and self-reported recovery in teleconsultations across India, with a focus on specialty-specific effects in culturally sensitive specialties.

Methods: We conducted a retrospective cross-sectional analysis of 286,196 anonymized teleconsultation records from a national telemedicine platform (January 2023–December 2024) spanning across 20 medical specialties using binary logistic regression. Records missing gender or satisfaction data were excluded from the analysis; recovery analyses included only consultations with completed day-21 follow-up surveys (n=1170, 0.4%). Outcomes included patient satisfaction (ratings 4 - 5 on a five-point scale) and self-reported recovery at follow-up. Logistic regression models (Stata 17.0) tested associations between GC and outcomes, controlling for consultation time, duration, and physician experience. Subgroup analyses were conducted for the top 5 specialties. Each record included key data on consultation duration, timing, physician experience, specialty type, patient satisfaction rating, and self-reported recovery status. The study excluded the pediatrics specialty from the analysis to control for the parental bias.

Results: Of the 286,196 consultations, 164,008 (60.4%) were gender-concordant. Overall, 261,213 of 286,196 (91.3%) patients reported good satisfaction. GC had a statistically significant negative association with patient satisfaction (odds ratio [OR] 0.87, 95% CI 0.85 - 0.90; *P*<.001). Across gender, the male doctor received higher satisfaction. In gynecology, female patient–female doctor pairs had significantly higher odds of reporting recovery (OR 4.53, 95% CI 0.8 - 25.3; *P*=.099). Overall, consultation timing (OR 0.99, 95% CI 0.998 - 0.999; *P*<.001) and patient satisfaction (OR 20.13, 95% CI 12.06 - 35.38; *P*<.001) were stronger predictors of self-reported recovery than GC.

Conclusions: GC in telemedicine has a context-dependent impact. While it does not independently predict clinical recovery, it meaningfully shapes patient satisfaction. These findings highlight that gender sensitivity training and context-specific communication approaches may enhance telemedicine experiences in culturally sensitive domains. Integrating awareness of gender dynamics into telehealth design and policy could strengthen patient trust and engagement in virtual care. Future research should explore specialty-specific dynamics and improve follow-up response rates to better assess clinical outcomes.

(J Med Internet Res 2026;28:e78311) doi:10.2196/78311

KEYWORDS

gender concordance; telemedicine; India; patient satisfaction; clinical outcomes; culturally sensitive specialties; physician-patient dyads

XSL•FO
RenderX

Introduction

Telemedicine and Gender Dynamics in the Indian Context

The COVID-19 crisis has triggered a revolutionary change in health care delivery and expedited the worldwide penetration of telemedicine as an acceptable mode of consultation [1]. In India, with sharp health care access inequalities, especially across rural and socioeconomically disadvantaged populations, telemedicine presented itself as a vital mechanism for overcoming geographic and resource-based barriers [2,3]. Latest figures indicate a sweeping rise in virtual consults, with millions of patients depending on telehealth facilities for medical advice, diagnosis, and treatment [4].

Despite the operational success and scalability of telemedicine in India, patient experience and perceived quality are varying [5]. While aspects like connectivity, clinical competence, and usability of the platform are extensively researched [2,6], one aspect that is not well examined is the influence of gender in the construction of the virtual doctor-patient interaction. More specifically, the way the gender match or mismatch between patient and provider, also known as gender concordance (GC), has an effect in the Indian telemedicine context.

GC, alignment between the gender identity of the health care provider and the patient, has been linked to improved trust, communication, and outcomes, particularly in intimate specialties like gynecology and psychiatry [7]. But the impact of GC in telemedicine, a medium that lacks presence and is based primarily on audio-only interactions, is not well understood, particularly against the Indian backdrop [8].

India's sociocultural landscape presents an additional level of complexity. Strongly rooted gender norms of modesty and patriarchal formations tend to shape health care–seeking behavior [7-10]. Women will be reluctant to share certain symptoms with a male physician, and male patients will find female health care providers more empathic across certain specialties [11,12]. Such factors pose essential questions regarding the impact of GC on patient satisfaction and perceived effectiveness in telemedicine.

Recent telehealth studies [7,11,13] show that GC may influence trust and comfort differently across cultural contexts, but evidence from non-Western settings remains scarce. Few large-scale analyses have examined patient-reported outcomes in virtual care, especially in India. Previous studies in this area have been carried out in Western societies with varying cultural values and health care infrastructure [14,15]. Consequently, it is not possible to transfer findings from such studies directly to India. Moreover, studies so far have concentrated mostly on clinical results or communication patterns, with a lack of proper incorporation of patient-reported measures like satisfaction or self-perceived recovery. There is an utmost need for large-scale evidence from a contextualized setting that can be used for developing culture-sensitive design and implementation of telemedicine services for India.

This paper attempts to fill this lacuna by investigating whether GC is a predictor of 2-principal patient-reported outcomes, consultation satisfaction, and self-reported recovery, in telemedicine consultations across a broad spectrum of medical specialties in India. Using a database of more than 286,000 anonymous telemedicine consultations conducted by one of India's biggest telemedicine platforms, we explore the interaction between gender dynamics and other factors, including consultation timing, physician experience, and specialty type, in shaping patient perceptions of quality of care.

GC in Health Care: Global Evidence

GC, the match between a patient's and a provider's gender, has typically been linked with increased patient trust, satisfaction, and communication in conventional health care situations, physical face-to-face consultations, especially in developed economies [6,15,16]. The important point to note here is that literature about GC in clinical settings has been using self-reported sex or gender by the patients and doctors as a gender variable and does not really differentiate between gender and sex of a person like some other social science studies would have done. For instance, GC in primary care is related to longer consultation times and more patient-oriented communication [16], and, in oncology, female patients tend to be more comfortable with female providers because of impressions of empathy and emotional support [17].

Yet the results are still unclear. Previous studies [14,18] discovered minimal or no effect of GC on clinical outcomes or satisfaction, which suggests that GC by itself is not a strong predictor of health outcomes. One of the studies also stresses the differences in GC effects depending on the clinical specialty and setting in a related systematic review [15].

With the development of telemedicine, questions arise as to whether the impact of GC is present in the virtual setting. While nonverbal cues are diminished in teleconsultations, there is evidence that rapport and trust can still be established [19]. When voice-only consultation is used, as it is in resource-poor environments such as rural India, GC can facilitate comfort and disclosure by alleviating gender-based communication barriers [18]. However, the inconclusive global evidence underscores the necessity for context-oriented, culture-sensitive studies.

Sociocultural Context and Gender Preferences in India

India is a specific context to study GC because of the entrenched gender norms, especially in specialties with sensitive health matters [20]. Research reveals that female patients are likely to seek female providers, particularly in psychiatry and gynecology, where personal revelation is necessary [2,14]. Hofstede cultural dimensions, specifically masculinity-femininity and power distance, can explain why the health care expectations of the patients in India tend to be influenced by gendered constructs of authority and nurturing [21]. These sociocultural undercurrents can shape whether GC is likely to build trust or create new tensions.

Specialty-Specific Insights

GC's impact varies by specialty, widely preferred in gynecology for perceived empathy, considered in pediatrics, and supportive of disclosure in mental health, especially for women [6,8,21,22]. Those differences highlight the need to examine GC as a

XSL•FO
RenderX

specialty-specific rather than a general concept. For pediatric consultations, GC was determined based on the gender of the primary caregiver who initiated the consultation, minimizing potential bias from child accounts.

Empirical Gaps in Indian Telemedicine

Despite growing health care reach and diminished geographic barriers [1,3], telemedicine's capacity for supporting patient-provider relationships within gender-sensitive frameworks is uncertain. More satisfaction was evidenced in GC-based video interactions by Verma et al [23], but outcomes of recovery did not vary significantly. More critical factors, for example, consultation duration and doctor experience, could play an even more powerful role in shaping outcomes, and so analyses including these factors with GC as control variables are needed.

Empirical Lacunae in Indian Research

With the increasing utilization of telemedicine in India [5], there is still a scarcity of empirical evidence regarding the impact of GC on the delivery of health care virtually, especially from the patient's side. Female patients in rural and traditional areas can avoid or underdeclare their problems during interactions with male practitioners, particularly for reproductive or sexual health [9]. Previous studies [2,3,22,24] underscored the importance of system design with a gender awareness that ensures equal access. No large-scale, quantitative studies so far have thoroughly analyzed the effect of GC on both satisfaction and health outcomes in telemedicine in India.

With the mixed international evidence, this research will bridge the gaps in existing research by investigating the outcomes of GC for patient satisfaction and self-perceived recovery from teleconsultation in various medical specialties in India. The issue under investigation is relevant to both policy and practice of allocation algorithms, provider communication and behavior, and training needs for providers working for telemedicine platforms in India. The research is relevant to all telemedicine platforms—public and private.

Study Aims

This research is directed to assess the role of GC between patient and doctor in customer satisfaction (CSAT) and patient-reported medical efficacy (PRME), recovery rate, in the Indian telemedicine context. The study hopes to provide evidence-based insights into the telemedicine encounter as influenced by sociodemographic congruence between providers and patients. We expect the findings to have direct practical applications for telehealth platform design, provider training, and patient matchmaking algorithms. Guided by expectation confirmation theory (ECT) and the patient-centered care model (PCCM), this study examines whether GC predicts (1) patient satisfaction and (2) self-reported recovery in teleconsultations across India. We hypothesize that GC may have specialty-specific effects, particularly in gynecology.

Methods

Data Source and Study Design

This study draws on a dataset provided by one of the leading private Indian telemedicine platforms, recognized for its extensive reach and service diversity across specialties [25]. The platform was selected due to its nationwide footprint and volume of consultations, making it a strong representative sample for analyzing virtual care patterns across India. As of 2024, the platform reports over 400 million registered users across 22 states, positioning it among the largest telehealth providers in the country. The platform provides close to 2 million teleconsultations annually.

This retrospective cross-sectional study used secondary, anonymized data from a leading Indian telemedicine platform. The dataset covered consultations from January 2023 to December 2024.

Sample Selection and Data Processing

The dataset consists of 286,196 telemedicine consultations from various specialties and systems of medicine, such as Allopathic, Ayurveda, and Homeopathy. The current study used the anonymized data provided by the telemedicine platform. The organization has robust processes to anonymize the data and does not share any personally identifiable information, even within the organization. The data were transferred to authors in a spreadsheet where each transaction was identified with a unique identification number. For all purposes, the data were secondary in nature for the investigators.

The platform, over the period of 2 years, had done close to 3 million teleconsultations. However, for the study purposes, the platform shared only the data for consultations for which the patient had responded to the patient satisfaction survey (CSAT). The health outcome medical efficacy (PRME) survey started in April 2024, and hence it is available in very limited proportion. The inclusion of transactions was based on the fact that the CSAT or PRME survey was responded back by the patient because outcome data were important for the study. To utilize the provided data, authors had to filter the data. The transactions with reported CSAT and PRME were separated. Patients' and doctors' genders were reported, so GC was derived as a variable and utilized to categorize the transactions.

Dataset Screening Process

A structured screening process was applied to 286,196 for missing data and to form the analytical sample.

- Exclusion criteria for satisfaction analysis were as follows:
  - Missing patient gender (n=14,578)
  - Missing doctor gender (n=151)
  - Unclear satisfaction rating response (n=3)
- Exclusion criteria for recovery analysis were as follows:
  - All consultations without a completed day-21 recovery follow-up (n=284,874)
  - Pediatric consultations (n=29,288) due to caregiver-provided responses

XSL•FO
RenderX

Patients accessed the platform through a web or mobile interface. The current structure of the telemedicine platform automatically assigns a doctor to a patient based on the specialty or symptom that the patient enters into the system. An algorithm that is based on the availability of doctors on the platform at that time connects a patient with a doctor. So, there is no choice from either the doctor or patient side involved in matching.

The current study excludes the pediatrics consultations from the assessment of the impact of GC on the CSAT and PRME. The reason here is the response gets confounded by the parent or parents' preferences or biases. The patient himself or herself is not the one responding to either of the surveys.

Variables and Measures

Dependent Variables

Patient Satisfaction [csat_good]

This binary variable indicates whether the patient reported high satisfaction with the consultation. This variable is created from the 1-5 scale-based question asked at the end of each consultation. The ratings of 1, 2, and 3 were considered to be dissatisfied (csat_good=0), and the ratings of 4 and 5 were considered to be satisfied (csat_good=1).

PRME-Recovery Rate

This binary variable indicates whether the patient reported improvement in symptoms after consultation. This recovery rate question is floated to all patients on the 21st day after consultation. Patients who reported recovery are captured as prme_good=1, and 0 otherwise.

Independent Variables

GC [gender_cc]

This binary variable indicates whether the patient's gender matched the physician's gender. The platform collects this demographic information as part of preconsultation user inputs; patients self-report their sex when registering or booking a consultation. Physicians do not collect or verify this information during the consultation process. While the system technically captures biological sex, we acknowledge that much of the existing literature in this domain uses the term "GC" rather than "sex concordance." In line with prevailing academic convention, we continue using the term GC throughout this study, while recognizing this conceptual limitation in our definition of the research question.

Time of Day [time_day]

This captures consultation timing, categorized into morning and night. The platform uses a 9 AM to 9 PM window as usual day timing from the doctor's practice perspective. Hence, consultation happening in this time period is marked as time_day=1.

Consultation Duration [total_dura]

This was the total duration [in seconds] of the teleconsultation.

Physician Experience [dr_experience]

This was measured in years of practice.

Data Analysis

The study started analysis with the finding of proportions of transactions reported with positive CSAT and positive PRME across patient-doctor gender-concordant dyads and nonconcordant dyads. The study employed a *z* test to check if the 2 groups were statistically different in CSAT or PRME.

To assess the impact of GC on patient outcomes, the study employed a logistic regression model. Exponentiated coefficients (odds ratios [OR]) were reported to interpret the impact of GC and other control variables. The analysis also included the regressions at the specialty level. These specialties were selected by consultation volume. These were identified using the metric of total number of completed consultations within the study period. Logistic regression analysis requires an adequate number of outcome events relative to the number of predictor variables to ensure the stability and reliability of parameter estimates. Following conventional guidelines, a minimum of 10 outcome events per predictor variable is recommended to avoid overfitting and ensure model validity [24]. Additionally, the Central Limit Theorem supports a minimum sample size of 30 as a general rule of thumb for approximating normality in sampling distributions [26]. These principles jointly informed the decision to set 30 observations as the minimum threshold for conducting specialty-specific regression analyses in this study. This gave us 9 specialties for CSAT and 7 specialties for PRME. However, the result tables reported only the 5 specialties that had the highest number of responses available for CSAT or PRME. Analyses were conducted using Stata (version 17.0; StataCorp LLC). Binary logistic regression models assessed associations between GC and 2 outcomes, patient satisfaction, and self-reported recovery, while controlling for consultation duration, physician experience, and time of day. Specialty-level regressions were conducted where event counts were ≥30.

Ethical Considerations

This study was approved by the Board of Research Ethics (BORE) of Goa Institute of Management (Approval No. GIM/BHCM042507). The study was conducted in compliance with ethical research standards. Since the data were anonymized and obtained with necessary permissions from the telemedicine platform, no personally identifiable information was accessed. The research adheres to data protection laws and ethical guidelines related to patient confidentiality. This retrospective study was approved by the institutional review board of Goa Institute of Management. Data used were deidentified secondary records collected with informed consent for service use; the platform's terms permit anonymized secondary analyses. All identifiers were removed prior to researcher access; only aggregate data were analyzed. No compensation was involved. No identifiable images or patient details are included.

Results

Overview of the Sample

Of the 286,196 consultations analyzed, 60.4% (n=164,008) were gender concordant and 47.8% (n=129,848) involved female patients. Table 1 shows the specialty-wise shares in total

XSL•FO
RenderX

transactions. Pediatrics is shown in descriptive statistics but excluded from outcome analyses due to caregiver response bias.

XSL•FO
RenderX

Table . Summary of data and variables.

Number of teleconsultations	Count (N=286,196)
Patient age (y), median (IQR)	30.0 (25-37)
Patient sex, n (%)	
Male	141,670 (52)
Female	129,848 (48)
Doctor experience (y), median (IQR)	8.0 (5-12)
Doctor sex, n (%)	
Male	169,363 (59)
Female	116,682 (41)
Specialty, n (%)	
Ayurveda	2021 (0.7)
Cancer	600 (0.2)
Cardiology	5290 (1.8)
Dental	3329 (1.2)
Dermatology	38,824 (14)
Diabetology	3729 (1.3)
Diet and nutrition	6069 (2.1)
Ear, Nose, Throat	11,282 (3.9)
Eye and Vision	5506 (1.9)
Gastroenterology	5654 (2)
General physician	65,794 (23)
General surgery	967 (0.3)
Gynecology	37,876 (13)
Homeopathy	2754 (1)
Nephrology	1643 (0.6)
Neurology	7212 (2.5)
Orthopedics	11,871 (4.1)
Pediatrics	29,288 (10)
Physiotherapy	1427 (0.5)
Psychiatry	5688 (2)
Psychological counseling	3993 (1.4)
Pulmonology	3216 (1.1)
Rheumatology	777 (0.3)
Sexology	10,192 (3.6)
Stomach and digestion	10,176 (3.6)
Urology	9798 (3.4)
Unknown	1220 (0.4)
Time of consultation, n (%)	
Peak hours—day (9 AM to 9 PM)	193,612 (68)
Nonpeak hours—night (9 PM to 9 AM)	92,583 (32)
Consultation duration (s), median (IQR)	382 (237-600)
CSAT[a] score (1-5), n (%)	
1	12,461 (4.4)

XSL•FO
RenderX

Number of teleconsultations	Count (N=286,196)
2	3732 (1.3)
3	8787 (3.1)
4	39,024 (14)
5	222,191 (78)
CSAT (good vs bad), n (%)	
Bad	24,980 (8.7)
Good	261,213 (91)
PRME[b] responses, n (%)	1322 (0.46)
PRME (good vs bad), n (%)	
Not recovered	322 (24)
Recovered	1000 (76)
GC[c], n (%)	
Discordant	107,369 (40)
Concordant	164,008 (60)

[a]CSAT: customer satisfaction.

[b]PRME: patient-reported medical efficacy.

[c]GC: gender concordance.

Table 2 shows that 135,755 out of 148,803 (91.23%) transactions with a gender-concordant patient-doctor dyad reported good CSAT compared with 86,680 out of 93,971 (92.24%) transactions where doctor-patient GC was not present. Even when the sample was categorized for patient gender, it showed similar results. Gender-concordant transactions showed lower CSAT compared with nonconcordant transactions.

Table . Statistical difference in CSAT[a] and PRME[b] across different groups.

	CSAT			PRME		
Group	GC[c]=0, n/N (%)	GC=1, n/N (%)	*z* score (*P* value)	GC=0, n/N (%)	GC=1, n/N (%)	*z* score (*P* value)
All	86,680/93,971 (92.24)	135,755/148,803 (91.23)	8.753 (<.001)	389/502 (77.49)	495/669 (73.99)	1.38 (.17)
Female patients	51,511/55,808 (92.3)	55,057/61,003 (90.25)	12.358 (<.001)	233/304 (76.64)	203/261 (77.78)	–0.32 (.75)
Male patients	35,169/38,163 (92.19)	80,698/87,800 (91.87)	1.469 (.14)	156/198 (78.79)	292/408 (71.57)	1.9 (.06)

[a]CSAT: customer satisfaction.

[b]PRME: patient-reported medical efficacy.

[c]GC: gender concordance.

For PRME, the available transactions were only 1322, and after excluding the pediatrics specialty, the available transactions were 1199. Out of the available data on PRME, 923 of 1199 (77.01%) transactions have good PRME; that is, 77.01% patients reported that they had recovered. Out of these 1199 transactions, 1170 transactions had data on patient and doctor gender. Among these, 669 (57.17%) transactions had patient-doctor GC. Table 2 shows that gender-concordant cohort of transactions reported that PRME is 495 out of 669 (73.99%), compared with 388 out of 501 (77.49%) in the nonconcordant cohort. This difference is not statistically significant, unlike what we found with CSAT. Patient gender-wise categorization also found insignificant difference between concordant and nonconcordant cohorts. However, for the female patient cohort, the gender-concordant group reported higher PRME compared to the nonconcordant group. In contrast, it was the other way around for the male patient cohort.

Table 3 shows that, on average, male patients report higher satisfaction compared with their female counterparts, and male doctors receive better satisfaction scores compared with female doctors. However, irrespective of patient gender, a gender-discordant dyad results in a better CSAT score.

XSL•FO
RenderX

Table . CSAT[a] for different doctor genders as per patient genders.

	Patient gender		
Doctor gender	All patients, n/N (%)	Female patients, n/N (%)	Male patients, n/N (%)
All doctors	222,564/242,914 (91.6)	106,681/116,935 (91.23)	115,883/125,979 (91.98)
Female doctors	12,398/13,713 (90.41)	7678/8592 (89.36)	4720/5121 (92.17)
Male doctors	210,166/229,201 (91.69)	99,003/108,343 (91.38)	111,163/120,858 (91.98)

[a]CSAT: customer satisfaction.

GC and Patient Satisfaction

This section is focused on the logistic regression results assessing the impact of GC and other independent variables on the CSAT. The study employed logistic regression on the overall sample and separately for 9 specialties individually. Table 4 reports the regression results for the overall sample and 5 specialties that had the highest number of transactions with CSAT reported.

Table . Relationship between CSAT[a] and gender concordance in telemedicine.

	csat_good					
Variable	All, OR[b] (95% CI; *P* value)	General physician, OR (95% CI; *P* value)	Gynecology, OR (95% CI; *P* value)	Dermatology, OR (95% CI; *P* value)	Orthopedics, OR (95% CI; *P* value)	ENT, OR (95% CI; *P* value)
Gender concordance	0.874 (0.81-0.94; <.001)	0.9796 (0.92-1.04; .49)	1.079 (0.97-1.20; .16)	0.917 (0.85-0.98; .02)	1.023 (0.89-1.18; .74)	0.841 (0.73-0.97; .02)
Consultation during office hours	1.081 (1.04-1.12; <.001)	1.114 (1.05-1.18; <.001)	1.038 (0.97-1.11; .31)	1.052 (0.98-1.13; .19)	0.867 (0.74-1.02; .08)	1.229 (1.06-1.42; <.001)
Duration of the consultation	1.0004 (1.0003-1.0005; <.001)	1.0003 (1.0002-1.0004; <.001)	1.001 (1.001-1.001; <.001)	1.001 (1.001-1.001; <.001)	1.000 (1.00-1.00; .05)	1.000 (1.00-1.00; .09)
Doctor experience	1.002 (1.00-1.002; .11)	1.003 (1.00-1.01; .06)	1.001 (1.00-1.01; .63)	0.995 (0.99-1; .16)	0.955 (0.93-1; <.001)	0.985 (0.97-1; .008)
Intercept	9.46 (8.55-10.47; <.001)	10.235 (8.96-11.69; <.001)	5.5912 (4.74-6.6; <.001)	8.6154 (7.28-10.2; <.001)	16.6125 (13.43-20.55; <.001)	11.9619 (10.27-13.93; <.001)
N	228,640	65,174	36,185	38,183	11,770	11,213

[a]CSAT: customer satisfaction.

[b]OR: odds ratio.

In the full sample, gender discordance was a statistically significant predictor of patient satisfaction (OR 0.874, 95% CI 0.85 - 0.90; *P*<.001), suggesting that the presence of GC results in a lower satisfaction rate. This aligns with the findings of Table 2, showing that female patient-male doctor and male patient-female doctor dyads showed higher satisfaction rates.

At the specialty level, GC was associated with lower satisfaction in dermatology (OR 0.917, 95% CI 0.85-0.98; *P*=.02) and ENT (OR 0.841, 95% CI 0.73-0.97; *P*=.02). In the specialty of Gynecology, where patient sensitivity about doctor gender is high [27,28], GC showed a positive but statistically nonsignificant association with satisfaction. The effect of GC was not uniform, suggesting that specialty context moderates its influence on patient experience.

Except for orthopedics, all other specialties showed that transaction happening during 9 AM to 9 PM had a positive impact on the CSAT. Except for the general physician group, other specialties showed that the younger (less experienced) the doctor, the better the CSAT rate.

GC and PRME

Table 5 reports the logistic regression results for the impact of GC on the PRME—the recovery rate. It presents 7 regressions: the first 2 for the whole sample for which all required variables were available, and the next 5 were separate specialty levels. The first model included CSAT as one of the explanatory variables, whereas the other 6 models did not include it.

XSL•FO
RenderX

Table . Relationship between patient-reported medical efficacy and gender concordance in telemedicine.

Variable	prme_good						
	All, OR[a] (95% CI; *P* value)	All, OR (95% CI; *P* value)	General physician, OR (95% CI; *P* value)	Gynecology, OR (95% CI; *P* value)	Dermatology, OR (95% CI; *P* value)	Orthopedics, OR (95% CI; *P* value)	ENT, OR (95% CI; *P* value)
Customer satisfaction	20.1334 (8.11-50.02; <.001)	—[b]	—	—	—	—	—
Gender concordance	0.8492 (0.63-1.15; .29)	0.8401 (0.64-1.01; .21)	0.855 (0.51-1.43; .55)	4.53 (0.75-27.26; .099)	1.102 (0.6-2.03; .76)	1.168 (0.33-4.11; .81)	1.733 (0.62-4.83; .29)
Consultation during office hours	1.142 (0.82-1.59; .43)	1.2524 (0.93-1.69; .14)	1.492 (0.84-2.64; .17)	2.099 (0.77-5.75; .15)	1.324 (0.68-2.57; .41)	0.423 (0.09-2.09; .29)	0.716 (0.23-2.28; .57)
Duration of the consultation	0.999 (0.99-1.00; <.001)	0.999 (0.99-1.00; .001)	0.999 (0.99-1.00; .18)	1.001 (1.00-1.002; .18)	1 (0.99-1.002; .95)	0.998 (0.995-1; .17)	0.998 (0.99-1; .19)
Doctor experience	1.0032 (0.98-1.02; .75)	1.007 (0.99-1.03; .44)	0.998 (0.97-1.02; .87)	1.09 (1.00-1.19; .04)	0.965 (0.91-1.02; .20)	0.999 (0.81-1.19; .99)	1.054 (0.97-1.14; .19)
Intercept	0.2854 (0.2-0.42; <.001)	3.5164 (2.4-5.15; <.001)	5.2211 (3.16-8.62; <.001)	0.1725 (0.02-1.58; .12)	3.448 (1.47-8.1; .005)	6.9175 (0.00-9.98; .91)	2.7454 (0.68-11.06; .16)
N	1170	1170	430	114	243	46	79

[a]OR: odds ratio.

[b]not applicable.

Based on the 1170 transactions, GC had no statistically significant effect on overall recovery outcomes (OR 0.84, 95% CI 0.64 - 1.1; *P*=.21). While marginal significance was noted in gynecology (*P*=.099), given the small number of transactions (n=1170) for recovery follow-ups, findings must be interpreted cautiously.

The first model showed that patient satisfaction was the strongest and most consistent predictor of self-reported recovery (OR 20.13, 95% CI 12.06 - 35.38; *P*<.001).

Longer consultations were weakly correlated with satisfaction, not recovery. The experience of a doctor did not show a positive impact on the patient-reported recovery rate, except in the gynecology specialty.

Discussion

Principal Findings

This study examined the impact of GC on patient satisfaction and self-reported recovery in more than 286,000 Indian telemedicine consultations. Contrary to conventional assumptions [7], GC did not uniformly enhance satisfaction or recovery. In fact, nonconcordant dyads reported higher satisfaction, challenging existing evidence and expectations of demographic matching improving health care experiences [6,15,16].

These findings align with 2 theories: ECT and PCCM.

ECT explains how satisfaction arises when experiences exceed expectations [26,27], as seen in male patient-female doctor dyads. In concordant dyads, higher expectations may have gone unmet, reducing perceived trust [29]. Multiple studies indicate that male patients interacting with female physicians frequently experience their expectations being exceeded, particularly in the domains of interpersonal communication and empathy [30], leading to higher trust. Research reveals that male patient-female doctor dyads are often characterized by higher patient-reported satisfaction and quality scores, largely due to the "expectation-surprise" effect: male patients, who may hold lower initial expectations of female physicians' technical and interpersonal abilities, report greatly enhanced perceptions when those expectations are surpassed in real clinical interactions [13]. Female doctors are consistently highlighted for their superior listening, attentiveness, and clarity, which not only challenge prevailing gender stereotypes but also foster increased trust and comfort among male patients. This dynamic has been observed across different care settings and cultural contexts, underscoring the potential benefits of such cross-gender provider-patient pairings for improving patient satisfaction and health care quality [31]. These superseded expectations may meet with better treatment adherence, leading to significantly different recovery rates in male patient-female doctor dyads compared to male-male dyads.

The female patient, and the overall sample, showed higher satisfaction rates when consulted with a male doctor, which is in corroboration with general perception of men being preferred for technical roles [32] and previous evidence of this female patients rating male doctors better than the female doctors [33].

PCCM emphasizes aligning care with patient preferences, values, and context [23,34]. In telemedicine, where nonverbal cues are limited, empathy, trust, and communication quality become more crucial than demographic similarity. This helps explain why patient satisfaction, rather than GC, was the strongest predictor of self-reported recovery [17]. Figure 1 illustrates how these 2 broad theories influence patient satisfaction and self-reported health outcomes through communication and trust.

XSL•FO
RenderX

Figure 1. Theoretical framework.

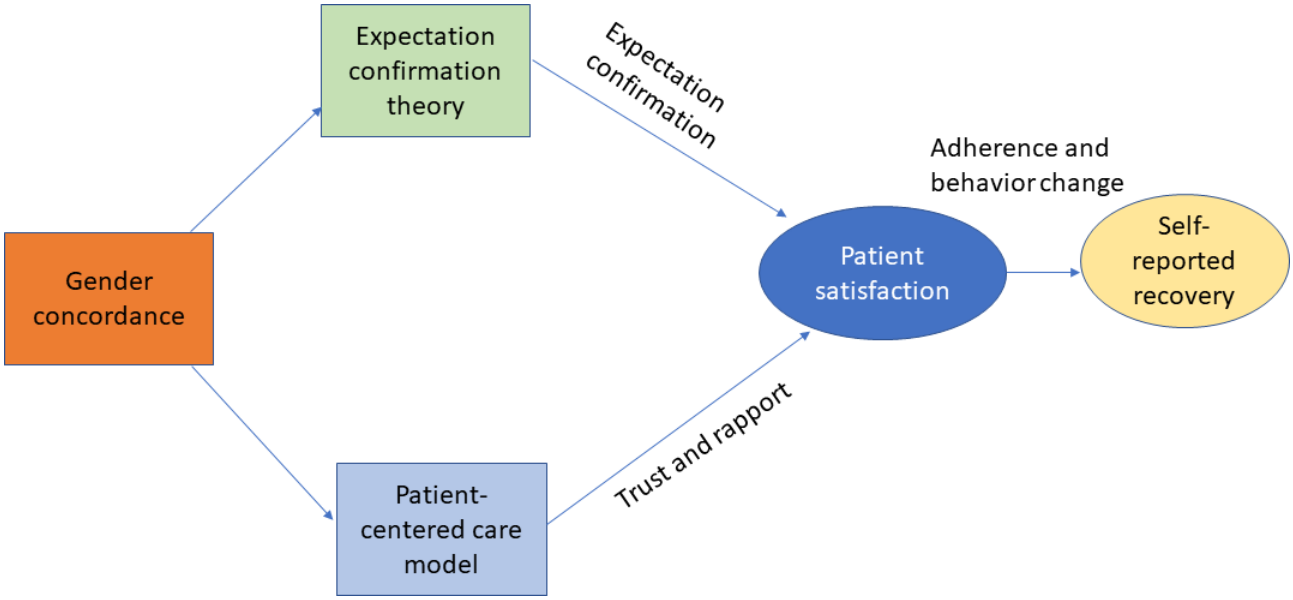

Comparison With Prior Work

These findings suggest telemedicine platforms, especially catering to Indian clients, should not default to GC-based matching. Instead, emphasis should be placed on enhancing communication skills and digital empathy, particularly in sensitive specialties such as gynecology. Male doctors are considered to be good at the technical skills and, hence, the trust and satisfaction [32,33]; and female doctors turn out to be more empathetic and better at communication, resulting in higher satisfaction [16,17]. Unlike evidence from the developed world, in the case of the Indian telemedicine context, GC is not playing any significant role. Hence, the training in active listening and patient engagement can be preferred tools over demographic alignment through matching algorithms [35].

Moreover, patient preferences, specialty, and combinations of them play an important role in final satisfaction and health outcome—recovery rate. The patient preferences should guide provider matching, particularly in specialties involving intimate care, without enforcing rigid defaults.

Strengths and Limitations

This study is among the first to explore GC and patient-reported outcomes using a very large real-world dataset from India, covering multiple specialties. Its strength lies in its scale, cross-specialty comparisons, and integration of both satisfaction and recovery indicators.

However, several limitations must be noted. First, PRME data represented <1% of consultations, limiting statistical power. Second, pediatric cases may involve caregiver responses, introducing bias; hence, the study could not include it in the analysis. Third, recall and nonresponse bias could affect postconsultation surveys. Fourth, this was a single-platform dataset, limiting generalizability. Moreover, generalizability to a very different country would also be limited, as the current evidence suggests that the Indian context behaved differently compared to Western evidence. Fifth, gender identity was operationalized as binary sex due to data limitations. The current findings are presented keeping in mind all these limiting factors. So, there is a need for further research, as described previously.

Future Directions

Further research is warranted to explore intersectional factors and evaluate objective health outcomes beyond self-reports. There is a definite need to also assess similar hypotheses of GC on health outcomes in physical outpatient department settings for understanding Indian evidence. Specialty-level research and insights would help inform not only patients but also the current medical students choosing their specialization and super-specialization. Future research should replicate this analysis across multiple telemedicine platforms, incorporating both quantitative and qualitative approaches to capture patients' subjective experiences. Including gender identity, language preference, and digital literacy could offer richer insights into patient-provider rapport.

Longitudinal or experimental designs could further assess how targeted interventions, such as empathy training or communication workshops for physicians, impact satisfaction and clinical outcomes.

Conclusions

This study examined how matching the gender of patients and doctors influenced satisfaction and patient-reported recovery rate, drawing on data from more than 286,000 telemedicine consultations across various medical specialties in India. The GC did not consistently predict recovery but did influence satisfaction. In fact, at the overall level, GC between patients and doctors resulted in lower satisfaction.

Patient satisfaction emerged as the strongest factor linked to self-reported recovery, underscoring how crucial trust, empathy, and clear communication are in virtual consultations [36]. However, because only a small fraction of consultations, less than 1% of the total data, included recovery data, conclusions drawn about health outcomes should be viewed with caution.

These findings suggest that Indian telemedicine platforms should prioritize strategies that enhance communication skills, cultural

XSL•FO
RenderX

sensitivity, and patient education. Indian patients apparently showed that gender-discordant dyads would be more effective. In specialties such as gynecology, where GC showed more pronounced effects on health outcomes, customizable matching features or patient-preference options may be warranted.

The current study has implications for telemedicine platforms, especially those catering to the Indian population. The evidence suggests that, unlike the Western world, patient satisfaction and health outcomes are not influenced by the GC; hence, there is no support for matching algorithms to consider this as a variable. The doctors need to be aware of varying preferences as per patient gender. There is also a case of further research on specialty-level evidence that may inform future doctors to choose their specializations. This may lead to a more effective patient-centric health care delivery.

In summary, while GC is not a major factor across the board in Indian telemedicine, it does matter in specific contexts. Its impact depends on the medical specialty, cultural expectations, and the quality of communication, highlighting the importance of personalized approaches to patient-provider interactions.

Acknowledgments

No generative AI tools were used in drafting or revising this manuscript.

All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The anonymized data that support this study are available from the corresponding author upon reasonable request but are not publicly shared due to confidentiality agreements with the telemedicine provider. The datasets of this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: NV, VJ
Data curation: VJ, NV
Formal analysis: VJ, NV
Writing – original draft: NV, VJ
Writing – review & editing: NV, VJ

Conflicts of Interest

None declared.

References

1. Perrin PB, Pierce BS, Elliott TR. COVID-19 and telemedicine: a revolution in healthcare delivery is at hand. Health Sci Rep 2020 Jun;3(2):e166. [doi: 10.1002/hsr2.166] [Medline: 32500101]
2. Kruse CS, Krowski N, Rodriguez B, Tran L, Vela J, Brooks M. Telehealth and patient satisfaction: a systematic review and narrative analysis. BMJ Open 2017 Aug 3;7(8):e016242. [doi: 10.1136/bmjopen-2017-016242] [Medline: 28775188]
3. Sharma P, Rao S, Krishna Kumar P, et al. Barriers and facilitators for the use of telehealth by healthcare providers in India-A systematic review. PLoS Digit Health 2024 Dec;3(12):e0000398. [doi: 10.1371/journal.pdig.0000398] [Medline: 39642108]
4. Mehrotra A, Ray K, Brockmeyer DM, Barnett ML, Bender JA. Rapidly converting to “virtual practices”: outpatient care in the era of COVID-19. NEJM Catal 2020 Apr 1;1(2). [doi: 10.1056/CAT.20.0091]
5. Aneja J, Goyal T, Verma M, Kaur G, Mirza M, Gupta S. Client satisfaction with telemedicine services during COVID-19 pandemic: a cross-sectional survey from a teaching institute of North India. J Family Med Prim Care 2022 Sep;11(9):5187-5193. [doi: 10.4103/jfmpc.jfmpc_2217_21] [Medline: 36505639]
6. Kaur KN, Niazi F, Thakur R, Saeed S, Rana S, Singh H. Patient satisfaction for telemedicine health services in the era of COVID-19 pandemic: a systematic review. Front Public Health 2022;10:1031867. [doi: 10.3389/fpubh.2022.1031867] [Medline: 36589954]
7. Kitole FA, Ali Z, Song J, et al. Exploring the gender preferences for healthcare providers and their influence on patient satisfaction. Healthcare (Basel) 2025 May 5;13(9):1063. [doi: 10.3390/healthcare13091063] [Medline: 40361841]
8. Pinedo-Torres I, Garcia-Villasante EJ, Gutierrez-Ortiz CC, et al. The doctor-patient relationship and barriers in non-verbal communication during teleconsultation in the era of COVID-19: a scoping review. F1000Res 2023 Jun 15;12:676. [doi: 10.12688/f1000research.129970.1]

XSL•FO
RenderX

9. Banerjee A, Sanyal D. Dynamics of doctor-patient relationship: a cross-sectional study on concordance, trust, and patient enablement. J Family Community Med 2012 Jan;19(1):12-19. [doi: 10.4103/2230-8229.94006] [Medline: 22518353]
10. Patel V, Rodrigues M, DeSouza N. Gender, poverty, and postnatal depression: a study of mothers in Goa, India. Am J Psychiatry 2002 Jan;159(1):43-47. [doi: 10.1176/appi.ajp.159.1.43] [Medline: 11772688]
11. Martinez KA, Rothberg MB. Physician gender and its association with patient satisfaction and visit length: an observational study in telemedicine. Cureus 2022 Sep;14(9):e29158. [doi: 10.7759/cureus.29158] [Medline: 36258932]
12. Dash S, Aarthy R, Mohan V. Telemedicine during COVID-19 in India-a new policy and its challenges. J Public Health Policy 2021 Sep;42(3):501-509. [doi: 10.1057/s41271-021-00287-w] [Medline: 34012012]
13. Si Y, Chen G, Zhou Z, Yip W, Chen X. The impact of physician-patient gender match on healthcare quality: an experiment in China. Soc Sci Med 2025 Sep;380:118166. [doi: 10.1016/j.socscimed.2025.118166] [Medline: 40451062]
14. Greenwood BN, Carnahan S, Huang L. Patient-physician gender concordance and increased mortality among female heart attack patients. Proc Natl Acad Sci U S A 2018 Aug 21;115(34):8569-8574. [doi: 10.1073/pnas.1800097115] [Medline: 30082406]
15. Lau ES, Hayes SN, Volgman AS, et al. Does patient-physician gender concordance influence patient perceptions or outcomes? J Am Coll Cardiol 2021 Mar 2;77(8):1135-1138. [doi: 10.1016/j.jacc.2020.12.031] [Medline: 33632488]
16. Kreps GL. Doctors talking with patients/patients talking with doctors: improving communication in medical visits. Health Commun 1995 Jan;7(1):67-71. [doi: 10.1207/s15327027hc0701_5]
17. Street RL, Makoul G, Arora NK, Epstein RM. How does communication heal? Pathways linking clinician-patient communication to health outcomes. Patient Educ Couns 2009 Mar;74(3):295-301. [doi: 10.1016/j.pec.2008.11.015] [Medline: 19150199]
18. Plichta JK, Williamson H, Sergesketter AR, et al. It's not you, It's me: the influence of patient and surgeon gender on patient satisfaction scores. Am J Surg 2020 Nov;220(5):1179-1188. [doi: 10.1016/j.amjsurg.2020.07.036] [Medline: 32847689]
19. Elsamadicy AA, Reddy GB, Nayar G, et al. Impact of gender disparities on short-term and long-term patient reported outcomes and satisfaction measures after elective lumbar spine surgery: a single institutional study of 384 patients. World Neurosurg 2017 Nov;107:952-958. [doi: 10.1016/j.wneu.2017.07.082] [Medline: 28743671]
20. Eggermont D, Kunst AE, Groenewegen PP, Verheij RA. Social concordance and patient reported experiences in countries with different gender equality: a multinational survey. BMC Prim Care 2024 Mar 23;25(1):97. [doi: 10.1186/s12875-024-02339-y] [Medline: 38521895]
21. Hofstede G. Dimensionalizing cultures: the Hofstede model in context. Online Read Psychol Cult 2011 Dec 1;2(1). [doi: 10.9707/2307-0919.1014]
22. Gideon J, Asthana S, Bisht R. Health systems in India: analysing barriers to inclusive health leadership through a gender lens. BMJ 2024 Jul 17;386:e078351. [doi: 10.1136/bmj-2023-078351] [Medline: 39019544]
23. Verma N, Buch B, Taralekar R, Acharya S. Diagnostic concordance of telemedicine as compared with face-to-face care in primary health care clinics in rural India: randomized crossover trial. JMIR Form Res 2023 Jun 23;7:e42775. [doi: 10.2196/42775] [Medline: 37130015]
24. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996 Dec;49(12):1373-1379. [doi: 10.1016/s0895-4356(96)00236-3] [Medline: 8970487]
25. Bagga B. Telemedicine in india: tech-driven democratization of healthcare. APAC News Network. 2024. URL: https://apacnewsnetwork.com/2024/09/telemedicine-in-india-tech-driven-democratization-of-healthcare/ [accessed 2026-01-02]
26. Kwak SG, Kim JH. Central limit theorem: the cornerstone of modern statistics. Korean J Anesthesiol 2017 Apr;70(2):144-156. [doi: 10.4097/kjae.2017.70.2.144] [Medline: 28367284]
27. Nimbi FM, Galizia R, Rossi R, et al. The biopsychosocial model and the sex-positive approach: an integrative perspective for sexology and general health care. Sex Res Soc Policy 2022 Sep;19(3):894-908. [doi: 10.1007/s13178-021-00647-x]
28. Aromaa A, Kero K, Grönlund J, et al. Let's talk about sexuality - a web-based survey of self-reported competence in sexual problems among obstetrician-gynecologists in Finland. Acta Obstet Gynecol Scand 2023 Feb;102(2):190-199. [doi: 10.1111/aogs.14492] [Medline: 36515100]
29. Zhu L, Jiang X, Cao J. Factors affecting continuance intention in non-face-to-face telemedicine services: trust typology and privacy concern perspectives. Healthcare (Basel) 2023 Jan 28;11(3):374. [doi: 10.3390/healthcare11030374] [Medline: 36766949]
30. Madanay F, Bundorf MK, Ubel PA. Physician gender and patient perceptions of interpersonal and technical skills in online reviews. JAMA Netw Open 2025 Feb 3;8(2):e2460018. [doi: 10.1001/jamanetworkopen.2024.60018] [Medline: 39951262]
31. Derose KP, Hays RD, McCaffrey DF, Baker DW. Does physician gender affect satisfaction of men and women visiting the emergency department? J Gen Intern Med 2001 Apr;16(4):218-226. [doi: 10.1046/j.1525-1497.2001.016004218.x] [Medline: 11318922]
32. AlSamhori JF, Rayyan R, Hammouri M, et al. Factors influencing gender preference towards surgeons among Jordanian adults: an investigation of healthcare bias. Sci Rep 2023 Jul 18;13(1):11614. [doi: 10.1038/s41598-023-38734-1] [Medline: 37464087]

XSL•FO
RenderX

33. Wiltshire J, Allison JJ, Brown R, Elder K. African American women perceptions of physician trustworthiness: a factorial survey analysis of physician race, gender and age. AIMS Public Health 2018;5(2):122-134. [doi: 10.3934/publichealth.2018.2.122] [Medline: 30094275]
34. Mead N, Bower P. Patient-centredness: a conceptual framework and review of the empirical literature. Soc Sci Med 2000 Oct;51(7):1087-1110. [doi: 10.1016/s0277-9536(00)00098-8] [Medline: 11005395]
35. Roter D, Hall JA. Doctor-patient communication: why and how communication contributes to the quality of medical care. In: Gellman MD, Turner JR, editors. Encyclopedia of Behavioral Medicine: Springer International Publishing; 2020:693-698. [doi: 10.1007/978-3-030-39903-0_1223]
36. Jameel A, Sahito N, Guo W, Khan S. Assessing patient satisfaction with practitioner communication: patient-centered care, hospital environment and patient trust in the public hospitals. Front Med (Lausanne) 2025;12:1544498. [doi: 10.3389/fmed.2025.1544498] [Medline: 40470044]

Abbreviations

CSAT: customer satisfaction
ECT: expectation confirmation theory
GC: gender concordance
OR: odds ratio
PCCM: patient-centered care model
PRME: patient-reported medical efficacy

Edited by A Schwartz, TDA Cardoso; submitted 30.May.2025; peer-reviewed by A Saxena, D Verran; revised version received 14.Nov.2025; accepted 24.Nov.2025; published 20.Jan.2026.

Please cite as:
Vaz N, Jani V
Gender Concordance and Patient Outcomes in Indian Telemedicine: Retrospective Cross-Sectional Quantitative Study of 286,000 Consultations
J Med Internet Res 2026;28:e78311
URL: https://www.jmir.org/2026/1/e78311
doi:10.2196/78311

© Nafisa Vaz, Vishalkumar Jani. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Virtual Delivery of Early Psychosis Care: Retrospective Cohort Study of Factors Associated With Initial Engagement

Trinity Vey[1,2], MSc; Nicole Kozloff[1,3,4], SM, MD; George Foussias[1,3], MD, PhD; Wei Wang[1,3], PhD; Albert HC Wong[1,3], MD, PhD; Aristotle N Voineskos[1,3], MD, PhD; Nicole Davis-Faroque[1,3], MD; Alexia Polillo[1,3,4], PhD

[1]Centre for Addiction and Mental Health, 1025 Queen Street West, Fourth Floor, Toronto, ON, Canada

[2]Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

[3]Department of Psychiatry, University of Toronto, Toronto, ON, Canada

[4]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Alexia Polillo, PhD
Centre for Addiction and Mental Health, 1025 Queen Street West, Fourth Floor, Toronto, ON, Canada

Abstract

Background: There is evidence that virtual delivery of early psychosis intervention (EPI) is well received by youth and has benefits such as reported improvements in accessibility, convenience, and comfort; however, potential barriers remain, including the digital divide and privacy concerns. Although initial engagement in EPI services is important for long-term recovery, little is known about initial engagement in the context of virtual care and the role of health equity and service use factors.

Objective: This study aimed to identify factors associated with attendance at the initial EPI consultation appointment when most were being delivered virtually.

Methods: This retrospective cohort study used electronic medical record data from patients aged 16 to 29 years who were referred to a large EPI program. The EPI program received 301 unique referrals that met study eligibility criteria from April to December 2020. Self-reported demographic variables were derived from the Centre for Addiction and Mental Health's structured health equity form and included age, gender, racial and ethnic group, country of birth, and sexual orientation. Service use factors derived from clinical documentation included referral source, days to consultation, and attendance at the consultation appointment, which was the primary outcome. Comparisons were made with 2018 to 2019 data from 999 participants from the same site prior to virtual care implementation using chi-square tests for categorical variables, independent *t* tests for continuous variables, and binary logistic regression.

Results: Patients had a mean age of 23.2 (SD 3.3) years, and 214 (71.1%) participants identified as male. Compared with pre–virtual care, there were significantly higher rates of inpatient referral (114/301, 37.9%) and lower rates of referral from outpatient and other providers (122/301, 40.5%) post–virtual care (χ^2_2=18.7, *P*<.001), with a small effect size and moderately narrow CI (Cramér *V*=0.120, 95% CI 0.06 to 0.17). Following univariable tests and stepwise backward selection, identifying as Black (odds ratio 0.45, 95% CI 0.21 to 0.97) and being referred from the emergency department or bridging clinic (odds ratio 0.24, 95% CI 0.08 to 0.72) were associated with decreased odds of attendance at the consultation appointment in the final adjusted model. All tests were 2-sided with an α level of .05.

Conclusions: This study is innovative in that it examines the self-reported health equity and service use factors that may contribute to nonattendance when most EPI appointments are delivered virtually, unlike previous studies that focused solely on differences in attendance rates. Although it was during the COVID-19 pandemic and may not be representative of virtual care in real-world practice, this study suggests that virtual care may improve initial engagement in EPI services; however, barriers to care still exist for Black patients and those referred from the emergency department. A hybrid model may improve connection to EPI, though targeted approaches are needed to bridge the digital divide and ensure that structurally marginalized and high-acuity patients have equitable access to care.

(J Med Internet Res 2026;28:e81313) doi:10.2196/81313

KEYWORDS

early psychosis intervention; psychosis; youth; virtual care; health equity; access to care; engagement

XSL•FO
RenderX

Introduction

The COVID-19 pandemic presented an urgent need to transition mental health services from in-person to virtual delivery, leading to research on feasibility and effectiveness. It also raised questions about the digital divide, referring to disparities between communities in access to technology, and the implications for structurally marginalized groups already facing barriers to accessing care [1]. Like many services, early psychosis intervention (EPI) programs rapidly shifted to virtual care during the COVID-19 pandemic, with the hope of maintaining access to and engagement in services. This was particularly crucial, given that approximately half of youth with psychosis do not access treatment and one-third disengage from EPI services early [2-5], despite the emphasis on identification and treatment early in the course of illness.

Research suggests that barriers preventing youth from accessing EPI treatment generally include stigma and a lack of knowledge about psychosis and where to seek treatment [6-8], with youth often viewing pathways to EPI care as complex [8]. Families can represent key sources of support throughout the help-seeking process and can help maintain engagement with services, particularly given the typical age of onset for psychosis [9,10]. Unfortunately, pathways to EPI treatment often begin with referrals from acute services, including emergency departments (EDs) or inpatient units [11,12], which also face barriers to ensuring timely outpatient follow-up. These include patient-level barriers, such as transportation or financial issues, and health systems–level issues, such as coordination between the ED and outpatient services, insufficient funding, and political disinterest [13,14]. Building partnerships between EPI and external services, including schools and shelters, can facilitate early detection, increase referrals from nonacute pathways, and improve access to vulnerable patients [14,15].

Emerging evidence suggests that virtual EPI care is well received by youth with psychosis. In recent cross-sectional studies, youth expressed satisfaction with the virtual delivery of EPI services and found it comparable to in-person treatment [16], highlighting the convenience, ease of use, accessibility, and comfort of virtual care [17]. Virtual care can also be conducive to a client-centered approach, bolstering youth autonomy and decision-making power over care [17]. While many youth have acknowledged the value of virtual care, some have reported feeling more isolated and disconnected from clinicians, finding it more difficult to express themselves during virtual appointments [17]. Other, more practical challenges include technological difficulties and privacy/confidentiality concerns. The extent to which the benefits of virtual care extend to improvements in initial EPI attendance is unclear.

Using service use data from EPI programs in the United States and Australia, 3 studies examined attendance rates before and after the implementation of virtual care, yielding inconsistent results. One study found an increase in missed appointments after virtual care implementation compared with pre–virtual care (13.3% and 7%, respectively) [17], whereas another found a 5% increase in attendance at EPI appointments following virtual care implementation [18]. A more recent study examined the ranges of missed appointments before and after virtual care implementation and found greater variability in the percentage of missed appointments in the post–virtual care (2.7% - 9%) than pre–virtual care implementation (2.8% - 6.4%) [19]. Another study examining EPI services before and after the pandemic noted a significant increase in video appointments offered postpandemic; however, no significant difference in video appointments attended [20]. Given these conflicting findings, more research is needed to evaluate the impact of virtual care implementation on EPI appointment attendance and identify factors that may lead to nonattendance.

We recently completed a study that examined factors associated with attendance at the EPI consultation appointment from 2018 to 2019, when this appointment was offered exclusively in person [5]. We found that older patients and those referred from the ED were less likely to attend the consultation appointment. Additionally, older patients and those who identified as Black or belonging to other racial and ethnic groups were more likely to be referred to EPI services from the ED compared with White patients, indicating that structurally marginalized groups are facing barriers accessing care early in the course of illness (before it becomes an emergency) [5]. While the digital divide and disparities in access to technology may be expected to increase barriers to care [19], it remains unclear if the aforementioned health equity and service use factors continue to affect access to EPI care when services are delivered virtually.

Using retrospective data from a large EPI program, we examined factors associated with attendance at the first EPI consultation appointment following the transition to most of these being scheduled virtually, with a focus on self-reported health equity and service use factors. Initial engagement was defined as attendance at the consultation appointment. We hypothesized that patients who were older, referred from the ED, or identified as Black would be less likely to attend the consultation appointment when it was delivered virtually [5].

Methods

Setting

The Centre for Addiction and Mental Health (CAMH) EPI program is the largest in Canada, serving downtown Toronto, a large urban center. The program provides consultation and 3 years of coordinated specialty care for people up to 29 years of age with affective, nonaffective, and substance-induced psychosis. The program receives referrals from the ED, CAMH inpatient units, CAMH outpatient psychiatrists, or externally through primary care providers or external inpatient and outpatient psychiatrists. The ED also houses a “bridging” clinic, where they may triage patients who are determined to be at lower acuity levels during business hours; it is also used to provide short-term follow-up after an inpatient discharge. External referrals are triaged by nurses to appropriate services. The program aims to offer consultation appointments within 2 weeks [21].

Study Design and Population

This was a retrospective cohort study that followed Strengthening the Reporting of Observational Studies in

XSL•FO
RenderX

Epidemiology (STROBE) reporting guidelines for cohort studies [22]. We replicated methods from our previous paper examining initial engagement with in-person services [5]. Electronic medical record (EMR) data were obtained for all patients aged 16 to 29 years who were referred to the CAMH EPI program from April 2020 to December 2020, the period immediately following virtual care implementation. We classified consultation appointments between 2018 and 2019 as "pre–virtual care," as they were offered exclusively in-person during this period. Appointments between April and December 2020 were classified as "post–virtual care," as the vast majority were conducted virtually, consistent with institutional practices and jurisdictional restrictions at the time. We focused on this specific period because in-person consultations resumed during subsequent phases of the pandemic. Referral data were excluded from the analysis if the referral was canceled for any reason and for patients who were previously enrolled in the CAMH EPI program. Missing data were manually extracted from EMR charts when possible, and data from second and third referrals were excluded from the analysis. Statistical comparisons were made with 2018 to 2019 data prior to virtual care implementation [5].

Variables and Missing Data

The primary outcome was rate of attendance at the first consultation appointment. Patients were coded as having not attended if they declined services, could not be reached to schedule their appointment, or did not attend their scheduled appointment. Demographic variables were self-reported by patients through CAMH's standardized health equity form that is routinely completed around the time of their first appointment and during ED visits and inpatient admissions. These variables included age, gender, racial and ethnic group ("other racial and ethnic groups" included Indigenous, Latin American, Middle Eastern, and other not specified), country of birth, and sexual orientation. Low prevalence categories were combined for data cells of fewer than 5 people to protect the privacy of participants and facilitate statistical comparisons. Recoded variables are outlined in Multimedia Appendix 1. Service use factors were derived from clinical documentation: referral source (inpatient, ED or bridging clinic, and those from outpatient psychiatrists, primary care providers, or other external providers) and days to consult (calculated as the number of days from referral to consultation appointment). We retained "do not know" or "prefer not to answer" responses in the analysis because we considered them valid responses to self-reported health equity questions. Nonresponse data were categorized as missing, and data were missing for three variables (25.9%-27.6%). We performed Little's MCAR test, which indicated that data were MCAR (χ^2_{13}=11.09, P=.60). However, we did not impute missing data because it is routinely completed by patients around the time of their first appointment and during ED visits and inpatient admissions. To minimize potential sources of bias, we applied consistent inclusion and exclusion criteria and used standardized EMR data fields to minimize misclassification.

Statistical Analysis

Descriptive statistics were calculated, including means, SDs, and percentages, to describe the demographic and service use characteristics of the full sample. We used chi-square tests for categorical variables and independent t tests for continuous variables to compare service use pre– versus post–virtual care. Effect sizes were calculated using Cohen d for continuous variables and Phi (2 × 2 tables) and Crame´r V (2 × 3 tables) for categorical variables. We used binary logistic regression to model the odds of attendance at the consultation appointment, controlling for demographic (ie, age, gender, racial and ethnic group, sexual orientation, and country of birth) and service use (ie, referral source and days to consult) factors. Simulation work by Vittinghoff and McCulloch (2007) demonstrated that logistic regression can yield reliable estimates with 5 to 9 events-per-variable (EPV) under commonly encountered modeling conditions (ie, when covariates are not extremely sparse, effect sizes are moderate, and covariates are not highly collinear). Our data satisfy all these characteristics, and the candidate covariates were prespecified based on theoretical and clinical consideration [23]. Univariable tests were performed and variables statistically significant at an a priori level of P<.20 were included in the multivariable model. Backward stepwise selection was used to determine the final adjusted model by removing variables with a P≥.20. This criterion was used as a heuristic to safeguard against type II error during model development, following recommendations suggesting that conservative thresholds (eg, .05 or .10) may prematurely exclude important variables [24]. However, we retained age and gender in the adjusted model, regardless of statistical significance, due to their theoretical and clinical importance. A 2-sided P<.05 was considered statistically significant. All statistical analyses were performed using Stata BE (version 17.0; StataCorp) [25].

Ethical Considerations

This study was approved by the Research Ethics Board at CAMH (#060/2020 - 01). Informed consent was not required by CAMH's Research Ethics Board, as retrospective deidentified EMR data were used for the analysis and posed minimal risk. Data were deidentified and additional efforts were taken to protect the privacy of participants, including combining data cells of fewer than 5 people for low prevalence categories. There was no compensation for participants in this study, as it was an analysis of retrospective EMR data. We have taken efforts to ensure that no identification of individual participants would be possible via the data in the results or supplementary material. There were no images included in the manuscript that could lead to identification of participants.

Results

Participant Characteristics

CAMH's EPI program received 383 patient referrals between April 2020 and December 2020. Referral data that did not meet study eligibility were excluded (70/383, 18.3%), and second and third referrals were filtered out of the analysis (12/313, 3.8%), leaving 301 unique patient referrals. Table 1 displays the demographic and service use characteristics of patients. Patients had a mean age of 23.2 (SD 3.3) years, 71.1% (214/301) identified as male, 29.2% (88/301) identified as White, 40.2% (121/301) identified as heterosexual, and 46.2% (139/301) were born in Canada.

XSL•FO
RenderX

Table . Characteristics of patients referred to the Centre for Addiction and Mental Health early psychosis intervention program in 2018 - 2019 and 2020.

Characteristics	All referrals	
	Pre–virtual care, 2018 - 2019 (n=999[a])	Post–virtual care, 2020 (n=301)
Age (y), mean (SD)	22.5 (3.5)	23.2 (3.3)
Age (y), median (IQR)	22 (20 - 25)	23 (21 - 26)
Gender, n (%)		
Male	654 (65.5)	214 (71.1)
Female	323 (32.3)	81 (26.9)
Trans, nonbinary, two-spirit, other, or prefer not to answer[b]	22 (2.2)	6 (2.0)
Racial and ethnic group, n (%)		
Asian	199 (19.9)	36 (12.0)
Black	176 (17.6)	55 (18.3)
White	384 (38.4)	88 (29.2)
Other racial and ethnic groups[c]	143 (14.3)	28 (9.3)
Do not know or prefer not to answer	24 (2.4)	16 (5.3)
Missing	73 (7.3)	78 (25.9)
Sexual orientation, n (%)		
Heterosexual	667 (66.8)	121 (40.2)
LGBTQ2S+[d,e]	159 (15.9)	41 (13.6)
Do not know or prefer not to answer[e]	54 (5.4)	56 (18.6)
Missing	119 (11.9)	83 (27.6)
Born in Canada, n (%)		
No	294 (29.4)	69 (22.9)
Yes	606 (60.7)	139 (46.2)
Do not know or prefer not to answer	26 (2.6)	15 (5.0)
Missing	73 (7.3)	78 (25.9)
Referral source, n (%)		
Outpatient psychiatrists, PCPs[f], or other external providers	525 (52.6)	122 (40.5)
ED[g]/Bridging	217 (21.7)	65 (21.6)
Inpatient	257 (25.7)	114 (37.9)
Days to consult, mean (SD)[h]	18.3 (13.9)	12.6 (8.9)
Days to consult, median (IQR)	15 (10-22)	11 (7-15)
Monthly referrals, mean (SD)	44.2 (9.9)	34 (4.4)

[a]Pre–virtual care data from Polillo et al [5].

[b]2018 - 2019 data includes two-spirit and prefer not to answer.

[c]Other racial and ethnic groups include Indigenous, Latin American, Middle Eastern, and other not specified.

[d]LGBTQ2S+: lesbian, gay, bisexual, trans, queer (or sometimes questioning), and two-spirited.

[e]2018 - 2019 data includes two-spirit and 2020 data includes prefer not to answer.

[f]PCP: primary care provider.

[g]ED: emergency department.

[h]Includes data only for participants who booked a consultation appointment (n=280).

XSL•FO
RenderX

Service Use

Service use characteristics are displayed in Tables 1 and 2. Approximately one-fifth of patients (65/301, 21.6%) were referred from the ED or bridging clinic, 37.9% (114/301) from inpatient units, and 40.5% (122/301) from other referral sources. There were significantly higher rates of inpatient referral (114/301, 37.9%) and lower rates of referral from outpatient and other providers (122/301, 40.5%) post–virtual care compared with pre–virtual care (χ^2_2=18.7, P<.001), with a small effect size and moderately narrow CI (Cramér V=0.120, 95% CI 0.06 to 0.17). The mean number of days from referral to consultation appointment was 12.6 (SD 8.9), which was significantly lower (t_{1149}=6.44, 95% CI 3.95 to 7.41; P<.001) than wait times pre–virtual care (mean 18.3, SD 13.9), with a moderate effect size and narrow CI (Cohen d=0.44, 95% CI 0.31 to 0.58). The mean number of monthly referrals post–virtual care (mean 34, SD 4.4) was significantly lower (t_{1298}=−17.3, 95% CI −10.9 to −8.75; P<.001) than pre–virtual care (mean 44.2, SD 9.9), with a large effect size and a narrow CI (Cohen d=−1.14, 95% CI −1.27 to −1.00). Overall, 84.1% (253/301) of patients attended their consultation appointment post–virtual care, which was significantly higher than attendance pre–virtual care (770/999, 77.1%); however, it was a small effect with a moderately wide CI (χ^2_1=6.71, ϕ=0.072, 95% CI 0.02 to 0.12; P=.01). Post–virtual care, patients referred from the ED or bridging clinic had the highest rate of nonattendance (18/65, 27.7%) at the consultation appointment compared with those referred from inpatient units (19/114, 16.7%) or other providers (11/122, 9%).

Table . Rates of attendance at early psychosis intervention consultation appointments at the Centre for Addiction and Mental Health by referral source in 2018 - 2019 and 2020.

Referral source	Outcome of referral					
	Pre–virtual care, 2018-2019[a]			Post–virtual care, 2020		
	Patients, n	Attended consult, n (%)	Did not attend consult[a], n (%)	Patients, n	Attended consult, n (%)	Did not attend consult[b], n (%)
All referral sources	999	770 (77.1)	229 (22.9)	301	253 (84.1)	48 (15.9)
Inpatient	257	215 (83.7)	42 (16.3)	114	95 (83.3)	19 (16.7)
ED[c]/Bridging	217	145 (66.8)	72 (33.2)	65	47 (72.3)	18 (27.7)
Outpatient psychiatrists, PCPs[d], or other external providers	525	410 (78.1)	115 (21.9)	122	111 (91.0)	11 (9.0)

[a]Pre–virtual care data from Polillo et al [5].

[b]Nonattendance at consult includes those who declined services, could not be reached for booking, or booked and did not attend.

[c]ED: emergency department.

[d]PCP: primary care provider.

Factors Associated With Attendance at Consultation Appointment

Following univariable tests and stepwise backward selection, identifying as Black (odds ratio 0.45, 95% CI 0.21 to 0.97) and being referred from the ED or bridging clinic (odds ratio 0.24, 95% CI 0.08 to 0.72) were associated with decreased odds of attendance at the consultation appointment in the final adjusted model (Table 3).

XSL•FO
RenderX

Table . Logistic regression analysis of factors associated with attendance at early psychosis intervention consultation appointment at the Centre for Addiction and Mental Health during virtual care delivery in 2020 (n=301).

Variable	Attendance at EPI[a] consultation appointment			
	Univariable		Multivariable	
	OR[b] (95% CI)	*P* value	OR (95% CI)	*P* value
Age (y)	0.94 (0.85 - 1.03)	.18	0.91 (0.80 - 1.02)	.11
Gender				
Male	Reference	—[c]	Reference	—
Female, trans, nonbinary, or other	0.98 (0.50 - 1.94)	.96	1.24 (0.51 - 3.00)	.64
Racial and ethnic group				
Asian[d]	0.65 (0.24 - 1.82)	.42	—	—
Black	0.42 (0.18 - 0.99)	.046	0.45 (0.21 - 0.97)	.04
White	Reference	—	Reference	—
Other racial and ethnic groups, do not know, or prefer not to answer[d,e]	1.58 (0.48 - 5.21)	.45	—	—
Sexual orientation				
Heterosexual	Reference	—	Reference	—
LGBTQ2S+[f], do not know, or prefer not to answer	1.39 (0.67 - 2.88)	.37	—	—
Born in Canada				
Yes	Reference	—	Reference	—
No, do not know, or prefer not to answer	0.91 (0.45 - 1.87)	.80	—	—
Referral source				
Outpatient psychiatrists, PCPs[g], or other external providers	Reference	—	Reference	—
ED[h]/Bridging	0.26 (0.11 - 0.59)	.001	0.24 (0.08 - 0.72)	.01
Inpatient	0.50 (0.22 - 1.09)	.08	0.48 (0.17 - 1.41)	.18
Days to consult	1.04 (0.98 - 1.10)	.22	—	—

[a]EPI: early psychosis intervention.

[b]OR: odds ratio

[c]Not applicable

[d]Variables removed from the adjusted model through backward stepwise selection.

[e]Other racial and ethnic groups include Indigenous, Latin American, Middle Eastern, and other not specified.

[f]LGBTQ2S+: lesbian, gay, bisexual, trans, queer (or sometimes questioning), and two-spirited.

[g]PCP: primary care provider

[h]ED: emergency department

Discussion

Principal Findings

Using retrospective cohort data from 301 patients, we found that patients referred to EPI services after the transition to virtual care had higher rates of attendance at the consultation appointment and encountered shorter wait times compared to those referred pre–virtual care. We also found higher rates of inpatient referral and lower rates of referral from outpatient and other providers post–virtual care compared to pre–virtual care. Approximately one-third of patients did not attend their consultation appointment; ED or bridging clinic referrals had the highest rate of nonattendance at the consultation appointment compared with other referral sources. Equity-related and service use factors, notably identifying as Black and being referred from

XSL•FO
RenderX

the ED or bridging clinic, were associated with decreased odds of attending the consultation appointment when it was mostly delivered virtually. These findings suggest that virtual care can improve initial engagement in EPI services; however, patients from structurally marginalized groups and those referred from acute sources still face barriers to care, even when appointments are delivered virtually.

Comparison With Prior Work

We observed a 7% increase in attendance at the EPI consultation appointment following the implementation of virtual care, which is higher than the 5% increase found across all appointments in a prior study [18]. This difference is likely attributable to the focus on the consultation appointment (rather than all appointments). Findings from both studies, when taken together, lend support to virtual care as a tool for improving initial engagement in EPI services, as well as throughout treatment; however, it is important to acknowledge that other factors, such as the COVID-19 lockdown which restricted nonessential outings and may have made youth more available to attend their appointments, may have influenced appointment attendance. We also observed a decrease in wait times from EPI referral to the consultation appointment post–virtual care compared to pre–virtual care, consistent with emerging evidence that virtual care may allow for increased efficiency and streamlined processes [17]. Although the association of shorter wait times with virtual care is an important finding, it is also important to note that monthly referrals to EPI were significantly lower post–virtual care compared with pre–virtual care, which, in turn, could be influencing the shorter wait times. Nevertheless, longer wait times for EPI services have been associated with worse patient outcomes, not only during the initial period of untreated psychosis but also later in treatment [26].

Our results also align with existing research on disparities in virtual care access. Patients referred from the ED or bridging clinic had decreased odds of attending the consultation appointment post–virtual care, consistent with our prior work showing lower attendance rates at in-person initial appointments for those referred from the ED or bridging clinic [5]. Individuals seeking care from acute services may experience greater symptom burden, such as disorganization, paranoia, and a fear of surveillance, which may affect the utilization of virtual care. For example, patients may be too disorganized to navigate connecting online at the prescribed time [27]. Additionally, these patients may be more likely to experience socioeconomic disadvantage, which is linked to increased ED usage, decreased engagement with outpatient mental health care [28], and digital equity barriers. In contrast to findings from our previous study examining in-person attendance [5], we also found that Black patients had a decreased odds of attending the consultation appointment post–virtual care, suggesting that virtual care may amplify health inequities through the digital divide. This is consistent with other research reporting that Black patients were less likely to use telehealth services compared to White patients [29,30], likely due to systemic barriers that can limit access to devices, the internet, or digital literacy skills [31,32]. This is layered upon a long history of racism, oppression, and stigmatization of Black people by the medical system, and ongoing inequities, including a greater risk of police involvement and involuntary hospitalization for Black youth experiencing psychosis [33]. Overcoming systemic barriers to psychosis care for Black patients will require systemic solutions, such as increasing the racial and ethnic diversity of clinicians, ensuring a trauma-informed and anti-racist approach to services, and including lived experience perspectives of Black patients and their families in EPI evaluation [34].

Implications

These findings have several implications for the design and delivery of EPI services in a postpandemic context. The use of a hybrid model in EPI, integrating virtual care with in-person appointments, can be a way to engage patients in treatment in a way that meets their needs while still maintaining the benefits of an in-person therapeutic relationship. Global surveys of clinicians and clinical leaders show strong support for hybrid models, though challenges exist around infrastructure, training, and digital literacy [35]. Mental health care organizations, including EPI programs, should consider building up technological infrastructure and capacity to support clinicians' ability to offer hybrid models of care, although we acknowledge that this may be challenging for more low-resourced settings. Additionally, our results showed a shift in referral patterns post–virtual care, with increased inpatient and decreased outpatient referrals. This may reflect delays in early detection of psychosis due to school closures or reduced access to primary care and other outpatient providers. It is clear that outpatient providers represent irreplaceable sources for early detection and EPI access. Leveraging hybrid care models can improve connection to outpatient providers, though consideration of systemic solutions to reduce inequities in digital access is required.

Limitations

This study has several limitations. Due to the wide CIs in the final adjusted model, the findings should be interpreted with caution. Variables, including race/ethnicity, sexual orientation, and country of birth, had more than 10% of missingness, which was not imputed due to it being self-reported health equity data. There were increases in the percentage of missing data for certain demographic categories from pre–virtual care to post–virtual care, including racial and ethnic group and sexual orientation. These changes highlight potential challenges with the reliability of self-reported data, particularly in the virtual care context. While the overall missingness is reported as more than 10%, the variability in the self-reported health equity data between the two timepoints may affect the accuracy and generalizability of the findings. Future research should explore methods to improve the completeness and reliability of self-reported data within virtual settings. The time period of the study was also a limitation, as it was during the COVID-19 pandemic and may not be representative of virtual care in real-world practice. Thus, findings may be partially related to pandemic effects rather than the transition to virtual care. The use of data from 2020 may also limit the relevance of our findings, given changes in healthcare delivery and policy over time [36]. We were not able to include other relevant patient- and system-level factors, such as socioeconomic status and illness severity, as covariates in the model. These factors may

XSL•FO
RenderX

better explain the associations observed in our final adjusted model. For example, lower referral volumes and treatment delays may have led to patients presenting to care when they were more ill and had a higher symptom burden [37]. As such, not having specific data on illness severity limits our ability to understand its potential impact on EPI attendance and engagement and should be an area for future investigation. It is also important to note that patients referred from the ED may differ from those referred through outpatient and inpatient settings [5], limiting the generalizability of our findings across referral sources. Another limitation is that the study setting was a well-resourced hospital with infrastructure and technological support dedicated to virtual care implementation, the findings may not be generalizable to lower-resourced mental health settings.

Conclusions

This retrospective cohort study of patients referred to EPI services is innovative in that it examines the self-reported health equity and service use factors that may contribute to nonattendance when most EPI appointments are delivered virtually. This is unlike previous studies, which solely focused on differences in attendance rates. Although the time period of the study is a limitation, as it was during the COVID-19 pandemic and may not be representative of virtual care in real-world practice, we found improved attendance rates at the first EPI consultation appointment and shorter wait times, suggesting that virtual care may improve initial engagement in EPI services. Despite this, barriers to care still exist for Black patients and those referred from the ED. The use of a hybrid model can be a way to improve connection to EPI, though targeted approaches are needed to improve the digital divide and ensure that structurally marginalized and high-acuity patients have equitable access to EPI care. Future research should examine virtual engagement in EPI services in a postpandemic context.

Acknowledgments

There was no use of generative artificial intelligence (AI) technology in the generation of text, figures, or other informational content of this manuscript.

Funding

This work was supported by the University of Toronto Department of Psychiatry’s Reasons for Hope Fund and the Canadian Institutes of Health Research.

Data Availability

The data that supports the findings of this study are available from the corresponding author, AP, upon reasonable request to protect the privacy and confidentiality of participants.

Authors' Contributions

Conceptualization**:** AP (lead), NK (equal)
Formal analysis**:** AP (lead), TV (equal)
Methodology**:** AP (lead), NK (equal), AHCW (support), GF (support), WW (support), AV (support)
Project administration**:** AP (lead), NK (support)
Supervision**:** AP
Visualization**:** TV (lead), AP (support)
Writing – original draft**:** TV (lead), AP (support)
Writing – review & editing**:** All authors

Conflicts of Interest

None declared.

Multimedia Appendix 1
Recoded variables.
[DOCX File, 15 KB - jmir_v28i1e81313_app1.docx]

Checklist 1
STROBE checklist.
[PDF File, 176 KB - jmir_v28i1e81313_app2.pdf]

References

1. Litchfield I, Shukla D, Greenfield S. Impact of COVID-19 on the digital divide: a rapid review. BMJ Open 2021 Oct 12;11(10):e053440. [doi: 10.1136/bmjopen-2021-053440] [Medline: 34642200]

XSL•FO
RenderX

2. Anderson KK, Norman R, MacDougall AG, et al. Disparities in access to early psychosis intervention services: comparison of service users and nonusers in health administrative data. Can J Psychiatry 2018 Jun;63(6):395-403. [doi: 10.1177/0706743718762101] [Medline: 29562748]
3. Doyle R, Turner N, Fanning F, et al. First-episode psychosis and disengagement from treatment: a systematic review. Psychiatr Serv 2014 May 1;65(5):603-611. [doi: 10.1176/appi.ps.201200570] [Medline: 24535333]
4. Mascayano F, van der Ven E, Martinez-Ales G, et al. Disengagement from early intervention services for psychosis: a systematic review. Psychiatr Serv 2021 Jan 1;72(1):49-60. [doi: 10.1176/appi.ps.201900375] [Medline: 33234052]
5. Polillo A, Foussias G, Wang W, et al. Care pathways and initial engagement in early psychosis intervention services among youths and young adults. JAMA Netw Open 2023 Sep 5;6(9):e2333526. [doi: 10.1001/jamanetworkopen.2023.33526] [Medline: 37703014]
6. dos Santos Martin I, Giacon BCC, Vedana KGG, Zanetti ACG, Fendrich L, Galera SAF. Where to seek help? Barriers to beginning treatment during the first-episode psychosis. Int J Nurs Sci 2018 Jul 10;5(3):249-254. [doi: 10.1016/j.ijnss.2018.06.007] [Medline: 31406833]
7. Tiller J, Maguire T, Newman-Taylor K. Early intervention in psychosis services: a systematic review and narrative synthesis of barriers and facilitators to seeking access. Eur Psychiatry 2023 Nov 6;66(1):e92. [doi: 10.1192/j.eurpsy.2023.2465] [Medline: 37929296]
8. Anderson KK, Fuhrer R, Malla AK. "There are too many steps before you get to where you need to be": help-seeking by patients with first-episode psychosis. J Ment Health 2013 Aug;22(4):384-395. [doi: 10.3109/09638237.2012.705922] [Medline: 22958140]
9. Del Vecchio V, Luciano M, Sampogna G, et al. The role of relatives in pathways to care of patients with a first episode of psychosis. Int J Soc Psychiatry 2015 Nov;61(7):631-637. [doi: 10.1177/0020764014568129] [Medline: 25614470]
10. Lucksted A, Stevenson J, Nossel I, Drapalski A, Piscitelli S, Dixon LB. Family member engagement with early psychosis specialty care. Early Interv Psychiatry 2018 Oct;12(5):922-927. [doi: 10.1111/eip.12403] [Medline: 27863039]
11. Anderson KK, Fuhrer R, Malla AK. The pathways to mental health care of first-episode psychosis patients: a systematic review. Psychol Med 2010 Oct;40(10):1585-1597. [doi: 10.1017/S0033291710000371] [Medline: 20236571]
12. Anderson KK, Fuhrer R, Wynant W, Abrahamowicz M, Buckeridge DL, Malla A. Patterns of health services use prior to a first diagnosis of psychosis: the importance of primary care. Soc Psychiatry Psychiatr Epidemiol 2013 Sep;48(9):1389-1398. [doi: 10.1007/s00127-013-0665-3] [Medline: 23429939]
13. Walker ER, Fukuda J, McMonigle M, Nguyen J, Druss BG. A qualitative study of barriers and facilitators to transitions from the emergency department to outpatient mental health care. Psychiatr Serv 2021 Nov 1;72(11):1311-1319. [doi: 10.1176/appi.ps.202000299] [Medline: 33887957]
14. O'Connell N, O'Connor K, McGrath D, et al. Early intervention in psychosis services: a systematic review and narrative synthesis of the barriers and facilitators to implementation. Eur Psychiatry 2021 Dec 16;65(1):e2. [doi: 10.1192/j.eurpsy.2021.2260] [Medline: 34913421]
15. Cheng C, Dewa CS, Goering P. Matryoshka Project: lessons learned about early intervention in psychosis programme development. Early Interv Psychiatry 2011 Feb;5(1):64-69. [doi: 10.1111/j.1751-7893.2010.00255.x] [Medline: 21272277]
16. Dong F, Jumper MBE, Becker-Haimes EM, et al. Tele-mental health transitions for pennsylvania coordinated specialty care programs for early psychosis during the COVID-19 pandemic. Psychiatr Q 2023 Jun;94(2):89-102. [doi: 10.1007/s11126-023-10015-0] [Medline: 36820952]
17. Randall LA, Raisin C, Waters F, Williams C, Shymko G, Davis D. Implementing telepsychiatry in an early psychosis service during COVID-19: experiences of young people and clinicians and changes in service utilization. Early Interv Psychiatry 2023 May;17(5):470-477. [doi: 10.1111/eip.13342] [Medline: 35943177]
18. Chaudhry S, Weiss A, Dillon G, O'Shea A, Hansel TC. Psychosis, telehealth, and COVID-19: successes and lessons learned from the first wave of the pandemic. Disaster Med Public Health Prep 2022 Oct;16(5):1785-1788. [doi: 10.1017/dmp.2021.42] [Medline: 33588969]
19. Nelson EB, Franco OH, Patton BA, et al. Telehealth of coordinated specialty care in early psychosis during COVID-19. J Clin Psychiatry 2022 Nov 30;84(1):21m14259. [doi: 10.4088/JCP.21m14259] [Medline: 36449475]
20. Duggan A, Buckley S, Fletcher E, et al. The impact of the COVID-19 pandemic on first-episode psychosis presentations in two early intervention in psychosis services. Ir J Psychol Med 2024 Dec 10:1-8. [doi: 10.1017/ipm.2024.61] [Medline: 39655482]
21. Early psychosis intervention program standards. : Ministry of Health and Long-Term Care; 2011 URL: https://epiontario.org/wp-content/uploads/2024/10/EPI-Program-Standards-2011_.pdf [accessed 2026-01-28]
22. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
23. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol 2007 Mar 15;165(6):710-718. [doi: 10.1093/aje/kwk052] [Medline: 17182981]
24. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression: John Wiley & Sons; 2013. URL: https://onlinelibrary.wiley.com/doi/book/10.1002/9781118548387 [accessed 2026-01-28]

XSL•FO
RenderX

25. Stata. URL: https://www.stata.com/ [accessed 2026-01-28]
26. Reichert A, Jacobs R. The impact of waiting time on patient outcomes: evidence from early intervention in psychosis services in England. Health Econ 2018 Nov;27(11):1772-1787. [doi: 10.1002/hec.3800] [Medline: 30014544]
27. Sugarman DE, Horvitz LE, Greenfield SF, Busch AB. Clinicians' perceptions of rapid scale-up of telehealth services in outpatient mental health treatment. Telemed J E Health 2021 Dec;27(12):1399-1408. [doi: 10.1089/tmj.2020.0481] [Medline: 33600272]
28. Sweeney S, Air T, Zannettino L, Galletly C. Psychosis, socioeconomic disadvantage, and health service use in South Australia: findings from the second Australian National Survey of Psychosis. Front Public Health 2015;3:259. [doi: 10.3389/fpubh.2015.00259] [Medline: 26636059]
29. Adepoju OE, Chae M, Ojinnaka CO, Shetty S, Angelocci T. Utilization gaps during the COVID-19 pandemic: racial and ethnic disparities in telemedicine uptake in federally qualified health center clinics. J Gen Intern Med 2022 Apr;37(5):1191-1197. [doi: 10.1007/s11606-021-07304-4] [Medline: 35112280]
30. Buis LR, Brown LK, Plegue MA, et al. Identifying inequities in video and audio telehealth services for primary care encounters during COVID-19: repeated cross-sectional, observational study. J Med Internet Res 2023 Sep 29;25:e49804. [doi: 10.2196/49804] [Medline: 37773609]
31. Curtis ME, Clingan SE, Guo H, Zhu Y, Mooney LJ, Hser YI. Disparities in digital access among American rural and urban households and implications for telemedicine-based services. J Rural Health 2022 Jun;38(3):512-518. [doi: 10.1111/jrh.12614] [Medline: 34355427]
32. Iasiello JA, Rajan A, Zervos E, Parikh AA, Snyder RA. Racial differences in patient-reported access to telehealth: an important and unmeasured social determinant of health. JCO Oncol Pract 2023 Dec;19(12):1215-1223. [doi: 10.1200/OP.23.00006] [Medline: 37844269]
33. Knight S, Jarvis GE, Ryder AG, Lashley M, Rousseau C. 'It just feels like an invasion': Black first-episode psychosis patients' experiences with coercive intervention and Its influence on help-seeking behaviours. J Black Psychol 2023 Apr;49(2):200-235. [doi: 10.1177/00957984221135377]
34. Davis B, Anglin DM, Oluwoye O, Keshavan M. The unfulfilled promise of equitable first episode care for Black-Americans: a way forward. Schizophr Res 2022 Mar;241:171-173. [doi: 10.1016/j.schres.2022.01.046] [Medline: 35124436]
35. Torous J. Hybrid and virtual care: interest outstrips infrastructure. NEJM Catal Innov Care Deliv 2025 Jan 15;6(2). [doi: 10.1056/CAT.25.0009]
36. Danseco E, Kurzawa J, Sundar P, Brown J, Huang C. Evaluating the sector-wide implementation of virtual child and youth mental health services in response to the COVID-19 pandemic: perspectives from service providers, agency leaders and clients. Implement Res Pract 2021;2:26334895211045690. [doi: 10.1177/26334895211045690] [Medline: 37089999]
37. Penttilä M, Jääskeläinen E, Hirvonen N, Isohanni M, Miettunen J. Duration of untreated psychosis as predictor of long-term outcome in schizophrenia: systematic review and meta-analysis. Br J Psychiatry 2014 Aug;205(2):88-94. [doi: 10.1192/bjp.bp.113.127753] [Medline: 25252316]

Abbreviations

CAMH: Centre for Addiction and Mental Health
ED: emergency department
EMR: electronic medical record
EPI: early psychosis intervention
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 29.Jul.2025; peer-reviewed by D Hilty, L Agrawal; revised version received 01.Dec.2025; accepted 04.Dec.2025; published 19.Feb.2026.

Please cite as:
Vey T, Kozloff N, Foussias G, Wang W, Wong AHC, Voineskos AN, Davis-Faroque N, Polillo A
Virtual Delivery of Early Psychosis Care: Retrospective Cohort Study of Factors Associated With Initial Engagement
J Med Internet Res 2026;28:e81313
URL: https://www.jmir.org/2026/1/e81313
doi:10.2196/81313

© Trinity Vey, Nicole Kozloff, George Foussias, Wei Wang, Albert HC Wong, Aristotle N Voineskos, Nicole Davis-Faroque, Alexia Polillo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Rising to the Challenge of Early Screening in Primary Health Care Through the Web Italian Network for Autism Spectrum Disorder (Win4ASD) in the Pediatric Population: Retrospective Observational Study

Noemi Buo[1*], MSc; Eleonora Rosi[1*], MSc; Silvia Busti Ceccarelli[1], MSc; Mariarosa Ferrario[2], MD; Valerio Maiorca[2]; Erika Morandi[2], MSc; Nicole Viganó[1], MSc; Ivan Limosani[3], MD; Laura Falcone[3], MSc; Massimo Molteni[1], MD; Paola Colombo[1], MSc, PhD

[1]Child Psychopathology Unit-Smart Lab, Scientific Institute, Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Eugenio Medea, Bosisio Parini, Lecco, Lazio, Italy

[2]Azienda Socio Sanitaria Territoriale della Valle Olona - Dipartimento di salute mentale e delle dipendenze, Gallarate, Varese, Italy

[3]Direzione Generale Welfare, Lombardy Region, Milan, Italy

[*]these authors contributed equally

Corresponding Author:
Eleonora Rosi, MSc
Child Psychopathology Unit-Smart Lab
Scientific Institute, Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Eugenio Medea
via don Luigi Monza 20
Bosisio Parini, Lecco, Lazio, 23842
Italy
Phone: 39 031877623
Email: eleonora.rosi@lanostrafamiglia.it

Abstract

Background: Early identification of autism spectrum disorder (ASD) and neurodevelopmental disorders is a key element in the ability to intervene early and improve children's outcomes. The American Academy of Pediatrics recommends routine autism screening for children aged 18 to 24 months. Telemedicine (health care services delivered remotely using digital technologies) is a valuable tool in supporting this objective. As highlighted in the literature, telemedicine increases the availability of treatment, reduces diagnostic waiting times, and supports the monitoring of neurodevelopmental disorders. The Web Italian Network for Autism Spectrum Disorder (Win4ASD) platform is a telemedicine-based screening tool designed to detect ASD early within the pediatric health care system in Italy.

Objective: This study investigated the integration of the Win4ASD platform into the Italian National Health Service, providing a detailed account of its operational features and use patterns among family pediatricians. It further examined the platform's contribution to reinforcing preventive health care strategies and enhancing coordination across child health services. Finally, through a retrospective observational analysis, this study evaluated the clinical effectiveness of Win4ASD in facilitating the early detection of ASD and improving screening outcomes.

Methods: The Win4ASD platform, which is active in the Lombardy region, helps family pediatricians screen toddlers for autism at an early age (18-24 months) using the Checklist for Autism in Toddlers. Furthermore, children identified as at risk are referred for further diagnostic evaluation to the local specialized autism unit—a mental health specialist team comprising child neuropsychiatrists, psychologists, and developmental therapists. This platform, which created an integrated network for early identification of autism, ensures rapid access to diagnostic and therapeutic services. This study analyzed data collected from the platform between January 2022 and September 2024 to assess family pediatricians' commitment and effectiveness in identifying children at risk of autism using descriptive analyses performed in SPSS.

Results: The first key finding was the commitment of family pediatricians from the Regional Public Health Services in Lombardy to use the platform. In total, 72.5% (909/1253) of all family pediatricians working for the Regional Public Health Services actively used the platform. The platform enabled the screening of 58,419 infants from the general population in Lombardy, of whom 596 (1%) were identified as "at risk" and referred for fast-track diagnostic evaluation, which is consistent with global prevalence

XSL•FO
RenderX

rates. In addition, diagnostic outcomes from a subsample are presented, showing the effectiveness of the platform in facilitating the early diagnosis of neurodevelopmental disorders.

Conclusions: Early identification of risk and timely intervention can prevent the progressive development of behavioral atypicalities, have a significant impact on the quality of life of the child and their family, and allow for a reduction in health care costs and a consequent improvement in the organization of the health care system.

(J Med Internet Res 2026;28:e74302) doi:10.2196/74302

KEYWORDS

screening Checklist for Autism in Toddlers; CHAT; online screening; tele–mental health; primary care; early detection; autism spectrum disorder; health care policy; surveillance system

Introduction

Prevention of Neurodevelopmental Disorders and Telemedicine

Telemedicine refers to the full range of health care services delivered through innovative digital technologies whereby the health professional and the patient are not in the same location [1,2]. Recently, telemedicine interventions have gained significant interest and expanded into many areas of medicine, including mental health, both before and after the COVID-19 pandemic [3-5]. The pandemic first increased the demand for effective health support, and the growing need for easy-to-access remote services has substantially boosted the use of telehealth since then [5,6]. According to recent reviews, telehealth can potentially increase treatment availability; reduce diagnostic waiting times; and aid in the monitoring of mental disorders, particularly neurodevelopmental disorders [5,7].

Neurodevelopmental disorders, such as autism spectrum disorder (ASD) and attention-deficit/hyperactivity disorder, typically begin in childhood and may persist as chronic conditions throughout life [8]. Neurodevelopmental conditions account for the highest estimated prevalence in children aged <15 years worldwide, ranging from 4.1% (<5 years) to 7% (10-14 years) [9]. The World Health Organization emphasizes that identifying children at risk of neurodevelopmental disorders is a critical first step to establishing a strong relationship between parents and health care professionals to ensure early intervention [10]. Early intervention aims to prevent or minimize motor, cognitive, and emotional impairments in young children disadvantaged by biological or environmental risk factors [11].

In summary, the high prevalence of neurodevelopmental disorders and the need for early intervention make prevention and screening more crucial than ever. Telemedicine can be an effective tool in identifying these disorders at an early stage.

The Health Care System in Italy

The Italian National Health Service (NHS) is a public, regionally based system organized at the national, regional, and local levels [12] and provides free primary care, inpatient care, and health screenings for all residents. Public health care is financed primarily through national and regional taxes (97%) and patient copayments. Private health insurance plays a limited role in Italy's health coverage system [13].

The pediatric health care system in Italy is part of the NHS and is organized in the traditional 3 levels of intervention: primary care (first access), secondary care (hospital care), and tertiary care (specialized hospital care) [14]. Children are required to register with a family pediatrician until the age of 6 years, with the option to remain under pediatric or general practitioner care until the age of 14 years. At 14 years, patients transition to general practitioners, although extensions with pediatricians may be granted for chronic or disabling conditions. Both family pediatricians and general practitioners serve as primary care providers within the system [15].

Children's development is monitored by family pediatricians at predetermined time points, totaling 10 visits from birth to the age of 13 years, during well-child checkups. Additionally, family pediatricians are responsible for implementing specific health promotion initiatives, providing parents with valuable information to assess their children's growth and health status, diagnosing and treating illnesses, prescribing medications and diagnostic tests when necessary, referring patients to specialized facilities, and collaborating with other health care professionals.

Early Prevention System for ASD

ASD is a neurodevelopmental disorder characterized by social and communication impairments and restricted or repetitive behaviors [8,16]. It has a median global prevalence of 65 in 10,000 and 1 in 31 children according to Centers for Disease Control and Prevention data [17-19]. Specifically, a recent systematic review reported a prevalence of ASD between 1.70% and 1.85% in American children aged 4 and 8 years, respectively, whereas the prevalence in Europe ranges between 0.38% and 1.55% [20]. In Italy, a very recent nationwide population study estimated a prevalence of 13.4 (CI 11.3-16.0) per 1000 children aged 7 to 9 years, with a male-to-female ratio of 4.4:1 [21].

The presentation and onset patterns of ASD are variable. In many cases, the core diagnostic features of autism can be identified during toddlerhood [22]. Interventions are more beneficial when delivered earlier in life [23,24], and the American Academy of Pediatrics recommends routine autism screening at 18 to 24 months [16]. Furthermore, over the past 12 years, the American Academy of Pediatrics has reported an increase in the prevalence rates of ASD in children, an expanded understanding of potential risk factors, improved awareness of co-occurring medical conditions and genetic contributions to the etiology, and substantial growth in research supporting evidence-based interventions [16]. Recent studies have emphasized the importance of early ASD detection through screening [7,23,24]. Standardized, digital, high-fidelity ASD

XSL•FO
RenderX

screening during pediatric well-child visits facilitates the identification of children at higher risk of autism at a younger age, including those with more subtle clinical manifestations [7,25-27].

In Italy, guidelines on the diagnosis and treatment of ASD in children and adolescents have been approved and published [28,29], emphasizing the importance of early diagnosis.

Given the importance of telemedicine in the prevention and active surveillance of mental health, a screening system, the Web Italian Network for Autism Spectrum Disorder (Win4ASD) platform [30,31], has been developed in Lombardy, a highly populated Italian region with 10,012,054 inhabitants [32]. It provides early identification of individuals at risk of neurodevelopmental disorders using the Checklist for Autism in Toddlers (CHAT) [33]. Owing to computerized scoring, children identified as being at risk are immediately referred by the platform to the local NHS child psychiatric services for a fast-track diagnostic evaluation. The platform aims to create an integrated, interdisciplinary network, enabling early identification of ASD risk; facilitating rapid activation of diagnostic pathways; and ensuring care, habilitation, and rehabilitation based on priorities and intensities determined by the child's age group and functioning profile.

On the basis of these premises, this paper aims to illustrate how the Win4ASD platform has become part of the public NHS and its potential as a preventive medicine model for the general population through a retrospective observational study.

The primary objective of this study was to illustrate the integration of the Win4ASD platform within the NHS, detailing its functionalities and analyzing the associated screening pathways. Furthermore, this study aimed to evaluate the platform's contribution to the reorganization of public health care services, the strengthening of preventive strategies, and the enhancement of coordination among health care professionals and facilities. Finally, the clinical effectiveness of the platform was assessed in a subsample of users through a retrospective observational study with the objective of determining how the use of Win4ASD may facilitate the early diagnosis of ASD and improve screening outcomes.

Methods

Overview

The Win4ASD platform enables early screening for autism in toddlers (from 18 months; –2 months to +2 months) by filling out the CHAT [33].

The family pediatrician accesses the platform to register the patient and record project consent and then administers the screening. The CHAT screening test involves the administration of 14 items: 9 questions directed to the parents and 5 direct observations of the child's behavior, which are assessed in person by the pediatrician, who provides the corresponding responses for these observational items and enters them into the platform. Upon completion, the platform provides automated scoring in real time; stores the results; and, based on the screening outcome, recommends appropriate follow-up actions for the family pediatrician to undertake.

If the screening outcome is "generic" or "medium risk," the assessment can be repeated after a month. In the case of a "high risk" outcome at the first screening and a "generic risk," "medium risk," or "high risk" outcome at the repeated screening, the Win4ASD platform allows for the patient's referral to fast-track diagnostic evaluation. This customized platform is designed to connect primary care (family pediatricians) and mental health specialist services in safe ways. Specifically, 8 health care protection agencies in Lombardy provided a list of all NHS pediatricians, who then received invitations to use the platform. Within the platform, each family pediatrician has the option to refer high-risk children to the local specialized autism unit (SAU), a multidisciplinary and specialized team focused on the diagnosis and treatment of ASD, based on the gold standard. Specifically, each diagnostic assessment includes standardized procedures—comprising parent or caregiver interviews, direct child observations, and administration of validated diagnostic instruments (eg, the Autism Diagnostic Observation Schedule, Second Edition [ADOS-2]; Autism Diagnostic Interview–Revised [ADI-R]; Vineland Adaptive Behavior Scales, Second Edition; Leiter International Performance Scale; Griffiths Scales of Child Development, Third Edition; Psychoeducational Profile, Third Edition; Protocol for a Preventive Assessment of the Motor Coordination Disorder; *Primo Vocabolario del Bambino*; and Adaptive Behavior Assessment System–II)—conducted by clinical psychologists and child psychiatrists. The final diagnosis is formulated by the child psychiatrist, who integrates all clinical and psychometric data to develop the child's functional profile and therapeutic plan. While the diagnostic process is standardized across local units, individual teams may incorporate additional assessments according to clinical needs [34-43].

The different services communicate through the platform bulletin board, where pediatricians and mental health specialists can write observations and attach documents or reports. In addition, e-learning training courses and information on how to use the platform for family pediatricians are visible and easily accessible from the platform dashboard. The dashboard aggregates key information for real-time monitoring of screening data, patient data, and the activity trends of family pediatricians. The platform is described in detail in the study by Colombo et al [31]. For this study, we considered family pediatricians' activities and screenings performed within the platform from January 1, 2022, to September 30, 2024. The effectiveness of the screening process was studied further by analyzing data from a subsample for which outcome data were available.

The platform is now being used throughout the Lombardy region following the approval of the Autism Operational Plan (regional deliberations XI/5415 of October 25, 2021, and XII/3686 of December 20, 2024) [34].

Ethical Considerations

This study was conducted in accordance with the ethical principles for research involving human participants and with the guidelines of the Declaration of Helsinki. It was reviewed and approved by the Eugenio Medea Scientific Institute Ethics

XSL•FO
RenderX

and Committee (protocol code 352; September 22, 2016). Informed consent was obtained from the caregivers of all participating children before data collection through pediatric clinical practice. All data used in this research were pseudonymized to ensure participants' privacy and confidentiality. No financial or material compensation was provided for participation in the study. The dataset analyzed does not contain any information that could allow for the identification of individual participants.

Materials

Win4ASD Platform

We used the input from the Win4ASD platform, representing "performance data" together with data resulting from screening. Specifically, we calculated the number of family pediatricians per participating district in the Lombardy region and subsequently categorized them into 3 groups: those who were active and regularly used the platform, those who had never accessed it, and those who had logged in but did not engage with it.

Screening Tool

The screening tool is the Italian version of the CHAT [33,44]—a checklist for detecting children aged 18 months (–2 months to +2 months) who are at risk of autism. It was validated in England with a population of 16,000 children and was designed to be administered by primary health care services. The checklist consists of 2 sections: section A comprises 9 questions for parents asked by the family pediatricians, while section B comprises 5 items completed by the pediatrician based on direct observation of the child's behavior. All items are rated as "yes" or "no." This standardized screening tool is available to family pediatricians through the web application. It takes 5 to 10 minutes to administer and requires the family pediatricians' direct observation and involvement in the entire process, thereby increasing their compliance. The screening result can be "high risk" (failure of all critical items), "medium risk" (failure of a few critical items), "generic risk" (a risk of other neurodevelopmental disorders; failure of a few noncritical items), or "no risk." If the outcome of the first evaluation is "medium risk" or "generic risk," the family pediatrician has the option of repeating the assessment after 1 month (second evaluation). If the outcome is "high risk" either at the first or repeated screening or "medium risk" or "generic risk" at the repeated evaluation, the family pediatrician refers the patient for fast-track diagnostic evaluation through the platform.

Diagnostic Outcome

The diagnostic results are available for a subsample derived from a specific SAU in a regional district (Valle Olona). The diagnostic procedures described above (eg, clinical interviews and observations of the child) were carried out, and in particular, the 2 gold standard tools for autism were used: the ADOS-2 [35] and ADI-R [45]. The ADOS-2 is a semistructured, standardized assessment of autism symptomatology across communication, reciprocal social interaction abilities, play, and restricted and repetitive behaviors and yields diagnostic classifications or concern classification scores. The ADI-R is an investigator-based semistructured diagnostic interview that was designed to provide a developmental history framework for a lifetime differential diagnosis of pervasive developmental disorders and information about current functioning (over the previous 3 months) for individuals with a mental age of ≥2 years from early childhood to adult life. It is usually conducted with the parents or main carer of the child.

Statistical Analysis

Analyses were performed both on the total Win4ASD platform sample and on a specific subsample from the Valle Olona SAU. Descriptive statistics were used to summarize demographic characteristics and screening outcomes. Continuous variables were expressed as means and SDs, and categorical variables were expressed as absolute and relative frequencies. The 95% CIs were calculated for the main outcome proportions to assess the precision of estimates. Additionally, comparisons between male and female participants were conducted to explore potential Sex-related differences in screening outcomes. All statistical analyses were performed using SPSS (version 26; IBM Corp) [46], and a 2-sided significance level of .05 was considered.

Data Exclusion

We excluded the CHATs completed outside the age range of 16 to 22 months from the overall sample. In addition, as mentioned previously, we focused on a time frame that excluded all CHATs completed before January 2022 and after September 30, 2024. Of the total sample, 16 participants were excluded from the repeat assessment because they repeated the CHAT despite being classified as "no risk" at the first assessment.

Results

Win4ASD Platform Data

Pediatricians' Commitment

A total of 1253 family pediatricians were registered on the platform during the indicated time frame, of whom 909 (72.5%) were active and regularly used the platform, 188 (15%) had never logged on, and 163 (13%) had logged on but were not using the platform. The average number of screening CHATs performed by family pediatricians was 70.5 (SD 53.9; range 1-317). Through the platform's interactive dashboard, all information is presented and synthesized simply and intuitively.

The data shown in Figure 1 illustrate the pediatricians' use of the platform across different regional districts and their inclusion in the e-learning program (n=346). The bar chart in Figure 2 provides information about family pediatricians' engagement over time.

XSL•FO
RenderX

Figure 1. Family pediatricians' commitment to the web platform (total number of pediatricians, pediatricians who have taken e-learning courses, pediatricians divided by local ATS department). ATS: Italian name for the 8 local health authorities.

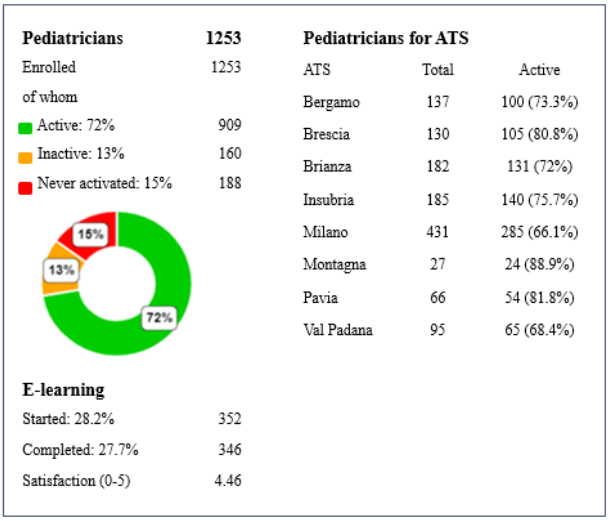

Figure 2. Family pediatricians' commitment over time: the x-axis shows the different months from January 2022 to September 2024, whereas the y-axis represents the number of pediatricians who registered on the platform and actively used it.

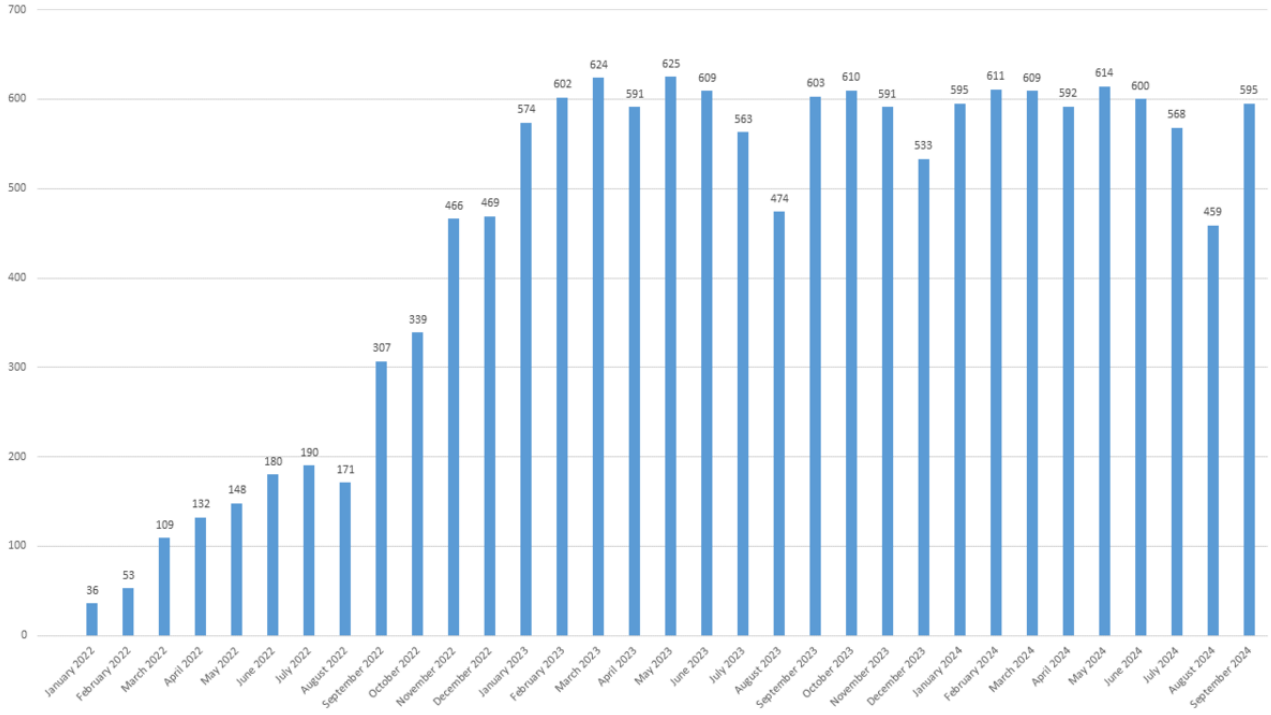

Early Screening System

A total of 58,419 toddlers from the general population in the Lombardy region were screened during the 16- or 18-month well-child checkup visit (n=29,980, 51.3% male and n=28,439, 48.7% female; mean age 17.32, SD 1.35; range 16-22 months).

From January 2022 to September 2024, the number of screenings gradually increased from 532 in the first quarter of 2022 to 5666 in the third quarter of 2024. The largest number of screenings was performed by family pediatricians in the first quarter of 2023, reaching 8775 CHATs (Figure 3).

A flowchart (Figure 4) shows the outcomes for patients with completed first and repeated screening CHATs, as well as referrals for a further diagnostic investigation.

XSL•FO
RenderX

Figure 3. Number of screenings by family pediatricians from January 2022 to September 2024 shown by quarterly trends.

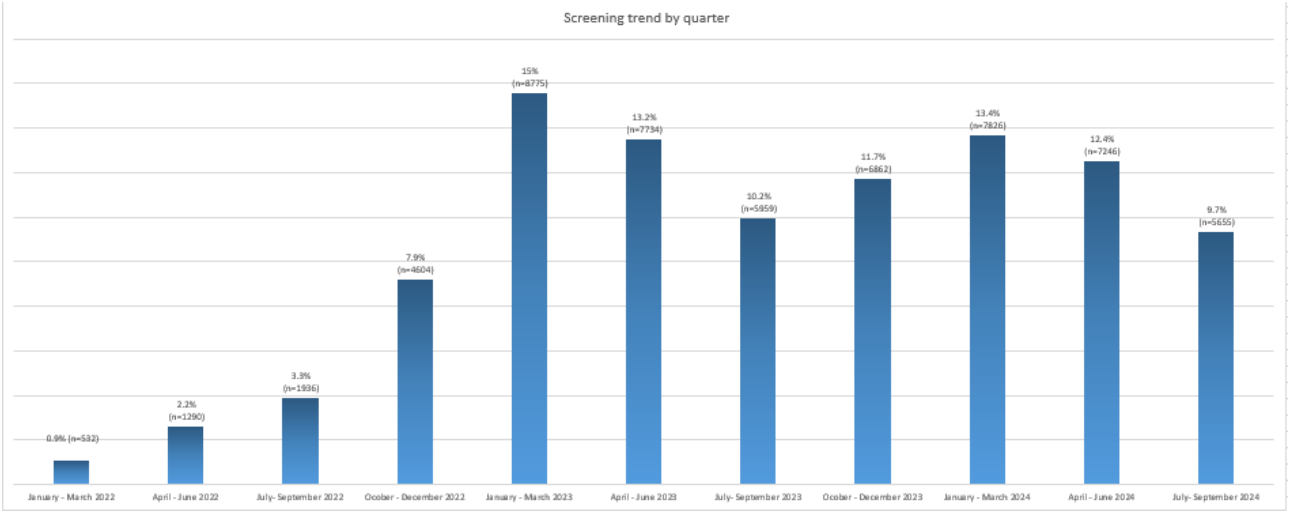

Figure 4. Screening outcomes: flowchart of screenings conducted by family pediatricians from January 2022 to September 2024. CHAT: Checklist for Autism in Toddlers.

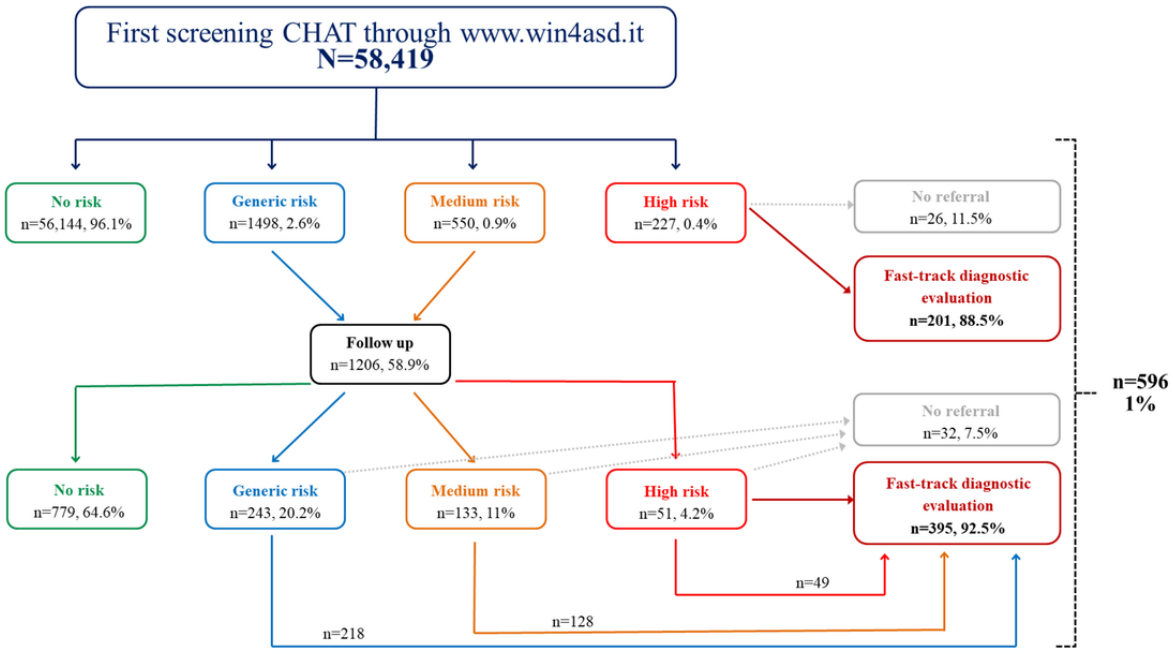

On the basis of the results, 96.1% (56,144/58,419) of the children were at "no risk" of a neurodevelopmental disorder at first screening, 2.6% (1498/58,419) had a "generic risk," 0.9% (550/58,419) were at "medium risk," and 0.4% (227/58,419) were at "high risk." Table 1 details the proportions of each screening outcome with corresponding 95% CIs in the overall screened population. A total of 0.3% (201/58,419) of the children screened in total (201/227, 88.5% of "high risk" children) were referred for fast-track diagnostic evaluation to the SAU. The detailed screening outcomes with a comparison of demographic variables is shown in Table 2.

Table 1. Proportions of each screening outcome with corresponding 95% CIs in the overall screened population (N=58,419).

Screening outcome	Participants, n (%; 95% CI)
No risk	56,144 (96.1; 95.95-96.26)
Generic risk	1498 (2.6; 2.44-2.69)
Medium risk	550 (0.9; 0.86-1.02)
High risk	227 (0.4; 0.34-0.44)

XSL•FO
RenderX

Table 2. Comparison of demographic and screening characteristics between male and female participants (N=58,419).

Variable	Male (n=29,980)	Female (n=28,439)	*P* value
Age at screening (months), mean (SD)	17.33 (1.36)	17.31 (1.33)	.03
Screening outcome, n (%)			<.001
No risk	28,552 (95.2)	27,592 (97)	
Generic risk	917 (3.1)	581 (2)	
Medium risk	581 (1.9)	192 (0.7)	
High risk	153 (0.5)	74 (0.3)	

In total, 11.5% (26/227) of the participants identified as "high risk" by the CHAT but who were not referred to the SAU were considered out of sample for several reasons. Of these 26 participants, 2 (7.7%) were excluded as they had moved away and were no longer receiving care at the appropriate SAU. In total, 23.1% (6/26) of the participants were already receiving care at other facilities. Regarding the remaining 69.2% (18/26), no information was available; lack of data may have been due to a family pediatrician's failure to use the platform correctly or a shared decision with the family not to proceed with referral of the child to the SAU.

Participants found to be at "generic risk" and "medium risk" (2048/58,419, 3.5%) repeated the screening after a month, as outlined in the protocol; among these 2048 participants, 1206 (58.9%) follow-ups were made.

Of the 1206 patients for whom screening was repeated (mean age at repeat screening 18.41 months), 779 (64.6%) were classified as "no risk," and 427 (35.4%) were identified as at risk. Of these 1206 patients, 243 (20.1%) were at generic risk, 133 (11%) were at medium risk, and 51 (4.2%) were at high risk. In total, 92.5% (395/427) of the patients at risk (395/58,419, 0.7% of the total sample) were referred for fast-track diagnostic evaluation to the SAU. In total, 1% (596/58,419) of the children in the total sample (399/596, 66.9% male; 197/596, 33.1% female) were referred for fast-track evaluation to the SAU.

In total, 7.5% (32/427) of the patients identified as "at risk" at the repeated assessment but who were not referred to the SAU were considered out of sample. No information was available. Again, the lack of data may have been due to a pediatrician's failure to use the platform or a shared decision with the family not to proceed with referral.

Assessing the Effectiveness of Screening

The performance of the CHAT screening test was evaluated by comparing the actual test results with the patients' true diagnostic status, as determined through a gold standard assessment. Among the children who underwent screening, diagnostic investigation outcomes were available for a subsample drawn from the Valle Olona SAU. This selection was driven by clinical and organizational considerations and, therefore, represents a convenience sampling [47]. Although this sampling approach may limit generalizability, a comparison between the Valle Olona SAU cohort and the rest of the population was performed using 2-tailed *t* tests and chi-square tests (Table 3). Given the large overall sample size (N=58,419), even minimal absolute differences reached statistical significance (eg, mean age 17.31, SD 1.34 months vs mean age 17.46, SD 1.42 months in the other SAUs and the Valle Olona SAU, respectively; *P*<.001).

Table 3. Comparison of demographic and screening characteristics among the overall screened population, the population at other centers, and the Valle Olona specialized autism unit (SAU; N=58,419).

Variable	Total sample	Valle Olona SAU (n=3599)	Other SAUs (n=54,820)	*P* value
Age at screening (months), mean (SD)	17.32 (1.35)	17.46 (1.42)	17.31 (1.34)	<.001
Sex, n (%)				.62
Male	29,980 (51.3)	1832 (50.9)	28,148 (51.3)	
Female	28,439 (48.7)	1767 (49.1)	26,672 (48.7)	
Screening outcome, n (%)				.02
No risk	56,144 (96.1)	3427 (95.2)	52,717 (96.2)	
Generic risk	1498 (2.6)	107 (3)	1391 (2.5)	
Medium risk	550 (0.9)	47 (1.3)	503 (0.9)	
High risk	227 (0.4)	18 (0.5)	209 (0.4)	

The Valle Olona SAU included 62 family pediatricians, of whom 56 (90%) regularly used the screening platform, 2 (3%) had never logged onto it, and 4 (6%) had logged in but did not use it actively. During the study period, 3599 CHAT screenings were performed by primary care pediatricians and specialists. Figure 5 shows the outcomes for patients who completed the CHAT screening at both assessments, including referrals for further diagnostic investigation. In total, 0.1% (2/3599) of the

XSL•FO
RenderX

patients were not referred because they had moved away and were no longer followed up on by the SAU. Overall, 1.5% (55/3599) of the total screened children (37/55, 67.3% male; 18/55, 32.7% female) were referred for fast-track diagnostic evaluation at the SAU. Of these 55 children, 27 (49.1%) were diagnosed with ASD and 20 (36.4%) were diagnosed with other neurodevelopmental disorders, including 3 (15%) language disorders (2 male and 1 female), 14 (70%) cases of intellectual disability (n=9, 64.3% male; n=5, 35.7% female), 1 (5%) genetic syndrome (male), and 2 (10%) unspecified neurodevelopmental disorders (1 male and 1 female). In total, 1.8% (1/55) of the children were still awaiting a diagnosis, whereas 12.7% (7/55) exhibited clinical signs but did not meet the diagnostic threshold; in these cases, targeted intervention programs were planned at home and at school, with ongoing 6- and 12-month clinical follow-ups.

Finally, a post hoc power analysis conducted using PASS 2024 (NCSS, LLC) [48] confirmed the adequacy of the Valle Olona SAU subsample (n=3599) to estimate the proportion of children identified as "at risk" (P=.01). At a 95% confidence level, the estimated interval ranged from 0.0070 to 0.0138, corresponding to a margin of error of –0.0034 to +0.0034, thus supporting the precision and robustness of the subsample-based estimates.

Figure 5. Effectiveness of screening: flowchart of screenings and diagnostic outcomes conducted by family pediatricians from January 2022 to September 2024 in the Valle Olona specialized autism unit subsample. ASD: autism spectrum disorder; CHAT: Checklist for Autism in Toddlers.

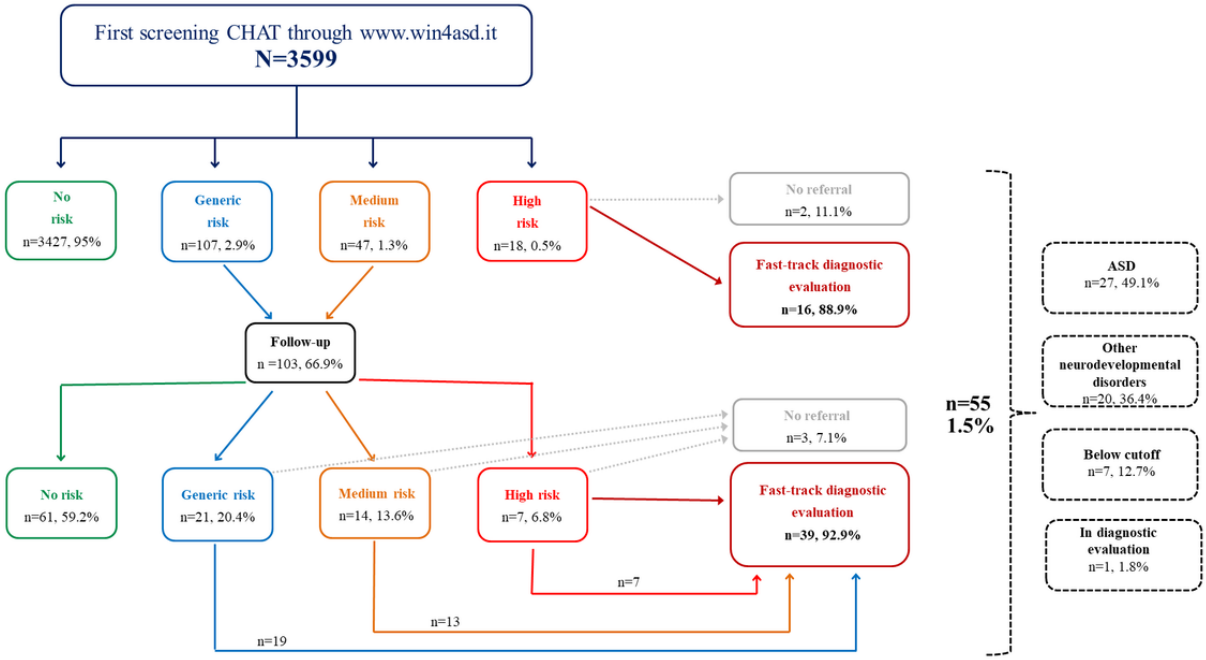

Discussion

Principal Findings

In this work, we presented the Win4ASD platform, an innovative digital tool designed to support early identification and diagnosis of children with ASD within the general population, as a model of preventive medicine. It holds significant potential for building a care network that supports children with ASD from early diagnosis to timely clinical intervention. We described how the platform was integrated into the regional NHS and demonstrated its clinical effectiveness for early ASD screening through a retrospective observational study. The results of this work highlight the political and institutional efforts behind the platform, as well as its clinical impact on the general population.

The first key finding is the commitment of family pediatricians to using the platform owing to the institutional agreement (a supplementary agreement for the Lombardy region) [49]. In total, 72.5% (909/1253) of all family pediatricians of the regional NHS were actively using the platform. This high level of participation makes ASD screening an integral part of the pediatric surveillance system through the systematic use of the CHAT. Furthermore, family pediatricians' engagement increased over time. We anticipate that continued use of the platform, along with ongoing dissemination and training through institutional channels, will further improve this percentage.

Additionally, our platform enabled the screening of 58,419 toddlers from the general population in Lombardy. According to Italian statistical data (Italian National Institute of Statistics), the estimated pediatric population aged 18 to 22 months between 2022 and 2024 was 69,474 [50]. On the basis of these figures, it can be estimated that 84.08% of children in Lombardy may have received well-child checkups and ASD screening. While a detailed analysis of screening adherence was not possible in this study, a large proportion of families in the region chose to use this free screening service. Moreover, compared with other similar studies that have screened for or diagnosed ASD through digital platforms or telemedicine, our study had a much larger sample size [5,25-27] and is innovative in terms of the network between family pediatricians and local care units for child psychiatry, which the platform succeeded in creating.

XSL•FO
RenderX

On the basis of the total sample (N=58,419), 1% (596/58,419) of the children were classified as "at risk" at the first or repeated screening and were referred for fast-track diagnostic evaluation. Although this proportion is numerically similar to global ASD prevalence estimates, it should be interpreted as the rate of screening positivity, not as a measure of confirmed ASD prevalence [18-20]. The results also show an overall higher number of male than female children. However, the statistically significant comparison between male and female children is probably due to the large sample size, as the absolute differences were quantitatively minimal.

Looking at the data on the effectiveness of screening from a specific SAU, 3599 screenings were performed, and 55 (1.5%) of these children were referred for fast-track diagnostic evaluation. This selection was driven by clinical and organizational considerations and represents a convenience sampling. From further comparative analyses with the general sample and from the post hoc analysis, it can be inferred that all differences between the 2 groups (general and Valle Olona SAU) were quantitatively negligible, supporting the overall homogeneity of the populations, the representativeness of the Valle Olona SAU subsample, and the precision and robustness of the subsample-based estimates.

In this SAU, 90% (56/62) of pediatricians actively used the platform, which led to an adherence rate higher than the regional average, making this SAU an effective model for the screening system. As mentioned above, the data for the total screenings and the percentage of children referred for further diagnosis in this SAU are in line with prevalence data [18-20]. Of the 55 patients referred for evaluation, 45 (81.8%) received a diagnosis, demonstrating the effectiveness of the screening system in identifying clinical signs. Specifically, of the 55 children, 27 (49.1%) received an ASD diagnosis and 20 (36.4%) were diagnosed with other neurodevelopmental disorders. This is in line with the findings of the authors of the CHAT, Baron-Cohen et al [51], who point out that the CHAT intercepts the risk of other neurodevelopmental disorders, especially in the medium or generic risk.

In short, the Win4ASD platform is a viable model for improving the NHS by enhancing continuity of care for children with ASD. Looking ahead, we anticipate even more efficient implementation of the platform, spreading this model further within the NHS and increasing screening coverage across different age groups of the child population. As part of upcoming developments, the Win4ASD platform is being enhanced with the integration of an additional screening tool, the Strengths and Difficulties Questionnaire (SDQ) [52]. The SDQ is a brief behavioral screening questionnaire for children aged 2 to 17 years. It comes in various versions to cater to the needs of researchers, clinicians, and educators. The version for children aged 2 to 4 years is now part of the platform. Similar to the CHAT, the platform allows pediatricians to complete all items of the SDQ, automatically score the screening, and facilitate seamless communication between pediatricians and the SAUs. The SDQ will be available on the platform for children aged 23 to 36 months, providing extended coverage for the active surveillance of neurodevelopmental disorders. The service is free of charge and serves as an additional screening layer within the NHS.

Limitations

This study has some limitations that must be considered when interpreting its findings. First, this work is not experimental research, which may introduce potential methodological bias. Specifically, we lacked statistical data on the number of children aged 18 to 22 months residing in Lombardy between 2022 and September 2024. Additionally, we did not have data on the number of children in Lombardy diagnosed with ASD or other neurodevelopmental disorders by public or private agencies to compare with the data from the platform.

In addition, we acknowledge that the use of data from a single SAU, along with the lack of information on children at generic, medium, or high risk who, for clinical reasons or parental choice, did not complete the screening and diagnostic process monitored by the platform, represents a limitation regarding the generalizability of our findings. To address this, we plan to expand the outcome dataset by incorporating cases from other SAUs, which will provide a broader and more representative perspective. Nevertheless, despite the potential of the platform, the emerging network, and the high number of screenings carried out, our results cannot yet be generalized to the wider regional context.

Furthermore, as illustrated in the flowcharts (Figures 4 and 5), some children identified as being at risk of neurodevelopmental disorders through the CHAT were not referred to SAU services for diagnostic assessment. This may be due to a variety of clinical and nonclinical factors, for instance, parental decisions regarding the continuation of diagnostic evaluation or a family's relocation to another region.

Conclusions

Owing to the political and institutional support that the platform has received over the years, it has been possible to make improvements that have led to an increasingly widespread dissemination of the screening system. Early detection of risk and timely intervention can prevent the progressive development of behavioral atypicalities; have a significant impact on the quality of life of the child and their family; and, finally, allow for a reduction in health care costs and a consequent improvement in the organization of the health care system.

Acknowledgments

The authors acknowledge all partners and collaborators of the project. For this paper, the generative AI tool ChatGPT (OpenAI [53]) was used to assist with language revision.

XSL•FO
RenderX

Funding

This work has been supported by the Italian Ministry of Health (Ricerca Corrente 2025-2026) and the *Fondo per la cura dei soggetti con disturbo dello spettro autistico* of the Italian Ministry of Health through the Regione Lombardia *Attivare Una risposta Territoriale INclusiva e Continuativa per le persone con Autismo* (AUTINCA) Project (deliberation XII/277; May 15, 2023).

Authors' Contributions

PC, NB, IL, LF, and MM designed the study. NB, ER, SBC, and NV conducted literature searches. NB, NV, SBC, MF, EM, and VM collected the data. NB, PC, and ER analyzed the data. NB and ER wrote the first draft of the manuscript. MM, IL, LF, and PC critically reviewed the manuscript. All authors have approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Telemedicina. Ministero della Salute. URL: https://tinyurl.com/523uxb4a [accessed 2025-01-18]
2. Telehealth: defining 21st century care. American Telemedicine Association. URL: https://www.americantelemed.org/resource/why-telemedicine/ [accessed 2025-02-21]
3. Galmiche S, Rahbe E, Fontanet A, Dinh A, Bénézit F, Lescure FX, et al. Implementation of a self-triage web application for suspected COVID-19 and its impact on emergency call centers: observational study. J Med Internet Res 2020 Nov 23;22(11):e22924 [FREE Full text] [doi: 10.2196/22924] [Medline: 33147165]
4. Basch E, Deal AM, Dueck AC, Scher HI, Kris MG, Hudis C, et al. Overall survival results of a trial assessing patient-reported outcomes for symptom monitoring during routine cancer treatment. JAMA 2017 Jul 11;318(2):197-198 [FREE Full text] [doi: 10.1001/jama.2017.7156] [Medline: 28586821]
5. Valentine AZ, Hall SS, Young E, Brown BJ, Groom MJ, Hollis C, et al. Implementation of telehealth services to assess, monitor, and treat neurodevelopmental disorders: systematic review. J Med Internet Res 2021 Jan 20;23(1):e22619 [FREE Full text] [doi: 10.2196/22619] [Medline: 33326409]
6. Holmes EA, O'Connor RC, Perry VH, Tracey I, Wessely S, Arseneault L, et al. Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. Lancet Psychiatry 2020 Jun;7(6):547-560 [FREE Full text] [doi: 10.1016/S2215-0366(20)30168-1] [Medline: 32304649]
7. Liu M, Ma Z. A systematic review of telehealth screening, assessment, and diagnosis of autism spectrum disorder. Child Adolesc Psychiatry Ment Health 2022 Oct 08;16(1):79. [doi: 10.1186/s13034-022-00514-6] [Medline: 36209100]
8. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5). Washington, DC: American Psychiatric Association; 2013.
9. Global report on children with developmental disabilities: from the margins to the mainstream. UNICEF. 2023 Sep 15. URL: https://www.unicef.org/documents/global-report-children-developmental-disabilities [accessed 2025-02-21]
10. World report on disability. World Health Organization. 2011 Dec 14. URL: https://www.who.int/publications/i/item/9789241564182 [accessed 2025-02-21]
11. Cioni G, Inguaggiato E, Sgandurra G. Early intervention in neurodevelopmental disorders: underlying neural mechanisms. Dev Med Child Neurol 2016 Mar 29;58 Suppl 4(S4):61-66 [FREE Full text] [doi: 10.1111/dmcn.13050] [Medline: 27027609]
12. Donatini A. The Italian health care system. Emilia Romagna Regional Health Authority, IT: International profiles of health care systems; 2015. URL: https://www.issuelab.org/resources/25100/25100.pdf#page=97 [accessed 2026-02-05]
13. France G, Taroni F, Donatini A. The Italian health-care system. Health Econ 2005 Sep;14(S1):S187-S202 [FREE Full text] [doi: 10.1002/hec.1035]
14. Corsello G, Ferrara P, Chiamenti G, Nigri L, Campanozzi A, Pettoello-Mantovani M. The child health care system in Italy. J Pediatr 2016 Oct;177S:S116-S126. [doi: 10.1016/j.jpeds.2016.04.048] [Medline: 27666260]
15. Crescere in salute: rete pediatrica. Agenzia di Tutela della Salute di Bergamo. URL: https://ats-bg.it/crescere-salute-rete-pediatrica-provincia-bergamo [accessed 2025-02-21]
16. Hyman SL, Levy SE, Myers SM, Council on Children With Disabilities, Section on Developmental and Behavioral Pediatrics. Identification, evaluation, and management of children with autism spectrum disorder. Pediatrics 2020 Jan;145(1):e20193447. [doi: 10.1542/peds.2019-3447] [Medline: 31843864]
17. Zeidan J, Fombonne E, Scorah J, Ibrahim A, Durkin MS, Saxena S, et al. Global prevalence of autism: a systematic review update. Autism Res 2022 May;15(5):778-790 [FREE Full text] [doi: 10.1002/aur.2696] [Medline: 35238171]
18. Maenner MJ, Warren Z, Williams AR, Amoakohene E, Bakian AV, Bilder DA, et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2020. MMWR Surveill Summ 2023 Mar 24;72(2):1-14 [FREE Full text] [doi: 10.15585/mmwr.ss7202a1] [Medline: 36952288]
19. Data and statistics on autism spectrum disorder. Centers for Disease Control and Prevention. 2025 May 27. URL: https://www.cdc.gov/autism/data-research/index.html [accessed 2025-02-21]

XSL•FO
RenderX

20. Bougeard C, Picarel-Blanchot F, Schmid R, Campbell R, Buitelaar J. Prevalence of autism spectrum disorder and co-morbidities in children and adolescents: a systematic literature review. Front Psychiatry 2021;12:744709 [FREE Full text] [doi: 10.3389/fpsyt.2021.744709] [Medline: 34777048]
21. Scattoni ML, Fatta LM, Micai M, Sali ME, Bellomo M, Salvitti T, et al. Autism spectrum disorder prevalence in Italy: a nationwide study promoted by the Ministry of Health. Child Adolesc Psychiatry Ment Health 2023 Oct 28;17(1):125 [FREE Full text] [doi: 10.1186/s13034-023-00673-0] [Medline: 37898807]
22. Lord C, Brugha TS, Charman T, Cusack J, Dumas G, Frazier T, et al. Autism spectrum disorder. Nat Rev Dis Primers 2020 Jan 16;6(1):5 [FREE Full text] [doi: 10.1038/s41572-019-0138-4] [Medline: 31949163]
23. Guthrie W, Wetherby AM, Woods J, Schatschneider C, Holland RD, Morgan L, et al. The earlier the better: an RCT of treatment timing effects for toddlers on the autism spectrum. Autism 2023 Mar 15;27(8):13623613231159153 [FREE Full text] [doi: 10.1177/13623613231159153] [Medline: 36922406]
24. Dimian AF, Symons FJ, Wolff JJ. Delay to early intensive behavioral intervention and educational outcomes for a Medicaid-enrolled cohort of children with autism. J Autism Dev Disord 2021 Apr;51(4):1054-1066. [doi: 10.1007/s10803-020-04586-1] [Medline: 32642958]
25. Vivanti G, Algur Y, Ryan V, McClure LA, Fein D, Stahmer AC, et al. The impact of using standardized autism screening on referral to specialist evaluation for young children on the autism spectrum: a cluster-randomized controlled trial. J Am Acad Child Adolesc Psychiatry 2025 Jun;64(6):686-698 [FREE Full text] [doi: 10.1016/j.jaac.2024.08.502] [Medline: 39454668]
26. Abbas H, Garberson F, Glover E, Wall DP. Machine learning approach for early detection of autism by combining questionnaire and home video screening. J Am Med Inform Assoc 2018 Aug 01;25(8):1000-1007 [FREE Full text] [doi: 10.1093/jamia/ocy039] [Medline: 29741630]
27. Abbas H, Garberson F, Liu-Mayo S, Glover E, Wall DP. Multi-modular AI approach to streamline autism diagnosis in young children. Sci Rep 2020 Mar 19;10(1):5014 [FREE Full text] [doi: 10.1038/s41598-020-61213-w] [Medline: 32193406]
28. Raccomandazioni della Linea Guida per la diagnosi e il trattamento di bambini e adolescenti con disturbo dello spettro autistico. Istituto Superiore di Sanità. 2021 Nov 30. URL: https://www.iss.it/-/raccomandazioni-lg-diagnosi-trattamento-di-bambini-adolescenti-con-asd [accessed 2025-02-21]
29. Morgano GP, Fulceri F, Nardocci F, Barbui C, Ostuzzi G, Papola D, et al. Introduction and methods of the evidence-based guidelines for the diagnosis and management of autism spectrum disorder by the Italian National Institute of Health. Health Qual Life Outcomes 2020 Mar 26;18(1):81 [FREE Full text] [doi: 10.1186/s12955-020-01320-4] [Medline: 32216833]
30. Win4ASD. URL: https://www.win4asd.it [accessed 2025-02-21]
31. Colombo P, Buo N, Busti Ceccarelli S, Molteni M. Integrating a new online platform in primary care for early detection, referral, and intervention in autism spectrum disorder: the first Italian pivotal project. Brain Sci 2022 Feb 12;12(2):256 [FREE Full text] [doi: 10.3390/brainsci12020256] [Medline: 35204020]
32. Dati sulla popolazione e l'occupazione. ISTAT. URL: https://esploradati.istat.it/databrowser/#/it/dw/categories [accessed 2025-02-11]
33. Baron-Cohen S, Allen J, Gillberg C. Can autism be detected at 18 months? The needle, the haystack, and the CHAT. Br J Psychiatry 1992 Dec 02;161(6):839-843. [doi: 10.1192/bjp.161.6.839] [Medline: 1483172]
34. Piano operativo regionale autismo. Regione Lombardia. 2024. URL: https://tinyurl.com/5n6bhue9 [accessed 2025-02-11]
35. Lord C, Rutter M, Dilavore P, Risi S, Gotham K, Bishop S. Autism Diagnostic Observation Schedule, Second Edition. Torrence, CA: Western Psychological Services; 2012.
36. Lanfranchi S, Rea M, Ferri R, Vianello R. Studio di validazione e standardizzazione italiana delle Griffiths III. Catalogo dei Prodotti Della Ricerca. 2019. URL: https://iris.uniroma1.it/handle/11573/1355472 [accessed 2025-02-21]
37. Roid GH, Miller LJ, Pomplun M, Koch C. Leiter International Performance Scale, Third Edition. Wood Dale, IL: Stoelting Company; 2013.
38. Schopler E, Lansing M, Reichler R, Marcus L. Psychoeducational Profile Third Edition (PEP‐3). Austin, TX: Pro-Ed Inc; 2004.
39. Schopler E, Lansing MD, Reichler RJ, Marcus LM. Pep 3 Examiner's Manual With Tables. Austin, TX: Pro-Ed Inc; 2005.
40. Sparrow SS, Cicchetti DV, Balla DA. Vineland Adaptive Behavior Scales, Second Edition (Vineland-II). London, UK: Pearson; 2005.
41. Harrison PL, Oakland T. ABAS-II assessment methods. In: Oakland T, Harrison PL, editors. Adaptive Behavior Assessment System-II. Cambridge, UK: Academic Press; 2008:37-49.
42. Sabbadini L. APCM-2 Abilità Prassiche e della Coordinazione Motoria – 2a Edizione. Göttingen, Germany: Hogrefe; 2015.
43. Caselli MC, Bello A, Rinaldi P, Stefanini S, Pasqualetti P. Il Primo Vocabolario Del Bambino: Gesti, Parole e Frasi. Milan, Italy: Franco Angeli; 2018.
44. CHAT – checklist for autism in toddler. Autism Research Centre. URL: https://docs.autismresearchcentre.com/tests/CHAT_Italian.pdf [accessed 2025-02-18]
45. Lord C, Rutter M, Le Couteur A. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. J Autism Dev Disord 1994 Oct;24(5):659-685. [doi: 10.1007/BF02172145] [Medline: 7814313]

XSL•FO
RenderX

46. IBM SPSS statistics for Windows; version 26.0. IBM Corp. 2019. URL: https://www.ibm.com/products/spss-statistics [accessed 2026-02-17]
47. Stratton S. Population research: convenience sampling strategies. Prehosp Disaster Med 2021 Jul 21;36(4):373-374 [FREE Full text] [doi: 10.1017/s1049023x21000649]
48. PASS 2026. NCSS Statistical Software. URL: https://www.ncss.com/ [accessed 2025-11-02]
49. Accordo integrativo regionale pediatria di famiglia. Regione Lombardia. 2025. URL: https://tinyurl.com/fwf77s44 [accessed 2025-02-18]
50. Italia, regioni, province. ISTAT. URL: https://esploradati.istat.it/databrowser/#/it/dw/categories/IT1,POP,1.0/POP_POPULATION/DCIS_POPRES1/IT1,22_289_DF_DCIS_POPRES1_1,1.0 [accessed 2025-02-03]
51. Baron-Cohen S, Wheelwright S, Cox A, Baird G, Charman T, Swettenham J, et al. Early identification of autism by the CHecklist for Autism in Toddlers (CHAT). J R Soc Med 2000 Oct 01;93(10):521-525 [FREE Full text] [doi: 10.1177/014107680009301007] [Medline: 11064690]
52. Goodman R. The Strengths and Difficulties Questionnaire: a research note. J Child Psychol Psychiatry 1997 Jul;38(5):581-586. [doi: 10.1111/j.1469-7610.1997.tb01545.x] [Medline: 9255702]
53. ChatGPT. OpenAI. URL: https://chatgpt.com/c/698b0793-c320-8389-b4d0-4f2f40f6f63d

Abbreviations

ADI-R: Autism Diagnostic Interview–Revised
ADOS-2: Autism Diagnostic Observation Schedule, Second Edition
ASD: autism spectrum disorderCHAT: Checklist for Autism in ToddlersNHS: National Health Service
SAU: specialized autism unit
SDQ: Strengths and Difficulties Questionnaire
Win4ASD: Web Italian Network for Autism Spectrum Disorder

Edited by A Stone; submitted 21.Mar.2025; peer-reviewed by RS Kucukardali, HC Koh; comments to author 09.Sep.2025; revised version received 21.Jan.2026; accepted 21.Jan.2026; published 24.Feb.2026.

Please cite as:
Buo N, Rosi E, Busti Ceccarelli S, Ferrario M, Maiorca V, Morandi E, Viganó N, Limosani I, Falcone L, Molteni M, Colombo P Rising to the Challenge of Early Screening in Primary Health Care Through the Web Italian Network for Autism Spectrum Disorder (Win4ASD) in the Pediatric Population: Retrospective Observational Study
J Med Internet Res 2026;28:e74302
URL: https://www.jmir.org/2026/1/e74302
doi:10.2196/74302
PMID:

©Noemi Buo, Eleonora Rosi, Silvia Busti Ceccarelli, Mariarosa Ferrario, Valerio Maiorca, Erika Morandi, Nicole Viganó, Ivan Limosani, Laura Falcone, Massimo Molteni, Paola Colombo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Effect of Veteran Race and Socioeconomic Status on Enrollment in Remote Patient Monitoring for Hypertension: Retrospective Observational Cross-Sectional Study

Hannah Robin Friedman[1], PhD; Hillary J Mull[1,2], MPP, PhD; Jenice Ria Guzman-Clark[3], GNP-BC, PhD; Daniel J Sturgeon[4,5], MS; Marva V Foster[1,2], RN, PhD

[1]Center for Health Optimization and Implementation Research, VA Boston Healthcare System, Boston, MA, United States

[2]Chobanian and Avedisian School of Medicine, Boston University, Boston, MA, United States

[3]Southern Arizona VA Health Care System, Tucson, AZ, United States

[4]Stanford Medicine, Palo Alto, CA, United States

[5]Health Economics Resource Center, VA Palo Alto Health Care System, Palo Alto, CA, United States

Corresponding Author:
Hannah Robin Friedman, PhD
Center for Health Optimization and Implementation Research
VA Boston Healthcare System
Lafayette City Center
2 Ave de Lafayette
Boston, MA, 02111
United States
Phone: 1 617 232 9500
Email: hannah.friedman2@va.gov

Abstract

Background: Black veterans and veterans from lower socioeconomic backgrounds are more likely to have uncontrolled hypertension. One potential explanatory factor is reduced access to specific treatments that result in improved chronic disease management. In the Veterans Health Administration (VHA), veterans with hypertension may enroll in a remote patient monitoring (RPM) program, which consists of patient education, daily home blood pressure (BP) monitoring, health coaching, and case management. Barriers for socioeconomically disadvantaged patients may exist for similar programs in other health systems; however, the VHA is an integrated health care system, and these barriers may differ for veteran populations.

Objective: The objective of this study was to assess the relationship between veteran race and neighborhood socioeconomic status and the likelihood of enrolling in the VHA RPM program.

Methods: The study sample included VHA-enrolled veterans with a diagnosis of hypertension (average BP >130/80 mm Hg on ≥2 BP readings) between fiscal years 2020 and 2023. We ran random-effects logistic regression models to assess the relationship between veteran race and Area Deprivation Index and RPM enrollment each year, controlling for potential demographic and clinical confounders. For sensitivity analysis, we limited our sample to veterans with stage 2 hypertension (BP >140/90 mm Hg) and on antihypertensive medication.

Results: Overall use of RPM was low, with only 4.1% (56,553/1,390,995; 95% CI 4%-4.1%) of veterans being enrolled in RPM. Black veterans, who represented 26.6% (n=35,096) of all veterans, were more likely (odds ratio [OR] 1.65, 95% CI 1.59-1.70) to enroll in RPM compared to White veterans. Asian American or Pacific Islander veterans were less likely to enroll (OR 0.83, 95% CI 0.74-0.94). We found no meaningful association between Area Deprivation Index and RPM enrollment (OR 1.00, 95% CI 0.99-1.00). When limiting our sample to those with stage or grade 2 hypertension, we found a similar association (OR 1.61, 95% CI 1.50-1.72) between Black race and RPM enrollment but no significant association with Asian or Pacific Islander race (OR 1.02, 95% CI 0.80-1.29).

Conclusions: Prior research on RPM in veterans has examined duration or outcomes of RPM enrollment but not the probability of initial enrollment. We found higher enrollment rates in the VHA RPM program among Black veterans but slightly lower enrollment among Asian American or Pacific Islander veterans. Higher enrollment among Black veterans and among those with higher comorbidity burden suggests that the VHA RPM program is successfully reaching those who could most benefit, despite

XSL•FO
RenderX

low overall enrollment. Given the low enrollment in RPM, future research should focus on improving uptake among veterans who could additionally benefit from the program. Non-VHA systems, particularly those serving low-income or socioeconomically disadvantaged areas, should explore subsidized or free RPM programs for eligible patients similar to the VHA's no-cost model for veterans.

(J Med Internet Res 2026;28:e78423) doi:10.2196/78423

KEYWORDS

hypertension; veterans' health; remote patient monitoring; telehealth; health equity

Introduction

Remote patient monitoring (RPM) is a form of telehealth that allows clinicians to track patient vital signs and health metrics outside traditional health care settings [1]. This can include using wireless-enabled blood pressure (BP) cuffs, thermometers, pulse oximeters, scales, and glucometers to transmit vital signs to a nurse from the patient's home on a regular basis. Integrating RPM into chronic disease management can potentially improve patient quality of life and mitigate health care costs by enabling the timely detection of health deterioration, thereby reducing unnecessary emergency department visits and hospital admissions [2,3]. The Veterans Health Administration (VHA) is an integrated health care system that provides health care for more than 9 million of the 18 million veterans in the United States [4]. As a comprehensive health care system, the VHA offers RPM services at no cost to veterans and has existed since the early 2000s serving veterans with a variety of chronic conditions [5]. The VHA RPM model is unique due to its nationwide coverage, centralized funding, and focus on a specific patient population. The essence of RPM, as implemented in the VHA, is ongoing assessment, monitoring, patient education, health coaching, and nurse case management of veterans in their place of residence. It also includes the provision of appropriate information to the veterans' primary care team or specialty care team, as well as the health care system, to enable timely care [5].

Veterans targeted for RPM are those who have high health service use (eg, frequent emergency room visits) due to chronic diseases, who have uncontrolled chronic diseases, and who agree to participate in daily submission of health data for monitoring [6,7]. Referrals can be made by any VHA-licensed independent clinician (physician, nurse practitioner, or physician assistant) or other clinicians when following an approved local protocol. Once enrolled, veterans are expected to submit vital sign and other relevant biometric data every day and respond to daily disease-specific questions.

Hypertension, or high BP, is one of the most common chronic medical conditions managed by the VHA RPM program [8]. Hypertension is a major contributor to cardiovascular morbidity and mortality in the United States and remains one of the most important modifiable cardiovascular risk factors [9]. Prevalence data indicate that 47% of adults in the United States have hypertension, which is equivalent to an estimated 122 million adults [10]. More than half of these individuals with hypertension have BP that is inadequately controlled [11]. There are racial and ethnic differences in hypertension prevalence and treatment in the United States, with more Black, Hispanic, and Asian adults having poor control of their BP compared to White adults despite Black adults having a higher prevalence of hypertension [12]. Hypertension is the most common chronic condition among patients in the VHA. Recent research has found that 87% of veterans have high BP and, of those, most (66%) are considered to have uncontrolled hypertension [13]. Effective control of high BP is of key importance for reducing the risk of hypertension-related cardiovascular disease complications.

However, several factors have been found to impact hypertension control, key among them being patient residence in socioeconomically disadvantaged neighborhoods [14-16]. In the VHA, Black veterans have the poorest control rates for hypertension compared to other ethnic backgrounds even after adjusting for access to care and socioeconomic status [17]. Black veterans also have considerably higher rates of cardiovascular-related hospitalizations, an indicator of poor hypertension control [18]. RPM has been posited as a potential intervention to decrease this racial and ethnic disparity in hypertension control and has been shown to improve hypertension control in patients of a large health care organization [3]; however, a small pilot study that compared a sample of primary care patients who mostly comprised minority groups or groups with low socioeconomic status who used RPM to a control group who did not use RPM found no difference in BP control after 90 days, but found that RPM enrollment was associated with decrease in all-cause clinic visits [19].

There is no singular socioeconomic determinant implicated in disadvantaged neighborhoods. Multidimensional factors such as economic (eg, low income and blue-collar jobs), social (eg, low educational level, nonvoter status, and high unemployment), and physical (eg, graffiti and low-rent housing) conditions, among others, can influence a neighborhood [20,21]. These neighborhood-level characteristics have been increasingly recognized as playing a critical role in health [22]. The use of RPM has been suggested as a potential solution to overcome the geographical limitations of health care services [23,24]. However, a review of existing literature on non-VHA populations has shown that some studies indicate the expected directionality (ie, greater neighborhood socioeconomic disadvantage is associated with decreased use of RPM), whereas others report mixed and null findings [25-28]. The pathway between neighborhood deprivation and RPM use for hypertension management in the VHA is poorly understood. Previous VHA studies have examined adherence to RPM and enrollment rates in RPM but have not examined the role of neighborhood deprivation on RPM enrollment [8,29-31]. Non-VHA studies have identified barriers to RPM among Black patients, including lower overall participation in clinical trials

XSL•FO
RenderX

and reduced precision and accuracy of wearable devices on people with darker skin tones [32,33]. The overall evidence points to the fact that living in a socioeconomically disadvantaged neighborhood may impact hypertension and possibly RPM use in non-VHA populations.

What is not known is whether neighborhood deprivation impacts enrollment in RPM among veterans; such knowledge will identify populations for whom appropriate interventions can be implemented. Therefore, the purpose of this study was to investigate whether veteran race or residence in a socioeconomically disadvantaged neighborhood was associated with the probability that a patient with hypertension will enroll in the VHA RPM program.

Methods

We used the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines to ensure completeness and accuracy when reporting our methods and results [34].

Population and Sample

This retrospective cross-sectional study included veterans with a diagnosis of hypertension between October 1, 2019, and September 30, 2023. We used data from the VHA's Corporate Data Warehouse to identify all covariates and outcomes for this study. Our initial sample consisted of all veterans with an *International Classification of Diseases, 10th Revision*, diagnosis code for hypertension on or before October 1, 2022 [35]. We excluded observations in which the veteran was not enrolled in the VHA at the start of the fiscal year (FY), did not have a diagnosis of hypertension before the start of the FY, or died before the end of the FY. We then limited our sample to veterans with uncontrolled hypertension.

Definition of Uncontrolled Hypertension

The American College of Cardiology and American Heart Association have defined hypertension as a BP of at least 130/80 mm Hg over a minimum of 2 BP readings [36,37]. To ensure that our sample consisted of those with hypertension in accordance with clinical practice guidelines, we limited the study sample to veterans with ≥2 days of recorded BP readings in the prior year and with an average BP reading of at least 130/80 mm Hg across all readings. For sensitivity analysis, we further limited the sample to those with an average BP of at least 140/90 mm Hg (hypertension stage or grade 2) despite being prescribed antihypertensive medications [37].

Measures

Outcome

Our outcome was any enrollment in the VHA's hypertension RPM program each year. Veterans were considered to have enrolled in RPM if any encounter had an RPM VHA stop code (683, 684, 685, or 686) included as a primary or secondary stop code in electronic health record data. VHA stop codes are administrative codes used to categorize and track outpatient clinical services and have been used in previous studies [38,39]. We did not differentiate between new enrollment or continued use of RPM from patients enrolled in RPM during the prior year. We also did not consider the duration of enrollment as the focus of this study was veteran enrollment in the RPM program.

Primary Independent Variables

For enrollment-adjusted regression models, our primary independent variables comprised veteran race (White; Black; Asian American, Native Hawaiian, Pacific Islander; American Indian, Alaska Native; and other race, unknown, or multiracial) and the Area Deprivation Index (ADI) of the census block area of the veteran's primary address. When possible, veteran race in the Corporate Data Warehouse is self-reported, and prior research has found high agreement between Corporate Data Warehouse race and self-reported race and ethnicity [40]. The ADI is an established composite measure of an area's overall socioeconomic vulnerability and includes measures of median home value, poverty rate, and education, among others. For analysis, the unit of change was a 10-percentile change in ADI national ranking.

Control Variables

We controlled for patients' overall health, specific comorbidities, age, sex, rurality, VHA priority group, and distance to the nearest VHA facility. Veterans' overall health was assessed using the Elixhauser Comorbidity Index, and we controlled separately for a diagnosis of diabetes mellitus due to complications associated with comorbid diabetes and hypertension, as well as the presence of a separate RPM program for veterans with diabetes. We also controlled for the type of hypertension (uncomplicated or with complications) as defined by the Elixhauser Comorbidity Index [41]. We described veterans' rurality using the VHA classification of urban, rural, and highly rural, which is originally based on Rural-Urban Commuting Area codes. A VHA priority group is a category assigned to veterans to determine the amount, if any, they may have to pay for health care. Veterans are classified into priority groups 1 to 8. We aggregated priority groups 1 to 3 (service-connected disability), priority group 4 (non–service-connected catastrophic disability or receiving VHA aid benefits), priority group 5 (low income or Medicaid eligible), and priority groups 6 to 8 (service history and copayments required for some services). All variables reflected the status of the patient at the beginning of the FY. For covariates for which data were missing (rurality, priority group, ADI, and comorbidities), we used multiple imputation to generate replacement values. Data processing and analysis were conducted using SAS Enterprise Guide (SAS Institute) and Stata (version 18; StataCorp).

Analysis

For descriptive analysis, we assessed the veteran characteristics by RPM enrollment; we used chi-square tests (for categorical variables) and Wilcoxon rank sum tests (for continuous variables) to assess the significance of differences between White veterans, Black veterans, and veterans of other races. We ran random-effects logistic regression models to assess the likelihood of a veteran being enrolled in RPM each year from FYs 2020-2023, with random intercepts for each individual veteran. We separated the cohort into person-years with 1 observation per patient per year for a maximum of 4

XSL•FO
RenderX

observations per patient, although not all patients were eligible for inclusion in all 4 years. Separating patients into person-years allowed for more accurate assessments of patients' potential confounders, such as comorbidities, age, location, and BP control, all of which could change over the course of 4 years. Using person-year as the unit of analysis also allowed us to account for the COVID-19 pandemic, which began in FY 2020 and substantially impacted health care use patterns, including RPM participation. We used multiple tests and metrics to assess model goodness of fit using log-likelihood ratio tests and also tested for potential collinearity using variance inflation factors.

Sensitivity Analysis

For sensitivity analysis, we limited our sample to those with stage or grade 2 hypertension (average BP reading>140/90 mm Hg) and presence of at least one hypertension medication in the 6 months before the start of the FY to focus on those who might be most in need of closer monitoring.

Ethical Considerations

This study was approved by the US Department of Veterans Affairs (VA) Boston Institutional Review Board (1765472-5; approval date: January 18, 2024). As this study involved only retrospective analysis of a large observational dataset derived from electronic health record data where risk was deemed minimal to study participants, we received a waiver for protected health information and a waiver of informed consent from the institutional review board. All identifiable study data were stored and analyzed on a secure server operated by the VHA Informatics and Computing Infrastructure, and no identifiable data were ever moved from that server. No identification of individual participants is included in this paper. As this study consisted of analysis of a secondary dataset, no participants received compensation.

Results

Overview

Among those with complete covariate data and who met the eligibility criteria (Figure 1), we identified 2,411,787 total person-years during FYs 2019-2023 among 1,390,995 unique veterans.

Figure 1. Veterans with hypertension eligible for inclusion in study sample between fiscal years 2019 and 2023.

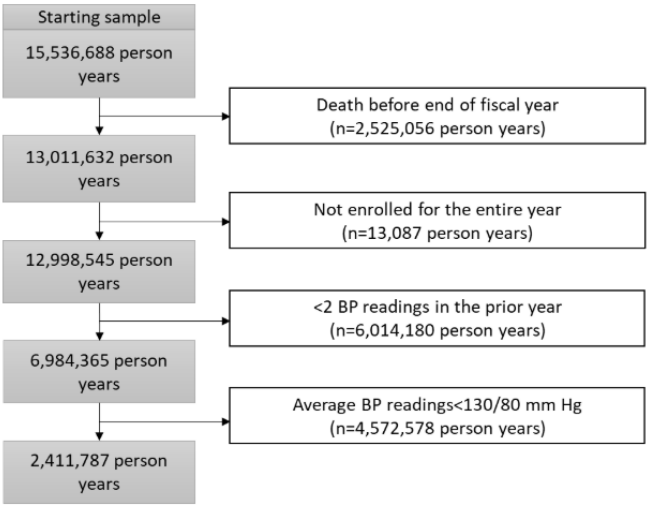

In Table 1, we share the characteristics of all veterans in our sample stratified by whether they were ever enrolled in RPM during our study period. If veterans appeared more than once in our sample, time-varying characteristics (eg, age and comorbidity) reflect their status during the first year they were eligible for inclusion in our sample. For priority group (n=6482 person-years), rurality (n=2320 person-years), ADI (n=36,351 person-years), and comorbidity (Elixhauser Comorbidity Index; diabetes and hypertension with complications; n=58,249 person-years), we used multiple imputation to replace missing values. Overall rates of RPM use were low, with 4.1% (56,553/1,390,995; 95% CI 4.0%-4.1%) of the total sample ever enrolling in RPM. There were statistically significant differences between those who enrolled and those who never enrolled across all veteran characteristics. Those with any RPM enrollment were more likely to be Black compared to individuals with no enrollment (19,457/56,533 veterans, 34.4%, 95% CI 34.0%-34.8% vs 350,391/1,334,462, 26.3%, 95% CI 26.2%-26.3%), and there was a very small but statistically significant difference by ADI percentile (the mean percentile was 0.3 higher for those with no RPM enrollment). Veterans who ever enrolled in RPM were more likely than those who never enrolled to be in VHA priority groups 1 to 3 (service-connected disability; 39,807/56,533, 70.4%, 95% CI 70%-70.8% vs 906,648/1,334,462, 67.9%, 95% CI 67.9%-68%), reside in urban areas (38,555/56,533, 68.2%, 95% CI 67.8%-68.6% vs 864,851/1,334,462 64.8%, 95% CI 64.7%-64.9%), and have 4 or more comorbidities (22,841/56,533, 40.4%, 95% CI 40.4%-40.8% vs 369,619/1,334,462, 27.7%, 95% CI 27.6%-27.8%).

XSL•FO
RenderX

Table 1. Characteristics of veterans with hypertension in the study sample between fiscal years 2020 and 2023 by enrollment in remote patient monitoring (N=1,390,995).

Characteristic	Never enrolled (n=1,334,462)	Ever enrolled (n=56,533)	Total	*P* value
Area Deprivation Index, mean (SD)	58.9 (25.2)	59.2 (25.2)	58.9 (25.2)	.001
Race n (%; 95% CI)				<.001
Hispanic, any race	83,061 (6.2; 6.2-6.3)	3421 (6.1; 5.9-6.3)	86,482 (6.2; 6.2-6.3)	
Non-Hispanic American Indian or Alaska Native	8,697 (0.7; 0.6-0.7)	309 (0.5; 0.5-0.6)	9006 (0.6; 0.6-0.7)	
Non-Hispanic Asian American and Pacific Islander	22,678 (1.7; 1.7-1.7)	747 (1.3; 1.2-1.4)	23,425 (1.7; 1.7-1.7)	
Non-Hispanic Black	350,391 (26.3; 26.2-26.3)	19,457 (34.4; 34.0-34.8)	369,848 (26.6; 26.5-26.7)	
Non-Hispanic White	805,018 (60.3; 60.2-60.4)	30,052 (53.2; 52.7-53.6)	835,070 (60; 60.0-60.1)	
Non-Hispanic other	64,617 (4.8; 4.8-4.9)	2548 (4.5; 4.3-4.7)	67,165 (4.8; 4.8-4.9)	
Fiscal year, n (%; 95% CI)				<.001
2020	636,001 (47.7; 47.6-47.7)	30,390 (53.8; 53.3-54.2)	666,391 (47.9; 47.8-48.0)	
2021	229,647 (17.2; 17.1-17.3)	10,252 (18.1; 17.8-18.5)	239,899 (17.2; 17.2-17.3)	
2022	233,225 (17.5; 17.4-17.5)	8772 (15.5; 15.2-15.8)	241,997 (17.4; 17.3-17.5)	
2023	235,589 (17.7; 17.6-17.7)	7120 (12.6; 12.3-12.9)	242,709 (17.4; 17.4-17.5)	
Sex, n (%; 95% CI)				<.001
Female	99,072 (7.4; 7.4-7.5)	5382 (9.5; 9.3-9.8)	104,454 (7.5; 7.5-7.6)	
Male	1,235,388 (92.6; 92.5-92.6)	51,151 (90.5; 90.2-90.7)	1,286,539 (92.5; 92.4-92.5)	
Age (y), n (%; 95% CI)				<.001
<65	673,661 (50.5; 50.4-50.6)	28,522 (50.5; 50.0-50.9)	702,183 (50.5; 50.4-50.6)	
65-74	499,645 (37.5; 37.4-37.5)	21,557 (38.1; 37.7-38.5)	521,202 (37.0; 37.4-37.6)	
75-84	124,783 (9.0; 9.3-9.4)	5172 (9.1; 8.9-9.4)	129,955 (9.3; 9.3-9.4)	
≥85	36,371 (2.8; 2.7-2.8)	1282 (2.3; 2.1-2.4)	37,653 (2.7; 2.7-2.7)	
Priority group, n (%; 95% CI)				<.001
1-3	906,648 (67.9; 67.9-68.0)	39,807 (70.4; 70.0-70.8)	946,455 (68.0; 68.0-68.1)	
4	15,105 (1.1; 1.1-1.2)	782 (1.4; 1.3-1.5)	15,887 (1.1; 1.1-1.2)	
5	211,435 (15.8; 15.8-15.9)	9138 (16.2; 15.9-16.5)	220,410 (15.8; 15.8-15.9)	
6-8	201,435 (15.1; 15.0-15.2)	6806 (12; 11.8-12.3)	208,241 (15.0; 14.9-15.0)	
Rurality, n (%; 95% CI)				<.001
Urban	864,851 (64.8; 64.7-64.9)	38,555 (68.2; 67.8-68.6)	903,406 (64.9; 64.9-65.0)	
Rural	415,852 (31.2; 31.1-31.2)	15,965 (28.2; 27.9-28.6)	431,817 (31.0; 31.0-31.1)	
Highly rural	53,759 (4.0; 4.0-4.1)	2014 (3.6; 3.4-3.7)	55,773 (4.0; 4.0-4.0)	
Elixhauser comorbidities, n (%; 95% CI)				<.001
1	262,970 (19.7; 19.6-19.8)	7055 (12.5; 12.2-12.8)	270,025 (19.4; 19.3-19.5)	
2	384,950 (28.8; 28.8-28.9)	13,145 (23.3; 22.9-23.6)	398,095 (28.6; 28.5-28.7)	
3	316,923 (23.7; 23.7-23.8)	13,492 (23.9; 23.5-24.2)	330,415 (23.8; 23.7-23.8)	
≥4	369,619 (27.7; 27.6-27.8)	22,841 (40.4; 40.0-40.8)	392,460 (28.2; 28.1-28.3)	
Hypertension type, n (%; 95% CI)				<.001
Uncomplicated	1,224,737 (91.8; 91.7-91.8)	49,383 (87.4; 87.1-87.6)	1,274,120 (91.6; 91.6-91.6)	
Complications[a]	109,725 (8.2; 8.2-8.3)	7150 (12.6; 12.4-12.9)	116,875 (8.4; 8.4-8.4)	

XSL•FO
RenderX

Characteristic	Never enrolled (n=1,334,462)	Ever enrolled (n=56,533)	Total	*P* value
Diabetes type 1 or 2, n (%; 95% CI)				<.001
No	899,914 (67.4; 67.4-67.5)	31,719 (56.1; 55.7-56.5)	931,633 (67.0; 66.9-67.1)	
Yes	434,548 (32.6; 32.5-32.6)	24,814 (43.9; 43.5-44.3)	459,362 (33.0; 32.9-33.1)	

[a]With hypertension-related heart disease or kidney disease.

In Table 2, we share the results of our first regression model. Compared to White veterans, we found that Black veterans (odds ratio [OR] 1.65, 95% CI 1.59-1.70) and veterans of "other" race (OR 1.11, 95% CI 1.03-1.19) had higher odds of RPM enrollment. Asian American and Pacific Islander (AAPI) veterans were slightly less likely (OR 0.83, 95% CI 0.74-0.94) to enroll. We found that, while technically statistically significant, there was no meaningful association between a 10-percentile change in ADI ranking and the odds (OR 1.00, 95% CI 0.99-1.00) of RPM enrollment. We tested different variable specifications of ADI (ADI quartiles and specific cutoff points [ie, 90th percentile]) and found that those in the highest quartile of ADI (most vulnerable) had slightly lower odds of RPM enrollment compared to those in the first quartile (least vulnerable) but that the magnitude (OR 0.93, 95% CI 0.90-0.97) was still quite small (Multimedia Appendix 1).

XSL•FO
RenderX

Table 2. Odds ratios from logistic regression modeling the odds of remote patient monitoring enrollment among veterans with hypertension from fiscal years 2020 to 2023.

Characteristic	Odds ratio (95% CI)
Race	
Hispanic, any race	1.09 (1.02-1.16)
Non-Hispanic American Indian or Alaska Native	0.90 (0.74-1.09)
Non-Hispanic Asian American and Pacific Islander	0.83 (0.74-0.94)
Non-Hispanic Black	1.65 (1.59-1.70)
Non-Hispanic White	__[a]
Non-Hispanic other	1.11 (1.03-1.19)
Area Deprivation Index (per 10-percentile change)	1.00 (0.99-1.00)
Fiscal year	
2020	—
2021	1.22 (1.18-1.26)
2022	1.12 (1.08-1.16)
2023	1.30 (1.26-1.34)
Sex	
Female	—
Male	0.72 (0.69-0.76)
Age (y)	
<65	—
65-74	1.04 (1.01-1.08)
75-84	1.12 (1.06-1.17)
≥85	0.95 (0.87-1.05)
Priority group	
1-3	—
4	1.20 (1.05-1.36)
5	1.01 (0.97-1.05)
6-8	0.83 (0.79-0.87)
Rurality	
Urban	—
Rural	0.91 (0.88-0.94)
Highly rural	0.94 (0.87-1.02)
Elixhauser comorbidities	
1	—
2	1.24 (1.18-1.29)
3	1.63 (1.56-1.71)
≥4	2.47 (2.36-2.58)
Hypertension type	
Uncomplicated	—
With complications[b]	1.57 (1.51-1.64)
Comorbid diabetes	
No	—
Yes	1.47 (1.42-1.51)

XSL•FO
RenderX

[a]Reference group.

[b]With hypertension-related heart disease or kidney disease.

In addition to veteran race, we found that male veterans were less likely (OR 0.72, 95% CI 0.69-0.76) to enroll in RPM and those between the ages of 65 and 84 years were slightly more likely to enroll in RPM than younger veterans. Among clinical covariates, veterans with diabetes (OR 1.47, 95% CI 1.42-1.51), hypertension-associated complications (eg, heart disease or kidney disease; OR 1.57, 95% CI 1.51-1.64), and more chronic health conditions (OR 1.24-2.47) were all significantly associated with higher odds of RPM enrollment.

Sensitivity Analysis

For sensitivity analyses, we further limited our sample to those with grade or stage 2 hypertension (average BP>140/80 mm Hg across ≥2 BP measurements) and who also were prescribed at least one antihypertensive medication. These requirements further limited our sample to 464,971 person-years across 348,265 unique veterans, or 19.3% (464,971 /2,411,782 person-years) of the sample in the original analysis. Among these veterans, compared to the initial sample, a higher proportion (17,762/348,265, 5.1%, 95% CI 5.0%-5.2%) were enrolled in RPM.

In the adjusted analysis for this population (Table 3), Black veterans still had higher odds (OR 1.61, 95% CI 1.50-1.72) of being enrolled in RPM compared to White veterans. There was no longer a statistically significant association between AAPI race and RPM enrollment. However, there was a slightly larger association between “other” race and RPM enrollment (OR 1.19, 95% CI 1.03-1.37). For ADI, the association with RPM enrollment remained very small. While most variables that were significant in the first model remained so in the second model, the magnitude was attenuated; for example, those with 4 or more comorbidities now had only 2.04 (95% CI 1.86-2.24) times the odds of RPM enrollment compared to 2.47 (95% CI 2.36-2.58) times the odds among the larger sample. The magnitude of other covariates also shifted, with some associations now not significant either due to differences in the patient population or due to the smaller sample size.

XSL•FO
RenderX

Table 3. Odds ratios from logistic regression modeling the odds of remote patient monitoring enrollment among veterans with stage 2 hypertension (blood pressure >140/90 mm Hg) and antihypertensive medication use from fiscal years 2020 to 2023.

Characteristic	Odds ratio (95% CI)
Race	
Hispanic, any race	1.27 (1.12-1.45)
Non-Hispanic American Indian or Alaska Native	1.12 (0.77-1.65)
Non-Hispanic Asian American and Pacific Islander	1.02 (0.80-1.29)
Non-Hispanic Black	1.61 (1.50-1.72)
Non-Hispanic White	__[a]
Non-Hispanic other	1.19 (1.03-1.37)
Area Deprivation Index (per 10-percentile change)	0.99 (0.98-1.00)
Fiscal year	
2020	—
2021	1.14 (1.05-1.22)
2022	1.07 (0.99-1.15)
2023	1.26 (1.17-1.36)
Sex	
Female	—
Male	0.68 (0.61-0.75)
Age (y)	
<65	—
65-74	0.98 (0.92-1.05)
75-84	1.07 (0.93-1.02)
≥85	1.03 (0.77-1.37)
Priority group	
1-3	—
4	1.25 (0.93-1.68)
5	0.94 (0.86-1.02)
6-8	0.92 (0.84-1.00)
Rurality	
Urban	—
Rural	0.94 (0.88-1.01)
Highly rural	0.80 (0.67-0.95)
Elixhauser comorbidities	
1	—
2	1.18 (1.08-1.29)
3	1.48 (1.35-1.63)
≥4	2.04 (1.86-2.24)
Hypertension type	
Uncomplicated	—
With complications[b]	1.63 (1.51-1.77)
Comorbid diabetes	
No	—
Yes	1.36 (1.28-1.45)

XSL•FO
RenderX

[a]Reference group.

[b]With hypertension-related heart disease or kidney disease.

Discussion

Principal Results

RPM for hypertension may provide disease management support; however, we found that only 4.1% (56,533/1,390,995) of potentially eligible veterans were ever enrolled during our study period. This finding aligns with the relatively low enrollment in RPM for other chronic diseases within the VHA [38,42]. While having different inclusion criteria, other studies examining RPM in the VHA have found relatively low enrollment [38]. In our retrospective national cross-sectional study of veterans with hypertension from FYs 2020-2023, we found mixed data on racial disparities in RPM enrollment and no association with the socioeconomic vulnerability of the neighborhood where the veterans lived. Black veterans and veterans of other races, including those identifying as multiracial or with unknown race, were more likely to be enrolled in RPM, whereas AAPI veterans were less likely to enroll. We found no meaningful difference in the likelihood of enrollment by neighborhood socioeconomic characteristics.

Comparison With Prior Work

Controlling for potential demographic and clinical confounders, we found that Black veterans were more likely than White veterans to enroll in RPM for hypertension and AAPI veterans were less likely to enroll. Previous research found that overall awareness of hypertension and hypertension control was lower among Asian American patients compared to non-Hispanic White patients, although there was considerable variation by ethnic group [14,43]. Prior research involving the use of RPM among patients without hypertension found that Black patients were more likely to be engaged in the program [44], whereas other research on different populations (nonveterans and those with heart failure) found that female patients and Black patients were less likely to enroll in RPM [32]. In our study, veterans with multiple comorbid conditions and those with diabetes or hypertension-associated complications were more likely to enroll in RPM. These findings indicate that VHA RPM is reaching the appropriate population as those targeted for additional support by VHA RPM are patients who have high health service use (eg, frequent emergency room visits) due to multimorbidity [5,31].

We found no meaningful difference in enrollment in RPM by ADI. This finding held true even when we tested different approaches to modeling, such as ADI quartiles or specified cutoffs. This finding supports previous research among rural and low-income patients that found that socioeconomic status was not a barrier to RPM use [45]. However, it contrasts other research that found that those in more disadvantaged neighborhoods are less likely to have access to and enroll in RPM [44,46-48]. These studies have often cited factors such as differential access to care, low health literacy, limited access to technology, or broadband availability (ie, internet access) as barriers to RPM use [49-51]. Barriers to RPM in the VHA are in part ameliorated because access to RPM is available to all eligible veterans at no cost and the technology used in the VHA does not necessitate broadband availability [5]. This may explain why ADI may not have impacted RPM use in our study population.

Among a smaller subset of patients with documented stage or grade 2 hypertension despite the use of hypertension medications, we still found that Black veterans were more likely than White veterans and AAPI veterans to enroll in RPM, but the magnitude of the association shrank, as it did for most clinical covariates. Perhaps the reason Black veterans were more likely to enroll in RPM was linked to their health and morbidity status as Black veterans are more likely to report circulatory-related chronic conditions than non-Hispanic White veterans [52]. Previous research has found that patients with hypertension and concomitant multimorbidity incur higher health care costs and have more complex health care needs [53-55]. While Black veterans were more likely than White veterans to enroll in the VHA RPM program, future research is needed to encourage uptake of RPM and determine whether RPM is equally effective across all participants. Additionally, given the low (approximately 4%) overall rate of RPM enrollment, RPM is likely one of many tools and approaches required to improve hypertension control in the larger population. Related interventions, such as video BP visits, have successfully reduced BP in veterans, but the intervention did not assess for whom the intervention was most effective [56]. Recent literature on hypertension among veterans shows that Black veterans, followed closely by American Indian and Alaska Native veterans, have the lowest overall rates of hypertension control [17]. Studies of RPM programs that specifically evaluate their effect on racial disparities in hypertension outcomes are promising but limited [57].

Limitations

While our sample included more than 1 million unique veterans over 4 years, our findings may not be generalizable to other populations. Veterans are not representative of the general American population as they are much more likely to be male, White, and older, although younger veteran populations are more diverse. Furthermore, we limited our sample to those with ≥2 BP readings on ≥2 occasions with an average BP of >130/80 mm Hg. This approach ensured that our study cohort had an active case of hypertension according to clinical guidelines but likely nonrandomly eliminated veterans who were healthier (and had fewer medical encounters in which BP was taken) and those who were less reliant on VHA care and less likely to have BP readings documented in VHA data. This approach further limits the generalizability of our study to the veteran population as a whole. We also excluded veterans who died, which likely eliminated veterans with poorer health, which also limits the implications of our study. Our study also only evaluated the VHA's RPM program, whereas other similar programs to monitor and control hypertension may be structured differently. Participants of other non-VHA programs may also have different barriers (such as lack of insurance coverage and changes in health care systems) that are less present in the VHA veteran

XSL•FO
RenderX

population, where many, though not all, veterans receive much of their care in the same system with low or no copayments. Those in lower priority groups (6-8) were the least likely to enroll in RPM, perhaps because they received less care at the VHA overall and were subject to copayments for some services [58,59].

Additionally, while we controlled for a number of clinical and demographic characteristics, there may still be some confounding information that we were unable to capture. Although widely used with established best practices at identifying the most accurate value, Corporate Data Warehouse data on race do not always consist of veteran self-reported race. Primary care provider awareness of RPM as an option for their patients may vary substantially, and there could also be differences between facilities in terms of awareness, uptake, and support for this program for veterans. Future research will examine these other factors. This study also did not evaluate how long veterans stayed in the RPM program, only whether they enrolled. Recent research has examined enrollment duration, and data on reasons for enrollment and disenrollment are stored externally to the VHA electronic health record by a third-party server, which was not accessible for this study [8]. Finally, while there are limitations to the use of ORs for reporting regression results, for rare outcomes such as ours, ORs approximate the risk ratio and eliminate these potential concerns [60].

Conclusions

Controlling for potential confounders, Black veterans were more likely than White veterans to be enrolled in the VHA hypertension RPM program, but we found no difference in RPM enrollment by neighborhood socioeconomic status. These results were consistent when limited to a smaller sample of veterans with hypertension that remained uncontrolled despite the use of antihypertensive medications. Prior research has assessed the duration of enrollment and outcomes after enrolling, but no prior studies have examined the probability of any enrollment in the RPM program for controlling hypertension in a veteran population. In an integrated health care system in which most patients do not pay for health care services, we found that patients from socioeconomically disadvantaged areas and Black veterans were not less likely to be enrolled in RPM. Our study results demonstrate that, within the VHA-enrolled veteran population, RPM is reaching an appropriate patient population (ie, patients with more comorbidities and patients with stage or grade 2 hypertension). However, overall enrollment was low, suggesting that RPM could still benefit additional patients who are not currently enrolled.

For non-VHA settings, reducing or eliminating financial barriers may improve access for similar nonveteran populations. As overall enrollment in RPM for hypertension was low, future research should focus on improving uptake among veterans who could additionally benefit from the program, including White and AAPI veterans. While the VHA is a unique health care system in the United States, being the largest integrated health care system in the country and serving a unique population, outside the VHA, RPM programs may be able to effectively reach minority populations and those living in socioeconomically disadvantaged areas. This study demonstrated that RPM use was not impacted by neighborhood socioeconomic status when the use of those services was either fully or partially paid for and the technology was not reliant on broadband availability. Non-VHA systems, particularly those serving low-income or socioeconomically disadvantaged areas, should explore subsidized or free RPM programs for eligible patients similar to the VHA's no-cost model for veterans.

Acknowledgments

The opinions expressed in this paper are the authors' own and do not necessarily represent the views of the US Department of Veterans Affairs or the US government. No generative artificial intelligence was used at any point in the development of this manuscript.

Funding

HRF was supported by the US Department of Veterans Affairs Office of Academic Affiliations Advanced Fellowship program in Health Systems Research at the Center for Healthcare Organization and Implementation Research, US Department of Veterans Affairs Boston Healthcare System.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to Veterans Health Administration policies but are available from the corresponding author on reasonable request. All requests are subject to compliance with all security requirements of the US Department of Veterans Affairs data sharing agreements.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sensitivity analysis.
[DOCX File , 14 KB - jmir_v28i1e78423_app1.docx]

XSL•FO
RenderX

References

1. Introduction to telehealth and remote patient monitoring. Telehealth.HHS.gov. URL: https://telehealth.hhs.gov/providers/best-practice-guides/telehealth-and-remote-patient-monitoring [accessed 2025-03-31]
2. Farias FA, Dagostini CM, Bicca YD, Falavigna VF, Falavigna A. Remote patient monitoring: a systematic review. Telemed J E Health 2020 May 17;26(5):576-583. [doi: 10.1089/tmj.2019.0066] [Medline: 31314689]
3. Smith W, Colbert BM, Namouz T, Caven D, Ewing JA, Albano AW. Remote patient monitoring is associated with improved outcomes in hypertension: a large, retrospective, cohort analysis. Healthcare (Basel) 2024 Aug 09;12(16):1583 [FREE Full text] [doi: 10.3390/healthcare12161583] [Medline: 39201142]
4. Rasmussen P, Farmer C. The promise and challenges of VA community care: veterans' issues in focus. Rand Health Q 2023 Jun;10(3):9 [FREE Full text] [Medline: 37333666]
5. Remote patient monitoring - home telehealth (RPMHT). U.S. Department of Veterans Affairs. URL: https://www.va.gov/lexington-health-care/programs/remote-patient-monitoring-home-telehealth-rpmht/ [accessed 2025-03-31]
6. At home. U.S. Department of Veterans Affairs. URL: https://telehealth.va.gov/type/home [accessed 2025-06-02]
7. Home telehealth operations manual. VHA Office of Connected Care. 2017 Dec. URL: https://www.vendorportal.ecms.va.gov/FBODocumentServer/DocumentServer.aspx?DocumentId=4869870&FileName=36C25718R0247-030.pdf [accessed 2025-12-21]
8. Gray KE, Rao M, Gunnink E, Eschenroeder L, Geyer JR, Nelson KM, et al. Home telehealth in the Veterans Health Administration: trends and correlates of length of enrollment from 2010 to 2017. J Gen Intern Med 2022 Sep;37(12):3089-3096 [FREE Full text] [doi: 10.1007/s11606-022-07452-1] [Medline: 35230624]
9. Bludorn J, Railey K. Hypertension guidelines and interventions. Prim Care 2024 Mar;51(1):41-52. [doi: 10.1016/j.pop.2023.07.002] [Medline: 38278572]
10. Tsao CW, Aday AW, Almarzooq ZI, Anderson CA, Arora P, Avery CL, et al. Heart disease and stroke statistics-2023 update: a report from the American Heart Association. Circulation 2023 Feb 21;147(8):e93-621. [doi: 10.1161/CIR.0000000000001123] [Medline: 36695182]
11. Chobufo MD, Gayam V, Soluny J, Rahman EU, Enoru S, Foryoung JB, et al. Prevalence and control rates of hypertension in the USA: 2017–2018. Int J Cardiol Hypertens 2020 Jul 31;6:100044. [doi: 10.1016/j.ijchy.2020.100044]
12. Aggarwal R, Chiu N, Wadhera RK, Moran AE, Raber I, Shen C, et al. Racial/ethnic disparities in hypertension prevalence, awareness, treatment, and control in the United States, 2013 to 2018. Hypertension 2021 Dec;78(6):1719-1726 [FREE Full text] [doi: 10.1161/HYPERTENSIONAHA.121.17570] [Medline: 34365809]
13. Yamada M, Wachsmuth J, Sambharia M, Griffin BR, Swee ML, Reisinger HS, et al. The prevalence and treatment of hypertension in Veterans Health Administration, assessing the impact of the updated clinical guidelines. J Hypertens 2023 Jun 01;41(6):995-1002 [FREE Full text] [doi: 10.1097/HJH.0000000000003424] [Medline: 37071434]
14. Abrahamowicz AA, Ebinger J, Whelton SP, Commodore-Mensah Y, Yang E. Racial and ethnic disparities in hypertension: barriers and opportunities to improve blood pressure control. Curr Cardiol Rep 2023 Jan 09;25(1):17-27. [doi: 10.1007/s11886-022-01826-x] [Medline: 36622491]
15. Xu J, Lawrence KG, O'Brien KM, Jackson CL, Sandler DP. Association between neighbourhood deprivation and hypertension in a US-wide Cohort. J Epidemiol Community Health 2022 Mar 17;76(3):268-273. [doi: 10.1136/jech-2021-216445] [Medline: 34789553]
16. Satapathy P, Khatib MN, Gaidhane S, Zahiruddin QS, Gaidhane AM, Rustagi S, et al. Association of neighborhood deprivation and hypertension: a systematic review and meta-analysis. Curr Probl Cardiol 2024 Apr;49(4):102438 [FREE Full text] [doi: 10.1016/j.cpcardiol.2024.102438] [Medline: 38301916]
17. Lind KE, Wong MS, Frochen SE, Yuan AH, Washington DL. Variation in hypertension control by race and ethnicity, and geography in US veterans. J Am Heart Assoc 2025 Jan 21;14(2):e035682 [FREE Full text] [doi: 10.1161/JAHA.123.035682] [Medline: 39791424]
18. Krishnamurthi N, Francis J, Fihn SD, Meyer CS, Whooley MA. Leading causes of cardiovascular hospitalization in 8.45 million US veterans. PLoS One 2018 Mar 22;13(3):e0193996. [doi: 10.1371/journal.pone.0193996] [Medline: 29566396]
19. Makutonin M, Dare J, Heekin M, Salancy A, Hood C, Dominguez LW. Remote patient monitoring for hypertension: feasibility and outcomes of a clinic-based pilot in a minority population. J Prim Care Community Health 2023;14:21501319231204586 [FREE Full text] [doi: 10.1177/21501319231204586] [Medline: 37815085]
20. Galobardes B, Shaw M, Lawlor DA, Lynch JW, Davey Smith G. Indicators of socioeconomic position (part 1). J Epidemiol Community Health 2006 Jan;60(1):7-12 [FREE Full text] [doi: 10.1136/jech.2004.023531] [Medline: 16361448]
21. Krieger N, Williams DR, Moss NE. Measuring social class in US public health research: concepts, methodologies, and guidelines. Annu Rev Public Health 1997 May;18(1):341-378. [doi: 10.1146/annurev.publhealth.18.1.341] [Medline: 9143723]
22. van Ham M, Manley D, Bailey N, Simpson L, Maclennan D. Neighbourhood Effects Research: New Perspectives. Dordrecht, The Netherlands: Springer; 2012.
23. Parati G, Ochoa JE, Postel-Vinay N, Pellegrini D, Torlasco C, Omboni S, et al. Home blood pressure telemonitoring: conventional approach and perspectives from mobile health technology. In: Stergiou G, Parati G, Mancia G, editors. Home Blood Pressure Monitoring. Cham, Switzerland: Springer; 2020:103-119.

XSL•FO
RenderX

24. Park SH, Shin JH, Park J, Choi WS. An updated meta-analysis of remote blood pressure monitoring in urban-dwelling patients with hypertension. Int J Environ Res Public Health 2021 Oct 09;18(20):10583 [FREE Full text] [doi: 10.3390/ijerph182010583] [Medline: 34682329]
25. Najarian M, Goudie A, Bona JP, Rezaeiahari M, Young SG, Bogulski CA, et al. Socioeconomic determinants of remote patient monitoring implementation among rural and urban hospitals. Telemed J E Health 2023 Nov 03;29(11):1624-1633. [doi: 10.1089/tmj.2022.0412] [Medline: 37010391]
26. Rivera V, Aldridge MD, Ornstein K, Moody KA, Chun A. RESEARCHRacial and socioeconomic disparities in access to telehealth. J Am Geriatr Soc 2021 Jan 24;69(1):44-45 [FREE Full text] [doi: 10.1111/jgs.16904] [Medline: 33075143]
27. Darrat I, Tam S, Boulis M, Williams AM. Socioeconomic disparities in patient use of telehealth during the coronavirus disease 2019 surge. JAMA Otolaryngol Head Neck Surg 2021 Mar 01;147(3):287-295 [FREE Full text] [doi: 10.1001/jamaoto.2020.5161] [Medline: 33443539]
28. Rocque GB, Caston NE, Hildreth K, Deng L, Henderson NL, Williams CP, et al. Engagement among diverse patient backgrounds in a remote symptom monitoring program. JCO Oncol Pract 2024 Oct;20(10):1426-1435. [doi: 10.1200/OP.24.00066] [Medline: 38917385]
29. Guzman-Clark J, Wakefield BJ, Farmer MM, Yefimova M, Viernes B, Lee ML, et al. Adherence to the use of home telehealth technologies and emergency room visits in veterans with heart failure. Telemed J E Health 2021 Sep 01;27(9):1003-1010 [FREE Full text] [doi: 10.1089/tmj.2020.0312] [Medline: 33275527]
30. Guzman-Clark JR, van Servellen G, Chang B, Mentes J, Hahn TJ. Predictors and outcomes of early adherence to the use of a home telehealth device by older veterans with heart failure. Telemed J E Health 2013 Mar;19(3):217-223. [doi: 10.1089/tmj.2012.0096] [Medline: 23268695]
31. Guzman-Clark J, Farmer MM, Wakefield BJ, Viernes B, Yefimova M, Lee ML, et al. Why patients stop using their home telehealth technologies over time: predictors of discontinuation in Veterans with heart failure. Nurs Outlook 2021 Mar;69(2):159-166 [FREE Full text] [doi: 10.1016/j.outlook.2020.11.004] [Medline: 33279151]
32. Mastoris I, DeFilippis EM, Martyn T, Morris AA, Van Spall HG, Sauer AJ. Remote patient monitoring for patients with heart failure: sex- and race-based disparities and opportunities. Card Fail Rev 2023 Feb 1;9:e02 [FREE Full text] [doi: 10.15420/cfr.2022.22] [Medline: 36891178]
33. Chen S, Li J. Participation of Black US residents in clinical trials of 24 cardiovascular drugs granted FDA approval, 2006-2020. JAMA Netw Open 2021 Mar 01;4(3):e212640 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.2640] [Medline: 33755163]
34. Ghaferi AA, Schwartz TA, Pawlik TM. STROBE reporting guidelines for observational studies. JAMA Surg 2021 Jun 01;156(6):577-578. [doi: 10.1001/jamasurg.2021.0528] [Medline: 33825815]
35. Beckman KD. How to document and code for hypertensive diseases in ICD-10. Fam Pract Manag 2014;21(2):5-9 [FREE Full text] [Medline: 24693838]
36. Reboussin DM, Allen NB, Griswold ME, Guallar E, Hong Y, Lackland DT, et al. Systematic review for the 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Hypertension 2018 Jun;71(6):e116-e135 [FREE Full text] [doi: 10.1161/HYP.0000000000000067] [Medline: 29133355]
37. Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison Himmelfarb C, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Hypertension 2018 Jun 26;71(6):e13-115 [FREE Full text] [doi: 10.1161/HYP.0000000000000065] [Medline: 29133356]
38. Mohr NM, Vakkalanka JP, Holcombe A, Carter KD, McCoy KD, Clark HM, et al. Effect of chronic disease home telehealth monitoring in the Veterans Health Administration on healthcare utilization and mortality. J Gen Intern Med 2023 Nov 08;38(15):3313-3320. [doi: 10.1007/s11606-023-08220-5] [Medline: 37157039]
39. Ferguson JM, Jacobs J, Yefimova M, Greene L, Heyworth L, Zulman DM. Virtual care expansion in the Veterans Health Administration during the COVID-19 pandemic: clinical services and patient characteristics associated with utilization. J Am Med Inform Assoc 2021 Mar 01;28(3):453-462 [FREE Full text] [doi: 10.1093/jamia/ocaa284] [Medline: 33125032]
40. Peltzman T, Rice K, Jones KT, Washington DL, Shiner B. Optimizing data on race and ethnicity for Veterans Affairs patients. Mil Med 2022 Jul 01;187(7-8):e955-e962. [doi: 10.1093/milmed/usac066] [Medline: 35323934]
41. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care 1998 Jan;36(1):8-27. [doi: 10.1097/00005650-199801000-00004] [Medline: 9431328]
42. Reddy A, Gunnink EJ, Rojas J, Nelson K, Wong ES. Impact of home telehealth expansion on high-cost utilization among Veterans Health Administration patients with diabetes. J Gen Intern Med 2025 May 31;40(6):1340-1349. [doi: 10.1007/s11606-024-09169-9] [Medline: 39482477]
43. Jung MY, Lee S, Thomas SB, Juon HS. Hypertension prevalence, treatment, and related behaviors among Asian Americans: an examination by method of measurement and disaggregated subgroups. J Racial Ethn Health Disparities 2019 Jun 7;6(3):584-593 [FREE Full text] [doi: 10.1007/s40615-018-00557-6] [Medline: 30618006]

XSL•FO
RenderX

44. Kirkland E, Schumann SO, Schreiner A, Heincelman M, Zhang J, Marsden J, et al. Patient demographics and clinic type are associated with patient engagement within a remote monitoring program. Telemed J E Health 2021 Aug 01;27(8):843-850 [FREE Full text] [doi: 10.1089/tmj.2020.0535] [Medline: 34115942]
45. Clark D3, Woods J, Zhang Y, Chandra S, Summers RL, Jones DW. Home blood pressure telemonitoring with remote hypertension management in a rural and low-income population. Hypertension 2021 Dec;78(6):1927-1929 [FREE Full text] [doi: 10.1161/HYPERTENSIONAHA.121.18153] [Medline: 34757773]
46. Fritz BA, Ramsey B, Taylor D, Shoup JP, Schmidt JM, Guinn M, et al. Association of race and neighborhood disadvantage with patient engagement in a home-based COVID-19 remote monitoring program. J Gen Intern Med 2022 Mar;37(4):838-846 [FREE Full text] [doi: 10.1007/s11606-021-07207-4] [Medline: 34993862]
47. Akar JG, Bao H, Jones P, Wang Y, Chaudhry SI, Varosy P, et al. Use of remote monitoring of newly implanted cardioverter-defibrillators: insights from the patient related determinants of ICD remote monitoring (PREDICT RM) study. Circulation 2013 Nov 26;128(22):2372-2383. [doi: 10.1161/CIRCULATIONAHA.113.002481] [Medline: 24043302]
48. Varma N, Piccini JP, Snell J, Fischer A, Dalal N, Mittal S. The relationship between level of adherence to automatic wireless remote monitoring and survival in pacemaker and defibrillator patients. J Am Coll Cardiol 2015 Jun 23;65(24):2601-2610 [FREE Full text] [doi: 10.1016/j.jacc.2015.04.033] [Medline: 25983008]
49. Foster M, Etchin A, Pope C, Hartmann CW, Emidio O, Bosworth HB. The impact of COVID-19 on hypertension and hypertension medication adherence among underrepresented racial and ethnic groups: a scoping review. Curr Hypertens Rep 2023 Nov 25;25(11):385-394. [doi: 10.1007/s11906-023-01262-4] [Medline: 37624472]
50. Patel P, Dhindsa D, Eapen DJ, Khera A, Gulati M, Stone NJ, et al. Optimizing the potential for telehealth in cardiovascular care (in the era of COVID-19): time will tell. Am J Med 2021 Aug;134(8):945-951. [doi: 10.1016/j.amjmed.2021.03.007] [Medline: 33845033]
51. Bress AP, Cohen JB, Anstey DE, Conroy MB, Ferdinand KC, Fontil V, et al. Inequities in hypertension control in the United States exposed and exacerbated by COVID-19 and the role of home blood pressure and virtual health care during and after the COVID-19 pandemic. J Am Heart Assoc 2021 Jun;10(11):e020997 [FREE Full text] [doi: 10.1161/JAHA.121.020997] [Medline: 34006116]
52. Ward RE, Nguyen XM, Li Y, Lord E, Lecky V, Song RJ, et al. Racial and ethnic disparities in U.S. veteran health characteristics. Int J Environ Res Public Health 2021 Mar 02;18(5):2411. [doi: 10.3390/ijerph18052411] [Medline: 33801200]
53. Wong MC, Wang HH, Cheung CS, Tong EL, Sek AC, Cheung NT, et al. Factors associated with multimorbidity and its link with poor blood pressure control among 223,286 hypertensive patients. Int J Cardiol 2014 Nov 15;177(1):202-208. [doi: 10.1016/j.ijcard.2014.09.021] [Medline: 25499379]
54. Nguyen TN, Kalia S, Hanlon P, Jani BD, Nicholl BI, Christie CD, et al. Multimorbidity and blood pressure control in patients attending primary care in Canada. J Prim Care Community Health 2023 Dec 14;14:21501319231215025 [FREE Full text] [doi: 10.1177/21501319231215025] [Medline: 38097504]
55. Alshahawey M, Jafari E, Smith SM, McDonough CW. Characterizing apparent treatment resistant hypertension in the United States: insights from the All of Us Research Program. J Am Med Inform Assoc 2024 Dec 01;31(12):2899-2907. [doi: 10.1093/jamia/ocae227] [Medline: 39181122]
56. Carrico M, Frosch C, Craig K, Carter M, Falk J, Guerrero S, et al. Implementation of video blood pressure visits in the Veterans Health Administration. Telemed J E Health 2024 Apr 01;30(4):1006-1012. [doi: 10.1089/tmj.2023.0269] [Medline: 37935031]
57. Khoong EC, Commodore-Mensah Y, Lyles CR, Fontil V. Use of self-measured blood pressure monitoring to improve hypertension equity. Curr Hypertens Rep 2022 Nov 24;24(11):599-613 [FREE Full text] [doi: 10.1007/s11906-022-01218-0] [Medline: 36001268]
58. VA priority groups. U.S. Department of Veterans Affairs. URL: https://www.va.gov/health-care/eligibility/priority-groups/ [accessed 2025-04-10]
59. Hynes DM, Koelling K, Stroupe K, Arnold N, Mallin K, Sohn MW, et al. Veterans' access to and use of Medicare and Veterans Affairs health care. Med Care 2007 Mar;45(3):214-223. [doi: 10.1097/01.mlr.0000244657.90074.b7] [Medline: 17304078]
60. Cummings P. The relative merits of risk ratios and odds ratios. Arch Pediatr Adolesc Med 2009 May 04;163(5):438-445. [doi: 10.1001/archpediatrics.2009.31] [Medline: 19414690]

Abbreviations

AAPI: Asian American and Pacific Islander
ADI: Area Deprivation Index
BP: blood pressure
FY: fiscal year
OR: odds ratio
RPM: remote patient monitoring
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

VA: US Department of Veterans Affairs
VHA: Veterans Health Administration

Edited by S Brini; submitted 02.Jun.2025; peer-reviewed by C Exeter, L Tieu, J Edwards; comments to author 05.Sep.2025; accepted 19.Nov.2025; published 09.Mar.2026.

Please cite as:
Friedman HR, Mull HJ, Guzman-Clark JR, Sturgeon DJ, Foster MV
The Effect of Veteran Race and Socioeconomic Status on Enrollment in Remote Patient Monitoring for Hypertension: Retrospective Observational Cross-Sectional Study
J Med Internet Res 2026;28:e78423
URL: https://www.jmir.org/2026/1/e78423
doi:10.2196/78423
PMID:

©Hannah Robin Friedman, Hillary J Mull, Jenice Ria Guzman-Clark, Daniel J Sturgeon, Marva V Foster. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact of Tele-Intensive Care Units on the Clinical Outcomes of Critically Ill Patients With COVID-19: Retrospective Cohort Study

Thais Dias Midega[1], MD; Fabio Barlem Hohmann[1], PhD; Renato Carneiro de Freitas Chaves[1,2], PhD; Guacyra Margarita Batista de Almeida[3], MD; Vivian Jaqueline Lima Leoneza[4], MD; Jennifer Ferreira Figueiredo Cabral[5]; Bianca Veloso Vitalino[6], MD; Emanuelle de Araújo Camboim[7], MBA; Nelma de Jesus Nogueira Machado[8]; Ricardo Fernando Batista de Melo[9], MD; Jorge Patrick Oliveira Feliciano[10]; Breno Mendes Cardoso[11], MD; Leonardo José Rolim Ferraz[1], MD; Thiago Domingos Corrêa[1], PhD; Maura Cristina dos Santos[12]; Renata Albaladejo Morbeck[12]; Adriano José Pereira[1,12], PhD

[1]Department of Critical Care Medicine, Hospital Israelita Albert Einstein, São Paulo, Sao Paulo, Brazil
[2]Massachusetts Institute of Technology, Cambridge, MA, United States
[3]Department of Critical Care Medicine, Hospital de Emergência Dr. Daniel Houly, Arapiraca, Brazil
[4]Department of Critical Care Medicine, Santa Casa de Misericórdia de Sorocaba, Sorocaba, Brazil
[5]Department of Critical Care Medicine, Hospital Regional do Cariri, Juazeiro do Norte, Brazil
[6]Department of Critical Care Medicine, Santa Casa de Misericórdia de São João del-Rei, São João del-Rei, Brazil
[7]Department of Critical Care Medicine, Hospital Metropolitano Sul Dom Helder Câmara, Cabo de Santo Agostinho, Brazil
[8]Department of Critical Care Medicine, Santa Casa de Misericórdia do Pará, Belém, Brazil
[9]Department of Critical Care Medicine, Hospital Santa Lúcia, Poços de Caldas, Brazil
[10]Department of Critical Care Medicine, Dom Orione Hospital, Araguaína, Brazil
[11]Department of Critical Care Medicine, Hospital Municipal Senhora Santana, Brasília de Minas, Brazil
[12]Department of Telemedicine, Hospital Israelita Albert Einstein., São Paulo, Brazil

Corresponding Author:
Thais Dias Midega, MD
Department of Critical Care Medicine
Hospital Israelita Albert Einstein
Rua Inhambu 1125
São Paulo, Sao Paulo
Brazil
Phone: 55 11997076207
Email: thais.dmidega@einstein.br

Abstract

Background: The COVID-19 pandemic imposed an unprecedented demand for intensive care unit (ICU) resources in Brazil, where shortages of trained intensivists prompted the implementation of telemedicine-based critical care support strategies.

Objective: This study aimed to evaluate the association between adherence to the Tele-ICU COVID-19 Brazil Program and clinical outcomes of ICU patients with COVID-19.

Methods: We conducted a retrospective cohort study including all ICUs participating in the Tele-ICU COVID-19 Brazil Program between April and December 2020. Program adherence was assessed at 2 levels: patient coverage, defined as the number of daily multidisciplinary rounds per patient divided by the patient's total ICU length of stay (LOS), and ICU coverage, defined as the number of daily multidisciplinary round days in the ICU divided by the total number of patient-days in that ICU. We compared outcomes between groups categorized by an empirically defined 50% cutoff: low patient coverage (<50%) versus high patient coverage (≥50%) and low ICU coverage (<50%) versus high ICU coverage (≥50%). Multilevel mixed-effects models accounting for ICU-level clustering were used to assess outcomes: logistic regression for ICU mortality (adjusted odds ratios) and linear mixed-effects regression with log-transformed ICU LOS (exponentiated coefficients, exp[β]).

Results: A total of 1680 patients were included. Compared with the low patient coverage group (<50%), patients in the high patient coverage (≥50%) had lower Sequential Organ Failure Assessment scores (median 2, IQR 0-5 vs median 3, IQR 0-6; *P*=.007); shorter ICU LOS (median 6, IQR 3-11 days vs median 11, IQR 6-20 days; *P*<.001); and shorter hospital LOS (median

XSL•FO
RenderX

9, IQR 5-16 days vs median 14, IQR 8-26 days; *P*<.001). In unadjusted analyses, ICU mortality did not differ significantly between the low and high patient coverage groups (50.1% vs 46.3%; *P*=.16). In multilevel analysis, mechanical ventilation and vasopressor use were independently associated with higher ICU mortality. Higher patient coverage was independently associated with lower ICU mortality (adjusted odds ratio 0.52, 95% CI 0.27-0.99; *P*=.048). In the log-transformed mixed-effects model for ICU LOS, a higher Sequential Organ Failure Assessment score (exp[β] 1.037, 95% CI 1.02-1.05; *P*<.001) and use of mechanical ventilation (exp[β] 1.23, 95% CI 1.05-1.43; *P*=.01) were associated with longer ICU LOS, whereas higher patient coverage was independently associated with shorter ICU LOS (exp[β] 0.17, 95% CI 0.13-0.21; *P*<.001). ICU coverage was not independently associated with ICU mortality or ICU LOS.

Conclusions: Greater patient-level coverage by remote intensivist–led multidisciplinary rounds within the Tele-ICU program was independently associated with lower ICU mortality and shorter ICU LOS. These findings support the potential contribution of tele–critical care strategies to expanding specialist support during public health emergencies.

(J Med Internet Res 2026;28:e64996) doi:10.2196/64996

KEYWORDS

coronavirus; COVID-19; SARS-CoV-2; intensive care unit; telemedicine; tele–intensive care unit; Tele-ICU; clinical outcomes; mortality

Introduction

Background

COVID-19 has become a major public health care concern, with almost half a billion cases diagnosed and more than 6 million deaths reported across the globe [1]. During the pandemic, waves of increased number of newly diagnosed patients with SARS-CoV-2 infection were reported worldwide, with varied degrees of disease severity [2]. The rapid spread of SARS-CoV-2 increased the demand for intensive care units (ICUs) and the pressure on health care systems [2].

In Brazil, the high demand for ICU beds coupled with the shortage of intensivists to care for patients who are critically ill—particularly in rural and remote regions far from major urban centers—motivated the search for alternative strategies to deliver efficient care [3]. In this scenario, in a partnership between the Brazilian Ministry of Health (via the Institutional Development Support Program of the Unified Health System) and some of the top-ranked Brazilian hospitals, the Tele-ICU COVID-19 Brazil Program was created to provide remote intensivists to guide daily multidisciplinary rounds (DMRs) in public COVID-19 ICUs from Brazil with patients with COVID-19. The program consisted of DMRs (structured patient-centered discussions) with the remote intensivist and the local ICU multidisciplinary team, aiming to review diagnostic hypotheses, list active problems, and establish joint therapeutic goals.

Different Tele-ICU models exist and vary in structure and intensity. The most common is the centralized “hub-and-spoke” model, where a central team remotely monitors several ICUs [4]. Alternatively, decentralized models allow remote specialists to provide support from multiple locations on demand. These configurations may include continuous monitoring, scheduled virtual rounds, or reactive consultations, depending on institutional needs and available infrastructure [4].

Observational studies indicated that implementing Tele-ICU programs may be an effective strategy for improving clinical outcomes [5-11]. However, reported impacts vary across studies, likely due to differences in Tele-ICU models and applications, including second opinion consultations, real-time evaluations, remotely conducted rounds, and others [12-14]. More recently, a randomized clinical trial evaluating an intensivist-led, telemedicine-based strategy that incorporated DMRs found no significant improvement in ICU length of stay (LOS) or patient mortality [15].

However, evidence regarding the impact of Tele-ICU interventions on clinical outcomes, particularly during large-scale health care crises such as the COVID-19 pandemic, remains limited [7,8]. In this context, this study aimed to evaluate the association of the Tele-ICU COVID-19 Brazil Program and clinical outcomes of patients with COVID-19. We hypothesized that higher adherence to the program’s structured DMRs would be associated with lower ICU mortality and shorter ICU LOS among patients admitted to public ICUs across Brazil during the COVID-19 pandemic.

Objectives

This study aimed to evaluate the association of the adherence to the Tele-ICU COVID-19 Brazil Program and clinical outcomes of patients with COVID-19.

Methods

Study Design

We performed a retrospective study to evaluate the impact of Tele-ICU COVID-19 Brazil Program on clinical outcomes of ICU patients with COVID-19.

Ethical Considerations

The study was approved by Brazilian National Ethics Committee on August 27, 2020, with waiver of informed consent (CAAE: 31459120.0.0000.0071, opinion number: 4.240.138). All data used in the study were deidentified before analysis to protect participant privacy and ensure confidentiality. No compensation or reimbursement was provided to participants.

Tele-ICU Program

The Tele-ICU COVID-19 Brazil Program was led by the Tele-ICU Department of the Hospital Israelita Albert Einstein, a private quaternary care hospital located in São Paulo, Brazil.

XSL•FO
RenderX

Participating ICUs that accepted the invitation (a convenience sample) were located in public hospitals across different geographic regions of the country (1 ICU per hospital) and were designated by the Brazilian Ministry of Health between April and December 2020. The recruitment process occurred at the institutional level and was coordinated directly by the Ministry of Health, which was responsible for identifying, evaluating, and selecting eligible ICUs. ICUs were eligible if they belonged to public or philanthropic hospitals and did not have a board-certified intensivist providing daily on-site care. The Ministry also established criteria for replacing units during the program: ICUs could be removed if they missed daily tele-rounds for 5 consecutive days or for more than 10 nonconsecutive days within a month.

As critical care specialists were not available daily in the participating ICUs during this period, DMRs were conducted entirely by the program's remote intensivists. Discussions were conducted by a board-certified intensivist trained in telemedicine, based at the remote center (tele-intensivist, in São Paulo, Brazil), in collaboration with the local multidisciplinary team (physician, nurse, and physiotherapist, in any part of the country). DMRs took place from Monday to Friday at a predetermined time (mostly during the mornings), from April to December 2020, and approached all patients admitted to the participating ICUs. The main objective of the DMR conducted by the tele-intensivist was to review diagnostic hypotheses, list active problems, and jointly create a treatment plan until the next DMR. Tele-intensivists made recommendations based on updated scientific evidence, suitable to the local context. Clinical protocols in texts and figures were made accessible through a dedicated application developed for the physicians and multidisciplinary team of the participating ICUs.

Study Participants

All the 17 participant ICUs of the program were invited to participate in this study. All patients admitted to these ICUs during the Tele-ICU program, from April to December 2020, were included in this analysis.

We excluded patients who meet the following criteria: (1) aged <18 years, (2) receiving exclusive palliative care at ICU admission, and (3) incomplete data about ICU outcomes.

Data Collection and Study Variables

All study data were collected by the Tele-ICU intensivist during the DMRs performed during the program and complemented by the local staff of the participant ICUs. Data were stored on a server at Hospital Israelita Albert Einstein, São Paulo, Brazil, and retrieved by the study authors after the approval of the ethics committee.

Collected data included demographics, comorbidities, Sequential Organ Failure Assessment (SOFA) score [16], number of DMRs, resource use, and organ support (vasopressors, noninvasive ventilation, and mechanical ventilation) during ICU stay, palliative care, ICU and hospital LOS, and ICU and in-hospital mortality. Variables were collected during the DMRs (and lately complemented by invited professionals of the participant ICU). SOFA score refers to the maximum value of SOFA during ICU stay.

Definitions

To evaluate the impact of the Tele-ICU program, we developed 2 variables: patient coverage and ICU coverage. *Patient coverage* quantifies how often each patient's case was discussed during DMRs based on the hypothesis that more frequent discussions could enhance clinical decision-making and continuity of care. *ICU coverage* reflects the overall participation of each ICU in the program, hypothesizing that higher engagement would promote local learning and greater adherence to evidence-based practices through regular interaction with remote intensivists. These variables were defined as follows:

- Patient coverage=(number of DMRs per patient)/(patient's ICU LOS)
- ICU coverage=(number of DMR days in the ICU)/(total ICU patient-days)

A cutoff of 50% was empirically defined to categorize groups. Patients were classified into *low patient coverage* (<50%) and *high patient coverage* (≥50%) groups based on their individual patient coverage. Similarly, ICUs were categorized into low ICU coverage (<50%) and high ICU coverage (≥50%) groups based on their overall ICU coverage.

Outcomes

Our primary outcome of interest was ICU mortality. The secondary outcomes were ICU LOS and use of resources (eg, mechanical ventilation, noninvasive ventilation, use of vasopressors, and renal replacement therapy).

Statistical Analysis

Categorical variables are presented as absolute and relative frequencies. Continuous variables are presented as median with IQR. Normality was assessed using the Shapiro-Wilk test.

Comparisons were made between groups: low patient coverage group vs high patient coverage group and low ICU coverage group vs high ICU coverage group. Continuous variables were compared using the independent 2-tailed *t* test or the Mann-Whitney *U* test in case of nonnormal distribution. Univariate analyses were performed to identify which predictors were associated with ICU mortality and ICU LOS.

We performed multivariate analyses using multilevel mixed modeling, with hospitals as random effect. All variables tested with a *P* value <.20 in the univariate analyses were included as fixed effects. We performed a multicollinearity analysis before the backward elimination procedure. A mixed model was undertaken to obtain adjusted odds ratio (OR) along with 95% CI and to define which variables were independently associated with ICU mortality.

A linear mixed-effects model with log-transformed ICU LOS as the outcome was used, with hospital as a random effect. Exp(β) along with the 95% CI were obtained to define which variables were independently associated with ICU LOS. To perform the analysis of the predictor ICU LOS, probability distribution analyses of the outcome were conducted. We carried out inferential tests to assess the fit of various probability distributions. The only distribution that demonstrated adherence

XSL•FO
RenderX

was the normal distribution for data transformed by natural logarithm. Thus, the outcome variable ICU LOS was transformed by taking the natural logarithm (ln).

All statistical tests were 2-tailed, and $P<.05$ was considered statistically significant. Analyses were performed using R software (version 4.1.0; R Foundation for Statistical Computing).

Results

From April to December 2020, a total of 1945 patients from 17 different ICUs across different Brazilian geographic regions were included in the database and had their plan of care defined by the remote tele-intensivist on at least 1 day of the Tele-ICU program. We excluded 233 (11.9%) patients who lacked data on ICU outcomes (leading to the exclusion of an entire ICU, as none of the patients from this ICU had available outcome data); 18 (0.9%) patients aged younger than 18 years; and 14 (0.7%) patients receiving palliative care since the ICU admission, as previously defined in our study protocol. Finally, 1680 patients from 16 different ICUs were included in the study.

Baseline characteristics of patients included in the study are presented in Table 1. The median age in this cohort was 66 (IQR 55-76) years, and 56.4% (865/1534) of the patients were men. The median SOFA score was 4 (IQR 1-9), and the median number of DMRs per patient was 3 (IQR 2-6). In this cohort, during ICU stay, 56.8% (769/1354) of the patients required mechanical ventilation, 46.3% (627/1354) of the patients required vasopressors, 16.9% (229/1354) of the patients required renal replacement therapy, and 9.7% (132/1361) of the patients required noninvasive ventilation. ICU mortality was 48.3% (812/1680), and the hospital mortality was 51.1% (847/1658). The median ICU and hospital LOS were, respectively, 8 (IQR 4-15) days and 11 (IQR 6-20) days (Table 1).

XSL•FO
RenderX

Table 1. Baseline characteristics and clinical outcomes of intensive care unit (ICU) patients with COVID-19 admitted to public ICUs participating in the Tele-ICU COVID-19 Brazil Program (retrospective multicenter cohort study, Brazil, April to December 2020; N=1680)a.

Characteristics	Patients
Male, n (%)	865 (56.4)
Age (years), median (IQR)	66 (55-76)
Comorbidities, n (%)	
Systemic hypertension	684 (62.8)
Diabetes mellitus	441 (40.5)
Congestive heart failure	74 (6.8)
Acute cerebral stroke	60 (5.5)
Coronary arterial disease	17 (1.6)
Previous myocardial infarction	21 (1.9)
Asthma	27 (2.5)
COPD[b]	86 (7.9)
Chronic kidney disease	44 (4)
Chronic kidney disease requiring RRT[c]	19 (1.7)
Locoregional cancer	4 (0.4)
Metastatic cancer	5 (0.5)
Hematologic cancer	1 (0.1)
Cognitive impairment and dementia	19 (1.7)
Liver cirrhosis	1 (0.1)
ICU mortality, n (%)	812 (48.3)
Hospital mortality, n (%)	847 (51.1)
ICU LOS[d], median (IQR)	8 (4-15)
Hospital LOS, median (IQR)	11 (6-20)
SOFA[e] score, median (IQR)	4 (1-9)
Number of DMRs[f] per patient, median (IQR)	3 (2-6)
Support on the first day of the Tele-ICU visit, n (%)	
Mechanical ventilation	639 (47.2)
Noninvasive ventilation	132 (9.7)
Vasopressors	425 (31.4)
RRT	128 (9.5)
Support in the first 24 hours of ICU stay, n (%)	
Mechanical ventilation	243 (36.9)
Noninvasive ventilation	71 (10.8)
Vasopressors	164 (24.9)
RRT	26 (3.9)
Support at any time during ICU stay, n (%)	
Mechanical ventilation	769 (56.8)
Noninvasive ventilation	168 (12.4)
Vasopressors	627 (46.3)
RRT	229 (16.9)

[a]Percentages were calculated based on available data for each variable.

XSL•FO
RenderX

[b]COPD: chronic obstructive pulmonary disease.

[c]RRT: renal replacement therapy.

[d]LOS: length of stay.

[e]SOFA score: Sequential Organ Failure Assessment score, ranges from 0 to 24, with higher scores indicating more severe organ dysfunction.

[f]DMR: daily multidisciplinary round.

Compared with the low patient coverage group, patients in the high patient coverage group had shorter ICU LOS (median 6, IQR 3-11 days vs median 11, IQR 6-20 days; *P*<.001) and hospital LOS (median 9, IQR 5-16 days vs median 14, IQR 8-26 days; *P*<.001; Table 2). SOFA score was higher in the low patient coverage group (median 3, IQR 0-6 vs median 2, IQR 0-5; *P*=.007). ICU mortality (50.1% vs 46.3%; *P*=.16) did not differ significantly between the low and high patient coverage groups. During ICU stay, patients in the high patient coverage group used mechanical ventilation less frequently, but used noninvasive ventilation more frequently when compared with the low patient coverage group.

Table 2. Comparison of clinical outcomes and resource use according to patient-level adherence to tele–intensive care unit (ICU) daily multidisciplinary rounds among ICU patients with COVID-19 (Brazil, April to December 2020).

	Low patient coverage group (n=673)	High patient coverage group (n=862)	*P* value
ICU deaths, n (%)	337 (50.1)	399 (46.3)	.16[a]
Hospital deaths, n (%)	351 (52.2)	417 (48.4)	.17[a]
ICU LOS[b], median (IQR)	11 (6-20)	6 (3-11)	<.001[c]
Hospital LOS, median (IQR)	14 (8-26)	9 (5-16)	<.001[c]
SOFA[d] score, median (IQR)	3 (0-6)	2 (0-5)	.007[c]
Support in the first 24 hours of ICU stay, n (%)			
Mechanical ventilation	86 (45.7)	154 (33)	.003[a]
Noninvasive ventilation	13 (6.9)	58 (12.4)	.06[a]
Vasopressors	56 (29.8)	106 (22.7)	.07[a]
RRT[e]	11 (5.9)	14 (3)	.14[a]
Support at any time during ICU stay, n (%)			
Mechanical ventilation	328 (60.3)	357 (51.6)	.003[a]
Noninvasive ventilation	58 (10.7)	108 (15.6)	.01[a]
Vasopressors	261 (48)	301 (43.5)	.13[a]
RRT	105 (19.3)	100 (14.5)	.03[a]

[a]*P* values were calculated using the chi-square test.

[b]LOS: length of stay.

[c]*P* values were calculated using the Mann-Whitney *U* test.

[d]SOFA score: Sequential Organ Failure Assessment score, ranges from 0 to 24, with higher scores indicating more severe organ dysfunction.

[e]RRT: renal replacement therapy.

Compared with the low ICU coverage group, patients in the high ICU coverage group had lower SOFA scores (median 2, IQR 0-4 vs median 3, IQR 1-7; *P*<.001), lower ICU mortality (44.3% vs 52.5%; *P*=.004), lower ICU LOS (median 7, IQR 4-12 days vs median 9, IQR 5-17 days; *P*<.001), and lower hospital LOS (median 8, IQR 5-16 days vs median 14, IQR 7-23 days; *P*<.001). During ICU stay, patients in the high ICU coverage group used mechanical ventilation, noninvasive ventilation, vasopressors, and renal replacement therapy less frequently (Table 3).

XSL•FO
RenderX

Table 3. Comparison of clinical outcomes and resource use according to intensive care unit (ICU)–level adherence to the Tele-ICU COVID-19 Brazil Program (Brazil, April to December 2020).

	Low ICU coverage group (n=844)	High ICU coverage group (n=836)	*P* value
ICU deaths, n (%)	442 (52.5)	370 (44.3)	.004[a]
Hospital deaths, n (%)	451 (53.5)	396 (48.7)	.05[a]
ICU LOS[b], median (IQR)	9 (5-17)	7 (4-12)	<.001[c]
Hospital LOS, median (IQR)	14 (7-23)	8 (5-16)	<.001[c]
SOFA[d] score, median (IQR)	3 (1-7)	2 (0-4)	<.001[c]
Palliative care, n (%)	146 (17.3)	123 (14.7)	.17[a]
Support in the first 24 hours of ICU stay, n (%)			
Mechanical ventilation	142 (44)	101 (30.1)	<.001[a]
Noninvasive ventilation	32 (9.9)	39 (11.6)	.56[a]
Vasopressors	93 (28.8)	71 (21.1)	.03[a]
RRT[e]	24 (7.4)	2 (0.6)	<.001[a]
Support at any time during ICU stay, n (%)			
Mechanical ventilation	471 (63.1)	298 (49)	<.001[a]
Noninvasive ventilation	109 (14.6)	59 (9.7)	.008[a]
Vasopressors	384 (51.5)	243 (40)	<.001[a]
RRT	168 (22.5)	61 (10)	<.001[a]

[a]*P* values were calculated using the chi-square test.

[b]LOS: length of stay.

[c]*P* values were calculated using the Mann-Whitney *U* test.

[d]SOFA score: Sequential Organ Failure Assessment score; ranges from 0 to 24, with higher scores indicating more severe organ dysfunction.

[e]RRT: renal replacement therapy.

Independent predictors of ICU mortality were the use of mechanical ventilation (OR 3.22, 95% CI 2.056-5.043; *P*<.001) and vasopressors (OR 1.87, 95% CI 1.236-2.824; *P*=.003) during ICU stay. The use of noninvasive ventilation (OR 0.54, 95% CI 0.331-0.881; *P*=.01) and patient coverage (OR 0.520, 95% CI 0.272-0.993; *P*=.048) were identified as a protective factors (Table 4). Neither ICU coverage nor total time of ICU participation in the project remained in the final model.

Table 4. Multilevel mixed-effects logistic regression analysis of factors associated with intensive care unit (ICU) mortality among critically ill patients with COVID-19 enrolled in the Tele-ICU COVID-19 Brazil Program (Brazil, April to December 2020).

Fixed effects	OR[a] (95% CI)	*P* value
SOFA[b] score	1.238 (0.860-1.317)	.052
Use of MV[c] during ICU stay	3.222 (2.056-5.043)	<.001
Use of NIV[d] during ICU stay	0.540 (0.331-0.881)	.01
Use of vasopressor during ICU stay	1.870 (1.236-2.824)	.003
Patient coverage	0.520 (0.272-0.993)	.048

[a]OR: odds ratio.

[b]SOFA: Sequential Organ Failure Assessment.

[c]MV: mechanical ventilation.

[d]NIV: noninvasive ventilation.

Independent predictors of higher ICU LOS included in the final adjusted model were SOFA (exp[β] 1.037, 95% CI 1.02-1.05; *P*<.001), use of mechanical ventilation during ICU stay (exp[β] 1.228, 95% CI 1.05-1.43; *P*=.01), and patient coverage as protective factor (exp[β] 0.166, 95% CI 0.13-0.21; *P*<.001; Table 5). As in the first model for ICU mortality, neither ICU

XSL•FO
RenderX

coverage nor total time of ICU participation in the project remained in this final adjusted model.

Table 5. Multilevel linear mixed-effects model evaluating factors associated with intensive care unit (ICU) length of stay among critically ill patients with COVID-19 participating in the Tele-ICU COVID-19 Brazil Program (Brazil, April to December 2020).

Fixed effects	Exp(β)[a] (95% CI)	*P* value
SOFA[b] score	1.037 (1.02-1.05)	<.001
Use of MV[c] during ICU stay	1.228 (1.05-1.43)	.01
Patient coverage	0.166 (0.13-0.21)	<.001

[a]Exp(β): exponentiated regression coefficient from the linear mixed-effects model with log-transformed ICU LOS; values >1 indicate longer ICU LOS, and values <1 indicate shorter ICU LOS.

[b]SOFA score: Sequential Organ Failure Assessment score; ranges from 0 to 24, with higher scores indicating more severe organ dysfunction.

[c]MV: mechanical ventilation.

Discussion

Principal Findings

The main finding of this study was that patients with higher patient coverage by Tele-ICU, defined as a greater proportion of ICU days during which patients were discussed in DMRs led by a remote intensivist during the Tele-ICU COVID-19 Brazil Program, were associated with lower ICU mortality and LOS. We postulate that Tele-ICU may have improved clinical outcomes by implementing multidisciplinary rounds, thereby enhancing adherence to best practices and promoting continuity of care in participating ICUs.

Our results are in line with some previous studies regarding the association between the use of Tele-ICU and lower mortality and LOS [5-10]. For instance, a prospective study found that the implementation of a Tele-ICU intervention, where the Tele-ICU team participated in critical care processes throughout the day, was associated with reduced adjusted odds of mortality and reduced hospital LOS, as well as with changes in best practice adherence and lower rates of preventable complications [17]. Similarly, a before-and-after study observed a notable rise in the percentage of patients receiving a daily sedative interruption with Tele-ICU support, which involved the incorporation of a third shift of Tele-ICU pharmacist assistance [18].

In addition, recently, a cluster randomized controlled trial found that telemedical quality improvement program increased adherence to 7 evidence-based German performance indicators in acute ICU care [19]. The quality indicators included were sedation, analgesia and delirium, ventilation, weaning from ventilation, infection management, enteral nutrition, patient and family communication, and early mobilization [19].

Nevertheless, there is significant heterogeneity among various Tele-ICU models, making these studies potentially noncomparable [8,9]. There are notable discrepancies concerning various characteristics in the previous studies regarding the technology and hospitals used, as well as the autonomy of Tele-ICU intensivists [8,9,18].

Tele-ICU COVID-19 Brazil Program was based on the implementation of multidisciplinary rounds conducted by a board-certified tele-intensivist in ICUs lacking daily critical care specialists. Daily rounds conducted by an intensivist have been previously associated with a 3-fold reduction in hospital mortality in surgical ICUs [20]. Furthermore, a study exploring organizational factors in Brazilian ICUs did not identify a direct impact of regular multidisciplinary rounds on mortality; however, the findings suggest that collaborative multidisciplinary efforts among ICU care providers positively influence patient outcomes [21].

The Tele-ICU program was also linked to an increase in the use of noninvasive ventilation and a decrease in the need of invasive mechanical ventilation. We hypothesize that the adoption of noninvasive ventilation increased as intensivists encouraged its use, and local teams observed that patients responding positively to noninvasive ventilation could safely be managed (especially considering the fear of aerosolization and spread of SARS-CoV-2 in the environment, which was highly prevalent in the first COVID-19 wave).

Interestingly, no independent association was found between overall ICU engagement (ICU coverage) or duration of ICU participation in the program and patient outcomes. Although we initially hypothesized that greater program exposure or broader ICU involvement would lead to better results, our findings suggest that the direct and consistent implementation of daily Tele-ICU rounds may have been the most influential factor, rather than general program duration or institutional engagement.

The program may also have enhanced continuity of care, as the remote intensivist overseeing the rounds remained mostly consistent from Monday to Friday, whereas the local team members often changed daily during the weekdays. Continuity of care is known to be a fundamental aspect of effective medical treatment and is linked to enhanced patient satisfaction, greater use of care services, and reduced health care costs [22]. Maintaining continuity of care in the ICU environment is particularly challenging, especially due to the prolonged duty hours that can impact the well-being of physicians [23]. In this scenario, the Tele-ICU solution emerges as a potential solution, providing a more stable and continuous presence to support ongoing patient care.

Limitations

Finally, our study has some limitations. First, it has a retrospective design, thus reporting associations rather than

XSL•FO
RenderX

cause-and-effect relationships. Second, we used the maximum SOFA score to quantify the severity of illness in patients, which may overestimate the severity of patients with COVID-19 included in our analysis. Third, although our study includes a considerable number of patients from multiple ICUs across different regions, the Tele-ICU program itself was coordinated by a single center, which may limit the external validity of the findings. Fourth, we acknowledge the potential variability in the presence of local intensivists across participating ICUs. While all selected ICUs lacked full-time intensivist coverage, which was a key inclusion criterion, there may have been occasional availability of specialists in some sites. This unmeasured variability may have influenced local care dynamics and should be considered as a potential confounding factor in interpreting the results. Finally, data on tracheostomy procedures were not collected in this study, which prevents evaluation of their potential impact on outcomes such as ICU mortality, duration of mechanical ventilation, and LOS [24]. Therefore, these should be considered when assessing generalizability, and the study should be considered as hypothesis generating.

Conclusions

During the COVID-19 pandemic, the Tele-ICU COVID-19 Brazil Program showed that greater exposure to DMRs led by remote intensivists was associated with lower ICU mortality and shorter ICU stays. These findings suggest that structured tele–critical care support may help extend specialized expertise, promote adherence to best practices, and enhance continuity of care in settings with limited intensivist availability. The results also highlight the potential value of Tele-ICU models in low- and middle-income countries and during periods of health care system strain. Further research is needed to confirm these findings, identify which program components drive the greatest benefit, and evaluate different Tele-ICU configurations and their cost-effectiveness to support future implementation efforts.

Acknowledgments

The authors would like to thank the intensive care unit (ICU) physicians, nursing staff, physical therapists, and all members of the multidisciplinary teams of the participating ICUs and of the Tele-ICU COVID-19 Brazil Program, who managed patients during the COVID-19 pandemic. The authors would also like to thank the Brazilian Ministry of Health for its support of the Tele-ICU COVID-19 Brazil Program. The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the Generative Artificial Intelligence Disclosure and Transparency Taxonomy (GAIDeT) taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: text and image generation. The GenAI tool used was ChatGPT (OpenAI). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes. ChatGPT was used to refine sentence structure and correct grammar and punctuation errors.

Funding

The Tele-ICU COVID-19 Brazil Program was primarily funded by the Brazilian Ministry of Health through the Institutional Development Program of the Unified Health System (PROADI-SUS), which covered the costs of physician (tele-consultant) services. No specific funding was received for this study, and the funder had no role in the study design, conduct, or publication.

Data Availability

The datasets analyzed during this study are not publicly available due to ethical and regulatory restrictions, including requirements from Brazilian regulatory agencies. However, anonymized data may be made available from the corresponding author upon reasonable request, subject to appropriate regulatory approvals.

Conflicts of Interest

None declared.

References

1. Weekly epidemiological update on COVID-19 - 1 September 2023. World Health Organization. 2023 Sep 1. URL: https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---1-september-2023 [accessed 2026-02-27]
2. Van Damme W, Dahake R, Delamou A, Ingelbeen B, Wouters E, Vanham G, et al. The COVID-19 pandemic: diverse contexts; different epidemics-how and why? BMJ Glob Health 2020 Jul 27;5(7):e003098 [FREE Full text] [doi: 10.1136/bmjgh-2020-003098] [Medline: 32718950]
3. Bastos LS, Ranzani OT, Souza TM, Hamacher S, Bozza FA. COVID-19 hospital admissions: Brazil's first and second waves compared. Lancet Respir Med 2021 Aug;9(8):e82-e83. [doi: 10.1016/s2213-2600(21)00287-3]
4. Merola R, Marra A, Simone SD, Vargas M. Telemedicine in intensive care unit: current practice and future prospect. J Intensive Care Med 2025 Apr;40(4):456-463. [doi: 10.1177/08850666251325782] [Medline: 40123239]
5. Rosenfeld BA, Dorman T, Breslow MJ, Pronovost P, Jenckes M, Zhang N, et al. Intensive care unit telemedicine: alternate paradigm for providing continuous intensivist care. Crit Care Med 2000 Dec;28(12):3925-3931. [doi: 10.1097/00003246-200012000-00034] [Medline: 11153637]

XSL•FO
RenderX

6. Thomas EJ, Lucke JF, Wueste L, Weavind L, Patel B. Association of telemedicine for remote monitoring of intensive care patients with mortality, complications, and length of stay. JAMA 2009 Dec 23;302(24):2671-2678. [doi: 10.1001/jama.2009.1902] [Medline: 20040555]
7. Wilcox ME, Adhikari NK. The effect of telemedicine in critically ill patients: systematic review and meta-analysis. Crit Care 2012 Jul 18;16(4):R127 [FREE Full text] [doi: 10.1186/cc11429] [Medline: 22809335]
8. Chen J, Sun D, Yang W, Liu M, Zhang S, Peng J, et al. Clinical and economic outcomes of telemedicine programs in the intensive care unit: a systematic review and meta-analysis. J Intensive Care Med 2018 Jul 22;33(7):383-393. [doi: 10.1177/0885066617726942] [Medline: 28826282]
9. Young LB, Chan PS, Lu X, Nallamothu BK, Sasson C, Cram PM. Impact of telemedicine intensive care unit coverage on patient outcomes: a systematic review and meta-analysis. Arch Intern Med 2011 Mar 28;171(6):498-506. [doi: 10.1001/archinternmed.2011.61] [Medline: 21444842]
10. Fusaro MV, Becker C, Scurlock C. Evaluating tele-ICU implementation based on observed and predicted ICU mortality: a systematic review and meta-analysis. Crit Care Med 2019 Apr;47(4):501-507. [doi: 10.1097/CCM.0000000000003627] [Medline: 30688718]
11. Watanabe T, Ohsugi K, Suminaga Y, Somei M, Kikuyama K, Mori M, et al. An evaluation of the impact of the implementation of the tele-ICU: a retrospective observational study. J Intensive Care 2023 Mar 07;11(1):9 [FREE Full text] [doi: 10.1186/s40560-023-00657-4] [Medline: 36882878]
12. Vranas KC, Slatore CG, Kerlin MP. Telemedicine coverage of intensive care units: a narrative review. Ann Am Thorac Soc 2018 Nov;15(11):1256-1264 [FREE Full text] [doi: 10.1513/AnnalsATS.201804-225CME] [Medline: 30382786]
13. Lilly CM, Fisher KA, Ries M, Pastores SM, Vender J, Pitts JA, et al. A national ICU telemedicine survey: validation and results. Chest 2012 Jul;142(1):40-47. [doi: 10.1378/chest.12-0310] [Medline: 22518025]
14. Fuhrman SA, Lilly CM. ICU telemedicine solutions. Clin Chest Med 2015 Sep;36(3):401-407. [doi: 10.1016/j.ccm.2015.05.004] [Medline: 26304277]
15. Pereira AJ, Noritomi DT, Dos Santos MC, Corrêa TD, Ferraz LJ, Schettino GP, et al. Effect of tele-ICU on clinical outcomes of critically ill patients: the TELESCOPE randomized clinical trial. JAMA 2024 Dec 03;332(21):1798-1807. [doi: 10.1001/jama.2024.20651] [Medline: 39382244]
16. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Med 1996 Jul;22(7):707-710. [doi: 10.1007/bf01709751]
17. Lilly CM, Cody S, Zhao H, Landry K, Baker SP, McIlwaine J, et al. Hospital mortality, length of stay, and preventable complications among critically ill patients before and after tele-ICU reengineering of critical care processes. JAMA 2011 Jun 01;305(21):2175-2183. [doi: 10.1001/jama.2011.697] [Medline: 21576622]
18. Forni A, Skehan N, Hartman CA, Yogaratnam D, Njoroge M, Schifferdecker C, et al. Evaluation of the impact of a tele-ICU pharmacist on the management of sedation in critically ill mechanically ventilated patients. Ann Pharmacother 2010 Mar;44(3):432-438. [doi: 10.1345/aph.1M576] [Medline: 20164471]
19. Spies CD, Paul N, Adrion C, Berger E, Busse R, Kraufmann B, et al. Effectiveness of an intensive care telehealth programme to improve process quality (ERIC): a multicentre stepped wedge cluster randomised controlled trial. Intensive Care Med 2023 Feb;49(2):191-204 [FREE Full text] [doi: 10.1007/s00134-022-06949-x] [Medline: 36645446]
20. Pronovost PJ, Jenckes MW, Dorman T, Garrett E, Breslow MJ, Rosenfeld BA, et al. Organizational characteristics of intensive care units related to outcomes of abdominal aortic surgery. JAMA 1999 Apr 14;281(14):1310-1317. [doi: 10.1001/jama.281.14.1310] [Medline: 10208147]
21. Soares M, Bozza FA, Angus DC, Japiassú AM, Viana WN, Costa R, et al. Organizational characteristics, outcomes, and resource use in 78 Brazilian intensive care units: the ORCHESTRA study. Intensive Care Med 2015 Dec 28;41(12):2149-2160. [doi: 10.1007/s00134-015-4076-7] [Medline: 26499477]
22. Sharma G, Freeman J, Zhang D, Goodwin JS. Continuity of care and intensive care unit use at the end of life. Arch Intern Med 2009 Jan 12;169(1):81-86 [FREE Full text] [doi: 10.1001/archinternmed.2008.514] [Medline: 19139328]
23. Parshuram CS, Amaral AC, Ferguson ND, Baker GR, Etchells EE, Flintoft V, et al. Patient safety, resident well-being and continuity of care with different resident duty schedules in the intensive care unit: a randomized trial. CMAJ 2015 Mar 17;187(5):321-329 [FREE Full text] [doi: 10.1503/cmaj.140752] [Medline: 25667258]
24. Merola R, Iacovazzo C, Troise S, Marra A, Formichella A, Servillo G, et al. Timing of tracheostomy in ICU patients: a systematic review and meta-analysis of randomized controlled trials. Life (Basel) 2024 Sep 14;14(9):1165 [FREE Full text] [doi: 10.3390/life14091165] [Medline: 39337948]

Abbreviations

DMR: daily multidisciplinary round
ICU: intensive care unit
LOS: length of stay
OR: odds ratio

XSL•FO
RenderX

SOFA: Sequential Organ Failure Assessment

Edited by A Mavragani, T de Azevedo Cardoso; submitted 01.Aug.2024; peer-reviewed by D Schnell, R Merola; comments to author 21.Apr.2025; revised version received 19.Dec.2025; accepted 30.Dec.2025; published 19.Mar.2026.

Please cite as:
Midega TD, Hohmann FB, Chaves RCDF, Almeida GMBD, Leoneza VJL, Cabral JFF, Vitalino BV, Camboim EDA, Machado NDJN, Melo RFBD, Feliciano JPO, Cardoso BM, Ferraz LJR, Corrêa TD, Santos MCD, Albaladejo Morbeck R, Pereira AJ
Impact of Tele-Intensive Care Units on the Clinical Outcomes of Critically Ill Patients With COVID-19: Retrospective Cohort Study
J Med Internet Res 2026;28:e64996
URL: https://www.jmir.org/2026/1/e64996
doi:10.2196/64996
PMID:

©Thais Dias Midega, Fabio Barlem Hohmann, Renato Carneiro de Freitas Chaves, Guacyra Margarita Batista de Almeida, Vivian Jaqueline Lima Leoneza, Jennifer Ferreira Figueiredo Cabral, Bianca Veloso Vitalino, Emanuelle de Araújo Camboim, Nelma de Jesus Nogueira Machado, Ricardo Fernando Batista de Melo, Jorge Patrick Oliveira Feliciano, Breno Mendes Cardoso, Leonardo José Rolim Ferraz, Thiago Domingos Corrêa, Maura Cristina dos Santos, Renata Albaladejo Morbeck, Adriano José Pereira. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Supporting Access to Care Through Peripheral Devices and Patient-Generated Health Data: Qualitative Study

Ashley C Griffin[1,2], PhD; Caroline Gray[1], PhD; Victoria Ngo[3], PhD; Camila Chaudhary[1], MPH; Stephanie L Shimada[4,5,6], PhD; Amanda M Midboe[1,7], PhD; Donna M Zulman[1,8], MD, MS

[1]Center for Innovation to Implementation, VA Palo Alto Health Care System, Menlo Park, CA, United States

[2]Center for Biomedical Informatics Research, School of Medicine, Stanford University, Stanford, CA, United States

[3]New England Geriatric Research Education and Clinical Center, VA Bedford Healthcare System, Bedford, MA, United States

[4]Center for Health Optimization & Implementation Research, VA Bedford Healthcare System, Bedford, MA, United States

[5]Department of Health Law, Policy, and Management, Boston University School of Public Health, Boston, MA, United States

[6]Department of Population and Quantitative Health Sciences, University of Massachusetts Chan Medical School, Worcester, MA, United States

[7]Department of Public Health Sciences, Division of Health Policy and Management, School of Medicine, University of California Davis, Davis, CA, United States

[8]Division of Primary Care and Population Health, Department of Medicine, School of Medicine, Stanford University, Stanford, CA, United States

Corresponding Author:
Ashley C Griffin, PhD
Center for Innovation to Implementation
VA Palo Alto Health Care System
795 Willow Road
Menlo Park, CA, 94025
United States
Phone: 1 650 493 5000
Email: griffina@stanford.edu

Abstract

Background: In 2016, the US Department of Veterans Affairs (VA) implemented a national initiative to distribute video-enabled tablets and peripheral devices, such as blood pressure monitors and weighing scales, to patients facing geographic, clinical, or socioeconomic challenges. Such patients could potentially benefit from health monitoring in conjunction with video-based care, as peripheral devices offer opportunities to enrich care received during a video visit and support tracking of health-related data collected outside of clinical care, or patient-generated health data. However, little is known about experiences with the devices and how they could support improved access to care.

Objective: We explored patients' experiences with VA-issued peripheral devices and their impact on video-based care and health monitoring outside of clinical visits.

Methods: We conducted in-depth semistructured interviews among patients who received VA-issued tablets and peripheral devices between 2023 and 2024. Purposive sampling was used to gather views based on gender, age, race or ethnicity, and rurality. Interviews were transcribed and analyzed using rapid qualitative analysis, guided by the Unified Theory of Acceptance and Use of Technology.

Results: Among 25 patients, most received a blood pressure monitor (21/25, 84%), a weight scale (14/25, 56%), and/or a pulse oximetry device (12/25, 48%). The majority reported using their peripheral devices (23/25, 92%) and tablets (19/25, 76%) to monitor their vital signs and attend video visits. Qualitative analysis yielded ten themes reflecting experiences and impacts of the devices, organized by the Unified Theory of Acceptance and Use of Technology constructs: "effort expectancy" consisted of (1) familiar and easy to use devices and (2) challenges of Bluetooth pairing and measurement; "performance expectancy" consisted of (3) integration with video visits, (4) health monitoring for peace of mind, (5) perceptions of improved vital signs and lifestyle behaviors, (6) removing obstacles to in-person care, and (7) desiring an overall picture of health; "social influence" consisted of (8) fostering care team connections and (9) promoting awareness of tablets and peripheral devices; and "facilitating conditions" consisted of (10) supportive help desk infrastructure. Overall, patients described using peripheral devices during virtual visits by syncing data to the tablet for real-time access by their care team. They also reported manually tracking and sharing patient-generated health data with their care team. Despite some challenges with Bluetooth pairing, patients found the devices easy to use and

XSL•FO
RenderX

contributed to improved health and motivation. Devices also reduced logistical burdens of in-person visits, especially for those with limited mobility, visual impairments, mental health needs, or transportation barriers.

Conclusions: Patients perceive that peripheral devices can enhance video-based care and support health care access and chronic disease management. Patients reported benefits to health, behavior, and communication with care teams. To maximize the impact, program enhancements should prioritize device interoperability, accessible training, and expanded outreach.

(J Med Internet Res 2026;28:e88405) doi:10.2196/88405

KEYWORDS

remote patient monitoring; self-management; health services accessibility; patient generated health data; veterans; mobile phone

Introduction

Background

Multiple forces are aligning to integrate patients' remote health data into care settings, showing strong potential to improve health outcomes. With 91% of adults in the United States owning a smartphone [1], the proliferation of consumer health technologies presents numerous opportunities for patients to track and monitor their health through apps and peripheral devices, such as blood pressure monitors, weighing scales, or pulse oximetry devices. A growing proportion of the US population is estimated to use a wearable device to monitor their health (40%), of which the majority (73%) are willing to share these data with their care team [2]. In addition to patient interest in tracking and sharing their health data, substantial progress has been made in the remote monitoring regulatory and interoperability ecosystems [3-5]. Notably, the Fast Healthcare Interoperability Resources standard [6] provides a framework to exchange data between health information systems, devices, apps, and electronic health records (EHRs). These regulatory and technological advancements have given rise to a plethora of patient-generated health data (PGHD), or data created and recorded by patients outside of the clinical setting [7].

Efforts within health systems to leverage PGHD are beginning to flourish [8-11], yet numerous studies have identified digital health disparities in access and use of peripheral devices [12-14]. Patients with lower socioeconomic status, limited internet access or connectivity, barriers to care, and older adults have been most impacted by the digital divide. These disparities can impact care and outcomes, as the use of peripheral devices and PGHD has demonstrated early value in personalized medicine and improvements in physiological measures and symptoms across a variety of chronic conditions [9,10,15,16]. Given the value of peripheral devices, system-wide efforts are needed to support greater access and more equitable use to allow all patients to receive these benefits.

To combat digital disparities, the US Department of Veterans Affairs (VA) distributes internet-connected tablets and peripheral devices to VA patients with barriers to accessing in-person care [17]. Patients are eligible for a tablet if they lack a device or connectivity, have a barrier to care (ie, geographical, clinical, or social), and require a VA video-based appointment within the next 90 days. In addition to the tablet, depending on the patient's clinical needs, they can receive 1 or more peripheral devices including a blood pressure monitor, pulse oximetry device, weighing scale, stethoscope, or thermometer. Peripheral devices are intended to be connected to the tablet via Bluetooth and used during a video visit to share a patient's vital signs with the care team. The tablets also contain a variety of preinstalled VA apps, allowing patients to engage with the VA patient portal or VA Mobile App Store [18]. Previous work has shown increased registration and use of VA's patient portal after receiving a tablet [19] and demonstrated how tablets foster opportunities to use VA apps for multiple health management needs [20].

Despite growing evidence supporting remote monitoring and PGHD integration, little is known about patient experiences using peripheral devices in conjunction with video-enabled tablets, and how they use these devices outside of clinical visits to collect and track health data. We also have a limited understanding of how these technologies are experienced by patients who face barriers to care. Previous work has examined patients' expectations and experiences with remote monitoring across various chronic conditions, highlighting how patients gain knowledge about their condition, feel reassured about their health, and how the devices support self-management and shared decision-making [21]. Patients have also reported reluctance to use new technologies, lack of trust in technology, concerns about costs, and fear of their data being lost or losing face-to-face communication with their care team [21]. Much of the existing telehealth and remote monitoring literature has focused on disease-specific interventions or on patient populations who already have access to technology [22-24].

National health system initiatives designed to mitigate these inequities, such as the VA tablet and peripheral device distribution program, aim to address structural barriers to care by providing devices, connectivity, and technical support to patients with documented challenges [17]. Previous evaluations of this initiative have demonstrated improvements in access to mental health services, continuity of care, and reductions in emergency department visits among VA patients receiving tablets [25,26]. These evaluations have mostly focused on usage and clinical outcomes rather than the patient's lived experience using peripheral devices and the data generated from them. We extend existing work within the context of patients who experience barriers to accessing health care services, including physical and mental health challenges, socioeconomic stressors, and transportation difficulties. Patients with these challenges potentially stand to gain substantial benefits from peripheral devices.

XSL•FO
RenderX

Objective

Among VA patients who received a tablet and 1 or more peripheral devices, we aimed to assess their experiences with the devices, including setup, use, technical and social support, and integration with their video-enabled tablet. We also examined how patients used PGHD generated from the peripheral devices to enhance their health, including through health monitoring, communication with care teams, and access to care.

Methods

Setting

The VA Office of Rural Health and Office of Connected Care tablet initiative provides tablets (ie, Apple iPads with T-Mobile [Deutsche Telekom] or Verizon [Verizon Communications Inc] data plans) and peripheral devices to VA patients based on an assessment by clinicians and social workers using a Digital Divide Consult template in the VA's EHR [27]. As of March 2026, more than 10,500 VA patients have been shipped both a peripheral device and a tablet. The primary goal of peripheral device distribution is to use the devices during a video visit, in which devices are paired via Bluetooth to the tablet. During the video visit, data are displayed to the care team within the VA Video Connect app and can be entered into the EHR. Devices can also be used without Bluetooth, but data are not shared with VA care teams when used without Bluetooth. Upon receipt of a device, patients receive a phone call from the VA Connected Device Support Team to assist with setting up and pairing the Bluetooth device with the tablet.

Ethical Considerations

This evaluation was conducted as part of the Virtual Access Quality Enhancement Research Initiative and was designated as nonresearch by the Stanford University institutional review board (#70128). All patients provided consent to participate in the interview and received a US $50 gift card upon completion. All recruitment and interview data were stored in a private and secure location to ensure confidentiality.

Sample and Sampling

To identify patients for in-depth semistructured interviews, we used EHR data from the VA Corporate Data Warehouse (CDW) that included sociodemographic and peripheral device characteristics. Patients were eligible if they were shipped a peripheral device between February 2023 and February 2024, and had a physical mailing address, email address, and phone number on file. We used purposive sampling to include perspectives from at least 25% (7/25) of women, older adults (older than 65 y), patients of diverse race or ethnicity, and those who resided in a rural area. Rurality was based on Census tract Rural Urban Commuting Areas codes corresponding to each patient's ZIP code [28].

We also aimed to gather views among patients who received different devices (ie, blood pressure cuff, weighing scale, pulse oximetry, thermometer, and stethoscope), those who received a device setup call from the VA Connected Device Support Team, and those who received a peripheral device and tablet but did not use it. Setup calls were tracked on a VA dashboard and then linked to our sample in the CDW. Use of the peripheral device and tablet was reported during the interview prescreening process and noted in our recruitment sheet. We also characterized the sample based on the most common chronic physical and mental health conditions from the CDW based on the list from Yoon et al [29].

Recruitment

Patients were first contacted through an email explaining the project and asking if they were interested in participating. If they expressed interest, they received a phone call to confirm that they had obtained a peripheral device and to schedule an interview. If there was no response to the initial recruitment email, 2 follow-up phone calls were attempted. Recruitment occurred between March 2024 and August 2024.

Interview Guide Development

To examine patients' experiences with the devices, an interview guide was developed using constructs from the Unified Theory of Acceptance and Use of Technology (UTAUT) [30], which is a well-established framework to assess acceptance and behavioral intention to use health technologies. Interview questions related to UTAUT constructs were developed to elicit input on various topics, including device setup and use, technical and social support, communication with the care team, integration with the tablet, use of data generated from the device, and improvements in health (Table 1).

XSL•FO
RenderX

Table 1. Interview questions mapped to constructs from the Unified Theory of Acceptance and Use of Technology.

Construct	Definition applied to this study	Questions
Effort expectancy	The degree of ease associated with using the peripheral device	• What was your experience like setting up the device? • What was it like learning to use the device? • How often do you use it?
Performance expectancy	The degree to which a patient believes that using the peripheral device will help them to improve their health	• Do you consider the device to be useful? (Why or why not?) • Have you used the device before or during a video visit or in-person visit with your care team? (If so, please describe how you used the device.) • How did receiving a device help you talk with your doctor? • How have you used the data generated from the device? Has it resulted in any lifestyle changes? • In your opinion, can the device improve your [blood pressure, weight, other health monitoring]?
Social influence	The degree to which a patient perceives that important others believe they should use the peripheral device	• Who is supportive of you using the device? • Do you have anyone in your support circle (family or friends) who helps you use the device?
Facilitating conditions	The degree to which a patient believes that organizational and technical infrastructure exists to support use of the peripheral devices	• Have you had prior training on how to set up and use the device? • What did you find most challenging about setting up the device? Using the device? • Are you able to get technical support on the device? • What other information or support do you need to use the device?

Interviews

Interviews were conducted by 2 trained qualitative researchers (ACG and CG) using Microsoft Teams software. Each interview included the interviewer and patient, with the exception of 1 interview that also involved the patient's caregiver and 1 interview with a late patient's family member. Thus, 26 individuals were interviewed to represent the 25 patients. Demographic information is presented for the 25 patients who received the peripheral devices. Interviews lasted for 30-60 minutes and were audio and/or video-recorded upon consent. Interviews were completed between April 2024 and August 2024.

Analysis

Interview data were transcribed verbatim by Keystrokes Transcription Service and imported into ATLAS.ti qualitative data analysis software (version 23; Lumivero). Patient narratives within the transcriptions were analyzed using rapid qualitative analysis, guided by the Planning for Assessing Rigor in Rapid Qualitative Analysis framework [31]. One team member (CG) reviewed the transcripts and interviewer's notes to summarize the content into a matrix based on the initial topics from the interview guide using ATLAS.ti. The matrix was then exported into Microsoft Excel, where a second team member (ACG) reviewed the data summaries and referenced the notes and recordings as needed. The resulting summaries were then discussed between the 2 team members (ACG and CG) and iteratively grouped into themes within Excel using a hybrid inductive-deductive approach to allow for emergent themes. The larger project team also provided input on the themes. Illustrative quotes were identified within each theme. To organize the themes and allow for comparison with related studies, we grouped each theme under the most relevant UTAUT construct.

The study team actively engaged in ongoing reflexive practices throughout data collection and analysis. We are aware that our personal and professional backgrounds shape interactions with patients and the interpretation of findings. Similarly, our role as researchers affiliated with the VA may influence patients' social desirability by potentially overreporting positive experiences or underreporting challenges with the devices because they perceive the interviewers as part of their care system. The study team had no established relationships with patients, and interviews were conducted using open-ended questions and neutral language. Iterative discussions with our multidisciplinary study team helped us remain aware of our positioning and avoid assumptions when interpreting the data.

Results

Patient Characteristics

Among 156 patients contacted, 25 (16%) participated in an interview. The mean age was 62 (SD 11) years, 32% (8/25) were female, 40% (10/25) were Black and/or African American, and 44% (11/25) lived in a rural area. Patients had a range of common chronic physical and mental health conditions (Table 2). Based on the EHR Digital Divide Consult, (1) a mental health diagnosis or participation in therapy (7/25, 28%) or (2) living farther than 30 miles from a VA Medical Center (6/25, 24%) were the most common reasons for receiving a tablet and peripheral device. The majority received a blood pressure monitor (21/25, 84%), weighing scale (14/25, 56%), and/or pulse oximetry device (12/25, 48%) from the VA (Table 2). Nearly all (23/25, 92%) reported using at least one of their peripheral devices. The 2 patients who were unable to use their peripheral devices received only a blood pressure monitor and tablet. They explained that the blood pressure monitor was missing a hardware component (ie, arm cuff and display

XSL•FO
RenderX

monitor) that prevented them from taking their blood pressure. Others reported using some of their devices but not all of them due to difficulty with Bluetooth setup, already having their own device at home, or not being sure why they received certain devices. Approximately three-quarters (19/25, 76%) reported using their tablet to do a video visit with their care team. Of the 18 patients who used both the tablet and peripheral devices, 9 (50%) were able to successfully connect them via Bluetooth so that their care team could access data in real-time during a video visit. Of the 6 patients who had not done a video visit, 3 described being unable to operate the tablet despite receiving a phone call from the VA Connected Device Support Team, 2 patients did not recall receiving a support call and were unable to start the video visit on their own, and 1 preferred using their desktop for video visits. According to the VA Connected Device Support Team dashboard, most received assistance to set up the peripheral devices and/or tablet, either by the VA initiating contact (15/25, 60%) or the patient initiating contact (6/25, 24%).

XSL•FO
RenderX

Table 2. Sociodemographic, clinical, social, and technological characteristics of patients who received a peripheral device (n=25).

Characteristics	Patients, n (%)
Sociodemographic	
Age group (y; mean 62, SD 11 y)	
18-44	3 (12)
45-64	10 (40)
65-74	9 (36)
75-84	3 (12)
Sex	
Female	8 (32)
Male	17 (68)
Race or ethnicity	
Black and/or African American	10 (40)
Hispanic or Latino	2 (8)
More than one race	1 (4)
Native American or Alaska Native	1 (4)
White	10 (40)
Unknown	1 (4)
Rurality	
Rural	11 (44)
Urban	14 (56)
Clinical and social	
Physical health conditions	
Hypertension	17 (68)
Heart failure, heart disease, or atrial fibrillation	9 (36)
Hyperlipidemia	16 (64)
Diabetes	9 (36)
Obesity	8 (32)
Lower back pain, fibromyalgia, or chronic migraine	17 (68)
Rheumatoid arthritis or osteoarthritis	6 (24)
Cataract, glaucoma, or visual impairments	8 (32)
Mental health conditions	
Depression	14 (56)
Anxiety	13 (52)
PTSD[a]	13 (52)
Drug or alcohol use disorder	6 (24)
Digital divide consult reasons (clinicians can select multiple reasons)	
Mental health diagnosis or participating in therapy	7 (28)
Lives father than 30 miles from VA[b] medical center	6 (24)
Health issues make the patient homebound	4 (16)
Difficulty attending visits at VA facility (eg, psychological distress or immunocompromised)	4 (16)
Work, school, or caregiver commitments make in-person visits challenging	4 (16)

XSL•FO
RenderX

Characteristics	Patients, n (%)
Other health issues (ie, hospitalized in last 90 days, hospice diagnosis, requires frequent video visits)	4 (16)
Other reasons (ie, social isolation, costs, or difficulty with public transport)	3 (12)
Technological	
Peripheral devices (patients can receive multiple devices)	
Blood pressure monitor	21 (84)
Weighing scale	14 (56)
Pulse oximetry device	12 (48)
Thermometer	7 (28)
Stethoscope	5 (20)
Ever used a peripheral device	23 (92)
Ever had a video visit on tablet	19 (76)
Tablet and/or peripheral device setup call	
VA-initiated setup complete	15 (60)
Patient-initiated setup complete	6 (24)
Patient requested in-person setup at local VA	2 (8)
Declined assistance or did not answer	2 (8)

[a]PTSD: posttraumatic stress disorder.

[b]VA: US Department of Veterans Affairs.

Experiences With Peripheral Devices

Overall, patients described using the peripheral devices for a variety of purposes, ranging from monitoring blood pressure post surgery, participating in virtual diabetes management classes, and engaging in home-based primary care programs, among other medical reasons. Patients reported that the devices offered many benefits when used in combination with the tablet, which was the intended use of receiving a peripheral device as part of the VA's national telehealth initiative. Patients also described valuable aspects of using the devices in between visits and having the ability to track and reflect on their health data over time. They discussed how the devices were simple to set up and use, and they highlighted opportunities to improve their distribution and usability. The following interviews revealed 10 themes reflecting patients' insights about and impacts of using the peripheral devices (Textbox 1).

XSL•FO
RenderX

Textbox 1. Themes for patient experiences with peripheral devices corresponding to constructs from the Unified Theory of Acceptance and Use of Technology.

Effort expectancy

- Familiar and easy to use devices
- Challenges of Bluetooth pairing and measurement

Performance expectancy

- Integration with video visits
- Health monitoring for peace of mind
- Perceptions of improved vital signs and lifestyle behaviors
- Removing obstacles of in-person care
- Desiring an overall picture of health

Social influence

- Fostering care team connections
- Promoting awareness of tablets and peripheral devices

Facilitating conditions

- Supportive help desk infrastructure

Effort Expectancy

Familiar and Easy to Use Devices

The devices were easy to set up and use, and patients discussed how they had used similar devices before. One patient described,

> *It was really easy. The first couple of times I learned it and just kept track of everything… I did that [set up] on my own, and the instructions were pretty easy.* [P5]

Among the 2 caregivers, both reported using similar devices and felt they were easy to operate. The caregiver of a home-based patient explained,

> *I've been taking care of him for a long time because I'm his caregiver plus his wife… I check his blood pressure, his temperature, and write it down in my little tablet. I'm up-to-date on it. You get the hang of everything after a while.* [P2]

Challenges of Bluetooth Pairing and Measurement

The greatest challenge with using the peripheral devices was connecting them to the iPad (Apple Inc) via Bluetooth. One patient felt,

> *It was just an issue connecting it… And I got frustrated and just used my own scale.* [P10]

Bluetooth pairing issues were mentioned across all types of peripheral devices, and some patients were unsure if their devices were Bluetooth-enabled. Patients who were unable to pair the device felt that having a Bluetooth connection would be useful so that their care team could see their readings.

There were also some concerns about the accuracy of the devices. Sometimes patients felt their blood pressure readings were too high or too low and brought in their devices to the clinic for help. One person reported,

> *I think I had been using [the blood pressure cuff] wrong at first because I kept getting these really low readings. So I called them and told them it didn't work, and when I went in and got to see the nurse in one of my appointments, she showed me how I should be using it. Now that I use it right and more regularly, I feel comfortable about the fact that I'm tracking my blood pressure regularly.* [P8]

It was recommended to have the option for classes to set up and understand how to take accurate measurements,

> *Sometimes you need to understand a little more about your blood pressure than just the cuff and the button that you push because it makes a difference on your feet being on the floor and stuff like that.* [P24]

Performance Expectancy

Integration With Video Visits

Patients described using the peripheral devices during video visits in which the devices were paired via Bluetooth to the tablet. This allowed patients to take their vital signs during the visit and display them automatically on the screen for their care team. These patients were usually part of a VA program, such as home-based primary care or a diabetes management program. One patient explained,

> *The weight scale is Bluetooth connected to the iPad. So when I step on it, it will just tell me to step off, and it'll put my weight on the iPad for the provider that I'm talking to so that they get the information. The same thing with the arm cuff… It automatically shows up on the screen.* [P4]

XSL•FO
RenderX

Patients could also take their vitals right before the visit, with one patient discussing how,

> *I describe my data to my care team and it automatically sends it to them too. It's all connected in one system there with the tablet.* [P16]

The stethoscope and thermometer, while less frequently discussed compared with the other types of devices, were commonly used in real-time during video visits.

Health Monitoring for Peace of Mind

Patients found the devices helpful for monitoring and feeling in control of their health. One patient stated,

> *I feel better and less anxious about my health because I have some control over some of the things. I don't have to wait for an appointment to get my blood pressure checked or to get weighed in and see if I gained or lost weight. So, it's just a feeling of more control and it reduces my anxiety and stress.* [P8]

The devices also provided reassurance for some patients that their medications were working properly and helped them feel less stressed about their blood pressure. Patients mentioned the benefits of monitoring their blood pressure after starting new medications (eg, steroids) and for certain medications that they were instructed not to take when their blood pressure was too low (eg, beta blockers).

The blood pressure monitor, weighing scale, and pulse oximeter were commonly used on a daily basis, such as to check blood pressure before taking medicine, before pacemaker battery replacement, or if something felt off. One patient explained,

> *On the oximeter, sometimes [I use it] two or three times a day. When I come back in from getting the groceries or whatever, and I have a hard time breathing, I take my oxygen level.* [P6]

Patients discussed how being able to take their vitals at home gave them comfort in helping to identify or distinguish between different health issues. For example, one patient described using the pulse oximeter to determine the difference between a panic attack versus shortness of breath for physical reasons. They explained how if they were experiencing mild shortness of breath, but their oxygen levels were high, they could use techniques to calm down and then let their doctor know about the episode. One patient also described using the blood pressure cuff to identify a headache from elevated blood pressure versus a migraine,

> *I get migraines but also get headaches from high blood pressure. So, [the blood pressure monitor] identifies just how high it is, do I need to call someone, and if it's a headache due to high blood pressure… If my blood pressure is high enough for a headache, I need to go in.* [P19]

Perceptions of Improved Vital Signs and Lifestyle Behaviors

Patients reflected on how the devices made a positive impact on their health and lifestyle behaviors. For example, monitoring blood pressure and weight fostered other healthy habits, such as eating less salt, drinking more water, and walking more. One patient reported,

> *I have increased my walking and watched what I'm eating, and last month I went to my doctor – he said I had lost 12 pounds.* [P17]

Others self-reported losing weight or lowering their blood pressure after receiving the devices. Patients also elaborated on how the devices strengthened their motivation to improve their health and helped them stay accountable since a member of their care team was checking on the data. Notably, one person indicated they were able to be taken off their current blood pressure medication, and another described how they did not have to start a medication.

Removing Obstacles of In-Person Care

Patients described how taking their vitals at home through the peripheral devices helped them avoid the challenges of attending in-person care. Patients described having mobility and visual impairments, needing to transport medical equipment (eg, oxygen tanks and wheelchairs) or arrange ambulatory transportation. One patient elaborated on using the peripheral devices in combination with video visits, which avoided some of the in-person challenges with their wheelchair.

> *I'm just very, very thankful that I can go through a video chat rather than have to load my powerchair up every time… it's kind of rough for me to get all that hooked up, and go up there and unhook and then come out and hook it back up again. And then if it's raining, I've got to fight with the cover that goes over the chair so it doesn't ruin.* [P16]

Other patients mentioned how attending in-person visits required them to schedule time off from work or drive long distances, which was especially challenging for some patients with anxiety.

> *I live in a smaller suburban area and most of my appointments are in the city, so that really gets my anxiety up… Sometimes it's more beneficial to be able to be in the safety of your own little habitat, to stay and do your appointment there. Sometimes that's beneficial because you're not distracted and all worked up and anxious by the time you get there. And then you have to spend half the appointment trying to calm down and focus on what you're doing there and nervous about having to make the drive back.* [P25]

Desiring an Overall Picture of Health

Patients were interested in viewing combined visualizations of their health within 1 interface, including their weight trends, blood pressure readings, and other vital signs. One patient described,

> *It'd be nice to be able to see something that I could reflect on over time, even if it was just six months at a time, to see the trends and how it's going because I suspect that everything's going to be worse when I'm depressed... Opening an app and putting that information in would be really awesome because then I could look at the whole entire picture.* [P19]

XSL•FO
RenderX

Patients also reported tracking their peripheral device data in combination with personal device data, such as physical activity, nutrition, or glucose, to reflect on improvements in their health. Another patient stated, "*knowledge is power*" [P24] when referring to their interest and perceived benefit in tracking and visualizing data from devices over time.

Social Influence

Fostering Care Team Connections

Patients explained how using the devices supported communication and connection with their care team. One patient stated,

> *I believe this [blood pressure monitor] device and having the connection with the physical therapist from cardiology improved my desire to get better. It had an immense effect on not just my mental side but the physical side, as well. And at that time, it had more to do with my mental health more than even the physical health.* [P23]

Patients also communicated with their care team as a result of data from the devices, such as asking for advice about fluctuating blood pressure measurements or weight. Patients also discussed how a combination of technologies supported communication with their care team. For example, the VA's My Health*e*Vet patient portal was used to communicate with their care team about data from the devices, such as sending messages or screenshots of their data when they had questions. Others reflected on reviewing their personal health information in the portal and their communication with their care team.

Promoting Awareness of VA-Issued Tablets and Peripheral Devices

Patients heard about receiving the tablets and devices from other veterans, which prompted them to reach out to their care team. For example,

> *I found out about [the devices] through another veteran, and I called right then and got the ball rolling… I try to spread the information as much as I can… The people in my building don't know that they can get a tablet and make it easier to communicate with the hospital.* [P24]

Patients recommended that more veterans need to know about this program and that their care team should discuss the opportunity to provide a tablet and any needed peripheral devices to support the veteran.

Facilitating Conditions

Supportive Help Desk Infrastructure

Patients described receiving a phone call from the VA help desk to assist with setting up the peripheral devices to ensure they could take their vitals at home and start a video visit. Patients found the help desk personnel to be very helpful and effective in training them to use the technologies, with one patient stating how,

> *I feel more confident now that when I go to log on I'm going to get right on [to the video visit].* [P8]

Patients could also contact the help desk at any time, as a contact number was provided on the back of the peripheral devices. Patients described calling the help desk for support with pairing the device via Bluetooth but were usually not able to set up the Bluetooth connection over the phone. Overall, patients did not need any ongoing technical support with the devices but reported knowing how to reach out to the help desk if any technical issues came up.

Discussion

Principal Findings

This qualitative study examined the experiences of patients who received VA-issued peripheral devices and tablets as part of a national telehealth initiative designed to address access barriers to care. Overall, nearly all patients who were interviewed reported using the peripheral devices and discussed the benefits for their health management, improved communication with their care team, and how the devices reduced their burden of attending in-person health care services. Patients frequently used peripheral devices to supplement virtual appointments by syncing data to the tablet for real-time clinician access or by reporting the data to their clinician during the appointment. Despite some usability challenges with Bluetooth data transmission, patients found the devices easy to use and integrated them into their daily routine. These findings support the role of peripheral devices and PGHD in supporting care and reducing access disparities among patients with complex medical and mental health conditions.

Our findings align with previous telehealth and remote monitoring literature, including the enabling role of home-based monitoring, the importance of technical and social support, and ongoing challenges related to interoperability [9,21,22]. Our study, however, extends existing research by demonstrating how and why these dynamics operate within a unique patient population supported by a national digital equity intervention. Unlike previous literature that examines early adopters or patients already engaged in digital health, our cohort consisted of patients identified through an EHR digital divide assessment as having significant barriers to accessing in-person care. These barriers included limited previous access to digital technologies, high prevalence of mental health conditions, mobility limitations, rural residence, and transportation challenges. Although these characteristics have historically been associated with lower engagement in telehealth and patient portals [14,22,32], we found that most patients engaged with the peripheral devices and incorporated them into day-to-day health management.

Guided by UTAUT, we identified several aspects that support the adoption and use of peripheral devices. Patients found the devices were helpful in managing their health (performance expectancy) and described numerous health benefits, including reduced anxiety, increased confidence in their medications, and improved diet and physical activity. These findings resonate with existing work showing that PGHD integration can increase self-management and activation, thereby leading to lifestyle changes and better adherence to treatment plans and medication regimens [33]. Overall, the devices were simple to set up and

XSL•FO
RenderX

operate (effort expectancy), which may be attributed to support from the VA's help desk (facilitating conditions) that was used by 84% (21/25) of patients. However, challenges with Bluetooth pairing and data syncing emerged as key concerns. These technical hurdles of usability and interoperability have been identified in previous studies as limiting factors for the effective adoption of peripheral devices [9,12]. Our findings provide new insight into how PGHD is used in real-world settings where technical integration is incomplete and suggest that PGHD can still retain clinical and relational value despite technical challenges. We also found that care team and social connections shape a patient's engagement with technology, including relationships with the care team, caregivers who assisted in device use, and fellow veterans who promoted awareness of the program (social influence), in accordance with existing work [34-36].

Peripheral Devices and PGHD Play a Role in Reducing Barriers to Care

Patients described how peripheral devices and tablets mitigated logistical barriers to in-person care. Patients with physical disabilities, mental health concerns, or long travel distances emphasized how home monitoring tools reduced the burden of attending in-person medical visits. These experiences reinforce the potential of virtual care to reduce care fragmentation and expand access to care for patients who face challenges in receiving in-person care. Patients expressed that avoiding stressful travel to appointments improved their ability to focus during visits. This is particularly important for patients with conditions that may impact driving or navigating public transportation, such as anxiety, limited mobility, or visual impairments. Our qualitative findings build upon our team's earlier quantitative evaluations of the VA's tablet distribution program, which showed reduced emergency department visits, improved continuity of mental health care, and demonstrated how tablets are reaching veterans from multiple high-need groups [25,26,37]. They also speak to how remote care and PGHD can foster more equitable access to care [13,38].

Patients also discussed using PGHD to spark conversations with their clinicians. In contrast to previous work documenting patient concerns with peripheral devices leading to less face-to-face time with care teams [21], patients in this study described how using their data helped facilitate more frequent and productive communication with care teams. Patients also noted improved understanding of data presented in the My Health*e*Vet patient portal, which highlights the complementary value of access to different digital tools [19,20]. These findings also touch upon considerations for integrating more technical and health literacy education into remote monitoring programs, as patients expressed interest in more in-depth training on device measurement techniques.

Opportunities for System Integration

An emergent theme in our study was the desire for an integrated platform that could aggregate and visualize data from different devices over time to facilitate self-management and health insights. Although the key focus of this telehealth initiative was for patients to use the peripheral device during a video visit, patients desired additional ways to use their data in meaningful ways. For example, patients expressed interest in tracking trends across multiple health domains (eg, blood pressure, glucose, and weight), which could enhance their understanding of personal health trajectories. These suggestions reflect a broader push in the informatics community toward integrated personal health records and highlight the importance of designing tools that offer patients a comprehensive view of their health [6,15].

Recently, the VA has prioritized integration of PGHD into research and clinical efforts, with the development of an enterprise-wide PGHD database that incorporates third-party apps, wearables, medical devices, and patient portal data [11]. Taking an interoperable approach, PGHD are standardized using Fast Healthcare Interoperability Resources to ensure data consistency across the different data sources so that it can be effectively linked with VA data. Robust system integration of heterogenous data streams can also set the foundation for future data-driven tools and algorithms that can improve patient care.

Limitations

These findings should be interpreted within the context of several limitations. First, our sample was purposively selected to include diverse perspectives (eg, age, race, gender, geography, and peripheral device or tablet use) but may not be representative of all VA patients who received peripheral devices. As the majority of patients used their peripheral devices (23/25, 92%) or tablet (19/25, 76%), our findings may overestimate positive experiences, including effort expectancy and ease of use, because nonusers were underrepresented. Study participants, and tablet and peripheral device recipients in general, may also have been more engaged, technologically inclined, or positively disposed toward digital tools than the overall VA patient population. Second, although we recruited patients who received a device within the previous 6 months, perspectives on health improvements may be subject to recall biases, which could impact the accuracy of self-reported improvements. Third, our interview guide focused on the core constructs of UTAUT and did not include topics on privacy, trust, or confidentiality of data generated from the peripheral devices. However, privacy concerns and perceived risks may also influence behavioral intention and use of technology [39,40]. Fourth, we did not conduct member checking to validate patient narratives or themes, but during the interviews, patients were asked to clarify responses as needed to ensure their perspectives were understood and properly documented. The insights generated from this work offer valuable real-world perspectives that can inform the ongoing design, implementation, and expansion of remote monitoring programs and PGHD use across large, national health systems.

Conclusion

As the use of peripheral devices and PGHD continues to grow, input from patients about their experiences can maximize the benefit of these tools for patients and health systems. Our findings suggest that peripheral devices can enhance health monitoring, reduce care barriers, and foster patient engagement, especially when accompanied by other virtual care tools and responsive technical support. These findings provide evidence to support continued investment in health system–wide connected care initiatives, with emphasis on device

XSL•FO
RenderX

interoperability, education on device use, and awareness to ensure equitable access.

Acknowledgments

We thank the veterans who shared their experiences and participated in this project. We also appreciate the analytic contributions of Liberty Green (MS), James Van Campen (MS), and Cindie Slightam (MPH). We also thank Emily Drucker (LCSW) and Nilesh Shah (MD) for their input on the manuscript.

Funding

This work was supported by a Stanford Medicine Spectrum Population Health Sciences pilot grant (principal investigator ACG) and a grant from the US Department of Veterans Affairs Health Systems Research (CDA 23-145, principal investigator ACG). The views expressed are those of the authors and do not represent those of the US Department of Veterans Affairs or those of the US government.

Data Availability

The qualitative data generated during this project are not publicly available to protect the privacy of individuals who participated. Deidentified data are available from the corresponding author upon reasonable request, subject to institutional or regulatory approval.

Authors' Contributions

All authors contributed to the conceptualization, design, and methodology. ACG and CG conducted the qualitative analysis. ACG was primarily responsible for the initial draft of the manuscript, with input from all authors. All authors participated in manuscript writing, revision, and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Mobile fact sheet. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2025-07-31]
2. Health information national trends survey 7. National Cancer Institute. 2024. URL: https://hints.cancer.gov/view-questions/all-hints-questions.aspx [accessed 2025-11-24]
3. 21st century cures act: Interoperability, information blocking, and the ONC health IT certification program. Health and Human Services Department. 2020. URL: https://www.federalregister.gov/documents/2020/05/01/2020-07419/21st-century-cures-act-interoperability-information-blocking-and-the-onc-health-it-certification [accessed 2025-07-31]
4. Interoperability and patient access final rule. Centers for Medicare & Medicaid Services. 2020. URL: https://www.cms.gov/priorities/burden-reduction/overview/interoperability/policies-and-regulations/cms-interoperability-and-patient-access-final-rule-cms-9115-f [accessed 2025-07-31]
5. 2024 State of FHIR survey results. Health Level Seven International and Firely. URL: https://www.hl7.org/documentcenter/public/white-papers/2024%20StateofFHIRSurveyResults_final.pdf(31 [accessed 2025-07-31]
6. Enabling health interoperability through FHIR. Health Level Seven International. 2025. URL: http://fhir.org/ [accessed 2025-07-31]
7. Patient-generated health data. Assistant Secretary for Technology Policy. 2025. URL: https://www.healthit.gov/topic/scientific-initiatives/patient-generated-health-data [accessed 2025-07-31]
8. Austin E, Lee JR, Amtmann D, Bloch R, Lawrence SO, McCall D, et al. Use of patient-generated health data across healthcare settings: implications for health systems. JAMIA Open 2020;3(1):70-76 [FREE Full text] [doi: 10.1093/jamiaopen/ooz065] [Medline: 32607489]
9. Tiase VL, Hull W, McFarland MM, Sward KA, Del Fiol G, Staes C, et al. Patient-generated health data and electronic health record integration: a scoping review. JAMIA Open 2020;3(4):619-627 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa052] [Medline: 33758798]
10. Foster C, Schinasi D, Kan K, Macy M, Wheeler D, Curfman A. Remote monitoring of patient- and family-generated health data in pediatrics. Pediatrics 2022;149(2):e2021054137 [FREE Full text] [doi: 10.1542/peds.2021-054137] [Medline: 35102417]
11. Newton TJ, Shah N, Lewis K, Zocchi MS, Bixler FR, Etingen B, et al. Developing infrastructure to realize the value of patient-generated health data in a large integrated health care system: the Veterans Health Administration experience. J Med Internet Res 2025;27:e70755 [FREE Full text] [doi: 10.2196/70755] [Medline: 40479588]

XSL•FO
RenderX

12. Grossman LV, Masterson Creber RM, Benda NC, Wright D, Vawdrey DK, Ancker JS. Interventions to increase patient portal use in vulnerable populations: a systematic review. J Am Med Inform Assoc 2019;26(8-9):855-870 [FREE Full text] [doi: 10.1093/jamia/ocz023] [Medline: 30958532]
13. Estacio EV, Whittle R, Protheroe J. The digital divide: Examining socio-demographic factors associated with health literacy, access and use of internet to seek health information. J Health Psychol 2019;24(12):1668-1675. [doi: 10.1177/1359105317695429] [Medline: 28810415]
14. Clare CA. Telehealth and the digital divide as a social determinant of health during the COVID-19 pandemic. Netw Model Anal Health Inform Bioinform 2021;10(1):26 [FREE Full text] [doi: 10.1007/s13721-021-00300-y] [Medline: 33842187]
15. Treadwell JR, Reston JT, Rouse B. Automated-Entry Patient-Generated Health Data for Chronic Conditions: The Evidence on Health Outcomes. Rockville (MD): Agency for Healthcare Research and Quality; 2021.
16. Gresham G, Hendifar AE, Spiegel B, Neeman E, Tuli R, Rimel BJ, et al. Wearable activity monitors to assess performance status and predict clinical outcomes in advanced cancer patients. NPJ Digit Med 2018;1:27 [FREE Full text] [doi: 10.1038/s41746-018-0032-6] [Medline: 31304309]
17. Zulman DM, Wong EP, Slightam C, Gregory A, Jacobs JC, Kimerling R, et al. Making connections: nationwide implementation of video telehealth tablets to address access barriers in veterans. JAMIA Open 2019;2(3):323-329 [FREE Full text] [doi: 10.1093/jamiaopen/ooz024] [Medline: 32766533]
18. VA app store. U.S. Department of Veterans Affairs. 2025. URL: https://mobile.va.gov/appstore [accessed 2025-07-31]
19. Griffin AC, Troszak LK, Van Campen J, Midboe AM, Zulman DM. Tablet distribution to veterans: an opportunity to increase patient portal adoption and use. J Am Med Inform Assoc 2022;30(1):73-82 [FREE Full text] [doi: 10.1093/jamia/ocac195] [Medline: 36269168]
20. Griffin AC, Greene L, Slightam C, Wray CM, Wielgosz J, Kuhn E, et al. Use of mobile applications and mental health services among veterans who received internet-connected tablets. JAMIA Open 2024;7(4):ooae119. [doi: 10.1093/jamiaopen/ooae119] [Medline: 39534216]
21. Walker RC, Tong A, Howard K, Palmer SC. Patient expectations and experiences of remote monitoring for chronic diseases: systematic review and thematic synthesis of qualitative studies. Int J Med Inform 2019;124:78-85 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.01.013] [Medline: 30784430]
22. Paul MM, Khera N, Elugunti PR, Ruff KC, Hommos MS, Thomas LF, et al. The state of remote patient monitoring for chronic disease management in the United States. J Med Internet Res 2025;27:e70422 [FREE Full text] [doi: 10.2196/70422] [Medline: 40570326]
23. Alvarez P, Sianis A, Brown J, Ali A, Briasoulis A. Chronic disease management in heart failure: focus on telemedicine and remote monitoring. Rev Cardiovasc Med 2021;22(2):403-413 [FREE Full text] [doi: 10.31083/j.rcm2202046] [Medline: 34258907]
24. Hirko K, Kerver J, Ford S, Szafranski C, Beckett J, Kitchen C, et al. Telehealth in response to the COVID-19 pandemic: implications for rural health disparities. J Am Med Inform Assoc 2020;27(11):1816-1818 [FREE Full text] [doi: 10.1093/jamia/ocaa156] [Medline: 32589735]
25. Jacobs JC, Blonigen DM, Kimerling R, Slightam C, Gregory AJ, Gurmessa T, et al. Increasing mental health care access, continuity, and efficiency for veterans through telehealth with video tablets. Psychiatr Serv 2019;70(11):976-982. [doi: 10.1176/appi.ps.201900104] [Medline: 31378193]
26. Gujral K, Van Campen J, Jacobs J, Kimerling R, Blonigen D, Zulman DM. Mental health service use, suicide behavior, and emergency department visits among rural US veterans who received video-enabled tablets during the COVID-19 pandemic. JAMA Netw Open 2022;5(4):e226250 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.6250] [Medline: 35385088]
27. Bridging the digital divide. U.S. Department of Veterans Affairs. 2025. URL: https://telehealth.va.gov/digital-divide [accessed 2025-07-31]
28. ZIP code RUCA approximation. Rural Health Research Center. URL: https://depts.washington.edu/uwruca/ruca-approx.php#:~:text=Because%20the%20smallest%20geographic%20identifiers,the%20ZIP%20code%20geographic%20unit [accessed 2025-07-31]
29. Yoon J, Chee CP, Su P, Almenoff P, Zulman DM, Wagner TH. Persistence of high health care costs among VA patients. Health Serv Res 2018;53(5):3898-3916 [FREE Full text] [doi: 10.1111/1475-6773.12989] [Medline: 29862504]
30. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: Toward a unified view. MIS Quarterly 2003;27(3):425-478. [doi: 10.5555/2017197.2017202]
31. Kowalski CP, Nevedal AL, Finley EP, Young JP, Lewinski AA, Midboe AM, et al. Planning for and assessing rigor in rapid qualitative analysis (PARRQA): a consensus-based framework for designing, conducting, and reporting. Implement Sci 2024;19(1):71 [FREE Full text] [doi: 10.1186/s13012-024-01397-1] [Medline: 39394597]
32. Nouri S, Khoong EC, Lyles CR. Addressing equity in telemedicine for chronic disease management during the COVID-19 pandemic. NEJM Catalyst Innovations in Care Delivery 2020;1(3). [doi: 10.1056/CAT.20.0123]
33. Conceptualizing a data infrastructure for the capture, use, and sharing of patient-generated health data in care delivery and research through 2024. The Office of the National Coordinator for Health Information Technology. 2018. URL: https://www.healthit.gov/sites/default/files/onc_pghd_final_white_paper.pdf [accessed 2025-07-31]

XSL•FO
RenderX

34. Fortuna KL, Brooks JM, Umucu E, Walker R, Chow PI. Peer support: a human factor to enhance engagement in digital health behavior change interventions. J Technol Behav Sci 2019;4(2):152-161 [FREE Full text] [doi: 10.1007/s41347-019-00105-x] [Medline: 34337145]
35. Peng R, Chang J, Zhu Y, Feng H, Cao Z. Unlocking the power of peer support in digital use and digital health interventions for older adults: a scoping review. J Adv Nurs 2026;82(1):162-173. [doi: 10.1111/jan.16963] [Medline: 40195282]
36. Tahsin F, Armas A, Kirakalaprathapan A, Kadu M, Sritharan J, Steele Gray C. Information and communications technologies enabling integrated primary care for patients with complex care needs: scoping review. J Med Internet Res 2023;25:e44035 [FREE Full text] [doi: 10.2196/44035] [Medline: 37074779]
37. Ferguson JM, Van Campen J, Slightam C, Greene L, Heyworth L, Zulman DM. Evaluation of the veterans health administration's digital divide consult for tablet distribution and telehealth adoption: cohort study. J Med Internet Res 2024;26:e59089 [FREE Full text] [doi: 10.2196/59089] [Medline: 39250183]
38. Campos-Castillo C, Anthony D. Racial and ethnic differences in self-reported telehealth use during the COVID-19 pandemic: a secondary analysis of a US survey of internet users from late March. J Am Med Inform Assoc 2021;28(1):119-125 [FREE Full text] [doi: 10.1093/jamia/ocaa221] [Medline: 32894772]
39. Liu Y, Lu X, Zhao G, Li C, Shi J. Adoption of mobile health services using the unified theory of acceptance and use of technology model: self-efficacy and privacy concerns. Front Psychol 2022;13:944976 [FREE Full text] [doi: 10.3389/fpsyg.2022.944976] [Medline: 36033004]
40. Su J, Wang Y, Liu H, Zhang Z, Wang Z, Li Z. Investigating the factors influencing users' adoption of artificial intelligence health assistants based on an extended UTAUT model. Sci Rep 2025;15(1):18215 [FREE Full text] [doi: 10.1038/s41598-025-01897-0] [Medline: 40414992]

Abbreviations

CDW: Corporate Data Warehouse
EHR: electronic health record
PGHD: patient-generated health data
UTAUT: Unified Theory of Acceptance and Use of Technology
VA: US Department of Veterans Affairs

Edited by A Coristine, A Stone; submitted 24.Nov.2025; peer-reviewed by S A, M Lennon; comments to author 22.Dec.2025; revised version received 09.Mar.2026; accepted 09.Mar.2026; published 06.Apr.2026.

Please cite as:
Griffin AC, Gray C, Ngo V, Chaudhary C, Shimada SL, Midboe AM, Zulman DM
Supporting Access to Care Through Peripheral Devices and Patient-Generated Health Data: Qualitative Study
J Med Internet Res 2026;28:e88405
URL: https://www.jmir.org/2026/1/e88405
doi:10.2196/88405
PMID:

©Ashley C Griffin, Caroline Gray, Victoria Ngo, Camila Chaudhary, Stephanie L Shimada, Amanda M Midboe, Donna M Zulman. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Telehealth Delivery of the Homeostasis–Enrichment–Plasticity Approach for Premature Infants With Developmental Risks: Exploratory Feasibility Study

Aymen Balikci[1], PhD; Gamze Cagla Sirma[2], MSc; Izgi-Miray Demirbag[1], MSc; Teresa A May-Benson[3], ScD; Hatice Gulhan Sozen[4], MD; Ayse Firdevs Aracikul Balikci[5], PhD; Gul Ilbay[6], Prof Dr

[1]Sense On Ltd, Istanbul, Turkey

[2]Department of Occupational Therapy, Faculty of Health Sciences, Fenerbahçe University, Istanbul, Istanbul, Turkey

[3]TMB Education, Norristown, PA, United States

[4]Department of Child Health and Diseases, Faculty of Medicine, Bahcesehir University, Istanbul, Turkey

[5]Department of Special Education, Faculty of Education, Istanbul Medipol University, Istanbul, Turkey

[6]Department of Physiology, Faculty of Medicine, Kocaeli University, Kocaeli, Turkey

Corresponding Author:
Gamze Cagla Sirma, MSc
Department of Occupational Therapy
Faculty of Health Sciences
Fenerbahçe University
Atatuk Mah Metropol Istanbul, Atasehir Blv.
Istanbul, Istanbul, 34758
Turkey
Phone: 90 2169101907
Email: cagladirgen@hotmail.com

Abstract

Background: Preterm delivery is an increasing worldwide health concern linked to increased neurodevelopmental risks. Early intervention is crucial for harnessing neuroplasticity to enhance developmental and functional performance outcomes; however, access to early intervention is frequently hindered by logistical, financial, and labor constraints. The Homeostasis–Enrichment–Plasticity (HEP) Approach is a family-centered early intervention model based on enriched environments, designed to improve infants' sensory-motor, cognitive, and socio-emotional development.

Objective: This study aimed to assess the feasibility, safety, acceptability, and outcomes sensitivity to change of implementing the HEP Approach through telehealth for premature infants at developmental risk.

Methods: A pre-post exploratory feasibility study was performed, including 16 preterm infants (aged 4-12 months corrected age), of whom 14 completed the study. The 12-week intervention included weekly remote sessions focused on environmental enrichment, active exploration, and parental guidance. The feasibility and acceptability were evaluated using a 24-item questionnaire. Developmental outcomes were assessed with the Young Children's Participation and Environment Measure, Ages and Stages Questionnaire (ASQ), Alberta Infant Motor Scale, Infant Motor Profile, and Depression Anxiety Stress Scales.

Results: High adherence (14/14, 100%) and retention (14/16, 87.5%) rates demonstrated robust feasibility. Parents indicated 86%-100% agreement across all feasible criteria, affirming safety, satisfaction, and acceptability. No adverse incidents were reported. Changes were identified in participation (Young Children's Participation and Environment Measure), motor development (Alberta Infant Motor Scale, Infant Motor Profile, and ASQ), communication and social-emotional domains (ASQ), and caregiver well-being (Depression Anxiety Stress Scales) (P<.05).

Conclusions: The telehealth implementation of the HEP Approach demonstrated feasibility, safety, and strong acceptance among families, along with quantifiable developmental and psychosocial changes. These initial findings endorse the model's viability as an accessible, family-oriented telehealth framework for infants born preterm. Future randomized controlled and longitudinal studies are necessary to validate intervention efficacy and scalability.

(J Med Internet Res 2026;28:e86883) doi:10.2196/86883

XSL•FO
RenderX

KEYWORDS

early intervention; environmental enrichment; premature infants; telehealth; adherence; retention

Introduction

Approximately 1 in every 10 children worldwide (9.9% of all births) is born preterm (<37 weeks of gestation) [1]. Preterm delivery is associated with substantial long-term health consequences, spanning multiple domains such as respiratory and cardiovascular health as well as neurodevelopmental concerns. Within the neurodevelopmental domain, impairments may range from severe conditions, such as cerebral palsy—particularly among those born at the lowest gestational ages—to less severe but still clinically significant outcomes [1-3]. Recent evidence indicates that even births occurring just a few weeks early increase the risk of adverse developmental outcomes [1,4,5]. This is particularly relevant as approximately 85% of preterm births occur between 32 and 37 weeks of gestation, meaning that a large proportion of preterm infants are at risk, challenging the earlier assumption that only extremely preterm infants face vulnerability [1,4,6]. Moreover, numerous studies have demonstrated that, compared with their full-term-born peers, preterm infants are at heightened risk for difficulties across multiple developmental domains, including cognitive [7-9], motor [7,10-12], sensory [11-13], language [11,14,15], and social-emotional functioning [5,15,16].

Overall, these data emphasize the importance of early developmental support for infants born preterm [17,18]. Given the significant neuroplasticity in early life [19,20], early intervention (EI) can influence multiple domains related to prematurity [17]. Evidence indicates that structured services during infancy improve cognitive, motor, sensory, language, and social-emotional outcomes, while also fostering long-term neurodevelopmental pathways and functional engagement [12,21-24]. Consequently, EI may influence neurodevelopmental trajectories, reducing risks associated with prematurity and enhancing the quality of life for infants and their families [12,17,18,24-26].

Although EI provides substantial benefits for the development of preterm infants, many families face persistent challenges in accessing these services. The most prominent barriers include shortages of qualified professionals, transportation difficulties, and the difficulty of organizing time to attend services—factors that are especially evident for families living in rural or underserved areas. While systemic issues such as delayed referrals or fragmented communication may also play a role, it is these fundamental constraints of workforce availability, accessibility, and time that most directly limit timely participation in EI and reduce its potential impact during critical developmental windows [27-30].

With advances in digital technologies, telehealth has become a promising alternative for addressing barriers to EI in preterm infants [31,32]. Studies during the COVID-19 pandemic showed that telemedicine maintained developmental follow-up and medical care when in-person visits were restricted, mitigating workforce and logistical challenges [33]. Telehealth services are valuable for infants discharged from the neonatal intensive care unit, as they decrease needless hospital visits and limit exposure to transmissible infections [31,32,34]. Furthermore, studies demonstrate that telehealth can be as efficacious as in-person care for many health conditions and interventions [31,35,36] and may offer a cost-effective health care delivery paradigm [31,32,37,38].

Building on the growing evidence for telehealth as a viable and cost-effective model of early intervention, the Homeostasis–Enrichment–Plasticity (HEP) Approach [39] represents a structured framework that can be adapted to remote delivery while preserving its core principles of parent coaching and enriched developmental support. The HEP Approach, developed by Balikci [39], is a comprehensive EI model designed to promote sensory-motor, cognitive, and social-emotional development in infants. The HEP Approach incorporates 10 fundamental principles—homeostasis, safety, multisensory experiences, novelty, sustained engagement, and active exploration—based on environmental enrichment paradigms, through a clinic-based program that includes parent coaching and concurrent home implementation [24,39,40]. The approach provides infants with enhanced sensory and motor experiences in the clinic, while parents obtain systematic coaching that is later integrated into everyday routines and natural settings, with telehealth follow-up reinforcing these procedures at home. Rooted in frameworks including Ecological Perception Theory, Perception–Action Theory, Sensory Integration, the Person–Environment–Occupation (PEO) Model, and Dynamic Systems Theory, the HEP Approach equips caregivers with adaptable, goal-focused strategies that enhance co-regulation, sensory-motor learning, and adaptive participation [24,39,40].

While previous research has demonstrated the feasibility and effectiveness of the HEP Approach in the traditional clinic-based/home-monitored format [24,40], its applicability through telehealth has not yet been systematically examined. This study represents the first attempt to evaluate the feasibility of delivering the HEP Approach via telehealth.

The aim of this study is to assess the feasibility of implementing the HEP Approach in a telehealth format and to determine whether this model provides an accessible, acceptable, and practical option for supporting the developmental outcomes of premature infants and their families.

Methods

Study Design

This is a pre-post feasibility experimental study. The study was conducted in Istanbul, Türkiye, with interventions delivered remotely via telehealth between January 2025 and July 2025.

Ethical Considerations

The study was conducted in accordance with the Declaration of Helsinki and approved by the Clinical Research Ethics Committee of Istanbul Medipol University

XSL•FO
RenderX

(E-10840098-202.3.02-7368). All caregivers provided written informed consent before participation and were informed of their right to withdraw from the study at any time. Participant privacy and confidentiality were strictly maintained, and all data were anonymized prior to analysis. No identifiable personal information was included in the manuscript. No financial or material compensation was provided to the participants.

Study Procedures

The study comprised 11 structured phases that guided the implementation of the HEP Approach via telehealth [40]. Phase 1 covered participant recruitment via physician referral and digital record review, while Phase 2 introduced families to the HEP philosophy and telehealth process. Phase 3 of the study comprised baseline evaluations completed exclusively remotely via Zoom. The evaluation process comprised caregiver interviews, unstructured observations, and structured observations tailored to the requirements of standardized assessment instruments. All evaluations, including the implementation of standardized assessments, were conducted remotely via videoconferencing (Zoom). In Phases 4-5, assessment findings were collaboratively interpreted with families, and hypotheses regarding sensory-motor, emotional, social, and environmental factors were formulated and discussed. Phase 6 focused on interactive goal setting, followed in Phases 7-8 by individualized intervention strategies delivered through weekly telehealth coaching, demonstrations, and integration into daily family routines, supported by video sharing via WhatsApp. Phase 9 emphasized adaptation and generalization of strategies, while Phase 10 ensured continuous monitoring with therapist feedback and reflective questioning to enhance family problem-solving. Finally, Phase 11 included postintervention assessments, feasibility questionnaires, and feedback interviews conducted via videoconferencing. A comprehensive description of each phase is provided in Table S1 in Multimedia Appendix 1.

Fidelity to the intervention was maintained through multiple strategies. Treating therapists completed 120 hours of theoretical and practical training, including supervision, provided by the first author, who developed the model. In addition, fidelity was reinforced through regular mentoring and the first author's systematic review of recorded telehealth sessions. All procedures were implemented in accordance with the HEP Approach Implementation Manual [22], and session recordings were evaluated using the fidelity checklists included in the manual to ensure high implementation reliability.

Participants and Recruitment

Convenience sampling, a nonprobability method, was used [40]. Neonatologists from 3 hospitals in Istanbul referred premature infants who met the designated inclusion criteria. The inclusion criteria were (1) corrected age ranging from 4 to 12 months, (2) birth at ≤36 weeks and 6 days of gestation, (3) absence of systemic diseases or congenital defects, and (4) familial consent to participate consistently in the study process. The exclusion criteria included the following conditions: (1) substantial impairment of vision or hearing; (2) history of febrile seizures; (3) medical conditions that impede active participation in the study, such as dependence on oxygen; and (4) prior participation in other experimental rehabilitation studies. Following assessment of the qualifying criteria, informed consent was secured.

A total of 39 preterm infants were assessed for eligibility. Eighteen did not meet the inclusion criteria, and five families declined to participate, resulting in 16 infants enrolled in the HEP telehealth intervention. All 16 received the intervention. During follow-up, 2 infants discontinued participation due to illness, leaving 14 who completed the program and were included in the final analysis. All pre- and postintervention evaluations were conducted by a physical therapist with over 5 years of pediatric experience and an advanced Master of Science degree. This therapist had received formal training in the administration and scoring of outcome measures as part of their graduate program and was blinded to the intervention. The HEP Approach itself was delivered by 2 physiotherapists, each with more than five years of pediatric clinical experience and advanced Master's-level training. The participant flow is illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. Flowchart of participant and outcome assessment through the trial. HEP: Homeostasis–Enrichment–Plasticity.

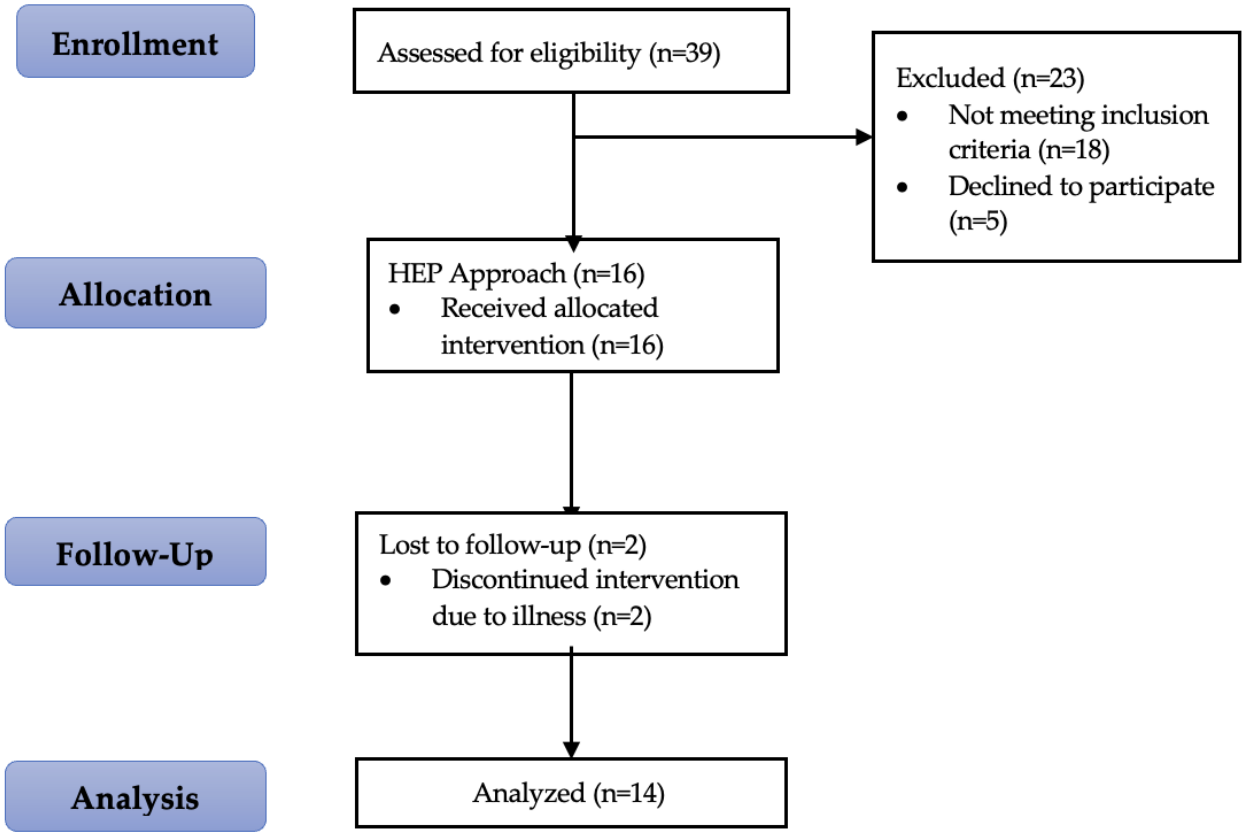

Feasibility Measure

Feasibility and acceptability of the telehealth-based HEP Approach were evaluated using a 24-item Likert-type questionnaire, adapted from the feasibility study by Balikci et al [40]. The survey was structured to assess four domains: feasibility (items 1-7), acceptability (items 8-12), satisfaction (items 13-19), and safety (items 20-24). Each item was rated on a five-point scale ranging from strongly agree to strongly disagree. For analytic purposes, responses of "Strongly Agree" and "Agree" were categorized as agreement, whereas all other responses were classified as disagreement. Alongside the structured items, 3 open-ended questions were incorporated at the conclusion of the form, inviting parents to provide their responses. The qualitative data were subsequently examined and thematically grouped according to the 3 questions. Feedback on the intervention's feasibility, safety, and perceived effectiveness was also collected from both therapists through postintervention reflections and monitoring logs maintained throughout the study period.

Outcome Measures

Caregivers first completed a demographic questionnaire capturing infant and maternal health, gestational and postnatal history, corrected age at enrollment, gender, birth weight, gestational age, maternal age, family income, and caregiver educational level.

Five standardized outcome measures were used to evaluate developmental progress across multiple domains. The Turkish translated version of the Young Children's Participation and Environment Measure (YC-PEM) was used to assess participation frequency, involvement, desire for change, and environmental support in both home and community contexts [41-43]. Broader developmental domains, including social-emotional, communication, gross motor, fine motor, problem-solving, and personal-social skills, were measured with the Turkish translated version of the Ages and Stages Questionnaire (ASQ) [44]. Motor development was further assessed using the Alberta Infant Motor Scale (AIMS) [45-49] and the Infant Motor Profile (IMP) [50-52], which captured dimensions such as variation, adaptability, symmetry, fluency, and performance. Finally, parental mental health and well-being were evaluated with the Depression Anxiety Stress Scales (DASS-21) [53]. The selection of outcome measures was guided by the study's feasibility objectives rather than by an intention to establish intervention efficacy. AIMS and IMP were selected as standardized, performance-based tools reported in remote or telehealth assessment contexts and among the limited motor assessments adaptable to video-based administration. In this study, these instruments were used to evaluate (1) whether they could be feasibly administered in a telehealth format within our sample and (2) whether they were capable of detecting measurable change over a 12-week period, regardless of whether such change reflected intervention-related or developmental factors.

Caregiver-reported measures (YC-PEM, ASQ, and DASS-21) were included to capture domains not directly assessable via remote standardized performance testing and to examine the feasibility of administering structured family-report instruments within a telehealth early intervention framework.

XSL•FO
RenderX

The Young Children's Participation and Environment Measure-Turkish Version

The Turkish translation of the YC-PEM was used in this study. The YC-PEM is a caregiver-completed measure designed to assess children's participation up to 5 years of age across home, daycare/preschool, and community settings. It evaluates four domains: frequency of participation, level of involvement, parental desire for change, and environmental supports and barriers. Previous studies demonstrated good internal consistency (α=0.67-0.96), fair-to-good test-retest reliability (ICC=0.59-0.94) in home and community contexts, and adequate construct validity in differentiating participation across age and disability groups [41,42]. The Turkish translation and cross-cultural adaptation of the YC-PEM were conducted following standardized procedures, and subsequent psychometric evaluation in a large sample of 367 children, both with and without disabilities, demonstrated good reliability (Cronbach α=0.68-0.94; ICC=0.69-0.89) together with evidence of discriminant and concurrent validity, thereby supporting the YC-PEM-T as a reliable and culturally appropriate instrument for assessing children's participation and environmental influences in Türkiye [43]. The YC-PEM was provided to families in electronic (PDF) format. Caregivers independently completed the measure at their convenience and returned the completed forms electronically.

The Ages and Stages Questionnaire-Turkish Version

The Turkish version of the Ages and Stages Questionnaire –Turkish Version (ASQ-TR) was used. The ASQ is a parent-completed developmental screening instrument designed for children aged 4 to 60 months (extended up to 72 months in some adaptations). Each form consists of age-specific items that evaluate five developmental domains: communication, gross motor, fine motor, problem-solving, and personal-social skills. Parents complete the questionnaires in approximately 10-15 minutes, based on their observations and knowledge of their children's performance, and respond to items with options such as "yes," "sometimes," or "not yet." Responses are scored to determine whether the child's development is on schedule, requires monitoring, or needs referral for further evaluation.

The original ASQ demonstrated good psychometric qualities, with reported sensitivity values typically ranging from 75% to 90% and specificity from 77% to 88%, alongside strong reliability indices and evidence of construct validity against standardized developmental assessments. The Turkish version (ASQ-TR), evaluated in a large sample of 833 children aged 3-72 months, demonstrated satisfactory reliability and validity. Concurrent validity analyses showed overall agreement rates of 70.95% when one domain was below the cutoff and 88.24% when 2 domains were below the cutoff, supporting the 2-domain criterion as a more accurate classification approach. Internal consistency coefficients (Cronbach α) ranged from 0.38 to 0.95 across domains and age forms, generally increasing with age. Domain-total correlations (r=0.54-0.96) and interdomain correlations (r=0.65-0.79) were all statistically significant. Test-retest reliability over a 2-week interval indicated high agreement across domains (ranging from 0.82 to 0.97), and interrater reliability between mothers and teachers had agreement levels between 73.68% and 87.18%, depending on the criterion used.

The use of a 2-domain cutoff criterion considerably improved diagnostic accuracy in this version of the measure, increasing specificity to 85% and positive predictive value to 75%, while maintaining high sensitivity (94%). These findings support the ASQ-TR as a reliable and culturally appropriate developmental screening tool for use in Türkiye. Families received the ASQ-TR in PDF format; caregivers completed the measure at their convenience and submitted the forms electronically.

The Alberta Infant Motor Scale

The AIMS is a standardized observational tool developed to identify motor delays and describe infants' gross motor performance. It evaluates spontaneous movements, posture, weight shifts, and antigravity control over a 20-30 minute assessment, often supported by age-appropriate toys to encourage activity. The scale includes 58 items across four positions (21 prone, 9 supine, 12 sitting, 16 standing), scored dichotomously and summarized on age-specific percentile charts ranging from the 5th to the 90th percentile. Based on these scores, infants' motor performance can be classified as delayed, suspect, or normal [45,46]. Due to its low cost, ease of administration, reproducibility, and minimal handling of the infant, AIMS is widely recognized as a practical measure in both clinical and community health contexts, and has been validated as reliable for high-risk populations The AIMS demonstrated strong concurrent validity with the Bayley Scales of Infant Development (r=0.95; P<.01) and high interrater reliability across all tested ages (intraclass correlation coefficients ranging from 0.76 to 0.99), supporting it as a valid and reliable tool for assessing motor development in high-risk infants [47]. More recently, studies have demonstrated that the AIMS can be successfully applied in telehealth settings, with strong agreement between remote and face-to-face assessments for neurodevelopmentally high-risk infants [48,49]. In this study, the AIMS was administered remotely via videoconferencing (Zoom). The evaluating specialist provided real-time verbal guidance to caregivers on how to position their child for each assessment item and scored the child's performance based on live observations. All sessions were video-recorded, enabling the evaluator to revisit the recordings when needed to verify or adjust scores. This administration procedure aligned with protocols reported in prior research [48,49].

The Infant Motor Profile

The IMP is a video-based observational tool designed to evaluate motor behavior in infants aged 3 to 18 months, or until a few months after independent walking is achieved. It consists of 80 items across five domains—movement variation, adaptability, symmetry, fluency, and performance—capturing both quantitative and qualitative aspects of motor development. Grounded in the Neuronal Group Selection Theory, the IMP emphasizes the importance of motor variability and adaptive strategy selection as indicators of central nervous system integrity [50].

The IMP has demonstrated strong psychometric properties, including acceptable interrater reliability in high-risk infants,

XSL•FO
RenderX

with Spearman correlation coefficients ranging from 0.47 to 0.94 across domains; reliability was highest for the performance domain (r=0.90-0.93) and lowest for symmetry (r=0.47-0.74), supporting its use as a reliable tool for assessing early motor development and identifying physiotherapy needs in the Turkish sample [51]. Moreover, the IMP has shown excellent concurrent validity with the AIMS (ρ=0.76; P<.001), a significant association with the General Movement Assessment, and high diagnostic accuracy in predicting neurodevelopmental disorders at 5 months (AUC=0.92), with strong sensitivity (93%) and specificity (81%), confirming its validity and clinical utility for evaluating early motor behavior and reflecting the severity of brain injury in at-risk infants. [52]. Additionally, studies indicate that the IMP is responsive to change, making it a valuable outcome measure in early intervention research [52]. Collectively, these findings support the IMP as a reliable and valid tool for assessing motor development and detecting early deviations in infants at risk of neurodevelopmental disorders. The IMP was administered by the examiner, who provided structured instructions for various positions and object manipulations during the initial session; the examiner recorded the session and took detailed notes while guiding the caregiver through the tasks, and later used the video recording to score and analyze the child's performance.

Depression Anxiety Stress Scales-21-Turkish Version

The DASS-21 is a self-report instrument designed to assess levels of depression, anxiety, and stress [53]. The Turkish cultural adaptation, validity, and reliability study of the DASS-21 was conducted by Yılmaz et al [54], who refined the existing Turkish translation of the scale through a comprehensive linguistic and cultural adaptation process and subsequently confirmed its psychometric soundness. Proven to be a valid and reliable tool for screening depressive and anxiety disorders, the DASS-21 effectively evaluates the severity of these 3 psychological conditions through 21 items. The scale consists of 3 subscales: depression (7 items), anxiety (7 items), and stress (7 items). Each item is rated on a 4-point Likert scale ranging from 0 (never) to 3 (always). Lower scores on the subscales indicate better psychological well-being [54]. The electronic (PDF) version of the DASS-21 in Turkish was given to families. At their convenience, caregivers completed the measure themselves and submitted the completed forms via WhatsApp.

Intervention

The HEP Approach, developed by Balıkcı [39], is a systematic early intervention model grounded in EE paradigms, principles of neuroplasticity, and ecological-dynamic theories of human development. Drawing on experimental EE research and mechanisms of brain plasticity, the model translates these principles into clinical practice through an integrated theoretical framework that combines Ecological Theory, Dynamic Systems Theory, Gibson's perception-action perspective, the Theory of Neuronal Group Selection, sensory integration theory, and the PEO model. Within this framework, development is conceptualized as emerging from the continuous interaction among the child, task demands, environmental affordances, and temporal processes [39,40].

Ecological and PEO perspectives inform the alignment of environmental supports and task structures with the infant's regulatory and developmental capacities, particularly within the zone of proximal development. Dynamic Systems Theory underpins the emphasis on non-linear change and self-organization, operationalized clinically through graded challenge, structured variation, and strategic manipulation of environmental and task constraints to facilitate spontaneous motor and behavioral reorganization. Gibson's perception-action framework guides the design of environments that invite active exploration, ensuring that sensory information is directly coupled with meaningful action. The Theory of Neuronal Group Selection and neuroplasticity principles support the systematic use of novelty, repetition within variation, and sustained enriched exposure to strengthen adaptive neural networks through experience-dependent plasticity.

Central to the HEP model is physiological homeostasis, conceptualized as a foundational prerequisite for exploration, engagement, and learning. Regulation of sleep, arousal, stress modulation, and coregulation processes is addressed prior to and throughout intervention to optimize the infant's readiness for adaptive change. Enriched environment principles are operationalized through 10 structured intervention components: physiological homeostasis, safety, provision of multisensory experiences, spatial adaptation, environmental and object novelty, just-right challenge, enjoyment, social opportunities, continuity, and active engagement and exploration. These components guide assessment, hypothesis generation, and intervention planning by systematically aligning environmental affordances, task demands, and individual sensory-motor and regulatory capacities.

In practice, the HEP Approach integrates structured therapist guidance with caregiver coaching to embed enriched, developmentally attuned experiences into daily routines and natural contexts. More detailed information regarding the theoretical foundations and conceptual framework of the HEP model is available in previously published studies [22,24,39,40].

The approach consists of 11 structured phases, beginning with referral and family introduction, followed by comprehensive assessment, identification of child and family strengths and challenges, hypothesis formulation, collaborative goal setting, selection of outcome measures, intervention planning, implementation, monitoring, and evaluation [40]. In this study, all phases of the HEP Approach were adapted for telehealth delivery and implemented over 12 one-hour weekly sessions across 3 months. Families, guided through Zoom, arranged their home environments according to their own resources, preferences, and the child's developmental needs, transforming everyday spaces into therapeutic contexts. Assessments and observations were conducted via structured caregiver interviews, live video interaction, and parent-recorded video clips, while therapists engaged parents in reflective discussions to interpret child behaviors and environmental affordances. Intervention planning and delivery progressed systematically: early sessions emphasized supporting homeostasis and self-regulation; subsequent sessions focused on reorganizing physical and social environments to foster active exploration; later sessions encouraged the diversification and generalization of newly

XSL•FO
RenderX

acquired skills across settings and materials; and final sessions concentrated on empowering parents to independently adapt environments and sustain enriched opportunities.

Each telehealth session also followed a five-step therapeutic sequence to ensure fidelity and caregiver engagement: (1) a check-in and reflection phase, where parents and therapists discussed the past week and reviewed parent-recorded videos; (2) an observation of natural parent-infant interaction, with therapists guiding camera placement to capture spontaneous exploration; (3) a guided reflective discussion, in which therapists and caregivers collaboratively interpreted the infant's strengths, challenges, and responses to environmental affordances; (4) a real-time application phase, where parents directly practiced strategies with their infant during the session under therapist support; and (5) a closure and collaborative planning phase, setting goals and strategies for home practice until the next meeting (Figure 2).

Figure 2. Therapeutic process cycle in telehealth sessions.

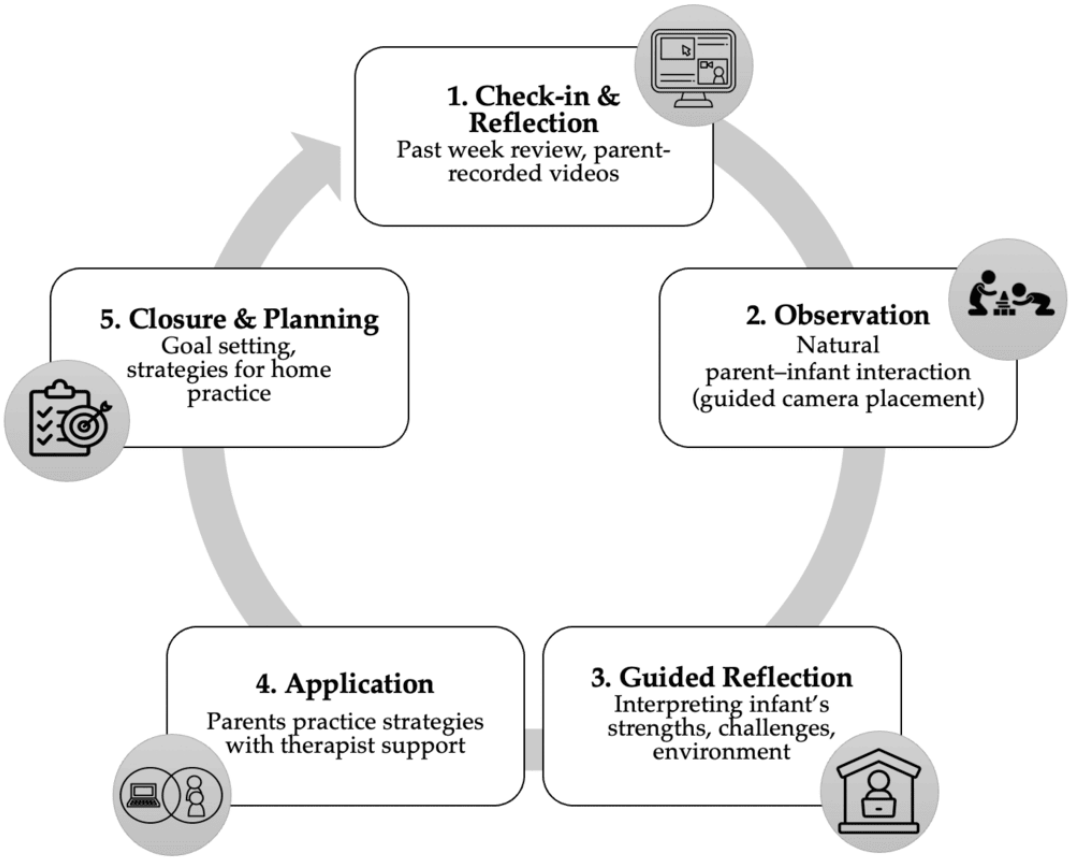

Grounded in 10 core principles—including maintenance of homeostasis, sensory experiences, safety, novelty, spatial adaptation, just-right challenge, enjoyment, social opportunities, continuous engagement, and active exploration—the intervention sought not only to promote self-regulation, sensory processing, motor and functional skills in children, but also to enhance parenting self-efficacy, thereby enabling families to provide ongoing enriched environments across contexts [24,39,40]. Fidelity to the approach was assured through adherence to the HEP manual [22].

Statistical Analysis

Study data were analyzed using IBM SPSS Statistics for Windows (version 22.0; SPSS Inc). Descriptive and demographic variables were summarized with frequency and percentage analyses. Outcome measure scores were examined using means and SDs. Normality of the research variables was assessed through skewness and kurtosis, with values between −1.5 and +1.5 for skewness and between −2.0 and +2.0 for kurtosis considered indicative of a normal distribution [55,56]. All variables met these criteria, confirming normal distribution. Accordingly, parametric statistical methods were applied. Two-sided paired *t* tests were conducted to compare repeated measures within dependent groups, and a *P* value of <0.05 was accepted as statistically significant. Given the exploratory feasibility design of the study, statistical analyses were conducted to examine whether measurable pre-post differences could be detected and to explore the sensitivity of the selected instruments to change within a telehealth context. No formal correction for multiple comparisons was applied, as the analyses were descriptive and hypothesis-generating rather than confirmatory.

Results

Demographic Characteristics of the Participants

The study included 14 infants, of whom 6 (42.9%) were female, and 8 (57.1%) were male. The mean maternal age at the time of birth was 33.42 (SD 3.69, range 25-39) years, while the mean

XSL•FO
RenderX

paternal age was 34.85 (SD 4.46, range 27-44) years. Mothers predominantly held a bachelor's degree (11/14, 78.6%), with a small proportion having completed high school (2/14, 14.3%) or postgraduate education (1/14, 7.1%). Similarly, most fathers had attained a bachelor's degree (12/14, 85.7%), while 7.1% (n=1) had completed high school and 7.1% (n=1) held postgraduate qualifications. The mean gestational age at birth was 32.14 (SD 3.03, range 26-36) weeks, and the mean birth weight was 1951.57 (SD 748.08, range 855-3400) grams. At baseline, the corrected mean infant age was 8.03 (SD 1.53, range 4.5-10.5) months. Detailed demographic characteristics of the participants are presented in Table 1.

Table 1. Participant demographic information (N=14).

Characteristic	Values
Sex, n (%)	
Female	6 (42.9)
Male	8 (57.1)
Mother's educational level, n (%)	
High school	2 (14.3)
Bachelor's degree	11 (78.6)
MSc or PhD	1 (7.1)
Father's educational level, n (%)	
Secondary school	0 (0)
High school	1 (7.1)
Bachelor's degree	12 (85.7)
MSc or PhD	1 (7.1)
Maternal age at infant's birth (years)	
Mean (SD)	33.42 (3.69)
Min-Max	25-39
Father's age at infant's birth (years)	
Mean (SD)	34.85 (4.46)
Min-Max	27-44
Birth gestational age (week)	
Mean (SD)	32.14 (3.03)
Min-Max	26-36
Corrected age at baseline (months)	
Mean (SD)	8.03 (1.53)
Min-Max	4.5-10.5
Birth weight (grams)	
Mean (SD)	1951.57 (748.08)
Min-Max	855-3400

Feasibility, Safety, Acceptability, and Parental Satisfaction in Telehealth

The trial enrolled 16 infants; however, 2 were withdrawn due to medical complications. One infant developed recurrent respiratory distress, and the other experienced feeding difficulties requiring hospitalization. As a result, both were unable to attend four consecutive telehealth sessions and were excluded from the study. Consequently, the intervention was completed with 14 infants who participated in all sessions and constituted the final sample for analysis. Over the 12-week period, all remaining caregivers and their infants attended the weekly sessions regularly and completed all 12 sessions without any disruptions.

The feasibility survey found highly positive results across all domains—feasibility, safety, acceptability, and parental satisfaction—regarding the implementation of the HEP Approach via telehealth. Overall, 86%-100% (n=12-14) of participants selected "Agree" or "Strongly Agree" across all survey items, indicating strong endorsement of the program. In terms of feasibility, 86% (n=12) of families agreed or strongly agreed that the HEP Approach was clear and understandable, and that the recommended activities and environmental arrangements were easy to apply within their daily routines.

XSL•FO
RenderX

Additionally, 93% (n=13) found participation in remote sessions convenient, and the process of regularly sharing home activity videos with the clinician practical and manageable. The remaining responses represented participants who chose neutral options, reflecting uncertainty or hesitation rather than disagreement, as these respondents did not indicate negative perceptions of the telehealth process or program content.

Regarding acceptability, all participants (100%) agreed or strongly agreed that the suggested activities and home arrangements were appropriate and suitable for their infants and families. They emphasized that the therapist's communication and approach toward both parents and infants were respectful and supportive, and that remote sessions were compatible with family dynamics and needs.

Within the satisfaction domain, 100% (N=14) of participants expressed that they were satisfied or highly satisfied with the program. They reported that the intervention effectively met their infants' developmental needs and enhanced their confidence and competence as parents. All families stated that they would recommend the program to other parents.

Regarding safety, 100% (N=14) of families agreed or strongly agreed that all recommended activities were safe for their infants, that they received sufficient guidance to ensure a secure home environment, and that the therapist consistently attended to the infants' and parents' sense of safety. Additionally, 86% (n=12) indicated that no adverse incidents (such as falls or collisions) occurred during the program, while the remaining 14% (n=2) neither agreed nor disagreed, reporting no negative experiences but maintaining a neutral stance.

Qualitative comments from caregivers corroborated these quantitative results. Parents highly valued the program's remote accessibility, the promptness of therapist replies, and the practicality of advice that could be incorporated into everyday routines. Initially, several families encountered difficulties integrating recommended activities into their routines and maintaining their infants' attention during remote sessions; however, these obstacles diminished with time as they adapted to the approach. Participants characterized the program as coherent, efficient, and highly advantageous, noting that it not only facilitated their infants' developmental progress but also enhanced their parental awareness, confidence, and involvement. Additionally, insights obtained from the 3 open-ended survey questions identified the following principal themes.

- The program's most valued features: Parents consistently emphasized that the curriculum was lucid, pragmatic, and easily implementable within the home setting, using resources readily accessible in everyday life. They highlighted that the therapist's instructions were clear and easily comprehensible, which enhanced their confidence in securely applying procedures. A mother remarked, "The activities were straightforward, and I was able to implement them immediately." For instance, I discovered that my infant required visual exposure to his hands before using them, and that maintaining an upright posture was crucial for engaging with his surroundings. Another participant remarked, "I appreciated that I could utilize materials we already possessed at home and did not require any special purchases." For example, I assisted my infant's sitting with a laundry basket, used cardboard boxes as a diminutive table, and crafted vibrant rattles using lentils and plastic bottles—these were inventive, enjoyable, and secure concepts." Numerous families appreciated the inclusive, home-centered format of the meetings, which facilitated the involvement of additional caregivers. One parent stated, "Due to its home-based nature, my husband and grandmother were able to participate, enhancing the experience and providing greater support for everyone involved."
- Suggestions for program improvement: Although parents expressed predominantly favorable opinions, a minority proposed slight enhancements to improve involvement and continuity. They suggested creating a straightforward activity tracking checklist and a visual family guide featuring images and sequential examples of activities. These parents suggested that these tools would help them recall, organize, and confidently recreate activities between sessions. One participant stated, "A straightforward checklist would assist me in adhering to the activities more consistently and monitoring our weekly progress." Another suggested, "A concise booklet featuring images that illustrate the proper positions and materials would enhance ease and safety in practice." These recommendations indicated families' need for organized assistance mechanisms to provide activities that enhance both safety and autonomy at home.
- Supplementary observations and pragmatic advantages: In their supplementary comments, parents underscored the practical and logistical benefits of the home-based arrangement. Avoiding traffic and congested public areas—particularly in major urban centers like Istanbul—was seen as alleviating stress, saving time, and reducing health risks for newborns. Families valued the flexibility to arrange lessons around their child's daily schedule, including nap and feeding times. A mother remarked, "We no longer had to endure prolonged traffic, resulting in a calmer environment for both the baby and us, thereby enhancing the productivity of the sessions." Another remarked, "The most advantageous aspect was the ability to select the most appropriate time for our child." If the baby became fatigued, we could postpone the session to another time, which enhanced its effectiveness.

These thoughts emphasized that the program's remote and home-based format not only improved safety and accessibility but also bolstered family engagement and comfort, strongly correlating with the favorable quantitative results.

Young Children's Participation and Environment Measure

Pre-post comparisons demonstrated statistically significant differences across both home and community participation domains. In the home setting, mean frequency scores increased from 3.67 (SD 0.37) to 4.32 (SD 0.62) (P<.001), and involvement scores increased from 2.43 (SD 0.57) to 3.49 (SD 0.47) (P<.001). Caregivers reported a decrease in desire for change (58.16, SD 5.63 to 21.43, SD 9.15; P<.001) and higher

XSL•FO
RenderX

environmental support scores (78.75, SD 13.39 to 87.36, SD 7.35; *P*=.002).

In the community setting, frequency and involvement scores also showed statistically significant increases (*P*<.001), while desire for change decreased (58.91, SD 10.20 to 39.31, SD 10.30; *P*<.001). Environmental support scores increased from 68.13 (SD 12.75) to 76.19 (SD 13.95) (*P*<.001).

Overall, the YC-PEM captured measurable pre-post changes over the 12-week telehealth delivery period. Further details are presented in Table 2.

XSL•FO
RenderX

Table 2. Outcome changes after 12 weeks of intervention (N=14).

	BI[a], mean (SD)	AI[b], mean (SD)	Estimate of effect (95% CI)	*t* test (*df*)	*P* value
YC-PEM[c] Home					
Frequency	3.67 (0.37)	4.32 (0.62)	–0.65 (–0.96 to –0.35)	–4.655 (13)	<.001[d]
Involvement	2.43 (0.57)	3.49 (0.47)	–1.06 (–1.32 to –0.80)	–8.720 (13)	<.001[d]
Desire for change	58.16 (5.63)	21.43 (9.15)	36.74 (30.25 to 43.22)	12.243 (13)	<.001[d]
Environmental support	78.75 (13.39)	87.36 (7.35)	–8.61 (–13.31 to –3.91)	–3.954 (13)	.002[d]
YC-PEM Community					
Frequency	1.55 (0.51)	2.00 (0.54)	–0.45 (–0.64 to –0.26)	–5.036 (13)	<.001[d]
Involvement	1.46 (0.46)	2.29 (0.78)	–0.83 (–1.07 to –0.58)	–7.351 (13)	<.001[d]
Desire for change	58.91 (10.20)	39.31 (10.30)	19.60 (11.52 to 27.69)	5.236 (13)	<.001[d]
Environmental support	68.13 (12.75)	76.19 (13.95)	–8.06 (–11.41 to –4.71)	–5.195 (13)	<.001[d]
ASQ[e]					
Social emotional	32.50 (16.37)	16.79 (10.85)	15.71 (6.56 to 20.89)	2.393 (13)	<.001[d]
Communication	45.36 (8.65)	50.71 (6.75)	–5.36 (–10.22 to –2.38)	–2.252 (13)	.03[d]
Gross motor	32.50 (16.02)	48.21 (10.30)	–15.71 (–23.55 to –4.33)	–3.626 (13)	.001[d]
Fine motor	44.64 (11.00)	54.64 (4.98)	–10.00 (–16.98 to –3.09)	–3.231 (13)	.009[d]
Problem solving	46.79 (12.19)	53.21 (4.64)	–6.43 (–15.52 to 2.66)	–1.528 (13)	.15
Personal-social	38.93 (12.89)	51.79 (5.75)	–12.86 (–21.16 to –3.35)	–3.841 (13)	.005[d]
AIMS[f]	31.07 (9.93)	47.50 (8.19)	–16.43 (–20.13 to –9.58)	–7.716 (13)	<.001[d]
IMP[g]					
Variation	81.93 (6.81)	96.50 (4.80)	–14.57 (–19.04 to –10.10)	–7.041 (13)	<.001[d]
Adaptability	73.25 (8.95)	91.79 (9.40)	–18.08 (–23.25 to –12.92)	–7.706 (13)	<.001[d]
Symmetry	98.21 (5.21)	99.14 (2.18)	–0.93 (–2.93 to 1.08)	–1.000 (13)	.34
Fluency	88.71 (14.04)	99.14 (3.21)	–10.43 (–18.22 to –2.64)	–2.893 (13)	.01[d]
Performance	70.14 (9.35)	88.93 (6.60)	–18.79 (–22.29 to –11.59)	–11.587 (13)	<.001[d]
Total	82.86 (5.32)	95.00 (4.37)	–12.14 (–14.66 to –10.41)	–10.408 (13)	<.001[d]
DASS-21[h]					
Depression	3.29 (2.40)	1.64 (1.01)	1.64 (0.50 to 2.79)	3.097 (13)	.008[d]
Anxiety	2.36 (0.84)	0.57 (0.65)	1.79 (1.38 to 2.19)	9.555 (13)	<.001[d]
Stress	6.57 (2.68)	4.00 (2.22)	2.57 (1.67 to 3.47)	6.188 (13)	<.001[d]
Total	12.36 (4.80)	6.21 (3.31)	6.14 (4.64 to 7.64)	8.849 (13)	<.001[d]

[a]BI: before intervention.

[b]AI: after intervention.

[c]YC-PEM: Young Children's Participation and Environment Measure.

[d]Statistically significant results (*P*<.05).

[e]ASQ: Ages and Stages Questionnaire.

[f]AIMS. Alberta Infant Motor Scale.

[g]IMP: Infant Motor Profile.

[h]DASS-21: Depression Anxiety Stress Scales.

XSL•FO
RenderX

Ages and Stages Questionnaire

Pre-post comparisons demonstrated statistically significant differences across several ASQ domains. Communication scores increased from 45.36 (SD 8.65) to 50.71 (SD 6.75) (P=.03), gross motor scores from 32.50 (SD 16.02) to 48.21 (SD 10.30) (P=.001), fine motor scores from 44.64 (SD 11.00) to 54.64 (SD 4.98) (P=.009), and personal-social scores from 38.93 (SD 12.89) to 51.79 (SD 5.75) (P=.005). Social-emotional scores decreased from 32.50 (SD 16.37) to 16.79 (SD 10.85) (P<.001).

Problem-solving scores increased over the 12-week period; however, this change did not reach statistical significance (P=.15). Overall, the ASQ detected measurable pre-post differences across multiple developmental domains during the telehealth delivery period. Further details are presented in Table 2.

Alberta Infant Motor Scale

Pre-post analysis revealed a statistically significant difference in AIMS total scores over the study period. The mean total score increased from 31.07 (SD 9.93) at baseline to 47.50 (SD 8.19) at follow-up (P<.001). These results reflect measurable changes in gross motor performance across the 12-week telehealth interval.

Infant Motor Profile

Pre-post comparisons indicated statistically significant differences across multiple IMP domains. Variation scores increased from 81.93 (SD 6.81) to 96.50 (SD 4.80) (P<.001), adaptability from 73.25 (SD 8.95) to 91.79 (SD 9.40) (P<.001), fluency from 88.71 (SD 14.04) to 99.14 (SD 3.21) (P=.01), and performance from 70.14 (SD 9.35) to 88.93 (SD 6.60) (P<.001). Symmetry scores did not differ significantly (P=.34). The overall IMP total score increased significantly from baseline to follow-up, from 82.86 (SD 5.32) to 95.00 (SD 4.37) (P<.001). Further details are presented in Table 2.

Depression Anxiety Stress Scales

Analysis of caregiver-reported outcomes showed statistically significant differences between baseline and follow-up scores. Depression scores declined from 3.29 (SD 2.40) to 1.64 (SD 1.01) (P=.008), anxiety scores from 2.36 (SD 0.84) to 0.57 (SD 0.65) (P<.001), and stress scores from 6.57 (SD 2.68) to 4.00 (SD 2.22) (P<.001). The total DASS-21 score decreased from 12.36 (SD 4.80) to 6.21 (SD 3.31) (P<.001). Detailed results are presented in Table 2.

Discussion

Principal Findings

This study evaluated the feasibility, acceptability, satisfaction, and safety of delivering the HEP Approach through telehealth for premature infants at developmental risk. Fourteen of 16 enrolled infants completed the 12-week program (87.5% retention), and session adherence reached 100%, indicating strong feasibility. Between 86% and 100% of caregivers endorsed all feasibility, acceptability, satisfaction, and safety items. No adverse events were reported. Therapists likewise confirmed that the intervention was practical to deliver remotely and supported appropriate environmental adaptations. Qualitative feedback further emphasized the clarity of instructions, ease of implementation with household materials, logistical advantages (eg, avoiding travel and exposure to illness), and opportunities for broader family involvement. These findings are consistent with previous clinic-based feasibility research and randomized controlled trial outcomes examining the HEP Approach [24,40], supporting the model's applicability across delivery formats.

The present results align with a growing body of telehealth-based early intervention literature demonstrating high parental engagement and satisfaction [31,33,34,57-59]. Telepractice models enabling routine-based home support have been well-received and promote active caregiver involvement [58]. Similar feasibility and satisfaction findings have been reported in the Baby Bridge telehealth model [32], neonatal intensive care unit follow-up telemedicine services [34], and telehealth physiotherapy programs for infants at risk of cerebral palsy [60,61]. Collectively, these studies reinforce the practicality and ecological validity of telehealth in early intervention contexts.

Significant improvements were observed in children's participation, as measured by the YC-PEM. Increases in home and community frequency and involvement, alongside reduced caregiver desire for change and improved environmental support, suggest meaningful enhancement in functional engagement. These findings are consistent with prior studies demonstrating the YC-PEM's sensitivity to intervention-related change across service models [62,63]. Similarly, previous HEP feasibility research using the Goal Attainment Scale reported functional improvements in daily routines [40]. Participation-focused enriched models such as the STEP protocol have also shown gains in YC-PEM outcomes [62,63], and early intervention intensity has been associated with improved home participation and environmental support [41]. Together, this evidence supports the relevance of participation-centered, family-focused intervention frameworks.

Motor development improved substantially. AIMS scores indicated significant gains in gross motor performance, consistent with previous clinic-based HEP studies [24,40]. These findings align with enriched environment-based interventions such as SAFE [21], STEP [62], and telehealth physiotherapy models [60], which have demonstrated positive effects on motor outcomes. The AIMS appears sensitive to motor change across both in-person and telehealth formats.

IMP findings further demonstrated significant improvements in variation, adaptability, fluency, performance, and total scores, reflecting enhancement in both motor quality and milestone acquisition. The absence of change in symmetry likely reflects high baseline symmetry, as similarly reported in CareToy studies [64]. Comparable improvements in IMP domains have been observed in enriched and family-centered models, including CareToy [64], COPCA [65], and EMI-Heart telehealth interventions [66]. Previous HEP studies have also demonstrated positive motor outcomes across multiple assessment tools [22,24,39,40,67]. Together, these results support the IMP's

XSL•FO
RenderX

responsiveness and reinforce the motor benefits of enriched, parent-guided early intervention.

ASQ findings revealed improvements in communication, gross motor, fine motor, personal-social, and social-emotional domains. These gains parallel performance-based motor improvements (AIMS and IMP) and are consistent with prior HEP research [24,40]. Similar multidomain improvements have been reported in enriched early intervention trials, such as the GAME [68,69] and structured physiotherapy programs [70], as well as in Alberta FICare models [71]. These findings suggest that family-centered enriched interventions may extend beyond motor domains, supporting broader sociocommunicative and cognitive development. In telehealth settings, the ASQ provides a practical, sensitive parent-reported measure for monitoring multidimensional progress.

Caregivers demonstrated significant reductions in depression, anxiety, and stress scores. These results mirror prior HEP findings [24] and align with evidence from other early intervention programs for parents of preterm infants [72]. The Mother-Infant Transaction Program [73], home-based postdischarge models [74], and family-centered neonatal care programs [75] have similarly reported improvements in parental psychological well-being. The consistent use of the DASS-21 across these studies supports its utility for detecting meaningful changes in caregiver emotional health.

Despite promising findings, several limitations must be acknowledged. The small sample size and lack of a control group limit causal inference and generalizability. Developmental maturation cannot be entirely excluded as a contributing factor. Long-term effects were not evaluated. Future studies should include larger randomized controlled trials, longitudinal follow-up, hybrid delivery models, cost-effectiveness analyses, and examination of cultural and socioeconomic influences on participation.

Conclusions

This exploratory feasibility research established that the telehealth implementation of the HEP Approach is a feasible, safe, and well-received early intervention paradigm for premature infants at developmental risk. High retention, complete session adherence, and elevated caregiver satisfaction percentages demonstrate that remote implementation of the HEP Approach is both feasible and based on family needs. In addition to feasibility, notable enhancements were evident across various developmental domains, including gross and fine motor skills, communication, social-emotional development, participation outcomes, and caregiver psychological well-being. The findings indicate that the fundamental elements of the HEP model, which prioritize environmental enrichment, parental guidance, and active engagement, can be effectively adapted to telehealth formats without undermining therapeutic integrity.

Despite the promising results, they must be approached with caution due to the limited sample size and the single-group design. Future research must include larger randomized controlled trials and longitudinal follow-up to assess the long-term effects, cost-effectiveness, and generalizability of the telehealth-based HEP Approach across diverse target groups and cultural contexts. The current findings, however, offer initial evidence that the HEP telehealth model constitutes a promising and scalable approach to accessible, family-centered early developmental care for preterm infants.

Acknowledgments

The authors thank all of the families who participated in this study.

Data Availability

The data are not publicly available due to ethical and privacy restrictions, but are available from the corresponding author upon reasonable request.

Funding

This research received no external funding.

Authors' Contributions

AB, AFAB, GCS, and TM-B contributed to the conceptualization of the study. The methodology was developed by HGS, GCS, IMD, AB, and TM-B. Formal analysis was conducted by AFAB, GI, HGS, and TM-B. The investigation was carried out by IMD and GCS. Resources were provided by AFAB, GI, HGS, IMD, TM-B, and AB. Data curation was performed by AFAB, HGS, GI, IMD, GCS, and AB. AB and GCS prepared the original draft of the manuscript. AB, GI, and TM-B contributed to the review and editing of the manuscript. Supervision was provided by AB and TM-B. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of the Homeostasis–Enrichment–Plasticity Approach phases in telehealth delivery.
[DOCX File , 19 KB - jmir_v28i1e86883_app1.docx]

XSL•FO
RenderX

References

1. Bradley E, Blencowe H, Moller A, Okwaraji YB, Sadler F, Gruending A, et al. Born too soon: global epidemiology of preterm birth and drivers for change. Reprod Health 2025;22(Suppl 2):105 [FREE Full text] [doi: 10.1186/s12978-025-02033-x] [Medline: 40551192]
2. Hee Chung E, Chou J, Brown KA. Neurodevelopmental outcomes of preterm infants: a recent literature review. Transl Pediatr 2020;9(Suppl 1):S3-S8 [FREE Full text] [doi: 10.21037/tp.2019.09.10] [Medline: 32206579]
3. Sarda S, Sarri G, Siffel C. Global prevalence of long-term neurodevelopmental impairment following extremely preterm birth: a systematic literature review. J Int Med Res 2021;49(7):3000605211028026 [FREE Full text] [doi: 10.1177/03000605211028026] [Medline: 34284680]
4. You J, Shamsi BH, Hao MC, Cao CH, Yang WY. A study on the neurodevelopment outcomes of late preterm infants. BMC Neurol 2019;19(1):108 [FREE Full text] [doi: 10.1186/s12883-019-1336-0] [Medline: 31146703]
5. Song I. Neurodevelopmental outcomes of preterm infants. Clin Exp Pediatr 2023;66(7):281-287 [FREE Full text] [doi: 10.3345/cep.2022.00822] [Medline: 36596743]
6. Chawanpaiboon S, Vogel JP, Moller AB, Lumbiganon P, Petzold M, Hogan D, et al. Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis. Lancet Glob Health 2019;7(1):e37-e46 [FREE Full text] [doi: 10.1016/S2214-109X(18)30451-0] [Medline: 30389451]
7. Khurana S, Kane AE, Brown SE, Tarver T, Dusing SC. Effect of neonatal therapy on the motor, cognitive, and behavioral development of infants born preterm: a systematic review. Dev Med Child Neurol 2020;62(6):684-692 [FREE Full text] [doi: 10.1111/dmcn.14485] [Medline: 32077096]
8. Zhang Q, Dong X, Jin W, Fan J. Early brain cognitive development in late preterm infants: an event-related potential and resting EEG study. Ital J Pediatr 2024;50(1):26 [FREE Full text] [doi: 10.1186/s13052-023-01567-9] [Medline: 38355639]
9. Nivins S, Padilla N, Kvanta H, Ådén U. Gestational age and cognitive development in childhood. JAMA Netw Open 2025;8(4):e254580 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.4580] [Medline: 40227687]
10. Ko J, Lim HK. Motor development comparison between preterm and full-term infants using alberta infant motor scale. Int J Environ Res Public Health 2023;20(5):3819 [FREE Full text] [doi: 10.3390/ijerph20053819] [Medline: 36900838]
11. Kobaş M, Kızıldere E, Doğan I, Aktan-Erciyes A, Demir-Lira Ö, Akman I, et al. Motor skills, language development, and visual processing in preterm and full-term infants. Curr Psychol 2023;42(15):12463-12475 [FREE Full text] [doi: 10.1007/s12144-021-02658-8]
12. Yildiz R, Yildiz A, Zorlular R, Elbasan B. Relationship between sensory processing skills and motor skills in 12-month-old infants. Brain Behav 2024;14(9):e70052 [FREE Full text] [doi: 10.1002/brb3.70052] [Medline: 39317996]
13. Sırtbaş-Işık G, Yardımcı-Lokmanoğlu BN, Livanelioğlu A, Mutlu A. Sensory processing and sleep characteristics in preterm infants in the early period of life. Sleep Med 2023;106:78-83. [doi: 10.1016/j.sleep.2023.03.025] [Medline: 37054558]
14. Vandormael C, Schoenhals L, Hüppi PS, Filippa M, Borradori Tolsa C. Language in preterm born children: atypical development and effects of early interventions on neuroplasticity. Neural Plast 2019;2019:6873270 [FREE Full text] [doi: 10.1155/2019/6873270] [Medline: 30930944]
15. Palomero-Sierra B, Sánchez-Gómez V, Magán-Maganto M, Bejarano-Martín Á, Ruiz-Ayúcar I, de Vena-Díez VB, et al. Early social communication and language development in moderate-to-late preterm infants: a longitudinal study. Front Psychol 2025;16:1556416 [FREE Full text] [doi: 10.3389/fpsyg.2025.1556416] [Medline: 40248834]
16. Hass J, Panceri C, Procianoy RS, Silveira RDC, Valentini NC. Risk factors for cognitive, motor and language development of preterm children in the first year of life. Rev Paul Pediatr 2022;41:e2021165 [FREE Full text] [doi: 10.1590/1984-0462/2023/41/2021165] [Medline: 36169521]
17. Fan J, Wang J, Zhang X, He R, He S, Yang M, et al. A home-based, post-discharge early intervention program promotes motor development and physical growth in the early preterm infants: a prospective, randomized controlled trial. BMC Pediatr 2021;21(1):162 [FREE Full text] [doi: 10.1186/s12887-021-02627-x] [Medline: 33827496]
18. Orton J, Doyle LW, Tripathi T, Boyd R, Anderson PJ, Spittle A. Early developmental intervention programmes provided post hospital discharge to prevent motor and cognitive impairment in preterm infants. Cochrane Database Syst Rev 2024;2(2):CD005495 [FREE Full text] [doi: 10.1002/14651858.CD005495.pub5] [Medline: 38348930]
19. DeMaster D, Bick J, Johnson U, Montroy JJ, Landry S, Duncan AF. Nurturing the preterm infant brain: leveraging neuroplasticity to improve neurobehavioral outcomes. Pediatr Res 2019;85(2):166-175. [doi: 10.1038/s41390-018-0203-9] [Medline: 30531968]
20. Rocha-Ferreira E, Hristova M. Plasticity in the neonatal brain following hypoxic-ischaemic injury. Neural Plast 2016;2016:4901014 [FREE Full text] [doi: 10.1155/2016/4901014] [Medline: 27047695]
21. Apaydın U, Yıldız R, Yıldız A, Acar Ş, Gücüyener K, Elbasan B. Short-term effects of SAFE early intervention approach in infants born preterm: a randomized controlled single-blinded study. Brain Behav 2023;13(10):e3199 [FREE Full text] [doi: 10.1002/brb3.3199] [Medline: 37534617]
22. Balikci A, Aracikul Balikci AF, Demirbag M, Sirma GC, Beaudry-Bellefeuille I, May-Benson TA. Implementation of the homeostasis-enrichment-plasticity (HEP®) approach for an infant at risk for autism spectrum disorder: a case report. J Child Sci 2025;15(1) [FREE Full text] [doi: 10.53391/2474-5871.1005]

XSL•FO
RenderX

23. Bernabe-Zuñiga JE, Rodriguez-Lucenilla MI, Alias-Castillo AJ, Rueda-Ruzafa L, Roman P, Del Mar Sanchez-Joya M. Early interventions with parental participation and their implications on the neurodevelopment of premature children: a systematic review and meta-analysis. Eur Child Adolesc Psychiatry 2025;34(3):853-865 [FREE Full text] [doi: 10.1007/s00787-024-02528-1] [Medline: 39028424]
24. Sirma G, Zengin Alpozgen A, Balikci A. Effects of the HEP® (Homeostasis-Enrichment-Plasticity) approach in preterm infants with increased developmental risk: a randomized controlled study. Front Pediatr 2025;13:1606490 [FREE Full text] [doi: 10.3389/fped.2025.1606490] [Medline: 41080046]
25. Filippa M, Lordier L, Lejeune F, De Almeida JS, Hüppi PS, Barcos-Munoz F, et al. Effect of an early music intervention on emotional and neurodevelopmental outcomes of preterm infants at 12 and 24 months. Front Psychol 2024;15:1443080 [FREE Full text] [doi: 10.3389/fpsyg.2024.1443080] [Medline: 39498332]
26. Smith L, McParlan A, Niss T, Fucile S. Longitudinal outcomes of a parent-administered sensorimotor intervention in preterm infants at 4 and 18 months. Early Hum Dev 2025;201:106171. [doi: 10.1016/j.earlhumdev.2024.106171] [Medline: 39662236]
27. Little A, Kamholz K, Corwin BK, Barrero-Castillero A, Wang CJ. Understanding barriers to early intervention services for preterm infants: lessons from two states. Acad Pediatr 2015;15(4):430-438 [FREE Full text] [doi: 10.1016/j.acap.2014.12.006] [Medline: 26142069]
28. Altunalan T, Sari Z, Saglam S, Çelik M, Utku Umut G. Investigating enrollment and continuity of parents to early childhood intervention services. Turk J Appl Soc Work 2020;3(2):45-54 [FREE Full text]
29. Decker K, Williams ER, Cook GA, Fry MM. The early intervention referral process for rural infants and toddlers with delays or disabilities: a family perspective. Matern Child Health J 2021;25(5):715-723 [FREE Full text] [doi: 10.1007/s10995-020-03067-2] [Medline: 33201452]
30. Namiiro F, Nolens B, Rujumba J, Kiguli S, Batte A, van den Akker T. "My baby is fine, no need for more clinic visits." Facilitators and barriers for utilisation of follow-up services for children born preterm in low-resource setting: parents' perceptions. Trop Med Int Health 2023;28(3):194-202 [FREE Full text] [doi: 10.1111/tmi.13857] [Medline: 36708235]
31. Sarik DA, Matsuda Y, Terrell EA, Sotolongo E, Hernandez M, Tena F, et al. J Pediatr Nurs 2022;67:139-147. [doi: 10.1016/j.pedn.2022.09.003] [Medline: 36116347]
32. Pineda R, Kellner P, Ghahramani S, Dusing S, Chapman R, Peden CJ. Feasibility of a telehealth model of therapy service delivery after neonatal intensive care unit (NICU) discharge. OTJR (Thorofare N J) 2025:15394492251362721. [doi: 10.1177/15394492251362721] [Medline: 40814234]
33. DeMauro S, Duncan AF, Hurt H. Telemedicine use in neonatal follow-up programs - what can we do and what we can't - lessons learned from COVID-19. Semin Perinatol 2021;45(5):151430 [FREE Full text] [doi: 10.1016/j.semperi.2021.151430] [Medline: 33892961]
34. Das A, Cina L, Mathew A, Aziz H, Aly H. Telemedicine, a tool for follow-up of infants discharged from the NICU? experience from a pilot project. J Perinatol 2020;40(6):875-880. [doi: 10.1038/s41372-020-0593-5] [Medline: 31959907]
35. Peskin A, Barth A, Andrew Rothenberg W, Turzi A, Formoso D, Garcia D, et al. New therapy for a new normal: comparing telehealth and in-person time-limited parent-child interaction therapy. Behav Ther 2024;55(1):106-121. [doi: 10.1016/j.beth.2023.05.002] [Medline: 38216225]
36. Rabner J, Norris LA, Olino TM, Kendall PC. A comparison of telehealth and in-person therapy for youth anxiety disorders. J Clin Child Adolesc Psychol 2025;54(5):555-566. [doi: 10.1080/15374416.2024.2372770] [Medline: 38995683]
37. Lung T, Hayes A, Wen LM, Xu H, Brown V, Baur LA, et al. Socioeconomic differences in the cost-effectiveness of a telephone-based intervention for obesity prevention in early childhood. Int J Obes (Lond) 2026;50(1):152-158. [doi: 10.1038/s41366-025-01904-4] [Medline: 40914767]
38. Salsabilla A, Azzahra AB, Syafitri RIP, Supadmi W, Suwantika AA. Cost-effectiveness of telemedicine in Asia: a scoping review. J Multidiscip Healthc 2021;14:3587-3596 [FREE Full text] [doi: 10.2147/JMDH.S332579] [Medline: 35002248]
39. Balıkcı A. Exploring effects of the HEP (homeostasis-enrichment-plasticity) approach as a comprehensive therapy intervention for an infant with cerebral palsy: a case report. J Child Sci 2022;12(01):e182-e195 [FREE Full text] [doi: 10.1055/s-0042-1757913]
40. Balikci A, May-Benson TA, Sirma GC, Kardas A, Demirbas D, Aracikul Balikci AF, et al. The homeostasis-enrichment-plasticity (HEP) approach for premature infants with developmental risks: a pre-post feasibility study. J Clin Med 2024;13(18):5374 [FREE Full text] [doi: 10.3390/jcm13185374] [Medline: 39336861]
41. Khetani M, Graham JE, Davies PL, Law MC, Simeonsson RJ. Psychometric properties of the young children's participation and environment measure. Arch Phys Med Rehabil 2015;96(2):307-316 [FREE Full text] [doi: 10.1016/j.apmr.2014.09.031] [Medline: 25449189]
42. Williams U, Law M, Hanna S, Gorter J. Using the young children’s participation and environment measure (YC-PEM) to describe young children’s participation and relationship to disability and complexity. J Dev Phys Disabil 2018;31(1):135-148 [FREE Full text] [doi: 10.1007/s10882-018-9637-6]
43. Seyhan-Bıyık K, Delioğlu K, Özal C, Üneş S, Tunçdemir M, Kerem-Günel M. Young children's participation and environment measure (YC-PEM): cross-cultural adaptation, reliability, and validity of Turkish version. Percept Mot Skills 2024;131(6):2125-2149. [doi: 10.1177/00315125241287971] [Medline: 39340534]

XSL•FO
RenderX

44. Kapci E, Kucuker S, Uslu R. How applicable are ages and stages questionnaires for use with Turkish children? Top Early Child Spec Educ 2010;30(3):176-188 [FREE Full text] [doi: 10.1177/0271121410373149]
45. Piper MC, Pinnell LE, Darrah J, Maguire T, Byrne PJ. Construction and validation of the Alberta infant motor scale (AIMS). J Public Health 1992;83(Suppl 2):S46-S50 [FREE Full text] [Medline: 1468050]
46. Piper MC, Darrah J. Motor Assessment of the Developing Infant. Philadelphia, PA: Saunders Co; 1994.
47. Almeida K, Dutra MVP, Mello RRD, Reis ABR, Martins PS. Concurrent validity and reliability of the Alberta infant motor scale in premature infants. J Pediatr (Rio J) 2008;84(5):442-448. [doi: 10.2223/JPED.1836] [Medline: 18923790]
48. Davies S, Lucas BR, Dwyer GM. Reliability of the Alberta infant motor scale (AIMS) when used via telehealth for neurodevelopmentally high-risk infants. Phys Occup Ther Pediatr 2025;45(4):558-571 [FREE Full text] [doi: 10.1080/01942638.2025.2451406] [Medline: 39865795]
49. Rawnsley K, Cheong JLY, Bennett K, Mahady ML, Smith SE, Zannino D, et al. Synchronous telehealth and face-to-face administration of the Alberta infant motor scale. Dev Med Child Neurol 2026;68(1):91-98 [FREE Full text] [doi: 10.1111/dmcn.16391] [Medline: 40576266]
50. Hadders-Algra M, Heineman KR. The Infant Motor Profile. New York, NY, USA: Routledge; 2021.
51. Akbaş AN, Çankaya Ö, Seyhan K, Günel MK. Inter-rater reliability of infant motor profile in 3-24-month-old high-risk infants: Turkey sample. Arc Health Sci Res 2020;6(7) [FREE Full text] [doi: 10.5152/archealthscires.2020.575433]
52. Rizzi R, Menici V, Cioni ML, Cecchi A, Barzacchi V, Beani E, Clinical CareToy-R Consortium. Concurrent and predictive validity of the infant motor profile in infants at risk of neurodevelopmental disorders. BMC Pediatr 2021;21(1):68 [FREE Full text] [doi: 10.1186/s12887-021-02522-5] [Medline: 33549070]
53. Lovibond PF, Lovibond SH. The structure of negative emotional states: comparison of the depression anxiety stress scales (DASS) with the beck depression and anxiety inventories. Behav Res Ther 1995;33(3):335-343. [doi: 10.1016/0005-7967(94)00075-u] [Medline: 7726811]
54. Yılmaz Ö, Boz H, Arslan A. Depresyon anksi˙yete stres ölçeği˙ni˙n (DASS 21) Türkçe kisa formunun geçerli˙li˙k-güveni˙li˙rli˙k çalişmasi [validity and reliability study of the Turkish short form of the Depression Anxiety Stress Scale (DASS-21)]. Res Financ Econ Soc Stud 2017;2:78-91 [FREE Full text]
55. Tabachnick B, Fidell L, Ullman J. Using Multivariate Statistics. Boston, MA, USA: Pearson; 2013.
56. George D, Mallery M. SPSS for Windows Step by Step: A Simple Guide and Reference. Boston, MA, USA: Pearson; 2010.
57. Cheung W, Aleman-Tovar J, Johnston AN, Little LM, Burke MM. A qualitative study exploring parental perceptions of telehealth in early intervention. J Dev Phys Disabil 2023;35(3):353-373 [FREE Full text] [doi: 10.1007/s10882-022-09853-w] [Medline: 35789590]
58. Hinton V, Akemoğlu Y, Tomeny K, McWilliam RA. Supporting families from a distance: implementing routines-based home visits via telepractice. Early Child Educ J 2023:1-8 [FREE Full text] [doi: 10.1007/s10643-023-01464-0] [Medline: 37360610]
59. Jimenez-Arberas E, Casais-Suarez Y, Fernandez-Mendez A, Menendez-Espina S, Rodriguez-Menendez S, Llosa JA, et al. Evidence-based implementation of the family-centered model and the use of tele-intervention in early childhood services: a systematic review. Healthcare (Basel) 2024;12(1):112 [FREE Full text] [doi: 10.3390/healthcare12010112] [Medline: 38201017]
60. Schlichting T, Martins da Silva K, Silva Moreira R, Marques de Moraes MV, Cicuto Ferreira Rocha NA, Boyd RN, et al. Telehealth program for infants at risk of cerebral palsy during the covid-19 pandemic: a pre-post feasibility experimental study. Phys Occup Ther Pediatr 2022;42(5):490-509. [doi: 10.1080/01942638.2022.2057209] [Medline: 35341469]
61. Hielkema T, Hamer EG, Boxum Anke G, La Bastide-Van Gemert S, Dirks T, Reinders-Messelink HA, L2M 0-2 Study Group. LEARN2MOVE 0-2 years, a randomized early intervention trial for infants at very high risk of cerebral palsy: neuromotor, cognitive, and behavioral outcome. Disabil Rehabil 2020;42(26):3752-3761 [FREE Full text] [doi: 10.1080/09638288.2019.1610508]
62. Lima C, Abreu RWFD, Verdério BN, Brugnaro BH, Santos MMD, Dos Santos AN, et al. Early intervention involving specific task-environment-participation (step) protocol for infants at risk: a feasibility study. Phys Occup Ther Pediatr 2023;43(3):303-320. [doi: 10.1080/01942638.2022.2142084] [Medline: 36329671]
63. Lima C, de Abreu RWF, Brugnaro BH, Rocha NACF. Participation of infants at biological risk is facilitated by remote intervention carried out by parents – step protocol: randomized clinical trial. Braz J Phys Ther 2024;28:100674 [FREE Full text] [doi: 10.1016/j.bjpt.2024.100674]
64. Sgandurra G, Lorentzen J, Inguaggiato E, Bartalena L, Beani E, Cecchi F, CareToy Consortium. A randomized clinical trial in preterm infants on the effects of a home-based early intervention with the 'CareToy System'. PLoS One 2017;12(3):e0173521 [FREE Full text] [doi: 10.1371/journal.pone.0173521] [Medline: 28328946]
65. Akhbari Ziegler S, von Rhein M, Meichtry A, Wirz M, Hielkema T, Hadders-Algra M. The coping with and caring for infants with special needs intervention was associated with improved motor development in preterm infants. Acta Paediatr 2021;110(4):1189-1200 [FREE Full text] [doi: 10.1111/apa.15619] [Medline: 33047325]
66. Mitteregger E, Dirks T, Theiler M, Kretschmar O, Latal B. The implementation of EMI-heart, a family-tailored early motor intervention in infants with complex congenital heart disease, in practice: a feasibility RCT. Pilot Feasibility Stud 2024;10(1):105 [FREE Full text] [doi: 10.1186/s40814-024-01532-1] [Medline: 39095881]

XSL•FO
RenderX

67. Balikci A, May-Benson TA, Sirma GC, Ilbay G. HEP (homeostasis-enrichment-plasticity) approach changes sensory-motor development trajectory and improves parental goals: a single subject study of an infant with hemiparetic cerebral palsy and twin anemia polycythemia sequence (TAPS). Children (Basel) 2024;11(7):876 [FREE Full text] [doi: 10.3390/children11070876] [Medline: 39062325]
68. Morgan C, Novak I, Dale RC, Guzzetta A, Badawi N. Single blind randomised controlled trial of GAME (goals - activity - motor enrichment) in infants at high risk of cerebral palsy. Res Dev Disabil 2016;55:256-267. [doi: 10.1016/j.ridd.2016.04.005] [Medline: 27164480]
69. Gündoğmuş E, Bumin G, Yalçın SS. Effect of early intervention on developmental domains and parent-child interaction among children with developmental delay: a randomized controlled study. Am J Occup Ther 2024;78(6):7806205110. [doi: 10.5014/ajot.2024.050706] [Medline: 39418650]
70. Ochandorena-Acha M, Terradas-Monllor M, López Sala L, Cazorla Sánchez ME, Fornaguera Marti M, Muñoz Pérez I, et al. Early physiotherapy intervention program for preterm infants and parents: a randomized, single-blind clinical trial. Children (Basel) 2022;9(6):895 [FREE Full text] [doi: 10.3390/children9060895] [Medline: 35740832]
71. Moe AM, Kurilova J, Afzal AR, Benzies KM. Effects of Alberta family integrated care (FICare) on preterm infant development: two studies at 2 months and between 6 and 24 months corrected age. J Clin Med 2022;11(6):1684 [FREE Full text] [doi: 10.3390/jcm11061684] [Medline: 35330009]
72. Finlayson F, Olsen J, Dusing SC, Guzzetta A, Eeles A, Spittle A. Supporting play, exploration, and early development intervention (SPEEDI) for preterm infants: a feasibility randomised controlled trial in an Australian context. Early Hum Dev 2020;151:105172. [doi: 10.1016/j.earlhumdev.2020.105172] [Medline: 33137579]
73. Landsem I, Handegård BH, Tunby J, Ulvund SE, Rønning JA. Early intervention program reduces stress in parents of preterms during childhood, a randomized controlled trial. Trials 2014;15:387 [FREE Full text] [doi: 10.1186/1745-6215-15-387] [Medline: 25282345]
74. Fan J, He R, He S, Yang M, Tao X, Zhou M, et al. Effectiveness of a home-based, post-discharge early intervention program for very preterm infants in reducing parental stress: a randomized controlled trial. BMC Public Health 2024;24(1):2476 [FREE Full text] [doi: 10.1186/s12889-024-19969-7] [Medline: 39261874]
75. Pourkhani S, Chehrzad MM, Reza Masouleh S, Kazem Nezhad Leyli E. The effect of family-based care on stress, anxiety, and depression of mothers with premature infants. J Holist Nurs Midwifery 2018;28(2):121-128 [FREE Full text] [doi: 10.29252/hnmj.28.2.121]

Abbreviations

AIMS: Alberta Infant Motor Scale
ASQ: Ages and Stages Questionnaire
DASS-21: Depression Anxiety Stress Scales
EI: early intervention
HEP: Homeostasis–Enrichment–Plasticity
IMP: Infant Motor Profile
PEO: Person–Environment–Occupation
YC-PEM: Young Children's Participation and Environment Measure

Edited by M Balcarras; submitted 31.Oct.2025; peer-reviewed by U Apaydin, B Elbasan; comments to author 11.Feb.2026; revised version received 04.Mar.2026; accepted 04.Mar.2026; published 07.Apr.2026.

Please cite as:
Balikci A, Sirma GC, Demirbag IM, May-Benson TA, Sozen HG, Aracikul Balikci AF, Ilbay G
Telehealth Delivery of the Homeostasis–Enrichment–Plasticity Approach for Premature Infants With Developmental Risks: Exploratory Feasibility Study
J Med Internet Res 2026;28:e86883
URL: https://www.jmir.org/2026/1/e86883
doi:10.2196/86883
PMID:

©Aymen Balikci, Gamze Cagla Sirma, Izgi-Miray Demirbag, Teresa A May-Benson, Hatice Gulhan Sozen, Ayse Firdevs Aracikul Balikci, Gul Ilbay. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Willingness to Share Internet Use Data for Research on Early Disease Detection: Cross-Sectional Survey

Christina Derksen[1], BSc, MSc, PhD; Mel Ramasawmy[1], BSc, MSc, PhD; Sanjula Arora[1], BSc, MSc; Wynsee Lau[2]; Suzanne E Scott[1], BSc, MSc, PhD

[1]Wolfson Institute of Population Health, Queen Mary University of London, Charterhouse Square, London, United Kingdom

[2]Faculty of Medicine, Imperial College London, London, United Kingdom

Corresponding Author:
Christina Derksen, BSc, MSc, PhD
Wolfson Institute of Population Health, Queen Mary University of London, Charterhouse Square, London, United Kingdom

Abstract

Background: Preliminary research has suggested that internet use data could offer digital signals of early disease and has the potential to facilitate early detection and improve patient outcomes. However, there are significant challenges in linking individual-level internet use data with health outcomes. One key aspect is that the public might not be willing to share data for research or that selective data sharing might create bias in datasets and increase inequalities.

Objective: Our study aimed to investigate the willingness of the public to share their internet use data for medical research and to identify key criteria that affect willingness to share.

Methods: We conducted a web-based, cross-sectional online survey with 2390 UK adults with and without a history of cancer, heart disease, and depression using quota sampling. Participants were randomly assigned to explore willingness to share different types of internet use data for 1 of 3 health conditions (cancer, heart disease, and depression) and for provision of a pictorial example of internet use data. Logistic regression analysis (α=.05) for each condition was used to determine key factors of willingness to share, including sociodemographics and attitudes toward sharing. Open-ended comments regarding facilitators of sharing and concerns were analyzed thematically.

Results: Willingness to share internet use data was high across conditions (74% - 77%, 95% CI 70.5% - 80.3%), especially for health app data (73% - 76%, 95% CI 69.8% - 79.1%). The pictorial example of browsing history did not affect willingness to share. For all conditions, factors consistently associated with willingness to share were perceived benefits (odds ratios [ORs] 5.692 - 8.850; all *P*<.001) and concerns (ORs 0.343 - 0.432; all *P*<.001). Key concerns were data privacy, potential for misuse, and lack of relevance. Suggestions to increase willingness to share included contributing to society and research, data security assurances, clarification of research purposes, and monetary incentives. Familiarity with internet use data was related to lower willingness to share for heart disease detection (OR 0.740, 95% CI 0.561 - 0.976). Asian ethnicity was associated with lower willingness to share internet use for cancer detection (OR 0.234, 95% CI 0.076 - 0.723). Younger age (OR 0.975, 95% CI 0.951 - 0.999) and male gender (OR 2.615, 95% CI 1.511 - 4.526) were associated with higher willingness to share data for depression detection.

Conclusions: This first large-scale assessment of public willingness to share internet use data for early disease detection adds novel insights by comparing conditions and examining sociodemographic factors alongside perceived benefits and risks. It highlights that understanding of internet data is limited yet willingness to share for research is high. Clear communication of benefits, strong privacy protections, and incentives may increase participation and reduce bias. The findings inform consent design, targeted outreach to underrepresented groups, and data governance for safe use of personal digital data. Future research should focus on improving public communication, particularly among less willing groups at risk of inequality.

(J Med Internet Res 2026;28:e85637) doi:10.2196/85637

KEYWORDS

early detection; internet use data; willingness to share online data; public perceptions; data privacy

Introduction

Across many health conditions, early diagnosis is crucial for successful treatment and long-term outcomes, potentially impacting survival and health care costs [1-3]. Early diagnosis depends on detection of signs and symptoms, presentation to health care professionals, and timely diagnostic tests. However, patients often present to health care weeks or months after the onset of symptoms, and a key reason for this delay is the time taken to interpret symptoms as potentially serious [4,5]. Thus,

XSL•FO
RenderX

the understanding of timing of symptom perception, symptom interpretation, and help-seeking behavior is vital in efforts to encourage early diagnosis.

"Digital footprints" capture huge amounts information, including online activity, in-store purchases, wearable device data, and social media engagement [6]. These digital data may provide insights into symptoms, changes in behavior, and the decision to seek help [7,8], and using these data may tackle current research limitations in this area. To date, researchers studying time to presentation have relied on retrospective self-report from patients recently diagnosed with the condition under investigation [9]. These data are subject to recall bias and could be affected by hindsight from the recent diagnosis [10,11].

In the United Kingdom, there are almost 50 million health-related searches per year using Google [12]. Internet use data include internet search history, calendar data, bookmarks, contacts, emails, photos, reminders, profiles, and documents stored online. Given the richness of these data, there is growing interest to explore the use of internet search data and other digital footprints as digital signals of serious disease. Preliminary research has begun to examine whether data on internet search history could be used to predict diagnosis of cancers and mental health conditions, with promising results. Changes in the frequency, content, and tone of online searches were evident up to 60 days before suicide attempts or general practitioner referral for gynecological cancers, up to 21 weeks before pancreatic cancer diagnosis, and up to 1 year before psychiatric hospitalization [13-16].

Despite the promise of the digital signals, there are multiple questions that need to be explored regarding the acceptability of this kind of research for patients and the public, especially if research involves sharing internet use data on an individual level and linking with future health outcomes such as a cancer diagnosis. Previous studies have mixed results regarding people's willingness to share different types of digital data over "traditional" types of health data such as biomedical samples or questionnaire data [17,18]. It is unclear if internet users are aware that these data exist, where they are stored, and how they could be used in research or if effective ways of explaining these issues to potential participants exist. This is key to informed consent.

Willingness to share might depend on several factors, including the type of data being shared, the way internet data are explained, the purpose of the research, privacy concerns, and sociodemographic factors [19]. If some groups are less willing to provide internet use data, findings from future research might not be representative and could thus perpetuate health inequalities. For example, gender, age, ethnicity, or socioeconomic status such as employment might influence willingness to share internet use data and other types of data such as location data or measures from health and fitness apps [19-21]. Further, the type of condition being studied and the relevance to an individual (eg, through personal history or experiences of friends and family being affected by a mental or physical health condition) can influence willingness to participate in medical research [22], yet it is unknown if this applies to willingness to share internet use data.

This study therefore investigated public perspectives on the use of internet use data for medical research to identify key criteria that could enhance acceptability and avoid inequalities in engagement. Our research questions were exploratory:

1. How willing are the public to share different types of internet use data for early diagnosis research?
2. Does willingness to share internet use data vary for different mental and physical health conditions?
3. Does the way internet use data is explained impact willingness to share internet use data?
4. Are sociodemographic characteristics, personal history of different mental and physical health conditions, familiarity with internet use data, and attitudes toward data sharing associated with willingness to share internet use data?

Methods

Study Design

We used a cross-sectional design to conduct an online survey on willingness to share different types of internet use data. We randomly assigned participants using an automated algorithm based on 2 factors: (1) the medical condition about which participants are asked whether they would be willing to share their data (cancer, heart disease, and depression) and (2) the way internet use data was explained (written text plus a screenshot as a pictorial example of internet use data vs written text only). Cancer, depression, and heart disease were selected, as these are common conditions for which early diagnosis is key and there is emerging evidence for use of digital data in prediction models [13-16]. We reported the study in line with Journal Article Reporting Standards Quantitative Research Design (JARS-Quant) standards [23].

Ethical Considerations

Ethical approval for research with human subjects was granted by the Queen Mary Ethics of Research Committee (reference number QME25.0924). Participants received the participant information sheet for download at the beginning of the survey and provided informed consent for participation and data sharing before they could access the survey. All data were collected anonymously. In line with Dynata's company policies, participants received tokens through the loyalty programs set up through Dynata partnerships automatically upon survey completion, which could be redeemed for vouchers or gifts. Incentives were valued at approximately £0.75 (US $1.00). No identification of individual participants in any images of the manuscript or appendices is possible.

Participants and Procedure

Participants were adult members of the public (18 years or older), both with and without a personal history of cancer, heart disease, or depression. The survey was hosted on the online platform Qualtrics. Participants were recruited through Dynata from participants who had previously signed up to an online panel from July 1, 2025, to July 23, 2025. As the panel was English-speaking and conducted initial checks with new participants, we assumed all participants would be proficient in the English language. Before full recruitment, the survey was pilot tested with patient and public representatives (see the

XSL•FO
RenderX

following sections). To ensure data quality, it was then released to 10% of the final population to prevent technical problems and unclear questions before full roll-out.

The survey was administered using quota sampling based on gender (approximately 50% male and 50% female), age (approximately 50% of the population over the age of 50 years), income (approximately 50% with disposable household income levels <£35,000), and ethnicity (maximum 85% of the population from a White ethnic group). A minimum of 300 participants with a history of cancer, heart disease, or depression was sought. We chose to sample for age to increase the proportion of older participants, who have a higher incidence of cancer and cardiovascular disease. Online survey panels often have a disproportionately higher share of younger respondents [24], and we anticipated that younger participants would be more likely to take part in a survey on the topic of internet use and data.

Measures

The survey assessed willingness to share data for early diagnosis research, attitudes toward data sharing, familiarity about internet use data, and sociodemographics. All items and the pictorial example can be found in Multimedia Appendix 1.

Willingness to Share Internet Use Data for Early Diagnosis Research and Attitudes Toward Data Sharing

The main outcome measure was willingness to share data, assessed as a single item adapted from Hirst et al [19] with 4 answer options (definitely yes to definitely no). Where appropriate, this was recoded to a binary variable (yes, no) for analyses. The same question was used to assess willingness to share specific types of internet use data (eg, internet search history or online purchases), with an option to choose "Not applicable (I do not use this)."

Open-ended questions were used to ask participants what would increase their willingness to share internet use data, as well as their concerns about data sharing.

Attitudes toward data sharing were assessed as perceived benefits and concerns or worries using an adaptation of scales by Sanderson et al [24] and Seltzer et al [18]. We used matrix questions to include perceived benefits and concerns and calculated mean scores (Cronbach α=0.96 for perceived benefits and 0.94 for concerns).

Familiarity With Internet Use

We assessed participants' self-reported familiarity with internet use data by adapting the scale by Hirst et al [19] on awareness about General Data Protection Regulation (GDPR). We asked whether participants had heard about internet use data before. Objective knowledge was assessed by asking what participants thought was part of internet use data from a list of examples (eg, internet search history as example of internet use data or content of emails as distractor). Correct answers were counted to calculate a knowledge score (range 1 - 10).

Sociodemographics and Current Internet Use

For sociodemographic characteristics, we assessed age, gender, educational level, employment, marital status, ethnicity, and preferred language. Items were adapted from the DISTINCT project [25]. We assessed disposable household income (<£35,000 or >£35,000) and postcodes (to determine levels of deprivation via English indices of deprivation 2019 [26], Postcode Deprivation Finder 2019 [27], and Multiple Deprivation Measures 2017 Lookup Tool [28]). Finally, we measured self-reported history of cancer, cardiovascular disease, and depression and whether their relatives had a history of these conditions. Answer options were adapted from the Health Survey for England, 2021 [29].

To assess current internet and online service use, we used 2 questions from the UK Digital Landscape survey [30] about frequency and purpose of internet use.

Data Preparation

There were no missing data in the study variables such as willingness to share internet use data, self-reported familiarity with internet use data, and perceived benefits or concerns, as items were set to mandatory. In the sociodemographic data, <1% were missing, except for disposable household income (2.3%) and deprivation quintile (7.8%). Online surveys are open to fraudulent responses [31,32]; as a precaution, we introduced a number of steps to ensure credibility of the data. An attention check was added halfway through the survey, asking participants to select "strongly agree" with answer options appearing in a random order. Participants were excluded if they failed the attention check. We also excluded participants who were flagged by Qualtrics as both likely bot *and* fraudulent respondent (Recaptcha Score<0.5 and Fraud Score>30) or as duplicate (flagged by Qualtrics or on inspection of the data), as well as participants who answered the survey in less than 2 minutes, which was considered to be insufficient time to read through the participant information sheet and questionnaire items. Finally, we excluded participants with the common placeholder "lorem ipsum" text or questionable open answers (eg, "Banana") on both open-ended questions.

Data Analysis

We descriptively report differences in the willingness to share internet use data for early diagnosis research using percentages of participants indicating they would be willing to share. We conducted χ^2 tests for differences in proportions willing to share internet use data between medical conditions, the way internet use data were explained, and history of the condition. Logistic regression analysis was used to determine how sociodemographic characteristics, perceived benefits and concerns, knowledge, and familiarity with internet use data affect willingness to share internet use data, coded as a binary outcome. Separate regression models were run for willingness to share internet use data for each medical condition (cancer, heart disease, depression).

Responses to open-ended questions were analyzed using a thematic content analysis approach and reported alongside how willing participants were to share internet use data. Categories were developed by two authors (CD and WL) and discussed with the research team including patient and public involvement (PPI) representatives. Answers not in English (one Romanian and one Polish) were translated using Google Translate, as

XSL•FO
RenderX

translated content fit the questions, showing that participants understood English but answered in a different preferred language.

The full syntax and the final datasets can be found on Open Science Framework (OSF) [33].

Power Analysis

Based on previous research [19], we made the conservative assumption that willingness to share internet use data would likely vary between 45% and 55% of participants being willing to share. For an a priori sample size calculation, we therefore assumed a difference of 0.1 in probabilities of being willing to share data or not for a binary factor (0.55 in one group, 0.45 in the other; odds ratio [OR] 1.49). To reach a power of 80% given an α-error level of .05, we determined a sample size of 786 participants per regression. As logistic regression analysis was repeated for each condition, we aimed for a sample size of 2358 participants (3 × 786). However, the sample sizes for the logistic regression analyses were lower than the targeted sample size, so analyses were repeated without the variable with the most missing data (level of deprivation; see the following sections) for sensitivity analysis.

Patient and Public Involvement

We presented initial ideas for survey items to 3 PPI representatives who were older than 55 years and from a minority background. They were recruited from the Centre for Cancer Screening, Prevention and Early Diagnosis PPI pool at Queen Mary University of London, which is a database of more than 150 people with an interest in contributing to research. They endorsed the research but asked to add questions about emotions (fear and embarrassment) to the items on attitudes toward data sharing. The group asked that the participant information sheet emphasized that data sharing was not required for participation in this study and to change the wording of some items to facilitate understanding. Following these changes, we piloted the survey with 2 PPI representatives in think-aloud sessions to ensure face validity [34]. The PPI group was also involved in interpreting the results and discussing implications for future research and practice.

Results

Participant Characteristics

Of the 2864 participants who gave informed consent and completed the survey, 474 were excluded: 365 failed the attention check, 32 were flagged as both likely bot and fraudulent respondent, 64 were duplicate respondents, 6 completed the survey in less than 2 minutes, and 7 gave questionable open answers. Hence, 2390 participants with a mean age of 50.2 (SD 16.86, range 18 - 90) years were included in the analyses. Participant characteristics are provided in Table 1.

XSL•FO
RenderX

Table . Participant characteristics from a cross-sectional survey in 2025 conducted with UK participants recruited through an online survey panel (N=2390).

Characteristic	Results, n (%)
Gender	
Man	1250 (52.3)
Woman	1135 (47.5)
Nonbinary	4 (0.2)
Prefer not to say	1 (<0.1)
Employment status	
Employed (including self-employed)	1433 (60)
Not employed	352 (14.7)
Retired	592 (24.8)
Prefer not to say	13 (0.5)
Highest educational qualifications	
No formal qualifications	73 (3.1)
GCSEs[a] or equivalent	408 (17.1)
AS[b], A level or equivalent	296 (12.4)
NVQ[c] up to level 3 or equivalent	256 (10.7)
Qualification at degree level or above	1250 (52.3)
Apprenticeship	51 (2.1)
Other or equivalent unknown	52 (2.2)
Prefer not to say	4 (0.2)
Disposable household income (£)	
<35,000	1206 (50.5)
>35,000	1127 (47.2)
Prefer not to say	57 (2.3)
Deprivation quintile based on postcode	
1 (most deprived)	464 (19.4)
2	522 (21.8)
3	425 (17.8)
4	385 (16.1)
5 (least deprived)	408 (17.1)
Missing	186 (7.8)
Marital status	
Single	662 (27.7)
Married or in legal relationship	1368 (57.2)
Widowed, separated or divorced	348 (14.6)
Prefer not to say	12 (0.5)
Ethnicity	
White	1978 (82.8)
Asian or Asian British	127 (5.3)
Black, Black British, Caribbean or African	216 (9)
Mixed or multiple ethnic groups	51 (2.1)
Other	8 (0.3)

XSL•FO
RenderX

Characteristic	Results, n (%)
Prefer not to say	10 (0.4)
Preferred language	
English	2376 (99.4)
Other	12 (0.5)
Prefer not to say	2 (0.1)
Personal history of medical conditions	
Cancer	398 (16.7)
Heart disease or vascular problems	422 (17.7)
Depression	745 (31.2)
None of the above	1047 (43.8)
Not sure/not formally diagnosed	126 (5.3)
Prefer not to say	16 (0.7)
Relatives' history of medical conditions	
Cancer	656 (27.4)
Heart disease or vascular problems	633 (26.5)
Depression	535 (22.4)
None of the above	1005 (42.1)
Not sure/not formally diagnosed	159 (6.7)
Prefer not to say	24 (1)
Knowledge[d]	
Internet search history	2143 (89.7)
Internet browsing history	2141 (89.6)
Health app data	1717 (63.5)
Online banking details (incorrect)	1021 (42.7)
App store data	1490 (62.3)
Content of emails (incorrect)	1050 (43.9)
Purchases made online	1650 (69)
Location history	1831 (76.6)
Internet searches in "incognito" mode (incorrect)	813 (34)
Video streaming history	1720 (72)
Self-reported familiarity with internet use data	
Not familiar	792 (33.1)
Heard about internet use data	850 (35.6)
Some knowledge	512 (21.4)
Extensive knowledge	236 (9.9)

[a]GCSE: General Certificate of Secondary Education.

[b]AS: Advanced Subsidiary.

[c]NVQ: National Vocational Qualification.

[d]Number and percentage of participants indicating this is part of internet use data.

Willingness to Share Internet Use Data

There were no differences in proportions of respondents willing to share internet use data between the 3 medical conditions (for early detection of cancer=77%; heart disease=74%; depression=74%). When asked about different types of data, participants were most willing to share their health app data (73% - 76%, 95% CI 69.8% - 79.1%), whereas participants were less willing to share other types of data, especially regarding purchases made online (52% - 57%, 95% CI

XSL•FO
RenderX

48.7% - 60.3%) and their location history (58% - 59%, 95% CI 54.2% - 62.2%). Details are provided in Table 2. There was variability in what participants understood to be a part of internet use data, with more than 40% of participants incorrectly believing that the content of their emails (1050/2390, 43.9%) and online banking history (1021/2390, 42.7%) would be included (see Table 1).

Table . Willingness to share internet use data for research on different medical conditions from a cross-sectional survey in 2025 with UK participants recruited through an online survey panel.

Type	Willingness to share internet use data for research on early detection of…[a]									Comparison, χ^2 (df)	*P* value
	Cancer			Heart disease			Depression				
	Total users, n	Users willing to share, n (%)	95% CI	Total users, n	Users willing to share, n (%)	95% CI	Total users, n	Users willing to share, n (%)	95% CI		
Internet use data in general	795	615 (77.4)	74.4 - 80.3	798	587 (73.6)	70.5 - 76.6	797	592 (74.3)	71.2 - 77.3	4.64 (2,2390)	.18
Internet search history	779	532 (68.3)	65.0 - 72.6	777	509 (65.5)	62.2 - 68.9	787	541 (68.7)	65.5 - 72.0	2.18 (2,2343)	.34
Internet browsing history	784	519 (66.2)	62.9 - 69.5	783	493 (63.0)	59.6 - 66.4	792	535 (67.6)	64.3 - 70.8	3.87 (2,2359)	.14
Health app data	682	498 (73.0)	69.7 - 76.4	670	508 (75.8)	72.6 - 79.1	688	503 (73.1)	69.8 - 76.4	1.78 (2,2040)	.41
App store data	730	446 (61.1)	57.6 - 64.6	732	455 (62.2)	58.6 - 65.7	748	490 (65.5)	62.1 - 68.9	3.37 (2,2210)	.19
Purchases made online	770	402 (52.2)	48.7 - 55.7	777	414 (53.3)	49.8 - 56.8	788	448 (56.9)	53.4 - 60.3	3.72 (2,2335)	.16
Location history	771	453 (58.8)	55.3 - 62.2	768	443 (57.7)	54.2 - 61.2	774	447 (57.8)	54.3 - 61.2	0.23 (2,2313)	.89
Video streaming history	750	463 (61.7)	58.3 - 65.2	751	471 (62.7)	59.3 - 66.2	756	499 (66.0)	62.6 - 69.4	3.26 (2,2257)	.20

[a]Total numbers and percentages for willingness to share include valid participants who are using these services and answered, “definitely yes” and “probably yes.”

Participants were descriptively less willing to share their internet use data if a pictorial example of internet use history was provided alongside written text compared with written text only, but this did not reach statistical significance for research on any of the 3 medical conditions (see Table 3).

Having a personal or family history of the condition being researched was not associated with willingness to share internet use history for that condition, with the exception of those with a personal history of depression. Participants who had a history of depression were more likely to be willing to share their internet use data for the early detection of depression (79.7%, 95% CI 74.5% - 84.8%) than participants without a history of depression (72.2%, 95% CI 67.8% - 76.5%; *P*=.03; Table 3).

XSL•FO
RenderX

Table . Willingness to share internet use data in relation to how the internet search history was described and personal and family history of the medical condition being researched from a cross-sectional survey in 2025 with UK participants recruited through an online survey panel.

Characteristic	Willingness to share internet use data for research on early detection of…[a]											
	Cancer				Heart disease				Depression			
	N (%)	95% CI	Comparison, χ^2 (df)	*P* value	N (%)	95% CI	Comparison, χ^2 (df)	*P* value	N (%)	95% CI	Comparison, χ^2 (df)	*P* value
Description of internet use			3.56 (1,795)	.06			2.05 (1,798)	.15			2.02 (1,797)	.16
Written description	312 (80.2)	76.2 - 84.2			298 (75.8)	71.6 - 80.1			294 (76.6)	72.3 - 80.8		
Pictorial example and written description	303 (74.6)	76.2 - 84.2			289 (71.4)	66.9 - 75.8			298 (72.2)	67.8 - 76.5		
Personal history of medical condition being researched			3.60 (1,790)	.058			0.36 (1,790)	.55			4.82 (1,794)	.03
Yes	116 (83.5)	77.2 - 89.7			103 (75.7)	68.4 - 83.0			188 (79.7)	74.5 - 84.8		
No	495 (76.0)	72.8 - 79.3			479 (73.2)	69.8 - 76.6			403 (72.2)	68.5 - 76.0		
Family history of medical condition being researched			3.17 (1,792)	.08			0.95 (1,789)	.33			3.17 (1,792)	.08
Yes	171 (81.8)	76.6 - 87.1			159 (76.4)	70.6 - 82.3			133 (76.9)	70.5 - 83.2		
No	442 (75.8)	72.3 - 79.3			424 (73.0)	69.4 - 76.6			449 (73.4)	69.9 - 76.9		

[a]Total numbers and percentages for willingness to share include participants answering, “definitely yes” and “probably yes.”

Key Factors Associated With Willingness to Share Internet Use Data

Multivariable logistic regression analysis on willingness to share internet use data for early detection of cancer indicated that willingness to share was associated with perceived benefits and perceived concerns of data sharing. Higher endorsement of benefits of sharing data was associated with a greater willingness to share (OR 8.588, 95% CI 5.687 - 12.970). A higher endorsement of concerns was associated with a lower willingness to share (OR 0.432, 95% CI 0.290 - 0.643). Ethnicity was significantly associated with willingness to share, with individuals of Asian or Asian British ethnicity demonstrating a significantly lower willingness to share than White participants (OR 0.234, 95% CI 0.076 - 0.723). No other demographic factors, including age, gender, employment, income, marital status, education level, personal or family history of cancer, or level of deprivation, were associated with a willingness to share in the multivariable regression analysis, although some (education, employment, ethnicity and income) had indicated associations in univariate analyses (see Multimedia Appendix 1). Factors related to digital literacy (knowledge about internet use data, the number of online services used, or self-reported familiarity with internet use data) were not significantly related to willingness to share internet use data for cancer research in the multivariable regression analysis.

Willingness to share internet use data for the early detection of heart disease and depression followed similar patterns, with perceived benefits being significantly associated with a higher willingness to share internet use data (heart disease: OR 5.692, 95% CI: 4.050 - 7.998; depression: OR 8.850, 95% CI 5.998 - 13.059) and perceived concerns being significantly associated with a lower willingness to share internet use data (heart disease: OR 0.424, 95% CI 0.292 - 0.617; depression: OR 0.343, 95% CI 0.228 - 0.516). For research on detection of heart disease, self-reported familiarity with internet use data was associated with a lower willingness to share internet use data (OR 0.740, 95% CI 0.561 - 0.976). For research on the early detection of depression, older participants were less willing to share their data (OR 0.975, 95% CI 0.951 - 0.999), and men were more likely to be willing to share than women (OR 2.615, 95% CI 1.511 - 4.526).

All results are displayed in Table 4.

Sample sizes for the logistic regression analyses were lower than intended due to missing sociodemographic data, especially for levels of deprivation. We therefore conducted sensitivity analyses by repeating the logistic regression analyses without levels of deprivation. Results were similar to the initial logistic regression analysis (see Multimedia Appendix 1), except for depression, where age was no longer associated with willingness to share their data (OR 0.981, 95% CI 0.959 - 1.003) and those with a higher knowledge score about internet use data were less willing to share internet use data (OR 0.842, 95% CI 0.723 - 0.981). For early detection of cancer, Asian and Asian

XSL•FO
RenderX

British were not significantly less likely to share their internet use data (OR 0.351, 95% CI 0.122 - 1.011).

XSL•FO
RenderX

Table . Logistic regression on willingness to share internet use data for research on the early detection of cancer from a cross-sectional survey in 2025 with UK participants recruited through an online survey panel.

Variable	Willingness to share internet use data for research on early detection of…											
	Cancer (n=670)[a]				Heart disease (n=646)[b]				Depression (n=673)[c]			
	Willing, n (%)	OR[d] (95% CI)	SE	*P* value	Willing, n (%)	OR (95% CI)	SE	*P* value	Willing, n (%)	OR (95% CI)	SE	*P* value
Description of internet use												
Written description	269 (80.5)	1	—[e]	—	246 (78.8)	1	—	—	244 (76.0)	1	—	—
Pictorial example and written description	257 (76.5)	0.617 (0.357-1.066)	0.279	.08	243 (72.8)	0.735 (0.450-1.198)	0.250	.22	263 (74.7)	1.200 (0.723-1.989)	0.258	.48
Gender												
Women	240 (80.5)	1	—	—	228 (74.8)	1	—	—	224 (70.2)	1	—	—
Men	286 (76.9)	0.980 (0.560-1.716)	0.286	.94	261 (76.5)	1.017 (0.607-1.704)	0.263	.95	283 (79.9)	2.615 (1.511-4.526)	0.280	<.001
Education												
Degree level or above	296 (80.9)	1	—	—	263 (79.7)	1	—	—	294 (80.1)	1	—	—
Basic education	98 (72.1)	0.590 (0.300-1.159)	0.344	.13	101 (71.6)	0.688 (0.343-1.380)	0.36	.29	90 (67.7)	0.614 (0.291-1.296)	0.381	.20
Advanced education	132 (78.6)	0.754 (0.390-1.455)	0.336	.40	125 (71.4)	0.627 (0.347-1.130)	0.301	.12	123 (71.1)	0.689 (0.362-1.313)	0.329	.26
Employment												
Employed	324 (83.7)	1	—	—	297 (76.3)	1	—	—	329 (78.5)	1	—	—
Unemployed	69 (67.6)	0.730 (0.353-1.509)	0.371	.40	60 (74.1)	1.172 (0.518-2.656)	0.417	.70	66 (75.0)	1.303 (0.575-2.950)	0.417	.53
Retired	133 (73.5)	1.194 (0.521-2.732)	0.422	.68	132 (75.0)	1.251 (0.583-2.687)	0.390	.57	112 (67.5)	1.770 (0.826-3.792)	0.389	.14
Income (£)												
<35,000	251 (74.5)	1	—	—	254 (73.6)	1	—	—	239 (72.2)	1	—	—
>35,000	275 (82.6)	1.327 (0.703-2.503)	0.324	.38	235 (78.1)	0.826 (0.459-1.487)	0.300	.52	268 (78.4)	0.775 (0.406-1.479)	0.330	.44
Marital status												

XSL•FO
RenderX

Variable	Willingness to share internet use data for research on early detection of…											
	Cancer (n=670)[a]				Heart disease (n=646)[b]				Depression (n=673)[c]			
	Willing, n (%)	OR[d] (95% CI)	SE	*P* value	Willing, n (%)	OR (95% CI)	SE	*P* value	Willing, n (%)	OR (95% CI)	SE	*P* value
Married or in legal partnership	310 (80.9)	1	—	—	299 (79.7)	1	—	—	314 (79.1)	1	—	—
Single	131 (74.4)	1.250 (0.625-2.499)	0.353	.53	123 (70.7)	0.669 (0.364-1.229)	0.310	.20	128 (69.6)	1.020 (0.537-1.936)	0.327	.95
Divorced/widowed/separated	85 (76.6)	1.057 (0.511-2.186)	0.371	.88	67 (69.1)	0.754 (0.362-1.571)	0.374	.45	65 (70.7)	1.294 (0.575-2.915)	0.414	.53
Ethnicity												
White	445 (78.1)	1	—	—	416 (75.0)	1	—	—	408 (73.8)	1	—	—
Asian or Asian British	21 (65.6)	0.234 (0.076-0.723)	0.576	.01	22 (81.5)	0.843 (0.231-3.084)	0.661	.80	32 (80.0)	1.277 (0.394-4.133)	0.599	.68
Black or Black British	48 (92.3)	0.806 (0.189-3.431)	0.739	.77	43 (82.7)	0.478 (0.177-1.291)	0.507	.15	58 (89.2)	0.682 (0.225-2.065)	0.565	.50
Mixed or multiple ethnic groups	12 (75.0)	0.251 (0.062-1.019)	0.714	.053	8 (66.7)	0.474 (0.098-2.283)	0.802	.35	9 (60.0)	0.273 (0.063-1.177)	0.745	.08
Level of deprivation												
IMD[f] Q1 (most deprived)	107 (80.5)	1	—	—	107 (78.7)	1	—	—	120 (78.4)	1	—	—
IMD Q2	130 (77.4)	1.664 (0.766-3.613)	0.396	.20	113 (75.8)	1.116 (0.541-2.302)	0.370	.77	114 (79.7)	2.043 (0.910-4.583)	0.412	.08
IMD Q3	93 (76.9)	0.965 (0.417-2.234)	0.428	.93	99 (72.3)	1.153 (0.531-2.503)	0.395	.72	93 (73.8)	1.450 (0.638-3.295)	0.419	.38
IMD Q4	94 (77.0)	1.162 (0.495-2.729)	0.436	.73	75 (75.0)	1.868 (0.809-4.315)	0.427	.14	83 (69.7)	1.896 (0.811-4.432)	0.433	.14
IMD Q5 (least deprived)	102 (81.0)	1.943 (0.793-4.763)	0.457	.15	95 (76.6)	1.155 (0.512-2.603)	0.415	.73	97 (73.5)	1.142 (0.496-2.628)	0.425	.76
Personal history of condition of interest												

XSL•FO
RenderX

Variable	Willingness to share internet use data for research on early detection of…											
	Cancer (n=670)[a]				Heart disease (n=646)[b]				Depression (n=673)[c]			
	Willing, n (%)	OR[d] (95% CI)	SE	*P* value	Willing, n (%)	OR (95% CI)	SE	*P* value	Willing, n (%)	OR (95% CI)	SE	*P* value
No	422 (77.3)	1	—	—	395 (75.1)	1	—	—	349 (73.0)	1	—	—
Yes	104 (83.9)	1.526 (0.720-3.236)	0.383	.27	94 (78.3)	0.892 (0.444-1.794)	0.357	.75	158 (81.0)	1.345 (0.720-2.513)	0.319	.35
Family history of condition of interest												
No	370 (76.8)	1	—	—	352 (73.9)	1	—	—	390 (74.4)	1	—	—
Yes	156 (83.0)	1.189 (0.651-2.171)	0.307	.57	137 (80.6)	1.037 (0.564-1.907)	0.311	.91	117 (78.5)	1.458 (0.758-2.804)	0.334	.26
Age	—	0.983 (0.958-1.009)	0.013	.21	—	0.984 (0.962-1.008)	0.012	.18	—	0.975 (0.951-0.999)	0.012	.04
Perceived benefits about sharing internet use data	—	8.588 (5.687-12.970)	0.210	<.001	—	5.692 (4.050-7.998)	0.174	<.001	—	8.850 (5.998-13.059)	0.198	<.001
Perceived concerns about sharing internet use data	—	0.432 (0.290-0.643)	0.203	<.001	—	0.424 (0.292-0.617)	0.191	<.001	—	0.343 (0.228-0.516)	0.208	<.001
Knowledge about internet use data (score)	—	1.061 (0.906-1.242)	0.080	.46	—	1.011 (0.866-1.180)	0.079	.891	—	0.858 (0.727-1.012)	0.084	.07
Number of online services used	—	0.923 (0.826-1.031)	0.057	.16	—	1.040 (0.942-1.148)	0.051	.440	—	1.033 (0.929-1.149)	0.054	.55

XSL•FO
RenderX

Variable	Willingness to share internet use data for research on early detection of…											
	Cancer (n=670)[a]				Heart disease (n=646)[b]				Depression (n=673)[c]			
	Willing, n (%)	OR[d] (95% CI)	SE	*P* value	Willing, n (%)	OR (95% CI)	SE	*P* value	Willing, n (%)	OR (95% CI)	SE	*P* value
Familiarity with internet use data	—	1.108 (0.814-1.509)	0.158	.52	—	0.740 (0.561-0.976)	0.141	.033	—	0.854 (0.625-1.167)	0.160	.32

[a]Nagelkerke R=0.573.

[b]Nagelkerke R=0.516

[c]Nagelkerke R=0.588.

[d]OR: odds ratio.

[e]Not applicable.

[f]IMD: index of multiple deprivation.

Statements on Willingness to Share and Concerns

The open-ended question on what would make them more willing to share their internet use data was answered by 2307 (2307/2390, 96.5%) participants. Percentages of statements relative to numbers of those definitely willing, probably willing, probably unwilling, and definitely unwilling to share are presented in Figure 1. The main suggestions given by those currently not willing to share were strong assurance of data security (66/596, 11.1%) and incentives (37/596, 6.2%); however, the majority of these participants stated “nothing” would make them more willing (definitely not: 189/236, 80.1%; probably not: 196/360, 54.4%). Those who were “probably” or “definitely” willing to share their internet use data offered a wider range of suggestions, including a sense of contribution to research and society (“helping others”: 480/1794, 26.8%); data security assurances (211/1794, 11.8%); clarification of research purposes (157/1794, 8.6%); incentives (110/1794, 6.1%); strict governance, personal control, and building trust (106/1794, 5.9%); and personal relevance (147/1794, 8.2%).

XSL•FO
RenderX

Figure 1. Suggestions to increase willingness to share internet use data from a cross-sectional survey in 2025 with UK participants recruited through an online survey panel. Statements could be coded as multiple categories. Participants saying nothing would make them more willing are not included in this figure.

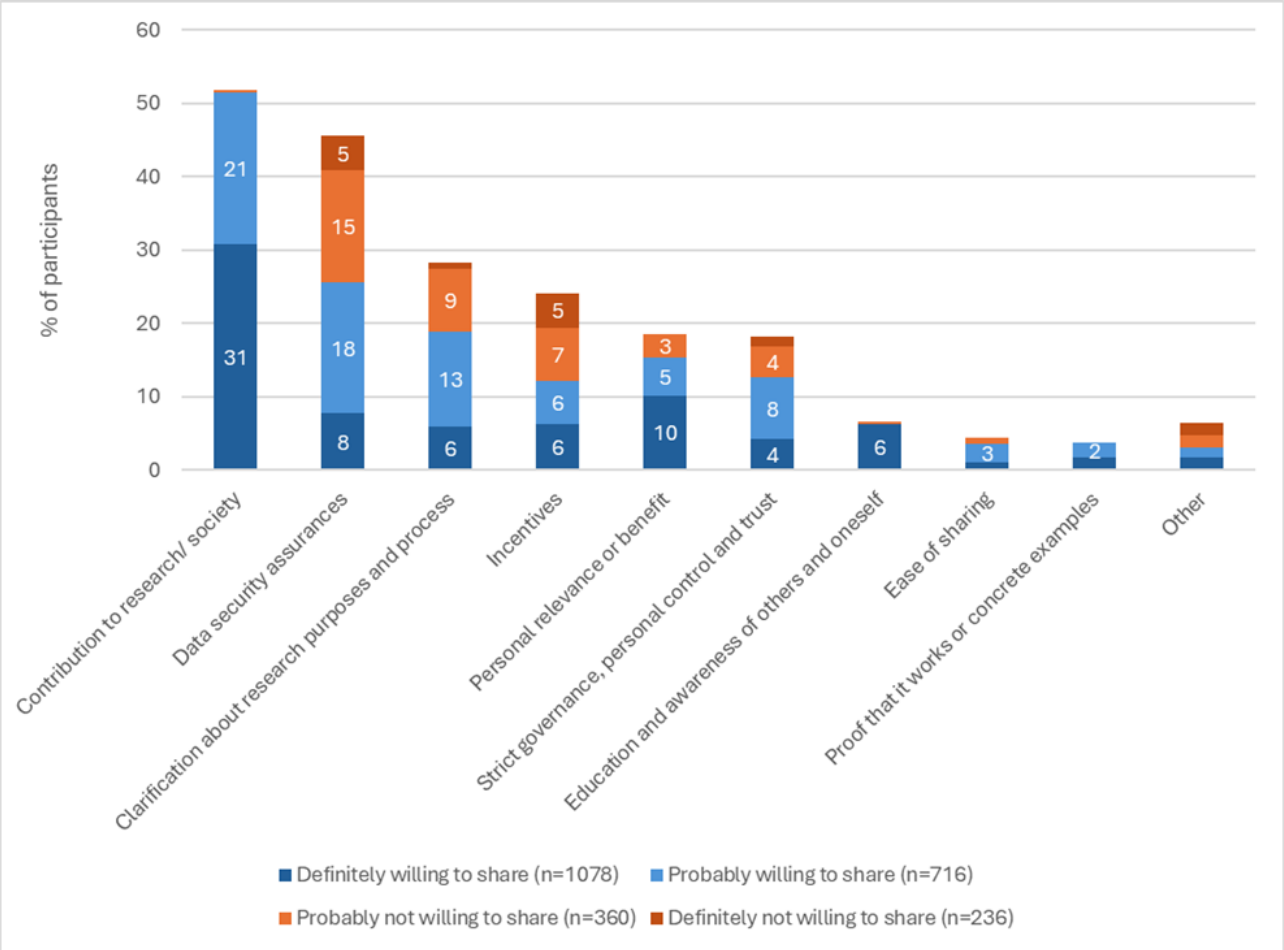

The question about concerns surrounding sharing internet use data was answered by 2318 (2318/2390, 97%) participants (Figure 2). Concerns of participants who were "probably not" or "definitely not" willing to share their internet use data mostly focused on data security and the invasion of privacy (160/596, 26.8%) and concerns about data misuse, including sharing with private (marketing) companies (39/596, 6.5%). Participants also questioned the relevance of their internet use data for early detection of disease (76/596, 12.8%). Participants who were "probably" or "definitely" willing to share their internet use data noted similar concerns, although many indicated no concerns (probably willing: 368/716, 51.4%; definitely willing: 794/1078, 73.7%). Examples are provided in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 2. Concerns about sharing internet use data from a cross-sectional survey in 2025 with UK participants recruited through an online survey panel. Statements could be coded as multiple categories. Participants saying they had no concerns are not included in this figure.

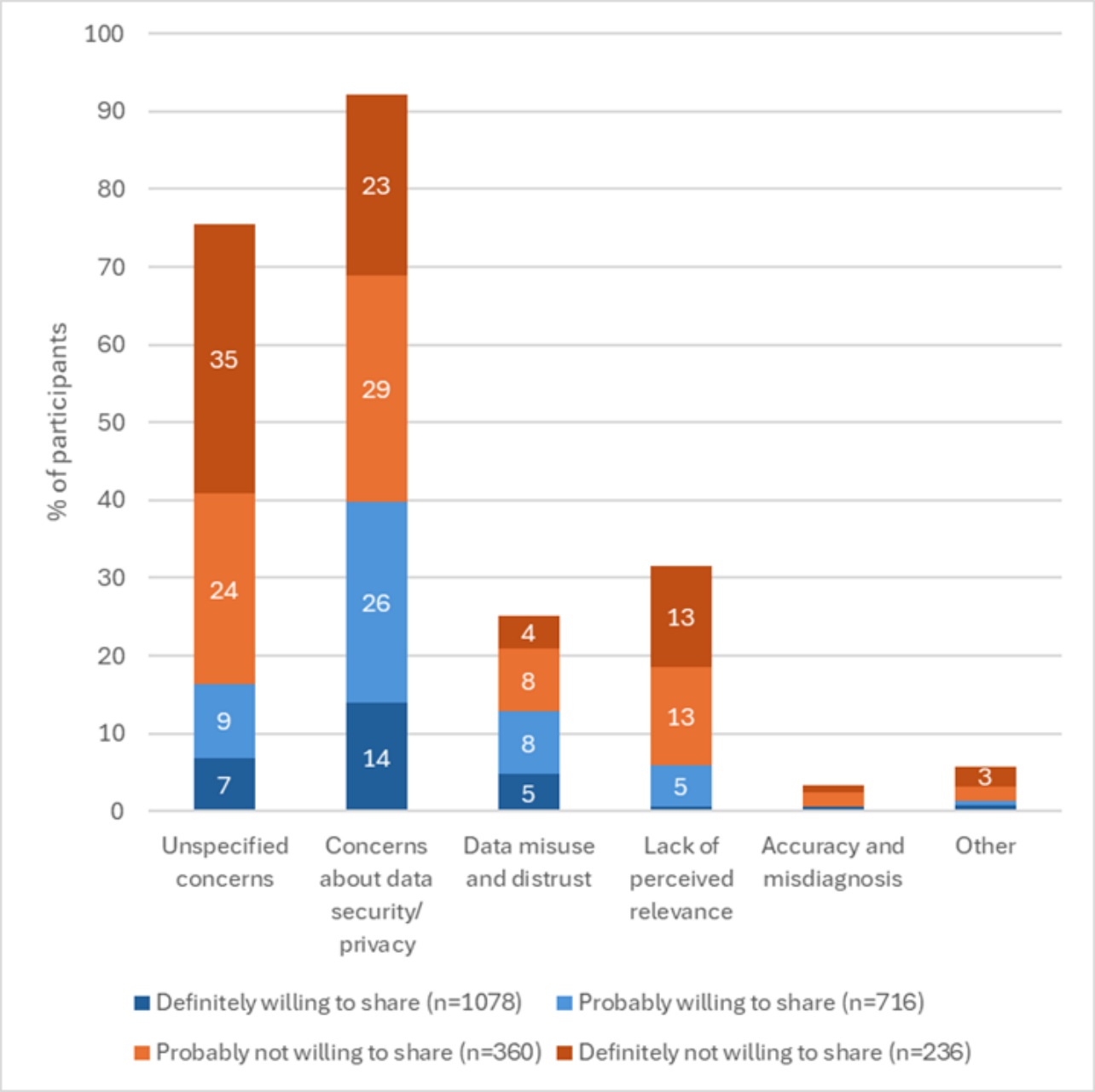

Discussion

Principal Findings

This study examined willingness to share internet use data for research into early detection of cancer, heart disease, and depression. Participants showed a similarly high willingness to share internet use data for all conditions. Willingness differed by data type, being lowest for online purchases and location history and highest for health app data. Viewing a screenshot of browsing history did not significantly affect willingness for any condition.

Perceived benefits and concerns were key factors associated with willingness across all conditions. Answering open questions about willingness to share, participants reported that contributing to research, personal relevance, data security, incentives, clear research purposes, concrete examples, strong governance, and trust in researchers would improve their willingness to share. In contrast, they reported that privacy concerns, fear of misuse, and low relevance could reduce willingness to share. Those who were more familiar with internet use data were less likely to share data for heart disease detection, but this was not associated with willingness to share data for research on cancer or depression. There were few sociodemographic factors associated with willingness to share, but some did exist: Asian and Asian British participants were less willing to share their internet use data for cancer detection (not replicated in the sensitivity analyses), while younger (not replicated in the sensitivity analyses) and male participants were more willing to share data for depression detection.

XSL•FO
RenderX

Comparison With Prior Work

We found a surprisingly high willingness to share internet use data for health research compared with recent studies. For example, Grande et al [20] classified approximately 13% of their participants as "agreeable to sharing," whereas about 45% of our participants indicated they were "definitely" willing to share. Similarly, Hirst et al [19] found that 61% of participants were willing to share their data with research organizations, and only 43% were willing to share with private organizations, compared with 75% in this study. Although we did not differentiate the kind of organization data would be shared with, we consistently found that a higher proportion of participants were willing to share across different types of data than in these studies. Our proportions are based on users of the respective types of data. However, willingness to share was similar when participants were asked more generally about "internet use data."

As we used a similar recruitment strategy as previous studies [19], it seems that the *purpose* of the data sharing described in our study may be responsible for the higher willingness to share internet use data. In our survey, we asked participants specifically about sharing their internet use data for the early detection of health conditions. Previous research has found public support for screening and early detection [35,36], especially if personally relevant [37]. This is line with open-ended comments regarding increased willingness to share if there were personal benefits (eg, receiving a quicker diagnosis). Beyond personal benefits, participants in our study who were willing to share their internet use data described a general sense of contribution to research and society, which our PPI group summarized as an act of "charity" in terms of "donating data." The Smart Data Research UK report [12] reports how the public grew more enthusiastic about using smart data in research for the "public good" when learning and discussing potential applications. However, information about the value of data sharing, as well as ensuring personal control over sharing potentially sensitive information is crucial to ensure acceptability of sharing data [38], which is reflected in our findings.

Another suggestion for improving willingness to share internet use data in our study was monetary incentives, which was not captured in our quantitative scales but reported in the free text. Even participants who were currently unwilling to share mentioned incentives as potentially increasing their willingness to share. This indicates a "trade-off" between privacy and payment, although the exact value of individual privacy is difficult to determine [39].

Our findings regarding higher willingness to share health app data than to share other types of internet use data are in line with previous research [20]. For example, Ackermann et al [40] found that a better "intuitive match" between the type of data requested and what the data are used for increases willingness to share internet use data. In this survey study, this would be the case for health app data but not necessarily for online purchases or location data, which participants were less willing to share. However, researchers developing risk prediction models based on or including internet use data might rely on changes in online behavior beyond concrete health app data (eg, if an individual repeatedly searches for nonspecific symptoms or home remedies). As individuals are likely to look up symptoms online or change the tone and semantics of their searches long before visiting their general practitioner, capturing these early changes might benefit early detection [13-16].

Types of data that are not directly related to medical contexts might also be seen as more personal and more likely to be misused, such as by marketing companies [19]. Concerns about privacy and misuse of data were associated with a lower willingness to share across all conditions in this study. They were commonly endorsed and reported in both the quantitative and qualitative part of the survey, even by participants who were "definitely" willing to share. The Smart Data Research UK report [12] also shows that the public is skeptical about private sector motivation and concerned about potential inequalities and data security.

Implications for Future Research and Practice

There is a clear need for transparent and open communication around the purpose of using internet data for early detection research, as well as data processing after it is shared, including which data, how, and why they are collected and stored [12,38]. Researchers and data scientists need to work toward clarifying which data are needed for what purpose, how to limit data sharing to what is necessary and filter out any other information, and how to communicate purpose and process of data sharing, especially concerning minority groups [41,42]. We found no effect of showing participants a pictorial example of internet use data on willingness to share. Our PPI group noted that the example only included a generic internet history without particularly sensitive data (eg, regarding financial details or a child's local school). Future research could examine the effect of different types of examples, preferably using participants' own internet use data.

Only 10% of participants in our sample noted that they knew a lot about internet use data, and 34% to 44% mistakenly believed that internet searches in "incognito" mode, online banking details, and content of emails are part of internet use data. This indicates that internet users might not be well-informed about the personal data they might be sharing online [43,44]. There is a need for more public information and education around what internet use data entail and what would need to be accessed for medical research to facilitate informed consent [45]. Our findings suggest that higher familiarity and, potentially, knowledge might negatively impact willingness to share internet use data for medical research. However, this should be interpreted with caution, as univariate analyses indicated the opposite trend. Actual familiarity and knowledge about internet use data and the consequences for willingness to share data should be investigated in more detail. Clear communication including which data are shared with whom and ensuring personal control are central to enable informed participation in internet research [45].

Our findings show how concerns around sharing internet use data might be more pronounced in some ethnic minority groups and corroborate research showing that cancer is associated with fear and stigma in minority communities [46]. Future research should focus on mitigation strategies to avoid inequalities in

XSL•FO
RenderX

this new stream of internet-based medical research, especially for vulnerable populations. Stigma around mental health conditions might drive the findings that younger age and male gender were associated with a higher willingness to share internet use data for the detection of depression [47]. As traditional norms of masculinity can reduce symptom expression and inhibit help-seeking for mental health conditions [48,49], sharing internet use data might be a more anonymous way to disclose potential symptoms.

Although many participants reported having concerns about sharing internet use data, a high proportion did not specify their concerns. This might have been enforced by the wording of the open-ended question around concerns but could also reflect a general unease with data sharing due to a lack of knowledge and transparency [50]. A qualitative study identified similar concerns around data security and misuse [51] as well as actions taken by individuals to protect their privacy, such as use of incognito mode for sensitive searches, that could impact future research. Future research, especially qualitative research, should examine how these concerns can be mitigated in more detail. Additionally, those who reported being "definitely not" willing to share data often said "nothing" would make them more willing. Future research should therefore investigate why some people would not be willing to share internet use data under any circumstances.

Strengths and Limitations

This study is a large survey with population-based quotas that builds on a previous qualitative study [51], previously used measures [18,19,24], and extensive patient and public input. However, the survey assessed self-reported willingness to share rather than the actual behavior, as well as self-reported diagnosis, which might have led to an overestimation of those with a confirmed clinical diagnosis, especially in those reporting to have depression [52]. It remains to be seen whether these results will be confirmed in future research and practice when shared internet use data are required.

We recruited through an online panel and therefore included participants who are comfortable answering online research surveys for incentives and attracted those with higher education (approximately 52% with at least a degree level of education, compared with the UK average of 33% - 34%, [53]). This might have contributed to the high willingness to share internet use data and frequent mentioning of incentives to increase willingness in the open-ended questions. Nevertheless, internet use data can only be shared by people who are using the internet in the first place, and internet use and knowledge about internet use data varied in the sample, suggesting the sample represents a range of internet users. The power for some of the logistic regression analyses might have been too low to detect differences. Although we ensured a representative percentage of the sample (approximately 17%) was from ethnic minority groups, the numbers within minority ethnic groups limit accurate comparisons.

Other potential methodological limitations due to the short survey design include the use of area-based index of multiple deprivation quintiles for individual-level data, potentially leading to ecological fallacy as individuals living in affluent areas may still experience deprivation and vice versa [54]. Nevertheless, it is still the most comprehensible method that could be applied in a short questionnaire, and sensitivity analyses were used to replicate logistic regression analyses without this deprivation variable [55]. We combined those "definitely" and "probably" (not) willing to share for analyses, although the groups might differ. Dichotomizing questions about willingness to share internet use data might have reduced the nuance of the data.

Finally, we had to exclude a substantial proportion (16.6%) of responses due to low data quality, including those flagged as bots and fraudulent responses and those failing attention checks. This does not negatively impact the quality of our final sample, as we were able to exclude and replace these participants during recruitment. However, this shows that researchers will need to account for "impostor" responses when planning for sample sizes for online surveys, implement strategies to detect low quality data (eg, attention checks or bot detection), and carefully check responses, leading to additional workload. If low-quality responses are not addressed, we may not be able to trust similar research results [56].

Conclusion

This study is innovative as the first large-scale survey, to the best of our knowledge, to examine public willingness to share internet use data for early detection research across multiple common health conditions while assessing perceived benefits, risks, and sociodemographic differences. This survey indicated willingness to share internet use data is high across research on early detection of different common health conditions. Willingness to share was associated with perceived benefits for oneself and society in general and limited by concerns around data privacy and potential for misuse. Participants were more willing to share data that were plausibly related to the purpose of early detection of health conditions (eg, health app data).

The survey results advance the understanding of participation in research using internet use data by identifying both barriers and enablers of data sharing. Although sociodemographic differences were limited, variations by age, gender, and ethnicity suggest potential inequities. The findings inform consent design, communication strategies, and data governance, highlighting the need for clear purpose limitation and robust data protection to foster trust and reduce inequalities. They emphasize the need to clarify which and how data are used for research into early detection and investigate how communication about internet use data could improve willingness to share while avoiding perpetuating inequalities. Data sharing needs to be limited to specific research purposes, and resolute data protection strategies need to be put in place.

XSL•FO
RenderX

Acknowledgments

We would like to thank our patient and public involvement (PPI) group for their valuable inputs during this project and for reviewing and testing the survey. We would also like to thank our participants for their time and interest in this study.

The Data Liberation Front (based at Google) was a collaborator on the study and advised on accurate description about internet use data. The research team members (CD, MR, SES) were the sole decision-makers on research design and outputs.

Funding

Google provided unrestricted funding for this research through a Privacy Research Faculty Award to SES. The funder had no role in the study design; collection, analysis, or interpretation of data; writing of the paper; and/or decision to submit for publication. SES is supported by Barts Charity (G-001520; MRC&U0036).

Data Availability

The full syntax and the final datasets can be found on OSF [56].

Authors' Contributions

SES acquired funding. SES, CD, MR and SA contributed to conceptualization. SES and CD were responsible for project administration. CD undertook the data collection and analysis, with input from SES, MR and WL. CD drafted the original manuscript. All authors contributed to writing – review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
WISER survey, additional regression analyses, and definitions of categories for suggestions to increase willingness to share and for concerns.
[DOCX File, 139 KB - jmir_v28i1e85637_app1.docx]

References

1. Oude Wolcherink MJ, Behr CM, Pouwels X, Doggen CJM, Koffijberg H. Health economic research assessing the value of early eetection of cardiovascular disease: a systematic review. Pharmacoeconomics 2023 Oct;41(10):1183-1203. [doi: 10.1007/s40273-023-01287-2] [Medline: 37328633]
2. Vogel B, Acevedo M, Appelman Y, et al. The Lancet Women and Cardiovascular Disease Commission: reducing the global burden by 2030. Lancet 2021 Jun 19;397(10292):2385-2438. [doi: 10.1016/S0140-6736(21)00684-X] [Medline: 34010613]
3. Crosby D, Bhatia S, Brindle KM, et al. Early detection of cancer. Science 2022 Mar 18;375(6586):eaay9040. [doi: 10.1126/science.aay9040] [Medline: 35298272]
4. Whitaker KL, Scott SE, Wardle J. Applying symptom appraisal models to understand sociodemographic differences in responses to possible cancer symptoms: a research agenda. Br J Cancer 2015 Mar 31;112 Suppl 1(Suppl 1):S27-S34. [doi: 10.1038/bjc.2015.39] [Medline: 25734385]
5. Pedersen SD, Nielsen BD, Assmann ML, Hauge EM, de Thurah A. Early diagnosis of rheumatoid arthritis: associations between patients' perceptions of initial symptoms and the timing of seeking help from the general practitioner. Scand J Rheumatol 2025 Jul;54(4):242-251. [doi: 10.1080/03009742.2025.2464457] [Medline: 40013732]
6. Asch JM, Asch DA, Klinger EV, Marks J, Sadek N, Merchant RM. Google search histories of patients presenting to an emergency department: an observational study. BMJ Open 2019 Feb 20;9(2):e024791. [doi: 10.1136/bmjopen-2018-024791] [Medline: 30787088]
7. Padrez KA, Ungar L, Schwartz HA, et al. Linking social media and medical record data: a study of adults presenting to an academic, urban emergency department. BMJ Qual Saf 2016 Jun;25(6):414-423. [doi: 10.1136/bmjqs-2015-004489] [Medline: 26464519]
8. D'Alfonso S, Coghlan S, Schmidt S, Mangelsdorf S. Ethical dimensions of digital phenotyping within the context of mental healthcare. J technol behav sci 2024 Jun 25;10(1):132-147. [doi: 10.1007/s41347-024-00423-9]
9. Koo MM, Hamilton W, Walter FM, Rubin GP, Lyratzopoulos G. Symptom signatures and diagnostic timeliness in cancer patients: a review of current evidence. Neoplasia 2018 Feb;20(2):165-174. [doi: 10.1016/j.neo.2017.11.005] [Medline: 29253839]
10. Ben-Zeev D, Young MA. Accuracy of hospitalized depressed patients' and healthy controls' retrospective symptom reports: an experience sampling study. J Nerv Ment Dis 2010 Apr;198(4):280-285. [doi: 10.1097/NMD.0b013e3181d6141f] [Medline: 20386257]

XSL•FO
RenderX

11. Lackner JM, Jaccard J, Keefer L, et al. The accuracy of patient-reported measures for GI symptoms: a comparison of real time and retrospective reports. Neurogastroenterol Motil 2014 Dec;26(12):1802-1811. [doi: 10.1111/nmo.12466] [Medline: 25424582]
12. Smart Data Research UK. Smart Data Research UK public dialogue report 2025. 2025 URL: https://www.sdruk.ukri.org/our-work/publications/public-dialogue-report/sdr-uk-public-dialogue-report-2025/ [accessed 2026-03-21]
13. Areán PA, Pratap A, Hsin H, et al. Perceived utility and characterization of personal Google search histories to detect data patterns proximal to a suicide attempt in individuals who previously attempted suicide: pilot cohort study. J Med Internet Res 2021 May 6;23(5):e27918. [doi: 10.2196/27918] [Medline: 33955838]
14. Barcroft JF, Yom-Tov E, Lampos V, et al. Using online search activity for earlier detection of gynaecological malignancy. BMC Public Health 2024 Mar 11;24(1):608. [doi: 10.1186/s12889-024-17673-0] [Medline: 38462622]
15. Paparrizos J, White RW, Horvitz E. Screening for pancreatic adenocarcinoma using signals from web search logs: feasibility study and results. J Oncol Pract 2016 Aug;12(8):737-744. [doi: 10.1200/JOP.2015.010504] [Medline: 27271506]
16. Birnbaum ML, Wen H, Van Meter A, et al. Identifying emerging mental illness utilizing search engine activity: a feasibility study. PLOS ONE 2020;15(10):e0240820. [doi: 10.1371/journal.pone.0240820] [Medline: 33064759]
17. Brall C, Berlin C, Zwahlen M, Ormond KE, Egger M, Vayena E. Public willingness to participate in personalized health research and biobanking: a large-scale Swiss survey. PLoS ONE 2021;16(4):e0249141. [doi: 10.1371/journal.pone.0249141] [Medline: 33793624]
18. Seltzer E, Goldshear J, Guntuku SC, et al. Patients' willingness to share digital health and non-health data for research: a cross-sectional study. BMC Med Inform Decis Mak 2019 Aug 8;19(1):157. [doi: 10.1186/s12911-019-0886-9] [Medline: 31395102]
19. Hirst Y, Stoffel ST, Brewer HR, Timotijevic L, Raats MM, Flanagan JM. Understanding public attitudes and willingness to share commercial data for health research: survey study in the United Kingdom. JMIR Public Health Surveill 2023 Mar 23;9:e40814. [doi: 10.2196/40814] [Medline: 36951929]
20. Grande D, Mitra N, Iyengar R, et al. Consumer willingness to share personal digital information for health-related uses. JAMA Netw Open 2022 Jan 4;5(1):e2144787. [doi: 10.1001/jamanetworkopen.2021.44787] [Medline: 35072717]
21. Shandhi MMH, Singh K, Janson N, et al. Assessment of ownership of smart devices and the acceptability of digital health data sharing. NPJ Digit Med 2024 Feb 22;7(1):44. [doi: 10.1038/s41746-024-01030-x] [Medline: 38388660]
22. Trauth JM, Musa D, Siminoff L, Jewell IK, Ricci E. Public attitudes regarding willingness to participate in medical research studies. J Health Soc Policy 2000;12(2):23-43. [doi: 10.1300/J045v12n02_02] [Medline: 11184441]
23. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
24. Sanderson SC, Brothers KB, Mercaldo ND, et al. Public attitudes toward consent and data sharing in biobank research: a large multi-site experimental survey in the US. Am J Hum Genet 2017 Mar 2;100(3):414-427. [doi: 10.1016/j.ajhg.2017.01.021] [Medline: 28190457]
25. Lidington E, Stiles M, Maudsley J, et al. Establishing a set of acceptable demographic questions for use in health research through public consultation. Res Involv Engagem 2026 Jan 28;12(1):24. [doi: 10.1186/s40900-026-00836-1] [Medline: 41606646]
26. English indices of deprivation 2019. Ministry of Housing, Communities & Local Government. 2019. URL: https://imd-by-postcode.opendatacommunities.org/imd/2019 [accessed 2026-03-21]
27. Arias J. Postcode Deprivation Finder. URL: https://alasdairrae.github.io/postcodez/ [accessed 2026-03-21]
28. NI multiple deprivation measures. Northern Ireland Statistics and Research Agency. URL: https://deprivation.nisra.gov.uk/ [accessed 2026-03-21]
29. Health Survey for England, 2021 part 2. NHS England. 2023. URL: https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/2021-part-2/adult-health-general-health [accessed 2026-03-21]
30. GOV UK. Research and analysis: Digital Landscape Research. 2012. URL: https://www.gov.uk/government/publications/digital-landscape-research/digital-landscape-research [accessed 2026-03-21]
31. Johnson MS, Adams VM, Byrne J. Addressing fraudulent responses in online surveys: insights from a web ‐ based participatory mapping study. People and Nature 2024 Feb;6(1):147-164 [FREE Full text] [doi: 10.1002/pan3.10557]
32. Lawlor J, Thomas C, Guhin AT, et al. Suspicious and fraudulent online survey participation: introducing the REAL framework. Methodological Innovations 2021 Sep;14(3):20597991211050468. [doi: 10.1177/20597991211050467]
33. Derksen C, Ramasawmy M, Scott S. WISER survey. Open Science Framework (OSF). URL: https://osf.io/bzn4q/overview [accessed 2026-03-21]
34. Willis GB, Artino AR. What do our respondents think we're asking? Using cognitive interviewing to improve medical education surveys. J Grad Med Educ 2013 Sep;5(3):353-356. [doi: 10.4300/JGME-D-13-00154.1] [Medline: 24404294]
35. Schmeising-Barnes N, Waller J, Marlow LAV. Attitudes to multi-cancer early detection (MCED) blood tests for population-based screening: a qualitative study in Great Britain. Soc Sci Med 2024 Apr;347:116762. [doi: 10.1016/j.socscimed.2024.116762] [Medline: 38520825]

XSL•FO
RenderX

36. Waller J, Osborne K, Wardle J. Enthusiasm for cancer screening in Great Britain: a general population survey. Br J Cancer 2015 Feb 3;112(3):562-566. [doi: 10.1038/bjc.2014.643] [Medline: 25535731]
37. Cheong AT, Khoo EM, Tong SF, Liew SM. To check or not to check? A qualitative study on how the public decides on health checks for cardiovascular disease prevention. PLoS ONE 2016;11(7):e0159438. [doi: 10.1371/journal.pone.0159438] [Medline: 27415432]
38. Shiells K, Di Cara N, Skatova A, et al. Participant acceptability of digital footprint data collection strategies: an exemplar approach to participant engagement and involvement in the ALSPAC birth cohort study. Int J Popul Data Sci 2020;5(3):1728. [doi: 10.23889/ijpds.v7i1.1728] [Medline: 35519823]
39. Skatova A, McDonald R, Ma S, Maple C. Unpacking privacy: valuation of personal data protection. PLoS ONE 2023;18(5):e0284581. [doi: 10.1371/journal.pone.0284581] [Medline: 37134067]
40. Ackermann KA, Burkhalter L, Mildenberger T, Frey M, Bearth A. Willingness to share data: contextual determinants of consumers' decisions to share private data with companies. J of Consumer Behaviour 2022 Mar;21(2):375-386 [FREE Full text] [doi: 10.1002/cb.2012]
41. Baines R, Stevens S, Austin D, et al. Patient and public willingness to share personal health data for third-party or secondary uses: systematic review. J Med Internet Res 2024 Mar 5;26:e50421. [doi: 10.2196/50421] [Medline: 38441944]
42. Yusuf ZK, Dixon WG, Sharp C, Cook L, Holm S, Sanders C. Building and sustaining public trust in health data sharing for musculoskeletal research: semistructured interview and focus group study. J Med Internet Res 2024 Oct 15;26:e53024. [doi: 10.2196/53024] [Medline: 39405526]
43. Bartol J, Prevodnik K, Vehovar V, Petrovčič A. The roles of perceived privacy control, Internet privacy concerns and Internet skills in the direct and indirect Internet uses of older adults: conceptual integration and empirical testing of a theoretical model. New Media & Society 2024 Aug;26(8):4490-4510. [doi: 10.1177/14614448221122734]
44. Raibulet C, Wang K. Awareness of privacy and data collection: exploring privacy policy effectiveness in Google Maps. Front Comput Sci 2025 Sep 24;7:1568179. [doi: 10.3389/fcomp.2025.1568179]
45. Bietz MJ, Bloss CS, Calvert S, et al. Opportunities and challenges in the use of personal health data for health research. J Am Med Inform Assoc 2016 Apr;23(e1):e42-e48. [doi: 10.1093/jamia/ocv118] [Medline: 26335984]
46. Licqurish S, Phillipson L, Chiang P, Walker J, Walter F, Emery J. Cancer beliefs in ethnic minority populations: a review and meta-synthesis of qualitative studies. Eur J Cancer Care 2017 Jan;26(1):e12556 [FREE Full text] [doi: 10.1111/ecc.12556]
47. Bradbury A. Mental health stigma: the impact of age and gender on attitudes. Community Ment Health J 2020 Jul;56(5):933-938. [doi: 10.1007/s10597-020-00559-x] [Medline: 32002711]
48. Seidler ZE, Dawes AJ, Rice SM, Oliffe JL, Dhillon HM. The role of masculinity in men's help-seeking for depression: a systematic review. Clin Psychol Rev 2016 Nov;49:106-118. [doi: 10.1016/j.cpr.2016.09.002] [Medline: 27664823]
49. Wirback T, Forsell Y, Larsson JO, Engström K, Edhborg M. Experiences of depression and help-seeking described by young Swedish men. Psychol Men Masc 2018 Jul;19(3):407-417. [doi: 10.1037/men0000110]
50. Baruh L, Secinti E, Cemalcilar Z. Online privacy concerns and privacy management: a meta-analytical review. J Commun 2017 Feb;67(1):26-53 [FREE Full text] [doi: 10.1111/jcom.12276]
51. Gradwell NC, Ramasawmy M, Arora S, et al. Acceptability of sharing internet browsing history for cancer research: think-aloud and interview study. JMIR Cancer 2026 Feb 2;12:e82009. [doi: 10.2196/82009] [Medline: 41628357]
52. Thombs BD, Kwakkenbos L, Levis AW, Benedetti A. Addressing overestimation of the prevalence of depression based on self-report screening questionnaires. CMAJ 2018 Jan 15;190(2):E44-E49. [doi: 10.1503/cmaj.170691] [Medline: 29335262]
53. Education, England and Wales: census 2021. Office for National Statistics. 2021. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/educationandchildcare/bulletins/educationenglandandwales/census2021 [accessed 2026-03-21]
54. Bryere J, Pornet C, Copin N, et al. Assessment of the ecological bias of seven aggregate social deprivation indices. BMC Public Health 2017 Jan 17;17(1):86. [doi: 10.1186/s12889-016-4007-8] [Medline: 28095815]
55. Capuano AW, Dawson JD, Ramirez MR, Wilson RS, Barnes LL, Fields RW. Modeling Likert scale outcomes with trend-proportional odds with and without cluster data. Methodology (Gott) 2016 Apr;12(2):33-43. [doi: 10.1027/1614-2241/a000106] [Medline: 37583928]
56. Comachio J, Poulsen A, Bamgboje-Ayodele A, et al. Identifying and counteracting fraudulent responses in online recruitment for health research: a scoping review. BMJ EBM 2025 Jun;30(3):173-182. [doi: 10.1136/bmjebm-2024-113170]

Abbreviations

GDPR: General Data Protection Regulation
JARS-Quant: Journal Article Reporting Standards Quantitative Research Design
OR: odds ratio
OSF: Open Science Framework
PPI: patient and public involvement

XSL•FO
RenderX

Edited by S Brini; submitted 10.Oct.2025; peer-reviewed by N Calanzani, N Howe; revised version received 07.Mar.2026; accepted 07.Mar.2026; published 25.Mar.2026.

Please cite as:
Derksen C, Ramasawmy M, Arora S, Lau W, Scott SE
Willingness to Share Internet Use Data for Research on Early Disease Detection: Cross-Sectional Survey
J Med Internet Res 2026;28:e85637
URL: https://www.jmir.org/2026/1/e85637
doi:10.2196/85637

© Christina Derksen, Mel Ramasawmy, Sanjula Arora, Wynsee Lau, Suzanne Scott. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

From Searching to Coping, How Chinese Patients With Breast Cancer Navigate Web-Based Health Information: Semistructured Interview Study

Jiahui Liu[1], MSc; Xingfeng Li[1], MSc; Xuying Li[2], PhD; Jiejun Chen[1], MSc; Dan Zhang[2], MSc; Cuiling Qiu[2], MSc

[1]Department of Breast Medicine, Hunan Cancer Hospital, 582 Xianjiahu Road, Yuelu District, Changsha, Hunan, China

[2]Department of Nursing, Hunan Cancer Hospital, 582 Xianjiahu Road, Yuelu District, Changsha, Hunan, China

Corresponding Author:
Cuiling Qiu, MSc
Department of Nursing, Hunan Cancer Hospital, 582 Xianjiahu Road, Yuelu District, Changsha, Hunan, China

Abstract

Background: With the development of digital health platforms, patients with breast cancer are increasingly relying on web-based resources to search for disease-related information. Proper usage of web-based health information by patients with breast cancer is crucial for understanding disease information and participating in treatment decisions. However, in the face of the large amount and complexity of information, it is still unclear how patients can make psychological adjustments and behavioral responses. Problems such as variable information quality and conflicting information are also affecting the cognitive and treatment decision-making process of patients with breast cancer.

Objective: This study aims to explore the real experiences of Chinese patients with breast cancer in their search for web-based health information from a phenomenological perspective, providing insights for optimizing future web-based health information support for patients.

Methods: This qualitative study used semistructured, in-depth face-to-face interviews to collect data. Through purposive and convenience sampling, 18 female patients with breast cancer were recruited from a tertiary cancer hospital in China. The data saturation principle was observed to determine the endpoint of data collection. The collected data were analyzed using thematic analysis.

Results: From 18 original interview documents, three themes and 11 subthemes were categorized as follows: (1) driving force of information search (emotion-based information search, problem-solving–oriented information search), (2) cognitive judgments amidst the information fog (interweaving of multichannel information, judgment of information authenticity, information applicability assessment, cognitive confusion in the context of information conflict, and construction of information meaning), and (3) adaptation under the pressure of web-based information (transform information into action, emotional regulatory coping, build a support network, and acceptance and adjustment of expectations).

Conclusions: This study reveals that the experiences of patients with breast cancer within web-based health information environments resemble an information navigation journey. Patients continuously search, evaluate, and adjust within the sea of information to maintain cognitive clarity and emotional equilibrium. The findings offer valuable insights for clinical health care providers, health information platform developers, and policymakers. They can help optimize digital health services and design personalized information support that better meets patients' needs.

(J Med Internet Res 2026;28:e80363) doi:10.2196/80363

KEYWORDS

breast cancer; information seeking; decision-making; coping strategy; qualitative research

Introduction

According to the latest global cancer data in 2020, breast cancer has surpassed lung cancer to become the world's leading cancer, being the main cause of cancer incidence, disability, and mortality among women worldwide, imposing a huge burden on global health services [1]. With the advancement of medical technology and the promotion of early screening, more and more patients with breast cancer can receive early diagnosis and treatment, and their survival prognosis has significantly improved compared with before [2,3]. However, due to the complex pathological characteristics of breast cancer, diverse treatment options, and long-term rehabilitation needs, patients' demand for health information is extremely urgent. A survey study found that among the patients with breast cancer interviewed, as many as 85% expressed a desire to obtain more information about their disease [4]. Sheehy et al [5] conducted a prospective longitudinal investigation and found that the

XSL•FO
RenderX

information needs of patients with breast cancer remained high throughout the entire follow-up period of 1 year, 3 years, and 5 years after diagnosis. The research conducted by Ludwigson et al [6] revealed that patients with breast cancer were dissatisfied with the health information they received regarding disease prognosis, symptom management, and surgical reconstruction. It can be seen that patients with breast cancer have long-term demands for health information, but these demands have not been met.

Therefore, with the rapid development of the internet and the increasing inclusiveness, accessibility, and comprehensibility of digital media, an increasing number of patients are beginning to actively seek health information through web-based channels to assist in their treatment decisions [7]. It has been reported that the internet has become the preferred source of information for patients with cancer, surpassing health care professionals [8]. According to relevant research reports, breast cancer is one of the most frequently searched cancer topics on the internet [9]. Online health information search (OHIS) plays a significant role in the health management of patients with chronic diseases, enabling them to understand their conditions and facilitating their participation in the decision-making process regarding treatment [10,11].

However, the current digital health information environment also presents numerous challenges. Given the large amount of inaccurate web-based information, people can easily become misinformed [12]. A study has shown that misleading web-based health information is associated with poor patient outcomes [13]. Moreover, there is an abundance of web-based health information, but the quality of this information is difficult to guarantee [14]. In the face of the challenges brought by web-based information, patients with breast cancer often need to make psychological trade-offs, forming a complex coping mechanism. The “Stress and Coping Theory” proposed by Lazarus and Folkman [15] provides a theoretical framework for understanding this phenomenon. According to this theory, when individuals encounter potentially threatening events, they undergo primary appraisal and secondary appraisal and, based on the situation, choose appropriate coping strategies [16]. This theory proposes two primary coping strategies: (1) problem-focused coping, aimed at resolving the problem or situation; and (2) emotion-focused/avoidance coping, aimed at regulating stress-related emotions [17]. Problem-focused coping involves acquiring relevant information, while emotion-focused coping entails avoiding thoughts about the threat or reevaluating the threat without altering the stressful reality. These 2 coping strategies may either promote or mutually hinder each other in stressful situations [18]. This theory has been widely applied to analyze the psychological adaptation mechanisms of patients in disease management and health behavior decision-making [19,20].

Although some studies have preliminarily explored the OHIS behaviors of patients with breast cancer, limited research has explored their subjective experiences, cognitive processes, and coping strategies—particularly within the Chinese cultural context [8,21]. Most of the existing literature focuses on the objective characteristics of information behavior. The perspective of phenomenological research emphasizes the genuine understanding of individual experiences [22]. Therefore, this study aims to explore the experiences of patients with breast cancer in web-based health information contexts from a phenomenological perspective. To facilitate a deeper understanding of the psychological mechanisms reflected in participants’ experiences, the Stress and Coping Theory was drawn upon as an interpretive reference. By filling this underexplored research gap, this study has deepened our understanding of how breast cancer patients search for, evaluate, and apply health information in a complex information environment. The research results provide practical guidance for designing personalized and operational web-based health interventions and information support strategies in clinical practice and may also offer insights for the development of digital health services for other chronic diseases or vulnerable groups.

Methods

Study Design

Guided by a phenomenological perspective and using qualitative research methods, this study aims to explore in depth the experiences and feelings of Chinese patients with breast cancer as they search for web-based health information during their disease journey [22]. Data were collected through semistructured face-to-face in-depth interviews. This study drew upon Lazarus and Folkman’s [15] stress-coping theory in designing the interview guide. Following thematic analysis, researchers used this theoretical framework to interpret patients’ cognitive evaluations and coping strategies within their web-based information-seeking behaviors, thereby enhancing the theoretical significance of the findings. Originally proposed in 1984, this classic psychological theory explains how individuals assess and respond to stressful events through a dynamic regulatory process involving person–environment interactions. The study reporting followed the COREQ (Consolidated Criteria for Reporting Qualitative Research Checklist; Checklist 1) [23].

Setting and Participants

From September 2024 to May 2025, participants were recruited through a combination of convenience and purposive sampling in 3 breast cancer inpatient wards at Hunan Cancer Hospital in China. Hunan Province is a typical province in central China, comprising 14 prefecture-level cities. Its diversity in language and culture, urban-rural structure, health care resources, economic development, and residents’ digital literacy provides a suitable setting for examining the contextual variations in information-seeking behaviors [24].

Two female researchers (LJH and ZD) were Advanced Practice Nurses in breast oncology with training in breast disease management and qualitative interviewing. The researcher has experience practicing in participants’ workplaces and is responsible for participant screening and interview scheduling. First, based on convenience sampling principles, the researcher reviewed daily inpatient records and made preliminary assessments of eligibility based on medical documentation. Subsequently, clinicians and charge nurses assisted in notifying eligible patients. Researchers then conducted face-to-face interactions with patients, explaining the study objectives,

XSL•FO
RenderX

procedures, timeline, and privacy safeguards to invite participation. To enrich the data dimensions, the research team purposefully included patients of varying ages, educational levels, disease stages, duration since diagnosis, and urban/rural residence after initial recruitment, continuing until data saturation was achieved.

Participant Recruitment and Enrollment

Participants were female patients diagnosed with breast cancer who had experience in searching for information on the internet. Inclusion criteria included the following: (1) aged 18 years and older; (2) diagnosed with breast cancer; (3) able to express themselves clearly, willing to participate in the study, and sign an informed consent form; and (4) previously actively searched for health information on the internet or mobile platforms during the disease. Exclusion criteria included the following: (1) comorbid serious mental illness affecting the interview communication; (2) those who refused to record or were unable to complete the interview.

Data Collection

Data were collected through semistructured, in-depth interviews. Based on the study aims, a review of relevant literature, and the theoretical framework of stress and coping, the interview guide was developed collaboratively by the research team members (LJH and QCL) with guidance from a qualitative research expert (LXY). To enhance clarity, exploratory questions were incorporated into the primary interview questions to better capture the depth and complexity of participants' experiences, as shown in Table 1. Two participants who met the inclusion criteria were invited for a pilot interview. Based on the feedback from the pilot participants, the wording and sequence of the interview questions were adjusted, and the exploratory questions were also optimized.

Table . Interview guide.

Main question	Exploratory questions
a. Please tell us about your motivation and initial experience of online health information searching.	• Under what circumstances did you begin using the internet to search for health information related to breast cancer? • What kind of information were you initially looking for? • What is the primary motivation for searching for information online?
b. Please tell us about the process of your information search.	• Which platforms or websites do you typically use to access online health information? • What challenges have you encountered during the information search process? • How do you screen and evaluate online health information? • Do you have any criteria or methods you rely on?
c. Please tell us about the impact of online health information on your emotions.	• How do you feel when searching for and reading online health information? • How has online health impacted you? • Could you describe your emotional experience in more detail?
d. Please tell us about the impact of online health information on your treatment decision-making.	• l How does online health information influence your treatment decisions? • How do you think searching for health information online affects your communication with your doctor? • Can you discuss how you question or offer suggestions to your doctor based on online health information?
e. Please tell us about your overall evaluation of and expectations for online health information.	• What role do you think online health information played in your journey to overcome breast cancer? • If given the opportunity, how would you improve your information-seeking process? • What kind of assistance or support would you like to receive?

The interview began after obtaining the participants' verbal and written informed consent. The entire process was audio-recorded after obtaining their consent, and at the same time, the researchers took field notes to capture nonverbal information and key immediate thoughts. During the interview, the researcher flexibly adjusted the order and depth of questions according to the participants' expressions and responses and encouraged patients to describe their information-seeking experience in-depth and truthfully. The average duration of the interviews is approximately 45 minutes (range 20 - 60 min). All interviews were conducted in quiet hospital spaces designated for patient–clinician communication to ensure privacy and comfort. To minimize interviewer bias and ensure the authenticity of participants' expressions, the researcher adopted a neutral and nonevaluative stance during the interviews, avoided leading or suggestive questions, and respected each participant's subjective experiences. When no new topics are generated, data collection is halted [25]. When no new significant themes emerge after analyzing 2 consecutive rounds of interviews, data saturation is achieved, and sampling ceases.

XSL•FO
RenderX

Data Analysis

The interview recordings were transcribed verbatim into Chinese text by the researcher (LJH) within 24 hours of completion and verified for accuracy by the researcher (ZD). Transcribed texts are managed using Excel, with each record containing participant ID, raw text, open coding, thematic categorization, and researcher notes. Excel files are uniformly stored in a controlled environment and encrypted to ensure participant information security and privacy.

Using thematic analysis as the methodological basis for theme summarization [26]. Two researchers (LJH and ZD) independently conducted the initial coding and then discussed and reached a consensus. In case of disagreement, a third-party expert (LXF) was invited to participate in the discussion to ensure the reliability and authenticity of the analysis. After coding completion, the researcher (LJH) categorized and merged relevant codes to form preliminary themes. These themes underwent discussion and review by all researchers to ensure alignment with the original data. Subsequently, themes were defined and named, with descriptive explanations drafted for each. Finally, themes were linked to corresponding original data citations. The researchers provided the preliminary analysis results to some participants for feedback and solicited their opinions to ensure the authenticity and representativeness of the analysis. To ensure the accuracy of language conversion, all citations and theme descriptions intended for publication were translated from the original Chinese into English by professional bilingual personnel with medical backgrounds. Subsequently, all members of the research team conducted back-translation and concept [27].

Reflexivity

Throughout the research process, the team maintained a reflective stance to enhance transparency, credibility, and rigor. Given that the interviewers had professional backgrounds in the breast department, face-to-face semistructured interviews fostered trust, enabling participants to be more willing to share their personal experiences. However, this also risked leading researchers to overinterpret participants' perspectives during interviews, potentially overlooking subtle variations in their experiences. Therefore, researchers maintained reflective journals before and after each interview, documenting their preconceptions, emotional responses, interview interactions, and any subjective impressions that might influence interpretation. These records were repeatedly reviewed during subsequent analysis. During data analysis, researchers repeatedly returned to original recordings to ensure initial coding stemmed from participants' authentic expressions. Researchers took particular care to separate clinical experience from the analytical process, avoiding the substitution of professional judgments for meanings presented in participants' narratives. Throughout theme development, the research team continuously challenged, compared, and refined interpretations through ongoing peer discussions and reflective dialogues. This multiperspective examination of coding consistency and thematic robustness enhanced the credibility and confirmation of research findings. Additionally, the team deliberately revisited earlier hypotheses and analytical pathways during the later stages of analysis [28].

Ethical Considerations

The study was approved by the Hunan Cancer Hospital Medical Ethics Review Committee (number SBQLL-2024 - 180). All participants signed written informed consent forms prior to the commencement of face-to-face interviews, ensuring their participation was voluntary, information was anonymized, and data were used solely for this study. Furthermore, no financial compensation was provided to any participants. All raw interview data and participant information underwent anonymization and were stored on an encrypted cloud storage drive to ensure confidentiality.

Results

Demographic Characteristics

A total of 18 female patients with breast cancer participated, originating from Wuhan, Hubei Province (n=1), Jiamusi City, Heilongjiang Province (n=1), and 9 prefecture-level cities in Hunan Province (Changsha, Yiyang, Changde, Hengyang, Loudi, Zhuzhou, Shaoyang, Yongzhou, and Yueyang) (n=16). All the participants were of Han ethnicity. All patients with breast cancer are currently undergoing treatment in the hospital. Except for one participant who had bilateral breast cancer, all the other participants had unilateral breast cancer. The time of cancer diagnosis for the participants ranged from 1 month to 20 years. The most commonly used digital search platform by the participants was Baidu. The general information of the participants is shown in Table 2.

XSL•FO
RenderX

Table . Characteristics of the participants.

Variables	Participants (N=18)
Age (years), mean (SD)	42.29 (10.60)
Education background, n (%)	
Junior high school and below	5（27.8）
Senior high school	3（16.7）
Bachelor degree	8（44.4）
Graduate degree	2（11.1）
Employment, n (%)	
Employed	6（33.3）
Unemployed	10（55.6）
Retired	2（11.1）
Marital status, n (%)	
Unmarried	1（5.56）
Married	15（83.3）
Divorced	2（11.1）
Medical insurance payment methods, n (%)	
Employee Medical Insurance	12（66.7）
Resident Medical Insurance	3（16.7）
New Rural Cooperative Medical Scheme	3（16.7）
Year of breast cancer diagnosis, n (%)	
＜1	9（50.0）
1-4	4（22.2）
5-9	2（11.1）
＞10	3（16.7）
Clinical stages, n (%)	
Stage Ⅰ-II	9（50.0）
Stage Ⅲ-IV	9（50.0）
Relapse and metastasis, n (%)	
Yes	9（50.0）
No	9（50.0）
Digital seeking platform, n (%)	
Baidu[a]	15（83.3）
XiaoHongShu[b]	5（27.8）
TikTok[c]	7（38.9）
Deepseek[d]	7（38.9）
WeChat[e]	4（22.2）
TengXunYuanBao[d]	1（5.56）

[a]Baidu: Comprehensive search engine.

[b]Xiaohongshu: Social sharing platform primarily featuring user-generated content.

[c]TikTok: Short-video-centric information dissemination platform.

[d]DeepSeek and TengXunYuanBao: Artificial intelligence-based intelligent service platform.

[e]WeChat: Comprehensive social platform.

XSL•FO
RenderX

This study identified three core themes and 11 subthemes of OHIS in patients with breast cancer through thematic analysis: (1) the driving forces of information search were emotion-driven and problem-oriented; (2) cognitive judgments in the information fog included the interweaving of multiple channels of information, assessment of information authenticity and applicability, cognitive confusion under information conflicts, and the construction of information meaning; (3) adaptive strategies under the pressure of network information encompassed information conversion into action, emotion regulation, support network construction, and expectation acceptance and adjustment. Overall, the OHIS behaviors of patients with breast cancer present a dynamic process driven by motivation, cognitive processing, and adaptive strategies. The overall coding framework, including the hierarchical relationship between themes and subthemes, is shown in Figure 1.

Figure 1. Coding framework of the web-based health information navigation process among patients with breast cancer.

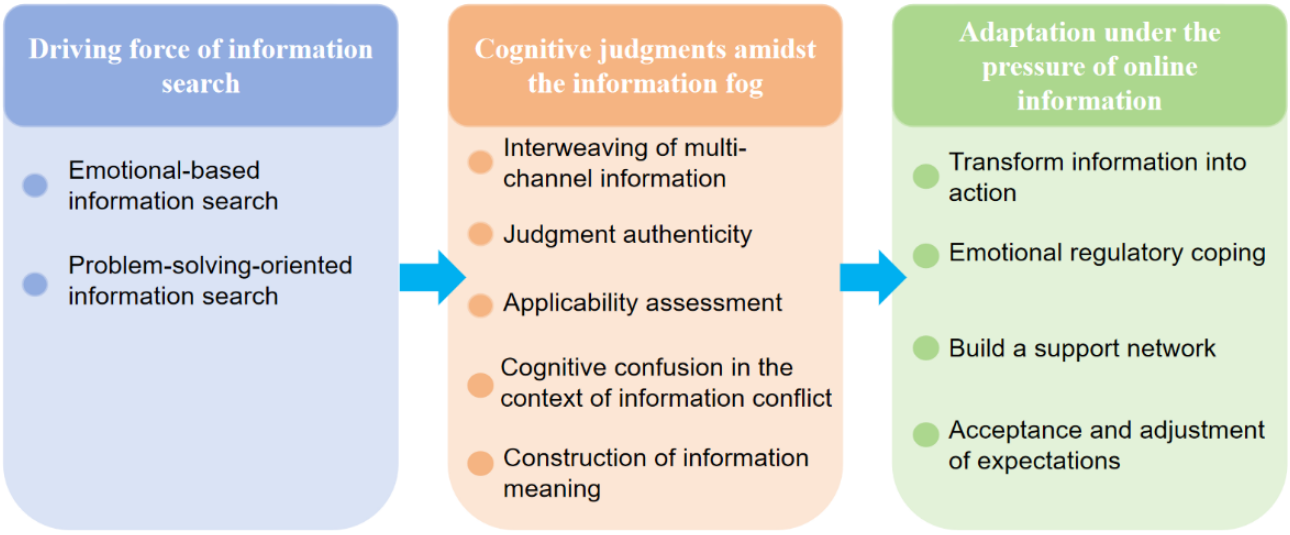

Driving Force of Information Search

Overview

This theme reflected the multiple drivers behind information-seeking behaviors of patients with breast cancer, stemming from the combined influence of emotional needs and practical concerns. On one hand, information retrieval served to alleviate uncertainty and provide emotional reassurance; on the other, patients also used web-based resources to understand their disease, support treatment decisions, and address daily management needs.

Emotion-Based Information Search

The information-seeking behavior of participants was an emotionally driven experience process. It began with the anxiety, fear, and uncertainty triggered by the diagnosis, and the participants attempted to obtain psychological comfort and a sense of control through searching.

> *I just wanna see if it's really breast cancer. It's just for some peace of mind. If it's not, I'll definitely feel better, right? I'm also really scared about myself…* [P6]

> *That kind of bad news really makes me anxious… Especially when I see stuff about having to take medicine long - term and go through all these treatments. It makes me super anxious, like the treatments just never end!* [P5]

> *I'm worried if there's still hope for a cure… 'Cause I've got old folks to take care of and kids to raise. Wanna check if there are any such cases…* [P1]

However, web-based health information not only serves a knowledge function; its content itself could directly influence patients' emotional states, becoming a significant trigger that amplifies positive or negative emotions. When search results were interpreted as "positive," patients often experienced a brief sense of hope and emotional uplift; conversely, negative or uncertain information could significantly exacerbate anxiety and emotional distress.

> *If the search results are good, I'll feel a bit happier. If they're bad, I'll definitely worry more.* [P6]

> *When you see something good, you'll feel great. But when you see something bad, it'll affect you. Your mood can easily be influenced by the internet. Especially when you come across something negative, you might feel depressed for days…* [P4]

In the collectivist culture of Chinese families, illness was regarded as a shared family event. Family members would proactively intervene in the patient's information-seeking behavior and act as an "emotional firewall" by discouraging the patient from searching for information. As a result, patients were protected and constrained by family emotions from the very beginning.

> *My husband doesn't want me to search randomly online. He said, "Don't look at random stuff on the Internet. Just follow what the doctor says.* [P6]

Problem-Solving–Oriented Information Search

The information-seeking behavior of participants revolved around a series of specific questions, reflecting their use of web-based resources as a vital supplement to inadequate clinical

XSL•FO
RenderX

information. Through information retrieval, patients sought to enhance their sense of control over the disease, understand diagnostic outcomes and disease staging, explore treatment options, and support decision-making, while also addressing symptoms, side effects, and issues related to recovery and lifestyle adjustments. The participants described their experience of understanding disease staging through web-based information.

> *I searched for it online. The doctor told me not to look it up, but I did anyway. T stands for the diameter, like the size of the tumor. So if it's T2, that means it's in category 2, right? And N1 means there's lymph node metastasis. I couldn't get an answer from the doctor, so later, when I checked my discharge record, I saw that I'm T1N1M0, which means there's no distant metastasis. After I found that out online, I was like, okay, I get it.* [P17]

In the management of symptoms and side effects, information retrieval was also used to guide practical interventions.

> *Mainly, I wanna search for the side effects of the injection. 'Cause I felt really awful after getting the red potion injected four times before. Every time, my stomach was killing me, and I'd throw up. So I wanna look it up online to see which time the side effects would be the worst and stuff. Then I'll see if the side effects will be less severe after I switch to a different medicine.* [P6]

> *Like, if my white blood cell count is low and I'm wondering what kind of medicine to take, or if I've got a bad taste in my mouth and my tongue is really white and thick and I wanna know how to relieve it, I'll just search for it online.* [P3]

Several participants described that the brevity of clinical consultations often left their information needs insufficiently addressed, leading them to rely on web-based resources for additional medical explanations.

> *You can see how busy the doctor is here. He can't talk to you for more than 10 minutes. He has so many patients, so there are many things you have to learn by yourself from other sources.* [P7]

Cognitive Judgments Amidst the Information Fog

Overview

When patients with breast cancer sought web-based health information across various platforms, they found themselves immersed in a landscape of multiple sources and diverse perspectives. This interweaving of information from different channels offered patients more clues for understanding their illness and treatment options, while also increasing the complexity of their cognitive judgment. Patients had to repeatedly weigh the authenticity of the information and its suitability for their specific disease stage and physical condition. When web-based information conflicted with existing knowledge or medical advice, they often experienced cognitive dissonance and indecision. Throughout this process, patients attempted to compare, filter, and reflect upon fragmented or conflicting information, assigning it personalized meaning in alignment with their own experiences.

Interweaving of Multichannel Information

Participants obtained health information through various web-based channels. Some initiated inquiries using general search engines, such as Baidu, and sought the truth by comparing multiple search results.

> *Yeah, I used Baidu, no other search engines. I think it's pretty reliable 'cause most of the stuff I saw was posted by doctors. Like which doctor from which hospital… For the same question, check out a few more results. If they're more or less the same, then it should be pretty accurate. Just search the same question a couple more times.* [P6]

Some participants searched for real-life experiences shared by other patients on social media platforms, while remaining highly vigilant against advertisements and potential scams.

> *Sometimes I also used Douyin. There are some people who are fellow patients, and they post videos directly on it. For example, like the Pyrrotinib I'm taking, they'll show its side - effects on Douyin, and it's more vivid there… What's on Douyin is a real reflection. Like some people do live - streaming to talk about their illness progress. They're quite honest and won't tell lies. Sometimes when I compare, it seems like I have the same symptoms too. Of course, some people might be just trying to sell stuff, so I stop watching after that…* [P7]

Some participants also viewed artificial intelligence (AI) tools as efficient and objective sources for interpreting medical terms and integrating health knowledge.

> *Besides the imaging data, I also had a carcinoembryonic antigen test. These are the two main factors, right? … I read on DeepSeek that it mainly looks at CA and CA153, 'cause these two things are characteristic of a high response to breast cancer. … My daughter checked both DeepSeek and ChatGPT for me, and the results were pretty much the same. … I think this stuff is great. It's all about the data, and now AI can be trusted. Isn't there an AI hospital opened in Shanghai now?* [P13]

Most participants tended to trust recommendations from doctors or hospital platforms, with their choices often determined by assessments of these channels' authority, convenience, and personal usage habits.

> *Can you still modify this interface? This was the essence of deep thinking. I don't often use Deepseek. I clicked it a few times, then stopped. Sometimes it got stuck, but once it got stuck, I couldn't get out of the interface.* [P12]

> *I think what Deepseek provides was a more objective perspective. They won't say something is definitely bad or good. There are no vested interests involved.* [P11]

XSL•FO
RenderX

Judgment of Information Authenticity

In assessing the authenticity of web-based health information, participants often cast themselves as information detectives, proactively using various cognitive strategies to discern truth from falsehood. Confronted with complex sources and uneven-quality web-based content, patients attempted to mitigate potential risks through personal experience and repeated verification. Some participants noted that on platforms dominated by user-generated content, the authenticity of information relied heavily on individual judgment. Since these platforms featured both content based on genuine experiences and information driven by advertising or commercial promotion, participants had to continually discern between them while reading.

> *On Xiaohongshu, you have to make your own judgment. Because there might be many that are relatively genuine, and there might also be some that are advertising-oriented, aiming to tell you what medicine to take or what to do. So they tend to write in that direction.* [P11]

To enhance the reliability of their judgments, some participants cross-verified information by conducting cross-platform or multiple searches. This reflected their active efforts to strengthen confidence in the authenticity of information through consistency validation.

> *For the same question, I would look at several results. If they are roughly the same, then they should not be too different. Search a few more times.* [P6]

Additionally, some participants emphasized the authority of information sources and the scientific rigor of content, often choosing to remain cautious when a foundation of trust was lacking.

> *What was said online can be quite confusing. You need to distinguish between what is true and what is false…Like those kinds of things on TikTok that are uncertain, where I don't have any real experience, or it's not with friends or anything like that. I don't dare to try them; I don't dare to take any risks or give up the treatment here and try something else (on the internet), because the stakes are too high.* [P2]

Several participants further developed verification strategies for authoritative personal accounts, reflecting patients' critical stance toward superficial authority.

> *First of all, I definitely need to check if this person is real… I'll look at his videos, the words he says to each patient. Where could he find so many shills? It's impossible. So many different patients.* [P14]

However, a portion of participants acknowledged that, when confronted with a complex information environment, their own judgment abilities were limited, which often left them feeling uncertain and powerless in verifying information.

> *I'm not sure whether the information on the internet is good or bad. Some of it is correct, while some is incorrect.* [P8]

Information Applicability Assessment

Upon obtaining information, participants typically did not adopt it immediately. Instead, they analyzed its applicability in light of their personal condition, treatment stage, and individual differences. One participant noted that when using generative AI tools, they did not directly follow the suggestions provided but rather treated them as supplementary reference material.

> *DeepSeek really thought things through and gave more detailed advice. For example, when I asked about eating peaches, it warned me to be cautious. Then, based on that advice and my own situation—like whether I had mouth ulcers or diarrhea—I decided whether to eat them or not. So, if I had diarrhea that day, I knew not to eat peaches because they might make it worse. I appreciated how it reminded me of these things so I could choose whether it was okay to eat or better to avoid.* [P18]

At the same time, participants recognized that web-based information was often based on group data and might not be fully applicable to individual circumstances. This awareness prompted them to maintain a certain distance from such information and to avoid applying it rigidly.

> *Regarding the internet, it is also based on big data, right? It's not an individual, either. For instance, how this person behaves might also be the case for you. It's just a proportion or percentage of the big data…* [P9]

Some participants prioritized seeking information that aligned closely with their personal needs, using this as the primary criterion for judging the information's value.

> *I won't browse random online information anymore. Whenever I come across any content suggesting a poor prognosis for triple-negative cases, I won't even bother to read it. I only look at those that are of the same type as mine and say the prognosis is relatively good.* [P14]

Several participants mentioned that they adjusted the content of their information search based on their needs at different times in treatment.

> *If after my radiotherapy is completed, how should I recover, how to exercise, learn to do sports, I will continue to search online, based on the actual situation. I need to change the search content according to my own disease development. For example, in the later stage, if I have just finished chemotherapy, I will start radiotherapy. What should I pay attention to?* [P3]

Moreover, the assessment of information relevance demonstrated clear phase-specific characteristics. Several participants indicated that the content of their information-seeking was continually adjusted throughout the treatment process.

> *After getting sick, I started searching for information about the disease online… I rested for half a year and didn't search again after returning to work. This time,*

XSL•FO
RenderX

the illness recurred, and I was more diligent in checking the information. [P1]

Cognitive Confusion in the Context of Information Conflict

When web-based health information conflicted with or contradicted treatment plans prescribed by their primary physicians, participants often experienced cognitive dissonance, which led to decision-making difficulties and a crisis of trust. In such situations, they had to repeatedly weigh conflicting advice from different authoritative sources, considerably heightening their psychological burden. Some participants described feeling confused and hesitant when they encountered discrepancies between web-based information and their doctors' advice.

> *The internet also says so. The doctors here say that radiotherapy is needed. Of course, there are doubts. Don't you think so? I'm just a patient, and I'm very confused. As long as someone says you don't have to have radiotherapy or chemotherapy, you'll be torn. It's all like this, right?* [P14]

> *It's still a very difficult decision-making process… The information on the internet is somewhat contrary to that of the doctor. It can affect the treatment plan your doctor gives you… Then your confidence in your attending doctor may decrease.* [P4]

However, when web-based information conflicted with what their doctors said, patients often kept their doubts to themselves, for fear of challenging medical authority. This reflected patients' reluctance to question professional authority within a high power-distance cultural context.

> *I didn't tell the doctor I got the info online 'cause I'm afraid the doc's gonna diss me. Like, the doc might say, "If you trust the internet so much, go ask those people online. Don't come to me.* [P4]

The direct consequence of information contradiction was a widespread deadlock in choice and a psychological burden. This contradiction left patients feeling caught in the middle.

> *Most doctors seem to say pretty much the same thing. But then it's like, traditional Chinese medicine has its own way of looking at it, and Western medicine has another.* [P2]

> *It's really tough to figure out what's true and what's not on your own—there's just so much conflicting info out there.* [P14]

Faced with decision-making difficulties, some participants would repeatedly verify and dwell on these contradictory pieces of information, reflecting their underlying desire for cognitive consistency.

> *...I've looked online, and some of them talk about how it's advanced and how it's impossible to cure this disease anyway, that you live with cancer, but one of the points I'm struggling with right now is that some of them talk about how your body will definitely break down if you do chemo, and that's 100% true. But if you don't do chemo, and this disease just continues, which option is better?* [P8]

Construction of Information Meaning

By integrating medical data with personal experiences, patients were able to derive new meaning from information. Dissatisfied with her doctor's brief explanation, one participant independently conducted a detailed assessment, comparing earlier imaging reports and using AI tools to perform calculations.

> *I myself have tested for carcinoembryonic antigen… Besides the imaging data, I also have the carcinoembryonic antigen. These are the two main factors, right?" … My daughter checked DeepSeek and ChatGPT for me, and the results were similar… We also calculated the volume… For example, for magnetic resonance imaging (MRI), I used MRI as a comparison… Then we calculated its volume… This indicates that the size has decreased… Overall, it has shrunk by 68.95%. I'm very happy that it's reduced by two-thirds… This proves that my own thinking and reaction were correct.* [P13]

After repeated comparisons and deliberation, participants developed a stable perception of the roles played by different information sources. One participant clearly distinguished the value of AI tools from that of doctors.

> *Deepseek is just a reference sometimes. It might explain things in a more detailed and nitpicky way. Doctors, on the other hand, might just go through things quickly. If you don't understand, you can ask. Docs talk so fast that you might not get what they mean at all. You can ask Deepseek, and it'll give you a more detailed explanation. But like those indicators I asked you about last time, well, you know, professionals should do professional jobs. It can't replace doctors and nurses.* [P18]

Some participants also constructed their own visions of the future based on the individual stories of their fellow patients.

> *You know, when you're scrolling through stuff these days, there are a bunch of really well - known breast cancer doctors with their own personal accounts… If you come across the good cases they share, it'll have a positive impact on you. But sometimes, if you see the not - so - good stuff, it can mess with your mood. There's this one doctor's video. He said one of his patients was in a similar situation to me, and it seemed like the prognosis was okay. That patient also lived a pretty long time.* [P4]

Through long-term engagement with information, patients gradually developed their own knowledge frameworks, transforming fragmented information into a coherent understanding of their illness.

> *I've been looking up new drugs and treatments—'cause, let's be real, everyone just wants to survive. When I first got diagnosed, I pretty much did all the research I could. I got a decent idea of the*

XSL•FO
RenderX

> *main treatment options—first-line, second-line meds, all that. So I kind of accepted, okay, I'll go through those during this year of treatment. But then I started thinking—what happens after that? Will there be any new drugs by then? Any new treatment approaches I could try next?* [P10]

Participants defined the process of information searching as an empowerment process, through which they enhanced their information literacy by means of continuous exploration and summarization.

> *Searching for reading materials is also a form of learning and it helps us, y'know. I think it's pretty good to have this kind of resource. At least it can guide me. Doctors are so busy taking care of so many patients. There's no way they can cater to every single patient's needs. Sometimes, you gotta rely on yourself.* [P7]

> *No matter how smart AI is, it's not like it can read your mind. But I do think it's something you have to learn how to use. The better you understand it, the more likely it is to give you the answers you actually want. I'm not super familiar with it yet—like, I haven't really studied it in a structured way or anything. But now there are more and more tools out there, like DeepSeek and other platforms, to help people learn.* [P11]

Through ongoing exploration, comparison, and summarization of information, these participants gradually built a firm understanding of their disease, treatment options, and personal responses. They subsequently translated this knowledge into actionable guidance for both decision-making and daily life.

Adaptation Under the Pressure of Web-Based Information

Overview

Facing the vast and complex web-based health information, patients with breast cancer have developed various adaptation strategies to alleviate the cognitive load, emotional stress, and decision-making difficulties brought about by the information. These strategies mainly fall into 4 categories: translating information into actions, emotional regulation and coping, establishing support networks, and acceptance and expectation adjustment.

Transform Information Into Action

This subtheme illustrated how participants translated acquired knowledge into concrete daily actions after obtaining, filtering, and verifying web-based health information. This process strengthened their self-management capacities and improved communication with health care providers. Some participants mentioned modifying their diets based on recommendations from tools like DeepSeek, tailoring the information to their individual health conditions.

> *Like, Deepseek says you gotta be careful when eating peaches. If you've got oral ulcers or diarrhea, don't eat 'em. So I'll judge based on how I'm feeling today. If I'm having diarrhea, I know not to eat 'em. But if my appetite's normal, no oral ulcers, and I don't feel nauseous, then I can eat 'em. So I can choose whether to eat 'em or not according to its suggestions.* [P18]

In information-driven self-care practices, some participants proactively integrated traditional Chinese health preservation wisdom, regarded as a crucial action resource, into their daily health management.

> *I watched a video on Video Account about the Three-Bean Drink. It seems pretty good. I mean, it's basically a food therapy, you know?* [P18]

In medical communication, participants drew on web-based information to engage with their physicians in seeking personalized guidance. This practice demonstrated how patients translated external information into concrete actions and interaction strategies, thereby obtaining a sense of psychological reassurance and decision support through such exchanges.

> *I saw someone on Xiaohongshu saying that taking Sanguisorba Tablets can increase white blood cell count. So I went to ask my doctor if I could take those pills. I was like, "Doc, I saw this on Xiaohongshu, and I even took a screenshot." Then I showed it to my doctor.* [P12]

Participants further integrated web-based rehabilitation information into their daily training routines, achieving a phased transition from cognition to action. Through practice, they continuously tested and adjusted the applicability of the information while cultivating a sense of self-efficacy. This process reinforced the connection between information empowerment and tangible rehabilitation outcomes.

> *Sometimes people on the Internet tell me to do exercises. My hands have recovered better than most people because I do exercises well. They tell me what exercises to do online, and I do them for 40 minutes. I jog for 20 minutes and do exercises for another 20 minutes. Actually, I stick to it every day. On the third and fourth days, I start with 10 minutes, then 15 minutes, and then 20 minutes. I gradually increase the time.* [P3]

Emotional Regulatory Coping

Faced with vast and often complex web-based health information, participants used multiple strategies to regulate their emotions, alleviate anxiety, and restore psychological equilibrium. These strategies encompassed both the active pursuit of positive information and the conscious limitation of exposure to content in order to mitigate the negative emotions elicited by web-based sources. For instance, some participants sought psychological comfort by deliberately reviewing positive content and reading about the recovery experiences of fellow patients.

> *I think the cure rate for breast cancer was relatively low before. I saw online that the survival period is now over five years, and the cure rate has reached several tens of percent. I think the current cure rate is quite high. I think this boosts my confidence… Because I have triple-negative breast cancer. But I*

XSL•FO
RenderX

> *saw that many of my friends with triple-negative cancer, as long as they survive for these five years, basically, they won't have a recurrence. This gives me a lot of confidence.* [P3]

Additionally, some participants maintained emotional stability through self-motivation and psychological self-regulation. They not only regulated their emotions by acquiring information but also reinforced positive feelings and beliefs through intentional behavioral practices.

> *I will browse the internet… I will copy some good sentences to encourage myself. Now, even a doctor shouting slogans has become quite famous on the video platform. On it, I have also learned to shout slogans. This mental victory method can change my inner emotional state.* [P13]

A small number of participants also halted their information retrieval, reduced their search frequency, and avoided specific platforms or negative topics to mitigate the emotional interference caused by web-based information.

> *I've searched before, and I don't do it anymore now. Now… even if I search, it won't be of any use.* [P5]

> *But the more I searched, the more anxious I became, and the more I didn't want to look.* [P1]

For some of the participants, the cell phone was never far away, and they could only reduce searching or reading web-based information by diverting their attention. However, there were also participants who had to passively accept web-based information.

> *I don't even want to search anymore. I just watch some funny videos every day.* [P 2]

> *Anyway, it's just for some gaming… Sometimes, it would come up on TikTok by itself, but I still wanted to check if there were any similar things related to me, like what others use for treatment. Just to see what others are posting on TikTok… Without searching, it just shows up on TikTok.* [P1]

Build a Support Network

Participants sought assistance from others within complex information environments to alleviate their decision-making pressure and emotional burdens.

> *I joined those patient groups. Then I'd listen to what they said and also do some searches on the group. There's a lot of info there. You know, I just wanna learn more about it.* [P17]

Patients often referred to the experiences of fellow patients and sought decision-making advice from doctors or nursing staff, or had family members screen and verify information on their behalf.

> *Seeing this information, one would surely be confused about how to choose this path and which one to take… But my daughter will always help me make the right decision.* [P8]

> *I talked to the doctor about online info. Like, I asked if I could have radiotherapy for bone metastases. He said I didn't need it and should just listen to doctors. Doctors have seen more cases. What I've seen is just one-sided.* [P7]

Acceptance and Adjustment of Expectations

Through information acquisition and practical experience, some participants gradually adjusted their expectations regarding treatment outcomes, shifting their focus from defeating the disease toward managing specific symptoms and maintaining daily functioning.

> *I just checked online, and some info said it's in the advanced stage. Anyway, it's impossible to cure your illness. You gotta live with this cancer…* [P8]

Upon recognizing the gap between their own knowledge and the professional system, participants learned to delegate decision-making authority to professional authorities after thoroughly exploring information.

> *Anyway, when you're in a tough spot trying to make a decision, I always think doctors have more experience than you, right? You've been looking up info for just a few months, but doctors have seen tons of patients. You gotta trust them as much as you can.* [P4]

Discussion

Summary of Results

The findings of this study authentically reflect the web-based information-seeking experiences of patients with breast cancer, which can be summarized under 3 core themes: drivers of information seeking, cognitive judgments amid information fog, and adaptive mechanisms under web-based information pressure. These themes are not isolated categories but rather depict a dynamic psychosocial process: how patients navigate a challenging digital information environment from their initial needs to ultimately develop personalized coping strategies.

Findings

This study found that information-seeking behavior among patients with breast cancer is driven by both emotional and problem-solving needs. The anxiety and uncertainty brought by a diagnosis drive emotional information-seeking centered on seeking reassurance, while simultaneously prompting problem-solving searches to understand the disease and participate in decision-making. Consistent with the classification of web-based help-seeking behaviors of patients with cancer in previous studies, the study by Feng et al [29] mainly focused on web-based health communities and categorized patients' help-seeking behaviors into informational and emotional support needs. This study further expands this perspective by extending the scope of analysis from help-seeking behaviors in web-based community interactions to a broader context of OHIS. Notably, these 2 drivers do not switch off and on simply; they often coexist and reinforce each other. This finding resonates with research by Liu et al [30], based on a dual-factor model, which confirmed that pediatric patients' OHISs are simultaneously influenced by complex factors—both facilitators (such as perceived stress) and inhibitors (such as learned helplessness),

XSL•FO
RenderX

revealing the inherent tension within behavioral drivers. In the Chinese context, this driving force is further moderated by the family system. Phenomena observed in this study, such as spousal discouragement and proxy searching by children, align with Feng's [31] findings that Chinese family communication patterns influence information-seeking behaviors undertaken for family members or oneself. This highlights how, within collectivist cultures, disease management is often viewed as a shared family unit event.

This study reveals patients' cognitive judgment processes within uncertain information environments. Patients develop strategies such as cross-verifying multiple channels, tracing information sources, and assessing applicability to construct personal cognitive safety nets amid information fog. The core paradox of this process lies in the dual role of information: while it may enhance patients' sense of control over their illness, excessive web-based information, poor quality content, and conflicting viewpoints can also trigger decision-making difficulties, diminished trust in information, and persistent anxiety [32,33]. This study particularly highlights the critical role of doctor-patient communication within the information fog. When patients conceal their web-based doubts for fear of challenging medical authority, conflicting information becomes internalized as a psychological burden. This finding strongly corroborates the study by Mahmood et al [34], which demonstrated that patient-centered communication significantly reduces frustration during information-seeking among cancer survivors. It cannot be overlooked that while OHIS has to some extent challenged the traditional paternalistic medical model, its role in clinical practice is increasingly being reinterpreted. Far from threatening physician authority, it may actually enhance the depth and quality of doctor-patient communication [35]. Furthermore, patients' heightened focus on dietary adjustments and nutritional therapies when evaluating information is deeply rooted in local health beliefs and traditional practices. This provides empirical support for moving beyond universal health literacy models and constructing culturally sensitive analytical frameworks.

In contrast to the results of the review by Chen et al [8], our study found that Chinese patients with breast cancer did not use the websites of national or nonprofit cancer organizations to search for health information. Chen et al [8] showed that patients with breast cancer, including those in China, often relied on these websites for information. This disparity may indicate that the patients in this sample have relatively low digital health literacy and have limited access to information through professional web-based channels. Digital health literacy specifically refers to a series of technical abilities and cognitive skills that individuals need to discover, understand, evaluate, and apply web-based health information [36]. The patients with breast cancer in this study's reliance on social media and general search engines may directly indicate their deficiencies in the 2 key aspects of digital literacy, namely identifying the authority of information and conducting effective web-based navigation. In this context, digital health care approaches, such as hospital official apps, web-based patient communities, or intelligent question-answering systems, can provide patients with alternative channels for information acquisition, thereby reducing the obstacles caused by insufficient digital literacy [37,38]. The demand of patients with cancer for digital health functions is closely related to their digital health literacy levels, and there are also differences in the preferences for digital health interventions among patients with different literacy levels. This indicates that digital literacy is an important factor influencing the use of digital nursing tools [39]. Through these platforms, patients can not only access timely and reliable health information but also pose questions, verify information, and receive personalized advice in an interactive environment, which in practice contributes to enhancing their ability to understand and apply information. Meanwhile, research indicates that there are significant disparities in digital health literacy among breast cancer patients. Patients with low literacy levels face greater difficulties in participating in eHealth interventions, highlighting the need to consider patients' literacy levels when designing digital nursing resources [40].

One of our findings is that the web-based information search behavior of patients with breast cancer changes with the disease stage, similar to the results of Madge, but their research focused on patients who had undergone breast cancer surgery [41]. Our participants had a wider range of disease stages, including both patients who had just undergone breast cancer surgery and those with recurrent breast cancer. The phased changes in information search behavior were more significant, suggesting that web-based health information intervention should be time-sensitive.

What is worth noting is that many participants with breast cancer, when searching for web-based health information, not only used traditional search engines and social media but also showed a high preference and trust in AI platforms. As shown in related studies, while traditional web-based resources remain dominant, LLM-based chatbots are becoming a resource for health information for some users, especially those who are younger and have a higher level of trust in AI [42]. However, this AI preference also comes with potential risks. The study found that the web-based health advice provided by AI models contains information misguidance and inaccurate information sources [43]. It can be seen that AI preference may trigger new cognitive trust traps. Therefore, we suggest that when promoting AI-assisted information services, it must be incorporated into the collaborative doctor-patient communication process, and through educational courses, web-based lectures, etc, improve patients' digital health literacy, thereby enhancing the accuracy of information acquisition and decision-making quality for patients [44].

Theoretical Extensions in Web-Based Health Information Contexts

This study's findings align with key processes in the stress-coping theory, thereby providing empirical support for understanding this theory in digital health information contexts. Concurrently, this study theorizes the complex web-based health information environment itself as a significant digital stressor, thereby expanding the application context of stress and coping theories. Findings indicate that for patients, this environment serves not only as a repository of resources for seeking support but may also constitute a persistent source of psychological stress due to information overload, contradictions, and

XSL•FO
RenderX

uncertainties. In essence, an overwhelming or inconsistent volume of information can trigger anxiety and stress responses, potentially leading to avoidance behaviors [45]. Research findings also indicate that information avoidance and reduced retrieval frequency are not merely passive reactions, but rather contextualized emotion regulation strategies. This discovery complements traditional coping classifications' understanding of avoidance strategies. Unlike the original theory's focus on immediate coping, this study found that patients construct meaning around their illness through repeated comparison and integration of information from multiple sources, thereby adjusting their expectations for the future. This demonstrates the long-term and developmental nature of the coping process. It is worth noting that since this theory was not originally designed for digital health information scenarios, some new phenomena related to web-based information conflict and digital uncertainty lack direct conceptual counterparts within the theory. The findings of this study provide new perspectives on the applicability and extension of this theory within digital health information environments.

Limitations

This study has the following limitations. Although most participants came from different cities and prefectures within Hunan Province of China, there are differences in language, economic development, and urban-rural structure within the province. However, the sample size of the qualitative research is limited, and it may still not cover the diversity of digital literacy and socioeconomic backgrounds. Participants were patients from a specialized cancer hospital, and these findings may not fully reflect the experiences or perspectives of health care providers in other medical systems or regions with differing technological infrastructure. Although semistructured in-depth interviews were adopted to obtain data, the respondents' recollections might have a selective bias, and some emotional or behavioral experiences might not be accurately expressed or recalled. At the same time, qualitative research itself has a strong interpretative and subjective nature, and different analysts might form different understandings of the same material. This study was conducted during a treatment stage of patients with breast cancer and did not longitudinally track the web-based information search process and behaviors of patients with breast cancer. Future research could further focus on the differences in information coping patterns among people from different cultures and with different health literacy levels, and also examine how coping strategies evolve dynamically over the course of the treatment from a longitudinal perspective. This study analyzed the research results based on the theory of stress and coping. It provided theoretical support for understanding web-based information-seeking behavior. However, these theories may have limitations in terms of applicability in the specific cultural context of China. Future research could further validate these theories by integrating cultural adaptation theories.

Implications of Research Findings for Policy and Future Research

This study reveals the complex experiences of patients with breast cancer in navigating web-based health information in the digital age. These findings not only deepen our understanding of patients' health information behaviors but also provide significant insights for improving health policies, optimizing clinical practices, and guiding future research directions. The key points are as follows:

1. Promote the construction of a health information support system: Health policy-makers and medical institutions should take the lead in building national or regional official disease information and support platforms. These platforms should integrate evidence-based medical guidelines, treatment pathway descriptions, side-effect management, and rehabilitation knowledge and present them in easily understandable forms such as videos and interactive question-and-answer sessions to reduce information risks.
2. Incorporate patient information behavior management into routine nursing and doctor-patient communication norms: It is recommended that modules for assessing and coaching patients' health information literacy be added to nursing practice guidelines and doctor-patient communication training. Health care providers should proactively inquire about patients' information sources and concerns and provide tailored information prescriptions. They should also guide patients in translating web-based information into safe, actionable self-management behaviors.
3. Conduct digital health literacy education: In the future, health information literacy education courses should be developed for patients with cancer. Special attention should be paid to digitally vulnerable groups such as older people and those with low educational levels, and barrier-free digital health literacy intervention measures should be designed to promote fairness in health information access.
4. Deepen local research and interdisciplinary cooperation: Future research should explore in depth how culture-specific factors shape information behavior patterns and carry out interdisciplinary cooperation among nursing, psychology, information science, and computer science. In the future, intelligent health information question and answer and decision-making assistance tools based on AI and in line with the Chinese cultural context can be developed.

Conclusions

The web-based information-seeking experience of Chinese patients with breast cancer is a dynamic process involving motivation, cognition, and adaptation. Their search behavior stems from dual needs: emotional drive and problem-solving. Within a complex and conflicting information environment, they assess authenticity and applicability while constructing meaning. Through strategies such as information transformation, emotional regulation, support network building, and expectation adjustment, they achieve individual adaptation. Theoretically, this finding enriches the research framework of digital health literacy and information behavior and provides empirical evidence for understanding patients' coping mechanisms in a complex information environment. In practice, this study suggests that clinical and nursing staff should pay attention to patients' emotional needs, cognitive judgment abilities, and adaptation strategies when designing digital health interventions and information support strategies. They should provide multichannel, easy-to-understand, and actionable information resources and combine social support and emotional regulation

XSL•FO
RenderX

guidance to help patients obtain, understand, and apply health information more effectively.

Acknowledgments

The authors extend their gratitude to the participants for sharing their experiences to make this study possible. As a nonnative English-speaking research team, we have made limited use of generative artificial intelligence tools for English language polishing and enhancement to ensure our research findings can be disseminated globally. Beyond this, no artificial intelligence was used in the creation of the remaining content.

Funding

This research was supported by the High-Level Talent Support Program of Hunan Cancer Hospital and the 2025 Hunan Normal University-Hunan Cancer Hospital Joint Research Project (grant: HNUCRI2025010).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

LJH and QCL contributed to conceptualization. LJH and ZD contributed to the data collection. LJH, LXF, and ZD conducted a thematic analysis. QCL contributed to project administration. LJH managed resources. QCL, CJJ, and LXY performed supervision. LJH and ZD contributed to writing the original draft. All authors contributed to reviewing and editing. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Checklist 1
COREQ checklist.
[DOCX File, 23 KB - jmir_v28i1e80363_app1.docx]

References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021 May;71(3):209-249. [doi: 10.3322/caac.21660] [Medline: 33538338]
2. Caetano Dos Santos FL, Michalek IM, Wojciechowska U, Didkowska J. Changes in the survival of patients with breast cancer: Poland, 2000-2019. Breast Cancer Res Treat 2023 Feb;197(3):623-631. [doi: 10.1007/s10549-022-06828-5] [Medline: 36509986]
3. Quang DT, Luong Thi T, Nguyen Di K, Vu Thi Quynh C, Nguyen Thi Hoa H, Phan Ngoc Q. Illuminating the breast cancer survival rates among Southeast Asian women: a systematic review and meta-analysis spanning four decades. Curr Probl Cancer 2024 Feb;48:101062. [doi: 10.1016/j.currproblcancer.2024.101062] [Medline: 38309146]
4. Kimiafar K, Sarbaz M, Shahid Sales S, Esmaeili M, Javame Ghazvini Z. Breast cancer patients' information needs and information-seeking behavior in a developing country. Breast 2016 Aug;28:156-160. [doi: 10.1016/j.breast.2016.05.011] [Medline: 27318171]
5. Sheehy EM, Lehane E, Quinn E, Livingstone V, Redmond HP, Corrigan MA. Information needs of patients with breast cancer at years one, three, and five after diagnosis. Clin Breast Cancer 2018 Dec;18(6):e1269-e1275. [doi: 10.1016/j.clbc.2018.06.007] [Medline: 30153977]
6. Ludwigson A, Huynh V, Vemuru S, et al. Characterizing informational needs and information seeking behavior of patients with breast cancer. Am J Surg 2024 Jan;227:100-105. [doi: 10.1016/j.amjsurg.2023.09.047] [Medline: 37863795]
7. Deng J, Savjani RR, Lee P. Online search behavior for cancer immunotherapy resources and readability analysis: an opportunity to aid in medical decision-making. Cureus 2019 Oct 7;11(10):e5857. [doi: 10.7759/cureus.5857] [Medline: 31763080]
8. Chen J, Duan Y, Xia H, Xiao R, Cai T, Yuan C. Online health information seeking behavior among breast cancer patients and survivors: a scoping review. BMC Womens Health 2025 Jan 3;25(1):1. [doi: 10.1186/s12905-024-03509-x] [Medline: 39754199]
9. Foroughi F, Lam AKY, Lim MSC, Saremi N, Ahmadvand A. "Googling" for cancer: an infodemiological assessment of online search interests in Australia, Canada, New Zealand, the United Kingdom, and the United States. JMIR Cancer 2016 May 4;2(1):e5. [doi: 10.2196/cancer.5212] [Medline: 28410185]

XSL•FO
RenderX

10. Ferraris G, Monzani D, Coppini V, et al. Barriers to and facilitators of online health information-seeking behaviours among cancer patients: a systematic review. Digit Health 2023;9:20552076231210663. [doi: 10.1177/20552076231210663] [Medline: 38107979]
11. Zhao YC, Zhao M, Song S. Online health information seeking among patients with chronic conditions: integrating the health belief model and social support theory. J Med Internet Res 2022 Nov 2;24(11):e42447. [doi: 10.2196/42447] [Medline: 36322124]
12. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. Annu Rev Public Health 2020 Apr 2;41:433-451. [doi: 10.1146/annurev-publhealth-040119-094127] [Medline: 31874069]
13. Southwell BG, Niederdeppe J, Cappella JN, et al. Misinformation as a misunderstood challenge to public health. Am J Prev Med 2019 Aug;57(2):282-285. [doi: 10.1016/j.amepre.2019.03.009] [Medline: 31248741]
14. Lim HM, Wong SS, Yip KC, et al. Online health information-seeking behaviour of patients attending a primary care clinic in Malaysia: a cross-sectional study. Fam Pract 2022 Jan 19;39(1):38-45. [doi: 10.1093/fampra/cmab099] [Medline: 34423368]
15. Lazarus RS, Folkman S. Stress, Appraisal, and Coping: Springer; 1984.
16. Folkman S, Moskowitz JT. Coping: pitfalls and promise. Annu Rev Psychol 2004;55(745-74):745-774. [doi: 10.1146/annurev.psych.55.090902.141456] [Medline: 14744233]
17. Lambert WW, Lazarus RS. Psychological stress and the coping process. Am J Psychol 1970 Dec;83(4):634. [doi: 10.2307/1420698]
18. Mehrabizadeh M, Zaremohzzabieh Z, Zarean M, Ahrari S, Ahmadi AR. Narratives of resilience: understanding Iranian breast cancer survivors through health belief model and stress-coping theory for enhanced interventions. BMC Womens Health 2024 Oct 8;24(1):552. [doi: 10.1186/s12905-024-03383-7] [Medline: 39375669]
19. Ko E, Lee Y. The effects of coping strategies between uncertainty and quality of life of Korean women with gynecological cancer: evaluation of uncertainty in illness theory and stress and coping theory. ANS Adv Nurs Sci 2024;47(3):E84-E95. [doi: 10.1097/ANS.0000000000000507] [Medline: 37498078]
20. Oh YS. Communications with health professionals and psychological distress in family caregivers to cancer patients: a model based on stress-coping theory. Appl Nurs Res 2017 Feb;33(5-9):5-9. [doi: 10.1016/j.apnr.2016.09.008] [Medline: 28096023]
21. Melhem SJ, Nabhani-Gebara S, Kayyali R. Evaluating online health information utilisation and its psychosocial implications among breast cancer survivors: qualitative explorations. Health Promot Perspect 2024 Mar;14(1):61-69. [doi: 10.34172/hpp.42682] [Medline: 38623349]
22. Neubauer BE, Witkop CT, Varpio L. How phenomenology can help us learn from the experiences of others. Perspect Med Educ 2019 Apr;8(2):90-97. [doi: 10.1007/s40037-019-0509-2] [Medline: 30953335]
23. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
24. Guo Y, Xi S, Zhang M. Provincial division of economic zones based on the improved urban gravity model: a case study of Hunan Province, China. PLoS ONE 2021;16(12):e0261205. [doi: 10.1371/journal.pone.0261205]
25. Kerr C, Nixon A, Wild D. Assessing and demonstrating data saturation in qualitative inquiry supporting patient-reported outcomes research. Expert Rev Pharmacoecon Outcomes Res 2010 Jun;10(3):269-281. [doi: 10.1586/erp.10.30] [Medline: 20545592]
26. Braun V, Clarke V, Hayfield N, Terry G. Thematic analysis. In: Liamputtong P, editor. Handbook of Research Methods in Health Social Sciences: Springer Singapore; 2019:843-860. [doi: 10.1007/978-981-10-5251-4_103]
27. Chen HY, Boore JR. Translation and back-translation in qualitative nursing research: methodological review. J Clin Nurs 2010 Jan;19(1-2):234-239. [doi: 10.1111/j.1365-2702.2009.02896.x] [Medline: 19886874]
28. Olmos-Vega FM, Stalmeijer RE, Varpio L, Kahlke R. A practical guide to reflexivity in qualitative research: AMEE guide no. 149. Med Teach 2022 Apr 7:1-11. [doi: 10.1080/0142159X.2022.2057287] [Medline: 35389310]
29. Feng X, Hu Y, Pfaff H, Liu S, Wang H, Wang Y. Emotional or informational? understanding cancer patients' multifaceted help-seeking behaviors in online health communities. Soc Sci Med 2025 Apr;371:117902. [doi: 10.1016/j.socscimed.2025.117902] [Medline: 40056535]
30. Liu F, Kong X, Xia T, Guo H. Exploring the impact of discharged patients' characteristics on online health information-seeking behaviors: insights from patients' dilemmas. Front Psychol 2025;16:1500627. [doi: 10.3389/fpsyg.2025.1500627] [Medline: 40066181]
31. Feng Y. For myself or for others? the influence of family communication patterns on family health history communication and online health information seeking. J Health Commun 2025;30(1-3):72-81. [doi: 10.1080/10810730.2025.2450617] [Medline: 39804582]
32. Wilhelm E, Vivilaki V, Calleja-Agius J, et al. Effects of the modern digital information environment on maternal health care professionals, the role of midwives, and the people in their care: scoping review. J Med Internet Res 2025 Feb 25;27:e70108. [doi: 10.2196/70108] [Medline: 39998875]

XSL•FO
RenderX

33. Zheng H, Chen X, Jiang S, Sun L. How does health information seeking from different online sources trigger cyberchondria? the roles of online information overload and information trust. Inf Process Manag 2023 Jul;60(4):103364. [doi: 10.1016/j.ipm.2023.103364]
34. Mahmood A, Kedia S, Ogunsanmi DO, Kabir U, Entwistle C. Patient-centered communication and cancer information-seeking experiences among cancer survivors: a population-based study in the United States. Patient Educ Couns 2025 Jun;135:108710. [doi: 10.1016/j.pec.2025.108710] [Medline: 40010060]
35. Starcevic V. The impact of online health information seeking on patients, clinicians, and patient-clinician relationship. Psychother Psychosom 2024;93(2):80-84. [doi: 10.1159/000538149] [Medline: 38531342]
36. van der Vaart R, Drossaert C. Development of the digital health literacy instrument: measuring a broad spectrum of health 1.0 and health 2.0 skills. J Med Internet Res 2017 Jan 24;19(1):e27. [doi: 10.2196/jmir.6709] [Medline: 28119275]
37. Lazarou I, Krooupa AM, Nikolopoulos S, et al. Cancer patients' perspectives and requirements of digital health technologies: a scoping literature review. Cancers (Basel) 2024 Jun 21;16(13):2293. [doi: 10.3390/cancers16132293] [Medline: 39001356]
38. Kirsch EP, Kunte SA, Wu KA, et al. Digital health platforms for breast cancer care: a scoping review. J Clin Med 2024 Mar 27;13(7):1937. [doi: 10.3390/jcm13071937] [Medline: 38610702]
39. Yoo SH, Sung JH, Lee K, et al. The needs for digital health and eHeath literacy of cancer patients, caregivers, and healthcare providers: a multicenter, descriptive correlational study. Eur J Oncol Nurs 2024 Jun;70:102581. [doi: 10.1016/j.ejon.2024.102581] [Medline: 38749385]
40. Moon Z, Zuchowski M, Moss-Morris R, Hunter MS, Norton S, Hughes LD. Disparities in access to mobile devices and e-health literacy among breast cancer survivors. Support Care Cancer 2022 Jan;30(1):117-126. [doi: 10.1007/s00520-021-06407-2] [Medline: 34236506]
41. Madge OL, Marinca AM, Daha C, Simion L. Health information seeking behaviour and decision making by patients undergoing breast cancer surgery: a qualitative study. Health Info Libr J 2025 Mar;42(1):42-53. [doi: 10.1111/hir.12480] [Medline: 36861854]
42. Yun HS, Bickmore T. Online health information-seeking in the era of large language models: cross-sectional web-based survey study. J Med Internet Res 2025 Mar 31;27:e68560. [doi: 10.2196/68560] [Medline: 40163112]
43. Yau JYS, Saadat S, Hsu E, et al. Accuracy of prospective assessments of 4 large language model chatbot responses to patient questions about emergency care: experimental comparative study. J Med Internet Res 2024 Nov 4;26:e60291. [doi: 10.2196/60291] [Medline: 39496149]
44. Álvarez-Pérez Y, Duarte-Díaz A, Toledo-Chávarri A, et al. Digital health literacy and person-centred care: co-creation of a massive open online course for women with breast cancer. Int J Environ Res Public Health 2023 Feb 22;20(5):3922. [doi: 10.3390/ijerph20053922] [Medline: 36900935]
45. Soroya SH, Farooq A, Mahmood K, Isoaho J, Zara SE. From information seeking to information avoidance: understanding the health information behavior during a global health crisis. Inf Process Manag 2021 Mar;58(2):102440. [doi: 10.1016/j.ipm.2020.102440] [Medline: 33281273]

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research Checklist
OHIS: online health information search

Edited by A Stone; submitted 09.Jul.2025; peer-reviewed by M Rosenlund, R Grasso; accepted 30.Jan.2026; published 26.Mar.2026.

Please cite as:
Liu J, Li X, Li X, Chen J, Zhang D, Qiu C
From Searching to Coping, How Chinese Patients With Breast Cancer Navigate Web-Based Health Information: Semistructured Interview Study
J Med Internet Res 2026;28:e80363
URL: https://www.jmir.org/2026/1/e80363
doi:10.2196/80363

©Jiahui Liu, Xingfeng Li, Xuying Li, Jiejun Chen, Dan Zhang, Cuiling Qiu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Sociodemographic Drivers of Recruitment and Attrition in Digital Neurological Research: Longitudinal Cohort Study

Peyman Nejat[1,2], MD; Ashley D Bachman[3], BS; Vicki M Stubbs[3], BS; Joseph R Duffy[3], PhD; John L Stricker[4], MS, PhD; Vitaly Herasevich[1,2], MD, PhD; David T Jones[3], MD; Rene L Utianski[3], PhD; Hugo Botha[2,3], MBChB

[1]Department of Anesthesiology and Perioperative Medicine, Mayo Clinic, Rochester, MN, United States

[2]Office of Digital Innovation, Center for Clinical And Translational Science, Mayo Clinic, Rochester, MN, United States

[3]Department of Neurology, Mayo Clinic, 200 First St SW, Rochester, MN, United States

[4]Department of Information Technologies, Mayo Clinic, Rochester, MN, United States

Corresponding Author:
Hugo Botha, MBChB
Office of Digital Innovation, Center for Clinical And Translational Science, Mayo Clinic, Rochester, MN, United States

Abstract

Background: Digital recruitment methods offer opportunities to address challenges in clinical research participation, particularly in neurology. However, the impact of digital approaches across socioeconomic and demographic groups remains inadequately understood.

Objective: This study investigates the influence of sociodemographic factors on recruitment and attrition in a remote neurological research cohort, mapping participation pathways and identifying disparities to inform inclusive digital strategies.

Methods: We conducted a nonexperimental, observational longitudinal cohort study at Mayo Clinic using patient-portal invitations between March and July 2024 as part of a remote speech capture study. Eligibility criteria included age 18 years and older, US residence, and English proficiency. Of 5846 invited patients, progression was tracked across checkpoints (invitation, eligibility screening, electronic consent, and task completion) using Epic (Epic Systems Corporation) to obtain demographic information, Qualtrics (Qualtrics, LLC) for screening, PTrax (a Mayo Clinic–developed Participant Tracking System) for consent tracking, and the recording platform. Socioeconomic context was assessed using the Housing-based Socioeconomic Status (HOUSES) index, where higher values indicate higher socioeconomic status, and the Area Deprivation Index (ADI), where higher values reflect greater neighborhood disadvantage. Data diagnostics included Anderson-Darling tests for non-normality and Little missing completely at random (MCAR) test to characterize missingness. Associations between participation outcomes and age, sex, urbanicity, and socioeconomic indices were examined using nonparametric tests. Exact *P* values and 95% CIs are reported. Analyses were conducted using BlueSky Statistics (BlueSky Statistics, LLC) and the Python *SciPy* package.

Results: Overall, 415 out of 5846 participants (7.1%) completed all study requirements. Completers were older (median age 66.4, IQR 56.0-72.5; 95% CI 65.1 - 67.6 years) than noncompleters (median age 62.8, IQR 47.5-72.7; 95% CI 62.2 - 63.2 years; *P*<.001). Participants from more socioeconomically disadvantaged neighborhoods were less likely to respond (invitation nonresponder median ADI 45.0, IQR 29.0-63.0 vs interested median ADI 42.0, IQR 27.0-59.0; *P*<.001), and completers had slightly lower ADI ranks than noncompleters (median 41.0, IQR 27.0-56.0 vs median 44.5, IQR 28.0-62.0; *P*=.04). Urban participants enrolled faster (median 32.0, IQR 9.0-58.0; 95% CI 31.0 - 37.0 days) than rural (median 41.0, IQR 22.0-65.0; 95% CI 37.0 - 49.0 days; *P*=.01). Female participants responded slower (median 38.5, IQR 14.8-66.3; 95% CI 35.0 - 41.0 days) than males (median 32.0, IQR 8.0-57.5; 95% CI 29.0 - 38.0 days; *P*=.01). No significant differences were observed for the HOUSES index, and device type was unrelated to completion or timelines. Missingness for key variables was completely at random (MCAR χ^2_3=3.45; *P*=.24).

Conclusions: Digital recruitment does not overcome traditional barriers to participation and may introduce new disparities related to age, urbanicity, and neighborhood disadvantage. These findings inform inclusive digital research strategies, including multichannel outreach, age-specific engagement, and rural technical support. This study applies longitudinal pathway analysis to digital neurology recruitment, offering actionable insights for improving inclusivity in remote research.

(J Med Internet Res 2026;28:e83432) doi:10.2196/83432

KEYWORDS

digital recruitment; neurological research; participation disparities; sociodemographic factors; digital divide

XSL•FO
RenderX

Introduction

Participant recruitment is a critical component of medical research studies, and the rate of participation varies significantly depending on the type of study, the setting, and the population involved. As an example, recent data suggest that enrollment rate in cancer trials is around 6.3% to 7.1% [1,2]. Enrollment rates may also differ among different populations; specifically, they can be lower among minority, pediatric, and geriatric populations [3,4]. Poor recruitment often results in underpowered studies, which lack the necessary sample size to detect meaningful differences between distinct groups. This can lead to statistically nonsignificant results even when there are clinically relevant effects [5]. Despite these efforts and the growing adoption of digital recruitment tools, recent analyses confirm that these challenges persist with participation patterns continuing to vary by age, sex, and socioeconomic status (SES) [6]. This underscores a critical point that digital methods may not fully mitigate traditional barriers and could introduce new inequities that influence who participates in research.

Digital tools, including social media platforms, mobile apps, electronic health records, patient portals, and electronic consent, have expanded the reach of research recruitment efforts by enabling targeted outreach to specific demographics and geographic areas [7-10]. These approaches can engage populations previously hard to reach through traditional methods, but evaluations report inconsistent effects on diversity and efficiency across settings [11]. A recent systematic review cataloged the spectrum of digital technologies deployed for recruitment and highlighted the still-limited evidence that any one approach reliably improves inclusion of underrepresented groups [6].

A central challenge for digital recruitment is the digital divide, defined as persistent inequities in broadband access, device availability, and digital literacy that map closely onto socioeconomic, geographic, and age-related lines [12]. These inequities can depress response rates and create systematic attrition in specific subgroups even when outreach is delivered digitally. Recent studies and reviews caution that digitalization can reproduce or widen participation gaps if considerations are not built into recruitment strategies [13]. Within health care systems, portal-based recruitment shows promise but exhibits differential response patterns across patient characteristics, underscoring the need to track impacts on disparity as programs scale. Moreover, scoping reviews in specific domains suggest that mixed, multichannel recruitment strategies (digital and offline) may improve inclusivity compared with relying on a single modality [11]. These disparities are often more pronounced among disadvantaged groups, including racial and ethnic minorities and older adults, who may have limited access to the internet and lower digital literacy levels [14-16]. For instance, studies have shown that older adults and African American patients are less likely to use digital health portals compared to their younger and White counterparts, highlighting a significant gap in technology use [16]. Here, the term “White” reflects the racial classification used in the original study, which reported race and ethnicity separately and did not provide ethnicity-specific breakdowns. Similarly, individuals from lower SES neighborhoods often have reduced access to the internet and lower health literacy, which can hinder their ability to engage with digital health technologies effectively [15]. While digital tools have the potential to improve the representativeness of trial participants, there is limited evidence supporting their effectiveness in recruiting underrepresented groups [8,17]. This underscores the need for targeted interventions and strategies to bridge the digital divide and ensure equitable access to digital health resources [18,19].

Digital recruitment has become especially popular with the increased interest in artificial intelligence (AI) in health care, which requires large and representative datasets to train. For example, the Bridge2AI-Voice program has the explicit goal to create “an ethically sourced flagship dataset to enable future research in artificial intelligence” [20]. Such datasets are typically assembled through digital recruitment workflows, underscoring the need to understand how these methods shape participant diversity. Our group has similarly endeavored to obtain a large bank of speech recordings focused on neurological disorders, with the goal of subsequently using these data to train AI models, through a primarily digital recruitment approach. Given concerns about the digital divide and its potential impact on representativeness, the broader speech capture study offered the ideal setting to formally investigate these issues in depth. This study aims to characterize recruitment and attrition patterns in a remote neurology cohort, quantify associations with sociodemographic factors (age, sex, neighborhood deprivation, housing-based SES, and urbanicity), and identify drop-off points to inform strategies for digital recruitment for all participants.

Methods

Study Design and Setting

This analysis was nonexperimental and observational, using a longitudinal cohort design to evaluate sociodemographic factors influencing recruitment and attrition in a remote speech capture study. This study was conducted in accordance with the American Psychological Association (APA) Journal Article Reporting Standards (JARS; American Psychological Association, 2018; refer to Multimedia Appendix 1 for the completed JARS-Quant checklist) [21]. The overarching speech capture study aims to remotely collect speech samples from patients with neurologic diseases to develop an easy-to-use and cost-effective screening tool for predicting disease progression. The parent speech capture study extends beyond the March-July 2024 analysis window; for this report, we used data collected within that period. The research was conducted at Mayo Clinic.

Inclusion and Exclusion Criteria

Individuals were eligible if they met the following criteria: (1) adults aged 18 years or older, (2) residing in the United States, and (3) able to communicate in English via spoken language.

Data Collection and Participant Characteristics

Patient identification was conducted using the Mayo Clinic Electronic Health Record (Epic, developed by Epic Systems Corporation), and invitations to complete an eligibility survey were sent via the patient portal using Qualtrics (developed by Qualtrics, LLC) [22]. The eligibility survey also assessed

XSL•FO
RenderX

participants' understanding of the study and asked whether they had a legally authorized representative responsible for financial or health care decisions. Once interest and eligibility were confirmed, participants received a PDF of the consent form to sign electronically via AdobeSign (Adobe Inc) through Mayo Clinic–developed Participant Tracking System (PTrax). PTrax is an institutional research software program designed to streamline informed consent processes, manage participant status, track enrollments and accruals, and provide reporting and analytics. Recruitment followed a convenience sampling approach, targeting patients with upcoming neurology appointments accessible via the institutional patient portal. After consent was obtained, participants were sent a secure message in the patient portal with a link to the speech recording platform and instructions.

We exported longitudinal record data for all eligible patients invited to the speech capture study between March and July 2024. This resulted in a total sample of 5846 participants, reflecting all patients meeting the inclusion criteria during the recruitment window. Race was extracted from the electronic health record, where "White" reflects a race category and is distinct from ethnicity. Ethnicity (eg, Hispanic or Latinx origin) was not included in the dataset for this secondary analysis; therefore, individuals categorized as "White" may include participants of diverse ethnic backgrounds. The invited cohort had a median (IQR) age of 63 (48-72) years, with 56.2% (3283/5846) female and 93.7% (5478/5846) identifying as White. Urbanicity distribution was 56.5% (3303/5846) urban, 23.3% (1363/5846) rural, and 20.2% (1180/5846) urban cluster. This secondary analysis included all eligible patients invited during the recruitment window of the overarching speech capture study. No a priori power calculation was performed because the analysis was observational and descriptive. Precision was conveyed using 95% CIs for medians and IQRs. Data were drawn from Epic (demographics), Qualtrics (survey responses), PTrax (consent tracking), and the speech recording platform (task completion).

Measures and Covariates

SES was assessed using the Housing-based Socioeconomic Status (HOUSES) index and the Area Deprivation Index (ADI) national rank. The HOUSES index is a practical and adaptable tool for assessing SES using housing data. It effectively correlates with traditional SES measures and predicts various health outcomes [23]. Higher HOUSES index values indicate higher SES, while lower values indicate lower SES. The ADI national rank measures neighborhood socioeconomic disadvantage based on factors such as income, education, employment, and housing characteristics [24,25]. Higher scores indicate greater disadvantage. ADI is widely used in public health research to assess the impact of socioeconomic context on health outcomes.

Primary outcomes included study completion and time to enrollment. Exposures and predictors included age, sex, urbanicity (urban, rural, or urban cluster), ADI national rank, and the HOUSES index. Potential confounders considered were device type and urbanicity. Urbanicity was treated as an exposure in primary analyses and as a potential confounder in models examining associations between socioeconomic indices and participation outcomes. The collected data included participants' age, sex assigned at birth (hereafter referred to as sex), date of invitation, recruitment process checkpoints and dates, and devices used for speech recording. Residence information was used to measure the HOUSES and ADI index and to classify participants by urbanicity (urban, rural, or urban cluster).

Quality of Measurements

This analysis relied exclusively on data exported from established operational systems (Epic for demographics, Qualtrics for eligibility responses, PTrax for participant status tracking, and the recording platform for task completion). No new training, instrumentation, or measurement procedures were implemented specifically for this analysis. All timestamps and statuses were generated by the source systems as part of routine workflows. Data were deidentified before analysis.

Analytic Strategy

The longitudinal time series data for each patient undergoing the participation process were standardized to align with specific checkpoints, as demonstrated in Figure 1. At each step, it was possible for patients not to respond to the research coordination team, which was defined as the "No Response" stage. After providing consent, participants were also free to withdraw it at any time. While most individuals followed one of the typical pathways depicted in Figure 1, there were 83 cases (1.4% of the total 5846 cases) in which participants deviated from these paths, necessitating intervention by research coordinators. These atypical cases were primarily attributed to personal circumstances or the involvement of a legally authorized representative.

XSL•FO
RenderX

Figure 1. Simplified view of the participation enrollment pathways in a longitudinal digital neurology research study. PTrax: Participant Tracking System.

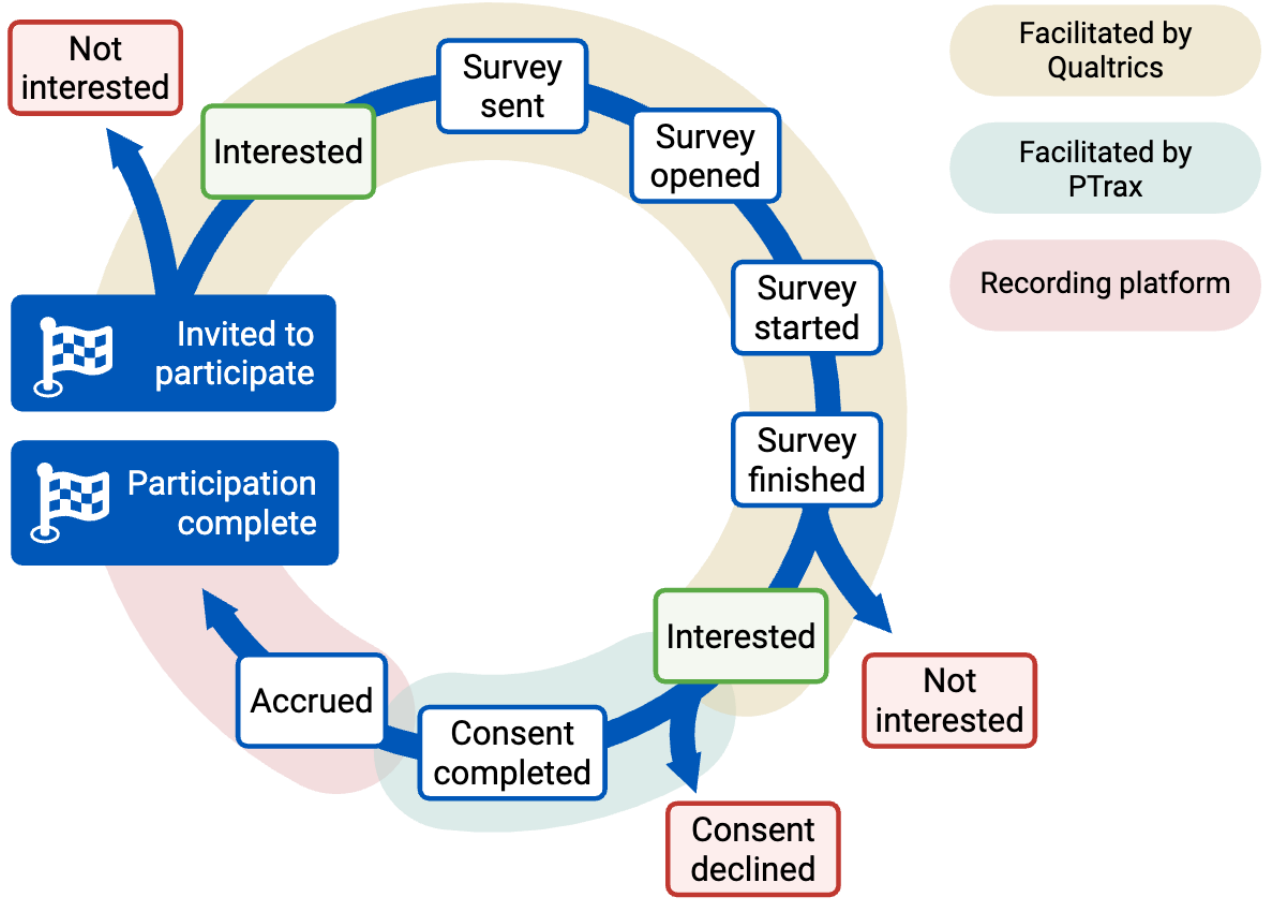

Kruskal-Wallis and Wilcoxon rank-sum tests were used to compare the median age, socioeconomic indices, and time taken to reach different steps of the study. These nonparametric tests were selected due to non-normal distribution of key variables, as confirmed by the Anderson-Darling test (P<.001). They are appropriate for comparing medians across groups and handling skewed data. Analysis was conducted at 2 levels: at each checkpoint and through an end-to-end investigation of participants who completed the study. For each path, a participant could take at each checkpoint, age, ADI national rank, and HOUSES index were compared to identify statistically significant differences, using a 2-sided alpha level of .05. At the end-to-end level, additional comparisons were made across sex, urbanicity (urban, rural, and urban cluster), ADI national rank, and HOUSES index to evaluate differences in time taken to complete participation, from initial invitation to accrual, and whether participants completed the study, regardless of the path taken. Analyses were conducted using BlueSky Statistics version 10.3.4 (developed by BlueSky Statistics, LLC) and Python *SciPy* package version 1.16. Exact P values are reported, with P<.001 used where appropriate. No effect sizes were calculated for analyses.

Data Diagnostics

Since both Kruskal-Wallis and Wilcoxon rank-sum tests require at least 5 data points in each comparison category, pathways with fewer than 5 participants without missing values were excluded from the analysis. This threshold was applied to ensure statistical validity and avoid unreliable comparisons in small subgroups, as no randomized assignments or masking strategies were implemented. To further assess the nature of missing data across key sociodemographic variables, including age, ADI national rank, and HOUSES index, we conducted Little's missing completely at random (MCAR) test [26]. This test evaluates whether data are MCAR, which informs the appropriate handling strategy. Based on the results, we adopted a pairwise deletion approach for statistical analyses, allowing each test to include all available cases for the specific variable of interest. This method was chosen to preserve sample size and maintain statistical power, particularly in subgroups with limited data. This method was chosen to preserve sample size and maintain statistical power, particularly in subgroups with limited data. No outlier removal or variable transformations were applied.

Ethical Considerations

The overarching speech capture study was reviewed and approved by the Mayo Clinic Institutional Review Board (#22-002430). Informed consent was obtained electronically for the primary speech capture study via Adobe Sign. For this secondary analysis of recruitment and attrition patterns, which used deidentified data from the primary study, the institutional review board determined that additional approval was not required. All data were deidentified prior to analysis to ensure participant confidentiality. No compensation was provided for

XSL•FO
RenderX

participation. No identifiable images of participants are included in the manuscript or any supplementary materials.

Results

Overview of the Invited Cohort

A total of 5846 patients were invited to participate in the study between March and July 2024. Of these participants, 3283 (56.2%) were female, 2560 (43.8%) were male, and 3 (0.1%) were unknown. Most participants (5478/5846, 93.7%) identified as White. The age distribution ranged from 18 to 96 years, with a median (IQR) age of 63 (48-72) years (N=5846; 95% CI 62 - 63). The narrow CI suggests high precision in estimating the median age of the invited cohort. Regarding urbanicity, 56.5% (3303/5846) of invited participants resided in urban areas, 23.3% (1363/5846) in rural areas, and 20.2% (1180/5846) in urban clusters (Table 1). Following accrual completion, participants used various devices to access the recording platform. Apple-based mobile devices (iPhone and iPad) were most frequently used (141/415, 34.0%), followed by Windows-based computers (134/415, 32.3%), Apple-based computers (82/415, 19.8%), Android-based mobile devices (52/415, 12.5%), and other devices (6/415, 1.4%).

Table . Demographic characteristics of 5846 patients invited to participate in a remote speech capture study for neurological research at Mayo Clinic between March and July 2024.

Characteristic	Participants, n (%)
Sex	
Female	3283 (56.2)
Male	2560 (43.8)
Unknown	3 (0.1)
Race	
White	5478 (93.7)
Black or African American	97 (1.7)
Choose not to disclose	65 (1.1)
Other	206 (3.5)
Age (years)	
18 - 33	569 (9.7)
34 - 49	999 (17.1)
50 - 64	1607 (27.5)
65 - 80	2203 (37.7)
≥81	468 (8.0)
Population	
Urban area	3303 (56.5)
Rural area	1363 (23.3)
Urban cluster	1180 (20.2)
Device used for participation	
Apple-based mobile device	141 (34.0)
Windows-based computer	134 (32.3)
Apple-based computer	82 (19.8)
Android-based mobile device	52 (12.5)
Other devices	6 (1.4)

The ADI national rank of participants spanned the entire range from 1 to 100, with a median (IQR) of 44 (28-61; n=5403; 95% CI 43 - 45), indicating representation across diverse socioeconomic backgrounds. Similarly, the HOUSES index percentile ranged from 1 to 100, with a median (IQR) of 70 (43-88; n=5439; 95% CI 69 - 71), demonstrating considerable variability in housing conditions among participants. The tight CI reflects reliable estimation despite the wide IQR, which indicates heterogeneity in housing-based SES. Neither age, ADI national rank, nor HOUSES index exhibited normal distribution according to the Anderson-Darling test (P<.001 for all variables), confirming the appropriateness of nonparametric statistical methods used in subsequent analyses. Figure 2 illustrates the distribution of the HOUSES index percentile and ADI national rank, with panel A showing ADI national ranks and panel B showing HOUSES index percentiles.

XSL•FO
RenderX

Figure 2. Distribution of socioeconomic measures among participants in a remote speech capture study for neurological research. ADI: Area Deprivation Index. HOUSES: Housing-based Socioeconomic Status.

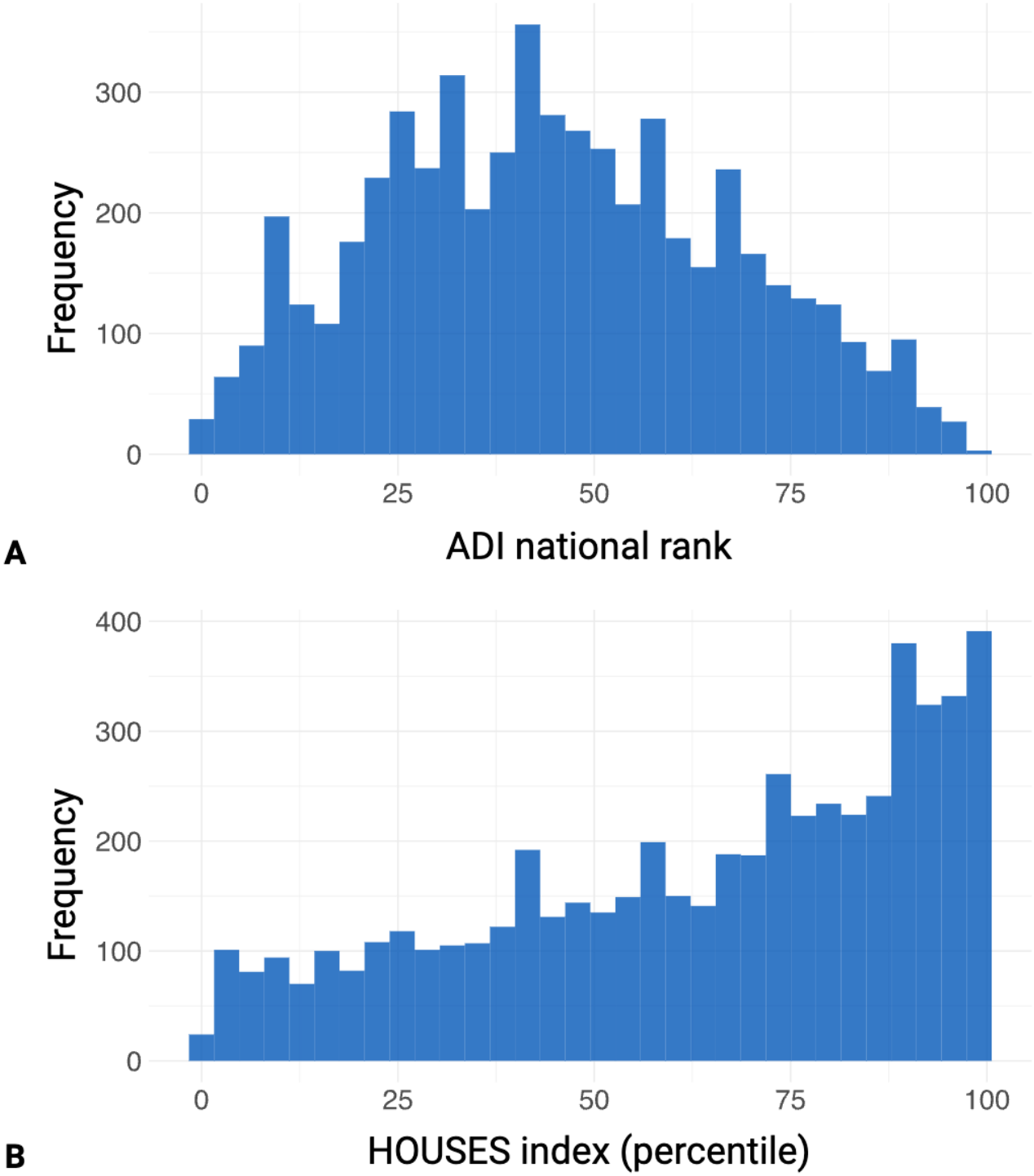

To assess the nature of missing data, Little MCAR test was conducted on age, ADI national rank, and HOUSES index and yielded χ^2_3=3.45; *P*=.24. This nonsignificant result indicates that missingness was not systematically related to other variables, supporting the assumption that data were MCAR. Accordingly, pairwise deletion was applied in subsequent analyses, a strategy chosen to preserve sample size and maintain statistical power while ensuring valid comparisons between study completers and noncompleters.

Comprehensive Participation Analysis

Significant differences were observed between participants who completed the study and those who did not across several demographic and socioeconomic factors. Age was higher among completers (median 66.4, IQR 56.0-72.5; 95% CI 65.1 - 67.6 years) compared to noncompleters (median 62.8, IQR 47.5-72.7; 95% CI 62.2 - 63.2 years; *P*<.001), suggesting that older individuals were more likely to participate (Table 2). The narrow CIs for both groups indicate high precision in estimating age differences, reinforcing the robustness of this finding.

XSL•FO
RenderX

Participants who completed the study also resided in slightly less socioeconomically disadvantaged areas, as indicated by lower ADI national ranks (median 41.0, IQR 27.0-56.0; 95% CI 39.0 - 43.0 vs median 44.5, IQR 28.0-62.0; 95% CI 43.0 - 45.0; *P*=.04). Although the difference is statistically significant, the overlapping CIs suggest that the magnitude of this effect is modest and should be interpreted cautiously. No significant differences were found in HOUSES percentiles, indicating that individual HOUSES may exert less influence on participation compared to neighborhood-level disadvantage.

Table . Comparison of socioeconomic and demographic characteristics between participants who completed the remote speech capture and those who did not. The Wilcoxon rank-sum test was applied to assess differences between groups.

Variable and group	Number of participants	Median (IQR)	95% CI	*P* value
Age (years)				<.001
Participation not complete	5431	62.8 (47.5 - 72.7)	62.2 - 63.2	
Participation complete	415	66.4 (56.0 - 72.5)	65.1 - 67.6	
ADI[a] national rank				.04
Participation not complete	5018	44.5 (28.0 - 62.0)	43.0 - 45.0	
Participation complete	385	41.0 (27.0 - 56.0)	39.0 - 43.0	
HOUSES[b] percentile				.76
Participation not complete	5052	70.0 (42.0 - 88.0)	68.0 - 71.0	
Participation complete	387	71.0 (44.0 - 88.0)	66.0 - 75.0	

[a]ADI: Area Deprivation Index.

[b]HOUSES: Housing-based Socioeconomic Status.

Sex differences in enrollment time were significant, with female participants taking longer to enroll than males (median 38.5, IQR 14.8-66.3; 95% CI 35.0 - 41.0 vs median 32.0, IQR 8.0-57.5; 95% CI 29.0 - 38.0 days; *P*=.01; Table 3). The relatively narrow CIs for both groups indicate precise estimates of enrollment time differences, reinforcing the statistical significance of this finding. Additionally, urbanicity influenced the time to complete enrollment. Participants from urban areas enrolled more quickly than those from rural or urban cluster regions (median 32.0, IQR 9.0-58.0; 95% CI 31.0 - 37.0 vs median 41.0, IQR 22.0-65.0; 95% CI 37.0 - 49.0 days and median 40.0, IQR 13.0-71.0; 95% CI 33.0 - 49.0, respectively; *P*=.01). The wider CIs for rural and urban cluster groups suggest greater variability in enrollment time compared to urban participants, possibly reflecting differences in access or engagement. However, no significant differences were found in completion time across sex or urbanicity. No significant associations were detected between HOUSES indices and study completion. Additionally, no significant relationship was found between the device type used for task completion and the completion or participation time. These null findings suggest that SES and device type did not influence completion dynamics in this cohort.

XSL•FO
RenderX

Table . Enrollment and completion times stratified by urbanicity and sex among participants in a remote speech capture study for neurological research. Statistical significance was assessed using the Wilcoxon rank-sum test or the Kruskal-Wallis test, depending on the category count.

Variable	Number of participants	Median (IQR)	95% CI	*P* value
Enrollment time (days)				.01
Rural	150	41.0 (22.0 - 65.0)	37.0 - 49.0	
Urban area	532	32.0 (9.0 - 58.0)	31.0 - 37.0	
Urban cluster	141	40.0 (13.0 - 71.0)	33.0 - 49.0	
Completion time (days)				.70
Rural	74	20.0 (12.3 - 36.0)	15.5 - 22.0	
Urban area	269	20.0 (10.0 - 32.0)	17.0 - 22.0	
Urban cluster	72	19.5 (7.0 - 38.8)	12.5 - 27.0	
Enrollment time (days)				.01
Male	343	32.0 (8.0 - 57.5)	29.0 - 38.0	
Female	480	38.5 (14.8 - 66.3)	35.0 - 41.0	
Completion time (days)				.95
Male	181	20.0 (10.0 - 32.0)	16.0 - 22.0	
Female	234	20.0 (9.0 - 35.0)	15.5 - 22.0	

Step-by-Step Participation Analysis

The majority of invited participants either did not read or did not respond to the initial invitation via Epic (n=2736) or expressed no interest (n=1752). Among the 1358 participants who initially expressed interest, 415 (30.6%) ultimately completed the study in its entirety. Throughout various stages of the recruitment process, a total of 3346 participants failed to respond to follow-up communications from the research coordination team.

Analysis of participant age across different pathways revealed that individuals who did not respond to the invitation or eligibility check were significantly younger than those who proceeded toward study completion. This age disparity contributed to an increase in the median age of participants completing the study (66.4, IQR 56.0-72.5; 95% CI 65.1 - 67.6) years compared to the overall invited cohort (62.8, IQR 47.5-72.7; 95% CI 62.2 - 63.2) years. The narrow CIs for both estimates indicate high precision, reinforcing confidence in the observed age-related attrition pattern. A similar pattern was observed among the 95 participants who withdrew consent after initially providing it but before being accrued for the recording session, with these participants having a median age of 55.8 years compared to 66.3 years for those who continued with the recording. This suggests that younger participants were disproportionately represented among early dropouts.

While no significant differences in the HOUSES index were observed across different participation pathways, participants who did not respond to the initial invitation had significantly higher ADI national ranks (median 45.0, IQR 29.0-63.0; 95% CI 44.0 - 46.0) compared to those who expressed interest (median 42.0, IQR 27.0-59.0; 95% CI 39.0 - 43.0), indicating residence in more socioeconomically disadvantaged neighborhoods. The narrow CIs for these ADI estimates suggest precise measurement of this disparity, underscoring the influence of neighborhood-level socioeconomic disadvantage on initial engagement. No other significant differences in ADI national ranks were observed across subsequent recruitment steps. Figure 3 details the participant flow and checkpoint-specific distributions.

XSL•FO
RenderX

Figure 3. Step-by-step analysis of participant recruitment and attrition in a longitudinal digital neurology research study conducted at Mayo Clinic between March and July 2024. The figure illustrates participant flow across recruitment checkpoints, with descriptive statistics for Area Deprivation Index (ADI), age, and Housing-based Socioeconomic Status (HOUSES) index reported for each pathway as median (IQR; 95% CI), arranged from top to bottom. Significant differences, based on a significance level of .05, are highlighted in green. Pie charts indicate the relative frequencies of participants at each stage. Paths with fewer than 5 participants are excluded from the analysis. LAR: Legally Authorized Representative.

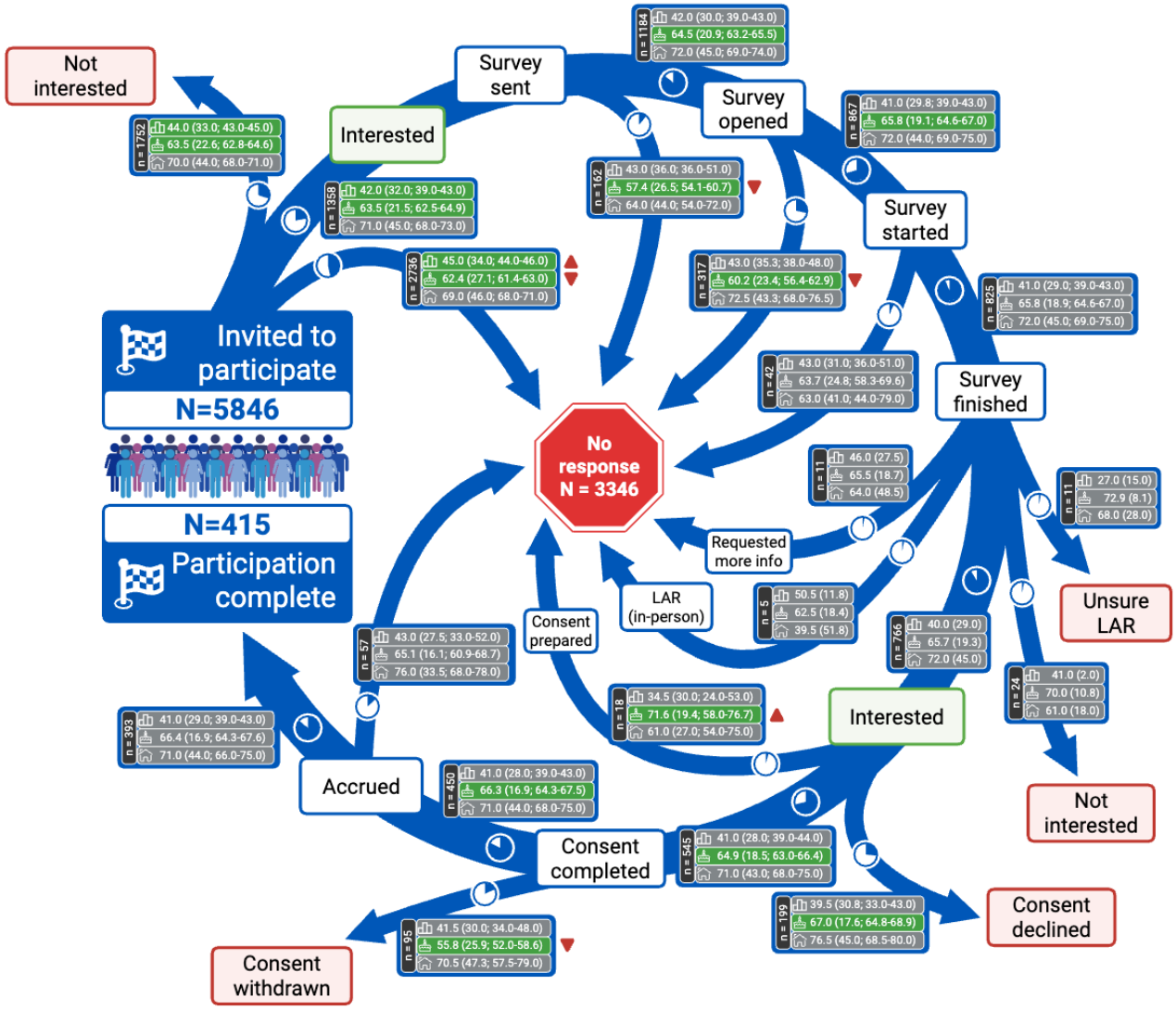

Discussion

Summary of Principal Findings

This study provides a comprehensive analysis of participant recruitment pathways in a digital speech research study, revealing important associations between sociodemographic factors and participation outcomes. Understanding these pathways is critical for improving equity and efficiency in digital clinical research [17]. First, recruitment remains a fundamental challenge in clinical research, with participation rates often below 10 percent in specialized studies, limiting generalizability and slowing innovation [1,2]. Second, as digital recruitment methods become increasingly prevalent, it is essential to assess whether these approaches truly reduce barriers or inadvertently perpetuate existing disparities, an area where evidence remains limited [8]. By analyzing longitudinal recruitment data from 5846 invited patients, our findings demonstrate that systematic pathway analysis can uncover patterns of participation and attrition that may not be apparent when examining only final enrollment outcomes. These results support our primary hypothesis that sociodemographic factors, including age, urbanicity, and neighborhood disadvantage, are associated with recruitment and attrition in digital research workflows. This finding challenges the assumption that digital methods inherently improve inclusivity and highlights the need for targeted strategies, such as age-specific engagement and rural digital support, to promote retention in remote research.

Age-Related Participation Patterns

The significant age differences observed at various dropout points suggest that digital recruitment methods may be less effective for younger populations. Our analysis showed that the median age increased from 63 (IQR 48-72; 95% CI 62 - 63) years in the invited cohort to 66.4 (IQR 56.0-72.5; 95% CI 65.1 - 67.6) years among completers, indicating a robust and precise trend toward older participant retention. This finding challenges the conventional wisdom that digital methods inherently appeal to younger participants and suggests that age-specific engagement strategies may be necessary throughout

XSL•FO
RenderX

the recruitment process. While younger individuals may be more comfortable with technology, our results and others suggest that younger participants are more likely to disengage or drop out of digital studies over time [27,28]. Possible explanations include competing priorities, lower perceived relevance of neurological research, and reduced tolerance for multistep enrollment processes [29]. This emphasizes the importance of age-sensitive retention strategies that extend beyond initial recruitment, including tailored messaging and incentives to support long-term engagement.

Disparities by Urbanicity in Enrollment Timing

The observation that participants from urban areas completed enrollment significantly faster than those from rural areas or urban clusters highlights potential disparities by urbanicity in research accessibility. Urban participants enrolled in a median of 32.0 (IQR 9.0-58.0; 95% CI 31.0 - 37.0) days compared to 41.0 (IQR 22.0-65.0; 95% CI 37.0 - 49.0) days for rural and 40.0 (IQR 13.0-71.0; 95% CI 33.0 - 49.0) days for urban clusters, underscoring a consistent and precise difference. This finding aligns with broader concerns about the urban-rural digital divide and suggests that digital recruitment, while theoretically boundaryless, may still be influenced by infrastructure, digital literacy, or health care engagement [30,31]. Rural participants may face slower internet speeds, limited device availability, and less familiarity with patient portals, which could delay enrollment [32]. This contrasts with some findings in telehealth adoption, which suggest that technology can overcome location-based care barriers [33]. To mitigate these disparities, future strategies should include technical support for rural participants, alternative enrollment options (eg, phone-based consent), and targeted outreach through local health care networks.

Socioeconomic Disadvantage and Participation

Perhaps most notably, our analysis revealed that participants from neighborhoods with higher socioeconomic disadvantage (higher ADI national ranks) were significantly less likely to respond to initial invitations. This may reflect limited broadband access, lower digital literacy, and competing priorities in disadvantaged neighborhoods [34]. This finding is consistent with prior research demonstrating that individuals from more disadvantaged areas experience greater barriers to engaging with digital health, including higher no-show rates and lower uptake of telehealth services [35,36]. Our study adds nuance by comparing neighborhood-level (ADI) and housing-based (HOUSES) SES measures, an approach rarely examined in digital recruitment literature. This finding suggests that digital recruitment methods may perpetuate existing socioeconomic disparities in research participation if not specifically designed to address these barriers. To mitigate these barriers, recruitment strategies should include targeted outreach in areas with high ADI, provide technical support, and incorporate offline options. The absence of significant differences in the HOUSES index across participation pathways, despite differences in the ADI national rank, suggests that neighborhood context likely influences digital engagement more than individual housing characteristics because infrastructure and community resources shape access and literacy.

Sex Differences in Enrollment Dynamics

Females took longer to complete the enrollment process than males, a difference that warrants further investigation. This may reflect differences in time availability, competing responsibilities, or engagement with digital health platforms that could impact recruitment strategies. Prior studies suggest that social and structural factors, such as unequal distribution of caregiving and household responsibilities, affect time availability for research participation, particularly in remote studies [37]. Recruitment workflows could incorporate flexible scheduling and simplified enrollment steps to reduce time burden. While underexplored in the current literature, these insights underscore the need for sex- and gender-aware design in digital recruitment, which may improve inclusivity and reduce attrition.

Recruitment Funnel Attrition

Of 5846 individuals invited, a large proportion did not respond or declined participation, and only 415 of all the invited individuals (7.1%) completed the study. This significant attrition, consistent with patterns observed in other digital recruitment efforts [38,39], may reflect perceived complexity of enrollment, lack of immediate incentives, and competing priorities among participants. These findings highlight the need for iterative, multitouch recruitment strategies that re-engage potential participants and address barriers to active enrollment. Such strategies could include reminder messages, simplified consent processes, and personalized follow-ups to maintain engagement. Our findings reinforce prior evidence that digital recruitment alone is insufficient and highlight the importance of hybrid approaches combining digital and traditional outreach strategies [40].

Limitations

Several limitations should be considered when interpreting these findings. First, this was an observational analysis conducted within a single academic health system, which may limit generalizability to other settings with different patient demographics or digital infrastructure. Second, the study relied on electronic health record–based digital recruitment and patient portal messaging, which presumes access to broadband internet and digital literacy; these factors were not directly measured and may have influenced participation patterns. Findings may generalize to similar academic health systems using portal-based recruitment but may differ in settings with lower portal adoption or different sociodemographic profiles. Third, the cohort was predominantly White (93.7%), limiting the ability to examine racial and ethnic disparities in digital recruitment. Future studies should prioritize inclusion of more diverse populations to assess whether similar sociodemographic patterns persist across racial and ethnic groups. Fourth, while we examined neighborhood-level (ADI) and housing-based (HOUSES) socioeconomic measures, other dimensions of socioeconomic status, such as income, education, and employment, were not available and could provide additional insights. Fifth, attrition analysis was based on recruitment checkpoints rather than qualitative data on participant motivations or barriers, which constrains the interpretation of underlying causes for dropout. Finally, the study period was relatively short (March to July

XSL•FO
RenderX

2024), and findings may not reflect seasonal or long-term trends in digital recruitment dynamics.

To address these challenges, we recommend several strategies to enhance inclusivity and effectiveness in digital recruitment. First, adopting a multichannel approach is essential. Digital methods should be complemented with traditional outreach, particularly for populations with limited online access, such as those in rural areas or socioeconomically disadvantaged communities. Second, age-specific engagement should be prioritized by developing tailored messaging and user experiences for different age groups, with special attention to reducing dropout rates among younger participants. Similarly, geographic barriers must be addressed by providing technical support and offering alternative participation options for both rural and urban clusters. To ensure continuous improvement, it is critical to track recruitment analytics. Monitoring data will help identify dropout points and demographic trends, enabling real-time adjustments to recruitment strategies. Additionally, efforts should be made to minimize participation burden by streamlining enrollment processes to reduce time demands, which is particularly beneficial for individuals with limited availability, such as those with caregiving responsibilities. Socioeconomic factors also require consideration. Recruitment materials should be inclusive, and incentives or support should be offered to offset participation costs, thereby improving socioeconomic accessibility. Furthermore, while device type did not affect completion time in our findings, platforms should still be optimized for device diversity, ensuring mobile-friendly and cross-device accessibility. Ultimately, intentional design and continuous evaluation are key to ensuring that digital methods promote inclusivity rather than hinder it. Future work should explore the mechanisms behind demographic and socioeconomic disparities, test interventions to address these patterns, and evaluate whether similar dynamics occur across other clinical domains.

Conclusion

This study demonstrates that digital recruitment methods in neurological research are subject to demographic, urbanicity, and socioeconomic influences that affect the representativeness of study populations. By mapping these factors, this study provides actionable insights for designing recruitment strategies that improve participation in remote neurological research and similar digital health initiatives. Our findings suggest that although digital recruitment expands reach, it does not eliminate traditional barriers and introduces new challenges. The digital divide appears to manifest in nuanced ways throughout the recruitment process, potentially influencing who participates in neurological research and, consequently, who benefits from its findings. The relatively low overall completion rate (7.1% of invited participants) underscores the persistent challenge of recruitment in specialized medical research, even with digital methods. Collectively, these findings reinforce the importance of refining digital recruitment strategies to bridge the persistent digital divide and promote research participation.

Acknowledgments

Some figures were created with BioRender. The authors declare the use of generative artificial intelligence (AI) in the research and writing process. According to the GAIDeT taxonomy (2025), tasks delegated to generative AI tools under full human supervision included literature search and systematization, code generation and optimization, proofreading and editing, text summarization, and reformatting. The generative AI tool used was GPT-5 within Microsoft Copilot. Responsibility for the final manuscript lies entirely with the authors; generative AI tools are not listed as authors and do not bear responsibility for the final outcomes. This declaration is submitted by PN.

Funding

This project was supported by Grant Number UL1TR002377 from the National Center for Advancing Translational Sciences (NCATS) and Grant Number R01AG083832 from the National Institute on Aging (NIA). The funding agencies had no role in the study design, data collection, analysis, interpretation of results, or writing of the manuscript. The contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to the sensitive nature of patient information and restrictions imposed by the Mayo Clinic Institutional Review Board. Deidentified data may be made available by the corresponding author upon reasonable request and contingent on appropriate institutional approvals.

Authors' Contributions

Conceptualization: HB (lead), PN (supporting)
Project administration: ADB, VMS
Data curation: ADB, VMS
Methodology: HB
Formal analysis: PN
Writing - original draft: PN
Writing - review & editing: JRD, JLS, VH, DTJ, RLU, HB

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
American Psychological Association JARS-Quant guidelines, completed checklist.
[PDF File, 193 KB - jmir_v28i1e83432_app1.pdf]

References

1. Unger JM, Fleury M. Nationally representative estimates of the participation of cancer patients in clinical research studies according to the Commission on Cancer. J Clin Oncol 2021 Oct 1;39(28):74. [doi: 10.1200/JCO.2020.39.28_suppl.74]
2. Unger JM, Shulman LN, Facktor MA, Nelson H, Fleury ME. National estimates of the participation of patients with cancer in clinical research studies based on Commission on Cancer accreditation data. J Clin Oncol 2024 Jun 20;42(18):2139-2148. [doi: 10.1200/JCO.23.01030] [Medline: 38564681]
3. Glickman SW, Anstrom KJ, Lin L, et al. Challenges in enrollment of minority, pediatric, and geriatric patients in emergency and acute care clinical research. Ann Emerg Med 2008 Jun;51(6):775-780. [doi: 10.1016/j.annemergmed.2007.11.002] [Medline: 18191297]
4. Khazak A, Parsa K, Elias E, Liaw BCH. Geriatric enrollment patterns in clinical trials at a National Cancer Institute. J Clin Oncol 2024 Jun 1;42(16_suppl):e13762-e13762. [doi: 10.1200/JCO.2024.42.16_suppl.e13762]
5. Ashery RS, McAuliffe WE. Implementation issues and techniques in randomized trials of outpatient psychosocial treatments for drug abusers: recruitment of subjects. Am J Drug Alcohol Abuse 1992;18(3):305-329. [doi: 10.3109/00952999209026069] [Medline: 1329493]
6. Bikou AG, Deligianni E, Dermiki-Gkana F, et al. Improving participant recruitment in clinical trials: comparative analysis of innovative digital platforms. J Med Internet Res 2024 Dec 18;26:e60504. [doi: 10.2196/60504] [Medline: 39693619]
7. Jackson J. Expanding opportunities to maximise research recruitment and data collection using digital tools. Int J Soc Res Methodol 2023 Nov 2;26(6):721-731. [doi: 10.1080/13645579.2022.2091258]
8. Kasahara A, Mitchell J, Yang J, Cuomo RE, McMann TJ, Mackey TK. Digital technologies used in clinical trial recruitment and enrollment including application to trial diversity and inclusion: a systematic review. Digit Health 2024;10:20552076241242390. [doi: 10.1177/20552076241242390] [Medline: 38559578]
9. Whitaker C, Stevelink S, Fear N. The use of Facebook in recruiting participants for health research purposes: a systematic review. J Med Internet Res 2017 Aug 28;19(8):e290. [doi: 10.2196/jmir.7071] [Medline: 28851679]
10. Schwinne M, Woods E, Chan B, et al. Transforming research recruitment: leveraging EHR systems and patient portals. J Clin Transl Sci 2025;9(1):e10. [doi: 10.1017/cts.2024.692] [Medline: 39911926]
11. Iflaifel M, Hall CL, Green HR, et al. Strategies to improve recruitment in mental health clinical trials: a scoping review (RE-MIND study). Trials 2024 Dec 18;25(1):832. [doi: 10.1186/s13063-024-08665-x] [Medline: 39695709]
12. Biga R, Nottebaum S, Kozlakidis Z, Psomiadis S. Digitalization of healthcare in lmics: digital health and the digital divide based on technological availability and development. In: Kozlakidis Z, Muradyan A, Sargsyan K, editors. Digitalization of Medicine in Low- and Middle-Income Countries: Paradigm Changes in Healthcare and Biomedical Research: Springer International Publishing; 2024:185-193. [doi: 10.1007/978-3-031-62332-5_18]
13. König LM, Krukowski RA, Kuntsche E, et al. Reducing intervention- and research-induced inequalities to tackle the digital divide in health promotion. Int J Equity Health 2023 Dec 4;22(1):249. [doi: 10.1186/s12939-023-02055-6] [Medline: 38049789]
14. Litchfield I, Shukla D, Greenfield S. Impact of COVID-19 on the digital divide: a rapid review. BMJ Open 2021 Oct 12;11(10):e053440. [doi: 10.1136/bmjopen-2021-053440] [Medline: 34642200]
15. Livingood WC, Bautista MAB, Smotherman C, et al. Comparative study of different SES neighborhood clinics for health literacy and internet access. Digit Health 2022;8:20552076221123715. [doi: 10.1177/20552076221123715] [Medline: 36081750]
16. Walker DM, Hefner JL, Fareed N, Huerta TR, McAlearney AS. Exploring the digital divide: age and race disparities in use of an inpatient portal. Telemed J E Health 2020 May;26(5):603-613. [doi: 10.1089/tmj.2019.0065] [Medline: 31313977]
17. Tomiwa T, Wong E, Miller HN, et al. Leveraging digital tools to enhance diversity and inclusion in clinical trial recruitment. Front Public Health 2024;12:1483367. [doi: 10.3389/fpubh.2024.1483367] [Medline: 39529717]
18. Bradway M, Årsand E. Diverse recruitment strategies are needed to reduce digital divide: results from a workshop addressing digital divide and effects of pandemic restrictions. Stud Health Technol Inform 2022 May 25;294:811-812. [doi: 10.3233/SHTI220592] [Medline: 35612212]
19. Friis-Healy EA, Nagy GA, Kollins SH. It is time to REACT: opportunities for digital mental health apps to reduce mental health disparities in racially and ethnically minoritized groups. JMIR Ment Health 2020;8(1):e25456. [doi: 10.2196/25456]
20. Bensoussan Y, Sigaras A, Rameau A, Elemento O, Powell M, Dorr D, et al. Bridge2AI-Voice: An Ethically-Sourced, Diverse Voice Dataset Linked to Health Information (Version 200): PhysioNet; 2025. [doi: 10.13026/3xt6-rf05]

XSL•FO
RenderX

21. APA Style Journal Article Reporting Standards (JARS). American Psychological Association. 2018. URL: https://apastyle.apa.org/jars [accessed 2026-01-29]
22. Qualtrics AI + XM platform_. Qualtrics. 2024. URL: https://www.qualtrics.com/en-au/ [accessed 2026-01-29]
23. Juhn YJ, Beebe TJ, Finnie DM, et al. Development and initial testing of a new socioeconomic status measure based on housing data. J Urban Health 2011 Oct;88(5):933-944. [doi: 10.1007/s11524-011-9572-7] [Medline: 21499815]
24. About the Neighbothood Atlas® and Area Deprivation Index. Center for Health Disparities Research University of Wisconsin School of Medicine and Public Health. 2024. URL: https://www.neighborhoodatlas.medicine.wisc.edu [accessed 2026-01-29]
25. Kind AJH, Buckingham WR. Making neighborhood-disadvantage metrics accessible - the Neighborhood Atlas. N Engl J Med 2018 Jun 28;378(26):2456-2458. [doi: 10.1056/NEJMp1802313] [Medline: 29949490]
26. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988 Dec;83(404):1198-1202. [doi: 10.1080/01621459.1988.10478722]
27. Cho PJ, Olaye IM, Shandhi MMH, Daza EJ, Foschini L, Dunn JP. Identification of key factors related to digital health observational study adherence and retention by data-driven approaches: an exploratory secondary analysis of two prospective longitudinal studies. Lancet Digit Health 2025 Jan;7(1):e23-e34. [doi: 10.1016/S2589-7500(24)00219-X] [Medline: 39722250]
28. Pratap A, Neto EC, Snyder P, et al. Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. NPJ Digit Med 2020;3(1):21. [doi: 10.1038/s41746-020-0224-8] [Medline: 32128451]
29. Gutiérrez Á, Cain R, Aranda MP. The digital divide exacerbates disparities in Latinx recruitment for Alzheimer's disease and related dementias online education during COVID-19. Gerontol Geriatr Med 2022;8:23337214221081372. [doi: 10.1177/23337214221081372] [Medline: 35252476]
30. Jongebloed H, Anderson K, Winter N, et al. The digital divide in rural and regional communities: a survey on the use of digital health technology and implications for supporting technology use. BMC Res Notes 2024 Mar 28;17(1):90. [doi: 10.1186/s13104-024-06687-x] [Medline: 38549176]
31. Kelpin S, Brockman TA, Decker PA, et al. Increasing digital equity to promote online smoking cessation program engagement among rural adults: a randomized controlled pilot trial. Commun Med (Lond) 2024 Oct 8;4(1):194. [doi: 10.1038/s43856-024-00624-6] [Medline: 39375517]
32. Quinton JK, Ong MK, Vangala S, et al. The association of broadband internet access and telemedicine utilization in rural western Tennessee: an observational study. BMC Health Serv Res 2021 Aug 3;21(1):765. [doi: 10.1186/s12913-021-06746-0] [Medline: 34344377]
33. Dorsey ER, Glidden AM, Holloway MR, Birbeck GL, Schwamm LH. Teleneurology and mobile technologies: the future of neurological care. Nat Rev Neurol 2018 May;14(5):285-297. [doi: 10.1038/nrneurol.2018.31] [Medline: 29623949]
34. Price-Haywood EG, Arnold C, Harden-Barrios J, Davis T. Stop the divide: facilitators and barriers to uptake of digital health interventions among socially disadvantaged populations. Ochsner J 2023;23(1):34-42. [doi: 10.31486/toj.22.0101] [Medline: 36936477]
35. Brown SH, Griffith ML, Kripalani S, Horst SN. Association of health literacy and area deprivation with initiation and completion of telehealth visits in adult medicine clinics across a large health care system. JAMA Netw Open 2022 Jul 1;5(7):e2223571. [doi: 10.1001/jamanetworkopen.2022.23571] [Medline: 35877125]
36. Ettman CK, Ringlein GV, Dohlman P, et al. Trends in mental health care and telehealth use across area deprivation: an analysis of electronic health records from 2016 to 2024. PNAS Nexus 2025 Feb;4(2):pgaf016. [doi: 10.1093/pnasnexus/pgaf016] [Medline: 39959710]
37. Jolliff A, Holden RJ, Valdez R, et al. Investigating the best practices for engagement in remote participatory design: mixed methods analysis of 4 remote studies with family caregivers. J Med Internet Res 2024 Dec 3;26:e60353. [doi: 10.2196/60353] [Medline: 39626228]
38. Brøgger-Mikkelsen M, Ali Z, Zibert JR, Andersen AD, Thomsen SF. Online patient recruitment in clinical trials: systematic review and meta-analysis. J Med Internet Res 2020 Nov 4;22(11):e22179. [doi: 10.2196/22179] [Medline: 33146627]
39. Pagliaro JA, Wash LK, Ly K, et al. Enrollment and retention outcomes from the Veterans Health Administration for a remote digital health study: multisite observational study. JMIR Form Res 2025 Aug 15;9:e68676. [doi: 10.2196/68676] [Medline: 40334212]
40. Frampton GK, Shepherd J, Pickett K, Griffiths G, Wyatt JC. Digital tools for the recruitment and retention of participants in randomised controlled trials: a systematic map. Trials 2020 Jun 5;21(1):478. [doi: 10.1186/s13063-020-04358-3] [Medline: 32498690]

Abbreviations

ADI: Area Deprivation Index
AI: artificial intelligence
APA: American Psychological Association
HOUSES: Housing-based Socioeconomic Status
JARS: Journal Article Reporting Standards

XSL•FO
RenderX

MCAR: missing completely at random
PTrax: Participant Tracking System
SES: socioeconomic status

Edited by S Brini; submitted 02.Sep.2025; peer-reviewed by B Agyare-Aggrey, C Ahmadu; revised version received 07.Jan.2026; accepted 09.Jan.2026; published 25.Feb.2026.

Please cite as:
Nejat P, Bachman AD, Stubbs VM, Duffy JR, Stricker JL, Herasevich V, Jones DT, Utianski RL, Botha H
Sociodemographic Drivers of Recruitment and Attrition in Digital Neurological Research: Longitudinal Cohort Study
J Med Internet Res 2026;28:e83432
URL: https://www.jmir.org/2026/1/e83432
doi:10.2196/83432

© Peyman Nejat, Ashley D Bachman, Vicki M Stubbs, Joseph R Duffy, John L Stricker, Vitaly Herasevich, David T Jones, Rene L Utianski, Hugo Botha. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Using Epidemiological Test Diagnostics to Select Fraud Detection Methods: Secondary Analysis of Quantitative Cross-Sectional Survey Data

Rachel Willard-Grace[1], MPH; Tali Klima[2], PhD; Mansi Dedhia[1], MPH, MS; Emily Lo[1], BA, BS; Annie Nisnevich[1], BS; Allison Gray[3], MA; Holly Henry[3], MHS, PhD

[1]Department of Family and Community Medicine, University of California, San Francisco, Box 1315 2540 23rd St., Floor 5, 5511, San Francisco, CA, United States

[2]Practical Research Solutions, Sunnyvale, CA, United States

[3]Lucile Packard Foundation for Children's Health, Palo Alto, CA, United States

Corresponding Author:
Rachel Willard-Grace, MPH
Department of Family and Community Medicine, University of California, San Francisco, Box 1315 2540 23rd St., Floor 5, 5511, San Francisco, CA, United States

Abstract

Background: Survey research has the potential to elevate the experiences and opinions of marginalized populations. The rising number of bot attacks, a method of participant fraud that creates multiple records in survey data using automated software, threatens to drown out those voices and produce inaccurate findings. Rapid identification and mitigation of bot attacks are vital; however, there is limited guidance for researchers on scalable approaches to address this problem.

Objective: This study aimed to assess how well recommended methods detect fraud using an epidemiological diagnostic test framework to inform web-based survey researchers on how best to identify and shut down bot attacks.

Methods: We analyzed data from a cross-sectional web-based statewide survey on access to pediatric subspecialty care in California that used Qualtrics survey software. Caregivers of children with chronic conditions were recruited through family resource centers (FRCs), nonprofit agencies serving families with developmental delays and chronic medical conditions. The survey was sent out to 17 FRCs, whose staff distributed anonymous links to their clients through listservs and flyers. Respondents who completed the survey received a US $30 gift card. Prior to launch, we designed a protocol to identify and respond to bot attacks and reviewed responses for markers of fraudulent activity. If markers were identified or there was a spike in responses, a senior member of our research team reviewed patterns among all submitted surveys for each FRC to look for signs of bot attacks. We calculated epidemiologic measures of diagnostic test accuracy, such as sensitivity, specificity, positive predictive value, and negative predictive value, which describe a test's ability to distinguish "disease" (in this case, fraudulent records) from normal cases, to better understand the utility of recommended strategies to identify bot attacks.

Results: We received 646 valid survey records and 905 fraudulent records resulting from bot attacks. The primary indicator of a bot attack was a sudden spike in responses to the survey. Differences in demographics and outcomes, including wait times for pediatric subspecialty care and use of health care services, between the valid and fraudulent data indicated that failure to remove fraudulent records would have substantially altered the survey results. Most recommended methods in the literature for identifying fraudulent responses had low sensitivity to detect bot attacks, and only 2 were better than chance alone at correctly identifying bot attacks. Combinations of fraud markers and blocks of repeated responses were particularly useful to identify bot attacks.

Conclusions: Fraudulent data entry using bots is increasing in survey research. Sharing flexible protocols to identify and mitigate them in a way that is responsive to their ever-changing nature is vital to ensuring that researchers elevate the voices of real people within survey research to inform policy and programmatic discussions.

(J Med Internet Res 2026;28:e85161) doi:10.2196/85161

KEYWORDS

health care surveys; bots; surveys; questionnaires; survey methodology; web-based data collection; specificity; sensitivity

XSL•FO
RenderX

Introduction

Survey research has the potential to elevate the experiences and opinions of individuals who may be difficult to reach. While ideal survey methods include the use of individualized, single-use survey links shared with a defined and verified set of respondents, in reality, there are important populations who are not neatly vetted and included in registries and, in some cases, whose safety depends on being invisible and unregistered [1,2]. Since creating single-use survey links matched to known populations is challenging in these situations, many surveys use generic survey links, which carry a greater risk of use by fraudulent responders for profit or manipulative purposes.

The rising number of fraudulent responses threatens to drown out those voices. Our nonexhaustive literature review of 13 studies exploring fraudulent responses found a range of 13% to 97% of survey responses were identified as fraudulent, with a pooled proportion of 61%. Eight of 13 studies reported that more than half of responses were fraudulent (Multimedia Appendix 1) [1-10]. The onslaught of fraudulent responses has a number of consequences; key among these is the distortion of study findings, leading to inaccurate results if fraudulent records are not excluded [4]. For the purpose of this paper, we define a "fraudulent record" as any record in which false information is presumed to be purposely provided. Bot attacks comprise a subset of these fraudulent records and are distinguished by the use of software to generate a large number of responses [11].

Rapid identification and mitigation of bot attacks are vital; however, there is limited guidance for survey researchers in the face of rapidly evolving and ever-more-sophisticated fraudulent attacks. We did not find any prior studies that used standard epidemiological methods to assess the effectiveness of various bot detection approaches. Standard recommendations include identifying duplicate responses or off-topic responses, inconsistencies between location stamp and location reported, similar or identical response or pattern of responses is provided across surveys, identifying inconsistencies between different survey items (eg, age and year of birth), embedding "honeypot" items that are visible to bots but not humans, or short survey completion time [10]. Most of these recommendations are based on assumptions of manual fraud, where individuals submit multiple surveys by hand (formerly known as "ballot box stuffing"), or relatively unintelligent technology that uses simple strategies and does not learn how to game the survey (eg, overcome screening items).

Based on our experience and anecdotal evidence from others, researchers have already observed failures of many of these methods, which are less applicable to modern bot capabilities. Moreover, among survey researchers seeking to reach populations with limited digital literacy, some suggested methods pose the additional danger of excluding eligible but less digitally savvy participants [6]. For example, we abandoned use of CAPTCHA screening (identification of images, such as motorcycles, in a matrix) on our survey platform after technologically savvy staff were consistently unable to pass the screening themselves.

Little is known about the ability of current fraud detection methods to successfully identify fraudulent records (sensitivity, in clinical epidemiological terms) or their ability to correctly identify valid records (ie, not to label a valid record as a fraudulent one, hereafter referred to as "specificity"). Using data from a California-wide survey and a diagnostic approach that treats bot attacks as a disease to be detected by various possible tests, we describe the ability of available tools to correctly identify "diseased cases" (positive predictive value) and the risk that they will miss such cases (negative predictive value). Insights from this analysis may provide guidance on how best to identify and shut down large-scale bot attacks on surveys.

Methods

We conducted a secondary analysis of data from a survey of caregivers of children with chronic conditions in California. Results of the study have been previously reported [12-15].

Population and Setting

We recruited caregivers of children with chronic conditions who were clients of 17 family resource centers (FRCs) across California. FRCs are nonprofit community agencies serving families of children with developmental delays and chronic medical conditions through education and advocacy. For example, they frequently provide classes and peer support to help caregivers of children with chronic conditions to secure social and educational services, and to connect with other families of children with similar conditions. Due to restricted budgets, they often have fewer organizational resources, including paid staff, and rely on volunteer caregivers to advise and provide services to families. They frequently communicate with clients through email listservs. Eligible participants were English- or Spanish-speaking adult caregivers residing in California whose child had a visit with a new pediatric subspecialist in the previous 12 months.

Distribution of the Survey

The study team initially proposed sharing a list of unique, single-use survey links with FRCs to be distributed to their clients, but FRCs were unable to use these unique links with their listserv software and were restricted from sharing their listservs with outside organizations, like the study team, which could have sent unique links directly to FRC clients. As a result, we generated a unique generic link for each FRC. FRCs distributed an invitation to the survey from the study team and the survey link through their email listservs of clients and printed flyers with a QR code link to the survey to post in their offices for access by families coming in for services. Data was collected between August and October 2023.

Fraud Detection Protocol

Based on the REAL (Reflect, Expect, Analyze, Label) framework for online survey fraud [16], we considered in advance how to prevent or identify fraudulent responses at each stage of the process, including pre–data collection (eg, recruitment through community-based partners), data collection (eg, integration of fraud detection questions in the survey design), and post–data collection (eg, a systematic review

XSL•FO
RenderX

process). We also borrowed ideas from Anomaly Detection Theory, grounded in engineering sciences, which posits that detection of anomalies requires identifying data that do not follow established patterns [17].

We anticipated that the distribution of the survey through an anonymous link to large groups of people could make the survey vulnerable to fraudulent responses, and in particular, to bot attacks in which an individual writes code to automatically submit multiple survey responses that appear to originate from eligible respondents. Based on sparse literature and consultation with colleagues, the team utilized 10 different fraud detection tools when the survey was fielded, including proprietary tools that can be turned on within the Qualtrics survey platform, such as invisible ReCAPTCHA technology, and other flags, like the Relevant ID Duplicate, Fraud scores, and Ballot Box Stuffing flags. In addition, we embedded questions that required internally consistent logic (eg, birth year and age), monitored open-ended questions for nonsensical or repeating responses, reviewed longitude and latitude to ensure respondent location in California, considered records based on completion times (including shorter-than-expected response time and duplicate submission times), and verified IP addresses of respondents through an external server of suspicious addresses (see Multimedia Appendix 2). The survey also included a series of 13 screening questions designed to remove ineligible participants before reaching the body of the survey, which we believed would discourage fraudulent responses.

The number of responses was monitored multiple times each day, and spikes in responses were immediately investigated. We completed a logbook entry for responses that triggered concern. Most indicators of fraudulent activity have caveats; for example, IP addresses triggering warnings from the fraudulent database could result from a previous infection of a legitimate device by malware that does not affect the survey (see Multimedia Appendix 2 for additional examples). As a result, no single flag was treated as a sole determinant of fraud. Classification of all records with flags was conducted through consensus between a senior researcher and a research associate. Fraudulent responses were often confirmed through patterns of responses, such as blocks of questions with the same responses (see Figure 1 for example). In the minority of cases in which the coding team was uncertain about a determination, the research associate followed up with the participant, requesting additional information that they could check against the survey to confirm that the response was valid.

Figure 1. Examples of fraudulent response patterns (repetition of blocks across multiple records) in a cross-sectional survey of caregivers of children with new pediatric subspecialty referrals (August-October 2023).

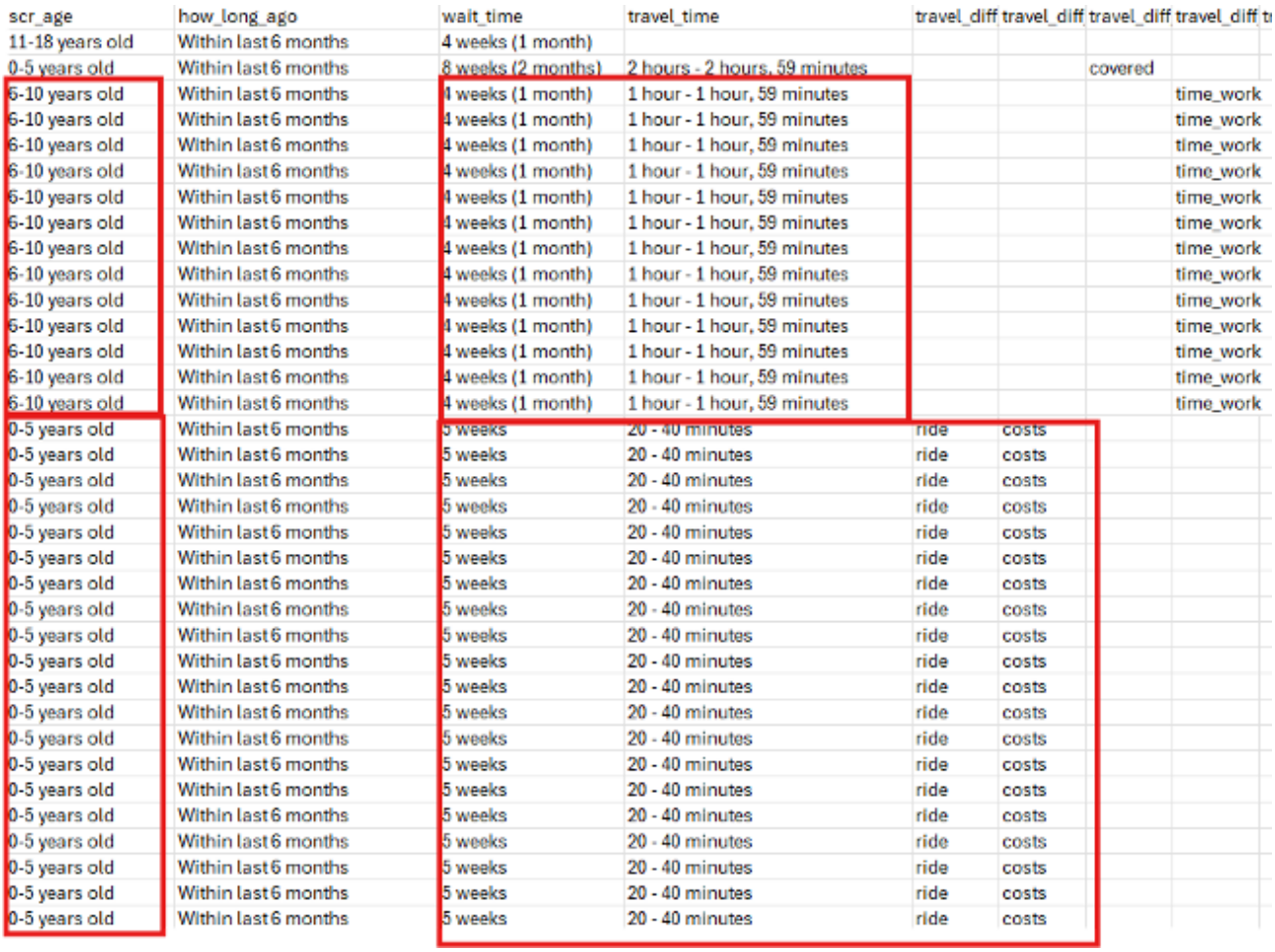

scr_age	how_long_ago	wait_time	travel_time	travel_diff	travel_diff	travel_diff	travel_diff	t
11-18 years old	Within last 6 months	4 weeks (1 month)						
0-5 years old	Within last 6 months	8 weeks (2 months)	2 hours - 2 hours, 59 minutes			covered		
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
6-10 years old	Within last 6 months	4 weeks (1 month)	1 hour - 1 hour, 59 minutes				time_work	
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			
0-5 years old	Within last 6 months	5 weeks	20 - 40 minutes	ride	costs			

XSL•FO
RenderX

Data Analysis

We cleaned datasets from each FRC and reviewed records for final disposition (eg, valid, ineligible based on various criteria, fraudulent). We examined differences in demographic characteristics and health care utilization (outcomes in the primary study) between valid and fraudulent records in order to ascertain how our results might differ if fraudulent data were not identified and removed.

To assess the validity of different fraud detection methods, we calculated the sensitivity, specificity, and positive and negative predictive value for each method (see Figure 2). Sensitivity is the ability of a test or method to correctly identify individuals with a disease, in this case, fraudulent records, while specificity is the ability of a test to correctly identify individuals without a disease, here, valid records [18]. Epidemiologists do not define universal thresholds of acceptable sensitivity and specificity, taking into consideration the danger of having false-positive or negative findings based on the severity and prevalence of the condition, for example. However, one rule of thumb is that the sum of the two values should be between 150% and 200%, meaning that it is at least more likely than chance alone to result in a correct identification [19]. We accordingly calculated the sum of sensitivity and specificity.

Figure 2. Epidemiological measures of diagnostic test accuracy.

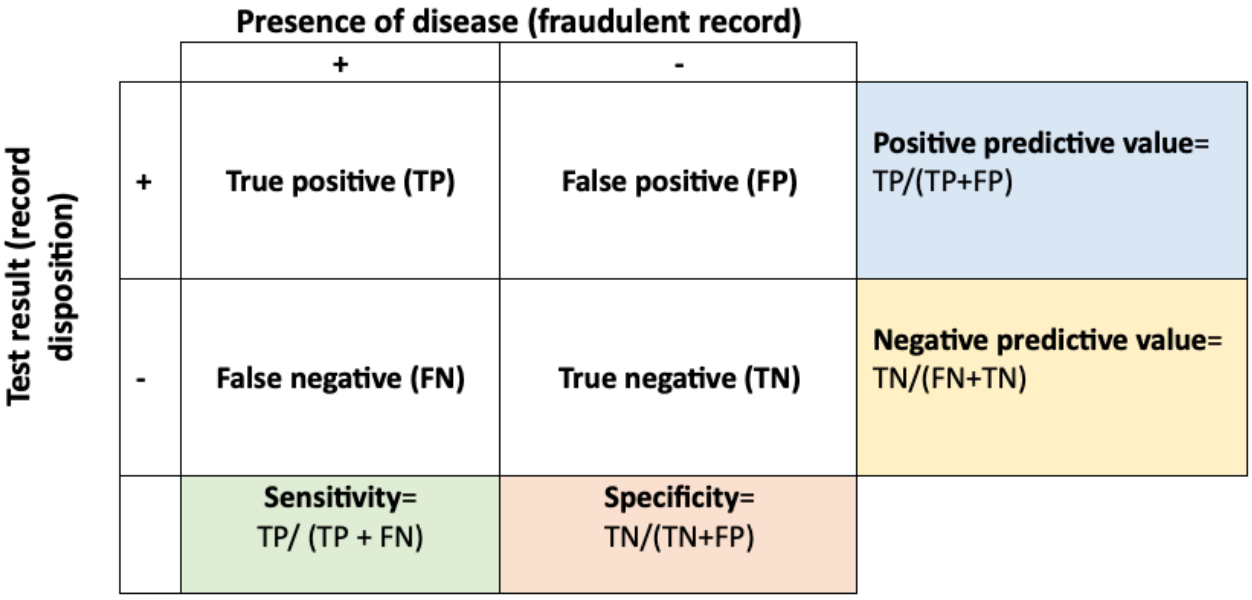

Positive predictive value is the proportion of individuals testing positive (given a fraudulent study disposition) who actually have the disease (are actually fraudulent). Negative predictive value is the proportion of individuals testing negative (given a valid disposition) who truly do not have the disease (are, in fact, valid records). Both positive and negative predictive values depend not only on the sensitivity and specificity of the tests, but also on the prevalence of the condition in the population (in this case, the prevalence of bot attacks in the pool of survey records).

Ethical Considerations

The protocol for the original study was approved by the University of California, San Francisco (UCSF) Institutional Review Board (23‐39316). The board approved implied consent for the study, whereby participants clicking on the survey link first viewed an information page approved by the board and then opted into participation by clicking on a link to take part in the survey. Data was deidentified prior to analysis. Respondents received a US $30 electronic gift card for completing the survey, once the validity of the survey was confirmed by the study team (see Multimedia Appendix 3).

Results

Of 2220 records, 646 (29%) were valid and complete (hereafter "valid"), 905 (41%) were classified as fraudulent by the protocol, 263 (12%) were valid but ineligible based on the eligibility screeners, 308 (14%) were incomplete surveys, and 98 (4%) were duplicate records. Duplicate records were distinguished from fraudulent records as those that made no effort to disguise their identity or device as identical; these appeared to result from respondents being unsure if their previous response had been received. Subsequent analyses were limited to valid and fraudulent records (n=1551).

There were a number of statistically significant differences in the reported demographics of valid versus fraudulent records (Table 1). When valid responses were compared with the overall responses (for valid + fraudulent records), valid responses were more likely to report having younger children (317/646, 49.1% aged 0‐5 y vs 536/1472, 36.4% for all responses); being Asian American and Pacific Islander (106/639, 16.4% vs 125/1368, 8.1%) or Latino/a (352/639, 55.1% vs 382/1368, 27.9%); living in a big city (218/635, 34.3% vs 697/1343, 51.9%); and having commercial health insurance (226/640, 35.3% vs 344/1353, 25.4%).

XSL•FO
RenderX

Table . Demographic characteristics of all, valid, and fraudulent respondents to a cross-sectional survey for caregivers of children with new pediatric subspecialty referrals in California (August-October 2023).

Characteristic	All (valid + fraudulent records), n/N (%)	Valid records, n/N (%)	Fraudulent records, n/N (%)	Difference between valid and fraudulent records	
				Pearson chi-square (*df*)	*P* value
Age (y)				104.12 (1)	<.001
0-5	536/1472 (36.4)	317/646 (49.1)	219/825 (26.5)		
6-10	527/1472 (35.8)	148/646 (22.9)	379/825 (45.9)		
11-20	408/1472 (27.7)	181/646 (28.0)	227/825 (27.7)		
Child's gender[a]				1.41 (1)	.24
Boy	829/1364 (60.8)	398/640 (62.2)	431/724 (59.5)		
Girl	530/1364 (38.9)	237/640 (37)	293/724 (40.5)		
Another gender	5/1364 (0.4)	5/640 (0.8)	0/724 (0.0)		
Race/ethnicity[b]					
Black/African American	190/1368 (13.9)	38/639 (5.9)	152/728 (20.9)	63.4 (1)	<.001
American Indian/Alaskan Native	34/1368 (2.5)	31/639 (4.8)	3/728 (0.3)	27.7 (1)	<.001
Asian/Pacific Islander	125/1368 (8.1)	106/639 (16.4)	19/728 (2.1)	104.2 (1)	<.001
Hispanic/Latino	382/1368 (27.9)	352/639 (55.1)	30/728 (4.1)	439 (1)	<.001
White	799/1368 (58.5)	258/639 (40.4)	541/728 (74.3)	161.4 (1)	<.001
Other	20/1368 (1.5)	18/639 (2.8)	2/728 (0.3)	—[c]	—
Urbanicity				149 (2)	<.001
Big city	697/1343 (51.9)	218/635 (34.3)	479/708 (67.7)		
Small city or suburb	557/1343 (41.5)	358/635 (56.4)	199/708 (28.1)		
Rural area	89/1343 (6.6)	59/635 (9.3)	30/708 (4.2)		
Insurance[b]					
Medicaid	986/1353 (72.9)	488/640 (76.3)	494/713 (69.9)	7 (1)	.008
Commercial	344/1353 (25.4)	226/640 (35.3)	118/713 (16.6)	62.6 (1)	<.001

[a]Reflects caregiver-reported gender. Because of small cell sizes, "another gender" was excluded from testing of differences across groups.

[b]May report more than 1 response, so responses may sum to more than 100%. For multiple response comparisons, used Pearson X^2 analysis to compare responses for each option across valid versus fraudulent responses.

[c]Not applicable.

With respect to health care utilization, which was the key outcome in the primary study, there were significant differences between valid versus fraudulent responses. Valid responses generally reported lower utilization than overall responses, such as having at least one hospitalization in the last year (172/639, 26.9% for valid vs 578/1352, 42.8% of overall responses; Table 2), at least one emergency department visit (324/638, 50.8% vs 799/1351, 59.1%), and at least 1 hour per week in medical visits (469/636, 73.7% vs 1139/1347, 84.6%). On the other hand, valid respondents were more likely to report 3-month waits or more for pediatric subspecialty care compared to the overall sample (156/638, 24.5% vs 271/1382, 19.6%).

XSL•FO
RenderX

Table . Outcomes for health care utilization of all, valid, and fraudulent respondents to a cross-sectional survey for caregivers of children with new pediatric subspecialty referrals in California (August-October 2023).

Outcome	All (valid + fraudulent) records, n/N (%)	Valid records, n/N (%)	Fraudulent records, n/N (%)	Difference between valid and fraudulent records	
				Pearson chi-square (*df*)	*P* value
Hospital visits in last year				124.1 (1)	<.001
0 times	774/1352 (57.3)	467/639 (73.1)	307/713 (43.1)		
1 or more times	578/1352 (42.8)	172/639 (26.9)	406/713 (56.9)		
ED[a] visits in last year				34.9 (1)	<.001
0 times	552/1351 (40.9)	314/638 (49.2)	238/713 (33.4)		
1 or more times	799/1351 (59.1)	324/638 (50.8)	475/713 (66.6)		
Wait time for pediatric subspecialist				17.6 (1)	<.001
2 mo or less	1111/1382 (80.4)	482/638 (75.6)	629/744 (84.5)		
3 mo or more	271/1382 (19.6)	156/638 (24.5)	115/744 (15.5)		
Hours per week in medical visits				108 (1)	<.001
<1 h/wk	208/1347 (15.4)	167/636 (26.3)	41/711 (5.8)		
1 or more hours/week	1139/1347 (84.6)	469/636 (73.7)	670/711 (94.2)		

[a]ED: emergency department.

We examined the validity of different fraud detection methods using epidemiological measures of diagnostic tests (Table 3). Most methods had low sensitivity (poor ability to detect fraudulent records), including comparisons of paired items (eg, age and y of birth), short completion time, duplicate IP addresses, and review of open-ended responses for duplicate or nonsensical answers. Even ReCAPTCHA technology, which relies on information that bots can see but humans cannot, only detected about half of fraudulent responses. The most sensitive indicators appeared to be a fraudulent IP address detection service, where a low threshold setting (ie, one that identified not only IP addresses with confirmed fraudulent activity but also those with suspicious activity) yielded a sensitivity of 88%, and the proprietary Fraud ID feature on the survey software, with a sensitivity of 67%.

XSL•FO
RenderX

Table . Sensitivity, specificity, positive and negative predictive value of fraud detection methods at identifying fraudulent responses in a cross-sectional survey for caregivers of children with new pediatric subspecialty referrals (August-October 2023) (n=1551).

Method of detecting fraudulent responses	True positive, n	False positive, n	False negative, n	True negative, n	Sensitivity (%)	Specificity (%)	Sensitivity + specificity (%)	Positive predictive value (%)	Negative predictive value (%)
Nonsensical open-ended responses	0	0	905	646	0	100	100	—[a]	41.7
Repeated open-ended responses	0	0	905	646	0	100	100	—	41.7
Qualtrics Ballot Box stuffing	0	0	905	646	0	100	100	—	41.7
Duplicate IP addresses	82	45	823	601	9.1	93	102.1	64.6	42.2
Short completion time (<4 min)	135	1	770	645	14.9	99.8	114.7	99.3	45.6
Discrepancy between paired items (age and year of birth)	183	15	642	631	22.2	97.7	119.9	92.4	49.6
Fraudulent IP address database (lower threshold)[b]	758	337	106	309	87.7	47.8	135.5	69.2	74.5
Longitude/latitude	389	9	637	516	37.9	98.3	136.2	97.7	44.8
Fraudulent IP address database (higher threshold)[b]	422	8	442	638	48.8	98.8	147.6	98.1	59.1
Qualtrics ReCAPTCHA functionality	466	0	439	646	51.5	100	151.5	100	59.5
Qualtrics Fraud ID	609	27	296	619	67.3	95.8	163.1	95.8	67.7
Combination: Qualtrics ReCAPTCHA or Fraud ID	675	37	230	609	74.6	94.3	168.9	94.8	72.6

[a]Not available.

[b]The service used to look up fraudulent IP addresses had two thresholds: a lower threshold indicating that the IP address had been used in suspicious activity, and a higher threshold, indicating that it had been used in fraudulent activity.

Sensitivity is typically balanced against specificity, the ability of a diagnostic test to accurately predict which people do not have a disease (in this case, to identify valid records). Almost all individual indicators yielded relatively high specificity, suggesting that they correctly identified the valid records in the sample, even as most of these methods could not identify the fraudulent ones well (Table 3).

When we examined the sum of sensitivity and specificity and used the proportional threshold of 150% or greater as an indicator of the ability of a test to identify cases, only fraud detection tools embedded in the survey software (ReCAPTCHA and Fraud ID within the Qualtrics system) met the criteria of being more useful than chance alone in identifying fraudulent records. Notably, each of these methods used alone would have missed between 33% and 48% of fraudulent records, based on

XSL•FO
RenderX

their modest sensitivity. Using the two measures together (ie, flagging if either was positive) would have resulted in a sensitivity of 75% and a specificity of 94%, with a combined total of 168.9%.

When we calculated predictive values for each method, we found that several had positive predictive values in excess of 90%, meaning that if they were activated (a positive test), it was a fairly good indicator that a record was indeed fraudulent, but negative predictive power less than 60%, indicating that the failure of a record to activate one of these indicators was not assurance that the record was valid. For example, activation of the survey software's ReCAPTCHA indicators was strongly suggestive of a fraudulent record (positive predictive values of 100%), but its negative predictive value (59.5%) suggests that it would miss around 40% of fraudulent records.

Discussion

Similar to past studies [1,3,5-10], in this large statewide survey, we observed a high rate of fraudulent records compared to valid ones (1.4 fraudulent records for every valid record) that would have significantly altered the characterization of the sample and study findings if undetected. Strategies recommended for survey fraud detection performed poorly as judged by epidemiological standards for detecting the "disease" of fraud. These included a number of tried-and-true methods, such as the use of paired items (age and year of birth) at opposite ends of the survey [2,5,8-10], which detected only 1 in 5 fraudulent records, or looking for duplicate IP addresses, which detected only 1 in 10 fraudulent records. Only 2 methods, both proprietary survey software technologies, performed better than chance alone in identifying fraudulent surveys, but even the combination of the 2 methods failed to detect about a quarter of fraudulent survey responses.

Fraudulent responses, if undetected, would have significantly altered the characterization of the sample as well as the study findings. For example, the responses from bot attacks indicated greater health care utilization and shorter wait times for pediatric subspecialty care compared to valid responses. Thus, failure to remove bot responses would have resulted in the incorrect conclusion that Californian families have greater access to pediatric subspecialty care than they actually do.

None of the 10 recommended fraud detection methods used in this study was effective as a stand-alone method capable of identifying fraudulent records, even those that other studies have reported to be highly effective. Moreover, our study represents a small number of bot attacks, and comparison to past literature suggests that bot attacks can have highly individualized signatures. For example, MacKinnon et al [20] were able to identify that over 31% of responses were fraudulent based on nonsensical open-ended questions, duplicate IP addresses, geocoding, and short completion time. In our study, these combined methods would have detected only 43% of fraudulent cases. Rather than use our results to suggest discarding or prioritizing particular measures, we echo the call of Dewitt et al [6] for the use of dynamic protocols, which use as many methods as possible to flag suspicious cases and human reviewers to manually review flagged records. The method is analogous to installing motion sensors, alarms at entry points, and safe meddling safeguards at a bank, but activating human guards to investigate if any of these detection mechanisms are activated.

Combating fraudulent survey responses poses a fundamental tension between data quality, inclusion, and resources. Manual review of responses and follow-up with suspect cases was highly labor-intensive in this study. Yet fraud detection protocols that require prior authentication of participants (eg, through requiring participants to upload documents [3]) or seek to expeditiously classify cases based on a priori algorithms risk exclusion of legitimate respondents. Populations with lower income, racial and ethnic minorities, and older adults are already underrepresented in internet research [21-23]. Adding additional barriers may pose not only an inconvenience but also a threat to the inclusion of historically marginalized voices [3]. In the short term, combating bot attacks will require greater allocation of resources. Developing strategies to improve bot detection, reviewing results in real time, and following up with participants when the authenticity of responses is in question are time-consuming and raise the cost of research. Ultimately, the field must invest in developing more cost-effective strategies for identifying and removing fraudulent records.

Advances in artificial intelligence (AI) will likely increase the frequency and sophistication of bot attacks, including mimicking human response biases, in an effort to evade detection. On the other hand, they may also contribute to the development of important bot detection tools that can co-evolve with bot attacks. Machine learning models may be well equipped over time to identify newly emerging patterns of fraudulent activity and to adapt as fraudulent tactics evolve. However, the emerging tool of AI does not make current research on patterns of fraudulent survey activity irrelevant, because development of technology for detecting bot attacks will initially rely on existing datasets, such as the one used for this study, using human classification of responses as valid or fraudulent. Anomaly Detection Theory advances a number of approaches to identify problem cases, such as a search for outliers [17]; however, when there are more fraudulent than valid responses to a survey, as was the case with our survey and has been with others, fraudulent cases may cease to be outliers. The need for accurate data for learning models may, therefore, lend even greater urgency to accurate identification within the datasets that are fed to the AI models.

Limitations of this study include a lack of a true gold standard for identifying fraudulent responses, which meant that we relied on manual review and consensus determination of case dispositions as reflective of the correct disposition of fraudulent or valid. On the other hand, we employed a high-effort strategy to distinguish valid from fraudulent responses, reviewing each response manually and following up with respondents when the validity of the response was in doubt, and we observed patterns unlikely to have occurred through true responses. Because fraud detection technologies such as Qualtrics ReCAPTCHA are proprietary, we cannot offer insight into the mechanisms of their ability to detect some fraudulent attacks or their failure to detect others. We collected data in August–October 2023, and bot-facilitating technology and methods may have evolved since that time. This study was conducted through web-based surveys

XSL•FO
RenderX

using anonymous links shared through community-based organizations serving families of children with chronic conditions. Further study is needed to generalize findings to understand patterns of response as bot attacks become more sophisticated.

The incidence of fraudulent responses to web-based surveys using bots has been rapidly increasing in survey research and threatens the integrity of important research. Our study identified two methods that performed better than chance alone in identifying fraudulent cases; the combination of these two tools, both proprietary tools of the survey software, identified 75% of fraudulent cases. Literature suggests that bot attacks may have diverse phenotypes, and our findings support calls for "dynamic protocols" that capture as many potential markers of fraud as possible for review and follow-up. To avoid the need to expend intensive human resources in future research, it is imperative that researchers begin to employ more sophisticated technology, investing in AI-enabled bot detection approaches that can evolve along with fraudulent attacks. As such, interdisciplinary collaborations with computer scientists and tech entrepreneurs who currently solve security problems could be fruitful to elevate the voices of real people within policy and programmatic discussions.

Acknowledgments

We acknowledge the caregiver participants who generously shared their time and expertise to take part in this study. We extend special thanks to the 17 family resource centers across California that took part in the study, as well as Dr Yvette Baptiste, Robert Rochin, Jacqui Knudsen, and Ana Seda from the Family Resource Center Network of California for helping to coordinate data collection. Alicia DiGiammarino provided invaluable support with copyediting.

Funding

Funding for data collection for the original study was provided by the Lucile Packard Foundation for Children's Health (grant 2023-09171). This secondary analysis was conducted without additional funding. Foundation staff contributed to the design of the original study but were not involved in design of this secondary analysis. Foundation staff AG and HH contributed to the manuscript revision and are listed as coauthors.

Data Availability

Requests for deidentified study data will be reviewed by the authors and made available where appropriate.

Authors' Contributions

This analysis was conceptualized and carried out by RW-G. The fraud detection protocol described herein was designed by TK and RW-G. AG, HH, TK, and RW-G contributed to instrument design and interpretation of analyses. RW-G and MD guided data collection and cleaning, with assistance from EL and AN. RW-G drafted the manuscript. All authors reviewed and provided feedback on the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Literature review, including methods tested, and proportion of fraudulent responses.
[DOCX File, 35 KB - jmir_v28i1e85161_app1.docx]

Multimedia Appendix 2
Approaches to detecting fraudulent responses.
[DOCX File, 22 KB - jmir_v28i1e85161_app2.docx]

Multimedia Appendix 3
Sample language in institutional review board/consent form.
[DOCX File, 19 KB - jmir_v28i1e85161_app3.docx]

References

1. Zuniga C, Ragosta S, Thompson TA. Recruiting foreign-born individuals who have sought an abortion in the United States: lessons from a feasibility study. Front Glob Womens Health 2023;4:1114820. [doi: 10.3389/fgwh.2023.1114820] [Medline: 37143605]

XSL•FO
RenderX

2. Myers KJ, Jaffe T, Kanda DA, et al. Reaching the “hard-to-reach” sexual and gender diverse communities for population-based research in cancer prevention and control: methods for online survey data collection and management. Front Oncol 2022;12:841951. [doi: 10.3389/fonc.2022.841951] [Medline: 35756657]
3. Campbell CK, Ndukwe S, Dubé K, Sauceda JA, Saberi P. Overcoming challenges of online research: measures to ensure enrollment of eligible participants. JAIDS, J Acquired Immune Defic Syndr 2022;91(2):232-236. [doi: 10.1097/QAI.0000000000003035]
4. Pratt-Chapman M, Moses J, Arem H. Strategies for the identification and prevention of survey fraud: data analysis of a web-based survey. JMIR Cancer 2021 Jul 16;7(3):e30730. [doi: 10.2196/30730] [Medline: 34269685]
5. Choi SH, Mitchell J, Lipkus I. Lessons learned from an online study with dual-smoker couples. Am J Health Behav 2017 Jan;41(1):61-66. [doi: 10.5993/ajhb.41.1.6] [Medline: 27935791]
6. Dewitt J, Capistrant B, Kohli N, et al. Addressing participant validity in a small internet health survey (the restore study): protocol and recommendations for survey response validation. JMIR Res Protoc 2018 Apr 24;7(4):e96. [doi: 10.2196/resprot.7655] [Medline: 29691203]
7. Griffin M, Martino RJ, LoSchiavo C, et al. Ensuring survey research data integrity in the era of internet bots. Qual Quant 2022;56(4):2841-2852. [doi: 10.1007/s11135-021-01252-1] [Medline: 34629553]
8. Mournet AM, Kleiman EM. Internet-based mental health survey research: navigating internet bots on Reddit. Cyberpsychol Behav Soc Netw 2023 Feb;26(2):73-79. [doi: 10.1089/cyber.2022.0173] [Medline: 36724303]
9. Pekarsky C, Skiffington J, Leijser LM, Slater D, Metcalfe A. Social media recruitment strategies to recruit pregnant women into a longitudinal observational cohort study: usability study. J Med Internet Res 2022 Dec 12;24(12):e40298. [doi: 10.2196/40298] [Medline: 36508244]
10. Pozzar R, Hammer MJ, Underhill-Blazey M, et al. Threats of bots and other bad actors to data quality following research participant recruitment through social media: cross-sectional questionnaire. J Med Internet Res 2020 Oct 7;22(10):e23021. [doi: 10.2196/23021] [Medline: 33026360]
11. Storozuk A, Ashley M, Delage V, Maloney EA. Got bots? Practical recommendations to protect online survey data from bot attacks. Quant Meth Psych 2020;16(5):472-481. [doi: 10.20982/tqmp.16.5.p472]
12. Klima T, Dedhia M, Willard-Grace R. A survey of caregivers of children and youth with special health care needs: family perspectives on peer support. : Lucile Packard Foundation for Children’s Health; 2024 URL: https://lpfch.org/wp-content/uploads/2024/06/20240520-ES-ATC-P2P-JT7.pdf [accessed 2026-02-17]
13. Access to care in California: the CYSHCN family experience. : Lucile Packard Foundation for Children’s Health; 2024 URL: https://childrens-coalition.org/wp-content/uploads/2024/02/Access-to-Care-in-California-The-CYSHCN-Family-Experience.pdf [accessed 2026-02-17]
14. Travel time and family preferences for in-person care. : Lucile Packard Foundation for Children’s Health; 2024 URL: https://lpfch.org/wp-content/uploads/2024/06/FINAL-Travel-Time_1pgr_JT5.pdf [accessed 2026-02-17]
15. Delays to access in pediatric subspecialty care in California. University of California San Francisco, Center for Excellence in Primary Care. URL: https://cepc.ucsf.edu/delays-access-pediatric-subspecialty-care-california [accessed 2026-02-19]
16. Lawlor J, Thomas C, Guhin AT, et al. Suspicious and fraudulent online survey participation: introducing the REAL framework. Methodol Innov 2021 Sep;14(3):20597991211050467. [doi: 10.1177/20597991211050467]
17. Dogo EM, Nwulu NI, Twala B, Aigbavboa C. A survey of machine learning methods applied to anomaly detection on drinking-water quality data. Urban Water J 2019 Mar 16;16(3):235-248. [doi: 10.1080/1573062X.2019.1637002]
18. Trevethan R. Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. Front Public Health 2017;5:307. [doi: 10.3389/fpubh.2017.00307] [Medline: 29209603]
19. Power M, Fell G, Wright M. Principles for high-quality, high-value testing. Evid Based Med 2013 Feb;18(1):5-10. [doi: 10.1136/eb-2012-100645] [Medline: 22740357]
20. MacKinnon KR, Khan N, Newman KM, et al. Introducing novel methods to identify fraudulent responses (sampling with Sisyphus): web-based LGBTQ2S+ mixed-methods study. J Med Internet Res 2025 Mar 17;27:e63252. [doi: 10.2196/63252] [Medline: 40096683]
21. Jang M, Vorderstrasse A. Socioeconomic status and racial or ethnic differences in participation: web-based survey. JMIR Res Protoc 2019 Apr 10;8(4):e11865. [doi: 10.2196/11865] [Medline: 30969173]
22. Kelfve S, Kivi M, Johansson B, Lindwall M. Going web or staying paper? The use of web-surveys among older people. BMC Med Res Methodol 2020 Oct 8;20(1):252. [doi: 10.1186/s12874-020-01138-0] [Medline: 33032531]
23. Corey KL, McCurry MK, Sethares KA, Bourbonniere M, Hirschman KB, Meghani SH. Utilizing internet-based recruitment and data collection to access different age groups of former family caregivers. Appl Nurs Res 2018 Dec;44:82-87. [doi: 10.1016/j.apnr.2018.10.005] [Medline: 30389065]

Abbreviations

AI: artificial intelligence
FRC: family resource center
REAL: Reflect, Expect, Analyze, Label

XSL•FO
RenderX

Edited by A Stone; submitted 02.Oct.2025; peer-reviewed by M Ekwueme, Y Deng; revised version received 11.Jan.2026; accepted 30.Jan.2026; published 05.Mar.2026.

Please cite as:
Willard-Grace R, Klima T, Dedhia M, Lo E, Nisnevich A, Gray A, Henry H
Using Epidemiological Test Diagnostics to Select Fraud Detection Methods: Secondary Analysis of Quantitative Cross-Sectional Survey Data
J Med Internet Res 2026;28:e85161
URL: https://www.jmir.org/2026/1/e85161
doi:10.2196/85161

© Rachel Willard-Grace, Tali Klima, Mansi Dedhia, Emily Lo, Annie Nisnevich, Allison Gray, Holly Henry. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Feasibility of Smartwatch Micro–Ecological Momentary Assessment for Tracking Eating Patterns of Malaysian Children and Adolescents in the South-East Asian Community Observatory Child Health Update 2020: Cross-Sectional Study

Richard Lane[1], PhD; Louise A C Millard[2], PhD; Ruth Salway[3], PhD; Chris J Stone[4], BSc (Hons), RSci; Andy L Skinner[5], PhD; Sophia M Brady[6], PhD; Jeevitha Mariapun[7], PhD; Sutha Rajakumar[7], PhD; Amutha Ramadas[7], PhD; Hussein Rizal[7], MSc; Laura Johnson[3], PhD; Tin Tin Su[7], DrMed; Miranda Elaine Glynis Armstrong[8,9], PhD

[1]Jean Golding Institute, University of Bristol, Bristol, United Kingdom

[2]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom

[3]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom

[4]Integrative Cancer Epidemiology Programme, School of Psychological Science, University of Bristol, Bristol, United Kingdom

[5]Bristol Medical School, Integrative Cancer Epidemiology Programme, University of Bristol, Bristol, United Kingdom

[6]Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, United Kingdom

[7]South East Asia Community Observatory (SEACO), Jeffrey Cheah School of Medicine and Health Sciences, Monash University Malaysia, Subang Jaya, Malaysia

[8]Department for Health, University of Bath, Bath, England, United Kingdom

[9]NIHR Bristol Biomedical Research Centre, University Hospitals Bristol and Weston NHS Foundation Trust, Bristol, United Kingdom

Corresponding Author:
Richard Lane, PhD
Jean Golding Institute
University of Bristol
Beacon House
Queen's Road
Bristol, BS8 1QU
United Kingdom
Phone: 44 117 928 9000
Email: mh19137@bristol.ac.uk

Abstract

Background: Mobile phone ecological momentary assessment (EMA) methods are a well-established measure of eating and drinking behaviors, but compliance can be poor. Micro-EMA (µEMA), which collects information with a single tap response to brief questions on smartwatches, offers a novel application that may improve response rates. To our knowledge, there is no data evaluating µEMA to measure eating habits in children or in low-to-middle-income countries.

Objective: In this study, we investigated the feasibility of micro-EMA to measure eating patterns in Malaysian children and adolescents.

Methods: We invited 100 children and adolescents aged 7-18 years in Segamat, Malaysia, to participate in 2021-2022. Smartwatches were distributed to 83 children and adolescents who agreed to participate. Participants were asked to wear the smartwatch for 8 days and respond to 12 prompts per day, hourly, from 9AM to 8PM, asking for information on their meals, snacks, and drinks consumed. A questionnaire captured their experiences using the smartwatch and µEMA interface. Response rate (proportion of prompts responded to) assessed participants' adherence. We explored associations between response rate with time of day, across days, age, and sex using multilevel binomial logistic regression modeling.

Results: Eighty-two participants provided usable smartwatch data. The median number (IQR) of meals, drinks, and snacks per day was 2 (2-4), 3 (1-5), and 1 (0-2), respectively, on the first day of the study. The median response rate across the study was 68% (IQR 50-83). The response rate decreased across study days from 74% (68-78) on Day 1 to 40% (30-50) on Day 7 (odds ratio [OR] per study day 0.73, 95% CI 0.64-0.83). Response rate was lowest at the start of the day and highest between the hours

XSL•FO
RenderX

of 12 PM and 2 PM. Female participants responded to more prompts than male participants (OR 1.72, 95% CI 1.03-2.86). There was no evidence of differential response by age (OR 0.73, 95% CI 0.41-1.28). Most participants (65%) rated their experience using the smartwatch positively, with 33% saying they were happy to participate in future studies using the smartwatch. For children that did not wear the smartwatch for the full study duration (n=22), discomfort was the most common complaint (41%).

Conclusions: In this study of the feasibility of µEMA on smartwatches to measure eating in Malaysian children, we found the method was acceptable. However, response rates declined across study days, resulting in substantial missingness. Future studies (eg, through focus groups) should explore approaches to improving response to event prompts, trial alternative devices to increase children's comfort, and evaluate revised protocols for reporting of intake events.

(J Med Internet Res 2026;28:e73435) doi:10.2196/73435

KEYWORDS

eating behavior; ecological momentary assessment; EMA; Malaysia; LMIC; children; adolescents; micro-interaction EMA; µEMA; smartwatch

Introduction

Noncommunicable diseases (NCDs) are the most common cause of death worldwide [1]. In Malaysia [2], NCDs particularly impact lower-income households [3]. Therefore, health surveillance in this population is required to better understand policy interventions that may improve health outcomes in Malaysia. Dietary risk factors accounted for 10% of all deaths globally in 2021 [4]; therefore, measuring eating is a crucial component of health surveillance. Traditional methods for measuring eating and dietary intake include food diaries, 24-hour recalls, diet histories, or food frequency questionnaires. While methods relying on memory of past behavior are subject to error like recall bias, prospective methods like diaries are affected by reactivity, where real or reported behavior is altered owing to the process of documenting food intake in real time. Underreporting is common in all existing methods, with an estimated 263 kcal per day typically missing from self-reported intakes compared with objective measures [5]. Underreporting varies with food type and eating occasion, with snacks and snack foods more likely to be left out of a self-reported record [6-9]. Online tool, such as Intake24 (Newcastle University), that guide users through a 24-hour recall process aim to reduce researcher burden in coding data collected. Photographic methods, where participants are asked to take pictures of their meals rather than write down each food and drink along with its portion size, aim to offer a more objective approach to add portion size estimation and reduced participant burden for capturing real-time food intake [10-12]. However, moving 24-hour recall online has not yet altered estimated underreporting [13,14], and issues with remembering to take photos before consuming foods as well as automating the estimation of foods and nutrients [15] in photos mean that outstanding challenges in dietary assessment methods remain [12]. Therefore, the feasibility of using alternative methods needs to be considered.

Ecological momentary assessment (EMA) is the repeated sampling of current behaviors in real-time in a natural environment [16]. EMA has evolved to be primarily delivered using mobile phones (mEMA), which have improved response rates compared with original pen and paper methods [17-19]. There is a large volume of literature on EMA using smartphones (n=796 studies) [20]. While diet is the second most commonly studied topic, it still only accounts for 4% (35/796) of these studies. Studies of diet using EMA in young people are primarily in the United States and Europe, with just 2 studies in Asia, in China, and Taiwan [18,21].

Liao [22] highlighted that response rates and compliance with EMA protocols were rarely reported. Since then, reporting of compliance has improved, but the response latency remains unknown from many studies [21]. Response rates to mEMA of diet are a median of 74% [23], which is similar to mEMA of all topics (mean 75% (IQR 64%-84%) [20]. A review of mEMA for diet in young people (16-30 years) showed response rates mostly exceeded 80% [21], whereas at younger ages poorer responses <80% are more often observed [18]. Lower response rates have also been associated with weekends versus weekdays [21], when participants receive more prompts during the day [17,20], and in males versus females [21].

Smartwatches are an emerging technology for collecting data alongside sensor data using micro-EMA (µEMA) protocols. This captures information using single-tap responses to brief questions, which is suitable for the small screens on these devices [19,23]. In adults, µEMA has been found to yield higher compliance rates despite more frequent sampling than mEMA and is perceived by users as less distracting [24]. Further, the use of µEMA significantly improves response rate (mean 72% vs 82%) but remains rare, with only 12 studies on any topic in any age group [20]. Despite these advantages, some limitations of µEMA have been reported in the literature, including limited battery life and technical problems such as problems with charging [25].

In children and adolescents, the use of pen and paper EMA to measure diet has typically been implemented outside of school hours [18]. Internet-connected devices such as mobile phones are often used for mEMA data collection [26]. These may be less suitable for child and adolescent populations, where 40% of education systems now ban the use of smartphones in school [27]. Further, devices such as smartwatches that can function without an internet connection may be better suited to rural, semirural, and low-resource settings where communication infrastructure may be less well-developed [28]. Therefore, smartwatches offer the potential to implement EMA across the whole day, with the potential for additional advantages such as improved compliance and response rates [20,24]. To our knowledge, only two diet studies involving adults in the United

XSL•FO
RenderX

Kingdom and the United States have reported on EMA with smartwatches, and none have involved children [18,23,29].

Therefore, this study investigates the feasibility of using smartwatch-based µEMA to record eating patterns in Malaysian children and adolescents. The collected µEMA data are used to examine the completeness of the collected data and factors associated with response rates, alongside survey responses assessing participants' experience during the study. Establishing the feasibility of this novel dietary measurement tool is an important first step to inform utility and any required refinement prior to deployment for dietary measurement more widely.

Methods

SEACO-CH20 Study

The South-East Asian Community Observatory (SEACO) health and demographic surveillance cohort is a dynamic cohort of 13,335 households in Segamat, a semirural region in the state of Johor Darul Takzim, Malaysia. The cohort was established in 2012, with surveys, blood tests, and physical measurement data collected from participants. In 2013 and 2018, health surveys were conducted on ~25,000 adults and children, 25,168 in 2013 and 24,710 in 2018.

All households (18,602) in 5 subdistricts, which SEACO operates, were invited to participate in the 2017 census. Altogether 11,617 households (40,184 residents) accepted our invitation. In 2018, participants who were involved in the 2017 census and were older than 5 years were invited to participate in the 2018 health round data survey, to which a total of 24,710 participants agreed. Potential participants from 3 subdistricts were preselected and approached via telephone using the existing health database (HR 2028). Participants' parents were approached via telephone for recruitment before the home visit.

Children and adolescents aged 7-18 years who were part of the SEACO cohort were invited to participate in the SEACO Child Health 2020 update (SEACO-CH20) study; a systematic review of EMA studies in youth recommended 7 as a lower age limit for EMA [26]. The eligibility of households was limited by location due to the safety measures implemented during the COVID-19 pandemic to reduce the risk to participants, households, and fieldworkers. Therefore, the 1993 children and adolescents invited to participate were from only 3 of the 5 SEACO subdistricts (Jabi, Sungai Segamat, and Gemereh) in the Segamat district.

Data collection visits to individual households were performed in person from November 1, 2021, to July 31, 2022. The data were collected as part of a larger study and included surveys, physical measurements, such as height, weight, blood pressure, waist and hip circumference, and blood sample collection. Participants were given wrist-worn Axivity AX6 6-axis accelerometers to monitor their physical activity, which were worn 24 hours per day over 7 days [30]. A random subset of the participants were also given TicWatch C2 (Mobvo) Android smartwatches to record eating and drinking with µEMA as part of this feasibility study, using a smartwatch µEMA system developed within the research team. The smartwatch system used was an adaptation of a µEMA system first used in a study involving high-temporal density longitudinal measurement of alcohol use [31] by a subset of the research team who developed this system. Participants wore these devices over the same 7-day period as the accelerometers. As they are both wrist-worn devices, this may have affected the acceptability of the smartwatch. Participants were briefed on the use of these devices by the data collectors, including how to charge them and how to replace them after charging. The original data collection plan can be seen in Figure S1 in Multimedia Appendix 1. A completed Checklist for Reporting Ecological Momentary Assessment Studies (CREMAS) [22] can be found in Multimedia Appendix 2.

Study Participants

Participants for the SEACO-CH20 study were selected from the larger SEACO cohort. Parents of participants provided consent for their children to participate in SEACO-CH20. A random subsample of 100 participants were each invited to wear a smartwatch.

Data Collection

SEACO-CH20 fieldworkers performed 2 home visits to collect the data. The smartwatch was distributed on the initial visit, and the participants were briefed on how to use and charge it. They were instructed to wear the device for the next 8 days, on "the wrist that [they] use to write." The smartwatches were collected during the second home visit, and the participants were asked to complete a questionnaire on their experience with the devices. Questionnaires assessed the participants' attitudes toward several aspects of the smartwatch study, including ease of use, their attitude toward charging, and their overall experience. Since children as young as 10 were asked to complete the questionnaires, a reduced set of acceptance questions was used to reduce burden. This was based on similar pilot work using novel methods in the ALSPAC G2 study [32] and included the following questions:

- Overall, how would you rate your experience of using the smartwatch during the study, on a scale from 1 (didn't like it at all) to 5 (really liked it)?
- If you were asked to use the smartwatch again in another study, would you participate?
- How many days in total did you wear the smartwatch for?
- If you wore the smartwatch for less than 8 days, what were the main reasons for not wearing it longer?

Parents of participants aged 7-9 years completed the survey on behalf of their children, while participants aged 10 years and older filled out the survey themselves. The full text of the survey can be found in Figure S2 in Multimedia Appendix 1.

Smartwatch µEMA Questions

During the study, the smartwatch prompted participants once every hour to enter any food or drink that they had consumed in the last hour. These prompts were scheduled to appear once every hour from 9 AM to 8 PM, so participants were expected to interact with the smartwatch 12 times throughout the day. We chose this hourly prompt frequency to maximize the chance that eating and drinking events were less likely to be missed and to capture more fine-grained temporal patterns in eating

XSL•FO
RenderX

behavior. As this was a feasibility study, this choice was justified, given that µEMA has been shown to improve compliance despite more prompts than mEMA in adults [26]. The smartwatch interface included the following 5 questions that the participants completed for each item consumed:

1. Have you had any food or drink in the last hour? Options: yes, no
2. What did you have? Options: meal, snack, drink
3. What size was it? Options: small, medium, large
4. What did you use to eat? Options: hands, fork/spoon, chopsticks
5. Where were you? Options: home, school, elsewhere

A possible future use of this methodology, once fully refined, could include automated eating detection [33]. Therefore, we included a question on the type of cutlery used ("What did you use to eat?"), as lack of information on utensil type has been highlighted as a limitation of some datasets used for algorithm development related to automated eating detection [34].

After entering this information for one item consumed, they were asked, "Any more food or drink to record?" and could then start again to add another entry. Therefore, each consumption entry either indicates that the participant did not eat or drink in the last hour or contains the answers to the above questions for a particular meal, drink, or snack, linked to an hour period within a day. If participants ignored the prompts, they would receive a reminder prompt after 1 minute; if they continued to ignore the prompt for a further 1 minute, the prompt would disappear and "no response" would be recorded by the smartwatch.

Participants could choose "back" on each question screen to return to a previous question and update their response. However, after submitting their answers for a particular item (ie, completing the "where were you" question for that item), they would not be able to return to that entry.

An additional prompt (the "catch-up") was scheduled every morning at 8 AM asking if they had consumed any food or drink on the previous day that had not been recorded on the smartwatch. If they indicated "yes," they were asked the same questions as above. Catch-up entries did not have an associated eating time but were labeled as catch-up–type events, indicating that they applied to the previous day.

The smartwatch study was co-created and piloted with the Malaysian research team and the English was translated into Malay for use on the smartwatch. All the original data collection was in Malay. The smartwatch protocol, including the prompts and possible responses, can be seen in Tables S1-S3 in Multimedia Appendix 1.

Smartwatch Data Cleaning

Smartwatches were distributed by fieldworkers partway through the day, and µEMA responses on this distribution day were removed from analyses. The study period is taken to be the subsequent 7 days after this distribution day.

The version of the EMA software we used did not save the hour period to which each entry belonged. Therefore, we needed to infer this from the submission timestamp, the date and time a particular entry was submitted. As entries for the same hour period are submitted one after the other, we used a time window to group nearby entries into a single "eating event," which is intended to capture the participants' responses to one prompt. A 30-minute window was chosen to group nearby prompts, as we expect this to collect entries from the same eating event without grouping prompts from adjacent hours. Previous work from diet diaries suggests that 30 minutes is a reasonable cut-point to distinguish independent eating occasions [35]. Occasionally, there may be participants with more than 12 eating events per day, for example, if they took more than 30 minutes to finish responding to a prompt. This occurred on 26 occasions, less than 5% of the total 574 (82 participants multiplied by 7 days) study days.

The µEMA data used was restricted to the 7 days after the distribution day. During the data review, we identified an issue with the collected data where there were sometimes multiple identical entries for a given intake event due to an issue with the µEMA software. Therefore, duplicate entries were identified as any pair of entries with identical contents (same meal type, portion size, utensil, and location), for the same hour period, and entered within 5 minutes of each other. The first such entry was kept in each case. Around 588 duplicate responses were removed of 10,539. Data cleaning was performed in Python (version 3.10.0; Python Software Foundation).

Response Rate

The response rate was calculated as the proportion of prompts responded to (with either at least one item consumed or an entry stating that they did not eat or drink anything in the previous hour). The response rate tells us the extent that participants engaged with the smartwatch app throughout the day but not the extent that the data entered are complete, that is, whether all intake events were recorded. We therefore summarized the number of each type of meal entry (meal, drink, snack, or no food or drink) submitted per day across our sample. For these summaries we included only participants who took part in the study outside the Ramadan fasting period (April 3, 2022-May 1, 2022; n=67), since fasting participants are likely to enter fewer eating events during the day. The mean of participants' response rates per day was recorded, and the median and quartiles of these were reported.

Attrition from the study was examined by identifying the last day each participant responded to any smartwatch prompt; participants who had ceased responding to the prompts are referred to as "inactive."

Statistical Analyses

We summarized response rates to each individual prompt of the smartwatch µEMA and the participants' experiences based on the survey questions. We used mean for continuous variables, n (%) for categorical variables, and median and IQR for ordinal or nonnormally distributed continuous variables.

We used a mixed-effect logistic regression model for the response (yes or no) to an individual prompt on a specific study day of data collection for each participant. A fixed effect term was included for study day (from the first to the seventh day) as a continuous linear trend. The time of day was also included

XSL•FO
RenderX

as a fixed effect to capture nonlinearity in response throughout the day, grouping the prompts by the nearest hour as follows to decrease the number of parameters in the model:

- Morning (9-11 AM)
- Lunchtime (12-2 PM)
- Afternoon (3-5 PM)
- Evening (6-8 PM)

Random intercept and random slope terms were included for study day within each participant. Estimates are provided as odds ratios (OR) and 95% CIs, interpreted as the multiplicative change in the odds of a participant responding to an individual prompt. The degree of difference between participants was summarized in the intraclass correlation coefficient.

To evaluate if changes in participation across wear days differed in boys versus girls, we repeated this base model, adding a fixed term for sex and an interaction term between sex and study day. Similarly, we explored differences by age group (Malaysian primary school age: 7-12 years versus secondary school age: 13-18 years) by adding a fixed term for age group and an interaction term between age group and study day to the base model.

Analyses were performed in Python version 3.10.0 and R (version 4.2.2; R Core Team) [36]. All of our analysis code is publicly available [37]. Git tag 3.0 (The Git Project) corresponds to the version of the analyses presented here.

Ethical Considerations

Written informed consent was obtained from parents or guardians on behalf of the participants. Children and adolescents were also asked to provide their written assent to participate in the study. Ethical approval was obtained from the Monash University Human Research Ethics Committee on March 17, 2020 (Project ID: 23271) and the University of Bristol REC Case no. 2020–4208 (ID nr: 1304255) prior to any data collection. The study was conducted in accordance with the Declaration of Helsinki for experiments involving humans.

Participants were given a token worth up to RM25 to compensate for their time participating in the study. This was divided into RM5 for completion of each of the following components: (1) questionnaires, (2) health check, (3) blood sample, (4) activity monitor, and (5) smartwatch. They were also given a free health screen and a direct referral to the government primary health care clinic if they were identified as high risk.

Study data have been deidentified and can be freely requested from SEACO, Monash University Malaysia Institutional Data Access at “mum.seaco@monash.edu” for researchers meeting the criteria for access to confidential data. Please refer to the web resource hosted on Monash University’s website [38] for more information.

Results

Participants

A flowchart showing the study participants can be seen in Figure 1. Parents of 728 participants consented to their children’s participation in SEACO-CH20, of which 626 provided demographic (age, sex, and ethnicity) and accelerometer data for the larger study. Of these, 100 participants were randomly invited to wear a smartwatch for this smaller feasibility study. Of the 100 participants invited to participate in the smartwatch substudy, 83 participants agreed. The reasons for nonparticipation included concern the device was not comfortable (n=3) and allergies (n=2). The remaining participants rejected the smartwatch without comment. One further participant accepted the smartwatch study but removed the EMA app from the watch during the study period, rendering their data unrecoverable, resulting in 82 participants who provided smartwatch data.

The sex, ethnicity, and age breakdown for all participants who took part in the smartwatch study can be seen in Table 1.

XSL•FO
RenderX

Figure 1. Study flowchart. Eligible participants were selected from the SEACO Health Round Survey 2018 (SEACO HR-2018) cohort. Reasons for rejecting the smartwatch study included concern about discomfort and allergies. SEACO-CH20: South-East Asian Community Observatory Child Health 2020; SEACO HR-2018: South-East Asian Community Observatory Health Round Survey 2018;.

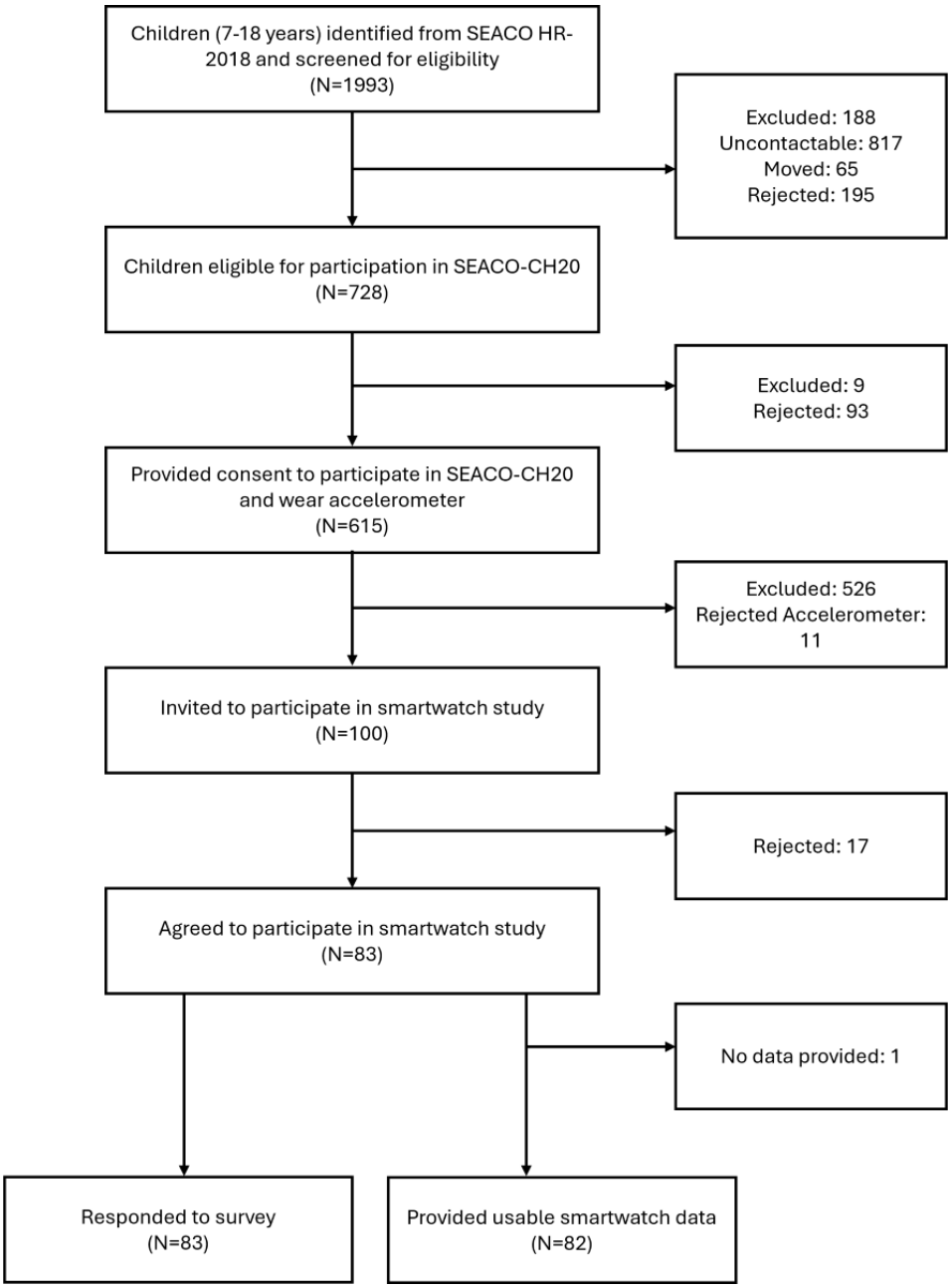

XSL•FO
RenderX

Table 1. Summary of participant demographics (N=83).

Participant characteristic	Smartwatch participants, n (%)
Sex	
Female	53 (64)
Male	30 (36)
Ethnicity	
Malay	73 (88)
Non-Malay	10 (12)
Age (years)	
7-12	24 (29)
13-15	28 (34)
16-18	31 (37)

Smartwatch Responses

The median prompt response rate was 69% (IQR 52%-82%). The number of participants who became inactive on each day can be seen in Figure 2. The majority (55/82, 67%) of participants were active until day 7, that is, they responded to at least 1 prompt on day 7.

Figure 2. The number of participants who became inactive on each day (N=82), that is, having responded to no µEMA prompts after this day. All participants were active for at least one day.

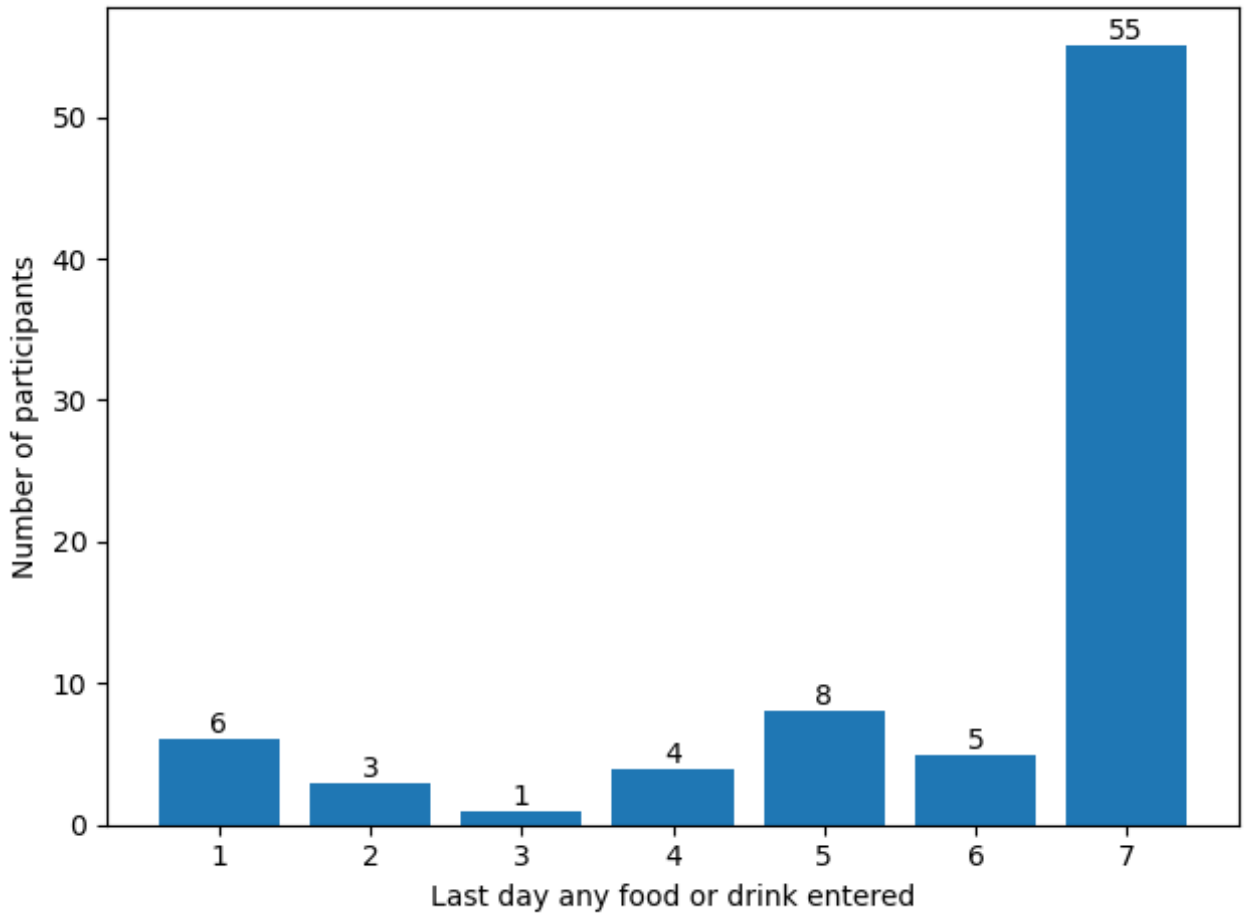

The median and IQR in the number of entries of each type made by each participant per day are summarized in Table 2. Fifteen participants took part (at least partially) during Ramadan and so were excluded from these summaries. Only a minority of intake events were submitted as catch-up entries (N=125 catch-up entries versus 4705 noncatch-up).

XSL•FO
RenderX

Table 2. The median and IQR of the number of noncatch-up entries per day per participant, for participants whose study period did not intersect with Ramadan (N=67).

Study day	Meal, median (IQR)	Drink, median (IQR)	Snack, median (IQR)
1	2 (2-4)	3 (1-5)	1 (0-2)
2	2 (1-3)	2 (1-5)	1 (0-2)
3	2 (1-3)	2 (0-3)	0 (0-1)
4	2 (1-3)	1 (0-3)	1 (0-1)
5	1 (0-2)	1 (0-3)	0 (0-1)
6	1 (0-2)	1 (0-2)	0 (0-1)
7	1 (0-2)	0 (0-2)	0 (0-1)

Response Rate Across and Within Study Days

The response rate for individual prompts had a median (IQR) of 67% (50-83). The response rate on each day ranged from 83% (66-92) on day 1 to 58% (33-75) on day 7.

Figure 3 shows the response rate with study day and time. The response rate decreased across study days (OR for each additional day of the study: 0.73 (95% CI 0.64-0.83). The response rate was lowest at the beginning of the day; the OR and 95% CIs are summarized in Table 3. The intraclass correlation coefficient was 0.207, which indicates that approximately 21% of the total variance in prompt response behavior was attributable to between-participant differences.

Figure 3. Participants' response rates against study day (left) and time of day (right; N=82).

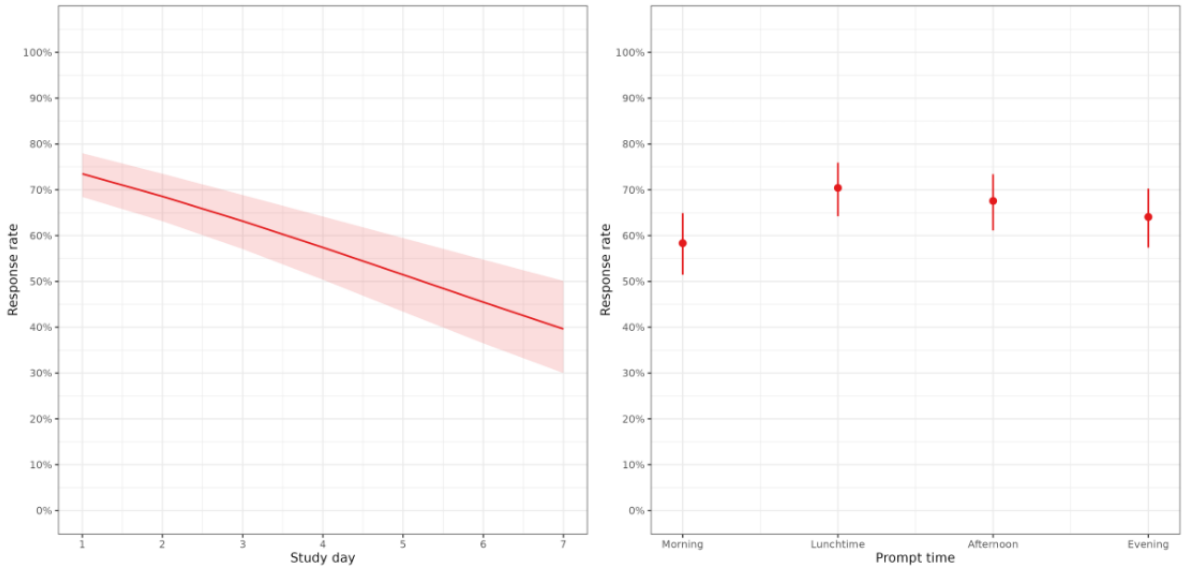

Table 3. The odds ratios for responses at different times of day, taking breakfast time (9-11AM) as the reference level. The response rate was lowest at the beginning of the day.

Time	Odds ratio (95% CI)
Breakfast (9-11 AM)	1[a] (0-0)
Lunchtime (12-2 PM)	1.69 (1.43-2.12)
Afternoon (3-5 PM)	1.49 (1.25-1.76)
Evening (6-8 PM)	1.27 (1.07-1.51)

[a]Reference level.

The results of analyses estimating differences due to sex and age are shown in Figure 4. Girls responded more often to the µEMA prompts compared with boys (OR 1.71, 95% CI 1.03-2.84). However, the daily patterns were similar for both sexes (interaction term OR 1.07, 95% CI 0.93-1.23). Response rate did not differ between age groups (OR 0.73, 95% CI 0.42-1.27), and daily response patterns were similar for the 2 age groups (study day-by-age interaction OR 1.11, 95% CI 0.95-1.29).

XSL•FO
RenderX

Figure 4. Association of micro-interaction ecological momentary assessment (μEMA) prompt response with study day stratified by sex (left) and age (right).

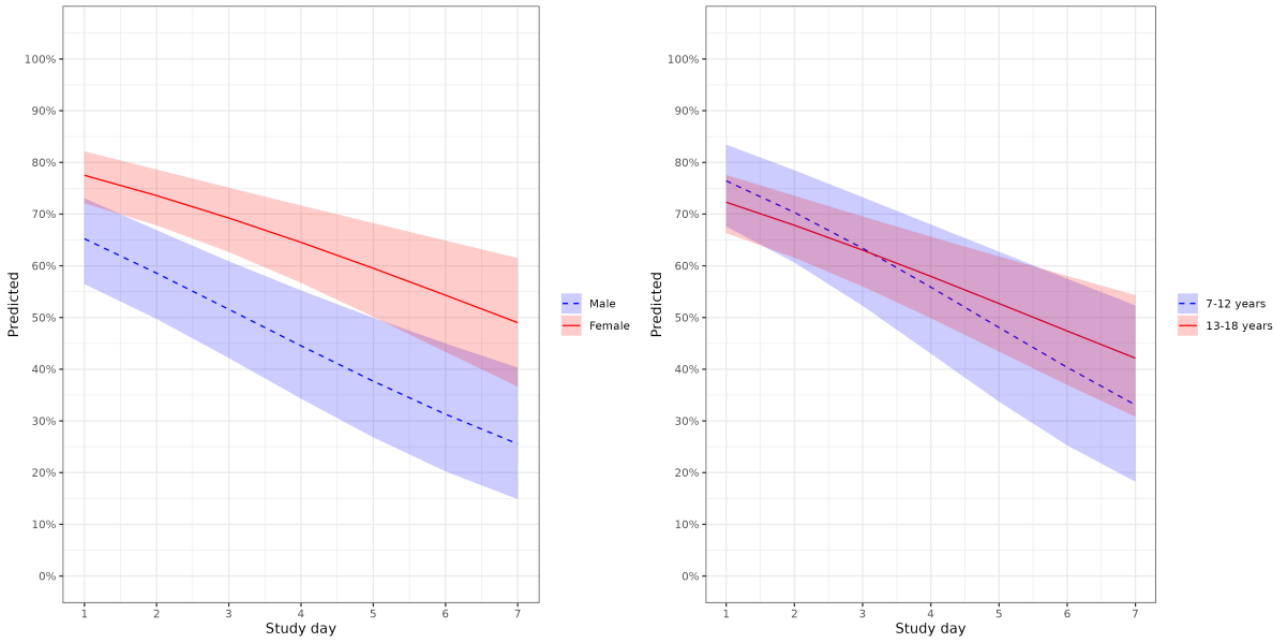

Evaluating Participants' Survey Responses on Acceptability

A summary of responses to questions about smartwatch acceptability and wear time is provided in Table 4.

A total of 54 out of 83 (65%) participants rated their experience using the smartwatch positively (a rating of 4 or 5 out of 5), 20 out of 83 (24%) gave a neutral rating (3 out of 5), and 8 out of 83 (10%) rated it negatively (1 or 2 out of 5). In addition, 27 out of 83 (33%) participants said that they would be happy to participate in future studies using the smartwatch, while 39 out of 83 (47%) said maybe, and 12 out of 83 (14%) said no. The majority of participants who responded (61/83, 73.5%) reported wearing the smartwatch for the entire duration of the study (8 days).

Table 4. The survey questions. Participants were directed to wear the smartwatch on the day that the watch was distributed and for the 7 subsequent days, making 8 days total. Missing data and participants who refused to respond are not included (N=83).

Question and response	Participants, n (%)
Overall, how would you rate your experience of using the smartwatch during the study, on a scale from 1 (didn't like it at all) to 5 (really liked it)?	
Negative (1 or 2)	8 (10)
Neutral (3)	20 (24)
Positive (4 or 5)	54 (65)
If you were asked to use the smartwatch again in another study, would you participate?	
No	12 (14)
Maybe	39 (47)
Yes	27 (33)
How many days in total did you wear the smartwatch for?	
5 or fewer	7 (9)
6	7 (8)
7	8 (10)
8	61 (73)

For those who reported that they did not wear the smartwatch for the entire duration of the study (22 participants), the most common reason was that they did not find it comfortable to wear (9/22, 41%). Other reasons included forgetting, that they did not see the benefit when they could not see the data, they were forbidden to wear it by school, and it ran out of battery (all 3 or fewer responses).

Summaries of the participants' responses to the remaining survey questions can be found in Tables S4-S8 and Figure S3 in Multimedia Appendix 1. Further summary statistics, including

XSL•FO
RenderX

catch-up events and participants who took part during Ramadan, can be found in Tables S9-S11 in Multimedia Appendix 1.

Discussion

Principal Findings

In this feasibility study of a smartwatch-based µEMA method to collect data on eating habits over 7 days in Malaysian children, we found that most participants (55/83, 67%) remained responsive to prompts up to the last day of the study. Participants were least likely to respond to prompts between 9 and 11 AM and most likely between 12 and 2 PM. The intraclass correlation coefficient was 20.7%, suggesting that while some variation in response pattern is attributable to between-participant differences, the majority of the variation (79.3%) was due to within-participant differences. The response rate dropped off day-on-day and was higher for female than male participants; no association was found between participant age group and response rate.

Our average response rate of 69% was lower than the average of 78% found in a meta-analysis of EMA in children and adolescents, including studies that prompted between 2 and 9 times daily [17]. That study found that prompting participants more often had a large negative effect on completion rate, which is further supported by Kraft et al [20], which found a negative correlation of –0.12 between increased number of prompts and response rate (P=.009). Participants in our study were prompted 12 times a day, plus an additional catch-up prompt in the morning. We justified our original prompt frequency choice, as µEMA has been shown in adults to improve compliance despite a higher number of prompts than mEMA. However, in our study using a child and adolescent population, it is likely a higher prompt frequency may have had a negative effect on response rate, especially in the case of repeated "No food/drink" entries. It has been reported [17] in nonclinical studies that "a higher average compliance rate was observed in studies that prompted participants 2-3 times daily (91.7%) compared with those that prompted participants more frequently (4-5 times: 77.4%; 6+ times: 75.0%)." This suggests that compliance may be improved by prompting participants less frequently, for example, by having 3 prompts daily at 11 AM, 3 PM, and 7 PM, although longer time intervals increase the reliance on memory, potentially affecting the completeness of recorded consumption events. It is also possible that our lower response rate might have been related to the number of questions asked in our µEMA protocol. A study that compared the deployment of 6 back-to-back multiple-choice questions delivered via a smartwatch versus a mobile phone found no difference in compliance between these 2 modalities. However, compliance was improved when single questions requiring a one-touch response were asked via a smartwatch, despite an increase in prompt frequency [39]. While on average, the participants (65%) rated the study protocol positively (either 4 or 5 out of 5), the response rate fell day-on-day. Participants were less likely to respond to prompts at the beginning or end of the day, compared with the middle of the day. Focus group or interview discussions were not feasible in this study due to COVID-19 restrictions but should be explored in future studies to determine the reasons for missing event prompts and nonresponses, which may include forgetting or being involved in a competing activity when the prompt is sent [40].

Female participants had a higher response rate than male participants, consistent with previous findings [21,41]. There was little evidence of difference in the relationship between response and study day for male versus female participants. An analysis of the SEACO-CH20 accelerometer dataset [30] found that a similar proportion of males and females had usable accelerometer data, suggesting that this was specific to the smartwatches rather than a difference with wrist-worn devices generally. Little evidence of a difference was found between response rates in the 7-12 years old and 13-18 years old age groups.

Although the subjective indicators suggested that most participants enjoyed wearing the smartwatch, only a minority of participants (27/83, 33%) indicated that they would be willing to participate in a similar study again; 39 out of 83 (47%) responded "Maybe." Potential changes to the protocol that may improve compliance could include only wearing the smartwatch instead of both the smartwatch and accelerometer.

Strengths and Limitations

This is the first study exploring the feasibility of using smartwatch-based EMA in a population of children and adolescents from a low-to-middle-income country. This study was part of the SEACO study, using the SEACO-CH20 dataset, and lays the foundation for an improved understanding of the potential for wearable devices for measuring relationships between eating and cardiometabolic health. Data on 24-hour eating behaviors are important for informing future policy that may reduce cardiometabolic risk among children and adolescents and prevent progression to cardiometabolic disease in adulthood. While a food frequency questionnaire was completed as part of the larger SEACO-CH20 study and reported elsewhere [42], this current study has assessed an alternative for recording behavior in real time.

However, this study did have some limitations. The number of questions in prompts may have affected compliance and should be discussed with participants to optimize the protocol for future studies in this population. It has been suggested [17] that compliance can be improved by incentivizing participants with a monetary reward or raffle entries. Since this study concerns young people, one potential incentive method could be to gamify the EMA process using a level-up or promotion system in the app [43]. Previous studies have explored adding an end-of-day catch-up prompt, which has been found to improve the reporting of dinner [40]. Replacing the morning catch-up prompt with one in the evening may improve response rate, especially given that we found that participants were more likely to respond to prompts in the evening than in the morning. Future studies may additionally consider using the catch-up entries to impute missed entries on the previous day, which could give more complete data. Another suggestion could be to incorporate a short period of training to improve response rates, where responses are monitored in real time by researchers and participants are prompted directly by researchers if missing responses are

XSL•FO
RenderX

common. Such an approach has been used to improve the accuracy of real-time food photography methods [12].

The questionnaires used for acceptability and acceptance were not standard because we were motivated to use a reduced set of questions (that have been used in a previous publication [31]) to reduce the burden on the young participants. Future studies should consider if the emerging, more standard approaches to exploring acceptability and acceptance for wearable devices (eg, those based on Technology Acceptance Models) have been developed to the point at which varying levels of participant burden can be accommodated.

Unbalanced statistics limited our ability to assess differences across age, sex, and ethnic group. The larger SEACO-CH20 accelerometer study [34] from which participants for this study were selected had more balanced statistics (49% female, 67% Malay, and 44% <13 years old), which suggests that the cohort used for smartwatch data may not represent the overall SEACO-CH20 cohort. In particular, we only had 4 participants aged 7-9 years, so further studies are required to better understand the feasibility of dietary µEMA in the younger participants. Participants in this study began wearing the devices on different days of the week, and it is possible that the day of the week could affect participation; for example, whether it is a weekend or weekday. A lower response rate on weekends has been previously documented by Battaglia [21]. We did not account for study start day due to the small sample size, and because schooling was disrupted throughout the study period due to the COVID-19 pandemic [44].

An issue with entry duplication meant that some entries may have been removed that were actual events, not due to the software issue. This bug with the smartwatch software has since been fixed. Additionally, only the response time of the participant was recorded, and not the time that the prompt was sent. This means we had to infer which hour window the entries corresponded to; this could be programmed in the software.

Discomfort was the most common reason for nonwear cited by the participants, which may be unique to our study protocol that required participants to wear 2 wrist-worn devices on the same arm. Furthermore, the smartwatch used in our study was not specifically designed to fit smaller children. Efforts to make smartwatches less intrusive, for example, by making them smaller, may further improve response rate and study uptake.

Ramadan, a culturally important event in Malaysian society, which includes fasting in some population groups, took place during the course of the data collection period. This is likely to have affected the eating behaviors of participants who took part during this time. To ensure that the number of EMA entries was not influenced by Ramadan, we excluded participants from our analyses who wore the smartwatch during the Ramadan period, thereby further limiting our sample size.

Conclusions

This study extends previous eating behavior studies by exploring the use of µEMA in a population of children and adolescents in Malaysia and is the first such study to do so. Willingness to take part in the µEMA study was high, but poor response rates suggest that the number of questions asked per prompt or the high number of prompts per day may be too burdensome. While our smartwatch-based EMA app was largely based on the µEMA methods originally developed by Intille et al [24], a key aspect of true µEMA implementation is the presentation of only one question at a time. In our approach, we chained questions to capture details on food and drink type, size, and consumption context, making it more accurately described as a modified µEMA. Further work is needed to explore different µEMA variations, including using fewer questions and/or fewer prompts, and identify devices that may be more comfortable for child and adolescent participants. The growing use of smartwatches amongst children, particularly in Southeast Asia may offer more opportunities for further study [45].

Acknowledgments

For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any author accepted manuscript version arising from this submission.

The authors would like to express their appreciation to the SEACO Field Teams and survey participants. The SEACO [46] funded the research detailed in this paper. The authors' opinions, however, are their own, and there is no real or implied sponsorship from SEACO.

Funding

The Medical Research Council (MR/T018984/1) and the Ministry of Higher Education/UK-MY Joint Partnership on Non-Communicable Diseases (2019/MR/T018984/), both provided funding in support of this research. The SEACO health and demographic surveillance system is supported by Monash University. The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript. Sophia M Brady is funded by the National Institute for Health and Care Research (NIHR) Applied Research Collaboration (ARC) North East and North Cumbria (NENC; NIHR200173). The NIHR Bristol Biomedical Research Centre funds Miranda EG Armstrong (NIHR203315). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

XSL•FO
RenderX

Data Availability

Data cannot be shared publicly for confidentiality and ethical reasons. Deidentified data are available and can be freely requested from the South East Asia Community Observatory, Monash University Malaysia Institutional Data Access at "mum.seaco@monash.edu" for researchers who meet the criteria for access to confidential data. For more information, please refer to [38].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary information, including the original data collection plan; survey instruments; smartwatch EMA protocols and prompts; and additional smartwatch and survey tables and figures.
[DOC File , 321 KB - jmir_v28i1e73435_app1.doc]

Multimedia Appendix 2
Checklist for Reporting Ecological Momentary Assessment Studies (CREMAS).
[PDF File (Adobe PDF File), 105 KB - jmir_v28i1e73435_app2.pdf]

References

1. Invisible numbers: the true extent of noncommunicable diseases and what to do about them. World Health Organization. 2022. URL: https://www.who.int/publications/i/item/9789240057661 [accessed 2025-03-20]
2. Low W, Lee Y, Samy AL. Non-communicable diseases in the Asia-Pacific region: prevalence, risk factors and community-based prevention. Int J Occup Med Environ Health 2015;28(1):20-26. [doi: 10.2478/s13382-014-0326-0] [Medline: 26159943]
3. Isamail N, Said R, Ismail NW, Haron SA. Non-communicable diseases impact low-income households in Malaysia. Malays J Med Sci 2024;31(1):124-139 [FREE Full text] [doi: 10.21315/mjms2024.31.1.11] [Medline: 38456101]
4. Diet. Institute for Health Metrics and Evaluation. URL: https://www.healthdata.org/research-analysis/health-risks-issues/diet-research-library [accessed 2025-03-25]
5. Mehranfar S, Jalilpiran Y, Jafari A, Jayedi A, Shab-Bidar S, Speakman JR, et al. Validity of dietary assessment methods compared with doubly labeled water in children: a systematic review and meta-analysis. Obes Rev 2024;25(9):e13768. [doi: 10.1111/obr.13768] [Medline: 38783784]
6. Poppitt SD, Swann D, Black AE, Prentice AM. Assessment of selective under-reporting of food intake by both obese and non-obese women in a metabolic facility. Int J Obes Relat Metab Disord 1998;22(4):303-311. [doi: 10.1038/sj.ijo.0800584] [Medline: 9578234]
7. Garden L, Clark H, Whybrow S, Stubbs RJ. Is misreporting of dietary intake by weighed food records or 24-hour recalls food specific? Eur J Clin Nutr 2018;72(7):1026-1034 [FREE Full text] [doi: 10.1038/s41430-018-0199-6] [Medline: 29789710]
8. Chan V, Davies A, Wellard-Cole L, Lu S, Ng H, Tsoi L, et al. Using wearable cameras to assess foods and beverages omitted in 24 hour dietary recalls and a text entry food record app. Nutrients 2021;13(6):1806 [FREE Full text] [doi: 10.3390/nu13061806] [Medline: 34073378]
9. Gemming L, Ni Mhurchu C. Dietary under-reporting: what foods and which meals are typically under-reported? Eur J Clin Nutr 2016;70(5):640-641. [doi: 10.1038/ejcn.2015.204] [Medline: 26669571]
10. Simpson E, Bradley J, Poliakov I, Jackson D, Olivier P, Adamson AJ, et al. Iterative development of an online dietary recall tool: INTAKE24. Nutrients 2017;9(2):118 [FREE Full text] [doi: 10.3390/nu9020118] [Medline: 28208763]
11. Gemming L, Utter J, Ni Mhurchu C. Image-assisted dietary assessment: a systematic review of the evidence. J Acad Nutr Diet 2015;115(1):64-77. [doi: 10.1016/j.jand.2014.09.015] [Medline: 25441955]
12. Höchsmann C, Martin CK. Review of the validity and feasibility of image-assisted methods for dietary assessment. Int J Obes (Lond) 2020;44(12):2358-2371 [FREE Full text] [doi: 10.1038/s41366-020-00693-2] [Medline: 33033394]
13. Park Y, Dodd KW, Kipnis V, Thompson FE, Potischman N, Schoeller DA, et al. Comparison of self-reported dietary intakes from the automated self-administered 24-h recall, 4-d food records, and food-frequency questionnaires against recovery biomarkers. Am J Clin Nutr 2018;107(1):80-93 [FREE Full text] [doi: 10.1093/ajcn/nqx002] [Medline: 29381789]
14. Foster E, Lee C, Imamura F, Hollidge SE, Westgate KL, Venables MC, et al. Validity and reliability of an online self-report 24-h dietary recall method (Intake24): a doubly labelled water study and repeated-measures analysis. J Nutr Sci 2019;8:e29 [FREE Full text] [doi: 10.1017/jns.2019.20] [Medline: 31501691]
15. Skinner A, Toumpakari Z, Stone C, Johnson L. Future directions for integrative objective assessment of eating using wearable sensing technology. Front Nutr 2020;7:80 [FREE Full text] [doi: 10.3389/fnut.2020.00080] [Medline: 32714939]

XSL•FO
RenderX

16. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
17. Wen CKF, Schneider S, Stone AA, Spruijt-Metz D. Compliance with mobile ecological momentary assessment protocols in children and adolescents: a systematic review and meta-analysis. J Med Internet Res 2017;19(4):e132 [FREE Full text] [doi: 10.2196/jmir.6641] [Medline: 28446418]
18. Mason TB, Do B, Wang S, Dunton GF. Ecological momentary assessment of eating and dietary intake behaviors in children and adolescents: a systematic review of the literature. Appetite 2020;144:104465 [FREE Full text] [doi: 10.1016/j.appet.2019.104465] [Medline: 31541670]
19. Berkman ET, Giuliani NR, Pruitt AK. Comparison of text messaging and paper-and-pencil for ecological momentary assessment of food craving and intake. Appetite 2014;81:131-137 [FREE Full text] [doi: 10.1016/j.appet.2014.06.010] [Medline: 24930596]
20. Kraft R, Reichert M, Pryss R. Mobile crowdsensing in ecological momentary assessment mHealth studies: a systematic review and analysis. Sensors (Basel) 2024;24(2):472 [FREE Full text] [doi: 10.3390/s24020472] [Medline: 38257567]
21. Battaglia B, Lee L, Jia SS, Partridge SR, Allman-Farinelli M. The use of mobile-based Ecological Momentary Assessment (mEMA) methodology to assess dietary intake, food consumption behaviours and context in young people: a systematic review. Healthcare (Basel) 2022;10(7):1329 [FREE Full text] [doi: 10.3390/healthcare10071329] [Medline: 35885855]
22. Liao Y, Skelton K, Dunton G, Bruening M. A systematic review of methods and procedures used in ecological momentary assessments of diet and physical activity research in youth: an adapted STROBE Checklist for Reporting EMA Studies (CREMAS). J Med Internet Res 2016;18(6):e151 [FREE Full text] [doi: 10.2196/jmir.4954] [Medline: 27328833]
23. Schembre SM, Liao Y, O'Connor SG, Hingle MD, Shen S, Hamoy KG, et al. Mobile ecological momentary diet assessment methods for behavioral research: systematic review. JMIR Mhealth Uhealth 2018;6(11):e11170 [FREE Full text] [doi: 10.2196/11170] [Medline: 30459148]
24. Intille S, Haynes C, Maniar D, Ponnada A, Manjourides J. µEMA: microinteraction-based Ecological Momentary Assessment (EMA) using a smartwatch. Proc ACM Int Conf Ubiquitous Comput 2016;2016:1124-1128 [FREE Full text] [doi: 10.1145/2971648.2971717] [Medline: 30238088]
25. Beukenhorst AL, Howells K, Cook L, McBeth J, O'Neill TW, Parkes MJ, et al. Engagement and participant experiences with consumer smartwatches for health research: longitudinal, observational feasibility study. JMIR Mhealth Uhealth 2020;8(1):e14368 [FREE Full text] [doi: 10.2196/14368] [Medline: 32012078]
26. Heron KE, Everhart RS, McHale SM, Smyth JM. Using mobile-technology-based Ecological Momentary Assessment (EMA) methods with youth: a systematic review and recommendations. J Pediatr Psychol 2017;42(10):1087-1107. [doi: 10.1093/jpepsy/jsx078] [Medline: 28475765]
27. To ban or not to ban? UNESCO. URL: https://www.unesco.org/en/articles/smartphones-school-only-when-they-clearly-support-learning [accessed 2025-11-04]
28. Dawood S. Digital divide and poverty eradication in the rural region of Northern Peninsular Malaysia. Indones J Geogr 2019;51(2):172.
29. Maugeri A, Barchitta M. A systematic review of ecological momentary assessment of diet: implications and perspectives for nutritional epidemiology. Nutrients 2019;11(11):2696 [FREE Full text] [doi: 10.3390/nu11112696] [Medline: 31703374]
30. Brady SM, Salway R, Mariapun J, Millard L, Ramadas A, Rizal H, et al. Accelerometer-measured 24-hour movement behaviours over 7 days in Malaysian children and adolescents: a cross-sectional study. PLoS One 2024;19(2):e0297102 [FREE Full text] [doi: 10.1371/journal.pone.0297102] [Medline: 38377079]
31. Stone C, Adams S, Wootton RE, Skinner A. Smartwatch-based ecological momentary assessment for high-temporal-density, longitudinal measurement of alcohol use (AlcoWatch): feasibility evaluation. JMIR Form Res 2025;9:e63184 [FREE Full text] [doi: 10.2196/63184] [Medline: 40131326]
32. Lawlor DA, Lewcock M, Rena-Jones L, Rollings C, Yip V, Smith D, ALSPAC Executive. The second generation of the Avon Longitudinal Study of Parents and Children (ALSPAC-G2): a cohort profile. Wellcome Open Res 2019;4:36 [FREE Full text] [doi: 10.12688/wellcomeopenres.15087.2] [Medline: 31984238]
33. Mekruksavanich S, Jantawong P, Jitpattanakul A. Smartwatch-based eating detection and cutlery classification using a deep residual network with squeeze-and-excitation module. Washington DC: IEEE; 2022 Presented at: 45th International Conference on Telecommunications and Signal Processing (TSP); 2022 July 13-15; Prague, Czech Republic. [doi: 10.1109/tsp55681.2022.9851333]
34. Stankoski S, Jordan M, Gjoreski H, Luštrek M. Smartwatch-based eating detection: data selection for machine learning from imbalanced data with imperfect labels. Sensors (Basel) 2021;21(5):1902 [FREE Full text] [doi: 10.3390/s21051902] [Medline: 33803121]
35. Ibacache F, Northstone K, Zou M, Johnson L. Investigating eating architecture and the impact of the precision of recorded eating time: a cross-sectional study. Am J Clin Nutr 2025;121(3):685-694 [FREE Full text] [doi: 10.1016/j.ajcnut.2025.01.012] [Medline: 39805560]
36. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2024.

XSL•FO
RenderX

37. JGIBristol / seaco_smartwatch_paper. Github. 2025. URL: https://github.com/JGIBristol/seaco_smartwatch_paper [accessed 2026-01-06]
38. How to Collaborate with SEACO. MONASH University. 2022. URL: https://www.monash.edu.my/seaco/research-and-training/how-to-collaborate-with-seaco [accessed 2025-11-18]
39. Ponnada A, Haynes C, Maniar D, Manjourides J, Intille S. Microinteraction ecological momentary assessment response rates: effect of microinteractions or the smartwatch? Proc ACM Interact Mob Wearable Ubiquitous Technol 2017;1(3):92 [FREE Full text] [doi: 10.1145/3130957] [Medline: 30198012]
40. Ziesemer K, König LM, Boushey CJ, Villinger K, Wahl DR, Butscher S, et al. Occurrence of and reasons for "Missing Events" in Mobile dietary assessments: results from three event-based ecological momentary assessment studies. JMIR Mhealth Uhealth 2020;8(10):e15430 [FREE Full text] [doi: 10.2196/15430] [Medline: 33052123]
41. Ashurst J, van Woerden I, Dunton G, Todd M, Ohri-Vachaspati P, Swan P, et al. The association among emotions and food choices in first-year college students using mobile-ecological momentary assessments. BMC Public Health 2018;18(1):573 [FREE Full text] [doi: 10.1186/s12889-018-5447-0] [Medline: 29716572]
42. Ramadas A, Rizal H, Rajakumar S, Mariapun J, Yasin MS, Armstrong MEG, et al. Dietary intake, obesity, and metabolic risk factors among children and adolescents in the SEACO-CH20 cross-sectional study. Sci Rep 2024;14(1):11265 [FREE Full text] [doi: 10.1038/s41598-024-61090-7] [Medline: 38760446]
43. van Berkel N, Goncalves J, Hosio S, Kostakos V. Gamification of mobile experience sampling improves data quality and quantity. Proc ACM Interact Mob Wearable Ubiquitous Technol 2017;1(3):1-21. [doi: 10.1145/3130972]
44. School sessions to cease rotation system starting today: Radzi. The Vibes.com. 2022. URL: https://www.thevibes.com/articles/education/58838/school-sessions-to-cease-rotation-system-starting-today-radzi [accessed 2023-02-24]
45. Global Smartwatch Shipments in 2024: Market Declines for First Time, China Leads for First Time. Counterpoint Technology Market Research. 2025. URL: https://www.counterpointresearch.com/insight/global-smartwatch-market-in-2024 [accessed 2025-04-01]
46. South East Asia Community Observatory. Monash University Malaysia. URL: https://www.monash.edu.my/seaco [accessed 2026-01-18]

Abbreviations

CREMAS: Checklist for Reporting Ecological Momentary Assessment Studies
EMA: ecological momentary assessment
mEMA: mobile ecological momentary assessment
NCD: noncommunicable disease
OR: odds ratio
SEACO: South-East Asian Community Observatory
SEACO-CH20: South-East Asian Community Observatory Child Health 2020
µEMA: micro-interaction ecological momentary assessment

Edited by A Mavragani; submitted 01.May.2025; peer-reviewed by C Wang, P Delir Haghighi; comments to author 17.Aug.2025; revised version received 18.Nov.2025; accepted 30.Dec.2025; published 06.Feb.2026.

Please cite as:
Lane R, Millard LAC, Salway R, Stone CJ, Skinner AL, Brady SM, Mariapun J, Rajakumar S, Ramadas A, Rizal H, Johnson L, Su TT, Armstrong MEG
The Feasibility of Smartwatch Micro–Ecological Momentary Assessment for Tracking Eating Patterns of Malaysian Children and Adolescents in the South-East Asian Community Observatory Child Health Update 2020: Cross-Sectional Study
J Med Internet Res 2026;28:e73435
URL: https://www.jmir.org/2026/1/e73435
doi:10.2196/73435
PMID:

©Richard Lane, Louise A C Millard, Ruth Salway, Chris J Stone, Andy L Skinner, Sophia M Brady, Jeevitha Mariapun, Sutha Rajakumar, Amutha Ramadas, Hussein Rizal, Laura Johnson, Tin Tin Su, Miranda Elaine Glynis Armstrong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Integrating a Large Language Model to Streamline Nursing Handover Documentation Across Multiple Hospitals in Taiwan: Development and Implementation Study

Ray-Jade Chen[1,2,3*], MD; Mai-Szu Wu[4,5,6*], MD; Lung-Wen Tsai[7,8,9], PhD; Shy-Shin Chang[10,11], MD, PhD; Shu-Tai Shen Hsiao[12,13], PhD; Yu-Sheng Lo[14,15], PhD

[1]Department of Surgery, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
[2]College of Medicine, Professional Master Program in Artificial Intelligence in Medicine, Taipei Medical University, Taipei, Taiwan
[3]Department of Surgery, Division of Pediatric Surgery, Taipei Medical University Hospital, Taipei, Taiwan
[4]Department of Internal Medicine, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
[5]Division of Nephrology, Department of Internal Medicine, Shuang Ho Hospital, New Taipei City, Taiwan
[6]Taipei Medical University Research Center of Urology and Kidney (TMU-RCUK), Taipei Medical University, Taipei, Taiwan
[7]Graduate Institute of Data Science, College of Management, Taipei Medical University, New Taipei City, Taiwan
[8]Department of Information Technology Office, Taipei Medical University Hospital, Taipei, Taiwan
[9]Department of Medicine Research, Taipei Medical University Hospital, Taipei, Taiwan
[10]Department of Family Medicine, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
[11]Department of Family Medicine, Taipei Medical University Hospital, Taipei, Taiwan
[12]Office of Superintendent, Taipei Medical University Hospital, Taipei, Taiwan
[13]School of Nursing, Taipei Medical University, Taipei, Taiwan
[14]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, New Taipei City, Taiwan
[15]Center of Digital Innovation and Development, Taipei Medical University, New Taipei City, Taiwan
[*]these authors contributed equally

Corresponding Author:
Yu-Sheng Lo, PhD
Graduate Institute of Biomedical Informatics
College of Medical Science and Technology
Taipei Medical University
No. 301, Yuantong Road, Zhonghe District
New Taipei City, 235
Taiwan
Phone: 886 6620 2589
Email: Loyusen@tmu.edu.tw

Abstract

Background: The global nursing shortage, exacerbated by heavy workloads and high turnover rates associated with the COVID-19 pandemic, continues to undermine care quality and nurse well-being. Although digital health technologies have enhanced coordination, improved communication, and reduced clinical errors in nursing practice, they have also increased nurses' documentation burden. Advances in large language models (LLMs) and other generative artificial intelligence (GenAI) tools facilitate the generation of accurate reports from electronic medical records (EMRs), thereby streamlining documentation workflows, saving time, and reducing nurses' workloads. Accordingly, integrating LLMs into electronic nursing documentation systems warrants further exploration.

Objective: This study examines the integration of an LLM into an in-house nursing information system (NIS) implemented across 3 hospitals in Taiwan to reduce the time and effort required for nursing handover documentation and to preliminarily assess the operational and economic implications of GenAI-assisted workflows.

Methods: A multidisciplinary team of nursing specialists and information technology experts at Taipei Medical University (TMU) restructured the organization's existing nursing handover documentation process to facilitate interaction with the LLM. The team also developed prompt-based interfaces to automatically generate section-specific content for the nursing handover

XSL•FO
RenderX

document. The LLM-integrated NIS was subsequently deployed across 3 hospitals in Taiwan: Taipei Medical University Hospital (TMUH), Wan Fang Hospital (WFH), and Shuang Ho Hospital (SHH). We then extracted and analyzed NIS log data to compare documentation times before and after LLM implementation, thereby quantifying time savings.

Results: Integration of the LLM into nursing handover documentation was associated with shorter per-patient documentation time in routine clinical use across TMUH, WFH, and SHH. Based on preintegration NIS logs (September 2024), the average handover document completion time per patient ranged from 3.45 (SD 3.82) to 4.32 (SD 4.48) minutes across hospitals and shifts, providing a preliminary baseline for subsequent comparisons. In postintegration NIS logs (October-December 2024), the overall handover document completion time per patient (mean) was substantially lower, ranging from 1.17 (SD 1.86) to 2.54 (SD 2.82) minutes across hospitals and shifts. Using monthly patient volume to estimate time savings, 113-273, 160-314, and 198-391 hours were saved per month at TMUH, WFH, and SHH, respectively, corresponding to aggregate savings of 474-981 hours per month across hospitals during the study period.

Conclusions: We integrated an LLM into an NIS to generate nursing handover documents without altering existing workflows. Across 3 hospitals within TMU's health system, GenAI assistance was associated with shorter documentation time and a positive net labor value from October to December 2024. Prompts were constrained, and nurse verification was required to mitigate hallucinations. Future work will enhance logging to capture reliability and editing metrics, compare LLM-generated drafts with nurse-finalized notes to inform prompt refinement, and assess generalizability to other documentation workflows.

(J Med Internet Res 2026;28:e81604) doi:10.2196/81604

KEYWORDS

nursing shortage; large language model; generative artificial intelligence; artificial intelligence; nursing information system; electronic medical records

Introduction

The global nursing shortage has been exacerbated by several factors and remains a critical challenge for health care systems, threatening health care quality and access [1,2]. The COVID-19 pandemic has increased nursing workloads, resulting in increased patient acuity, greater infection control demands, and longer working hours—all of which have contributed to nurses experiencing physical exhaustion, moral distress, and mental health challenges [3,4]. Moreover, the pandemic worsened the nursing shortage, with 100,000 nurses leaving the US health care workforce during the pandemic and an additional 900,000 planning to exit by 2027 [5]. This shortage has compromised care quality and increased burnout and turnover risks. According to the World Health Organization (WHO), although the global nursing workforce grew from 27.9 million in 2018 to 29.8 million in 2023, a shortage of 5.8 million nurses persists, which is projected to decrease to 4.1 million by 2030 [6].

Although Taiwan experienced early success in controlling the COVID-19 pandemic, the pandemic exposed structural problems within the country's nursing sector. Nursing shortages, particularly in intensive care and emergency departments, persisted during the late-pandemic and postpandemic periods. High turnover rates and uneven workloads contributed to increased work-related stress [7]. Furthermore, interviews with public health nurses revealed that frequent changes in temporary assignments and information systems during the pandemic exacerbated fatigue and created additional challenges for frontline nurses [8].

The ongoing digital transformation in health care has rendered digital technologies essential to nursing practice. Studies have indicated that the adoption of digital tools by health care professionals improves care quality, enhances patient participation in their own care [9,10], fosters better coordination among health care professionals [11], reduces medication errors and adverse events [12,13], and improves documentation accuracy [14,15]. However, this shift has led to more time being spent on indirect tasks, particularly activities related to maintaining electronic medical records (EMRs), including data collection, entry, and confirmation [16]. Documentation accounts for a considerable portion of nurses' work time, and this burden is exacerbated by high patient turnover [17] and the use of multiple recording systems [18,19]. A systematic review revealed that nurses spend between 13% and 31% of their work time on documentation [20]; thus, considering nurses' heavy workloads, reducing documentation time has become a critical issue in nursing.

Large language models (LLMs), a category of artificial intelligence (AI) tools, perform deep learning with millions or billions of parameters on extensive data to generate human-like text [21]. LLM platforms, such as ChatGPT (OpenAI), Gemini (Google LLC), and Llama (Meta Platforms, Inc.) [22,23], are applied in various domains, including text generation, question answering, translation, and creative content creation. Moreover, large multimodal models have been developed to process several types of data, such as text, images, video, and audio, thereby broadening the application potential of generative AI (GenAI) tools [24,25]. In health care, LLMs and other GenAI tools enable coherent text generation for tasks such as drug discovery, personalized medicine, radiology diagnosis, automated clinical documentation, medical research, and health information consultation [26-34]. These tools are driving advancements in diverse health care applications; however, ensuring the accuracy, reliability, and fairness of the generated results remains critical [35].

AI applications in nursing are rapidly expanding, and AI tools have the potential to become essential parts of nursing practice [36,37]. By integrating clinical data with language generation tools, GenAI systems can improve documentation quality,

XSL•FO
RenderX

support clinical decision-making, and reduce administrative workload. Additionally, these systems can considerably shorten documentation time and enhance handover efficiency by rapidly producing consistent records, thereby allowing nurses to devote more time to patient care and ultimately improving the quality and safety of health care delivery [38,39].

To explore the potential of incorporating LLMs into electronic nursing documentation systems, this study integrated an LLM into an in-house nursing information system (NIS) to reduce the time and effort required for nursing handover documentation. We evaluated the LLM-integrated NIS by comparing the time spent by nurses on nursing handover documentation before and after integration. We additionally conducted a preliminary cost-effectiveness assessment and quantified the revision rate of GenAI-generated drafts before final submission.

Methods

Settings

This study was conducted at Taipei Medical University (TMU), which operates 3 hospitals: Taipei Medical University Hospital (TMUH), Wan Fang Hospital (WFH), and Shuang Ho Hospital (SHH). The combined capacity of these hospitals is 2675 acute beds (TMUH: 735; WFH: 761; SHH: 1179), and they handle a total of over 400,000 outpatient visits per month. Owing to nursing shortages, operational capacity decreased to 2195 acute beds as of May 2025 (TMUH: 608; WFH: 576; SHH: 1011). Nursing workflows are supported by an in-house NIS managed by the TMU Center of Digital Innovation and Development.

Google Gemini

We selected a HIPAA (Health Insurance Portability and Accountability Act)-compliant enterprise instance of Gemini 1.5 Pro (Google LLC) for integration into the in-house NIS. Gemini 1.5 Pro is a mid-sized multimodal LLM optimized for a broad range of reasoning tasks and can generate text output in response to diverse inputs, including text, image, video, and audio. Moreover, it offers capabilities for output streaming, conversational interactions, and system-level instruction handling. The model can be used to conduct various language-related tasks, such as text generation, text editing, code generation, problem-solving, recommendation generation, information extraction, and data generation [40].

Conventional Workflow in Nursing Handover Documentation

In hospitals across Taiwan, nursing schedules are typically structured around 3 daily shifts, and nurses are required to complete handover documentation twice per day. The nursing handover document follows the WHO-endorsed ISBAR (Identify, Situation, Background, Assessment, Recommendation) framework structured into 5 key components [41].

Figure 1 illustrates the conventional workflow for nursing handover documentation with the in-house NIS. Initially, the NIS displays a list of patients requiring handover documentation. Subsequently, nurses collect and analyze individualized patient information using various EMR systems. They then sequentially write a nursing handover document for each patient. Finally, the documented handover information is manually summarized to formulate the corresponding nursing recommendations for each patient.

In general, nurses prepare handover documents for 5-7 patients per shift, and senior nurses estimated an average completion time of approximately 5 minutes per patient. Accordingly, each nurse dedicates approximately 1 hour to complete patient handover tasks; this time frame includes both the documentation process and the subsequent verification and explanation of patient details to the receiving nurse.

XSL•FO
RenderX

Figure 1. Conventional workflow for nursing handover documentation using the in-house nursing information system (NIS) at Taipei Medical University. EMR: electronic medical record.

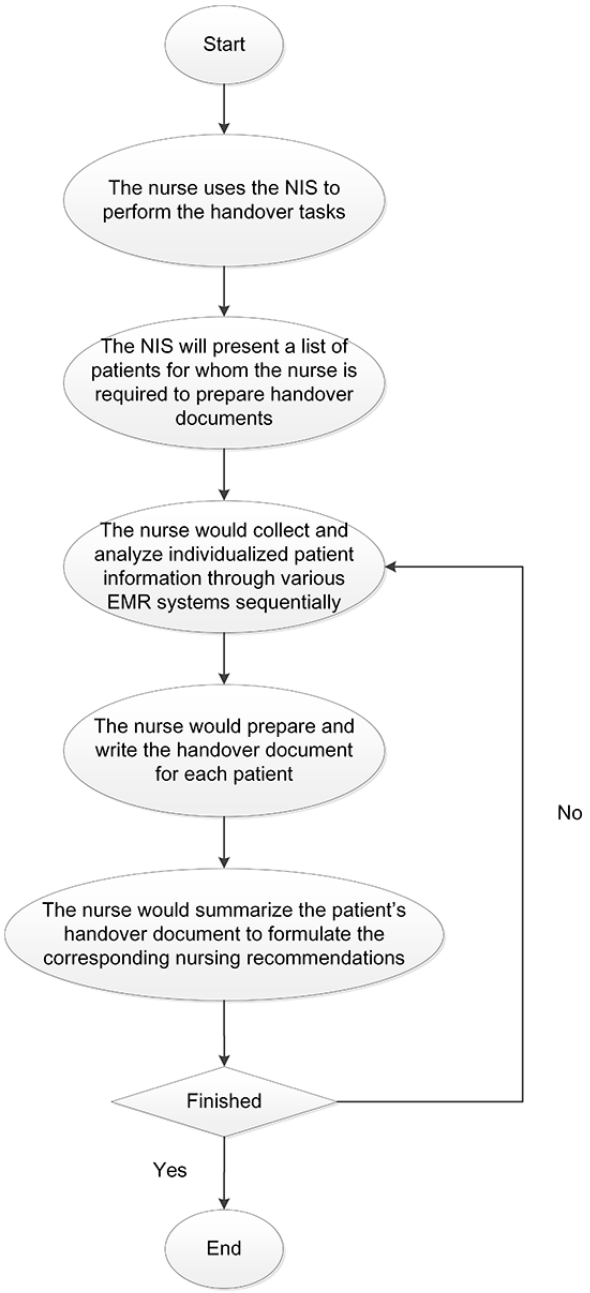

Prompt Design for Nursing Handover Documentation

A multidisciplinary team of nursing specialists and information technology professionals collaboratively outlined the key components of the nursing handover document, as defined in the existing EMR systems. Each component was divided into several sections, which were systematically evaluated to assess the extracted content and identify potential problems. The most critical problem identified was related to the manual collection, verification, transcription, and integration of patient data from various EMR systems. Thus, we reorganized the structure of the nursing handover document. Tables 1 and 2, respectively, present the sections of the original and revised (for the LLM-integrated NIS) nursing handover documents across 5 key components according to the ISBAR framework.

We developed section-specific prompts to support nurses' use of the LLM, targeting 4 tasks: (1) extracting information from the provided input, (2) summarizing input text, (3) interpreting or translating text (eg, English to Mandarin), and (4) formatting content based on the input. When executed, these prompts generate draft content for the corresponding handover sections. Prompt development followed an iterative, human-in-the-loop process: draft prompts were repeatedly tested using deidentified EMRs, and outputs were reviewed by nursing professionals for clinical relevance, completeness, and adherence to the intended structure. Feedback-informed refinements were made to prompt wording, section constraints, and output formatting until the drafts were suitable for routine handover use. All prompts instructed the LLM to use only the provided input and to avoid introducing new values, assumptions, or inferences. No patient

XSL•FO
RenderX

identifiers or personally identifiable information was transmitted to the LLM.

For example, in the laboratory results section, the NIS queried the laboratory information system to retrieve each patient's test results and corresponding reference ranges from the preceding 48 hours. A tailored prompt was then sent via the LLM application programming interface (API) to identify abnormal results by comparing values with reference ranges and to present the extracted abnormalities in a structured table. Table 3 summarizes the tailored prompts by section and the corresponding EMR source systems, including the inpatient system, medical documentation system, NIS, laboratory information system, examination information system, radiology information system, patient schedule system, operating room system, and blood bank system.

Table 1. The original nursing handover document.

Components	Sections
Identity	Introduction
Situation	Vital signs, blood glucose, laboratory test results, culture results, diagnostic examination reports, radiology reports, pending diagnostic examination tests or radiology tests, surgical records, and blood products
Background	Admission reason, medical history
Assessment	Wound records, consultation records
Recommendations	A blank field for manual data entry
Others	A blank field for manual data entry

Table 2. The revised (for the large language model–integrated nursing information system) nursing handover document.

Components	Sections
Identity	Introduction
Handover records[a]	Vital signs, blood glucose, laboratory test results, culture results, diagnostic examination reports, radiology reports, pending diagnostic examination tests or radiology tests, surgical records, blood products, wound records, consultation records, and recommendations
Background	Admission reason, medical history
Others	A blank field for manual data entry

[a]Handover records include records related to the situation (vital signs, blood glucose, laboratory test results, culture results, diagnostic examination reports, radiology reports, pending diagnostic examination tests or radiology tests, surgical records, and blood products), assessment (wound records and consultation records), and recommendations.

XSL•FO
RenderX

Table 3. Tailored prompts by handover section and corresponding electronic medical record source systems.

Components and sections	Electronic medical record source systems	Tailored prompts
Identity		
Introduction	Inpatient system	The nursing information system already provides information (manual data entry is not necessary).
Background		
Admission reason and medical history	Inpatient system and medical documentation system	Summarize admission details into a concise paragraph. Highlight the primary reason for admission and include key clinical details.
Handover records		
Vital signs	Nursing information system	Extract and structure the following vital signs data from the past 8 hours from the provided data: date, time, temperature, pulse, respiration, blood pressure, oxygen level, and Glasgow Coma Scale values. Maintain chronological order.
Blood glucose	Laboratory information system	Extract and structure blood glucose readings from the past 8 hours, including date, time, and measurement value. Organize the data in reverse chronological order.
Laboratory test results	Laboratory information system	Extract the following laboratory results from the past 48 hours: test item, value, and reference range. Mark abnormal values (H for high and L for low), and leave normal values unmarked. Organize the results chronologically and return them in a structured JSON format.
Culture results	Laboratory information system	Extract and structure key information from culture data within 6 days. For each data sample, identify and output the following fields in a structured format: collection date, specimen type, culture results, and status (completed or pending).
Diagnostic examination reports	Examination information system	Summarize diagnostic findings from the past 24 hours, emphasizing significant abnormal results in Standard Mandarin. If such findings are unavailable, state “No examination reports available.”
Radiology reports	Radiology information system	Summarize key radiological findings from the past 24 hours, highlighting clinically relevant abnormalities in Standard Mandarin. If such findings are unavailable, state “No radiology reports available.”
Pending diagnostic examination tests or radiology tests	Patient schedule system	List upcoming diagnostic and radiology tests in the following format: procedure name and scheduled date.
Surgical records	Operating room system	Extract the following information from surgical records from the past 48 hours: procedure name, start time, end time, blood loss, and anesthesia method. If no records are available, state “No surgical records available.”
Blood products	Blood bank system	Extract the following information on blood product orders within 48 hours: order number, blood type and Rh[a] factor, order date, product name, quantity ordered/used, and transfusion date. If such information is unavailable, state “No blood product records available.”
Wound records	Nursing information system	Extract the following wound assessment data: type, location, assessment date/time, and dimensions. If these data are unavailable, state “No wound assessment records available.”
Consultation records	Medical documentation system	Summarize consultation findings and recommendations in Standard Mandarin. If such information is unavailable, state “No consultation records available.”
Recommendations	Blank field	Generate comprehensive nursing care recommendations based on the provided data. Present recommendations as bullet points in only Standard Mandarin. Include information on priority interventions, monitoring requirements, preventive measures, and areas for special attention.
Others		
Others	Blank field for manual data entry	N/A[b]

[a]Rh factor: Rhesus factor.

[b]N/A: not applicable.

XSL•FO
RenderX

Interactions Between the NIS and the LLM in the Designed Workflow for Nursing Handover Documentation

Google Gemini 1.5 Pro offers an API that provides access to its LLM, which was integrated into the in-house NIS to automatically generate nursing handover documents.

Figure 2 illustrates the integrated workflow of the LLM-integrated NIS. For each section of the nursing handover document, the NIS sends a request to the relevant EMR system to retrieve patient data through asynchronous calls. The retrieved data and the corresponding tailored prompt are then used as input to call the LLM API. Upon receiving the LLM response, the NIS organizes the content and formats it according to the appropriate style for each section. The final output generated by the NIS is a structured and comprehensive handover document that can be manually reviewed and validated by nursing staff.

Figure 2. Integrated workflow for the large language model (LLM)–integrated nursing information system (NIS). API: application programming interface; EMR: electronic medical record.

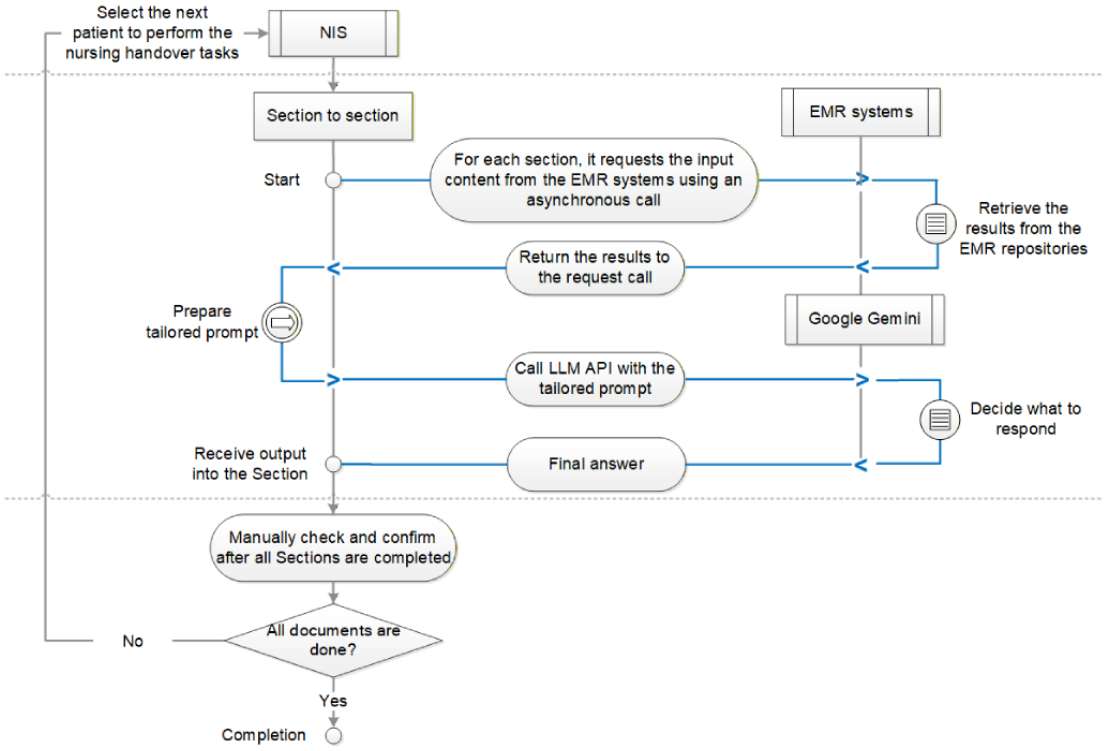

Log Data Collection and Analysis of the Nursing Handover Document in the NIS

In this study, log data from the LLM-integrated NIS, implemented in October 2024, were used to conduct a preliminary pre-post comparison of nursing handover document completion times. Handover document completion time was defined as the interval from when a nurse selected a patient by clicking the handover button, through manual entry before GenAI integration or automatic draft generation after integration, to final confirmation, when the nurse completed review and revision and clicked the “Confirm” button for that patient.

To support subsequent analyses and system optimization, step-level log data were collected across the full handover documentation workflow. We retrospectively extracted NIS logs from September 2024 (preintegration) and from October-December 2024 (postintegration) for the 3 TMU-managed hospitals. To reduce the influence of extreme values, log data with handover document completion times exceeding 20 minutes were excluded from the analyses. In addition, we logged both the automatically GenAI-generated handover draft and the final nurse-confirmed version to preliminarily quantify the document revision rate. As nurse identifiers were unavailable, we used hospital × shift as the clustering unit and applied robust (sandwich) SEs to account for within-cluster correlation.

Ethical Considerations

This retrospective study of deidentified NIS logs was approved by the TMU-Joint Institutional Review Board (TMU-JIRB; approval number N202506094). All data were deidentified to protect participant privacy and confidentiality, and no personally identifiable information was collected or disclosed. The TMU-JIRB waived the requirement for informed consent. All images included in the manuscript were reviewed to ensure that no individual participants or users could be identified. The hospitals also conduct routine qualitative and quantitative medical record audits, providing organizational accountability and quality assurance for AI-assisted documentation.

Results

Descriptive Statistics of Nursing-Related Factors for the Examined Hospitals

From September 1 to December 31, 2024, we collected NIS log data from the 3 hospitals during daily nursing handovers across

XSL•FO
RenderX

3 shifts (day, night, and overnight). At TMU, nursing handover documents must be prepared for most inpatients; however, certain patient categories are exempt from this requirement. These categories include patients undergoing postoperative observation, day hospitalization, and 1-day chemotherapy at TMUH; nursing home residents and day-care patients at WFH; and day-care, psychiatric ward, nursing home, and postpartum home care patients at SHH.

During the 4-month study period, the number of nurses working across the 3 daily shifts ranged from 501 to 524 at TMUH, 468 to 489 at WFH, and 862 to 892 at SHH. Monthly inpatient case volumes ranged from 7477 to 7733 at TMUH, 5406 to 5977 at WFH, and 8872 to 9814 at SHH. The mean monthly length of stay was 4.7-4.9 days at TMUH, 4.7-5.1 days at WFH, and 4.9-5.0 days at SHH. Across all shifts, total monthly shift counts ranged from 35,527 to 37,892 at TMUH, 27,500 to 30,835 at WFH, and 42,535 to 44,945 at SHH. Multimedia Appendix 1 summarizes descriptive statistics for the 3 hospitals included in this study.

Comparison of the Original and GenAI-Assisted Nursing Handover Documentation Functions in the NIS

In accordance with the workflow shown in Figure 1, the LLM-integrated NIS initially displays a table listing patients, enabling nurses to generate handover documents sequentially by clicking a selection checkbox (Figure 3). Each row presents key patient attributes, including the patient's name, medical record number, bed number, birth date and age, admission date and length of stay, department, and attending physician. The table also includes a handover button to initiate the nursing handover process.

Figure 4 depicts the nursing handover documentation interface in the original NIS, displayed when a nurse selects a patient and initiates the handover process. Nurses complete 5 handover sections: Introduction (auto-populated; no manual entry required), Situation, Background, Assessment, and Recommendation and Other information.

Figure 5 depicts the revised nursing handover document in the LLM-integrated NIS. In contrast to the original NIS, selecting a patient for handover triggers an API call to the LLM. After the response is returned, the LLM-integrated NIS automatically populates draft content for each handover section. Nurses then review, revise, and confirm the draft. Once finalized, the nurse clicks "Confirm" to submit the handover document for that patient (Figure 6).

Figure 3. Screenshot of the patient list interface used by nurses to perform handover tasks in the nursing information system.



XSL•FO
RenderX

Figure 4. Screenshot of the Identify, Situation, Background, Assessment, Recommendation interface used for nursing handover tasks in the original nursing information system.

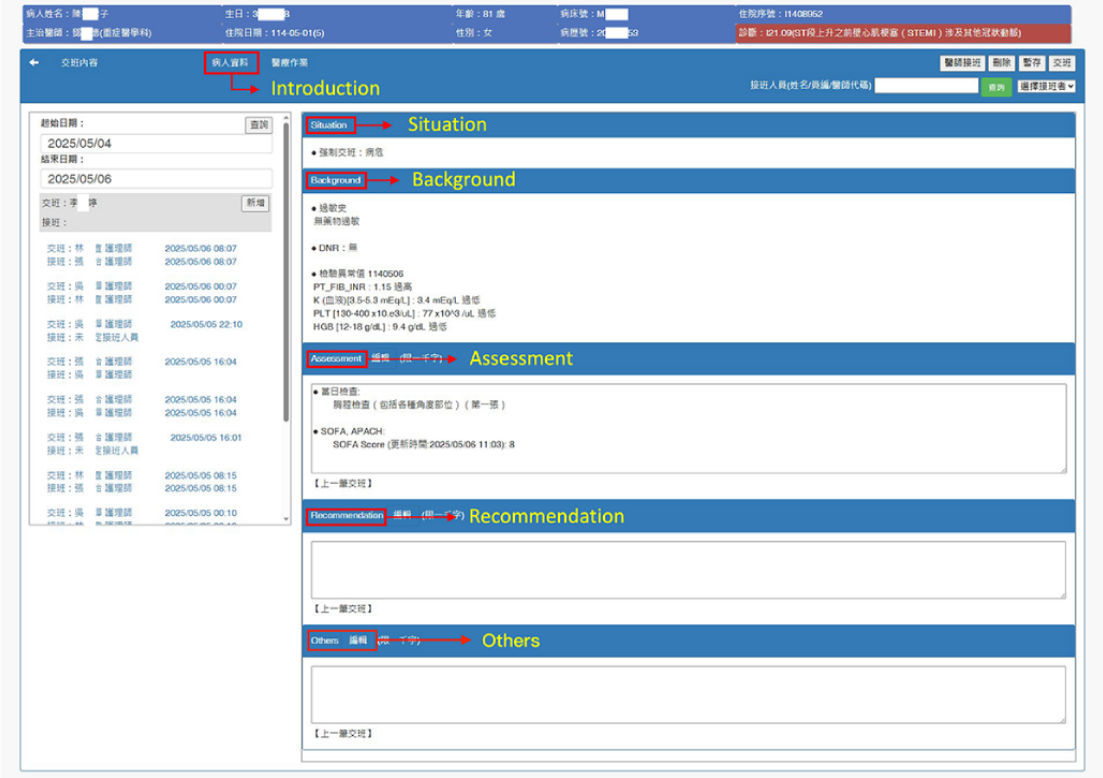

Figure 5. Screenshot of the revised nursing handover document in the large language model–integrated nursing information system.

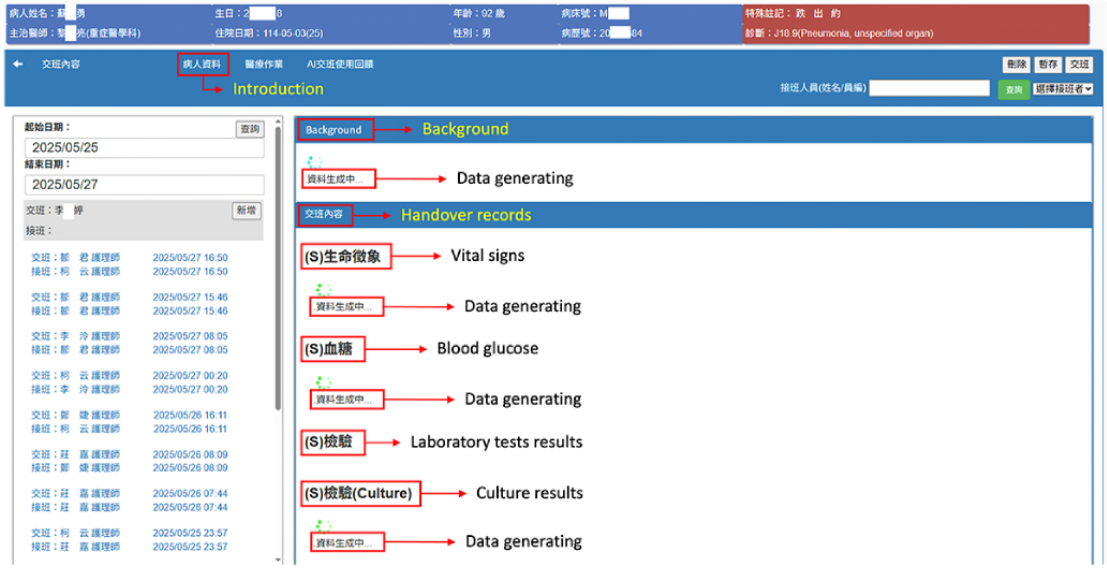

XSL•FO
RenderX

Figure 6. Screenshot of a complete nursing handover document automatically generated by the large language model–integrated nursing information system.

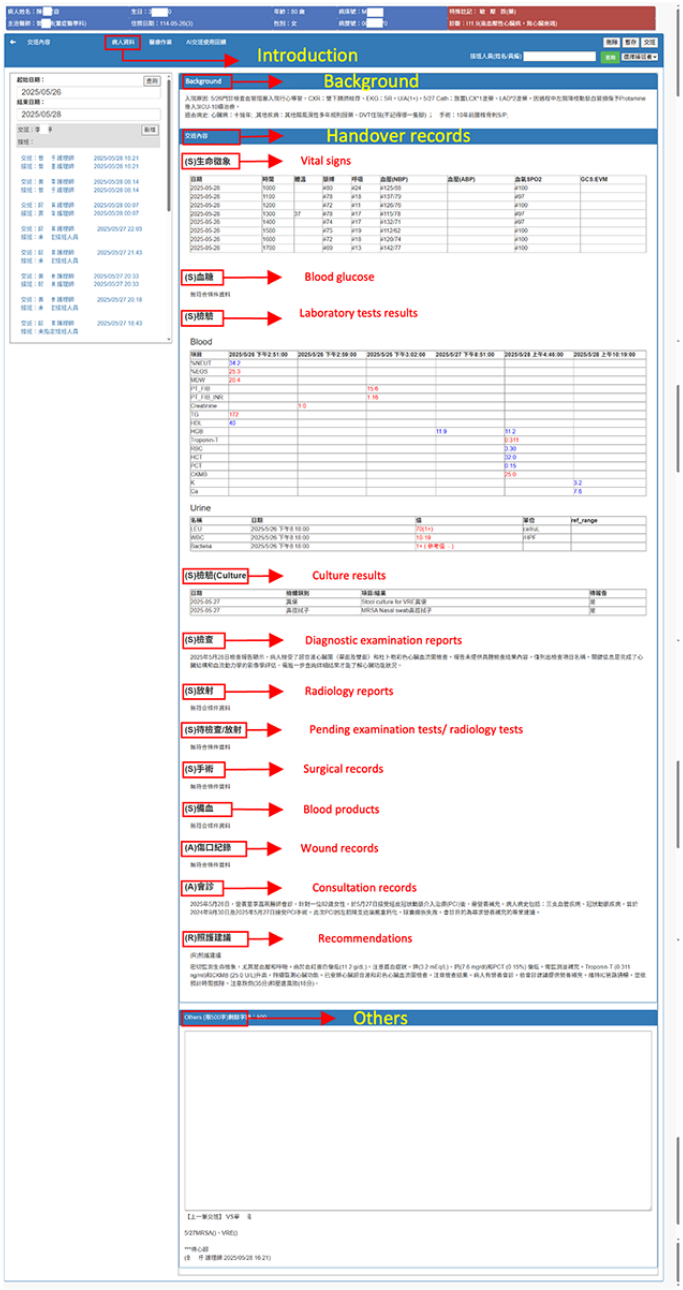

Analysis of NIS Log Data for Pre-LLM Integration

Table 4 presents detailed NIS log data from September 2024 for TMUH, WFH, and SHH, including the number of shifts and the mean (SD) of handover document completion time per patient for each shift. Before LLM integration, nurses often had to pause documentation to gather additional patient information or respond to unplanned clinical events (eg, bedside care, family inquiries, or urgent clinical duties). Accordingly, the NIS included a draft-save function that allowed handover content to be temporarily stored and completed later.

A total of 110,457 shift-level handover logs were analyzed (TMUH: 37,087; WFH: 30,835; and SHH: 42,535), representing large-scale real-world clinical operations. Across hospitals and shifts, the mean handover document completion time per patient ranged from 3.45 (SD 3.82) to 4.32 (SD 4.48) minutes. At TMUH, mean completion times were 4.10 (SD 4.55) minutes (day shift), 4.32 (SD 4.48) minutes (night shift), and 3.45 (SD 3.82) minutes (overnight shift). At WFH, the corresponding values were 4.26 (SD 5.08), 3.70 (SD 4.47), and 4.10 (SD 5.29) minutes, respectively, whereas at SHH they were 3.67 (SD 4.31), 4.09 (SD 4.25), and 3.81 (SD 4.23) minutes, respectively.

Although only 1 month of preintegration log data were available, these results provide a preliminary baseline of nursing handover documentation performance against which the impact of subsequent GenAI-assisted workflows can be quantitatively evaluated.

XSL•FO
RenderX

Table 4. Shift counts and mean document completion time, stratified by hospital and shift type before large language model integration (September 2024).

Hospital and shift	September 24	
	Number of shifts, n	Document completion time (minutes), mean (SD)
Taipei Medical University Hospital (n=37,087)		
Day	12,418	4.10 (4.55)
Night	12,203	4.32 (4.48)
Overnight	12,466	3.45 (3.82)
Wan Fang Hospital (n=30,835)		
Day	9379	4.26 (5.08)
Night	10,547	3.70 (4.47)
Overnight	10,909	4.10 (5.29)
Shuang Ho Hospital (n=42,535)		
Day	13,868	3.67 (4.31)
Night	14,466	4.09 (4.25)
Overnight	14,201	3.81 (4.23)

Analysis of NIS Log Data for LLM Responses

Table 5 summarizes shift-level log data from October to December 2024 for TMUH, WFH, and SHH, including the number of shifts, mean GenAI document generation time per patient, mean nurse manual (review-revision-confirmation) time per patient, and mean overall document completion time per patient. SHH had the highest shift volume (43,473 in October to 44,945 in December) and the largest monthly patient volume (8872-8989; Multimedia Appendix 1), with mean overall completion times ranging from 1.74 (SD 2.5) to 2.11 (SD 2.82) minutes. WFH recorded the lowest shift volume (27,500-28,730) and the shortest mean overall completion times (mean 1.17, SD 1.86 minutes to mean 1.78, SD 2.53 minutes) while managing 5746-5977 patients per month (Multimedia Appendix 1). TMUH showed intermediate shift volume (35,890-37,892) and patient volume (7477-7733; Multimedia Appendix 1), with mean overall completion times of 2.13 (SD 2.5) to 2.54 (SD 2.82) minutes.

We further conducted analyses using joint Wald tests to assess whether hospital, shift, and month were associated with statistically significant differences. As individual nurse identifiers were not available in the log data, clustering at the nurse level was not feasible. Therefore, we defined the clustering unit as hospital × shift, and the corresponding results are presented in Table 6.

In Table 6, joint Wald tests indicated that all 3 outcomes (GenAI time, manual time, and document completion time) differed significantly across hospitals and shifts (hospital: Wald χ^2_2=52.85, P<.001; shift: Wald χ^2_2=10.39, P=.006). By contrast, the effect of month was not statistically significant (Wald χ^2_2=5.74, P=.057).

Moreover, we used generalized estimating equations (GEEs) to account for within-cluster correlation arising from the clustered and longitudinal structure of the data. Separate GEE models were fitted for GenAI time, manual time, and document completion time. Each model included month, shift, and hospital as categorical covariates and used an exchangeable working correlation structure with robust SEs; thus, estimates represent population-averaged (system-level) effects rather than individual-level effects. Table 7 reports the population-averaged GEE results for GenAI time, manual time, and document completion time (β and SE). Relative to SHH, TMUH had longer times across all outcomes (GenAI time: β=3.05; manual time: β=20.35; document completion time: β=23.40; all P<.001), whereas WFH had shorter times (GenAI time: β=−3.47; manual time: β=−23.16; document completion time: β=−26.64; all P<.001). December (vs October) was associated with small increases across outcomes (P=.02), whereas November (vs October; P=.41) and shift (P=.44) were not statistically significant.

XSL•FO
RenderX

Table 5. Monthly shift-level patient volume and mean per-patient generative artificial intelligence, manual, and overall completion times across 3 Taipei Medical University hospitals (October-December 2024).

Hospital and shift	October 24				November 24				December 24			
	Number of shifts, n	Generative artificial intelligence time (seconds)[a], mean (SD)	Manual time (minutes)[b], mean (SD)	Document completion time (minutes)[c], mean (SD)	Number of shifts, n	Generative artificial intelligence time (seconds), mean (SD)	Manual time (minutes), mean (SD)	Document completion time (minutes), mean (SD)	Number of shifts, n	Generative artificial intelligence time (seconds), mean (SD)	Manual time (minutes), mean (SD)	Document completion time (minutes), mean (SD)
Taipei Medical University Hospital[d]												
Day	12,043	17.0 (20.0)	1.89 (2.23)	2.17 (2.56)	11,882	16.70 (19.60)	1.85 (2.18)	2.13 (2.50)	12,510	17.00 (19.60)	1.88 (2.17)	2.17 (2.50)
Night	11,928	19.6 (22.1)	2.18 (2.45)	2.50 (2.82)	11,900	18.70 (22.10)	2.08 (2.45)	2.40 (2.82)	12,740	19.80 (22.10)	2.20 (2.45)	2.54 (2.82)
Overnight	11,918	19.0 (21.6)	2.11 (2.40)	2.43 (2.76)	11,745	19.00 (21.60)	2.11 (2.39)	2.42 (2.75)	12,642	18.60 (20.80)	2.07 (2.31)	2.38 (2.66)
Wan Fang Hospital[e]												
Day	8291	12.1 (17.9)	1.34 (1.99)	1.54 (2.29)	8620	13.20 (19.50)	1.47 (2.17)	1.68 (2.50)	8905	13.90 (19.80)	1.55 (2.20)	1.78 (2.53)
Night	9612	12.5 (18.7)	1.39 (2.08)	1.60 (2.39)	10,035	12.50 (18.70)	1.38 (2.07)	1.59 (2.39)	10,190	12.50 (18.60)	1.39 (2.06)	1.60 (2.37)
Overnight	9597	9.2 (14.6)	1.02 (1.62)	1.17 (1.86)	9438	10.10 (15.40)	1.12 (1.72)	1.28 (1.97)	9635	9.60 (15.10)	1.07 (1.68)	1.23 (1.93)
Shuang Ho Hospital[f]												
Day	13,103	14.8 (20.2)	1.65 (2.24)	1.89 (2.58)	13,931	15.40 (20.70)	1.71 (2.30)	1.97 (2.65)	13,875	15.30 (20.90)	1.70 (2.33)	1.96 (2.67)
Night	15,317	15.7 (21.3)	1.74 (2.36)	2.01 (2.72)	15,268	16.10 (21.50)	1.79 (2.39)	2.06 (2.75)	16,040	16.50 (22.00)	1.84 (2.45)	2.11 (2.82)
Overnight	15,053	13.7 (19.6)	1.52 (2.18)	1.74 (2.50)	15,337	14.10 (20.60)	1.57 (2.29)	1.80 (2.64)	15,030	15.60 (22.40)	1.74 (2.49)	2.00 (2.86)

[a]Mean generative artificial intelligence document generation time per patient.

[b]Mean nurse manual (review-revision-confirmation) time per patient.

[c]Mean overall document completion time per patient.

[d]The total number of shifts for October, November, and December was 35,890, 35,527, and 37,892, respectively.

[e]The total number of shifts for October, November, and December was 27,500, 28,092, and 28,730, respectively.

[f]The total number of shifts for October, November, and December was 43,473, 44,536, and 44,945, respectively.

XSL•FO
RenderX

Table 6. Joint Wald tests for month, shift, and hospital effects on generative artificial intelligence time, manual time, and document completion time.

Outcome	Factor	Wald chi-square statistic (*df*)	*P* value
Generative artificial intelligence document generation time	Month	5.74 (2)	.06
Manual time[a]	Month	5.74 (2)	.06
Document completion time[b]	Month	5.74 (2)	.06
Generative artificial intelligence document generation time	Shift	10.39 (2)	.006
Manual time	Shift	10.39 (2)	.006
Document completion time	Shift	10.39 (2)	.006
Generative artificial intelligence document generation time	Hospital	52.85 (2)	<.001
Manual time	Hospital	52.85 (2)	<.001
Document completion time	Hospital	52.85 (2)	<.001

[a]Nurse manual (review-revision-confirmation) time.

[b]Overall document completion time per patient.

Table 7. Generalized estimating equation model coefficients for generative artificial intelligence document generation time, manual time, and document completion time.

Outcome	Factor	β (seconds)	95% CI	*P* value
Generative artificial intelligence document generation time	Intercept	15.09	14.26 to 15.92	<.001
Generative artificial intelligence document generation time	Hospital: TMUH[a] vs SHH[b]	3.05	1.80 to 4.31	<.001
Generative artificial intelligence document generation time	Hospital: WFH[c] vs SHH	–3.47	–4.72 to –2.23	<.001
Manual time[d]	Intercept	100.60	95.05 to 106.15	<.001
Manual time	Hospital: TMUH vs SHH	20.35	11.97 to 28.72	<.001
Manual time	Hospital: WFH vs SHH	–23.16	–31.45 to –14.88	<.001
Document completion time[e]	Intercept	115.69	109.31 to 122.07	<.001
Document completion time	Hospital: TMUH vs SHH	23.40	13.77 to 33.03	<.001
Document completion time	Hospital: WFH vs SHH	–26.64	–36.16 to –17.11	<.001
Generative artificial intelligence document generation time	Month: December vs October	.56	0.09 to 1.04	.02
Manual time	Month: December vs October	3.75	0.58 to 6.92	.02
Document completion time	Month: December vs October	4.31	0.66 to 7.95	.02
Generative artificial intelligence document generation time	Shift: night vs day	1.05	–0.24 to 2.34	.11
Manual time	Shift: night vs day	6.99	–1.61 to 15.60	.11
Document completion time	Shift: night vs day	8.04	–1.85 to 17.93	.11
Generative artificial intelligence document generation time	Month: November vs October	.16	–0.23 to 0.55	.41
Manual time	Month: November vs October	1.09	–1.50 to 3.69	.41
Document completion time	Month: November vs October	1.26	–1.73 to 4.24	.41
Generative artificial intelligence document generation time	Shift: overnight vs day	–.69	–2.45 to 1.07	.44
Manual time	Shift: overnight vs day	–4.59	–16.34 to 7.15	.44
Document completion time	Shift: overnight vs day	–5.28	–18.79 to 8.22	.44

[a]TMUH: Taipei Medical University Hospital.

[b]SHH: Shuang Ho Hospital.

[c]WFH: Wan Fang Hospital.

[d]Nurse manual (review-revision-confirmation) time.

[e]Overall document completion time per patient.

XSL•FO
RenderX

Comparison of NIS Log Data Before and After LLM Integration

Using NIS logs from the preintegration period (September 2024) and the postintegration period (October-December 2024), we compared handover document completion time. Table 8 reports population-averaged GEE estimates for document completion time (seconds). Document completion time was significantly longer preintegration than postintegration (β=63.48 seconds; 95% CI 53.81-73.16; *P*<.001), and the overall period effect was significant (χ^2_1=165.24; *P*<.001). Relative to SHH, completion time was longer at TMUH (β=19.10 seconds; 95% CI 10.34-27.85; *P*<.001) and shorter at WFH (β=−25.67 seconds; 95% CI −34.42 to −16.92; *P*<.001). Shift effects were not statistically significant (night vs day, *P*=.15; overnight vs day, *P*=.33).

Table 8. Generalized estimating equation model coefficients for handover document completion time comparing pre- and postintegration periods.

Outcome	Factor	β (seconds)	95% CI	*P* value
Document completion time[a]	Intercept	118.01	112.21 to 123.81	<.001
Document completion time	Pre[b] vs Post[c]	63.48	53.81 to 73.16	<.001
Document completion time	Taipei Medical University Hospital vs Shuang Ho Hospital	19.10	10.34 to 27.85	<.001
Document completion time	Wan Fang Hospital vs Shuang Ho Hospital	–25.67	–34.42 to –16.92	<.001
Document completion time	Night vs day	6.44	–2.35 to 15.22	.15
Document completion time	Overnight vs day	–5.98	–17.99 to 6.03	.33

[a]Overall document completion time per patient.

[b]Preintegration document completion time.

[c]Postintegration document completion time.

Revision Rate of GenAI-Generated Handover Drafts

Using NIS logs from October to December 2024, we examined revisions of GenAI-generated handover drafts by hospital and shift. Table 9 presents, for each hospital-shift stratum, the number of patient-level handover documents in which nurses revised the GenAI-generated draft (number of documents with revisions), the total number of patient-level handover documents (number of documents), and the corresponding revision rate (number of documents with revisions/number of documents, %).

Revisions were common across hospitals and months. At TMUH, revision rates were higher during night (9267/12,740-9162/11,928, 72.74%-76.81%) and overnight shifts (9551/11,918-10,830/12,642, 80.14%-85.67%) than during day shifts (8343/12,510-8164/12,043, 66.69%-67.79%). At WFH, day-shift revision rates increased from nearly 50% (4215/8291, 50.84%) in October to over 60% (5334/8620-5562/8905, 61.88%-62.46%) in November-December, whereas night-shift revision rates decreased from nearly 79% (7684/9612, 79.94%) in October to nearly 60% (6000/10,035-6178/10,190, 59.79%-60.63%) thereafter; overnight-shift revision rates remained high (6960/9438-7804/9597, 73.74%-81.32%). At SHH, revision rates were relatively stable, with day shifts at 68.86%-74.16% (9554/13,875-9717/13,103) and night and overnight shifts at 73.95%-78.95% (11,861/16,040-11,884/15,053).

XSL•FO
RenderX

Table 9. Monthly patient-level handover document counts and revision rates of generative artificial intelligence–generated drafts by hospital and shift (October-December 2024).[a]

Hospital and shift	October 24			November 24			December 24		
	Number of documents with revisions[b], n	Number of documents[c], n	Revision rate[d], %	Number of documents with revisions, n	Number of documents, n	Revision rate, %	Number of documents with revisions, n	Number of documents, n	Revision rate, %
Taipei Medical University Hospital									
Day	8164	12,043	67.79	8001	11,882	67.34	8343	12,510	66.69
Night	9162	11,928	76.81	9000	11,900	75.63	9267	12,740	72.74
Overnight	9551	11,918	80.14	9744	11,745	82.96	10,830	12,642	85.67
Wan Fang Hospital									
Day	4215	8291	50.84	5334	8620	61.88	5562	8905	62.46
Night	7684	9612	79.94	6000	10,035	59.79	6178	10,190	60.63
Overnight	7804	9597	81.32	6960	9438	73.74	7810	9635	81.06
Shuang Ho Hospital									
Day	9717	13,103	74.16	9717	13,931	69.75	9554	13,875	68.86
Night	11,886	15,317	77.60	11,886	15,268	77.85	11,861	16,040	73.95
Overnight	11,884	15,053	78.95	11,884	15,337	77.49	11,759	15,030	78.24

[a]Each log entry corresponds to 1 handover document for a single patient within a shift.

[b]"Number of documents with revisions" denotes documents in which the generative artificial intelligence–generated draft was edited before confirmation.

[c]"Number of documents" denotes the total number of patient-level handover documents.

[d]The revision rate (%) was calculated as documents with revisions/number of documents × 100.

Estimated Time Savings in Handover Documentation Before and After LLM Integration

Nurses at TMUH, WFH, and SHH estimated that they typically completed handover documentation for 5-7 patients per shift when using the original NIS. Based on the available preintegration logs (limited to 1 month), the average documentation completion time ranged from 3.45 (SD 3.82) to 4.32 (SD 4.48) minutes per patient across hospitals and shifts (Table 4). Postintegration logs from October to December 2024 (Table 5) further characterized workflow performance by decomposing document completion time into GenAI generation time and nurse review-revision-confirmation time; the average overall completion time per patient ranged from 1.17 (SD 1.86) to 2.54 (SD 2.82) minutes. Taken together, these figures suggest a potential absolute reduction of approximately 0.91-3.15 minutes per patient ([3.45-2.54]/3.45-[4.32-1.17]/4.32, 26%-73% relative reduction). Moreover, using monthly patient volumes and the observed reduction in per-patient completion time, the LLM-integrated workflow corresponds to estimated savings of approximately 470-980 hours per month across the 3 hospitals (Table 10). Full-time equivalent (FTE) is a standard unit used to express workload or staffing capacity as the work time of 1 full-time worker over a specified period. FTE equivalents were calculated assuming 160 nursing work hours per FTE-month. At TMUH, the projected reduction was approximately 113-282 hours per month (equivalent to 0.7-1.8 FTE/month, assuming 160 working hours per FTE). At WFH, the estimated savings were approximately 160-314 hours per month (1.0-2.0 FTE/month). At SHH, which had the highest patient volume, the estimated savings were approximately 198-387 hours per month (1.2-2.4 FTE/month).

XSL•FO
RenderX

Table 10. Estimated monthly nursing time savings after large language model integration, by hospital (hours and full-time equivalents).[a]

Hospital	October 24, hours (full-time equivalent)[b]	November 24, hours (full-time equivalent)[b]	December 24, hours (full-time equivalent)[b]	Average, hours (full-time equivalent)[b]
Taipei Medical University Hospital	113-273 (0.71-1.71)	115-276 (0.72-1.72)	117-282 (0.73-1.76)	115-277 (0.72-1.73)
Wan Fang Hospital	163-307 (1.02-1.92)	166-314 (1.04-1.96)	160-302 (1.00-1.89)	163-308 (1.02-1.92)
Shuang Ho Hospital	198-381 (1.24-2.38)	203-391 (1.27-2.44)	201-387 (1.25-2.42)	201-386 (1.25-2.41)
Total	474-962 (2.96-6.01)	484-981 (3.02-6.13)	478-970 (2.99-6.07)	479-971 (2.99-6.07)

[a]Hours saved/month = patients × (minutes saved/patient)/60. Preintegration times (September 2024 only; Table 3) were 3.45-4.32 minutes/patient. Postintegration times (October-December 2024; Table 4) were as follows: Taipei Medical University Hospital, 2.13-2.54 minutes/patient; Wan Fang Hospital, 1.17-1.78 minutes/patient; and Shuang Ho Hospital 1.74-2.11 minutes/patient. Bounds were computed as pre-min – post-max and pre-max – post-min.

[b]Values are listed as ranges.

Cost-Effectiveness of GenAI-Assisted Handover Documentation

Preliminary GenAI token cost-effectiveness was evaluated by comparing GenAI operating expenditures with estimated nursing time savings (Table 11). Monthly patient volume was used to approximate documentation time saved, and nursing time was valued at US $10 per hour to estimate the equivalent labor value (opportunity cost) of time freed for clinical work (ie, this does not represent direct budget savings unless staffing levels or overtime are reduced). On a monthly basis, allocated GenAI operating costs were approximately US $493-US $986 per hospital, corresponding to an estimated 113-391 hours saved per hospital per month. The resulting monthly net labor value ranged from US $322 to US $2100 at TMUH, US $970 to US $2595 at WFH, and US $1024 to US $3082 at SHH. Aggregated across hospitals, the estimated time savings were 474-981 hours per month, equivalent to US $4740-US $9810 in labor value. After subtracting monthly GenAI operating expenditures (US $1871-US $2464), the implied net labor value was approximately US $2276-US $7939 per month. Over the 3-month period, the total GenAI cost was US $6396.97, with an estimated 1436-2913 hours saved, yielding an overall net labor value of US $7963-US $22,733.

XSL•FO
RenderX

Table 11. Estimated monthly generative artificial intelligence operating costs, hours saved, equivalent labor value, and net labor value by hospital (October-December 2024).

Hospital and measure	October 24	November 24	December 24	Overall (October to December)
Taipei Medical University Hospital				
Allocated generative artificial intelligence cost, US $[a]	630.18	688.92	848.02	2167.11
Time saved (hours)[b], range	113-273	115-276	117-282	345-831
Equivalent labor value (US $10/hour), range	1130-2730	1150-2760	1170-2820	3450-8310
Net labor value[c] (US $), range	499.82-2099.82	461.08-2071.08	321.98-1971.98	1282.89-6142.89
Wan Fang Hospital				
Allocated generative artificial intelligence cost, US $	493.14	544.74	630.12	1667.99
Time saved (hours), range	163-307	166-314	160-302	489-923
Equivalent labor value (US $10/hour), range	1630-3070	1660-3140	1600-3020	4890-9230
Net labor value (US $), range	1136.86-2576.86	1115.26-2595.26	969.88-2389.88	3222.01-7562.01
Shuang Ho Hospital				
Allocated generative artificial intelligence cost, US $	747.75	828.36	985.75	2561.87
Time saved (hours), range	198-381	203-391	201-387	602-1159
Equivalent labor value (US $10/hour), range	1980-3810	2030-3910	2010-3870	6020-11,590
Net labor value (US $), range	1232.25-3062.25	1201.64-3081.64	1024.25-2884.25	3458.13-9028.13
All				
Allocated generative artificial intelligence cost, US $	1871.07	2062.02	2463.89	6396.97
Time saved (hours), range	474-961	484-981	478-971	1436-2913
Equivalent labor value (US $10/hour), range	4740-9610	4840-9810	4780-9710	14,360-29,130
Net labor value (US $), range	2868.93-7738.93	2777.98-7747.98	2316.11-7246.11	7963.03-22,733.03

[a]Allocated generative artificial intelligence costs were distributed according to each hospital's monthly patient-volume share.

[b]Time saved was estimated from monthly patient volumes and the reduction in per-patient handover document completion time.

[c]Net labor value was calculated as equivalent labor value − allocated generative artificial intelligence cost.

Discussion

Principal Findings

In this study, we integrated Google Gemini 1.5 Pro into the in-house NIS of TMU Health to improve the efficiency of nursing handover documentation across 3 hospitals in Taiwan. In the LLM-integrated workflow, a nurse initiates handover by selecting a patient, after which the LLM-integrated NIS retrieves relevant patient data and submits section-specific, tailored prompts to the LLM. The LLM generates draft content for each handover section, and the compiled document is returned to the LLM-integrated NIS for nurse review, revision, and finalization. This integration streamlines handover documentation by reducing manual data entry and supporting a more efficient routine workflow.

The LLM-integrated NIS retains the functionalities of the original in-house NIS, thereby enabling seamless incorporation of the LLM into the existing nursing handover documentation workflow. Nurses remain familiar with the system interfaces, obviating the need for adaptation and minimizing challenges related to digital skills training and system usability. Furthermore, nurses can readily interact with the LLM-integrated NIS and access the generated content, ensuring that the system does not interrupt workflow [42,43]. Consequently, nurses can maintain their focus on patient care while benefiting from the time-saving features of the LLM, ultimately improving both clinical and operational outcomes.

Across 3 TMU Health hospitals, postintegration NIS logs showed shorter per-patient handover documentation completion times than the limited 1-month preintegration baseline, corresponding to an estimated reduction of 0.91-3.15 minutes per patient ([3.45-2.54]/3.45-[4.32-1.17]/4.32, 26%-73%). When scaled by monthly patient volumes (Table 10), this reduction translated to approximately 470-980 hours saved per month across hospitals (approximately 0.7-1.8 FTE/month, assuming 160 working hours per FTE). A preliminary token cost evaluation suggested that GenAI operating expenditures (US

XSL•FO
RenderX

$493-US $986 per hospital per month) were modest relative to the estimated labor or opportunity value of nursing time freed for clinical work, yielding a positive net labor value across October-December 2024 (Table 11). Overall, these findings suggest that GenAI-assisted handover documentation may reduce documentation burden and return clinically meaningful nursing time that could be redirected to direct patient care, which is particularly relevant in the context of nursing workforce shortages.

Safety Considerations and Hallucination Mitigation

Hallucinations are a well-recognized limitation of GenAI, particularly in clinical documentation tasks in which factual accuracy is critical [44-46]. During initial development and testing of the LLM-integrated NIS, we observed occasional hallucinations when the LLM was used to extract and structure vital signs from NIS inputs; for example, the model sometimes generated an average body temperature that was not explicitly documented in the source record. To mitigate hallucination risk, we implemented multiple safeguards. First, we iteratively refined prompts to restrict the LLM to reformatting and summarizing only the provided input and to prohibit the introduction of new values or inferences. Second, we required nurse review, correction, and confirmation before finalizing each handover note, thereby adding an additional verification layer. Despite these measures, we considered the residual risk of LLM-generated structured vital-sign outputs to be clinically unacceptable. Accordingly, we discontinued LLM-based structuring of vital signs and instead directly displayed the patient's vital signs in the NIS. By contrast, the LLM appeared more reliable for extracting and synthesizing narrative information (eg, abnormal findings and clinically relevant descriptions) from radiology and examination reports, in which the task emphasized highlighting salient text rather than generating or imputing numeric values.

Limitations

First, only 1 month of preintegration log data were available, limiting the robustness of the baseline for manual handover documentation and constraining pre-post comparisons. Accordingly, the preintegration results should be interpreted as a preliminary benchmark for estimating the potential impact of subsequent GenAI-assisted workflows. Second, the NIS log schema was not sufficiently granular to support several secondary analyses. Specifically, we could not quantify technical reliability indicators (eg, connection failures, API time-outs, error returns, failed generations, or regeneration attempts), nor could we stratify performance by patient complexity (eg, intensive care unit vs general ward) or clinical acuity proxies such as length of stay. Consequently, the reported time and cost estimates reflect system-level averages and may mask heterogeneity in performance for complex cases or under high-load conditions. Third, we did not conduct a formal natural language processing evaluation of generated text quality using standard metrics or a gold-standard reference set [47,48]. Although nursing professionals reviewed and verified the LLM-generated handover content as part of routine workflow, the absence of quantitative natural language processing evaluation limits comparability with prior summarization studies and precludes objective assessment of section-level content fidelity. Fourth, although we quantified the revision rate of GenAI-generated drafts, we did not classify edit types (eg, corrections vs clarifications vs additions) or examine the reasons underlying nurse modifications. Therefore, we could not determine which content areas most frequently required changes or whether revisions primarily reflected model limitations, workflow preferences, or patient- and context-specific factors. Fifth, we did not conduct a formal usability or user experience evaluation. Specifically, we did not administer validated instruments (eg, the System Usability Scale or Technology Acceptance Model) or conduct structured interviews to assess nurses' perceived workload, trust in GenAI outputs, or overall satisfaction, limiting conclusions about user acceptability beyond log-based performance metrics.

Future Directions

Our preliminary findings suggest that LLMs can support rapid, automated generation of nursing handover documents; however, nurse review and revision remain necessary before finalization. First, future work will expand the NIS logging framework to capture reliability and workflow metrics (eg, time-outs, error codes, retries, regeneration events, and editing behaviors) and enable more granular performance evaluation, including downstream operational outcomes such as nursing overtime hours. Second, we will compare LLM-generated drafts with nurse-finalized handovers to characterize error patterns, quantify revision types, and inform iterative prompt and model refinement. These analyses will be guided by predefined quality indicators and evaluation frameworks (eg, quantitative and qualitative ISBAR-based assessments) in future studies. Collectively, these data will support improvements in factual accuracy and the development of additional safeguards, including rule-based checks and retrieval-augmented generation strategies [49]. Third, to evaluate usability and acceptance, we will incorporate standardized instruments and structured qualitative feedback to assess nurse-perceived usability, trust, cognitive load, and workflow impact alongside log-derived efficiency outcomes. Fourth, to improve transferability, we will evaluate interoperability in settings with different EMR architectures and develop a configurable prompt-and-template layer (eg, leveraging HL7/FHIR [Health Level 7/Fast Healthcare Interoperability Resources] where available) to reduce site-specific customization. Specifically, we aim to reduce institution-specific effort related to data mapping; integration with source systems (eg, laboratory information system/radiology information system/examination information system); authentication and authorization; local handover templates and documentation policies; and network or operational configurations. In parallel, we will preserve and further validate transferable components, including ISBAR-aligned section prompts, the human-in-the-loop workflow (draft generation with nurse review-revision-confirmation), safety safeguards (eg, prompt constraints and use of source-of-truth structured data), and a log-based evaluation framework for efficiency, revision behavior, and reliability metrics. Finally, we will assess generalizability to other clinical documentation workflows,

XSL•FO
RenderX

including automated note drafting and radiology report generation.

Conclusions

In this study, we integrated an LLM into an existing NIS to generate nursing handover documents in routine practice without modifying the user interface or disrupting established workflows. Across 3 hospitals within TMU Health, GenAI-assisted handover documentation was associated with reduced per-patient documentation completion time in real-world use. A preliminary cost-effectiveness analysis suggested that GenAI operating expenditures were modest relative to the estimated labor value of nursing time freed for clinical work from October to December 2024. To mitigate hallucination risk, particularly for structured numeric data, we constrained prompts and embedded nurse review, editing, and confirmation within the workflow. Overall, these findings support the operational feasibility of integrating GenAI into hospital documentation workflows while retaining human verification. Future work will expand NIS logging to quantify reliability and editing metrics, compare LLM-generated drafts with nurse-finalized documents to identify error patterns and inform prompt and model refinement, and evaluate interoperability and generalizability to other clinical documentation tasks.

Acknowledgments

We thank all the nursing staff of TMU Health for their support. This research was supported by TMU in Taiwan (grant TMU 109-5400-025-400). OpenAI's ChatGPT (GPT-5.2) was used solely for grammar and spelling checks. All edits were reviewed by the authors, and no patient-identifiable data were entered into the tool.

Data Availability

The data are available upon reasonable request from the corresponding author (YSL) and with TMU-JIRB approval; they are not publicly available due to ethical and privacy restrictions.

Authors' Contributions

Conceptualization: YSL
Formal analysis: LWT
Investigation: SSC, STSH
Methodology: YSL
Supervision: RJC, MSW
Writing – original draft: YSL

Conflicts of Interest

None declared.

Multimedia Appendix 1
Monthly patient volume, nurse staff levels, shift counts, and mean length of stay, stratified by hospital and shift type (September-December 2024).
[DOCX File , 24 KB - jmir_v28i1e81604_app1.docx]

References

1. Oulton JA. The global nursing shortage: an overview of issues and actions. Policy Polit Nurs Pract 2006 Aug;7(3 Suppl):34S-39S. [doi: 10.1177/1527154406293968] [Medline: 17071693]
2. Marć M, Bartosiewicz A, Burzyńska J, Chmiel Z, Januszewicz P. A nursing shortage - a prospect of global and local policies. Int Nurs Rev 2019 Mar;66(1):9-16. [doi: 10.1111/inr.12473] [Medline: 30039849]
3. Qureshi SM, Bookey-Bassett S, Purdy N, Greig MA, Kelly H, Neumann WP. Modelling the impacts of COVID-19 on nurse workload and quality of care using process simulation. PLoS One 2022;17(10):e0275890 [FREE Full text] [doi: 10.1371/journal.pone.0275890] [Medline: 36228015]
4. Ribeiro OMPL, Coimbra VMO, Pereira SCDA, Faria ADCA, Teles PJFC, Rocha CGD. Impact of COVID-19 on the environments of professional nursing practice and nurses' job satisfaction. Int J Environ Res Public Health 2022 Dec 16;19(24):ijerph192416908 [FREE Full text] [doi: 10.3390/ijerph192416908] [Medline: 36554789]
5. NCSBN research projects significant nursing workforce shortages and crisis. National Council of State Boards of Nursing. URL: https://www.ncsbn.org/news/ncsbn-research-projects-significant-nursing-workforce-shortages-and-crisis [accessed 2026-01-26]
6. State of the world's nursing report 2025: investing in education, jobs, leadership and service delivery. World Health Organization. URL: https://www.who.int/publications/i/item/9789240110236 [accessed 2026-01-26]

XSL•FO
RenderX

7. Liao Z, Sun S, Clarissa C, Aviles L, Lin C, Kao CT, et al. Exploring the challenges of Taiwanese nurses in the COVID-19 post-pandemic era. J Formos Med Assoc 2024 Nov;123 Suppl 3:S215-S222 [FREE Full text] [doi: 10.1016/j.jfma.2024.03.020] [Medline: 38570237]
8. Wu F, Sun J, Wu H, Su M, Lin C. The experiences of public health nurses during the COVID-19 in Taiwan: a qualitative study. Nurs Open 2024 Jan;11(1):e2091 [FREE Full text] [doi: 10.1002/nop2.2091] [Medline: 38268278]
9. Navarro Martínez O, Igual García J, Salcedo VT. Nurses' view of benefits, enablers and constraints to the use of digital health tools with patients: a cross-sectional study. Digit Health 2023 Sep 04;9:20552076231197339 [FREE Full text] [doi: 10.1177/20552076231197339] [Medline: 37675060]
10. Huter K, Krick T, Domhoff D, Seibert K, Wolf-Ostermann K, Rothgang H. Effectiveness of digital technologies to support nursing care: results of a scoping review. J Multidiscip Healthc 2020;13:1905-1926 [FREE Full text] [doi: 10.2147/JMDH.S286193] [Medline: 33328736]
11. Abd El Mongy Ismail B, Mamdouh Elkholy S, Mohamed Said R, Said Elwakeil E. Digital nurse' competencies in relation to patient safety and teamwork communication. Egyptian Journal of Health Care 2025 Mar 01;16(1):705-717. [doi: 10.21608/ejhc.2025.417311]
12. Kutney-Lee A, Kelly D. The effect of hospital electronic health record adoption on nurse-assessed quality of care and patient safety. J Nurs Adm 2011 Nov;41(11):466-472. [doi: 10.1097/NNA.0b013e3182346e4b] [Medline: 22033316]
13. Alotaibi YK, Federico F. The impact of health information technology on patient safety. Saudi Med J 2017 Dec;38(12):1173-1180 [FREE Full text] [doi: 10.15537/smj.2017.12.20631] [Medline: 29209664]
14. McCarthy B, Fitzgerald S, O'Shea M, Condon C, Hartnett-Collins G, Clancy M, et al. Electronic nursing documentation interventions to promote or improve patient safety and quality care: a systematic review. J Nurs Manag 2019 Apr;27(3):491-501. [doi: 10.1111/jonm.12727] [Medline: 30387215]
15. Douma MC, Rejeb MB, Zardoub N, Braham A, Chouchene H, Bouallegue O, et al. Impact of implementing electronic nursing records on quality and safety indicators in care. Libyan J Med 2024 Dec 31;19(1):2421625 [FREE Full text] [doi: 10.1080/19932820.2024.2421625] [Medline: 39570988]
16. Bakhoum N, Gerhart C, Schremp E, Jeffrey AD, Anders S, France D, et al. A time and motion analysis of nursing workload and electronic health record use in the emergency department. J Emerg Nurs 2021 Sep;47(5):733-741. [doi: 10.1016/j.jen.2021.03.007] [Medline: 33888334]
17. Duffield C, Diers D, O'Brien-Pallas L, Aisbett C, Roche M, King M, et al. Nursing staffing, nursing workload, the work environment and patient outcomes. Appl Nurs Res 2011 Nov;24(4):244-255. [doi: 10.1016/j.apnr.2009.12.004] [Medline: 20974086]
18. Blay N, Roche MA, Duffield C, Gallagher R. Intrahospital transfers and the impact on nursing workload. J Clin Nurs 2017 Dec;26(23-24):4822-4829. [doi: 10.1111/jocn.13838] [Medline: 28382638]
19. Shihundla RC, Lebese RT, Maputle MS. Effects of increased nurses' workload on quality documentation of patient information at selected primary health care facilities in Vhembe District, Limpopo Province. Curationis 2016 May 13;39(1):1545 [FREE Full text] [doi: 10.4102/curationis.v39i1.1545] [Medline: 27246793]
20. Lavander P, Meriläinen M, Turkki L. Working time use and division of labour among nurses and health-care workers in hospitals - a systematic review. J Nurs Manag 2016 Nov;24(8):1027-1040. [doi: 10.1111/jonm.12423] [Medline: 27581093]
21. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 06;6(1):120. [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
22. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
23. Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton E, et al. Applications and concerns of ChatGPT and other conversational large language models in health care: systematic review. J Med Internet Res 2024 Nov 07;26:e22769 [FREE Full text] [doi: 10.2196/22769] [Medline: 39509695]
24. Meskó B. The impact of multimodal large language models on health care's future. J Med Internet Res 2023 Nov 02;25:e52865 [FREE Full text] [doi: 10.2196/52865] [Medline: 37917126]
25. Zhang K, Meng X, Yan X, Ji J, Liu J, Xu H, et al. Revolutionizing health care: the transformative impact of large language models in medicine. J Med Internet Res 2025 Jan 07;27:e59069 [FREE Full text] [doi: 10.2196/59069] [Medline: 39773666]
26. Vert J. How will generative AI disrupt data science in drug discovery? Nat Biotechnol 2023 Jun;41(6):750-751. [doi: 10.1038/s41587-023-01789-6] [Medline: 37156917]
27. Ghebrehiwet I, Zaki N, Damseh R, Mohamad MS. Revolutionizing personalized medicine with generative AI: a systematic review. Artif Intell Rev 2024 Apr 25;57(5):128. [doi: 10.1007/s10462-024-10768-5]
28. Patrinos GP, Sarhangi N, Sarrami B, Khodayari N, Larijani B, Hasanzad M. Using ChatGPT to predict the future of personalized medicine. Pharmacogenomics J 2023 Nov 19;23(6):178-184. [doi: 10.1038/s41397-023-00316-9] [Medline: 37726551]
29. Huang J, Neill L, Wittbrodt M, Melnick D, Klug M, Thompson M, et al. Generative artificial intelligence for chest radiograph interpretation in the emergency department. JAMA Netw Open 2023 Oct 02;6(10):e2336100 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36100] [Medline: 37796505]

XSL•FO
RenderX

30. Hong EK, Roh B, Park B, Jo J, Bae W, Soung Park J, et al. Value of using a generative ai model in chest radiography reporting: a reader study. Radiology 2025 Mar;314(3):e241646. [doi: 10.1148/radiol.241646] [Medline: 40067108]
31. Baxter SL, Longhurst CA, Millen M, Sitapati AM, Tai-Seale M. Generative artificial intelligence responses to patient messages in the electronic health record: early lessons learned. JAMIA Open 2024 Jul;7(2):ooae028 [FREE Full text] [doi: 10.1093/jamiaopen/ooae028] [Medline: 38601475]
32. Hobensack M, von Gerich H, Vyas P, Withall J, Peltonen L, Block LJ, et al. A rapid review on current and potential uses of large language models in nursing. Int J Nurs Stud 2024 Jun;154:104753. [doi: 10.1016/j.ijnurstu.2024.104753] [Medline: 38560958]
33. Zong H, Wu R, Cha J, Wang J, Wu E, Li J, et al. Large language models in worldwide medical exams: platform development and comprehensive analysis. J Med Internet Res 2024 Dec 27;26:e66114 [FREE Full text] [doi: 10.2196/66114] [Medline: 39729356]
34. Yun HS, Bickmore T. Online health information-seeking in the era of large language models: cross-sectional web-based survey study. J Med Internet Res 2025 Mar 31;27:e68560 [FREE Full text] [doi: 10.2196/68560] [Medline: 40163112]
35. Tan D, Huang Y, Liu M, Li Z, Wu X, Huang C. Identification of online health information using large pretrained language models: mixed methods study. J Med Internet Res 2025 May 14;27:e70733 [FREE Full text] [doi: 10.2196/70733] [Medline: 40367512]
36. Seibert K, Domhoff D, Bruch D, Schulte-Althoff M, Fürstenau D, Biessmann F, et al. Application scenarios for artificial intelligence in nursing care: rapid review. J Med Internet Res 2021 Nov 29;23(11):e26522 [FREE Full text] [doi: 10.2196/26522] [Medline: 34847057]
37. Ronquillo CE, Peltonen L, Pruinelli L, Chu CH, Bakken S, Beduschi A, et al. Artificial intelligence in nursing: priorities and opportunities from an international invitational think-tank of the Nursing and Artificial Intelligence Leadership Collaborative. J Adv Nurs 2021 Sep;77(9):3707-3717 [FREE Full text] [doi: 10.1111/jan.14855] [Medline: 34003504]
38. Yelne S, Chaudhary M, Dod K, Sayyad A, Sharma R. Harnessing the power of AI: a comprehensive review of its impact and challenges in nursing science and healthcare. Cureus 2023 Nov;15(11):e49252 [FREE Full text] [doi: 10.7759/cureus.49252] [Medline: 38143615]
39. El Arab RA, Al Moosa OA, Abuadas FH, Somerville J. The role of AI in nursing education and practice: umbrella review. J Med Internet Res 2025 Apr 04;27:e69881 [FREE Full text] [doi: 10.2196/69881] [Medline: 40072926]
40. Gemini API: Google AI for developers. Google LLC. URL: https://ai.google.dev/gemini-api/docs [accessed 2026-01-28]
41. Burgess A, van Diggele C, Roberts C, Mellis C. Teaching clinical handover with ISBAR. BMC Med Educ 2020 Dec 03;20(Suppl 2):459 [FREE Full text] [doi: 10.1186/s12909-020-02285-0] [Medline: 33272274]
42. Yoo HJ, Lee H. Critical role of information and communication technology in nursing during the COVID-19 pandemic: a qualitative study. J Nurs Manag 2022 Nov;30(8):3677-3685 [FREE Full text] [doi: 10.1111/jonm.13880] [Medline: 36325914]
43. Abdolkhani R, Petersen S, Walter R, Zhao L, Butler-Henderson K, Livesay K. The impact of digital health transformation driven by COVID-19 on nursing practice: systematic literature review. JMIR Nurs 2022 Aug 30;5(1):e40348 [FREE Full text] [doi: 10.2196/40348] [Medline: 35867838]
44. Birhane A, Kasirzadeh A, Leslie D, Wachter S. Science in the age of large language models. Nat Rev Phys 2023 Apr 26;5(5):277-280. [doi: 10.1038/s42254-023-00581-4]
45. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
46. Goodman KE, Yi PH, Morgan DJ. AI-generated clinical summaries require more than accuracy. JAMA 2024 Feb 27;331(8):637-638. [doi: 10.1001/jama.2024.0555] [Medline: 38285439]
47. Bala S, Keniston A, Burden M. Patient perception of plain-language medical notes generated using artificial intelligence software: pilot mixed-methods study. JMIR Form Res 2020 Jun 05;4(6):e16670 [FREE Full text] [doi: 10.2196/16670] [Medline: 32442148]
48. Abbasian M, Khatibi E, Azimi I, Oniani D, Shakeri Hossein Abad Z, Thieme A, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digit Med 2024 Mar 29;7(1):82 [FREE Full text] [doi: 10.1038/s41746-024-01074-z] [Medline: 38553625]
49. Jeong M, Sohn J, Sung M, Kang J. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. Bioinformatics 2024 Jun 28;40(Suppl 1):i119-i129 [FREE Full text] [doi: 10.1093/bioinformatics/btae238] [Medline: 38940167]

Abbreviations

AI: artificial intelligence
API: application programming interface
EMR: electronic medical record
FHIR: Fast Healthcare Interoperability Resources

XSL•FO
RenderX

FTE: full-time equivalent
GEE: generalized estimating equation
GenAI: generative artificial intelligence
HIPAA: Health Insurance Portability and Accountability Act
HL7: Health Level 7
ISBAR: Identify, Situation, Background, Assessment, Recommendation
JIRB: Joint Institutional Review Board
LLM: large language model
NIS: nursing information system
RIS: radiology information system
SHH: Shuang Ho Hospital
TMU: Taipei Medical University
TMUH: Taipei Medical University Hospital
WFH: Wan Fang Hospital
WHO: World Health Organization

Edited by A Coristine; submitted 01.Aug.2025; peer-reviewed by KE Zavotsky, N Bievre, F Al Dhabbari, S Al Mukhtar; comments to author 02.Dec.2025; revised version received 26.Feb.2026; accepted 26.Feb.2026; published 12.Mar.2026.

Please cite as:
Chen RJ, Wu MS, Tsai LW, Chang SS, Shen Hsiao ST, Lo YS
Integrating a Large Language Model to Streamline Nursing Handover Documentation Across Multiple Hospitals in Taiwan: Development and Implementation Study
J Med Internet Res 2026;28:e81604
URL: https://www.jmir.org/2026/1/e81604
doi:10.2196/81604
PMID:

©Ray-Jade Chen, Mai-Szu Wu, Lung-Wen Tsai, Shy-Shin Chang, Shu-Tai Shen Hsiao, Yu-Sheng Lo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Health Technology Use Among Rehabilitation Professionals in China: Multi-Province Cross-Sectional Survey

Shuning Duan[1,2], MSc; Zi-Ru Wang[1], PhD; Xier Chen[2], MSc; Guoxin Ni[1], MD, PhD

[1]Department of Rehabilitation Medicine, The First Affiliated Hospital of Xiamen University, School of Medicine, Xiamen University, Xiamen, China
[2]School of Sports Medicine, Wuhan Sports University, Wuhan, Hubei, China

Corresponding Author:
Guoxin Ni, MD, PhD
Department of Rehabilitation Medicine
The First Affiliated Hospital of Xiamen University
School of Medicine, Xiamen University
No. 55 Zhenhai Road
Siming District
Xiamen, 361003
China
Phone: 86 5922137084
Email: nigx@xmu.edu.cn

Abstract

Background: The rapid expansion of rehabilitation needs in China has intensified pressure on a workforce that remains unevenly distributed. Digital health technologies (DHTs) offer potential to increase service reach and efficiency. However, little is known about how rehabilitation professionals currently gather and document clinical information, nor about their readiness to integrate digital tools into routine practice within China's rapidly digitalizing health system.

Objective: This study aimed to describe how rehabilitation professionals in China collect subjective and objective clinical information, document patient data in routine practice, and assess their willingness to use DHTs in clinical settings.

Methods: We conducted a multi-province observational cross-sectional survey using a culturally adapted questionnaire based on the World Health Organization Digital Health Interventions framework. The instrument assessed participant characteristics, information collection methods, documentation practices, and willingness to adopt digital functions across rehabilitation activities. Descriptive analyses and subgroup comparisons were performed on 324 complete responses from certified rehabilitation professionals. The multi-province cross-sectional online survey was conducted among licensed rehabilitation professionals in China with internet access. Participants were recruited through professional networks and social media platforms.

Results: Respondents were drawn from 20 provincial-level administrative regions across China, including Fujian (n=72), Guangdong (n=77), and Shanxi (n=45), among others, with 82.7% (268/324) employed in public sector rehabilitation services. Traditional methods dominated clinical work. Face-to-face communication was used frequently for subjective assessment by 96.3% (312/324) of respondents, whereas digital channels such as email (22/324, 6.8%) and telephone (47/324, 14.5%) saw limited use. For objective information, visual observation (271/324, 83.7%) and manual measurement tools (195/324, 60.2%) remained the primary approaches, while motion capture technology (45/324, 13.8%) and wearable sensors (13/324, 4%) were rarely used. Documentation practices also relied heavily on analogue formats, with 82.1% (266/324) using handwritten notes and 60.2% (195/324) using paper templates. In contrast, willingness to adopt DHTs was consistently high, with 80.6% (261/324) of respondents indicating readiness to use digital systems for identity verification, 79.0% (256/324) for progress tracking, and 78.1% (253/324) for outcome measurement. Subgroup analyses revealed that educational level significantly influenced the adoption of advanced technologies, with master's or doctoral degree holders reporting higher use of sensor-based assessment, motion capture, and wearable devices. In contrast, professional title and clinical specialty showed limited influence, with no significant differences observed for most digital health functions.

Conclusions: Rehabilitation professionals in China demonstrate strong readiness to use DHTs, yet their routine practice remains largely paper-based and analogue. These findings provide evidence to inform implementation strategies, workforce training, and system-level planning aimed at accelerating digital transformation in rehabilitation services.

(J Med Internet Res 2026;28:e90623) doi:10.2196/90623

XSL•FO
RenderX

KEYWORDS

digital health technologies; telehealth; rehabilitation; online survey; health technology acceptance

Introduction

Rehabilitation is widely recognized as a core component of modern health systems. Global estimates indicate that more than 2 billion people live with health conditions that could benefit from rehabilitation, and this number continues to grow as populations age and chronic conditions become more common [1,2]. The role of rehabilitation in restoring function, promoting independence, and supporting healthy aging has been increasingly acknowledged in global public health priorities [3,4]. Rehabilitation is a series of interventions for individuals who are experiencing or may experience limitations in everyday functioning to help them obtain and maintain the optimum functional status [5,6]. Globally, rehabilitation services primarily address functional impairments resulting from injury, illness, or chronic conditions, encompassing disciplines such as physiotherapy, occupational therapy, and speech-language therapy [5,7,8]. Within China's health care system, this scope aligns with the professional classifications established by the National Health Commission (NHC) and the Chinese Association of Rehabilitation Medicine, focusing on physical and functional rehabilitation rather than substance use disorder or mental health rehabilitation [9,10]. Insights into rehabilitation professionals' current digital health practices and readiness can guide the successful implementation of these technologies in clinical settings.

Despite this growing demand, access to rehabilitation is uneven across regions. The World Health Organization (WHO) has noted that the density of rehabilitation professionals remains far below required levels in many settings, especially in low-and-middle income regions [11,12]. Previous studies have revealed that China ranked first globally in rehabilitation needs [13], yet faces severe supply shortages and regional disparities. China faces a severe workforce shortage, with only an estimated 150,000 physiotherapists serving a population of 1.4 billion. This equates to approximately 10.7 therapists per 100,000 population [14], far below the density in Australia, where physiotherapists alone account for 153.2 per 100,000 population [15,16]. Evidence indicates that China's digital health infrastructure has expanded rapidly, with 900 internet hospitals and over 24,000 telemedicine facilities established nationwide by October 2020 [17]. However, persistent institutional barriers remain. Information silos across different levels of rehabilitation institutions impede data sharing and care coordination [18,19]. Coupled with this, gaps in the payment system exclude most digital solutions from national health insurance coverage [20,21]. In such large and complex health systems, the scale of the population and the fragmentation of service delivery further complicate efforts to expand rehabilitation capacity in a timely and coordinated manner. These challenges place additional strain on existing health care infrastructure and highlight the need for a more sustainable and scalable approach to delivering rehabilitation that can extend reach, support continuity of care, and make efficient use of limited human resources.

Digital health technologies (DHTs), as defined by the WHO, involve the use of digital and smart technologies, including artificial intelligence, the internet of things, big data, robotics, and connected devices, to improve health outcomes [22]. These technologies offer significant advantages in efficiency, scalability, and user accessibility. Evidence suggests that digital tools can support communication, monitoring, and self-management; contribute to improved outcomes; and reduce costs [23-25]. For example, in musculoskeletal rehabilitation, DHTs have emerged as transformative solutions to address the limitations of conventional care, with recent evidence demonstrating that artificial intelligence-driven motion analysis, wearable sensors, and tele-rehabilitation platforms can achieve outcomes comparable or superior to traditional approaches, alongside adherence gains of 15-40% and cost reductions of approximately 30-40% [26]. In post-trauma rehabilitation, remote digital interventions delivered via video conferencing and virtual reality (VR) have shown positive effects on physical and psychosocial outcomes for individuals with spinal cord injuries [27]. Wearable sensor technology has also gained traction in routine rehabilitation care; studies indicate that both clinicians and patients recognize its value in objectively measuring activity outside clinical settings, with most clinicians preferring continuous monitoring integrated into electronic health records to optimize workflow [28,29]. In stroke rehabilitation, home-based DHTs, including wearable devices, smartphone apps, and sensor-based solutions, have demonstrated high feasibility and usability for monitoring mobility, upper extremity function, and daily living activities [30]. Digital avatars represent another emerging modality. These approaches may be especially valuable where access to rehabilitation is limited by distance, concentration of services in urban centers, or shortages of specialized staff.

Despite the potential of DHTs to transform rehabilitation services, their development and application among rehabilitation professionals in China remain limited. Cross-disciplinary collaboration among policymakers, health care providers, technology developers, and end-users is essential to ensure that these digital solutions align with the needs of rehabilitation professionals and the patients they serve. Existing evidence suggests that digital health adoption varies across health systems and professional groups, influenced by factors such as workflow integration, technical infrastructure, and system-level support [31-33]. However, most of the existing work has focused on general practice or nursing professions, while the DHT practices of rehabilitation professionals remain comparatively underexamined. This gap is particularly notable in middle-income countries, where digital health infrastructure is expanding but the role of these tools in everyday rehabilitation practice has not been well documented [34,35]. Generating empirical data from such settings can help clarify current patterns of DHTs use and inform future implementation strategies that support appropriate and equitable integration of digital health within rehabilitation services. The findings will

XSL•FO
RenderX

help guide future research in the design of DHTs as well as policymakers.

Methods

Study Design

A multi-province cross-sectional survey was conducted to describe the use of DHTs among rehabilitation professionals in China. The study was reported in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [36] (Multimedia Appendix 1). Only complete and eligible survey responses were included in the analysis.

Instrument Development and Reliability

The survey was adapted from the clinician version of the digital health interventions (DHI) questionnaire developed by Merolli et al [37], which was created through qualitative item review of the WHO DHI classification framework [38]. The questionnaire was translated into Chinese and culturally adapted following established guidelines. The adaptation process included forward translation by 2 bilingual researchers, expert review by 5 Chinese rehabilitation specialists, back translation to ensure conceptual equivalence, cognitive interviews with 10 rehabilitation therapists, and pilot testing with 30 therapists. All adaptations focused on maintaining conceptual equivalence while ensuring clarity and relevance for rehabilitation professionals working in the Chinese health system.

The final instrument comprised four sections: (1) demographic and professional characteristics, (2) methods for obtaining subjective and objective clinical information, (3) documentation practices, and (4) willingness to use digital technologies for 32 rehabilitation-related clinical functions. All items assessing frequency and willingness used 5-point Likert scales, adapted from the approach used by Merolli et al [37]. For frequency items, responses were scored from 1 (“Never”) to 5 (“Always”). For willingness items, responses were scored from 1 (“Not at all willing”) to 5 (“very much willing”). In both cases, higher scores indicate greater frequency of use or stronger willingness to adopt. For analysis, responses of 4 or 5 were classified as “users” (for frequency items) or “willing” (for willingness items), following the approach of Merolli et al [37]. A sensitivity analysis using an alternative classification threshold (responses 3-5 defined as users and willing) confirmed that the core findings, particularly the substantial gap between willingness and actual use, remained robust regardless of the cut-off point chosen. The complete Chinese survey instrument and its English translation are provided in Multimedia Appendix 2 and Multimedia Appendix 3, respectively.

Internal consistency was assessed using Cronbach α coefficient. The results showed: subjective information collection domain (9 items): α=0.782; objective information collection domain (10 items): α=0.878; documentation practices domain (12 items): α=0.852; willingness to use digital technologies domain (32 items): α=0.986; total scale (63 items): α=0.963. All reliability coefficients exceeded the acceptable threshold of 0.70, indicating good to excellent internal consistency of the adapted questionnaire in this sample.

Setting and Participants

This cross-sectional study was conducted between April and July 2025 among rehabilitation professionals practicing in mainland China. Eligibility was determined through 2 sequential screening questions at the beginning of the online questionnaire. The first question asked: “Have you passed the national health professional qualification examination organized by the NHC and the Ministry of Human Resources and Social Security, obtained a rehabilitation therapy qualification certificate, and are you currently providing clinical rehabilitation services in China?” (Yes or No). The remaining proceeded to the second screening question: “In the past six months, have you treated or managed an average of at least five rehabilitation patients per week?” (Yes or No). Participants represented all professional levels and practice settings, including general hospitals, rehabilitation specialty hospitals, community health centers, older adult care institutions, and sports medicine facilities.

A combination of convenience and snowball sampling was used. The survey targeted rehabilitation professionals practicing in mainland China, including physical therapists, occupational therapists, speech and language therapists, and other recognized rehabilitation therapy practitioners. In China, professional roles in rehabilitation can sometimes overlap, with practitioners potentially assuming responsibilities across multiple specialty areas. To address this complexity and ensure accurate classification of participants’ primary professional focus, we adopted a classification framework based on core clinical responsibilities. This framework was informed by the professional definitions of World Physiotherapy and the World Federation of Occupational Therapists [39,40], and adapted to align with the specialty classifications of the Chinese Association of Rehabilitation Medicine [41,42] Specifically, respondents were asked to select their primary practice area from 9 predefined categories (eg, neurological rehabilitation, musculoskeletal rehabilitation, and pediatric rehabilitation) and to describe their routine responsibilities in an open-ended question. Based on their responses, we classified participants into the most appropriate professional category for analysis.

Recruitment and Data Collection

Data were collected between April and July 2025 using an anonymous online survey hosted on Questionnaire Star (Changsha Ranxing Science and Technology Co, Ltd), a widely used Chinese survey platform. Recruitment occurred through multiple channels, including professional networks, departmental distribution within hospitals, social media groups for rehabilitation therapists, and peer referral. Potential participants accessed the survey through an open link. The front page of the survey clearly stated the research purpose, assured anonymity, and emphasized voluntary participation; proceeding to the questionnaire implied informed consent. After screening questions confirmed eligibility, participants were able to proceed to the full survey. Duplicate submissions were limited using single-device submission control, and no incentives were offered.

XSL•FO
RenderX

Quality Control

To ensure data quality, we implemented multilayered quality control measures throughout the recruitment process. As illustrated in the participant flowchart, of the 393 initial responses received, 52 were excluded for not meeting eligibility criteria (lack of national qualification or not in clinical practice), and an additional 17 were excluded for not meeting the minimum caseload requirement (fewer than 5 patients per week). Following eligibility screening, responses were further scrutinized through automated checks, including IP address to detect duplicate submissions, and manual review of completion times and response patterns. After these quality control procedures, 324 valid responses were retained for final analysis.

Data Analysis

Data from completed surveys were analyzed descriptively using SPSS (version 27; IBM Corp). Categorical variables were summarized using frequencies and percentages. For Likert-scale items, responses of always or frequently were classified as users, and responses of very much willing or quite a bit were classified as willing, following the approach used by Merolli et al [37]. Descriptive secondary comparisons were conducted across demographic and professional subgroups to explore variation in DHT use and willingness. For all subgroup comparisons, Kruskal-Wallis tests were used due to the non-normal distribution of the data. Post-hoc pairwise comparisons were conducted using Dunn test with Bonferroni correction to control for Type I error. Effect sizes (ε^2) were calculated manually using the formula $\varepsilon^2 = (H - k + 1) / (N - k)$, where k represents the number of groups and N the total sample size. Effect sizes were interpreted as small (0.01), medium (0.06), or large (0.14). Open-ended responses were reviewed using a directed qualitative content approach to summarize commonly reported barriers and facilitators.

Missing data were handled through listwise deletion based on the study's eligibility criteria. After applying the predefined exclusion standards (qualification, practice status, and minimum caseload requirements), 324 complete and valid responses were retained for final analysis, with no missing values in any key variables.

Ethical Considerations

Ethics approval for this study was granted by the ethics committee of the First Affiliated Hospital of Xiamen University (approval XMFHIIT-2025SL023) in accordance with institutional guidelines. Informed consent was obtained from all the participants, and data were anonymized to ensure privacy and confidentiality. No compensation was provided for participation.

Results

Overview

This survey collected a total of 393 questionnaires from 20 provinces across China, of which 324 met the eligibility criteria. The exclusion criterion was failure to pass the NHC qualification examination, not currently practicing in mainland China, or having treated or managed an average of fewer than 5 rehabilitation patients per week over the 6 months preceding the survey. Upon receipt of the research participant invitation letter and relevant study information, voluntary participation in the survey constituted informed consent. Figure 1 presents the detailed participant flow diagram for this study.

XSL•FO
RenderX

Figure 1. Recruitment flowchart for this study.

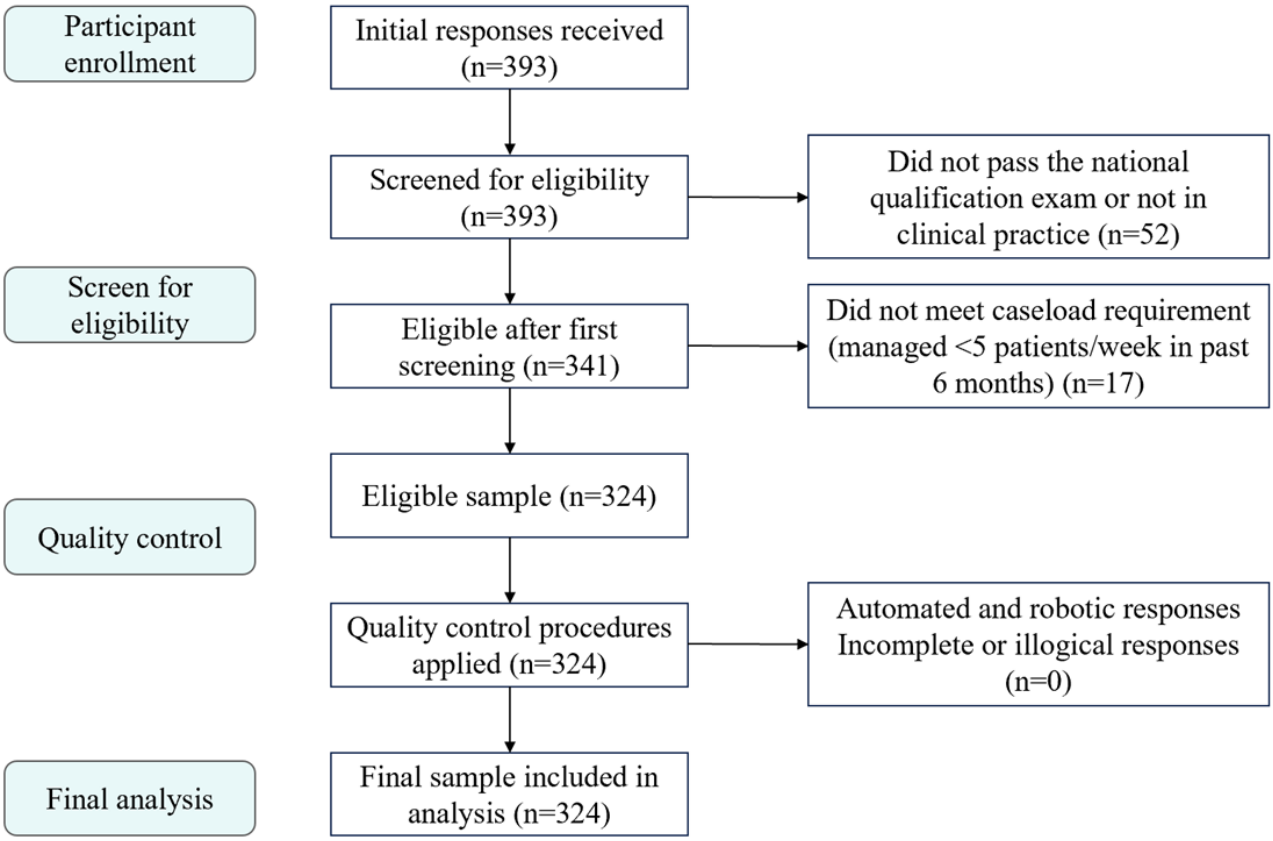

Participant Characteristics

A total of 324 certified rehabilitation professionals were included in the analysis. Most were aged 20 to 29 years (209/324, 64.5%) and female (173/324, 53.4%). Nearly half held the title rehabilitation therapist (158/324, 48.8%), with 26.9% (87/324) classified as senior rehabilitation therapists. Clinical experience was most commonly 2-5 years (157/324, 48.5%), and the majority held a bachelor's degree (209/324, 64.5%). Most participants worked within public-sector general hospital rehabilitation departments (268/324, 82.7%), and internet access was near-universal, primarily via mobile phone (322/324, 99.4%). Table 1 summarizes demographic characteristics. The geographic distribution of respondents is illustrated in Figure 2, with the highest proportions from Guangdong (77/324, 23.8%), Fujian (72/324, 22.2%), and Shanxi (45/324, 13.9%).

XSL•FO
RenderX

Table 1. Participant characteristics (N=324).

Characteristics	Participants, n (%)
Sex	
Male	151 (46.6)
Female	173 (53.4)
Age (years)	
20-29	209 (64.5)
30-39	96 (29.6)
40-49	18 (5.6)
50-59	1 (0.3)
Professional title	
Rehabilitation assistant	60 (18.5)
Rehabilitation therapist	158 (48.8)
Senior rehabilitation therapist	87 (26.9)
Associate chief rehabilitation therapist	16 (4.9)
Chief rehabilitation therapist	3 (0.9)
Years of clinical experience	
<2 years	52 (16)
2-5 year	157 (48.5)
6-10 years	68 (21)
11-15 years	27 (8.3)
16-20 years	13 (4)
>20 years	7 (2.2)
Highest education level	
Associate degree	44 (13.6)
Bachelor’s degree	209 (64.5)
Master’s degree	62 (19.1)
Doctoral degree	9 (2.8)
Work sector (institutional ownership)	
Private institution	55 (17)
Public institution	253 (78.1)
Both public and private	13 (4)
Other	3 (0.9)
Primary work setting (institution type)	
General hospital rehabilitation department	268 (82.7)
Rehabilitation specialty hospital	28 (8.6)
Community health center	6 (1.9)
Elder care institution	12 (3.7)
Sports team or sports medicine center	1 (0.3)
Other	9 (2.8)
Weekly clinical hours	
<5	14 (4.3)
6-10	64 (19.8)
11-20	11 (3.4)

XSL•FO
RenderX

Characteristics	Participants, n (%)
21-30	20 (6.2)
31-40	128 (39.5)
>40	87 (26.9)
Primary clinical area	
Neurological rehabilitation	138 (42.6)
Musculoskeletal rehabilitation	69 (21.3)
Cardiorespiratory rehabilitation	11 (3.4)
Pediatric rehabilitation	35 (10.8)
Geriatric rehabilitation	26 (8)
Trauma rehabilitation	2 (0.6)
Mental health rehabilitation	3 (0.9)
Speech, swallowing, and hearing rehabilitation	29 (9)
Occupational rehabilitation	2 (0.6)
Others	9 (2.8)
Province	
Fujian	72 (22.2)
Guizhou	2 (0.6)
Guangdong	77 (23)
Shanxi	45 (13.9)
Beijing	12 (3.7)
Henan	5 (1.5)
Sichuan	13 (4)
Jilin	4 (1.2)
Jiangxi	5 (1.5)
Shaanxi	10 (3.1)
Gansu	1 (0.3)
Shandong	21 (6.5)
Chongqing	4 (1.2)
Tianjin	2 (0.6)
Hubei	9 (2.8)
Shanghai	13 (4)
Liaoning	1 (0.3)
Anhui	8 (2.5)
Zhejiang	2 (0.6)
Jiangsu	18 (5.5)

XSL•FO
RenderX

Figure 2. Geographic distribution of survey respondents (N=324).



Collection of Subjective Clinical Information

Figure 3 illustrates a marked reliance on in-person communication, with 96.3% (312/324) reporting frequent or constant use of face-to-face discussions, far exceeding any other method. Digital communication channels were rarely incorporated into routine assessment. Only 14.2% (46/324) used e-messages, 14.5% (47/324) used phone calls, and 6.8% (22/324) used email with any regularity. Paper-based patient-reported outcome measures (PROMs) were far more established (202/324, 62.4% frequent use) than electronic formats (99/324, 30.6%), indicating that digital patient-reported data capture has not yet been scaled into routine practice.

Figure 3. Frequency of methods and tools used to collect subjective clinical information. EHR: electronic health record; PROM: patient-reported outcome measure.

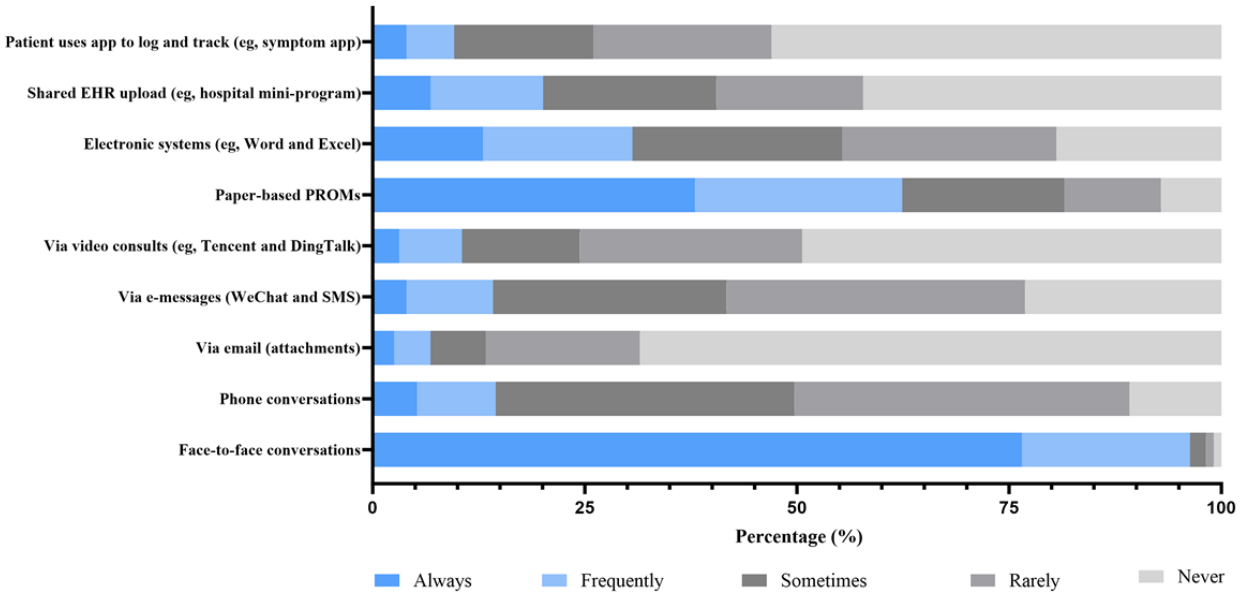

XSL•FO
RenderX

Collection of Objective Clinical Information

Figure 4 demonstrates similar dominance of traditional approaches for objective assessment. Visual observation was the primary method (271/324, 83.7% frequently or always), while manual measurement tools such as goniometers and dynamometers were also widely used (195/324, 60.2%). In contrast, digital motion-capture technologies, VR-based tools, wearable sensors, and patient-generated device data showed limited integration, with frequent use ranging between 4 to 14% across categories. More than 70% (227/324) rarely or never accessed wearable-generated metrics, and video-based assessment remained infrequent. These patterns reflect early-stage digital adoption rather than widespread integration of technology-supported movement analysis.

Figure 4. Frequency of methods and tools used to collect objective clinical information. VR: virtual reality.

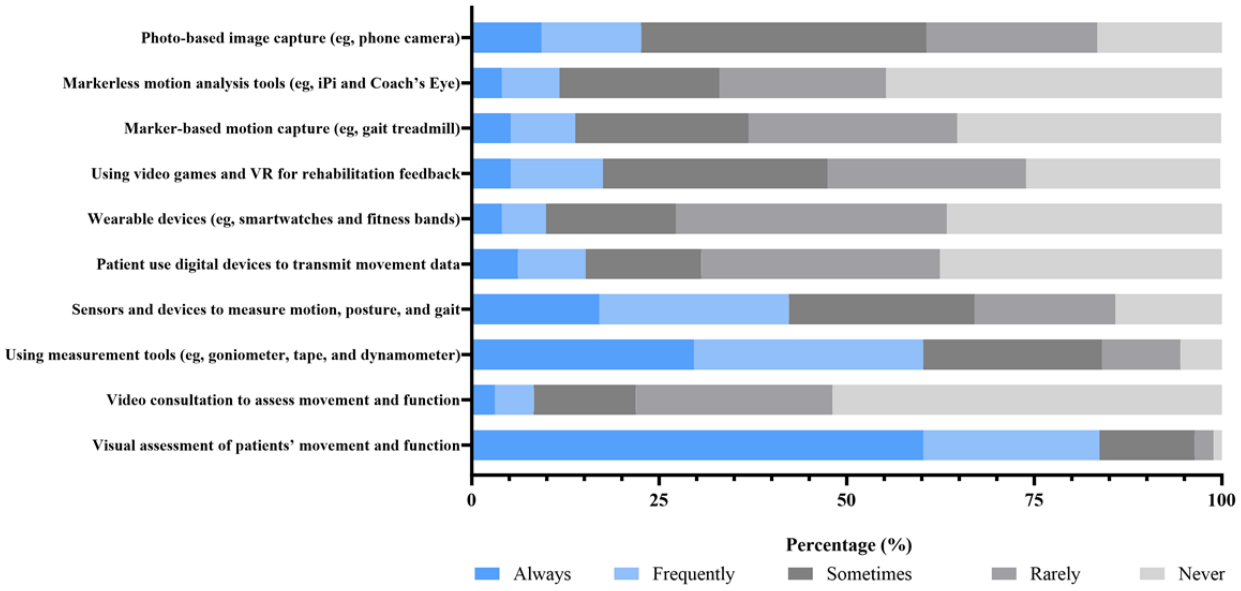

Documentation of Clinical Assessment Information

As displayed in Figure 5, handwritten free-text notes (266/324, 82.1%) and structured paper templates (195/324, 60.2%) formed the dominant documentation format. Unstructured EMR entry was more commonly used than structured coding (183/324, 56.5% vs 69/324, 21.3%), suggesting fragmented rather than standardized digital documentation. Microsoft Word or Microsoft Excel documentation was used occasionally (121/324, 37.4% combined frequent or occasional use), but mobile app–based entries, patient-entered digital notes, audio files, and video and image storage remained minimal, with most respondents reporting rare or no use. The predominance of paper systems suggests a slow transition toward digitalized information environments.

XSL•FO
RenderX

Figure 5. Frequency of methods and tools used to document clinical assessment information. EHR: electronic health record; EMR: electronic medical record; ICD-10: International Statistical Classification of Diseases, Tenth Revision; ICF: International Classification of Functioning, Disability and Health.

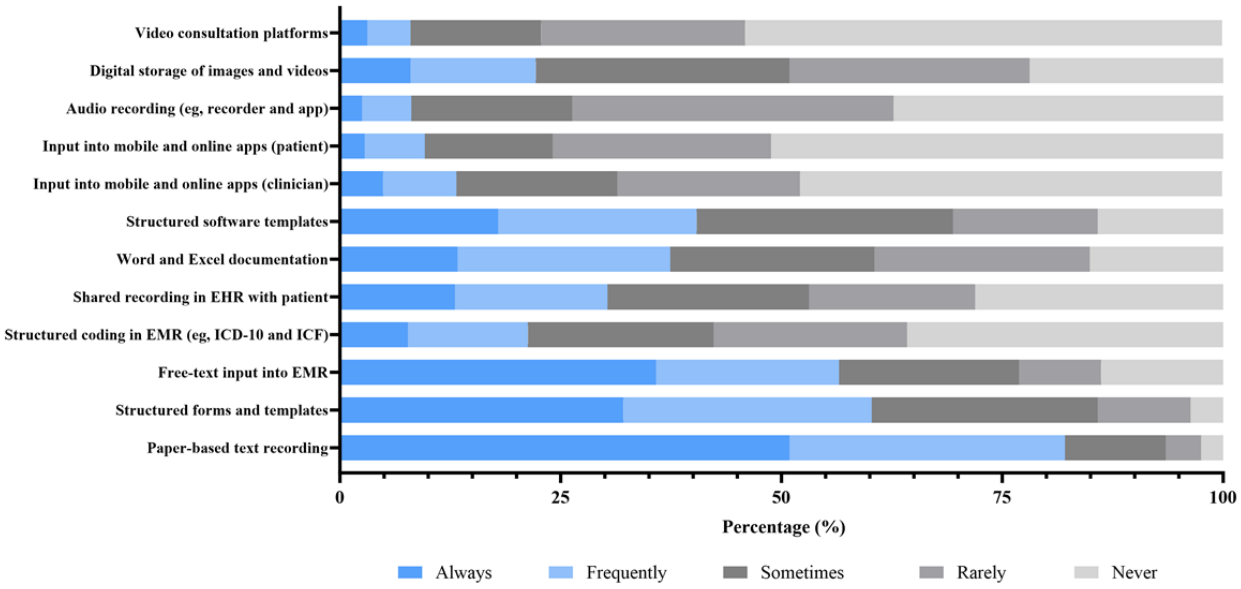

Willingness to Use Digital Technologies

Figure 6 shows generally high willingness across nearly all digital functions assessed. High proportions of respondents were willing to use digital systems for identity verification (261/324, 80.6%), progress tracking (256/324, 79.0%), and outcome measurement (253/324, 78.1%), all exceeding 75%. Willingness also exceeded 70% for accessing imaging (235/324, 72.5%), receiving clinical alerts (231/324, 71.3%), and recording treatment changes (228/324, 70.4%). Lower but still favorable levels were observed for teleconsultation (179/324, 55.2%) and digital referral coordination (185/324, 57.1%). The overall mean willingness score across these digital functions was 4.02, with moderate SD values (0.90-1.11), indicating consistent support for digital adoption within the cohort. Overall, expressed willingness was consistently higher than reported current use across digital domains, indicating a gap between readiness and practice.

XSL•FO
RenderX

Figure 6. Willingness to use digital technologies for rehabilitation-related clinical activities. CT: computed tomography; ICD-10: International Statistical Classification of Diseases, Tenth Revision; ICF: International Classification of Functioning, Disability and Health; MRI: magnetic resonance imaging.

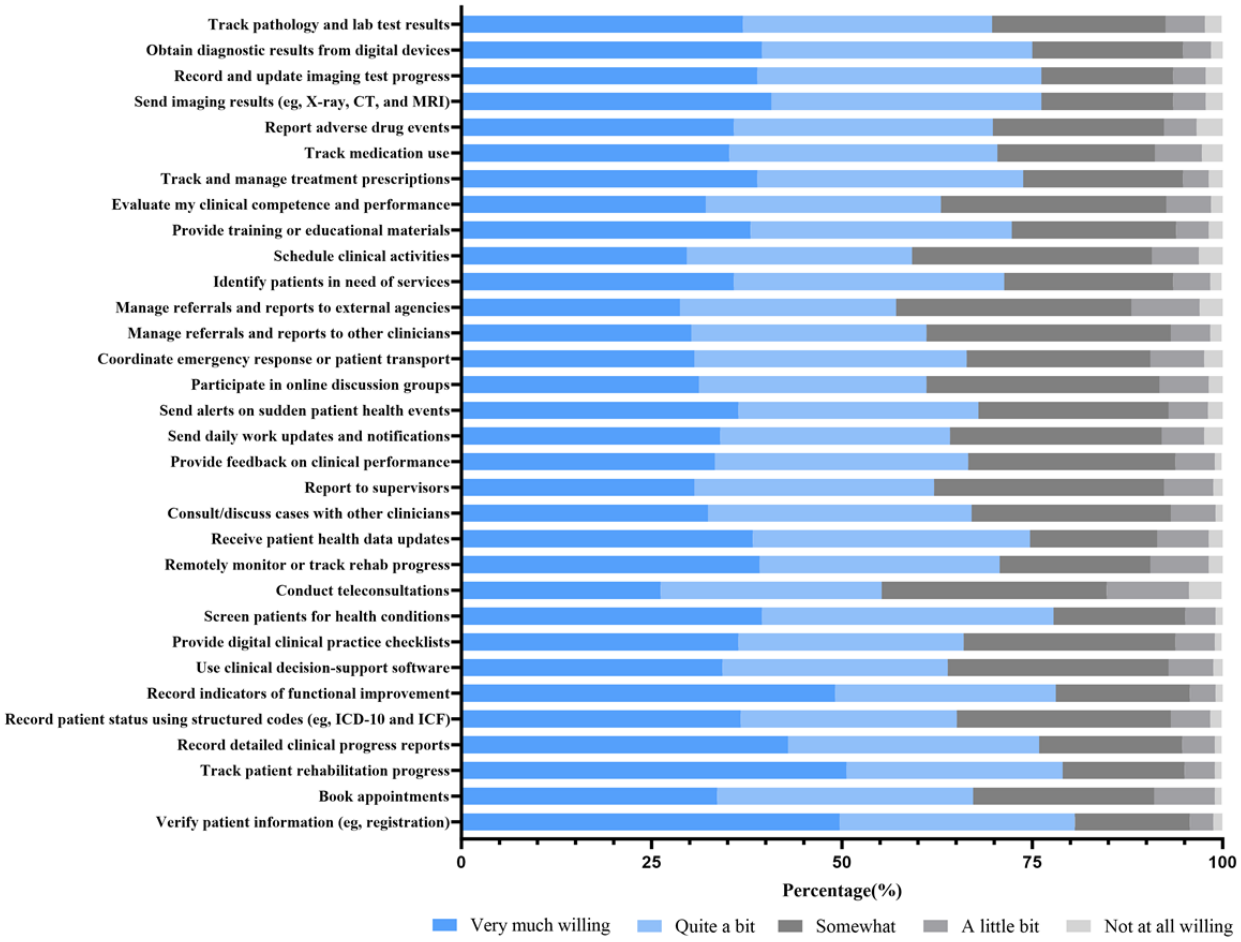

Subgroup Analyses by Demographic and Clinical Specialty

Subgroup analyses showed that educational level significantly influenced the adoption of advanced technologies, while professional title and clinical specialty had limited effects. Full details of the Kruskal-Wallis test results, including H values, *P* values, effect sizes, and post-hoc comparisons, are provided in Multimedia Appendix 4.

Findings revealed significant differences across educational levels for several outcomes. Sensor-based assessment (H=24.801, *P*<.001, ε^2=0.068), marker-based motion capture (H=23.052, *P*<.001, ε^2=0.063), wearable device data (H=13.695, *P*=.003, ε^2=0.033), electronic systems (H=10.823, *P*=.01, ε^2=0.024), and unstructured free-text EMR entry (H=12.214, *P*=.007, ε^2=0.029) all varied significantly, with post-hoc comparisons showing that therapists with master's or doctoral degrees consistently reported higher use than those with associate or bachelor's degrees. By professional title, significant differences were found for sensor-based assessment (H=18.904, *P*<.001, ε^2=0.047), wearable device data (H=13.142, *P*=.01, ε^2=0.029), unstructured free-text EMR entry (H=11.851, *P*=.02, ε^2=0.025), and teleconsultation willingness (H=11.802, *P*=.02, ε^2=0.024). Post-hoc analyses indicated that rehabilitation assistants and therapists generally reported lower use and willingness than senior or associate chief therapists. Across clinical specialties, limited differences were observed: sensor-based assessment (H=25.507, *P*=.002, ε^2=0.052) and face-to-face communication (H=25.134, *P*=.003, ε^2=0.051) showed significant variation, with speech and language therapists using sensors more frequently than neurological or musculoskeletal therapists, and mental health therapists using face-to-face communication less frequently than several other specialties. Unstructured free-text EMR entry (H=18.112, *P*=.03, ε^2=0.029) and willingness to send imaging results (H=18.075, *P*=.03, ε^2=0.029) reached overall significance, but no pairwise comparisons remained significant after correction. No other outcomes differed significantly across groups (all *P*>.05).

Summary of Open-Ended Responses

Open-ended responses were reviewed to provide additional context to the quantitative findings. Commonly reported barriers to digital health use included limited interoperability between information systems, continued reliance on paper-based workflows, insufficient institutional support, and time constraints in routine clinical practice. Respondents also highlighted perceived benefits of digital technologies, particularly in improving efficiency, data continuity, and long-term patient monitoring. These responses did not introduce themes that contradicted the quantitative results.

XSL•FO
RenderX

Discussion

Principal Findings

This cross-sectional survey provides new evidence on how rehabilitation professionals in China use DHTs in routine practice. Given that implementation of DHIs represents a cultural transformation of traditional health care [43], understanding the confidence and attitudes of health care professionals regarding the use of DHTs is a growing priority [44]. In this sample, in-person communication (312/324, 96.3%) and paper-based documentation (266/324, 82.1%) were the dominant methods for obtaining clinical information. In contrast, the adoption of digital tools remained limited across all domains. For instance, the frequent use of email (22/324, 6.8%), patient-initiated app-based data sharing (32/324, 9.6%), and video consultations for objective assessment (27/324, 8.3%) each fell below 10%. Despite this limited uptake, respondents expressed strong willingness to incorporate digital technologies into their clinical work, with more than 70% reporting positive attitudes toward most digital functions. This high readiness stands in stark contrast to the low levels of actual use, revealing a substantial willingness-practice gap. For example, although 79% (256/324) of respondents were willing to digitally track patient rehabilitation progress, only 30.6% (99/324) frequently used electronic systems to collect subjective information. Similarly, while 55.2% (179/324) expressed willingness to conduct teleconsultations, the actual frequent use of video-based objective assessment was only 8.3% (27/324). These findings underscore a significant opportunity to advance the digital transformation of rehabilitation services.

Previous research in other regions shows both alignment and divergence with our findings. An Australian survey using the original version of this instrument [37] reported widespread use of EMR-based free-text documentation and strong willingness to adopt digital tools for tasks such as scheduling, imaging tracking, and teleconsultation, yet therapists still relied predominantly on face-to-face communication for subjective assessment and visual observation for objective evaluation. In comparison, Chinese therapists showed similar dependence on direct communication but demonstrated broader, more experimental uptake of digital methods, including photo, video, and sensor-based assessment, while maintaining heavier reliance on paper-based records. Although willingness to use digital technologies was high in both contexts, the emphasis differed: Australian clinicians prioritized tools that streamline communication and workflow, whereas Chinese therapists expressed greater interest in outcome monitoring, identity verification, and progress tracking rather than patient-facing teleconsultation. These differences likely reflect cultural preferences for relationship-centered interaction, earlier digitization of health records in Australia, and variability in hospital information systems across Chinese provinces [45].

Findings from other regions further contextualize these patterns. An Italian survey reported high availability of digital equipment but low levels of perceived integration and knowledge, alongside marked regional disparities in resources and experience [46]. Similarly, Norwegian physiotherapists identified telephone consultation as the most commonly used digital communication modality, with limited uptake of video consultation, gaming, or VR technologies [47]. Studies from the Middle East, North Africa, and other middle-income settings likewise describe early-stage digital adoption within rehabilitation systems, characterized by strong enthusiasm but persistent barriers related to software availability, interoperability, digital literacy, and organizational readiness [48-50]. Collectively, this body of evidence suggests that globally, clinician enthusiasm for digital rehabilitation often outpaces the degree of practical implementation [51-53]. Our study extends this literature by providing empirical data from a rapidly digitizing health system, highlighting both shared challenges and substantial untapped potential.

The substantial willingness-practice gap observed in our study is consistent with recent findings from other rehabilitation settings in China. A survey of orthopedic rehabilitation medical staff found that while support for digital therapeutics was high, only 21% had actually used them, with concerns about privacy, interpersonal relationships, and time constraints limiting adoption [54]. Similarly, a study of autism spectrum disorder digital health services across 2 Chinese provinces reported that despite high demand among rehabilitation therapists (90.0%), actual use remained critically low (21.6%) [55]. Together, these findings suggest that the gap between readiness and implementation is not unique to our sample but reflects a broader systemic challenge across diverse rehabilitation disciplines in China. Our results also suggest that the strong demand for digital health training (234/324, 72.3%) highlights a critical leverage point. This finding aligns with previous research indicating that the absence of systematic, formal education on digital technologies limits the readiness of current and future health professionals for digital transformation [56]. It suggests that rehabilitation professionals are not only ready to adopt digital tools but are also actively seeking opportunities to develop the skills required for their effective implementation. Bridging this educational gap could accelerate the translation of high willingness into tangible clinical practice by integrating digital health competencies into entry-level curricula and establishing continuing professional development programs.

This study also examined the influence of educational background, professional title, and clinical specialty on digital health use patterns. Subgroup analyses revealed that educational level was the most significant factor influencing the adoption of advanced technologies, with therapists holding master's or doctoral degrees reporting significantly higher use of sensor-based assessment, motion capture technology, and wearable devices compared to those with associate or bachelor's degrees. In contrast, the influence of professional title was limited; although significant differences emerged in a few indicators, most post-hoc comparisons were nonsignificant after adjustment, and professional title had no effect on the vast majority of digital health functions. Similarly, differences across clinical specialties were minimal. The absence of differences across most domains suggests that readiness for digital rehabilitation is broadly shared across specialties, rather than being driven by discipline-specific factors. This pattern points toward the influence of system-level conditions rather than

XSL•FO
RenderX

professional background alone in shaping digital health adoption in rehabilitation practice. Insights from open-ended responses further support the interpretation that barriers to digital rehabilitation are primarily structural and organizational rather than related to clinician willingness. The implications of these findings are considerable. Given China's large population, rising burden of disabling conditions, and uneven geographic distribution of rehabilitation resources, the high level of willingness among therapists represents a critical opportunity [57-59].

Scaling up digital health could help extend services to underserved regions, improve continuity of care, reduce travel burden for patients, and partially mitigate workforce shortages [60,61]. Evidence from digital health implementation research indicates that the development of supportive policy environments, including reimbursement arrangements, data governance structures, and standardized documentation frameworks, can enable more durable adoption of technology in rehabilitation services [62-64]. Research in digital health education also highlights the value of structured training, with multiple evaluations reporting that integration of telehealth and digital assessment skills into both preservice and continuing professional programs improves clinician confidence and routine application [65,66]. Evidence from programs in musculoskeletal rehabilitation further suggests that remote teleconsultations can have the same effect as in-person physio consultations and can maintain clinical outcomes while expanding reach to populations with limited physical access [24]. Meanwhile, secondary analyses in our study revealed a positive trend in willingness to engage in teleconsultation across increasing professional titles, with higher-ranking therapists reporting greater willingness. This finding underscores the need for tiered digital training strategies tailored to rehabilitation professionals at different career stages, with particular attention to supporting lower- and mid-level practitioners. In addition, studies from low-resource regions show that software availability, device access, and internet reliability strongly influence digital function, with rural and underresourced settings experiencing the greatest barriers [67,68].

Taken together, this study demonstrates that limited use of DHTs in rehabilitation practice should not be interpreted as a lack of clinician readiness. Instead, our findings indicate a widespread willingness to engage with digital rehabilitation across specialties and professional levels, alongside persistent structural and organizational constraints that limit implementation. By providing evidence from China, this study contributes to a growing body of international research showing that the digital transformation of rehabilitation is primarily shaped by system-level conditions rather than individual clinician attitudes. These insights underscore the need for implementation strategies that move beyond profession-specific solutions and address the broader institutional, infrastructural, and policy environments in which rehabilitation services are delivered.

Limitations

This study has several limitations. This study has several limitations. Due to open-link recruitment, the response rate could not be calculated, precluding assessment of nonresponse bias. The online survey mode inherently introduced selection bias by systematically excluding professionals with limited internet access or digital skills, and participation itself implied a certain level of comfort with digital technologies. This self-selection mechanism likely overestimates the overall population's willingness and actual use levels, particularly for those in remote areas or primary care settings. Geographic distribution bias further limits generalizability, as respondents were predominantly from several provinces, with underrepresentation from central and western regions and primary care institutions. Additionally, all data were based on self-reports, which may be subject to social desirability and recall biases. The cross-sectional design captures technology use at a single time point and cannot reveal dynamic changes in adoption or causal pathways, underscoring the need for longitudinal studies.

Conclusions

In conclusion, clinical practice in China continues to rely primarily on traditional in-person and paper-based methods, yet clinicians express clear readiness to adopt digital health. This readiness indicates significant potential for targeted policy actions, education programs, and implementation planning. Addressing practical and structural barriers could accelerate digital transformation in rehabilitation and support improved access and more sustainable delivery of rehabilitation services.

Acknowledgments

The authors are grateful to those who participated in this study. We sincerely thank Professor Mark Merolli from the University of Melbourne for his valuable suggestions and feedback during the preparation of this study. All authors contributed to drafting the manuscript or revising it critically for important intellectual content and approved the final version for submission. The authors declare that no generative artificial intelligence was used in the creation of this manuscript.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Funding

This work was supported by the National Social Science Fund of China (Grant No. 23BTY117), which funded the corresponding author GN.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: SD (equal), ZRW (equal), GN (equal)
Data curation: SD (equal), ZRW (equal), XC (equal)
Formal analysis: SD (equal), ZRW (equal)
Funding acquisition: GN
Investigation: SD (equal), XC (equal)
Methodology: SD (equal), ZRW (equal)
Project administration: SD (lead), ZRW (equal), XC (supporting)
Resources: GN, ZRW
Supervision: GN
Visualization: SD (supporting), ZRW (lead)
Writing – original draft: SD (lead), ZRW (supporting)
Writing – review and editing: SD (lead), ZRW (supporting), GN (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
STROBE checklist.
[DOCX File , 34 KB - jmir_v28i1e90623_app1.docx]

Multimedia Appendix 2
Survey instrument for rehabilitation therapists on digital health applications (Chinese version).
[DOCX File , 35 KB - jmir_v28i1e90623_app2.docx]

Multimedia Appendix 3
Survey instrument for rehabilitation therapists on digital health applications (Chinese back-translation to English for review).
[DOCX File , 32 KB - jmir_v28i1e90623_app3.docx]

Multimedia Appendix 4
Subgroup comparisons of digital health technology use by educational level, professional title, and clinical specialty.
[DOCX File , 19 KB - jmir_v28i1e90623_app4.docx]

References

1. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the global burden of disease study 2019: a systematic analysis for the global burden of disease study 2019. Lancet 2021;396(10267):2006-2017 [FREE Full text] [doi: 10.1016/S0140-6736(20)32340-0] [Medline: 33275908]
2. Jette AM. Global prevalence of disability and need for rehabilitation. Phys Ther 2021;101(2):pzab004. [doi: 10.1093/ptj/pzab004] [Medline: 33556184]
3. To Dutka J, Gans BM, Bracciano A, Bharadwaj S, Akinwuntan A, Mauk K, et al. Delivering rehabilitation care around the world: voices from the field. Arch Phys Med Rehabil 2023;104(9):1385-1393 [FREE Full text] [doi: 10.1016/j.apmr.2023.03.009] [Medline: 36997079]
4. Taylor R, Zwisler AD, Uddin J. Global health-care systems must prioritise rehabilitation. Lancet 2021;396(10267):1946-1947 [FREE Full text] [doi: 10.1016/S0140-6736(20)32533-2] [Medline: 33275909]
5. Krahn GL. WHO world report on disability: a review. Disabil Health J 2011;4(3):141-142. [doi: 10.1016/j.dhjo.2011.05.001] [Medline: 21723520]
6. Rehabilitation in health systems. World Health Organization. 2017. URL: https://www.who.int/publications/i/item/9789241549974 [accessed 2023-05-03]
7. Rehabilitation competency framework. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240008281 [accessed 2026-03-06]
8. Rehabilitation: key for health in the 21st century. World Health Organization. 2017. URL: https://www.who.int/publications/i/item/9789241515115 [accessed 2026-03-06]
9. Basic standards and management guidelines for rehabilitation medical centers (Trial) (Guoweiyifa [2017] No. 51). National Health and Family Planning Commission. 2017. URL: http://www.nhfpc.gov.cn/yzygj/s3577/201711/fac102fd386a41f1ab545315d7c26045.shtml [accessed 2026-03-06]

XSL•FO
RenderX

10. Notice on issuing opinions on accelerating the development of rehabilitation medicine [Web Page in Chinese]. National Health Commission of the People's Republic of China. 2021. URL: https://www.gov.cn/zhengce/zhengceku/2021-06/17/content_5618767.htm [accessed 2026-03-06]
11. Rehabilitation. World Health Organization. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/rehabilitation [accessed 2026-03-20]
12. Htwe O, Yuliawiratman BS, Tannor AY, Nor Asikin MZ, Soh E, DE Groote W, et al. Barriers and facilitators for increased accessibility to quality rehabilitation services in low- and middle- income countries: a systematic review. Eur J Phys Rehabil Med 2024;60(3):514-522 [FREE Full text] [doi: 10.23736/S1973-9087.24.08154-1] [Medline: 38551518]
13. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the global burden of disease study 2019: a systematic analysis for the global burden of disease study 2019. Lancet 2021;396(10267):2006-2017 [FREE Full text] [doi: 10.1016/S0140-6736(20)32340-0] [Medline: 33275908]
14. Chinese Physical Therapy Association. World Physiotherapy. 2023. URL: https://world.physio/membership/china [accessed 2026-03-06]
15. Yisma E, Versace VL, Jones M, Walsh S, Jones S, May E, et al. The distribution of registered occupational therapists, physiotherapists, and podiatrists in Australia. PLoS One 2023;18(9):e0291962 [FREE Full text] [doi: 10.1371/journal.pone.0291962] [Medline: 37733814]
16. Australian Health Practitioner Regulation Agency. Physiotherapists - a snapshot as at 30 June 2023. Ahpra & National Boards. Melbourne, Australia: Australian Health Practitioner Regulation Agency; 2023. URL: https://www.ahpra.gov.au/Publications/Registrant-data/Physiotherapists.aspx [accessed 2026-03-06]
17. Transcript of the press conference on October 28, 2020. National Health Commission of the PRC. 2020. URL: https://www.nhc.gov.cn/xcs/c100122/202010/96cb6d028e5f4c9380059bc7ac83bb21.shtml [accessed 2020-08-01]
18. Jin Y, Hu J. The current status, challenges and future of medical and health data sharing in China: lessons from the European health data space. Eur J Health Law 2025;32(3):239-280. [doi: 10.1163/15718093-bja10147] [Medline: 40541395]
19. Li X, Cong Y. A systematic literature review of ethical challenges related to medical and public health data sharing in China. J Empir Res Hum Res Ethics 2021;16(5):537-554. [doi: 10.1177/15562646211040299] [Medline: 34516325]
20. Li Q, Cheng F, Zeng H, Xu J. Health insurance payment for telehealth services: scoping review and narrative synthesis. J Med Internet Res 2024;26:e56699 [FREE Full text] [doi: 10.2196/56699] [Medline: 39652868]
21. Wu G, Gong M, Wu Y, Liu L, Shi B, Zeng Z. Advancing digital health in China: aligning challenges, opportunities, and solutions with the global initiative on digital health (GIDH). Health Care Sci 2024;3(5):365-369. [doi: 10.1002/hcs2.118] [Medline: 39479272]
22. World Health Organization. Global Strategy on Digital Health 2020-2025. Geneva: World Health Organization; 2021.
23. Kizilkilic SE, Ramakers W, Falter M, Scherrenberg M, Bonneux C, Pieters Z, et al. A digitally-supported shared decision making approach for patients during cardiac rehabilitation: a randomized controlled trial. Eur J Prev Cardiol 2025:zwaf537. [doi: 10.1093/eurjpc/zwaf537] [Medline: 40874533]
24. Hinman RS, Campbell PK, Kimp AJ, Russell T, Foster NE, Kasza J, et al. Telerehabilitation consultations with a physiotherapist for chronic knee pain versus in-person consultations in Australia: the PEAK non-inferiority randomised controlled trial. Lancet 2024;403(10433):1267-1278. [doi: 10.1016/S0140-6736(23)02630-2] [Medline: 38461844]
25. Grochtdreis T, König HH, Renner J, Sehner S, Dekker A, Briken P, et al. Cost-effectiveness of an interdisciplinary, internet-based transgender health care program in Germany: economic evaluation alongside a randomized controlled trial. J Med Internet Res 2025;27:e66371 [FREE Full text] [doi: 10.2196/66371] [Medline: 40537093]
26. Lee YK, Yoon EJ, Kim TH, Kim JI, Kim JH. Musculoskeletal digital therapeutics and digital health rehabilitation: a global paradigm shift in orthopedic care. J Clin Med 2025;14(23):8467 [FREE Full text] [doi: 10.3390/jcm14238467] [Medline: 41375770]
27. Al-Jabr H, Salt E, Stephenson J, Hamdan E, Helliwell T. Can remote digital health interventions improve rehabilitation following major trauma? Results of a systematic review. Arch Phys Med Rehabil 2025;S0003-9993(25):01002 [FREE Full text] [doi: 10.1016/j.apmr.2025.10.012] [Medline: 41187799]
28. Gu B, Kim HS, Kim H, Yoo J. Advancements in wearable sensor technologies for health monitoring in terms of clinical applications, rehabilitation, and disease risk assessment: systematic review. JMIR Mhealth Uhealth 2026;14:e76084 [FREE Full text] [doi: 10.2196/76084] [Medline: 41511829]
29. Wang Q, Markopoulos P, Yu B, Chen W, Timmermans A. Interactive wearable systems for upper body rehabilitation: a systematic review. J Neuroeng Rehabil 2017;14(1):20 [FREE Full text] [doi: 10.1186/s12984-017-0229-y] [Medline: 28284228]
30. Gooch HJ, Jarvis KA, Stockley RC. Behavior change approaches in digital technology-based physical rehabilitation interventions following stroke: scoping review. J Med Internet Res 2024;26:e48725 [FREE Full text] [doi: 10.2196/48725] [Medline: 38656777]
31. Jeilani A, Hussein A. Impact of digital health technologies adoption on healthcare workers' performance and workload: perspective with DOI and TOE models. BMC Health Serv Res 2025;25(1):271 [FREE Full text] [doi: 10.1186/s12913-025-12414-4] [Medline: 39966874]

XSL•FO
RenderX

32. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, Martinez EZ, Zucoloto ML, Østengaard L, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023 Sep 18;6(1):161. [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
33. Weik L, Fehring L, Mortsiefer A, Meister S. Understanding inherent influencing factors to digital health adoption in general practices through a mixed-methods analysis. NPJ Digit Med 2024 Feb 27;7(1):47. [doi: 10.1038/s41746-024-01049-0] [Medline: 38413767]
34. Sylla B, Ismaila O, Diallo G. 25 years of digital health toward universal health coverage in low- and middle-income countries: rapid systematic review. J Med Internet Res 2025;27:e59042 [FREE Full text] [doi: 10.2196/59042] [Medline: 40440696]
35. Portz J, Moore S, Bull S. Evolutionary trends in the adoption, adaptation, and abandonment of mobile health technologies: viewpoint based on 25 years of research. J Med Internet Res 2024;26:e62790 [FREE Full text] [doi: 10.2196/62790] [Medline: 39331463]
36. Cuschieri S. The STROBE guidelines. Saudi J Anaesth 2019;13(Suppl 1):S31-S34 [FREE Full text] [doi: 10.4103/sja.SJA_543_18] [Medline: 30930717]
37. Merolli M, Gray K, Choo D, Lawford BJ, Hinman RS. Use, and acceptability, of digital health technologies in musculoskeletal physical therapy: a survey of physical therapists and patients. Musculoskeletal Care 2022;20(3):641-659 [FREE Full text] [doi: 10.1002/msc.1627] [Medline: 35278266]
38. Classification of digital health interventions v1.0: a shared language to describe the uses of digital technology for health. World Health Organization. 2018. URL: https://www.who.int/publications/i/item/WHO-RHR-18.06 [accessed 2026-03-20]
39. Description of physical therapy: policy statement. World Physiotherapy. 2019. URL: https://world.physio/sites/default/files/2020-04/PS-2019-Description-of-physical-therapy.pdf [accessed 2026-03-20]
40. Definitions of occupational therapy from member organisations. World Federation of Occupational Therapists. 2012. URL: https://wfot.org/resources/definitions-of-occupational-therapy-from-member-organisations [accessed 2026-03-20]
41. Assessment guideline for international classification of functioning, disability and health-rehabilitation set (ICF-RS) (T/CARM 001?2020). Chinese Association of Rehabilitation Medicine. 2020. URL: https://www.ttbz.org.cn/StandardManage/Detail/41167/ [accessed 2026-03-07]
42. Li J, Chen L. Guidelines for the standardization of rehabilitation medicine association standards (2025 Edition). Chin J Rehab Med 2025;40(1):1. [doi: 10.3870/j.issn.1001-1242.2025.01.001]
43. Meskó B, Drobni Z, Bényei É, Gergely B, Győrffy Z. Digital health is a cultural transformation of traditional healthcare. Mhealth 2017;3:38 [FREE Full text] [doi: 10.21037/mhealth.2017.08.07] [Medline: 29184890]
44. Kuek A, Hakkennes S. Healthcare staff digital literacy levels and their attitudes towards information systems. Health Informatics J 2020;26(1):592-612 [FREE Full text] [doi: 10.1177/1460458219839613] [Medline: 30983476]
45. National digital health strategy advancing care for all Australians. Australian Digital Health Agency. 2025. URL: https://www.digitalhealth.gov.au/newsroom/media/national-digital-health-strategy-advancing-care-for-all-australians [accessed 2026-03-05]
46. Capecci M, Gandolfi M, Straudi S, Calabrò RS, Baldini N, Pepa L, et al. Shaping the future: an Italian survey unveils the unmet need to empower physical medicine and rehabilitation professionals with technological skills. Eur J Phys Rehabil Med 2024;60(3):540-543 [FREE Full text] [doi: 10.23736/S1973-9087.24.08376-X] [Medline: 38618695]
47. Martinsen L, Østerås N, Moseng T, Tveter AT. Usage, attitudes, facilitators, and barriers toward digital health technologies in musculoskeletal care: survey among primary care physiotherapists in Norway. JMIR Rehabil Assist Technol 2024;11:e54116 [FREE Full text] [doi: 10.2196/54116] [Medline: 39283661]
48. El-Jardali F, Bou-Karroum L, Jabbour M, Bou-Karroum K, Aoun A, Salameh S, et al. Digital health in fragile states in the middle east and north Africa (MENA) region: a scoping review of the literature. PLoS One 2023;18(4):e0285226 [FREE Full text] [doi: 10.1371/journal.pone.0285226] [Medline: 37115778]
49. Qasrawi R, Al Sabbah H, Issa G, Thwib S, Amro M, Atari S, et al. Assessing the impact of digital health literacy on health management practices in Arab middle eastern and north African countries: insights from predictive modeling. Front Digit Health 2025;7:1555436 [FREE Full text] [doi: 10.3389/fdgth.2025.1555436] [Medline: 41112210]
50. Aljabri NQ, Bulkeley K, Cusick A. Telerehabilitation in the middle east North Africa Region: A structured review. Int J Telerehabil 2021;13(2):e6401 [FREE Full text] [doi: 10.5195/ijt.2021.6401] [Medline: 35646235]
51. Yew SQ, Trivedi D, Adanan NIH, Chew BH. Facilitators and barriers to the implementation of digital health technologies in hospital settings in lower- and middle-income countries since the onset of the COVID-19 pandemic: scoping review. J Med Internet Res 2025;27:e63482 [FREE Full text] [doi: 10.2196/63482] [Medline: 40053793]
52. Nicora G, Parimbelli E, Mauro MC, Falchini F, Germanotta M, Fasano A, et al. Healthcare practitioners and robotic-assisted rehabilitation: understanding needs and barriers. J Neuroeng Rehabil 2025;22(1):78 [FREE Full text] [doi: 10.1186/s12984-025-01593-0] [Medline: 40205389]
53. Scott Kruse C, Karem P, Shifflett K, Vegi L, Ravi K, Brooks M. Evaluating barriers to adopting telemedicine worldwide: a systematic review. J Telemed Telecare 2018;24(1):4-12 [FREE Full text] [doi: 10.1177/1357633X16674087] [Medline: 29320966]

XSL•FO
RenderX

54. Liu FT, Chong T, Zhang K, Shi X, Mu J, Wang G, et al. Awareness of and willingness to use digital therapeutics among orthopedic rehabilitation medical staff: a questionnaire survey. Chinese Journal of Clinicians (Electronic Edition) 2024;18(12):1125-1131. [doi: 10.3877/cma.j.issn.1674-0785.2024.12.008]
55. Zou M, Gong X, Feng L, Li S, Lu C, Liu Z, et al. Regional disparities in the use and demand for digital health services for autism spectrum disorder in China: cross-sectional survey of stakeholder perspectives. J Med Internet Res 2025;27:e77157 [FREE Full text] [doi: 10.2196/77157] [Medline: 41043146]
56. Zhang Y, Liu J. Does education affect economic growth? A re-examination of empirical data from China. Sustainability 2022 Dec 06;14(23):16289. [doi: 10.3390/su142316289]
57. Chen X, Giles J, Yao Y, Yip W, Meng Q, Berkman L, et al. The path to healthy ageing in China: a peking university-lancet commission. Lancet 2022;400(10367):1967-2006 [FREE Full text] [doi: 10.1016/S0140-6736(22)01546-X] [Medline: 36423650]
58. Capecci M, Gandolfi M, Straudi S, Calabrò RS, Baldini N, Pepa L, et al. Advancing public health through technological rehabilitation: insights from a national clinician survey. BMC Health Serv Res 2024;24(1):1626 [FREE Full text] [doi: 10.1186/s12913-024-11991-0] [Medline: 39702315]
59. Jones SE, Campbell PK, Kimp AJ, Bennell K, Foster NE, Russell T, et al. Evaluation of a novel e-learning program for physiotherapists to manage knee osteoarthritis via telehealth: qualitative study nested in the PEAK (physiotherapy exercise and physical activity for knee osteoarthritis) randomized controlled trial. J Med Internet Res 2021;23(4):e25872 [FREE Full text] [doi: 10.2196/25872] [Medline: 33929326]
60. Walle AD, Butta FW, Kassie SY, Chereka AA, Kanfe SG, Dubale AT, et al. Healthcare professionals' attitude to using mobile health technology and its associated factors in a resource-limited country-an implication for digital health implementers: a cross sectional study. Biomed Res Int 2024;2024:1631376 [FREE Full text] [doi: 10.1155/2024/1631376] [Medline: 39035773]
61. Wosny M, Strasser LM, Hastings J. Experience of health care professionals using digital tools in the hospital: qualitative systematic review. JMIR Hum Factors 2023;10:e50357 [FREE Full text] [doi: 10.2196/50357] [Medline: 37847535]
62. Gordon WJ, Landman A, Zhang H, Bates DW. Beyond validation: getting health apps into clinical practice. NPJ Digit Med 2020;3:14. [doi: 10.1038/s41746-019-0212-z] [Medline: 32047860]
63. Elser A, Lange M, Kopkow C, Schäfer AG. Barriers and facilitators to the implementation of virtual reality interventions for people with chronic pain: scoping review. JMIR XR Spat Comput 2024;1:e53129. [doi: 10.2196/53129] [Medline: 41341873]
64. Krahe MA, Larkins SL, Adams N. Digital health implementation in Australia: a scientometric review of the research. Digit Health 2024;10:20552076241297729 [FREE Full text] [doi: 10.1177/20552076241297729] [Medline: 39539722]
65. Rettinger L, Maul L, Putz P, Ertelt-Bach V, Huber A, Javorszky SM, et al. Telehealth acceptance and perceived barriers among health professionals: pre-post evaluation of a web-based telehealth course. JMIR Hum Factors 2025;12:e74107 [FREE Full text] [doi: 10.2196/74107] [Medline: 40902201]
66. Davies L, Hinman RS, Russell T, Lawford B, Bennell K. An international core capability framework for physiotherapists delivering telephone-based care. J Physiother 2022;68(2):136-141 [FREE Full text] [doi: 10.1016/j.jphys.2022.02.002] [Medline: 35283052]
67. Dukuzimana MJ, Ukwizabigira JB, Muhire C, Jokinen K, Tumusiime DK, Murtonen KP, et al. Rehabilitation needs at rural primary health care settings: perspectives of health center nurses in burera district of Rwanda. BMC Prim Care 2025;26(1):224. [doi: 10.1186/s12875-025-02921-y] [Medline: 40660097]
68. Motiwala F, Ezezika O. Barriers to scaling health technologies in sub-Saharan Africa: Lessons from Ethiopia, Nigeria, and Rwanda. Afr J Sci Technol Innovation Devel 2021 Nov 28;14(7):1788-1797. [doi: 10.1080/20421338.2021.1985203]

Abbreviations

DHI: digital health interventions
DHT: digital health technology
NHC: National Health Commission
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
VR: virtual reality
WHO: World Health Organization

XSL•FO
RenderX

Edited by M Balcarras; submitted 31.Dec.2025; peer-reviewed by W Qin, Z Liu, O Ndibe, Y Hu; comments to author 12.Feb.2026; revised version received 10.Mar.2026; accepted 17.Mar.2026; published 09.Apr.2026.

Please cite as:
Duan S, Wang ZR, Chen X, Ni G
Digital Health Technology Use Among Rehabilitation Professionals in China: Multi-Province Cross-Sectional Survey
J Med Internet Res 2026;28:e90623
URL: https://www.jmir.org/2026/1/e90623
doi:10.2196/90623
PMID:

©Shuning Duan, Zi-Ru Wang, Xier Chen, Guoxin Ni. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Associations Between Short-Video Platform Use and Health Across Health Distribution and Usage Behaviors in China: Cross-Sectional Questionnaire Study

Yangyang Pan[1], MSc; Kangkang Zhang[1], MSc, PhD; Yilin Wei[1], MA, PhD; Yangzhen Huang[1], MSc; Chengxu Long[1,2], MSc, PhD; Chenxin Yang[1], MSc, PhD; Shangfeng Tang[1,2], MSc, PhD

[1]School of Medicine and Health Management, Huazhong University of Science and Technology, No. 13, Aviation Road, Qiaokou District, Wuhan, China

[2]Research Center for Rural Health Service, Key Research Institute of Humanities & Social Sciences of Hubei Provincial Department of Education, Wuhan, China

Corresponding Author:
Shangfeng Tang, MSc, PhD
School of Medicine and Health Management, Huazhong University of Science and Technology, No. 13, Aviation Road, Qiaokou District, Wuhan, China

Abstract

Background: Short-video platforms, characterized by algorithmic curation and passive consumption, have emerged as dominant components of digital life. However, the associations between short-video platform use and health across different groups and usage behaviors remain understudied.

Objective: This study investigates associations between short-video platform use and health, examining whether these relationships vary across health status, usage behaviors, and socioeconomic status.

Methods: A cross-sectional study was conducted using multistage stratified sampling across eastern, central, and western China from July to September 2024. The inclusion criteria were age 18 years or older, ability to communicate effectively, and no cognitive disorders or mental disturbance. Of 7725 participants enrolled, 46.96% (n=3628) were male, and the average age was 65.49 (SD 8.39) years. The data were collected via face-to-face interviews using a structured questionnaire. Self-rated health and relative health deprivation (Kakwani index) were used to measure health. Quantile regression explored associations between whether using short-video platform and health varies across the health distribution, while linear regression examined associations of years, frequency, daily duration, and purpose diversity of short-video platform use with health. Moderating effect analysis explored the role of socioeconomic status in the relationship between the daily duration of use and health.

Results: Coefficients were tested using 2-tailed *t* tests, and statistical significance was defined as a 2-sided *P* value less than .05. Quantile regression revealed heterogeneous associations. Compared to nonusers, short-video platform users had better self-rated health at the 70th to 90th quantiles and lower relative health deprivation at the 10th to 30th quantiles. However, the users at the 10th quantile of self-rated health had worse self-rated health (β=−2.224, 95% CI −3.835 to −0.613). Longer engagement (≥3 y) correlated with lower relative health deprivation (β=1.970, 95% CI 0.308-3.632), while daily use of 1 - 4 hours was associated with poorer self-rated health (β=−3.385, 95% CI −4.872 to −1.898; β=−3.038, 95% CI −5.054 to −1.022) and higher relative health deprivation (β=0.035, 95% CI 0.021-0.050; *P*<.001; β=0.034, 95% CI 0.014-0.054). Compared to no purposeful use, using with 2 purposes was associated with better self-rated health (β=6.082, 95% CI 0.250-11.914) and lower relative health deprivation (β=−0.063, 95% CI −0.120 to −0.005). The association was stronger for use with 3 or more purposes. Socioeconomic status moderated the relationship between daily duration of use and health.

Conclusions: This study provides a more specific investigation of how these associations vary across health strata and usage patterns. The findings reveal patterns of benefit and risk across population subgroups, underscoring that how and why individuals engage with platforms matter more than mere access or frequency. These insights necessitate targeted digital well-being policies that protect vulnerable groups, particularly those in poor health or with lower socioeconomic status. Furthermore, policies should actively encourage intentional, functionally grounded use to reduce health inequities and advance equitable digital inclusion.

(J Med Internet Res 2026;28:e86526) doi:10.2196/86526

KEYWORDS

short-video platform; self-rated health; quantile regression; health disparities; socioeconomic status

XSL•FO
RenderX

Introduction

The rapid proliferation of digital technologies has fundamentally reconfigured human behavior and social dynamics, establishing digital engagement as a pivotal social determinant of health [1]. Emerging evidence suggests that digital engagement patterns exhibit complex relationships with physical and mental well-being [2-6]. Some studies indicate that digital engagement promotes social support and information access [7,8], while others suggest that it correlates with anxiety, depression, and sedentary behavior [9,10]. This paradoxical finding underscores the necessity to consider technological characteristics and user experience variability when assessing the relationship between digital engagement and health.

Among digital platforms, short-video platforms, such as TikTok, Douyin, and Reels, represent a unique research frontier given their algorithmic-driven content delivery, passive consumption modalities, and entertainment-centric design [11]. In China, over 1 billion users engage with short-video content, accounting for 95.2% of internet users as of June 2023 [12]. However, high-frequency passive browsing and performative interactions features of short-video platforms can amplify the potential benefits of digital engagement, as well as exacerbate their risks. These platforms mediate health-related behaviors through 3 intersecting pathways: information, behavior, and psychology. Information pathways primarily shape individuals' health knowledge and literacy. Studies have shown that internet use for health information was associated with higher health literacy and health service use [13]. But perceived health information overload discourages enhancing health literacy [14]. Behavioral pathways encompass habitual usage behaviors and health-promoting behaviors. Access to health information facilitates the development of health behaviors [15,16], while sedentary behavior and addictive behavior during internet use may adversely affect health [17,18]. Psychological pathways involve platform-based social interactions and psychological symptoms. On the one hand, the internet can expand social networks, providing social support and entertainment [19,20]. On the other hand, problematic internet use may lead to symptoms of anxiety and depression [21]. Crucially, the relationship between these pathways and health is unlikely to be uniform across all individuals but rather varies according to their existing resources and health status.

However, existing studies predominantly focus on aggregate associations between digital engagement and health outcomes. The heterogeneous association between short-video platform use and health regarding different population subgroups remains understudied. One key issue is how its multidimensional characteristics, including the time, frequency, duration, and purpose of short-video platform use, are related to individual health. Moreover, engagement with short-video platforms has been an emerging behavioral determinant nested within broader socioeconomic and environmental structures. The social determinants of health theory holds that health and health inequalities are determined by social, economic, and environmental factors [22]. The digital health paradox further sharpens this lens, highlighting that people excluded by digital technology face the risk of being abandoned by society in most cases. So, people who have very limited access to the internet are precisely those who need the most health services and bear the heaviest health burden, such as low-income populations and long-term ill individuals [23]. Therefore, it is necessary to consider whether the association between short-video platform use and health exhibits heterogeneity across different health distributions and socioeconomic status (SES) groups.

To comprehensively understand the relationship between short-video platform use and individual health, the measurement of health also requires a multidimensional perspective. Self-rated health has been validated to comprehensively reflect overall health status and effectively predict mortality [24], yet it fails to capture the distribution and accumulation of health conditions. To address this limitation, we incorporate relative health deprivation, operationalized through the Kakwani index [25]. Relative deprivation theory holds that an individual's health level is negatively correlated with health disadvantages within a specific group, which means the lower the health level, the greater the health disadvantage and the higher the degree of relative deprivation of health faced. By calculating the health deprivation index, it is possible to reflect the long-term and structural accumulation of health disadvantages.

This study aims to investigate the heterogeneous associations between short-video platform use and individual health across different health status and usage behaviors. Specifically, we aim to (1) examine the association between whether using short-video platform and health varies across the health distribution, (2) analyze the associations between multidimensional usage behaviors (years, frequency, daily duration, and purpose diversity) and health, and (3) explore the moderating role of SES in the relationship between daily duration of use and health. By identifying vulnerable subgroups involved in digital participation, this study provides critical, actionable insights for designing targeted digital health interventions and equitable platform governance policies.

Methods

Inclusion and Exclusion

Individuals who met the following criteria were included in this survey: (1) 18 years of age or older and (2) were able to communicate effectively with investigators. The exclusion criteria included (1) no cognitive disorders and (2) no mental disturbance. The data were collected using a structured questionnaire. Questionnaires with completion times below 5 minutes or above 50 minutes were excluded, as pilot testing indicated these thresholds captured inattentive or interrupted responses. Questionnaires with a response missing rate exceeding 50% for the purpose of quality control were also excluded.

Participant Characteristics

Of the 7725 participants included in the final analysis, 3628 (46.96%) were male, and the average age was 65.49 (SD 9.39) years. A total of 4403 (57.0%) participants were users of short-video platforms. Further detailed characteristics of the participants, including demographic, socioeconomic,

XSL•FO
RenderX

health-related, and social support factors, are presented in Table 1.

XSL•FO
RenderX

Table . Characteristics of the participants before propensity matching.

Characteristics	All	Short-video platforms		Self-rated health		Relative health deprivation	
		Users	Nonusers	Mean (SD)	*P* value	Mean (SD)	*P* value
N	7725 (100)	4403 (57.00)	3322 (43.00)	—[a]	—	—	—
Age (y), mean (SD)	65.49 (9.39)	61.92 (8.68)	70.21 (8.12)	65.59 (17.17)	<.001	0.14 (0.17)	<.001
Sex, n (%)					.007		.09
Male	3628 (46.96)	2270 (51.56)	1358 (40.88)	66.15 (17.40)		0.14 (0.17)	
Female	4097 (53.04)	2133 (48.44)	1964 (59.12)	65.09 (16.96)		0.15 (0.17)	
Marriage, n (%)					<.001		<.001
Married or cohabiting	6492 (84.04)	3966 (90.07)	2526 (76.04)	66.17 (16.98)		0.14 (0.16)	
Other marital status[b]	1233 (15.96)	437 (9.93)	796 (23.96)	62.52 (17.85)		0.17 (0.18)	
Urban-rural residence, n (%)					<.001		<.001
Urban	2342 (30.32)	1332 (30.25)	1010 (30.40)	67.93 (16.02)		0.12 (0.15)	
Rural	5383 (69.68)	3071 (69.75)	2312 (69.60)	65.59 (17.56)		0.15 (0.17)	
Socioeconomic status, n (%)					<.001		<.001
Low	2880 (37.28)	1622 (36.84)	1258 (37.87)	61.07 (18.19)		0.19 (0.19)	
Middle	4310 (55.79)	2416 (54.87)	1894 (57.01)	68.00 (15.86)		0.12 (0.14)	
High	535 (6.93)	365 (8.29)	170 (5.12)	70.39 (16.49)		0.11 (0.15)	
Education, n (%)					<.001		<.001
Illiterate or semiliterate	2960 (38.32)	1104 (25.07)	1856 (55.87)	63.45 (16.86)		0.16 (0.17)	
Primary school	2277 (29.48)	1336 (30.34)	941 (28.33)	65.42 (17.56)		0.15 (0.17)	
Middle school	1619 (20.96)	1281 (29.09)	338 (10.17)	67.46 (16.96)		0.13 (0.16)	
High school or vocational school	686 (8.88)	535 (12.15)	151 (4.55)	69.40 (16.72)		0.11 (0.16)	
College degree / Bachelor's degree or above	183 (2.37)	147 (3.34)	36 (1.08)	71.30 (15.26)		0.10 (0.14)	
Employment status, n (%)					<.001		<.001
Unemployed	4539 (58.76)	2337 (53.08)	2202 (66.29)	64.22 (17.51)		0.16 (0.18)	
Employed	3186 (41.24)	2066 (46.92)	1120 (33.71)	67.53 (16.49)		0.13 (0.15)	
Convenience of obtaining medical services, n (%)					<.001		<.001
Low	22 (0.29)	15 (0.34)	7 (0.21)	55.18 (27.16)		0.28 (0.32)	
Medium low	284 (3.68)	104 (2.36)	180 (5.42)	55.10 (19.23)		0.25 (0.21)	
Middle	900 (11.67)	378 (8.59)	522 (15.71)	59.49 (18.08)		0.20 (0.19)	
Medium high	4394 (57.00)	2448 (55.60)	1946 (58.58)	66.55 (16.46)		0.13 (0.16)	
High	2109 (27.36)	1458 (33.11)	667 (20.08)	67.74 (16.72)		0.13 (0.16)	
Number of persons living together, n (%)					.008		<.001
0	740 (9.58)	319 (7.25)	421 (12.67)	63.92 (16.88)		0.16 (0.17)	
1-3	3961 (51.28)	2331 (52.94)	1630 (49.07)	65.22 (17.48)		0.15 (0.17)	
4-6	2657 (34.39)	1543 (35.04)	1114 (33.53)	66.36 (16.96)		0.14 (0.16)	
>6	367 (4.75)	210 (4.77)	157 (4.73)	67.34 (15.42)		0.12 (0.14)	

XSL•FO
RenderX

Characteristics	All	Short-video platforms		Self-rated health		Relative health deprivation	
		Users	Nonusers	Mean (SD)	*P* value	Mean (SD)	*P* value
Number of persons contacting with regularly, n (%)					<.001		<.001
0	307 (3.97)	151 (3.43)	156 (4.70)	58.57 (19.56)		0.22 (0.21)	
1	1348 (17.45)	583 (13.24)	765 (23.03)	64.94 (17.34)		0.15 (0.18)	
2-3	2855 (3.96)	1600 (36.34)	1255 (37.78)	65.76 (1660)		0.14 (0.16)	
4-5	1567 (20.28)	915 (20.78)	652 (19.63)	66.44 (15.91)		0.13 (0.15)	
≥6	1648 (21.33)	1154 (26.21)	494 (14.87)	66.31 (18.36)		0.14 (0.18)	
Multimorbidity-weighted index, mean (SD)	2.51 (2.02)	2.40 (1.96)	2.66 (2.09)	65.59 (17.17)	<.001	0.14 (0.17)	<.001
Depression, mean (SD)	5.87 (4.78)	5.28 (4.50)	6.65 (5.02)	65.59 (17.17)	<.001	0.14 (0.17)	<.001
Anxiety, mean (SD)	2.07 (3.45)	1.82 (3.22)	2.39 (3.71)	65.59 (17.17)	<.001	0.14 (0.17)	<.001
Loneliness, mean (SD)	3.70 (1.29)	3.54 (1.19)	3.90 (1.39)	65.59 (17.17)	<.001	0.14 (0.17)	<.001

[a]Not applicable.

[b]Other marital status includes being widowed, separated, divorced, or never married.

Sampling Procedures

Multistage stratified sampling was applied to select participants in the survey from July to September 2024. First, initial stratification was conducted based on geographical region. Fujian province, Hubei province, and Chongqing municipality were chosen as the survey areas, which are respectively located in the eastern, central, and western regions of China. Second, within each province, 6‐12 urban subdistricts and rural townships were randomly selected as secondary sampling units, ultimately covering 24 subdistricts or townships. Finally, through the community registries, 300‐400 participants were randomly selected from each subdistrict or township. Sampling weights were not applied in the analyses, and all estimates are presented as unweighted statistics. The sampling design aimed to ensure structural representativeness across regional and urban-rural strata within the selected provinces. The data were collected through face-to-face interviews with the respondents by uniformly trained investigators, and the questionnaires were filled out simultaneously. This study was conducted and reported following the Journal Article Reporting Standards for quantitative research [26].

Sample Size, Power, and Precision

A total of 7725 valid samples were included from the survey, yielding an effective recovery rate of 85.48%. All the analyses were conducted on this sample of valid questionnaires, and there were no missing data for the variables required by this study. Sample size estimation followed the standard formula for proportion-based calculations: [n=uα2pq⁄d2], where uα represents the critical value for a 95% CI (uα=1.96), *p* is the estimated proportion of short-video platform users based on the 52nd Statistical Report on Internet Development in China—Research on Internet Development [12], *q* is the complementary proportion (q=1-p), and d denotes the allowable error (1.2%). Using these parameters, the minimum required sample size was calculated as 5283.

Measures and Covariates

Self-Rated Health

Self-rated health was applied to reflect an individual's health status. Self-rated health was assessed using a visual analog scale. Participants were asked the following question: "Overall, how would you rate your current health status?" They responded by selecting a point on a horizontal scale ranging from 0 to 100, with a higher score indicating a better level of health.

Relative Health Deprivation

We measured relative health deprivation by calculating the Kakwani relative deprivation index using individuals' self-rated health, referring to the research of Kakwani [25]. The first step is to assume *G* as a reference group containing *k* sample sizes and rank the self-rated health levels of individuals in this group in ascending order. Then we can obtain $G = (g_1,g_2,g_3, \ldots ,g_{k-1},g_k)$, which is the distribution vector of self-rated health. Therefore, compared with the self-rated health of individual *i* with that of *j*, the relative health deprivation of individual *i* is as follows:

(1)RD(gj−gi)={gj−giif gj≥gi0if gj<gi

On this basis, the average relative health deprivation for individual *i* can be calculated as follows:

(2)RD(gi)=1kλG(kg+gi+−kg+gi)

where λ_G is the average self-rated health of all samples within the reference group *G*. ***kgi+*** is the number of individuals whose self-rated health exceeds that of individual *i*. λgi+ is the average self-rated health of individuals who have higher self-rated health than *i*.

XSL•FO
RenderX

The reference group for calculating relative health deprivation was defined as all participants within the same geographically stratified sampling group. This allows the deprivation measure to reflect within-region health comparisons.

Short-Video Platform Use

We took "whether to use short video software" as one of the independent variables, with the options containing yes and no. Besides, due to the multidimensional characteristics of digital engagement, years, frequency, daily duration, and purpose diversity of short-video platform use were selected to measure the usage patterns. The assessment methods for the four dimensions are as follows: (1) years of use included "≤1 year," "1-3 years," and "≥3 years," with the assigned values being 1, 2, and 3; (2) frequency of use included "several times a year," "several times a month," "1-4 times a week," and "almost every day," with values ranging from 1 to 4; (3) daily duration of use included: "≤30 minutes," "31-60 minutes," "61-120 minutes," "120-240 minutes," and "≥240 minutes," with values ranging from 1 to 5; and (4) the purpose diversity of use mainly measures the number of intended uses, regarding information acquisition, entertainment, social interaction, creative sharing, and economic acquisition. The options of number of purposes included "0," "1," "2," and "≥3," with values ranging from 1 to 4.

SES

To further analyze the role of socioeconomic factors in the relationship between short-video platform use and health, SES was incorporated into the moderating effect analysis, and it was set at 3 levels: low, middle, and high. The SES level was reported by the participants themselves.

Covariates

We used a series of demographic and socioeconomic factors, social support factors, and health-related factors as covariates. Demographic and socioeconomic factors included age, gender, urban-rural residence, SES, education, and employment status. Social support factors measured marriage, number of persons living together, and number of persons contacted regularly, following the framework proposed by van den Brink et al [27]. Among health-related factors, the Chinese multimorbidity–weighted index was used to control multimorbidity burden, which was calculated according to Hu et al [28]. Moreover, the convenience of obtaining medical services and mental health, including depression (as assessed with 10-item Center for Epidemiologic Studies Depression Scale) [29], anxiety (as assessed with 7-item Generalized Anxiety Disorder) [30], and loneliness (as assessed with University of California at Los Angeles Loneliness Scale) [31], was also controlled. The options included by these variables are shown in Table 1.

Statistical Analysis

First, 2-tailed *t* test and variance analysis were used to analyze the differences in self-rated health and relative health deprivation among different groups. In the second step, propensity score matching analysis with no replacement was performed to control for selection bias based on the probability of short-video platform use (users and nonusers). Afterward, we fitted quantile regression models on the 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, and 90th quantiles of the self-rated health and relative health deprivation distribution to investigate the differences in health between short-video platform users and nonusers. Quantile indicates the location of individuals within the outcome distribution [32]. For example, the individual in the 80th quantile of the conditional distribution has a self-rated health larger than 79% of those with similar short-video platform use behaviors. Quantile regression is a flexible and powerful tool that can assess how health relates to short-video platform use at any quantile distribution. Then, a linear regression model was applied to explore the association of different dimensions of short-video platform use with health, using the full sample. Independent variables consisted of the years, frequency, daily duration, and purpose diversity of use. Finally, we used moderating effect analysis to explore the moderating role of SES between the daily duration of use and health. A series of sensitivity analyses was conducted to test the robustness of the results. Quantile regression was conducted again based on propensity score matching with replacement. In the linear regression model, years, frequency, daily duration, and purpose diversity of use were regarded as continuous variables to be analyzed. Further, we limited the analysis of these dimensions to users of short-video platforms, so as to make the coverage more targeted. The urban-rural residence was used to replace the self-assessed SES to evaluate the moderating role of socioeconomic factors. In all models, coefficients were tested using 2-tailed *t* tests.

Stata 18.0 was used for all analyses. Statistical significance was defined as a 2-sided *P* value less than .05.

Ethical Considerations

Ethics Review and Approval

This study involved human participants and was conducted in accordance with the Declaration of Helsinki. The study protocol was reviewed and approved by the Medical Ethics Committee of Tongji Medical College, Huazhong University of Science and Technology (approval No. [2023] Lun Shen Zi (S193)).

Informed Consent

Prior to participation, all individuals were provided with detailed information about the study's purpose, procedures, potential risks, and benefits. Written informed consent was obtained from each participant.

Privacy and Confidentiality

To protect participant privacy, all collected data were anonymized at the point of collection. Personal identifiers were removed, and each participant was assigned a unique study code. The data were stored on password-protected secure servers accessible only to the authorized research team.

Participant Compensation

Participants did not receive financial compensation for participation. However, small tokens of appreciation (toothpaste of RMB ¥10 [approximately US $1.42]) were provided in accordance with local ethical norms to acknowledge participants' time.

XSL•FO
RenderX

Image and Identifiable Data Protection

This manuscript and its supplementary materials do not contain any images, videos, or personal details that could lead to the identification of an individual participant.

Results

Characteristics of Participants

The final sample consisted of 7725 participants in this study (Figure 1). Characteristics of eligible participants are presented in Table 1. Of the 7725 participants, 4403 (57%) used short-video platforms, 3628 (46.96%) were male, 6492 (84.04%) were married or cohabiting, and 2342 (30.32%) lived in urban areas. Participants who were spousal (n=3966, 90.07% vs n=2526, 76.04%) and had received education (n=6621, 74.93% vs n=5869, 44.13%) were more likely to use short-video platforms. Compared with nonusers of short-video platforms, users had lower scores of multimorbidity-weighted index (2.40, SD 1.96 vs 2.66, SD 2.09), depression (5.28, SD 4.50 vs 6.65, SD 5.02), anxiety (1.82, SD 3.22 vs 2.39, SD 3.71), and loneliness (3.54, SD 1.19 vs 3.90, SD 1.39). The differences in scores of self-rated health and relative health deprivation among different groups are shown in Table 1.

Figure 1. Flow diagram of participant recruitment and data exclusion.

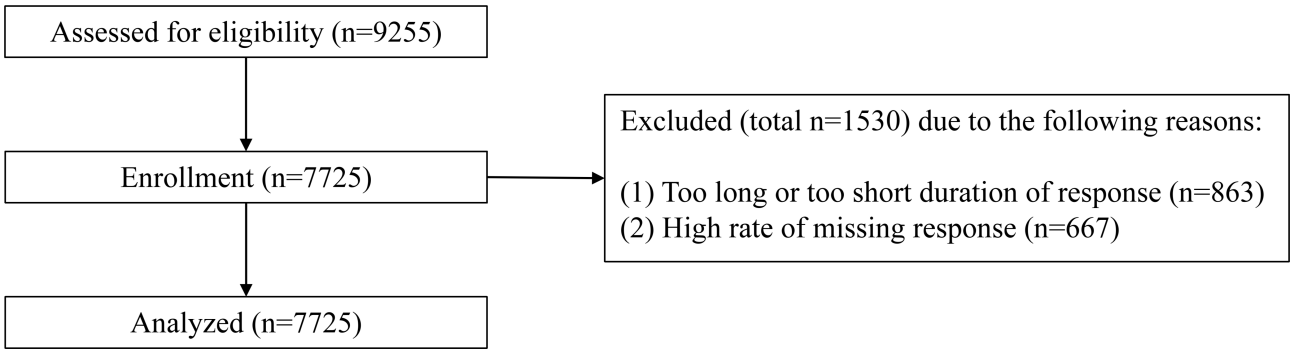

Quantile Regression Results

Figures S1-S2 in Multimedia Appendix 1 show the distribution of self-rated health and relative health deprivation in quantiles. The 2 purple curves fluctuate to a large extent near the gray uniform line, indicating that the scores of self-rated health and relative health deprivation are not normally distributed. So, it is necessary to apply quantile regression. In total, 3322 pairs of users of short-video platforms versus nonusers of short-video platforms were formed based on the propensity matching scores. The characteristics of the participants in the propensity-matched dataset are provided in Table S1 in Multimedia Appendix 1.

The quantile regression results in the propensity-matched dataset are reported in Table 2. Compared with the nonusers of short-video platforms, the self-rated health of users was higher at the 70th to 90th quantile (P<.05), and the score of the relative health deprivation of users was lower at the 10th to 30th quantile (P<.05). For self-rated health, the correlation was the strongest at the 90th quantile (β=1.931, 95% CI 1.299-2.563; P<.001). Compared to the users at the 20th and 30th quantiles, the score of relative health deprivation decreased to a lesser extent among those at the 10th percentile (β=−0.003, 95% CI −0.005 to −0.001; P<.001). Moreover, the self-rated health of users is lower than nonusers below the 10th quantile (β=−2.224, 95% CI −3.835 to −0.613; P=.007), while no significant association was observed at the 90th quantile of relative health deprivation. There was no significant association at the 20th to 60th quantile of self-rated health and the 40th to 90th quantile of relative health deprivation.

XSL•FO
RenderX

Table . Results for quantile regression analysis of the association between the use of short-video platforms and self-rated health, and relative health deprivation in propensity-matched dataset.

Quantile (reference: nonusers)	Self-rated health[a]		Relative health deprivation[b]	
	Coefficient (95% CI)	*P* value	Coefficient (95% CI)	*P* value
0.1	–2.224 (–3.835 to –0.613)	.007	–0.003 (–0.005 to –0.001)	<.001
0.2	–0.013 (–1.375 to 1.348)	.99	–0.005 (–0.007 to –0.002)	.001
0.3	0.813 (–0.355 to 1.982)	.17	–0.005 (–0.007 to –0.002)	<.001
0.4	0.544 (–0.521 to 1.608)	.32	–0.002 (–0.007 to 0.003)	.38
0.5	0.772 (–0.600 to 2.145)	.27	–0.001 (–0.006 to 0.003)	.58
0.6	0.627 (–0.304 to 1.558)	.19	–0.005 (–0.012 to 0.002)	.15
0.7	0.990 (0.067 to 1.913)	.04	–0.008 (–0.020 to 0.003)	.15
0.8	1.423 (0.600 to 2.246)	.001	–0.002 (–0.019 to 0.014)	.78
0.9	1.931 (1.299 to 2.563)	<.001	0.021 (–0.001 to 0.043)	.07

[a]Adjusted R^2: 0.179; *F*-value: 97.67; *F* significance: <.001.

[b]Adjusted R^2: 0.17; *F*-value: 91.78; *F* significance: <.001

Linear Regression Results

Correlation analysis indicated that there was a correlation among years, frequency, daily duration, and the number of use purposes of short-video platforms usage. The results of the multicollinearity test showed that there was no collinearity among these variables (variance inflation factor<5). The specific results are shown in Tables S2-S3 in Multimedia Appendix 1.

As shown in Table 3, compared with participants who have used short-video platforms for ≤1 year, those who have used them for ≥3 years showed significantly better self-rated health (β=1.970, 95% CI 0.308-3.632; *P*<.05). Yet, after adjusting for the control variables, there was no significant correlation between the years of use and relative health deprivation. There was no significant correlation between frequency of use and self-rated health and relative health deprivation. Participants who spent 61 - 120 minutes and 120 - 240 minutes per day had significantly worse self-rated health (β=−3.385, 95% CI −4.872 to −1.898; *P*<.001; β=−3.038, 95% CI −5.054 to −1.022; *P*<.01) and higher score of relative health deprivation (β=0.035, 95% CI 0.021-0.050; *P*<.001; β=0.034, 95% CI 0.014-0.054; *P*<.01). However, there was no significant relationship between the participants' daily duration of use exceeding 240 minutes and their self-rated health or relative health deprivation. Compared to no purposeful use, using short-video platforms for 2 purposes was associated with better self-rated health (β=6.082, 95% CI 0.250-11.914; *P*<.05) and lower relative health deprivation (β=−0.063, 95% CI −0.120 to −0.005; *P*<.05). The association was stronger for use with ≥3 purposes (self-rated health: β=6.213, 95% CI 0.276-12.151; *P*<.05; relative health deprivation: β=−.068, 95% CI −0.127 to −0.009; *P*<.05).

XSL•FO
RenderX

Table . Association between the usage behaviors of short-video platforms and self-rated health and relative health deprivation.

Variables	Self-rated health		Relative health deprivation	
	Model 1, coefficient (95% CI)	Model 2, coefficient (95% CI)	Model 3, coefficient (95% CI)	Model 4, coefficient (95% CI)
Years of use (reference: ≤1)				
1-3	1.775 (–0.012 to 3.562)	1.156 (–0.493 to 2.805)	–0.011 (–0.029 to 0.006)	–0.006 (–0.023 to 0.010)
≥3	4.116 (2.332 to 5.900)[a]	1.970 (0.308 to 3.632)[b]	–0.032 (–0.050 to –0.015)[a]	–0.014 (–0.031 to 0.002)
Frequency of use (reference: Several times a year)				
Several times a month	–4.718 (–11.378 to 1.941)	–5.033 (–11.13 to 1.074)	0.050 (–0.016 to 0.115)	0.053 (–0.007 to 0.114)
1-4 times a week	–4.399 (–10.485 to 1.687)	–4.673 (–10.255 to 0.907)	0.039 (–0.020 to 0.099)	0.043 (–0.012 to 0.098)
Almost every day	–4.983 (–10.997 to 1.032)	–5.250 (–10.763 to 0.262)	0.049 (–0.010 to 0.108)	0.052 (–0.002 to 0.107)
Daily duration of use (min; reference: ≤30)				
31-60	1.021 (–0.396 to 2.437)	0.046 (–1.259 to 1.350)	–0.008 (–0.022 to 0.006)	0.002 (–0.011 to 0.015)
61-120	–2.451 (–4.067 to –0.834)[c]	–3.385 (–4.872 to –1.898)[a]	0.026 (0.010 to 0.042)[c]	0.035 (0.021 to 0.050)[a]
120-240	–2.856 (–5.055, –0.657)[b]	–3.038 (–5.054 to –1.022)[c]	0.033 (0.011 to 0.054)[c]	0.034 (0.014 to 0.054)[c]
≥240	0.202 (–3.220 to 2.815)	–0.352 (–3.119 to 2.415)	0.008 (–0.022 to 0.038)	0.009 (–0.018 to 0.037)
Numbers of purposes of use (reference: 0)				
1	5.694 (–0.564 to 11.953)	4.970 (–0.773 to 10.712)	–0.055 (–0.117 to 0.006)	–0.051 (–0.108 to 0.006)
2	7.263 (0.910 to 13.616)[b]	6.082 (0.250 to 11.914)[b]	–0.071 (–0.133 to –0.008)[b]	–0.063 (–0.120 to –0.005)[b]
≥3	8.349 (1.890 to 14.807)[b]	6.213 (0.276 to 12.151)[b]	–0.084 (–0.147 to –0.020)[b]	–0.068 (–0.127 to –0.009)[b]
Control variables	No	Yes	No	Yes
Adjusted R^2	0.021	0.18	0.016	0.166
F-value	14.82	66.1	11.21	59.95
F significance	<.001	<.001	<.001	<.001

[a]*P*<.001.

[b]*P*<.05.

[c]*P*<.01.

Moderating Effect Analysis Results

As shown in Figures 2 and 3, SES was a significant moderator of the association between daily duration of use and health (all *P* for interaction <.05). The corresponding coefficients and *P* for interaction can be found in Table S4 in Multimedia Appendix 1. Specifically, compared to the high SES group, participants with middle SES showed significantly lower self-rated health (β=−2.257, 95% CI −3.540 to −0.975; *P*=.001) and higher relative health deprivation (β=0.018, 95% CI 0.005-0.031; *P*=.005). Similarly, the low SES group showed significantly poorer self-rated health (β=−2.022, 95% CI −3.349 to −0.695; *P*=.003) and a higher score of health deprivation (β=0.014, 95% CI 0.001-0.027; *P*=.04).

XSL•FO
RenderX

Figure 2. Results for moderating effect analysis between daily duration of short-video platforms usage and self-rated health. “≤30 minutes,” “31-60 minutes,” “61-120 minutes,” “120-240 minutes,” and “≥240 minutes” indicate daily duration of use. Error bars represent 95% CI.

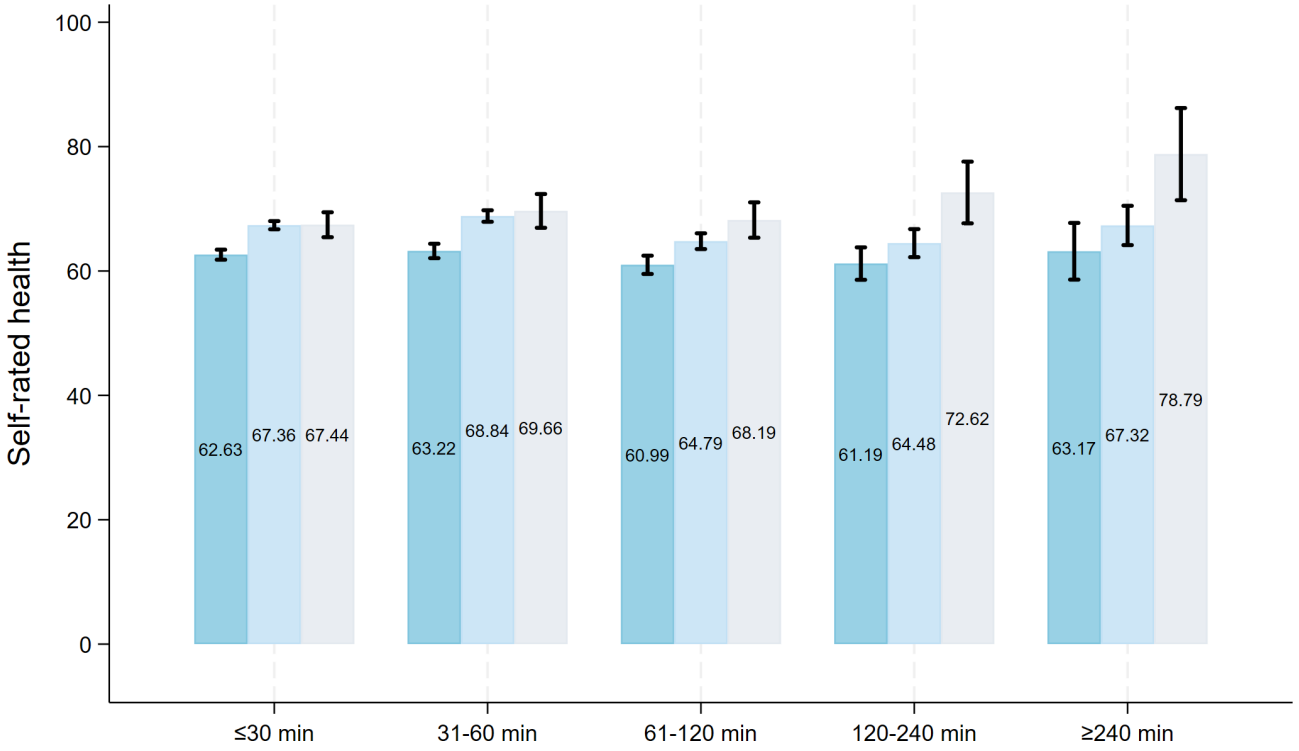

Figure 3. Results for moderating effect analysis between daily duration of short-video platforms usage and relative health deprivation. “≤30 minutes,” “31-60 minutes,” “61-120 minutes,” “120-240 minutes,” and “≥240 minutes” indicate daily duration of use. Error bars represent 95% CI.

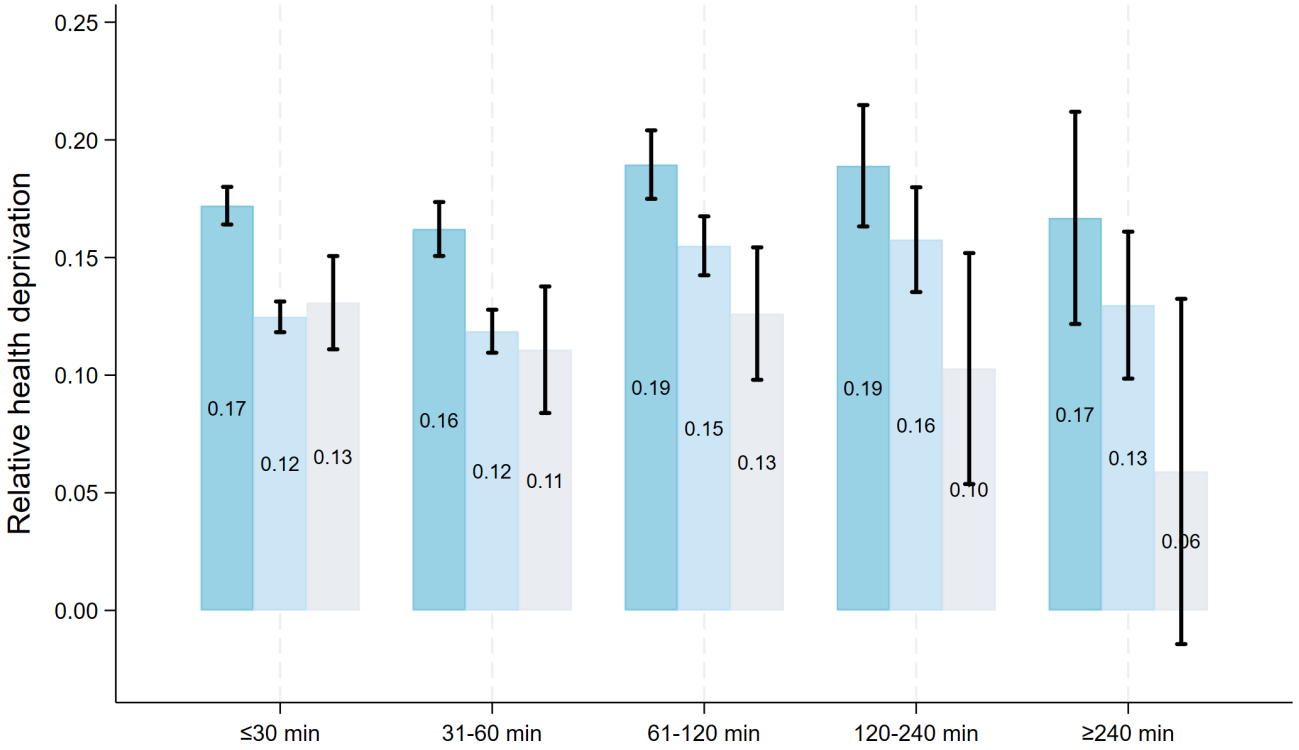

Sensitivity Analysis Results

After changing the propensity score matching method, the conclusions of the quantile regression remained unchanged. Regarding years, frequency, daily duration, and purpose diversity of usage as continuous variables in linear regression, the relevant conclusions are broadly consistent. The analysis results for the users of short-video platforms also found similar results. Moreover, compared to the urban residence group, the rural residence group had significantly lower self-rated health (β=−0.890, 95% CI −1.667 to −.113; *P*=0.03) and higher relative health deprivation (β=0.008, 95% CI 0.000-0.016; *P*=.038). This is also consistent with the previous analysis of the moderating role. All sensitivity analysis results are provided in Tables S5-S8 in Multimedia Appendix 1.

XSL•FO
RenderX

Discussion

Principal Results

This study explored the heterogeneous association between multidimensional short-video platform use and individual health among Chinese populations. The results reveal a series of nuanced findings. Only among the health-advantaged group with relatively high self-rated health (70th-90th quantile) and low relative deprivation (10th-30th quantile), using short-video platforms was associated with better self-rated health and lower relative health deprivation compared to nonuse. Moreover, short-video platform use was linked to worse self-rated health in the least healthy group (10th percentile) yet showed no significant association with health deprivation in the most relatively deprived group (90th percentile). Another finding showed that greater purpose diversity (≥2 purposes) and longer engagement (≥3 y) with short-video platforms were positively associated with better self-rated health. However, frequency of use showed no significant link to health. A daily use duration of 1 to 4 hours was associated with poorer self-rated health and higher relative deprivation, yet no such negative association was found for duration exceeding 4 hours. SES significantly moderated the association between daily use duration and relative health deprivation, with the strength of this association being more pronounced among middle- and low-SES groups than in the high-SES group.

Interpretation

Our findings demonstrate theoretical congruence with the diffusion of innovations framework. This theory posits that technology adoption follows a social gradient, often privileging early adopters who possess greater resources [33]. The health-advantaged group has preexisting health and social resources that are better equipped to navigate digital environments selectively. They can make use of these advantages for information, social connection, and stress relief, thereby further enhancing their well-being [34]. Their lower baseline deprivation may also buffer them against negative social comparison, a known risk of short-video platforms use. Conversely, the significant negative association observed in the least healthy group (10th percentile of self-rated health) suggests potential pathways of harm. Individuals in poor health may be more vulnerable to algorithmic traps and engage in passive consumption of content. Their use may exacerbate sedentary behavior or increase exposure to distressing or misleading health information [35-37]. Another noteworthy finding is that there is no significant correlation between short-video platform use and relative health deprivation among the group with the highest level of health deprivation (10th percentile). It indicates that for individuals who perceive themselves as severely disadvantaged compared to others, engagement with short-video platforms does little to relieve this sense of inequity [38]. Their deprivation likely depends on structural factors, such as the severity of the disease and socioeconomic disadvantage. These cannot be easily relieved by the fragmented social support and information on short-video platforms. For individuals at the 20th to 60th percentiles of self-rated health and at the 40th to 90th percentiles of relative health deprivation, no significant association was observed. This group typically fits the description of the early majority as defined in diffusion of innovations theory [33]. Their habits of use are more likely to be influenced by the inherent characteristics of short videos, which are typically characterized by frequent but brief, passive, and entertainment-oriented consumption [39]. Together, this pattern of use tends to make short videos a relatively insignificant supplementary tool. Its association with cognitive, emotional, and behavioral remains diffuse and thus below the threshold of detectable statistical association in a cross-sectional design. Notably, our finding of a modest reduction in relative health deprivation ($\beta=-0.003$) at the 10th percentile requires careful interpretation. While this change is small at the individual level, its consistent detection at the lower tail of the deprivation distribution holds broader significance. It suggests that even among those who already feel relatively less deprived, short-video platform use may contribute a marginal, yet statistically robust, psychological sense of gain. From a population health perspective, a small shift that is reliably present across a large subgroup can translate into a meaningful aggregate benefit [40].

The differentiated associations observed between behaviors of short-video platform use and health suggest that how and why individuals engage with these platforms may matter more than how often they do so. The positive association of purpose diversity (≥2 purposes) with health underscores the importance of an active, agentic attitude for short-video platforms use. Short-video platforms are engineered for high engagement through infinite scroll and personalized feeds, which typically promote passive viewing [41]. Users who could actively use the platform for information seeking, social interaction, and content creation likely mitigate risks of mindless scrolling [42-44]. The association linked to longer engagement (≥3 y) may reflect digital literacy increases with the accumulation of use time. New users need to spend time exploring and mastering the functions of the short-video platforms, while long-term users may develop more regular and proactive habits. They might be able to use the platform's functions more flexibly and benefit from it [45,46]. No significant correlation between frequency of use and health is observed, which is inconsistent with the previous studies regarding internet usage [10,47,48]. It might be due to the inherent characteristics of the short-video platforms. Unlike traditional media consumption, where each session constitutes a discrete, sustained period of attention, short-video use is characterized by brief and repetitive engagements [49]. Therefore, the frequency of use may not accurately reflect its association with health, as it fails to distinguish between these short usage periods and more prolonged, immersive usage periods. For the daily duration of use, using short-video platforms for 1 to 4 hours per day is associated with poorer health. This might indicate the psychological fatigue caused by excessive use and the adverse physical activities, such as sedentary behavior [5,50,51]. The daily duration of use exceeding 4 hours was not significantly associated with self-rated health and relative health deprivation. This might be related to the user's occupation, pattern of use, and so on. This requires further analysis in subsequent studies.

The moderating role of SES observed in this study also aligns with the innovation diffusion theory and the digital divide theory

XSL•FO
RenderX

[33]. From the perspective of innovation diffusion, high-SES individuals often act as early adopters. Their greater material and cognitive resources facilitate not only access but also more adaptive and beneficial engagement with new technologies, such as short-video platforms. This translates into higher digital health literacy, better self-regulatory capacity, and the ability to use these platforms more selectively [52-54]. Conversely, the theory of the second-level digital divide explains the heightened vulnerability of middle- and low-SES groups [55]. Disparities in digital skills and resources often confine their engagement to more passive, entertainment-focused consumption patterns [56]. This pattern of use, characterized by prolonged but less intentional exposure, makes them more susceptible to the platforms' risks, such as unfavorable social comparison, algorithmic traps, and sedentary behavior [57]. Thus, preexisting socioeconomic inequalities are reproduced and potentially amplified in the digital realm, exacerbating feelings of relative health deprivation among the already disadvantaged.

Limitations

Several limitations of this study should be acknowledged. First, despite the application of propensity score matching to mitigate selection bias, the inherent constraints of cross-sectional observational design preclude definitive causal inferences regarding short-video platform use and health outcomes [24]. Second, while multistage stratified sampling was implemented to ensure regional diversity, potential sampling biases may exist due to nonprobability sampling methods and self-reported participation rates, limiting the generalizability of the findings to the broader Chinese adult population. Third, self-rated health as a subjective measure is susceptible to social desirability bias and cultural interpretation variance, potentially underestimating or overestimating actual health disparities. Fourth, although multiple covariates were statistically controlled, residual confounding effects from unmeasured variables, such as cognitive function and digital literacy, cannot be entirely ruled out, particularly given their demonstrated associations with both health behaviors and technology engagement patterns. Fifth, SES relying on self-report may have measurement error. Future studies would benefit from incorporating objective indicators, such as household income brackets, educational attainment levels, and occupational prestige scores to enhance construct validity. Sixth, the categorization of screen time into discrete intervals limits our ability to model dose-response relationships or detect nonlinear threshold effects within usage categories. This limitation could be addressed through device-based time-use monitoring in subsequent research. Seven, the lack of content type differentiation obscures the understanding of how specific platform features link to health, an issue that could be resolved through qualitative content analysis. Finally, the study focused on short-video platforms and did not capture the potential association between other digital media forms and health. Future longitudinal studies should incorporate specific contents of usage to better identify for whom and under what conditions digital engagement supports health.

Conclusions

This study advances the field by moving beyond the examination of overall associations in the novel context of short-video platforms. It investigates for whom and under what conditions the use of these platforms is linked to health. By exploring multidimensional usage behaviors, it identifies distinct patterns of benefit and risk across population subgroups, providing a nuanced understanding of the relationship between digital engagement and health. Self-rated health and relative health deprivation help to clarify an individual's subjective perception of health, as well as the sense of health inequality derived from social comparisons. This study carries significant associations between short-video platform use and individual health, which are heterogeneous across population subgroups. Policy must therefore shift from ensuring universal access to actively governing for equitable digital well-being. This requires targeted protections for vulnerable groups, especially those in poor health or with lower SES. They are most susceptible to harm from passive or prolonged use. Meanwhile, the observed disparities in relative health deprivation suggest public policy should aim not only to improve material health access but also to intentionally shape digital environments to address both material and perceived health disparities.

Acknowledgments

We thank all the volunteer participants for their patience and participation. The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing. The GAI tool used was DeepSeek-V3.2. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was supported by National Key R&D Program in China (2022YFE0133000), National Nature Science Foundation of China (72374079), Hubei Provincial Department of Science and Technology (JCZRYB202500352). The funders had no involvement in the study design, data collection, data analysis, interpretation of results, or the writing of the manuscript.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to institutional data protection policies and participant confidentiality regulations. The data may be made available from the corresponding author upon reasonable request by qualified researchers for academic purposes.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: YP
Data curation: KZ
Formal analysis: YP
Funding acquisition: ST
Investigation: KZ, YH, YW
Methodology: ST
Project administration: ST
Resources: ST
Software: YP
Supervision: ST
Validation: YW
Visualization: YH
Writing – original draft: YP
Writing – review and editing: CL, CY, YP

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional results.
[DOCX File, 201 KB - jmir_v28i1e86526_app1.docx]

References

1. Altanlar A, Eren İ, Aktuğlu Aktan E. Digital technologies and public life in Istanbul: socio-demographic patterns and urban implications. Cities 2025 Oct;165:106179. [doi: 10.1016/j.cities.2025.106179]
2. Tavares AI. Self-assessed health among older people in Europe and internet use. Int J Med Inform 2020 Sep;141:104240. [doi: 10.1016/j.ijmedinf.2020.104240] [Medline: 32739610]
3. Li X, Liu J, Huang N, Zhao W, He H. Association between internet use and sleep health among middle-aged and older Chinese individuals: nationwide longitudinal study. J Med Internet Res 2025;27:e71030. [doi: 10.2196/71030] [Medline: 40239202]
4. Wen W, Zhang Y, Shi W, Li J. Association between internet use and physical health, mental health, and subjective health in middle-aged and older adults: nationally representative cross-sectional survey in China. J Med Internet Res 2023 Mar 21;25:e40956. [doi: 10.2196/40956] [Medline: 36943368]
5. Choudhury A, Asan O, Alelyani T. Exploring the role of the internet, care quality and communication in shaping mental health: analysis of the Health Information National Trends Survey. IEEE J Biomed Health Inform 2022 Jan;26(1):468-477. [doi: 10.1109/JBHI.2021.3087083] [Medline: 34097623]
6. Ma JQ, Sheng L. Internet use time and mental health among rural adolescents in China: a longitudinal study. J Affect Disord 2023 Sep 15;337(18-26):18-26. [doi: 10.1016/j.jad.2023.05.054] [Medline: 37236271]
7. Tirado-Morueta R, Aguaded-Gómez JI, Ortíz-Sobrino M, Rodríguez-Martín A, Álvarez-Arregui E. Determinants of social gratifications obtained by older adults moderated by public supports for Internet access in Spain. Telemat Inform 2020 Jun;49:101363. [doi: 10.1016/j.tele.2020.101363]
8. Shim H, Ailshire J, Zelinski E, Crimmins E. The Health and Retirement Study: analysis of associations between use of the internet for health information and use of health services at multiple time points. J Med Internet Res 2018 May 25;20(5):e200. [doi: 10.2196/jmir.8203] [Medline: 29802088]
9. Thom RP, Bickham DS, Rich M. Internet use, depression, and anxiety in a healthy adolescent population: prospective cohort study. JMIR Ment Health 2018 May 22;5(2):e44. [doi: 10.2196/mental.8471] [Medline: 29789282]
10. Rana J. 1156 Associations between internet use and overweight/obesity in Nepal: do gender and urbanity modify these associations? Int J Epidemiol 2021 Sep 1;50(Supplement_1). [doi: 10.1093/ije/dyab168.559]
11. Guan JL, Xia SH, Zhao K, et al. Videos in short-video sharing platforms as sources of information on colorectal polyps: cross-sectional content analysis study. J Med Internet Res 2024 Oct 29;26:e51655. [doi: 10.2196/51655] [Medline: 39470708]
12. The 52nd statistical report on internet development in China. : China Internet Network Information Center; 2023 URL: https://www.cnnic.com.cn/IDR/ReportDownloads/202311/P020231121355042476714.pdf [accessed 2026-02-20]
13. Wu J, Peng H, Fan Z, Zhao H. Family guidance weakens an effect of old people's educational attainment on eHealth literacy by internet experience. Innov Aging 2023 Dec 21;7(Supplement_1):953-954. [doi: 10.1093/geroni/igad104.3063]
14. Jiang S, Beaudoin CE. Health literacy and the internet: an exploratory study on the 2013 HINTS survey. Comput Human Behav 2016 May;58:240-248. [doi: 10.1016/j.chb.2016.01.007]

XSL•FO
RenderX

15. Mitsutake S, Shibata A, Ishii K, Oka K. Associations of eHealth literacy with health behavior among adult internet users. J Med Internet Res 2016 Jul 18;18(7):e192. [doi: 10.2196/jmir.5413] [Medline: 27432783]
16. Nakagomi A, Shiba K, Kawachi I, et al. Internet use and subsequent health and well-being in older adults: an outcome-wide analysis. Comput Human Behav 2022 May;130:107156. [doi: 10.1016/j.chb.2021.107156]
17. Szreder E, Krysta K. Internet and mobile phone addiction in individuals aged 15-50 years. Eur Psychiatr 2025 Apr;68(S1):S465-S465. [doi: 10.1192/j.eurpsy.2025.963]
18. Kovačić Petrović Z, Peraica T, Kozarić-Kovačić D, Palavra IR. Internet use and internet-based addictive behaviours during coronavirus pandemic. Curr Opin Psychiatry 2022 Sep 1;35(5):324-331. [doi: 10.1097/YCO.0000000000000804] [Medline: 35855504]
19. Mitchell ME, Lebow JR, Uribe R, Grathouse H, Shoger W. Internet use, happiness, social support and introversion: a more fine grained analysis of person variables and internet activity. Comput Human Behav 2011 Sep;27(5):1857-1861. [doi: 10.1016/j.chb.2011.04.008]
20. Zhao H, Peng HM, Fan ZY, Wu JX. Different types of Internet use and social adaptation among older adults. Innov Aging 2023 Dec 21;7(Supplement_1):669. [doi: 10.1093/geroni/igad104.2175]
21. Hernández C, Cottin M, Parada F, et al. Watching the world from my screen: a longitudinal evaluation of the influence of a problematic use of the internet on depressive symptomatology. Comput Human Behav 2022 Jan;126:106995. [doi: 10.1016/j.chb.2021.106995]
22. Dahlgren G, Whitehead M. The Dahlgren-Whitehead model of health determinants: 30 years on and still chasing rainbows. Public Health 2021 Oct;199:20-24. [doi: 10.1016/j.puhe.2021.08.009] [Medline: 34534885]
23. Gann B. Combating digital health inequality in the time of coronavirus. J Consum Health Internet 2020 Jul 2;24(3):278-284. [doi: 10.1080/15398285.2020.1791670]
24. Garbarski D. Research in and prospects for the measurement of health using self-rated health. Public Opin Q 2016;80(4):977-997. [doi: 10.1093/poq/nfw033] [Medline: 27833212]
25. Kakwani N. The relative deprivation curve and its applications. J Bus Econ Stat 1984 Oct;2(4):384-394. [doi: 10.1080/07350015.1984.10509412]
26. Levitt HM, Bamberg M, Creswell JW, Frost DM, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):26-46. [doi: 10.1037/amp0000151] [Medline: 29345485]
27. van den Brink RHS, Schutter N, Hanssen DJC, et al. Prognostic significance of social network, social support and loneliness for course of major depressive disorder in adulthood and old age. Epidemiol Psychiatr Sci 2018 Jun;27(3):266-277. [doi: 10.1017/S2045796017000014] [Medline: 28183368]
28. Hu WH, Liu YY, Yang CH, et al. Developing and validating a Chinese multimorbidity-weighted index for middle-aged and older community-dwelling individuals. Age Ageing 2022 Feb 2;51(2):afab274. [doi: 10.1093/ageing/afab274] [Medline: 35211718]
29. Andresen EM, Malmgren JA, Carter WB, Patrick DL. Screening for depression in well older adults: evaluation of a short form of the CES-D. Am J Prev Med 1994 Mar;10(2):77-84. [doi: 10.1016/S0749-3797(18)30622-6]
30. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
31. Russell D, Peplau LA, Cutrona CE. The revised UCLA Loneliness Scale: concurrent and discriminant validity evidence. J Pers Soc Psychol 1980 Sep;39(3):472-480. [doi: 10.1037//0022-3514.39.3.472] [Medline: 7431205]
32. Koenker R, Bassett G. Regression quantiles. Econometrica 1978 Jan;46(1):33. [doi: 10.2307/1913643]
33. Rogers EM. Diffusion of Innovations, 5th edition: Free Press; 2003. URL: https://www.simonandschuster.com/books/Diffusion-of-Innovations-5th-Edition/Everett-M-Rogers/9780743222099?utm_source=chatgpt.com [accessed 2026-02-20]
34. Rademakers J, Heijmans M. Beyond reading and understanding: health literacy as the capacity to act. Int J Environ Res Public Health 2018 Aug 7;15(8):1676. [doi: 10.3390/ijerph15081676] [Medline: 30087254]
35. Theodoratou M, Patiri I, Katsarou D, et al. The link between problematic internet use, physical activity, and mental health: implications for depression and anxiety. Eur Psychiatr 2025 Apr;68(S1):S806-S807. [doi: 10.1192/j.eurpsy.2025.1641]
36. Cline RJ, Haynes KM. Consumer health information seeking on the Internet: the state of the art. Health Educ Res 2001 Dec;16(6):671-692. [doi: 10.1093/her/16.6.671] [Medline: 11780707]
37. Ang S, Lim E, Malhotra R. Digital independence: health-related difficulty in internet use and how it affects older adults' quality of life. Innov Aging 2019 Nov 8;3(Supplement_1):S442-S443. [doi: 10.1093/geroni/igz038.1661]
38. McCloud RF, Okechukwu CA, Sorensen G, Viswanath K. Beyond access: barriers to internet health information seeking among the urban poor. J Am Med Inform Assoc 2016 Nov;23(6):1053-1059. [doi: 10.1093/jamia/ocv204] [Medline: 27206459]
39. Pinho C, Franco M, Mendes L. Application of innovation diffusion theory to the E-learning process: higher education context. Educ Inf Technol 2021 Jan;26(1):421-440. [doi: 10.1007/s10639-020-10269-2]
40. Thompson C. Rose's prevention paradox. J Applied Philosophy 2018 May;35(2):242-256. [doi: 10.1111/japp.12177]

XSL•FO
RenderX

41. Xu J, Ismail AB, Jamal JB. The impact of mobile short video platform characteristics on users' continuous use behavior in China: a stimulus-organism-response perspective [Article in Russian]. Multidiscip Sci J 2024 Jul;7(1):2025019. [doi: 10.31893/multiscience.2025019]
42. Ghweeba M, Lindenmeyer A, Shishi S, Abbas M, Waheed A, Amer S. What predicts online health information-seeking behavior among Egyptian adults? A cross-sectional study. J Med Internet Res 2017 Jun 22;19(6):e216. [doi: 10.2196/jmir.6855] [Medline: 28642216]
43. Vošner HB, Bobek S, Kokol P, Krečič MJ. Attitudes of active older Internet users towards online social networking. Comput Human Behav 2016 Feb;55:230-241. [doi: 10.1016/j.chb.2015.09.014]
44. Lee S, Kim BG. User, system, and social related factors affecting perceived usefulness for continuance usage intention of mobile apps. Int J Mob Commun 2021;19(2):190-217. [doi: 10.1504/IJMC.2021.113457]
45. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
46. Arcury TA, Sandberg JC, Melius KP, et al. Older adult internet use and eHealth literacy. J Appl Gerontol 2020 Feb;39(2):141-150. [doi: 10.1177/0733464818807468] [Medline: 30353776]
47. Falk Erhag H, Ahlner F, Rydberg Sterner T, Skoog I, Bergström A. Internet use and self-rated health among Swedish 70-year-olds: a cross-sectional study. BMC Geriatr 2019 Dec 23;19(1):365. [doi: 10.1186/s12877-019-1392-8] [Medline: 31870313]
48. Luo Y, Yip PSF, Zhang Q. Positive association between Internet use and mental health among adults aged ≥50 years in 23 countries. Nat Hum Behav 2025 Jan;9(1):90-100. [doi: 10.1038/s41562-024-02048-7] [Medline: 39558112]
49. Yang Y, Liu RD, Ding Y, Lin J, Ding Z, Yang X. Time distortion for short-form video users. Comput Human Behav 2024 Feb;151:108009. [doi: 10.1016/j.chb.2023.108009]
50. Kraut R, Patterson M, Lundmark V, Kiesler S, Mukopadhyay T, Scherlis W. Internet paradox. A social technology that reduces social involvement and psychological well-being? Am Psychol 1998 Sep;53(9):1017-1031. [doi: 10.1037//0003-066x.53.9.1017] [Medline: 9841579]
51. Bessière K, Pressman S, Kiesler S, Kraut R. Effects of internet use on health and depression: a longitudinal study. J Med Internet Res 2010 Feb 28;12(1):e6. [doi: 10.2196/jmir.1149] [Medline: 20228047]
52. Anrijs S, Mariën I, De Marez L, Ponnet K. Excluded from essential internet services: examining associations between digital exclusion, socio-economic resources and internet resources. Technol Soc 2023 May;73:102211. [doi: 10.1016/j.techsoc.2023.102211]
53. Kobul MK. Socioeconomic status influences Turkish digital natives' internet use habitus. Behav Inf Technol 2023 Apr 4;42(5):624-642. [doi: 10.1080/0144929X.2022.2034970]
54. Shin YM. Internet addiction in children and adolescents. In: Internet Addiction in Children and Adolescents: Risk Factors, Assessment, and Treatment: Springer Publishing Company; 2017:29-45. [doi: 10.1891/9780826133731.0002]
55. Van Dijk J. The Digital Divide: Polity Press; 2019. URL: https://www.researchgate.net/publication/336775102_The_Digital_Divide [accessed 2026-02-20]
56. Ndasauka Y, Kayange GM. Introduction: making a case for addiction in Africa. In: Ndasauka Y, Kayange GM, editors. Addiction in South and East Africa: Interdisciplinary Approaches: Palgrave Macmillan Cham; 2019:1-6. [doi: 10.1007/978-3-030-13593-5_1]
57. Scheerder A, van Deursen A, van Dijk J. Determinants of Internet skills, uses and outcomes. A systematic review of the second- and third-level digital divide. Telemat Inform 2017 Dec;34(8):1607-1624. [doi: 10.1016/j.tele.2017.07.007]

Abbreviations

SES: socioeconomic status

Edited by S Brini; submitted 26.Oct.2025; peer-reviewed by A Famotire, B Najand, PA Okooboh, WS Babalola; accepted 29.Jan.2026; published 12.Mar.2026.

Please cite as:
Pan Y, Zhang K, Wei Y, Huang Y, Long C, Yang C, Tang S
Associations Between Short-Video Platform Use and Health Across Health Distribution and Usage Behaviors in China: Cross-Sectional Questionnaire Study
J Med Internet Res 2026;28:e86526
URL: https://www.jmir.org/2026/1/e86526
doi:10.2196/86526

XSL•FO
RenderX

© Yangyang Pan, Kangkang Zhang, Yilin Wei, Yangzhen Huang, Chengxu Long, Chenxin Yang, Shangfeng Tang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

User Profiles and Engagement in a Hypertension Self-Management App: Cross-Sectional Survey

Felix Muehlensiepen[1,2,3,4], MPH, PhD, Dr rer medic habil; Susann May[1,2,3], MPH, Dr rer medic; Frances Seifert[1,2,3], MSc; Eileen Wengemuth[1], Dr rer nat; Olen Johannsen[5], MSc; Martin Middeke[6], Prof, Dr; Martin Heinze[1,7], Prof, Dr; Pascal Petit[4], PhD; Nicolas Vuillerme[4], Prof, Dr; Sebastian Spethmann[2,3*], Prof, Dr; Dunja Bruch[8*], MSc

[1]Center for Health Services Research, Faculty of Health Sciences Brandenburg, Brandenburg Medical School, Rüdersdorf, Seebad 82/83, Ruedersdorf bei Berlin, Germany

[2]Department of Cardiology, Angiology and Intensive Care Medicine, Deutsches Herzzentrum der Charité, Berlin, Germany

[3]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[4]Université Grenoble Alpes, CNRS, Grenoble INP, LIG Sangria, Grenoble, France

[5]Hypertension Care UG, Ingolstadt, Germany

[6]Hypertension Center Munich, Munich, Germany

[7]Center for Mental Health, Brandenburg Medical School, Immanuel Klinik Rüdersdorf, Ruedersdorf bei Berlin, Germany

[8]Department of Cardiovascular Surgery, Heart Center Brandenburg, Faculty of Health Sciences, Brandenburg Medical School Theodor Fontane, Bernau, Germany

[*]these authors contributed equally

Corresponding Author:
Felix Muehlensiepen, MPH, PhD, Dr rer medic habil
Center for Health Services Research, Faculty of Health Sciences Brandenburg, Brandenburg Medical School, Rüdersdorf, Seebad 82/83, Ruedersdorf bei Berlin, Germany

Related Article:

This is a corrected version. See correction statement: https://www.jmir.org/2026/1/e93515

Abstract

Background: Mobile health (mHealth) technologies can improve hypertension self-management, yet real-world adoption remains limited and unequally distributed.

Objective: This study aimed to characterize the profiles, usage patterns, and engagement of active users of a hypertension self-management app (*Hypertension.APP*) in Germany, with a focus on user engagement and potential digital divides.

Methods: We conducted a cross-sectional online survey among adult users of *Hypertension.APP* in Germany between January and September 2023. An 88-item questionnaire assessed app usage patterns, perceived utility, integration into clinical care, sociodemographic and clinical data, and digital health literacy (eHealth Literacy Scale; scores 16 - 40). Digital health literacy was categorized as low (16 - 23.99), moderate (24 - 31.99), or high (32-40). Descriptive statistics and univariable ordinal logistic regression were used to explore associations between sociodemographic and clinical variables and app usage frequency.

Results: Of 254 respondents, the mean age was 53.6 years, and 54.3% (138/254) were male. A total of 44.5% (113/254) had a university or technical college degree, and 44.5% (113/254) reported a monthly net income higher than €2500 (US $2950). Most participants (224/254, 88.2%) reported access to at least two digital devices. Overall, 88.2% (224/254) had moderate or high digital health literacy (eHealth Literacy Scale ≥24). App engagement was high: 80.7% (205/254) reported using the app at least weekly, and 52.4% (133/254) reported using the app to prepare for medical visits. However, only 20.1% (51/254) reported that the app was formally integrated into their medical care, and 11.8% (30/254) indicated that medication had been adjusted based on app data. In univariable ordinal logistic regression analyses, higher education, longer duration of hypertension, and living in a small town (5000 - 20,000 inhabitants) were associated with more frequent app use, whereas systolic blood pressure of 140 mm Hg or higher was associated with less frequent use. Digital health literacy was not clearly associated with app usage frequency among current users.

Conclusions: Users of this hypertension self-management app were predominantly well-educated, digitally literate individuals with established hypertension, reinforcing concerns about a persistent digital divide. While app usability and engagement were high, formal clinical integration remained limited. Simply making an app available is insufficient; strategies to promote equitable

XSL•FO
RenderX

access, strengthen clinical integration, and support patients with lower digital health literacy are needed for mHealth to contribute effectively to hypertension management.

Trial Registration: German Clinical Trials Register DRKS00029761; https://drks.de/search/en/trial/DRKS00029761/details

(J Med Internet Res 2026;28:e83075) doi:10.2196/83075

KEYWORDS

digital health; mHealth; digital divide; hypertension; cardiology; chronic disease; questionnaire; health services research; user perspectives; equity; digital public health; mobile health

Introduction

Cardiovascular diseases remain the leading cause of death worldwide, despite the availability of evidence-based strategies for prevention and treatment [1]. Arterial hypertension—affecting more than 1 billion people globally—is a major modifiable risk factor [2]. In addition to genetic predisposition, the development and progression of high blood pressure are strongly influenced by individual lifestyle factors, health literacy, and behaviors [3], as well as by structural determinants such as access to and quality of medical care [4].

In recent years, digital health technologies have gained increasing relevance in the prevention and management of chronic conditions, including hypertension [5]. Mobile health (mHealth) applications, wearable devices, and other digital tools offer new opportunities for individuals to monitor their health, adopt healthier behaviors, and take a more active role in managing their condition. These tools also provide health professionals with novel ways to support daily care routines and improve treatment adherence [6].

Multiple studies have shown that mHealth interventions can contribute to better blood pressure control [7]. Most recently, Prendergast et al [8] demonstrated in a landmark single-center randomized clinical trial that an emergency department–based education and mHealth empowerment intervention for patients with hypertension led to significantly greater reductions in systolic blood pressure (SBP) at 6 months compared with standard discharge care.

However, findings from controlled trials may not translate to routine practice. Evidence from other areas of digital health suggests that adherence to mHealth interventions in everyday practice tends to be low [9-11]. Common barriers include limited awareness of mHealth benefits among both patients and health care professionals [12-14], insufficient reimbursement structures [15], and difficulties integrating digital tools into existing care pathways [16]. While these challenges have been widely discussed from the perspective of health systems and providers, less is known about the individuals who actively use such tools. Gaining insight into who uses hypertension apps, how they are used, and what users expect from them can support the development of more effective and user-centered digital health solutions.

In a preceding qualitative study [17], we explored patient users' perspectives on a hypertension self-management app (*Hypertension.APP*), focusing on perceived benefits, potential risks, and the app's role in everyday care. Patients valued features such as continuous monitoring and personalized feedback but reported limited integration into formal health care and expressed concerns about data privacy. Moreover, the findings suggested that users tended to be digitally literate, self-effective, and relatively well-educated. These qualitative insights provided important initial evidence but were based on a small sample (n=20). The present study aims to build on these findings using a quantitative approach to provide a broader and more generalizable understanding of app users in hypertension care.

The objective of this cross-sectional study was to investigate the profiles, usage patterns, and engagement of active users of a hypertension self-management app. Using a standardized survey, we examined user demographics, usage patterns, usability, and clinical integration of *Hypertension.APP*.

Methods

Study Design

This study was conducted as a cross-sectional study. It examines how users of *Hypertension.APP* evaluate their experiences and the app's perceived impact on their hypertension care. The data provide insights into user attitudes, app engagement, and the role of digital health technologies in hypertension management.

This study is part of the Digital Preventive Measures for Arterial Hypertension project [18], which examines the structural and individual factors of patients and stakeholders in the use of digital preventive measures among patients with arterial hypertension in Germany.

Setting

The study was conducted via online recruitment, targeting adult users of *Hypertension.APP* across Germany. Recruitment and data collection were carried out between January 1 and September 30, 2023. Users were approached within their real-world use context of the app. Potential participants were invited via 3 digital channels: (1) in-app notifications within *Hypertension.APP*, (2) email invitations sent to registered users, and (3) social media posts on channels maintained by the app provider and project partners. All invitations briefly described the purpose of the study, stated the eligibility criteria, and provided a link to the online questionnaire.

Participants

Eligibility criteria included individuals aged 18 years or older who had used the app for at least 1 month. Exclusion criteria included individuals without access to a smartphone or tablet and those unable to provide informed consent. After clicking on the survey link, interested individuals were first presented

XSL•FO
RenderX

with an online information sheet and an electronic informed consent form. Only participants who provided consent were forwarded to the actual questionnaire. Participation was voluntary and anonymous, and respondents could skip individual questions if they did not wish to answer them.

Sample Size

The planned sample size was set at N=200, based on an a priori power analysis (G*Power) for the chi-square test with the following assumptions: medium effect (w=0.2), α=.05, power (1-β)=.80, *df*=1 (required sample size n=197). Due to the high willingness of app users to participate, the sample size was exceeded.

mHealth App: *Hypertension.APP*

Based on expert recommendations (Prof Dr Martin Middeke and Prof Dr med Miklós Illyés), *Hypertension.APP* was selected as an example and examined through a standardized user survey. *Hypertension.APP* was developed and is maintained by Hypertension Care UG (Ingolstadt) [19]. The application had been downloaded in app stores nearly 11,000 times as of July 2025. It includes a diary function to document blood pressure data, a personalized guide that provides information based on the user's health profile, an export function to generate reports for medical consultation, a reminder function, and interactive exercises. The app is freely accessible in various app stores. There is also a purchasable premium version that includes additional features.

Two authors (OJ and MM) are the founders and/or chief executive officers of Hypertension Care UG, the company that developed *Hypertension.APP*, the app evaluated in this study. To minimize potential conflicts of interest, their role in the research process was restricted. The academic research team (FM, SM, DB, PP, NV, SS) independently developed the questionnaire, conducted data collection, managed the dataset, and performed all statistical analyses. The app developers did not have access to raw data and were not involved in data analysis or interpretation. All authors contributed to manuscript drafting, but final decisions on analysis and interpretation were made by the independent academic team.

Questionnaire

The development of the 88-item questionnaire followed a structured, multistep process. First, an initial item pool was generated based on preliminary findings from our preceding qualitative interview study on hypertension app use [17], existing literature on mHealth adoption, and theoretical frameworks related to self-management and digital health literacy. To ensure content validity, we integrated established measurement instruments, including the German version of the eHealth Literacy Scale (eHEALS) [20] and system usability scale (SUS) [21], and supplemented these with newly developed items capturing app-specific use cases, clinical integration, and user expectations. Second, the research team (FM, SM, EW, DB) iteratively refined item wording and domain structure to improve face validity, comprehensiveness, and consistency. This process included expert review by clinicians and digital health researchers with experience in hypertension management and questionnaire development (OJ, MM, SS). Third, the questionnaire was pilot-tested with a sample of 10 adult hypertension app users to evaluate clarity, item relevance, and response burden. Feedback from the pilot phase resulted in minor adjustments to wording, response options, and the sequence of questions to enhance comprehensibility and usability.

The final instrument (German version: Multimedia Appendix 1; English version: Multimedia Appendix 2) covered multiple domains: app usage patterns, integration into medical care, perceived utility of specific app features, usability (SUS), digital health literacy (eHEALS), clinical characteristics, health behaviors, and sociodemographic information.

Variables

Primary variables included user demographics, eHealth literacy, app engagement patterns, and perceptions of app utility. eHealth literacy was assessed using the German version of the eHEALS scale [20] and categorized into "Low" (16 - 23.99), "Moderate" (24 - 31.99), and "High" (32–40) levels. These thresholds follow the categorization approach proposed by Monkman et al [22], who examined conceptual distinctions within the scale and recommended these score ranges to differentiate levels of digital health literacy. App utility was measured using the German version of the SUS [21]. App engagement metrics, such as frequency and duration of use, were assessed alongside subjective evaluations of the app's role in hypertension care. Sociodemographic data, including age, gender, education, and place of residence, were also collected.

Data Sources and Measurement

All questionnaire data were self-reported, including blood pressure values. Participants provided their most recent blood pressure measurement, and no standardized measurement protocol was imposed. Participants were instructed to report values documented in *Hypertension.APP* or measured through their usual home monitoring routines. Accordingly, systolic and diastolic values reflect self-reported blood pressure at or near the time of survey completion rather than standardized study measurements.

Bias

Efforts to mitigate bias included the use of validated instruments, such as eHEALS scale [20] and SUS [22], and standardized survey procedures. Missing data were handled by including only observations with less than 20% missing values.

Statistical Methods

Data analysis was performed using R statistical software (version 2023.06.2; R Core Team, 2023). Descriptive statistics were used to summarize participant characteristics, app usage patterns, and SUS scores. Subgroup analyses were conducted for app usage frequency to explore variations by demographic factors (age, gender, place of residence, and education level). SUS scores were reported descriptively for the overall sample without subgroup analyses. To examine whether sociodemographic variables and digital health literacy (eHEALS) predicted app usage frequency, a univariable ordinal logistic regression analysis was performed using the MASS package [23].

XSL•FO
RenderX

Ethical Considerations

The study was approved by the Ethics Committee of Brandenburg Medical School Theodor Fontane (E-02 - 20220620). The study was registered as part of the Digital Preventive Measures for Arterial Hypertension project at the German Clinical Trials Register (DRKS00029761). Primary data collection was conducted in compliance with the current data protection regulations of the General Data Protection Regulation and the Declaration of Helsinki. Participants provided informed consent before being forwarded the actual questionnaire. Data collection was anonymous, and no conclusions can be drawn about individuals in the presentation of results. Participants received €15.00 (US $17.53) as an incentive and compensation for participation. The anonymity of the survey data during incentivization was ensured by using a separate database and individual passwords, with the two datasets stored separately. Personal data used for incentivization were deleted. There was no formal affiliation or relationship between the hypertension app examined in this study and the overarching research project. The app was developed and operated independently, and the collection and analysis of usage data were conducted solely for scientific purposes within the scope of this study.

Results

Demographic Characteristics

A total of 254 individuals completed the survey. The mean age was 53.6 (SD 13.9) years, ranging from 18 to 83 years. The majority of participants identified as male (139/254, 54.7%) or female (109/254, 42.9%), with a small proportion identifying as nonbinary (2/254, 0.8%). Most respondents reported having completed secondary or higher education, with 113/254 (44.5%) holding a university or technical college degree. Regarding monthly net income, 113/254 (44.5%) reported earnings above €2500 (US $3000), while 36/254 (14.2%) earned less than €1500 (US $1800) per month. Income was not specified by 49/254 (19.3%) of participants.

Participants were geographically diverse. Over two-fifths (105/254, 41.3%) reported living in large cities (>100,000 inhabitants), while 38/254 (15%) resided in rural areas (<5000 inhabitants). Digital health literacy was measured using the eHEALS scale. Among participants, 183/254 (72%) demonstrated moderate or higher levels of digital health literacy. Specifically, 20/254 (8%) of users fell into the “Low” category, 163/254 (64%) were in the “Moderate” category, and 71/254 (28%) were classified as having “High” eHealth literacy. Overall, 216/254 (85%) of participants reported knowing how to find health-related information online, but only 90/254 (35.4%) felt confident making health decisions based on online information.

An overview of sociodemographic characteristics is provided in Table 1.

XSL•FO
RenderX

Table . Sociodemographic characteristics of adult users of *Hypertension.APP* participating in a cross-sectional online survey in Germany (N=254; January–September 2023). The table includes age, sex, education, income, place of residence, and digital health literacy.

Characteristic	Values, n (%)
Sex	
Male	139 (54.7)
Female	109 (42.9)
Nonbinary	2 (0.8)
Missing	4 (1.6)
Education level	
No degree	2 (0.8)
Secondary school certificate	27 (10.6)
Intermediate school diploma	42 (16.5)
High school diploma	56 (22)
University/technical college degree	113 (44.5)
Other degree	7 (2.8)
Missing	7 (2.8)
Income (monthly net income)	
<€1000 (US $1200)	16 (6.3)
€1000 - 1500 (US $1200-1800)	20 (7.9)
€1500 - 2500 (US $1800-3000)	51 (20.1)
>€2500 (US $3000)	113 (44.5)
Not specified	49 (19.3)
Missing	5 (2)
Size of place of residence	
Rural (<5000 inhabitants)	38 (15)
Small town (5000-20,000)	64 (25.2)
Medium-sized town (20,000-100,000)	42 (16.5)
Large city (>100,000)	105 (41.3)
Missing	5 (2)
Digital health literacy (eHEALS[a])	
Low (16 - 23.99)	26 (10.2)
Moderate (24 - 31.99)	98 (38.6)
High (32-40)	126 (49.6)
Missing	4 (1.6)
Availability of digital technologies	
Smartphone	244 (96.1)
Computer or laptop	223 (87.8)
Tablet	173 (68.1)
Landline phone	164 (64.6)
Smartwatch/fitness tracker	124 (48.8)

[a]eHEALS: eHealth literacy scale.

Clinical Characteristics

Table 2 summarizes the clinical profile of participants. The majority of respondents reported SBP values between 120 and 139 mm Hg (158/254, 62.2%), and 24.8% (63/254) reported values at or above 140 mm Hg. Diastolic blood pressure was most frequently reported between 80 and 99 mm Hg (185/254,

XSL•FO
RenderX

72.8%), with 24.4% (62/254) indicating levels of 100 mm Hg or higher. Most participants (192/254, 75.6%) were currently taking antihypertensive medication. Regarding hypertension history, 21.3% (54/254) reported a diagnosis within the past year, while 35% (89/254) had been diagnosed more than 5 years ago. Notably, 6.3% (16/254) indicated that they had never been diagnosed with high blood pressure or no longer had the condition. In terms of smoking behavior, 9.1% (23/254) of respondents were current smokers, 32.3% (82/254) were former smokers, and 57.9% (147/254) had never smoked regularly.

XSL•FO
RenderX

Table . Clinical characteristics of adult users of *Hypertension.APP* in Germany, collected as part of a cross-sectional online survey (January–September 2023). Variables include blood pressure categories, comorbidities, medication use, duration of hypertension, and smoking status.

Characteristic	Values, n (%)
Systolic blood pressure	
<120 mm Hg	31 (12.2)
120 - 129 mm Hg	67 (26.4)
130 - 139 mm Hg	91 (35.8)
140 - 159 mm Hg	54 (21.3)
160 - 179 mm Hg	9 (3.5)
>180 mm Hg	0 (0)
Missing	2 (0.8)
Diastolic blood pressure	
<80 mm Hg	3 (1.2)
80 - 84 mm Hg	68 (26.8)
85 - 89 mm Hg	57 (22.4)
90 - 99 mm Hg	60 (23.6)
100 - 109 mm Hg	56 (22)
>110 mm Hg	6 (2.4)
Missing	4 (1.6)
Antihypertensive medication use	
Yes	192 (75.6)
No	60 (23.6)
Missing	2 (0.8)
Time since hypertension diagnosis	
<1 y	54 (21.3)
1 - 5 y	93 (36.6)
5 - 10 y	29 (11.4)
>10 y	60 (23.6)
Never diagnosed or no longer hypertensive	16 (6.3)
Missing	2 (0.8)
Smoking status	
Current smoker	23 (9.1)
Former smoker	82 (32.3)
Never smoked regularly	147 (57.9)
Missing	2 (0.8)
Comorbidities	
Elevated cholesterol levels	69 (27.2)
Other chronic or serious medical conditions	48 (18.9)
Physical disability	28 (11)
History of heart attack or stroke	22 (8.66)
History of cardiac or vascular surgery	12 (4.72)
Diagnosed with diabetes mellitus	8 (3.15)
Diagnosed with chronic kidney disease	6 (2.36)
Means of transportation to go to the doctor	

XSL•FO
RenderX

Characteristic	Values, n (%)
Car	180 (70.9)
Public transportation	116 (45.7)
Bike	109 (42.9)
Other means of transport	32 (12.6)

The mean reported travel time to the doctor was 18.4 (SD 18.8) minutes, ranging from 0 to 159 minutes.

App Usage and User Experience

Figure 1 summarizes the results of the 10-item SUS for all 254 participants. *Hypertension.APP* achieved a mean SUS score of 82 (SD 11.98; range 38 - 100), indicating "excellent" usability according to established benchmarks.

Figure 1. System usability scale (SUS) ratings of *Hypertension.App* among adult users in Germany participating in a cross-sectional survey (January-September 2023). The figure displays item-level responses (left) and the distribution of total SUS scores (right). Values represent mean (range)

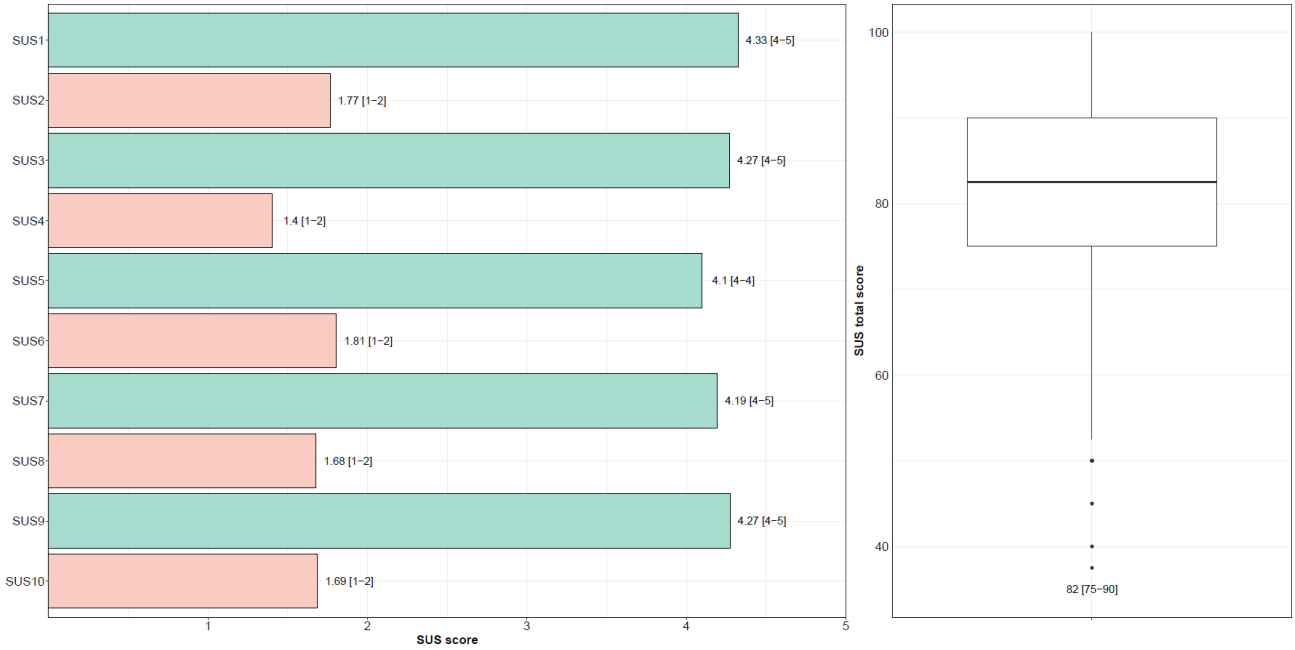

Most respondents (239/254, 94.1%) reported actively using the app at the time of the survey; 52.4% (133/254) used it daily and 80.7% (205/254) at least once per week. Table 3 provides an overview of motivations for app use, specific use cases, and the extent to which the app was integrated into medical care.

XSL•FO
RenderX

Table . App usage patterns, reasons for use, integration into medical care, and perceived usefulness among adult users of *Hypertension.APP* in Germany, based on a cross-sectional online survey conducted between January and September 2023 (N=254).

Characteristic (multiple responses possible/top 5)	Values, n (%)
Source of app awareness	
Independently searched for blood pressure apps online or in the app store	112 (44.1)
Recommendation by a doctor	45 (17.7)
Advertisement in print media	30 (11.8)
Recommendation from friends/family	27 (10.6)
Online advertisement	20 (7.9)
Learned about the app in another way	31 (12.2)
Reasons for using the app (user motivations)	
To document blood pressure	220 (86.6)
To get information about high blood pressure	107 (42.1)
To do something for one's health	76 (29.9)
Own curiosity	75 (29.5)
To try deep breathing function	34 (13.4)
App use cases (features actually used)	
Documentation blood pressure values	242 (95.3)
Personalized health information suggestions	169 (66.5)
Diary function	102 (40.2)
General health information	96 (37.8)
To create reports for doctor's visits	86 (33.9)
App integrated into medical treatment	
Using blood pressure records to prepare for doctor's visits	133 (52.4)
App not yet integrated, but planned for future use	76 (29.9)
App not currently integrated into care	51 (20.1)
Medication adjusted by physician based on patient-reported entries	30 (11.8)
Discussing advice or health information with physician	14 (5.5)
Best feature of the app	
Blood pressure documentation	230 (90.6)
Direct feedback on blood pressure values	138 (54.3)
Preparation of reports for doctor's visit	111 (43.7)
Personalized health information	98 (38.6)
Diary entries	77 (30.3)

To explore potential predictors of usage frequency, we conducted univariable ordinal logistic regression analyses (full results are provided in Multimedia Appendix 3). Of the 10 sociodemographic and clinical variables examined, 4 were associated with app usage frequency. Having a SBP of 140 mm Hg or higher was associated with less frequent app use, whereas living in a small town (5000 - 20,000 inhabitants), having received an initial hypertension diagnosis more than 1 year ago, and having a higher level of education were associated with more frequent app use.

Participants reported using the app for multiple purposes. The most common use cases were documentation of blood pressure values (242/254, 95.3%), accessing personalized health information (169/254, 66.5%), and using the diary function (102/254, 40.2%). More than half of participants (133/254,

XSL•FO
RenderX

52.4%) indicated that they used the app to prepare for doctor's visits. However, only a minority reported that the app had been formally integrated into their medical care (51/254, 20.1%) or that their physicians had adjusted medications based on app data (30/254, 11.8%). When asked about the app's most helpful features, participants most frequently cited blood pressure documentation (230/254, 90.6%), direct feedback on blood pressure values (138/254, 54.3%), and preparation of reports for medical visits (111/254, 43.7%).

Awareness of the app was most commonly gained through self-initiated searches in the app store or online (112/254, 44.1%), followed by recommendations from health care professionals (45/254, 17.7%), advertisements in print media (30/254, 11.8%), and recommendations from friends or family (27/254, 10.6%).

Discussion

Overview

This cross-sectional survey provides insight into the profiles, usage patterns, and engagement of active users of a mobile app in hypertension care. Most users were digitally literate, middle-aged or older adults with higher education levels and above-average income. While app utility ratings and engagement were generally high, especially for documentation and preparation of doctor visits, formal integration of the app into clinical care remained limited.

Who Uses the App? The Usual Suspects

Taken together, our data underscore a persistent digital divide [24]. The user base of the app largely reflects the "usual suspects" of digital health adoption: individuals with higher levels of education and income, often residing in urban or semiurban areas, and possessing moderate to high digital health literacy. These users are typically proactive in seeking out digital tools, show high engagement, and report favorable usability experiences (SUS). In contrast, individuals with lower digital competencies or limited access to technology were underrepresented, even though such groups often bear a disproportionate hypertension burden and face greater challenges in self-management [25].

This pattern aligns closely with prior research. A 2025 Dutch study demonstrated that people with a low socioeconomic position had significantly less access to and use of eHealth, whereas individuals with a higher socioeconomic position showed substantially greater utilization [26]. One key explanatory factor is eHealth literacy: highly educated individuals generally have stronger eHealth literacy and comfort with technology [26], which lowers barriers to using health apps. Conversely, individuals from disadvantaged backgrounds often face multiple obstacles, including limited skills, fewer resources, or lower confidence, and are therefore more likely to adopt new digital health tools later—or not at all [26].

Interestingly, digital health literacy was not associated with app usage frequency in our sample. At first glance, this appears to contradict previous research. However, our sample consisted exclusively of active app users, which implies a pronounced self-selection process: individuals with lower digital health literacy are less likely to adopt such tools in the first place. Among those who already use an app, variability in digital literacy is typically reduced, and literacy may no longer influence usage intensity. Moreover, eHEALS measures perceived rather than actual digital skills, which may further attenuate associations. Taken together, these factors likely explain why digital health literacy did not emerge as a significant predictor of usage frequency in this study.

Consistent with our findings, prior surveys in the United States and Europe have similarly found that health app users are typically younger or middle-aged adults with higher income and education levels [27,28]. Thus, despite the widespread availability of mHealth apps, their use remains socially patterned and disproportionately concentrated among individuals with higher socioeconomic status and digital competencies, as noted by multiple researchers [24,29]. Our study reinforces this point: access alone is not uniform, and digital interventions in hypertension care may be inadvertently catering to an already empowered subset of patients.

Notably, participants with SBP of 140 mm Hg or higher used the app less frequently, suggesting that worse current control may correlate with lower digital engagement. Interestingly, a noteworthy proportion of participants reported diastolic blood pressure values above recommended thresholds. While we did not perform subgroup analyses exploring associations between location, sociodemographic features, or other predictors and blood pressure levels, this finding merits attention. Elevated diastolic readings may reflect limited hypertension control among specific subgroups, which could affect both engagement with self-monitoring tools and clinical outcomes. Future studies should explore whether such patterns are associated with geographic, educational, or behavioral factors and assess whether app usage may contribute to improved blood pressure control over time.

While causal inferences cannot be drawn, similar paradoxical patterns have been observed elsewhere. For example, a German survey reported lower health app use among individuals who perceived themselves at high diabetes risk [30]. Even after adjusting for demographics, high perceived risk was associated with lower app usage, suggesting that individuals with greater health concerns or disease burden might rely more on traditional care or face motivational barriers to using apps. These findings highlight the need for targeted strategies to better support and engage higher-risk patients in digital self-management.

Implications for Research and Practice

Our results highlight an unmet need to make digital health technologies more inclusive and accessible. Currently, mHealth interventions often serve the "tech-savvy" but miss those with limited digital experience or lower health literacy, a pattern that can exacerbate health disparities [29]. To ensure these tools benefit a broader patient population, developers and health care providers should consider strategies to lower the barriers to entry for less digitally literate users. Recommended approaches include the following:

- Usability and language: Apps should use clear navigation and plain language. Short sentences improve

XSL•FO
RenderX

comprehension, and intuitive icons and visuals can support users who struggle with text-heavy information. Furthermore, providing content and instructions in multiple languages is crucial in diverse populations [31].

- Low-threshold onboarding and education: It should be easy for new users to get started. This could involve in-app tutorials, community workshops, or integration with primary care so that health care staff can introduce and demonstrate the app. Such onboarding can build confidence for those who are less technologically experienced. Digital navigators [32] could serve as an important bridge between patients and technology and could play a key role in this process; however, they have yet to be widely adopted in routine care.
- Collaboration with health care providers (clinical integration): Stronger links between app data and primary care workflows would benefit both patients and clinicians. Integrating the app with electronic health records and enabling data sharing with physicians would enable routine use of these data for monitoring and decision-making during clinic visits.

From a health systems perspective, clinical integration is pivotal. Although patient engagement was high, incorporation of the app into routine care by providers was low—a disconnect commonly seen with mHealth tools. Experts argue that integrating patient-facing apps into standard clinical practice is essential for digital health to reach its full potential [33]. At present, most consumer health apps operate in isolation—patients may track their data, but it often stays on their phone and is not mandatorily reviewed by their clinician. In line with this, few physicians today formally recommend or prescribe health apps to their patients, often due to limited time and lack of information, reimbursement, or guidelines [12-16]. Addressing these barriers will require collaboration among developers, health care providers, and policy makers to ensure apps meet clinical standards, fit seamlessly into workflows, and address patient needs. Ultimately, stronger collaboration and integration are necessary for mHealth to move from a personal self-management aid to a recognized element of health care delivery.

Limitations and Generalizability

To our knowledge, this is one of the first quantitative studies to systematically characterize real-world users of a hypertension self-management app in Germany. Prior work in this field has primarily relied on small qualitative samples or clinical trial settings, which often do not reflect everyday app usage. By combining detailed sociodemographic, clinical, and digital literacy data with real-world app engagement patterns, our study advances understanding of who uses hypertension apps and how these tools are integrated—or not integrated—into routine care. The finding that a digital divide persists even among active app users, as well as the inverse association between elevated SBP and frequency of app use, provides novel insights that have not been reported in previous hypertension mHealth research.

Yet, it is important to acknowledge the study's limitations. First, our data come from a cross-sectional survey based on self-reported app use, behaviors, and blood pressure values. Self-reported data are susceptible to recall bias (participants might not remember their usage accurately) and social desirability bias (participants might overstate "good" behaviors, like app use, or underreport problems). In addition, blood pressure measurements were self-reported by users and not clinically validated, which may have introduced measurement inaccuracies or reporting bias. This is a common caveat in mHealth surveys, and while we took steps to anonymize responses, some bias cannot be ruled out.

Second, the sample is not representative of all individuals with hypertension or of all app users. Due to the self-selection recruitment process, the study inherently attracted participants who were receptive to research, which might also be associated with academic backgrounds. Furthermore, the convenience sample—recruited within *Hypertension.APP*—limits representativeness for all people with hypertension or all hypertension app users.

Third, Germany's high-resource, digitally mature context may not reflect lower-resource settings. What we observed in an urban European context (high smartphone penetration, moderate-to-high digital literacy) might not directly translate to settings where access to smartphones, the internet, or health education is more limited.

A further limitation is that important sociodemographic factors known to influence mHealth adoption, such as ethnicity, migration background, and primary language, were not collected. This restricts our ability to assess disparities along these dimensions and may obscure differential patterns of digital engagement. Finally, income was measured with a coarse, self-developed scale that bunched many participants into the highest category, limiting socioeconomic resolution. A more granular income measure would have provided a clearer picture of the socioeconomic background of app users.

Conclusion

In summary, users of the hypertension self-management app in Germany were predominantly well-educated, digitally literate individuals with a history of hypertension. While the app was rated as useful and showed high engagement, particularly for tracking health data and preparing for medical appointments, its formal integration into clinical care remained limited. Furthermore, user profiles point to a persistent digital divide, suggesting that socioeconomically disadvantaged groups are underrepresented. Simply making an app available is insufficient; who uses it, how it is used, and whether it is integrated into care are critical. Future research should test strategies to engage underrepresented groups, strengthen confidence in digital health information, and link patient-facing apps more effectively with clinical care so that mHealth can be more inclusive and effective in managing hypertension.

XSL•FO
RenderX

Acknowledgments

The authors would like to deeply thank the participants of this study. For language editing and readability, the authors used ChatGPT (version 5) during the preparation of this manuscript. All content was subsequently reviewed and revised by the authors, who take full responsibility for the final version of the publication.

Funding

This project was supported by a grant of the Federal Joint Committee (Gemeinsamer Bundesausschuss, G-BA) in Germany from May 2022 to April 2025 (Innovation fund grant number 01VSF21042). This work was also partially supported by funding from the French government, managed by the National Research Agency (ANR) under the France 2030 program (ANR-23-IACL-0006) and under the “Investissements d’avenir” program (ANR-10-AIRT-0005 and ANR-15-IDEX-0002). The funders of the study had no influence in the study design.

Data Availability

The datasets generated or analyzed during this study are not publicly available as participants did not provide consent for data sharing. For further questions regarding the reuse of data, please contact the corresponding author.

Authors' Contributions

All authors were involved in drafting the article and critically revising it for important intellectual content, and all authors approved the final version to be submitted for publication.
FM had full access to all the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis.
Study conception and design: FM, SM, EW, OJ, SS, DB.
Acquisition of data: FM, DB, OJ.
Analysis and interpretation of data: FM, PP, NV.

Conflicts of Interest

OJ is CEO of *Hypertension Care UG*, which offers *Hypertension.APP*. OJ and MM are founders of *Hypertension.APP*. These authors contributed to the conception of the study and supported recruitment but were not involved in the statistical analysis or interpretation of the data. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Multimedia Appendix 1
German questionnaire.
[PDF File, 637 KB - jmir_v28i1e83075_app1.pdf]

Multimedia Appendix 2
English version of the questionnaire.
[PDF File, 649 KB - jmir_v28i1e83075_app2.pdf]

Multimedia Appendix 3
Association between app usage frequency and sociodemographic variables.
[PDF File, 210 KB - jmir_v28i1e83075_app3.pdf]

References

1. Roth GA, Mensah GA, Johnson CO, et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. J Am Coll Cardiol 2020 Dec 22;76(25):2982-3021. [doi: 10.1016/j.jacc.2020.11.010] [Medline: 33309175]
2. Forouzanfar MH, Liu P, Roth GA, et al. Global burden of hypertension and systolic blood pressure of at least 110 to 115 mm Hg, 1990-2015. JAMA 2017 Jan 10;317(2):165-182. [doi: 10.1001/jama.2016.19043] [Medline: 28097354]
3. Williams B, Mancia G, Spiering W, et al. 2018 ESC/ESH Guidelines for the management of arterial hypertension. Eur Heart J 2018 Sep 1;39(33):3021-3104. [doi: 10.1093/eurheartj/ehy339] [Medline: 30165516]
4. Egan BM, Li J, Sutherland SE, Rakotz MK, Wozniak GD. Hypertension control in the United States 2009 to 2018: factors underlying falling control rates during 2015 to 2018 across age- and race-ethnicity groups. Hypertension 2021 Sep;78(3):578-587. [doi: 10.1161/HYPERTENSIONAHA.120.16418] [Medline: 34120453]

XSL•FO
RenderX

5. Parati G, Goncalves A, Soergel D, et al. New perspectives for hypertension management: progress in methodological and technological developments. Eur J Prev Cardiol 2023 Jan 11;30(1):48-60. [doi: 10.1093/eurjpc/zwac203] [Medline: 36073370]
6. Kuwabara A, Su S, Krauss J. Utilizing digital health technologies for patient education in lifestyle medicine. Am J Lifestyle Med 2019;14(2):137-142. [doi: 10.1177/1559827619892547] [Medline: 32231478]
7. Mikulski BS, Bellei EA, Biduski D, De Marchi ACB. Mobile health applications and medication adherence of patients with hypertension: a systematic review and meta-analysis. Am J Prev Med 2022 Apr;62(4):626-634. [doi: 10.1016/j.amepre.2021.11.003] [Medline: 34963562]
8. Prendergast H, Kitsiou S, Petzel Gimbar R, et al. Emergency department–based education and mHealth empowerment intervention for hypertension: the TOUCHED randomized clinical trial. JAMA Cardiol 2025 Jul 1;10(7):657-665. [doi: 10.1001/jamacardio.2025.0675] [Medline: 40266598]
9. Labinsky H, Gupta L, Raimondo MG, Schett G, Knitza J. Real-world usage of digital health applications (DiGA) in rheumatology: results from a German patient survey. Rheumatol Int 2023 Apr;43(4):713-719. [doi: 10.1007/s00296-022-05261-7] [Medline: 36543961]
10. Boy K, May S, Labinsky H, et al. Back on track - digital health applications to treat back pain of rheumatic patients? Results of a qualitative interview study. Rheumatol Int 2024 Dec;44(12):2977-2987. [doi: 10.1007/s00296-024-05726-x] [Medline: 39340666]
11. Albrecht A, Taubmann J, Minopoulou I, et al. Real-world-evidence of digital health applications (DiGAs) in rheumatology: insights from the DiGAReal registry. Rheumatol Ther 2025 Apr;12(2):267-282. [doi: 10.1007/s40744-025-00744-y] [Medline: 39836353]
12. Peng W, Kanthawala S, Yuan S, Hussain SA. A qualitative study of user perceptions of mobile health apps. BMC Public Health 2016 Nov 14;16(1):1158. [doi: 10.1186/s12889-016-3808-0] [Medline: 27842533]
13. May S, Muehlensiepen F, Wengemuth E, et al. Benefits and barriers to mHealth in hypertension care: qualitative study with German health care professionals. JMIR Hum Factors 2025 Mar 10;12:e52544. [doi: 10.2196/52544] [Medline: 40063928]
14. May S, Seifert F, Bruch D, Heinze M, Spethmann S, Muehlensiepen F. Insights into how mHealth applications could be introduced into standard hypertension care in Germany: qualitative study with German cardiologists and general practitioners. JMIR mHealth uHealth 2025 Mar 28;13:e56666. [doi: 10.2196/56666] [Medline: 40153776]
15. Melchiorre MG, Lamura G, Barbabella F, ICARE4EU Consortium. eHealth for people with multimorbidity: results from the ICARE4EU project and insights from the “10 e’s” by Gunther Eysenbach. PLoS One 2018;13(11):e0207292. [doi: 10.1371/journal.pone.0207292] [Medline: 30427924]
16. Ardito V, Golubev G, Ciani O, Tarricone R. Evaluating barriers and facilitators to the uptake of mHealth apps in cancer care using the consolidated framework for implementation research: scoping literature review. JMIR Cancer 2023 Mar 30;9:e42092. [doi: 10.2196/42092] [Medline: 36995750]
17. Muehlensiepen F, Bruch D, Seifert F, et al. mHealth apps for hypertension self-management: interview study among patient-users. JMIR Form Res 2024 Sep 27;8:e56162. [doi: 10.2196/56162] [Medline: 39331954]
18. Bruch D, Muehlensiepen F, May S, et al. Digital Preventive Measures for Arterial Hypertension (DiPaH) - a mixed-methods study protocol for health services research. Front Cardiovasc Med 2022;9:1089968. [doi: 10.3389/fcvm.2022.1089968] [Medline: 36704456]
19. Hypertension Care UG. HypertensionAPP. 2025. URL: https://www.hypertonie.app/?lang=en [accessed 2026-01-27]
20. Soellner R, Huber S, Reder M. The concept of eHealth literacy and its measurement: German translation of the eHEALS. J Media Psychol 2014;26(1):29-38. [doi: 10.1027/1864-1105/a000104]
21. Gao M, Kortum P, Oswald FL. Multi-language toolkit for the system usability scale. Int J Hum Comput Interact 2020 Dec 13;36(20):1883-1901. [doi: 10.1080/10447318.2020.1801173]
22. Monkman H, Kushniruk AW, Barnett J, Borycki EM, Greiner LE, Sheets D. Are health literacy and eHealth literacy the same or different? Stud Health Technol Inform 2017;245:178-182. [doi: 10.3233/978-1-61499-830-3-178] [Medline: 29295077]
23. Venables WN, Ripley BD. Modern Applied Statistics with S, 4th edition: Springer; 2002. [doi: 10.1007/978-0-387-21706-2]
24. Equity within digital health technology within the WHO European Region: a scoping review. : World Health Organization Regional Office for Europe; 2022 URL: https://iris.who.int/server/api/core/bitstreams/c0b312f6-4877-4d2a-aa66-769e84625226/content [accessed 2026-01-27]
25. Leng B, Jin Y, Li G, Chen L, Jin N. Socioeconomic status and hypertension: a meta-analysis. J Hypertens 2015 Feb;33(2):221-229. [doi: 10.1097/HJH.0000000000000428] [Medline: 25479029]
26. Standaar L, van Tuyl L, Suijkerbuijk A, Brabers A, Friele R. Differences in eHealth access, use, and perceived benefit between different socioeconomic groups in the Dutch context: secondary cross-sectional study. JMIR Form Res 2025 Jan 7;9:e49585. [doi: 10.2196/49585] [Medline: 39773883]
27. Ernsting C, Dombrowski SU, Oedekoven M, et al. Using smartphones and health apps to change and manage health behaviors: a population-based survey. J Med Internet Res 2017 Apr 5;19(4):e101. [doi: 10.2196/jmir.6838] [Medline: 28381394]

XSL•FO
RenderX

28. Krebs P, Duncan DT. Health app use among US mobile phone owners: a national survey. JMIR mHealth uHealth 2015 Nov 4;3(4):e101. [doi: 10.2196/mhealth.4924] [Medline: 26537656]
29. Hengst TM, Lechner L, Dohmen D, Bolman CA. The facilitators and barriers of mHealth adoption and use among people with a low socio-economic position: a scoping review. Digit Health 2023;9:20552076231198702. [doi: 10.1177/20552076231198702] [Medline: 37691766]
30. Stühmann LM, Paprott R, Heidemann C, et al. Health app use and its correlates among individuals with and without type 2 diabetes: nationwide population-based survey. JMIR Diabetes 2020 May 20;5(2):e14396. [doi: 10.2196/14396] [Medline: 32432555]
31. eHealth literacy. Centers for Disease Control and Prevention. 2024 Jun 21. URL: https://www.cdc.gov/health-literacy/php/research-summaries/ehealth.html [accessed 2026-01-27]
32. DigiNavi.: MHB IAG Mental Health Policy & Digitization; 2025. URL: https://www.diginavi.de/en/ [accessed 2025-07-27]
33. Gordon WJ, Landman A, Zhang H, Bates DW. Beyond validation: getting health apps into clinical practice. NPJ Digit Med 2020;3:14. [doi: 10.1038/s41746-019-0212-z] [Medline: 32047860]

Abbreviations

eHEALS: eHealth Literacy Scale
mHealth: mobile health
SBP: systolic blood pressure
SUS: system usability scale

Edited by A Stone; submitted 27.Aug.2025; peer-reviewed by G Kerr, K Lakshminarayan; accepted 16.Dec.2025; published 11.Feb.2026.

Please cite as:
Muehlensiepen F, May S, Seifert F, Wengemuth E, Johannsen O, Middeke M, Heinze M, Petit P, Vuillerme N, Spethmann S, Bruch D
User Profiles and Engagement in a Hypertension Self-Management App: Cross-Sectional Survey
J Med Internet Res 2026;28:e83075
URL: https://www.jmir.org/2026/1/e83075
doi:10.2196/83075

© Felix Muehlensiepen, Susann May, Frances Seifert, Eileen Wengemuth, Olen Johannsen, Martin Middeke, Martin Heinze, Pascal Petit, Nicolas Vuillerme, Sebastian Spethmann, Dunja Bruch. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Understanding User Intent in Code-Mixed Sexual and Reproductive Health Queries in Urban India: Hierarchical Classification Approach Using Large Language Models

Sumon Kanti Dey[1], BSc; Manvi S[1], MPH; Aradhana Thapa[2], MPH; Meet Shah[3], MS; Zeel Mehta[3], MSc; Shraddha Kale Kapile[3], MSW; Tanvi Divate[3], BA; Suhani Jalota[4], PhD; Azra Ismail[1], PhD

[1]Department of Biomedical Informatics, School of Medicine, Emory University, 101 Woodruff Circle, Atlanta, GA, United States

[2]Hubert Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, GA, United States

[3]Myna Mahila Foundation, Mumbai, India

[4]Hoover Institution, Stanford University, Stanford, CA, United States

Corresponding Author:
Sumon Kanti Dey, BSc
Department of Biomedical Informatics, School of Medicine, Emory University, 101 Woodruff Circle, Atlanta, GA, United States

Abstract

Background: Sexual and reproductive health (SRH) remains a stigmatized and taboo topic globally, limiting access to reliable information. These challenges are heightened in the Global South, where linguistic and cultural diversity further complicates information access. In India (the study context), many individuals express SRH concerns in code-mixed language, such as Hinglish (code-mixed Hindi and English), and use colloquial terms. Large language models (LLMs) could help answer SRH questions, but most are trained for English and may perform poorly on code-mixed text and miss cultural nuances. Our research aims to address this gap by assessing the current state of LLMs in understanding user intent in SRH queries for a low-resource language.

Objective: We evaluate the effectiveness of proprietary, multilingual open-weight, and Indic LLMs in zero-shot settings for identifying user intent in code-mixed Hinglish SRH queries. Our goal is to assess how well LLMs assign correct labels in a 2-level hierarchical classification (topic and subtopic). We take a hierarchical approach because SRH queries are complex and context-dependent; flat labels may obscure clinically important distinctions and lead to misdirected guidance. We also characterize common error types driving misclassification.

Methods: We analyzed 4161 deidentified questions about SRH in Hinglish, collected by our partner nonprofit organization (Myna Mahila Foundation) in an underserved community in urban Mumbai. Queries were annotated into 8 topics and 40 subtopics using a hierarchical framework that captured linguistic, cultural, and contextual variation. We evaluated proprietary, multilingual open-weight, and Indic-specific LLMs in zero-shot settings. Performance was measured using hierarchical F_1 (hF_1), Exact Match, and topic- and subtopic-level accuracy.

Results: Proprietary models achieved the strongest results, with GPT-5 performing best overall (hF_1= 0.784). Among open-weight systems, Sarvam-M emerged as the top-performing Indic model (hF_1=0.757), ranking just below the top-performing proprietary model and performing comparably with Claude-3.5-Sonnet (0.745; Anthropic) as well as large multilingual systems such as Llama-3.3-70B-Instruct (0.742; Meta) and Gemma-3-27B-IT (0.739; Google). Other Indic models performed considerably lower (eg, Llama-3-Gaja-Hindi-8B [0.596; CognitiveLab], Krutrim-2-Instruct [0.558; OLA Krutrim Team], and Airavata [0.404; AI4Bharat]). Smaller multilingual open-weight models, including Mixtral-8 × 7B-Instruct (0.593), Llama-3.1-8B-Instruct (0.630), Gemma-2-9B-IT (0.657), consistently outperformed them, showing that parameter size alone does not explain performance gaps. While models generally captured broad topical intent, they frequently failed at fine-grained intent recognition, especially with euphemisms, colloquial expressions, and locally or culturally situated questions.

Conclusions: Hierarchical classification revealed persistent gaps in how LLMs handle code-mixed queries. Proprietary models performed best, but Sarvam-M shows that open-weight Indic systems can achieve performance near state-of-the-art models when supported by robust training data, cultural adaptation, and appropriate scale. These findings highlight the potential of localized, culturally aligned models to advance linguistically inclusive artificial intelligence tools and expand equitable access to SRH information in underserved populations globally.

(J Med Internet Res 2026;28:e86545) doi:10.2196/86545

KEYWORDS

sexual and reproductive health; large language models; code-mixing; Hinglish; hierarchical classification; conversational agents

XSL•FO
RenderX

Introduction

Background

Recent advancements in large language models (LLMs) present an opportunity to address significant gaps in health care information delivery. LLMs could be leveraged to simplify complex medical information, respond to patient queries, and enhance health literacy among the general population [1,2]. Despite these advancements, there remain significant disparities in the performance of these models across languages, especially in health care tasks [3]. Most LLMs are predominantly centered on the English language [4,5]. They can fail to recognize local dialects, cultural nuances, and speaking patterns, especially for non-English speaking populations that are less represented online. Health communication is further shaped by social dynamics, including gender, educational status, functional literacy, and cultural context [6-8]. Our research focuses on addressing these challenges in the context of India by evaluating the performance of LLMs in detecting user intent in the context of sexual and reproductive health (SRH).

SRH presents unique challenges for health information delivery, given that stigma, misinformation, and social barriers can restrict individuals—especially women—from seeking reliable health information and medical support [9,10]. For instance, 78% of the 15 million abortions in India take place outside medical facilities, highlighting the need for better access to reproductive health services and information [11]. Deep-rooted societal taboos also prevent open discussions about sex-related topics, exacerbating barriers to SRH awareness and services [12]. At the same time, SRH is a time-sensitive domain in which delays can have serious consequences. In India, lack of early pregnancy care leads to undetected and unmanaged conditions such as anemia, diabetes, hypertension, and infections, which are significant causes of pregnancy loss and maternal mortality [13,14].

While a multitude of online platforms exist where individuals can engage with health care professionals or even community members (eg, Reddit; Reddit, Inc), the reach of these platforms remains low in underserved communities in India. The absence of such inclusive platforms for linguistically diverse, resource-constrained communities further contributes to health inequities, leaving many women without access to critical SRH-related knowledge and services. Prior work by Wang et al [15] has shown the potential of a rule-based conversational chatbot for SRH support among young people in India (SnehAI), demonstrating strong user engagement and information-seeking behavior. We build on this work by evaluating the potential of LLMs to support the understanding of user intent in code-mixed SRH queries for such interventions.

Adding to these challenges is the extensive linguistic diversity within India, which significantly influences communication patterns, particularly in informal and colloquial settings. A notable linguistic phenomenon prevalent across India is code-mixing, the practice of blending multiple languages within a single conversation or utterance. We focus on Hinglish—a popular form of code-mixing involving Hindi and English, where individuals use the phonetic Latin script instead of the Devanagari script to write Hindi words [16]. This was often the preferred mode of typing for the population on which we focus in this study and has been documented in prior research in other Hindi-speaking populations in India [17,18]. Code-mixing remains a long-standing challenge in natural language understanding research, with several publicly available LLMs still struggling to interpret and generate code-switched text [19,20].

Another significant challenge is the inherently layered and context-dependent nature of questions that individuals frequently ask about SRH. Broader concerns related to pregnancy may branch into distinct subtopics such as antepartum emergency, postpartum pain, infertility, or abortion, reflecting the complex structure of real-world health inquiries. Traditional flat classification approaches may collapse these distinctions [21], leading to the generation of misleading responses that may fail to capture critical aspects of care and support. Our research aims to address this gap by taking a hierarchical classification approach to understanding user intent in SRH queries (Figure 1).

Figure 1. Hierarchical sexual and reproductive health (SRH) intent classification framework. This figure illustrates our approach to processing a code-mixed SRH user query (an English translation of the original Hinglish text is shown for clarity) using a large language model (LLM) to infer intent. The model maps the query to both a topic and its corresponding subtopic, producing structured intent classification. JSON: JavaScript Object Notation.

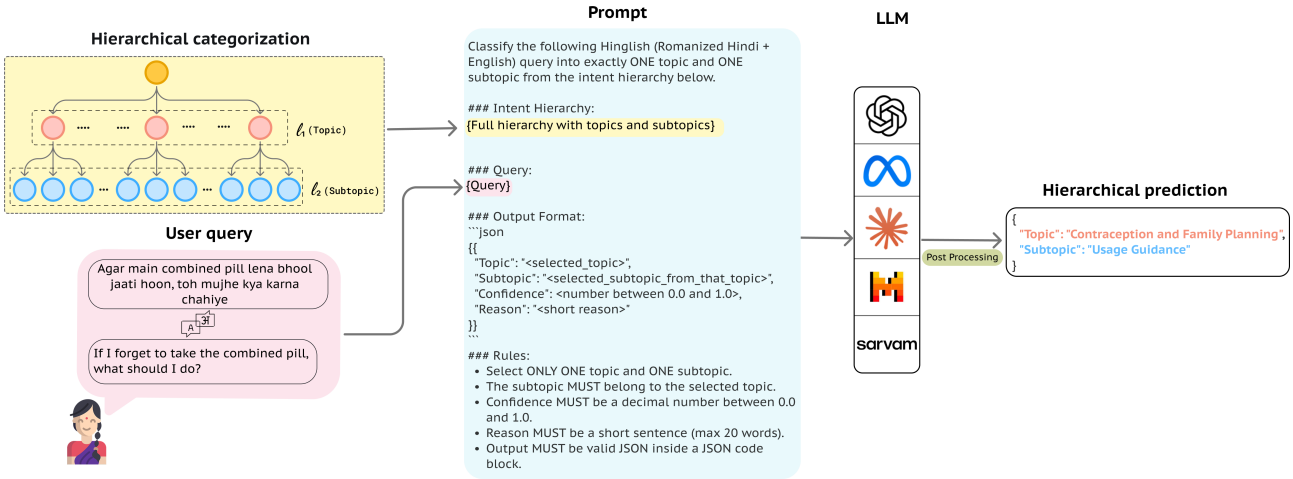

XSL•FO
RenderX

In summary, this study contributes the following: (1) we designed a hierarchical intent structure for SRH queries with 2 levels (topic and subtopic) to capture user intent in code-mixed Hinglish, providing clear label descriptions to support consistent annotation and evaluation, (2) we evaluated several state-of-the-art proprietary and open-weight multilingual LLMs (including Indic LLMs) to measure how effectively they handle hierarchical intent classification of SRH-related queries expressed in Hinglish. This evaluation also provides insight into how well LLMs can handle code-mixed text, ensuring broader applicability in the real-world health care contexts, and (3) we conducted a qualitative error analysis with thematic coding across models to examine misclassifications in code-mixed SRH queries. We explain why these errors occur and how models misread culturally sensitive terms and euphemisms. Representative annotated examples from the dataset are provided in Table 1.

Table . Sample annotated user queries from the dataset. The examples below present both transliteration and code-mixing between Hindi and English.

Query	English translation	Topic	Subtopic
Periods agar monthly na aaye to kya karen?	What should I do if my periods are not regular every month?	Menstrual Health	Menstrual Cycle Information
Gabhnirodhak goliyan lene se pehle mujhe doctor se kya puchna chahiye?	What should I ask the doctor before taking birth control pills?	Contraception and Family Planning	Usage Guidance
Agar PCOD[a] ka 3 mahina dawa chalane ke baad bhi shi nahi hua to kya kare?	If PCOD[a] does not improve after three months of medication, what should I do?	PCOS[b] or PCOD[a]	Management
Agar abhi bacha nahi rakhna hai aur fir bhi pregnant ho gaye to abortion ke liye kya kare?	If I do not want to have a baby right now but still get pregnant, what should I do for an abortion?	Pregnancy and PNC[c]	Abortion

[a]PCOD: polycystic ovarian disease.

[b]PCOS: polycystic ovary syndrome.

[c]PNC: postnatal care.

Related Work

Hierarchical classification is a method of assigning items to categories organized within a hierarchy of classes [22,23]. This approach has been broadly used in a range of domains where layered structures are important, including e-commerce user query categorization [24,25], health care question answering [26,27], clinical guideline classification [21], and personalized health care analysis of women's menstrual health disorders [28]. Its advantages lie in its ability to model the inherent complexity of user intent, particularly in health care contexts where individuals begin their health care concerns by first referencing broader domains and then subsequently focusing on specific details or contexts. While flat classification systems are widely used as classification baselines, their lack of hierarchical awareness limits their effectiveness, especially in domains where understanding class structure is essential [21,29]. Although hierarchical classification has demonstrated clear benefits across diverse applications, its potential remains underexplored in the context of code-mixed languages and SRH, where understanding layered and implicit user intent is especially important.

The recent development of highly parameterized language models, such as GPT-5 (OpenAI) [30], GPT-4o (OpenAI) [31], Llama 3 (Meta) [32], Gemma 3 (Google) [33], among many others, has significantly enhanced health care reasoning, understanding, and summarization [34]. These models can generate concise and user-friendly summaries and have the potential to make medical information more accessible to the general public. Despite these developments, a major limitation remains their reliance on training data in English, making them less effective for regional languages [12,35]. For instance, Llama 3 sources 90% of its pretraining data from English sources [4], while a substantial portion of GPT-4's pretraining data is similarly English-dominant [5]. This linguistic bias creates information disparities across languages, where equivalent questions in different languages may produce inconsistent and inaccurate responses [36]. Our study seeks to address this gap.

Recent research has increasingly focused on using chatbots to provide health care assistance, particularly around SRH. For example, Wang et al [15] designed an SRH-focused rule-based chatbot on Facebook Messenger (Meta Platforms, Inc) to analyze how users engage with artificial intelligence (AI) to seek health information. Their study found that users frequently shared personal concerns and SRH-related queries in code-mixed languages, highlighting the need for linguistically adaptable health care AI models. There are several reasons why LLMs can be useful for SRH information delivery, including gaps in sexual education [37] and societal stigma and taboo around sex-related topics [12]. However, recent studies indicate that LLMs struggle with Indic languages, particularly when queries are code-mixed and culturally situated [38]. To analyze these issues systematically, we collected a dataset called sexual and reproductive health queries (SRHQ)-India, which captures real-world Indian SRH user queries in a code-mixed format.

As queries in SRH domains are often linguistically and contextually diverse, the hierarchical categorization approach offers the potential for more accurate and interpretable results. Recent advancements in LLMs have made significant improvements in flat text classification, especially in the health care domain. In our work, we seek to evaluate LLMs' understanding of the hierarchical classification task. In this

XSL•FO
RenderX

regard, we explore the use of a zero-shot learning technique that involves crafting prompts to enable language models to generate useful responses without prior examples, relying entirely on pretrained knowledge to tackle the new task. This technique has been used in several health care applications, such as capturing the context of clinical text [39] and classifying health care queries [40]. In our work, we leverage zero-shot learning to evaluate the efficiency of language models in processing code-mixed user queries.

Methods

Overview

In this study, we used a dataset gathered by the Myna Mahila Foundation, a nonprofit women's health organization in India. The data were collected through a preliminary SRH chatbot prototype that was developed and piloted with 488 women in informal settlements of Mumbai, India, from October 2023 to December 2024 to capture real-world user queries on SRH, including both text and voice input. This is part of their ongoing effort to improve access to SRH information for women and girls in urban slums. Below in "Dataset Generation and Curation" and "Data Annotation" subheadings, we detail our approach to structuring and annotating the dataset. We also outline our evaluation framework for model performance.

Dataset Generation and Curation

The dataset originally contained 4858 queries and was refined to reduce redundancy using cosine similarity computed over term frequency–inverse document frequency representations of the text. Queries were first normalized using standard text preprocessing, including Unicode correction, lowercasing, whitespace normalization, and removal of basic punctuation. Cosine similarity was then used to identify syntactically similar queries, and pairs with a similarity score of 90% or higher were considered near duplicates, and only one was retained. For example, "kya saheli tablet lene se periods ka date badal jata hai?" or "saheli tablet se periods ka date badal jata hai kya" convey the same meaning: "Does taking the Saheli tablet (a nonhormonal contraceptive pill) change the period date?" Term frequency–inverse document frequency–based lexical similarity was chosen instead of pretrained sentence embeddings because many widely used embedding models are trained predominantly on monolingual or English-centric data and perform poorly on code-mixed Indian languages [41,42]. Moreover, our aim was to remove lexically rephrased queries rather than infer semantic equivalence. In addition, we removed queries containing fewer than three words to exclude vague or incomplete inputs. We also filtered out greeting-based queries, such as "Namaste" or "Aap kaise hai," which means "How are you," which do not contribute to intent classification.

After filtering, we obtained a final dataset of 4161 queries, primarily in Hinglish. The dataset consists of queries with misspelled English words, transliterated Hindi terms, and borrowed English medical words that are common in everyday SRH talk (eg, periods, condom, pregnancy, in-vitro fertilization [IVF], and polycystic ovary syndrome). We also observed queries related to cultural practices or religious beliefs. For example, "Masik pali aane se Bhagvan ke pas kau nahi jana chahiye?" which in English translates to "Why should one avoid going to god or temples during periods?" This query centers around the religious norm of how women are advised not to visit the temple during their periods in certain cultures and communities in India. Additionally, the query "Det ke pahele piryat q aata hai" translates to "Why periods come before the date?" where det (date) and piryat (period) are misspelled terms. Such variations highlight the real-world linguistic challenges in SRH-related conversations and underscore the need for models capable of accurately processing noisy, code-mixed, and culturally or religiously influenced queries.

Beyond linguistic complexities, the dataset also captures deeply ingrained cultural myths and misinformation, particularly related to pregnancy and gender beliefs. For instance, the query "Main pregnant hu or jinko sirf ladkiya hai kya unko dekte rahne se kya muje ladki hogi?" translates to "I am pregnant, if I keep seeing women who have only given birth to daughters, will I also have a daughter?" Similarly, "Kya muje sirf ladka hi chahiye to uske liye khuch upay hai kya?" means "Is there any way to have only a boy?" These queries reflect long-standing cultural expectations in some communities in India, where socioeconomic structures reinforce a preference for sons. Such myths not only influence reproductive decisions but also contribute to gender-based discrimination and misinformation.

We adopted a hierarchical thematic structure for annotation. Queries were first classified into broad topics and then further refined into subcategories to capture the complexity of SRH-related concerns. For example, questions related to postpartum, antepartum, and abortion fall under the broad category of Pregnancy and Postnatal Care (PNC) category, while questions related to the Menstrual Health category are further classified based on their context to subtopics, including Menstrual Cycle Information, Menstrual Flow, or Period Pain Management. We have also included Mental Health and Wellness, with subtopics including Stress Management and Safety Concerns, because stress, sleep, mood, and anxiety issues often co-occur with SRH questions and influence help-seeking and guidance needs. Placing these queries within the SRH hierarchy keeps related concerns together and improves intent interpretation. Table 2 provides the annotation framework in detail. The annotation guidelines were developed through an iterative process of feedback from doctors and program staff at the Myna Mahila Foundation. This helped standardize interpretations and improve annotation reliability across the dataset. Key descriptive statistics of the dataset are provided in Table 3. In addition, we quantify the degree of intrasentential Hindi-English code-mixing using the Code-Mixing Index (CMI); dataset-level CMI statistics are reported in Table 3, and the metric is described in the following "CMI Calculation" subheading.

XSL•FO
RenderX

Table . Annotated dataset distribution by topic and subtopic.

Topic and subtopic	Queries (N=4161), n (%)
Contraception and Family Planning	1417 (34.1)
Family Planning Queries	522 (36.8)
Usage Guidance	291 (20.5)
Types of Contraceptives	209 (14.7)
Sterilization	184 (13)
Effectiveness and Duration	127 (9)
Side Effects	84 (5.9)
Menstrual Health	1104 (26.5)
Menstrual Cycle Information	673 (61)
Period Pain Management	325 (29.4)
Sanitary Products and Hygiene	62 (5.6)
Menstrual Flow	44 (4)
Pregnancy and PNC[a]	547 (13.1)
Pregnancy Information	290 (53)
Antepartum	120 (21.9)
Infertility	53 (9.7)
Abortion	26 (4.8)
Postpartum	23 (4.2)
Miscarriage	20 (3.7)
Breastfeeding	15 (2.7)
Sexual and Vaginal Health	322 (7.7)
Sex-Related Queries	128 (39.8)
Vaginal Health and Discharge	127 (39.4)
Reproductive Anatomy	26 (8.1)
Urinary Tract Infections (UTI)	16 (5)
Sexually Transmitted Infections (STI or STD)	15 (4.7)
Vaginal or Uterine Infections	10 (3.1)
PCOS or PCOD[bc]	101 (2.4)
Information	49 (48.5)
Management	28 (27.7)
Symptoms	24 (23.8)
HIV[d]	52 (1.2)
Stigma and Awareness	14 (26.9)
Treatment	14 (26.9)
Prevention	12 (23.1)
Symptoms and Early Detection	12 (23.1)
Mental Health and Wellness	22 (0.5)
Stress Management	13 (59.1)
Information and Safety Concerns	9 (40.9)
Other	596 (14.3)
General Health Queries	361 (60.6)

XSL•FO
RenderX

Topic and subtopic	Queries (N=4161), n (%)
Diet and Nutrition	97 (16.3)
Exercise and Fitness	46 (7.7)
Health Equity and Access	38 (6.4)
Marriage and Relationships	21 (3.5)
Misconceptions and Myths	13 (2.2)
Child Health	12 (2)
Cultural, Religious, or Moral Norms	8 (1.3)

[a]PNC: postnatal care.

[b]PCOS: polycystic ovary syndrome.

[c]PCOD: polycystic ovarian disease.

[d]HIV: human immunodeficiency virus.

Table . Statistics of the dataset.

Statistic	Values
Total number of topics	8
Total number of subtopics	40
Total number of queries	4161
Maximum question length (words)	71 words
Average question length (words)	10.56 words
% queries in transliterated Hindi[a]	61.53
% queries in intrasentential mixing[b]	38.37
Code-Mixing Index (CMI), %, mean (SD)	37.38 (10.43)

[a]Transliterated Hindi refers to text written in Hindi using the Latin script, often incorporating English loan words common in sexual and reproductive health (SRH) contexts (eg, periods, condom, and pregnancy).

[b]Intrasentential mixing refers to text where Hindi and English elements are combined within the same sentence, written in Latin script.

Our approach to annotation serves 2 purposes. First, hierarchical classification offers a deeper perspective on SRH concerns, allowing the identification of common themes and trends in user queries. It strengthens decision support for stakeholders, including health care professionals, policymakers, and organizations, by offering a well-organized dataset that can inform better resource allocation and intervention strategies. Second, hierarchical classification can enhance user guidance and response accuracy based on user intent, ensuring that individuals are directed to relevant and specific information tailored to their concerns.

CMI Calculation

The CMI [43] is an utterance-level, ratio-based metric that quantifies intrasentential code-mixing by measuring the proportion of lexical tokens that do not belong to the matrix (dominant) language.

CMI is defined as:

$$CMI=100\times(1-\max(wi)n-u), \text{if } n>u$$

Where n is the total number of tokens, u is the number of language-independent tokens, n-u is the sum of the number of tokens from N languages, and max {wi}is the highest number of words belonging to a particular language. If an utterance contains only language-independent tokens (n=u), the CMI is defined as zero.

We used the pretrained FastText (Meta AI) language identification model, which supports 176 languages [44], to automatically assign language labels at the token level. FastText has been widely used for language identification in multilingual and code-mixed text due to its efficiency and robustness on short lexical units, and its use of subword information [45,46]. Most queries in the dataset are bilingual in nature, with one dominant language, either English or Romanized Hindi, serving as the matrix language within each utterance. We used the model to identify English tokens (and assumed the remaining tokens to be Hindi) due to potential concerns about its performance on Hindi words. Numerals, punctuation, and symbols were treated as language-independent. This automated procedure enabled consistent computation of sentence-level CMI scores across all 4161 queries.

Data Annotation

We annotated the user queries into 8 categories, which were further divided into subcategories to capture the depth and complexity of the topics. The annotation of our code-mixed dataset on SRH required both strong linguistic competence and domain expertise. Since the data are code-mixed (Hindi and

XSL•FO
RenderX

English), they present unique represents unique challenges such as nonstandard spelling, switching between languages in the same query, and the cultural nuances embedded in users' expressions. To meet these complexities, this study included 2 experienced annotators with backgrounds in public health and global health. Both annotators are native or fluent speakers of Indian languages and coauthors of this paper, ensuring familiarity with cultural and linguistic contexts relevant to SRH queries. Both were guided by a comprehensive annotation manual, refined by subject matter experts (public health care professionals and medical doctors). Some user questions could overlap with multiple topics or subtopics (for instance, a question about convincing a partner to use contraceptives could fall under the "Family Planning" or "Marriage and Relationship" category). In such cases, annotators selected the closest-fitting category. The full annotation guidelines are provided in Multimedia Appendix 1.

To ensure annotation quality and consistency, we adopted an iterative annotation process. Initially, 10% of the dataset was annotated independently by both annotators. This step was particularly critical, as it allowed us to refine annotation guidelines and address challenges arising from orthographic variations, code-mixing, and semantic ambiguities. Discrepancies were resolved through discussion and led to refinements in the annotation guidelines. We continued the iterative co-annotation process until the annotators consistently achieved a 95% agreement threshold during the pilot phase. Once the threshold was met, the remaining data were divided between the 2 annotators for independent annotation. We used Cohen kappa [47] across overlapping annotated samples to compute interannotator agreement. The resulting score of 83% indicated substantial agreement, reflecting almost perfect agreement [48]. The distribution of the query categories and subcategories is provided in Table 2.

Evaluation Metrics

To evaluate model performance on our hierarchical classification task, we adopted the Hierarchical F_1-score (hF_1) proposed by Kosmopoulos et al [23]. It extends traditional metrics by explicitly considering the hierarchical relationship among labels. To compute this metric, we augment both the predicted labels (Y^) and the ground truth labels (Y) by incorporating their ancestor labels from the hierarchy, resulting in augmented sets Y^aug*and*Yaug. This augmentation ensures that the evaluation accurately reflects hierarchical dependencies between topics and subtopics.

The hierarchical F_1-score (hF_1) is defined as:

$$hF_1 = \frac{2 \cdot hP \cdot hR}{hP + hR}, \text{where } hP = \frac{|\hat{Y}_{aug} \cap Y_{aug}|}{|\hat{Y}_{aug}|}, hR = \frac{|\hat{Y}_{aug} \cap Y_{aug}|}{|Y_{aug}|}$$

Here, hPr and hRe (hierarchical precision and recall) evaluate the proportion of correctly predicted hierarchical labels among all predicted and actual labels, respectively. Additionally, we used supporting metrics, including Exact Match, which evaluates the percentage of queries where both the topic and subtopic predictions exactly match the ground truth labels. Accuracy@l_1 measures the topic-level accuracy, reflecting the proportion of samples where the predicted topic is correct. Accuracy@l_2 measures conditional subtopic accuracy, defined as the proportion of correct subtopic predictions given that the topic was correctly predicted. These complementary metrics collectively provide a comprehensive evaluation of the model's performance, capturing strict end-to-end correctness (Exact Match), coarse-grained topic identification (Accuracy@l_1), fine-grained subtopic discrimination (Accuracy@l_2), and hierarchical sensitivity (hF_1).

Experimental Analysis

In this study, we evaluate the performance of language models on a hierarchical classification task in a zero-shot setting, where the model attempts to generalize output patterns at the topic and subtopic levels for data it has not encountered during training. Zero-shot evaluation is particularly challenging because it requires models to infer patterns without explicit task-specific fine-tuning.

In our setup, models were provided with the hierarchical structure and the list of possible topic and subtopic labels, but not with the detailed semantic definitions used in the human annotation guideline. This was a deliberate design choice to evaluate the models' intrinsic ability to interpret real-world, code-mixed SRH queries without additional semantic instruction. It also reflects practical deployment scenarios where the model must generalize from raw user input rather than rely on curated taxonomies at the time of inference.

Our evaluation includes 7 open-weight multilingual models, 5 Indic-specific models, and 3 proprietary models. The open-weight multilingual models are Mixtral-8 × 7B-Instruct (Mistral AI) [49], Llama-3.1-8B-Instruct (Meta) [32], Llama-3.3-70B-Instruct (Meta), Gemma-2-9B-IT (Google) [50], Gemma-3-27B-IT (Google) [33], Qwen-2.5-7B-Instruct (Alibaba Group) [51], and Aya-Expanse-8B (Cohere Labs) [52]. The technical reports of these models demonstrate their multilingual capabilities on benchmark datasets, making them strong candidates for handling code-mixed data.

The Indic models are Airavata [53], an instruction-tuned Hindi language model developed by AI4Bharat; AryaBhatta-GemmaGenZ-Vikas-Merged [54] (referred to as AryaBhatta [GenVR Research]), a model trained on 9 Indic languages that excels particularly in Hindi reasoning and literature tasks; and Llama-3-Gaja-Hindi-8B [55], a bilingual Hindi-English LLM specialized in Indic language understanding. We also evaluate Krutrim-2-Instruct [56], a Mistral-NeMo (Mistral AI and NVIDIA)–based model trained on diverse domains, including Indic languages, and fine-tuned with direct preference optimization to improve alignment and reasoning for Indian contexts. For proprietary baselines, we include GPT-5 [30], GPT-4o [31], and Claude-3.5-Sonnet [57], which serve as state-of-the-art performance references despite their closed parameter counts.

For open-weight multilingual models, we include a range spanning smaller (7B-9B) to larger (27B-70B) architectures. Alongside these, we also include Sarvam-M (Sarvam AI) [58], a state-of-the-art open-source hybrid Indic LLM built on Mistral-Small, designed to enhance reasoning in Indian languages through supervised fine-tuning and reward-based reinforcement learning. Since Indic LLMs are still emerging,

XSL•FO
RenderX

our evaluation also includes small to mid-sized versions currently available (7B-12B). This selection allows for a balanced comparison across Indic, multilingual, and proprietary systems, highlighting the relative strengths and limitations of each category.

With LLMs, a precise and well-structured prompt is paramount for shaping responses and ensuring that the model focuses on the most relevant aspects of the input. Recent research [59] has shown that even minor lexical variation—sometimes just a single-word difference—in prompts can significantly impact model performance on downstream tasks. To accurately determine topic and subtopic levels in our task, we experimented with multiple prompt variations on a small, representative subset of our dataset. This iterative process allowed us to refine the prompts efficiently before running the large-scale evaluation. Smaller open-weight models were evaluated locally on a graphics processing unit with 48GB of memory (eg, NVIDIA RTX A6000), while proprietary models were accessed via application programming interface–based inference. The final prompt template used across all models is shown in Table S1 in Multimedia Appendix 2. The prompt template also requested a self-reported confidence score (0.0 - 1.0); however, this value was extracted for future analysis but was not used in this evaluation.

Qualitative Error Analysis

To better understand the strengths and limitations of the models, we conducted a qualitative error analysis of the 2 best-performing models in each family (proprietary, multilingual open-weight, and Indic models). For each model, we sampled 50 misclassified test queries stratified by topic and subtopic. The same 2 annotators who created the dataset served as reviewers and independently coded thematic errors; disagreements were resolved by consensus. Results are summarized in the “Error Analysis” section, with examples in Table 4.

XSL•FO
RenderX

Table . Performance of 6 models on representative sexual and reproductive health queries (SRHQ) in the SRHQ-India dataset queries at the topic-subtopic level.

Error types	Query (Hinglish→English)	Ground truth (Topic→Subtopic)	GPT-5	Claude-3.5-Sonnet	Llama-3.3-70B-Instruct	Gemma-3-27B-IT	Sarvam-M	Llama-3-Gaja-Hindi-8B
C1	Safaiya kaise karwate hain? (How is an abortion done?)	Pregnancy and [a]PNC→Abortion	✓[b]	✓	—[c]	—	—	—
C2	Pregnancy me 5 month me vomiting hoti h to kya kre (What should I do if I am vomiting in the 5th month of pregnancy?)	Pregnancy and PNC→Antepartum	✓	✓	—	—	✓	—
C3	Family planning Muslim community me accept hain kya? (Is family planning accepted in the Muslim community?)	Contraception and Family Planning→Family Planning Queries	—	—	✓	✓	—	✓
C3	Masik pali aane se Bhagvan ke pas kau nahi jana chahiye? (Why should one avoid going to god/temples during periods?)	Other→Cultural, Religious, or Moral Norms	✓	✓	—	—	✓	—
C4	1 sal se bacha rukhne ke liye try kar rahe hai lekin nahi rukh raha hai to kya karna padega? (We have been trying to have a child for one year, but it has not happened. What should we do?)	Pregnancy and PNC→Infertility	✓	✓	—	—	—	—

XSL•FO
RenderX

Error types	Query (Hinglish→English)	Ground truth (Topic→Subtopic)	GPT-5	Claude-3.5-Sonnet	Llama-3.3-70B-Instruct	Gemma-3-27B-IT	Sarvam-M	Llama-3-Gaja-Hindi-8B
C4	Family planning may agar koi mahila test tube karvati hai to kya hota hai? (In family planning, if a woman undergoes test-tube treatment, what happens?)	Pregnancy and PNC→Infertility	—	—	—	—	—	—
C5	Kitni der jinda rahte hain sperm? (How long do sperm stay alive?)	Sexual and Vaginal Health→Reproductive Anatomy	—	✓	—	—	—	—

[a]PNC: postnatal care.

[b]✓ indicates correct topic-subtopic classification.

[c]— indicates incorrect topic-subtopic classification

Ethical Considerations

We used deidentified secondary data shared by the Myna Mahila Foundation. Since the dataset was fully deidentified before being shared, no personally identifiable information was accessible to the research team at Emory University. This study did not involve direct interaction with human participants, and no additional data collection was conducted. Approval for the use of this secondary dataset was obtained from Emory University's Institutional Review Board (Protocol #2025P011010). Data security measures were implemented to maintain confidentiality, and the dataset was stored on a secure, password-protected system. Only authorized personnel had access to the data. As the study relies on pre-existing deidentified data, the risk to individuals is minimal.

Results

Overview

We now present the results of the hierarchical classification of SRHQ with LLMs. Table 5 compares model performance on the hierarchical classification across 3 types of models: open-weight, Indic (LLMs fine-tuned for Indian languages), and proprietary. Figure 2 provides a visual comparison across 4 evaluation metrics: hierarchical F_1(hF_1), Exact Match, and accuracy at the topic (l_1) and subtopic (l_2) levels. Although we also evaluated GPT-4o, its performance was less than 1.5% lower than GPT-5 across metrics. Bootstrap-based 95% CIs further show substantial overlap between GPT-5 (hF_1=0.784, 95% CI 0.774 - 0.795) and GPT-4o (hF_1=0.779, 95% CI 0.768 - 0.789), indicating that performance differences between the 2 proprietary systems are not statistically distinguishable (Table S2 in Multimedia Appendix 2). For consistency and balanced cross-category comparisons, we focus here on GPT-5 and Claude-3.5-Sonnet as representative proprietary models.

XSL•FO
RenderX

Table . Performance comparison of different models in a zero-shot setting on the sexual and reproductive health queries (SRHQ) dataset. Open-weight, proprietary, and Indic models are distinctly highlighted for clarity. All Indic models are open weights. In each category, the best-performing model is highlighted in **bold**, and the second-best is underlined.

Category and model	#Params	hF_1 (hierarchical $F1$)	Exact Match	Accuracy@l_1	Accuracy@l_2
Open-weight models					
Mixtral-8 × 7B-Instruct	7B	0.593	0.453	0.733	0.617
Llama-3.1-8B-Instruct	8B	0.630	0.491	0.769	0.638
Qwen-2.5-7b-Instruct	7B	0.605	0.463	0.747	0.619
Aya-Expanse-8B	8B	0.528	0.411	0.646	0.636
Gemma-2-9B-IT	9B	0.657	0.544	0.770	0.706
Gemma-3-27B-IT	27B	0.739	0.629	0.849	0.741
Llama-3.3-70B-Instruct	70B	**0.742**	**0.630**	**0.853**	**0.738**
Indic models					
Airavata	7B	0.404	0.226	0.581	0.389
Llama-3-Gaja-Hindi-8B	8B	0.596	0.452	0.740	0.610
AryaBhatta	8.5B	0.365	0.157	0.574	0.273
Krutrim-2-Instruct	12B	0.558	0.386	0.731	0.527
Sarvam-M	24B	**0.757**	**0.647**	**0.867**	**0.747**
Proprietary models					
GPT-5	—[a]	**0.784**	**0.683**	**0.886**	**0.771**
GPT-4o[b]	—	0.779	0.675	0.882	0.764
Claude-3.5-Sonnet	—	0.745	0.639	0.851	0.751

[a]Not available.

[b]Although GPT-4o was the second-best-performing proprietary model, its performance differed from GPT-5 by less than 1.5% across metrics. Therefore, GPT-5 and Claude-3.5-Sonnet were selected as representative proprietary models for balanced cross-category comparisons in the main analysis.

XSL•FO
RenderX

Figure 2. Visual comparison of model performance on the sexual and reproductive health queries (SRHQ) dataset across 4 evaluation metrics: hierarchical F_1(hF_1), Exact Match, Accuracy@l_1, and Accuracy@l_2. The figure highlights relative strengths and differences among open-weight, Indic, and proprietary models.

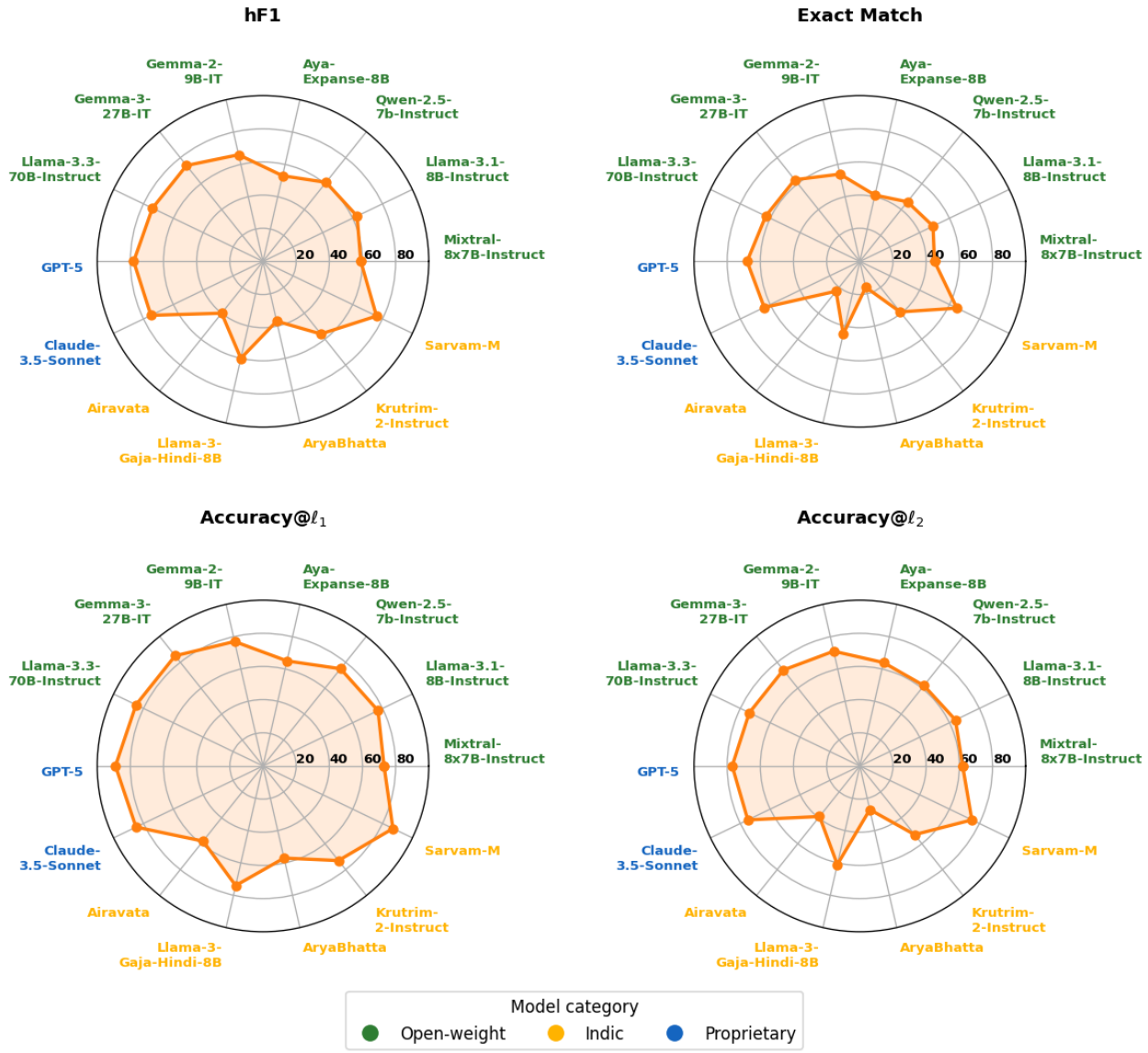

Proprietary systems delivered the highest and most consistent performance, with GPT-5 achieving the best results across all metrics, with hF_1=0.784, Exact Match=0.683, and accuracies of 0.886 at l_1 and 0.771 at l_2. Claude-3.5-Sonnet closely followed (hF_1=0.745 and Exact Match=0.639), exceeding 0.85 topic-level accuracy and 0.75 subtopic-level accuracy.

Among open-weight models, the strongest performers were Llama-3.3-70B-Instruct and Gemma-3-27B-IT, which achieved performance close to the proprietary models, with hF_1 scores of 0.742 and 0.739 and Exact Match scores of 0.630 and 0.629, respectively. Both models exceeded 0.84 in topic-level accuracy (l_1) and maintained strong subtopic accuracy ($l_2 \approx 0.74$) with scores of 0.738 and 0.741. Mid-sized models such as Gemma-2-9B-IT also performed competitively (hF_1=0.657, Exact Match=0.544, l_1=0.770, and l_2=0.706), whereas smaller models such as Aya-Expanse-8B underperformed across metrics (hF_1=0.528, Exact Match=0.411) and struggled particularly at the subtopic level (l_2=0.636). Figure 2 illustrates this scaling effect clearly, with larger multilingual models covering a broader area across hF_1, Exact Match, and both accuracy levels compared to their smaller counterparts.

Indic models showed the widest variability, with Sarvam-M emerging as a strong outlier, achieving hF_1=0.757 and Exact Match=0.647. Compared to GPT-5, Sarvam-M fell short by only 3.4% in hF_1. At the topic level, it reached an accuracy of l_1=0.867, and at the subtopic level l_2=0.747, only slightly below the top-performing model GPT-5 (0.886 and 0.771, respectively). This indicates that Sarvam-M is the strongest free and open-weight alternative to proprietary models, considering the restricted access and cost associated with proprietary models. Other Indic models performed substantially lower; the second-highest-performing Indic model was Llama-3-Gaja-Hindi-8B, achieving an hF_1 of 0.596, an Exact Match of 0.452, and accuracies of 0.740 at l_1 and 0.610 at l_2. Interestingly, it outperformed the larger Krutrim-2-Instruct

XSL•FO
RenderX

(12B; OLA Krutrim Team) (hF_1=0.558, Exact Match=0.386, l_1=0.731, and l_2=0.527), while Airavata demonstrated lower performance across metrics (hF_1=0.404, Exact Match=0.226, l_1=0.581, and l_2=0.389). These comparisons indicate that parameter count alone does not guarantee superior performance; factors such as model design and training data quality appear to play a more decisive role. By contrast, Airavata and AryaBhatta consistently trailed across all metrics, clustering at the lowest end of performance in both topic- and subtopic-level accuracy.

The results highlight a consistent performance hierarchy across model categories. Proprietary systems dominated overall, while large open-weight multilingual models narrowed the gap across all metrics, and Indic models demonstrated both promise for local contexts and limitations. Sarvam-M stood out as a competitive alternative, while other Indic models struggled to identify user intent in code-mixed settings. The combination of Table 5 and Figure 2 underscores not only differences in overall performance but also the critical difficulty of achieving accurate subtopic-level predictions in mixed-language SRH queries.

To further examine performance differences, Figures 3 and 4 present F_1 scores across SRHQ-India topics. Figure 3 highlights per-topic F_1 scores alongside model-wise averages, where the reported average F_1 is the macro-average of flat, topic-level F_1 scores computed across all topics for each model. Proprietary model GPT-5 achieved the highest overall average F_1 score (0.85), while Sarvam-M emerged as the strongest open-weight Indic alternative with an average of 0.82, with performance comparable to Claude-3.5-Sonnet (0.79) and other open-weight models such as Llama-3.3-70B (0.78) and Gemma-3-27B (0.79). Other Indic models, such as Llama-3-Gaja-Hindi-8B (CognitiveLab), showed weaker averages and wider variability, particularly in nuanced domains.

Figure 3. Model-wise topic-level F_1 scores across the sexual and reproductive health queries (SRHQ)-India dataset. Each bar represents the flat F_1 score for a specific topic, while the dotted lines indicate the macro-average of topic-level F_1 scores across all topics for each model. Results highlight differences between proprietary, open-weight, and Indic models. HIV: human immunodeficiency virus; PCOD: polycystic ovarian disease; PCOS: polycystic ovary syndrome; PNC: postnatal care.

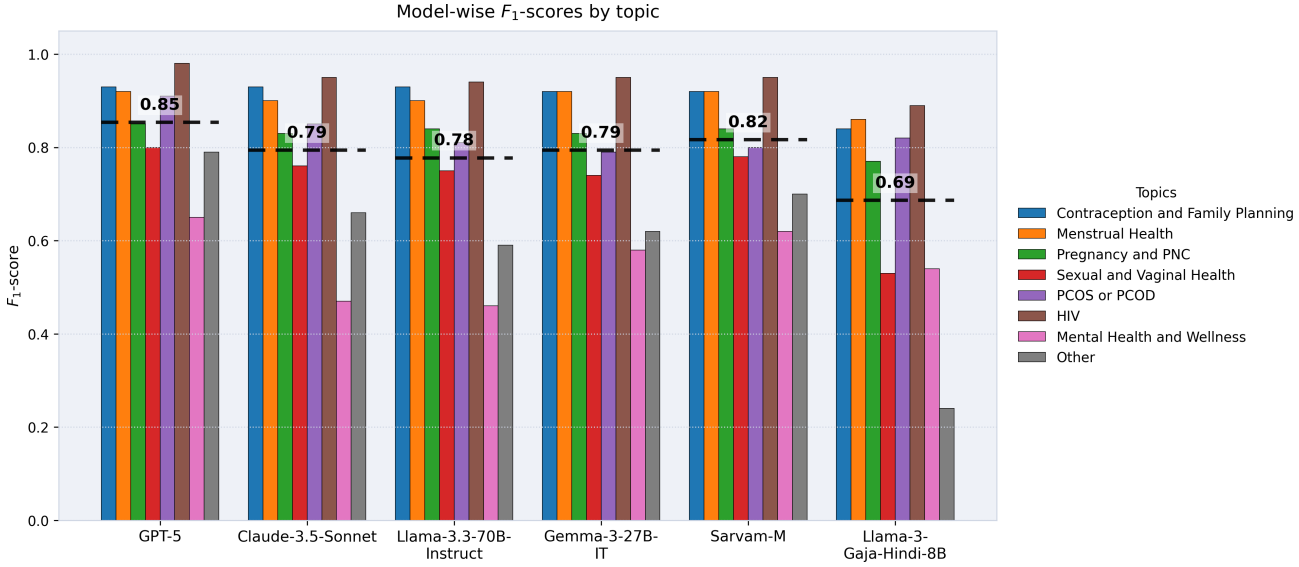

XSL•FO
RenderX

Figure 4. Topic-level F_1 scores with 95% CIs across sexual and reproductive health queries (SRHQ)-India. Each bar represents the F_1 score for a specific topic, with error bars indicating 95% CIs estimated via bootstrap resampling. Results highlight performance differences across proprietary, open-weight, and Indic models. HIV: human immunodeficiency virus; PCOD: polycystic ovarian disease; PCOS: polycystic ovary syndrome; PNC: postnatal care.

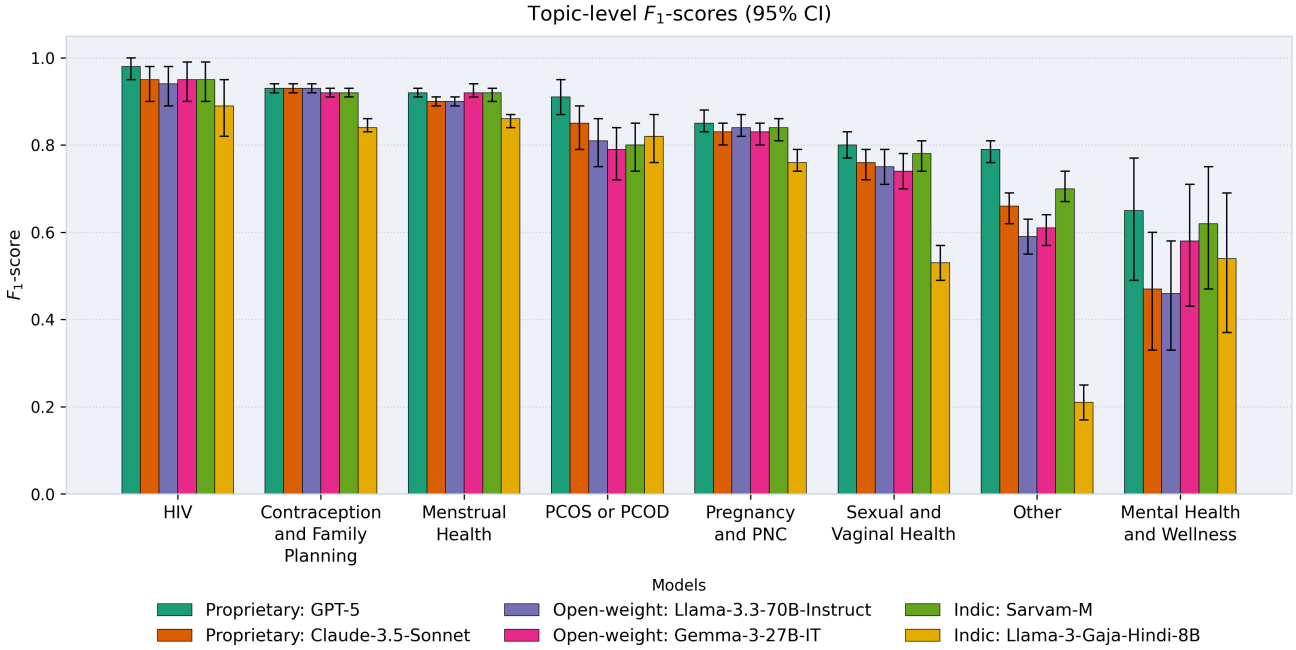

Figure 4 provides a closer look at topic-level differences across models. Proprietary systems maintained stability across nearly all categories, performing strongest in human immunodeficiency virus, Contraception and Family Planning, and Menstrual Health, which require precise and time-sensitive classification. Although human immunodeficiency virus was represented in fewer queries, performance remained consistent across models, suggesting less variability in how these queries are expressed. Sarvam-M again demonstrated competitive results, closely tracking proprietary and large multilingual models in categories such as Contraception and Family Planning, Menstrual Health, and polycystic ovary syndrome or polycystic ovarian disease. However, all models showed relative declines in Sexual and Vaginal Health, and Mental Health and Wellness, where queries frequently overlap with adjacent categories and require additional context for a clear-label decision. In these topics, diverse expressions of intent and limited sample sizes contribute to greater uncertainty, particularly for Mental Health and Wellness, where even a small number of misclassifications can substantially affect performance estimates. These findings highlight that while proprietary systems remain strongest, large multilingual open-weight models substantially close the gap, and Sarvam-M stands out as a promising Indic alternative. Yet, across all families, accurate subtopic-level predictions in code-mixed SRH queries continue to be the most persistent challenge.

To further examine where models succeed and fail, Figure 5 presents confusion matrices for proprietary, open-weight, and Indic models, highlighting recurring misclassification patterns. Consistent weaknesses appeared in overlap-prone and context-dependent topics such as Other, Mental Health and Wellness, and Sexual and Vaginal Health, where misclassifications were frequent and often collapsed into broader categories. Proprietary systems, particularly GPT-5, demonstrated the most balanced performance across all topics, while Claude-3.5-Sonnet showed sharper drops in Other categories despite strong results elsewhere. Open-weight models (Gemma-3-27B and Llama-3.3-70B) performed competitively but exhibited greater instability in culturally sensitive areas, including frequent misclassification of Sexual and Vaginal Health queries into the Pregnancy and PNC or Other category. Among Indic models, Sarvam-M emerged as the strongest, closely mirroring proprietary models' performance by correctly classifying over 90% of queries in Menstrual Health, Contraception and Family Planning, and Mental Health and Wellness. However, it still struggled in fine-grained intent recognition, with frequent misclassification of Contraception and Family Planning into the Pregnancy and PNC or Menstrual Health category. By contrast, smaller Indic systems such as Llama-3-Gaja-Hindi-8B displayed widespread confusions across sensitive categories and a notable tendency to produce "NotValid" predictions, where models generated labels not included in the predefined SRH hierarchy shown in the prompt. For clarity in Figure 5, predictions that fall outside the predefined SRH hierarchy are grouped under a "NotValid" category. These out-of-hierarchy outputs receive no credit in the hierarchical evaluation. These cases underscore the limited ability of smaller Indic models to consistently map inputs to the defined label structure and reflect the challenges related to limited training data and weaker alignment with the task structure.

XSL•FO
RenderX

Figure 5. Topic-level confusion matrices for 6 models representing proprietary, open-weight, and Indic categories. Each cell shows counts (black) and row-normalized percentages. Correct classifications appear in green diagonal cells with percentages in blue. Misclassifications appear in off-diagonal yellow cells, with percentages shown in red. "NotValid" denotes predictions that do not belong to any valid topic. HIV: human immunodeficiency virus; PCOD: polycystic ovarian disease; PCOS: polycystic ovary syndrome; PNC: postnatal care.

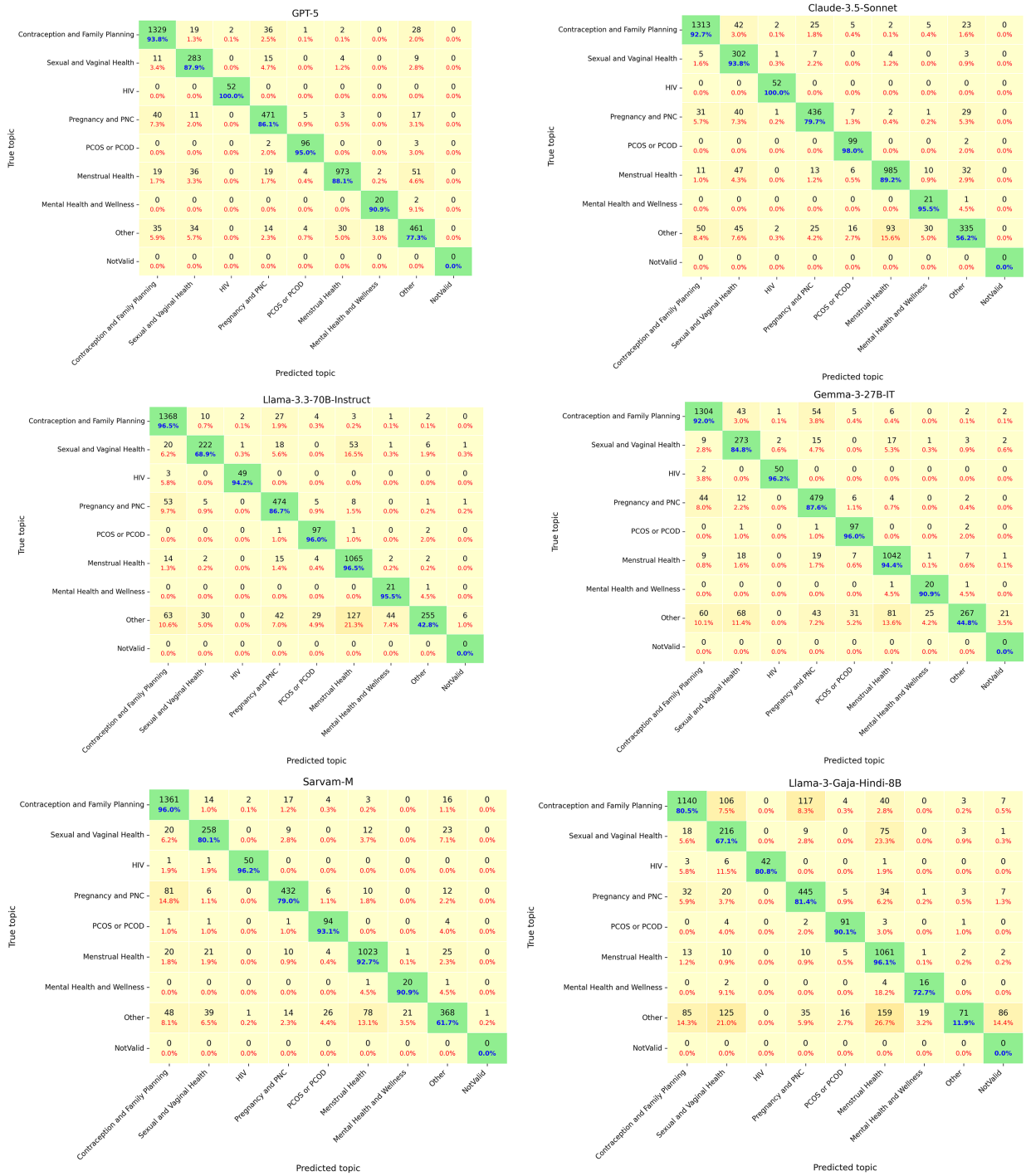

Error Analysis

To better understand model behavior beyond aggregate metrics, we conducted a qualitative analysis of errors from 6 representative models (GPT-5, Claude-3.5-Sonnet, Llama-3.3-70B-Instruct, Gemma-3-27B-IT, Sarvam-M, and Llama-3-Gaja-Hindi-8B) spanning proprietary, multilingual open-weight, and Indic categories. For each model, we sampled 50 misclassified queries and performed a detailed review of their errors. Table 4 presents examples of misclassification. Additionally, we included error analysis for GPT-5 and GPT-4o in Table S3 in Multimedia Appendix 2, highlighting cases where GPT-4o misclassified queries that GPT-5 handled correctly.

C1: Lexical Ambiguity and Euphemisms

In Table 4, row 1, the query "safaiya kaise karwate hain?" translates to "How is an abortion done?" and should be classified as Pregnancy and PNC→Abortion. However, only the proprietary models assigned the topic and subtopic correctly; both open-weight and Indic models misclassified the query

XSL•FO
RenderX

under Sexual and Vaginal Health, Menstrual Health, or General Health Queries. As detailed in Table S4 in Multimedia Appendix 3, model-generated explanations indicate that these errors stem from the literal interpretation of the colloquial term “safaiya” as hygiene, sanitary products, or general cleaning practices. In practice, however, the term is commonly used to refer to the abortion evacuation procedure in the Indian context [60]. This example highlights how culturally embedded euphemisms widely used in Indian contexts can mislead models if culturally specific meanings are not recognized.

C2: Misclassifying Stage in the Reproductive Health Journey

As shown in Table 4, row 2, models frequently confused different pregnancy stages. The query “Pregnancy me 5 month me vomiting hoti hya to kya kre” translates to “What should I do if I am vomiting in the 5th month of pregnancy?” which pertains to the antepartum stage (pregnancy period). Of the 6 best-performing models, only GPT-5, Claude-3.5-Sonnet, and Sarvam-M predicted this correctly. The remaining models classified this query under more general Pregnancy Information, thereby failing to distinguish the specific stage of the reproductive health journey.

As illustrated in Table S5 in Multimedia Appendix 3, these models recognized the pregnancy context but did not incorporate gestational timing into their subtopic selection. This distinction is clinically important, as symptoms such as vomiting have different implications and recommended guidance during the antepartum vs postpartum periods. Assigning such queries in overly broad subcategories such as “Pregnancy Information” dilutes the specificity of the advice that models are expected to provide and risks misalignment with clinical best practices.

C3: Cultural and Religious Context Error

In Table 4, row 3, the user query “Family planning Muslim community me accept hain kya?” translates to “Is family planning accepted in the Muslim community?” reflects a user’s religious concern regarding the acceptability of family planning. In our evaluation, both open-weight models correctly assigned the topic and subtopic, while others—including proprietary and Indic models—placed the query under Cultural, Religious, or Moral Norms. Model-generated reasons (Table S6 in Multimedia Appendix 3) indicate that these models focused on the religious framing of the questions rather than their underlying health-related intent. Similarly, in row 4, the query “Masik pali aane se Bhagvan ke pas kyo nahi jana chahiye?” (Why should one avoid going to temples during menstruation?) reflects a culturally and religiously grounded belief rather than a medical concern. GPT-5, Claude-3.5-Sonnet, and Sarvam-M correctly classified this query under Cultural, Religious, or Moral Norms. In contrast, 3 other models, including both open-weight and Indic models, classified it under health-related categories such as Menstrual Health or Misconceptions and Myths. Model-generated reasons (Table S6 in Multimedia Appendix 3) suggest that these systems focused on biological or psychological interpretations of menstruation rather than the user’s culturally grounded concern.

In both examples above, the alternative categorizations are semantically plausible given the cultural framing of the user queries; therefore, we treat these cases as a routing-level issue rather than a linguistic error. We later discuss how we might deal with such ambiguous cases when classifying user intent in the “Limitations and Future Works” subheading.

C4: Misunderstanding Reproductive Intent

This category captures cases where models failed to recognize the user’s reproductive intent, often reversing it from wanting to conceive to wanting to avoid conception. For example, in Table 4, row 5, the query “1 sal se bacha rukhne ke liye try kar rahe hai lekin nahi rukh raha hai to kya karna padega?” translates to “We have been trying to have a child for one year, but it has not happened. What should we do?“ and reflects the issue of infertility, a deeply personal and emotionally sensitive topic. However, only the proprietary model classified this query correctly. Model-generated reasons from other systems (Table S7 in Multimedia Appendix 3) indicate that the colloquial phrase “bacha rukna,” which in India commonly means conceiving, was interpreted literally as preventing pregnancy. As a result, these models prioritized Contraception and Family Planning subtopics such as Family Planning Queries and Contraceptive Effectiveness and Duration. While such interpretations are not inherently logical given the taxonomy, they diverge from the user’s underlying intent of seeking fertility-related guidance.

A similar misalignment occurred in row 6 with the query “Family planning me agar koi mahila test tube karvati hai to kya hota hai?” (“In family planning, if a woman undergoes a test tube procedure, what happens?”). Here, the user was referring to IVF, a fertility treatment aimed at achieving pregnancy. Yet all models categorized the query under Family Planning or Sterilization rather than Infertility. This demonstrates how culturally specific terms such as “test tube karvati hai” *can invert* user intent, classifying conception-seeking queries as if they were about contraception.

From an SRH perspective, such errors are critical, since infertility queries require guidance on fertility assessment and treatment. Misclassifying these queries risks invalidating user concerns, reinforcing stigma, as infertility in India is often associated with blame and silence, particularly for women [61]. Treating these queries as matters of pregnancy avoidance could further marginalize users seeking support.

C5: Misclassifying Questions About Bodily Anatomy or Processes

In Table 4, row 7, the query “Kitni der jinda rahte hain sperm?” translates to “How long do sperm stay alive?“ and should be categorized under Sexual and Vaginal Health, subtopic Reproductive Anatomy. Only 1 model (Claude-3.5-Sonnet) correctly identified both the topic and subtopic. Other models routed the query to categories such as Sexual and Vaginal Health, Contraception and Family Planning, Pregnancy and PNC, or General Health Queries (Table S8 in Multimedia Appendix 3). Model-generated reasons indicate that many systems overassociate the term sperm with sexual behavior, contraception, or pregnancy risk rather than recognizing it as a biological question. Such routing decisions are consequential,

XSL•FO
RenderX

as users seeking factual information about sperm viability may instead receive guidance about sexual behavior or contraception, reducing clarity and potentially misdirecting them away from accurate, evidence-based biological information.

Discussion

Principal Findings

Our evaluation shows that both model family and training data alignment play an important role in hierarchical SRH intent classification, alongside model scale. Proprietary systems still lead, but the gap narrows when models are trained on culturally aligned and domain-relevant data. Sarvam-M, an Indic-first model, emerged as the strongest Indic system and achieved performance comparable to top-performing proprietary models and the best large open-weight baselines such as Llama-3.3-70B-Instruct and Gemma-3-27B-IT. It is also important to consider that Sarvam-M is larger than the other Indic models evaluated, and its performance likely reflects a combination of increased model capacity and Indic-focused training data. While these results do not fully disentangle the effects of scale and data quality, they suggest that culturally aligned resources, rather than parameter count, are the primary drivers of performance on code-mixed intent classification.

Across the 6 best-performing models, we observed a consistent pattern of subtopic misclassification. In Pregnancy and PNC, antepartum and postpartum concerns are frequently conflated, and abortion-related queries are sometimes assigned to the "Pregnancy Information" subcategory. In "Sexual and Vaginal Health," euphemisms such as "safaiya or safai" (colloquial for abortion) were misinterpreted literally as "cleaning." Similarly, culturally embedded phrases such as "bacha rukna" (conceiving) are sometimes misinterpreted as preventing pregnancy, and "test tube karvati hai" (referring to IVF) is sometimes treated as sterilization or contraception [62]. This reflects how health communication in rural, low-literacy settings often relies on indirect, gendered, and culturally mediated channels [63]. Accounting for these cultural expressions is essential to improve classification accuracy and ensure that clinically relevant questions are not overlooked. We also observed that SRH queries could be misrouted as Cultural, Religious, or Moral Norms, even when the underlying intent is to seek health guidance or clinical information. A user query could also plausibly fit into multiple other categories, as we observed. Future work may explore culturally informed prompt design and lightweight in-context examples that expose models to common SRH euphemisms and colloquial expressions. Additionally, to address overlapping topics, researchers could explore alternative approaches that account for such ambiguity, such as ranked classification [64].

In the "Other" category, even strong models such as Llama-3-Gaja-Hindi-8B and Gemma-3-27B-IT frequently assigned queries to new topics outside the defined hierarchy: we consider that kind of misclassification as "Not Valid" predictions (Figure 5). These cases arise when models generate unintended or unsupported labels that are not part of the hierarchical classification structure, reflecting models' tendency toward hallucinations [65] or schema misalignment [66]. Our hierarchical framework requires a correct topic prediction before the subtopic evaluation. Accordingly, "NotValid" predictions are treated as complete classification failures and receive no credit in the hierarchical F_1 score. Overall, these patterns indicate that models struggle to reliably map culturally and contextually grounded queries to the specified hierarchy. This limitation is essentially critical in SRH contexts, where errors at the topic level prevent meaningful subtopic interpretation and risk misrouting sensitive queries. High precision at the fine-grained intent level is paramount to ensure safe guidance and avoid reinforcing stigma.

When compared to prior chatbot efforts, our findings highlight the importance of a hierarchical classification approach. SnehAI [15], an SRH-focused conversational agent deployed in India, demonstrated feasibility but relied on rule-based methods, which restrict its ability to process open-ended, code-mixed, or culturally nuanced queries. Health-Pariksha [38], while not designed specifically for SRH, evaluated several LLMs on real-world health care queries, measuring the factual correctness, semantic similarity, coherence, and conciseness of the model responses. Although this approach contributed to improving trustworthiness, it did not address the challenges of intent recognition or hierarchical classification. In contrast, our framework directly targets the layered structure of SRH queries, enabling distinction between broad categories and their fine-grained subcategories. This supports classifications that are both clinically appropriate and culturally relevant. Through our evaluation, proprietary models emerged as the most reliable for hierarchical classification, but notably, Sarvam-M emerged as the second-highest performing model overall—outscoring Claude-3.5-Sonnet and closely matching Llama-3.3-70B-Instruct. Given that Sarvam-M is open weight and free to use, it represents a promising alternative to proprietary systems, showing that with robust data and cultural adaptation, Indic models can deliver competitive performance.

Limitations and Future Works

This study has several limitations. First, we focused on Hinglish as an illustrative case of code-mixed SRH queries, while many other languages and code-mixing patterns are used in India. Second, the hierarchical intent schema was developed for our setting and user population; researchers working in other regions should adapt label definitions to their context and target users. Third, although we evaluated many more models, several were excluded from the final set presented in this paper, either due to poor performance on code-mixed data or the cost of highly parameterized models. Fourth, our dataset is limited in size; certain topics and subtopics are underrepresented, resulting in lower statistical power for these categories, and performance estimates for these topics should be interpreted as preliminary. Fifth, our evaluation was conducted in a strict zero-shot setting, where models were not provided with the detailed subtopic definitions used by human annotators. This design allows us to assess how models infer intent from real-world, code-mixed queries without semantic scaffolding. Finally, our evaluation treats intent as a single topic, even though some queries plausibly map to more than 1 topic, motivating the exploration of alternative modeling approaches that can better capture such ambiguity.

XSL•FO
RenderX

In the future, we plan to extend our work to support additional Indian languages and code-mixing patterns, enriching both the breadth of topics and the depth of subtopic coverage in collaboration with the Myna Mahila Foundation. We also plan to evaluate in-context prompting strategies and to explore ranked classification [67] and hierarchical selective classification [64] approaches, which allow models to better account for ambiguity in hierarchical intent prediction.

Implications

This study highlights the importance of hierarchical classification for SRH applications in low-resource settings. By capturing intent at both the topic and subtopic levels, this framework improves contextual precision, supports safer dialogue systems, and enables public health organizations to design targeted interventions. At the same time, our findings show that LLMs often struggle with euphemisms and sociocultural framing, highlighting the need for systems that are socially attuned as well as technically accurate. Our evaluation offers a starting benchmark that can guide the development of culturally aligned, open-source LLM tools for low-resource health care contexts. More broadly, this work contributes to the discourse on health equity in AI by addressing the linguistic and cultural barriers that limit access to reliable health information for marginalized populations.

Conclusions

In this study, we evaluated the zero-shot performance of proprietary, open-weight multilingual, and instruction-tuned Indic LLMs on the hierarchical classification of SRH queries. Our findings reveal that the proprietary model (GPT-5) was the most reliable, while Sarvam-M emerged as the strongest Indic model and performed competitively with Claude-3.5-Sonnet and large open-weight models such as Llama-3.3-70B-Instruct. This highlights the importance of culturally aligned and domain-relevant training data alongside model scale. Our error analysis further revealed that models frequently misclassified queries that involved euphemisms, cultural or religious language, or time-sensitive concerns, underscoring the importance of capturing nuance at both topic and subtopic levels. Misclassifications in these areas risk unsafe guidance and reinforcing stigma, particularly around sensitive issues such as abortion and infertility. By introducing a benchmark evaluation framework, this study supports the development of open-source multilingual models for SRH and advances culturally aligned, socially responsive AI systems to better serve the health information needs of underserved communities.

The source code used in this study, including scripts for model evaluation, hierarchical metric calculation (hF_1), bootstrap resampling, and CMI calculation, is publicly available at a GitHub repository [68] to support future research.

Acknowledgments

We sincerely thank the Myna Mahila Foundation for their invaluable support throughout this study. We would also like to thank the community members who were engaged in the data collection process with the Myna Mahila Foundation, and the editor and reviewers for their thoughtful and detailed comments.

Funding

This research was made possible with the support of the Google Award for Inclusion Research and the Agency Fund through Myna Mahila USA (SA24-03-03).

Data Availability

The sexual and reproductive health queries (SRHQ)-India dataset introduced in this study contains sensitive queries related to sexual and reproductive health, expressed in code-mixed Hinglish. To ensure privacy and ethical use, the dataset will not be publicly released in full. However, deidentified subsets or access to the full dataset can be provided upon reasonable request to the corresponding author for academic and noncommercial research purposes, subject to data use agreements and ethical review.

Authors' Contributions

SKD and MS conceptualized the study and designed the data framework. AT and MS performed data annotation. SKD carried out the formal analysis, developed the methodology, and created visualizations. ZM, SJ, TD, and MS provided resources for the study. AI and SJ supervised the research. The original draft was prepared by SKD, and all authors contributed to reviewing and editing the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Annotation guidelines for hierarchical topic-subtopic labeling of the sexual and reproductive health queries (SRHQ-India dataset).
[DOCX File, 46 KB - jmir_v28i1e86545_app1.docx]

Multimedia Appendix 2

XSL•FO
RenderX

Prompt template and statistical performance analysis of zero-shot hierarchical classification performance across open-weight, Indic, and proprietary large language models (LLMs), including a comparative analysis of GPT-5 and GPT-4o on representative user queries.
[DOCX File, 36 KB - jmir_v28i1e86545_app2.docx]

Multimedia Appendix 3
Model-generated reasons for predicted hierarchical classifications across representative sexual and reproductive health (SRH) queries, reported for all evaluated models.
[DOCX File, 38 KB - jmir_v28i1e86545_app3.docx]

References

1. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023 May 17;10:e47564. [doi: 10.2196/47564] [Medline: 37195756]
2. Wang W, Xu B, Fang T, Zhang L, Song Y. KnowComp at semeval-2023 task 7: fine-tuning pre-trained language models for clinical trial entailment identification. In: E S, AS D, H TM, R K, E S, editors. Presented at: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023); Jul 13-14, 2023; Toronto, Canada p. 1-9 URL: https://aclanthology.org/2023.semeval-1 [accessed 2026-03-11] [doi: 10.18653/v1/2023.semeval-1.1]
3. Verma G, Mujumdar R, Wang ZJ, Choudhury MD, Kumar S. Overcoming language disparity in online content classification with multimodal learning. 2022 Presented at: Proceedings of the International AAAI Conference on Web and Social Media; Jun 6-9, 2022; Atlanta, GA p. 1040-1051. [doi: 10.1609/icwsm.v16i1.19356]
4. Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. arXiv. Preprint posted online on Feb 27, 2023. [doi: 10.48550/arXiv.2302.13971]
5. OpenAI AJ, Adler S, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 15, 2024. [doi: 10.48550/arXiv.2303.08774]
6. Das M, Angeli F, Krumeich AJSM, van Schayck OCP. The gendered experience with respect to health-seeking behaviour in an urban slum of Kolkata, India. Int J Equity Health 2018 Feb 14;17(1):24. [doi: 10.1186/s12939-018-0738-8] [Medline: 29444674]
7. Deva R, Ramani D, Divate T, Jalota S, Ismail A. “Kya family planning after marriage hoti hai?”: integrating cultural sensitivity in an LLM chatbot for reproductive health. Presented at: CHI ’25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems; Apr 26 to May 1, 2025; Yokohama Japan p. 1-23 URL: https://dl.acm.org/doi/proceedings/10.1145/3706598 [doi: 10.1145/3706598.3713362]
8. Omar M, Soffer S, Agbareia R, et al. Sociodemographic biases in medical decision making by large language models. Nat Med 2025 Jun;31(6):1873-1881. [doi: 10.1038/s41591-025-03626-6] [Medline: 40195448]
9. Purnat TD, Wilhelm E, Scales D, et al. Impacts of sexual and reproductive health and rights misinformation in digital spaces on human rights protection and promotion: scoping review. JMIR Infodemiology 2025 Dec 30;5:e83747. [doi: 10.2196/83747] [Medline: 41468582]
10. John JN, Gorman S, Scales D, Gorman J. Online misleading information about women’s reproductive health: a narrative review. J Gen Intern Med 2025 Apr;40(5):1123-1131. [doi: 10.1007/s11606-024-09118-6] [Medline: 39511120]
11. Singh S, Shekhar C, Acharya R, et al. The incidence of abortion and unintended pregnancy in India, 2015. Lancet Glob Health 2018 Jan;6(1):e111-e120. [doi: 10.1016/S2214-109X(17)30453-9] [Medline: 29241602]
12. Singh SM, Garg SK, Misra A, Seth A, Chakraborty T. SUKHSANDESH: an avatar therapeutic question answering platform for sexual education in rural india. In: Larson K, editor. Presented at: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24); Aug 3-9, 2024; Jeju, South Korea p. 7465-7473. [doi: 10.24963/ijcai.2024/826]
13. Singh D, Singh T, Kumar D. Complications during pregnancy among Indian tribal women: a mini-review. JMMS 2025;2(2):128-131. [doi: 10.51219/JMMS/Kumar-D/25]
14. Joyce CM, Sharma D, Mukherji A, Nandi A. Socioeconomic inequalities in adverse pregnancy outcomes in India: 2004-2019. PLOS Glob Public Health 2024;4(9):e0003701. [doi: 10.1371/journal.pgph.0003701] [Medline: 39292712]
15. Wang H, Gupta S, Singhal A, et al. An artificial intelligence chatbot for young people’s sexual and reproductive health in India (SnehAI): instrumental case study. J Med Internet Res 2022 Jan 3;24(1):e29969. [doi: 10.2196/29969] [Medline: 34982034]
16. Patwa P, Aguilar G, Kar S, et al. SemEval-2020 task 9: overview of sentiment analysis of code-mixed tweets. In: Herbelot A, Zhu X, Palmer A, Schneider N, May J, Shutova E, editors. Presented at: Proceedings of the Fourteenth Workshop on Semantic Evaluation; Dec 12-13, 2020; Barcelona (online) p. 774-790 URL: https://aclanthology.org/2020.semeval-1 [doi: 10.18653/v1/2020.semeval-1.100]
17. Parshad RD, Bhowmick S, Chand V, Kumari N, Sinha N. What is India speaking? Exploring the “Hinglish” invasion. Phys A: Stat Mech Appl 2016 May;449:375-389. [doi: 10.1016/j.physa.2016.01.015]

XSL•FO
RenderX

18. Sengupta A, Das S, Akhtar MS, Chakraborty T. Social, economic, and demographic factors drive the emergence of Hinglish code-mixing on social media. Humanit Soc Sci Commun 2024;11(1):606. [doi: 10.1057/s41599-024-03058-6]
19. Heredia M, Labaka G, Barnes J, Soroa A. Conditioning llms to generate code-switched text. arXiv. Preprint posted online on Feb 18, 2025. [doi: 10.48550/arXiv.2502.12924]
20. Yong ZX, Zhang R, Forde J, et al. Prompting multilingual large language models to generate code-mixed texts: the case of south east asian languages. Presented at: Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching; Dec 6-10, 2023; Singapore p. 43-63 URL: https://aclanthology.org/2023.calcs-1 [accessed 2026-02-02] [doi: 10.18653/v1/2023.calcs-1.5]
21. Moskovitch R, Cohen-Kashi S, Dror U, Levy I, Maimon A, Shahar Y. Multiple hierarchical classification of free-text clinical guidelines. Artif Intell Med 2006 Jul;37(3):177-190. [doi: 10.1016/j.artmed.2006.04.001] [Medline: 16730962]
22. Ee-Peng L. Hierarchical text classification and evaluation. Presented at: 2001 IEEE International Conference on Data Mining; Nov 29 to Dec 2, 2001; San Jose, CA p. 521-528. [doi: 10.1109/ICDM.2001.989560]
23. Kosmopoulos A, Partalas I, Gaussier E, Paliouras G, Androutsopoulos I. Evaluation measures for hierarchical classification: a unified view and novel approaches. Data Min Knowl Discov 2015 May;29(3):820-865. [doi: 10.1007/s10618-014-0382-x]
24. He B, Nag S, Cui L, et al. Hierarchical query classification in e-commerce search. Presented at: WWW '24: Companion Proceedings of the ACM Web Conference 2024; May 13-17, 2024; Singapore p. 338-345 URL: https://dl.acm.org/doi/proceedings/10.1145/3589335 [accessed 2026-03-11] [doi: 10.1145/3589335.3648332]
25. Cao H, Hu DH, Shen D, et al. Context-aware query classification. Presented at: SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval; Jul 19-23, 2009; Boston, MA p. 3-10 URL: https://dl.acm.org/doi/proceedings/10.1145/1571941 [accessed 2026-03-11] [doi: 10.1145/1571941.1571945]
26. Zhu M, Ahuja A, Wei W, Reddy CK. A hierarchical attention retrieval model for healthcare question answering. Presented at: WWW '19: The World Wide Web Conference; May 13-17, 2019; San Francisco, CA p. 2472-2482 URL: https://dl.acm.org/doi/proceedings/10.1145/3308558 [accessed 2026-03-11] [doi: 10.1145/3308558.3313699]
27. Deng Y, Denecke K. Classification of user queries according to a hierarchical medical procedure encoding system using an ensemble classifier. Front Artif Intell 2022;5:1000283. [doi: 10.3389/frai.2022.1000283] [Medline: 36406473]
28. Sosnowski Ł, Żuławińska J, Dutta S, Szymusik I, Zyguła A, Bambul-Mazurek E. Artificial intelligence in personalized healthcare analysis for womens' menstrual health disorders. Presented at: 2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS); Sep 4-7, 2022; Sofia, Bulgaria p. 751-760. [doi: 10.15439/2022F59]
29. Chen S, Bouadjenek MR, Naseem U, et al. Leveraging taxonomy and LLMs for improved multimodal hierarchical classification. In: Rambow O, Wanner L, Apidianaki M, Al-Khalifa H, Eugenio BD, Schockaert S, editors. Presented at: Proceedings of the 31st International Conference on Computational Linguistics Association for Computational Linguistics; Jan 19-24, 2025; Abu Dhabi, UAE p. 6244-6254 URL: https://aclanthology.org/2025.coling-main.417/ [accessed 2025-09-14]
30. GPT-5 system card. OpenAI. URL: https://cdn.openai.com/gpt-5-system-card.pdf [accessed 2025-09-20]
31. OpenAI HA, Lerer A, et al. GPT-4o system card. arXiv. Preprint posted online on Oct 25, 2024. [doi: 10.48550/arXiv.2410.21276]
32. Grattafiori A, Dubey A, Jauhri A, et al. The Llama 3 herd of models. arXiv. Preprint posted online on Jul 31, 2024. [doi: 10.48550/arXiv.2407.21783]
33. Kamath A, Ferret J, et al. Gemma 3 technical report. arXiv. Preprint posted online on Mar 25, 2025. [doi: 10.48550/arXiv.2503.19786]
34. Wang L, Wan Z, Ni C, et al. A systematic review of chatgpt and other conversational large language models in healthcare. medRxiv. Preprint posted online on Apr 27, 2024. [doi: 10.1101/2024.04.26.24306390]
35. Jin W, Mao H, Li Z, et al. Amazon-M2: a multilingual multi-locale shopping session dataset for recommendation and text generation. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, editors. Presented at: NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems; Dec 10-16, 2023; New Orleans, LA p. 8006-8026 URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/193df57a2366d032fb18dcac0698d09a-Paper-Datasets_and_Benchmarks.pdf [accessed 2026-03-11] [doi: 10.52202/075280-0351]
36. Jin Y, Chandra M, Verma G, Hu Y, De Choudhury M, Kumar S. Better to ask in English: cross-lingual evaluation of large language models for healthcare queries. Presented at: WWW '24: Proceedings of the ACM Web Conference 2024; May 13-17, 2024; Singapore p. 2627-2638 URL: https://dl.acm.org/doi/proceedings/10.1145/3589334 [accessed 2026-03-11] [doi: 10.1145/3589334.3645643]
37. Tripathi N, Sekher TV. Youth in India ready for sex education? Emerging evidence from national surveys. In: Murthy AK, editor. PLoS ONE 2013;8(8):e71584. [doi: 10.1371/journal.pone.0071584] [Medline: 23951197]
38. Gumma V, Raghunath A, Jain M, Sitaram S. HEALTH-PARIKSHA: assessing RAG models for HEALTH chatbots in real-world multilingual settings. arXiv. Preprint posted online on Oct 17, 2024. [doi: 10.48550/arXiv.2410.13671]
39. Sivarajkumar S, Wang Y. HealthPrompt: a zero-shot learning paradigm for clinical natural language processing. AMIA Annu Symp Proc 2022;2022:972-981. [Medline: 37128372]

XSL•FO
RenderX

40. Guo Y, Ovadje A, Al-Garadi MA, Sarker A. Evaluating large language models for health-related text classification tasks with public social media data. J Am Med Inform Assoc 2024 Oct 1;31(10):2181-2189. [doi: 10.1093/jamia/ocae210] [Medline: 39121174]
41. Singh P, Lefever E. Sentiment analysis for hinglish code-mixed tweets by means of cross-lingual word embeddings. In: Solorio T, Choudhury M, Bali K, Sitaram S, Das A, Diab M, editors. Presented at: Proceedings of the 4th Workshop on Computational Approaches to Code Switching European Language Resources Association; May 11-16, 2020; Marseille, France p. 45-51 URL: https://aclanthology.org/2020.calcs-1.6/ [accessed 2025-12-17]
42. Chanda S, Mishra A, Pal S. Sentiment analysis of code-mixed Dravidian languages leveraging pretrained model and word-level language tag. Nat lang process 2025 Mar;31(2):477-499. [doi: 10.1017/nlp.2024.30]
43. Gambäck B, Das A, et al. Comparing the level of code-switching in corpora. In: Calzolari N, Choukri K, Declerck T, editors. Presented at: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) European Language Resources Association (ELRA); May 23-28, 2016; Portorož, Slovenia p. 1850-1855 URL: https://aclanthology.org/L16-1292/
44. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. Presented at: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers; Apr 3-7, 2017; Valencia, Spain p. 427-431 URL: http://aclweb.org/anthology/E17-2 [accessed 2025-02-02] [doi: 10.18653/v1/E17-2068]
45. Bañón M, Ramírez-Sánchez G, Zaragoza-Bernabeu J, FastSpell ORS. The langid magic spell. In: Calzolari N, Kan MY, Hoste V, Lenci A, Sakti S, Xue N, editors. Presented at: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) ELRA and ICCL; May 20-25, 2024; Torino, Italia p. 7133-7140 URL: https://aclanthology.org/2024.lrec-main.626/ [accessed 2026-03-11] [doi: 10.63317/3w889497us8n]
46. Ibrahim M. CUFE@NLU of devanagari script languages 2025: language identification using fasttext. In: Sarveswaran K, Vaidya A, Krishna Bal B, Shams S, Thapa S, editors. Presented at: Proceedings of the First Workshop on Challenges in Processing South Asian Languages; Jan 19-24, 2025; Abu Dhabi, UAE p. 273-277 URL: https://aclanthology.org/2025.chipsal-1.30/ [accessed 2026-03-11]
47. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46. [doi: 10.1177/001316446002000104]
48. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005 May;37(5):360-363. [Medline: 15883903]
49. Jiang AQ, Sablayrolles A, Roux A, et al. Mixtral of experts. arXiv. Preprint posted online on Jan 8, 2024. [doi: 10.48550/arXiv.2401.04088]
50. Riviere M, Pathak S, et al. Gemma 2: improving open language models at a practical size. arXiv. Preprint posted online on Jul 31, 2024. [doi: 10.48550/arXiv.2408.00118]
51. Qwen YA, Yang B, et al. Qwen2.5 technical report. arXiv. Preprint posted online on Dec 19, 2025. [doi: 10.48550/arXiv.2412.15115]
52. Dang J, Singh S, D'souza D, et al. Aya expanse: combining research breakthroughs for a new multilingual frontier. arXiv. Preprint posted online on Dec 5, 2024. [doi: 10.48550/arXiv.2412.04261]
53. Gala J, Jayakumar T, Husain JA, et al. Airavata: introducing hindi instruction-tuned LLM. arXiv. Preprint posted online on Jan 26, 2024. [doi: 10.48550/arXiv.2401.15006]
54. GenVRadmin/aryabhatta-gemmagenz-vikas-merged. Hugging Face. URL: https://huggingface.co/GenVRadmin/AryaBhatta-GemmaGenZ-Vikas-Merged [accessed 2026-03-11]
55. Cognitive-lab/llama3-gaja-hindi-8B-v01hugging face. Hugging Face. URL: https://huggingface.co/Cognitive-Lab/LLama3-Gaja-Hindi-8B-v0.1 [accessed 2025-10-27]
56. Kallappa A, Kamble P, Ravi A, et al. Krutrim LLM: multilingual foundational model for over a billion people. arXiv. Preprint posted online on Feb 10, 2025. [doi: 10.48550/arXiv.2502.09642]
57. Claude 3.5 Sonnet. Anthropic. URL: https://www.anthropic.com/news/claude-3-5-sonnet [accessed 2025-09-20]
58. Sarvam-M. Sarvam. URL: https://www.sarvam.ai/blogs/sarvam-m [accessed 2025-09-20]
59. Zhan P, Xu Z, Tan Q, Song J, Xie R. Unveiling the lexical sensitivity of llms: combinatorial optimization for prompt enhancement. In: Al-Onaizan Y, Bansal M, Chen YN, editors. Presented at: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Nov 12-16, 2024; Miami, FL p. 5128-5154 URL: https://aclanthology.org/2024.emnlp-main [accessed 2026-03-11] [doi: 10.18653/v1/2024.emnlp-main.295]
60. Bhattacharya S, Bashar MA, Singh A. So near, yet so far: access to safe abortion services remains elusive for poor women in India. BMJ Case Rep 2017 Oct 13;2017:bcr2017220980. [doi: 10.1136/bcr-2017-220980] [Medline: 29030364]
61. Roberts L, Renati S, Solomon S, Montgomery S. Women and infertility in a pronatalist culture: mental health in the slums of Mumbai. Int J Womens Health 2020;12:993-1003. [doi: 10.2147/IJWH.S273149] [Medline: 33192102]
62. Patra S, Unisa S. Addressing reproductive health knowledge, infertility and coping strategies among rural women in India. J Biosoc Sci 2021 Jul;53(4):557-565. [doi: 10.1017/S0021932020000371] [Medline: 32677598]

XSL•FO
RenderX

63. Husain Z, Dutta M, Ghosh S. Contraceptive use among illiterate women in India: does proximate illiteracy matter?.: MPRA Pap; 2011. URL: https://ideas.repec.org//p/pra/mprapa/30790.html [accessed 2025-10-16]
64. El-Yaniv R, Galil I, Goren S. Hierarchical selective classification. Presented at: Advances in Neural Information Processing Systems 37; Dec 10-15, 2024 URL: http://www.proceedings.com/79017.html [accessed 2026-03-11] [doi: 10.52202/079017-3526]
65. Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. Interact J Med Res 2025 Jan 28;14:e59823. [doi: 10.2196/59823] [Medline: 39874574]
66. Yang K, Tao G, Chen X, Xu J. Alleviating the fear of losing alignment in LLM fine-tuning. Presented at: 2025 IEEE Symposium on Security and Privacy (SP); May 12-14, 2025; San Francisco, CA p. 2152-2170. [doi: 10.1109/SP61157.2025.00171]
67. Atapour-Abarghouei A, Bonner S, McGough AS. Rank over class: the untapped potential of ranking in natural language processing. Presented at: 2021 IEEE International Conference on Big Data (Big Data); Dec 15-18, 2021; Orlando, FL p. 3950-3959. [doi: 10.1109/BigData52589.2021.9671386]
68. EmoryCareLab/hierarchical-srh-intent. GitHub. 2026. URL: https://github.com/EmoryCareLab/hierarchical-srh-intent [accessed 2026-02-02]

Abbreviations

AI: artificial intelligence
CMI: Code-Mixing Index
IVF: in-vitro fertilization
LLM: large language model
PNC: postnatal care
SRH: sexual and reproductive health
SRHQ: sexual and reproductive health queries

Edited by A Coristine; submitted 30.Oct.2025; peer-reviewed by AK Singh, FT Brito, M Borovic; accepted 18.Feb.2026; published 24.Mar.2026.

Please cite as:
Dey SK, S M, Thapa A, Shah M, Mehta Z, Kapile SK, Divate T, Jalota S, Ismail A
Understanding User Intent in Code-Mixed Sexual and Reproductive Health Queries in Urban India: Hierarchical Classification Approach Using Large Language Models
J Med Internet Res 2026;28:e86545
URL: https://www.jmir.org/2026/1/e86545
doi:10.2196/86545

© Sumon Kanti Dey, Manvi S, Aradhana Thapa, Meet Shah, Zeel Mehta, Shraddha Kale Kapile, Tanvi Divate, Suhani Jalota, Azra Ismail. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Longitudinal Effects of a Smartphone Game (Tumaini) for HIV Prevention Among Kenyan Adolescents: 45-Month Trajectories of Condom Use–Related Proximal Outcomes From a Randomized Controlled Trial

Kate Winskell[1], PhD; Gaëlle Sabben[1], MPH; Haowen Qin[2], MSPH; Calvin Mbeda[3], BSc; Sophie Goldenberg[4], MPH; Ken Ondeng'e[3], MPH; Richard Ndivo[3], BSc; Judith Arego[3], BA; Xinwei He[2], MS; Robert A Bednarczyk[1], PhD; Kelli Komro[5], PhD, MPH; Robert H Lyles[2], PhD; Victor Mudhune[3], PhD

[1]Hubert Department of Global Health, Rollins School of Public Health at Emory University, Atlanta, GA, United States

[2]Department of Biostatistics and Bioinformatics, Rollins School of Public Health at Emory University, Atlanta, GA, United States

[3]HIV Research Division, Centre for Global Health Research, Kenya Medical Research Institute, Kisumu, Kenya

[4]Department of Epidemiology, Rollins School of Public Health at Emory University, Atlanta, GA, United States

[5]Department of Behavioral, Social, and Health Education Sciences, Rollins School of Public Health at Emory University, Atlanta, GA, United States

Corresponding Author:
Kate Winskell, PhD
Hubert Department of Global Health
Rollins School of Public Health at Emory University
1518 Clifton Road, NE
Atlanta, GA, 30322
United States
Phone: 1 404 727 5286
Email: swinske@emory.edu

Abstract

Background: African adolescents and young adults account for a disproportionate number of new HIV infections. There is an urgent need to identify scalable and cost-effective behavioral HIV prevention strategies for this population. Using a condom at first sex is associated with a higher likelihood of consistent use later. Tumaini ("Hope for the Future" in Swahili; Emory University) is a choose-your-own-adventure smartphone game that has been shown to reduce the risk of unprotected first sex by end line in a 45-month randomized controlled trial in western Kenya.

Objective: This study aimed to assess the impact of Tumaini on proximal outcomes related to condom use at first sex (specifically, behavioral intentions, self-efficacy, attitudes, and knowledge) longitudinally across mid-adolescence in the above trial.

Methods: Adolescent participants (n=996, mean baseline age 14, SD 0.56 years) were randomized 1:1 to receive either a smartphone loaded with Tumaini or an attention-control math game for 5 to 7 weeks at 3 time points (mean age 14.0, SD 0.56; 15.3, SD 0.55; and 16.0, SD 0.56 years, respectively). They completed a behavioral survey at 13 time points, through mean age 17.7 (SD 0.56) years. Using generalized estimating equations and controlling for age at baseline, we modeled mean scores (overall and stratified by gender) on a range of condom-related survey items over time to assess mean differences at specific time points. We applied appropriate Bonferroni corrections to inferences about cross-arm differences in mean changes relative to baseline at 4 time points (after each intervention period and at end line; α=.05/4) and within-arm mean changes relative to baseline at each of the 12 post-baseline time points (α=.05/12). Analyses were conducted as intent-to-treat.

Results: At end line, 97.8% (n=974) of the sample had been retained. Participants in both arms dedicated a mean total of >30 hours to their assigned game. There was significant improvement across all condom-related proximal outcomes in the intervention arm relative to the control arm immediately after initial intervention exposure. For almost all outcomes, a significant cross-arm difference was also present at end line and for most outcomes at the 2 intervening comparison time points. Some outcomes saw stronger intervention effects on female participants (eg, self-efficacy to refuse unprotected sex) or male participants (eg, knowledge that condoms are an effective way to prevent HIV). In each arm, intention to use a condom at first sex was consistently higher among male participants; however, female intervention-arm scores overtook male control-arm scores following initial intervention exposure.

XSL•FO
RenderX

Conclusions: Tumaini significantly improved theory-based proximal outcomes related to condom use, with effects sustained 45 months post initial exposure and 16 months post most recent exposure. Adolescents benefited from even short-term exposure, though repeated exposure generally sustained and reinforced intervention effects. As access to smartphones increases, Tumaini has potential for high scalability and impact on condom-related outcomes.

Trial Registration: ClinicalTrials.gov NCT04437667; https://clinicaltrials.gov/study/NCT04437667

International Registered Report Identifier (IRRID): RR2-10.2196/35117

(J Med Internet Res 2026;28:e83982) doi:10.2196/83982

KEYWORDS

adolescents; HIV; prevention; sexual health; sub-Saharan Africa; Kenya; behavioral intervention; serious game; game for health; interactive narrative; smartphone; randomized controlled trial; mHealth; condom; behavioral intention; self-efficacy; attitude; knowledge

Introduction

Worldwide, a third of all new adult (aged 15-49 years) HIV infections occur in eastern and southern Africa, and over a third of these infections occur in young people aged 15-24 years [1]. HIV-related morbidity and mortality in this age group are disproportionately high. In addition, as the largest-ever generation of young Africans is now entering adolescence, it is feared that a resurgence in the global epidemic could ensue unless more is done to prevent HIV in this demographic [2]. There is an urgent need to develop and rigorously test scalable and cost-effective behavioral interventions to prevent HIV among African adolescents and young adults, particularly adolescent girls and young women who are disproportionately affected [1].

Reaching adolescents with information and skills before sexual debut may help establish long-term patterns of safer sexual behavior [3]. For example, studies have shown that those who use condoms at first sex are more likely to use them consistently in the future [4]. There is a need for a stronger evidence base on the longitudinal effects of early and ongoing intervention, as few studies evaluating any HIV prevention interventions for adolescents and young adults include long-term follow-up [5].

Scaling up in-person, group-based HIV prevention interventions is challenging [6]. Digital behavioral interventions are likely to be easier to deliver at scale with consistent quality than those delivered in person. As smartphone penetration grows across sub-Saharan Africa, especially among young people, digital HIV prevention interventions are becoming increasingly viable. Across the continent, 81% of mobile phone connections will be smartphones by 2030 [7].

Electronic games are a promising subset of digital behavioral interventions [8]. However, few have been rigorously designed and evaluated [9]. Specifically, games leveraging interactive narrative can provide a level of experiential learning unparalleled by most other interventions by allowing players to experience real agency in a virtual and safe environment.

Tumaini (“hope for the future” in Swahili) is a “choose-your-own-adventure” game app for inexpensive Android smartphones. It was developed by specialists in adolescent sexual health (KW, GS, with others) in collaboration with a commercial game developer with input from Kenyan adolescents and their parents. The game’s theoretical framework [10] draws on Social Cognitive Theory [11], existing evidence-based interventions for youth HIV prevention, and literature on games for health and Entertainment-Education [12,13]. It is grounded in research on HIV-themed narratives written by thousands of young Africans [14-16].

Tumaini uses interactive narrative to promote observational learning, cognitive and behavioral rehearsal, and problem-solving [10]. It provides over 12 hours of discrete gameplay and comprises (1) an engaging role-playing narrative in which players make choices for 6 characters (3 female, 3 male) as they proceed through adolescence and observe short- and long-term consequences in the characters’ lives; (2) mini-games to reinforce knowledge and skills; and (3) My Story, in which players set life goals, reflect on how the game relates to their lives, customize a personal avatar, and collect rewards. The narrative uses branching logic for 40 possible long-term outcomes (“epilogues”) across the 6 characters and is designed to be replayed, allowing players to experience different outcomes based on their in-game decisions. Players must assume the role of characters different from themselves in terms of sex and serostatus, thereby challenging stigma and harmful gender norms.

In a randomized feasibility study with 60 adolescents in Kenya, the intervention arm showed significant gains in sexual health-related knowledge and self-efficacy and behavioral intention for risk avoidance strategies and sexual risk communication when compared with the control arm 6 weeks postintervention [17]. Intervention-arm participants spent on average over 50% longer playing the game than instructed, and quantitative and qualitative data on user engagement and on game appeal, relevance, and acceptability among adolescents and parents were extremely positive [18].

A randomized controlled trial with a larger sample size (n=996) and longer follow-up (45 months) was undertaken with the objective of determining the intervention’s effects on (1) age at sexual debut and condom use at first sex and (2) related proximal outcomes (behavioral intentions, self-efficacy, attitudes, and knowledge). Tumaini demonstrated significant impact on the primary trial outcome (high-risk sexual debut vs low-risk or no sexual debut), attributable to increased condom use at first sex, at end line in intent-to-treat analyses [19].

XSL•FO
RenderX

Significant effects were also observed across a wide range of other outcomes at the end line.

Participants completed a behavioral survey at 13 time points over the 45-month duration of the trial, with a mean participant age of 14.0 to 17.7 (SD 0.56) years. The survey included items measuring a range of theory-based proximal outcomes related to condom use. These data and their longitudinal trajectories are the focus of the analyses presented in this manuscript. As the primary behavioral outcomes are related to sexual debut and developmentally influenced, hence statistically underpowered prior to end line, measures of proximal outcomes are of particular importance to understanding the effects of the intervention between trial baseline and end line. The purpose of this manuscript is to leverage these longitudinal data to evaluate Tumaini's impact on proximal outcomes related to condom use across 45 months of mid-adolescence, a key developmental phase, while assessing differential effects by gender.

Methods

Study Setting

Kenya has the 7th largest HIV burden globally and an HIV epidemic that is generalized throughout the population. Kisumu County has Kenya's second-highest prevalence (15.5%) and annual number of new infections [20]. The study took place in Kisumu Town, Kenya's third largest city, where youth aged 15-24 years comprise 48% of new HIV infections [21].

Study Design

This was a 45-month, 2-arm, individually randomized controlled trial with 996 participants, with a mean age of 14.0 (SD 0.6) years at baseline [22]. The study took place from October 2020 (start of enrollment) to August 2024 (last quantitative data collection). Participants were randomized 1:1 to play either Tumaini or a commercially available attention-control math game, Brainilis (Appgeneration – Software Technologies, Lda). They were provided with a low-cost Android smartphone loaded with their assigned game at mean participant age 14.0 (SD 0.56), 15.3 (SD 0.55), and 16.0 (SD 0.56) years, that is, between baseline data collection (T1) and second data collection time point (T2), 6th data collection time point (T6) and 7th data collection time point (T7), and 8th data collection time point (T8) and 9th data collection time point (T9). Gameplay periods lasted 5-7 weeks and took place during school holidays; they were unevenly spaced due to COVID-19–related disruptions to the school calendar. Participants completed behavioral surveys at 13 time points (baseline plus 12 follow-up surveys).

Recruitment and Randomization

Participant recruitment was conducted through a combination of school- and community-based strategies in urban and periurban Kisumu. Eligibility criteria for participation were being aged 12-14 years at the time of enrollment, being a resident in Kisumu Town, having basic English literacy (Grade 3-4 on the Flesch-Kincaid Reading Scale), and having no prior exposure to Tumaini-related activities.

Recruitment slots were used to balance participant gender, age, and residence (in the 16 wards that comprise Kisumu East, West, and Central). Randomization to the study arms was conducted by a blinded member of the data analysis team with no contact with participants. The success of randomization was confirmed through calculation and cross-arm comparison of descriptive statistics for adolescent demographics.

Due to the nature of the intervention, participants were not blinded to group assignment, which was revealed following completion of all baseline activities. Data analysts remained blinded to study arm assignment until completion of all primary outcome data analysis.

Data Collection and Analysis

Behavioral surveys were self-administered at 13 time points (T1-T13; 13th data collection time point [endline]), via tablet-based Open Data Kit. Participants completed surveys with headphones in a private space. Access to audio of all questions was provided to maximize understandability. Survey items were designed to evaluate the impact of exposure to the intervention on a range of primary and secondary outcomes. The study's primary outcomes assessed experience of sexual debut and condom use at first sex separately and also combined them in a composite binary variable of high-risk sexual debut (no condom use or "I don't know") vs low-risk sexual debut (condom use) or no risk (not yet initiated sex). Sex was defined as penile-vaginal ("when a boy or man puts his penis inside a girl's or woman's vagina"). It was specified that condoms were understood to be male condoms, defined as "a thin rubber covering worn by a man on his erect penis when he has sex," due to the low availability of female condoms (or "internal condoms"). Endline outcome analyses are presented elsewhere [19].

Survey items also assessed theory-based and intervention-targeted proximal influences, namely knowledge, attitudes, self-efficacy, and behavioral intentions, of these primary outcomes. The behavioral survey instrument was adapted from the version used in the Tumaini feasibility study [17]. Individual measures were drawn from existing instruments, with priority given to those previously used with sub-Saharan African youth populations [23-25]. The chosen items were adapted for age, linguistic, and cultural appropriateness and consistency of formatting, and supplemented with additional measures where necessary. In the interest of age appropriateness, some questions were framed as hypothetical risk scenarios ("*Imagine that*…").

Here, we focus on the longitudinal trajectories of the proximal outcomes related to condom-use behavior. These outcomes, their exact wording, and response options (binary and 3 or 4-point Likert scales) are presented in Table 1 (additional items are detailed in Table S1 in Multimedia Appendix 1). Each item was scored on a 0-1 scale, with higher values indicating more desirable or, in the case of knowledge, correct answers. This standardization of scores facilitated comparison across items. Mean scores were calculated for individual items at each time point. Only participants who had responded to the question at that time point were included in mean score calculations; missing values were not imputed. Only those participants

XSL•FO
RenderX

reporting never having had sex were asked about their intention to use a condom at first sex.

Table 1. Condom-related proximal outcomes, behavioral survey question wording, and response options for Tumaini efficacy trial.

Condom-related proximal outcome	Question wording	Response options
Behavioral intention		
Intention to use a condom at first sex	"When you have sex for the first time, do you plan to use a condom?"	Definitely yes, maybe yes, maybe no, definitely no
Intention to talk to a partner about HIV prevention	"Imagine that either in the future or now you are thinking about having sex with someone and you are not ready to become a parent. […] Would you talk with him/her about preventing HIV?"	Definitely yes, maybe yes, maybe no, definitely no
Self-efficacy		
Self-efficacy to use a condom correctly	"I am sure that I could use a condom correctly."	I strongly agree, I agree, I disagree, I strongly agree
Self-efficacy to refuse unprotected sex	"Imagine that, either in the future or now, you are in a relationship. He/ she says she wants to have sex with you, but she does not want to use a condom. […] If I did not want to have sex with him/her without a condom, I am sure that I could say no firmly."	I strongly agree, I agree, I disagree, I strongly agree
Attitudes		
Endorsement that condom use is a sign of respect	"Using a condom is a sign that you respect your partner"	I strongly agree, I agree, I disagree, I strongly agree
Endorsement of gender parity in condom initiation	"A woman can suggest using condoms just like a man can"	I strongly agree, I agree, I disagree, I strongly agree
Knowledge		
Knowledge of how to use a condom correctly	"Do you know how to use a condom correctly?"	Yes, no
Knowledge that condoms are effective for HIV prevention	"Are condoms an effective way to prevent HIV?"	Yes, no, I don't know

We used generalized estimating equations [26] to model these proximal outcome variables measured repeatedly over time and to facilitate statistical inferences about cross-arm (intervention vs control) mean differences at specific time points. Specifically, the generalized estimating equations model controlled for participants' baseline age (centered at the average value; due to the success of randomization, we did not control for any other covariates) and included indicator variables for arm (intervention as reference) and time (T1–T13, with T1 as reference), together with the 12 group-by-time interaction terms. This approach accounts for the correlated nature of the data, using an exchangeable working correlation structure that was selected via initial comparisons of alternatives based on the quasi-likelihood information criterion [27]. The granular nature of the model provides a close fit to the sample means of the outcome for each arm at each time, while facilitating direct inferences about the change in the mean outcome relative to baseline for the intervention arm at each time point (T2–T13) and the cross-arm differences in these mean changes. Of primary interest in these analyses are Bonferroni-corrected (significance level .05/4) inferences about the cross-arm differences in mean changes relative to baseline at 4 specific time points (T2, T7, T9, and T13). The inferences at T2 and T13 compare the average initial and late (end line) change in the mean response relative to baseline for the intervention and control arms. The inferences at T7 and T9 are designed to investigate potential enhancements in the cross-arm differences in change relative to baseline at the 2 time points following reintroduction of the intervention. Secondary to these cross-arm comparisons, we also report Bonferroni-corrected (significance level .05/12) inferences to assess mean within-arm changes relative to baseline at each of the 12 subsequent time points (T2–T13). These analyses were conducted overall and separately for female and male participants.

All analyses were conducted as intent-to-treat and were completed using SAS (version 9.4; SAS Institute Inc).

Ethical Considerations

The Institutional Review Board of Emory University (IRB00108404) and the Scientific and Ethics Review Unit of the Kenya Medical Research Institute (KEMRI; SERU: KEMRI/SERU/CGHR/11/3812) approved this study. Written informed consent was secured from one parent or guardian, and the participant provided their written assent prior to study baseline. To ensure participant privacy, datasets were deidentified; all participants were identified by a randomly assigned study identifier, which was the only identifier used in analysis datasets and to link data across time points and sources. Data were stored and transferred on encrypted devices. For each study visit, adolescent participants received KShs 500 (~US $5) as compensation for their time.

XSL•FO
RenderX

Results

Overview

A participant flowchart is shown in Figure 1, and baseline characteristics of the sample are presented in Table 2. Participants represented a range of socioeconomic and food security profiles. At end line, 97.8% of the sample (n=974) had been retained. There were no adverse events associated with study participation.

Figure 1. Trial flow diagram of study activities and participant totals for 45-month efficacy trial of a smartphone game (Tumaini) to increase age and condom use at first sex among young Kenyans.

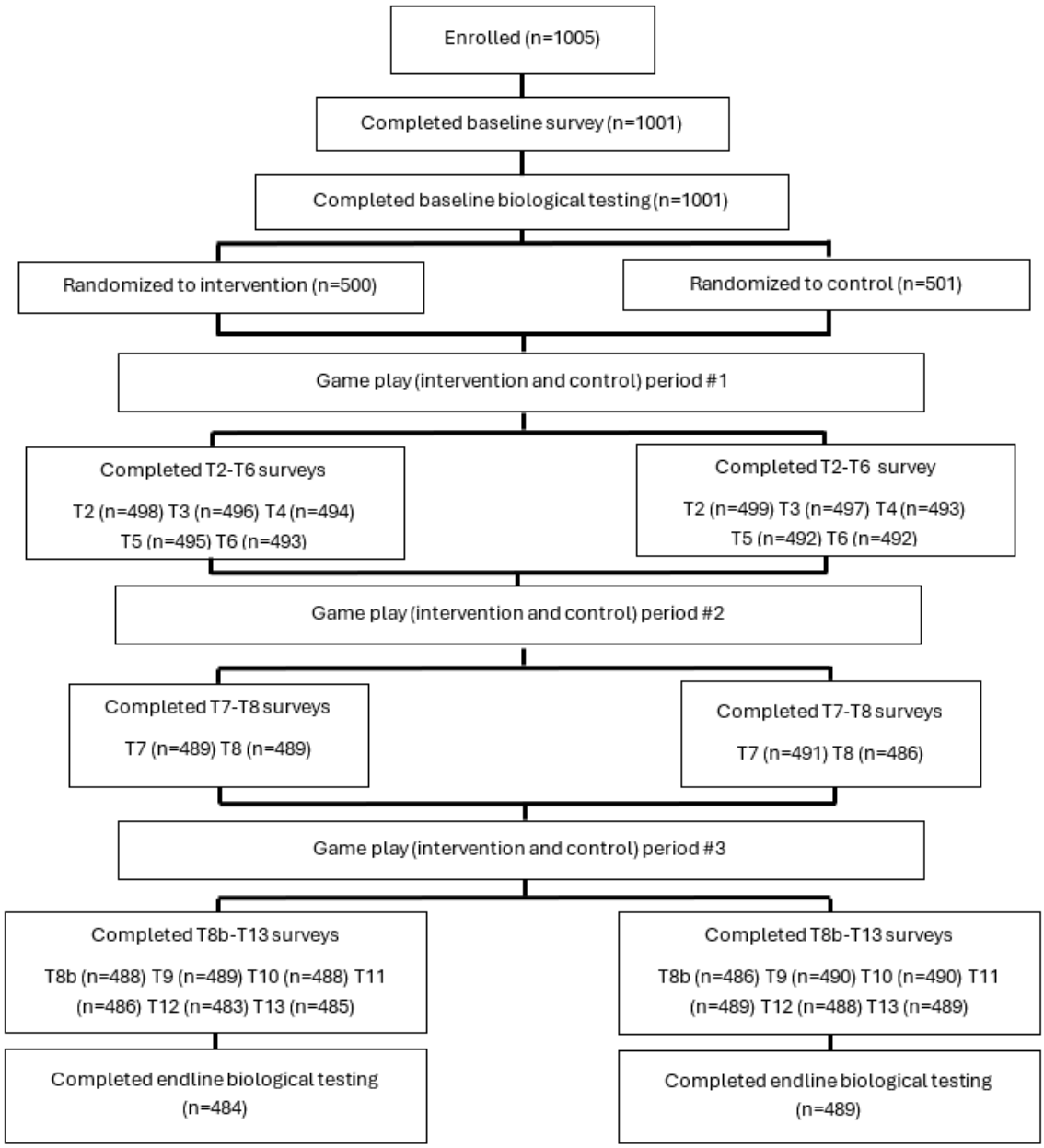

XSL•FO
RenderX

Table 2. Baseline characteristics of the intent-to-treat population (n=996) in the Tumaini efficacy trial.

Demographic characteristics	Intervention (n=499)	Control (n=497)
Sex, n (%)		
Female	249 (50)	250 (50)
Male	250 (50)	247 (50)
Age (years), mean (SD)	13.97 (0.57)	13.99 (0.54)
Enrolled in school, n (%)	499 (100)	493 (99)
Socioeconomic status, n (%)		
Lower socioeconomic status	241 (49)	241 (49)
Higher socioeconomic status	254 (51)	251 (51)
Food insecurity, n (%)		
None	270 (54)	260 (52)
Low	138 (28)	149 (30)
Medium	73 (15)	73 (15)
High	18 (4)	15 (3)
Living with HIV, n (%)	8 (2)	8 (2)
Experienced sexual debut, n (%)	76 (15)	76 (15)

Self-reported engagement in the intervention and control games was high at the end of each of the 3 intervention periods, and there was 100% exposure of each arm to their allocated game by T2. Participants in each arm dedicated comparable amounts of time to their assigned game across the 3 intervention periods, totaling 33.56 hours (SD 17.75) to Tumaini and 32.37 hours (SD 18.41) to the math game, indicating that the latter functioned as a highly effective attention-control condition. Fifty-eight percent of the control arm reported having played some of Tumaini by T9 (mean hours for those exposed: 19.77, SD 15.03) [19], with 35% exposed at T2 and 52% at T7.

We provide an overview of cross-arm differences in change since baseline at 4 time points (Table 3). Mean scores for each outcome at each time point and associated *P* values for within-arm change since baseline are presented in Tables S2-S9 in Multimedia Appendix 1.

XSL•FO
RenderX

Table 3. Cross-arm difference in mean change in proximal outcomes relative to baseline and associated *P* values at 4 time points in Tumaini efficacy trial.

Time point	T2[a]		T7[b]		T9[c]		T13[d]	
	Difference in change in mean score	*P* value	Difference in change in mean score	*P* value	Difference in change in mean score	*P* value	Difference in change in mean score	*P* value
Mean age (years)	14.1 (0.56)	_[e]	15.6 (0.55)	–	16.4 (0.56)	–	17.7 (0.56)	–
Behavioral intentions								
Intention to use a condom at first sex								
All	0.118	.005[f]	0.200	.008[f]	0.201	.01	0.164	.007[f]
Female	0.083	.02	0.088	.03	0.099	.02	0.103	.02
Male	0.055	.11	0.059	.11	0.045	.25	0.059	.14
Intention to talk to a partner about HIV prevention								
All	0.107	<.001[f]	0.085	<.001[f]	0.081	<.001[f]	0.076	<.001[f]
Female	0.115	<.001[f]	0.047	.11	0.048	.10	0.057	.05
Male	0.099	<.001[f]	0.123	<.001[f]	0.115	<.001[f]	0.095	.002[f]
Self-efficacy								
Self-efficacy to use a condom correctly								
All	0.215	<.001[f]	0.283	<.001[f]	0.273	<.001[f]	0.254	<.001[f]
Female	0.216	<.001[f]	0.278	<.001[f]	0.263	<.001[f]	0.283	<.001[f]
Male	0.213	<.001[f]	0.287	<.001[f]	0.282	<.001[f]	0.224	<.001[f]
Self-efficacy to refuse unprotected sex								
All	0.072	.006[f]	0.039	.13	0.042	.11	–0.037	.41
Female	0.107	.004[f]	0.028	.43	0.063	.09	0.040	.28
Male	0.037	.32	0.051	.17	0.021	.57	0.002	.96
Attitudes								
Endorsement that condom use is a sign of respect								
All	0.218	<.001[f]	0.201	<.001[f]	0.201	<.001[f]	0.164	<.001[f]
Female	0.234	<.001[f]	0.187	<.001[f]	0.201	<.001[f]	0.191	<.001[f]
Male	0.203	<.001[f]	0.214	<.001[f]	0.202	<.001[f]	0.138	<.001[f]
Endorsement of gender parity in condom initiation								
All	0.151	<.001[f]	0.205	<.001[f]	0.221	<.001[f]	0.192	<.001[f]
Female	0.146	<.001[f]	0.220	<.001[f]	0.218	<.001[f]	0.203	<.001[f]
Male	0.156	<.001[f]	0.189	<.001[f]	0.224	<.001[f]	0.181	<.001[f]
Knowledge								
Knowledge of how to use a condom correctly								
All	0.372	<.001[f]	0.429	<.001[f]	0.464	<.001[f]	0.458	<.001[f]
Female	0.376	<.001[f]	0.440	<.001[f]	0.462	<.001[f]	0.462	<.001[f]
Male	0.367	<.001[f]	0.417	<.001[f]	0.465	<.001[f]	0.453	<.001[f]
Knowledge that condoms are effective for HIV prevention								
All	0.096	.001[f]	0.093	.01[f]	0.085	.02	0.110	.003[f]

XSL•FO
RenderX

Time point	T2[a]		T7[b]		T9[c]		T13[d]	
	Difference in change in mean score	*P* value	Difference in change in mean score	*P* value	Difference in change in mean score	*P* value	Difference in change in mean score	*P* value
Female	0.116	.006[f]	0.123	.01	0.097	.05	0.158	.002[f]
Male	0.076	.07	0.059	.25	0.072	.17	0.062	.25

[a]T2: second data collection time point.

[b]T7: 7th data collection time point.

[c]T9: 9th data collection time point.

[d]T13: 13th data collection time point (end line).

[e]Not applicable.

[f]Indicates significance after Bonferroni correction (significance level .05/4, ie, .0125)

For each outcome, we present longitudinal plotlines below for all participants and stratified by gender. In the plots, solid lines track estimated mean responses (marked with triangles) for the intervention group over time; dashed lines (with circles) track estimated means for the control group. Solid markers (triangles for the intervention arm, circles for the control arm) indicate that the within-arm difference in mean response relative to baseline at that time point is significant after Bonferroni correction (ie, at α=.05/12 level); open symbols indicate lack of significance relative to baseline. Inferences addressing the 4 primary cross-arm comparisons of mean change relative to baseline are indicated with *P* values at the top of the plot. Asterisks after the *P* value indicate significance after Bonferroni correction (ie, at α=.05/4 level).

Behavioral Intentions

Intention to Use a Condom at First Sex

Figure 2 summarizes the results of the longitudinal analysis of the condom use at first sex behavioral intention outcome.

Cross-arm difference: when all participants were included in the analysis, there was significantly greater improvement in the intervention arm than the control arm immediately after initial intervention exposure (*P*=.005), following the second intervention period (ie, at T7; *P*=.008), and at end line (*P*=.007). Significance was marginal after Bonferroni correction following the third intervention period at T9 (*P*=.01) and not present when participants were stratified by gender, suggesting insufficient power once the sample was halved.

Within-arm change: solid triangles at T2-T13 for all participants indicate a Bonferroni-corrected statistically significant difference in the mean relative to baseline for the intervention arm at every subsequent time point. In addition to the steep uptick in mean score following initial intervention exposure, the intervention arm showed an uptick following the second and third intervention periods. In contrast, significant within-arm differences from baseline were present only occasionally in the control arm, with mean scores at end line closely resembling those at baseline.

Difference by gender: intention to use a condom at first sex was higher from baseline among male than female participants in both arms. Female intervention-arm participants saw a steeper uptick at T2 and significant improvement relative to baseline for all subsequent time points, such that it overtook the male control arm. After an initial upward trajectory, the control-arm plot trended downwards for female participants from T5 (age 14.8, SD 0.55).

Figure 2. Longitudinal trajectory of intention to use a condom at first sex at 13 timepoints among all, female, and male participants not reporting sexual debut.

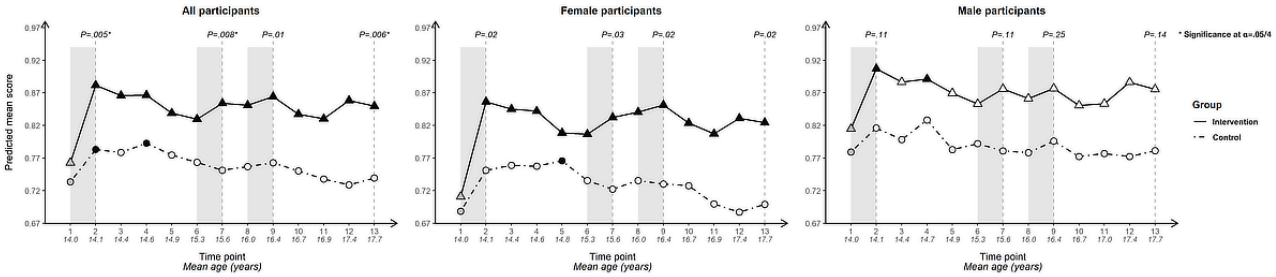

Intention to Talk to a Partner About HIV Prevention

Cross-arm difference: significant cross-arm differences in increased intention to talk to a partner about HIV prevention (Figure 3) were present at all 4 time points for all participants and for male participants.

XSL•FO
RenderX

Figure 3. Longitudinal trajectory of intention to talk to a partner about HIV prevention at 13 time points among all, female, and male participants.

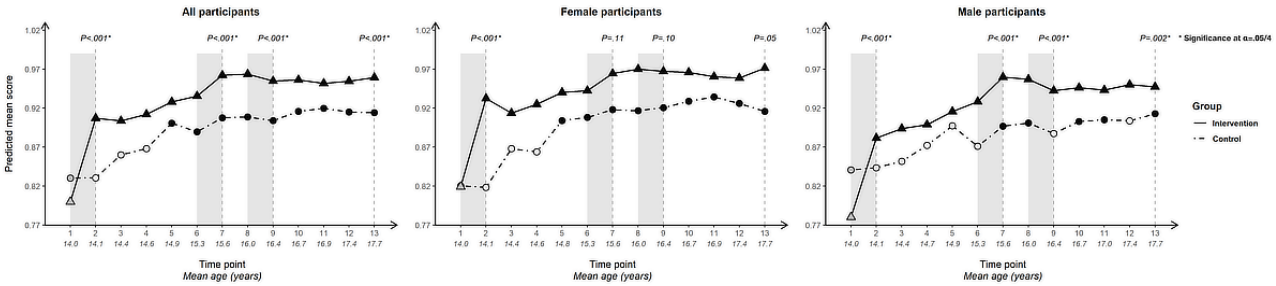

Despite an uptick in the intervention arm during the second intervention period, cross-arm differences were not significant at T7, T9, and end line for female participants. The plots suggest that this may be due to a combination of ceiling effects (mean scores in the intervention arm were extremely high from T7 [age 15.6, SD 0.55] onwards, limiting scope for improvement) and developmental gains in the control arm.

Within-arm change: for intervention-arm participants, within-arm change since baseline was significant for all time points for all female and male participants. For control-arm participants, difference relative to baseline mean became significant from T5 (age 14.9, SD 0.55) onwards for all and female participants. Male control-arm participants saw sporadic significant change since baseline from T7 (age 15.6, SD 0.56) onwards. Unlike intention to use a condom at first sex, mean scores followed an upward trajectory over time for both arms, suggesting a developmental effect.

A related outcome, self-efficacy to talk to a partner about HIV prevention (Figure S1 in Multimedia Appendix 2 and Table S10 in Multimedia Appendix 1), showed a similar trajectory to intention to talk to a partner about HIV prevention, with control-arm change since baseline for male participants achieving significance from mean age 14.9 (SD 0.56) years and end line values approximating those of the intervention arm.

Self-Efficacy

Self-Efficacy to Use a Condom Correctly

Cross-arm difference: baseline mean scores on self-efficacy to use a condom correctly (Figure 4) were lower and intervention-arm gains greater than for all other condom-related behavioral determinants except knowledge of correct condom use (which shows a similar trajectory, particularly in the intervention arm). Cross-arm differences in change since baseline were highly significant (P<.001) at all 4 time points for all participants and when stratified by gender.

Within-arm change from baseline was significant for intervention-arm participants at all subsequent time points, following a sharp uptick in mean score after the first intervention period and smaller upticks after the other intervention periods. For control-arm participants, a significant difference relative to baseline only occurred at one time point for each group: negatively at T7 (age 15.6, SD 0.55) for all and for female participants, and positively at end line for male participants.

Difference by gender: the overall upward trajectory among intervention-arm participants from T6 (age 15.4, SD 0.55) was steeper for male than for female participants. A similar upward trajectory was observed for male participants in the control arm, though at a much lower level, suggesting a developmental influence. The trajectory for female control-arm participants, by contrast, trended downwards.

Figure 4. Longitudinal trajectory of self-efficacy to use a condom correctly at 13 time points among all, female, and male participants.

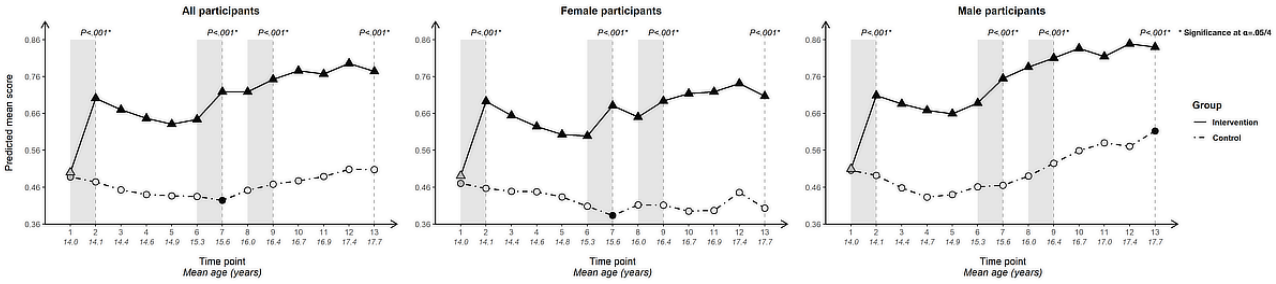

Self-Efficacy to Refuse Unprotected Sex

Cross-arm difference in increased self-efficacy to refuse unprotected sex (Figure 5) was significant following initial exposure to the intervention for all participants (P=.006) and for female participants (P=.004), but not for male participants (P=.32).

XSL•FO
RenderX

Figure 5. Longitudinal trajectory of self-efficacy to refuse unprotected sex at 13 time points among all, female, and male participants.

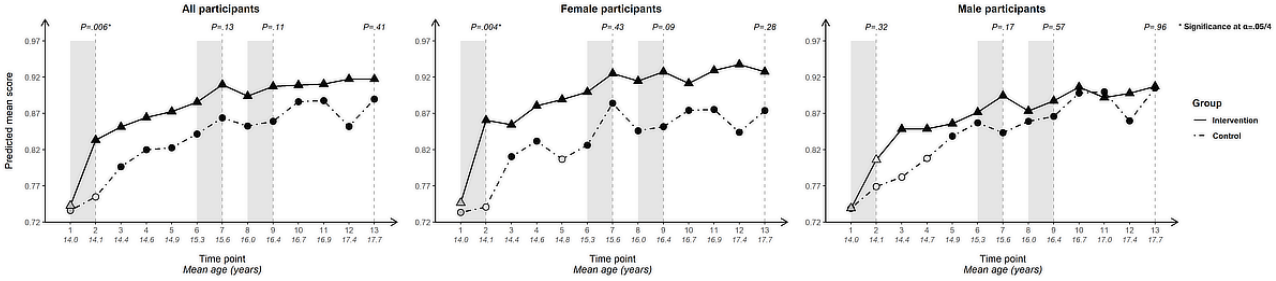

It was not significant at subsequent time points for any of the population groupings, despite upticks in the intervention arm following the second and third intervention periods. The upward trend in the intervention arm was mirrored by a similar trajectory, albeit at a generally lower level, in the control arm, suggesting a positive developmental influence for this outcome.

For the intervention arm, within-arm change since baseline was significant across all time points for all participants and when stratified by gender, except at T2 for male participants. The within-arm difference was significant for all control-arm participants from T3 (age 14.4, SD 0.56) onwards and for male participants from T5 onwards. For female participants, it was generally significant from T3 onwards.

Difference by gender: for male participants, the intervention-arm plotline was crossed by the control-arm plotline at T11 (age 17.0, SD 0.56; across all determinants, this was the only instance of this happening). End line values for both arms were nearly identical among male participants.

Attitudes: Endorsement of "Using a Condom Is a Sign You Respect Your Partner" and "A Woman Can Suggest Using Condoms Just Like a Man Can"

The survey included 2 condom-related attitude items, asking participants to indicate their degree of agreement with "using a condom is a sign that you respect your partner" (Figure 6) and "a woman can suggest using a condom just like a man can" (Figure 7).

Cross-arm difference in change since baseline was highly significant for endorsement of both condom-related attitude questions at each time point and for all population groupings.

Significant within-arm change since baseline was present in the intervention arm at every time point for both questions and for all population groupings. The intervention-arm uptick following initial exposure to Tumaini was more pronounced for the first, gender-neutral question, with moderate increases across subsequent time points. For the second question, which challenged traditional cultural norms around women initiating condom use, the longitudinal trajectory for the intervention arm showed a steady rise from T2 that was reflected at lower mean scores but a similar slope in the control arm, suggesting a more pronounced intersection of intervention and developmental influences. Both questions saw an uptick in intervention-arm scores following the second and third intervention periods.

Figure 6. Longitudinal trajectory of endorsement of "Using a condom is a sign that you respect your partner" at 13 time points for all, female, and male participants.

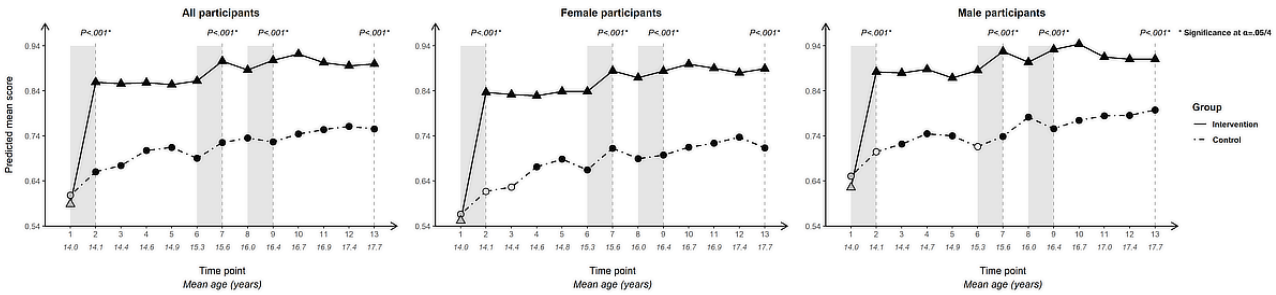

XSL•FO
RenderX

Figure 7. Longitudinal trajectory of endorsement of "A woman can suggest using condoms just like a man can" at 13 time points for all, female, and male participants.

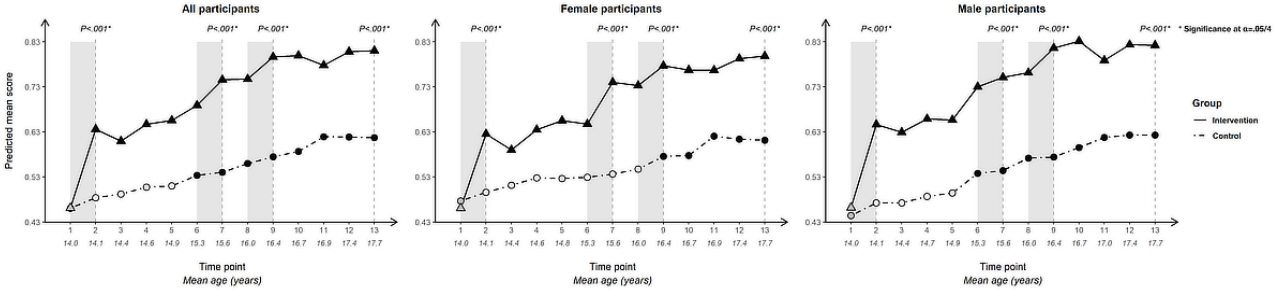

Knowledge

Knowledge of How to Use a Condom Correctly

The 2 condom-related knowledge questions show variations in trajectory. For knowledge of correct condom use (Figure 8), the cross-arm change in score since baseline differed highly significantly (P<.001) at all 4 time points and across all population groupings.

Baseline scores were the lowest for any condom-use determinant, particularly among female participants, but saw a spike in the intervention arm following initial exposure that was built upon at the second and third intervention periods.

The within-arm difference since baseline in the intervention arm was significant at all 12 post-baseline time points across all population groupings. The control arm only saw significance from T8 (age 16.0, SD 0.55; T10: age 16.7, SD 0.55, for female participants) onward, with a slight upward trend at later ages, suggesting a mild developmental influence.

Figure 8. Longitudinal trajectory of knowledge of correct condom use at 13 time points for all, female, and male participants.

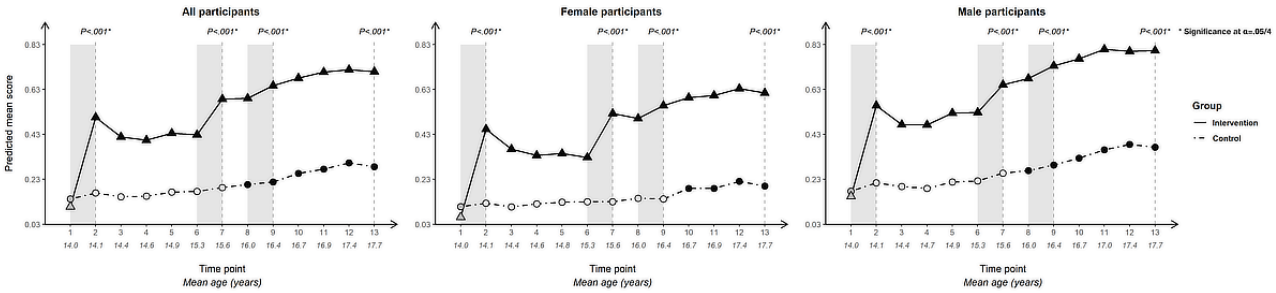

Knowledge That Condoms Are Effective at Preventing HIV

Cross-arm difference in change in knowledge that condoms are an effective way to prevent HIV (Figure 9) was significant following Bonferroni correction immediately after initial intervention exposure and at end line for all participants and for female participants, and at T7 for all participants. It was not significant at any time point for male participants, who had higher scores throughout than female participants.

The within-arm change since baseline was significant for the intervention arm at all time points for all participants.

This was true when stratifying by gender, except for at T5 and T6 for female participants and T3-T5 for male participants. In addition to a steep uptick after initial intervention exposure, the intervention arm saw upticks for the 2 subsequent intervention periods. This is particularly striking for female participants whose mean scores dip between intervention periods but recover following re-exposure to the intervention. Mean scores then stabilize with a slight upward trajectory from this point forward, while the control arm dips, indicating the value of repeated intervention exposure across mid-adolescence for female participants for this outcome. The control arm was nonsignificant for within-arm change since baseline at all time points except for male participants from T10 (age 16.7, SD 0.56) onward.

XSL•FO
RenderX

Figure 9. Longitudinal trajectory of knowledge that condoms are effective at preventing HIV at 13 time points for all, female, and male participants.

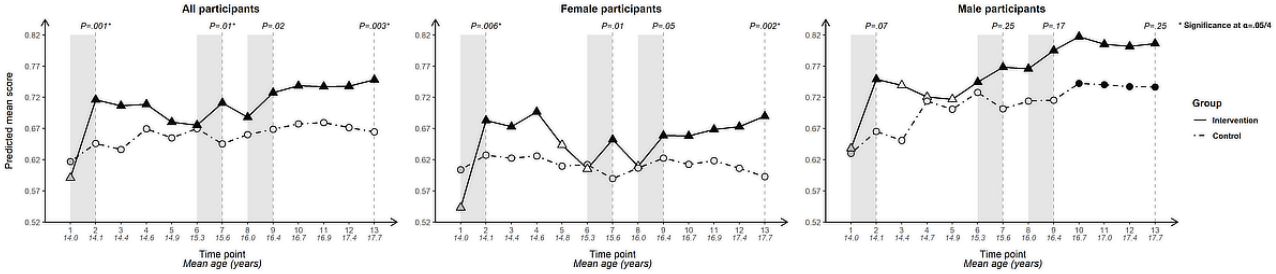

Discussion

Principal Findings

We evaluated the longitudinal impact of a theory-based choose-your-own-adventure smartphone game on proximal outcomes related to condom use across 45 months of mid-adolescence. These proximal outcomes are of particular importance given that assessment of primary behavioral outcomes, related to sexual debut, was not powered until end line, 45 months after participants' initial exposure to Tumaini and 16 months after their most recent exposure. Intervention-arm participants devoted over 30 hours in total to playing Tumaini, while the control-arm game attracted similar levels of engagement, thereby functioning as an effective attention-control condition.

Cross-arm difference: despite cross-arm contamination, Tumaini significantly improved all intervention-targeted proximal influences of condom use relative to the control arm immediately after initial exposure. For almost all proximal outcomes, a significant cross-arm difference was also present at end line and, for most outcomes, at the 2 intervening time points when cross-arm comparison was undertaken, following re-exposure to the intervention.

Within-arm difference: for the intervention arm, when all participants were included in the analysis, the within-arm mean change since baseline was significant for every outcome at every time point. All outcomes saw a spike in intervention-arm mean score immediately after initial exposure to Tumaini. Repeat exposure to the intervention at ages 15.3 (SD 0.55) and 16.0 (SD 0.55) years reinforced these effects in almost every case.

These significant findings and their timing relative to the intervention periods provide compelling evidence of Tumaini's impact and suggest that adolescents may benefit from even short-term exposure, though repeated exposure frequently sustained and strengthened effects. They are also consistent with our primary behavioral outcomes at end line; among participants reporting sexual debut between baseline and end line, 50% (41/82) of control-arm participants had unprotected first sex compared to 28% (23/81) of intervention-arm participants (relative risk [RR] 0.57, 95% CI 0.38-0.85) [19].

Multiple frameworks across health communication, social psychology, and behavioral science support a hierarchy of effects in behavioral interventions, for example [28], whereby knowledge is least resistant to change, followed by attitudes, self-efficacy, behavioral intentions, and finally behavior itself. In line with this hierarchy, while behavioral HIV prevention interventions for adolescents in sub-Saharan Africa have improved condom-related knowledge, attitudes, or self-efficacy, comparatively few have been shown in rigorous studies to reduce sexual risk by increasing condom use [5,29-31]. Our own findings are positive across the hierarchy of effects, from knowledge through attitudes, self-efficacy, and intentions to behavior.

Findings from the present analyses indicate that Tumaini's condom-related content was most effective at strengthening self-efficacy to use a condom (especially among male participants), which is highly dependent on knowledge of how to use a condom correctly, and at informing positive condom-related attitudes. While findings are extremely positive across the board, they also highlight areas where the intervention could potentially be revised to achieve even stronger effects. For example, they suggest that Tumaini could do more to strengthen male participants' self-efficacy to refuse unprotected sex. Trajectories for knowledge that condoms are an effective way to prevent HIV also differ substantially by gender. However, lower scores for female participants may relate less to condom efficacy, which is strongly emphasized throughout the intervention, than to the different levels of agency exercised by female versus male adolescents in condom use.

In addition to being the main strategy for HIV prevention for the population addressed in our trial, condoms are also the primary method of contraception [32]. Tumaini emphasizes the advantages of condoms for preventing pregnancy in addition to HIV. Pregnancy is an outcome that is a more immediate, and hence often more resonant, threat for adolescents—especially girls and young women—than HIV, even in high-prevalence parts of sub-Saharan Africa. Leveraging the long-term consequences of unplanned pregnancy for adolescents' future plans and dreams in condom promotion efforts also has the advantage of being less likely to contribute to HIV stigma. It is notable that trajectories for pregnancy-related proximal outcomes (Figures S2-S4 in Multimedia Appendix 2) were very similar to those for HIV prevention.

In their 2010 systematic review and meta-analysis of HIV prevention for sub-Saharan African youth, Michielsen et al [33] found that effects on condom use were larger in male participants in almost all studies where these were reported as significant. Gender stratification of our primary behavioral outcomes at end line, in contrast, revealed a stronger impact on female participants, for whom the risk of unprotected first sex was reduced significantly from 57% (21/37) in the control arm

XSL•FO
RenderX

to 21% (7/33) in the intervention arm (RR 0.37, 95% CI 0.18-0.76) [19]. Male participants also saw a reduction, but it was nonsignificant, from 44% (20/45) in the control arm to 33% (16/48) in the intervention arm (RR 0.75, 95% CI 0.45-1.26) despite them scoring higher in the present analyses at every time point for intention to use a condom at first sex. Intention may be lower for many adolescent girls and young women in Kenya because they envisage their sexual debut happening within marriage, a context in which they foresee less need to rely on a condom. Lower intervention impact on male participants' condom use at first sex suggests that the translation of intention into behavior may operate differently for male than for female adolescents. It is possible that lower self-efficacy to refuse unprotected sex may be a factor and/or that male participants struggle in the emotionally charged context of first sex when performance anxiety may be exacerbated by condom use.

Alongside those referenced above, multiple structural and contextual factors have been identified as potentially militating against translation of condom-use proximal outcomes up the hierarchy of effects, from knowledge to behavior. Sociocultural factors include, for example, the association of condoms with immorality; gender norms promoting "submissiveness" in adolescent girls and young women; perceptions among young men that condoms reduce sexual pleasure; and a range of myths and misconceptions [15,16,34,35]. Among behavioral theories, social cognitive theory, in which Tumaini is grounded, has the advantage of foregrounding the role played by the social environment in individual behavior. Tumaini's narrative-based approach, drawing on research on thousands of stories contributed by young Africans to scriptwriting competitions [15,16,36], ensures that contextual and structural factors are extensively engaged with.

In addition, although delivered at the individual level, Tumaini either supports or has the potential to support engagement at higher socioecological levels of analysis. It catalyzes considerable household (including parent-child) interaction [37], which is modeled and nurtured in the game by the presence of parent, other adults, and older peer characters who function as "virtual mentors." Further, the high levels of engagement it causes and sustains suggest that, when widely disseminated, its effects could extend to the community level with the potential to support normative change. Depending on population needs, moreover, its effects could be enhanced through dissemination in combination with structural or biomedical programming, alongside anticipated direct-to-consumer dissemination via an app store.

By 2030, smartphones will account for over 80% of all mobile connections in sub-Saharan Africa [7]. Although growth is uneven and there are cross-national, urban-rural, and gender disparities in mobile phone ownership generally and particularly for more expensive phones, penetration is especially rapid among young people [38,39]. At present, over 3 quarters of all mobile phones in Kenya are smartphones (compared to 61% across all of sub-Saharan Africa). While there is no gender disparity in ownership of mobile phones in general in Kenya, there is one for smartphones (49% of men compared to 43% of women own one). In rural areas, a combination of lower data access and cost limits smartphone ownership. Nonetheless, smartphone access and equity in access are increasing in Kenya and across sub-Saharan Africa, and this trajectory is unlikely to change in the near future.

With access barriers in mind, we built Tumaini for a low-cost smartphone as an app that does not need ongoing data access or regular updates. It is designed to be played by adolescents on the phones of parents, siblings, and other family members, or on their own phones or those of their friends; hence, players need access to a smartphone, but they do not need to own one themselves. At the study end line in 2024, 11% of efficacy trial participants (n=106) indicated that no one in their household owned a smartphone, and 17% (n=168) reported owning their own smartphone (though there was a gender disparity, with 21% of boys compared to 14% of girls owning one). Tumaini's scalability and impact outside of efficacy trial conditions warrants further investigation (and ongoing investigation as smartphone access increases).

Strengths and Limitations

Some longitudinal observational studies have followed cohorts of young Africans through adolescence, tracking sexual and reproductive health outcomes [40,41]. However, longitudinal intervention evaluations are rare. An exception is "Let Us Protect Our Future!" with follow-ups at 3, 6, 12, 42, and 54 months [42,43]. Our efficacy trial is notable for the length and frequency of follow-up (13 surveys over 45 months), in addition to its extremely high levels of participant retention (97.5% at end line), its effective attention-control condition, and its high and sustained intervention engagement.

The longitudinal trajectories presented here allow us to observe Tumaini's influence on a range of theory-based, condom-related proximal outcomes across mid-adolescence and across the 3 intervention periods. The within-arm comparisons, with plotlines incorporating intervention periods, provide invaluable contextualization for the cross-arm comparative data and facilitate identification of gender-specific and developmental influences. These analyses thus provide richer insights into the effects of Tumaini, including how these may be optimized, than end line estimates of relative risk alone.

In addition to these strengths, the study has limitations. First, there is a risk that repeated participant exposure to the same survey items over time might blunt the sensitivity of those questions. However, cross-arm differences, upticks in intervention-arm mean scores at each intervention period, and diversity in longitudinal trajectory by outcome suggest that this was not a major concern in this study. Second, while we drew our survey items from scales that had been used and, where possible, also validated with similar populations, we were unable to use the complete validated scales due to the age of our participants (some as young as 12 at baseline) and the risk of overburdening them with an extensive questionnaire or exposing them to non–age-appropriate content. Third, by their nature, our proximal outcomes are self-reported. Specifically, self-efficacy to use a condom correctly and knowledge of how to use a condom correctly are self-assessments, and it is possible that these measures may not ultimately mediate correct condom use in real-life circumstances. Fourth, the study provided

XSL•FO
RenderX

participants with smartphones with parental restrictions blocking internet access; additional research is needed to assess Tumaini's effectiveness under more "real-world" conditions. We are planning a hybrid type 2 effectiveness-implementation study. Fifth, as parents self-selected to learn more about the study, the sample may be biased toward participants whose parents are more open to sexual and reproductive health education for adolescents. Finally, cross-arm contamination may have reduced our ability to detect some intervention effects.

Future Research

Ongoing qualitative research with this cohort may shed additional light on contextual factors informing these trajectories. Mediation analysis currently underway will allow us to pinpoint which proximal outcomes mediated our behavioral outcomes at end line for those who had reached sexual debut; there is no scope for mediation analysis at earlier time points because these outcomes are developmentally influenced, resulting in small cell sizes at younger participant ages. Through this additional research, we will seek to optimize Tumaini prior to our planned hybrid type 2 effectiveness-implementation trial.

Conclusions

Tumaini significantly improved theory-based proximal influences of condom use, with effects sustained 45 months post-initial exposure and 16 months post-most recent exposure. Spikes in intervention-arm scores immediately after initial exposure indicate that adolescents benefit from even short-term exposure, though repeated exposure frequently sustained and reinforced effects. At a time when threats to global HIV funding highlight the necessity of preventing new infections among young people, study findings support the promise of reaching adolescents at an early age with information and skills to support a safer sexual debut. In the context of rapidly increasing access to smartphones in sub-Saharan Africa and at a time when cost-effectiveness is at a premium, Tumaini has potential for high scalability and impact on condom-related outcomes.

Acknowledgments

The authors gratefully acknowledge our research and administrative teams in Kenya and the United States, and especially all communities and participants involved. We thank the Henry Jackson Foundation Medical Research Institute for their support, in particular, Drs Daniel Ochiel and Tiffany Hamm. We extend special thanks to Dr Victor Akelo and Dr Gary W Harper for their contributions to this study.

Funding

Research reported in this publication was supported by the National Institute of Mental Health of the US National Institutes of Health under Award Number 5R01MH118982 (PI: KW). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research was also supported by the Emory Center for AIDS Research (P30 AI050409). The sponsors played no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Data Availability

Data underlying these analyses are not publicly available as the analyses are ongoing. Data are available from the corresponding author on reasonable request.

Authors' Contributions

KW led the design of the study, and GS, KK, RAB, RHL, VM, and KO contributed to the design of the study. VM, KO, JA, and CM contributed to protocol implementation, data collection, and management. SG, HQ, and GS conducted data analyses, guided by RHL, KW, RAB, and KK. KW drafted the manuscript; all other authors edited and reviewed the manuscript. KW is the overall principal investigator, while VM is the site principal investigator for this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental tables with additional outcomes and mean outcome scores at 13 time points and associated within- and cross-arm change.
[DOCX File , 61 KB - jmir_v28i1e83982_app1.docx]

Multimedia Appendix 2
Longitudinal trajectories of additional survey items with relevance to condom use at 13 time points among all, female, and male participants.
[DOCX File , 669 KB - jmir_v28i1e83982_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 625 KB - jmir_v28i1e83982_app3.pdf]

Multimedia Appendix 4
Higher-resolution versions of figures 2-9.
[ZIP File (Zip Archive), 1339 KB - jmir_v28i1e83982_app4.zip]

References

1. The urgency of now: AIDS at a crossroads. Joint United Nations Programme on HIV/AIDS (UNAIDS). 2024. URL: https://www.unaids.org/en/resources/documents/2024/global-aids-update-2024?utm_source=chatgpt.com [accessed 2026-01-16]
2. Benevides R, Logan J, Patel P. Introduction to the youth HIV prevention supplement. AIDS Educ Prev 2023;35:1-3. [doi: 10.1521/aeap.2023.35.suppa.1]
3. Poobalan AS, Pitchforth E, Imamura M, Tucker JS, Philip K, Spratt J, et al. Characteristics of effective interventions in improving young people's sexual health: a review of reviews. Sex Educ 2009;9(3):319-336. [doi: 10.1080/14681810903059185]
4. Pedlow CT, Carey MP. Developmentally appropriate sexual risk reduction interventions for adolescents: rationale, review of interventions, and recommendations for research and practice. Ann Behav Med 2004;27(3):172-184 [FREE Full text] [doi: 10.1207/s15324796abm2703_5] [Medline: 15184093]
5. Shangase N, Kharsany ABM, Ntombela NP, Pettifor A, McKinnon LR. A systematic review of randomized controlled trials of school based interventions on sexual risk behaviors and sexually transmitted infections among young adolescents in sub-Saharan Africa. AIDS Behav 2021;25(11):3669-3686 [FREE Full text] [doi: 10.1007/s10461-021-03242-8] [Medline: 33772695]
6. Li DH, Macapagal K, Mongrella M, Saber R, Mustanski B. "Your package could not be delivered": the state of digital HIV intervention implementation in the US. Curr HIV/AIDS Rep 2024;21(3):152-167. [doi: 10.1007/s11904-024-00693-1] [Medline: 38502421]
7. The mobile economy sub-Saharan Africa. GSMA. 2024. URL: https://www.gsma.com/solutions-and-impact/connectivity-for-good/mobile-economy/sub-saharan-africa-2024/?utm_source=chatgpt.com [accessed 2026-01-16]
8. Hightow-Weidman L, Muessig K, Bauermeister J, LeGrand S, Fiellin L. The future of digital games for HIV prevention and care. Curr Opin HIV AIDS 2017;12(5):501-507 [FREE Full text] [doi: 10.1097/COH.0000000000000399] [Medline: 28692490]
9. DeSmet A, Baranowski T, Thompson D, Shirong Lu A. Fantasy, facts and fun: digital health games for impact and implementation. In: de Bruijn GJ, Vandebosch H, editors. Handbook on Health, Media and Communications. Berlin, Boston: De Gruyter Mouton; 2025:287-308.
10. Winskell K, Sabben G, Obong'o C. Interactive narrative in a mobile health behavioral intervention (Tumaini): theoretical grounding and structure of a smartphone game to prevent HIV among young Africans. JMIR Serious Games 2019;7(2):e13037 [FREE Full text] [doi: 10.2196/13037] [Medline: 31066692]
11. Bandura A. Social foundations of thought and action: a social-cognitive view. Englewood Cliffs, NJ: Prentice-Hall; 1986.
12. Sood S, Menard T, Witte K. The theory behind entertainment-education. In: Singhal A, Cody M, Rogers E, Sabido M, editors. Entertainment-Education and Social Change. Mahwah, NJ and London: Lawrence Erlbaum Associates; 2004.
13. Singhal A, Cody M, Roger E, Sabido M. Entertainment-Education and Social Change: History, Research, and Practice. Mahwah, NJ: Lawrence Erlbaum Associates; 2004.
14. Winskell K. Social representations theory and young Africans' creative narratives about HIV/AIDS, 1997-2014. J Theory Soc Behav 2021;51(1):164-182 [FREE Full text] [doi: 10.1111/jtsb.12270] [Medline: 34483357]
15. Winskell K, Obyerodhyambo O, Stephenson R. Making sense of condoms: social representations in young people's HIV-related narratives from six African countries. Soc Sci Med 2011;72(6):953-961 [FREE Full text] [doi: 10.1016/j.socscimed.2011.01.014] [Medline: 21388731]
16. Winskell K, Singleton R, Sabben G, Tiendrébéogo G, Obong'o C, Dia FL, et al. Social representations of the prevention of heterosexual transmission of HIV among young Africans from five countries, 1997-2014. PLoS One 2020;15(3):e0227878 [FREE Full text] [doi: 10.1371/journal.pone.0227878] [Medline: 32142515]
17. Winskell K, Sabben G, Akelo V, Ondeng'e K, Obong'o C, Stephenson R, et al. A smartphone game-based intervention (Tumaini) to prevent HIV among young Africans: pilot randomized controlled trial. JMIR Mhealth Uhealth 2018;6(8):e10482 [FREE Full text] [doi: 10.2196/10482] [Medline: 30068501]
18. Sabben G, Mudhune V, Ondeng'e K, Odero I, Ndivo R, Akelo V, et al. A smartphone game to prevent HIV among young Africans (Tumaini): assessing intervention and study acceptability among adolescents and their parents in a randomized controlled trial. JMIR Mhealth Uhealth 2019 May 21;7(5):e13049 [FREE Full text] [doi: 10.2196/13049] [Medline: 31115348]

XSL•FO
RenderX

19. Winskell K, Sabben G, Mbeda C, Ondeng'e K, Arego J, Qin H, et al. Efficacy of a smartphone game (Tumaini) to increase age and condom use at first sex among adolescents in Kenya: a randomised controlled trial. Lancet Child Adoles Health 2026;10:273-283. [doi: 10.1016/S2352-4642(25)00353-0]
20. Banadakoppa Manjappa R, Bhattacharjee P, Shaw SY, Gitonga J, Kioko J, Songok F, et al. A sub-national HIV epidemic appraisal in Kenya: a new approach for identifying priority geographies, populations and programmes for optimizing coverage for HIV prevention. J Int AIDS Soc 2024;27 Suppl 2(Suppl 2):e26245. [doi: 10.1002/jia2.26245] [Medline: 38982894]
21. Truong HM, Guzé MA, Ouma D, Bushman D, Mocello AR, Kadede K, et al. Community-based HIV testing for urban youth in Western Kenya. AIDS Behav 2022;26(3):814-821 [FREE Full text] [doi: 10.1007/s10461-021-03441-3] [Medline: 34518938]
22. Mudhune V, Sabben G, Ondenge K, Mbeda C, Morales M, Lyles RH, et al. The efficacy of a smartphone game to prevent HIV among young Africans: protocol for a randomized controlled trial in the context of COVID-19. JMIR Res Protoc 2022;11(3):e35117 [FREE Full text] [doi: 10.2196/35117] [Medline: 35030090]
23. Vandenhoudt H, Miller KS, Ochura J, Wyckoff SC, Obong'o CO, Otwoma NJ, et al. Evaluation of a U.S. evidence-based parenting intervention in rural Western Kenya: from parents matter! To families matter!. AIDS Educ Prev 2010;22(4):328-343. [doi: 10.1521/aeap.2010.22.4.328] [Medline: 20707693]
24. Learning from adolescents to prevent HIV and unintended pregnancy. In: Issues Brief (Alan Guttmacher Inst). New York, NY: Guttmacher Institute; 2007.
25. Ndebele M, Kasese-Hara M, Greyling M. Application of the information, motivation and behavioural skills model for targeting HIV risk behaviour amongst adolescent learners in South Africa. SAHARA J 2012;9 Suppl 1:S37-S47 [FREE Full text] [doi: 10.1080/17290376.2012.744903] [Medline: 23234378]
26. Liang K, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73(1):13-22 [FREE Full text] [doi: 10.1093/biomet/73.1.13]
27. Pan W. Akaike's information criterion in generalized estimating equations. Biometrics 2001;57(1):120-125. [doi: 10.1111/j.0006-341x.2001.00120.x] [Medline: 11252586]
28. McGuire W. Theoretical foundations of campaigns. In: Rice R, Atkin C, editors. Public communication campaigns. Newbury Park, CA: Sage; 1989:43-65.
29. Mwale M, Muula AS. Systematic review: a review of adolescent behavior change interventions [BCI] and their effectiveness in HIV and AIDS prevention in sub-Saharan Africa. BMC Public Health 2017;17(1):718 [FREE Full text] [doi: 10.1186/s12889-017-4729-2] [Medline: 28923040]
30. Muthoni CN, Kneipp SM, Gichane MW, Caiola CE, Pettifor AE, Williams JR. A systematic review of HIV interventions for young women in sub-Saharan Africa. AIDS Behav 2020;24(12):3395-3413. [doi: 10.1007/s10461-020-02914-1] [Medline: 32410052]
31. Sani AS, Abraham C, Denford S, Ball S. School-based sexual health education interventions to prevent STI/HIV in sub-Saharan Africa: a systematic review and meta-analysis. BMC Public Health 2016;16(1):1069 [FREE Full text] [doi: 10.1186/s12889-016-3715-4] [Medline: 27724886]
32. Kenya demographic and health survey. KNBS (Kenya National Bureau of Statistics). 2022. URL: https://www.knbs.or.ke/reports/kdhs-2022/?utm_source=chatgpt.com [accessed 2026-01-16]
33. Michielsen K, Chersich MF, Luchters S, De Koker P, Van Rossem R, Temmerman M. Effectiveness of HIV prevention for youth in sub-Saharan Africa: systematic review and meta-analysis of randomized and nonrandomized trials. AIDS 2010;24(8):1193-1202. [doi: 10.1097/QAD.0b013e3283384791] [Medline: 20375876]
34. Casmir E, Daniel AK, Ongolly F, Thuo N, Oluoch L, Kiptinness C, et al. Protection at first sexual intercourse among adolescent girls and young women in Kenya. Arch Sex Behav 2021;50(1):219-227 [FREE Full text] [doi: 10.1007/s10508-020-01785-w] [Medline: 32720186]
35. Tavrow P, Karei EM, Obbuyi A, Omollo V. Community norms about youth condom use in Western Kenya: is transition occurring? Afr J Reprod Health 2012;16(2):241-252. [Medline: 22916556]
36. Winskell K, Beres LK, Hill E, Mbakwem BC, Obyerodhyambo O. Making sense of abstinence: social representations in young Africans' HIV-related narratives from six countries. Cult Health Sex 2011;13(8):945-959 [FREE Full text] [doi: 10.1080/13691058.2011.591431] [Medline: 21787256]
37. Winskell K, Sabben G, Ondeng'e K, Odero I, Akelo V, Mudhune V. A smartphone game to prevent HIV among young Kenyans: household dynamics of gameplay in a feasibility study. Health Educ J 2019;78(5):595-606 [FREE Full text] [doi: 10.1177/0017896919832344] [Medline: 34219796]
38. The mobile gender gap report. GSMA. 2024. URL: https://www.gsma.com/r/gender-gap/ [accessed 2026-01-16]
39. Kenya demographic and health survey 2022. Kenya National Bureau of Statistic and ICF. 2023. URL: https://www.knbs.or.ke/reports/kdhs-2022/ [accessed 2026-01-20]
40. Richter L, Norris S, Pettifor J, Yach D, Cameron N. Cohort profile: Mandela's children: the 1990 birth to twenty study in South Africa. Int J Epidemiol 2007;36(3):504-511 [FREE Full text] [doi: 10.1093/ije/dym016] [Medline: 17355979]
41. Baird S, Chirwa E, McIntosh C, Özler B. The short‐term impacts of a schooling conditional cash transfer program on the sexual behavior of young women. Health Economics 2010;19(S1):55-68. [doi: 10.1002/hec.1569]

XSL•FO
RenderX

42. Jemmott JB, Jemmott LS, O'Leary A, Ngwane Z, Icard LD, Bellamy SL, et al. School-based randomized controlled trial of an HIV/STD risk-reduction intervention for South African adolescents. Arch Pediatr Adolesc Med 2010 Oct;164(10):923-929 [FREE Full text] [doi: 10.1001/archpediatrics.2010.176] [Medline: 20921349]
43. Jemmott JB, Jemmott LS, O'Leary A, Ngwane Z, Lewis DA, Bellamy SL, et al. HIV/STI risk-reduction intervention efficacy with South African adolescents over 54 months. Health Psychol 2015;34(6):610-621 [FREE Full text] [doi: 10.1037/hea0000140] [Medline: 25110841]

Abbreviations

KEMRI: Kenya Medical Research Institute
RR: relative risk

Edited by A Stone; submitted 15.Sep.2025; peer-reviewed by P Kreniske, S Naar; comments to author 30.Oct.2025; revised version received 19.Dec.2025; accepted 30.Dec.2025; published 10.Mar.2026.

Please cite as:
Winskell K, Sabben G, Qin H, Mbeda C, Goldenberg S, Ondeng'e K, Ndivo R, Arego J, He X, Bednarczyk RA, Komro K, Lyles RH, Mudhune V
Longitudinal Effects of a Smartphone Game (Tumaini) for HIV Prevention Among Kenyan Adolescents: 45-Month Trajectories of Condom Use–Related Proximal Outcomes From a Randomized Controlled Trial
J Med Internet Res 2026;28:e83982
URL: https://www.jmir.org/2026/1/e83982
doi:10.2196/83982
PMID:

©Kate Winskell, Gaëlle Sabben, Haowen Qin, Calvin Mbeda, Sophie Goldenberg, Ken Ondeng'e, Richard Ndivo, Judith Arego, Xinwei He, Robert A Bednarczyk, Kelli Komro, Robert H Lyles, Victor Mudhune. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Effect of Text Messaging on the Postoperative Pain Experience in Pediatric Patients Undergoing Thoracic Surgery: Randomized Controlled Trial

Yun Shi*, MD; Qianqiu Wang*, BA; Aihua Liu, BA; Li Jiang, MD

Department of Anesthesiology, Children's Hospital of Fudan University, 399 Wanyuan Road, Minhang District, Shanghai, China

*these authors contributed equally

Corresponding Author:
Li Jiang, MD
Department of Anesthesiology, Children's Hospital of Fudan University, 399 Wanyuan Road, Minhang District, Shanghai, China

Abstract

Background: Inadequately controlled postoperative pain continues to pose a significant clinical challenge in pediatric patients undergoing thoracic surgery.

Objective: This randomized controlled study aimed to investigate the effectiveness of SMS-based educational support for postoperative pain management on patients' pain experience.

Methods: A total of 100 pediatric patients undergoing thoracic surgery were enrolled between December 2, 2023, and January 28, 2025. Patients in the intervention group (group 1) received structured postoperative pain management education via SMS text messages, whereas those in the control group (group 2) received standard oral education. Pain intensity and pain-related interference were assessed using the Brief Pain Inventory, and self-efficacy was measured using the Pain Self-Efficacy Scale.

Results: The number of patients with moderate-to-severe average pain was significantly lower in group 1 than in group 2 (n=19, 18% vs n=19, 38%; *P*=.04), and group 1 had significantly higher patient self-efficacy scores (mean 29.3, SD 7.5 vs mean 25.2, SD 8.7*; P*=.01). Least pain scores were lower in group 1 compared with group 2 across all 3 consecutive postoperative days (PODs): POD1 (mean 2.8, SD 0.8 vs mean 3.3, SD 0.7; *P*<.001), POD2 (mean 2.2, SD 0.8 vs mean 2.7, SD 0.5*; P*<.001), and POD3 (mean 1.7, SD 0.7 vs mean2.2, SD 0.4; *P*<.001). Similarly, average pain intensity was lower in group 1 compared with group 2 across all 3 consecutive PODs: POD1 (mean 3.7, SD 0.9 vs mean 4.7, SD 0.6; *P*<.001); POD2 (mean 3.3, SD 0.8 vs mean 3.6, SD 0.5; *P*=.01); and POD3 (mean 2.5, SD 0.8 vs mean 3.1, SD 0.5; *P*<.001). General activity was significantly less affected in group 1 on POD 1 (mean 4.3, SD 1.0 vs mean 5.0, SD 1.5; *P*=.004) and POD 2 (mean 3.1, SD 0.7 vs mean 3.7, SD 1.3; *P*=.009).

Conclusions: The use of an SMS-based educational intervention significantly improved the postoperative pain experience of pediatric patients undergoing thoracic surgery. Further research is needed to clarify its impact on clinical outcomes and to better understand the mechanisms underlying improved pain management.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2300076554; https://www.chictr.org.cn/showprojEN.html?proj=203355

(J Med Internet Res 2026;28:e81806) doi:10.2196/81806

KEYWORDS

SMS; thoracic surgery; pain; pain self-efficacy; pediatric; text messaging

Introduction

Thoracic surgery is typically associated with substantial postoperative pain. To optimize analgesia, a multimodal pain management strategy is widely adopted, combining nonsteroidal anti-inflammatory drugs such as ibuprofen [1] and paracetamol [2], regional anesthesia [3], and a patient-controlled analgesia pump (PCAP) [4]. Within this framework, both patients and caregivers are encouraged to participate actively in postoperative pain management, including pain assessment and analgesic administration via a PCAP. Despite these efforts, pain following thoracic surgery often remains inadequately controlled. Previous studies have reported that approximately 54% of pediatric patients experience moderate-to-severe pain on the day of surgery, and both patients and parents frequently express dissatisfaction with postoperative pain management [5]. Pain has a profound impact on the quality of postoperative recovery. As a negative and inherently subjective experience, pain is perceived directly by children. Inadequate analgesia compromises effective breathing, coughing, and moving, thereby increasing the risk of respiratory complications and delaying

XSL•FO
RenderX

postoperative recovery [6,7]. Furthermore, insufficient control of acute postoperative pain has been associated with an elevated risk of developing chronic pain after discharge [8]. Several factors may contribute to suboptimal pain control in children, even in the presence of a PCAP. These include difficulties in accurately assessing pain severity, uncertainty regarding the appropriate timing of analgesic administration, challenges in adhering to medication regimens, and concerns regarding potential side effects of analgesics [9-11]. Collectively, these barriers may be attributed to insufficient educational information and/or ineffective educational support for patients and their caregivers [12].

With the rapid development of mobile technologies, internet-based medical support has emerged as an increasingly prominent component of health care delivery. The effectiveness of digital health interventions in enhancing self-management among pediatric patients with chronic diseases has been well documented [13,14]. These approaches emphasize patient-centered self-care, with health care professionals facilitating self-management by providing informational, emotional, and social support. Recently, emerging studies have investigated the use of SMS text messaging in pediatric pain management, with most focusing on pain assessment and monitoring. The feasibility and acceptability of SMS-based interventions have been widely demonstrated, as reflected by high response rates [15-17], strong caregiver acceptance [18], improved medication adherence, and reduced anxiety levels [19]. Several studies suggest that participants demonstrate improved pain coping and reduced anxiety and depression [20]. However, the effect of SMS on pain outcomes, particularly pain severity, remains unclear. More broadly, evidence regarding the impact of digital health tools, including mobile apps and other implementation strategies, on pediatric pain severity is limited and inconsistent. Some studies have reported improved pain self-management among pediatric patients [16,21-23], whereas others have found no significant benefit in pain control [20,24]. Notably, one study reported greater pain severity among patients in the intervention group [25]. Consequently, the use of SMS for postoperative pain management in pediatric populations remains in its early stages. Further research is warranted to evaluate the effects of SMS-based interventions on postoperative pain outcomes in pediatric populations.

In this prospective randomized controlled trial, an intervention group (group 1), which received SMS-based educational support for postoperative pain management, was compared with a control group (group 2), which received conventional educational support, to evaluate the effectiveness of SMS in helping manage the postoperative pain experience in pediatric patients undergoing thoracic surgery. The hypothesis was that pediatric patients receiving SMS-based educational support would demonstrate improved pain coping and reduced postoperative pain intensity.

Methods

Study Participants

Pediatric patients who underwent elective thoracic surgery at the Children's Hospital of Fudan University were consecutively recruited for this randomized controlled trial. Eligible patients were (1) aged 7 to 18 years, (2) classified as American Society of Anesthesiologists (ASA) physical status I or II, (3) expected to receive a combination of a preoperative paravertebral block and a PCAP for postoperative pain management, and (4) had access to a smartphone (iOS or Android). Exclusion criteria included (1) chronic pain (defined as pain persisting or recurring for more than 3 months), (2) a history of psychiatric disorders, (3) planned video-assisted thoracoscopic surgery, (4) inability to assess pain severity using a numeric rating scale (NRS), (5) communication difficulties, (6) any prior surgeries, (7) postoperative transfer to the intensive care unit (ICU), or (8) refusal to participate.

Study Design

Patients who met the inclusion criteria were consecutively enrolled during a preoperative visit. Participants were assigned identification numbers in the order of recruitment. An independent statistician generated the random allocation sequence using SPSS software (version 20; IBM Corp). The allocation sequence was placed in sealed envelopes, which were opened only after participant enrollment. Separate investigators were responsible for patient enrollment and follow-up. All patients underwent surgery under general anesthesia. Due to the study design, participants were not blinded to the intervention in either group, but they did not have information about each other. The statistician remained blinded to group allocation.

Procedure

Participants received comprehensive educational information on postoperative pain management. The information covered the cause of postoperative pain, advantages and disadvantages of PCAP, indications for PCAP, operational principles of PCAP, self-assessment of pain, pain management goals, proper use of PCAP, nonpharmacological pain relief methods, scheduled medication reminders, and guidance on early postoperative mobilization. Group 1 received information through a combination of text, illustrative images (eg, pain assessment scales), and instructional videos (eg, demonstrating how to operate the PCAP). Concise (Multimedia Appendix 1) and detailed (Multimedia Appendix 2) versions of the educational content were provided to facilitate participant understanding. The concise version emphasized practical operation and workflow, whereas the detailed version provided more comprehensive explanations to support a deeper understanding. The information was sent out manually on the day prior to surgery. Group 2 received conventional oral education support that provided the same information. All participants were instructed on the importance of self-monitoring and managing postoperative pain to promote engagement with pain self-management. Participants were encouraged to actively assess and manage their own pain, with parental assistance permitted as needed. They were advised to contact health care providers whenever necessary. However, participants in group 1 were not allowed to communicate with the investigators via SMS.

XSL•FO
RenderX

Outcomes and Measures

All patients underwent surgery under general anesthesia. A preoperative paravertebral block was performed, and a PCAP was initiated upon discharge from the postoperative anesthesia care unit. Demographic data, including age, sex, place of residence, and parental educational level, were collected prior to surgery. Patients were followed for the first 3 postoperative days (PODs) by a research nurse at 24, 48, and 72 hours after surgery. At each time point, the cumulative dose of hydromorphone administered via PCAP was retrieved from the system and recorded. If pain was inadequately controlled, additional rescue analgesic doses were permitted, and dosages were verified through patient medical records. As hydromorphone doses were calculated according to body weight, the cumulative dosage was analyzed and compared on a per-day, per–body weight basis.

Patient-reported pain intensity and pain interference were assessed over the 3 consecutive PODs using the Brief Pain Inventory (BPI) questionnaire. The BPI, recommended by the National Cancer Institute, is used to evaluate pain severity using an NRS ranging from 0 (no pain) to 10 (worst pain), capturing the highest, lowest, and average pain experienced. It also assesses the impact of pain on daily functioning, including activities, mood, sleep, interpersonal relationships, and enjoyment of life. Consequently, the BPI is widely used in clinical practice [26] and research to provide detailed information on pain experience [27-29]. Pain self-efficacy refers to a patient's belief and confidence regarding their ability to effectively manage pain and attain desired outcomes. Pain self-efficacy is recognized as an important resilience factor for pain, particularly in children and adolescents [30]. The Pain Self-Efficacy Scale (PSES) [31] was used to evaluate the patient's confidence in managing pain. The scale consists of 10 items, each scored from 1 to 6, and was administered both before surgery and on POD3. Patient satisfaction with postoperative pain management was evaluated on POD3 using an NRS (0=not satisfied at all, 10=very satisfied). Analgesic-related complications—including postoperative nausea and vomiting, urine retention, constipation, severe sedation, and respiratory depression—were also recorded.

The primary outcome of the study was the incidence of moderate-to-severe average pain on POD1, defined as an NRS score of 4 or greater. Secondary outcomes included cumulative hydromorphone consumption, pain experiences and their impact on daily life, pain self-efficacy, analgesic-related complications, and patient satisfaction.

Sample Size Estimation

A pilot study conducted at the Children's Hospital of Fudan University reported that the incidence of moderate-to-severe average pain on POD1 was 39.5%. A clinically meaningful reduction was defined as a one-third decrease in this incidence with the implementation of SMS for postoperative pain management. Assuming 80% power and a significance level of .05, the minimum required sample size was calculated as 45 patients per group. To account for an anticipated 10% loss to follow-up, the sample size was increased to 50 patients per group.

Data Analysis

Only participants who completed the entire questionnaire were included in the analysis. Data were analyzed using GraphPad Prism (version 9; Graphpad Corp) and SPSS software (version 20; IBM Corp). Continuous variables were tested for normality. Normally distributed data were expressed as means with SDs and compared using independent *t* tests. Conversely, nonnormally distributed data were expressed as medians with IQRs and compared using the Mann-Whitney *U* test. Categorical data were expressed as numbers and percentages and analyzed using the chi-square test. Univariate association between the incidence of moderate-to-severe average pain and PSES score was analyzed using the Poisson correlation.

Ethical Considerations

The study was conducted at the Children's Hospital of Fudan University and approved by the research ethics board (202328; approval date June 20, 2023; chairperson Li-Ling Qian). The study was registered at Chictr.org.cn (ChiCTR2300076554). Eligible patients who met the inclusion criteria and their parents were approached in a separate interview room 1 day before surgery. Detailed explanations of the study objectives, methods, potential benefits, and risks were provided. Participants were informed that their medical data would be reported publicly, but no personally identifiable information would be included. They were assured that participation would not affect the quality of care and that withdrawal from the study was permitted at any time. All the responders volunteered to participate free of charge. Signed informed consent was obtained from the parents of participants who agreed to participate in the study.

Results

Study Population

From December 2, 2023, to January 28, 2025, a total of 183 patients were assessed for eligibility. Of these, 134 patients met the inclusion criteria, and 100 (76.4%) agreed to participate. None of the patients were excluded from the analysis owing to missing data or loss to follow-up (Figure 1). No significant differences were observed between the groups in terms of age, sex, weight, diagnosis, and parental educational level (Table 1).

XSL•FO
RenderX

Figure 1. Study flow diagram.

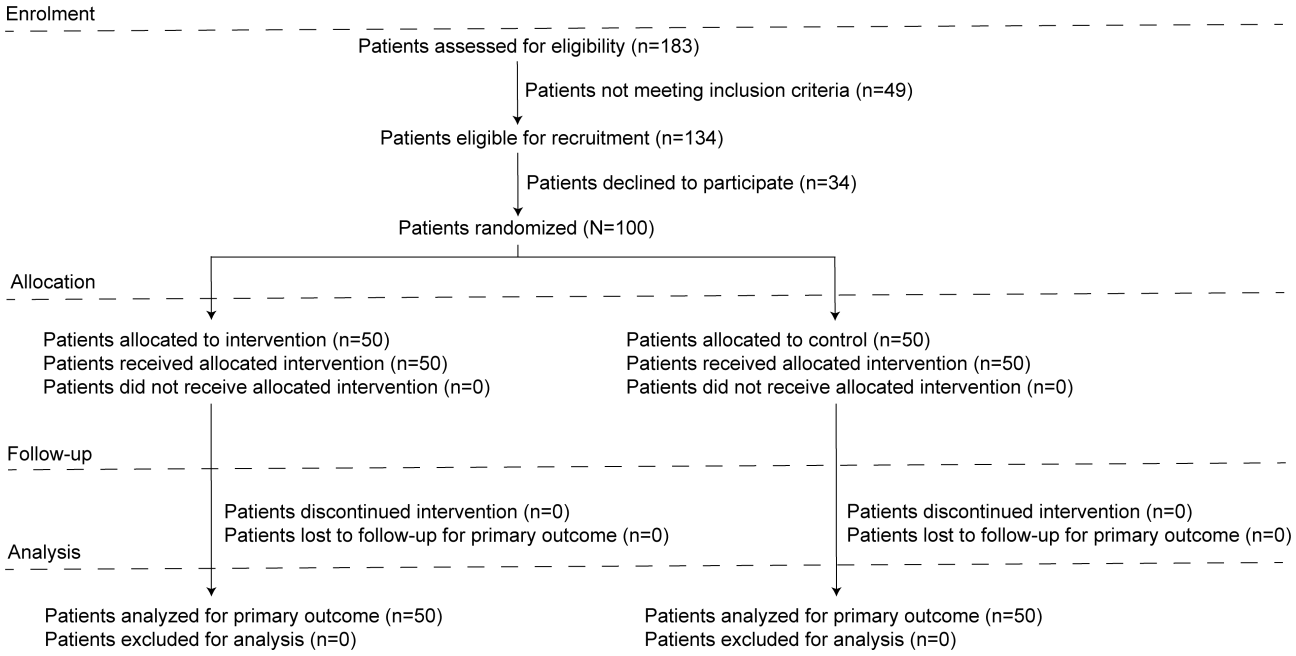

XSL•FO
RenderX

Table . Participant characteristics (N=100).

		Intervention group (n=50)	Control group (n=50)	P value
Age (y), mean (SD)		10.7 (2.8)	11.1 (2.9)	.48
Gender, n (%)				>.99
	Male	33 (66)	32 (64)	
	Female	17 (34)	18 (36)	
Weight (kg), mean (SD)		39.9 (11.7)	40.1 (14.2)	.94
Parent education, n (%)				.61
	Middle school	8 (16)	10 (20)	
	High school	10 (20)	12 (24)	
	Associate degree	18 (36)	15 (30)	
	Undergraduate degree	12 (24)	13 (26)	
	Master’s degree and above	2 (4)	0 (0)	
Diagnosis, n (%)				.17
	Rib lesions	3 (6)	2 (4)	
	Diaphragmatic lesions	3 (6)	2 (4)	
	Pectus excavatum/carinatum	12 (24)	24 (48)	
	Mediastinal mass	16 (32)	12 (24)	
	Pulmonary lesions	16 (32)	10 (20)	
History of surgery, n (%)				.68
	Yes	22 (44)	19 (38)	
	No	28 (56)	31 (62)	
Experience with PCAP, n (%)				.15
	Yes	15 (30)	8 (16)	
	No	35 (70)	42 (84)	
PCAP operator, n (%)				.69
	Patient	0 (0)	0 (0)	
	Parent	22 (44)	25 (50)	
	Both	28 (56)	25 (50)	

[a]PCAP: patient-controlled analgesia pump.

Primary Outcome

The incidence of moderate-to-severe average pain on POD1 was significantly lower in group 1 than in group 2 (n=9, 18% vs n=19, 38%; *P*=.04), whereas no difference was detected between the groups on POD2 and POD3 (Table 2).

XSL•FO
RenderX

Table . Pain experience among pediatric patients after thoracic surgery (N=100).

	Postoperative day 1			Postoperative day 2			Postoperative day 3		
	Group 1 (n=50)	Group 2 (n=50)	*P* value	Group 1 (n=50)	Group 2 (n=50)	*P* value	Group 1 (n=50)	Group 2 (n=50)	*P* value
Primary outcome, n (%)									
Patients with moderate-to-severe average pain, n (%)	9 (18)	19 (38)	.04	0 (0)	1 (2)	>.99	0 (0)	1 (2)	>.99
Secondary outcomes									
NRS[a] score, mean (SD)									
Worst pain	6.4 (1.1)	6.1 (0.7)	.58	5.3 (0.9)	5.2 (0.7)	.40	4.7 (1.1)	4.7 (0.6)	.80
Least pain	2.8 (0.8)	3.3 (0.7)	<.001	2.2 (0.8)	2.7 (0.5)	<.001	1.7 (0.7)	2.2 (0.4)	<.001
Average pain	3.7 (0.9)	4.7 (0.6)	<.001	3.3 (0.8)	3.6 (0.5)	.01	2.5 (0.8)	3.1 (0.5)	<.001
Patients reporting pain trends, n (%)			.005			.83			.03
Decreasing	11 (22)	2 (4)		35 (70)	34 (68)		41 (82)	30 (60)	
Increasing	20 (40)	34 (68)		0 (0)	0 (0)		0 (0)	0 (0)	
Unchanged	19 (38)	14 (28)		15 (30)	16 (32)		9 (18)	20 (40)	
Pain interference with daily functioning (0-10 scale), mean (SD)									
General activities	4.3 (1.0)	5.0 (1.5)	.004	3.1 (0.7)	3.7 (1.3)	.009	2.6 (1.3)	2.5 (0.7)	.69
Mood	4.3 (2.4)	3.8 (1.7)	.30	2.4 (0.8)	2.8 (2.1)	.19	1.8 (1.2)	1.7 (0.7)	.65
Walking	9.9 (0.4)	9.9 (0.3)	.72	8.8 (0.9)	8.9 (0.6)	.52	7.4 (1.4)	8.2 (0.7)	<.001
Sleeping	3.9 (1.8)	4.1 (1.3)	.50	2.3 (1.0)	2.8 (0.7)	.11	1.3 (0.8)	2.2 (0.5)	<.001
Relationship with others	1.9 (1.2)	2.3 (1.0)	.14	1.3 (1.0)	1.5 (0.6)	.39	0.7 (0.8)	1.0 (0.4)	<.001
Enjoyment of life	2.4 (1.1)	2.7 (1.0)	.09	1.6 (0.9)	1.8 (0.5)	.12	1.0 (0.01)	1.2 (0.4)	.01
Cumulative consumption of hydromorphone (μg/kg/d)	43.9 (8.3)	38.7 (7.0)	<.001	33.1 (10.8)	32.8 (6.8)	.86	28.7 (9.6)	27.6 (6.2)	.53

XSL•FO
RenderX

	Postoperative day 1			Postoperative day 2			Postoperative day 3		
	Group 1 (n=50)	Group 2 (n=50)	*P* value	Group 1 (n=50)	Group 2 (n=50)	*P* value	Group 1 (n=50)	Group 2 (n=50)	*P* value
Patients with analgesic-related complications, n (%)			.23			>.99			>.99
Postoperative nausea and vomiting	29 (58)	21 (42)		5 (10)	4 (8)		0 (0)	0 (0)	
Other	0 (0)	0 (0)		0 (0)	0 (0)		0 (0)	0 (0)	

[a]NRS: numeric rating scale.

Secondary Outcomes

Pain experiences—including pain intensity, treatment, analgesic consumption, and complications—and their impact on patients' daily lives are summarized in Table 2. Specifically, the intensity of least pain was significantly lower in group 1 versus group 2 on each of the 3 consecutive days: POD1 (mean 2.8, SD 0.8 vs mean 3.3, SD 0.7; *P*<.001), POD2 (mean 2.2, SD 0.8 vs mean 2.7, SD 0.5; *P*<.001), and POD3 (mean 1.7, SD 0.7 vs mean 2.2, SD 0.4; *P*<.001). Similarly, average pain was lower in group 1 versus group 2 across the 3 consecutive days: POD1 (mean 3.7, SD 0.9 vs mean 4.7, SD 0.6; *P*<.001), POD2 (mean 3.3, SD 0.8 vs mean 3.6, SD 0.5; *P*=.01), and POD3 (mean 2.5, SD 0.8 vs mean 3.1, SD 0.5; *P*<.001).

By contrast, no differences were observed in the intensity of worst pain between the groups. Although NRS scores for worst, least, and average pain decreased significantly over time in both groups, a significantly higher proportion of patients in group 1 than in group 2 reported pain relief on POD1 (n=11, 22% vs n=2, 4%; *P*=.005) and POD 3 (n=41, 82% vs n=30, 60%; *P*=.03). The impact of pain on daily activities was significantly lower in group 1 than in group 2 on POD1 (mean 4.3, SD 1.0 vs mean 5.0, SD 1.5; *P*=.004) and POD2 (mean 3.1, SD 0.7 vs mean 3.7, SD 1.3; *P*=.009). No significant between-group differences were observed in the impact of pain on mood, walking, sleeping, relationships with others, and enjoyment of life, except on POD3. Cumulative hydromorphone consumption on POD1 was significantly higher in group 1 than in group 2 (mean 43.9, SD 8.3 vs mean 38.7, SD 7.0; *P*<.001).

Patient self-efficacy for postoperative pain was evaluated using the PSES (Table 3). No significant differences were observed between the 2 groups prior to surgery; however, the PSES score was significantly higher in group 1 than in group 2 on POD3 (mean 29.3, SD 7.5 vs mean 25.2, SD 8.7; *P*=.01). No significant differences were found between the 2 groups in patient satisfaction on POD3 (mean 8.0, SD 1.6 vs mean 7.7, SD 1.9; *P*=.40) or in the duration of the hospital stay (mean 8.4, SD 0.97 vs mean 8.1, SD 0.89 days; *P*=.11). To explore the correlation between pain self-efficacy and pain severity, a univariate logistic regression analysis was performed. The incidence of moderate-to-severe pain was not significantly associated with the PSES score (*P*=.59). No differences were detected in the incidence of analgesic-related complications across the 3 PODs. The recorded adverse events consisted primarily of postoperative nausea and vomiting (Table 2), all of which resolved spontaneously.

Table . Patient Self-Efficacy Scale scores in postoperative pain assessment.

	Group 1	Group 2	*P* value
Presurgery, mean (SD)	26.8 (7.8)	24.4 (7.7)	.12
Postoperative day 3, mean (SD)	29.3 (7.5)	25.2 (8.7)	.01

Discussion

Principal Findings

Overall, this study demonstrated that the number of patients with moderate-to-severe average pain on POD1 after thoracic surgery in pediatric patients was significantly lower in group 1 than in group 2 following the implementation of SMS-based education (n=9, 18% vs n=19, 38%). Furthermore, both the lowest and average NRS scores decreased over the first 3 PODs, and a greater proportion of patients in group 1 than in group 2 reported a trend of decreasing pain (n=11, 22% vs n=2, 4%). Collectively, these findings indicate that the implementation of SMS-based education led to an improvement in postoperative pain management in pediatric patients. Notably, group 1 not only exhibited lower absolute NRS scores than group 2, but these scores also reflected a clinically relevant difference in pain severity: group 1 experienced mild pain, whereas group 2 experienced moderate pain (mean 3.7, SD 0.9 vs mean 4.7, SD 0.6). Together with the primary outcome showing that significantly more patients in group 2 experienced moderate-to-severe average pain, these improvements in pain experience are clinically meaningful. This interpretation is further supported by the significantly reduced impact of pain on general activities observed in group 1.

XSL•FO
RenderX

Group 1 demonstrated a higher cumulative consumption of analgesics on POD1, indicating that this group received more rescue analgesic doses through PCAP. This finding may partly explain the observed improvement in postoperative pain control in participants. Owing to the limited time available for preoperative education and the difficulty in effectively explaining or retaining detailed information, most patients and parents have an incomplete understanding of pain assessment and the goals of pain management [9,10,32]. Moreover, concerns regarding potential side effects often lead parents to be overly cautious about analgesic administration [11,33,34], resulting in a tendency to encourage children to tolerate pain rather than to administer analgesics proactively. Consequently, even with the PCAP, pain relief in pediatric patients may remain inadequate [5,35]. By contrast, parents and patients in group 1 were able to consult mobile push messages at any time and receive ongoing educational support when questions arose. This access facilitated more accurate assessment of pain intensity and timely administration of rescue analgesic doses in accordance with medical staff expectations.

Pain self-efficacy may represent another important determinant of pain experienced in this study. Although evidence regarding the impact of pain self-efficacy on postoperative pain in pediatric patients remains limited, studies in adults have demonstrated that enhancing pain self-efficacy can reduce postoperative pain intensity [36] and improve pain-related outcomes and patients' overall recovery quality [37]. Moreover, studies conducted in pediatric patients with chronic pain have consistently reported beneficial effects of pain self-efficacy on children's pain and quality of life [38-40]. Patients who are more confident in their pain management strategies typically experience less pain and distress. Accordingly, the lower pain intensity and improved general activity observed in group 1 may, at least in part, be attributable to enhanced pain self-efficacy, which is consistent with the findings of previous studies. However, this study did not detect a causal relationship between pain self-efficacy and pain severity. This may reflect the multifactorial nature of pain and be related to the relatively small sample size. Consequently, pain self-efficacy could not be confirmed as a variable independently associated with pain severity in this study. Further studies are warranted to clarify the underlying mechanisms.

Finally, pain is a highly subjective experience influenced not only by pharmacological and nonpharmacological analgesic strategies for analgesia but also by various other factors [41]. Previous studies have reported that placebo effects can alleviate pain by enhancing patients' expectations of beneficial outcomes [42]. In this study, given the ubiquity of mobile phone use, the participants in group 1 could easily access relevant information anytime and anywhere, which may have strengthened their expectations of effective pain control. Moreover, because parents assume primary responsibility for postoperative pain management in children, their cognition, attitudes toward pain, and psychological states—such as helplessness, stress, depression, or confidence—inevitably influence children's perceptions of pain. Parental responses to children's pain and their sense of security have been shown to affect pediatric pain experiences. Accordingly, the observed changes in pain experience in this study may be partly attributable to the sense of security experienced by both parents and children, knowing that guidance and support were readily accessible through SMS.

Although general activity was significantly better in group 1 than in group 2, no significant between-group differences were observed in other domains of pain interference. Although group 1 reported better sleep, social functioning, and enjoyment of life compared with group 2 on POD3, scores in both groups remained at a satisfactorily low level. Therefore, further studies are warranted to determine whether these differences have a significant impact on clinical outcomes despite more favorable subjective evaluations. In addition, the highest NRS scores in both groups remained above 4 points over the 3 PODs, with no significant between-group differences. Similarly, walking ability scores in both groups remained above 5 points during this period, although group 1 performed significantly better than group 2 on POD3. These findings suggest that pain during movement was inadequately controlled in both groups. This limitation may explain the absence of a significant difference in overall patient satisfaction despite improvements in average pain intensity and general activity in group 1. It may also be related to the relatively low analgesic consumption observed among study participants compared with that reported in previous studies [43,44]. Further studies are needed to optimize pain management and achieve more effective postoperative analgesia in patients.

Overall, the data indicate that implementing SMS for postoperative pain management in pediatric patients undergoing thoracic surgery significantly enhanced pain self-efficacy, effectively reduced pain intensity, and improved daily functioning. These findings indicate the importance of patients' subjective experiences in clinical practice and highlight the role of interventions that increase patient engagement and confidence in pain self-management. This study provides a framework for progressing interventions from feasible and acceptable to effective. Given the limitations of existing research, future studies should examine a broader range of clinical scenarios and more diverse patient populations. This could include exploring different modalities (eg, SMS, mobile apps, and web-based platforms), varying disease contexts (eg, chronic pain, cancer, and postoperative pain), different care settings (eg, hospital and home), and relevant comorbidities (eg, depression and anxiety) to validate findings and enhance generalizability. Such efforts would support the translation of research findings and facilitate the clinical implementation of high-quality SMS interventions for pain management in pediatric populations.

Study Limitations

This study has some limitations that should be considered when interpreting the results. First, the study was conducted at a single center, the Children's Hospital of Fudan University, one of the largest pediatric medical centers in an economically developed region of China. Parents of patients at this center generally have a higher socioeconomic status, better level of education, and greater compliance with medical directives. Consequently, the single-center design and relatively small sample size may limit the generalizability of the findings. Second, the study focused solely on the impact of SMS text messages on postoperative acute pain management in pediatric patients classified as ASA

XSL•FO
RenderX

physical status I or II. Inclusion of patients with ASA physical status III or higher could have broadened the applicability of the results. However, at the Children's Hospital of Fudan University, the pediatric patients were transferred to the intensive care unit, where more frequent and detailed pain assessments and interventions are performed by medical staff. Such intensive management would inevitably influence patients' self-management and the study outcomes, thereby justifying the exclusion of patients with more severe conditions.

Second, there is a significant concomitant relationship between pain and psychiatric disorders, with complex interactions. For example, pediatric patients with attention-deficit/hyperactivity disorder are more prone to pain [45], which may be related to increased pain sensitivity [46]. Pain and pain-related dysfunction may be further exacerbated in patients with depression and anxiety [47], while patients with autism spectrum disorders may exhibit abnormal pain perception [48]. Thus, pain management for pediatric patients with comorbid psychiatric disorders is a distinct and complex topic; however, it was not the focus of this study. Therefore, children with psychiatric disorders were excluded from this study. Owing to variations in pain characteristics and patient populations, the results obtained in this specific cohort may not be generalizable to broader populations. However, the findings can serve as a reference for calculating sample sizes in future studies.

Third, self-reporting bias is another consideration. Pain, the primary outcome, was assessed using self-reported measures, which may be influenced by memory errors, social desirability bias, or comprehension differences. Randomization was used during the study design to ensure balanced baseline characteristics and minimize selection bias. Although participants were not blinded to group allocation, data collectors were blinded, reducing the risk of researcher expectation bias and its potential impact on participants' reported outcomes. During the data collection phase, validated standardized questionnaires and scales were used to quantify the participants' self-reported results, ensuring consistency in data collection and reducing differences in subjective interpretation. Furthermore, pain experience was evaluated from multiple perspectives. In addition to the primary outcome of moderate-to-severe pain incidence, secondary assessments included quantified pain scores, daily pain intensity trends, pain-related functional impairment, and quality of life measures, such as general activity, mood, and sleep. Continuous follow-up over 3 PODs enabled a comprehensive assessment across multiple dimensions and reduced reliance on single self-reports. Participants also received thorough education regarding research objectives and reporting procedures, thereby enhancing compliance and mitigating single-report bias. However, reporting bias could not be entirely eliminated.

Finally, the study did not investigate the underlying causes of pain and/or contributing factors, including individual pain perception, nonpharmacological strategies, psychological status, or the interaction of analgesic use and pain experience, which limited insight into the mechanisms influencing the observed outcomes.

Conclusions

The implementation of SMS-based postoperative pain management in pediatric patients undergoing thoracic surgery improved clinical outcomes. However, its effects on patient outcomes and the underlying mechanisms require further investigation.

Acknowledgments

We would like to thank Jun Xia for enrolling participants, Peng Shi for assisting in data analysis, and all the nursing staff in the thoracic surgery ward for recording patient-reported complications. No generative artificial intelligence was used in any portion of the manuscript preparation.

Funding

The authors declared no financial support was received for this work.

Data Availability

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy and ethical restrictions.

Authors' Contributions

Conceptualization: LJ, YS
Data curation: QW, YS
Formal analysis: AL, QW
Investigation: QW, AL
Methodology: YS
Project administration: LJ
Supervision: LJ, YS
Validation: LJ, YS
Writing—original draft: QW
Writing—review and editing: LJ, YS

XSL•FO
RenderX

The manuscript was approved by all authors, and the order of authors listed in the manuscript has been approved by all authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example of the educational information sent to participants in group 1.
[PNG File, 337 KB - jmir_v28i1e81806_app1.png]

Multimedia Appendix 2
Detailed version of the information sent to participants, using self-assessment methods for pain (translated to English).
[PNG File, 272 KB - jmir_v28i1e81806_app2.png]

Checklist 1
CONSORT-eHEALTH checklist (V1.6.1).
[PDF File, 870 KB - jmir_v28i1e81806_app3.pdf]

References

1. Poddighe D, Brambilla I, Licari A, Marseglia GL. Ibuprofen for pain control in children: new value for an old molecule. Pediatr Emerg Care 2019 Jun;35(6):448-453. [doi: 10.1097/PEC.0000000000001505] [Medline: 29912084]
2. Macario A, Royal MA. A literature review of randomized clinical trials of intravenous acetaminophen (paracetamol) for acute postoperative pain. Pain Pract 2011;11(3):290-296. [doi: 10.1111/j.1533-2500.2010.00426.x] [Medline: 21114616]
3. He Y, Xu M, Li Z, Deng L, Kang Y, Zuo Y. Safety and feasibility of ultrasound-guided serratus anterior plane block and intercostal nerve block for management of post-sternotomy pain in pediatric cardiac patients: a prospective, randomized trial. Anaesth Crit Care Pain Med 2023 Dec;42(6):101268. [doi: 10.1016/j.accpm.2023.101268] [Medline: 37364851]
4. El-Boghdadly K, Levy NA, Fawcett WJ, et al. Peri-operative pain management in adults: a multidisciplinary consensus statement from the Association of Anaesthetists and the British Pain Society. Anaesthesia 2024 Nov;79(11):1220-1236. [doi: 10.1111/anae.16391] [Medline: 39319373]
5. Tomaszek L, Fenikowski D, Cież-Piekarczyk N, Mędrzycka-Dąbrowska W. Maximum pain at rest in pediatric patients undergoing elective thoracic surgery and the predictors of moderate-to-severe pain-secondary data analysis. J Clin Med 2024 Feb 1;13(3):844. [doi: 10.3390/jcm13030844] [Medline: 38337538]
6. Sniecinski RM, Bennett-Guerrero E, Shore-Lesserson L. Anticoagulation management and heparin resistance during cardiopulmonary bypass: a survey of Society of Cardiovascular Anesthesiologists members. Anesth Analg 2019 Aug;129(2):e41-e44. [doi: 10.1213/ANE.0000000000003981] [Medline: 30540616]
7. Miyazaki T, Doi R, Matsumoto K. Post-thoracotomy pain syndrome in the era of minimally invasive thoracic surgery. J Thorac Dis 2024 May 31;16(5):3422-3430. [doi: 10.21037/jtd-24-158] [Medline: 38883660]
8. Rabbitts JA, Fisher E, Rosenbloom BN, Palermo TM. Prevalence and predictors of chronic postsurgical pain in children: a systematic review and meta-analysis. J Pain 2017 Jun;18(6):605-614. [doi: 10.1016/j.jpain.2017.03.007] [Medline: 28363861]
9. Palomaa AK, Hakala M, Pölkki T. Parents' perceptions of their child's pain assessment in hospital care: a cross-sectional study. J Pediatr Nurs 2023;71:79-87. [doi: 10.1016/j.pedn.2023.03.012] [Medline: 37030017]
10. Escobar-Castellanos M, Míguez-Navarro MC, García-Mancebo J, et al. How much do parents know about pain in their children? Pediatr Emerg Care 2023 Jan 1;39(1):40-44. [doi: 10.1097/PEC.0000000000002585] [Medline: 36580891]
11. Yan C, Cheung RSY, Wong CL, et al. Stress and perception of procedural pain management in Chinese parents of children with cancer. J Pain Symptom Manage 2021 Jan;61(1):90-102. [doi: 10.1016/j.jpainsymman.2020.06.028] [Medline: 32640278]
12. Hu Y, Shi Y. Differences between parents and nurses in perceived educational support for pediatric postoperative pain management. J Eval Clin Pract 2025 Dec;31(8):e70338. [doi: 10.1111/jep.70338] [Medline: 41308016]
13. Bentsen L, Hanghøj S, Hjerming M, et al. Development of quality of life in adolescents and young adults with cancer using a patient support smartphone app: prepost interventional study. JMIR Cancer 2023 Dec 4;9:e49735. [doi: 10.2196/49735] [Medline: 38048144]
14. Versteegh LA, Chang AB, Chirgwin S, Tenorio FP, Wilson CA, McCallum GB. Multi-lingual "asthma app" improves health knowledge of asthma among Australian First Nations carers of children with asthma. Front Pediatr 2022;10:925189. [doi: 10.3389/fped.2022.925189] [Medline: 36110111]
15. Chen Y, Chin M, Greenberg S, Johnstone C, McGuinness J. Post-tonsillectomy pain in 24 children - utilising short message service (SMS) to assess postoperative outcomes. Clin Otolaryngol 2012 Oct;37(5):412-414. [doi: 10.1111/j.1749-4486.2012.02521.x] [Medline: 23164270]

XSL•FO
RenderX

16. Nelson SE, Adams AJ, Buczek MJ, Anthony CA, Shah AS. Postoperative pain and opioid use in children with supracondylar humeral fractures: balancing analgesia and opioid stewardship. J Bone Joint Surg Am 2019 Jan 16;101(2):119-126. [doi: 10.2106/JBJS.18.00657] [Medline: 30653041]
17. Drake-Brockman TFE, Smallbone HE, Sommerfield D, von Ungern-Sternberg BS. Remote after-care using smartphones: a feasibility study of monitoring children's pain with automated SMS messaging. Paediatr Anaesth 2022 Aug;32(8):954-960. [doi: 10.1111/pan.14481] [Medline: 35569012]
18. Schroeder K, Maiarana J, Gisiri M, et al. Caregiver acceptability of mobile phone use for pediatric cancer care in Tanzania: cross-sectional questionnaire study. JMIR Pediatr Parent 2021 Dec 8;4(4):e27988. [doi: 10.2196/27988] [Medline: 34889763]
19. Newton L, Sulman C. Use of text messaging to improve patient experience and communication with pediatric tonsillectomy patients. Int J Pediatr Otorhinolaryngol 2018 Oct;113:213-217. [doi: 10.1016/j.ijporl.2018.07.048]
20. Grasaas E, Helseth S, Fegran L, et al. App-based intervention among adolescents with persistent pain: a pilot feasibility randomized controlled trial. Pilot Feasibility Stud 2022 Jul 27;8(1):158. [doi: 10.1186/s40814-022-01113-0] [Medline: 35897086]
21. Palermo TM, de la Vega R, Murray C, Law E, Zhou C. A digital health psychological intervention (WebMAP Mobile) for children and adolescents with chronic pain: results of a hybrid effectiveness-implementation stepped-wedge cluster randomized trial. Pain 2020 Dec;161(12):2763-2774. [doi: 10.1097/j.pain.0000000000001994] [Medline: 32658147]
22. Simon JDHP, Schepers SA, van Gorp M, et al. Pain monitoring app leads to less pain in children with cancer at home: results of a randomized controlled trial. Cancer 2024 Jul 1;130(13):2339-2350. [doi: 10.1002/cncr.35100] [Medline: 37947136]
23. Shaygan M, Jaberi A. The effect of a smartphone-based pain management application on pain intensity and quality of life in adolescents with chronic pain. Sci Rep 2021 Mar 23;11(1):6588. [doi: 10.1038/s41598-021-86156-8] [Medline: 33758322]
24. Lalloo C, Harris LR, Hundert AS, et al. The iCanCope pain self-management application for adolescents with juvenile idiopathic arthritis: a pilot randomized controlled trial. Rheumatology (Oxford) 2021 Jan 5;60(1):196-206. [doi: 10.1093/rheumatology/keaa178] [Medline: 32613229]
25. Tiozzo E, Biagioli V, Brancaccio M, et al. Using an app to monitor postoperative pain at home in pediatric patients. J Child Health Care 2021 Jun;25(2):212-224. [doi: 10.1177/1367493520919313] [Medline: 33571013]
26. Ohashi K, Suzuki H, Sata Y, et al. Postoperative pain and quality of life after lung cancer surgery: a prospective observational study. Ann Palliat Med 2023 Mar;12(2):346-355. [doi: 10.21037/apm-22-207] [Medline: 36627847]
27. Lu JJ, Zhang FY, Feng Y, Yang YF, Guo Y, Feng YZ. The Chinese version of the Brief Pain Inventory-Facial (BPI-F) among different populations: factor structure and measurement invariance. J Dent 2023 Dec;139:104771. [doi: 10.1016/j.jdent.2023.104771] [Medline: 37925049]
28. Wang XS, Mendoza TR, Gao SZ, Cleeland CS. The Chinese version of the Brief Pain Inventory (BPI-C): its development and use in a study of cancer pain. Pain 1996 Oct;67(2-3):407-416. [doi: 10.1016/0304-3959(96)03147-8] [Medline: 8951936]
29. Tan G, Jensen MP, Thornby JI, Shanti BF. Validation of the Brief Pain Inventory for chronic nonmalignant pain. J Pain 2004 Mar;5(2):133-137. [doi: 10.1016/j.jpain.2003.12.005] [Medline: 15042521]
30. Stahlschmidt L, Hübner-Möhler B, Dogan M, Wager J. Pain self-efficacy measures for children and adolescents: a systematic review. J Pediatr Psychol 2019 Jun 1;44(5):530-541. [doi: 10.1093/jpepsy/jsz002] [Medline: 30802913]
31. Stahlschmidt L, Dogan M, Hübner-Möhler B, et al. Development and validation of the Scale for Pain Self-Efficacy (SPaSE) in German and English languages for children and adolescents. J Pain 2023 Jun;24(6):1069-1079. [doi: 10.1016/j.jpain.2023.01.007] [Medline: 36646401]
32. Valizadeh F, Ahmadi F, Zarea K. Neglect of postoperative pain management in children: a qualitative study based on the experiences of parents. J Pediatr Nurs 2016;31(4):439-448. [doi: 10.1016/j.pedn.2016.02.012] [Medline: 27033024]
33. Rony RYZ, Fortier MA, Chorney JM, Perret D, Kain ZN. Parental postoperative pain management: attitudes, assessment, and management. Pediatrics 2010 Jun;125(6):e1372-e1378. [doi: 10.1542/peds.2009-2632] [Medline: 20498177]
34. Zhang SM, Chan YY, Lai LL, Liang PR, Lin M. Parental postoperative pain management perceptions, attitudes, and practices in pediatric limb fractures. Pain Manag Nurs 2023 Oct;24(5):506-512. [doi: 10.1016/j.pmn.2023.07.006] [Medline: 37574333]
35. Gan TJ. Poorly controlled postoperative pain: prevalence, consequences, and prevention. J Pain Res 2017;10:2287-2298. [doi: 10.2147/JPR.S144066] [Medline: 29026331]
36. Wang L, Qin F, Liu H, Lu XH, Zhen L, Li GX. Pain sensitivity and acute postoperative pain in patients undergoing abdominal surgery: the mediating roles of pain self-efficacy and pain catastrophizing. Pain Manag Nurs 2024 Apr;25(2):e108-e114. [doi: 10.1016/j.pmn.2023.12.001] [Medline: 38184383]
37. Lee YJ, Bettick D, Rosenberg C. Improving pain self-efficacy in orthopedic surgery patients through video-based education: a quality improvement project. Pain Manag Nurs 2024 Oct;25(5):451-458. [doi: 10.1016/j.pmn.2024.04.008] [Medline: 38719657]
38. Tomlinson RM, Bax KC, Ashok D, McMurtry CM. Health-related quality of life in youth with abdominal pain: an examination of optimism and pain self-efficacy. J Psychosom Res 2021 Aug;147:110531. [doi: 10.1016/j.jpsychores.2021.110531] [Medline: 34082155]

XSL•FO
RenderX

39. Nabbijohn AN, Newby-Clark IR, Mack D, Stintzi A, McMurtry CM. Pain trajectories in pediatric inflammatory bowel disease: disease severity, optimism, and pain self-efficacy. Clin J Pain 2025 May 1;41(5):e1279. [doi: 10.1097/AJP.0000000000001279] [Medline: 40052200]
40. Haraldstad K, Stea TH. Associations between pain, self-efficacy, sleep duration, and symptoms of depression in adolescents: a cross-sectional survey. BMC Public Health 2021 Sep 5;21(1):1617. [doi: 10.1186/s12889-021-11680-1] [Medline: 34482826]
41. Colloca L, Barsky AJ. Placebo and nocebo effects. N Engl J Med 2020 Feb 6;382(6):554-561. [doi: 10.1056/NEJMra1907805] [Medline: 32023375]
42. Okusogu C, Colloca L. Placebo hypoalgesia: above and beyond expectancy and conditioning. Curr Opin Behav Sci 2019 Apr;26:75-81. [doi: 10.1016/j.cobeha.2018.10.008] [Medline: 31538101]
43. Abourezk AR, Abdelbaser I, Badran A, Abdelfattah M. Ultrasound-guided mid-point transverse process to pleura block versus thoracic paravertebral block in pediatric open-heart surgery: a randomized controlled non-inferiority study. J Clin Anesth 2024 Oct;97:111507. [doi: 10.1016/j.jclinane.2024.111507] [Medline: 38852396]
44. Qi J, Du B, Gurnaney H, Lu P, Zuo Y. A prospective randomized observer-blinded study to assess postoperative analgesia provided by an ultrasound-guided bilateral thoracic paravertebral block for children undergoing the Nuss procedure. Reg Anesth Pain Med 2014;39(3):208-213. [doi: 10.1097/AAP.0000000000000071] [Medline: 24682079]
45. Mundal I, Schei J, Lydersen S, Thomsen PH, Nøvik TS, Kvitland LR. Prevalence of chronic and multisite pain in adolescents and young adults with ADHD: a comparative study between clinical and general population samples (the HUNT study). Eur Child Adolesc Psychiatry 2024 May;33(5):1433-1442. [doi: 10.1007/s00787-023-02249-x] [Medline: 37386203]
46. Bouchatta O, Aby F, Sifeddine W, et al. Pain hypersensitivity in a pharmacological mouse model of attention-deficit/hyperactivity disorder. Proc Natl Acad Sci U S A 2022 Jul 26;119(30):e2114094119. [doi: 10.1073/pnas.2114094119] [Medline: 35858441]
47. Dudeney J, Aaron RV, Hathway T, et al. Anxiety and depression in youth with chronic pain: a systematic review and meta-analysis. JAMA Pediatr 2024 Nov 1;178(11):1114-1123. [doi: 10.1001/jamapediatrics.2024.3039] [Medline: 39250143]
48. Balasco L, Provenzano G, Bozzi Y. Sensory abnormalities in autism spectrum disorders: a focus on the tactile domain, from genetic mouse models to the clinic. Front Psychiatry 2019;10:1016. [doi: 10.3389/fpsyt.2019.01016] [Medline: 32047448]

Abbreviations

ASA: American Society of Anesthesiologists
BPI: Brief Pain Inventory
NRS: numeric rating scale
PCAP: patient-controlled analgesia pump
POD: postoperative day
PSES: Pain Self-Efficacy Scale

Edited by A Schwartz, M Balcarras; submitted 04.Aug.2025; peer-reviewed by C Brimeyer, D Poddighe; revised version received 08.Feb.2026; accepted 09.Feb.2026; published 24.Mar.2026.

Please cite as:
Shi Y, Wang Q, Liu A, Jiang L
The Effect of Text Messaging on the Postoperative Pain Experience in Pediatric Patients Undergoing Thoracic Surgery: Randomized Controlled Trial
J Med Internet Res 2026;28:e81806
URL: https://www.jmir.org/2026/1/e81806
doi:10.2196/81806

© Yun Shi, Qianqiu Wang, Aihua Liu, Li Jiang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Designing Electronic Problem-Solving Training for Individuals With Traumatic Brain Injury: Mixed Methods, Community-Based, Participatory Research Case Study

Matthew Schmidt[1,2], PhD; Yueqi Weng[2], MEd; Shannon Juengst[3], PhD, CRC; Alexandra Holland[3], BS

[1]College of Pharmacy, Department of Clinical and Administrative Pharmacy, University of Georgia, River's Crossing, 215, Athens, GA, United States

[2]Department of Workforce Education and Instructional Technology, Mary Francis Early College of Education, University of Georgia, Athens, GA, United States

[3]Brain Injury Research Center, TIRR Memorial Hermann, Houston, TX, United States

Corresponding Author:
Matthew Schmidt, PhD
College of Pharmacy, Department of Clinical and Administrative Pharmacy, University of Georgia, River's Crossing, 215, Athens, GA, United States

Abstract

Background: Traditional rehabilitation research often excludes the voices of individuals with lived experience of traumatic brain injury (TBI), resulting in interventions that lack relevance, accessibility, and effectiveness. Community-based participatory research (CBPR) offers an alternative framework that emphasizes collaboration, power sharing, and sustained engagement with patients, caregivers, and clinicians.

Objective: This study aimed to apply CBPR to guide front-end design (empathy interviews, empathy mapping, personas) and to evaluate the sociotechnical-pedagogical usability of the Electronic Problem-Solving Training (ePST) mobile health (mHealth) intervention with TBI partners.

Methods: A multistep, mixed methods design case methodology was adopted, guided by CBPR principles and learning experience design. Participatory mechanisms included a 33-member Community Advisory Board and 10 Community Engagement Studios that engaged TBI survivors, caregivers, clinicians, and researchers throughout the Discover, Define, Develop, and Deliver phases of the Double Diamond model. Iterative activities included empathy interviews (n=14), persona development (n=10), rapid prototyping, and usability testing with 5 participants with TBI using think-aloud protocols and the Comprehensive Assessment of Usability for Learning Technologies instrument.

Results: The co-design process successfully translated community feedback into an empathy-informed, user-centered prototype and systematically identified design considerations that single-partner approaches overlook. TBI-specific design requirements emerged, including the need for linear content progression over branching navigation, higher technical performance standards, and explicit content signaling with clarity prioritized over novel interface design. Think-aloud protocols revealed that participants struggled with mobile navigation and branching structures but excelled with sequential content progression. In addition, the input from individuals with TBI, caregivers, clinicians, and researchers led to practical refinements such as shorter microlearning lessons (5 - 12 min), clearer voiceover tone, and simplified navigation, directly addressing the study's objective of improving accessibility and emotional resonance. Overall usability was high, measured using the Comprehensive Assessment of Usability for Learning Technologies (CAUSLT), with an average score of 4.25 out of 5 (SD 0.72; 95% CI 3.36 - 5.15; n=5). Knowledge accuracy was 80% (8/10 items; 95% CI 49% - 94%; n=5 participants; 2 items each), indicating that the system effectively supported learning and comprehension. Module completion was 100% (5/5; 95% CI 56.6% - 100%). Average time-on-task for 10 lesson completions was 11.47 (SD 5.28; range 4.6 - 21.42) minutes per lesson, demonstrating strong task efficiency and engagement. Highest ratings were observed in the pedagogical usability domain, reflecting that the interface was clear, intuitive, and conducive to learning. Collectively, these findings suggest that applying CBPR across all design stages produced a technically sound, easy-to-use, and pedagogically meaningful mHealth tool specifically tailored for individuals with TBI.

Conclusions: Sustained CBPR across full design and development cycles resulted in high usability for ePST for individuals with TBI. Ultimately, this study operationalized a full-cycle pipeline that links sustained community partnership to measured usability outcomes, producing community-informed design principles and a reproducible mixed methods approach for formative mHealth development for TBI.

(J Med Internet Res 2026;28:e83995) doi:10.2196/83995

XSL•FO
RenderX

KEYWORDS

traumatic brain injury; community-based participatory research; user-centered design; usability; mHealth; rehabilitation; problem-solving training

Introduction

Traumatic Brain Injury Rehabilitation

Traditional rehabilitation research underrepresents people with lived experience of disability, including traumatic brain injury (TBI), yielding interventions misaligned with patient contexts [1-4]. TBI has acute and chronic sequelae [5,6] affecting cognition, emotion, and social functioning [5-10] that adversely affect learning and care access [11,12]. When design ignores these constraints, relevance and engagement drop [13]. Although recent studies demonstrated the feasibility of participatory adaptations in TBI rehabilitation [14,15], such approaches remain rare [16]. Emerging protocols increasingly incorporate caregiver and community voices through community-based participatory research (CBPR) frameworks [17], yet broader adoption remains limited [18]. Indeed, although chronic challenges faced by individuals with TBI are increasingly recognized, rehabilitation research rarely translates this awareness into meaningful community engagement or integration of practitioner perspectives [19,20]. Provider- and institution-centered models continue to dominate, reinforcing inequities [21] and limiting collaboration between researchers, clinicians, and patients [22,23]. This results in less responsive interventions, lower user satisfaction, and reduced effectiveness [5,24-26]. Despite growing support for participatory approaches, provider-centric norms persist [27]. This study responds to those gaps by modeling a collaborative, community-informed design process [18,28].

Power sharing and collaborative decision-making are critical to designing effective, context-responsive interventions [28,29]. Rehabilitation requires real-world interaction, collaboration, and adaptability to individual needs [30]. Participatory approaches shift decision-making toward community members [31], having produced measurable improvements in health outcomes and patient-reported measures [32]. For example, ethnographic work by Manhas and colleagues [33] showed shared decision-making in rehabilitation enhances patient satisfaction, understanding, goal attainment, and self-reported outcomes. This contrasts with provider-driven models that limit patient involvement and flexibility. Indeed, challenges such as limited community engagement, asymmetrical decision-making, and provider-centered research can undermine the relevance and impact of TBI rehabilitation efforts [19,21-23]. These issues call for more inclusive approaches that prioritize symmetrical decision-making and meaningful collaboration with the TBI community [33].

This paper presents a case example of the formative design and evaluation of Electronic Problem-Solving Training (ePST), a metacognitive, evidence-based mobile health (mHealth) problem-solving intervention. ePST is based on PST, a cognitive-behavioral approach with proven efficacy for neurodevelopmental and psychological conditions that is grounded in some of the strongest evidence in cognitive rehabilitation [34,35]. PST and comparable approaches are widely used in psychology to improve problem-solving skills and mindset [36,37] and have shown promise for preventing and treating cognitive deficits [38] through numerous clinical trials [39]. Research suggests PST can be especially beneficial for long-term or multifaceted health issues, such as TBI [40,41]. A robust body of evidence shows that such problem-solving approaches lead to meaningful reductions in symptoms, strengthen individuals' confidence in managing their health, and enhance adherence to prescribed regimens [42-44]. ePST was developed using learning experience design and a CBPR framework to ensure accessibility, community-driven decision-making, and iterative co-design [18,45,46]. Learning experience design and CBPR guided front-end activities and the sociotechnical-pedagogical usability evaluation reported here.

Background and Rationale

Rehabilitation research often centers around clinician and designer perspectives over patient input, reducing relevance, effectiveness, and adaptability for individuals with TBI [47,48]. Correa and colleagues [49] showed that interventions lacking patient involvement can be misaligned with how patients perceive risks, benefits, and treatment goals, undermining recruitment and randomization. Such problems suggest a need for adaptive, patient-informed approaches, which CBPR can provide in a context-sensitive and ethical manner [46]. CBPR helps researchers understand lived experience and co-create interventions that are more relevant, acceptable, and effective. For example, Quilico and colleagues [28] partnered with people with TBI and caregivers to adapt a physical activity program, producing changes that improved relevance, outcomes, and engagement. Groussard and colleagues [50] involved users with lived TBI experience and caregivers in developing and evaluating a cognitive support system, yielding improved user satisfaction and greater autonomy. However, participation alone is insufficient. CBPR requires reciprocal relationships among community members, academics, and practice partners to draw on diverse strengths [51]. As a case-in-point, Springer and Skolarus [52] specifically distinguished between the "community-based" and "participatory" components of CBPR to clarify how all components of this approach are needed to promote sustained, power-sharing partnerships.

As CBPR is applied increasingly to digital health interventions like ePST, new design and evaluation demands emerge. For example, reporting in CBPR remains inconsistent, and implementation is uneven, especially in rehabilitation contexts [16]. Usability and contextual fit present persistent barriers to adoption in eHealth and mHealth, reinforcing the need for community-informed design and iterative testing cycles [53]. In addition, promoting sustained engagement remains a challenge [18,54], which supports the use of innovative pedagogical strategies such as microlearning, a design approach shown to improve engagement and learning outcomes in health applications when lessons are limited to a length of 5 minutes

XSL•FO
RenderX

to 12 minutes [55]. In parallel, sociotechnical frameworks have been recommended for evaluating patient-facing tools, supporting our integration of CBPR, learning experience design, and the sociotechnical-pedagogical framework (discussed in the next section) [56]. Collectively, these gaps suggest a need to balance TBI rehabilitation complexity with the provision of usable, accessible, and engaging interventions. We illustrate our approach to achieving this balance through the conceptual framework we present in the following section.

Conceptual Framework

We developed a conceptual model that places CBPR at the methodological core, pairs participatory mechanisms (Community Advisory Board or Community Engagement Studios) with learning experience design to convert community partner input into design principles, and maps these strands onto the Double Diamond (Figure 1) for iterative development and sociotechnical-pedagogical evaluation [57-61].

Figure 1. Conceptual framework integrating community-based participatory research, the Community Advisory Board (CAB), Community Engagement Studios (CES), and learning experience design (LXD) mapped onto the Double Diamond design framework for individuals with traumatic brain injury (TBI).

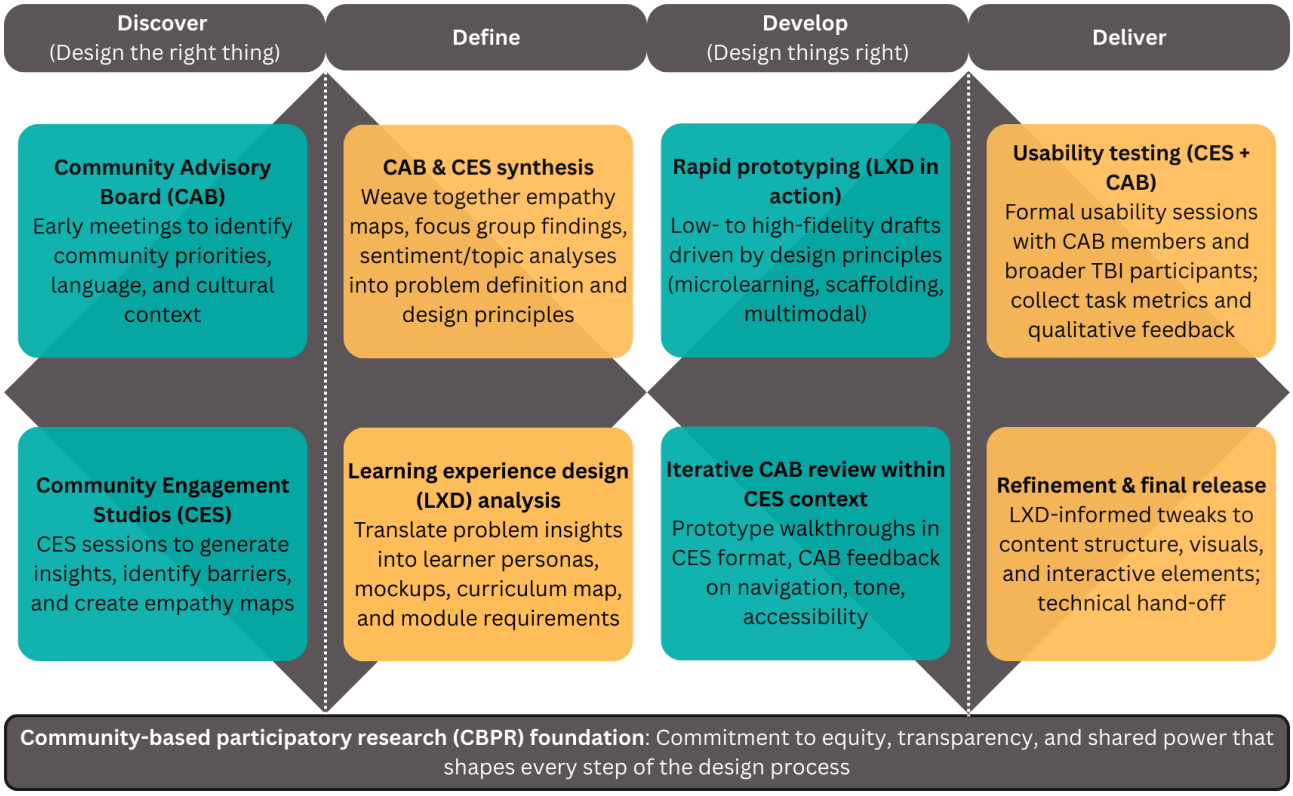

CBPR as the Foundational Ethos

We adopted CBPR as a foundational ethos to foster inclusive, patient-centered rehabilitation design and to translate community priorities into practice [62-65]. Central to CBPR are collaboration and balanced partnerships that share decision-making responsibility [66]. Unlike short-term, investigator-led studies, CBPR emphasizes long-term reciprocal relationships that promote ethical research practices and improved outcomes [67]. This is important because top-down, limited-duration studies can erode trust and exclude local needs, with standardized practices that do not accommodate community input tending to perpetuate these problems [68,69]. CBPR's emphasis on shared decision-making across all phases of the research process provides one avenue to address these problems [70]. Collaboration through structured partnerships allows community members to inform priorities, participate in knowledge creation, and strengthen the real-world applicability of interventions [31]. These approaches move research beyond expert-driven agendas by integrating the lived experiences, priorities, and contextual knowledge of community members into the design and implementation process [71].

Participatory Mechanisms

Community Advisory Board

Community Advisory Boards are structured, ongoing partnerships that integrate people with lived experience into research, providing authentic representation and culturally grounded input across the project lifecycle [72-75]. Unlike short-term focus groups, Community Advisory Boards meet regularly to co-develop research strategy, advise on ethics and context, and guide intervention refinement, fostering shared leadership, trust, and power sharing [76-79].

Community Engagement Studios

Community Engagement Studios are structured, facilitated consultations in which researchers obtain targeted feedback from panels of community experts, caregivers, and clinicians. Unlike advisory boards or focus groups, Community Engagement Studios use focused, iterative sessions to promote dialogue, reciprocal learning, and sustained community involvement [59,80]. Originating with the Meharry-Vanderbilt Community-Engaged Research Core [58], Community Engagement Studios were developed to overcome participation

XSL•FO
RenderX

barriers in clinical and rehabilitation research, including mistrust from historical unethical practices and social inequities [81]. By positioning community members as consultants and experts rather than passive subjects, Community Engagement Studios help identify barriers, adapt interventions to community needs, and build trust with underrepresented groups [82,83]. Community Engagement Studios can enhance cultural adaptation [84], increase minority participation [82,83], and reduce power imbalances between researchers and community members [85,86].

Operationalization via Learning Experience Design

Learning experience design is a learner-centered, theoretically grounded framework that integrates instructional design, cognitive science, user experience, and participatory approaches [87,88]. Learning experience design emphasizes designing engaging and inclusive learning environments that respond to learners' real-world needs and experience [89,90]. Learning experience design focuses on the cognitive, emotional, and perceptual influences of learner interactions with content, tools, and people across the learning process [91-93]. Learning experience design guided ePST's Double Diamond workflow. In *Discover*, empathy interviews identified core needs and constraints; in *Define*, those insights plus Community Advisory Board input shaped personas, module structure, and mock content [94,95]. To address TBI-specific cognitive limits (eg, memory, fatigue), the *Develop* phase adopted microlearning (ie, short, digestible lessons lasting 5 minutes to 12 minutes) intended to lower cognitive load, promote encoding, and support retention [96-99]. *Deliver* used iterative usability testing to validate designs and drive refinements. Multimodal strategies (text, visuals, voiceover, interactivity) and gamification (badges, progress indicators, interactive tasks) supported diverse preferences and motivation [100,101].

Sociotechnical-Pedagogical Framework

The sociotechnical-pedagogical framework conceptualizes learner experience as the alignment of 3 interdependent domains: technological, pedagogical, and sociocultural [102,103]. The technological domain covers reliability, accessibility, device compatibility, navigability, and error tolerance; the pedagogical domain covers alignment of objectives, materials, activities, and assessment, plus clarity, scaffolding, cognitive load management, and feedback quality; and the sociocultural domain addresses presence, identity, communication, cultural responsiveness, and scenario authenticity. The sociotechnical-pedagogical framework serves as both a design and evaluation lens, operationalized via dimension-specific heuristics validated against course evaluations that identified 195 distinct problems consolidated into nonoverlapping heuristics spanning the 3 domains [102]. This approach is critical in neurorehabilitation because traditional usability frameworks often miss interactions among cognitive, social, and technical factors [88]. For people with TBI, technological design must go beyond basic accessibility to reduce cognitive load (simplified interfaces, memory supports, fatigue accommodations) and ensure assistive-technology compatibility. Pedagogical design should address executive function limits via clear structure, predictable flows, compensatory strategies, repetition, and metacognitive scaffolds to support transfer. Sociocultural design must attend to stigma, identity shifts after injury, peer and family involvement, and social-context fit. The sociotechnical-pedagogical framework reveals problems that purely technical reviews miss.

Intervention Description

ePST is a cross-platform, community-informed mHealth intervention tailored to the cognitive and emotional needs of adults with TBI. Built on microlearning, it delivers short (5 - 12 minutes), chunked lessons with built-in progress tracking to reduce cognitive load. Engagement features include motivational messaging derived from empathy interviews; a virtual coach ("Ruth"); personalized learning pathways; embedded reminders; and gamified elements (badges, certificates) to support memory, reinforce learning, and sustain motivation (Figure 2). ePST is grounded in problem-solving training and operationalizes the 6-step ABCDEF mnemonic (Figure 3): A, assess the problem; B, brainstorm solutions; C, consider and choose; D, develop and do; E, evaluate; and F, flex. ePST translates these steps into scaffolded modules that teach structured decision-making and problem-solving strategies tailored to adults with TBI. A description of the ePST learning modules is provided in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 2. Representative screenshots from the Electronic Problem-Solving Training (ePST) prototype and final user interface, captured during usability testing with adults with traumatic brain injury: (1) progress tracker, (2) virtual coach "Ruth" interface, and (3) reminder/notification panel.

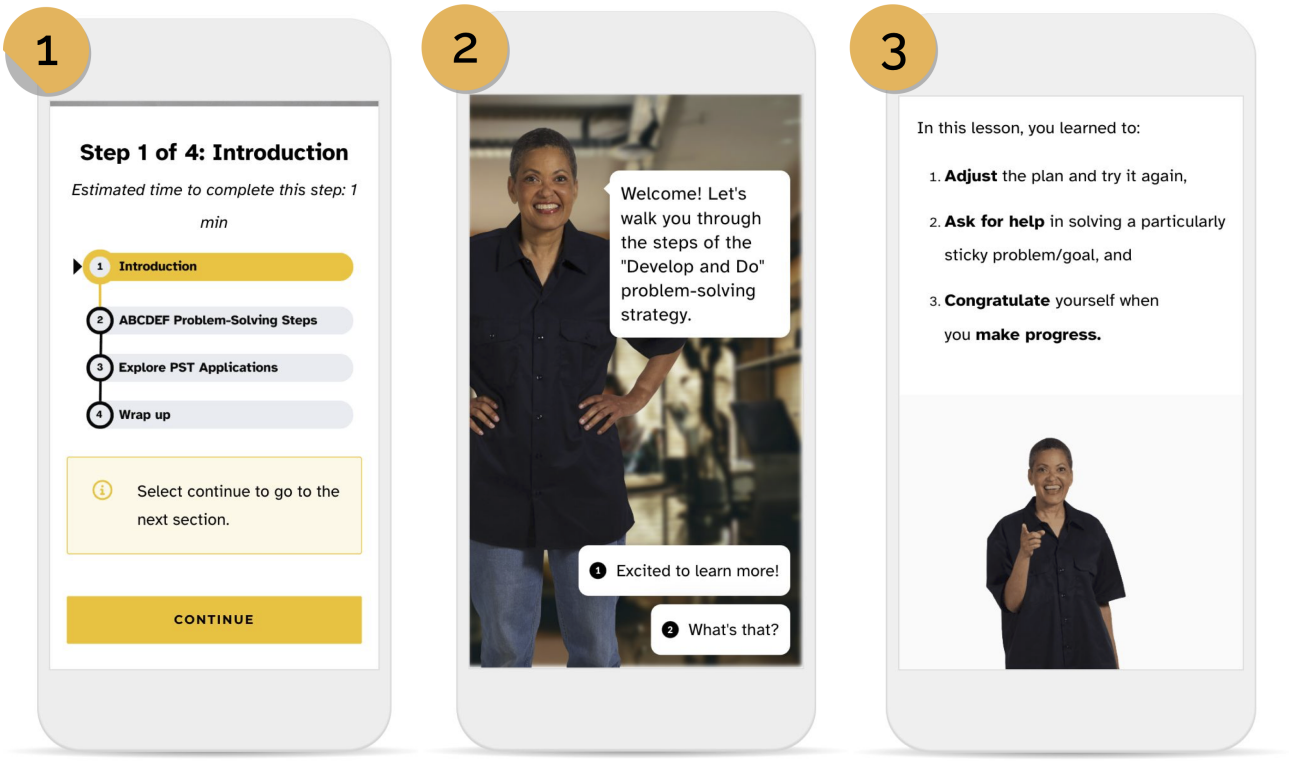

Figure 3. Problem-solving training strategy diagram showing the ABCDEF 6-step metacognitive process implemented in Electronic Problem-Solving Training (ePST): assess, brainstorm, consider and choose, develop and do, evaluate, and flex.

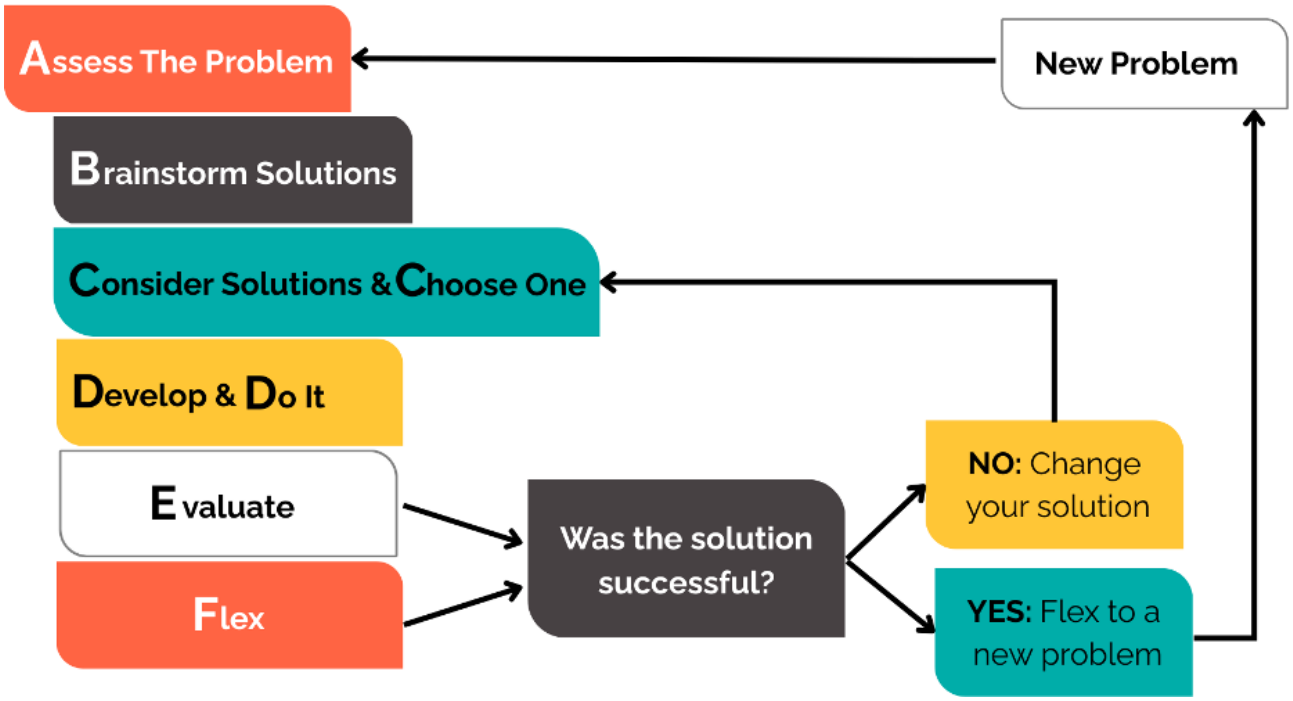

Purpose and Research Questions

The purpose of this iterative, multimethod formative design and evaluation case study was to use a CBPR approach to guide front-end design activities (ie, empathy interviews, empathy mapping, and persona development) and to evaluate the sociotechnical-pedagogical usability of the ePST intervention with TBI community members (ie, Community Advisory Board members, families, providers, and individuals with lived TBI experience) at a large public university and a large medical center in the southern United States. The questions that guided this research included: research question (RQ) 1: What themes related to learning needs, barriers, and preferences emerge from front-end design activities (empathy interviews, empathy mapping, and persona development) with TBI community members? RQ 2: How did individuals with TBI perceive the

XSL•FO
RenderX

sociocultural, technological, and pedagogical usability aspects of their experience with ePST during testing? RQ 3: How were identified sociotechnical-pedagogical usability issues addressed through design refinements?

Methods

Double Diamond Approach

This multimethod formative design and evaluation study followed the Double Diamond approach (Discover, Define, Develop, Deliver) and ran from February 2024 through July 2024. In *Discover*, we established the Community Advisory Board, drafted initial design principles, and conducted Community Engagement Studio empathy interviews with people with TBI. In *Define*, we developed learner personas, produced a curriculum map, and iteratively refined priorities via Community Engagement Studios and Community Advisory Board reviews. In *Develop*, we translated principles into low- to high-fidelity prototypes and internal subject matter expert review. In *Deliver*, we conducted iterative usability testing with people with lived TBI experience and implemented refinements after each round. We reported patient and public involvement using the GRIPP2 Short Form (GRIPP2-SF) [104]. A 1-page mapping table linking GRIPP2-SF items to manuscript locations is provided in Multimedia Appendix 1.

Participants

Community Advisory Board Participants

The Community Advisory Board (n=33) was purposively assembled to include people with lived TBI experience, caregivers, clinicians, researchers, industry representatives, advocates, and members of minoritized groups. Members were identified via professional networks, partner clinics, and community organizations and invited by email. Selection criteria were TBI or digital health expertise, lived experience, demographic diversity, and advocacy and service representation. Community Advisory Board members received US $25 per meeting. Community Advisory Board composition is provided in Multimedia Appendix 1.

Empathy Interview Participants

Empathy interview participants (n=14) were recruited via clinician referral and community outreach at a large tertiary rehabilitation center in the southern United States in February 2024 and March 2024. Inclusion criteria were age ≥18 years, proficiency in English, and either (1) documented TBI confirmed by clinician referral or review of medical records when available or (2) self-reported TBI with screening confirmation of capacity to participate. Exclusion criteria were severe communication impairments or acute medical instability that precluded informed consent or participation. Caregivers and providers were eligible if they provided regular care or clinical services to adults with TBI.

Usability Testing Participants

Usability testing participants (n=5) were recruited purposively from the same clinical and community sources in July 2024 to capture variation in technology experience and time since injury. Inclusion criteria included age ≥18 years, English fluency, history of TBI (clinician referral or medical record when available), ability to use a smartphone or computer without assistance, and capacity to provide informed consent and follow study tasks. Exclusion criteria included acute medical or psychiatric instability and severe receptive or expressive communication impairments that prevented participation. Five participants is standard for early-stage formative usability tests, with 80% of usability problems identified via small samples [105,106]. This low number limits statistical generalizability but is conventional in heuristic-based usability work intended for problem identification [107,108]. This approach has substantial precedent in digital health formative studies that use small, purposive usability samples to drive iterative refinements [109-111]. To increase rigor and reduce bias from the small sample, we purposively sampled, triangulated findings, and applied an iterative refine-and-retest logic.

Ethical Considerations

The study protocol was approved by the Human Research Protection Program at the University of Georgia (IRB #00009943) on June 17, 2024, and was deemed exempt. Written informed consent was obtained electronically via Qualtrics. Participants reviewed the full consent document, typed their full name and the date to indicate agreement, and submitted the consent form. The consent form covered study purpose and procedures, audio and video recording, foreseeable risks and benefits, right to withdraw, compensation, and data handling. All study data were de-identified and stored on encrypted University of Georgia servers with access limited to authorized study personnel. The master linking list and raw recordings will be destroyed at study end; de-identified data may be used for future research but will not be deposited in a public repository. Participants received US $25 per activity, and Community Advisory Board members were paid US $25 per meeting, with payments issued via Clincard after each session. Payments were institutional review board–approved and described in the consent forms. No identifying information was included in this paper or multimedia appendices.

Procedures

Given the iterative nature of the Double Diamond approach, analysis was multimodal and occurred across all phases of design, with analysis falling into 3 broad categories: (1) qualitative, (2) quantitative, and (3) computational. No data were missing for any of the reported analyses.

Discover Phase Procedures

Establishment of the Community Advisory Board

Community Advisory Board members were recruited purposively [112] from the community, academia, industry, and medical-related institutes based on 4 criteria: (1) professional expertise in TBI rehabilitation, assistive technology, and or digital health; (2) lived experience with TBI; (3) demographic diversity across age, gender, socioeconomic status, and geography; and (4) community member representation including those with lived experience, clinicians, researchers, technology developers, and advocacy organizations. Quarterly Community Advisory Board meetings were held across Phases 1 - 3 (total n=16).

XSL•FO
RenderX

Establishment of Preliminary Design Principles

We drew on findings from a prior study, *Caregivers in Dementia PST and DSJ* (CaDeS), which tested coach-delivered PST [113]. Open-ended responses to overall intervention satisfaction were analyzed using machine learning techniques, including sentiment analysis and latent Dirichlet allocation, to generate an initial set of 7 design principles for ePST, which were reviewed and refined in a subsequent Community Engagement Studio session.

Empathy Interviews

Empathy interviews were conducted with 3 groups of participants. Groups 1 (n=6) and 2 (n=3) consisted of TBI survivors. Group 3 consisted of care partners and providers (n=5). Interviews were guided by the 4-phase empathy framework from Kouprie and Visser [114]. All interviews were approximately 75 minutes and conducted online using Zoom web conferencing software. Questions focused on (1) learning challenges, (2) effective therapies, (3) the impact of others' stories, (4) group-specific challenges, (5) building trust through shared expertise, and (6) motivational messages. Interviews were recorded and transcribed using Zoom.

Define Phase Procedures

Empathy Mapping

Empathy mapping guided learner analysis and informed the design of ePST [115]. Empathy mapping involved synthesizing participants' responses into 4 core domains ("Says," "Thinks," "Does," and "Feels") to foster understanding of their motivations, challenges, and learning preferences. This allowed capture of nuanced information about participants' cognitive, emotional, and behavioral experiences. A total of 9 empathy maps were created (see Multimedia Appendix 1). These were then used to generate learner personas and referenced to inform design.

Persona Development

Personas are fictional, data-informed archetypes that represent individuals within the target population [116]. Our personas provided summaries of representative descriptors based on information that was synthesized from empathy maps. Personas were presented to the Community Advisory Board, reviewed, and revised. Initial designs included TBI severity; however, this was removed at the recommendation of the Community Advisory Board, as severity was an inadequate method to represent nuanced TBI characteristics, especially chronically. The final set of personas (n=10) is provided in Multimedia Appendix 1.

Refinement of Design Principles

Design principles were refined based on a structured empathy interview with 5 caregivers and providers. Analysis comprised a discussion-based analytic process to identify key insights from the transcripts. The design principles were then reviewed in a Community Engagement Studio session with the Community Advisory Board, who provided feedback on clarity, relevance, and completeness. Analysis did not focus on achieving saturation but instead prioritized triangulation across data sources and methods for development of design principles.

Develop Phase Procedures

Community Engagement Studios

Structured Community Engagement Studio sessions were used to elicit structured feedback during Community Advisory Board meetings. Community Engagement Studio sessions (n=10) focused on usability challenges, content clarity, and delivery preferences. Community Engagement Studio sessions were between 60 minutes and 90 minutes, included 6 to 8 participants, and followed a structured protocol. A trained moderator guided discussion. Discussion foci varied depending on which design artifacts were being reviewed, Participants reflected on design artifacts' clarity, relevance, and usability. Sessions were conducted, audio recorded, and transcribed using Zoom. Transcripts and notes were then synthesized into actionable design recommendations.

Rapid Prototyping

Rapid prototyping is an iterative design approach that quickly develops and refines working models based on user feedback [117]. This approach was used to transform insights from the *Define* phase into working prototypes. Initial design concepts were explored through low-fidelity mockups then iteratively refined into medium- and high-fidelity prototypes, with emphasis on flexibility and responsiveness to user input [117]. Designs were regularly reviewed during Community Engagement Studios for issues such as navigation, language complexity, and content pacing.

Deliver Phase Procedures

Usability tests (n=5) were conducted by a trained graduate student and a university professor usability expert. Testing followed a semistructured, task-based research protocol. Sessions were between 60 minutes and 75 minutes and were conducted, recorded, and transcribed using Zoom. Participants completed 5 structured usability tasks per session while thinking aloud and sharing their screens. Tasks assessed both technological usability (eg, navigation, multimedia interaction) and pedagogical usability (eg, clarity of content, microlearning structure). Participants then completed the Comprehensive Assessment of Usability for Learning Technologies (CAUSLT) instrument [118]. Data were analyzed using an integrated approach that combined observational, survey, and efficiency metrics. Think-aloud transcripts and observer notes were reviewed and discussed by two team members to identify barriers. Responses to the CAUSLT instrument were summarized using descriptive statistics and disaggregated across the 3 instrument factors. Design flaws were prioritized using Nielsen's severity scale [119]. Efficiency data were extracted from session recordings. Findings were documented in a report that was reviewed with the Community Advisory Board, whose feedback guided refinements in areas such as voiceover quality, mobile navigation, and content clarity for cognitive accessibility.

XSL•FO
RenderX

Results

Discover Phase Results

Composition of Community Advisory Board

The composition of the Community Advisory Board is presented in Table 1. Some individuals were represented in more than one group, as some Community Advisory Board members self-identified with more than one category.

Table . Composition of the Community Advisory Board (n=33) including counts and role categories for members recruited purposively from clinical partners, community organizations, academic networks, industry, and advocacy groups.

Category	Composition	Total representatives, n
Academic researchers	PhD researchers (n=12), psychologists (n=8), educational technology experts (n=2), graduate students (n=2)	24
Industry professionals	Software developer (n=1), software designers (n=5)	6
Individuals with lived experience and advocates	Individuals with TBI[a] (n=6), TBI care partners (n=4), disability advocates (n=6)	16
Rehabilitation and clinical professionals	Occupational therapists (n=4), social workers (n=3), rehabilitation counselors (n=3)	10
Individuals with physical disabilities	Blind (n=1), deaf (n=1)	2
Individuals from minoritized groups	LGBTQAI+[b] (n=4), minoritized racial and ethnic groups (n=6)	10

[a]TBI: traumatic brain injury.

[b]LGBTQAI+: lesbian, gay, bisexual, transgender, queer (or questioning), asexual (or allied), intersex, plus

Preliminary Design Principles

Preliminary design principles were established based on results of prior research and Community Advisory Board input. Principles emphasized accessibility, emotional resonance, clarity of messaging, and personalization, serving as our foundation for early prototypes, visual design, and engagement strategies. These preliminary principles were later expanded and structured into a comprehensive hierarchy, reported in the Define Phase Results section.

Empathy Interview Participant Demographics

We recruited 14 participants (Table 2) for empathy interviews, including individuals with TBI (n=9) and caregivers and providers (n=5).

Table . Participant demographics for empathy interviews (n=14), including de-identified breakdown by participant group, race or ethnicity, age bands, and gender.

Characteristics	Individuals with TBI[a], n	Caregivers and providers, n
Race		
Hispanic	1	1
Caucasian or White	5	3
African American	2	0
Asian	0	1
Age (years)		
30-39	4	0
40-49	3	1
50-59	1	0
60-69	0	3
≥70	0	1
Gender		
Female	7	5
Male	2	0

[a]TBI: traumatic brain injury.

XSL•FO
RenderX

Define Phase Results

Empathy Maps

Empathy maps were created (n=9), with each map including brief descriptors in the categories “Says,” “Thinks,” “Feels,” and “Does.” Analysis revealed 4 key themes characterizing the post-TBI experience: Participants experienced (1) frustration and disorientation with everyday tasks, (2) loss of self-identity and nostalgia for pre-injury life, (3) physical exhaustion from therapy that decreased motivation, and (4) social isolation due to perceived lack of family understanding. In response, participants developed adaptive strategies including structured skill relearning through rehabilitation and memory aids such as sticky notes. The full set of empathy maps is provided in Multimedia Appendix 1.

Personas

A set of personas (n=10) was created to guide design. Personas highlighted varied life contexts, recovery journeys, and learning needs across individuals such as veterans, students, professionals, and retirees. Each reflected unique combinations of cognitive, emotional, and physical challenges, along with personal goals like regaining independence, improving memory, or reducing stigma. Common facilitators included family support, adaptive tools, storytelling, and professional guidance. Despite varied barriers ranging from aphasia to fatigue to discrimination, all personas demonstrated resilience and motivation to recover. The complete set of personas is provided in Multimedia Appendix 1.

Refined Design Principles

A refined set of design principles was created in the *Define* phase, incorporating the preliminary set created during the *Discover* phase. Using a framework proposed by Kali [120], the design team organized these insights into a 3-tiered hierarchy (specific, pragmatic, and metaprinciples). Pragmatic principles reflected actionable guidance relevant to the learning design. These pragmatic principles were grouped into 6 broader metaprinciples, such as accessibility, emotional support, motivation, personalization, cultural relevance, and evidence-based action. Where applicable, specific principles (eg, interface features, content structures) were also identified to illustrate how the pragmatic principles would translate into concrete design decisions (Table 3).

XSL•FO
RenderX

Table . Design principles for Electronic Problem-Solving Training (ePST) module development, including metaprinciples, pragmatic principles, and specific principles derived from empathy interviews, empathy maps, Community Advisory Board and Community Engagement Studio feedback, and persona development.

Metaprinciple	Pragmatic principle	Specific principles
1. Ensure accessibility and usability	Design for cognitive and physical inclusion	Use clear, concise, jargon-free language; include closed captioning; support mobile-first navigation; design intuitive interaction patterns
1. Ensure accessibility and usability	Support memory and comprehension	Reinforce key concepts with reminders and visual anchors; use chunked content and repeated exposure
1. Ensure accessibility and usability	Allow flexible engagement	Enable learners to proceed at their own pace; allow pausing and resuming lessons easily
2. Support emotional and behavioral needs	Encourage emotional regulation	Include calming activities (eg, music, mindfulness cues); normalize behavioral variability in content
2. Support emotional and behavioral needs	Empathize with behavioral and communication challenges	Acknowledge and adapt for speech and behavioral limitations; use neutral, nonjudgmental tone
2. Support emotional and behavioral needs	Promote self-awareness and acceptance	Include prompts or reflection activities to build insight into strengths and limitations
3. Foster motivation and engagement	Use positive reinforcement	Integrate badges, rewards, and affirming feedback
3. Foster motivation and engagement	Emphasize goal setting and achievement	Provide explicit opportunities to set and track goals
3. Foster motivation and engagement	Provide regular feedback	Visual progress indicators; summary pages at lesson or module completion
4. Enable personalized and multimodal learning	Use varied sensory inputs	Combine visuals, audio narration, and interactivity
4. Enable personalized and multimodal learning	Allow for autonomy and independence	Design lessons that can be completed without facilitator support; scaffold progressively to reduce reliance on help
4. Enable personalized and multimodal learning	Tailor content for diverse learners	Include customizable avatars or pathways; vary representation and examples by demographic relevance
5. Establish credibility and cultural relevance	Include lived experience	Use testimonials from TBI[a] survivors and care partners; embed quotes and real-world scenarios
5. Establish credibility and cultural relevance	Partner with trusted organizations	Reference TIRR[b], advocacy groups, and clinical partners in content
5. Establish credibility and cultural relevance	Practice inclusive and representative design	Include diverse racial, ethnic, and gender identities; adapt content for veterans and other priority subgroups
6. Ground content in evidence and action	Communicate evidence accessibly	Present supporting research in simplified language or visuals; avoid academic jargon
6. Ground content in evidence and action	Use motivating calls to action	End modules with clear next steps (eg, “Enroll,” “Learn more”); include clickable links or guided follow-ups

[a]TBI: traumatic brain injury.

[b]TIRR:The Institute for Rehabilitation and Research.

Develop Phase Results

During the *Develop* phase, design artifacts progressed from low-fidelity storyboards to high-fidelity interactive prototypes (Figure 4). Low-fidelity mockups were iteratively refined into functional prototypes via structured Community Advisory Board feedback focused on usability, content clarity, accessibility, and delivery preferences. Key outputs included finalized lesson content, assessments, a cohesive visual design system, and functional prototypes. Community Engagement Studio sessions generated actionable recommendations that were synthesized into successive prototype iterations.

XSL•FO
RenderX

Figure 4. Evolution of selected Electronic Problem-Solving Training (ePST) design elements from prototypes to final product with panels illustrating iterative, prioritized changes driven by Community Advisory Board, Community Engagement Studios, and usability feedback (eg, badge redesign, microlearning length, voiceover tone, navigation simplification).

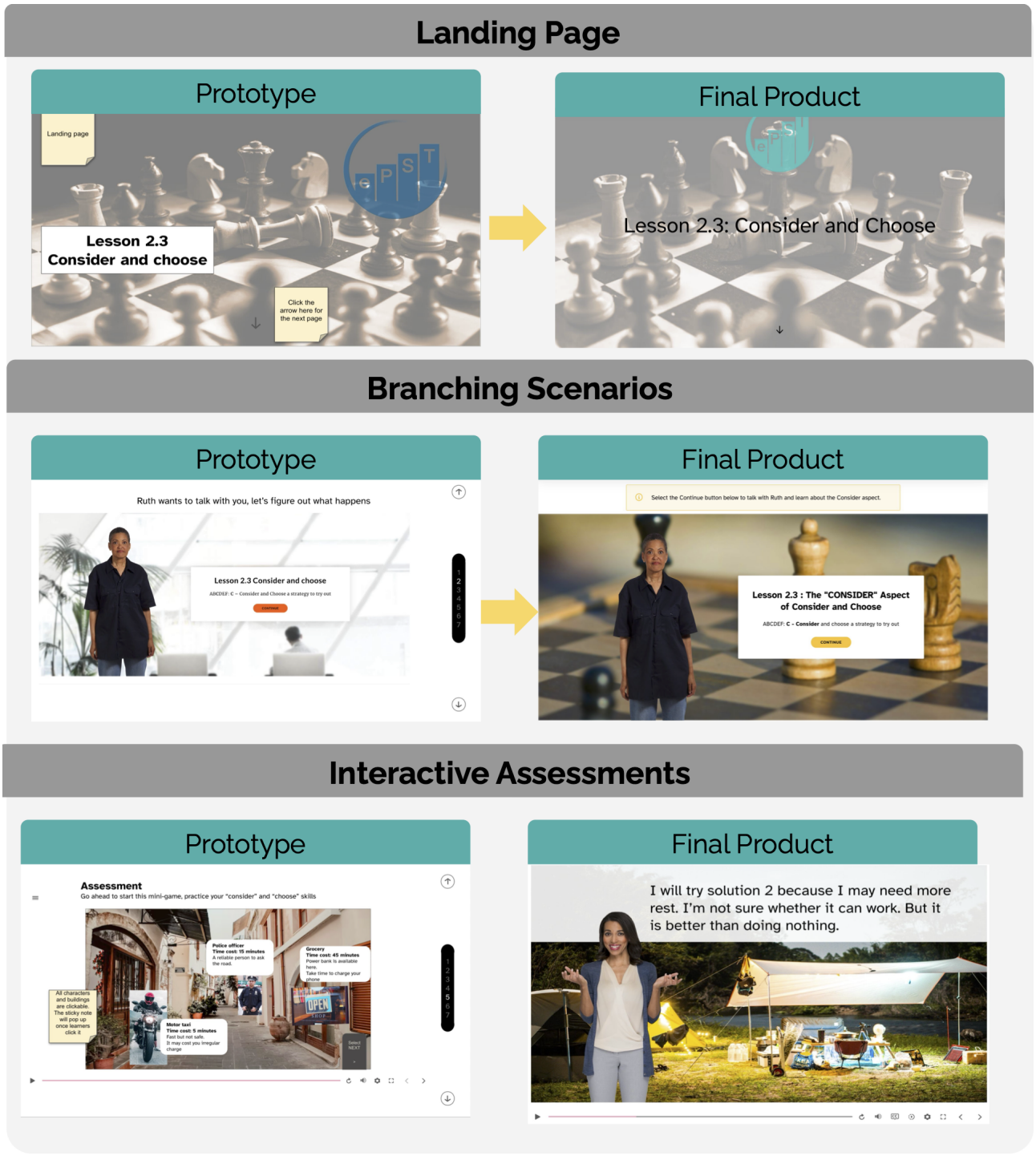

Deliver Phase Results

Deliver Phase Participant Demographics

Participant demographics for the usability study are presented in Table 4.

XSL•FO
RenderX

Table . Usability study participant demographics (n=5) including individual-level characteristics of age, gender, race or ethnicity, education, employment, years since injury, and baseline technology experience.

Participant[a]	Age (years)	Gender	Race or ethnicity	Education	Employment status	Time since injury (years)	Technology experience
Leo	47	Male	Hispanic	High school	Permanent disability	17	Capable user, no eHealth experience
Morgan	31	Female	White	Some college	Stay-at-home spouse	11	Capable user, occasional eHealth use
Alexis	50	Female	Black	Bachelor's degree	Permanent disability	20	Experienced user, frequent eHealth use
Riley	47	Male	White	Some college	Permanent disability	14	Experienced user, occasional eHealth use
Emma	36	Female	White	Some college	Stay-at-home parent	13	Experienced user, occasional eHealth use

[a]Participant names are pseudonyms.

Performance Metrics

All participants (n=5) finished every module (95% CI 56.6% - 100%). Lessons were completed efficiently, with participants spending about an average of 11.5 (SD 5.3; range 4.6 - 21.4) minutes for 10 lesson completions. Knowledge checks showed solid comprehension (8/10 items correct; 95% CI 49% - 94%; n=5, 2 items each), meeting our objectives for task efficiency and learning support. Performance metrics are summarized in Table 5.

Table . Usability performance results and knowledge assessment, including task and efficiency measures (lesson completion time, tasks per lesson, task completion rate) and knowledge-item accuracy derived from recorded usability sessions (n=5).

Metric	Result
Efficiency measures	
Lesson completion time (minutes), mean (SD)	11.47 (5.28)
Completion time - Module 2 (minutes), mean	10.50
Completion time - Module 3 (minutes), mean	13.10
Time (minutes), range	4.6 - 21.42
Tasks per user, mean	22.8
Tasks per lesson, mean	11.4
Task completion rate (tasks per minute), mean	0.996
Knowledge assessment	
Overall accuracy (% correct)	80
Question 1 accuracy (% correct)	60
Question 2 accuracy (% correct)	100

CAUSLT Usability Assessment

Participants completed the CAUSLT, which evaluates 3 dimensions of usability in educational technology using a 5-point Likert scale (1=Strongly Disagree, 5=Strongly Agree). Overall usability was high on the CAUSLT, with a mean score of 4.25 out of 5 (SD 0.72; 95% CI 3.36 - 5.15; n=5), supporting our objective that the prototype be easy to use and learn. Results are presented in Table 6 and illustrated in Figures 5-7 .

XSL•FO
RenderX

Table . Usability scores by sociotechnical-pedagogical domain for technological, pedagogical, and sociocultural usability as measured using the Comprehensive Assessment of Usability for Learning Technologies (CAUSLT).

Usability dimension	Score, mean (SD)	Score, range
Technological usability	4.06 (0.95)	3.50-5.00
Pedagogical usability	4.34 (0.77)	3.43-5.00
Sociocultural usability	4.13 (0.87)	3.00-5.00

Figure 5. Technological usability responses collected during usability testing using the Comprehensive Assessment of Usability for Learning Technologies (CAUSLT), showing domain and item-level means and SDs for the technological domain (navigation, performance, error tolerance).

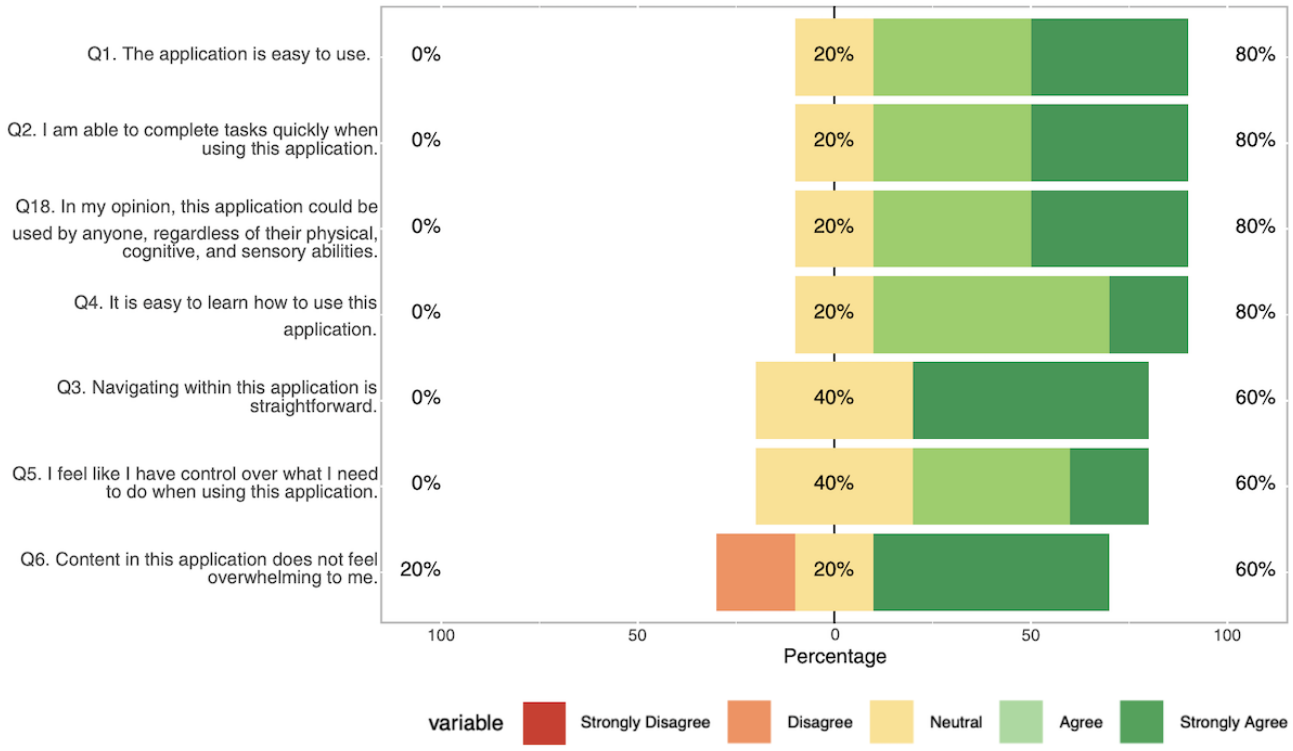

XSL•FO
RenderX

Figure 6. Pedagogical usability responses collected during usability testing using the Comprehensive Assessment of Usability for Learning Technologies (CAUSLT), showing domain and item-level means and SDs for pedagogical measures (ease of learning, clarity, learning support, engagement).

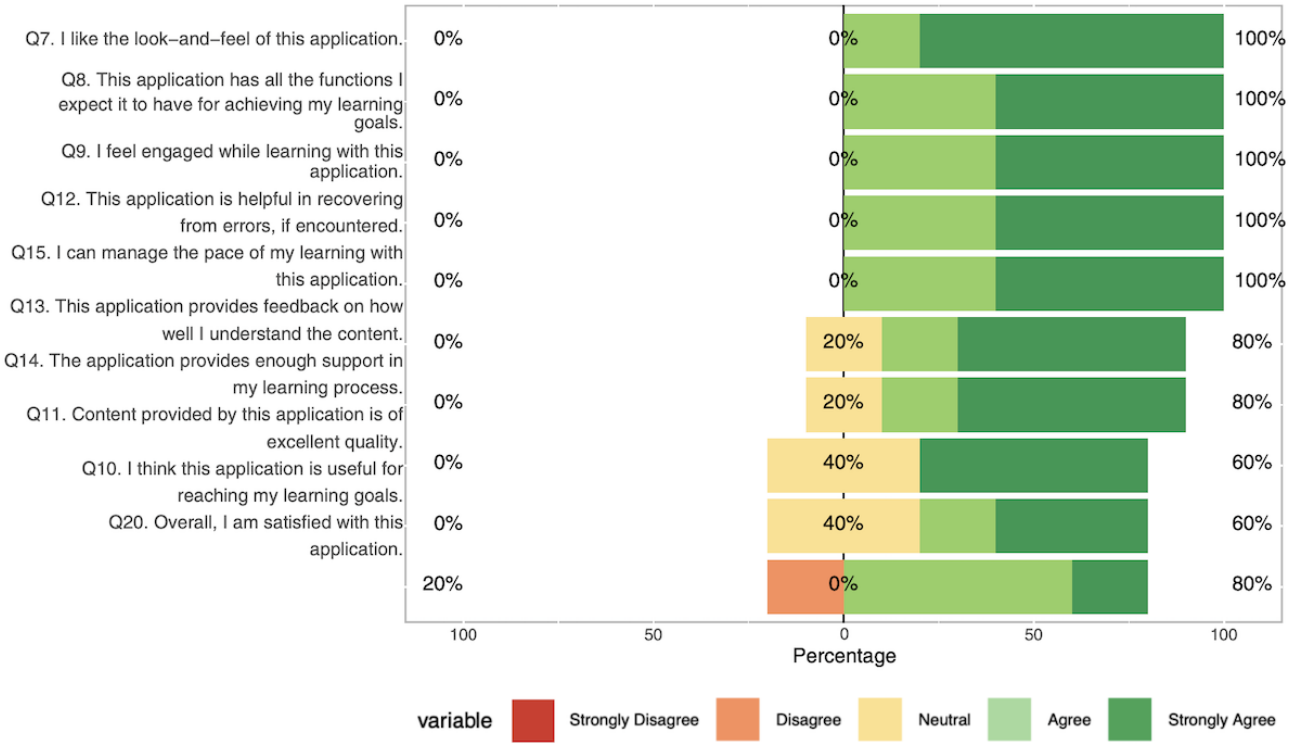

Figure 7. Sociocultural usability responses collected during usability testing using the Comprehensive Assessment of Usability for Learning Technologies (CAUSLT), showing domain means and SDs for sociocultural presence, accessibility, and relevance.

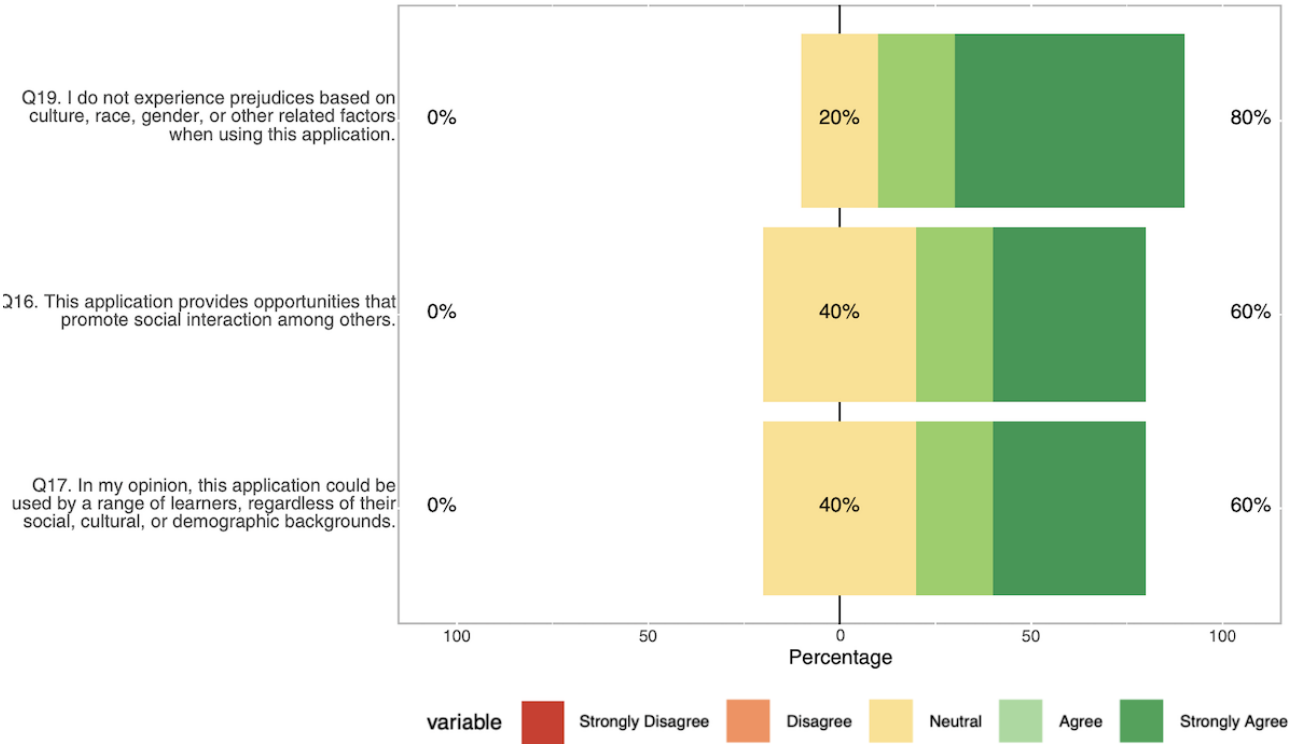

Pedagogical usability received the highest ratings (mean 4.34), with participants particularly valuing the application's look and feel (mean 4.8) and core learning functions including engagement, error recovery, and pace management (mean 4.6). Of the participants, 100% (5/5) agreed or strongly agreed on items related to learning engagement, pacing, and functional adequacy, with only feedback-related items showing some neutral responses (1/5, 20%). Technological usability scored well overall (mean 4.06), with strongest ratings for ease of use, task completion speed, and accessibility across different abilities. We found 80% (4/5) agreement across most technological items, though navigation and user control received slightly lower ratings (3/5, 60% agreement), and content overwhelm was the only item receiving any disagreement (1/5, 20%). Sociocultural

XSL•FO
RenderX

usability (mean 4.13) showed more variability. Participants were most confident about avoiding cultural prejudices (4/5, 80% agreement), while social interaction opportunities and cross-demographic accessibility both received 60% (3/5) agreement, with higher levels of neutrality (2/5, 40%), indicating potential areas for enhancement.

Key Usability Findings

Qualitative analysis of think-aloud transcripts and observation notes revealed both significant strengths and areas requiring refinement in ePST's usability. Content analysis identified patterns across participants' experiences that highlight the application's effectiveness at engaging users with TBI while revealing specific technical and interface challenges that impact user experience (Table 7). These findings provided actionable insights for iterative design improvements, which were incorporated between each usability testing session.

Table . Usability strengths and priority areas for improvement, summarizing positively rated features and recurrent problems identified from think-aloud protocols, observer notes, Comprehensive Assessment of Usability for Learning Technologies (CAUSLT) responses, and Community Advisory Board and Community Engagement Studio review.

Finding	Description
Usability strengths	
Intuitive interface design	Participants navigated the application easily and found interactive elements engaging. One participant described the storyline object as "pretty cool," indicating positive reception of multimedia components.
Effective progress tracking	Badge system and progress indicators were clearly understood and valued by users. Representative quotes: "It looks like I've received one badge 1,2,3, and five more to go" and "It proves to me that I've done something."
Engaging multimedia elements	Varied voiceover tones, storytelling approach, and visual design elements received positive feedback. Participants appreciated the narrative-based learning style and accessibility features.
Successful error recovery	Adaptive feedback mechanisms enabled users to recover from errors in knowledge checks without significant frustration, maintaining learning continuity.
Areas for improvement	
Mobile navigation issues	Users experienced confusion with mobile interface controls and activity progression. Representative quotes: "I see a button on the bottom right that looks like a back arrow" and "Wait, where was 3.2?"
Content comprehension	Some participants misunderstood instructions or content elements. One participant stated "Wait, this is not a question" when encountering a Storyline component.
Technical performance	Loading delays and playback issues disrupted user experience. Representative quotes: "Oh, it has to load all over again" and "I can't see the whole screen" (mobile display problems).

Discussion

Principal Findings

We applied a CBPR and learning experience design–guided formative design process to develop and evaluate ePST and addressed 3 core questions about front-end needs, sociotechnical-pedagogical usability, and how identified issues were resolved through design refinements. Usability was high across all domains, knowledge accuracy was 80% (an encouraging result for formative testing suggesting acceptable immediate comprehension), and mean time-on-task was 11.47 minutes per lesson while engaging in the think-aloud protocol. Participatory activities produced concrete design changes (ie, microlearning 5-12–minute lessons, badge refinements, voiceover adjustments), helped identify partner-specific priorities (ie, caregiver, clinician, lived experience perspectives), and revealed TBI-specific requirements (ie, linear progression, higher technical performance, explicit content signaling).

Taken together, these findings suggest that sustained community engagement can yield measurable usability improvements and actionable implementation guidance for TBI mHealth interventions. These outcomes map directly to established digital health usability constructs of effectiveness, efficiency, and satisfaction (ISO 9241 - 11) [121] and to mHealth-specific evaluation guidance such as the validated mHealth App Usability Questionnaire [122]. Our combined questionnaire plus think-aloud pipeline also follows human factors and usability engineering recommendations for medical and mHealth systems (IEC 62366) [123,124] and recent mHealth usability reviews [125,126].

Iterative Community Feedback Enhanced Technical Usability

The Community Advisory Board structure (33 diverse community partners) and structured Community Engagement Studio sessions (n=10) enabled systematic integration of community input across development phases. CAUSLT scores averaged 87.3 out of 100, with pedagogical usability receiving

XSL•FO
RenderX

the highest ratings. Participants completed lessons efficiently and achieved 80% accuracy on knowledge assessments, comparing favorably to cognitive rehabilitation intervention outcomes reported in systematic reviews [127]. Empathy interviews with TBI survivors revealed specific cognitive load concerns that directly informed the microlearning approach (5‐12–minute lessons) and progress tracking features. Community Advisory Board feedback on early prototypes resulted in modification of the badge system design and influenced voiceover tone selection to reduce perceived condescension. These modifications were fundamental design decisions that addressed cognitive accessibility requirements identified through community input [128]. Importantly, usability issues identified through think-aloud protocols mapped directly to areas where Community Advisory Board input had been limited or where technical constraints overrode community recommendations, suggesting that user involvement depth correlates with usability outcomes [129].

Multistakeholder Representation Identified Comprehensive Design Requirements

The Community Advisory Board's composition systematically identified design considerations that single partner approaches typically overlook. Caregivers identified family involvement features, while clinicians contributed evidence-based content validation, and individuals with lived experience prioritized autonomy and stigma reduction elements. This multiperspective input directly shaped the sociocultural usability features that scored highly in evaluation, particularly around cultural responsiveness and inclusive design [130]. Unlike traditional focus groups or surveys, the sustained Community Advisory Board engagement spanning the entire development cycle allowed for iterative refinement based on evolving understanding of user needs. This depth of engagement appeared to contribute to high pedagogical usability scores and enabled authentic relationship-building rather than extractive consultation [72].

TBI-Specific Technology Design Requirements Emerged

The usability evaluation revealed specific design requirements for cognitive rehabilitation technology that extend beyond general accessibility guidelines. Analysis of user interactions demonstrated that traditional e-learning design principles require significant adaptation for users with cognitive impairments, consistent with cognitive load theory applications in special populations [131]. The 21-minute range in task completion times (range 4.6‐21.42 min) revealed that cognitive processing variability in TBI populations requires deliberate architectural choices rather than standard responsive design. Participants performed optimally with linear content progression and struggled with branching navigation structures, suggesting that linear content progression may reduce cognitive demands relative to complex navigation structures for users with executive function deficits [127].

Navigation issues identified in think-aloud protocols were predominantly mobile-specific, with participants reporting confusion about interface cues ("I see a button on the bottom right that looks like a back arrow") and progression sequences. Technical performance issues disproportionately disrupted learning flow for participants with attention deficits, suggesting that cognitive rehabilitation technology requires higher technical performance standards than typical educational applications [132]. Although participants appreciated multimedia elements and voiceover variety, content comprehension issues arose when instructional clarity was sacrificed for engagement, suggesting that TBI rehabilitation technology might require explicit signaling of content types and interaction expectations, with clarity taking precedence over novel interface design [133].

Methodological Contributions

This study contributes methodological insights for implementing CBPR in rehabilitation technology development. The integration of Community Advisory Board and Community Engagement Studio structures with learning experience design principles demonstrates how participatory research can move beyond consultation to systematic co-design. The sustained engagement model (33 diverse partners across the entire development cycle) provides a replicable framework for authentic community involvement, providing an actionable alternative to extractive research practices. The mapping of community input to specific design modifications illustrates how participatory methods can produce measurable technical improvements, not merely ethical and accessibility compliance. Findings support the claim that CBPR's value extends beyond moral imperatives to offer practical advantages in rehabilitation technology effectiveness.

The literature consistently supports that technology development through iterative user-centered design is associated with higher adherence and lower abandonment [134-139]. This more frequent and consistent engagement leads to clinical benefits [134,140]. Additionally, high usability facilitates scale-up and sustainability [135]. Despite this, development of digital health care technologies often fails to include patient, client, and clinician voices through early and ongoing user-engagement [134,141,142]. A recent scoping review [139] on reasons for abandonment of behavioral and mental health mobile interventions found 6 categories of reasons for abandonment, 3 of which could be directly addressed through user-centered and participatory design: (1) poor user experience, (2) evolving user needs and goals, and (3) content and features.

There is a growing body of literature specifically in rehabilitation supporting that usability, acceptability, and user-centered design contribute to implementation and sustainability of remote, technology-support interventions [28,143-146], but substantial work still needs to be done. A systematic review of cognitive rehabilitation interventions for older adults found that usability and user experience often explained mixed effectiveness of these technology-based interventions [147]. Though an even smaller body of research, a few studies have examined user-centered design for assistive technology and cognitive rehabilitation interventions for people with TBI [28,143-146]. These papers, consistent with our own findings, emphasized the importance of (1) tailoring the technology to reduce cognitive load; (2) having high error tolerance and easy error correction; (3) including multimodal prompts; and (4) involving clinicians, care partners, and survivors in technology design. Evidence in TBI is smaller and more heterogeneous than in general digital mental health, but

XSL•FO
RenderX

findings consistently point to usability as a facilitating factor for adoption and benefit.

This study's contribution is integrative rather than disciplinary. We operationalized a full-cycle pipeline that combines community-based participatory research with learning experience design; mapped participatory inputs onto a sociotechnical-pedagogical evaluation lens; and triangulated think-aloud, task, and survey metrics to produce community-informed design principles for TBI mHealth. Taken together, this cross-disciplinary operationalization provides a reproducible, pragmatic approach for formative mHealth development in cognitive rehabilitation and offers concrete, testable design guidance for teams working at the intersection of participatory methods, instructional design, and digital health.

Limitations and Future Directions

Several limitations constrain the generalizability of our findings. Usability testing used a small, purposive sample (n=5) appropriate for formative evaluation but insufficient for population-level inferences. Consequently, the effect estimates (eg, CAUSLT mean, accuracy) had wide confidence intervals; therefore, subgroup effects could not be assessed. Thus CAUSLT mean, knowledge accuracy, and completion rates should be viewed as exploratory. Our design mitigations were purposive sampling for heterogeneity, triangulation across qualitative and quantitative data streams, and sustained Community Advisory Board engagement to improve ecological validity. Nonetheless, future work should evaluate ePST in larger, more diverse TBI samples to quantify variability across injury characteristics, device types, demographic groups, and contexts of use and to permit powered hypothesis testing and subgroup analysis, which is the focus of our current feasibility study. Further, some reported technical issues may reflect device-specific limitations rather than design flaws, indicating need for expanded cross-platform testing. In addition to this, the TBI-specific design features reported here may not transfer directly to other neurological populations, requiring investigation of how CBPR-based approaches perform across different rehabilitation contexts. Although initial usability testing revealed strong satisfaction, sustainability of engagement remains unknown, suggesting a need for longitudinal metrics capturing retention, adherence, and health outcome durability. Future research should focus on evaluating barriers and facilitators to adoption and abandonment and how this engagement (or lack thereof) affects scale-up and sustainability of health care technologies using digital health frameworks such as the NASSS (nonadoption, abandonment, scale-up, spread, and sustainability) framework [135].

A direction for future research is how participatory practices might influence long-term health outcomes and treatment adherence beyond usability metrics. Integration of adaptive technologies such as artificial intelligence–driven personalization, voice-guided prompts, and real-time support could represent promising directions for accommodating cognitive variability in neurological populations. Additionally, examining the scalability of intensive CBPR approaches across diverse rehabilitation contexts (ie, stroke recovery, spinal cord injury) could advance understanding of participatory design's broader applicability.

Conclusions

This study demonstrated that systematic application of CBPR principles can produce both qualitative and quantitative improvements in rehabilitation technology usability through iterative community feedback, diverse stakeholder representation, and sustained engagement processes. The development of ePST illustrates how participatory methods can address specific design requirements for cognitive accessibility while maintaining high user satisfaction. The findings suggest that cognitive rehabilitation technology can benefit from specific design considerations including attention to cognitive load, clear navigation patterns, and explicit content signaling to address TBI-related challenges. This work provides further support for CBPR as a practical methodology in rehabilitation technology development, enhancing ethical research practices as well as technical outcomes. Investigation of long-term engagement sustainability and adaptive technology integration remains a direction for future research with promise for advancing understanding of how participatory approaches might contribute to more equitable, personalized, and effective rehabilitation interventions.

Acknowledgments

We used ChatGPT to assist with editing for grammar, creating figure and table captions, and enhancing readability of lengthy and complex sentence formulations. All artificial intelligence (AI) outputs were reviewed, edited, and approved by the named authors.

Funding

This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs through the Congressionally Directed Medical Research Programs TBI and Psychological Health program under award number HT9425-23-1-0567. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense.

Data Availability

All data are available in the manuscript and multimedia appendices.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Traumatic brain injury (TBI) personas and empathy maps for user-centered design: 11 de-identified, data-informed personas and associated empathy-map summaries created from Community Advisory Board (CAB)/Community Engagement Studios (CES) and empathy-interview data to guide Electronic Problem-Solving Training (ePST) module design and accessibility decisions. Each persona includes demographics (age, language, location), TBI characteristics and time-since-injury, goals, behaviors, attitudes, motivations, barriers, facilitators, and concise "key attributes" used to prioritize features (eg, linear lesson flow, memory supports, fatigue accommodations, family involvement).
[PDF File, 2560 KB - jmir_v28i1e83995_app1.pdf]

Checklist 1
GRIPP 2 Short Form checklist.
[PDF File, 77 KB - jmir_v28i1e83995_app2.pdf]

References

1. Toro-Hernández ML, Mondragón-Barrera A, Múnera-Orozco S, Villa-Torres L, Camelo-Castillo W. Experiences with rehabilitation and impact on community participation among adults with physical disability in Colombia: perspectives from stakeholders using a community based research approach. Int J Equity Health 2019 Jun 3;18(1):18. [doi: 10.1186/s12939-019-0923-4] [Medline: 31155006]
2. Khayatzadeh-Mahani A, Wittevrongel K, Nicholas DB, Zwicker JD. Prioritizing barriers and solutions to improve employment for persons with developmental disabilities. Disabil Rehabil 2020 Sep;42(19):2696-2706. [doi: 10.1080/09638288.2019.1570356] [Medline: 30856355]
3. Kersey J, Garcia P, Evans E, et al. Underrepresentation of participants from marginalized racial and ethnic groups: a secondary analysis of the cognitive rehabilitation literature. Arch Rehabil Res Clin Transl 2025 Jun;7(2):100431. [doi: 10.1016/j.arrct.2025.100431] [Medline: 40678291]
4. Omar S, Williams CC, Bugg LB, Colantonio A. "Somewhere along the line, your mask isn't going to be fitting right": institutional racism in Black narratives of traumatic brain injury rehabilitation across the practice continuum. BMC Health Serv Res 2024 Jul 24;24(1):834. [doi: 10.1186/s12913-024-10986-1] [Medline: 39049041]
5. Maas AIR, Menon DK, Manley GT, et al. Traumatic brain injury: progress and challenges in prevention, clinical care, and research. Lancet Neurol 2022 Nov;21(11):1004-1060. [doi: 10.1016/S1474-4422(22)00309-X] [Medline: 36183712]
6. Taylor CA, Bell JM, Breiding MJ, Xu L. Traumatic brain injury-related emergency department visits, hospitalizations, and deaths - United States, 2007 and 2013. MMWR Surveill Summ 2017 Mar 17;66(9):1-16. [doi: 10.15585/mmwr.ss6609a1] [Medline: 28301451]
7. Polinder S, Cnossen MC, Real RGL, et al. A Multidimensional approach to post-concussion symptoms in mild traumatic brain injury. Front Neurol 2018;9:1113. [doi: 10.3389/fneur.2018.01113] [Medline: 30619066]
8. Azouvi P, Arnould A, Dromer E, Vallat-Azouvi C. Neuropsychology of traumatic brain injury: an expert overview. Rev Neurol (Paris) 2017;173(7-8):461-472. [doi: 10.1016/j.neurol.2017.07.006] [Medline: 28847474]
9. Milders M. Relationship between social cognition and social behaviour following traumatic brain injury. Brain Inj 2019;33(1):62-68. [doi: 10.1080/02699052.2018.1531301] [Medline: 30325217]
10. Semple BD, Zamani A, Rayner G, Shultz SR, Jones NC. Affective, neurocognitive and psychosocial disorders associated with traumatic brain injury and post-traumatic epilepsy. Neurobiol Dis 2019 Mar;123:27-41. [doi: 10.1016/j.nbd.2018.07.018] [Medline: 30059725]
11. Ertas-Spantgar F, Korabova S, Gabel A, Schiering I, Müller SV. Guiding patients with traumatic brain injury through the instrumental activities of daily living with the RehaGoal App: a feasibility study. Disabil Rehabil Assist Technol 2024 Feb;19(2):254-265. [doi: 10.1080/17483107.2022.2080290] [Medline: 35713480]
12. Hou Y, Zhou A, Brooks L, Reid D, Turkstra L, MacDonald S. Rehabilitation access for individuals with cognitive-communication challenges after traumatic brain injury: a co-design study with persons with lived experience. Int J Lang Commun Disord 2024;59(2):648-664. [doi: 10.1111/1460-6984.12895] [Medline: 37189286]
13. Novak I, te Velde A, Hines A, et al. Rehabilitation evidence-based decision-making: the READ model. Front Rehabilit Sci 2021 Oct 5;2. [doi: 10.3389/fresc.2021.726410]
14. Doerwald F, Stalling I, Recke C, et al. A rapid review of digital approaches for the participatory development of health-related interventions. Front Public Health 2024;12:1461422. [doi: 10.3389/fpubh.2024.1461422] [Medline: 39678234]
15. Benz C, Scott-Jeffs W, McKercher KA, et al. Community-based participatory-research through co-design: supporting collaboration from all sides of disability. Res Involv Engagem 2024 May 10;10(1):47. [doi: 10.1186/s40900-024-00573-3] [Medline: 38730283]

XSL•FO
RenderX

16. Kilfoy A, Hsu TCC, Stockton-Powdrell C, Whelan P, Chu CH, Jibb L. An umbrella review on how digital health intervention co-design is conducted and described. NPJ Digit Med 2024 Dec 23;7(1):374. [doi: 10.1038/s41746-024-01385-1] [Medline: 39715947]
17. Perrin PB, Haun JN, Klyce DW, et al. Efficacy and implementation planning across the Veterans Affairs polytrauma system of care: protocol for the REACH intervention for caregivers of veterans and service members with traumatic brain injury. JMIR Res Protoc 2024 Aug 15;13:e57692. [doi: 10.2196/57692] [Medline: 39145996]
18. Castro-Figueroa E, Rosario-Maldonado FJ, Asencio-Toro G, et al. Empowering community partners in health disparities research: refining a community-based participatory research (CBPR) training curriculum. Pedagogy Health Promot 2025. [doi: 10.1177/23733799251345430]
19. Mohatt NV, Kreisel CJ, Brenner LA, CRITICAL Team. Engaging those living with moderate to severe TBI and their caregivers in research. J Patient Exp 2021;8(2374373521998852):2374373521998852. [doi: 10.1177/2374373521998852] [Medline: 34179408]
20. Haun JN, Nakase-Richardson R, Cotner BA, et al. Stakeholder engagement to identify implementation strategies to overcome barriers to delivering chronic pain treatments: a NIDILRR and VA TBI model systems collaborative project. J Head Trauma Rehabil 2024;39(1):E29-E40. [doi: 10.1097/HTR.0000000000000920] [Medline: 38167720]
21. Eady K, Moreau KA, Marshall S, Egan M. Patient, family, and health professional perspectives of how families are involved in adult inpatient traumatic brain injury rehabilitation. Patient Exp J 2024 Nov 14;11(3):29-36. [doi: 10.35680/2372-0247.1891]
22. Cameron LJ, Somerville LM, Naismith CE, Watterson D, Maric V, Lannin NA. A qualitative investigation into the patient-centered goal-setting practices of allied health clinicians working in rehabilitation. Clin Rehabil 2018 Jun;32(6):827-840. [doi: 10.1177/0269215517752488] [Medline: 29327603]
23. Hoffman JM, Curran M, Barber J, Lucas S, Fann JR, Zumsteg JM. Collaborative care for chronic pain after traumatic brain injury: a randomized clinical trial. JAMA Netw Open 2024 Jun 3;7(6):e2413459. [doi: 10.1001/jamanetworkopen.2024.13459] [Medline: 38829619]
24. D'Cruz K, Antonopoulos S, Rothman R, Douglas J, Winkler D, Oliver S. Co-designing with adults with acquired neurological disability in the community: a scoping review protocol. BMJ Open 2022 Dec 8;12(12):e064921. [doi: 10.1136/bmjopen-2022-064921] [Medline: 36600382]
25. Curran MC, Lucas S, Fann JR, Zumsteg JM, Hoffman JM. Chronic pain after traumatic brain injury: a collaborative care approach. Front Rehabil Sci 2024;5:1398856. [doi: 10.3389/fresc.2024.1398856] [Medline: 39253025]
26. Morrow EL, Mayberry LS, Duff MC. The growing gap: a study of sleep, encoding, and consolidation of new words in chronic traumatic brain injury. Neuropsychologia 2023 Jun 6;184:108518. [doi: 10.1016/j.neuropsychologia.2023.108518] [Medline: 36804844]
27. Manley K, Saunders K, Wilkinson D, Faruqui R, Sakel M. Co-creating system-wide improvement for people with traumatic brain injury across one integrated care system in the United Kingdom to initiate a transformation journey through co-production. Health Expect 2023 Apr;26(2):869-881. [doi: 10.1111/hex.13712] [Medline: 36715266]
28. Quilico E, Wilkinson S, Duncan L, et al. Participatory co-creation of an adapted physical activity program for adults with moderate-to-severe traumatic brain injury. Front Rehabil Sci 2022;3:900178. [doi: 10.3389/fresc.2022.900178] [Medline: 36188895]
29. Power E, Morrow R. Digital, co-created implementation of communication partner training programs for stroke, brain injury, and dementia: past, present, and future. Int J Speech Lang Pathol 2024 Jun;26(3):317-333. [doi: 10.1080/17549507.2024.2362856] [Medline: 38962904]
30. Karhula M, Saukkonen S, Kinnunen A, Heiskanen T, Xiong E, Anttila H. ICF-luokituksen yksilötekijöiden kuvaus on osa toimintakyvyn laaja-alaista arviointia: kartoittava kirjallisuuskatsaus ICF-yksilötekijöitä käsittelevistä tutkimuksista. Kuntoutus 2021 Jun 18;44(2):9-24. [doi: 10.37451/kuntoutus.109476]
31. Salsberg J, Macridis S, Garcia Bengoechea E, Macaulay AC, Moore S, KSDPP School Travel Planning Committee. The shifting dynamics of social roles and project ownership over the lifecycle of a community-based participatory research project. Fam Pract 2017 Jun 1;34(3):305-312. [doi: 10.1093/fampra/cmx006] [Medline: 28334748]
32. Burduladze N, Jones LP, Jones BD, et al. Exploring power and power sharing in participatory health research partnerships: a scoping review protocol. PLOS ONE 2024;19(7):e0303799. [doi: 10.1371/journal.pone.0303799] [Medline: 39024348]
33. Manhas KP, Olson K, Churchill K, Vohra S, Wasylak T. Experiences of shared decision-making in community rehabilitation: a focused ethnography. BMC Health Serv Res 2020 Apr 19;20(1):329. [doi: 10.1186/s12913-020-05223-4] [Medline: 32306972]
34. Nezu AM, Nezu CM. Problem solving. In: Norcross JC, VandenBos GR, Freedheim DK, Pole N, editors. APA Handbook of Clinical Psychology: Psychopathology and Health: American Psychological Association; 2016:449-460. [doi: 10.1037/14862-019]
35. D'Zurilla TJ, Nezu AM, Maydeu-Olivares A. Social problem solving: theory and assessment. In: Chang EC, D'Zurilla TJ, Sanna LJ, editors. Social Problem Solving: Theory, Research, and Training 2004. URL: http://psycnet.apa.org/psycinfo/2004-14507-001

XSL•FO
RenderX

36. D'Zurilla TJ, Nezu AM. Problem-Solving Therapy: A Positive Approach to Clinical Intervention, 3rd edition: Springer Publishing Company; 2007.
37. D'Zurilla TJ, Goldfried MR. Problem solving and behavior modification. J Abnorm Psychol 1971 Aug;78(1):107-126. [doi: 10.1037/h0031360] [Medline: 4938262]
38. Jiang C, Zhou H, Chen L, Zhou Z. Problem solving therapy improves effortful cognition in major depression. Front Psychiatry 2021;12:607718. [doi: 10.3389/fpsyt.2021.607718] [Medline: 33897483]
39. Cuijpers P, de Wit L, Kleiboer A, Karyotaki E, Ebert DD. Problem-solving therapy for adult depression: an updated meta-analysis. Eur psychiatr 2018;48(1):27-37. [doi: 10.1016/j.eurpsy.2017.11.006]
40. Narad ME, Raj S, Yeates KO, et al. Randomized controlled trial of an online problem-solving intervention following adolescent traumatic brain injury: family outcomes. Arch Phys Med Rehabil 2019 May;100(5):811-820. [doi: 10.1016/j.apmr.2019.01.010] [Medline: 30738021]
41. Zhang N, Kaizar EE, Narad ME, et al. Examination of injury, host, and social-environmental moderators of online family problem solving treatment efficacy for pediatric traumatic brain injury Using an individual participant data meta-analytic approach. J Neurotrauma 2019 Apr 1;36(7):1147-1155. [doi: 10.1089/neu.2018.5885] [Medline: 30328749]
42. Palermo TM, Law EF, Bromberg M, Fales J, Eccleston C, Wilson AC. Problem-solving skills training for parents of children with chronic pain: a pilot randomized controlled trial. Pain 2016 Jun;157(6):1213-1223. [doi: 10.1097/j.pain.0000000000000508] [Medline: 26845525]
43. Economides M, Ranta K, Nazander A, et al. Long-term outcomes of a therapist-supported, smartphone-based intervention for elevated symptoms of depression and anxiety: quasiexperimental, pre-postintervention study. JMIR Mhealth Uhealth 2019 Aug 26;7(8):e14284. [doi: 10.2196/14284] [Medline: 31452521]
44. Ghanbari E, Yektatalab S, Mehrabi M. Effects of psychoeducational interventions using mobile apps and mobile-based online group discussions on anxiety and self-esteem in women with breast cancer: randomized controlled trial. JMIR Mhealth Uhealth 2021 May 18;9(5):e19262. [doi: 10.2196/19262] [Medline: 34003138]
45. Floor N. This Is Learning Experience Design: What It Is, How It Works, and Why It Matters: New Riders; 2023.
46. Israel BA, Eng E, Schulz AJ, Parker EA. Methods for Community-Based Participatory Research for Health: John Wiley & Sons; 2005.
47. Lim H, Kakonge L, Hu Y, et al. So, i can feel normal: participatory design for accessible social media sites for individuals with traumatic brain injury. 2023 Apr 19 Presented at: CHI '23; Apr 19, 2023; Hamburg Germany p. 1-19 URL: https://dl.acm.org/doi/proceedings/10.1145/3544548 [doi: 10.1145/3544548.3581222]
48. Lorenz EA, Bråten Støen A, Lie Fridheim M, Alsos OA. Design recommendations for XR-based motor rehabilitation exergames at home. Front Virtual Real 2024 Jan 22;5. [doi: 10.3389/frvir.2024.1340072]
49. Correa DJ, Kwon CS, Connors S, et al. Applying participatory action research in traumatic brain injury studies to prevent post-traumatic epilepsy. Neurobiol Dis 2019 Mar;123:137-144. [doi: 10.1016/j.nbd.2018.07.007] [Medline: 30031158]
50. Groussard PY, Pigot H, Giroux S. From conception to evaluation of mobile services for people with head injury: a participatory design perspective. Neuropsychol Rehabil 2018 Jul;28(5):667-688. [doi: 10.1080/09602011.2015.1117499] [Medline: 26679473]
51. Coombe CM, Schulz AJ, Guluma L, et al. Enhancing capacity of community-academic partnerships to achieve health equity: results from the CBPR Partnership Academy. Health Promot Pract 2020 Jul;21(4):552-563. [doi: 10.1177/1524839918818830] [Medline: 30596283]
52. Springer MV, Skolarus LE. Community-based participatory research. Stroke 2019 Mar;50(3):e48-e50. [doi: 10.1161/STROKEAHA.118.024241] [Medline: 30661505]
53. Bonn MM, Graham LJ, Marrocco S, Jeske S, Moran B, Wolfe DL. Usability evaluation of a self-management mobile application for individuals with a mild traumatic brain injury. Digit Health 2023;9:20552076231183555. [doi: 10.1177/20552076231183555] [Medline: 37426589]
54. Smith KA, Ward T, Lambe S, et al. Engagement and attrition in digital mental health: current challenges and potential solutions. NPJ Digit Med 2025 Jul 2;8(1):398. [doi: 10.1038/s41746-025-01778-w] [Medline: 40604240]
55. Abbasalizadeh M, Farsi Z, Sajadi SA, Atashi A. The effect of mobile health application training based on micro-learning method on the level of resilience and happiness among intensive care nurses: a randomized controlled trial. BMC Psychiatry 2024 Dec 27;24(1):954. [doi: 10.1186/s12888-024-06429-0] [Medline: 39731084]
56. Jacob C, Lindeque J, Müller R, et al. A sociotechnical framework to assess patient-facing eHealth tools: results of a modified Delphi process. NPJ Digit Med 2023 Dec 15;6(1):232. [doi: 10.1038/s41746-023-00982-w] [Medline: 38102323]
57. Lin MCM, Vasarhelyi K, Wong KLY, et al. Engaging community to co-design learning health systems: lessons from storytelling and Design Jam, a community case study from British Columbia, Canada. Front Health Serv 2025;5:1620659. [doi: 10.3389/frhs.2025.1620659] [Medline: 40800072]
58. Joosten YA, Israel TL, Williams NA, et al. Community Engagement Studios: a structured approach to obtaining meaningful input from stakeholders to inform research. Acad Med 2015 Dec;90(12):1646-1650. [doi: 10.1097/ACM.0000000000000794] [Medline: 26107879]

XSL•FO
RenderX

59. Zisman-Ilani Y, Buell J, Mazel S, Hennig S, Nicholson J. Virtual Community Engagement Studio (V-CES): engaging mothers with mental health and substance use conditions in research. Front Psychiatry 2022;13:805781. [doi: 10.3389/fpsyt.2022.805781] [Medline: 35782439]
60. Vial S, Boudhraâ S, Dumont M. Human-centered design approaches in digital mental health interventions: exploratory mapping review. JMIR Ment Health 2022 Jun 7;9(6):e35591. [doi: 10.2196/35591] [Medline: 35671081]
61. Banbury A, Pedell S, Parkinson L, Byrne L. Using the Double Diamond model to co-design a dementia caregivers telehealth peer support program. J Telemed Telecare 2021 Dec;27(10):667-673. [doi: 10.1177/1357633X211048980] [Medline: 34726994]
62. Kersey J, Alimi E, McArthur AR, et al. ENGAGE-TBI: adaptation of a community-based intervention to improve social participation after brain injury. Brain Inj 2025 May 12;39(6):518-525. [doi: 10.1080/02699052.2025.2449927] [Medline: 39773100]
63. Zhang Y, Xie YJ, Yang L, et al. Community-based participatory research (CBPR) approaches in vaccination promotion: a scoping review. Int J Equity Health 2024 Nov 5;23(1):227. [doi: 10.1186/s12939-024-02278-1] [Medline: 39501299]
64. Kayes NM, Martin RA, Bright FA, Kersten P, Pollock A. Optimizing the real-world impact of rehabilitation reviews: increasing the relevance and usability of systematic reviews in rehabilitation. Eur J Phys Rehabil Med 2019 Jun;55(3):331-341. [doi: 10.23736/S1973-9087.19.05793-9] [Medline: 30990002]
65. Grindell C, Sanders T, Bec R, Mary Tod A, Wolstenholme D. Improving knowledge mobilisation in healthcare: a qualitative exploration of creative co-design methods. Evid Policy 2022 May 1;18(2):265-290. [doi: 10.1332/174426421X16436512504633]
66. Israel BA, Schulz AJ, Coombe CM, et al. Community-based participatory research: an approach to research in the urban context. In: Galea S, Ettman CK, Vlahov D, editors. Urban Health: Oxford University Press; 2019. [doi: 10.1093/oso/9780190915858.003.0029]
67. Ferreira MP, Gendron F. Community-based participatory research with traditional and indigenous communities of the Americas: historical context and future directions. International Journal of Critical Pedagogy ;3(3) [FREE Full text]
68. Wallerstein N, Muhammad M, Sanchez-Youngman S, et al. Power dynamics in community-based participatory research: a multiple-case study analysis of partnering contexts, histories, and practices. Health Educ Behav 2019 Oct;46(1_suppl):19S-32S. [doi: 10.1177/1090198119852998] [Medline: 31549557]
69. Plamondon K, Ndumbe-Eyoh S, Shahram S. Equity, power, and transformative research coproduction. In: Graham ID, Rycroft-Malone J, Kothari A, McCutcheon C, editors. Research Co‐Production in Healthcare 2022:34-53. [doi: 10.1002/9781119757269.ch3]
70. Collins SE, Clifasefi SL, Stanton J, et al. Community-based participatory research (CBPR): towards equitable involvement of community in psychology research. Am Psychol 2018 Oct;73(7):884-898. [doi: 10.1037/amp0000167] [Medline: 29355352]
71. Agyepong IA, Godt S, Sombie I, Binka C, Okine V, Ingabire MG. Strengthening capacities and resource allocation for co-production of health research in low and middle income countries. BMJ 2021 Feb 15;372:n166. [doi: 10.1136/bmj.n166] [Medline: 33593725]
72. Newman SD, Andrews JO, Magwood GS, Jenkins C, Cox MJ, Williamson DC. Community advisory boards in community-based participatory research: a synthesis of best processes. Prev Chronic Dis 2011 May;8(3):A70. [Medline: 21477510]
73. Safo S, Cunningham C, Beckman A, Haughton L, Starrels JL. "A place at the table:" a qualitative analysis of community board members' experiences with academic HIV/AIDS research. BMC Med Res Methodol 2016 Jul 11;16(1):80. [doi: 10.1186/s12874-016-0181-8] [Medline: 27401678]
74. Biondo J, Johnson N. The process and significance of convening a community advisory board with individuals with severe mental illness. Journal of Participatory Research Methods 2025;6(2). [doi: 10.35844/001c.130021]
75. Duke M. Community-based participatory research. Oxford Research Encyclopedia of Anthropology 2020. [doi: 10.1093/acrefore/9780190854584.013.225]
76. Abraham H, Anyetei-Anum GP, Krogman A, et al. The HEALERS: a patient, community, and stakeholder advisory board focus group series to refine a novel virtual world-based cardiac rehabilitation intervention and clinical trial. Front Digit Health 2025;7:1427539. [doi: 10.3389/fdgth.2025.1427539] [Medline: 40735343]
77. Hornbuckle LM, Rauer A. Engaging a community advisory board to inform an exercise intervention in older African-American couples. J Prim Prev 2020 Jun;41(3):261-278. [doi: 10.1007/s10935-020-00589-x] [Medline: 32410065]
78. Brockman TA, Balls-Berry JE, West IW, et al. Researchers' experiences working with community advisory boards: how community member feedback impacted the research. J Clin Trans Sci 2021;5(1):e117. [doi: 10.1017/cts.2021.22]
79. Ogunsanya ME, Kaninjing E, Morton DJ, Dwyer K, Young ME, Odedina FT. Bridging the gap: a community advisory board promoting community engagement in cancer research for ethnically diverse populations. Am J Mens Health 2024;18(5):15579883241280826. [doi: 10.1177/15579883241280826] [Medline: 39340388]
80. Quinn ED, Cotter K, Kurin K, Brown K. Conducting a Community Engagement Studio to adapt enhanced milieu teaching. Am J Speech Lang Pathol 2022 May 10;31(3):1095-1113. [doi: 10.1044/2021_AJSLP-21-00100] [Medline: 35007426]

XSL•FO
RenderX

81. Killough CM, Martinez J, Mata H, et al. New horizons in community engagement: virtual community engagement studios amplifying community voices about health research in New Mexico. J Clin Trans Sci 2024;8(1):e140. [doi: 10.1017/cts.2024.608]
82. Nielson C, Huang Y, Kull CA, Park AH. Utilizing Community Engagement Studios to inform patient experience in a multicenter randomized control trial. Int J Pediatr Otorhinolaryngol 2020 Jun;133:110007. [doi: 10.1016/j.ijporl.2020.110007] [Medline: 32208178]
83. Johnson DA, Joosten YA, Wilkins CH, Shibao CA. Case study: community engagement and clinical trial success: outreach to African American women. Clinical Translational Sci 2015 Aug;8(4):388-390 [FREE Full text] [doi: 10.1111/cts.12264]
84. Skiba MB, Badger TA, Garcia DO, Chilton FH, Winters-Stone KM. Adapting a dyadic exercise program to be culturally relevant for Hispanic men with prostate cancer using community engagement studio: a brief report. Front Psychol 2024;15:1294546. [doi: 10.3389/fpsyg.2024.1294546] [Medline: 38716273]
85. Scheffey K, Avelis J, Patel M, Oon AL, Evans C, Glanz K. Use of Community Engagement Studios to adapt a hybrid effectiveness-implementation study of social incentives and physical activity for the STEP Together study. Health Promot Pract 2024 Mar;25(2):285-292. [doi: 10.1177/15248399221113863] [Medline: 35899691]
86. Stock MR, Ceïde ME, Lounsbury DW, Zwerling J. Utilizing community engagement studios to inform clinical trial design at a Center of Excellence for Alzheimer's Disease. J Clin Trans Sci 2022;6(1):e73. [doi: 10.1017/cts.2022.388]
87. Schmidt M, Huang R. Defining learning experience design: voices from the field of learning design & technology. TechTrends 2022 Mar;66(2):141-158. [doi: 10.1007/s11528-021-00656-y]
88. Jahnke I, Schmidt M, Earnshaw Y, Tawfik A. Theoretical considerations of learning experience design. In: Theories to Influence the Future of Learning Design and Technology: EdTech Books; 2022. [doi: 10.59668/534]
89. Aloizou V, Ioannou A, Boloudakis M, Retalis S. A learning experience design framework for multimodal learning in the early childhood. Smart Learn Environ 2025;12(1). [doi: 10.1186/s40561-025-00376-3]
90. Hernández R, Kilar-Magdziarz E. Learning experience design (LXD) of language and content modules: insights from students and instructors. In: Kumar P, Eisenberg J, editors. Synchronous and Asynchronous Approaches to Teaching: Palgrave Macmillan; 2023:269-290. [doi: 10.1007/978-3-031-17841-2_13]
91. Conceição SCO, Howles L. Designing the Online Learning Experience: Evidence-Based Principles and Strategies: Taylor & Francis; 2023.
92. Kensing F, Blomberg J. Participatory design: issues and concerns. Computer Supported Cooperative Work (CSCW) 1998 Sep;7(3-4):167-185. [doi: 10.1023/A:1008689307411]
93. Bowen K, Forssell KS, Rosier S. Theories of change in learning experience (LX) design. In: Tawfik A, Earnshaw Y, Jahnke I, editors. Learner and User Experience Research: An Introduction for the Field of Learning Design and Technology 2020.
94. Georgsson M, Staggers N. An evaluation of patients' experienced usability of a diabetes mHealth system using a multi-method approach. J Biomed Inform 2016 Feb;59:115-129. [doi: 10.1016/j.jbi.2015.11.008] [Medline: 26639894]
95. McDonald J, Westerberg T. Learning experience design as an orienting guide for practice: insights from designing for expertise. jaid 2023. [doi: 10.59668/515.12898]
96. McInnes K, Friesen CL, MacKenzie DE, Westwood DA, Boe SG. Mild traumatic brain injury (mTBI) and chronic cognitive impairment: a scoping review. PLoS ONE 2017;12(4):e0174847. [doi: 10.1371/journal.pone.0174847] [Medline: 28399158]
97. Mavroudis I, Ciobica A, Bejenariu AC, et al. Cognitive impairment following mild traumatic brain injury (mTBI): a review. Med Bogota Colomb 2024;60(3):380. [doi: 10.3390/medicina60030380]
98. Jahnke I, Lee YM, Pham M, He H, Austin L. Unpacking the inherent design principles of mobile microlearning. Tech Know Learn 2020 Sep;25(3):585-619. [doi: 10.1007/s10758-019-09413-w]
99. Lopez S. The impact of cognitive load theory on the effectiveness of microlearning modules. European Journal of Education and Pedagogy 2024;5(2):29-35 [FREE Full text] [doi: 10.24018/ejedu.2024.5.2.799]
100. Mayer RE. Cognitive theory of multimedia learning. In: Mayer R, editor. The Cambridge Handbook of Multimedia Learning: Cambridge University Press; 2005:31-48. [doi: 10.1017/CBO9780511816819.004]
101. Rahmani-Katigari M, Mohammadian F, Shahmoradi L. Development of a serious game-based cognitive rehabilitation system for patients with brain injury. BMC Psychiatry 2023 Nov 29;23(1):893. [doi: 10.1186/s12888-023-05396-2] [Medline: 38031072]
102. Jahnke I, Riedel N, Singh K, Moore J. Advancing sociotechnical-pedagogical heuristics for the usability evaluation of online courses for adult learners. OLJ 2021;25(4):337-360. [doi: 10.24059/olj.v25i4.2439]
103. Schmidt M, Earnshaw Y, Jahnke I, Tawfik AA. Entangled eclecticism: a sociotechnical-pedagogical systems theory approach to learning experience design. Education Tech Research Dev 2024 Jun;72(3):1483-1505. [doi: 10.1007/s11423-024-10353-1]
104. Staniszewska S, Brett J, Simera I, et al. GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. Res Involv Engagem 2017;3(1):13. [doi: 10.1186/s40900-017-0062-2] [Medline: 29062538]
105. Virzi RA. Refining the test phase of usability evaluation: how many subjects is enough? Hum Factors 1992 Aug;34(4):457-468. [doi: 10.1177/001872089203400407]
106. Nielsen J, Landauer TK. A mathematical model of the finding of usability problems. 1993 Presented at: the SIGCHI conference; May 1, 1993; Amsterdam, The Netherlands p. 206-213 URL: http://portal.acm.org/citation.cfm?doid=169059 [doi: 10.1145/169059.169166]

XSL•FO
RenderX

107. Faulkner L. Beyond the five-user assumption: benefits of increased sample sizes in usability testing. Behav Res Methods Instrum Comput 2003 Aug;35(3):379-383. [doi: 10.3758/bf03195514] [Medline: 14587545]
108. Lewis JR. Sample sizes for usability studies: additional considerations. Hum Factors 1994 Jun;36(2):368-378. [doi: 10.1177/001872089403600215] [Medline: 8070799]
109. Fowler LA, Vázquez MM, DePietro B, Wilfley DE, Fitzsimmons-Craft EE. Development, usability, and preliminary efficacy of a virtual reality experience to promote healthy lifestyle behaviors in children: pilot randomized controlled trial. Mhealth 2024;10:29. [doi: 10.21037/mhealth-24-24] [Medline: 39534453]
110. Papadopoulos-Nydam G, Rieger JM, Constantinescu G. Usability testing of a mHealth system for swallowing therapy in patients following stroke. Perspect ASHA SIGs 2021 Oct 20;6(5):1205-1211. [doi: 10.1044/2021_PERSP-21-00075]
111. Constantinescu G, Kuffel K, King B, Hodgetts W, Rieger J. Usability testing of an mHealth device for swallowing therapy in head and neck cancer survivors. Health Informatics J 2019 Dec;25(4):1373-1382. [doi: 10.1177/1460458218766574] [Medline: 29618274]
112. Patton MQ. Qualitative Research & Evaluation Methods: Integrating Theory and Practice: SAGE Publications, Inc; 2015.
113. Juengst S, Supnet C, Kew CLN, et al. Bilingual problem-solving training for caregivers of adults with dementia: a randomized, factorial-design protocol for the CaDeS trial. Contemp Clin Trials 2021 Sep;108:106506. [doi: 10.1016/j.cct.2021.106506] [Medline: 34273551]
114. Kouprie M, Visser FS. A framework for empathy in design: stepping into and out of the user's life. Journal of Engineering Design 2009 Oct;20(5):437-448. [doi: 10.1080/09544820902875033]
115. Siricharoen WV. Using empathy mapping in design thinking process for personas discovering. In: Vinh PC, Rakib A, editors. : Springer International Publishing; 2021 Presented at: Context-Aware Systems and Applications, and Nature of Computation and Communication. ICCASA ICTCC 2020 p. 182-191. [doi: 10.1007/978-3-030-67101-3_15]
116. Schmidt M, Tawfik A. Activity theory as a lens for developing and applying personas and scenarios in learning experience design. JAID 2022. [doi: 10.59668/354.5904]
117. Tripp SD, Bichelmeyer B. Rapid prototyping: an alternative instructional design strategy. ETR&D 1990 Mar;38(1):31-44. [doi: 10.1007/BF02298246]
118. Lu J, Schmidt M, Shin J. Beyond technological usability: exploratory factor analysis of the comprehensive assessment of usability scale for learning technologies (CAUSLT). arXiv. Preprint posted online on Feb 3, 2025. [doi: 10.48550/arXiv.2501.18754]
119. Nielsen J. Severity Ratings for Usability Problems.: Nielsen Norman Group URL: https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/ [accessed 2025-09-10]
120. Kali Y. The design principles database as a means for promoting design-based research. In: Handbook of Design Research Methods in Education: Routledge; 2008:423-438. [doi: 10.4324/9781315759593]
121. ISO 9241-11:2018: ergonomics of human-system interaction part 11: usability: definitions and concepts. International Organization for Standardization. 2018. URL: https://www.iso.org/standard/63500.html [accessed 2026-01-10]
122. Zhou L, Bao J, Setiawan IMA, Saptono A, Parmanto B. The mHealth App Usability Questionnaire (MAUQ): development and validation study. JMIR Mhealth Uhealth 2019 Apr 11;7(4):e11500. [doi: 10.2196/11500] [Medline: 30973342]
123. IEC 62366-1:2015 — Medical devices — Part 1: Application of usability engineering to medical devices. International Electrotechnical Commission. 2015. URL: https://webstore.iec.ch/en/publication/21863 [accessed 2026-01-03]
124. IEC 62366-1:2015/AMD1:2020 — Amendment 1 - medical devices - part 1: application of usability engineering to medical devices. International Electrotechnical Commission. 2020. URL: https://webstore.iec.ch/en/publication/59980 [accessed 2026-01-03]
125. Deniz-Garcia A, Fabelo H, Rodriguez-Almeida AJ, et al. Quality, usability, and effectiveness of mHealth apps and the role of artificial intelligence: current scenario and challenges. J Med Internet Res 2023 May 4;25:e44030. [doi: 10.2196/44030] [Medline: 37140973]
126. Hach S, Alder G, Stavric V, Taylor D, Signal N. Usability assessment methods for mobile apps for physical rehabilitation: umbrella review. JMIR Mhealth Uhealth 2024 Oct 4;12(1):e49449. [doi: 10.2196/49449] [Medline: 39365988]
127. Cicerone KD, Goldin Y, Ganci K, et al. Evidence-based cognitive rehabilitation: systematic review of the literature from 2009 through 2014. Arch Phys Med Rehabil 2019 Aug;100(8):1515-1533. [doi: 10.1016/j.apmr.2019.02.011] [Medline: 30926291]
128. Quintero C. A review: accessible technology through participatory design. Disabil Rehabil Assist Technol 2022 May;17(4):369-375. [doi: 10.1080/17483107.2020.1785564] [Medline: 32620068]
129. Fischer B, Peine A, Östlund B. The importance of user involvement: a systematic review of involving older users in technology design. Gerontologist 2020 Sep 15;60(7):e513-e523. [doi: 10.1093/geront/gnz163] [Medline: 31773145]
130. Sanders EN, Stappers PJ. Co-creation and the new landscapes of design. International Journal of CoCreation in Design and the Arts 2008;4:5-18. [doi: 10.1080/15710880701875068]
131. de Jong T. Cognitive load theory, educational research, and instructional design: some food for thought. Instr Sci 2010 Mar;38(2):105-134. [doi: 10.1007/s11251-009-9110-0]
132. Alahmadi T, Drew S. Subjective evaluation of website accessibility and usability: a survey for people with sensory disabilities. W4A '17: Proceedings of the 14th International Web for All Conference 2017:1-4. [doi: 10.1145/3058555.3058579]

XSL•FO
RenderX

133. Bonn MM, Graham LJ, Marrocco S, Jeske S, Moran B, Wolfe DL. Usability evaluation of a self-management mobile application for individuals with a mild traumatic brain injury. Digit Health 2023;9(20552076231183555):20552076231183555. [doi: 10.1177/20552076231183555] [Medline: 37426589]
134. Portz J, Moore S, Bull S. Evolutionary trends in the adoption, adaptation, and abandonment of mobile health technologies: viewpoint based on 25 years of research. J Med Internet Res 2024 Sep 27;26(1):e62790. [doi: 10.2196/62790] [Medline: 39331463]
135. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
136. Kujala S. User involvement: a review of the benefits and challenges. Behav Inf Technol 2003 Jan;22(1):1-16. [doi: 10.1080/01449290301782]
137. Patel S, Akhtar A, Malins S, et al. The acceptability and usability of digital health interventions for adults with depression, anxiety, and somatoform disorders: qualitative systematic review and meta-synthesis. J Med Internet Res 2020 Jul 6;22(7):e16228. [doi: 10.2196/16228] [Medline: 32628116]
138. Phillips B, Zhao H. Predictors of assistive technology abandonment. Assist Technol 1993;5(1):36-45. [doi: 10.1080/10400435.1993.10132205] [Medline: 10171664]
139. Kidman PG, Curtis RG, Watson A, Maher CA. When and why adults abandon lifestyle behavior and mental health mobile apps: scoping review. J Med Internet Res 2024 Dec 18;26(1):e56897. [doi: 10.2196/56897] [Medline: 39693620]
140. Forbes A, Keleher MR, Venditto M, DiBiasi F. Assessing patient adherence to and engagement with digital interventions for depression in clinical trials: systematic literature review. J Med Internet Res 2023 Aug 11;25:e43727. [doi: 10.2196/43727] [Medline: 37566447]
141. Johnston M, Mobasheri M, King D, Darzi A. The Imperial Clarify, Design and Evaluate (CDE) approach to mHealth app development. BMJ Innov 2015 Apr;1(2):39-42. [doi: 10.1136/bmjinnov-2014-000020]
142. Ikwunne T, Hederman L, Wall PJ. Design processes for user engagement with mobile health: a systematic review. IJACSA 2022;13(2). [doi: 10.14569/IJACSA.2022.0130235]
143. Martin S, Armstrong E, Thomson E, et al. A qualitative study adopting a user-centered approach to design and validate a brain computer interface for cognitive rehabilitation for people with brain injury. Assist Technol 2018;30(5):233-241. [doi: 10.1080/10400435.2017.1317675] [Medline: 28708963]
144. Pinard S, Bottari C, Laliberté C, et al. Development of an assistive technology for cognition to support meal preparation in severe traumatic brain injury: user-centered design study. JMIR Hum Factors 2022 Aug 4;9(3):e34821. [doi: 10.2196/34821] [Medline: 35925663]
145. Schmidt M, Cheng L, Raj S, Wade S. Formative design and evaluation of a responsive eHealth/mHealth intervention for positive family adaptation following pediatric traumatic brain injury. J Form Des Learn 2020 Dec;4(2):88-106. [doi: 10.1007/s41686-020-00049-z]
146. Schmidt M, Babcock L, Kurowski BG, Cassedy A, Sidol C, Wade SL. Usage patterns of an mHealth symptom monitoring app among adolescents with acute mild traumatic brain injuries. J Head Trauma Rehabil 2022;37(3):134-143. [doi: 10.1097/HTR.0000000000000768] [Medline: 35125434]
147. Kraaijkamp JJM, van Dam van Isselt EF, Persoon A, Versluis A, Chavannes NH, Achterberg WP. eHealth in geriatric rehabilitation: systematic review of effectiveness, feasibility, and usability. J Med Internet Res 2021 Aug 19;23(8):e24015. [doi: 10.2196/24015] [Medline: 34420918]

Abbreviations

CaDeS: Caregivers in Dementia PST and DSJ
CBPR: community-based participatory research
ePST: Electronic Problem-Solving Training
mHealth: mobile health
NASSS: nonadoption, abandonment, scale-up, spread, and sustainability
PST: Problem-Solving Training
TBI: traumatic brain injury

XSL•FO
RenderX

Edited by S Brini; submitted 11.Sep.2025; peer-reviewed by OC Orwa, GLP Bodagala, O Akhadelor; accepted 11.Nov.2025; published 20.Jan.2026.

Please cite as:
Schmidt M, Weng Y, Juengst S, Holland A
Designing Electronic Problem-Solving Training for Individuals With Traumatic Brain Injury: Mixed Methods, Community-Based, Participatory Research Case Study
J Med Internet Res 2026;28:e83995
URL: https://www.jmir.org/2026/1/e83995
doi:10.2196/83995

© Matthew Schmidt, Yueqi Weng, Shannon Juengst, Alexandra Holland. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Toward Patient-Centric Digital Monitoring of Obstructive Sleep Apnea: Mixed Methods Study

James Kenneth Timmis[1,2*], PhD; Kerstin Alexandra Schorr[3*], MSc; Rana Yüksel[1], MSc; Tim van den Broek[3], MSc; Sebastiaan Overeem[4,5], MD, PhD; Dagmar Josine Smid[6], MSc; Willem Johan van den Brink[3], PhD; Nina Leonie Haring[3], PhD

[1]Department of Political Science, University of Freiburg, Freiburg, Baden-Wurttemberg, Germany

[2]Athena Institute for Research on Innovation and Communication in Health and Life Sciences, Vrije Universiteit Amsterdam, Amsterdam, North Holland, The Netherlands

[3]Research Group Microbiology and Systems Biology, Netherlands Organization for Applied Scientific Research (TNO), Leiden, The Netherlands

[4]Sleep Medicine Center Kempenhaeghe, Heeze, The Netherlands

[5]Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

[6]Research Group Human Performance, Netherlands Organisation for Applied Scientific Research (TNO), Soesterberg, The Netherlands

[*]these authors contributed equally

Corresponding Author:
Nina Leonie Haring, PhD
Research Group Microbiology and Systems Biology
Netherlands Organization for Applied Scientific Research (TNO)
Sylviusweg 71
Leiden, 2333 BE
The Netherlands
Phone: 31 627995667
Email: nina.haring@tno.nl

Abstract

Background: Obstructive sleep apnea (OSA) is a sleep disorder characterized by repeated breathing disruptions during sleep. Remote patient monitoring (RPM) of OSA is important, yet contemporary methods are limited. Sensor-based digital health technologies (sDHTs) promise a step advance in OSA RPM, but must provide meaningful, actionable, and usable outputs for patients. While the centrality of considering patient views in sDHT development is widely acknowledged, patient perspectives and priorities are rarely assessed.

Objective: This study aimed to identify patient-prioritized health aspects and preferences for digital measures and RPM to enhance OSA care quality and patient experience, guided by the digital measures that matter framework.

Methods: We used a mixed methods design combining quantitative and qualitative approaches. Individuals with a formal OSA diagnosis and persistent sleep problems (n=223) completed a survey in which they ranked items related to treatment burdens and health priorities, and responded to open-ended questions about restoring previous quality-of-life elements and desired health goals. To gain deeper qualitative insights, we conducted semistructured interviews with patients with OSA, patient advocates, and health care professionals (n=11), focusing on follow-up care, attitudes toward sDHTs and RPM, and preferences for future OSA-related sDHTs and metrics. Quantitative data were analyzed using bootstrap-aggregated Borda counts (broad support) and Plackett-Luce modeling (intense prioritization), while qualitative data from surveys and interviews were analyzed thematically.

Results: Key meaningful aspects of health included the improvement of subjective sleep quality (top-ranked burden; health goal for 46.5%, 93/200 of participants), an increase in daytime energy (quality-of-life aspect to restore for 35.6%, 72/202 and health goal for 25.5%, 51/200 of participants), and physical activity (quality-of-life aspect to restore for 24.7%, 50/202 and health goal for 16.5%, 33/200 of participants). Sleep characteristics and daytime energy were priority targets for digital measure development. Smartwatches, sleep mats, and smart rings were preferred modalities for integration into RPM. Participants' priorities for enhancing monitoring included (1) expanding metrics beyond the Apnea-Hypopnea Index (AHI; 36.6%, 52/142), (2) improving measurement accuracy (20.4%, 29/142), and (3) ensuring outputs are meaningful, understandable (18.3%, 26/142), and actionable (9.2%, 13/142). Patients also reported difficulty interpreting RPM data to determine if and when follow-up care is needed and what type of care is appropriate.

XSL•FO
RenderX

Conclusions: RPM solutions for OSA should expand beyond AHI, ensure accuracy and interpretability, and provide actionable insights to support comprehensive patient-centric management.

(J Med Internet Res 2026;28:e82460) doi:10.2196/82460

KEYWORDS

obstructive sleep apnea; digital biomarkers; remote patient monitoring; patient-centric care; wearable sensors; digital outcome measures; patient-reported outcomes

Introduction

Obstructive sleep apnea (OSA) is a sleep-related breathing disorder characterized by repeated episodes of partial or complete blockage of the upper airway during sleep, leading to oxygen desaturation events and reduced sleep quality [1]. Common symptoms include loud snoring, waking up gasping for air or choking, and excessive daytime sleepiness. OSA is a chronic condition, and approximately 14% of the global population is affected [2]. Due to aging populations and increasing rates of obesity, the prevalence of OSA is expected to rise in the upcoming years. The chronic nature of OSA, combined with its links to other health conditions and multimorbidity, imposes a high burden on patients and significantly strains health system resources [3,4].

The Apnea-Hypopnea Index (AHI) is the standard for classifying OSA severity by counting the number of apneas (complete pauses in breathing) and hypopneas (partial reductions in breathing) per hour of sleep. While AHI is widely used, it has limitations, such as not accounting for the duration and depth of breathing interruptions, or their impact on oxygen levels and sleep stages [5]. Consequently, AHI scores show poor correlation with disease burden and treatment outcomes [6,7]. Moreover, AHI lacks meaning from the patient perspective, not only because it fails to capture symptoms or daily functional impairments, but also because many patients struggle to interpret clinical measures and understand what pertinent changes in their scores actually imply for their health [8].

The most common treatment options for OSA (in the Netherlands) include continuous positive airway pressure (CPAP) and mandibular advancement devices (MADs) [9]. CPAP is routinely supported by remote patient monitoring (RPM) solutions. RPM refers to systems that allow health care professionals to assess, monitor, and care for patients virtually, often in extraclinical settings [10]. While CPAP RPM platforms provide data on use, mask leaks, and AHI scores [11], they fail to capture patient-centric outcomes, and often do not prevent nonadherence [12]. In contrast, most MADs lack embedded sensors, so adherence and treatment effectiveness are typically assessed subjectively during follow-up visits [13].

Patient-reported outcome measures (PROMs) offer invaluable insights into the subjective experiences of individuals with OSA, especially regarding disease-related quality of life [14]. However, PROMs are limited by respondent burden and various biases, such as nonresponse, fatigue, and recall bias [15,16]. Sensor-based digital health technologies (sDHTs) can be defined as (often wearable) devices that use sensors (for instance, accelerometers or photoplethysmography) to capture health measures, such as symptoms and functional states, continuously [17,18]. They have the potential to (partially) replace, or complement, existing PROMs and thereby provide more objective, real-time insights into patients' health [19,20]. In the context of OSA, sDHTs may be integrated directly into treatment devices or used independently (eg, watches and sleep mats). By passively and continuously capturing objective, longitudinal data based on digital biomarkers, sDHTs also hold potential to facilitate the prediction of treatment responses, optimization of titration, and enhancement of adherence. This supports the transition toward person-centered OSA care, which empowers patients to manage their own condition [21]. Additionally, patient-centric digital end points are becoming increasingly important for clinical research [22].

However, studies across a wide range of different settings have shown that, for (new generations of) technologies to be properly adopted, patients must consider them meaningful, actionable, and usable [18,22-25]. Although the importance of considering patient priorities in sDHT development is widely acknowledged, the former are rarely assessed and integrated in new sDHTs [26]. Efforts led by the Digital Medicine Society (DiMe) are paving the way for the development of sDHTs that are truly patient-centric [18,23]. According to DiMe's digital measures that matter framework, this process begins with the identification of meaningful aspects of health—aspects of a condition that patients wish to improve, arrest, or prevent—which then guide the selection of measurable concepts that reflect patients' lived experiences and priorities [18]. Noteworthy is that the perspectives of patients with OSA are rarely reported in scientific literature, and we are not aware of literature published on the priorities of patients with OSA for sDHTs. To contribute toward addressing this knowledge gap, this study aimed to identify meaningful aspects of health, patient and clinician priorities, and preferences regarding sDHTs for OSA management using a mixed methods design.

Methods

Study Design

This observational, exploratory study used a sequential mixed methods design, integrating survey-based quantitative data with qualitative data collected in semistructured interviews to gather insights from individuals with OSA reporting persistent sleep problems and health care professionals involved in OSA care. Quantitative data were collected through a survey completed by n=223 respondents (which is a sample size considered sufficient in published studies with similar study aims, designs, and settings) [27-29], between January and March 2024, while qualitative insights were obtained from semistructured interviews conducted in May-June 2024 with 6 patients, 2

XSL•FO
RenderX

patient advocates, and 3 health care professionals (Table S1 in Multimedia Appendix 1).

Survey

The survey targeted a broad population of individuals with self-reported sleep problems; participants were included who were aged 18 years and older and experiencing sleep problems at least 3 times per week for a minimum of 3 consecutive months. Moreover, we included in this analysis only those respondents who reported receiving a formal OSA diagnosis from a health care professional. Survey participants were recruited via social media posts, flyers distributed in primary and secondary sleep care settings (including physiotherapy practices and sleep clinics), and a newsletter announcement by the Dutch Apnea Association (ApneuVereniging). The survey (originally developed in Dutch and translated into English for this manuscript) consisted of 18 questions (multiple choice, ranking, and open-ended) distributed through the online platform Survalyzer (Multimedia Appendix 1). Respondents were only able to participate after providing informed consent. The survey included 3 themes. The first theme covered demographic and background information, including medical history and sleep disorder profile (questions 1-8). Questions 1 to 4 were adapted from the "Netherlands working conditions" survey conducted by the Netherlands Organization for Applied Scientific Research (TNO) and Statistics Netherlands [30]. Questions 5 to 8 were developed based on relevant literature and expert input from sleep health care professionals. The second theme focused on meaningful aspects of health (questions 9-12), with questions developed using the digital measures that matter framework by Manta et al [18], which provides patient-centered question formulations to identify aspects of health most meaningful to individuals. The third theme explored preferences and experiences with sDHTs, adapted to the sleep field from a survey exploring this theme in patients in the cardiovascular risk management care pathway, using expert input and supporting literature. The original survey is currently being prepared for publication. In open-ended questions, participants were asked to provide details on, for example, health goals or the aspects that positively drove their quality of life (before their development of OSA), which they would like to restore. Participants had no word limit. While pretesting is a standard step to ensure clarity and reliability, it was not feasible in this instance due to time constraints. Face validity of the survey was assessed by an internal panel of experts experienced in survey design.

Survey Data Analysis

The data analysis for this study primarily involved descriptive statistics, including the calculation of frequencies and percentages to summarize the data and identify patterns and trends within the dataset. Ranking items were analyzed using both Borda counts and the Plackett-Luce model. Using both methods allowed us to combine the accessibility of Borda counts with the statistical rigor of Plackett-Luce and to assess consistency across approaches. The Borda count is a point-based voting method in which each item receives a score based on its rank position, with scores aggregated across participants to produce a consensus ranking [31]. We included Borda counts because they provide a simple and easily interpretable descriptive summary of rankings that has been widely used in previous work. Within each question, we calculated mean scores and 95% CIs. This method assumes complete rankings from all participants. However, some ranking questions in our survey elicited partial rankings, as participants were asked to rank only their top 3 items from a larger set, potentially introducing bias. To address this, the Plackett-Luce model was used as a complementary approach, as it does not penalize unranked items [32]. The Plackett-Luce model estimates the relative *worth* or preference strength of each item based on observed rankings, including partial ones. Each item is assigned a positive worth parameter, which is interpreted comparatively: an item with a higher worth than others is more likely to be systematically preferred. Observations were treated as independent despite potential within-subject correlation, as sparse data precluded models accounting for this, and results should be interpreted accordingly. All computations were performed in R (version 4.4.0; R Foundation for Statistical Computing) using the *PlackettLuce* (version 0.4.3) and *emmeans* (version *1.11.1*) packages.

Open-ended survey responses were analyzed using thematic analysis as outlined by Braun and Clarke [33]. Coding was performed by one researcher and independently reviewed by a second researcher to ensure consistency. Themes were developed based on the frequency, emphasis, and contextual richness of participant responses. In some cases, subdividing themes into subthemes was necessary to provide a deeper understanding of the data.

Interviews

We used purposive sampling to select adequate participants for the interviews. The inclusion criteria were as follows: Dutch patients who have been formally diagnosed with OSA and have experience with (digital) RPM for OSA (current or discontinued use; the latter, to elicit challenges for general use and adherence); health care professionals with a specialization in OSA and who have experience with (digital) RPM for OSA, to elicit their perspective on clinical workflows and patient interactions; and patient advocates for OSA, to elicit perspectives on the pain points and needs of the broader community of patients with OSA in the Netherlands. Our sample included individuals from (1) patients with OSA from the survey who provided email addresses for follow-up, (2) TNO's network from previous OSA collaborations, and (3) clinicians identified through online searches for OSA expertise. The recruitment email included detailed information on the study, including the background of the research, its objectives, and the informed consent form. If individuals agreed to participate, they were able to choose between an in-person or online interview via Microsoft Teams.

The interviews, which on average took about 60 minutes, were conducted and recorded via Microsoft Teams by RY, who used an interview guide that had been expert checked (by NLH and JKT) and piloted a priori (Multimedia Appendix 1). The interview guide was based on a conceptual framework closely aligned with the technology acceptance model, which had been adapted to systematically elicit the perceived usefulness and

XSL•FO
RenderX

ease of use of, perceived needs for, and willingness to engage with OSA RPM and pertinent communication and reporting mechanisms in inter alia OSA follow-up care. The conceptual framework was also used to create the initial deductive code book.

Qualitative Data Analysis

The interviews were manually transcribed (verbatim) by RY. To support analysis, the transcripts were imported into the computer-assisted qualitative data analysis software (CAQDAS) Atlas.TI (version 8.02; Scientific Software Development). The transcripts were analyzed by RY based on the 6 steps of thematic analysis using deductive and inductive (complementary) coding approaches as outlined by Braun and Clarke [33]. To improve the credibility of the analysis, 10% of the transcripts (in this case, n=2 interviews) were independently coded by DS, see acknowledgments. We determined intercoder reliability (approximately 90%) by manually reviewing the codes created and assigned by both coders for both interview transcripts. To increase dependability, codes were carefully tabulated and aggregated into themes based closely on the initial conceptual framework, analyzed and discussed with NLH and JKT, and finally reported by RY. RY also performed member checks (providing a summary of the preliminary analysis of an interview together with a couple of key quotes to the pertinent interviewer, and asking for comment) with multiple interviewees to improve confirmability. The code book and coding were reviewed by another researcher. Finally, the process of selective coding focused on selecting the most important and representative codes to develop overarching themes that addressed the research subquestions of this study [34]. Atlas.Ti was used for the coding process. While the number of possible interviews was limited by time and resource constraints, we, by tendency, reached data saturation in interview 7. However, one further major concept was revealed in interview 9. Interviews 10 and 11 revealed no additional concepts. In consequence, data saturation can therefore not be considered formally achieved. To enhance rigor, we used member checks, peer debriefing, and strategies to minimize social desirability and interview bias (eg, building rapport). For the qualitative component of this study, we carefully considered the 4 established quality criteria of trustworthiness in qualitative research, namely credibility, transferability, dependability, and confirmability, to enhance the overall quality of the study [35].

Ethical Considerations

This study was reviewed in accordance with institutional and national ethical standards for research involving human participants. The research protocol was submitted to TNO's ethical review board, and ethical approval was obtained from TNO's ethical review board for both the survey (study 2023-103) and the interviews (study 2024-026). Informed consent was obtained from participants before data collection. Participants received no compensation for participation. Data used in this study were anonymized, and no personally identifiable information was retained. All data were stored on secure, access-controlled servers in compliance with data protection regulations. No identifiable images or personal information of participants are included in this manuscript or supplementary materials. Ethical approval was further granted by TNO's ethical review board for making data available in a repository (study 2023-103).

Data Management and Availability

The datasets generated and analyzed during this study are publicly available in the Harvard Dataverse repository under the title "Replication Data for: Toward Patient-Centric Digital Health Solutions for Obstructive Sleep Apnea Monitoring: Perspectives from Dutch Patients and Healthcare Professionals – a mixed-method study" [36]. The repository includes an anonymized survey dataset (n=223) containing quantitative responses on meaningful aspects of health, attitudes toward remote patient monitoring, and digital health technology preferences. To protect participant confidentiality, direct identifiers have been removed, and free-text fields have been redacted to avoid inadvertent disclosure of personal information.

Code Availability

The analysis code used to generate the quantitative results reported in this study is publicly available in the same Harvard Dataverse repository [36].

Reporting Guideline

This study adhered to the Mixed Methods Reporting in Rehabilitation and Health Sciences guideline, and the completed checklist is provided in Multimedia Appendix 2.

Results

The survey focused on the themes of meaningful aspects of health, current use of and attitudes toward sDHTs or RPM, and preferences for future OSA-specific sDHTs and RPM solutions. To gain deeper qualitative insights, we also conducted semistructured interviews with formally diagnosed patients with OSA, patient advocates, and health care professionals. The interviews covered the themes follow-up care, attitudes toward sDHTs and RPM, and preferences for future OSA-related sDHTs and metrics; Table S2 in Multimedia Appendix 1.

Demographic Information

A total of 404 individuals initiated the survey; after applying eligibility criteria, the final analytic sample comprised 223 Dutch patients with a formal OSA diagnosis (Multimedia Appendix 3). A total of 48.4% were female, and the mean age was 65 (SD 9) years (Table 1). The majority of the sample (133/223, 59.6%) was highly educated, and 67.3% (150/223) of the participants worked for less than 1 day per week or not at all. A total of 34.1% (76/223) were formally diagnosed with at least one other sleep-related illness besides OSA. Also, other comorbid conditions were reported by 71% (158/223), with obesity and cardiovascular disease being the most prevalent. All interview participants were Dutch and purposefully selected; they had to be at least 18 years of age and have preexisting experience with OSA RPM solutions (Table S1 in Multimedia Appendix 1). We interviewed 6 Dutch individuals with a formal OSA diagnosis, 2 patient advocates affiliated with the national association for patients with OSA (ApneuVereniging), and 3 health care professionals experienced with managing OSA. This purposive sample was designed to capture a range of

XSL•FO
RenderX

perspectives from key stakeholder groups (patients, advocates, and clinicians; Table S2 in Multimedia Appendix 1).

Table 1. Respondent characteristics of surveyed cohort (n=223).

Respondent characteristics	Values
Age (years), median (range)	67 (28-86)
Sex, n (%)	
Male	48 (108)
Female	52 (115)
Education, n (%)	
Secondary education	6.7 (15)
Secondary vocational education	33.6 (75)
Higher professional education and academic education	59.6 (133)
OSA care provider visited within the last year, n (%)	
General practitioner	38 (84)
Medical specialist (outside sleep clinic)	50 (111)
Company doctor	6 (13)
Psychologist	6 (13)
Sleep therapist	4 (9)
Sleep clinic	27 (59)
Other	3 (6)
None	30 (67)
Sleep-related diagnoses, n (%)	
Obstructive sleep apnea	100 (223)
Insomnia	10 (22)
Hypersomnia	12 (27)
Sleep rhythm disorder	1 (2)
Parasomnia	4 (10)
Sleep-related movement disorder	15 (34)
Other	3 (6)
Other diagnoses, n (%)	
Obesity	33 (73)
Cardiovascular disease	37 (82)
Diabetes type 2	15 (34)
Depression	9 (20)
Other	18 (42)
None	29 (64)

Meaningful Aspects of Health

Overall, we found that meaningful aspects of health for individuals with OSA and persistent sleep problems encompass both physical limitations and psychological burdens. Subjective sleep quality, daytime energy levels, and physical activity consistently emerged as key priorities—highlighted as important, considerable burdens when impaired, and as goals for improvement or resumption. In addition, psychological concerns such as worrying about health impacts and difficulties concentrating reflect the broader mental burden of OSA. We present results from ranking exercises of prespecified health aspects and, separately, thematic analyses of responses to open-ended questions.

Burdens of Living With OSA (Ranking)

Worrying about OSA health impacts, sleep interruptions, and problems concentrating were ranked highest (Figure 1, parts A and B). Based on Borda counts, these items showed broad support across the cohort, while Plackett-Luce modeling revealed that certain concerns—such as falling asleep while driving or problems concentrating – were intensely prioritized

XSL•FO
RenderX

by subsets of respondents. Being too tired for hobbies or sport was also considered an important burden (fourth place for both methods).

General Health Priorities (Ranking)

While physical fitness attracted the strongest support overall, it came in second regarding rank concentration (Figure 1, parts C and D). This means that while physical fitness was considered most important overall, the agreement on specific ranks selected was less consistent than, for example, mental and emotional well-being or chronic disease management. Subjective sleep quality was ranked second overall, and first, based on rank agreement (Figure 1, part D), so the reversed relationship was observed when compared with physical fitness. Subjective sleep quality refers to the experienced quality of sleep as perceived by the individual, including aspects such as sleep fragmentation, feelings of restorative sleep, and, more generally, what respondents described as "good" sleep without providing a fixed definition. Mental and emotional well-being, chronic disease management, and diet came in third, fourth, and fifth, respectively, for both analytical methods, respectively.

Figure 1. Meaningful aspects of health identified through survey questions. Parts A and B were derived from a 3-item ranking question from 9 prespecified items (n=221). Parts B and C were derived from a 2-item ranking question from 5 prespecified items (n=219). OSA: obstructive sleep apnea.

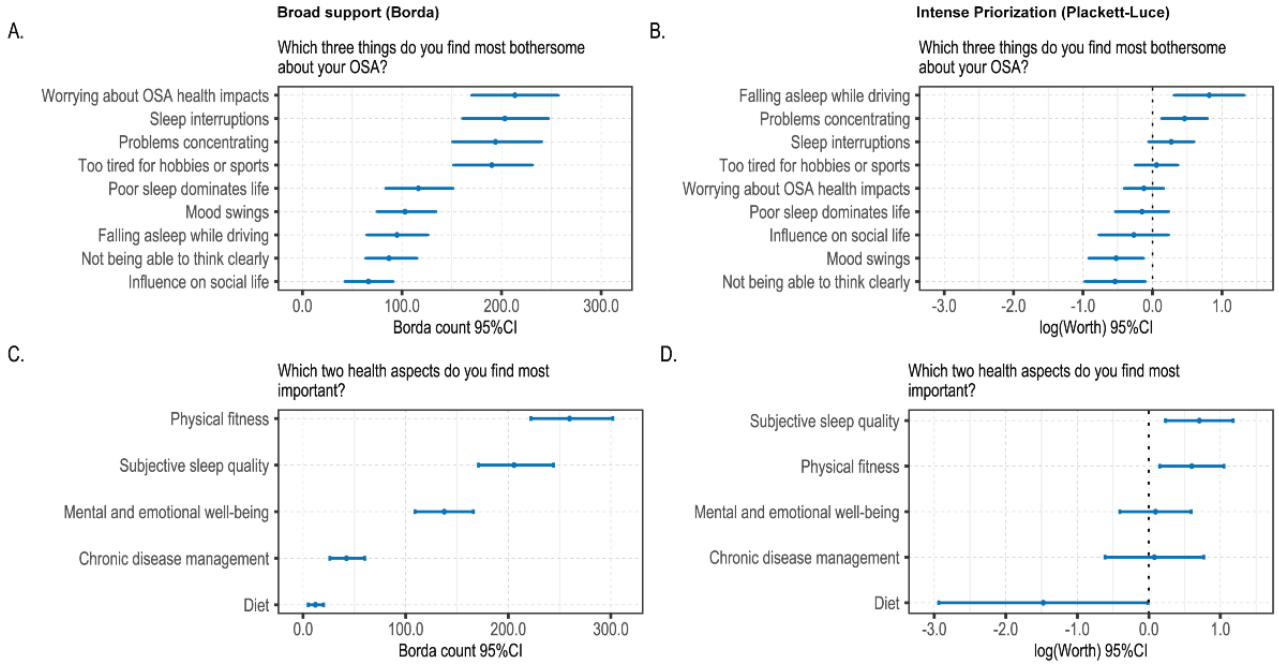

Restoring Previous Quality of Life (Open-Ended)

A broad desire for higher daytime energy levels was expressed by 35.5% (72/202) of respondents (Table 2). For example, one participant reported that they missed "getting through a whole day without constraints from extreme tiredness." Another explained: "I still do all things normally, only it [takes much longer] because you are so terribly tired." Various types of physical activities were described by 24.8% (50/202) of respondents as aspects that they missed from their quality of life before disease onset, such as hiking, cycling, or exercise more generally. Interestingly, "None" was the third most common theme (17.8%, 36/202), reflecting respondents who reported nothing from their previous quality of life they wished to restore. Social activities (16.8%, 34/202), activities requiring concentration, such as reading a book (15.8%, 32/202), and restorative sleep (14.9%, 30/202) were also mentioned.

XSL•FO
RenderX

Table 2. Aspects of their previous quality of life that respondents wished to restore (derived from open-ended questions; n=202).

Items	Respondents, n (%)
Having more energy	72 (35.6)
Physical activity	50 (24.8)
None	36 (17.8)
Social life and leisure	34 (16.8)
Concentration	32 (15.8)
Subjective sleep quality	30 (14.9)
Feeling more productive	22 (10.9)
Daily activities	16 (7.9)
Physical health	16 (7.9)
Being intervention-free	3 (1.5)

Health Goals Relating to OSA Symptoms (Open-Ended)

As shown in Table 3, "Subjective sleep quality" was the top health goal for our sample (46.5%, 93/200). For example, one participant explained that their goal was: "no more lying awake after going to the toilet at night." The themes weight loss and daytime energy levels both came in second (25.2%, 51/200). Participants explained their desire to feel less tired and more energetic throughout the day. The third most frequently mentioned aspect was physical activity (16.3%, 33/200). Additionally, 14.4% (29/200) of participants wish to improve "health aspects related to physical health," including, for example, addressing night sweats, palpitations, or restless leg syndrome.

Table 3. Health goals related to obstructive sleep apnea symptoms derived from open-ended questions (n=200).

Items	Respondents, n (%)
Subjective sleep quality	93 (46.5)
Daytime energy levels	51 (25.2)
Weight loss	51 (25.2)
Physical activity	33 (16.3)
Physical symptoms	29 (14.4)
Improved treatment	22 (10.9)
Concentration	16 (7.9)
Mental health	13 (6.4)
Social life and leisure	8 (4)
Other	5 (2.4)
None	3 (1.5)

OSA Follow-Up Care in the Netherlands

Based on the survey data, half of the participants reported contact with medical specialists, and 38% with general practitioners, respectively. A total of 30% (67/223) indicated that they had no physical follow-up health care visits. 26.4% (59/223) indicated that they had visited sleep centers during the past year (Table 1). Notably, several respondents who selected the option "other" explained that they had not attended physical follow-up visits but were, instead, relying on remote monitoring by their CPAP devices or self-monitoring of their condition. For example, one participant explained:

> *No. I haven't seen a specialist since I got the [CPAP device]. According to the pulmonologist, everything was under control, and there was no need to return.*

According to some interviewees, follow-up care is straightforward but limited, as exemplified here by the following comment:

> *[I was] diagnosed [with] sleep apnea [and] provided...with the CPAP mask. A year later, I had a follow-up appointment...to ensure my mask fit well and that everything was in order. After that, I received no follow-up care anymore.* [Participant B8; patient with OSA who is also a general practitioner]

B3 (patient with OSA) highlighted that they were not even sure what type of care they should be receiving: "I received no follow-up care. I don't even know what follow-up care you should get." Participant B11 (a patient advocate) explained that this was likely due to resource constraints:

XSL•FO
RenderX

> *Clinics don't have the capacity to provide follow-up care for all these patients. There are long waiting lists, and those waiting for OSA diagnosis are prioritized, meaning that those who should receive follow-up care need to wait.* [Participant B11; patient advocate]

By the same token, the two interviewed somnologists highlighted that, on the one hand, patients have a role in taking charge if issues occur, and, on the other, follow-up care can indeed vary greatly based on patients' needs and the severity of their OSA. Participant B7 (a somnologist) explained, "Doctors expect that patients will initiate contact if there's a problem....Effective communication and reporting in follow-up care...also requires patients to allocate time." Participant B9 (pulmonologist/somnologist) provided a top-level view of her patient-centered approach to follow-up care:

> *I do an initial evaluation after 6-8 weeks post-diagnosis, then plan further check-ups based on the patient's condition and needs. If a patient is doing well, I schedule a follow-up after a year..... If there are any issues, I see them sooner, using RPM data from their CPAP device to guide decisions.* [Participant B9]

Nevertheless, participant B5 (a patient with OSA) explained, other hurdles might exist in achieving good continuity of care: "I called my supplier, who said that they had sent a report of my monitored data to the hospital, but the hospital claimed they that hadn't received any report from my supplier."

sDHTs Used and Attitudes Toward RPM

Based on the survey data, 86% (182/212) of participants agreed that sDHTs could contribute toward improving management of their OSA. Table 4 shows that, in our sample, the most frequently used device for self-monitoring was CPAP (87.1%, 155/178), followed by weight scales (47.8%, 85/178), blood pressure monitors (46.6%, 83/178), and smart watches (35.4%, 63/178 for heart rate measurements and 19.7%, 35/178 for sleep tracking). Sleeping mats were used least frequently (n=1). Figure 2, parts A and B, shows that, when asked to rank different form factors according to their preference, smart watches were ranked first, both in terms of overall frequency and subgroup concentration. The least preferred technology was clothing with integrated sensors. Notably, only a single participant reported using a sleep mat, yet when participants were asked to rank their preferred technologies, sleeping mats were ranked among the top three. While the majority of participants (67%, 119/178) agreed that both they themselves, as well as their care providers, should have access to data collected with sDHTs, 28% (49/178) thought their care providers should have access only under certain conditions—merely 4% (7/178) indicated that they should be the only party with access to their data.

Also, the majority of our interviewees held positive views toward the usability of RPM technologies for supporting OSA management. Participant B6 (a patient with OSA), for example, was particularly enthusiastic: "RPM tools [are] a great addition because of the shortage of health care professionals now, so I think it's fantastic that we're focusing on that." By the same token, some patients with OSA had concerns regarding the dependability of the RPM ecosystem and the actionability of information contained therein. For example, a patient with OSA stated:

> *I appreciate RPM, but it's only effective if I receive feedback from the RPM technology supplier or hospital. Ideally, it should alert you promptly if issues like stopped breathing arise, which can ensure constant surveillance and timely intervention.* [Participant B3; patient with OSA]

Moreover, there exists an unmet need regarding the interpretation of data: A patient with OSA explained:

> *I do see my sleep apnea values, for example, an AHI score of 11.3, [but] is that good? What does this score mean for my health, and how can I improve it? I'm missing this information, and I think many do.* [Participant B4; patient with OSA]

Table 4. Currently used technologies to monitor health are derived from multiple-choice questions (n=178).

Items	Respondents, n (%)
CPAP[a]	155 (87.1)
Weight scale	85 (47.8)
Blood pressure monitor	83 (46.6)
Smartwatch (heart rate)	63 (35.4)
Pulse oximeter	46 (25.8)
Smartwatch (sleep)	35 (19.7)
Health app	21 (11.8)
Glucose monitor	16 (9)
Other	5 (2.8)
Sleep mat	1 (0.6)

[a]CPAP: continuous positive airway pressure. Images of devices are shown in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 2. Form factor preferences. Parts A and B were derived from the full ranking of 7 prespecified items (n=109).

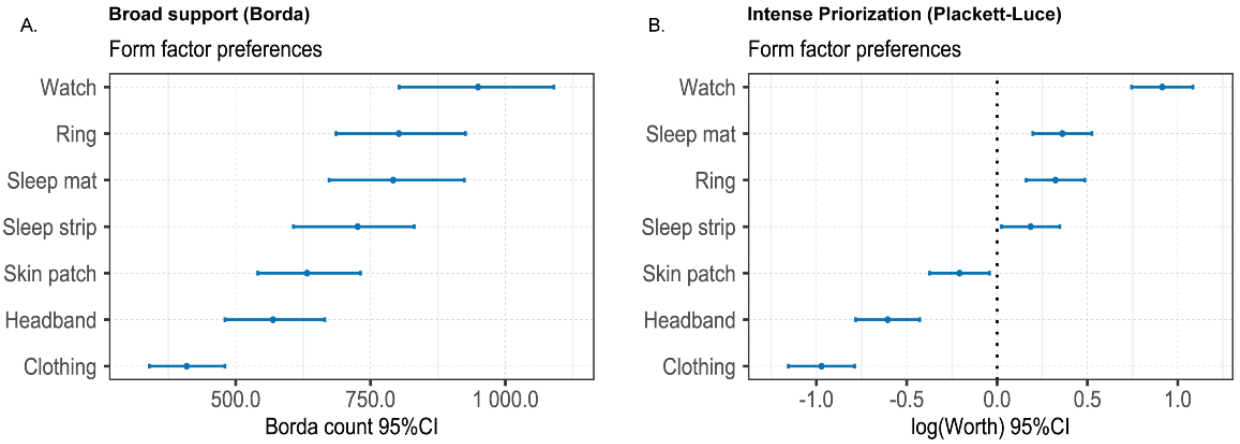

This was somewhat echoed, albeit in more general terms, by the somnologist (Participant B7), who was of the view that current RPM solutions have not yet reached the degree of maturity required for effective OSA remote management:

> *Follow-up care cannot be provided as effectively through digital health technologies like RPM compared to clinical care. As a doctor, you need to see, touch, and smell a patient to provide optimal care. You miss the human aspect of care when everything is digitized and managed through technologies that provide RPM.* [Participant B7]

Nevertheless, our data indicate that some patients are actively engaged with their RPM data. Participant B2 (a patient with OSA) explained:

> *I read my reported data from the screen of my device. Then I have the DreamMapper app [patient-facing app showing CPAP data]. I take out the SD card of my [CPAP] device and plug it into my computer, on which I have installed the Oscar program. This program gives me a comprehensive report on my monitored sleep apnea. Oscar is usually a very good program, but the hospitals don't want to use it because it's not validated.* [Participant B2]

This quote quite vividly illustrates that, while solutions exist that can be meaningful for patients with OSA who want to dive deeper into their data and better understand their disease, these might be located outside of traditional clinical care pathways (and therefore lack clinical oversight and supervision).

Improving Digital Measures for the Future of OSA RPM

Participants were asked to rank which three digital measures would help them gain more insight into their OSA. Sleep characteristics ranked highest overall. However, broad support in the Borda count analysis and rank agreement in the Plackett-Luce model for both sleep characteristics and daytime energy levels (second rank) indicate that these are the top targets for future digital measures (Figure 3, parts A and B). Physical activity was the only other item that came in at the identical rank (8) for both methods of analysis. When asked how current OSA measurement practices could be improved, the most frequently given answer related to suggestions for additional measurements (mentioned by 52/142, 36.6% of respondents; Table 5). Specifically, the majority of survey participants were interested in OSA vitals beyond AHI, such as heart rate–derived and saturation-based measures. Obtaining insights into sleep characteristics, such as sleep stages, was also mentioned by several survey participants. Other suggestions included improved accuracy and reliability of measurements, and a more comprehensible presentation of monitoring data. For example, one survey participant highlighted the need for more clarity in analysis reports, such as presentation in layman's terms, whereas another patient expressed the desire for more monitoring and guidance from OSA professionals and home care providers. Also, the more general value of digital measures for OSA monitoring was underscored. For example, one survey participant indicated that they would like to add monitoring options to their MAD-based treatment:

> *Many people have MADs but have no proof – except through complaints (snoring) or feeling fit – [indicating] whether it helps..... It would be great if...[a wearable or smart watch], or other technology could do that* [track and report]

XSL•FO
RenderX

Figure 3. Health aspects participants would like to understand better. Parts A and B were derived from a 3-item ranking of 9 prespecified items (n=213). OSA: obstructive sleep apnea.

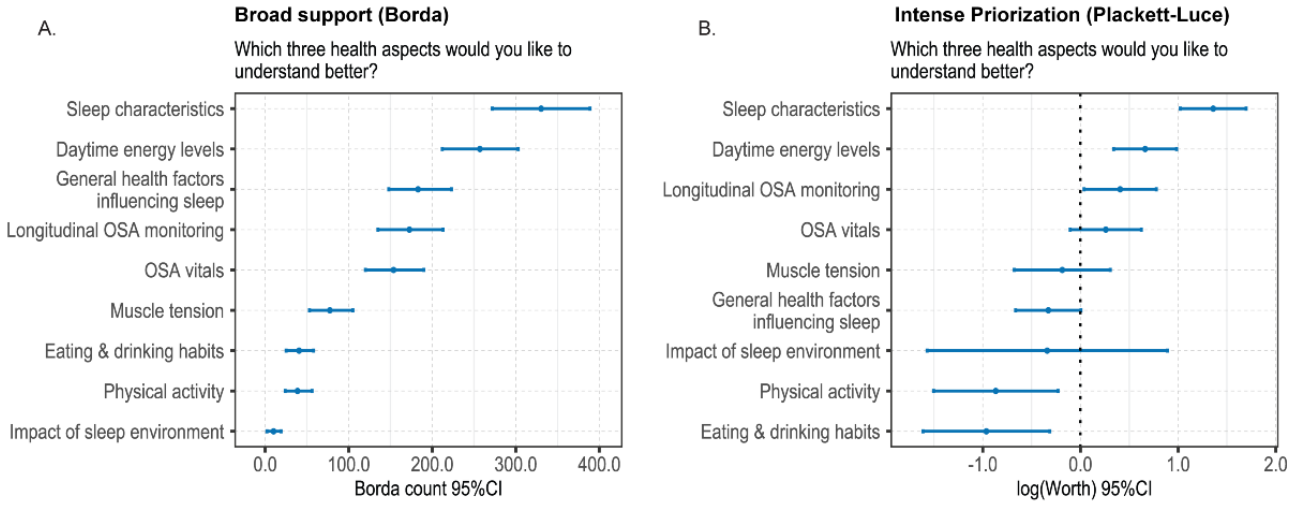

Table 5. Suggestions for improving obstructive sleep apnea remote patient monitoring. Thematically analyzed responses to an open-ended question. Obstructive sleep apnea management advice includes lifestyle advice and contact with a health care provider (n=142).

Items	Respondents, n (%)
Additional measurements	
All	52 (36.6)
OSA[a] severity measures[b]	21 (40.4)
Sleep characteristics	16 (30.8)
Metabolic health	11 (21.2)
Unspecified	4 (7.7)
None	32 (22.5)
Accuracy and reliability	29 (20.4)
More comprehensible presentation	26 (18.3)
OSA management advice	13 (9.2)
Product design	5 (3.5)
Other	6 (4.2)

[a]OSA: obstructive sleep apnea.

[b]OSA severity measures: heart rate–derived measures, breathing, saturation.

The interview data provided additional detail on a range of potential improvements for future OSA management. The majority of patients with OSA preferred noninvasive sDHTs and emphasized the need for a stronger product design focus on less disruptive and more convenient devices for daily use, and better integration with daily attire. A patient with OSA explained:

> *I already use a [smart watch]...for my heart monitoring, but it would be nice if future RPM technologies are non-invasive, like an App or a small device beside the bed. Large or intrusive devices are uncomfortable for light sleepers like me.* [Participant B6; patient with OSA]

Participant B4 (a patient with OSA) also mentioned the, inter alia, visual product design of wearable RPM technology: “A smartwatch-like device that tracks my heart rate and oxygen saturation would be ideal, as long as it’s non-invasive and looks good.” Regarding specific measures, participant B4 (a patient with OSA) said: “I would like to measure my oxygen saturation during the night. This could be an improvement regarding RPM metrics for OSA.”

Discussion

Principal Findings

This mixed methods study, guided by DiMe’s digital measures that matter framework, aimed to, first, identify meaningful aspects of health and care needs of individuals with OSA, and, second, inform the improvement of existing, and the development of new, pertinent digital measures and RPM

XSL•FO
RenderX

solutions. Based on our two Dutch cohorts (patients, their representatives, and specialized health care professionals), our principal findings are that, first, improving subjective sleep quality, increasing physical activity, and increasing daytime energy levels are key meaningful aspects. Second, sleep characteristics (particularly sleep fragmentation), daytime energy (especially fatigue and excessive daytime sleepiness), and nighttime oxygen saturation are priority targets for digital measure development. Third, smartwatches, sleep mats, and smart rings are strongly preferred as modalities for sDHTs that can be integrated into future RPM solutions. Fourth, current digital monitoring practices should be enhanced by focusing on expanding metrics beyond AHI, improving measurement accuracy, and ensuring that digital measures are meaningful, understandable, and actionable for end users. Finally, patients lack the ability to determine from RPM output whether they need to seek follow-up care and, if so, what type of care is appropriate.

Comparison With Previous Work

Our principal findings highlight that improving subjective sleep quality, increasing physical activity, and enhancing daytime energy levels are key meaningful aspects of health for patients with OSA. Each can (potentially) be assessed using sDHT-derived metrics, though these technologies vary widely in maturity and clinical readiness.

Sleep was a consistently prioritized health aspect, essential to daily functioning and overall well-being. Participants expressed skepticism about the accuracy and clinical utility of sleep data from personal devices, echoing World Sleep Society recommendations that, while sDHTs hold potential for monitoring sleep patterns, their proprietary algorithms often lack validation in sleep disorders [37]. Recent studies in OSA populations show promising results for tracking metrics like total sleep time and sleep efficiency, with some devices showing moderate-to-strong agreement with polysomnography; however, broader validation is still needed to ensure accuracy and clinical relevance [38-42].

Physical activity emerged as another key health aspect, with participants identifying it as a health priority and an area they wished to improve. Physical activity is both a meaningful health goal and a key factor in OSA management, as increasing activity can help reduce OSA severity while effective treatment may, in turn, support higher activity levels. Structured exercise interventions can reduce AHI, improve oxygenation, and alleviate daytime sleepiness [43-45]. However, evidence for physical activity levels as an outcome of CPAP therapy is mixed, with studies reporting, despite improved symptoms, modest or no changes in physical activity levels [46-48]. sDHTs could help monitor physical activity as a potential biomarker of functional gains and support behavior change in OSA care.

Daytime energy levels emerged as a top-rated health aspect, with participants identifying it as a key priority, an important health goal, and the most valued quality-of-life factor to restore. This reflects both excessive daytime sleepiness and fatigue, which present heterogeneously across patients and persist in some even after optimal therapy [49-51]. While these symptoms are typically assessed using PROMs, such as the Epworth Sleepiness Scale and Fatigue Severity Scale, their use is limited by respondent burden, recall bias, and low temporal resolution [51]. sDHTs offer a promising, noninvasive alternative for continuous monitoring of these symptoms. Metrics, such as heart rate variability and physical activity, are actively being explored, yet validated digital measures for routine monitoring are lacking—making this a clear priority for innovation [51,52].

Implications for Practice and Policy

For Developers of sDHTs and RPM Solutions

Developing new digital measures and RPM solutions is complex and resource-intensive, often requiring significant time before their impact reaches clinical practice. In the short term, the rapid rise of consumer wearables creates an opportunity to validate and optimize digital measures—such as sleep and physical activity—for specific populations like people with OSA. This should be prioritized as a practical step toward innovation. Beyond creating new measures, improving existing RPM tools is equally urgent. Enhancing reporting platforms with clear explanations, contextualized feedback, and actionable suggestions could better support patients in managing their condition. Interfaces that let users toggle between simplified and detailed views would help accommodate diverse preferences and levels of health literacy.

For Policymakers and Payers

On a broader scale, policymakers and payers should consider setting standards to ensure RPM systems deliver accurate, patient-centric, harmonized, and actionable data presentations. A key requirement is that digital measures embedded in RPM solutions are validated for accuracy, reliability, usability, and clinical relevance in target populations, such as patients with OSA [22,23]. Establishing such validation criteria will help ensure that these tools provide meaningful insights and support clinical decision-making. These standards could also guide reimbursement decisions and accelerate the adoption of RPM solutions that empower patients, improve treatment outcomes, and reduce health care burden.

For Clinical Practice

Our findings underscore several opportunities to strengthen clinical care. First, follow-up care for patients with OSA is often limited and fragmented, highlighting the need for better alignment between home care providers and clinicians and a clearer definition of follow-up pathways. Telemedicine offers a promising, cost-effective avenue to facilitate structured and timely follow-up. Second, patients frequently report difficulties in interpreting RPM data; clinicians and home care providers should therefore provide guidance to ensure patients understand the outputs from their CPAP. Finally, as patients increasingly adopt self-monitoring solutions outside formal care pathways, clinicians should be supported in evaluating the reliability of these tools and integrating relevant patient-generated data into care when appropriate.

Limitations and Strengths

Our study has several limitations. First, the sample was skewed toward more highly educated individuals and included a higher proportion of women, likely reflecting the greater prevalence

XSL•FO
RenderX

of comorbid insomnia in female patients with OSA [53-55]. Second, the survey was not pretested, which might impact its validity. Third, the number of participating health care professionals was limited to 3 (1 somnologist, 1 somnologist/pulmonologist, and 1 general practitioner), which means their perspectives may not fully capture the diversity of clinical views and could reflect individual experiences. However, several of their insights were echoed by patients, supporting their relevance to the themes identified. Furthermore, the surveyed cohort focused on individuals with OSA who report persistent sleep problems despite treatment. While this subgroup may not fully represent the broader OSA population, it highlights a group with substantial unmet needs and provides valuable insights into priorities for digital health innovations and RPM solutions. Another strength is that our patient-centric approach aligns with value-based health care principles and was guided by the digital measures that matter framework to identify outcomes that matter most to individuals with OSA. The combination of quantitative survey data and qualitative interviews, including input from patient representatives, ensured the perspectives captured reflect a broad OSA population.

Conclusion

Rather than relying solely on clinical end points such as AHI, our findings suggest that outcomes such as physical activity, restorative sleep, and daily functioning are central to patients' lived experiences—and are therefore critical targets for sDHTs and RPM metric innovation. Developing and validating new digital measures that capture these experiences will require time, interdisciplinary collaboration, and ongoing involvement of patients to ensure relevance and usability. In the meantime, existing RPM systems can be strengthened by improving transparency, accessibility, and contextual interpretation of currently collected data, making these platforms more meaningful and actionable for patients. The broad patient priorities identified in this study can serve as an excellent starting point for defining patient-centric digital measures, allowing for comprehensive disease management.

Acknowledgments

The authors would like to thank Marian Schoone, Suzan Wopereis, Hardy van der Ven, Koen Hogenelst, Elsbeth de Korte, and André Boorsma for their valuable contributions and insightful discussions, which greatly supported the development of this work. Portions of the text were revised with the assistance of the generative AI language model ChatGPT (GPT-5; OpenAI), and all content was subsequently reviewed and approved by the authors.

Data Availability

The datasets generated and analyzed during this study are publicly available in the Harvard Dataverse repository under the title "Replication Data for: Toward Patient-Centric Digital Health Solutions for Obstructive Sleep Apnea Monitoring: Perspectives from Dutch Patients and Healthcare Professionals – a mixed-method study" [36].

Funding

This research was supported by internal funding from the Netherlands Organization for Applied Scientific Research (TNO). The article processing fee was covered by TNO's internal funds. No external grants or commercial funding were received for this work.

Authors' Contributions

Conceptualization: NLH, WJvdB
Data curation: NLH, KAS, RY, TvdB, DJS
Formal analysis: NLH, KAS, RY, TvdB, DJS
Funding acquisition: WJvdB
Investigation: NLH, RY
Methodology: NLH, WJvdB, TvdB, JKT
Project administration: NLH
Software: KAS, TvdB
Supervision: JKT, NLH
Validation: WJvdB, DJS
Visualization: KAS, TvdB
Writing—original draft: JKT, KAS, NLH
Writing—review and editing: SO, JKT, NLH

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Overview of themes covered in mixed methods components, characteristics of interviewed cohort (n=11), survey (original in Dutch, for this manuscript translated to English), and interview guide (original in Dutch, for this manuscript translated to English).
[DOCX File , 519 KB - jmir_v28i1e82460_app1.docx]

Multimedia Appendix 2
Mixed Methods Reporting in Rehabilitation and Health Sciences checklist.
[PDF File (Adobe PDF File), 697 KB - jmir_v28i1e82460_app2.pdf]

Multimedia Appendix 3
Survey response flow diagram. The final sample was defined as respondents who completed demographic questions (Q1-8) and at least one study question (Q9-18).
[PNG File , 12 KB - jmir_v28i1e82460_app3.png]

References

1. Lv R, Liu X, Zhang Y, Dong N, Wang X, He Y, et al. Pathophysiological mechanisms and therapeutic approaches in obstructive sleep apnea syndrome. Signal Transduct Target Ther 2023;8(1):218 [FREE Full text] [doi: 10.1038/s41392-023-01496-3] [Medline: 37230968]
2. Benjafield AV, Ayas NT, Eastwood PR, Heinzer R, Ip MSM, Morrell MJ, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. Lancet Respir Med 2019;7(8):687-698 [FREE Full text] [doi: 10.1016/S2213-2600(19)30198-5] [Medline: 31300334]
3. Heatley EM, Harris M, Battersby M, McEvoy RD, Chai-Coetzer CL, Antic NA. Obstructive sleep apnoea in adults: a common chronic condition in need of a comprehensive chronic condition management approach. Sleep Med Rev 2013;17(5):349-355. [doi: 10.1016/j.smrv.2012.09.004] [Medline: 23434125]
4. Ehsan Z, Ingram DG. Economic and social costs of sleep apnea. Curr Pulmonol Rep 2016;5(2):111-115. [doi: 10.1007/s13665-016-0142-z]
5. Malhotra A, Ayappa I, Ayas N, Collop N, Kirsch D, Mcardle N, et al. Metrics of sleep apnea severity: beyond the apnea-hypopnea index. Sleep 2021;44(7):zsab030 [FREE Full text] [doi: 10.1093/sleep/zsab030] [Medline: 33693939]
6. Vanderveken OM, Gagnadoux F. Comparative outcomes in obstructive sleep apnea therapy: mean disease alleviation as a more appropriate measure-it's about time. Sleep 2023;46(10):zsad210 [FREE Full text] [doi: 10.1093/sleep/zsad210] [Medline: 37549352]
7. Won CHJ. When will we ditch the AHI? J Clin Sleep Med 2020;16(7):1001-1003 [FREE Full text] [doi: 10.5664/jcsm.8594] [Medline: 32441250]
8. Redline S, Baker-Goodwin S, Bakker JP, Epstein M, Hanes S, Hanson M, Sleep Apnea Patient-Centered Outcomes Network. Patient partnerships transforming sleep medicine research and clinical care: perspectives from the sleep apnea patient-centered outcomes network. J Clin Sleep Med 2016;12(7):1053-1058 [FREE Full text] [doi: 10.5664/jcsm.5948] [Medline: 27166300]
9. van Mechelen PH, Damhuis E, Hazeleger E. Monograph about the care pathway for sleep apnea: a research report from ApneuVereniging (Dutch Apnea Association) in cooperation with Kantar vereniging apneu. ApneuVereniging (Dutch Apnea Association). 2021. URL: https://administratie.apneuvereniging.nl/files/2021-monograph-about-dutch-care-pathway-osa.pdf [accessed 2025-12-03]
10. RPM definition. US Food and Drug Administration. 2023. URL: https://tinyurl.com/3j2yubte [accessed 2025-11-27]
11. Schwab RJ, Badr SM, Epstein LJ, Gay PC, Gozal D, Kohler M, ATS Subcommittee on CPAP Adherence Tracking Systems. An official American Thoracic Society statement: continuous positive airway pressure adherence tracking systems. The optimal monitoring strategies and outcome measures in adults. Am J Respir Crit Care Med 2013;188(5):613-620 [FREE Full text] [doi: 10.1164/rccm.201307-1282ST] [Medline: 23992588]
12. Phillips CL, Grunstein RR, Darendeliler MA, Mihailidou AS, Srinivasan VK, Yee BJ, et al. Health outcomes of continuous positive airway pressure versus oral appliance treatment for obstructive sleep apnea: a randomized controlled trial. Am J Respir Crit Care Med 2013;187(8):879-887. [doi: 10.1164/rccm.201212-2223OC] [Medline: 23413266]
13. Dieltjens M, Braem MJ, Op de Beeck S, Vroegop AVMT, Kazemeini E, Van de Perck E, et al. Remotely controlled mandibular positioning of oral appliance therapy during polysomnography and drug-induced sleep endoscopy compared with conventional subjective titration in patients with obstructive sleep apnea: protocol for a randomized crossover trial. Trials 2019;20(1):615 [FREE Full text] [doi: 10.1186/s13063-019-3698-4] [Medline: 31665059]
14. Abma IL, van der Wees PJ, Veer V, Westert GP, Rovers M. Measurement properties of patient-reported outcome measures (PROMs) in adults with obstructive sleep apnea (OSA): a systematic review. Sleep Med Rev 2016;28:18-31 [FREE Full text] [doi: 10.1016/j.smrv.2015.07.006] [Medline: 26433776]
15. Zini MLL, Banfi G. A narrative literature review of bias in collecting patient reported outcomes measures (PROMs). Int J Environ Res Public Health 2021;18(23):12445 [FREE Full text] [doi: 10.3390/ijerph182312445] [Medline: 34886170]

XSL•FO
RenderX

16. Aiyegbusi OL, Roydhouse J, Rivera SC, Kamudoni P, Schache P, Wilson R, et al. Key considerations to reduce or address respondent burden in patient-reported outcome (PRO) data collection. Nat Commun 2022;13(1):6026 [FREE Full text] [doi: 10.1038/s41467-022-33826-4] [Medline: 36224187]
17. Digital health technologies (DHTs) for drug development. US Food and Drug Administration. URL: https://www.fda.gov/science-research/science-and-research-special-topics/digital-health-technologies-dhts-drug-development [accessed 2025-11-27]
18. Manta C, Patrick-Lake B, Goldsack J. Digital measures that matter to patients: a framework to guide the selection and development of digital measures of health. Digit Biomark 2020;4(3):69-77 [FREE Full text] [doi: 10.1159/000509725] [Medline: 33083687]
19. Kolk MZH, Frodi DM, Langford J, Meskers CJ, Andersen TO, Jacobsen PK, et al. Behavioural digital biomarkers enable real-time monitoring of patient-reported outcomes: a substudy of the multicentre, prospective observational SafeHeart study. Eur Heart J Qual Care Clin Outcomes 2024;10(6):531-542. [doi: 10.1093/ehjqcco/qcad069] [Medline: 38059857]
20. Heros R, Patterson D, Huygen F, Skaribas I, Schultz D, Wilson D, et al. Objective wearable measures and subjective questionnaires for predicting response to neurostimulation in people with chronic pain. Bioelectron Med 2023;9(1):13 [FREE Full text] [doi: 10.1186/s42234-023-00115-4] [Medline: 37340467]
21. Lim DC, Sutherland K, Cistulli PA, Pack AI. P4 medicine approach to obstructive sleep apnoea. Respirology 2017;22(5):849-860 [FREE Full text] [doi: 10.1111/resp.13063] [Medline: 28477347]
22. Aryal S, Blankenship JM, Bachman SL, Hwang S, Zhai Y, Richards JC, et al. Patient-centricity in digital measure development: co-evolution of best practice and regulatory guidance. NPJ Digit Med 2024;7(1):128 [FREE Full text] [doi: 10.1038/s41746-024-01110-y] [Medline: 38755349]
23. Bakker JP, Barge R, Centra J, Cobb B, Cota C, Guo CC, et al. V3+ extends the V3 framework to ensure user-centricity and scalability of sensor-based digital health technologies. NPJ Digit Med 2025;8(1):51 [FREE Full text] [doi: 10.1038/s41746-024-01322-2] [Medline: 39856145]
24. Goldsack JC, Coravos A, Bakker JP, Bent B, Dowling AV, Fitzer-Attas C, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for biometric monitoring technologies (BioMeTs). NPJ Digit Med 2020;3:55 [FREE Full text] [doi: 10.1038/s41746-020-0260-4] [Medline: 32337371]
25. Taylor KI, Staunton H, Lipsmeier F, Nobbs D, Lindemann M. Outcome measures based on digital health technology sensor data: data- and patient-centric approaches. NPJ Digit Med 2020;3:97 [FREE Full text] [doi: 10.1038/s41746-020-0305-8] [Medline: 32715091]
26. Baines R, Bradwell H, Edwards K, Stevens S, Prime S, Tredinnick-Rowe J, et al. Meaningful patient and public involvement in digital health innovation, implementation and evaluation: a systematic review. Health Expect 2022;25(4):1232-1245 [FREE Full text] [doi: 10.1111/hex.13506] [Medline: 35526274]
27. Lazarevic N, Pizzuti C, Rosic G, Bœhm C, Williams K, Caillaud C. A mixed-methods study exploring women's perceptions and recommendations for a pregnancy app with monitoring tools. NPJ Digit Med 2023;6(1):50 [FREE Full text] [doi: 10.1038/s41746-023-00792-0] [Medline: 36964179]
28. Renn BN, Hoeft TJ, Lee HS, Bauer AM, Areán PA. Preference for in-person psychotherapy versus digital psychotherapy options for depression: survey of adults in the U.S. NPJ Digit Med 2019;2:6 [FREE Full text] [doi: 10.1038/s41746-019-0077-1] [Medline: 31304356]
29. Weik L, Fehring L, Mortsiefer A, Meister S. Understanding inherent influencing factors to digital health adoption in general practices through a mixed-methods analysis. NPJ Digit Med 2024;7(1):47 [FREE Full text] [doi: 10.1038/s41746-024-01049-0] [Medline: 38413767]
30. Mars GMJ, van den Heuvel SG, Knops JCM, de Vroome EMM, Gielen WJM, van Dam LMC, et al. National Survey of Working Conditions (NEA) 2023 - Research description [article in Dutch]. Centraal Bureau voor de Statistiek. URL: https://www.cbs.nl/nl-nl/longread/rapportages/2024/nationale-enquete-arbeidsomstandigheden--nea---2023-onderzoeksbeschrijving [accessed 2025-10-28]
31. Saari DG. Geometry of voting. In: Arrow KJ, Sen AK, Suzumura K, editors. Handbook of Social Choice and Welfare. Amsterdam, The Netherlands: Elsevier Science; 2011.
32. Turner H, van Etten J, Firth D, Kosmidis I. Modelling rankings in R: the PlackettLuce package. Comput Stat 2020 Feb 12;35(3):1027-1057. [doi: 10.1007/s00180-020-00959-3]
33. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
34. Alhassan AI. Analyzing the application of mixed method methodology in medical education: a qualitative study. BMC Med Educ 2024;24(1):225 [FREE Full text] [doi: 10.1186/s12909-024-05242-3] [Medline: 38438987]
35. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Methods 2017;16(1). [doi: 10.1177/1609406917733847]
36. van den Broek T, Timmis JK, Schorr KA, Yüksel R, Overeem S, Smid D, et al. Replication data for: toward patient-centric digital health solutions for obstructive sleep apnea monitoring: perspectives from Dutch patients and healthcare professionals: a mixed-method study. Harvard Dataverse 2021.

XSL•FO
RenderX

37. Chee MW, Baumert M, Scott H, Cellini N, Goldstein C, Baron K, World Sleep Society Sleep Tracker Task Force. World Sleep Society recommendations for the use of wearable consumer health trackers that monitor sleep. Sleep Med 2025;131:106506. [doi: 10.1016/j.sleep.2025.106506] [Medline: 40300398]
38. Gruwez A, Bruyneel AV, Bruyneel M. The validity of two commercially-available sleep trackers and actigraphy for assessment of sleep parameters in obstructive sleep apnea patients. PLoS One 2019;14(1):e0210569 [FREE Full text] [doi: 10.1371/journal.pone.0210569] [Medline: 30625225]
39. Kuhn J, Schiphorst LRB, Wulterkens BM, Asin J, Duis N, Overeem S, et al. Multi-night home assessment of total sleep time misperception in obstructive sleep apnea with and without insomnia symptoms. Clocks Sleep 2024;6(4):777-788 [FREE Full text] [doi: 10.3390/clockssleep6040050] [Medline: 39727626]
40. Strumpf Z, Gu W, Tsai C, Chen P, Yeh E, Leung L, et al. Belun Ring (Belun Sleep System BLS-100): deep learning-facilitated wearable enables obstructive sleep apnea detection, apnea severity categorization, and sleep stage classification in patients suspected of obstructive sleep apnea. Sleep Health 2023;9(4):430-440 [FREE Full text] [doi: 10.1016/j.sleh.2023.05.001] [Medline: 37380590]
41. Cai Y, Zheng YJ, Cheng CM, Strohl KP, Mason AE, Chang JL. Impact of hypoglossal nerve stimulation on consumer sleep technology metrics and patient symptoms. Laryngoscope 2024;134(7):3406-3411. [doi: 10.1002/lary.31398] [Medline: 38516821]
42. Byun J, Noh KC, Shin WC. Performance of the Fitbit Charge 2 and Galaxy Watch 2 compared with polysomnography in assessing patients with obstructive sleep apnoea. Chronobiol Int 2023;40(5):596-602. [doi: 10.1080/07420528.2023.2191720] [Medline: 36971253]
43. Peng J, Yuan Y, Zhao Y, Ren H. Effects of exercise on patients with obstructive sleep apnea: a systematic review and meta-analysis. Int J Environ Res Public Health 2022;19(17):10845 [FREE Full text] [doi: 10.3390/ijerph191710845] [Medline: 36078558]
44. Mendelson M, Bailly S, Marillier M, Flore P, Borel JC, Vivodtzev I, et al. Obstructive sleep apnea syndrome, objectively measured physical activity and exercise training interventions: a systematic review and meta-analysis. Front Neurol 2018;9:73 [FREE Full text] [doi: 10.3389/fneur.2018.00073] [Medline: 29520251]
45. Pawar M, Venkatesan P, Mysore S, Bhat G. Effectiveness of aerobic exercise training in patients with obstructive sleep apnea: a systematic review and meta-analysis. Eur Arch Otorhinolaryngol 2025. [doi: 10.1007/s00405-025-09436-3] [Medline: 40329037]
46. Feng Y, Maislin D, Keenan BT, Gislason T, Arnardottir ES, Benediktsdottir B, et al. Physical activity following positive airway pressure treatment in adults with and without obesity and with moderate-severe obstructive sleep apnea. J Clin Sleep Med 2018;14(10):1705-1715 [FREE Full text] [doi: 10.5664/jcsm.7378] [Medline: 30353806]
47. Stevens D, Loffler KA, Buman MP, Dunstan DW, Luo Y, Lorenzi-Filho G, SAVE investigators. CPAP increases physical activity in obstructive sleep apnea with cardiovascular disease. J Clin Sleep Med 2021;17(2):141-148 [FREE Full text] [doi: 10.5664/jcsm.8792] [Medline: 32951632]
48. West SD, Kohler M, Nicoll DJ, Stradling JR. The effect of continuous positive airway pressure treatment on physical activity in patients with obstructive sleep apnoea: a randomised controlled trial. Sleep Med 2009;10(9):1056-1058. [doi: 10.1016/j.sleep.2008.11.007] [Medline: 19427263]
49. Schwarz EI, Schiza S. Sex differences in sleep and sleep-disordered breathing. Curr Opin Pulm Med 2024;30(6):593-599. [doi: 10.1097/MCP.0000000000001116] [Medline: 39189037]
50. Pyun SY, Choi SJ, Jo H, Hwang Y, Cho JW, Joo EY. Gender differences in Korean patients with obstructive sleep apnea. Sleep Med Res 2020;11(2):121-128. [doi: 10.17241/smr.2020.00556]
51. Steier JS, Bogan RK, Cano-Pumarega IM, Fleetham JA, Insalaco G, Lal C, et al. Recommendations for clinical management of excessive daytime sleepiness in obstructive sleep apnoea - A Delphi consensus study. Sleep Med 2023;112:104-115 [FREE Full text] [doi: 10.1016/j.sleep.2023.10.001] [Medline: 37839271]
52. Adão Martins NR, Annaheim S, Spengler CM, Rossi RM. Fatigue monitoring through wearables: a state-of-the-art review. Front Physiol 2021;12:790292 [FREE Full text] [doi: 10.3389/fphys.2021.790292] [Medline: 34975541]
53. Lee J, Ahn S. Polysomnographic findings and psychiatric symptoms in patients with comorbid insomnia and sleep apnea: a retrospective study focusing on sex differences. Sleep Breath 2025;29(1):78. [doi: 10.1007/s11325-025-03248-9] [Medline: 39808352]
54. Subramanian S, Guntupalli B, Murugan T, Bopparaju S, Chanamolu S, Casturi L, et al. Gender and ethnic differences in prevalence of self-reported insomnia among patients with obstructive sleep apnea. Sleep Breath 2011;15(4):711-715. [doi: 10.1007/s11325-010-0426-4] [Medline: 20953842]
55. Saaresranta T, Hedner J, Bonsignore MR, Riha RL, McNicholas WT, Penzel T, ESADA Study Group. Clinical phenotypes and comorbidity in European sleep apnoea patients. PLoS One 2016;11(10):e0163439 [FREE Full text] [doi: 10.1371/journal.pone.0163439] [Medline: 27701416]

Abbreviations

AHI: Apnea-Hypopnea Index

XSL•FO
RenderX

CAQDAS: computer-assisted qualitative data analysis software
CPAP: continuous positive airway pressure
DiMe: Digital Medicine Society
MAD: mandibular advancement device
OSA: obstructive sleep apnea
PROM: patient-reported outcome measure
RPM: remote patient monitoring
sDHT: sensor-based digital health technology
TNO: Netherlands Organization for Applied Scientific Research

Edited by A Stone, A Mavragani; submitted 15.Aug.2025; peer-reviewed by W Ahmed, O Oluwole; comments to author 08.Sep.2025; revised version received 28.Oct.2025; accepted 28.Oct.2025; published 08.Jan.2026.

Please cite as:
Timmis JK, Schorr KA, Yüksel R, van den Broek T, Overeem S, Smid DJ, van den Brink WJ, Haring NL
Toward Patient-Centric Digital Monitoring of Obstructive Sleep Apnea: Mixed Methods Study
J Med Internet Res 2026;28:e82460
URL: https://www.jmir.org/2026/1/e82460
doi:10.2196/82460
PMID:

©James Kenneth Timmis, Kerstin Alexandra Schorr, Rana Yüksel, Tim van den Broek, Sebastiaan Overeem, Dagmar Josine Smid, Willem Johan van den Brink, Nina Leonie Haring. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of Educational Videos in Encouraging Preferences for Guideline-Based Cancer Screening in Japan: Three-Arm Pseudorandomized Controlled Trial

Kosuke Sakai[1], MD, PhD; Yuko Furuya[1], MD, PhD; Shoko Nakazawa[1], MD, PhD; Kota Fukai[1], MD, PhD; Kei Sano[2], MD, PhD; Masayuki Tatemichi[1], MD, PhD

[1]Department of Preventive Medicine, Tokai University, School of Medicine, Kanagawa, Japan

[2]Department of Ophthalmology, The Jikei University, School of Medicine, Tokyo, Japan

Corresponding Author:
Kota Fukai, MD, PhD
Department of Preventive Medicine
Tokai University, School of Medicine
143 Shimokasuya
Isehara
Kanagawa, 259-1193
Japan
Phone: 81 463 93 1121 ext 2612
Fax: 81 463 92 3549
Email: kota229@tokai.ac.jp

Abstract

Background: Although cancer screening is essential for early detection and an improved prognosis, screening beyond the recommended guidelines may increase the risk of false-positive results. Consequently, educating individuals about the potential harm of non–guideline-based cancer screening is essential; however, effective communication methods remain unclear.

Objective: This study aimed to evaluate the effectiveness of different types of educational videos in encouraging preferences for guideline-based cancer screening.

Methods: This 3-arm pseudorandomized controlled trial was conducted in June 2025 using a Japanese online survey platform. Eligible respondents were working adults aged 30 to 60 years with no history of major cancer. Respondents were assigned to 1 of the following 3 video conditions: video A, which provided a logical explanation of false-positive risks; video B, which presented the narrative of a woman who received a false-positive result from breast cancer screening; and video C, which depicted a man who underwent unnecessary follow-up testing after tumor marker screening. The primary outcome was the preference for guideline-based cancer screening after watching the videos. The secondary outcomes included 7 self-reported video evaluation items, such as perceived relevance and clarity, assessed using a 5-point Likert scale. Differences in the primary outcome between video groups were analyzed using multivariable logistic regression with adjustment for covariates. Means and 95% CIs were calculated for each secondary outcome according to sex and video group. In addition, before-and-after changes in screening preferences were assessed using McNemar test, with a significance level of .05.

Results: In total, 1200 respondents (400 per group) completed the survey. No statistically significant differences in the primary outcome were observed among the video groups. With reference to video A, the adjusted odds ratios for preferring guideline-based screening were 0.89 (95% CI 0.59-1.32) for video B and 0.98 (95% CI 0.65-1.46) for video C. Regarding secondary outcomes, male respondents rated video B less favorably than female respondents in terms of relevance and willingness to undergo guideline-based screening. The before-and-after comparison showed a significant change in preference for guideline-based screening (P=.04). These videos appeared to be more effective for individuals with an annual history of colorectal cancer screening than for those without such a history.

Conclusions: Educational videos have the potential to influence cancer screening preferences; however, no single video format has demonstrated clear superiority. These findings underscore the importance of tailoring educational materials to the target audience characteristics. Further research is required to develop effective strategies for encouraging guideline-based cancer screening.

XSL•FO
RenderX

Trial Registration: University Hospital Medical Information Network Clinical Trials Registry UMIN000060549; https://center6.umin.ac.jp/cgi-open-bin/ctr/ctr_view.cgi?recptno=R000066119

(J Med Internet Res 2026;28:e82322) doi:10.2196/82322

KEYWORDS

health education; randomized controlled trial; cancer screening; false positive; decision-making

Introduction

The strategic implementation of cancer screening facilitates early cancer detection, thereby improving patient survival through less invasive treatment options [1]. National guidelines, such as those developed by the US Preventive Services Task Force and the Japan National Cancer Center, clearly define the populations that benefit from screening and the tests recommended for them, based on robust evidence of mortality reduction [2,3]. Despite these guidelines, many individuals still undergo nonrecommended screenings, including breast cancer screening in people younger than the recommended age and whole-body positron emission tomography-computed tomography (PET-CT) imaging [4,5].

These non–guideline-based screenings often cause harm to participants through false-positive results, overdiagnosis, and overtreatment [6]. False-positive results often lead to negative psychosocial consequences [7]. In addition, prostate, lung, colorectal, and ovarian cancer screening trials reported that undergoing up to 14 cancer screening tests over a 3-year period resulted in cumulative risks of 60.4% and 48.8% for experiencing at least 1 false-positive result in men and women, respectively [8]. Moreover, approximately 10% of Japanese women in their 40s have been reported to experience false-positive results during breast cancer screening [9]. This risk is considered further elevated when screening is conducted at a younger age [10]. Furthermore, previous studies [11,12] have suggested that false-positive results can result in decreased adherence to future cancer screening. Accordingly, health care professionals play a crucial role in facilitating informed decision-making in cancer screening by helping participants understand both the benefits and harms of screening [13,14].

Previous studies [14,15] have evaluated the effects of various educational programs aimed at enhancing knowledge of the benefits and harms of cancer screening. For instance, multicomponent interventions based on behavior change models were implemented at the community level for breast cancer education [15]. These interventions were reported to be successful when various factors were considered, including the use of video consultations, provision of sessions in multiple languages, and consideration of the socioeconomic status of the community [16]. However, these educational efforts were designed to increase cancer screening uptake and not to alert people to the harmful effects of non–guideline-based cancer screening.

In Japan, the commercialization of health care has led to a strong emphasis on the benefits of early cancer detection and additional testing as much as possible. A survey on the implementation status of cancer screening in Japan [17] found that approximately 70% of companies offered colorectal cancer screening to individuals younger than 40 years, and 45% offered breast cancer screening to individuals younger than 40 years. A national survey in Japan [18] has shown that more than 50,000 healthy individuals in 2005 and approximately 39,000 per year from 2006 to 2009 underwent PET-CT cancer screening outside of guideline-based programs. Therefore, educating individuals about the potential harms of non–guideline-based cancer screening is essential for informed decision-making; however, effective communication methods remain unclear. One educational method can be to raise awareness about the risks associated with false-positive results.

We aimed to compare the short-term impact of 3 brief educational videos on preferences for guideline-based cancer screening. This intervention was grounded in the narrative transportation theory, which suggests that stories transport audiences into a narrative world, focusing attention, emotion, and imagery on the story and thereby reducing counterarguing and shaping beliefs and attitudes [19]. In this framework, we hypothesized that narrative videos would be more effective than an explanatory format in shifting preferences toward guideline-based screening. By comparing the effectiveness of the 3 types of educational videos, we gained deeper insight into how approaches are most effective for specific characteristics of the target audience.

Methods

Study Design and Respondents

This study was conducted according to the protocol registered in the University Hospital Medical Information Network Clinical Trials Registry (UMIN000060549). In June 2025, we conducted a 3-arm pseudorandomized controlled trial using a Japanese online survey platform. This trial was supported by Rakuten Insight Inc, a web-based research company located in Tokyo. The respondents were recruited from the company's registry via email. This manuscript is reported in accordance with the Consolidated Standards of Reporting Trials (CONSORT) 2010 statement (Multimedia Appendix 1).

The inclusion criteria were working individuals aged between 30 and 60 years who agreed to participate in the study. The exclusion criteria were a history of stomach, lung, colon, breast, or cervical cancer. In addition, individuals who reported undergoing cancer screening recommended by local governments within the past 3 years were excluded from the study, as these programs only offer screening options that adhere to national guidelines. The participants were administered a questionnaire comprising 27 questions. Respondents who required <5 minutes or >30 minutes to complete the

XSL•FO
RenderX

questionnaire were considered to have given inappropriate answers and were excluded.

Randomization and Procedure

The respondents were not individually randomized. Instead, separate web-based questionnaires were prepared for each of the 3 video conditions (A, B, and C). Invitations were sent to participants until 400 valid responses were obtained for each video group. Invitations for each group were distributed sequentially and exclusively, and individuals who were invited to one video condition were not invited to any other condition. The target population included 200 men and 200 women in 1 group, resulting in 3 groups with a total target sample size of 1200 respondents. Each group viewed a different video. We ensured that the respondents were unable to answer the next question unless they watched the video until the end, to ensure data quality. The sample size was determined based on previous studies of educational videos using animation [20].

Intervention

We developed 3 educational videos to inform people about the potential harm caused by false-positive results in cancer screening and to encourage guideline-based screening decisions. Each educational video was developed based on the health belief model [21], which explained that health-related behavioral change can be influenced by individuals' perceptions of expected benefits and costs. In the context of cancer screening, the perceived benefits of guideline-based screening were detecting cancer at an earlier stage, enabling less invasive treatment, and potentially reducing cancer mortality. The costs included financial and time burdens, psychological distress triggered by false-positive results, and potential harm from unnecessary follow-up procedures such as avoidable radiation exposure. By presenting the benefits of guideline-concordant screening and the potential harms of non–guideline-based screening, the videos were designed to support informed decision-making and encourage preferences aligned with screening guidelines.

All videos had a duration of approximately 100 seconds. Video A illustrated the disadvantages of excessive cancer screening using charts and numerical data. Video B showed the story of a woman in her 30s who underwent breast cancer screening, was referred for further testing, and ultimately received a false-positive result. Video C showed the story of a working man who underwent multiple tumor marker tests, received positive results in several categories, and was subsequently required to undergo numerous further tests. The structures of the 3 videos are shown in Multimedia Appendix 2. Video A is shown in Multimedia Appendix 3; video B is shown in Multimedia Appendix 4; and video C is shown in Multimedia Appendix 5. The duration of video A was 1 minute and 24 seconds, whereas videos B and C lasted for 1 minute and 17 seconds each.

Outcome Measures

The primary outcome was preference for guideline-based cancer screening after watching the videos. Respondents were asked what type of cancer screening they would prefer, with the following options: (1) "I want to undergo cancer screening as recommended by the guidelines," (2) "I want to undergo cancer screening at an earlier age than recommended," (3) "I want to undergo different types of cancer screening than recommended (eg, tumor markers or PET-CT imaging)," (4) "I want to undergo cancer screening at intervals shorter or longer than the recommended interval," (5) "I do not want to undergo any cancer screening," and (6) "I have no idea." Options 2, 3, and 4 were not mutually exclusive, and if any of these options were chosen, the response was interpreted as a preference for non–guideline-based cancer screening. The primary outcome measured an immediate postexposure preference and was intended as a proximal indicator of decision-making.

We included a 7-item evaluation survey as a brief process evaluation to assess how participants received the educational messages and to aid interpretation of the primary outcome. Conceptually, the items included (1) perceived personal relevance and engagement with the message (relevance), which is central to narrative persuasion and transportation; (2) comprehensibility and cognitive processing of the content (clarity), which is a prerequisite for informed decision-making; (3) perceived informational sufficiency for making a screening decision (informativeness), corresponding to the "knowledge or understanding" component needed to weigh benefits and harms; (4) perceived acceptability as educational material (acceptability), reflecting practical feasibility and appropriateness for dissemination; and (5) negative affect and avoidance tendency (aversion), which relates to potential unintended emotional burden and message avoidance [22]. Respondents rated these aspects on a 5-point Likert scale (1=strongly disagree, 2=disagree, 3=neutral, 4=agree, and 5=strongly agree): relevance ("This video is relevant to me"), clarity ("This video is clear and easy to understand"), informativeness ("This video provides all the necessary information for screening participants"), acceptability ("This video is acceptable as educational material"), and aversion ("I do not want to think about this video"). In addition, respondents were asked about the influence of the videos on 2 aspects: motivating them to discuss their content with family or friends and undergoing guideline-based cancer screening using the same 5-point scale.

Covariates

Covariates included sex, age, education, and preferences for cancer screening before viewing the video. Sex was based on self-reported information, and age was determined from the date of birth. Education was categorized into middle school, high school, vocational schools, junior college or technical college, university, graduate school, currently enrolled, and other, and respondents were asked to select the most appropriate option. However, for analysis, education was treated as a binary variable: less than a university degree vs a university degree or higher. The preferences for cancer screening before viewing the video were obtained using the same method as that used for the primary outcome.

Statistical Analysis

The distributions of age, marital status, education, income, and preferences for cancer screening before watching the video were tabulated for the 3 groups to compare the characteristics of the

XSL•FO
RenderX

respondents. The primary outcome variable was coded as 1 if the respondent selected the option "I want to undergo guideline-based cancer screening" after watching the video, and 0 otherwise. The adjusted odds ratios (aORs) and 95% CIs for videos B and C, relative to video A, were estimated using multivariate logistic regression analysis. The covariates included sex, age, education, and a preference for cancer screening before watching the video.

The means and 95% CIs were calculated for each of the secondary outcomes, which were the 7 self-reported video evaluation indicators, stratified by sex. In addition to the prespecified primary between-group comparison, we conducted an exploratory, post hoc before-and-after analysis of screening preferences immediately before and after video viewing across all participants. This analysis was not prespecified in the trial protocol and should be interpreted as descriptive and hypothesis generating. A 2-sided McNemar test was performed at a significance level of .05. Changes in cancer screening preferences before and after viewing the videos were visualized using a Sankey diagram.

We compared 3 groups to examine the characteristics of participants who perceived the video as effective: individuals who shifted toward guideline-based screening (effective group), those whose preferences did not change (neutral group), and those who shifted away from guideline-based screening (counterproductive group). To compare participant characteristics across the 3 groups, *P* values were calculated using Welch 1-way ANOVA for continuous variables and the chi-square test for categorical variables, with a 2-sided significance level set at 5%. These *P* values were provided for descriptive purposes; no adjustment for multiple comparisons was applied. Age and total response time were treated as continuous variables, while age group (10-year categories), sex, marital status, education, income, and history of colorectal cancer screening were treated as categorical variables. All statistical analyses were performed using R (version 4.4.2; R Foundation for Statistical Computing). Because all questions were mandatory, the dataset contained no missing values.

Ethical Contributions

This study was approved by the Research Ethics Committee of Tokai University School of Medicine (24R099). Informed consent was obtained from all respondents before they responded to the questionnaire. Because cancer-related educational materials may unintentionally elicit fear or distress, potential adverse emotional effects could not be fully excluded. During video production, we intentionally avoided sensational expressions. The video materials were submitted for institutional ethics review and were evaluated from a third-party perspective. Before participation, respondents were informed that the survey included videos addressing the potential harms and benefits of cancer screening. Respondents were also informed that participation was voluntary, that they could discontinue the survey at any time, and that discontinuation would not result in any disadvantage. Participation was incentivized by compensating respondents who completed the survey with points that could be exchanged for merchandise on online platforms. To protect respondents' privacy, we avoided collecting personally identifiable information. We confirmed that the data contained no personal information before conducting our analysis.

Results

Characteristics of Respondents

Table 1 summarizes the characteristics of the overall cohort and the individual video groups. The mean age of 1200 participants was 48.9 (SD 7.4) years, and the most common age group was 50 to 60 years (n=647, 53.9%). The proportion of unmarried individuals was 43.4% (n=521), and the most common educational attainment was a university degree (n=502, 41.8%). The most frequently reported income level was ≤JP ¥4 million (US $1.00=JP ¥145 is used throughout the text), with 310 (25.8%) respondents. The most common preference for cancer screening before watching the video was guideline-based screening (n=565, 47.1%).

XSL•FO
RenderX

Table 1. Characteristics of all respondents (N=1200).

Characteristics	All videos (n=1200)	Video A (n=400)	Video B (n=400)	Video C (n=400)
Age (years), mean (SD)	48.9 (7.4)	49.13 (7.5)	48.31 (7.8)	49.17 (7.3)
Age (years), n (%)				
30-39	162 (13.5)	48 (12.0)	65 (16.2)	49 (12.2)
40-49	391 (32.6)	131 (32.8)	133 (33.2)	127 (31.8)
50-60	647 (53.9)	221 (55.2)	202 (50.5)	224 (56.0)
Marital status, n (%)				
Unmarried	521 (43.4)	180 (45.0)	173 (43.2)	168 (42.0)
Education, n (%)				
Junior high school	9 (0.8)	3 (0.8)	2 (0.5)	4 (1.0)
High school	299 (24.9)	110 (27.5)	101 (25.2)	88 (22.0)
Vocational school graduate	160 (13.3)	53 (13.2)	54 (13.5)	53 (13.2)
Junior college or technical college graduate	139 (11.6)	45 (11.2)	46 (11.5)	48 (12.0)
University graduate	502 (41.8)	161 (40.2)	166 (41.5)	175 (43.8)
Graduate school graduate	82 (6.8)	25 (6.2)	29 (7.2)	28 (7.0)
Currently enrolled	9 (0.8)	3 (0.8)	2 (0.5)	4 (1.0)
Income (Japanese yen; million , US $1.00 = ¥145), n (%)				
≤4.00	310 (25.8)	115 (28.7)	99 (24.8)	96 (24.0)
4.01-6.00	270 (22.5)	90 (22.5)	100 (25.0)	80 (20.0)
6.01-8.00	240 (20.0)	80 (20.0)	71 (17.8)	89 (22.2)
8.01-10.00	163 (13.6)	48 (12.0)	58 (14.5)	57 (14.2)
10.01-12.00	98 (8.2)	24 (6.0)	32 (8.0)	42 (10.5)
12.01-15.00	70 (5.8)	27 (6.8)	26 (6.5)	17 (4.2)
≥15.01	49 (4.1)	16 (4.0)	14 (3.5)	19 (4.8)
Preference for cancer screening before watching the video, n (%)				
“I want to undergo cancer screening in accordance with the guidelines.”	565 (47.1)	177 (44.2)	183 (45.8)	205 (51.2)
“I want to undergo cancer screening that differs from the guidelines.”	383 (31.9)	135 (33.8)	136 (34.0)	112 (28.0)
“I don’t want to get screened for cancer.”	67 (5.6)	18 (4.5)	25 (6.2)	24 (6.0)
“I have no idea.”	185 (15.4)	70 (17.5)	56 (14.0)	59 (14.8)

Effect of Video Viewing on Cancer Screening Preferences Based on Guidelines

Figure 1 presents the aORs and 95% CIs for the primary outcome based on logistic regression analysis. Compared with video A, the aOR for video B was 0.89 (95% CI 0.59-1.32) and that for video C was 0.98 (95% CI 0.65-1.46), indicating no statistically significant differences. Multimedia Appendix 6 presents the results for all variables included in the logistic regression analysis.

XSL•FO
RenderX

Figure 1. Effect of videos on guideline-based screening preference. Analyses were adjusted for sex, age, education (university degree or lower), and preference for cancer screening before watching the video.

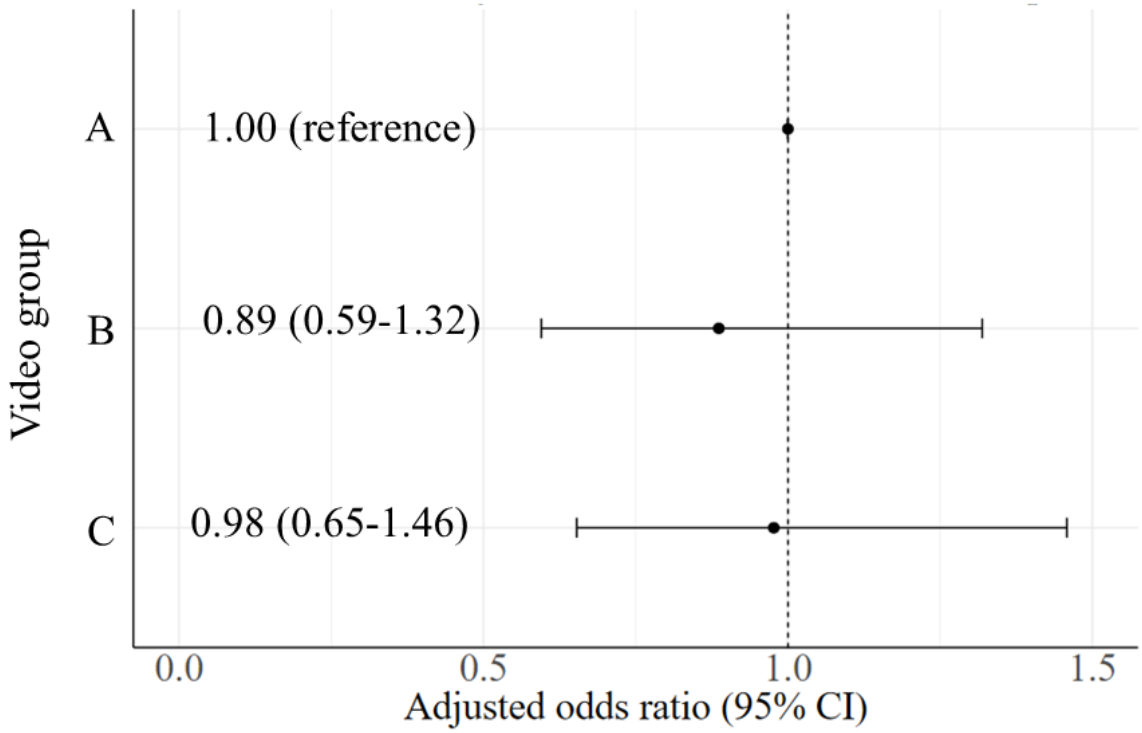

Evaluation of the Videos

Figure 2 shows the results of the secondary outcomes (7-item evaluation) for each video stratified by sex. The perceived relevance of video B was significantly lower among men (mean 2.5; 95% CI 2.4-2.7) than among women (mean 3.5; 95% CI 3.4-3.7). Similarly, willingness to undergo guideline-based screening was significantly lower among men (mean 3.3; 95% CI 3.1-3.4) than among women (mean 3.7; 95% CI 3.5-3.8). All scores, rounded to the second decimal place, are presented in Multimedia Appendix 7. Although no statistically significant differences were observed, the point estimates for all evaluation items among the women tended to be equal to or higher for video B than for video C.

XSL•FO
RenderX

Figure 2. Mean score with 95% CI for each video and sex group.

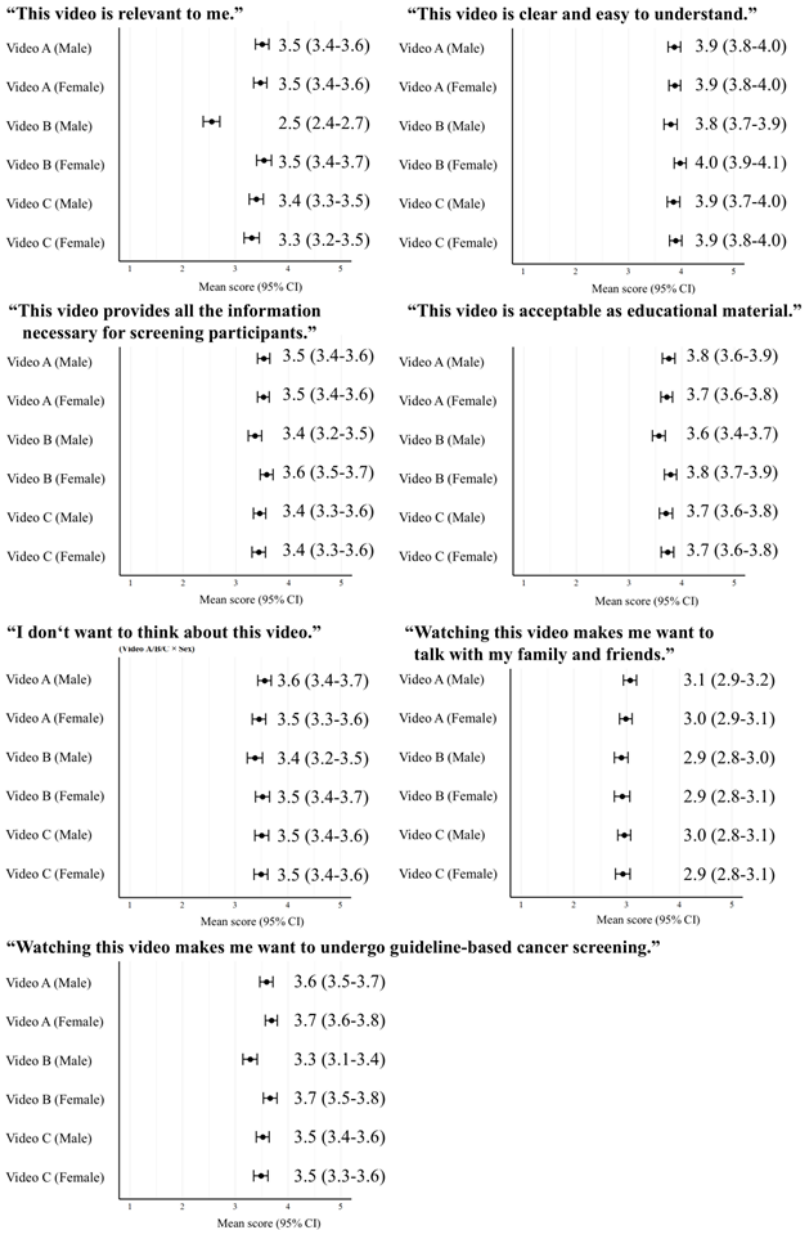

Total Changes in Cancer Screening Preferences Before and After Viewing the Video

Figure 3 illustrates the changes in screening preferences before and after watching the video using a Sankey diagram. Among the respondents, 91 shifted toward guideline-based screening (effective group), 283 maintained non–guideline-based preferences (neutral group), and 57 transitioned from guideline-based to non–guideline-based screening preferences (counterproductive group). Table 2 presents the characteristics of these 3 groups. A history of annual colorectal cancer screening over the previous 2 years was reported to be 57.1% (52/91) in the effective group, 52.7% (149/283) in the neutral group, and 43.9% (25/57) in the counterproductive group (P=.04). Although no statistically significant difference was observed in median total response time, the counterproductive group tended to have a shorter response time (436 seconds) than the effective group (459 seconds) and the neutral group (464 seconds). The density distribution of the total response times for these groups is shown in Multimedia Appendix 8. Before watching the video, 383 respondents preferred non–guideline-based cancer screening, and after viewing the video, 91 (23.8%) shifted to a preference for guideline-based screening.

XSL•FO
RenderX

Figure 3. Changes in preferences regarding cancer screening before and after watching the videos. Flow widths represent the proportion of participants who were consistent in or transitioned between preference categories. Colors denote response categories and are consistent across the "before and after watching the video" nodes. Blue represents “I want to undergo guideline-based cancer screening.” Orange represents "I want to undergo cancer screening that differs from the guidelines," which included the following survey response options: “I want to undergo cancer screening at an earlier age than recommended,” “I want to undergo different types of cancer screening than recommended (eg, tumor markers or PET-CT imaging),” and “I want to undergo cancer screening at intervals shorter or longer than the recommended interval.” Green represents “I do not want to undergo any cancer screening.” Purple represents “I have no idea.”.

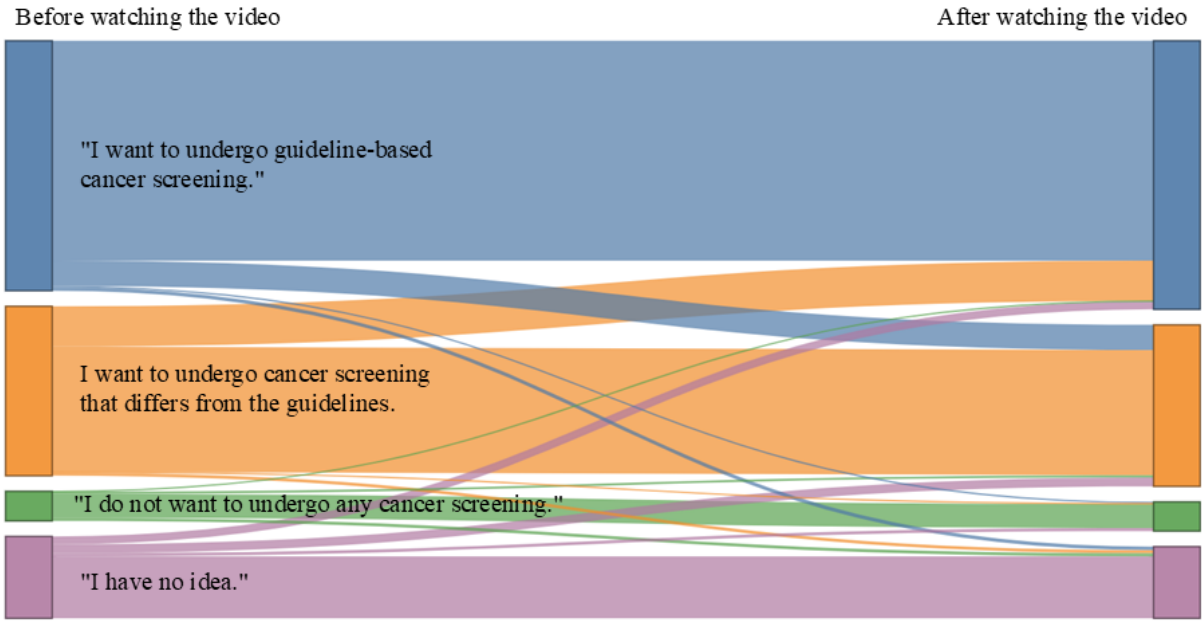

XSL•FO
RenderX

Table 2. Characteristics of respondents in the effective, neutral, and counterproductive groups (N=431).

Characteristics	Effective group (n=91)	Neutral group (n=283)	Counterproductive group (n=57)	*P* value
Age (years), mean (SD)	50.0 (7.1)	48.5 (7.6)	48.8 (7.8)	.26
Age (years), n (%)				.61
30-39	8 (8.8)	43 (15.2)	7 (12.3)	
40-49	31 (34.1)	85 (30.0)	17 (29.8)	
50-60	52 (57.1)	155 (54.8)	33 (57.9)	
Sex, n (%)				
Female	49 (53.8)	138 (48.8)	29 (50.9)	.70
Marital status, n (%)				
Unmarried	34 (37.4)	101 (35.7)	20 (35.1)	.95
Education, n (%)				
University degree or higher	48 (52.7)	152 (53.7)	23 (40.4)	.18
Income (Japanese yen; million; US $1.00 = ¥145), n (%)				.93
≤4.00	22 (24.2)	66 (23.3)	16 (28.1)	
4.01-6.00	17 (18.7)	56 (19.8)	12 (21.1)	
6.01-8.00	23 (25.3)	52 (18.4)	12 (21.1)	
8.01-10.00	13 (14.3)	43 (15.2)	6 (10.5)	
10.01-12.00	6 (6.6)	29 (10.2)	5 (8.8)	
12.01-15.00	7 (7.7)	19 (6.7)	4 (7.0)	
≥15.01	3 (3.3)	18 (6.4)	2 (3.5)	
History of colorectal cancer screening, n (%)				.04
Annually (past 2 years)	52 (57.1)	149 (52.7)	25 (43.9)	
Once (past 2 years)	12 (13.2)	15 (5.3)	4 (7.0)	
Earlier than past 2 years	4 (4.4)	38 (13.4)	11 (19.3)	
Never received	23 (25.3)	76 (26.9)	16 (28.1)	
Do not know	0 (0.0)	5 (1.8)	1 (1.8)	
Total response time (seconds), median (IQR)	459 (366–603)	464 (383–600)	436 (356–522)	.22

Discussion

Principal Findings

In this study, we aimed to identify effective educational strategies to convey the risk of false-positive results associated with non–guideline-based cancer screening. No statistically significant differences were observed in preferences for guideline-based cancer screening among the 3 videos designed to educate participants regarding the harm caused by false-positive results. The subjective evaluation provided additional insights into the effectiveness of video-based educational methods beyond the primary preference outcome. Among men, the evaluation of the narrative regarding breast cancer screening yielded lower scores for relevance and willingness to undergo guideline-based screening compared with evaluations among women. In contrast, the breast cancer screening narrative was generally received more favorably by women than the tumor marker narrative presented by a male character. Although not specified in the protocol beforehand, the McNemar test indicated a significant change in preferences after watching the videos. Among the 383 respondents who initially preferred non–guideline-based cancer screening, only 91 (23.8%) changed their preference to guideline-based screening after watching videos. The effective group underwent annual colorectal cancer screening. Meanwhile, the counterproductive group showed shorter response times than the effective and neutral groups.

Narrative educational videos are not expected to be more effective than logical or explanatory videos in cancer screening education. However, according to a systematic review of health education using videos [23], narrative presentations that evoke real-world situations are more effective in promoting behavioral change than the simple didactic delivery of information. Therefore, we created a control video that provided information and 2 narrative videos and hypothesized that narrative videos would be more effective than the control video; however, our results did not support this hypothesis. The findings of this study are consistent with those of a previous study examining the impact of different animation presentation formats on educational outcomes [24]. The previous study presented

XSL•FO
RenderX

animated videos that explained the advantages and disadvantages of colorectal cancer screening using 3 formats: videos with animated pictographs, videos with static pictographs, and audio booklets accompanied by static pictographs. No significant differences were observed among the 3 formats in terms of knowledge acquisition or screening participation [24]. In contrast, an intervention study on lung cancer screening [25] reported that presenting an informational film in addition to a booklet was more effective than presenting only a booklet. The addition of films was found to enhance knowledge acquisition, both objectively and subjectively [25]. The absence of a statistically significant difference observed in this study may be explained by the fact that the outcome measure focused on changes in preference rather than knowledge acquisition. A systematic review of the effectiveness of animated videos found that they consistently led to greater knowledge acquisition than other educational materials and standard care; however, no clear evidence was available regarding their effectiveness in influencing attitudes and perceptions [20]. Therefore, future intervention studies are required to identify appropriate methods for presenting animated videos.

The absence of statistically significant differences in the primary outcome may also be explained in other ways. The intervention intensity may have been insufficient: all videos were brief and delivered in a single exposure, which might limit their effect on preferences. Measurement sensitivity may have been limited. The primary outcome was dichotomized into guideline-based preference vs all other options immediately after exposure, which may have masked more nuanced shifts, including the avoidance of non–guideline-based options. In addition, baseline preferences and audience heterogeneity may have contributed to the null findings. Nearly half of the respondents (565/ 1200, 47.1%) preferred guideline-based screening even before viewing the videos, which may have reduced the room for improvement (ceiling effects). Conversely, respondents with strong preexisting beliefs favoring early or intensive screening may require more tailored or repeated messaging to shift preferences. Considering these reasons, future intervention studies should be designed to examine differences in effectiveness across educational methods.

Narrative videos that focus on breast cancer may be less effective or appropriate for male audiences. The scores for relevance and willingness to undergo guideline-based screening for video B were lower among men than among women. Narrative transportation theory suggests that narrative messages are more persuasive when viewers perceive the story as personally relevant and can identify with the protagonist [19]. A breast cancer–focused narrative may have been less self-relevant for male respondents, potentially reducing narrative engagement and transportation and thereby explaining the lower ratings for perceived relevance and willingness to undergo guideline-based screening observed among men for video B. Audience engagement has been identified as a key factor even in the effectiveness of educational videos [26]. Because breast cancer is typically not viewed as a personal health issue by men, it may be less effective in fostering engagement among men. This interpretation is further supported by the finding that women tended to rate the breast cancer video more favorably than video C, which showed a male undergoing multiple tumor marker tests.

The harmful effects of screening are often regarded by screening participants as important prior knowledge when making informed decisions. A qualitative study in Australia [14] used focus group discussions on harms regarding mammography and reported that several participants considered this information important for making informed choices and expressed a strong desire to be informed in advance. However, a randomized controlled trial conducted in Australia [27] showed that informing participants about the harm of breast cancer screening did not reduce their willingness to undergo screening and helped maintain their level of knowledge. Although providing balanced information is important, public information sources often fail to adequately address the associated risks [28]. In this regard, video-based education offers repeatability and accessibility across geographic boundaries [29]. Accordingly, video B in this study helps resolve this issue for women.

A before-and-after comparison of video-based education suggests a potential influence on participants' intentions. Although the videos were only 100 seconds long, an increase in the preference for guideline-based cancer screening was observed. Individuals who undergo annual colorectal cancer screening are generally more likely to actively participate in screening for other cancers. If such individuals become more inclined to follow guideline-based screening after learning about false-positive results, it may indicate that the educational videos were effectively communicated to an appropriate audience.

Notably, education regarding the risks of non–guideline-based screening had a limited impact on changing individuals' preferences for guideline-based cancer screening. In this study, fewer than 1 in 4 respondents changed their preference from non–guideline-based to guideline-based screening after watching the videos. This may be partly explained by the misleading positive feedback loops that are intrinsic to cancer screening [30]. When cancer is detected through screening, the benefit of early detection tends to be emphasized, whereas a negative result reinforces the perception that one is responsible for one's health. Moreover, the cost and pain associated with follow-up examinations are often perceived as acceptable trade-offs for survival [30]. In addition, negative feedback from personal screening experiences is rare, which may help sustain positive perceptions. From a physician's perspective, considering the seriousness of identifying prostate cancer after not undergoing screening, the decision to undergo screening is a no-loss choice compared with not undergoing screening [31]. Furthermore, specific harms, such as fear, pain, and financial burden, are prioritized by physicians to explain the harms of screening. Consequently, problem-based learning curricula aimed at helping medical students understand harm are being developed and evaluated for effectiveness to address this unconscious bias [32].

This study contributes to the literature on cancer screening education in certain ways. First, although many educational video interventions primarily aim to increase screening participation, our videos specifically communicated the potential harms of non–guideline-based screening and evaluated changes

XSL•FO
RenderX

in preference for guideline-based screening. Second, we compared 3 brief videos in online delivery settings and assessed multiple perceptions that can guide the development of video-based educational materials. Third, observed sex differences in perceived relevance for the breast cancer narrative suggest that matching content to the target audience is important. Fourth, considering the trend toward earlier and excessive cancer screening in Japan, the finding that a large number of participants still preferred screening beyond recommended levels even after viewing the educational materials is an important insight. Future efforts should focus on clarifying the underlying problems and exploring effective intervention methods.

Limitations

Despite the positive outcomes, this study has some limitations. First, the generalizability of the results may be limited. The respondents in this study were individuals who were able to spend at least 5 minutes completing the online survey. Therefore, the findings may not be generalizable to populations that are less inclined to participate in online surveys. Second, the primary outcome was a self-reported preference assessed immediately after video exposure, which may not translate into actual screening behaviors in real-world settings. This limits the external validity and practical applicability of our findings. Future studies should incorporate longer follow-up periods and objective end points, such as medical records from hospitals and clinics, to evaluate whether brief video-based education results in behavioral change. Third, the circumstances under which respondents watched the educational videos during the survey, likely influencing their engagement and responses, are unclear. At the beginning of the survey, the respondents were instructed to watch the video until the end while listening to the audio; however, we were unable to verify whether they followed these instructions. For example, if the participants joined the survey while on a train without using headphones, they may have completed it without listening to the audio. Fourth, the before-and-after analysis was exploratory and not prespecified in the trial protocol. Although this analysis provides insight into immediate within-person shifts after exposure, it may be susceptible to analytic flexibility and should be interpreted cautiously. Therefore, the primary inference of this study is based on the prespecified between-group comparison of video formats.

Conclusions

In this online pseudorandomized trial, narrative videos did not demonstrate a greater effect than explanatory videos in increasing preferences for guideline-based cancer screening. This negative finding suggests that concise, single-video interventions are unlikely to produce large format-specific effects and that preferences for excessive screening may persist despite education about the harms of false-positive results. To address beliefs supporting screening beyond recommended levels, future interventions should consider increasing intensity and tailoring education content to individual characteristics.

Acknowledgments

The authors acknowledge the support of Rakuten Insight for facilitating participant recruitment and administering the web-based survey using their proprietary online panels. The authors also thank Tetsu Kuribayashi and colleagues at ABEJA Inc for their assistance in producing the educational videos. The authors are grateful to Editage for editing and reviewing this manuscript for English-language clarity. During manuscript preparation, GPT-5.2 [33] was used by authors to assist with English-language editing.

Funding

This study was supported by a Health Labour Sciences Research Grant from the Ministry of Health, Labour and Welfare (grant 23EA1003).

Data Availability

Data sharing is not applicable to this article as no data sets were generated or analyzed during this study. Reasonable requests for additional information regarding the study can be directed to the corresponding author.

Authors' Contributions

K Sakai and MT conceptualized and designed this study; K Sakai collected the data and performed the statistical analysis; and KF, YF, SN, K Sano, and MT contributed to the interpretation of the results and drafted the manuscript. All authors have reviewed, revised, and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT 2025 checklist.
[DOCX File , 35 KB - jmir_v28i1e82322_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Structures of 3 types of cancer screening promotion videos.
[DOCX File , 5100 KB - jmir_v28i1e82322_app2.docx]

Multimedia Appendix 3
Video A. Explanatory video on guideline-based cancer screening and false positives (84 seconds). The presenter uses a whiteboard and figures to describe the screening pathway from screening to confirmatory testing and to explain why false positives are more common when cancer prevalence is low. The video summarizes potential harm, such as anxiety, time, and costs of follow-up testing, alongside benefits of early detection and concludes with the message to follow evidence-based screening guidelines.
[MP4 File (MP4 Video), 98225 KB - jmir_v28i1e82322_app3.mp4]

Multimedia Appendix 4
Video B. Narrative video showing a false-positive experience after early-age breast cancer screening (77 seconds). A female employee receives an abnormal result, struggles to arrange follow-up, and experiences stress and practical burdens, such as repeated visits and time off work, before learning that no cancer is present. The video highlights that breast cancer risk is low at age 30 and ends by recommending starting the screening at the recommended age.
[MP4 File (MP4 Video), 158564 KB - jmir_v28i1e82322_app4.mp4]

Multimedia Appendix 5
Video C. Narrative video depicting false-positive results after tumor-marker testing (77 seconds). A male employee undergoes tumor-marker testing, receives abnormal results, and undergoes multiple follow-up examinations over several months, resulting in no abnormality. The video emphasizes the harms of unnecessary testing, such as monetary costs and repeated radiation exposure, and recommends to choose guideline-based cancer screening.
[MP4 File (MP4 Video), 74441 KB - jmir_v28i1e82322_app5.mp4]

Multimedia Appendix 6
Complete results from the analysis for the primary outcome.
[DOCX File , 22 KB - jmir_v28i1e82322_app6.docx]

Multimedia Appendix 7
All results from the 7-item evaluation with mean score with 95% CI for each video.
[DOCX File , 21 KB - jmir_v28i1e82322_app7.docx]

Multimedia Appendix 8
Density distribution of total response time in 3 change groups.
[DOCX File , 613 KB - jmir_v28i1e82322_app8.docx]

References

1. Crosby D, Bhatia S, Brindle KM, Coussens LM, Dive C, Emberton M, et al. Early detection of cancer. Science 2022 Mar 18;375(6586):eaay9040. [doi: 10.1126/science.aay9040] [Medline: 35298272]
2. Wright K, Shkabari B, Booth C, Knopf K, Tregear M, Gyawali B. Congruence of cancer screening recommendations between the USPSTF and the top ten US cancer centers: a cross sectional study. EClinicalMedicine 2025 Apr;82:103169 [FREE Full text] [doi: 10.1016/j.eclinm.2025.103169] [Medline: 40235947]
3. Hamashima C. Cancer screening guidelines and policy making: 15 years of experience in cancer screening guideline development in Japan. Jpn J Clin Oncol 2018 Mar 01;48(3):278-286. [doi: 10.1093/jjco/hyx190] [Medline: 29315389]
4. Leavy MB, Starzyk K, Myers E, Curhan G, Gliklich R. Using real-world evidence to support a changing paradigm for cancer screening: a commentary. Pharmacoepidemiol Drug Saf 2020 Oct;29(10):1312-1315. [doi: 10.1002/pds.5104] [Medline: 32902061]
5. Ghotbi N, Iwanaga M, Ohtsuru A, Ogawa Y, Yamashita S. Cancer screening with whole-body PET/CT for healthy asymptomatic people in Japan: re-evaluation of its test validity and radiation exposure. Asian Pac J Cancer Prev 2007;8(1):93-97 [FREE Full text] [Medline: 17477780]
6. Rintala S, Dahlstrom KR, Franco EL, Louvanto K. A synthesis of evidence for cancer-specific screening interventions: a preventive medicine golden jubilee review. Prev Med 2023 Feb;167:107395 [FREE Full text] [doi: 10.1016/j.ypmed.2022.107395] [Medline: 36565859]
7. Rasmussen JF, Siersma V, Malmqvist J, Brodersen J. Psychosocial consequences of false positives in the Danish Lung Cancer CT Screening Trial: a nested matched cohort study. BMJ Open 2020 Jun 04;10(6):e034682 [FREE Full text] [doi: 10.1136/bmjopen-2019-034682] [Medline: 32503869]

XSL•FO
RenderX

8. Croswell JM, Kramer BS, Kreimer AR, Prorok PC, Xu J, Baker SG, et al. Cumulative incidence of false-positive results in repeated, multimodal cancer screening. Ann Fam Med 2009;7(3):212-222 [FREE Full text] [doi: 10.1370/afm.942] [Medline: 19433838]
9. Kasahara Y, Kawai M, Tsuji I, Tohno E, Yokoe T, Irahara M, et al. Harms of screening mammography for breast cancer in Japanese women. Breast Cancer 2013 Oct;20(4):310-315. [doi: 10.1007/s12282-012-0333-6] [Medline: 22282164]
10. Tsunematsu M, Kakehashi M. An analysis of mass screening strategies using a mathematical model: comparison of breast cancer screening in Japan and the United States. J Epidemiol 2015;25(2):162-171 [FREE Full text] [doi: 10.2188/jea.JE20140047] [Medline: 25483105]
11. Ford ME, Havstad SL, Flickinger L, Johnson CC. Examining the effects of false positive lung cancer screening results on subsequent lung cancer screening adherence. Cancer Epidemiol Biomarkers Prev 2003 Jan;12(1):28-33. [Medline: 12540500]
12. Miglioretti DL, Abraham L, Sprague BL, Lee CI, Bissell MC, Ho TH, et al. Association between false-positive results and return to screening mammography in the breast cancer surveillance consortium cohort. Ann Intern Med 2024 Oct;177(10):1297-1307. [doi: 10.7326/M24-0123] [Medline: 39222505]
13. Klabunde CN, Vernon SW, Nadel MR, Breen N, Seeff LC, Brown ML. Barriers to colorectal cancer screening: a comparison of reports from primary care physicians and average-risk adults. Med Care 2005 Sep;43(9):939-944. [doi: 10.1097/01.mlr.0000173599.67470.ba] [Medline: 16116360]
14. Hersch J, Jansen J, Barratt A, Irwig L, Houssami N, Howard K, et al. Women's views on overdiagnosis in breast cancer screening: a qualitative study. BMJ 2013 Jan 23;346:f158 [FREE Full text] [doi: 10.1136/bmj.f158] [Medline: 23344309]
15. Noman S, Shahar HK, Abdul Rahman H, Ismail S, Abdulwahid Al-Jaberi M, Azzani M. The effectiveness of educational interventions on breast cancer screening uptake, knowledge, and beliefs among women: a systematic review. Int J Environ Res Public Health 2020 Dec 31;18(1):e33 [FREE Full text] [doi: 10.3390/ijerph18010263] [Medline: 33396424]
16. Wu S, Chalela P, Ramirez AG. Changes in knowledge and awareness for a community-based cancer screening educational program. Arch Public Health 2023 Jul 12;81(1):130 [FREE Full text] [doi: 10.1186/s13690-023-01144-w] [Medline: 37438840]
17. Minamitani M, Tatemichi M, Mukai T, Katano A, Ohira S, Nakagawa K. Adherence to national guidelines for colorectal, breast, and cervical cancer screenings in Japanese workplaces: a survey-based classification of enterprises' practices into "overscreening," "underscreening," and "guideline-adherence screening". BMC Public Health 2024 Aug 15;24(1):2223 [FREE Full text] [doi: 10.1186/s12889-024-19775-1] [Medline: 39148101]
18. Minamimoto R, Senda M, Jinnouchi S, Terauchi T, Yoshida T, Murano T, et al. The current status of an FDG-PET cancer screening program in Japan, based on a 4-year (2006-2009) nationwide survey. Ann Nucl Med 2013 Jan 20;27(1):46-57 [FREE Full text] [doi: 10.1007/s12149-012-0660-x] [Medline: 23086544]
19. Green MC, Fitzgerald K. Transportation theory applied to health and risk messaging. In: Nussbaum JF, editor. Oxford Research Encyclopedia of Communication. Oxford, UK: Oxford University Press; 2017.
20. Moe-Byrne T, Evans E, Benhebil N, Knapp P. The effectiveness of video animations as information tools for patients and the general public: a systematic review. Front Digit Health 2022;4:1010779 [FREE Full text] [doi: 10.3389/fdgth.2022.1010779] [Medline: 36777110]
21. Lau J, Lim TZ, Jianlin Wong G, Tan KK. The health belief model and colorectal cancer screening in the general population: a systematic review. Prev Med Rep 2020 Dec;20:101223 [FREE Full text] [doi: 10.1016/j.pmedr.2020.101223] [Medline: 33088680]
22. Correia D, Kokole D, Rehm J, Tran A, Ferreira-Borges C, Galea G, et al. Effect of alcohol health warning labels on knowledge related to the ill effects of alcohol on cancer risk and their public perceptions in 14 European countries: an online survey experiment. Lancet Public Health 2024 Jul;9(7):e470-e480 [FREE Full text] [doi: 10.1016/S2468-2667(24)00102-6] [Medline: 38942558]
23. Abu Abed M, Himmel W, Vormfelde S, Koschack J. Video-assisted patient education to modify behavior: a systematic review. Patient Educ Couns 2014 Oct;97(1):16-22. [doi: 10.1016/j.pec.2014.06.015] [Medline: 25043785]
24. Housten AJ, Kamath GR, Bevers TB, Cantor SB, Dixon N, Hite A, et al. Does animation improve comprehension of risk information in patients with low health literacy? A randomized trial. Med Decis Making 2020 Jan;40(1):17-28 [FREE Full text] [doi: 10.1177/0272989X19890296] [Medline: 31795820]
25. Ruparel M, Quaife SL, Ghimire B, Dickson JL, Bhowmik A, Navani N, et al. Impact of a lung cancer screening information film on informed decision-making: a randomized trial. Ann Am Thorac Soc 2019 Jun;16(6):744-751 [FREE Full text] [doi: 10.1513/AnnalsATS.201811-841OC] [Medline: 31082267]
26. Brame CJ. Effective educational videos: principles and guidelines for maximizing student learning from video content. CBE Life Sci Educ 2016;15(4):33 [FREE Full text] [doi: 10.1187/cbe.16-03-0125] [Medline: 27789532]
27. Hersch J, Barratt A, McGeechan K, Jansen J, Houssami N, Dhillon H, et al. Informing women about overdetection in breast cancer screening: two-year outcomes from a randomized trial. J Natl Cancer Inst 2021 Nov 02;113(11):1523-1530 [FREE Full text] [doi: 10.1093/jnci/djab083] [Medline: 33871631]
28. Attena F, Cancellieri M, Pelullo CP. Scarce information about breast cancer screening: an Italian websites analysis. Medicine (Baltimore) 2016 Dec;95(50):e5615 [FREE Full text] [doi: 10.1097/MD.0000000000005615] [Medline: 27977602]

XSL•FO
RenderX

29. Chatterjee A, Strong G, Meinert E, Milne-Ives M, Halkes M, Wyatt-Haines E. The use of video for patient information and education: a scoping review of the variability and effectiveness of interventions. Patient Educ Couns 2021 Sep;104(9):2189-2199. [doi: 10.1016/j.pec.2021.02.009] [Medline: 33741233]
30. Kramer BS, Croswell JM. Cancer screening: the clash of science and intuition. Annu Rev Med 2009;60:125-137. [doi: 10.1146/annurev.med.60.101107.134802] [Medline: 18803476]
31. Ransohoff DF, McNaughton Collins M, Fowler FJ. Why is prostate cancer screening so common when the evidence is so uncertain? A system without negative feedback. Am J Med 2002 Dec 01;113(8):663-667. [doi: 10.1016/s0002-9343(02)01235-4] [Medline: 12505117]
32. Ekhlas S, Lang E, Dickinson J. Optimizing overdiagnosis education for medical students. BMJ Evid Based Med 2023;28(Suppl 1):A29. [doi: 10.1136/EBM-2023-POD.60]
33. GPT-5.2. OpenAI. URL: https://openai.com/index/introducing-gpt-5-2/ [accessed 2025-05-25]

Abbreviations

aOR: adjusted odds ratio
PET-CT: positron emission tomography-computed tomography

Edited by A Schwartz, M Balcarras; submitted 13.Aug.2025; peer-reviewed by S Osera, Y Igarashi; comments to author 16.Dec.2025; accepted 19.Jan.2026; published 12.Feb.2026.

Please cite as:
Sakai K, Furuya Y, Nakazawa S, Fukai K, Sano K, Tatemichi M
Effectiveness of Educational Videos in Encouraging Preferences for Guideline-Based Cancer Screening in Japan: Three-Arm Pseudorandomized Controlled Trial
J Med Internet Res 2026;28:e82322
URL: https://www.jmir.org/2026/1/e82322
doi:10.2196/82322
PMID:

©Kosuke Sakai, Yuko Furuya, Shoko Nakazawa, Kota Fukai, Kei Sano, Masayuki Tatemichi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Seeking But Not Discussing Online Health Information With Physicians: Cross-Sectional Survey Study of eHealth Literacy–Empowerment Profiles and Patient-Centered Communication

Qianfeng Lu[1], PhD; Wen Jiao[2], PhD; Angela Chang[3], PhD; Peter Johannes Schulz[1,4], PhD

[1]Faculty of Communication, Culture and Society, Università della Svizzera italiana, Lugano, Switzerland

[2]School of Communication, Soochow University, Soochow, China

[3]Faculty of Social Sciences, University of Macau, Macau, China

[4]Department of Communication and Media, Ewha Womans University, Seoul, Republic of Korea

Corresponding Author:
Peter Johannes Schulz, PhD
Faculty of Communication, Culture and Society
Università della Svizzera italiana
Via Giuseppe Buffi 13
Lugano, 6900
Switzerland
Phone: 41 586664724
Email: schulzp@usi.ch

Abstract

Background: Patients frequently search for health information online and value physician support in evaluating and interpreting their findings, yet many hesitate to share their online searches with their physicians. This hesitation hinders shared decision-making and compromises patient care. While extensive research has examined patients' online health information–seeking behaviors, little has focused on patients' disclosure of this information to their physicians during consultations.

Objective: Guided by the Health Empowerment Model and the Linguistic Model of Patient Participation in Care, this study aims to (1) identify distinct patient profiles based on eHealth literacy and psychological health empowerment levels, (2) examine how these patient profiles differ in online health information seeking and disclosure to physicians, and (3) investigate whether patient-centered communication (PCC) promotes information disclosure and whether this effect varies by patient profile.

Methods: This cross-sectional study surveyed 2001 Chinese participants recruited through convenience sampling. Patient profiles were identified using *k*-means cluster analysis with standardized *z* scores of eHealth literacy and psychological health empowerment. Differences between profiles in information behaviors were examined using 1-way Welch ANOVA, chi-square tests, and pairwise comparisons. Regression analyses examined the association between PCC and disclosure of online health information. Moderation analyses using the Hayes PROCESS macro assessed whether this association varied across patient profiles.

Results: Four distinct patient profiles were identified: effective self-managers (996/2001, 49.8%), moderate-needs dependent patients (408/2001, 20.4%), high-needs patients (68/2001, 3.4%), and dangerous self-managers (529/2001, 26.4%). Profiles differed significantly in information-seeking intentions ($F_{3,289}$=62.09; P<.001; η^2=0.12) and disclosure intentions ($F_{3,299.41}$=66.08; P<.001; η^2=0.09). "Effective self-managers" showed the highest seeking (mean 4.01, 95% CI 3.96-4.06) and disclosure intentions (mean 3.43, 95% CI 3.36-3.50), while "high-needs patients" showed the lowest intentions for both behaviors. Actual information-seeking rates also differed significantly across profiles (χ^2_3=103.4; P<.001), with "effective self-managers" having the highest rate (800/996, 80.3%) and "high-needs patients" the lowest (25/68, 36.8%). Among seekers, disclosure rates varied significantly (χ^2_3=23.1; P<.001), with "high-needs patients" showing the highest disclosure (16/25, 64%) despite having the lowest seeking rate. PCC was positively associated with actual information disclosure behavior (odds ratio 1.26, 95% CI 1.04-1.53; P=.02), with no significant moderation by patient profiles (χ^2_3=1.7; P=.64).

XSL•FO
RenderX

Conclusions: This study extends existing literature from information-seeking behavior to patients' disclosure of online findings to physicians. Unlike prior research that examined eHealth literacy and psychological health empowerment separately, this study integrated these constructs to identify meaningful patient profiles with distinct information behavior patterns. PCC facilitates disclosure regardless of patient profile. For practice, physicians should adopt a PCC that acknowledges patients' online research efforts, promoting safer information use and stronger patient-physician relationships.

(J Med Internet Res 2026;28:e78836) doi:10.2196/78836

KEYWORDS

empowerment; literacy; patient-provider communication; patient-centered communication; health information; misinformation

Introduction

Background

The use of digital platforms and online information channels is becoming increasingly common for patients seeking health-related information [1,2]. Accessing online health information offers various benefits to patients. By seeking health information online, patients can gain social support, be better informed, and become more prepared to participate in their health care [3-5]. However, drawbacks exist. Online searches may increase patient anxiety, create confusion, and even lead to medical nonadherence [6,7]. The risks are associated with the potential misuse of online health information [8,9], as a significant portion of online content contains misinformation [10] or may be incomplete or unsuitable for a patient's specific condition [11].

Physicians can play a crucial role in guiding patients to navigate online health information [12]. Patient perspectives showed that they generally want physicians to help them understand and assess the suitability of online health information for their specific condition [4,13]. Medical consultations provide valuable opportunities for patients to receive professional advice tailored to their specific conditions. When physicians discuss online health information with patients during medical encounters, they can identify and possibly correct misconceptions patients acquired through internet searches, potentially preventing the misuse of online materials [14]. Additionally, when physicians and patients discuss internet-sourced health information, it helps doctors better understand patients' perceptions of their conditions and health care needs [15]. This understanding, even when it reveals discrepancies between professional medical advice and patients' personal beliefs, creates opportunities for forming more engaged and shared decision-making [14]. Interview studies have shown that conversations about internet-sourced information can enhance patient engagement during consultations [16], strengthen the physician-patient relationship [17], and foster mutual trust [18].

Despite these benefits, many patients still withhold internet-sourced health information from their physicians [19,20]. Barriers to discussing online findings include limited consultation time, concerns about online information quality (such as conflicting or inaccurate content), and patients' perceptions that physicians may be unwilling to engage in such discussions [21,22]. With patients increasingly turning to online resources as their first source of health information prior to consultations [23], it is essential to understand not only their search behaviors but also what influences their decision to disclose or withhold these findings during medical encounters. However, the existing literature has primarily focused on search behaviors while largely neglecting the disclosure aspects [21].

This study examines patients' online health information seeking and their disclosure of these findings to physicians during medical encounters. We draw on the Health Empowerment Model (HEM) [24] to understand how patient characteristics shape information behaviors, and the Linguistic Model of Patient Participation in Care [25] to examine how physician communication influences disclosure.

Theoretical Framework

Health Empowerment Model

The HEM recognizes the distinct but complementary roles of psychological health empowerment and health literacy in influencing health behaviors [24]. Health literacy refers to the ability to access, understand, and evaluate health information [26], and psychological health empowerment is more about the patient's sense of control, confidence, and motivation to act actively in managing their health [27]. Without literacy skills, patients may struggle to identify reliable health information; without empowerment, even accurate information may go underused [24]. Together, these two factors enable patients to make informed decisions and actively manage their health [24]. This model is especially valuable in the digital age, where the use of online health information has increased rapidly, intensifying the need for patients to possess both the literacy skills to accurately evaluate online information (eHealth literacy) and the motivation and confidence to seek and use it (psychological health empowerment) [28]. Depending on patients' ability to navigate online content, online health information can empower but may also misinform them [29].

The HEM delineates 4 distinct combinations of empowerment and health literacy, each of which can be associated with health information behaviors uniquely [30]. The "effective self-manager" represents the ideal scenario: an empowered patient with high eHealth literacy who proactively seeks out and acts on that online health information, making informed decisions based on their ability to evaluate information accurately. In contrast, the "dangerous self-manager" is empowered but has low eHealth literacy. They are proactive in seeking out online information, yet lack the skills to evaluate it accurately, which may lead to poor or even harmful health decisions [24]. The "needlessly dependent patient" has the high literacy skills to understand online health information but lacks the empowerment to act on it, often deferring decisions to others despite having the knowledge to engage more actively in their

XSL•FO
RenderX

care. Finally, the "high-needs patient" is deficient in both empowerment and eHealth literacy, typically taking a passive role in their health care and requiring support to find, understand, and apply online health information [24].

The HEM has been applied to examine patient health status [31], help-seeking behaviors [32], medical decision-making [33], chronic disease management [34,35], and participation in online health communities [36]. These studies consistently demonstrate that patients require both high empowerment and health literacy to effectively manage their health and use health resources, including online health information [37]. Deficiencies in either factor lead to suboptimal health behaviors and poorer outcomes. The interplay between empowerment and literacy creates different patient profiles with varied ways of managing health, seeking information, and engaging with health care providers [24]. Building on this framework, this study aims to identify distinct patient categories based on the combination of these two factors in the context of online health information behaviors.

eHealth Literacy

The traditional concept of health literacy is defined as the personal cognitive and social skills to access, understand, and use health information. eHealth literacy extends it by emphasizing the knowledge and competencies required to navigate electronic information platforms [38-40]. The increasing prevalence of health information and proliferation of misinformation on the internet have made these skills more urgent than ever [41]. Individuals with higher eHealth literacy are more likely to seek health information online [42,43]. For example, empirical studies have found that individuals with higher levels of eHealth literacy are more likely to use health applications to acquire information [44,45], access social media platforms to find and share health-related content [43], and explore national health information portals for health care [46].

Patients' eHealth literacy skills shape both the quality of information they find online and their communication with health care providers [47,48]. A systematic review found that patients with low health literacy had more difficulty evaluating online health information and showed less trust in such information [29]. Conversely, those with higher literacy communicate more clearly with providers as they can articulate their problems and information needs more precisely [47]. While these review findings were based on general health literacy, they likely extend to eHealth literacy, given that fundamental health literacy skills are also core components of eHealth literacy [39]. Empirical findings on whether eHealth literacy leads patients to disclose their internet findings with providers are mixed: some studies show a positive association [40], whereas others report no significant link [49]. These inconsistencies indicate that, beyond the ability to handle online information, a motivational driver may also be important to consider. Drawing on the HEM, we include psychological health empowerment as a complementary determinant to examine alongside eHealth literacy.

Psychological Health Empowerment

A concept that often intertwines with health literacy is psychological health empowerment [50]. Psychological health empowerment is also frequently associated with online health information behavior but often studied separately from eHealth literacy [4]. It is sometimes referred to as health empowerment or patient empowerment [27,51]. Health empowerment has been consistently associated with more active online health information seeking. Patients view online information-seeking as a means to be informed and prepared to manage their health conditions [4]. By accessing online health resources, patients perceive greater control over their conditions [52] and increased confidence when communicating with physicians during medical consultations [53,54]. In essence, while empowered patients are more likely to seek information, the act of seeking and finding relevant health information online further enhances their sense of empowerment [37].

Health empowerment also determines how information is integrated into health decision-making [55]. When patients feel empowered, they actively engage in decision-making with health care providers [56]. This sense of empowerment motivates patients to take an active role in managing their health, including seeking and using online health resources. Empowered patients are more likely to bring their online findings into clinical discussions with physicians as a means to address their information needs, clarify uncertainties, and participate in decision-making [57]. In contrast, patients with low empowerment often remain passive in health care interactions with their providers [57,58]. Despite potentially having found the same information online, these less-empowered patients may hesitate to disclose their findings during consultations, instead deferring entirely to physician direction [24].

To sum up, eHealth literacy equips patients with the skills needed to critically assess the reliability and relevance of online content they encounter [39]. Psychological health empowerment, on the other hand, reflects patients' motivation and confidence to act on this information and actively participate in their health care [59]. The HEM posits that both eHealth literacy and psychological health empowerment are necessary for effective use of health information and that their combination determines how patients engage with health information. Although previous research has examined these two concepts, they were typically studied separately, with studies often focusing on either health literacy or health empowerment but failing to integrate both [37,60]. To fill this gap, this study also examines health information behaviors across different eHealth literacy and empowerment profiles, from online searching before consultations to information sharing during physician encounters.

Patient-Centered Communication

Despite recognizing the potential benefits of discussing online health information with physicians, many patients hesitate to share their internet findings during consultations [21]. This reluctance stems largely from patients' concerns about how physicians will react to their online search [22,61]. Empirical evidence demonstrates that patients' hesitation to disclose

XSL•FO
RenderX

internet-sourced information is directly related to their communication experience with physicians [4].

Patient-centered communication (PCC)—defined as a clinical communication approach where providers seek to understand patients' perspectives, reach shared understanding, and involve patients as partners in care decisions [62]—offers a promising approach to encourage information sharing. It differs fundamentally from the paternalistic model by positioning patients as equal participants in the clinical relationship. Research shows that PCC promotes patient health care participation and is linked to improved patient trust and satisfaction with providers [63,64]. When physicians adopt patient-centered approaches, patients feel their perspectives are valued [65] and are more likely to share internet-sourced information [21]. Conversely, when patients perceive physicians as unreceptive or threatened by their online research, they are unwilling to disclose [21,52,66].

The importance of physician communication is supported by the Linguistic Model of Patient Participation in Care [25]. This model identifies 3 key determinants of patient participation in medical encounters: predisposing factors, enabling factors, and physician communication behaviors. Predisposing factors encompass patients' sociodemographic and psychosocial characteristics, including psychological health empowerment, that influence their propensity to participate actively in health care encounters. Enabling factors refer to patients' communicative resources and capabilities, such as eHealth literacy skills, that determine their capacity to engage effectively with health care providers and articulate their needs. Physician communication behaviors include verbal and nonverbal practices, such as providing clear explanations and demonstrating empathy, that can either facilitate or inhibit patient participation. The model posits that these 3 factors interact rather than operate independently [25]. It suggests that even patients with high empowerment and literacy may withhold information if physician communication does not support their disclosure. Thus, PCC serves as a crucial contextual factor that can affect information disclosure, regardless of patients' individual characteristics.

PCC may not affect all patients equally. The HEM suggests that patients with different literacy and empowerment levels rely on their physicians differently [24]. Those with high empowerment and literacy may already feel confident sharing online findings, while patients with lower empowerment and literacy may be more sensitive to physician communication styles and depend more heavily on physician encouragement. This raises the question of whether the effect of PCC on disclosure varies across patient profiles. Therefore, the final aim of this study is to examine the impact of PCC on patient information disclosure and whether it varies across different patient profiles.

Based on the above, this study poses the following research questions (RQs):

- RQ1: What distinct patient profiles can be identified based on eHealth literacy and psychological health empowerment?
- RQ2: How do patients' online information-seeking behaviors prior to consultations and their subsequent disclosures to physicians differ across the identified patient profiles?
- RQ3a: How does patient-centered communication relate to patients' disclosure of online health information to physicians? RQ3b: How does this relationship differ across patient profiles?

Methods

This cross-sectional survey study is reported according to CHERRIES (Checklist for Reporting Results of Internet e-Surveys) guidelines [67].

Recruitment

This study focuses on China, the world's largest digital community, with approximately 1.08 billion internet users in 2023 [68]. This cross-sectional survey study was conducted as part of a larger research project examining health information-seeking behavior and health care use in China.

Data collection was conducted in Mainland China and Macau (Special Administrative Region of China) between September and November 2022. The survey was distributed online through Chinese social media platforms, including WeChat, QQ, and Sina Weibo. The survey was created using the Qualtrics platform and took approximately 15 minutes to complete.

Eligible participants were required to be Chinese, aged 18 years or older, and residing in Mainland China or Macau. Snowball convenience sampling was used. Initial recruitment began on the University of Macau campus, where research team members approached individuals in public areas (eg, libraries and classrooms) and invited them to participate. Participants who completed the survey were asked to share it with their friends and family members who met the eligibility criteria.

A total of 3387 respondents answered the survey. The final analytical sample included 2001 participants who provided complete responses on all focal variables (eHealth literacy, health empowerment, and information seeking and disclosure behavior). Among these, 1569 participants had received health care in the previous year and could provide responses regarding physician communication, forming the subsample for PCC analyses. No formal sample size calculation was conducted, and the sample size was determined by the number of complete responses obtained during the data collection period. Survey design and recruitment procedures are also published in Jiao et al [37].

Ethical Considerations

This study received ethical approval from the University of Macau's Ethical Committee for the Social Sciences and Humanities (application code: SSHRE22-APP093-FSS), which covered the complete data collection procedure. The original ethical approval included permission for secondary analyses of the collected data; therefore, no additional approval was required for this study. All participants provided informed consent prior to completing the survey. Participants were informed that the study aimed to help researchers and policymakers better understand population health behaviors related to internet use for health purposes. Participant privacy and confidentiality were

XSL•FO
RenderX

protected through anonymous data collection. No personally identifiable information was collected, and all survey responses were deidentified before data analysis. No compensation was provided to participants. Participation was entirely voluntary, and participants could withdraw at any time without consequence. No images or materials in this manuscript contain identifiable participant information.

Measures

eHealth Literacy

The eHealth Literacy Scale (eHEALS), a validated 8-item instrument developed by Norman and Skinner [69], was used to evaluate participants' eHealth literacy. It uses the operational definition of eHealth literacy as "an individual's ability to seek, find, understand, appraise, and apply health information from electronic sources to address or solve a health problem" and uses self-reported measures to assess perceived competence across these domains. Despite being developed nearly two decades ago, eHEALS continues to receive research attention with recent validation studies confirming its robustness across diverse populations [70,71]. The core competencies it measures (seeking, understanding, appraising, and applying health information from electronic sources) remain essential regardless of the online information source and represent foundational skills needed to navigate today's digital health information landscape.

We used the Chinese translation of eHEALS validated by Chang and Schulz [72], which demonstrated good psychometric properties in Chinese populations and has been empirically applied in studies among Chinese-speaking populations [73,74]. Responses were measured on a 5-point Likert scale, with higher scores indicating greater eHealth literacy. The mean of all item scores was calculated to represent participants' eHealth literacy levels. The reliability of eHEALS was assessed using Cronbach α. Additionally, we reported McDonald ω, which is less affected by the number of items (a limitation of Cronbach α) and does not assume equal factor loadings of items. Both reliability tests showed good internal consistency (Cronbach α=0.93; McDonald ω=0.93). The complete survey wording can be found in Multimedia Appendix 1.

Psychological Health Empowerment

The Psychological Health Empowerment Scale (PHES) was used to measure patient empowerment [27]. This established 8-item instrument assesses individuals' perceived control over one's health, perceived competence of understanding of one's health condition, and their motivation to achieve health goals. Sample items include statements such as "I can motivate myself to manage my health and make a better life" and "I can make every possible effort to achieve health goals." The PHES has been widely adopted across diverse populations to measure health empowerment [75]. We used the Chinese translation by Jiang and Street [76], which has subsequently been used in other studies with Chinese-speaking populations [77,78]. Responses were recorded on a 5-point Likert scale, with higher scores indicating a greater sense of health empowerment. The overall health empowerment level was calculated by averaging the 8 items (Cronbach α=0.90; McDonald ω=0.90).

Patient-Centered Communication

Participants who had received health care in the preceding year (n=1569) were asked about their communication experiences with physicians. The 7-Item Patient-Centered Communication Scale based on Epstein and Street's [79] model was used to assess how frequently physicians (1) provided opportunities to ask health-related questions, (2) paid attention to feelings and emotions, (3) involved patients in health care decisions to their desired extent, (4) ensured understanding of health management tasks, (5) explained medical information comprehensibly, (6) spent adequate time with patients, and (7) helped manage health-related uncertainty. This scale has been widely used to estimate PCC [80,81]. We used the Chinese translation from the Health Information National Trends Survey in China (HINTS-China) [82]. Responses were recorded on a 5-point Likert scale. The overall patient-physician communication was calculated by averaging the 7 items, with higher scores indicating better PCC (Cronbach α=0.89; McDonald ω=0.89).

Information Seeking and Disclosure Behavior

The two facets of participants' health information-seeking behavior in the clinic context, intended versus actual, were assessed. We inquired about their past behaviors regarding whether participants have sought out health information on the internet before consulting with a doctor (actual seeking) and whether such information was subsequently discussed during medical visits (actual disclosure). The two specific survey questions are "Reflecting on your recent medical consultations, did you search for health information on the Internet about your condition before meeting with your doctor?" For those who answered positively (n=1428), they were further queried, "Furthermore, did you discuss the information you found on the Internet with your doctor during the consultation?" A binary response of "yes" (=1) or "no" or "I am not sure" (=0) was applied.

The hypothetical scenarios aimed to elicit intended behaviors. Participants were presented with a situation in which they hypothetically experienced a new health symptom (eg, chest pain) and were asked to predict their likelihood of searching for health information online before a scheduled doctor's appointment (intended seeking) and their propensity to share the found information with the doctor (intended disclosure). The corresponding survey questions were as follows: "Imagine experiencing chest pain for one week, a symptom you've never had before. As a result, you decide to schedule a medical appointment with a new doctor. How likely are you to search for information about your condition on the Internet before visiting the doctor? And how likely are you to share the information you found on the Internet with the doctor?" A 5-point Likert scale ranging from "very unlikely" (=1) to "very likely" (=5) was used. A pilot study with university adults in Macau and feedback from the pilot study were used to refine these survey questions.

Control Variables

We collected participants' demographic information, including age, gender, and education level. Age was recorded in years, while education level was categorized based on the highest

XSL•FO
RenderX

grade completed, ranging from primary school and below (coded as 1) to a bachelor's degree and above (coded as 6). Gender was represented as a binary variable, with "female" coded as 0. In addition, self-rated general health status and trust in internet-sourced health information were included as control variables to minimize confounding effects. Participants were asked to rate their overall health status on a scale from 1 (poor) to 5 (excellent). Trust in health information from social media, websites, search engines, and apps was measured with a 5-point scale ("1=not at all" to "5=a lot"). The mean scores of these trust responses were calculated to provide an overall trust in online health information (Cronbach α=0.89; McDonald ω=0.80).

Statistical Analysis

Descriptive statistics were conducted to ascertain the overall sample characteristics. Then, a *k*-means cluster analysis was performed using the standardized *z* scores of 2 key variables: eHealth Literacy and Psychological Health Empowerment. This analysis was carried out in the R program (R Foundation for Statistical Computing), using the "NbClust" package to identify the optimal number of clusters [83]. The NbClust package determines the optimal number of clusters by computing different cluster validity indices, such as the Silhouette index, Dunn index, and Calinski-Harabasz criterion. Rather than relying on a single criterion, it evaluates multiple indices and provides a consensus solution based on the majority rule. NbClust has been widely used in previous empirical studies across various disciplines. Following the cluster analysis, the distinct patient eHealth Literacy and Empowerment Profiles (eHL-E Profiles) were established.

Then, 1-way ANOVA, using Welch correction to account for potential heterogeneity of variances, was conducted to examine differences in intended information-seeking and disclosure behaviors across the different eHL-E Profiles. Games-Howell post hoc comparisons were performed to identify specific between-group differences. Additionally, chi-square tests were used to assess differences in actual information-seeking and disclosure behaviors among the eHL-E Profiles. Pairwise comparisons with Bonferroni adjustment were further conducted to identify which specific eHL-E Profiles differed significantly from each other in their actual information behaviors.

Finally, regression analyses were performed to examine the association between PCC and patient information disclosure. For the dichotomous outcome actual disclosure (0=no, 1=yes), we fitted a binary logistic regression with PCC as the focal predictor. For the continuous outcome intended disclosure (interval scaled), we ran a multiple linear regression. All models controlled for eHealth literacy, health empowerment, self-rated health, overall trust in online health information, and demographic variables (age, gender, and education) to mitigate confounding bias. To test whether the effect of PCC on disclosure varied by patient profile groups, we estimated a simple moderation model (model 1) in the Hayes PROCESS macro (version 4.3), controlling for self-rated health, trust in online health information, and demographic variables. We requested 5000 bootstrap samples to obtain bias-corrected 95% CIs for all conditional effects.

Results

Sample Characteristics

Descriptive statistics are displayed in Table 1. The mean age of participants was 31.78 (SD 13.58) years, 64.3% (1202/1870) were female, and 82.5% (1508/1829) had received a college degree or higher education. General health status was rated, on average, as good (mean 2.98, SD 0.81). Overall trust in online health information was moderate (mean 2.48, SD 0.69).

XSL•FO
RenderX

Table 1. Demographic and health-related characteristics of study participants (N=2001).

Variables	Values
Age[a] (years), mean (SD)	31.78 (13.58)
Gender[b] , n (%)	
Female	1202 (64.3)
Male	668 (35.7)
Education[c]	
Mean (SD)	4.60 (1.07)
Primary school or below, n (%)	24 (1.3)
Junior middle school, n (%)	58 (3.1)
High school, n (%)	239 (13.1)
Junior college, n (%)	258 (14.1)
Bachelor's degree, n (%)	982 (53.7)
Master's degree or above, n (%)	268 (14.7)
Health status, mean (SD)	2.98 (0.81)
Trust in online health information, mean (SD)	2.48 (0.69)
eHealth literacy, mean (SD)	3.48 (0.70)
Health empowerment, mean (SD)	3.77 (0.59)
Patient-centered communication[d], mean (SD)	2.53 (0.63)
Intended seeking, mean (SD)	3.79 (1.01)
Intended disclosure, mean (SD)	3.19 (1.09)
Actual seeking (yes), n (%)	1428 (71.4)
Actual disclosure[e] (yes), n (%)	800 (56.0)

[a]Among the total sample of 2001 participants, missing data were observed for 139 participants.

[b]Among the total sample of 2001 participants, missing data were observed for 131 participants.

[c]Among the total sample of 2001 participants, missing data were observed for 172 participants.

[d]A total of 1569 participants who had received health care in the previous year provided answers regarding their communication with the physician.

[e]A total of 1428 participants who had searched for health information on the internet before meeting their physicians responded about whether they had discussed this information with their physician.

Patient Profiles Based on eHealth Literacy and Psychological Health Empowerment

RQ1 asked what distinct profiles of patients could be identified based on their eHealth literacy and psychological health empowerment. *K*-means cluster analysis identified a 5-cluster solution as the optimal fit. Two clusters exhibited high scores in both eHealth literacy and health empowerment. Specifically, one had mean scores for eHealth literacy and health empowerment of 4.27 (95% CI 4.21-4.34, SD 0.51) and 4.61 (95% CI 4.57-4.65, SD 0.31), respectively; another had corresponding scores of 3.90 (95% CI 3.88-3.92, SD 0.28) and 3.89 (95% CI 3.87-3.90, SD 0.23). Aligned with the HEM, these clusters exemplify a group with both high literacy and high empowerment, and can be labeled as the "effective self-manager." Therefore, they were amalgamated into a single profile, profile 1, which encompasses 996 participants (996/2001, 49.8%). This aggregated profile demonstrated mean eHealth literacy and health empowerment scores of 3.98 (95% CI 3.96-4.01, SD 0.38) and 4.05 (95% CI 4.02-4.07, SD 0.39), respectively, signifying high levels of both eHealth literacy and empowerment.

The remaining 3 clusters exhibited distinct combinations of eHealth literacy and empowerment levels, and thus we retained these unique eHealth literacy–empowerment profiles without merging. Figure 1 displays the *z* scores for eHealth literacy and health empowerment across the 4 profiles ultimately identified, including the aggregated profile 1, the effective self-manager. Table 2 lists their mean scores.

Profile 2 (408/2001, 20.4%) exhibited a comparatively low level of eHealth literacy (mean 3.25, 95% CI 3.21-3.29, SD 0.41) and an even lower level of empowerment (mean 3.09, 95% CI 3.06-3.12, SD 0.3). In the HEM, a group characterized by high literacy but low empowerment is labeled as a needlessly dependent patient. However, the current profile's eHealth literacy level, being below the average, cannot be considered high. Therefore, it is labeled as the "moderate-needs dependent patient."

XSL•FO
RenderX

Profile 3 (68/2001, 3.4%) showed the lowest levels of eHealth literacy (mean 1.67, 95% CI 1.52-1.81, SD 0.62) and health empowerment (mean 2.33, 95% CI 2.14-2.51, SD 0.76), indicating substantial deficiencies in both literacy and empowerment. According to HEM, it is labeled as the "high-needs patient."

Finally, profile 4 (529/2001, 26.4%) exhibited mean scores for eHealth literacy of 2.95 (95% CI 2.92-2.98, SD 0.41) and health empowerment of 3.95 (95% CI 3.92-3.97, SD 0.28). This suggests individuals in this profile possess a combination of low literacy with high empowerment, labeled as the "dangerous self-manager" according to the HEM.

Beyond having distinct combinations of eHealth literacy and empowerment levels, Welch ANOVA confirmed significant differences in the values of both eHealth literacy ($F_{3,287.48}$=1103.55; P<.001) and health empowerment ($F_{3,289.77}$=991.5; P<.001) among the 4 labeled profiles, with all pairwise comparisons reaching statistical significance (all P<.001, Table 2), confirming that these profiles are statistically distinct in both their combination patterns and magnitude levels of eHealth literacy and empowerment.

Figure 1. Standardized z scores for eHealth literacy and psychological health empowerment by patient profile.

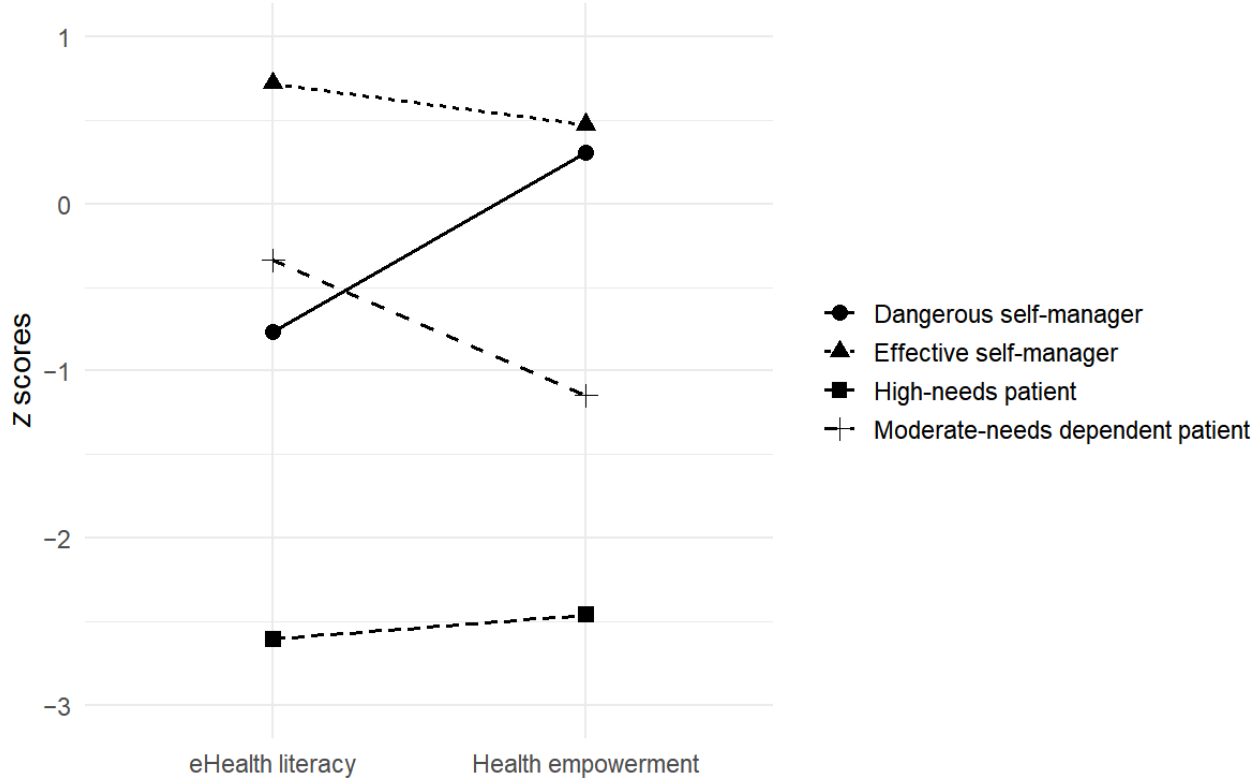

Table 2. Descriptive statistics and group comparisons for eHealth literacy and psychological health empowerment by patient profile[a].

Variables	Profile 1 (n=996): Effective self-manager	Profile 2 (n=408): Moderate-needs dependent patient	Profile 3 (n=68): High-needs patient	Profile 4 (n=529): Dangerous self-manager	Welch ANOVA: *F* test (*df*)
eHealth literacy[b], mean (SD)	3.98 (0.38)$_A$	3.25 (0.41)$_B$	1.67 (0.62)$_C$	2.95 (0.41)$_D$	1103.55 (3, 287.48)[c]
Health empowerment[b], mean (SD)	4.05 (0.39)$_A$	3.09 (0.30)$_B$	2.33 (0.76)$_C$	3.95 (0.28)$_D$	991.50 (3, 289.77)[c]

[a]Welch ANOVA and Games-Howell post hoc tests were conducted due to the violation of the assumption of homogeneity of variance.

[b]Differing subscripts (A, B, C, D) indicate significant mean difference in a row at the significance level of P<.001.

[c]P<.001.

Information Behavior Across Patient Profiles

Welch ANOVA indicated significant differences in intended information-seeking ($F_{3,289}$=62.09; P<.001; η^2=0.12) and intended disclosure ($F_{3,299.41}$=66.08; P<.001; η^2=0.09) across the 4 eHL-E Profiles (see Table 3).

Games-Howell post hoc tests provided detailed results for all pairwise comparisons. Regarding intended information-seeking, the "effective self-manager" group exhibited the highest intention (mean 4.01, 95% CI 3.96-4.06, SD 0.88), while the "high-needs patient" group had the lowest seeking intention (mean 2.13, 95% CI 1.83-2.44, SD 1.26). The "moderate-needs dependent patient" (mean 3.66, 95% CI 3.57-3.76, SD 0.98)

XSL•FO
RenderX

and the "dangerous self-manager" (mean 3.69, 95% CI 3.61-3.78, SD 1.01) groups displayed similar seeking intentions. Except for the comparison between the "moderate-needs dependent patient" and the "dangerous self-manager" groups (mean difference –0.03, 95% CI –0.20 to 0.14; *P*=.97), all other mean differences were statistically significant at *P*<.001.

Regarding intended information disclosure, similar patterns emerged. The "effective self-manager" group once again had the greatest intention (mean 3.43, 95% CI 3.36-3.50, SD 1.02) to share online health information with their physicians. Conversely, the "high-needs patient" group exhibited the lowest intention to disclose (mean 1.85, 95% CI 1.6-2.1, SD 1). The "moderate-needs dependent patient" (mean 3, 95% CI 2.91-3.10, SD 0.98) and "dangerous self-manager" (mean 3.05, 95% CI 2.95-3.14, SD 1.12) groups were comparable. Except for the comparison between the "moderate-needs dependent patient" and the "dangerous self-manager" groups (mean difference 0.04, 95% CI –0.14 to 0.22; *P*=.93), all other group comparisons yielded significant mean differences (*P*<.001).

For actual information-seeking, the "effective self-manager" again had the highest proportion of individuals who reported seeking health information regarding their health condition before seeing the physician (800/996, 80.3%). The "moderate-needs dependent patient" group had a lower rate of 66.7% (272/408), and the "dangerous self-manager" group had a similar rate of 62.6% (331/529). In contrast, the "high-needs patient" group, characterized by the lowest eHealth literacy and empowerment, had the lowest rate, with only 36.8% (25/68) reporting actual seeking behavior. The chi-square test was significant (χ^2_3=103.4; *P*<.001) with a Cramer *V* of 0.227, indicating a moderate association (see Table 4). Pairwise comparisons with Bonferroni adjustment revealed that the "effective self-manager" had significantly higher seeking rates than all other profiles, while the "high-needs patient" had significantly lower rates than all other profiles. The "moderate-needs dependent patient" and "dangerous self-manager" profiles did not differ significantly from each other.

Among those who answered "yes" to actual seeking behavior, their actual disclosure behavior was examined (n=1428). The "effective self-manager" profile had a high proportion of individuals who disclosed health information (465/800, 58.1%). The "moderate-needs dependent patient" profile had a lower rate of 43.8% (119/272), and the "dangerous self-manager" profile had a similar rate of 47.4% (157/331). Interestingly, the "high-needs patient" profile showed the highest proportion of disclosures, with 64% (16/25) of those who searched for information online choosing to disclose it to their physicians. The chi-square test was significant (χ^2_3=23.1; *P*<.001) with a Cramer *V* of 0.127, suggesting a small to moderate association. Pairwise comparisons with Bonferroni adjustment showed that the "effective self-manager" profile had significantly higher disclosure rates than both the "moderate-needs dependent patient" and "dangerous self-manager" profiles (*P*<.05). These latter two profiles did not differ significantly from each other. The "high-needs patient" profile did not differ significantly from any other profile, which is likely due to the small sample size (n=16) limiting statistical power (Table 4).

Table 3. Descriptive statistics and group comparisons of intended information-seeking and disclosure behaviors across 4 patient profiles[a].

Intended information behavior	Profile 1 (n=996): Effective self-manager, mean (SD)	Profile 2 (n=408): Moderate-needs dependent patient, mean (SD)	Profile 3 (n=68): High-needs patient, mean (SD)	Profile 4 (n=529): Dangerous self-manager, mean (SD)	Welch ANOVA: *F* test (*df*)
Intended seeking[b]	4.01 (0.88)$_A$	3.66 (0.98)$_B$	2.13 (1.26)$_C$	3.69 (1.01)$_B$	62.09 (3, 289)[c,d]
Intended disclosure[b]	3.43 (1.02)$_A$	3.00 (0.98)$_B$	1.85 (1.00)$_C$	3.05 (1.12)$_B$	66.08 (3, 299.41)[d,e]

[a]Welch ANOVA and Games-Howell post hoc tests were conducted due to the violation of the assumption of homogeneity of variance.

[b]Differing subscripts (A, B, C) indicate significant mean difference in a row at the significance level of *P*<.001.

[c]η^2=0.12.

[d]*P*<.001.

[e]η^2=0.09.

XSL•FO
RenderX

Table 4. Descriptive statistics and group comparisons of actual information-seeking and disclosure behaviors across 4 patient profiles[a].

Actual information behaviors	Profile 1 (n=996): Effective self-manager, n (%)	Profile 2 (n=408): Moderate-needs dependent patient, n (%)	Profile 3 (n=68): High-needs patient, n (%)	Profile 4 (n=529): Dangerous self-manager, n (%)	Chi-square (*df*)
Actual seeking (yes)[b]	800 (80.3)$_A$	272 (66.7)$_B$	25 (36.8)$_C$	331 (62.6)$_B$	103.4 (3)[c,d]
Actual disclosure (yes)[b,e]	465 (58.1)$_A$	119 (43.8)$_B$	16 (64.0)$_{A,B}$	157 (47.4)$_B$	23.1 (3)[d,f]

[a]Different subscripts (A, B, C) within a row indicate significant differences between profiles at $P<.05$ or lower.

[b]Represents the frequency and percentage for dichotomous variable.

[c]Cramer $V=0.227$.

[d]$P<.001$.

[e]Only the participants who answered "yes" in actual seeking are included; hence, n=1428.

[f]Cramer $V=0.127$.

The Role of PCC in Information Disclosure

We expect that controlling for patient predisposing (health empowerment) and enabling factors (eHealth literacy), PCC would be associated with patients' disclosure of internet-sourced information to physicians. To examine this relationship, we conducted regression analyses. Intended disclosure showed no significant association with PCC (β=–0.007, 95% CI –0.09 to 0.07; $P=.76$). However, actual disclosure behavior demonstrated a positive association with PCC (odds ratio 1.26, 95% CI 1.04-1.53; $P=.02$). This indicates a positive relationship between PCC and patients' disclosure of internet-sourced information to their physicians in medical encounters (see Table 5).

Two single moderator models using the Hayes PROCESS macro (model 1 with a multicategorical moderator) tested whether PCC related differently to information disclosure across the 4 patient-profile clusters. For intended disclosure (continuous outcome), the PCC × patient profile interaction was nonsignificant ($F_{3,1413}=0.24$; $P=.87$); none of the 3 contrast-specific interaction terms was significant ($|\beta|\leq0.07$, all $P\geq.5$). The results were similar for actual disclosure (binary outcome): the interaction did not reach significance ($\chi^2_3=1.7$; $P=.64$) and each individual interaction term was likewise nonsignificant (all $P\geq.33$). These results indicate that the association between PCC and information disclosure did not significantly vary across patient profiles, suggesting a consistent relationship regardless of empowerment and literacy profiles.

Table 5. Multiple linear regression and logistic regression results for intended and actual information disclosure with physicians[a].

Variables	Intended disclosure[b]		Actual disclosure[c]	
	Coefficients (95% CI)	*P* value	Odds ratio (95% CI)	*P* value
Age	0.00 (0.000 to 0.008)	.06	1.02[e] (1.01 to 1.03)	.003
Gender	0.02 (–0.09 to 0.13)	.76	0.93 (0.72 to 1.21)	.59
Education	0.10[f] (0.05 to 0.15)	<.001	1.01 (0.89 to 1.15)	.88
Health status	–0.04 (–0.11 to 0.03)	.27	1.19[g] (1.01 to 1.39)	.04
Trust in online health information	0.18[f] (0.10 to 0.26)	<.001	1.11 (0.92 to 1.34)	.30
eHealth literacy	0.31[f] (0.22 to 0.40)	<.001	1.32[g] (1.07 to 1.64)	.01
Health empowerment	0.20[f] (0.10 to 0.30)	<.001	1.05 (0.82 to 1.34)	.69
PCC[d]	–0.01 (–0.09 to 0.07)	.76	1.26[g] (1.04 to 1.53)	.02

[a]Mean, median, or mode imputation was used for missing values on age, gender, and education.

[b]$R^2=0.11$.

[c]Nagelkerke $R^2=0.04$.

[d]PCC: patient-centered communication.

[e]$P<.01$.

[f]$P<.001$.

[g]$P<.05$.

XSL•FO
RenderX

Discussion

Principal Results

This study identified 4 distinct patient profiles based on eHealth literacy and psychological health empowerment: effective self-managers, moderate-needs dependent patients, high-needs patients, and dangerous self-managers. These profiles demonstrated significantly different patterns in both online health information seeking prior to consultations and disclosure of this information to physicians. PCC was positively associated with actual information disclosure, and this association did not vary across patient profiles. Together, these findings demonstrate that eHealth literacy and empowerment create meaningful patient groups with distinct information behaviors, and PCC plays a consistent role in encouraging disclosure regardless of patient characteristics. Below, we discuss these findings in relation to existing literature and their implications for theory and practice.

First, our identification of 4 distinct patient profiles supports the HEM's core premise that patients can be meaningfully categorized based on their literacy and empowerment levels. Three of these profiles—effective self-managers, high-needs patients, and dangerous self-managers—matched HEM predictions closely [24]. However, we found an interesting deviation. The HEM predicts a "needlessly dependent patient" with high literacy but low empowerment. Instead, we identified "moderate-needs dependent patients" who had both below-average literacy and low empowerment. This finding appears consistent with other empirical applications of HEM. For instance, a recent study classified 269 participants using depression literacy and empowerment; they only identified 2 of the 4 theoretical profiles [36]. Such findings suggest that while the HEM provides a valuable conceptual framework, researchers should anticipate that patient populations may not always exhibit all 4 theorized combinations. The specific profiles that emerge may depend on factors such as the targeted literacy used and the characteristics of the population of interest. For instance, when studying populations that are relatively educated and familiar with online technologies (thus having high eHealth literacy), the theoretically posited profiles of "high-needs patients" and "needlessly dependent patients" with low literacy may not be identified due to the overall high literacy baseline of the sample.

Interestingly, our sample split nearly evenly between "balanced" profiles (effective self-managers and high-needs patients, around 53%, 1064/2001 of the total sample) and "imbalanced" profiles where literacy and empowerment diverged (moderate-needs dependent patients and dangerous self-managers, around 47%, 937/2001). This finding suggests that eHealth literacy and psychological health empowerment do not necessarily develop in parallel. In practice, digital health initiatives aiming for patient empowerment often assume patients have the eHealth literacy needed to grasp their content [84], while online interventions that focus on improving patients' literacy skills often treat empowerment as one of the primary outcomes [85]. Our findings suggest that the assumption of high eHealth literacy inherently leading to empowerment and effective patient engagement with eHealth resources may not hold true. Therefore, targeting health literacy alone may not guarantee corresponding improvements in empowerment and vice versa.

Second, we found significant behavioral differences across patient profiles. These patient profiles not just represent differences in their literacy and empowerment but also differ with their online health information seeking and disclosure patterns. One interesting finding emerged from the "high-needs patient" profile. They present a particularly intriguing information behavior pattern: while they showed the lowest information-seeking behavior, they demonstrated the highest disclosure rates when they did search. Rather than being passive or disengaged, "high-needs patients" appear highly selective in their information-seeking but deeply reliant on physician guidance. This suggests that when patients with low eHealth literacy and empowerment do seek information, they may place exceptional value on professional validation and interpretation.

The "dangerous self-manager" profile raises particular clinical concerns. These patients demonstrate moderately high information-seeking behavior but relatively low disclosure rates, creating a gap between information gathering and sharing with providers. Given their lower eHealth literacy levels, this combination of active seeking with limited disclosure could pose safety risks, as they may be making health decisions based on misinterpreted information without professional guidance.

In contrast, "effective self-managers" demonstrated the most comprehensive use of online health information, showing both high seeking behavior and high disclosure rates with physicians. This finding aligns with prior research linking high literacy and empowerment to more active health care participation and effective patient-provider collaboration [33,86].

Our analysis showed an interesting pattern regarding PCC and information disclosure. While participants' retrospective assessments of PCC quality from past health care encounters showed no association with their intended disclosure behavior in hypothetical future scenarios, PCC was significantly linked to their actual disclosure behavior in those past encounters. This suggests that patients may form intentions about sharing online health information based on factors other than their previous communication experiences. However, the quality of physician communication during actual encounters does influence whether patients disclose their online findings in real time. Therefore, disclosure behavior appears more responsive to the immediate communication context rather than retrospective evaluations of past clinical experiences. This finding aligns with empirical studies showing that physicians' communicative behavior during encounters affects whether patients share their online findings [21,66,87].

Furthermore, PCC demonstrated a statistically significant positive association with information disclosure behavior that was fairly consistent across all 4 patient profiles. This finding is particularly notable given the substantial information on behavioral differences we observed. This universal responsiveness suggests that PCC can promote engagement across diverse patient profiles, regardless of their health literacy or empowerment levels.

XSL•FO
RenderX

Theoretical Implications

This study makes theoretical contributions to the understanding of patient eHealth literacy, empowerment, and the HEM. First, our findings provide empirical support for HEM. Our identification of patient profiles aligns closely with the model's proposed combinations of high and low empowerment and health literacy. These profiles exhibited distinct information behaviors, from comprehensive information seeking and disclosure among "effective self-managers" to limited engagement among "high-needs patients" and "dangerous self-managers." While previous HEM studies are primarily based on Western health contexts, our study demonstrates the model's validity in Chinese health care settings. Additionally, our study extends HEM by integrating physician-patient communication as a contextual factor, which was not originally included in the model. Specifically, we demonstrated that PCC encourages patients to disclose online health information to their physicians, regardless of their empowerment or literacy combination. This finding indicates that future applications of HEM should incorporate interpersonal communication aspects as relevant contextual variables when examining patient health behaviors in clinical settings.

Second, our work advances the eHealth literacy and patient empowerment literature by demonstrating that both cognitive capabilities (eHealth literacy) and motivational agency (psychological health empowerment) are necessary for effective use of online health information. These constructs are often studied separately, yet our findings suggest they operate together. Future research should integrate both when modeling patient information behaviors to develop a more complete understanding of how patients access, evaluate, and act on health information.

Practical Implications

Our findings offer practical insights into health education and clinical care. The identification of distinct eHL-E Profiles highlights the potential for tailored communication strategies for each patient group. Effective self-managers, with their high literacy and empowerment, can benefit from more autonomous decision-making opportunities and acknowledgment to continue engaging with reliable online health information. High-needs patients, in contrast, require interventions that enhance both their literacy skills and psychological empowerment; health care providers should therefore offer these patients greater support in navigating health information rather than solely expecting autonomous or shared decision-making. For dangerous self-managers, health programs should focus primarily on building literacy and information evaluation skills; providers should guide them in assessing online health information and encourage them to share and discuss their findings with physicians to prevent potential misuses.

In the digital age, patients commonly arrive at medical consultations with information gathered online [5]. Yet, our study, alongside works by others [88,89], highlights a concerning trend: patients often withhold this online-acquired information from their physicians. This reluctance to share online findings prevents opportunities to correct misconceptions from online searches and diminishes chances to better understand patients' views and perceptions about their health. To encourage patients' sharing, we recommend that physicians adopt PCC strategies. These include acknowledging patients' efforts to search for information online as preparation for their visits rather than devaluing this information [90,91], asking open-ended questions about what patients have learned online, and encouraging patients to discuss any concerns about information they have found [15]. By fostering these open dialogues, patients can benefit from both the wealth of online health information and the expertise of medical professionals.

The study findings also have implications for digital health initiatives on eHealth literacy or empowerment. With the proliferation of digital tools and artificial intelligence (AI) technologies in health promotion [92,93], training programs should address both skill development and confidence-building. Specifically, digital health platforms can implement strategies such as differentiated information presentation based on users' health literacy levels, gradually building eHealth literacy through scaffolded learning. Additionally, empowerment-focused features, such as personalized goal setting and progress tracking, can be integrated into online health applications to encourage patient participation.

Limitations

Our study has several limitations. First, there is no established measure for estimating patient information disclosure behavior. Survey studies on this topic typically devised their own measures based on their specific focus [49]. The measurement of actual behavior, including information seeking and disclosure, relied on participants' recall of previous medical visits with physicians. This could introduce recall bias to their responses.

Second, our sample was skewed toward young, highly educated individuals with college degrees or higher. This is also reflected in the overrepresentation of the effective self-manager group and underrepresentation of the high-needs patient group. The small size of the high-needs patient group may bias our results and limit their generalizability to those with low eHealth literacy and empowerment. The unbalanced sample may stem from our survey's online distribution, which is more accessible to digital platform users. Future research should apply more inclusive sampling for those less digitally skilled. Additionally, our sample has a significant gender imbalance, with females representing approximately twice as many participants as males. This overrepresentation of female participants limits the generalizability of our findings across gender groups. Future studies should use stratified sampling methods to ensure more balanced gender representation.

Third, there is ongoing debate in the literature regarding health literacy measurement, which can be subjective (self-rated) or objective (performance-based) [94]. Our study used the eHEALS, which relies on self-reported perceptions of competence rather than objective assessment of actual skills. While eHEALS is widely validated and predictive of health behaviors and outcomes, we acknowledge that self-reported measures may not fully capture individuals' actual eHealth literacy performance. Besides, there is growing recognition that traditional eHealth literacy measures like eHEALS may not fully address contemporary digital health challenges, and some

XSL•FO
RenderX

advocate for a broader digital health literacy concept that encompasses the competence of using digital health tools and additional measurements to incorporate AI literacy. Future studies should consider including these newer measurements to capture the full spectrum of digital health competencies.

Fourth, our data were collected between September and November 2022, prior to the public release of ChatGPT (November 30, 2022) and before AI chatbots became widely available for health information seeking. Therefore, AI-generated health information use was not prevalent among participants and was not included in our study design. This represents a limitation, as the current digital health information landscape has evolved rapidly since our data collection. Future research should examine how AI literacy and the use of AI-generated health information shape patient-physician communication and information behavior patterns.

Conclusions

This study extends existing literature by moving beyond online health information seeking to examine patients' disclosure of online findings to physicians. Unlike prior research that examined eHealth literacy and psychological health empowerment separately, this study integrated these constructs to identify meaningful patient profiles with distinct information behavior patterns. Findings indicate that active health information seeking and disclosure require both adequate eHealth literacy skills and psychological health empowerment. Furthermore, PCC facilitates disclosure regardless of patient profile. For practice, physicians should adopt a PCC that acknowledges patients' online research efforts and actively supports their eHealth literacy and empowerment. This can create a clinical environment where patients feel comfortable sharing their online findings, ultimately contributing to safer use of online health information and stronger patient-physician relationships.

Acknowledgments

During the preparation of this manuscript, ChatGPT-4 (OpenAI) was used to identify grammatical errors and suggest improvements for sentence readability. No original concepts, data analysis, or interpretations were generated by the AI; all intellectual content is solely that of the authors. The authors reviewed all suggested changes and take full responsibility for the published paper.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Data Availability

The dataset used in this study is available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: QL, PJS
Data curation: WJ, AC
Formal analysis: QL
Writing – original draft: QL
Writing – review & editing: QL, PJS, AC

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete measurement scales.
[DOCX File , 22 KB - jmir_v28i1e78836_app1.docx]

References

1. Wang X, Shi J, Kong H. Online health information seeking: a review and meta-analysis. Health Commun 2021;36(10):1163-1175. [doi: 10.1080/10410236.2020.1748829] [Medline: 32290679]
2. Swoboda CM, Van Hulle JM, McAlearney A, Huerta T. Odds of talking to healthcare providers as the initial source of healthcare information: updated cross-sectional results from the Health Information National Trends Survey (HINTS). BMC Fam Pract 2018;19(1):146 [FREE Full text] [doi: 10.1186/s12875-018-0805-7] [Medline: 30157770]
3. Wang J, Ashvetiya T, Quaye E, Parakh K, Martin S. Online health searches and their perceived effects on patients and patient-clinician relationships: a systematic review. Am J Med 2018;131(10):1250.e1-1250.e10. [doi: 10.1016/j.amjmed.2018.04.019] [Medline: 29730361]
4. Tan SSL, Goonawardene N. Internet health information seeking and the patient-physician relationship: a systematic review. J Med Internet Res 2017;19(1):e9 [FREE Full text] [doi: 10.2196/jmir.5729] [Medline: 28104579]

XSL•FO
RenderX

5. Luo A, Qin L, Yuan Y, Yang Z, Liu F, Huang P, et al. The effect of online health information seeking on physician-patient relationships: systematic review. J Med Internet Res 2022;24(2):e23354 [FREE Full text] [doi: 10.2196/23354] [Medline: 35142620]
6. Linn AJ, van Weert JCM, Gebeyehu BG, Sanders R, Diviani N, Smit EG, et al. Patients'online information-seeking behavior throughout treatment: the impact on medication beliefs and medication adherence. Health Commun 2019;34(12):1461-1468 [FREE Full text] [doi: 10.1080/10410236.2018.1500430] [Medline: 30052088]
7. Te Poel F, Baumgartner S, Hartmann T, Tanis M. The curious case of cyberchondria: a longitudinal study on the reciprocal relationship between health anxiety and online health information seeking. J Anxiety Disord 2016;43:32-40. [doi: 10.1016/j.janxdis.2016.07.009] [Medline: 27497667]
8. Walter N, Brooks J, Saucier C, Suresh S. Evaluating the impact of attempts to correct health misinformation on social media: a meta-analysis. Health Commun 2021;36(13):1776-1784. [doi: 10.1080/10410236.2020.1794553] [Medline: 32762260]
9. Lu J, Xiao Y. Heuristic information processing as a mediating factor in the process of exposure to COVID-19 vaccine information and misinformation sharing on social media. Health Commun 2024;39(12):2779-2792. [doi: 10.1080/10410236.2023.2288373] [Medline: 38016931]
10. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021;23(1):e17187 [FREE Full text] [doi: 10.2196/17187] [Medline: 33470931]
11. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. Annu Rev Public Health 2020;41:433-451 [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094127] [Medline: 31874069]
12. Chen YY, Li C, Liang J, Tsai C. Health information obtained from the internet and changes in medical decision making: questionnaire development and cross-sectional survey. J Med Internet Res 2018;20(2):e47 [FREE Full text] [doi: 10.2196/jmir.9370] [Medline: 29434017]
13. Bowes P, Stevenson F, Ahluwalia S, Murray E. 'I need her to be a doctor': patients' experiences of presenting health information from the internet in GP consultations. Br J Gen Pract 2012;62(604):e732-e738 [FREE Full text] [doi: 10.3399/bjgp12X658250] [Medline: 23211176]
14. Lu Q, Schulz P. Physician perspectives on internet-informed patients: systematic review. J Med Internet Res 2024;26:e47620 [FREE Full text] [doi: 10.2196/47620] [Medline: 38842920]
15. Shachar L. "You become a slightly better doctor": doctors adopting integrated medical expertise through interactions with e-patients. Soc Sci Med 2022;305:115038. [doi: 10.1016/j.socscimed.2022.115038] [Medline: 35598443]
16. Schick TS, Höllerl L, Biedermann T, Zink A, Ziehfreund S. Impact of digital media on the patient journey and patient-physician relationship among dermatologists and adult patients with skin diseases: qualitative interview study. J Med Internet Res 2023;25:e44129 [FREE Full text] [doi: 10.2196/44129] [Medline: 37738078]
17. Győrffy Z, Radó N, Mesko B. Digitally engaged physicians about the digital health transition. PLoS One 2020;15(9):e0238658 [FREE Full text] [doi: 10.1371/journal.pone.0238658] [Medline: 32986733]
18. Shen MJ, Dyson R, D'Agostino TA, Ostroff J, Dickler M, Heerdt A, et al. Cancer-related internet information communication between oncologists and patients with breast cancer: a qualitative study. Psychooncology 2015;24(11):1439-1447 [FREE Full text] [doi: 10.1002/pon.3752] [Medline: 25631285]
19. Kim J, Kim S. Physicians' perception of the effects of Internet health information on the doctor-patient relationship. Inform Health Soc Care 2009;34(3):136-148. [doi: 10.1080/17538150903102422] [Medline: 19670004]
20. George GC, Iwuanyanwu E, Buford A, Piha-Paul S, Subbiah V, Fu S, et al. Cancer-related internet use and its association with patient decision making and trust in physicians among patients in an early drug development clinic: a questionnaire-based cross-sectional observational study. J Med Internet Res 2019;21(3):e10348 [FREE Full text] [doi: 10.2196/10348] [Medline: 30869638]
21. Linn AJ, van Weert JCM, Bylund C. Are they important? Patients' communication barriers to discussing online health information during consultations. Health Commun 2024;39(5):945-950 [FREE Full text] [doi: 10.1080/10410236.2023.2197311] [Medline: 37072688]
22. Chiu YC. Probing, impelling, but not offending doctors: the role of the internet as an information source for patients' interactions with doctors. Qual Health Res 2011;21(12):1658-1666. [doi: 10.1177/1049732311417455] [Medline: 21799204]
23. Jia X, Pang Y, Liu L. Online health information seeking behavior: a systematic review. Healthcare (Basel) 2021;9(12):1740 [FREE Full text] [doi: 10.3390/healthcare9121740] [Medline: 34946466]
24. Schulz PJ, Nakamoto K. Health literacy and patient empowerment in health communication: the importance of separating conjoined twins. Patient Educ Couns 2013;90(1):4-11. [doi: 10.1016/j.pec.2012.09.006] [Medline: 23063359]
25. Street R. Active patients as powerful communicators. In: Robinson P, Giles W, editors. The New Handbook of Language and Social Psychology. New Jersey: Wiley; 2001:541-560.
26. Schulz PJ. The concept of health literacy. In: Zeyer A, Kyburz-Graber R, editors. Science | Environment | Health: Towards a Renewed Pedagogy for Science Education. Dordrecht: Springer; 2012:69-84.
27. Menon ST. Toward a model of psychological health empowerment: implications for health care in multicultural communities. Nurse Educ Today 2002;22(1):28-39; discussion 40. [doi: 10.1054/nedt.2001.0721] [Medline: 11886225]

XSL•FO
RenderX

28. Schulz PJ, Nakamoto K. The perils of misinformation: when health literacy goes awry. Nat Rev Nephrol 2022;18(3):135-136 [FREE Full text] [doi: 10.1038/s41581-021-00534-z] [Medline: 34997231]
29. Diviani N, van den Putte B, Giani S, van Weert JC. Low health literacy and evaluation of online health information: a systematic review of the literature. J Med Internet Res 2015;17(5):e112 [FREE Full text] [doi: 10.2196/jmir.4018] [Medline: 25953147]
30. Schulz PJ, Nakamoto K. Worse Than Ignorance: The Challenge of Health Misinformation. Cambridge: Cambridge University Press; 2024.
31. Náfrádi L, Nakamoto K, Csabai M, Papp-Zipernovszky O, Schulz PJ. An empirical test of the health empowerment model: does patient empowerment moderate the effect of health literacy on health status? Patient Educ Couns 2018;101(3):511-517. [doi: 10.1016/j.pec.2017.09.004] [Medline: 28899712]
32. Schulz PJ, Crosignani F, Petrocchi S. Critical test of the beneficial consequences of lifting the ban on direct-to-consumer advertising for prescription drugs in Italy: experimental exposure and questionnaire study. J Med Internet Res 2023;25:e40616 [FREE Full text] [doi: 10.2196/40616] [Medline: 37459159]
33. Sak G, Rothenfluh F, Schulz P. Assessing the predictive power of psychological empowerment and health literacy for older patients' participation in health care: a cross-sectional population-based study. BMC Geriatr 2017;17(1):59 [FREE Full text] [doi: 10.1186/s12877-017-0448-x] [Medline: 28219334]
34. Camerini L, Schulz P, Nakamoto K. Differential effects of health knowledge and health empowerment over patients' self-management and health outcomes: a cross-sectional evaluation. Patient Educ Couns 2012;89(2):337-344. [doi: 10.1016/j.pec.2012.08.005] [Medline: 22959333]
35. Wang RH, Hsu H, Lee Y, Shin S, Lin KD, An L. Patient empowerment interacts with health literacy to associate with subsequent self-management behaviors in patients with type 2 diabetes: a prospective study in Taiwan. Patient Educ Couns 2016;99(10):1626-1631. [doi: 10.1016/j.pec.2016.04.001] [Medline: 27083406]
36. Bizzotto N, de Bruijn GJ, Schulz P. Clusters of patient empowerment and mental health literacy differentiate professional help-seeking attitudes in online mental health communities users. Health Expect 2025;28(1):e70153 [FREE Full text] [doi: 10.1111/hex.70153] [Medline: 39815681]
37. Jiao W, Chang A, Ho M, Lu Q, Liu M, Schulz P. Predicting and empowering health for Generation Z by comparing health information seeking and digital health literacy: cross-sectional questionnaire study. J Med Internet Res 2023;25:e47595 [FREE Full text] [doi: 10.2196/47595] [Medline: 37902832]
38. Norman C. eHealth Literacy 2.0: problems and opportunities with an evolving concept. J Med Internet Res 2011;13(4):e125 [FREE Full text] [doi: 10.2196/jmir.2035] [Medline: 22193243]
39. Norman CD, Skinner H. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
40. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012;14(1):e19 [FREE Full text] [doi: 10.2196/jmir.1619] [Medline: 22357448]
41. Moretti V, Brunelli L, Conte A, Valdi G, Guelfi M, Masoni M, et al. A web tool to help counter the spread of misinformation and fake news: pre-post study among medical students to increase digital health literacy. JMIR Med Educ 2023;9:e38377 [FREE Full text] [doi: 10.2196/38377] [Medline: 36996010]
42. Hu X, Bell R, Kravitz R, Orrange S. The prepared patient: information seeking of online support group members before their medical appointments. J Health Commun 2012;17(8):960-978. [doi: 10.1080/10810730.2011.650828] [Medline: 22574697]
43. Tennant B, Stellefson M, Dodd V, Chaney B, Chaney D, Paige S, et al. eHealth literacy and Web 2.0 health information seeking behaviors among baby boomers and older adults. J Med Internet Res 2015;17(3):e70 [FREE Full text] [doi: 10.2196/jmir.3992] [Medline: 25783036]
44. Lee HY, Jin S, Henning-Smith C, Lee J, Lee J. Role of health literacy in health-related information-seeking behavior online: cross-sectional study. J Med Internet Res 2021;23(1):e14088 [FREE Full text] [doi: 10.2196/14088] [Medline: 33502332]
45. Melhem SJ, Nabhani-Gebara S, Kayyali R. Digital trends, digital literacy, and e-health engagement predictors of breast and colorectal cancer survivors: a population-based cross-sectional survey. Int J Environ Res Public Health 2023;20(2):1472. [doi: 10.3390/ijerph20021472]
46. Sundell E, Wångdahl J, Grauman Å. Health literacy and digital health information-seeking behavior: a cross-sectional study among highly educated Swedes. BMC Public Health 2022;22(1):2278 [FREE Full text] [doi: 10.1186/s12889-022-14751-z] [Medline: 36471284]
47. Vamos S, Vine M, Gordon I, Alaimo D. The role of health literacy in patient-physician communication: a scoping review. Health Behav and Policy Rev 2024;11(4):1624-1654. [doi: 10.14485/hbpr.11.4.2]
48. Damian AJ, Gallo JJ. Promoting health literacy during the COVID-19 pandemic: a call to action for healthcare professionals. HKS Misinfo Review 2020;1(3). [doi: 10.37016/mr-2020-027]
49. Li S, Wang K. Sharing online health information with physicians: understanding the associations among patient characteristics, directness of sharing, and physician-patient relationship. Front Psychol 2022;13:839723 [FREE Full text] [doi: 10.3389/fpsyg.2022.839723] [Medline: 35432074]

XSL•FO
RenderX

50. Crondahl K, Karlsson LE. The nexus between health literacy and empowerment: a scoping review. Sage Open 2016;6(2). [doi: 10.1177/2158244016646410]
51. Salmon P, Hall G. Patient empowerment and control: a psychological discourse in the service of medicine. Soc Sci Med 2003;57:1969-1980. [doi: 10.1016/s0277-9536(03)00063-7]
52. Broom A. Virtually He@lthy: the impact of internet use on disease experience and the doctor-patient relationship. Qual Health Res 2005;15(3):325-345. [doi: 10.1177/1049732304272916] [Medline: 15761103]
53. Kivits J. Informed patients and the internet: a mediated context for consultations with health professionals. J Health Psychol 2006;11(2):269-282. [doi: 10.1177/1359105306061186] [Medline: 16464924]
54. Sillence E, Briggs P, Harris PR, Fishwick L. How do patients evaluate and make use of online health information? Soc Sci Med 2007;64(9):1853-1862. [doi: 10.1016/j.socscimed.2007.01.012] [Medline: 17328998]
55. Jiang F, Liu Y, Hu J, Chen X. Understanding health empowerment from the perspective of information processing: questionnaire study. J Med Internet Res 2022;24(1):e27178 [FREE Full text] [doi: 10.2196/27178] [Medline: 35014957]
56. Stovell D, Morrison A, Panayiotou M, Hutton P. Shared treatment decision-making and empowerment-related outcomes in psychosis: systematic review and meta-analysis. Br J Psychiatry 2016;209(1):23-28. [doi: 10.1192/bjp.bp.114.158931] [Medline: 27198483]
57. Hickmann E, Richter P, Schlieter H. All together now – patient engagement, patient empowerment, and associated terms in personal healthcare. BMC Health Serv Res 2022;22:1116 [FREE Full text] [doi: 10.1186/s12913-022-08501-5] [Medline: 36056354]
58. Thiruvengadam SK, Ashvetiya T, Stone N, Blumenthal R, Martin S. Shared decision-making and patient empowerment in preventive cardiology. Curr Cardiol Rep 2016;18(5):49. [doi: 10.1007/s11886-016-0729-6] [Medline: 27098670]
59. Aujoulat I, Marcolongo R, Bonadiman L, Deccache A. Reconsidering patient empowerment in chronic illness: a critique of models of self-efficacy and bodily control. Soc Sci Med 2008;66(5):1228-1239. [doi: 10.1016/j.socscimed.2007.11.034] [Medline: 18155338]
60. Papp-Zipernovszky O, Horváth MD, Schulz PJ, Csabai M. Generation gaps in digital health literacy and their impact on health information seeking behavior and health empowerment in Hungary. Front Public Health 2021;9:635943 [FREE Full text] [doi: 10.3389/fpubh.2021.635943] [Medline: 34055714]
61. Schrank B, Sibitz I, Unger A, Amering M. How patients with schizophrenia use the internet: qualitative study. J Med Internet Res 2010;12(5):e70 [FREE Full text] [doi: 10.2196/jmir.1550] [Medline: 21169176]
62. Epstein RM, Street RL. Patient-Centered Communication in Cancer Care: Promoting Healing and Reducing Suffering. NIH Publication No. 07-6225. Bethesda, MD: National Cancer Institute; 2007.
63. Wanzer MB, Booth-Butterfield M, Gruber K. Perceptions of health care providers' communication: relationships between patient-centered communication and satisfaction. Health Commun 2004;16(3):363-383. [doi: 10.1207/S15327027HC1603_6] [Medline: 15265756]
64. Siebinga VY, Driever E, Stiggelbout A, Brand P. Shared decision making, patient-centered communication and patient satisfaction – a cross-sectional analysis. Patient Educ Couns 2022;105(7):2145-2150. [doi: 10.1016/j.pec.2022.03.012] [Medline: 35337712]
65. Brand PLP, Stiggelbout AM. Effective follow-up consultations: the importance of patient-centered communication and shared decision making. Paediatr Respir Rev 2013;14(4):224-228. [doi: 10.1016/j.prrv.2013.01.002] [Medline: 23434177]
66. Silver MP. Patient perspectives on online health information and communication with doctors: a qualitative study of patients 50 years old and over. J Med Internet Res 2015;17(1):e19 [FREE Full text] [doi: 10.2196/jmir.3588] [Medline: 25586865]
67. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet e-Surveys (CHERRIES). J Med Internet Res 2004;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
68. China's internet users total 1.079 bln with vibrant digital infrastructure, internet app growth. The State Council, The People's Republic of China. 2023 Aug 28. URL: http://english.www.gov.cn/archive/statistics/202308/28/content_WS64ec5126c6d0868f4e8dee21.html [accessed 2026-01-27]
69. Norman CD, Skinner H. eHEALS: the eHealth literacy scale. J Med Internet Res 2006;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
70. Chung SY, Park B, Nahm E. The Korean eHealth Literacy Scale (K-eHEALS): reliability and validity testing in younger adults recruited online. J Med Internet Res 2018;20(4):e138 [FREE Full text] [doi: 10.2196/jmir.8759] [Medline: 29678800]
71. Brørs G, Wentzel-Larsen T, Dalen H, Hansen T, Norman C, Wahl A, CONCARD Investigators. Psychometric properties of the Norwegian version of the Electronic Health Literacy Scale (eHEALS) among patients after percutaneous coronary intervention: cross-sectional validation study. J Med Internet Res 2020;22(7):e17312 [FREE Full text] [doi: 10.2196/17312] [Medline: 32720900]
72. Chang A, Schulz PJ. The measurements and an elaborated understanding of Chinese eHealth Literacy (C-eHEALS) in chronic patients in China. Int J Environ Res Public Health 2018;15(7):1553 [FREE Full text] [doi: 10.3390/ijerph15071553] [Medline: 30041420]
73. Lwin MO, Panchapakesan C, Sheldenkar A, Calvert G, Lim L, Lu J. Determinants of eHealth literacy among adults in China. J Health Commun 2020;25(5):385-393. [doi: 10.1080/10810730.2020.1776422] [Medline: 32552607]

XSL•FO
RenderX

74. Cui GH, Li S, Yin Y, Chen L, Li J, Liang F, et al. The relationship among social capital, eHealth literacy and health behaviours in Chinese elderly people: a cross-sectional study. BMC Public Health 2021;21(1):45 [FREE Full text] [doi: 10.1186/s12889-020-10037-4] [Medline: 33407275]
75. Oh HJ, Lee B. The effect of computer-mediated social support in online communities on patient empowerment and doctor-patient communication. Health Commun 2012;27(1):30-41. [doi: 10.1080/10410236.2011.567449] [Medline: 21797714]
76. Jiang S, Street RL. The effects of patient-centered communication, social capital, and internet use on patient empowerment: a cross-sectional study in China. Glob Health Promot 2019;26(4):33-43. [doi: 10.1177/1757975917749197] [Medline: 29521571]
77. Jiang S, Liu PL, Chia CWJ. Can online patient-provider communication improve emotional well-being? Examining the roles of social presence and patient empowerment. Cyberpsychol Behav Soc Netw 2023;26(5):366-370. [doi: 10.1089/cyber.2022.0289] [Medline: 36800227]
78. Jiang S, Ng AYK, Ngien A. The effects of social media information discussion, perceived information overload and patient empowerment in influencing HPV Knowledge. J Health Commun 2022;27(6):407-415. [doi: 10.1080/10810730.2022.2115591] [Medline: 35996841]
79. Epstein RM, Street RL. The values and value of patient-centered care. Ann Fam Med 2011;9(2):100-103 [FREE Full text] [doi: 10.1370/afm.1239] [Medline: 21403134]
80. Trivedi N, Moser RP, Breslau ES, Chou WYS. Predictors of patient-centered communication among U.S. adults: analysis of the 2017-2018 Health Information National Trends Survey (HINTS). J Health Commun 2021;26(1):57-64. [doi: 10.1080/10810730.2021.1878400] [Medline: 33648425]
81. McKeown L, Hong YA, Kreps GL, Xue H. Trends and differences in perceptions of patient-centered communication among adults in the US. Patient Educ Couns 2023;106:128-134. [doi: 10.1016/j.pec.2022.10.010] [Medline: 36270858]
82. Liu PL, Jiang S. Patient-centered communication mediates the relationship between health information acquisition and patient trust in physicians: a five-year comparison in China. Health Commun 2021;36(2):207-216. [doi: 10.1080/10410236.2019.1673948] [Medline: 31617412]
83. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: an R package for determining the relevant number of clusters in a data set. J Stat Soft 2014;61(6):1-36. [doi: 10.18637/jss.v061.i06]
84. Cheng C, Beauchamp A, Elsworth GR, Osborne RH. Applying the electronic health literacy lens: systematic review of electronic health interventions targeted at socially disadvantaged groups. J Med Internet Res 2020;22(8):e18476 [FREE Full text] [doi: 10.2196/18476] [Medline: 32788144]
85. Samoocha D, Bruinvels DJ, Elbers NA, Anema JR, van der Beek AJ. Effectiveness of web-based interventions on patient empowerment: a systematic review and meta-analysis. J Med Internet Res 2010;12(2):e23 [FREE Full text] [doi: 10.2196/jmir.1286] [Medline: 20581001]
86. Çiftci N, Yıldız M, Yıldırım Ö. The effect of health literacy and health empowerment on quality of life in the elderly. Psychogeriatrics 2023;23(4):609-620. [doi: 10.1111/psyg.12969] [Medline: 37186342]
87. Stevenson FA, Kerr C, Murray E, Nazareth I. Information from the internet and the doctor-patient relationship: the patient perspective—a qualitative study. BMC Fam Pract 2007;8:47 [FREE Full text] [doi: 10.1186/1471-2296-8-47] [Medline: 17705836]
88. Wong DKK, Cheung MK. Online health information seeking and ehealth literacy among patients attending a primary care clinic in Hong Kong: a cross-sectional survey. J Med Internet Res 2019;21(3):e10831 [FREE Full text] [doi: 10.2196/10831] [Medline: 30916666]
89. Waring ME, McManus DD, Amante DJ, Darling CE, Kiefe CI. Online health information seeking by adults hospitalized for acute coronary syndromes: who looks for information, and who discusses it with healthcare providers? Patient Educ Couns 2018;101(11):1973-1981 [FREE Full text] [doi: 10.1016/j.pec.2018.06.016] [Medline: 30305253]
90. Fujioka Y, Stewart E. How do physicians discuss e-health with patients? The relationship of physicians' e-health beliefs to physician mediation styles. Health Commun 2013;28(4):317-328. [doi: 10.1080/10410236.2012.682971] [Medline: 22716050]
91. Walker L, Sillence E. Examining healthcare professionals' communication around decision-making with internet-informed patients. Health Commun 2024;39(6):1094-1101. [doi: 10.1080/10410236.2023.2204580] [Medline: 37150845]
92. Guo W, Chen Y. Investigating whether AI will replace human physicians and understanding the interplay of the source of consultation, health-related stigma, and explanations of diagnoses on patients' evaluations of medical consultations: randomized factorial experiment. J Med Internet Res 2025;27:e66760 [FREE Full text] [doi: 10.2196/66760] [Medline: 40053785]
93. Yun HS, Bickmore T. Online health information-seeking in the era of large language models: cross-sectional web-based survey study. J Med Internet Res 2025;27:e68560 [FREE Full text] [doi: 10.2196/68560] [Medline: 40163112]
94. Goodman MS, Griffey RT, Carpenter CR, Blanchard M, Kaphingst KA. Do subjective measures improve the ability to identify limited health literacy in a clinical setting? J Am Board Fam Med 2015;28(5):584-594 [FREE Full text] [doi: 10.3122/jabfm.2015.05.150037] [Medline: 26355130]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
CHERRIES: Checklist for Reporting Results of Internet e-Surveys
eHEALS: eHealth Literacy Scale
eHL-E Profiles: eHealth Literacy and Empowerment Profiles
HEM: Health Empowerment Model
HINTS-China: Health Information National Trends Survey in China
PCC: patient-centered communication
PHES: Psychological Health Empowerment Scale
RQ: research question

Edited by S Brini; submitted 12.Jul.2025; peer-reviewed by L Shang, T Vagg; comments to author 09.Sep.2025; revised version received 09.Dec.2025; accepted 10.Dec.2025; published 18.Feb.2026.

Please cite as:
Lu Q, Jiao W, Chang A, Schulz PJ
Seeking But Not Discussing Online Health Information With Physicians: Cross-Sectional Survey Study of eHealth Literacy–Empowerment Profiles and Patient-Centered Communication
J Med Internet Res 2026;28:e78836
URL: https://www.jmir.org/2026/1/e78836
doi:10.2196/78836
PMID:

©Qianfeng Lu, Wen Jiao, Angela Chang, Peter Johannes Schulz. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact of GPT-4–Generated Discharge Letters on Patients' Medical Comprehension: Prospective Crossover Study

Friederike Holderried[1*], MD; Alessandra Sonanini[1*], MD; Christian Stegemann–Philipps[1], MSc; Anne Herrmann–Werner[1], MD; Philipp Spitzer[2], MD; Martina Guthoff[3], MD; Nils Heyne[4], MD; Konstantin Sering[1], PhD; Martin Holderried[5], MD; Felix Eisinger[4], MD

[1]Tübingen Institute for Medical Education (TIME), University of Tübingen, Tübingen, Germany

[2]Department of Psychiatry and Psychotherapy, University of Erlangen-Nürnberg, Erlangen, Germany

[3]Nephrological Center Villingen-Schwenningen, Villingen-Schwenningen, Germany

[4]Department of Diabetology, Endocrinology, Nephrology, University of Tübingen, Tübingen, Germany

[5]Karlsruhe Municipal Hospital, Karlsruhe, Germany

[*]these authors contributed equally

Corresponding Author:
Felix Eisinger, MD
Department of Diabetology, Endocrinology, Nephrology
University of Tübingen
Ottfried-Müller-Straße 10
Tübingen, 72076
Germany
Phone: 49 07071 29 82712
Email: Felix.Eisinger@med.uni-tuebingen.de

Abstract

Background: Patients often struggle to understand standard hospital discharge letters, increasing the risk of medication errors and misunderstandings. According to cognitive load theory (CLT), complex, information-dense texts can overload working memory and impair comprehension. Artificial intelligence tools that generate patient-centered versions could help reduce extraneous cognitive load and bridge this gap. However, evidence for their effectiveness remains limited.

Objective: This study aimed to evaluate whether GPT-4 (OpenAI)–generated patient-centered letters improve standardized patients' retention and understanding of safety-relevant medical information compared with standard hospital discharge letters, and to explore potential effects on cognitive load as described by CLT.

Methods: In this prospective, randomized, crossover study, 48 trained standardized patients received a conventional discharge letter for an assigned disease (out of 3) and its matching GPT-4–generated patient-centered letter. Participants read one version first, identified predefined safety-relevant "learning objectives," and then repeated the task with the alternate version. The primary outcome was the proportion of learning objectives fully, partially, or not reported. In a secondary analysis, results were stratified by content field (Medication, Organization, Prevention of Complications, Lifestyle/Disease Management) and Bloom taxonomy level ("Remember," "Understand").

Results: The letter type significantly influenced comprehension (odds ratio [OR] 1.74, 95% CI 1.45-2.08; P<.001). Patient letters, compared with discharge letters, led to higher rates of fully (490/1073, 45.7% vs 413/1073, 38.5%) or partially (322/1073, 30% vs 287/1073, 26.7%) stated learning objectives and fewer omissions (261/1073, 24.3% vs 373/1073, 34.8%). Participants performed better on "Remember" than on "Understand" learning objectives, regardless of letter type (OR 3.33, 95% CI 1.96-5.88; P<.001). Compared with standard hospital discharge letters, patient letters consistently improved results at both cognitive levels ("Remember": 278/545, 51% vs 242/545, 44.4%; "Understand": 212/528, 40.2% vs 171/528, 32.4% fully stated). The effect of patient letters varied by content field (P<.001). The greatest improvements were observed for "Medication" (170/254, 66.9% vs 129/254, 50.8% fully stated) and "Organization" (78/158, 49.4% vs 62/158, 39.2% fully stated). Improvements in the content field "Prevention of Complications" were modest, and those for "Lifestyle/Disease Management" were even smaller across all conditions. A total of 24.3% (261/1073) of key information remained unrecognized.

Conclusions: In this explanatory study, GPT-4–generated patient letters improved comprehension of safety-relevant discharge information among standardized patients, particularly regarding medication and organizational aspects. However, they were less

XSL•FO
RenderX

effective in supporting higher-order understanding, such as risk prevention or lifestyle management. These hypothesis-driven findings can be interpreted within a CLT framework and may motivate prospective evaluation of multimodal, iterative supports.

(J Med Internet Res 2026;28:e81243) doi:10.2196/81243

KEYWORDS

artificial intelligence; GPT-4; health care communication; large language models; patient education; patient letters; patient safety

Introduction

The transition from inpatient to outpatient care represents a critical juncture for information exchange and patient safety, often marked by communication errors and incomplete understanding [1]. Nearly half (49%) of patients experience at least 1 postdischarge medical error, most frequently concerning medication management, diagnostic workup, or test result follow-up [2]. Effective patient understanding during this phase is crucial, as misinterpretation or omission of medical instructions can directly impact adherence, follow-up care, and safety outcomes [3,4]. Studies estimate that up to 78% of patients misunderstand key elements of their hospital care and discharge instructions [5], and approximately half (50.8%) of patients receiving polypharmacy misinterpret at least 1 dosage instruction [6].

Patients' comprehension of discharge information is influenced by multiple factors, including health literacy, educational background, prior knowledge, and emotional stress during this vulnerable phase [7,8]. Considering that 36% of US adults have limited health literacy [9], a substantial proportion of patients may struggle to fully understand and follow postdischarge instructions, which are typically written for communication between health care professionals and often fail to meet patients' informational needs [10]. Therefore, there is a clear need for discharge documents that are patient-centered, clearly structured, and accessible; these will be referred to as patient letters throughout this manuscript.

Recent advances in artificial intelligence (AI), particularly large language models (LLMs) such as GPT-4 (OpenAI), offer a promising approach to address this need [11]. By transforming complex medical documents into plain-language patient letters, AI has the potential to improve patients' comprehension and recall of important medical information. Several feasibility studies have demonstrated that LLMs can generate readable and accurate patient letters from standard discharge documents, highlighting their technical capability to simplify language, structure information, and provide consistent formatting [12-14]. To date, only a small number of studies have investigated the impact of GPT-generated patient letters on patient understanding [15-17]. These studies have primarily relied on semistructured interviews or multiple-choice assessments, offering only quantitative evaluations of patient knowledge. A more nuanced understanding of patient comprehension is still lacking, particularly regarding which types of information are retained, how the cognitive complexity of content influences understanding, and whether comprehension differs across medical content domains. Accordingly, evidence on the educational impact and cognitive benefits of AI-generated patient communications remains limited.

This study is based on our previous work, in which we explored the feasibility of transforming discharge letters into patient letters using GPT-4 [12]. That study primarily focused on the accuracy of the generated letters and measures of patient-centeredness, demonstrating that overall, GPT-4 can produce readable and structured patient-directed documents, with some limitations including omissions. Building on these findings, this study shifts the focus to an application-oriented perspective. It investigates whether GPT-4–generated patient letters improve comprehension, using standardized patients (SPs) to simulate real-world understanding and follow-up tasks.

To systematically evaluate comprehension, we framed patient understanding as a learning process informed by cognitive load theory (CLT) [18]. According to CLT, learning efficiency depends on three types of cognitive load:

1. Intrinsic load refers to the inherent complexity of the information and is dependent on the learner's level of expertise. In the context of discharge letters, intrinsic load is generally high due to complex medical content and procedural details, and it can only be modified to a limited extent.
2. Extraneous load reflects mental resources allocated to elements that do not contribute to learning or schema acquisition. High extraneous load can arise from poorly structured information, excessive jargon, or frequent use of abbreviations [18]. In our prior study, we demonstrated that GPT-4–generated patient letters mitigated these contributing factors by providing increased repetition of key points, fewer abbreviations, and reduced use of medical terminology [12].
3. Germane load represents the mental resources devoted to acquiring and automating schemata in long-term memory. In our context, the goal is to maximize germane load so that patients allocate as much cognitive energy as possible to truly understand, retain, and integrate the medical information.

Guided by this theoretical framework, the study addressed the following research questions:

- Does the reduction of extraneous load through AI-generated patient letters translate into improved comprehension of safety-relevant information?
- Are comprehension gains consistent across different content domains (eg, Medication, Organization, Prevention of Complications, and Lifestyle)?
- Do comprehension improvements differ between lower-order ("Remember") and higher-order ("Understand") cognitive processes as defined by Bloom taxonomy [19]?

By combining a patient-centered perspective with a learning-theoretical framework, this study examines the

XSL•FO
RenderX

educational potential and limitations of AI-generated patient letters and considers how they might contribute to patient safety during vulnerable care transitions.

Methods

Study Design and Setting

In our previous study, we developed a methodology for generating GPT-4–based patient letters, demonstrating the feasibility of transforming discharge letters into patient-friendly formats while maintaining medical accuracy and patient-centeredness. The prompting strategy and human review process were described in detail in that publication [12].

Building directly on this work, this prospective, randomized, crossover study with SPs used 3 discharge letters and the corresponding GPT-4–generated patient letters from our prior study, each representing a common chronic condition: arterial hypertension, type 2 diabetes mellitus, and diabetic kidney disease. These diagnoses were selected because they are highly prevalent and frequently occur in both inpatient and outpatient care. The letters are provided in Multimedia Appendix 1 in German (original) and English (translated).

Both letter formats were presented to 48 SPs using a crossover design (Figure 1). We compared perception and retention of key medical information—defined as "learning objectives"—presented in the letters.

XSL•FO
RenderX

Figure 1. Overview of study design and rating procedure in a prospective, randomized crossover study comparing the retention of predefined learning objectives by standardized patients after reading either a standard hospital discharge letter or a GPT-4–generated patient letter. AI: artificial intelligence; LO: learning objectives.

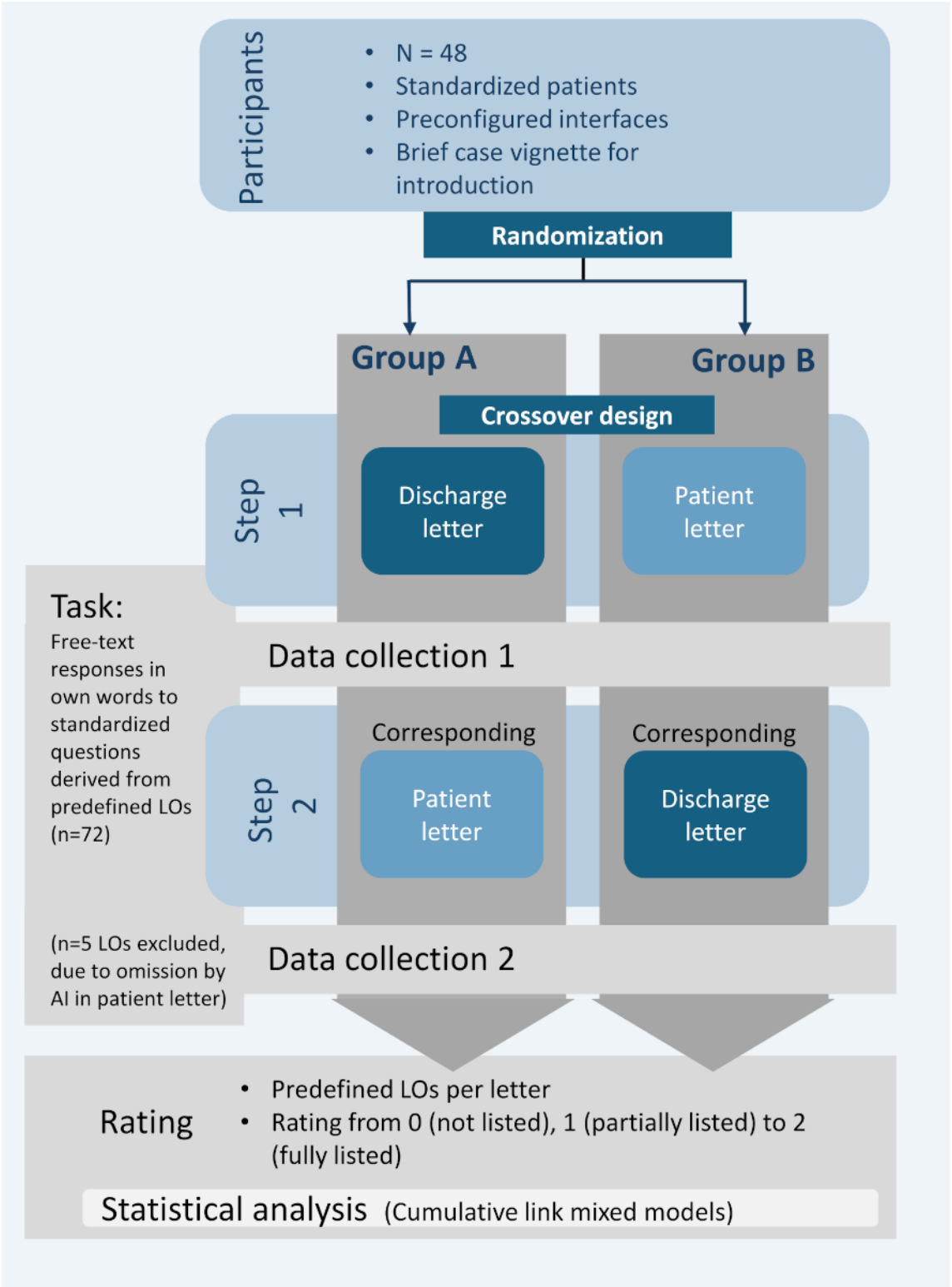

Participants and Study Procedures

A total of 48 SPs participated in our study (Multimedia Appendix 2). SPs are individuals who are specifically trained as patient actors to portray medical patients in a consistent and realistic manner, and they participated in the study in their specific role [20,21]. The sample size was determined by the number of participants available for recruitment and was not calculated prior to the study. The SPs were recruited from the SP programs of the Universities of Tübingen and Erlangen, Germany. Data were collected between May and September 2024. After reading a brief case vignette covering basic medical and socioeconomic information, each participant was randomized to the discharge letter (group A) or its patient-friendly version (group B). Participants were not

XSL•FO
RenderX

informed about the sequence in which they would receive the letters. However, the differences between the discharge letter and the GPT-4–generated patient letter were apparent, making full blinding impossible. Therefore, no formal assessment of blinding adequacy was performed.

At both sites, letters were presented digitally through a web-based interface. Participants were instructed to read each letter with care and were permitted to scroll back to the text while answering the open-ended questions. This setting mirrors the real-world access to written discharge information. The SPs extracted key information from 4 content fields and answered predefined open questions (Multimedia Appendix 3). This approach was chosen because continuous access limits short-term memory bias and mirrors real practice [22]. To foster active knowledge construction, participants were instructed to answer in their own words without copying. Letters were then crossed over, and the task was repeated with the alternate version.

Learning Objectives and Content Fields

For each of the 3 clinical cases (arterial hypertension, diabetes mellitus, and diabetic kidney disease), we defined 24 safety-related learning objectives that cover information considered essential by the treating physicians for patient self-management after discharge (eg, knowledge of medication changes, warning symptoms, and follow-up appointments). The learning objectives were designed and revised by an interdisciplinary committee of 4 physicians (specialized in internal medicine, nephrology, diabetology, general medicine, and medical education) based on current clinical guidelines [23-27] and their professional experience. They were formulated as short, observable statements (eg, “The patient can ... list” and “The patient understands that ...”) and checked for comprehensibility and redundancy.

For each case, the 24 learning objectives were evenly distributed across the levels of the revised Bloom taxonomy, with 12 learning objectives in the “Remember” category and 12 in the “Understand” category. In Bloom taxonomy, the “Remember” category refers to the ability to recall facts and basic concepts, while the “Understand” category goes beyond recall and involves interpreting, explaining, and making sense of the information [19,28]. Each learning objective was further mapped to 1 of 4 content fields: “Organization” (follow-up tasks), “Medication” (drug changes), “Prevention of Complications” (early detection and management of disease-related complications), and “Lifestyle/Disease Management” (self-care advice/lifestyle modifications; Multimedia Appendix 3). This structure ensured comparable cognitive demands and content coverage across all 3 cases and enabled stratified analyses by content area and Bloom category. The process of defining learning objectives and creating the original discharge letters has been described in detail in our previous publication [12].

Across the 3 letters used in this study, 72 learning objectives were assessed in total (24 per letter). A total of 5 of these 72 learning objectives were not represented in the GPT-4–generated patient letters because the model did not carry them over from the original discharge letters. These learning objectives were therefore excluded from the analysis for both letter types (Multimedia Appendix 4). The quality and completeness of the translation from discharge letter to patient letter were examined in detail in our previous publication and were therefore not reassessed within the scope of this study [12]. After this exclusion, 34 of the 67 remaining learning objectives were classified as “Remember” and 33 as “Understand” according to Bloom taxonomy (Multimedia Appendix 3).

Assessment Approach

Reading the discharge or patient letter and then answering the case questions was designed as a structured learning assessment sequence [29]: the letters provided the information to be learned, and the SPs’ responses reflected what they had actually understood. To capture constructed understanding rather than simple recognition, we used open-ended questions that required participants to formulate their answers in their own words rather than choosing from predefined options.

Open-ended response formats are less prone to guessing than multiple-choice questions and are better suited for assessing higher levels of processing for learning objectives at the “Understand” level. This approach closely aligns the assessment with the predefined learning objectives and the intended depth of understanding. This choice is supported by assessment literature [30].

Outcome Measures

The primary outcome was the understanding of safety-related information at the level of individual learning objectives. Each participant response (participant × learning objective × letter) was rated on an ordinal scale with 3 categories reflecting the degree of goal achievement in the open-ended response (0=not specified, 1=partially specified, 2=fully specified). The main comparison of interest was the difference between the patient letters generated by GPT-4 and the standard discharge letters with respect to this ordinal measure of SPs’ understanding.

Secondary end points were the distribution of this ordinal comprehension metric across the 4 predefined content areas (Organization, Medication, Prevention of Complications, and Lifestyle/Disease Management) and the distribution across the levels of Bloom taxonomy (“Remember” vs “Understand”). As an additional secondary end point, the frequency of medically incorrect answers was examined.

Randomization

Participants at the University Hospital Tübingen were allocated to the 3 clinical cases and the order of letter presentation (discharge letter first vs patient letter first) according to a pregenerated randomization list. This list was created prior to participant recruitment using a randomly generated allocation sequence that specified both case assignment and presentation order, with approximately balanced group sizes. Participants were enrolled consecutively and assigned to the next available entry on the randomization list. Allocation was not known in advance to either participants or study personnel conducting the sessions, thereby preserving allocation concealment at the participant level.

Participants from the University of Erlangen, who took part remotely via a web-based study platform, were randomized

XSL•FO
RenderX

using a computer-generated 1:1 allocation algorithm implemented within the platform. Upon enrollment, the platform automatically assigned participants to the letter-order sequence and the clinical case.

The use of 2 different randomization methods resulted from the different study sites: on-site participation in Tübingen required an offline procedure, whereas the Erlangen cohort participated exclusively online. Randomization was conducted separately within each site and not across sites. Although randomization procedures differed between sites, both approaches ensured allocation concealment at the individual level. Importantly, the substantive data collection procedure (including exposure to the letters and assessment of learning objectives) was identical at both locations.

Data Analysis

Rating Process

The answers of the SPs were independently rated by 2 experienced clinicians (FH and AS) to determine whether the predefined learning objectives were correctly and completely identified and reported. The rating process followed a predefined, ordinally scaled rating structure, ranging from "not listed" over "partially listed" to "fully listed," and disagreements were solved by discussion. A specific definition was established in advance for each learning objective (Multimedia Appendix 3). Raters were blinded to both the letter type (discharge letter or patient letter) and the order in which it was presented. No formal assessment of blinding adequacy was performed. The raters achieved a Cohen κ of 0.97. Due to the high level of agreement, one of the two ratings was selected at random for use in the statistical analysis.

Statistical Analysis

Due to the ordinal nature of the ratings of the learning objectives (comprehensively reported, partially reported, and not reported), cumulative link mixed models were fitted to the data. Random intercepts for participant and learning objective were used to account for participant and learning objective random effects (Multimedia Appendix 5). The fixed effects of letter type, Bloom category, and content field were investigated as variables of interest with model comparisons. The fixed effects disease type and time point of measurement were included as controls. Relevant models were assessed, and alternatives were considered. *P* values for differences in individual content fields were calculated by subsetting the data to each content field. Models were compared according to the Akaike Information Criterion and likelihood ratio tests. No missing data were present. The analyses were conducted using R software (version 4.4; R Foundation for Statistical Computing) and the ordinal package version 2023.12-4.1 [31,32].

Reporting Standards

The reporting of this study followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guideline for cohort studies [33]. A reporting checklist can be found in the Multimedia Appendix 6.

Ethical Considerations

This study was approved by the local ethics committee (778/2023BO2) of Tübingen University Hospital, Germany. Written informed consent was obtained from the SPs before inclusion in the study.

No direct identifiers (eg, names or dates of birth) were collected in the study. A short demographic questionnaire (age, gender, educational background, and prior experience as an SP) was completed by the SPs; these results were stored separately and not linked to individual response data. During data collection, response data were temporarily saved on secured cloud infrastructure; evaluation data and the separate demographic file were stored in different locations with a role-based access control. Reidentification was not possible at any time. Due to data loss at one site, demographics were available for 38 out of 48 SPs. As demographics were not used in the analysis of learning objective performance, and there is no evidence of systematic bias between sites, this missing information does not affect the validity of the study outcomes. The study was conducted in accordance with the Declaration of Helsinki.

Results

Demographic Data

The median age of the participants was 46.7 (IQR 28.3-65.0) years, with a female-to-male ratio of approximately 1.92:1. The majority (24/38, 63.2%) had a university degree. Task order did not influence performance (*P*=.80), although participants performed slightly better on their second attempt (odds ratio [OR] 1.23, 95% CI 1.03-1.47; *P*=.02). The type of disease did not have a significant main effect on the retention rate of learning objectives (*P*=.96; Multimedia Appendix 5). Therefore, all patient data were pooled for the subsequent comparison of the discharge letter and the patient letter.

Patient Letters vs Discharge Letters

Overall, patient letters improved information uptake: SPs retained more learning objectives after reading the patient letter compared with the discharge letter (OR 1.74, 95% CI 1.45-2.08; *P*<.001; Figure 2). Comprehensive reporting rose to 45.7% (490/1073) with the patient letter compared to 38.5% (413/1073) with the discharge letter. Partial reporting also improved (322/1073, 30% vs 287/1073, 26.7%), while omissions decreased from 34.8% (373/1073) to 24.3% (261/1073; Figure 2; Multimedia Appendix 7). Remarkably, despite overall gains, those 24.3% (261/1073) of all ratings remained "not reported" after reading the patient letter. Because omissions varied by participant, we conducted an item-level analysis to identify objectives consistently missed (predefined as "not reported" in ≥50% of the cases). A total of 6 (9%) of 67 objectives met this criterion. By Bloom level, 5 out of 6 were "Understand" and 1/6 "Remember"; by content field, 3 out of 6 fell under "Prevention of Complications," 2 out of 6 under "Lifestyle/Disease Management," and 1 out of 6 under "Organization" (Multimedia Appendix 8). On the other hand, the overall proportion of medically incorrect reports by the participants was low, with no significant difference between the letter types (*P*=.55; Figure 3).

XSL•FO
RenderX

Figure 2. Percentage of predefined learning objectives that were comprehensively reported, partially reported, or not reported by standardized patients after reading either a standard hospital discharge letter (DL) or a GPT-4–generated patient letter (PL). Results are shown overall (left) and separately (right) for the Bloom taxonomy levels "Remember" and "Understand." Statistical comparison was performed using cumulative link mixed models. Overall, standardized patients retained significantly more learning objectives with the patient letter than with the discharge letter. Across both letter types, performance was significantly higher for "Remember" than for "Understand" learning objectives.

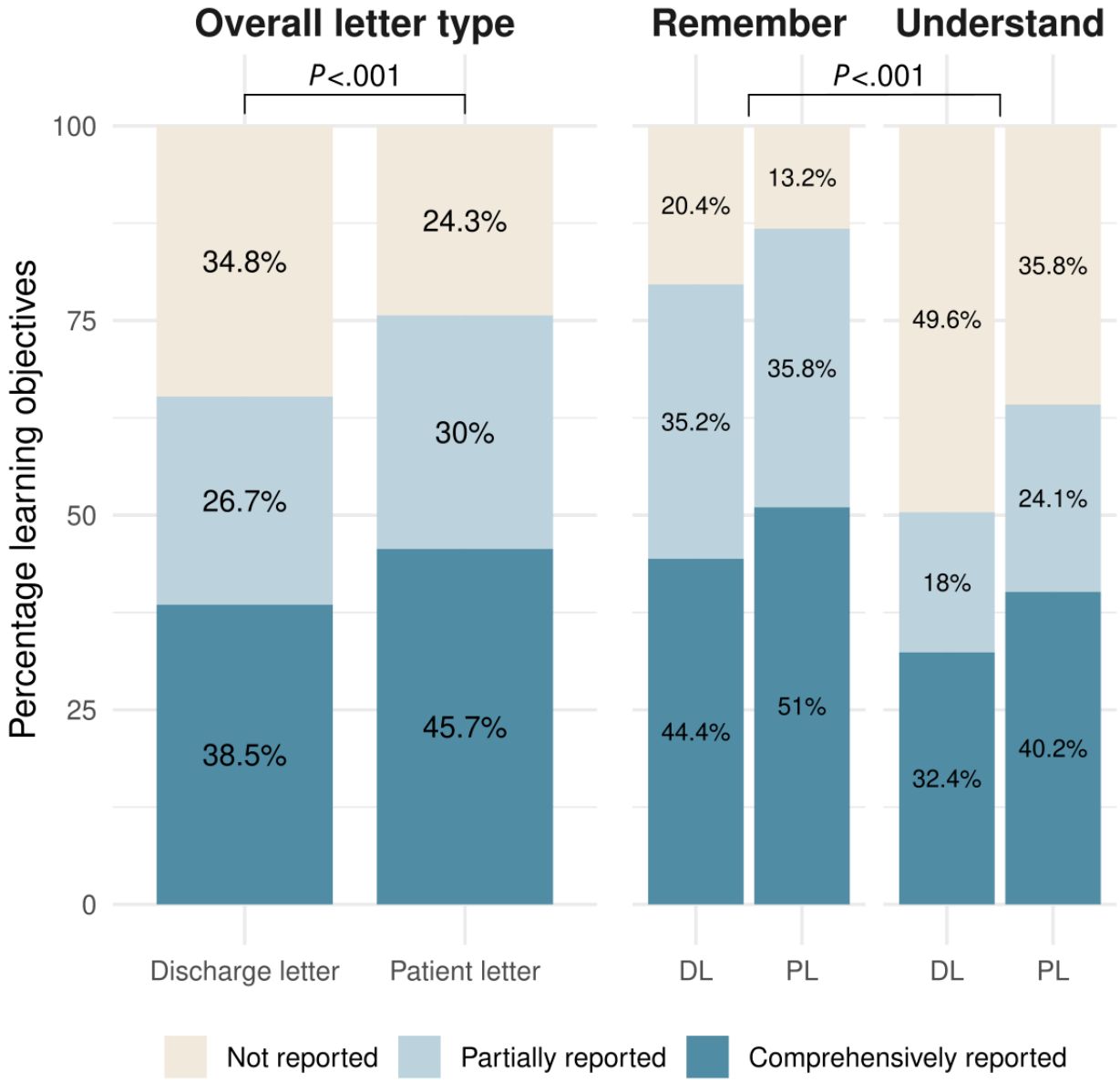

XSL•FO
RenderX

Figure 3. Medical correctness of standardized patient responses after reading either a standard hospital discharge letter or a GPT-4–generated patient letter. The figure illustrates the percentage of responses rated as medically accurate, medically inaccurate, or not applicable (statements without medical content). Statistical comparison was conducted using cumulative link mixed models.

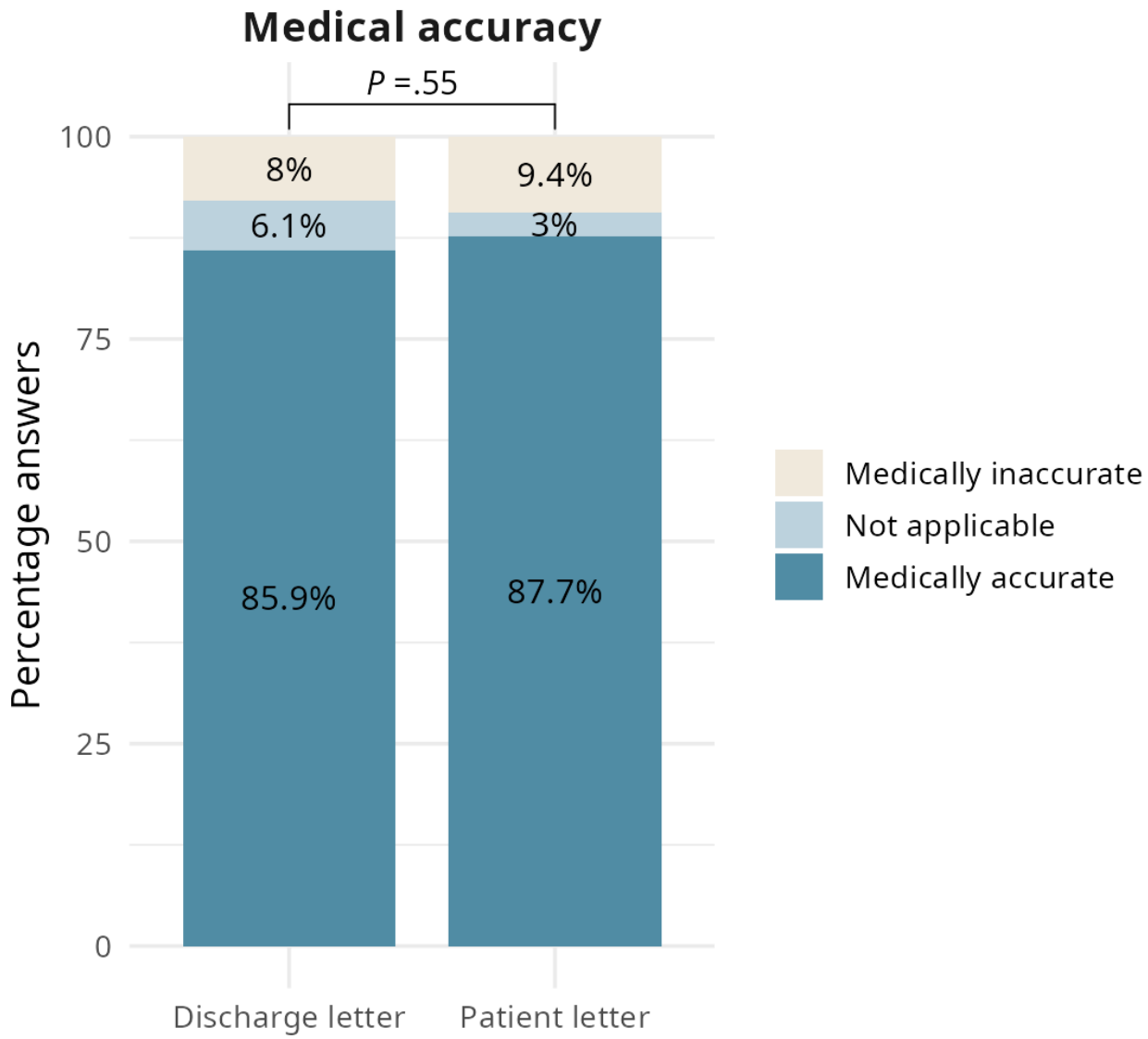

Impact of Cognitive Demand (Bloom Taxonomy) on Information Uptake

A subgroup analysis of learning objectives based on Bloom taxonomy revealed that, regardless of letter type, participants performed better on learning objectives classified under the "Remember" category compared to those classified under the "Understand" category (OR 3.33, 95% CI 1.96-5.88; *P*<.001; Figure 2). In both categories, the proportion of comprehensively reported learning objectives was higher when presented in the patient letter format compared to the discharge letter format (278/545, 51% vs 242/545, 44.4% for "Remember" and 212/528, 40.2% vs 171/528, 32.4% for "Understand"; Figure 2; Multimedia Appendix 7). A subsequent analysis showed no significant interaction between Bloom taxonomy level and letter type (*P*=.13), indicating that the observed benefits of the patient letter were consistent across both cognitive levels.

Impact of Content Field on Information Uptake

In a subgroup analysis based on content fields, we observed a statistically significant interaction effect between the type of letter and the content field (*P*<.001), indicating that the benefit of the patient letter depends on the content field. Additional values for effect size, OR, and CI for the results in this paragraph are illustrated in Multimedia Appendix 9. A substantial increase in comprehension associated with the patient letter was observed for the domain of "Medication" (*P*<.001), with a rate of accurately reported learning objectives of 66.9% (170/254) compared to 50.8% (129/254) for the discharge letter. In the GPT-4–generated patient letter, the changes and purpose of each drug were explicitly stated in plain language (eg, "You were also given two other medications, metformin and empagliflozin, which help control blood sugar"), whereas the standard discharge letter contained this information in a condensed table format without further explanation. For further details on the translation process from discharge to patient letters, we refer to our previous work [12].

XSL•FO
RenderX

The second largest improvement was seen in the domain of "Organization" (78/158, 49.4% vs 62/158, 39.2%; *P*<.001; Figure 4; Multimedia Appendix 7). In contrast, improvements observed with the patient letter in the content fields "Lifestyle/Disease Management" (*P*=.38) and "Prevention of Complications" (*P*<.01) were small and medium. In the content field "Prevention of Complications," the patient letter resulted in a minor increase in the proportion of partially addressed learning objectives compared to the discharge letter (102/311, 32.8% vs 70/311, 22.5%). However, the rate of completely reported learning objectives remained the lowest compared to the other content fields (Figure 4; Multimedia Appendix 7). Regarding the content field "Lifestyle/Disease Management," there was even a small decrease observed for the partially reported learning objectives (100/350, 28.6% in the patient letter vs 107/350, 30.6% in the discharge letter). However, the amount of correctly stated learning objectives still showed a small increase (149/350, 42.6% in the patient letter vs 139/350, 39.7% in the discharge letter; Figure 4; Multimedia Appendix 7).

Figure 4. Percentage of predefined learning objectives that were comprehensively reported, partially reported, or not reported by standardized patients after reading either a standard hospital discharge letter (DL) or a GPT-4–generated patient letter (PL). Results are displayed across 4 different content categories: Organization, Medication, Prevention of Complications, and Lifestyle/Disease Management. Differences between letter types were evaluated using cumulative link mixed models. Significant improvements were achieved with the patient letter in the content fields "Medication" and "Organization," whereas gains were modest for "Prevention of Complications" and minimal for "Lifestyle/Disease Management.".

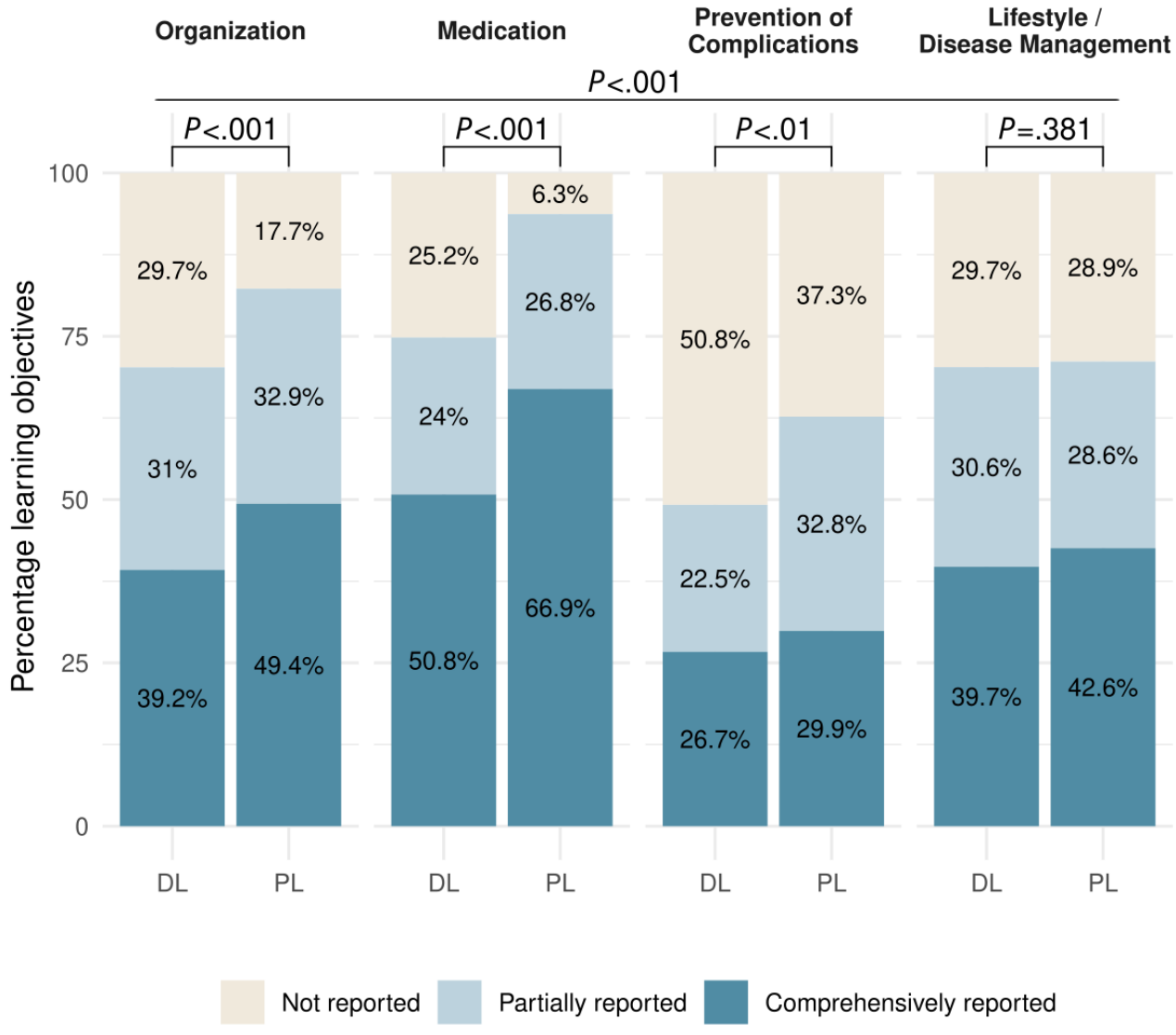

Discussion

Principal Findings

This prospective, explanatory study with SPs showed that AI-generated patient letters increased the retrieval of medication-related learning objectives to 66.9% (170/254) and organizational objectives to 49.4% (78/158), compared with 50.8% (129/254) and 39.2% (62/158), respectively, after reading the standard hospital discharge letters. These gains are larger than those reported in earlier single-arm or pre-post studies of human or AI-generated patient letters that lacked counterbalancing [13,34,35]. By scoring open, free-text answers, we also captured a broader slice of comprehension than recent

XSL•FO
RenderX

GPT-based studies that relied on recognition formats [15-17]. However, the reliance on predefined learning objectives, evaluated by clinicians, may still miss nuances of SP understanding. In particular, misconceptions and emotional interpretations may be underreported, and these remain important aspects for future qualitative work.

Content- and Bloom-Specific Insights

Although overall performance improved, SPs still failed to report, on average, nearly a quarter of the key information, highlighting a persistent patient safety gap for patients. Gains clustered in the areas of Medication and Organization, which are long-recognized weak spots: <50% of patients understand their discharge medicines [36], and medication errors account for roughly half of preventable harm [37]. By increasing the retention of medication-related information, we hypothesize that patient letters have the potential to improve medication management and reduce medication errors, thereby potentially contributing to enhanced patient safety. Similarly, improving awareness of necessary follow-up appointments may also support patient safety by reducing missed visits and associated complications. Previous research supports these mechanisms: written patient information material has been shown to increase medication adherence [38], missed follow-up appointments predict readmissions and emergency visits [39-41], and low health literacy is associated with higher readmission rates [42]. Moreover, communication interventions at hospital discharge are significantly linked to fewer readmissions and higher treatment adherence [4]. While these findings provide a rationale for the use of AI-generated patient letters, the effects on actual patient safety outcomes remain to be empirically evaluated in future research.

In contrast, improvements in the content field "Prevention of Complications" were modest, and those for "Lifestyle/Disease Management" were even smaller. Notably, 37.3% (116/311) of the key information regarding "Prevention of Complications" remained unknown. While higher cognitive demand likely contributed, other factors such as complex sentence structures, domain-specific phrasing, multistep reasoning, and the absence of visual aids may have further limited comprehension. As shown in our previous study, GPT-4 also faced greater challenges translating "Prevention of Complications" content compared with other domains [12]. Interestingly, although GPT-4 more successfully simplified "Lifestyle/Disease Management" information, SP performance remained low, suggesting that simplification alone does not guarantee comprehension when integration into practical self-management is required, and suggesting a distinction between content that is challenging for the AI to translate versus content that is difficult for SPs to understand and apply. Stratifying learning objectives by Bloom taxonomy showed higher performance for "Remember" than for "Understand" items (278/545, 51% vs 212/528, 40.2% correct with the patient letter), and 5 out of 6 objectives omitted by ≥50% of participants fell into the "Understand" category. This difference is unsurprising given that "Remember" tasks primarily involve locating explicit information, whereas "Understand" tasks require higher-order cognitive processes. Importantly, the patient letter improved performance in both categories compared with the discharge letter ("Remember": 278/545, 51% vs 242/545, 44.4%; Understand: 212/528, 40.2% vs 171/528, 32.4%), and there was no significant interaction between letter type and Bloom level (P=.13), indicating that the observed benefit of the patient letter was approximately equal for both cognitive levels.

These findings suggest that GPT-4–generated patient letters can support not only the retrieval of factual information but may also modestly enhance comprehension-oriented learning objectives, though higher-level understanding remains more challenging. This is consistent with prior longitudinal research demonstrating that discharge instructions decay rapidly without reinforcement [43-45].

The clustering of consistently missed objectives in "Prevention of Complications" and, secondarily, "Lifestyle/Disease Management," together with the predominance of "Understand"-level items, suggests that a static, text-only letter may not be sufficient in contexts where patients need to link symptoms to actions, reason about conditional risks (eg, sleep apnea as source of high blood pressure), or maintain self-management routines (eg, home blood pressure documentation or dietary change).

Didactic Implications (CLT)

CLT provides a helpful framework for interpreting our findings [18]. Recent work applying CLT to clinical practice shows that stress, emotion, and uncertainty increase intrinsic cognitive load and that working memory depletion can impair clinical performance [46]. Against this background, it seems plausible that linguistic simplification and better structure primarily reduce extraneous cognitive load. Our prior work demonstrated that GPT-4 can reorganize discharge letters into patient letters mainly by removing jargon and chunking information [12], which likely supports this reduction. While discharge letters are generally already well-structured, AI-generated patient letters further optimize content organization in a patient-centered manner.

We did not directly measure intrinsic, extraneous, or germane cognitive load in this study; CLT is therefore used as an interpretive framework rather than an empirically tested mechanism. Our learning objectives were deliberately designed so that each content field contained both "Remember" and "Understand" objectives. However, objectives that remained most difficult (defined as those omitted by ≥50% of participants) were predominantly from the "Understand" category in the area of "Prevention of Complications" and, to a lesser extent, in the area of "Lifestyle/Disease Management." These learning objectives (eg, hypoglycemia symptom-action mapping, recognition of statin-associated muscle symptoms, dehydration-related kidney risk, or the contribution of home blood pressure monitoring and Mediterranean diet to long-term disease control) require the integration of multiple information elements, conditional if-then thinking, and sustained self-regulation. Within the framework of CLT, they can therefore be interpreted as tasks with a comparatively high intrinsic cognitive load.

In contrast, the largest gains from patient letters were observed predominantly in factual learning objectives in the areas of "Medication" or "Organization," which required participants

XSL•FO
RenderX

to find and reproduce explicitly stated information (eg, medication changes or appointments). Such learning objectives are likely to be particularly sensitive to a reduction in extrinsic load through simpler language and clearer structure.

Based on these findings, we conclude that AI-based translation into patient letters may reduce extraneous cognitive load, thereby possibly freeing working-memory resources for the processing of concrete, action-oriented tasks (eg, tablet strength and appointment date). However, for learning objectives with high intrinsic cognitive load, such as symptom-to-action coupling (eg, tremors-hypoglycemia compensation required), conditional risk thinking (eg, allergy present notification to the care team), or permanent self-regulation (eg, home blood pressure monitoring and dietary changes) reducing extrinsic cognitive load alone may be not sufficient to free up enough relevant resources (germane cognitive load) for deeper understanding and behavioral change.

Taken together, these CLT-based interpretations are hypothesis-generating and should be tested in future studies that directly measure cognitive load. They further suggest that patient letters may be most effective for immediate, concrete instructions, whereas higher-order learning objectives may require complementary, multimodal supports (eg, infographics, microvideos, interactive prompts, and teach-back). These, however, need prospective evaluation. Evidence for each adjunct already exists in discharge-education and adherence research [38,47,48].

Beyond CLT, factors such as stress-induced cognitive depletion, low health literacy, and limited self-efficacy may also reduce patients' ability to process complex medical content [49-51]. Tailoring follow-up communication to these variables will be essential [52].

Clinical Implications

The implementation of AI-generated patient letters requires careful attention to ethical, legal, and social aspects [53]. In general, human oversight ("human-in-the-loop") remains essential when using AI to translate clinician-written discharge letters into patient letters. Our study indicates that the patient letter format alone does not yield sufficient comprehension gains for cognitively demanding learning objectives, particularly those classified as "Understand" by Bloom taxonomy and thematically related to prevention of complications or lifestyle/disease management. It is further underscored by omissions and hallucinations observed in the prior translation process [12], which required the exclusion of 5 learning objectives in this study. Therefore, establishing practical governance structures is essential for clinical practice. By identifying the specific content domains most prone to errors or omissions, our results can help to target quality-assurance efforts efficiently, making oversight both feasible and focused on the areas critical for patient safety.

From a legal perspective, robust data protection is critical, as personal patient information constitutes a central component of these letters [54,55]. Socially, simplified language may improve accessibility for patients with low health literacy and reduce language barriers [56]. While the format does not accommodate individual literacy needs or reading difficulties, it may still help reduce social barriers in discharge communication. Finally, practical and feasibility considerations include technical integration into existing clinical workflows, usability for health care staff, and costs associated with development, implementation, and maintenance. However, manual letter production is time-consuming and costly, whereas AI offers a potentially scalable solution. Effective measures without AI can be found in the literature, including patient letters written in simple language and structured summaries that reduce technical jargon and organize content for easy recall [57], portal access (facilitating review but not guaranteeing understanding) [58,59], supplementary written materials (linked with better medication adherence) [38], audio recordings of consultations (supporting recall) [43], and interactive elements such as brief teach-back or app-based prompts (which enhance encoding and memory) [47,49]. However, implementation is frequently constrained by time and workload, limited organizational support and training, workflow barriers, and gaps in digital access and use [58,59]. These problems are documented across teach-back implementation (time/sustainability barriers), discharge letter production (time pressure, authorship level, and workload effects), and portal-based delivery (digital divide: differences in access, skills, and use) [60-62].

Our study isolates the contribution of AI-generated formulations and was not designed as a head-to-head comparison. In this context, AI-generated patient letters, with clinician oversight, may offer a practical, scalable way to produce patient-centered text for domains where our data indicate clear gains (eg, medication and organization), and future work should address workflow integration, usability for clinicians, maintenance costs, and robust data protection.

Limitations

This study has several limitations. First, we used only 1 GPT-4 version and focused on 3 chronic diseases (hypertension, diabetes, and diabetic kidney disease), which limits generalizability to other LLM versions and medical conditions. Although participants and clinician raters were blinded to letter type and order as far as possible, full blinding was not achievable, and the success of blinding was not formally assessed for either group. Because the stylistic differences between discharge letters and patient letters are recognizable, which was unavoidable given the study content, participants and raters may have inferred the format, introducing potential expectancy effects. Such effects could operate in both directions: some participants might assume the patient letter to be easier, whereas others may approach AI-generated text with skepticism, potentially reducing trust or attention. These bidirectional biases could either amplify or attenuate format-related differences. While standardized instructions, structured scoring procedures, and uniform training were used to minimize subjective influence, residual bias cannot be excluded and should be considered when interpreting the magnitude of effects. Additionally, we did not formally assess the feasibility or acceptability of patient letters through validated usability or satisfaction scales. These aspects should be addressed in future implementation studies. Furthermore, our study used standardized rather than real-world patients to examine the effects of AI-generated patient letters.

XSL•FO
RenderX

This approach ensured high procedural consistency and enabled a controlled comparison between AI-generated and human-written letters, as reflected in recent high-impact AI studies using SPs [21]. However, it limits external validity; SPs are not able to fully represent real patients' stress levels, emotional states, cognitive vulnerability, and health literacy diversity, especially in the context of acute illness or hospitalization. In clinical practice, such factors may negatively affect patient comprehension due to reduced germane cognitive load. Consequently, although the direction of the observed effects (improved understanding through AI-generated patient letters) is expected to remain, the overall level of understanding and recall would likely be lower among real patient populations for both letter types. In addition, letters were written in German and later translated for publication, limiting the generalizability of our findings to other languages. Another potential limitation of this study is the exclusion of 5 predefined learning objectives that were not present in the GPT-4–generated patient letters due to omissions, which could introduce bias in favor of the AI-generated format. However, a statistical analysis including all 72 learning objectives showed that the overall pattern of results remained unchanged despite the absence of these 5 items in the AI-generated patient letters. This suggests that the exclusion did not materially affect the observed effects. Additionally, the sample size in our study was determined by availability rather than by formal power calculation, which may limit the ability to detect subtle letter-type differences, particularly in subgroup analyses. Our findings should therefore be interpreted with caution. Finally, given the potential for hallucinations or omissions in AI-generated texts, as observed in the translation process [12], clinical review and governance are essential if AI-based tools are deployed in practice, and attention to AI transparency and the potential for unexpected outputs remains critical.

Conclusions

This hypothesis-driven, explanatory study showed that AI-generated patient letters improved the uptake of basic, action-oriented information in SPs. This effect might reflect a reduction in extraneous cognitive load. We hypothesize that, in clinical practice, this could reduce common sources of postdischarge harm and improve patient safety. However, patient letters alone did not close gaps in higher-order understanding, particularly in risk prevention and lifestyle change. Addressing such complex learning goals may require multimodal, interactive support (eg, explanatory graphics, short videos, or app-based follow-ups [47,48,63]) that reinforce the text-based patient letter.

Future studies should therefore evaluate integrative models that layer such tools onto patient letters and tailor content to health literacy, self-efficacy, and situational resources [64], ideally within a didactic framework (eg, Observing Patient Involvement Scale or Patient Education Materials Assessment Tool [65]) to enhance actionability. Studies including more diverse medical conditions, real patient populations with varying health literacy levels, different LLMs, and multiple languages are needed to improve generalizability. We also suggest incorporating qualitative approaches in future studies (eg, think-aloud protocols or teach-back methods) to better capture nuances in patient understanding and potential contributors to medical error.

Acknowledgments

We wish to thank the University of Tübingen and Eric Nazarenus for supporting our study.

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading, editing, and translation.

The GenAI tools used were: ChatGPT by OpenAI, DeepL SE.

Responsibility for the final manuscript lies entirely with the authors.

GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Funding

This work was supported by the Open Access Publishing Fund of the University of Tübingen.

Authors' Contributions

FE had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: FE, FH, AS
Acquisition, analysis, or interpretation of data: all authors
Drafting of the manuscript: FE, AS, FH
Critical review of the manuscript for important intellectual content: all authors
Statistical analysis: KS

XSL•FO
RenderX

Administrative, technical, or material support: CSP, KS, PS
Supervision: AHW, MH

Conflicts of Interest

None declared.

Multimedia Appendix 1
GPT-4–generated patient letters (German original and English translation).
[PDF File (Adobe PDF File), 319 KB - jmir_v28i1e81243_app1.pdf]

Multimedia Appendix 2
Study flowchart. Overview of participant enrollment, randomization, crossover procedure, and analysis population.
[PDF File (Adobe PDF File), 86 KB - jmir_v28i1e81243_app2.pdf]

Multimedia Appendix 3
Overview of the learning objectives, standardized questions, and the corresponding rating scheme applied to evaluate information retention and comprehension across study conditions.
[XLSX File (Microsoft Excel File), 18 KB - jmir_v28i1e81243_app3.xlsx]

Multimedia Appendix 4
List of the learning objectives that were not represented in the GPT-4–generated patient letters because the model did not carry them over from the original discharge letters. These objectives were excluded from the analysis for both letter types.
[DOCX File , 22 KB - jmir_v28i1e81243_app4.docx]

Multimedia Appendix 5
Likelihood ratio tests.
[DOCX File , 27 KB - jmir_v28i1e81243_app5.docx]

Multimedia Appendix 6
STROBE checklist detailing how each reporting item was fulfilled and where it can be found in the manuscript.
[DOCX File , 33 KB - jmir_v28i1e81243_app6.docx]

Multimedia Appendix 7
Number of learning outcomes achieved, stratified by letter type (patient letter vs discharge letter).
[DOCX File , 25 KB - jmir_v28i1e81243_app7.docx]

Multimedia Appendix 8
Item-level analysis identifying learning objectives that were consistently missed, predefined as those “not reported” by ≥50% of participants.
[DOCX File , 22 KB - jmir_v28i1e81243_app8.docx]

Multimedia Appendix 9
Odds ratios from the best-fitting cumulative link mixed model. Odds ratios represent the interaction between letter type and content field, estimated relative to an arbitrary reference category. The reference category is defined as letter type = discharge letter (DL) and content field = Lifestyle / Disease Management. *P* values for the odds ratios are calculated in comparison with this reference category.
[PNG File , 106 KB - jmir_v28i1e81243_app9.png]

References

1. Marsall M, Hornung T, Bäuerle A, Weigl M. Quality of care transition, patient safety incidents, and patients' health status: a structural equation model on the complexity of the discharge process. BMC Health Serv Res 2024;24(1):576 [FREE Full text] [doi: 10.1186/s12913-024-11047-3] [Medline: 38702719]
2. Moore C, Wisnivesky J, Williams S, McGinn T. Medical errors related to discontinuity of care from an inpatient to an outpatient setting. J Gen Intern Med 2003;18(8):646-651 [FREE Full text] [doi: 10.1046/j.1525-1497.2003.20722.x] [Medline: 12911647]

XSL•FO
RenderX

3. Steele EM, Burney HN, Freije SL, Zellars RC, Lautenschlaeger T, Holmes JA. Missed follow-up is associated with worse survival in stage I lung cancer: results from a large multi-site academic hospital system. Sci Rep 2024;14(1):17710 [FREE Full text] [doi: 10.1038/s41598-024-68351-5] [Medline: 39085317]
4. Becker C, Zumbrunn S, Beck K, Vincent A, Loretz N, Müller J, et al. Interventions to improve communication at hospital discharge and rates of readmission: a systematic review and meta-analysis. JAMA Netw Open 2021;4(8):e2119346 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.19346] [Medline: 34448868]
5. Engel KG, Heisler M, Smith DM, Robinson CH, Forman JH, Ubel PA. Patient comprehension of emergency department care and instructions: are patients aware of when they do not understand? Ann Emerg Med 2009;53(4):454-461.e15. [doi: 10.1016/j.annemergmed.2008.05.016] [Medline: 18619710]
6. Belachew EA, Sendekie AK, Wondm SA, Ayele EM, Netere AK. Misunderstanding of dosing regimen instructions among patients with chronic diseases receiving polypharmacy at the University of Gondar comprehensive specialized hospital. PLoS One 2023;18(1):e0280204 [FREE Full text] [doi: 10.1371/journal.pone.0280204] [Medline: 36634103]
7. Kreps GL. Promoting patient comprehension of relevant health information. Isr J Health Policy Res 2018;7(1):56 [FREE Full text] [doi: 10.1186/s13584-018-0250-z] [Medline: 30236157]
8. Toole J, Kohansieh M, Khan U, Romero S, Ghali M, Zeltser R, et al. Does your patient understand their treatment plan? Factors affecting patient understanding of their medical care treatment plan in the inpatient setting. J Patient Exp 2020;7(6):1151-1157 [FREE Full text] [doi: 10.1177/2374373520948400] [Medline: 33457558]
9. Magnani JW, Mujahid MS, Aronow HD, Cené CW, Dickson VV, Havranek E, American Heart Association Council on EpidemiologyPrevention; Council on Cardiovascular Disease in the Young; Council on CardiovascularStroke Nursing; Council on Peripheral Vascular Disease; Council on Quality of CareOutcomes Research;Stroke Council. Health literacy and cardiovascular disease: fundamental relevance to primary and secondary prevention: a scientific statement from the American Heart Association. Circulation 2018;138(2):e48-e74 [FREE Full text] [doi: 10.1161/CIR.0000000000000579] [Medline: 29866648]
10. Choudhry AJ, Baghdadi YMK, Wagie AE, Habermann EB, Heller SF, Jenkins DH, et al. Readability of discharge summaries: with what level of information are we dismissing our patients? Am J Surg 2016;211(3):631-636 [FREE Full text] [doi: 10.1016/j.amjsurg.2015.12.005] [Medline: 26794665]
11. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health 2023;5(4):e179-e181 [FREE Full text] [doi: 10.1016/S2589-7500(23)00048-1] [Medline: 36894409]
12. Eisinger F, Holderried F, Mahling M, Stegemann-Philipps C, Herrmann-Werner A, Nazarenus E, et al. What's going on with me and how can I better manage my health? The potential of GPT-4 to transform discharge letters into patient-centered letters to enhance patient safety: prospective, exploratory study. J Med Internet Res 2025;27:e67143 [FREE Full text] [doi: 10.2196/67143] [Medline: 39836954]
13. Zaretsky J, Kim JM, Baskharoun S, Zhao Y, Austrian J, Aphinyanaphongs Y, et al. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. JAMA Netw Open 2024;7(3):e240357 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.0357] [Medline: 38466307]
14. Shahnam A, Nindra U, Hitchen N, Tang J, Hong M, Hong JH, et al. Application of generative artificial intelligence for physician and patient oncology letters—AI-OncLetters. JCO Clin Cancer Inform 2025(9):e2400323. [doi: 10.1200/cci-24-00323]
15. Kumar A, Wang H, Muir KW, Mishra V, Engelhard M. A cross-sectional study of GPT-4–based plain language translation of clinical notes to improve patient comprehension of disease course and management. NEJM AI 2025;2(2):AIoa2400402. [doi: 10.1056/aioa2400402]
16. van Driel MHE, Blok N, van den Brand JAJG, van de Sande D, de Vries M, Eijlers B, et al. Leveraging GPT-4 enables patient comprehension of radiology reports. Eur J Radiol 2025;187:112111 [FREE Full text] [doi: 10.1016/j.ejrad.2025.112111] [Medline: 40294519]
17. Kim SI, Park J, Kim T, Seo W, Kim T, Cha WC, et al. Enhancing patient participation in emergency department through patient-friendly clinical notes generated by large language models. Sci Rep 2025;16(1):1409 [FREE Full text] [doi: 10.1038/s41598-025-31113-y] [Medline: 41350380]
18. Sweller J. Cognitive load during problem solving: effects on learning. Cogn Sci 2010;12(2):257-285 [FREE Full text] [doi: 10.1207/s15516709cog1202_4]
19. Adams NE. Bloom's taxonomy of cognitive learning objectives. J Med Libr Assoc 2015;103(3):152-153 [FREE Full text] [doi: 10.3163/1536-5050.103.3.010] [Medline: 26213509]
20. Felix HM, Simon LV. Types of Standardized Patients and Recruitment in Medical Simulation. Treasure Island (FL): StatPearls; 2025.
21. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature 2025;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
22. Hellyer S. Supplementary report: frequency of stimulus presentation and short-term decrement in recall. J Exp Psychol 1962;64(6):650-650. [doi: 10.1037/h0043117] [Medline: 13953638]

XSL•FO
RenderX

23. McEvoy JW, McCarthy CP, Bruno RM, Brouwers S, Canavan MD, Ceconi C, ESC Scientific Document Group. 2024 ESC guidelines for the management of elevated blood pressure and hypertension. Eur Heart J 2024;45(38):3912-4018. [doi: 10.1093/eurheartj/ehae178] [Medline: 39210715]
24. American Diabetes Association Professional Practice Committee. 6. Glycemic goals and hypoglycemia: standards of care in diabetes-2024. Diabetes Care 2024;47(Suppl 1):S111-S125. [doi: 10.2337/dc24-S006] [Medline: 38078586]
25. American Diabetes Association Professional Practice Committee. 2. Diagnosis and classification of diabetes: standards of care in diabetes-2024. Diabetes Care 2024;47(Suppl 1):S20-S42. [doi: 10.2337/dc24-S002] [Medline: 38078589]
26. American Diabetes Association Professional Practice Committee. 3. Prevention or delay of diabetes and associated comorbidities: standards of care in diabetes-2024. Diabetes Care 2024;47(Suppl 1):S43-S51. [doi: 10.2337/dc24-S003] [Medline: 38078581]
27. American Diabetes Association Professional Practice Committee. 11. Chronic kidney disease and risk management: standards of care in diabetes-2024. Diabetes Care 2024;47(Suppl 1):S219-S230. [doi: 10.2337/dc24-S011] [Medline: 38078574]
28. Krathwohl DR. A revision of Bloom's taxonomy: an overview. Theory Pract 2010;41(4):212-218. [doi: 10.1207/s15430421tip4104_2]
29. Biggs JB, Tang CSK. Teaching for Quality Learning at University What the Student Does. 4th ed. Maidenhead: Society for Research into Higher Education; 2011.
30. Abdolreza GZ, Mansourzadeh A, Montazeri KA, Saeidi M. Reflections on using open-ended questions. Med Educ Bull 2022;3(2):475-482. [doi: 10.22034/meb.2022.333518.1054]
31. Haubo R, Christensen B, editors. Cumulative Link Models for Ordinal Regression With the R Package ordinal. Denmark: Christensen Statistics; 2018.
32. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Core Team, R Foundation for Statistical Computing; 2025.
33. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med 2007;4(10):e296 [FREE Full text] [doi: 10.1371/journal.pmed.0040296] [Medline: 17941714]
34. Hoffmann H, Jonietz A, Gräfe W, Zenker R, Voigt K, Riemenschneider H. Associations of an easy-to-understand patient letter on the health literacy of patients after discharge from hospital: results of a randomized controlled intervention study. Gesundheitswesen 2023;85(S 03):S183-S188 [FREE Full text] [doi: 10.1055/a-2130-2374] [Medline: 37751754]
35. Lin R, Gallagher R, Spinaze M, Najoumian H, Dennis C, Clifton-Bligh R, et al. Effect of a patient-directed discharge letter on patient understanding of their hospitalisation. Intern Med J 2014;44(9):851-857. [doi: 10.1111/imj.12482] [Medline: 24863954]
36. Makaryus AN, Friedman EA. Patients' understanding of their treatment plans and diagnosis at discharge. Mayo Clin Proc 2005;80(8):991-994. [doi: 10.4065/80.8.991] [Medline: 16092576]
37. Hodkinson A, Tyler N, Ashcroft DM, Keers RN, Khan K, Phipps D, et al. Preventable medication harm across health care settings: a systematic review and meta-analysis. BMC Med 2020;18(1):313 [FREE Full text] [doi: 10.1186/s12916-020-01774-9] [Medline: 33153451]
38. Al Jeraisy M, Alshammari H, Albassam M, Al Aamer K, Abolfotouh MA. Utility of patient information leaflet and perceived impact of its use on medication adherence. BMC Public Health 2023;23(1):488 [FREE Full text] [doi: 10.1186/s12889-023-15346-y] [Medline: 36918823]
39. Coppa K, Kim EJ, Oppenheim MI, Bock KR, Conigliaro J, Hirsch JS. Examination of post-discharge follow-up appointment status and 30-day readmission. J Gen Intern Med 2021;36(5):1214-1221 [FREE Full text] [doi: 10.1007/s11606-020-06569-5] [Medline: 33469750]
40. McAlister FA, Youngson E, Bakal JA, Kaul P, Ezekowitz J, van Walraven C. Impact of physician continuity on death or urgent readmission after discharge among patients with heart failure. CMAJ 2013;185(14):E681-E689 [FREE Full text] [doi: 10.1503/cmaj.130048] [Medline: 23959284]
41. Baky V, Moran D, Warwick T, George A, Williams T, McWilliams E, et al. Obtaining a follow-up appointment before discharge protects against readmission for patients with acute coronary syndrome and heart failure: a quality improvement project. Int J Cardiol 2018;257:12-15. [doi: 10.1016/j.ijcard.2017.10.036] [Medline: 29506682]
42. Shahid R, Shoker M, Chu LM, Frehlick R, Ward H, Pahwa P. Impact of low health literacy on patients' health outcomes: a multicenter cohort study. BMC Health Serv Res 2022;22(1):1148 [FREE Full text] [doi: 10.1186/s12913-022-08527-9] [Medline: 36096793]
43. Wolderslund M, Kofoed PE, Holst R, Waidtløw K, Ammentorp J. Outpatients' recall of information when provided with an audio recording: a mixed-methods study. Patient Educ Couns 2020;103(1):63-70. [doi: 10.1016/j.pec.2019.08.030] [Medline: 31473043]
44. Laws MB, Lee Y, Taubin T, Rogers WH, Wilson IB. Factors associated with patient recall of key information in ambulatory specialty care visits: results of an innovative methodology. PLoS One 2018;13(2):e0191940 [FREE Full text] [doi: 10.1371/journal.pone.0191940] [Medline: 29389994]
45. Kessels RPC. Patients' memory for medical information. J R Soc Med 2003;96(5):219-222 [FREE Full text] [doi: 10.1177/014107680309600504] [Medline: 12724430]

XSL•FO
RenderX

46. Szulewski A, Howes D, van Merriënboer JJG, Sweller J. From theory to practice: the application of cognitive load theory to the practice of medicine. Acad Med 2021;96(1):24-30. [doi: 10.1097/ACM.0000000000003524] [Medline: 32496287]
47. Xu Q, Sundar SS. Interactivity and memory: Information processing of interactive versus non-interactive content. Comput Hum Behav 2016;63:620-629 [FREE Full text] [doi: 10.1016/j.chb.2016.05.046]
48. Dada A, Koras OA, Bauer M, Butler A, Smith KE, Kleesiek J. MeDiSumQA: patient-oriented question-answer generation from discharge letters. arXiv. Preprint posted online on February 5, 2025 2025. [doi: 10.18653/v1/2025.cl4health-1.10]
49. Baumeister RF, André N, Southwick DA, Tice DM. Self-control and limited willpower: current status of ego depletion theory and research. Curr Opin Psychol 2024;60:101882 [FREE Full text] [doi: 10.1016/j.copsyc.2024.101882] [Medline: 39278166]
50. Cui Y, Li R, Yang T, Wang H, Jin S, Liu N, et al. Influence of positive and negative affect on self-management among patients with early chronic kidney disease during the COVID-19 pandemic: the mediating and suppressing effect of ego depletion. Front Psychiatry 2022;13:992404 [FREE Full text] [doi: 10.3389/fpsyt.2022.992404] [Medline: 36245863]
51. Yumiya Y, Goto A, Konta T. Association between health literacy and understanding of doctors' explanations: the Yamagata study. Health Lit Res Pract 2024;8(3):e175-e183 [FREE Full text] [doi: 10.3928/24748307-20240819-03] [Medline: 39251187]
52. Stormacq C, Wosinski J, Boillat E, Van den Broucke S. Effects of health literacy interventions on health-related outcomes in socioeconomically disadvantaged adults living in the community: a systematic review. JBI Evid Synth 2020;18(7):1389-1469. [doi: 10.11124/JBISRIR-D-18-00023] [Medline: 32813388]
53. Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. Nat Med 2023;29(10):2396-2398. [doi: 10.1038/s41591-023-02412-6] [Medline: 37391665]
54. Gorelik AJ, Li M, Hahne J, Wang J, Ren Y, Yang L, et al. Ethics of AI in healthcare: a scoping review demonstrating applicability of a foundational framework. Front Digit Health 2025;7:1662642 [FREE Full text] [doi: 10.3389/fdgth.2025.1662642] [Medline: 41000336]
55. Park MK, Ashwood N, Capes N. Ethics of artificial intelligence in medicine. Cureus 2025;17(5):e83567. [doi: 10.7759/cureus.83567] [Medline: 40486469]
56. Kong M, Fernandez A, Bains J, Milisavljevic A, Brooks KC, Shanmugam A, et al. Evaluation of the accuracy and safety of machine translation of patient-specific discharge instructions: a comparative analysis. BMJ Qual Saf 2025:bmjqs-2024-018384 [FREE Full text] [doi: 10.1136/bmjqs-2024-018384] [Medline: 40633961]
57. Siegrist V, Mata R, Langewitz W, Gerger H, Furger S, Hertwig R, et al. Does information structuring improve recall of discharge information? A cluster randomized clinical trial. PLoS One 2021;16(10):e0257656 [FREE Full text] [doi: 10.1371/journal.pone.0257656] [Medline: 34662341]
58. Brands MR, Gouw SC, Beestrum M, Cronin RM, Fijnvandraat K, Badawy SM. Patient-centered digital health records and their effects on health outcomes: systematic review. J Med Internet Res 2022;24(12):e43086 [FREE Full text] [doi: 10.2196/43086] [Medline: 36548034]
59. Neves AL, Freise L, Laranjo L, Carter AW, Darzi A, Mayer E. Impact of providing patients access to electronic health records on quality and safety of care: a systematic review and meta-analysis. BMJ Qual Saf 2020;29(12):1019-1032 [FREE Full text] [doi: 10.1136/bmjqs-2019-010581] [Medline: 32532814]
60. Talevski J, Wong Shee A, Rasmussen B, Kemp G, Beauchamp A. Teach-back: a systematic review of implementation and impacts. PLoS One 2020;15(4):e0231350 [FREE Full text] [doi: 10.1371/journal.pone.0231350] [Medline: 32287296]
61. Weetman K, Spencer R, Dale J, Scott E, Schnurr S. What makes a "successful" or "unsuccessful" discharge letter? Hospital clinician and general practitioner assessments of the quality of discharge letters. BMC Health Serv Res 2021;21(1):349 [FREE Full text] [doi: 10.1186/s12913-021-06345-z] [Medline: 33858383]
62. Coit MH, Katz JT, McMahon GT. The effect of workload reduction on the quality of residents' discharge summaries. J Gen Intern Med 2011;26(1):28-32 [FREE Full text] [doi: 10.1007/s11606-010-1465-z] [Medline: 20697967]
63. Carris NW, Bullers K, McKee M, Schanze J, Eubanks T, Epperson C, et al. Automated lifestyle interventions and weight loss: a systematic review of randomized controlled trials. Int J Obes (Lond) 2025;49(6):992-997. [doi: 10.1038/s41366-025-01746-0] [Medline: 40158054]
64. Sajjadi SL, Ghafourifard M, Khosroshahi HT. The effect of individualized education on learning needs of patients undergoing hemodialysis: a randomized controlled clinical trial. BMC Nephrol 2024;25(1):452 [FREE Full text] [doi: 10.1186/s12882-024-03886-3] [Medline: 39695998]
65. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014;96(3):395-403. [doi: 10.1016/j.pec.2014.05.027] [Medline: 24973195]

Abbreviations

AI: artificial intelligence
CLT: cognitive load theory
LLM: large language model

XSL•FO
RenderX

OR: odds ratio
SP: standardized patient
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by J Sarvestan; submitted 26.Jul.2025; peer-reviewed by F Al Dhabbari, Z Vojvodic, K Fiscella; comments to author 01.Sep.2025; revised version received 23.Dec.2025; accepted 24.Dec.2025; published 26.Feb.2026.

Please cite as:
Holderried F, Sonanini A, Stegemann–Philipps C, Herrmann–Werner A, Spitzer P, Guthoff M, Heyne N, Sering K, Holderried M, Eisinger F
Impact of GPT-4–Generated Discharge Letters on Patients' Medical Comprehension: Prospective Crossover Study
J Med Internet Res 2026;28:e81243
URL: https://www.jmir.org/2026/1/e81243
doi:10.2196/81243
PMID:41746691

©Friederike Holderried, Alessandra Sonanini, Christian Stegemann–Philipps, Anne Herrmann–Werner, Philipp Spitzer, Martina Guthoff, Nils Heyne, Konstantin Sering, Martin Holderried, Felix Eisinger. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Use, Utility, and User Experience of Cloud-Based Medical Imaging in Pulmonary Nodule Care in China: Mixed Methods Study

Ziguo Chen[1*], MM; Junhan Wu[1*], MD; Yunying Chen[2*], MM; Weitao Zhuang[3], MD; Jiaxuan Huang[4], MM; Weifeng Zhong[1], MM; Zijie Li[5], MM; Shuhuan Xie[1], MD; Chaofan Liu[1], MM; Guojie Lu[1], MM; Guibin Qiao[1], MD, PhD

[1]Department of Thoracic Surgery, Zhujiang Hospital, Southern Medical University, No. 253 Industrial Avenue Middle, Haizhu District, Guangzhou, China

[2]The Fifth Affiliated Hospital, Sun Yat-sen University, Zhuhai, China

[3]State Key Laboratory of Respiratory Diseases, National Clinical Research Center for Respiratory Disease, National Center for Respiratory Medicine, Department of Pulmonary and Critical Care Medicine, Guangzhou Institute of Respiratory Health, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China

[4]School of Medicine, South China University of Technology, Guangzhou, China

[5]Shantou University Medical College, Shantou University, Shantou, China

[*]these authors contributed equally

Corresponding Author:
Guibin Qiao, MD, PhD
Department of Thoracic Surgery, Zhujiang Hospital, Southern Medical University, No. 253 Industrial Avenue Middle, Haizhu District, Guangzhou, China

Abstract

Background: The detection of pulmonary nodules (PNs) has increased with the use of low-dose computed tomography screening. Effective management requires timely longitudinal surveillance and reliable comparison with prior examinations, yet access to previous imaging across institutions is often fragmented, leading to delays and potentially unnecessary repeat scans and costs. Cloud-based medical imaging (CMI) solutions offer a potential means of improving access and facilitating cross-institutional data exchange. However, the adoption and utility of CMI in PN care, especially in China, remain underexplored.

Objective: This study aims to evaluate the possession, use, and impact of CMI on health care utilization, patient knowledge, and financial burden, as well as to identify usability and interoperability barriers through qualitative investigation.

Methods: A mixed methods cross-sectional study was conducted from October 2022 to May 2024. The study involved 701 patients with PNs who completed structured surveys, and 20 participants (10 patients and 10 physicians) were interviewed. CMI use was defined as self-reported ability to view radiological images on a mobile device. We compared CMI users and nonusers and estimated adjusted odds ratios using multivariable logistic regression, then applied 1:1 propensity score matching to examine associations between CMI use and health care utilization, costs, and patient perceptions, and qualitative interviews were analyzed for usability themes.

Results: The study found that 611 (87.2%) out of 701 patients had obtained CMI, with 404 (57.6%) out of 701 patients actively using it. In multivariable analysis, older age was independently associated with lower CMI use (odds ratios 0.985, 95% CI 0.972 - 0.999). After 1:1 propensity score matching, CMI users accessed more internet hospitals, consulted more physicians, and reported lower health care costs compared to nonusers. Users also demonstrated higher disease knowledge. Qualitative data identified key barriers, including poor system usability, limited retention time for images, and weak interoperability. CMI was perceived as beneficial for patient convenience and clinical efficiency, though concerns over image quality and system fragmentation were prevalent.

Conclusions: While CMI is widely available, its usage remains suboptimal. Increased use is associated with enhanced health care engagement and reduced costs, suggesting that improving system usability and ensuring consistent access to imaging may help realize potential benefits of CMI. Future improvements should focus on ensuring long-term access, better retention protocols, and overcoming interoperability issues.

(J Med Internet Res 2026;28:e86745) doi:10.2196/86745

KEYWORDS

cloud-based medical imaging; pulmonary nodules; health care utilization; telemedicine; qualitative interview

XSL•FO
RenderX

Introduction

With the widespread adoption of low-dose computed tomography screening, the detection rate of pulmonary nodules (PNs) continues to increase [1]. Longitudinal imaging surveillance and comparison with prior examinations are foundational to PN management and are consistently emphasized in major guidelines [2,3]. In routine practice, patients often receive care across multiple institutions and seek additional opinions; when prior images are not retrievable promptly and reliably, repeat scanning and delays in care become likely. Evidence indicates that cross-regional or interinstitutional health information exchange (HIE) is associated with fewer duplicate imaging examinations and lower related costs [4,5].

In recent years, internet hospitals that commonly refer to hospital-affiliated online clinical services that extend routine care pathways through digital platforms, rather than stand-alone telemedicine encounters, have become an important complement to in-person clinics for remote consultation and follow-up. Real-world data suggest that their use is associated with fewer visits and lower costs [6,7], yet such encounters rely even more heavily on HIE. Enabling cross-institutional and cross-channel reuse of imaging requires cloud-based tools built on the Digital Imaging and Communications in Medicine (DICOM) or DICOMweb standards, which are collectively referred to as cloud-based medical imaging (CMI), and they support interoperable data exchange and network retrieval [8]. Recent implementations show that DICOMweb-based workflows can deliver web-based image viewing, improving visit efficiency. Recent surveys of web-based DICOM viewers suggest that DICOMweb support is increasingly common, but implementation depth and performance vary across products and settings [9]. Meanwhile, direct patient access to images is increasingly common. Although this can enhance knowledge and engagement, the absence of patient-facing explanations may trigger or amplify "scanxiety" [10]. Accordingly, information availability and interpretability represent distinct challenges. Recent studies showed that patient engagement with radiology reports and images is heterogeneous with respect to uptake, timing, and depth and differs across patients [11]. When patients review radiology results, they often encounter clinician-oriented terminology with limited general explanation, which may hinder comprehension and contribute to concern even when results are available in a timely manner [12,13]. A recent systematic review reported that patient-oriented reporting features, including plain-language summaries, glossaries, and visual aids, are associated with improved comprehension and satisfaction and may reduce avoidable anxiety [12].

In this imaging-dependent clinical context, timely retrieval and cross-channel reuse of imaging studies are prerequisites for effective longitudinal management. However, empirical evidence specific to populations with PNs remains sparse regarding CMI availability versus actual use, its associations with health care utilization and financial burden, user experience, and interoperability barriers. While digital image accessibility has been studied in general contexts, this research represents the first mixed methods study to focus on the unique PN care landscape in China. In this context, effective management is heavily dependent on long-term, high-frequency longitudinal imaging comparisons, making current limitations, such as brief data retention and system fragmentation, particularly impactful. Furthermore, unlike the unified patient portals common in other regions, the Chinese mobile health ecosystem is characterized by fragmented, QR code–based access, which presents distinct usability challenges. Beyond describing accessibility, this study provides a robust analysis of the correlations between actual CMI use and health care utilization, financial burden, and patient cognition, offering a holistic perspective on the "possession-to-use gap" that has not been previously characterized. This study aimed to characterize, within PN care, the landscape of CMI availability and use; quantify its associations with health care utilization and costs; assess relationships with disease perceptions; and, through qualitative inquiry, identify key usability and interoperability barriers to inform optimization and scale-up.

Methods

Study Design and Participants

We used an explanatory sequential mixed methods design. Survey findings informed the interview guide and purposive sampling, and integration was conducted through a prespecified comparison framework and meta-inferences based on convergence, complementarity, and discrepancies. Integrated findings are reported using narrative weaving that links quantitative results with qualitative themes and illustrative quotes. The reporting of the quantitative component follows the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guideline, and the qualitative component was designed, conducted, and reported with reference to the COREQ (Consolidated Criteria for Reporting Qualitative Research). From October 2022 to May 2024, consecutive patients with PNs were recruited via a thoracic surgery e-consultation clinic on a national telemedicine platform [14]. In total, 701 participants completed the structured survey. Eligibility criteria were newly identified or undersurveillance PNs, age ≥18 years, and ability to complete the questionnaire. We excluded individuals who had already undergone resection or had a definitive pathological diagnosis.

To complement and contextualize the quantitative findings, we conducted semistructured interviews with a purposive, maximum-variation sample of 10 patients and 10 physicians. Patient participants varied by age (26 - 73 y), sex, and educational level. Physician participants were thoracic surgeons spanning junior, mid-career, and senior ranks.

Ethical Considerations

This study was approved by the Research Ethics Committee of Guangdong Provincial People's Hospital (approval number: 2025-KY-122 - 01). All participants provided written informed consent under institutional ethics approval. All questionnaire and interview data were deidentified before analysis, and access was restricted to the research team. The research team collected questionnaire responses and qualitative interview materials; we did not collect, download, store, or process participants' DICOM files or other raw radiology image files as part of the study

XSL•FO
RenderX

procedures. Participants received no compensation for participation.

Measures and Variables

In China, CMI is commonly obtained or shared as DICOM files viewable in a browser, QR codes appended to radiology reports, or links on hospital websites. For QR code or web link access, the code or link typically directs patients to an institution-hosted or vendor-hosted web viewer. Depending on the implementation, the viewer may be DICOM-based (including zero-footprint viewers) or may rely on proprietary web formats; therefore, QR or link access should not be assumed to be DICOMweb-compliant. Some patients receive only a textual report and/or film, with no CMI access. We defined use of CMI as the self-reported ability to view radiological images on a mobile device. This measure does not differentiate between passive access and more active or intensive use, nor does it quantify frequency or depth of use. Sociodemographic and clinical variables included age, sex, education (primary, middle school, high school, junior college, undergraduate, and graduate), and surveillance status. Care-seeking and cost measures comprised the number of hospitals used (internet and offline), the total number of physicians consulted, and annual out-of-pocket expenditures (care-related fees, transportation, lodging). Self-efficacy and disease perceptions included perceived disease severity, level of disease awareness, and change in anxiety level.

Qualitative Interviews: Design, Data Collection, and Management

We developed separate interview guides for patients and physicians (Multimedia Appendix 1). The guides covered CMI awareness and use, preferences of CMI format, perceived clinical benefits and limitations, workflow usability and pain points, psychological responses, and priorities for system improvement. Trained qualitative researchers conducted in-person interviews (20 - 40 min). With permission, sessions were audio-recorded and accompanied by field notes. Recordings were transcribed verbatim, deidentified, and stored on password-protected devices. Texts were managed and queried in NVivo (version 11; QSR International Pty Ltd). Bilingual researchers translated quotations for manuscript presentation with dual review; discrepancies were resolved by revisiting the Chinese original. Additionally, interviews and analysis proceeded iteratively. During concurrent coding, we tracked the emergence of new codes and concluded interviewing when successive interviews no longer generated new major themes and the coding framework had stabilized.

Data Analysis

Quantitative Analysis

Continuous variables were summarized using medians and interquartile ranges, whereas categorical variables were summarized as frequencies and percentages. The primary outcome was the use of CMI, defined as whether participants accessed and viewed their imaging data through CMI. Group comparisons were performed using chi-square or Mann-Whitney *U* tests. To quantify associations with CMI use, odds ratios with 95% CIs were estimated using univariable and multivariable logistic regression. To minimize confounding, propensity score matching was conducted with CMI use as the dependent variable. Prespecified covariates included age, sex, education, and surveillance duration. Nearest-neighbor 1:1 matching was applied using a caliper width of 0.02 times the SD of the logit of the propensity score. Covariate balance was assessed using standardized mean differences (<0.10). In the matched sample, health care utilization, financial burden, and disease perceptions were compared between CMI users and nonusers. Analyses were performed using SPSS Statistics (version 26.0; IBM Corp), and visualizations were created in GraphPad Prism (version 8.0).

Qualitative Analysis and Integration

Two researchers independently reviewed all transcripts, wrote reflexive memos, and conducted line-by-line open coding using in vivo codes where appropriate. After independent coding, the team discussed and resolved discrepancies; merged overlapping codes; refined code boundaries; and compiled a shared, version-controlled codebook. A constant comparative approach was used to group semantically related codes into categories within each participant group and to compare categories across groups using a patient-versus-physician matrix, identifying areas of convergence and divergence. Potentially disconfirming cases were examined to test the robustness of themes, which were iteratively refined based on clarity, coherence, and distinctiveness. Approximately 20% of the transcripts were double-coded for quality assurance, with disagreements resolved through discussion. An auditable analytic trail and reflexive notes were maintained throughout. Member checking was not undertaken. However, integration was performed through using the quantitative results to inform interview focus and an explicit joint display and narrative weaving to connect quantitative patterns with qualitative explanations and identify areas of convergence and discrepancy.

Results

CMI Possession and Use

Overview

Among 701 patients with PNs, 87.2% (n=611) had obtained CMI. Of these, 47.1% (n=330) acquired it by requesting from hospitals, while 40.1% (n=281) received it proactively from hospitals; 12.8% (n=90) had only reports or films without CMI. The main modalities were QR code or web link (n=352, 50.3%) and DICOM files (n=259, 36.9%). Overall, 57.6% (n=404) were CMI users, and 42.4% (n=297) were nonusers (Table 1).

XSL•FO
RenderX

Table . Possession and application of CMI[a] among patients with PNs[b].

Possession categories, application of CMI, and variables	Patients, n (%)
Possession of CMI	
No, I only got radiological report and/or film	90 (12.8)
Radiological reports and films	79 (11.3)
Radiological reports	11 (1.6)
Yes, I got it by applying to the hospital	330 (47.1)
DICOM[c] files	259 (36.9)
Yes, hospitals took initiative to provide it	281 (40.1)
QR code or web links	352 (50.3)
Application of CMI	
Users of CMI	404 (57.6)
Nonusers of CMI	297 (42.4)

[a]CMI: cloud medical images.

[b]PN: pulmonary nodule.

[c]DICOM: Digital Imaging and Communications in Medicine.

Baseline Differences and Independent Correlates of CMI Use

In the unmatched sample (Table 2), users were younger than nonusers (median 46, IQR 38 - 55 y vs median 48, IQR 39 - 57 y; *P*=.001), and distributions differed for education (overall *P*=.004) and prior internet-clinic experience (overall *P*=.03). Sex (*P*=.60) and time since nodule detection (*P*=.09) did not differ. In multivariable logistic regression (Table 3), age was independently and inversely associated with CMI use (odds ratios 0.985, 95% CI 0.972 - 0.999; *P*=.04), while education was not significant after adjustment (*P*=.21).

Table . Comparison of baseline characteristics between users and nonusers of CMI[a].

Variable	Overall (N=701)	Users (n=404, 57.6%)	Nonusers (n=297, 42.4%)	*P* value
Age (y), median (IQR)	48 (39-57)	46 (38-55)	48 (39-57)	.001
Sex, n (%)				.60
Female	435 (62.1)	254 (62.9)	184 (60.9)	
Male	266 (37.9)	151 (37.1)	105 (39.1)	
Education attainment, n (%)				.004
Primary school	24 (3.4)	8 (2.0)	16 (5.4)	
Junior high school	68 (9.7)	41 (10.1)	27 (9.1)	
Senior high school	88 (12.6)	43 (10.6)	45 (15.2)	
Junior college	150 (21.4)	79 (19.6)	71 (23.9)	
Undergraduate	291 (41.5)	181 (44.8)	110 (37.0)	
Graduate	80 (11.4)	52 (12.9)	28 (9.4)	
Online consultation, n (%)				.03
Never	207 (29.5)	99 (24.5)	108 (36.4)	
Ever but need assistance	288 (41.1)	184 (45.5)	104 (35.0)	
Ever and not need assistance	266 (37.9)	121 (30.0)	85 (28.6)	
Time to the first detection of PNs[b] (mo), median (IQR)	12 (3-32)	14 (4-32)	12 (2-33)	.09

[a]CMI: cloud medical images.

[b]PNs: pulmonary nodules.

XSL•FO
RenderX

Table . Binomial logistic regression analysis for influencing factors of CMI[a] application in patients with PNs[b].

Variable	Univariate analyses		Multivariate analysis	
	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Age (y)	0.979 (0.967 - 0.991)	.001	0.985 (0.972 - 0.999)	.04
Time (mo)	0.998 (0.993 - 1.004)	.58	—[c]	—
Sex (reference: female)		.60	—	—
Male	0.921 (0.677 - 1.254)			
Education attainment (reference: primary school)	—	.02	—	.21
Junior high school	3.037 (1.142 - 8.075)	.03	2.655 (0.989 - 7.123)	.05
Senior high school	1.911 (0.742 - 4.922)	.18	1.651 (0.634 - 4.300)	.30
Junior college	2.225 (0.898 - 5.513)	.08	1.833 (0.726 - 4.626)	.20
Undergraduate	3.291 (1.363 - 7.943)	.008	2.449 (0.973 - 6.162)	.06
Graduate	3.714 (1.415 - 9.750)	.008	2.583. (0.929 - 7.178)	.07

[a]CMI: cloud-based medical images.

[b]PN: pulmonary nodule.

[c]Not applicable.

Health Care Utilization, Costs, and Perceptions After Matching

After 1:1 propensity score matching, baseline balance was achieved, the main directional differences observed prematching generally persisted (Table S1 in Multimedia Appendix 2) and the covariate balance before and after propensity score matching persisted in Multimedia Appendix 3 . In the matched cohort, as shown in Figure 1A, the distribution for number of internet hospitals (median 1, IQR 0 - 1 vs median 1, IQR 0 - 1; *P*=.01) used differed significantly between users and nonusers, whereas, as shown in Figure 1B, the number of offline hospitals (median 2, IQR 1 - 2 vs median 1, IQR 1 - 2; *P*=.19) did not differ. The number of physicians consulted (median 3, IQR 1 - 3 vs median 2, IQR 1 - 3; *P*=.02) is also displayed in Figure 1C. CMI users were significantly more likely to spend 0 - 2500 Chinese Yuan (CNY; US $1=6.87 CNY) on treatment and diagnostic costs (229/291, 78.7% vs 201/291, 69.1%; *P*=.004) and on travel and accommodation expenses (249/291, 85.6% vs 229/291, 78.7%; *P*=.03), suggesting they incurred lower expenses compared to nonusers. As presented in Figure 1F-H, regarding disease perceptions, users reported higher levels of disease awareness (overall *P*=.009), whereas perceived disease severity (overall *P*=.97) and anxiety change (overall *P*=.81) did not differ (Figure 1).

XSL•FO
RenderX

Figure 1. Health care utilization, costs, and perceptions after 1:1 propensity score matching. Note that 100% stacked bar charts compare cloud-based medical imaging (CMI) Users and Nonusers on 8 outcomes: (A) number of internet hospitals, (B) number of offline hospitals, (C) number of consulting physicians, (D) treatment and diagnostic costs, (E) travel and accommodation expenses, (F) perceived disease severity, (G) level of disease awareness, and (H) change in anxiety level. Bars show the percentage distribution of categories within each group (cost bands: Chinese Yuan [CNY] 0 - 2500, 2500 - 5000, 5000 - 7500, and 7500 - 10,000; ≥10,000; other category labels as shown in panel legends; all costs reported in CNY were converted to US dollars using an exchange rate of US $1=6.87 CNY). *P* values above panels compare distributions between the groups using 2-sided chi-square tests. Significant differences were observed for internet hospitals (*P*=.01), consulting physicians (*P*=.02), treatment or diagnostic costs (*P*=.004), travel or accommodation expenses (*P*=.03), and disease awareness (*P*=.009); no differences were found for offline hospitals (*P*=.19), perceived severity (*P*=.97), or change in anxiety (*P*=.81). Baseline balance after matching is shown in Table S1 in Multimedia Appendix 2.

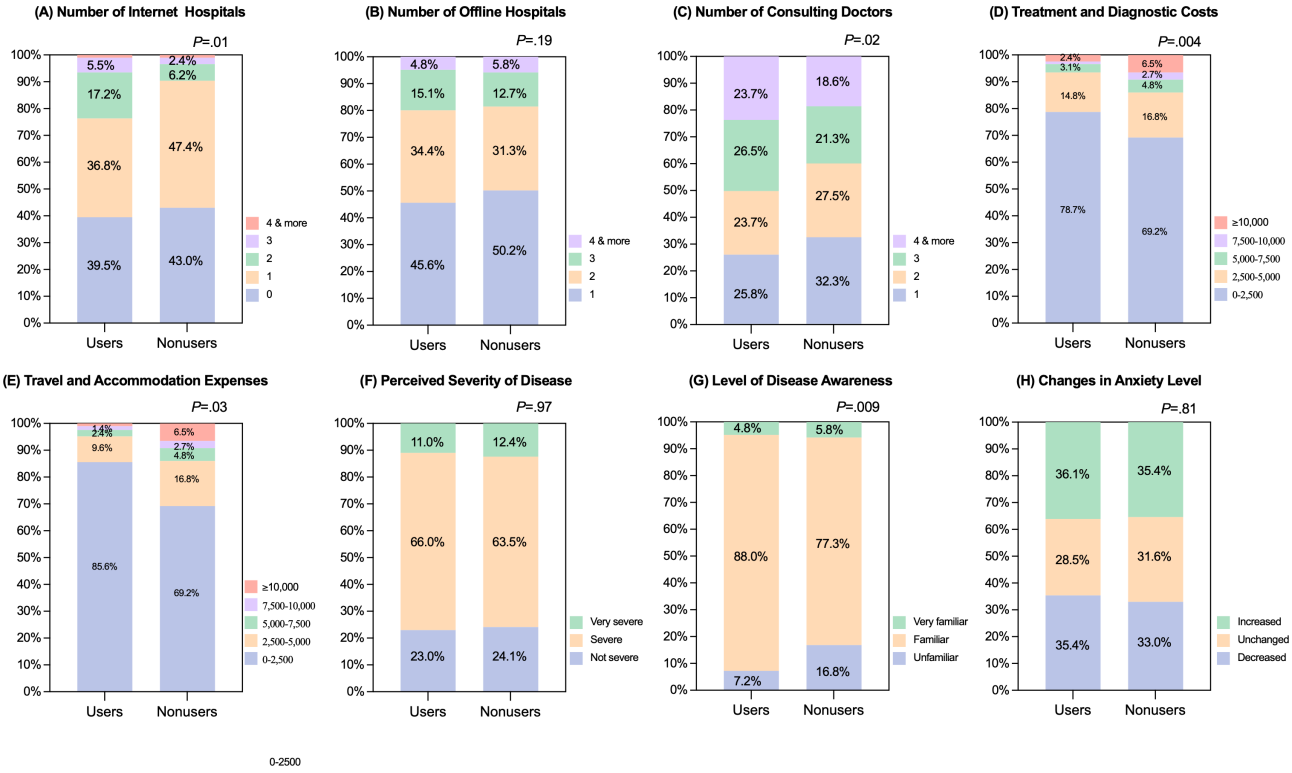

Stratification by Access Modality (DICOM vs QR Code or Web Link) Among Users

Among CMI users, in Figure 2A, the DICOM group differed from the QR or link group in distributions of the number of offline hospitals (median 2, IQR 1 - 2 vs median 2, IQR 1 - 2; *P*=.04) and the number of physicians consulted (median 3, IQR 2 - 4 vs median 3, IQR 1 - 3; *P*=.02) as depicted in Figure 2B, whereas as shown in Figure 2C, the number of internet hospitals (median 1, IQR 0 - 2 vs median 1, IQR 0 - 2; *P*=.99) did not differ. The DICOM group had no difference in treatment and diagnostic costs (95/132, 72.0% vs 209/261, 80.1%; *P*=.09) and travel and accommodation expenses (108/132, 82.6% vs 223/261, 85.4%; *P*=.52) with the majority spending 0 - 2500 CNY, but higher disease knowledge (overall *P*=.004) as presented in Figure 2G. The groups did not differ in perceived severity (overall *P*=.17) or anxiety change (overall *P*=.19; Figure 2).

XSL•FO
RenderX

Figure 2. Outcomes among cloud-based medical imaging (CMI) users stratified by access modality (Digital Imaging and Communications in Medicine [DICOM] vs QR code or web link). Among CMI users only, 100% stacked bar charts compare those accessing the service via DICOM versus QR code or web link across the same 8 outcomes: (A) number of internet hospitals, (B) number of offline hospitals, (C) number of consulting physicians, (D) treatment and diagnostic costs, (E) travel and accommodation expenses, (F) perceived disease severity, (G) level of disease awareness, and (H) change in anxiety level. Bars display percentage distributions for each category (cost bands identical to Figure 1). *P* values above panels indicate 2-sided chi-square tests comparing modalities. Significant differences were seen for offline hospitals (*P*=.04), consulting physicians (*P*=.02), and disease awareness (*P*=.004). No significant differences were observed for internet hospitals (*P*=.99), treatment or diagnostic costs (*P*=.09), travel or accommodation expenses (*P*=.52), perceived severity (*P*=.17), or change in anxiety (*P*=.19).

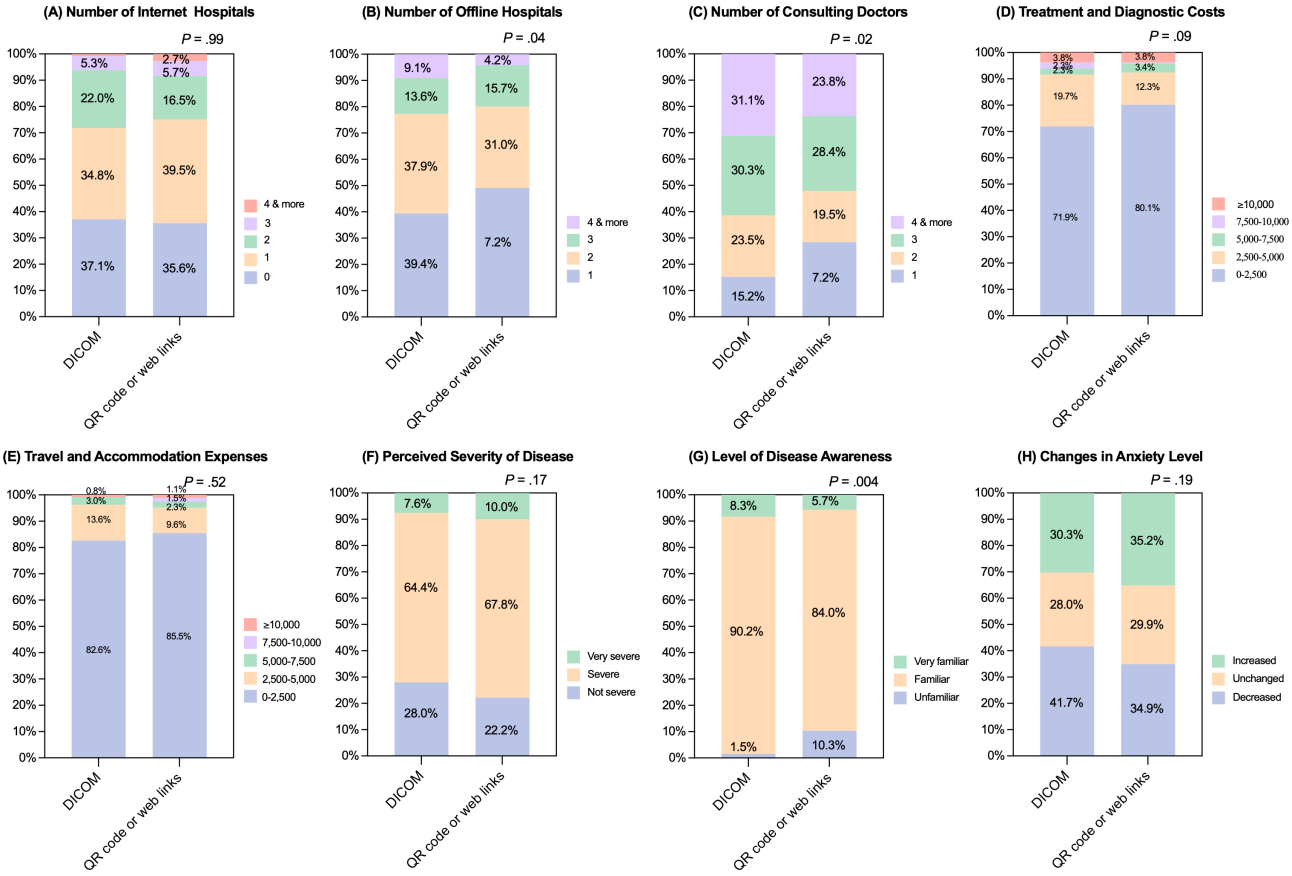

Integrated Findings Across Quantitative and Qualitative Strands

Across both strands, we observed a consistent possession-to-use gap: most participants obtained CMI, but only a subset reported using it. The interview data indicate that this gap reflects not only access but also day-to-day usability barriers and system limitations, including burdensome authentication steps, inconsistent workflows across hospitals, and short retention periods that make access links unreliable. After matching, the observed patterns in health care use and self-reported costs aligned with the convenience and efficiency benefits described in interviews, while psychological responses to viewing images varied in direction. A joint display linking key quantitative results with qualitative themes and illustrative quotes is provided in Table S2 in Multimedia Appendix 4.

Qualitative Themes

Overview

We identified five overarching themes: (1) convenience and clinical value, (2) heterogeneous user experience and system usability, (3) limited retention time and poor interoperability, (4) psychological effects and communication, and (5) improvement priorities and governance. Representative quotes and themes are provided below (Table 4) and in Multimedia Appendix 1.

XSL•FO
RenderX

Table . Themes and representative quotes.

Theme	Representative quotes
Care convenience and clinical value	• "I used to carry film to every visit; now showing the images on my phone is enough." (Patient) • ["When I follow up at another hospital, clinicians can view my prior images, saving many unnecessary steps." (Patient) • "Cloud-based imaging turns pulmonary nodule surveillance from discontinuous checkpoints into a continuous trajectory." (Physician)
Divergent user experience and system usability	• "Different hospitals require different accounts; I often can't remember them." (Patient) • "When the connection is poor, the whole process fails." (Patient) • "There are too many steps to access the images; during busy clinics this disrupts workflow." (Physician) • "Window width and window level are difficult to adjust, which affects assessment of key details." (Physician)
Limited retention and poor interoperability	• "QR codes expire after a while, so at the next visit I have to start over." (Patient) • "Restrictions on the number of downloads are inconvenient." (Patient) • "Lack of interoperability creates information silos, making repeat imaging hard to avoid." (Physician) • "Without unified standards, retrieving outside images feels like assembling a jigsaw." (Physician)
Psychological effects and communication	• "Seeing the imaging changes makes me feel more in control." (Patient) • "There is too much medical terminology I don't understand, which makes me more anxious." (Patient) • "We need to 'translate' images into conclusions and recommendations that patients can understand." (Physician)
Improvement priorities and governance	• "Ideally, a single platform would be usable nationwide." (Patient) • "Prioritize speed, clarity, stability, and privacy first; artificial intelligence (AI) features can come later." (Physician) • "It would be ideal to place history, reports, and images together in one 'case folder.'" (Physician)

Theme 1: Convenience and Clinical Value

Patients and physicians agreed that CMI reduces patient burden and improves clinical efficiency. Patients emphasized not needing to carry films and smoother cross-institution visits; physicians highlighted longitudinal comparison of nodules, fewer unnecessary repeat scans, shorter visit times, and improved access for distant patients.

> ***Patient:*** *I no longer carry a stack of films—showing the images on my phone is enough*
>
> ***Patient:*** *When I switch hospitals, prior images can be pulled up, and it's easier for the doctor to decide on follow-up.*
>
> ***Physician:*** *Having previous images side-by-side makes surveillance decisions more coherent and reduces repeat CTs.*
>
> ***Physician:*** *For out-of-town patients, remote consults with shared images are much more efficient.*

Theme 2: Heterogeneous User Experience and Usability

Experience diverged: younger or more digitally literate patients navigated smoothly, whereas older patients struggled with login, verification, account linking, scanning steps, bandwidth, and device limitations. Physicians cited slow loading, complex workflows, suboptimal image quality, windowing controls, and heterogeneity across hospital platforms.

> ***Patient:*** *Some QR codes open directly; others require registration and verification—too many steps.*
>
> ***Patient:*** *If the network is poor, I just give up.*
>
> ***Physician:*** *When the clinic network slows, image loading disrupts the flow of consultation.*
>
> ***Physician:*** *Different platforms work differently; the learning burden prevents a stable routine.*

Theme 3: Limited Retention and Interoperability Gaps

Short retention periods, expired links or QR codes, download limits, and weak cross-institution interoperability were common pain points. Patients could not retrieve older studies for long-term surveillance; physicians faced siloed platforms and resorted to repeat testing.

> ***Patient:*** *Links stop working after a year or two, making longitudinal comparison difficult.*
>
> ***Patient:*** *Limits on how many times the link can be opened are especially inconvenient.*

XSL•FO
RenderX

> ***Physician:*** *Platform fragmentation leads to information silos and, inevitably, duplicate imaging.*
>
> ***Physician:*** *Without unified interfaces, retrieving outside images is inefficient.*

Theme 4: Psychological Effects and Communication

CMI had bidirectional psychological effects: some patients felt reassured by ready access; others became anxious due to misinterpretation. Physicians noted that clear explanation is essential to avoid overinterpretation and undue worry.

> ***Patient:*** *Seeing my images makes me feel more in control.*
>
> ***Patient:*** *The terminology is confusing; the more I look, the more anxious I feel.*
>
> ***Physician:*** *Without explanations, patients may overread images and worry unnecessarily.*

Theme 5: Improvement Priorities and Governance

Patients and physicians broadly agreed on priorities but with different emphases. Patients wanted longer retention, removal of access limits, simpler flows, and cross-institution sharing. Physicians emphasized speed and clarity, unified standards or platforms, integration of history and reports, clear privacy and accountability, and gradual introduction of artificial intelligence–assisted localization and comparison.

> ***Patient:*** *Ideally, images would be stored long-term without repeated registration.*
>
> ***Physician:*** *We need unified data standards and interfaces, integrating imaging, reports, and history into one workflow.*
>
> ***Physician:*** *Baseline requirements are clarity, loading performance, stability, and privacy; these should be solved first.*

Discussion

Principal Findings

This mixed methods study shows that most hospitals can provide CMI in varied formats, yet a substantial possession-to-use gap persists. Compared with nonusers, CMI users engaged more with internet hospitals and physicians, reported lower care-related and travel or lodging costs, and self-reported higher disease knowledge. Qualitative findings pinpointed system usability, retention, timeliness, and interoperability as principal limitations. Overall, our results align with recent evidence on the accessibility and cost advantages of internet hospitals [6] and the association between interoperability and reduced duplicate imaging [5], offering practical directions for improving workflows.

Older age was independently associated with lower CMI use, whereas education was not significant after adjustment. This aligns with reviews showing that older adults face barriers with digital health tools—account/identity verification complexity, dense information, jargon, input burden, and screen design can all hinder adoption and sustained use [15,16]. Our interviews echoed concrete obstacles, such as multiplatform authentication, link expiry, and loading instability. For this population, the most effective improvements are those that reduce usability barriers rather than simply expanding access. Examples include single sign-on with unified entry points, clinic-side preloading combined with direct viewer access, caregiver proxy accounts, and concise instructional micro-guides, all of which have been shown to enhance eHealth participation among older adults [17-19].

Quantitatively, users interacted with more internet hospitals and physicians, while the number of offline hospitals did not change, suggesting that CMI may improve the efficiency and quality of care pathways rather than simply reducing in-person visits. Lower patient-reported costs for users are consistent with the reduced service frequency and improved accessibility observed with internet hospitals [6]. From a continuity-of-information perspective, integrated longitudinal viewers and HIE platforms reduce repeat imaging and patient burden [5], aligning with our cost-direction findings. Although users reported greater disease knowledge, perceived severity and anxiety did not differ, echoing prior work that immediate online access can increase knowledge but may sustain “scanxiety” in the absence of explanatory support [20]. Qualitative findings further contextualize these patterns by illustrating how CMI facilitates more efficient remote consultations and transforms surveillance into a continuous trajectory. These insights help interpret the observed cost reductions as a direct result of fewer unnecessary repeat scans and improved clinical efficiency during cross-institutional visits. The qualitative theme of limited retention and poor interoperability may also help explain the possession-to-use gap observed in our quantitative results, that even when patients have access to CMI, short retention periods and cross-institution barriers can reduce the practical usefulness of CMI and discourage actual use. Finally, the qualitative theme of bidirectional psychological effects explains why increased knowledge may not lower anxiety, as patients reported feeling more in control while simultaneously feeling confused by complex terminology in the absence of professional translation.

Stratification by access modality indicated advantages for DICOM over QR or link access in engagement and knowledge—plausibly due to DICOM’s interactivity, longer-term retention, and standardization, which facilitate time-series comparison and remote consultation. Technically, mature DICOMweb workflows support browser-based import and zero-footprint viewing suited to both lightweight patient access and professional clinical review [9]. At a systems level, better regional informatics and interoperability correlate with improved health service performance [21].

Our interviews highlighted short retention periods, inconsistent rules, and fragile links and QR codes as major obstacles, which undermine in-clinic retrieval. Consistent with research on patient portals, stable and renewable access points with clear audit trails and proactive expiry reminders are foundational for practical use, especially among older adults [22]. Clinically, immediate simultaneous access to prior images during the visit is critical. Rather than adding new entry points, default patient-centered retrieval organized by patient, chronological timeline, and imaging series, together with online comparison, should be standard capabilities. Workflows based on DICOMweb and zero footprint viewers already enable standardized retrieval and

XSL•FO
RenderX

reliable online comparison without local installation [8], and access to prior images reliably reduces repeat testing and delays [23].

Our findings suggest that interoperability and durable access in CMI are not only technical issues but also governance issues that require consistent rules for minimum retention periods, renewable access, audit trails, and accountability across institutions [24]. China's national health informatization agenda has emphasized standardization and interoperability maturity certification, which provides an actionable foundation to operationalize imaging-specific retention and exchange requirements at scale [25]. Building on recent analyses that identify standards, interoperability, legislation, and compliance as the core foundations of China's digital health governance for the Healthy China 2030 initiative, the clinical management pathway for pulmonary nodule surveillance can be strengthened by aligning platform requirements with these governance elements [24,26]. Finally, equity-oriented implementation is needed to ensure that national digital health strategies translate into broadly accessible longitudinal imaging services rather than widening existing digital divides [26].

Limitations

This cross-sectional design precludes causal inference. Visit counts, costs, and disease perceptions were self-reported and thus subject to recall and reporting bias. Recruitment through online and hospital settings may have overrepresented digitally literate individuals, limiting external validity. Consequently, the 57.6% usage rate likely represents an upper bound for this population, and the lack of regional stratification may mask even steeper barriers in underconnected or rural areas. Therefore, our findings are most applicable to digitally connected patients with PNs who engage in cross-institutional care and remote consultation and should be extrapolated cautiously to offline-only pathways and underserved settings. Although propensity score matching mitigated confounding, residual confounding cannot be excluded. The measure of CMI use does not distinguish passive access from more active use, and we did not capture frequency or depth of use. The qualitative sample was modest and did not include participant checking. Future studies could improve construct validity by incorporating intensity indicators ideally complemented by objective usage logs where feasible. Future multisite sampling, longitudinal, or quasi-experimental studies are also warranted to validate these findings and assess long-term adoption trajectories.

Conclusions

In PN management, CMI is widely available yet remains underused. Actual use was associated with greater engagement in online care, lower related costs, and higher self-reported disease knowledge. Qualitative findings identified 3 actionable constraints—usability thresholds, unstable data retention, and inconsistent clinical viewing functions. Addressing these limitations requires ensuring renewable, auditable, and stable access, together with default support for longitudinal, in-browser comparison.

Acknowledgments

We acknowledge that ChatGPT (OpenAI) was used to assist in language editing and improving the clarity of the manuscript.

Funding

This study was supported by the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0529400 & 2024ZD0529401) and National Key R&D Program of China (2022YFE0133100).

Data Availability

The datasets generated or analyzed during this study are not publicly available due to privacy restrictions and institutional policies on patient data sharing but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: GL, GQ
Formal analysis: JH, JW, WZ, ZL
Funding acquisition: GQ
Investigation: CL, SX, YC, ZC
Methodology: GL, GQ, WZ, ZC
Software: JW, ZC
Supervision: GL, GQ
Resources: GQ
Visualization: WZ, ZC
Writing – original draft: YC, ZC
Writing – review & editing: GL, GQ, JW
All authors had full access to all of the data (including statistical reports and tables) in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
Author GL is the co-corresponding author (phone: 86 18680494527; email: 152537065@qq.com)

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guides.
[DOCX File, 20 KB - jmir_v28i1e86745_app1.docx]

Multimedia Appendix 2
Baseline characteristics for 701 patients in pulmonary nodules (PNs) after propensity score matching (PSM).
[DOCX File, 18 KB - jmir_v28i1e86745_app2.docx]

Multimedia Appendix 3
Covariate balance before and after propensity score matching.
[DOCX File, 29 KB - jmir_v28i1e86745_app3.docx]

Multimedia Appendix 4
Joint display of integrated quantitative and qualitative findings.
[DOCX File, 19 KB - jmir_v28i1e86745_app4.docx]

Checklist 1
COREQ checklist.
[PDF File, 380 KB - jmir_v28i1e86745_app5.pdf]

Checklist 2
STROBE checklist.
[DOC File, 84 KB - jmir_v28i1e86745_app6.doc]

References

1. Gould MK, Tang T, Liu ILA, et al. Recent trends in the identification of incidental pulmonary nodules. Am J Respir Crit Care Med 2015 Nov 15;192(10):1208-1214. [doi: 10.1164/rccm.201505-0990OC] [Medline: 26214244]
2. MacMahon H, Naidich DP, Goo JM, et al. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. Radiology 2017 Jul;284(1):228-243. [doi: 10.1148/radiol.2017161659] [Medline: 28240562]
3. Wood DE, Kazerooni EA, Baum SL, et al. Lung cancer screening, version 3.2018, NCCN clinical practice guidelines in oncology. J Natl Compr Canc Netw 2018 Apr;16(4):412-441. [doi: 10.6004/jnccn.2018.0020] [Medline: 29632061]
4. Jung HY, Vest JR, Unruh MA, Kern LM, Kaushal R, HITEC Investigators. Use of health information exchange and repeat imaging costs. J Am Coll Radiol 2015 Dec;12(12 Pt B):1364-1370. [doi: 10.1016/j.jacr.2015.09.010] [Medline: 26614881]
5. Yuan Y, Price M, Schmidt DF, Ward M, Nebeker J, Pizer S. Integrated health record viewers and reduction in duplicate medical imaging: retrospective observational analysis. JMIR Med Inform 2022 May 20;10(5):e32168. [doi: 10.2196/32168] [Medline: 35594070]
6. Liu Y, Jin H, Yu Z, Tong Y. Impact of internet hospital consultations on outpatient visits and expenses: quasi-experimental study. J Med Internet Res 2024 Nov 11;26:e57609. [doi: 10.2196/57609] [Medline: 39527807]
7. Wang K, Zou W, Lai Y, et al. Accessibility, cost, and quality of an online regular follow-up visit service at an internet hospital in China: mixed methods study. J Med Internet Res 2024 Oct 21;26:e54902. [doi: 10.2196/54902] [Medline: 39432365]
8. Genereaux BW, Dennison DK, Ho K, et al. DICOMweb™: background and application of the web standard for medical imaging. J Digit Imaging 2018 Jun;31(3):321-326. [doi: 10.1007/s10278-018-0073-z] [Medline: 29748852]
9. Pereira H, Romero L, Miguel Faria P. Web-based DICOM viewers: a survey and a performance classification. J Imaging Inform Med 2025 Jun;38(3):1304-1322. [doi: 10.1007/s10278-024-01216-5] [Medline: 39349783]
10. Norris EC, Halaska C, Sachs PB, Lin CT, Sanfilippo K, Honce JM. Understanding patient experiences, opinions, and actions taken after viewing their own radiology images online: web-based survey. JMIR Form Res 2022 Apr 25;6(4):e29496. [doi: 10.2196/29496] [Medline: 35468086]
11. Wang J, Goldberg JE, Block T, et al. Patterns of access to radiology reports and images through a patient portal. J Imaging Inform Med 2024 Apr;37(2):504-509. [doi: 10.1007/s10278-024-00996-0] [Medline: 38315344]
12. van der Mee FAM, Ottenheijm RPG, Gentry EGS, et al. The impact of different radiology report formats on patient information processing: a systematic review. Eur Radiol 2025 May;35(5):2644-2657. [doi: 10.1007/s00330-024-11165-w] [Medline: 39545980]

XSL•FO
RenderX

13. Chau M. Alone with the diagnosis: a reflective analysis on imaging report access and emotional burden. Radiography (Lond) 2025 Dec;31 Suppl 2:103104. [doi: 10.1016/j.radi.2025.103104] [Medline: 40764216]
14. HaoDF Online [Website in Simplified Chinese]. URL: https://www.haodf.com [accessed 2026-03-11]
15. Ahmad NA, Mat Ludin AF, Shahar S, Mohd Noah SA, Mohd Tohit N. Willingness, perceived barriers and motivators in adopting mobile applications for health-related interventions among older adults: a scoping review. BMJ Open 2022 Mar 9;12(3):e054561. [doi: 10.1136/bmjopen-2021-054561] [Medline: 35264349]
16. Lemos M, Henriques AR, Lopes DG, et al. Usability and utility of a mobile app to deliver health-related content to an older adult population: pilot noncontrolled quasi-experimental study. JMIR Form Res 2024 May 17;8:e46151. [doi: 10.2196/46151] [Medline: 38758585]
17. Grossman LV, Masterson Creber RM, Benda NC, Wright D, Vawdrey DK, Ancker JS. Interventions to increase patient portal use in vulnerable populations: a systematic review. J Am Med Inform Assoc 2019 Aug 1;26(8-9):855-870. [doi: 10.1093/jamia/ocz023] [Medline: 30958532]
18. Turner K, Clary A, Hong YR, Alishahi Tabriz A, Shea CM. Patient portal barriers and group differences: cross-sectional national survey study. J Med Internet Res 2020 Sep 17;22(9):e18870. [doi: 10.2196/18870] [Medline: 32940620]
19. Wolff JL, Dukhanin V, Burgdorf JG, DesRoches CM. Shared access to patient portals for older adults: implications for privacy and digital health equity. JMIR Aging 2022 May 4;5(2):e34628. [doi: 10.2196/34628] [Medline: 35507405]
20. Derry-Vick HM, Heathcote LC, Glesby N, et al. Scanxiety among adults with cancer: a scoping review to guide research and interventions. Cancers (Basel) 2023 Feb 22;15(5):1381. [doi: 10.3390/cancers15051381] [Medline: 36900174]
21. Ye A, Deng Y, Li X, Shao G. The impact of informatization development on healthcare services in China. Sci Rep 2024 Dec 28;14(1):31041. [doi: 10.1038/s41598-024-82268-z] [Medline: 39730846]
22. Son EH, Nahm ES. Adult patients' experiences of using a patient portal with a focus on perceived benefits and difficulties, and perceptions on privacy and security: qualitative descriptive study. JMIR Hum Factors 2023 Jul 25;10:e46044. [doi: 10.2196/46044] [Medline: 37490316]
23. Bailey JE, Pope RA, Elliott EC, Wan JY, Waters TM, Frisse ME. Health information exchange reduces repeated diagnostic imaging for back pain. Ann Emerg Med 2013 Jul;62(1):16-24. [doi: 10.1016/j.annemergmed.2013.01.006] [Medline: 23465552]
24. Wang M, Lu X, Du Y, et al. Digital health governance in China by a whole-of-society approach. NPJ Digit Med 2025 Aug 2;8(1):496. [doi: 10.1038/s41746-025-01876-9] [Medline: 40750663]
25. Li C, Xu X, Zhou G, et al. Implementation of national health informatization in China: survey about the status quo. JMIR Med Inform 2019 Feb 21;7(1):e12238. [doi: 10.2196/12238] [Medline: 30789350]
26. Wu G, Gong M, Wu Y, Liu L, Shi B, Zeng Z. Advancing digital health in China: aligning challenges, opportunities, and solutions with the Global Initiative on Digital Health (GIDH). Health Care Sci 2024 Oct;3(5):365-369. [doi: 10.1002/hcs2.118] [Medline: 39479272]

Abbreviations

CMI: cloud-based medical imaging
CNY: Chinese Yuan
DICOM: Digital Imaging and Communications in Medicine
HIE: health information exchange
PN: pulmonary nodule
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Stone; submitted 29.Oct.2025; peer-reviewed by O Akinpeloye, T Jimoh; revised version received 25.Feb.2026; accepted 25.Feb.2026; published 30.Mar.2026.

Please cite as:
Chen Z, Wu J, Chen Y, Zhuang W, Huang J, Zhong W, Li Z, Xie S, Liu C, Lu G, Qiao G
Use, Utility, and User Experience of Cloud-Based Medical Imaging in Pulmonary Nodule Care in China: Mixed Methods Study
J Med Internet Res 2026;28:e86745
URL: https://www.jmir.org/2026/1/e86745
doi:10.2196/86745

© Ziguo Chen, Junhan Wu, Yunying Chen, Weitao Zhuang, Jiaxuan Huang, Weifeng Zhong, Zijie Li, Shuhuan Xie, Chaofan Liu, Guojie Lu, Guibin Qiao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The WONE Index as a Multidimensional Assessment of Stress Resilience: A Development and Validation Study

Lydia Genevieve Roos[1], PhD; Destiny Gilliland[1,2], BA; Kelsey Julian[1,3], MA; Reeva Misra[1], MA

[1]Walking on Earth, Ltd, International House 36-38 Cornhill, London, United Kingdom

[2]Department of Biobehavioral Health, College of Human Health and Development, The Pennsylvania State University, University Park, PA, United States

[3]Health Psychology PhD Program, College of Humanities and Earth and Social Sciences, University of North Carolina at Charlotte, Charlotte, NC, United States

Corresponding Author:
Lydia Genevieve Roos, PhD
Walking on Earth, Ltd, International House 36-38 Cornhill, London, United Kingdom

Abstract

Background: Stress resilience is a dynamic process shaped by the interaction between demands and adaptive resources. Existing measures assess stress and resilience as separate constructs, limiting their use in digital health and workplace interventions. An integrated measure capturing both domains is needed.

Objective: We developed and validated the WONE Index, a multidimensional stress resilience tool designed to measure both current stress load and adaptive resources among full-time working adults.

Methods: We developed the 32-item WONE Index through literature review, expert consultation, and iterative refinement to assess stress load and resilience resources across behavioral, cognitive, and social domains. Phase 1 (N=1005; United States– or United Kingdom–based full-time employees) evaluated the initial item pool using exploratory and confirmatory factor analyses to establish the preliminary factor structure and assess reliability and validity. Phase 2 (N=306; United States–based adults) expanded underperforming domains, refined items, and tested incremental validity, test-retest reliability, and measurement invariance. Data were collected online through CloudResearch (Connect) and Prolific (Prolific Academic Ltd) using secure survey platforms.

Results: Phase 1 supported a 2-domain structure: a Stress Load factor (Work Stress, Personal Stress, and Burnout) and a Resilience Resources factor (Emotion Regulation and Coping, Social Connectedness, and Sleep). Model fit indices were excellent (comparative fit index, CFI=0.95; Tucker-Lewis index, TLI=0.94; and root mean square error of approximation, RMSEA=0.05). Phase 2 replicated and extended this structure, expanding Resilience Resources into 7 domains (adding Purpose and Prosociality, Physical Activity, Dietary Intake, and Perseverative Thinking). Confirmatory factor analyses supported a 2-domain structure, comprising a higher-order Stress Load factor with 3 subdomains (Work Stress, Personal Stress, and Burnout) and a higher-order Resilience Resources factor with 7 subdomains (Emotion Regulation and Coping, Social Connectedness, Purpose and Prosociality, Sleep, Physical Activity, Dietary Intake, and Perseverative Thinking). The Stress Load model demonstrated excellent fit (χ^2_{33}=64.18; P=.01; CFI=0.99; TLI=0.98; RMSEA=0.06; and standardized root mean square residual=0.05), and the Resilience Resources model also fit well (χ^2_{443}=745.20, P<.001; CFI=0.94; TLI=0.94; RMSEA=0.05; and standardized root mean square residual=0.06). All subscales showed strong internal consistency (composite reliability: mean 0.84, SD 0.10; range 0.74 - 0.93) and excellent test-retest reliability over 3 weeks (intraclass correlation coefficients 0.77 - 0.90, 95% CI 0.87-0.93). The index showed strong convergent validity (r=0.73 with Connor-Davidson Resilience Scale and r=–0.66 with Perceived Stress Scale-4) and explained additional variance beyond established measures in predicting depression, anxiety, and well-being (ΔR^2=0.07 - 0.11; P<.001).

Conclusions: The WONE Index provides a psychometrically robust tool for assessing stress resilience capacity in working adults. Its integrated structure captures dynamic relationships between stress exposure and resilience resources, thereby supporting personalized intervention delivery in digital health platforms and organizational well-being programs.

(J Med Internet Res 2026;28:e81714) doi:10.2196/81714

KEYWORDS

psychological stress; occupational stress; psychological resilience; occupational health; psychometrics; factor analysis; measurement validation; instrumentation

XSL•FO
RenderX

Introduction

Background

Stress exposure is a pervasive challenge affecting individuals across work, personal, and societal domains, with well-documented impacts on psychological functioning, physical health, and quality of life [1,2]. While stress is common, individuals vary dramatically in their capacity to maintain well-being and adapt effectively under adversity—a quality captured by the concept of stress resilience [3-5]. As digital platforms increasingly seek to deliver personalized mental health support at scale, the ability to comprehensively assess individual resilience capacity has become critical [6,7]. However, existing resilience measures face fundamental limitations: they typically assess stress and resilience as separate constructs, provide insufficient detail to guide personalized interventions, and fail to capture the temporal dynamics through which resilience operates [8]. Addressing these measurement gaps is essential for advancing both resilience science and the development of effective digital mental health interventions across diverse contexts.

Resilience as a Dynamic, Multifaceted Process

We define stress resilience as the dynamic capacity to effectively respond to, recover from, and adapt in the face of adversity while maintaining psychological well-being and functional capacity [9-11]. Rather than reducing resilience to a fixed trait or the simple absence of pathology, this view conceptualizes it as an active and multifaceted process [9]. This process emerges from the interaction between current stress exposure and available protective resources or vulnerabilities, spanning cognitive, behavioral, social, affective, and physiological domains [4,5].

Job Demands-Resources Framework

This conceptualization aligns with the well-established job demands-resources (JD-R) model, which posits that psychological well-being results from the balance between demands that require sustained effort and resources that help manage those demands [12,13]. While originally developed for occupational contexts, this framework has been successfully extended to general stress and resilience processes, demonstrating that the demands-resources balance operates across life domains [14].

When demands exceed available resources, individuals experience strain and potential burnout. When resources are sufficient to meet demands, individuals can maintain well-being and even experience growth. This framework recognizes that resilience capacity is fundamentally determined by 2 interrelated but distinct domains: Stress Load (current psychological demands that require sustained effort) and Adaptive Resilience Resources (capacities that buffer demands and facilitate recovery).

Resilience Resources encompass the behavioral, cognitive, affective, and social capacities that enable adaptive responding to demands and facilitate recovery from adversity. These include emotion regulation skills, social support networks, physical health behaviors, cognitive flexibility, and meaning-making abilities. These resources function as protective factors against stress-related impairment [11]. Critically, many of these resources are modifiable through intervention, making them valuable targets for resilience-building efforts.

Having established the resource component of the JD-R framework, we now turn to the demands side of this balance, focusing on perceived stress. The JD-R framework aligns closely with the transactional model of stress [15-17], which emphasizes that the subjective appraisal of situational demands, rather than their objective frequency or intensity, is a strong determinant of psychological outcomes. We thus emphasize perceived stress rather than only objective stressors because individual appraisals of situational demands are more predictive of well-being and adaptation than event counts [16,17]. Within this perspective, cognitive appraisal serves as a regulatory mechanism that shapes whether a demand functions as a challenge that mobilizes engagement or as a hindrance that accelerates depletion.

The assessment of perceived stress presents both opportunities and challenges for resilience measurement. Perceived stress captures the subjectively experienced burden that directly impacts psychological functioning, making it more predictive of mental health outcomes than objective stressor inventories [18]. However, stress perceptions are themselves influenced by available resilience resources—individuals with stronger emotion regulation skills, social support, or coping strategies may appraise the same objective situation as less threatening [16,17]. This creates a dynamic relationship where resilience resources both protect against stress impact and influence stress perceptions themselves. Rather than viewing this complex relationship as a hindrance to accurate measurement, our framework recognizes this interdependence as fundamental to understanding stress resilience capacity.

Bringing these elements together, the demands-resources framework explains how stress resilience emerges from the interaction between these domains. High current demands deplete available psychological resources and can overwhelm coping capacity. On the other hand, strong resilience resources buffer against demand-related impact and facilitate faster recovery [14,19]. This dynamic interaction means that resilience capacity cannot be accurately assessed by measuring either demands or resources alone; comprehensive evaluation requires understanding both current demands (stress load) and available adaptive capacities (resilience resources).

However, a critical limitation remains in the classic JD-R model: it does not specify the temporal dynamics of these processes—the timescales over which demands erode resources or resources recover. This limitation is consequential for measurement. Because resilience represents an evolving balance rather than a fixed trait, capturing resilience capacity requires understanding both someone's current demands and resources as well as how that balance changes over time. An individual experiencing high demands with adequate resources today may show progressive depletion over subsequent weeks, while another may maintain equilibrium or recover capacity despite similar initial states.

XSL•FO
RenderX

The Need for New Measurement

The dynamic, interconnected nature of stress resilience processes creates significant measurement challenges. Traditional approaches that assess stress and resilience as separate, static constructs fail to capture the fundamental demands-resources interactions that determine resilience capacity [9]. This measurement gap is particularly problematic in digital mental health contexts. Assessment tools must simultaneously provide accurate evaluation of current psychological states and generate actionable insights for personalized intervention delivery.

Current measures have 3 critical limitations that prevent them from capturing these complex dynamics. First and most fundamentally, they fail to simultaneously assess both current perceived stress demands and available protective resources within the same framework. Traditional approaches measure either stress or resilience in isolation, missing the critical interplay between demands and adaptive capacities that determines whether individuals can maintain equilibrium, experience depletion, or build capacity. This separation makes it difficult to understand how an individual's current psychological burden relates to their available coping capacities, limiting the ability to identify whether intervention should focus on stress reduction, resource building, or both.

Second, among measures that assess resilience, most are either too general or too narrow to guide personalized intervention. General measures, such as the Connor-Davidson Resilience Scale (CD-RISC [20]), effectively assess overall resilience characteristics but do not provide the detailed mapping of specific modifiable capacities—such as emotion regulation skills, social support quality, sleep patterns, or physical activity levels—that practitioners and digital platforms need to develop targeted intervention plans. Conversely, aspect-specific measures, such as the Brief Resilience Scale (BRS [21]), assess particular processes, such as stress recovery, effectively but do not capture the full range of behavioral, cognitive, and social factors that contribute to overall resilience capacity.

Third, existing measures do not adequately capture the temporal dynamics of resilience processes. As established earlier, resilience represents an evolving balance rather than a fixed trait, yet current measures typically provide only single-time-point snapshots rather than enabling tracking of how the balance between demands and resources changes over time.

Digital mental health platforms require assessment tools that address all 3 limitations by bridging the gap between comprehensive scientific measurement and practical intervention guidance. Such measures must simultaneously assess both stress and resources within a unified framework, provide sufficient detail to guide personalized interventions across multiple modifiable domains, and enable tracking of resilience trajectories over time through repeated administration. Critically, these tools must also be optimized for digital delivery contexts—brief enough for repeated mobile administration without user burden, structured to enable automated scoring and feedback, and designed to inform real-time algorithmic personalization of intervention content. The need for theoretically grounded measures meeting these criteria has become increasingly urgent as digital platforms seek to provide effective, scalable mental health interventions that address the dynamic nature of stress resilience capacity.

To address these measurement and intervention challenges, we developed the Walking on Earth (WONE; Walking on Earth, Ltd) Index as the assessment foundation for WONE, a digital stress resilience platform. Unlike traditional assessment tools developed for paper-and-pencil or clinical interview administration, the index was specifically designed for digital delivery via the WONE platform using mobile-optimized item presentation, response formats suited to digital interfaces, and brevity enabling repeated monthly assessment without user fatigue.

Two-Phase Validation Strategy

The validation of the WONE Index used a systematic 2-phase approach designed to address the complexity of developing a theoretically grounded yet multidimensional measure. This strategy allowed empirical findings from Phase 1 to inform the refinement of Phase 2.

Phase 1 served as an exploratory foundation study with four primary aims: (1) establishing the initial factor structure through traditional psychometric approaches, (2) identifying items and domains with adequate psychometric properties, (3) assessing convergent and criterion validity with established measures, and (4) revealing measurement challenges requiring refinement to inform Phase 2.

Phase 2 functioned as a confirmatory refinement phase with four key objectives: (1) implementing refined measurement models based on Phase 1 findings, (2) addressing identified measurement limitations through expanded item development, (3) establishing comprehensive psychometric properties, including temporal stability, and (4) demonstrating incremental validity beyond existing gold-standard measures.

Methods

Methods (Phase 1)

Participants

Participants (N=1005; 502 in the United States, and 503 in the United Kingdom) were recruited through CloudResearch (US sample) and Prolific (UK sample) using nonprobability quota sampling. Inclusion criteria included being aged 18 - 65 years and working full-time (because the WONE user base targets adults working full-time), living in the United States or the United Kingdom (as the WONE platform currently operates primarily with companies based in these regions), and being fluent in English. We set quotas for gender within the CloudResearch and Prolific platforms to ensure adequate representation, such that neither men nor women are considered less than 45% of the study sample. Power analyses are provided in the "Data Analysis" section. Participants were balanced by gender (51% women in the United States and 48% in the United Kingdom), predominantly White (71% in the United States and 70% in the United Kingdom), with a mean age of 37.2 (SD 9.6) years in the United States and 36 (SD 10.4) years in the United

XSL•FO
RenderX

Kingdom. Complete demographic characteristics for Phase 1 are provided in Table S1 in Multimedia Appendix 1 [22-24].

Procedures

All study procedures were conducted remotely using Qualtrics (Qualtrics, LLC) survey software. Participants accessed the survey through secure links distributed via CloudResearch (US sample) and Prolific (UK sample). After providing informed consent, participants completed the WONE Index and validation measures in a single sitting (approximately 20 - 30 minutes). The survey was compatible with desktop and mobile devices. Built-in attention checks were included to ensure data quality. Participants who were unable to complete the survey in 1 session, except those who reported technical difficulties, or who did not pass at least 75% of attention checks, were disqualified. Participants were compensated for their time with US $6.25 or £4.50 for the US and UK samples, respectively.

Initial Item Development

The WONE Index item pool was developed through a targeted literature review and iterative expert consultation. We focused the literature review on identifying core constructs empirically linked to stress adaptation and resilience, drawing from theoretical and measurement work in stress appraisal, coping, self-regulation, social support, health behaviors, and well-being. We synthesized findings to determine the primary domains of adaptive capacity most consistently associated with mental health and performance outcomes.

Four health psychologists with expertise in stress and resilience then identified the constructs to be represented across both stress load and resilience resource domains, drawing on existing measures of perceived stress, burnout, resilience, coping, and health behaviors. In addition, we created new items to address domains not captured in previous tools, particularly those relevant to digital health and workplace contexts.

Through several collaborative meetings, the experts reviewed empirical evidence, debated conceptual boundaries, and progressively refined a broad initial pool into a smaller set of candidate items. Two specialists in digital health survey design then reviewed these items to evaluate clarity, readability, and redundancy for digital administration, after which additional consolidation produced the final 32-item instrument encompassing both current stress experiences and resilience resources.

Rather than imposing a fixed structure, the item pool was intentionally designed to allow empirical testing of whether stress load- and resource-related elements would remain distinct, overlap, or converge into higher-order domains. This theoretically grounded yet empirically flexible approach reflects our view of resilience as a dynamic, multidimensional construct rather than a static trait.

Current Stress Experiences

We tested 10 items to assess current stress load and burnout. For current stress, we assessed the following constructs: perceived stress, anxiety, and overwhelm at work and perceived stress, anxiety, and overwhelm at home/in one's personal life. Our approach to stress measurement for perceived work and personal stress acknowledges that individuals conceptualize and express psychological strain differently in real-world contexts [25,26]. Rather than relying on a single "stress" indicator, we assessed 3 related but distinct constructs—stress, anxiety, and overwhelm—within both work and personal life domains.

Research demonstrates that individuals vary considerably in how they conceptualize and articulate psychological distress, with significant cultural and individual differences in whether experiences are described as "stress," "anxiety," "overwhelm," or other terms [25,27]. Although academic literature often treats these as separate constructs, everyday users may not make such distinctions when describing their experiences [28]. Cultural, educational, and personal factors fundamentally influence how psychological experiences are articulated and understood [29,30].

For example, some individuals may more readily identify with feeling "overwhelmed" by their responsibilities, while others may describe similar experiences as "stress" or "anxiety." By capturing all 3 dimensions, the WONE Index ensures a comprehensive assessment regardless of how someone naturally conceptualizes their distress, while also reflecting the interconnected nature of these stress-related experiences in daily life [31]. This inclusive measurement strategy maximizes the tool's utility across diverse user populations and contexts.

We also included 4 indicators of burnout, which assessed cynicism, disengagement, reduced productivity, and mental exhaustion related to work, reflecting the core dimensions of occupational burnout as conceptualized in the Maslach Burnout Inventory framework [32,33]. These burnout indicators capture chronic workplace stress that differs qualitatively from acute stress experiences.

Resilience Resources

Resilience resources represent the psychological, social, and behavioral capacities that enable individuals to adapt to stress, recover from strain, and maintain well-being. This domain reflects the resources side of the WONE Index framework, balancing the stress-load indicators as described in the "Current Stress Experiences" section above.

Altogether, we tested 22 items to measure resilience resources. Nine emotion regulation and coping items consisted of emotion regulation, perceived ability to cope, tendency to bounce back after a stressor, adaptability, stress-related growth mindset, effective coping techniques, perspective-taking ability, frequency of positive affect, and perceived life meaning and purpose. Three self-efficacy and perceived control items assessed: self-efficacy beliefs, perceived control, and the sense that things were going smoothly. Two social support items assessed: trusted support availability and support satisfaction. Four sleep and energy items assessed: sleep quality, duration, latency, disturbances, and overall energy levels. Two dietary intake items included alcohol consumption and nutritious diet quality. Finally, 2 physical activity items assessed moderate and vigorous physical activity and sedentary time.

XSL•FO
RenderX

Response Formats

Items used varying response formats optimized for each domain, all on 1 - 5 point scales with domain-appropriate anchors, where higher numbers indicate better resilience. For example, response options included frequency scales (eg, "Never" to "Always"), agreement scales (eg, "Strongly Disagree" to "Strongly Agree"), and intensity scales (eg, "Not at all" to "Extremely"), selected based on item content and established measurement practices for each construct.

The WONE Index is organized into sections that pair construct-specific introductory instructions with domain-appropriate response formats. For example, stress-related items begin with "In the past month, how often have you felt…," burnout items use agreement anchors, and health behavior items use categorical duration or quality scales. These variations are intentional, aligning response structure with the theoretical nature of each construct to ensure interpretability across diverse domains of stress and resilience. All items used a standardized 1-month timeframe (eg, "Over the past month..." or "In an average week over the past month…") to ensure consistency, capture relatively stable patterns, and remain sensitive to change.

External Validation Measures

Validated external scales were administered to assess criterion validity across stress, resilience, and mental health domains. These measures were used in both studies 1 and 2.

Stress Assessment

Perceived stress was assessed using the Perceived Stress Scale-4 (PSS-4 [18]), a brief version of one of the most widely used instruments for measuring stress perception [34]. The PSS-4 has demonstrated acceptable internal consistency (α=.60-.82 [34]) and good test-retest reliability over 4 weeks (r=0.73 [35]). It also shows strong convergent validity with measures of distress and discriminant validity from indicators of well-being [18]. The PSS-4 was selected for its brevity and suitability for rapid data collection while maintaining adequate psychometric robustness.

Resilience Assessment

Two established measures were used to assess resilience. The 10-item CD-RISC-10 [36] measures the ability to cope with adversity and demonstrates excellent internal consistency (α=.85-.92) as well as good convergent and discriminant validity across populations [36]. Although test-retest reliability was not re-examined for the 10-item version, the original 25-item CD-RISC shows strong temporal stability (r=0.87 [20]).

The 6-item BRS [21] measures the ability to recover or "bounce back" from stress and demonstrates high internal consistency (α=.80-.91), 1-month test-retest reliability (r=0.69), and robust convergent and discriminant validity with measures of affect, optimism, and neuroticism [21].

Mental Health and Well-being Outcomes

Depressive symptoms were measured using the Patient-Reported Outcomes Measurement Information System Short Form-8a (PROMIS-SF-8a [37]), which demonstrates excellent internal consistency (α>.90) and robust convergent validity with legacy measures such as the Patient Health Questionnaire-9 (PHQ-9) and Center for Epidemiologic Studies Depression (CES-D) [38] scales. Although test-retest data are not yet available for the fixed 8-item form, the PROMIS Depression computerized adaptive test shows high temporal stability (intraclass correlation coefficient [ICC]=0.86 [39]).

Anxiety was assessed with the Generalized Anxiety Disorder-7 (GAD-7 [40]), which has high internal consistency (α=.89-.92), good test-retest reliability (r=.83), and robust convergent and discriminant validity relative to other anxiety and depression scales [41].

Well-being was measured with the World Health Organization-5 Well-Being Index (WHO-5 [42]), which shows good internal consistency (α=.80-.90) and strong 5-day test-retest reliability (ICC=0.87 [43]), as well as robust convergent validity with measures of life satisfaction and positive affect [42].

Data Analysis

Statistical Software

Statistical analyses were conducted using IBM SPSS Statistics (version 29.0). Confirmatory factor analyses were performed using IBM SPSS AMOS (version 22.0).

Data Quality, Screening, and Analytic Assumptions

Participants' data were excluded from analyses if their responses to qualitative screening questions were nonsensical or irrelevant to the questions asked, as these were indicators of either automated responses or insufficient English fluency for valid participation. All survey items were required responses, resulting in no missing data for the final analytic sample.

Descriptive statistics and normality assessments were conducted for all key variables. Skewness and kurtosis values were all within the acceptable range of ±1.0, indicating reasonably symmetric distributions without problematic tail behavior. While Shapiro-Wilk tests indicated significant departures from normality for all variables (all Ps<.001), this is expected given the large sample sizes, as these tests become overly sensitive to minor deviations with substantial samples [44]. Visual inspection of histograms and Q-Q plots confirmed that distributions were approximately normal with no severe violations.

Outlier analysis using boxplot methods identified a small number of extreme cases across variables (1 - 11 outliers per variable, representing approximately 1% - 4% of cases). These outliers were retained as they represented valid responses within the expected range of the constructs and did not appear to result from data entry errors. Given the acceptable skewness and kurtosis values, robust nature of maximum likelihood estimation, and large sample sizes, the data were deemed appropriate for the planned factor analyses and correlational analyses.

Analytic Procedures

Exploratory Factor Analysis

Exploratory factor analysis (EFA) was conducted using principal axis factoring with Promax rotation to identify the underlying

XSL•FO
RenderX

factor structure. Factor retention was determined using multiple criteria: (1) eigenvalues >1.0 (Kaiser criterion), (2) scree plot inspection, (3) theoretical interpretability, and (4) total variance explained. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett test of sphericity were used to assess data appropriateness for factor analysis. We systematically evaluated items and retained them based on the following criteria: primary factor loadings ≥0.40, cross-loadings <0.30, difference between primary and secondary loadings ≥0.15, communalities ≥0.30, and theoretical coherence with factor interpretation.

Confirmatory Factor Analysis

We conducted a higher-order confirmatory factor analysis (CFA) using maximum likelihood estimation with robust SEs (Huber-White). The higher-order factors in the model consisted of latent factors related to (1) current stress experiences and (2) resilience resources. We evaluated model fit using multiple indices: χ^2/df ratio (<5.0 acceptable and <3.0 good), comparative fit index (CFI ≥0.90 acceptable and ≥0.95 good), Tucker-Lewis index (TLI ≥0.90 acceptable and ≥0.95 good), root mean square error of approximation (RMSEA ≤0.08 acceptable and ≤0.06 good) and standardized root mean square residual (SRMR ≤0.08 acceptable and ≤0.06 good).

Reliability Assessment (Phase 1)

We assessed reliability using composite reliability (CR) estimates, which are more appropriate than Cronbach alpha for factor-based models, particularly in structural equation modeling frameworks [45]. CR was calculated using standardized factor loadings: $CR=(\Sigma\lambda)^2/[(\Sigma\lambda)^2 + \Sigma(1-\lambda^2)]$ [46]. Reliability ≥0.70 was considered acceptable for subscales and ≥0.90 for total scales.

Validity Testing (Phase 1)

Convergent validity was assessed through multiple approaches: (1) within the CFA by examining factor loading magnitudes (≥0.60 preferred and ≥0.40 minimum), statistical significance, and average variance extracted (AVE ≥0.50); and (2) through Pearson correlations with theoretically related established measures administered concurrently.

Discriminant validity was evaluated through multiple methods: (1) examining the CFA factor structure for cross-loadings and modification indices (MIs) suggesting misspecification, (2) computing the heterotrait-monotrait (HTMT) ratio with conservative (0.85) and liberal (0.90) thresholds, and (3) evaluating correlation patterns demonstrating stronger relationships between similar constructs than dissimilar constructs.

Concurrent validity was assessed using Pearson correlations with established measures administered simultaneously, specifically examining relationships between WONE Index scales and corresponding validated measures of the same constructs (eg, WONE Resilience Index with CD-RISC and BRS).

Criterion-related validity was evaluated by examining correlations between WONE Index scales and important mental health and well-being outcomes, including depression, anxiety, and psychological well-being.

Power Analysis

We conducted comprehensive a priori power analyses to determine the appropriateness of our targeted sample size of 1000 for the planned statistical tests, including factor analysis, reliability testing, validity assessment, and measurement invariance testing. We used a 2-tailed significance level of α=.05 for all analyses. Detailed power analyses are provided in Multimedia Appendix 1.

Methods (Phase 2)

Participants and Procedure

We used the same nonprobability quota sampling recruitment strategy as Phase 1, except that we focused on US participants and recruited through CloudResearch. Participants were asked to complete surveys at 2 time points spaced 3 weeks apart. Participants were compensated US $5 at Time 1 and US $4 at Time 2; payment at Time 2 was slightly lower due to fewer items (eg, demographics). The analyses presented here focus on Time 1 data (N=306) for all EFAs and CFAs, with Time 2 used to assess test-retest reliability. Power analyses are provided within the “Data Analysis” section.

Theoretical Framework

Phase 2 was designed to refine and confirm the factor structure of the WONE Index based on the findings from Phase 1.

Phase 1 demonstrated that stress-related items formed a coherent, well-fitting factor structure, while several resilience domains—particularly health behaviors—required additional development. The strong validity evidence for the core stress and resilience skills components provided a solid foundation for expansion, while the challenges with nutrition, physical activity, and other behavioral domains indicated the need for larger item pools and more comprehensive domain coverage.

Phase 2 addresses these findings through several key design features. First, we substantially expanded the item pools for domains that showed promise but required development, particularly health behavior and social connection domains. Second, we adopted a domain-specific analytical approach, rather than the hierarchical approach used in Phase 1, allowing for comprehensive factor identification within each theoretical area while reducing model complexity. Third, we maintained the successful stress measurement structure from Phase 1 while expanding the resilience assessment to capture the full breadth of protective factors identified in resilience literature.

Domain Structure, Changes, and Measurement Strategy

Based on the findings and item performance from Phase 1, we adopted a staged, domain-specific approach that reflects both empirical patterns and theoretical considerations. The item pool was divided into 2 conceptually distinct domains for psychometric modeling and was analyzed within their respective models.

XSL•FO
RenderX

Resilience Resources Domain

Included items assessing behavioral and psychological capacities expected to support stress resilience. These spanned Emotion Regulation and Coping, Social Connectedness, Prosocial Values, Sleep, Physical Activity, and Dietary Intake. This comprehensive approach allowed for detailed examination of each resilience pathway while maintaining theoretical coherence across domains.

Stress Domain

We included the previously tested items under the Stress domain, reflecting stress experiences in one's work and personal life, with strategic modifications based on Phase 1 findings and contemporary contextual factors. In addition, recognizing the chronic stress that has characterized daily life since the COVID-19 pandemic [47], including polarizing societal and political experiences that have become the norm within the past decade [48], global conflicts, and economic uncertainty, we added 1 item specifically assessing stress, anxiety, and overwhelm due to current societal, political, and economic issues. This addition acknowledges that contemporary stress experiences extend beyond traditional work and personal life domains to include broader societal stressors [49] that have become increasingly prominent in recent years.

Measurement Strategy

Items were scored such that higher scores indicate greater stress or resilience resources when assessed separately, and the stress-load items were reverse-coded when assessed together so that higher overall scores reflected greater resilience. We analyzed both stress and resilience within their own respective models to reduce model complexity while still maintaining theoretical relevance. This staged approach is supported by previous psychometric validation practices, especially in multidimensional well-being assessments [50], and it allows for comprehensive development of each theoretical domain while maintaining the overarching framework that proved successful in Phase 1.

The domain-specific approach also identifies optimal factor structures for each area without the constraints imposed by simultaneous analysis of all domains. This approach is particularly important for health behavior domains that showed measurement challenges in Phase 1, allowing for detailed examination of their factor structure and item performance.

The WONE Index was developed and is copyrighted by Walking on Earth, Inc. The full administration materials, including detailed instructions and response options, are proprietary and therefore not publicly available. However, the measure may be available for use upon reasonable request to WONE via the website.

External Validation Measures

The same validated external scales from Phase 1 were administered in Phase 2 to assess validity, with 2 additional measures included.

Depressive Symptoms

In addition to the PROMIS-SF-8a [37], which was used in Phase 1, we also administered the PHQ-8 [51], a modified version of the PHQ-9 that omits the ninth item on thoughts of death or self-harm. The PHQ-8 demonstrates high internal consistency (α=.82-.89 [51,52]) and strong short-term test-retest reliability (r=0.89 [52]). It also demonstrates diagnostic performance comparable to the PHQ-9, with nearly identical sensitivity, specificity, and overall accuracy [52,53]. We included this measure to enable comparison with the digital health literature that frequently uses PHQ-based instruments and to triangulate depression assessment using complementary approaches. While the PROMIS-SF-8a provides an abstract, mood-focused assessment with reduced somatic bias, the PHQ-8 offers a behaviorally anchored assessment directly mapping onto *DSM* (*Diagnostic and Statistical Manual of Mental Disorders*) criteria [51,54]. Together, the 2 instruments strengthen construct validation by capturing depressive symptoms from 2 distinct measurement approaches.

Personality Assessment

To assess discriminant validity, we included the brief Big Five Inventory (BFI-10 [55]). This short-form instrument measures extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience, using 2 items per domain. The BFI-10 demonstrates acceptable internal consistency across traits (α=.65-.82), strong test-retest reliability (r=0.72-0.84), and high convergent validity with the full 44-item BFI (r=0.91-0.96 [55]).

Data Quality, Screening, and Analytic Assumptions

Participants' data were excluded from analyses if their responses to qualitative screening questions were nonsensical or irrelevant to the questions asked, as these were indicators of either automated responses or insufficient English fluency for valid participation. All survey items were required responses, resulting in no missing data for the final analytic samples (N=1005 for Phase 1 and N=306 for Phase 2).

Descriptive statistics and normality assessments were conducted for all key variables in both studies. Skewness and kurtosis values were all within the acceptable range of ±1.0 across both samples, indicating reasonably symmetric distributions without problematic tail behavior. Shapiro-Wilk tests indicated significant departures from normality for all variables in Phase 1 (all P<.001), which is expected given the large sample size. In Phase 2, normality tests were significant for external validation measures but not significant for the full WONE Index or Stress Subscale, with the Resilience Subscale showing marginal significance (P=.05). These patterns reflect the expected sensitivity of normality tests to sample size, with larger samples detecting minor deviations that are not practically meaningful [44]. Visual inspection of histograms and Q-Q plots confirmed that distributions were approximately normal with no severe violations.

Outlier analysis using boxplot methods identified minimal extreme cases: Phase 1 showed 1 - 11 outliers per variable (representing 1% - 4% of cases), while Phase 2 showed substantially fewer outliers (0 - 1 outliers per variable,

XSL•FO
RenderX

representing <1% of cases). These outliers were retained as they represented valid responses within the expected range of the constructs and did not appear to result from data entry errors. Given the acceptable skewness and kurtosis values, the robust nature of maximum likelihood estimation, and adequate sample sizes, the data were deemed appropriate for the planned factor analyses and correlational analyses.

Analytic Procedures

EFA

We used a staged EFA approach to identify the underlying factor structure within each domain. Principal axis factoring with Promax (oblique) rotation was used to accommodate the anticipated correlations between factors within each domain. For each EFA, sampling adequacy was assessed via the KMO statistic and Bartlett test of sphericity. Factor retention decisions were guided by eigenvalues >1.0, scree plot inspection, and theoretical coherence.

Items were retained if they met the following criteria: (1) primary loading ≥0.30, (2) minimal cross-loadings, and (3) thematic and theoretical fit with the emerging factor.

CFA

Following EFA, we specified and tested confirmatory factor models for each domain using maximum likelihood estimation with robust SEs. This domain-specific modeling strategy was selected to preserve theoretical clarity, allow for comprehensive factor development within each area, and build on the empirical structure observed in Phase 1, where resilience-related indicators and stress indicators showed distinct loading patterns.

Current Stress Model

Items loaded onto 3 first-order latent factors identified in Phase 1: Work Stress, Personal Stress, and Burnout. This model replicates the successful stress structure from Phase 1 while incorporating any additional stress-related items developed for Phase 2.

Resilience Resources Model

Items loaded onto 7 first-order latent factors representing expanded domains: Emotion Regulation and Coping, Social Connectedness, Purpose and Prosociality, Sleep, Physical Activity, Dietary Intake, and Perseverative Cognition. This expanded structure addresses the measurement challenges identified in Phase 1 by providing comprehensive coverage of resilience domains with sufficient items for stable factor identification.

The separate modeling approach allows for optimal factor structure identification within each domain while maintaining the theoretical framework established in Phase 1. Following successful domain-specific CFAs, the models can be integrated to test the overarching 2-factor higher-order structure (Stress and Resilience Resources) that demonstrated excellent validity in Phase 1. Model fit was evaluated using standard indices: CFI ≥0.90, RMSEA ≤0.08, and SRMR ≤0.08, consistent with conventional guidelines [56,57]. Model modifications were considered only when theoretically justifiable and indicated by MIs ≥10. When MIs suggested potential improvements, we prioritized residual covariances and factor-to-error covariances over cross-loadings to maintain interpretable factor structures while acknowledging theoretically expected relationships among constructs.

Reliability Assessment (Phase 2)

Reliability assessment was performed using the same method as described above for phase 1.

Validity Testing (Phase 2)

Convergent validity, concurrent validity, and criterion-related validity was assessed through multiple approaches, as described above for phase 1.

Test-Retest Reliability

A subset of participants completed the WONE Index at a second time point approximately 3 weeks after their initial assessment to evaluate temporal stability. Test-retest reliability was assessed using ICC with a 2-way mixed-effects model for consistency of agreement. Single-measure ICCs (ICC[2,1]) were calculated for the full WONE Index, Stress Subscale, and Resilience Subscale. ICC values were interpreted using established guidelines: <0.50=poor, 0.50 - 0.75=moderate, 0.75 - 0.90=good, and >0.90=excellent reliability [58].

Incremental Validity

Incremental validity was assessed through hierarchical multiple regression analyses to determine whether the WONE Index provides a meaningful prediction of mental health outcomes beyond established measures. Models tested incremental validity beyond: (1) the Perceived Stress Scale, (2) the CD-RISC, (3) the BRS, and (4) combined established measures. Mental health outcomes included depressive symptoms, anxiety symptoms, and well-being. Incremental validity was demonstrated by statistically significant R^2 change (ΔR^2) when the WONE Index was added to models.

Measurement Invariance Testing

We assessed measurement invariance of the Current Stress model using multigroup CFA, following the standard stepwise approach: configural, metric, scalar, and strict invariance. Each step imposed increasingly restrictive equality constraints across groups. Model fit was evaluated using commonly accepted thresholds for change in fit indices (ΔCFI ≤0.01, ΔRMSEA ≤0.015, and ΔSRMR ≤0.03 for metric invariance and ≤0.01 for scalar and strict invariance [59]). Fit was considered adequate when CFI and TLI were ≥0.95, RMSEA ≤0.06, and SRMR ≤0.08.

Power Analysis

We conducted comprehensive a priori power analyses to determine the appropriateness of our targeted sample size of 300 for the planned statistical tests. We used a 2-tailed significance level of α=.05 for all analyses. Detailed power analyses are provided in Multimedia Appendix 1.

XSL•FO
RenderX

Weighting Methodology

Hybrid Weighting Strategy Development

The WONE Index used a novel hybrid weighting approach that systematically integrates empirical prediction with theoretical importance and practical actionability. This methodology addresses a critical limitation in traditional psychometric approaches, in which purely data-driven weighting may undervalue theoretically important but empirically complex constructs, particularly in intervention-focused applications.

Empirical Weight Derivation

Empirical weights were derived through multiple regression analyses using each validation measure as a dependent variable and the 10 WONE factor scores as predictors. Standardized beta coefficients were extracted from regression models predicting 6 priority outcomes: CD-RISC-10, BRS, PSS-4, PHQ-8, GAD-7, and WHO-5.

Final empirical weights were calculated using a weighted average of standardized beta coefficients, with priority weights assigned based on theoretical importance: CD-RISC (25%), BRS (22%), PSS-4 (18%), PHQ-8 (15%), GAD-7 (12%), and WHO-5 (8%). For measures where higher scores indicate poorer outcomes (PSS-4, PHQ-8, and GAD-7), beta coefficients were reverse-scored to ensure all factors contributed positively to overall resilience scoring.

Theoretical Weight Framework

Theoretical weights were developed through a literature review [11,60-62] and author consensus, guided by the WONE Index's core conceptual model emphasizing resilience capacity through perceived stress (resource depletion) and modifiable protective/risk factors (resource building/depletion). Theoretical weights were assigned based on three primary criteria: (1) strength of empirical evidence linking the construct to stress resilience outcomes, (2) theoretical importance within established resilience frameworks, and (3) behavioral actionability for intervention purposes.

Hybrid Integration and Final Weights

The final hybrid weights were calculated using a 50/50 integration formula: Final Weight = (Theoretical Weight×0.5) + (Empirical Weight×0.5). This approach preserves empirical predictive validity while ensuring meaningful representation of theoretically important and behaviorally actionable constructs. A minimum weight floor of 5% was applied to ensure all constructs maintain sufficient influence for generating actionable user recommendations in digital health application contexts. Constructs falling below this threshold after hybrid calculation were elevated to 5%, with proportional reductions applied to higher-weighted constructs to maintain a total weight sum of 100%.

Ethical Considerations

This investigation was conducted in accordance with the Declaration of Helsinki and was determined to be exempt from Institutional Review Board (IRB) review after meeting ethical standards as evaluated by the Western Clinical Group IRB under 45 CFR § 46.104(d)(2) on November 27, 2024 (IRB ID 20244893). All participants provided informed consent before participation. Surveys were completed using SurveyMonkey (SurveyMonkey Inc), a GDPR-compliant platform with data encryption in transit and at rest, password-protected access controls, and secure data storage. Phase 1 participation was fully anonymous; no personally identifiable information was collected, and responses were recorded without linkage to any identifiers. In Phase 2, participant tracking across survey waves was managed through CloudResearch's "Waves" feature, which assigns a randomly generated 32-character hexadecimal string as an identifier. This procedure enabled longitudinal matching of responses while preventing access to identifying information. Researchers did not have access to participants' identities, and because surveys were administered in SurveyMonkey independently of CloudResearch, neither platform possessed both identity and response data. All data were stored on encrypted servers and analyzed in deidentified form.

Results

Phase 1: Results

Sample Demographics

Phase 1 included 1005 adults from the United States and the United Kingdom. Both samples were balanced by gender and predominantly White (refer to Table S1 in Multimedia Appendix 1).

Reliability of External Validation Measures

Stress Assessment

The PSS-4 demonstrated acceptable to good internal consistency, with reliability improving from Phase 1 (Cronbach α=0.79, 95% CI 0.77 - 0.81; McDonald ω=0.84) to Phase 2 (Cronbach α=0.85, 95% CI 0.82 - 0.88; McDonald ω=0.88). These reliability coefficients exceeded conventional thresholds for research use, supporting the measure's suitability for criterion validity analyses.

Resilience Assessment

Both resilience measures exhibited excellent internal consistency across studies. The CD-RISC-10 maintained consistent reliability (Phase 1: Cronbach α=0.91, 95% CI 0.90 - 0.92; McDonald ω=0.92 and Phase 2: Cronbach α=0.91, 95% CI 0.90 - 0.93; McDonald ω=0.93), while the BRS improved from Phase 1 (Cronbach α=0.88, 95% CI 0.87 - 0.90; McDonald ω=0.92) to Phase 2 (Cronbach α=0.92, 95% CI 0.91 - 0.94; McDonald ω=0.94). The convergence between reliability indices confirms robust psychometric performance and establishes a strong foundation for resilience-related criterion validity.

Mental Health and Well-Being Outcomes

All mental health measures demonstrated exceptional reliability with coefficients consistently exceeding 0.90. The PROMIS Depression Short Form-8a achieved the highest consistency (Phase 1: Cronbach α=0.95, 95% CI 0.95 - 0.96; McDonald ω=0.97 and Phase 2: Cronbach α=0.96, 95% CI 0.95 - 0.97; McDonald ω=0.97), followed by the Generalized Anxiety Disorder-7 (Phase 1: Cronbach α=0.92, 95% CI 0.91 - 0.93; McDonald ω=0.94 and Phase 2: Cronbach α=0.93, 95% CI

XSL•FO
RenderX

0.92 - 0.94; McDonald ω=0.96) and WHO-5 (Phase 1: Cronbach α=0.91, 95% CI 0.90 - 0.92; McDonald ω=0.93 and Phase 2: Cronbach α=0.92, 95% CI 0.91 - 0.94; McDonald ω=0.92). These high reliability values ensure that criterion validity relationships reflect true associations rather than measurement error.

EFA

Data demonstrated excellent suitability for factor analysis (KMO=0.926; Bartlett test: χ^2_{351}=13002.56; *P*<.001). Initially, 7 factors emerged—Resilience Skills and Beliefs, Work Stress, Personal Stress, Sleep, Burnout, Social Support, and Control—which accounted for 60.4% of the combined variance. Items related to alcohol, nutrition, sedentary time, and physical activity did not load onto any factor at ≥0.3, so they were removed. These findings indicated that health behavior domains required expanded item pools with more comprehensive measurement to achieve stable factor identification, which informed item development for Phase 2. The energy item loaded erroneously onto the Control factor and demonstrated a low Sleep loading, so it was also removed.

The final EFA with 27 items yielded a clear, interpretable 6-factor solution explaining 64.2% of the total variance: Resilience Skills and Beliefs, Work Stress, Personal Stress, Sleep, Burnout, and Social Support. Notably, the mental exhaustion item loaded onto the Work Stress factor instead of the Burnout factor. Although unexpected, mental exhaustion is typically the first symptom of burnout [63] and aligns conceptually with workplace stress; thus, the item was retained within the Work Stress factor. These items were then tested using the same factor placements in CFAs.

CFA

A 2-factor higher-order CFA was conducted to test the proposed measurement model. The model specified Stress as a second-order latent construct comprising 3 first-order factors (Work Stress, Personal Stress, and Burnout) and Resilience Resources as a second-order latent construct comprising 3 first-order factors (Resilience Skills and Beliefs, Social Support, and Sleep Quality). The model included 33 observed indicators across 6 first-order factors. The initial higher-order model demonstrated adequate fit (χ^2_{317}=11,593.27; *P*<.001; CFI=0.90; TLI=0.89; RMSEA=0.07, 95% CI 0.06 - 0.07; SRMR=0.058).

Following examination of MIs, residual covariances were freed for 15 conceptually related item pairs with MIs exceeding 20. The refined model achieved excellent fit (χ^2_{302}=993.30; *P*<.001; CFI=0.95; TLI=0.94; RMSEA=0.049, 95% CI 0.046 - 0.052; SRMR=0.052). Although the chi-square test remained statistically significant—consistent with large sample size sensitivity [64]—all practical fit indices met or exceeded conventional standards for excellent model fit [56].

The Resilience Skills and Beliefs factor demonstrated strong overall performance, with loadings ranging from acceptable to strong (0.40≤ λ ≤0.76). Social Support emerged as the most cohesive first-order factor, with both indicators demonstrating exceptionally strong loadings (λ >0.80). Sleep showed a mixed but interpretable pattern, with 1 dominant indicator (sleep quality; λ=0.87) and 3 secondary indicators with more modest loadings (0.52≤ λ ≤0.63). The 3 stress factors (Work Stress, Personal Stress, and Burnout) all demonstrated strong psychometric properties (0.40≤ λ ≤0.76; refer to Table S3 in Multimedia Appendix 1).

Second-Order Factor Structure

The higher-order structure received strong empirical support, with all second-order loadings exceeding 0.60 and demonstrating appropriate magnitudes. The Stress factor showed balanced contributions from its 3 constituent factors, with loadings ranging from 0.71 to 0.741, indicating that Work Stress, Personal Stress, and Burnout represent relatively equal contributors to the overarching stress construct.

Resilience Resources exhibited a more hierarchical structure, with Resilience Skills and Beliefs serving as the dominant indicator (λ=0.94) and Social Support and Sleep Quality contributing more modestly but substantially (λ=0.67 and 0.62, respectively). This pattern suggests that while all 3 resources are important components of resilience, individual resilience skills may serve as the core organizing feature of this higher-order construct.

Reliability and Validity Testing

Internal Consistency Reliability

All factors demonstrated acceptable to excellent internal consistency reliability, as provided in Table S4 in Multimedia Appendix 1. CR estimates exceeded established thresholds across all factors. The full WONE Index showed excellent reliability, while subscale reliability was consistently strong across all first-order factors. All reliability estimates met or exceeded the 0.70 threshold for acceptable reliability, with most surpassing the 0.80 standard for good reliability.

Convergent Validity

The WONE Index demonstrated strong convergent validity with theoretically related established measures via correlations with externally validated measures. Resilience constructs showed substantial positive correlations with validated resilience measures: the Resilience Index correlated strongly with the CD-RISC (*r*=0.74; *P*<.001) and moderately with the BRS (*r*=0.68; *P*<.001). These correlations fall within the large effect size range [65], indicating that the WONE Resilience Index captures similar underlying resilience constructs while maintaining distinctiveness.

Stress constructs exhibited strong convergent validity with established stress and mental health measures. The Stress subscale correlated substantially with the PSS (*r*=0.66; *P*<.001), demonstrating convergent validity for the overall stress construct. Individual stress components showed expected relationships: positive correlations with depressive symptoms, anxiety symptoms, and perceived stress, as well as a negative correlation with well-being, were observed, supporting the validity of the stress measurement model.

AVE estimates, calculated as the mean of the communalities (R^2) for each factor's items (refer to Table S3 in Multimedia Appendix 1), provided evidence of convergent validity for most

XSL•FO
RenderX

factors. Four factors met or exceeded the 0.50 criterion, while 2 factors showed AVE values below 0.50 but above 0.40. While AVE values below 0.50 suggest some concern for convergent validity, the strong CR estimates and substantial factor loadings discussed previously indicate that convergent validity is still adequate for these factors [46].

Discriminant Validity

HTMT analysis provided strong evidence for discriminant validity. All HTMT values fell well below the conservative 0.85 threshold, with an average value of 0.49, indicating excellent discriminant validity between all factor pairs (refer to Table S4 in Multimedia Appendix 1).

Concurrent Validity

The WONE Index demonstrated excellent concurrent validity with established criterion measures (refer to Table S5 in Multimedia Appendix 1). The Resilience subscale showed strong concurrent validity with both the CD-RISC and BRS, 2 well-validated resilience measures with different theoretical emphases. The Stress subscale demonstrated substantial concurrent validity with the PSS, a widely used measure of subjective stress appraisal.

Criterion-Related Validity

The WONE Index showed strong criterion-related validity with important mental health and well-being outcomes (refer to Table S5 in Multimedia Appendix 1). The full WONE Index demonstrated robust associations with all criterion measures, showing positive correlations with well-being and negative correlations with both depressive and anxiety symptoms.

The Stress subscale exhibited strong criterion-related validity, with negative associations with well-being and positive associations with both depressive and anxiety symptoms. The Resilience subscale demonstrated complementary criterion-related validity, with positive associations with well-being and negative associations with both depressive and anxiety symptoms.

Phase 2 Results

Sample Demographics

Phase 2 included 306 US adults. The sample was balanced by gender and predominantly White (refer to Table S6 in Multimedia Appendix 1). Although the study was open to participants aged 18 - 64 years, the youngest participant was aged 20 years, and the oldest was aged 63 years, resulting in an age range of 20 - 63 years.

Reliability of External Validation Measures

Depressive Symptoms

The PHQ-8 demonstrated excellent internal consistency (Cronbach α=0.90, 95% CI 0.88-0.91; McDonald ω=0.93). This strong reliability supports its use as a complementary measure to the PROMIS-SF-8a and strengthens construct validation through convergent evidence across different measurement approaches.

Personality

Reliability for the BFI-10 subscales was modest, consistent with expectations for short-form personality measures. Extraversion (α=.71 and ω=0.71) and Neuroticism (α=.77 and ω=0.77) showed adequate reliability, while Conscientiousness (α=.58 and ω=0.63), Agreeableness (α=.57 and ω=0.57), and Openness (α=.58 and ω=0.63) were lower. These results are consistent with previous findings indicating that short-form personality measures often exhibit attenuated reliability due to their limited item coverage per construct.

EFA Results

Stage 1: Resilience Factors Analysis

Data Suitability and Sample Characteristics

Analysis was conducted on 306 participants using 51 resilience-related items. Data demonstrated excellent suitability for factor analysis: KMO measure of sampling adequacy=0.91 (excellent), and Bartlett test of sphericity was highly significant (χ^2_{528}=5790.89; *P*<.001), confirming the appropriateness of the correlation matrix for factorization.

Factor Extraction and Retention Strategy

Based on theoretical considerations, we specified a maximum of 7 factors corresponding to hypothesized resilience domains: (1) Emotion Regulation and Coping Tendencies, (2) Resilience-Related Beliefs, (3) Positive Psychology Buffers, (4) Social Connectedness, (5) Dietary Intake, (6) Physical Activity, and (7) Sleep.

Item Reduction and Refinement Process

Through iterative analysis, we applied stringent psychometric criteria for item retention. Items were systematically removed based on (1) low communalities (<0.30), (2) inadequate factor loadings (<0.40), and (3) theoretically inappropriate cross-loadings. This process resulted in the removal of 15 items.

Final Factor Structure and Variance Explained

The final 7-factor solution with 33 items explained 66.04% of the total variance, demonstrating substantial explanatory power. Factor loadings ranged from 0.42 to 1.00, with the majority of items (26/33, 78.79%) exceeding 0.50, indicating strong factor-item relationships. Alternative solutions with 5 or 6 factors were examined but yielded weaker, less interpretable factor structures.

Stage 2: Stress Factors EFA

Data Suitability and Factor Extraction

Analyses using 13 stress-related items with the same sample (N=306) demonstrated good data suitability (KMO=0.87; Bartlett test χ^2_{55}=2186.66; *P*<.001). Based on theoretical considerations and Phase 1 findings, we specified 3 stress factors corresponding to distinct stress domains, assuming that the new item related to societal, political, and economic stress would align with the Personal Stress factor. Principal axis factoring with Promax rotation converged in 6 iterations.

XSL•FO
RenderX

Factor Structure and Item Performance

The 3-factor solution explained 74.76% of total variance. Two items were removed due to substantial cross-loadings across stress domains: Perceived Control ("Felt that you have control over the important things in your life") and Smooth ("That things were going smoothly for you") both loaded significantly onto both Work Stress and Personal Stress factors, indicating these items captured general perceived control and life satisfaction rather than domain-specific stressors. The final 11 items demonstrated a clear factor structure with strong psychometric properties.

Notably, the exhaustion item loaded onto the Work Stress rather than the Burnout factor, which is consistent with theoretical models positioning exhaustion as an early indicator of work-related stress that may precede full burnout syndrome.

Integrated Factor Structure Summary

The staged EFA approach successfully identified a robust 10-factor structure encompassing 43 items: 7 resilience factors (32 items) and 3 stress factors (11 items). This comprehensive framework captures both current stress exposure and modifiable factors that influence resilience capacity, providing a theoretically grounded and empirically supported foundation for comprehensive stress resilience assessment.

The excellent psychometric properties, substantial variance explained (66.04% for resilience factors and 74.76% for stress factors), and strong reliability estimates support the validity of this multidimensional approach to measuring stress resilience capacity in applied settings.

CFA Results

Current Stress Model

We conducted a CFA to evaluate the structure of the Current Stress domain, which specified a higher-order latent factor (Stress) comprised of 3 latent factors: Work Stress, Personal Stress, and Burnout. The initial model demonstrated adequate fit to the data (χ^2_{41}=204.52; P<.001; CFI=0.92; TLI=0.90; RMSEA=0.11; and SRMR=0.07). We used MIs ≥10 to identify potential covariances between item residuals and added the following theoretically justified covariances to improve model fit without altering the core factor structure (refer to Multimedia Appendix 1).

These modifications reflected close semantic or contextual overlap among items and improved the overall model fit. Factor-to-error covariances were preferred over cross-loadings to maintain clear factor interpretation while acknowledging systematic relationships between theoretically related stress constructs.

The revised model demonstrated excellent fit to the data (χ^2_{33}=64.18; P=.001; CFI=0.99; TLI=0.98; RMSEA=0.06; and SRMR=0.05). All standardized factor loadings were statistically significant and strong, ranging from 0.54 to 0.92 across the 11 items. Most loadings exceeded 0.75, supporting the construct validity of the latent factors. Squared multiple correlations (R^2) showed that most items explained a substantial proportion of variance, ranging from 0.29 to 0.84. Full standardized factor loadings, SEs, and communalities are provided in Table 1. The WONE Index is a proprietary measure copyrighted by Walking on Earth, Inc. Full administration materials are available upon reasonable request.

XSL•FO
RenderX

Table . Standardized factor loadings and communalities for Phase 2 confirmatory factor analysis (CFA) models.

Construct	Factor	Domain	Standardized loading[a]	SE	Communality (R^2)
Stress (personal)	Personal Stress	Stress	0.92	—[b]	0.84
Anxiousness (personal)	Personal Stress	Stress	0.87	0.05	0.75
Overwhelm (personal)	Personal Stress	Stress	0.86	0.05	0.74
Societal, political, and economic stress	Personal Stress	Stress	0.54	0.06	0.29
Stress (work)	Work Stress	Stress	0.90	—	0.82
Anxiousness (work)	Work Stress	Stress	0.76	0.06	0.58
Overwhelm (work)	Work Stress	Stress	0.83	0.06	0.69
Mental exhaustion	Work Stress	Stress	0.79	0.07	0.63
Disengagement	Burnout	Stress	0.86	—	0.74
Cynicism	Burnout	Stress	0.75	0.08	0.56
Lack of productivity	Burnout	Stress	0.75	0.08	0.56
Vigorous physical activity	Physical Activity	Resilience Resources	0.99	0.11	1.00
Moderate physical activity	Physical Activity	Resilience Resources	0.63	—	0.40
Sleep quality	Sleep	Resilience Resources	0.85	—	0.72
Fatigue	Sleep	Resilience Resources	0.74	0.07	0.55
Sleep duration	Sleep	Resilience Resources	0.47	0.10	0.22
Sleep latency	Sleep	Resilience Resources	0.62	0.08	0.38
Sleep disturbances	Sleep	Resilience Resources	0.64	0.08	0.41
Nutritious food intake	Dietary Intake	Resilience Resources	0.65	—	0.42
Processed food intake	Dietary Intake	Resilience Resources	0.50	0.13	0.25
Caffeine intake	Dietary Intake	Resilience Resources	0.32	0.13	0.10
Caffeine reliance	Dietary Intake	Resilience Resources	0.52	0.19	0.27
Emotion regulation	Emotion Regulation and Coping	Resilience Resources	0.74	0.06	0.68
Emotional understanding	Emotion Regulation and Coping	Resilience Resources	0.61	0.06	0.37
Distress tolerance	Emotion Regulation and Coping	Resilience Resources	0.70	0.06	0.49
Acceptance	Emotion Regulation and Coping	Resilience Resources	0.50	0.06	0.25
Ability to bounce back	Emotion Regulation and Coping	Resilience Resources	0.76	0.06	0.58
Adaptability	Emotion Regulation and Coping	Resilience Resources	0.73	0.05	0.53
Effective coping	Emotion Regulation and Coping	Resilience Resources	0.83	—	0.69
Self-efficacy	Emotion Regulation and Coping	Resilience Resources	0.68	0.06	0.47
Cognitive flexibility	Emotion Regulation and Coping	Resilience Resources	0.66	0.06	0.43
Dwelling	Perseverative Thinking	Resilience Resources	0.91	—	0.82
Worrying	Perseverative Thinking	Resilience Resources	0.88	0.06	0.78

XSL•FO
RenderX

Construct	Factor	Domain	Standardized loading[a]	SE	Communality (R^2)
Meaning and purpose	Purpose and Prosociality	Resilience Resources	0.54	0.10	0.44
Gratitude	Purpose and Prosociality	Resilience Resources	0.78	0.09	0.61
Compassion for others	Purpose and Prosociality	Resilience Resources	0.63	0.09	0.40
Consideration for others	Purpose and Prosociality	Resilience Resources	0.67	0.07	0.45
Making a positive difference	Purpose and Prosociality	Resilience Resources	0.79	—	0.63
Trusted support system	Social Connection	Resilience Resources	0.89	0.05	0.79
Support satisfaction	Social Connection	Resilience Resources	0.87	0.06	0.76
Strength from close others	Social Connection	Resilience Resources	0.87	—	0.76
Belongingness	Social Connection	Resilience Resources	0.85	0.05	0.72
Loneliness	Social Connection	Resilience Resources	0.83	0.06	0.69

[a]All factor loadings are statistically significant at $P<.001$.

[b]SE not reported when factor loading was fixed to 1 for model identification.

Second-Order Factor Structure

The higher-order Stress factor was well supported by strong loadings from all 3 first-order factors. Work Stress showed the strongest loading (λ=0.80 and SE=0.12), followed by Personal Stress (λ=0.76) and Burnout (λ=0.67 and SE=0.13). The second-order loadings ranged from 0.67 to 0.80, with an average of 0.74, indicating that all 3 stress domains contribute substantially and relatively equally to the overarching stress construct. The squared multiple correlations for the first-order factors (0.44-0.65) demonstrate that the higher-order factor explains a substantial proportion of variance in each stress domain.

Resilience Resources Model

We then conducted a CFA to assess the structure of the Resilience Resources model, which specified a higher-order latent factor (Resources) comprised of 7 first-order latent domains: Emotion Regulation and Coping, Social Connectedness, Compassion and Gratitude, Sleep, Physical Activity, Dietary Intake, and Perseverative Thinking. The initial model demonstrated adequate fit to the data ($\chi^2_{458}=1081.71$; $P<.001$; CFI=0.88; TLI=0.87; RMSEA=0.07; and SRMR=0.09). After reviewing MIs≥10, we added theoretically grounded modifications (refer to Multimedia Appendix 1 for specifics).

These modifications reflect meaningful psychological relationships among resilience constructs and improved model fit substantially while preserving the primary factor structure. This approach was chosen over alternative specifications (eg, cross-loadings) to maintain interpretability while acknowledging the theoretically expected interconnections among resilience domains.

The revised model fit the data well ($\chi^2_{443}=745.20$; $P<.001$; CFI=0.94; TLI=0.94; RMSEA=0.05; and SRMR=0.06). All standardized factor loadings were statistically significant, ranging from 0.31 to 0.99. Most loadings exceeded 0.60, indicating strong relationships between latent constructs and their corresponding indicators. Squared multiple correlations (R^2) demonstrated good explanatory power for most observed variables, supporting the reliability and coherence of the factor structure. The confirmed 10-factor structure represents the final, validated framework of the WONE Index. Each factor captures specific theoretical domains as described in the “Second-Order Factor Structure" section. Item-level psychometric details are provided in Table 1, and theoretical descriptions of the factors are included in the “Discussion” section for Phase 2.

Second-Order Factor Structure

The higher-order Resilience Resources factor showed a more varied pattern of loadings from the 7 first-order factors, ranging from 0.31 to 0.88 (average=0.59). Emotion Regulation and Coping emerged as the dominant contributor (λ=0.88), followed by Perseverative Thinking (λ=0.73) and Sleep (λ=0.78). Social Connection showed a moderate loading (λ=0.55), while Purpose and Prosociality (λ=0.37) and Physical Activity (λ=0.31) contributed more modestly to the overarching construct. This hierarchical pattern suggests that while multiple domains contribute to resilience resources, cognitive-emotional regulatory capacities serve as the primary organizing feature, with behavioral and social factors providing important but secondary contributions.

Although each domain was analyzed separately for model estimation and reporting, together these 2 validated higher-order structures conceptually represent a correlated higher-order framework consistent with JD-R theory. This framework conceptualizes Stress Load and Resilience Resources as distinct but interrelated systems that cannot be reduced to a single overarching factor. The correlated-factor interpretation was retained over a bifactor alternative because it better reflects the theoretical position that resilience emerges from dynamic

XSL•FO
RenderX

interactions among multiple interdependent subsystems rather than from a single general dimension.

The 2 higher-order factors were also strongly correlated at the scale level, which indicates that higher stress load was statistically associated with lower resilience resources, although this relationship was not modeled within the same CFA structure. The squared multiple correlations for the first-order factors ranged from 0.10 to 0.78, indicating substantial variation in how well the higher-order factor explains variance across different resilience domains.

Validity and Reliability

All correlations between the WONE Index and external measures are provided in Table 2.

XSL•FO
RenderX

Table . Phase 2 correlations between the Walking on Earth (WONE) Index and established measures.

Variable	Full WONE Index	WONE stress subscale	WONE resilience subscale	PSS[a]	CD-RISC[b]	BRS[c]	PHQ-8[d]	PROMIS-SF-8a[e]	GAD-7[f]	WHO-5[g]	BFI-E[h]	BFI-A[i]	BFI-C[j]	BFI-N[k]	BFI-O[l]
Full WONE Index															
r	1	–0.84	0.98	–0.81	0.75	0.74	–0.77	–0.79	–0.77	0.83	0.31	0.46	0.45	–0.74	–0.03
P value	—[m]	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.56
WONE stress subscale															
r	–0.84a	1	–0.71	0.73	–0.52	–0.57	0.69	0.71	0.75	–0.68	–0.21	–0.31	–0.39	0.66	0.03
P value	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.59
WONE resilience subscale															
r	0.98a	–0.71	1	–0.78	0.77	0.75	–0.73	–0.76	–0.71	0.82	0.33	0.47	0.44	–0.71	–0.03
P value	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.58
PSS															
r	–0.81	0.73	–0.78	1	–0.66	–0.68	0.72	0.80	0.76	–0.74	–0.26	–0.36	–0.41	0.61	0.02
P value	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.74
CD-RISC															
r	0.75	–0.52	0.77	–0.66	1	0.83	–0.52	–0.57	–0.53	0.66	0.33	0.33	0.42	–0.65	–0.06
P value	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.27
BRS															
r	0.74	–0.57	0.75	–0.68	0.83	1	–0.53	–0.57	–0.58	0.62	0.27	0.40	0.39	–0.68	–0.05
P value	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.44
PHQ-8															
r	–0.77	0.69	–0.73	0.72	–0.52	–0.53	1	0.85	0.81	–0.75	–0.22	–0.34	–0.41	0.55	0.05
P value	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.36
PROMIS-SF-8a															
r	–0.79	0.71	–0.76	0.80	–0.57	–0.57	0.85	1	0.78	–0.76	–0.29	–0.40	–0.41	0.55	0.04
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	.45
GAD-7															
r	–0.77	0.75	–0.71	0.76	–0.53	–0.58	0.81	0.78	1	–0.67	–0.19	–0.30	–0.33	0.67	0.06
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	.27
WHO-5															
r	0.83	–0.68	0.82	–0.74	0.66	0.62	–0.76	–0.75	–0.67	1	0.32	0.42	0.43	–0.61	–0.04
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	.53
BFI-E															
r	0.31	–0.21	0.33	–0.26	0.33	0.27	–0.29	–0.22	–0.19	0.32	1	0.30	0.25	–0.33	0.02

XSL•FO
RenderX

Variable		Full WONE Index	WONE stress subscale	WONE resilience subscale	PSS[a]	CD-RISC[b]	BRS[c]	PHQ-8[d]	PROMIS-SF-8a[e]	GAD-7[f]	WHO-5[g]	BFI-E[h]	BFI-A[i]	BFI-C[j]	BFI-N[k]	BFI-O[l]
	P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	.75
BFI-A																
	r	0.46a	–0.31	0.47	–0.34	0.33	0.40	–0.40	–0.35	–0.30	0.42	0.30	1	0.34	–0.33	–0.05
	P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	.34
BFI-C																
	r	0.45a	–0.39	0.44	–0.41	0.42	0.34	–0.41	–0.41	–0.33	0.43	0.25	0.34	1	–0.38	0.02
	P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	.76
BFI-N																
	r	–0.74	0.66	–0.71	0.61	–0.65	–0.68	0.55	0.55	0.67	–0.61	–0.33	–0.33	–0.38	1	0.02
	P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	.67
BFI-O																
	r	–0.03	0.03	–0.03	0.02	–0.06	–0.05	0.04	0.05	0.06	–0.04	0.02	–0.05	0.02	0.02	1
	P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—

[a]PSS: Perceived Stress Scale.

[b]CD-RISC: Connor-Davidson Resilience Scale.

[c]BRS: Brief Resilience Scale.

[d]PHQ-8: Patient Health Questionnaire-8.

[e]PROMIS-SF-8a: Patient-Reported Outcomes Measurement Information System Short Form 8a.

[f]GAD-7: Generalized Anxiety Disorder-7.

[g]WHO-5: World Health Organization-5 Well-Being Index.

[h]BFI-E: Big Five Inventory-Extraversion

[i]BFI-A: Big Five Inventory-Agreeableness

[j]BFI-C: Big Five Inventory-Conscientiousness

[k]BFI-N: Big Five Inventory-Neuroticism

[l]BFI-O: Big Five Inventory-Openness

[m]Not applicable.

Construct-Level Convergent Validity

The WONE Index demonstrated strong construct-level convergent validity with established measures of stress and resilience. The Stress Subscale showed a large positive correlation with the PSS, indicating that higher scores on our stress measure align closely with higher perceived stress as measured by this gold-standard instrument.

For resilience measures, the Resilience Subscale demonstrated excellent convergent validity with both the CD-RISC-10 and the BRS. Similarly, the Full WONE Index showed strong positive correlations with both resilience measures. These large correlations (all exceeding Cohen benchmark for large effects [65]) with established resilience measures are consistent with the WONE Index measuring conceptually related but distinct aspects of stress resilience, as would be expected given its emphasis on specific behavioral resources and demands-resources integration versus the trait-based conceptualizations of the CD-RISC and BRS. These patterns support construct validity while indicating that the WONE Index captures additional variance beyond established measures.

Concurrent Validity

The measure demonstrated excellent concurrent validity with theoretically related mental health and well-being outcomes, with all correlations in the expected directions and magnitudes.

Stress-Mental Health Relationships

The Stress Subscale showed strong positive correlations with both depressive symptom measures (PHQ-8 and PROMIS-SF-8a) and anxiety symptoms (GAD-7). Conversely, the Stress Subscale demonstrated a strong negative correlation with well-being (WHO-5), indicating that higher stress is associated with lower well-being, as expected.

XSL•FO
RenderX

Resilience-Mental Health Relationships

The Resilience Subscale showed strong protective relationships with mental health outcomes. Higher resilience scores were associated with lower depressive symptoms on both measures, lower anxiety symptoms, and higher well-being.

Full WONE Index Concurrent Validity

The Full WONE Index demonstrated excellent concurrent validity across all mental health outcomes, with large negative correlations with depressive and anxiety symptoms and a large positive correlation with well-being. These strong relationships across all mental health domains support the Full Index as a comprehensive measure of stress resilience capacity.

Comparative Performance

Notably, the WONE Index demonstrated stronger correlations with mental health outcomes than established resilience measures and comparable correlations with the PSS for stress-related outcomes (refer to Table 2).

Discriminant Validity

The WONE Index showed appropriate discriminant validity when examined against personality traits, with one theoretically meaningful exception that actually supports construct validity.

Neuroticism (Supporting Evidence for Construct Validity)

The Full WONE Index demonstrated a strong negative correlation with neuroticism. Rather than representing a discriminant validity concern, this relationship provides important construct validity evidence. Neuroticism, characterized by heightened stress reactivity and emotional instability, would be expected to show a strong negative relationship with stress resilience capacity. This correlation suggests our measure successfully captures individual differences in stress vulnerability and resilience resources, making it particularly valuable for identifying individuals who may benefit most from resilience-building interventions.

Other Personality Traits

The Index showed moderate correlations with agreeableness and conscientiousness, suggesting some overlap with these adaptive personality traits, though correlations remained below the threshold indicating construct redundancy. The measure showed a smaller correlation with extraversion and was essentially uncorrelated with openness, demonstrating good discrimination from these personality dimensions.

Discriminant Validity Between Domains

HTMT ratios were calculated to assess discriminant validity between measurement domains. All HTMT values fell below the conservative threshold of 0.85, with most below 0.70, supporting discriminant validity between domains while confirming their theoretical relationships.

Stress Domain Relationships

The 3 stress domains showed moderate to strong relationships, confirming that they capture related but distinct stress experiences (refer to Table S7 in Multimedia Appendix 1).

Resilience Domain Relationships

Resilience domains demonstrated appropriate discriminant validity. The strongest relationships were observed between conceptually related domains (Emotion Regulation and Coping ↔ Sleep; Purpose and Prosociality ↔ Social Connection), supporting the theoretical structure while confirming domain distinctiveness (refer to Table S8 in Multimedia Appendix 1).

Internal Consistency

The CR estimate for the overall scale was 0.84. Estimates for individual factors ranged from 0.62 to 0.96, with 9 of 10 factors exceeding the 0.70 threshold for acceptable reliability. Eight factors achieved good to excellent reliability (CR ≥0.80; refer to Table 3).

XSL•FO
RenderX

Table . Factor-level test-retest reliability and internal consistency.

Subscale and factor	ICC[a] (95% CI)	CR[b]	AVE[b]
WONE[c] Index	0.90 (0.87-0.93)	—	—
Stress subscale	0.84 (0.80-0.88)	—	—
Personal Stress factor	0.90 (0.87-0.92)	0.90	0.64
Work Stress factor	0.89 (0.85-0.91)	0.91	0.67
Burnout factor	0.87 (0.82-0.90)	0.84	0.58
Resilience subscale	0.90 (0.87-0.92)	—	—
Emotion Regulation and Coping factor	0.95 (0.93-0.96)	0.89	0.49
Social Connectedness factor	0.93 (0.91-0.95)	0.94	0.74
Purpose and Prosociality factor	0.93 (0.91-0.95)	0.83	0.49
Sleep factor	0.90 (0.87-0.93)	0.83	0.51
Dietary Intake factor	0.90 (0.86-0.92)	0.62	0.31
Physical Activity factor	0.81 (0.74-0.85)	0.81	0.70
Perseverative Thinking factor	0.79 (0.72-0.84)	0.89	0.80

[a]ICC: intraclass correlation coefficient.

[b]Composite reliability (CR) and average variance extracted (AVE) estimates are not applicable to the full WONE Index, as it is calculated as a composite of 2 separate confirmatory factor models, rather than being estimated as a latent factor within a single measurement model.

[c]WONE: Walking on Earth

The Dietary Intake factor showed marginal reliability (CR=0.62 and AVE=0.31), likely reflecting the conceptual diversity of dietary behaviors assessed (caffeine reliance, processed food consumption, and nutritious diet quality). The marginal reliability of the Dietary Intake factor likely reflects the conceptual diversity of dietary behaviors rather than measurement inadequacy. Despite the lower statistical indicators, these items were retained due to their theoretical importance in the resilience framework and the nascent state of dietary behavior measurement in resilience research.

Test-Retest Reliability

The WONE Index demonstrated excellent temporal stability across all major scales over 3 weeks (Table 3). The Full WONE Index showed excellent test-retest reliability, indicating highly consistent measurement of overall stress resilience capacity over time. Both subscales also demonstrated strong temporal stability, with the Resilience Subscale exhibiting excellent reliability and the Stress Subscale showing good reliability.

These findings indicate that the WONE Index provides a stable and consistent measurement of both stress and resilience resources across time. The excellent temporal stability of the WONE Index supports its use as a reliable assessment tool for tracking stress resilience capacity. The high test-retest reliability coefficients indicate that observed changes in WONE scores over time reflect true changes in an individual's stress resilience capacity rather than measurement error, making the instrument suitable for monitoring intervention effects and tracking progress.

Individual WONE factors also showed strong to excellent test-retest reliability across the 3-week interval (Table 3). Nine of 10 factors achieved excellent temporal stability (ICC >0.85).

Incremental Validity Results

The WONE Index provided significant incremental prediction over a combination of existing gold-standard measures. All hierarchical regression models demonstrated statistically significant incremental validity, indicating that the WONE Index provides meaningful prediction beyond established measures across all mental health outcomes tested. The WONE Index's superior correlational performance relative to established resilience measures and comparable performance to the PSS provides the foundation for these incremental validity findings.

Model 1: Beyond Perceived Stress Scale

The WONE Full Index provided significant incremental prediction beyond the PSS across all outcomes. Even after accounting for perceived stress, the WONE explained additional variance in depressive symptoms (PHQ-8: ΔR^2=0.10, P<.001; PROMIS-SF-8a: ΔR^2=0.06, P<.001; GAD-7: ΔR^2=0.07, P<.001; and WHO-5: ΔR^2=0.16, P<.001).

Model 2: Beyond Established Resilience Measures

Model 2a (Beyond CD-RISC)

The WONE Index demonstrated substantial incremental validity beyond the CD-RISC across all outcomes, with particularly strong incremental prediction for depressive symptoms and well-being (PHQ-8: ΔR^2=0.32, P<.001; PROMIS-SF-8a: ΔR^2=0.31, P<.001; GAD-7: ΔR^2=0.31, P<.001; and WHO-5: ΔR^2=0.26, P<.001).

Model 2b (Beyond BRS)

Similar patterns were observed when testing incremental validity beyond the BRS, with significant ΔR^2 values across all mental health outcomes (PHQ-8: ΔR^2=0.31, P<.001; PROMIS-SF-8a:

XSL•FO
RenderX

ΔR^2=0.31, *P*<.001; GAD-7: ΔR^2=0.25, *P*<.001; and WHO-5: ΔR^2=0.31, *P*<.001).

Model 3: Beyond Combined Established Measures

Most importantly, the WONE Index provided significant incremental prediction even beyond the combination of both established stress and resilience measures (PSS, CD-RISC, and BRS). This represents the most stringent test of incremental validity, as it demonstrates that WONE adds meaningful prediction beyond current best practices that would use both stress and resilience assessments (PHQ-8: ΔR^2=0.11, *P*<.001; PROMIS-SF-8a: ΔR^2=0.07, *P*<.001; GAD-7: ΔR^2=0.07, *P*<.001; and WHO-5: ΔR^2=0.11, *P*<.001).

Measurement Invariance for Stress

Measurement Invariance Testing Across Gender

Multigroup CFA indicated good configural fit across women (n=164) and men (n=142), suggesting a consistent factor structure (χ^2_{66}=115.71; *P*<.001; CFI=0.98; TLI=0.96; RMSEA=0.05; and SRMR=0.06). Constraining factor loadings (metric model) maintained a strong fit (χ^2_{76}=125.58; *P*<.001; CFI=0.98; TLI=0.97; RMSEA=0.05; and SRMR=0.06), supporting metric invariance. Adding intercept constraints (scalar model) yielded an acceptable fit (χ^2_{86}=151.50; *P*<.001; CFI=0.97; TLI=0.96; RMSEA=0.05; and SRMR=0.07), supporting latent mean comparisons. The strict model (residual variances constrained) also fit well (χ^2_{99}=163.14; *P*<.001; CFI=0.97; TLI=0.97; RMSEA=0.05; and SRMR=0.07), supporting full invariance across gender.

Measurement Invariance Across Race

Participants were grouped as White (n=211) and non-White (n=95). Configural fit was strong (χ^2_{66}=100.62; *P*=.01; CFI=0.99; TLI=0.98; RMSEA=0.04; and SRMR=0.05), indicating consistent structure. The metric model (loadings constrained) also showed excellent fit (χ^2_{76}=100.62; *P*=.03; CFI=0.99; TLI=0.98; RMSEA=0.03; and SRMR=0.05). Scalar constraints (intercepts) retained good fit (χ^2_{86}=111.65; *P*=.03; CFI=0.99; TLI=0.99; RMSEA=0.03; and SRMR=0.04), allowing for latent mean comparisons. The strict model (residuals constrained) showed acceptable fit (χ^2_{99}=129.24; *P*=.02; CFI=0.97; TLI=0.98; RMSEA=0.03; and SRMR=0.045), indicating full invariance across race.

Measurement Invariance Across Age Groups

Participants were split into ages 20 - 39 years (n=175) and 40 - 64 years (n=131). The configural model demonstrated good fit (χ^2_{66}=95.23; *P*=.01; CFI=0.99; TLI=0.98; RMSEA=0.04; and SRMR=0.06). Metric invariance was supported with excellent fit when loadings were constrained (χ^2_{76}=102.98; *P*=.02; CFI=0.99; TLI=0.98; RMSEA=0.03; and SRMR=0.06). The scalar model (adding intercept constraints) showed a strong fit (χ^2_{86}=127.18; *P*=.003; CFI=0.98; TLI=0.98; RMSEA=0.04; and SRMR=0.06), permitting latent mean comparisons. The strict model (residuals constrained) also fit well (χ^2_{99}=143.23; *P*<.002; CFI=0.98; TLI=0.98; RMSEA=0.04; and SRMR=0.06), indicating strict invariance across age groups.

Measurement Invariance for Resilience Resources

Measurement Invariance Across Gender

We evaluated measurement invariance for the Resilience Resources model across gender (women: n=164 and men: n=142) using multigroup CFA. The configural model demonstrated good fit (χ^2_{886}=1223.58; *P*<.001; CFI=0.94; TLI=0.93; RMSEA=0.04; and SRMR=0.08), indicating a consistent factor structure across groups. Constraining factor loadings in the metric model yielded similar fit (χ^2_{917}=1288.44; *P*<.001; CFI=0.93; TLI=0.92; RMSEA=0.04; and SRMR=0.08), supporting equivalence of loadings. The scalar model, which additionally constrained item intercepts, also showed good fit (χ^2_{949}=1360.84; *P*<.001; CFI=0.92; TLI=0.92; RMSEA=0.04; and SRMR=0.09). The strict model, which further constrained residual variances, continued to meet fit thresholds (χ^2_{988}=1413.10; *P*<.001; CFI=0.92; TLI=0.92; RMSEA=0.04; and SRMR=0.09), supporting full measurement invariance across gender.

Measurement Invariance Across Race

We next assessed invariance across race, comparing White (n=211) and non-White (n=95) participants. The configural model demonstrated good fit (χ^2_{886}=1326.95; *P*=.01; CFI=0.92; TLI=0.91; RMSEA=0.04; and SRMR=0.07), indicating a consistent factor structure. The metric model, which constrained factor loadings, showed similar fit (χ^2_{917}=1363.49; *P*<.001; CFI=0.92; TLI=0.91; RMSEA=0.04; SRMR=0.07). The scalar model, which additionally constrained item intercepts, maintained good fit (χ^2_{949}=1433.49; *P*<.001; CFI=0.91; TLI=0.91; RMSEA=0.04; and SRMR=0.07). The strict model, which further constrained residual variances, also met fit thresholds (χ^2_{988}=1540.42; *P*<.001; CFI=0.90; TLI=0.90; RMSEA=0.04; and SRMR=0.07), supporting full measurement invariance across race.

Measurement Invariance Across Age Groups

Participants were grouped by age 20 - 39 years (n=175) and 40 - 64 years (n=131). The configural model demonstrated acceptable fit (χ^2_{886}=1352.55; *P*<.001; CFI=0.92; TLI=0.91; RMSEA=0.04; and SRMR=0.07). Metric invariance was supported with excellent fit when loadings were constrained (χ^2_{917}=1385.98; *P*<.001; CFI=0.92; TLI=0.91; RMSEA=0.04; and SRMR=0.08). The scalar model (adding intercept constraints) showed strong fit (χ^2_{949}=1458.55; *P*<.001; CFI=0.91; TLI=0.90; RMSEA=0.04; and SRMR=0.08), permitting latent mean comparisons. The strict model (residuals constrained) also fit well (χ^2_{988}=1542.30; *P*<.001; CFI=0.90; TLI=0.90; RMSEA=0.04; and SRMR=0.08), indicating strict invariance across age groups.

Weighting

Empirical Weight Distribution

The regression-derived empirical weights revealed a clear hierarchical structure dominated by psychological and stress-related constructs (Table 4). Emotion Regulation and Coping emerged as the most critical predictor (44.6%), followed

XSL•FO
RenderX

by Personal Stress (16.3%) and Perseverative Cognition (8.3%). Notably, Physical Activity received minimal empirical weighting (0.2%), likely reflecting mediation through other measured constructs rather than a lack of theoretical importance.

Table . Empirical, theoretical, and final hybrid weights.

Construct	Empirical weight, %	Theoretical weight, %	Hybrid weight, %	Final weight, %
Emotion Regulation and Coping	44.6	20.0	32.3	31
Personal Stress	16.3	10.0	13.2	13
Social Connectedness	7.8	18.0	12.9	12
Perseverative Thinking	8.3	14.0	11.2	11
Sleep	6.9	9.0	7.9	8
Dietary Intake	6.2	5.0	5.6	5
Burnout	5.3	5.0	5.2	5
Physical Activity	0.2	7.0	3.6	5
Work Stress	1.0	7.0	4.0	5
Purpose and Prosociality	3.9	5.0	4.5	5

Theoretical Weight Rationale

Theoretical weights addressed empirical limitations while ensuring comprehensive coverage of modifiable resilience factors. Social Connectedness received high theoretical weighting based on robust meta-analytic evidence linking social support to stress resilience outcomes [66,67]. Sleep and Physical Activity received substantial theoretical weights given their well-documented roles in stress buffering and physiological recovery mechanisms [68-70].

Work Stress received meaningful theoretical weighting despite minimal empirical contribution, reflecting complex nonlinear relationships observed in organizational research, where higher work demands correlate positively with resilience in certain populations [71], suggesting stress inoculation effects [72] or self-selection mechanisms.

Hybrid Weight Validation

Cross-sectional validation demonstrated superior performance of the hybrid weighting approach across all criterion measures (Table 5). The weighted approach showed consistent improvements over unweighted alternatives, with particularly strong gains for primary resilience measures: CD-RISC (+0.06 vs unweighted) and BRS (+0.08 vs unweighted).

Table . Validation performance comparison.

Outcome measure	Hybrid weighted	Unweighted	Item-level	Best performing version
CD-RISC[a]	0.75[b]	0.68[b]	0.73[b]	Weighted
BRS[c]	0.74[b]	0.67[b]	0.71[b]	Weighted
PSS-4[d]	−0.81[b]	−0.77[b]	−0.79[b]	Weighted
PHQ-8[e]	−0.77[b]	−0.76[b]	−0.77[b]	Item-level
PROMIS-SF-8a[f]	−0.79[b]	−0.77[b]	−0.79[b]	Weighted
GAD-7[g]	−0.77[b]	−0.74[b]	−0.73[b]	Weighted
WHO-5[h]	0.83[b]	0.81[b]	0.84[b]	Item-level

[a]CD-RISC: Connor-Davidson Resilience Scale.

[b]$P<.001$

[c]BRS: Brief Resilience Scale.

[d]PSS-4: Perceived Stress Scale-4.

[e]PHQ-8: Patient Health Questionnaire-8.

[f]PROMIS-SF-8a: Patient-Reported Outcomes Measurement Information System Short Form 8a.

[g]GAD-7: Generalized Anxiety Disorder-7.

[h]WHO-5: World Health Organization-5 Well-Being Index.

XSL•FO
RenderX

Validation Performance Summary

The hybrid approach demonstrated meaningful improvements across resilience-related outcomes while maintaining strong correlations with mental health indicators. The greatest improvements occurred for measures most directly assessing resilience capacity (CD-RISC and BRS), suggesting the theoretical components successfully enhanced the measurement of core resilience constructs.

Notably, the hybrid approach outperformed both purely unweighted and item-level averaging approaches, indicating that the systematic integration of empirical prediction with theoretical importance creates optimal measurement properties. The modest improvement for depression (PHQ-8) likely reflects the already strong empirical weighting of constructs most relevant to depressive symptoms.

Final Weight Structure and Clinical Implications

The final hybrid weights create a theoretically coherent hierarchy that preserves predictive validity while ensuring comprehensive intervention guidance. Emotion Regulation and Coping maintains prominence, consistent with its central role in stress resilience frameworks. Core stress and cognitive constructs receive substantial representation: Personal Stress, Social Connectedness, and Perseverative Thinking.

Critically, all behavioral and contextual factors achieve meaningful weighting for intervention purposes: Sleep, Physical Activity, Work Stress, Dietary Intake, Burnout, and Compassion/Gratitude/Meaning. This distribution supports the WONE Index's dual function as both a predictive assessment and behavioral intervention guidance system.

Longitudinal validation using Time 1 WONE scores to predict Time 2 outcomes (n=203) further confirmed the superior predictive performance of the weighted approach across all validation measures, though detailed longitudinal analyses are beyond the scope of the current phase.

Discussion

Principal Findings

The WONE Index was developed and validated through a 2-phase process to progressively refine its measurement model. Phase 1 provided strong initial psychometric evidence, establishing a 6-factor structure organized into 2 higher-order domains: Stress Load (Work Stress, Personal Stress, and Burnout) and Resilience Resources (Resilience Skills and Beliefs, Social Support, and Sleep). While statistically robust, this model did not fully capture the multidimensional and dynamic nature of resilience, particularly with respect to health behaviors and cognitive processes.

Phase 2 expanded and refined this framework into a 10-factor structure organized within 2 higher-order domains, which we propose as the finalized structure. Seven resource-related factors (Emotion Regulation and Coping, Social Connection, Purpose and Prosociality, Sleep, Physical Activity, Dietary Intake, and Perseverative Thinking) and 3 stress factors (Work Stress, Personal Stress, and Burnout) were identified and grouped into the higher-order domains of Stress Load and Resilience Resources. In both phases, we followed standard CFA practice by incorporating theoretically justified modifications suggested by MIs, including correlated residuals within conceptually related item sets. These refinements improved model fit while preserving the theoretical structure. Model fit indices for both Phase 1 and Phase 2 were strong, but the Phase 2 model offered comparable psychometric performance while adding domains that improved interpretability and intervention relevance, making it the preferred structure. By incorporating additional factors, the final model allows for greater precision in identifying areas of vulnerability and creates more actionable pathways for intervention.

Importantly, the Index allows us to move beyond identifying whether someone is "resilient" to understanding how their system is functioning. It captures systemic imbalances in how individuals perceive, feel, adapt, and respond to stress, thereby holistically measuring resilience and highlighting where interventions should target. For example, someone experiencing high stress load but with depleted resources may benefit most from micro-moment stress-reduction strategies that support daily functioning before resource-building can take hold. Conversely, individuals with low external stressors but insufficient resilience resources may require preventive interventions to expand their repertoire of coping and regulatory skills.

This perspective also acknowledges paradoxical cases such as "skin-deep resilience," in which individuals may appear psychologically resilient to chronic stress, yet their bodies still carry physiological costs (eg, accelerated aging, immune dysregulation, and cardiovascular risk). In such cases, building resilience resources remains essential, as resource strengthening may buffer against hidden biological wear-and-tear even when stress load is not consciously perceived as high [73,74]. By integrating both stress load and resilience resources, the WONE Index provides a nuanced lens for identifying vulnerability, tailoring interventions, and supporting resilience development across populations and contexts.

The WONE Index also demonstrated excellent temporal stability, strong reliability across scales, and robust validity with multiple established measures of resilience, stress, and mental health outcomes. Together, this evidence supports the Index as both a rigorous scientific tool and a practical framework for understanding resilience in diverse populations.

The hybrid weighting methodology is a key innovation of this work. It integrates empirical weights derived from predictive modeling of mental health and well-being outcomes with theoretical weights grounded in resilience science and dynamical systems constructs (eg, Sleep, Social Connection, and Emotion Regulation). This dual approach produces scores that are both predictively robust and have practical utility in digital health and applied psychology.

Conceptually, this framework aligns with multidimensional outcome systems such as the Treatment Outcome Package [75], which emphasizes comprehensive assessment across multiple domains to guide individualized feedback and treatment. The WONE Index adopts a similar logic by balancing empirical evidence with theoretical modifiability, ensuring that domain

XSL•FO
RenderX

scores capture both predictive value and potential for change through intervention. This approach enhances the Index's relevance for digital health applications by identifying high-impact targets while also laying the foundation for future integration of predictive modeling approaches to further enhance precision and personalization. In doing so, it bridges traditional clinical feedback models with data-driven precision frameworks, offering a scalable approach to personalized resilience assessment.

Together, these findings establish the WONE Index as a comprehensive, multidimensional, and psychometrically rigorous measure of resilience, conceptualized as adaptive capacity emerging from the balance between stress load and resilience resources, and designed to both measure resilience and inform targeted intervention.

Comparison With Previous Work

O'Donohue et al [76] identified 45 distinct measurement approaches with substantial heterogeneity in stress resilience conceptualization and operationalization. Critical limitations documented across existing measures include: (1) most measures assess stress or resilience separately rather than their dynamic interaction, with only 17.5% using resilience measures and many relying solely on stress indices; (2) predominant use of trait-based rather than process-based conceptualizations (eg, CD-RISC, BRS, and RSA which focus on stable characteristics), limiting sensitivity to temporal change; and (3) limited validation for repeated measurement contexts, with most instruments designed for single-timepoint assessment [8]. In addition, widely used resilience measures—including the CD-RISC, BRS, and RSA—focus predominantly on perceptual attitudes ("I can deal with whatever comes") rather than behaviorally specific, modifiable resources, and do not integrate health behavior determinants—including sleep, physical activity, and nutrition—documented as protective factors that buffer stress and promote resilience [20,21,77-80].

The WONE Index addresses these gaps through 3 key distinctions. First, it integrates stress load and resilience resources within a unified framework, operationalizing the demands-resources balance emphasized by JD-R theory rather than measuring constructs separately [12]. Second, it provides comprehensive domain coverage spanning psychological resources (emotion regulation, coping, and perseverative cognition), social resources (connection and support), meaning-oriented resources (purpose, gratitude, and prosociality), and health behavioral resources (sleep, physical activity, and nutrition)—a combination rarely found in existing validated measures to our knowledge. Third, it emphasizes behaviorally specific, modifiable capacities providing concrete intervention targets for digital health applications. However, longitudinal validation is needed to establish sensitivity to intervention-related change, and comparative effectiveness research must determine whether this integrated approach provides advantages over using established measures in combination, while independent replication by researchers without commercial affiliations is essential for confirming generalizability across diverse populations and contexts.

The WONE Index contributes to a growing literature emphasizing that stress and resilience are not opposite poles of a single construct but distinct domains with unique predictive value. Similar to existing resilience frameworks [4,11], our results underscore the centrality of regulatory skills, social connection, emotion, and meaning-making, but the Index also extends previous tools in several important ways. Unlike widely used measures that treat demands and resources as isolated constructs (eg, PSS [18], CD-RISC [20], and BRS [21]), the Index embeds them in a unified structure reflecting how these factors interact as interdependent subsystems. This operationalization enables the measure to serve dual purposes: scientifically, it captures multidomain adaptive capacity aligned with JD-R and systems-based models; practically, it functions as an intervention guidance system that balances measurement rigor with intervention utility.

The 10-factor structure within 2 higher-order domains provides a multidimensional systems approach to resilience that begins to approximate a network of interdependent resilience domains by capturing distinct contributions from cognitive, behavioral, social, and emotional domains, each of which showed robust factor loadings and validity evidence in Phase 2. Cognitive (eg, perseverative thinking), social (social connection and purpose/prosociality), emotion (eg, emotion regulation and positive affect), health behavior (eg, sleep, physical activity, and diet), and stress exposures (eg, personal, work, and burnout) are all included, creating a systemic framework rather than a 1D scale, which has been popular in established measures of resilience.

This expanded domain coverage forms the foundation of its systems perspective. This aligns with allostatic models, JD-R, and complex dynamical systems models of psychopathology, which emphasize how individuals may become "stuck" in maladaptive states or transition toward adaptive states depending on system-level dynamics [81]. By conceptualizing resilience as adaptive capacity emerging from the balance between stress load and resilience resources, the WONE Index provides a practical operationalization of these theoretical models.

As mentioned in the previous paragraph, the WONE index builds directly on the JD-R framework [12,13], which posits that health, well-being, and performance are determined by the interplay between demands (eg, workload, role conflict, and emotional strain) and resources (eg, social support, coping strategies, and recovery experiences). Our higher-order domains map closely to this model: Stress Load reflects job and life demands, while Resilience Resources reflect the protective assets available to meet them. Importantly, the JD-R framework distinguishes acute strain driven by demands from the chronic exhaustion that develops into burnout. The WONE Index mirrors this theoretical distinction by separately modeling work stress and burnout, offering a psychometric tool for examining how short-term strain may evolve into longer-term depletion and maladaptive system states.

Beyond alignment, the Index extends JD-R in 2 important ways. First, it broadens the scope by incorporating personal, societal, and health behavior domains, increasing relevance across multiple life contexts rather than being limited to occupational

XSL•FO
RenderX

settings. Second, the hybrid weighting methodology adds precision to JD-R by quantifying the relative impact of different resources. Whereas JD-R theory broadly emphasizes the buffering role of resources, the WONE Index specifies which domains—such as sleep, coping, or social connection—are both theoretically central and empirically predictive of outcomes such as depression, anxiety, and well-being. In this way, the Index operationalizes the JD-R concept of resources while also extending it to a wider range of life demands and contexts.

Applications to Digital Mental Health

The WONE Index advances digital mental health measurement in several important ways.

First, it was developed and validated specifically within digital delivery contexts, ensuring its psychometric properties hold when administered via web and mobile interfaces rather than assuming transferability from traditional formats. The index was validated via a digital survey/platform, providing confidence that observed reliability and validity reflect real-world digital administration rather than performance under idealized controlled conditions.

Second, the Index enables a measurement-based care approach at scale. Traditional assessment typically occurs at discrete time points (eg, intake and discharge) with results informing clinician-delivered interventions. In contrast, the WONE Index supports continuous assessment cycles where monthly administration generates updated resilience profiles that automatically inform algorithmic personalization of intervention content. This creates feedback loops between assessment and intervention that are difficult to achieve in traditional service delivery models, allowing platforms to adaptively adjust recommendations as users' stress-resource balance shifts over time.

Third, the multidimensional structure provides actionable granularity for automated intervention matching. Rather than producing a single resilience score requiring clinician interpretation to determine intervention targets, the 10 factors within 2 overarching domains assessed in the WONE Index directly map onto intervention modules within the platform. Low scores on emotion regulation trigger recommendations for regulation-focused techniques, depleted social connection scores prompt social relationship-focused interventions, and poor sleep activates sleep hygiene interventions. This direct assessment-to-intervention mapping enables digital platforms to deliver truly personalized care pathways without requiring human clinical judgment for every user.

Fourth, the measure's design supports population-level insights alongside individual assessment. Aggregated anonymized data reveal organizational stress patterns, high-risk subgroups, and systemic factors affecting workforce resilience. This dual-level utility, individual personalization plus population surveillance, distinguishes the Index from traditional measures designed solely for individual clinical assessment. Organizations can identify when specific teams show declining resilience resources or elevated stress load, enabling proactive organizational interventions before individual crises emerge.

Finally, the hybrid weighting methodology reflects digital health's unique requirements. Pure empirical weighting maximizes prediction but may undervalue behaviorally modifiable domains showing weaker cross-sectional associations yet greater intervention responsiveness. Theoretical weighting ensures the Index prioritizes domains where digital platforms can deliver effective interventions (eg, sleep, physical activity, and social connection) even when these show modest concurrent prediction. This methodology acknowledges that digital health assessment serves intervention guidance, not just prediction, requiring weights that balance predictive validity with behavioral actionability.

Together, these features distinguish the WONE Index from existing stress and resilience measures, which typically assess either protective or risk factors in isolation, rely on static trait-based designs, and lack validation for digital health or workplace contexts. By integrating stress load and resilience resources within a unified, multidomain structure specifically designed for digital mental health contexts and linking assessment directly to intervention personalization, the WONE Index operationalizes the demands-resources framework for temporal tracking through repeated administration in applied digital settings.

Strengths

This study has several notable strengths. First, it used a multiphase design, beginning with exploratory development and culminating in confirmatory validation. This sequential approach provides both a rigorous psychometric foundation and evidence of generalizability. Second, the WONE Index advances the field by integrating multiple domains of resilience into a hierarchical structure, capturing stress load alongside resilience resources. This systemic framing reflects how resilience operates across cognitive, social, behavioral, and meaning-oriented processes rather than as a single dimension.

Third, the Index demonstrated robust psychometric qualities, including strong reliability, validity across multiple criteria, and temporal stability, supporting its use for both research and applied purposes. Fourth, the hybrid weighting methodology represents a novel contribution to measuring development, balancing empirical prediction with theoretical modifiability and enhancing the practical utility of final scores.

Limitations

Several limitations should also be noted. First, the data were cross-sectional, which precludes causal inference. While test-retest analyses support temporal stability, the design limits the ability to determine whether changes in stress load or resilience resources precede improvements in outcomes. Future longitudinal studies are needed to examine how resilience processes evolve over time and in response to intervention.

Second, the study relied on self-report measures, which introduces potential for bias, such as recall inaccuracy or social desirability effects. We sought to mitigate these issues by using validated instruments with well-established psychometric properties and by using a brief, single-session survey format to minimize fatigue-related error. Nevertheless, self-report bias could have influenced observed relationships among constructs.

XSL•FO
RenderX

Future studies should integrate behavioral and physiological indicators (eg, ecological momentary assessment, wearables, and biomarkers) to strengthen validity and reduce reliance on self-report alone.

Third, although the sample was diverse and measurement invariance was supported across gender, race, and age, the use of online convenience sampling (CloudResearch and Prolific) may limit generalizability to digitally-engaged populations. Participants were primarily from the United States and the United Kingdom and were predominantly White, which may constrain cultural generalizability. Accordingly, future research should aim to validate the WONE Index across more diverse cultural, linguistic, and global populations to ensure measurement equivalence and capture potential cultural differences in how stress load and resilience resources manifest.

Fourth, brief item pools for some domains such as diet and physical activity, while necessary for reducing participant burden, may have limited construct breadth. The Dietary Intake factor, in particular, showed lower reliability, representing a known measurement limitation. We retained this domain due to its theoretical relevance and behavioral actionability in digital health contexts.

In addition, these health behavior domains may capture variance from aspects of resilience not fully represented in other adaptive resource domains (eg, energy balance or embodied forms of adaptation). It is also possible that shared variance among closely related constructs may have partially attenuated their distinct contributions. Future research should be mindful that diet and physical activity may relate more strongly to multimodal and objective data sources (eg, wearables and ecological monitoring) given their greater overlap with physiological processes compared to psychosocial resources.

Further, while the hybrid weighting system enhances both predictive power and practical relevance, the process of assigning theoretical weights remains partly subjective. We sought to balance theoretical rationale with empirical model fit, but weighting decisions could still introduce bias in the relative importance of certain domains. Future work should refine these weightings through stakeholder engagement and cross-validation in applied contexts.

Finally, the WONE Index was developed within a digital health implementation context, providing access to large-scale user samples and enabling iterative refinement based on real-world engagement. While this design addresses documented measurement gaps through integrated assessment of stress and resilience resources, establishing whether this approach provides practical advantages over existing measures requires additional validation. Future research should examine predictive validity for clinical outcomes and intervention response, longitudinal sensitivity to change in intervention contexts, and generalizability across demographic and cultural groups.

Comparative effectiveness research will be valuable to determine whether integrated assessment offers advantages over domain-specific measures used in combination, or whether different measurement approaches serve complementary purposes depending on the implementation context. Independent validation by researchers without commercial affiliations will also be valuable for confirming generalizability and providing an unbiased evaluation of the measure's comparative utility across different use cases.

Future Directions

This work lays the foundation for several important lines of future research. First, longitudinal studies are needed to examine how resilience resources and stress load interact over time and to test whether the WONE Index is sensitive to intervention-related changes. Such research will help clarify the dynamic nature of adaptive capacity and resilience plasticity and help examine whether certain domains play different roles in resilience decline versus recovery. This work may also reveal that the factors contributing to stress-related problems differ from those most critical for recovery, information that could further refine how digital platforms prioritize and sequence intervention recommendations to maximize impact during different phases of the resilience process.

Second, clinical applications represent a critical next step. Embedding the WONE Index into treatment contexts could provide insights into how resilience factors contribute to therapeutic outcomes, as well as identify which domains serve as leverage points for recovery. The hierarchical, multidomain structure is particularly well-suited for monitoring differential change across domains during intervention.

Third, contextual and multimethod integration holds promise for advancing ecological validity. Pairing the Index with ecological momentary assessment could capture the contexts in which stress triggers arise and the adaptive strategies mobilized in response, illuminating how resilience unfolds in daily life. This could be further enhanced through pairing with objective markers, including wearable-derived measures (eg, sleep patterns, activity, and heart rate variability) and biomarkers (eg, cortisol and inflammatory cytokines). Such multimethod approaches would reduce reliance on self-report alone, allow researchers to examine how subjective and biological processes align or diverge, and provide a more complete picture of adaptive capacity as a multilevel system.

Fourth, cross-cultural research is needed to explore how resilience processes manifest across diverse sociocultural contexts. While the Index demonstrated measurement invariance across gender, race, and age, expanding to other cultural and occupational groups would strengthen its generalizability and highlight context-specific resilience mechanisms.

Finally, continued methodological innovation will be essential for advancing resilience science. While the hybrid weighting system provides a balanced approach to scoring, the WONE Index also creates opportunities for more advanced analytic methods that can further expand resilience science. Machine learning can extend beyond traditional statistical models by prioritizing prediction and generalizability. It can leverage the Index's structure to detect complex nonlinear interactions, uncover latent resilience profiles through clustering methods, and integrate multimodal data (eg, self-report, wearables, and biomarkers). Machine learning's emphasis on out-of-sample

XSL•FO
RenderX

accuracy and scalability makes it particularly useful for precision prediction in digital health contexts.

Future research can also incorporate Bayesian and systems-based approaches to deepen mechanistic understanding. Bayesian approaches can model uncertainty around resilience estimates, incorporate previous theoretical knowledge, and improve estimation in smaller or intensive longitudinal samples such as ecological momentary assessment. Additional frameworks, including network analysis, dynamical systems modeling, and mixture modeling, could further illuminate how resilience operates as a system, identify leverage points, and detect early warning signals of maladaptive change. Together, these approaches position the WONE Index as both a measurement tool and a platform for advancing predictive and mechanistic models of resilience.

Conclusion

The WONE Index successfully bridges scientific rigor and practical utility, addressing the assessment-intervention gap that has limited resilience research impact in applied settings. Strong incremental validity beyond gold-standard measures demonstrates that the Index captures unique aspects of stress-resilience capacity not assessed by existing tools. The methodological innovation of hybrid weighting—balancing empirical prediction with theoretical modifiability—establishes a strengthened standard for intervention-focused measurement development.

By simultaneously assessing stress load and resilience resources within a unified framework specifically designed for digital delivery, the Index enables personalized intervention matching at scale while maintaining rigorous psychometric standards. As digital mental health expands, measures that satisfy both scientific and practical requirements will be essential for enhancing treatment effectiveness and accessibility. The WONE Index provides a scientifically grounded foundation for this evolution, serving as both a research tool and a platform-integrated assessment system.

Acknowledgments

We would like to thank Carolina Estevao for lending her expertise to an earlier version of the WONE Index, as well as Gabriella Bergin-Cartwright for her support in managing data collection for this project. We would also like to thank the participants of this study, as well as the employers who make this platform available to their employees. Portions of this manuscript were developed with the assistance of generative artificial intelligence (ChatGPT-4 from OpenAI; Claude AI, Anthropic). These tools were used to assist in preparing the manuscript's initial outline, resolving statistical software troubleshooting, and editorial refinement. Any AI-generated content served purely as a drafting aid and was critically reviewed and verified for accuracy by the authors and reworded by the authors prior to inclusion. The authors take full responsibility for the integrity and accuracy of the final manuscript content.

Funding

This study was funded by Walking on Earth.

Data Availability

Individual deidentified data that underlie the results reported in this manuscript can be shared privately for research purposes upon receipt of a methodologically sound proposal whose proposed use of the data is approved by the authors and Walking on Earth's legal and security teams. To gain access, requesters will need to submit a proposal to the corresponding author and sign a data access agreement that includes a commitment to (1) using the data only for research purposes; (2) not attempting to or actually reidentifying any individual; (3) securing the data using appropriate safeguards; and (4) destroying or returning the data after analyses are completed. The WONE Index is a proprietary measure copyrighted by Walking on Earth, Inc. Full administration materials, including detailed instructions and response options, are not publicly available. However, Walking on Earth welcomes collaboration and research partnerships, and the measure may be made available for use upon reasonable request via the company's website or by emailing science@walkingonearth.com.

Authors' Contributions

LGR led the conceptualization of the study and was responsible for data curation, formal analysis, investigation, and methodology development. LG also oversaw project administration and supervision and contributed to both the original drafting and the review and editing of the manuscript. DG contributed to the study through formal analysis, methodological support, and visualization, and participated in writing the original draft as well as reviewing and editing the manuscript. KJ contributed to the visualization and participated in writing the original draft, along with reviewing and editing the manuscript. RM contributed to the study's conceptualization and provided key resources, in addition to participating in the review and editing of the manuscript.

Conflicts of Interest

LGR and RM are employees of Walking on Earth and have received a salary and stock options from the company.

Multimedia Appendix 1

XSL•FO
RenderX

Power analyses for Phase 1 (N=1005) and Phase 2 (N=306), Phase 1 and Phase 2 participant demographics (Tables S1-S2), Phase 1 CFA factor loadings and communalities (Table S3), Phase 1 internal consistency statistics (Table S4), Phase 1 HTMT discriminant validity matrix (Table S5), correlations with established measures (Table S6), comparative fit index (CFA) model modifications, and **Phase 2** heterotrait-monotrait (HTMT) discriminant validity results (Tables S7-S8).
[PDF File, 230 KB - jmir_v28i1e81714_app1.pdf]

References

1. Mahmud S, Mohsin M, Dewan MN, Muyeed A. The global prevalence of depression, anxiety, stress, and insomnia among general population during COVID-19 pandemic: a systematic review and meta-analysis. Trends Psychol 2023;31(1):143-170. [doi: 10.1007/s43076-021-00116-9] [Medline: 40477944]
2. Slavich GM. Social Safety Theory: understanding social stress, disease risk, resilience, and behavior during the COVID-19 pandemic and beyond. Curr Opin Psychol 2022 Jun;45:101299. [doi: 10.1016/j.copsyc.2022.101299] [Medline: 35219156]
3. Masten AS. Resilience in developmental systems: principles, pathways, and protective processes in research and practice. In: Multisystemic Resilience: Adaptation and Transformation in Contexts of Change: Oxford University Press; 2021:113-134.
4. Bonanno GA, Romero SA, Klein SI. The temporal elements of psychological resilience: an integrative framework for the study of individuals, families, and communities. Psychol Inq 2015 Apr 3;26(2):139-169. [doi: 10.1080/1047840X.2015.992677]
5. Southwick SM, Charney DS. The science of resilience: implications for the prevention and treatment of depression. Science 2012 Oct 5;338(6103):79-82. [doi: 10.1126/science.1222942] [Medline: 23042887]
6. Insel T. Digital mental health care: five lessons from Act 1 and a preview of Acts 2-5. NPJ Digit Med 2023 Jan 26;6(1):9. [doi: 10.1038/s41746-023-00760-8] [Medline: 36702920]
7. Torous J, Linardon J, Goldberg SB, et al. The evolving field of digital mental health: current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. World Psychiatry 2025 Jun;24(2):156-174. [doi: 10.1002/wps.21299] [Medline: 40371757]
8. Windle G, Bennett KM, Noyes J. A methodological review of resilience measurement scales. Health Qual Life Outcomes 2011 Feb 4;9(1):8. [doi: 10.1186/1477-7525-9-8] [Medline: 21294858]
9. Hill Y, Dolezal ML, Nordbeck PC, et al. Moving from traits to the dynamic process: the next steps in research on human resilience. J Aggress Maltreat Trauma 2025 Jul 3;34(7):971-989. [doi: 10.1080/10926771.2024.2431733]
10. Luthar SS, Cicchetti D, Becker B. The construct of resilience: a critical evaluation and guidelines for future work. Child Dev 2000;71(3):543-562. [doi: 10.1111/1467-8624.00164] [Medline: 10953923]
11. Kalisch R, Baker DG, Basten U, et al. The resilience framework as a strategy to combat stress-related disorders. Nat Hum Behav 2017 Nov;1(11):784-790. [doi: 10.1038/s41562-017-0200-8] [Medline: 31024125]
12. Bakker AB, Demerouti E. Job demands-resources theory: taking stock and looking forward. J Occup Health Psychol 2017 Jul;22(3):273-285. [doi: 10.1037/ocp0000056] [Medline: 27732008]
13. Demerouti E, Bakker AB, Nachreiner F, Schaufeli WB. The job demands-resources model of burnout. J Appl Psychol 2001 Jun;86(3):499-512. [doi: 10.1037/0021-9010.86.3.499] [Medline: 11419809]
14. Hobfoll SE, Halbesleben J, Neveu JP, Westman M. Conservation of resources in the organizational context: the reality of resources and their consequences. Annu Rev Organ Psychol Organ Behav 2018 Jan 21;5(1):103-128 [FREE Full text] [doi: 10.1146/annurev-orgpsych-032117-104640]
15. Biggs A, Brough P, Drummond S. Lazarus and Folkman's psychological stress and coping theory. In: Cooper CL, Quick JC, editors. The Handbook of Stress and Health, 1st edition 2017. [doi: 10.1002/9781118993811]
16. Folkman S, Moskowitz JT. Coping: pitfalls and promise. Annu Rev Psychol 2004;55(1):745-774. [doi: 10.1146/annurev.psych.55.090902.141456] [Medline: 14744233]
17. Lazarus RS, Folkman S. Stress, Appraisal, and Coping: Springer; 1984.
18. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [doi: 10.2307/2136404] [Medline: 6668417]
19. Cohen S, Wills TA. Stress, social support, and the buffering hypothesis. Psychol Bull 1985 Sep;98(2):310-357. [doi: 10.1037/0033-2909.98.2.310] [Medline: 3901065]
20. Connor KM, Davidson JRT. Development of a new resilience scale: the Connor-Davidson Resilience Scale (CD-RISC). Depress Anxiety 2003;18(2):76-82. [doi: 10.1002/da.10113] [Medline: 12964174]
21. Smith BW, Dalen J, Wiggins K, Tooley E, Christopher P, Bernard J. The brief resilience scale: assessing the ability to bounce back. Int J Behav Med 2008;15(3):194-200. [doi: 10.1080/10705500802222972] [Medline: 18696313]
22. MacCallum RC, Browne MW, Sugawara HM. Power analysis and determination of sample size for covariance structure modeling. Psychol Methods 1996;1(2):130-149. [doi: 10.1037//1082-989X.1.2.130]
23. Bonett DG. Sample size requirements for testing and estimating coefficient alpha. J Educ Behav Stat 2002 Dec;27(4):335-340. [doi: 10.3102/10769986027004335]
24. Zou GY. Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. Stat Med 2012 Dec 20;31(29):3972-3981. [doi: 10.1002/sim.5466] [Medline: 22764084]

XSL•FO
RenderX

25. Kaiser BN, Haroz EE, Kohrt BA, Bolton PA, Bass JK, Hinton DE. "Thinking too much": a systematic review of a common idiom of distress. Soc Sci Med 2015 Dec;147:170-183. [doi: 10.1016/j.socscimed.2015.10.044] [Medline: 26584235]
26. Lewis-Fernández R, Kirmayer LJ. Cultural concepts of distress and psychiatric disorders: understanding symptom experience and expression in context. Transcult Psychiatry 2019 Aug;56(4):786-803. [doi: 10.1177/1363461519861795] [Medline: 31347476]
27. Schmid RF, Thomas J, Rentzsch K. Individual differences in parasympathetic nervous system reactivity in response to everyday stress are associated with momentary emotional exhaustion. Sci Rep 2024 Nov 4;14(1):26662. [doi: 10.1038/s41598-024-74873-9] [Medline: 39496636]
28. Jerotic S, Ignjatovic N, Maric NP, et al. A comparative study on mental disorder conceptualization: a cross-disciplinary analysis. Community Ment Health J 2024 May;60(4):813-825. [doi: 10.1007/s10597-024-01240-3] [Medline: 38319528]
29. Tweed RG, White K, Lehman DR. Culture, stress, and coping: internally- and externally-targeted control strategies of European Canadians, East Asian Canadians, and Japanese. J Cross-Cult Psychol 2004;35(6):652-668. [doi: 10.1177/0022022104270109]
30. Chentsova-Dutton YE, Ryder AG, Tsai JL. Understanding depression across cultural contexts. In: Handbook of Depression, 3rd edition: The Guilford Press; 2014.
31. Matthews G, Szalma J, Panganiban AR, Neubauer C, Warm JS. Profiling task stress with the dundee stress state questionnaire. In: Cavalcanti L, Azevedo S, editors. Psychology of Stress: Nova; 2013:49-91.
32. Maslach C, Schaufeli WB, Leiter MP. Job burnout. Annu Rev Psychol 2001;52(1):397-422. [doi: 10.1146/annurev.psych.52.1.397] [Medline: 11148311]
33. Schaufeli WB, Leiter MP, Maslach C. Burnout: 35 years of research and practice. Career Development International 2009 Jun 19;14(3):204-220. [doi: 10.1108/13620430910966406]
34. Lee EH. Review of the psychometric evidence of the Perceived Stress Scale. Asian Nurs Res (Korean Soc Nurs Sci) 2012 Dec;6(4):121-127. [doi: 10.1016/j.anr.2012.08.004] [Medline: 25031113]
35. Figalová N, Charvát M, Univerzita Palackého v Olomouci, Czech Republic. The Perceived Stress Scale: reliability and validity study in the Czech Republic. Ceskoslov psychol 2021;65(1):46-59. [doi: 10.51561/cspsych.65.1.46]
36. Campbell-Sills L, Stein MB. Psychometric analysis and refinement of the Connor-Davidson Resilience Scale (CD-RISC): validation of a 10-item measure of resilience. J Trauma Stress 2007 Dec;20(6):1019-1028. [doi: 10.1002/jts.20271] [Medline: 18157881]
37. Pilkonis PA, Choi SW, Reise SP, et al. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. Assessment 2011 Sep;18(3):263-283. [doi: 10.1177/1073191111411667] [Medline: 21697139]
38. Schalet BD, Pilkonis PA, Yu L, et al. Clinical validity of PROMIS depression, anxiety, and anger across diverse clinical samples. J Clin Epidemiol 2016 May;73:119-127. [doi: 10.1016/j.jclinepi.2015.08.036] [Medline: 26931289]
39. Moazzami M, Katz P, Bonilla D, et al. Validity and reliability of patient reported outcomes measurement information system computerized adaptive tests in systemic lupus erythematous. Lupus (Los Angel) 2021 Nov;30(13):2102-2113. [doi: 10.1177/09612033211051275]
40. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
41. Löwe B, Decker O, Müller S, et al. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. Med Care 2008 Mar;46(3):266-274. [doi: 10.1097/MLR.0b013e318160d093] [Medline: 18388841]
42. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167-176. [doi: 10.1159/000376585] [Medline: 25831962]
43. Hajos TRS, Pouwer F, Skovlund SE, et al. Psychometric and screening properties of the WHO-5 well-being index in adult outpatients with type 1 or type 2 diabetes mellitus. Diabet Med 2013 Feb;30(2):e63-e69. [doi: 10.1111/dme.12040] [Medline: 23072401]
44. Tabachnick BG, Fidell LS. Using Multivariate Statistics, 7th edition: Pearson; 2019.
45. Fu Y, Wen Z, Wang Y. A comparison of reliability estimation based on confirmatory factor analysis and exploratory structural equation models. Educ Psychol Meas 2022 Apr;82(2):205-224. [doi: 10.1177/00131644211008953] [Medline: 35185157]
46. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. J Mark Res 1981 Feb;18(1):39-50. [doi: 10.1177/002224378101800104]
47. Pfefferbaum B, North CS. Mental health and the Covid-19 pandemic. N Engl J Med 2020 Aug 6;383(6):510-512. [doi: 10.1056/NEJMp2008017] [Medline: 32283003]
48. Stress in America 2023: a nation recovering from collective trauma. American Psychological Association. 2023. URL: https://www.apa.org/news/press/releases/stress/2023/collective-trauma-recovery [accessed 2025-12-13]
49. Slavich GM. Social Safety Theory: a biologically based evolutionary perspective on life stress, health, and behavior. Annu Rev Clin Psychol 2020 May 7;16(1):265-295. [doi: 10.1146/annurev-clinpsy-032816-045159] [Medline: 32141764]
50. Clark LA, Watson D. Constructing validity: basic issues in objective scale development. Psychol Assess 1995;7(3):309-319. [doi: 10.1037//1040-3590.7.3.309]

XSL•FO
RenderX

51. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
52. Shin C, Lee SH, Han KM, Yoon HK, Han C. Comparison of the usefulness of the PHQ-8 and PHQ-9 for screening for major depressive disorder: analysis of psychiatric outpatient data. Psychiatry Investig 2019 Apr;16(4):300-305. [doi: 10.30773/pi.2019.02.01] [Medline: 31042692]
53. Wu Y, Levis B, Riehm KE, et al. Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. Psychol Med 2020 Jun;50(8):1368-1380. [doi: 10.1017/S0033291719001314] [Medline: 31298180]
54. Wells TS, Horton JL, LeardMann CA, Jacobson IG, Boyko EJ. A comparison of the PRIME-MD PHQ-9 and PHQ-8 in a large military prospective study, the Millennium Cohort Study. J Affect Disord 2013 May 15;148(1):77-83. [doi: 10.1016/j.jad.2012.11.052] [Medline: 23246365]
55. Rammstedt B, John OP. Measuring personality in one minute or less: a 10-item short version of the Big Five Inventory in English and German. J Res Pers 2007 Feb;41(1):203-212. [doi: 10.1016/j.jrp.2006.02.001]
56. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modeling 1999 Jan;6(1):1-55. [doi: 10.1080/10705519909540118]
57. Kline RB. Principles and Practice of Structural Equation Modeling, 4th edition: Guilford Press; 2016.
58. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163. [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
59. Chen FF. Sensitivity of goodness of fit indexes to lack of measurement invariance. Struct Equ Modeling 2007 Jul 31;14(3):464-504. [doi: 10.1080/10705510701301834]
60. Liu JJW, Reed M, Girard TA. Advancing resilience: an integrative, multi-system model of resilience. Pers Individ Dif 2017 Jun;111:111-118. [doi: 10.1016/j.paid.2017.02.007]
61. McEwen BS, Gray J, Nasca C. Recognizing resilience: learning from the effects of stress on the brain. Neurobiol Stress 2015 Jan 1;1:1-11. [doi: 10.1016/j.ynstr.2014.09.001] [Medline: 25506601]
62. Epel ES, Crosswell AD, Mayer SE, et al. More than a feeling: a unified view of stress measurement for population science. Front Neuroendocrinol 2018 Apr;49:146-169. [doi: 10.1016/j.yfrne.2018.03.001] [Medline: 29551356]
63. Maslach C, Leiter MP. Understanding the burnout experience: recent research and its implications for psychiatry. World Psychiatry 2016 Jun;15(2):103-111. [doi: 10.1002/wps.20311] [Medline: 27265691]
64. Yuan KH, Bentler PM. 5. Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. Sociol Methodol 2000 Aug;30(1):165-200. [doi: 10.1111/0081-1750.00078]
65. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Lawrence Erlbaum Associates; 1988. URL: https://utstat.utoronto.ca/~brunner/oldclass/378f16/readings/CohenPower.pdf [accessed 2025-12-13]
66. Zell E, Stockus CA. Social support and psychological adjustment: a quantitative synthesis of 60 meta-analyses. Am Psychol 2025 Jan;80(1):33-46. [doi: 10.1037/amp0001323] [Medline: 38695783]
67. Vila J. Social support and longevity: meta-analysis-based evidence and psychobiological mechanisms. Front Psychol 2021;12:717164. [doi: 10.3389/fpsyg.2021.717164] [Medline: 34589025]
68. Chaput JP, Dutil C, Featherstone R, et al. Sleep timing, sleep consistency, and health in adults: a systematic review. Appl Physiol Nutr Metab 2020 Oct;45(10 (Suppl. 2)):S232-S247. [doi: 10.1139/apnm-2020-0032] [Medline: 33054339]
69. Flueckiger L, Lieb R, Meyer AH, Witthauer C, Mata J. The importance of physical activity and sleep for affect on stressful days: two intensive longitudinal studies. Emotion 2016 Jun;16(4):488-497. [doi: 10.1037/emo0000143] [Medline: 26709860]
70. Mücke M, Ludyga S, Colledge F, Gerber M. Influence of regular physical activity and fitness on stress reactivity as measured with the trier social stress test protocol: a systematic review. Sports Med 2018 Nov;48(11):2607-2622. [doi: 10.1007/s40279-018-0979-0] [Medline: 30159718]
71. Hakanen JJ, Schaufeli WB. Do burnout and work engagement predict depressive symptoms and life satisfaction? A three-wave seven-year prospective study. J Affect Disord 2012 Dec 10;141(2-3):415-424. [doi: 10.1016/j.jad.2012.02.043] [Medline: 22445702]
72. Seery MD. Resilience: a silver lining to experiencing adverse life events? Curr Dir Psychol Sci 2011;20(6):390-394. [doi: 10.1177/0963721411424740]
73. Chung KW, Martz CD, Lutz B, et al. Skin-deep resilience in the Black women's experiences living with lupus study. Health Psychol 2025 Aug;44(8):800-809. [doi: 10.1037/hea0001469] [Medline: 40193435]
74. Lowe SR. Embracing complexity in resilience research. Nat Mental Health 2025;3(4):391-392. [doi: 10.1038/s44220-025-00403-9]
75. Boswell JF, Kraus DR, Castonguay LG, Youn SJ. Treatment outcome package: measuring and facilitating multidimensional change. Psychotherapy (Chic) 2015 Dec;52(4):422-431. [doi: 10.1037/pst0000028] [Medline: 26641372]
76. O'Donohue JS, Mesagno C, O'Brien B. How can stress resilience be monitored? A systematic review of measurement in humans. Curr Psychol 2021 Jun;40(6):2853-2876. [doi: 10.1007/s12144-019-00226-9]
77. Friborg O, Hjemdal O, Rosenvinge JH, Martinussen M. A new rating scale for adult resilience: what are the central protective resources behind healthy adjustment? Int J Methods Psych Res 2003 Jun;12(2):65-76. [doi: 10.1002/mpr.143]

XSL•FO
RenderX

78. Leipold B, Klier K, Dapperger E, Schmidt A. Physical activity and nutrition in relation to resilience: a cross-sectional study. Sci Rep 2024 Jan 27;14(1):2272. [doi: 10.1038/s41598-024-52753-6] [Medline: 38280920]
79. Bertollo AG, Capitanio MZ, Schuh LA, Pradella N, Ignácio ZM. Habits and vulnerability or resilience to stress - impact on depressive disorders. Behav Brain Res 2025 Jul 26;490:115630. [doi: 10.1016/j.bbr.2025.115630] [Medline: 40334944]
80. Rink LC, Silva SG, Adair KC, Oyesanya TO, Humphreys JC, Sexton JB. The association between well-being behaviors and resilience in health care workers. West J Nurs Res 2022 Aug;44(8):743-754. [doi: 10.1177/01939459211017515] [Medline: 34039117]
81. Bringmann L, Helmich M, Eronen M, Voelkle M. Complex systems approaches to psychopathology. In: Krueger RF, Blaney PH, editors. Oxford Textbook of Psychopathology: Oxford University Press; 2023:103-122.

Abbreviations

AVE: average variance extracted
BFI-10: Big Five Inventory-10
BRS: Brief Resilience Scale
CD-RISC: Connor-Davidson Resilience Scale
CES-D: Center for Epidemiologic Studies Depression
CFA: confirmatory factor analysis
CFI: comparative fit index
CR: composite reliability
***DSM*:** *Diagnostic and Statistical Manual of Mental Disorders*
EFA: exploratory factor analysis
GAD-7: Generalized Anxiety Disorder-7
HTMT: heterotrait-monotrait
ICC: intraclass correlation coefficient
IRB: Institutional Review Board
JD-R: job demands-resources
KMO: Kaiser-Meyer-Olkin
MI: modification index
PHQ: Patient Health Questionnaire
PROMIS-SF-8a: Patient-Reported Outcomes Measurement Information System Short Form 8a
PSS: Perceived Stress Scale
RMSEA: root mean square error of approximation
SRMR: standardized root mean square residual
TLI: Tucker-Lewis index
WHO-5: World Health Organization-5 Well-Being Index
WONE: Walking on Earth
ΔR^2: R^2 change

Edited by N Cahill; submitted 01.Aug.2025; peer-reviewed by H Mahmoodi, NH Zainal; revised version received 07.Nov.2025; accepted 10.Nov.2025; published 05.Jan.2026.

Please cite as:
Roos LG, Gilliland D, Julian K, Misra R
The WONE Index as a Multidimensional Assessment of Stress Resilience: A Development and Validation Study
J Med Internet Res 2026;28:e81714
URL: https://www.jmir.org/2026/1/e81714
doi:10.2196/81714

© Lydia Genevieve Roos, Destiny Gilliland, Kelsey Julian, Reeva Misra. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Program Theory and Core Outcome Set Development for a Technology-Assisted Counseling Intervention in Dementia: Multimethods Study

Dorothee Bauernschmidt, Dr rer medic; Anja Bieber, Dr rer medic; Ronja Hubrich, MSc; Janina Wittmann, MSc; Gabriele Meyer, Dr phil

Institute of Health, Midwifery and Nursing Science, Medical Faculty, University Medicine Halle, Martin Luther University Halle-Wittenberg, Magdeburger Straße 8, Halle (Saale), Germany

Corresponding Author:
Dorothee Bauernschmidt, Dr rer medic
Institute of Health, Midwifery and Nursing Science, Medical Faculty, University Medicine Halle, Martin Luther University Halle-Wittenberg, Magdeburger Straße 8, Halle (Saale), Germany

Abstract

Background: Counseling in family dementia care aims to support caregivers in mastering challenges. The use of information and communication technologies (ICT) to administer counseling can improve accessibility. Evidence syntheses report inconsistent findings on the effectiveness of technology-assisted counseling. There is a considerable heterogeneity in outcomes assessed in clinical trials, and approaches to develop and evaluate interventions are not guided by theory in most cases.

Objective: This study aims to develop an initial program theory of a technology-assisted counseling intervention for family dementia caregivers and to create the data basis for the consensus process of a core outcome set.

Methods: We integrated the methodological strands for the development of a program theory and a core outcome set in an innovative way. A scoping review was conducted to collect data on characteristics and theoretical foundations of technology-mediated counseling interventions as well as outcomes of clinical studies. We explored the lived experience of relevant interest-holders and conducted semistructured interviews applying a phenomenological approach to data analysis. Synthesis of findings was performed by developing a logic model and formulating an initial program theory.

Results: We included 69 records reporting on 34 interventions. Designs and other study characteristics vary, and interventions are heterogeneous in terms of components and ICT used for delivering counseling. We conducted interviews with 15 family caregivers and 12 counselors. The themes *being affected*, *feeling insecure and helpless in the face of the health care system*, and *search for information and communicative exchange* illustrate the caregivers' lifeworld perception. Themes identified in counselors' interviews comprise *work attitude and standards*, *unpredictability*, *expectations*, *working conditions*, *organizational influence*, and *tools: techniques and networking*. The constitutive pattern of *having/being somebody to count on* was incorporated into the program theory. In the theory of change, we describe the way to a sustainable supportive cooperation between caregivers and counselors ensuring ongoing support throughout the caregiving process. We explicate the effects of the technology-assisted counseling intervention such as improved knowledge, attitude, and interaction, as well as stability and safety of care in the outcomes chain. The theory of action comprises the inputs, activities, and outputs of the intervention. The graphical synthesis of findings is presented in the logic model.

Conclusions: To effectively develop, implement, and evaluate technology-assisted counseling in family dementia care, a theory-led approach is essential. A carefully modeled intervention that combines technological options with in-person counseling may help to overcome disparities in access to health care and improve accessibility to counseling. A supportive working environment for counselors, in which artificial intelligence is used to reduce time spent on documentation and administrative tasks, may help mitigate the effects of the growing shortage of skilled professionals.

Trial Registration: Core Outcome Measures in Effectiveness Trials (COMET) Initiative 2884; https://www.comet-initiative.org/Studies/Details/2884

(J Med Internet Res 2026;28:e81669) doi:10.2196/81669

KEYWORDS

theory; core outcome set; logic model; counseling; family caregiver; dementia

XSL•FO
RenderX

Introduction

Counseling interventions in family dementia care aim to support caregivers in mastering challenges in caregiving. Information and communication technologies (ICT) are used to deliver counseling in an easily accessible way [1-3].

Counseling may contribute to mitigating the negative impact family dementia care can exert on caregivers, which is described in terms of burden, depression, decreased health and quality of life, as well as social isolation [4,5]. In light of the predicted increase in the number of persons with dementia [6], the need for support interventions for family caregivers is also likely to increase. Provided by professionals, counseling can be defined as the "use of an interactive helping process focusing on the needs, problems, or feelings of the patient and significant others to enhance or support coping, problem-solving, and interpersonal relationships" [7]. The use of ICT is discussed to reduce barriers to utilization by overcoming distances, enabling persons who are homebound or living in rural areas to participate in services, offering anonymous counseling, or fostering asynchronous communication beneficial for persons who are employed [3,8]. Types of ICT have expanded over time: long-established helpline services use the telephone as a widespread and undemanding technology. More recently, counseling is provided via videoconferencing software, email, or chats using mobile devices and applications, thus leading to increased demands on technological infrastructure, equipment, and digital literacy [8].

Technology-assisted counseling (also referred to as "technology-based" in previous publications) is a complex intervention [9], and the development, implementation, and evaluation of complex interventions are challenging. The effectiveness of technology-assisted counseling in dementia has not yet been proven: Evidence syntheses of technology-assisted psychosocial interventions including counseling for caregivers of persons with dementia show inconsistent findings on positive effects on outcomes such as burden, depression, or quality of life [1-3,10]. Our own meta-analyses revealed no significant effects of technology-assisted counseling on depressive symptoms, burden, and self-efficacy or mastery perceived by family caregivers of persons with dementia [11]. We found a considerable heterogeneity in outcomes and a lack of theoretical approaches guiding the development, implementation, and evaluation of the interventions [8,11].

Therefore, we conducted the ProCOS project: "Development and Evaluation of a Technology-Assisted Counseling Intervention for Family Caregivers of Persons With Dementia – Program Theory and Preparation of a Core Outcome Set" [12]. Within the 12-month project, we aimed at creating the foundation for the future consensus process of a core outcome set (COS) and at developing a program theory by combining the methodological strands of the two developmental processes in an innovative way.

A COS is "an agreed standardized collection of outcomes which should be measured and reported, as a minimum, in all trials for a specific clinical area" [13]. The use of a COS reduces heterogeneity in clinical trials and enhances comparability and thus synthesis of evidence [13]. Core outcome sets for the evaluation of health care interventions and of psychosocial community-based interventions for persons with dementia living at home predominately focus on outcomes of persons with dementia [14,15]. A set of measures has been recommended to evaluate a broad range of psychosocial interventions for persons with dementia and their family caregivers [16]. There is no COS that specifically focuses on technology-assisted counseling interventions for family dementia caregivers.

The Framework for Developing and Evaluating Complex Interventions identifies program theory as a core element of complex interventions [9]. This underlines the importance of theoretical approaches to successfully develop, implement, and evaluate complex interventions. A program theory is an "explicit theory of how an intervention is understood to contribute to its intended or observed outcomes" [17]. Following the approach introduced by Funnell and Rogers [17], we developed a "purposeful program theory" comprising the theory of change, the outcomes chain, and the theory of action. The theory of change explicates the central mechanism of how the intended changes can be achieved, and the theory of action explains how the intervention is designed to initiate the theory of change. These elements are linked by the outcomes chain comprising the immediate and intermediate outcomes and the impact of the intervention, as well as hypothesized relationships between outcomes [17]. In line with the updated UK Medical Research Council framework, we understand a program theory as a detailed textual description [18]. Logic models serve as visual representations of program theories [17,18]. Detailed logic models graphically illustrate the (assumed) causal mechanisms through which an intervention is expected to produce outcomes, as well as contextual dependencies and preconditions [18]. Developing a program theory that incorporates perspectives of diverse interest-holders and integrates theoretical and empirical knowledge at the beginning of interventional research is considered best practice [9]. Program theories and logic models are adapted and refined throughout the development, implementation, and evaluation of complex interventions to address the question, "What works in which circumstances and how?"—thus applying a theory-based approach [9].

As the methodological strands both for developing a program theory and a COS integrate knowledge obtained from literature and perspectives from different interest-holders [13,17], we brought these two processes together to develop a program theory and prepare the consensus process of a COS of a technology-assisted counseling intervention for caregivers of persons with dementia. The central instrument for this innovative approach is the logic model. As graphical representations of program theories [17], logic models have been used to synthesize data [19,20] and to visualize assumed causal relationships and mechanisms of action of complex interventions [21]. By fostering a shared understanding among interest-holders [9,17], we used the logic model for data synthesis, and we will integrate it in the future consensus process of the COS, allowing interest-holders to critically review the quality of the program theory. To our knowledge, this approach has not yet been implemented.

By drawing on previous work on the effectiveness and implementation success of technology-assisted counseling in

XSL•FO
RenderX

dementia [8,11,22], we therefore aimed at addressing gaps in knowledge. Thus, our first objective was to develop a program theory of a technology-assisted counseling intervention for family dementia caregivers. Our secondary objective was to compile lists of (potential) outcomes that have been identified by caregivers or counselors or assessed in clinical studies.

The following two sets of questions guided our research:

1. What interventions that use ICT to provide counseling for family dementia caregivers are described in literature? What are the characteristics of these interventions? What theoretical underpinnings for intervention development and implementation are explicated in the form of theoretical references, program theories, or logic models? What outcomes have been examined in clinical trials? What assessment instruments have been used?
2. How do family caregivers for persons with dementia and counselors experience counseling services? What expectations do persons seeking or providing counseling have of each other? Which outcomes should or could be achieved through counseling, and how can these outcomes be achieved? Which factors have an impact on the effectiveness of counseling? What are possible outcomes for assessing the effectiveness of counseling interventions?

Methods

The ProCOS project was registered with the Core Outcome Measures in Effectiveness Trials (COMET) Initiative [23]. The study protocol has been published [12].

Study Design

We adopted a multimethod design including a literature review and a qualitative substudy [24]. Results were graphically synthesized into a logic model and an initial program theory was formulated comprising the theory of change, the outcomes chain, and the theory of action [17].

We applied the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses, Extension for Scoping Reviews) [25], the SRQR (Standards for Reporting Qualitative Research) [26], and the COS-STAR (Core Outcome Set—Standards for Reporting) [27] to structure our report.

Scoping Review

We aimed to map the knowledge on technology-assisted counseling interventions for family caregivers of persons with dementia and followed the Joanna Briggs Institute methodological guidance for scoping reviews [28].

Eligibility Criteria

We included studies on interventions using ICT to deliver individualized counseling to caregivers of persons with dementia. Publications in the German or English language were accepted, irrespective of their design. Detailed inclusion and exclusion criteria according to the Population-Concept-Context scheme are provided in the study protocol [12].

Search Strategy and Information Sources

By updating the literature search of a previous systematic review [8,11,22], we searched the databases CINAHL, MEDLINE via PubMed, Cochrane Library, and PsycINFO (December 2023) and conducted additional forward and backward citation searching as well as a free web search (April 2024). The database-specific search strategies and the search terms of the free web search are provided in Multimedia Appendix 1.

Selection of Sources of Evidence

Titles, abstracts, and full texts were screened independently by two researchers (DB and RH) using the Rayyan web app [29]. Discrepancies in decisions were resolved by discussion.

Data Charting Process and Data Items

We extended a previously developed data extraction sheet and extracted data on study characteristics (year of publication, country of study conduct, design and methods, and number of participants) and on outcomes examined as well as assessment instruments used in included studies. We also applied criteria from the Template for Intervention Description and Replication (TIDieR) checklist [30] and from the revised Criteria for Reporting the Development and Evaluation of Complex Interventions (CReDECI 2) guideline [31] in order to collect information on objectives, components, theoretical underpinnings of counseling interventions, technology and materials used for delivering counseling, frequency and duration of sessions, and implementation issues. Data extraction was performed by one reviewer (DB). A cross-check of extracted data was conducted by another researcher (RH) for 20% of included interventions indicating accuracy and completeness of data extraction.

Qualitative Substudy

With the aim to explore the lived experience of counseling in dementia care, we conducted interviews with interest-holders in caregiving and counseling in the context of family dementia care.

Qualitative Approach and Research Paradigm

We adopted a phenomenological perspective [32] and focused on the lifeworld experiences of caregivers and counselors.

Context, Researcher Characteristics, and Reflexivity

All interviews and a preliminary analysis were conducted by the primary investigator (DB), who is experienced in phenomenological research. Results were then discussed within the research team. Team members are nursing scientists with extensive experience in nursing practice and dementia research, and a physiotherapist.

Units of Study and Sampling Strategy

Interviews were conducted with family caregivers and counselors. We used a purposive sampling strategy [33] to obtain a heterogeneous sample. Predefined criteria for recruitment of dementia caregivers were age, gender, socioeconomic status, and family relationship, as well as spatial and emotional proximity to the person receiving care. Predefined criteria for recruitment of counselors were disciplinary background, professional qualifications, duration of counseling

XSL•FO
RenderX

experience, and characteristics of employing organizations. We included persons who have received or delivered counseling via technology, in-person, or both. While our efforts to recruit a heterogeneous sample of family dementia caregivers had limited success, we succeeded more in engaging counselors with diverse characteristics.

Invitations to participate and study information were distributed via existing contacts and networks established through previous research projects and the State Competence Center for Dementia of the German federal state of Saxony-Anhalt. In addition, we contacted counseling services and self-help organizations of family caregivers throughout Germany. Recruitment was completed when no new information emerged from interviews indicating that data saturation was achieved [34,35].

Data Collection Methods and Data Processing

We performed semistructured interviews using an interview guide with open-ended questions to give interviewees room to share their experiences [36]. Questions focusing on experiences in receiving and providing counseling were asked in the course of the interview. Finally, additional data on sociodemographic characteristics, information on the care arrangement, or on the professional situation were collected. The translated interview guide is provided in Multimedia Appendix 2.

Interviews were arranged at participants' convenience to minimize time exposure and burden for participants. We planned to conduct individual interviews, but switched to group interviews at the request of some participating caregivers and counselors.

Recordings of interviews were transcribed verbatim using f4 transcription software [37]. Transcripts were checked for accuracy and pseudonymized by one researcher (RH).

Data Analysis

We conducted an interpretive phenomenological analysis applying the following modified working steps described by Diekelmann [38]:

1. The transcripts are read several times to obtain an overall understanding.
2. An interpretive summary of each interview is written.
3. Interpretive summaries are analyzed and discussed to identify emerging themes.
4. Disagreements in interpretation are resolved by returning to the text.
5. Through comparing and contrasting texts, the themes that recurred and reflected the shared practices and common meanings are identified and described.
6. As themes are compared, a constitutive pattern emerges that links the themes and is present in all interviews.
7. The themes and the constitutive pattern are described using quotations to illustrate findings.

Themes and the constitutive pattern illuminate the lifeworld perception of receiving or providing counseling in the context of family dementia care.

In addition, outcomes designated by caregivers and counselors were independently extracted from the transcripts by two researchers (DB and JW) and discussed intensively in order to include them as accessible statements [39] into the future consensus-building process.

Synthesis of Data Through Logic Model Development and Formulation of a Preliminary Program Theory

Contrary to the original plan to conduct the data synthesis into the logic model and the formulation of the program theory consecutively, we brought these two working steps together in an iterative process. We switched back and forth between writing memos and graphically synthesizing the data. Following the recommendation made by Funnell and Rogers [17], we combined the inductive and deductive approach with articulating interest-holders' mental models to develop the program theory. Mental models are diverse interest-holders' beliefs about how a program achieves its results [17]. We integrated the mental models explicated by caregivers and counselors with data derived from literature and with information on programs operating in practice. We treated data extracted from studies included in the scoping review as qualitative data [20]. Data were charted and categorized [20], and assigned to the elements of the program theory [17].

The starting point for the development of the theory of change was the situation analysis for which we applied the guiding questions formulated by Funnell and Rogers [17]. Key aspects of the situation analysis are depicted as macro-, meso-, and microcontext in the logic model.

We then explicate our assumptions of how the intended changes can be achieved by formulating the theory of change: (hypothesized) mechanisms of effective counseling, as described by interview partners or identified in literature, were incorporated into the theory of change.

Outcomes derived from clinical studies or mentioned by participants were summarized in tables and served as the foundation of the outcomes chain. The development of the outcomes chain followed the steps outlined by Funnell and Rogers [17]: (1) listing of possible outcomes, (2) clustering outcomes and assigning working labels to each cluster, (3) arranging outcomes in a chain of if-then statements, (4) identifying feedback loops, and (5) validating the outcomes chain.

By explicating how the intervention is designed to initiate the theory of change, we developed the theory of action and identified resources (inputs), activities, and outputs [17]. We incorporated intervention components and strategies that were reported in the literature or by study participants as helpful or effective in achieving the intended outcomes in the theory of action.

Patient and Public Involvement

We established a study advisory board consisting of a person with long-term experience in caring for family members with dementia and engagement in an informal support network for family caregivers, a person with extensive counseling experience, and an experienced researcher in the field of dementia care. Representing different groups of interest-holders, the members of the study advisory board critically reviewed

XSL•FO
RenderX

the instruments and approaches for data collection and were involved in the discussion of results and conclusions for the design of a technology-assisted counseling intervention. Three virtual meetings were held in addition to feedback provided via email over the course of the 12-month project.

Techniques to Enhance Trustworthiness

We applied the strategies of expert consultation and peer debriefing by consulting the members of the study advisory board throughout the research process.

Ethical Considerations

The ethics committee of the Medical Faculty of the Martin Luther University Halle-Wittenberg approved the ProCOS study (no. 2023 - 093). Persons interested in participating received written information on procedures prior to the interview. The time and place of the interviews were determined based on the participants' preferences. We obtained written informed consent from the participants who were informed that the consent to participate can be withdrawn at any time. We maintained the security of data by storing data protected from access by persons who are not involved in the project. Audio recordings of interviews were pseudonymized during the transcription process. Study participants received a small gift (equivalent to US $5), but no financial compensation.

Results

Scoping Review

Selection of Sources of Evidence

The updated database search yielded 3995 records. An additional 1965 records were identified through complementary search strategies. After the removal of 1448 duplicates of reports identified via database search and 332 duplicates of reports identified via other methods, 4180 titles and abstracts were screened. A total of 141 full texts were screened for eligibility. Of these, 17 records reporting on 7 interventions were included. These new studies were combined with 52 records on 27 interventions from the previous review, resulting in 69 records reporting on 34 interventions that were included in the scoping review (Figure 1).

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses, Extension for Scoping Reviews) flow diagram for updated reviews [40].

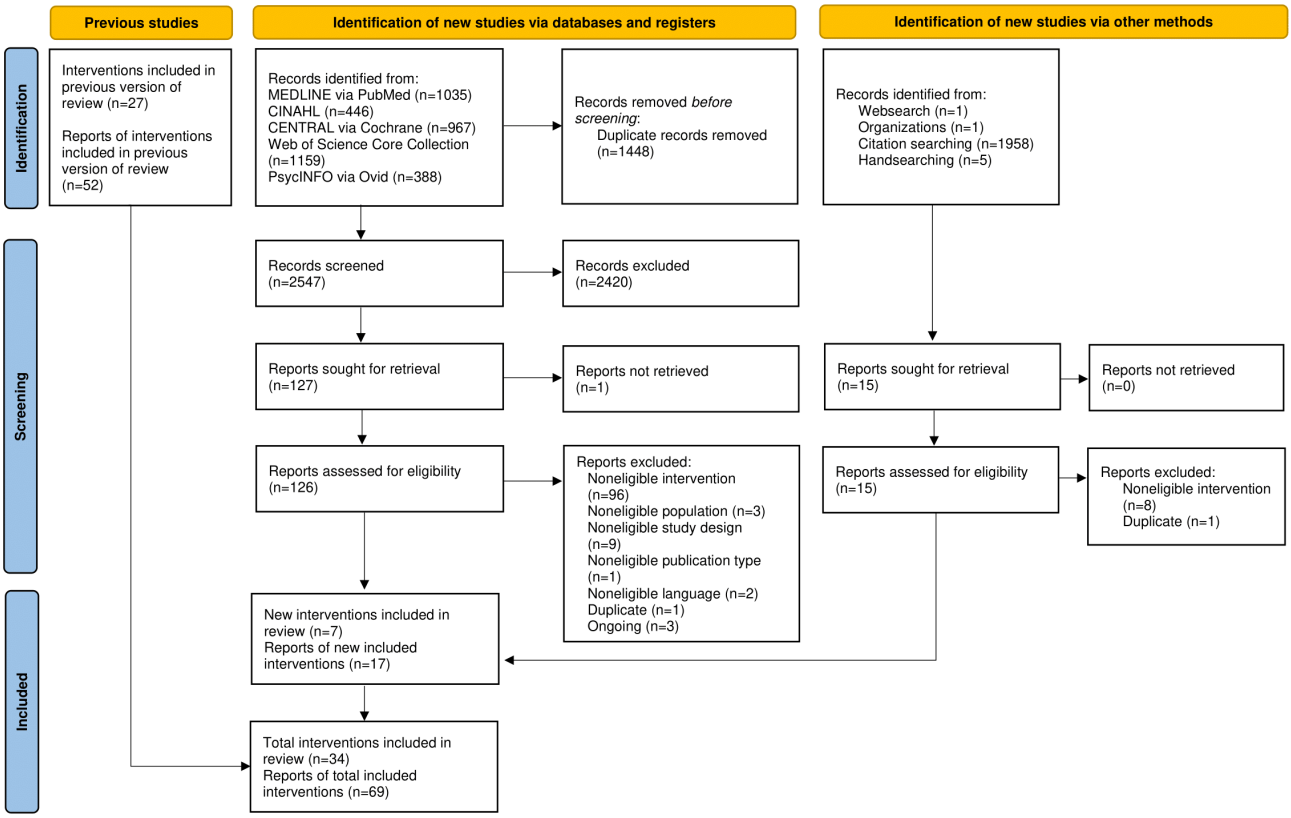

Characteristics of Sources of Evidence

The studies were published between 1993 and 2024 (last search April 2024) and were conducted in 10 countries (Australia, Canada, Germany, Israel, Italy, Japan, the Netherlands, Sweden, the United Kingdom, and the United States).

As we included studies irrespective of their design, topics, and methodological approaches varied. The effectiveness of 7 interventions on outcomes such as depressive symptoms, burden, and self-efficacy (refer to Table 1) has been evaluated by randomized controlled trials (RCTs) [41-63], including a small-scale RCT in the context of a pilot and feasibility study [64,65]. Nonrandomized trials [66-68] and a pre-post-intervention trial [69] assessed the effects on caregivers' ability in disease management, burden, or mental health; quantitative descriptive approaches were conducted to examine sociodemographic characteristics of service users, reasons for use and topics discussed, advice provided, and satisfaction with services, as well as ways of access to

XSL•FO
RenderX

counseling [70-82]. Qualitative approaches were used to explore expectations and experiences of caregivers with technology-assisted counseling services [59,65,70,83-92]. Case studies [93-95] examined individual counseling processes reflecting on the complexity of delivering counseling via ICT, as well as skills and knowledge needed to provide effective support. We included two process evaluation studies focusing on implementation issues when linking eHealth interventions to existing support [96] and combining a cognitive rehabilitation program with education and counseling [97]. Mixed methods approaches were applied to achieve an in-depth understanding of how the intervention works, to assess usability, and to describe and explain (non-)usage of services [56,57,98-103]. Detailed information on included references is presented in Multimedia Appendix 3 [8,30,31,41-109].

XSL•FO
RenderX

Table . Outcomes identified in the literature.

Outcomes and assessment instruments	Intervention[a]	References
Outcomes for caregivers		
Depression		
Zung Self-Rating Depression Scale	Coyne[a] comparator or experimental	[42]
Center for Epidemiology Studies Depression Scale (CES-D)	FITT-C[b]	[44-50]
	RCTM[c]	[54-60,64,65]
Center for Epidemiologic Studies Depression Scale (CES-D-10)	CuidaTEXT	[98,104-106]
Center for Epidemiologic Studies Depression Scale–Revised (CESD-R)	TeleFAMILIES[d]	[69]
Geriatric Depression Scale (GDS)	FITT-D[e]	[43]
Mood Assessment Scale (MAD)	RCTM[c]	[54-60]
Grief		
Caregiver Grief Scale (CGS)	RCTM[c]	[64,65]
Guilt		
Caregiver Guilt Questionnaire (CGQ)	RCTM[c]	[64,65]
Anxiety		
Geriatric Anxiety Inventory (GAI)	RCTM[c]	[64,65]
Burden or caregiver distress		
Zarit Burden Interview (ZBI)	Coyne[a] comparator or experimental	[42]
	FITT-C[b]	[44-50]
	FITT-D[e]	[43]
	TeleFAMILIES[d]	[69]
Zarit Burden Interview (ZBI-6)	CuidaTEXT	[98,104-106]
Burden Scale for Family Caregivers (BSFC)	ICSS[f]	[90-92,99-101]
Caregiver Burden Inventory (CBI)	Natale[a]	[66]
	De Cola[a]	[82]
Subjective stress		
Not specified (7-item measure of care-related strain; 1-item measure assessing the caregiver's difficulty with the relative's mental or emotional state); Adapted Zarit Burden Interview (ZBI)	RCTM[c]	[54-60]
Perceived stress		
Perceived Stress Scale (PSS)	RCTM[c]	[64,65]
Caregiver strain		
Modified Caregiver Strain Index (CSI)	CuidaTEXT	[98,104-106]
Reaction to care-receiver behavior		
Revised Memory and Behavior Problem Checklist (RMBPC)	FITT-C[b]	[44-50]
	FITT-D[e]	[43]
	TeleFAMILIES[d]	[69]

XSL•FO
RenderX

Outcomes and assessment instruments	Intervention[a]	References
Emotional impact of neuropsychiatric symptoms		
Neuropsychiatric Inventory (NPI) Burden subscale	Dementelcoach	[61-63,67,68,96]
Upset		
Caregiver Behavioral Occurrence and Upset Scale (modeled on the Agitated Behavior in Dementia Scale)	Laver[a]	[51-53]
Behavioral symptom severity–distress		
Neuropsychiatric Inventory–Questionnaire (NPI-Q) Distress	CuidaTEXT	[98,104-106]
Coping		
Coping Orientation to Problems Experienced Inventory (COPE-28)	CuidaTEXT	[98,104-106]
Affect		
Scale of Positive and Negative Experience (SPANE)	CuidaTEXT	[98,104-106]
Self-perception of negative and positive aspects of caregiving		
Carers of Older People in Europe (COPE) Index	InformCare	[102,103,107]
Positive aspects of caregiving		
Positive Aspects of Caregiving (PAC) Scale	FITT-C[b]	[44-50]
	CuidaTEXT	[98,104-106]
Self-efficacy		
PROMIS Self-Efficacy for Emotions (derived from the Self-Efficacy for Managing Chronic Disease (SEMCD) questionnaire)	Care Consultation or Care Consultation Plus	[41]
Self-Efficacy Questionnaire (SEQ)	FITT-C[b]	[44-50]
	FITT-D[e]	[43]
	RCTM[c]	[54-60]
Mastery		
Caregiving Mastery Index (CMI)	Laver[a]	[51-53]
Sense of capability		
Short Sense of Competence Scale (SSCQ)	Dementelcoach	[61-63,67,68,96]
	RCTM[c]	[54-60]
Caregiving competence		
Preparedness for Caregiving Scale (PCS)	CuidaTEXT	[98,104-106]
Psychological well-being		
World Health Organization Well-Being Index (WHO-5)	InformCare	[102,103,107]
Happiness		
TOPICS-MDS[g] item	Dementelcoach	[61-63,67,68,96]
Quality of life		
Euro Quality of Life Visual Analog Scale (EQ-5D VAS)	FITT-C[b]	[44-50]

XSL•FO
RenderX

Outcomes and assessment instruments	Intervention[a]	References
TOPICS-MDS[g] item	Dementelcoach	[61-63,67,68,96]
Health		
PROMIS Global Health: Global Physical Health (GPH) and Global Mental Health (GMH)	Care Consultation or Care Consultation Plus	[41]
General Health Questionnaire (GHQ-28)	Dementelcoach	[61-63,67,68,96]
SF 36 General Health	FITT-D[e]	[43]
Not specified (1-item question)	CuidaTEXT	[98,104-106]
Unmet needs		
Measure of Unmet Needs (UN)	CuidaTEXT	[98,104-106]
Perceived social support or support for caring		
Multidimensional Scale of Perceived Social Support (MSPSS)	FITT-D[e]	[43]
	InformCare	[102,103,107]
Interpersonal Support Evaluation List (ISEL-12)	CuidaTEXT	[98,104-106]
Support for Caring subscale of the Adult Carer Quality of Life Questionnaire (AC-QoL)	RCTM[c]	[64,65]
Family functioning		
Family Assessment Device (FAD)	FITT-C[b]	[44-50]
	FITT-D[e]	[43]
Knowledge		
Alzheimer's Disease Knowledge Test	FITT-D[e]	[43]
Epidemiology/Etiology Disease Scale (EEDS)	CuidaTEXT	[98,104-106]
Perceived change		
Perceived Change Scale (PCS)	Laver[a]	[51-53]
Secondary role strains		
Not specified (2 single-item ratings of the caregiver's and the relative's adjustment to residential long-term care placement)	RCTM[c]	[54-60]
Residential care stress		
Not specified (6-item measure for perceptions of staff communication with family, 5-item measure for staff support for family, 10-item measure assessing 5 positive and 5 negative types of caregiver interactions with their relative, staff, other family; item caregivers' upset to see their relative in a residential care setting); Family Involvement Interview	RCTM[c]	[54-60]
Resource use		
Number of community or health services used	Coyne[a] comparator or experimental	[42]
	FITT-C[b]	[44-50]
	FITT-D[e]	[43]
Outcomes for persons with dementia		
Depression		

XSL•FO
RenderX

Outcomes and assessment instruments	Intervention[a]	References
Geriatric Depression Scale (GDS)	De Cola[a]	[82]
Global cognitive state		
Mini-Mental State Examination (MMSE)	De Cola[a]	[82]
Cognitive impairment		
Bedford Alzheimer Nursing Severity Scale (BANSS)	De Cola[a]	[82]
Functional dependency		
Activities of Daily Living (ADL) and Instrumental Activities of Daily Living Scale (IADL)	De Cola[a]	[82]
Instrumental activities of daily living		
Functional Activities Questionnaire (FAQ)	CuidaTEXT	[98,104-106]
Functionality		
Caregiver Assessment of Function and Upset (CAFU)	Laver[a]	[51-53]
Neuropsychiatric symptoms or behavioral symptom severity		
Neuropsychiatric Inventory (NPI) Total symptoms	Dementelcoach	[61-63,67,68,96]
Neuropsychiatric Inventory-Questionnaire (NPI-Q) Severity	Dementelcoach	[61-63,67,68,96]
	CuidaTEXT	[98,104-106]
Psychiatric symptoms		
Brief Psychiatric Rating Scale (BPRS)	De Cola[a]	[82]
Number of behaviors such as verbal aggression, refusing care, restlessness, anxiety, waking overnight, and repetitive questioning		
Caregiver Behavioral Occurrence and Upset Scale (modeled on the Agitated Behavior in Dementia Scale)	Laver[a]	[51-53]
Resource use		
Number of community or health services used	Coyne[a] comparator or experimental	[42]
	FITT-C[b]	[44-50]

[a]When no name is reported, the name of the first author was assigned to the intervention.

[b]FITT-C: Family Intervention: Telephone Tracking–Caregiver.

[c]RCTM: Residential Care Transition Module.

[d]TeleFAMILIES: Telehealth-Administered Families Access to Memory Impairment and Loss Information, Engagement, and Supports.

[e]FITT-D: Family Intervention: Telephone Tracking–Dementia.

[f]ICSS: Internet-Based Caregiver Support Service.

[g]TOPICS-MDS: The Older Persons and Informal Caregivers Survey Minimum DataSet.

Results of Individual Sources of Evidence

We summarized the included interventions into groups that were formed based on the ICT used and the additional components of the interventions. We found 17 interventions that delivered counseling via telephone, email, or videoconferencing [41-50,66,70-78,83-87,93,94,108]. Three interventions provided counseling via videoconferencing [69,79,88]. One intervention each was assigned to the groups "counseling via email" [95], "counseling via SMS text messaging" [98,104-106], and "counseling via an interactive mobile app combined with additional features" [89]. We included 4 web-based psychosocial interventions that combined information, communication, and counseling [80,81,90-92,99-103,107]. Four videoconference- or telephone-based interventions combined counseling services with telemonitoring or psychoeducation [51-60,64,65,82,109]. Three technology-assisted interventions offered counseling as

XSL•FO
RenderX

part of a comprehensive program with non-technology-assisted components [61-63,67,68,85-87,96,97].

The interventions aim to support family caregivers and persons with dementia, to provide information and education, and to enhance coping with and managing of the caregiving process. Theoretical foundations of interventions such as psychological or psychosocial concepts are referred to for 10 interventions [43-65,67,68,70,93-98,104-106,108]. The health service usage behavior has been theorized for an internet-based information support and personalized email intervention [90-92,99-101]. A program theory is mentioned for 1 intervention [83].

In the included reports, various components of the interventions are described, which are combined in different ways: intervention manuals or guidelines are made available for counselors. In some cases, special training is provided on dementia, counseling strategies, or the use of ICT in counseling. Supervision and coaching are offered to persons delivering counseling. There are various modes of documenting the content of the counseling sessions (eg, protocols and log sheets) and making it available to counselees (eg, letters and script proposals). In addition, information and educational material is offered as brochures, manuals, databases, websites, and videotapes. In some cases, the necessary technical equipment or assistance with its use is provided. Persons delivering counseling have qualifications in nursing, social work, social science, psychology, mental health, geriatrics, or occupational therapy. Additional experiences in gerontological, psychogeriatric, or dementia care, as well as experiences as family caregivers, are described.

Data on implementation issues is available to varying degrees. Implementation strategies were described for 2 interventions [55,96]. The reported modifications of interventions are mainly related to adaptations to the needs of new target groups and to pandemic-related restrictions. Barriers to implementation are technical issues (eg, lack of hardware, limited software functionality or usability, limited internet access, and suboptimal infrastructure), lack of digital skills (of persons delivering and receiving counseling), security issues, and aspects resulting from the type of technology used in each case (eg, loss of context, lack of prompts from the surroundings, and not being able to follow up). Facilitating factors are special trainings (eg, technical support and conversation strategies), ongoing support (eg, regular team meetings and supervision), features of interventions (eg, customized to sociocultural preferences of target groups and undemanding technology), and commitment of management of implementing institutions. External conditions of implementation are described as established organizations in which intervention programs are embedded, collaborations with interest-holders in dementia care, and the SARS-CoV-2 pandemic.

A detailed description of the included intervention programs is provided in Multimedia Appendix 3. Intervention components that have been incorporated into the program theory and that are represented in the logic model are referred to in the section Logic Model and Program Theory.

Outcomes of caregivers or persons with dementia examined and the assessment instruments used in the included studies are listed in Table 1.

Qualitative Substudy

Sample Characteristics

We conducted 4 individual and 3 group interviews lasting from 1 to 3 hours (total duration: 12 hours 41 minutes) with 15 family caregivers (14 women and 1 man). One participant cared for a person with Parkinson disease; all other caregivers cared for a person with dementia. One family had a migration background; the caregiving arrangement no longer existed in 3 cases due to the death of the person receiving care.

In addition, 8 individual and 2 group interviews lasting from 50 to 100 minutes (total duration: 12 hours 14 minutes) with 12 counselors (11 women and 1 man) were conducted. Participants had been working as counselors for an average of 9 years and had qualifications in nursing, health and nursing science, social work, psychology, and administration. The providing institutions of participating counselors varied. Details on the characteristics of participants are provided in Table 2.

XSL•FO
RenderX

Table . Sociodemographic information on interview partners.

Characteristics	Values
Caregivers (n=15)	
Age in years, mean (range)	70 (44-83)
Gender, n	
Women	14
Man	1
Care-receiving person, n	
Husband or spouse	10
Mother	3
Father	1
Brother	1
Caregiving arrangement, n	
Home-based care	12
Institutionalized care	3
Years of caregiving, mean (range)	6 (1-20)
Counselors (n=12)	
Age in years, mean (range)	51 (33-61)
Gender, n	
Women	11
Man	1
Years of counseling, mean (range)	9 (1-30)
Providing organization, n	
Municipality	4
Welfare organization	2
Self-help organization	2
Research institution	1
University hospital	1
Registered association	1
Self-employed counselor	1

Themes Identified in Interviews With Caregivers

We identified the following 3 themes in the caregivers' statements: being affected, feeling insecure and helpless in the face of the health care system, and search for information and communicative exchange.

Being Affected

> *...and then I cried a lot and had really bad thoughts.* [Caregiver 15]

Participants' descriptions of how they are affected by the demands of the caregiving situation made up a large part of the interviews. The caregivers report conflicts with the persons receiving care encompassing harassment and insults, as well as verbal and physical attacks: "Back in November, my husband pushed me down the stairs" [Caregiver 15]. They also describe their grief over the loss of a person close to them due to the changes evoked by dementia, as well as the loss of life plans and perspectives. In addition, the demands of caregiving pose a burden on caregivers: "...that you just can't do it anymore" [Caregiver 15].

Feeling Insecure and Helpless in the Face of the Health Care System

> *And nobody tells you all this.* [Caregiver 15]

The interviewees state a lack of information and a lack of support from persons or institutions involved in health care such as primary care physicians or health insurances:

> *There's the 24-hour care, every single day. Yeah, and then if you spend the whole morning on the phone just trying to get a simple answer, well, that's just how it is for me.* [Caregiver 15]

High financial expenses often limit the use of support services.

XSL•FO
RenderX

Search for Information and Communicative Exchange

> *That made a huge difference for me, that I stopped thinking that I was somehow (...) weird.* [Caregivers 3-5]

Caregivers report beneficial experiences with empathetic counseling focused on problem-solving:

> *She [counselor] asks, "So, how are you doing?" Finally, someone actually asks how I'm handling all this. (...) And then she says, "Okay, so here's what you can do..." – and that's the kind of support that really does you good.* [Caregivers 12-14]

Participants are not always aware of the differing objectives of services such as counseling, support, or self-help groups; however, they emphasize the importance of continuity in support services:

> *It has to be done continuously. (...) Even though I was really impressed by my first two meetings here – and they were incredibly helpful – if I had stopped then, I wouldn't have this feeling now.* [Caregivers 3-5]

Interviewees indicate a preference for face-to-face meetings and express a reluctance to use ICT for counseling services: "You have to deal with all that tech stuff, and honestly, no – it's just too much for me" [Caregiver 15]. Participants, however, make use of familiar technologies to maintain ongoing contact with their counselors: "I have her (counselor) on (messenging service), and she texts too – like she'll wish me a nice weekend and ask if everything's okay or if something's going on" [Caregiver 15].

Themes Identified in Interviews With Counselors

We identified the following themes in counselors' interviews: work attitude and standards, unpredictability, expectations, working conditions, organizational influence, and tools: techniques and networking.

Work Attitude and Standards

> *Good counseling is when you understand the person in their (...) lifeworld (...) when they feel that their issues are taken seriously.* [Counselor 7]

The interviewees show a high level of commitment, appreciation, and empathy toward those seeking counseling. They describe how their work has fostered both personal and professional growth, and they highlight specific skills required for different counseling formats, such as anonymous helpline services. To further ensure the quality of counseling services, training and support offered to counselors are considered essential:

> *So, as the head of the Support Center for Family Caregivers, I really keep an eye on that. What do my colleagues need? And they get it. It's really important to me that the quality here is good.* [Counselor 6]

To establish a common understanding of counseling, conceptual frameworks or manuals are developed to document the shared values: "I'm gonna write down what the core things in our counseling are – what's really important to us" [Counselors 3 and 4]. Efforts are being made to reach out to family caregivers, for instance, by maintaining a presence in public spaces: "I've had this idea for a while – that we should have a place right in the middle of the city, like in a department store or a shopping center, somewhere people actually go" [Counselor 6]. This seems necessary, since family caregivers are often unable to initiate contact, according to the counselors:

> *(Caregiver) says, "Yeah, your address has been sitting here for two years before I even call." Happens all the time. All the time.* [Counselor 6]

Unpredictability

> *And that they [caregivers] come with such a mountain of problems.* [Counselor 12]

Study participants indicate that the concerns and the state of mind of the caregivers are not known before the counseling and cover a wide spectrum—from "calm" to "nervous breakdown." In some cases, counselees are overwhelmed and unable to speak:

> *Sometimes they just sit there and start to cry. Then there's something to drink, a tissue, and maybe even a piece of chocolate.* [Counselor 6]

Expectations

> *I wish they [caregivers] would open up so that they could really be helped properly.* [Counselor 10]

We found that counselors have various expectations of those seeking counseling. While some of the participants stated that they had no expectations at all, others expected caregivers' openness to share their concerns and experiences with the counselors. In some cases, counselors expect the caregivers to be willing to implement counselors' advice.

Working Conditions

> *Our doors are always open, and a colleague can come and say, "I have to tell you this."* [Counselor 6]

The description of the working conditions is a frequent theme in the interviews. The study participants highlight the importance of mutual support from colleagues as described in the headline quote. An important aspect is whether the participant is working in a team or as a "lone fighter":

> *And because there's no real exchange with others, you have the chance to think about things and maybe do them differently – but you can never really share it with anyone.* [Counselor 7]

Teams in which different professional groups are represented and whose members support each other are considered as an essential element in coping with the demands of counseling work:

> *We've got people from all sorts of backgrounds – social work, sociology, (...) nursing background (...) psychologist, (...) gerontology and even theology. (...) we're a really diverse team and can bring in all kinds of different perspectives.* [Counselors 3 and 4]

The theme also covers personnel and technical equipment, which varies greatly among participants:

> *I'd like to be able to scan some documents and bring them here to save people a trip. Maybe use a small*

XSL•FO
RenderX

> *mobile printer or something like that (...) but there's just no way.* [Counselor 9]

In addition, the opportunity to independently organize and design their everyday working life is considered a crucial factor of the working conditions.

Organizational Influence

> *Then they [superiors] would clearly say, "That's not your job!"* [Counselor 7]

The providing organization has a significant influence on counselors' performance in the sense of a spillover effect. Lack of agreement between managers and counselors on the purpose and the design of counseling services has a negative impact on its implementation.

Tools: Techniques and Networking

> *To pave the way for people.* [Counselor 9]

Study participants describe techniques they are using in their professional lives. These include conversation management, individual practices to ease the situation for counselees (eg, offering something to drink and asking questions that facilitate the dialogue), as well as assessment tools, documentation aids, and information materials. ICT is used to enhance access to counseling services:

> *If people live further away, that's totally normal. Or if they're working, they'll say, "Okay, I can chat with you for fifteen minutes on (videoconferencing software)." We're happy to do that.* [Counselor 6]

Professional networks are used to help caregivers with problems that fall outside the counselors' scope of responsibility (eg, medical and legal issues). Cooperation (rather than competition) between different service providers and the combination of various support services is seen as key to providing effective support:

> *Well, just counseling on its own or just training doesn't really work. It has to be a combination of different things – then it can work.* [Counselor 8]

Constitutive Pattern

The common ground in the lifeworld perceptions of the two interest-holder groups forms the constitutive pattern of *somebody to count on*. While caregivers express the need for *having somebody to count on*, who guides them empathically through the challenges of family dementia care, counselors aim at *being somebody to count on* by competently assisting caregivers to master those challenges.

Outcomes Identified by Interview Partners

Caregivers and counselors described effects, which could be used to measure effectiveness of counseling. The potential outcomes are listed in Table 3.

XSL•FO
RenderX

Table . Outcomes identified by interview partners (caregivers or counselors).

Outcomes	Identifying interview partner
Level of knowledge	Ca[a]; Co[b]
Knowledge about the disease and its (potential) course	Ca; Co
Knowledge about how to deal with changed behavior of the person with dementia	Ca; Co
Knowledge about what to do to improve well-being and safety of the person with dementia	Ca; Co
Knowledge about available support services	Ca; Co
Knowledge about legal regulations and financial support options	Co
Level of caregiver burden	Co
Level of caregiver distress	Co
Level of emotional strain	Co
Level of emotional overload	Co
Stability of caregiver perceived burden (level is maintained despite increasing care needs)	Co
Sense of relief	Ca; Co
Sense of relief through empathetic understanding	Ca; Co
Sense of relief from having a counselor to fall back on (reliability, accountability of counseling)	Ca; Co
Sense of relief through active support (eg, support in filling out application forms)	Ca; Co
Sense of order (plan in mind about what to do next or how to get along with caregiving)	Ca; Co
Caregiver depressive symptoms	Ca
Stability of caregiver perceived depressive symptoms (level is maintained despite increasing care needs)	Co
Caregiver guilt	Co
Feeling of helplessness	Ca
Quality of life (caregiver)	Co
Quality of life (person with dementia)	Co
Caregiver well-being	Co
Caregiver (physical or psychological) health	Co
Caregiver satisfaction	Co
Congruousness of caregiving arrangement with caregiver's preferences and wishes	Co
Caregiver attitude toward caregiving	Co
Sense of confidence	Co
Sense of confidence in decision-making	Co
Sense of confidence in managing caregiving challenges	Co
Caregiver mastery	Co
Empowerment (eg, in finding support by oneself)	Co
Feeling well advised	Co
Problem-solving ability	Ca; Co
Realization of counselors' suggestions (eg, changes in the home environment, use of support services, creation or use of personal space, and hobbies)	Co

XSL•FO
RenderX

Outcomes	Identifying interview partner
Balance between caregiving and self-care	Co
Recognition of personal limits in caregiving capacity	Co
Safety of care	Co
Stability of caregiving arrangement over time	Co
Number of support services used	Co
Duration of caregiving arrangement	Co
Time to transition to long-term care facility	Co
Extent and stability of caregiver's support network	Co
Number of conflicts with care-receiving person	Co
Cost of care	Co

[a]Ca: caregivers.

[b]Co: counselors.

Taken together, the statements of the two participant groups on effective counseling include the following characteristics: an empathetic, appreciative, and compassionate attitude; knowledge transfer; provision of materials and help with bureaucracy; providing a view from the "outside": classifying, weighting, and sorting caregiving issues; recommendations for action (eg, dealing with changes in behavior); support in decision-making and in defining boundaries; and sustainability and continuity. Comments and suggestions of participating caregivers and counselors that have been incorporated into the program theory and that are represented in the logic model are referred to in the section Logic Model and Program Theory.

Logic Model and Program Theory

The logic model of the technology-assisted counseling intervention is provided in Figure 2.

Figure 2. Logic model of the technology-assisted counseling intervention. AI: artificial intelligence; ICT: information and communication technology. *The theory of change, the outcomes chain, and the outcomes listed in Tables 1 and 3 will be subject to a future consensus-building process.

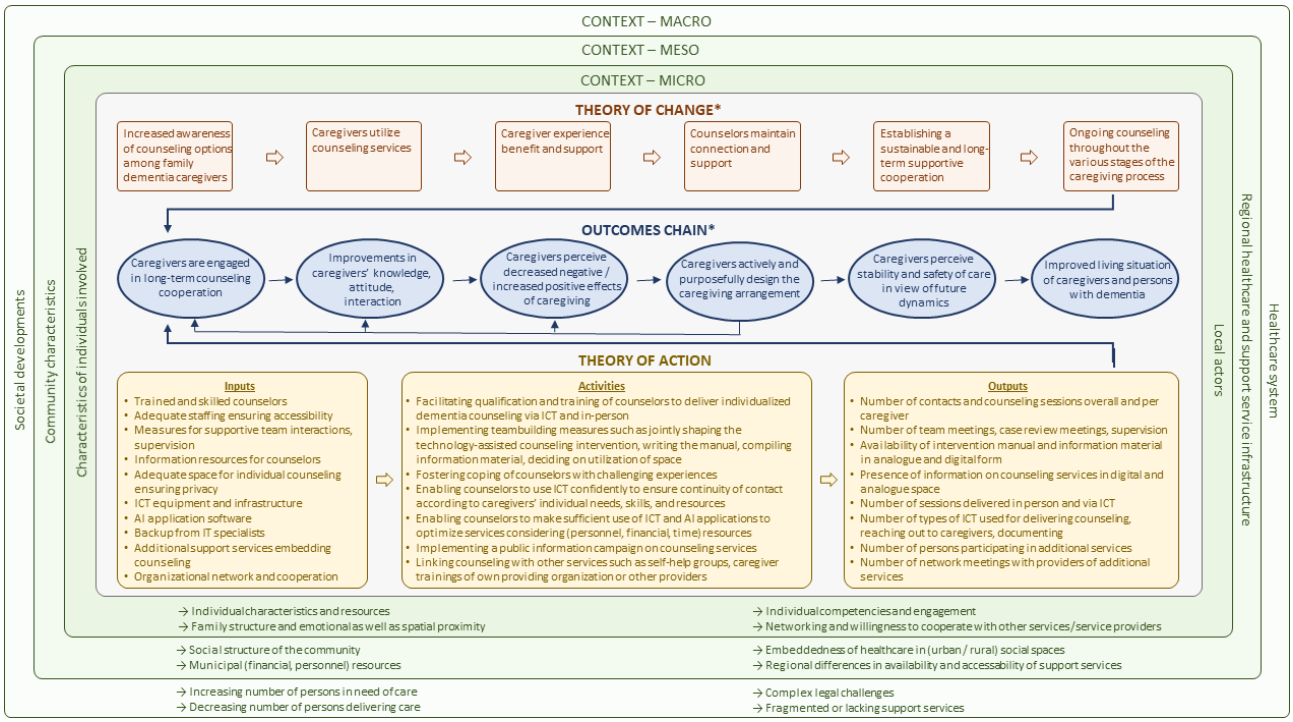

In the following section, we outline the program theory. The detailed version can be obtained from the authors on request.

The contextual factors of family dementia care in Germany on the macro-, meso-, and microlevels are illustrated in green in the logic model: Ongoing demographic developments lead to an imbalance between care needs and families' capacity to provide care. Challenges faced by individuals assuming caregiving responsibilities can be exacerbated by fragmented health care services and complex legal regulations. The unequal distribution of support services across communities with varying resources can result in inadequate support for family caregivers of persons with dementia.

The problem to be addressed is the inadequate use of counseling services by family caregivers of persons with dementia. Study

XSL•FO
RenderX

participants indicated that caregivers are rarely made aware of the counseling services available. Moreover, caregivers often do not actively seek contact with services because they feel overwhelmed by caregiving tasks, experience feelings of shame, or hesitate due to the perceived effort involved.

In our theory of change, we outline how a technology-assisted counseling intervention may help to address this issue: We hypothesize that an increased awareness of counseling services among family dementia caregivers leads to an initial use associated with a beneficial and supportive experience.

As both caregivers and counselors emphasize the importance of continuous counseling and support, we propose that counselors adopt unobtrusive strategies to maintain engagement with caregivers. Using diverse types of ICT that caregivers are familiar with facilitates the ongoing contact with counselors. This approach fosters a sustainable, long-term supportive cooperation that enables consistent counseling throughout all stages of the caregiving process.

In the outcomes chain, we describe the immediate and intermediate outcomes, which were derived by clustering outcomes identified in the literature and those mentioned by study participants (refer to Tables 1 and 3). We assume that long-term counseling cooperation leads to improved caregiver knowledge about caregiving aspects (eg, the (potential) course of dementia, how to deal with changed behavior of the person with dementia, and available support services), to an enhanced attitude toward caregiving, and to improved interaction with the care-receiving person. This entails an altered impact of caregiving, in that the caregivers perceive decreased negative effects such as burden, distress, feeling of helplessness, and increased positive effects such as sense of confidence, empowerment, and satisfaction. Caregivers are thereby enabled to actively and purposefully shape the caregiving arrangement, for example, by using support services and to achieve congruousness of the caregiving arrangement with their own preferences and values. Caregivers perceive stability of the caregiving arrangement and safety of care despite unpredictable developments and dynamics and are aware that they may approach counselors any time needed. This leads to an improved living situation for both the caregivers and the persons with dementia. As indicated by the arrows in the logic model, the levels of the outcomes chain can occur in a recursive process.

Within the theory of action, we specify the inputs and actions through which the theory of change can be initiated [17]. The inputs specified in the theory of action are derived from the statements of study participants presented above and from intervention components described in the literature (refer to Multimedia Appendix 3). As counselors emphasized the importance of adequate working conditions, we incorporated aspects such as sufficient staffing, measures to promote a supportive working environment, and access to informational resources, including libraries and databases. Appropriate technological equipment and infrastructure, along with IT support, enable the reliable delivery of technology-assisted counseling. Furthermore, networks for cooperation with complementary services—either within the organization or in collaboration with external providers—are crucial for providing effective support. Program activities are directed at enabling counselors to provide technology-assisted counseling to family dementia caregivers competently and confidently. The activities depicted in the logic model further promote supportive working relationships by fostering mutual support within the team and articulating shared values in their work, for example, through manuals or counseling concepts. The implementation of a public information campaign may help to raise awareness of counseling services and facilitate outreach to family caregivers. In addition, diverse services are integrated to ensure timely, individualized support. Considering limited personnel, financial, and time resources, counselors are empowered to use ICT and artificial intelligence (AI) applications sufficiently to optimize the efficiency of services. Outputs represent the tangible, measurable products of activities [17]. We included the availability of the intervention manual and the information material, as well as the number of persons using counseling or other services and the number of events such as meetings to enhance cooperation within the team, within the organization, and with other organizations, as well as counts of types of ICT.

Discussion

Principal Findings

In this paper, we present the results of a scoping review and a qualitative study synthesized into a logic model and a program theory of a technology-assisted counseling intervention for family caregivers of persons with dementia. Information obtained from literature, empirical findings from the qualitative substudy, and theoretical approaches have been merged to guide the future COS consensus process and the modeling of the intervention.

Theoretical Approaches and Implications of Evolving Technological Modalities Informing the Program Theory

Presumed mechanisms of the technology-assisted counseling intervention were informed by theoretical approaches underpinning the interventions identified in the scoping review: The model of determinants of subjective burden of caregivers of persons with dementia forms the theoretical framework of the Dementelcoach intervention [68]. Aspects such as caregivers' personal characteristics, material and social circumstances, and the support they receive are integrated into the program theory. In addition, the concept map of the Internet-Based Caregiver Support Service (ICSS) described by Chiu and Eysenbach [92] to conceptualize usage behavior of family caregivers by integrating 3 theoretical approaches (Anderson's model of health service utilization, Venkatesh's theory of technology acceptance, and Chatman's and Wilson's information behavior theories) was used to shape our theoretical understanding. Factors such as dynamic and individual caregiver needs, perceived efforts of ICT options, and preferences in using ICT [92] have been taken into account.

Further findings of the scoping review illustrate how technologies used for delivering counseling have evolved over time from technology-assisted counseling provided exclusively via telephone to counseling via chats or SMS text messaging.

XSL•FO
RenderX

In addition, multiple technologies are used in some interventions for establishing contact or delivering counseling. This diversification has led to a partial overlap in the categories formulated to group interventions in this paper.

The appropriateness of technological options for administering counseling varies with respect to availability, flexibility, and requirements for technological equipment and skills [3,8]. In addition, the suitability of the various technological approaches for discussing personal issues in depth or addressing specific requests directly is perceived differently [3,8]. Therefore, we will integrate multiple technological ways to access counseling in the development of the intervention to accommodate caregivers' individual needs and resources and to enhance technology acceptance.

The combination of various ICT systems poses higher demands on counselors' technological literacy and counseling competencies. To effectively switch between various modes of delivery, counselors require specific qualifications and skills, for example, to compensate for the lack of visual clues during telephone counseling, to facilitate openness in conversations via videoconferencing software, or to address counselees' concerns in written asynchronous communication such as email or chat [8]. To enable counselors to use ICT confidently, inputs of the intervention aim at promoting an effective working environment by including adequate technological equipment and infrastructure as well as available support by IT specialists for training counselors and troubleshooting.

Overcoming Utilization Barriers Through Integrated Access Modalities and Targeted Outreach

Caregivers and counselors participating in the qualitative study largely expressed a preference for in-person counseling. The reluctance to use ICT that we observed despite the shift toward technology-assisted services during the SARS-CoV-2 pandemic [110] could have been partly influenced by the older age of the participating caregivers. Nevertheless, Gonzalez-Fraile et al [3] state in their systematic review that remotely delivered training and support interventions appeared to be less acceptable than control interventions, as assessed by attrition rates. Therefore, we integrated face-to-face services offered at counseling centers or in caregivers' home environments into the program theory of the technology-assisted counseling intervention.

We hypothesize that the combination of diverse technologies in addition to in-person counseling contributes to overcoming barriers to utilization of counseling services. The different characteristics of technological options allow for the targeting of distinct user groups. Findings of the qualitative study indicate that SMS text messaging is rather highly accepted among middle-aged or older persons and can be used to maintain contact and emotional involvement by sharing photos or invitations to events. Results from the literature also show that SMS text messaging is suitable for addressing disparities in access to caregiving support for persons belonging to minorities [98,106]. The needs of ethnic groups may also be met by counseling via email, offering an alternative service model [99,100]. Participants of the qualitative substudy stated that nonnative speaking persons may benefit from conversations via email as the asynchronous communication gives them more time to consider counselors' suggestions.

As outlined above, caregivers described in detail the consequences arising from the caregiving responsibility summarized in the theme "being affected." Feeling overwhelmed and stressed can prevent caregivers from actively seeking information and support [111]. In addition, the lack of awareness as well as the regional lack of availability contributes to the underuse of support services such as counseling [111]. We therefore included a campaign in the program theory to increase awareness among public and professional communities about counseling opportunities. By reaching out through an enhanced digital and analog presence, we aim to encourage family dementia caregivers to initiate contact to counseling services.

Embedding Counseling Services for Maintaining Ongoing Contact

Inquiries about the caregivers' mental models during the interviews indicate that family dementia caregivers often do not distinguish between the different objectives of various support services. While professionals intend to give room for sharing caregiving experiences in support groups, for example, some of the participants also expected information to be conveyed at these meetings. Embedding counseling in a variety of support services may contribute to supplementing the supportive effect of individual interventions, to maintaining contact, and to actively offering support when need occurs. Interventions offering counseling as part of a comprehensive program with non-technology-assisted components provide examples of such an approach: One example combines telephone counseling with a group activity program for community-dwelling individuals with dementia, aiming to empower persons with dementia and their caregivers [97]. Another combines telephone coaching (Dementelcoach) with respite care, thereby enhancing the effectiveness of the intervention [68]. A third example links two eHealth caregiver interventions to existing Meeting Centers for persons with dementia and their family caregivers [63,96].

Integrating AI Technology to Mitigate the Shortage of Skilled Health Care Professionals

With care needs increasing due to the demographic development, the shortage of skilled health care workers results in higher workloads and accelerated turnover rates [112]. Various strategies are described for retaining health care staff [112]. Authors of the Dementelcoach study reported that professionals delivering the intervention were part-time workers in the psychogeriatric sector. These professionals were encouraged to expand their work hours to provide counseling in conjunction with their existing jobs [68]. Offering a complementary activity may help to meet existing support needs in the face of the shortage of skilled staff [68].

All participating counselors reported a high workload, and the growing scarcity of skilled professionals may further exacerbate work-related stress [113]. To reduce time spent on documentation and administrative tasks, AI technology will be integrated into the intervention at the level of partial automation by augmenting human performance [113]. In using large

XSL•FO
RenderX

language models for transcribing counseling sessions and composing summaries, counselors are relieved of documentation tasks and allowed to focus solely on counselees' concerns during sessions. AI can also be used for scheduling tasks and to compile lists of regionally available support services for caregivers based on an AI-augmented database [114]. AI-drafted content will be evaluated by counselors to ensure accuracy and quality of information. This may help to alleviate concerns associated with the use of AI in the context of vulnerable groups—in particular, concerns about the reliability of information, privacy, and data security [113].

In sum, technology can make an important contribution to overcoming problems arising from the increasing need for support due to the growing number of persons with dementia [6], the growing shortage of skilled health care professionals [115], and social and infrastructural inequalities [116]. It is essential that technology-assisted counseling interventions are designed with consideration of individual and situational requirements in order to enhance their acceptability and feasibility from the perspectives of both recipients and providers [117].

Strengths and Limitations

We established a profound database for formulating an initial program theory and preparing for the development of a COS by integrating knowledge obtained from literature and the lived experience of two interest-holder groups. Participants in the qualitative study were recruited all over Germany. Caregivers living in urban and rural regions shared their experiences with counseling services, but we did not succeed in recruiting a heterogeneous sample in terms of age and gender. This may impose limitations on findings, as perspectives of younger and diverse caregivers, who might be more technologically savvy, are underrepresented. Younger family caregivers who are also employed, particularly long-distance caregivers, may be more inclined to use ICT to access counseling services. By combining various technological access options with in-person counseling, we advocate for an inclusive approach that addresses the needs of diverse target groups.

We were able to identify commonalities and differences in the interest-holders' expectations and mental models, which were incorporated into the program theory.

Another limitation is the inclusion of only two groups of interest-holders. Due to the limited duration of the project, we were not able to expand the study population of the qualitative inquiry. Representatives of providing organizations and policymakers will be included in the consensus process of the COS.

We followed accepted methodological guidance [13,17] and integrated methodological approaches in an innovative way. This proved to be beneficial: When asked about potential outcomes for measuring effectiveness of counseling, interviewees reflected on and critically questioned their own ideas and expectations. We found it advantageous to consider adequate outcomes from the beginning of the development process, to debate underlying (hypothesized) causal relationships, and to visualize causal assumptions in the logic model. This theory-led approach to the development, implementation, and evaluation of a technology-assisted counseling intervention for family caregivers of persons with dementia is consistent with the recommendations of the updated UK Medical Research Council guidance [9].

Future Directions

We will perform a theory-led approach to modeling, implementing, and evaluating the technology-assisted counseling intervention for family dementia caregivers. In a first step, the "long list" of outcomes will be created in accordance with the COS development methodology [13]. The participants of the qualitative study found it surprisingly difficult to name potential outcomes for assessing the effectiveness of counseling interventions, which led to a high number of (partially) overlapping outcomes. In a follow-up study, we will involve representatives of interest-holder groups such as family caregivers, counselors, managers of provider organizations, and researchers to integrate outcomes reported in literature (Table 1) and outcomes identified by interview partners (Table 3): Redundant outcomes will be removed; for example, outcomes such as "burden," "guilt," and "stress or distress" were identified in both the literature and interviews. The remaining outcomes will then be grouped into outcome domains applying ontologies, as proposed by Williamson et al [13]. The resulting "long list" of clustered outcomes forms the basis for the consensus process of the COS [13]. We will apply the Delphi approach involving groups of relevant interest-holders to determine important clinical outcomes [13]. To establish a shared understanding among participating interest-holders, elements of the program theory will be gradually integrated into the consensus process. At the outset of this process, we will use the logic model to visualize and critically discuss the theory of change. Subsequently, the importance of individual outcomes will be rated in Delphi rounds and finally consented in a consensus conference [13].

The refined logic model and program theory will then be used for modeling the intervention in collaboration with practice partners in a future research project. The logic model will serve to foster a shared understanding among the individuals involved and to systematically document the adaptation of key components to the specific context [9,18].

Conclusions

We developed an initial program theory of a technology-assisted counseling intervention for family caregivers of persons with dementia by introducing a methodological innovation. Findings obtained from interest-holder groups and literature are synthesized into a program theory and visualized by a logic model. We have also compiled a comprehensive list of potential outcomes, which includes the outcomes examined in clinical studies and those that are relevant from the perspective of interest-holders. This enables the consensus process for finalizing the COS for technology-assisted counseling interventions.

These results will inform the theory-led modeling, implementation, and evaluation of the intervention, which will include a customized ICT package. This package has the

XSL•FO
RenderX

potential to improve accessibility to counseling for caregivers by overcoming disparities in access to health care services. In addition, the design of the intervention can positively impact work conditions for health care professionals delivering support and improve the efficiency of services.

Acknowledgments

We thank the caregivers and counselors who participated in the qualitative study for devoting their time and sharing their experiences. In addition, we thank the members of the study advisory board for their valuable contributions and feedback. We also acknowledge the financial support of the Open Access Publication Fund of Martin Luther University Halle-Wittenberg.

Funding

This work was supported by the German Research Foundation (DFG, project number 533628714). The funding source had no role in the design and conduct of the study and preparation of the manuscript.

Data Availability

Data extracted from studies included in the scoping review are included in Multimedia Appendix 3.

Authors' Contributions

Conceptualization: DB, GM
Formal analysis: DB
Funding acquisition: DB, GM
Investigation: DB
Supervision: GM
Validation: RH, JW, GM, AB
Writing – original draft: DB
Writing – review & editing: GM, AB, JW, RH

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies.
[DOCX File, 52 KB - jmir_v28i1e81669_app1.docx]

Multimedia Appendix 2
Interview guide.
[DOCX File, 39 KB - jmir_v28i1e81669_app2.docx]

Multimedia Appendix 3
Description of reports and interventions.
[DOCX File, 206 KB - jmir_v28i1e81669_app3.docx]

Checklist 1
PRISMA-ScR, SRQR, and COS-STAR checklists.
[DOCX File, 72 KB - jmir_v28i1e81669_app4.docx]

References

1. Ferrero-Sereno P, Mendoza-Muñoz M, Palomo-López P, et al. A systematic review of the effectiveness of technological interventions for caregivers of people with dementia: effects on quality of life and psychoemotional variables. Front Public Health 2025;13:1579239. [doi: 10.3389/fpubh.2025.1579239] [Medline: 40438051]
2. Saragih ID, Tonapa SI, Porta CM, Lee BO. Effects of telehealth intervention for people with dementia and their carers: a systematic review and meta‐analysis of randomized controlled studies. J Nurs Scholarsh 2022 Nov;54(6):704-719. [doi: 10.1111/jnu.12797] [Medline: 35769007]
3. González-Fraile E, Ballesteros J, Rueda JR, Santos-Zorrozúa B, Solà I, McCleery J. Remotely delivered information, training and support for informal caregivers of people with dementia. Cochrane Database Syst Rev 2021 Jan 4;1(1):CD006440. [doi: 10.1002/14651858.CD006440.pub3] [Medline: 33417236]

XSL•FO
RenderX

4. Gilsenan J, Gorman C, Shevlin M. Explaining caregiver burden in a large sample of UK dementia caregivers: the role of contextual factors, behavioural problems, psychological resilience, and anticipatory grief. Aging Ment Health 2023;27(7):1274-1281. [doi: 10.1080/13607863.2022.2102138] [Medline: 35881027]
5. Lindeza P, Rodrigues M, Costa J, Guerreiro M, Rosa MM. Impact of dementia on informal care: a systematic review of family caregivers' perceptions. BMJ Support Palliat Care 2020 Oct 14. [doi: 10.1136/bmjspcare-2020-002242] [Medline: 33055092]
6. GBD 2019 Dementia Forecasting Collaborators. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. Lancet Public Health 2022 Feb;7(2):e105-e125. [doi: 10.1016/S2468-2667(21)00249-8] [Medline: 34998485]
7. Butcher HK, Bulechek GM, Dochterman JM, Wagner CM, editors. Nursing Interventions Classification (NIC), 7th edition: Elsevier; 2018.
8. Bauernschmidt D, Wittmann J, Hirt J, Meyer G, Bieber A. The implementation success of technology-based counseling in dementia care: scoping review. JMIR Aging 2024 Jan 25;7:e51544. [doi: 10.2196/51544] [Medline: 38271050]
9. Skivington K, Matthews L, Simpson SA, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061. [doi: 10.1136/bmj.n2061] [Medline: 34593508]
10. Cheng JY, Nurul S, Cheng LJ, He HG. Effectiveness of technology-delivered psychosocial interventions for family caregivers of patients with dementia: a systematic review, meta-analysis and meta-regression. Int J Ment Health Nurs 2024 Dec;33(6):1796-1816. [doi: 10.1111/inm.13390] [Medline: 39034437]
11. Bauernschmidt D, Hirt J, Langer G, et al. Technology-based counselling for people with dementia and their informal carers: a systematic review and meta-analysis. J Alzheimers Dis 2023;93(3):891-906. [doi: 10.3233/JAD-221194] [Medline: 37125549]
12. Bauernschmidt D, Wittmann J, Bieber A, Meyer G. Integrating programme theory into the development of a core outcome set for technology-assisted counselling interventions in dementia: study protocol of the ProCOS study. BMJ Open 2024 Aug 6;14(8):e081526. [doi: 10.1136/bmjopen-2023-081526] [Medline: 39107024]
13. Williamson PR, Altman DG, Bagley H, et al. The COMET Handbook: version 1.0. Trials 2017 Jun 20;18(Suppl 3):280. [doi: 10.1186/s13063-017-1978-4] [Medline: 28681707]
14. Core Outcome Measures in Effectiveness Trials. COMET Database. URL: https://www.comet-initiative.org/Studies [accessed 2025-12-20]
15. Reilly ST, Harding AJE, Morbey H, et al. What is important to people with dementia living at home? A set of core outcome items for use in the evaluation of non-pharmacological community-based health and social care interventions. Age Ageing 2020 Jul 1;49(4):664-671. [doi: 10.1093/ageing/afaa015]
16. Dementia outcome measures: charting new territory; report of a JPND Working Group on Longitudinal Cohorts. : Neurodegenerative Disease Research; 2015 Oct URL: https://www.neurodegenerationresearch.eu/wp-content/uploads/2015/10/JPND-Report-Fountain.pdf [accessed 2025-12-20]
17. Funnell S, Rogers PJ. Purposeful Program Theory: Effective Use of Theories of Change and Logic Models: Jossey-Bass; 2011.
18. Skivington K, Matthews L, Simpson SA, et al. Framework for the development and evaluation of complex interventions: gap analysis, workshop and consultation-informed update. Health Technol Assess 2021 Sep;25(57):1-132. [doi: 10.3310/hta25570] [Medline: 34590577]
19. Anderson LM, Petticrew M, Rehfuess E, et al. Using logic models to capture complexity in systematic reviews. Res Synth Methods 2011 Mar;2(1):33-42. [doi: 10.1002/jrsm.32] [Medline: 26061598]
20. Baxter SK, Blank L, Woods HB, Payne N, Rimmer M, Goyder E. Using logic model methods in systematic review synthesis: describing complex pathways in referral management interventions. BMC Med Res Methodol 2014 May 10;14:62. [doi: 10.1186/1471-2288-14-62] [Medline: 24885751]
21. Bleijenberg N, de Man-van Ginkel JM, Trappenburg JCA, et al. Increasing value and reducing waste by optimizing the development of complex interventions: enriching the development phase of the Medical Research Council (MRC) framework. Int J Nurs Stud 2018 Mar;79:86-93. [doi: 10.1016/j.ijnurstu.2017.12.001] [Medline: 29220738]
22. Hirt J, Langer G, Wilde F, Bauernschmidt D, Meyer G, Bieber A. Technology-based counselling in dementia (TeCoDem): study protocol of a mixed-methods systematic review with qualitative comparative analysis and meta-analysis. BMJ Open 2021 Dec 8;11(12):e054157. [doi: 10.1136/bmjopen-2021-054157] [Medline: 34880025]
23. Core Outcome Measures in Effectiveness Trials. Registration ProCOS. URL: https://www.comet-initiative.org/Studies/Details/2884 [accessed 2025-12-20]
24. Anguera MT, Blanco-Villaseñor A, Losada JL, Sánchez-Algarra P, Onwuegbuzie AJ. Revisiting the difference between mixed methods and multimethods: is it all in the name? Qual Quant 2018 Nov;52(6):2757-2770. [doi: 10.1007/s11135-018-0700-2]
25. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
26. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]

XSL•FO
RenderX

27. Kirkham JJ, Gorst S, Altman DG, et al. Core Outcome Set—Standards for Reporting: the COS-STAR statement. PLoS Med 2016 Oct;13(10):e1002148. [doi: 10.1371/journal.pmed.1002148] [Medline: 27755541]
28. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evidence Synthesis 2020;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167]
29. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev 2016 Dec 5;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
30. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: Template for Intervention Description and Replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348:g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]
31. Möhler R, Köpke S, Meyer G. Criteria for Reporting the Development and Evaluation of Complex Interventions in healthcare: revised guideline (CReDECI 2). Trials 2015 Dec;16(1). [doi: 10.1186/s13063-015-0709-y]
32. Zahavi D, Martiny KMM. Phenomenology in nursing studies: new perspectives. Int J Nurs Stud 2019 May;93:155-162. [doi: 10.1016/j.ijnurstu.2019.01.014] [Medline: 30795899]
33. Akremi L. Stichprobenziehung in der qualitativen Sozialforschung. In: Baur N, Blasius J, editors. Handbuch Methoden Der Empirischen Sozialforschung: Springer; 2022:405-424. [doi: 10.1007/978-3-658-37985-8_26]
34. Schreier M. Fallauswahl. In: Mey G, Mruck K, editors. Handbuch Qualitative Forschung in der Psychologie: Band 2: Designs und Verfahren: Springer; 2020:19-39. [doi: 10.1007/978-3-658-26887-9]
35. Rahimi S, Khatooni M. Saturation in qualitative research: an evolutionary concept analysis. Int J Nurs Stud Adv 2024 Jun;6(100174):100174. [doi: 10.1016/j.ijnsa.2024.100174] [Medline: 38746797]
36. Helfferich C. Leitfaden- und Experteninterviews. In: Baur N, Blasius J, editors. Handbuch Methoden der empirischen Sozialforschung: Springer; 2022:875-892. [doi: 10.1007/978-3-658-37985-8_55]
37. Audiotranskription. Audiotranskription. URL: https://www.audiotranskription.de [accessed 2025-12-20]
38. Diekelmann NL. Learning-as-testing: a Heideggerian hermeneutical analysis of the lived experiences of students and teachers in nursing. ANS Adv Nurs Sci 1992 Mar;14(3):72-83. [doi: 10.1097/00012272-199203000-00010] [Medline: 1550334]
39. Morbey H, Harding AJE, Swarbrick C, et al. Involving people living with dementia in research: an accessible modified Delphi survey for core outcome set development. Trials 2019 Jan 6;20(1):12. [doi: 10.1186/s13063-018-3069-6] [Medline: 30612587]
40. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. [doi: 10.1136/bmj.n71]
41. Hodgson NA, Petrovsky DV, Finegan K, Kallmyer BA, Pike J, Fazio S. One call makes a difference: an evaluation of the Alzheimer's Association National Helpline on dementia caregiver outcomes. Patient Educ Couns 2021 Apr;104(4):896-902. [doi: 10.1016/j.pec.2020.09.026] [Medline: 33004235]
42. Coyne AC, Potenza M, Nose MAB. Caregiving and dementia: the impact of telephone helpline services. Am J Alzheimers Dis (Columbia) 1995 Jul;10(4):27-32. [doi: 10.1177/153331759501000407]
43. Tremont G, Davis JD, Bishop DS, Fortinsky RH. Telephone-delivered psychosocial intervention reduces burden in dementia caregivers. Dementia (London) 2008;7(4):503-520. [doi: 10.1177/1471301208096632] [Medline: 20228893]
44. Telephone support for dementia caregivers: NCT00735800. National Institutes of Health. 2008. URL: https://clinicaltrials.gov/ct2/show/NCT00735800 [accessed 2025-12-20]
45. Tremont G, Davis J, O'Connor K, et al. Relationship between expectancy/credibility and early response to telephone-based dementia caregiver interventions. Alzheimer's & Dementia 2011 Jul;7(4S_Part_12):S435. [doi: 10.1016/j.jalz.2011.05.1257]
46. Tremont G, Davis J, Grover C, et al. Randomized controlled trial of a telephone‐delivered intervention (FITT‐Caregiver) for dementia caregivers. Alzheimer's & Dementia 2013 Jul;9(4S_Part_8):324-P325. [doi: 10.1016/j.jalz.2013.04.161]
47. Tremont G, Davis JD, Papandonatos GD, et al. A telephone intervention for dementia caregivers: background, design, and baseline characteristics. Contemp Clin Trials 2013 Nov;36(2):338-347. [doi: 10.1016/j.cct.2013.07.011] [Medline: 23916916]
48. Tremont G, Davis J, Bryant K, et al. Effect of a telephone‐based dementia caregiver intervention on use of community support services and health care resources. Alzheimer's & Dementia 2014 Jul;10(4S_Part_3):226-P227. [doi: 10.1016/j.jalz.2014.04.319]
49. Tremont G, Davis JD, Papandonatos GD, et al. Psychosocial telephone intervention for dementia caregivers: a randomized, controlled trial. Alzheimers Dement 2015 May;11(5):541-548. [doi: 10.1016/j.jalz.2014.05.1752] [Medline: 25074341]
50. Tremont G, Davis JD, Ott BR, et al. Randomized trial of the Family Intervention: Telephone Tracking-Caregiver for dementia caregivers: use of community and healthcare resources. J Am Geriatr Soc 2017 May;65(5):924-930. [doi: 10.1111/jgs.14684] [Medline: 28008609]
51. Does telehealth delivery of the COPE program provide a non-inferior alternative to face-to-face treatment for community dwelling people with cognitive impairment?: ACTRN12617000117314. Australian New Zealand Clinical Trials Registry. 2017. URL: https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?id=372140&isReview=true [accessed 2025-12-20]
52. Laver K, Liu E, Clemson L, et al. Does telehealth delivery of a dyadic dementia care program provide a noninferior alternative to face-to-face delivery of the same program? A randomized, controlled trial. Am J Geriatr Psychiatry 2020 Jun;28(6):673-682. [doi: 10.1016/j.jagp.2020.02.009] [Medline: 32234275]

XSL•FO
RenderX

53. Laver K. Lessons learned from offering a telehealth intervention for people with dementia and their caregivers. Alzheimers Dement 2021;17(S7):e051024. [doi: 10.1002/alz.051024]
54. The Residential Care Transition Module (RCTM): NCT02915939. National Institutes of Health. URL: https://clinicaltrials.gov/ct2/show/NCT02915939 [accessed 2025-12-20]
55. Gaugler JE, Statz TL, Birkeland RW, et al. The Residential Care Transition Module: a single-blinded randomized controlled evaluation of a telehealth support intervention for family caregivers of persons with dementia living in residential long-term care. BMC Geriatr 2020 Apr 15;20(1):133. [doi: 10.1186/s12877-020-01542-7] [Medline: 32293314]
56. Statz TL, Peterson CM, Birkeland RW, et al. "We moved her too soon": navigating guilt among adult child and spousal caregivers of persons living with dementia following a move into residential long-term care. Couple Family Psychol 2022 Dec;11(4):300-314. [doi: 10.1037/cfp0000150] [Medline: 36743783]
57. Zmora R, Statz TL, Birkeland RW, et al. Transitioning to long-term care: family caregiver experiences of dementia, communities, and counseling. J Aging Health 2021 Jan;33(1-2):133-146. [doi: 10.1177/0898264320963588] [Medline: 32990494]
58. Zmora R, Statz TL, Birkeland RW, et al. Corrigendum: transitioning to long-term care: family caregiver experiences of dementia, communities, and counseling. J Aging Health 2021 Jan;33(1-2):133-146. [doi: 10.1177/0898264320963588]
59. Albers EA, Birkeland RW, Louwagie KW, et al. A qualitative analysis of mechanisms of benefit in the residential care transition module: a telehealth intervention for caregivers of relatives with dementia living in residential long-term care. Inquiry 2023;60. [doi: 10.1177/00469580231217981] [Medline: 38142369]
60. Gaugler JE, Birkeland RW, Albers EA, et al. Efficacy of the residential care transition module: a telehealth intervention for dementia family caregivers of relatives living in residential long-term care settings. Psychol Aging 2024 Aug;39(5):565-577. [doi: 10.1037/pag0000820] [Medline: 38753405]
61. Meeting Centers 3.0; Evaluation of the implementation and costeffectiveness of three new tailored interventions for people with dementia and their carers: dementalent, dementelcoach and STAR-online training: NTR5521. Netherlands Trial Register. URL: https://www.trialregister.nl/trial/5396 [accessed 2025-12-20]
62. Droes RM, Rijn A, Bosmans J, Meiland F. The individualized Meeting Centers Support Program; evaluation of benefits and costs [Abstract]. In: Book of Abstracts 19th IPA International Congress 2019, Vol. 31. [doi: 10.1017/S1041610219001339]
63. Dröes RM, van Rijn A, Rus E, Dacier S, Meiland F. Utilization, effect, and benefit of the individualized Meeting Centers Support Program for people with dementia and caregivers. Clin Interv Aging 2019;14:1527-1553. [doi: 10.2147/CIA.S212852] [Medline: 31692559]
64. Can the residential care transition module improve the psychological health of family carers of people with dementia during the residential care placement process in Australia?: ACTRN12621001462875. Australian New Zealand Clinical Trials Registry. 2021. URL: https://anzctr.org.au/Trial/Registration/TrialReview.aspx?id=382745&isReview=true [accessed 2025-12-20]
65. Brooks D, Wyles K, Pachana NA, Beattie E, Gaugler JE. Tailored videoconferencing counselling program to support family carers of people living with dementia during the transition to permanent residential care: a pilot and feasibility randomised trial. BMC Geriatr 2024 Apr 26;24(1):375. [doi: 10.1186/s12877-024-04907-4] [Medline: 38671367]
66. Natale G, Zigoura E, Carpaneto V, et al. Telephone calls provide effective support for most caregivers of patients with dementia but not for all. Int J Geriatr Psychiatry 2012 Feb;27(2):215-216. [doi: 10.1002/gps.2710] [Medline: 22223146]
67. Mierlo L, Meiland F, Droes R. Dementelcoach: effect of telephone coaching on informal and professional carers of community dwelling people with dementia [Poster]. Int Psychogeriatr 2011;23:378-379. [doi: 10.1017/S1041610211001293]
68. van Mierlo LD, Meiland FJM, Dröes RM. Dementelcoach: effect of telephone coaching on carers of community-dwelling people with dementia. Int Psychogeriatr 2012 Feb;24(2):212-222. [doi: 10.1017/S1041610211001827] [Medline: 21995966]
69. Rice JD, Sperling SA, Brown DS, Mittleman MS, Manning CA. Evaluating the efficacy of TeleFAMILIES: a telehealth intervention for caregivers of community-dwelling people with dementia. Aging Ment Health 2022 Aug;26(8):1613-1619. [doi: 10.1080/13607863.2021.1935462] [Medline: 34125635]
70. Wilkinson S. An analysis and evaluation of the admiral nurse dementia helpline. : Dementia UK; 2016.
71. Gilliard J, Keady J, Evers C, Milton S. Telephone helplines for people with dementia. Int J Geriatr Psychiatry 1998;13(10):734-735. [doi: 10.1002/(SICI)1099-1166(1998100)13:10<734::AID-GPS846>3.0.CO;2-X] [Medline: 9818311]
72. Jansen S. Fünf Jahre Alzheimer-Telefon der Deutschen Alzheimer Gesellschaft. Nervenheilkunde 2007;26(08):685-689. [doi: 10.1055/s-0038-1626915]
73. Pendergrass A, Weiß S, Gräßel E. Zu welchen Themen suchen Angehörige von Demenzbetroffenen telefonische oder emailbasierte Beratung auf? Aktuelle Ergebnisse des bundesweiten Beratungsangebots der Deutschen Alzheimer Gesellschaft e.V. Gesundheitswesen 2019 Dec;81(12):1018-1021. [doi: 10.1055/a-0586-8809] [Medline: 29672813]
74. Kurz A, Fischer IM, Dogan V, Kurz C. Telefonische Hilfe zur Selbsthilfe: Das Alzheimer-Telefon aus der Sicht der Nutzer:innen. Ger Med Sci 2024. [doi: 10.3205/000331] [Medline: 38883339]
75. Harvey R, Roques PK, Fox NC, Rossor MN. CANDID - Counselling and Diagnosis in Dementia: a national telemedicine service supporting the care of younger patients with dementia. Int J Geriatr Psychiatry 1998;13(6):381-388. [doi: 10.1002/(sici)1099-1166(199806)13:63.0.co;2-m]

XSL•FO
RenderX

76. Silverstein NM, Kennedy K, McCormick D. A telephone helpline for Alzheimer 's disease: Information, referral, and support. American Journal of Alzheimer's Care and Related Disorders & Research 1993 Sep;8(5):28-36. [doi: 10.1177/153331759300800507]
77. Rozani V, Pilko A, Kagan I. Provision of integrated care by a National Call Support Program for Alzheimer's patients and their caregivers. Int J Integr Care 2022;22(S3):91. [doi: 10.5334/ijic.ICIC22294]
78. Nakano Y, Hishikawa N, Sakamoto K, et al. A unique telephone support system for dementia patients and their caregivers managed in Japan (Okayama Dementia Call Center, ODCC). Neurology & Clinical Neurosc 2018 Jul;6(4):100-103 [FREE Full text] [doi: 10.1111/ncn3.12200]
79. Tousi B, Kanetsky C, Udelson N. ALZ i-Connect. Am J Alzheimers Dis Other Demen 2017 Feb;32(1):63-66. [doi: 10.1177/1533317516677615] [Medline: 27899432]
80. Kelly B. Link2Care: internet-based information and support for caregivers. Generations 2003;27(4):87-88 [FREE Full text]
81. Rentz M, Hoene A. Online coaching for caregivers: using technology to provide support and information. Alzheimers Care 2010;11(3):206-209 [FREE Full text] [doi: 10.1097/ACQ.0b013e3181ebc878]
82. De Cola MC, De Luca R, Bramanti A, Bertè F, Bramanti P, Calabrò RS. Tele-health services for the elderly: a novel southern Italy family needs-oriented model. J Telemed Telecare 2016 Sep;22(6):356-362. [doi: 10.1177/1357633X15604290] [Medline: 26377125]
83. Wheat H, Griffiths S, Gude A, et al. Practitioners' ability to remotely develop understanding for personalised care and support planning: a thematic analysis of multiple data sources from the feasibility phase of the Dementia Personalised Care Team (D-PACT) intervention. Dementia (London) 2023 Oct;22(7):1461-1486. [doi: 10.1177/14713012231185281] [Medline: 37354084]
84. Chang BL, Nitta S, Carter PA, Markham YK. Perceived helpfulness of telephone calls. J Gerontol Nurs 2004 Sep;30(9):14-21. [doi: 10.3928/0098-9134-20040901-05]
85. Salfi J. Seeking to understand telephone support for dementia caregivers: a qualitative case study [Dissertation]. : McMaster University (Canada); 2004 Sep URL: https://prod-ms-be.lib.mcmaster.ca/server/api/core/bitstreams/3c270094-5285-4532-a851-a9d4b77a3ba0/content [accessed 2025-12-20]
86. Salfi J, Ploeg J, Black ME. Seeking to understand telephone support for dementia caregivers. West J Nurs Res 2005 Oct;27(6):701-721. [doi: 10.1177/0193945905276882] [Medline: 16157943]
87. Spilsbury K. Telephone support met the perceived needs of dementia caregivers for convenient access to information, referral, and emotional support. Evid Based Nurs 2006 Jul;9(3):94. [doi: 10.1136/ebn.9.3.94] [Medline: 16865845]
88. Madden G, Rose T, Crystal L. Using video consultations to support family carers of people living with dementia. Nurs Older People 2022 Feb 1;34(1):28-33. [doi: 10.7748/nop.2021.e1346] [Medline: 34431259]
89. Dorell Å, Konradsen H, Kallström AP, Kabir ZN. "A friend during troubled times": experiences of family caregivers to persons with dementia when receiving professional support via a mobile app. PLoS ONE 2022;17(8):e0271972. [doi: 10.1371/journal.pone.0271972] [Medline: 35917295]
90. Chiu T, Lottridge D. Development and iterative refinement of an internet-based service for Chinese family caregivers of people with Alzheimer disease. AMIA Annu Symp Proc 2005;2005:919. [Medline: 16779206]
91. Chiu TML, Marziali E, Tang M, Colantonio A, Carswell A. Client-centred concepts in a personalized e-mail support intervention designed for Chinese caregivers of family members with dementia: a qualitative study. Hong Kong J Occup Ther 2010 Dec;20(2):87-93. [doi: 10.1016/S1569-1861(11)70008-0]
92. Chiu TML, Eysenbach G. Theorizing the health service usage behavior of family caregivers: a qualitative study of an internet-based intervention. Int J Med Inform 2011 Nov;80(11):754-764. [doi: 10.1016/j.ijmedinf.2011.08.010]
93. Brown P, Oliver E, Dening KH. Supporting family carers via the Admiral Nurse Dementia Helpline: reflection on a case study. Nurs Older People 2020 Sep 22;32(5):16-20. [doi: 10.7748/nop.2020.e1248] [Medline: 32400141]
94. Drayton S, Dening KH. Achieving positive outcomes in complex cases: the Admiral Nurse Dementia Helpline (Innovative Practice). Dementia (London) 2020 May;19(4):1308-1315. [doi: 10.1177/1471301217740005] [Medline: 29132219]
95. Sabat SR. Flourishing of the self while caregiving for a person with dementia: a case study of education, counseling, and psychosocial support via email. Dementia (London) 2011 Feb;10(1):81-97. [doi: 10.1177/1471301210392986]
96. van Rijn A, Meiland F, Dröes RM. Linking two new e-health caregiver interventions to meeting centres for people with dementia and their carers; a process evaluation. Aging Ment Health 2020 Aug;24(8):1316-1325. [doi: 10.1080/13607863.2019.1617243] [Medline: 31119946]
97. Nomura M, Makimoto K, Kato M, et al. Empowering older people with early dementia and family caregivers: a participatory action research study. Int J Nurs Stud 2009 Apr;46(4):431-441. [doi: 10.1016/j.ijnurstu.2007.09.009] [Medline: 17983619]
98. Perales-Puchalt J, Acosta-Rullán M, Ramírez-Mantilla M, et al. A text messaging intervention to support Latinx family caregivers of individuals with dementia (CuidaTEXT): development and usability study. JMIR Aging 2022;5(2):e35625. [doi: 10.2196/35625]
99. Chiu TML. Usage and non-usage behaviour of ehealth services among Chinese Canadians caring for a family member with dementia [Dissertation]. : University of Toronto; 2008 Jul URL: https://utoronto.scholaris.ca/items/1ef09030-a269-4a8c-9950-7e85019059d3 [accessed 2025-12-20]

XSL•FO
RenderX

100. Chiu T, Marziali E, Colantonio A, et al. Internet-based caregiver support for Chinese Canadians taking care of a family member with Alzheimer disease and related dementia. Can J Aging 2009 Dec;28(4):323-336. [doi: 10.1017/S0714980809990158]
101. Chiu TML, Eysenbach G. Stages of use: consideration, initiation, utilization, and outcomes of an internet-mediated intervention. BMC Med Inform Decis Mak 2010 Nov 23;10(1):73. [doi: 10.1186/1472-6947-10-73] [Medline: 21092275]
102. Barbabella F, Poli A, Andréasson F, et al. A web-based psychosocial intervention for family caregivers of older people: results from a mixed-methods study in three European countries. JMIR Res Protoc 2016 Oct 6;5(4):e196. [doi: 10.2196/resprot.5847] [Medline: 27713113]
103. Barbabella F, Poli A, Hanson E, et al. Usage and usability of a web-based program for family caregivers of older people in three European countries: a mixed-methods evaluation. Comput Inform Nurs 2018 May;36(5):232-241. [doi: 10.1097/CIN.0000000000000422] [Medline: 29505433]
104. CuidaTXT: a text message dementia-caregiver intervention for Latinos: NCT04316104. National Institutes of Health. 2020. URL: https://clinicaltrials.gov/study/NCT04316104 [accessed 2025-12-20]
105. Puchalt JP. CuidaTXT: a text message dementia-caregiver intervention for Latinos. Innov Aging 2020;4(Suppl 1):769. [doi: 10.1093/geroni/igaa057.2777]
106. Perales-Puchalt J, Ramírez-Mantilla M, Fracachán-Cabrera M, et al. A text message intervention to support Latino dementia family caregivers (CuidaTEXT): feasibility study. Clin Gerontol 2024;47(1):50-65. [doi: 10.1080/07317115.2022.2137449] [Medline: 36268684]
107. Lamura G, Poli A, Yghemonos S, Barbabella F. InformCare: the European information hub on family care. International Journal of Care and Caring 2017;1(3):409-413. [doi: 10.1332/239788217X15018371295074]
108. Brown P, Oliver E, Dening KH. Increasing need for telehealth services for families affected by dementia as a result of Covid-19. J Community Nurs 2020;34(5):59-64 [FREE Full text]
109. Twaddle IKB, Hattori-Uchima MP, Orallo RG, Gutierrez NJ. Telehealth outreach programming in the Pacific Island of Guam: providing access to dementia care support services during the COVID-19 pandemic. Alzheimers Dement 2021 Dec;17 Suppl 8(Suppl 8):e050134. [doi: 10.1002/alz.050134] [Medline: 34971289]
110. Caprioli T, Mason S, Tetlow H, Limbert S, Reilly S, Giebel C. "Necessity is the mother of invention": experiences of accessing and delivering dementia-related support services by information communication technology during the pandemic in the UK. Dementia (London) 2025 Feb;24(2):323-343. [doi: 10.1177/14713012241272906] [Medline: 39117353]
111. Sorrentino M, Fiorilla C, Mercogliano M, et al. Barriers for access and utilization of dementia care services in Europe: a systematic review. BMC Geriatr 2025;25(1). [doi: 10.1186/s12877-025-05805-z]
112. De Vries N, Lavreysen O, Boone A, et al. Retaining healthcare workers: a systematic review of strategies for sustaining power in the workplace. Healthcare (Basel) 2023;11(13):1887. [doi: 10.3390/healthcare11131887]
113. Bienefeld N, Keller E, Grote G. AI interventions to alleviate healthcare shortages and enhance work conditions in critical care: qualitative analysis. J Med Internet Res 2025;27:e50852. [doi: 10.2196/50852]
114. Li H, Fu JF, Python A. Implementing large language models in health care: clinician-focused review with interactive guideline. J Med Internet Res 2025;27:e71916. [doi: 10.2196/71916]
115. Scheffler RM, Arnold DR. Projecting shortages and surpluses of doctors and nurses in the OECD: what looms ahead. Health Econ Policy Law 2019 Apr;14(2):274-290. [doi: 10.1017/S174413311700055X] [Medline: 29357954]
116. Mistry SK, Shaw M, Raffan F, et al. Inequity in access and delivery of virtual care interventions: a scoping review. Int J Environ Res Public Health 2022;19(15). [doi: 10.3390/ijerph19159411] [Medline: 35954768]
117. Coetzer JA, Loukili I, Goedhart NS, et al. The potential and paradoxes of eHealth research for digitally marginalised groups: a qualitative meta-review. Soc Sci Med 2024 Jun;350:116895. [doi: 10.1016/j.socscimed.2024.116895] [Medline: 38710135]

Abbreviations

AI: artificial intelligence
COMET: Core Outcome Measures in Effectiveness Trials Initiative
COS: core outcome set
COS-STAR: Core Outcome Set—Standards for Reporting
CReDECI 2: Revised Criteria for Reporting the Development and Evaluation of Complex Interventions
ICSS: Internet-Based Caregiver Support Service
ICT: information and communication technologies
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses, Extension for Scoping Reviews
ProCOS: Program Theory and Preparation of a Core Outcome Set Project
RCT: randomized controlled trial
SRQR: Standards for Reporting Qualitative Research
TIDieR: Template for Intervention Description and Replication

XSL•FO
RenderX

Edited by A Stone; submitted 01.Aug.2025; peer-reviewed by K Zhang, N Bievre, O Akhadelor; revised version received 25.Nov.2025; accepted 25.Nov.2025; published 20.Jan.2026.

Please cite as:
Bauernschmidt D, Bieber A, Hubrich R, Wittmann J, Meyer G
Program Theory and Core Outcome Set Development for a Technology-Assisted Counseling Intervention in Dementia: Multimethods Study
J Med Internet Res 2026;28:e81669
URL: https://www.jmir.org/2026/1/e81669
doi:10.2196/81669

© Dorothee Bauernschmidt, Anja Bieber, Ronja Hubrich, Janina Wittmann, Gabriele Meyer. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Validation of the Simplified Chinese Clear Communication Index Using Diabetes Education Materials: Instrument Adaptation and Validation Study

Lu Liu[1*], MSN; Yi Chen[2*], PhD; Xiaoqi Li[1], MSN; Hanyu Du[1], MSN; Rong Gui[1], MSN; Huiqing Jin[1], MSN; Fengfei Chen[1], MSN; Yujun Lin[1], MSN; Qiuhuan Huang[3,4], BSc; Yingge Tong[1], PhD

[1]Department of Nursing, School of Public Health and Nursing, Hangzhou Normal University, No. 2318, Yuhangtang Road, Hangzhou, China

[2]Department of Biomedicine and Prevention, University of Rome Tor Vergata, Rome, Italy

[3]Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, Guangxi Zhuang Autonomous Region, China

[4]Guangxi Clinical Medical Research Center for Hepatobiliary Diseases, Baise, Guangxi Zhuang Autonomous Region, China

[*]these authors contributed equally

Corresponding Author:
Yingge Tong, PhD
Department of Nursing, School of Public Health and Nursing, Hangzhou Normal University, No. 2318, Yuhangtang Road, Hangzhou, China

Abstract

Background: In the postpandemic context, the surge of digital health information has intensified public demand for clear and practical communication, particularly in China, where health literacy disparities persist. The Clear Communication Index, developed by the US Centers for Disease Control and Prevention (CDC), is a standardized tool for assessing the clarity and actionability of health materials, but no version adapted to the simplified Chinese context has been established.

Objective: This study aims to translate and culturally adapt the Clear Communication Index into simplified Chinese, and subsequently validate its psychometric properties using diabetes health communication materials from provincial CDCs across Mainland China, a disease area with a substantial public health burden and strong reliance on health education.

Methods: Following a standardized cross-cultural process (forward-back translation and expert review), we developed the simplified Chinese version of the Clear Communication Index (C-CCI) and finalized a 12-item scale across 4 dimensions (Main Message and Call to Action, Behavioral Recommendations, Numbers, Risk) with yes or no scoring (0 - 100). One top-ranked diabetes health education material was sampled from each provincial CDC website in Mainland China (30/31 included; 96.8%) on May 18, 2025. Twelve raters with multidisciplinary backgrounds completed a 3-week standardized training program and independently evaluated each article (360 ratings). Structural validity was examined using exploratory factor analysis and confirmatory factor analysis. Content validity was assessed by the item-level content validity index and the scale-level content validity index/average. Reliability was evaluated by internal consistency (Cronbach α) and inter-rater agreement (Fleiss kappa), while convergent/discriminant validity was assessed using composite reliability and average variance extracted (AVE).

Results: The 4-factor structure was supported. Content validity was high, with the scale-level content validity index/average values of 0.976 for clarity and 1 for relevance. Overall reliability was acceptable (Cronbach α=0.837), with particularly strong internal consistency in the Risk dimension (Cronbach α=0.910). Inter-rater agreement was substantial (κ=0.624). Convergent validity (composite reliability=0.897 - 0.914; AVE=0.645 - 0.831) and discriminant validity were satisfactory. Application to 30 provincial CDC websites yielded a mean C-CCI score of 53.84 (SD 29.74), well below the recommended threshold of 90. No significant regional differences were observed in total scores; however, Behavioral Recommendations scored slightly higher in western provinces than in eastern and central regions (η^2=0.034), representing a small effect size with limited practical significance.

Conclusions: The C-CCI demonstrated good validity, reliability, and feasibility for evaluating simplified Chinese health communication materials. These findings underscore the need to strengthen health communication practices in China and encourage provincial CDCs to align material development with national health literacy goals. Integrating the C-CCI into routine CDC review protocols could support evidence-based quality assurance and advance clearer, more actionable public health communication nationwide.

(J Med Internet Res 2026;28:e83935) doi:10.2196/83935

XSL•FO
RenderX

KEYWORDS

clear communication index; psychometric validation; cross-cultural adaptation; health literacy; health education materials; public health communication; diabetes education; simplified Chinese

Introduction

Effective communication of health information is critical to facilitating accurate public understanding and proper application [1]. In the postpandemic era, both public attention to and demand for health information have increased significantly [2]. However, limited health literacy remains a widespread challenge worldwide, affecting about 88% of adults in the United States, nearly 70% in China, and up to 62.1% in parts of Europe [3-5]. Complex medical terminology and dense sentence structures often impede comprehension, especially among populations with limited literacy, contributing to poor adherence, adverse outcomes, and higher mortality [6,7]. These challenges highlight the need for clear, actionable, and audience-appropriate health materials.

To address these challenges, several standardized tools have been developed internationally to evaluate the quality of health communication materials, including the Suitability Assessment of Materials (SAM) [8], the Patient Education Materials Assessment Tool (PEMAT) [9], and the Clear Communication Index (CCI) [10]. These instruments differ in both scope and analytical emphasis. SAM adopts a broad suitability-oriented framework emphasizing presentation and cultural appropriateness, while PEMAT focuses on understandability and actionability. However, neither tool explicitly prioritizes key messages or systematically integrates numeric and risk communication.

In addition, several tools, including the Credible and Usable Evaluation of patient education tool for websites [11], the Health Information Website Evaluation Tool [12], the Health Information Quality Assessment Tool [13], and the QUality Evaluation Scoring Tool [14], have been developed to assess the quality and credibility of health information, particularly for professional or web-based contexts. However, their primary focus on credibility and usability limits their applicability to audience-centered, action-oriented evaluation of public health materials.

In contrast to these approaches, the CCI incorporates audience-centered principles, prioritizing key messages, numerical clarity, and behavioral guidance, making it a more comprehensive and action-oriented evaluation framework [10]. Among these instruments, the SAM and PEMAT have been translated and validated in simplified Chinese [15,16]; however, a validated simplified Chinese version of the Clear Communication Index (C-CCI) has not yet been developed.

Developed by the US Centers for Disease Control and Prevention (CDC) in 2014, the CCI provides a practical and evidence-based method for assessing the clarity of health materials [10]. It is available in a full version (20 items across 7 dimensions) for longer materials and a modified version (13 items across 4 dimensions) designed for short-format communications, both using binary scoring with a recommended minimum score of 90 to indicate acceptable clarity and usability [10,17,18]. The CCI has been widely applied across health domains, including patient education, cancer communication, and environmental health [19-27], and cross-culturally validated in countries, such as Brazil and Thailand [28,29]. A recent systematic review classified the CCI as a Grade B recommended instrument for health communication evaluation [30].

However, no validated simplified C-CCI currently exists. Given linguistic and cultural differences between Chinese and English, direct adoption may limit its accuracy and relevance [31]. In China, health communication is largely delivered through a nationwide network of CDCs that translate national health policies into public education, particularly for chronic disease prevention and management [32]. Diabetes, as a major noncommunicable disease, represents a central focus of CDC health education across provinces. Given provincial CDCs' mandate to bridge policy objectives and public engagement across regions, a standardized, context-sensitive framework for evaluating their educational materials is needed.

To address this gap, this study aimed to translate and culturally adapt the original CCI into a simplified Chinese version (C-CCI) through a standardized cross-cultural process and to evaluate its reliability and validity using health communication materials from provincial CDCs across Mainland China.

Methods

Translation and Cultural Adaptation of the CCI Tool

After obtaining formal authorization to use and adapt the original CCI, the research team followed the standardized cross-cultural adaptation model proposed by Sperber et al [33]. The process was conducted in 4 stages. First, 2 native simplified Chinese translators with expertise in health communication independently produced forward translations of the instrument (T1 and T2). Then, these were compared and reconciled by a senior bilingual translator with a medical background, who synthesized them into a single version (T3). To verify the linguistic equivalence, T3 was subsequently back-translated into English (T4) by an independent professional bilingual translator who had no prior knowledge of the tool and was not involved in the forward translation process, thereby minimizing potential bias and ensuring semantic fidelity. Finally, to further evaluate translation quality, we applied the linguistic equivalence assessment method recommended by Sperber et al [33]. This method involved comparing the original English version with the back-translated version across 2 dimensions. The first dimension is comparability of language, which refers to the similarity of words, phrases, and sentence structures. The second dimension is similarity of interpretability, which refers to the extent to which different wordings convey equivalent meaning. A linguist with a doctoral degree conducted this assessment using 4-point ordinal scales (1=extremely comparable or similar; 2=comparable or similar; 3=not comparable or similar; 4=not at all comparable or similar). Items rated as 3 or 4 on either dimension were reviewed by an expert

XSL•FO
RenderX

panel consisting of one health education specialist and 2 nurses, all of whom had formal training in health communication and substantial experience in patient-facing health education and counseling. Revisions were made through iterative panel discussions until consensus was reached. When the initial agreement was unattainable, additional rounds of discussion were used to determine the most contextually and semantically appropriate wording. The finalized items were then incorporated into the final C-CCI. The overall process is summarized in Figure 1.

Figure 1. Process of developing the modified simplified Chinese version of Clear Communication Index.

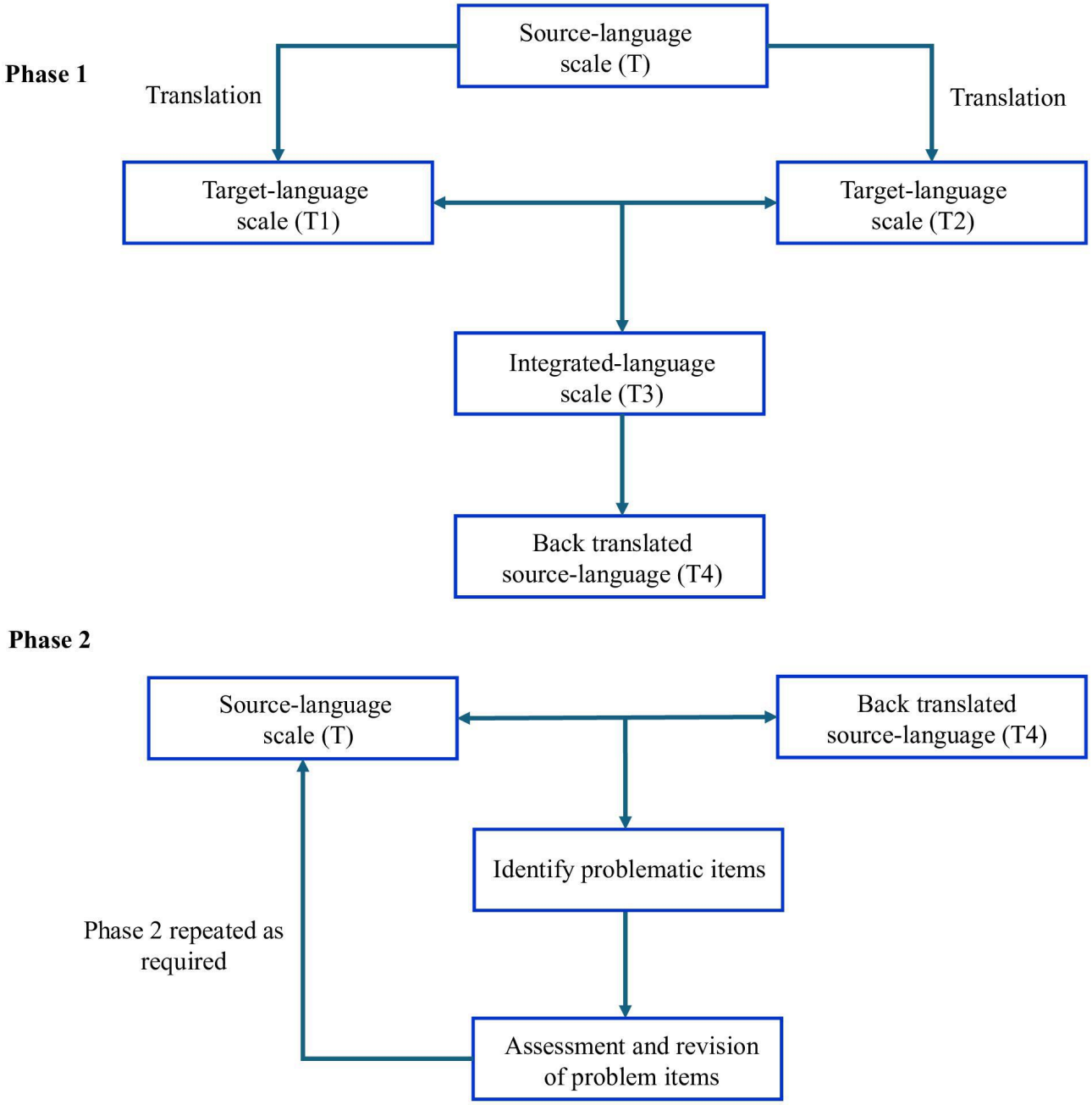

Reliability and Validity Evaluation of the C-CCI

Content Validity

Following Polit and Beck's guidelines for establishing content validity [34], 7 experts were recruited to ensure methodological rigor and representativeness. The expert panel comprised 7 raters (2 male and 5 female), with a mean age of 41 years (SD 5.8). All experts held senior professional titles. Their backgrounds in health education, chronic disease nursing, and nursing education provided a balance of clinical and educational perspectives for evaluating the C-CCI items. The evaluation focused on 2 dimensions: language clarity and content relevance. Experts independently rated each item using a 4-point Likert scale. For clarity, scores ranged from 1=unclear to 4=clear; for relevance, from 1=irrelevant to 4=relevant. Based on these ratings, both the item-level content validity index (I-CVI) and the scale-level content validity index/average (S-CVI/Ave) were calculated. The I-CVI was computed as the proportion of experts rating an item as 3 or 4. The S-CVI/Ave was calculated as the average item-level CVIs, reflecting the overall content validity of the instrument. According to published recommendations,

XSL•FO
RenderX

an S-CVI/Ave above 0.90 is considered acceptable; for panels of ≤5 experts, an I-CVI of 1 is required, whereas for panels of >5 experts, the minimum acceptable I-CVI is 0.78 [34,35].

Exploratory Factor Analysis

An exploratory factor analysis (EFA) was conducted to investigate the latent structure of the C-CCI scale. Prior to factor extraction, the Kaiser–Meyer–Olkin (KMO) test and Bartlett test of sphericity were performed to evaluate the suitability of the data. When the KMO value exceeded 0.60 and the Bartlett test yielded statistically significant results (P<.05), the data were deemed suitable for factor analysis [36]. Factors were extracted using principal component analysis. As prespecified in the analysis plan, the choice of rotation method was determined by the inter-factor correlations: when correlation coefficients exceeded 0.30, oblique rotation was applied; otherwise, orthogonal rotation was used [37]. Items were retained according to the following criteria: factor loading below 0.40, cross-loadings on multiple factors with a difference of less than 0.25. Items meeting any of these conditions were flagged for potential deletion [38].

Confirmatory Factor Analysis

Based on the factor structure derived from the EFA, a confirmatory factor analysis (CFA) was conducted to assess model fit. The assessment was performed at 2 levels: preliminary fit and overall fit indices. Preliminary criteria required all error variances to be positive, t-values of parameter estimates to reach statistical significance (P<.05), standard errors to be small, and factor loadings to fall within the range of 0.50 - 0.95.

Overall model fit was primarily evaluated using the root mean square error of approximation (RMSEA) and the comparative fit index (CFI), which are widely recommended as core indices for assessing model adequacy [39,40]. RMSEA values below 0.08 were considered indicative of acceptable fit, and values of CFI≥0.90 reflected good model fit. As complementary indices, the standardized root mean square residual (<0.08), the incremental fit index and Tucker–Lewis index (≥0.90), as well as the goodness-of-fit index (≥0.90), were also examined to provide a more comprehensive evaluation of model performance [39-41]. In addition, chi-square to degrees of freedom ratio (χ^2/df)<3 was considered indicative of good fit, whereas values between 3 and 5 were regarded as acceptable [36].

Convergent and Discriminant Validity

Composite reliability (CR) and average variance extracted (AVE) were calculated to evaluate convergent validity [42]. Convergent validity was considered strong when CR exceeded 0.70 and AVE was greater than 0.50, and acceptable when AVE ranged between 0.36 and 0.50 [43]. Discriminant validity was assessed by comparing the square root of the AVE for each construct with the inter-construct correlations. When the square root of AVE for a construct was greater than its correlations with other constructs, discriminant validity was deemed satisfactory [42].

Reliability

To assess the reliability of the C-CCI, Fleiss kappa (κ) test was applied to evaluate inter-rater agreement, and Cronbach α coefficient was computed to examine internal consistency. A κ value between 0.61 and 0.80 was interpreted as indicating substantial agreement [44], whereas a Cronbach α coefficient greater than 0.70 was deemed indicative of acceptable internal consistency [45].

Data Collection

A comprehensive search of the official websites of all 31 provincial CDCs in Mainland China was conducted, focusing on diabetes as the evaluation theme given its global rise, major public health burden, and reliance on health education for prevention and management [46,47]. The search was completed on May 18, 2025.

The keyword “diabetes” was used in the internal search engines of the provincial CDC websites to ensure accuracy and consistency. Consistent with the “least effort” principle, whereby the public typically focuses on top-ranked results [48], the highest-ranked diabetes-related health article from each provincial CDC was selected for analysis. Exclusion criteria included government reports, bidding information, and policy or regulatory documents exceeding 1000 words. The Uniform Resource Locators of the selected articles were compiled into an Excel database for systematic evaluation. The Jilin Provincial CDC website did not provide relevant content and was therefore excluded. In total, 30 provincial-level CDC websites were included, covering regions with diverse economic development levels and cultural backgrounds, thereby ensuring a nationally representative sample.

Nine graduate nursing students and 3 senior nurses with more than 5 years of endocrinology nursing experience were invited as evaluators. Prior to the formal evaluation, all evaluators completed a 3-week systematic training that covered the theoretical foundation of the tool, scoring criteria, and application methods. Training included multiple discussions and simulated scoring sessions to enhance the consistency and understanding of the tool’s dimensions. After the training, twelve raters independently evaluated health education materials collected from 30 provinces, producing a total of 360 data entries. To ensure stable and reliable results, factor analysis requires an adequate sample size. Recommendations vary, with common rules of thumb suggesting a minimum of 100 to 200 participants, or a subject-to-item ratio of at least 5:1 or 10:1 [49]. CFA, in particular, requires larger samples to achieve stable parameter estimates and sufficient statistical power [50]. Simulation-based studies using Monte Carlo methods indicate that sample size requirements for CFA depend more on model complexity, factor loadings, and indicator structure than on fixed numerical thresholds [50]. Under typical psychometric conditions, about 200 participants generally provide sufficient power for models of moderate complexity [50]. Accordingly, this study used a 2-step validation process using 2 independent samples: used 2 independent samples: 150 cases for EFA and 210 for CFA, the latter yielding more than 0.80 power for detecting RMSEA deviations at 0.08 [50], ensuring model stability and reliable fit evaluation.

XSL•FO
RenderX

Data Analysis

All statistical analyses were conducted using IBM SPSS Statistics (version 27.0; IBM Corp) and IBM SPSS Amos (version 28.0; IBM Corp). Based on the study design, consistency tests, reliability analyses, EFA, and CFA were performed to assess the scale's reliability and validity. All analyses were 2-tailed, and a *P* value <.05 was considered statistically significant.

Ethical Considerations

This study was approved by the Ethics Committee of Hangzhou Normal University (Approval No. 2025063). All participants provided informed consent before participating. All rater data were anonymized and deidentified prior to analysis, and the evaluated health education materials were obtained from publicly accessible sources. All evaluation data were coded, securely stored, and used solely for research purposes. The study was conducted in accordance with relevant institutional guidelines and the Declaration of Helsinki. All participants involved in this study were provided a small token of appreciation in the form of a gift worth approximately 30 RMB (US $4.13).

Results

Comparison Between Original and Back-Translation, and Cultural Adaptation

Comparison of the original and back-translated versions showed strong conceptual and semantic equivalence (Table 1). Across all items, the mean score for comparability of language was 1.308 (SD 0.63), and the mean score for SI was 1.077 (SD 0.28), indicating a high level of linguistic and interpretive consistency between the 2 versions. Overall variation across items was minimal, with the exception of item 3. While the phrase "policy change" can be directly translated, the research team cautioned that, within the context of public health in Mainland China, it may be misconstrued as implying radical "reform." To avoid this potential misinterpretation, the phrase was revised to "alignment with national health policies," thereby softening the connotation of drastic change and emphasizing policy continuity and optimization.

XSL•FO
RenderX

Table . This section compares the original and back-translated English versions of the Clear Communication Index Score sheet in terms of language comparability and interpretability similarity.

Original English version	Back-translated English version	Comparability of language	Similarity of interpretability
Does the material contain one main message statement? A main message is the one thing you want to communicate to a person or group that they must remember.	Does the material have a core message? The core message is what you want to convey to individuals or groups and wish them to remember.	2	1
Is the main message at the top, beginning, or front of the material? The main message must be in the first paragraph or section. A section is a block of text between headings.	Is the core message at the top, beginning, or front of the material? It must appear in the first paragraph or section. A paragraph is a block of text between headings.	1	1
Does the material include one or more calls to action for the primary audience? If the material includes a specific behavioral recommendation, a prompt to get more information, a request to share information with someone else, or a broad call for program or policy change, answer yes.	Does the material contain one or more calls to action for the main audience? If the material includes any of the following—for example, specific behavioral recommendations, guidance on how to access authoritative health information, encouragement to share the information with others, or a call to action aligned with national health policy directives.	3	1
Do both the main message and the call to action use the active voice?	Are both the core message and call to action in active voice?	1	1
Does the material always use words the primary audience uses?	Is the wording always appropriate to the main audience's habits?	2	2
Is the most important information the primary audience needs summarized in the first paragraph or section?	Is the most important information for the target audience outlined in the first paragraph or section?	1	1
Does the material include one or more behavioral recommendations for the primary audience?	Does the material offer one or more behavior recommendations for the target audience?	1	1
Does the material explain why the behavioral recommendation(s) is important to the primary audience?	Does it explain the importance of the recommended actions to the main audience?	1	1
Does the material always present numbers the primary audience uses?	Are numbers in the material always presented in a way that's familiar to the target audience?	1	1
Does the audience have to conduct mathematical calculations?	Does the audience need to do math?	1	1
Does the material explain the nature of the risk?	Does the material explain the nature of the risk?	1	1
Does the material address both the risks and benefits of the recommended behaviors? This includes actual risks and benefits and those perceived by your audience.	Does it describe the precautions and benefits of the recommended actions? This includes actual and perceived risks and benefits.	1	1
If the material uses numeric probability to describe risk, is the probability also explained with words or a visual?	If the material uses numbers (eg, 1/5 or 20%) to describe risks and also explains these probabilities with words or illustrations, select "yes".	1	1

To further enhance comprehensibility and applicability in the Mainland simplified Chinese context, culturally specific expressions were localized while maintaining consistency with the original scoring framework. For example, in the “NOTE” section of the scale, “Facebook posts,” “Twitter messages,” “scripts for podcasts” and “call center responses” were replaced with “WeChat official account articles,” “Weibo posts” “short video scripts” and “customer service responses” respectively. These adaptations enhanced cultural relevance and user understanding.

XSL•FO
RenderX

Validity and Reliability of the C-CCI

Content Validity

The scale demonstrated high content validity, with overall CVIs of 0.976 for clarity and 0.934 for relevance. Item-level CVIs ranged from 0.857 to 1 for clarity and from 0.124 to 1 for relevance. The notably low relevance score was attributable to Item 4 under the dimension "Main Message and Call to Action," which focused on the use of passive voice. Passive constructions, though acceptable in English health communication, are uncommon in simplified Chinese and may reduce clarity and perceived relevance. This reflects a cross-linguistic mismatch between formal and functional equivalence, consistent with Nida's theory of dynamic equivalence. Accordingly, Item 4 was deemed culturally unsuitable and removed [51]. After its removal, all items achieved a relevance index of 1, yielding a final scale of 12 items (Table 2).

Table . The reliability and content validity of the simplified Chinese version of the Clear Communication Index.

Variables	Cronbach α (N=360)	I-CVI[a] (n=7)	S-CVI[b] (n=7)
Main message and call to action	0.861	—[c]	—
Behavioral recommendations	0.894	—	—
Numbers	0.827	—	—
Risk	0.910	—	—
Total	0.837	0.857~1.000[d] or 1[e]	0.976[d] or 1[e]

[a]I-CVI: item-level content validity index.

[b]S-CVI: scale-level content validity index.

[c]Not applicable.

[d]For language clarity.

[e]For content relevance.

EFA

The KMO value was 0.769, and Bartlett test of sphericity was significant (χ^2_{66}=1034.141; P<.001), indicating the data's suitability for factor analysis. Four factors with eigenvalues greater than 1 were extracted, accounting for 78.7% of the total variance, consistent with the factor structure of the original scale. The scree plot also supported a 4-factor solution, with a clear inflection after the fourth factor (Figure 2).

Figure 2. Scree plot of Eigenvalues for exploratory factor analysis of the simplified Chinese version of the Clear Communication Index.

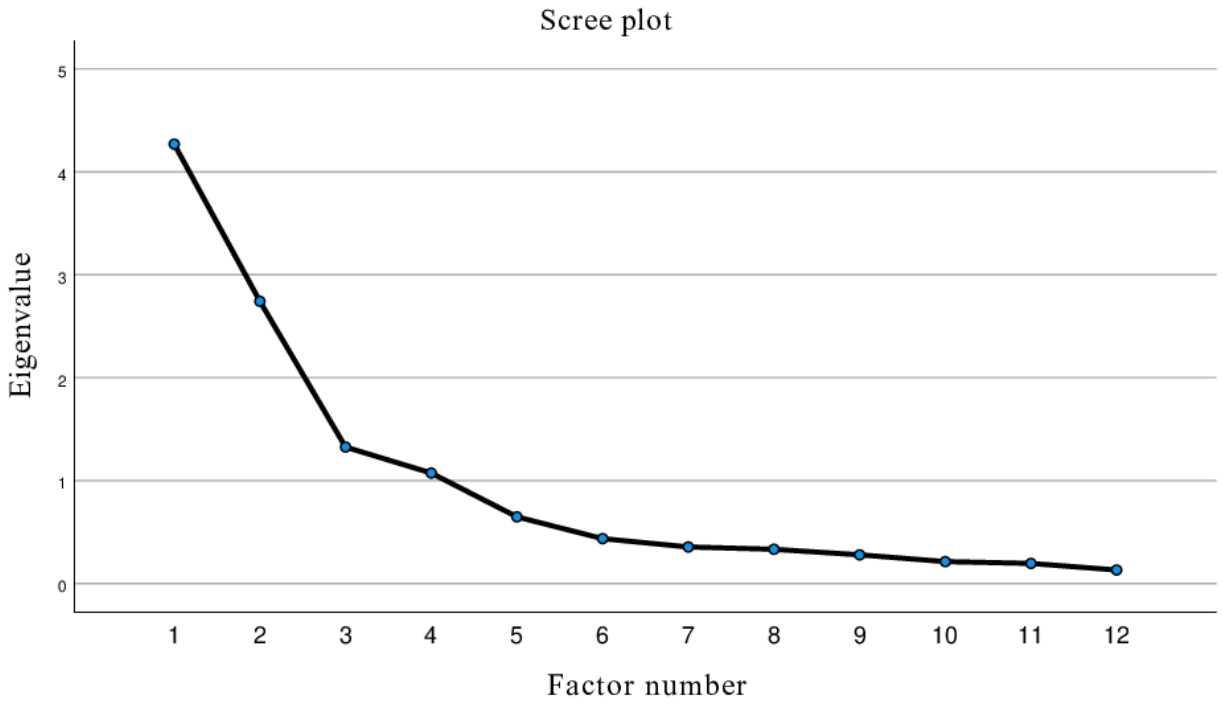

Item loadings (Table 3) were generally high and well defined, and all items showed satisfactory communalities, ranging from 0.475 to 0.903 [52]. Although item 4 had a slightly lower loading (0.484), it exceeded the 0.40 retention threshold and

XSL•FO
RenderX

demonstrated adequate communality and semantic alignment within the "Main Message and Call to Action" dimension; thus, it was retained. Overall, the EFA indicated a clear and stable factor structure with well-defined dimensions, providing strong support for the construct validity of the C-CCI.

Table . Factor loadings of exploratory factor analysis for the simplified Chinese version of the Clear Communication Index.

Items	Factor 1	Factor 2	Factor 3	Factor 4	Communality
A1	0.810	—[a]	—	—	0.711
A2	0.835	—	—	—	0.705
A3	0.863	—	—	—	0.769
A4	0.484	—	0.435	—	0.475
A5	0.878	—	—	—	0.810
B6	—	0.905	—	—	0.888
B7	—	0.922	—	—	0.903
C8	—	—	0.862	—	0.789
C9	—	—	0.890	—	0.830
D10	—	—	—	0.913	0.899
D11	—	—	—	0.819	0.785
D12	—	—	—	0.911	0.847

[a]Loadings < 0.40 were suppressed.

CFA

As shown in Figure 3, the CFA results supported the 4-factor structure of the C-CCI. Model fit indices indicated an excellent fit: $\chi^2_{48}=90.95$ ($\chi^2/df=1.895$), suggesting strong model adequacy; RMSEA=0.065, further supporting the good fit assumption. The CFI value was 0.972, exceeding the conventional cutoff of 0.90 and indicating strong model fit. Other indices also performed well: goodness-of-fit index=0.934, incremental fit index=0.972, Tucker–Lewis index=0.961, and standardized root mean square residual=0.072 (Table 4). All standardized factor loadings were substantial, ranging from 0.70 to 0.95, with each indicator loading strongly on its intended latent construct. No post hoc model respecifications were made based on modification indices, and the final model was retained as theoretically specified. These results demonstrate an excellent overall model fit, with stable and theoretically coherent parameter estimates across the 4-factor structure.

Figure 3. Standardized 4-factor model of the simplified Chinese version of Clear Communication Index.

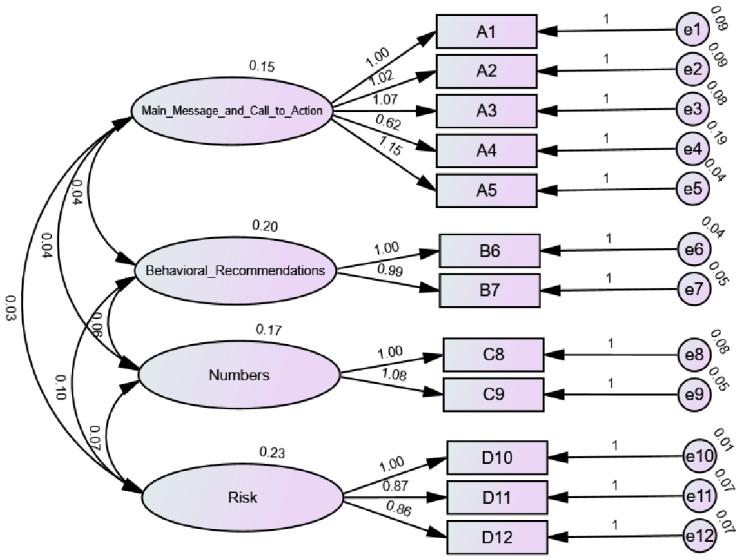

XSL•FO
RenderX

Table . Model fit index for the simplified Chinese version of the Clear Communication Index.

Index	Four-factor	Evaluation standard
χ^2 (*df*)	90.95 (48)	≤3
GFI[a]	0.934	>0.900
CFI[b]	0.972	>0.900
IFI[c]	0.972	>0.900
TLI[d]	0.961	>0.900
RMSEA[e]	0.065	<0.080
SRMR[f]	0.072	<0.080

[a]GFI: goodness-of-fit index.

[b]CFI: comparative fit index.

[c]IFI: incremental fit index.

[d]TLI: Tucker–Lewis index.

[e]RMSEA: root mean square error of approximation.

[f]SRMR: standardized root mean square residual.

Convergent and Discriminant Validity

CR values were 0.907, 0.897, 0.914, and 0.897 for Main Message and Call to Action, Behavioral Recommendations, Numbers, and Risk, respectively. The AVE values were 0.831, 0.814, 0.774, and 0.645, demonstrating strong convergent validity. Discriminant validity was assessed using the Fornell–Larcker criterion [43], comparing the square root of each construct's AVE with its inter-construct correlations (Table 5). In all cases, the square root of AVE exceeded its highest correlation with any other factor. Specifically, the Risk construct exhibited a square root of AVE of 0.803, showed a higher value than its strongest correlation with Behavioral Recommendations (r=0.442), and the same pattern was observed for Main Message and Call to Action (0.911), Behavioral Recommendations (0.902), and Numbers (0.880). Although moderate correlations were observed among some constructs, these findings indicate that the 4 dimensions are empirically distinct while remaining theoretically related.

Table . Convergent validity and discriminant validity of the simplified Chinese version of the Clear Communication Index.

Convergent validity and discriminant validity	Risk	Numbers	Behavioral recommendations	Main message and call to action	AVE[a]	CR[b]
Risk	0.803	—	—	—	0.645	0.897
Numbers	0.371[c]	0.880	—	—	0.774	0.914
Behavioral recommendations	0.442[c]	0.320[c]	0.902	—	0.814	0.897
Main message and call to action	0.152[c]	0.221[c]	0.241[c]	0.911	0.831	0.907

[a]AVE: average variance extracted.

[b]CR: composite reliability.

[c]P<.01.

Reliability

Expert ratings demonstrated substantial consistency (κ=0.624). The scale demonstrated high internal consistency overall (Cronbach α=0.837), with dimension-specific coefficients of 0.861 (Main Message and Call to Action), 0.894 (Behavioral Recommendations), 0.827 (Numbers), and 0.910 (Risk), confirming strong reliability across dimensions.

Evaluation of Diabetes Health Communication by Provincial CDCs in Mainland China

Of the 31 provincial-level administrative divisions in Mainland China, 30 CDC websites were included (with a coverage of 96.8%), ensuring strong national representativeness. Among diabetes-related health education materials, one province (Qinghai) provided video content with text captions, 8 used text–image formats, and 21 (70%) provided text-only content. Provinces were categorized into eastern, central, and western economic regions in Mainland China [53]. Mean C-CCI scores did not differ significantly across regions (East: mean 50.06, SD 29.06; Central: mean 55.26, SD 35.52; and West: mean

XSL•FO
RenderX

56.48, SD 26.32). The national mean score was 53.84 (SD 29.74), which fell well below the recommended threshold of 90, indicating considerable room for improvement in health communication quality. No regional differences were found across dimensions except for Behavioral Recommendations, where scores in the western region were significantly higher than those in the eastern and central regions (P=.002). However, the effect size was small (η^2=0.034), suggesting that regional grouping accounted for only a minor proportion of the variance in Behavioral Recommendations (Table 6).

Table . Comparison of scores across dimensions in Eastern, Central, and Western Regions of Mainland China (N=360).

Dimensions and regions	n (%)	Mean (SD)	F test (df)	P value	LSD[a]
Main message and call to Action			0.152（2）	.86	—[b]
East	132 (36.7)	2.7 (1.91)			
Central	84 (23.3)	2.83 (2.21)			
West	144 (40)	2.82 (1.94)			
Overall	360 (100)	2.78 (1.99)			
Behavioral recommendations[c]			6.271（2）	.002	East＜West, P=.001; Central＜West, P=.008
East	132 (36.7)	0.98 (0.95)			
Central	84 (23.3)	1 (0.96)			
West	144 (40)	1.34 (0.89)			
Overall	360 (100)	1.13 (0.94)			
Numbers			1.399 (2)	.25	—
East	132 (36.7)	1.09 (0.89)			
Central	84 (23.3)	1.3 (0.86)			
West	144 (40)	1.13 (0.96)			
Overall	360 (100)	1.16 (0.91)			
Risk			1.457 (2)	.23	—
East	132 (36.7)	1.23 (1.46)			
Central	84 (23.3)	1.5 (1.41)			
West	144 (40)	1.49 (1.27)			
Overall	360 (100)	1.4 (1.38)			
Total score			1.735 (2)	.18	—
East	132 (36.7)	50.06 (29.06)			
Central	84 (23.3)	55.26 (35.52)			
West	144 (40)	56.48 (26.32)			
Overall	360 (100)	53.84 (29.74)			

[a]LSD: least significant difference.

[b]Not significant ($P \geq .05$)

[c]η^2 represents the effect size, indicating the proportion of total variance explained by regional grouping. For the Behavioral Recommendations dimension, $\eta^2 = 0.034$, indicating a small effect size.

Discussion

Reliability and Validity of the C-CCI Under a Rigorous Sinicization Process

To our knowledge, this study represents the first rigorous and systematic cross-cultural adaptation process to translate, culturally adapt, and psychometrically validate the CCI, resulting in the C-CCI. Raters with substantial professional experience underwent structured training encompassing theoretical foundations, scoring standards, and simulated evaluations. This process enhanced scoring consistency and minimized subjective bias, strengthening the methodological rigor of data collection. The C-CCI retained the 4 original dimensions: Main Message and Call to Action, Behavioral Recommendations, Numbers, and Risk, thereby preserving theoretical coherence, a structure that was further supported by EFA and confirmed through CFA. Targeted revisions improved cultural applicability, including revising item 3 from “policy change” to “alignment with national health policies” and deleting item 4 due to the incompatibility

XSL•FO
RenderX

of passive voice in simplified Chinese. The deletion of items does not alter the scoring rules and, consequently, has no impact on the scoring threshold of the C-CCI. Following these adjustments, expert content validity scores improved, confirming the scientific soundness of the modifications. Psychometrically, the C-CCI demonstrated robust validity and reliability that not only met but exceeded international benchmarks for comparable Health Information Quality Assessment Tools. Prior reviews have shown that most instruments remain at Grade B due to limited structural validation and modest explained variance [30]. In contrast, the C-CCI achieved a cumulative variance of 78.7% in EFA and excellent model fit in CFA (RMSEA=0.065 and CFI=0.972), exceeding Grade B thresholds and reflecting superior structural coherence. Convergent validity was further supported by high CR and AVE values, while strong internal consistency (Cronbach α>.80) and substantial interrater agreement (κ=0.624) confirmed the instrument's stability and reproducibility.

Applicability and Promotion Potential of the C-CCI

Using diabetes health communication materials from 30 provincial CDC websites in Mainland China, including text-only, text-image, and video formats, this study systematically evaluated the real-world applicability of the C-CCI. The tool showed a clear structure and specific scoring dimensions, and raters achieved strong inter-rater reliability after brief training, demonstrating strong practicality and replicability in applied settings. The C-CCI effectively identified weaknesses in health communication materials across 4 dimensions, demonstrating strong problem detection capacity and intervention guidance value. Its stable performance across different regions and formats (text only, text-image, and video) further supports its adaptability to diverse contexts.

Importantly, the C-CCI aligns well with China's evolving digital communication ecosystem, in which health information is increasingly disseminated through official WeChat (Tencent) accounts, TikTok (ByteDance) videos, and other interactive platforms. Although these channels differ in format and engagement mechanisms, they share core demands for message clarity, concise information hierarchy, and actionable behavioral guidance, principles directly reflected in the C-CCI framework. From an implementation perspective, provincial CDCs may use the tool as a formative evaluation instrument during content development and internal review, using it to guide message structuring and action-oriented framing prior to dissemination. This approach enables consistent quality assurance across regions while maintaining flexibility in adapting to platform-specific communication styles.

Beyond the Mainland China context, the C-CCI's methodological foundation may also hold relevance for other Chinese-speaking regions that share similar linguistic and readability challenges, mirroring Mainland China communication patterns noted in previous studies [54-58]. However, its broader application should be guided by further adaptation studies that consider local policy environments and platform ecosystems, ensuring evidence-based contextualization rather than direct transfer.

Comparison With International Adaptations and Added Value of This Study

The CCI has been adapted into multiple languages and contexts worldwide. For instance, the Brazilian Portuguese version demonstrated applicability among primary health care professionals through field validation [28], while the Thai version emphasized dissemination of the theoretical framework and facilitated adoption in local health education [29]. These studies collectively established the feasibility of cross-cultural adaptation but were limited in the scope of psychometric testing and in the range of evaluated materials. Compared with prior validations, this study advances the methodological frontier of CCI research through multiformat empirical testing across 30 provincial CDC websites. By systematically evaluating diabetes-related materials in text-only, text-image, and video formats, the study assessed whether the measurement properties of the C-CCI remained stable across media that differ substantially in information organization. This format inclusivity, supported by consistent reliability across modalities, broadens the empirical use of the CCI framework beyond single-medium applications.

In addition, the study incorporated a more comprehensive psychometric evaluation strategy, combining exploratory and confirmatory factor analyses with rigorous validity and reliability testing. The convergence of large-scale institutional sampling and cross-format validation provides a robust methodological blueprint for future adaptations in diverse health communication settings. This study also contributes a transferable model for standardizing and assessing health information quality across varied media environments.

Quality of Diabetes Health Communication in Mainland China and Improvement Strategies

Using the C-CCI, this study systematically assessed the quality of diabetes health communication in mainland China, identifying overall quality levels, regional variation, and areas for improvement. Median scores across the 4 dimensions were approximately half of the maximum, revealing substantial room for enhancement. The national mean score of 53.84 (SD 29.74) fell below the recommended threshold of 90 [10], underscoring deficiencies in information completeness, structural coherence, and balance across dimensions.

Regional comparisons showed that western provinces scored significantly higher in Behavioral Recommendations than eastern and central provinces. However, the magnitude of this difference was small (η^2=0.034), indicating that regional grouping explained only a limited proportion of the variance in Behavioral Recommendations. This pattern may be partly influenced by unmeasured contextual factors (eg, regional economic capacity, resource distribution, and health communication infrastructure), pending future covariate adjustment and mediation analyses. Given the heterogeneous developmental landscape across and within China's provinces, aggregation into broad east–central–west categories may obscure underlying structural disparities and inflate within-group variance. For example, within the eastern region, provinces, such as Hainan and Hebei, may exhibit comparatively lower development profiles than other eastern provinces, while within

XSL•FO
RenderX

the western region, provinces and municipalities, such as Sichuan and Chongqing, may demonstrate development levels comparable to, or even exceeding, those of some central or eastern provinces. Similar internal disparities may also be present within the central region. These findings underscore the importance of interpreting statistically significant regional differences alongside effect size estimates. Hence, future research should adopt more fine-grained, mechanism-sensitive grouping strategies or multilevel analytical frameworks, incorporating appropriate covariate adjustment and mediation analyses to distinguish structural influences from context-specific practices. Based on standardized percentage scores, Risk emerged as the lowest-performing dimension, followed by Main Message and Call to Action, with Behavioral Recommendations and Numbers showing relatively better performance. Accordingly, improvement efforts should be prioritized in alignment with these empirical deficits: (1) enriching risk communication should be the foremost priority. This could involve contextualizing both absolute and relative risk information in line with evidence-based communication principles, supported by standardized templates and iterative user testing; (2) strengthening the clarity and authority of main messages should follow, through improved headline structuring, message salience testing (eg, digital A/B testing), and inclusion of credible source attribution; (3) improving numerical presentation by using absolute frequencies, consistent denominators, and simplified visual aids can promote better comprehension across literacy levels; and (4) structuring Behavioral Recommendations into concise, sequential action steps with links to relevant resources can reinforce practical applicability. These refinements would elevate both the scientific quality and communicative impact of diabetes health materials, thereby increasing the effectiveness of public health interventions.

Limitations

This study has several limitations. First, the sample was restricted to provincial CDC websites and focused exclusively on diabetes-related materials, without including hospitals, nongovernmental organizations, or social media platforms, which are also important health communication channels. Future research should expand to a broader range of health topics and communication platforms to test the generalizability of the C-CCI. Second, all raters had medical or nursing backgrounds. While this ensured professional rigor, it may have introduced certain biases. Rater training yielded acceptable interrater consistency (κ=0.624), although potential bias from disciplinary homogeneity remains. Future studies could strengthen methodological reflexivity by adopting hybrid rating panels that combine professional experts with lay raters to systematically compare expert-oriented and user-centered assessments. In this approach, lay raters could be purposively sampled to reflect variation in health literacy, education, and regional backgrounds, enabling a more representative appraisal of communication quality under real-world conditions. Such iterative and diversified designs would not only enhance measurement robustness but also model continuous improvement in cross-rater reliability and interpretive balance.

Conclusions

This study completed the cross-cultural translation and psychometric validation of the C-CCI, establishing a reliable and context-sensitive instrument for evaluating the clarity and actionability of health communication in Mainland China. Beyond measurement performance, the findings highlight the C-CCI's potential as a standardized quality assurance tool to support systematic improvement in public health communication, in line with national efforts to advance health literacy and equitable access to health information [59].

By demonstrating stable performance across text, text-image, and video formats in diabetes-related materials from provincial CDC websites, the C-CCI shows strong potential for routine application in CDC practice in chronic disease communication. As exemplified in diabetes education, this framework can be extended to broader chronic disease portfolios and used to support longitudinal monitoring of health communication quality over time. Its integration into content development and review workflows could facilitate evidence-informed standardization of health materials nationwide, helping to reduce avoidable comprehension gaps and strengthen public engagement. Future research should further validate this approach across disease areas and communication contexts, consolidating the role of the C-CCI within China's public health system.

Acknowledgments

The authors sincerely thank all participants who contributed to this study but did not meet the criteria for authorship. We are also grateful to the US Centers for Disease Control and Prevention for granting permission to translate and use the Clear Communication Index for research purposes.

Funding

This research received funding from the General Programs of Humanities and Social Sciences,Ministry of Education of China (Grant No. 23YJAZH136).

Data Availability

The data are available from the corresponding author, upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

LL, CY, and HQH contributed to the conceptualization of the study. Data curation was performed by LL, XQL, HYD, and RG. Formal analysis was conducted by LL and YC. The methodology was developed by YC. The investigation was carried out by LL, XQL, HYD, RG, HQJ, FFC, and YJL. Resources were provided by XQL, HYD, RG, HQJ, FFC, and YJL. Funding acquisition was the responsibility of YGT, who together with QHH, supervised the study. The original draft was written by LL, YC, and YGT. The article was revised by LL, YC, YGT, and QHH. This article has two corresponding authors, Qiuhuan Huang and YGT, with YGT serving as the primary corresponding author.

Conflicts of Interest

None declared.

References

1. National Academies of Sciences, Engineering, and Medicine. Effective Health Communication Within the Current Information Environment and the Role of the Federal Government: Proceedings of a Workshop: Washington, DC: The National Academies Press; 2023. [doi: 10.17226/27210] [Medline: 37878739]
2. Fu Y, Han P, Wang J, Shahzad F. Digital pathways to healthcare: a systematic review for unveiling the trends and insights in online health information-seeking behavior. Front Public Health 2025;13:1497025. [doi: 10.3389/fpubh.2025.1497025] [Medline: 40013047]
3. Brach C. AHRQ Health Literacy Universal Precautions Toolkit, 3rd edition: Rockville, MD: Agency for Healthcare Research and Quality; 2023.
4. National Health Commission of the People's Republic of China. Monitoring report on Chinese residents' health literacy 2024. URL: https://www.nhc.gov.cn/xcs/c100122/202501/18ecbeb9c42942bea9e0fced7a963299.shtml [accessed 2025-06-21]
5. The HLS19 Consortium of the WHO Action Network M-POHL. International report on the methodology, results, and recommendations of the European Health Literacy Population Survey 2019–2021 (HLS19) of M-POHL.: Vienna: Austrian National Public Health Institute; 2021 URL: https://m-pohl.net/Results [accessed 2025-08-25]
6. Nutbeam D, Lloyd JE. Understanding and responding to health literacy as a social determinant of health. Annu Rev Public Health 2021 Apr 1;42:159-173. [doi: 10.1146/annurev-publhealth-090419-102529] [Medline: 33035427]
7. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med 2011 Jul 19;155(2):97-107. [doi: 10.7326/0003-4819-155-2-201107190-00005] [Medline: 21768583]
8. Doak CC, Doak LG, Root JH. Teaching Patients With Low Literacy Skills, 2nd edition: J B Lippincott; 1996.
9. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014 Sep;96(3):395-403. [doi: 10.1016/j.pec.2014.05.027] [Medline: 24973195]
10. Baur C, Prue C. The CDC Clear Communication Index is a new evidence-based tool to prepare and review health information. Health Promot Pract 2014 Sep;15(5):629-637. [doi: 10.1177/1524839914538969] [Medline: 24951489]
11. Klompstra L, Liljeroos M, Lundgren J, Ingadottir B. A Clinical Tool (CUE-tool) for health care professionals to assess the usability and quality of the content of medical information websites: electronic delphi study. J Med Internet Res 2021 Feb 17;23(2):e22668. [doi: 10.2196/22668] [Medline: 33595439]
12. Zubiena L, Lewin O, Coleman R, et al. Development and testing of the health information website evaluation tool on neck pain websites - an analysis of reliability, validity, and utility. Patient Educ Couns 2023 Aug;113:107762. [doi: 10.1016/j.pec.2023.107762] [Medline: 37087877]
13. Lee N, Oh SW, Cho B, Myung SK, Hwang SS, Yoon GH. A Health Information Quality Assessment Tool for Korean online newspaper articles: development study. J Med Internet Res 2021 Jul 29;23(7):e24436. [doi: 10.2196/24436] [Medline: 34326038]
14. Robillard JM, Jun JH, Lai JA, Feng TL. The QUEST for quality online health information: validation of a short quantitative tool. BMC Med Inform Decis Mak 2018 Oct 19;18(1):87. [doi: 10.1186/s12911-018-0668-9] [Medline: 30340488]
15. Shan Y, Ji M, Xing Z, Dong Z, Wang D, Cao X. The simplified Chinese version of the suitability assessment of materials for the evaluation of health-related information for adults: translation and validation study. JMIR Form Res 2023 Jul 7;7(1):e41609. [doi: 10.2196/41609] [Medline: 37418296]
16. Shan Y, Ji M, Dong Z, Xing Z, Wang D, Cao X. The Chinese version of the Patient Education Materials Assessment Tool for printable materials: translation, adaptation, and validation study. J Med Internet Res 2023 May 18;25:e39808. [doi: 10.2196/39808] [Medline: 37200085]
17. Centers for Disease Control and Prevention (CDC). CDC clear communication index score sheet. 2014. URL: https://www.cdc.gov/ccindex/ [accessed 2025-08-25]
18. Centers for Disease Control and Prevention (CDC). Modified CDC clear communication index score sheet. 2014. URL: https://www.cdc.gov/ccindex/ [accessed 2025-08-25]

XSL•FO
RenderX

19. McClure E, Ng J, Vitzthum K, Rudd R. A mismatch between patient education materials about sickle cell disease and the literacy level of their intended audience. Prev Chronic Dis 2016 May 12;13:E64. [doi: 10.5888/pcd13.150478] [Medline: 27172259]
20. Tran BNN, Ruan QZ, Epstein S, Ricci JA, Rudd RE, Lee BT. Literacy analysis of national comprehensive cancer network patient guidelines for the most common malignancies in the United States. Cancer 2018 Feb 15;124(4):769-774. [doi: 10.1002/cncr.31113] [Medline: 29178322]
21. Modiri O, Guha D, Alotaibi NM, Ibrahim GM, Lipsman N, Fallah A. Readability and quality of wikipedia pages on neurosurgical topics. Clin Neurol Neurosurg 2018 Mar;166:66-70. [doi: 10.1016/j.clineuro.2018.01.021] [Medline: 29408776]
22. Phetxumphou K, Roy S, Davy BM, Estabrooks PA, You W, Dietrich AM. Assessing clarity of message communication for mandated USEPA drinking water quality reports. J Water Health 2016 Apr;14(2):223-235. [doi: 10.2166/wh.2015.134] [Medline: 27105408]
23. Alpert JM, Desens L, Krist AH, Aycock RA, Kreps GL. Measuring health literacy levels of a patient portal using the CDC's Clear Communication Index. Health Promot Pract 2017 Jan;18(1):140-149. [doi: 10.1177/1524839916643703] [Medline: 27188894]
24. Saeed F, Anderson I. Evaluating the quality and readability of internet information on meningiomas. World Neurosurg 2017 Jan;97:312-316. [doi: 10.1016/j.wneu.2016.10.001] [Medline: 27742505]
25. Porter KJ, Alexander R, Perzynski KM, Kruzliakova N, Zoellner JM. Using the clear communication index to improve materials for a behavioral intervention. Health Commun 2019 Jun;34(7):782-788. [doi: 10.1080/10410236.2018.1436383] [Medline: 29419320]
26. Etienne J, Chirico S, Gunabalasingham T, Jarvis A, ICF. Final report: clear communications and uncertainty. EFS3 2018;15(5):1412E. [doi: 10.2903/sp.efsa.2018.EN-1412]
27. Suka M, Yamauchi T, Yanagisawa H. Comparing responses to differently framed and formatted persuasive messages to encourage help-seeking for depression in Japanese adults: a cross-sectional study with 2-month follow-up. BMJ Open 2018 Nov 12;8(11):e020823. [doi: 10.1136/bmjopen-2017-020823] [Medline: 30420341]
28. Marinho A, Baur C, Ferreira FM, Borges-Oliveira AC, Abreu MD. Cross-cultural adaptation of the Clear Communication Index to Brazilian Portuguese. Rev Saude Publica 2020;54:26. [doi: 10.11606/s1518-8787.2020054001561] [Medline: 32187313]
29. Fakkheaw L, Suwanwaree J, Moungngern Y, Thintip K, Mirattanaphrai S. The CDC Clear Communication Index Health Education Materials Assessment Tool. SMB 2021;14(3):41-47 [FREE Full text] [doi: 10.33192/smb.v14i3.251326]
30. Li Y, Ouyang H, Lin G, Peng Y, Yao J, Chen Y. Evaluation of the measurement properties of online health information quality assessment tools: a systematic review. Int J Nurs Sci 2025 Mar;12(2):130-136. [doi: 10.1016/j.ijnss.2025.02.015] [Medline: 40241869]
31. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine (Phila Pa 1986) 2000 Dec;25(24):3186-3191. [doi: 10.1097/00007632-200012150-00014]
32. Chinese Center for Disease Control and Prevention. China CDC. URL: https://en.chinacdc.cn/about/overview/ [accessed 2026-01-01]
33. Sperber AD, Devellis RF, Boehlecke B. Cross-cultural translation: methodology and validation. J Cross-Cult Psychol 1994;25(4):501-524. [doi: 10.1177/0022022194254006]
34. Polit DF, Beck CT, Owen SV. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Res Nurs Health 2007 Aug;30(4):459-467. [doi: 10.1002/nur.20199] [Medline: 17654487]
35. Polit DF, Beck CT. The content validity index: are you sure you know what's being reported? Critique and recommendations. Res Nurs Health 2006 Oct;29(5):489-497. [doi: 10.1002/nur.20147] [Medline: 16977646]
36. Kaiser HF. An index of factorial simplicity. Psychometrika 1974 Mar;39(1):31-36. [doi: 10.1007/BF02291575]
37. Costello AB, Osborne J. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. Pract Assess Res Eval 2005;10(1). [doi: 10.7275/jyj1-4868]
38. Worthington RL, Whittaker TA. Scale development research: a content analysis and recommendations for best practices. Couns Psychol 2006;34(6):806-838. [doi: 10.1177/0011000006288127]
39. Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. Evaluating the use of exploratory factor analysis in psychological research. Psychol Methods 1999;4(3):272-299. [doi: 10.1037/1082-989X.4.3.272]
40. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modeling 1999 Jan;6(1):1-55. [doi: 10.1080/10705519909540118]
41. Jöreskog KG, Sörbom D. LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language: Chicago, IL: Scientific Software International; 1993.
42. Sahoo M. Structural equation modeling: threshold criteria for assessing model fit. In: Subudhi RN, Mishra S, editors. Methodological Issues in Management Research: Advances, Challenges, and the Way Ahead: Emerald Publishing Limited, Vol. 2019:269-276. [doi: 10.1108/978-1-78973-973-220191016]
43. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. Journal of Marketing Research 1981 Feb;18(1):39. [doi: 10.2307/3151312]

XSL•FO
RenderX

44. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]
45. DeVellis RF, Thorpe CT. Scale Development: Theory and Applications, 5th edition: SAGE Publications, Inc; 2022.
46. Hossain MJ, Al-Mamun M, Islam MR. Diabetes mellitus, the fastest growing global public health concern: early detection should be focused. Health Sci Rep 2024 Mar;7(3):e2004. [doi: 10.1002/hsr2.2004] [Medline: 38524769]
47. Butayeva J, Ratan ZA, Downie S, Hosseinzadeh H. The impact of health literacy interventions on glycemic control and self-management outcomes among type 2 diabetes mellitus: a systematic review. J Diabetes 2023 Sep;15(9):724-735. [doi: 10.1111/1753-0407.13436] [Medline: 37407516]
48. Pang PCI, Chang S, Verspoor K, Pearce J. Designing health websites based on users' web-based information-seeking behaviors: a mixed-method observational study. J Med Internet Res 2016 Jun 6;18(6):e145. [doi: 10.2196/jmir.5661] [Medline: 27267955]
49. Worthington RL. Scale development research: a content analysis and recommendations for best practices. Couns Psychol 2006;34(6):806-838. [doi: 10.1177/0011000006288127]
50. Wolf EJ, Harrington KM, Clark SL, Miller MW. Sample size requirements for structural equation models: an evaluation of power, bias, and solution propriety. Educ Psychol Meas 2013 Dec;76(6):913-934. [doi: 10.1177/0013164413495237] [Medline: 25705052]
51. Nida EA. Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating: EJ Brill; 1964.
52. Field A. Discovering Statistics Using IBM SPSS Statistics, 6th edition: Sage Publications; 2024.
53. Liu W, Liu Y, Twum P, Li S. National equity of health resource allocation in China: data from 2009 to 2013. Int J Equity Health 2016 Apr 19;15(1):68. [doi: 10.1186/s12939-016-0357-1] [Medline: 27093960]
54. Jin K, Ding D, Gullick J, Koo F, Neubeck L. A Chinese immigrant paradox? Low coronary heart disease incidence but higher short-term mortality in western-dwelling Chinese immigrants: a systematic review and meta-analysis. J Am Heart Assoc 2015 Dec 18;4(12):e002568. [doi: 10.1161/JAHA.115.002568] [Medline: 26683217]
55. Kuan AS, Chen TJ, Lee WC. Barriers to health care services in migrants and potential strategies to improve accessibility: a qualitative analysis. J Chin Med Assoc 2020 Jan;83(1):95-101. [doi: 10.1097/JCMA.0000000000000224] [Medline: 31714441]
56. Zhu C, Song B, Walsh CA, Ekoh PC, Qiao X, Xu A. Barriers to accessing health care of older Chinese immigrants in Canada: a scoping review. Front Public Health 2024;12:1445964. [doi: 10.3389/fpubh.2024.1445964] [Medline: 39564368]
57. Zhang L, Ding D. Coronary heart disease prevention among Chinese immigrants: the importance of health literacy. Health Literacy and Communication Open 2025 Dec 31;3(1). [doi: 10.1080/28355245.2025.2476134]
58. Zhang L, Ding D, Neubeck L, Gallagher R. Health literacy as a predictor of emergency department visits and self-rated health among Chinese immigrants: findings from an Australian survey. Patient Educ Couns 2020 Nov;103(11):2353-2360. [doi: 10.1016/j.pec.2020.04.017] [Medline: 32456982]
59. State Council of the People's Republic of China. Healthy China Action (2019–2030). URL: https://www.gov.cn/xinwen/2019-07/15/content_5409694.htm [accessed 2025-09-22]

Abbreviations

AVE: average variance extracted
C-CCI: Chinese version of Clear Communication Index
CCI: Clear Communication Index
CDC: Centers for Disease Control and Prevention
CFA: confirmatory factor analysis
CFI: comparative fit index
CR: composite reliability
EFA: exploratory factor analysis
I-CVI: Item-level Content Validity Index
KMO: Kaiser–Meyer–Olkin
PEMAT: Patient Education Materials Assessment Tool
RMSEA: root mean square error of approximation
S-CVI/Ave: Scale-level Content Validity Index/Average
SAM: Suitability Assessment of Materials

XSL•FO
RenderX

Edited by A Schwartz, M Balcarras; submitted 15.Sep.2025; peer-reviewed by C Ahmadu, R Nooripour; accepted 23.Jan.2026; published 20.Feb.2026.

Please cite as:
Liu L, Chen Y, Li X, Du H, Gui R, Jin H, Chen F, Lin Y, Huang Q, Tong Y
Validation of the Simplified Chinese Clear Communication Index Using Diabetes Education Materials: Instrument Adaptation and Validation Study
J Med Internet Res 2026;28:e83935
URL: https://www.jmir.org/2026/1/e83935
doi:10.2196/83935

© Lu Liu, Yi Chen, Xiaoqi Li, Hanyu Du, Rong Gui, Huiqing Jin, Fengfei Chen, Yujun Lin, Qiuhuan Huang, Yingge Tong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Digital Tool for Assessing the Distinct Effects of Depression, Anxiety, and Attention-Deficit/Hyperactivity Disorder (ADHD) on Children's Emotional Cognitive Bias: Cross-Sectional Study

Sang-Eon Park[1,2], PhD; Jisu Chung[1]; Sang Ah Lee[1,3], PhD

[1]Department of Brain and Cognitive Sciences, Seoul National University, Seoul, Republic of Korea

[2]Brain Imaging Center, Seoul National University, Seoul, Republic of Korea

[3]Institute for Data Innovation in Science, Seoul National University, Seoul, Republic of Korea

Corresponding Author:
Sang Ah Lee, PhD
Department of Brain and Cognitive Sciences
Seoul National University
Gwanakro 1
Seoul,
Republic of Korea
Phone: 82 28802863
Email: sangahlee@snu.ac.kr

Abstract

Background: Emotional wellness and healthy neurocognitive development are crucial from early childhood. An imbalance in attentional and emotional regulation system is associated with an increased risk of depression, anxiety, and attention-deficit/hyperactivity disorder (ADHD). Early assessment of these risks is essential, but it is difficult to conduct cognitive tests that are both child-friendly and able to dissociate different behavioral biases.

Objective: This study aimed to develop a digital app–based tool designed for young children to objectively assess cognition-affect interactions and examine the association with standardized scales for anxiety, depression, and ADHD.

Methods: In this cross-sectional study, 78 healthy children (36 female) aged 4-10 (mean 7.2, SD 1.4) years with no history of mental illness were recruited from the local community center and children's museum. Emotional regulation and attentional control were assessed using an animated emotional Flanker task, emotional Stroop task, and emotional Go/No-Go task on a touchscreen computer. Children's current mental health was measured using self-reported depression and anxiety states through the Center for Epidemiological Studies Depression Scale for Children (CES-DC) and State-Trait Anxiety Inventory for Children (STAI-CH), while the ADHD risk was assessed using the Korean ADHD Rating Scale (K-ARS) for parents. Principal component analysis was applied to behavioral measures across the tasks to group them by similarities and extract 3 abstract scores ("E-scores") representing different aspects of cognitive function (attention, selective inhibition, and emotional sensitivity). Associations between E-scores and mental health or ADHD risk were then tested.

Results: There was a significant improvement in general attention across development (Pearson correlation between E-score 1 and age: r=–0.75; P<.001; 2-sided α=.05) but not emotion-attention interactions. Performance was also correlated with mental health scales. First, children with higher depression symptoms (ie, higher CES-DC) were slower in their responses in general (ie, higher E-score 1; Pearson correlation after controlling for age: r=0.29; P=.04). Second, both anxious and depressed (ie, higher STAI-CH and CES-DC) children demonstrated reduced attention selectively to the emotional stimuli as indicated by elongated RT and lower accuracy (ie, higher E-score 2; anxiety: r=0.34; P=.02; depression: r=0.51; P<.001). Lastly, children with higher ADHD scales (ie, higher Korean ADHD Rating Scale [K-ARS]) showed lower accuracy across the three tasks, particularly for emotional stimuli (ie, lower E-score 3; r=–0.32; P=.03).

Conclusions: By combining well-established emotional cognitive tasks with our dimensionality reduction techniques, we extracted individual affective-cognitive characteristics from diverse but noisy behavioral patterns and identified their association with mental health and ADHD-related symptoms in children. These results demonstrate the scientific validity, versatility, and translational potential of our gamified digital assessment tool for monitoring young children's affective and cognitive health in daily life. Future longitudinal studies in children with formal clinical diagnoses will further strengthen the generalizability of these findings.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e86286) doi:10.2196/86286

KEYWORDS

ADHD; attention; attention-deficit/hyperactivity disorder; eHealth; emotion; mental health

Introduction

Early childhood is a crucial phase of cognitive and emotional development that sets the stage for children's well-being throughout life. Disruptions in the healthy processing of emotional stimuli not only raise the risk of mental disorders such as depression, anxiety, and attention-deficit/hyperactivity disorder (ADHD), but also make it challenging for children to deploy and regulate attentional control, causing difficulties in children's learning and school performance [1]. However, because these symptoms often begin at a young age [2], early monitoring and objective identification of individual vulnerability have been challenging based on traditional parent reports and clinical interviews.

Understanding how emotional processing in the brain develops during childhood provides insight into how such interactions between emotional and cognitive function are related to mental health. During childhood, the brain first undergoes subcortical development in emotion-processing regions such as the amygdala [3] and reward-processing regions such as the ventral striatum [4]. Across maturation, these regions form stronger network connectivity to the prefrontal cortex (PFC) [5], which develops gradually for modulating arousal and attention [5,6], ensuring that emotional stimuli are assessed and managed in a context-appropriate manner. Due to the immature regulatory function of the PFC, younger children rely more on subcortical function. This developmental imbalance causes them to exhibit greater sensitivity to emotional stimuli [7,8], which can be detrimental to both emotional wellness and cognitive performance from early childhood [9].

Based on findings from developmental cognitive neuroscience, the effects of depression, anxiety, and ADHD on cognitive abilities such as emotion recognition, attention, and inhibitory control may be dissociable despite their high comorbidity [10]. For instance, while depression is mainly characterized by feelings of sadness, low mood, and loss of interest in their usual activities [11], it is also associated with slower cognitive processing speed [12] and difficulty disengaging from negative emotional information [13]. Anxiety, on the other hand, is more strongly associated with heightened reactivity to potential threat [14,15], leading to an attentional bias toward emotional stimuli and an impairment in one's ability to focus on nonthreatening, neutral stimuli [16]. Finally, although ADHD is marked by deficits in cognitive control, impulse regulation, and hyperactivity in daily life patterns [17], children with ADHD tend to be particularly more impulsive in response to rewarding or positively valenced stimuli [18,19].

Serious games can serve as digital monitoring tools that can be applied in daily life [20] and are particularly effective for young children who may be less receptive to repeated administrations of self-report surveys and cognitive tests. Recent studies have demonstrated that digital tools can be developed by leveraging associations between cognitive profiles and mental health or ADHD symptoms [21]. For example, one study assessed ADHD risk in children aged 4-8 years by examining cognitive functions such as reward processing using computer-based games [22]. Other studies have measured attention and working memory to examine their associations with depression [23] and anxiety [24] symptoms in adults. In addition, some studies did not directly evaluate the association between performance of digital cognitive tasks and mental health scales but instead assessed related cognitive domains, such as executive functions [25] and cognitive control [26]. However, most prior digital task–based studies have focused on single symptom domains, without accounting for the high comorbidity among ADHD, depression, and anxiety, which may not only necessitate separate tests for each but also limit the interpretability of assessments.

Unfortunately, despite the overwhelming evidence that mental health is simultaneously influenced by emotional sensitivity and cognitive processing [27-30], there is still a lack of digital tasks that capture such interactions within the scope of a game-like application that can be widely used by the public. More importantly, prior studies showed critical limitations in distinguishing the impacts of depression, anxiety, and ADHD simultaneously, using a single versatile assessment tool, despite the differences in the neurocognitive mechanisms that underlie their specific symptomatology [31]. The duration and user-friendliness of such tasks are especially critical for testing young children, as multiple lengthy traditional laboratory-based tests often fail to sustain children's interest or use excessively violent or provocative emotional stimuli that are inappropriate for young children [32].

Given the above limitations, we set out to develop and validate a novel digital assessment tool that can characterize the relationships between children's mental health or ADHD survey scores and emotional cognitive performance, an integration that has been largely absent from previous studies. Based on our previous study on a diverse range of both healthy participants and depressed adults [33], we designed a gamified task that is engaging and suitable for young children, which can be administered easily in about 30 minutes. A unique and novel aspect of our approach was to calculate abstract componential measures extracted from a range of behavioral patterns, which allowed us to identify individual cognitive-affective characteristics that were differentially associated with anxiety, depression, and ADHD symptoms. Unlike prior digital task–based tools that focus on isolated symptom domains, our framework enables simultaneous assessment of mental health and ADHD symptoms within a single test.

We developed modified versions of the emotional Flanker task (eFlanker), emotional Go/No-Go task (eGoNoGo), and emotional Stroop task (eStroop) using animated faces on a touchscreen interface. The eFlanker measured children's attentional control in quickly assessing the perceptual stimuli and responding to the target stimulus without being influenced

XSL•FO
RenderX

by nearby distractors. The eGoNoGo measured children's ability to remember and select particular combinations of emotional and perceptual features, while asserting cognitive control to suppress responses to nontarget lures. Additionally, we developed a novel eStroop, requiring simple responses to the target features while inhibiting the irrelevant information. These 3 tasks demonstrated their validity as a psychometric test for children (aged 4-6 years) in previous studies [34-36]. Mental health and ADHD symptoms were assessed using standardized surveys that are widely used in both research and clinical settings.

To ensure the task was usable and relevant for the general public, we assessed children who may be at elevated risk for cognitive-affective difficulties but had not yet been diagnosed with a mental disorder. Recruitment in a community-based context allowed us to obtain a broad and representative dataset spanning healthy to at-risk individuals while minimizing explicit selection biases. We chose a target age range of 4-10 years, which corresponds to early childhood, a developmental period marked by ongoing maturation of prefrontal cortical systems supporting emotional regulation and cognitive control. This period is also when individual affective-cognitive characteristics related to anxiety, depression, and ADHD begin to emerge [37-39]. Given that adolescence is associated with distinct neurobiological mechanisms in processes such as risk-taking and emotion regulation [40], we limited recruitment to children younger than 10 years of age, in accordance with the definition of adolescence provided by the World Health Organization [41].

Methods

Participants

Children aged between 4 and 10 years without any history of mental illness were recruited from a local community center and a local children's museum (Gyeonggi Children's Museum; N=78; 36 female, mean age 7.2, SD 1.4 years). A convenience sampling approach in a public, community-based context was used to recruit children from diverse environments while minimizing explicit selection biases. Data were collected over a 6-month period (August 2023 to February 2024). This schedule was chosen to avoid the annual grade transition in March within the Korean education system, which can represent a substantial environmental change for children. Children who could not use the touchscreen were excluded from the study.

Sample Size Determination

To determine the required sample size for this study, we referred to previous studies reporting associations between attention-related behavioral features and mental health or ADHD scales. In our previous work [33], emotion-related slowing of reaction time (RT) in the Flanker task (relative to the neutral condition) showed Pearson correlations of r=0.342 with depression symptoms and r=0.359 with anxiety symptoms. A power analysis showed that the minimum sample sizes required to achieve 80% statistical power (β=.2) with a 95% confidence level for these correlation coefficients were 64 and 59, respectively. In addition, we referred to another study reporting an association between error rate in the Flanker task and ADHD rating scale scores (Pearson correlation r=0.52) [42], for which the minimum required sample size under the same criteria was 27. To account for potential participant loss and missing data that could reduce statistical power, the target sample size was increased by 20%, resulting in a final sample size of 78 participants.

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki, and all experimental procedures were approved by the institutional review board for human research of Seoul National University (2501/003-021 and 2407/001-014). Written informed consent, which included all general elements required for research involving children, such as the purpose of the study, experimental procedures, potential benefits and discomforts, confidentiality, and the right to withdraw, was obtained from the parent or legal guardian. All study data were deidentified immediately after collection. All participants were compensated with 20,000 KRW (US $14) for 30 minutes of participation.

Emotional Cognitive Tasks

We designed 3 emotional cognitive tasks (eFlanker, eGoNoGo, and eStroop) to measure attention, working memory, and selective inhibition along with emotional processing (Figure 1). All tasks were developed based on the PsychoPy software (version 23.1.2; Open Science Tools Ltd) and implemented on a computer with responses recorded using a touchscreen monitor.

XSL•FO
RenderX

Figure 1. (A) The 3 emotional cognitive tasks used in this study. A diagram with an example screenshot of the emotional cognitive tasks: (B) emotional Flanker task (eFlanker), (C) emotional Go/No-Go task (eGoNoGo), and (D) emotional Stroop task (eStroop).

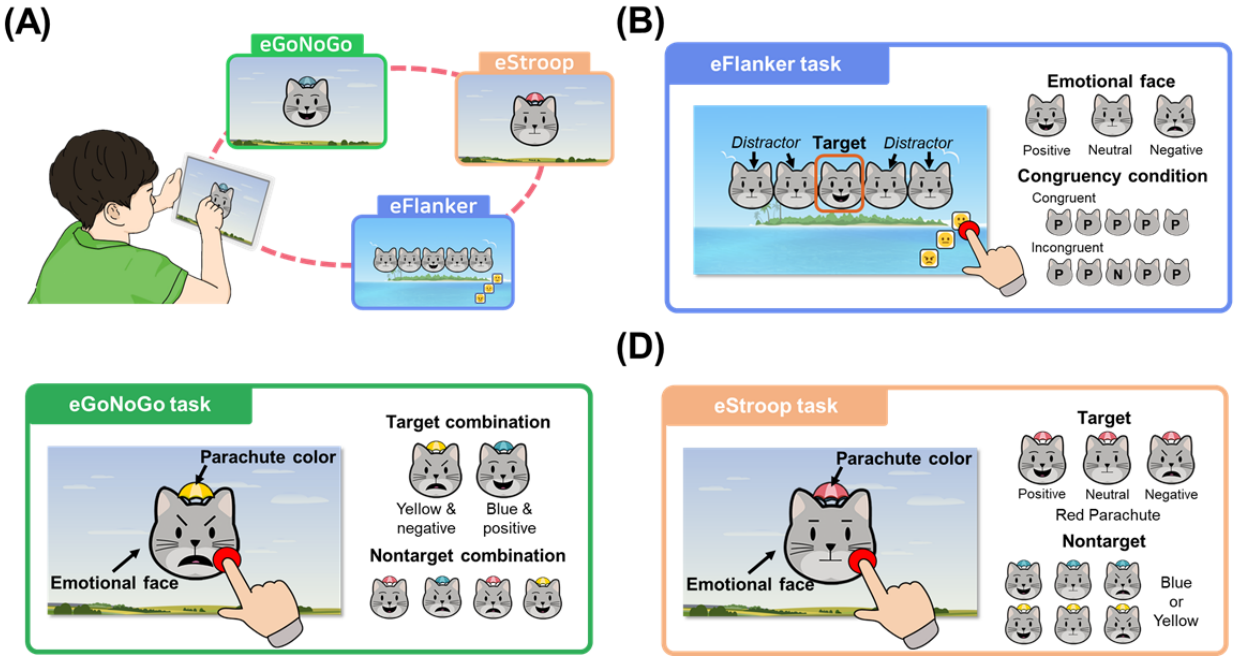

eFlanker

The eFlanker was designed to test for attention and its interactions with emotional stimuli [43,44]. Children were instructed to choose the face (located at the bottom right side of the screen) expressing the same emotion as the target, which was presented in the middle, with 2 distractors on both sides of the wings. The target and distractors were presented after a brief fixation cross (1000 ms), and feedback was provided on the screen once the child responded. The target and distractor consisted of 3 emotional (positive, neutral, and negative) cat faces, and there were a total of 9 possible combinations for their presentation as the target and distractor. Children performed 18 trials of practice blocks to familiarize themselves with the task, and then conducted 2 main blocks with 90 trials each (total 180 trials, 20 trials for each condition), with trial types presented in a random order. A break time was given between the main blocks.

eGoNoGo

The eGoNoGo was designed to measure emotional bias in executive function and memory by testing one's ability to choose the target stimuli while withholding responses to the nontarget objects [45,46]. Out of 6 possible combinations from 3 parachute colors (yellow, red, and blue) and 2 emotions (positive and negative), children were instructed to respond to 2 target combinations of a specific parachute color and a specific emotional cat face (eg, blue parachute with the positive cat face and the yellow parachute with the negative cat face). The color-emotion combination was counterbalanced and randomly assigned across participants. After the fixation cross was presented for 1000 milliseconds, a stimulus was presented in the middle of the screen. The nontarget stimuli were presented for 1500 milliseconds, and the target stimuli were presented until the participants responded. If participants responded to the nontarget (incorrect response) or responded later than 2000 milliseconds, visual feedback on their performance ("Incorrect" or "Respond quickly") was displayed for 1000 milliseconds. Children performed 20 practice trials and then 2 main blocks with 60 trials each (total 120 trials). We adjusted the frequency of the target trials to be slightly higher, to 40% of the total trials, to offset the imbalance between the target (2 out of 6 possible combinations) and nontarget trials (4 out of 6) to keep the children engaged in the task. A break time was given between the main blocks.

eStroop

The eStroop was designed to test inhibitory control by regulating children's response to the nontarget object. Children were asked to touch the face stimuli with a specific parachute color (either red, blue, or yellow) while ignoring the other colors, regardless of the emotion of the face (positive, neutral, and negative). The color-emotion combination was counterbalanced across participants. After a fixation cross was presented for 1000 milliseconds, the nontarget trial was presented for 2500 milliseconds, and the target trial was presented until the participants responded. If participants responded to the nontarget (incorrect response) or responded to the target later than 2000 milliseconds, visual feedback ("Incorrect" or "Respond quickly") was displayed for 1000 milliseconds. Children performed 12 practice trials and then 3 main blocks with 36 trials each (total 108 trials). A break time was given between the main blocks.

Procedures

The experiments were conducted in a quiet room either in the laboratory or in the local children's museum. Both experimental environments were child-friendly and were arranged to be similar to each other. Anxiety and depression levels were measured using the State-Trait Anxiety Inventory for Children (STAI-CH) [47] and Center for Epidemiological Studies Depression Scale for Children (CES-DC) [48] scales, which

XSL•FO
RenderX

were designed specifically for children. In addition, their caregivers completed the ADHD questionnaire using the K-ARS [49] based on *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition] [50]).

Children and their caregivers were provided with informed consent and a description of the overall task. After a brief instruction on the tasks, followed by practice sessions, the children performed the 3 emotional cognitive tasks in a mixed order. During this time, the caregivers completed the ADHD survey (K-ARS) about their children. After the completion of the main tasks, the children were surveyed on questions about their mental health (CES-DC and STAI-CH) either through a written form on an electronic device or through conversation with the experimenter, based on their preference. We did not observe significant differences in CES-DC or STAI-CH scores between the self-completed and experimenter-administered responses (CES-DC: t_{73}=–0.725; *P*=.47; STAI-CH: t_{73}=0.09; *P*=.93; 2-sample *t* test).

Missing Data

Some children missed 4 or more consecutive trials due to distraction or an explicit request to discontinue participation during the task. In total, 14 children met these criteria. Specifically, the number of missing cases was 2 for the eFlanker, 4 for the eGoNoGo, and 10 for the eStroop. These data were excluded from the entire analysis.

Behavioral Indices from the Emotional Cognitive Tasks

Accuracy and RT were calculated from the emotional cognitive tasks to compare behavioral performance across conditions and children. We averaged RTs from the correct trials, removing those shorter than 200 milliseconds to eliminate accidentally pressed responses. Outliers for RT or accuracy in each emotional condition were excluded at the subject level. Outliers were defined as values exceeding 3 times the median absolute deviation from the median, where the median absolute deviation was computed as the median of the absolute deviations from the median.

Extracting E-Scores as an Abstract Measure of Task Performance

Abstract scores were required to summarize a range of behavioral indices in the children's performance on the emotional cognitive tasks. The behavioral indices from 64 children were used for these analyses, after removing 14 children who lost interest during 1 of the 3 tasks (missing data), resulting in RT or accuracy values that were not appropriate for measuring an association with their mental health or ADHD. First, we calculated 82 behavioral indices consisting of RT and accuracy from various conditions of the 3 tasks (eFlanker, eGoNoGo, and eStroop), including averaged RT/accuracy of whole trials or emotion-related RT/accuracy difference between emotional conditions or between congruent and incongruent conditions (Table S1 in Multimedia Appendix 1). Second, to extract the most informative components while minimizing collinearity, 2D latent features were derived by extracting the first and second principal components (PCs) of the *z*-scored behavioral indices [33,51,52]. To avoid having the principal component analysis (PCA) dominated by the overrepresentation of outliers, *z*-scored indices with an absolute value over 2.5 were reduced to 2.5. Only the first two PCs explained more than 10% of the variance (Figure S1 in Multimedia Appendix 1). Then, the k-means clustering method was applied to the 2 PCs for dividing the latent features into several groups based on their squared Euclidean distance [53,54]. The optimal number of groups to maximally separate the latent features was determined by identifying the point at which the rate of decrease in inertia (within-cluster sum of squared errors) slowed down [55] (Figure S1 in Multimedia Appendix 1). Behavioral indices within the same group represented a similar pattern across the participants (eg, RT in the eFlanker was highly correlated with RT in the eStroop). Lastly, we calculated E-scores ranging from 1 to 3 by calculating the first PC of *z*-scored behavioral indices from each of the 3 groups.

Statistical Tests

Behavioral performance across emotional conditions was compared using paired *t* tests with a 2-sided significance level of α=.05. We also performed analysis of covariance (ANCOVA) on task performance measures (RT and accuracy) and E-scores, with gender as a fixed factor and age as a continuous covariate. Pearson partial correlation analyses were conducted to assess associations between E-scores and mental health or ADHD scales while controlling for age.

Results

Depression, Anxiety, and ADHD Scales Across Development

We investigated whether children's anxiety, depression, and ADHD scores were correlated with one another and how they differed across age. First, there was a strong association between depression (CES-DC) and anxiety (STAI-CH; Pearson correlation *r*=0.493; *P*<.001), while the risks for ADHD (K-ARS) were not correlated with either anxiety or depression (K-ARS and CES-DC: *r*=0.068; *P*=.58; K-ARS and STAI-CH: *r*=0.062; *P*=.61). When we conducted an ANCOVA with gender as a fixed factor and age as a covariate, we found higher depression symptoms in younger ($F_{1,72}$=7.044; *P*=.01) and male children ($F_{1,72}$=5.822, male 95% CI 12.8-19, female 95% CI 9.79-13.4; *P*=.02). To avoid the possibility that the CES-DC and STAI-CH self-report scales was too difficult for 1 child aged <5 years, we conducted the same analysis without that child and found that the effect of age on the depression symptoms still showed a strong trend but was not significant ($F_{1,71}$=3.495; *P*=.07), while the gender effect was still significant ($F_{1,71}$=4.68; *P*=.03). Self-reported anxiety and parent-reported ADHD levels were not significantly different across age and gender (anxiety: age effect $F_{1,71}$=0.628; *P*=.43 and gender effect $F_{1,71}$=0.009; *P*=.93; ADHD: age effect $F_{1,70}$=0.719; *P*=.4 and gender effect $F_{1,70}$=0.001; *P*=.99).

Emotion-Related Effects in Emotional Cognitive Task Performance

We found a strong emotion-related effect on children's attention in the eFlanker (Figure 2A). First, RT was significantly higher in both positive (95% CI 1330-1469 ms) and negative (95% CI

XSL•FO
RenderX

1329-1463 ms) conditions compared to the neutral (95% CI 1167-1290 ms) condition (positive vs neutral t_{68}=11.404; P<.001; negative vs neutral: t_{67}=9.821; P<.001). Similarly, accuracy was lower for the 2 emotional conditions (positive: 95% CI 0.961-0.975; negative: 95% CI 0.958-0.972) than the neutral condition (95% CI 0.974-0.986; positive vs neutral: t_{63}=3.414; P=.001; negative vs neutral: t_{65}=5.859; P<.001), indicating that children struggled to suppress the distraction induced by emotional stimuli. However, there were no significant differences between congruent and incongruent conditions in both RT (congruent: 95% CI 1272-1402 ms; incongruent: 95% CI 1284-1414 ms) and accuracy (congruent: 95% CI 0.964-0.976; incongruent: 95% CI 0.963-0.976) on the eFlanker (Figure 2B; RT: t_{68}=0.265; P=.79; accuracy: t_{64}=0.331; P=.74). This lack of a significant congruency effect may be attributed to the dominant effects of emotional face stimuli, as reported in a previous study [56]. In the eGoNoGo, children made significantly more errors in the positive condition (positive: 95% CI 0.898-0.932; negative: 95% CI 0.951-0.97; t_{59}=4.776; P<.001; Figure 2C), which was possibly a result of failure to inhibit a response to the rewarding stimuli. RT was not different between positive and negative conditions (positive: 95% CI 1175-1300 ms; negative: 95% CI 1170-1273 ms). Similarly, in the eStroop, accuracy in the positive condition (95% CI 0.956-0.983) was significantly lower compared to the neutral condition (95% CI 0.975-0.994; t_{67}=2.6; P=.01; Figure 2D), while RT was not different across emotion (positive: 95% CI 793-861 ms; negative: 95% CI 812-886 ms) and neutral conditions (95% CI 800-872 ms).

Figure 2. Behavioral results (reaction time [RT] and accuracy) from all participants in the 3 emotional cognitive tasks. (A) RT (left) and accuracy (right) across 3 emotional conditions (positive [Pos]; neutral [Neu]; and negative [Neg]) in the emotional Flanker task (eFlanker). (B) RT and accuracy in congruent (Cong) and incongruent (Incong) conditions in the eFlanker. (C) RT and accuracy across positive and negative conditions in the emotional Go/No-Go task (eGoNoGo). (D) RT and accuracy across positive, neutral, and negative conditions in the emotional Stroop task (eStroop). Error bars indicate SE. ns: not significant. *P<.05, **P<.01, ***P<.001.

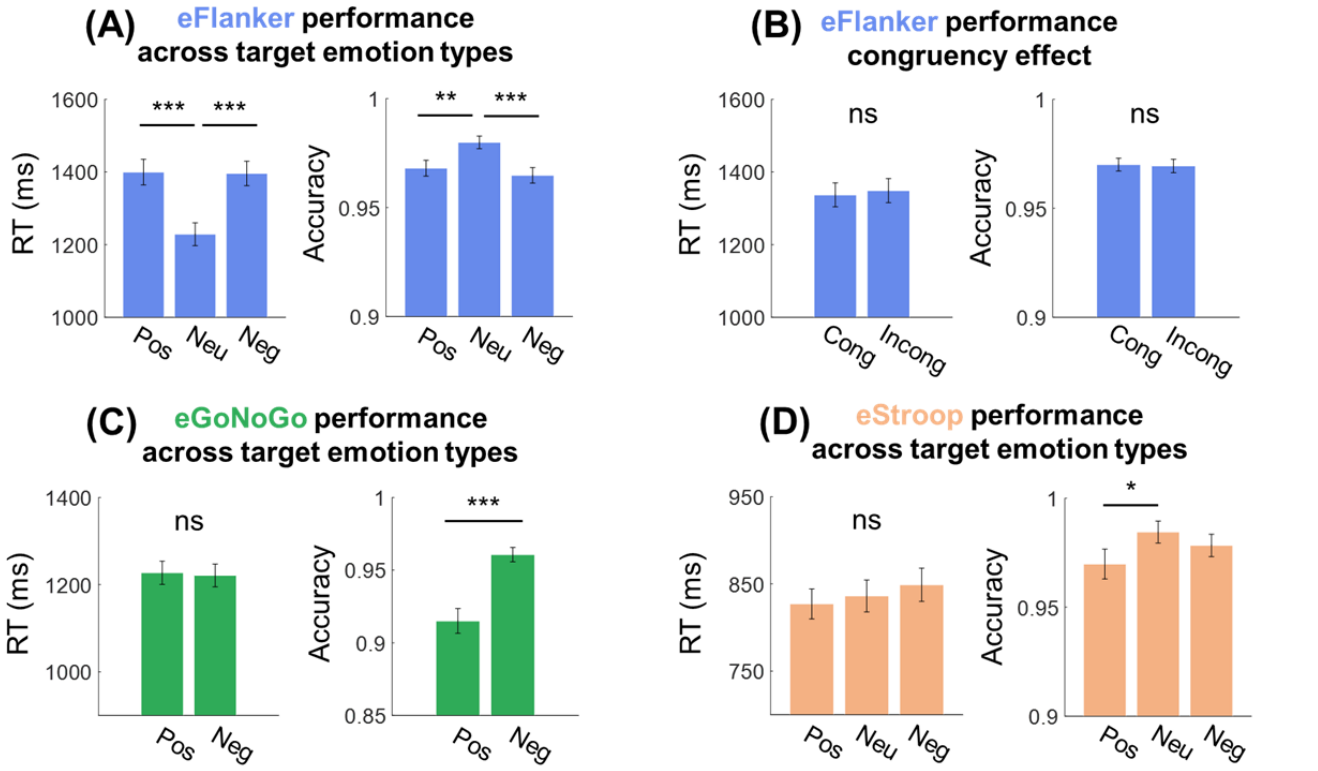

Emotional Cognitive Task Performance Across Development

ANCOVA tests were conducted on general task performance measured by RT and accuracy with gender as a fixed factor and age as a covariate. General task performance in all tasks was strongly correlated with age, confirming an improvement in attention-related functions between age 4 and 10 years (eFlanker RT: $F_{1,66}$=50.817; P<.001; accuracy: $F_{1,66}$=22.349; P<.001; eGoNoGo RT: $F_{1,63}$=19.513; P<.001; accuracy: $F_{1,62}$=8.281; P=.005; eStroop RT: $F_{1,62}$=16.585; P<.001; accuracy: $F_{1,61}$=5.879; P=.02; see Pearson correlation coefficients in Figures 3A-3D). We found a faster RT in male children in the eStroop RT (male: 95% CI 765-843 ms, female: 95% CI 838-949 ms; $F_{1,62}$=4.933; P=.03), but no gender differences were found in the other behavioral measures.

XSL•FO
RenderX

Figure 3. Developmental changes in overall performance, congruency effect, and emotional sensitivity were examined using a cross-sectional dataset. (A-D) Scatter plots showing significant relationships between age and performance in the 3 emotional cognitive tasks. (E-H) The congruency effect in the emotional Flanker task (eFlanker) and emotional sensitivity in reaction time (RT) across all tasks were not correlated with age. Cong: congruent; eGoNoGo: emotional Go/No-Go task; Emo: emotional; eStroop: emotional Stroop task; Incong: incongruent; Neg: negative; Neu: neutral; Pos: positive.

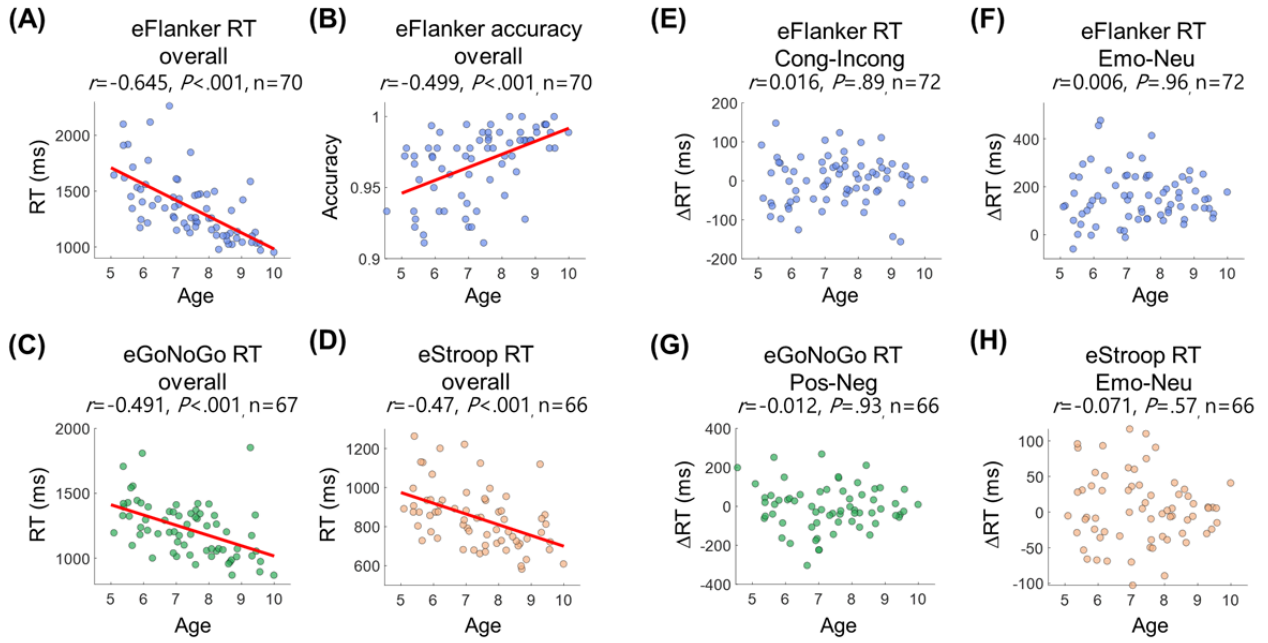

The congruency effect (ie, RT and accuracy differences between congruent and incongruent conditions) or performance difference between emotional and neutral targets in the eFlanker did not change in this age range (Figures 3E and 3F). Additionally, the RT or accuracy difference between positive and negative conditions in the eGoNoGo did not show any changes across age (Figure 3G). Similarly, the difference in eStroop performance (RT and accuracy) between the emotional and neutral conditions was not correlated with age (Figure 3H). None of these behavioral indices showed between-gender differences.

Relationship Between Emotional Cognitive Task Performance and Mental Health Measures

Among the children recruited for this study, 28 (36%) children (CES-DC >15 [57]) were in the high-risk group of depression based on the CES-DC scale (Figure 4A). Depressed mental health was associated with emotional processing during the eFlanker, as RT difference between emotion and neutral target conditions was higher for children with a higher depression scale (Pearson correlation between CES-DC and ΔRT emotional-neutral: r=0.281; P=.02 after controlling for age; Table 1). In addition, the congruency effect on accuracy (ie, a higher accuracy in the congruent condition compared to the incongruent condition) in the eFlanker was reduced in children with higher CES-DC (Figure 4B; r=–0.255; P=.03; n=70).

Figure 4. Relationship between task performance and mental health/attention-deficit/hyperactivity disorder (ADHD) scales. (A) Distribution of children's self-reported depression scales (Center for Epidemiological Studies Depression Scale for Children [CES-DC]). (B) Correlation between the CES-DC scale and the congruency effect on the emotional Flanker task (eFlanker) accuracy after covarying out age. A lower congruency effect was observed in children with a higher CES-DC level. (C) Distribution of children's self-reported anxiety scales (State-Trait Anxiety Inventory for Children [STAI-CH]). (D) Lower eFlanker accuracy in the negative (Neg) condition was associated with a higher STAI-CH level. (E) Distribution of children's ADHD scales, based on parent reports (Korean ADHD Rating Scale [K-ARS]). (F) A lower eFlanker accuracy in the emotional (Emo) condition was observed in children with a higher K-ARS level.

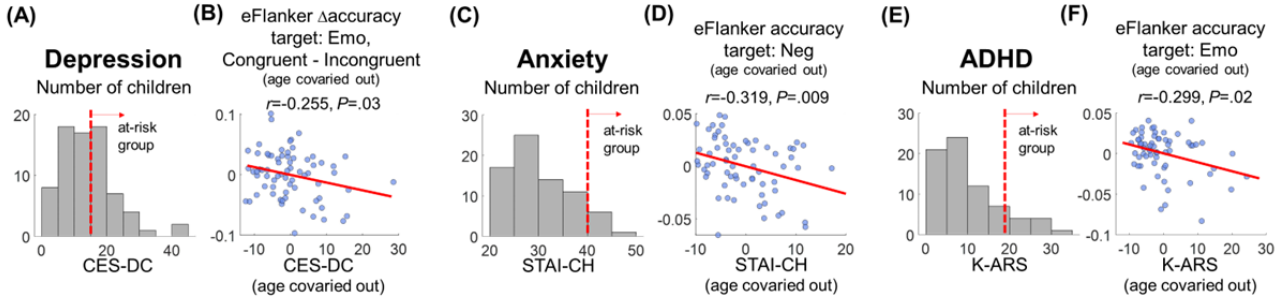

XSL•FO
RenderX

Table 1. Behavioral indices derived from the emotional cognitive tasks, grouped according to principal component analysis (PCA), with coefficients for E-score calculation. Only indices that showed a significant Pearson correlation with mental health measures (Center for Epidemiological Studies Depression Scale for Children [CES-DC] and State-Trait Anxiety Inventory for Children [STAI-CH]) or attention-deficit/hyperactivity disorder symptoms (Korean ADHD Rating Scale [K-ARS]) after covarying out age or overall task reaction time/accuracy on each task are presented. The complete lists are provided in Table S1 in Multimedia Appendix 1.

Group	Task	Feature	PCA coefficient	Age, *r*	CES-DC, *r*[a]	STAI-CH, *r*[a]	K-ARS, *r*[a]
1	eFlanker[b]	Overall reaction time	0.201	–0.645[c]	0.106	–0.019	–0.131
1	eFlanker	Reaction time (negative target, neutral distractor)	0.188	–0.714[c]	0.237[c]	0.113	–0.101
1	eFlanker	Reaction time (negative target, negative distractor)	0.192	–0.667[c]	0.262[c]	–0.094	–0.062
1	eFlanker	Reaction time (emotional target, congruent distractor)	0.210	–0.632[c]	0.293[c]	0.034	–0.028
1	eGoNo-Go[d]	Overall reaction time	0.146	–0.491[c]	0.176	0.132	–0.119
1	eStroop[e]	Overall reaction time	0.137	–0.47[c]	0.039	0.174	–0.177
2	eFlanker	Reaction time difference (emotional vs neutral targets; all distractors)	0.267	0.006	0.282[c]	0.158	–0.053
2	eFlanker	Accuracy (negative target, negative distractor)	–0.211	0.232[c]	–0.397[c]	–0.354[c]	–0.108
2	eFlanker	Accuracy difference (positive vs neutral targets; same-valence distractors)	–0.180	0.105	–0.263[c]	–0.186	–0.090
2	eFlanker	Accuracy difference (negative vs neutral targets; same-valence distractors)	–0.220	0.015	–0.235[c]	–0.232	–0.059
2	eFlanker	Accuracy difference (emotional-congruent vs neutral-neutral trials)	–0.250	0.131	–0.303[c]	–0.223	–0.053
3	eFlanker	Overall accuracy	0.285	0.499[c]	–0.013	–0.167	–0.255[c]
3	eFlanker	Accuracy (emotional target, incongruent distractor)	0.273	0.390[c]	0.143	–0.014	–0.329[c]
3	eFlanker	Accuracy (positive target, incongruent distractor)	0.274	0.344[c]	–0.060	–0.003	–0.383[c]
3	eFlanker	Accuracy (negative target, incongruent distractor)	0.224	0.381[c]	–0.096	–0.218	–0.29[c]
3	eFlanker	Accuracy difference (negative targets; incongruent vs congruent distractors)	0.062	0.009	0.129	0.220	–0.249[c]
3	eFlanker	Accuracy difference (emotional targets; incongruent vs congruent distractors)	0.146	0.204	0.257[c]	0.218	–0.142
3	eFlanker	Accuracy (all targets, incongruent distractor)	0.297	0.5[c]	0.034	–0.108	–0.333[c]
3	eFlanker	Accuracy (emotional targets)	0.287	0.432[c]	–0.071	–0.125	–0.322[c]
3	eGoNo-Go	Overall accuracy	0.188	0.333[c]	–0.078	0.036	–0.299[c]
3	eGoNo-Go	Accuracy (positive target)	0.139	0.423[c]	–0.113	–0.018	–0.292[c]
3	eStroop	Overall accuracy	0.165	0.292[c]	0.072	0.091	–0.145

[a]Age was covaried when calculating the Pearson correlation coefficients.

[b]eFlanker: emotional Flanker task.

[c]Indicates a statistical significance with *P* value of <.05.

[d]eGoNoGo: emotional Go/No-Go task.

[e]eStroop: emotional Stroop task.

Compared to depression, the number of children in the moderate to high anxiety group was lower (n=7, 9%; Figure 4C; STAI-CH >39 [58]). Accuracy on the eFlanker in the negative condition was the only behavioral measure correlated with the STAI-CH scale (Figure 4D; *r*=0.319; *P*=.009 after controlling for age; n=67). This may be due to a lower number of children with high anxiety, resulting in less clear across-individual contrasts for each behavioral index.

Lastly, a total of 10 (12.8%) children (K-ARS >18 [59]) were in the high-risk group for ADHD (Figure 4E). These children showed less accurate performance during the eFlanker and

XSL•FO
RenderX

eGoNoGo, especially in the emotional condition, compared to children with lower K-ARS scores (Figure 4F; r=–0.299; P=.02 after controlling for age; n=64).

Extracting "E(motional)-Scores" Using Behavioral Indices From the Emotional Cognitive Tasks

Given the nature of the tasks and the age of the participants, each behavioral index showed high variance across trials. In addition, we observed high redundancy as many behavioral measures were highly correlated with one another. Therefore, we opted to extract the common pattern spread over many behavioral indices, while maintaining the amount of information. We applied dimensionality reduction to the behavioral indices from the 3 emotional cognitive tasks to extract condensed information of individual performance by reducing repetitiveness (Figure 5A). Behavioral features (D=82) were divided into 3 groups by applying the k-means clustering method (Figure 5B), and E-scores were calculated by extracting the first PC of behavioral indices from each group (Figure 5C). We confirmed the robustness of the clustering using leave-one-out cross-validation, which assigned each excluded sample to the nearest group center and yielded labels identical to the original grouping.

Figure 5. Grouping of features (general reaction time [RT] and accuracy, the effects of emotion, and congruency effects) to calculate E-scores. (A) Dimensionality reduction of features based on their similarity across participants by principal component analysis (PCA). (B) The result of feature grouping after k-means clustering visualized on a 2D plane for the first 2 principal components (PCs). (C) Calculating E-scores using the first PC from each feature group. (D) Relationship between children's individual characteristics (ie, age, depression, anxiety, and attention-deficit/hyperactivity disorder [ADHD] scales) and E-scores. Only statistically significant (P<.05) linear relationships are visualized with a best-fit line. (E-G) Correlation between E-scores and age. (H-J) Correlation between each E-score and mental health/ADHD scales after controlling for age. CES-DC: Center for Epidemiological Studies Depression Scale for Children; FDR: false discovery rate; K-ARS: Korean ADHD Rating Scale; STAI-CH: State-Trait Anxiety Inventory for Children.

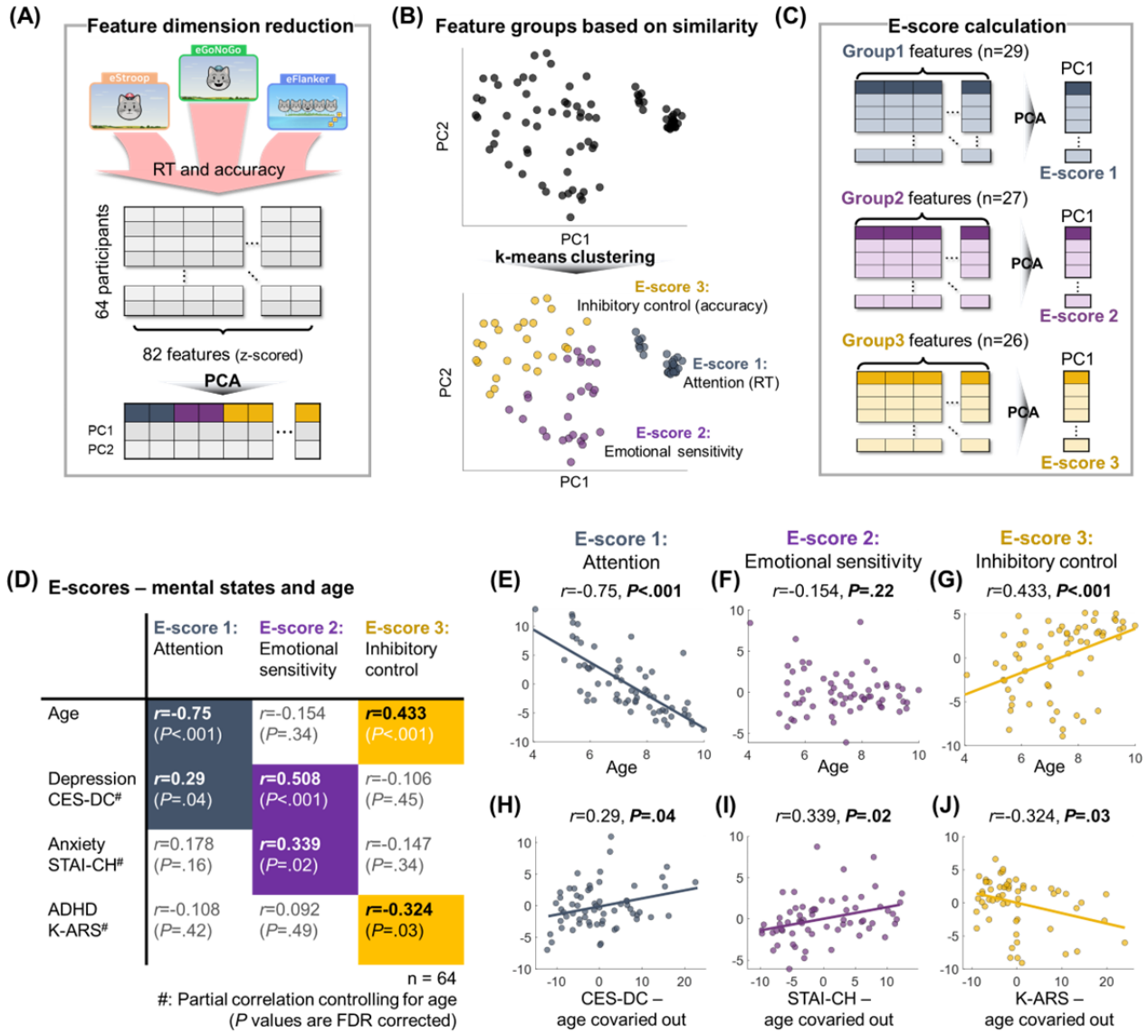

The first feature group contained RT features in various conditions across the eFlanker, eGoNoGo, and eStroop, thus representing the attention function (see Table S1 in Multimedia Appendix 1 for the whole list of behavioral features and coefficients for the first PC calculation). Positive PC coefficients of RT features for E-score 1 indicated that a higher E-score 1

XSL•FO
RenderX

represented decreased attention. Next, the second feature group represented complex interactions between emotion and attention. Coefficients of the first PC indicated that a higher E-score 2 was associated with a longer RT and lower accuracy, particularly in the emotional conditions, as well as higher congruency effects on RT in the eFlanker. Additionally, longer RT in the negative condition of the eStroop compared to the neutral condition also contributed to E-score 2. The third feature group represented accuracy on the tasks, as coefficients of accuracy features across various conditions in all 3 tasks positively contributed to E-score 3. Additionally, higher accuracy in the incongruent condition (ie, lower congruency effect) of the eFlanker also contributed to E-score 3.

Relationships Between E-Scores and Depression, Anxiety, or ADHD Symptoms

We evaluated the relevance of E-scores as individual emotional-cognitive characteristics by examining their correlations with mental health and ADHD survey scales (Figure 5D). First, ANCOVA tests were applied on E-scores with gender as a fixed factor and age as a continuous covariate. E-scores 1 and 3, which represented attention and executive function based on RT and accuracy, were significantly associated with age (E-score 1: $F_{1,60}$=78.404; P<.001; E-score 3: $F_{1,60}$=14.673; P<.001; all P values here and below are corrected for false discovery rate; see Pearson correlation coefficients in Figures 5E and 5G). In contrast, E-score 2, representing emotion-attention interactions, was not correlated with age (Figure 5F; $F_{1,60}$=1.493; P=.23). None of the E-scores showed significant between-gender differences. Self-reported depression level measured by the CES-DC was associated with E-score 1 (Figure 5H; r=0.29; P=.04), which replicated previous studies showing that children with depressive symptoms showed lower attention function represented by a longer RT [60]. In addition, a significant positive correlation between the CES-DC scale and E-score 2 (r=0.508; P<.001) indicated that children with higher depression were more sensitive to the emotional stimuli, resulting in longer RT and lower accuracy. E-score 2 was also positively correlated with the anxiety scale (ie, STAI-CH; Figure 5I; r=0.339; P=.02), showing similar cognitive characteristics induced by anxiety and depression. No differences in these association patterns were observed between the 2 collection methods (self-completed vs experimenter-guided; see Table S2 in Multimedia Appendix 1 for details). Lastly, E-score 3, which represented accuracy, was negatively correlated with the ADHD scale (ie, K-ARS; Figure 5J; r=–0.324; P=.03), suggesting that children with a higher risk for ADHD were more impulsive and inaccurate in their responses to emotional stimuli.

To further disentangle the effects of depression, anxiety, and ADHD on task performance (ie, E-scores), we conducted partial correlation analyses by controlling for the other scales and age (Table S3 in Multimedia Appendix 1). Every correlation remained significant except for E-score 2 and STAI-CH, which suggested that the E-score 2 does not reflect an anxiety-specific effect but rather overlapping characteristics shared between the anxiety and depression measures.

The dimensionality reduction technique revealed common and distinct behavioral patterns across the 3 emotional cognitive tasks. Extracting the common patterns contributed to reducing noise in the behavioral data and thus compensated for the short performance time of the tasks (10-15 minutes for each task), while the distinctiveness contributed to collecting various affective-cognitive profiles of children. Therefore, a range of behavioral indices was reduced to condense features, which were translated into a more robust, composite measure for individual affective-cognitive characteristics that is not susceptible to biases in a single measure. The results highlight the dissociability across depression, anxiety, and ADHD risk, given that each scale was correlated with different E-scores.

Discussion

Principal Findings

In this study, we aimed to investigate the distinct impacts of depression, anxiety, and ADHD on emotion recognition and attention in children aged between 4 and 10 years, using child-friendly gamified eFlanker, eGoNoGo, and eStroop. Our results highlight that various task performance measures related to emotion and cognition can be summarized into three distinctive components (attention, selective inhibition, and emotional sensitivity). Based on this, we introduced E-scores to incorporate meaningful information about the participant while minimizing noisy data distributed across the various features. The E-scores provided enhanced interpretability of behavioral task performance and were associated with mental health (anxiety and depression) or ADHD symptoms. The emotional cognitive assessments of our study are advantageous in that they allow monitoring of children's neurocognitive development, which may help address current limitations of measurement through self-report surveys.

Comparison to Prior Literature

The main novelty of this study lies in the development and validation of digital tasks that measure interactions between individual cognitive and affective characteristics and their associations with age-, mental health–, and ADHD-related measures within a unified framework. This approach offers advantages over previous studies that focused on single symptom domains [22-24], as it facilitates interpretation of children's behavioral outcomes while reducing potential confounds arising from comorbidity. Administering 3 digital cognitive tasks across multiple emotional conditions generated a wide range of behavioral indices; to address the resulting analytical complexity, we applied dimensionality reduction techniques to extract informative features across these indices. Together, these differences in study aims and methodological approach distinguish this work from prior studies.

As reported in many other studies [61], we observed a high correlation between depression and anxiety symptoms in our sample. Both anxiety and depression were associated with a larger emotional interference in attention functions (ie, RT increase and accuracy decrease in the emotional compared to the neutral conditions), typically higher for negative stimuli (Figure 4D). This is consistent with our previous study in which emotional bias served as a marker of both depression and anxiety [33]. Anxious and depressed individuals often show decreased inhibitory control [62] and higher emotional sensitivity [63],

XSL•FO
RenderX

and thus they have a higher chance of being distracted by emotional stimuli [33,64]. This pattern has been linked to a hyperactive amygdala, which is also associated with weaker top-down control and reduced activation in the dorsolateral PFC and ventromedial PFC [65,66].

ADHD risk measured by the K-ARS scale survey was independent of the depression or anxiety scales. Similarly, unlike depression or anxiety, ADHD was strongly associated with accuracy across various conditions in the 3 tasks. Children with ADHD have difficulty inhibiting irrelevant stimuli, reflecting deficits in selective attention and executive control [67]. The results in the eFlanker suggested that children with higher K-ARS scores showed reduced selective inhibition toward emotional stimuli (compared to neutral ones), which is consistent with their higher impulsivity [68,69]. Similarly, the lower accuracy observed in the eGoNoGo and eStroop likely resulted from disrupted working memory and executive control due to higher impulsivity. In particular, the general tendency for young children to respond to an incorrect positive stimulus in the eGoNoGo and eStroop are consistent with that interpretation. Interestingly, children's ADHD score was only associated with accuracy and not RT; this may be related to the fact that impulsive individuals tend to prioritize speed over accuracy, according to the well-known speed-accuracy trade-off, especially in the Flanker task [70].

Previously, we found that children's performance in the emotional cognitive tasks can be characterized as longer RT and lower accuracy compared to adults [33]. In this study, a cross-sectional investigation with a wider range of age showed that older children performed better than younger children, indicating that their general attention functions develop across childhood (4-10 years of age) with the maturation of the PFC [71]. Another characteristic of the emotional cognitive tasks is a general tendency for longer RT and lower accuracy in the emotional conditions. In particular, children exhibited more errors for positive stimuli than for neutral or negative stimuli, likely reflecting greater attentional bias toward positive cues, which can be associated with stronger approach motivation and heightened reward sensitivity [72]. However, these emotion-attention interactions remained unchanged across ages in our study. Previous studies reported that the emotional sensitivity of children is relatively higher until the age of around 10 years, when the amygdala reactivity decreases along with PFC maturation [73,74]. Thus, it is possible that all children in our study were significantly affected by emotional stimuli.

One difference between the results of our eFlanker with young children and our previous study [33] with adults is the reduction in the congruency effect. This discrepancy can be explained by developmental attention mechanisms that describe children's attentional bias toward more salient or emotional stimuli [7,8]. Congruency effects, characterized by longer RT and lower accuracy in the incongruent conditions, are induced by increased cognitive resources needed to resolve conflict in the incongruent condition and are typically accompanied by increased activity in the anterior cingulate cortex and other frontal regions [75]. However, as emotional targets are likely to attract more attention in children, flankers may be equally distracting in both congruent and incongruent conditions. Consistent with this interpretation, a reduced or absent congruency effect has also been reported in a previous study using a complex Flanker task with emotional faces as either target or distractor stimuli [56].

Strengths

The prevalence of childhood depression and ADHD has drastically increased over the past few years [76,77], and thus, a daily monitoring tool is essential for risk detection within the optimal time window for early intervention. We demonstrated that a simple emotional cognitive assessment tool can provide a stronger scientific basis for identifying affective-cognitive characteristics related to mental health and ADHD. Our approach highlights the efficacy of simple and easy, child-friendly, touchscreen-based tasks that are enjoyable even for young children, which is a critical factor for inducing accurate measurement of their attention, emotional sensitivity, and inhibitory function [78,79]. Furthermore, the average duration of gameplay was around 30 minutes, which is manageable and sustainable for repeated testing in children. These advantages of our tool enable an easy but quantifiable measurement of cognitive symptoms related to children's mental health and ADHD symptoms (eg, during school activities) for parents and clinicians.

Limitations and Future Directions

One limitation of this study is that it is based on a subclinical sample of children and that the mental health and ADHD measures were based on self-report and parent surveys. To enhance its applicability and generalizability to patient populations, future studies should validate the results using a formal clinical diagnosis. Nevertheless, in our previous study [33], we demonstrated that the relationships between task performance and mental health observed in healthy adults were also present in individuals diagnosed with depression. Second, parent-reported assessments of ADHD can also be subject to bias influenced by factors such as cultural context, parental mental health, and variability in the interpretation of symptoms. To mitigate this, we provided standardized instructions and guided administration to minimize variability in parental interpretation of symptom items. Third, while our age range was within the recommended period for the ADHD assessment, it may be relatively young for reliably assessing depression and anxiety, particularly for children aged <5 years. Therefore, our prediction based on behavioral performance may indicate a risk for later developmental stages. Fourth, cultural differences could influence emotion-cognition interactions in young children. Although this factor was minimized in this study, as all participants were Korean, cultural considerations may need to be considered when applying the tool internationally. Fifth, some children lost interest in the task during testing, which limited our ability to accurately assess their affective-cognitive characteristics. Including these data would likely weaken the observed associations between behavioral indices and mental health or ADHD survey scales; therefore, these cases were treated as missing data, resulting in a reduced sample size. For future applications in daily-life, such as deployment as a mobile or web-based application, online algorithms to detect loss of engagement will be necessary. In addition, appropriate real-time outlier detection and data-quality control procedures will be

XSL•FO
RenderX

required to reliably capture users' cognitive and affective characteristics. Lastly, a longitudinal study would be useful to investigate the effectiveness of the tool in assessing and predicting mental health and ADHD over an extended period, into adolescence.

Conclusion

Our gamified digital monitoring tool offers a practical and objective way to connect individual behavioral profiles with their affective and cognitive health, particularly for young children who may benefit greatly from early intervention but are difficult to accurately assess using traditional clinical interviews. While requiring further standardization and a well-established protocol to ensure data privacy, the ability to perform a rapid assessment also makes this tool well-suited for practical use in a variety of settings.

Acknowledgments

We would like to thank Minwoo Kim for his help with creating the task. We also thank Yu Jin Rah, Gakyung Kim, and the Gyeonggi Children's Museum, and all of the children and families that participated in this study.

Funding

SAL received funding from the Creative-Pioneering Researchers Program through Seoul National University, SYEDU, Samsung Electronics Co, Ltd (A0426-20220108), the National Research Foundation of Korea (2023R1A2C2005587, 2021M3A9E4080780, RS-2023-00301976), and the Korea Basic Science Institute (National Research Facilities and Equipment Center; RS-2024-00435727). Funders did not play any role in the study design, data collection, and analysis.

Data Availability

The data that support the findings of this study are publicly available on GitHub [80].

Authors' Contributions

Conceptualization: SEP (lead) and SAL (equal)
Data curation: JC
Formal analysis: SEP (lead), JC (supporting)
Funding acquisition: SAL
Methodology: SEP
Project administration: SAL
Supervision: SAL
Validation: SEP (lead) and SAL (equal)
Visualization: SEP
Writing – original draft: SEP
Writing – review and editing: SAL (lead) and JC (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional figure and tables.
[PDF File (Adobe PDF File), 414 KB - jmir_v28i1e86286_app1.pdf]

References

1. Muris P. Relation of attention control and school performance in normal children. Percept Mot Skills 2006;102(1):78-80. [doi: 10.2466/pms.102.1.78-80] [Medline: 16671601]
2. Kessler RC, Amminger GP, Aguilar-Gaxiola S, Alonso J, Lee S, Ustün TB. Age of onset of mental disorders: a review of recent literature. Curr Opin Psychiatry 2007;20(4):359-364 [FREE Full text] [doi: 10.1097/YCO.0b013e32816ebc8c] [Medline: 17551351]
3. Gaffrey MS, Barch DM, Luby JL, Petersen SE. Amygdala functional connectivity is associated with emotion regulation and amygdala reactivity in 4- to 6-year-olds. J Am Acad Child Adolesc Psychiatry 2021;60(1):176-185 [FREE Full text] [doi: 10.1016/j.jaac.2020.01.024] [Medline: 32119912]
4. Galvan A. Adolescent development of the reward system. Front Hum Neurosci 2010;4:6 [FREE Full text] [doi: 10.3389/neuro.09.006.2010] [Medline: 20179786]

XSL•FO
RenderX

5. Wu M, Kujawa A, Lu LH, Fitzgerald DA, Klumpp H, Fitzgerald KD, et al. Age-related changes in amygdala-frontal connectivity during emotional face processing from childhood into young adulthood. Hum Brain Mapp 2016;37(5):1684-1695 [FREE Full text] [doi: 10.1002/hbm.23129] [Medline: 26931629]
6. Hare TA, Tottenham N, Galvan A, Voss HU, Glover GH, Casey BJ. Biological substrates of emotional reactivity and regulation in adolescence during an emotional go-nogo task. Biol Psychiatry 2008;63(10):927-934 [FREE Full text] [doi: 10.1016/j.biopsych.2008.03.015] [Medline: 18452757]
7. Silvers JA, McRae K, Gabrieli JDE, Gross JJ, Remy KA, Ochsner KN. Age-related differences in emotional reactivity, regulation, and rejection sensitivity in adolescence. Emotion 2012;12(6):1235-1247 [FREE Full text] [doi: 10.1037/a0028297] [Medline: 22642356]
8. Arain M, Haque M, Johal L, Mathur P, Nel W, Rais A, et al. Maturation of the adolescent brain. Neuropsychiatr Dis Treat 2013;9:449-461 [FREE Full text] [doi: 10.2147/NDT.S39776] [Medline: 23579318]
9. Ahmed SP, Bittencourt-Hewitt A, Sebastian CL. Neurocognitive bases of emotion regulation development in adolescence. Dev Cogn Neurosci 2015;15:11-25 [FREE Full text] [doi: 10.1016/j.dcn.2015.07.006] [Medline: 26340451]
10. Michielsen M, Comijs HC, Semeijn EJ, Beekman ATF, Deeg DJH, Sandra Kooij JJ. The comorbidity of anxiety and depressive symptoms in older adults with attention-deficit/hyperactivity disorder: a longitudinal study. J Affect Disord 2013;148(2-3):220-227. [doi: 10.1016/j.jad.2012.11.063] [Medline: 23267726]
11. Kennedy SH. Core symptoms of major depressive disorder: relevance to diagnosis and treatment. Dialogues Clin Neurosci 2008;10(3):271-277 [FREE Full text] [doi: 10.31887/DCNS.2008.10.3/shkennedy] [Medline: 18979940]
12. Tsourtos G, Thompson JC, Stough C. Evidence of an early information processing speed deficit in unipolar major depression. Psychol Med 2002;32(2):259-265. [doi: 10.1017/s0033291701005001] [Medline: 11866321]
13. Gotlib IH, Joormann J. Cognition and depression: current status and future directions. Annu Rev Clin Psychol 2010;6:285-312 [FREE Full text] [doi: 10.1146/annurev.clinpsy.121208.131305] [Medline: 20192795]
14. Kujawa A, MacNamara A, Fitzgerald KD, Monk CS, Phan KL. Enhanced neural reactivity to threatening faces in anxious youth: evidence from event-related potentials. J Abnorm Child Psychol 2015;43(8):1493-1501 [FREE Full text] [doi: 10.1007/s10802-015-0029-4] [Medline: 25943264]
15. Eldar S, Yankelevitch R, Lamy D, Bar-Haim Y. Enhanced neural reactivity and selective attention to threat in anxiety. Biol Psychol 2010;85(2):252-257. [doi: 10.1016/j.biopsycho.2010.07.010] [Medline: 20655976]
16. Cisler JM, Koster EHW. Mechanisms of attentional biases towards threat in anxiety disorders: an integrative review. Clin Psychol Rev 2010;30(2):203-216 [FREE Full text] [doi: 10.1016/j.cpr.2009.11.003] [Medline: 20005616]
17. Barkley RA, Wasserstein J. ADHD and the nature of self-control. J Cogn Psychother 2000;14(1):111-113. [doi: 10.1891/0889-8391.14.1.111]
18. Iaboni F, Douglas VI, Ditto B. Psychophysiological response of ADHD children to reward and extinction. Psychophysiology 1997;34(1):116-123. [doi: 10.1111/j.1469-8986.1997.tb02422.x] [Medline: 9009815]
19. Patros CH, Alderson RM, Kasper LJ, Tarle SJ, Lea SE, Hudec KL. Choice-impulsivity in children and adolescents with attention-deficit/hyperactivity disorder (ADHD): a meta-analytic review. Clin Psychol Rev 2016;43:162-174. [doi: 10.1016/j.cpr.2015.11.001] [Medline: 26602954]
20. Wang K, Varma DS, Prosperi M. A systematic review of the effectiveness of mobile apps for monitoring and management of mental health symptoms or disorders. J Psychiatr Res 2018;107:73-78. [doi: 10.1016/j.jpsychires.2018.10.006] [Medline: 30347316]
21. Wiley K, Robinson R, Mandryk RL. The making and evaluation of digital games used for the assessment of attention: systematic review. JMIR Serious Games 2021;9(3):e26449 [FREE Full text] [doi: 10.2196/26449] [Medline: 34383674]
22. Peijnenborgh JCAW, Hurks PPM, Aldenkamp AP, van der Spek ED, Rauterberg GWM, Vles JSH, et al. A study on the validity of a computer-based game to assess cognitive processes, reward mechanisms, and time perception in children aged 4-8 years. JMIR Serious Games 2016;4(2):e15 [FREE Full text] [doi: 10.2196/games.5997] [Medline: 27658428]
23. Mandryk RL, Birk MV, Vedress S, Wiley K, Reid E, Berger P, et al. Remote assessment of depression using digital biomarkers from cognitive tasks. Front Psychol 2021;12:767507 [FREE Full text] [doi: 10.3389/fpsyg.2021.767507] [Medline: 34975656]
24. Voinescu A, Petrini K, Stanton Fraser D, Lazarovicz R, Papavă I, Fodor LA, et al. The effectiveness of a virtual reality attention task to predict depression and anxiety in comparison with current clinical measures. Virtual Real 2021;27(1):119-140. [doi: 10.1007/s10055-021-00520-7]
25. O'Laughlin KD, Cheng BH, Volponi JJ, Lorentz JDA, Obregon SA, Younger JW, et al. Validation of an adaptive assessment of executive functions (Adaptive Cognitive Evaluation-Explorer): longitudinal and cross-sectional analyses of cognitive task performance. J Med Internet Res 2025;27:e60041 [FREE Full text] [doi: 10.2196/60041] [Medline: 40258271]
26. Song H, Yi D, Park H. Validation of a mobile game-based assessment of cognitive control among children and adolescents. PLoS One 2020;15(3):e0230498 [FREE Full text] [doi: 10.1371/journal.pone.0230498] [Medline: 32196531]
27. Beck AT. The evolution of the cognitive model of depression and its neurobiological correlates. Am J Psychiatry 2008;165(8):969-977. [doi: 10.1176/appi.ajp.2008.08050721] [Medline: 18628348]
28. Robinson ESJ, Roiser JP. Affective biases in humans and animals. Curr Top Behav Neurosci 2016. [doi: 10.1007/7854_2015_5011] [Medline: 27311761]

XSL•FO
RenderX

29. Roiser JP, Elliott R, Sahakian BJ. Cognitive mechanisms of treatment in depression. Neuropsychopharmacology 2012;37(1):117-136 [FREE Full text] [doi: 10.1038/npp.2011.183] [Medline: 21976044]
30. Snyder HR, Miyake A, Hankin BL. Advancing understanding of executive function impairments and psychopathology: bridging the gap between clinical and cognitive approaches. Front Psychol 2015;6:328 [FREE Full text] [doi: 10.3389/fpsyg.2015.00328] [Medline: 25859234]
31. Mogg K, Salum GA, Bradley BP, Gadelha A, Pan P, Alvarenga P, et al. Attention network functioning in children with anxiety disorders, attention-deficit/hyperactivity disorder and non-clinical anxiety. Psychol Med 2015;45(12):2633-2646 [FREE Full text] [doi: 10.1017/S0033291715000586] [Medline: 26234806]
32. Mrug S, Madan A, Windle M. Emotional desensitization to violence contributes to adolescents' violent behavior. J Abnorm Child Psychol 2016;44(1):75-86 [FREE Full text] [doi: 10.1007/s10802-015-9986-x] [Medline: 25684447]
33. Park SE, Chung J, Lee J, Kim MJ, Kim J, Jeon HJ, et al. Digital assessment of cognitive-affective biases related to mental health. PLOS Digit Health 2024;3(8):e0000595. [doi: 10.1371/journal.pdig.0000595] [Medline: 39208388]
34. Abdul Rahman A, Wiebe SA. Valence matters: an electrophysiological study on how emotions influence cognitive performance in children. Dev Psychobiol 2019;61(2):290-303. [doi: 10.1002/dev.21813] [Medline: 30548999]
35. Vaughan AM, Perone S. Higher fearfulness in 4- to 6-year-old children related to better performance on an emotional than neutral stroop task. J Cogn Dev 2025;26(1):1-20. [doi: 10.1080/15248372.2025.2553532]
36. Ezpeleta L, Navarro JB, de la Osa N, Penelo E, Trepat E, Martin V, et al. Attention to emotion through a go/no-go task in children with oppositionality and callous-unemotional traits. Compr Psychiatry 2017;75:35-45 [FREE Full text] [doi: 10.1016/j.comppsych.2017.02.004] [Medline: 28301801]
37. Dudeney J, Sharpe L, Hunt C. Attentional bias towards threatening stimuli in children with anxiety: a meta-analysis. Clin Psychol Rev 2015;40:66-75. [doi: 10.1016/j.cpr.2015.05.007] [Medline: 26071667]
38. Harrison AJ, Gibb BE. Attentional biases in currently depressed children: an eye-tracking study of biases in sustained attention to emotional stimuli. J Clin Child Adolesc Psychol 2015;44(6):1008-1014 [FREE Full text] [doi: 10.1080/15374416.2014.930688] [Medline: 25010145]
39. Murray AL, Russell A, Calderón Alfaro FA. Early emotion regulation developmental trajectories and ADHD, internalizing, and conduct problems symptoms in childhood. Dev Psychopathol 2025;37(3):1474-1481. [doi: 10.1017/S0954579424001263] [Medline: 39282723]
40. Casey BJ, Jones RM, Somerville LH. Braking and accelerating of the adolescent brain. J Res Adolesc 2011;21(1):21-33. [doi: 10.1111/j.1532-7795.2010.00712.x] [Medline: 21475613]
41. Dick B, Ferguson BJ. Health for the world's adolescents: a second chance in the second decade. J Adolesc Health 2015;56(1):3-6. [doi: 10.1016/j.jadohealth.2014.10.260] [Medline: 25530601]
42. Marquardt L, Eichele H, Lundervold AJ, Haavik J, Eichele T. Event-related-potential (ERP) correlates of performance monitoring in adults with attention-deficit hyperactivity disorder (ADHD). Front Psychol 2018;9:485 [FREE Full text] [doi: 10.3389/fpsyg.2018.00485] [Medline: 29706908]
43. Pe ML, Vandekerckhove J, Kuppens P. A diffusion model account of the relationship between the emotional flanker task and rumination and depression. Emotion 2013;13(4):739-747 [FREE Full text] [doi: 10.1037/a0031628] [Medline: 23527499]
44. Zetsche U, D'Avanzato C, Joormann J. Depression and rumination: relation to components of inhibition. Cogn Emot 2012;26(4):758-767. [doi: 10.1080/02699931.2011.613919] [Medline: 21970297]
45. Gutiérrez-Cobo MJ, Cabello R, Fernández-Berrocal P. The three models of emotional intelligence and performance in a hot and cool go/no-go task in undergraduate students. Front Behav Neurosci 2017;11:33 [FREE Full text] [doi: 10.3389/fnbeh.2017.00033] [Medline: 28275343]
46. Schulz KP, Fan J, Magidina O, Marks DJ, Hahn B, Halperin JM. Does the emotional go/no-go task really measure behavioral inhibition? Convergence with measures on a non-emotional analog. Arch Clin Neuropsychol 2007;22(2):151-160 [FREE Full text] [doi: 10.1016/j.acn.2006.12.001] [Medline: 17207962]
47. Spielberger CD. State-Trait Anxiety Inventory for Children. Palo Alto, CA: Consulting Psychologist Press; 1973.
48. Weissman MM, Orvaschel H, Padian N. Children's symptom and social functioning self-report scales. Comparison of mothers' and children's reports. J Nerv Ment Dis 1980;168(12):736-740. [doi: 10.1097/00005053-198012000-00005] [Medline: 7452212]
49. DuPaul GJ, Power TJ, Anastopoulos AD, Reid R. ADHD Rating Scale-5 for Children and Adolescents: Checklists, Norms, and Clinical Interpretation. New York: The Guilford Press; 2016.
50. So YK, Noh JS, Kim YS, Ko SG, Koh YJ. The reliability and validity of Korean parent and teacher ADHD rating scale. J Korean Neuropsychiatr Assoc 2016;41(2):283-289. [doi: 10.4306/JKNPA.2002.41.2.283]
51. Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip Rev Comput Stat 2010;2(4):433-459. [doi: 10.1002/(issn)1939-0068]
52. Park SE, Laxpati NG, Gutekunst CA, Connolly MJ, Tung J, Berglund K, et al. A machine learning approach to characterize the modulation of the hippocampal rhythms via optogenetic stimulation of the medial septum. Int J Neural Syst 2019;29(10):1950020. [doi: 10.1142/S0129065719500205] [Medline: 31505977]
53. Arthur D, Vassilvitskii S. K-means++: the advantages of careful seeding. 2007 Presented at: SODA '07: Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms; 2007 January 7-9; New Orleans Louisiana p. 1027-1035.

XSL•FO
RenderX

54. Park SE, Lee J, Park JH, Hwang MH, Lee SA. Cognotyping by what-where-when retrieval reveals the potential role of temporal memory and its neural correlates in understanding individual differences across aging and alzheimer disease. J Cogn Neurosci 2023;35(11):1773-1787. [doi: 10.1162/jocn_a_02039] [Medline: 37584602]
55. Rykov A, De Amorim RC, Makarenkov V, Mirkin B. Inertia-based indices to determine the number of clusters in k-means: an experimental evaluation. IEEE Access 2024;12:11761-11773. [doi: 10.1109/access.2024.3350791]
56. Montalti M, Mirabella G. Unveiling the influence of task-relevance of emotional faces on behavioral reactions in a multi-face context using a novel flanker-go/no-go task. Sci Rep 2023;13(1):20183. [doi: 10.1038/s41598-023-47385-1] [Medline: 37978229]
57. Fendrich M, Weissman MM, Warner V. Screening for depressive disorder in children and adolescents: validating the center for epidemiologic studies depression scale for children. Am J Epidemiol 1990;131(3):538-551. [doi: 10.1093/oxfordjournals.aje.a115529] [Medline: 2301363]
58. Julian LJ. Measures of anxiety. Arthritis Care Res (Hoboken) 2011;63(suppl 11):S467-S472. [doi: 10.1002/acr.20561.Measures]
59. Cutler AJ, Brams M, Bukstein O, Mattingly G, McBurnett K, White C, et al. Response/remission with guanfacine extended-release and psychostimulants in children and adolescents with attention-deficit/hyperactivity disorder. J Am Acad Child Adolesc Psychiatry 2014;53(10):1092-1101. [doi: 10.1016/j.jaac.2014.08.001] [Medline: 25245353]
60. Wagner S, Müller C, Helmreich I, Huss M, Tadić A. A meta-analysis of cognitive functions in children and adolescents with major depressive disorder. Eur Child Adolesc Psychiatry 2015;24(1):5-19. [doi: 10.1007/s00787-014-0559-2] [Medline: 24869711]
61. McElroy E, Fearon P, Belsky J, Fonagy P, Patalay P. Networks of depression and anxiety symptoms across development. J Am Acad Child Adolesc Psychiatry 2018;57(12):964-973 [FREE Full text] [doi: 10.1016/j.jaac.2018.05.027] [Medline: 30522742]
62. Snyder HR, Friedman NP, Hankin BL. Transdiagnostic mechanisms of psychopathology in youth: executive functions, dependent stress, and rumination. Cognit Ther Res 2019;43(5):834-851 [FREE Full text] [doi: 10.1007/s10608-019-10016-z] [Medline: 31551642]
63. Watson D. Differentiating the mood and anxiety disorders: a quadripartite model. Annu Rev Clin Psychol 2009;5:221-247. [doi: 10.1146/annurev.clinpsy.032408.153510] [Medline: 19327030]
64. Gibb BE, McGeary JE, Beevers CG. Attentional biases to emotional stimuli: key components of the RDoC constructs of sustained threat and loss. Am J Med Genet B Neuropsychiatr Genet 2016;171B(1):65-80. [doi: 10.1002/ajmg.b.32383] [Medline: 26369836]
65. Motzkin JC, Philippi CL, Wolf RC, Baskaya MK, Koenigs M. Ventromedial prefrontal cortex is critical for the regulation of amygdala activity in humans. Biol Psychiatry 2015;77(3):276-284. [doi: 10.1016/j.biopsych.2014.02.014] [Medline: 24673881]
66. Nejati V, Majdi R, Salehinejad MA, Nitsche MA. The role of dorsolateral and ventromedial prefrontal cortex in the processing of emotional dimensions. Sci Rep 2021;11(1):1971 [FREE Full text] [doi: 10.1038/s41598-021-81454-7] [Medline: 33479323]
67. Tenenbaum RB, Musser ED, Morris S, Ward AR, Raiker JS, Coles EK, et al. Response inhibition, response execution, and emotion regulation among children with attention-deficit/hyperactivity disorder. J Abnorm Child Psychol 2019;47(4):589-603 [FREE Full text] [doi: 10.1007/s10802-018-0466-y] [Medline: 30112596]
68. Rosen PJ, Factor PI. Emotional impulsivity and emotional and behavioral difficulties among children with ADHD: an ecological momentary assessment study. J Atten Disord 2015;19(9):779-793. [doi: 10.1177/1087054712463064] [Medline: 23172248]
69. Özbaran B, Kalyoncu T, Köse S. Theory of mind and emotion regulation difficulties in children with ADHD. Psychiatry Res 2018;270:117-122. [doi: 10.1016/j.psychres.2018.09.034] [Medline: 30245374]
70. Dickman SJ, Meyer DE. Impulsivity and speed-accuracy tradeoffs in information processing. J Pers Soc Psychol 1988;54(2):274-290. [doi: 10.1037//0022-3514.54.2.274] [Medline: 3346814]
71. Uytun MC. Development period of prefrontal cortex. In: Prefrontal Cortex. London: Intechopen; 2018.
72. Pool E, Brosch T, Delplanque S, Sander D. Attentional bias for positive emotional stimuli: a meta-analytic investigation. Psychol Bull 2016;142(1):79-106. [doi: 10.1037/bul0000026] [Medline: 26390266]
73. Gee DG, Humphreys KL, Flannery J, Goff B, Telzer EH, Shapiro M, et al. A developmental shift from positive to negative connectivity in human amygdala-prefrontal circuitry. J Neurosci 2013;33(10):4584-4593 [FREE Full text] [doi: 10.1523/JNEUROSCI.3446-12.2013] [Medline: 23467374]
74. Gee DG, Hanson C, Caglar LR, Fareri DS, Gabard-Durnam LJ, Mills-Finnerty C, et al. Experimental evidence for a child-to-adolescent switch in human amygdala-prefrontal cortex communication: a cross-sectional pilot study. Dev Sci 2022;25(4):e13238 [FREE Full text] [doi: 10.1111/desc.13238] [Medline: 35080089]
75. Borsa VM, Della Rosa PA, Catricalà E, Canini M, Iadanza A, Falini A, et al. Interference and conflict monitoring in individuals with amnestic mild cognitive impairment: a structural study of the anterior cingulate cortex. J Neuropsychol 2018;12(1):23-40. [doi: 10.1111/jnp.12105] [Medline: 27147117]

XSL•FO
RenderX

76. Goodwin RD, Dierker LC, Wu M, Galea S, Hoven CW, Weinberger AH. Trends in U.S. depression prevalence from 2015 to 2020: the widening treatment gap. Am J Prev Med 2022;63(5):726-733 [FREE Full text] [doi: 10.1016/j.amepre.2022.05.014] [Medline: 36272761]
77. Danielson ML, Holbrook JR, Bitsko RH, Newsome K, Charania SN, McCord RF, et al. State-level estimates of the prevalence of parent-reported adhd diagnosis and treatment among U.S. children and adolescents, 2016 to 2019. J Atten Disord 2022;26(13):1685-1697. [doi: 10.1177/10870547221099961] [Medline: 35603751]
78. Lindqvist S, Thorell LB. Brief report: manipulation of task difficulty in inhibitory control tasks. Child Neuropsychol 2009;15(1):1-7. [doi: 10.1080/09297040701793647] [Medline: 18608218]
79. Simmering VR, Andrews CM, Leuenberger R, Kovack-Lesh KA. Inconsistent flanker congruency effects across stimulus types and age groups: a cautionary tale. Behav Res Methods 2023;55(4):1778-1817. [doi: 10.3758/s13428-022-01889-2] [Medline: 35768744]
80. Digital_Therapy_Baby. GitHub. URL: https://github.com/LeelabSNU/Digital_Therapy_Baby [accessed 2026-02-11]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
ANCOVA: analysis of covariance
CES-DC: Center for Epidemiological Studies Depression Scale for Children
DSM-5: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition)
eFlanker: emotional Flanker task
eGoNoGo: emotional Go/No-Go task
eStroop: emotional Stroop task
K-ARS: Korean ADHD Rating Scale
PC: principal component
PCA: principal component analysis
PFC: prefrontal cortex
RT: reaction time
STAI-CH: State-Trait Anxiety Inventory for Children

Edited by T Leung, S Brini; submitted 22.Oct.2025; peer-reviewed by A Barutcu, S Ajayi, S Ye; comments to author 21.Nov.2025; revised version received 23.Jan.2026; accepted 31.Jan.2026; published 25.Feb.2026.

Please cite as:
Park SE, Chung J, Lee SA
A Digital Tool for Assessing the Distinct Effects of Depression, Anxiety, and Attention-Deficit/Hyperactivity Disorder (ADHD) on Children's Emotional Cognitive Bias: Cross-Sectional Study
J Med Internet Res 2026;28:e86286
URL: https://www.jmir.org/2026/1/e86286
doi:10.2196/86286
PMID:

©Sang-Eon Park, Jisu Chung, Sang Ah Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Factors Influencing Continuance Intention for Online Consultations Among Survivors of Cancer: Grounded Theory Study

Yutang Yao[1*], MD; Musi Zhang[2*], BSN; Shanshan Peng[2], BMed; Zhuzhong Cheng[1], MD; Yun Duan[1], BMed

[1]Department of Nuclear Medicine, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

[2]Department of Radiation Oncology, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

[*]these authors contributed equally

Corresponding Author:
Yun Duan, BMed
Department of Nuclear Medicine
Sichuan Cancer Hospital & Institute, Sichuan Cancer Center
University of Electronic Science and Technology of China
55 Renmin South Rd Section 4
Chengdu, Sichuan, 610041
China
Phone: 86 02885420214
Email: duanyun163@126.com

Abstract

Background: Online consultation platforms have become an important component of survivorship care for patients with cancer, offering flexible access to oncology expertise between scheduled visits. However, evidence on what drives the willingness of survivors of cancer to continue using online consultations after initial adoption remains limited in China. A better understanding of continuance intention is needed to inform survivor-centered digital health strategies.

Objective: This study aimed to explore the influencing factors of continued use of online consultations among survivors of cancer in southwest China and develop a grounded theoretical model explaining continuance intention.

Methods: A grounded theory qualitative design was used. A total of 26 adult survivors of cancer with diverse demographic and clinical characteristics were purposively recruited from a tertiary cancer center in southwest China. All participants had used online consultations at least once in the preceding year. Semistructured telephone interviews were audio recorded; transcribed verbatim; and analyzed using open, axial, and selective coding with constant comparison until theoretical saturation was reached. During selective coding, categories and their relationships were integrated and iteratively refined to construct a grounded theoretical model of continuance intention.

Results: Six interrelated domains influenced survivors' continued use of online consultation platforms: platform quality, physician competence, user perception, individual condition, external context, and privacy concerns. Platform quality and physician competence influenced user perception of usefulness, reassurance, and trust, which functioned as a mediator of continued use. Individual condition, including health status, health literacy, and psychological needs, influenced both perceived usefulness and reliance on online consultations. External context, especially family encouragement, peer recommendations, and availability of local oncology services, directly facilitated or constrained continued use. Privacy concerns moderated how survivors balanced perceived benefits against risks of data misuse, stigma, and unwanted disclosure of cancer history. Survivors described online consultations as offering rapid guidance and emotional support that complemented hospital-based care but reported discontinuation when interactions were delayed or impersonal or when perceived privacy risks outweighed the benefits.

Conclusions: The willingness of survivors of cancer to continue using online consultation platforms depends on multiple interrelated factors beyond traditional technological usability. Sustained engagement is shaped by survivors' perceptions of usefulness and trust, physician empathy and timeliness, family encouragement, and acceptance of privacy trade-offs. The theoretical model advances understanding of digital health continuance in oncology and offers practical guidance for developing survivor-centered online consultation services.

(J Med Internet Res 2026;28:e84644) doi:10.2196/84644

XSL•FO
RenderX

KEYWORDS

survivors of cancer; online consultation; telemedicine; digital health; grounded theory; continuance intention; patient engagement; physician-patient communication; privacy concerns; survivorship care

Introduction

Background

Cancer remains one of the leading causes of morbidity and mortality worldwide, placing a tremendous burden on health systems, families, and individuals. Advances in diagnosis and treatment have improved survival rates, and as a result, the population of survivors of cancer continues to expand [1-3]. However, survivorship is not simply the period after treatment. It also includes ongoing challenges such as long-term side effects, fear of recurrence, psychosocial adjustment, and the need for continuous follow-up care. The World Health Organization and many national cancer control programs have emphasized the importance of survivorship care as an integral component of comprehensive cancer management [4,5].

Digital health technologies have become increasingly important in bridging gaps in access to care, especially for patients requiring long-term monitoring [6-8]. Online consultation platforms, accessed through mobile apps or web portals, allow survivors to communicate with oncology professionals, seek advice on symptoms or lifestyle modifications, and receive psychological support. For survivors living in regions with limited oncology resources, online consultation provides a vital link to specialists otherwise geographically inaccessible [9-12]. Even in urban centers with advanced hospitals, the platforms offer convenience and continuity between scheduled in-person visits [13,14].

The COVID-19 pandemic further accelerated the adoption of telemedicine and online health services, demonstrating their potential to supplement traditional care [14-16]. Survivors of cancer, often immunocompromised, benefitted from reduced exposure to hospital environments through digital consultations. As health systems adapt to postpandemic realities, online consultation is positioned as a permanent component of integrated cancer survivorship care. However, the long-term viability of such services depends not merely on initial adoption but also on survivors' willingness to continue using them. Therefore, understanding the factors that shape continued use is essential for sustaining the role of online consultation in oncology care.

Research on online consultation and telemedicine has expanded rapidly in recent years, with a growing focus on use patterns and determinants of adoption [17-19]. Several theoretical models have guided this scholarship. The technology acceptance model (TAM) and expectation confirmation model (ECM) are frequently used to examine perceived usefulness, ease of use, and satisfaction as drivers of continued use. Studies across various digital health contexts, including mobile health apps, wearable devices, and patient portals, have confirmed that perceived usefulness and trust are strong predictors of sustained engagement [19-22].

In oncology, early work has highlighted the potential of online consultation to support symptom management, medication adherence, and communication between patients and health care professionals. For example, randomized controlled trials involving survivors of breast cancer have demonstrated that digital follow-up systems improve quality of life and reduce hospital visits [23-26]. Observational studies in Europe and Asia have reported that online consultation platforms are especially effective in providing dietary advice, managing side effects, and delivering psychosocial support to survivors [27-29].

A growing body of literature has investigated continuance intention in digital health contexts. In nononcology populations, continuance is influenced by habit formation, social influence, and perceived quality of service. For instance, research on chronic disease management apps has found that individuals with strong digital literacy and positive reinforcement from peers are more likely to continue use [30-32]. Other studies have identified cost, accessibility, and integration with offline care as determinants of sustained use [33,34]. However, in cancer populations, empirical evidence remains fragmented. Some studies of survivors of breast cancer have revealed that online consultation provides emotional reassurance and complements offline care [35-37], whereas others have found that survivors discontinue use due to inconsistent physician availability and lack of trust [38].

In this study, continuance intention refers to survivors' intention to keep using an online consultation platform after initial adoption rather than a single episode of use. In digital health, continuance intention is essential because the benefits of telemedicine and online consultation usually accumulate through repeated contacts, ongoing symptom management, and sustained relationships with clinicians. If survivors discontinue use after early trials, online services may show good initial uptake yet fail to deliver long-term gains in symptom control, psychological support, or care coordination.

While prior studies provide valuable insights into adoption of and satisfaction with online consultations, few have systematically investigated the factors influencing continued use by survivors of cancer over time. Most quantitative research has relied on preexisting models such as the TAM or ECM, which capture perceptions of technology but may not adequately reflect the lived realities of survivorship. Survivors navigate a complex interplay of medical, psychological, social, and technological factors. Their willingness to sustain engagement with online consultation platforms cannot be reduced to perceived usefulness alone.

There is also a lack of theory-building research specific to oncology survivorship in the Chinese context. Most studies of digital health continuance have applied established frameworks such as the TAM and ECM with a focus on perceived usefulness, ease of use, and satisfaction. Such models were not developed for the complex realities of cancer survivorship, where survivors simultaneously manage late effects, fear of

XSL•FO
RenderX

recurrence, family expectations, and resource constraints in the health system. Without theory grounded in survivors' lived experiences, especially within Chinese sociocultural and institutional settings, it is difficult to design online consultation services that are both acceptable and sustainable. Theory-building work can clarify which technological, relational, and contextual influences matter most for continuance and how they interact during survivorship care.

Objectives

Therefore, this study aimed to explore the factors that influence the continued use of online consultation platforms by survivors of cancer in southwest China and develop a grounded theoretical model explaining continuance intention in this context. Specifically, this study sought to identify individual, relational, and contextual influences on continued use and articulate how such influences interrelate within survivorship care.

Methods

Study Design

This study used a grounded theory qualitative design to examine factors influencing continuance intention regarding online consultations among survivors of cancer. The approach by Strauss and Corbin [39] was adopted, with iterative cycles of data collection and analysis and a 3-stage process of open coding, axial coding, and selective coding [40]. Constant comparative analysis was used to compare incidents within and across interviews and refine categories and their properties as the theoretical model developed [41].

Study Setting

This study was conducted at the Sichuan Cancer Hospital and Institute, Sichuan Cancer Center, the Affiliated Cancer Hospital of the University of Electronic Science and Technology of China. This institution is the largest national tertiary cancer hospital in southwest China, with comprehensive functions in cancer prevention, treatment, rehabilitation, research, and education. Its extensive clinical services and diverse patient population made it an ideal setting for examining the experiences of survivors of cancer with online consultation platforms.

Participants and Recruitment

Participants were adult survivors of cancer with diverse cancer types and in diverse survivorship stages who had used an online consultation platform at least once in the preceding year. Inclusion criteria were age of ≥18 years, confirmed cancer diagnosis, completion of initial treatment or in active follow-up, at least one prior online consultation related to cancer care within the previous year, and ability to communicate via telephone in Mandarin. Survivors with severe cognitive impairment or acute clinical distress or who were unable to communicate effectively via phone were excluded. The criterion of at least one recent online consultation ensured that participants could describe concrete experiences with platform use and decisions about continued use rather than hypothetical views.

Recruitment followed purposive sampling to maximize diversity in age, gender, cancer type, treatment stage, and place of residence. During outpatient follow-up appointments, clinical nurses and oncologists briefly introduced the study to eligible survivors and, with permission, shared contact details with the research team. The team then telephoned interested survivors to provide detailed study information, confirm eligibility, and arrange an interview time. In survivorship support groups and patient-led online forums, a short study notice invited interested survivors to contact the team directly. All invitations were active rather than open public advertisements. Degree of experience with online consultation and digital literacy were not used as formal sampling strata. Instead, the focus was on variation in survivorship trajectories and clinical backgrounds. Digital literacy was not assessed using a standardized scale, which is acknowledged as a limitation when interpreting differences in continuance intention.

Data Collection

Data were collected through semistructured telephone interviews, which were chosen to accommodate survivors living in different regions and reduce travel burden. After eligibility was confirmed, interviews were scheduled at times convenient for participants, usually outside routine clinic visits. Two trained oncology nurses (YY and MZ) conducted all interviews. Both interviewers worked at the same tertiary cancer center but were not part of the clinical team directly responsible for participants' current treatment. At the start of each interview, the interviewer introduced their professional background, clarified the voluntary nature of participation, and emphasized that decisions about care would not be affected by participation. Interviews followed a flexible guide covering experiences with online consultations, reasons for continued use or discontinuation, family and social influences, and privacy concerns. (Multimedia Appendix 1) Interviews lasted between 25 and 40 minutes, were audio recorded with permission, and were transcribed verbatim in Mandarin. Each transcript was anonymized and assigned an identifier (C01 to C26).

Data Analysis

Data analysis began after the first interviews and proceeded concurrently with ongoing data collection. Two researchers (YY and MZ) conducted line-by-line open coding on an initial set of transcripts to identify concepts related to survivors' experiences of online consultations, perceived benefits and drawbacks, relational and family influences, and privacy concerns. (Multimedia Appendix 2) Codes were compared, merged, and refined in regular meetings, and a preliminary coding framework was developed.

During axial coding, conceptually similar codes were clustered into categories, and relationships among categories were explored using constant comparison across participants and time points. Analytic memos documented emerging ideas about potential mediators, moderators, and contextual conditions influencing continuance intention. In the selective coding stage, categories were integrated into 6 higher-level domains and a core category of continuance intention. The developing theoretical model was iteratively checked against the data, including accounts that appeared to deviate from early interpretations, and revised until it accounted for the range of observed patterns.

XSL•FO
RenderX

Coding was conducted independently by the 2 researchers, and discrepancies were discussed and resolved through consensus, with a third researcher (YD) consulted when needed. An audit trail including codebooks, memos, and diagrams was maintained to support transparency and dependability. Sampling became increasingly focused as analysis progressed, for example, by recruiting survivors from different age groups and residential areas once the importance of family involvement and privacy concerns became clear. Theoretical saturation was assessed after 23 interviews when no new categories were identified and relationships between domains appeared stable. Three additional interviews were conducted to confirm saturation and ensure that the model held true for more recent cases.

Rigor and Trustworthiness

Credibility and trustworthiness were supported through multiple strategies. First, purposive and iterative sampling captured survivors with diverse demographic and clinical backgrounds to enhance variation in experiences. Second, 2 researchers independently coded transcripts and compared interpretations in regular analysis meetings, with disagreements resolved through discussion and involvement of a third researcher where required. Third, constant comparison and negative case analysis were used to test whether the evolving model could accommodate accounts that challenged early assumptions. Fourth, an audit trail of coding decisions, memos, and diagrams was maintained to support dependability and confirmability. Finally, brief member checking was conducted with 4 participants who were invited to comment on thematic summaries; they confirmed that the domains and relationships reflected their experiences. Reporting follows the Standards for Reporting Qualitative Research and is informed by the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines.

Researcher Positionality

The research team consisted of oncology clinicians and nursing researchers working in a tertiary cancer center in southwest China. The interviewers were oncology nurses with long-standing experience caring for survivors during treatment and follow-up, which facilitated rapport but may also have shaped the topics explored and the way in which participants described their care. The senior author is a nuclear medicine physician with experience in survivorship care and digital health initiatives in the hospital where this study was conducted. The team acknowledges that familiarity with hospital-affiliated online platforms and a generally positive view of digital health could influence interpretation of the data. To address this, assumptions were documented in analytic memos, and team discussions explicitly considered alternative explanations and accounts that did not align with expectations.

Ethical Considerations

The study protocol was reviewed and approved by the ethics committee of Sichuan Cancer Hospital and Institute (approval SCCHEC-02-2020-036). All participants received verbal and written information about the study and provided informed verbal consent before the interviews. Participation was voluntary, and survivors could decline questions or withdraw at any time without consequences for their clinical care.

To protect privacy and confidentiality, audio recordings were stored on password-protected devices accessible only to the research team, and transcripts were deidentified by removing names and other direct identifiers. Potentially identifying combinations of demographic and clinical details were aggregated in reporting so that individual participants could not be recognized. No financial incentives or material compensation were provided for participation.

Results

Participant Characteristics

The 26 participants included 15 (58%) women and 11 (42%) men, with most being middle-aged (40-49 years: n=8, 31%; 50-59 years: n=7, 27%). Survivors of breast cancer constituted the largest group (n=7, 27%), followed by lung (n=4, 15%) and colorectal (n=3, 12%) cancer, whereas other cancer types such as cervical, ovarian, prostate, gastric, and thyroid cancer each represented 8% (n=2), and liver and kidney cancer each represented 4% (n=1). Survivorship stages were balanced, with 31% (n=8) in active treatment and 35% (n=9) each in remission and long-term survivorship. Educational levels ranged from 23% (n=6) with a high school or lower level to 12% (n=3) with postgraduate education. Most participants (n=11, 42%) were employed, and 62% (n=16) lived in urban areas as presented in Table 1.

Analysis identified 6 domains shaping the continued use by survivors of cancer of online consultation platforms: platform quality, physician competence, user perception, individual condition, external context, and privacy concerns. The interrelationships among the domains formed a theoretical model explaining continuance intention.

XSL•FO
RenderX

Table 1. Sociodemographic and clinical characteristics of survivors of cancer participating in grounded theory interviews on continuance intention regarding online consultations in southwest China (N=26).

Category	Participants, n (%)
Gender	
Female	15 (58)
Male	11 (42)
Age (years)	
30-39	5 (19)
40-49	8 (31)
50-59	7 (27)
≥60	6 (23)
Cancer type	
Breast	7 (27)
Lung	4 (15)
Colorectal	3 (12)
Cervical	2 (8)
Ovarian	2 (8)
Prostate	2 (8)
Gastric	2 (8)
Thyroid	2 (8)
Liver	1 (4)
Kidney	1 (4)
Treatment stage	
Active treatment	8 (31)
Remission	9 (35)
Long-term survivorship	9 (35)
Educational level	
High school or lower	6 (23)
College or junior college	10 (38)
Bachelor’s degree	7 (27)
Master’s degree or higher	3 (12)
Occupation status	
Employed	11 (42)
Retired	7 (27)
Homemaker	3 (12)
Self-employed	3 (12)
Unemployed	2 (8)
Geographic location	
Urban	16 (62)
Semiurban or rural	10 (38)

Platform Quality

Platform quality encompassed service quality and information quality. Survivors valued consistent access to oncology specialists, transparent consultation fees, reliable technical support, and effective complaint resolution. Information quality included completeness of physician profiles, clarity of treatment explanations, and access to trustworthy cancer-related educational content as presented in Table 2.

XSL•FO
RenderX

Table 2. Themes and subthemes related to platform quality influencing continuance intention regarding online consultations among survivors of cancer.

Subtheme	Key insights	Quote
Service quality	Survivors expected efficient systems, reasonable fees, and transparent policies.	"The charges were clear. I could reach the same oncologist again. The convenience made me continue." [C04; male; survivor of gastric cancer]
Information quality	Accurate physician and treatment information increased trust.	"I can see which doctors are experts or which are younger doctors. It's easier than going to the hospital." [C21; female; survivor of ovarian cancer]

Physician Competence

Physician competence emerged as central. Survivors highlighted professional expertise, communication ability, and timeliness. Expertise reassured patients on side effect management and recurrence risks. Compassionate communication provided emotional support. Timely responses were critical, especially during treatment as presented in Table 3.

Table 3. Themes and subthemes related to physician competence influencing continuance intention regarding online consultations among survivors of cancer.

Subtheme	Key insights	Quote
Professional expertise	Detailed, tailored explanations increased trust.	"When the doctor explained why my fatigue persisted, it felt like guidance I could trust." [C18; male; survivor of lung cancer]
Communication ability	Survivors valued empathy and encouragement.	"She told me that anxiety was normal, and suddenly I felt less alone. That made me come back." [C09; female; survivor of breast cancer]
Timeliness	Delays discouraged use; quick replies reinforced reliance.	"During chemo, hours felt like days. A fast reply was the reason I kept using the app." [C23; male; survivor of colorectal cancer]

User Perception

User perception comprised ease of use, perceived usefulness, and positive expectation. Survivors emphasized the importance of intuitive interfaces and smooth navigation. Usefulness was defined as reassurance between hospital visits, management of treatment side effects, and guidance for lifestyle adaptation. Positive expectation referred to trust in the future development of digital health services.

Survivors described online consultations as a lifeline during uncertain recovery phases. One participant noted the following:

> *Even when nothing urgent, knowing I could reach professional doctors quickly gave me comfort.* [C06; female; survivor of cervical cancer]

Individual Condition

Individual condition referred to health literacy, health status, and personal needs. Survivors with higher health literacy were more confident in engaging online, whereas those with limited literacy faced challenges in sustained use. Health status influenced patterns: those experiencing lingering treatment effects frequently sought reassurance, whereas long-term survivors without active symptoms used consultation more sporadically. Needs extended beyond medical queries to psychological reassurance and lifestyle advice.

One participant said the following:

> *After surgery I still had numbness in my hands, and I didn't always know if it was normal. Having the online consultation gave me a way to check quickly without waiting weeks for my hospital appointment.* [C02; female; aged 52 years; survivor of breast cancer]

External Context

External context included family encouragement, peer influence, and offline health care alternatives. Family members often facilitated continued use, particularly adult children assisting older survivors with technology. Peer groups shared recommendations of reliable platforms, which reinforced trust. Survivors in rural areas with limited oncology services relied heavily on online consultations, whereas those living near tertiary hospitals sometimes preferred in-person visits.

For instance, a participant mentioned the following:

> *My son insisted I keep using the app. He said it was safer than traveling two hours to the hospital every time I worried about something small, and he even helped me learn how to pay for consultations.* [C09; male; aged 63 years; survivor of colorectal cancer]

Privacy Concerns

Privacy concerns centered on disclosure of medical records, genetic test results, and sensitive images. Survivors expressed anxiety about misuse of their cancer history but balanced this against perceived benefits. When consultations provided timely reassurance, survivors accepted privacy trade-offs.

One survivor explained the following:

> *I hesitated before uploading my CT scans, but the doctor's advice was worth it.* [C12; male; survivor of prostate cancer]

Interrelationships Among Domains

The findings revealed that platform quality and physician competence had a strong influence on user perception of online consultations, including perceived usefulness, reassurance, and trust. User perception, in turn, mediated the relationship between these domains and continuance intention. Survivors' decisions

XSL•FO
RenderX

to continue or discontinue online consultations depended on how they interpreted their cumulative experiences rather than on isolated platform attributes. Individual condition, including health status, health literacy, and psychological needs, directly shaped both perceived usefulness and reliance on online consultations. External context, such as family encouragement, peer influence, and accessibility of local oncology services, provided structural conditions that either facilitated or constrained continued use. Privacy concerns moderated the influence of other domains by intensifying or weakening the impact of perceived benefits. Survivors with high privacy concerns sometimes restricted or stopped use despite recognizing benefits, whereas institutional trust in hospital-affiliated platforms could soften privacy fears and support continued engagement (Table 4).

Table 4. Representative relationships among domains in the grounded theory model and their descriptions.

Relationship	Description
User perception→intention	Perceptions of usefulness and trust shaped continuity.
Platform quality→perception	Service quality enhanced trust and ease of use.
Physician competence→perception	Physician expertise built reassurance.
External context→intention	Peer recommendations influenced adoption.
Individual condition→intention	Active symptoms increased reliance.
Privacy concern→intention	Anxiety reduced willingness.

Theoretical Model

The final theoretical model positions continuance intention at the center, influenced by 6 interconnected domains. Platform quality and physician competence directly influence user perception, which mediates the relationship with continuance. Individual condition and external context exert both direct and indirect effects. Privacy concerns act as a contextual moderator, influencing the strength of survivors' willingness to continue (Figure 1).

Figure 1. Grounded theory model of continuance intention for online consultations among survivors of cancer. Platform quality and physician competence influence continuance intention indirectly through user perception (mediator). Individual condition and external context have direct effects on continuance intention. Privacy concerns have a direct negative effect on continuance intention and moderate the association between user perception and continuance intention (N=26).

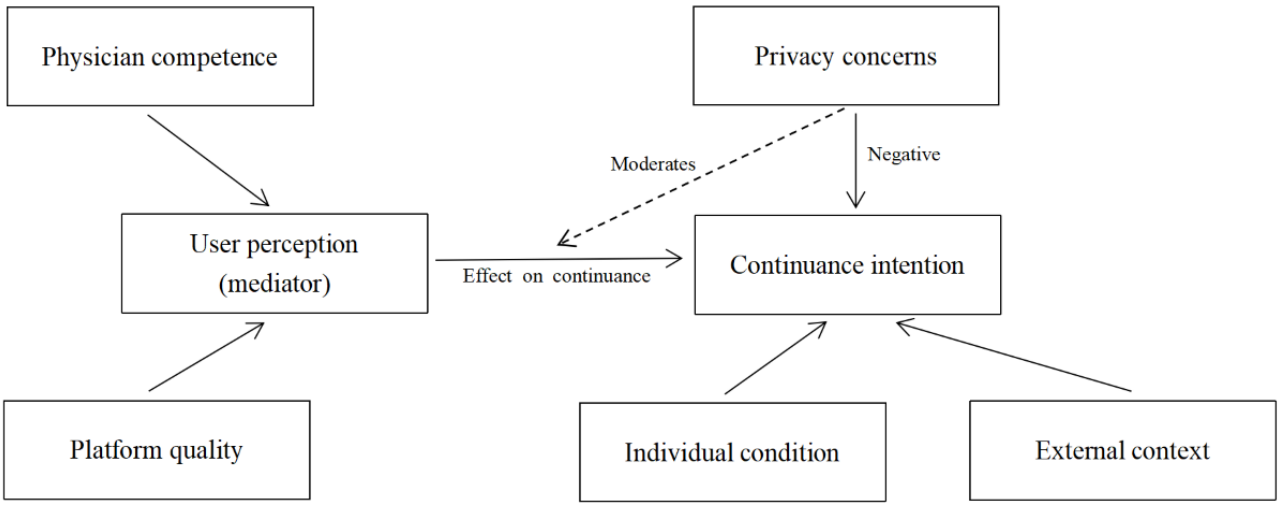

Discussion

Principal Findings

Analysis of in-depth interviews with 26 Chinese survivors of cancer identified 6 interrelated domains that shaped continuance intention: platform quality, physician competence, user perception, individual condition, external context, and privacy concerns. The domains were integrated into a theoretical model illustrating how survivors' decisions to continue using online consultation platforms were mediated by perceptions of usefulness, trust, and accessibility while being simultaneously shaped by their personal health status, family influences, and privacy considerations. The findings underscore that continued use is not determined solely by technological ease or usefulness but emerges from a dynamic negotiation among survivors' individual needs, relational contexts, and systemic constraints.

Comparison With Prior Work

Previous studies have consistently identified perceived usefulness, trust, and habit formation as determinants of sustained use. For example, research on mobile disease management apps has demonstrated that convenience and ease of monitoring reinforce long-term engagement [42,43]. Other studies have linked continuance intention to perceptions of quality of physician feedback and integration into daily routines [44,45]. This suggests that the mechanisms of continuance

XSL•FO
RenderX

identified in our study are not unique to oncology but resonate with broader digital health behaviors.

Survivors of cancer often experience uncertainty long after completion of primary treatment, with concerns about recurrence, lingering side effects, and psychosocial adjustment [46-48]. Unlike users of general wellness or chronic disease apps, survivors in our study emphasized the importance of empathetic physician communication and reassurance as key reasons for continuing online consultations. This aligns with studies on survivors of breast cancer, which have found that online consultations offer emotional support and alleviate isolation [49,50]. By centering survivors' voices, our grounded theory extends beyond existing quantitative models that have often prioritized technological features over relational care.

Privacy concerns emerged as a moderator of continuance intention. Prior work has documented anxiety about data security in telemedicine broadly [51,52], but our findings reveal that survivors actively weigh the risks of disclosure against perceived benefits of reassurance. For example, some participants expressed hesitation about uploading genetic test results or sensitive images but, ultimately, continued use when the consultation reduced uncertainty. This dynamic risk-benefit trade-off is less discussed in the literature but resonates with recent studies in Chinese mobile health contexts, which report that patients are more willing to share personal data when platforms are affiliated with reputable hospitals [53]. Our study shows that this negotiation is particularly salient in oncology, where personal health information carries social and familial implications.

The findings need to be interpreted within the sociocultural context of southwest China. Survivors' decisions about online consultations were often negotiated within families rather than made individually. Adult children in particular encouraged continued use, assisted with technical tasks, and sometimes controlled access to online services. Trust in public hospital-affiliated platforms reduced worries about data misuse compared with commercial platforms and supported acceptance of privacy trade-offs. At the same time, differences in digital literacy and internet access between urban and rural areas and between younger and older survivors shaped how readily survivors could use and benefit from online consultations. Such features of the Chinese context influence the transferability of the model to other settings and illustrate the value of theory development that is grounded in local sociocultural dynamics.

The configuration of the model would likely differ in health care contexts characterized by lower institutional trust or more individualistic decision-making. In low-trust settings, institutional affiliation may be less able to buffer privacy concerns, and privacy concerns may exert a more direct negative effect on continuance rather than primarily moderating perceived benefits. In more individualistic contexts, the external context pathway via family facilitation may be weaker, whereas individual appraisal of risk, autonomy, and personal preference may play a stronger role in shaping continuance intention.

Theoretical and Practical Implications

This study advances understanding of digital oncology care in several ways. It focused on continuance intention rather than on initial adoption or satisfaction, emphasizing the long-term dynamics of survivor engagement. It developed a theory-informed framework grounded in the lived experiences of survivors of cancer in China, addressing an important gap in survivorship research. The model highlights how technological features, physician competence, survivor perceptions, individual conditions, family and social contexts, and privacy concerns interact to shape continued use of online consultation platforms. As a result, it offers a more comprehensive account of digital health continuance and practical guidance for policymakers, health care institutions, and technology developers who seek to design sustainable services for survivors of cancer.

Theoretically, this study shows the value of grounded theory for examining continuance intention in digital health and extends work based on the TAM and the ECM. Existing frameworks foreground technological usability, perceived usefulness, and satisfaction. In contrast, this model integrates relational, contextual, and personal health factors as core elements of continuance, thereby expanding the conceptual vocabulary for studying sustained engagement with health technologies. Unlike consumer technologies where continuance is often driven by convenience, entertainment, or routine fit, continuance intention in oncology survivorship is structurally anchored in clinical uncertainty, perceived vulnerability, and reliance on professional reassurance between episodic in-person visits. Survivors re-engage when symptoms, late effects, or fear of recurrence create a need for interpretation and emotional stabilization, making the relationship with clinicians and the perceived safety of the channel central to sustained use. In this context, reassurance and trust operate as distinct relational mechanisms rather than as subcomponents of perceived usefulness. Reassurance reflects reduction of uncertainty and emotional distress through clinician responsiveness and empathy, whereas trust reflects confidence in physician competence and institutional credibility. Both shape whether survivors interpret online consultations as safe and legitimate for ongoing survivorship management, which explains why user perception mediates continuance intention through reassurance and trust in addition to instrumental usefulness.

The model also clarifies the role of user perception. Survivors' willingness to continue was shaped not only by platform quality or physician competence in isolation but also by how combined experiences influenced perceptions of usefulness, trust, and reassurance. User perception functioned as an active interpretive process that mediated the impact of system attributes and survivorship trajectories on continuance intention. In addition, this study offers a more refined understanding of privacy concerns in digital health. Privacy concerns operated as a dynamic moderator that affected the strength of continuance intention depending on how survivors weighed perceived benefits against perceived risks. This view moves beyond simple classifications of users as either concerned or unconcerned and encourages consideration of continuance as an ongoing negotiation.

XSL•FO
RenderX

From a practical standpoint, the findings suggest several priorities for online consultation services in oncology. Platforms should ensure clear physician profiles, transparent fee structures, and reliable technical support so that survivors can form stable expectations about service quality. Clinicians who provide online consultations require support and training in empathetic communication to address survivors' psychological and emotional needs in addition to clinical questions. Timely responses are particularly critical for survivors in active treatment, which supports the integration of triage protocols, clear coverage arrangements, and notification systems that reduce avoidable delays. Addressing privacy concerns calls for transparent communication about data protection policies and visible affiliation with trusted institutions; survivors in this study expressed greater confidence in platforms linked to tertiary hospitals. Finally, given the central role of family encouragement, platform features that facilitate caregiver involvement, such as options for shared access or family-oriented consultation modes, may align with prevailing cultural practices and help sustain survivor engagement over time.

Limitations

First, this study was conducted in a single regional tertiary cancer center in southwest China. Although the institution serves a large and diverse catchment area, experiences in other regions or health systems may differ. Second, the sample comprised survivors who were willing and able to participate in telephone interviews and who had had at least one prior online consultation, which may bias findings toward survivors who are more engaged with digital services. Standardized measures of digital literacy and prior telehealth experience were not collected, limiting the ability to quantify their influence on continuance intention. Third, interviews were conducted by oncology nurses affiliated with the same institution, which may have introduced social desirability bias or inhibited criticism of hospital-based platforms despite efforts to emphasize independence from clinical decisions. Finally, qualitative findings are interpretive and context specific; future research using mixed methods and larger samples is needed to test and refine the model in other oncological and cultural settings.

Future Directions

Future research should test the theoretical model developed in this study through quantitative methods. Large-scale surveys could examine the relative influence of the 6 domains and validate their interrelationships. Cross-cultural comparative studies would be valuable to assess whether the role of family support and privacy negotiation is unique to China or generalizable to other contexts. Furthermore, longitudinal research could explore how continuance intention evolves across different phases of survivorship, from active treatment to long-term follow-up.

Intervention studies could also evaluate strategies to strengthen continuance. For instance, training programs for physicians in digital empathy or platform designs that allow for caregiver participation may enhance perceptions of usefulness and trust. Policy research should investigate frameworks for data protection that balance survivors' privacy with the need for effective digital oncology services. As cancer survivorship continues to grow worldwide, such research is critical for integrating online consultations into sustainable, patient-centered models of care.

Conclusions

Six interrelated domains influenced the continuance intention of survivors of cancer to use online consultation platforms: platform quality, physician competence, user perception, individual condition, external context, and privacy concerns. The grounded theory model shows that continued use results from a dynamic negotiation between perceived benefits and perceived risks that is shaped by survivors' health needs, family roles, and institutional trust. Beyond summarizing determinants of service use, the model provides a theory-based foundation for designing and evaluating online consultation services that are more responsive to survivors' long-term needs. By clarifying how technological, relational, and contextual influences interact in a Chinese oncology setting, this study contributes to broader efforts to build sustainable and equitable digital health systems in cancer survivorship care.

Acknowledgments

During the preparation and revision of this manuscript, we used generative artificial intelligence tools, namely Kimi Chat (version K2; Moonshot AI) and GPT-5 and ChatGPT-5.1 (OpenAI) to assist with original draft translation, language polishing, and restructuring, which were further reviewed and revised by the study group.

Funding

This study was supported by the Sichuan Medical and Health Care Promotion Institute Scientific Research Project (grant). The funder had no role in the study design, data collection, data analysis, decision to publish, or preparation of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: YY, MZ, YD

XSL•FO
RenderX

Formal analysis: YY, MZ, SP, ZC, YD
Funding acquisition: YY
Investigation and data collection: YY, MZ, SP, ZC
Methodology: YY, MZ, SP
Supervision: YY, MZ
Visualization: YY, ZC
Writing—original draft: YY, MZ, SP
Writing—review and editing: YY, MZ, SP, ZC, YD

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semistructured interview guide.
[PDF File (Adobe PDF File), 75 KB - jmir_v28i1e84644_app1.pdf]

Multimedia Appendix 2
Open coding concepts.
[XLSX File (Microsoft Excel File), 12 KB - jmir_v28i1e84644_app2.xlsx]

References

1. Chen ZD, Zhang PF, Xi H, Wei B, Chen L, Tang Y. Recent advances in the diagnosis, staging, treatment, and prognosis of advanced gastric cancer: a literature review. Front Med (Lausanne) 2021 Oct 26;8:744839. [doi: 10.3389/fmed.2021.744839] [Medline: 34765619]
2. Zeineddine FA, Zeineddine MA, Yousef A, Gu Y, Chowdhury S, Dasari A, et al. Survival improvement for patients with metastatic colorectal cancer over twenty years. NPJ Precis Oncol 2023 Feb 13;7(1):16 [FREE Full text] [doi: 10.1038/s41698-023-00353-4] [Medline: 36781990]
3. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. CA Cancer J Clin 2021 Jan 12;71(1):7-33 [FREE Full text] [doi: 10.3322/caac.21654] [Medline: 33433946]
4. Jacobsen PB, DeRosa AP, Henderson TO, Mayer DK, Moskowitz CS, Paskett ED, et al. Systematic review of the impact of cancer survivorship care plans on health outcomes and health care delivery. J Clin Oncol 2018 Jul 10;36(20):2088-2100 [FREE Full text] [doi: 10.1200/JCO.2018.77.7482] [Medline: 29775389]
5. Chan RJ, Crawford-Williams F, Crichton M, Joseph R, Hart NH, Milley K, et al. Effectiveness and implementation of models of cancer survivorship care: an overview of systematic reviews. J Cancer Surviv 2023 Feb 16;17(1):197-221 [FREE Full text] [doi: 10.1007/s11764-021-01128-1] [Medline: 34786652]
6. Marthick M, McGregor D, Alison J, Cheema B, Dhillon H, Shaw T. Supportive care interventions for people with cancer assisted by digital technology: systematic review. J Med Internet Res 2021 Oct 29;23(10):e24722 [FREE Full text] [doi: 10.2196/24722] [Medline: 34714246]
7. Garg S, Williams NL, Ip A, Dicker AP. Clinical Integration of digital solutions in health care: an overview of the current landscape of digital technologies in cancer care. JCO Clin Cancer Inform 2018 Dec;2:1-9. [doi: 10.1200/CCI.17.00159] [Medline: 30652580]
8. Mikles SP, Griffin AC, Chung AE. Health information technology to support cancer survivorship care planning: a systematic review. J Am Med Inform Assoc 2021 Sep 18;28(10):2277-2286 [FREE Full text] [doi: 10.1093/jamia/ocab134] [Medline: 34333588]
9. Morris BB, Rossi B, Fuemmeler B. The role of digital health technology in rural cancer care delivery: a systematic review. J Rural Health 2022 Jun;38(3):493-511 [FREE Full text] [doi: 10.1111/jrh.12619] [Medline: 34480506]
10. Shi J, Robinson TA, Loomba P, Murley B, Strong LL, Basen-Engquist K, et al. "Everything cannot be handled virtually": understanding intention and use of digital health technologies among rural adults and rural cancer survivors. J Rural Health 2025 Jan 29;41(1):e12926. [doi: 10.1111/jrh.12926] [Medline: 39878383]
11. Hemming L, Duijts SF, Zomerdijk N, Cockburn C, Yuen E, Hardman R, et al. A systematic review of peer support interventions to improve psychosocial functioning among cancer survivors: can findings be translated to survivors with a rare cancer living rurally? Orphanet J Rare Dis 2024 Dec 20;19(1):473 [FREE Full text] [doi: 10.1186/s13023-024-03477-3] [Medline: 39707418]
12. Pradhan P, Sharman AR, Palme CE, Elliott MS, Clark JR, Venchiarutti RL. Models of survivorship care in patients with head and neck cancer in regional, rural, and remote areas: a systematic review. J Cancer Surviv (Forthcoming) 2024 Jul 20. [doi: 10.1007/s11764-024-01643-x] [Medline: 39031309]

XSL•FO
RenderX

13. Bin Ramlee MH, Abdul Satar NF. The feasibility and preliminary efficacy of a locally developed digital health intervention in improving breast cancer survivor's quality of life (QOL): qualitative and quantitative analysis result. J Clin Oncol 2024 Jun 01;42(16_suppl):e13803. [doi: 10.1200/jco.2024.42.16_suppl.e13803]
14. Onyeaka HK, Zambrano J, Longley RM, Celano CM, Naslund JA, Amonoo HL. Use of digital health tools for health promotion in cancer survivors. Psychooncology 2021 Aug;30(8):1302-1310 [FREE Full text] [doi: 10.1002/pon.5677] [Medline: 33742737]
15. Toni E, Ayatollahi H. An insight into the use of telemedicine technology for cancer patients during the COVID-19 pandemic: a scoping review. BMC Med Inform Decis Mak 2024 Apr 19;24(1):104 [FREE Full text] [doi: 10.1186/s12911-024-02507-1] [Medline: 38641567]
16. Royce TJ, Sanoff HK, Rewari A. Telemedicine for cancer care in the time of COVID-19. JAMA Oncol 2020 Nov 01;6(11):1698-1699 [FREE Full text] [doi: 10.1001/jamaoncol.2020.2684] [Medline: 32672821]
17. Pang N, Lau J, Fong SY, Wong CY, Tan KK. Telemedicine acceptance among older adult patients with cancer: scoping review. J Med Internet Res 2022 Mar 29;24(3):e28724 [FREE Full text] [doi: 10.2196/28724] [Medline: 35348462]
18. Irurita-Morales P, Soto-Ruiz N, Martín-Rodríguez LS, Escalada-Hernández P, García-Vivar C. Use of telehealth among cancer survivors: a scoping review. Telemed J E Health 2023 Jul 01;29(7):956-985. [doi: 10.1089/tmj.2022.0351] [Medline: 36445755]
19. Cox A, Lucas G, Marcu A, Piano M, Grosvenor W, Mold F, et al. Cancer survivors' experience with telehealth: a systematic review and thematic synthesis. J Med Internet Res 2017 Jan 09;19(1):e11 [FREE Full text] [doi: 10.2196/jmir.6575] [Medline: 28069561]
20. Mou J, Shin DD, Cohen J. Understanding trust and perceived usefulness in the consumer acceptance of an e-service: a longitudinal investigation. Behav Inf Technol 2016 Jul 22;36(2):125-139. [doi: 10.1080/0144929X.2016.1203024]
21. Adjekum A, Blasimme A, Vayena E. Elements of trust in digital health systems: scoping review. J Med Internet Res 2018 Dec 13;20(12):e11254 [FREE Full text] [doi: 10.2196/11254] [Medline: 30545807]
22. Liu K, Tao D. The roles of trust, personalization, loss of privacy, and anthropomorphism in public acceptance of smart healthcare services. Comput Human Behav 2022 Feb;127:107026. [doi: 10.1016/j.chb.2021.107026]
23. Pimentel-Parra GA, Soto-Ruiz MN, San Martín-Rodríguez L, Escalada-Hernández P, García-Vivar C. Effectiveness of digital health on the quality of life of long-term breast cancer survivors: a systematic review. Semin Oncol Nurs 2023 Aug;39(4):151418 [FREE Full text] [doi: 10.1016/j.soncn.2023.151418] [Medline: 37045645]
24. Saltbæk L, Bidstrup PE, Karlsen RV, Høeg BL, Horsboel TA, Belmonte F, et al. Nurse-led individualized follow-up versus regular physician-led visits after early breast cancer (MyHealth): a phase III randomized, controlled trial. J Clin Oncol 2024 Jun 10;42(17):2038-2049. [doi: 10.1200/JCO.23.01447] [Medline: 38498781]
25. Saevarsdottir SR, Gudmundsdottir SL. Mobile apps and quality of life in patients with breast cancer and survivors: systematic literature review. J Med Internet Res 2023 Jul 26;25:e42852 [FREE Full text] [doi: 10.2196/42852] [Medline: 37494111]
26. Little P, Bradbury K, Stuart B, Barnett J, Krusche A, Steele M, et al. Digital intervention (renewed) to support symptom management, wellbeing, and quality of life among cancer survivors in primary care: a randomised controlled trial. Br J Gen Pract 2025 May;75(754):e357-e365. [doi: 10.3399/BJGP.2023.0262] [Medline: 38164562]
27. Lazar DE, Postolica R, Hanganu B, Mocanu V, Ioan BG. Web-based nutrition: a useful resource for cancer patients? Front Nutr 2023;10:1134793 [FREE Full text] [doi: 10.3389/fnut.2023.1134793] [Medline: 37457987]
28. Medina JC, Flix-Valle A, Rodríguez-Ortega A, Hernández-Ribas R, Lleras de Frutos M, Ochoa-Arnedo C. ICOnnecta't: development and initial results of a stepped psychosocial ehealth ecosystem to facilitate risk assessment and prevention of early emotional distress in breast cancer survivors' journey. Cancers (Basel) 2022 Feb 15;14(4):974 [FREE Full text] [doi: 10.3390/cancers14040974] [Medline: 35205722]
29. Walsh CA, Al Achkar M. A qualitative study of online support communities for lung cancer survivors on targeted therapies. Support Care Cancer 2021 Aug;29(8):4493-4500 [FREE Full text] [doi: 10.1007/s00520-021-05989-1] [Medline: 33458808]
30. Taylor ML, Thomas EE, Vitangcol K, Marx W, Campbell KL, Caffery LJ, et al. Digital health experiences reported in chronic disease management: an umbrella review of qualitative studies. J Telemed Telecare 2022 Nov 08;28(10):705-717. [doi: 10.1177/1357633x221119620]
31. Verweel L, Newman A, Michaelchuk W, Packham T, Goldstein R, Brooks D. The effect of digital interventions on related health literacy and skills for individuals living with chronic diseases: a systematic review and meta-analysis. Int J Med Inform 2023 Sep;177:105114. [doi: 10.1016/j.ijmedinf.2023.105114] [Medline: 37329765]
32. Li Z, Lu F, Wu J, Bao R, Rao Y, Yang Y, et al. Usability and effectiveness of eHealth and mHealth interventions that support self-management and health care transition in adolescents and young adults with chronic disease: systematic review. J Med Internet Res 2024 Nov 26;26:e56556 [FREE Full text] [doi: 10.2196/56556] [Medline: 39589770]
33. Endalamaw A, Zewdie A, Wolka E, Assefa Y. A scoping review of digital health technologies in multimorbidity management: mechanisms, outcomes, challenges, and strategies. BMC Health Serv Res 2025 Mar 15;25(1):382 [FREE Full text] [doi: 10.1186/s12913-025-12548-5] [Medline: 40089752]
34. Gentili A, Failla G, Melnyk A, Puleo V, Tanna GL, Ricciardi W, et al. The cost-effectiveness of digital health interventions: a systematic review of the literature. Front Public Health 2022;10:787135 [FREE Full text] [doi: 10.3389/fpubh.2022.787135] [Medline: 36033812]

XSL•FO
RenderX

35. Akkol-Solakoglu S, Hevey D. Internet-delivered cognitive behavioural therapy for depression and anxiety in breast cancer survivors: results from a randomised controlled trial. Psychooncology 2023 Mar 30;32(3):446-456. [doi: 10.1002/pon.6097] [Medline: 36635249]
36. Peng L, Yang Y, Chen M, Xu C, Chen Y, Liu R, et al. Effects of an online mindfulness-based intervention on fear of cancer recurrence and quality of life among Chinese breast cancer survivors. Complement Ther Clin Pract 2022 Nov;49:101686. [doi: 10.1016/j.ctcp.2022.101686] [Medline: 36347151]
37. Smith SK, Westbrook K, MacDermott K, Amarasekara S, LeBlanc M, Pan W. Four conversations: a randomized controlled trial of an online, personalized coping and decision aid for metastatic breast cancer patients. J Palliat Med 2020 Mar 01;23(3):353-358. [doi: 10.1089/jpm.2019.0234] [Medline: 31638448]
38. Yang M, Jiang J, Kiang M, Yuan F. Re-examining the impact of multidimensional trust on patients' online medical consultation service continuance decision. Inf Syst Front 2022 Mar 04;24(3):983-1007 [FREE Full text] [doi: 10.1007/s10796-021-10117-9] [Medline: 33688300]
39. Gosby JR. Media reviews: basics of qualitative research - techniques and procedures for developing grounded theory 2nd edition by A. Strauss and J. Corbin. Sage Publications. J Adv Nurs 2003 Jan 15;32(4):1036. [doi: 10.1046/j.1365-2648.2000.t01-7-01210.x]
40. Chun Tie Y, Birks M, Francis K. Grounded theory research: a design framework for novice researchers. SAGE Open Med 2019;7:2050312118822927 [FREE Full text] [doi: 10.1177/2050312118822927] [Medline: 30637106]
41. McCrae N, Purssell E. Is it really theoretical? A review of sampling in grounded theory studies in nursing journals. J Adv Nurs 2016 Oct 26;72(10):2284-2293. [doi: 10.1111/jan.12986] [Medline: 27113800]
42. Adu MD, Malabu UH, Malau-Aduli AE, Drovandi A, Malau-Aduli BS. User retention and engagement with a mobile app intervention to support self-management in Australians with type 1 or type 2 diabetes (my care hub): mixed methods study. JMIR Mhealth Uhealth 2020 Jun 11;8(6):e17802 [FREE Full text] [doi: 10.2196/17802] [Medline: 32525491]
43. Crafoord MT, Fjell M, Sundberg K, Nilsson M, Langius-Eklöf A. Engagement in an interactive app for symptom self-management during treatment in patients with breast or prostate cancer: mixed methods study. J Med Internet Res 2020 Aug 10;22(8):e17058 [FREE Full text] [doi: 10.2196/17058] [Medline: 32663140]
44. Yan M, Filieri R, Raguseo E, Gorton M. Mobile apps for healthy living: factors influencing continuance intention for health apps. Technol Forecast Soc Change 2021 May;166:120644. [doi: 10.1016/j.techfore.2021.120644]
45. Wang T, Wang W, Liang J, Nuo M, Wen Q, Wei W, et al. Identifying major impact factors affecting the continuance intention of mHealth: a systematic review and multi-subgroup meta-analysis. NPJ Digit Med 2022 Sep 15;5(1):145 [FREE Full text] [doi: 10.1038/s41746-022-00692-9] [Medline: 36109594]
46. Demoor-Goldschmidt C, Porro B. Editorial: young adults or adults who are survivors of childhood cancer: psychosocial side effects, education, and employment. Front Psychol 2024 Nov 8;15:1510822 [FREE Full text] [doi: 10.3389/fpsyg.2024.1510822] [Medline: 39582993]
47. Luo X, Xu H, Zhang Y, Liu S, Xu S, Xie Y, et al. Identifying the unmet needs of post-treatment colorectal cancer survivors: a critical literature review. Eur J Oncol Nurs 2024 Jun;70:102570 [FREE Full text] [doi: 10.1016/j.ejon.2024.102570] [Medline: 38574419]
48. Antoni MH, Moreno PI, Penedo FJ. Stress management interventions to facilitate psychological and physiological adaptation and optimal health outcomes in cancer patients and survivors. Annu Rev Psychol 2023 Jan 18;74(1):423-455 [FREE Full text] [doi: 10.1146/annurev-psych-030122-124119] [Medline: 35961041]
49. Niu Z, Bhurosy T, Heckman C. Cancer survivors' emotional well-being: roles of internet information seeking, patient-centered communication, and social support. J Health Commun 2021 Jul 03;26(7):514-522. [doi: 10.1080/10810730.2021.1966685] [Medline: 34435927]
50. Suarez NR, Morrow AS, LaVecchia CM, Dugas M, Carnovale V, Maraboto A, et al. Connected and supported: a scoping review of how online communities provide social support for breast cancer survivors. J Cancer Surviv (Forthcoming) 2024 Aug 28. [doi: 10.1007/s11764-024-01660-w] [Medline: 39196462]
51. Magdy M, Hosny KM, Ghali NI, Ghoniemy S. Security of medical images for telemedicine: a systematic review. Multimed Tools Appl 2022 Mar 22;81(18):25101-25145 [FREE Full text] [doi: 10.1007/s11042-022-11956-7] [Medline: 35342327]
52. Zanke P, Sontakke D. Safeguarding patient confidentiality in telemedicine: a systematic review of privacy and security risks, and best practices for data protection. Int J Curr Sci Res Rev 2024 Jun 19;07(06):3910-3922. [doi: 10.47191/ijcsrr/v7-i6-42]
53. Li X, Cong Y. A systematic literature review of ethical challenges related to medical and public health data sharing in China. J Empir Res Hum Res Ethics 2021 Dec 13;16(5):537-554. [doi: 10.1177/15562646211040299] [Medline: 34516325]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
ECM: expectation confirmation model
TAM: technology acceptance model

XSL•FO
RenderX

Edited by A Mavragani, N Cahill; submitted 23.Sep.2025; peer-reviewed by R Akinniranye, O Akhadelor, E Emokpae; comments to author 26.Nov.2025; revised version received 19.Dec.2025; accepted 23.Dec.2025; published 07.Jan.2026.

Please cite as:
Yao Y, Zhang M, Peng S, Cheng Z, Duan Y
Factors Influencing Continuance Intention for Online Consultations Among Survivors of Cancer: Grounded Theory Study
J Med Internet Res 2026;28:e84644
URL: https://www.jmir.org/2026/1/e84644
doi:10.2196/84644
PMID:41432715

©Yutang Yao, Musi Zhang, Shanshan Peng, Zhuzhong Cheng, Yun Duan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Structure of Psychopathology on Reddit: Network Analysis of Mental Health Communities in Relation to the ICD Diagnostic System

Bojan Evkoski[1], MSc; Srebrenka Letina[2], PhD; Petra Kralj Novak[1,3], PhD

[1]Department of Network and Data Science, Central European University, Vienna, Austria

[2]Department of Psychology, University of Limerick, Limerick, Ireland

[3]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

Corresponding Author:
Bojan Evkoski, MSc
Department of Network and Data Science
Central European University
Quellenstraße 51
Vienna, 1100
Austria
Phone: 43 1252307111
Email: evkoski_bojan@phd.ceu.edu

Abstract

Background: Social media platforms such as Reddit have become important spaces where individuals articulate their distress, seek support, and explore alternative ways of understanding mental health outside traditional institutional frameworks. These environments provide an opportunity to examine mental health discourse at scale, offering perspectives that extend beyond traditional clinical and research settings.

Objective: This study aims to examine the structure of mental health communities on Reddit by identifying patterns of association between mental disorders reflected in user activity and assessing how these relationships align with established diagnostic categories in the *ICD* (*International Classification of Diseases*).

Methods: We manually curated 114 Reddit communities focused on specific mental health conditions from the 20,000 most active subreddits in 2022. Each community was labeled into 49 disorders and categorized under 9 *ICD* diagnostic categories within the group of mental and behavioral disorders, collectively known as the F codes. We constructed a disorder association network by identifying statistically significant user overlaps based on coposting across subreddit pairs using a bipartite configuration model, with Bonferroni-corrected significance (P<.001). We analyzed the connectivity of the network within and across diagnostic categories, examining inter- and intracategory links. Finally, we compared the structure of disorder associations inferred from Reddit with the *ICD* classification derived from diagnostic criteria using hierarchical clustering.

Results: The inferred Reddit network of psychopathology revealed an interconnected structure (density=0.135), with all but 6 disorders forming a single giant component that spans across all 9 diagnostic categories. The most prominent disorders by number of users included hyperkinetic disorders (85,000), depressive episodes and recurrent depressive disorders (73,000), habit and impulse disorders (69,000), pervasive developmental disorders (52,000), and generalized anxiety disorder (44,000). In terms of connectivity, posttraumatic stress disorder (17/48 of all possible connections), obsessive-compulsive disorder (16/48), and depersonalization-derealization disorder (15/48) emerged as the most central in the network of positive disorder associations, while schizotypal disorder, avoidant personality disorder, and agoraphobia were the most central when accounting for the association strength. At the level of disorder categories, several disorders, such as bipolar disorder and premenstrual dysphoric disorder, displayed high intercategory associations but weak intracategory ties, indicating blurred diagnostic boundaries. The network of negative coposting associations revealed a divergence from the expectations of past research; for instance, addiction-related communities (eg, alcohol and opioids) were negatively associated with much of the broader mental health discourse. Finally, hierarchical comparisons showed moderate overlap between the Reddit network of disorder associations and the *ICD* network of diagnostic criteria, both in pairwise edge similarity (13% of edges present in both networks) and overall clustering (Adjusted Rand Index=0.295).

XSL•FO
RenderX

Conclusions: Reddit-based mental health communities reveal a complementary structure of disorder associations shaped by lived experience, often diverging from formal diagnostic criteria and exhibiting patterns of association that do not align with established diagnostic boundaries.

(J Med Internet Res 2026;28:e80958) doi:10.2196/80958

KEYWORDS

psychopathology; disorder associations; online mental health support; Reddit; network analysis; ICD-10

Introduction

Mental health problems have remained one of the main public health concerns, especially among younger populations [1-3]. This trend has unfolded amid rapid technological advancements and broader societal changes, including the lasting disruptions brought by the COVID-19 pandemic [4,5]. Yet, the pace of these societal and technological shifts has not been matched by corresponding adaptability in the broader health care system.

The challenge of adaptability is particularly evident in the diagnostic frameworks of psychopathology. Clinical taxonomies such as the *DSM* (*Diagnostic and Statistical Manual of Mental Disorders*) and the *ICD* (*International Classification of Diseases*) provide standardized frameworks for identifying, labeling, and treating psychological conditions. In these systems, disorders are defined and grouped into diagnostic categories based on symptom profiles, potential causal mechanisms or course of illness. These categories serve essential roles in practice: they guide diagnostic and treatment decisions, shape insurance coverage, structure research protocols, and enable communication across professionals. However, despite their utility, these frameworks and their diagnostic categories are built on standardized assumptions that inevitably generalize and oversimplify inherently subjective and dynamic experiences. Consequently, they have faced ongoing criticism regarding the ambiguity of their boundaries and their limited clarity, stability, and cultural relevance across diverse contexts [6-8]. Recent reform efforts, such as the Research Domain Criteria and the Hierarchical Taxonomy of Psychopathology, aim to address these limitations, but also underscore the complexity and ongoing contestation surrounding psychiatric classification in the pursuit of a more fluid, flexible, and context-sensitive approach to understanding mental illness [9-11]. However, such efforts cannot be fully undertaken in isolation from the shifting social and technological environments that shape how symptoms are expressed, experienced, and interpreted.

Over the past 2 decades, digital technologies have transformed nearly every facet of daily life, including how individuals relate to their own mental health. Mobile connectivity and online social media play central roles in how people articulate their experiences, shape their identities, and seek support [12,13]. Platforms such as Reddit have emerged as key infrastructures for navigating psychological problems, particularly for individuals who may not have access to, or trust in, formal health systems [14]. The appeal of these spaces is amplified by the limitations of traditional care: clinical services often prioritize severe cases, socioeconomic barriers restrict access, and not all individuals are equally willing or able to seek professional help. Even when accessed, brief and episodic consultations may leave little room to capture the full scope of ongoing psychological struggles [15-17]. By contrast, the anonymity and decentralization of digital health communities enable frank discussions of stigmatized experiences, which allows users to articulate symptoms and explore possible explanations. In doing so, these platforms contribute to a contemporary, living discourse around mental health.

Because users can engage freely, anonymously, and repeatedly over time, digital health communities provide a naturally occurring data source for examining both the expression of diverse conditions and their interrelations [18,19]. While research on online platforms has typically emphasized diagnostic tools or peer support within isolated communities or narrow disorder sets [20-22], recent cross-community studies have started studying online mental health communities through comparative and cross-community perspectives. For example, Morini et al [23] analyzed the content of 67 mental health communities on Reddit and showed that support-seeking and venting are dominant posting intents, and that community feedback shapes subsequent participation. In parallel, Jin and Zhu [24] constructed a multimorbidity network linking diabetes communities to 88 other disease subreddits, revealing connections to mental health and weight-management forums and showing that discussion of physical illness can extend into mental health–focused communities. Such cross-community studies have the potential to complement traditional research and open the way for examining large-scale, system-level patterns of psychopathology and the diagnostic frameworks that represent them.

Empirical research on diagnostic frameworks and relationships between disorders has largely relied on 2 sources: surveys and clinical registers. Surveys capture subsets of symptoms in specific populations and depend on self-report, which is difficult to validate at scale [25-27]. On the other hand, while clinical registers are based on verified diagnoses, they typically contain biases based on severity, health care access and even diagnostic conventions [28,29]. As such, both perspectives offer important yet partial views on the structure of psychopathology. We position digital mental health communities, specifically Reddit, as a complementary source that provides behavioral signals of perceived relatedness among mental disorders. These data are not confined to small samples or predefined diagnostic categories, and they are not restricted to clinically severe cases. Instead, they capture large-scale, naturally occurring expressions of personal experience that cannot replace survey or register approaches, yet can broaden the empirical basis for psychopathology research and enable triangulation across complementary sources.

XSL•FO
RenderX

We study the structure of psychopathology on Reddit by analyzing users' coposting activity in condition-specific mental health communities. The dataset comprises 114 subreddits, each centered on a distinct disorder, collectively covering more than half a million users and over 1.5 million posts. By tracing associations across disorders through shared patterns of coposting, we adopt a data-driven network perspective on how conditions are interconnected in contemporary contexts. This perspective draws on tools from network analysis that have grown influential in mental health research, particularly network psychometrics, where disorders are modeled as systems of interacting symptoms rather than latent disease categories [30-32]. Whereas network psychometrics emphasizes within-disorder architecture, we shift the focus to relationships between disorders as reflected in cross-community engagement. In doing so, we complement symptom-level approaches of psychopathology with a view of how people navigate multiple diagnostic ideas at once, negotiating meaning and seeking support across disorder boundaries.

Our main objective is to infer a significance-based network of associations between mental health conditions as expressed through online engagement. We identify pairs that co-occur more or less often than expected, as well as mental health conditions that act as high-degree bridges across diagnostic categories. We highlight clusters of co-occurring disorders and examine strong cross-category connections that may signal transdiagnostic roles not anticipated by existing taxonomies. Finally, we compare the Reddit-derived structure with the hierarchy encoded in the *ICD-10* (*International Statistical Classification of Diseases, 10th Revision*) diagnostic criteria, outlining conceptual and structural differences that contribute to the broader discussion on psychiatric nosology. In this way, Reddit functions as both a site of peer support and a window into how people collectively make sense of mental health beyond formal clinical narratives.

Methods

Study Design

The study design centers on the structure of user coposting across disorder-specific subreddits as a behavioral proxy for latent relationships between mental health disorders. We treat statistically significant user overlap between subreddit pairs as evidence that seeking advice for one disorder is associated with seeking advice for another, indicating that the 2 disorders are conceptually proximate. Such proximity may rise from several mechanisms, such as comorbidity, where users experience or suspect multiple concurrent disorders; diagnostic progression, where users transition between diagnoses over time; and misdiagnosis, where users reconsider or question a diagnosis, whether self-identified or clinically provided. Although some overlap could reflect general cross-community engagement, previous research shows that most mental health content on Reddit centers on first-person struggles rather than general discussion [14,23]. Our own validation supports this pattern: more than two-thirds of posts contain more self-referential pronouns ("I," "me," and "myself") than references to others (refer to Multimedia Appendix 1), consistent with help-seeking grounded in personal experience. To limit residual unrelated activity, we infer edges only when observed user overlap exceeds expectations under a conservative bipartite configuration model with stringent multiple-comparison control (refer to the "Network Inference" subheading below). As a final major design choice, we decided to focus on activity through posts rather than comments. Posts in disorder-specific communities are the primary venue for sharing experiences and seeking help, whereas comments are more reactive, often offering advice or feedback. While comments provide valuable perspectives on peer interaction and information diffusion, their scale and heterogeneity risk diluting the signal of the help-seeking behavior we aim to capture. Together, these design choices justify interpreting coposting as a cautious but credible signal of perceived relatedness between disorders without interpreting activity as evidence of diagnosis or medical history.

Data

We collected Reddit data through the Pushshift application programming interface by looking into the 20,000 biggest subreddits [33,34]. From this corpus, we manually curated a list of subreddits whose primary focus is to provide information, support, and shared lived experience related to specific mental health conditions. Each candidate subreddit was reviewed individually, considering both its description and posted content, and was included only if those corresponded to a distinct disorder as described in the *ICD-10* (2019 release) [35].

Each included subreddit was annotated according to its corresponding *ICD-10* diagnostic code at level 4 granularity (eg, F48.1, depersonalization-derealization syndrome). To study the complete hierarchy of psychopathology, we further annotated each community to higher-level categories of the *ICD* taxonomy, including level 3 codes (eg, F48, other neurotic disorders) and level 2 codes (eg, F4, neurotic, stress-related, and somatoform disorders). For completeness and future interoperability, we also provided mappings to the equivalent codes in the *ICD-11* (*International Classification of Diseases, 11th Revision*), a taxonomy that is yet to be used in practice. In total, 114 condition-specific mental health subreddits were identified and classified into 49 unique *ICD-10* disorders, covering 9 level 2 diagnostic categories of mental and behavioral disorders (from F0 to F9). A full list of annotated subreddits and their hierarchical coding is provided in Multimedia Appendix 2 and is publicly available for reuse, along with a separate table listing all disorders and their corresponding *ICD-10* codes.

We restricted our analyses to posts from 2022 to ensure temporal consistency and avoid confounding effects from major platform-level disruptions. This includes discontinuities associated with the COVID-19 pandemic and subsequent policy or moderation shifts, as well as more recent artifacts linked to the rise of generative artificial intelligence [36,37]. The resulting dataset comprised 1,513,016 posts authored by 545,330 unique users across all included mental health subreddits. On average, each user contributed 2.77 posts, with 96,742 (17.74%) users posting in more than 1 mental health subreddit, which forms the foundation for our analysis of disorder co-occurrence (refer to Multimedia Appendix 3 regarding the distribution of posts across disorder categories and subreddits). To enhance data

XSL•FO
RenderX

quality, we excluded accounts indicative of automated or spam-like activity. Specifically, the dataset excluded users who posted more than 365 times in the study year (corresponding to a rate of more than 1 post per day), as well as known automated accounts listed in the publicly available bot directory BotRank [38]. Ultimately, the data collection procedure resulted in a comprehensive representation of Reddit's mental health discourse by systematically covering the major condition-specific communities active on the platform.

Network Inference

Overview

The main objective of this work is to infer relationships between mental health conditions as they emerge from patterns of shared user participation across the studied disorder-related subreddits. To this end, we constructed a weighted network in which each node represents 1 of the 49 classified mental health disorders, defined at the level 4 granularity of *ICD-10* codes, and edges capture statistically significant user coposting overlaps between the sets of subreddits corresponding to each disorder pair. The construction involved 4 key steps: computing user overlap between disorder pairs, estimating a null model for expected coposting, determining statistical significance, and assigning edge weights based on deviation from null expectations.

Disorder Association Metric

For each pair of disorders, we measured the strength of coposting association by calculating the overlap coefficient between the sets of users who contributed to subreddits of each corresponding disorder. This coefficient captures the size of the user intersection normalized by the smaller of the 2 user sets:



where x and y denote the sets of unique users who posted in the subreddits associated with each disorder.

The overlap coefficient is particularly suited for capturing asymmetric coengagement. By measuring the proportion of shared users relative to the smaller community, we can address cases where participation in one subreddit could be almost entirely embedded within another. This property aligns well with the hierarchical and overlapping conceptualization of many mental health conditions, where narrower or less prevalent disorders often exist within the broader spectrum of more common ones.

Null Model and Statistical Testing

To assess the significance of observed user overlaps, we used a binary bipartite configuration model as the null model. The user-disorder bipartite network was constructed with one set of nodes representing users and the other representing mental health disorders, where an edge denotes a post by a user in a subreddit labeled with a specific disorder. To generate the null distribution, we randomly rewired the bipartite network 10,000 times while preserving the degree distributions of both users and disorders and preventing multiedge cases. Each rewired network underwent several edge swaps equal to 10× the total number of edges, ensuring sufficient randomization while maintaining the original participation heterogeneity. This approach follows established practices in bipartite network analysis, where degree-preserving randomization is commonly used to construct realistic null models for statistical inference [39,40].

For each disorder pair, we computed the distribution of overlap coefficients across the 10,000 null replicates and evaluated the statistical deviation of the observed overlap using a standard z score. To correct for multiple comparisons across all possible disorder pairs ($n \times [n-1]/2$), we applied a Bonferroni-corrected significance threshold of $P<.001$ [41]. Based on this, each disorder pair falls into one of 3 categories:

- Positive association: observed overlap is significantly higher than expected.
- Negative association: observed overlap is significantly lower than expected.
- No evidence for association: observed overlap does not differ significantly from null expectations.

Edge Weights and Network Representation

For each statistically significant association, we assigned a weight equal to the difference between the observed and the mean expected overlap coefficient. This measure reflects the magnitude of deviation from the null, with higher values indicating stronger-than-expected co-occurring between disorders. Theoretically, the weights range from 0 (no deviation) to 1, with larger values signaling greater empirical association strength relative to what would be expected by chance under the null model.

The resulting structure was represented as two undirected weighted networks of positive and negative disorder associations. This network representation enabled intuitive interpretation of pairwise association patterns, capturing the relational landscape of psychopathology. An interactive online version of the network visualization was also developed to facilitate further exploration (refer to Multimedia Appendices 4 and 5) [42-44].

Node-Level Metrics

To characterize the centrality of disorders within the inferred network, we computed 2 measures of connectedness:

- Unweighted degree: the total number of significant associations (edges) for a given disorder. This captures how broadly a condition is connected across the mental health landscape.
- Weighted degree: the sum of edge weights for all associations of a disorder. This emphasizes the cumulative strength of its connections, highlighting conditions that might not have many links, yet maintain particularly strong associations.

Across analyses, we used both unweighted and weighted edges depending on the methodological requirements of each approach. Comprehensive data for both unweighted and weighted node degrees of the Reddit network are provided in Multimedia Appendix 2.

XSL•FO
RenderX

ICD-10 Network of Diagnostic Criteria

As a reference point for the Reddit-derived association network, we constructed a network of mental disorders based on formal diagnostic criteria, referred to as the *ICD-10* diagnostic criteria network. This network was built using data curated by Tio et al [45], who systematically extracted diagnostic symptoms for each *ICD-10* code falling into Chapter F: mental and behavioral disorders. This dataset provides a standardized operationalization of disorder-level symptom profiles for *ICD-10* codes that remain in clinical use [46]. It offers a unique opportunity to analyze relationships between disorders grounded in clinical definitions, which are otherwise difficult to access in a machine-readable or systematically coded form. As such, we used this network for comparing the *ICD-10* diagnostic structure with the association patterns inferred from the coposting activity on Reddit.

In the *ICD-10* diagnostic criteria network, each node (disorder) was represented as a set of diagnostic criteria, and the edge weight (strength of connection) between 2 disorders was calculated as the overlap coefficient between their diagnostic criteria sets. This formulation mirrors our approach to the Reddit network, enabling a consistent comparison across both systems. To ensure a valid basis for comparison, we restricted this network to include only those disorders for which a corresponding Reddit community had been identified in our manual curation.

Hierarchical Clustering of Association Networks

We used hierarchical clustering to study the potential modular and hierarchical structure of the association network derived from Reddit, as well as that of the *ICD-10* diagnostic criteria network. This approach, which has previously been used to uncover grouping patterns in association networks [47-49], allowed us to infer latent groupings of mental health conditions based on observed similarity patterns. Importantly, it enabled a system-level comparison of the 2 networks, moving beyond pairwise overlap to examine how broader patterns of connectivity and disorder organization differ between Reddit discourse and the formal *ICD-10* diagnostic structure, akin to comparing hierarchical network communities.

We applied standard agglomerative clustering with average linkage, where 2 clusters were merged based on the average distance between all pairs of disorders across the 2 clusters [50]. Since the weighted edges in both networks represent similarity (rather than distance), we defined the distance between any 2 disorders x and y as:

$$d(x,y) = max(e(x,y)) - e(x,y)$$

Where $e(x,y)$ is the observed similarity (edge weight) and $max(e)$ ensures that distances are positive and properly scaled for clustering.

To assess the extent to which each network exhibits clustered structure, we first computed the weighted modularity, a standard quality function that quantifies how much edge weight lies within clusters compared to between clusters, relative to the expectation under a weighted degree-preserving null model [51]. This allowed us to evaluate and compare the overall modular organization of the Reddit network and the *ICD-10* diagnostic criteria network.

For a clustering π (nodes grouped into clusters), the weighted modularity is defined as:



Where $w_{i,j}$ is the observed edge weight, $s_i=\sum_j w_{i,j}$ is the weighted degree of node i, $2W=\sum_{ij} w_{ij}$ is the total edge weight, and $1\{\cdot\}$ equals 1 when i and j are assigned to the same cluster. Intuitively, higher weighted modularity values indicate a clearer community structure. Q_w is large when within-cluster connections carry more weight than expected under a null model preserving node strengths.

Since raw modularity depends on network density and degree or strength heterogeneity, we estimated the normalized modularity for each network under a degree-sequence–preserving null model to enable a fair comparison between the Reddit network and the *ICD-10* network of diagnostic criteria. For each network, we first generated an ensemble of randomized topologies by repeatedly swapping edge pairs and their associated weights, thereby preserving the overall degree sequence and weight distribution while randomizing the network. Then, for each randomized graph G', we evaluated $Q_w(G',\pi(\tau))$ on the same partition $\pi(\tau)$ obtained from the observed network; this isolates how expected the observed within-cluster concentration of weight is, given the degree and weight distribution.

We then calculated the normalized modularity, defined as the difference between the observed modularity and the expected value under the null model:

$$\Delta Q_w = Q_w(G,\pi(\tau)) - E[Q_w(G',\pi(\tau))]$$

where the latter was estimated from 1000 randomized realizations. Finally, for each dendrogram height produced through hierarchical clustering, we evaluated the normalized weighted modularity and compared the Reddit network with the *ICD-10* network of diagnostic criteria, focusing on the clustering results corresponding to the dendrogram cuts with the highest normalized weighted modularity in each network. In other words, the final clustering of a network is obtained as the cut τ that maximizes ΔQ_w across all admissible cuts:



To evaluate the similarity between the final clusters of the 2 networks, we used 2 standard comparison indices: the Adjusted Rand Index (ARI) and the normalized mutual information (NMI) [52]. ARI measures the agreement between 2 clustering results by quantifying how often pairs of nodes are grouped or separated in the same way, adjusted for chance. NMI captures the amount of shared information between the clusters of the networks and is less sensitive to differences in the number or size of clusters. Both metrics range from 0 (no agreement beyond chance) to 1 (perfect correspondence).

XSL•FO
RenderX

Ethical Considerations

The subreddits analyzed in this study are publicly accessible and do not require login credentials. Posts are shared under pseudonymous accounts, and all usernames were further pseudoanonymized before analysis to protect privacy. Given the sensitive and personal nature of the disclosures that may appear in these vulnerable communities, we applied strict privacy safeguards and limited all reporting to aggregated results, without focusing on individual cases and basing our analysis only on coposting rather than the content of the posts themselves. The study was purely observational, involving no interaction or intervention with users. We did not attempt to contact individuals, and no analyses were conducted that could enable reidentification of participants. Our approach followed widely recognized ethical frameworks for internet-mediated research [53-56]. In particular, we adhered to the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2), which emphasizes proportional safeguards for minimizing risks when analyzing data from public platforms where individuals can reasonably expect to be observed without explicit consent [57]. This research was also reviewed and approved by the Department of Network and Data Science Ethical Research Committee at the Central European University (reference no 2024-2025/16/ RD/DNDS).

Results

The Reddit Network of Psychopathology

We constructed a data-driven network of mental health disorder associations based on user coposting behavior across 114 Reddit communities. Each node in the network corresponds to a disorder-level *ICD-10* code (Chapter F), and edges represent statistically significant coposting links, derived from observed versus expected user overlap.

The derived Reddit network of positive associations consists of 49 nodes and 159 edges, with a density of 0.135. Despite this relative sparsity, the network forms a single giant component encompassing 43 of the 49 disorders, exemplifying the interconnectedness of mental health conditions. Only 3 disorders remained isolated without any associations, indicating a lack of shared user engagement with other conditions. These isolated disorders are F00-F03 (dementia), F63.0 (gambling addiction), and F98.5 (stuttering).

Figure 1 (top panel) shows a circular layout of the Reddit network. Node size corresponds to the total number of users active in that disorder's subreddits, and node color denotes its *ICD-10* diagnostic category. The disorders with the highest number of users include F10 (alcohol addiction), F32-F33 (depressive episodes and recurrent depression), F63.8 (other habit and impulse disorders, including excessive masturbation and pornography addiction), F84 (pervasive developmental disorders in *ICD-10*; termed autism spectrum disorder in *ICD-11*), and F90.0 (hyperkinetic disorders in *ICD-10*; termed attention-deficit/hyperactivity disorder [ADHD] in *ICD-11*). While the distribution of user activity is heterogeneous, we observed no significant association between the number of users engaging with a given disorder and the number of connections in the network. This lack of association between volume and connectivity supports the robustness of the inference method, suggesting that network centrality reflects patterns of coposting that cannot be explained simply by subreddit size (refer to Multimedia Appendix 6 for correlation results between degree and volume).

XSL•FO
RenderX

Figure 1. Network of positive associations among disorders derived from Reddit coposting activity in 2022, covering 114 condition-specific communities (subreddits) mapped to ICD-10 (International Classification of Diseases, 10th Revision) Chapter F (Mental and Behavioral Disorders). The network includes 49 disorder nodes linked by 159 edges. Top: a circular (radial) network visualization. Each node represents a distinct ICD-10 disorder, grouped and color-coded by higher-level diagnostic category (F1-F9). Node size indicates the number of unique users who posted at least once in the corresponding subreddits. Edges connect pairs of disorders that share a statistically significant overlap in user activity, estimated with a bipartite configuration null model and corrected for multiple testing (Bonferroni-adjusted, $P<.001$). Edge density and cross-category links illustrate the extent of interconnectedness across diagnostic boundaries. Bottom: disorder-level statistics. The center panel shows the number of users associated with each disorder. The bottom panel shows the degree of each disorder, defined as the number of significant co-occurring associations, broken down into intracategory associations (links within the same ICD-10 level-2 group) and intercategory associations (links connecting disorders across different ICD-10 groups; eg, between F1 and F5). Colored bars highlight the maximum values within each diagnostic category. Together, these panels summarize both the structure of the Reddit-based disorder network and the relative prominence of individual disorders by participation size and connectivity.

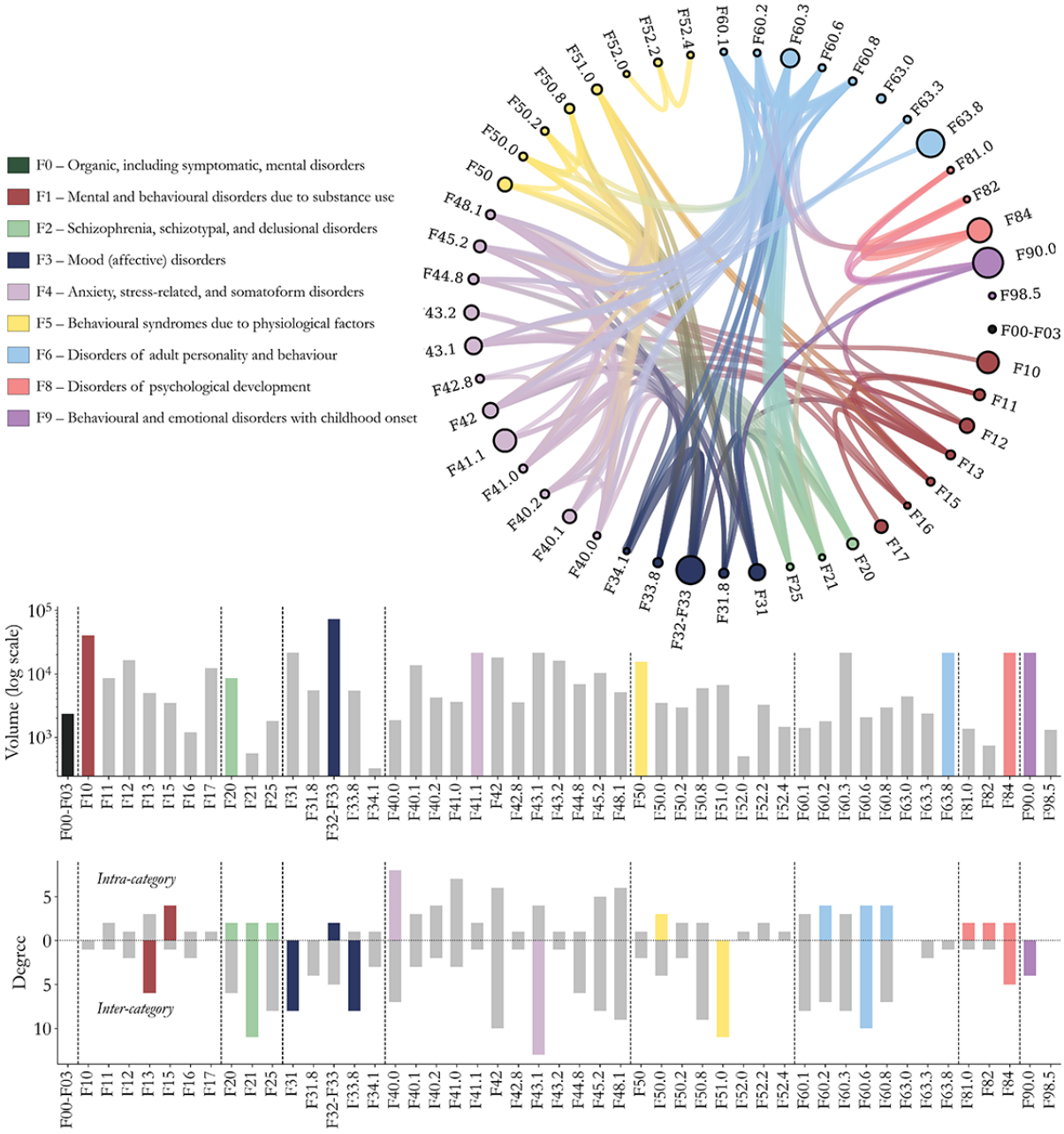

Figure 1 (bottom panel) shows the degree distribution of nodes in the network, revealing substantial heterogeneity in how strongly different disorders are connected, ranging from highly linked hubs to sparsely connected nodes (refer to Multimedia Appendix 2 for results on weighted degrees). The most connected disorders include F43.1 (posttraumatic stress disorder [PTSD]), F42 (obsessive-compulsive disorder), F40.0 (agoraphobia), F48.1 (depersonalization-derealization syndrome), and F60.6 (avoidant personality disorder). These disorders emerge as transdiagnostic hubs, a concept central to psychiatric nosology, with associations spanning a wide range of other conditions. In contrast, disorders with minimal connectivity, such as F10 (alcohol addiction), F17 (tobacco addiction), and F52.0 (lack or loss of sexual desire), may reflect more segmented or marginalized user groups, narrower community focus, or greater diagnostic specificity, similar to the 3 isolated disorders identified before in this subsection.

XSL•FO
RenderX

The network of positive associations reveals a strong presence of cross-category links, with 106 of 159 edges connecting disorders classified in different *ICD-10* diagnostic categories. Notably, some disorders show markedly higher intercategory associations than intracategory ones. For example, F43.1 (PTSD), F51.0 (nonorganic insomnia), and F33.8 (used here to denote premenstrual dysphoric disorder) exhibit the largest discrepancies, with strong cross-category ties but weak integration within their respective *ICD-10* groups. This pattern also extends to entire higher-level diagnostic categories: all conditions in F2 (schizophrenia, schizotypal, and delusional disorders) and F3 (mood and affective disorders) display more intercategory than intracategory connections, highlighting particularly fluid boundaries for these diagnostic categories.

Negative Associations

While our primary focus was on positive associations, indicative of shared user bases and potential comorbidity patterns, we also identified a set of negative associations, where user coposting occurred less frequently than expected under the null model. However, due to methodological limitations, such results should be interpreted at the level of individual edges only, not aggregated across nodes. Specifically, the inference strategy using a null model and the overlap coefficient to measure association becomes increasingly insensitive to low coposting rates in smaller or less active subreddits, introducing a lower-bound bias that distorts node-level summaries of negative associations (refer to Multimedia Appendix 7).

Despite this limitation, a consistent pattern emerges. Negative associations were most commonly observed between impulse-related disorders (eg, F10 for alcohol use, F11 for opioid use, and F63.8 for other habit and impulse disorders) and other mental health categories. These patterns may reflect distinct user populations, stigma-driven disengagement, or divergent framings of psychological distress and its management. The presence of such negative ties underscores that disorder communities on Reddit are not only interconnected but also socially and discursively fragmented in ways that do not always align with theoretical similarity and comorbidity reported in the literature. A notable example is F11 (opioid use), which showed negative associations with 7 of 12 nodes in the F4 category (neurotic, stress-related, and somatoform disorders), despite previous research suggesting strong links between anxiety and opioid use [58,59]. A full list of negative edges and their weights is provided in Multimedia Appendix 2. Considering the limitations of our analysis on negative associations, these findings warrant further investigation using alternative methodological approaches better suited to capture negative associations in smaller samples.

The Hierarchical Structure of Psychopathology: Comparison With ICD-10 Diagnostic Criteria

To better understand the structural logic underlying user-inferred relationships between mental health conditions, we examined the emergent hierarchical organization of the Reddit coposting network and compared it with a clinically derived alternative based on diagnostic criteria overlap. While pairwise associations are informative, a hierarchical view enables assessment of whether larger clusters of disorders emerge in user behavior and how these clusters compare with the taxonomy in the *ICD-10* system.

Figure 2 (top) presents both the Reddit and diagnostic criteria–based networks using circular layouts. Both networks contain the same set of 49 *ICD-10* codes but differ in how connections are formed: the Reddit network uses statistically significant coposting links based on user overlap, while the diagnostic criteria network connects disorders based on shared clinical features, as curated by Tio et al [45]. Edges in both networks are weighted by the overlap coefficient, which quantifies the proportion of shared elements between 2 sets relative to the smaller set.

XSL•FO
RenderX

Figure 2. Comparison between 2 networks of mental health disorder relationships: 1 derived from Reddit user activity and 1 constructed from clinical diagnostic definitions, illustrating both overlaps and divergences between lived-experience associations emerging from online peer support and the formal structures encoded in psychiatric nosology. Top panels: circular network visualizations. The network on the left is based on coposting behavior across 128 condition-specific Reddit communities (subreddits) active in 2022, mapped to International Classification of Diseases, 10th Revision (ICD-10) Chapter F (Mental and Behavioral Disorders). Two disorders are linked if users post in both corresponding subreddits more often than expected by chance. The network on the right is the diagnostic criteria network, based on a curated list of disorders and diagnostic criteria of ICD-10, where disorders are linked if they share at least 1 formal diagnostic feature. In both networks, node size reflects connectivity (degree), and color indicates the ICD-10 diagnostic category (F0-F9). Black edges highlight associations that are common to both networks, while colored edges are unique to either Reddit (orange) or the diagnostic criteria (blue). Bottom panels: hierarchical clustering of disorders in both networks. Agglomerative clustering with average linkage was applied to identify higher-level clusters, with the cut-off chosen to maximize normalized weighted modularity (a measure of how well the network separates into cohesive clusters beyond chance). The resulting normalized weighted modularity values were ΔQ_w=0.420 for Reddit and ΔQ_w=0.521 for the diagnostic criteria network, showing that the Reddit network is more interconnected across diagnostic categories compared with the more modular diagnostic criteria–based network. Cluster labels are simplified, human-readable summaries of underlying ICD-10 codes. ADHD: attention-deficit/hyperactivity disorder; ARI: Adjusted Rand Index; NMI: normalized mutual information; OCD: obsessive-compulsive disorder; PTSD: posttraumatic stress disorder.

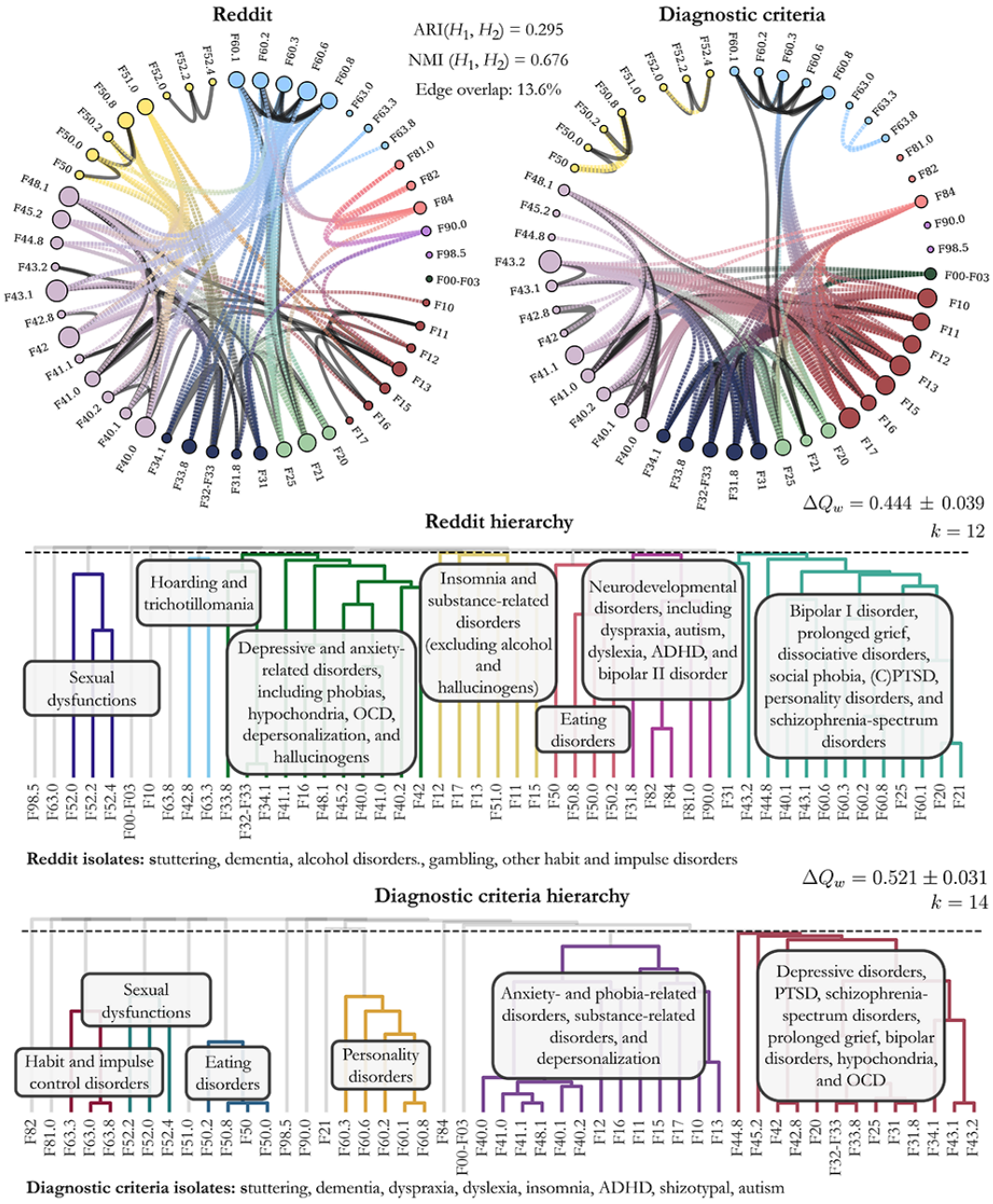

We applied agglomerative clustering with average linkage to both networks and identified the optimal clustering by cutting the dendrogram at the level of maximum normalized weighted modularity (ΔQ_{max}). The resulting modularity scores were 0.444 (12 clusters) for Reddit and 0.521 (14 clusters) for the diagnostic criteria network (Figure 2, bottom). This finding suggests that the Reddit network is less modular, reflecting less distinct clusters and more overlapping patterns of association compared with the more compartmentalized structure based on symptom overlap.

XSL•FO
RenderX

Overall, the comparison of the 2 hierarchies showed partial alignment between Reddit-based coposting behaviors and the formal diagnostic structure, with areas of both convergence and divergence. Only 13% of links were shared between the 2 networks, highlighting a substantial dissimilarity in their underlying associations. This limited edge overlap suggests that user-inferred connections diverge notably from those based on diagnostic criteria. However, when comparing their hierarchical structure, we observe more alignment. The ARI of 0.295 and the NMI of 0.676 indicate moderate similarity between the clusters of the 2 networks, pointing to partial similarity in how disorders are categorized despite the underlying differences in pairwise associations.

The Reddit network also presents a more convoluted structure, with only 5 disorders remaining unclustered at the modularity-optimal cut: F98.5 (stuttering), F00-F03 (dementia), F10 (alcohol addiction), F63.0 (gambling addiction), and F63.8 (other habit and impulse disorders). In contrast, the diagnostic criteria network leaves 8 disorders unclustered, with only 2 repeated from the Reddit network (stuttering and dementia). Importantly, some of the conditions that remain unclustered in the *ICD-10*–based network occupy more central positions in the hierarchy of the Reddit network. These include F21 (schizotypal disorder), F43.1 (PTSD), F44.81 (dissociative identity disorder), F84 (autism spectrum), and F90.0 (ADHD). Although still considered peripheral in traditional diagnostic systems, several of these conditions have received growing attention in recent years, either for their proposed transdiagnostic relevance [60,61] or for their apparent connection to broader technological and social changes, including increased screen exposure and digital media use [62-64]. Their presence in Reddit-based clusters may indicate that users organize mental health experiences in ways that diverge from formal diagnostic structures, shaped instead by evolving discussions around neurodivergence, trauma, and identity.

Some of the clusters in the Reddit hierarchy echo broader dimensional models of psychopathology. For example, F32-F33 (depressive disorders), F41.0-F41.1 (anxiety disorders), and F42 (obsessive-compulsive disorder) form a prominent cluster, an intersection that is absent from the *ICD-10* diagnostic structure but aligns with internalizing spectra described in dimensional models [10]. Another Reddit-derived cluster connects psychotic conditions such as F20 (schizophrenia) and F25 (schizoaffective disorder) with trauma-related disorders such as F43.1 (PTSD) and F44.81 (dissociative identity disorder), as well as personality disorders such as F60.3 (emotionally unstable personality disorder). This forms a cluster of thought disorders situated at the boundary between conditions typically conceptualized as internalizing or externalizing.

Discussion

Principal Findings and Relevance to Psychopathology Research

To examine how anonymous users collectively make sense of mental health problems and how this organization relates to clinical diagnostic manuals, we analyzed activity in 114 disorder-focused Reddit communities involving 545,000 users and more than 1.5M posts. We inferred a significance-based network of associations among 49 disorders spanning 9 *ICD-10* mental and behavioral disorder categories (F0-F9), derived from patterns of coposting across communities. We then compared this Reddit-based structure with a network constructed from overlaps in diagnostic criteria defined in the *ICD-10*. The Reddit network revealed a highly interconnected organization that crossed traditional diagnostic boundaries. Several disorders occupied central positions, acting as hubs that linked otherwise distinct diagnostic categories, while others, particularly substance- and behavior-related addictions, appeared less integrated into the broader mental health discourse. Comparisons of the hierarchical structure showed only partial correspondence between the Reddit network and *ICD-10*, indicating that patterns of association emerging from lived experience differ in systematic ways from those encoded in formal taxonomies. Viewed in this large-scale context, digital mental health communities do not simply mirror existing diagnostic structures. Instead, they organize mental distress through shared experience and collective interpretation, producing a socially situated view of how disorders relate to one another. This perspective extends beyond what can be captured through surveys or clinical registers alone and can help refine how comorbidity patterns and disorder boundaries are understood outside formal clinical settings.

In more detail, the statistical inference of the Reddit network revealed 159 associations between the 49 studied disorders. Despite its moderate density (0.135), the network formed a large, connected component that encompassed 43 of the 49 conditions, indicating that most conditions were linked to one another through chains of significant associations that cut across traditional diagnostic categories. Complementary patterns have been reported in symptom-level network research, where psychopathology appears as a highly interconnected system rather than clearly separated clusters [65].

Further looking into the structure of the Reddit network, several disorders emerged as central transdiagnostic hubs, including F43.1 (PTSD), F42 (obsessive-compulsive disorder), F40.0 (agoraphobia), F48.1 (depersonalization-derealization syndrome), and F60.6 (avoidant personality disorder). Their high centrality reflects symptoms that cut across diagnostic boundaries, such as intrusive thoughts, avoidance, and disturbances of identity. Addressing these transdiagnostic overlapping symptoms may therefore be key for treatment and intervention, particularly in younger populations where online peer support heavily shapes help-seeking behavior.

When considering negative associations, the Reddit network revealed specific points of disconnect between certain disorders and the broader network. Disorders related to substance use and behavioral addictions consistently appeared underconnected or negatively associated with other mental health conditions. This is in sharp contrast with comorbidity estimates typically reported in previous research, according to which substance use disorders often co-occur with other mental health conditions. For example, population-based estimates indicate that roughly 1 in 4 individuals with a substance use disorder have a comorbid mental disorder [66,67]. Our contrasting results might reflect a tendency of individuals in these communities to focus more

XSL•FO
RenderX

narrowly on managing acute behavioral symptoms or crises rather than looking into their mental health more broadly. Their relative isolation may also stem from prevailing stigmatization and limited self-recognition or acknowledgment of co-occurring mental health issues, which together tend to position externalizing disorders outside the domain of conventional psychological problems [68-70]. Whether driven by a narrow focus on symptom management, social framing, or self-stigma, these isolating mechanisms risk reinforcing silos in both peer support and clinical care, obscuring potential links between addiction and other forms of psychopathology and making it more difficult to approach treatment holistically.

Beyond overall centrality, some disorders showed distinct patterns of connectivity, forming disproportionately strong links across the F code diagnostic categories while remaining relatively weakly integrated within their own. These bridging conditions included F43.1 (PTSD), F51.0 (nonorganic insomnia), and F33.8 (used to denote premenstrual dysphoric disorder here). They illustrate transdiagnostic mechanisms that current categorical frameworks do not explicitly capture, whether through hormonally linked mood dysregulation, sleep disturbances that cut across almost all clinical categories, or trauma-related symptoms that span affective, anxiety, dissociative, and personality domains [71,72]. Notably, this pattern extended beyond individual conditions to entire diagnostic categories: all disorders in F2 (schizophrenia, schizotypal, and delusional disorders), F3 (mood and affective disorders), and the F60 subcategory (personality disorders) displayed more intercategory than intracategory links. Such patterns suggest particularly permeable diagnostic boundaries for these categories, a result that has also been observed at the level of diagnostic criteria [73].

The looseness of diagnostic boundaries was also observed through the comparative analysis between the Reddit network and the *ICD-10* network based on diagnostic criteria. The Reddit network displayed only slightly higher intercategory connectivity than the *ICD* diagnostic criteria network (68% vs 65% of all edges). However, the comparative hierarchical clustering analysis revealed that the Reddit network produced a substantially different hierarchy of disorders that only partially aligns with *ICD-10* diagnostic structures. The 2 networks showed only low to moderate similarity in their clustering (ARI=0.295 and NMI=0.676), with only 13% of edges present in both networks. Reddit also exhibited lower modularity than *ICD-10*, meaning that these clusters were less separated by diagnostic category and more interconnected across them (normalized weighted modularity of 0.444 compared to 0.521). In addition, key divergences emerged at the level of disorders. Conditions such as F21 (schizotypal disorder), F84 (autism spectrum), and F90.0 (ADHD) were central and well-integrated within the Reddit network but did not form cohesive clusters within the *ICD*-based hierarchy of diagnostic criteria. Qualitative analyses of Reddit discussions similarly report tensions between lay and professional expertise [74], while quantitative comparisons indicate that certain conditions (such as anxiety-related and affective disorders) are disproportionately represented relative to registry data [75]. Such discrepancies may stem from the platform's affordances of anonymity, its demographic composition, and its emphasis on peer support, but they may also signal blind spots in clinical frameworks. However, rather than contradicting established evidence, these divergences show how online data can surface experiential transdiagnostic mechanisms that remain underrepresented within formal diagnostic systems.

Despite clear differences between the Reddit network and the ICD-based network of diagnostic criteria, both point to the difficulty of fitting mental disorders into rigid, discrete categories [76,77]. While neither network should be treated as a ground truth of interdisorder relationships, the interconnected structure observed in both aligns with longstanding critiques that current psychiatric nosology underestimates the interconnected nature of psychopathology [78-81] and supports discussions of alternative paradigms, such as dimensional models that emphasize broad spectra and shared underlying features [82].

Limitations and Future Directions

While our findings provide a large-scale view of the interconnected structure of psychopathology through mental health communities, several limitations should also be acknowledged. Our analysis is based on Reddit, a platform whose user base skews toward younger, Western, male, and digitally literate populations, which limits generalizability [83,84]. Reddit-specific dynamics such as anonymity, community norms, and moderation practices also shape what is shared and who participates. These features increase accessibility and ease of self-disclosure, but they also raise validity concerns, including account transience and the use of throwaway accounts that make authenticity difficult to assess [14,85]. Previous work also suggests that individuals with broader lay concepts of disorders are more likely to self-diagnose [86], which may amplify certain demographic biases. While these factors may contribute to divergences between the Reddit-derived and clinical structures, future work should clarify whether they reflect robust disorder associations or platform-specific outcomes. Hence, observed results should not be interpreted as verified comorbidities, but rather as behavioral signals that approximate perceived relatedness between disorders within digital contexts.

The methodological decisions delimit the scope of our results. We analyze posts, but not comments, made in 2022 within mental health support subreddits, and include only communities that map to a distinct disorder category in the *ICD-10*. Focusing on posts aligns with our aim to capture self-disclosure and help-seeking at the point of initiation, though it necessarily excludes peer interaction and information diffusion occurring in comment threads. Restricting to *ICD*-mapped subreddits increases construct clarity but may underrepresent transdiagnostic communities (such as r/MentalHealthSupport) or symptom-focused communities (such as r/SuicideWatch). Limiting the data to 2022 reduces temporal confounding from earlier structural instability, demographic shifts, and disruptions during the COVID-19 pandemic, and it ensures comparability across subreddits within a single, more mature phase of the platform. However, this temporal focus also constrains interpretation: network patterns and discourse are dynamic, and

XSL•FO
RenderX

analyses spanning multiple years could reveal different structures as community composition and cultural context evolve. The year 2022 was chosen to provide a stable and interpretable baseline, but not to imply that the resulting associations are fixed over time.

Looking ahead, several promising directions emerge for future research. Reddit data offers a powerful proxy for examining mental health from multiple perspectives. Incorporating replies would add a complementary interaction layer, as comment threads reveal who engages with whom, what types of support are exchanged (eg, validation or advice), and how these interactions relate to subsequent posting trajectories. Temporal analyses could further enrich this view on 2 fronts. At the societal level, they would capture how mental health concepts and discourse evolve with cultural and technological change. At the individual level, following users over time could help distinguish between comorbidity, diagnostic progression, and broader help-seeking patterns, adding precision to how disorder associations are interpreted. Extending this approach beyond Reddit to platforms with different affordances and user bases would test how platform design shapes the organization of mental health discourse. We chose *ICD-10* as the reference framework for its international coverage and compatibility with registry data. However, comparable analyses using *DSM-5* (Diagnostic and Statistical Manual of Mental Disorders [Fifth Edition]), the forthcoming *ICD-11*, and other evolving diagnostic systems will be essential to assess psychiatry's continuing efforts toward more coherent and empirically grounded concepts of mental health [11]. Such extensions could further reinforce the value of online mental health data as a bridge between peer discourse and clinical knowledge.

Conclusions

This work maps a large-scale structure of mental health communities as they currently grow outside clinical settings, highlighting the necessity of perspectives that extend beyond formal diagnostic frameworks to achieve a more complete population-level understanding of psychopathology. Diagnostic frameworks remain essential, but they capture only part of how distress is articulated and managed in practice. Outside formal care, people navigate symptoms and negotiate meaning while seeking peer communities that mediate their mental health challenges. Digital platforms such as Reddit have become central to this process: they provide spaces for disclosure and support while also shaping the categories, language, and norms through which psychological distress is understood. In this sense, they are not only mirrors of cultural shifts but also infrastructures that reorganize how mental health is lived and discussed. Neglecting these platforms as legitimate sites of knowledge risks leaving research and practice poorly aligned with mental health needs amid rapid technological change.

Acknowledgments

BE was responsible for data collection, experimental design, analysis, and writing the manuscript. SL and PKN contributed to the initial study conceptualization and provided critical feedback during manuscript revisions. SL assisted in validating the *ICD-10* annotations. All authors reviewed and approved the final version of the manuscript.

Generative artificial intelligence (ChatGPT 5; OpenAI) was used to support linguistic refinement and coherence in the manuscript text. All conceptual contributions, interpretations, and substantive arguments presented in this manuscript are solely those of the authors.

Funding

PKN acknowledges partial funding by the research program Knowledge Technologies (P2-0103). This work is the result of research conducted at Central European University, a private university, with open-access provided through the CEU Open Access Fund.

Data Availability

Subreddit metadata and aggregated data supporting the findings of this study are provided in the corresponding Multimedia Appendices. Due to the platform's terms of service, raw Reddit post content cannot be publicly shared. Additional aggregated data may be made available upon reasonable request by contacting the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1

(Top left) Distribution of the number of self-focused pronouns (I, me, myself) and other-focused pronouns used per post in the selected subreddits, capped at 20 occurrences. The density on the y-axis represents the relative proportion of posts at each pronoun count, rather than raw counts. Posts tend to include more self-focused pronouns (blue) compared to other-focused pronouns (red). (Top right) Overall proportion of posts showing different pronoun balance patterns: more self-focused pronouns (Self > Other), more other-focused pronouns (Other > Self), or balanced/none (Equal/Zero). The majority of posts are self-focused. (Bottom) Pronoun balance patterns stratified by ICD-10 diagnostic categories (Level 2). Each bar represents the proportion of posts within

XSL•FO
RenderX

a diagnostic category that are more self-focused, more other-focused, or balanced. Across nearly all categories, self-focused pronouns dominate, but the relative proportions vary slightly between diagnostic groups.
[PNG File , 490 KB - jmir_v28i1e80958_app1.png]

Multimedia Appendix 2
Supplementary Tables (.xlsx): - ICD-10 Disorder Reference List - Mapping of Subreddits to ICD-10 and ICD-11 - List of (Weighted) Degrees (The Reddit Network of Psychopathology, Positive Associations) - List of Negative Edges and their Weights (Reddit Network of Psychopathology, Negative Associations). - ICD-10 Diagnostic Criteria.
[XLSX File (Microsoft Excel File), 115 KB - jmir_v28i1e80958_app2.xlsx]

Multimedia Appendix 3
Number of posts and subreddits across ICD-10 (International Classification of Diseases, 10th Revision) Chapter F mental health–related categories.
[PNG File , 73 KB - jmir_v28i1e80958_app3.png]

Multimedia Appendix 4
Online Interactive Tool - Map of Associations: Screenshot of the interactive Retina interface showing the node F32–F33 Depression (episodic and recurrent) selected as an example. The left-hand panel displays node-level metrics, including the number of users who posted in the corresponding subreddit(s), the weighted degree (sum of edge weights), the clustering coefficient (indicating local cohesiveness), and the number of triangles (three-node loops) formed around the node. This illustrates how the interface allows for exploratory analysis of individual disorder communities within the broader network structure.
[PNG File , 72 KB - jmir_v28i1e80958_app4.png]

Multimedia Appendix 5
Online Interactive Tool (Map of Disorder Associations): An interactive web-based visualization of the inferred disorder association network, allowing users to explore positive and negative associations, node connectivity, and diagnostic groupings derived from Reddit coposting activity.
[DOCX File , 89 KB - jmir_v28i1e80958_app5.docx]

Multimedia Appendix 6
No evidence of correlation between the number of users active in relation to a specific disorder (node size) and the number of links (node degree) in the Reddit network of psychopathology associations. The result supports the robustness of the inference method, suggesting that node centrality is not driven by subreddit size. Spearman correlation: r=0.068, p=0.64.
[PNG File , 79 KB - jmir_v28i1e80958_app6.png]

Multimedia Appendix 7 [PNG File , 74 KB - jmir_v28i1e80958_app7.png]

References

1. Twenge JM, Gentile B, DeWall CN, Ma D, Lacefield K, Schurtz DR. Birth cohort increases in psychopathology among young Americans, 1938-2007: a cross-temporal meta-analysis of the MMPI. Clin Psychol Rev 2010;30(2):145-154. [doi: 10.1016/j.cpr.2009.10.005] [Medline: 19945203]
2. Twenge JM, Joiner TE, Rogers ML, Martin GN. Increases in depressive symptoms, suicide-related outcomes, and suicide rates among U.S. adolescents after 2010 and links to increased new media screen time. Clinical Psychological Science 2017;6(1):3-17. [doi: 10.1177/2167702617723376]
3. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
4. Sacco R, Camilleri N, Eberhardt J, Umla-Runge K, Newbury-Birch D. A systematic review and meta-analysis on the prevalence of mental disorders among children and adolescents in Europe. Eur Child Adolesc Psychiatry 2024;33(9):2877-2894 [FREE Full text] [doi: 10.1007/s00787-022-02131-2] [Medline: 36581685]
5. Hawes MT, Szenczy AK, Klein DN, Hajcak G, Nelson BD. Increases in depression and anxiety symptoms in adolescents and young adults during the COVID-19 pandemic. Psychol Med 2022;52(14):3222-3230 [FREE Full text] [doi: 10.1017/S0033291720005358] [Medline: 33436120]
6. Hyman SE. The diagnosis of mental disorders: the problem of reification. Annu Rev Clin Psychol 2010;6:155-179. [doi: 10.1146/annurev.clinpsy.3.022806.091532] [Medline: 17716032]
7. Wakefield JC. Klerman's "credo" reconsidered: neo-Kraepelinianism, Spitzer's views, and what we can learn from the past. World Psychiatry 2022;21(1):4-25 [FREE Full text] [doi: 10.1002/wps.20942] [Medline: 35015356]

XSL•FO
RenderX

8. Kendell R, Jablensky A. Distinguishing between the validity and utility of psychiatric diagnoses. Am J Psychiatry 2003;160(1):4-12. [doi: 10.1176/appi.ajp.160.1.4] [Medline: 12505793]
9. Morris SE, Sanislow CA, Pacheco J, Vaidyanathan U, Gordon JA, Cuthbert BN. Revisiting the seven pillars of RDoC. BMC Med 2022;20(1):220 [FREE Full text] [doi: 10.1186/s12916-022-02414-0] [Medline: 35768815]
10. Kotov R, Krueger RF, Watson D, Achenbach TM, Althoff RR, Bagby RM, et al. The hierarchical taxonomy of psychopathology (HiTOP): a dimensional alternative to traditional nosologies. J Abnorm Psychol 2017;126(4):454-477. [doi: 10.1037/abn0000258] [Medline: 28333488]
11. Zachar P. A Metaphysics of Psychopathology. Cambridge, Massachusetts: MIT Press; 2014.
12. Naslund JA, Aschbrenner KA, Marsch LA, Bartels SJ. The future of mental health care: peer-to-peer support and social media. Epidemiol Psychiatr Sci 2016;25(2):113-122 [FREE Full text] [doi: 10.1017/S2045796015001067] [Medline: 26744309]
13. Rayland A, Andrews J. From social network to peer support network: opportunities to explore mechanisms of online peer support for mental health. JMIR Ment Health 2023;10:e41855 [FREE Full text] [doi: 10.2196/41855] [Medline: 36853738]
14. De Choudhury M, De S. Mental health discourse on reddit: self-disclosure, social support, and anonymity. ICWSM 2014;8(1):71-80. [doi: 10.1609/icwsm.v8i1.14526]
15. Kirkbride JB, Anglin DM, Colman I, Dykxhoorn J, Jones PB, Patalay P, et al. The social determinants of mental health and disorder: evidence, prevention and recommendations. World Psychiatry 2024;23(1):58-90 [FREE Full text] [doi: 10.1002/wps.21160] [Medline: 38214615]
16. Dickson SJ, Bussey K, Kangas M, Grocott S, Rapee RM. Barriers to accessing and engaging with mental health services for low-income families in Australia: a qualitative evaluation. J Child Fam Stud 2025;34(11):2862-2877. [doi: 10.1007/s10826-025-03172-2]
17. Lowther-Payne HJ, Ushakova A, Beckwith A, Liberty C, Edge R, Lobban F. Understanding inequalities in access to adult mental health services in the UK: a systematic mapping review. BMC Health Serv Res 2023;23(1):1042 [FREE Full text] [doi: 10.1186/s12913-023-10030-8] [Medline: 37773154]
18. Montag C, Duke É, Markowetz A. Toward psychoinformatics: computer science meets psychology. Comput Math Methods Med 2016;2016:2983685 [FREE Full text] [doi: 10.1155/2016/2983685] [Medline: 27403204]
19. Conway M, O'Connor D. Social media, big data, and mental health: current advances and ethical implications. Curr Opin Psychol 2016;9:77-82 [FREE Full text] [doi: 10.1016/j.copsyc.2016.01.004] [Medline: 27042689]
20. Chancellor S, De Choudhury M. Methods in predictive techniques for mental health status on social media: a critical review. NPJ Digit Med 2020;3:43 [FREE Full text] [doi: 10.1038/s41746-020-0233-7] [Medline: 32219184]
21. Feldhege J, Moessner M, Bauer S. Who says what? Content and participation characteristics in an online depression community. J Affect Disord 2020;263:521-527. [doi: 10.1016/j.jad.2019.11.007] [Medline: 31780138]
22. Low DM, Rumker L, Talkar T, Torous J, Cecchi G, Ghosh SS. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-19: observational study. J Med Internet Res 2020;22(10):e22635 [FREE Full text] [doi: 10.2196/22635] [Medline: 32936777]
23. Morini V, Sansoni M, Rossetti G, Pedreschi D, Castillo C. Participant behavior and community response in online mental health communities: insights from reddit. Computers in Human Behavior 2025;165:108544. [doi: 10.1016/j.chb.2024.108544]
24. Jin C, Zhu Z. Multimorbidity patterns and early signals of diabetes in online communities. JAMIA Open 2025;8(3):ooaf049. [doi: 10.1093/jamiaopen/ooaf049] [Medline: 40453764]
25. McElroy E, Shevlin M, Murphy J, McBride O. Co-occurring internalizing and externalizing psychopathology in childhood and adolescence: a network approach. Eur Child Adolesc Psychiatry 2018;27(11):1449-1457 [FREE Full text] [doi: 10.1007/s00787-018-1128-x] [Medline: 29520540]
26. Boschloo L, Schoevers RA, van Borkulo CD, Borsboom D, Oldehinkel AJ. The network structure of psychopathology in a community sample of preadolescents. J Abnorm Psychol 2016;125(4):599-606. [doi: 10.1037/abn0000150] [Medline: 27030994]
27. Forbes MK. Reconstructing psychopathology: a data-driven reorganization of the symptoms in DSM-5. Clin Psychol Sci 2023;13(3). [doi: 10.31234/osf.io/7um9a]
28. Plana-Ripoll O, Pedersen CB, Holtz Y, Benros ME, Dalsgaard S, de Jonge P, et al. Exploring comorbidity within mental disorders among a Danish national population. JAMA Psychiatry 2019;76(3):259-270 [FREE Full text] [doi: 10.1001/jamapsychiatry.2018.3658] [Medline: 30649197]
29. Dervić E, Sorger J, Yang L, Leutner M, Kautzky A, Thurner S, et al. Unraveling cradle-to-grave disease trajectories from multilayer comorbidity networks. NPJ Digit Med 2024;7(1):56 [FREE Full text] [doi: 10.1038/s41746-024-01015-w] [Medline: 38454004]
30. Dalgleish T, Black M, Johnston D, Bevan A. Transdiagnostic approaches to mental health problems: current status and future directions. J Consult Clin Psychol 2020;88(3):179-195. [doi: 10.1037/ccp0000482] [Medline: 32068421]
31. Borsboom D, Cramer AOJ. Network analysis: an integrative approach to the structure of psychopathology. Annu Rev Clin Psychol 2013;9:91-121. [doi: 10.1146/annurev-clinpsy-050212-185608] [Medline: 23537483]
32. Borsboom D. A network theory of mental disorders. World Psychiatry 2017;16(1):5-13 [FREE Full text] [doi: 10.1002/wps.20375] [Medline: 28127906]

XSL•FO
RenderX

33. Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J. The pushshift reddit dataset. In: Proceedings of the International AAAI Conference on Web and Social Media. 2020 Presented at: Proceedings of the International AAAI Conference on Web and Social Media; June 8-11, 2020; Atlanta, GA p. 830-839. [doi: 10.1609/icwsm.v14i1.7347]
34. List of reddit communities (sorted by users). Reddit. 2025. URL: https://www.reddit.com/best/communities/1/ [accessed 2025-07-14]
35. ICD-10 version: 2019. World Health Organization (WHO). 2019. URL: https://icd.who.int/browse10/2019/en [accessed 2025-07-14]
36. Burtch G, Lee D, Chen Z. The consequences of generative AI for online knowledge communities. Sci Rep 2024;14(1):10413 [FREE Full text] [doi: 10.1038/s41598-024-61221-0] [Medline: 38710885]
37. Møller AG, Romero DM, Jurgens D, Aiello LM. The impact of generative AI on social media: an experimental study. arXiv 2025. [doi: 10.48550/arXiv.2506.14295]
38. Home. Bot Rank. URL: https://botrank.com [accessed 2025-07-14]
39. Neal Z. The backbone of bipartite projections: inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors. Social Networks 2014;39:84-97. [doi: 10.1016/j.socnet.2014.06.001]
40. Saracco F, Straka MJ, Clemente RD, Gabrielli A, Caldarelli G, Squartini T. Inferring monopartite projections of bipartite networks: an entropy-based approach. New J. Phys 2017;19(5):053022. [doi: 10.1088/1367-2630/aa6b38]
41. Shaffer JP. Multiple hypothesis testing. Annu Rev Psychol 2024;46(1):561-584. [doi: 10.1146/annurev.psych.46.1.561]
42. Reddit Network of Psychopathology - Positive Associations (Interactive). Retina. URL: https://ouestware.gitlab.io/retina/beta/#/graph/?url=https%3A%2F%2Fgist.githubusercontent.com%2Fboevkoski%2F293ebeb33e7a5462cd6597c5079033%2Fraw%2Ff219e5b5343fe3375 1cc2df80ccf3cbd78d889%2Freddit_psychopathology_positive_associations.gexf [accessed 2025-07-14]
43. Reddit Network of Psychopathology - Negative Associations (Interactive). Retina. URL: https://ouestware.gitlab.io/retina/beta/#/graph/?url=https%3A%2F%2Fgist.githubusercontent.com%2Fboevkoski%2F4b35e1816e0706a1aa80ad03a4e1a2b%2Fraw%2F2d122d931cfe944c4d9c51713d7357a6d44e47%2Freddit_psychopathology_negative_associations.gexf [accessed 2025-07-14]
44. Network of psychopathology based on ICD-10 diagnostic criteria (Interactive). Retina. URL: https://ouestware.gitlab.io/retina/beta/#/graph/?url=https://gist.githubusercontent.com/boevkoski/4423ac95663e168e2d355b12d96a7a6f/raw/feee54a83186335facaf89c5133f6673a95ed6b8/ICD10_psychopathology_diagnostic_criteria_overlaps.gexf [accessed 2025-07-14]
45. Tio P, Epskamp S, Noordhof A, Borsboom D. Mapping the manuals of madness: comparing the ICD-10 and DSM-IV-TR using a network approach. Int J Methods Psychiatr Res 2016;25(4):267-276 [FREE Full text] [doi: 10.1002/mpr.1503] [Medline: 27028040]
46. The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines. Geneva, Switzerland: World Health Organization; 1992.
47. Yim O, Ramdeen KT. Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. TQMP 2015;11(1):8-21. [doi: 10.20982/tqmp.11.1.p008]
48. Mammone N, Ieracitano C, Adeli H, Bramanti A, Morabito FC. Permutation jaccard distance-based hierarchical clustering to estimate EEG network density modifications in MCI subjects. IEEE Trans Neural Netw Learn Syst 2018;29:5122-5135. [doi: 10.1109/TNNLS.2018.2791644] [Medline: 29994428]
49. Liu X, Zhu XH, Qiu P, Chen W. A correlation-matrix-based hierarchical clustering method for functional connectivity analysis. J Neurosci Methods 2012;211(1):94-102 [FREE Full text] [doi: 10.1016/j.jneumeth.2012.08.016] [Medline: 22939920]
50. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. WIREs Data Min & Knowl 2011;2(1):86-97. [doi: 10.1002/widm.53]
51. Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci U S A 2006;103(23):8577-8582 [FREE Full text] [doi: 10.1073/pnas.0601602103] [Medline: 16723398]
52. Wagner S, Wagner D. Comparing clusterings - an overview. University of Science and Technology of China (USTC). URL: http://staff.ustc.edu.cn/~zwp/teach/MVA/cluster_validation.pdf [accessed 2007-01-12]
53. King SA. Researching internet communities: proposed ethical guidelines for the reporting of results. The Information Society 1996;12(2):119-128. [doi: 10.1080/713856145]
54. Eysenbach G, Till JE. Ethical issues in qualitative research on internet communities. BMJ 2001;323(7321):1103-1105 [FREE Full text] [doi: 10.1136/bmj.323.7321.1103] [Medline: 11701577]
55. Moreno MA, Goniu N, Moreno PS, Diekema D. Ethics of social media research: common concerns and practical considerations. Cyberpsychol Behav Soc Netw 2013;16(9):708-713 [FREE Full text] [doi: 10.1089/cyber.2012.0334] [Medline: 23679571]
56. Chancellor S, Birnbaum M, Caine E, Silenzio V, De CM. A taxonomy of ethical tensions in inferring mental health states from social media. 2019 Presented at: FAT* '19: Conference on Fairness, Accountability, and Transparency; January 29-31, 2019; Atlanta, GA p. 79-88. [doi: 10.1145/3287560.3287587]

XSL•FO
RenderX

57. Tri-Council Policy Statement: ethical conduct for research involving humans. Government of Canada. 2022. URL: https://ethics.gc.ca/eng/policy-politique_tcps2-eptc2_2022.html [accessed 2025-12-23]
58. Kosten TR, George TP. The neurobiology of opioid dependence: implications for treatment. Sci Pract Perspect 2002;1(1):13-20 [FREE Full text] [doi: 10.1151/spp021113] [Medline: 18567959]
59. Martins SS, Fenton MC, Keyes KM, Blanco C, Zhu H, Storr CL. Mood and anxiety disorders and their association with non-medical prescription opioid use and prescription opioid-use disorder: longitudinal evidence from the National Epidemiologic Study on Alcohol and Related Conditions. Psychol Med 2012;42(6):1261-1272 [FREE Full text] [doi: 10.1017/S0033291711002145] [Medline: 21999943]
60. McLaughlin KA, Colich NL, Rodman AM, Weissman DG. Mechanisms linking childhood trauma exposure and psychopathology: a transdiagnostic model of risk and resilience. BMC Med 2020;18(1):96 [FREE Full text] [doi: 10.1186/s12916-020-01561-6] [Medline: 32238167]
61. Ellickson-Larew S, Stasik-O'Brien SM, Stanton K, Watson D. Dissociation as a multidimensional transdiagnostic symptom. PConsci-TRP 2020;7(2):126-150. [doi: 10.1037/cns0000218]
62. Wu JB, Yang Y, Zhou Q, Li J, Yang WK, Yin X, et al. The relationship between screen time, screen content for children aged 1-3, and the risk of ADHD in preschools. PLoS One 2025;20(4):e0312654 [FREE Full text] [doi: 10.1371/journal.pone.0312654] [Medline: 40267918]
63. Fekih-Romdhane F, Jahrami H, Away R, Trabelsi K, Pandi-Perumal SR, Seeman MV, et al. The relationship between technology addictions and schizotypal traits: mediating roles of depression, anxiety, and stress. BMC Psychiatry 2023;23(1):67 [FREE Full text] [doi: 10.1186/s12888-023-04563-9] [Medline: 36698079]
64. Dong HY, Wang B, Li HH, Yue XJ, Jia FY. Correlation between screen time and autistic symptoms as well as development quotients in children with autism spectrum disorder. Front Psychiatry 2021;12:619994 [FREE Full text] [doi: 10.3389/fpsyt.2021.619994] [Medline: 33664683]
65. Borsboom D, Cramer AOJ, Schmittmann VD, Epskamp S, Waldorp LJ. The small world of psychopathology. PLoS One 2011;6(11):e27407 [FREE Full text] [doi: 10.1371/journal.pone.0027407] [Medline: 22114671]
66. Grant BF, Goldstein RB, Saha TD, Chou SP, Jung J, Zhang H, et al. Epidemiology of DSM-5 alcohol use disorder: results from the National Epidemiologic Survey on Alcohol and Related Conditions III. JAMA Psychiatry 2015;72(8):757-766 [FREE Full text] [doi: 10.1001/jamapsychiatry.2015.0584] [Medline: 26039070]
67. Kessler RC, Nelson CB, McGonagle KA, Edlund MJ, Frank RG, Leaf PJ. The epidemiology of co-occurring addictive and mental disorders: implications for prevention and service utilization. Am J Orthopsychiatry 1996;66(1):17-31 [FREE Full text] [doi: 10.1037/h0080151] [Medline: 8720638]
68. Corrigan PW, Kuwabara S, O'Shaughnessy J. The public stigma of mental illness and drug addiction. J Soc Work 2009;9(2):139-147. [doi: 10.1177/1468017308101818]
69. Room R. Stigma, social inequality and alcohol and drug use. Drug Alcohol Rev 2005;24(2):143-155. [doi: 10.1080/09595230500102434] [Medline: 16076584]
70. Hing N, Russell AMT, Gainsbury SM, Nuske E. The public stigma of problem gambling: its nature and relative intensity compared to other health conditions. J Gambl Stud 2016;32(3):847-864 [FREE Full text] [doi: 10.1007/s10899-015-9580-8] [Medline: 26487344]
71. Hogg B, Gardoki-Souto I, Valiente-Gómez A, Rosa AR, Fortea L, Radua J, et al. Psychological trauma as a transdiagnostic risk factor for mental disorder: an umbrella meta-analysis. Eur Arch Psychiatry Clin Neurosci 2023;273(2):397-410. [doi: 10.1007/s00406-022-01495-5] [Medline: 36208317]
72. Evkoski B, Letina S, Kralj Novak P, Riddell J. Premenstrual dysphoric disorder in online peer support communities: a Reddit case study. Sci Rep 2025;15(1):34300 [FREE Full text] [doi: 10.1038/s41598-025-19220-2] [Medline: 41034480]
73. Forbes MK, Neo B, Nezami OM, Fried EI, Faure K, Michelsen B, et al. Elemental psychopathology: distilling constituent symptoms and patterns of repetition in the diagnostic criteria of the DSM-5. Psychol Med 2024;54(5):886-894. [doi: 10.1017/S0033291723002544] [Medline: 37665038]
74. Underhill R, Foulkes L. Self-diagnosis of mental disorders: a qualitative study of attitudes on Reddit. Qual Health Res 2025;35(7):779-792 [FREE Full text] [doi: 10.1177/10497323241288785] [Medline: 39422576]
75. Chan GJ, Fung M, Warrington J, Nowak SA. Understanding health-related discussions on reddit: development of a topic assignment method and exploratory analysis. JMIR Form Res 2025;9:e55309 [FREE Full text] [doi: 10.2196/55309] [Medline: 39879094]
76. Kendler KS. The nature of psychiatric disorders. World Psychiatry 2016;15(1):5-12 [FREE Full text] [doi: 10.1002/wps.20292] [Medline: 26833596]
77. Zachar P. Psychological Concepts and Biological Psychiatry: A Philosophical Analysis. Amsterdam, Netherlands: John Benjamins Publishing Company; 2000.
78. Zachar P, Kendler KS. Psychiatric disorders: a conceptual taxonomy. Am J Psychiatry 2007;164(4):557-565. [doi: 10.1176/ajp.2007.164.4.557] [Medline: 17403967]
79. Lahey BB, Tiemeier H, Krueger RF. Seven reasons why binary diagnostic categories should be replaced with empirically sounder and less stigmatizing dimensions. JCPP Adv 2022;2(4):e12108 [FREE Full text] [doi: 10.1002/jcv2.12108] [Medline: 37431412]

XSL•FO
RenderX

80. McGorry PD, Hickie IB, Kotov R, Schmaal L, Wood SJ, Allan SM, et al. New diagnosis in psychiatry: beyond heuristics. Psychol Med 2025;55:e26. [doi: 10.1017/S003329172400223X] [Medline: 39911018]
81. Ringwald WR, Forbes MK, Wright AGC. Meta-analysis of structural evidence for the Hierarchical Taxonomy of Psychopathology (HiTOP) model. Psychol Med 2023;53(2):533-546. [doi: 10.1017/S0033291721001902] [Medline: 33988108]
82. Ruggero CJ, Kotov R, Hopwood CJ, First M, Clark LA, Skodol AE, et al. Integrating the Hierarchical Taxonomy of Psychopathology (HiTOP) into clinical practice. J Consult Clin Psychol 2019;87(12):1069-1084 [FREE Full text] [doi: 10.1037/ccp0000452] [Medline: 31724426]
83. Finlay SC. Age and gender in Reddit commenting and success. J Inf Sci Theory Pract 2014;2(3):18-28. [doi: 10.1633/jistap.2014.2.3.2]
84. Reddit users by country. World Population Review. 2025. URL: https://worldpopulationreview.com/country-rankings/reddit-users-by-country [accessed 2025-07-14]
85. Proferes N, Jones N, Gilbert S, Fiesler C, Zimmer M. Studying Reddit: a systematic overview of disciplines, approaches, methods, and ethics. Social Media + Society 2021;7(2):205630512110190. [doi: 10.1177/20563051211019004]
86. Tse JSY, Haslam N. Broad concepts of mental disorder predict self-diagnosis. SSM - Mental Health 2024;6:100326. [doi: 10.1016/j.ssmmh.2024.100326]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
ARI: Adjusted Rand Index
DSM: Diagnostic and Statistical Manual of Mental Disorders
DSM-5: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition)
ICD: International Classification of Diseases
ICD-10: International Statistical Classification of Diseases, 10th Revision
ICD-11: International Classification of Diseases, 11th Revision
NMI: normalized mutual information
PTSD: posttraumatic stress disorder
TCPS 2: Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans

Edited by A Mavragani; submitted 19.Jul.2025; peer-reviewed by J Kang, V Morini, Z Zhu, X Liang, A Markovits; comments to author 02.Sep.2025; accepted 02.Dec.2025; published 30.Jan.2026.

Please cite as:
Evkoski B, Letina S, Kralj Novak P
The Structure of Psychopathology on Reddit: Network Analysis of Mental Health Communities in Relation to the ICD Diagnostic System
J Med Internet Res 2026;28:e80958
URL: https://www.jmir.org/2026/1/e80958
doi:10.2196/80958
PMID:

©Bojan Evkoski, Srebrenka Letina, Petra Kralj Novak. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Experiences of Alert Fatigue and Its Contributing Factors in Hospitals: Qualitative Study

Nicki Newton[1], BSc (Hons), Grad Cert, PhD; Adeola Bamgboje-Ayodele[2], PhD; Rowena Forsyth[3], PhD; Amina Tariq[4], PhD; Jing Huang[5], MIM; Ramyasri Yannam[5], MBBS; Daniel J Lalor[5], MBA; Andrew James Sobey[5], MHSM; Melissa T Baysari[1], PhD

[1]Digital Health Human Factors Group, Susan Wakil School of Nursing and Midwifery, Faculty of Medicine and Health, The University of Sydney, Susan Wakil Health Building, Camperdown, Australia

[2]Discipline of Design, School of Architecture, Design and Planning, The University of Sydney, Sydney, Australia

[3]Cyberpsychology Research Group, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia

[4]Australian Centre for Health Services Innovation and Centre for Healthcare Transformation, School of Public Health and Social Work, Faculty of Health, Queensland University of Technology, Brisbane, Australia

[5]Canberra Health Services, Act Health, Canberra, Australia

Corresponding Author:
Nicki Newton, BSc (Hons), Grad Cert, PhD
Digital Health Human Factors Group, Susan Wakil School of Nursing and Midwifery, Faculty of Medicine and Health, The University of Sydney, Susan Wakil Health Building, Camperdown, Australia

Abstract

Background: Alerts, a key feature of electronic health record systems, intend to improve patient safety by providing timely information at the point of care. However, many electronic health record systems generate excessive alerts that are not immediately clinically relevant and that contribute to alert fatigue. Despite growing recognition of alert fatigue as a safety concern, clinicians' experiences of alert fatigue and the broader system-level factors that contribute to it being experienced are not well understood.

Objective: This study aims to use a human factors approach to (1) comprehensively explore how alert fatigue is experienced by junior doctors; (2) identify factors that contribute to experiences of alert fatigue; (3) identify perceived impacts of alert fatigue on employees, organizations, and patients; and (4) identify strategies to reduce alert fatigue in practice.

Methods: Semistructured interviews were conducted with junior doctors working in hospitals across Australia. Data were thematically analyzed using a hybrid inductive and deductive approach, informed by the Systems Engineering Initiative for Patient Safety and an information processing model.

Results: A total of 20 junior doctors were interviewed. Alert fatigue was described to occur at different stages of information processing, including when alerts were not detected, were superficially processed using mental shortcuts, or required excessive cognitive effort to interpret. When alerts were not detected or thoroughly processed, participants more often perceived impacts on patient safety and care quality due to the potential to miss important information. Further, when alerts required excessive cognitive effort, participants frequently reported interruptions, frustration, and time and effort loss as impacts. Factors influencing experiences of alert fatigue were identified in all Systems Engineering Initiative for Patient Safety work system domains, including those related to people, tasks, the environment, tools and technologies, and the organization. Key contributors included the design and clinical relevance of alerts, institutional norms and expectations, and information overload from system alerts as well as other alerts and tasks. Alert fatigue was also described to be experienced differently depending on provider characteristics, such as experiences with and knowledge of alerts, mood, and personality, and organizational factors, including culture, shift type, and time of day.

Conclusions: Alert fatigue is not a binary concept but is instead experienced on a continuum and influenced by interacting individual, technical, and contextual factors. Future research should incorporate clinician self-reports to evaluate experiences of alert fatigue in addition to objective measures. Addressing alert fatigue requires tailored interventions that target its different causes and outcomes. These could include technical and design improvements, changes to organizational practices, and individual customization to reduce experiences of fatigue and accommodate differences in clinicians' needs.

(J Med Internet Res 2026;28:e78676) doi:10.2196/78676

KEYWORDS

alert fatigue; clinical decision support systems; alerts; hospitals; junior doctors; human factors; digital health

XSL•FO
RenderX

Introduction

Clinical decision support (CDS) systems have proliferated in health care, forming a core component of many electronic health record (EHR) systems now used in routine practice [1-3]. Within these systems, alerts represent the primary mode of CDS delivery. Common alerts include those that notify clinicians of potential drug-drug interactions, allergies, and dosing errors and those promoting adherence to best practice care [4,5]. Alerts can interrupt clinicians' workflows during their interactions with the EHR (interruptive alerts) or passively display CDS through icons or flags that require users to actively seek information out (passive alerts) [6]. While alerts are designed to support clinical decision-making and enhance patient safety by providing relevant information at the point of care, many systems generate excessive alerts that are not immediately clinically relevant [4,7-10].

Alert fatigue is a widely reported concern arising from excessive alerting, but as a concept, it remains poorly understood and inconsistently measured [2,11]. Existing research has largely relied on observable behaviors, such as alert overrides and response times, to infer the presence of alert fatigue, assuming that clinicians experiencing alert fatigue will dismiss alerts without considering their content [12,13]. While some studies suggest that alert fatigue arises when clinicians become cognitively overloaded by alerts [4], others report it to be driven by desensitization that develops following repeated exposure to the same alert over time [14,15].

These inconsistencies highlight the limitations of using behavioral proxies to measure what is fundamentally a *subjective* mental state [16,17]. Observable behaviors cannot fully capture clinicians' cognitive experiences and offer limited insight into the broader factors that contribute to fatigue [18]. While alert fatigue is frequently identified as a barrier to clinicians' use of alerts in qualitative studies, these studies are typically concerned with overall alert experiences rather than examining alert fatigue itself and the factors that shape it [19-21]. As a result, definitions of alert fatigue and its perceived causes and impacts remain superficial and lack a systemic perspective.

Human factors methods have been widely applied in health care to understand and address system-level factors that impact care delivery, patient safety, and performance outcomes [22]. These approaches are well suited to investigating complex, dynamic issues like alert fatigue, as they account for the cognitive and behavioral processes that occur during clinicians' interactions with technologies, within the broader context of the tasks, organizational settings, and environments in which these interactions occur [23,24]. Accordingly, this study aimed to (1) comprehensively explore how alert fatigue is experienced by junior doctors; (2) identify factors that contribute to experiences of alert fatigue; (3) identify perceived impacts of alert fatigue on employees, organizations, and patients; and (4) identify strategies to reduce alert fatigue in practice.

Methods

Study Design

This study is reported following the SRQR (Standards for Reporting Qualitative Research) guidelines (Checklist 1) [25]. A qualitative design, using semistructured individual interviews, was selected to explore the subjective and nuanced nature of alert fatigue in depth [26,27].

Theoretical Frameworks

The Systems Engineering Initiative for Patient Safety (SEIPS) is a human factors model used in health care settings to identify and address safety concerns [24]. The SEIPS describes the "work system," which comprises 5 domains, including the *person* (both patients and health care providers) at the center of care, who performs *tasks*, using various *tools and technologies* within a *work environment* under specific *organizational conditions*. Interactions between components of the work system influence the delivery of care and other *process*es, which in turn impact *patient*, *employee*, and *organizational outcomes*. The SEIPS model was used to interpret and present data.

We used an information processing model as outlined by Wickens and Carswell [28] to further interpret the *process* by which alert fatigue was experienced by participants. The model describes information processing to occur across 4 stages: *selecting information* (detecting stimuli from the environment), *perception and data interpretation* (recognizing and interpreting the stimuli), *comprehension and cognition* (processing and integrating information with prior knowledge to assess relevance and meaning), and *action selection* (choosing and executing a response based on the processed information). Thus, this model was well suited to understanding the mental processes of alert fatigue and associated actions arising from its presence. Although alert fatigue has previously been described in the context of information processing [29], an analysis of how alert fatigue manifests across different stages has not been explored. A schematic of the integrated models is presented in Figure 1.

XSL•FO
RenderX

Figure 1. Schematic of the integrated Systems Engineering Initiative for Patient Safety (SEIPS) and information processing models used to interpret data.

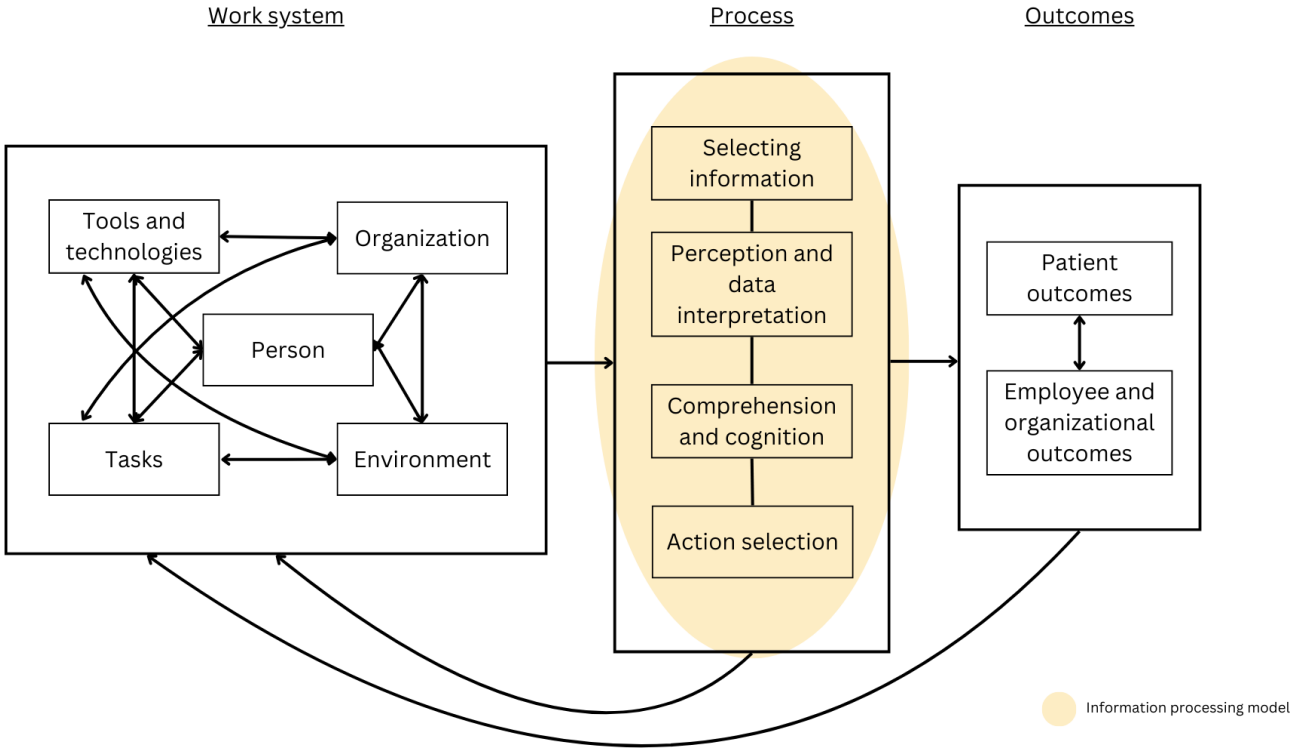

Study Team

The study team included researchers with expertise across human factors, digital health, health services, sociology, and psychology; and hospital-based pharmacy directors. Additionally, 2 members of the study team were junior doctors, who provided lived experience insights.

Participants and Recruitment

Eligible participants were practicing junior doctors, also referred to as junior medical officers (JMOs) in Australian hospital systems, defined as interns (postgraduate year [PGY] 1), residents (PGY2), and registrars (PGY3+), who had interacted with alerts in clinical information systems (eg, EHR systems) in hospital settings in Australia. Junior doctors were selected for the study as they are frequent recipients of alerts in hospital settings. A convenience sample of participants was recruited by disseminating details about the study through email distribution lists, posters, and presentations at JMO training sessions, at 2 hospitals within the same hospital network. Additional participants were recruited via snowball sampling through the authors' professional networks, including word of mouth, LinkedIn posts, and emails to hospital contacts who circulated the study to JMO networks. Prospective participants were encouraged to contact the researchers via email if they were interested in participating.

Data Collection

A semistructured interview guide was developed using insights from existing research on alert and alarm fatigue and discussion among all authors [4,14,30]. Two pilot interviews were conducted by 1 author (NN) with JMO coresearchers via Microsoft Teams, which resulted in minor refinements to the interview guide, such as adding new questions and changing the order of questions to enhance flow. Questions centered around participants' use of alerts, perceptions and experiences of alert fatigue, factors influencing alert fatigue, the impact of alert fatigue, and recommendations to reduce experiences of alert fatigue in practice (Multimedia Appendix 1). Individual interviews were conducted with participants by one author (NN) using Microsoft Teams, Zoom (Zoom Communications, Inc), or a phone. Three participants were known to the interviewer prior to the interviews through their professional and personal networks.

Interviews were audio recorded, transcribed using Otter AI software (Otter.ai, Inc) [31], and checked for accuracy. Transcription and analysis occurred in parallel to data collection, allowing for further refinement of the interview guide to explore themes in depth as interviews progressed. Interviews continued until data saturation was reached; that is, no new themes emerged in the final interviews, and only new examples or perspectives related to existing themes were identified [32].

Data Analysis

Transcripts were imported into NVivo (version 14.24.1; Lumivero) [33] and thematically analyzed using a combined inductive and deductive approach [34]. The first 6 transcripts were independently analyzed by 2 authors (NN and MB or AB or RF) who inductively coded quotes. Themes were deductively mapped to 4 categories: defining alert fatigue, factors influencing experiences of alert fatigue, the impact of alert fatigue, and strategies to reduce alert fatigue. Each pair of authors met to discuss alignment and resolve discrepancies.

Following discussions, 3 of the 4 original categories were restructured under the SEIPS model, where codes related to defining alert fatigue were moved to *processes*, factors

XSL•FO
RenderX

influencing experiences of alert fatigue were moved to the *work system*, and the impact of alert fatigue was moved to *outcomes* [24]. Themes related to the process of alert fatigue were further interpreted using the information processing model [28].

Remaining transcripts were analyzed by 1 author (NN) using the SEIPS model, adding new codes and themes as they emerged. Upon completion of the analysis of all interviews, 4 authors (NN, MB, AB, and RF) participated in a workshop to discuss the categorization of codes and themes, which resulted in some codes being renamed, merged, and/or recategorized. The coding framework was then presented to all authors and select participants for member checking, who provided input and agreed upon the final thematic structure.

Ethical Considerations

Ethical approval was received from the Australian Capital Territory Health Low Risk Ethics Committee (protocol: 2024.LRE.00024). All participants provided written informed consent and did not receive compensation for participating in the study. Interview data were deidentified at the point of transcription, with all potentially identifying details removed. Demographic characteristics are presented in aggregate, and no identifiable participant information is included in this study.

Results

Summary

A total of 20 interviews were conducted between August 2024 and March 2025. Participants included 9 interns (PGY1), 6 residents (PGY2), and 5 registrars (PGY3+), working at 10 hospital sites, across 3 Australian states (Table 1). Interviews lasted between 18 and 56 minutes, with an average length of 33 (SD 11; range 18-56) minutes. Two interviews were conducted via phone and the remainder via videoconference.

XSL•FO
RenderX

Table . Participant characteristics

Characteristics	Participants, n (%)
Role (PGY[a])	
Intern (PGY1)	9 (45)
Resident (PGY2)	6 (30)
Registrar (PGY3+)	5 (25)
Sex	
Male	12 (60)
Female	8 (40)
Length of time at current hospital	
Less than 2 months	3 (15)
2 months to 1 year	6 (30)
1 to 2 years	4 (20)
2+ years	7 (35)
Clinical experience total (including prior clinical work)	
Less than 2 months	2 (10)
2 months to 1 year	3 (15)
1 to 2 years	4 (20)
2-5 years	5 (25)
5+ years	6 (30)
State	
NSW[b]	11 (55)
ACT[c]	7 (35)
SA[d]	2 (10)
Systems used for alerts	
Commercial	18 (90)
Commercial and homegrown	2 (10)

[a]PGY: postgraduate year.

[b]NSW: New South Wales.

[c]ACT: Australian Capital Territory.

[d]SA: South Australia.

Participants received alerts from EHR, electronic medical record (EMR), and/or electronic medication management systems. All participants were users of commercially available EHR or EMR systems (eg, Epic, Cerner, Sunrise, and MedChart), while 2 participants also used homegrown systems. Alert types included passive and interruptive alerts such as those for medications (eg, allergy, drug-drug interactions, and dosing), and best practice (eg, venous thromboembolism prophylaxis, advanced care directives, and resuscitation plans). Alerts popped up during tasks such as medication ordering, opening patient charts, and opening the EHR system. Alerts outside of standard clinical information systems ("system") were also described.

In the following sections, we present the clinicians' perceptions of alert fatigue (process), contributing factors (work system), impacts (outcomes), and potential strategies for addressing alert fatigue. Figure 2 shows a schematic of themes aligned to the SEIPS model.

XSL•FO
RenderX

Figure 2. Schematic of themes related to experiences of alert fatigue within the Systems Engineering Initiative for Patient Safety (SEIPS) model.

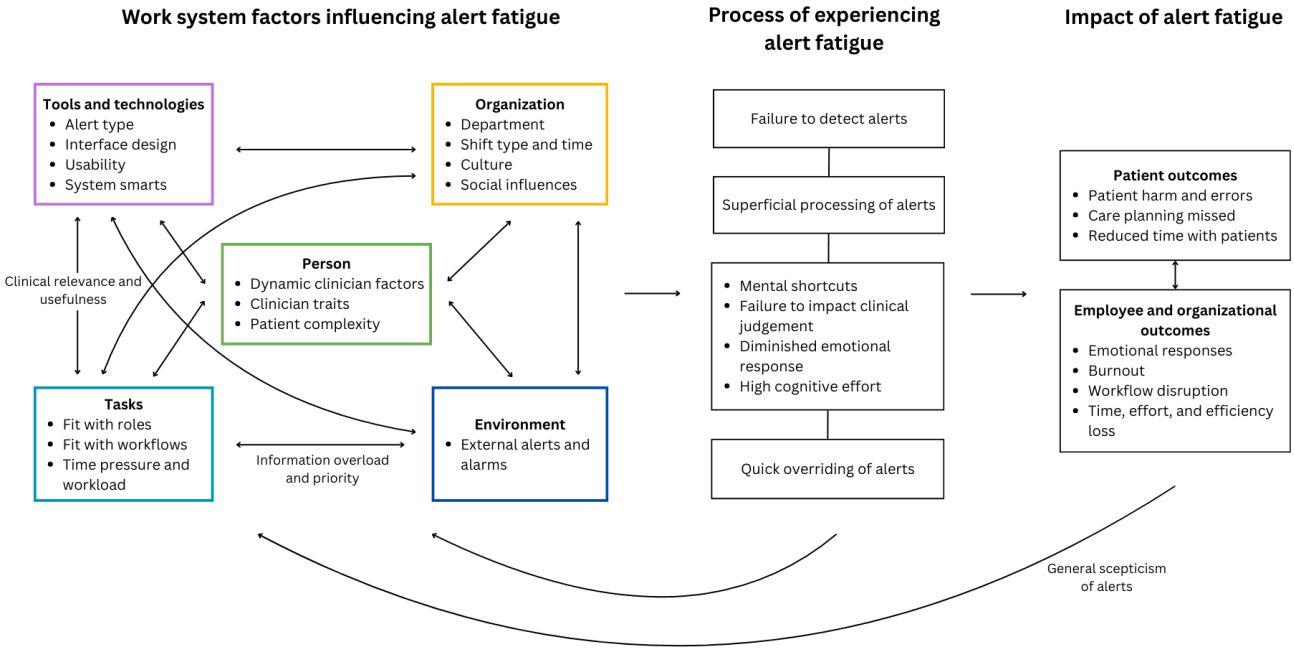

The Process of Experiencing Alert Fatigue

Selecting Information

Alert fatigue manifested during the information selection stage as a failure to notice alerts or detect alert stimuli. This occurred mainly for passive alerts, which were often described to be inconspicuous, but also for interruptive alerts: “The idea of alerts being there is just part of the job, it’s just part of the furniture…I probably at least once will miss the fact that [medications are] expiring, my eyes will gloss over it” (P1).

Perception and Data Interpretation

Perception and interpretation involved recognizing and processing alerts. Clinicians described superficially processing alerts by only partially viewing or reading them before responding. Participants explained that alerts were often pre-empted based on their interactions with the system before an alert was triggered or identified by design features such as titles, keywords, fonts, colors, and layouts. Many participants felt that they could quickly and effectively recognize these alerts and often did not feel “fatigued” by alerts that were processed in this manner: “it’s so mindless….All you process is, it’s an ECG one, cancel, it’s an ACD one, cancel” (P14).

Comprehension and Cognition

Comprehension and cognition involved assessing the relevance of alerts to inform decision-making. While participants initially described reading and actively interpreting the value of alerts, alert fatigue occurred where active interpretation of alerts and influence on clinical judgment ceased and mental shortcuts instead guided participants:

> *I would have read the alert the first time, but then…you go, “okay, I know what this alert is about,” and the next day you would have forgotten which drugs are even mentioned. I just know that this is not relevant to me.* [P13]

Emotional responses to alerts, such as anxiety and stress, were reported to be common early on. However, this diminished over time as users reported becoming desensitized and “complacent.” Conversely, some clinicians reported experiencing alert fatigue due to the cognitive effort and time required to interpret and appropriately respond to alerts.

Action Selection

Action selection comprised users’ responses to alerts. Alerts that were quickly and habitually overridden signified alert fatigue at this stage. For example, participants described being “so used to clicking boxes” (P6) that “you just click through them automatically” (P13) and that “it’s like whack-a-mole. You’re just getting rid of everything as fast as you can” (P14).

Some participants said that they broke habitual responding patterns, for example, when their role changed, or they made an error due to experiencing alert fatigue. However, these patterns were likely to return: “after a while, you desensitize yourself again after that going well, it hasn’t happened for a long time, and your brain kind of becomes complacent again, with going Yes, yes, yes, through all the tick boxes” (P18). Some alerts that were quickly overridden were however sometimes still perceived to be useful as they prompted consideration and did not take excessive cognitive effort to process.

Work System Factors Influencing Experiences of Alert Fatigue

Factors were coded to all work system domains of the SEIPS. Multimedia Appendix 2 shows the distribution of coded references within each domain.

Person

Patients’ clinical complexity was reported to influence alert fatigue. Some participants described paying closer attention to alerts for unstable patients, while others described increased

XSL•FO
RenderX

fatigue due to the volume of alerts that complex patients triggered.

Clinician traits, including clinician type, clinical experience, and personality, also influenced alert fatigue. While many participants believed that senior clinicians experienced greater fatigue, more senior participants reported experiencing less fatigue than when junior, as they had more confidence and consequently used less cognitive effort to appropriately respond: "when I was more junior and didn't quite understand as well, I would often just go, 'advice ignored,' without really understanding the reason why" (P13). Personality traits were also raised, where participants who were more meticulous described paying more attention to alerts: "for me personally, I don't like just dismissing alerts without actually thinking through it" (P19).

Alert fatigue was influenced by dynamic factors that evolved based on individuals' experiences with alerts over time. Those who possessed a positive attitude toward alerts generally reported experiencing less fatigue:

> *I just view these alerts as another opportunity to make sure you have the whole picture.* [P10]

> *if you think it's a pain in the ass, then it's something that's going to irritate you every time. But if your attitude is, which mine is, is- "it's just part of the job," so you just get on with it.* [P15]

Alert fatigue was influenced by participants' knowledge and perceived importance of the risk the alert was targeting, including personal experiences of missing information or making errors: "if I don't understand or agree with the importance, I'm much more likely to dismiss it with so many other competing priorities" (P17). Additionally, some clinicians explained they did not know how to correctly respond to alerts:

> *no one told me how to…no one knows. So you just choose something that makes sense.* [P2]

Others described remaining cautious and attentive to alerts due to their awareness of alert fatigue.

Alert fatigue was also reported to be influenced by clinicians' emotional state, including their mood and whether they were experiencing other forms of fatigue: "you feel more emotionally drained, less sympathetic, less compassionate, and you might dismiss things more" (P1).

Tasks

Alert fatigue was influenced by workflow alignment, clinical relevance, fit with roles and responsibilities, time pressure, and information overload. Fatigue was often reported to occur where alerts were misaligned with organizational and decision-making processes. Poorly timed alerts, such as those triggered upon opening a patient's chart, were typically reported to be quickly dismissed: "I intend to view their chart with a very specific goal and updating the resuscitation right then is not a priority" (P16).

Clinical relevance and usefulness relative to clinicians' roles and responsibilities were often mentioned. While some alert types were seen as entirely irrelevant, others were inconsistently useful. Fatigue also arose where relevant alerts were lacking, forcing users to manually search for information: "it has taken me six clicks to get one answer, which is highly inefficient" (P16). Useful alerts were described to help with cognitive unloading, save time, and provide unfamiliar or easily forgotten information:

> *doctors would really appreciate if there was one alert saying all the blood tests that have been ordered are now back…because we spend a significant portion of our cognitive time in the day just thinking about that.* [P16]

Interestingly, some participants described experiencing less fatigue from alerts with poor utility: "we do not get effective alerts from EMR at the moment, so we don't rely on those alerts, so there's no fatigue involved there" (P16).

Alerts perceived to be well aligned with roles and responsibilities often did not induce alert fatigue even when they were often disregarded. This was particularly evident where clinicians had significant individual responsibility: "I'm always very grateful to get [discharge medications] alerts because the consequences of getting something wrong in that way can be disastrous, and it is on you legally" (P1).

In contrast, alerts received by a team of clinicians created a diffusion of responsibility, which led to these alerts being deprioritized: "if it was sent to everyone, it's everyone's job" (P17). Participants also described automatically clicking through alerts where they had lower autonomy: "someone's told you to chart that med, so you just do it" (P13).

Situations in which clinicians faced greater time pressure and higher workloads, such as during ward rounds, were described to trigger alert fatigue even in participants who were typically unaffected:

> *I don't feel like I'm at a stage where any alert that pops up- it's tapped into that alert fatigue. For me, it sort of comes in waves when I'm quite busy within the hospital and there's a lot of tasks that have built up.* [P10]

Information overload also contributed, with participants feeling overwhelmed by excess information or many alerts received at once:

> *it'll often give you up to 10 different alerts [in a single encounter]…it's hard to tell which ones are actually worth stopping for and thinking about, and which ones you just ignore.* [P4]

Other information presented on screen, such as task lists and message functions within the system, also contributed to overload.

Environment

Only one factor was coded to the work environment, which included information overload from *multisensory* nonsystem alerts or individual clinicians' digital ecosystem, including pagers, emails, WhatsApp messages, auditory alarms, phone calls, and in-person prompts. System alerts were perceived as follows:

XSL•FO
RenderX

> *yet another thing that the junior doctor is responsible for or is bombarded with…there's just so many different ways to message a JMO about a task, or to put a task somewhere for them to do, to page us, to come and find us on the ward…there seems to be a lot of over stimulation.* [P4]

The level of priority given to system alerts over other tasks was often reported to be low: "with the assumption that if it's automated, then it's not as pressing" (P17).

Tools and Technologies

Alert type, interface design, system usability, and volume influenced alert fatigue. Participants typically described alert fatigue to develop for specific types of alerts. However, similar-looking alerts could cause confusion or require extra cognitive effort to interpret: "allergies and drug-drug interactions often pop up looking the same, and it takes a little bit longer to work out which one it is" (P10). Participants described interface design features, including the use of colors, font, position, and layout, that could draw their attention; however, a participant stated:

> *once you get used to a certain font or format or typeface, the same problem will just happen. Regardless of whether it's in red or rainbow or purple, once you've had it pop up 20 times in one day…you just stop paying attention to it.* [P13]

Alerts with excess words and jargon contributed to alert fatigue as it could "blur what you're actually trying to pick up on in the writing as important" (P18), versus shorter alerts which were quickly processed: "the alert I pay most attention to is one that's about 10 words, regardless of the style, if it's short and something that your brain can process in three seconds, then it is probably going to be read" (P13).

Requiring override reasons was perceived positively for important alerts, forcing clinicians to "skim them because…you can't just say ignore. You've got to say, why you ignored, so they force you to put a little bit of accountability against it" (P14). Alert importance was sometimes interpreted based on the presence or absence of additional steps to dismiss the alert: "if it was so important, you couldn't click x, right?...I think everyone, whether subconsciously or consciously, knows that if it was the be all and end all, there wouldn't be a not now button" (P20).

Poor usability of alerts and clinical information systems, such as systems being "slow" and "clunky," was reported to contribute to alert fatigue by disrupting workflows. Clinicians also described interruptions where alerts blocked access to other areas of the system: "I need to cancel that order, stop what I'm doing, go into the allergy section and then interpret that information" (P10).

Alerts lacking system "smarts," for example, those flagging allergy and drug-drug interactions for patients who were already prescribed the drug in question, were described to contribute to fatigue. Similarly, constant repetition was perceived to create a "boy who cried wolf" (P17) effect. While an excessive volume of alerts was mentioned, it was only described as problematic where alerts were repetitive, irrelevant, or where multiple alerts were received at once.

Organization

Organizational factors included department, shift length and type, institutional culture, social influences, and team dynamics. While experiences of alert fatigue were generally consistent between departments and rotations, differences sometimes arose due to variations in the workload, time pressure, types of tasks and alerts received, and autonomy over tasks. Fatigue was described to increase on after-hours shifts, due to time pressure and the need to cover multiple wards, and at certain times of the day, including late at night. Participants also cited increased alert fatigue at the end of a shift due to workload demands and at the very beginning of a shift where they had insufficient information to appropriately respond.

Institutional and departmental culture and peer behaviors shaped role expectations, which in turn impacted alert fatigue. Alerts aligned with institutional norms were not necessarily less frustrating but were reported to be more widely accepted. Expectations of roles were described to be learned informally from colleagues:

> *You rock up and people just start asking you to do things…if you haven't been asked to do it, especially by a senior doctor on your team, then it's probably not your job.* [P17]

This was reinforced by observing colleagues dismissing alerts:

> *they're like, "Oh, this is so annoying." Click, click, click. It sets the scene of "Oh, this is just such an inconvenience." Skip, skip, skip.* [P8]

New JMOs reported asking colleagues and senior doctors for advice on how to respond to alerts. Alert fatigue was heightened due to the "unspoken pressure about the efficiency that you're required to have as a junior" (P4), where alerts were perceived to decrease efficiency.

Impact of Alert Fatigue

Employee and Organizational Outcomes

Alert fatigue was associated with feelings of frustration and annoyance, and along with other work stressors, could contribute to burnout and general fatigue. Some participants additionally described guilt; for example, "you just feel as though maybe you're not doing as good a job…you hate to think that you're not giving 100% to your patients" (P15). While alert fatigue was often tied to specific alerts, some participants developed broader skepticism, feeling "suspicious of alerts in general" (P6), creating a "sense of mistrust in the system" (P9) after encountering many alerts that were not relevant.

Some participants felt that fatigued users might miss opportunities to streamline their workflows. However, excess alerts were more often described to disrupt workflows, resulting in a loss of time, cognitive effort, and efficiency: "when I'm already holding six pieces of information in my head, then I might lose two, which is the most frustrating aspect" (P9). These interruptions were perceived to impact patient flow (eg, "if things don't move at a certain speed, the system breaks down,

XSL•FO
RenderX

or patients get stuck somewhere, or miss out on discharging home at a particular time" [P4]), lead to mistakes or unnecessary responses, and shift attention away from core issues (eg, "it leads to uncertainty, and I think often clouds the issue that the patient's in hospital for" [P9]).

Patient Outcomes

Participants felt alert fatigue could increase the risk of errors and result in actual patient harm, particularly when important alerts were missed or misinterpreted. For example, some participants observed doses being missed, patients being prescribed medications they were allergic to, and pre-emptive care planning being missed. Alert fatigue was also perceived to reduce time spent with patients and inadvertently impact care by reducing clinicians' patience.

Strategies to Reduce Alert Fatigue

Strategies to reduce alert fatigue included technical changes such as improving the interface design, personalizing and increasing the relevance of alerts; organizational strategies such as monitoring and evaluation initiatives and educating clinicians about alert fatigue; as well as personal strategies such as improving overall well-being and using workarounds. A full list of strategies is presented in Multimedia Appendix 3.

Discussion

Principal Findings

This study provides a comprehensive exploration of junior doctors' experiences of alert fatigue in hospital settings. Importantly, alert fatigue was found to manifest at different stages of information processing where participants failed to detect alerts, superficially processed alerts, developed mental models, and used excessive cognitive effort to interpret alerts, which led to different response patterns. These experiences were perceived to impact patient safety, such as increasing the risk of errors and contributing to patient harm, and clinician well-being, such as workflow interruptions and reduced efficiency.

The SEIPS model highlighted complex interactions between factors contributing to experiences of alert fatigue, including alert logic and design, organizational culture, and individual traits and emotional states. We observed junior doctors at the same hospital, who received similar alerts, to experience alert fatigue differently depending on their experience, role, and personality. For example, where junior doctors had positive attitudes and alerts were tightly aligned with responsibilities, even frequently disregarded alerts were not perceived as fatiguing. While poor clinical relevance remains a fundamentally important factor in the development of alert fatigue, our study reveals that doctors may have different levels of tolerance depending on these personal, organizational, and task factors. We also observed feedback loops between factors contributing to alert fatigue and its outcomes. For instance, repeated exposure to low-value alerts could lead to negative attitudes, subsequently reducing trust and influencing attitudes toward future alerts.

Expanding Definitions of Alert Fatigue

In our study, alert fatigue was rarely experienced in an all-or-nothing fashion. Instead, participants often described alert fatigue to arise in specific contexts, such as when under time pressure or for certain alert types. While some participants reported generalized negative perceptions of alerts, none described being universally fatigued or entirely unaffected. Similar to existing studies, we also found alert fatigue to develop for both passive and interruptive alerts [35].

Our findings support the view that alert fatigue is not simply an action, but a mental state that is subjectively experienced at conscious and subconscious levels [16,29]. Participants described 3 distinct information processing pathways in which alert fatigue occurred: one where alert stimuli were not perceived at all, another in which alerts were recognized but superficially processed based on mental shortcuts, and a third where alerts required high cognitive effort to interpret and respond to. The first 2 pathways, in line with existing conceptualizations of alert fatigue, were often associated with alerts being rapidly dismissed [4,12,13]. Despite this, mental shortcuts were perceived as an efficient way to process information and were often not associated with conscious experiences of fatigue. While many participants believed they were able to rapidly interpret alerts, patient safety risks could however still arise due to the potential to misinterpret information, particularly when the alerts lacked visual distinctiveness.

In contrast, the third pathway was consistently described as fatiguing. Participants reported feeling frustrated due to the cognitive burden of alerts and related impacts on efficiency, which arose particularly where alerts were ambiguous, time-consuming, or difficult to action. For certain alert types or individuals, this burden was described to gradually lead to cognitive disengagement, with clinicians eventually defaulting to rapid override as a strategy to avoid further mental effort. This finding mirrors alarm fatigue literature that suggests fatigue develops as a coping mechanism for cognitive burden [30,36]. While burnout is frequently linked to alert fatigue [37], our findings suggest that rapidly overriding alerts may conversely serve as a protective mechanism to manage cognitive load and avoid burnout, consistent with literature linking longer EHR use to increased fatigue and burnout [38,39].

These findings suggest that override rates and response times alone are insufficient measures of alert fatigue. Aligning with prior critiques, our work confirms that frequently overridden alerts are sometimes still attended to and can be perceived as helpful [13,40]. Similarly, although quick dismissals can indicate alert fatigue, they can also represent efficient pattern recognition, while longer interactions could both indicate appropriate engagement with alerts and signal cognitive burden. We therefore recommend supplementing system metrics with clinician self-reports to better understand both conscious and subconscious experiences of fatigue and applying solutions that appropriately address the specific type or types of fatigue experienced.

XSL•FO
RenderX

Work System Factors Contributing to Alert Fatigue

In line with existing literature, we found exposure to clinically irrelevant, repetitive, and poorly timed alerts contributes to alert fatigue [4,15,21,41,42]. Consistent with quantitative studies, the total volume of system alerts was rarely identified as a cause of alert fatigue alone, but instead became problematic when clinicians received many clinically irrelevant and repetitive alerts, and multiple alerts within a single clinical encounter [4]. We also identified novel influences, including personality traits, mood, social learning, and exposure to external alerts. Many participants described cognitive overload as a result of not just system alerts, but also notifications from pagers, emails, phone calls, and verbal prompts, which cumulatively contributed to experiences of fatigue and dismissal of system alerts. While existing literature has sought to improve total alert burden to reduce alert fatigue [43,44], our study highlights the need to consider system alerts within the broader ecosystem of alerts and tasks that clinicians receive.

Social and organizational factors are underexplored in existing alert fatigue literature but were found to collectively shape experiences in our study. Junior doctors often looked to peers to guide their own responses to alerts and expectations of their roles, affecting not just alert responses but also the framing of alerts as helpful or annoying. Organizational and cultural expectations of efficiency further increased alert fatigue, as managing alerts was often seen as a barrier to achieving other clinical tasks. While alerts are likely to offer the most benefit to junior doctors, our study substantiates prior research that suggests this group experiences more fatigue than their senior counterparts [45]. Participants with more clinical experience had greater confidence and therefore could interpret and respond to alerts with less cognitive effort, whereas those with less experience reported more ambiguity and higher effort expended in interpreting information.

Recommendations for Addressing Alert Fatigue

Our results highlighted several strategies for reducing alert fatigue that align with prior literature, including removing ineffective alerts and increasing clinical relevance [46], incorporating contextual awareness [47], applying human factors design principles [48,49], and implementing alerts in line with the 5 rights of CDS (right information, person, format, channel, and time) [6,7,50,51]. Our findings build on this literature by providing insight into how and why strategies are likely to be effective, and how they can be targeted to the different causes and impacts of alert fatigue. Several studies testing these strategies have reported their effectiveness in decreasing the overall volume and increasing the acceptance rate of alerts [50]. However, we argue that future research must also evaluate whether clinicians' *experiences* of fatigue are reduced. Given that our results showed junior doctors are likely to experience greater levels of alert fatigue than more senior doctors, we emphasize the importance of including this group in redesign efforts and evaluation of fatigue reduction.

We also identify and present new strategies that could improve experiences of alert fatigue. These include allowing clinicians to personalize alerts, education on alert fatigue, supporting overall emotional well-being, and streamlining the system and other alerts. Due to the individual variability in experiences of alert fatigue, personalization may be particularly valuable in fostering a greater sense of control and responsibility, thus improving engagement with alerts that individuals deem to be relevant. Further research is needed, however, to evaluate the effectiveness of personalization and other strategies in reducing alert fatigue and improving patient safety.

Interestingly, in some cases, experiences of fatigue were not associated with the presence of alerts, but by their absence, where critical information was difficult and time-consuming to locate. This finding underscores the need for balance between over- and underalerting, ensuring alerts effectively address the clinical or safety need. Implementers should carefully consider the issue they aim to target, evaluating whether CDS and alerts specifically are an appropriate solution, or if different strategies or forms of CDS, such as passive alerts or search functionality, should instead be considered [3,6].

Limitations

Our participants were junior doctors in urban Australian hospitals who primarily used commercial information systems. Findings therefore may not generalize to other clinician groups, specialties, or settings, such as rural hospitals or those using homegrown systems. Additionally, we did not incorporate observational or system usage data, which may limit insights into subconscious behaviors. Future studies should incorporate mixed methods, that is, both clinicians' experiences of alert fatigue and actual responses to alerts and explore experiences of alert fatigue among different user groups.

To our knowledge, participants in our study used only knowledge-based alerts, limiting conclusions regarding experiences of alert fatigue for smarter alerts, such as those that are artificial intelligence (AI)-based. Our results can, however, offer important considerations. Although AI-based alerts can significantly improve clinical relevance and specificity [52], they may also increase cognitive burden if information is complex or constantly changing. Alternatively, given the fast-paced nature of the setting in which alerts are presented, AI-based alerts may be superficially processed and, due to their high positive predictive value, be automatically accepted rather than rejected due to automation bias [53]. Further research is needed to explore how alert fatigue manifests for smarter alerts in hospital settings and its impact on patient safety.

Conclusions

This study highlights alert fatigue as a dynamic, context-dependent experience that is not easily captured through studying behavior alone. Our findings reinforce the need to move beyond using singular metrics to measure alert fatigue and instead address alert fatigue using tailored strategies that target the specific causes and outcomes involved. While reducing low-value alerts and improving interface design remain important, broader approaches that support clinician well-being, foster positive organizational culture, and consider the impact of competing priorities and tasks may be equally critical.

XSL•FO
RenderX

Funding

NN is supported by a Digital Health Cooperative Research Centre (CRC) Limited scholarship. Digital Health CRC Limited is funded under the Australian Commonwealth's CRC Program. The funder had no involvement in the study design, data collection, analysis, interpretation, or writing of the manuscript.

Data Availability

The data analyzed in this study are available from the corresponding author upon reasonable request.

Authors' Contributions

NN (in consultation with all authors) conceptualized the study. NN collected data, transcribed data, and prepared data for the analysis. NN, MTB, ABA, and RF analyzed the data. NN drafted the manuscript, with all authors providing feedback. All authors provided input into study conception, design, and interpretation of findings and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semistructured interview guide.
[DOCX File, 28 KB - jmir_v28i1e78676_app1.docx]

Multimedia Appendix 2
Proportion of coded references related to alert fatigue mapped to each Systems Engineering Initiative for Patient Safety (SEIPS) work system domain.
[DOCX File, 45 KB - jmir_v28i1e78676_app2.docx]

Multimedia Appendix 3
Strategies discussed by study participants for reducing alert fatigue in practice.
[DOCX File, 24 KB - jmir_v28i1e78676_app3.docx]

Checklist 1
SRQR checklist.
[PDF File, 153 KB - jmir_v28i1e78676_app4.pdf]

References

1. Wright A, Aaron S, Seger DL, Samal L, Schiff GD, Bates DW. Reduced effectiveness of interruptive drug-drug interaction alerts after conversion to a commercial electronic health record. J Gen Intern Med 2018 Nov;33(11):1868-1876. [doi: 10.1007/s11606-018-4415-9] [Medline: 29766382]
2. Kane-Gill SL, O'Connor MF, Rothschild JM, et al. Technologic distractions (part 1): summary of approaches to manage alert quantity with intent to reduce alert fatigue and suggestions for alert fatigue metrics. Crit Care Med 2017 Sep;45(9):1481-1488. [doi: 10.1097/CCM.0000000000002580] [Medline: 28682835]
3. Awad S, Loveday T, Lau R, Baysari MT. Development of a human factors–based guideline to support the design, evaluation, and continuous improvement of clinical decision support. Mayo Clin Proc Digit Health 2025 Mar;3(1):100182. [doi: 10.1016/j.mcpdig.2024.11.003] [Medline: 40206997]
4. Ancker JS, Edwards A, Nosal S, et al. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak 2017 Apr 10;17(1):36. [doi: 10.1186/s12911-017-0430-8] [Medline: 28395667]
5. Newton N, Bamgboje-Ayodele A, Forsyth R, Tariq A, Baysari MT. A systematic review of clinicians' acceptance and use of clinical decision support systems over time. NPJ Digit Med 2025 May 26;8(1):309. [doi: 10.1038/s41746-025-01662-7] [Medline: 40419669]
6. Blecker S, Pandya R, Stork S, et al. Interruptive versus noninterruptive clinical decision support: usability study. JMIR Hum Factors 2019 Apr 17;6(2):e12469. [doi: 10.2196/12469] [Medline: 30994460]
7. Osheroff J, Teich J, Levick D, et al. Improving Outcomes With Clinical Decision Support: An Implementer's Guide, 3rd edition: HIMSS Publishing; 2012:348.
8. Poly TN, Islam MM, Yang HC, Li YCJ. Appropriateness of overridden alerts in computerized physician order entry: systematic review. JMIR Med Inform 2020 Jul 20;8(7):e15653. [doi: 10.2196/15653] [Medline: 32706721]
9. Bittmann JA, Haefeli WE, Seidling HM. Modulators influencing medication alert acceptance: an explorative review. Appl Clin Inform 2022 Mar;13(2):468-485. [doi: 10.1055/s-0042-1748146] [Medline: 35981555]

XSL•FO
RenderX

10. Elias P, Peterson E, Wachter B, Ward C, Poon E, Navar AM. Evaluating the impact of interruptive alerts within a health system: use, response time, and cumulative time burden. Appl Clin Inform 2019 Oct;10(5):909-917. [doi: 10.1055/s-0039-1700869] [Medline: 31777057]
11. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17. [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
12. McDaniel RB, Burlison JD, Baker DK, et al. Alert dwell time: introduction of a measure to evaluate interruptive clinical decision support alerts. J Am Med Inform Assoc 2016 Apr;23(e1):e138-e141. [doi: 10.1093/jamia/ocv144] [Medline: 26499101]
13. McGreevey JD III, Mallozzi CP, Perkins RM, Shelov E, Schreiber R. Reducing alert burden in electronic health records: state of the art recommendations from four health systems. Appl Clin Inform 2020 Jan;11(1):1-12. [doi: 10.1055/s-0039-3402715] [Medline: 31893559]
14. Embi PJ, Leonard AC. Evaluating alert fatigue over time to EHR-based clinical trial alerts: findings from a randomized controlled study. J Am Med Inform Assoc 2012 Jun;19(e1):e145-e148. [doi: 10.1136/amiajnl-2011-000743] [Medline: 22534081]
15. Wang L, Goh KH, Yeow A, et al. Habit and automaticity in medical alert override: cohort study. J Med Internet Res 2022 Feb 16;24(2):e23355. [doi: 10.2196/23355] [Medline: 35171102]
16. Baysari MT, Tariq A, Day RO, Westbrook JI. Alert override as a habitual behavior—a new perspective on a persistent problem. J Am Med Inform Assoc 2017 Mar 1;24(2):409-412. [doi: 10.1093/jamia/ocw072] [Medline: 27274015]
17. Ash JS, Sittig DF, Campbell EM, Guappone KP, Dykstra RH. Some unintended consequences of clinical decision support systems. AMIA Annu Symp Proc 2007 Oct 11;2007:26-30. [Medline: 18693791]
18. Beurskens AJ, Bültmann U, Kant I, Vercoulen JH, Bleijenberg G, Swaen GM. Fatigue among working people: validity of a questionnaire measure. Occup Environ Med 2000 May;57(5):353-357. [doi: 10.1136/oem.57.5.353] [Medline: 10769302]
19. Ford E, Edelman N, Somers L, et al. Barriers and facilitators to the adoption of electronic clinical decision support systems: a qualitative interview study with UK general practitioners. BMC Med Inform Decis Mak 2021 Jun 21;21(1):193. [doi: 10.1186/s12911-021-01557-z] [Medline: 34154580]
20. Feldstein A, Simon SR, Schneider J, et al. How to design computerized alerts to safe prescribing practices. Jt Comm J Qual Saf 2004 Nov;30(11):602-613. [doi: 10.1016/s1549-3741(04)30071-7] [Medline: 15565759]
21. Russ AL, Zillich AJ, McManus MS, Doebbeling BN, Saleem JJ. Prescribers' interactions with medication alerts at the point of prescribing: a multi-method, in situ investigation of the human-computer interaction. Int J Med Inform 2012 Apr;81(4):232-243. [doi: 10.1016/j.ijmedinf.2012.01.002] [Medline: 22296761]
22. Holden RJ, Carayon P, Gurses AP, et al. SEIPS 2.0: a human factors framework for studying and improving the work of healthcare professionals and patients. Ergonomics 2013;56(11):1669-1686. [doi: 10.1080/00140139.2013.838643] [Medline: 24088063]
23. Russ AL, Fairbanks RJ, Karsh BT, Militello LG, Saleem JJ, Wears RL. The science of human factors: separating fact from fiction. BMJ Qual Saf 2013 Oct;22(10):802-808. [doi: 10.1136/bmjqs-2012-001450] [Medline: 23592760]
24. Carayon P, Schoofs Hundt A, Karsh BT, et al. Work system design for patient safety: the SEIPS model. Qual Saf Health Care 2006 Dec;15 Suppl 1(Suppl 1):i50-i58. [doi: 10.1136/qshc.2005.015842] [Medline: 17142610]
25. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
26. Adams WC. Conducting semi‐structured interviews. In: Newcomer KE, Hatry HP, Wholey JS, editors. Handbook of Practical Program Evaluation: Wiley; 2015:492-505. [doi: 10.1002/9781119171386]
27. Fossey E, Harvey C, McDermott F, Davidson L. Understanding and evaluating qualitative research. Aust N Z J Psychiatry 2002 Dec;36(6):717-732. [doi: 10.1046/j.1440-1614.2002.01100.x] [Medline: 12406114]
28. Wickens CD, Carswell CM. Information processing. In: Salvendy G, Karwowski W, editors. Handbook of Human Factors and Ergonomics, 5th edition: Wiley; 2021:114-158. [doi: 10.1002/9781119636113]
29. Joseph AL, Borycki EM, Kushniruk AW. Alert fatigue and errors caused by technology: a scoping review and introduction to the flow of cognitive processing model. Knowl Manag E-Learn 2021 Dec 30(4):500-521. [doi: 10.34105/j.kmel.2021.13.027]
30. Movahedi A, Sadooghiasl A, Ahmadi F, Vaismoradi M. A grounded theory study of alarm fatigue among nurses in intensive care units. Aust Crit Care 2023 Nov;36(6):980-988. [doi: 10.1016/j.aucc.2022.12.004] [Medline: 36737263]
31. Otter AI. 2025. URL: https://otter.ai/ [accessed 2026-02-03]
32. Saunders B, Sim J, Kingstone T, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907. [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
33. NVivo 14.24.1. Lumivero. 2024. URL: https://lumivero.com/products/nvivo/ [accessed 2026-02-03]
34. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
35. Kizzier-Carnahan V, Artis KA, Mohan V, Gold JA. Frequency of passive EHR alerts in the ICU: another form of alert fatigue? J Patient Saf 2019 Sep;15(3):246-250. [doi: 10.1097/PTS.0000000000000270] [Medline: 27331600]

XSL•FO
RenderX

36. Obisesan O, Barber E, Martin P, Brougham N, Tymkew H. Original research: alarm fatigue: exploring the adaptive and maladaptive coping strategies of nurses. Am J Nurs 2024 Oct 1;124(10):24-30. [doi: 10.1097/01.NAJ.0001063808.07614.8d] [Medline: 39259867]
37. Gani I, Litchfield I, Shukla D, Delanerolle G, Cockburn N, Pathmanathan A. Understanding "alert fatigue" in primary care: qualitative systematic review of general practitioners' attitudes and experiences of clinical alerts, prompts, and reminders. J Med Internet Res 2025 Feb 7;27:e62763. [doi: 10.2196/62763] [Medline: 39918864]
38. Khairat S, Coleman C, Ottmar P, Jayachander DI, Bice T, Carson SS. Association of electronic health record use with physician fatigue and efficiency. JAMA Netw Open 2020 Jun 1;3(6):e207385. [doi: 10.1001/jamanetworkopen.2020.7385] [Medline: 32515799]
39. Wu Y, Wu M, Wang C, Lin J, Liu J, Liu S. Evaluating the prevalence of burnout among health care professionals related to electronic health record use: systematic review and meta-analysis. JMIR Med Inform 2024 Jun 12;12:e54811. [doi: 10.2196/54811] [Medline: 38865188]
40. McCoy AB, Waitman LR, Lewis JB, et al. A framework for evaluating the appropriateness of clinical decision support alerts and responses. J Am Med Inform Assoc 2012;19(3):346-352. [doi: 10.1136/amiajnl-2011-000185] [Medline: 21849334]
41. van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. J Am Med Inform Assoc 2006;13(2):138-147. [doi: 10.1197/jamia.M1809] [Medline: 16357358]
42. Salwei ME, Carayon P, Hoonakker PLT, et al. Workflow integration analysis of a human factors–based clinical decision support in the emergency department. Appl Ergon 2021 Nov;97:103498. [doi: 10.1016/j.apergo.2021.103498] [Medline: 34182430]
43. Murad DA, Tsugawa Y, Elashoff DA, Baldwin KM, Bell DS. Distinct components of alert fatigue in physicians' responses to a noninterruptive clinical decision support alert. J Am Med Inform Assoc 2022 Dec 13;30(1):64-72. [doi: 10.1093/jamia/ocac191] [Medline: 36264258]
44. Orenstein EW, Kandaswamy S, Muthu N, et al. Alert burden in pediatric hospitals: a cross-sectional analysis of six academic pediatric health systems using novel metrics. J Am Med Inform Assoc 2021 Nov 25;28(12):2654-2660. [doi: 10.1093/jamia/ocab179] [Medline: 34664664]
45. Baysari MT, Dort BAV, Zheng WY, et al. Prescribers' reported acceptance and use of drug-drug interaction alerts: an Australian survey. Health Informatics J 2022;28(2):14604582221100678. [doi: 10.1177/14604582221100678] [Medline: 35531625]
46. Blackley SV, Lo YC, Varghese S, et al. Building an allergy reconciliation module to eliminate allergy discrepancies in electronic health records. J Am Med Inform Assoc 2025 Apr 1;32(4):648-655. [doi: 10.1093/jamia/ocaf022] [Medline: 39969950]
47. Chien SC, Yang HC, Chen CY, et al. Using alert dwell time to filter universal clinical alerts: a machine learning approach. Comput Methods Programs Biomed 2023 Oct;240:107696. [doi: 10.1016/j.cmpb.2023.107696] [Medline: 37480643]
48. Zachariah M, Phansalkar S, Seidling HM, et al. Development and preliminary evidence for the validity of an instrument assessing implementation of human-factors principles in medication-related decision-support systems—I-MeDeSA. J Am Med Inform Assoc 2011 Dec;18 Suppl 1(Suppl 1):i62-i72. [doi: 10.1136/amiajnl-2011-000362] [Medline: 21946241]
49. Phansalkar S, Zachariah M, Seidling HM, Mendes C, Volk L, Bates DW. Evaluation of medication alerts in electronic health records for compliance with human factors principles. J Am Med Inform Assoc 2014 Oct;21(e2):e332-e340. [doi: 10.1136/amiajnl-2013-002279] [Medline: 24780721]
50. Ng HJH, Kansal A, Abdul Naseer JF, et al. Optimizing Best Practice Advisory alerts in electronic medical records with a multi-pronged strategy at a tertiary care hospital in Singapore. JAMIA Open 2023 Aug 1;6(3):ooad056. [doi: 10.1093/jamiaopen/ooad056] [Medline: 37538232]
51. Hussain MI, Reynolds TL, Zheng K. Medication safety alert fatigue may be reduced via interaction design and clinical role tailoring: a systematic review. J Am Med Inform Assoc 2019 Oct 1;26(10):1141-1149. [doi: 10.1093/jamia/ocz095] [Medline: 31206159]
52. Graafsma J, Murphy RM, van de Garde EMW, et al. The use of artificial intelligence to optimize medication alerts generated by clinical decision support systems: a scoping review. J Am Med Inform Assoc 2024 May 20;31(6):1411-1422. [doi: 10.1093/jamia/ocae076] [Medline: 38641410]
53. Abdelwanis M, Alarafati HK, Tammam MMS, Simsekler MCE. Exploring the risks of automation bias in healthcare artificial intelligence applications: a Bowtie analysis. J Saf Sci Resil 2024 Dec;5(4):460-469. [doi: 10.1016/j.jnlssr.2024.06.001]

Abbreviations

AI: artificial intelligence
CDS: clinical decision support
CRC: Cooperative Research Centre
EHR: electronic health record
EMR: electronic medical record

XSL•FO
RenderX

JMO: junior medical officer
PGY: postgraduate year
SEIPS: Systems Engineering Initiative for Patient Safety
SRQR: Standards for Reporting Qualitative Research

Edited by A Stone, TDA Cardoso; submitted 06.Jun.2025; peer-reviewed by J Aarts, PV Asaro, R Marshall, S Lloyd; revised version received 27.Oct.2025; accepted 11.Nov.2025; published 19.Feb.2026.

Please cite as:
Newton N, Bamgboje-Ayodele A, Forsyth R, Tariq A, Huang J, Yannam R, Lalor DJ, Sobey AJ, Baysari MT
Experiences of Alert Fatigue and Its Contributing Factors in Hospitals: Qualitative Study
J Med Internet Res 2026;28:e78676
URL: https://www.jmir.org/2026/1/e78676
doi:10.2196/78676

© Nicki Newton, Adeola Bamgboje-Ayodele, Rowena Forsyth, Amina Tariq, Jing Huang, Ramyasri Yannam, Daniel J Lalor, Andrew James Sobey, Melissa T Baysari. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Improving the Understandability of Clinical Guidelines: Development and Evaluation of a GPT-4–Based Pipeline

Matthew D Jones[1], PhD; Melissa Torgbi[2], MSc; Harish Tayyar Madabushi[2], PhD

[1]Department of Life Sciences, University of Bath, Claverton Down, Bath, United Kingdom

[2]Department of Computer Science, University of Bath, Bath, United Kingdom

Corresponding Author:
Matthew D Jones, PhD
Department of Life Sciences, University of Bath, Claverton Down, Bath, United Kingdom

Abstract

Background: Difficulty in finding and understanding information in clinical guidelines contributes to medication errors. Large language models (LLMs) can simplify complex text to aid in understanding, but this approach to improving the quality of guidelines has not been investigated. However, LLMs are also known to hallucinate or generate outputs that may not align with reality.

Objective: This study aimed to develop and evaluate an LLM pipeline to improve the readability of clinical guidelines while ensuring the preservation of critical content.

Methods: To align LLM revisions with research evidence and enable comparison with manual editing, the National Health Service Injectable Medicines Guide (IMG) was used as a case study, to which a GPT-4–based pipeline was applied, with prompts based on user testing–derived recommendations for IMG authors. This enabled readability comparisons between various IMG guideline versions: original, manually revised, or GPT-4–revised using the user testing–derived recommendations, and fully user tested. Readability was evaluated using readability metrics and ratings from 3 expert pharmacists. Content similarity before and after LLM revision was assessed using BERT (bidirectional encoder representations from transformers) scores and expert pharmacist review.

Results: Considering 20 IMG guidelines used in practice, BERT scores indicated high semantic similarity between the original and LLM-revised guidelines (0.88-0.96). An omission, addition, or change in meaning was identified by at least one pharmacist in 30 (20%), 7 (5%), and 18 (12%) of the 153 guideline subsections, respectively. The SMOG (Simple Measure of Gobbledygook) grade showed a small but significant improvement in readability for the LLM-revised guidelines (mean difference 0.32, 95% CI 0.10 - 0.55; *P*=.02) and the manually revised versions (mean difference 0.46, 95% CI 0.13 - 0.79; *P*=.03). There was no significant difference between the LLM and manually revised versions (*P*>.99). There were no significant differences between Flesch-Kincaid reading grades (*P*=.91). Expert ratings favored the LLM-revised versions for understandability. Considering 2 IMG guidelines from previous research, user testing produced a greater improvement in readability than LLM revision.

Conclusions: Authors should not use current LLMs to modify clinical guidelines without carefully checking the revised text for unintended omissions, additions, or changes in meaning. Further work should investigate the potential of LLMs to augment manual user testing and reduce the barriers to the wider use of this approach to improve the safety of clinical guidelines.

(J Med Internet Res 2026;28:e81915) doi:10.2196/81915

KEYWORDS

artificial intelligence; clinical guidelines; comprehension; guidelines as topic; health care professionals; health personnel; large language models; patient safety; readability; understandability

Introduction

Medication errors are a leading cause of avoidable harm in health care systems, with an estimated international cost of US $42 billion per annum [1]. Numerous factors can lead to a medication error [2], one of which is difficulty in finding or understanding prescribing and medication administration information in clinical guidelines [3-5]. Consequently, guidelines that are contradictory, incomprehensible, or of poor quality are included in frameworks of the causes of patient safety incidents [6]. Clarity of writing and formatting also influences guideline adoption [7].

Few studies have explored ways to help clinicians more easily find and understand guidelines. Some evidence suggests that iterative user testing and redesign can improve guideline usability by identifying and addressing problem areas [8,9]. In one study, user-tested injectable medicines guidelines led to a 2.5-fold increase in error-free intravenous administration and a 96% probability of cost-effectiveness [10,11]. Such improvements result from both the testing process and the

XSL•FO
RenderX

application of good practice in information writing and design [12]. Consequently, in 2020, format changes and clear writing advice were added to the authors' instructions for the United Kingdom's National Health Service Injectable Medicines Guide (IMG). This is an online guide to preparing and administering injectable medicines, mainly used by nurses. Subsequently, more than 300 drug guidelines have been manually updated, but not user tested.

Most guideline-related medication errors involve documents produced by local health care organizations [3]. User testing and revising these "local guidelines" could therefore yield the greatest safety improvements. As there are many thousands of these documents within a single country, limited expertise in information writing and design within health care organizations may hinder this approach.

Large language models (LLMs), such as GPT-4, may help overcome this barrier by simplifying complex text. This approach has been tested on health information for patients initially written by subject experts. Early results are mixed: LLMs can improve readability metrics [13-18] and patient understanding [15] but can also worsen readability metrics [19]. These gains may come at a cost, with some content omitted and up to 21% of revised content deemed clinically inappropriate [13-15,19]. LLMs have also been used to improve understanding of specialist medical records [20]. However, they have not yet been applied to improve the readability of clinical guidelines for health care professionals; therefore, it remains unclear whether LLMs can enhance readability without compromising content accuracy or safety.

Therefore, this study evaluated a GPT-4–based pipeline to improve guideline readability while safeguarding against omission, addition, or change of meaning of safety-critical content. To enable comparison with manual revision and align with previous research, the pipeline was applied to the IMG. This enabled various readability comparisons (Figure 1), as well as comparison of the content of guidelines before and after LLM revision.

Specific objectives were therefore to (1) develop a GPT-4–based pipeline to improve guideline clarity with built-in safeguards; (2) evaluate content similarity of 20 guidelines before and after revision by the GPT-4 pipeline; (3) compare the readability of 20 equivalent original guidelines, LLM-revised guidelines, and manually revised guidelines; and (4) compare the readability of original, LLM-revised, and user-tested aminophylline and voriconazole guidelines.

Figure 1. Readability comparisons reported in this study. "Original" versions were the guidelines in use in the Injectable Medicines Guide (IMG) in 2019. "LLM-revised" versions were developed from the original guidelines by the large language model (LLM) pipeline. "Manually revised" versions were the guidelines in use in the IMG in 2024, which had been updated by guideline authors in line with revised authors' instructions introduced in 2020. "User-tested" versions were developed from the original guidelines via a user-testing process during a previous study [9].

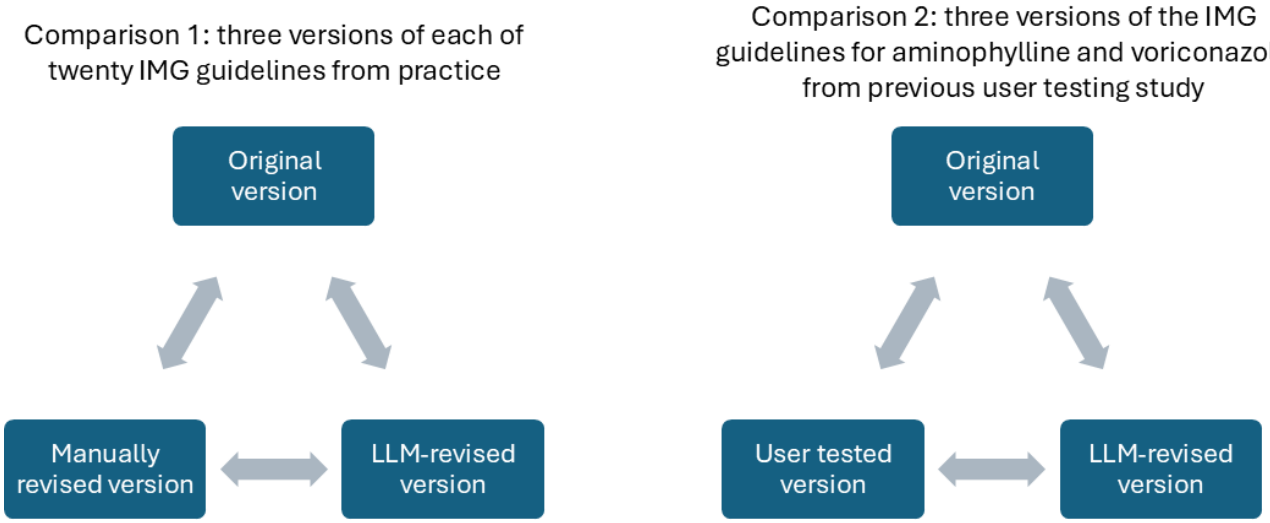

Methods

This study is reported in accordance with the TRIPOD-LLM (Transparent Reporting of a Multivariable Model for Individual Prognosis or Diagnosis - Large Language Model) guideline [21].

Development and Application of the GPT-4 Pipeline

IMG guidelines were converted to Markdown and paired with iteratively developed prompts for input to the LLM pipeline. Prompts instructed the LLM to improve clarity and readability without altering content. The LLM pipeline used the OpenAI API and the GPT-4 model. GPT-4–generated "LLM-revised" Markdown versions of the original guidelines, which were then converted to static HTML for display.

Prompting Strategies

Prompting was based on the Reframing Principles for Instruction-Following LLMs [22], which offer tested strategies for constructing effective prompts, including using low-level patterns, examples, simplified sentence structures, and clear constraints. Initially, researchers with expertise in computer science (MT and HTM) constructed prompts based on the user testing–derived recommendations incorporated into the IMG authors' instructions in 2020 [9]. To prevent omissions or changes in meaning, prompts required a semantic similarity score above 0.9 before proceeding. While GPT-4 cannot

XSL•FO
RenderX

compute semantic accuracy, we hypothesized that it would interpret this instruction to ensure edits were semantically close to the original and that this reference would make the requirement more explicit. Selected outputs of these initial prompts were reviewed by a pharmacist with expertise in intravenous medication administration and information design (MDJ) and discussed with the wider research team. These initial tests found that overly prescriptive instructions for human authors caused overgeneralization and errors, including omissions and additions. Prompt instructions were then iteratively refined to improve effectiveness. The final prompts (Multimedia Appendix 1, online supplement) used for both comparisons were a product of this iterative development, incorporating insights from our explorations while omitting some instructions originally intended for human authors.

GPT-4–generated instructions were also explored for comparison 2 (Figure 1). The original and user-tested guidelines for voriconazole and aminophylline were given to GPT-4, which was instructed to generate prompts for another model to revise the original guidelines to be similar to the user-tested versions, without directly referencing them. However, these prompts were ineffective and were therefore excluded from final experiments.

Selection of IMG Guidelines

Twenty adult IMG guidelines were selected in consultation with the IMG editorial team to cover a range of drugs, administration methods (infusion and injection), reconstitution or dilution needs, and higher-risk or complex cases. Versions from 2019 ("original") and 2024 ("manually revised," updated in line with revised IMG authors' instructions) were obtained. Voriconazole and aminophylline guidelines from 2019 and their user-tested revisions were also sourced from prior research [9].

Guideline Content Similarity

Content similarity of the 20 IMG guidelines from practice before and after LLM revision was assessed using the bidirectional encoder representations from transformers (BERT) score [23] and expert review. The BERT score calculates the similarity between each word or token in a candidate and reference sentence using cosine similarity and the contextual embeddings from the BERT model [20]. We used the standard BERT score library to perform this calculation [24]. Scores range from 0 to 1, with higher values indicating greater similarity.

Three intensive care pharmacists independently assessed 9 sections of LLM-revised guidelines (method of administration, reconstitution, dilution, expiry, flushing, adverse effects, extravasation, other comments, and compatibility) for clinically relevant omissions, additions, or changes in meaning compared with the originals. The pharmacists were known to the research team through professional networks and selected for their significant experience in intensive care pharmacy, although this was not formally defined. All invited pharmacists participated and were aware of the study aims, although they were blinded to guideline version while assessing content.

In addition, the outputs were evaluated manually by the research team for modifications to medication names or their order of presentation.

Guideline Readability

Readability was assessed with 2 metrics and expert review. The readability metrics used were the SMOG (Simple Measure of Gobbledygook) grade and the Flesch-Kincaid reading grade. SMOG was chosen for its validation against full comprehension, while Flesch-Kincaid is widely used in health research [25]. Both estimate the years of education required to understand a text, so larger scores indicate lower readability. Scores were calculated with the Textstat Python library using the entire guideline text.

The relative readability of the 20 IMG guidelines from practice before and after LLM revision was also independently assessed by the 3 intensive care pharmacists. Comparing the original and revised versions of each guideline, they rated relative overall and subsection-level understandability using a 5-point Likert scale.

Statistical Analysis

Analyses were performed in Stata v18.0 (StataCorp LLC). Readability metrics and BERT scores were summarized as mean (SD), with differences tested using repeated measures ANOVA with post-hoc Bonferroni-corrected paired *t* tests. Expert ratings were summarized descriptively, and interrater reliability assessed using Gwet's AC to account for skewed distributions [26].

Ethical Considerations

This study was assessed as having low potential to do harm and thus reviewed and given a favorable opinion by proportionate review by the Department of Life Science Departmental Research Ethics Officer at the University of Bath on June 22, 2023 (number 442).

Results

Pipeline Development

The LLM pipeline successfully produced output text showing revised versions of the 20 IMG guidelines from practice and the voriconazole and aminophylline guidelines from previous user-testing research. These revised guidelines were then evaluated for content similarity using the methods described earlier.

Content Similarity of the Set of 20 Guidelines From Practice

Table 1 presents the BERT scores comparing the semantic similarity of the 20 IMG guidelines from practice before and after LLM revision. These ranged from 0.88 to 0.96, indicating a high degree of similarity. As a benchmark, Hanna and Bojar [27] reported an average BERT score of 0.815 when comparing 2 different, professional human translations of the same source sentences across several distinct linguistic phenomena. The 3 pharmacists' ratings of clinically relevant omissions, additions, or changes in meaning following LLM revision of the 20 guidelines from practice showed a high degree of interrater reliability, with Gwet's ACs of 0.87 (omissions), 0.97 (additions), and 0.92 (changes in meaning). Figures 2-4 show the number of pharmacist reviewers who identified an omission, addition, or change in meaning in each subsection. Of 459

XSL•FO
RenderX

individual subsection ratings (153 subsections each rated by 3 pharmacists), 54 (11.8%) identified an omission, 8 (1.7%) identified an addition, and 20 (4.4%) identified a change in meaning. An omission, addition, or change in meaning was identified by at least one pharmacist in 30 (20%), 7 (5%), and 18 (12%), respectively, of the 153 individual subsections. However, only 8 subsections had an omission identified by all 3 reviewers, and no additions or changes of meaning were identified by all 3 reviewers (Table 2). Overall, 65% (36/55) of subsection omissions, additions, or changes of meaning were only identified by 1 reviewer. These differences were concentrated in certain subsections (method of administration, reconstitution, dilution, adverse effects, and other comments) and guidelines (amiodarone, amoxicillin, furosemide, and propofol). Only one subsection (flushing) and two guidelines (levetiracetam and paracetamol) had no omissions, additions, or changes in meaning identified by any pharmacist.

Table . BERT (bidirectional encoder representations from transformers) scores comparing the semantic similarity of the 20 Injectable Medicines Guide guidelines from practice before and after large language model revision.[a]

Guideline	BERT score
Amiodarone	0.95
Amoxicillin	0.91
Ceftriaxone	0.95
Cyclizine	0.91
Fentanyl	0.96
Flucloxacillin	0.90
Furosemide	0.88
Gentamicin	0.93
Levetiracetam	0.92
Magnesium sulfate	0.90
Meropenem	0.93
Metronidazole	0.90
Noradrenaline	0.94
Omeprazole	0.94
Paracetamol	0.94
Phenytoin	0.88
Piperacillin and tazobactam	0.95
Propofol	0.91
Teicoplanin	0.93
Vancomycin	0.88

[a]Mean (SD): 0.92 (0.02).

XSL•FO
RenderX

Figure 2. Heatmap showing the number of pharmacist reviewers who identified an omission of information following large language model revision of each subsection of each of the 20 original guidelines. Blank subsections were not present in the relevant guideline. AE: adverse effects; Compat.: compatibility; Expiry: expiry time; Extrav.: extravasation; Flush: flushing; Method: method of administration; Other: other comments; Recon.: reconstitution.

Guideline	Subsection								
	Method	Recon.	Dilution	Expiry	Flush	AEs	Extrav.	Other	Compat.
Amiodarone	1		3	0	0	0	0	0	3
Amoxicillin	3	1	2		0	0	0	3	1
Ceftriaxone	0	1	0		0	0	0	0	0
Cyclizine	0		1		0	0	0	0	0
Fentanyl	0		0	0	0	0	0	0	3
Flucloxacillin	0	0	0		0	0	0	0	0
Furosemide	2		0	0	0	0	2	1	0
Gentamicin	0		0		0	0	0	0	0
Levetiracetam	0		0		0	0	0	0	0
Magnesium sulfate	0		0	0	0	0	0	0	0
Meropenem	0	0	1		0	0	0	0	0
Metronidazole	0				0	0	0	0	0
Noradrenaline	0		0	0	0	0	0	0	1
Omeprazole	0	1	3	0	0	0	0	0	1
Paracetamol	0				0	0	0	0	0
Phenytoin	0		0		0	0		3	0
Piperacillin and tazobactam	0	0	0		0	0	0	1	0
Propofol	1		0	0	0	0	0	1	2
Teicoplanin	0	2	0		0	0	0	2	0
Vancomycin	0	2	1	0	0	0	2	3	0

Figure 3. Heatmap showing the number of pharmacist reviewers who identified an addition of information following large language model revision of each subsection of each of the 20 original guidelines. Blank subsections were not present in the relevant guideline. AE: adverse effects; Compat.: compatibility; Expiry: expiry time; Extrav.: extravasation; Flush: flushing; Method: method of administration; Other: other comments; Recon.: reconstitution.

Guideline	Subsection								
	Method	Recon.	Dilution	Expiry	Flush	AEs	Extrav.	Other	Compat.
Amiodarone	0		0	0	0	0	0	0	0
Amoxicillin	0	0	0		0	2	0	1	0
Ceftriaxone	0	0	0		0	0	0	0	0
Cyclizine	0		0		0	0	0	0	0
Fentanyl	0		0	0	0	0	0	0	0
Flucloxacillin	0	0	1		0	0	0	0	0
Furosemide	0		0	0	0	0	0	0	0
Gentamicin	0		0		0	0	0	0	0
Levetiracetam	0		0		0	0	0	0	0
Magnesium sulfate	0		0	1	0	1	0	0	0
Meropenem	0	0	0		0	0	0	0	0
Metronidazole	0				0	0	0	0	0
Noradrenaline	0		0	0	0	0	0	0	0
Omeprazole	0	0	0	0	0	0	0	0	0
Paracetamol	0				0	0	0	0	0
Phenytoin	0		0		0	1		0	0
Piperacillin and tazobactam	0	0	0		0	0	0	0	0
Propofol	0		0	0	0	0	0	1	0
Teicoplanin	0	0	0		0	0	0	0	0
Vancomycin	0	0	0	0	0	0	0	0	0

XSL•FO
RenderX

Figure 4. Heatmap showing the number of pharmacist reviewers who identified a change in meaning following large language model revision of each subsection of each of the 20 original guidelines. Blank subsections were not present in the relevant guideline. AEs: adverse effects; Compat.: compatibility; Expiry: expiry time; Extrav.: extravasation; Flush: flushing; Method: method of administration; Other: other comments; Recon.: reconstitution.

Guideline	Subsection								
	Method	Recon.	Dilution	Expiry	Flush	AEs	Extrav.	Other	Compat.
Amiodarone	0		2	0	0	0	0	0	2
Amoxicillin	1	1	1		0	1	0	1	1
Ceftriaxone	0	0	0		0	0	0	0	0
Cyclizine	0		0		0	0	0	0	0
Fentanyl	0		0	0	0	0	0	0	1
Flucloxacillin	0	0	0		0	0	0	0	0
Furosemide	0		0	0	0	1	0	1	0
Gentamicin	1		0		0	1	0	0	0
Levetiracetam	0		0		0	0	0	0	0
Magnesium sulfate	0		0	0	0	0	0	0	0
Meropenem	0	1	0		0	0	0	0	0
Metronidazole	0				0	0	0	0	1
Noradrenaline	0		0	0	0	0	0	0	0
Omeprazole	0	0	1	0	0	0	0	0	0
Paracetamol	0				0	0	0	0	0
Phenytoin	0		0		0	0		0	0
Piperacillin and tazobactam	0	0	0		0	0	0	0	0
Propofol	0		0	0	0	0	0	1	0
Teicoplanin	1	0	0		0	0	0	0	0
Vancomycin	0	0	0	0	0	0	0	0	0

Table . The number of subsections with an identified omission, addition, or change of meaning following large language model revision of each of the 20 original guidelines.

	The number of subsections with an addition, omission, or change of meaning		
	Identified by 1 reviewer	Identified by 2 reviewers	Identified by 3 reviewers
Omission	14	8	8
Addition	6	1	0
Change	16	2	0

Key differences described in the pharmacists' free text comments included the occasional omission of the contents of an entire subsection (eg, extravasation, other comments, and compatibility), the omission of one specific detail (eg, information relating to one of several doses of a medicine), objective changes in meaning (eg, "hypotension" changed to "hypertension" and "mg/mL" changed to "mg") and subjective changes in interpretation (eg, the revised wording implying an adverse effect is probable rather than possible or lessening a safety warning).

Manual checks identified no alterations to medication names or order of presentation.

Readability of the 20 Guidelines From Practice

Table 3 presents overall readability metrics for the 3 different versions of the IMG from practice. Table S1 in Multimedia Appendix 1 shows the metrics calculated for each individual guideline. Repeated measures ANOVA identified a significant difference between the SMOG grades of at least 2 guideline versions (P=.007; df=59), so post hoc paired t tests with Bonferroni corrections were performed. The mean difference in SMOG grade between the original and LLM-revised versions was 0.32 (95% CI 0.10 - 0.55; P=.02; df=19), indicating a small but significant improvement in readability following LLM revision. Similarly, the mean difference in SMOG grade between the original and manually revised versions was 0.46 (95% CI 0.13 - 0.79; P=.03; df=19), indicating significant improvement in the readability of the guidelines following changes to the writing guide in 2020. However, the mean difference in SMOG grade between the LLM-revised and manually revised versions was 0.14 (95% CI −0.18 to 0.45; P>.99; df=19), suggesting no difference in the readability of these 2 versions. In contrast, repeated measures ANOVA did not identify differences between the Flesch-Kincaid reading grades of the original, LLM-revised, and manually revised versions (P=.91; df=59).

XSL•FO
RenderX

Table . Summary readability metrics for 3 different versions of injectable medicines guides from practice for 20 intravenous drugs.

	SMOG[a] grade, mean (SD)	Flesch-Kincaid grade, mean (SD)
Original version	12.3 (0.2)	12.1 (0.4)
LLM[b]-revised version	12.0 (0.2)	12.0 (0.4)
Manually revised version	11.8 (0.1)	11.9 (0.3)

[a]SMOG: Simple Measure of Gobbledygook.

[b]LLM: large language model.

The 3 pharmacists' ratings of the relative readability of the 20 guidelines from practice before and after LLM revision showed a high degree of interrater reliability, with Gwet's ACs of 0.75 (overall guideline) and 0.85 (subsections). Table 4 presents the pharmacists' ratings of the overall understandability of the guidelines, which favored the LLM-revised versions being easier to understand (26/60 ratings, 43.3%) compared with the original versions (11/60 ratings, 18.3%). Figure S1 (multimedia supplement) summarizes the pharmacists' ratings of the relative readability of each subsection. Of 459 individual ratings, 56 favored the original version and 125 favored the LLM-revised version. For 6 of the 9 subsections and 17 of the 20 drugs, there were more ratings that favored the LLM-revised version than the original version. The pharmacists' free text comments on relative readability frequently mentioned more concise language and a clearer layout for the LLM-revised versions resulting especially from the use of bullet points. However, they also noted that this was sometimes achieved through the omission of information.

Table . The 3 pharmacists' ratings of the overall relative understandability of the 20 original and large language model (LLM)–revised guidelines.

	Number of ratings, n (%)[a]
2019 version is much easier to understand	1 (1.7)
2019 version is slightly easier to understand	10 (16.7)
Both versions are equally easy to understand	23 (38.3)
LLM-revised version is slightly easier to understand	25 (41.7)
LLM-revised version is much easier to understand	1 (1.7)

[a]Three pharmacist reviewers each independently rated 20 guidelines, giving a total of 60 ratings.

Readability of the User- Tested Guidelines

Table 5 presents the readability metrics for 3 different versions of injectable medicines guides for aminophylline and voriconazole that underwent user testing and subsequent revision in a previous study [9]. For aminophylline, LLM revision did not decrease the reading grade, but following user testing and revision, both readability metrics fell. For voriconazole, LLM revision decreased both readability grades, but user testing and subsequent revision decreased them even further.

Table . Guideline readability metrics for three different versions of injectable medicines guides for aminophylline and voriconazole that underwent user testing and subsequent revision in a previous study [9].

	Original version	LLM[a]-revised version	User-tested version
SMOG[b] grade			
Aminophylline	11.7	11.8	11.0
Voriconazole	11.9	10.9	10.6
Flesch-Kincaid grade			
Aminophylline	9.7	9.8	8.3
Voriconazole	10.5	10.3	7.9

[a]LLM: large language model.

[b]SMOG: Simple Measure of Gobbledygook.

Discussion

Principal Findings

The GPT-4–based pipeline did not introduce clinically relevant omissions, additions, or changes of meaning to≥80% of individual subsections within the 20 guidelines from practice. However, content was entirely similar in only a few individual guidelines or subsection types, and differences (especially omissions) were concentrated in certain drugs or subsections in an unpredictable way. It is noteworthy that in the cases of

XSL•FO
RenderX

furosemide, phenytoin, and vancomycin, GPT-4 did not maintain a BERT score above 0.9.

SMOG grades suggested that when following similar instructions or prompts, the LLM-based pipeline was as effective as manual editing in producing a small improvement in readability, and this was supported by the expert reviewers' ratings of understandability. In contrast, the Flesch-Kincaid reading grade did not find any change in readability. This may be because the Flesch-Kincaid reading grade reflects a lower standard of comprehension than the SMOG grade [25] but reinforces the marginal nature of the readability improvement. In addition, the GPT-4 pipeline did not produce improvements in either readability metric as large as those resulting from user testing and subsequent document revision.

Comparison With Prior Work

These results are aligned with those of studies that have applied LLMs to improve the readability of patient information, which have demonstrated improvements in readability metrics [13-18] that are often associated with the omission of information [13,15,19]. This association between readability and omission of content is unsurprising, as the removal of complex concepts is likely to make a text easier to understand. However, in a safety-critical context, such as health care, the omission of essential information must be avoided, and evidence to date suggests that LLMs are currently unable to achieve this reliably. In contrast, an iterative process of user testing and subsequent document revision has a lower risk of introducing unintended content changes. In addition, while this study has demonstrated that user testing can improve the readability metrics of clinical guidelines, it has also been shown to have beneficial effects on more important user- or patient-centered outcomes, such as reducing difficulties finding and understanding information [8,9,28], and subsequently preventing guideline-related medication errors [11].

Therefore, user testing remains the guideline improvement technique with the strongest evidence base. However, the findings of this study suggest that specifically designed LLM-based pipelines with appropriate safeguards do have potential to augment (but not replace) a manual user testing process, by providing suggestions for potential improvements during the document revision process. These safeguards should be included as explicit clinician-guided, hard-coded programmatic feature checks (eg, verifying that all medications are included and in the correct order) rather than as an instruction to the LLM, which may be ignored. This approach has the potential to partially address the limited availability of knowledge and skills in information writing and design within health care organizations.

Strengths and Limitations

This study has several strengths, including the use of specific prompts based on previous research and the use of both manually revised and user-tested monographs as comparators. In addition, it evaluated both the content and readability of the guidelines using a combination of calculated metrics and human evaluation, an approach recently recommended for studies in this field [29]. However, there are also a number of limitations. First, user- and patient-centered outcomes might have been evaluated (eg, the ability of end users to find and understand information in the guidelines; the frequency of medication errors when using revised guidelines), so it is not clear whether the observed small improvements in readability would result in improved patient care. This could be addressed by future randomized studies comparing original and LLM-revised guidelines in actual or simulated patient care. Second, this study used a general LLM rather than a bespoke model trained with health care–specific information [29], which might have resulted in greater improvements in readability and/or fewer content changes. Such models should be used in future studies. Third, while the results were replicable among the 22 drug guides tested in this study, only one type of guideline relating to intravenous drug administration was used. Therefore, results may not be replicable to other types of guideline, especially as the prompting strategy was based on user testing of these specific guidelines. Other types of guidelines and more general prompting strategies should be investigated in future studies. Fourth, the inclusion of a prompt based on semantic similarity did not provide a reliable safeguard against content changes, as the BERT score for 3 guidelines was <0.9 (Table 1). Future studies should implement semantic similarity thresholds as programmatic post hoc checks (compute the BERT score externally and rerun or reject LLM outputs) and investigate more extensive use of clinician-guided hard-coded programmatic feature checks (eg, verifying that all medications are included and in the correct order). Fifth, the prompting strategy was developed by only a small team, so it is more likely to have been affected by participant subjectivity. This could be addressed by developing prompts among a larger team with a greater range of expertise. Sixth, instruction 9 in the GPT-4 prompts ("remove information about the size of the bags") was based on a misinterpretation of previous user testing–derived recommendations. Six of 12 content review panel comments describing omissions from the "Dilution" subsection related to the clinically relevant omission of infusion bag volume, suggesting this error contributed to some of the unintended omissions identified in this study.

Implications for Practice and Future Research

The present findings suggest that guideline authors should not use current LLMs to modify clinical guidelines without carefully checking the revised text for unintended omissions, additions, or changes of meaning. Similarly, clinicians should not use current LLMs to answer clinical questions without carefully checking the results against reliable sources of information. However, LLMs do have the potential to augment manual user testing and reduce the barriers to the wider use of this approach to improve the safety of the thousands of local clinical guidelines currently in use. Further work is required to embed LLM functionality into guideline authoring systems and subsequently demonstrate the safety and effectiveness of this approach.

Conclusions

LLMs are able to revise the contents of clinical guidelines to produce small improvements in readability, but with unintended omissions, additions, or changes of meaning to a minority of the information. Small improvements in readability were evident

XSL•FO
RenderX

using the SMOG grade and expert review, but not with the Flesch-Kincaid reading grade. It is not known whether these readability improvements would lead to better understanding of, or adherence to, clinical guidelines when used in practice. Omissions of information accounted for the majority of unintended changes to the information contained within the guidelines. Guideline-related medication errors continue to occur but may be prevented by the combination of human expertise, user feedback, and appropriate technologies.

Acknowledgments

The authors are grateful to members of the Injectable Medicines Guide editorial team, the pharmacist reviewers, Holly Wilson (Onaya Science), and Joseph Marvin Imperial (University of Bath) for their support of this study. Microsoft Copilot (GPT-5 model) was used during manuscript revision to edit the Introduction and Methods sections for brevity.

Funding

This project was funded by a Pump Priming Award from the Department of Life Sciences, University of Bath. Additional funding was received from the Academic Secondment Scheme of Research and Innovation Services at the University of Bath. These funders had no role in study design; collection, analysis, and interpretation of data; writing of the paper; and/or decision to submit for publication.

Data Availability

Data and code are available in a public, open access repository. Data created during this research that are not presented in full in this paper are openly available from [30].

Authors' Contributions

MDJ: conceptualization, methodology, formal analysis, investigation, writing – original draft, funding acquisition
MT: software, investigation, writing – review and editing
HTM: conceptualization, methodology, resources, writing – review and editing
MDJ is the guarantor.

Conflicts of Interest

MDJ is a member of the National Health Service Injectable Medicines Guide advisory board and held an honorary position in the editorial team at the start of this project. The other authors have no competing interests.

Multimedia Appendix 1
Large language model prompts and additional data.
[PDF File, 134 KB - jmir_v28i1e81915_app1.pdf]

References

1. Medication without harm - global patient safety challenge on medication safety. : World Health Organization; 2017 URL: https://iris.who.int/server/api/core/bitstreams/15520c4f-89d1-4a8f-9bab-6b771aa30acb/content [accessed 2026-01-29]
2. Naseralallah L, Stewart D, Azfar Ali R, Paudyal V. An umbrella review of systematic reviews on contributory factors to medication errors in health-care settings. Expert Opin Drug Saf 2022 Nov;21(11):1379-1399. [doi: 10.1080/14740338.2022.2147921] [Medline: 36408597]
3. Jones MD, Liu S, Powell F, et al. Exploring the role of guidelines in contributing to medication errors: a descriptive analysis of national patient safety incident data. Drug Saf 2024 Apr;47(4):389-400. [doi: 10.1007/s40264-024-01396-7] [Medline: 38308152]
4. Mahomedradja RF, Schinkel M, Sigaloff KCE, et al. Factors influencing in-hospital prescribing errors: a systematic review. Br J Clin Pharmacol 2023 Jun;89(6):1724-1735. [doi: 10.1111/bcp.15694] [Medline: 36805648]
5. Jones MD, Clarke J, Feather C, et al. Use of pediatric injectable medicines guidelines and associated medication administration errors: a human reliability analysis. Ann Pharmacother 2021 Nov;55(11):1333-1340. [doi: 10.1177/1060028021999647] [Medline: 33641479]
6. Lawton R, McEachan RRC, Giles SJ, Sirriyeh R, Watt IS, Wright J. Development of an evidence-based framework of factors contributing to patient safety incidents in hospital settings: a systematic review. BMJ Qual Saf 2012 May;21(5):369-380. [doi: 10.1136/bmjqs-2011-000443] [Medline: 22421911]
7. Kastner M, Bhattacharyya O, Hayden L, et al. Guideline uptake is influenced by six implementability domains for creating and communicating guidelines: a realist review. J Clin Epidemiol 2015 May;68(5):498-509. [doi: 10.1016/j.jclinepi.2014.12.013] [Medline: 25684154]

XSL•FO
RenderX

8. Raynor DK, Veene PD, Bryant D. The effectiveness of the summary of product characteristics (SmPC) and recommendations for improvement. Ther Innov Regul Sci 2014 Mar;48(2):255-265. [doi: 10.1177/2168479013501311] [Medline: 30227505]
9. Jones MD, Franklin BD, Watson MC, Raynor DK. User testing to improve retrieval and comprehension of information in guidelines to improve medicines safety. J Patient Saf 2022 Jan 1;18(1):e172-e179. [doi: 10.1097/PTS.0000000000000723] [Medline: 32569098]
10. Jones MD, Franklin BD, Raynor DK, Thom H, Watson MC, Kandiyali R. Costs and cost-effectiveness of user-testing of health professionals' guidelines to reduce the frequency of intravenous medicines administration errors by nurses in the United Kingdom: a probabilistic model based on voriconazole administration. Appl Health Econ Health Policy 2022 Jan;20(1):91-104. [doi: 10.1007/s40258-021-00675-z] [Medline: 34403128]
11. Jones MD, McGrogan A, Raynor DK, Watson MC, Franklin BD. User-testing guidelines to improve the safety of intravenous medicines administration: a randomised in situ simulation study. BMJ Qual Saf 2021 Jan;30(1):17-26. [doi: 10.1136/bmjqs-2020-010884] [Medline: 32606212]
12. Theo Raynor DK, Blackwell K, Middleton W. What do writers need to know about user testing? Med Writ 2015 Dec;24(4):215-218. [doi: 10.1179/2047480615Z.000000000326]
13. Moons P, Van Bulck L. Using ChatGPT and Google Bard to improve the readability of written patient information: a proof of concept. Eur J Cardiovasc Nurs 2024 Mar 12;23(2):122-126. [doi: 10.1093/eurjcn/zvad087] [Medline: 37603843]
14. Haver HL, Lin CT, Sirajuddin A, Yi PH, Jeudy J. Use of ChatGPT, GPT-4, and Bard to improve readability of ChatGPT's answers to common questions about lung cancer and lung cancer screening. AJR Am J Roentgenol 2023 Nov;221(5):701-704. [doi: 10.2214/AJR.23.29622] [Medline: 37341179]
15. Salam B, Kravchenko D, Mesropyan N, et al. ChatGPT makes cardiovascular MRI reports easy-to-understand: a feasibility study. J Cardiovasc Magn Reson 2024;26:100881. [doi: 10.1016/j.jocmr.2024.100881]
16. Patel EA, Fleischer L, Filip P, et al. The use of artificial intelligence to improve readability of otolaryngology patient education materials. Otolaryngol Head Neck Surg 2024 Aug;171(2):603-608. [doi: 10.1002/ohn.816] [Medline: 38751109]
17. Kianian R, Sun D, Rojas-Carabali W, Agrawal R, Tsui E. Large language models may help patients understand peer-reviewed scientific articles about ophthalmology: development and usability study. J Med Internet Res 2024 Dec 24;26:e59843. [doi: 10.2196/59843] [Medline: 39719077]
18. Swisher AR, Wu AW, Liu GC, Lee MK, Carle TR, Tang DM. Enhancing health literacy: evaluating the readability of patient handouts revised by ChatGPT's large language model. Otolaryngol Head Neck Surg 2024 Dec;171(6):1751-1757. [doi: 10.1002/ohn.927] [Medline: 39105460]
19. Stoneham AC, Walker LC, Newman MJ, Nicholls A, Avis D. Can artificial intelligence make elective hand clinic letters easier for patients to understand? J Hand Surg Eur Vol 2024 Nov;49(10):1269-1270. [doi: 10.1177/17531934241246479] [Medline: 38641940]
20. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); Jun 2-7, 2019; Minneapolis, Minnesota p. 4171-4186 URL: https://aclanthology.org/N19-1423/ [accessed 2026-01-29] [doi: 10.18653/v1/N19-1423]
21. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med 2025 Jan;31(1):60-69. [doi: 10.1038/s41591-024-03425-5] [Medline: 39779929]
22. Mishra S, Khashabi D, Baral C, Choi Y, Hajishirzi H. Reframing instructional prompts to GPTk's language. 2021 Presented at: Findings of the Association for Computational Linguistics: ACL 2022; May 22-27, 2022 URL: https://aclanthology.org/2022.findings-acl.50/ [accessed 2026-01-29] [doi: 10.18653/v1/2022.findings-acl.50]
23. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. arXiv. Preprint posted online on Apr 21, 2019. [doi: 10.48550/arXiv.1904.09675]
24. BERT score for text generation. GitHub, Inc. 2023. URL: https://github.com/Tiiiger/bert_score [accessed 2026-01-29]
25. Wang LW, Miller MJ, Schmitt MR, Wen FK. Assessing readability formula differences with written health information materials: application, results, and recommendations. Res Social Adm Pharm 2013;9(5):503-516. [doi: 10.1016/j.sapharm.2012.05.009] [Medline: 22835706]
26. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008 May;61(Pt 1):29-48. [doi: 10.1348/000711006X126600] [Medline: 18482474]
27. Hanna M, Bojar O. A fine-grained analysis of BERT score. In: Barrault L, Bojar O, Bougares F, Chatterjee R, Costa-jussa MR, Federmann C, editors. Presented at: Proceedings of the Sixth Conference on Machine Translation; Nov 10-11, 2021 p. 507-517 URL: https://aclanthology.org/2021.wmt-1.59/ [accessed 2026-01-29]
28. Verhoeven F, Steehouder MF, Hendrix RMG, Van Gemert-Pijnen JEWC. From expert-driven to user-oriented communication of infection control guidelines. Int J Hum Comput Stud 2010 Jun;68(6):328-343. [doi: 10.1016/j.ijhcs.2009.07.003]
29. Madabushi HT, Jones MD. Large language models in healthcare information research: making progress in an emerging field. BMJ Qual Saf 2025 Jan 28;34(2):73-76. [doi: 10.1136/bmjqs-2024-017896] [Medline: 39443104]
30. Jones MD, Torgbi M, Tayyar Madabushi H. Large language models to improve the understandability of clinical guidelines: an evaluation of readability improvements and unintended content changes produced by GPT-4. GitHub. 2026. URL: https://github.com/melissatorgbi/LLM-Clinical-Guideline-Understandability [accessed 2026-02-19]

XSL•FO
RenderX

Abbreviations

BERT: bidirectional encoder representations from transformers
IMG: Injectable Medicines Guide
LLM: large language model
SMOG: Simple Measure of Gobbledygook
TRIPOD-LLM: Transparent Reporting of a Multivariable Model for Individual Prognosis or Diagnosis - Large Language Model

Edited by A Coristine; submitted 05.Aug.2025; peer-reviewed by C Gissel, J Mira; revised version received 21.Jan.2026; accepted 22.Jan.2026; published 23.Feb.2026.

Please cite as:
Jones MD, Torgbi M, Tayyar Madabushi H
Improving the Understandability of Clinical Guidelines: Development and Evaluation of a GPT-4–Based Pipeline
J Med Internet Res 2026;28:e81915
URL: https://www.jmir.org/2026/1/e81915
doi:10.2196/81915

© Matthew D Jones, Melissa Torgbi, Harish Tayyar Madabushi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Interorganizational Mechanisms for Developing and Implementing Clinical Decision Support Systems in Primary Care: Exploratory, Qualitative Case Study

Jesse J M M Santema[1], MSc; Jeroen D H van Wijngaarden[2], MSc, PhD; Eric G Hiddink[3,4], MSc, PharmD; Fleur Deken[5], MSc, PhD; Maaike Kleinsmann[1,6], MSc, PhD; Hendrikus J A van Os[3,7], MD, PhD

[1]Industrial Design Engineering, Delft University of Technology, Mekelweg 5, Delft, South Holland, The Netherlands

[2]Health Services Management & Organization, Erasmus School of Health Policy and Management, Rotterdam, South Holland, The Netherlands

[3]National eHealth Living Lab, Department of Public Health & Primary Care, Leiden University Medical Center, Leiden, South Holland, The Netherlands

[4]Health Base Foundation, Houten, Utrecht, The Netherlands

[5]School of Business and Economics, Knowledge, Information and Innovation, Vrije Universiteit Amsterdam, Amsterdam, North Holland, The Netherlands

[6]Department of Cardiology, Leiden University Medical Center, Leiden, South Holland, The Netherlands

[7]Department of Psychiatry, University Medical Center Utrecht, Utrecht, Utrecht, The Netherlands

Corresponding Author:
Jesse J M M Santema, MSc
Industrial Design Engineering, Delft University of Technology, Mekelweg 5, Delft, South Holland, The Netherlands

Abstract

Background: Clinical decision support systems (CDSS) have the potential to improve patient safety and reduce costs in primary care. However, CDSS adoption remains limited due to development and implementation challenges. CDSSs are complex interventions involving multiple interacting components that require technological innovation and behavioral and organizational change. Additionally, the primary care context is considered a complex system with high care demand, fragmented structures, and many independent yet interdependent organizations. Established determinant frameworks for implementing and scaling up complex health care interventions support the identification of implementation determinants. However, they offer limited guidance on the underlying processes of these determinants, such as the implementation processes involved in complex interorganizational collaboration in primary care.

Objective: This study examined how an interorganizational collaboration in Dutch primary care (*Gezonde zorg*, *Gezonde regio* [*GzGr*]) achieved an iterative CDSS development and implementation. We aimed to identify the mechanisms that supported the collaboration in overcoming challenges.

Methods: We performed an exploratory process-level case study. Data were collected through 15 semistructured interviews. The nonadoption, abandonment, scale-up, spread, and sustainability framework was used to ensure comprehensive topic coverage during the interviews, but not as an analytical framework. We triangulated the interviews with internal and external documents and expert input. Using a thematic, inductive approach, we developed a chronological overview of the collaboration and identified mechanisms offering insights into how GzGr navigated complexity in the development and implementation of CDSS.

Results: We identified two mechanisms: (1) enacting an interorganizational value model and (2) iterative, co-creative experimentation. First, GzGr was driven by a coalition of the willing (ie, individuals willing to take an extra step), with shared goals that prioritized collective benefit while respecting organizational values. They established shared principles that translated the broad GzGr mission into concrete CDSS development choices, while also guiding strategic expansion by involving mission-aligned, innovative organizations. Second, after initial prototypes, GzGr established an iterative learning and improvement experimentation for both the technology and the collaboration. This process allowed for rapid feedback, validation of added value, and ongoing refinement. Additionally, this experimentation approached the development and implementation phase as a continuous process involving multistakeholders, supporting both the technology and the collaboration.

Conclusions: This study identified 2 mechanisms that sustained interorganizational collaboration and CDSS development. These mechanisms connected collaborative and technical changes across people, technology, and organizational levels, enabling technological viability, stakeholder value, and multilevel support. The mechanisms operated both within and between organizations through iterative cycles of development and implementation. Practical implications include involving multilevel, innovative, and influential stakeholders; maintaining alignment through an orchestrating actor; and adopting an iterative approach between development and implementation. Our findings extend existing determinant frameworks by offering process-level insights into

XSL•FO
RenderX

how such mechanisms help overcome challenges in the development and implementation of CDSS within interorganizational collaborations.

(J Med Internet Res 2026;28:e83084) doi:10.2196/83084

KEYWORDS

clinical decision support systems; interorganizational collaboration; mechanisms of change; multilevel stakeholder involvement; integration of people; integration of technology; integration of business; iterative development and implementation; process-level analysis

Introduction

Clinical decision support systems (CDSS) hold the potential to improve patient safety, diagnostic accuracy, and clinical decision-making and reduce health care costs [1]. CDSSs achieve these aims by using rule-based or artificial intelligence (AI) algorithms to deliver patient-specific alerts, reminders, and dosage support, especially in primary care [2-4]. Despite its potential, CDSS adoption remains limited due to interdependent development and implementation challenges [1,5-7], particularly in primary care [2,8,9]. These challenges partially stem from the complexity of CDSS, as they involve multiple interacting components, often necessitate behavioral changes from end users, and affect clinical and economic outcomes [10,11]. The complexity of CDSS also results in interconnections between users and their social and organizational context.

Traditional implementation approaches assess interventions under controlled conditions before applying them in practice [10,12-15]. This approach can lead to unintended consequences during real-world implementation due to the intervention's complexity [14]. Therefore, digital health and innovation scholars advocate for real-world implementation with iterative cycles consisting of development and implementation that proceed [10,12-14,16-18]. These iterative cycles allow contextual adaptation and gradual refinement, which improves implementation outcomes and mitigates unintended consequences [13,14]. Examples of models for an iterative process are the Centre for eHealth and Wellbeing Research roadmap, UK Medical Research Council guidelines, and the nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) framework [10,12,13,18]. These models provide valuable insights into determinants for the implementation of digital health interventions in single organizations (while recognizing the wider contextual factors influencing implementation). However, these models pay less attention to the interactions among determinants and processes underlying these determinants [14,19].

Insights into these underlying processes are especially relevant in the context of CDSS in primary care, where both the intervention and the setting are inherently complex. Primary care in the Netherlands involves many independent yet interdependent organizations, such as general practices, care groups, payers, and technology developers [2,14,20,21]. These stakeholders differ in roles, interests, and values owing to separate financing structures, although they must collaborate across organizational boundaries to ensure continuity and accessibility of care. This results in a highly fragmented and complex care landscape. Stakeholders and their interactions are illustrated in Figure 1. Research suggests that a context with multiple independent and interdependent organizations calls for interorganizational collaboration to enable iterative development and implementation [11,16]. Such collaboration aims for coordinated efforts between organizations working toward a shared goal [22,23]. However, these studies do not elaborate on how an interorganizational collaboration can be established in practice.

XSL•FO
RenderX

Figure 1. Simplified overview of 6 independent organizations in Dutch primary care and their connections. The connecting lines highlight complexity. Every organization must interact with every other organization for the continuity and accessibility of care. GP: general practitioner.

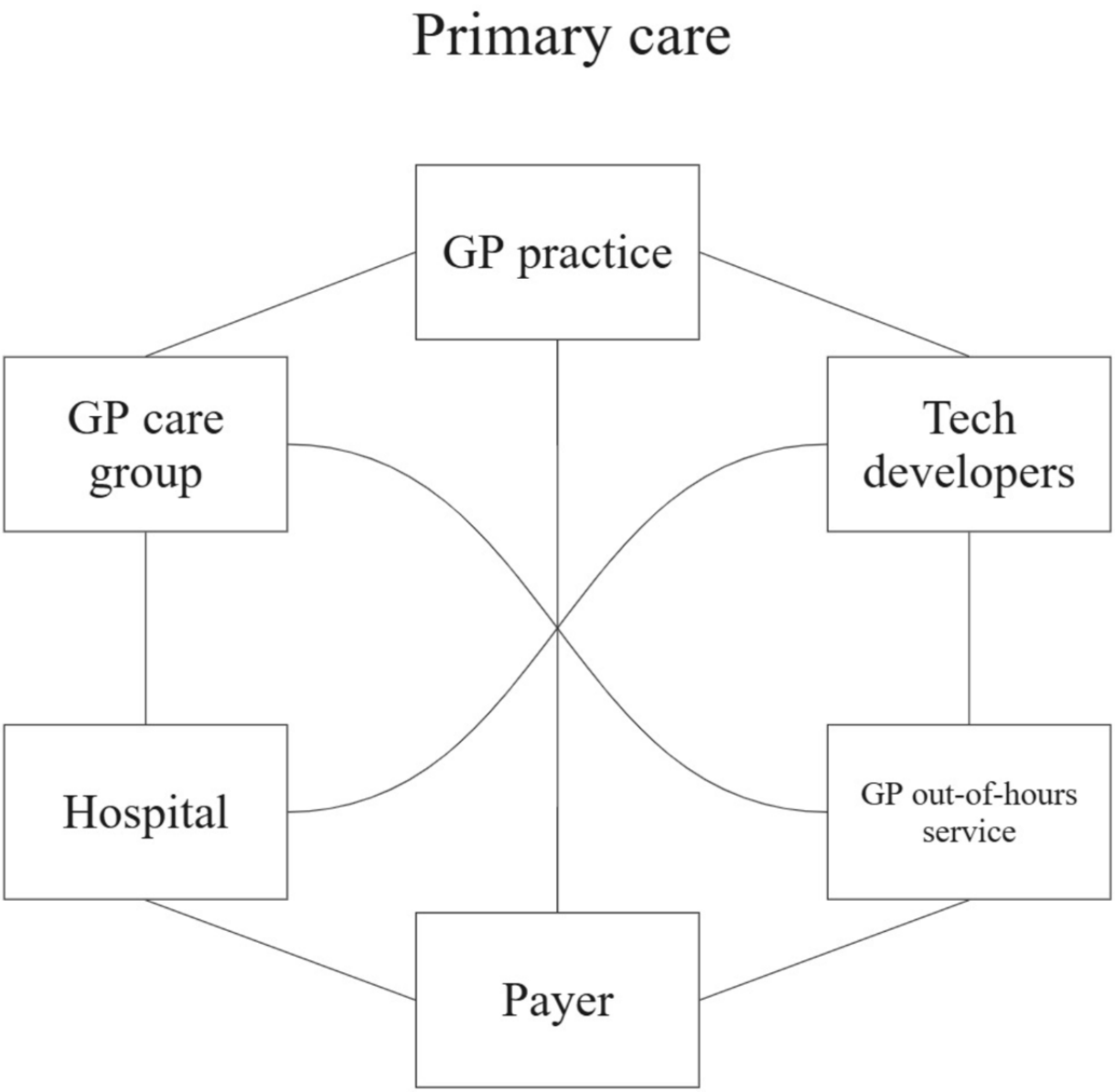

We built on prior determinant research conducted in single-organization implementation contexts [10,12,13,18], which emphasized the need for an iterative CDSS development and implementation process. Our study extended this line of research by examining how challenges encountered during development and implementation were addressed in an interorganizational primary care collaboration. Specifically, we focused on identifying mechanisms that explain the underlying processes through which interventions are developed and implemented [24]. This focus is particularly relevant in complex settings such as primary care. Our research question was as follows: What mechanisms facilitate overcoming challenges in the development and implementation of CDSS within an interorganizational collaboration in primary care? We aimed to identify mechanisms that enabled CDSS development and implementation within an interorganizational collaboration in primary care. We conducted a qualitative process analysis of a Dutch primary care collaboration, allowing us to focus on how challenges and actions developed over time and formed mechanisms.

Methods

Study Design

We performed an exploratory case study to identify which mechanisms enabled *Gezonde zorg, Gezonde regio* (GzGr) to successfully develop and implement CDSS in Dutch primary care [25]. We define mechanisms in the context of this paper as sequential patterns of actions that explain how change occurs, both between and within organizations in relation to technological interventions. These actions represent how GzGr created the conditions and behaviors needed to cope with challenges that occurred over time in the development and implementation of complex digital health solutions in complex

XSL•FO
RenderX

settings. The sequential aspect emphasizes how actions build on previous actions or challenges over time, indicating that their order may be causally related to their final effect. To identify how, why, and when mechanisms were applied, we used qualitative process analysis by tracing actions and decisions over time [26-28]. While the mechanisms captured what drove change, our accompanying narrative illustrated how these processes unfolded in practice. This study was designed and reported in line with the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist (Checklist 1) [29].

Case Study Selection

We researched "Healthy care, Healthy region" (in Dutch: *Gezonde zorg, Gezonde regio* [*GzGr*]). We identified GzGr as a unique case study via field exploration through researchers' networks and via internet searches. Case selection was based on the following criteria: (1) successful regional implementation of a CDSS across multiple primary care practices, (2) full integration in primary care workflow, including IT integration; and (3) collaboration involving all relevant stakeholders in primary care.

Case Description

The GzGr initiative involves a diverse range of organizations: general practitioner (GP) practices, regional primary care organizations, a regional hospital, a health care insurance firm, data processing firms, technology and content development firms (collectively referred to as technology partners), and advisory groups responsible for implementation support and technology evaluation [30]. Figure 2 provides a visual representation of the GzGr structure. GzGr developed multiple rule-based and data-driven CDSS to provide real-time alerts to GPs regarding patient health risks [31]. The steering committee, consisting of executives from key partner organizations, sets the strategic direction and oversees key decisions. The working group, consisting of health care purchasers and innovation managers from the care groups, manages day-to-day operations and the collaboration and advises the steering committee. The editorial committee consists of end users, technology developers, and representatives from the health care insurance firm and is responsible for identifying widely experienced problems and overseeing CDSS development.

Figure 2. Visual representation of the GzGr structure, with health care and supporting organizations (represented by the horizontal squares). The vertical squares represent the governance structure. GzGr: *Gezonde zorg, Gezonde regio*.

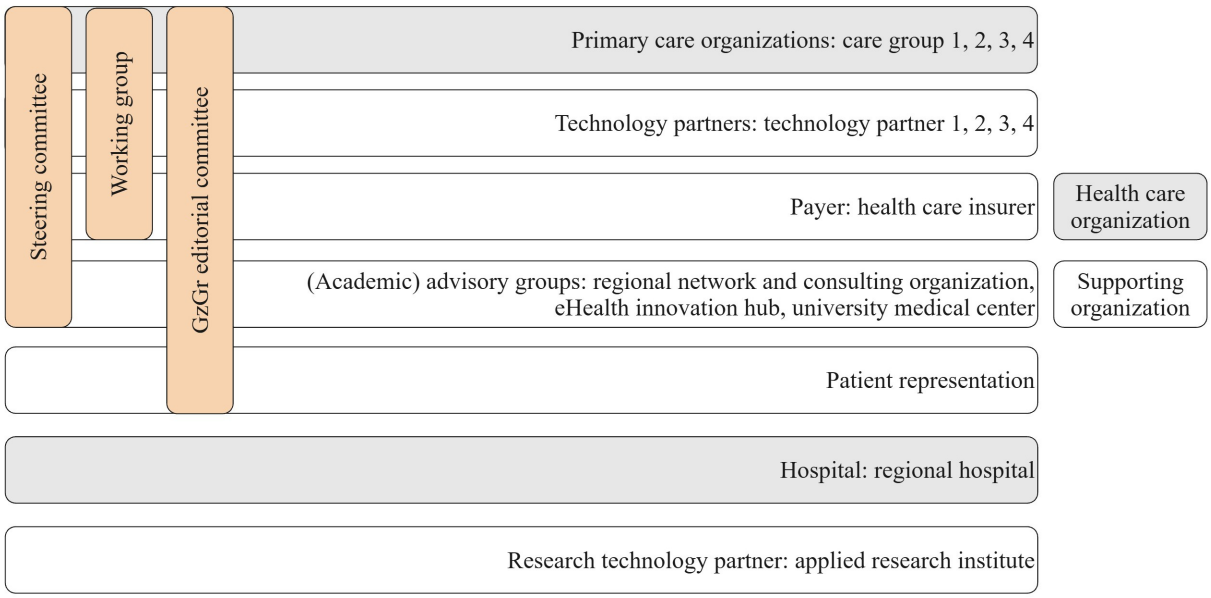

Data Collection

We collected multiple sources of data, including interviews and documents, aiming for data triangulation to increase the validity of the data [32]. In total, we recruited and interviewed 15 unique respondents from December 2021 to September 2022 by JDHvW. Interviews were conducted until data saturation was reached, meaning that interviews confirmed findings rather than yielded new findings [33]. Data saturation determined our study size. We interviewed at least one participant from each partner organization involved in GzGr. An overview of participants is given in Multimedia Appendix 1. Our eligibility criteria for participants were their (1) active involvement in the operational development of the CDSS, by working at one of the technology partners, at a general practice or a care group (defined as an organization that coordinates and supports GPs with administrative burden); and/or (2) involvement in the strategic decision-making of GzGr, either through participation in the steering committee or by holding an executive mandate within an organization.

We sampled the respondents using snowball sampling and purposive sampling and used the networks of participants, partner organizations of GzGr, and researchers [34]. JDHvW contacted participants by email through a GzGr contact person, after which additional participants were identified by previously interviewed respondents. Although this approach introduced potential selection bias at first, snowball sampling helped recruit a diverse and representative group. No participants declined participation.

XSL•FO
RenderX

We followed a semistructured interview approach to examine which mechanisms were used by GzGr to ensure progress in development and implementation. We composed an interview topic guide based on the NASSS framework [12]. Topics were based on the following domains: intervention features, the expected value, adopter readiness, the broader context, adaptability over time, and the interplay among domains. We omitted the "condition" domain, as the developed CDSS were designed to apply across multiple clinical conditions. Furthermore, we asked about involved stakeholders, organizational characteristics, and the development process. We explored how GzGr addressed related challenges in these domains to identify underlying mechanisms, which formed the basis for identifying mechanisms.

Interviews were conducted via video call. Each interview lasted between 45 and 60 minutes, and the respondent(s) and JDHvW were present. The interviews were audio recorded and afterward transcribed verbatim. Transcripts were made available to respondents for review upon request. Moreover, we collected internal and external documents of the organizations via the participants and the internet, from September 2023 to January 2024. All documents were stored in Atlas.ti (ATLAS.ti Scientific Software Development GmbH) for analysis.

Data Analysis

To identify mechanisms that facilitated overcoming challenges in the development and implementation of CDSS, our data analysis consisted of multiple steps, namely (1) composing a chronological overview; (2) thematic and inductive coding to identify mechanisms; (3) performing a member check; and (4) constructing a conceptualized narrative. We used Atlas.ti 9 to support the coding process. Further elaboration on the coding process and the codebook is given in Multimedia Appendix 2.

1. We created a chronological overview of challenges and actions related to development, implementation, or collaboration encountered or undertaken by GzGr to provide deeper insight into how the stakeholders developed and implemented CDSS. The chronological overview is given in Multimedia Appendix 3.
2. Next, we conducted an in-depth analysis to identify mechanisms, consisting of patterns of actions taken by GzGr. We used thematic analysis and inductive coding to reveal new insights and underlying patterns in the data [35,36]. We performed a second round of open coding of the transcripts. During open coding, codes that had been generated in the earlier transcripts were applicable to later interviews. At that stage, no new codes were introduced, which indicated that we reached code saturation. These codes were used for composing a codebook (see Multimedia Appendix 2) and for coding all the transcripts. Then, we identified 2 mechanisms, by linking challenges and determining the sequential patterns of actions that, according to participants, had led to the successful development and implementation of CDSS. Mechanism 1, enacting an interorganizational value model, is defined as aligning independent yet interdependent organizational goals and values into regional goals and values and translating them into practical actions. The mechanism 2, iterative, co-creative experimentation, is defined as continuously, jointly developing and adapting solutions with stakeholders, treating development and implementation as one interconnected, ongoing process. In Multimedia Appendix 3, we further elaborated on the link between challenges and actions and how the mechanisms emerged or evolved over time.
3. We performed member checks with the steering committee of GzGr to reach consensus on the mechanisms deployed. Additionally, EGH and HJAvO, who participated in the GzGr collaboration as (scientific) experts, supported the interpretation of qualitative data.
4. The final step in the data analysis was constructing a narrative of the challenges encountered and the actions taken by GzGr, supported by quotations of the participants. At the end of every period, we conceptualized the sequential patterns of actions that formed a mechanism. Furthermore, we bundled the sequential patterns of actions per mechanism in Table 1. In the discussion, we compared these findings with prior work.

Table . Mechanisms consisting of sequential patterns of actions to develop and implement complex digital health solutions in complex settings.

Mechanism	Sequential processes
Enacting an interorganizational value model	• Forming a coalition of the willing • Defining and aligning on principles for organization selection and development early in the process (ownership, pricing, co-creation, transparency, guidelines, embedding in workflows) • Streamline decision-making and maintain focus during development due to aligned principles
Iterative, co-creative experimentation	• Co-creation in structured feedback loops between end users, developers, and decision-makers • Taking ownership of communication by the collaboration • Technological adjustments combined with measurable scientific outcomes

Ethical Considerations

This study was conducted in accordance with ethical guidelines for qualitative research involving human participants. Ethical approval was obtained from the Research Ethics Review Committee of the Erasmus School of Health Policy & Management (reference number 21 - 014), ensuring compliance

XSL•FO
RenderX

with relevant regulations on data protection and participant rights.

Additionally, participants signed an informed consent form before participating in interviews, which stated the goal of the research, the researchers, and the interview; privacy measures (including pseudonymizing data); and confidentiality measures (raw data available upon reasonable request) and included consent for audio recording. No identifying information or images of individual participants are included in this manuscript or its supplementary materials. Participants did not receive any compensation.

Results

Overview

In this section, we described the chronological sequence of the development and implementation challenges and actions. Table 2 summarizes the main challenges faced by GzGr and the resulting actions that led to the development and implementation of CDSS. In retrospect, we distinguished 3 periods, each with distinct challenges and focus. The chronological overview in Multimedia Appendix 3 is a more elaborated version of these challenges, actions, and their connection to the mechanism and provides a brief overview of the timeline.

XSL•FO
RenderX

Table . Summary overview of challenges and actions across 3 periods that contributed to enacting an interorganizational value model and to applying iterative, co-creative experimentation.

Period, challenge, and actions	Mechanism leveraged
2010 - 2014: forming a coalition of the willing	
Staff shortages within GzGr[a]	
Started collaboration with influential individuals	1
Exploring technological needs, unexpected outcomes, and collaboration stopped	
Research institute measured the baseline population funding	1
Engaged, innovative alignment is difficult	
Composing coalition of the willing and aligning principles for collaboration expansion	1
Engaged, innovative alignment is difficult	
Technology partner 1 and regional network and consulting organization joined the collaboration	1
Mission changed: enhancing preventive care by developing CDSS[b]	1
Technology partners 2 and 3 joined the organization based on aligned principles	1
First pilot CDSS development	1
Not all organizations were innovative	
Shift in focus to involving innovative GP[c] practices and organizations	1
2015 - 2017: develop understandable CDSS	
Engaged, innovative alignment is difficult	
Establishing CDSS development principles	1
Building understandable CDSS	1
Lack of workflow integration, PDF notifications for CDSS	
Technology partner 4 joined, ensuring integration of CDSS in GP information system	1
Six knowledge rules deployed, five in development	1, 2
2018 - 2021: Implement CDSS in daily practice	
Use of CDSS low	
More focus on implementation	2
Communication responsibility of individual care groups	
GzGr improved information material, developed training, and informed community	2
Focus shift to implementation	
Established editorial committee for identifying implementation problems	2
Focus shift to implementation	
Implementation staff assigned, improving the digital skills of health care professionals	2
Improve CDSS	
Testing CDSS prototypes and retrieving feedback from GPs	1, 2
Developed dashboard with insights into the use of CDSS	2
Strategic organizational dialogues about opportunities for individual organizations and GzGr	1, 2

[a]GzGr: *Gezonde zorg, Gezonde regio* (Healthy care, Healthy region).

[b]CDSS: clinical decision support system.

XSL•FO
RenderX

[c]GP: general practitioner.

Period 1 (2010-2014): Forming a Coalition of the Willing

In this first period, we explain how GzGr started their collaboration with a mission, formed a mission-driven coalition of the willing, and established principles for expanding the collaboration.

The GzGr collaboration was initiated by 3 senior influential individuals in health care. One was a senior executive at the largest regional health care insurer, securing long-term funding. The other two were executives from a primary care group and a regional hospital, providing managerial commitment. These individuals were considered influential in the region because they could allocate resources, influence organizational decisions, and mobilize support (financial and people). The aim was to provide more efficient care to maintain regional health care accessibility, with a focus on prevention, the use of routine data, and in line with shared savings funding models. In their view, delivering future-proof care required close collaboration between care organizations and funders to foster joint responsibility for a specific population. As an initial attempt toward operationalizing this mission, a research-focused technology partner explored the feasibility of population-based funding. However, it became evident that this organization lacked expertise in operationalizing the outcomes, posing a challenge for GzGr, leading GzGr to stop the collaboration.

Realizing the long-term mission to aim for more efficient care and involving partnering organizations was also experienced as challenging for GzGr; therefore, they decided to refocus their efforts on a more short-term, actionable mission: enhancing preventive care interventions in GP practices by developing and implementing CDSS. This system would leverage routine care data to support GPs in patient monitoring and the early detection of health risks.

To realize the development, GzGr invited 2 new individuals from organizations to join the collaboration: an innovation manager working at technology partner 1, contributing to clinical content and CDSS development, and a project manager from the regional network and consulting organization. According to respondents, the addition of these 2 organizations was key in translating GzGr's broad mission into an actionable, concrete CDSS. Together, these 5 individuals brought extensive professional networks, complementary expertise, and executive backing from their organizations:

> *Our success lies in the combination of having people who deeply understand the data and work with it daily, others who know the primary process well, a few executives who are genuinely enthusiastic about this, and a healthcare insurer that recognizes its importance.* [Professor of data science, technology partner 1]

More importantly, respondents emphasized that these individuals consistently prioritized the shared, collaborative mission. At the same time, they ensured alignment with their own organizational interests, without aiming to maximize individual organizational gain. GzGr referred to the individuals as a "coalition of the willing," marking the emergence of a mission-driven collaboration within GzGr:

> *Based on knowing a few enthusiastic people, being aware of the expertise available, having a shared goal, and the motivation to pursue it: we just went ahead and did it. We developed a proof of concept and demonstrated that it works. Even stakeholders said: This will accelerate things.* [Policy adviser, national health care insurance umbrella organization]

To make progress, GzGr recognized the need for a data infrastructure to support CDSS implementation across GP practices and enable data integration and evaluation. This meant that additional technology partners were needed. Identifying technology partners that aligned with GzGr's mission and collaborative approach proved challenging. Therefore, GzGr formed principles for organization selection and task allocation: willingness to experiment, previous experience in CDSS development, knowledge of legal and regulatory standards, possession of appropriate certifications, a willingness to work at a cost-plus model (ie, a pricing approach where providers are reimbursed based on actual costs plus a fixed margin), and agreeing upon the shared mission. These principles ensured that new participating organizations aligned with GzGr's chosen direction, allowing flexibility in how the mission was executed but preserving its overall purpose, which prevented mission drift and delay in progress.

Through technology partner 1's network, technology partners 2 and 3 joined. Together, GzGr began developing CDSS that could be implemented at scale. One of the first successful pilots was a data-driven CDSS for cardiovascular pharmacotherapy, identifying cost-effective treatment options with equal or better outcomes. Savings were shared between providers and insurers, illustrating an early shared savings model. Figure 3 provides a visual overview of the contributing organizations (squares), their roles, interaction (arrows), and an example of CDSS (circle). Data from multiple national registries are integrated by technology partners (double-headed arrow) and linked to GP electronic health records and the GP (single-headed arrow), where CDSS are embedded to support clinical decisions. Finally, CDSS are evaluated with insurers and academic partners. The figure also includes an example of a cardiovascular risk management rule, showing how data generated by GPs trigger actions for patient follow-up, supported by continuous evaluation and feedback between organizations.

XSL•FO
RenderX

Figure 3. Organizations involved in developing and implementing CDSS within GzGr, showing data integration, embedding in GP records, and continuous evaluation. The squares represent participating organizations, the circles represent an example of CDSS, and the arrows represent the data infrastructure. CDSS: clinical decision support system; CVRM: cardiovascular risk management; EHR: electronic health record; GP: general practitioner.

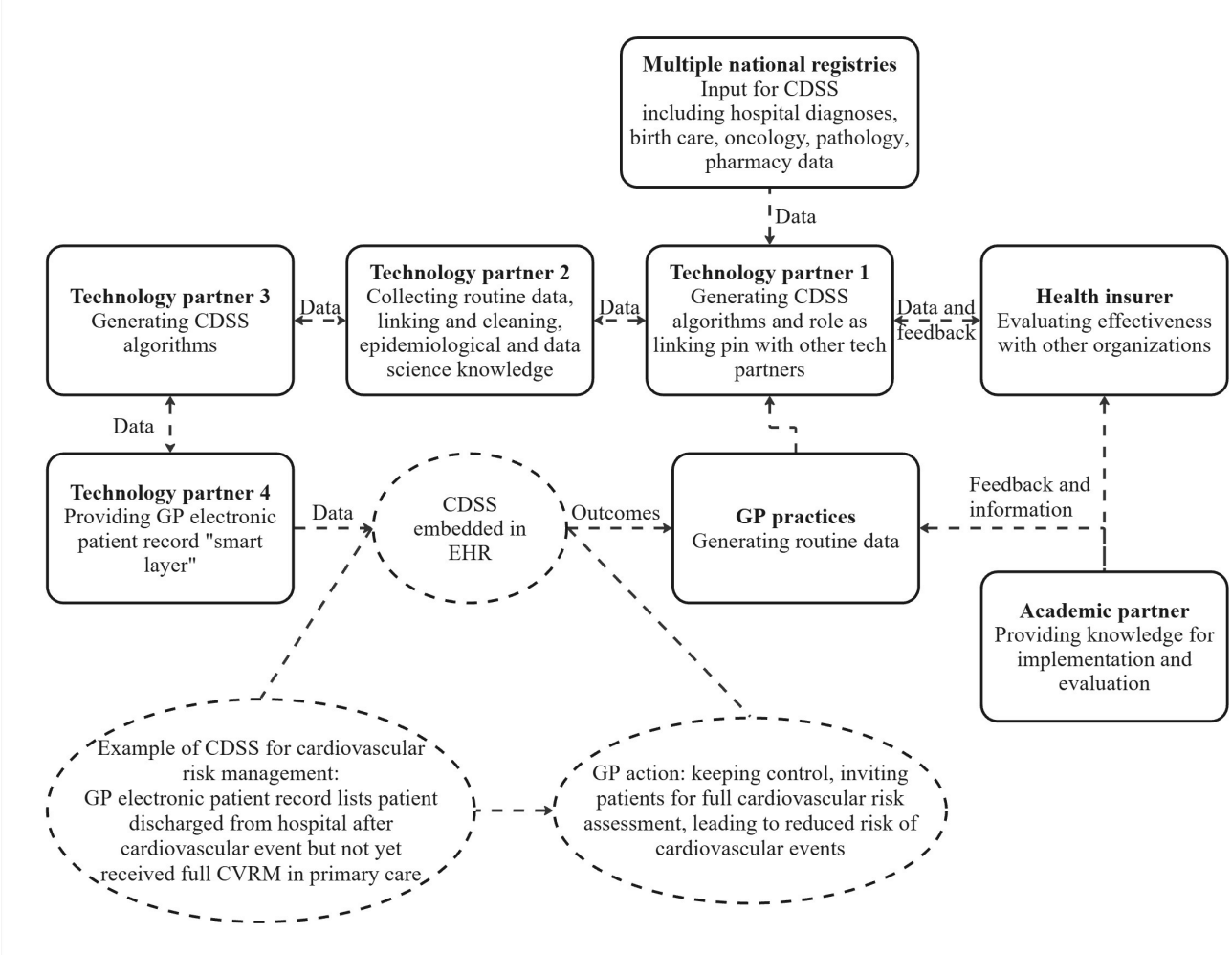

From the outset, CDSS and infrastructure development were co-funded by health care insurers and technology partners, who believed in the potential of CDSS and reimbursed GPs for using CDSS, even in the absence of traditional scientific validation. Gaining internal support within the insurer required ongoing efforts from the individual involved in GzGr from the start, who actively used the shared mission to convince colleagues of the value of GzGr:

> *We have invited a colleague working at the healthcare insurer, because internal support [for GzGr at the insurer] was needed to keep justifying, once again, why funding should continue.* [Strategic adviser, regional network, and consulting organization]

To increase the impact of the CDSS, GzGr sought to expand the collaboration with additional GP care groups in the region. However, not all care groups were willing to experiment and change their internal processes to accommodate the CDSS. Therefore, GzGr decided to prioritize GP practices with an "innovative mindset" and willingness to experiment. Additionally, not all collaborations within GzGr endured. A regional hospital withdrew after the departure of the executive initiating GzGr, which respondents linked to a misalignment between individual commitment and organizational priorities.

To conclude, GzGr operationalized their broad mission to enhance preventative care by developing and implementing CDSS. The first mechanism was used: enacting an interorganizational value model, which built on shared value and accountability. We identified two sequential patterns of actions, forming this mechanism: (1) forming a coalition of the willing that shared a common mission, combined their expertise, and mobilized resources; and (2) defining principles for organization selection to ensure mutual value, emphasizing innovation, compliance, cost-plus funding model, and mission alignment. Table 1 presents the sequential patterns of actions that form the mechanism.

Period 2 (2015-2017): Develop Understandable Algorithms

In the second period, GzGr further developed the interorganizational value model by aligning on development principles, preventing disagreements during or after the CDSS development. Management, together with end users, developed these principles to ensure broad implementation, feasibility, and usability in multiple levels within the organizations. The principles for developing CDSS were as follows.

First, agreement was needed on the pricing structure and ownership of CDSS. For GPs and the insurer, it was not

XSL•FO
RenderX

acceptable to invest their time and share sensitive data, while the technology partners would make high profits, own CDSS, and if prices would fluctuate. At the same time, the technology partners needed to work for profit to survive. Eventually, it was agreed that the technology partners were allowed to make a small profit based on a cost-plus price model and that CDSS would be free of property rights:

> *They [technology partners] need to make profit otherwise the company will not survive. But we discussed we do not want value-based pricing [assessing the price based on the costs the technology saves], but a cost-plus model...so we can set a fixed price per patient and make it more attractive for the healthcare insurers because the prices are predictable.* [Innovation manager, technology partner 3]

Second, GzGr fundamentally chose co-creation as the basis for CDSS development, ensuring that CDSS were driven by clinical problems that were prioritized by both GPs, practice nurses, and patients. While clinicians identified problems in practice related to delivering care, technology developers explored what was technically feasible, and funders assessed financial viability. This ensured the integration of the requirements from people, technology, and business. Initially, this human-centered, problem-driven approach challenged technology partners, who were used to a supplier-client model with predefined requests. Now, they had to share responsibility with end users in identifying clinical problems, requiring them to adapt their development processes:

> *We apply a problem-driven principle, and because we develop together, you can feel a real drive to actually use it [CDSS] and make the most of it.* [Director primary care, technology partner 4]

Third, the GzGr organizations agreed that transparency of CDSS and its development was key. CDSS' decision paths (even for AI, using logistic regression) could be followed by GPs, and the functionality was made explainable, allowing GPs to verify accuracy and build confidence in the use of CDSS. This transparency supports GPs in meeting legal requirements that state that health care professionals must be able to explain clinical decisions.

Fourth, aligning with legal, regulatory, and clinical guidelines was essential for GzGr. This meant that GPs retained control over clinical decision-making, with the ability to override CDSS recommendations. Although aligning with guidelines was costly, it ensured quality, scalability, and support from GPs:

> *I am pleased with...legal and regulatory guidelines, what might sound paradoxical, but it forces you to demonstrate how your algorithm works,...leading to transparency.* [Innovation manager, technology partner 3]

Fifth, it was agreed that seamless workflow integration was essential by embedding CDSS in GP electronic health records to avoid system switching. The coalition of the willing convinced the largest GP information system developer (technology partner 4) to join, which made it possible that CDSSs could be integrated within their systems. Respondents noted that achieving technical integration was a significant but successfully managed challenge, as they had experienced how a nonintegrated system was rarely used:

> *The initial PDFs were barely used, people had to open a PDF, log into a new system again. This just did not work in practice.* [Account manager, health care insurer]

While following these 5 principles, 7 CDSS were developed in the GP landscape and were used by 160 GP practices, reaching up to 550,000 patients. An additional 5 CDSS were under development and were tested in 1 to 5 practices, with a reach of 10,000 to 25,000 patients (in 2022; Textbox 1). The development of CDSS started in 2014 and is ongoing.

XSL•FO
RenderX

Textbox 1. Overview of the 7 developed and 5 clinical decision support systems (CDSS) under development in 2022, based on the aligned development principles, including adoption metrics (number of general practitioner [GP] practices and patients).

Developed by *Gezonde zorg, Gezonde regio*

- CDSS applications
- Patients at risk of being unserved in cardiovascular risk management.
- Referral protocol for patients with chest pain.
- Feedback on chest pain referral practices.
- Referral protocol for peripheral arterial disease (PAD).
- Peripheral arterial disease: case identification and follow-up.
- Feedback on referral behavior in PAD cases.
- Assessment of indication for goals-of-care conversations in frail older adults.
- Adoption metrics
- Developed CDSS were implemented in 160 GP practices, reaching approximately 550,000 patients.

Development by GzGr

- CDSS applications
- Patients with chronic obstructive pulmonary disease potentially overtreated with inhaled corticosteroids.
- Patients on direct oral anticoagulants/non–vitamin K antagonist oral anticoagulants with chronic kidney disease (chronic kidney disease epidemiology collaboration) for over 12 months.
- Ensuring appropriate laboratory monitoring frequency for patients on central nervous system medications.
- Treatment adjustments based on a significant decline in kidney function.
- Identification of lithium users at high risk of toxicity.
- Adoption metrics
- CDSS under development were implemented in 1 to 5 practices, reaching between 10,000 and 25,000 patients.

In conclusion, the organizations concretized the interorganizational value model by aligning early on developing principles (the sequential process): (1) agreements on ownership and pricing, (2) co-creation, (3) transparency and guideline alignment, and (4) embedding CDSS in daily workflows. These principles translated the coalition's mission into daily practice, helping organizations navigate conflicts, build trust, and coordinate decisions while preserving their individual and regional value. Additionally, the mechanism of iterative, co-creative experimentation emerged with the first prototypes, driven by structured feedback loops between end users, developers, and decision-makers (the sequential process). The sequential patterns of actions that form these mechanisms are presented in Table 1.

Period 3 (2018-2021): Implement CDSS in Daily Practice

After the first CDSS was developed, it became clear that the use of CDSS by GPs was lower than initially expected. In response, the focus shifted to integrating CDSS into daily practice and optimizing use, learning from previous experience, and iteratively adapting the strategy, with the same mission.

Initially, during implementation, primary care groups managed communication within their own organizations, but this proved challenging and ineffective. As a result, GzGr started developing materials, webinars, and introduced on-site implementation staff to strengthen end user skills and integration into workflows. Identifying relevant problems for CDSS also remained difficult, leading to the formation of an editorial committee, comprising technology partners, the insurer, and innovative GPs and assistants. They identified widely experienced implementation problems, composed solutions and an implementation plan, and promoted CDSS use among peers.

As the health care insurer and technology partners had substantially invested in the development and implementation, they wanted a guarantee that GPs used CDSS. Therefore, implementation outcomes (eg, CDSS usage rate and adoption of regional GP practices) were developed and agreed upon, obliging the GzGr organizations to act when algorithm usage rates dropped below a certain threshold:

> *It's no longer optional to use the algorithms [CDSS]. We have invested significantly, and the strategic goals must be met. While some front runners are making real progress, others still hesitate. That is why we are now fully focusing on implementation,...healthcare organizations must take responsibility for internal adoption.* [Account manager, health care insurer]

XSL•FO
RenderX

Additionally, in this period, practice-oriented academic partners joined GzGr to facilitate research, testing, and evaluation, without losing momentum and to ensure a pragmatic process. The results demonstrated that CDSS contributed to time efficiency, improved care quality, and increased adoption by bringing a scientific perspective. The steering committee assessed implementation outcomes across people, technology, and business to ensure stakeholder value. Values were not only monetary; namely, for a GP, saving time was essential, and the health care insurer had to fulfill their duty of care.

GzGr used the implementation and scientific outcomes and user feedback for a continuous iterative learning and improvement experimentation process to refine both the CDSS and the collaboration. The process was as follows: every first version of CDSS is tested in 1 or 2 GP practices (serving between 2000 and 5000 patients). The first tests were done in 5 to 10 GP practices (serving between 10,000 and 25,000 patients). When a CDSS was approved by GzGr, it was implemented in all practices (serving more than 500,000 patients). This process led to adaptations in both the technology and the mission-driven collaboration. Early feedback enabled timely adjustments, strengthened collaboration, and increased support among end users. Respondents noted that this iterative process became a functioning "engine for innovation" within and between the development and implementation phase: creating the opportunity and confidence to develop other technologies and expand the collaboration with a hospital and more care groups.

To conclude, the mechanism "iterative co-creative experimentation" further developed as GzGr shifted from developing technology to embedding CDSS in daily practice. The sequential patterns of actions were (1) co-creation in structured feedback loops, ensuring that people, technology, and organizational perspectives directly informed further development; (2) GzGr took ownership of communication; and (3) technological adjustments were tied to measurable, scientific outcomes. Together, these sequential patterns of actions strengthened shared commitment to GzGr's mission, while the interorganizational value model streamlined decision-making and maintained focus during iterative development and implementation cycles. Table 1 presents the sequential patterns of actions that shape the mechanisms.

Discussion

Principal Study Findings

We conducted a qualitative process analysis of our exploratory case study data to identify which mechanisms helped overcome challenges in the development and implementation of CDSS within GzGr, an interorganizational collaboration. This analysis shed light on sequential patterns of actions by identifying the mechanisms through which challenges were addressed across 3 subsequent periods. We identified 2 mechanisms: (1) enacting an interorganizational value model and (2) iterative, co-creative experimentation. The 3 periods show that the mechanisms reinforced each other and how the underlying actions built on each other over time. Iterative, co-creative experimentation refined the interorganizational value model. In turn, this shared value model made iterative experimentation more focused and effective.

Comparison With Prior Work

We elaborate on how these mechanisms that explain CDSS can be developed and implemented in an interorganizational context and compare them to previous research.

Enacting an Interorganizational Value Model for Mission-Driven Interorganizational Collaboration and CDSS Development

One of the key perceived challenges by GzGr in developing and implementing CDSS across multiple organizations was the fragmentation of responsibilities, interests, and values. This reflects the complexity of the setting, characterized by many independent yet interdependent organizations. GzGr addressed the setting's complexity by forming an interorganizational value model. Underpinning this value model was a mission-driven coalition of the willing. This coalition enabled GzGr to move beyond organizational silos and foster collaboration by spreading the shared mission and value across individual organizations. Earlier research shows that collaborative performance tends to be strongest when a closely connected group with clear leadership and a shared purpose works together, supported by wider networks that enable collaboration across boundaries [37]. Additionally, GzGr showed that early alignment on shared principles is key to addressing implementation challenges. Agreements about ownership and pricing, transparency, and traceability of CDSS, and embedding them into daily workflow are earlier mentioned in single-organizational literature as facilitating the development and implementation of digital health interventions [16,20]. However, existing literature does not explain how shared principles can be operationalized early into an interorganizational value model, which GzGr's sequential patterns of actions address.

Embedding people, technology, and organizational interests within this value model enabled GzGr to balance diverse values, address challenges in the primary care process, ensure technical feasibility, and maintain financial viability. Specifically, different organizations and different levels within organizations operate with their own structures, priorities, and financial logic, leading to differences in the perceived value of the collaboration and outcomes [38]. Taking all organizational values into account will help identify priorities, problems, and solutions that can improve real-world implementation of contextually adapted technologies [13,18].

In addition, GzGr continuously assessed value, monetary, or otherwise (such as improved knowledge, diagnostics, or patient care) to enact the interorganizational value model. An earlier study showed that moving beyond monetary value and keeping people, technology, and business in mind can lead to alternative, sustainable payment models [39]. Literature also highlights the importance of integrating people and technology, such as aligning workflows to improve eHealth usability [1,40], but often overlooks the business dimension. While business model design based on desirability, feasibility, and viability of interventions is seen as key to success [13,18,41], financial

XSL•FO
RenderX

interdependencies between organizations are seen as a problem for business modeling [12].

The contribution of our study lies in identifying the sequential patterns of actions for enacting an interorganizational value model. This model integrates people, technology, and business perspectives to align principles for CDSS development and guide coordinated actions across independent organizations.

Iterative, Co-Creative Experimentation That Spans Development and Implementation

CDSS are considered complex interventions due to the multiple interacting components and the connection between end users and their context. Developing and implementing a complex intervention in a complex system requires an iterative process of development and implementation [10]. In this research, we found that the boundary between development and implementation was blurred and that iterative experimentation was an ongoing process of the sequential patterns of actions. This led to GzGr building and sustaining the involvement of multiple organizations, instead of only one-off interactions between developers and users.

For GzGr, the learning process took place in small-scale, short-cycle experiments with multilevel stakeholders, which is not common in the implementation in Dutch health care. Small-scale, short-cycle testing is especially relevant for eHealth technologies, which evolve rapidly and require adaptation to social and contextual factors for adoption [42]. Therefore, scholars promote small-scale iteration and agile methods such as "fail early, fail often," without long-term, extensive datasets, to reduce delays and costs [12,41,43]. However, such an iterative approach conflicts with traditional development and lengthy outcome measure processes in the health care industry (eg, medical device regulations or randomized controlled trials). Health care is considered a risk-averse and regulated industry [43]. GzGr involved practice-oriented academic partners, which can foster implementation in a risk-averse industry [44].

Another difference between the technological and health care industries is how implementation is perceived. In the digital industry, implementation is often seen as the final phase and end point, whereas in health care, it is a complex, ongoing process and should not be treated as the last step [36]. We advocate for the mechanism of iterative, co-creative experimentation that bridges development and implementation, rather than traditional, phased models. The National Health Service also implemented a continuous learning and adapting iterative process, aligning the technology with adopters, which requires iterative refinement and an agile approach, leading to regional adoption [45].

In such an iterative process (rather common in digital innovation in other industries [46]), it is crucial to involve all relevant stakeholders in the design and test phase to improve the implementation success of digital health interventions, realizing a co-creative process [1,40,47,48]. By involving end users in the development, a collaboration can anticipate implementation problems. Additionally, consensus on CDSS functionality and added value supports implementation [48,49]. The sequential pattern of actions included user feedback, which led to rapid technological improvements, strengthening trust in both the mission and the collaboration before broader implementation to more risk-averse groups. This method is in line with AI adoption within the National Health Service [20].

Previously mentioned studies mainly emphasized end user involvement (particularly health care professionals) during development in controlled settings. However, involving multilevel stakeholders during this process helps generate buy-in by aligning intervention design with strategic goals, thereby supporting adoption and keeping service models central [1,13,41,47,50,51]. In doing so, GzGr illustrates how to overcome the common pitfall of "pilotitis," where digital health innovations stall in the pilot phase without scaling [52].

Our findings contribute by identifying the sequential patterns of actions that underlie an iterative, co-creative experimentation in a risk-averse setting. GzGr created a continuous, co-creative experimentation process spanning development and implementation and sustainably involving multilevel stakeholders, who kept the mission in mind and did not treat development and implementation as separate phases. This illustrates how joint progress can be achieved by connecting social change and technological development, which was crucial for achieving sustainable implementation of CDSS.

Limitations

To interpret the results of this study, we have reflected on several limitations in our research. First, the focus on a single case in the Netherlands restricts the generalizability of the findings to other contexts. CDSSs have been scaled to other organizations and are currently being expanded to additional regions. However, it remains uncertain to what extent the collaboration's approach to CDSS development and implementation is dependent on context or how this compares to other countries. Further research is needed to validate these findings across different organizations, regions, and countries.

Second, reflecting on our use of qualitative data and a retrospective analysis, recalling bias and interpretative variation among respondents might exist. We mitigated this limitation by combining data sources (interviews, documents, and GzGr experts), making use of data triangulation. Moreover, our transdisciplinary research team enriched qualitative data analysis by incorporating diverse perspectives, including medical expertise, innovation ecosystems, organizational governance, design, and implementation. This multiperspective approach enhances the study's applicability and relevance, despite its single-case focus. Additionally, the retrospective character of this research gave us the opportunity to compose a chronological overview. Due to this nature, the study has a descriptive rather than a predictive value, as formally no causality between actions, sequences of actions, and effects was assessed. Future research may provide predictive insights into how the timeline of the sequential patterns of actions and mechanisms unfolds during implementation.

Third, our identification of 2 mechanisms may reflect a degree of interpretation of the researchers, as is inherent to qualitative research. We addressed this by documenting our observations in a chronological overview table (Multimedia Appendix 3),

XSL•FO
RenderX

performing member checks, and linking our findings to existing literature. These factors provide a thorough understanding of the mechanisms underlying CDSS development and implementation in an interorganizational context.

Implications and Future Research

This research identified mechanisms that help interorganizational collaborations overcome challenges in the iterative development and implementation of CDSS. In this perspective, our study extends the determinant-level frameworks NASSS, Centre for eHealth and Wellbeing Research, and UK Medical Research Council by adding an interorganizational, process-level perspective that captures how collaboration unfolds across independent organizations. It shows how organizations, while maintaining their own interests, continuously realign around a shared goal through iterative steps to compose an interorganizational value model that sustains collective progress.

Our practical implications consist of 3 elements. First, to navigate complexity, the development and implementation of CDSS in interorganizational collaborations should involve multilevel stakeholders from the start of the collaboration, especially innovative and influential individuals who are willing to go beyond formal roles to advance shared goals and form a coalition of the willing. Second, an orchestrating actor, such as GzGr, is needed to facilitate communication, alignment, and technical support across individual organizations within interorganizational collaborations. Third, an iterative approach should be adopted, where feedback from multilevel stakeholders informs continuous technological and collaborative improvements that will support both adoption and long-term collaboration.

Future research should enhance the generalizability of our findings by comparing this case with other interorganizational collaborations to identify shared mechanisms. Additionally, further investigation is needed into the scalability of CDSS, distinguishing context-dependent elements. Finally, future research should explore how these mechanisms can be systematically applied to support adaptive implementation in research and practice.

Conclusions

Our study offered insights into the mechanisms for developing and implementing CDSS in an interorganizational collaboration in primary care. Our findings extend the determinant-level implementation frameworks for complex digital health interventions by offering process-level insights into sequential patterns of actions that show how change unfolds in complex, real-world settings. We identified two mechanisms: (1) enacting an interorganizational value model and (2) iterative, co-creative experimentation. These mechanisms supported both technological and organizational change, through the approach of integrating people, technology, and business aspects for all and across organizations. Through this approach, CDSS were generated and implemented via an iterative, co-creative experiment. This resulted in CDSS that were (financially) viable and that generated added value per individual stakeholder and organization. Additionally, CDSS could be embedded in the individual organizations as they fulfilled their values, needs, and interests. Therefore, CDSSs reached multilevel support and were well adopted.

Overall, our study enhances the understanding of how sequential patterns of actions enable digital health innovations to be embedded and sustained in an interorganizational collaboration. It highlights that successful implementation in complex systems depends on establishing shared principles early, forming the basis for an interorganizational value model. Simultaneously, an iterative collaboration spanning development and implementation is required. Throughout this process, continuous alignment of social, technological, and organizational dimensions is essential. We offered both practical and conceptual contributions to understanding how digital health interventions can be embedded and sustained in interorganizational collaboration.

Acknowledgments

The authors would like to thank all participants and members of GzGr for sharing their time, insights, and openness to reflect on their development and implementation process. The authors would also like to thank Prof Dr Anne Marie Weggelaar for her role in securing the funding that made this project possible. Additionally, the authors appreciate the constructive input of our interdisciplinary author team throughout the research and writing process. Throughout manuscript development, we conducted this review with the assistance of ChatGPT, versions 4 and 5, from OpenAI in 2024 and 2025. We used generative artificial intelligence as a sparring partner (eg, for conceptual distinctions) to improve English phrasing and condense text. We carefully reviewed and refined all content to ensure our perceptions, analytical depth, and conceptual coherence.

Funding

Data collection for this study was made possible through funding provided by the BigMedilytics program. The authors were supported by their respective institutions; no additional external funding was obtained for this work. The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request. The authors chose this approach because the transcripts include personal details and context-specific reflections, some of which were not relevant for the analysis and may be sensitive for collaborating partners.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: JJMMS, JDHvW, EGH, MK, FD, HJAvO
Data curation: JJMMS
Formal analysis: JJMMS, JDHvW, HJAvO
Investigation: JDHvW, JJMMS
Methodology: MK, FD, JJMMS
Supervision: MK, FD
Validation: HJAvO, EGH
Writing – original draft: JJMMS, HJAvO, JDHvW
Writing – review & editing: MK, FD, EH

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of the participants.
[DOCX File, 53 KB - jmir_v28i1e83084_app1.docx]

Multimedia Appendix 2
Elaboration on coding and codebook process.
[DOCX File, 57 KB - jmir_v28i1e83084_app2.docx]

Multimedia Appendix 3
Chronological overview of events, challenges, and mechanisms encountered by *Gezonde zorg, Gezonde regio*.
[DOCX File, 57 KB - jmir_v28i1e83084_app3.docx]

Checklist 1
COREQ checklist.
[PDF File, 491 KB - jmir_v28i1e83084_app4.pdf]

References

1. Abell B, Naicker S, Rodwell D, et al. Identifying barriers and facilitators to successful implementation of computerized clinical decision support systems in hospitals: a NASSS framework-informed scoping review. Implement Sci 2023 Jul 26;18(1):32. [doi: 10.1186/s13012-023-01287-y] [Medline: 37495997]
2. Harada T, Miyagami T, Kunitomo K, Shimizu T. Clinical decision support systems for diagnosis in primary care: a scoping review. Int J Environ Res Public Health 2021 Aug 10;18(16):8435. [doi: 10.3390/ijerph18168435] [Medline: 34444182]
3. Ji M, Genchev GZ, Huang H, Xu T, Lu H, Yu G. Evaluation framework for successful artificial intelligence-enabled clinical decision support systems: mixed methods study. J Med Internet Res 2021 Jun 2;23(6):e25929. [doi: 10.2196/25929] [Medline: 34076581]
4. Jamian L, Wheless L, Crofford LJ, Barnado A. Rule-based and machine learning algorithms identify patients with systemic sclerosis accurately in the electronic health record. Arthritis Res Ther 2019 Dec 30;21(1):305. [doi: 10.1186/s13075-019-2092-7] [Medline: 31888720]
5. Devaraj S, Sharma SK, Fausto DJ, Viernes S, Kharrazi H. Barriers and facilitators to clinical decision support systems adoption: a systematic review. J Bus Adm Res 2014;3(2):36. [doi: 10.5430/jbar.v3n2p36]
6. Laka M, Milazzo A, Merlin T. Factors that impact the adoption of clinical decision support systems (CDSS) for antibiotic management. Int J Environ Res Public Health 2021 Feb 16;18(4):1901. [doi: 10.3390/ijerph18041901] [Medline: 33669353]
7. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3(1):17. [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
8. Meunier PY, Raynaud C, Guimaraes E, Gueyffier F, Letrilliart L. Barriers and facilitators to the use of clinical decision support systems in primary care: a mixed-methods systematic review. Ann Fam Med 2023;21(1):57-69. [doi: 10.1370/afm.2908] [Medline: 36690490]
9. Toth-Pal E, Wårdh I, Strender LE, Nilsson G. Implementing a clinical decision-support system in practice: a qualitative analysis of influencing attitudes and characteristics among general practitioners. Inform Health Soc Care 2008 Mar;33(1):39-54. [doi: 10.1080/17538150801956754] [Medline: 18604761]
10. O'Cathain A, Croot L, Duncan E, et al. Guidance on how to develop complex interventions to improve health and healthcare. BMJ Open 2019 Aug 15;9(8):e029954. [doi: 10.1136/bmjopen-2019-029954] [Medline: 31420394]

XSL•FO
RenderX

11. van Dyk L. A review of telehealth service implementation frameworks. Int J Environ Res Public Health 2014 Jan 23;11(2):1279-1298. [doi: 10.3390/ijerph110201279] [Medline: 24464237]
12. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
13. van Gemert-Pijnen JEWC, Nijland N, van Limburg M, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011 Dec 5;13(4):e111. [doi: 10.2196/jmir.1672] [Medline: 22155738]
14. Doolin B. Implementing e-health. In: The Oxford Handbook of Health Care Management: Oxford University Press; 2016:352-374 URL: https://books.google.co.in/ [accessed 2025-07-25]
15. Plsek PE, Greenhalgh T. The challenge of complexity in health care. BMJ 2001 Sep 15;323(7313):625-628. [doi: 10.1136/bmj.323.7313.625]
16. van Limburg M, van Gemert-Pijnen JEWC, Nijland N, Ossebaard HC, Hendrix RMG, Seydel ER. Why business modeling is crucial in the development of eHealth technologies. J Med Internet Res 2011 Dec 28;13(4):e124. [doi: 10.2196/jmir.1674] [Medline: 22204896]
17. Cherns A. The principles of sociotechnical design. Hum Relat 1976 Aug;29(8):783-792. [doi: 10.1177/001872677602900806]
18. Kip H, Beerlage-de Jong N, van Gemert-Pijnen LJEWC, Kelders SM. The CeHRes roadmap 2.0: update of a holistic framework for development, implementation, and evaluation of eHealth technologies. J Med Internet Res 2025 Jan 13;27:e59601. [doi: 10.2196/59601] [Medline: 39805104]
19. Nilsen P. Making sense of implementation theories, models and frameworks. Implement Sci 2015 Apr 21;10(1):53. [doi: 10.1186/s13012-015-0242-0] [Medline: 25895742]
20. Morrison K. Artificial intelligence and the NHS: a qualitative exploration of the factors influencing adoption. Future Healthc J 2021 Nov;8(3):e648-e654. [doi: 10.7861/fhj.2020-0258] [Medline: 34888459]
21. May CR, Johnson M, Finch T. Implementation, context and complexity. Implement Sci 2016 Oct 19;11(1):141. [doi: 10.1186/s13012-016-0506-3] [Medline: 27756414]
22. Aunger JA, Millar R, Greenhalgh J, Mannion R, Rafferty AM, McLeod H. Why do some inter-organisational collaborations in healthcare work when others do not? A realist review. Syst Rev 2021 Mar 22;10(1):82. [doi: 10.1186/s13643-021-01630-8] [Medline: 33752755]
23. Karam M, Brault I, Van Durme T, Macq J. Comparing interprofessional and interorganizational collaboration in healthcare: a systematic review of the qualitative research. Int J Nurs Stud 2018 Mar;79:70-83. [doi: 10.1016/j.ijnurstu.2017.11.002] [Medline: 29202313]
24. Moullin JC, Dickson KS, Stadnick NA, et al. Ten recommendations for using implementation frameworks in research and practice. Implement Sci Commun 2020;1:42. [doi: 10.1186/s43058-020-00023-7] [Medline: 32885199]
25. Yin RK. Case Study Research: Design and Methods: Sage; 2009. [doi: 10.33524/cjar.v14i1.73]
26. Tracy SJ. Qualitative Research Methods: Collecting Evidence, Crafting Analysis, Communicating Impact: Wiley-Blackwell; 2024.
27. Bryman A. Social Research Methods, 5th edition: Oxford University Press; 2016. URL: https://archive.org/details/socialresearchme0000brym_w6b0 [accessed 2026-02-06]
28. Langley A. Strategies for theorizing from process data. Acad Manag Rev 1999 Oct;24(4):691. [doi: 10.5465/amr.1999.2553248]
29. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
30. Organisatie en partners [Website in Dutch]. Gezonde zorg, Gezonde regio. 2024. URL: https://gzgr.nl/over-gzgr/organisatie-en-partners [accessed 2024-09-13]
31. Zo gebruiken wij data [Website in Dutch]. Gezonde zorg, Gezonde regio. 2024. URL: https://gzgr.nl/over-gzgr/zo-gebruiken-wij-data [accessed 2024-09-13]
32. Flick U. Designing Qualitative Research, 2nd edition: SAGE Publications; 2018. [doi: 10.4135/9781529622737]
33. Aguboshim FC. Adequacy of sample size in a qualitative case study and the dilemma of data saturation: a narrative review. World J Adv Res Rev 2021;10(3):180-187. [doi: 10.30574/wjarr.2021.10.3.0277]
34. Parker C, Scott S, Geddes A. In: Cernat A, Sakshaug JW, Delamon S, editors. Snowball Sampling: SAGE Publications Ltd; 2019. [doi: 10.4135/9781526421036831710]
35. Braun V, Clarke V, Hayfield N, Terry G. Thematic analysis. In: Liamputtong P, editor. Handbook of Research Methods in Health Social Sciences: Springer; 2019:843-860. [doi: 10.1007/978-981-10-5251-4_103]
36. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]

XSL•FO
RenderX

37. Bohnett E, Vacca R, Hu Y, Hulse D, Varda D. Resilience and fragmentation in healthcare coalitions: the link between resource contributions and centrality in health-related interorganizational networks. Soc Networks 2022 Oct;71:87-95. [doi: 10.1016/j.socnet.2022.07.004] [Medline: 36060606]
38. Porter ME. What is value in health care? N Engl J Med 2010 Dec 23;363(26):2477-2481. [doi: 10.1056/NEJMp1011024] [Medline: 21142528]
39. Rakers MM, van Os HJA, Recourt K, Mosis G, Chavannes NH, Struijs JN. Perceived barriers and facilitators of structural reimbursement for remote patient monitoring, an exploratory qualitative study. Health Policy Technol 2023 Mar;12(1):100718. [doi: 10.1016/j.hlpt.2022.100718]
40. Seljelid B, Varsi C, Solberg Nes L, Øystese KA, Børøsund E. A digital patient-provider communication intervention (InvolveMe): qualitative study on the implementation preparation based on identified facilitators and barriers. J Med Internet Res 2021 Apr 8;23(4):e22399. [doi: 10.2196/22399] [Medline: 33830063]
41. Christie HL, Boots LMM, Hermans I, et al. Business models of eHealth interventions to support informal caregivers of people with dementia in the Netherlands: analysis of case studies. JMIR Aging 2021 Jun 3;4(2):e24724. [doi: 10.2196/24724] [Medline: 34081009]
42. Kumar S, Nilsen WJ, Abernethy A, et al. Mobile health technology evaluation: the mHealth evidence workshop. Am J Prev Med 2013 Aug;45(2):228-236. [doi: 10.1016/j.amepre.2013.03.017] [Medline: 23867031]
43. Duffy A, Christie GJ, Moreno S. The challenges toward real-world implementation of digital health design approaches: narrative review. JMIR Hum Factors 2022 Sep 9;9(3):e35693. [doi: 10.2196/35693] [Medline: 36083628]
44. Mervyn K, Amoo N, Malby R. Challenges and insights in inter-organizational collaborative healthcare networks: an empirical case study of a place-based network. Int J Organ Anal 2019;27(4):875-902. [doi: 10.1108/IJOA-05-2018-1415]
45. Elizondo A, Williams R, Anderson S, Cresswell K. Implementing integrated care infrastructure: a longitudinal study on the interplay of policies, interorganizational arrangements and interoperability in NHS England. Health Policy 2025 Oct;160:105237. [doi: 10.1016/j.healthpol.2024.105237] [Medline: 39706716]
46. Brown T. Design thinking. Harv Bus Rev 2008 Jun;86(6):84-92. [Medline: 18605031]
47. Ackerhans S, Huynh T, Kaiser C, Schultz C. Exploring the role of professional identity in the implementation of clinical decision support systems-a narrative review. Implement Sci 2024 Feb 12;19(1):11. [doi: 10.1186/s13012-024-01339-x] [Medline: 38347525]
48. Eichner JS, Das M. Challenges and barriers to clinical decision support (CDS) design and implementation experienced in the Agency for Healthcare Research and Quality CDS demonstrations (prepared for the AHRQ National Resource Center for Health Information Technology under contract no. 290-04-0016). : Agency for Healthcare Research and Quality; 2012 URL: https://digital.ahrq.gov/sites/default/files/docs/page/CDS_challenges_and_barriers.pdf [accessed 2025-08-15]
49. Schroeder D, Luig T, Finch TL, Beesoon S, Campbell-Scherer DL. Understanding implementation context and social processes through integrating normalization process theory (NPT) and the consolidated framework for implementation research (CFIR). Implement Sci Commun 2022 Feb 9;3(1):13. [doi: 10.1186/s43058-022-00264-8] [Medline: 35139915]
50. Tossaint-Schoenmakers R, Kasteleyn MJ, Rauwerdink A, Chavannes N, Willems S, Talboom-Kamp EPWA. Development of a quality management model and self-assessment questionnaire for hybrid health care: concept mapping study. JMIR Form Res 2022 Jul 7;6(7):e38683. [doi: 10.2196/38683] [Medline: 35797097]
51. van Velsen L, Ludden G, Grünloh C. The limitations of user-and human-centered design in an eHealth context and how to move beyond them. J Med Internet Res 2022 Oct 5;24(10):e37341. [doi: 10.2196/37341] [Medline: 36197718]
52. Pannunzio V. Towards a convergent approach to the use of data in digital health design [Dissertation]. : Delft University of Technology; 2023 URL: https://research.tudelft.nl/en/publications/towards-a-convergent-approach-to-the-use-of-data-in-digital-healt/ [accessed 2026-01-30]

ABBREVIATIONS

CDSS: clinical decision support systems
COREQ: Consolidated Criteria for Reporting Qualitative Research
GP: general practitioner
GzGr: *Gezonde zorg*, *Gezonde regio* (Healthy care, Healthy region)
NASSS: Nonadoption, abandonment, scale-up, spread, and sustainability

XSL•FO
RenderX

Edited by S Brini; submitted 27.Aug.2025; peer-reviewed by R Akinniranye, RS Mannak, Y Wu; accepted 11.Nov.2025; published 05.Mar.2026.

Please cite as:
Santema JJMM, van Wijngaarden JDH, Hiddink EG, Deken F, Kleinsmann M, van Os HJA
Interorganizational Mechanisms for Developing and Implementing Clinical Decision Support Systems in Primary Care: Exploratory, Qualitative Case Study
J Med Internet Res 2026;28:e83084
URL: https://www.jmir.org/2026/1/e83084
doi:10.2196/83084

© Jesse J M M Santema, Jeroen D H van Wijngaarden, Eric G Hiddink, Fleur Deken, Maaike Kleinsmann, Hendrikus J A van Os. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessing Health Care Professionals' Perceptions of a New System in Clinical Workflows: Systems Engineering Initiative for Patient Safety–Based Consensual Qualitative Research

Ye-Eun Park[1*], BA; Minsu Ock[2*], MD, PhD; Jae-Ho Lee[1,3], MD, PhD; Dae-Hyun Ko[4], MD, PhD; Hak-Jae Lee[5], MD, PhD; Taezoon Park[6], PhD; Junsang Yoo[7], PhD; Yura Lee[1], MD, PhD

[1]Department of Information Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[2]Department of Preventive Medicine, Ulsan University Hospital, University of Ulsan College of Medicine, Ulsan, Republic of Korea
[3]Department of Emergency Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[4]Department of Laboratory Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[5]Division of Acute Care Surgery, Department of Surgery, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[6]Department of Industrial & Information Systems Engineering, Soongsil University, Seoul, Republic of Korea
[7]Department of Digital Health, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Yura Lee, MD, PhD
Department of Information Medicine
Asan Medical Center
University of Ulsan College of Medicine
Seoul
Republic of Korea
Phone: 82 2 3010 1498
Fax: 82 2 3010 2531
Email: haepary@naver.com

Abstract

Background: Artificial intelligence (AI)–enabled clinical decision support systems (CDSSs) are increasingly embedded within electronic health record (EHR) environments; however, their introduction can disrupt existing workflows and raise patient safety concerns, particularly in high-stakes settings such as surgical transfusion. Limited qualitative evidence exists regarding how frontline professionals anticipate the clinical, organizational, and workflow implications of such systems before wider deployment.

Objective: This study aims to qualitatively examine the anticipated clinical, organizational, and workflow-level implications of implementing personalized Maximum Surgical Blood Order Schedule—Thoracic Surgery (pMSBOS-TS), an AI-enabled CDSS for personalized surgical blood ordering, before large-scale deployment.

Methods: We conducted a consensual qualitative study with 14 multidisciplinary health care professionals involved in transfusion-related tasks at a large tertiary hospital. Following 1 pilot focus group to refine the interview guide and workflow diagram, 2 semistructured focus group discussions were held with 14 participants (5 physicians, 6 nurses, and 3 blood bank staff). Transcripts were analyzed using the Systems Engineering Initiative for Patient Safety (SEIPS) 101 framework, focusing on People, Environment, Tools, and Tasks, and were supported by task- and workflow-based analyses of transfusion processes. Member checking was conducted with participants and external clinicians to enhance validity.

Results: A total of 189 semantic units and 61 core ideas were identified across 18 subdomains and 7 overarching domains. Participants anticipated that pMSBOS-TS could reduce unwarranted variation in blood ordering and planning, provided that algorithmic performance is reliable and the interface is tightly integrated into existing EHR workflows. At the same time, they expressed concerns regarding increased verification burden, system limitations in unexpected clinical scenarios, and potential communication bottlenecks between clinical units and the blood bank. Organizational culture, governance structures, and local transfusion logistics were viewed as critical determinants of whether the system would reduce or inadvertently increase workload and blood product waste.

XSL•FO
RenderX

Conclusions: This preimplementation, SEIPS-based qualitative evaluation suggests that the successful adoption of an AI-enabled transfusion CDSS depends not only on predictive performance but also on sociotechnical readiness, including user trust, workflow fit, and organizational support. These findings provide practice-based insights to inform staged implementation, training, and governance strategies aimed at safely integrating predictive transfusion CDSSs into EHR-supported surgical workflows.

(J Med Internet Res 2026;28:e86166) doi:10.2196/86166

KEYWORDS

decision support systems (clinical); precision medicine, patient safety; blood transfusion; algorithms (artificial intelligence); information system

Introduction

The integration of artificial intelligence (AI) and digital systems into clinical practice is transforming health care delivery. As health care professionals increasingly encounter emerging technologies, their acceptance and perceptions of these systems substantially influence clinical efficiency and implementation success [1-3]. Despite growing interest in digital innovations, including clinical decision support systems (CDSSs), their adoption can disrupt existing workflows, necessitating careful evaluation of organizational- and user-level impacts [4,5].

Although prior workflow analyses have predominantly focused on identifying the root causes of patient safety incidents [6-8], there is an increasing need for proactive assessments during the early stages of system deployment. In particular, the rapid proliferation of electronic health record (EHR)–based applications requires a deeper understanding of how novel systems interface with existing clinical processes. Nevertheless, comprehensive workflow analyses addressing the multifaceted challenges of system adoption remain limited.

The widespread implementation of EHRs, accelerated by initiatives such as the Meaningful Use Program under the Health Information Technology for Economic and Clinical Health Act [9-13], has catalyzed the development of embedded CDSSs to enhance care quality and operational efficiency [14-16]. However, the adoption and integration of such tools remain complex, particularly in high-stakes settings such as surgical transfusion.

Traditional evaluation approaches, such as user satisfaction surveys or system log data, are useful for capturing surface-level feedback but often fall short in explaining context-dependent interactions among health care professionals and between users and systems [17]. Consequently, rigorous qualitative methodologies are essential for understanding the nuanced implications of technology integration [18].

Accordingly, this consensual qualitative research (CQR) study aimed to conduct a preimplementation evaluation of the personalized Maximum Surgical Blood Order Schedule—Thoracic Surgery (pMSBOS-TS) system by examining its anticipated implications for workflows, usability, and organizational conditions, and by identifying factors that may support its safe and effective integration into clinical practice [19]. To our knowledge, only a few studies have examined AI-enabled CDSSs for transfusion planning using a structured preimplementation evaluation; most existing CDSS research has focused on postdeployment clinician acceptance and use rather than prospective workflow and system-integration assessments [20]. By applying CQR alongside a sociotechnical framework before system deployment, this study provides a multistakeholder assessment of how an AI-based transfusion tool may influence workflows, communication patterns, and organizational processes.

Methods

System Description (pMSBOS-TS)

pMSBOS-TS is a machine learning–based CDSS developed to generate personalized maximum blood-ordering recommendations for thoracic surgery patients by integrating patient-, laboratory-, and procedure-specific predictors [21]. The underlying algorithms were developed and validated in a prior work by the collaborating investigators [19], and this study evaluates the system's anticipated effects in a real-world clinical context.

Study Site

This study was conducted at Asan Medical Center, a 2764-bed tertiary hospital that performs approximately 70,892 surgeries annually as of 2023.

Study Design

This qualitative study examined the anticipated effects of applying the pMSBOS-TS system for health care professionals involved in transfusion tasks, under the assumption that the system was integrated into the existing electronic medical record (EMR). The assessment used the Systems Engineering Initiative for Patient Safety (SEIPS) framework to evaluate potential impacts across the domains of People, Environment, Tools, and Tasks (PETT) [22,23].

Originally introduced by Carayon et al [22] and subsequently expanded, SEIPS conceptualizes health care as comprising interacting components—person, task, technology/tools, organization, and environment—that shape care processes and, ultimately, outcomes such as patient safety and care quality [23,24]. SEIPS 2.0 incorporated patients and families as active participants and placed greater emphasis on processes, while SEIPS 3.0 expanded the framework to the patient-journey level, including cross-setting transitions.

To enhance accessibility, Holden and Carayon [23] proposed "SEIPS 101," a simplified, practice-oriented adaptation of the model. SEIPS 101 retains the core elements of the work system, process, and outcomes, while streamlining the work system into 4 primary components (PETT). In this study, *people* refer to

XSL•FO
RenderX

health care stakeholders (clinicians and patients) and their capabilities or needs; *tasks* denote activities and workflows; *tools/technology* include equipment, information technology (IT) systems (CDSS/health information systems), and other job aids; and *environment* encompasses both the physical setting (eg, layout, lighting, and noise) and the social or organizational context (eg, culture, policies, and teamwork).

By focusing on these elements and their interactions, a SEIPS/PETT-based evaluation can map how a new CDSS influences clinical work—for instance, how it reshapes clinicians' tasks, introduces tool-related issues, or alters team dynamics—and how these changes affect care processes and outcomes. This sociotechnical perspective is particularly useful for identifying system-level issues that purely clinical or IT-focused evaluations may overlook [25].

A key advantage of this approach is its holistic orientation toward workflow and safety, providing a structured method for analyzing and designing health care processes. The PETT scan can serve as both a checklist and a documentation tool to ensure that all relevant components of the work system are systematically addressed. Accordingly, we applied the PETT scan to identify barriers and facilitators across components and to explore their interactions, thereby providing a comprehensive overview of the work system. Our methodology systematically applied the PETT scan tool from the SEIPS framework to assess users, the surrounding environment, tasks, and tools/technology (Figure 1).

Three group sessions were conducted between October 19, 2023, and May 2, 2024, following a structured 5-stage process: participant recruitment, workflow feedback, system introduction, PETT-based evaluation, and postsession feedback (Figure 2).

Figure 1. Simplified SEIPS 101 model of work systems, processes, and outcomes. SEIPS: Systems Engineering Initiative for Patient Safety.

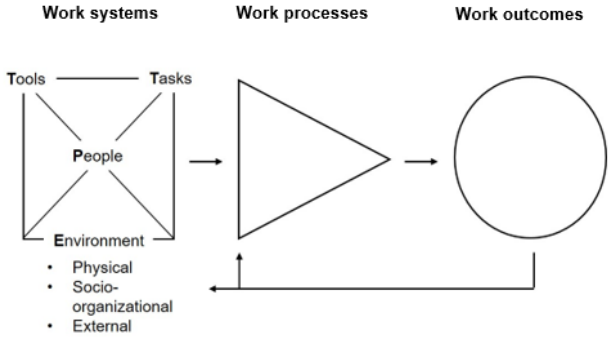

Figure 2. Overview of the participant engagement and SEIPS-PETT analysis process. PETT: People, Environment, Tools, and Tasks; pMSBOS-TS: personalized Maximum Surgical Blood Order Schedule—Thoracic Surgery; SEIPS: Systems Engineering Initiative for Patient Safety.

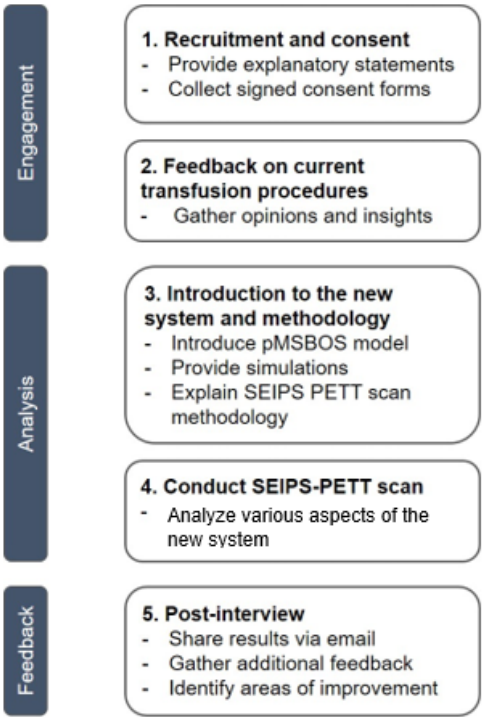

XSL•FO
RenderX

Recruitment

To assess the usability and workflow implications of the pMSBOS-TS system, we recruited health care professionals currently involved in transfusion-related tasks and identified as potential end users of the system. Recruitment was conducted via the hospital's online bulletin board and direct outreach. Eligible participants were those with more than 1 year of experience in procedures or surgeries with a high likelihood of requiring transfusion and who had experience handling blood products.

Inclusion criteria were surgeons prescribing blood products, nurses in surgical wards and operating rooms handling blood requests and transfusions, and blood bank staff managing the release and distribution of blood products.

Three focus group discussions (FGDs) were conducted for this study: 1 pilot session and 2 main sessions, each comprising 3-8 participants. Group composition was balanced across roles (eg, surgeons, nurses, and blood bank personnel) to gather diverse perspectives while maintaining thematic consistency. A semistructured interview guide was used for all sessions, and each participant took part in only 1 session.

Data Collection

All FGDs were conducted in person in a private meeting room within the hospital. No individual interviews were conducted; all qualitative data were collected through FGDs. Each session was facilitated by YL (female), a trained moderator with expertise in qualitative health research, and assisted by YEP (female), who managed logistics, note-taking, and audio recording. Although the researchers and participants worked within the same institution, no direct personal or supervisory relationships existed. Participants were informed of the study objectives and were aware that the moderator was a member of the research team; no personal goals or interests of the researchers were disclosed.

A semistructured interview guide was used in all sessions. Each session lasted 60-80 minutes and was audio-recorded with participant permission. Participants were encouraged to describe specific cases and reflect on workflow steps using printed workflow diagrams and sticky-note categorization. All participants were invited to speak, write, categorize notes, and present their views during the sessions.

All interviews were conducted in Korean, transcribed verbatim, reviewed by participants, and subsequently translated into English. Translation was performed by bilingual researchers fluent in both Korean and English. To ensure semantic equivalence, translated transcripts were reviewed by a second bilingual researcher, and discrepancies were resolved through discussion. To minimize interpretive bias, participants were emailed the FGD transcripts organized by content unit to verify accuracy and identify any distortions or omissions in meaning. No participants had a prior personal relationship with the moderator.

The methodology and reporting of the qualitative findings followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines [26]. We adhered to the COREQ guidelines in describing the study design, data collection, and qualitative analysis procedures.

Qualitative Analysis

Overview

The qualitative analysis was conducted in accordance with the principles of CQR to minimize interpretive distortion and to derive in-depth insights from clinical field experts regarding the use of pMSBOS-TS [27].

Initial Stage of Qualitative Analysis: Preliminary Discussions and Hypothesis Formation

The research team hypothesized that personalized prediction could support more efficient use of blood products. Based on a literature review of the adoption of new medical information systems and expert discussions, we concluded that, in addition to understanding the risks and benefits of pMSBOS-TS, factors related to its successful implementation would be essential [28,29].

Accordingly, the following core questions were developed:

- "What positive or negative impacts could be expected if pMSBOS-TS is introduced into the clinical field?"
- "What factors are important for the successful adaptation and implementation of pMSBOS-TS in clinical practice?"

To conduct a multidimensional evaluation of the risks and impacts associated with introducing a new application, we employed the PETT scan framework. Prior research using the SEIPS model was reviewed to inform the development of the interview guide [22,23]. Participants actively engaged with the PETT framework by manually categorizing insights using sticky notes, which facilitated recall and reflection. They also participated in a transfusion-related workflow analysis to identify potential impacts at each step.

A pilot focus group session was conducted with clinical professionals to examine transfusion-related processes in greater detail and to identify clinical scenarios relevant to the application of pMSBOS-TS. This pilot session served multiple purposes: it validated the structure and appropriateness of the PETT-based interview methodology, supported the development of a draft transfusion workflow diagram, and informed the refinement of the initial domain-subcategory-core idea table. Although the pilot session provided useful foundational insights, its data were excluded from the final cross-case analysis to ensure consistency with the main dataset.

Intracase Analysis: Task Analysis and Development of Domains and Subdomains

To ensure comprehensive capture of participants' views on the impacts and potential harms of pMSBOS-TS at each stage of the transfusion process, a detailed task analysis was performed [30,31]. The analysis team, comprising 2 PhD-level researchers and 1 research assistant with a bachelor's degree, developed the following domains, subdomains, and core ideas based on the core questions:

XSL•FO
RenderX

- Domains: transfusion-related work experiences (case examples), pMSBOS-TS–related opinions, and major categories derived from the PETT scan.
- Subdomains: specific subcomponents of the PETT scan and preliminary workflow elements identified during the pilot focus group sessions.
- Core ideas: derived by reviewing interview transcripts and postinterview notes; preliminary categorizations were validated through consultation with 2 domain experts.

Cross-Case Analysis

Two analysis team members (YEP and YL) independently reviewed and coded meaningful units from the transcripts and sticky notes into subdomains and core ideas. Coding discrepancies were resolved through consensus discussion or, when necessary, adjudicated by a third team member. Core ideas that did not fit existing categories were temporarily placed in an "Other" category and redefined as needed.

A frequency-based coding scheme was employed to reflect both sentiment and the prevalence of opinions: positive opinions were marked with plus signs (+), and negative opinions with minus signs (–). Frequency was indicated as follows: 1 symbol for 1 participant, 2 symbols for 2-3 participants, and 3 symbols for 4 or more participants. Expression formats and interpretation strategies were discussed and agreed upon for divergent opinions within the same theme.

Categories were defined as follows:

- General: common across all sessions (– – –/+ + +).
- Typical: consistent within a session but not endorsed by all participants.
- Variant: mixed opinions or views that appeared in only a subset of cases.

For validation, the finalized core ideas were shared via email with the original participants for member checking to ensure accuracy and minimize interpretive distortion. In addition, 3 clinical professionals from other medical institutions (meeting the same inclusion criteria as the interview participants) were contacted. After explaining the study objectives and methodology, the core ideas and task-analysis flow diagram were shared via email to confirm that the interpretations were neither biased nor incomplete. Data saturation was assessed during cross-case analysis; no new core ideas emerged from experts outside the interview group, indicating thematic saturation consistent with CQR guidelines.

Themes and core ideas were derived inductively from the data in accordance with the CQR approach, whereas the overarching domains were organized using the SEIPS/PETT framework. No themes were predetermined before analysis.

Ethics Statement

This study was approved by the Institutional Review Board (IRB) of Asan Medical Center, Korea (IRB 2023-0724), and conducted in accordance with relevant ethical guidelines. Informed consent was obtained from all participants before their involvement, with assurances of anonymity and confidentiality. Participants were briefed on the study's objectives and the intended use of the collected data. Participants received compensation for their participation.

Results

Participants

Fourteen participants took part in 2 FGDs, comprising 5 physicians (3 fellows/professors and 2 residents), 6 nurses (4 from the surgical intensive care unit, 1 from the anesthesia and recovery unit, and 1 from the internal medicine ward), and 3 blood bank staff members (Table 1). Detailed participant characteristics, including participant identifiers, are provided in Multimedia Appendix 1.

Table 1. Demographics of the participants.

Variables	Doctor (n=5)	Medical technologist (n=3)	Nurse (n=6)	Total (N=14)
Sex, n				
Male	2	1	0	3
Female	3	2	6	11
Age, mean (SD)	30 (11)	40 (14.1)	28.3 (6.9)	31.4 (11.7)
Transfusion experience, mean (SD)	8.2 (9.4)	20.3 (10.1)	13 (7.2)	13.53 (10.12)
Interviews participated in, n				
First	2	3	3	8
Second	3	0	3	6

We organized the results of a task analysis of transfusion-related processes using a swim-lane approach (Multimedia Appendix 2). From the workflow analysis and FGDs, we extracted 189 unique semantic units (81 in category A and 108 in category B, excluding duplicates). These units were organized according to their relevance to transfusion workflows. Semantic units describing inefficient use or cases involving blood products, regardless of whether pMSBOS-TS was applied, were classified as category A. By contrast, statements concerning anticipated impacts of pMSBOS-TS implementation or potential risks associated with its use were classified as category B.

The semantic units were further condensed into 61 core ideas: 21 from category A and 40 from category B. These were then organized into 18 subdomains (3 in category A and 15 in category B) and 7 overarching domains (1 in category A and 6 in category B). The complete analytical framework, including

XSL•FO
RenderX

domains, subdomains, and core ideas, is summarized in Multimedia Appendix 3.

In category A, titled "(In)Efficient Use of Blood Components," 3 subdomains were identified: *Inaccurate Dosage (Over- or Under-Dosing)*, *Inefficient Blood Usage and Processes*, and *Other Transfusion Process Errors*.

Category B comprised 6 primary domains: *People*, *Environment*, *Tools*, *Tasks*, *Work Processes*, and *Work Outcomes*. Within the People domain, 2 subdomains were identified: *Prescribing Physicians* (3 general and 1 variant core idea) and *Other Transfusion Stakeholders* (1 general and 1 variant core idea). The Environment domain included 2 subdomains: *Interprofessional Communication* and *Socio-organizational Context*. The Tools domain encompassed aspects related to algorithmic performance, interface design, and perceived reliability. The Tasks domain included *Achieving Algorithm Objectives*, *Task Complexity and Variability*, and *Handling Unanticipated Situations*.

In the Work Processes domain, responses were classified based on tasks associated with the application phase of pMSBOS-TS (eg, order entry and related processes). The Work Outcomes domain addressed participants' perceptions of the consequences or results associated with the algorithm's use.

Category A: Notable Experiences in Transfusion Practice or Cases of Inefficient Blood Product Use

Of the 21 core ideas from category A, 6 were classified as *typical* and another 6 as *variant*. Within the subdomain "Inaccurate Dosage (Over- or Under-Dosing)," 2 typical and 2 variant core ideas were identified. Participants primarily highlighted experiences related to the prescribing and preparation phases (Table 2).

Regarding unintentional over- or underprescription, several participants referred to habitual or standardized prescribing practices in their departments that led to unnecessary transfusions.

> *We routinely prescribed three units of RBCs, even when we did not actually end up transfusing them...* [Participant 2_5]

A view that routine surgeries rarely require transfusions was also expressed. During the preparation stage, including order verification on the ward and confirmation/preparation in the blood bank, participants noted discrepancies between prescribed and requested volumes.

> *When a patient is going into surgery, they often pre-prescribe fresh frozen plasma, and the blood bank will call us saying it was too much and ask to cancel it. It is a tough situation...* [Participant 1_8]

By contrast, a blood bank staff member pointed out that excessive orders often stemmed from physicians' personal preferences:

> *Sometimes the physician asks us to save much more blood than needed. So we end up placing unnecessary orders. It makes our inventory look abundant to the central blood center, but when we actually have an emergency, there is not much usable blood left.* [Participant 1_2]

In the subdomain of "Inefficient Blood Usage and Processes," 2 typical and 2 variant core ideas were identified. Participants noted issues in communication and procedural transparency during the transfusion preparation process.

> *I often feel there is a disconnect between prescription and preparation. There is a timing issue when blood is prepared, and we are supposed to call the physician directly—but sometimes we miss it. If the doctor does not say, "I have prescribed blood for tomorrow," we will not know until we ask. Then they just say, "Do not worry, it is for tomorrow." But in those communication gaps, we sometimes fail to have the blood ready when the patient actually needs it.* [Participant 1_4]

In addition, logistical difficulties in transporting blood products within the hospital were cited, including challenges related to elevators, distance, and location.

> *Sometimes we cannot even get an elevator. I think it is pretty hard for the ward staff to go pick up the blood themselves. We end up waiting a long time for the blood to arrive...* [Participant 1_4]

Some participants also described inconveniences in transfusion process workflows, including managing low blood inventory levels, emergency dispatch of blood transport vehicles, and cumbersome steps such as cross-matching or additional compatibility testing.

In the subdomain of "Other Transfusion Process Errors," 2 typical and 1 variant core ideas were identified. These primarily involved unexpected clinical events during procedures. For example, some participants described scenarios in which unanticipated vascular damage occurred during surgery.

> *One that really stands out is during CS or liver transplant surgeries—suddenly the aorta or artery tears, and we are hooking up multiple blood bags to the massive transfusion machine. It is just intense during those massive transfusions.* [Participant 1_6]

Other participants reflected on communication breakdowns and disorganized processes during emergencies:

> *Nobody is talking to each other, everyone is busy...Trying to get consent from the caregiver, some people are preparing the transfusion, others are trying to set up IV lines for massive transfusion, even when the lines are not ready yet—they just keep trying anyway...* [Participant 1_5]

XSL•FO
RenderX

Table 2. Cross-case analysis for inefficient use of blood components (category A).

Domain, subdomain, and core idea	Frequency[a]		Total frequency
	Group 1	Group 2	
1. Efficient Use of Blood Components			
1.1. Inaccurate Dosage (Over-/Underdosing)			
1.1.1. Prescription: clinician-driven over- or underordering	+ +	0	Variant
1.1.2. Prescription: over- or underordering not based on individual discretion	+	+ +	Typical
1.1.3. Preparation: over-/underpreparation during ward request or blood bank confirmation	+ +	+	Typical
1.1.4. Administration: clinically unnecessary or inadequate transfusions	+	0	Variant
1.2. Inefficient Blood Usage and Processes			
1.2.1. Lack of transparency in communication or procedures during transfusion preparation	+	+ +	Typical
1.2.2. Issues in subprocess tasks for order fulfillment	+	+	Typical
1.2.3. Logistics issues in internal transport of blood products	+ +	0	Variant
1.2.4. Blood management challenges	+ +	0	Variant
1.2.5. Preparation of transfusion support devices	0	0	N/A[b]
1.2.6 Blood product wastage	0	0	N/A
1.3. Other Transfusion Process Errors			
1.3.1. Patient identity and data checks	0	0	N/A
1.3.2. Transfusion ordering	0	+	Variant
1.3.3. Order verification	0	0	N/A
1.3.4. Ward-level blood preparation	+	0	Variant
1.3.5. Blood bank unit preparation	0	0	N/A
1.3.6. Blood transport and arrival confirmation	0	0	N/A
1.3.7. Transfusion administration	+	+	Typical
1.3.8. Transfusion completion	0	0	N/A
1.3.9. Other: unforeseen clinical scenarios	+ +	+	Typical

[a]Frequency indicators reflect both sentiment and prevalence of opinions. Positive opinions are denoted by plus signs (+) and negative opinions by minus signs (–). One symbol indicates 1 participant, 2 symbols indicate 2-3 participants, and 3 symbols indicate 4 or more participants. “0” indicates that no relevant statements were identified for that core idea.

[b]N/A: not applicable (also see footnote “a”).

Category B: Perceptions and Opinions Regarding pMSBOS-TS

People

Regarding the impact of pMSBOS-TS on individual users, participants generally agreed that the system could reduce variation in blood ordering volumes attributable to the prescribing physician’s personal tendencies (Table 3).

We used to just prescribe based on our assumptions, but now this app (pMSBOS-TS) prescribes based on what it studied through machine learning. [Participant 2_5]

Reduced deviation according to the characteristics of the physicians’ prescribing tendencies (stable/adventurous). [Sticky note]

XSL•FO
RenderX

Table 3. Cross-case analysis for the 6 primary domains: People, Environment, Tools, Tasks, Work Processes, and Work Outcomes (category B).

Domain, subdomain, and core idea	Frequency[a]		Total frequency
	Group 1	Group 2	
1. People			
1.1. Prescribing Physicians			
1.1.1. Variability in prescribed volume based on the physician's experience or skillset (may decrease variation [P[b]] or increase it [N[c]])	+ +	+ +	General
1.1.2. Adaptation gap according to the physician's proficiency with the ordering system	+ +	+ +	General
1.1.3. Clinician perceptions of CDSS[d] affecting adoption of the new system	+ +	+ +	General
1.1.4. Final verification of calculated transfusion requirement and prescribing responsibility must remain with the ordering physician	0	+ + +	Variant
1.2. Other Transfusion Stakeholders			
1.2.1. Blood bank staff, nurses, and other transfusion team members gain improved demand forecasting through the system	+ +	+ + +	General
1.2.2. Other: potential to reduce nonuser-initiated over- or underordering	0	+	Variant
2. Environment			
2.1. Interprofessional Communication			
2.1.1. Clear communication between clinical staff and the blood bank is critical for successful system adoption	+ +	0	Variant
2.1.2. System may enhance transparency (P) but could introduce additional confirmation steps or confusion (N)	– – –	– –	General
2.2. Socio-Organizational Context			
2.2.1. Organization's culture around individual variation in transfusion demand: blame culture may decrease (P), or lack of clear norms may increase confusion (N)	– –	+	Variant
2.2.2. Organizational climate and policies influence uptake	+ +	+ +	General
2.2.3. Institutional blood management challenges affect implementation	+ +	+	Typical
2.2.4. Clarity of governance is critical for program sustainability	+	+ +	Typical
2.2.5. Other: variations in system impact and blood management practices expected based on health care facility size and infrastructure	0	+	Variant
2.3. Physical Environment			
2.3.1. Anticipated effective utilization in settings with high concentrations of clinical staff and resources	+	0	Variant
2.3.2. Physical environment factors are critical for successful adoption	+	0	Variant
3. Tools			
3.1. Algorithm Performance			

XSL•FO
RenderX

Domain, subdomain, and core idea	Frequency[a]		Total frequency
	Group 1	Group 2	
3.1.1. System response time is a key determinant for adoption	0	+ + +	Variant
3.1.2. Ability to incorporate a wide range of clinical input variables	+ +	+ + +	General
3.1.3. Capability to generate tailored recommendations for various blood components	+	+ + +	Typical
3.1.4. Predictive accuracy of the algorithm is essential	+ +	+ +	General
3.2. Usability and System Design			
3.2.1. Interface layout and input mechanisms must support efficient use	+ +	+ +	General
3.2.2. Workflow integration should not interrupt clinical tasks	+	+	Typical
3.2.3. Seamless electronic medical record interoperability to prepopulate patient data and eliminate manual entry	+ +	+ + +	General
3.2.4. Other: support for operation across diverse platforms	+	0	Variant
3.3. Trust in the Tool			
3.3.1. Confidence in the pMSBOS-TS[e] model is critical for sustained use	+ +	+	Typical
3.3.2. Other: difficulty establishing confidence in the algorithm due to the inherently opaque machine-learning inference process	0	+	Variant
4. Tasks			
4.1 Achieving Algorithm Objectives			
4.1.1. Impact on returns/wastage: personalized demand forecasts should reduce waste (P) or, if overestimation/mistrust occurs, increase waste (N)	+	+/– – –	Variant
4.1.2. Supply from the National Blood Service: procurement may become easier (P) or more difficult (N)	–	0	Variant
4.1.3. Other: improved blood inventory management may help reduce delays in transfusion	0	+	Variant
4.2. Task Complexity and Variability			
4.2.1. System introduction may decrease (P) or increase (N) required time and effort	+/–	++/–	Variant
4.2.2. Must accommodate complex cases	+ +	0	Variant
4.3. Handling Unanticipated Situations			
4.3.1. Flexibility to manage unforeseen variables not captured by the algorithm	+ +	+ + +	General
4.3.2. Impact from overlapping work-system issues	+ +	+ +	General
4.3.3. Limiting tool application to defined scenarios ensures safe and effective use	0	+ +	Variant
4.3.4. Other: malfunctions or unintended consequences may arise if users do not fully understand the task or intended use of the system.	+	0	Variant
5. Work Processes			
5.1. Linked Diagnostic Orders			
5.1.1. Integration of ancillary tests with the transfusion program is critical	+	+ +	Typical
5.2. Preparation of Related Procedures/Equipment			
5.2.1. System may facilitate preparation of downstream tasks (P) or, conversely, increase complexity of subsequent steps (N)	–	+ +/– –	Variant
6. Work Outcomes			
6.1. Efficiency of Blood Use and Management			

XSL•FO
RenderX

Domain, subdomain, and core idea	Frequency[a]		Total frequency
	Group 1	Group 2	
6.1.1. Contribution of pMSBOS-TS to efficient utilization and inventory control: positive or negative	+ +	+/– – –	Variant
6.1.2. Long-term impact: accumulation of usage data to further refine predictive accuracy and inform adoption strategies	+	+ +	Typical
6.1.3. Indirect clinical benefits: supports expedited detection of abnormal laboratory findings, thereby enhancing overall patient management	+	+ +	Typical
6.2. Organizational Culture and Processes for Personalized Transfusion			
6.2.1. Role of pMSBOS-TS in fostering a culture and workflow for personalized maximum transfusion prediction: positive impact or none/negative impact	– –	–	Typical

[a]Frequency indicators reflect both sentiment and prevalence of opinions. Positive opinions are denoted by plus signs (+) and negative opinions by minus signs (–). One symbol indicates 1 participant, 2 symbols indicate 2-3 participants, and 3 symbols indicate 4 or more participants. “0” indicates that no relevant statements were identified for that core idea.

[b]P: positive.

[c]N: negative.

[d]CDSS: clinical decision support system.

[e]pMSBOS-TS: personalized Maximum Surgical Blood Order Schedule—Thoracic Surgery.

However, it was also noted that the effectiveness of pMSBOS-TS would depend on the prescriber’s level of proficiency with the system and their trust in its recommendations.

> *Someone who judges based on his/her own experience and is not familiar with a new tool, may pursue the existing method.* [Sticky note]

In one group, it was emphasized that final confirmation and responsibility for blood ordering based on calculated transfusion requirements should remain with the physician. Among the broader group of health care professionals involved in the transfusion process, there was a general expectation that pMSBOS-TS would support more accurate predictions of required blood volumes.

Environment

Regarding the environmental impact of pMSBOS-TS, participants expressed concern that its implementation might lead to increased verification steps or procedural confusion during blood preparation and release.

> *There used to be a set standard of 3 or 2 (example of the number of blood packs in order set), but if it changes for each patient...like, this patient needs to prepare 8, this patient needs to prepare 5...If our nurses had to check with the prescribing doctor each time and prepare a certain number of bloods, I thought there could be confusion.* [Participant 1_5]

There was also a shared perception that the success of integration would depend heavily on the organizational culture and internal dynamics of the institution.

> *I think that the surgical nursing team may focus more on the team atmosphere when the pMSBOS-TS is implemented.* [Participant 2_3]

Opinions were divided between the 2 groups regarding whether pMSBOS-TS would positively or negatively influence the development of an organizational culture that accommodates interindividual variability in transfusion needs.

Tools

Regarding pMSBOS-TS as a tool, participants broadly agreed that the algorithm should be capable of incorporating a wider range of clinical input variables, and that its predictive performance—specifically, its accuracy—would be crucial to successful adoption.

> *Even for the same disease, the bleeding risk differs depending on the severity, and since previous abdominal surgery has a big influence on the tissue adhesion, it seems likely that more blood transfusions will be needed. I wonder to what extent this will be reflected in the algorithm.* [Participant 1_7]

With respect to usability, participants consistently emphasized that the interface should be intuitive, with screen layouts and input controls designed for ease of use. They also stressed that patient information should be automatically integrated from the EMR to minimize manual data entry by users.

> *If (the data input of pMSBOS-TS is) not linked to EHR, additional workload is possible/input error is possible.* [Sticky note]

The importance of system response speed was highlighted frequently in only 1 of the 2 groups.

Tasks

Regarding the interaction between pMSBOS-TS and task performance, participants generally agreed that the system’s ability to flexibly accommodate unexpected clinical scenarios or variables not accounted for by the algorithm would be critical.

> *Variation in the skill of the surgeon performing the surgery/The possibility of an unexpected worse situation occurring for the patient.* [Sticky note]

Task performance was also noted to be influenced when issues from other components of the broader work system intersected

XSL•FO
RenderX

with the use of pMSBOS-TS. Opinions were divided across both groups regarding the anticipated impact of pMSBOS-TS on blood product returns or waste, as well as on the complexity and performance of transfusion-related tasks.

Work Processes and Work Outcomes

No dominant themes emerged concerning the influence of pMSBOS-TS on work processes or outcomes. Opinions were mixed regarding whether the system would affect subsequent tasks or contribute meaningfully to the management and utilization of blood products.

> *The pMSBOS-TS would be helpful in assessing the appropriateness of blood transfusion (for example, for health insurance's claim eligibility review)* [Participant 1_1]
>
> *If more blood is prescribed to account for the risk of bleeding, there may be more problems with distribution, such as returns or disposal.* [Participant 2_6]

Discussion

Principal Findings

This study identified clinicians' anticipated benefits and concerns regarding pMSBOS-TS across workflow, usability, and organizational domains. Consistent with prior CDSS research, such systems can offer clear benefits when successfully implemented, enhancing patient safety and supporting clinical decision-making. Through a qualitative investigation involving frontline health care professionals, we explored anticipated impacts—both positive and negative—associated with implementing pMSBOS-TS, a prediction-based CDSS for maximum surgical blood ordering in preoperative patients.

Findings revealed several core concepts that extend beyond transfusion tasks and are broadly applicable to CDSS implementation. These include the importance of users' (ie, physicians') system proficiency and trust, the influence of organizational culture, the accuracy of task performance, and overall system usability. By engaging diverse stakeholders and applying the PETT framework, we examined the hypothesis that successful and safe CDSS implementation depends not only on algorithmic performance but also on workflow integration and interprofessional interactions.

Interpretation and Implications

These observations illustrate how pMSBOS-TS may interact with existing transfusion workflows and sociotechnical structures. Whereas traditional MSBOS approaches are often static (eg, "for procedure X, always have Y units ready"), machine learning models can personalize recommendations by incorporating patient-specific factors [19]. Given the urgent and collaborative nature of transfusion processes, participants noted concerns regarding potential unintended effects on interdisciplinary collaboration and emphasized the need for flexibility within the system's functionality. These insights suggest that, for a newly introduced CDSS to achieve its intended impact, attention must be paid not only to algorithmic performance but also to the specific characteristics of the task environment, end users, and interface design.

Task analysis has long been used to address patient safety issues and guide quality improvement initiatives in health care, including the design, implementation, and optimization of health IT systems [30,32,33]. During the development of the swim-lane diagram, we observed that transfusion workflows extend beyond a simple physician-patient interaction, involving complex coordination among physicians, nurses, and blood bank staff. Findings from category A confirmed that inefficiencies arise not only from interprofessional issues but also from interdepartmental and spatial constraints. Compared with medication administration, transfusion processes are more resource-constrained and require specialized procedures—matching, delivery, storage, and multistep verification. As a result, discrepancies in prescribing or preparation can lead to significant resource waste and additional workload. Moreover, blood banks require specialized expertise and must coordinate directly with central blood suppliers, making communication breakdowns a potential source of operational strain.

This complexity was reflected in the cross-case analysis for category A, which identified inefficient practices such as over- or underprescribing influenced by order sets or department-specific habits. These findings align with participants' views that organizational culture will play a central role in shaping the success of pMSBOS-TS implementation. Concerns regarding inconsistent blood preparation, lack of transparency, and potential confusion during rollout further underscore the importance of attention to workflow readiness. Additionally, the need for compatibility testing, highlighted in both categories A and B, emphasizes that evaluations must account for the entire sequence of related workflows, including upstream preparation and downstream follow-up.

The core question—"What positive or negative impacts could be expected if pMSBOS-TS is introduced into clinical practice?"—elicited both supportive and critical responses. Anticipated benefits included reduced prescription variability and improved accuracy of blood use predictions, aligning with the system's development goals. However, participants expressed divergent views regarding downstream tasks, such as adjusting over- or underprescriptions driven by order sets and managing transfusion-related tools. Differences also emerged regarding the expected benefits of personalized blood ordering—short term (eg, time savings and reduced blood waste) versus long term (eg, improved blood reserve management).

The most frequently cited concern was the potential for increased workload due to additional confirmation (double-checking) steps. Given that transfusions often occur in life-threatening situations, this emphasis on safety is understandable. The introduction of a confirmation mechanism is an essential safeguard, and participants' concerns reflect a broadly shared and appropriate level of caution. In addition, several participants noted that patient-specific transfusion considerations—such as previous alloimmunization, allergic reactions, or rare blood requirements—are not represented in the current algorithm and would continue to require clinical

XSL•FO
RenderX

judgment. These factors delineate important boundaries for safe use and underscore the need for guidance on when algorithmic recommendations should be supplemented by clinician expertise.

Participants also voiced skepticism regarding the impact of personalized services on clinical workflows, a sentiment commonly reported in the adoption of precision medicine tools. This resistance may partly stem from the lack of distinction between short- and long-term effects in the interview questions. As with many new technologies, early implementation often increases user workload. Thus, realizing the intended benefits of systems such as pMSBOS-TS requires not only user engagement but also effective communication with stakeholders, workflow adaptation, and careful integration into existing processes.

In response to the question, "What factors are important for the successful adaptation and implementation of pMSBOS-TS in clinical practice?," participants commonly cited human and environmental factors, such as clinician trust and organizational culture, alongside functional considerations, including input/output variables, interface usability, and EMR integration. These findings align with prior research identifying such factors as critical to successful system adoption [28,29]. Given that transfusions often occur in emergencies, such as massive intraoperative bleeding or rapid clinical deterioration, participants emphasized the need for system flexibility and attention to overlapping workflows.

These insights suggest that successful implementation requires not only optimization of algorithmic performance and user-centered interface design but also enhanced user training and a supportive organizational environment. System adoption is not linear but cyclical and interactive, involving system use, feedback, and iterative refinement. Accordingly, the relationship between algorithms and workflows should be understood as bidirectional.

Furthermore, adaptation should be evaluated in stages during system implementation. Conflicting opinions regarding expected effects—such as concerns about increased confirmation steps or confusion during blood withdrawal or preparation—reflect common apprehensions in clinical settings. This underscores the need for ongoing, structured efforts to minimize unintended consequences and ensure effective integration into routine workflows.

Although participants received preinterview materials and program demonstrations, many required further clarification, particularly regarding how personalized transfusion calculations could reduce overall prescription volume. The machine learning processes underlying these calculations were frequently described as a "black box," and comprehension remained limited even with explainable AI functions. Although participants commonly perceived the model as a black box, not all machine learning approaches lack transparency; for example, tree-based or regression-based models may provide interpretable feature contributions, though such details were beyond the scope of this preimplementation evaluation [21]. This highlights the need for further research on clinician education, comprehension, and acceptance when introducing AI-driven clinical tools.

Implementing a novel AI-driven CDSS in a complex health care environment is not merely a technical deployment—it represents a sociotechnical change that reverberates through clinicians' routines, tasks, and organizational structures. A SEIPS-based evaluation captures this full spectrum of impacts by examining PETT across each stage of the workflow, enabling identification of both anticipated benefits and unanticipated drawbacks [23]. Although other analytic approaches provide useful insights, SEIPS offers a unifying systems perspective that is particularly valuable for complex, high-risk interventions such as CDSSs in clinical care. We also emphasize the importance of balancing approaches by supplementing SEIPS findings with quantitative measures from prior studies and by ensuring practical applicability through the simplified SEIPS 101 model, which facilitates stakeholder engagement [19,23].

Strengths and Limitations

This study has several limitations. It was conducted in a single country and institution, and our FGD findings—derived from the SEIPS framework and task analysis—are qualitative and context-specific, which may limit replicability. In addition, the SEIPS framework has been criticized for its complexity and limited applicability to macro-level evaluation [23]. To mitigate these concerns, we performed content validation with medical professionals from external institutions who were not involved in the interviews. Given the characteristics of tertiary hospitals, some findings may not generalize to smaller institutions with fewer personnel dedicated to transfusion management. Nevertheless, because regular transfusion-related surgeries are less common in smaller hospitals and transfusion practices involve multiple specialties, future work should include a broader range of settings and professional roles.

As a result of the FGD-based task analysis and qualitative design, recall bias and incomplete reporting are possible. Follow-up research (eg, system usage log analyses) could address these limitations; however, such data were not available for this preimplementation evaluation of a new CDSS [34]. This may represent a common constraint in predeployment CDSS impact or risk assessments. Future studies should complement qualitative findings with quantitative, log-based evaluations as these data become available.

As we targeted actively practicing medical professionals, the physicians in our sample had shorter career durations than other groups. This reflects the staffing structure of the clinical environment and is unlikely to affect the analysis of practical workflows. However, future research should examine how professional seniority may influence perceptions related to organizational decision-making and AI governance.

Given the specific nature of transfusion workflows, some requirements identified in this study may not generalize to other AI applications. Moreover, to accurately assess the impact of new systems, future evaluations should differentiate expected effects across stages of system adoption and adaptation. This limitation became apparent during our analysis and warrants attention in subsequent research. Although successful AI deployment depends heavily on the human-system interface, current literature lacks theory-driven qualitative evaluations examining how AI fits within complex sociotechnical systems.

XSL•FO
RenderX

We addressed these challenges by engaging diverse stakeholders, applying a structured sociotechnical framework, and adhering to rigorous qualitative research methods.

Overall Contribution

This study contributes to CDSS implementation research by providing one of the few preimplementation evaluations of an AI-based transfusion decision support tool. Using CQR and the SEIPS 101 framework, we captured practical concerns and anticipated impacts across physicians, nurses, and blood bank staff, offering concise, context-specific insights that complement findings from prior postdeployment CDSS studies.

Conclusions

As systems and workflows interact dynamically, it is essential to consider both tool performance and contextual adaptation. Future research should segment the adaptation process and examine interactions among users, their roles, and organizational dynamics.

Acknowledgments

This study was funded by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, supported by the Ministry of Health & Welfare, Republic of Korea (grant RS-2022-KH130250). The funder had no role in the study design, data collection, data analysis or interpretation, or manuscript preparation.

Data Sharing

The datasets generated and analyzed during this study are not publicly available due to the lack of prior consideration for data sharing during study design and Institutional Review Board approval. However, data are available from the corresponding author (YL) upon reasonable request and with approval from the Asan Medical Center Institutional Data Access/Ethics Committee.

Authors' Contributions

Conceptualization: YL, MO
Data curation: YEP
Formal analysis: YEP
Funding acquisition: YL
Investigation: YEP, YL
Methodology: YL, MO
Project administration: YL
Resources: DHK, HJL
Supervision: YL
Validation: YL, DHK, HJL, JHL, TP, JY
Visualization: YL, YEP
Writing – original draft: YEP, MO
Writing – review & editing: YL, MO
All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Demographic and professional characteristics of focus group discussion participants.
[DOCX File , 22 KB - jmir_v28i1e86166_app1.docx]

Multimedia Appendix 2
Integrated workflow for preoperative blood ordering, preparation, and transfusion in surgical care settings.
[DOCX File , 401 KB - jmir_v28i1e86166_app2.docx]

Multimedia Appendix 3
Analytical framework, including domains, subdomains, and core ideas.
[DOCX File , 26 KB - jmir_v28i1e86166_app3.docx]

References

XSL•FO
RenderX

1. Liyanage H, Liaw S, Jonnagaddala J, Schreiber R, Kuziemsky C, Terry AL, et al. Artificial intelligence in primary health care: perceptions, issues, and challenges. Yearb Med Inform 2019 Aug 25;28(1):41-46 [FREE Full text] [doi: 10.1055/s-0039-1677901] [Medline: 31022751]
2. Chen P, Lin C, Wu W. Big data management in healthcare: adoption challenges and implications. International Journal of Information Management 2020 Aug;53:102078 [FREE Full text] [doi: 10.1016/j.ijinfomgt.2020.102078]
3. Staras S, Tauscher JS, Rich N, Samarah E, Thompson LA, Vinson MM, et al. Using a clinical workflow analysis to enhance eHealth implementation planning: tutorial and case study. JMIR Mhealth Uhealth 2021 Mar 31;9(3):e18534 [FREE Full text] [doi: 10.2196/18534] [Medline: 33626016]
4. Mohan K, Ahlemann F. Understanding acceptance of information system development and management methodologies by actual users: A review and assessment of existing literature. International Journal of Information Management 2013 Oct;33(5):831-839 [FREE Full text] [doi: 10.1016/j.ijinfomgt.2013.06.003]
5. Salwei ME, Carayon P, Hoonakker PLT, Hundt AS, Wiegmann D, Pulia M, et al. Workflow integration analysis of a human factors-based clinical decision support in the emergency department. Appl Ergon 2021 Nov 01;97(4):103498-103497 [FREE Full text] [doi: 10.1016/j.apergo.2021.103498] [Medline: 34182430]
6. Paradis KC, Naheedy KW, Matuszak MM, Kashani R, Burger P, Moran JM. The fusion of incident learning and failure mode and effects analysis for data-driven patient safety improvements. Pract Radiat Oncol 2021 Jan;11(1):e106-e113. [doi: 10.1016/j.prro.2020.02.015] [Medline: 32201319]
7. Baartmans MC, Van Schoten SM, Wagner C. Generic analysis method to learn from serious adverse events in Dutch hospitals: a human factors perspective. BMJ Open Qual 2022 Feb 01;11(1):e001637 [FREE Full text] [doi: 10.1136/bmjoq-2021-001637] [Medline: 35105550]
8. Gray T, Antolak A, Ahmed S, Magnelli A, Lu L, Cho Y, et al. Implementing failure mode and effect analysis to improve the safety of volumetric modulated arc therapy for total body irradiation. Med Phys 2023 Jul 02;50(7):4092-4104. [doi: 10.1002/mp.16466] [Medline: 37265031]
9. Charles D, King J, Patel V, Furukawa M. Adoption of electronic health record systems among U.S. non-federal acute care hospitals. In: ASTP Health IT Data Brief. Washington, DC: Office of the National Coordinator for Health Inforamtion Technology; 2013.
10. Woldemariam MT, Jimma W. Adoption of electronic health record systems to enhance the quality of healthcare in low-income countries: a systematic review. BMJ Health Care Inform 2023 Jun 12;30(1):e100704 [FREE Full text] [doi: 10.1136/bmjhci-2022-100704] [Medline: 37308185]
11. Office of the National Coordinator for Health IT. Health IT legislation. Health IT. URL: https://www.healthit.gov/topic/laws-regulation-and-policy/health-it-legislation [accessed 2024-09-05]
12. Vos JFJ, Boonstra A, Kooistra A, Seelen M, van Offenbeek M. The influence of electronic health record use on collaboration among medical specialties. BMC Health Serv Res 2020 Jul 22;20(1):676 [FREE Full text] [doi: 10.1186/s12913-020-05542-6] [Medline: 32698807]
13. National trends in hospital and physician adoption of electronic health records. Health IT. URL: https://www.healthit.gov/data/quickstats/national-trends-hospital-and-physician-adoption-electronic-health-records [accessed 2024-09-05]
14. Stewart C. Global clinical decision support systems (CDSS) market in 2018, and a forecast for 2028. Statista. 2021. URL: https://www.statista.com/statistics/871216/cdss-market-value-worldwide/?srsltid=AfmBOooY6zdqBtqWh9YjFwuaTFYY9MfnYNYXbBgOCAfxvFQ7KKIqKHqT [accessed 2024-09-05]
15. Amjad A, Kordel P, Fernandes G. A review on innovation in healthcare sector (telehealth) through artificial intelligence. Sustainability 2023 Apr 14;15(8):6655 [FREE Full text] [doi: 10.3390/su15086655] [Medline: 26701262]
16. Hak F, Guimarães T, Santos M. Towards effective clinical decision support systems: a systematic review. PLoS One 2022 Aug 15;17(8):e0272846 [FREE Full text] [doi: 10.1371/journal.pone.0272846] [Medline: 35969526]
17. Liberati EG, Ruggiero F, Galuppo L, Gorli M, González-Lorenzo M, Maraldi M, et al. What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. Implement Sci 2017 Sep 15;12(1):113 [FREE Full text] [doi: 10.1186/s13012-017-0644-2] [Medline: 28915822]
18. Pyo J, Lee W, Choi EY, Jang SG, Ock M. Qualitative research in healthcare: necessity and characteristics. J Prev Med Public Health 2023 Jan;56(1):12-20 [FREE Full text] [doi: 10.3961/jpmph.22.451] [Medline: 36746418]
19. Hur S, Yoo J, Min JY, Jeon YJ, Cho JH, Seo JY, et al. Development, validation, and usability evaluation of machine learning algorithms for predicting personalized red blood cell demand among thoracic surgery patients. Int J Med Inform 2024 Nov;191:105543 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105543] [Medline: 39084087]
20. Newton N, Bamgboje-Ayodele A, Forsyth R, Tariq A, Baysari MT. A systematic review of clinicians' acceptance and use of clinical decision support systems over time. NPJ Digit Med 2025 May 26;8(1):309 [FREE Full text] [doi: 10.1038/s41746-025-01662-7] [Medline: 40419669]
21. Hur S, Lee Y, Park J, Jeon YJ, Cho JH, Cho D, et al. Comparison of SHAP and clinician friendly explanations reveals effects on clinical decision behaviour. NPJ Digit Med 2025 Sep 26;8(1):578 [FREE Full text] [doi: 10.1038/s41746-025-01958-8] [Medline: 41006498]

XSL•FO
RenderX

22. Carayon P, Schoofs Hundt A, Karsh B, Gurses AP, Alvarado CJ, Smith M, et al. Work system design for patient safety: the SEIPS model. Qual Saf Health Care 2006 Dec;15 Suppl 1(Suppl 1):i50-i58 [FREE Full text] [doi: 10.1136/qshc.2005.015842] [Medline: 17142610]
23. Holden RJ, Carayon P. SEIPS 101 and seven simple SEIPS tools. BMJ Qual Saf 2021 Nov 26;30(11):901-910 [FREE Full text] [doi: 10.1136/bmjqs-2020-012538] [Medline: 34039748]
24. Cho I. Frameworks for evaluating the impact of safety technology use. Healthc Inform Res 2023 Apr;29(2):89-92 [FREE Full text] [doi: 10.4258/hir.2023.29.2.89] [Medline: 37190732]
25. Dubé M, Hron JD, Biesbroek S, Chan-MacRae M, Shearer A, Landi R, et al. Human factors and systems simulation methods to optimize peri-operative EHR design and implementation. Adv Simul (Lond) 2025 Apr 23;10(1):23 [FREE Full text] [doi: 10.1186/s41077-025-00349-z] [Medline: 40269997]
26. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
27. Im D, Pyo J, Lee H, Jung H, Ock M. Qualitative research in healthcare: data analysis. J Prev Med Public Health 2023 Mar;56(2):100-110 [FREE Full text] [doi: 10.3961/jpmph.22.471] [Medline: 37055353]
28. Laka M, Milazzo A, Merlin T. Factors that impact the adoption of clinical decision support systems (CDSS) for antibiotic management. Int J Environ Res Public Health 2021 Feb 16;18(4):1901 [FREE Full text] [doi: 10.3390/ijerph18041901] [Medline: 33669353]
29. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
30. Saitwal H, Feng X, Walji M, Patel V, Zhang J. Assessing performance of an electronic health record (EHR) using cognitive task analysis. Int J Med Inform 2010 Jul;79(7):501-506. [doi: 10.1016/j.ijmedinf.2010.04.001] [Medline: 20452274]
31. Singprasong R, Eldabi T. An integrated methodology for process improvement and delivery system visualization at a multidisciplinary cancer center. J Healthc Qual 2013;35(2):24-32. [doi: 10.1111/j.1945-1474.2011.00174.x] [Medline: 22092497]
32. Russ AL, Militello LG, Glassman PA, Arthur KJ, Zillich AJ, Weiner M. Adapting cognitive task analysis to investigate clinical decision making and medication safety incidents. J Patient Saf 2017 May 3;15(3):191-197. [doi: 10.1097/pts.0000000000000324]
33. Graham LA, Gray C, Wagner TH, Illarmo S, Hawn MT, Wren SM, et al. Applying cognitive task analysis to health services research. Health Serv Res 2023 Apr 09;58(2):415-422 [FREE Full text] [doi: 10.1111/1475-6773.14106] [Medline: 36421922]
34. Zheng K, Ratwani RM, Adler-Milstein J. Studying workflow and workarounds in electronic health record–supported work to improve health system performance. Annals of Internal Medicine 2020 Jun 02;172(11_Supplement):S116-S122. [doi: 10.7326/m19-0871]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
COREQ: Consolidated Criteria for Reporting Qualitative Research
CQR: consensual qualitative research
EHR: electronic health record
EMR: electronic medical record
FGD: focus group discussion
IRB: institutional review board
IT: information technology
PETT: People, Environment, Tools, and Tasks
pMSBOS-TS: personalized Maximum Surgical Blood Order Schedule—Thoracic Surgery
SEIPS: Systems Engineering Initiative for Patient Safety

XSL•FO
RenderX

Edited by A Stone; submitted 20.Oct.2025; peer-reviewed by S Waldvogel Abramowski, CJ Engstrom; comments to author 20.Nov.2025; revised version received 22.Dec.2025; accepted 24.Dec.2025; published 23.Jan.2026.

Please cite as:
Park YE, Ock M, Lee JH, Ko DH, Lee HJ, Park T, Yoo J, Lee Y
Assessing Health Care Professionals' Perceptions of a New System in Clinical Workflows: Systems Engineering Initiative for Patient Safety–Based Consensual Qualitative Research
J Med Internet Res 2026;28:e86166
URL: https://www.jmir.org/2026/1/e86166
doi:10.2196/86166
PMID:

©Ye-Eun Park, Minsu Ock, Jae-Ho Lee, Dae-Hyun Ko, Hak-Jae Lee, Taezoon Park, Junsang Yoo, Yura Lee. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

Machine Learning for Predicting Stroke Risk Stratification Using Multiomics Data: Systematic Review

Hae Young Yoo[1], RN, PhD; Hyerim Shin[1], RN, BSN; Eun-Jung Kim[1], RN, MSN; Youn-Jung Son[1], RN, PhD, CCAPN, FAAN

Chung-Ang University, Seoul, Republic of Korea

Corresponding Author:
Youn-Jung Son, RN, PhD, CCAPN, FAAN
Chung-Ang University
84 Heukseok-roDongjak District, Red Cross College of Nursing
Seoul, 06974
Republic of Korea
Phone: 82 2 820 5198
Fax: 82 2 824 7961
Email: yjson@cau.ac.kr

Abstract

Background: Stroke is a complex, multidimensional disorder influenced by interacting inflammatory, immune, coagulation, endothelial, and metabolic pathways. Single-omics approaches seldom capture this complexity, whereas multiomics techniques provide complementary insights but generate high-dimensional and correlated feature spaces. Machine learning (ML) offers strategies to manage these challenges; however, the predictive accuracy and reproducibility of multiomics-based ML models for stroke remain poorly characterized.

Objective: This review aimed to conduct a systematic evaluation of ML models using multiomics data for stroke risk stratification and comprehensive patterns in discriminatory performance, integration strategies, and validation and reporting practices to inform future methodological development.

Methods: We conducted a comprehensive literature search following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 recommendations. Studies published from January 2000 to July 2025 were identified across 9 databases, including PubMed, MEDLINE Ultimate, EMBASE, CINAHL, Web of Science, Scopus, Cochrane CENTRAL, ACM Digital Library, and IEEE Xplore. Eligible studies included adults with ischemic, hemorrhagic, or unspecified stroke as the prediction target; applied at least 2 omics layers; and reported ML performance metrics. Risk of bias was assessed using the Prediction Model Risk of Bias Assessment Tool, while reporting quality was evaluated using Minimum Information for Medical AI Reporting. The primary outcome was the area under the receiver operating characteristic curve.

Results: A total of 7 studies (n=40,274) published between 2022 and 2025 fulfilled the inclusion criteria. All studies combined 2 omics layers, most often using middle-level integration with dyads such as metabolomics-proteomics and metabolomics-lipidomics. Supervised ML algorithms across studies included support vector machines, tree-based ensembles, generalized linear models, and deep learning architectures. Three studies reported external validation of the integrated multiomics model, while 1 study conducted only an external assessment of a single marker rather than validation of the integrated model. Three studies reported an assessment of calibration, and clinically prespecified operating points were rarely described. Reported areas under the receiver operating characteristic curve varied by prediction task, ranging from 0.75 to 0.96 for acute diagnosis models and from 0.75 to 0.97 for onset risk prediction models; the highest externally validated performance was achieved by a support vector machine trained on a metabolomics-proteomics dyad in mixed stroke types (ischemic and hemorrhagic).

Conclusions: Multiomics ML models showed high apparent discrimination for stroke risk stratification, but current evidence remains methodologically limited. Small sample sizes, heterogeneous designs, and incomplete reporting currently hinder the reproducibility and generalizability of multiomics ML models for stroke risk prediction. To advance the field, future studies should adopt leakage-resistant evaluation frameworks, conduct site-specific external validations, and benchmark against both single-omics and clinical baselines to demonstrate incremental value. Well-designed, transparently reported investigations will be essential to move multiomics ML models from exploratory promise toward clinically actionable tools in precision stroke care.

(J Med Internet Res 2026;28:e85654) doi:10.2196/85654

XSL•FO
RenderX

KEYWORDS

deep learning; epigenomics; genomics; lipidomics; machine learning; metabolomics; ML; multiomics; proteomics; risk stratification; stroke; transcriptomics

Introduction

Stroke is defined by acute neurological deficits resulting from vascular brain injury [1,2]. Globally, this condition is responsible for about 6.6 million deaths annually, making it the second leading cause of mortality. Its contribution to disability-adjusted life years is anticipated to rise considerably by 2050 [3]. As stroke is associated with long-term functional impairment, cognitive decline, and a high risk of recurrence, early risk stratification remains critical [3-6]. Multiple layers of interacting risk factors shape the occurrence and progression of stroke [7]. The prediction of stroke occurrence and clinical course is influenced by a number of factors at the molecular level, including genetic predisposition, epigenetic alterations, chronic inflammatory responses, and metabolic abnormalities [8-13]. Although this multilayered pathophysiology helps explain heterogeneity in etiological classification, its complexity makes it difficult to capture fully using only clinical variables or a single biomarker [13-17].

Recently, multiomics approaches have been used to characterize cross-layer relationships—genetic, epigenetic, transcriptomic, proteomic, metabolomic, and lipidomic—within integrated analytic frameworks [13,15,18,19]. This reflects the inherently multilayered biology of stroke, in which inflammatory, immune, coagulation, and metabolic axes operate concurrently [17,20]. Prior evidence indicates that multiomics strategies may outperform single-omics approaches in clinical applications, including stroke subtype classification, risk prediction, and acute diagnosis [19,21]. Simultaneously, machine learning (ML) has proven particularly effective for analyzing multiomics data, which are marked by high dimensionality and nonlinearity [22,23]. ML enables the flexible integration of clinical variables with omics layers, supports reproducible analytic pipelines, and strengthens the performance and generalizability of prediction models [23,24]. Accordingly, there is increasing interest in developing precision ML models that leverage multidimensional biological data.

Although studies applying multiomics and ML to stroke have increased rapidly, substantial heterogeneity is evident across investigations in terms of the types of omics incorporated, integration strategies, prediction end points, predictor variables, evaluation metrics, disease spectra, validation methods, and sample sizes [25,26]. This variability limits cross-model comparability and complicates assessments of clinical applicability [27]. Nevertheless, comprehensive comparative and quantitative evaluations of the relative effectiveness of these approaches for stroke prediction remain scarce, and no systematic review synthesizing this evidence has been published to date.

This study aimed to perform a systematic analysis of ML-based investigations using multiomics data for stroke risk stratification and identify comprehensive performance patterns across the included studies to inform future methodological development.

Methods

Study Design

This review was conducted and reported in line with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 recommendations (Multimedia Appendix 1) [28]. Prior to conducting the review, the study protocol was prospectively registered in the PROSPERO (International Prospective Register of Systematic Reviews; registration number CRD420251089823).

Data Sources and Search Strategy

A systematic search was conducted in PubMed, MEDLINE Ultimate (EBSCOhost), EMBASE, CINAHL, Web of Science, Scopus, Cochrane CENTRAL, ACM Digital Library, and IEEE Xplore for records published between January 1, 2000, and July 31, 2025. The start date was chosen a priori to capture early applications of omics-based prediction and to avoid arbitrarily excluding potentially relevant studies at the interface of omics and ML. Because multiomics and ML terms are inconsistently indexed across databases, we deliberately adopted a highly sensitive search strategy using multiple overlapping biomedical and multidisciplinary databases. Search strategies combined controlled vocabulary and free-text terms related to stroke, omics, and ML. Controlled vocabularies included Medical Subject Headings, Emtree, and CINAHL Headings; database-specific syntax and Boolean operators were adapted to each platform. Full search details for all databases are available in Multimedia Appendix 2. Gray literature—major preprint servers, including medRxiv, bioRxiv, and Google Scholar—were searched to check for any potentially missed or in-press studies; however, no additional records meeting the inclusion criteria were identified (final search date: July 31, 2025).

Inclusion and Exclusion Criteria

Study eligibility was determined through the PICO (participants, interventions, comparisons, and outcomes) framework and is summarized in Table 1. Studies were included if they (1) enrolled adults (aged ≥18 years) with stroke or at risk of stroke (ischemic, hemorrhagic, or unspecified), (2) used multiomics data to develop ML models, and (3) reported model performance for at least 1 stroke-related prediction outcome. In this review, stroke-related outcomes included incident stroke risk, acute stroke diagnosis among patients with suspected stroke, and etiological subtype classification. In addition, we adopted a pragmatic, prediction-oriented definition of ML. Eligible models included supervised algorithms that used multiomics data as input to predict the prespecified outcomes. Studies were excluded if they targeted combined conditions rather than stroke alone, did not aim to predict stroke, substituted nonomics data for molecular omics, applied only omics or only ML without integration, or evaluated treatment effects. Conventional regression analyses conducted exclusively for explanatory or associative purposes without a specified predictive task were

XSL•FO
RenderX

excluded, as they were not considered ML. The publication period (January 1, 2000, to July 31, 2025) was chosen to capture the evolution of multiomics- and ML-based stroke risk prediction from its early emergence to the most recent developments.

Table 1. Eligibility criteria for the systematic review using the PICO (participant, intervention, comparison, and outcomes) framework.

Item	Inclusion criteria	Exclusion criteria
Participants	Studies enrolling adults (aged ≥18 years) with stroke or at risk of stroke (ischemic, hemorrhagic, or unspecified)	• Studies involving children or adolescents (aged <18 years) • Studies restricted to demographic-only strata (eg, only men) • Studies focusing on combined conditions (eg, stroke with dementia or atrial fibrillation) rather than stroke alone
Intervention	Studies using multiomics data (genomics, transcriptomics, proteomics, metabolomics, etc) to develop ML[a] models for stroke prediction	• Studies that do not aim to predict stroke • Studies applying only omics or only ML, without integration of both • Studies evaluating treatment effects (eg, drugs, procedures, or interventions) • Studies using nonomics data (eg, imaging data such as CTb, MRIc, EEGd, or ultrasound) instead of molecular omics for stroke prediction
Comparison	No restrictions based on comparison	• No restrictions based on comparison
Outcomes	Studies reporting prediction performance of ML models for stroke-related outcomes	• Studies that do not report prediction performance
Language	No restrictions based on language	• No restrictions based on language
Publication period	Studies published from January 1, 2000, to July 31, 2025	• Studies published before January 1, 2000
Study designs	Original research articles, including observational studies, retrospective or prospective cohort studies, secondary data analyses, or basic science studies	• Reviews, meta-analyses, protocols, editorials, or nonhuman studies • Feasibility-only pilot studies without full model development or performance reporting
Publication types	Full published original articles in peer-reviewed journals	• Abstract-only records, letters, withdrawn papers, or documents without full text

[a]ML: machine learning.

[b]CT: computed tomography.

[c]MRI: magnetic resonance imaging.

[d]EEG: electroencephalogram.

Study Selection, Data Extraction, and Synthesis

Records retrieved from all sources were imported into EndNote 21 (Clarivate Plc) for deduplication and screening management, with tracking and extraction maintained in Microsoft Excel 2021 (Microsoft Corp). Titles and abstracts were independently screened in duplicate by 2 reviewers (HYY and HS) based on predefined eligibility criteria. Full texts of potentially eligible reports and risk of bias were also assessed independently in duplicate by 2 reviewers (HYY and HS). Disagreements between reviewers were resolved through discussion and, when necessary, adjudicated by a third reviewer (Y-JS). Data were extracted independently by 2 authors (HYY and HS) using a standardized piloted form and cross-verified by a third author (Y-JS). Extracted items included study title, authors, publication year, country, multiomics studied, sample size, data source, omics type, stroke type, prediction task, omics analytic methods, ML algorithms, outcome metrics and summary results, and validation details (internal/external and strategy). Model performance metrics were extracted and summarized descriptively. The primary measure of discrimination was the area under the receiver operating characteristic curve (AUC) with 95% CIs.

When discrimination was reported using Harrell C statistic for a binary outcome rather than a receiver operating characteristic (ROC)–based AUC, we treated the C statistics as numerically equivalent to the AUC for the purpose of pooling [29,30]. Publication bias is typically assessed when at least 10 studies are available; because fewer studies met the inclusion criteria in this review, publication bias was not formally evaluated, and no funnel plots were generated [31].

A meta-analysis was not performed because the limited number of eligible studies and the substantial between-study heterogeneity precluded meaningful quantitative synthesis. The included studies varied markedly in study design (retrospective vs prospective), sample size, and data sources (single-center cohorts, community-based datasets, and biobank data). In addition, the prediction tasks differed considerably, encompassing stroke onset risk prediction, acute diagnosis, and subtype classification. Given this conceptual, methodological, and statistical heterogeneity, quantitative pooling of effect estimates was judged inappropriate. Instead, we conducted a structured narrative synthesis to systematically summarize study characteristics, omics layers evaluated, ML approaches applied, and clinical prediction targets.

XSL•FO
RenderX

Quality and Risk of Bias Assessment

Quality appraisal and risk of bias assessment were conducted using 2 frameworks suited to prediction modeling and medical artificial intelligence (AI) reporting: the Prediction Model Risk of Bias Assessment Tool (PROBAST) [32], and the Minimum Information for Medical AI Reporting (MINIMAR) [33]. PROBAST was used for the primary risk of bias classification, and was applied across the participants, predictors, outcomes, and analysis domains, with judgments at both the domain and overall levels categorized as low risk, high risk, or unclear [32]. MINIMAR was adopted as the checklist for AI and ML studies, as it represents an expert-developed framework suited to evaluating reporting quality in omics-based ML research, particularly in early translational health care contexts [33,34]. Each item was coded Y (adequate), P (partial), N (not reported), or X (not applicable), and compliance (%) was calculated as the proportion of Y among applicable items.

Ethical Considerations

The institutional review board of Chung-Ang University (number 1041078-20250715-HR-225) approved the study protocol.

Results

Literature Search Results

The database search across 9 sources yielded 2960 records. Of these, 881 duplicates were excluded, leaving 2079 records to be screened. During title and abstract screening, 1413 records were excluded. Full texts were obtained for 666 reports, of which 530 were excluded as off-topic following full-text assessment.

The remaining 136 studies underwent full eligibility evaluation. Of these, 129 were excluded for the following reasons: insufficient sample size (<30; n=5), not a stroke prediction study (n=25), not multiomics (n=70), outcome outside scope (n=6), full text unavailable (n=1), and failure to meet inclusion criteria on full-text review (n=22). A total of 7 studies were identified as meeting the inclusion criteria and were subsequently analyzed in this review (Figure 1) [35-41]. Gray literature searches yielded no additional records; therefore, the PRISMA flow diagram reflects only the results from database searches.

XSL•FO
RenderX

Figure 1. PRISMA flow diagram of study selection for systematic review of multiomics-based machine learning models for stroke prediction.

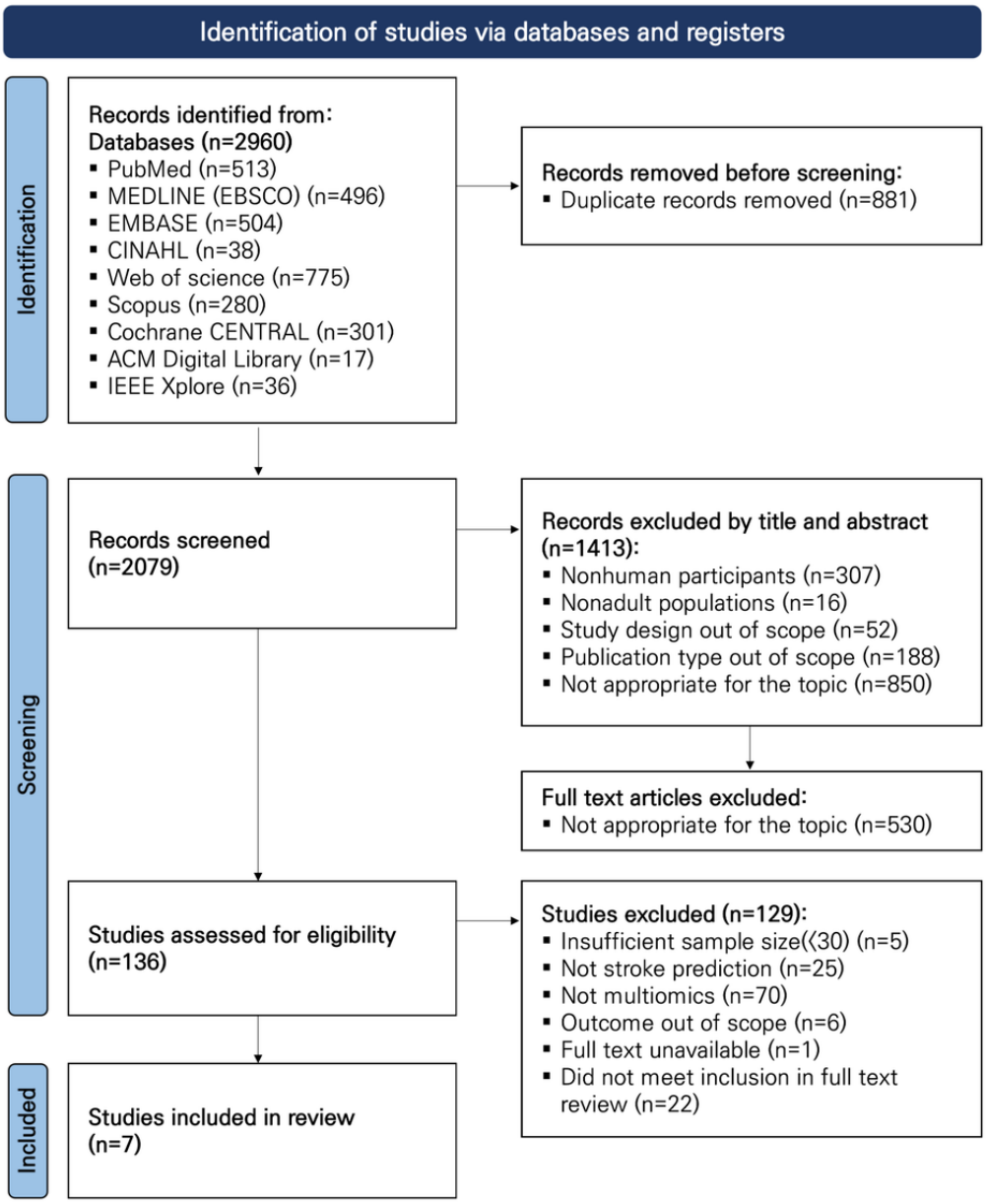

Quality Assessment of the Included Studies

Study appraisals are summarized in Multimedia Appendix 3. On PROBAST, 4 studies were rated as low risk of bias [36,37,40,41], 2 studies as high risk [38,39], and 1 study as unclear overall [35]. Most concerns were concentrated in the analysis domain, whereas participants, predictors, and outcomes were generally judged to be at low risk. Applicability issues were minimal across the included studies. On MINIMAR, overall reporting compliance ranged from 11 of 21 items [35] to 17 of 21 items [41], with a median value of 14 of 21. Information on demographic characteristics (race, ethnicity, and socioeconomic status), intended users, data missingness, external validation, and data transparency was often insufficiently reported or entirely absent.

Characteristics of the Included Studies

A total of 7 studies published between 2022 [38] and 2025 [36,41] were included in this review (Table 2). The studies represented 3 countries: China (n=4) [36-38,40], Spain (n=1) [39], and the United Kingdom (n=1; from England, Scotland, and Wales) [41]. Studies that used public repositories were classified as multinational [35,36]. Data sources were grouped into 3 categories: clinical cohorts recruited from hospitals or community health service centers (n=4) [37-40], public datasets (n=2) [35,41], and mixed sources (n=1) [36]. In the mixed-sources study [36], models were developed using gene expression omnibus datasets [42] and externally validated in independently recruited local clinical cohorts. Study designs were predominantly prospective (n=4) [37-39,41], with retrospective (n=2) [35,36], and mixed retrospective-prospective (n=1) [40]. Sample sizes varied widely, ranging from 30

XSL•FO
RenderX

participants in a hospital cohort [39] to 38,060 participants in a large public dataset [41].

Reported ML approaches included random forest, support vector machine (SVM), gradient boosting (extreme gradient boosting; adaptive boosting), generalized linear and logistic models, Bayesian logistic regression, the least absolute shrinkage and selection operator (LASSO), including the LASSO-penalized Cox proportional hazards model, and deep learning architectures such as the biological multilayer graph neural network. All studies integrated 2 omics layers, and the most frequent dyads were lipidomics-metabolomics (n=2) [37,38].

Regarding stroke type, 6 studies [35-39,41] specifically investigated ischemic stroke, whereas 1 study [40] included all stroke types. Prediction targets were grouped into categories: onset risk prediction (n=3) [35,40,41], acute diagnosis (n=2) [37,38], subtype classification (n=1) [39], and both reporting of diagnosis and subtype classification (n=1) [36].

Table 2. Characteristics of the included studies (N=7) of multiomics-based machine learning models for stroke prediction. The table summarizes key design features of each included study. Sample size (n) refers to the development cohort size for each study.

Study (year)	Country	Data source	Study design	Sample size, n	Machine learning algorithms	Omics types	Prediction task	
							Stroke type	Target
Liu et al [35] (2024)	Multinational	IEU[a] and GEO[b] database	Retrospective	103	RF[c], SVM[d], XGB[e], and GLM[f]	Genomics and transcriptomics	Ischemic stroke	Onset risk prediction
Chen et al [36] (2025)	Multinational and China	GEO database, 2022-2023 single hospital	Retrospective	1785	RF and LASSO[g]	Transcriptomics and proteomics	Ischemic stroke	Acute diagnosis; subtype classification
Zhao et al [37] (2023)	China	2017-2019 CCPTP[h]	Prospective	90	RF and BLR[i]	Lipidomics and metabolomics	Ischemic stroke	Acute diagnosis
Ye et al [38] (2022)	China	Single hospital	Prospective	106	RF, SVM, and LR[j]	Lipidomics and metabolomics	Ischemic stroke	Acute diagnosis
Labarga et al [39] (2024)	Spain	2015-2016 single hospital	Prospective	30	BioMGNN[k] and XGB	Transcriptomics and epigenomics	Ischemic stroke	Subtype classification
Zeng et al [40] (2024)	China	2010-2011 community health service centers, 2021-2023 single hospital	Prospective and retrospective	100	XGB, SVM, GNB[l], RF, AdaBoost[m], and L1-regularized LR	Metabolomics and proteomics	Stroke	Onset risk prediction
Gan et al [41] (2025)	England, Scotland, and Wales	UKB[n]	Prospective	38,060	LASSO-Cox[o]	Proteomics and genomics	Ischemic stroke	Onset risk prediction

[a]IEU: integrative epidemiology unit.

[b]GEO: gene expression omnibus.

[c]RF: random forest.

[d]SVM: support vector machine.

[e]XGB: extreme gradient boosting.

[f]GLM: generalized linear model.

[g]LASSO: least absolute shrinkage and selection operator regression.

[h]CCPTP: community comprehensive prevention and treatment project.

[i]BLR: binary logistic regression.

[j]LR: logistic regression.

[k]BioMGNN: biological multilayer graph neural network.

[l]GNB: Gaussian naïve bayes.

[m]AdaBoost: adaptive boosting.

[n]UKB: UK biobank.

[o]LASSO-Cox: least absolute shrinkage and selection operator-penalized Cox proportional hazards model.

Omics Features and Integration Strategies in Multiomics Stroke Modeling

Table 3 summarizes the characteristics of omics data, their analytic roles, and the integration strategies applied in each study. Omics type lists all molecular layers included per study, and biospecimen records the biological source of each layer as reported. The analysis unit indicates the level of the final features entered into the model.

XSL•FO
RenderX

Table 3. Omics features and integration strategies for multiomics stroke modeling in the included studies (n=7).

Study (year) and omics data type	Biospecimen	Analysis unit	Top-ranked feature biomarkers	Role in pipeline	Feature selection method	Integration strategy
Liu et al [35] (2024)						Middle integration
Genomics	GWAS[a] summary data	SNP[b] to gene-level	*FURIN (OMIM 136950), TOMM40 (OMIM 608061), HDDC3 (NCBI Gene 374659), ALDH2 (OMIM 100650), MAN2A2 (OMIM 600988)*	FS[c], FP[d], RS[e]	LD[f] clumping with genome-wide significance and F-stat	
Transcriptomics	Whole-blood mRNA[g]	Gene-level expression	*FURIN, TOMM40, HDDC3, ALDH2, MAN2A2*	FS, FP, VAL[h]	DEAGs[i]	
Chen et al [36] (2025)						Late integration
Transcriptomics	Serum miRNA[j], whole-blood mRNA	miRNA expression, gene-level mRNA expression	hsa-miR-3646, hsa-miR-4669, hsa-miR-4721, hsa-miR-504-3p, hsa-miR-8085, hsa-miR-3064-5p, hsa-miR-6736-5p, hsa-miR-6835-5p, hsa-miR-6793-5p	FS, FP	DEMs[k], ML[l] selection	
Proteomics	Plasma protein	Protein level	ITGAM[m]	FS, VAL	Cross-omics prioritization, targeted validation	
Zhao et al [37] (2023)						Middle integration
Lipidomics	Serum	LC-MS[n] untargeted metabolite features	PE[o] (18:0p/18:2), PE (16:0e/20:4), OAHFA[p] (36:3), PE (16:0p/20:3), PE (18:1p/18:2)	FS, FP, VAL	PLS-DA[q], ML selection	
Metabolomics	Serum	LC-MS untargeted lipid species	4-hydroxyphenylpyruvic acid, cafestol	FS, FP, VAL	PLS-DA, ML selection	
Ye et al [38] (2022)						Middle integration
Lipidomics	Serum	LC-MS untargeted metabolite features	PE (22:6/P-18:0), Cer[r] 34:2, GlcCer[s] (d18:0/18:0), DG[t] 44:0, LysoPC[u] (16:0), 22:6-OH[v]/LysoPC, TAG[w] 58:7-FA[x] 22:4	FS, FP	OPLS-DA[y], receiver operating characteristic-based evaluation	
Metabolomics	Serum	LC-MS untargeted lipid species, targeted lipid species	Taurine, oleoylcarnitine, creatinine	FS, FP	OPLS-DA, receiver operating characteristic-based evaluation	
Labarga et al [39] (2024)						Middle integration
Transcriptomics	Whole-blood mRNA, miRNA, circRNA[z]	Gene-level expression, miRNA expression, circRNA expression	Not reported (high-dimensional omics features; candidate biomarkers mentioned but no explicit ranking provided)	FS, FP, VAL	GraphSAGE[aa]-based embedding, attention weight	
Epigenomics	Whole-blood DNA methylation	CpG[ab] β-values	Not reported (high-dimensional omics features; candidate biomarkers mentioned but no explicit ranking provided)	FS, FP, VAL	Probe-level quality control, GraphSAGE-based embedding, attention weight	
Zeng et al [40] (2024)						Early integration

XSL•FO
RenderX

Study (year) and omics data type	Biospecimen	Analysis unit	Top-ranked feature biomarkers	Role in pipeline	Feature selection method	Integration strategy
Metabolomics	Plasma	LC-MS untargeted metabolite features	Caprolactam	FS, FP, VAL	OPLS-DA, ML selection	
Proteomics	Plasma	DIA[ac] protein groups	C4BPA[ad], COL15A1[ae], HBB[af]	FS, FP, VAL	DEPs[ag], ML selection, cross-model intersection	
Gan et al [41] (2025)						Late integration
Proteomics	Plasma	Antibody-based proteomics analytes	17-protein score (GDF15[ah], PLAUR[ai], NT-proBNP[aj], IGFBP4[ak], BCAN[al] +12 proteins)	FS, FP, RS, VAL	LASSO-Cox[am], 17-protein score derived	

XSL•FO
RenderX

Study (year) and omics data type	Biospecimen	Analysis unit	Top-ranked feature biomarkers	Role in pipeline	Feature selection method	Integration strategy
Genomics	Whole-blood DNA genotypes	PRS[an]	Not reported	FP, RS, VAL	Precomputed PRS	

[a]GWAS: genome-wide association study.
[b]SNP: single nucleotide polymorphism.
[c]FS: feature selection.
[d]FP: final predictors.
[e]RS: risk score.
[f]LD: linkage disequilibrium.
[g]mRNA: messenger RNA.
[h]VAL: validation.
[i]DEAGs: differentially expressed associated genes.
[j]miRNA: microRNA, microRNA of human origin.
[k]DEMs: differentially expressed miRNAs.
[l]ML: machine learning.
[m]ITGAM: integrin subunit alpha M.
[n]LC-MS: liquid chromatography-mass spectrometry.
[o]PE: phosphatidylethanolamine.
[p]OAHFA: (O-acyl)-1-hydroxy fatty acid.
[q]PLS-DA: partial least squares-discriminant analysis.
[r]Cer: ceramide.
[s]GlcCer: glucosylceramide.
[t]DG: diacylglycerol.
[u]LysoPC: lysophosphatidylcholine.
[v]OH: hydroxyl group.
[w]TAG: triacylglycerol.
[x]FA: fatty acid.
[y]OPLS-DA: orthogonal partial least squares-discriminant analysis.
[z]circRNA: circular RNA.
[aa]GraphSAGE: graph sample and aggregate.
[ab]CpG: cytosine phosphate guanine.
[ac]DIA: data-independent acquisition.
[ad]C4BPA: complement component 4 binding protein alpha.
[ae]COL15A1: collagen type XV alpha 1 chain.
[af]HBB: hemoglobin subunit beta.
[ag]DEPs: differentially expressed proteins.
[ah]GDF15: growth differentiation factor 15.
[ai]PLAUR: plasminogen activator, urokinase receptor.
[aj]NT-proBNP: N-terminal pro b-type natriuretic peptide.
[ak]IGFBP4: insulin-like growth factor binding protein 4.
[al]BCAN: brevican.
[am]LASSO-Cox: least absolute shrinkage and selection operator-penalized Cox proportional hazards model.
[an]PRS: polygenic risk score.

Top-ranked feature biomarkers represent the features or panels identified by authors as having the highest contributions. Role in pipeline indicates how each biomarker was used and is coded as feature selection, final predictors, validation, or risk score. Multiple codes may apply when roles differ across study phases. The feature selection method summarizes the dominant selector applied before model fitting. Integration method classifies how multiomics layers were combined, following established taxonomies [43,44]: early integration (simple concatenation of preprocessed omics into a single matrix for a downstream learner), middle integration (a joint latent representation learned directly across omics followed by modeling), and late integration (separate omics-specific models or scores combined at the decision or score level).

Across the 7 included studies, multiomics inputs comprised transcriptomics (n=3) [35,36,39], proteomics (n=3) [36,40,41], metabolomics (n=3) [37,38,40], lipidomics (n=2) [37,38], epigenomics (n=1) [39], and genomics (n=2) [35,41] (Table 3). Biospecimens were predominantly blood derived—whole blood for messenger RNA, circular RNA expression, and

XSL•FO
RenderX

DNA-methylation; serum or whole blood for micro RNA (miRNA) profiling; and serum or plasma for protein panels and liquid chromatography–mass spectrometry (LC-MS)–based metabolite and lipid features. Genomic data were sourced from genome-wide association study summary statistics in 1 study [35] and from whole blood DNA genotypes in another [41].

Most studies (n=6) reported top-ranked biomarkers or panels. By contrast, 1 study [39] used deep learning–based representation learning to process high-dimensional transcriptomic and epigenomic inputs directly as raw features, subsequently deriving multiomics embeddings from the full feature space. Consequently, no ranked lists of biomarkers were generated in this study. Regarding analytic roles, omics variables were consistently used for feature selection and as final predictors. A dedicated validation step—internal and/or external—was described in 5 studies [36-40], while a risk score stage was incorporated in 2 studies [35,41].

Given the heterogeneity of feature selection approaches, Table 3 provides a concise method label for each study. Regarding integration strategies, most studies used middle-level integration (n=4) [35,37-39], while late-level integration was reported in 2 studies [36,41] and early-level integration in 1 study [40]. Middle integration was implemented in three main forms: (1) cross-omics filtering, in which genome-wide association study signals were intersected with transcriptomic differentially expressed genes to define the feature set for ML; (2) projection-based selection, where lipidomic and metabolomic data were reduced to compact multiomics panels; and (3) graph-based joint embeddings. Late integration was used in two distinct ways: (1) validation-stage fusion, in which transcriptomic discoveries were linked with an independent proteomic readout to confirm cross-omics consistency; and (2) score-level fusion, where a protein risk score and a polygenic risk score were jointly incorporated into prediction models.

A study [40] used a mixed stroke end point combining ischemic and hemorrhagic stroke. Because this outcome definition merges etiologically distinct conditions, the identified features are interpreted as mixed-population signals rather than subtype-specific molecular biomarkers. Methodologically, the study performed early integration by analysis-specific quality control and univariate screening, followed by concatenation of the retained features into a single matrix for common selection (LASSO) and modeling (eg, SVM).

Model Performance of Risk Prediction for Stroke

Predictive performance is summarized in Table 4. Discrimination was characterized primarily using AUC with 95% CIs; when available, sensitivity, specificity, validation method, and calibration details were also extracted. The representative performance shown for each study does not necessarily correspond to the numerically highest AUC. Primary estimates were selected to align with the objectives of this review. Multiomics integration and the use of clinically comparable control groups were prioritized as key criteria, and external validation results were adopted whenever available. If external validation was not reported, the best internally validated AUC was recorded. For Gan et al [41], the proteomics plus polygenic risk score specification was designated as the primary estimate because it most reflected multiomics integration, consistent with the inclusion criteria of this review. Further details are provided in the footnotes of Table 4.

XSL•FO
RenderX

Table 4. Predictive performance of multiomics-based machine learning models for stroke in adult population (N=7). For the primary model selected from each study, the table summarizes omics final predictors, validation approach, and key performance metrics. When external validation was reported, "n" indicates the sample size of the external validation cohort, reported separately from the development cohort.

Study (year)	Final predictors, n	Cross-validation	External validation	Calibration	Machine learning model performance			
					Model	AUC[a] (95% CI)	Sensitivity	Specificity
Liu et al [35] (2024)	5	Not reported	Yes (n=12)	Yes[b]	GLM[c]	1.000 (1.000-1.000)	Not reported	Not reported
Chen et al [36] (2025)	1[d]	10-fold	Partial (n=44)	Not reported	LASSO+RF[e]	0.750 (0.601-0.899)	0.567	0.857
Zhao et al [37] (2023)	7	7-fold	No	Yes	RF+LR[f]	0.837[g] (0.732-0.942)	0.897	0.733
Ye et al [38] (2022)	10	Not reported	No	Not reported	SVM[h]	0.963 (0.884-1.000)	Not reported	0.917
Labarga et al [39] (2024)	High-dimensional omics features	5-fold	No	Not reported	BioMGNN[i]	0.950 (not reported)	0.950	Not reported
Zeng et al [40] (2024)	4	10-fold	Yes (n=44)	Yes	SVM	0.973 (0.921-0.999)	0.864	1.000
Gan et al [41] (2025)	2[j]	10-fold	Yes (n=4970)	Not reported	LASSO-Cox[k]	0.748[l] (0.707-0.789)	Not reported	Not reported

[a]AUC: area under the receiver operating characteristic curve.

[b]Calibration was reported in the Methods, but no plot or statistic was presented.

[c]GLM: generalized linear model.

[d]External AUC from integrin alpha M (ITGAM) measured in an independent clinical cohort.

[e]LASSO+RF: least absolute shrinkage and selection operator regression–based feature selection by random forest as the final classifier.

[f]RF+LR: random forest feature selection followed by logistic regression as the final classifier.

[g]Value shown is hypertensive ischemic stroke versus healthy controls.

[h]SVM: support vector machine.

[i]BioMGNN: biological multilayer graph neural network.

[j]Score-level fusion of proteomic and genetic risk scores.

[k]LASSO-Cox: least absolute shrinkage and selection operator-penalized Cox proportional hazards model.

[l]Harrell C statistic for a boundary outcome, treated as numerically equivalent to AUC for quantitative synthesis. The estimate corresponds to protein risk score + polygenic risk score (PRS) to reflect multiomics integration.

Final predictors in Table 4 refer to the predictors actually entered into the final model (eg, selected feature panel, single markers, or score-level inputs), excluding clinical covariates. Across studies, most models relied on compact omics-derived inputs, while 1 study [39] analyzed high-dimensional feature sets. For late-level integration, the final predictors correspond to the score-level [41] or marker-level [36] inputs entered into the model, rather than the underlying omics features used to construct them. Ten-fold cross-validation was the most commonly applied procedure (n=3) [36,40,41], while 7-fold [37] and 5-fold [39] schemes were also reported. Several studies [35,38] did not specify the resampling method. The algorithms most frequently applied were SVM and random forest, often coupled with LASSO for feature selection. Additional approaches included generalized linear models, the LASSO-penalized Cox proportional hazards model, and deep learning architectures such as a biological multilayer graph neural network.

A total of 3 studies [35,40,41] reported external validation of the integrated multiomics model, whereas Chen et al [36] reported an external assessment limited to a single biomarker, integrin alpha M. Three studies [35,37,40] reported calibration assessment. Although Liu et al [35] stated that calibration curves were generated, no plot or statistic was presented. Among externally validated models—excluding 1 boundary estimate (AUC=1.00) [35]—AUC values ranged from 0.75 [41] to 0.97 [40], both of which were onset risk prediction models. The only acute diagnosis model with external validation was Chen et al [36] (AUC=0.75); however, the external validation was conducted as a cross-omics single-marker assessment—integrin alpha M measured at the proteomic level—rather than validation of an integrated multiomics model. When all reported AUCs were stratified by clinical task regardless of validation type, acute diagnosis models reported AUCs ranging from 0.75 [36] to 0.96 [38], whereas onset risk prediction models reported AUCs ranging from 0.75 [41] to 0.97 [40]. Notably, the SVM model in Zeng et al [40] achieved the highest externally validated discrimination (AUC=0.97) using a metabolomics-proteomics combination. However, because this estimate was derived from a mixed stroke end point, it does not reflect performance for a specific stroke subtype. Where

XSL•FO
RenderX

reported, sensitivity and specificity at the chosen decision thresholds are summarized in Table 4. A total of 5 studies [36-40] reported sensitivity and/or specificity, and only 3 studies [36,37,40] reported both metrics at prespecified decision thresholds. Because reporting was incomplete and thresholds and reporting formats were heterogeneous across studies, attempts at separate quantitative synthesis of sensitivity or specificity were not conducted. Reporting of model performance and validation procedures varies across studies, which limited comparability and may have contributed to optimistic estimates. Only 1 study [41] evaluated model performance against standard clinical predictors. In Gan et al [41], the multiomics model had a C statistic of 0.765 in internal validation and 0.748 in external validation, whereas the clinical predictor model showed corresponding values of 0.753 and 0.754, respectively.

Discussion

Overview

This systematic review synthesized 7 studies of multiomics-based ML models for stroke-related prediction tasks. To our knowledge, it is one of the first reviews to focus specifically on multiomics ML in the context of stroke risk stratification. Although many of the included models reported apparently high discrimination, these findings arise from small, heterogeneous, and methodologically limited evidence base characterized by high-dimensional omics inputs, limited external validation, and inconsistent reporting. Accordingly, the primary contribution of this review is not to demonstrate that current multiomics models are ready for clinically meaningful or generalizable accuracy, but rather to show that the current evidence base is methodologically limited, prone to performance inflation, and in need of substantially more rigorous and transparent evaluation.

Discrimination and Methodological Quality in Multiomics-Based Machine Learning Stroke Prediction

On face value, most included multiomics ML models reported high apparent discrimination, particularly when compact biomarker panels were constructed from biologically coherent omics dyads. However, these apparently strong AUCs must be interpreted with extreme caution. Across studies, small-to-moderate sample sizes, high-dimensional inputs, and optimistic validation strategies created ideal conditions for severe overfitting and data leakage. Within this constrained context, some technical patterns were still observable. For example, a metabolomics-proteomics SVM reported an externally validated AUC of 0.97 with a boundary CI (95% CI 0.92-1.00) [40]. This finding is often interpreted as evidence that compact and mechanically coherent biomarker panels can match or exceed more complex multilayered approaches [45]. At the algorithmic level, the prominence of SVM in the best-performing models is also broadly consistent with previous meta-analytic work in cardiovascular ML, in which SVM achieved slightly higher pooled AUCs for stroke outcomes than boosting algorithms or convolutional neural networks [46]. In the specific metabolomics-proteomics SVM model in this review, the combination of a modest sample size and almost upper-limit confidence bound suggests that residual overfitting and optimism remain likely, even with external testing.

A deep learning model applied to transcriptomic-epigenomic matrix also reported an AUC approaching 0.95 [39]; however, this estimate was generated in a cohort of only 30 participants in a pronounced "p≫n" biomarker-discovery setting. This high AUC is therefore best viewed as a methodological artifact rather than a valid predictive performance. Without leakage-robust pipelines and adequately powered external validation, this result should be considered exploratory and should not be used to claim strong discrimination. In this context, the combination of high-capacity architecture, high-dimensional feature space, and limited effective sample size makes severe overfitting and information leakage not merely possible but statistically expected. In addition, the study did not report specificity at prespecified decision thresholds, which makes it difficult to relate the reported sensitivity to false-positive rates and limits the clinical interpretability of the findings [47]. The generalized linear model–based study reported an AUC of 1.00 [35]. Given the small sample size and high-dimensional feature space, this boundary estimate is more appropriately interpreted as an optimistic methodological artifact rather than valid or generalizable predictive performance [48,49]. In addition, calibration was described but not shown; the absence of explicit plots or statistics limits contextual interpretation of the apparent perfect separation. Taken together, these findings reinforce that near-ceiling discrimination estimates in this review are not reliable indicators of clinical validity. Overall, the reported performance should be viewed as preliminary and potentially optimistic given the methodological constraints of the included studies.

Observed Performance Patterns by Integration Strategy and Omics Layers

The performance of multiomics ML models appears to be influenced by both the integration strategy and the specific omics layers used. Within this small and heterogeneous set of studies, most applied middle-level integration, which offered a pragmatic balance for limited sample sizes and heterogeneous data platforms by generating joint feature representations prior to classification [50,51]. Early-level integration, while more vulnerable to dimensionality challenges, produced strong results when combined with stringent feature selection and relatively careful validation [52], as demonstrated in one of the higher-performing studies in this review. By contrast, late-level integration, implemented either as score-level fusion or cross-omics validation, reported more modest AUC values (0.75) [36,41]. Notably, Gan et al [41] used score-level fusion by combining a proteomic risk score with a polygenic risk score, rather than feature-level integration. This formulation may be less able to capture complex interomics relations [44,53]. Accordingly, several methodological studies have suggested that late-level integration may underperform because it cannot fully capture synergistic cross-layer interactions [51,52]. In this review, however, these differences cannot be attributed to integration strategy alone, because the late-integration studies also differed in prediction task, sample size, and model specification. Thus, the apparent underperformance of late-level integration in this review should be regarded as a

XSL•FO
RenderX

hypothesis-generating observation rather than a general conclusion about the inherent limitation of late-integration approaches.

Beyond integration strategies, the selection of omics layers fundamentally shaped predictive utility. Consistent with prior research, transcriptomic-epigenomic pairs were generally used to represent relatively stable regulatory states, whereas lipidomic-metabolomic dyads were used to capture pathway dynamics and acute physiological responses [54-56]. Genomic data mostly contributed a predispositional context rather than serving as direct predictive variables. Additionally, the definition of the analysis unit—for example, gene-level expression, CpG probe values, or LC-MS features—determined the imputation burden, multiplicity, and scope of feasible feature selection methods, thereby influencing both model performance and reproducibility [55,57,58]. Collectively, these observations underscore that methodological design choices in omics integration and feature representation are not simply technical decisions but pivotal determinants of translational potential.

Cross-Study Heterogeneity and Barriers to Quantitative Synthesis

Across the included studies, differences in reported performance primarily reflected variations in study design and analytic choices rather than random fluctuation. Substantial heterogeneity was observed across study populations, prediction targets, stroke end points, and analytic units. In this context, the evidence was synthesized narratively rather than summarized as a single pooled effect size.

More specifically, formal pooling of AUCs was not undertaken because it would have required combining performance estimates derived from substantially different data structures and outcome definitions. With respect to the omics inputs, some studies modeled gene- or protein-level summaries, whereas others analyzed large sets of individual probes or untargeted LC-MS features, leading to marked differences in dimensionality and in the feature-selection and validation strategies that were feasible. At the outcome level, 1 study modeled composite “all-stroke” end points that combined ischemic and hemorrhagic events, whereas others focused exclusively on ischemic stroke. Because these subtypes have distinct pathophysiology and omics profiles, mixed end points limit subtype-specific interpretation and contribute to heterogeneity across studies. Consequently, similar AUC values may reflect different biological mechanisms, case mixes, and follow-up structures, and a single pooled estimate would provide limited additional clinical or methodological insight. Overall, this pattern of heterogeneity is consistent with an early, exploratory phase of multiomics ML research in stroke and highlights the need for more standardized designs and reporting.

Comparable patterns of methodological divergence have been described in broader reviews of ML-based prediction models, in which variation in datasets, predictors, analytic pipelines, and algorithms has constrained the interpretability of pooled accuracy estimates [46,59-61]. Our findings are consistent with this broader literature: the included studies are sufficiently mature for qualitative comparison but not yet harmonized to support robust quantitative synthesis.

Implications for Practice and Future Research

For time-critical decisions in hyperacute stroke care, integrating current multiomics workflows into the minute-to-hour decision windows that govern acute management remains challenging. In particular, most acute diagnosis studies in this review relied on omics profiling that generally depends on centralized laboratory workflows (eg, LC-MS–based proteomic/metabolomic assays). Given current turnaround times, these approaches are unlikely to be clinically translatable within hyperacute treatment windows, limiting near-term clinical feasibility [62]. From a broader risk-stratification perspective, however, multiomics ML approaches may be better suited to longer time horizons, including identification of individuals at elevated future stroke risk, refinement of etiologic classification after the acute phase, and delineation of biologically homogeneous patient subgroups that are not apparent from clinical variables or imaging alone [63,64]. Any future role in acute pathways would require both stronger predictive evidence and assay platforms with substantially shorter turnaround times and lower costs.

Within this context, the current evidence does not yet provide a sufficiently robust foundation for routine implementation of multiomics-based ML models for stroke risk stratification. These models are more accurately regarded as exploratory prototypes, demonstrating potential capabilities within an optimized analytical framework. Compact, blood-derived biomarker panels constructed from biologically coherent omics dyads may offer a plausible route for future clinical application, provided that their performance can be reliably replicated and calibrated in adequately powered, prospectively designed studies using appropriate analytic platforms [65,66].

From a methodological perspective, the studies reviewed highlight critical domains requiring standardization. Repository-based cohorts provide statistical power and facilitate discovery, but they may obscure preanalytical variability and heterogeneity in investigator judgment [67,68]. By contrast, locally recruited cohorts enable clinically focused hypotheses but often suffer from limited sample sizes [69]. Furthermore, the choice of analytic unit—whether gene, CpG probe, protein, or LC-MS feature level—emerged as a critical determinant of which feature selection and integration techniques were feasible, thereby exerting a direct influence on predictive performance. Collectively, these observations underscore the imperative for harmonized analytic frameworks and transparent reporting standards to strengthen reproducibility and enable meaningful cross-study comparability.

Prospective, multisite validation of compact and biologically coherent panels across diverse populations should be a priority. Such studies must incorporate leakage-robust validation pipelines, systematically report calibration and decision-curve analyses, clinically relevant operating points (including sensitivity and specificity at prespecified thresholds), and benchmark performance against both single-omics models and established clinical baselines to quantify true incremental value [70,71]. In future work, both primary studies and evidence syntheses should avoid pooling ischemic and hemorrhagic stroke into undifferentiated “all-stroke” end points when quantifying

XSL•FO
RenderX

model performance, given their distinct pathophysiological and multiomics profiles. These observations collectively support subtype-specific development and synthesis as a more appropriate strategy for future multiomics ML research in stroke. Expanding prediction targets to encompass hemorrhagic stroke, postevent prognostic outcomes, stroke recurrence in secondary prediction context, and health-economic evaluations will also be essential to define the broader clinical utility of multiomics ML approaches. With rigorous methodological safeguards and robust external validation, these models hold the potential to transition from exploratory promise to clinically actionable tools in precision cerebrovascular medicine.

Limitations

This review has several limitations. First, although we initially planned to conduct a meta-analysis, we ultimately judged that conducting a meta-analysis would be inappropriate because the included studies differed markedly in sample size, data sources, validation strategies, and outcome definitions (eg, stroke onset risk prediction, acute diagnosis, and subtype classification); these discrepancies collectively precluded meaningful quantitative synthesis and robust pooled effect estimates. Second, external validation was conducted in only half of the studies, while calibration and clinically relevant operating points were reported in only a minority, limiting confidence in both reproducibility and clinical applicability. Third, most cohorts focused on ischemic stroke and blood-based assays, with only limited data for hemorrhagic or mixed all-stroke populations and alternative biospecimens, restricting generalizability and likely contributing further to clinical and biological heterogeneity. Fourth, reporting gaps—particularly concerning missing-data mechanisms, demographic characteristics, and preanalytical procedures—were frequent and constrained assessment of bias and transportability. Fifth, differences in prediction objectives and performance metrics complicated the cross-study interpretation. Furthermore, the selection of a single representative model in each study, while designed to ensure consistency, introduced subjectivity and may have underestimated the comprehensive range of model performance reported in individual studies. Regarding literature retrieval, the IEEE Xplore search relied on broad umbrella terms rather than a comprehensive set of omics layer–specific keywords, which may have reduced sensitivity for some technical records. Sixth, although we excluded studies with a sample size under 30, the included analyses involved small cohorts relative to the dimensionality of the omics data, making them highly susceptible to overfitting and inflated performance even when cross-validation was reported. Finally, methodological variability across biospecimens, analytic levels, and computational platforms further limited comparability across studies. Collectively, these limitations—including the small number of eligible studies, limited external validation, incomplete reporting, and methodological heterogeneity—underscore the need for standardized study designs, rigorous external validation, and transparent reporting practices to advance the field toward reproducible and clinically meaningful applications.

Conclusion

This systematic review shows that multiomics-based ML models for stroke-related prediction tasks frequently reported high apparent discrimination, but almost exclusively in small, heterogeneous, and high-dimensional studies with limited external validation and incomplete reporting. Current performance estimates are therefore likely optimistic and are not sufficient to justify routine clinical use or to demonstrate superiority over single-omics or purely clinical models. Even the highest externally validated result should be viewed as a promising proof of concept rather than evidence of robust and clinically generalizable performance. Looking ahead, multiomics integration remains a plausible basis for more precise stroke risk stratification, but only if future models are developed and evaluated in adequately powered, prospectively planned, and transparently reported studies. Such work will require leakage-resistant validation frameworks, clear calibration and clinically relevant operating points, and systematic benchmarking against simpler alternatives. Under these conditions, multiomics-based ML models may progress from exploratory prototypes to robust tools that support earlier diagnosis and more tailored intervention in stroke care.

Acknowledgments

The authors declare that generative artificial intelligence tools were not used in the preparation of this manuscript.

Funding

This work was supported by a grant from the National Research Foundation of Korea, funded by the Korean government (Ministry of Science and ICT; grant RS-2025-00563902).

Data Availability

The data underlying the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

HYY contributed to conceptualization, data curation, validation, formal analysis, writing the original draft, and reviewing and editing the manuscript. HS contributed to data curation, validation, software development, visualization, formal analysis, writing the original draft, and reviewing and editing the manuscript. EJK contributed to data curation, validation, software development,

XSL•FO
RenderX

visualization, formal analysis, writing the original draft, and reviewing and editing the manuscript. YJS contributed to conceptualization, methodology, validation, supervision, project administration, writing the original draft, and reviewing and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA checklist for systematic review.
[DOCX File , 33 KB - jmir_v28i1e85654_app1.docx]

Multimedia Appendix 2
Literature search strategy.
[DOCX File , 36 KB - jmir_v28i1e85654_app2.docx]

Multimedia Appendix 3
Critical appraisal and assessment of bias among studies.
[DOCX File , 28 KB - jmir_v28i1e85654_app3.docx]

References

1. Sacco RL, Kasner SE, Broderick JP, Caplan LR, Connors JJB, Culebras A. An updated definition of stroke for the 21st century: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 2013;44(7):2064-2089 [FREE Full text] [doi: 10.1161/STR.0b013e318296aeca] [Medline: 23652265]
2. Bersano A, Gatti L. Pathophysiology and treatment of stroke: present status and future perspectives. Int J Mol Sci 2023;24(19):14848 [FREE Full text] [doi: 10.3390/ijms241914848] [Medline: 37834297]
3. Feigin VL, Owolabi MO, World Stroke Organization–Lancet Neurology Commission Stroke Collaboration Group. Pragmatic solutions to reduce the global burden of stroke: a World Stroke Organization–Lancet Neurology Commission. Lancet Neurol 2023;22(12):1160-1206. [doi: 10.1016/S1474-4422(23)00277-6] [Medline: 37827183]
4. Campbell BCV, De Silva DA, Macleod MR, Coutts SB, Schwamm LH, Davis SM. Ischaemic stroke. Nat Rev Dis Primers 2019;5(1):70. [doi: 10.1038/s41572-019-0118-8] [Medline: 31601801]
5. Lin B, Zhang Z, Mei Y, Wang C, Xu H, Liu L, et al. Cumulative risk of stroke recurrence over the last 10 years: a systematic review and meta-analysis. Neurol Sci 2021;42(1):61-71. [doi: 10.1007/s10072-020-04797-5] [Medline: 33040195]
6. Khosla A, Cao Y, Lin CCY, Chiu HK, Hu J, Lee H. An integrated machine learning approach to stroke prediction. 2010 Presented at: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; July 25-28, 2010; Washington p. 183-192. [doi: 10.1145/1835804.1835830]
7. Zhang S, Zhang W, Zhou G. Extended risk factors for stroke prevention. J Natl Med Assoc 2019;111(4):447-456. [doi: 10.1016/j.jnma.2019.02.004] [Medline: 30878142]
8. Misra S, Singh P, Nath M, Bhalla D, Sengupta S, Kumar A. Blood-based protein biomarkers for the diagnosis of acute stroke: a discovery-based SWATH-MS proteomic approach. Front Neurol 2022;13:989856 [FREE Full text] [doi: 10.3389/fneur.2022.989856] [Medline: 36237606]
9. Misra S, Kawamura Y, Singh P, Sengupta S, Nath M, Rahman Z. Prognostic biomarkers of intracerebral hemorrhage identified using targeted proteomics and machine learning algorithms. PLoS One 2024;19(6):e0296616 [FREE Full text] [doi: 10.1371/journal.pone.0296616] [Medline: 38829877]
10. Cheng S, Xu Z, Bian S, Chen X, Shi Y, Li Y. The STROMICS genome study: deep whole-genome sequencing and analysis of 10K Chinese patients with ischemic stroke reveal complex genetic and phenotypic interplay. Cell Discov 2023;9(1):75 [FREE Full text] [doi: 10.1038/s41421-023-00582-8] [Medline: 37479695]
11. Balasubramanian R, Hu J, Guasch-Ferre M, Li J, Sorond F, Zhao Y. Metabolomic profiles associated with incident ischemic stroke. Neurology 2022;98(5):e483-e492 [FREE Full text] [doi: 10.1212/WNL.0000000000013129] [Medline: 34853177]
12. Hochrainer K, Yang W. Stroke proteomics: from discovery to diagnostic and therapeutic applications. Circ Res 2022;130(8):1145-1166 [FREE Full text] [doi: 10.1161/CIRCRESAHA.122.320110] [Medline: 35420912]
13. Alemu R, Sharew NT, Arsano YY, Ahmed M, Tekola-Ayele F, Mersha TB. Multi-omics approaches for understanding gene-environment interactions in noncommunicable diseases: techniques, translation, and equity issues. Hum Genomics 2025;19(1):8 [FREE Full text] [doi: 10.1186/s40246-025-00718-9] [Medline: 39891174]
14. Wekesa JS, Kimwele M. A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment. Front Genet 2023;14:1199087 [FREE Full text] [doi: 10.3389/fgene.2023.1199087] [Medline: 37547471]
15. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol 2017;18(1):83 [FREE Full text] [doi: 10.1186/s13059-017-1215-1] [Medline: 28476144]

XSL•FO
RenderX

16. Zeng J, Zhou H, Wan H, Yang J. Single-cell omics: moving towards a new era in ischemic stroke research. Eur J Pharmacol 2025;1000:177725. [doi: 10.1016/j.ejphar.2025.177725] [Medline: 40350018]
17. Li W, Shao C, Zhou H, Du H, Chen H, Wan H. Multi-omics research strategies in ischemic stroke: a multidimensional perspective. Ageing Res Rev 2022;81:101730. [doi: 10.1016/j.arr.2022.101730] [Medline: 36087702]
18. Montaner J, Ramiro L, Simats A, Tiedt S, Makris K, Jickling GC. Multilevel omics for the discovery of biomarkers and therapeutic targets for stroke. Nat Rev Neurol 2020;16(5):247-264. [doi: 10.1038/s41582-020-0350-6] [Medline: 32322099]
19. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights 2020;14:1177932219899051 [FREE Full text] [doi: 10.1177/1177932219899051] [Medline: 32076369]
20. Lehmann ALCF, Alfieri DF, de Araújo MCM, Trevisani ER, Nagao MR, Pesente FS. Immune-inflammatory, coagulation, adhesion, and imaging biomarkers combined in machine learning models improve the prediction of death 1 year after ischemic stroke. Clin Exp Med 2022;22(1):111-123. [doi: 10.1007/s10238-021-00732-w] [Medline: 34120242]
21. Mukherjee A, Abraham S, Singh A, Balaji S, Mukunthan KS. From data to cure: a comprehensive exploration of multi-omics data analysis for targeted therapies. Mol Biotechnol 2025;67(4):1269-1289. [doi: 10.1007/s12033-024-01133-6] [Medline: 38565775]
22. Lee J, Park KM, Park S. Interpretable machine learning for prediction of clinical outcomes in acute ischemic stroke. Front Neurol 2023;14:1234046 [FREE Full text] [doi: 10.3389/fneur.2023.1234046] [Medline: 37745661]
23. Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. Brief Bioinform 2022;23(1):bbab454 [FREE Full text] [doi: 10.1093/bib/bbab454] [Medline: 34791014]
24. Li R, Li L, Xu Y, Yang J. Machine learning meets omics: applications and perspectives. Brief Bioinform 2022;23(1):bbab460. [doi: 10.1093/bib/bbab460] [Medline: 34791021]
25. Chen C, Wang J, Pan D, Wang X, Xu Y, Yan J. Applications of multi-omics analysis in human diseases. MedComm (2020) 2023;4(4):e315 [FREE Full text] [doi: 10.1002/mco2.315] [Medline: 37533767]
26. Tarazona S, Arzalluz-Luque A, Conesa A. Undisclosed, unmet and neglected challenges in multi-omics studies. Nat Comput Sci 2021;1(6):395-402. [doi: 10.1038/s43588-021-00086-z] [Medline: 38217236]
27. Sarker IH. Machine learning: Algorithms, real-world applications and research directions. SN Comput Sci 2021;2(3):160 [FREE Full text] [doi: 10.1007/s42979-021-00592-x] [Medline: 33778771]
28. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
29. Cook NR, Paynter NP. Performance of reclassification statistics in comparing risk prediction models. Biom J 2011;53(2):237-258 [FREE Full text] [doi: 10.1002/bimj.201000078] [Medline: 21294152]
30. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115(7):928-935. [doi: 10.1161/CIRCULATIONAHA.106.672402] [Medline: 17309939]
31. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions. Version 6.5. Cochrane. 2024. URL: https://www.cochrane.org/handbook [accessed 2025-08-29]
32. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170(1):51-58 [FREE Full text] [doi: 10.7326/M18-1376] [Medline: 30596875]
33. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc 2020;27(12):2011-2015 [FREE Full text] [doi: 10.1093/jamia/ocaa088] [Medline: 32594179]
34. Kolbinger FR, Veldhuizen GP, Zhu J, Truhn D, Kather JN. Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis. Commun Med (Lond) 2024;4(1):71 [FREE Full text] [doi: 10.1038/s43856-024-00492-0] [Medline: 38605106]
35. Liu Y, Wang W, Cui X, Lyu J, Xie Y. Exploring genetic associations of 3 types of risk factors with ischemic stroke: an integrated bioinformatics study. Stroke 2024;55(6):1619-1628. [doi: 10.1161/STROKEAHA.123.044424] [Medline: 38591222]
36. Chen Z, Zhang X, Qi X, Zheng J, He N, Zheng B. Identification of a potential miRNA-mRNA regulatory network for ischemic stroke by using bioinformatics methods: a retrospective study based on the Gene Expression Omnibus database. Front Immunol 2025;16:1467865 [FREE Full text] [doi: 10.3389/fimmu.2025.1467865] [Medline: 40297585]
37. Zhao T, Yan Q, Wang C, Zeng J, Zhang R, Wang H. Identification of serum biomarkers of ischemic stroke in a hypertensive population based on metabolomics and lipidomics. Neuroscience 2023;533:22-35. [doi: 10.1016/j.neuroscience.2023.09.017] [Medline: 37806545]
38. Ye X, Zhu B, Chen Y, Wang Y, Wang D, Zhao Z. Integrated metabolomics and lipidomics approach for the study of metabolic network and early diagnosis in cerebral infarction. J Proteome Res 2022;21(11):2635-2646. [doi: 10.1021/acs.jproteome.2c00348] [Medline: 36264770]
39. Labarga A, Martínez-Gonzalez J, Barajas M. Integrative multi-omics analysis for etiology classification and biomarker discovery in stroke: advancing towards precision medicine. Biology (Basel) 2024;13(5):338 [FREE Full text] [doi: 10.3390/biology13050338] [Medline: 38785820]

XSL•FO
RenderX

40. Zeng J, Wang C, Guo J, Zhao T, Wang H, Zhang R. Multiomics profiling of plasma reveals molecular alterations prior to a diagnosis with stroke among Chinese hypertension patients. J Proteome Res 2024;23(12):5421-5437. [doi: 10.1021/acs.jproteome.4c00559] [Medline: 39466185]
41. Gan X, Yang S, Zhang Y, Ye Z, Zhang Y, Xiang H. Large-scale plasma proteomics profiles for predicting ischemic stroke risk in the general population. Stroke 2025;56:456-464. [doi: 10.1161/STROKEAHA.124.048654]
42. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF. NCBI GEO: archive for functional genomics data sets—10 years on. Nucleic Acids Res 2011;39(Database issue):D1005-D1010 [FREE Full text] [doi: 10.1093/nar/gkq1184] [Medline: 21097893]
43. Abdelaziz EH, Ismail R, Mabrouk MS, Amin E. Multi-omics data integration and analysis pipeline for precision medicine: systematic review. Comput Biol Chem 2024;113:108254. [doi: 10.1016/j.compbiolchem.2024.108254] [Medline: 39447405]
44. Gligorijević V, Pržulj N. Methods for biological data integration: perspectives and challenges. J R Soc Interface 2015;12(112):20150571 [FREE Full text] [doi: 10.1098/rsif.2015.0571] [Medline: 26490630]
45. Wang M, Yu G, Ressom HW. Integrative analysis of proteomic, glycomic, and metabolomic data for biomarker discovery. IEEE J Biomed Health Inform 2016;20(5):1225-1231 [FREE Full text] [doi: 10.1109/JBHI.2016.2574201] [Medline: 27249841]
46. Krittanawong C, Virk HUH, Bangalore S, Wang Z, Johnson KW, Pinotti R. Machine learning prediction in cardiovascular diseases: a meta-analysis. Sci Rep 2020;10(1):16057 [FREE Full text] [doi: 10.1038/s41598-020-72685-1] [Medline: 32994452]
47. Erickson BJ, Kitamura F. Magician's corner: 9. Performance metrics for machine learning models. Radiol Artif Intell 2021;3(3):e200126 [FREE Full text] [doi: 10.1148/ryai.2021200126] [Medline: 34136815]
48. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One 2019;14(11):e0224365 [FREE Full text] [doi: 10.1371/journal.pone.0224365] [Medline: 31697686]
49. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB. Calculating the sample size required for developing a clinical prediction model. BMJ 2020;368:m441. [doi: 10.1136/bmj.m441] [Medline: 32188600]
50. Cai Z, Poulos RC, Liu J, Zhong Q. Machine learning for multi-omics data integration in cancer. iScience 2022;25(2):103798 [FREE Full text] [doi: 10.1016/j.isci.2022.103798] [Medline: 35169688]
51. Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J 2021;19:3735-3746 [FREE Full text] [doi: 10.1016/j.csbj.2021.06.030] [Medline: 34285775]
52. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res 2018;46(20):10546-10562 [FREE Full text] [doi: 10.1093/nar/gky889] [Medline: 30295871]
53. Wen G, Li L. FGCNSurv: dually fused graph convolutional network for multi-omics survival prediction. Bioinformatics 2023;39(8):btad472 [FREE Full text] [doi: 10.1093/bioinformatics/btad472] [Medline: 37522887]
54. Wang R, Li B, Lam SM, Shui G. Integration of lipidomics and metabolomics for in-depth understanding of cellular mechanism and disease progression. J Genet Genomics 2020;47(2):69-83. [doi: 10.1016/j.jgg.2019.11.009] [Medline: 32178981]
55. Athieniti E, Spyrou GM. A guide to multi-omics data collection and integration for translational medicine. Comput Struct Biotechnol J 2023;21:134-149 [FREE Full text] [doi: 10.1016/j.csbj.2022.11.050] [Medline: 36544480]
56. Pineda S, Gomez-Rubio P, Picornell A, Bessonov K, Márquez M, Kogevinas M. Framework for the integration of genomics, epigenomics and transcriptomics in complex diseases. Hum Hered 2015;79(3-4):124-136. [doi: 10.1159/000381184] [Medline: 26201698]
57. Flores JE, Claborne DM, Weller ZD, Webb-Robertson BM, Waters KM, Bramer LM. Missing data in multi-omics integration: recent advances through artificial intelligence. Front Artif Intell 2023;6:1098308 [FREE Full text] [doi: 10.3389/frai.2023.1098308] [Medline: 36844425]
58. Song M, Greenbaum J, Luttrell J, Zhou W, Wu C, Shen H. A review of integrative imputation for multi-omics datasets. Front Genet 2020;11:570255 [FREE Full text] [doi: 10.3389/fgene.2020.570255] [Medline: 33193667]
59. Lei N, Zhang X, Wei M, Lao B, Xu X, Zhang M. Machine learning algorithms' accuracy in predicting kidney disease progression: a systematic review and meta-analysis. BMC Med Inform Decis Mak 2022;22(1):205 [FREE Full text] [doi: 10.1186/s12911-022-01951-1] [Medline: 35915457]
60. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 2020;46(3):383-400 [FREE Full text] [doi: 10.1007/s00134-019-05872-y] [Medline: 31965266]
61. Wang Q, Chang Z, Liu X, Wang Y, Feng C, Ping Y. Predictive value of machine learning for platinum chemotherapy responses in ovarian cancer: systematic review and meta-analysis. J Med Internet Res 2024;26:e48527 [FREE Full text] [doi: 10.2196/48527] [Medline: 38252469]
62. Simpkins AN, Janowski M, Oz HS, Roberts J, Bix G, Doré S. Biomarker application for precision medicine in stroke. Transl Stroke Res 2020;11(4):615-627 [FREE Full text] [doi: 10.1007/s12975-019-00762-3] [Medline: 31848851]

XSL•FO
RenderX

63. Li Q, Zhao L, Chan CL, Zhang Y, Tong SW, Zhang X. Multi-level biomarkers for early diagnosis of ischaemic stroke: a systematic review and meta-analysis. Int J Mol Sci 2023;24(18):13821 [FREE Full text] [doi: 10.3390/ijms241813821] [Medline: 37762122]
64. Dagonnier M, Donnan GA, Davis SM, Dewey HM, Howells DW. Acute stroke biomarkers: are we there yet? Front Neurol 2021;12:619721 [FREE Full text] [doi: 10.3389/fneur.2021.619721] [Medline: 33633673]
65. Zhang WH, Wang WQ, Han X, Gao HL, Li TJ, Xu SS. Advances on diagnostic biomarkers of pancreatic ductal adenocarcinoma: a systems biology perspective. Comput Struct Biotechnol J 2020;18:3606-3614 [FREE Full text] [doi: 10.1016/j.csbj.2020.11.018] [Medline: 33304458]
66. Gallitto G, Englert R, Kincses B, Kotikalapudi R, Li J, Hoffschlag K. External validation of machine learning models-registered models and adaptive sample splitting. Gigascience 2025;14:giaf036 [FREE Full text] [doi: 10.1093/gigascience/giaf036] [Medline: 40366867]
67. Michalska-Falkowska A, Niklinski J, Juhl H, Sulewska A, Kisluk J, Charkiewicz R. Applied molecular-based quality control of biobanked samples for multi-omics approach. Cancers (Basel) 2023;15(14):3742 [FREE Full text] [doi: 10.3390/cancers15143742] [Medline: 37509403]
68. Rozera T, Pasolli E, Segata N, Ianiro G. Machine learning and artificial intelligence in the multi-omics approach to gut microbiota. Gastroenterology 2025;169(3):487-501 [FREE Full text] [doi: 10.1053/j.gastro.2025.02.035] [Medline: 40118220]
69. Mataraso SJ, Espinosa CA, Seong D, Reincke SM, Berson E, Reiss JD. A machine learning approach to leveraging electronic health records for enhanced omics analysis. Nat Mach Intell 2025;7(2):293-306. [doi: 10.1038/s42256-024-00974-9] [Medline: 40008295]
70. Hassan A, Gulzar Ahmad S, Ullah Munir E, Ali Khan I, Ramzan N. Predictive modelling and identification of key risk factors for stroke using machine learning. Sci Rep 2024;14(1):11498 [FREE Full text] [doi: 10.1038/s41598-024-61665-4] [Medline: 38769427]
71. Glaab E, Rauschenberger A, Banzi R, Gerardi C, Garcia P, Demotes J. Biomarker discovery studies for patient stratification using machine learning analysis of omics data: a scoping review. BMJ Open 2021;11(12):e053674 [FREE Full text] [doi: 10.1136/bmjopen-2021-053674] [Medline: 34873011]

Abbreviations

AI: artificial intelligence
AUC: area under the receiver operating characteristic curve
LASSO: least absolute shrinkage and selection operator
LC-MS: liquid chromatography–mass spectrometry
MINIMAR: Minimum Information for Medical AI Reporting
ML: machine learning
PICO: participants, interventions, comparisons, and outcomes
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction Model Risk of Bias Assessment Tool
PROSPERO: International Prospective Register of Systematic Reviews
ROC: receiver operating characteristic
SVM: support vector machine

Edited by A Coristine; submitted 10.Oct.2025; peer-reviewed by T Liu, H Lv, M Gandhi; comments to author 10.Nov.2025; revised version received 27.Jan.2026; accepted 29.Jan.2026; published 19.Feb.2026.

Please cite as:
Yoo HY, Shin H, Kim EJ, Son YJ
Machine Learning for Predicting Stroke Risk Stratification Using Multiomics Data: Systematic Review
J Med Internet Res 2026;28:e85654
URL: https://www.jmir.org/2026/1/e85654
doi:10.2196/85654
PMID:

©Hae Young Yoo, Hyerim Shin, Eun-Jung Kim, Youn-Jung Son. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly

XSL•FO
RenderX

cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Advancing Nursing Data Integration Through a Nursing Minimum Dataset for the Conceptual and Technical Development of a "Fall Prevention" Data Module: Development Study

Sarah Milkov[1], MSc; Antonia Schmidt[2], MSc; Anja Burmann[3], PhD; Niklas Tschorn[2], MSc; Marcel Klötgen[2], MSc; Wolfgang Deiters[1], Prof Dr; Christian Potthoff[4]; Kirsten Neveling[4], BSc; Yvonne Weber[5], MSc; Maren Keuchel[3], MSc; Daniela Holle[1], Prof Dr

[1]Hochschule Bochum, Bochum, Germany

[2]Fraunhofer Institute for Software and Systems Engineering ISST, Dortmund, Germany

[3]Fraunhofer Institute for Software and Systems Engineering, Dortmund, North Rhine-Westphalia, Germany

[4]Diakonie Michaelshoven Pflege und Wohnen gGmbH, Köln, Germany

[5]Connext Communication GmbH, Paderborn, Germany

Corresponding Author:
Sarah Milkov, MSc
Hochschule Bochum
Gesundheitscampus 6-8
Bochum, 44801
Germany
Phone: 49 23477727701
Email: Sarah.Milkov@hs-bochum.de

Abstract

Background: In aging populations, the demand for care, including care delivery in long-term care (LTC) facilities, is increasing. This situation highlights the need to optimize care processes through continuous scientific evaluation. The use of artificial intelligence (AI) has the potential for use in nursing research, but it experiences a lack of standardization and structuring of nursing data. Although solutions such as standardized nursing terminologies exist, their use in practice has thus far not been widespread and is often associated with high documentation costs.

Objective: This paper presents the conceptual and technical development of a nursing minimum dataset that focuses on a specific "fall prevention" use case. The aim of this work was to improve data standardization and usability for research and AI-based analysis in LTC settings.

Methods: A representation of the "fall prevention" use case was developed using literature analyses, co-design workshops, and a quantitative survey (n=158). Technical indexing was conducted by translating the results into the technical terminology of the Health Level Seven International Fast Healthcare Interoperability Resources standard.

Results: The "fall prevention" use case was developed as part of a German nursing minimum dataset for long-term residential care with 8 basic modules (patient or client demographics) and 11 extension modules (nursing care elements). The module of the "fall prevention" use case includes fall risk factors, interventions, and outcomes. The literature analysis included 4 international fall guidelines and 17 practice and transfer documents established in German LTC. In total, 12 experts from the fields of management, quality management, technical application support, nursing service management, department management, and members of the PFLIP (Pflege-Kerndatensatz und Intersektorales Pflegedaten-Repository [Nursing Minimum Data Set and Intersectoral Nursing Data Repository]) research project participated in the workshops. A total of 158 people participated in the quantitative survey, the majority of whom were female (117/158, 74%), with 63% (100/158) working directly in nursing care and an average of 24.9 years of professional experience, mainly in LTC (63/158, 40%), outpatient care (37/158, 23%), and hospitals (14/158, 9%). The relevant content, in the sense of a minimum set of items, was identified and prioritized in collaboration with nursing experts and translated into a Fast Healthcare Interoperability Resources–based implementation guide.

Conclusions: This approach addresses the lack of structured nursing data for AI and research and can serve as an example for interoperable, cross-sector solutions in global LTC.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e82417) doi:10.2196/82417

KEYWORDS

nursing; minimum dataset; core dataset; fall; standardization; decision-making

Introduction

Background

Demographic change in Germany and Europe is leading to a steadily growing number of people in need of care, especially older people with multiple health problems [1]. A growing problem in this age group is falls, which place considerable physical and psychological strain on both those affected and the care systems and increase the need for treatment [2,3]. The continuous scientific development of nursing processes is essential for improving the quality of care, promoting evidence-based knowledge, and further developing nursing practices [4,5].

The application of artificial intelligence (AI) provides new possibilities for nursing research by enabling the efficient analysis of large amounts of care data, recognizing patterns, and obtaining scientifically useful insights [6]. This strengthened nursing research contributes significantly to improving and facilitating nursing care, as scientific findings can be used to develop evidence-based care strategies [7]. In turn, data from nursing care form the basis for AI-supported analyses (Figure 1) [8]. The cyclical process between (AI-based) research, health care and nursing care, and AI innovations, with data serving as the central interface, is shown in Figure 1. Using the example of fall prevention, AI innovations can enhance research opportunities, such as studies on falls in long-term care (LTC). The findings from this research can then improve the quality of nursing care and help prevent falls. In nursing care, data are continuously generated, for example, when falls are documented, which feeds back into research and supports further AI-driven innovations in fall prevention. The challenge is to bring the data collected in the health care system into a uniform format that is accessible for research.

Figure 1. Nursing research, care, and AI innovation cycle. AI: artificial intelligence; HC: health care; NC: nursing care.

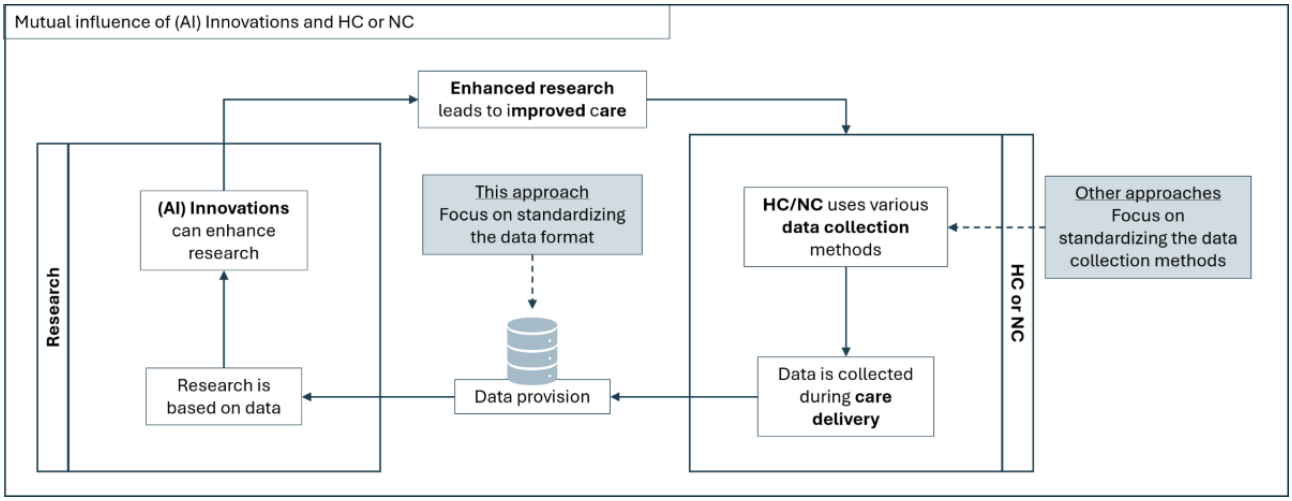

The secondary use of existing documentation data offers an alternative to traditional prospective studies for research purposes. However, accessing and harmonizing these data are challenging due to heterogeneous systems and varying documentation practices, which often result in unstructured or nonstandardized data.

Existing solutions, such as internationally recognized nursing terminologies or so-called standardized nursing languages, aim to standardize the documentation of nursing data by specifying certain terms, topics, and terminologies [9]. However, the use of standardized nursing terminology and specialist languages is not common across Europe. The use of structured nursing documentation varies widely across countries, and studies by Thoroddsen et al [10] and Dos Santos et al [11] suggest that nursing diagnoses and interventions are used most frequently.

The general use of standardized nursing terminology is still lacking, which makes access to nursing data more difficult [10]. A cross-sectional study from 2020 in the Netherlands revealed that approximately half of the 667 nurses and assistants surveyed would use standardized terminology in electronic documentation [12]. The reasons for this include burdensome documentation (use barrier) and a still predominant use of medical terminology (terminology barrier), although medical terminology is insufficient for describing nursing processes, such as assessing unsteadiness or unsuitable footwear. Other studies also show that complex documentation is inefficient in care practice, consumes resources necessary to perform care measures, and leads to a considerable documentation burden for nursing staff [11,13-16].

In the case of Germany, LTC facilities took a step away from standardization toward more patient-oriented documentation in 2015 with the introduction of a law to reduce bureaucracy in care documentation, which led to an increased use of unstructured free-text entries [17,18]. The absence of legal provisions governing the use of standardized terminology constitutes a structural barrier to the implementation of such concepts [19]. Furthermore, there is currently a paucity of

XSL•FO
RenderX

terminology that has been specifically developed for German LTC, with the literature consisting of translations of foreign approaches. Nursing care systems are characterized by a high degree of heterogeneity in different countries, which reinforces the use barrier [20].

This paper addresses a research gap in German LTC caused by the lack of integration and routine use of nursing terminology. This situation gives rise to a paucity of structured and standardized data, while existing approaches to data use remain ineffective, resulting in the untapped potential of valuable data for research purposes remaining unused. Our approach aims to standardize the format of research data rather than the methods of data collection, allowing nursing staff to continue working as usual, while their collected data are converted for research purposes (Figure 1).

The following research questions are addressed:

- What conceptual, nursing, and technical requirements are necessary to develop a nursing minimum dataset (NMDS) based on existing nursing care data from routine documentation using the example of the "fall prevention" use case?
- How can a minimum modular nursing dataset contribute to the harmonization and integration of nursing data for cross-sectoral care and data-based research?

The conceptual and technical development of an NMDS according to the definition by Werley et al [21] for the context of LTC is presented via the "fall prevention" use case, and its potential is discussed in terms of the challenges mentioned. In our use case, an analysis is conducted of data on risk factors, interventions, and outcomes of falls with the aim of preventing such incidents. In addition, a possible practical application in the context of the AI-supported analysis of nursing data is presented.

Objectives

This paper presents a new approach to enable the secondary use of nursing data without additional effort in nursing practice. To this end, the following objectives are pursued: (1) outline the necessary steps for the development of an NMDS based on the "fall prevention" use case, guided by the literature and expert input; (2) describe the technical approach to systematically record fall risk factors, including the collection and prioritization of key factors; and (3) provide a perspective on an NMDS for data collection and harmonization, demonstrating its applicability for research and AI-based analysis.

Related Work

Some related work must be accounted for when addressing the research objectives. The German research project PFLIP (Pflege-Kerndatensatz und Intersektorales Pflegedaten-Repository [Nursing Minimum Data Set and Intersectoral Nursing Data Repository]), on which the following results rely, aims to standardize previously collected nursing data to make the data available for research and AI applications and to offer the possibility of harmonizing data across sectors [22]. This project is based on the German Medical Informatics Initiative (MII), which has developed a medical core dataset on which the NMDS presented in this paper is based [22,23].

With respect to the scientific foundations of fall risk assessment, the "fall prevention expert standard" guidelines served as a foundation for the development of a use case. In Germany, the nursing guidelines that define best practices and quality benchmarks to ensure high standards of nursing care are called "expert standards" [2].

With respect to standardization and interoperability, the National Association of Statutory Health Insurance Physicians in Germany publishes relevant health care transition documents for interoperable data exchange in a telematic infrastructure. The work of this organization is used as a basis for the development of conceptual and technical use cases [24].

To enable standardized fall data to be used on a broader basis, a suitable data format must be chosen. The Health Level Seven International *(HL7)* Fast Healthcare Interoperability Resources *(*FHIR*) standard* is an internationally recognized framework for the exchange of health information, and its technical terminology is used by health care providers, software developers, and governments worldwide to improve interoperability between different systems and facilitate access to health data. The active community of FHIR includes professionals from various health care fields, who work together to further develop and implement the standard to optimize global health care.

This paper examines the development of an NMDS for LTCs using the example of the "fall prevention" use case. Building on initiatives such as the MII, German expert standards, and international standards such as HL7 FHIR, the NMDS aims to harmonize nursing data, integrate it across sectors, and make it usable for both practice and research, assuming that such an approach is feasible, relevant to practice, and sustainable to implement.

Methods

Overview

The methodology for achieving our research goals consists of an approach of conceptual standardization, which we combine with and translate into technical standardization. An overview of the procedure is presented in Figure 2.

XSL•FO
RenderX

Figure 2. Methodological approach. AI: artificial intelligence; FHIR: Fast Healthcare Interoperability Resources; HL7: Health Level Seven International; MII: Medical Informatics Initiative; NMDS: nursing minimum dataset.

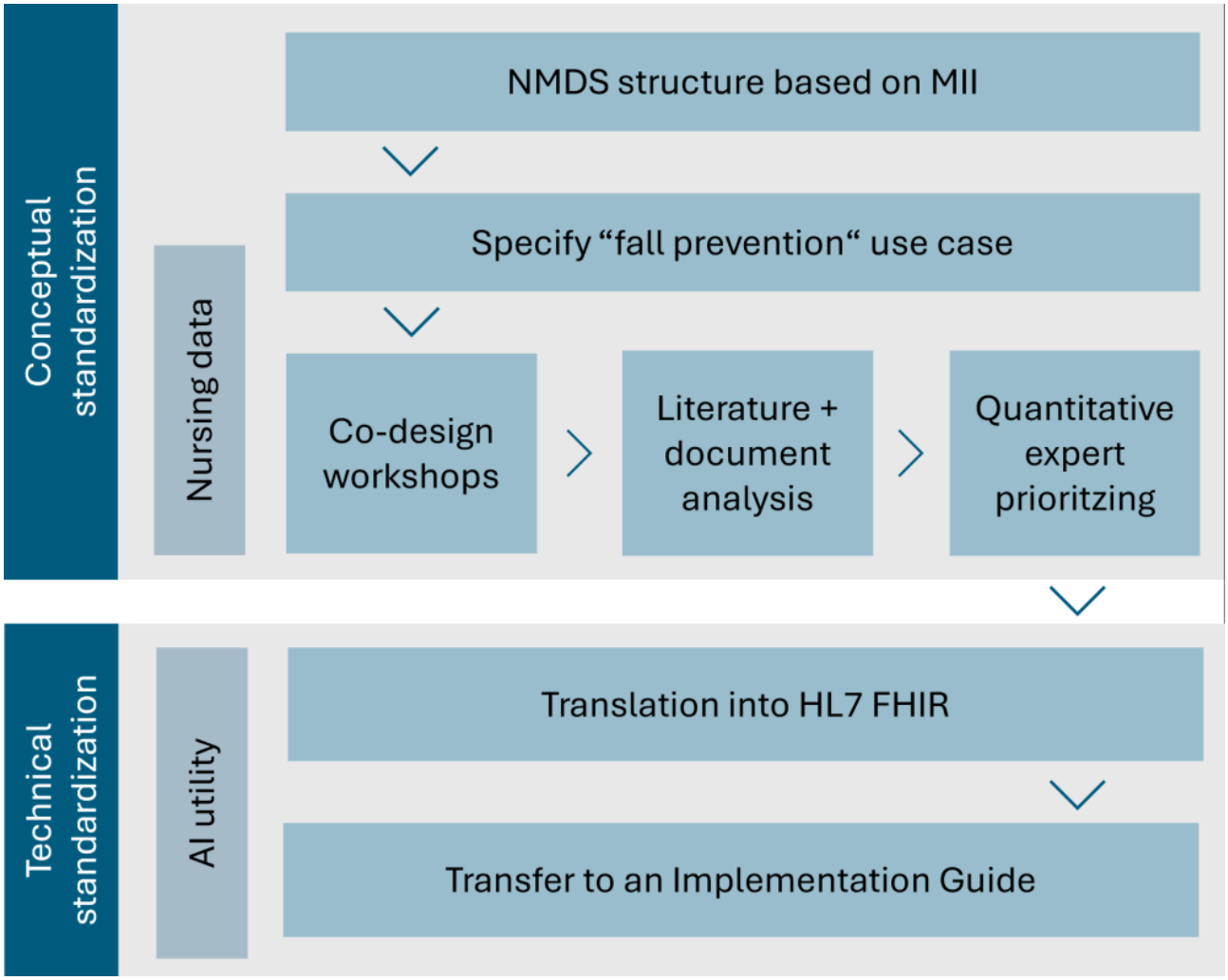

Conceptual Standardization

Overview

A co-design approach was used to develop the content of the "fall prevention" use case of an NMDS. Co-design, according to Sanders and Stappers [25], is an approach that describes collective creativity throughout an entire design process. Designers collaborate to develop innovative solutions by integrating diverse perspectives and skills into the design process [22]. Within this approach, an NMDS structure was developed, and the use case was selected and developed in a participatory manner. The development was supplemented by literature and document analyses and a quantitative survey of nursing staff, and the initially broad approach was narrowed down to a minimum set of items.

Co-Design Development of an NMDS Based on the "Fall Prevention" Use Case

The aim was to design a structure for an NMDS and specify the "fall prevention" use case. The use case was considered particularly suitable because a fall event is extensively documented for insurance-related reasons; therefore, extensive data are available [2].

The foundation for this work was a model developed within the PFLIP research project, drawing on the MII medical core dataset and the German guideline "Expert Standard of Nursing," which served as a basis for this study, while details of its development are not reported here (Multimedia Appendix 1) [20,23]. A qualitative, participatory design in the form of a co-designed expert workshop was chosen to achieve the stated goal.

Participants were recruited through the management of the participating nursing facility in writing and through personal contact. The sample was specifically selected to reflect a range of nursing expertise, and the criteria for participation were work in the professional fields of general management, quality management, technical application support, nursing service management, nursing department management, and contact with fall documentation in everyday work. No restrictions were imposed regarding sociodemographic data. Before the workshop took place, all the participants provided their consent for the procedure and data processing. Participation was voluntary and could be discontinued or revoked at any time. The data were processed anonymously; thus, no conclusions could be drawn about individuals, and the data were handled in accordance with the General Data Protection Regulation.

The study was not approved by an institutional review board. The aim of the study was to gather opinions and assessments

XSL•FO
RenderX

from employees of a nursing facility as experts on a fictitious use case. This is not interventional or experimental research on humans, and no personal data were collected.

The "fall prevention expert standard" [26] served as the foundation for the use case and was initially evaluated for completeness by the nursing staff during a workshop; its content was visualized for the participants, which allowed them to prioritize the listed fall-related items according to perceived relevance and to suggest additional aspects they considered important in the context of fall prevention. The results were compiled verbally and documented in writing by the research team. Additionally, patient journey mapping [27] was developed to understand how residents of nursing facilities navigate the nursing system, to illustrate the interfaces in the cross-sectoral documentation of a fall, and to identify additional documentation tools that provide information on the current standard fall documentation.

The results were evaluated using qualitative content analysis according to Mayring [28]. Coding was performed deductively using predefined categories derived from the expert standard. This approach enables a structured and comprehensible analysis of the data.

Literature and Document Search

Additional international guidelines for fall events were researched on the platform Google Scholar and in national guideline registers. Practice documents and legal requirements were evaluated to identify current documentation of falls in nursing care. The documents were searched directly on the website of providers for transfer documents in medicine and nursing care as well as an LTC facility in Germany. Finally, a targeted search for standardized nursing terminology was conducted to incorporate international standards. This search included an analysis of documents on the International Classification of Nursing Practice [29].

Guidelines that addressed the topic of "fall prevention" in LTC settings and among older adults were included. The guidelines were subsequently analyzed in terms of content (risk factors, interventions, and outcomes). All the documents were screened according to these criteria.

Relevant information from the guidelines and documents was systematically extracted and organized according to predefined categories of risk factors, interventions, and outcomes and presented in tabular form, ensuring that the guideline-based requirements relating to falls in LTC were clearly and comprehensively presented.

Quantitative Online Survey

An anonymous online survey was conducted to specify broad findings in a practice-relevant manner. Below, we report on the study following the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist [30].

The target population included people working in the nursing sector who were chosen for open convenience sampling. The questionnaire was distributed nationwide to nursing facilities in various settings as well as to nursing staff with different professional backgrounds and experience levels.

The survey was created with the software LimeSurvey (LimeSurvey GmbH) and can be found in Multimedia Appendix 2. The aim of the questionnaire was to prioritize fall events identified by nursing staff. The questionnaire included demographic questions and a Likert scale on which the participants rated the relevance of various fall-related items (risk factors, interventions, and outcomes). The participants also indicated whether the items mentioned had previously been documented in their care facility and, if so, whether they had been documented digitally. The findings from the previous steps were used as a base set of questions for the survey.

The survey responses were descriptively analyzed and graphically presented via the software Looker Studio (Google LLC). As the survey was created in LimeSurvey, it was possible to set it up so that no incomplete questionnaires were saved; this ensured that no incomplete answers were included in the evaluation.

Technical Standardization

In the next step, the interim results were translated into the technical terminology of the HL7 FHIR standard and made available for use in digital applications and AI-based systems.

Systematic Capture of Fall Risk Factors and Outcomes

To enable broader use of the standardized fall data, HL7 FHIR was chosen as a suitable data format. The methodology used to translate the identified fall parameters into HL7 FHIR involved a series of systematic processes. The focus was initially limited to fall risks that were documented digitally in our cooperating LTC facility and received more than 80% agreement on their relevance for fall detection by the nursing staff, as rated in the Quantitative Online Survey section. This restriction, including the deliberate exclusion of interventions, was made to conserve time- and effort-related resources. Initially, a thorough screening of the FHIR base resources was conducted to determine which resources were most suitable for representing the identified fall risk factors. Subsequently, existing implementation guides (IGs), such as the MII core dataset, were searched for eligible profiles that could be used as a basis for further profiling. Fall risk factors and fall outcomes were subsequently mapped to appropriate codes in the Systematized Nomenclature of Medicine Clinical Terms [31] and the *ICD-10* (*International Statistical Classification of Diseases and Related Health Problems*) [32] to ensure that each risk factor was accurately represented in a standardized format that can be widely recognized and used in the health care industry. An example of this is shown in Multimedia Appendix 3. Finally, FHIR Shorthand was used during the profiling process to create a concise, human-readable, and computable representation of the mapped fall risk factors; this significantly simplified the process of creating and maintaining FHIR resources and profiles.

Perspective on a "Fall Prevention" NMDS for Data Collection and the Harmonization of Nursing Data for Research Purposes

In addition to the systematic capture of fall risk factors, the definition of "fall" NMDS is an important part of the reusability of the data and its structure. As discussed in the previous sections, this secondary use of data plays an essential role in

XSL•FO
RenderX

research on health services. These aspects are also recognized by research data networks such as the German MII, which relies on FHIR. To be compatible with these structures in the future, the NMDS "fall" use case for data collection and the harmonization of nursing data were transformed to FHIR.

An IG was created that provides a detailed, step-by-step manual that effectively communicates how to implement the mapped fall risk factors in FHIR. The IG offers a topic-specific bundled presentation of profiles, value sets, and other FHIR data structures. The need for this IG is based on recognition of the complexity of health care data and the challenges that can arise when standardized health care data models such as FHIR are implemented.

Another argument in favor of following the example of existing research data networks is the prospect of being connected to them. Considering and incorporating the FHIR-based core dataset of the MII ensure that the collection of fall risk factors can be included in the research data network and made available to researchers.

Ethical Considerations

This study did not require approval by an institutional review board, as it did not involve interventional or experimental research with human participants. The study aimed to collect opinions and assessments from nurses as experts regarding fictitious fall risks derived from the literature. All participants received written information about the purpose, scope, and procedures of the questionnaire. Participation was entirely voluntary, and informed consent was obtained prior to data collection. Participants were informed that their data would be processed anonymously. The study was conducted in accordance with the principles of the Declaration of Helsinki, the guidelines of the Council for International Organizations of Medical Sciences, the World Health Organization, the European General Data Protection Regulation, the German Federal Data Protection Act, and the Data Protection Act of the State of North Rhine-Westphalia (Landesdatenschutzgesetz NRW). As all survey data were fully anonymized and no information was collected that would allow the identification of individual participants, the study does not involve the processing of personal data as defined in Article 4(1) of the European General Data Protection Regulation. All data were stored on protected servers. All participating institutions approved the study as part of a data protection agreement, with the involvement of data protection experts.

Results

Conceptual Standardization

The "fall prevention" use case was conceptually developed on the basis of an NMDS based on the MII core data (Multimedia Appendix 2).

Co-Design Development of an NMDS Based on the "Fall Prevention" Use Case

The workshops took place in August 2022. In total, 12 participants from the fields of management, quality management, technical application support, nursing service management, and department management and members of the PFLIP research project participated in the co-design workshops. Experts from the field of nursing science also played a key role in developing patient journey mapping. During this process, participants observed that while the expert standard provides valuable insights, it could be further enriched by considering additional factors to capture the full complexity of falls, leading to an expansion of the topic. Patient journey mapping helps to identify critical interfaces, relevant documentation, and transfer forms that are essential for a comprehensive representation of fall-related processes. The results are presented in Multimedia Appendix 4. On this basis, a search for these documents was initiated to incorporate them into the development of the use case.

Literature and Document Search

As part of the literature analysis, 4 additional international fall guidelines, 5 documents that are routinely applied in German LTC practice, 6 documents from a provider of transfer documents in Germany, 6 transfer documents from a selected LTC facility in Germany, and the International Classification of Nursing Practice falls catalogue were included. Many of these documents depict fall events rather generally. Thus far, no study has comprehensively addressed falls along with associated nursing-relevant risk factors and intervention options.

Quantitative Survey

The quantitative online survey was conducted from March 17, 2023, to July 30, 2023, with a total of 158 participants, 117 of 158 (74%) of whom were identified as female. Among these participants, 63% (100/158) were directly involved in nursing care. The participants had an average of 24.86 years of professional experience, ranging from 0.5 to 45 years. Most respondents (63/158, 40%) worked in LTC at the time of the survey, followed by home care (37/158, 23%) and hospital settings (14/158, 9%).

Risk factors were identified as items in the following categories: diseases, disease-related changes, environment, medications, and other factors. As interventions, items were identified in the following categories: physical measures, clothing adjustment measures, environmental adjustment measures, training and counseling measures, measures involving other professional groups, measures for the use of assistive devices for fall and injury prevention, the use of technology-based measures, and additional nursing measures. As outcomes, items were identified in the category of information about fall events. The results, including the identified content of the "fall prevention" use case, are presented in Multimedia Appendix 5. Table 1 shows the prioritized risk factors, including the percentage of agreement, which were further processed in the subsequent steps.

XSL•FO
RenderX

Table 1. Prioritized fall risk factors.

Risk factor	Rated importance, n (%)
1. Dizziness (vertigo)	158 (100)
2. Parkinson disease	158 (100)
3. Gait instability	158 (100)
4. Poststroke condition	157 (99)
5. History of falls	157 (99)
6. Overestimation of abilities	157 (99)
7. Mobility impairment	156 (99)
8. Antidepressants	156 (99)
9. Multiple sclerosis	156 (99)
10. Musculoskeletal impairment	155 (98)
11. Assistive devices (eg, rollator)	155 (98)
12. Visual impairment	154 (97)
13. Delirium (confused state)	154 (97)
14. Blood pressure fluctuations	154 (97)
15. Joint paresis (paralysis)	153 (97)
16. Confusion	151 (96)
17. Dementia	151 (95)
18. Fear of falling	150 (95)
19. Cognitive impairment	149 (94)
20. Sensorimotor impairment	148 (94)
21. History of fractures	148 (94)
22. Polypharmacy	147 (93)
23. Multimorbidity (multiple conditions)	146 (92)
24. Blood sugar fluctuations	146 (92)
25. Inadequate activities	146 (92)
26. Compliance	145 (92)
27. Alcohol or drug or nicotine abuse	143 (90)
28. Limb amputation	142 (90)
29. Antihypertensives (blood pressure–lowering drugs)	142 (90)
30. Psychotropics (affecting the psyche)	140 (89)
31. Pain	140 (88)
32. Anticonvulsants (to prevent epileptic seizures)	139 (88)
33. Cardiac arrhythmia	137 (86)
34. Measures involving deprivation of liberty	136 (86)
35. Wandering behavior	133 (84)
36. Antiarrhythmic medications	131 (83)
37. Incontinence	130 (82)
38. Sleep disorder	127 (80)
39. Osteoarthritis	127 (80)

XSL•FO
RenderX

Technical Standardization

Systematic Capture of Fall Risk Factors

An overview of the risk factors considered for translation into FHIR can be found in Table 1. The objectives were considered because these received more than 80% agreement on relevance and were currently documented in our cooperating LTC facility.

When the risk factors were translated into the appropriate FHIR resources, existing profiles were screened. For risk factors involving a disease diagnosis, it was necessary to distinguish between medical and nursing diagnoses, because the term "diagnosis" is legally protected and relevant for billing purposes in the German health care system. In total, 60% (n=21) of the risk factors can be mapped as diagnoses, of which 24% (n=10) are nursing diagnoses and 26% (n=11) are medical diagnoses, whereby a total of 4 risk factors may be specified as both medical and nursing diagnoses. For 21% (n=9) of the risk factors, a nursing observation profile was created, which can be used to describe an occurrence during normal nursing activities. A further 12% (n=5) could be mapped to the MII medication profile. More specific profiles were created for the remaining 14% (n=7) of the risk factors, as they had more precise requirements and required differentiation. For example, the risk factor for incontinence was divided into fecal and urinary incontinence, and the necessity of a decision was included for measures involving deprivation of liberty.

Perspective on a "Fall Prevention" NMDS for Data Collection and Harmonization of Nursing Data for Research Purposes

At the heart of the developed IG is a questionnaire, represented by the FHIR resource questionnaire and questionnaire response, which was specifically designed for the "fall prevention" use case. The guide serves as an essential tool for ensuring consistency and accuracy in the implementation process across different systems and platforms. The response form composed of these 2 FHIR resources allows for the systematic documentation of all fall-relevant risk factors. The IG offers a collection of FHIR resources to capture previously identified risk factors for a fall. The order of the questionnaire response items reflects the prioritization of risk factors from the survey (Table 1). In addition to profiles designed solely for one or a few other risk factors, 3 more comprehensive profiles have been developed that can depict a multitude of risk factors (Tables 2 and 3) [33].

Table 2. Self-defined profiles in the PFLIP[a] implementation guide.

HL7[b] FHIR[c] basic resource	Profile designation
Bundle	• Multi-Morbidity (original Multimorbidität)
Condition	• Medical-Diagnosis (original Medizinische-Diagnose) • Nursing-Diagnosis (original Pflege-Diagnose)
Observation	• Nursing-Observation (original Pflege-Beobachtung) • Nursing-Observation-Urinary-Incontinence (original Pflege-Beobachtung-Harn-Inkontinenz) • Nursing-Observation-Faecal-Incontinence (original Pflege-Beobachtung-Stuhl-Inkontinenz) • Survey-Fall-History-6-Months (original Erhebung-Sturzhistorie-6-Monate) • Deprivation-of-Liberty (original Freiheitsentziehende-Maßnahme)
Device	• Aids (original Hilfsmittel)
List	• Multi-Medication (original Multimedikation)
Questionnaire	• Risk-Assessment (original Risikoermittlung)
QuestionnaireResponse	• Risk-Assessment-Fall (original Risikoermittlung-Sturz)
Patient	• Resident-Pseudonymized (original BewohnerIn-Pseudonymisiert)

[a]PFLIP: Pflege-Kerndatensatz und Intersektorales Pflegedaten-Repository (Nursing Minimum Data Set and Intersectoral Nursing Data Repository).

[b]HL7: Health Level Seven International.

[c]FHIR: Fast Healthcare Interoperability Resources.

XSL•FO
RenderX

Table 3. Extensions within the PFLIP[a] implementation guide.

Name	Context of use
VoluntaryConsentExists (original *FreiwilligeEinwilligungLiegtVor*)	Deprivation-of-Liberty (original *Freiheitsentziehende-Maßnahme*)
CourtOrderExists (original *RichterlicheAnordnungLiegtVor*)	Deprivation-of-Liberty (original *Freiheitsentziehende-Maßnahme*)
DateOfBirth_mm_yyyy (original *Geburtstag_mm_yyyy*)	Resident-Pseudonymized (original *BewohnerIn-Pseudonymisiert*)
FullSurveyPeriod (original *VollständigerErhebungszeitraum*)	Survey-Fall-History-6-Months (original *Erhebung-Sturzhistorie-6-Monate*)

[a]PFLIP: Pflege-Kerndatensatz und Intersektorales Pflegedaten-Repository (Nursing Minimum Data Set and Intersectoral Nursing Data Repository).

Discussion

Principal Findings

The development of a German NMDS use case for fall prevention in LTC is based on a literature analysis of international guidelines and German practice documents, resulting in fall risk factors, interventions, and outcomes, and was developed in collaboration with 12 experts from various nursing fields. The relevant minimum items were prioritized by experts, reflected upon in a quantitative survey with 158 participants, and finally transferred into an FHIR-based IG.

The developed use case can benefit nursing research by making standardized, documented data available for secondary use and enabling AI-supported analyses. Since the dataset is based on routine documentation in nursing facilities and standardizes already documented data, it has the potential to support problems that have arisen thus far in efforts to standardize care documentation, such as a lack of manageability and the effort of documentation by nursing staff [13-16].

In contrast to broader international NMDS initiatives, the approach presented here takes a more detailed approach to a specific use case, resulting in practical, actionable information, better reflecting clinical variability, and creating a solid foundation for digital applications and AI.

In addition, the NMDS has been linked to the technical standardization HL7 FHIR from the outset, ensuring interoperability and direct technical implementability and guaranteeing connectivity to modern digital systems.

However, it should be noted that the dataset does not automatically ensure that documented data are also transferred to standards (such as FHIR). Data that are documented in other facilities and care documentation systems are not always automatically available in FHIR and may be in different formats. To counteract this limitation, a system solution, such as a data repository, is needed to automatically standardize routine data. The standardization of content and technology also facilitates the exchange of data between sectors and professions. In addition, the structural orientation toward an existing medical core dataset represents a step toward harmonizing nursing and medical data. This can support the need to give nursing terminology equivalent weight in documentation as medical terminology and integrate them instead of overwriting them [9,23].

Unlike previous approaches in LTC, which often failed to be implemented, the NMDS developed here is based exclusively on existing routine documentation to promote acceptance and long-term use in everyday nursing care without increasing the workload [34]. In accordance with extant recommendations in the literature, the development of NMDS should be undertaken in a multidisciplinary manner and integrated into (digital) documentation systems, with a focus on nursing content as opposed to medical content; thus, it should be oriented toward nursing standards, which we followed and built a base for [34,35].

The co-design approach chosen to develop the use case indicates a way of standardizing nursing data while maintaining a high degree of user centricity. This gives this NMDS an advantage over previous NMDSs and assessment instruments. In the future, a person-centered approach can be ensured by broadening the perspective to include affected persons, that is, residents of LTC facilities. The results in the NMDS table (Multimedia Appendix 5) illustrate the differences between risk factors that are perceived as relevant by practitioners and the scientific literature and those that are documented in a German geriatric care facility; this can indicate which important fall risk factors have not yet been mapped in routine documentation, such as “environmental factors” in this case. This information highlights the need for further research.

Although the results were collected nationally, findings from other European countries demonstrate difficulties in introducing standardized care documentation systems and standardization. This approach can serve as a model for the development of NMDSs for other countries [13-15].

There lies a challenge in bringing the data collected in the health care sector into a uniform format that is accessible for research. Once the data are accessible for research, AI innovations can benefit from a much larger data pool that is also very close to the point of care; this significantly improves the results of AI applications. The FHIR translation makes the data usable for AI applications that can assess and predict falls. The IG thus serves as a valuable resource for health care providers and researchers alike and provides a structured and standardized approach to capturing and analyzing fall risk factors. It is hoped that the future application of this guide will contribute to the development of more effective and personalized fall prevention strategies and ultimately improve patient outcomes. The developed IG also provides a methodological basis for expanding the minimum dataset to include other nursing topics in addition to falls as well as other care sectors and is internationally applicable.

XSL•FO
RenderX

Proof of Concept: Practical Application Example—An AI-Based Nursing Decision Support System

An AI-based nursing decision support system (NDSS) for elderly people to predict falls in nursing homes can be used as an example of an application of the content and technology developed for our NMDS.

Developed in a user-centered design process based on International Organization for Standardization (ISO) 9241-210-2019, the NDSS covers routinely documented aspects of falls in German nursing homes and reflects the most relevant fall-related factors from the literature and practice [36]. Nursing staff with different levels of education and specialized training, administrative employees from an outpatient care facility and day care, and people who worked in the field of nursing science were asked about the practical suitability of the proposed NDSS in 4 workshops. Based on the workshop results, the NDSS was improved iteratively. The decision support system is designed to predict the risk of falling using a 5-step scale and focuses on the interpretability of AI so that predictions are comprehensible and transparent for users. Therefore, the NDSS can provide information about recognized risk factors that are responsible for corresponding fall risk. The proposed decision support system also suggests prophylaxis to reduce the risk of falling and show the progression of this risk over several months.

Limitations

The results have several limitations. For reasons of feasibility, we focused on one use case. In this use case, we initially focused on a limited number of risk factors that were technically implemented. In the future, we plan to expand NMDS thematically. It must also be considered that there are different contexts, causes, and types of falls. Falls are presented in a highly standardized manner in this paper. In addition, to date, only aspects of inpatient LTC have been examined, but we plan to extend the findings to different settings.

It can be assumed that both the participants in the co-design workshops and the participants in the quantitative survey showed a certain level of interest in and affinity for the topic. As participation was voluntary, it is possible that less-interested participants did not take part due to participation bias.

Another potential bias in the quantitative study arises from the fact that participant evaluations were used as a criterion for item selection. Since documentation routines, levels of knowledge, and familiarity with existing evidence vary in everyday clinical practice, there is a risk that empirically relevant content will be underestimated or overlooked; this can lead to an overemphasis on subjective assessments, while evidence-based aspects remain potentially underrepresented. Moreover, we deliberately choose a user-centered approach to differentiate ourselves from previous solutions and ensure greater practical relevance. To reduce that bias, a broad survey covering a range of backgrounds and nursing staff was conducted. In the future, the perspectives of residents within the designated dataset should be incorporated. Furthermore, the participants' ratings were based on international nursing standards and fall prevention guidelines and were not indiscriminately determined by the participants.

Nevertheless, it is believed that this approach is effective in enabling a user-centered methodology to be adopted, thus counteracting the lack of use and implementation of NMDS and nursing terminology and standards, as outlined in the problem statement in the introduction. It is important to validate the results in future studies.

The results are oriented toward national standards, practical procedures, and documentation in German LTC facilities. However, the procedure and methodology can be applied internationally.

Conclusions

The "fall prevention" use case developed for a German NMDS in LTC shows how nursing data can be recorded in a standardized and user-centric way. The use of previously documented routine data reduces the documentation burden for nursing staff and facilitates cross-sector data exchange. This concept creates a valuable basis for nursing research by providing high-quality, structured data for secondary use. It also offers the potential for international application and can serve as a model for the development of nursing datasets in other subject areas, particularly through FHIR translation for use in AI-based applications.

The IG and the profiles it contains are to be submitted for an HL7 FHIR balloting. This structured mechanism is designed to solicit feedback from members of the HL7 community to ensure quality and relevance. The suggestions for changes and improvements collected in this process can be evaluated during further efforts and, if approved, can be integrated into the IG to promote the joint creation and further development of the NMDS.

In conclusion, our approach shows that co-design makes user-centered and applicable NMDS development feasible. Moreover, FHIR enables interoperability. This work represents a preliminary step toward harmonizing medical and nursing care in Germany by aligning the data format with MII structures, with future efforts needed to fully integrate the nursing sector. This paper enables us to adapt the approach internationally for AI-ready nursing data.

Acknowledgments

HL7, FHIR and the FHIR [FLAME DESIGN] are the registered trademarks of Health Level Seven International and their use does not constitute endorsement by HL7.

XSL•FO
RenderX

Funding

The paper was produced as part of the PFLIP (Pflege-Kerndatensatz und Intersektorales Pflegedaten-Repository [Nursing Minimum Data Set and Intersectoral Nursing Data Repository]) research project funded by the Federal Ministry of Research, Technology and Space in Germany (16SV8883).

Data Availability

The datasets generated during this study are not publicly available, as they were collected for internal project purposes; however, requests for data access may be submitted to the corresponding author and will be subject to institutional restrictions.

Authors' Contributions

SM and AS wrote the main manuscript text. NT contributed to the writing of the manuscript. DH, AB, M Klötgen, WD, AS, and SM were involved in the conception and design of the submission as well as the substantial revision of the manuscript. M Keuchel was involved in the substantial revision of the manuscript. CP, KN, YW, DH, AB, WD, AS, NT, M Klötgen, M Keuchel, and SM were involved in the acquisition, analysis, or interpretation of the data. All authors approved the submission of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Nursing minimum dataset (NMDS). Visualization of the integration of an NMDS draft into the structures of the Medical Informatics Initiative (MII) core dataset as an extension module, whereby the MII core dataset is divided into basic and extension modules. The basic modules are defined in a cross-disciplinary manner. The extension modules map data from specific applications or specialist areas.
[PNG File , 73 KB - jmir_v28i1e82417_app1.png]

Multimedia Appendix 2
Online survey.
[PDF File (Adobe PDF File), 120 KB - jmir_v28i1e82417_app2.pdf]

Multimedia Appendix 3
Example of assigning a fall risk factor to Fast Healthcare Interoperability Resources and Systematized Nomenclature of Medicine Clinical Terms codes.
[PNG File , 9 KB - jmir_v28i1e82417_app3.png]

Multimedia Appendix 4
Patient journey mapping.
[PNG File , 270 KB - jmir_v28i1e82417_app4.png]

Multimedia Appendix 5
Results: “fall prevention” use case.
[PNG File , 240 KB - jmir_v28i1e82417_app5.png]

References

1. Nagode M, Lebar L. Trends and challenges in long-term care in Europe. RSP 2019;26(2):255-262. [doi: 10.3935/rsp.v26i2.1655]
2. Expertenstandards und Auditinstrumente.: Deutsches Netzwerk für Qualitätsentwicklung in der Pflege URL: https://www.dnqp.de/expertenstandards-und-auditinstrumente/ [accessed 2025-04-16]
3. Xu Q, Ou X, Li J. The risk of falls among the aging population: a systematic review and meta-analysis. Front Public Health 2022;10:902599 [FREE Full text] [doi: 10.3389/fpubh.2022.902599] [Medline: 36324472]
4. Eckardt P, Culley JM, Corwin E, Richmond T, Dougherty C, Pickler RH, et al. National nursing science priorities: creating a shared vision. Nurs Outlook 2017;65(6):726-736. [doi: 10.1016/j.outlook.2017.06.002] [Medline: 28711216]
5. Rodgers BL, Jacelon CS, Knafl KA. Concept analysis and the advance of nursing knowledge: state of the science. J Nurs Scholarsh 2018;50(4):451-459. [doi: 10.1111/jnu.12386] [Medline: 29689127]
6. Yasin YM, Al-Hamad A, Metersky K, Kehyayan V. Incorporation of artificial intelligence into nursing research: a scoping review. Int Nurs Rev 2025;72(1):e13013. [doi: 10.1111/inr.13013] [Medline: 38967044]

XSL•FO
RenderX

7. Curtis K, Fry M, Shaban RZ, Considine J. Translating research findings to clinical nursing practice. J Clin Nurs 2017;26(5-6):862-872 [FREE Full text] [doi: 10.1111/jocn.13586] [Medline: 27649522]
8. Warnecke M, Holle D, Burmann A. Enabling data-driven nursing innovations: user-centered development of a nursing data module. Curr Dir Biomed Eng 2023;9(1):339-342. [doi: 10.1515/cdbme-2023-1085]
9. Muntlin Athlin Å. Methods, metrics and research gaps around minimum data sets for nursing practice and fundamental care: a scoping literature review. J Clin Nurs 2018;27(11-12):2230-2247. [doi: 10.1111/jocn.14155] [Medline: 29119641]
10. Thoroddsen A, Ehrenberg A, Sermeus W, Saranto K. A survey of nursing documentation, terminologies and standards in European countries. NI 2012 (2012) 2012;2012:406 [FREE Full text] [Medline: 24199130]
11. Dos Santos FC, D'Agostino F, Härkönen M, Nantschev R, Christensen B, Müller-Staub M, et al. Improving the quality of nursing care through standardized nursing languages: call to action across European countries. Int J Med Inform 2024;192:105627. [doi: 10.1016/j.ijmedinf.2024.105627] [Medline: 39306908]
12. De Groot K, De Veer AJE, Paans W, Francke AL. Use of electronic health records and standardized terminologies: a nationwide survey of nursing staff experiences. Int J Nurs Stud 2020;104:103523. [doi: 10.1016/j.ijnurstu.2020.103523] [Medline: 32086028]
13. Laukvik LB, Lyngstad M, Rotegård AK, Fossum M. Utilizing nursing standards in electronic health records: a descriptive qualitative study. Int J Med Inform 2024;184:105350 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105350] [Medline: 38306850]
14. Gesner E, Dykes PC, Zhang L, Gazarian P. Documentation burden in nursing and its role in clinician burnout syndrome. Appl Clin Inform 2022;13(5):983-990 [FREE Full text] [doi: 10.1055/s-0042-1757157] [Medline: 36261113]
15. Lindsay MR, Lytle K. Implementing best practices to redesign workflow and optimize nursing documentation in the electronic health record. Appl Clin Inform 2022;13(3):711-719 [FREE Full text] [doi: 10.1055/a-1868-6431] [Medline: 35668677]
16. De Groot K, De Veer AJE, Munster AM, Francke AL, Paans W. Nursing documentation and its relationship with perceived nursing workload: a mixed-methods study among community nurses. BMC Nurs 2022;21(1):34 [FREE Full text] [doi: 10.1186/s12912-022-00811-7] [Medline: 35090442]
17. Beikirch-Korporal E, Nolting HD, Wipp M. Dokumentieren mit Dem Strukturmodell: Grundlagen - Einführung - Management. Hannover: Vincentz Network; 2017.
18. Panka C. Pflegedokumentation entbürokratisiert: reorganisation der Altenpflege mit dem strukturmodell. In: Entbürokratisierte Pflegedokumentation. Göttingen: Hogrefe AG; 2017.
19. Pflegedokumentation stationär: das handbuch für die pflegeleitung. Bundesministerium für Familie, Senioren, Frauen und Jugend (BMFSFJ). 2015. URL: https://www.bmfsfj.de/resource/blob/79108/391791ec4d2bee12f694780cc595f845/handbuch-pflegedokumentation-data.pdf [accessed 2026-01-17]
20. Giap T, Park M, Bui LK. A comprehensive picture of using standardized nursing languages in long-term care systems: an integrative review. Int J Nurs Knowl 2025;36(2):193-208. [doi: 10.1111/2047-3095.12478] [Medline: 38951041]
21. Werley HH, Devine EC, Zorn CR, Ryan P, Westra BL. The nursing minimum data set: abstraction tool for standardized, comparable, essential data. Am J Public Health 1991;81(4):421-426. [doi: 10.2105/ajph.81.4.421] [Medline: 2003618]
22. PFLIP project. Fraunhofer ISST. URL: https://www.isst.fraunhofer.de/en/departments/healthcare/projects/PFLIP.html [accessed 2025-04-21]
23. The Medical Informatics Initiative's core data set. Medical Informatics Initiative. URL: https://www.medizininformatik-initiative.de/en/medical-informatics-initiatives-core-data-set [accessed 2025-04-21]
24. Telematikinfrastruktur. Kassenärztliche Bundesvereinigung. URL: https://www.kbv.de/praxis/digitalisierung/telematikinfrastruktur [accessed 2025-04-21]
25. Sanders EBN, Stappers PJ. Co-creation and the new landscapes of design. CoDesign 2008;4(1):5-18. [doi: 10.1080/15710880701875068]
26. Expertenstandard sturzprophylaxe in der pflege. Deutsches Netzwerk für Qualitätsentwicklung in der Pflege. 2022. URL: https://www.dnqp.de/fileadmin/HSOS/Homepages/DNQP/Dateien/Expertenstandards/Sturzprophylaxe/Sturz_2Akt_Auszug.pdf [accessed 2026-01-17]
27. Bulto LN, Davies E, Kelly J, Hendriks JM. Patient journey mapping: emerging methods for understanding and improving patient experiences of health systems and services. Eur J Cardiovasc Nurs 2024;23(4):429-433. [doi: 10.1093/eurjcn/zvae012] [Medline: 38306596]
28. Mayring P. Qualitative content analysis. In: Handbook of Qualitative Research in Psychology. Germany: VS Publishing House for Social Sciences; 2010:601-613.
29. Cho I, Kim J, Chae JS, Jung M, Kim YH. Development of ICNP-based inpatient falls prevention catalogue. Int Nurs Rev 2020;67(2):239-248. [doi: 10.1111/inr.12566] [Medline: 31762026]
30. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
31. SNOMED CT—the global clinical terminology. SNOMED International. URL: https://www.snomed.org/ [accessed 2025-04-25]

XSL•FO
RenderX

32. ICD-10-WHO: internationale statistische klassifikation der krankheiten und verwandter gesundheitsprobleme. Bundesinstitut für Arzneimittel und Medizinprodukte. URL: https://www.bfarm.de/DE/Kodiersysteme/Klassifikationen/ICD/ICD-10-WHO/_node.html [accessed 2026-01-17]
33. PFLIP-Pflegedaten HL7 FHIR implementation guide. Fraunhofer ISST. 2025. URL: https://pflip-ig.isst.fraunhofer.de [accessed 2025-04-21]
34. Milkov S, Linders L, Napetschnig A, Deiters W, Sirsch E, Holle D. Nursing minimum datasets in long-term care settings: scoping review. J Med Internet Res 2025;27:e68670 [FREE Full text] [doi: 10.2196/68670] [Medline: 41089026]
35. Freguia F, Danielis M, Moreale R, Palese A. Nursing minimum data sets: findings from an umbrella review. Health Informatics J 2022;28(2):14604582221099826 [FREE Full text] [doi: 10.1177/14604582221099826] [Medline: 35634983]
36. Tschorn N, Keuchel M, Müller I, Milkov S, Potthoff C, Windrath D, et al. Iterative design of a decision support system for fall risk detection in residential care facilities. Curr Dir Biomed Eng 2024;10(4):653-656. [doi: 10.1515/cdbme-2024-2160]

Abbreviations

AI: artificial intelligence
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level Seven International
ICD-10: International Statistical Classification of Diseases and Related Health Problems
IG: implementation guide
ISO: International Organization for Standardization
LTC: long-term care
MII: Medical Informatics Initiative
NDSS: nursing decision-support system
NMDS: nursing minimum dataset
PFLIP: Pflege-Kerndatensatz und Intersektorales Pflegedaten-Repository (Nursing Minimum Data Set and Intersectoral Nursing Data Repository)

Edited by MP Gagnon; submitted 14.Aug.2025; peer-reviewed by K Seibert, J Olsen, F Akhter; comments to author 29.Oct.2025; revised version received 17.Dec.2025; accepted 09.Jan.2026; published 17.Mar.2026.

Please cite as:
Milkov S, Schmidt A, Burmann A, Tschorn N, Klötgen M, Deiters W, Potthoff C, Neveling K, Weber Y, Keuchel M, Holle D
Advancing Nursing Data Integration Through a Nursing Minimum Dataset for the Conceptual and Technical Development of a "Fall Prevention" Data Module: Development Study
J Med Internet Res 2026;28:e82417
URL: https://www.jmir.org/2026/1/e82417
doi:10.2196/82417
PMID:

©Sarah Milkov, Antonia Schmidt, Anja Burmann, Niklas Tschorn, Marcel Klötgen, Wolfgang Deiters, Christian Potthoff, Kirsten Neveling, Yvonne Weber, Maren Keuchel, Daniela Holle. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Challenges of Standard Pediatric Epilepsy Monitoring and the Potential Benefits of Contactless Sensor Technologies: Exploratory Qualitative Study

Natalie Öhl[1], MA; Dilbar Strack[2], Dr med; Stefan Martin Kusnik[2], Dr med; Isabell Link[1], Dr rer biol hum; Marlene Hofmann[2], MSc; Tobias Steigleder[1], Dr med; Alina Weisser[1], MSc; Christoph Ostgathe[1], Prof Dr Med; Regina Trollmann[2], Prof Dr Med; Maria Heckel[1], PD, Dr phil, Dr rer biol hum

[1]Department of Palliative Medicine, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Werner-von-Siemens Str. 34, Erlangen, Germany

[2]Department of Pediatric and Adolescent Medicine, Division of Pediatric Neurology and Social Pediatrics, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Corresponding Author:
Natalie Öhl, MA
Department of Palliative Medicine, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Werner-von-Siemens Str. 34, Erlangen, Germany

Abstract

Background: Epilepsy is a common neurological condition in children, and accurate detection of seizures and their frequency is essential for diagnosis and treatment. Standard monitoring using electroencephalography alongside clinical observation is often burdensome in pediatric settings, as electrodes can cause discomfort and restrict mobility. Contactless sensor technologies may offer a promising supplement by enabling monitoring without physical contact.

Objective: This study aims to explore challenges in standard pediatric epilepsy monitoring from the perspective of health care professionals and examines the potential benefits and requirements of supplementary contactless sensor technologies in this setting.

Methods: Participant observation of routine processes in standard pediatric epilepsy monitoring was conducted at a German university hospital. Field notes from 40 observed procedures were analyzed using structuring content analysis. Building on these findings, a focus group with pediatric neurologists, nurses, and medical technical assistants (n=6) explored the potential benefits and implementation requirements of contactless sensor technologies. Focus group data were analyzed using focus group illustration maps.

Results: A reference workflow of standard pediatric epilepsy monitoring was derived, revealing psychosocial, medical, and organizational challenges faced by health care professionals. Electroencephalography recordings and clinical observation required considerable reassurance of patients and parents or carers, were vulnerable to movement artifacts and incomplete seizure documentation, and were labor- and resource-intensive. Focus group participants viewed contactless sensor technologies as a potentially valuable supplement by enabling continuous long-term monitoring with minimal additional burden.

Conclusions: By identifying challenges associated with standard pediatric epilepsy monitoring, this study provides a foundation for the needs-based development and implementation of supplementary contactless sensor technologies. Such technologies should be tailored to the clinical setting and designed to address existing burdens, with the potential to complement standard monitoring.

Trial Registration: Deutsches Register Klinischer Studien (DRKS) DRKS00027017; https://drks.de/search/de/trial/DRKS00027017

(J Med Internet Res 2026;28:e83009) doi:10.2196/83009

KEYWORDS

pediatrics; epilepsy; monitoring; technology; qualitative research

Introduction

Epilepsy is one of the most common neurological conditions in children and is characterized by an enduring risk of epileptic seizures [1]. Detecting seizures and their frequency is essential for diagnosis, treatment, and therapy monitoring [2]. The current gold standard for epilepsy diagnosis and monitoring is electroencephalography (EEG) alongside clinical observation [3,4].

XSL•FO
RenderX

EEG procedures typically take place in clinical settings and range from short-term recordings of approximately 30 minutes to long-term monitoring over several hours [5]. The electrodes attached to the patient's head often lead to discomfort and restricted mobility [5,6]. This is particularly relevant in pediatric populations, where resistance to EEG technology may further complicate the procedure. These challenges place considerable demands on health care professionals and add complexity to clinical workflows. While challenges associated with EEG recordings and long-term monitoring in general have been studied [7-9], there remains a lack of empirical knowledge on the specific challenges of standard pediatric epilepsy monitoring in clinical practice from the perspectives of the professionals involved. Qualitative studies on this topic are scarce, and observational studies remain particularly limited [10-12]. Yet, observational approaches are essential in this context, as they allow direct capture of real-world clinical processes, interaction patterns, and workflow constraints, rather than relying on retrospective self-report [13,14].

Various technologies supporting seizure detection have been developed or are under investigation, including video-based approaches, mattress sensors, and mobile applications [6,7]. These systems monitor physiological signals relevant to epilepsy, such as cardiovascular changes, respiration, electrodermal activity, motor activity, audio, eye movements, temperature changes, body pressure, and moisture [6]. Although many are noninvasive, they are typically contact-based [6] and may be difficult to tolerate for children, who cannot meaningfully weigh potential benefits against discomfort and may instinctively try to remove them. Therefore, contactless sensor technologies may be particularly promising in pediatric epilepsy monitoring [6] by reducing discomfort and improving adherence.

One example of a contactless approach is a new type of radar technology based on interferometry [15], currently being developed within federally funded research projects, such as BrainEpP (BMBF: 13GW0295D) and EmpkinS (DFG: CRC 1483 - 442419336) [16-18]. Positioned near the child's chest, for example, under the bed, the radar continuously records the patient's heartbeat [19-21]. Changes in heart rate and heart rate variability (HRV) are known to occur before and during epileptic seizures [22]. Machine-learning algorithms derive these parameters from radar motion signals [23] and use this information to support seizure detection. In adult populations, the heart rate calculated on the basis of radar data is of high accuracy when compared to the electrocardiogram [16,24,25].

Despite growing interest in contactless seizure monitoring, empirical evidence on how health care professionals perceive their benefits, limitations, and integration into pediatric clinical practice remains limited. Their perspectives are essential to ensure usability, feasibility, and meaningful clinical benefit [6].

To deliver clinical value, innovative technologies such as contactless sensors should be compatible with real-world clinical workflow and constraints [26]. This requires first identifying challenges in current pediatric epilepsy monitoring practices and, building on these insights, understanding how health care professionals envision technological support addressing these challenges. Such a sequential and integrated approach is relevant for the responsible development and implementation of new technologies [27].

Accordingly, this study addresses two research gaps: (1) it explores challenges in standard pediatric epilepsy monitoring at a university hospital epilepsy center from the perspective of health care professionals, to identify opportunities for workflow improvements and technological support; and (2) based on these insights, it examines the potential benefits and requirements of supplementary contactless sensor technologies in this clinical setting from the perspective of health care professionals.

Methods

Study Design

We used a qualitative design, informed by a constructivist epistemological approach, drawing on principles of ethnography to understand the complex social and organizational processes in pediatric epilepsy monitoring. Between November 2021 and January 2022, we conducted ethnographic fieldwork [13,28] using participant observation [13,14] to examine routine processes in standard pediatric epilepsy monitoring. This method allowed us to capture real-world workflows, interactions, and challenges in context. Building on these observations, we conducted a focus group [29] with health care professionals in June 2022 to explore the potential benefits and requirements of supplementary contactless sensor technologies (Figure 1). The study process adhered to COREQ (Consolidated Criteria for Reporting Qualitative Studies) guidelines (Checklist 1) [30], ensuring transparency and rigor in reporting qualitative research.

XSL•FO
RenderX

Figure 1. Synthesized methods used in study design.

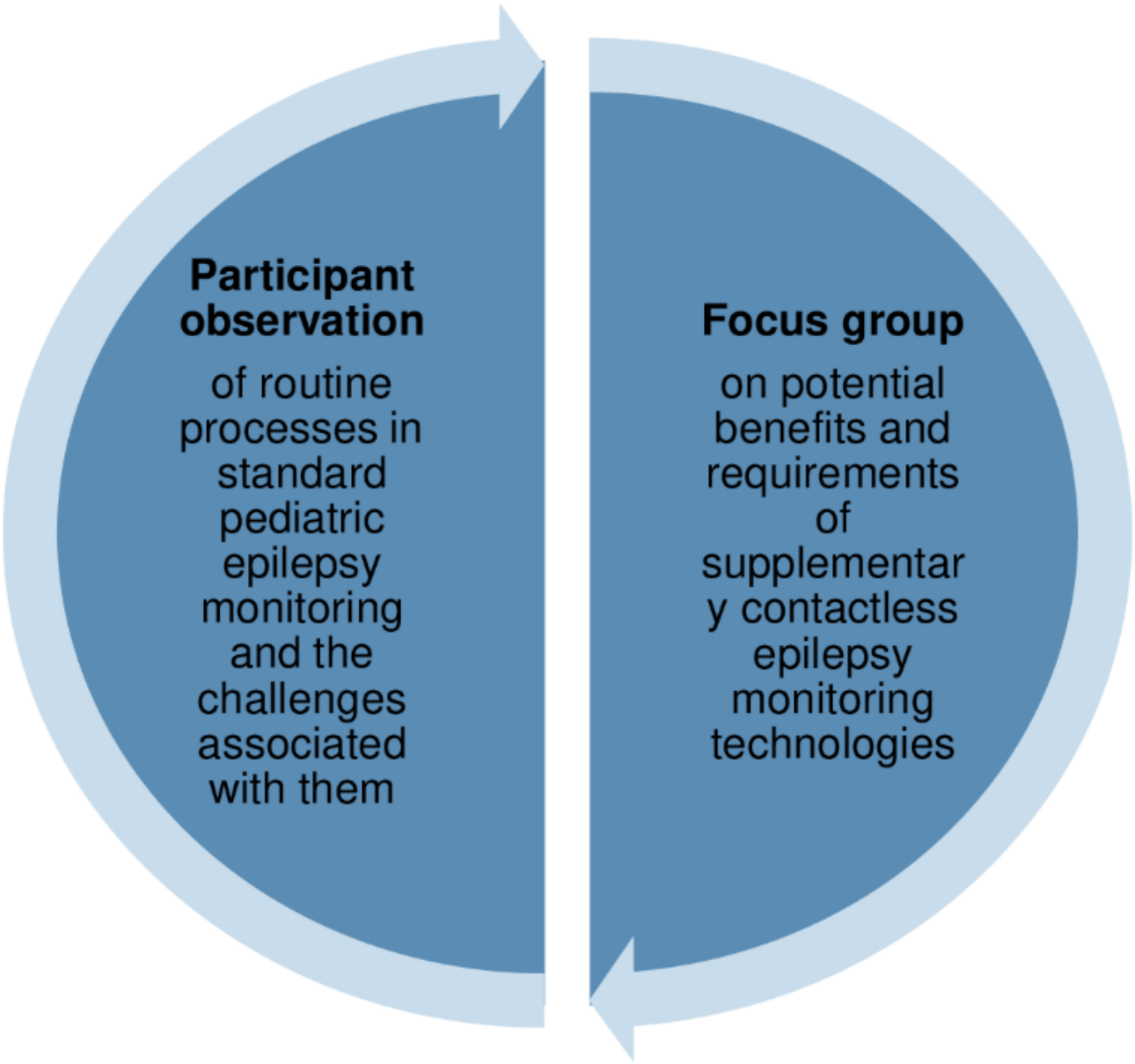

Setting

The data were collected in the pediatric epilepsy center of the Division of Pediatric Neurology at University Hospital Erlangen, Germany. EEG recording (using the 10 - 20 system) adhered to the guidelines issued by the DGKN (German Society for Clinical Neurophysiology and Functional Imaging) [31].

Research Team and Reflexivity

Participant observation and focus group were carried out by IL, a study coordinator, research associate in the Department of Palliative Medicine, and medical technician, in consultation with MaHo, a research associate in the Division of Pediatric Neurology. Prior to conducting the observational study, IL received training in participant observation from an experienced researcher.

Participant Observation

Data Collection

With informed consent, the actions and standard procedures of health care professionals (pediatric neurologists, nurses, and medical technical assistants, MTAs) were observed during routine pediatric epilepsy monitoring. This included diagnostic, long-term, and seizure-detection EEG as well as clinical observation. We chose situations for observation in consultation with the nurse and MTA in charge of the relevant department.

The observer (IL) was not known to the health care professionals beforehand; she introduced herself and gave brief details on the study at the beginning of each situation of interest. The observation included routine EEG recording processes, working environment (structures and organization of the ward; furnishings, equipment, and facilities in patient rooms), type of diagnostic methods, devices, and materials used, and how they were used, roles and responsibilities of the stakeholders involved, communication and collaboration processes,

XSL•FO
RenderX

information provision and decision-making, and duration of the monitoring. A guide detailed the components and the focus of the observation (Multimedia Appendix 1). Throughout the observations, health care professionals explained procedures and provided information about processes and experiences. Besides, conversations that arose spontaneously with health care professionals were also documented in the field notes. Between November 2021 and January 2022 (on 8 days, 4 hours each), a total of 40 situations were observed. Documentation was done with field notes and pictures of the technical facilities, with the pictures showing only the equipment and no children or staff. Data collection continued until saturation was reached in terms of varied and typical pediatric epilepsy monitoring situations.

Data Analysis

On the basis of the field notes, IL created 40 pseudonymized descriptions of the situations and imported them to MAXQDA [32], setting document variables. We used the MAXQDA software (version 20; VERBI Software) for the analysis. Each description comprised details in reference to each component of the observation guide (Multimedia Appendix 1) relating to the care of 1 patient with epilepsy. We ensured interobserver reliability by having multiple researchers independently analyzing the field notes, thereby minimizing potential bias. NO, IL, and MH conducted structured content analysis using content categories as defined by Mayring [33,34]. The units of analysis were determined by the description of situations observed. Coding took place using a subsumption strategy, assigning data to main categories and subcategories. We used the components of the observation guide as deductive main categories and then inductively added further main and subcategories (Table 1).

Table . Coding scheme for descriptions of situations observed.

Main categories	Subcategories (inductive)
Routine processes (deductive)	• Patient admission • Medical information and informed consent • Patient logistics • EEG[a] electrode placement • EEG recording • Post EEG procedures • EEG interpretation • Working routines on the ward
Information sharing (deductive)	• Information sharing between parents or carers or patients and pediatric neurologists • Information sharing between nurses and pediatric neurologists • Documentation
Human-technology interaction (inductive)	• Handling movement artifacts and reapplying EEG electrodes • Setting markers in EEG recordings
Human-human interaction (inductive)	• Instructing and calming patients during EEG electrode placement and recording • Instructing and calming parents or carers during clinical observation
Resource requirements (inductive)	• Human resources • Consumables • Time • Facilities

[a]EEG: electroencephalography.

Coding was done by one coder (NO) (Figure 2). Due to the heterogeneity of the situations observed, the coder carried out deductive coding of 100 percent of situations before adjusting the system of categories and performing inductive coding. Three feedback loops, involving NO, MH, and IL, took place to enable us to attain the maximum possible intersubjectivity and to verify the categories. The adaptation of the coding scheme was an iterative process that went through the material multiple times. To minimize intersubjective differences in interpretation, we drew up coding guidelines ensuring our codes were clearly defined and included anchor examples. A reference workflow emerged on the basis of the main categories and subcategories; we had a pediatric neurologist (DS) check the workflow for correctness and completeness.

XSL•FO
RenderX

Figure 2. Data analysis using the method of structuring content analysis set out by Mayring [33,34].

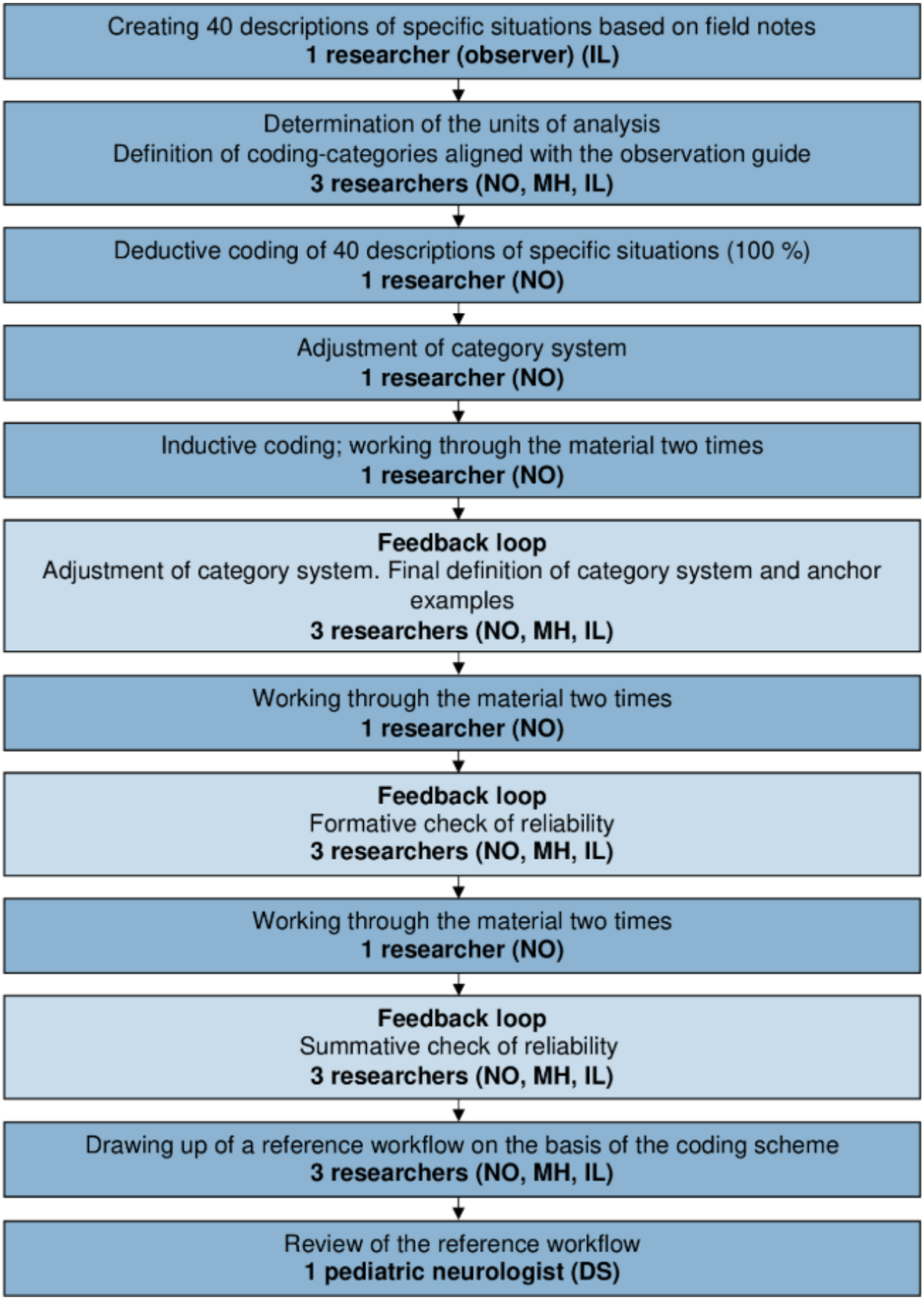

XSL•FO
RenderX

Focus Group

A focus group discussion with health care professionals was performed to explore the potential integration of supplementary contactless sensor technologies into pediatric epilepsy monitoring, including their perceived benefits and requirements.

Recruitment of Participants

Health care professionals of the Division of Pediatric Neurology, University Hospital Erlangen, all directly involved in the care of pediatric patients with epilepsy, were invited. Participants included pediatric neurologists (n=4), a nurse (n=1), and an MTA (n=1), thereby covering all relevant perspectives and offering a sample of the broader population of epilepsy care providers. One of the authors (MaHo) approached prospective participants, whom we had specifically identified as meeting these criteria, face to face; IL provided information to those who showed interest in taking part. All those approached gave written consent to participate. The researcher conducting the focus group (IL) was an external research associate who was fleetingly personally known to the participants.

Data Collection

We created and piloted a focus group guide, informed by preliminary results from the participant observation. The facilitator (IL) received prior training to conduct the focus group. To support the facilitator, MaHo and AW were present during the focus group. The focus group took place in June 2022 at University Hospital Erlangen, with a duration of 60 minutes and was audio-recorded. We printed out key questions in advance of the discussion and made them available to all participants. The focus group commenced with an outline of the research topic and the aims of the research. To support the discussion, we gave a short presentation on an exemplary contactless sensor technology for supporting epilepsy monitoring, involving a flyer detailing the technology. Specifically, we introduced a new type of radar that is currently under development in the research project BrainEpP (BMBF: 13GW0295D).

Data Analysis

We analyzed the focus group using focus group illustration maps, a method that integrates knowledge mapping principles with focus group procedures [35]. Focus group illustration maps combine data collection and analysis in one unified, structured, and documented process.

Creating the Maps During the Focus Group

During the focus group, one researcher (AW) documented participants' statements in real time on a flipchart. Statements were summarized, assigned to thematic clusters, and placed as nodes within a network; connecting lines indicated conceptual links among statements. This graphical documentation supported structuring and condensation of participants' contributions, preserved relational connections between arguments, and enabled immediate consensual validation, as participants could verify, correct, or expand written content on the spot. By the end of the focus group, a structured and participant-validated visual summary existed for each guiding question.

Post-Session Refinement

After the focus group, the flipchart graphics were transferred into Miro Board [36], and digital illustration maps were created for each key question. Two researchers (NO and AW) independently reviewed the maps, comparing them with the audio recordings to ensure completeness, correctness, and logical consistency.

As a final step, 2 researchers (NO and IL) aligned the results of the participant observation with the focus group illustration maps to identify connections between the 2 data sources.

Ethical Considerations

The study was approved by the Ethics Committee of the Faculty of Medicine at Friedrich-Alexander-Universität Erlangen-Nürnberg (350_21 B) on September 20, 2021. The trial was registered with Deutsches Register Klinischer Studien (DRKS; DRKS00027017) on November 16, 2021. All participants received full information on the study and on their right to withdraw consent to participate at any time without giving reasons and were able to ask questions prior to the participant observation and the focus group. All participants gave informed consent to participate in the study. All data were pseudonymized prior to analysis, ensuring that no information could be linked back to individuals, and access to the pseudonymized dataset was restricted to authorized members of the research team only. Participants did not receive any form of compensation for their involvement in the study.

Results

Reference Workflow

To contextualize the challenges identified in standard pediatric epilepsy monitoring within a university hospital epilepsy center, we first present a reference workflow that provides an overview of the process. The coding scheme of the participant observation (Table 1) led to this reference workflow (Figure 3). It outlines the clinical stages of pediatric epilepsy monitoring, the stakeholders involved, the work steps, information flows, resources required, and the interactions that characterize routine practice. The process varies depending on the type of EEG (eg, routine vs long-term video EEG procedure) and whether it takes place on an inpatient or outpatient basis. Routine EEGs entail at least 20 minutes of recording, whereas long-term EEGs require approximately 24 hours of recording in the inpatient clinic.

Within this reference workflow, psychosocial, medical, and organizational challenges in standard pediatric epilepsy monitoring procedures encountered by health care professionals were identified (Figure 4). These challenges highlight where workflow improvements or supportive technologies could be most beneficial.

XSL•FO
RenderX

Figure 3. Reference workflow for standard pediatric epilepsy monitoring in a university hospital epilepsy center based on observations of routine processes (own illustration). EEG: electroencephalography.

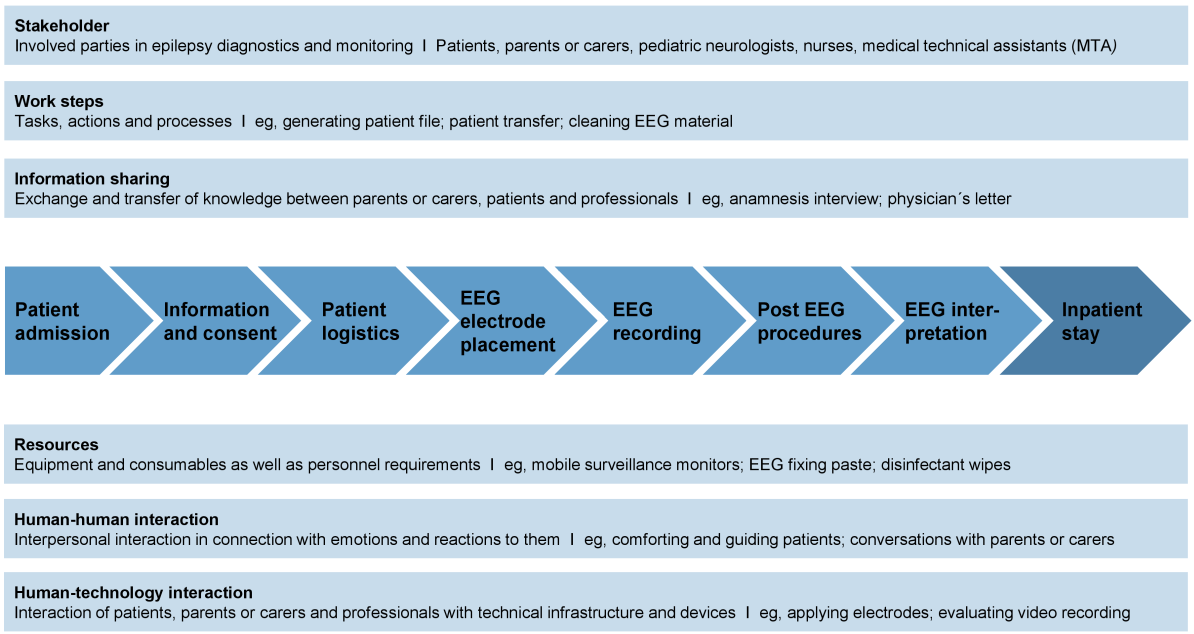

Figure 4. Identified challenges in standard pediatric epilepsy monitoring in a university hospital epilepsy center. EEG: electroencephalography.

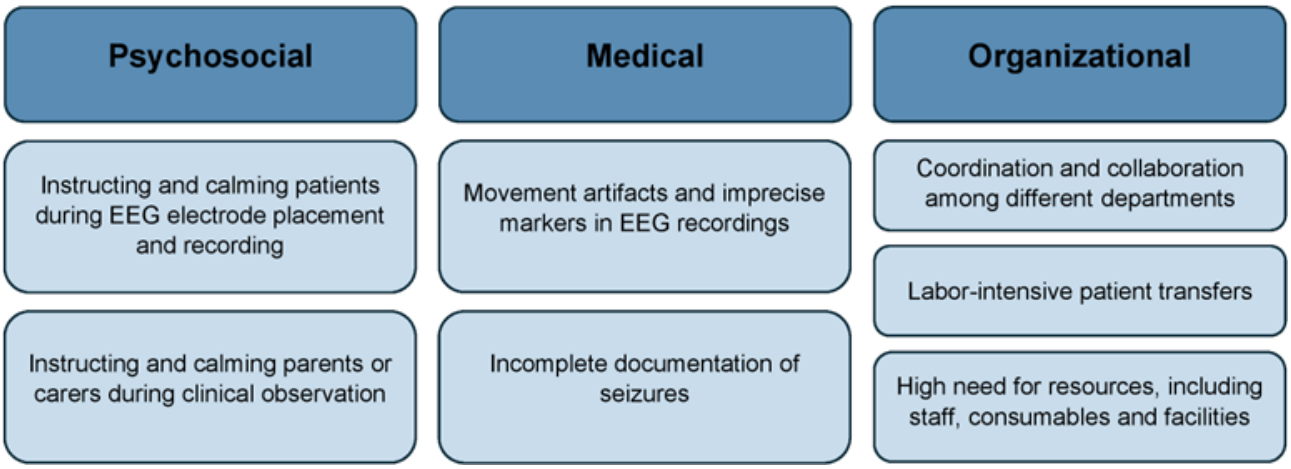

Psychosocial Challenges

Instructing and Calming Patients During EEG Electrode Placement and Recording

During the observation, health care professionals explained that patients must remain still during the application of electrodes and the EEG recording, as physical movement can cause artifacts (see "Medical challenges" section below). While many patients tolerated the equipment well, some showed signs of stress or fear, such as screaming, crying, and irregular breathing:

> *The child starts to wheeze and shows irregular breathing (holding her breath, jerky exhalations). She taps/drums her hands on the examination table cushion.* [S24]

Some resisted the recording technology by shaking their heads or turning away. In several observational situations, MTAs involved in the workflow instructed patients and parents or carers (individuals actively involved in care who have legal parental responsibilities or legal guardianship) to ensure the technology remained in place. Patients with disabilities, limited communication abilities, or cognitive impairments occasionally had difficulties following the MTAs' instructions. MTAs tried to comfort patients by interacting with them empathetically, using elements of play and physical contact.

> *Throughout the examination, the medical technical assistant is slightly bent over the child, her arm rests reassuringly on the child's upper body and she tells a comforting story.* [S22]

MTAs also encouraged parents or carers to calm or distract their children during this process. This was very demanding for

XSL•FO
RenderX

MTAs, who had to cope with children crying and screaming while maintaining patience and empathy.

Instructing and Calming Parents or Carers During Clinical Observation

In addition to EEG recordings, clinical observation of seizure activity is essential. Health care professionals request parents or carers to document seizures characteristic of the patient's epilepsy. During the observation, health care professionals noted the generally high motivation of parents or carers to engage in this process. However, they also acknowledged that it represents an additional responsibility for parents or carers in an already stressful situation. Health care professionals explained that parents or carers often experience pressure to do everything "right," although they are not medically trained, and are acutely aware that their child's diagnosis may depend on the quality of their observations. Health care professionals explained that they face the challenge of calming parents or carers in this demanding situation while ensuring that the documentation is done correctly. This can be further complicated in cases involving language barriers:

> *Due to the limited language proficiency of some non-native parents, they do not fully understand their tasks and have difficulties knowing what and how to document.* [S13]

Medical Challenges

Movement Artifacts and Imprecise Markers in EEG Recordings

During the observation, patients' movements in the course of EEG recordings led to artifacts and thus reduced the recordings' informational value in medical terms.

> *The patient is extremely tense, resulting in an EEG that is heavily affected by artifacts.* [S10]

When patients resisted the procedure, electrodes and probes became detached, fell off, and required reapplication; this interrupted the course of monitoring and reduced the overall time it covered. MTAs sometimes struggled to obtain analyzable recordings. They documented movement artifacts that occurred by placing a marker in the recording; the necessity of paying simultaneous attention to several parallel tasks occasionally resulted in markers being set imprecisely.

Incomplete Documentation of Seizures

The documentation of epileptic seizures by health care professionals and parents or carers is of crucial significance. However, the observation showed that its completeness was occasionally limited due to the difficulties that even specialists experience in explaining different types of seizures to laypeople. Health care professionals explained that parents or carers lack professional knowledge, which impedes correct documentation. Language barriers were an additional source of misunderstanding. The information loss may impact treatment:

> *Information on seizures is completely missing in the template and the further diagnosis has to be made without this additional information.* [S13]

Health care professionals explained that specialized epilepsy centers typically video-record long-term EEGs, which enables pediatric neurologists to examine the video material when parents or carers have documented a seizure incompletely. However, it is extremely time-consuming, and not all seizure patterns can be identified.

Organizational Challenges

The observation showed that epilepsy monitoring in a pediatric epilepsy center is a labor- and resource-intensive procedure that makes heavy demands on the management of procedures and routines, including appointments and patient transport for EEG recordings.

Coordination and Collaboration Among Different Departments

Coordination of appointments for inpatients took place via telephone conversations with the ward. When the ward was busy or an emergency arose, delays could cause inpatients to arrive late. If a patient was late and the monitoring room had been prepared specifically for this patient, it was usually not possible to fill the gap with another patient; the delay caused delays to other patients and occasionally resulted in postponed appointments. This dynamic situation required staff to respond flexibly and make rapid decisions on the order of patients and who to prioritize.

> *An MTA repeatedly calls the ward and asks whether the patient has already been admitted and explains that all preparations have been made (...). Referring to the digital schedule of appointments, two MTAs begin to discuss which patients' appointments could be moved forward during the waiting period.* [S15]

Labor-Intensive Patient Transfers

The clinical processes vary based on which EEG is required (routine or long-term). Health care professionals explained that, depending on the process used for EEG electrode placement and recording, patients need to be transferred between the pediatric neurology ward, the outpatient clinic, and, on occasion, the EEG monitoring room.

> *Patient transport depends on the patient's health status, prior experience, and age, and may require one or two medical assistants, or in special cases with high seizure frequency a licensed nurse.* [S34]

Patient transfers took up a substantial amount of time, reducing flexibility in patient management and contributing to delays.

High Need for Resources, Including Staff, Consumables, and Facilities

The process of standard epilepsy monitoring engages a large amount of resources, including staff, consumables, and facilities. The observation showed that the preparation and application of EEGs were particularly labor-intensive, depending on the patient's ability to cooperate. Consumables needed included abrasive paste and cotton swabs to clean the scalp, conductive or fixing paste for attaching the electrodes, and, in some cases, a syringe to insert conductive paste into the electrode caps; general hygiene material will also be required:

XSL•FO
RenderX

> *The medical assistant first cleans the patient's scalp with the abrasive paste and a cotton swab and then attaches the cup electrode to the scalp with the fixing paste.* [S16]

Long-term EEGs required specially equipped monitoring rooms. Transfer of patients from the ward to these rooms required mobile surveillance monitors and mobile blood pressure monitors.

Potential Benefits of Supplementary Epilepsy Monitoring With Contactless Sensor Technologies

In the participant observation, challenges of standard pediatric epilepsy monitoring from the perspectives of health care professionals were identified. In the subsequent focus group, health care professionals discussed the potential benefits and requirements of using contactless sensor technologies for supplementary monitoring.

Overall, participants discussed several aspects that need to be considered when implementing such technology. They emphasized that training is required not only for clinical staff but also for patients and parents or carers. Without proper instruction, additional workload may arise for clinical staff due to questions from parents or carers. Participants suggested that refresher training should be provided after some time, along with a re-evaluation to gather feedback on needs and any problems encountered. There should also be a clear point of contact for issues related to the use of the technology.

> *Training and a clear contact person are essential to avoid extra workload and ensure smooth use.*

The responsibilities of clinical staff need to be clearly defined. Participants noted that implementing this technology requires additional effort, and clinics must allocate sufficient time and resources to support this new task. Depending on the type of contactless technology and the physiological signals being monitored, participants noted that the constraints and technical limitations must be carefully evaluated and considered during implementation.

Continuous Long-Term Epilepsy Monitoring Data

Participants described EEG as a significant, long-established, and irreplaceable method for epilepsy diagnosis. However, contactless sensor technologies were seen as a useful complement to the gold standard of epilepsy diagnosis (EEG and clinical observation), in providing long-term monitoring data. Participants emphasized that such technology can only complement, but not replace, EEG.

Sensors work without requiring direct physical contact with the patient, making monitoring easier and less labor-intensive for staff. Specifically, the radar-based technology introduced might allow monitoring biomarkers such as respiration and HRV. Participants discussed its potential to detect seizures and measure their duration, frequency, and type. They argued that continuous monitoring data over a longer period could support diagnosis and guide treatment more effectively than it does currently.

> *With continuous monitoring, we could gather information to better understand seizure patterns, manage treatment, and assess how medications are working.*

The exemplary introduced radar-based technology, which can detect changes in HRV up to 3 minutes before the onset of the seizure [20]. Participants see a benefit in an alarm that could be programmed to sound when a change in biomarkers takes place, enabling timely intervention. However, they noted that there is a risk of false alarms or frequent alerts even for minor seizures, so the system needs to be carefully calibrated.

> *There is a risk of constant alarms and continuous beeping, which could cause stress for parents and clinical staff, and place a burden on them.*

Freedom of Movement for Patients During Monitoring

The potential for contactless sensor technologies to avoid restricting patients' freedom of movement during monitoring was highlighted. This could help prevent additional emotional distress, especially to pediatric patients, and thus be relieving for the health care professionals involved. Specifically, for the introduced radar technology, participants noted that installing it in the patient's room or near their bed allows patients to move freely within the radar's field of view, thereby enhancing patient comfort.

> *Compared to other monitors, it's contactless, making it easier to handle with patients causing less physical strain on them.*

Security for Parents or Carers Observing Seizures

It was suggested that it would be valuable if contactless sensor technologies could automatically document seizures and trigger an alarm. Participants emphasized that parents or carers will continue to play an essential role in pediatric epilepsy diagnosis and monitoring by observing and documenting epileptic seizures; however, such technologies might have the potential to provide them with support and greater certainty in this endeavor. Particularly at night, sensors could reduce anxiety about missing a seizure and reduce the pressure of constant vigilance.

> *With this technology, we don't have to rely solely on parents to spot every seizure. It helps make things more objective, reduces anxiety about missing a seizure, and allows parents to sleep more peacefully, even when they are not present.*

Supplementary Monitoring in the Outpatient Setting

Focus group participants highlighted that contactless sensor technologies could be particularly suitable for monitoring at home. They suggested that long-term epilepsy monitoring data outside the clinical setting could open up new perspectives for epilepsy management. Participants noted that nonintrusive epilepsy monitoring in the child's home may improve adherence to monitoring and reduce the stigmatization associated with frequent hospital visits. However, acceptance among patients and parents or carers in the outpatient setting was discussed. Participants emphasized that it may depend on the child's age and on alarm settings, suggesting that alarms might be most useful for longer seizures. They also highlighted several requirements for outpatient use, including reliability, quick and

XSL•FO
RenderX

easy operation, clear instructions and training, and the availability of a contact person in case of problems, especially during the night.

> *A hotline at night for technical problems could be very helpful.*

Discussion

Principal Findings

Participant observation revealed a broad set of psychosocial, medical, and organizational challenges faced by health care professionals during standard pediatric epilepsy monitoring in a university hospital epilepsy center. Building on these findings, the focus group highlighted how contactless sensor technologies might address selected aspects of these challenges. Overall, such technologies were consistently perceived as a complementary rather than replacement tool for EEG and clinical observation, aligning with literature emphasizing the value of multimodal approaches in epilepsy monitoring [6,7].

Psychosocial Perspective

Standard pediatric epilepsy monitoring imposes considerable psychosocial demands on health care professionals, who must continuously instruct, reassure, and calm patients and their parents or carers during EEG procedures and clinical observation. This sustained emotional labor contributes to workload and stress in an already demanding clinical environment, with potential implications for care quality and staff burnout [37]. Here, the focus group shows that contactless sensor technologies' value lies primarily in avoiding additional strain rather than reducing existing burdens.

Observations also highlighted stress among parents or carers due to the fear of missing seizure events, a burden well documented in the literature [10]. Participants suggested that supplementary contactless monitoring could increase families' sense of security by enabling continuous surveillance, while clearly emphasizing that clinical observation remains essential.

Medical Perspective

Participant observation identified challenges in standard pediatric EEG procedures, including movement artifacts, imprecise seizure markers, and electrode displacement, particularly in children who have difficulty tolerating electrodes or remaining still. Incomplete seizure documentation further compromised the quality and diagnostic yield, with direct implications for clinical decision-making [38,39]. Focus group participants perceived contactless sensor technologies as potentially valuable for providing continuous long-term data to supplement standard monitoring by capturing additional physiological correlates of seizures and informing clinical decisions. At the same time, such technologies have inherent technical and medical limitations, including susceptibility to motion artifacts, limited seizure-type differentiation [40], and the absence of direct cortical signal acquisition. These constraints vary depending on the underlying physiological signals and algorithms, reinforcing the role of contactless sensors as an adjunct rather than an alternative to EEG. Ensuring technical accuracy, patient comfort, and continuous availability is therefore essential and requires further investigation [7]. While this study primarily explored perceived benefits and requirements related to the noncontact design of such technologies, further research is required to evaluate their medical relevance and performance across different use cases, sensor types, and physiological signals. In line with established clinical EEG practice, the integration of supplementary contactless monitoring technologies requires clearly defined minimum standards to ensure patient safety and data quality [38-40]. These include reliable signal accuracy, continuous availability, clear assistance procedures, predefined contraindications, transparent communication of benefits and risks, and well-defined indications for specific patient groups and seizure types [38-40]. Our findings suggest that addressing these requirements is an important implication to ensure that supplementary contactless technologies meaningfully support clinical care without compromising established neurophysiological standards.

Organizational Perspective

Standard pediatric epilepsy monitoring was found to be labor- and resource-intensive, requiring high coordination across departments and staff-intensive patient transfers. Participants did not view contactless sensor technologies as a solution to these structural challenges, as standard monitoring procedures would remain necessary. Organizational burdens are therefore more likely to be addressed through workflow or structural changes or through alternative technical or nontechnical innovations [41-44].

Potential for Outpatient Monitoring

Although this study focused on inpatient monitoring, participants noted the potential relevance of contactless technologies in outpatient settings. This aligns with evidence that long-term home monitoring can complement diagnosis and treatment planning [45], particularly given the variability and underreporting of seizures by patients and parents or carers [46]. While prior studies have explored patient and caregiver preferences for seizure detection devices in outpatient settings [47], pediatric-specific research on home-monitoring burdens and acceptance remains limited. Future research should examine the feasibility, perceived benefits, and user acceptance of contactless technologies among pediatric patients and their families, building on similar work from adult epilepsy care with wearable technologies [48].

Strengths and Limitations

Research on new digital technologies is challenging and requires methodological innovations [27,49,50]. In this qualitative study, we applied a sequential, integrated design, using participant observation to identify challenges in standard pediatric epilepsy monitoring, followed by a focus group to explore the potential role of supplementary contactless sensor technologies. This ensured that technological innovation is grounded in real-world clinical needs and constraints [26].

A key strength of this study is the use of participant observation conducted by an external researcher, supplemented by verbal explanations from the health care professionals involved, which provides a more objective perspective than relying solely on

XSL•FO
RenderX

verbal accounts from those involved in the procedures [13,14]. Observations have proven particularly relevant in the context of health technologies [51], as they allow researchers to move beyond subjective intended use and capture actual practices. However, the observer's presence and concurrent explanations may have influenced behavior.

Subjective bias was mitigated through a structured observation guide and by involving an external researcher rather than a member of the pediatric neurology department. Additionally, the observer's background in medical engineering may have shaped attention toward technical aspects, which we tried to balance by following the observation guide and consulting with a neuropediatrician when needed.

The focus group captured expert perspectives from health care professionals directly involved in pediatric epilepsy care, ensuring high clinical relevance. However, the exclusively professional perspective and the predominance of pediatric neurologists may have influenced the findings. Future studies should therefore include a broader range of health care professionals, as well as patients and parents or carers, to capture diverse experiences and expectations.

This study was conducted at a single tertiary hospital in Germany, where resources, staffing, and EEG logistics may differ from other settings. Potential differences in such contextual factors and internal processes make it impossible to generalize the findings completely or apply them directly to other epilepsy centers; however, we may safely assume that the challenges identified can reasonably be expected in comparable settings. Future research should include multisite studies across different care contexts (eg, outpatient or non-university hospital settings) to capture a broader range of contextual factors.

Conclusions

We have found standard pediatric epilepsy monitoring in a university hospital epilepsy center to entail psychosocial, medical, and organizational challenges for health care professionals. Having identified these, we consider that we have created an important foundation for the development of new supportive contactless sensor technologies for this setting. Any such new technology needs to fit into the specific care setting and seek to address the existing challenges and burdens in order to deliver a needs-based benefit. The capacity of contactless sensor technologies to monitor seizures without physical contact with the patient suggests that they have great potential to support standard pediatric epilepsy monitoring in both in- and outpatient settings with long-term monitoring data.

Acknowledgments

The present work was performed by Natalie Öhl in (partial) fulfillment of the requirements for obtaining the degree "Dr. rer. biol. hum" at the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The authors extend their thanks to Sandra Grimminger for providing training in participant observation to Isabell Link. The authors thank the project partners in BrainEpP for collaboration: Geratherm Respiratory GmbH, DeMeTec GmbH, Voigtmann GmbH, Silicon Radar GmbH, Hamburg University of Technology (TUHH): Institute for High Frequency Technology, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU): Department of Electrical Engineering, Department of Criminal Law, Criminal Procedural Law, Commercial Criminal Law, and Medical Criminal Law. The manuscript was language-edited by the Language Service of the Friedrich-Alexander-Universität Erlangen-Nürnberg. The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing. The GenAI tool used was ChatGPT 4.5. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was funded by the German Federal Ministry of Education and Research (BMBF) (13GW0295D) and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant SFB 1483 – Project-ID 442419336.

Data Availability

The datasets used and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: CO, RT, and TS initiated the project and developed the overall study design for BrainEpP.
Formal analysis: NÖ conducted the analysis and interpretation of the data. IL and MH contributed substantially to the data analysis and interpretation. DS, TS, CO, and RT provided their medical expertise to support the interpretation of the data.
Funding acquisition: CO, RT, and TS wrote the funding proposal.
Investigation: IL and MaHo collected the study data. MaHo was responsible for participant recruitment for the focus group and for coordinating the participant observation.
AW assisted with data collection during the focus group. DS and SK supported the data collection for the participant observation.
Methodology: MH and IL developed the study protocol for the observational field study and the focus group. DS, SK, TS, CO, and RT contributed their medical expertise to the study design.
Project administration: IL and MaHo coordinated the study activities.

XSL•FO
RenderX

Supervision: MH, CO, and RT supervised the study.
Validation: NÖ, MH, and IL verified the consistency and reproducibility of the study results and research outputs.
Visualization: NÖ created visualizations.
Writing – original draft: NÖ wrote the original manuscript draft.
All authors thoroughly reviewed the manuscript draft.

Conflicts of Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Multimedia Appendix 1
Observation guide.
[DOCX File, 15 KB - jmir_v28i1e83009_app1.docx]

Checklist 1
COREQ Checklist.
[PDF File, 429 KB - jmir_v28i1e83009_app2.pdf]

References

1. Fisher RS, Acevedo C, Arzimanoglou A, et al. ILAE official report: a practical clinical definition of epilepsy. Epilepsia 2014 Apr;55(4):475-482. [doi: 10.1111/epi.12550] [Medline: 24730690]
2. Hartl E, Noachtar S. EEG: Stellenwerte und technische Voraussetzungen der EEG-Untersuchung [Article in German]. Klin Neurophysiol 2018 Feb;49(1):48-55. [doi: 10.1055/s-0043-124742]
3. Rosenow F, Klein KM, Hamer HM. Non-invasive EEG evaluation in epilepsy diagnosis. Expert Rev Neurother 2015 Apr;15(4):425-444. [doi: 10.1586/14737175.2015.1025382] [Medline: 25779862]
4. Bandopadhyay R, Singh T, Ghoneim MM, et al. Recent developments in diagnosis of epilepsy: scope of MicroRNA and technological advancements. Biology (Basel) 2021 Oct 25;10(11):1097. [doi: 10.3390/biology10111097] [Medline: 34827090]
5. Marathe M. Inpatient experiences of long-term monitoring: the case of EEG. : ACM; 2024 Nov 11 Presented at: CSCW: '24The 27th ACM Conference on Computer-Supported Cooperative Work and Social Computing; Nov 9-13, 2024; San Jose Costa Rica p. 511-518. [doi: 10.1145/3678884.3681899]
6. Van de Vel A, Cuppens K, Bonroy B, et al. Non-EEG seizure detection systems and potential SUDEP prevention: state of the art: review and update. In: Huffel S, editor. Seizure 2016 Oct;41:141-153. [doi: 10.1016/j.seizure.2016.07.012] [Medline: 27567266]
7. Ramgopal S, Thome-Souza S, Jackson M, et al. Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy. Epilepsy Behav 2014 Aug;37:291-307. [doi: 10.1016/j.yebeh.2014.06.023] [Medline: 25174001]
8. Sansevere AJ, Duncan ED, Libenson MH, Loddenkemper T, Pearl PL, Tasker RC. Continuous EEG in pediatric critical care: yield and efficiency of seizure detection. J Clin Neurophysiol 2017 Sep;34(5):421-426. [doi: 10.1097/WNP.0000000000000379] [Medline: 28430674]
9. Sullivan JE, Corcoran-Donnelly M, Dlugos DJ. Challenges in pediatric video-EEG monitoring. Am J Electroneurodiagnostic Technol 2007 Jun;47(2):127-139. [doi: 10.1080/1086508X.2007.11079619] [Medline: 17679580]
10. van Westrhenen A, de Lange WFM, Hagebeuk EEO, Lazeron RHC, Thijs RD, Kars MC. Parental experiences and perspectives on the value of seizure detection while caring for a child with epilepsy: a qualitative study. Epilepsy Behav 2021 Sep 28;124:108323. [doi: 10.1016/j.yebeh.2021.108323] [Medline: 34598099]
11. Harden J, Black R, Chin RFM. Families' experiences of living with pediatric epilepsy: a qualitative systematic review. Epilepsy Behav 2016 Jul;60:225-237. [doi: 10.1016/j.yebeh.2016.04.034] [Medline: 27240309]
12. Chong L, Jamieson NJ, Gill D, et al. Children's experiences of epilepsy: a systematic review of qualitative studies. Pediatrics 2016 Sep;138(3):e20160658. [doi: 10.1542/peds.2016-0658] [Medline: 27511947]
13. Hammersley M, Atkinson P. Ethnography: Principles in Practice, 4th edition: Routledge; 2019. [doi: 10.4324/9781315146027]
14. Kawulich B. Participant observation as a data collection method. Forum: Qual Soc Res Sozialforschung 2005;6. [doi: 10.17169/fqs-6.2.466]
15. Liebetruth M, Kehe K, Steinritz D, Sammito S. Systematic literature review regarding heart rate and respiratory rate measurement by means of radar technology. Sensors (Basel) 2024 Feb 4;24(3):1003. [doi: 10.3390/s24031003] [Medline: 38339721]
16. Grießhammer SG, Malessa A, Lu H, et al. Contactless radar-based heart rate estimation in palliative care - a feasibility study and possible use in symptom management. BMC Palliat Care 2024 Nov 30;23(1):273. [doi: 10.1186/s12904-024-01592-3] [Medline: 39616332]
17. Mejdani D, Bräunig J, GrießHammer SG, et al. Radar-based tremor quantification using deep learning for improved Parkinson's and palliative care assessment. Trans Rad Sys 2024;2:1174-1185. [doi: 10.1109/TRS.2024.3494473]

XSL•FO
RenderX

18. Braeunig J, Mejdani D, Krauss D, et al. Radar-based recognition of activities of daily living in the palliative care context using deep learning. : IEEE; 2023 Presented at: 2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI); Oct 15-18, 2023; Pittsburgh, PA, USA p. 1-4. [doi: 10.1109/BHI58575.2023.10313506]
19. Linz S, Vinci G, Mann S, et al. A compact, versatile six-port radar module for industrial and medical applications. Journal of Electrical and Computer Engineering 2013;2013:1-10. [doi: 10.1155/2013/382913]
20. Vinci G, Lindner S, Barbon F, et al. Six-port radar sensor for remote respiration rate and heartbeat vital-sign monitoring. IEEE Trans Microwave Theory Techn 2013;61(5):2093-2100. [doi: 10.1109/TMTT.2013.2247055]
21. Islam SMM. Radar-based remote physiological sensing: progress, challenges, and opportunities. Front Physiol 2022;13:955208. [doi: 10.3389/fphys.2022.955208] [Medline: 36304581]
22. Sevcencu C, Struijk JJ. Autonomic alterations and cardiac changes in epilepsy. Epilepsia 2010 May;51(5):725-737. [doi: 10.1111/j.1528-1167.2009.02479.x] [Medline: 20067509]
23. Lu H, Kusnik S, Mammadova D, Trollmann R, Koelpin A. HRV-based monitoring of neonatal seizures with machine learning. : IEEE; 2024 Presented at: 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); Jul 15-19, 2024; Orlando, FL, USA p. 1-4. [doi: 10.1109/EMBC53108.2024.10782590]
24. Shi K, Schellenberger S, Will C, et al. A dataset of radar-recorded heart sounds and vital signs including synchronised reference sensor signals. Sci Data 2020 Feb 13;7(1):50. [doi: 10.1038/s41597-020-0390-1] [Medline: 32054854]
25. Shi K, Steigleder T, Schellenberger S, et al. Contactless analysis of heart rate variability during cold pressor test using radar interferometry and bidirectional LSTM networks. Sci Rep 2021 Feb 4;11(1):3025. [doi: 10.1038/s41598-021-81101-1] [Medline: 33542260]
26. Yusof MM, Kuljis J, Papazafeiropoulou A, Stergioulas LK. An evaluation framework for health information systems: human, organization and technology-fit factors (HOT-fit). Int J Med Inform 2008 Jun;77(6):386-398. [doi: 10.1016/j.ijmedinf.2007.08.011] [Medline: 17964851]
27. Weisser A, Öhl N, Steigleder T, et al. Soziale und ethische Implikationen neuer technologien in der palliativversorgung: kriterien für eine personenzentrierte methodenwahl. Zeitschrift für Palliativmedizin 2024 Sep;25(5):e79-e79. [doi: 10.1055/s-0044-1788537]
28. O'Reilly K. Ethnographic Methods, 2nd edition: Routledge; 2012:255. [doi: 10.4324/9780203864722]
29. Barbour R. Doing Focus Groups: SAGE Publications Ltd; 2018. [doi: 10.4135/9781526441836]
30. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
31. Walter U, Brandt S, Förderreuther S, et al. Empfehlungen der Deutschen Gesellschaft für Klinische Neurophysiologie und Funktionelle Bildgebung zur Diagnostik des irreversiblen Hirnfunktionsausfalls [Article in German]. Klin Neurophysiol 2019 Mar;50(1):17-22. [doi: 10.1055/a-0792-0986]
32. MAXQDA: VERBI Software; 2022. URL: https://www.maxqda.com [accessed 2025-08-01]
33. Mayring P. Qualitative Content Analysis. Forum Qual Soc Res 2000;1(2). [doi: 10.17169/fqs-1.2.1089]
34. Mayring P. Qualitative Content Analysis: A Step-by-Step Guide: SAGE Publications; 2021:978-1529701975. [doi: 10.4135/9781036231798]
35. Pelz C, Schmitt A, Meis M. Knowledge mapping as a tool for analyzing focus groups and presenting their results in market and evaluation research. Forum Qual Soc Res 2004;5(2). [doi: 10.17169/fqs-5.2.601]
36. Miro. 2023. URL: https://miro.com [accessed 2025-08-01]
37. Lu M, Speri E, Lauks A, Leibson C, Rao SC. Understanding health care provider burnout when caring for patients with refractory epilepsy in the United States. Neurol Clin Pract 2024 Apr;14(2):e200260. [doi: 10.1212/CPJ.0000000000200260] [Medline: 38585442]
38. Peltola ME, Leitinger M, Halford JJ, et al. Routine and sleep EEG: minimum recording standards of the International Federation of Clinical Neurophysiology and the International League Against Epilepsy. Epilepsia 2023 Mar;64(3):602-618. [doi: 10.1111/epi.17448] [Medline: 36762397]
39. Tatum WO, Mani J, Jin K, et al. Minimum standards for inpatient long-term video-EEG monitoring: a clinical practice guideline of International League Against Epilepsy and the International Federation of Clinical Neurophysiology. Clin Neurophysiol 2022 Feb;134:111-128. [doi: 10.1016/j.clinph.2021.07.016] [Medline: 34955428]
40. Beniczky S, Wiebe S, Jeppesen J, et al. Automated seizure detection using wearable devices: a clinical practice guideline of the International League Against Epilepsy and the International Federation of Clinical Neurophysiology. Clin Neurophysiol 2021 May;132(5):1173-1184. [doi: 10.1016/j.clinph.2020.12.009] [Medline: 33678577]
41. Kiriakopoulos ET, Chu F, Jobst BC, Kobau R. Innovating the practice of epilepsy care to improve patient outcomes: a model for community health worker integration. Front Neurol 2025;16:1654687. [doi: 10.3389/fneur.2025.1654687] [Medline: 41127292]
42. Martz G, Gurung I, Li YH. Implementation of clinic-based data capture in a real-world epilepsy clinic. Epilepsy Behav 2025 Nov;172:110730. [doi: 10.1016/j.yebeh.2025.110730] [Medline: 41043241]
43. Tao S, Lhatoo S, Hampson J, Cui L, Zhang GQ. A bespoke electronic health record for epilepsy care (EpiToMe): development and qualitative evaluation. J Med Internet Res 2021 Feb 12;23(2):e22939. [doi: 10.2196/22939] [Medline: 33576745]

XSL•FO
RenderX

44. Pellinen J, Foster EC, Wilmshurst JM, Zuberi SM, French J. Improving epilepsy diagnosis across the lifespan: approaches and innovations. Lancet Neurol 2024 May;23(5):511-521. [doi: 10.1016/S1474-4422(24)00079-6] [Medline: 38631767]
45. Velis D, Plouin P, Gotman J, da Silva FL, ILAE DMC Subcommittee on Neurophysiology. Recommendations regarding the requirements and applications for long-term recordings in epilepsy. Epilepsia 2007 Feb;48(2):379-384. [doi: 10.1111/j.1528-1167.2007.00920.x] [Medline: 17295634]
46. Elger CE, Hoppe C. Diagnostic challenges in epilepsy: seizure under-reporting and seizure detection. Lancet Neurol 2018 Mar;17(3):279-288. [doi: 10.1016/S1474-4422(18)30038-3] [Medline: 29452687]
47. Bruno E, Viana PF, Sperling MR, Richardson MP. Seizure detection at home: do devices on the market match the needs of people living with epilepsy and their caregivers? Epilepsia 2020 Nov;61 Suppl 1(S1):S11-S24. [doi: 10.1111/epi.16521] [Medline: 32385909]
48. Bruno E, Simblett S, Lang A, et al. Wearable technology in epilepsy: the views of patients, caregivers, and healthcare professionals. Epilepsy Behav 2018 Aug;85:141-149. [doi: 10.1016/j.yebeh.2018.05.044] [Medline: 29940377]
49. Murray E, Hekler EB, Andersson G, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016 Nov;51(5):843-851. [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
50. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions-a call for innovative evidence generation approaches. NPJ Digit Med 2020;3(1):110. [doi: 10.1038/s41746-020-00314-2] [Medline: 32904379]
51. Zanaboni P, Ngangue P, Mbemba GIC, Schopf TR, Bergmo TS, Gagnon MP. Methods to evaluate the effects of internet-based digital health interventions for citizens: systematic review of reviews. J Med Internet Res 2018 Jun 7;20(6):e10202. [doi: 10.2196/10202] [Medline: 29880470]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Studies
EEG: electroencephalography
HRV: heart rate variability
MTA: medical technical assistant

Edited by A Stone; submitted 26.Aug.2025; peer-reviewed by DF Borges, J Gehrmann; revised version received 27.Jan.2026; accepted 27.Jan.2026; published 24.Mar.2026.

Please cite as:
Öhl N, Strack D, Kusnik SM, Link I, Hofmann M, Steigleder T, Weisser A, Ostgathe C, Trollmann R, Heckel M
Challenges of Standard Pediatric Epilepsy Monitoring and the Potential Benefits of Contactless Sensor Technologies: Exploratory Qualitative Study
J Med Internet Res 2026;28:e83009
URL: https://www.jmir.org/2026/1/e83009
doi:10.2196/83009

© Natalie Öhl, Dilbar Strack, Stefan Martin Kusnik, Isabell Link, Marlene Hofmann, Tobias Steigleder, Alina Weisser, Christoph Ostgathe, Regina Trollmann, Maria Heckel. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and User-Centered Evaluation of Smart Systems for Loneliness Monitoring in Older Adults: Mixed Methods Study

Yi Zhou[1], PhD; Jessica Rees[2], PhD; Faith Matcham[3], PhD; Ashay Patel[4], PhD; Michela Antonelli[4], PhD; Anthea Tinker[2†], PhD; Sebastien Ourselin[4], PhD; Wei Liu[1], PhD

[1]Department of Engineering, King's College London, London, United Kingdom

[2]Department of Global Health and Social Medicine, King's College London, London, United Kingdom

[3]School of Psychology, Faculty of Science, Engineering and Medicine, University of Sussex, Brighton, United Kingdom

[4]School of Biomedical Engineering & Imaging Sciences, King's College London, London, United Kingdom

[†]deceased

Corresponding Author:
Wei Liu, PhD
Department of Engineering
King's College London
Strand Campus
London, WC2R 2LS
United Kingdom
Phone: 44 020 7836 5454
Email: wei.liu@kcl.ac.uk

Abstract

Background: Loneliness is a critical issue among older adults and constitutes a significant risk factor for a range of physical and mental health conditions. However, current assessment methods primarily rely on self-report questionnaires and clinical evaluations, which are susceptible to recall bias and social desirability bias, highlighting the need for more objective and continuous assessment approaches. Recent studies have reported associations between physiological and behavioral indicators and the experience of loneliness in older adults. While these technologies have demonstrated correlations between physiological and behavioral sensor data and the experience of loneliness, their implementation has been limited. Most systems rely on fixed-location sensors or smartphone apps, with little attention given to the integration of these tools into users' daily routines. To date, no published studies have applied smart textile technology, which integrates sensing capabilities directly into garments or furniture, as a medium for loneliness detection. This study addresses that gap by exploring the usability, experiential acceptability, and ethical considerations of smart textile-based monitoring systems.

Objective: This study aims to assess the perceived usability, acceptability, and emotional resonance of a smart loneliness monitoring system integrating sensing garments, furniture, and a mobile app and identify design implications to guide future improvement and promote sustained engagement among older adults.

Methods: Building on earlier conceptual research, a functional prototype system was developed and evaluated through 2 immersive in-person workshops with older adults (N=10). A mixed methods approach was applied, combining structured questionnaires, sensory ethnographic observations, focus group discussions, and experience-based co-design. Quantitative data were analyzed descriptively, and qualitative data were analyzed thematically to explore user perceptions related to system usability, emotional response, lifestyle compatibility, and ethical considerations.

Results: Quantitative data indicated high user satisfaction in dimensions such as comfort, ease of use, and feedback clarity. However, trust in long-term monitoring and willingness to use the system regularly varied. Thematic analysis revealed 4 main areas influencing acceptance, including wearability, usability, and daily integration; trust, privacy, and data control; perceptions of loneliness and the limits of detection; and adoption, applicability, and ethical futures. Participants emphasized the need for discretion, personalization, and human oversight in system feedback and data-sharing mechanisms.

Conclusions: The resulting prototype was positively received, demonstrating the potential of smart systems for passive and personalized loneliness monitoring among older adults. However, adoption is influenced by perceptions of autonomy, emotional sensitivity, and contextual integration. Future development should focus on modularity, transparency, and integration within care infrastructures to ensure ethical and sustainable deployment.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e81027) doi:10.2196/81027

KEYWORDS

older adults; mental health; loneliness; monitoring; user-centered design; smart textile; wearable and ambient technology; wearable technology

Introduction

Loneliness and social isolation have been identified as significant global mental health challenges, with particularly profound effects on older adults. There are up to a quarter of older individuals worldwide experiencing social isolation [1-3]. As people age, they may reduce their social interactions due to various factors such as reduced mobility, retirement, or the loss of partners, leading to social isolation and intensify feelings of loneliness [4-6]. A growing body of research has shown that prolonged loneliness is associated with increased risks of depression, cognitive decline, cardiovascular disease, and mortality rates, which poses health risks comparable to smoking and obesity [7,8]. Beyond its impact on individuals, loneliness also places substantial strain on health care systems by increasing demand for clinical care and long-term support services [4].

Despite a growing understanding of the impact of loneliness on health, it remains a challenge to accurately measure and monitor loneliness, particularly in nonclinical and home-based settings [2,7]. While clinical settings may allow for structured assessments by health professionals, such measurements are often constrained by time, context, or the presence of other comorbidities. Traditional assessment methods mainly rely on self-report questionnaires such as the University of California, Los Angeles 3-Item Loneliness Scale, the De Jong Gierveld Loneliness Scale for older adults, or clinician-administered questionnaires [9,10]. While these tools are well-validated, they are susceptible to recall and social desirability bias, particularly among older adults who may underreport emotional distress due to stigma or generational attitudes toward mental health [11,12]. Moreover, such methods often provide snapshot assessments rather than capturing the dynamic, fluctuating nature of loneliness as experienced in daily life [13]. These limitations emphasize the need for continuous, objective, and context-aware detection of loneliness, enabling more timely and personalized interventions.

Sensor-based technologies, especially wearable and ambient sensing systems, have advanced considerably in mental health monitoring, offering new opportunities to detect loneliness through physiological and behavioral data [14,15]. Recent studies have reported several behavioral patterns and physiological indicators associated with loneliness, including reduced physical activity [16,17], sleep disturbances [5], binge or comfort eating [18], elevated blood pressure [19], and increased average salivary cortisol levels [20,21]. While these findings do not establish diagnostic relationships, they suggest measurable correlates that may guide the design of future sensing-based systems. Wearable devices such as smartwatches and fitness bands have demonstrated the ability to monitor many of these indicators. When analyzed over time, these data can provide inferences about an individual's psychological well-being and deviations from their baseline states [20,22]. Additionally, various sensing systems have been applied to monitor loneliness and social isolation in older adults. These include vision-based motion capture systems for activity level tracking [23], ambient light and sound sensors for detecting social behaviors [24,25], and smartwatches (including accelerometers or inclinometers) to track posture and sedentary behavior [16,26]. However, camera-based systems often raise privacy concerns, and fixture-mounted sensors such as wall-mounted light and sound sensors may lack the portability required for continuous monitoring at home and in community settings [15]. While wearable devices such as smartwatches offer portable sensing, they present their own limitations in older adults, including discomfort with wrist-worn devices, low personalization, and poor integration into daily domestic routines [27,28]. Moreover, many existing monitoring systems focus solely on physical movement or posture and fail to capture the complex emotional and physiological dimensions of loneliness [14,29].

Textile-based sensing technologies, which integrate sensors and conductive materials into fabrics, are able to offer a comfortable and effective solution for long-term mental health monitoring. By integrating sensing capabilities into flexible fabrics, sensing textile systems can passively and continuously collect data without interfering with users' daily routines or drawing attention to the monitoring process [22,30]. Furthermore, electronic textiles can seamlessly embed into familiar objects such as garments or home furnishings, enhancing both physical comfort and acceptability for older users [31-33]. Despite advances in smart textiles for health monitoring, no textile-based sensing system has been developed specifically for loneliness detection. One prior study explored the use of a textile band to capture speech frequency as an indicator of social interaction, but it did not attempt to evaluate the subjective experience of loneliness or integrate these signals into a mental health monitoring framework [22]. While a growing body of pervasive computing and ambient-assisted living research has focused on detecting loneliness through environmental sensors and wearables [34-36], these systems have largely relied on noncustomizable and device-centric approaches with limited integration into user experiences. In contrast, textile-based systems offer the potential for more seamless, passive, and embodied interaction. However, despite these advantages, they remain underexplored in loneliness-related apps. Previous research has highlighted the critical role of design factors such as format, materials, and sensor placement in user acceptance and sustained engagement with textile-based systems [37,38]. However, few studies have directly integrated co-design, lived experience research, and user-driven evaluation into the development of such systems for older adults. As loneliness is not only just a behavioral state but also a subjective and socially situated experience, the design and development of textile sensing systems need to go beyond engineering efficacy to

XSL•FO
RenderX

reflect the emotional, ethical, and contextual needs of users [12]. This study addresses this gap by placing older adults' voices at the center of system development and evaluation.

In our previous research, we conducted interviews and collected feedback from older users and stakeholders to understand the design requirements and expectations of smart loneliness monitoring systems for older adults [12,27,28]. These earlier works were primarily conceptual, exploring hypothetical interactions and preferences prior to the existence of a working prototype. In contrast, this study advances this body of work by designing and evaluating an integrated smart loneliness monitoring system, comprising sensing garments, furniture, and a companion mobile app through immersive user engagement. Additionally, this study makes a novel contribution by combining prototype-led experience, sensory ethnography, structured quantitative feedback, and co-design outputs to generate both actionable design insights and a deeper understanding of the emotional and ethical responses of older adults. By combining quantitative and qualitative analyses, we examined different dimensions of user acceptance in the context of smart loneliness detection, including wearability, emotional trust, loneliness perception, and pathways to adoption. Finally, we discussed the design implications, ethical considerations, and future directions for integrating smart textiles into the everyday mental health care of older adults.

Methods

Overview

This study builds upon our previous interview and co-design research conducted with older adults and stakeholders, which identified essential user needs and expectations regarding the design and development of smart systems to monitor loneliness in later life [12,27,28]. These early insights informed the development of our smart loneliness monitoring systems, which comprise sensing garments and sensing furniture designed to unobtrusively capture physiological and behavioral indicators associated with loneliness. The resulting prototype was evaluated in the current focus group study.

To evaluate and further improve the system, 2 in-person evaluation workshops were held, each involving 5 older adults aged 65 years and older who had experienced loneliness (N=10). The aim of the workshops was to gather experiential feedback and design suggestions from older users. A mixed methods approach was used, integrating sensory observations, self-report questionnaires, focus group discussions, and co-design activities. These methods enabled a comprehensive exploration of users' practical and emotional feedback to the system, providing pragmatic design insights to guide future development.

Participants

Participants were recruited using a combination of convenience and purposive sampling strategies, with the aim to engage older adults from diverse life backgrounds and with varying levels of technological adaptability. Recruitment was conducted through 2 main channels. First, the research team directly contacted individuals who had previously expressed interest in the DELONELINESS project [39]. Second, a study invitation was distributed via the PROTECT study newsletter, which reaches over 20,000 older adults across the United Kingdom. For logistical feasibility, only individuals residing within a 50-mile radius of central London were considered from the PROTECT email list. Eligibility criteria included being aged 65 years or older, fluent in spoken English, and having experienced loneliness at some point during their later life (postretirement age). Individuals diagnosed with cognitive impairments or dementia were excluded from participation to ensure that participants could provide informed consent and fully engage with the system interaction and co-design activities. Participants were screened by researchers trained in applying the principles of the Mental Capacity Act (2005) [40], which enabled them to assess an individual's capacity to participate during the recruitment stage.

Loneliness severity was assessed using the University of California, Los Angeles Loneliness Scale [41], which has been linked to various health outcomes and functional limitations. In total, 10 participants who met the eligibility criteria took part in the study and completed both workshop sessions. This sample size was determined based on recommendations for user-centered qualitative evaluations, which typically involve 3 to 15 participants to obtain experiential insights and design implications in early-stage technology development [42]. Similar sample sizes have been used in published co-design and feasibility studies involving older adults and digital health technologies [43-45]. A total of 10 participants consented and completed both workshop sessions. While small in size, this sample enabled in-depth participatory engagement, iterative feedback, and contextual exploration, which are central goals of this exploratory mixed methods study. Additionally, data collection concluded after 10 participants, as thematic saturation was observed across the 2 workshop sessions, with recurring patterns and consistent feedback emerging during the analysis phase.

Technology Description

The smart loneliness monitoring systems evaluated in this study consisted of 2 key components, including the sensing garment and sensing furniture. In the previous co-design workshops with older adults and stakeholders, we identified key design factors influencing the older adults' acceptance of monitoring technologies [28]. Additionally, our prior qualitative research exploring the psychological experience of loneliness in later life informed sensing technology selection and development [12,39]. Building on these insights, our systems were constructed to continuously and noninvasively track physiological and behavioral signals associated with loneliness while remaining compatible with the daily lives and domestic environments of older users.

The sensing garment was mainly intended for physiological signal monitoring. To ensure both wearability and sensing accuracy, we developed 3 different sizes (small, medium, and large) of long-sleeved zip-up shirts made from a breathable elastic textile blend (92% polyester and 8% elastane). The shirts were designed to be worn over regular clothing to facilitate dressing and undressing during workshop sessions. The sensing system of the garment included a fabric-based conductive circuit,

XSL•FO
RenderX

modular sensor units, and a data acquisition module. As shown in Figure 1A, textile circuits were mapped along the garment seams and encapsulated using thermoplastic polyurethane via heat-pressing to ensure durability, washability, and smoothness, minimizing tactile discomfort or abrasion risks for older users with sensitive skin.

Figure 1. Components of the smart loneliness monitoring system: (A) sensing garment embedded with conductive textile circuits and modular sensors for pulse rate, respiration, and temperature monitoring. (B) Smart furniture including a sensing pillow and seat pad integrated with pressure sensor module for posture and behavioral monitoring.

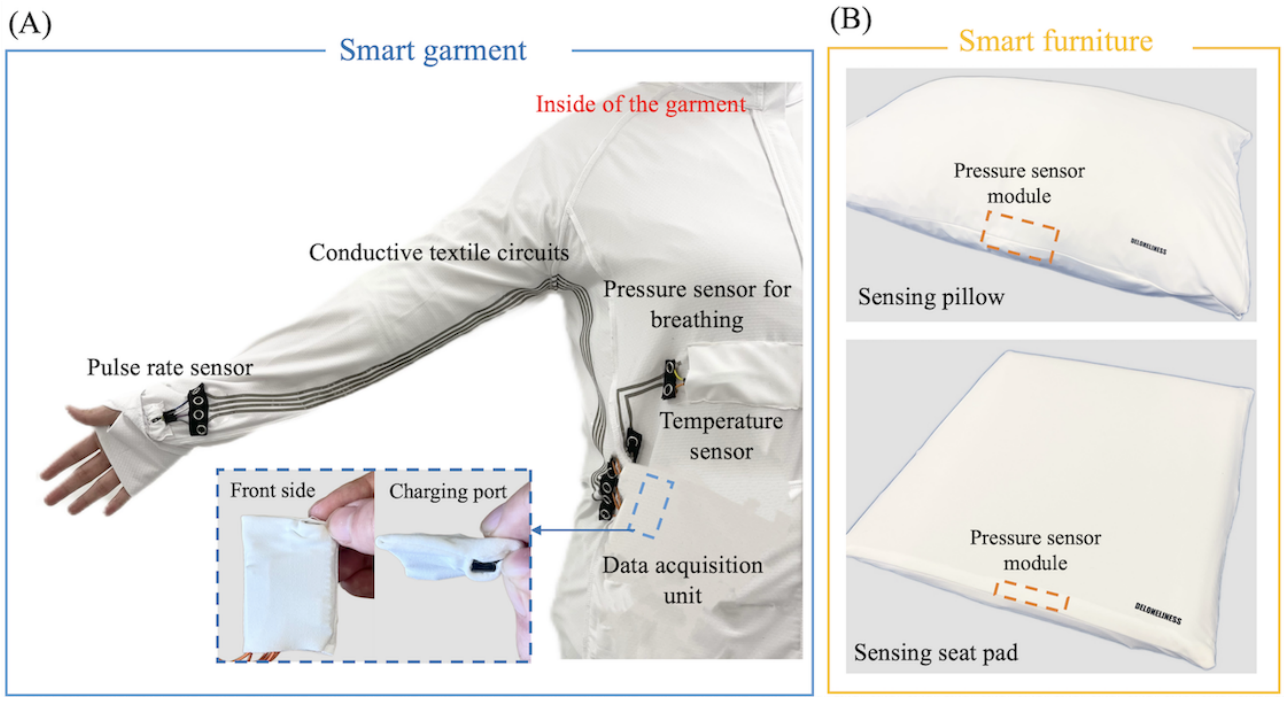

To support individual autonomy and improve independence, a modular design was applied using metal press-fit snaps, allowing users or caregivers to easily attach, detach, or reposition sensing components without technical expertise. Specifically, a pulse rate sensor was placed at the wrist cuff to enable accurate heart rate monitoring. A temperature sensor and a pressure sensor for respiration were embedded on the interior side of the chest and abdominal regions to collect real-time body temperature and breathing rate data. The data acquisition unit included a built-in inertial measurement unit housed in a soft fabric casing and integrated into a garment pocket. The sensing system was powered by a commercially available rechargeable battery and can be conveniently charged via an external charging port without needing to open the casing, thereby reducing the cognitive and physical burden during maintenance. The proposed modular design allowed older users to customize their configurations by selecting and combining the sensing components most relevant to their individual needs.

The sensing furniture was designed based on everyday household items such as a pillow (sensing pillow) and a seat cushion (sensing seat pad) with custom-developed textile covers (Figure 1B). Pressure sensor modules were embedded into designated internal regions of the pillow and seat pad to enable continuous monitoring of posture and pressure distribution while sitting or lying down. Similar to our sensing garment, the furniture also applied a modular design that allowed sensor modules and the data acquisition unit to be attached via press-fit snaps, simplifying removal for maintenance or cleaning of the textile surfaces.

Figure 2 demonstrates the system architecture. The proposed smart loneliness monitoring system was designed to continuously collect physiological and behavioral data through the sensing garment and furniture. These raw signals would be transmitted via a smartphone or communication gateway, which performed preliminary signal preprocessing such as noise filtering and timestamping before uploading to a secure cloud environment. In future iterations, advanced machine learning algorithms would be applied in the cloud to identify potential indicators of loneliness, such as irregular activity patterns, reduced physiological variability, or prolonged inactivity. Thresholds for generating loneliness-related feedback have not yet been predefined. Based on participant input during the co-design stage, the system was intended to include adaptive feedback mechanisms in user interfaces, such as customizable mood prompts and check-in features, allowing users to confirm, dismiss, or annotate inferred states.

XSL•FO
RenderX

Figure 2. Proposed system architecture and data flow.

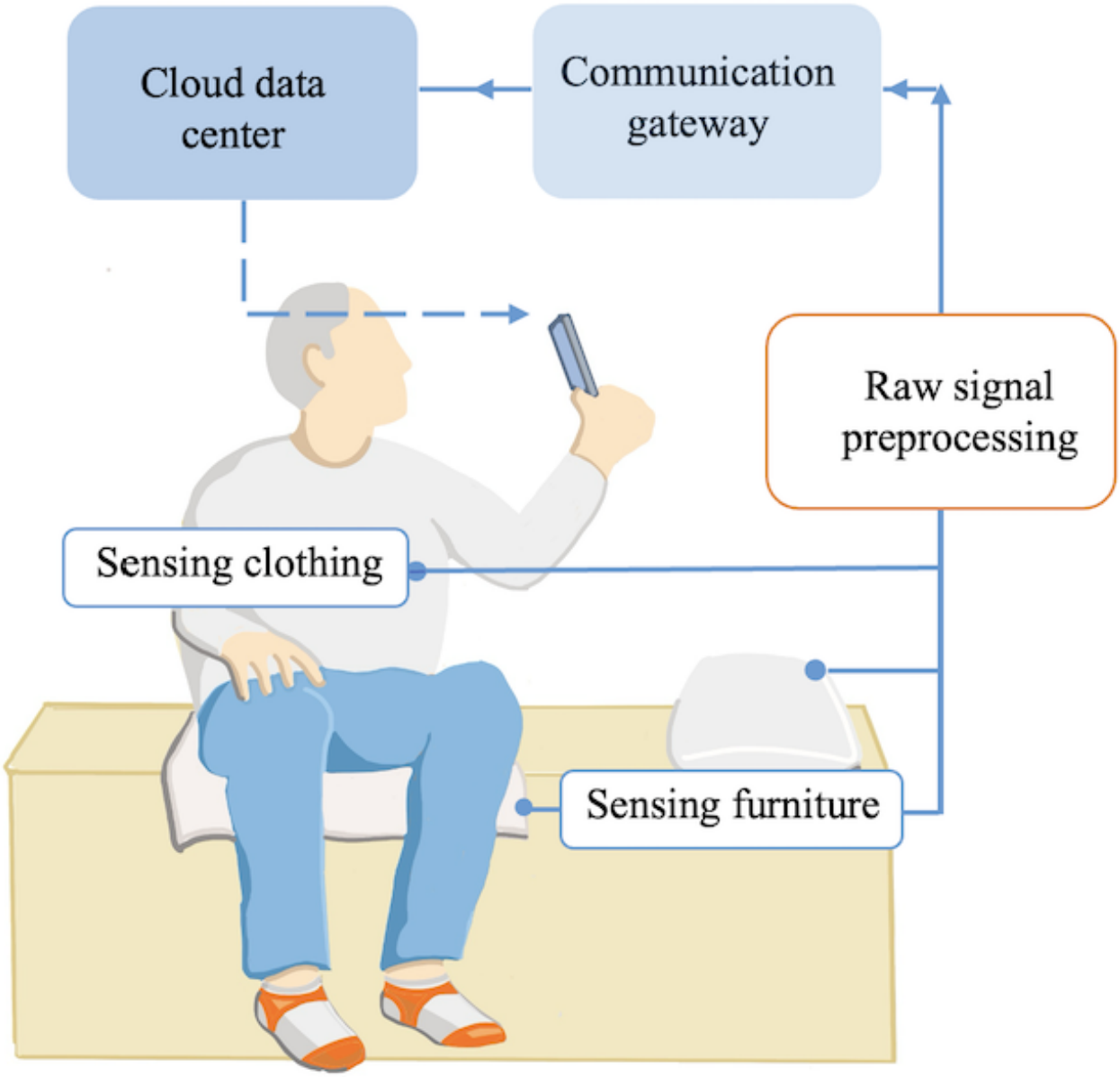

Figure 3 demonstrates our user interfaces from various user perspectives. From the user's perspective (Figure 3A), the app provided real-time feedback on heart rate, respiration rate, skin temperature, sleep duration, and daily activity levels, while also visualizing current loneliness status and offering recommendations for health-promoting activities. The system further included companion interfaces for family members and caregivers (Figure 3B). Family members can monitor their loved one's loneliness status, view summarized statistics, and check upcoming social or medical appointments, with integrated communication options such as direct calling or messaging. For health care professionals (Figure 3C), the app offered detailed health data and aggregated loneliness levels across users, facilitating targeted service recommendations based on individual needs.

XSL•FO
RenderX

Figure 3. User interfaces of the smart loneliness monitoring system. (A) User interface displaying real-time physiological data, loneliness status, and activity recommendations. (B) Family member interface providing an overview of emotional status, scheduled activities, and communication access. (C) Health care professionals interface showing loneliness metrics across individuals and the service referral interface.

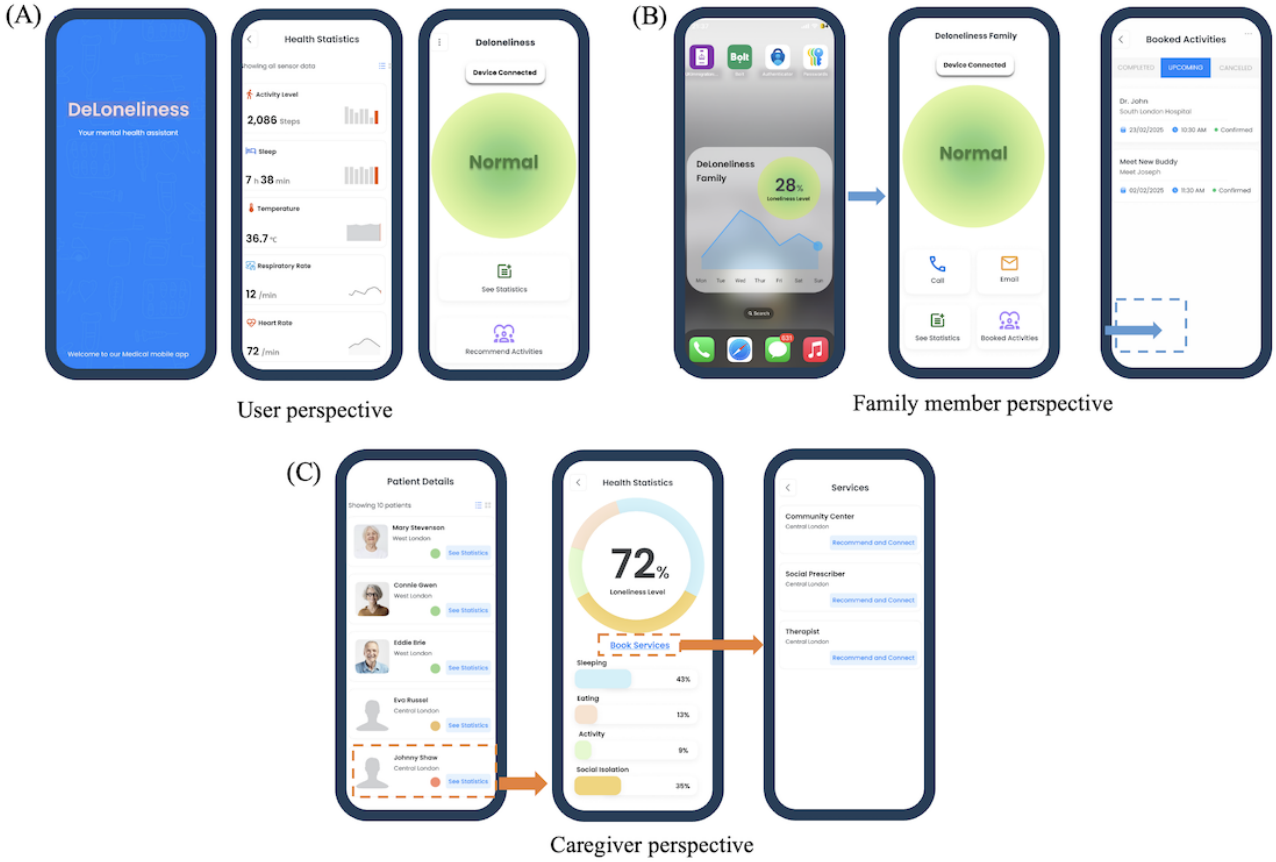

Procedure

Overview

This study was conducted through in-person workshops consisting of a series of structured participatory activities designed to evaluate and improve the smart loneliness monitoring system. The research procedure was divided into 2 sequential components including an experiential evaluation of the system and an experience-based co-design for improvement session (Figure 4).

XSL•FO
RenderX

Figure 4. Workshop procedure.

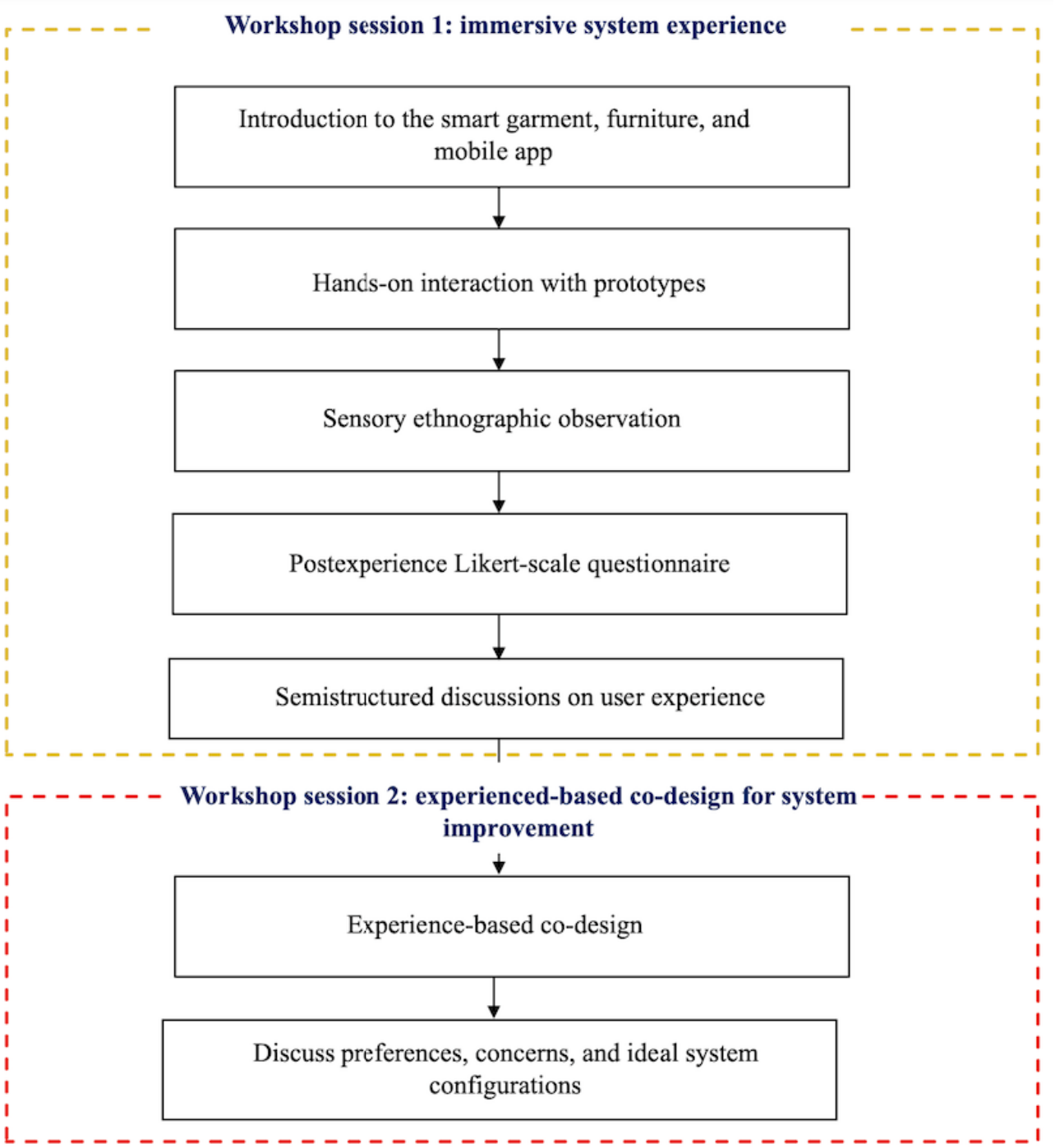

Session 1: Experiencing the Smart Loneliness Monitoring System

The first session of the workshop was designed to allow participants to interact directly with the smart loneliness monitoring system. This session comprised 3 main stages: an introductory demonstration, an immersive trial, and a postexperience feedback phase including a questionnaire and discussion. The session aimed to understand users' initial perceptions, sensory impressions, and evaluations of the system within a supportive and participatory environment.

The session began with an introduction to the smart loneliness monitoring system by the lead author (YZ), who presented the components of the system, including the sensing garment and sensing furniture. Participants were shown a short video demonstrating the system's functional workflow and potential use cases in everyday life. This was followed by a live demonstration, during which researchers explained how to operate, maintain, clean, and charge the system. Participants were encouraged to ask questions at any time during the demonstration, ensuring that the purpose, use, and operational aspects of the system were clarified. This stage served as a foundation for the participants' subsequent independent and immersive system interaction.

All the participants were then invited to try on the sensing garment and interact with the sensing furniture (sitting on the

XSL•FO
RenderX

sensing seat pad or lying on the pillow) and operate the accompanying mobile app to see the real-time physiological feedback. This immersive experience was guided by principles of sensory ethnography, which is a commonly applied method in the evaluation of wearable and environmental technologies to capture the embodied and affective dimensions of user experience [46-48]. Participants were encouraged to attend to their physical, emotional, and sensory responses such as how the garment felt on their skin, the sensations caused by the furniture, and their affective responses to the system feedback while experiencing the systems. Three research assistants (YZ, JR, and WL) used structured observation forms (Multimedia Appendix 1) to document participants' verbal and nonverbal reactions, behaviors, and interactions with the materials.

After the immersive trial, participants completed a Likert-scale questionnaire, designed to assess key dimensions such as ease of use, comfort, perceived usefulness, trust in the system, and concerns about data privacy. The questionnaire was informed by the technology acceptance model (TAM), which is a widely used theoretical framework that explains users' acceptance of technology based on perceived usefulness and perceived ease of use [49]. Specifically, we adapted core constructs from TAM, including perceived usefulness, ease of use, and behavioral intention to use. To reflect the characteristics of smart textile and ambient sensing systems, we also supplemented items derived from established usability and health technology evaluation frameworks such as comfort, aesthetics, data privacy, and daily life integration [50,51]. A full list of questionnaire items is provided in Multimedia Appendix 2. The experience session concluded with a semistructured focus group discussion. The discussion guide was also informed by the TAM and themes identified in our previous co-design research, which served to ensure comprehensive topic coverage rather than dictate interpretive categories [27]. Topics explored included the system's integration into daily life, emotional responses to the monitoring experience, clarity and interpretability of feedback, comfort level of the systems, and concerns around data sharing and privacy.

Session 2: Experience-Based Co-Design for System Improvement

The second part of the workshop focused on collecting user-driven suggestions for system improvements through an experience-based co-design activity [52]. Based on the insights and impressions gained during the initial system trial, participants were invited by the lead investigator (YZ) to reimagine various aspects of the smart systems. This session aims to empower older adults to become cocreators, allowing them to contribute their experiential knowledge and preferences toward the future development of a more practical, comfortable, and acceptable solution.

Each participant was provided with a design toolkit, which included body and furniture layout co-design template (Multimedia Appendix 3), sensor placement icons, and a range of fabric samples representing commonly used textile compositions (100% cotton, 100% wool, 80% cotton/20% polyester, 80% wool/20% polyester, and 92% polyester/8% elastane). The toolkit also includes materials for conceptualizing sensor integration methods, such as Velcro, magnetic clasps, hooks, and press-fit snaps, which are commonly used to attach electronic modules in textile-based systems.

The session began with a short reflective exercise, in which participants were asked to individually outline their daily routines, experience prompts, and situational preferences related to loneliness and technology use. This laid the foundation for them to anchor design thinking in their own life experience. Participants were then encouraged to engage in open-ended visual prototyping using the templates and materials provided. They annotated body outlines and home layouts with preferred sensor placement zones, marked areas to be avoided for comfort or privacy reasons, and proposed new features such as adjustable garment structures, softer fabric options, or additional sensing functions to enhance comfort, usability, and emotional acceptance.

The co-design session ended with a group sharing activity, where each participant presented their redesigned concept to the group and explained the rationale behind their design decisions. These presentations provided valuable qualitative insights into the different preferences and expectations of older users, which can inform iterative improvements in the future system design.

Data Analysis

Overview

A mixed methods analytical approach was applied to analyze the data collected during the workshops. The study was structured in sequential phases, where the initial quantitative questionnaire was used to prompt structured participant reflection, and the subsequent qualitative phase provided a deeper contextual understanding. Quantitative data from the Likert-scale questionnaire were analyzed descriptively. The qualitative data including audio recordings from focus group discussions, co-design artifacts, field notes from sensory ethnographic observations, and participants' reflective comments following the co-design session were analyzed using inductive thematic analysis [53,54].

Quantitative Data Analysis

Quantitative data from the postexperience Likert-scale questionnaires were entered into SPSS (version 29; IBM Corp) and analyzed using descriptive statistics, including means, SDs, and frequency distributions. These data provided an initial understanding of participants' perceptions and satisfaction with the system and served as a foundation for subsequent qualitative discussions.

Qualitative Data Analysis

Audio recordings from the focus group discussions and the co-design sharing presentations were transcribed collaboratively by the first and second authors (YZ and JR). In addition, co-design artifacts such as annotated body and home layout templates, sticky notes, and sketches were digitized and treated as supplementary qualitative data. Field notes from sensory ethnographic observations were also integrated into the qualitative findings to contextualize participants' behaviors and embodied responses during interaction with the system.

XSL•FO
RenderX

Thematic analysis followed Braun and Clarke's [55] 6-phase framework and was independently conducted by 2 researchers (YZ and WL). The researchers first familiarized themselves with the data through review of transcripts and design materials, followed by open coding using NVivo (version 14; Lumivero). Codes were iteratively grouped into broader themes, which were refined through collaborative discussions within the research team. Although the focus group guide was informed by constructs from the TAM and themes identified in prior co-design research, data analysis proceeded entirely inductively. Themes were generated directly from participants' perspectives and lived experiences, rather than being constrained by a predetermined theoretical framework.

While the earlier quantitative assessment was structured around 8 predefined dimensions, the subsequent qualitative analysis adopted an open-coding approach. This was to ensure that participants' language, emotional nuance, and contextual expressions were not shaped or limited by prior assumptions. Although some thematic overlap with the questionnaire domains was observed, the 2 phases were intentionally not aligned, allowing for the emergence of novel concerns and experiential insights.

Ethical Considerations

This study received ethics approval from the King's College London Research Ethics Committee (reference: LRS/DP-24/25-34602). Prior to participation, all individuals were provided with a participant information sheet and a consent form, which has explained the purpose of the study, their rights, and the voluntary nature of participation. Participants were informed that they could withdraw from the study at any time. Written informed consent was obtained on the day of the workshop. During data processing, all the data were anonymized by assigning each participant a unique and nonidentifiable identification number. A password-protected file containing participants' names and contact details was stored separately from research data and was only accessible to the core research team. No direct financial compensation was provided for participation. However, refreshments and lunch were offered on the day of the workshop, and travel expenses were reimbursed.

Results

Participants' Characteristics

A total of 10 participants meeting the eligibility criteria took part in the study. Table 1 summarizes their demographic characteristics, including age range, sex, living arrangements, levels of technology use, prior experience with health-monitoring technologies, and loneliness score.

XSL•FO
RenderX

Table 1. Characteristics of the participants (N=10).

Characteristics	Values
Age (years)	
Mean (SD)	68.8 (4.2)
Range	65-79
Sex, n (%)	
Female	8 (80)
Male	2 (20)
Highest education level, n (%)	
Secondary education or below	6 (60)
Postsecondary education	4 (40)
Employment status, n (%)	
Retired	9 (90)
Employed	1 (10)
Living arrangement, n (%)	
Living alone	5 (50)
Living with others	5 (50)
Technology use, n (%)	
Low	2 (20)
High	8 (80)
Experiencing in using health monitoring technology	
Yes	5 (50)
No	5 (50)
Loneliness score (UCLA[a])	
Mean (SD)	5.1 (1.2)
Range	3-7

[a]UCLA: University of California, Los Angeles.

Descriptive Statistics

We conducted a quantitative analysis of the postexperience questionnaires completed by participants following their interaction with the smart loneliness monitoring system. The questionnaire comprised 18 items spanning 7 key dimensions, including ease of use, integration into daily life, perceived usefulness, understanding and clarity of feedback, trust in system functionality, concerns of data privacy, comfort level, and overall system acceptance. Each item was rated on a 5-point Likert scale starting from 1=strongly disagree" to 5=strongly agree.

Table 2 summarizes the descriptive statistics for user perceptions of the smart loneliness monitoring system. Overall, the responses reflected a positive attitude toward the system, with participants tending to "agree" or "strongly agree" with the statements across various dimensions. The highest-rated dimensions were ease of use, understanding and clarity of feedback, and comfort level, with average scores of 4.0 (SD 0.73), 4.5 (SD 0.52), and 4.4 (SD 0.53), respectively. However, responses related to the system's long-term integration, trust, and data sharing showed greater variability, indicating that evaluations of the system's role in users' everyday life were more complex and individualized. These divergences will be further explored in the qualitative findings presented in the Thematic Analysis section.

XSL•FO
RenderX

Table 2. Descriptive statistics for user perceptions of the smart loneliness monitoring systems.

Item description	Mean (SD)	FD[a]-1	FD-2	FD-3	FD-4	FD-5
1. Ease of use						
System is easy and convenient to use	3.8 (0.92)	0	1	2	5	2
Garment easy to put on or take off	4.4 (0.70)	0	0	1	4	5
Maintenance process is manageable	3.9 (0.57)	0	0	2	7	1
2. Integration into daily life						
Willing to use regularly	3.5 (1.18)	0	2	4	1	3
System fits into daily routines	3.7 (0.67)	0	0	4	5	1
3. Usefulness						
Useful for personal well-being	3.8 (0.82)	0	1	2	6	1
Useful for detecting loneliness	3.9 (0.87)	0	0	4	3	3
4. Understanding and clarity of feedback						
App feedback was easy to understand and interpret	4.5 (0.52)	0	0	0	5	5
5. Trust and reliability						
Trust in system performance	3.6 (0.69)	0	0	5	4	1
System reliably monitors loneliness conditions	3.4 (0.51)	0	0	6	4	0
6. Privacy and data concerns						
Comfortable sharing data	3.6 (0.84)	0	1	3	5	1
Comfortable being continuously monitored	3.8 (1.13)	0	2	1	4	3
7. Comfort level						
Comfort of sensing garment	4.3 (0.48)	0	0	0	7	3
Comfort of sensing furniture	4.9 (0.32)	0	0	0	1	9
Comfort of electronic textile component	4.3 (0.48)	0	0	0	7	3
Systems feels emotionally supportive	4.1 (0.87)	0	0	3	3	4
8. Overall acceptance						
Overall system acceptability	3.9 (0.73)	0	0	3	5	2
Willingness to recommend to others	3.7 (0.67)	0	0	4	5	1

[a]FD: frequency distribution.

Specifically, the ease of use received relatively high ratings. The highest-scoring item was “The garment was easy to put on and take off” (mean 4.4, SD 0.70), followed by “The maintenance process is manageable” (mean 3.9, SD 0.57). These results suggest that most of the participants were able to interact with the system independently and comfortably. In contrast, the integration into daily life revealed different perspectives. While most participants agreed that the system could be incorporated into their routine (mean 3.7, SD 0.67), their willingness to use the system regularly over time varied more significantly (mean 3.5, SD 1.18). This points to the complexity of sustained engagement and highlights participants’ desire for personalization and control. These themes are also demonstrated in their co-design for improvement session (see Wearability, Usability, and Daily Integration section). In terms of perceived usefulness, most participants believed that the system would be beneficial to them (mean 3.8, SD 0.82) and could effectively detect indicators of loneliness (mean 3.9, SD 0.87). These ratings support the conceptual value of the system. However, several participants expressed concerns about the transparency of how loneliness was inferred, especially in relation to physiological data. This issue was elaborated further in the focus group discussions (see Trust, Privacy, and Data Control section). Additionally, the dimension of understanding and clarity of feedback was rated highly. All participants reported being able to interpret the outputs provided by the accompanying mobile app (mean 4.5, SD 0.52), indicating a positive perception of the interface’s communicative clarity. However, ratings related to trust and privacy were more mixed. Participants expressed moderate trust in the system’s ability to reliably monitor loneliness (mean 3.4, SD 0.51). Responses to data sharing (mean 3.6, SD 0.84) and continuous monitoring (mean 3.8, SD 1.13) were generally positive, but the higher SDs suggest a divergence in acceptance of long-term monitoring technologies among participants. These differences were explored in depth during focus group discussions, where participants reflected on the roles of caregivers and family members in accessing sensitive data and expressed a wide range of perspectives on appropriate

XSL•FO
RenderX

data governance (see the Adoption, Applicability, and Ethical Futures section). Furthermore, the comfort dimension received the most consistently high ratings. Items relating to the comfort of the sensing garment (mean 4.3, SD 0.48), sensing furniture (mean 4.9, SD 0.32), and electronic textile materials (mean 4.3, SD 0.48) showed both high mean scores and low variability, indicating strong consensus among participants. Observational data also support these findings, with most participants exhibiting relaxed body language, positive comments, and tactile exploratory behaviors consistent with physical ease and embodied comfort. Finally, the dimension of overall system acceptance was rated positively (mean 3.9, SD 0.73), and participants reported moderate willingness to recommend the system to others (mean 3.7, SD 0.67). These findings suggest general acceptance and openness to the concept while also indicating further specific design improvement required to support long-term compliance.

Overall, these quantitative results provide an initial understanding of participants' functional and emotional responses to the system. They also helped shape the focus of the thematic analysis by identifying areas of strong consensus and divergence, which are further explored through focus group discussions, co-design artifacts, and sensory ethnographic observations in the Thematic Analysis section.

Thematic Analysis

Overview

Thematic analysis of the focus group discussions, co-design artifacts, and sensory ethnographic observations resulted in the identification of 4 main themes, each encompassing multiple subthemes, including wearability, usability, and daily integration; trust, privacy, and data control; perceptions of loneliness and the limits of detection; and adoption, applicability, and ethical futures.

Wearability, Usability, and Daily Integration

Garment Preferences and Adaptive Design

Participants expressed various preferences regarding the design, material, and format of the smart garment. These preferences emphasize the value of adaptable design that can align with individual lifestyles, seasonal changes, and social settings.

One of the most frequently raised concerns was about high temperature during warm weather. While participants acknowledged that the current version of the smart garment featured long sleeves to facilitate wearability within the workshop context, some participants questioned its year-round practicality: "I was just wondering how practical it would be in the summer to wear when it is hot."

Participants further expressed their expectations for varied design options based on seasonal needs and daily routines:

> *In the winter, a vest would be something of choice. But in the summer, maybe a t shirt. But then obviously, you'd have to provide different styles.*

A couple of participants suggested to design the sensing garment into sleeveless styles such as vests and sports bras, which would allow users to retain their preferred outerwear while still benefiting from the system's embedded sensing functions:

> *I would want something that more comfortable and I can put my own clothes on. So for me, I would prefer a vest or something like a sports bra.*

Additionally, some older people concerned about the visibility of sensing components, such as metal press-fit snaps and the data acquisition unit. For example, too many metal snaps on the outside of clothing were considered potentially offensive, as participants indicated that they would not want others to know that they were wearing a system designed to monitor loneliness. This reflected broader sensitivities around emotional health and a strong preference for unobtrusive and socially invisible technologies. While some participants found these acceptable on activewear, they felt such elements appeared out of place on more casual garments like t-shirts.

These concerns also extended beyond aesthetic considerations, but also included social signaling and potential stigma. Participants expressed a preference for discreet designs that would not attract unnecessary attention or provoke inquiries:

> *In a perfect world, I'd like it to be invisible, because otherwise you may spend half your life explaining ... people are going to say "what's that?" And do you want to discuss the fact that you're being monitored for loneliness with people you're not necessarily that close to?*

These findings were further supported during the co-design sessions. Participants proposed alternative sensor attachment mechanisms beyond the current snap-on method, suggesting modular sensor units that could integrate with their existing personal clothing. Outcomes from the co-design activities included vest sketches with internal linings to conceal sensor modules and annotations, indicating preferences for "subtle seams" and "concealed fasteners." One participant also suggested the concept of a "pin-on sensor," reflecting a desire for wearables that conform to users' existing dressing habits, rather than imposing new ones. This highlights the importance of designing smart loneliness monitoring systems that seamlessly integrate into users' daily lives and personal style, thereby enhancing the likelihood of long-term adoption.

Sensor Placement and Alternative Technology

During the focus group discussions, participants expressed a range of concerns and preferences regarding the placement of sensors and the physical dimensions of the monitoring components. While most participants were satisfied with the flexibility and comfort of the electronic textile circuit, a key issue identified was the discomfort caused by the rigid data acquisition module within the sensing garment:

> *About the little hard board (data collection unit), that is to me kind annoying. It's just wherever you put it.*

Some participants also asked whether the hardware could be miniaturized and made less obtrusive, suggesting that improvements in physical design could significantly enhance adoption. This perspective was reflected in several co-design artifacts, where participants reimagined the data module as a smaller patch or accessory. During the co-design sessions, many

XSL•FO
RenderX

participants proposed relocating the data unit to the back or side of the garment, thereby reducing bulk and improving comfort around the front torso area.

Beyond physical discomfort, the focus group also revealed diverse preferences regarding sensor location, particularly in relation to wrist-based monitoring. Some participants clearly expressed aversion to wearing anything on their wrist: "I don't wear anything on my wrist and I don't particularly want to."

In contrast, others found wrist-based sensing beneficial, especially those already familiar with commercial wearable devices: "I'm already using my watch to collect my health data, and I find it very convenient."

These conflicting attitudes were also observed in the sensory ethnographic field notes, which recorded moments of hesitation and hand gestures when participants explored the heart rate sensor embedded within the sleeve of the prototype garment. This further highlights the tactile and cognitive responses that influence user acceptance.

Moreover, several participants raised concerns about the technical requirements for physiological data accuracy. One question reflected a broader concern about the need for reliable biometric sensing: "Do the garment and furniture sensors need to be in close contact with the body to get accurate results?"

This suggests a critical design tension in smart wearable technologies between comfort and data accuracy. Older users preferred sensor systems that offer customizability, modularity, and interchangeable placement. Future iterations of the system should therefore not only miniaturize key components but also offer multiple sensor placement options, enabling users to select configurations that best align with their comfort, personal habits, and lifestyle.

Modular Usability

Modularity is a core feature of the system prototype, designed to enhance ease of maintenance, personal adaptability, and user autonomy. Across focus group discussions, co-design outcomes, and sensory ethnographic observations, participants generally endorsed the modular design principle while also highlighting practical challenges related to charging, cleaning, and reassembly.

Some older participants raised concerns about whether individuals with physical limitations would be able to perform these tasks independently:

> *I didn't have trouble taking the components out and putting them back, but if one hand wasn't very agile, it would be hard to do. These steps really require both hands. For anyone older, or with arthritis, it might be difficult to pull these things out of such a small, tight pocket.*

This concern was also reflected in co-design artifacts, where participants proposed simplified fastening mechanisms or introduced concepts such as magnetic snap-in connectors to reduce the burden of fine motor control. Ethnographic observations further documented moments of hesitation, uncertainty, or participants seeking assistance when attempting to detach or reattach sensor modules.

Additionally, some participants suggested that the system should include charging notifications within the app:

> *It's quite good to know that this stuff doesn't need to be charged every day like a smartwatch, but should there be a reminder in the app? You know, because you don't charge it daily, you might forget when it does need charging.*

Despite these operational challenges, participants consistently affirmed the value of modularity. They appreciated the separation of electronic components from the textile base not only for practical purposes, such as maintaining and laundering, but also for the potential long-term benefits, such as upgrading or replacing individual components over time. In co-design templates, several participants proposed personalized configurations, suggesting that different sensor modules could be swapped or added according to evolving health needs.

Material Comfort and Sensory Feedback

Material comfort is a critical factor influencing participants' responses to the smart loneliness monitoring system, particularly in relation to fabric texture, thermal regulation, and skin contact. Several participants expressed discomfort with synthetic textiles, especially when worn directly against the skin:

> *Wearing it as an outer layer is fine, I think it's quite comfortable. But if I were to wear it close to my skin, like a vest or a t-shirt, I definitely wouldn't want it to be polyester, because I find polyester too hot and sweaty when worn directly against the skin.*

During the co-design activities, participants were given fabric samples, including 100% cotton, cotton blends, wool blends, and polyester-spandex, and they were invited to annotate their preferences directly onto garment outline templates. Cotton and cotton-blend fabrics were the most frequently selected, accompanied by annotations such as "soft," "breathable," and "not itchy."

Comfort was also related to personal experiences and medical histories. For example, one participant noted that clothing design needs to consider changes in tactile sensitivity in the postoperative area:

> *I had a mastectomy. I don't have any breasts. I hate anything that scratches. It's just the normal is not normal anymore. You know what I mean?*

Compared to the sensing garments, the sensing furniture components such as the sensing pillow and cushion were generally perceived as more comfortable and less intrusive. This distinction was further supported by sensory ethnographic observations, which captured some participants' nuanced physical interactions with the garments. While wearing the smart clothing, some participants often made subtle adjustments to collars, pulled at sleeves, or ran their fingers along seam lines. Some participants hesitated before putting on the garment or asked whether the sensors would touch their skin directly. In contrast, interactions with the smart furniture were more relaxed. Most participants sat down without instruction, leaned back comfortably, and engaged in conversation while using the seat pad.

XSL•FO
RenderX

Contextual Integration and Lifestyle Compatibility

Participant feedback indicated that the acceptability of the system was closely related to its ability to seamlessly integrate into users' everyday lives, domestic environments, and personal routines. For example, smart furniture components were generally perceived as less intrusive and more acceptable for long-term and low-effort engagement:

> *People do tend to sit in the same seat every day, in the same place to watch TV. If someone was sitting there all day watching TV or doing something else, you'd find it very useful.*

In contrast, concerns were raised about the disruption that smart garments might cause to the unpredictability of daily routines. One participant described how the demand for continuous wear might not be compatible with their lifestyle:

> *I do so many things during the day. When I come back from the garden, I might have sweated or gotten dirty and need to change clothes and then, oops, I might forget to put on the smart garment again.*

This issue was further reflected in sensory ethnographic field notes, which captured several participants expressing uncertainty about whether they were "wearing it correctly" or whether the sensors would still function properly after shifting position.

During the co-design sessions, participants engaged with home layout templates, marking preferred sensor locations. They often placed sensors in areas associated with habitual furniture use, such as a reading chair, dining table, or frequently used seating areas. Some participants also proposed to integrate the system as part of standard domestic infrastructures in care homes: "You just install it when someone moves in for safety."

These insights suggest that smart loneliness monitoring systems should not only be conceived as stand-alone technologies, but rather as components within a broader ecosystem of smart living, with potential to be embedded into existing domestic practices and infrastructural frameworks.

Trust, Privacy, and Data Control

Conditional Trust and System Reliability

Participants across both workshops expressed a degree of trust in the smart loneliness monitoring system but consistently emphasized that trust in such technologies is not taken for granted. Several participants highlighted that trust would need to be earned over time, through demonstrable functionality and accuracy in real-life use:

> *I think I'd have to actually wear it and see it identify (my loneliness) without me saying anything. That would be the only way to really learn that it was working. What would happen at the end of the day? Would it give you a ping? Something to say "woo! Looks like you're lonely at the moment." And if I was in the middle of having a conversation with somebody and feeling perfectly fine, I know I would not trust it.*

Moreover, past experiences with commercial wearable monitoring devices appeared to shape users' current skepticism. One participant referenced their partner's experience:

> *My husband uses a Fitbit, but it clearly doesn't record his steps accurately. So I can't fully trust it.*

These encounters with inaccurate sensing may contribute to reserved user fatigue or doubts about the credibility of wearable sensors. These doubts also extended to smart monitoring systems, especially when applied to emotionally complex and subjective states like loneliness.

While most participants felt confident interpreting the outputs presented in the accompanying mobile app, they nonetheless highlighted a need for greater transparency and interpretability in system feedback. During the co-design sessions, participants proposed a range of suggestions to improve algorithmic explainability, including the addition of visual indicators such as:

> *Why is it saying I'm lonely?*
>
> *What data triggered this message?*

Another recommendation was the inclusion of a feedback confirmation mechanism, enabling users to validate the system feedback, thus contributing to a dynamic and trust-building model. One participant proposed a "check-in" feature, whereby if the system identified them as potentially lonely, they could choose to confirm or dismiss the notification. Over time, such feedback loops would allow the algorithm to learn from user responses, thereby refining its accuracy and building user confidence.

Customizable Data-Sharing Preferences

While participants acknowledged the potential value of sharing data with others particularly in the context of health or emotional support, several older adults highlighted the importance of retaining control over their data-sharing choices:

> *I feel that collecting this data all the time is quite intrusive. I might not want my daughter to know how I'm feeling at that moment, and I certainly don't want her using an app to monitor me. I want to stay in control of my emotions, and I feel this would take that control away.*

Conversely, another participant expressed openness to sibling-based support, provided that geographic distance warranted it. Interestingly, when asked, "You wouldn't want your daughter to see your data, would you want to see your mother's?" the participant hesitated. These responses highlight how data sharing preferences are relational and context-dependent and may vary depending on the role of the recipient as caregiver or care recipient.

Additionally, several participants expressed a preference for conditional sharing models, where data would only be shared under specific circumstances:

> *If I'm having a breakdown at home because I'm feeling lonely and haven't seen anyone for three weeks, then I do want them to know. But if I'm just feeling a bit off and can't be bothered to go to the coffee shop, I don't necessarily want to share that.*

During the co-design sessions, participants proposed improvements to the current permissions interface. Users expressed a desire to select which types of data such as

XSL•FO
RenderX

emotional states, activity levels, or physiological indicators would be visible to specific recipients, including family members, clinicians, or community caregivers. Some also suggested adding a dashboard that clearly shows "who can see what," along with visual indicators to support transparency and ease of management.

Ethical Concerns and Data Ownership

Across both workshops, while the majority of participants expressed openness toward the use of monitoring technologies, they also emphasized that trust in such systems depends not only on accuracy but also on transparency of purpose, data flow, and long-term governance.

A recurring concern was "Who owns the data?" and often followed by anxiety about potential commercial exploitation. As one participant asked: "What if the data gets sold? Who's to say it won't be sold?"

The prospect that personal health or emotional data might be commodified was troubling for many older adults, particularly given the lack of clear regulation surrounding data collected outside clinical systems.

Concerns also extended to the broader implications of artificial intelligence governance in connection with wearable monitoring systems: "I understand that these technologies are developed with good intentions, but I do worry that they could be repurposed for surveillance or behavioral manipulation."

During the co-design sessions, many older adults indicated that these ethical concerns did not necessarily result in rejection of the system. Rather, they expressed a desire for greater clarity regarding the system's governance model, data stewardship, and effective plans for future use. These findings suggest that ethical acceptability is not solely a matter of obtaining "informed consent" at the point of use but requires ongoing transparency and participatory data governance. Future iterations of the system should consider the development of interactive tools that clearly and accessibly communicate key information regarding data provenance ownership and rights.

Perceptions of Loneliness and the Limits of Detection

The Subjectivity of Loneliness

Some participants believed that loneliness is fundamentally a subjective and emotionally complex experience, one that cannot be directly inferred from behavioral or physiological signals alone. This perspective occurred in our earlier focus group discussions, where several participants questioned the assumption that sensor data could reliably infer emotional states: "What you feel inside can't be monitored by anyone, no sensor can pick that up."

Participants also noted that loneliness is highly individualized and not necessarily linked to physical solitude: "You can feel lonely in a crowded room but feel fine when you're alone."

This further raised questions among participants about whether the algorithm could accurately identify individualized feelings of loneliness. In response, the investigator (JR) clarified that while current algorithmic models may be developed from broad datasets, their core functionality is intended to adapt to individual patterns. The system is designed to learn over time with correct or incorrect feedback, establishing and refining a personalized emotional profile, thereby improving its accuracy.

Moreover, participants noted that loneliness is not exclusive to older adults but occurs across different ages and life stages, highlighting the need for the system to detect emotional nuance rather than demographic generalization.

In the co-design sessions, some participants expressed discomfort with the term "loneliness," describing it as "too strong" or "too negative." They proposed softer alternatives, such as "reflective state" or "well-being indicator." Others annotated their interface sketches with prompts like "How are you feeling today?" in place of system-generated loneliness labels. These design annotations suggest that users may prefer tools that prompt self-reflection, rather than systems that presume to define their emotional states on their behalf.

Algorithmic Assumptions and Multimodal Analysis

As discussed in the previous subsection, participants questioned the logic behind the algorithm used for loneliness detection. For instance, they expressed concern that feedback based solely on low physical activity might lead to false positives: "You're detected as being very still, but you're not lonely, you're just enjoying your book."

In response, researchers explained that the current system applies multimodal integration, combining physiological signals with behavioral data for a more comprehensive analysis. While participants appreciated this approach, they also asked whether the system could incorporate additional personalized factors to further improve detection accuracy: "If I'm motionless, and those factors you mentioned lead the system to assume I'm lonely, then I'm wondering, what other factors could be added?"

This desire for individualized calibration was also evident in the co-design artifacts. On the smart garment and furniture templates, some participants annotated notes such as "Don't assume lack of movement means I'm feeling low." Others, particularly older adults, proposed that the system should seek user confirmation before tagging as a loneliness event, effectively embedding emotional subjectivity into the algorithmic process.

Adoption, Applicability, and Ethical Futures

Adoption Among the Older Adults

While participants generally recognized the value of the systems in their later life, they also noted that those who might benefit most should be very old adults or individuals strongly invested in their independence (may be the least willing to adopt it). Factors such as psychological resistance and identity-related concerns were seen as major potential barriers to real-world implementation.

Most participants in both of the workshops were still socially active and engaged but reflected on family members who were socially isolated yet unwilling to engage with monitoring technologies or accept support:

XSL•FO
RenderX

> *My mum is very old and lives on her own. She's obviously lonely, but there's no way she'd get involved in something like this.*

Others expressed similar insight:

> *She won't accept any help. Yes, even though she's clearly spiraling because of it.*

These reflections highlight a tension between need, self-perception, and autonomy. The concept of "independent living" was both a source of pride and a practical barrier. Older adults may perceive acknowledging loneliness or using assistive technologies as a threat to their autonomy or even as an admission of vulnerability.

This resistance was also observed in the sensory ethnographic field notes. While most participants interacted with system components during the experience, a small number showed hesitation or disengagement when asked to wear the garment or respond to app prompts. Some made dismissive remarks such as: "This seems something for people who need looking after."

In co-design sessions, some participants openly stated that they found it a bit difficult to imagine themselves using such a system even hypothetically because it felt "irrelevant" or "only for people in worse situations." These views may reflect a desire to maintain a sense of competence and independence, even in the face of known risks or increasing social withdrawal.

Timing of App

During the focus group discussions, participants also reflected on when these technologies should be applied. Some participants thought it would be best to introduce the system before severe loneliness or functional decline occurs: "Maybe you need to catch people before they get to that point, so you can do it as a prevention rather than a late intervention."

Some participants thought that if people fall into severe isolation, their willingness to engage with new technologies may be significantly reduced: "By the time they need it, they may not want to learn it."

In addition, participants also mentioned that they may be more willing to use the system when they still feel in control, an insight that is consistent with the discussion in the Trust, Privacy, and Data Control section that trust is built through gradual voluntary participation rather than sudden or mandatory use.

Some participants proposed a "onboarding stage" during the co-design session, such as "Let me start with one feature first" or "Phase 1: Activity tracking only, no reminders yet." Others suggested setting up a "trial period" for the system so that users can experience it during this period without worrying about data being misunderstood or shared too early.

Expanded Health Monitoring

While loneliness detection was the primary aim of the system, many participants expressed strong interest in expanding its functionality to support broader physiological and behavior monitoring: "It's good to know the system can monitor so many different things, but could it also include more features?"

Some participants also highlighted their specific personal monitoring needs: "I often need to drink water, otherwise I get heart palpitations. could it tell me if I'm dehydrated?"

These ideas were further developed during the co-design activities, where participants added additional monitoring points to the smart garment framework diagrams. Suggestions included integrating electrocardiography and blood pressure tracking and hydration-level detection. Rather than replacing the core functionality related to loneliness, these suggestions were seen as complementary integrations that could increase the system's everyday utility and relevance by addressing users' broader well-being in a more holistic and meaningful way.

Linked Intervention

When participants reflected on the system's potential to monitor physiological and behavioral signals, the discussion naturally extended to scenarios involving health emergencies and critical incidents. Many raised concerns about automated alerts and connected interventions:

> *If you're in a state of severe loneliness or at some kind of risk, the next level of concern is whether social services would actually respond, whether your GP would be notified, or a nurse, or a district nurse would come out. That's the real worry.*

While many participants appreciated the system's potential to issue alerts or prompts, they expressed hesitancy about fully autonomous system actions, particularly in the context of emotional monitoring. Some expressed discomfort with the idea of automated triage, instead showing a preference for human-mediated intervention: "I would like a caregiver or clinician to review the data before contacting me."

In the co-design sessions, participants proposed customizable alert settings, including adjustable emergency thresholds, such as "Only trigger an alert if an abnormal signal persists for more than 10 minutes." and "Notify family members first, then professionals." These suggestions reflect a strong desire for tiered intervention logic, whereby users can define the severity of signals required to activate alerts as well as the order and type of recipients to be notified. This approach highlights the need for a personalized response, rather than a one-size-fits-all automation model.

Furthermore, although stakeholders such as care providers were not present in either workshop, participants actively imagined various relational configurations. Some participants preferred family members as first contacts, while others, particularly those living alone, favored designated professional care networks.

Discussion

Principal Findings

This study examined the acceptability and perceived usability of a novel smart loneliness monitoring system for older adults, comprising sensing garments, furniture, and a companion mobile app. Through user-centered evaluation and experience-based co-design, this study aimed to comprehensively explore users' practical, emotional, and ethical concerns to the system and to provide actionable design insights for future development. While

XSL•FO
RenderX

prior studies in pervasive computing and ambient-assisted living have focused on detecting behavioral correlates of social isolation, few have investigated how older adults themselves experience, interpret, and negotiate such monitoring systems [34-36]. Therefore, our findings contributed to bridging the gap between technical feasibility and user acceptability in the emerging field of smart mental health textiles.

Previous research has shown that older adults' acceptance of smart monitoring systems is heavily affected by perceived usefulness, particularly whether the data collected support meaningful or supportive interventions [17,56,57]. Given the practical and societal significance of loneliness monitoring, our prior research had explored older adults' initial design needs alongside stakeholder perspectives at the conceptual level [27]. Building on this foundation, this study combines smart textile design and sensing technology to develop prototypes that allow participants to physically experience, evaluate, and reimagine the system. This study identified 4 main domains affecting user acceptance and future design improvement, including wearability, usability and integration; trust, privacy and data control; limitations of loneliness monitoring; and future adoption, applicability, and ethical considerations.

Our findings highlighted the importance of adaptability and lifestyle compatibility in determining system acceptability, which are key aspects emphasized in previous research on wearable health technologies for older adults. Prior studies have shown that comfort, convenience, and discretion strongly influence engagement with wearable devices, particularly among older users who may have heightened sensitivities to fabric texture or skin contact [15]. In our study, although most participants were satisfied with synthetic materials for outerwear use, individual preferences and health history highlighted the value of personalized textile options. Participants appreciated the modular design and comfortable electronic textile integration, which enhanced wearability. However, unlike previous work that primarily assessed ergonomics or fit, our findings underscored the importance of emotional comfort and social invisibility. Participants expressed concerns about the visibility of sensing components and the potential stigma associated with being perceived as "monitored." This highlights an important extension of the current understanding of usability from physical comfort to psychosocial comfort. Additionally, the discussion around seasonal practicality and clothing preferences, such as the suggestion to adopt undergarment formats like vests or sports bras, introduces a novel consideration for thermal and social appropriateness and compatibility with individual everyday routines. While some previous smart garment research explored aesthetic design [58], our findings suggest that adaptability to seasonal routines and existing clothing habits is critical for long-term adoption. Moreover, the discomfort caused by rigid sensor modules further emphasizes the need for miniaturized and flexible electronics, a challenge also noted in emerging literature on e-textile scalability and integration [59,60]. Compared to sensing garments, sensing furniture was perceived as more comfortable, less obtrusive, and better aligned with habitual behaviors.

Trust, privacy, and data control were also key to user acceptance. Prior studies on digital health and remote monitoring technologies have consistently shown that trust and perceived data security are prerequisites for sustained engagement among older adults [61]. Our findings support these observations but extend them by revealing that participants' conditional trust depended not only on privacy assurances but also on the reliability, interpretability, and transparency of system feedback. These aspects highlighted in this study collaborated with previous studies on smart home and telehealth systems. However, unlike earlier work that primarily emphasized the role of institutional trust in health care providers [62], our participants focused on personal data sovereignty, and they want to see, understand, and adjust what the system infers about them in real time. While many older adults found the app's real-time feedback intuitive and easy to use, their past experiences with commercial wearables, such as smartwatches and fitness trackers, triggered skepticism regarding its accuracy when interpreting subtle emotional or behavioral changes. This echoes concerns in prior research that algorithmic opacity undermines user confidence in affective or well-being monitoring [63]. Our findings show that trust must be earned gradually through use and supported by transparent feedback mechanisms that allow users to confirm or challenge system outputs, which is an important element rarely addressed and discussed in earlier studies. Users have also shown a great preference to granular control over data sharing, allowing users to tailor access to different stakeholders such as family members, clinicians, or caregivers. While prior work on privacy in older adults has discussed consent management in general terms [64], our participants preferred a dynamic and contextual control that could match their current mental state and relationships. These insights extend the current literature by emphasizing that ethical acceptability is not achieved solely through initial consent, but through ongoing transparency, accountability, and user agency in data governance. Future development of wearable monitoring technologies should move beyond data protection compliance to include user-facing transparency features and active participatory data management frameworks.

Strengths and Limitations

The key strengths of this study lie in its multimethod integration of quantitative and qualitative approaches, combining structured questionnaires, focus groups, sensory ethnography, and experience-based co-design. This enabled a multifaceted understanding of older adults' experiences with the smart loneliness monitoring system and helped identify comprehensive user-driven directions for iterative improvement. However, several limitations also need to be acknowledged. First, all the participants were recruited from the United Kingdom, which may introduce a degree of regional bias and limit the generalizability of the findings to older adults in other geographic, cultural, or health care settings. In addition, the sample included relatively few male participants. Given that sex may influence perspectives on technology, privacy, and well-being, future research should aim to increase sex diversity to improve the representativeness of findings. Participants with diagnosed cognitive impairment were excluded during recruitment. Researchers applied the principles of the Mental Capacity Act (2005) to assess participants' capacity to understand the study and provide informed consent [40]. While

XSL•FO
RenderX

this ensured ethical participation and meaningful engagement with the system, the study does not include the perspectives of older adults living with cognitive impairment or dementia. Future work should explore how smart loneliness monitoring systems can be adapted or tailored for these populations, who may have different usability needs and vulnerabilities. Additionally, although participants met the inclusion criteria of being aged 65 years and older and having experienced loneliness, the majority of participants remained socially active, potentially limiting the generalizability of findings to more isolated or vulnerable populations. Moreover, the overall sample size was small. While small-scale qualitative studies can offer rich insight, findings should be interpreted as exploratory and hypothesis-generating rather than definitive. Finally, while the workshops were conducted in a controlled environment to allow in-depth interaction and observation, this setting may not fully capture the complexity of real-world use. To further strengthen the validity and practical relevance of the system, future research should involve longitudinal field testing in home or community settings. This would allow for continuous data collection over an extended period, enabling more robust analysis of behavioral and physiological patterns, as well as user engagement over time.

Conclusions

This study examined older adults' experiences with smart loneliness monitoring systems, including sensing garments, furniture, and a companion mobile app. Through immersive workshops combining structured questionnaires, focus groups, sensory ethnographic observation, and experience-based co-design, we investigated older users' practical, emotional, and ethical responses to the system, providing actionable design insights to inform future development. These findings indicate that while participants generally accepted the concept of monitoring via smart textile wearables and furniture, their willingness to adopt such systems over time is highly dependent on usability, personalization, lifestyle compatibility, and perceived control. Participants consistently emphasized the importance of adaptable design that respects bodily comfort, domestic routines, and personal identity. Moreover, the modular smart system developed in this study demonstrates strong potential as a discreet and passive sensing platform for psychological and social health indicators. However, continued user-centered iteration, broader real community testing, and deeper integration with health care infrastructure are required to ensure its future success in real-world applications.

Acknowledgments

The authors would like to thank all participants and collaborators who contributed to this study.

Funding

This work was supported by the UK Engineering and Physical Sciences Research Council and the National Institute of Health and Care Research (grant EP/W031434/1).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

YZ was responsible for conceptualization, methodology, investigation, formal analysis, data curation, writing the original draft, and editing the manuscript. JR was responsible for recruitment, formal analysis, data curation, and reviewing and editing the manuscript. FM was responsible for conceptualization, methodology, and reviewing and editing the manuscript. AP was responsible for conceptualization and reviewing and editing the manuscript. MA and AT was responsible for funding acquisition and reviewing and editing the manuscript. SO was responsible for supervision and funding acquisition. WL was responsible for supervision, funding acquisition, resources, and reviewing and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Structured observation form.
[PDF File (Adobe PDF File), 32 KB - jmir_v28i1e81027_app1.pdf]

Multimedia Appendix 2
Likert-scale questionnaire.
[PDF File (Adobe PDF File), 1514 KB - jmir_v28i1e81027_app2.pdf]

Multimedia Appendix 3
Co-design template.

XSL•FO
RenderX

[PDF File (Adobe PDF File), 340 KB - jmir_v28i1e81027_app3.pdf]

References

1. Guglielmi V, Colangeli L, Parrotta ME, Ciammariconi A, Milani I, D'Adamo M, et al. Social isolation and loneliness in non-communicable chronic diseases: impact of COVID-19 pandemic, population aging and technological progress. Nutr Metab Cardiovasc Dis 2025;35(6):104015 [FREE Full text] [doi: 10.1016/j.numecd.2025.104015] [Medline: 40189996]
2. Walsh BE, Rottenberg J, Schlauch RC. Why loneliness requires a multidimensional approach: a critical narrative review. Nat Mental Health 2025;3(2):175-184. [doi: 10.1038/s44220-024-00382-3]
3. Teo RH, Cheng WH, Cheng LJ, Lau Y, Lau ST. Global prevalence of social isolation among community-dwelling older adults: a systematic review and meta-analysis. Arch Gerontol Geriatr 2023;107:104904. [doi: 10.1016/j.archger.2022.104904] [Medline: 36563614]
4. Aalto UL, Bonin-Guillaume S. Loneliness among older people exacerbated by the COVID-19 pandemic. J Nutr Health Aging 2023;27(8):617-618. [doi: 10.1007/s12603-023-1968-z]
5. Benson JA, McSorley VE, Hawkley LC, Lauderdale DS. Associations of loneliness and social isolation with actigraph and self-reported sleep quality in a national sample of older adults. Sleep 2021;44(1):zsaa140 [FREE Full text] [doi: 10.1093/sleep/zsaa140] [Medline: 32691067]
6. Emerson E, Fortune N, Llewellyn G, Stancliffe R. Loneliness, social support, social isolation and wellbeing among working age adults with and without disability: cross-sectional study. Disabil Health J 2021;14(1):100965 [FREE Full text] [doi: 10.1016/j.dhjo.2020.100965] [Medline: 32843311]
7. Latikka R, Rubio-Hernández R, Lohan ES, Rantala J, Nieto Fernández F, Laitinen A, et al. Older adults' loneliness, social isolation, and physical information and communication technology in the era of ambient assisted living: a systematic literature review. J Med Internet Res 2021;23(12):e28022. [doi: 10.2196/28022] [Medline: 34967760]
8. Mushtaq R, Shoib S, Shah T, Mushtaq S. Relationship between loneliness, psychiatric disorders and physical health? A review on the psychological aspects of loneliness. J Clin Diagn Res 2014;8(9):WE01-WE04 [FREE Full text] [doi: 10.7860/JCDR/2014/10077.4828] [Medline: 25386507]
9. Prabhu D, Kholghi M, Sandhu M, Lu W, Packer K, Higgins L, et al. Sensor-based assessment of social isolation and loneliness in older adults: a survey. Sensors (Basel) 2022 Dec 16;22(24):9944 [FREE Full text] [doi: 10.3390/s22249944] [Medline: 36560312]
10. Fakoya OA, McCorry NK, Donnelly M. Loneliness and social isolation interventions for older adults: a scoping review of reviews. BMC Public Health 2020;20(1):129 [FREE Full text] [doi: 10.1186/s12889-020-8251-6] [Medline: 32054474]
11. Ratcliffe J, Kanaan M, Galdas P. Reconceptualising men's loneliness: an interpretivist interview study of UK-based men. Soc Sci Med 2023;332:116129 [FREE Full text] [doi: 10.1016/j.socscimed.2023.116129] [Medline: 37531906]
12. Rees J, Liu W, Canson J, Crosby L, Tinker A, Probst F, et al. Qualitative exploration of the lived experiences of loneliness in later life to inform technology development. Int J Qual Stud Health Well-Being 2024;19(1):2398259 [FREE Full text] [doi: 10.1080/17482631.2024.2398259] [Medline: 39305060]
13. Johnson KT, Zawadzki MJ, Kho C. Loneliness and sleep in everyday life: using ecological momentary assessment to characterize the shape of daily loneliness experience. Sleep Health 2024;10(4):508-514 [FREE Full text] [doi: 10.1016/j.sleh.2024.04.003] [Medline: 38839482]
14. Gedam S, Paul S. A review on mental stress detection using wearable sensors and machine learning techniques. IEEE Access 2021;9:84045-84066. [doi: 10.1109/access.2021.3085502]
15. Moore K, O'Shea E, Kenny L, Barton J, Tedesco S, Sica M, et al. Older adults' experiences with using wearable devices: qualitative systematic review and meta-synthesis. JMIR Mhealth Uhealth 2021;9(6):e23832 [FREE Full text] [doi: 10.2196/23832] [Medline: 34081020]
16. Schrempft S, Jackowska M, Hamer M, Steptoe A. Associations between social isolation, loneliness, and objective physical activity in older men and women. BMC Public Health 2019;19(1):74 [FREE Full text] [doi: 10.1186/s12889-019-6424-y] [Medline: 30651092]
17. Meredith SJ, Cox N, Ibrahim K, Higson J, McNiff J, Mitchell S, et al. Factors that influence older adults' participation in physical activity: a systematic review of qualitative studies. Age Ageing 2023;52(8):afad145. [doi: 10.1093/ageing/afac125.002]
18. Frayn M, Livshits S, Knäuper B. Emotional eating and weight regulation: a qualitative study of compensatory behaviors and concerns. J Eat Disord 2018;6:23 [FREE Full text] [doi: 10.1186/s40337-018-0210-6] [Medline: 30221002]
19. Hawkley LC, Cacioppo JT. Loneliness and blood pressure in older adults: defining connections. Aging Health 2010;6(4):415-418. [doi: 10.2217/ahe.10.41]
20. Gao F, Liu C, Zhang L, Liu T, Wang Z, Song Z, et al. Wearable and flexible electrochemical sensors for sweat analysis: a review. Microsyst Nanoeng 2023;9:1-21 [FREE Full text] [doi: 10.1038/s41378-022-00443-6] [Medline: 36597511]
21. Schutter N, Holwerda T, Stek M, Dekker J, Rhebergen D, Comijs H. Loneliness in older adults is associated with diminished cortisol output. J Psychosom Res 2017;95:19-25. [doi: 10.1016/j.jpsychores.2017.02.002]
22. Ejupi A, Menon C. Detection of talking in respiratory signals: a feasibility study using machine learning and wearable textile-based sensors. Sensors (Basel) 2018;18(8):2474 [FREE Full text] [doi: 10.3390/s18082474] [Medline: 30065177]

XSL•FO
RenderX

23. Kim M, Hong S, Youm S. Development of an intrinsic health risk prediction model for camera-based monitoring of older adults living alone. Sci Rep 2022;12(1):18855 [FREE Full text] [doi: 10.1038/s41598-022-23663-2] [Medline: 36344806]
24. Vuegen L, Van DBB, Karsmakers P, Van HH, Vanrumste B. Monitoring activities of daily living using Wireless Acoustic Sensor Networks in clean and noisy conditions. 2015 Nov 05 Presented at: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); August 25-29, 2015; Milan, Italy p. 4966-4969. [doi: 10.1109/embc.2015.7319506]
25. van den Boer J, van der Lee A, Zhou L, Papapanagiotou V, Diou C, Delopoulos A, et al. The SPLENDID eating detection sensor: development and feasibility study. JMIR Mhealth Uhealth 2018;6(9):e9781 [FREE Full text] [doi: 10.2196/mhealth.9781] [Medline: 30181111]
26. Kalisch T, Theil C, Gosheger G, Ackmann T, Schoenhals I, Moellenbeck B. Measuring sedentary behavior using waist- and thigh-worn accelerometers and inclinometers—are the results comparable? Ther Adv Musculoskelet Dis 2022;14:1759720X221079256 [FREE Full text] [doi: 10.1177/1759720X221079256] [Medline: 35310836]
27. Probst F, Rees J, Aslam Z, Mexia N, Molteni E, Matcham F, et al. Evaluating a smart textile loneliness monitoring system for older people: co-design and qualitative focus group study. JMIR Aging 2024;7:e57622 [FREE Full text] [doi: 10.2196/57622] [Medline: 39688889]
28. Probst F, Liu W. Co-design of a loneliness monitoring system with older people and stakeholders. 2024 Presented at: DRS Biennial Conference Series; June 23-28, 2024; Boston, MA, United States URL: https://dl.designresearchsociety.org/drs-conference-papers/drs2024/researchpapers/134 [doi: 10.21606/drs.2024.545]
29. Dzedzickis A, Kaklauskas A, Bucinskas V. Human emotion recognition: review of sensors and methods. Sensors (Basel) 2020;20(3):592 [FREE Full text] [doi: 10.3390/s20030592] [Medline: 31973140]
30. Cosma G, Brown D, Battersby S, Kettley S, Kettley R. Analysis of multimodal data obtained from users of smart textiles designed for mental wellbeing. 2017 Aug 14 Presented at: 2017 International Conference on Internet of Things for the Global Community (IoTGC); July 10-13, 2017; Funchal, Portugal. [doi: 10.1109/iotgc.2017.8008974]
31. Khundaqji H, Hing W, Furness J, Climstein M. Smart shirts for monitoring physiological parameters: scoping review. JMIR Mhealth Uhealth 2020;8(5):e18092 [FREE Full text] [doi: 10.2196/18092] [Medline: 32348279]
32. Chen M, Ma Y, Song J, Lai C, Hu B. Smart clothing: connecting human with clouds and big data for sustainable health monitoring. Mobile Netw Appl 2016;21(5):825-845. [doi: 10.1007/s11036-016-0745-1]
33. Vuohijoki T, Ihalainen T, Virkki J. Smart clothing and furniture for supporting participation-co-creation concepts for daily living. SN Appl Sci 2023;5(4):110. [doi: 10.1007/s42452-023-05315-w]
34. Lodder GMA, Scholte RHJ, Goossens L, Engels RCME, Verhagen M. Loneliness and the social monitoring system: emotion recognition and eye gaze in a real-life conversation. Br J Psychol 2016;107(1):135-153. [doi: 10.1111/bjop.12131] [Medline: 25854912]
35. Jafarlou S, Azimi I, Lai J, Wang Y, Labbaf S, Nguyen B, et al. Objective monitoring of loneliness levels using smart devices: a multi-device approach for mental health applications. PLoS One 2024;19(6):e0298949 [FREE Full text] [doi: 10.1371/journal.pone.0298949] [Medline: 38900745]
36. Sarhaddi F, Azimi I, Niela-Vilen H, Axelin A, Liljeberg P, Rahmani AM. Maternal social loneliness detection using passive sensing through continuous monitoring in everyday settings: longitudinal study. JMIR Form Res 2023;7:e47950 [FREE Full text] [doi: 10.2196/47950] [Medline: 37556183]
37. Fernández-Caramés T, Fraga-Lamas P. Towards the internet-of-smart-clothing: a review on IoT wearables and garments for creating intelligent connected e-textiles. Electronics 2018;7(12):405. [doi: 10.3390/electronics7120405]
38. Tsai T, Lin W, Chang Y, Chang P, Lee M. Technology anxiety and resistance to change behavioral study of a wearable cardiac warming system using an extended TAM for older adults. PLoS One 2020;15(1):e0227270 [FREE Full text] [doi: 10.1371/journal.pone.0227270] [Medline: 31929560]
39. Rees J, Liu W, Ourselin S, Shi Y, Probst F, Antonelli M, et al. Understanding the psychological experiences of loneliness in later life: qualitative protocol to inform technology development. BMJ Open 2023;13(6):e072420 [FREE Full text] [doi: 10.1136/bmjopen-2023-072420] [Medline: 37336536]
40. Mental Capacity Act 2005. Statute Law Database. URL: https://www.legislation.gov.uk/ukpga/2005/9/section/1 [accessed 2025-12-23]
41. Hughes ME, Waite LJ, Hawkley LC, Cacioppo JT. A short scale for measuring loneliness in large surveys: results from two population-based studies. Res Aging 2004;26(6):655-672 [FREE Full text] [doi: 10.1177/0164027504268574] [Medline: 18504506]
42. Courage C, Baxter K. A Practical Guide to User Requirements Methods, Tools, and Techniques. Houston, TX: Gulf Professional Publishing; 2005.
43. Thell M, Edvardsson K, Aljeshy R, Ibrahim K, Warner G. A trauma support app for young people: co-design and usability study. JMIR Form Res 2025 Mar 18;9:e57789. [doi: 10.2196/57789]
44. Grieve N, Braaten K, MacPherson M, Liu S, Jung ME. Involving end users in the development and usability testing of a smartphone app designed for individuals with prediabetes: mixed-methods focus group study. JMIR Form Res 2025;9:e59386 [FREE Full text] [doi: 10.2196/59386] [Medline: 39935015]

XSL•FO
RenderX

45. Steen-Olsen EB, Pappot H, Hjerming M, Hanghoej S, Holländer-Mieritz C. Monitoring adolescent and young adult patients with cancer via a smart t-shirt: prospective, single-cohort, mixed methods feasibility study (OncoSmartShirt Study). JMIR Mhealth Uhealth 2024;12:e50620 [FREE Full text] [doi: 10.2196/50620] [Medline: 38717366]
46. Lupton D. Editorial: Towards sensory studies of digital health. Digit Health 2017;3:2055207617740090. [doi: 10.1177/2055207617740090]
47. Valtonen A, Markuksela V, Moisander J. Doing sensory ethnography in consumer research. Int J Consumer Studies 2010 Jun 09;34(4):375-380. [doi: 10.1111/j.1470-6431.2010.00876.x] [Medline: 22489612]
48. Sunderland N, Bristed H, Gudes O, Boddy J, Da Silva M. What does it feel like to live here? Exploring sensory ethnography as a collaborative methodology for investigating social determinants of health in place. Health Place 2012;18(5):1056-1067. [doi: 10.1016/j.healthplace.2012.05.007] [Medline: 22722015]
49. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989;13(3):319-340. [doi: 10.2307/249008]
50. Brooke J. SUS: A Quick and Dirty Usability Scale. Boca Raton, FL: CRC Press; 1996.
51. Venkatesh, Morris, Davis, Davis. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425. [doi: 10.2307/30036540]
52. Donetto S, Pierri P, Tsianakas V, Robert G. Experience-based co-design and healthcare improvement: realizing participatory design in the public sector. Des J 2015;18(2):227-248. [doi: 10.2752/175630615X14212498964312]
53. Mishra P, Pandey C, Singh U, Gupta A, Sahu C, Keshri A. Descriptive statistics and normality tests for statistical data. Ann Card Anaesth 2019;22(1):67. [doi: 10.4103/aca.aca_157_18]
54. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
55. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa] [Medline: 32100154]
56. Brauner P, Heek JV, Ziefle M. Age, gender, and technology attitude as factors for acceptance of smart interactive textiles in home environments—towards a smart textile technology acceptance model. 2017 Presented at: Proceedings of the 3rd International Conference on Information and Communication Technologies for Ageing Well and e-Health ICT4AWE; April 28-29, 2017; Porto, Portugal p. 13-24 URL: https://www.scitepress.org/Link.aspx?doi=10.5220/0006255600130024 [doi: 10.5220/0006255600130024]
57. Choi YK, Thompson HJ, Demiris G. Use of an internet-of-things smart home system for healthy aging in older adults in residential settings: pilot feasibility study. JMIR Aging 2020 Nov 10;3(2):e21964 [FREE Full text] [doi: 10.2196/21964] [Medline: 33170128]
58. Imbesi S, Scataglini S. A user centered methodology for the design of smart apparel for older users. Sensors (Basel) 2021;21(8):2804 [FREE Full text] [doi: 10.3390/s21082804] [Medline: 33923514]
59. Meena JS, Choi SB, Jung S, Kim JW. Electronic textiles: new age of wearable technology for healthcare and fitness solutions. Mater Today Bio 2023;19:100565 [FREE Full text] [doi: 10.1016/j.mtbio.2023.100565] [Medline: 36816602]
60. Zhou Y, Ratcliffe J, Molteni E, Patel A, Liu J, Mexia N, et al. Smart textile systems for loneliness monitoring in older people care: a review of sensing and design innovations. Adv Elect Mater 2025;11(16):e00300. [doi: 10.1002/aelm.202500300]
61. Kebede AS, Ozolins L, Holst H, Galvin K. Digital engagement of older adults: scoping review. J Med Internet Res 2022;24(12):e40192. [doi: 10.2196/40192] [Medline: 36477006]
62. Hou G, Li X, Wang H. How to improve older adults' trust and intentions to use virtual health agents: an extended technology acceptance model. Humanit Soc Sci Commun 2024 Dec 18;11(1):1677. [doi: 10.1057/s41599-024-04232-6]
63. Wanner J, Herm L, Heinrich K, Janiesch C. The effect of transparency and trust on intelligent system acceptance: evidence from a user-based study. Electron Mark 2022;32(4):2079-2102. [doi: 10.1007/s12525-022-00593-5]
64. Altawalbeh SM, Alkhateeb FM, Attarabeen OF. Ethical issues in consenting older adults: academic researchers and community perspectives. J Pharm Health Serv Res 2020;11(1):25-32 [FREE Full text] [doi: 10.1111/jphs.12327] [Medline: 33042231]

Abbreviations

TAM: technology acceptance model

XSL•FO
RenderX

Edited by A Stone; submitted 21.Jul.2025; peer-reviewed by HC Chiu, MDG Pimentel; comments to author 24.Oct.2025; revised version received 24.Dec.2025; accepted 29.Dec.2025; published 28.Jan.2026.

Please cite as:
Zhou Y, Rees J, Matcham F, Patel A, Antonelli M, Tinker A, Ourselin S, Liu W
Development and User-Centered Evaluation of Smart Systems for Loneliness Monitoring in Older Adults: Mixed Methods Study
J Med Internet Res 2026;28:e81027
URL: https://www.jmir.org/2026/1/e81027
doi:10.2196/81027
PMID:

©Yi Zhou, Jessica Rees, Faith Matcham, Ashay Patel, Michela Antonelli, Anthea Tinker, Sebastien Ourselin, Wei Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Guide on Selection of Optimal Motivational Themes for Use in a Clinical Trial Recruiting Black US Adults: Survey Study

Tamunotonye Harry[1*], MS; Jingyi Cao[2], MPH; Zaib Hussain[3], MBBS, MPH; Ruth-Alma N Turkson-Ocran[2], MPH, APRN, PhD; Stephen P Juraschek[2], MD, PhD; Timothy P Lahey[4], MD, MMSc; Timothy B Plante[4], MD, MHS; Yuanyuan Feng[1*], PhD

[1]Department of Computer Science, University of Vermont, 85 South Prospect Street, Burlington, VT, United States

[2]Division of General Medicine, Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, United States

[3]Division of Cardiology, Johns Hopkins University School of Medicine, Baltimore, MD, United States

[4]Department of Medicine, Larner College of Medicine, University of Vermont, Burlington, VT, United States

[*]these authors contributed equally

Corresponding Author:
Tamunotonye Harry, MS
Department of Computer Science, University of Vermont, 85 South Prospect Street, Burlington, VT, United States

Abstract

Background: Black adults in the United States face significant cardiovascular health disparities, which are likely exacerbated by the underrepresentation of Black adults in cardiovascular clinical trials. The Black adult US population has experienced unique historical events, discriminatory practices, and practical obstacles that might contribute to this underrepresentation in clinical trials. An improved understanding of motivations that encourage or discourage participation in cardiovascular clinical trials can lead to more effective clinical trial recruitment and help mitigate these cardiovascular health disparities.

Objective: This study aimed to determine, using an online survey, which motivational themes in clinical trial recruitment advertisements were most effective in encouraging Black adults to participate in a hypertension-focused trial. We also examined how trust in health care and various demographic factors influenced their decision to participate.

Methods: We conducted an online survey with 829 self-identified Black adults in the United States, using a between-subject design to test 4 literature-derived motivational themes in clinical trial recruitment advertisements: (1) contribution to science, (2) helping the community, (3) lowering blood pressure, and (4) access to perks (US $500 worth of groceries or an equivalent cash amount). We assessed advertisement appeal, willingness to participate, and willingness to recommend clinical trial participation to others using cumulative link mixed models.

Results: Demographic factors played a more significant role than motivational themes in predicting advertisement effectiveness. Adults aged 40 - 59 years and individuals diagnosed with high blood pressure were more likely to find the advertisements appealing and express willingness to participate. Urban residents engaged more with the advertisements compared to those in suburban or rural areas. Participants with liberal (odds ratio [OR] 1.37, 95% CI 1.01 - 1.85; *P*=.04) and conservative (OR 1.62, 95% CI 1.09 - 2.40; *P*=.02) political views were more willing to participate in the clinical trial compared to those with moderate views. However, the “lowering my blood pressure” theme was less effective among individuals who distrusted health care institutions (OR 0.40, 95% CI 0.16 - 0.97; *P*=.04) and also reduced willingness to recommend the trial (OR 0.36, 95% CI 0.15 - 0.85; *P*=.02). In addition, higher trust levels were unexpectedly associated with lower willingness to participate when exposed to this theme (OR 0.41, 95% CI 0.17 - 0.98; *P*=.04).

Conclusions: Demographic targeting (age, health status, and geographic location) is more critical than generic motivational messaging in recruiting Black adults to clinical trials. Successful digital health recruitment requires targeting specific demographic segments with tailored messages, as effectiveness varies significantly across sociodemographic factors. The online survey methodology offers researchers a rapid, scalable tool for pretesting recruitment strategies, though it should complement rather than replace community engagement. These insights can help reduce cardiovascular health disparities by improving clinical trial representativeness.

(J Med Internet Res 2026;28:e75857) doi:10.2196/75857

KEYWORDS

clinical trials; recruitment; African American; digital advertising; health disparities; cardiovascular health

XSL•FO
RenderX

Introduction

Background

In the United States, Black adults face poorer cardiovascular health outcomes compared to other groups [1] and are underrepresented in clinical trials [2] that may benefit them. While some large cardiovascular trials have successfully recruited adequate numbers of Black participants through intensive community engagement [3,4], persistent underrepresentation remains a widespread challenge across most trials [2,5,6], particularly outside major academic medical centers. The drivers of the underrepresentation of Black adults in cardiovascular clinical trials are diverse. While the Tuskegee Syphilis Study [7] is often cited as a contributor to medical mistrust among Black adults, research suggests that knowledge of Tuskegee does not universally deter participation and that its impact varies considerably [8,9]. Contemporary factors, such as ongoing experiences with discriminatory practices and physician bias based on race or ethnicity, mental health status, education, income, and sex within the US health care system, play more substantial roles in shaping trust and participation decisions [10,11]. In addition, practical obstacles to participation, such as a lack of information about clinical trials, health insurance, time, and financial constraints, are significant barriers to participation [12,13], contributing to the failure of clinical trials in the United States to achieve their demographic representation goals [2,5,6]. These contemporary, structural barriers often outweigh historical events in determining participation. Despite the federal government's efforts to ensure inclusion in clinical trials [14], underrepresentation persists, resulting in disparities in health outcomes for minority racial and ethnic groups. Targeted advertisements may improve the recruitment of US Black adults to trials, but it is unclear what advertisement content might be most beneficial.

To support the successful recruitment of Black adults into a community-based participatory trial focused on recruiting Black urban adults (Groceries for Black Residents of Boston to Stop Hypertension [GoFresh]), we conducted an online survey with 829 self-identified Black adults in the United States to evaluate the effectiveness of a flyer-style trial recruitment advertisement with 4 literature-derived motivational themes. The goal of this study is twofold (1) to identify the optimal recruitment motivational themes for recruiting Black adults into the GoFresh trials and (2) to examine other factors (eg, trust in health care systems and demographic characteristics) that may impact Black adults' intentions to participate in clinical trials. In addition, this study demonstrates how online survey platforms paired with digital advertisement testing can serve as a rapid, cost-effective tool for pretesting recruitment messages before large-scale deployment. This approach allows researchers to iterate on messaging strategies and identify demographic-specific preferences, complementing traditional community engagement approaches. This digital health approach is particularly relevant as clinical trials increasingly use digital channels (eg, social media and online advertisements) to reach diverse populations, yet often lack systematic evidence about which digital messages are most effective for specific audiences.

We seek to answer a key research question (RQ):

- RQ: What motivates Black adults to participate in clinical trials, and how do demographic factors, health status, and trust in health care influence their responses to different recruitment messages?

Based on prior literature, we developed the following hypotheses:

- H1: Different motivational themes in clinical trial advertisements result in different levels of appeal, willingness to participate, and willingness to recommend the trial among Black adults.
- H2: The effectiveness of motivational themes is influenced by an individual's trust levels in health care systems.
- H3: Demographic characteristics influence responses to clinical trial recruitment through both direct effects on participation interest and by moderating the effectiveness of different motivational themes.

These hypotheses are grounded in prior research demonstrating that (1) message framing affects health behavior [15,16], (2) trust in health care systems significantly influences Black adults' willingness to participate in research [11,17,18], and (3) demographic characteristics shape health priorities and information processing [19,20].

Previous Work (Literature Review)

Clinical trials are essential for discovering new prevention, diagnostic, or treatment approaches and improving population health. Without accounting for diversity, resulting interventions may prove ineffective for broader populations, undermining trust and potentially increasing health costs [21,22]. US clinical trials have consistently underrepresented minority groups despite government policies to improve recruitment [23]. Beyond investigator bias [24], possible reasons include lack of awareness, mistrust of the health care system due to experiences with discrimination or bias, and financial and insurance-related obstacles [12,13,24]. A fundamental barrier to Black adults' participation in clinical trials is that they are often never explicitly asked or invited to participate [25,26]. When Black adults are approached, they are frequently deemed ineligible due to comorbid conditions, which are prevalent in this population due to longstanding health disparities [27]. This structural exclusion, whether through lack of outreach or restrictive eligibility criteria, represents a critical gap that precedes any consideration of motivational messaging.

Current solutions, such as population-specific trials and community engagement, have increased participation from underrepresented groups [28]. However, these are costly in human resources, time-consuming, and risk selection bias, making them unsustainable long-term [28,29]. Most clinical trials rely on traditional recruitment methods such as recruitment by investigators, local doctors, patient databases, printed materials (eg, flyers and posters), and community organizations, with limited data on the performance of digital strategies (eg, websites, social media marketing, and online advertisements) [30]. As a result, clinical trial recruitment recommendations are primarily based on case studies and individual practitioners' experiences [30,31]. Despite existing research on recruitment

XSL•FO
RenderX

strategy effectiveness, there remains a significant gap in data-driven approaches to systematically understand what truly improves clinical trial recruitment.

Big data and web analytics have transformed advertising across industries [32-34], enabling data-driven decision-making. Initial evidence suggests that these approaches can boost participation among underrepresented groups [35,36], yet most US clinical trials fail to leverage these methods to enhance recruitment diversity [37]. Given this opportunity, clinical trial advertising can be tailored, using data-driven methods, to the values and motivations of the target population. Nevertheless, data-driven strategies carry inherent risks of bias when underlying data underrepresent or stereotypically portray minoritized groups [38,39].

The literature [12,19,20] highlights how specific factors in individuals' lives influence their decision to participate in clinical trials. Recognizing each community's unique values allows researchers to develop resonant recruitment strategies, improving enrollment and retention [40]. Integrating big data and web analytics into clinical trial recruitment may help identify patterns and preferences among diverse populations. By analyzing these data, researchers can determine which messages and platforms most effectively reach different demographics, leading to more efficient and inclusive recruitment efforts. This study uses a literature-derived approach to develop 4 motivational themes for the GoFresh trials [12,41-43]. These themes reflect both intrinsic motivations (personal values and community impact) [44-46] and extrinsic motivations (health outcomes and financial benefits) [47,48] documented in prior qualitative and quantitative studies of trial recruitment barriers and facilitators.

Study Aims

This study has 3 primary aims:

- Aim 1: Determine which motivational themes in clinical trial recruitment advertisements are most effective in encouraging Black adults to participate in a cardiovascular health trial, and identify how these preferences vary by demographic characteristics.
- Aim 2: Examine how trust in health care systems moderates Black adults' responses to different recruitment message themes.
- Aim 3: Demonstrate proof of concept for using an online survey methodology as a rapid, cost-effective tool for pretesting clinical trial recruitment messages before large-scale deployment.

Methods

GoFresh Trials

This study is a collaboration with the GoFresh trials. The GoFresh trials (ClinicalTrials.gov nos NCT05121337 [GoFresh] and NCT05393232 [GoFreshRx]) are investigating the impact of home-delivered healthy groceries on blood pressure in Black adults with untreated and treated high blood pressure. In this study, participants are randomly assigned to receive either 12 weeks of Dietary Approaches to Stop Hypertension–patterned [49] grocery delivery with dietitian assistance or self-directed shopping with a monthly stipend over 3 months. After the intervention period, both groups are observed for 9 more months to determine barriers and facilitators for maintaining the Dietary Approaches to Stop Hypertension diet. The ultimate goal of these trials is to improve the cardiovascular health of Black adults living in healthy food priority areas, with and without treated hypertension. Findings from the trials will provide practical insights into effective and scalable strategies to prevent hypertension among Black adults [50].

Survey Design

The 4 motivational themes were derived from prior research on factors influencing clinical trial participation [12,41-43]. These themes include contributing to science, helping the community, lowering my blood pressure, and accessing perks (Textbox 1). Flyer-style advertisements were developed for each of these motivational themes (Figures 1-4). The advertisement content and design were reviewed by a community advisory board at Morgan State University to ensure cultural appropriateness and relevance before being tested in the survey. These variations included the GoFresh logo but did not include real contact information for the trial.

Each participant was randomly assigned to 1 of the 4 advertisements and asked to answer the same set of survey questions, rating the level of advertisement appeal, their willingness to participate in the trial, and their willingness to recommend the trial to others. The survey also collected data on participants' trust in health care systems using a validated scale from Rose et al [51] and Armstrong et al [52], as well as their basic demographic characteristics (refer to Multimedia Appendix 1).

XSL•FO
RenderX

Textbox 1. List of independent variables.

Motivational themes:

- The 4 motivational themes in the trial recruitment advertisement:
- Contribution to science (science): motivating participants to support scientific research.
- Helping the community (community): encouraging participants to benefit their local community.
- Lowering my blood pressure (blood pressure): promoting personal health improvement.
- Access to perks (perks): highlighting tangible benefits for participation.

Trust levels in health care:

- Participants' trust levels calculated using the Healthcare System Distrust Scale (10-50) [32,33]:
- Distrust (10-30): lowest tertile of trust scores.
- Neutral (31-35): middle tertile of trust scores.
- Trust (36-50): high tertile of trust scores.

Demographic backgrounds:

- Potential confounding factors:
- Age group
- Political view
- Gender
- Location (urbanity)

Figure 1. Motivational theme of the advertisement variant: contribution to science.



XSL•FO
RenderX

Figure 2. Motivational theme of the advertisement variant: helping the community.



Figure 3. Motivational theme of the advertisement variant: lowering my blood pressure.



XSL•FO
RenderX

Figure 4. Motivational theme of the advertisement variant: access to perks.



Study Population

We used the Prolific platform (Prolific Academic Ltd) [53] for web-based participant recruitment. Prolific is an online research platform that connects researchers with high-quality, prescreened participants for surveys and experiments, ensuring ethical standards, fair compensation, and reliable data collection. The platform's prescreening filter ensured that only individuals who self-identified as Black or African American or mixed race were recruited. Multimedia Appendix 2 shows the inclusion and exclusion criteria used in the GoFresh trials. The survey was deployed on February 2, 2024, and was closed 3 days later. We targeted 1000 respondents. Excluding instances of obvious survey abuse (n=60), participants who provided complete responses were included in the analysis. We did not restrict our sample to individuals with diagnosed hypertension because examining how health status influences recruitment message responses was a key RQ. This also reflects real-world recruitment for cardiovascular trials targeting both prevention and treatment populations.

Variables and Measurements

Dependent Variables

We operationalized the effectiveness of the trial recruitment advertisement as three variables: (1) the appeal of the advertisement—the attractiveness of the advertisement, (2) willingness to participate—readiness to join the advertised clinical trial, and (3) willingness to recommend—likelihood of sharing the clinical trial with others.

The primary goal of advertising appeal is to effectively communicate with a target audience and motivate them to take action. It can be subjective depending on messaging, visuals, or media type (direct mail, social media, print, etc) that are attractive to one's cognition, emotions, or both, which are rooted in past and present experiences [54]. In clinical trials, an effective trial advertisement motivates (appeals to) the target audience to participate in or recommend the trial within their network.

Independent Variables

Participants were exposed to 1 of 4 motivational appeals in recruitment advertisements, each emphasizing a distinct reason for participation: contributing to scientific research (science), helping one's community (community), improving personal health outcomes, such as lowering blood pressure (blood pressure), or gaining access to tangible benefits (perks). Alongside motivation, trust in the health care system was considered a critical factor in recruitment engagement. Participants' trust levels were measured using the Healthcare System Distrust Scale [51,52] with scores ranging from 10 to 50. To ensure an even distribution, trust was categorized into 3 levels using tertile cutoffs calculated in RStudio software (Posit PBC) [55]: distrust (10-30), neutral (31-35), and trust (36-50).

These variables provide context for differences in responses to motivational themes and trust levels, helping to refine strategies for trial recruitment. Together, motivation, trust, and demographic characteristics shape participant engagement and the overall effectiveness of recruitment advertisements.

Statistical Analysis

Sociodemographics and model covariates were tabulated with descriptive statistics. A cumulative link mixed model, a type of multilevel ordered logistic regression model, was used in our analysis. This approach allows estimation of model effects on an ordinal dependent variable while accommodating for random effects (ie, repeated measures). Models were adjusted for demographic factors, including age, political views, gender, and urbanity (location).

Main Effect Predictors

We included predictors for advertisement themes, trust levels, and prior diagnosis of high blood pressure by a doctor (high

XSL•FO
RenderX

blood pressure status), as well as demographic variables such as age group, income, political view, and gender.

Effect Modifier

We used effect modifiers to examine the interactions among independent variables. We evaluated different interaction combinations among important independent variables to address our research hypothesis. After comparing the statistical significance of these interactions, we included interactions between advertisement themes and trust levels (ie, advertisement theme × trust level) to better elucidate how the effectiveness of each advertisement theme varied among people with different levels of trust in health care systems.

Predictor Baselines

We set the baseline value for the motivational theme as "contribution to science" since it is the fundamental goal of clinical trials. For other predictors, we selected the neutral or most common values as baselines. For example, for the trust in health care systems predictor, we selected the "Neutral" category as the baseline because it represented the most frequently chosen response within the trust variable. This ensures that the baseline reflects the predominant sentiment among participants, providing a solid reference point for analysis.

Data Cleaning

To ensure model convergence, we performed the following steps: (1) converted responses such as "not sure," "prefer not to answer," and uncommon responses (eg, gender; "binary") to the "not available (N/A)" placeholder value, which also made the data more uniform; and (2) combined categories within certain variables, such as trust in the health care system and political views, to simplify and standardize the data for modeling.

Final Models

We fit 18 ordered logistic regression models, 6 for each effectiveness metric: appeal, willingness to participate, and willingness to recommend. From these, we selected the best model for each metric (3 models in total) based on a balance between predictor complexity and performance, as measured by the Akaike information criterion score.

Results Reporting

For predictors with statistical significance (P<.05), we reported the odds ratios (ORs), calculated as the exponential of the model estimate (β), along with the 95% CI. We also calculated the marginal 95% CIs for these interactions to evaluate their statistical significance and range of effect.

Our statistical analyses were conducted using R (version 4.4.0; R Foundation for Statistical Computing) with the ordinal package [56]. No adjustments were made for multiple comparisons.

Ethical Considerations

This study was approved by the University of Vermont Institutional Review Board (STUDY00002558). All participants provided electronic informed consent before beginning the survey, after being informed about the study's purpose, procedures, risks, benefits, and their right to withdraw at any time. We anonymized participants' data before using the remaining 829 survey responses for the final analysis. All data were deidentified, and no personally identifiable information was collected or linked to survey responses. Data were stored securely on password-protected servers accessible only to the research team. Participants who provided complete responses were compensated US $1.60 each, corresponding to an average rate of US $9.58 per hour, through the Prolific platform [53].

Results

Analytical Sample

A total of 913 surveys were started, with 897 completed (did not complete: n=16). After removing entries with incomplete fields (n=44), 853 valid responses remained. As shown in Multimedia Appendix 2, we further excluded 24 participants for the following reasons: not identifying as Black, self-reporting a lack of understanding of clinical trials despite being provided with a definition, providing unreasonable or nonrandomized answers, or being age outliers. This resulted in a final analytical sample of 829 responses.

Analytical Sample Sociodemographics

The sociodemographics of the analytical sample are provided in Multimedia Appendix 3. The 829 valid responses were from participants who self-identified as Black or African American on Prolific; 7 of these participants identified as mixed race. The mean age was 40 (SD 13.0) years. Educational levels varied widely, and most participants (474/829) reported annual incomes between US $35,000 and US $129,999. Participants reporting female and male genders were equally represented (female: 50.2%, 416/829, and male: 49.1%, 407/829). Most participants lived in urban or suburban areas in the United States (large city: 28.2%, 234/829; city: 33.7%, 279/829; and town or suburb: 32.2%, 267/829) and reported liberal (54.6%, 453/829), moderate (26.8%, 222/829), or conservative (18.6%, 154/829) views.

In terms of health status, 29.6% (245/829) reported a prior diagnosis of high blood pressure. The mean score on the Healthcare System Distrust Scale was 32.7 (SD 4.9; range 10-50; distrust [10-30], neutral [31-35], and trust [36-50]). Responses to individual items on the scale showed that 33.4% (277/829) of participants expressed distrust, while 30.3% (251/829) expressed trust in the health care system.

Appeal of Advertisement

As shown in Figure 5, all motivational themes had a high level of appeal, with 76% - 79% (155/204-165/209) of respondents somewhat or strongly agreeing. Notably, the "access to perks" theme was the most appealing, with 79% (165/209) agreeing. In contrast, the "helping the community" theme had the highest level of disagreement, with 15% (31/206) of respondents somewhat or strongly disagreeing.

XSL•FO
RenderX

Figure 5. Percentage of participant responses to the appeal of motivational themes in clinical trial advertisements.

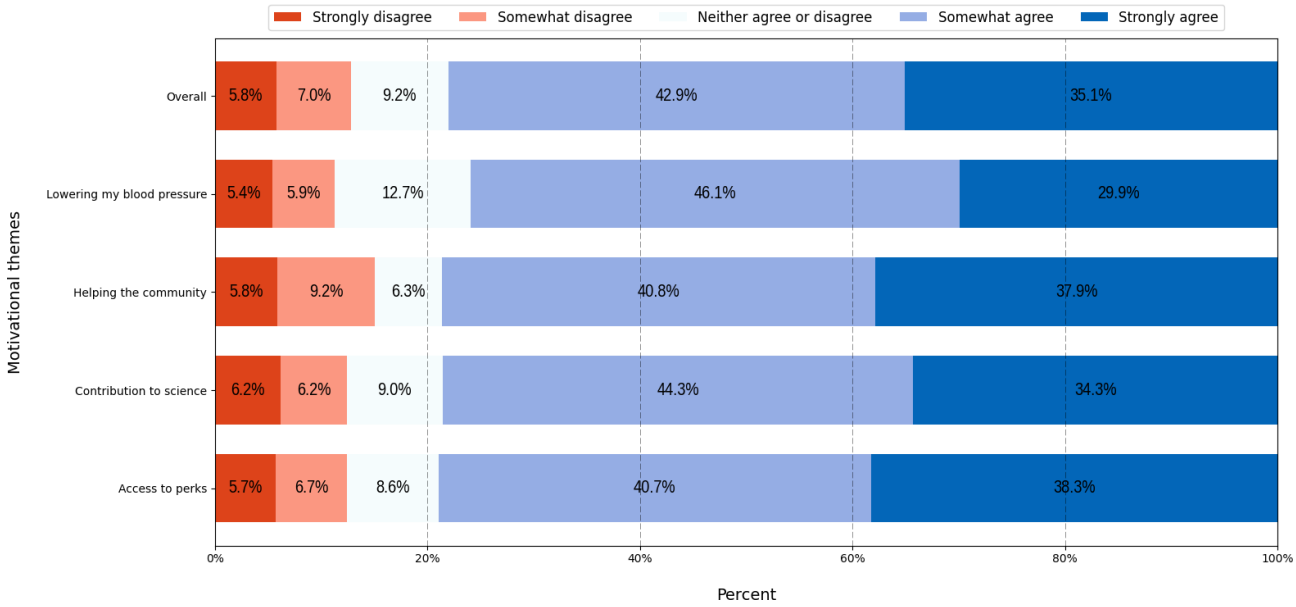

Willingness to Participate

As shown in Figure 6, scores for willingness to participate were lower than the appeal of the advertisement, with lower overall agreement to join clinical trials across all 4 motivational themes, ranging from 46.6% to 62.6% (95/204-131/209). The “access to perks” theme yielded the highest willingness to participate at 62.6% (131/209) and the lowest level of disagreement at 23% (48/209). Alternatively, the “lowering my blood pressure” theme had the lowest willingness at 46.6% and the highest level of disagreement, with 33.8% (69/204) strongly or somewhat disagreeing.

Figure 6. Percentage of participants’ willingness to participate in clinical trials by motivational theme.

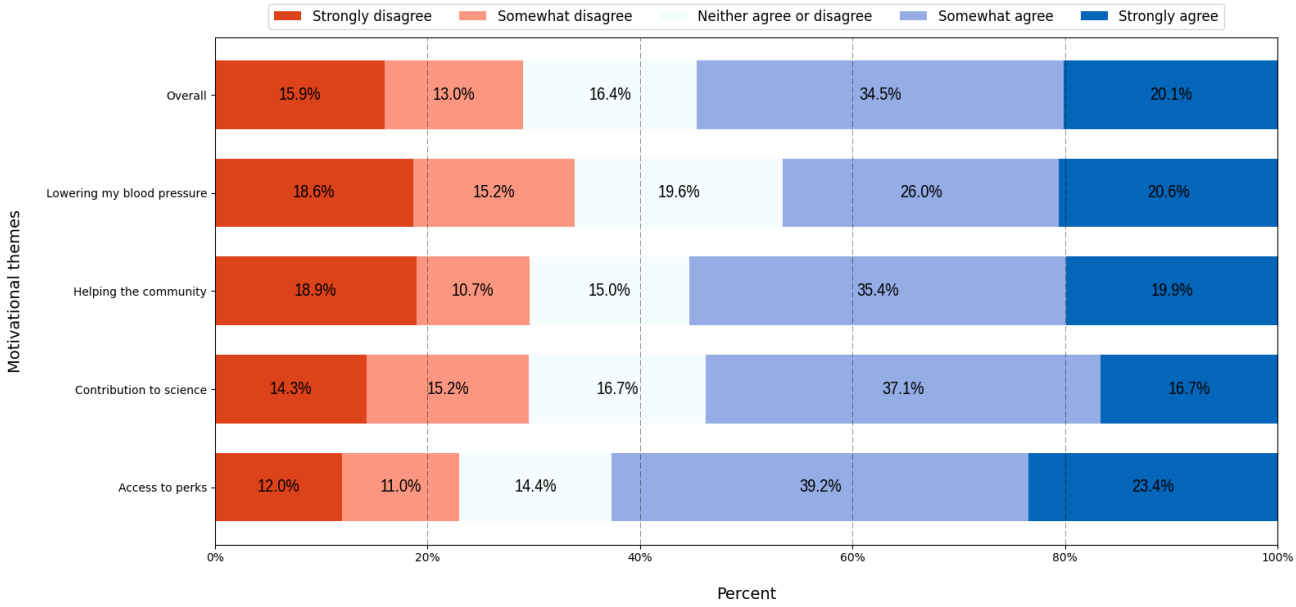

Willingness to Recommend

As shown in Figure 7, willingness to recommend was generally higher than willingness to participate for most themes, ranging from 58.3% to 68.4% (119/204-143/209) of participants somewhat or strongly agreeing. The highest willingness to recommend was for the “access to perks” theme at 68.4% (143/209). On the other hand, 25% (51/204) of respondents who viewed the “lowering my blood pressure” theme were unwilling (somewhat or strongly disagreed) to recommend the advertisement, the highest level of unwillingness, followed by the “contribution to science” theme at 21.9% (46/210).

XSL•FO
RenderX

Figure 7. Percentage of participants' likelihood to recommend clinical trial participation by motivational theme.

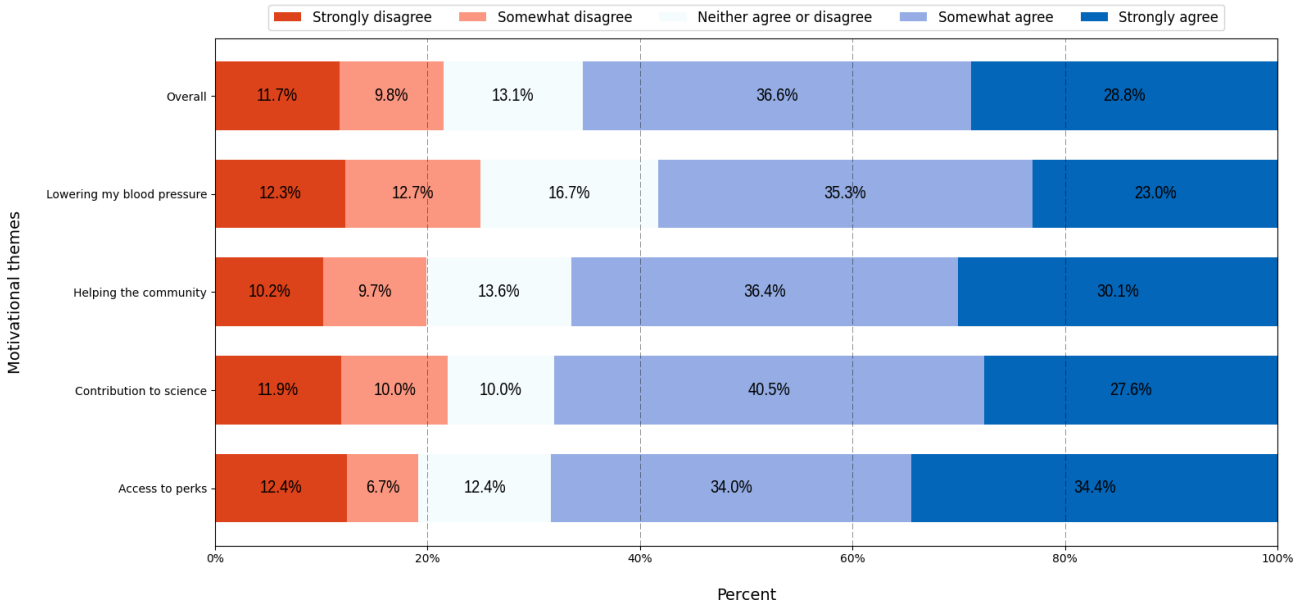

Inferential Statistics

Appeal of Advertisement

The model coefficients for the appeal of advertisements are provided in Table 1. Although all groups were compared, certain groups exhibited higher or lower odds of finding a theme appealing.

XSL•FO
RenderX

Table . Model coefficients for the appeal of motivational themes.

Predictors and effect modifiers	Estimate (β)	Value, OR[a] (95% CI)	*P* value
Motivational themes			
Research	0	1	—
Blood pressure	0.34	1.41 (0.78-2.54)	.25
Community	0.25	1.29 (0.7-2.37)	.41
Perks	0.05	1.05 (0.57-1.94)	.88
Trust levels			
Neutral	0	1	—
Trust	0.41	1.51 (0.80-2.84)	.21
Distrust	0.52	1.68 (0.90-3.14)	.11
High blood pressure diagnosis			
No	0	1	—
Yes	0.50	1.65 (1.20-2.25)	<.001
Age (years)			
18 - 29	0	1	—
30 - 39	0.09	1.09 (0.76-1.57)	.64
40 - 49	0.42	1.52 (1.01-2.29)	.04
50 - 59	0.45	1.57 (1.01-2.43)	.04
60 - 69	−0.10	0.91 (0.51-1.61)	.75
>70	0.10	1.1 (0.44-2.73)	.84
Political view			
Moderate	0	1	—
Liberal	0.45	1.56 (1.14-2.13)	.01
Conservative	0.30	1.35 (0.89-2.03)	.15
Sex			
Female	0	1	—
Male	−0.08	0.93 (0.71-1.21)	.58
Location			
Large city	0	1	—
City	−0.2	0.82 (0.58-1.15)	.24
Rural area	−0.2	0.82 (0.45-1.50)	.52
Town or suburb	−0.54	0.58 (0.41-0.83)	<.001
Motivational themes × trust levels			
Blood pressure × trust	−0.87	0.42 (0.17-1.03)	.06
Community × trust	−0.73	0.48 (0.19-1.21)	.12
Perks × trust	−0.15	0.86 (0.34-2.18)	.76
Blood pressure × distrust	−0.92	0.40 (0.16-0.97)	.04
Community × distrust	−0.15	0.86 (0.35-2.11)	.74
Perks × distrust	0.30	1.35 (0.56-3.30)	.51

[a]OR: odds ratio.

XSL•FO
RenderX

Main Effects

None of the motivational themes or trust levels had a significant effect on appeal. However, there were significant associations with appeal among certain groups. Participants who reported being diagnosed with high blood pressure had 65% higher odds of finding the themes appealing compared to those without such a diagnosis (OR 1.65, 95% CI 1.20 - 2.25; *P*=.002). People aged 40 - 49 years (OR 1.52, 95% CI 1.01 - 2.29; *P*=.04) and 50 - 59 years (OR 1.57, 95% CI 1.01 - 2.43; *P*=.04) had 52% and 57% higher odds, respectively, of finding the themes appealing compared to those aged 18 - 29 years. Participants with liberal political views had 56% higher odds of finding the themes appealing (OR 1.56, 95% CI 1.14 - 2.13; *P*=.005) compared to those with moderate views. Living in a town or suburb was associated with 42% lower odds of finding the themes appealing compared to those living in large cities (OR 0.58, 95% CI 0.41 - 0.83; *P*=.002).

Interaction Between Themes and Trust Levels

While the main effects of motivational themes and trust levels were not significantly associated with appeal, their interaction was significant. There was a significant interaction between the "lowering my blood pressure" theme and the "distrust" level (OR 0.40, 95% CI 0.16 - 0.97; *P*=.04). This indicates that individuals who viewed the "lowering my blood pressure" theme and expressed distrust in health care systems were 60% less likely to find the clinical trial advertisement appealing. Although this interaction was statistically significant, the broad CI suggests uncertainty regarding the precise magnitude of this effect.

Willingness to Participate

The model coefficients for willingness to participate are provided in Table 2. Although all groups were compared, certain groups exhibited higher or lower odds of participating in a clinical trial.

XSL•FO
RenderX

Table . Model coefficients for willingness to participate in clinical trials.

Predictors and effect modifiers	Estimate (β)	Value, OR[a] (95% CI)	*P* value
Motivational themes			
Research	0	1	—
Blood pressure	0.32	1.38 (0.78 - 2.44)	.27
Community	0.16	1.17 (0.65 - 2.10)	.60
Perks	0.29	1.34 (0.75 - 2.39)	.31
Trust levels			
Neutral	0	1	—
Trust	0.51	1.66 (0.92 - 3.00)	.09
Distrust	0.51	1.66 (0.91 - 3.04)	.10
High blood pressure diagnosis			
No	0	1	—
Yes	0.63	1.88 (1.39 - 2.53)	<.001
Age (years)			
18 - 29	0	1	—
30 - 39	0.11	1.11 (0.78 - 1.59)	.56
40 - 49	0.38	1.46 (0.99 - 2.16)	.05
50 - 59	0.55	1.74 (1.14 - 2.65)	.01
60 - 69	−0.14	0.87 (0.50 - 1.50)	.61
>70	−0.03	0.97 (0.42 - 2.24)	.95
Political view			
Moderate	0	1	—
Liberal	0.31	1.37 (1.01 - 1.85)	.04
Conservative	0.48	1.62 (1.09 - 2.40)	.02
Sex			
Female	0	1	—
Male	0.19	1.21 (0.94 - 1.57)	.14
Location			
Large city	0	1	—
City	−0.27	0.77 (0.55 - 1.07)	.12
Rural area	0.11	1.11 (0.62 - 1.99)	.72
Town or suburb	−0.56	0.57 (0.41 - 0.80)	<.001
Motivational themes × trust levels			
Blood pressure × trust	−0.88	0.41 (0.17 - 0.98)	.04
Community × trust	−0.46	0.63 (0.26 - 1.50)	.30
Perks × trust	−0.13	0.88 (0.37 - 2.09)	.77
Blood pressure × distrust	−0.70	0.5 (0.21 - 1.17)	.11
Community × distrust	−0.25	0.78 (0.33 - 1.85)	.57
Perks × distrust	0.31	1.37 (0.58 - 3.20)	.47

[a]OR: odds ratio.

XSL•FO
RenderX

Main Effects

The motivational themes and trust levels did not have a significant effect on willingness to participate. There were significant demographic associations. Participants who reported being diagnosed with high blood pressure (OR 1.88, 95% CI 1.39 - 2.53; *P*<.001) had 88% higher odds of being willing to participate in the clinical trial focused on high blood pressure compared to those who had not been diagnosed. Individuals aged 50 - 59 years had 74% higher odds of willingness to participate in the clinical trial compared to those aged 18 - 29 years (OR 1.74, 95% CI: 1.14 - 2.65, *P*=.01). Participants with both liberal (OR 1.37, 95% CI 1.01 - 1.85; *P*=.04) and conservative (OR 1.62, 95% CI 1.09 - 2.40; *P*=.02) views had 37% and 62% higher odds, respectively, of being willing to participate in the clinical trial focused on high blood pressure compared to those with moderate views. Living in a town or suburb was associated with 43% lower odds of being willing to participate in clinical trials compared to those living in large cities (OR 0.57, 95% CI 0.41 - 0.80; *P*=.001).

Interaction Between Themes and Trust Levels

The interaction between the "lowering my blood pressure" motivational theme and trust level was found to be significant (OR 0.41, 95% CI 0.17 - 0.98; *P*=.04), indicating that individuals who viewed the "lowering my blood pressure" theme and had higher levels of trust were 59% less willing to participate in the clinical trial. Although the interaction was statistically significant, the wide CI indicates uncertainty about the exact size of this effect.

Willingness to Recommend

The model coefficients for willingness to recommend are provided in Table 3. All groups were compared to each other, but certain groups exhibited higher or lower odds of recommending a clinical trial to others.

XSL•FO
RenderX

Table . Model coefficients for willingness to recommend clinical trials.

Predictors and effect modifiers	Estimate (β)	Value, OR[a] (95% CI)	*P* value
Motivational themes			
Research	0	1	—
Blood pressure	0.19	1.21 (0.68 - 2.16)	.51
Community	0.18	1.2 (0.66 - 2.17)	.56
Perks	0.03	1.03 (0.57 - 1.86)	.93
Trust levels			
Neutral	0	1	—
Trust	0.28	1.32 (0.72 - 2.42)	.37
Distrust	0.50	1.65 (0.89 - 3.06)	.11
High blood pressure diagnosis			
No	0	1	—
Yes	0.31	1.37 (1.01 - 1.85)	.04
Age (years)			
18 - 29	0	1	—
30 - 39	−0.09	0.92 (0.64 - 1.32)	.64
40 - 49	0.27	1.31 (0.88 - 1.95)	.19
50 - 59	0.28	1.32 (0.87 - 2.01)	.19
60 - 69	−0.21	0.81 (0.47 - 1.40)	.45
>70	0.44	1.55 (0.64 - 3.72)	.33
Political view			
Moderate	0	1	—
Liberal	0.3	1.35 (1.00 - 1.83)	.05
Conservative	0.38	1.46 (0.99 - 2.16)	.06
Sex			
Female	0	1	—
Male	0.08	1.08 (0.83 - 1.41)	.54
Location			
Large city	0	1	—
City	−0.34	0.71 (0.51 - 1.00)	.05
Rural area	−0.13	0.88 (0.5 - 1.57)	.67
Town or suburb	−0.48	0.62 (0.44 - 0.88)	.01
Motivational themes × trust levels			
Blood pressure × trust	−0.56	0.57 (0.24 - 1.36)	.20
Community × trust	−0.6	0.55 (0.23 - 1.33)	.18
Perks × trust	0.3	1.35 (0.55 - 3.30)	.52
Blood pressure × distrust	−1.03	0.36 (0.15 - 0.85)	.02
Community × distrust	−0.21	0.81 (0.34 - 1.95)	.64
Perks × distrust	0.09	1.09 (0.46 - 2.61)	.84

[a]OR: odds ratio.

XSL•FO
RenderX

Main Effects

Motivational themes and trust levels did not have a significant effect on willingness to recommend the advertised clinical trial, although there were other demographic associations. A diagnosis of high blood pressure was significantly associated with 37% higher odds of recommending the trial, regardless of the themes (OR 1.37, 95% CI 1.01 - 1.85; *P*=.04). Living in a city (OR 0.71, 95% CI 0.51 - 1.00; *P*=.05) and in a town or suburban areas (OR 0.62, 95% CI 0.44 - 0.88; *P*=.006) was associated with 29% and 38% lower odds of recommending clinical trial advertisements, respectively, compared to living in large cities. While the result for individuals living in a city is statistically significant (*P*=.05), the upper bound of the CI is 1, suggesting the effect could range from a decreased willingness to recommend to no effect.

Interaction Between Themes and Trust Levels

The interaction between the "lowering my blood pressure" theme and the "distrust" level was statistically significant (OR 0.36, 95% CI 0.15 - 0.85; *P*=.02). This suggests that individuals who viewed the "lowering my blood pressure" theme and expressed distrust in health care systems were 64% less likely to recommend the clinical trial.

Discussion

While trials such as the Jackson Heart Study [3] and REasons for Geographic And Racial Differences in Stroke study [4] achieved strong Black representation through intensive community engagement, the majority of clinical trials continue to struggle with recruitment diversity [12,13,24]. Our study builds on these successes by providing a scalable methodology for optimizing recruitment messaging across different trial contexts and resource levels.

Principal Findings

This study makes 3 key contributions to clinical trial recruitment science. First, demographic factors (age, health status, and geographic location) are stronger predictors of Black adults' interest in cardiovascular trials than motivational message themes themselves, suggesting demographic segmentation should be prioritized in recruitment planning. Second, we demonstrate complex trust dynamics, whereby the "lowering blood pressure" theme (personal health benefit messaging) was less effective among both low-trust and high-trust individuals. Third, we provide proof of concept for using online surveys as a rapid, cost-effective tool for pretesting recruitment messages, offering a scalable methodology that complements community engagement approaches.

Targeting individuals based solely on racial identity overlooks diversity within racial groups, as factors such as socioeconomic status, health conditions, geographic location, age, and political beliefs shape perspectives and motivations to enroll in clinical trials [19,57,58]. Even within the same demographic group, individuals have different needs and motivations. Our research identifies key considerations impacting how Black adults engage with digital clinical trial advertisements: individual-level demographic influences on advertisement effectiveness and trust and information dynamics. Considering these factors can improve design and marketing strategies, increasing the likelihood that recruitment advertisements will resonate with potential participants and enhance the effectiveness of clinical trial recruitment.

Demographic Influences on Advertisement Effectiveness

Demographic factors such as age, health status, political views, and location significantly influence Black adults' responses to clinical trial advertisements [19,58,59]. Our findings show hypertension-related trial advertisements resonate most with adults aged 40 - 59 years and individuals diagnosed with high blood pressure, suggesting that targeted approaches based on age and relevant health conditions are effective [59]. These individuals, facing age-related health challenges, may actively seek new treatments, increasing their likelihood of participation [44].

While clinical trial advertisements appealed slightly more to Black adults with liberal views, both liberals and conservatives showed willingness to participate. This adds context to how political ideology affects participation, as prior studies [60-62] often overlooked practical considerations driving engagement in underrepresented populations. Our study demonstrates that Black adults' participation decisions are primarily influenced by pragmatic factors such as incentives, safety, trust, access to treatments, and health assurances, rather than political ideology. This suggests prioritizing tangible benefits during recruitment alongside clear, transparent informed consent.

Black adults living in suburban and rural areas were less likely to engage with clinical trial advertisements compared to those in cities, aligning with literature showing that urban areas have more trial sites and health care facilities. Rural and suburban communities remain underrepresented, have limited trial awareness, face logistical challenges such as transportation barriers, and may harbor mistrust [63-66]. Outreach strategies in these areas should address specific access- and trust-related challenges.

Trust and Information Dynamics

For clinical trial recruitment, trust fundamentally shapes how potential participants perceive trial motives and their willingness to participate. This manifests through information credibility, transparency regarding risks and benefits, and perceived trustworthiness of sponsoring organizations [67-69]. Our study revealed that trust in the health care system interacted significantly with the "lowering my blood pressure" theme, indicating that for Black adults, trust remains a critical issue influenced by varying motivations and socioeconomic factors, compounded by historical contexts and personal health care experiences [7]. Given these findings, recruitment strategies should prioritize messaging that builds trust with this population.

Motivational Themes

Black adults indicated reluctance to participate in clinical trials even when they found certain themes appealing. This is reflected in the overall lower levels of agreement in willingness to participate (28.9%) and willingness to recommend (21.5%) compared to the appeal of the themes (12.8%). These results

XSL•FO
RenderX

also imply that additional motivators, such as tangible perks, may be necessary to boost participation, as personal health benefit themes such as "lowering my blood pressure" seem less motivating compared to "access to perks," which appears to be a stronger draw than health benefits alone. Participants also showed a higher willingness to recommend clinical trials with a promise of perks than to participate themselves. This also indicates that participants are more comfortable endorsing or suggesting trials that offer monetary incentives compared to "lowering my blood pressure," indicating that personal health benefits may not be a popular reason for recommending clinical trials.

Notably, the low willingness to participate or recommend the "lowering my blood pressure" theme may be partly due to the fact that 68.4% (567/829) of participants reported never having been diagnosed with high blood pressure, compared to 29.6% (245/829) who reported having been diagnosed.

We recently published the performance of these themes on a GoFresh clinical trial recruitment website [70]. We found that "lowering my blood pressure," "access to perks," and "helping my community" messages had a similar proportion of website visitors who completed the signup form on the GoFresh website. These 3 themes all outperformed the "contribution to science" theme. It is unclear how these themes relate to more relevant outcomes, including clinical trial enrollment and completion. Motivational theme preferences identified in this study might also reflect actual trial enrollment and completion rates. Work to determine these outcomes by theme is ongoing.

Practical Implications for Researchers

Our findings offer actionable recommendations for clinical trial recruitment:

- Prioritize demographic segmentation: demographic factors (age, health status, and location) are stronger predictors than message themes alone. However, certain themes (eg, "access to perks") may resonate particularly well with minority populations and should be strategically incorporated when targeting these groups.
- Use online surveys for rapid message testing: this methodology enables cost-effective, rapid pretesting of recruitment messages before investing in full-scale campaigns, allowing iterative refinement based on empirical data.
- Navigate trust dynamics carefully: the relationship between message type and trust varies by community. Among minoritized populations, practical considerations may outweigh health-benefit appeals, particularly when institutional trust is low.
- Combine quantitative and community-engaged approaches: we developed messages with community advisory board input, then quantitatively tested them. This combined approach ensures both cultural relevance and demographic insights.
- Adapt to your specific context: use this methodology as a framework for testing message effectiveness within your own target population, as recruitment patterns are likely context-dependent.

While our study focuses on optimizing motivational messaging, we acknowledge that the most effective recruitment strategy begins with actively and explicitly inviting Black adults to participate. Recruitment messaging, regardless of how well-optimized, cannot address the fundamental barrier of never being asked. Our findings complement, rather than replace, the critical need for proactive outreach and inclusive eligibility criteria that account for the higher prevalence of comorbidities in Black populations.

Limitations and Future Work

We recognize certain constraints in our study, which we discuss to provide context and ensure a clear interpretation of our results.

This study relied on self-reported intentions rather than actual trial participation behavior. Future research incorporating observed behavioral outcomes could enhance generalizability. Our use of an online panel (Prolific) may limit generalizability to Black adults who lack internet access, are less digitally engaged, or do not participate in online research. The study focused specifically on dietary intervention for blood pressure, and findings may not generalize to other clinical trial types. Future research should explore a broader range of trials involving different medical conditions, interventions, and populations to strengthen applicability. Future work can compare how different audiences respond to messages directly addressing concerns vs neutral messaging to improve communication strategies. We attempted to stratify data by hypertension status, but the skewed distribution (only 245/829, 29.6% observations for the hypertensive group) led to statistical model instability due to sparse data, causing Hessian errors. Simplifying the model would not fully address our RQs, as our goal was to analyze a sample representative of Black adults who might see these advertisements. Future research focused specifically on the hypertensive demographic could examine this group in greater detail, providing more reliable estimates and statistical stability.

A significant limitation is the lack of message variance. We tested 1 message per theme rather than multiple variants. We cannot definitively determine whether effects are attributable to themes or to specific message characteristics. Future research should test multiple message variants per theme to isolate theme-level from message-specific effects [71]. Our quantitative approach risks oversimplifying complex cultural factors that qualitative methods might better capture. Our findings should inform or enhance community-engaged recruitment and not replace it. This methodology served our proof-of-concept goal of rapidly testing recruitment messages before deployment.

While these limitations should be considered, they do not undermine the research validity. The study provides valuable insights that future research can build upon to further enhance the applicability and impact of these findings.

Conclusion

This study contributes new empirical evidence to the science of inclusive trial recruitment and provides an evidence-based methodology to identify effective advertisements for improving participant diversity. It emphasizes the need for thoughtful design of digital clinical trial advertisements targeting Black

XSL•FO
RenderX

adults in the United States. Simply targeting individuals based on racial identity ignores diverse factors such as age, socioeconomic status, location, health conditions, and political beliefs that shape motivations. Effective advertisements must address these differences, offering broadly appealing messages that also meet specific needs. Recruitment messages should be diverse and tailored to address multiple motivations rather than seeking a single, universal driving factor. Addressing trust and cultural sensitivity in recruitment strategies can boost participation, build stronger relationships between health care institutions and communities, and lead to more equitable health research outcomes.

Acknowledgments

The authors are grateful to Maria Sckolnick at the University of Vermont for her guidance. Generative artificial intelligence (AI) tools (such as Claude Opus-4.1 and ChatGPT-3.5) were used only to refine language clarity and grammar in portions of the manuscript. All scientific content, study design, data analysis, interpretation of results, and conclusions were developed by the authors without AI assistance.

Funding

This research was funded by the American Heart Association (award 953782) as part of the Improving Participation Among Diverse Populations in Cardiovascular Clinical Trials (IMPACT) network's Digital Advertisements for Inclusive Trial Accrual (DAITA) project, within the Science of Diversity in Clinical Trials Strategically-Focused Research Network.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

TH contributed to methodology, formal analysis, investigation, data curation, writing the original draft, writing–review and editing, and visualization. JC, ZH, and RANTO contributed to writing–review and editing. SPJ contributed to conceptualization, writing–review and editing, supervision, and funding acquisition. TPL contributed to conceptualization and writing–review and editing. TBP contributed to conceptualization, writing–review and editing, supervision, and funding acquisition. YF contributed to conceptualization, writing–review and editing, supervision, funding acquisition, and project administration.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey design.
[DOCX File, 12 KB - jmir_v28i1e75857_app1.docx]

Multimedia Appendix 2
Flowchart of participant inclusion and exclusion criteria for this study.
[PNG File, 96 KB - jmir_v28i1e75857_app2.png]

Multimedia Appendix 3
Sociodemographic characteristics of the analytical sample and distribution of motivational themes in clinical trial advertisements by demographic factors.
[DOCX File, 11 KB - jmir_v28i1e75857_app3.docx]

References

1. Fryar CD, Ostchega Y, Hales CM, Zhang G, Kruszon-Moran D. Hypertension prevalence and control among adults: United States, 2015-2016. NCHS Data Brief 2017 Oct(289):1-8. [Medline: 29155682]
2. Turner BE, Steinberg JR, Weeks BT, Rodriguez F, Cullen MR. Race/ethnicity reporting and representation in US clinical trials: a cohort study. Lancet Reg Health Am 2022 Jul;11:100252. [doi: 10.1016/j.lana.2022.100252] [Medline: 35875251]
3. Jackson Heart Study. University of Mississippi Medical Center. URL: https://umc.edu/Healthcare/Heart/jackson-heart-study.html [accessed 2025-11-18]
4. Howard VJ, Cushman M, Pulley L, et al. The reasons for geographic and racial differences in stroke study: objectives and design. Neuroepidemiology 2005;25(3):135-143. [doi: 10.1159/000086678] [Medline: 15990444]

XSL•FO
RenderX

5. Steinberg JR, Turner BE, DiTosto JD, et al. Race and ethnicity reporting and representation in obstetrics and gynecology clinical trials and publications from 2007-2020. JAMA Surg 2023 Feb 1;158(2):181-190. [doi: 10.1001/jamasurg.2022.6600] [Medline: 36542396]
6. Loree JM, Anand S, Dasari A, et al. Disparity of race reporting and representation in clinical trials leading to cancer drug approvals from 2008 to 2018. JAMA Oncol 2019 Oct 1;5(10):e191870. [doi: 10.1001/jamaoncol.2019.1870] [Medline: 31415071]
7. Scharff DP, Mathews KJ, Jackson P, Hoffsuemmer J, Martin E, Edwards D. More than Tuskegee: understanding mistrust about research participation. J Health Care Poor Underserved 2010 Aug;21(3):879-897. [doi: 10.1353/hpu.0.0323] [Medline: 20693733]
8. Brandon DT, Isaac LA, LaVeist TA. The legacy of Tuskegee and trust in medical care: is Tuskegee responsible for race differences in mistrust of medical care? J Natl Med Assoc 2005 Jul;97(7):951-956. [Medline: 16080664]
9. Katz RV, Kegeles SS, Kressin NR, et al. The Tuskegee Legacy Project: willingness of minorities to participate in biomedical research. J Health Care Poor Underserved 2006 Nov;17(4):698-715. [doi: 10.1353/hpu.2006.0126] [Medline: 17242525]
10. Nong P, Raj M, Creary M, Kardia SLR, Platt JE. Patient-Reported Experiences of Discrimination in the US Health Care System. JAMA Netw Open 2020 Dec 1;3(12):e2029650. [doi: 10.1001/jamanetworkopen.2020.29650] [Medline: 33320264]
11. Armstrong K, Putt M, Halbert CH, et al. Prior experiences of racial discrimination and racial differences in health care system distrust. Med Care 2013 Feb;51(2):144-150. [doi: 10.1097/MLR.0b013e31827310a1] [Medline: 23222499]
12. Ketchum FB, Erickson CM, Chin NA, et al. What influences the willingness of Blacks and African Americans to enroll in preclinical Alzheimer's disease biomarker research? A Qualitative vignette analysis. J Alzheimers Dis 2022;87(3):1167-1179. [doi: 10.3233/JAD-215521] [Medline: 35466937]
13. George S, Duran N, Norris K. A systematic review of barriers and facilitators to minority research participation among African Americans, Latinos, Asian Americans, and Pacific Islanders. Am J Public Health 2014 Feb;104(2):e16-e31. [doi: 10.2105/AJPH.2013.301706] [Medline: 24328648]
14. NOT-OD-18-014: Revision: NIH Policy and Guidelines on the Inclusion of Women and Minorities as Subjects in Clinical Research. National Institutes of Health. URL: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-18-014.html [accessed 2024-04-08]
15. Gallagher KM, Updegraff JA. Health message framing effects on attitudes, intentions, and behavior: a meta-analytic review. Ann Behav Med 2012 Feb;43(1):101-116. [doi: 10.1007/s12160-011-9308-7] [Medline: 21993844]
16. Rothman AJ, Salovey P, Antone C, Keough K, Martin CD. The influence of message framing on intentions to perform health behaviors. J Exp Soc Psychol 1993 Sep;29(5):408-433. [doi: 10.1006/jesp.1993.1019]
17. Thompson HS, Manning M, Mitchell J, et al. Factors associated with racial/ethnic group-based medical mistrust and perspectives on COVID-19 vaccine trial participation and vaccine uptake in the US. JAMA Netw Open 2021 May 3;4(5):e2111629. [doi: 10.1001/jamanetworkopen.2021.11629] [Medline: 34042990]
18. Taylor LA, Nong P, Platt J. Fifty years of trust research in health care: a synthetic review. Milbank Q 2023 Mar;101(1):126-178. [doi: 10.1111/1468-0009.12598] [Medline: 36689251]
19. Kim JY, Florez M, Botto E, Belgrave X, Grace C, Getz K. The influence of socioeconomic status on individual attitudes and experience with clinical trials. Commun Med (Lond) 2024 Sep 5;4(1):172. [doi: 10.1038/s43856-024-00586-9] [Medline: 39237734]
20. Wilkinson AV, Vasudevan V, Honn SE, Spitz MR, Chamberlain RM. Sociodemographic characteristics, health beliefs, and the accuracy of cancer knowledge. J Cancer Educ 2009;24(1):58-64. [doi: 10.1080/08858190802664834] [Medline: 19259867]
21. Woods-Burnham L, Johnson JR, Hooker SE, Bedell FW, Dorff TB, Kittles RA. The role of diverse populations in US clinical trials. Med 2021 Jan 15;2(1):21-24. [doi: 10.1016/j.medj.2020.12.009] [Medline: 35590131]
22. Why diverse representation in clinical research matters and the current state of representation within the clinical research ecosystem. In: Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups: National Academies Press; 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK584396/ [accessed 2026-03-05]
23. Branson RD, Davis K, Butler KL. African Americans' participation in clinical research: importance, barriers, and solutions. Am J Surg 2007 Jan;193(1):32-39. [doi: 10.1016/j.amjsurg.2005.11.007] [Medline: 17188084]
24. Niranjan SJ, Wenzel JA, Martin MY, et al. Perceived institutional barriers among clinical and research professionals: minority participation in oncology clinical trials. JCO Oncol Pract 2021 May;17(5):e666-e675. [doi: 10.1200/OP.20.00970] [Medline: 33974821]
25. Barriers to representation of underrepresented and excluded populations in clinical research. In: Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups: National Academies Press; 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK584407/ [accessed 2026-03-05]
26. Garber M, Hanusa BH, Switzer GE, Mellors J, Arnold RM. HIV-infected African Americans are willing to participate in HIV treatment trials. J Gen Intern Med 2007 Jan;22(1):17-42. [doi: 10.1007/s11606-007-0121-8] [Medline: 17351837]
27. Smedley BD, Stith AY, Nelson AR, editors. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care: National Academies Press; 2003. URL: https://www.ncbi.nlm.nih.gov/books/NBK220358/ [accessed 2026-03-05]

XSL•FO
RenderX

28. Taani MH, Zabler B, Fendrich M, Schiffman R. Lessons learned for recruitment and retention of low-income African Americans. Contemp Clin Trials Commun 2020 Mar;17:100533. [doi: 10.1016/j.conctc.2020.100533] [Medline: 32211558]
29. Hamel LM, Penner LA, Albrecht TL, Heath E, Gwede CK, Eggly S. Barriers to clinical trial enrollment in racial and ethnic minority patients with cancer. Cancer Control 2016 Oct;23(4):327-337. [doi: 10.1177/107327481602300404] [Medline: 27842322]
30. Tanner A, Kim SH, Friedman DB, Foster C, Bergeron CD. Promoting clinical research to medically underserved communities: current practices and perceptions about clinical trial recruiting strategies. Contemp Clin Trials 2015 Mar;41:39-44. [doi: 10.1016/j.cct.2014.12.010] [Medline: 25542611]
31. Andrasik MP, Broder GB, Wallace SE, et al. Increasing Black, Indigenous and People of Color participation in clinical trials through community engagement and recruitment goal establishment. PLoS One 2021;16(10):e0258858. [doi: 10.1371/journal.pone.0258858] [Medline: 34665829]
32. Crowe S, Cresswell K, Robertson A, Huby G, Avery A, Sheikh A. The case study approach. BMC Med Res Methodol 2011 Jun 27;11:100. [doi: 10.1186/1471-2288-11-100] [Medline: 21707982]
33. Hung JL, He W, Shen J. Big data analytics for supply chain relationship in banking. Industrial Marketing Management 2020 Apr;86:144-153. [doi: 10.1016/j.indmarman.2019.11.001]
34. Guo C, Chen J. Big data analytics in healthcare. In: Nakamori Y, editor. Knowledge Technology and Systems: Springer Nature; 2023:27-70. [doi: 10.1007/978-981-99-1075-5_2]
35. Berman R, Israeli A. The value of descriptive analytics: evidence from online retailers. Marketing Science 2022 Nov;41(6):1074-1096. [doi: 10.1287/mksc.2022.1352]
36. Tsaltskan V, Nguyen K, Eaglin C, Holers VM, Deane KD, Firestein GS. Comparison of web-based advertising and a social media platform as recruitment tools for underserved and hard-to-reach populations in rheumatology clinical research. ACR Open Rheumatol 2022 Jul;4(7):623-630. [doi: 10.1002/acr2.11448] [Medline: 35536990]
37. Inan OT, Tenaerts P, Prindiville SA, et al. Digitizing clinical trials. NPJ Digit Med 2020;3(1):101. [doi: 10.1038/s41746-020-0302-y] [Medline: 32821856]
38. Bærøe K, Gundersen T, Henden E, Rommetveit K. Can medical algorithms be fair? Three ethical quandaries and one dilemma. BMJ Health Care Inform 2022 Apr;29(1):e100445. [doi: 10.1136/bmjhci-2021-100445] [Medline: 35396245]
39. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
40. Anastasi JK, Capili B, Norton M, McMahon DJ, Marder K. Recruitment and retention of clinical trial participants: understanding motivations of patients with chronic pain and other populations. Front Pain Res (Lausanne) 2023;4:1330937. [doi: 10.3389/fpain.2023.1330937] [Medline: 38606348]
41. Williams CP, Senft Everson N, Shelburne N, Norton WE. Demographic and health behavior factors associated with clinical trial invitation and participation in the United States. JAMA Netw Open 2021 Sep 1;4(9):e2127792. [doi: 10.1001/jamanetworkopen.2021.27792] [Medline: 34586365]
42. Le D, Ozbeki H, Salazar S, Berl M, Turner MM, Price OA. Improving African American women's engagement in clinical research: a systematic review of barriers to participation in clinical trials. J Natl Med Assoc 2022 Jun;114(3):324-339. [doi: 10.1016/j.jnma.2022.02.004] [Medline: 35279325]
43. Cunningham-Erves J, Kusnoor SV, Villalta-Gil V, et al. Development and pilot implementation of guidelines for culturally tailored research recruitment materials for African Americans and Latinos. BMC Med Res Methodol 2022 Sep 24;22(1):248. [doi: 10.1186/s12874-022-01724-4] [Medline: 36153481]
44. Wendler D, Krohmal B, Emanuel EJ, Grady C, Group E. Why patients continue to participate in clinical research. Arch Intern Med 2008 Jun 23;168(12):1294-1299. [doi: 10.1001/archinte.168.12.1294] [Medline: 18574086]
45. Corbie-Smith G, Thomas SB, Williams MV, Moody-Ayers S. Attitudes and beliefs of African Americans toward participation in medical research. J Gen Intern Med 1999 Sep;14(9):537-546. [doi: 10.1046/j.1525-1497.1999.07048.x] [Medline: 10491242]
46. Carrera JS, Brown P, Brody JG, Morello-Frosch R. Research altruism as motivation for participation in community-centered environmental health research. Soc Sci Med 2018 Jan;196:175-181. [doi: 10.1016/j.socscimed.2017.11.028]
47. Largent EA, Eriksen W, Barg FK, Greysen SR, Halpern SD. Participants' perspectives on payment for research participation: A qualitative study. Ethics & Human Research 2022 Nov;44(6):14-22 [FREE Full text] [doi: 10.1002/eahr.500147]
48. Stunkel L, Grady C. More than the money: a review of the literature examining healthy volunteer motivations. Contemp Clin Trials 2011 May;32(3):342-352. [doi: 10.1016/j.cct.2010.12.003] [Medline: 21146635]
49. Siervo M, Lara J, Chowdhury S, Ashor A, Oggioni C, Mathers JC. Effects of the Dietary Approach to Stop Hypertension (DASH) diet on cardiovascular risk factors: a systematic review and meta-analysis. Br J Nutr 2015 Jan 14;113(1):1-15. [doi: 10.1017/S0007114514003341] [Medline: 25430608]
50. Turkson-Ocran RAN, Cluett JL, Fitzpatrick SL, et al. Rationale and design of the Groceries for Black Residents of Boston to Stop Hypertension among Adults without Treated Hypertension (GoFresh) Trial. Am J Hypertens 2023 Apr 15;36(5):256-263. [doi: 10.1093/ajh/hpad008] [Medline: 37061794]
51. Rose A, Peters N, Shea JA, Armstrong K. Development and testing of the health care system distrust scale. J Gen Intern Med 2004 Jan;19(1):57-63. [doi: 10.1111/j.1525-1497.2004.21146.x] [Medline: 14748861]

XSL•FO
RenderX

52. Armstrong K, Rose A, Peters N, Long JA, McMurphy S, Shea JA. Distrust of the health care system and self-reported health in the United States. J Gen Intern Med 2006 Apr;21(4):292-297. [doi: 10.1111/j.1525-1497.2006.00396.x] [Medline: 16686803]
53. Real human data. Ready in minutes.. Prolific. 2024. URL: https://www.prolific.com [accessed 2025-03-03]
54. Yousef M, Rundle-Thiele S, Dietrich T. Advertising appeals effectiveness: a systematic literature review. Health Promot Int 2023 Aug 1;38(4):daab204. [doi: 10.1093/heapro/daab204] [Medline: 34931233]
55. The data science code editor. Posit. URL: https://www.posit.co/ [accessed 2025-03-27]
56. ordinal: regression models for ordinal data. Comprehensive R Archive Network (CRAN). 2024. URL: https://cran.r-project.org/web/packages/ordinal/index.html [accessed 2025-03-27]
57. Unger JM, Cook E, Tai E, Bleyer A. The role of clinical trial participation in cancer research: barriers, evidence, and strategies. Am Soc Clin Oncol Educ Book 2016;35(36):185-198. [doi: 10.1200/EDBK_156686] [Medline: 27249699]
58. Unger JM, Barlow WE, Martin DP, et al. Comparison of survival outcomes among cancer patients treated in and out of clinical trials. J Natl Cancer Inst 2014 Mar;106(3):dju002. [doi: 10.1093/jnci/dju002] [Medline: 24627276]
59. Espeland MA, Crimmins EM, Grossardt BR, et al. Clinical trials targeting aging and age-related multimorbidity. J Gerontol A Biol Sci Med Sci 2017 Mar 1;72(3):355-361. [doi: 10.1093/gerona/glw220] [Medline: 28364543]
60. Gabel M, Gooblar J, Roe CM, Selsor NJ, Morris JC. Political ideology and participation in clinical studies. Alzheimer Dis Assoc Disord 2018;32(3):179-184. [doi: 10.1097/WAD.0000000000000244] [Medline: 29351092]
61. Gabel M, Gooblar J, Roe CM, Morris JC. The ideological divide in confidence in science and participation in medical research. Sci Rep 2021 Feb 4;11(1):3120. [doi: 10.1038/s41598-021-82516-6] [Medline: 33542334]
62. Onyeaka H, Weber DB, Chido-Amajuoyi O, Muoghalu C, Amonoo HL. The influence of political ideology on clinical trial knowledge, invitation, and participation among adults in the United States. Clin Trials 2023 Dec;20(6):708-713. [doi: 10.1177/17407745231178790] [Medline: 37345562]
63. Seidler EM, Keshaviah A, Brown C, Wood E, Granick L, Kimball AB. Geographic distribution of clinical trials may lead to inequities in access. Clin Investig (Lond) 2014 Apr;4(4):373-380. [doi: 10.4155/cli.14.21]
64. Unger JM, Moseley A, Symington B, Chavez-MacGregor M, Ramsey SD, Hershman DL. Geographic distribution and survival outcomes for rural patients with cancer treated in clinical trials. JAMA Netw Open 2018 Aug 3;1(4):e181235. [doi: 10.1001/jamanetworkopen.2018.1235] [Medline: 30646114]
65. Kim SH, Tanner A, Friedman DB, Foster C, Bergeron CD. Barriers to clinical trial participation: a comparison of rural and urban communities in South Carolina. J Community Health 2014 Jun;39(3):562-571. [doi: 10.1007/s10900-013-9798-2] [Medline: 24310703]
66. Dunlop AL, Leroy ZC, Logue KM, Glanz K, Dunlop BW. Preconsent education about research processes improved African Americans' willingness to participate in clinical research. J Clin Epidemiol 2011 Aug;64(8):872-877. [doi: 10.1016/j.jclinepi.2010.11.008] [Medline: 21330104]
67. Morgan SE, Harrison TR, Wright KO, Jia X, Deal B, Malova K. The role of perceived expertise and trustworthiness in research study and clinical trial recruitment: perspectives of clinical research coordinators and African American and Black Caribbean patients. PLoS ONE 2023;18(6):e0275770. [doi: 10.1371/journal.pone.0275770] [Medline: 37342999]
68. Chatio ST, Ganle JK, Adongo PB, et al. Strategies to build trust in the conduct of clinical trials: stakeholders' views in a qualitative study in Ghana. PLOS Glob Public Health 2025;5(4):e0003201. [doi: 10.1371/journal.pgph.0003201] [Medline: 40198690]
69. Fisher JA. Institutional mistrust in the organization of pharmaceutical clinical trials. Med Health Care Philos 2008 Dec;11(4):403-413. [doi: 10.1007/s11019-008-9154-y] [Medline: 18633728]
70. Harry T, Hussain Z, Cao J, et al. Randomized comparison of online motivational themes in clinical trial recruitment. Circ Cardiovasc Qual Outcomes 2025 Dec;18(12):e012945. [doi: 10.1161/CIRCOUTCOMES.125.012945] [Medline: 41201904]
71. Thorson E, Wicks R, Leshner G. Experimental methodology in journalism and mass communication research. Journal Mass Commun Q 2012 Mar;89(1):112-124. [doi: 10.1177/1077699011430066]

Abbreviations

GoFresh: Groceries for Black Residents of Boston to Stop Hypertension
OR: odds ratio
RQ: research question

XSL•FO
RenderX

Edited by A Stone; submitted 12.Apr.2025; peer-reviewed by T Aisha, C Kirkpatrick, Y Zhang; revised version received 25.Nov.2025; accepted 19.Dec.2025; published 19.Mar.2026.

Please cite as:
Harry T, Cao J, Hussain Z, Turkson-Ocran RAN, Juraschek SP, Lahey TP, Plante TB, Feng Y
Guide on Selection of Optimal Motivational Themes for Use in a Clinical Trial Recruiting Black US Adults: Survey Study
J Med Internet Res 2026;28:e75857
URL: https://www.jmir.org/2026/1/e75857
doi:10.2196/75857

© Tamunotonye Harry, Jingyi Cao, Zaib Hussain, Ruth-Alma N Turkson-Ocran, Stephen P Juraschek, Timothy P Lahey, Timothy B Plante, Yuanyuan Feng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Ethical Handling of Occupational Health and Safety Data in the Fire Service: Empirical Interview and Focus Group Study of Firefighter and Fire Service Leadership Privacy Preferences

Rachel J Topazian[1], PhD; Aleksandra Wec[2], BA; Joseph Ali[3], JD; Shannon Frattaroli[2], MPH, PhD; Cassandra K Crifasi[2], MPH, PhD

[1]Division of General Internal Medicine, Center for Bioethics and Humanities, University of Colorado, 13080 E. 19th Ave, Mail Stop B137, Aurora, CO, United States

[2]Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States

[3]Department of International Health and Berman Institute of Bioethics, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States

Corresponding Author:
Rachel J Topazian, PhD
Division of General Internal Medicine, Center for Bioethics and Humanities, University of Colorado, 13080 E. 19th Ave, Mail Stop B137, Aurora, CO, United States

Abstract

Background: There are ongoing efforts to collect larger and higher-quality amounts of occupational health and safety data to better understand and prevent injuries and fatalities among high-risk workers, such as firefighters. Digital health systems including wearable technologies, mobile apps, or internet-based data collection platforms could collect large amounts of sensitive data, but there is little evidence on worker and employer perspectives on data privacy in the fire service.

Objective: Our study examined firefighters' and fire service leadership's preferences regarding occupational health and safety data privacy.

Methods: We conducted interviews and focus groups with career firefighters in Maryland and Virginia; interviews with union representatives and department-level leaders in each state; and interviews with national-level fire service leaders in advocacy, government, and research organizations (March to November 2023). Interviews and focus groups were audio recorded and transcribed. We analyzed transcripts using thematic analysis.

Results: The sample included 31 career firefighters, 2 union leaders, 11 national leaders, and 21 department-level leaders (65 total participants from 35 interviews and 4 focus groups). We identified 4 themes: acceptability of data access, sharing, and reporting practices; data sharing and access preferences; appropriate use of firefighter data; and the need for improved communication. Leaders described firefighters' concerns about job loss and loss of privacy. Firefighters expressed general preferences that their data be deidentified and not shared widely, and they identified mental health data as important but particularly sensitive information. Firefighters also expressed frustration about sharing data with researchers or their departments without knowing the purpose or outcomes. Both firefighters and leaders emphasized the need for enhanced communication and translation of data for firefighters.

Conclusions: Fire service leaders held more concerns about the use and sharing of occupational health and safety data than firefighters, but both groups identified ways to further safeguard firefighter data and improve communication about health and safety data. Future fire service data collection should incorporate privacy protections, such as limiting the collection of identifiable information and restricting data access. Data collection should be accompanied by clear communication about the purpose of the data collection, how firefighter data will be used and accessed, and the interpretation of the results. Future digital health interventions should integrate these data privacy protections to respect firefighter preferences and contribute to acceptability and uptake.

(J Med Internet Res 2026;28:e84465) doi:10.2196/84465

KEYWORDS

firefighter; occupational health and safety; data privacy; mental health data; data access; data sharing

XSL•FO
RenderX

Introduction

Firefighters take on great personal risks to provide essential services for their communities. There are approximately 1 million firefighters in the United States who respond to over 25 million calls each year [1,2]. Approximately two-thirds of these firefighters are volunteer or unpaid, and one-third are career ("paid") firefighters [1]. Firefighters encounter a vast array of incidents, including fires, natural disasters, and medical emergencies requiring emergency medical services [2]. US firefighters sustain high rates of injury and fatality in the line of duty: in 2022, there were over 65,000 injuries and 96 fatalities that occurred at work [3,4]. Beyond on-duty incidents, firefighting has been linked to long-term adverse health outcomes, including elevated rates of certain cancers [5,6].

There are growing calls from fire service leaders, government officials, and researchers to utilize firefighter health and safety data to better understand the hazards of firefighting and design interventions to protect future generations [7-12]. Proposed data sources include national injury or disease registries, biometric or exposure data from wearable devices, and self-reported risk factors or health outcomes from other digital health platforms, such as smartphone apps [8,13,14]. While recent data collection efforts have focused on injury and disease surveillance [8,9], researchers have identified flaws in existing data infrastructure, such as large swaths of missing occupational data in cancer and burn registries, highlighting the need for more sophisticated data reporting and data linkages [9,10,12]. Federal agencies are supporting new data platforms to enhance injury surveillance and research. In 2023, the US Fire Administration began developing the National Emergency Response Information System (NERIS), a new fire service analytics platform, and the National Institute for Occupational Safety and Health created the National Firefighter Registry for Cancer [7,15,16].

While these new data initiatives may bolster health and safety efforts, expanding the collection of or access to firefighter health and safety data raises ethical concerns [8]. Additional firefighter data might help improve health and safety, but collecting these data, particularly within fire departments, raises concerns about privacy, as firefighters might be asked or required to share sensitive information with their employers, government agencies, or researchers. The fire service's hierarchical organization—fire departments are frequently described as paramilitary organizations—adds additional complexity around data privacy, as firefighters of lower ranks may feel they have less power or choice as to whether or not they want to share personal health information with their employers.

There is little research on data privacy in the fire service, and existing studies focus on the acceptance of new technology in fire departments. Research from our study team examined perceptions of wearable data in the fire service, concluding that while leaders were optimistic about the potential for wearable data to ensure health and safety, firefighters had concerns about privacy, autonomy, and job loss [17]. Although firefighters identified major tradeoffs associated with wearable data, firefighters and leaders agreed on the need for data protections, such as deidentifying data, limiting wearable use to specific settings, providing firefighters access to their data, and preventing punitive data use. Another study examined the implementation of a biometric earpiece that firefighters could wear at a live fire, which would provide an incident commander with data including heart rate, body temperature, calories burned, and exertion estimates [18]. The study concluded that firefighters were concerned about the impact of technology on their autonomy and that firefighters were more skeptical than department leaders. Other researchers developed a biometric privacy framework based on firefighter attitudes toward a personnel management system that would use biometric information (eg, a fingerprint) to manage personnel at a fire. This study identified concerns related to workplace accountability, data access, and employer trust [19]. Collectively, these studies suggest that firefighters may have complex views on how their health and safety data are handled in the workplace, but we lack robust evidence on their privacy preferences and whether or not these preferences are understood and honored in fire departments.

Our study examined fire service members' and their partners' views regarding occupational health and safety data privacy. We explored data sharing, access, and communication preferences among career firefighters, fire department leaders, and national fire service leaders. Our findings highlight opportunities to improve current data privacy practices and ensure that new occupational health and safety digital health platforms, from wearables to mobile health to web-based surveys, are ethically implemented.

Methods

Sample and Recruitment

We purposively recruited participants from 3 groups: career (eg, "paid") firefighters (including union leaders), fire department–level leaders, and national-level leaders. We focused on career as opposed to volunteer firefighters because we wanted to understand views on data privacy among full-time employees. We sampled fire departments that had participated in the Firefighter Organizational Culture of Safety survey. The Firefighter Organizational Culture of Safety survey is used in fire departments to assess safety climate [20]. We recruited 4 fire departments: 2 in Maryland and 2 in Virginia. We targeted departments of varying sizes that were in different regions of each state to facilitate collecting a range of perspectives. We recruited firefighters and department leaders from within the same fire departments to allow for comparisons between management and the rank-and-file within an organization. For participants within fire departments (department leaders, union leaders, firefighters), we asked a point of contact to suggest participants in specific roles whom we then invited to participate. Department leaders and union leaders were invited to participate in interviews, while career rank-and-file firefighters were invited to participate in focus group discussions. Participants were recruited via email. We encountered data collection challenges in Maryland that limited the size of the firefighter sample (see more in the *Results* and *Discussion* sections).

XSL•FO
RenderX

We conducted interviews with leaders within fire departments, including those in overall leadership, health and safety roles, or human resources. Leadership participants included civilians and sworn members of the fire service. For the national-level leaders, we conducted purposive sampling of individuals from government, advocacy, and research organizations. We also conducted snowball sampling for both department and national leaders. We believed that firefighters might hold a range of privacy views but have less experience discussing their privacy preferences. Although some individuals might feel less comfortable sharing their views in a focus group, we chose to conduct focus groups with firefighters to allow for cross talk and the exchange of ideas on an infrequently discussed topic. We anticipated that leaders might have stronger preconceived notions of privacy norms specific to their roles and chose to conduct interviews to investigate their individual views in depth.

Data Collection

We developed interview and focus group guides to elicit views on data privacy. We aimed to understand how participants perceived several privacy domains: data collection, data use, and data access or sharing [21]. We developed questions that examined issues within these domains in the fire service and asked all participants questions within these domains, but we tailored questions as needed based on role. The guides were reviewed and refined by the study team. We collected demographic information from all participants. Multimedia Appendix 1 shows the interview and focus group discussion topics, including the key domains and high-level questions. Regardless of the data collection mode (interview vs focus group), we asked all participants about the same domains and then asked several follow-up questions to probe more deeply into interesting lines of inquiry related to the overall topic.

We conducted focus groups in person at fire stations. For some focus groups, firefighters were on duty and in service, meaning that they were available to respond to calls. If participants received a call during the discussion, it was paused or continued with the remaining participants until they returned. Interviews were conducted over Zoom and via phone. Data were collected from March to November 2023.

Consistent with qualitative research practices [22], the interviewer or moderator wrote a memo after each focus group or interview to capture initial themes and observations. All interviews and focus groups were audio recorded and transcribed verbatim by a professional transcription service.

Analysis

We analyzed interview and focus group transcripts using thematic analysis [23]. To develop an initial codebook, 1 team member reviewed the discussion guides, memos, and transcripts and drafted an initial set of codes capturing key concepts. A second team member reviewed the guides, memos, and transcripts and then updated the draft codebook. These 2 team members both coded a subset of interview and focus group transcripts, revised the codebook, and then independently coded the remaining transcripts. They continued to meet to discuss questions about their application of the codes and whether additional codes or refinements to the definitions were needed throughout the coding process. After completing a majority of the coding, the team members examined individual codes and groupings of codes that related to data privacy. The coding team identified themes capturing core concepts and trends in the data based on these codes and groupings of codes. We determined that we had reached thematic saturation within each stakeholder group when no new themes were emerging in later transcripts. The codebook, example code application, and themes were shared with the full study team periodically to obtain their input. Coding was conducted in Dedoose (version 10.0.59; SocioCultural Research Consultants, LLC) [24].

To compare perspectives across the fire service community, we grouped participants into 2 categories: firefighters (career firefighters and union leaders) and leadership (department leaders and national leaders). These 2 categories reflect conceptual similarities within groups (eg, union leaders are themselves firefighters who represent firefighters' interests), and there were similar results within each group. When describing our findings, we primarily refer to these 2 groups.

None of the coauthors have experience as volunteer or career firefighters, which may have impacted the study and the interpretation of results. The interviewer or moderator participated in ride-alongs with Maryland firefighters to gain insights into firefighting and fire service culture and discussed data privacy issues with local fire service collaborators in Maryland and Virginia to inform the study materials and analyses. One author is a commissioner of a Maryland fire department and may have brought a management perspective to the study. The authors discussed the study findings with study partners in the fire service to help interpret the results. The coders also consulted researchers with extensive experience conducting research with the fire service. This study follows the COREQ (Consolidated Criteria for Reporting Qualitative Research) guideline (Checklist 1) [25].

Ethical Considerations

This research was approved by the Johns Hopkins Bloomberg School of Public Health Institutional Review Board (24091). We obtained oral consent from all participants. Participants were not compensated for their participation. We deidentified interview transcripts to protect participant privacy.

Results

Sample

The sample included a total of 65 participants from 35 interviews and 4 focus groups (5 - 9 participants in each): 31 career firefighters, 2 union leaders, 21 department leaders, and 11 national leaders (Table 1).

XSL•FO
RenderX

Table . Interview and focus group participant roles by jurisdiction[a].

Participants	Virginia, n	Maryland, n	National, n	Total, n
Firefighters				
Career firefighters	30	1	—[b]	31
Union leaders	1	1	—	2
Leadership				
Department	7	14	—	21
Government	—	—	3	3
Advocacy	—	—	6	6
Research	—	—	2	2
Total	38	16	11	65

[a]The data were collected from March to November 2023.

[b]Not applicable.

We encountered data collection challenges in Maryland, which limited the firefighter sample in that state (1 Maryland firefighter participated in a key informant interview compared to 30 Virginia firefighters in focus groups); this limitation is discussed in the *Discussion* section. Overall, our sample was largely White and male, which reflects the fire service's gender and race demographics [26]. The leadership sample was more diverse in terms of race and gender: 25% (8/32) of leaders were non-White versus 9% (3/33) of firefighters, and 34% (11/32) of the leadership participants were women versus 12% (4/33) of firefighters. Leaders were also older and more experienced than firefighters: the mean age of leadership participants was 47.7 (SD 10.2) years versus 35.9 (SD 7.5) years, and the mean years of experience was 24.6 (SD 11.9) years versus 11.8 (SD 8.8) years. Table 2 contains self-reported demographic characteristics. Focus groups were, on average, 77 minutes. Firefighter interviews averaged 39 minutes, and leadership interviews averaged 45 minutes.

Table . Interview and focus group participant characteristics by role[a].

Participant characteristics	Firefighters (n=33)	Leadership (n=32)
Gender, n (%)		
Male	29 (88)	21 (66)
Female	4 (12)	11 (34)
Race, n (%)		
White	30 (91)	24 (75)
Non-White	3 (9)	8 (25)
Age (y), mean (SD)	35.9 (7.5)	47.7 (10.2)
Experience (y), mean (SD)	11.8 (8.8)	24.6 (11.9)

[a]The table shows self-reported demographic characteristics. One firefighter did not report their age. Two leadership participants did not report their years of experience. The data were collected from March to November 2023.

Themes

Overview

We identified four themes: (1) acceptability of current data access, sharing, and reporting practices; (2) data sharing and access preferences; (3) appropriate use of firefighter data; and (4) need for improved communication about data. Table 3 shows the themes and provides an overview of themes by group (firefighters and leadership).

XSL•FO
RenderX

Table . Overview of themes[a].

Themes	Leadership	Firefighters
Acceptability of current data access, sharing, and reporting practices	All participants described limited sharing of identifiable data and adequate protections for most health and safety data.	All participants described limited sharing of identifiable data and adequate protections for most health and safety data.
Data sharing and access preferences	Leaders described concerns about data sharing and ideal conditions for facilitating data collection, such as deidentification and limiting sharing.	Firefighters largely did not share these concerns but favored deidentification and identified sensitive kinds of information, such as mental health data.
Appropriate use of firefighter data	Leaders believed that firefighters were resistant to sharing information based on fear of termination or job impacts; many linked this to fears about losing identity as a firefighter.	For the most part, firefighters did not share these concerns.
Need for improved communication about data	Leaders acknowledged the importance of communicating well about data.	Firefighters and advocates emphasized the need for data interpretation and dissemination.

[a]The table shows themes from focus groups and interviews with firefighters and fire service leaders (March to November 2023).

Acceptability of Current Data Access, Sharing, and Reporting Practices

Within departments, firefighters and leadership described limited sharing of firefighter health and safety data, and most believed that such data were adequately protected by their departments. Nearly all leaders indicated that access to identifiable firefighter health or safety information was granted on a need-to-know basis. For instance, injury information was accessible only to health and safety staff. Most firefighters were required to complete an annual physical exam, but most departments did not receive detailed reports of members' annual physical results, just a notification of whether they passed or failed.

Externally, department leaders mentioned being legally required to report injuries to government agencies, such as the Occupational Safety and Health Administration, and third parties, such as the National Fire Protection Association and the International Association of Fire Fighters union. Most firefighters were uncertain about how the data might be reported externally but assumed that injury information was shared with government agencies or third parties and viewed this sharing as acceptable.

For the most part, firefighters appeared comfortable with the kinds and amounts of information they shared with their departments, as well as how this information was accessed by department leaders. Some simply did not express strong views on current data sharing practices, whereas others shared that efforts to limit data sharing were sufficient. Others stated that sharing personal information related to their physical health simply came with working in a high-risk, public job.

> *I don't have any real experiences or bad or positive experiences with my data in the department, except for just our health as in our yearly physicals and then stress tests and things like that. And I haven't had any issues with those or anything.* [Firefighter]

> *I'm sure no one loves to give all their personal information to an employer, but it's probably necessary in this particular field.* [Firefighter]

While privacy protocols and confidentiality were reported to be formally upheld, the culture of informal information sharing within departments differed; many firefighters and a few leaders noted that news about injuries traveled quickly.

> *Just like with the injury thing, you can decipher who's, we heard, oh, such and such went to [...] the mental health rehab center. Oh, this guy's on injury leave for the next three months. Wonder where he is. We can find out things by digging.* [Firefighter]

Although a few individuals were troubled by how quickly news about health diagnoses or injuries could spread in their departments, they pointed out that this typically occurred through informal channels in which firefighters shared information with each other, not through leadership.

Data Sharing and Access Preferences

While most firefighters were comfortable with their departmental data sharing practices, some expressed preferences regarding the kinds of information they were comfortable sharing and with whom they were comfortable sharing it. Many had an ideological preference for limiting access to identifiable information and opting for aggregated data sharing both within and outside of departments. Union representatives concurred—they described encountering broad resistance among firefighters to sharing medical information, although the firefighters in our sample did not express these concerns. Firefighters were generally supportive of sharing information with researchers conducting health and safety research, although some described their frustration that they did not receive individual results or overall results of studies in which they participated (described in more detail below).

Reluctance to share identifiable information appeared to be heightened based on the sensitivity of the data in question. One female firefighter pointed out that she viewed reproductive health information as more private than physical injury information. Mental health data were highly valued by most participants—firefighters emphasized the need for more mental health research and services—but they characterized mental health data as complicated to collect and access. Firefighters generally agreed that mental health data were closely held, both

XSL•FO
RenderX

among leadership and peer networks, but participants were still concerned about who might have access to identifiable sensitive information. In one focus group, firefighters described their dismay when asked to complete a mental health research survey that was theoretically anonymous but asked for information that participants believed could be used to identify them. Several firefighters explained:

> *Participant 4: Well, but that survey we just did, it has a lot of pretty in-depth. If you take it seriously, it has a lot of in-depth personal mental health [questions]. Like, for example, I think there's some on there like, have you ever considered committing suicide? Like, how many times in the last week have you felt like, that you were too depressed to come to work? ...*
>
> *Participant 3: And you don't know where it's going.*
>
> *Participant 4: You don't know where it's going.*
>
> *Many: Yeah.*
>
> *Participant 4: Then you get to the end. It's like basically you got to describe your eye color, you know. [laughter]* [Firefighters]

Leadership was aligned with firefighters in preferring that the data be deidentified or aggregated for research and prevention. For some, this was a matter of principle. Others thought that firefighters were less likely to share information if they believed it would be identifiable:

> *Even if you tell them it's going to nobody but me or nobody but this specific office, they are less likely to participate if there is identifying information.* [Department leader]

Others indicated that firefighters were reluctant to share mental or behavioral health information due to fear of reprisal and the sensitive nature of these data. Some research and advocacy participants thought that firefighters were more likely to share information with external entities than within their departments. There were mixed views among leaders about whether firefighters were opposed to sharing data within departments on principle, with some describing widespread privacy concerns and others indicating that such views were rare.

Beyond individual-level data collection and sharing, a small number of leaders expressed a desire to expand access to population-level data across the fire service. These participants described a tension between using data to advance health and safety priorities while respecting individual privacy rights. Several cited structural challenges in securely storing or accessing data, which prevented this kind of population-level analysis, but they believed it would be useful in the future.

Appropriate Use of Firefighter Data

Firefighters in our sample expressed few concerns about how their data were currently being used, whereas leadership described their views on appropriate ways of utilizing firefighter data and recalled instances in which they had heard concerns from firefighters. Leaders emphasized the importance of using firefighter data to improve health and safety but acknowledged that this could raise concerns about data being used against firefighters. Several department leaders stressed that health and safety data should be used in nonpunitive ways:

> *We try really hard in our office, at least the safety office, that we are not a punitive type of section of the department. Because once we become punitive, people stop talking to us. People stop trusting us. And our job is to try to make things safer for them. And we can't make things safer if we don't really truly understand what the problems are, you know?* [Department leader]

Leaders across all subgroups, as well as union representatives, believed that firefighters were fearful that sharing health or injury information could result in job loss or changes in job status, ultimately impacting their identities as firefighters. Many described encountering concerns among some firefighters that if they did not pass medical or fitness exams, they would be terminated. Most department leaders believed these views were unfounded and expressed a desire to use these exams for health and safety purposes:

> *Because you have firefighters that will be like, "a machine or a doctor or this METS [metabolic equivalents] does not dictate on how I can fight fire," which is true ... [but] all the research they've done with this, says that if your METS are below eight, you are at a higher risk of dying of an MI [myocardial infarction] or something out on fire scene. So that's the biggest issue I feel in the fire service. Everyone thinks that we're trying to take their job away or take a ton of money out of their pocket, what they don't understand is, no, we're trying to keep you healthy and safe and everything like that.* [Department leader]

In other cases, leaders anticipated firefighter concerns that sharing health information would adversely affect their day-to-day work even if their employment status remained unchanged:

> *I know a couple cases [of] people that never reported anything because they didn't want to go through the process. Because you got to go to the county clinic and then you got go to the headquarters, and most people get into this job for adrenaline rush and helping people. And then you're on modified duty, sitting behind a desk with limitations.* [Department leader]

Several individuals framed these concerns as fears of losing their identity as a firefighter:

> *Their fear isn't so much losing their job, their fear is losing their identity, who they are.* [Advocate and department leader]

While these concerns were widespread among leaders, most firefighters did not discuss job impacts. A few were worried that their annual physicals had become so in-depth that they might identify health risks for which they could be terminated or moved to light duty, although this perspective arose in only one focus group. Union leaders attributed hesitation about sharing medical information to underlying concerns about fitness

XSL•FO
RenderX

for duty and privacy but were still supportive of requiring annual physical exams.

Calls for Improved Communication About Data

As noted earlier, many participants advocated for sharing data with firefighters. However, participants also described shortcomings in how departments or researchers currently communicate with firefighters.

Regardless of how data originated, department and national leaders strongly endorsed the importance of returning individual data and interpretations of data back to firefighters. Many supported sharing information directly with individual firefighters, such as returning each individual's medical data from an annual exam to empower them to make healthy choices. Participants also emphasized the importance of communicating aggregated and analyzed data to provide education on topics, such as injury prevention or fitness practices. Department leaders and advocates highlighted the importance of partnering with trusted leaders, such as union officials, and communicating effectively about the purpose of data collection. A small number of participants described challenges in sharing data effectively, noting that the fire service or their department did not have a strong track record of effectively communicating about data.

Despite this widespread endorsement of high-quality communication from leadership, many firefighters noted that there was significant room for improvement. Firefighters were vocal about the need for better communication regarding data in general and better interpretation of the data that they provided for departmental or research purposes. Several firefighters wanted more information about the purpose behind data collection efforts or the significance of individual results. For instance, some described not knowing how to interpret annual physical exam results:

> *I'm going to voice off. I think the value sucks, to be honest. They [the fire department] do blood work and they do a full physical but nine times out of 10 we get a paper back that says, hey, there was something abnormal about your blood work or something we found, and that's it. No follow up, no telling you what it actually is. And to actually follow up with them to find out is really hard. They do the work, but they don't actually give you the results.* [Firefighter]

Others were frustrated with the lack of research results described earlier. The lack of communication about their own data and study results led them to conclude that a large share of data collection has no purpose or ultimate benefit to firefighters.

> *Participant 2: Well look at the urine samples that they [researchers] would collect, stuff like that. They had us give a urine sample after we got back from fires to see, hey, do you have whatever…*
>
> *Participant 9: This has been absorbed into your bloodstream.*
>
> *Participant 2: Arsenic, you know, absorbed into your body.*
>
> *Participant 3: We did it for two years and got…*
>
> *Participant 5: And they lost funding and...*
>
> *Participant 2: Right. They lost funding and it went away. Or every sample you gave, you got zero back from…*
>
> *Participant 5: No idea.* [Firefighters]

Discussion

Our study examined firefighters' and fire service leaders' data privacy preferences and identified opportunities for fire departments and researchers to ensure firefighter data are appropriately safeguarded and utilized. Leaders had more concerns about data use and data sharing, attributed in part to anticipated firefighter privacy concerns, whereas firefighters held general preferences for deidentifying data and limiting sharing, particularly for sensitive information, such as mental health data. Both groups called for improved data communication.

Interestingly, leaders in our sample voiced more concerns about current data practices than firefighters, and leaders were more attuned to the need for and barriers to ethical data collection, use, and sharing. This might reflect the fact that leadership was more likely to encounter diverse views on data than firefighters working in a single fire station or to hear concerns held by a small number of firefighters but elevated to management. It might also reflect the responsibilities that accompanied their roles—leaders who regularly interacted with firefighter data, or who were responsible for policy implementation, may have had more opportunities to consider the ethical and legal obligations surrounding such data. This aligns with research from our study team examining perceptions of the value of data in the fire service, which found that fire service leaders identified more use cases for occupational health and safety data than firefighters [27]. Leaders were likely more familiar with occupational health and safety data, more likely to value it, and therefore more likely to have strong views on the privacy of that data. Although leaders had more specific concerns about privacy practices than firefighters, both groups identified opportunities to limit access to sensitive data and improve communication.

The launch of new systems, such as NERIS, offers opportunities to adopt a "privacy by design" or "value-sensitive design" approach in which safeguards for privacy or other values are built directly into digital health systems [28-30]. Privacy by design focuses on privacy protections, whereas value-sensitive design accounts for moral values such as privacy, in addition to other values such as ease of use or safety benefits. Value-sensitive design may be a useful framework for fire service technology developers, who may be designing technologies to enhance one moral end (eg, protecting worker health and safety) but need to balance other moral values (eg, protecting privacy) and practical design considerations (eg, usability of a smartphone app for firefighters to track injuries).

For instance, firefighters in our sample expressed hesitancy to share sensitive information about their mental health despite recognizing a need for more mental health research and screening. A value-sensitive mental health data collection system in a mobile app could be designed to anonymize responses, impose limits on collecting sensitive data (eg, suicidal ideation),

XSL•FO
RenderX

or place limits on who has access to this information (eg, union leaders instead of department leaders). Leadership identified a broader array of potential data protections, such as limiting identifiable data collection and avoiding punitive use of health and safety data. The developers of digital health interventions for the fire service could take these preferences into account. For example, the developers of data surveillance systems, such as NERIS, could address these priorities by limiting the collection of identifiable information when gathering population-level injury data. Fire service technology companies could design wearable devices and smartphone apps that incorporate privacy features, such as allowing firefighters to opt into sharing specific identifiers or kinds of data with developers, their employers, researchers, or government officials. While these findings have implications for a range of digital health technologies and developers, fire departments, researchers, and government agencies can also help develop policies and practices accompanying these technical safeguards to ensure that health and safety data are used in protective, not punitive, ways.

Our findings also emphasize the importance of transparent and high-quality communication about firefighter health and safety data, both in research and practice settings. Many of the firefighters we spoke with were broadly supportive of research on topics, such as cancer, but were apathetic about their own contributions. Several described their frustration that after sharing sensitive information, such as symptoms of psychological distress or biospecimens, they never heard about the purpose or outcome of a study in which they participated. There are clear and well-established ethical imperatives for communicating the purpose of a study to research participants [31]. Our findings suggest there is strong demand from firefighters for researchers to clearly communicate about the objectives and conclusions of their work. Understanding these kinds of communication best practices will be particularly important as the National Institute for Occupational Safety and Health rolls out the National Firefighter Registry, a national-level longitudinal research effort. There are fewer well-established norms for communication about this kind of public health surveillance as opposed to individual research studies. Additional research on this topic could help establish guidelines on if and how firefighters want to receive information from surveillance activities such as the National Firefighter Registry.

Communicating individual research results is a more complicated endeavor; there is robust debate about when it is appropriate to return individual research results to participants, based on the kinds of data in question and how results might be communicated [32,33]. However, there is growing consensus that study participants want to receive medically significant and actionable findings, which is consistent with the views of firefighters in our sample [32,34]. While researchers cannot always ensure that findings are translated into practice, they can commit to providing research participants with individual results that might increase awareness about their own health and to sharing the conclusions of a study. Research ethics literature offers best practices for how researchers can communicate with participants, including offering context, using clear and simple language, and providing opportunities for questions [35,36].

Our findings also reveal the need for more research on the best manner in which to return health or medical information collected in occupational settings (not for research), for instance, from an annual physical exam or a department-led safety initiative. Researchers and employers operate in different environments, and the nature of data collection will vary accordingly. For example, firefighters may be required to participate in an annual exam as part of the terms of their employment, whereas research participants voluntarily enroll in a study. Research studies aim to produce generalizable knowledge and societal benefits, whereas a department's safety efforts might have limited population-level benefits. More work is needed to determine the ethical obligations of employers in communicating health and safety data to their employees. Ensuring ethical, high-quality communication about these kinds of data may not only benefit individual firefighters but could also help bolster trust in data collection efforts within the fire service writ large.

Finally, our findings reflect a need for robust ethical guidance on how to appropriately handle health and safety data in occupational settings, including the establishment of ethical principles as well as detailed, practical guidance on privacy topics. Although our participants did not raise major alarms about unethical data practices, they described legitimate concerns about ensuring firefighter data privacy while calling for large amounts of new data collection. The concerns that health and safety data could be used in punitive ways—to fire someone or change their job responsibilities without cause—reflect serious potential repercussions from insufficient data privacy protections. A few experts have discussed the need for confidential data practices for health and safety data, but guidelines for ethical data practices in any occupational setting remain general [37]. The COVID-19 pandemic revealed the need for ethics and data governance guidance for employers, as they began to deploy biometric screenings and digital contact tracing apps to attempt to control disease spread in workplaces, often with little ethical guidance [38,39]. While there were efforts to develop guidelines specific to COVID-19 [40], more work is needed to establish guidelines for the routine collection of occupational health and safety data, outside of a public health emergency. The continued refinement of ethical frameworks specific to occupational health and safety, as well as practical guidance on how to implement these frameworks, will help employers implement new data efforts in an ethical manner.

Our results should be interpreted with several key limitations in mind. Our findings reflect the specific geographic regions and fire departments in our sample. Fire departments were recruited from study team networks and had previously participated in research, which may have influenced their views toward research and data privacy. We encountered recruitment challenges in Maryland that limited the size of our firefighter sample in that state. We found similar results across Maryland and Virginia, but a larger sample may have captured additional viewpoints. Attitudes toward data privacy may differ based on department culture, trust in leadership, staffing (eg, combination, volunteer, or career), and geography. Future research should

XSL•FO
RenderX

explore this topic in a wider range of departments. Our sample was also largely White and male. While this reflects the overall demographics of the fire service [26], evidence from other settings suggests that women and racial minorities experience data and technology in distinct and often inequitable ways and thus may have unique privacy preferences [41-44]. More research is needed to understand how perspectives on occupational health and safety data privacy might vary based on gender and race.

Our findings offer insights into pathways for enhancing firefighter data privacy. Incorporating firefighters' privacy preferences into digital health data systems, for instance, by deidentifying firefighter data and limiting data access, may build confidence and trust in data collection efforts. Improving communication about data will address firefighters' desire to hear more about the purpose and impact of data collection. By addressing firefighters' privacy concerns and conducting high-quality, tailored communication, government, research, and department-level leaders can foster trust in data collection efforts and ensure that firefighter data are used in ethical and acceptable ways to protect their health and safety.

Acknowledgments

Generative AI was not used at any stage in the preparation of this manuscript.

Funding

RJT received tuition and stipend support from a grant from the US Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health to the Johns Hopkins Education and Research Center for Occupational Safety and Health (award T42 OH0008428), as well as the Susan P. Baker Scholarship in Injury Prevention and Control and Center for Qualitative Studies in Health and Medicine Dissertation Enhancement Award. AW was supported by grant T32AG066576 from the National Institute on Aging, National Institutes of Health. Data collection and analysis were funded by a pilot project award from the National Institute for Occupational Safety and Health Johns Hopkins Education and Research Center for Occupational Safety and Health (award T42 OH0008428).

Data Availability

The data are not available to protect participant confidentiality. The fire service communities studied are small, raising concerns about identifiability.

Authors' Contributions

Conceptualization: CKC, JA, RJT, SF
Data curation: AW, RJT
Formal analysis: AW, RJT
Funding acquisition: CKC, RJT
Investigation: AW, RJT
Methodology: AW, CKC, JA, RJT, SF
Project administration: RJT
Supervision: CKC, JA, SF
Writing – original draft: RJT

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview and focus group discussion topics.
[DOCX File, 31 KB - jmir_v28i1e84465_app1.docx]

Checklist 1
COREQ checklist.
[PDF File, 190 KB - jmir_v28i1e84465_app2.pdf]

References

1. Fahy R, Evarts B, Stein GP. U.S. fire department profile. National Fire Protection Association. 2022. URL: https://www.nfpa.org/education-and-research/research/nfpa-research/fire-statistical-reports/us-fire-department-profile [accessed 2023-12-01]

XSL•FO
RenderX

2. Fire department overall run profile as reported to the national fire incident reporting system (2020). : U.S. Fire Administration; 2020 URL: https://www.usfa.fema.gov/downloads/pdf/statistics/v22i1-fire-department-run-profile.pdf [accessed 2026-03-17]
3. Campbell R, Hall S, Petrillo J. United States firefighter injuries. National Fire Protection Association. URL: https://www.nfpa.org/education-and-research/research/nfpa-research/fire-statistical-reports/firefighter-injuries-in-the-united-states [accessed 2026-03-17]
4. Campbell R, Petrillo JT. Fatal firefighter injuries in the US in 2022. : National Fire Protection Association; 2023 Jun 18 URL: https://img1.wsimg.com/blobby/go/95969465-abe3-4680-bcb1-af5b16fc7ebd/downloads/2022%20LODD%20Report.pdf?ver=1741883121806 [accessed 2026-03-17]
5. Lee DJ, Ahn S, McClure LA, et al. Cancer risk and mortality among firefighters: a meta-analytic review. Front Oncol 2023;13:1130754. [doi: 10.3389/fonc.2023.1130754] [Medline: 37251928]
6. Laroche E, L'Espérance S. Cancer incidence and mortality among firefighters: an overview of epidemiologic systematic reviews. Int J Environ Res Public Health 2021 Mar 3;18(5):2519. [doi: 10.3390/ijerph18052519] [Medline: 33802629]
7. National firefighter registry (NFR) for cancer protocol. : Centers for Disease Control and Prevention (CDC); 2025 URL: https://www.cdc.gov/niosh/media/pdfs/2025/05/NFR-Protocol-Aprl_2025_OMB.pdf [accessed 2026-03-23]
8. Grant C, Hamins A, Bryner N, Jones AW, Koepke GH. Research roadmap for smart fire fighting. : National Institute of Standards and Technology; 2015 Jun URL: https://nvlpubs.nist.gov/nistpubs/specialpublications/nist.sp.1191.pdf [accessed 2026-03-17]
9. Widman SA, LeVasseur MT, Tabb LP, Taylor JA. The benefits of data linkage for firefighter injury surveillance. Inj Prev 2018 Feb;24(1):19-28. [doi: 10.1136/injuryprev-2016-042213] [Medline: 28196830]
10. Kahn SA, Mosier M. Firefighter injury data are wrought with discrepancy: pearls from the national burn repository. J Burn Care Res 2018 Jun 13;39(4):623-625. [doi: 10.1093/jbcr/irx025] [Medline: 29860394]
11. Averill J, Moore L, Notarianni K, Santos R, Wissoker D. Proceedings of the national fire service data summit. : National Institute of Standards and Technology; 2011 URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=908244 [accessed 2026-03-23]
12. McClure LA, Koru-Sengul T, Hernandez MN, et al. Availability and accuracy of occupation in cancer registry data among Florida firefighters. PLoS ONE 2019;14(4):e0215867. [doi: 10.1371/journal.pone.0215867] [Medline: 31039169]
13. Weidinger J. What is known and what remains unexplored: a review of the firefighter information technologies literature. Int J Disaster Risk Reduct ;78(2022):103115. [doi: 10.1016/j.ijdrr.2022.103115]
14. Taborri J, Pasinetti S, Cardinali L, Perroni F, Rossi S. Preventing and monitoring work-related diseases in firefighters: a literature review on sensor-based systems and future perspectives in robotic devices. Int J Environ Res Public Health 2021 Sep 15;18(18):9723. [doi: 10.3390/ijerph18189723] [Medline: 34574646]
15. About the National Emergency Response Information System (NERIS). US Fire Administration. 2023. URL: https://www.usfa.fema.gov/nfirs/neris/about-neris/ [accessed 2023-07-23]
16. National Firefighter Registry (NFR) for cancer. CDC. URL: https://www.cdc.gov/niosh/firefighters/registry.html [accessed 2022-07-22]
17. Topazian RJ, Wec A, Ali J, et al. Ethical use of wearable device data in occupational settings: perspectives from the fire service. J Occup Environ Med 2025 Sep 1;67(9):e621-e629. [doi: 10.1097/JOM.0000000000003446] [Medline: 40359076]
18. Rumsey A, Le Dantec CA. Clearing the smoke: the changing identities and work in firefighters. 2019 Presented at: DIS '19: Proceedings of the 2019 on Designing Interactive Systems Conference; Jun 23-28, 2019; San Diego CA, USA p. 581-592. [doi: 10.1145/3322276.3322292]
19. Carpenter D, McLeod A, Hicks C, Maasberg M. Privacy and biometrics: an empirical examination of employee concerns. Inf Syst Front 2018 Feb;20(1):91-110. [doi: 10.1007/s10796-016-9667-5]
20. Taylor JA, Davis AL, Shepler LJ, et al. Development and validation of the fire service safety climate scale. Saf Sci 2019 Oct;118:126-144. [doi: 10.1016/j.ssci.2019.05.007]
21. Richards N. Why Privacy Matters: Oxford University Press; 2021. [doi: 10.1093/oso/9780190939045.001.0001]
22. Miles MB, Huberman MA. Qualitative Data Analysis: An Expanded Sourcebook, 2nd edition: SAGE Publications, Inc; 1994. URL: https://openlibrary.org/books/OL21978508M/Qualitative_data_analysis [accessed 2026-03-17]
23. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
24. dedoose. URL: https://www.dedoose.com/ [accessed 2026-03-17]
25. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
26. Employed people by detailed occupation, sex, race, and Hispanic or Latino ethnicity. US Bureau of Labor Statistics. URL: https://www.bls.gov/cps/cpsaat11.htm [accessed 2024-01-03]
27. Topazian RJ, Wec A, Ali J, et al. The value of occupational health and safety data: perceptions of current data practices in the fire service. J Safety Res 2025 Sep;94:120-127. [doi: 10.1016/j.jsr.2025.06.016] [Medline: 40930628]

XSL•FO
RenderX

28. Cavoukian A. Privacy by design: the 7 foundational principles—implementation and mapping of fair information practices. : Office of the Information & Privacy Commissioner of Ontario; 2006 URL: https://student.cs.uwaterloo.ca/~cs492/papers/7foundationalprinciples_longer.pdf [accessed 2026-03-17]
29. Schaar P. Privacy by design. Ident Inf Soc 2010 Aug;3(2):267-274. [doi: 10.1007/s12394-010-0055-x]
30. Friedman B, Kahn PH, Borning A. Value sensitive design and information systems. In: The Handbook of Information and Computer Ethics: John Wiley & Sons, Inc; 2008:69-101. [doi: 10.1002/9780470281819]
31. The Belmont Report: ethical principles and guidelines for the protection of human subjects of research. : U.S. Department of Health & Human Services; 1979 URL: https://www.hhs.gov/ohrp/sites/default/files/the-belmont-report-508c_FINAL.pdf [accessed 2024-08-30]
32. Botkin JR, Mancher M, Busta ER, Downey AS, editors. Returning Individual Research Results to Participants: Guidance for a New Research Paradigm: National Academies Press; 2018. [doi: 10.17226/25094] [Medline: 30001048]
33. Wong CA, Hernandez AF, Califf RM. Return of research results to study participants: uncharted and untested. JAMA 2018 Aug 7;320(5):435-436. [doi: 10.1001/jama.2018.7898] [Medline: 29931289]
34. Vears DF, Minion JT, Roberts SJ, et al. Return of individual research results from genomic research: a systematic review of stakeholder perspectives. PLOS ONE 2021;16(11):e0258646. [doi: 10.1371/journal.pone.0258646] [Medline: 34748551]
35. Hintz EA, Dean M. Best practices for returning research findings to participants: methodological and ethical considerations for communication researchers. Commun Methods Meas 2020 Jan 2;14(1):38-54. [doi: 10.1080/19312458.2019.1650165]
36. Downey AS, Busta ER, Mancher M, Botkin JR, editors. Principles for the return of individual research results: ethical and societal considerations. In: Returning Individual Research Results to Participants: Guidance for a New Research Paradigm: National Academies Press; 2018:59-92 URL: https://www.ncbi.nlm.nih.gov/books/NBK525079 [accessed 2024-08-30]
37. Rogers B, Schill AL. Ethics and Total Worker Health®: constructs for ethical decision-making and competencies for professional practice. Int J Environ Res Public Health 2021 Sep 24;18(19):10030. [doi: 10.3390/ijerph181910030] [Medline: 34639332]
38. Gidengil C, Fischer S, Broten N. A framework for evaluating approaches to symptom screening in the workplace during the COVID-19 pandemic. : RAND Corporation; 2020 URL: https://www.rand.org/pubs/perspectives/PEA653-1.html [accessed 2026-03-17]
39. Ingram C, Downey V, Roe M, et al. COVID-19 prevention and control measures in workplace settings: a rapid review and meta-analysis. Int J Environ Res Public Health 2021 Jul 24;18(15):7847. [doi: 10.3390/ijerph18157847] [Medline: 34360142]
40. Kahn JP. Digital Contact Tracing for Pandemic Response: Ethics and Governance Guidance: Johns Hopkins University Press; 2020. URL: https://muse.jhu.edu/pub/1/oa_monograph/book/75831 [accessed 2024-10-01]
41. Benjamin R. Race after Technology: Abolitionist Tools for the New Jim Code: Polity; 2019. URL: https://www.politybooks.com/bookdetail?book_slug=race-after-technology-abolitionist-tools-for-the-new-jim-code--9781509526390 [accessed 2026-03-17]
42. Allen AL. Dismantling the “black opticon”: privacy, race, equity, and online data-protection reform. Yale Law Journal Forum. 2021 Nov. URL: https://www.yalelawjournal.org/forum/dismantling-the-black-opticon [accessed 2026-03-17]
43. Stark L, Stanhaus A, Anthony DL. “I don’t want someone to watch me while I’m working”: gendered views of facial recognition technology in workplace surveillance. J Assoc Inf Sci Technol 2020 Sep;71(9):1074-1088. [doi: 10.1002/asi.24342]
44. Jacobs JV, Hettinger LJ, Huang YH, et al. Employee acceptance of wearable technology in the workplace. Appl Ergon 2019 Jul;78:148-156. [doi: 10.1016/j.apergo.2019.03.003] [Medline: 31046946]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
NERIS: National Emergency Response Information System

Edited by A Stone; submitted 23.Sep.2025; peer-reviewed by KF Hollis, S Mohanadas, SJ Kim; revised version received 20.Feb.2026; accepted 23.Feb.2026; published 02.Apr.2026.

Please cite as:
Topazian RJ, Wec A, Ali J, Frattaroli S, Crifasi CK
Ethical Handling of Occupational Health and Safety Data in the Fire Service: Empirical Interview and Focus Group Study of Firefighter and Fire Service Leadership Privacy Preferences
J Med Internet Res 2026;28:e84465
URL: https://www.jmir.org/2026/1/e84465
doi:10.2196/84465

XSL•FO
RenderX

© Rachel J Topazian, Aleksandra Wec, Joseph Ali, Shannon Frattaroli, Cassandra K Crifasi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Ethics of Leveraging Routinely Collected Patient Data for AI Development: Mixed Methods Study

Menno T Maris[1,2], MSc; Joanna E Klopotowska[2,3], PhD, PharmD; Ronald Cornet[2,3], PhD; Mariëtte A Van den Hoven[2,4], PhD; Joris E Lieverse[2,3], PharmD; Daniel Fernández-Llaneza[2,3], MSc; Marieke A R Bak[1,2,5], PhD; LEAPfROG Consortium[6,7]

[1]Department of Ethics, Law and Humanities, Amsterdam UMC Location University of Amsterdam, Meibergdreef 9, Amsterdam, The Netherlands

[2]Amsterdam Public Health Institute, Amsterdam, The Netherlands

[3]Department of Medical Informatics, Amsterdam UMC location University of Amsterdam, Amsterdam, The Netherlands

[4]Department of Ethics, Law and Humanities, Amsterdam UMC location Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[5]Institute of History and Ethics in Medicine, TUM School of Medicine, Technical University of Munich, Munich, Germany

[6]See Acknowledgments

[7]

Corresponding Author:
Menno T Maris, MSc
Department of Ethics, Law and Humanities, Amsterdam UMC Location University of Amsterdam, Meibergdreef 9, Amsterdam, The Netherlands

Abstract

Background: Electronic health record (EHR) data, a key form of routinely collected patient data, offer great potential for medical research and the development of artificial intelligence (AI) tools. However, because these data are primarily gathered for health care rather than research, it often lacks the quality needed for AI training, raising both methodological and ethical concerns. While previous studies have reviewed the ethical implications of both routinely collected patient data and AI separately, their intersection, where AI is applied to such data, remains largely unexplored.

Objective: This study aimed to examine the ethical challenges that arise at the intersection of EHR data and AI development and to derive practice-oriented recommendations using the Dutch LEAPfROG (Leveraging Real-World Data to Optimize Pharmacotherapy Outcomes in Multimorbid Patients Using Machine Learning and Knowledge Representation Methods) project as a guiding case.

Methods: We used a mixed methods design combining a scoping literature review with a systematic search and 2 stakeholder workshops structured by the guidance ethics approach, reflecting a staged and iterative process aligned with the LEAPfROG project's development phases. The review identified 25 relevant publications from 2014 to 2024. The workshops, conducted with 17 and 13 participants respectively, included patients, clinicians, ethicists, data officers, and AI developers. Both workshops used dialogue to identify ethical values, impacts, and action points, focusing on a case study of drug-induced acute kidney injury.

Results: The analysis highlighted four major themes: (1) data privacy, transparency, and consent, including challenges of meaningful consent and risks of reidentification; (2) public trust and regulatory challenges, such as fragmented oversight and inconsistent governance; (3) fair representation and model generalizability, where incomplete or biased data may perpetuate health inequities; and (4) responsible AI integration in clinical practice, including concerns about clinical tropism, administrative burden, and the risk of overreliance on AI outputs. Both literature and stakeholder perspectives underscore the risk of decontextualization when EHR data are reused and emphasize the importance of clearly defining the purpose of data reuse to ensure real-world applicability and foster trust.

Conclusions: Responsible AI development requires explicit attention to how EHR data are produced, interpreted, and governed in practice, recognizing that data quality and meaning are shaped by the clinical, institutional, and social contexts in which they originate. Technical solutions or top-down regulation alone are insufficient. Instead, stakeholder-led and context-sensitive approaches are needed to define the purposes, risks, and benefits of medical AI. Grounded in the realities of health care practice and in the perspectives of patients, clinicians, and data custodians, these approaches can strengthen transparency, fairness, and clinical relevance, ensuring that EHR data are used ethically and effectively to support equitable and trustworthy AI innovation.

(J Med Internet Res 2026;28:e79863) doi:10.2196/79863

XSL•FO
RenderX

KEYWORDS

artificial intelligence; routinely collected health data; ethics; stakeholder participation; medical informatics; pharmacotherapy

Introduction

Both routinely collected patient data and artificial intelligence (AI) are widely regarded as promising resources for advancing medical research. AI methods are valued for their ability to manage and analyze large, diverse datasets, making them potentially well-suited to maximize the potential of routinely collected patient data. Studies leveraging such data offer several advantages over other study designs, including lower administrative costs, adaptability to evolving practice patterns, and access to larger sample sizes [1]. One of the fields where leveraging routinely collected data with AI seems particularly promising is individualized pharmacotherapy, focusing on the "rational use of drugs" to tailor pharmacotherapy to patients' clinical needs, thereby maximizing benefits and minimizing harm [2]. Effective innovation in pharmacotherapy is currently especially challenging for patients with multimorbidity, characterized by the presence of 2 or more chronic diseases and a frequent need for multiple medications, a phenomenon known as polypharmacy [3,4]. Despite the increasing prevalence of multimorbidity, partly due to population aging and improvements in health care services [5], clinical programs, guidelines, and research continue to primarily focus on managing individual diseases [3,6].

By leveraging routinely collected patient data from electronic health records (EHRs) with AI, the Dutch LEAPfROG Project (Leveraging Real-World Data to Optimize Pharmacotherapy Outcomes in Multimorbid Patients Using Machine Learning and Knowledge Representation Methods) seeks to enhance medication safety for patients with multimorbidity. A central focus of the project is the clinically urgent use case of drug-induced acute kidney injury (DAKI) in patients with chronic kidney disease (CKD). Ultimately, LEAPfROG aims to develop high-quality and fit-for-purpose EHR data, adhering to the principles of findability, accessibility, interoperability, and reusability [7] and to develop robust analytic methods, providing real-world evidence on pharmacotherapy outcomes and contributing to learning health systems that offer personalized treatments, improve quality of life, and reduce health care costs (Textbox 1).

Textbox 1. LEAPfROG (Leveraging Real-World Data to Optimize Pharmacotherapy Outcomes in Multimorbid Patients Using Machine Learning and Knowledge Representation Methods) project: improving drug safety in chronic kidney disease (CKD) through electronic health record (EHR) data and artificial intelligence (AI).

Background

The LEAPfROG project leverages routinely collected EHR data and structured domain knowledge to improve medication safety for patients with multimorbidity, specifically aiming to reduce the risk of drug-induced acute kidney injury (DAKI) in patients with CKD.

The LEAPfROG approach combines data from 3 EHR sources: detailed EHRs from Amsterdam UMC hospitals, EHRs from the Academic Network of General Practices Amsterdam (ANHA), and EHRs from general practices and outpatient pharmacies within the PHARMQ Database Network, covering general practices, outpatient pharmacies, and hospitals. Together, these datasets provide a rich source of insights into patient pharmacotherapy outcomes and drug exposures. To enhance these insights, LEAPfROG integrates domain knowledge from resources such as drug monographs, medication safety alerts, and scientific literature.

A key goal of the project is to develop transparent, explainable tools that can assist clinicians in making well-informed prescribing decisions. LEAPfROG's AI models combine machine learning with knowledge representation methods to detect DAKI, assess DAKI risk in patients with CKD, and clarify the underlying factors contributing to kidney safety issues. By doing so, the project aims to improve the detection of adverse drug events such as DAKI and enhance patient outcomes, particularly for patients with multimorbidity, such as those with CKD.

Causal machine learning model for retrospective diagnosis of DAKI

One of LEAPfROG's aims is the development of a causal machine learning model to assist physicians in determining whether acute kidney injury was caused by a drug (combination), a process known as "retrospective diagnosis." This model aims not only to identify CKD patients who developed DAKI but also to explain cause-and-effect relationships underlying the DAKI diagnosis. Through this effort, LEAPfROG aims to create a large dataset of patients who did and did not develop DAKI. This dataset is critical for training future machine learning models capable of identifying patients at risk of DAKI ("prognostic models"), ultimately helping to prevent such adverse drug events in future CKD patients.

Because EHR data are collected during routine care for clinical purposes, they often have quality limitations when reused for research, lacking the consistency required by clinical research standards [8,9]. Integrating AI with EHR data may therefore exacerbate or further obscure existing data shortcomings, ultimately impacting their meaningful use in clinical practice [10,11].

Previous studies have systematically reviewed ethical issues related to leveraging health data for medical research, including concerns related to the sharing and linkage of health data [12], public attitudes towards the reuse of health data for research [13], and broader ethical considerations of using EHRs for biomedical research [14]. Regarding the ethical implications of AI in health care, previous studies have reviewed both epistemic ethical issues (ie, the type of knowledge AI generates) and normative ethical concerns [15], along with empirical investigations into ethical considerations specific to medical AI [16]. However, literature reviews on the ethical considerations inherent in the integration of EHR data and AI technologies are currently nonexistent.

In this study, we aim to explore the ethical implications of reusing EHR data for AI-driven research and innovation in health care and to provide actionable recommendations based on our use case, the LEAPfROG project. This article presents

XSL•FO
RenderX

the results of a mixed methods approach that combines a scoping literature review with a systematic search and 2 innovative stakeholder engagement workshops grounded in the guidance ethics approach (GEA) [17]. The scoping review informed the design of the stakeholder workshops, which in turn generated empirical insights that complemented and contextualized the review's findings.

Methods

Study Design

The study followed an iterative, participatory design that combined a scoping review and stakeholder engagement workshops to explore the ethical implications of reusing EHR data for AI-driven health research. The scoping review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) extension for scoping reviews [18]. No protocol for the scoping review component of this study was registered. The stakeholder workshops were structured using the GEA for evaluating emerging technologies [17]. The overall design was staged and sequential rather than triangulated, with each step reflecting a different stage of the LEAPfROG project. This iterative process reflected the real-world complexity of AI ethics, allowing themes to unfold progressively, from broad reflection on EHR reuse to more specific considerations of AI development and implementation.

The workshops were integrated into the staged structure of the project, with their timing and number determined in advance and aligned with key project phases. Rather than aiming for thematic saturation, the workshops served as structured points of ethical reflection tied to the specific technology and context [17]. The scoping review identified ethical themes in the literature, providing a broad understanding of existing ethical considerations, while the workshops enabled further exploration and contextualization within the real-world context of the LEAPfROG project.

In the first workshop, participants discussed the ethical implications of reusing EHR data in the Dutch context, focusing on DAKI in patients with CKD. As these discussions aligned with initial literature findings, the subsequent scoping review shifted focus from the ethics of EHR reuse to ethical considerations at the intersection of EHR data and AI. In the second workshop, preliminary findings from the literature were presented to validate, contextualize, and further refine the identified themes. Furthermore, this workshop centered on a concrete AI-based tool within the LEAPfROG project: a causal machine learning model for the retrospective diagnosis of DAKI in patients with CKD.

Scoping Review

Search Strategy

MTM, JEK, RC, and MB jointly developed the search strategy to identify articles that thoroughly examine or discuss the ethical considerations associated with leveraging EHR data for advancing medicine using AI. A comprehensive search was conducted by MTM in PubMed, CINAHL (via EBSCO), and Web of Science, restricted to results between March 2014 and March 2024. The search was structured around 4 key concepts: "Artificial Intelligence," "Ethics," "Research," and "Routinely collected patient data" (Multimedia Appendix 1). Additionally, the most relevant MeSH (Medical Subject Headings) terms, synonyms, and related terms were identified, leading to the development of final search strings for each database. Results from the 3 databases were merged in EndNote (version 21; Clarivate Analytics), and duplicates were removed.

Selection Process and Analysis

One researcher (MTM) screened titles and abstracts using Rayyan [19]. Eligible studies were written in English and addressed ethical considerations at the intersection of AI, including machine learning and natural language processing, that analyze routinely collected patient data. Following Benchimol et al [20], such data were defined as health data collected without specific a priori research questions, including observational data from EHRs and disease registries. Particular emphasis during selection was placed on studies using EHRs or data comparable in origin, as these form the primary data source within the LEAPfROG project and are central to the ethical questions explored in this study. Because terminology varies across publications, the search strategy was intentionally broad to capture studies referring to similar data under related terms (eg, clinical or electronic medical records). Studies that also discussed other data sources (eg, clinical trial or wearable data) were included, provided the focus was sufficiently on the reuse of EHRs. Any uncertainties during full-text assessment were resolved by consensus between 2 researchers (MTM and MB). Reference lists were screened for additional relevant studies.

We uploaded full-text articles to MaxQDA and performed open coding to explore the ethical implications of using routinely collected EHR data for AI-driven research and innovation. Data extraction and coding were treated as the data charting process for the scoping review. Using a subset of 10 articles, one researcher (MTM) developed a coding scheme based on overlapping ethical considerations and emerging themes, such as "privacy" and "doctor-patient relationship." This framework was discussed with the second researcher (MB) throughout the process and guided the coding of the remaining articles by MTM.

Insights from the scoping review informed both the design and thematic focus of the second stakeholder workshop. The resulting data were analyzed inductively to explore how participants interpreted and prioritized ethical themes. This analysis refined and contextualized the review's findings, emphasizing those most pertinent to real-world practice within the LEAPfROG project and completing the staged, iterative process.

Guidance Ethics Workshops

Workshop Design

We organized 2 stakeholder engagement workshops using the GEA (translated from the Dutch "Aanpak Begeleidingsethiek"), a relatively new method for addressing the ethical aspects of technological innovation [17]. The workshops were conducted

XSL•FO
RenderX

in Dutch. The GEA uses dialogue to map the most pressing moral considerations in technology implementation, as identified by stakeholders [21]. The GEA primarily serves to guide the development of a technology by prioritizing ethical considerations relevant to the specific case, while striving to align with stakeholder values.

Both workshops, which lasted 3 hours each, followed the GEA approach and comprised three separate stages: (1) providing a clear description of the technology and its context, (2) identifying stakeholders, potential impacts, and underlying values, and (3) generating a list of options for responsible innovation and implementation from 3 distinct perspectives: technology, environment, and user (Textbox 2). The first workshop was prepared by JEK, RC, MTM, and MB and moderated by external facilitators from Stichting ECP. The second was prepared by JEK, MTM, MB, MvdH, and DF-L and moderated by JEK, MTM, MvdH, JEL, and MB.

Textbox 2. Stakeholder workshops design "guidance ethics approach."

Stage 1: Technology in context

i. Introduction by project leaders: Description and use context.

ii. Stakeholder inquiry: Opportunity for participants to gain a deeper understanding of and explore the details of the intended technology.

Stage 2: Dialogue

i. Actors: Individuals or groups involved in or affected by the proposed technology or solution.

ii. Effects: Potential benefits and harms arising from the development and implementation of the technology or solution.

iii. Values: Examination of the most pertinent ethical considerations related to the effects from step II.

Stage 3: Action points

Action points to promote responsible innovation at 3 levels of abstraction:

i. Technology*:* How should it be designed?

ii. Environment: Includes both tangible elements (eg, physical infrastructure) and systematic aspects (eg, policy).

iii. User: What can users do?

Participant Recruitment

Recruitment took place within the Dutch health care and research context through 2 structured routes. First, invitations were distributed via the LEAPfROG consortium network, including affiliated partners and organizations, following the project's progress. Second, patient participants were recruited through the Dutch Kidney Patients Association (Nederlandse Vereniging voor Nierpatiënten [NVN]). When invitees were unavailable, comparable participants were identified outside these routes to maintain diversity across stakeholder groups. All invited participants received a written background document explaining the LEAPfROG project, the purpose of the workshops, and their expected contribution. No prior preparation was required for participation. Patients were invited as experiential experts rather than professionals and were offered an introductory meeting to ask questions and familiarize themselves with the process. All participants provided informed consent before the workshops took place.

Before the workshops, participants were notified of the intention to publish a report on the session results and were given the opportunity to raise objections. Prior to publication, they were able to request changes and provide consent for the final reports, which are available in Dutch as Multimedia Appendices 2 and 3. The first workshop, led by external moderators from Stichting ECP, was not audio-recorded. To further strengthen the accuracy and completeness of the documentation, the second workshop was audio recorded, with participants informed in advance and given the opportunity to object.

Documentation and Validation

In both workshops, discussions were documented on flip charts or a shared screen, allowing participants to review and refine notes in real time. Notes from moderators and researchers were later compared and synthesized into validated reports. The GEA emphasizes collective ethical reflection, in which differing perspectives are explicitly explored rather than reconciled into consensus. In the first workshop, notes were taken independently by 2 external moderators from Stichting ECP, alongside 2 researchers (JEK and MTM). The moderators' notes from the first session formed the basis for the initial draft of the first workshop report. Notes from JEK and MTM were compared, reconciled, and extensively discussed within the research team to ensure completeness and accuracy. This iterative process led to the final validated report. The first draft of the second workshop report was prepared by MTM and subsequently finalized through joint review by MTM, JEK, and JEL.

Workshop Data Analysis

The workshop discussions, facilitator notes, and summary reports were thematically analyzed using the GEA framework, which emphasizes dialogue and collective reflection among diverse stakeholders. The analysis focused on how participants articulated ethical and practical considerations across the 3 reflective phases of the GEA: mapping the technology in context, ethical reflection, and translation to action. MTM led the review and synthesis of the workshop materials, identifying recurring themes and insights that emerged through these dialogues. The resulting reports were designed as standalone

XSL•FO
RenderX

outputs for dissemination within the LEAPfROG consortium. For the purpose of this paper, the analysis focused on themes most relevant to the research gap at the intersection of EHR data and AI development. The coding framework developed from the scoping review provided the analytical foundation, guiding the thematic structure while allowing refinement and contextualization based on stakeholder perspectives. Themes were subsequently reviewed and refined with the coauthors during manuscript preparation to ensure consistency and alignment with the broader study.

Ethical Considerations

The overall LEAPfROG project protocol was reviewed by the Medical Ethics Committee of Amsterdam UMC (the Netherlands). The committee granted a waiver of formal approval (W22_340 #22.412) because the LEAPfROG project, including all substudies, does not fall within the scope of the Dutch Medical Research Involving Human Subjects Act (WMO). All workshop participants were informed about the study objectives and procedures and provided informed consent to participate. Participation was voluntary and unpaid. Travel expenses were reimbursed where applicable. Participants were informed that workshop reports would be made publicly available, and any attribution of names or roles in those reports occurred with participants' knowledge and consent.

Results

Included Literature

The results of the literature screening are presented in a PRISMA flow diagram (Figure 1). In total, 25 articles met the inclusion criteria. Most were commentary-style publications (eg, perspectives and debates; n=14), alongside review-type articles (n=8) and a smaller number of qualitative studies (n=3). In terms of medical focus, the largest group of articles explored themes within general health care and clinical informatics (n=9). Other represented areas included medical ethics and health equity (n=5), primary care (n=3), public health and health disparities (n=2), geriatrics (n=1), psychiatry (n=1), nephrology (n=1), rheumatology (n=1), pediatrics (n=1), and orthopedics (n=1).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) flow diagram of study selection. AI: artificial intelligence.

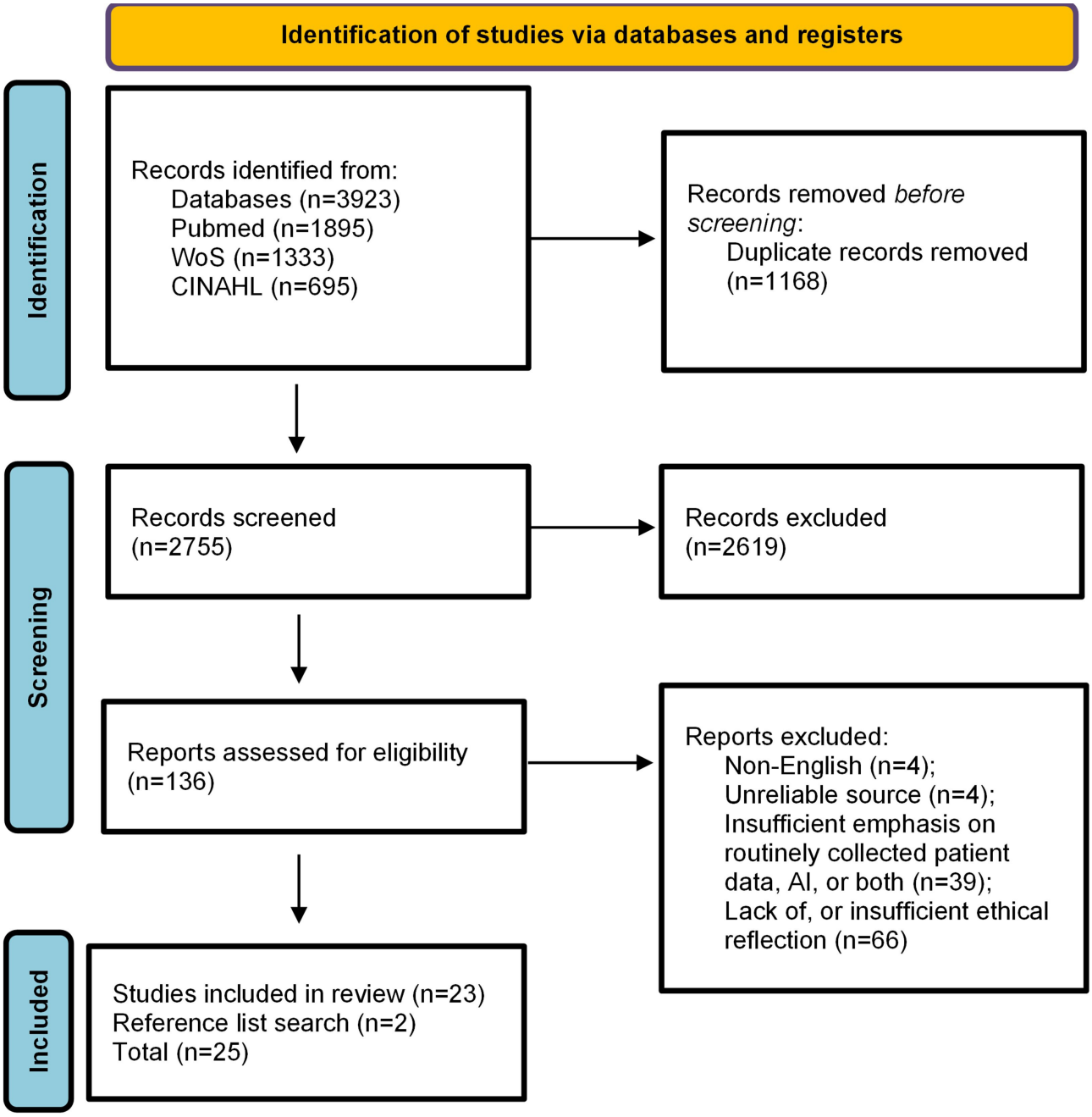

In terms of geographical distribution, the majority of articles were published by authors based in North America (n=14), followed by Europe (n=8), Asia (n=2), and Australia (n=1). A detailed overview of article characteristics is available in Multimedia Appendix 4 [22-46].

Workshop Participants

Two stakeholder workshops were held, with 17 participants in the first and 13 in the second, with some attending both. Participants included medical ethicists (n=3), medical informatics experts and AI engineers (n=9), patients associated with the Kidney Patients Association Netherlands (NVN) (n=2), patient representatives from the NVN (n=2), a nephrologist, a chief medical information officer (n=1), a chief scientific information officer and endocrinologist (n=1), a general practitioner (n=1), pharmacists (n=2), a digital transition manager, data protection officers (n=2), and a digital ethics manager.

Key Themes From the Literature and Integration of Stakeholder Workshops

This section is structured around 4 key themes: data privacy, transparency, and consent; public trust and regulatory challenges; fair representation and model generalizability; and responsible AI integration in clinical practice. This structure reflects the staged design of the LEAPfROG project and the cyclical nature of AI development in health care, which typically extends across data governance and quality, model generalizability, and clinical implementation, with each dimension influencing and being influenced by the others across policy, data, AI, and practice. The themes raised during the second stakeholder session aligned with the coding scheme

XSL•FO
RenderX

developed from the scoping review, reinforcing the relevance of the literature to stakeholder perspectives. The workshops played a crucial role in grounding the analysis in real-world experience, ensuring its practical relevance within the LEAPfROG project.

To illustrate how thematic emphasis shifted across the different study components, Table 1 presents the staged development of these themes across the scoping review and stakeholder workshops.

Table . Shifts in thematic emphasis across study components of the staged LEAPfROG (Leveraging Real-World Data to Optimize Pharmacotherapy Outcomes in Multimorbid Patients Using Machine Learning and Knowledge Representation Methods) design.

Key themes / stage	Workshop 1 (EHR[a] data reuse)	Scoping review (EHR–AI[b] Integration)	Workshop 2 (AI development and implementation)
1. Data privacy, transparency, and consent	• Patient autonomy and pseudonymization challenges • Clear consent for secondary use of EHR data	• Privacy risks amplified by data linkage and free-text inputs; persistent reidentification vulnerabilities • Tensions between ethically robust consent and notions of a civic duty to share EHR data	• Fostering trust through transparency across AI model development, explainability, and outputs
2. Public trust and regulatory challenges	• Institutional fragmentation and suboptimal national EHR infrastructure • Inconsistent policy and legal interpretations among stakeholders (including data custodians and researchers)	• Siloed governance and fragmented oversight • Regulatory uncertainty limiting data sharing and accountability	• Stakeholder-led, value-driven governance • Shared vision on responsible EHR data reuse and multistakeholder collaboration
3. Fair representation and model generalizability	• Data heterogeneity and variability in EHR documentation practices • Digital and social inequities shaping clinicians' use of EHRs and patients' access to their health data	• Missing or biased data reinforcing inequities • Lack of demographic diversity and transparency in reporting	• Variation in clinical practices (eg, workflow diversity) shaping EHR data patterns and subsequent model outputs and interpretation • Need to recontextualize EHR-derived data across hospital departments and care settings
4. AI integration in clinical practice	• Documentation burden and workflow diversity in CKD[c]/DAKI[d] care • Impact on EHR data reliability and clinician workload	• Model opacity, "clinical tropism" • Risk of overreliance and clinician deskilling • Ethical need for interpretability and accountability	• Timing and interpretability of AI model outputs • Risk of overly general or narrow outputs with limited clinical relevance • Model could enhance clinicians' understanding of drug side effects and support shared decision-making

[a]EHR: electronic health record.

[b]AI: artificial intelligence.

[c]CKD: chronic kidney disease.

[d]DAKI: drug-induced acute kidney injury.

We connected insights from the literature and stakeholder discussions to situate the key themes within the Dutch health care landscape and the specific context of the LEAPfROG project. Each of the following sections elaborates on one of the 4 key themes and concludes with a reflection on how that theme was grounded and contextualized through stakeholder discussions.

Data Privacy, Transparency, and Consent

Privacy Risks and Security Challenges

The reviewed literature highlights growing privacy concerns around the reuse of EHR data for AI research [22-34]. EHR data originate from everyday clinical care and often contain sensitive personal and medical information recorded without explicit patient consent, raising concerns about misuse and breaches of confidentiality [30]. In contrast, data from controlled research settings such as randomized trials follow predefined consent procedures and strict regulatory oversight.

Safeguarding privacy in this context requires clear data access agreements, secure systems for storage and transmission, and transparent mechanisms for patient control and consent, in line with legal and ethical standards. The risks described in the literature include security breaches, such as hacking or malware attacks, that could lead to data exploitation, intentional discrimination, or identity theft [27,31,35]. Additionally, combining multiple data sources, such as EHRs and disease

XSL•FO
RenderX

registries, amplifies privacy risks through greater exposure to breaches, inconsistencies, and loss of data integrity [23,27,28,30,34,36]. Advances in AI have further intensified the risk of misuse and harm by enabling sophisticated reidentification, tracking, and profiling techniques that threaten patient confidentiality, including methods capable of inferring sensitive information from nonsensitive data. These developments heighten the risk of misuse and harm [28,35]. Even privacy-preserving approaches remain vulnerable to inference attacks that can reveal attributes of training data, posing ongoing risks to confidentiality [35].

Moreover, free text fields in EHRs often include both medical and personal information. Persisting difficulties in deidentifying this information increase the potential for reidentification and unintended disclosure when used for research purposes, possibly even beyond the individual patient [27]. Ford et al [27] discuss the challenge of balancing automated and human-led deidentification of free text, noting that participants in citizen jury discussions expressed concerns that automated methods might leave identifiers, while human-led processes, on the other hand, could introduce bias or error due to inconsistency or workload. Others emphasize that the use of AI methods to extract information from unstructured EHR text often lacks standardized guidelines and the level of scrutiny applied to other medical interventions, further exacerbating privacy risks [33].

Informed Consent: Is There a Civic Duty to Share Data for AI Health Research?

Most included studies identify challenges related to the need for explicit patient consent in reusing EHR data for AI-driven research [23,25,27-31,33-35,37,38]. Patients are often unaware of how their data may be used over time, raising questions about the validity of prior or implicit consent given for clinical care or research. Evolving AI models may introduce new applications that extend beyond the original scope of consent, repurposing data in ways that patients could not have anticipated [31,38]. These concerns intensify when multiple data sources are linked, making it difficult to ensure meaningful consent at each stage of data integration [22,23,38,39]. Without clear processes for obtaining and maintaining consent, AI-driven research risks undermining patient autonomy and trust in health data governance, compromising its moral legitimacy.

Among the included studies, Müller critically examines the view that citizens have a moral duty to share their health data for medical AI development. He argues against this by outlining key contrasts between common assumptions underlying such a duty and the ethical realities of data sharing [28]. First, the "rule to rescue," which asserts a duty to help others in immediate danger, does not apply here, as data sharing lacks the direct, person-to-person moral obligation implied in such situations. Second, the "low risks, high benefits" argument claims that societal gains justify data sharing. He argues that this view overlooks how risks such as privacy breaches and erosion of trust in health care are unequally distributed and more immediate than potential benefits. Third, the "property rights argument" holds that the involvement of public institutions in generating health data justifies limiting individual ownership. Müller disputes this, emphasizing that health data are inherently personal and sensitive, and that treating them as a public commodity undermines autonomy and confidentiality.

Instead of a universal moral duty, Müller proposes a civic responsibility approach grounded in transparency, value alignment, and voluntary participation, emphasizing consent mechanisms that allow individuals to manage how their data are shared. In line with this perspective, other authors argue that even when data reuse does not legally require consent, researchers remain ethically obligated to communicate clearly how data are used, how AI models operate, and how decisions are reached, as legal compliance does not substitute for transparency or patient and public engagement [33].

Stakeholder Perspectives: Data Privacy, Transparency, and Consent

In line with the literature, the stakeholder group discussed potential risks to patient privacy in the reuse of EHR data for research. They identified data exchange and EHR systems linkages as particularly vulnerable points, along with the risk of unauthorized data use by third parties such as health care insurers and private companies. To address these concerns, stakeholders proposed several measures, including the use of virtual research data environments that strictly comply with data protection regulations, ensuring that research data are pseudonymized or anonymized to reduce the risk of misuse or breaches, and enabling data providers to audit researchers' actions on the data.

Other suggestions involved giving patients greater control over how their EHR data is reused, including the ability to decide whether, and which parts of, their records should be used for research. While this approach could enhance patient autonomy, concerns were raised about the risk of overburdening patients with repeated consent requests, especially when the research offers no clear or direct benefit at the individual level and does not immediately reflect patient priorities.

Public Trust and Regulatory Challenges

Trust and transparency

The majority of the studies highlight trust as crucial in the relationship between various publics and those who collect, control, access, and use EHR data for AI research [23-25,27,29,31-36,38,40-42]. Many studies identify low public awareness as a key factor undermining trust in these practices [24,27,32,36]. Patients are often unaware that their data are shared across organizations within integrated systems or for what specific purposes [31]. Low public trust in institutions handling health data may reduce participation in AI research, ultimately introducing self-selection bias and weakening model quality and representativeness [28,31,34]. Given these concerns, many studies underscore the ethical importance of transparency, not only as a means to foster trustworthiness but also for ensuring legitimate and accountable data governance [22,24,25,27-34,38,40,41].

Regulatory fragmentation and legal uncertainty

Several studies identify regulatory and legal uncertainties as major obstacles to the adoption of health AI tools [30,41,42]. A fragmented legal landscape, marked by inconsistencies across

XSL•FO
RenderX

institutions and jurisdictions, creates compliance and accountability challenges [29,31,33]. In Germany, for example, data-sharing initiatives are complicated by regulatory fragmentation and varying requirements across sectors [29]. Siloed regulation, in which separate and uncoordinated frameworks govern different types of health data and medical technologies, fragments oversight, leading to regulatory gaps (areas not clearly covered by any authority), overlaps (conflicting or duplicated requirements), and broader compliance hurdles that hinder implementation [33,38]. Regulatory uncertainty often deters data custodians—responsible for data storage, integrity, and authorized access—from sharing EHR data beyond clinical settings. This reflects a broader tension between oversight predominantly aimed at safeguarding privacy and fostering the flexibility required for AI innovation. For instance, strict interpretations of data protection laws, such as the General Data Protection Regulation, can limit researchers' access to patient data, influencing the trajectory of developments in the field [30,38].

Stakeholder Workshops: Public Trust and Regulatory Challenges

In the Netherlands, EHR data exchange remains limited due to the fragmented health care system, where institutions rely on diverse and often incompatible standards. For example, hospitals may use different clinical coding systems or inconsistent data formats for diagnoses, making integration across systems technically and semantically challenging. This challenge is further compounded by a decentralized governance framework that grants considerable institutional autonomy, resulting in varied requirements and interpretations across institutional policies. In line with the literature, stakeholders noted that unclear and inconsistent interpretations of laws and regulations create barriers to collaboration within a complex network of actors, including hospitals, health insurers, data vendors, EHR system providers, and research funders, each imposing specific conditions for data availability and reuse.

A key suggestion that emerged during the workshops was to develop a shared vision within collaborative networks on the "why" of reusing EHR data for research. This could be achieved, for example, through consensus workshops involving patients, health care providers, researchers, data managers, and data protection officers, with attention to values such as trustworthiness, solidarity, quality of care, and privacy. Participants also emphasized the importance of establishing shared agreements within these networks on which data should be recorded in the EHR, using predefined terminologies, and defining how and where these should be implemented.

Fair Representation and Model Generalizability

Health Inequities Reflected in the Data

Most included studies warn that missing or incomplete patient data can compromise model performance and perpetuate existing health care inequities, disproportionately affecting marginalized groups [22-25,28,30,31,33,37,39-46]. These inequities often stem from social determinants of health, such as financial barriers, limited digital and health literacy, and structural marginalization. Failing to adequately consider these factors can result in flawed models that underestimate disease burden for certain groups and undermine the equitable distribution of benefits [25,36,37,39,42-44]. Several studies further highlight the need to account for the historical context of research on race and ethnicity, ancestry, and sex and gender minorities, particularly in the United States, where past exploitation contributes to ongoing distrust in health care systems [40,42].

Institutional and Structural Variability in Data and Its Impact on AI Data Quality

Several studies highlight how institution-specific practices, norms, and patient populations shape the composition of EHR data [22,24,25,28-33,35,37,40,42-46]. Factors such as patient proximity to clinics, visit frequency, insurance coverage, and differences in clinical protocols and diagnostic resources influence data representation [22,24,25]. For example, some institutions may request diagnostic tests earlier [30] or use activity-based financing, leading to an overrepresentation of codes tied to higher reimbursement, reflecting billing priorities rather than actual clinical complexity [22,25]. Sampling bias can occur when chronically ill patients, who interact more frequently with the health care system, are overrepresented, skewing patterns that do not reflect the broader population. Within institutions, departmental priorities and documentation practices may differ, contributing to inconsistencies in data capture [25]. Increasing reliance on data-intensive health care can reduce direct communication between medical staff, heightening the risk of incomplete or selective documentation [24].

While standardizing EHR data through structured formats and predefined categories may reduce inconsistencies, it also risks removing essential clinical context [27,34,36,44]. Excluding physicians' notes and clinical judgments, for example, may lead to the loss of valuable information such as symptom descriptions, diagnostic reasoning, and the context preceding symptoms. This increases the likelihood of misinterpretation by researchers and AI models, potentially resulting in biased or unsafe decisions [27,30,36,44,46]. Yet, as Knevel and Liao [30] note, some gaps in EHR data can hold clinical meaning. For instance, the absence of specific test results may suggest that a particular diagnosis was deemed unlikely. Interpreting such nuances requires strong domain knowledge and clinician involvement to ensure meaningful analysis.

Variability in EHR systems across institutions and countries continues to challenge the development of generalizable AI models [30]. The absence of centralized or standardized data-sharing mechanisms, such as those based on findability, accessibility, interoperability, and reusability principles, exacerbates this problem by hindering interoperability and seamless data integration across settings [23,33,34,36,38]. To address structural and institutional fragmentation, federated learning has been proposed as a promising approach aimed at preserving patient privacy while enabling AI models to be trained across institutions without sharing raw data. Such methods seek to facilitate collaboration across decentralized datasets while maintaining institutional autonomy. Related privacy-preserving techniques include differential privacy and homomorphic encryption, which are designed to enhance data

XSL•FO
RenderX

protection but remain largely experimental and face challenges of scalability, performance, and validation in clinical settings [35].

Challenges in AI Model Reliability and Potential Harm

Several studies raise concerns about harm from unreliable or biased AI systems, particularly when embedded in clinical decision-making. Deliberate and unintentional choices in AI development can introduce algorithmic bias, reinforcing EHR data limitations and distorting model outputs [22,24,25,27,28,30-33,35,37,42,43,45]. Examples include data integrity issues from reusing and merging databases, such as data duplication [24,35], model design choices reflecting provider bias [32], and preselection of training data based on physician or developer preferences [28,46].

Technical flaws throughout model development can also result in harmful clinical outcomes, even when using high-quality data [33,37,43]. For example, models may overfit datasets that do not reflect relevant clinical questions, undermining generalizability and reliability [33,37,43]. Optimistic predictions can lead to futile or even harmful interventions, while overly pessimistic ones may cause unnecessary withholding of care [32]. In such cases, the ethical imperative to minimize harm may justify simpler models that generalize more safely, even if less accurate. Some studies argue that the appropriate level of model explainability should be calibrated according to the clinical risks and the extent to which domain experts can interpret the model's output [24,33,37,46]. Several studies also highlight the lack of standardized frameworks or tools to assess and evaluate AI's real-world clinical impact, particularly regarding patient safety, value alignment, and health equity, thereby limiting consistent assessment of ethical implications and potential harm [25,26,29-34,38,40,43-45].

Barriers to Inclusive AI Development

Several studies identify deeper systemic and infrastructural barriers that undermine inclusive and equitable AI. Adapting AI to local environments requires significant investment in data integration, testing, infrastructure, and clinical expertise [22,30]. However, many health organizations lack the secure, scalable infrastructure needed for effective AI deployment, a gap often overlooked in policies that emphasize AI's benefits while downplaying resource demands [22]. These barriers also raise questions of equity in data governance, including whether and how data contributors should be compensated or share in AI's benefits [31].

Another persistent challenge is the unclear definition of target populations and the inconsistent representation of demographic groups in AI development [22,25,28,41,45]. One study found, for example, inconsistent demographic reporting across 164 EHR-based AI studies, undermining model representativeness and reproducibility [45]. Limited data sharing, closed-source models, and poor interoperability hinder validation across diverse clinical settings, while low digital adoption and data vendor monopolies further restrict datasets to less representative populations [25,28,29,32-34,36,38,41,42,45].

Paulus and Kent [42] argue that fairness in algorithmic decision-making is context-dependent, requiring value judgments and stakeholder consensus. No universal definition exists, and ethical concerns over protected attributes such as race persist. Others similarly highlight concerns about the values and priorities embedded in AI models, which may implicitly reflect biases or unexamined assumptions [44]. This risk intensifies when health data is used for generating profit, for example, potentially placing financial incentives above equitable patient care and responsible data stewardship [30,31,33,34]. In response, many studies advocate for sustained stakeholder engagement throughout AI development efforts, including participatory co-design and mixed methods approaches that integrate qualitative and quantitative insights rather than relying on externally imposed solutions [22-25,27,28,32,33,37,40-44,46].

Stakeholder Workshops: Fair Representation and Model Generalizability

Stakeholder discussions highlighted the complex realities of clinical practice in the context of CKD and DAKI, which are inherently mirrored in the data. For instance, when treating sepsis with antibiotics such as vancomycin, known to potentially cause acute kidney injury (AKI), clinicians may consciously accept the risk of AKI as a necessary trade-off to, in some cases, save the patient's life. Additionally, when a side effect is common and widely acknowledged, physicians may not document or provide less detailed documentation, focusing instead on more pressing clinical problems. Moreover, these considerations can also vary across medical specialties, with clinical practices such as prescribing practices differing, for example, between internal medicine and cardiology.

In short, understanding a potential causal relationship between a drug and AKI provides only part of the story. During the workshops, it was emphasized that gaining a clear understanding of how variability from different clinical settings is represented in the EHR data is crucial, as is considering how such complexity could be effectively incorporated into an AI-driven decision-support model for retrospective DAKI diagnosis. At the same time, consistent with Knevel and Liao [30], it was noted that research using EHR data can offer valuable insights into clinical routines and decision-making patterns, such as differences in prescribing practices across specialties. However, uncovering these insights is highly resource-intensive and time-consuming, as it requires not only technical expertise but also deep contextual knowledge and sustained engagement with clinical practice.

Responsible AI Integration in Clinical Practice

Administrative Workload

While efforts to structure and standardize EHR data are central to enabling AI-driven research, several studies emphasize the unintended increase in administrative burden for health care providers [24,33,35-38,40,41,44]. Processes such as careful data assessment and logging, essential for ensuring data quality, reducing fragmentation, and addressing issues like concept drift in longitudinal datasets, are time-consuming and resource-intensive [33,37,41]. These demands raise ethical concerns related to clinician well-being, sustainable work practices, and their impact on patient care [24,29,36,41].

XSL•FO
RenderX

Furthermore, several studies note that professionals' earlier experiences with EHR implementations often failed to deliver the anticipated quality improvements despite significant increases in workload associated with data entry and system management, contributing to skepticism toward data-driven AI solutions [24,29,41]. Finally, many health care settings still lack fully digitized, structured records, complicating the integration of AI into clinical workflows [41]. Together, these challenges highlight a tension between the technical demands of fit-for-purpose data in AI-based research and the practical realities of clinical work, underscoring the need for implementation strategies that balance data quality with provider sustainability.

Clinical Tropism

A major barrier to AI adoption in health care is the limited transparency and interpretability of complex, "black-box," models [22,24-28,30-34,36-41,43-46]. When AI-driven decisions lack clear explanations, clinicians and patients struggle to trust them, undermining ethical requirements for transparency, informed decision-making, and accountability in clinical care. Without insight into how these models function, users cannot detect or correct biases, increasing the risk of erroneous or harmful outcomes [22,39,41]. Distinguishing causal from correlative relationships is essential in medical decision-making to avoid ineffective or harmful interventions [28,39]. These risks are further amplified by the fact that harmful AI-related errors can occur at scale, exceeding the impact of individual provider errors [22].

As discussed earlier, many studies highlight challenges of representativeness and model generalizability, emphasizing how incomplete or biased datasets can perpetuate existing inequalities. Beyond these population-level concerns, several studies draw attention to how AI systems mirror the clinical contexts in which they are developed. Alami et al [22] describe this as clinical tropism, referring to the tendency of AI systems to reproduce the specific practices, routines, and priorities of the settings that generated their training data. They argue that models trained on localized data may reflect institutional protocols, workflows, or device infrastructures that are not easily transferable to other contexts. Over time, such systems risk becoming repositories of established clinical practices, developing what Alami et al [22] term a "clinical mind," rather than generating new insights. This dynamic may lead professionals to place disproportionate trust in AI outputs that merely echo existing patterns of care. Furthermore, because clinical practice continually evolves through new treatments and protocol updates, EHRs are inherently temporal, making input data prone to diverge from the contexts represented in their underlying sources [22].

Overreliance and the Risk of Deskilling

Several studies highlight the risk of deskilling among medical professionals as AI adoption increases, potentially weakening clinical judgment, autonomy, and the ability to make independent decisions [22,25,26,29-32,37-39,41,43]. Overreliance on AI may also reduce adaptability across clinical contexts and limit clinicians' capacity to account for patient preferences and values, raising concerns about the long-term impact on clinical competency and patient care [22,25,26,32,37,43].

Empirical evidence suggests that caregivers (including nurses, midwives, clinicians, and pharmacists) are generally less likely to question algorithm-driven diagnostic results, increasing the error risk and further eroding clinical judgment over time [22,25,28,37]. Conversely, when AI-generated results deviate considerably from a clinician's assessment, there is a risk that the model's recommendations will be dismissed outright, potentially limiting AI's practical value in clinical settings [22,43]. These dynamics underscore the need for careful integration of AI into clinical workflows, with sustained attention to its effects on professional roles, decision-making, and oversight.

Stakeholder Workshops: Responsible AI Integration in Clinical Practice

In the complex care setting of CKD and DAKI, precise documentation of medication use and adverse drug events is essential for generating reliable data to support AI-driven research. However, DAKI is difficult to diagnose because of its multifactorial causes, subtle onset, and patient-specific variability, often involving combinations of dehydration, sepsis, and drug toxicity. During stakeholder workshops, participants stressed that documenting this complexity is time-consuming. One general practitioner noted that recording side effects alone can take up to 30 minutes, a significant burden given current time pressures in Dutch health care. Although general practitioners are legally required to provide digital access to medical records upon request, they are not obliged to use an EHR system. One kidney patient shared that her general practitioner still prefers pen and paper, relying on digital systems only when necessary, illustrating how variability in documentation practices continues to limit the availability of structured data for AI.

Participants saw potential for the model to have a learning effect, improving clinicians' understanding of drug side effects over time. This could enhance care quality and, if communicated clearly, support shared decision-making. However, they stressed that model outputs must always be evaluated alongside clinical expertise to ensure safe and appropriate care. To improve clinical relevance, they suggested linking outputs to relevant guidelines and scientific literature to help physicians interpret findings in context and assess their significance for individual patients. At the same time, they cautioned that the time required to interpret outputs might increase or shift clinical workload rather than reduce it.

Stakeholders raised concerns consistent with the concept of "clinical tropism" of Alami et al [22], which describes how AI models may reproduce the narrow or context-specific knowledge embedded in the data on which they are trained. They noted that this tendency could be reinforced by the steering effect of clinical guidelines and institutional protocols that shape both data collection and model performance. Participants further noted that publication bias in the studies informing model development can compound these effects. When only positive or statistically significant results are published, the evidence base fails to reflect the full range of clinical outcomes, leading

XSL•FO
RenderX

models to learn a distorted view of clinical reality, thereby reinforcing rather than mitigating clinical tropism. Together, publication bias and the standardizing influence of clinical guidelines narrow the evidentiary foundation on which AI systems are built, privileging typical or well-studied cases while obscuring clinical variability and local practice. Participants also cautioned that integrating diverse data sources may produce overly general outputs with limited clinical relevance, lacking actionable value for clinicians and therefore likely to be disregarded.

Finally, participants stressed that assessing the causes of AKI requires attention to factors beyond medications and careful interpretation of predicted probabilities within individual patient contexts. For instance, a 61% probability that an antibiotic caused AKI might be clinically insignificant in one case but highly relevant in another, depending on the patient's circumstances. Participants emphasized that the model should be developed and implemented primarily as a decision-support tool to help clinicians assess whether a drug likely contributed to AKI.

Discussion

Principal Findings

This study used a mixed methods approach to explore the ethical considerations of using EHR data for AI-driven health research and innovation, with the LEAPfROG project serving as a guiding case. We combined a scoping literature review, a systematic search, and 2 stakeholder engagement workshops, all informed by the GEA [17].

Previous studies have examined the reuse of EHRs and the ethical implications of AI development separately. Research on data reuse has addressed governance, consent, and public trust, showing how challenges of transparency, fairness, and oversight are amplified by the data's variable quality, uncertain provenance, and fragmented governance [12-14]. Studies on medical AI have focused on fairness, transparency, and accountability, mapping the ethical and epistemic risks that arise within algorithmic systems [15,16]. Our study brings these strands together by examining how recurring ethical challenges emerge and take shape at the intersection of EHR data and AI development, while also situating these issues within a real-world clinical and institutional setting (the LEAPfROG project).

By integrating stakeholder perspectives with existing ethical analyses, this study contributes an empirically grounded account of how recurring ethical challenges emerge in practice when AI systems are developed using EHR data. In this context, issues of consent, privacy, and governance become more complex as EHR data is repurposed and leveraged using AI, while the opacity and scale of AI systems introduce new risks of decontextualization and fragmented accountability across policy and oversight domains.

Building on this analysis, we advocate for an approach to data governance and AI development that is centered on and led by stakeholders. We stress the importance of involving all relevant parties not only in evaluating benefits and risks, but also in shaping the goals and direction of AI innovation. Accordingly, we structure the following discussion as follows: first, we present a set of practical recommendations derived from the literature and stakeholder workshops, intended to guide future reflection, engagement, and implementation. We then explore the concept of data work and examine the ethical risks associated with removing data from its original context and reusing it for AI development. Finally, we review current practices in stakeholder engagement and highlight the need for more inclusive and participatory forms of governance in future AI initiatives.

Practical Recommendations

In addition to the central themes, both the scoping review and stakeholder workshops yielded practical recommendations for addressing ethical concerns related to AI and the use of EHRs (Table 2). Although these action points were not the primary focus of this study, they offer opportunities and potential entry points for addressing the broader challenges we aimed to explore. These insights reflect the perspectives and priorities identified in the sources consulted and do not necessarily represent the position of the author group. As the findings draw on international literature, some recommendations, such as patient representation in ethics committees, may already be established in contexts like the Netherlands. Nevertheless, they offer a useful supplement for future reflection, engagement, and implementation.

XSL•FO
RenderX

Table . Key recommendations from literature and stakeholder workshops.

Key themes and recommendations	Main actors	Source (key articles)
Data privacy, transparency, and consent		
Strengthen AI[a] and data literacy among patients, clinicians, and institutional review board (IRB) members, emphasizing how EHR[b] quality and AI training may affect privacy, consent, and data integrity.	Board of Directors of health care organizations	[38,41]+ workshops
Implement dynamic consent mechanisms that reflect the iterative reuse of EHR data in AI development and ensure institutional governance structures can accommodate these updates.	Board of Directors of health care organizations	[27,28,31]
Integrate consent management tools into EHR interfaces so that patients can review and modify how their data is used in research and model training, with clinical staff facilitating informed communication.	Board of Directors of health care organizations and EHR system vendors	[25,31]+ workshops
Require transparent documentation of AI provenance, including funding sources, proprietary model components, and training data, to prevent hidden conflicts of interest and foster public trust.	Researchers involved in AI projects like the LEAPfROG[c] project	[23,25,29]
Public trust and regulatory challenges		
Adopt standardized, FAIR-aligned[d] data practices and establish cross-institutional agreements that clarify ownership, accountability, and data flow between hospitals and general practices.	National research infrastructure organizations like Health-RI in the Netherlands	[33,34,36,40]+ workshops
Establish joint oversight frameworks that include clinicians, data stewards, developers, and regulators to coordinate decision-making on data use, ownership, and third-party involvement, reducing fragmented accountability.	National research infrastructure organizations like Health-RI in the Netherlands	[23,25,29,31]+ workshops
Embed patient representatives (ideally trained in AI ethics and data governance) in ethics committees and regulatory boards to strengthen inclusiveness and transparency in oversight processes.	National research infrastructure organizations like Health-RI in the Netherlands and Board of Directors of health care organizations	[25,38,41].
Fair representation and model generalizability		
Validate AI models in real clinical environments (for example, CKD[e] and AKI[f] contexts) and across diverse patient subgroups to ensure representativeness and clinical relevance.	Researchers involved in AI projects like the LEAPfROG project	[37,42,43,46]
Conduct fairness audits to explicitly assess whether model outputs are biased by demographic differences or by specific clinical routines and documentation practices that vary across institutions.	Researchers involved in AI projects like the LEAPfROG project	[32,40,42,43]
Adopt mixed methods evaluation combining quantitative metrics with qualitative insights to understand how AI affects workflows, professional judgment, and patient experience.	Researchers involved in AI projects like the LEAPfROG project	[30,40,42,43]+ workshop 2
Implement continuous postdeployment monitoring: reassess clinical impact on patient outcomes over time to ensure alignment with evolving clinical practice and technologies.	Board of Directors of health care organizations	[25,38,46]+ workshop 2

XSL•FO
RenderX

Key themes and recommendations	Main actors	Source (key articles)
Appraise both clinical and economic outcomes through frameworks that evaluate how AI contributes to care quality, efficiency, and sustainability of health care delivery.	Researchers involved in AI projects like the LEAPfROG project	[22,39,46]
Responsible AI integration in clinical practice		
Align AI development with clinical workflows, prioritizing usability and practical relevance over purely technical performance metrics.	Researchers involved in AI projects like the LEAPfROG project	[22,24]+ workshops
Ensure interpretability supports clinical judgment by designing models that reflect clinical complexity, communicate uncertainty, and complement human expertise.	Researchers involved in AI projects like the LEAPfROG project	[22,26,37]+ workshop 2
Develop learning-oriented AI systems that enhance, rather than erode, professional expertise, and support reflective clinical practice.	Researchers involved in AI projects like the LEAPfROG project	Workshop 2
Design EHR/AI tools collaboratively with stakeholders throughout the data and model lifecycle to clarify purposes, risks, and benefits, aligning these with shared values and patient expectations.	Researchers involved in AI projects like the LEAPfROG project and Board of Directors of health care organizations	[25,33,38,41]+ workshop 1

[a]AI: artificial intelligence.

[b]EHR: electronic health record.

[c]LEAPfROG: Leveraging Real-World Data to Optimize Pharmacotherapy Outcomes in Multimorbid Patients Using Machine Learning and Knowledge Representation Methods.

[d]FAIR: findability, accessibility, interoperability, and reusability.

[e]CKD: chronic kidney disease.

[f]AKI: acute kidney injury.

Making Data Work

While much of the public and academic debate on AI in health care focuses on model explainability or algorithmic opacity, these discussions often overlook a more foundational issue: the nature of the data on which AI depends. Earlier studies on health data reuse have shown [12-14] that questions of trust, consent, and governance already hinge on how data is produced, managed, and interpreted. Our findings suggest that without a deeper understanding of how health data is actually generated, structured, and used in practice, efforts to build fair, accurate, and trustworthy AI systems will remain incomplete. Data is rarely a neutral input, as it is deeply shaped by clinical routines, institutional diversity, and social context. Documentation practices, system design, and socio-ethical expectations determine what gets recorded and how it is organized. Recognizing this "data work" is not merely a technical concern but an ethical one, since it determines whose experiences are recorded and how decisions are justified.

Our findings reveal the complexity and effort involved in making EHR data fit for AI-driven research. Both the reviewed literature and stakeholder insights point to inconsistencies in EHR documentation, variability in clinical workflows, and the labor-intensive nature of data preparation and standardization. Such challenges underscore that data is produced through ongoing, situated work rather than passively collected. This aligns with the literature review on data-related activities by Bertelsen et al [47], or "data work," which emphasizes its multifaceted and situated nature. They group data work into 3 interrelated categories: data collection (eg, capturing, discovering, requesting, and self-tracking), data production (eg, coding, entry, and digitization), and data use and sharing (eg, providing, exchanging, and disseminating).

As data move through different stages and are reused for AI training, they often lose the clinical and institutional contexts that once gave them meaning. This process, referred to here as decontextualization, involves detaching data from the social and professional environments in which they were produced, allowing them to be reinterpreted, or even misinterpreted. Hoeyer [48] describes this as a separation between the "data work of production" and the "data work of analysis," where separation from the original context renders data fragile and open to misinterpretation. In AI development, this dynamic has ethical as well as epistemic consequences: models may treat context-dependent judgments or incomplete entries as objective facts, obscuring the human and institutional processes that shaped them. This dynamic is also reflected in the concept of "broken data," which highlights the ongoing, often hidden repair and improvisation required to keep data fit for purpose [49], as well as in studies showing how context loss invites interpretive uncertainty, with trust, visual design, and perceived credibility shaping how data is understood [50].

When AI systems trained on such data are applied in clinical settings, the same loss of context can reappear, shaping how models interact with the environments that produced their data and often reinforcing the very patterns they were meant to improve. Stakeholders noted that AI systems may mirror existing

XSL•FO
RenderX

prescribing and documentation practices, reproducing the same routines they aim to change. This recursive dynamic reflects what emerged from the literature as clinical tropism, where models embed themselves in the patterns most strongly represented in their training data and in the settings where those patterns persist. When that influence begins to shape the outcomes it predicts, it can result in a self-fulfilling prophecy [51]. In such cases, models appear accurate precisely because their use helps make their predictions true. Clinical tropism creates the conditions, while the self-fulfilling prophecy describes the result. Together, they show how decontextualization can evolve from an issue within data to one that reshapes clinical practice, as models begin to reproduce the very patterns they were designed to improve. These concerns are mirrored in debates on explainable AI, where current approaches have been criticized for creating a false sense of understanding and trust, offering surface descriptions of model behavior without clarifying whether decisions are reasonable or justified [52]. Rather than mitigating bias, such transparency can obscure how context loss shapes both model development and evaluation.

Extending the discussion of decontextualization, Hoeyer [48] draws attention to a deeper issue often overlooked in discussions of data work: the ways in which data are interpreted and mobilized in pursuit of narratives that are often implicit and conflicting. He argues that datafication fragments patient information into pieces that can be recombined to serve diverse aims, such as clinical care, research, governance, or performance monitoring. In this process, the patient may be obscured, and data detached from the context of its production. Clinicians may document with specific knowledge and intent, while analysts or policymakers may act on institutional logics shaped by different priorities. These tensions are not resolved through technical solutions alone but often reappear, in new forms, within the very tools developed to address them. For instance, even privacy-preserving approaches such as federated learning, which are designed to mitigate data sharing risks, can inadvertently deepen decontextualization by abstracting data use from the institutional and clinical settings in which they originate [35,53]. By enabling model training across institutions without sharing raw data, these methods aim to enhance privacy but may nevertheless reproduce institutional differences and context loss, introducing additional ethical concerns that extend beyond the scope of this paper but are highlighted elsewhere [53].

As is evident in the LEAPfROG project, AI initiatives involving EHR data span a diverse network of stakeholders, including health care providers, data custodians, EHR vendors, patients, and institutions, whose interests may not align. These misalignments are not incidental but intrinsic to how data is leveraged, representing structural and arguably unavoidable realities that must be acknowledged. Developing responsible AI, therefore, requires collaborations that explicitly recognize these dynamics and openly address how stakeholders themselves navigate and negotiate data use in practice, while articulating the competing aims, values, and uses of data within their own processes [47]. Doing so requires more than technical expertise: it depends on collaboration across domains to understand how clinical, organizational, and social contexts shape data and influence AI training and deployment. Together, these insights underscore that making data work is as much a social and ethical practice as a technical one.

Beyond Compliance: Meaningful Stakeholder Involvement

In current debates on health data governance, the public good is often invoked as a guiding principle, with values such as solidarity promoted as ethical foundations [54]. Prainsack et al [55], for instance, frame data solidarity as a means to promote justice, prevent harm, and ensure collective benefit. However, critics argue that such values, while well-intentioned, are often applied in top-down, unreflective ways that limit their ethical force [54,56,57] and risk overlooking deeper power asymmetries, such as the lack of patient representation in governance and decision-making structures. Rather than rejecting these principles, they have called for alternative models grounded in mutual aid and collective action, fostering more equitable relationships between data providers, users, and affected communities.

Taken together, these critiques reveal a broader structural issue: legal and policy alignment, while necessary, is not sufficient for the ethical use of health data or responsible AI development. As our results show, many studies point to the need for participatory, transparent governance that goes beyond compliance or rhetorical appeals to ethics. Concerns about fragmented oversight, unclear consent procedures, and opaque decision-making illustrate how legitimacy is undermined when stakeholders are not meaningfully involved. Without sustained, inclusive engagement, especially with those most directly affected, governance risks remain superficial and biased toward institutional priorities.

At a practical level, these structural shortcomings are reinforced by persistent power asymmetries in engagement processes, which risk prioritizing influential groups such as government officials, large nongovernmental organizations, or scientists [58]. These “elite” actors, with greater access to time, resources, and expertise, tend to set the terms of participation, sidelining less-resourced groups and limiting patient involvement. This is especially problematic in the context of CKD, where patients face complex treatments, uncertainty, and symptoms that can limit participation in research [59]. As a result, they are often underrepresented in studies that fail to reflect patient or caregiver priorities. Critical participation frameworks emphasize that such unequal conditions make engagement prone to tokenism, particularly when structural inequities like corporate influence or health care disparities remain unaddressed [59,60].

These dynamics become even more pronounced in the context of AI development. Here, technological optimism often frames innovation as inherently beneficial to patient care [61], which can obscure deeper ethical and social tensions. By emphasizing efficiency, accuracy, and progress, this framing tends to narrow debate to technical or policy issues, while sidelining questions of power and lived experience. Consequently, participation in AI development tends to be tightly controlled, serving institutional efficiency or public acceptance rather than fostering meaningful, value-driven dialogue. In this way, engagement

XSL•FO
RenderX

risks becoming instrumental, legitimizing AI initiatives rather than engaging with lived experience or long-term ethical and practical implications.

In light of these challenges, we were particularly mindful of how stakeholder engagement was approached in the LEAPfROG workshops. Recognizing the potential influence of technological optimism, we sought to create space for more critical, grounded conversations about the role of AI in health care. By grounding ethical reflections in concrete use cases and fostering open-ended, collaborative discussions, we aimed to avoid tokenism and reduce the influence of more resourced stakeholders. Our mixed methods approach, which included a literature review alongside stakeholder workshops, helped broaden the scope to address systemic issues such as health care inequities and the resource-intensive nature of AI development.

As an alternative to top-down regulation and rhetorical appeals to ethics, more participatory and inclusive governance models may offer a promising path forward. One such approach involves shifting toward stakeholder-led governance within consortia, supported by coproduced guidelines or codes of conduct. A key suggestion from our workshops was to develop a shared vision around the reuse of EHR data, including its purpose, risks, and benefits, through collaborative processes involving patients, clinicians, researchers, data managers, and data protection officers. Such efforts could also include agreements on what data should be recorded in EHRs and how. Rather than treating participation as a means to secure legitimacy for predetermined goals, these practices aim to embed values like trustworthiness, solidarity, and privacy into the infrastructure of data governance itself. By continuing our collaboration with the Dutch Kidney Association (NVN), we hope to support a model of engagement that centers patient perspectives, addresses power imbalances, and ultimately aligns data-driven innovation with the needs of those most affected.

Limitations

While time constraints, limited participant availability, and last-minute cancellations are common challenges in stakeholder workshops, a more critical limitation was ensuring balanced representation among stakeholders. Although the workshops were conducted in Dutch to promote inclusivity across diverse education levels and professional backgrounds, this choice inadvertently excluded non-Dutch-speaking participants, including some within the LEAPfROG consortium. Similarly, the scoping review included only English-language publications, potentially introducing language bias and omitting relevant studies in other languages. However, only 4 non-English publications were identified during screening, suggesting that such literature remains limited in this field.

In addition, the predominance of commentaries and reviews (14 out of 25) over empirical studies may limit the depth of evidence available to substantiate claims about real-world implementation, stakeholder experiences, or the practical implications. However, this imbalance also reflects the early and largely conceptual stage of ethical reflection at the intersection of AI and EHR data. Furthermore, many of the reviewed publications were authored by researchers without specialized legal or regulatory expertise, such as legal scholars deeply familiar with the GDPR (General Data Protection Regulation). This may have shaped how certain challenges were framed and further limited discussion of broader ethical issues, such as whether EHR data should be used for AI-driven research at all.

As mentioned above, our efforts to include diverse stakeholders may also have unintentionally reinforced power or knowledge imbalances, particularly due to the underrepresentation of patients. However, as described in the methods, patient participants were briefed in advance to support meaningful engagement. To enhance transparency and relevance, all workshop participants were also invited to review and comment on the workshop reports prior to publication. In addition, while policy perspectives were well represented in the workshops, participants with formal legal or regulatory expertise were underrepresented. This limitation may have influenced how governance and compliance challenges were discussed.

Furthermore, thematic convergence between the literature and the stakeholder workshop discussions may reflect the dominance of Global North perspectives, particularly those shaped by Dutch socioeconomic and institutional norms. While this alignment across methods strengthens the internal coherence of our findings, it also highlights a potential blind spot. Both the literature reviewed and the workshops conducted were situated within contexts that may not capture the unique challenges and priorities of the Global South. As others have noted, failing to account for these limitations risks overlooking perspectives that remain largely absent from this paper [62].

Beyond the institutional contexts examined here, large-scale health data initiatives (for example, the All of Us Research Program [63] and MIMIC-IV [64]) have likewise revealed persistent ethical challenges related to decontextualization, governance, and accountability. Similar questions are now emerging within Europe through the development of the European Health Data Space [54], which aims to harmonize secondary data use across member states. These examples demonstrate that the issues identified in this study re-emerge even in large-scale data initiatives with formal governance frameworks, suggesting that future research should further examine how context-specific governance practices can inform responsible data sharing at scale.

Empirical evaluation and the development of metrics to assess how the proposed ethical measures influence factors such as AI model fairness, data quality, or stakeholder trust fall outside the scope of this study. The goal of this work was to derive ethical measures through a novel approach that combines a scoping review with workshops based on the GEA. Future studies in other settings are needed to further develop and test evaluation frameworks that can operationalize these ethical measures in practice.

Our study deliberately focused on the Dutch and European legal and policy context to support LEAPfROG's goals. The transferability of our stakeholder workshop methods to other countries or institutional settings was beyond its scope and could be explored in future empirical research. Future work could also broaden these discussions by addressing themes such as data sovereignty, the ethical dimensions of data labor in the Global South, and the potential for data colonialism [62,65].

XSL•FO
RenderX

Conclusions

The use of EHR data in AI development is situated at the intersection of health care, data science, ethics, and policy, where competing priorities, values, and practices converge. Given this complex intersection, building responsible AI-based tools requires more than technical solutions and compliance. Our findings point to the need to begin with a critical understanding of the data, for which domain knowledge is essential. EHR data are not neutral but shaped by clinical routines, institutional constraints, and documentation practices. When removed from context, it is vulnerable to misinterpretation, which can cause AI systems to reproduce inequities and reinforce clinical patterns instead of supporting informed decision-making. Addressing these risks is essential to ensure accuracy and accountability. Our mixed methods approach, which included multistakeholder guidance ethics workshops, underscored the need for inclusive, value-sensitive development grounded in the realities of health care practice.

Acknowledgments

Special thanks to Daniël Tijink and Pieter van Kuilenburg for preparing and moderating the first guidance ethics workshop, and to Giovanni Cinà for his support in preparing the second workshop. We would also like to thank the participants of the workshops for their valuable contributions regarding the next steps for the LEAPfROG consortium, as well as the development of this paper.

Finally, the LEAPfROG consortium is a unique cross-sectional collaboration between patients, health care professionals, epidemiologists, medical informatics, and computer science researchers, as well as partners from the medical technology industry, regulatory bodies, and health care insurance. Principal investigator: Dr Joanna E Klopotowska. Work package (WP) leaders: Dr Ronald Cornet (WP1), Prof Dr Annette ten Teije (WP2), Dr Joanna E Klopotowska (WP3), and Dr Stephanie Medlock (WP4). Beneficiaries of the LEAPfROG consortium: Amsterdam University Medical Center, leadership and coordination; Vrije Universiteit Amsterdam, and Open University Heerlen. Ongoing LEAPfROG PhDs: Daniel Fernández-Llaneza, Romy M P Vos, Joris E Lieverse, and Menno Maris (Amsterdam University Medical Center), The LEAPfROG Consortium members include: Ameen Abu-Hanna (Amsterdam University Medical Center), Birgit A Damoiseaux (Amsterdam University Medical Center), Cornelis Boersma (Open Universiteit), Dave A Dongelmans (Nationale Intensive Care Evaluatie foundation), David H de Koning (Amsterdam University Medical Center), Frank van Harmelen (Vrije Universiteit Amsterdam), Gerty Holla (Amsterdam Economic Board), Heiralde Marck (Koninklijke Nederlandse Maatschappij ter bevordering der Pharmacie [KNMP] Geneesmiddelen Informatie Centrum), Iacopo Vagliano (Amsterdam University Medical Center), Jan Pander (AstraZeneca), Jurjen van der Schans (Open Universiteit), Giovanni Cinà (Amsterdam University Medical Center), Kitty J Jager (Amsterdam University Medical Center, European Renal Association-European Dialysis and Transplantation Association [ERA-EDTA] Registry), Leonora van Drop - Grandia (Z-Index), Linda Dusseljee-Peute (Amsterdam University Medical Center), Luuk B Hilbrands (Radboud University Medical Centre), Marcel SG Kwa (College ter Beoordeling van Geneesmiddelen), Marieke A R Bak (Amsterdam University Medical Center), Mariette van den Hoven (Amsterdam University Medical Centre), Martijn G Kersloot (Castor), Nicolette F de Keizer (Amsterdam University Medical Center), Otto R Maarsingh (Amsterdam University Medical Center), Paul Blank (NWO), Piet Heingraaf (Pitts.AI), Renée de Wildt (Nier Vereniging Nederland), Ron Herings (PHARMO Institute for Drug Outcomes Research & Amsterdam University Medical Center), Ron J Keizer (InsightRX), Ruben Boyd (IXA), Sebastiaan L Knijnenburg (Castor), Sipke Visser (Digital Health Link), Stijn Gremmen (Dutch Kidney Foundation/Nierstichting), Teun van Gelder (Leiden University Medical Center), Tjerk S Heijmens Visser (CZ Health Insurance, Zorgverzekeraars Nederland), and Vianda S Stel (Amsterdam University Medical Center, ERA-EDTA Registry). Generative artificial intelligence (ChatGPT, OpenAI) was used solely to provide limited assistance with grammar and phrasing. The authors confirm that all writing, analysis, and interpretation were carried out by the authors without the use of AI tools.

Funding

The study was funded by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO; Dutch Research Council) (KICH1.ST01.20.011). It was co-funded in cash by the Dutch Kidney Foundation and the National Intensive Care Evaluation (NICE) foundation, and in kind by the PHARMO Institute for Drug Outcomes Research, Castor, InsightRX, Z-Index, and Digital Health Link.

Authors' Contributions

MTM prepared the original draft of the manuscript and contributed to conceptualization, formal analysis, investigation, methodology, verification, resources, and visualization. MB, JEK, and RC contributed to conceptualization, investigation, methodology, supervision, verification (MB and JEK), resources (MB and JEK), and review and editing of the manuscript. JEK, RC, and MvdH acquired funding, and MvdH also contributed to the investigation, resources, and review and editing. DF-L contributed to methodology and review and editing. JEL contributed to verification and review and editing. All authors and LEAPfROG consortium members gave final approval of the submitted version.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search query details for systematic search.
[PDF File, 172 KB - jmir_v28i1e79863_app1.pdf]

Multimedia Appendix 2
Report of stakeholder workshop 1.
[PDF File, 660 KB - jmir_v28i1e79863_app2.pdf]

Multimedia Appendix 3
Report of stakeholder workshop 2.
[PDF File, 456 KB - jmir_v28i1e79863_app3.pdf]

Multimedia Appendix 4
Characteristics of included studies.
[PDF File, 313 KB - jmir_v28i1e79863_app4.pdf]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 40 KB - jmir_v28i1e79863_app5.docx]

References

1. Hammack-Aviran CM, Brelsford KM, McKenna KC, Graham RD, Lampron ZM, Beskow LM. Research use of electronic health records: patients' views on alternative approaches to permission. AJOB Empir Bioeth 2020;11(3):172-186. [doi: 10.1080/23294515.2020.1755383] [Medline: 32338567]
2. Jørgensen JT. Twenty years with personalized medicine: past, present, and future of individualized pharmacotherapy. Oncologist 2019 Jul;24(7):e432-e440. [doi: 10.1634/theoncologist.2019-0054] [Medline: 30940745]
3. Fraccaro P, Arguello Casteleiro M, Ainsworth J, Buchan I. Adoption of clinical decision support in multimorbidity: a systematic review. JMIR Med Inform 2015 Jan 7;3(1):e4. [doi: 10.2196/medinform.3503] [Medline: 25785897]
4. Masnoon N, Shakib S, Kalisch-Ellett L, Caughey GE. What is polypharmacy? a systematic review of definitions. BMC Geriatr 2017 Oct 10;17(1):230. [doi: 10.1186/s12877-017-0621-2] [Medline: 29017448]
5. Chowdhury SR, Chandra Das D, Sunna TC, Beyene J, Hossain A. Global and regional prevalence of multimorbidity in the adult population in community settings: a systematic review and meta-analysis. EClinicalMedicine 2023 Mar;57:101860. [doi: 10.1016/j.eclinm.2023.101860] [Medline: 36864977]
6. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. The Lancet 2012 Jul;380(9836):37-43. [doi: 10.1016/S0140-6736(12)60240-2]
7. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3(1):160018. [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
8. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care 2013 Aug;51(8 Suppl 3):S30-S37. [doi: 10.1097/MLR.0b013e31829b1dbd] [Medline: 23774517]
9. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017 Jan 1;24(1):198-208. [doi: 10.1093/jamia/ocw042]
10. Magrabi F, Ammenwerth E, McNair JB, et al. Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications. Yearb Med Inform 2019 Aug;28(1):128-134. [doi: 10.1055/s-0039-1677903] [Medline: 31022752]
11. Kristiansen TB, Kristensen K, Uffelmann J, Brandslund I. Erroneous data: the achilles' heel of AI and personalized medicine. Front Digit Health 2022;4:862095. [doi: 10.3389/fdgth.2022.862095] [Medline: 35937419]
12. Aitken M, de St Jorre J, Pagliari C, Jepson R, Cunningham-Burley S. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. BMC Med Ethics 2016 Nov 10;17(1):73. [doi: 10.1186/s12910-016-0153-x] [Medline: 27832780]

XSL•FO
RenderX

13. Skovgaard LL, Wadmann S, Hoeyer K. A review of attitudes towards the reuse of health data among people in the European Union: the primacy of purpose and the common good. Health Policy 2019 Jun;123(6):564-571. [doi: 10.1016/j.healthpol.2019.03.012] [Medline: 30961905]
14. Piasecki J, Walkiewicz-Żarek E, Figas-Skrzypulec J, Kordecka A, Dranseika V. Ethical issues in biomedical research using electronic health records: a systematic review. Med Health Care Philos 2021 Dec;24(4):633-658. [doi: 10.1007/s11019-021-10031-6] [Medline: 34146228]
15. Morley J, Machado CCV, Burr C, et al. The ethics of AI in health care: a mapping review. Soc Sci Med 2020 Sep;260:113172. [doi: 10.1016/j.socscimed.2020.113172]
16. Tang L, Li J, Fantus S. Medical artificial intelligence ethics: a systematic review of empirical studies. Digit Health 2023;9:20552076231186064. [doi: 10.1177/20552076231186064] [Medline: 37434728]
17. Verbeek PP, Tijink D. Guidance ethics approach: an ethical dialogue about technology with perspective on actions. University of Twente. 2020. URL: https://ris.utwente.nl/ws/portalfiles/portal/247401391/060_002_Boek_Guidance_ethics_approach_Digital_EN.pdf [accessed 2026-02-04]
18. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
19. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 5;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
20. Benchimol EI, Smeeth L, Guttmann A, et al. The reporting of studies conducted using observational routinely-collected health data (RECORD) statement. PLoS Med 2015 Oct;12(10):e1001885. [doi: 10.1371/journal.pmed.1001885] [Medline: 26440803]
21. Krom A, de Boer A, Geurtzen R, de Vries MC. Capabilities and stakeholders - two ways of enriching the ethical debate on artificial womb technology. Am J Bioeth 2023 May;23(5):110-113. [doi: 10.1080/15265161.2023.2191028] [Medline: 37130420]
22. Alami H, Lehoux P, Auclair Y, et al. Artificial intelligence and health technology assessment: anticipating a new level of complexity. J Med Internet Res 2020 Jul 7;22(7):e17707. [doi: 10.2196/17707] [Medline: 32406850]
23. Baric-Parker J, Anderson EE. Patient data-sharing for AI: ethical challenges, catholic solutions. Linacre Q 2020 Nov;87(4):471-481. [doi: 10.1177/0024363920922690] [Medline: 33100395]
24. Bednorz A, Mak JKL, Jylhävä J, Religa D. Use of electronic medical records (EMR) in gerontology: benefits, considerations and a promising future. Clin Interv Aging 2023;18:2171-2183. [doi: 10.2147/CIA.S400887] [Medline: 38152074]
25. Cohen IG, Amarasingham R, Shah A, Xie B, Lo B. The legal and ethical concerns that arise from using complex predictive analytics in health care. Health Aff (Millwood) 2014 Jul;33(7):1139-1147. [doi: 10.1377/hlthaff.2014.0048]
26. Fischer T, Brothers KB, Erdmann P, Langanke M. Clinical decision-making and secondary findings in systems medicine. BMC Med Ethics 2016 May 21;17(1):32. [doi: 10.1186/s12910-016-0113-5] [Medline: 27209083]
27. Ford E, Oswald M, Hassan L, Bozentko K, Nenadic G, Cassell J. Should free-text data in electronic medical records be shared for research? a citizens' jury study in the UK. J Med Ethics 2020 Jun;46(6):367-377. [doi: 10.1136/medethics-2019-105472]
28. Müller S. Is there a civic duty to support medical AI development by sharing electronic health records? BMC Med Ethics 2022 Dec 10;23(1):134. [doi: 10.1186/s12910-022-00871-z] [Medline: 36496427]
29. Kamradt M, Poß-Doering R, Szecsenyi J. Exploring physician perspectives on using real-world care data for the development of artificial intelligence-based technologies in health care: qualitative study. JMIR Form Res 2022 May 18;6(5):e35367. [doi: 10.2196/35367] [Medline: 35583921]
30. Knevel R, Liao KP. From real-world electronic health record data to real-world results using artificial intelligence. Ann Rheum Dis 2023 Mar;82(3):306-311. [doi: 10.1136/ard-2022-222626] [Medline: 36150748]
31. Liaw ST, Liyanage H, Kuziemsky C, et al. Ethical use of electronic health record data and artificial intelligence: recommendations of the primary care informatics working group of the international medical informatics association. Yearb Med Inform 2020 Aug;29(1):51-57. [doi: 10.1055/s-0040-1701980] [Medline: 32303098]
32. Anom BY. Ethics of big data and artificial intelligence in medicine. Ethics Med Public Health 2020 Oct;15:100568. [doi: 10.1016/j.jemep.2020.100568]
33. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 2020 Mar 20;368:l6927. [doi: 10.1136/bmj.l6927] [Medline: 32198138]
34. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol 2022 Nov 5;22(1):287. [doi: 10.1186/s12874-022-01768-6] [Medline: 36335315]
35. Khalid N, Qayyum A, Bilal M, Al-Fuqaha A, Qadir J. Privacy-preserving artificial intelligence in healthcare: techniques and applications. Comput Biol Med 2023 May;158:106848. [doi: 10.1016/j.compbiomed.2023.106848] [Medline: 37044052]
36. Lovis C. Unlocking the power of artificial intelligence and big data in medicine. J Med Internet Res 2019 Nov 8;21(11):e16607. [doi: 10.2196/16607] [Medline: 31702565]
37. Chekroud AM, Bondar J, Delgadillo J, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry 2021 Jun;20(2):154-170. [doi: 10.1002/wps.20882] [Medline: 34002503]

XSL•FO
RenderX

38. Ho CL, Caals K, editors. A Call for an Ethics and Governance Action Plan to Harness the Power of Artificial Intelligence and Digitalization in Nephrology Seminars in Nephrology: Elsevier; 2021. [doi: 10.1016/j.semnephrol.2021.05.009]
39. Atkinson JG, Atkinson EG. Machine learning and health care: potential benefits and issues. J Ambul Care Manage 2023;46(2):114-120. [doi: 10.1097/JAC.0000000000000453] [Medline: 36649491]
40. Breen N, Berrigan D, Jackson JS, et al. Translational health disparities research in a data-rich world. Health Equity 2019;3(1):588-600. [doi: 10.1089/heq.2019.0042] [Medline: 31720554]
41. Darcel K, Upshaw T, Craig-Neil A, et al. Implementing artificial intelligence in Canadian primary care: barriers and strategies identified through a national deliberative dialogue. PLoS ONE 2023;18(2):e0281733. [doi: 10.1371/journal.pone.0281733] [Medline: 36848339]
42. Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. NPJ Digit Med 2020;3(1):99. [doi: 10.1038/s41746-020-0304-9] [Medline: 32821854]
43. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018 Nov 1;178(11):1544-1547. [doi: 10.1001/jamainternmed.2018.3763] [Medline: 30128552]
44. Landau AY, Ferrarello S, Blanchard A, et al. Developing machine learning-based models to help identify child abuse and neglect: key ethical challenges and recommended solutions. J Am Med Inform Assoc 2022 Jan 29;29(3):576-580. [doi: 10.1093/jamia/ocab286] [Medline: 35024859]
45. Bozkurt S, Cahan EM, Seneviratne MG, et al. Reporting of demographic data and representativeness in machine learning models using electronic health records. J Am Med Inform Assoc 2020 Dec 9;27(12):1878-1884. [doi: 10.1093/jamia/ocaa164] [Medline: 32935131]
46. Rubinger L, Gazendam A, Ekhtiari S, Bhandari M. Machine learning and artificial intelligence in research and healthcare. Injury 2023 May;54 Suppl 3:S69-S73. [doi: 10.1016/j.injury.2022.01.046] [Medline: 35135685]
47. Bertelsen PS, Bossen C, Knudsen C, Pedersen AM. Data work and practices in healthcare: a scoping review. Int J Med Inform 2024 Apr;184:105348. [doi: 10.1016/j.ijmedinf.2024.105348] [Medline: 38309238]
48. Hoeyer K. Data Paradoxes: The Politics of Intensified Data Sourcing in Contemporary Healthcare: MIT Press; 2023.
49. Pink S, Ruckenstein M, Willim R, Duque M. Broken data: conceptualising data in an emerging world. Big Data Soc 2018 Jan;5(1):2053951717753228. [doi: 10.1177/2053951717753228]
50. Kennedy H, Hill RL. The feeling of numbers: emotions in everyday engagements with data and their visualisation. Sociology 2018 Aug;52(4):830-848. [doi: 10.1177/0038038516674675]
51. van Amsterdam WAC, van Geloven N, Krijthe JH, Ranganath R, Cinà G. When accurate prediction models yield harmful self-fulfilling prophecies. Patterns (N Y) 2025 Apr 11;6(4):101229. [doi: 10.1016/j.patter.2025.101229] [Medline: 40264961]
52. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021 Nov;3(11):e745-e750. [doi: 10.1016/S2589-7500(21)00208-9] [Medline: 34711379]
53. Bak M, Madai VI, Celi LA, et al. Federated learning is not a cure-all for data ethics. Nat Mach Intell 2024;6(4):370-372. [doi: 10.1038/s42256-024-00813-x]
54. Staunton C, Shabani M, Mascalzoni D, Mežinska S, Slokenberga S. Ethical and social reflections on the proposed European Health Data Space. Eur J Hum Genet 2024 May;32(5):498-505. [doi: 10.1038/s41431-024-01543-9] [Medline: 38355959]
55. Prainsack B, El-Sayed S, Forgó N, Szoszkiewicz Ł, Baumer P. Data solidarity: a blueprint for governing health futures. Lancet Digit Health 2022 Nov;4(11):e773-e774. [doi: 10.1016/S2589-7500(22)00189-3] [Medline: 36307191]
56. Bak MAR, Ploem MC, Tan HL, Blom MT, Willems DL. Towards trust-based governance of health data research. Med Health Care Philos 2023 Jun;26(2):185-200. [doi: 10.1007/s11019-022-10134-8] [Medline: 36633724]
57. Proost MD. Data solidarity disrupted: musings on the overlooked role of mutual aid in data-driven medicine. Kennedy Inst Ethics J 2023;33(4):401-419. [doi: 10.1353/ken.2023.a931052] [Medline: 38973484]
58. Turnhout E, Metze T, Wyborn C, Klenk N, Louder E. The politics of co-production: participation, power, and transformation. Curr Opin Environ Sustain 2020 Feb;42:15-21. [doi: 10.1016/j.cosust.2019.11.009]
59. Cazzolli R, Sluiter A, Guha C, et al. Partnering with patients and caregivers to enrich research and care in kidney disease: values and strategies. Clin Kidney J 2023 Sep;16(Suppl 1):i57-i68. [doi: 10.1093/ckj/sfad063] [Medline: 37711636]
60. Dedding C, Aussems K. Participatie, het verschil tussen een methode en een kritisch paradigma [participation, the difference between a method and a critical paradigm]. TSG Tijdschr Gezondheidswet [TSG J Health Sci] 2024;102(3):81-87. [doi: 10.1007/s12508-024-00439-9]
61. Lysen F, Wyatt S. Refusing participation: hesitations about designing responsible patient engagement with artificial intelligence in healthcare. J Responsible Innov 2024 Dec 31;11(1):2300161. [doi: 10.1080/23299460.2023.2300161]
62. Roche C, Lewis D, Wall P. Artificial intelligence ethics: an inclusive global discourse. arXiv. Preprint posted online on Aug 23, 2021. [doi: 10.48550/arXiv.2108.09959]
63. Ramirez AH, Sulieman L, Schlueter DJ, et al. The All of Us research program: data quality, utility, and diversity. Patterns (N Y) 2022 Aug 12;3(8):100570. [doi: 10.1016/j.patter.2022.100570] [Medline: 36033590]
64. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023 Jan 3;10(1):1. [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]

XSL•FO
RenderX

65. Arora A, Barrett M, Lee E, Oborn E, Prince K. Risk and the future of AI: algorithmic bias, data colonialism, and marginalization. Inf Organ 2023 Sep;33(3):100478. [doi: 10.1016/j.infoandorg.2023.100478]

Abbreviations

AI: artificial intelligence
AKI: acute kidney injury
CKD: chronic kidney disease
DAKI: drug-induced acute kidney injury
EHR: electronic health record
GDPR: General Data Protection Regulation
GEA: Guidance Ethics Approach
LEAPfROG: Leveraging Real-World Data to Optimize Pharmacotherapy Outcomes in Multimorbid Patients Using Machine Learning and Knowledge Representation Methods
MeSH: Medical Subject Headings
NVN: Nierpatiënten Vereniging Nederland (Kidney Patients Association Netherlands)
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis

Edited by A Stone; submitted 30.Jun.2025; peer-reviewed by C Wu, DS Jaisankar, N Gupta; revised version received 25.Nov.2025; accepted 25.Nov.2025; published 02.Mar.2026.

Please cite as:
Maris MT, Klopotowska JE, Cornet R, Van den Hoven MA, Lieverse JE, Fernández-Llaneza D, Bak MAR, LEAPfROG Consortium
The Ethics of Leveraging Routinely Collected Patient Data for AI Development: Mixed Methods Study
J Med Internet Res 2026;28:e79863
URL: https://www.jmir.org/2026/1/e79863
doi:10.2196/79863

© Menno T Maris, Joanna E Klopotowska, Ronald Cornet, Mariëtte A Van den Hoven, Joris E Lieverse, Daniel Fernández-Llaneza, Marieke A R Bak, LEAPfROG Consortium. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Ethical Knowledge, Challenges, and Institutional Strategies Among Medical AI Developers and Researchers: Focus Group Study

Sophia Fantus[1], PhD; Jinxu Li[2], MA; Tianci Wang[3], MS; Lu Tang[2], PhD

[1]School of Social Work, The University of Texas at Arlington, Arlington, TX, United States

[2]Department of Communication & Journalism, Texas A&M University, College Station, TX, United States

[3]Burnett School of Medicine, Texas Christian University, Fort Worth, TX, United States

Corresponding Author:
Lu Tang, PhD
Department of Communication & Journalism
Texas A&M University
4234 TAMU
College Station, TX, 77843-4234
United States
Phone: 1 2136754090
Email: ltang@tamu.edu

Abstract

Background: As artificial intelligence (AI) becomes increasingly embedded in clinical decision-making and preventive care, it is urgent to address ethical concerns such as bias, privacy, and transparency to protect clinician and patient populations. Although prior research has examined the perspectives of medical AI stakeholders, including clinicians, patients, and health system leaders, far less is known about how medical AI developers and researchers understand and engage with ethical challenges as they develop AI tools. This gap is consequential because developers' ethical awareness, decision-making, and institutional environments influence how AI tools are conceptualized and deployed in practice. Thus, it is essential to understand how developers perceive these issues and what supports they identify as necessary for ethical AI development.

Objective: The objectives of the study were twofold: (1) to examine medical AI developers' and researchers' knowledge, attitudes, and experiences with AI ethics; and (2) to identify recommendations to enhance and strengthen interpersonal and institutional ethics-focused training and support.

Methods: We conducted 2 semistructured focus groups (60-90 minutes each) in 2024 with 13 AI developers and researchers affiliated with 5 US-based academic institutions. Participants' work spanned a wide variety of medical AI applications, including Alzheimer disease prediction, clinical imaging, electronic health records analysis, digital health, counseling and behavioral health, and genotype–phenotype modeling. Focus groups were conducted via Microsoft Teams, recorded, and transcribed verbatim. We applied conventional qualitative content analysis to inductively identify emerging concepts, categories, and themes. Coding was performed independently by 3 researchers, with consensus reached through iterative team meetings.

Results: The analysis identified four key themes: (1) AI ethics knowledge acquisition: participants reported learning about ethics informally through peer-reviewed literature, reviewer feedback, social media, and mentorship rather than through structured training; (2) ethical encounters: participants described recurring ethical challenges related to data bias, patient privacy, generative AI use, commercialization pressures, and a tendency for research environments to prioritize model accuracy over ethical reflection; (3) reflections on ethical implications: participants expressed concern about downstream effects on patient care and clinician autonomy, and model generalizability, noting that rapid technological innovation outpaces regulatory and evaluative processes; and (4) strategies to mitigate ethical concerns: recommendations included clearer institutional guidelines, ethics checklists, interdisciplinary collaboration, multi-institutional data sharing, enhanced institutional review board support, and the inclusion of bioethicists as members of the AI research team.

Conclusions: Medical AI developers and researchers recognize significant ethical challenges in their work but lack structured training, resources, and institutional mechanisms to address them. Findings of this study underscore the need for institutions to consider embedding ethics into research processes through practical tools, mentorship, and interdisciplinary partnerships. Strengthening these supports is essential to preparing the next generation of developers to design and deploy ethical AI in health care.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e79613) doi:10.2196/79613

KEYWORDS

AI ethics; AI; artificial intelligence; ethics; focus group; medical

Introduction

Artificial Intelligence in Health and Medicine

Artificial intelligence (AI) is reshaping health care; AI tools are aimed at reducing costs [1], streamlining clinical workflow [2], and facilitating clinician and patient experiences [3]. Current AI applications may include assistance with clinical decisions, image-based diagnosis, self-diagnosis, mental health screening, and chronic disease management [2]. For example, electronic health records (EHRs) use natural language processing to support clinical decisions [4], and at-home AI monitoring systems assist older adults and those with long-term chronic illnesses, potentially alleviating caregiver burden [5]. Health care providers have started to use AI for medical imaging, diagnosis and disease screening, and prediction [6-9]. Furthermore, emerging scholarship demonstrates that AI has shown, to some extent, faster diagnostic speed and higher accuracy than human experts in image analysis and precision medicine [1,10]. The speed at which AI has been integrated and accepted into health care networks and its ease of access for users are unprecedented.

Despite these benefits, there are significant obstacles to AI implementation in clinical practice. Concerns about patient data security, privacy, clinician and patient autonomy, and decision-making may erode trust in AI outputs [11,12]. In addition, the pace of AI technology innovation often surpasses regulatory guidance at the federal and state levels [13]. Identifying and understanding ethical challenges may help establish practice and policy guidelines across health systems. Such mechanisms may ensure that AI developers and researchers, along with clinicians, patients, and health system leadership and administration, adapt and integrate AI that considers the needs and perspectives of all stakeholders. The purpose of this study was to explore AI developers' and researchers' understanding of and training in medical AI ethics.

Ethical Attitudes and Knowledge of Medical AI

Recent scholarship has shown rising ethical concerns among various stakeholders such as clinicians, patients, families, and policymakers who engage with medical AI [12,14-16]. Clinicians report that AI tools serve as a time-saving benefit in completing administrative tasks, which may effectively increase clinical productivity and patient engagement [6,17]. Yet, they are concerned with patient data privacy, the impacts on the clinician-patient relationship, and the possibility that the financial burdens of AI tools may heighten health inequities [7,17,18]. Patients and families voice similar concerns, focusing on patient autonomy and shared decision-making [19,20]. Patients articulate unease with the application of AI in treatment recommendations, medication administration, and surgical procedures [20].

Gap in Scholarship

The attitudes and perspectives of medical AI ethics among health care stakeholders are important. Yet, perspectives of other key stakeholders, such as AI developers and researchers, are underrepresented in research [12]. Previous studies demonstrate that ethical issues, including privacy and data security, fairness, transparency, and reliability of machine learning predictive analytics, are encountered by AI developers [21-23]. Yet, the perceived responsibility of developers and researchers to mitigate potential harms varies widely; some AI developers are cognizant of the broader societal impacts of AI (beyond technical considerations and optimization), while others feel disconnected and detached from direct patient and clinician outcomes [14,24,25]. Many AI developers report barriers to mitigating AI harms, including limited authority to make such decisions, and external pressures to deliver products quickly, all of which can hinder ethical reflexivity [26].

With a paucity of evidence-informed data on developers' AI knowledge and attitudes, further research is necessary to understand how AI ethics is addressed prior to deployment. Academic institutions play a central role in AI research and development, which lays the foundation for industry's AI design and application [14,24]. Academia has an important role in educating, training, and shaping the future generation of AI developers. This study presents a unique opportunity to guide policy, practice, and education efforts in research institutes that are aligned with the needs of AI developers and consider the deployment of ethical AI across health systems.

Research Rationale

AI developers and researchers work on algorithms that ultimately shape medical AI tools. Yet, clinicians often assume that AI tools used in clinical settings have been ethically scrutinized prior to deployment [27]. To understand ethical encounters of AI design, this study identifies the knowledge, attitudes, and training in medical AI ethics among AI researchers and developers. As an exploratory, pilot study, this work aims to offer an initial, in-depth understanding of how developers and researchers experience and navigate ethical challenges in medical AI. Rather than seeking generalizability, our goal was to capture diverse perspectives across academic contexts to illuminate key issues and inform the design of future large-scale, quantitative investigations on AI ethics training and institutional practices. Findings may inform strategies to facilitate AI ethics integration in development.

Methods

Recruitment and Sampling

The research team qualitatively explored perspectives of medical AI ethics among AI researchers and developers who were employed at academic institutions in the United States. Members of the research team had expertise in health communications,

XSL•FO
RenderX

AI and health promotion, and bioethics. Participants' inclusion criteria were (1) aged ≥18 years; (2) ability to read, understand, and communicate in English; (3) employed at an academic institution; (4) involved in medical AI research and development; and (5) consent to participate in a focus group. Participants were recruited through purposive sampling and chain referral methods to appropriately reach individuals who had an academic background in working with medical AI tools and the creation of algorithms. A study announcement and blurb were sent through listserv emails, through networks and contacts of the research team, and by word-of-mouth. Participants responded by email, and a focus group was scheduled that met participants' availability through an anonymous When2Meet Poll.

Data Collection

Focus groups (n=2; 60 minutes each) were held over Microsoft Teams and facilitated by the lead author. The facilitator had no pre-existing relationships with any participants. The lead author introduced herself to the research team and explained her background in clinical ethics, bioethics, and health scholarship prior to initiation. A semistructured interview guide was used in each focus group to reflect on participants' knowledge, attitudes, and encounters with AI ethics, and practical strategies to enhance or improve ethics education and training. Participants did not receive the interview guide prior to the scheduled focus group. Some example questions included the following: (1) What do you think is the extent of your AI ethics knowledge? (2) What is your prior experience with AI ethics? (3) What are the ethical concerns you have when conducting AI research? (4) How as a research team do you deliberate ongoing ethical issues you face? (5) In your current workplace, what training or learning opportunities are there with AI ethics? (6) What can your supervisor or the institution or university do to support you in understanding and identifying ethical issues in AI research? The interview guide was developed and piloted by members of the research team.

All focus groups (roughly 60-90 minutes each) were recorded and transcribed verbatim. No repeat interviews or focus groups were conducted. Transcripts were cleaned for errors and deidentified to protect participant confidentiality. Recordings and transcripts were stored on the university's secure password-protected server, and only members of the research team had access to the data. Thematic saturation was assessed through iterative review during and after the second focus group. At that stage, no substantively new themes emerged, and only minor variations of existing concepts were observed. We therefore determined that thematic saturation had been sufficiently achieved for the purposes of this pilot, exploratory study, and data collection concluded accordingly. In addition, members of the research team contacted different principal investigators within their respective academic units to enhance diversity in disciplinary backgrounds, institutional affiliations, and research areas. This strategy broadened the participant pool and helped capture a wider range of perspectives on AI ethics while maintaining the feasibility of this exploratory qualitative study.

Data Analysis

We applied conventional content analysis, an inductive qualitative approach used when limited theory or research exists on a topic [28]. Analysis occurs directly from the data without the use of preconceived frameworks or codebooks, allowing researchers to conduct in-depth exploration of raw data [28]. All focus group transcripts (roughly 70 pages of transcript data) were redacted and anonymized, and participants were given a specific numerical code to ensure the accuracy of responses. The 2 focus group transcripts were disseminated to 3 members of the research team (known as the coding team) to review independently from one another. Transcripts were reviewed, and data were coded inductively to form new insights and perspectives on medical AI ethics. Initial coding was conducted first to highlight exact words or phrases to denote emerging concepts. The whole research team met to discuss initial thoughts and impressions from the transcript and to develop a codebook. The transcripts were reviewed a few more times by the three independent coders to organize: (1) codes into categories, (2) categories into clusters, and (3) clusters into emerging themes. The coding team met frequently to finalize the codebook and to reach consensus on emerging themes and patterns from the data. Discrepancies in coding were discussed by the coding team to reach consensus. Codes were iteratively clustered into broader conceptual categories and then synthesized into higher-order themes that reflected shared meanings across participants. Throughout this process, the team also noted and discussed negative or divergent cases to ensure that contrasting perspectives were represented and that the final themes captured the full range of participant experiences. Once the codebook was finalized, the whole research team met to review findings and finalize themes. Rigor and trustworthiness were attained through peer debriefing with other AI and data science experts. The themes were grounded in participant data to capture their perspectives, thoughts, and insights on medical AI ethics.

Ethical Considerations

The study was approved by the Institutional Review Boards (IRBs) of Texas A&M University (approval number: IRB2023-0396D) and the University of Texas at Arlington (approval number: 2023-0234) prior to participant recruitment. Before the scheduled focus groups, participants received a copy of an informed consent form, which they signed and returned electronically. At the beginning of each focus group, the research team reviewed the consent information again orally, provided time for clarifying questions, and reiterated confidentiality limits and group norms. To protect the privacy and confidentiality of the participants, we transcribed the video recordings and removed all identifying information, including names, geographic location, or university affiliations. We conducted the data analysis based on the anonymized transcripts instead of videos. Each participant was paid US $20 in an Amazon gift card as compensation for their time.

XSL•FO
RenderX

Results

Overview

We interviewed a total of 13 participants employed in medical AI research and development. No participants declined or withdrew participation either before or during the focus group. Six (46%) participants were women, and 7 (54%) were men. Six (46%) participants identified as Chinese, and other participants identified as Asian Indian, Middle Eastern, Egyptian, Bangladeshi, Pacific Islander, or Taiwanese; only 1 participant identified as White. Participants held a range of positions, including research faculty (5/13, 38%), graduate students or research assistants (7/13, 54%), and programmer (1/13, 0.07%). Focus group participants represented 5 distinct academic institutes from different regions in the United States.

Participants' AI research included (1) disease or surgery outcomes prediction, (2) prediction and optimization of treatment, (3) analysis of electronic medical records or diagnostic imaging, (4) genetic analysis and genotype–phenotype correlation, (5) AI in counseling and behavioral health, and (6) AI in digital health and clinical trial work.

Four themes emerged from the analysis of focus group transcripts: (1) AI ethics knowledge acquisition that demonstrates how and where participants obtain AI ethics information, (2) ethical encounters that identify the main ethical issues that arise from algorithm development and design, (3) perceptions of ethical encounters to understand the implications of unresolved ethical encounters, and (4) recommendations and strategies to facilitate ethical deliberation and debrief in the workplace (Table 1).

XSL•FO
RenderX

Table 1. Themes and examples identified through conventional content analysis of focus groups on medical artificial intelligence (AI) ethics among US-based AI developers and researchers (2024).

Themes (innovation-decision process) and subthemes	Description	Quote
AI knowledge acquisition (knowledge)		
Peer-reviewed publications	Learning about AI ethics from published studies discussing bias, fairness, and responsible AI.	"So I was following up some publications, so I started to see the trend of new publications coming up and talking about like as I mentioned AI bias…"
Reviewers' feedback	Gaining awareness of ethical issues through reviewers' comments during the publication process.	"The first time I realized it was when I submitted my manuscript to Nature-like journals; most reviewers pointed it out, and that's when I started thinking about ethics seriously. "
Social media and AI policy updates	Following experts and organizations online to stay current with national and international AI ethics guidelines.	"I don't know if you've seen the recent news, like DeepMind's phone app for ChatGPT."
Informal mentorship and seminars	Receiving ethics training through informal networks, research supervisors, and academic workshops.	"Some competitions from big tech companies like Microsoft and Meta discussed these topics, and in our school, we also have weekly seminars about them."
Lack of formal training	Having little to no structured ethics education leads to uncertainty about ethical risks in AI research.	"I haven't received much training in AI ethics, so sometimes I don't even know what the problems are. Getting more training would help me recognize the issues and address them better."
Ethical encounters		
Data bias and fairness	Challenges related to underrepresentation in training data and unfair model outputs.	"Sometimes the data we use don't really represent everyone, so the model ends up being unfair to certain groups."
Privacy concerns	Issues with using patient data without proper consent or beyond the original intended use.	"Using patient data can be tricky; we're not always sure if we have full consent or if it's okay to reuse it for other purposes."
Use of generative AI	Concerns about researchers using tools like ChatGPT to fabricate or skip steps.	"Some people just ask ChatGPT to write sections for them, and that really blurs the line between help and fabrication."
Commercialization pressures	Ethical concerns regarding profit-driven deployment by tech companies over academic integrity.	"Once big companies get involved, the focus often shifts from research integrity to making profits."
Focus on accuracy over ethics	Some researchers prioritize performance over ethical considerations.	"Everyone talks about model accuracy, but barely anyone mentions the ethical side of it."
Reflections on ethical implications		
Model generalizability and explainability	Ethical concerns arise when models cannot be applied broadly or are not easily interpretable.	"Sometimes the model works great on one dataset but fails completely on another, and we don't really know why."
Impact on patient care	Researchers worry about AI models causing harm or failing to help diverse patient populations.	"If the model gives the wrong prediction, it could actually harm patients instead of helping them."
Clinician autonomy and displacement	Fears that AI may replace doctors or alter clinician-patient relationships.	"Some doctors worry that AI might start making decisions for them or replace parts of their job."
Technological pace vs evaluation speed	Difficulty in evaluating AI tools quickly enough to match their development speed.	"AI is moving so fast that our evaluation methods can't really keep up."
Ethical burden on researchers	Responsibility to address AI ethics falls heavily on developers without adequate support.	"We're the ones expected to think about ethics, but no one really gives us the tools or training to do it properly."
Strategies to mitigate ethical concerns (implementation and confirmation)		
Guideline communication	Improve access to updated ethical AI guidelines and standards.	"We really need clearer and more accessible guidelines on AI ethics; sometimes it's hard to even find the latest ones."
Ethics checklists and scenarios	Using predefined lists or cases to test and evaluate model ethics and bias.	"Having a checklist or real cases to go through would make it easier to see where our model might go wrong ethically."
Data collaboration and diversity	Partnering with other institutions to diversify datasets and reduce bias.	"If we could share data across more institutions, the models would be less biased and more reliable."
IRB[a] support and governance	Having AI-specific ethics experts within IRBs to guide responsible research.	"It would help a lot if IRBs had someone who actually understands AI to guide us on the ethical parts."

XSL•FO
RenderX

Themes (innovation-decision process) and subthemes	Description	Quote
Inclusion of bioethicists	Adding bioethics experts to AI teams to help identify and resolve ethical issues.	"Having a bioethicist on the team would make us think about these issues more seriously from the start."

[a]IRB: Institutional Review Board.

Medical AI Ethics Knowledge Acquisition

Participants discussed various avenues through which they sought AI ethics information and knowledge, including peer-reviewed publications, journal feedback, social media, and informal institutional learning. Several participants mentioned journal submission guidelines or peer review feedback that relayed information on AI ethics or included statements that mentioned AI use and plagiarism:

> *Although I had studied ethics during my medical school, I never paid attention during machine learning research. The first time I came to know was when I submitted my manuscript…the first thing they [reviewers] pointed out was about this [ethics]…most of the reviewers were concerned about it. So that's when I started thinking about it more seriously. I came to start thinking about ethics because if you publish in good journals, people will point out those things.* [P2]

Others observed a rise in publications on AI ethics and begun to read peer-reviewed articles for information: "I started seeing the trend of new publications coming up, talking about bias, fairness, and then AI ethics" (P5).

Other participants relied on social media to inform AI ethics knowledge:

> *I've been following [on X] anyone that has their hands in AI ethics or AI policy and checking all the guidelines, not just institutional levels, but national and international levels. It's hard to keep up with all the literature that's being pumped out right now. But it's important to at least familiarize yourself with some of the different pieces…and what the relevant concerns are that are transcending that international sphere.* [P6]

Social media was perceived as more current and relevant than peer-reviewed publications, able to keep up with fast-paced developments.

A small number of participants described that AI research ethics knowledge derived from informal discussions with supervisors or participation in university seminars or workshops:

> *I have really lucked out into having good people in my circle and training me. I think that's a huge resource in terms of understanding ethics and AI, and then also intentional engagement with current guidelines that are being put out.* [P6]

Yet, one participant remained silent during this discussion. When prompted, the participant stated having little to no knowledge about AI ethics:

> *I have received not too much training in AI ethics, so that is why I don't even know what the problems are. Even if I'm making some mistakes, I don't know if those are things that I should have been careful about.* [P2]

This participant's experience is important to elucidate, as it shows potential training gaps and impacts on students and researchers.

Ethical Encounters in AI Development

Participants reflected on ethical challenges encountered in their research environments, including concerns about data bias, privacy, commercialization, and the use of generative AI tools. Participants discussed the fabrication of data, in which the reliance on AI-generated tools, like ChatGPT, has enabled colleagues to skip steps through automated written responses. Yet, the primary ethical obstacle in medical AI research, as reported by participants, was bias and underrepresentation within training datasets. The ethical issues were consequences related to predictive modeling and fairness, especially how data omission could disproportionately exclude people of color. As a participant who worked on radiation therapy, observed:

> *When we build the predictive model, if our model is just purely based on the data we collected, it seems like it's not very fair for Black people or Asian people. That's the issue we are currently facing.* [P4]

Another participant agreed, stating:

> *Most of the data are coming from European [and] the therapeutics will be ultimately optimized for a certain group of people [so] it can't be generalized. If you are not careful with what kind of metric you're using to assess and evaluate that model, you're pretty much classifying everything as negative. And institutions like to incorporate these models into their systems. If you put more weight on them [the positive cases] you might be identifying the white skin tone but not the darker skin tone.* [P1]

There were also concerns of data security and the risk of breaching patient data privacy:

> *If you're using patient data without their proper consent or you use data that trains a model that is then used for something else that's not within the previously defined scope, that is not ethical.*

A participant discussed that privacy was also the "need to test it [the model] and then be transparent to the community and [provide] the proper instructions of the model's performance degree" (P10) to ensure that the model is explainable and interpretable to key stakeholders.

Participants deliberated on the dangers of the commercialization of AI technologies and limited regulations:

XSL•FO
RenderX

> *What I am really concerned with is that these big tech companies are pushing very hard to deploy their AI model into the hospital system. It's linked to profit, a very profitable market; if those big tech companies want to push their product, I don't know whether they will do it with the same level of checks and balances because it's profit-driven, and they can promise a lot of money, and we cannot make that promise.* [P9]

The overall ethical concern related to commercialization was the fast-paced development of AI and the time sensitivity of implementing AI into health care spaces; for-profit companies will be selected over evidence-informed AI research programs.

> *The perceived competitiveness between big tech corporate research and also academic research and I feel like they are not playing by the rules because they can skirt and essentially do things that we have to abide by like privacy issues and so forth.* [P8]

In contrast, there were participants who did not perceive these as ethics issues but rather as an accuracy issue: "I'm not really focused on ethics. I tend to focus more on accuracy, something that will make the model better but not actually the ethics" (P3).

Reflections on Ethical Implications

Participants described how the ethical encounters stated above influenced research design and modeling choices and raised broader concerns about patient care, clinician roles, and the future of health care. For example, issues related to fairness and bias influenced generalizability:

> *from a data scientist perspective, it's an issue; you cannot have a very accurate model with very high bias. You can build your model, but we want the model to have higher generalizability; we need to take this issue from a data scientist perspective.* [P4]

Participants noted that limits to explainability of AI impeded solutions to resolve ethical issues:

> *with so much advancement in AI technology, there is still no standard correct ways of evaluating my model because I haven't understood my data or the distribution of the data yet.* [P1]

Other participants considered ways ethical issues in AI development may impact patient care and physician interactions. For example, participants who worked on large language models deliberated on how AI tools can generate clinical notes for the patient and questioned the accuracy of how "clinical notes could be to the specific patient…. we don't know how that benefits the patients" (P7). Questions related to predictive modeling also drew fears of perpetuating patient harms:

> *How do we balance advancing healthcare to truly help patients in this unprecedented way, but also make sure that we're not exploiting them or using models that aren't appropriate for them?* [P6]

Participants reported that physician autonomy and patient-physician relationships were another important area to identify the ethical implications of AI deployment in health care. One participant asked, "whether AI is going to replace certain jobs and tasks and maybe even eventually replace doctors; that's a conversation I have about my research" (P8). The timing to evaluate technology was an added concern that could impact patient care:

> *By the time you come up with a standard metric that you need to satisfy your AI model to be deployed in a healthcare facility…maybe the technology has completely changed by then. I don't know what the solution would be, but clinicians, researchers, lawmakers, you know, everybody needs to be on board because they can no longer take that long to evaluate a technology.* [P1]

The perceived impacts of AI ethics placed added burdens and responsibilities on AI researchers and developers. Heightened attention on AI ethics placed more obligations on AI developers and researchers to resolve these issues, yet without training or learned mitigation strategies.

Strategies to Mitigate Ethical Concerns

Participants described a range of strategies to address AI ethics in research, including individual practices, team-based approaches, and institutional-level interventions. Participants suggested individual and team-based approaches that foster transparent communication and knowledge mobilization. For example, a participant emphasized the need for

> *…good communication about the latest guidelines that are available from different communities. If that becomes available to use as students and even as faculty that will be more helpful to make us more compliant with those regulations.* [P5]

Guidelines, in turn, can assist in the design and development of checklists or critical scenarios to mitigate biased models. One participant stated

> *…maybe we can have a checklist on what we need to see before doing something that's more concrete. I know it's difficult to do that in AI ethics because we don't directly use it for patient outcomes right now. But I think it will be a good starting point to have some kind of checklist on what we should be careful about.* [P2]

Another participant wanted the actual model to counteract biases:

> *The first check should be done on the data and how the data biases have been handled by the AI models. And last, what are the abusive ways this model can be used? We should have some critical scenario by which we can test our model, like some exerted test on the product to see whether this product is stable up to two years…whether it is up to our expectation.* [P1]

Participants advocated for organizational and institutional strategies to support ethical AI development. One approach mentioned was multi-institutional collaboration to improve data quality and diversity and mitigate bias by increasing access to larger, more representative datasets. One participant said: "Where you don't have enough data to support the deep learning [models], we have to collaborate with other institutes to not

XSL•FO
RenderX

only expand the sample size but also introduce diversity into the data" (P4). Participants also called for ethics consortiums to foster ethical awareness and skill development. For instance, a participant said: "I feel like getting more training will help me more to even identify what the problems are and then address them" (P2).

An added strategy was to equip the IRB with AI-specific guidance or AI ethics expertise on regulatory committees to enhance regulation through a uniform approach and facilitate adherence to best practices. IRB involvement could promote a more consistent approach to oversight and improve adherence to best practices. However, several participants expressed frustration that current assistance or guidance sought from the IRB often resulted in confusion rather than clarity:

> *We have those IRB boards and maybe some better governance…to have somebody also on AI ethics and being responsible for sharing that awareness as well as ensuring that we are going through the guidelines and sticking through the regulations.* [P5]

Finally, for some participants, a bioethicist or ethics expert should be involved as a potential interdisciplinary member of the research team:

> *When you don't have a bioethicist at your beck and call or infused in the research in some degree, that makes it really tricky too because you might not have the checks and balances that are appropriate in maybe expanding your research or getting into the right market.* [P6]

Discussion

Overview

The integration of medical AI in preventive care and clinical decision-making means that researchers, data scientists, and those involved in the design and development of AI need to start becoming attuned to its clinical impacts. This study aims to address gaps in scholarship by examining AI researchers' ethics perspectives. Academic institutions, such as universities and research institutes, play a central role in educating the future generation of AI developers on AI ethics and design.

Findings from the study inform how medical AI may be diffused into health care settings and how its use communicated effectively between physicians, staff, and patients. The themes from this study may be adapted into Rogers' diffusion of innovation framework [29]. Rogers maps a 5-stage decision-making process to evaluate and adopt AI innovation into practice. The series of stages is as follows: (1) knowledge (gains understanding), (2) persuasion (reflect on attitudes), (3) decision (activities and experiences that lead to choice), (4) implementation (its actual use in practice), and (5) confirmation (to avoid dissonance and conflict).

Our findings ought to be conceptualized within the diffusion of innovation framework to understand how perspectives of AI researchers and developers offer insight into the steps, attitudes, and barriers that influence the decision-making of medical AI adoption and integration. Findings from this study may inform policy, practice, and education efforts to readily prepare AI researchers and developers to identify and examine ethical encounters in their work and to illustrate how medical AI attitudes, perceptions, and support may influence adoption or rejection [29].

AI Ethics Knowledge Acquisition

In the knowledge stage of individual decision-making, participants in this study received information about AI ethics from a multitude of sources, including social media, peer review journal commentary and publications, and voluntary workshops and seminars. Students' particular focus on social media as an access point for AI knowledge may be an important consideration to assess (1) the type of accuracy of messaging received and (2) the ethical issues being described and disseminated. Knowledge garnered through social media may filter into how students understand and evaluate their own research and ethical encounters, including how early adopters may rely more on social media than on peer-reviewed sources. For example, participants who described issues as rooted in accuracy rather than ethics may benefit from conversations and messaging that deciphers ethical issues from technical issues; the ability to understand how to identify and label issues as ethical (rather than solely technical) could enhance medical AI ethics knowledge and lead to a more nuanced and robust deliberation on how ethics may impact medical AI design, development, and deployment.

Ethical Encounters and Resolution Strategies

Past research experiences of participants impacted their attitudes and perspectives of medical AI. Bias and fairness were central ethical challenges identified by participants, particularly the underrepresentation of people of color in training datasets. Such omissions risk reinforcing structural inequities and limiting the generalizability of medical AI systems. To address these issues, future research should prioritize diversifying medical datasets and integrating fairness auditing across development, supported by multi-institutional collaborations and community engagement to ensure representativeness, transparency, and accountability [30-32]. Together, these efforts can help mitigate the disproportionate exclusion of marginalized groups and promote more equitable AI-driven health outcomes to address negative attitudes toward medical AI innovation.

Extant scholarship echoes the current study's findings by demonstrating that AI developers possess some awareness of key ethical principles, such as fairness, data security, transparency, and reliability [14], that may persuade their decision-making in adoption. Nichol et al [14], for example, conducted semistructured interviews with 40 employees from AI organizations. Participants in the study identified potential impacts of ethical issues, such as violation of patients' privacy, misdirected health care practices, and disrupted health care systems. Other studies have similarly shown that some AI developers are sensitive to broader societal impacts of ethical AI [24,25], beyond technical issues and optimization of their algorithms [33,34].

Although participants in this study were able to identify emerging ethical issues and had thoughtfully evaluated how

XSL•FO
RenderX

these issues would impact patient and clinician experiences, there were limited resolution strategies. These experiences impacted AI researchers' decision-making, including not knowing whether to adopt or reject innovations in their work. For participants, ethical encounters were often left unresolved with no clear direction on how to proceed. The lack of informed decision-making was rooted in a lack of clarity from institutions and left participants feeling that AI researchers and developers held an undue burden in deploying ethical AI tools without further scrutiny or analysis. The added pressure on AI developers and researchers to perform was perceived as a challenge, and our findings suggest that other key stakeholders (including physicians or clinicians and health systems) ought to contribute to ethical decision-making when AI is used in practice. Thus, findings from this study show that decisions of whether to adopt or reject AI ought to include diverse perspectives to allow for more information, to identify problems, and to have support [29].

AI models must continue to be questioned and analyzed by stakeholders even after deployment. With AI technologies changing so rapidly, participants struggled to balance the fast-paced development of AI algorithms with the ethical concerns that arose. This led participants to articulate that medical AI tools ought to be continuously reviewed and evaluated.

The barriers and limited support indicate that implementation and confirmation, the final stages in the innovation-decision process, may be difficult to reach. Participants described that in the development of medical AI innovation, they often evaluated long-term impacts on patients and families and desired further support from mentors, supervisors, and organizational leadership. The perspectives of participants show that there continue to be conflicting messages and dissonance among researchers and developers regarding the adoption and use of medical AI in practice settings. Further organizational practices and policies ought to be considered to assist in decision-making activities to facilitate a more robust and comprehensive adoption process.

This study's findings echo prior work wherein AI developers voiced confusion regarding their own roles and responsibilities in mitigating the potential harm of their tools, compounded by perceived limited authority, external pressures to produce, and the difficulty of balancing productivity and ethical considerations [26,33]. Algorithm development is highly complex and iterative, making it difficult for researchers to predict its ethical impact and apply oversight in the process. As participants in this study noted, transparency and explainability were key ethical issues, and a gap in accessible checklists or guidelines heightened obstacles to elucidating datasets and explaining patterns to clinicians and patients who may rely on these algorithms for diagnosis and treatment. The issue here is that resolving ethical encounters requires additional time and energy from AI developers, which may be an added challenge in a high-stress environment that is at odds with the fast-paced development of commercial AI tools [14]. The capacity to build collaborative environments, hold ethics consortiums, and have a robust network of people and resources to support AI ethics awareness, knowledge, and action is critical to support AI developers and researchers. Relieving some of the burdens on AI developers and researchers with institutional mechanisms can model an environment that supports ethical rigor and deliberation and lead to reinforcement and confirmation of medical AI technologies in practice settings.

As pointed out by Mittelstadt [35], compared with medicine, the field of AI research is much more heterogeneous, without defined common aims, fiduciary duties, or historical professional norms. The constant changes and shifts related to AI policy and procedure create difficulties in outlining consistent guidelines or measures to follow. Additionally, AI developers typically have backgrounds in computer science with limited training in ethics. The relative unfamiliarity with ethical principles and their implications could add further barriers to ethical medical AI development, potentially leading to ethically flawed AI products that could impose unintended harm to patients [34]. Thus, multisite collaborations, interdisciplinary communication, and IRB guidance and best standards may help to reduce the burden on AI developers, create more teachable moments, and establish more thoughtful and intentional mechanisms for deliberating ethical encounters, along with clear resolution pathways to facilitate implementation and confirmation.

Bioethics-Informed Guidance

The inclusion of bioethicists on research teams, as stated by participants, has been suggested in prior theoretical scholarship. For example, McLennan et al [36] proposed the concept of "embedded ethics," a collaborative approach that creates interdisciplinary research teams whereby AI developers and ethicists can anticipate, identify, and address ethical issues as they arise in the development process. Other studies have suggested a practical ethics checklist for AI developers [37] that recognizes ethical and social responsibility within AI development [38] or ethics guidelines and review processes specifically designed for AI developers [39,40] to support research design and analysis. These efforts offer improvements by having a refined focus on the practicality of how to use ethics recommendations and an emphasis on frontline AI developers, who can help mitigate ethical issues prior to AI deployment and use. This study's findings demonstrate that AI developers are interested in gaining knowledge about AI ethics, are already deliberating on the ethical encounters in their design and development, and are thoughtful about the longer-term practical implications of their work in health systems. Future research ought to consider strategies to mitigate ethical encounters and to advocate for heightened ethics knowledge, training, and conversations specifically targeted for AI developers and researchers. Foundational seminars on how to identify and label an issue as ethical (as opposed to technical) are a critical first step in training to ensure all developers and researchers can recognize these encounters in practice. Supervisors and managers must consider ways to encourage ethical dialogue and empower students and faculty to seek ways to mitigate ethical concerns and bridge their work to practice. This may involve bringing in bioethicists or other ethics experts who can speak diligently, thoughtfully, and comprehensively about these topics. AI developers and researchers should not be working in siloes but rather placed in communities with other medical AI stakeholders to heighten ethical dialogue and theorize novel

XSL•FO
RenderX

mitigation strategies. Additionally, to enhance the actionability of these recommendations, institutions could develop sample ethics checklists (eg, addressing data representativeness, model explainability, and patient privacy), workflow templates that map ethical review points within the AI development process, and metrics to evaluate ongoing compliance. Such practical tools can help translate ethical principles into consistent, operational practices for research offices and AI teams. These steps may facilitate the diffusion of innovation processes to allow for an easier and more transparent decision, implementation, and confirmation process that can lead to ethical adoption of medical AI. It is clear that ethical conflict and encounters of ethical dilemmas in the development and deployment of medical AI have stark impacts on the diffusion of innovation and the ability to effectively implement and reinforce the decision to adopt. Future research may seek to understand how this process may infiltrate the decision-making and ethical attitudes and perspectives of other key stakeholders, including physicians, allied health workers, health care administrators, patients, and families.

Limitations

This is one of the first studies in North America to examine AI developers' knowledge, encounters, and recommendations of medical AI ethics. Yet, there are several limitations. The small and relatively homogenous sample limits the diversity and generalizability of the findings. Future research should pursue broader and more inclusive investigations that capture perspectives from a wider range of disciplines, institutions, and demographic backgrounds across the United States. The representation of only 5 academic institutions may narrow the findings and may overlook other ethical concerns that emerge in distinct research areas. This was also a qualitative focus group study, and, thus, participants may have had concerns regarding privacy and confidentiality, reputation, and status in responses, and the emergence of potential power imbalances with student participants. Furthermore, as with most focus group studies, participants may have provided more socially desirable responses due to the group setting or the presence of peers, which could have influenced the depth or candor of some discussions. The focus group facilitator mitigated any ethical concerns by setting group norms, ensuring privacy and confidentiality, and piloting focus group questions and prompts. Future research may consider an anonymous survey to broadly examine ethical encounters in medical AI research.

Conclusions

As an exploratory pilot study, the current findings provide preliminary insights that can guide future empirical and institutional efforts. Findings from this study are important to determine the next steps to facilitate ethical decision-making among medical AI developers and researchers. There ought to be strategies to effectively deliberate about AI ethics across research teams and create opportunities for multisite collaboration, IRB debriefs and guidelines, protocol checklists and testing mechanisms, and the involvement of key stakeholders in deliberation, including bioethicists, clinicians, patients, and hospital leadership or administration with AI research teams. These initial insights lay the groundwork for larger-scale, multi-institutional investigations that can further validate and expand on the patterns identified here. The perspectives of key stakeholders may inform stages in the innovation-decision process and gain insight into barriers, supports, and resources necessary to ethically adopt medical AI into practice.

Funding

This work was supported by the National Institute of Health, United States (grant number: 3U01AG070112-02S2) as well as the Texas A&M Institute of Data Sciences.

Data Availability

The data that support the findings of this study are not publicly available due to privacy and confidentiality agreements.

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ Checklist.
[PDF File (Adobe PDF File), 423 KB - jmir_v28i1e79613_app1.pdf]

References

1. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. In: Artificial Intelligence in Healthcare. California: Academic Press; 2020:25-60.
2. Chew HSJ, Achananuparp P. Perceptions and needs of artificial intelligence in health care to increase adoption: scoping review. J Med Internet Res 2022;24(1):e32939 [FREE Full text] [doi: 10.2196/32939] [Medline: 35029538]
3. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]

XSL•FO
RenderX

4. Afshar M, Adelaine S, Resnik F, Mundt MP, Long J, Leaf M. Deployment of real-time natural language processing and deep learning clinical decision support in the electronic health record: pipeline implementation for an opioid misuse screener in hospitalized adults. JMIR Med Inform 2023;11:e44977 [FREE Full text] [doi: 10.2196/44977] [Medline: 37079367]
5. Sapci AH, Sapci HA. Innovative assisted living tools, remote monitoring technologies, artificial intelligence-driven solutions, and robotic systems for aging societies: systematic review. JMIR Aging 2019;2(2):e15429 [FREE Full text] [doi: 10.2196/15429] [Medline: 31782740]
6. Chen Y, Stavropoulou C, Narasinkan R, Baker A, Scarbrough H. Professionals' responses to the introduction of AI innovations in radiology and their implications for future adoption: a qualitative study. BMC Health Serv Res 2021;21(1):813 [FREE Full text] [doi: 10.1186/s12913-021-06861-y] [Medline: 34389014]
7. Wangmo T, Lipps M, Kressig RW, Ienca M. Ethical concerns with the use of intelligent assistive technology: findings from a qualitative study with professional stakeholders. BMC Med Ethics 2019;20(1):98 [FREE Full text] [doi: 10.1186/s12910-019-0437-z] [Medline: 31856798]
8. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021;27(12):2176-2182 [FREE Full text] [doi: 10.1038/s41591-021-01595-0] [Medline: 34893776]
9. Hendrix N, Hauber B, Lee CI, Bansal A, Veenstra DL. Artificial intelligence in breast cancer screening: primary care provider preferences. J Am Med Inform Assoc 2021;28(6):1117-1124 [FREE Full text] [doi: 10.1093/jamia/ocaa292] [Medline: 33367670]
10. Miller DD, Brown EW. Artificial intelligence in medical practice: the question to the answer? Am J Med 2018;131(2):129-133. [doi: 10.1016/j.amjmed.2017.10.035] [Medline: 29126825]
11. Keskinbora KH. Medical ethics considerations on artificial intelligence. J Clin Neurosci 2019;64:277-282. [doi: 10.1016/j.jocn.2019.03.001] [Medline: 30878282]
12. Tang L, Li J, Fantus S. Medical artificial intelligence ethics: a systematic review of empirical studies. Digit Health 2023;9 [FREE Full text] [doi: 10.1177/20552076231186064] [Medline: 37434728]
13. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 2020;3:118 [FREE Full text] [doi: 10.1038/s41746-020-00324-0] [Medline: 32984550]
14. Nichol AA, Sankar PL, Halley MC, Federico CA, Cho MK. Developer perspectives on potential harms of machine learning predictive analytics in health care: qualitative analysis. J Med Internet Res 2023;25:e47609 [FREE Full text] [doi: 10.2196/47609] [Medline: 37971798]
15. Čartolovni A, Malešević A, Poslon L. Critical analysis of the AI impact on the patient-physician relationship: a multi-stakeholder qualitative study. Digit Health 2023;9:20552076231220833 [FREE Full text] [doi: 10.1177/20552076231220833] [Medline: 38130798]
16. Bergquist M, Rolandsson B, Gryska E, Laesser M, Hoefling N, Heckemann R. Trust and stakeholder perspectives on the implementation of AI tools in clinical radiology. Eur Radiol 2024;34(1):338-347 [FREE Full text] [doi: 10.1007/s00330-023-09967-5] [Medline: 37505245]
17. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD, Halamka JD, DesRoches CM. Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. J Med Internet Res 2019;21(3):e12802 [FREE Full text] [doi: 10.2196/12802] [Medline: 30892270]
18. Bourla A, Ferreri F, Ogorzelec L, Peretti C, Guinchard C, Mouchabac S. Psychiatrists' attitudes toward disruptive new technologies: mixed-methods study. JMIR Ment Health 2018;5(4):e10240 [FREE Full text] [doi: 10.2196/10240] [Medline: 30552086]
19. Ongena YP, Haan M, Yakar D, Kwee TC. Patients' views on the implementation of artificial intelligence in radiology: development and validation of a standardized questionnaire. Eur Radiol 2020;30(2):1033-1040 [FREE Full text] [doi: 10.1007/s00330-019-06486-0] [Medline: 31705254]
20. Witkowski K, Dougherty RB, Neely SR. Public perceptions of artificial intelligence in healthcare: ethical concerns and opportunities for patient-centered care. BMC Med Ethics 2024;25(1):74 [FREE Full text] [doi: 10.1186/s12910-024-01066-4] [Medline: 38909180]
21. Akinrinola O, Okoye CC, Ofodile OC, Ugochukwu CE. Navigating and reviewing ethical dilemmas in AI development: strategies for transparency, fairness, and accountability. GSC Adv Res Rev 2024;18(3):050-058. [doi: 10.30574/gscarr.2024.18.3.0088]
22. Drabiak K, Kyzer S, Nemov V, El Naqa I. AI and machine learning ethics, law, diversity, and global impact. Br J Radiol 2023;96(1150). [doi: 10.1259/bjr.20220934] [Medline: 37191072]
23. Mathrani A, Susnjak T, Ramaswami G, Barczak A. Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics. Comput Educ Open 2021;2. [doi: 10.1016/j.caeo.2021.100060]
24. Morley J, Kinsey L, Elhalal A, Garcia F, Ziosi M, Floridi L. Operationalising AI ethics: barriers, enablers and next steps. AI Soc 2021;38(1):411-423. [doi: 10.1007/s00146-021-01308-8]
25. Sanderson C, Douglas D, Lu Q, Schleiger E, Whittle J, Lacey J. AI ethics principles in practice: perspectives of designers and developers. IEEE Trans Technol Soc 2023;4(2):171-187. [doi: 10.1109/tts.2023.3257303]

XSL•FO
RenderX

26. Griffin TA, Green BP, Welie JVM. The ethical agency of AI developers. AI Ethics 2023;4(2):179-188. [doi: 10.1007/s43681-022-00256-3]
27. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: implications for clinical decision-making. PLOS Digit Health 2024;3(11). [doi: 10.1371/journal.pdig.0000651] [Medline: 39509461]
28. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
29. Rogers EM. Diffusion of Innovation. 4th Edition. New York: Free Press; 1995.
30. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: a call for open science. Patterns (N Y) 2021;2(10):100347 [FREE Full text] [doi: 10.1016/j.patter.2021.100347] [Medline: 34693373]
31. Drukker K, Chen W, Gichoya J, Gruszauskas N, Kalpathy-Cramer J, Koyejo S. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. J Med Imaging (Bellingham) 2023;10(6):061104 [FREE Full text] [doi: 10.1117/1.JMI.10.6.061104] [Medline: 37125409]
32. Chinta SV, Wang Z, Palikhe A, Zhang X, Kashif A, Smith MA. AI-driven healthcare: fairness in AI healthcare: a survey. PLOS Digit Health 2025;4(5):e0000864. [doi: 10.1371/journal.pdig.0000864] [Medline: 40392801]
33. Griffin TA, Green BP, Welie JV. The ethical wisdom of AI developers. AI Ethics 2024;5(2):1087-1097. [doi: 10.1007/s43681-024-00458-x]
34. Kim JP, Ryan K, Kasun M, Hogg J, Dunn LB, Roberts LW. Physicians' and machine learning researchers' perspectives on ethical issues in the early development of clinical machine learning tools: qualitative interview study. JMIR AI 2023;2:e47449 [FREE Full text] [doi: 10.2196/47449] [Medline: 38875536]
35. Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell 2019;1(11):501-507. [doi: 10.1038/s42256-019-0114-4]
36. McLennan S, Fiske A, Tigard D, Müller R, Haddadin S, Buyx A. Embedded ethics: a proposal for integrating ethics into the development of medical AI. BMC Med Ethics 2022;23(1):6 [FREE Full text] [doi: 10.1186/s12910-022-00746-3] [Medline: 35081955]
37. Morley J, Cowls J, Taddeo M, Floridi L. Ethical guidelines for COVID-19 tracing apps. Nature 2020;582(7810):29-31. [doi: 10.1038/d41586-020-01578-0] [Medline: 32467596]
38. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1(5):206-215 [FREE Full text] [doi: 10.1038/s42256-019-0048-x] [Medline: 35603010]
39. Ryan M, Stahl BC. Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. J Inf Commun Ethics Soc 2020;19(1):61-86. [doi: 10.1108/jices-12-2019-0138]
40. Srikumar M, Finlay R, Abuhamad G, Ashurst C, Campbell R, Campbell-Ratcliffe E. Advancing ethics review practices in AI research. Nat Mach Intell 2022;4(12):1061-1064. [doi: 10.1038/s42256-022-00585-2]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
IRB: institutional review board

Edited by A Stone; submitted 24.Jun.2025; peer-reviewed by I Adefolaju, X Liang, LP Gorrepati; comments to author 06.Aug.2025; revised version received 15.Dec.2025; accepted 29.Dec.2025; published 28.Jan.2026.

Please cite as:
Fantus S, Li J, Wang T, Tang L
Ethical Knowledge, Challenges, and Institutional Strategies Among Medical AI Developers and Researchers: Focus Group Study
J Med Internet Res 2026;28:e79613
URL: https://www.jmir.org/2026/1/e79613
doi:10.2196/79613
PMID:

©Sophia Fantus, Jinxu Li, Tianci Wang, Lu Tang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Educating Students About Digital Health Research Ethics: Curricula Review and Expert Interview Study

Yier Zhu[1], MPH; Brian McInnis[1], PhD; Tanya Punater[1], BS; Brittany York[1], MPH; Camille Nebeker[1,2,3], EdD, MS

[1]Herbert Wertheim School of Public Health and Human Longevity Science, University of California San Diego, La Jolla, CA, United States

[2]The Design Lab, University of California San Diego, La Jolla, CA, United States

[3]Qualcomm Institute, University of California San Diego, La Jolla, CA, United States

Corresponding Author:
Camille Nebeker, EdD, MS
Herbert Wertheim School of Public Health and Human Longevity Science
University of California San Diego
9500 Gilman Drive, MC 0811
La Jolla, CA, 92093
United States
Phone: 1 619 518 2875
Email: nebeker@health.ucsd.edu

Abstract

Background: The rapid growth of digital health research, involving wearable devices, mobile apps, and sociotechnical health systems, raises complex ethical, legal, and social considerations. While institutional review boards and research ethics frameworks address some concerns, less is known about how students and trainees in digital health are systematically educated to recognize and navigate these challenges. Understanding the scope and content of ethics training is critical to ensuring the responsible development and application of digital health technologies.

Objective: This study investigated how college students are trained to identify and address ethical considerations in digital health research through an analysis of formal curricula and expert perspectives.

Methods: Researchers reviewed 132 syllabi from 76 academic programs across 62 universities and conducted semistructured interviews with 6 leading digital health scholars. All syllabi were coded for instructional content and learning objectives. Researchers conducted open coding and collaboratively applied affinity diagramming to organize the data into hierarchical themes.

Results: All syllabi included instructional content, and most included explicit learning objectives. Analysis identified 7 key themes, which captured both explicit knowledge imparted through formal instruction and tacit knowledge cultivated through laboratory work, mentorship, and applied experiences. Findings highlighted gaps between formal ethics instruction and the realities of research practice.

Conclusions: Ethics education in digital health research develops through the interplay of formal coursework and practice-based training, each fostering complementary skills needed for data-intensive and collaborative environments. Together, these pathways support students in identifying ethical issues, applying principles contextually, anticipating emerging risks, and communicating across disciplines; however, access to experiential learning opportunities remains inconsistent. Strengthening ethics training will require expanding structured early research engagement, cultivating communities of practice, and translating tacit ethical reasoning into accessible resources. Integrating ethical reflection into routine research activities may better prepare future digital health researchers to responsibly design and govern sociotechnical health systems.

(J Med Internet Res 2026;28:e82861) doi:10.2196/82861

KEYWORDS

digital health research; ethics education; practice-based learning; qualitative research; research ethics; research training; sociotechnical health systems

XSL•FO
RenderX

Introduction

As digital technologies become integral to health research and clinical care, students must be prepared to examine related ethical, social, and organizational implications. Digital health tools, from mobile apps and wearables to artificial intelligence (AI) systems, can improve care; yet, their implementation often outpaces understanding of effects on patients, providers, and health systems [1,2]. Our research focuses on digital health research ethics as a sociotechnical concern: technologies both shape and are shaped by institutional contexts, stakeholder relationships, and data infrastructures to coordinate people, groups, and computing resources with a goal of improving patient health and well-being [3,4]. For example, a conversational agent located in a patient's room could be used to take notes during a clinical visit [5], notify nursing teams when the patient is experiencing increased anxiety [6], and converse with the patient to reduce their feelings of social isolation [7]. Digital health tools, like wearable sensors, offer new ways to monitor patients, both in and outside of the clinic, which have important ethical implications [8]. By adopting a sociotechnical lens, we attend to system components (ie, technologies, data practices, and institutional policies) and social interactions between team workflows and patient/participant engagement to examine emerging ethical issues, including unanticipated privacy and data management risks.

Academic education and training programs play a critical role in preparing digital health researchers and clinicians to navigate ethical challenges like algorithmic bias, data privacy, and access inequities, yet little empirical work has examined the availability and scope of such programs across organizations contributing to digital health research [6]. Unlike general research ethics, digital health research includes characteristics that can strain traditional academic ethics models. These differences include streams of continuous, granular individual-level data that may remain identifiable and blurred lines between research and industry data flows [9-12]. Importantly, informed consent communications are challenged by difficulties explaining passive sensing and other pervasive technologies, data management practices, and evolving algorithms [13-15].

Researchers learn to recognize and mitigate ethical challenges through university courses, workshops, and readings (explicit knowledge) as well as through firsthand opportunities to learn by training with expert research practitioners and working with clinicians, caregivers, and digital health technology designers in the field (tacit knowledge). As digital health ethical and social implications relate to the design and deployment of technology within social, cultural, and organizational contexts, it is critical for students to learn through both explicit coursework and experience-based learning (tacit knowledge). However, less is known about the specific components of this education among disciplines involved in the design and testing of digital health tools. Explicit training supplies common concepts and language; experiential learning develops judgment for context-specific decisions as teams design, deploy, and evaluate sociotechnical systems in health settings [10,11,16]. Because these sociotechnical properties scale through computing infrastructures, the stakes and failure modes in digital health differ from those typically addressed in general research ethics training.

To examine how students gain ethical competencies in digital health research, we analyzed university syllabi to characterize explicit knowledge conveyed through coursework and conducted interviews with leading digital health researchers to understand how tacit knowledge is conveyed through the conduct of research, which can include decisions about consent, data protection, stakeholder engagement, and dissemination [17,18]. Our goals for this study were to (1) evaluate how programs that contribute to digital health research socialize students to ethical and responsible research practices through coursework and (2) describe how expert researchers convey tacit ethical practices to trainees through hands-on training. We do not test hypotheses; rather, this exploratory, mixed evidence synthesis (syllabi and interviews) maps explicit and tacit learning opportunities in digital health research ethics.

Methods

Overview

This study used a multimethod qualitative design, integrating 2 complementary components to examine how ethical competencies in digital health research are taught and learned. We analyzed course syllabi to characterize explicit ethics instruction and conducted semistructured interviews with experienced digital health researchers to understand how tacit ethical knowledge is conveyed through research practice. We used the sociotechnical framework to integrate insights from syllabi (explicit knowledge) and interviews (tacit knowledge), examining how instructional content aligned with or diverged from the ethical challenges researchers encounter in sociotechnical research environments. These 2 data sources provide a comprehensive synthesis of explicit and experiential learning.

Ethical Considerations

The research activities involved minimal risk to participants, and the protocol was reviewed and verified as exempt under 45 CFR 46.104(d)2 by the institutional review board (IRB) at the University of California San Diego (exemption verification 210941). All interview participants received recruitment materials and provided verbal informed consent prior to audio recording. Participants did not receive compensation. Reporting followed the Standards for Reporting Qualitative Research (SRQR), identified by the EQUATOR (Enhancing the Quality and Transparency of Health Research) Network.

Audio recordings and transcripts were stored in University of California San Diego's secure Microsoft Teams environment, which provides encrypted, access-restricted cloud storage; only authorized study personnel with institutional credentials could access study files to ensure participant privacy and confidentiality.

Study Design

The study used 2 complementary components to examine how ethical competencies in digital health research are taught and learned. First, a syllabi analysis was conducted to characterize

XSL•FO
RenderX

explicit ethics instruction delivered through university coursework. Second, expert interviews were conducted to examine tacit ethics knowledge conveyed through hands-on research practice. Although designed and analyzed independently, the 2 components together highlight complementary dimensions of ethics education in sociotechnical digital health research. To identify academic programs and researchers relevant to digital health ethics training, this study drew upon a preliminary literature review of sociotechnical digital health research in computing and information science venues. This review informed the selection of academic units for the syllabi search as well as the identification of researchers eligible for the interview component.

Course Syllabi Collection—Explicit Knowledge

Institutional websites and departmental materials were examined to determine whether units associated with the publications included in the systematic literature review (forthcoming) offered formal or informal training opportunities focused on ethical and responsible research practices. These included required ethics coursework, elective offerings, workshops, journal clubs, and other structured learning activities relevant to sociotechnical systems and digital health research.

A total of 132 course syllabi were collected from academic units identified through the search strategy. Of these, 98 syllabi contained learning objectives and were therefore included in the learning objective analysis. The remaining syllabi were excluded because they lacked learning objectives or did not provide sufficient detail to determine ethics-related instructional content. The final analytic set represents 64 academic programs from universities in the United States and internationally.

When course information was not publicly available, research staff contacted instructors or administrative personnel to request syllabi, course descriptions, and information about relevant instructional activities. Syllabi served as the primary data source for examining explicit ethics-related knowledge conveyed through coursework.

Criteria for identifying ethics-related courses were derived from the syllabi analysis protocol, which specified how the research team determined whether a course contained substantial ethics content. Courses were classified as ethics-focused when the title, description, learning outcomes, or listed topics indicated a primary emphasis on ethical issues relevant to sociotechnical systems that contribute to digital health research. Courses in which ethics was addressed only briefly within a broader technical curriculum were not classified as ethics-focused. When syllabi were unavailable, these criteria were applied to publicly accessible catalog descriptions and course topic summaries.

To determine who to contact within each academic program, research staff reviewed departmental web pages, faculty directories, graduate program administrator listings, and instructional staff profiles to identify individuals most likely to have knowledge of course offerings. Depending on the structure of each department, emails were directed to course instructors, department chairs, program coordinators, graduate advisors, or general departmental accounts.

Beginning on August 18, 2023, and concluding on December 8, 2023, research staff sent up to 4 email communications to relevant instructors and administrative staff to solicit course information. The communication provided an overview of the study and requested (1) syllabi for targeted courses, (2) course descriptions or lists of required graduate courses, and (3) information about formal and informal training opportunities, such as workshops, seminars, journal clubs, or other instructional activities. Recipients were also invited to forward the request to colleagues who might have access to course materials or additional information.

Syllabi Analysis

After collecting the course syllabi, the content was analyzed to explore the breadth of learning objectives and course topics related to digital health research ethics. Learning objectives, also known as learning goals or outcomes, are statements that define what students should learn or be able to do after completing a course or activity. Course topics refer to the sequence of subject areas covered during a course, often presented along with a schedule for the course lectures and assignments or summarized in a course description. To analyze the collection of course syllabi, our study applied an approach similar to recent studies of courses in technology ethics [17] and social computing [18]. Specifically, 3 researchers reviewed each syllabus to identify learning objectives and course topics. Initial coding was conducted independently by multiple members of the research team, followed by regular consensus-based meetings to review interpretations, resolve discrepancies, and refine analytic decisions. To identify themes, the research team applied an affinity diagramming method [19] to qualitatively organize the learning objectives and course topics into hierarchical clusters. Members of the research team met twice per week for 6 weeks using FigJam (Figma Inc) [20] to support collaborative synthesis. The resulting themes characterize explicit ethics-related knowledge conveyed through coursework and are later interpreted alongside interview findings.

Interview Recruitment and Procedures—Tacit Knowledge

Semistructured interviews were conducted between August 17 and September 7, 2023. Recruitment involved outreach to a pool of 32 eligible individuals, with up to 3 contact attempts per person. Recruitment emails summarized the study purpose and the general topical domains from which participants could select areas they felt most prepared to discuss. Authors were eligible to participate if they were corresponding or lead authors on included publications and could speak to training, ethical decision-making, and research team practices in digital health research. A total of 6 individuals consented to participate and completed the interview. All interviews were analyzed. Together, the 6 interviewees had 516 publications and 85,252 citations, reflecting their standing in the field.

Eligible researchers received an email invitation describing the study purpose, interview procedures, and plans to audio-record the session. Interviews were conducted via Zoom (Zoom Communications) by a trained graduate student researcher, with a supervising team member present to support consistency in

XSL•FO
RenderX

procedures. Verbal informed consent was obtained at the beginning of each interview. All interviews were audio-recorded with participant consent and transcribed using Zoom's built-in transcription feature.

The semistructured interview guide was informed by our prior literature review and publication-specific analyses of sociotechnical health research practices (forthcoming). Questions were organized around the following five themes identified in our prior analyses: (1) prototyping designs, including how research teams collaboratively prototype systems, protocols, and materials; (2) engaging stakeholders, including how stakeholders are involved throughout the research life cycle; (3) developing informed consent procedures and data protections, including how teams design consent processes and data safeguards; (4) mentoring research, including the role of mentorship in training and guiding research practices; and (5) sharing research, including how research teams disseminate findings and return results to participants and stakeholders. The guide included open-ended prompts within each theme to elicit concrete examples from participants' experiences. Experts were invited to select two to three themes they felt most prepared to discuss.

Interview Transcripts Analysis

An iterative, inductive thematic approach was used to analyze interview transcripts. Although the analysis was inductive, interpretation was informed by a digital health research ethics perspective, which sensitized the team to ethical considerations arising from the design and use of engineering-based digital health systems within research and institutional contexts. Following each interview, researchers drafted analytic memos documenting preliminary impressions, emerging concepts, and notable patterns [21]. These memos and associated transcript excerpts were reviewed by supervising researchers to ensure accuracy, deepen interpretation, and maintain analytic rigor.

Building on this initial memoing process, 1 researcher conducted detailed inductive coding to identify recurring concepts, and a second researcher independently reviewed these summaries and coded transcript segments to refine and validate emerging interpretations. The student interviewers and supervising researchers met regularly to discuss analytic decisions, resolve discrepancies through consensus, and iteratively refine the thematic structure; four researchers contributed to the overall analysis. Because the full set of interviews represented a small sample of senior digital health researchers, the team analyzed all available data and refined themes until shared conceptual coherence was reached through regular consensus meetings. Topics from the interview guide are summarized with example prompts in Textbox 1.

Textbox 1. Interview focus areas and example prompts.

Ethics-related research practices discussed during interviews

- Prototyping technology for health research
- Engaging with community stakeholders
- Developing informed consent procedures and data protection policies
- Sharing research data and findings

Reflective prompts about field development and ethics priorities

- How do you find and connect with other researchers in your field?
- Where do you look for funding for your research?
- How do you see this research field evolving over the next few years?
- What do you see as key research ethics topics to include in training programs?
- Who sets the guidelines and standards for ethics training in your field?

Results

Overview

Across our analysis of 98 course syllabi and 6 expert interviews, we identified seven themes that characterize how ethical competencies in digital health research are taught and learned: (1) broad understanding of ethical principles, (2) applied ethics, (3) research in context, (4) ethics in technology design, (5) anticipating the future, (6) communication practice, and (7) professional responsibility. Some conceptual overlap across themes occurs across communication, team collaboration, and professional responsibility. This is expected given the sociotechnical context of digital health research. Ethical competence is developed and demonstrated across interpersonal interaction, team workflows, and organizational norms, as these are interdependent domains. The sections that follow describe the programs along with each theme and illustrate how explicit and tacit learning opportunities manifest across programs and research environments.

Academic Programs

The academic programs identified through the search reflect a breadth of disciplines. A majority (n=48, 77%) of the universities represented by the syllabus collection are located within the United States, with 14 (22%) located internationally. The 132 courses include graduate-level instruction for doctoral and medical school students as well as students in a range of

XSL•FO
RenderX

master's degree programs (Table 1). Courses were open to undergraduate and graduate students (n=24, 18%), with the majority (n=64, 48%) limited to undergraduate enrollment.

Table 1. Total courses by academic discipline based on the name of the university department.

Academic discipline	Count, n
Bioengineering	2
Biology	2
Cognitive science	2
Communication	2
Computer science	62
Digital media	8
Education	1
Engineering	10
Global health	1
Human-centered computing	3
Industrial design	1
Informatics	18
Information science	11
Innovation	1
Nursing	1
Pediatrics	1
Philosophy	1
Public policy	1
Sociology	1
Total	132

Sample Size

We analyzed 98 syllabi that included learning objectives to identify explicit ethics-related learning opportunities and identified 7 themes, as summarized in Table 2. Within the 98 syllabi analyzed, 5 universities provided instruction for 5 of the 7 themes, whereas 12 universities incorporated only 1 (19%) theme. None of the academic programs offered a course series that encompassed all 7 learning-objective themes. The following sections describe each theme. Additional examples of learning objectives and theme-specific references are provided in Multimedia Appendix 1.

XSL•FO
RenderX

Table 2. Descriptions of themes with the number of learning objectives and courses represented in each theme.

Theme descriptions	Objectives, n	Courses (n=98), n (%)
Theme 1: broad understanding. Gain a broad understanding of ethical principles and guidelines [22-25]		
Learn to apply ethical principles in research, in general	20	18 (18.4)
Explore how ethical principles have been applied in different contexts	33	18 (18.4)
Summary	53	31 (31.6)
Theme 2: applied ethics. Practice applying ethical principles to real-world cases and scenarios [26-29]		
Identify health informatics issues from various perspectives	12	4 (4.1)
Identify ethical challenges related to technology	42	27 (27.6)
Identify ethical considerations related to privacy and security	17	13 (13.3)
Learn about ethical practices in data science	28	19 (19.4)
Summary	99	43 (43.9)
Theme 3: research in context. Build research skills necessary to investigate ethical questions within specific context [30-33]		
Explain uses of AI[a]	19	7 (7.1)
Learn about data science methods	23	13 (13.3)
Describe security and privacy processes	37	13 (13.3)
Develop research and analysis skills	33	16 (16.3)
Summary	112	42 (42.9)
Theme 4: ethics in technology design. Learn how ethical principles can play into decisions about human-computer interaction [34-37]		
Develop foundational skills in the design process	28	16 (16.3)
Gain skills in human-computer interaction design	18	6 (6.1)
Learn about ethical design principles	19	7 (7.1)
Learn about the design of digital health systems	16	5 (5.1)
Summary	81	27 (27.6)
Theme 5: anticipating the future. Learn how to identify and assess ethical issues tied to emerging and futuristic technologies [38-46]		
Explore diverse perspectives in technology	15	11 (11.2)
Explore the societal impact of health technology	14	7 (7.1)
Explore the societal impact of novel and future technology	24	21 (21.4)
Explore ELSIs[b] of technology	36	16 (16.3)
Introduced to an overview of computing research	33	19 (19.4)
Gain basic knowledge in business	1	1 (1)
Summary	123	50 (51)
Theme 6: communication practice. Develop skills to communicate ethical concerns in group discussions, presentations, and written reports [47-51]		
Develop professional communication skills	26	18 (18.4)
Learn about engaging with diverse audiences	22	18 (18.4)
Developing critical thinking and reasoning skills	17	13 (13.3)
Summary	65	42 (42.9)
Theme 7: professional responsibility. Understand the ethical and professional standards expected of researchers conducting research in the public interest [52-56]		
Describe ethical and professional responsibilities in computing	17	14 (14.3)
Learn about professional ethics in workplaces	11	7 (7.1)
Develop teamwork and team management skills	13	10 (10.2)
Summary	41	28 (29)

XSL•FO
RenderX

[a]AI: artificial intelligence.

[b]ELSI: ethical, legal, and social implications.

Theme 1: Broad Understanding. Principles and Guidelines (eg, Bioethics and Human Rights)

This theme captures how courses orient students to ethical traditions (bioethics, research ethics, and rights-based perspectives) and institutional guardrails (eg, IRB training). The emphasis is on mastering shared vocabulary and analytic lenses that later anchor applied judgments in digital health contexts.

Our analysis found that a third of all courses (31/98) include learning objectives that emphasize training in diverse philosophical approaches to ethical thinking (Table 2, Theme 1). As summarized in Table 2 (Theme 1), foundational objectives appear across a substantial subset of courses but vary in depth from survey-style introductions to more rigorous treatments that require analytic writing and IRB certification. Programs rarely sequence these foundations with later, hands-on components, creating discontinuities that reappear in downstream themes. One course articulates this by stating that students will “develop a working understanding of major traditions of philosophy and ethical theory to provide foundational tools of analysis.” (C108)

Foundational principles serve as a grounding reference point for decisions about informed consent, risk assessment, and mitigation, and expectations for fairness as digital health systems are implemented in new settings. Differences in depth point to gaps in scaffolding from foundational principles to applied and design contexts. These foundational principles matter because they establish the ethical vocabulary and conceptual grounding students later rely upon when making context-specific decisions in data-intensive health environments. Without this grounding, students may struggle to recognize how abstract principles translate into concrete ethical challenges that emerge in sociotechnical systems.

Theme 2: Applied Ethics. Practice Applying Ethical Principles to Real-World Sociotechnical Contexts

Courses in this theme move beyond abstraction and ask students to apply ethical principles to concrete cases involving algorithmic decision-making, data privacy and security, and responsible AI deployment in health contexts. Applied ethics is positioned as a bridge between frameworks and situated decisions made during design, analysis, and deployment, with related learning objectives appearing in 43 (44%) of 98 courses. These courses emphasized students' ability to analyze ethical challenges (ie, information storage, exchange, security, and privacy) across various contexts, such as health informatics, data science, and computer science (Table 2, Theme 2). For example, “Apply the philosophies and theories covered to computer science problems and scenarios, including risks associated with large language models and algorithmic decision-making” (C55). As shown in Table 2 (Theme 2), applied ethics objectives are prevalent yet unevenly embedded. Some courses integrate iterative casework and debate across the term; others contain a single module within broader technical instruction. Assignments commonly require students to reason through value trade-offs across stakeholders and technical constraints, making ethical salience visible at practical decision points.

Applied ethics training equips students to recognize and navigate tensions where harms materialize, such as in data curation, model evaluation, and deployment pipelines. This applied focus is essential in digital health because ethical tensions often arise at the level of implementation, such as privacy trade-offs in data pipelines or fairness issues in algorithmic modeling. These courses help students identify where harms may materialize and how to reason through competing stakeholder values, trade-offs in data pipelines, or fairness issues in algorithmic modeling.

Theme 3: Research in Context. Methods to Study Sociotechnical Interactions

This theme reflects the pairing of technical content (AI/machine learning, security, and data science) with research design and analysis (literature reviews, proposal writing, and qualitative/quantitative reasoning) so students can empirically study sociotechnical systems in health. The focus is not coding per se, but the methodological capacity to investigate ethical questions in context. Results reveal 43% (42/98) of our course syllabi (Table 2, Theme 3) depict broad coverage of research skills objectives. We found that some syllabi emphasize technical fluency, while others center on social scientific inquiry. Few programs explicitly connect these skill sets to institutional realities (eg, governance and clinical workflows), a linkage that becomes crucial in later themes. As an example, C66 states, “This course covers the fundamental principles of the scientific method, and the core skills of planning, designing, executing, evaluating, and presenting research.”

Blending computing and social science methods reveals how risks emerge from interactions across data, tools, institutions, and people; however, explicit integration of where risks and benefits can accrue across clinical and organizational constraints appears inconsistent. Syllabi linked to this theme aim to develop competence by enabling students to investigate how risks, biases, and unintended consequences occur, reflecting the sociotechnical nature of digital health systems. Understanding context is critical for anticipating where ethical safeguards may fail.

Theme 4: Ethics in Technology Design. Learn How Ethical Principles Can Play Into Decisions About Human-Computer Interaction

Courses link values (accessibility, equity, and sustainability) to concrete design choices for digital health systems. Here, ethics is framed as an up-front design requirement, which may include incorporating inclusive methods, accessibility standards, and sensitivity to domain-specific issues such as digital biomarkers or health disparities, rather than a post hoc audit. One effective way to bridge the gap between ethical principles in theory and the practice of system design is through the incorporation of human-centered design (HCD) and human-computer interaction

XSL•FO
RenderX

concepts. This includes examining "how digital health technologies can be designed to achieve health equity [with a] central focus on applying human-centered design concepts to the domain of digital health equity" (C19). Our analysis found that 27% (n=26) of the courses teach students HCD processes (Table 2, Theme 4). Unlike Theme 3, courses in Theme 4 prioritize understanding system design concepts, though systems are also often incorporated into specific research methods.

Introducing criteria into design processes that increase awareness of ethical practices can help teams to operationalize what respect for persons and justice look like in the research design phase. Programs that connect HCD/human-computer interaction to clinical governance appear better positioned to align design intent with downstream demonstration of ethical practices.

Explicit and Tacit Knowledge About Research Ethics Through Curriculum and Hands-on Training

As described in the Methods section, digital health research experts, selected based on their prior citation frequency and prominence within the digital health research community, were interviewed to gather their perspectives on the digital health field and to speculate about the ethics training students will need in the future. Quotes from the interviews are uniquely identified by code, P1-P6, with each code corresponding to a specific expert. Because the interview data aligned closely with Themes 5-7, the following theme descriptions integrate evidence from both course syllabi and experts' perspectives.

Theme 5: Anticipating the Future. Learn How to Identify and Assess Ethical Issues Tied to Emerging and Futuristic Technologies

Courses in Theme 5 ask students to anticipate ethical, legal, and social trade-offs of emerging or speculative technologies before systems are deployed, with attention to governance, unintended consequences, and positionality. As summarized in Table 2 (Theme 5) , this theme appears in 50 (51%) of 98 syllabi, making it the most prevalent in the corpus; yet, programs vary in how anticipation is taught. Some emphasize legal and policy analysis, others use scenario debates and stakeholder role-play to surface value trade-offs. One course, for example, requires students to "propose recommendations...that represent the perspectives of stakeholders to develop a resolution." (C2).

Interview accounts show how this anticipatory work becomes a tacit judgment in practice. Researchers described teaching students to reason through downstream risks under uncertainty , particularly around data practices that are easy to overlook but consequential for participants. For instance, with passive sensing of GPS data, students are asked to grapple with consent memory and continuous collection: "it gets turned on, and it runs...continues to pull those data until the person uninstalls the app.... How do you manage that?" (P2). Teams use these moments to clarify what to do when participants later realize some data feel too sensitive to be shared, and to revisit data sharing procedures so that participants are aware of protections throughout a study.

Senior researchers also coach students to communicate the salience of risk when peers view data as "just data." As one expert noted, "Students might ask, 'This is just data; why are you making it so hard to access?' I bring examples of how location data alone can be very compromising or reveal intimate parts of someone's life.... It can have serious consequences if used beyond consent." (P6). Another expert emphasized that even deidentified datasets can create reidentification risk when clinical features are unique, requiring thoughtful protection while retaining research value (P4). In parallel, experts voiced concern that AI/large language model adoption in health research is moving faster than frameworks and training, urging explicit conversations about safe, responsible use. As one explained, "we should...figure out ways to talk about how to use these technologies in a safe, ethical, and responsible way." (P6).

Anticipatory ethics is particularly important in digital health, where emerging technologies outpace regulation and policy. Learning objectives that aim to prepare students to spot plausible harms and governance gaps before deployment, but curricula differ in modality and depth. Interview evidence shows how mentors transform anticipation into on-the-ground judgment, teaching students to (1) notice subtle data risks (eg, passive GPS), (2) communicate why those risks matter to participant welfare, and (3) revisit data management choices as technologies and contexts evolve.

Theme 6: Communication Practice. Develop Skills to Communicate Ethical Concerns in Group Discussions, Presentations, and Written Reports

Theme 6 centers on the individual's capacity to surface and reason about ethical issues through argumentation, critique, writing, and public speaking. As shown in Table 2 (Theme 6), communication-focused learning objectives appear in 42 (43%) of the 98 syllabi, but their sophistication varies. This spans from basic presentation mechanics to rigorous argumentation that engages counterpositions with evidence and adapts messages to different audiences.

Many courses build these competencies by having students read extensively, write analytically, and lead structured discussions. One syllabus puts it plainly: students should "read a lot, write a fair amount, and become more comfortable participating in oral discussions and giving an oral presentation." (C49). Another emphasizes the standard to which this participation is held: "articulate your ethical beliefs precisely and persuasively." (C116). Together, these objectives position rhetoric as the mechanism by which ethical reasoning becomes public, contestable, and actionable in research settings.

Communication is foundational in ethical practice. Ethical issues in digital health typically surface through dialogue ; these syllabi describe tools to practice talking about ethical challenges. The interview narratives show those tools at work explaining risk, negotiating disagreement, and recording decisions so teams remain accountable to participants and stakeholders.

Theme 7: Professional Responsibility. Understand the Ethical and Professional Standards Expected of Researchers

This theme focuses on organization-level norms, including professional codes, role clarity, escalation paths for ethical

XSL•FO
RenderX

concerns, cross-disciplinary collaboration, and stewardship responsibilities for managing volumes of individual-level sensitive data. As indicated in Table 2 (Theme 7), such objectives appear in 28 (29%) of 98 syllabi, less frequently than communication or anticipation themes. Syllabi in this theme ask students to analyze professional association policies and to consider how responsibilities translate into daily practice in research teams and health organizations; for example, "What special responsibilities do we have as computing professionals? How can we apply the Software Engineering Code of Ethics and ACM Code of Ethics in our daily practice?" (C21).

Tacit practices, described by experts, make the organizational mechanics of responsibility concrete. Senior researchers emphasized modeling norms and setting expectations that are then diffused through laboratories and professional communities. For example, "we have to show that this is the way we do the work." (P5). Experts also highlighted the need to work across disciplines and with stakeholders (eg, clinicians, patients, and community partners) so that design decisions reflect health system realities and community values; collaboration remains essential but complex, requiring time, care, and clear alignment with funder constraints (P4 and P1).

Experts train laboratory members to conduct stakeholder-engaged digital health research, which includes building relationships, listening across perspectives, and giving back to communities. These are competencies that extend professional responsibility beyond the laboratory into the settings where digital health systems are designed and implemented (P4 and P5).

Professional responsibility is enacted through governance structures and norms (codes, roles, and processes) and through collaboration with diverse stakeholders. Courses introduce these expectations; interviews show how senior researchers make governance visible in day-to-day work, translating ethical intent into accountable team practices.

Discussion

Principal Findings

Digital health research operates within complex sociotechnical systems [57], making ethical competence foundational to responsible innovation. This study examined how research ethics education prepares trainees for the sociotechnical realities of digital health research. Findings indicate that ethical competence develops through the interaction of classroom instruction and experiential learning; yet, gaps remain in translating abstract principles into research practice. Addressing these gaps will require instructional strategies that more intentionally integrate ethics training with the collaborative, data-intensive environments in which digital health research occurs.

Research ethics education in digital health relies on both explicit instruction and tacit learning. Classroom teaching introduces guiding ethical principles, analytic tools, and case studies [58], while participation on research teams enables students to identify, discuss, and resolve ethical issues as they arise [59]. Experts in our study described mentoring trainees through decisions affecting participant data privacy, allowing students to observe how choices related to data storage, sharing, and stewardship unfold across a study. These complementary experiences reinforce the reciprocal relationship between teaching and research practice.

Our thematic results illustrate how ethical competence in digital health research emerges through sociotechnical interactions rather than individual reasoning alone. For example, the "applied ethics" and "research in context" themes show how ethical issues arise within data pipelines, modeling practices, and clinical workflows, where technical and organizational factors intersect. Themes such as "communication practice" and "professional responsibility" highlight how interpersonal coordination, role clarity, and governance structures shape ethical action within research teams. Our results confirm that ethics education must prepare trainees to navigate these interconnected sociotechnical dynamics rather than isolated ethical dilemmas.

Ethics Education as a Sociotechnical Process

Our findings reveal a persistent sociotechnical tension in ethics education. Coursework often emphasizes individual reasoning and abstract principles, whereas ethical challenges in digital health research are collective and infrastructural, emerging through team workflows, data pipelines, institutional requirements, and evolving stakeholder expectations. Ethical competence, therefore, extends beyond cognitive skill to a situated practice shaped by organizational routines, technical affordances, and supervisory norms [60]. Senior researchers play an important role by modeling everyday ethical decision-making and making these processes visible, enabling norms to diffuse across laboratories and professional communities [61]. Cross-training activities, such as journal clubs, further support this process by cultivating a shared vocabulary for reasoning about risks, roles, and constraints.

Because digital health research frequently involves health care systems and community partners, stakeholder communication and team science skills are prioritized in both coursework and experiential training [62]. Experts reported connecting trainees with community collaborators to build practical experience working across sectors, while classroom exercises in perspective-taking and public speaking reinforce these competencies. Such preparation is critical for sustaining multipartner research, where navigating disagreement is often central to maintaining trust.

Despite these complementary pathways, reliance on experiential learning raises concerns about consistency in training. Ethical norms are frequently transmitted within research groups, meaning that uneven access to hands-on sociotechnical decision-making could reinforce inequities in students' ability to develop ethical judgment [63]. Our syllabus analysis supports this interpretation in that courses commonly address privacy, fairness, and professional responsibilities; however, fewer examine how these principles are enacted within algorithmic pipelines, wearable sensing systems, or clinical governance processes. This gap helps explain why experts view experiential learning as essential for translating abstract principles into practice while underscoring the need for more intentional instructional supports.

XSL•FO
RenderX

Expanding Experiential Learning Through Instructional Infrastructure

One strategy for expanding access to practice-based learning is to scaffold undergraduate participation through a structured critique modeled after the ethics and society review [64]. Students who have completed research ethics coursework could contribute to proposal review in bounded, time-limited ways. Although ethics and society review relies on expert reviewers and may be difficult to scale, prior work in human-computer interaction demonstrates that scaffolding approaches, such as design critique and historical analysis, help novice learners complete complex evaluative tasks while strengthening ethical reasoning [59,65,66]. Instructors could adopt similar workflows by training students to collaboratively analyze case studies [67]. However, implementation challenges remain, including limited institutional support, uneven student preparedness for complex sociotechnical contexts, and the risk that simplified workflows may obscure ethical tensions. Addressing these constraints will require continued evaluation and the development of shared educational resources supported by professional associations and research sponsors.

Translating Tacit Insights Into Shareable Resources

Beyond scaffolding, our findings suggest the importance of translating practice-based insight into reusable instructional materials. Tacit knowledge develops through experience and is often conveyed as intuition, making it difficult to formalize. Yet ethical concerns that emerge during research offer opportunities for documentation. Existing tools, such as ethics checklists (eg, the Digital Health Checklist [68]) and design artifacts such as card decks that embed ethical principles into computing and AI research [69,70] demonstrate how abstract guidance can be externalized into prompts that support reflection in practice.

Building on this approach, researchers could create brief ethics discussion prompts when salient concerns arise during a study. Drawing on Parker's framework [71], such prompts would include a problem statement, a description of the solution space, and a guiding question, accompanied by reflection on how the issue was addressed. Unlike traditional case studies, which often isolate single decisions, these prompts could capture ethical judgment as it unfolds within research workflows. When shared alongside research outputs, they have the potential to convert tacit experience into concise, transferable instructional resources. Adoption, however, raises important questions about documenting sensitive issues, ownership of materials, and incentives for participation, which create tensions that must be addressed to normalize reflective ethical practice in digital health research.

Theoretical Implications for Digital Health Research

Our findings reconnect research ethics training to sociotechnical theory by illustrating that ethical competence emerges through the interaction of data infrastructures, research workflows, and institutional expectations rather than through formal instruction alone. Experts described guiding trainees through real-time decisions about passive data collection, such as GPS and sensor-based monitoring, where ethical judgment developed in relation to both continuous data flows and governance requirements surrounding consent and stewardship. Viewing ethics education through a sociotechnical lens shifts the focus from individual responsibility toward the systems and environments that shape ethical action. By foregrounding ethics education as a sociotechnical phenomenon, this study extends prior work that frames ethical decision-making as an individual competency. Instead, our findings suggest that ethical capacity is distributed across training environments, research infrastructures, and supervisory practices, underscoring the importance of designing educational systems that support ethical action.

Implications for Research Ethics Training in Digital Health

Responding to calls for more practice-oriented ethics training in digital health, this study makes 3 primary contributions. First, it provides a thematic analysis of explicit ethics instruction across 98 syllabi representing 76 international academic programs, offering one of the most comprehensive mappings of current curricular approaches. Second, it advances understanding of how trainees develop ethical competence by identifying experiential pathways through which tacit knowledge is cultivated in research settings. Third, it proposes actionable strategies for strengthening instructional infrastructure, including closer integration of classroom learning with research practice, scaffolded opportunities for student participation, educator-researcher learning communities, and the use of embedded ethics discussion prompts to support learning reflective practice.

Collectively, our findings position research ethics education not simply as a pedagogical challenge but as an instructional infrastructure issue within digital health research. Ethical competence is shaped by the environments in which researchers are trained, suggesting that strengthening this infrastructure requires closer integration of classroom learning with structured experiential opportunities, expanded scaffolded participation in ethical review, and mechanisms for translating practice-based insight into shared instructional resources. Advancing these efforts could promote more consistent training, reduce inequities in access to experiential learning, and better prepare the next generation of researchers to navigate the sociotechnical complexities of data-intensive health research.

Limitations

This study has several limitations related to our sampling strategy. Digital health remains an emerging sociotechnical field with few formal degree programs and no standardized course titles (eg, "digital health research ethics"), making traditional catalog searches impractical. We therefore identified researchers and affiliated programs through Association for Computing Machinery publications to center the study on active research environments; however, this approach may have excluded relevant curricula at teaching-focused colleges and medical schools that prioritize instruction over research. Our sample also reflects a geographic concentration in the United States and Europe, introducing a potential Western bias. Teaching priorities and approaches to ethics education may differ globally, highlighting the need for more internationally

XSL•FO
RenderX

representative research. Additionally, syllabi provide insight into intended instruction but do not capture how courses are experienced by instructors or students. Although interviews with senior researchers offered perspectives on mentoring and training, we did not include trainees or individuals who left academic pathways, limiting our understanding of how ethics education is enacted and received. Future research should examine classroom practices and team-based research environments to better identify effective approaches to teaching ethical digital health research.

Conclusion

Our study demonstrates that ethics education and learning emerge through the interaction of formal coursework and practice-based training, each cultivating distinct capabilities needed to navigate data-intensive and collaborative research environments. While these complementary pathways help prepare students to recognize ethical issues, apply principles in context, anticipate emerging risks, and communicate across disciplines, access to experiential ethics training remains uneven. Strengthening ethics education will therefore require expanding structured opportunities for early research engagement, fostering communities of practice among educators and researchers, and developing mechanisms that translate tacit ethical reasoning into shareable resources. Designing training environments that integrate ethical reflection into everyday research practice may better equip the next generation of digital health researchers to responsibly shape the sociotechnical health systems they both study and build.

Acknowledgments

The authors wish to thank Sneha Jaiswal for her contributions to the research process and to Daniela G Vital for her administrative program management over the course of the project.

Funding

This material is based upon work supported by the National Science Foundation under grant 2124975. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: CN (lead), BM (equal)
Data curation: YZ, BM, TP
Formal analysis: BM, YZ, TP (supporting)
Funding acquisition: CN (lead)
Investigation: YZ, TP, BY
Methodology: BM
Project administration: BY (lead)
Resources: CN
Supervision: BM, CN, BY
Validation: YZ, TP
Visualization: YZ (lead), TP (supporting)
Writing—original draft: YZ (lead), CN, BY (supporting)
Writing—review & editing: YZ, BM, TP, BY, CN

Conflicts of Interest

None declared.

Multimedia Appendix 1
Readings to support themes identified in the Results section.
[DOCX File , 40 KB - jmir_v28i1e82861_app1.docx]

References

1. Lupton D. Critical perspectives on digital health technologies. Sociol Compass 2014;8(12):1344-1359. [doi: 10.1111/soc4.12226]
2. Topol E. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. USA: Basic Books, Inc; 2019.

XSL•FO
RenderX

3. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017;19(11):e8775 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
4. Sittig DF, Singh H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. Qual Saf Health Care 2010;19 Suppl 3(Suppl 3):i68-i74 [FREE Full text] [doi: 10.1136/qshc.2010.042085] [Medline: 20959322]
5. Wilcox L, Brewer R, Diaz F. AI consent futures: a case study on voice data collection with clinicians. Proc ACM Hum-Comput Interact 2023;7(CSCW2):1-30. [doi: 10.1145/3610107]
6. Hernandez-Cruz N, Garcia-Constantino M, Beltran-Marquez J, Cruz-Sandoval D, Lopez-Nava I, Cleland I, et al. Study design of an environmental smart microphone system to detect anxiety in patients with dementia. USA: Association for Computing Machinery; 2019 Presented at: Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare; 2029 May 20; New York, NY, USA p. 383-388. [doi: 10.1145/3329189.3329234]
7. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
8. Dimond JP, Fiesler C, Bruckman AS. Domestic violence and information communication technologies. Interact Comput 2011;23(5):413-421. [doi: 10.1016/j.intcom.2011.04.006]
9. Saleheen N, Ullah MA, Chakraborty S, Ones DS, Srivastava M, Kumar S. WristPrint: characterizing user re-identification risks from wrist-worn accelerometry data. Conf Comput Commun Secur 2021;2021:2807-2823 [FREE Full text] [doi: 10.1145/3460120.3484799] [Medline: 36883116]
10. Bloss C, Nebeker C, Bietz M, Bae D, Bigby B, Devereaux M, et al. Reimagining human research protections for 21st century science. J Med Internet Res 2016;18(12):e329 [FREE Full text] [doi: 10.2196/jmir.6634] [Medline: 28007687]
11. Dunseath S, Weibel N, Bloss CS, Nebeker C. NIH support of mobile, imaging, pervasive sensing, social media and location tracking (MISST) research: laying the foundation to examine research ethics in the digital age. NPJ Digit Med 2018;1(1):20171 [FREE Full text] [doi: 10.1038/s41746-017-0001-5] [Medline: 31304345]
12. Shilton K, Greene D. Linking platforms, practices, and developer ethics: levers for privacy discourse in mobile application development. J Bus Ethics 2017;155(1):131-146. [doi: 10.1007/s10551-017-3504-8]
13. McInnis BJ, Pindus R, Kareem D, Vital DG, Hekler EB, Nebeker C. Factors influencing informed consent preferences in digital health research: survey study of prospective participants. J Med Internet Res 2025;27:e63349 [FREE Full text] [doi: 10.2196/63349] [Medline: 39847412]
14. McInnis BJ, Pindus R, Kareem D, Nebeker C. Considerations for the design of informed consent in digital health research: participant perspectives. J Empir Res Hum Res Ethics 2024;19(4-5):175-185 [FREE Full text] [doi: 10.1177/15562646241290078] [Medline: 39397541]
15. McInnis BJ, Pindus R, Kareem D, Gamboa S, Nebeker C. Exploring the future of informed consent: applying a service design approach. Proc ACM Hum-Comput Interact 2024;8(CSCW1):1-31. [doi: 10.1145/3637330]
16. Nebeker C, Harlow J, Espinoza Giacinto R, Orozco-Linares R, Bloss CS, Weibel N. Ethical and regulatory challenges of research using pervasive sensing and other emerging technologies: IRB perspectives. AJOB Empir Bioeth 2017;8(4):266-276. [doi: 10.1080/23294515.2017.1403980] [Medline: 29125425]
17. Fiesler C, Garrett N, Beard N. What do we teach when we teach tech ethics? A syllabi analysis. USA: Association for Computing Machinery; 2020 Presented at: Proceedings of the 51st ACM Technical Symposium on Computer Science Education; 2020 Feb 26; New York, NY, USA p. 289-295. [doi: 10.1145/3328778.3366825]
18. Delcourt C, Venkatagiri S, Chandrasekharan E. What's in a social computing course: analyzing computer and information science syllabi. USA: Association for Computing Machinery; 2024 Presented at: Proceedings of the 6th Annual Symposium on HCI Education; 2024 Jun 05; New York, NY, USA p. 1-8. [doi: 10.1145/3658619.3658623]
19. Holtzblatt K, Beyer H. The affinity diagram. In: Holtzblatt K, Beyer H, editors. Context Des Second Ed Boston: Morgan Kaufmann. Context Des Second Ed Boston: Morgan Kaufmann; 2017:127-146.
20. FigJam: the online collaborative whiteboard for teams. Figma. 2025. URL: https://www.figma.com/figjam/ [accessed 2025-08-08]
21. Birks M, Chapman Y, Francis K. Memoing in qualitative research: probing data and processes. J Res Nurs 2008;13(1):68-75. [doi: 10.1177/1744987107081254]
22. Integrity in scientific research video series. American Association for the Advancement of Science. URL: https://www.aaas.org/resources/integrated-public-use-microdata-series-ipums [accessed 2026-03-10]
23. Jones CP. Confronting institutionalized racism. Phylon (1960) 2002;50(1/2):7. [doi: 10.2307/4149999]
24. Ess C. Digital media ethics. In: Oxford Research Encyclopedia of Communication. Oxford, UK: Oxford University Press; 2017.
25. Responsible conduct of research training. NIH Office of Intramural Research. URL: https://oir.nih.gov/sourcebook/ethical-conduct/responsible-conduct-research-training [accessed 2026-03-10]
26. Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, et al. The moral machine experiment. Nature 2018;563(7729):59-64. [doi: 10.1038/s41586-018-0637-6] [Medline: 30356211]

XSL•FO
RenderX

27. Kolata G. When doctors use a chatbot to improve their bedside manner. The New York Times. 2023. URL: https://www.nytimes.com/2023/06/12/health/doctors-chatgpt-artificial-intelligence.html [accessed 2026-03-10]
28. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464):447-453 [FREE Full text] [doi: 10.1126/science.aax2342] [Medline: 31649194]
29. Waldman AE. Cognitive biases, dark patterns, and the 'privacy paradox'. Curr Opin Psychol 2020;31:105-109. [doi: 10.1016/j.copsyc.2019.08.025] [Medline: 31590106]
30. Foxwell HJ. Creating Good Data: A Guide to Dataset Structure and Data Representation. New York, NY: Apress; 2020.
31. James G, Witten D, Hastie T, Tibshirani R. Introduction to Statistical Learning: With Applications in R. 2nd ed. New York, NY: Springer; 2021.
32. O'Neil C. The ethical data scientist. Slate. 2016. URL: https://slate.com/technology/2016/02/how-to-bring-better-ethics-to-data-science.html [accessed 2026-03-10]
33. Conklin AW. Principles of Computer Security: CompTIA Security+ and Beyond. 5th ed. New York, NY: McGraw-Hill Education; 2026.
34. Dix A. Human–computer interaction: a stable discipline, a nascent science, and the growth of the long tail. Interacting with Computers 2010;22(1):13-27. [doi: 10.1016/j.intcom.2009.11.007]
35. Garrett JJ. The Elements of User Experience: User-Centered Design for the Web and Beyond. 2nd ed. Berkeley, CA: New Riders; 2011.
36. Horton S, Quesenbery W. A Web for Everyone: Designing Accessible User Experiences. New York, NY: Rosenfeld Media; 2014.
37. Lazar J, Goldstein DF, Taylor A. Ensuring Digital Accessibility through Process and Policy. 1st ed. San Francisco, CA: Morgan Kaufmann Publishers Inc; 2015.
38. Barocas S, Selbst AD. Big data's disparate impact. Calif L Rev 2016. [doi: 10.15779/Z38BG31]
39. Calo R. Artificial intelligence policy: a roadmap. SSRN 2017. [doi: 10.2139/ssrn.3015350]
40. Friedler SA, Scheidegger C, Venkatasubramanian S. The (Im)possibility of fairness. Commun ACM 2021;64(4):136-143. [doi: 10.1145/3433949]
41. Kirkpatrick N. Making a living in the toxic world of discarded electronics. The Washington Post. 2015. URL: https://www.washingtonpost.com/news/in-sight/wp/2015/04/15/the-children-who-make-a-living-in-the-toxic-world-of-discarded-electronics [accessed 2026-03-10]
42. Noble SU. Algorithms of Oppression: How Search Engines Reinforce Racism. New York, NY: New York University Press; 2018.
43. Rotman D. The disparity between the rich and everyone else is larger than ever in the United States and increasing in much of Europe. Why? MIT Technology Review. 2014. URL: https://www.technologyreview.com/2014/10/21/170679/technology-and-inequality/ [accessed 2026-03-10]
44. Shaw J. Confronting pitfalls of machine learning, artificial intelligence. Harvard Magazine. 2018. URL: https://www.harvardmagazine.com/2018/12/artificial-intelligence-limitations [accessed 2026-03-10]
45. Wallach W, Allen C. Moral Machines: Teaching Robots Right from Wrong. Oxford, UK: Oxford University Press; 2009.
46. Winner L. Trust and terror: the vulnerability of complex socio‐technical systems. Sci Cult 2004;13(2):155-172. [doi: 10.1080/0950543042000226594]
47. Dainow DB. SIENNA D6.3: Methods for translating ethical analysis into instruments for the ethical development and deployment of emerging technologies. Zenodo 2021. [doi: 10.5281/ZENODO.5541538]
48. Niki A. How to nail the pitch: 17 storytelling tips for startups. The Founder Institute. URL: https://FI.co/insight/how-to-nail-the-pitch-17-storytelling-tips-for-startups [accessed 2026-03-10]
49. Michaelson HB. Creative aspects of engineering writing. IRE Trans Eng Writing Speech 1961;4(3):77-79. [doi: 10.1109/TEWS.1961.4322737]
50. Tebeaux E, Dragga S. The Essentials of Technical Communication. 5th ed. Oxford, UK: Oxford University Press; 2020.
51. Trimble J. Writing with Style: Conversations on the Art of Writing. 3rd ed. Munich: Pearson; 2010.
52. Gotterbarn D, Brinkman B, Flick C, Kirkpatrick MS, Miller K, Varansky K, et al. ACM code of ethics and professional conduct. Association for Computing Machinery. URL: https://www.acm.org/code-of-ethics [accessed 2026-03-10]
53. Compact between postdoctoral appointees and their mentors. Association of American medical colleges. Association of American Medical Colleges. 2017. URL: https://www.aamc.org/about-us/mission-areas/biomedical-research/post-doc-compact [accessed 2026-03-10]
54. Davis M. Thinking Like an Engineer: Studies in the Ethics of a Profession. 1st ed. New York, NY: Oxford University Press; 1998.
55. Hill LA. Becoming a Manager: How New Managers Master the Challenges of Leadership. Boston: Harvard Business Review Press; 2003.
56. National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs, Board on Higher Education and Workforce, Committee on Effective Mentoring in STEMM. In: Dahlberg ML, Byars-Winston A, editors. The Science of Effective Mentorship in STEMM. Washington, DC: National Academies Press; 2019.

XSL•FO
RenderX

57. Baxter G, Sommerville I. Socio-technical systems: from design methods to systems engineering. Interact Comput 2011;23(1):4-7 [FREE Full text] [doi: 10.1002/iub.1366] [Medline: 25904163]
58. Todd EM, Torrence BS, Watts LL, Mulhearn TJ, Connelly S, Mumford MD. Effective practices in the delivery of research ethics education: a qualitative review of instructional methods. Account Res 2017;24(5):297-321. [doi: 10.1080/08989621.2017.1301210] [Medline: 28332862]
59. Luther K, Tolentino J, Wu W, Pavel A, Bailey B, Agrawala M, et al. Structuring, aggregating, and evaluating crowdsourced design critique. USA: Association for Computing Machinery; 2015 Presented at: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing; 2025 Feb 28; New York, NY, USA p. 473-485. [doi: 10.1145/2675133.2675283]
60. Eraut M. Non-formal learning and tacit knowledge in professional work. Br J Educ Psychol 2000;70(1):113-136. [doi: 10.1348/000709900158001] [Medline: 10765570]
61. Lave J, Wenger E. Situated Learning: Legitimate Peripheral Participation. United Kingdom: Cambridge University Press; 1991.
62. Hall KL, Vogel AL, Huang GC, Serrano KJ, Rice EL, Tsakraklides SP, et al. The science of team science: a review of the empirical evidence and research gaps on collaboration in science. Am Psychol 2018;73(4):532-548. [doi: 10.1037/amp0000319] [Medline: 29792466]
63. Hafferty FW, Franks R. The hidden curriculum, ethics teaching, and the structure of medical education. Acad Med 1994;69(11):861-871. [doi: 10.1097/00001888-199411000-00001] [Medline: 7945681]
64. Bernstein MS, Levi M, Magnus D, Rajala BA, Satz D, Waeiss Q. Ethics and society review: ethics reflection as a precondition to research funding. Proc Natl Acad Sci USA 2021;118(52):e2117261118 [FREE Full text] [doi: 10.1073/pnas.2117261118] [Medline: 34934006]
65. Luther K, Pavel A, Wu W, Tolentino J, Agrawala M, Hartmann B, et al. CrowdCrit: crowdsourcing aggregating visual design critique. USA: Association for Computing Machinery; 2014 Presented at: Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing; 2014 Feb 15; New York, NY, USA p. 21-24. [doi: 10.1145/2556420.2556788]
66. Wang NC, Hicks D, Luther K. Exploring trade-offs between learning and productivity in crowdsourced history. 2018 Presented at: Proceedings of the ACM on Human-Computer Interaction; 2018 Nov 1; Canada p. 1-24. [doi: 10.1145/3274447]
67. Feldman MQ, McInnis BJ. How we write with crowds. 2021 Presented at: Proceedings of the ACM on Human-Computer Interaction; 2021 Jan 5; New York NY USA p. 1-31. [doi: 10.1145/3432928]
68. Nebeker C, Bartlett Ellis RJ, Torous J. Development of a decision-making checklist tool to support technology selection in digital health research. Transl Behav Med 2020;10(4):1004-1015 [FREE Full text] [doi: 10.1093/tbm/ibz074] [Medline: 31120511]
69. Hsieh G, Halperin BA, Schmitz E, Chew YN, Tseng YC. What is in the cardsxploring uses, patterns, trends in design cards. USA: Association for Computing Machinery; 2023 Presented at: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems; 2023 Apr 23; New York, NY, USA p. 1-18. [doi: 10.1145/3544548.3580712]
70. Sadek M, Constantinides M, Quercia D, Mougenot C. Guidelines for Integrating Value Sensitive Design in Responsible AI Toolkits. USA: Association for Computing Machinery; 2024 Presented at: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems; 2024 Feb 29; New York, NY, USA p. 1-20. [doi: 10.1145/3613904.3642810]
71. Parker WC. Public discourses in schools: Purposes, problems, possibilities. Educ Res 2006;35(8):11-18. [doi: 10.3102/0013189x035008011]

Abbreviations

AI: artificial intelligence
EQUATOR: Enhancing the Quality and Transparency of Health Research
HCD: human-centered design
IRB: institutional review board
SRQR: Standards for Reporting Qualitative Research

Edited by A Stone; submitted 22.Aug.2025; peer-reviewed by H Matisonn, S Arabi; comments to author 29.Oct.2025; accepted 18.Feb.2026; published 26.Mar.2026.

Please cite as:
Zhu Y, McInnis B, Punater T, York B, Nebeker C
Educating Students About Digital Health Research Ethics: Curricula Review and Expert Interview Study
J Med Internet Res 2026;28:e82861
URL: https://www.jmir.org/2026/1/e82861
doi:10.2196/82861
PMID:

XSL•FO
RenderX

©Yier Zhu, Brian McInnis, Tanya Punater, Brittany York, Camille Nebeker. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Factors Influencing Universal Coverage of AI-Assisted Cervical Cancer Screening: Qualitative Study Based on the Macro Model of Health System

Lu Ji[1*], PhD; Xinke Zhou[2*], MD; Lan Yao[2], PhD

[1]School of Nursing, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

[2]School of Medicine and Health Management, Tongji Medical College, Huazhong University of Science and Technology, 13 HangKong Road, Wuhan, China

[*]these authors contributed equally

Corresponding Author:
Lan Yao, PhD
School of Medicine and Health Management, Tongji Medical College, Huazhong University of Science and Technology, 13 HangKong Road, Wuhan, China

Abstract

Background: Improving screening coverage is a central goal of the global strategy to eliminate cervical cancer. In resource-constrained settings, insufficient service accessibility remains a key barrier to expanding coverage. Supported by artificial intelligence (AI)–assisted diagnostic technology, Hubei province has pioneered China's first provincial-level population-wide cervical cancer screening program, serving 12.67 million eligible women. This initiative provides an innovative practice for addressing such challenges.

Objective: This study systematically examines major factors influencing the achievement of universal screening coverage targets through interviews with core managers and implementers of Hubei province's screening program. It aims to provide empirical evidence and strategic recommendations for applying AI technologies in cervical cancer screening and enhancing screening coverage rates.

Methods: The interview guide was developed under the guidance of the macro model of health system. A combination of purposive sampling and multistage stratified sampling was used to capture provincial-level overviews and understand regional implementation variations, respectively. Guided by the macro model of health system, interview outlines were developed. Semistructured interviews were conducted between January and August 2024 with key project personnel (one per institution) from 14 relevant institutions. Interview data were analyzed using thematic analysis, with systematic coding and management facilitated by the NVivo software.

Results: Key informants reported that comprehensive screening has been largely achieved. The analysis identified government stewardship, AI-assisted screening technology, screening funding, and health literacy as the major factors for achieving universal screening coverage. Among these, government leadership and the application of AI-assisted diagnostic technologies provide significant driving factors. Additional factors encompassed structural dimensions, including multisectoral coordination, trained screening technicians, and information systems; process dimensions, such as institutional service delivery capacity, quality control measures, and community mobilization; along with outcome dimensions comprising population coverage, cytology positivity rate, follow-up, and treatment rate.

Conclusions: Achieving large-scale cervical cancer screening requires coordinated efforts across four dimensions: government stewardship, screening technology, screening funding, and health literacy. Government stewardship served as the core driver in advancing population-wide screening coverage. Its mechanisms included coordinated procurement of AI-assisted screening services, secured financial investment, formulation of targeted policies, promotion of multi-sectoral collaboration, and optimization of service delivery models. These efforts systematically improved the accessibility and utilization of screening services, ultimately encouraging and facilitating active participation among residents.

(J Med Internet Res 2026;28:e75372) doi:10.2196/75372

KEYWORDS

cervical cancer screening; screening; qualitative research; health systems; artificial intelligence; AI

XSL•FO
RenderX

Introduction

Cervical cancer remains one of the major health threats for women worldwide. In 2022, there were approximately 660,000 new cases globally, accounting for 3.3% of total cancer cases, with a mortality rate as high as 56% [1]. Data from the China Cancer Registry Annual Report 2022 show a cervical cancer incidence rate of 13.8 per 100,000 women, with approximately 150,000 new cases annually [2]. In recent years, both incidence and mortality rates have shown an upward trend [3]. The World Health Organization launched a global strategy in 2020 aiming to eliminate cervical cancer through human papillomavirus vaccination, cervical cancer screening, and treatment [4]. Cervical cancer screening is a key focus of policy interventions in China. The current screening rate is approximately 36.8%; however, China aims to achieve a 70% screening rate by 2030 [5].

Improving screening coverage depends on multiple factors, particularly in resource-limited settings where accessible and affordable screening technologies are critically needed [6,7]. Significant disparities in screening coverage exist across nations and regions [8-11], influenced by a combination of factors including sociocultural norms, economic status, health care policies, health awareness, educational resources, and technological capabilities [12-17]. In low- and middle-income countries (LMICs), women commonly face multiple barriers, including limited health literacy, inadequate access to services, insufficient funding, and scarce medical resources, which in turn constrain the improvement of screening coverage rates [18-21]. Both the World Health Organization and the National Institutes of Health recommend the use of artificial intelligence (AI) technology in resource-limited settings to effectively improve access to screening services.

AI-assisted diagnostic technology offers advantages in accessibility, cost, and diagnostic efficiency. Leveraging these strengths, China has implemented its first provincial-level cervical cancer screening program with universal coverage. The shortage of pathologists makes it difficult to implement traditional cytological screening effectively in primary health care institutions, limiting its capacity to adequately meet the population's screening needs [22,23]. In Hubei province, the adoption of AI-assisted cervical cancer screening technology achieved a coverage rate of 13.45% within 6 months. Calculated with the same number of pathologists, its efficiency was 87.5 times that of manual slide interpretation [24,25]. The AI-assisted approach showed remarkable diagnostic accuracy and cost-effectiveness, with an average per-person cost of 49 RMB [26-28]. This innovative technology has significantly enhanced screening coverage, particularly in resource-limited settings.

Current research on AI technology predominantly focuses on the diagnostic performance of screening techniques, with a notable scarcity of studies investigating its large-scale implementation and associated influencing factors [29-33]. The macro model of health system (MMHS) applies systems thinking to analyze the interplay and interdependence among various components within a health system, making it well-suited for policy interventions and health system diagnostics. This study developed interview guidelines based on the MMHS theoretical framework and conducted interviews with managers and health care providers involved in Hubei's cervical cancer screening program. The aim was to identify the factors influencing the achievement of universal coverage screening, thereby providing reference suggestions for the wider application of AI-assisted diagnostic technology. Given the core bottleneck of limited cervical cancer screening services due to insufficient pathologist resources in primary health care institutions, this study proposes the following exploratory hypothesis: the introduction of cost-effective and quality-reliable AI-assisted screening technology is expected to substantially overcome the primary barriers to achieving universal screening coverage, thereby accelerating the attainment of coverage targets. Meanwhile, maintaining an open research stance, we developed a theory-driven interview guide to capture the diverse complexities and obstacles that may arise during the implementation of universal coverage programs.

Methods

Overview

This study adopts a qualitative approach following Consolidated Criteria for Reporting Qualitative Research (COREQ), the standardized and widely recognized checklist for qualitative research (Checklist 1), and conducted semistructured interviews with managers and implementers of cervical cancer screening programs from 14 institutions. To systematically explore the factors influencing universal cervical cancer screening, the interview guide was developed based on the MMHS. Thematic analysis was performed following a 6-step analytic framework [34], through which textual data were integrated, categorized, and synthesized to ultimately identify the factors affecting the achievement of universal cervical cancer screening coverage. The researchers' reflections on the predefined statements are presented in detail in the *Discussion* section.

Macro Model of Health System

Framework

The macro model of health system (MMHS) is a theoretical and methodological framework that applies systems thinking to analyze the interactions and interdependencies among the various components within a health system and to reveal its operational patterns [35]. This model integrates systems science, complexity theory, and the theory of social determinants of health to comprehensively explain the combined impact of multiple factors on the health system. The model consists of 2 subsystems: an external submodel and an internal submodel. The external submodel encompasses external factors affecting health, such as politics, economy, culture, population needs, biology, environment, and behavior. The internal submodel represents the internal drivers of health, following the logical framework of structure-process-outcome (Figure 1). Within this framework, administrative, organizational, and resource factors at the structural level influence service delivery processes, which in turn affect system outcomes and ultimately determine health outcomes.

XSL•FO
RenderX

The MMHS is applicable to policy intervention research and health system diagnostics. The comprehensive cervical cancer screening program represents a concrete manifestation of public health policy. The implementation of cervical cancer screening programs is influenced by external factors, such as the policy environment and economic conditions, as well as internal factors including administrative processes, organizational structures, and resource allocation affecting service provision and utilization. These factors ultimately impact screening coverage and population cervical health outcomes. Therefore, this study uses the theoretical framework of the MMHS to design research tools and diagnose the factors influencing the implementation of universal cervical cancer screening.

Figure 1. Conceptual framework of the macro model of health system.

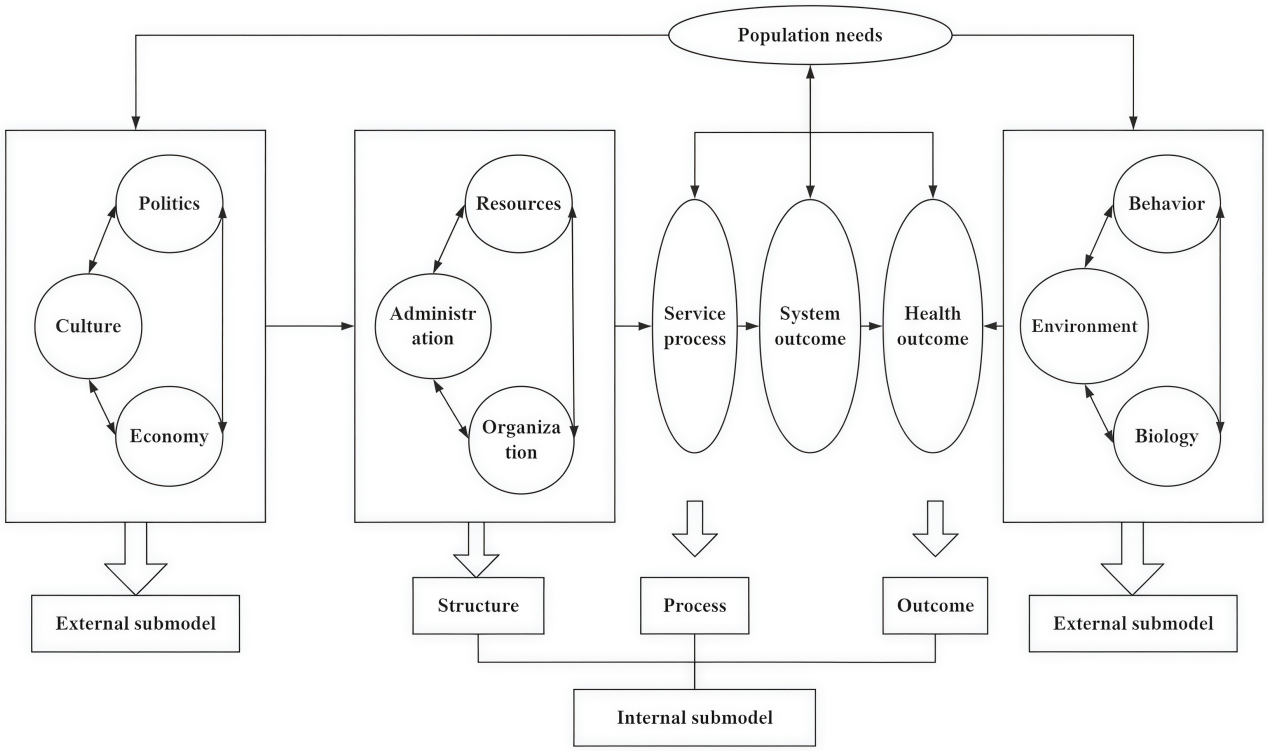

AI-Assisted Cervical Cancer Screening

AI-assisted diagnostic technology has been extensively implemented in China's free cervical cancer screening programs and has facilitated the completion of the nation's first provincial-level universal cervical cancer screening project, which provided screening services to approximately 12.67 million eligible women within 3 years. The most notable advantage of this technology lies in its ability to deliver cervical cancer screening of equivalent quality with high efficiency and low cost in resource-limited settings, thereby significantly advancing health equity.

Within the screening workflow, primary health care institutions are only required to collect cervical exfoliated cell samples from participants. The samples are then transported via logistics to laboratories, where they undergo digitization to generate cellular images uploaded to a cloud platform. Diagnosis is performed through a collaborative effort involving AI algorithms, cytotechnicians, and pathologists. Screening participants can quickly access their test results by scanning a QR code [24].

Study Participants

This study was conducted in Hubei province, China, from January to August 2024. A combination of purposive sampling and multistage stratified sampling was used to obtain an information-rich and representative sample. Hubei province was selected as the study area because it is the first region in China to implement province-wide universal cervical cancer screening.

The sampling procedure was conducted in 2 phases. In the first phase, purposive sampling was applied at the provincial and municipal levels to capture macro-level perspectives on policy-making and program management. Five key administrators were invited from the core management institutions involved in the program. These included 2 officials from the Maternal and Child Health Department of the Provincial/Municipal Health Commission, 2 project managers from the Provincial/Municipal Women's Federation, and 1 project director from the Provincial Maternal and Child Health Hospital.

In the second phase, a multistage stratified sampling method was used to select county-level implementation sites to investigate influencing factors at the grassroots level. The multistage stratified sampling primarily considers factors such as screening rounds, cervical cancer detection rates, and follow-up rates. The detailed sampling flowchart is presented in Figure 2. Based on this framework, 9 county or district-level maternal and child health hospitals were selected as study sites. These hospitals are the sole organizing and implementing institutions for cervical cancer screening within their respective jurisdictions; therefore, their project managers possess

XSL•FO
RenderX

comprehensive knowledge of local screening management and operations. One project manager directly responsible for the program was invited from each hospital to participate in an interview. Thus, this study included 14 key informants as interview participants, whose role distribution ensured multilevel perspective coverage from policy formulation to grassroots implementation.

Figure 2. Flowchart of the multistage stratified sampling process.

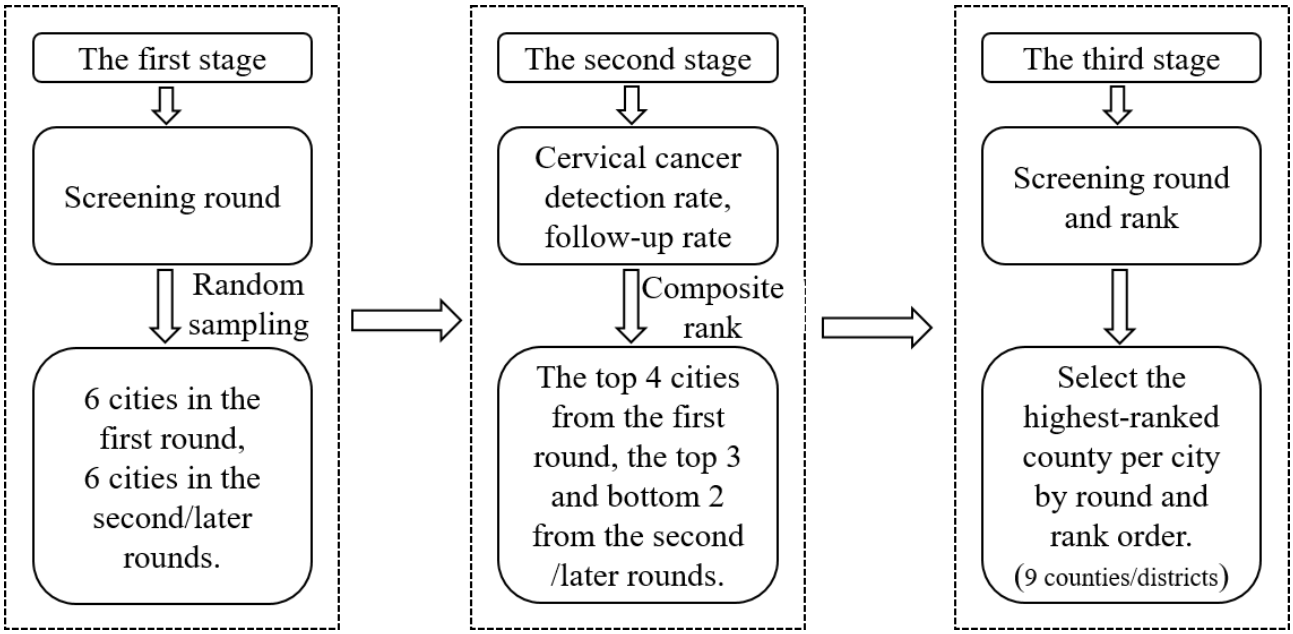

Data Collection

All participants were informed of the researcher's institutional affiliation and the purpose of the qualitative study, which was to understand the implementation of the cervical cancer screening program. Following the provision of verbal informed consent by all participants, the data were collected between January and August 2024. All semistructured interviews were conducted by the same researcher (the first author, holding a PhD in management) to ensure consistency in the interview process and interaction style. The interviewer was female, possessed 7 years of qualitative research experience, and had previously conducted specialized interviews related to cervical cancer screening, ensuring a deep understanding of the research context. Of the 14 participants, 2 had prior professional contact with the interviewer from a 2021 project; to mitigate potential bias toward official stances, the interviewer emphasized academic neutrality and proactively probed for challenges, while the remaining 12 were new contacts. Interviews were carried out either face-to-face or by telephone, in private offices or meeting rooms to protect participant confidentiality. Each interview lasted approximately 60 to 80 minutes and was audio-recorded.

Data Management and Analysis

Within 24 hours post interview, a research team member transcribed the audio recordings verbatim. A second researcher cross-verified transcripts against original recordings to ensure completeness and accuracy. We replaced all personal identifiers with unique codes to anonymize field notes, recordings, and documents. Transcripts were imported into the NVivo 12 (QSR International) software for organization and coding. During the dynamic determination process, we adopted the constant comparative method of "interviewing, coding, and comparing simultaneously." Two professionals independently completed the transcription and coding of interviews; if there were discrepancies in coding, the experienced interview host made the final decision, which effectively avoided subjective biases caused by different researchers. Based on the inductive thematic analysis method under the MMHS framework, we have analyzed and identified the key influencing factors and have also obtained a comprehensive and systematic set of influencing factors.

Thematic saturation was achieved in this study, with no new themes emerging after the interview with the second participant. In the early stage of the study, the applicability of the MMHS framework was verified through a systematic literature review and expert consultations, and a semi-structured interview guide was developed based on this framework. The guide includes sequential questions targeting influencing factors at different levels, supplemented by keyword prompts to minimize potential information gaps, thereby ensuring the systematicity and completeness of the interview content (Multimedia Appendix 1). Coupled with the subsequent adoption of inductive thematic analysis, which enabled the rapid categorization of data into corresponding thematic categories, the study achieved thematic saturation efficiently.

Furthermore, the findings of the thematic analysis were presented and validated for feedback at a cervical cancer screening seminar with 2 provincial-level key informants, who were also participants in this study. Their perspectives were highly consistent with the study's findings, and no revision suggestions were proposed.

Ethical Considerations

Ethics approval was obtained from the Tongji Medical College, Huazhong University of Science and Technology (approval

XSL•FO
RenderX

number [2023] IEC(A235)). Prior to the start of each interview—whether conducted face-to-face or via telephone—verbal informed consent was formally obtained from all participants, who were fully informed of the study purpose, procedures, potential risks, and right to withdraw at any time without any negative consequences before their consent was confirmed. All participant data were anonymized and treated in strict confidence. Each participant was offered a financial compensation of 500 RMB (US $70.60) as a recognition of the time they spent (60 - 80 minutes per interview) and the professional responses they provided. A portion of participants voluntarily chose to participate on an unpaid basis, aiming to promote the smooth implementation of the full-coverage project to serve women's health. The compensation was solely intended to offset participants' time costs and acknowledge their professional inputs, without serving as an incentive to influence the authenticity of their responses.

Results

Participant Characteristics

No participants refused or withdrew during the process. Between January and August 2024, we conducted interviews with 14 administrators from the Health Commission (HC), Women's Federation (WF), and maternal and child health (MCH) hospitals. Participants comprised 9 program managers from MCH hospitals, 2 program directors from provincial HC, and 2 representatives from provincial WF. Among the participants, 92.86% (n=13) reported having more than 10 years of work experience, with 42.86% (n=6) having over 20 years. Most participants held bachelor's degrees, with 3 participants having completed postgraduate education. Detailed participant characteristics are provided in Table 1.

Table . Sociodemographic and occupational characteristics of the participants.

Characteristic	Participants, n (%)
Organization	
The Women's Federation	2 (14.29)
The Maternal and Child Health Hospitals	10 (71.42)
The Health and Wellness Administration	2 (14.29)
Sex	
Male	4 (28.57)
Female	10 (71.42)
Age (y)	
30 - 39	2 (14.29)
40 - 49	9 (64.29)
50 - 59	3 (21.43)
Education	
Associate degree	1 (7.14)
College	10 (71.42)
Master	3 (21.43)
Work experience (y)	
3 - 9	1 (7.14)
10 - 19	7 (50)
20 - 29	6 (42.86)

Thematic Analysis

Overview

Thematic analysis revealed that the core drivers can be summarized into 4 key factors: government stewardship serves as the fundamental guarantee, screening technology provides the foundational support, screening funding supplies sustained momentum, and improving public health literacy is the intrinsic key to achieving broad participation. Further systematic analysis indicates that these factors can be integrated into a comprehensive "structure-process-outcome" framework. Structural factors include government stewardship, multisectoral collaboration, funding, information system, screening technology, and trained screening technicians. Process factors include service delivery capacity, community mobilization, health literacy, and quality control. Outcome measures include coverage, cytology positivity rate, and follow-up rate treatment rate. Detailed results are presented in Table 2.

XSL•FO
RenderX

Table . Themes and Subthemes.

Themes and subthemes	Tertiary theme
Structure	
Administration	• Government stewardship
Organization	• Multisectoral collaboration
Resources	• Funding for screening • Information system • Screening technology • Trained screening technicians
Process	
Service provision and utilization	• Service delivery capacity • Community mobilization • Health literacy
Outcomes	
Screening results	• Coverage • Cytology positivity rate • Follow-up rate • Treatment rate

Government Stewardship

Participants indicated that strong government promotion was a key factor enabling the project to achieve comprehensive coverage. This governmental driving force was primarily manifested by incorporating the screening program into regional government performance metrics, with local government departments overseeing HC or WF and MHC in completing the screening tasks. Participants consistently emphasized:

> *The government's proactive approach demonstrates the decisive role of political will. Sustained political support has been the key driver for program implementation, and policy endorsement of AI technology was indispensable for scaling up this AI-assisted screening initiative.* [MCH01,02,03,04,07,08; HC01]

Another participant noted:

> *The provincial government prioritized free cervical cancer screening, incorporating it into the ' public service delivery' performance metrics. Screening completion rates became key evaluation criteria for local governments. Without such government mobilization, universal screening coverage would be unattainable.* [WF01]

Additionally, the government's leading role is reflected in its support for screening funding and the application of AI-assisted diagnostic technologies.

Screening Technology

Participants indicated that AI-assisted cervical cancer screening technology serves as a critical tool for achieving universal coverage. Previously, primary health care institutions struggled to meet demands of residents due to a shortage of pathologists. By facilitating the decentralization of remote pathological resources, AI-assisted diagnostic technology has enhanced service accessibility and addressed screening bottlenecks in areas with limited health care resources. Furthermore, the screening efficiency of AI-assisted diagnostic technology is remarkably high, enabling significant improvements in screening coverage within a short timeframe.

> *AI-based image analysis reduces physicians' workload and ensures high-quality task completion in primary care settings with insufficient pathology personnel. The technology can automatically handle large volumes of repetitive tasks, such as preliminary screening and marking abnormal areas, thereby alleviating physicians' workload. Under traditional methods, a pathologist could screen a maximum of 100 individuals per day, whereas digitization increases this capacity to nearly 10,000.* [MCH03]

> *The automated system processes cases much faster than manual methods, and the generated reports are highly accurate. Our AI big data cloud diagnosis platform generates up to 50,000 daily diagnoses.* [MCH07]

> *In our region, manual screening previously allowed only 4,000 cases annually, but with AI technology, we can now screen over 73,000 individuals per year.* [MCH10]

> *Some regions face significant resource limitations, particularly in mountainous areas where specialized pathologists are scarce. The introduction of AI-assisted screening technology helps mitigate these disparities. Through cloud-based API calls, primary care institutions can access advanced diagnostic technology, improving the accessibility of cervical cancer screening.* [MCH01]

XSL•FO
RenderX

Screening Funding

Insufficient screening funding, inadequate matching funds, and unclear allocation collectively constitute the primary financial challenges for the screening program. These issues not only hinder smooth program implementation but also strain health care institutions' financial capacity and compromise service quality. Some screening institutions report that large-scale screening has caused more severe funding shortages than before.

> *The per-case subsidy decreased 25.8% (from 66 RMB to 49 RMB) following AI implementation. But we must cover multiple costs including disposable supplies, field visits of personnel and equipment, and follow-up activities, costs that were underestimated or unaccounted for in the budget, making the 49 RMB subsidy insufficient to cover actual costs. We consistently operate at a loss when conducting screenings.* [MCH02,07,10]
>
> *Healthcare institutions receive less than 40% government funding, relying entirely on self-generated revenue. However, state funding for program operations remains severely inadequate, and trade union funding has been reduced in recent years—we're expected to do more work with significantly less funding.* [MCH04]

Furthermore, insufficient matching funds remain an issue.

> *We face tremendous financial pressure. The government assigned heavy screening targets without allocating sufficient funds, requiring advance payments that overextend local government finances.* [MCH05,08]

Unclear fund allocation exacerbates the funding shortage, remaining a key barrier to achieving universal screening coverage.

> *The core funding issue stems from policy-mandated cost estimates set unrealistically low, resulting in systemic underfunding.* [HC01]
>
> *Current lump-sum allocations lack itemized funding for dual cancers (cervical or breast), with no specific indicators for individual budget lines…* [MCH07]
>
> *Lump-sum funding becomes problematic at county-level redistribution, with no clear documentation on allocation plans or expenditure purposes.* [MCH01]

Health Literacy

Participants noted that insufficient health literacy among residents is a significant barrier, particularly in rural areas. Although AI-assisted diagnostic technologies for cervical cancer screening have substantially improved service accessibility, residents' willingness to utilize free screening services remains low. Officials from MCH hospitals emphasized that promoting cervical cancer prevention awareness, disseminating information about free screening, and mobilizing residents to participate require dedicated and proactive efforts, rather than merely providing information about free screening availability.

> *If individuals receive invitations, they may participate; however, without direct contact, they might overlook group notifications.* [MCH08]
>
> *Many are constrained by cognitive limitations and traditional cultural beliefs, while others decline participation due to factors like travel distance to townships or family responsibilities.* [MCH09]
>
> *Due to insufficient recognition of its importance, incentives such as transportation assistance and free breast cancer screenings are necessary to achieve screening rates.* [MCH04]
>
> *Project success fundamentally depends on population health literacy—once literacy improves, residents will proactively seek screening without promotional campaigns.* [MCH07]
>
> *Some women refuse follow-up after initial negative results, mistakenly viewing normal outcomes as permanent risk exemption—reflecting inadequate screening awareness.* [MCH01]

Multisectoral Collaboration

Participants indicated that a multisectoral collaboration mechanism had been established prior to the implementation of the comprehensive coverage project. During the execution of the project, under government stewardship, the responsibilities of each department were further clarified, detailed, and refined. In the process of implementing the comprehensive coverage plan, the MCH were responsible for providing screening services, and the HC oversaw personnel training, project progress, and quality control, while the WF was tasked with mobilizing residents to actively utilize cervical cancer screening services.

> *This project is spearheaded by the Women's Federation and the Hubei Provincial Health Commission, with joint participation from the Trade Union and the Finance Department. The Women's Federation is responsible for publicity efforts, while the Health Commission provides staff training for maternal and child health hospitals , oversees project progress, and ensures quality control. The Finance Department provides financial support.* [HC02; MCH01]

Information System

Screening programs require the collection of extensive information, and the AI-assisted diagnostic system's information platform streamlines the entire process for residents—from initial screening to receiving results. The implementation of this information platform has enhanced the management efficiency of screening programs. However, several challenges persist, including incomplete system development, data collection gaps, data sharing difficulties, and limited usability at grassroots levels. Participants reported:

> *The system cannot export positive cases, requiring manual entry for each one—extremely inconvenient.* [MCH02]

Another user noted:

XSL•FO
RenderX

> *The system operation is overly complex. A simplified system design would enhance usability.* [MCH06]

Additionally, concerns were raised about data completeness:

> *The system's statistical data isn't entirely complete … trade union and corporate data aren't reported into this system.* [MCH01]

Trained Screening Technicians

Professional personnel required for cervical cancer screening include screening operators, cytopathologists, and diagnostic pathologists, with the shortage of the latter 2 categories being particularly pronounced. Project managers reported that the introduction of AI-assisted diagnostic technology has addressed the shortage of specialized technical personnel in cytological and pathological diagnosis.

> *Current staffing remains suboptimal with inadequate professional qualifications. Nevertheless, AI technology can address challenges in cytology and pathology diagnosis, effectively resolving the difficulties caused by a shortage of specialized personnel.* [MCH01,07; WF02]

> *We conduct annual colposcopy training requiring physicians to pass certification exams. Additionally, the provincial MCH hospital offers specialized three-month training programs to ensure sampling proficiency.* [MCH05,06; HC02]

Service Delivery Capacity

In the comprehensive cervical cancer screening program, MCH's service capacity is primarily reflected in its capabilities for cervical cancer screening and treatment. The screening program encompasses both urban and rural regions, involving provincial or county MCH and township health centers. District or county-level centers exhibit good screening and treatment capabilities. However, township centers face challenges stemming from professional and resource limitations, resulting in restricted screening and treatment capacities. Notably, AI technology has significantly enhanced their screening capabilities.

> *Regarding service capacity, most institutions can adequately manage general gynecological diseases and precancerous lesions.* [MCH02,04,07]

> *Township facilities are experiencing downsizing with insufficient obstetrician and gynecologist, limiting maternal and child health follow-up capacity. However, AI-enabled sample referral to provincial centers now ensures successful rural screening implementation.* [MCH03; HC01]

Quality Control

Quality control is essential for ensuring both progress and quality in achieving universal cervical cancer screening coverage. It involves technical accuracy and reliability at the operational level. In addition to on-site inspections, remote monitoring is conducted through the information system, primarily to track completion rates across all screening sites.

> *We monitor progress through the information system and reported data, including the number of individuals screened, treated, and managed. These metrics ultimately generate performance rankings.* [MCH01,03,05]

Community Mobilization

Community mobilization campaigns, particularly those driven by administrative initiatives and free voucher distribution, have proven effective in enhancing awareness and increasing cervical cancer screening participation rates. Large-scale community mobilization campaigns can improve public understanding of cervical cancer, thereby increasing screening willingness and advancing the goal of universal screening coverage. These strategies have significantly contributed to achieving comprehensive screening coverage.

> *The free cervical cancer screening program will be promoted through flyers, short videos, official social media accounts, television programs, small-scale lectures, and live-streaming sessions.* [MCH04,05,09]

> *Free screening vouchers were printed and distributed to village clinics and women's committee directors, serving both promotional and participatory purposes. Last year, 100,000 vouchers were printed and distributed based on demand.* [MCH02]

Coverage and Cytology Positivity Rate

Nearly all regions reported achieving comprehensive coverage of cervical cancer screening, primarily attributed to the enhanced service accessibility and efficiency facilitated by AI-assisted technology. Compared to previous methods, the technological upgrades in screening have also significantly increased the cytological positive detection rate. This has further heightened health awareness and willingness to proactively utilize screening services. Coverage and positive detection rates not only directly affect early diagnosis and treatment rates in the current screening cycle but also dynamically shape whether the goal of sustained universal coverage can be achieved.

> *The adoption of AI assistance has significantly improved overall detection rates compared to previous methods.* [MCH01]

With advancements in screening technology and updated equipment, the positive detection rate for cervical cancer screening has increased markedly compared to historical levels. Multiple respondents confirmed:

> *Universal coverage has essentially been achieved.* [MCH02-10]

Follow-Up and Treatment Rates

Cervical cancer screening should be conducted at regular intervals to detect lesions at an early stage and enable timely intervention. Participants emphasized that achieving comprehensive screening coverage requires attention not only to the screening rate but, more importantly, to follow-up management and treatment rates. Follow-up and treatment rates are key indicators for the closed-loop management of screening programs, directly influencing their effectiveness and affecting

XSL•FO
RenderX

whether residents will continue to participate in screening regularly in the future. Although universal screening coverage has been achieved, significant challenges remain in the follow-up of positive cases, which may, to some extent, affect future screening rates.

> *Our follow-up reaches approximately 70% of targets, as many patients don't respond to phone calls … with higher loss-to-follow-up rates in rural areas.* [MCH05,10]

Primary challenges include patient noncompliance and disparities between screening, re-screening, and treatment locations. These difficulties complicate treatment rate documentation.

> *Some migrant workers relocate before results are available, while many screen-positive patients seek treatment at top-tier hospitals elsewhere—making data tracking challenging though we maintain documentation efforts.* [WF01,02]

Discussion

Principal Findings

Through interviews, we have identified factors influencing the achievement of universal coverage in cervical cancer screening, which span 3 dimensions: structural, service-process, and outcome-related. Among these, government stewardship is considered the primary determinant, followed by the application of AI-assisted diagnostic technology, sufficient funding, and public health literacy, all of which jointly contribute to the realization of comprehensive screening coverage. During implementation, government-led initiatives provide strong support for the promotion of AI-assisted diagnostic technology, multisectoral collaboration, and funding allocation, while also facilitating public access to screening services. Concurrently, AI-assisted diagnostic technology helps alleviate shortages of trained screening technicians, establishes information management platforms, reduces screening costs, and thereby enhances institutional service capacity. To further advance the goal of universal coverage, it is essential to ensure sustained financial investment, improve public health literacy, and strengthen follow-up and treatment for positive cases.

Government stewardship serves as the critical enabler for implementing large-scale cervical cancer screening programs. In developed nations such as the United States, New Zealand, and the United Kingdom, cervical cancer prevention systems are legally mandated, whereas China adopts national policy-driven approaches [36]. Hubei's universal screening program, integrated into the provincial "Public Service Initiatives," ties screening targets to local government performance metrics, ensuring comprehensive policy support and fiscal allocation [37]. The program established a multisector framework led by the government, coordinated by the Women's Federation, with participation from health care, financial, and labor unions. It is noteworthy that even when funding management and public health literacy failed to exert a significant positive influence, the government effectively achieved the phased objective of "testing all who should be tested" through the impetus of clearly defined assessment targets. In contrast, most LMICs lack both legal frameworks and substantive policy support for cervical cancer prevention [38]. Screening in LMICs remains opportunistic, lacking systematic implementation strategies. Thus, robust governmental commitment and policy infrastructure are prerequisites for population-wide screening.

AI-assisted screening technology represents a pivotal component in achieving large-scale cervical cancer screening. It substantially enhances screening efficiency and coverage while reducing costs, enabling Hubei province to complete population-wide screening within compressed timelines. In 2016, a single Chinese pathologist could manually process approximately 100 cytology screenings daily—a rate that would leave millions of women lifetime-unserved by screening programs [39]. However, AI technology allows a single pathologist to conduct roughly 10,000 screenings per day, significantly cutting down mass screening time. The technology also improves accessibility, allowing rural women to undergo sample collection at local clinics without travel burdens. Economically, AI-assisted screening demonstrates superior cost-effectiveness, reducing per-capita costs from 100-200 RMB (medical service fees) to 49 RMB, enabling broader eligibility for free screening. These advantages make universal screening coverage a feasible public health goal.

Financially, the government has allocated approximately 600 million RMB to advance the universal cervical cancer screening initiative. However, it is noteworthy that screening institutions continue to face challenges in funding security, manifested through 3 dimensions: insufficient total funding, lack of matching funds, and ambiguous allocation mechanisms. Funding shortages are particularly acute in LMICs, where systemic underinvestment persists [40,41]. Despite AI's cost-saving potential, funding gaps endure due to insufficient government reimbursement and overreliance on singular revenue streams [42,43]. Sustainable implementation requires diversified financing mechanisms and evidence-based fiscal policies [44]. Current tiered matching-fund systems fail to alleviate local funding deficits, as reported by stakeholders. This mismatch disproportionately affects economically disadvantaged regions, where heavy reliance on central transfers, limited fiscal autonomy, and low economic capacity exacerbate funding inadequacies [45]. Solutions include needs-based allocation models incorporating regional disparities in fiscal capacity, population size, and geographic coverage. Long-term economic development remains critical for fiscal resilience [46]. China's lump-sum public health funding—which includes screening budgets—lacks transparent suballocation guidelines [47]. Ambiguities persist regarding cost coverage and interagency fund distribution. Unclear allocation risks resource concentration, compromising efficiency and equity [48]. Standardized operational guidelines for fund distribution are thus imperative for program sustainability [49].

Population health literacy has emerged as a critical determinant of overall health status, reflecting growing public health awareness. It directly influences both health service utilization and treatment-seeking behaviors [50]. This study identified multidimensional barriers to screening participation: low

XSL•FO
RenderX

awareness, embarrassment, geographic accessibility issues, and competing household priorities—necessitating incentives, such as gifts or transportation. Knowledge gaps about cervical cancer persist universally across income settings [51]. Despite understanding triennial screening recommendations, women often discontinue after negative results, reflecting insufficient risk perception. Targeted education interventions demonstrably improve screening knowledge and participation rates [52]. Sustainable screening expansion thus requires dual strategies: technological or service optimization and literacy-focused health promotion. Multimodal education can foster proactive health behaviors and sustained screening engagement [53,54].

In the initial phase of our research, we viewed AI technology as the primary driver for achieving universal cervical cancer screening coverage. We believed that addressing the supply bottleneck caused by a shortage of pathologists through technological means would enable rapid attainment of coverage targets. However, through systematic interview analysis, we discovered that while AI technology can effectively enhance service delivery capacity, its effectiveness is highly dependent on systemic support. Government leadership emerged as the most critical factor in achieving screening coverage. By integrating screening into government performance evaluation systems, introducing new technologies (AI-assisted diagnosis), and providing sustained financial support, the government effectively coordinated resources. It mobilized multiple departments—including WFs and MCH hospitals—to collaborate on outreach and organizational efforts, significantly boosting residents' willingness to participate and actual screening rates. This reflection prompted a fundamental shift in our cognitive framework: moving from an initial "technology-centric" singular assumption to recognizing a multilayered, systematic advancement logic characterized by "government leadership, technological support, multistakeholder collaboration, and competency enhancement." This shift lends our research conclusions greater explanatory power and policy relevance.

Strengths and Limitations

Improving cervical cancer screening coverage is a critical public health priority in China and other LMICs. The emergence of AI technology offers a promising solution to enhance the accessibility of these services. Within the context of AI applications in China's basic public health services, this study systematically investigates the challenges faced in its practical implementation for increasing screening coverage. Furthermore, by applying the MMHS, we provide a comprehensive analysis of the determinants of screening, which offers a robust framework for developing the interview guide. However, a limitation of this study is that it only considers the influencing factors from the perspectives of project managers and providers. In actual program implementation, the perspective of the demand side is equally crucial, and its absence may result in factors affecting program execution being identified incompletely.

Conclusions

The achievement of comprehensive cervical cancer screening coverage is mainly influenced by government stewardship, AI-assisted screening, screening funding, and health literacy. Among these, government stewardship serves as the core driving force, specifically reflected in the inclusion of screening in government performance evaluations, which drives policy implementation, technological adoption, funding security, multisectoral collaboration, and activities aimed at improving residents' health literacy and screening willingness. The application of AI-assisted diagnostic technology effectively mitigates constraints, such as shortages of trained screening technicians, limited funding, and insufficient service delivery capacity. To ensure the smooth implementation of the project and achieve preventive goals, coordinated efforts are required in 3 key areas: at the policy level, clear screening plans and supportive measures should be formulated; at the financial level, sufficient funding must be guaranteed, with coordinated multilevel and multidepartmental resources and clearly defined allocation mechanisms; at the service delivery level, diversified health education should be carried out to enhance residents' health literacy, cultivate regular screening habits, and promote the application of AI technology.

Acknowledgments

We extend our gratitude to all the stakeholders in Hubei province who contributed their insights and perspectives on the research topic.

Funding

This work was supported by Health Commission of Hubei Province, China (WJ2023Z008).

Data Availability

All data generated or analyzed during this study are included in this published paper.

Authors' Contributions

LY is the principal investigator of the study and responsible for the overall content as guarantor. LJ designed the study, drafted the study documents, and participated in both data collection and analysis. XZ assisted in drafting the study documents and coding and analyzing the transcripts. LY and LJ provided feedback on the study protocol draft and related documents and reviewed the quality of the data, and XZ assisted data collection. LJ and XZ supported the development of data extraction processes and the

XSL•FO
RenderX

interpretation of the findings. XZ authored the first draft of the manuscript with support from LJ and LY. All other authors contributed feedback on the manuscript and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[DOCX File, 22 KB - jmir_v28i1e75372_app1.docx]

Checklist 1
COREQ checklist.
[DOC File, 73 KB - jmir_v28i1e75372_app2.doc]

References

1. Ferlay J, Ervik M, Lam F, et al. Cancer today. Global Cancer Observatory (GCO). URL: https://gco.iarc.who.int/today [accessed 2024-09-24]
2. Cancer today–China. : Global Cancer Observatory (GCO), International Agency for Research on Cancer (IARC), World Health Organization (WHO); 2024 URL: https://gco.iarc.who.int/media/globocan/factsheets/populations/160-china-fact-sheet.pdf [accessed 2024-09-24]
3. Zheng R, Zhang S, Zeng H, et al. Cancer incidence and mortality in China, 2016. J Natl Cancer Cent 2022 Mar;2(1):1-9. [doi: 10.1016/j.jncc.2022.02.002] [Medline: 39035212]
4. Global strategy to accelerate the elimination of cervical cancer as a public health problem. : World Health Organization (WHO); 2020 URL: https://www.who.int/publications/i/item/9789240014107 [accessed 2024-10-28]
5. Zhang X, Hu S, Zhao F, et al. Cervical cancer screening service policies in China, 2006–2023: a quantitative analysis of government documents. Chinese J Public Health 2024;40(5):521-527. [doi: 10.11847/zgggws1143537]
6. Campos NG, Tsu V, Jeronimo J, Mvundura M, Lee K, Kim JJ. To expand coverage, or increase frequency: quantifying the tradeoffs between equity and efficiency facing cervical cancer screening programs in low-resource settings. Int J Cancer 2017 Mar 15;140(6):1293-1305. [doi: 10.1002/ijc.30551] [Medline: 27925175]
7. Lin S, Gao K, Gu S, et al. Worldwide trends in cervical cancer incidence and mortality, with predictions for the next 15 years. Cancer 2021 Nov 1;127(21):4030-4039. [doi: 10.1002/cncr.33795] [Medline: 34368955]
8. Yagi A, Ueda Y, Kakuda M, et al. Epidemiologic and clinical analysis of cervical cancer using data from the population-based Osaka Cancer Registry. Cancer Res 2019 Mar 15;79(6):1252-1259. [doi: 10.1158/0008-5472.CAN-18-3109] [Medline: 30635276]
9. Ojamaa K, Innos K, Baburin A, Everaus H, Veerus P. Trends in cervical cancer incidence and survival in Estonia from 1995 to 2014. BMC Cancer 2018 Nov 7;18(1):1075. [doi: 10.1186/s12885-018-5006-1] [Medline: 30404606]
10. Bruni L, Serrano B, Roura E, et al. Cervical cancer screening programmes and age-specific coverage estimates for 202 countries and territories worldwide: a review and synthetic analysis. Lancet Glob Health 2022 Aug;10(8):e1115-e1127. [doi: 10.1016/S2214-109X(22)00241-8] [Medline: 35839811]
11. Zhang M, Zhong Y, Wang L, et al. Cervical cancer screening coverage—China, 2018-2019. China CDC Wkly 2022 Dec 2;4(48):1077-1082. [doi: 10.46234/ccdcw2022.217] [Medline: 36751373]
12. Lahole BK, Woldamlak M, Kussia W. Determinants of cervical cancer screening intention among reproductive age women in Ethiopia: a systematic review and meta-analysis. PLoS One 2024;19(10):e0312449. [doi: 10.1371/journal.pone.0312449] [Medline: 39480772]
13. Wang M, Chen H, Wong MCS, Huang J, Jin Y, Zheng ZJ. Effects of screening coverage and screening quality assurance on cervical cancer mortality: implication for integrated framework to monitor global implementation of cervical cancer screening programmes. J Glob Health 2024 Oct 25;14:04189. [doi: 10.7189/jogh.14.04189] [Medline: 39450607]
14. Salehiniya H, Momenimovahed Z, Allahqoli L, Momenimovahed S, Alkatout I. Factors related to cervical cancer screening among Asian women. Eur Rev Med Pharmacol Sci 2021 Oct;25(19):6109-6122. [doi: 10.26355/eurrev_202110_26889] [Medline: 34661271]
15. Sabeena S, Ravishankar N, Kalpana MM. Implementation strategies of cervical cancer screening in South Asia: a systematic review. Int J Gynaecol Obstet 2024 Aug;166(2):483-493. [doi: 10.1002/ijgo.15366] [Medline: 38268412]
16. Surendran S, Foo CD, Tan DHY, et al. Understanding barriers and facilitators of breast and cervical cancer screening among Singapore women: a qualitative approach. Asian Pac J Cancer Prev 2023 Mar 1;24(3):889-895. [doi: 10.31557/APJCP.2023.24.3.889] [Medline: 36974542]
17. Suk R, Hong YR, Rajan SS, Xie Z, Zhu Y, Spencer JC. Assessment of us preventive services task force guideline-concordant cervical cancer screening rates and reasons for underscreening by age, race and ethnicity, sexual orientation, rurality, and

XSL•FO
RenderX

insurance, 2005 to 2019. JAMA Netw Open 2022 Jan 4;5(1):e2143582. [doi: 10.1001/jamanetworkopen.2021.43582] [Medline: 35040970]
18. Shrestha AD, Andersen JG, Gyawali B, et al. Cervical cancer screening utilization, and associated factors, in Nepal: a systematic review and meta-analysis. Public Health 2022 Sep;210:16-25. [doi: 10.1016/j.puhe.2022.06.007] [Medline: 35863158]
19. Mezei AK, Armstrong HL, Pedersen HN, et al. Cost-effectiveness of cervical cancer screening methods in low- and middle-income countries: a systematic review. Int J Cancer 2017 Aug 1;141(3):437-446. [doi: 10.1002/ijc.30695] [Medline: 28297074]
20. Allanson ER, Schmeler KM. Cervical cancer prevention in low- and middle-income countries. Clin Obstet Gynecol 2021 Sep 1;64(3):501-518. [doi: 10.1097/GRF.0000000000000629] [Medline: 34120126]
21. Martinez ME, Schmeler KM, Lajous M, Newman LA. Cancer screening in low- and middle-income countries. Am Soc Clin Oncol Educ Book 2024 Jun;44(3):e431272. [doi: 10.1200/EDBK_431272] [Medline: 38843475]
22. Di J, Rutherford S, Chu C. Review of the cervical cancer burden and population-based cervical cancer screening in China. Asian Pac J Cancer Prev 2015;16(17):7401-7407. [doi: 10.7314/apjcp.2015.16.17.7401] [Medline: 26625735]
23. Ji L, Chen M, Yao L. Strategies to eliminate cervical cancer in China. Front Oncol 2023;13:1105468. [doi: 10.3389/fonc.2023.1105468] [Medline: 37333817]
24. Ji L, Yao Y, Yu D, et al. Performance of a full-coverage cervical cancer screening program using on an artificial intelligence- and cloud-based diagnostic system: observational study of an ultralarge population. J Med Internet Res 2024 Nov 20;26:e51477. [doi: 10.2196/51477] [Medline: 39566061]
25. Bao H, Sun X, Zhang Y, et al. The artificial intelligence-assisted cytology diagnostic system in large-scale cervical cancer screening: a population-based cohort study of 0.7 million women. Cancer Med 2020 Sep;9(18):6896-6906. [doi: 10.1002/cam4.3296] [Medline: 32697872]
26. Bao H, Bi H, Zhang X, et al. Artificial intelligence-assisted cytology for detection of cervical intraepithelial neoplasia or invasive cancer: a multicenter, clinical-based, observational study. Gynecol Oncol 2020 Oct;159(1):171-178. [doi: 10.1016/j.ygyno.2020.07.099] [Medline: 32814641]
27. Chen H, Liu J, Wen QM, et al. CytoBrain: cervical cancer screening system based on deep learning technology. J Comput Sci Technol 2021 Apr;36(2):347-360. [doi: 10.1007/s11390-021-0849-3]
28. Cervical cancer screening can be done at township health centers [Article in Simplified Chinese]. Hubei Provincial Health Commission. 2025. URL: https://wjw.hubei.gov.cn/bmdt/dtyw/202410/t20241021_5377559.shtml [accessed 2025-01-09]
29. Hou X, Shen G, Zhou L, Li Y, Wang T, Ma X. Artificial intelligence in cervical cancer screening and diagnosis. Front Oncol 2022;12:851367. [doi: 10.3389/fonc.2022.851367] [Medline: 35359358]
30. Bi WL, Hosny A, Schabath MB, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. CA Cancer J Clin 2019 Mar;69(2):127-157. [doi: 10.3322/caac.21552] [Medline: 30720861]
31. Adams SJ, Stone E, Baldwin DR, Vliegenthart R, Lee P, Fintelmann FJ. Lung cancer screening. Lancet 2023 Feb 4;401(10374):390-408. [doi: 10.1016/S0140-6736(22)01694-4] [Medline: 36563698]
32. Wu T, Lucas E, Zhao F, Basu P, Qiao Y. Artificial intelligence strengthens cervical cancer screening—present and future. Cancer Biol Med 2024 Sep 19;21(10):864-879. [doi: 10.20892/j.issn.2095-3941.2024.0198] [Medline: 39297572]
33. Gastounioti A, Desai S, Ahluwalia VS, Conant EF, Kontos D. Artificial intelligence in mammographic phenotyping of breast cancer risk: a narrative review. Breast Cancer Res 2022 Feb 20;24(1):14. [doi: 10.1186/s13058-022-01509-z] [Medline: 35184757]
34. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
35. Hao M. Health Policy [Book in Chinese], 2nd edition: People's Medical Publishing House; 2013. URL: https://product.dangdang.com/11956642953.html [accessed 2026-03-23]
36. Bao HL, Fang LW, Wang LH. The strategy for establishment of comprehensive cervical cancer prevention and control in the world [Article in Chinese]. Zhonghua Yu Fang Yi Xue Za Zhi 2017 Jan 6;51(1):96-100. [doi: 10.3760/cma.j.issn.0253-9624.2017.01.019] [Medline: 28056279]
37. Dee EC, Eala MAB, Robredo JPG, et al. Leveraging national and global political determinants of health to promote equity in cancer care. J Natl Cancer Inst 2023 Oct 9;115(10):1157-1163. [doi: 10.1093/jnci/djad123] [Medline: 37402623]
38. Lee J, Ismail-Pratt I, Machalek DA, Kumarasamy S, Garland SM. From barriers to opportunities from COVID-19 pandemic: stakeholder perspectives on cervical cancer screening programs in LMICs of the Asia-Pacific region. PLOS Glob Public Health 2024;4(10):e0003768. [doi: 10.1371/journal.pgph.0003768] [Medline: 39365795]
39. Li Y, Ma L, Yang C, et al. A study on service capacity of primary medical and health institutions for cervical cancer screening in urban and rural areas in China. Chin J Cancer Res 2019 Oct;31(5):838-848. [doi: 10.21147/j.issn.1000-9604.2019.05.13] [Medline: 31814687]
40. Petersen Z, Jaca A, Ginindza TG, et al. Barriers to uptake of cervical cancer screening services in low-and-middle-income countries: a systematic review. BMC Womens Health 2022 Dec 2;22(1):486. [doi: 10.1186/s12905-022-02043-y] [Medline: 36461001]

XSL•FO
RenderX

41. Devarapalli P, Labani S, Nagarjuna N, Panchal P, Asthana S. Barriers affecting uptake of cervical cancer screening in low and middle income countries: a systematic review. Indian J Cancer 2018;55(4):318-326. [doi: 10.4103/ijc.IJC_253_18] [Medline: 30829264]
42. Luo YB, Jia LG, Liu JA, et al. Management on national basic public health service projects and its related issues in Hubei province [Article in Chinese]. Chin J Public Health 2019;35(6):676-679. [doi: 10.11847/zgggws1120669]
43. Zhu X, Yao Q, Dai W, et al. Cervical cancer screening aided by artificial intelligence, China. Bull World Health Organ 2023 Jun 1;101(6):381-390. [doi: 10.2471/BLT.22.289061] [Medline: 37265676]
44. Kutzin J. Health financing for universal coverage and health system performance: concepts and implications for policy. Bull World Health Organ 2013 Aug 1;91(8):602-611. [doi: 10.2471/BLT.12.113985] [Medline: 23940408]
45. Feng G. Local finance from the perspective of government intervention: a literature review [Article in Chinese]. Financial 2016(3):67-79125. [doi: 10.3969/j.issn.1674-7690.2016.03.005]
46. Zhang Y. Reflections on the optimization of China's central finance to local transfer payment system: current situation, question and countermeasure [Article in Chinese]. Finance 2024;14(5):1640-1647. [doi: 10.12677/fin.2024.145168]
47. Chen H, Tao HB, Luo LX, et al. The situation and analysis of the allocation and performance management system of community health service funds at home and abroad. Chin Gen Pract 2011(16):7. [doi: 10.3969/j.issn.1007-9572.2011.16.005]
48. Zhi Q, Meng T. Funding allocation, inequality, and scientific research output: an empirical study based on the life science sector of natural science foundation of China. Scientometrics 2016 Feb;106(2):603-628. [doi: 10.1007/s11192-015-1773-5]
49. Wang LM, Yu J. Analysis of the allocation mechanism of basic public health services in Wuhan. Med Soc (Berkeley) 2013(7):56-59. [doi: 10.3870/YXYSH.2013.07.019]
50. Kim K, Han HR. Potential links between health literacy and cervical cancer screening behaviors: a systematic review. Psychooncology 2016 Feb;25(2):122-130. [doi: 10.1002/pon.3883] [Medline: 26086119]
51. Kung TPH, Gordon JR, Abdullahi A, et al. "My husband says this: if you are alive, you can be someone…": facilitators and barriers to cervical cancer screening among women living with HIV in India. Cancer Causes Control 2019 Apr;30(4):365-374. [doi: 10.1007/s10552-019-01145-7] [Medline: 30809741]
52. Mukama T, Ndejjo R, Musabyimana A, Halage AA, Musoke D. Women's knowledge and attitudes towards cervical cancer prevention: a cross sectional study in Eastern Uganda. BMC Womens Health 2017 Jan 31;17(1):9. [doi: 10.1186/s12905-017-0365-3] [Medline: 28137260]
53. Popalis ML, Ramirez SI, Leach KM, Granzow ME, Stoltzfus KC, Moss JL. Improving cervical cancer screening rates: a scoping review of resources and interventions. Cancer Causes Control 2022 Nov;33(11):1325-1333. [doi: 10.1007/s10552-022-01618-2] [Medline: 35980511]
54. Holme F, Maldonado F, Martinez-Granera OB, et al. HPV-based cervical cancer screening in Nicaragua: from testing to treatment. BMC Public Health 2020 Apr 15;20(1):495. [doi: 10.1186/s12889-020-08601-z] [Medline: 32295562]

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research
HC: Health Commission
LMIC: low- and middle-income country
MCH: maternal and child health
MMHS: macro model of health system
WF: Women's Federation

Edited by A Stone; submitted 02.Apr.2025; peer-reviewed by C Long, D Tang, G Raut, M Temmerman, S Tang; accepted 23.Feb.2026; published 01.Apr.2026.

Please cite as:
Ji L, Zhou X, Yao L
Factors Influencing Universal Coverage of AI-Assisted Cervical Cancer Screening: Qualitative Study Based on the Macro Model of Health System
J Med Internet Res 2026;28:e75372
URL: https://www.jmir.org/2026/1/e75372
doi:10.2196/75372

© Lu Ji, Xinke Zhou, Lan Yao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Strategy for Hepatitis B and C Virus Testing Campaigns Through Web Services and Digital Advertising in Japan: Nationwide Cross-Sectional Study With Correspondence Analysis

Kosuke Sakai[1], MD, PhD; Shoko Nakazawa[1], MD, PhD; Yuko Furuya[1], MD, PhD; Kota Fukai[1], MD, PhD; Kei Sano[2], MD, PhD; Masaaki Korenaga[3], MD, PhD; Masayuki Tatemichi[1], MD, PhD

[1]Department of Preventive Medicine, School of Medicine, Tokai University, 143 Shimokasuya, Isehara, Japan

[2]Department of Ophthalmology, The Jikei University School of Medicine, Minatoku, Japan

[3]Hepatitis Information Center, The Research Center for Hepatitis and Immunology, National Institute of Global Health and Medicine, Japan Institute for Health Security, Ichikawa, Japan

Corresponding Author:
Kota Fukai, MD, PhD
Department of Preventive Medicine, School of Medicine, Tokai University, 143 Shimokasuya, Isehara, Japan

Abstract

Background: Public awareness campaigns and testing promotion must be strengthened to eliminate infections with hepatitis B and C viruses (HBV and HCV, respectively) by 2030. Although public health campaigns using various forms of advertising are widely implemented, the most appropriate channels for viral hepatitis testing remain unclear.

Objective: This study aims to identify web services and digital advertising channels appropriate for promoting HBV and HCV testing, segmented by prior testing history and the desire for hepatitis virus testing.

Methods: A nationwide cross-sectional online survey of Japanese adults aged 20 to 69 years was conducted. The respondents answered questions regarding viral hepatitis testing status, routinely used web services (180 options), and exposure to digital advertising (25 options). Correspondence analysis was used to visualize relationships among testing segments, web services, and digital advertising. For individuals classified as “never having been tested and wishing to be tested,” channel-specific alignment was quantified using cosine θ. Sensitivity analyses were conducted by repeating the correspondence analysis after excluding respondents uncertain about their testing history and by fitting modified Poisson regression models with robust variance to estimate prevalence ratios and 95% CIs.

Results: Of the 2000 respondents (1011 male and 989 female), 18% (n=359) reported prior HBV and HCV testing, and 22.1% (n=441) were unsure whether they had ever been tested. Web services characteristically associated with “never having been tested and wishing to be tested” included Lawson (convenience store: cosine θ=0.989) and Cosme (shopping: cosine θ=0.987). The corresponding digital advertising channels included in-store and storefront screens at Welcia (pharmacy chain: θ=0.994) and Lawson (cosine θ=0.937). Segment-specific patterns varied according to age group and sex. Sensitivity analyses excluding the unsure group showed similar patterns. Modified Poisson regression results were also consistent; for example, Lawson web service use was associated with a desire for hepatitis virus testing (prevalence ratio 1.75, 95% CI 1.22 - 2.52).

Conclusions: In Japan, the convenience store chain Lawson was a frequently used touchpoint, both online and offline, among individuals seeking viral hepatitis testing. Future studies are needed to determine whether implementing awareness-raising activities through Lawson can increase the uptake of testing and subsequent treatment.

(J Med Internet Res 2026;28:e89585) doi:10.2196/89585

KEYWORDS

hepatitis B; hepatitis C; digital advertising; web services; correspondence analysis

Introduction

The World Health Organization (WHO) has set the goal of eliminating hepatitis B virus (HBV) and hepatitis C virus (HCV) infections by 2030 [1]. In Japan, over 70% of patients with hepatocellular carcinoma are infected with HBV or HCV [2,3]. In 2015, approximately 2 million individuals were infected with HBV or HCV, of whom approximately 1 million were undiagnosed or untreated [4]. From 2015 to 2022, a decreasing trend in acute-onset HBV and HCV hepatitis was reported, partially attributable to strengthened surveillance systems and, in the case of HCV, to remarkable advances in treatment with direct-acting antivirals, which have dramatically reduced the disease burden [5]. However, the burden of viral hepatitis is

XSL•FO
RenderX

predicted to persist until 2035, and more public health measures are required to achieve this goal [4].

In Japan, viral hepatitis testing is offered through community screening programs, health checkups, public health centers, prenatal care, and blood donations [6]. Among these, annual workplace health checkups, which cover many working-age adults, represent a particularly important opportunity for hepatitis virus testing [7]. Several companies have used this opportunity to implement HBV and HCV testing. Nevertheless, viral hepatitis screening is not part of legally required tests, resulting in only a subset of employees undergoing testing [7]. Furthermore, testing is frequently avoided due to social stigma and workplace discrimination [8]. To effectively address these challenges and enhance the response to viral hepatitis, the Ministry of Health, Labor and Welfare has formally urged employers and relevant organizations nationwide to collaborate to facilitate testing and ensure that employees who test positive receive appropriate medical treatment [9]. Hence, our research team has conducted awareness campaigns targeting top management and staff responsible for workplace health checkups within the human resources and general affairs departments [10]. To strengthen countermeasures against viral hepatitis, it is necessary to promote awareness campaigns among not only key workers but also the general population.

Recently, health awareness campaigns using websites and social media advertisements have become increasingly prevalent [11]. For example, as part of a viral hepatitis awareness campaign, educational columns and videos have been distributed via websites such as “Let’s Learn About Hepatitis” [12]. Information on the hepatitis virus has been widely disseminated to the public. However, effective health campaigns require accurately defining the target population and delivering content that is appropriately tailored to their specific attributes [13]. A scoping review of HCV media campaigns found that their effectiveness varied, emphasizing the need for customized approaches [14]. Accordingly, HCV care has been incorporated into opioid treatment programs to improve access for high-risk groups [15]. Additionally, social media strategies have been explored to identify individuals with HBV [16]. Overall, these observations indicate that designing interventions specifically for different populations is crucial for success.

Public health marketing for viral hepatitis should be tailored to a person’s testing history and results. One effective strategy to encourage more people to get tested is to guide those who have never been tested but are interested toward specific locations where they can access testing. However, there is limited knowledge regarding the most effective ways to share information about hepatitis testing with individuals who seek it.

This study aims to identify which web services and digital advertisements are most effective for promoting hepatitis testing in Japan, using groups defined by their prior testing history and interest in testing. The results of this study will inform the development of digital advertising strategies to more effectively disseminate information and promote hepatitis testing. These findings on public health advertising strategies can be applied to a range of health issues.

Methods

Study Design and Selection of Respondents

This nationwide cross-sectional study was conducted in August 2025 using a Japanese-language online questionnaire. Rakuten Insight Inc, a Tokyo-based web research firm, supported the survey. Panelists were users who registered through the company’s standard process and could be invited to specific surveys based on their profile information. Rakuten Insight has been used in previous peer-reviewed public health research. Respondents were recruited from the company’s registry via email and app push notifications. Invitations were randomly sent to individuals with active accounts in the past year who met the inclusion criteria as follows: aged 20 to 69 years, residing in Japan, and consenting to participate. Participants first answered screening questions; only those eligible completed the full questionnaire. Stratified sampling was carried out across 20 segments, divided by sex (male and female) and 5-year age strata, reflecting population estimates from Japan’s 2020 Population Census [17]. Surveys continued until 2000 valid responses were obtained, with target quotas set for each segment. The sample size was determined by the study budget and the maximum feasible recruitment. This study followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for cross-sectional observational studies (Checklist 1) [18], and web survey reporting adhered to the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) standards (Checklist 2) [19].

As this was a closed survey administered by a commercial online survey company, respondents were required to log in using their unique panel numbers before accessing the questionnaire. Once a respondent completed the survey, the questionnaire was no longer displayed. The IP addresses or cookies were not included in the survey. The survey consisted of 1 question per page, totaling 8 questions. No final review of responses was conducted.

Assessment of HBV and HCV Testing History and Desire for Testing

Information regarding the history of hepatitis virus testing and the desire for testing was obtained using a questionnaire. The respondents were asked to indicate the statement that best described their HBV and HCV testing status: (1) having been tested and knowing the results, (2) having been tested but not knowing the results, (3) never having been tested and wishing to be tested, (4) never having been tested and not wishing to be tested, and (5) unsure if tested. In Japan, public guidance recommends that all individuals undergo hepatitis virus testing at least once in their lifetime; therefore, this study focused on ever-tested status and the current desire for testing [7].

Assessment of Daily Used Web Services

Respondents were asked to complete multiple-choice items on regularly used web services. The survey was conducted annually by the research firm with all panel respondents. Data collected over 1 year, from November 2024 (when the questionnaire was distributed) through the completion of this survey, were analyzed. The items were synthesized from 28 genres: video

XSL•FO
RenderX

streaming, video sharing, music, social networking services, search engines, news, newspapers, magazines, books, calendars, hobbies, learning, maps, points, beauty, finance, general e-commerce, electronics stores, retail, manufacturer e-commerce, online shopping, flea markets, games, cooking recipes, restaurants, food delivery, travel, and weather. Multimedia Appendix 1 lists 180 types of web services, including Google, YouTube, Amazon, Netflix, Wikipedia, Facebook, Instagram, X (formerly known as Twitter), LinkedIn, and Spotify.

Assessment of Daily Shown Digital Advertising

Respondents were asked about their routine exposure to digital advertising by broad genres rather than brand names, such as convenience stores, supermarkets, shopping malls, drugstores, discount stores, transit media including in-car train screens, stations, bus stops, taxis, large outdoor video screens, vending machines, other noncommercial facilities, and other digital advertising (Multimedia Appendix 2).

Other Measurements

To characterize the respondents, information was collected on sex, age, educational background, household income, marital status, and employment status. Sex was self-reported, with respondents indicating whether they were male or female. Age was determined using the age on the date of response, based on the date of birth.

Regarding their educational background, respondents were asked to select the most appropriate option from the following: middle school, high school, vocational school, junior or technical college, university, graduate school, currently enrolled, or other.

The participants were further asked to select the most appropriate option from the following choices regarding pretax household income for 2024: up to JP ¥4 million, JP ¥4.01 - 6 million, JP ¥6.01 - 8 million, JP ¥8.01 - 10 million, JP ¥10.01 - 12 million, JP ¥12.01 - 15 million, and over JP ¥15 million (JP ¥159=US $1).

Marital status was defined either as married or unmarried. Regarding employment status, individuals who were currently working were asked about their main occupations and industries. Occupations were classified into 12 categories based on the Japan Standard Occupational Classification (Fifth Revision, December 2009) [20] as follows: administrative and managerial workers; professional and engineering workers; clerical workers; sales workers; service workers; security workers; agriculture, forestry, and fishery workers; manufacturing process workers; transport and machine operation workers; construction and mining workers; carrying, cleaning, packaging, and related workers; and workers unclassified by occupation [20]. Industries were classified into 20 categories based on the Japan Standard Industrial Classification: agriculture and forestry; fisheries; mining and quarrying of stone and gravel; construction; manufacturing; electricity, gas, heat, and water; information and communications; transport and postal services; wholesale and retail trade; finance and insurance; real estate and goods rental and leasing; scientific research, professional, and technical services; accommodation, food, and drink services; living-related, personal, and amusement services; education and learning support; medical, health care, and welfare; compound services; other services; government; and unclassified industries [21].

Statistical Analysis

The numbers and percentages of respondents were compiled by sex, age, education, income, marital status, and HBV or HCV testing status. To assess selection bias in the survey, we compared the distributions of respondents' educational backgrounds, income, occupations, industries, and residential regions with those from the 2020 Population Census [17]. The respondents' selection status was tabulated for 180 types of web services and 25 types of digital advertisements. As all questionnaire items were set as mandatory, no data were missing.

Correspondence analysis was used to identify the web services and digital advertising appropriate for raising awareness, tailored to the history of viral hepatitis testing and the current desire for testing. Correspondence analysis is frequently used in marketing analysis [22], as it can handle large amounts of categorical data and has been employed to examine relationships between purchasing trends across numerous products and buyer attributes. In public health studies, this method has also been applied to verify the distribution of categorical data [23]. Our previous research targeting key individuals for workplace health checkups also employed this method, successfully identifying prioritized information on viral hepatitis and appropriate media for conducting campaigns [10]. Ideally, a comprehensive representation of the distribution across the 5 HBV or HCV response categories would necessitate 4 dimensions to capture the total variance (100% inertia). However, as 4D models are not practically interpretable for visualization, we selected 3 dimensions as the most suitable option for effective exploratory mapping. Statistical analyses were performed using KH Coder software (version 3.0, SCREEN Advanced System Solutions Co Ltd) [24].

To confirm the validity of the responses, a correspondence analysis was performed to visualize web service usage patterns by respondents' sex and age. A 3D visualization of the correspondence analysis results was generated to characterize web service usage patterns associated with each viral hepatitis testing history and testing desire. To quantitatively evaluate the web service characteristics of groups willing to undergo testing (willing group), we assessed the deviation from the origin of each web service relative to the willing group using cosine θ. The same analysis was conducted for the digital advertisements. Since digital service usage patterns can differ by sex and age, the analysis was stratified by sex (male and female) and age group (20 - 49 y and 50 - 69 y).

Responses regarding HBV or HCV testing history and intention were classified into 5 mutually exclusive categories: tested and known result, tested but unknown result, never tested and willing to be tested, never tested and unwilling to be tested, and unsure if previously tested. The category "unsure if tested" was treated as an independent variable, as uncertainty about prior testing was considered meaningful for public health communication and may reflect distinct factors, such as limited recall and incomplete communication of test results.

XSL•FO
RenderX

During the sensitivity analysis, we excluded respondents who selected "unsure if tested" and subsequently performed correspondence analysis to identify web services and digital advertising channels associated with the willing group. Additionally, we fitted a series of modified Poisson regression models with robust variance, treating each of the web services selected by at least 5% (100/2000) and each of the digital advertisements as binary outcomes. Although the correspondence analysis used all web services, the multivariate regression analysis included only selected web services to adjust for covariates. The main independent variable was the desire for viral hepatitis testing (never tested but want to be tested vs all other responses). All models were adjusted for sex, age, household income, educational background (less than university vs university or higher), and marital status. Models with web services as outcomes were further adjusted for the total number of web services selected (0‐180); models with digital advertisements as outcomes were further adjusted for the total number of advertisements selected (0‐25). Age, sex, household income, educational attainment, and marital status were included as covariates as they likely influence prior testing history and desire for hepatitis virus testing [25]. Moreover, the total number of selected web services or digital advertising channels was included, as respondents demonstrating a higher overall propensity to choose multiple digital options are, by definition, more likely to exhibit positive results for any specific outcome. The detailed variable definitions and coding schemes (web services and digital advertising channels) are presented in Multimedia Appendix 3. For each outcome, we estimated prevalence ratios (PRs) and 95% CIs for the desire for viral hepatitis testing, with a 2-sided significance level of 5%. The statistical software used was R (version 4.4.2; R Foundation for Statistical Computing) [26].

Ethical Considerations

This study was approved by the research ethics committee of the Tokai University School of Medicine (23R039). Informed consent was obtained from all participants before they responded to the questionnaire. The consent screen outlined the study's purpose, procedures, voluntary participation, and the process for withdrawing consent. Respondents were allowed to withdraw their consent up to 3 months after survey completion. To protect privacy, Rakuten Insight removed all personally identifiable information before data delivery. These measures were designed to minimize the risk of information leakage. Participants earned panel points based on the provider's standard incentive system: those who completed the survey received 5 points, each equivalent to JP ¥1 (US $0.6), which were redeemable within the provider's partner network.

Results

Participants

Of the 2,379,232 registered individuals in the survey panel, 357,272 were invited to participate in the study via email and push notifications. Of these, 5621 responded to the screening question (Figure 1). A total of 3278 individuals were excluded due to declining participation (n=2676), incomplete responses, or reaching quota targets (n=602). Among the 2343 individuals who answered all questions, 2000 were selected for analysis using the following process. First, respondents were excluded if their response times were below a threshold defined as the median response time multiplied by a device factor (0.3 for PC and 0.2 for smartphone). Subsequently, a random number was assigned to each valid sample using the RAND function in Microsoft Excel. The samples were ranked in ascending order according to assigned random numbers, and the top 2000 were selected for inclusion in the final dataset. The invitation-to-screening proportion was 1.6% (5621/357,272), while the screening-to-analysis proportion was 85.4% (2000/2343).

XSL•FO
RenderX

Figure 1. Flowchart of participant selection for analysis.

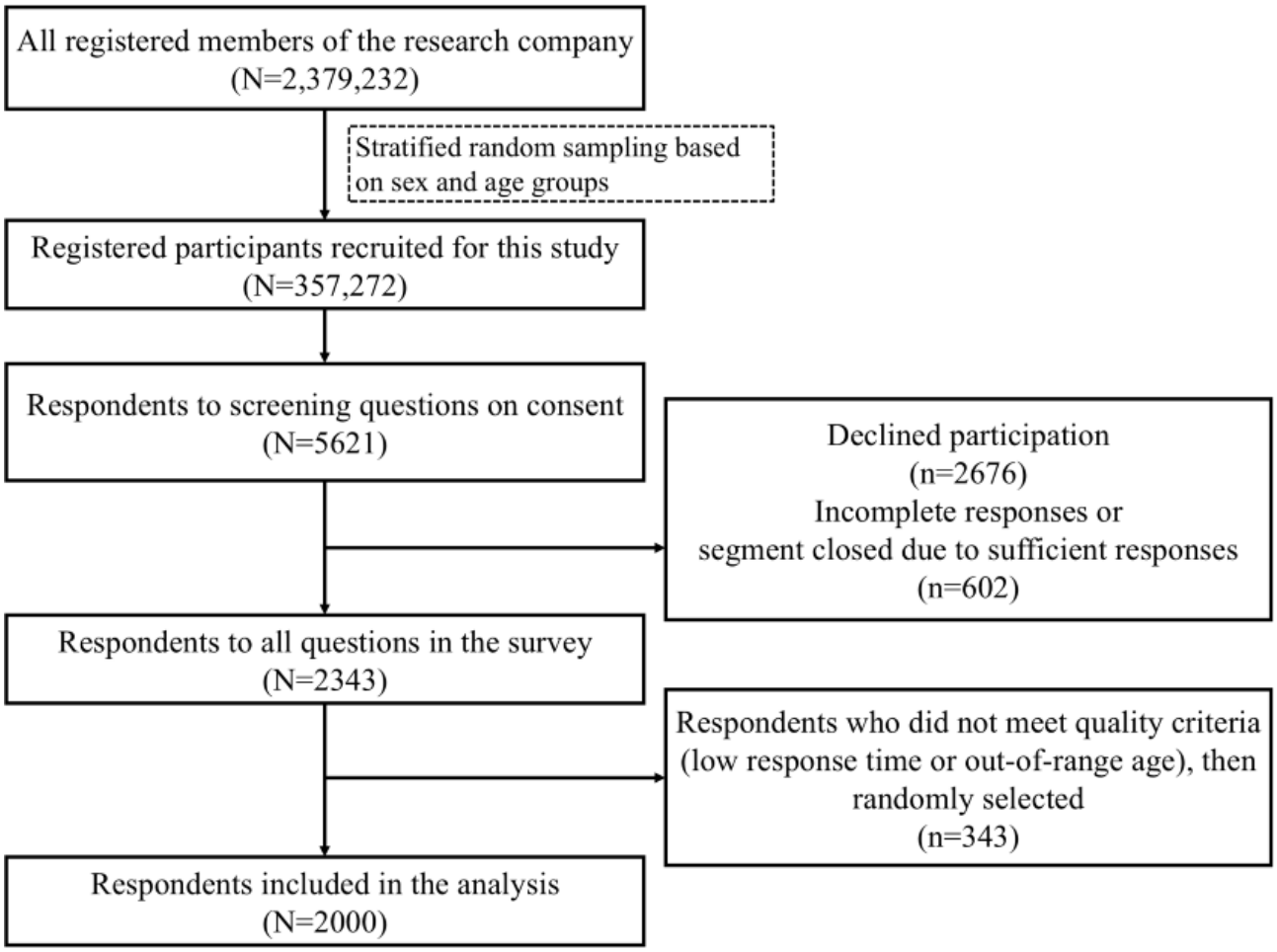

Characteristics of the Respondents

Responses were obtained from 1011 male and 989 female participants (Table 1). Most respondents (484/2000, 24.2%) were aged 50 to 59 years. The educational background was university or graduate school for 56.3% (569/1011) of the male and 37.6% (372/989) of the female participants. A household income of less than JP ¥4 million was the most common, accounting for 31.4% (317/1011) of male and 41.2% (407/989) of female participants. Regarding marital status, 48.4% (489/1011) of male and 52.1% (515/989) of female participants were married. The percentages of respondents who had been tested for HBV or HCV and knew their results were 14.4% (146/1011) among male and 18.7% (185/989) among female participants. The percentage of individuals who were never tested but wanted to be tested was 26% (263/1011) among male and 23.4% (231/989) among female participants.

To assess selection bias, we compared the distributions of educational background (Multimedia Appendix 4), household income (Multimedia Appendix 5), occupation and industry (Multimedia Appendix 6), and residential region (Multimedia Appendix 7) with those from the 2020 Population Census of Japan.

Regarding educational attainment, national statistics indicate that 27.8% of male and 17.9% of female participants hold university degrees, whereas in this study, 50.1% of male and 34.8% of female participants hold university degrees.

Regarding household income, the census showed that 48.9% of households earned less than JP ¥4 million, whereas only 36.2% of respondents in this study were in this income bracket. Among male participants, administrative and managerial workers accounted for 2.1% in the census and 14.5% in this study, sales workers accounted for 12.4% in the census and 5.8% in this study, and manufacturing process workers accounted for 13.4% in the census and 6.3% in this study. The distribution of industries among the respondents in this study was similar to that in the census.

Regarding residential regions, participants were overrepresented in Tokyo (male: +2.41% points; female: +2.7% points), Kanagawa (male: +2.24% points; female: +2.47% points), and Osaka (male: +2.13% points; female: +0.70% points) compared with the national population distribution.

XSL•FO
RenderX

Table . Respondent characteristics.

Characteristics	All (N=2000), n (%)	Male (n=1011), n (%)	Female (n=989), n (%)
Age (y)			
20 - 29	339 (17)	175 (17.3)	164 (16.6)
30 - 39	351 (17.6)	180 (17.8)	171 (17.3)
40 - 49	433 (21.7)	219 (21.7)	214 (21.6)
50 - 59	484 (24.2)	244 (24.1)	240 (24.3)
60 - 69	393 (19.7)	193 (19.1)	200 (20.2)
Educational background			
Middle or high school	534 (26.7)	254 (25.1)	280 (28.3)
Vocational school, junior college, or technical college	434 (21.7)	130 (12.9)	304 (30.7)
University or graduate school	941 (47.1)	569 (56.3)	372 (37.6)
Currently enrolled	73 (3.7)	47 (4.6)	26 (2.6)
Other	18 (0.9)	11 (1.1)	7 (0.7)
Household income (JP ¥[a])			
≤4 million	724 (36.2)	317 (31.4)	407 (41.2)
4.01 - 6 million	414 (20.7)	205 (20.3)	209 (21.1)
6.01 - 8 million	358 (17.9)	192 (19)	166 (16.8)
8.01 - 10 million	236 (11.8)	138 (13.6)	98 (9.9)
10.01 - 12 million	133 (6.7)	79 (7.8)	54 (5.5)
12.01 - 15 million	73 (3.7)	42 (4.2)	31 (3.1)
>15 million	62 (3.1)	38 (3.8)	24 (2.4)
Marital status			
Married	1004 (50.2)	489 (48.4)	515 (52.1)
Unmarried	996 (49.8)	522 (51.6)	474 (47.9)
History of HBV[b] and HCV[c] testing and current desire for testing			
Tested and known result (known)	331 (16.6)	146 (14.4)	185 (18.7)
Tested but unknown result (unknown)	28 (1.4)	19 (1.9)	9 (0.9)
Never tested, want to be (willing)	494 (24.7)	263 (26)	231 (23.4)
Never tested, do not want to be (unwilling)	706 (35.3)	358 (35.4)	348 (35.2)
Unsure if tested (unsure)	441 (22.1)	225 (22.3)	216 (21.8)

[a]JP ¥159=US $1.

[b]HBV: hepatitis B virus.

[c]HCV: hepatitis C virus

Characteristics of Web Services and Digital Advertising

The top 50 selected web services are presented in Multimedia Appendix 8. The top 10 selected services were the general e-commerce company Rakuten Ichiba (n=863), Google (n=740), Amazon (n=690), YouTube (n=669), the search engine Yahoo! Japan (n=658), Yahoo! News (n=624), Google Maps (n=589), Yahoo! Weather & Disaster (n=503), the travel reservation site Rakuten Travel (n=455), and the restaurant review site Tabelog (n=420).

The ranking of the 25 types of digital advertising is presented in Multimedia Appendix 9. The most frequent responses were “Nothing particular” at 53.1% (1062/2000), “FamilyMart in-store and storefront digital advertising” at 18.3% (365/2000), “Digital advertising inside trains” at 17.9% (358/2000), “Digital advertising at stations and inside station facilities” at 17.3% (346/2000), and “Seven-Eleven in-store and storefront digital advertising” at 16.1% (322/2000).

XSL•FO
RenderX

Correspondence Analysis Between Respondent Characteristics and Web Services or Digital Advertisements

The results of the correspondence analysis between sex and web services are presented in Multimedia Appendix 10. The specific web services used by male participants included internet technology services such as IT media and DMM.com, as well as newspaper sites such as Nikkei and Yomiuri. Among female participants, the most common services were cosmetic services and cooking recipe sites.

The results of the correspondence analysis between age groups and web services are shown in Multimedia Appendix 11. In their 20s, Uber Eats (food delivery), TikTok (video sharing), Spotify (music), and gaming services were prominent; in their 30s and 40s, Cosme (cosmetic shopping), Rakuten Fashion (general e-commerce), HOT PEPPER Beauty (beauty salon booking and review site), and Cookpad (recipe sharing and cooking website); in their 50s, Yahoo sites (search engine and relevant services) and Facebook; and in their 60s, Japanet (TV shopping), MSN Japan (online news and portal site), and Belluna (mail-order and online shopping site).

Correspondence Analysis Between HBV and HCV Testing Characteristics and Web Services or Digital Advertising

Figure 2 shows the 3 dimensions: correspondence analysis results for the respondents' HBV and HCV testing statuses, the 60 web services farthest from the origin, and all 25 digital advertisements. Web services specific to people who had not been tested and wanted to be tested were Lawson (convenience store) with a cosine θ of 0.989 and Cosme (cosmetic shopping) with a cosine θ of 0.987. Digital advertising targeted at individuals who had not been tested and wanted to be tested included in-store and storefront advertising at Welcia (pharmacy chain; cosine θ=0.994) and Lawson (convenience store; cosine θ=0.937).

Figure 2. Associations of viral testing history and the desire for testing with (A) web services and (B) digital advertising.

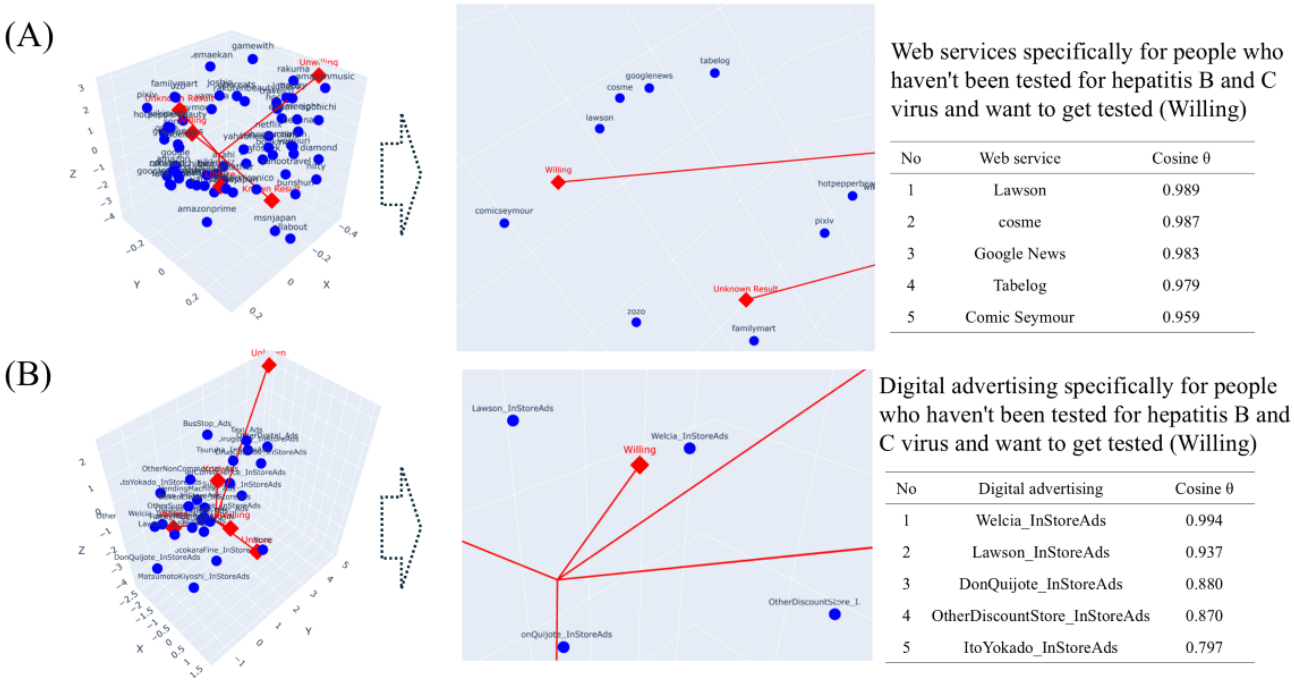

No	Web service	Cosine θ
1	Lawson	0.989
2	cosme	0.987
3	Google News	0.983
4	Tabelog	0.979
5	Comic Seymour	0.959

No	Digital advertising	Cosine θ
1	Welcia_InStoreAds	0.994
2	Lawson_InStoreAds	0.937
3	DonQuijote_InStoreAds	0.880
4	OtherDiscountStore_InStoreAds	0.870
5	ItoYokado_InStoreAds	0.797

The results of the correspondence analysis between HBV and HCV testing status and web services by sex (male and female) and age group (<50 y or >50 y) are shown in Figures 3 and 4, respectively. Depending on sex and age group, web services specific to individuals who had not been tested and who wanted to be tested were identified. Among individuals who intended to undergo testing, the most frequently used were Nikkei Biz (news media) and Yahoo-related services (search engine) for male participants, Demae-can (food delivery) and e-book platforms (Comic Seymour and Mecha Comic) for female participants, Google-related services for those aged 20 to 49 years, and Zozotown (fashion e-commerce site) and Booking.com (hotel booking site) for those aged 50 to 69 years. In the correspondence analysis, the first 2 dimensions explained most of the variability in the distribution of web services: among male participants, 73.8% of the total inertia (56.3% on axis 1 and 17.5% on axis 2), and among female participants, 71.6% (51% on axis 1 and 20.6% on axis 2). For participants aged 20 to 49 years, the first 2 dimensions accounted for 82.1% of total inertia (70.5% on axis 1 and 11.6% on axis 2); for those aged 50 to 69 years, they accounted for 68% (40.7% on axis 1 and 27.3% on axis 2).

XSL•FO
RenderX

Figure 3. Relationship between hepatitis B virus (HBV) or hepatitis C virus (HCV) screening experience and web services by sex. (A) Results for male participants. (B) Results for female participants.

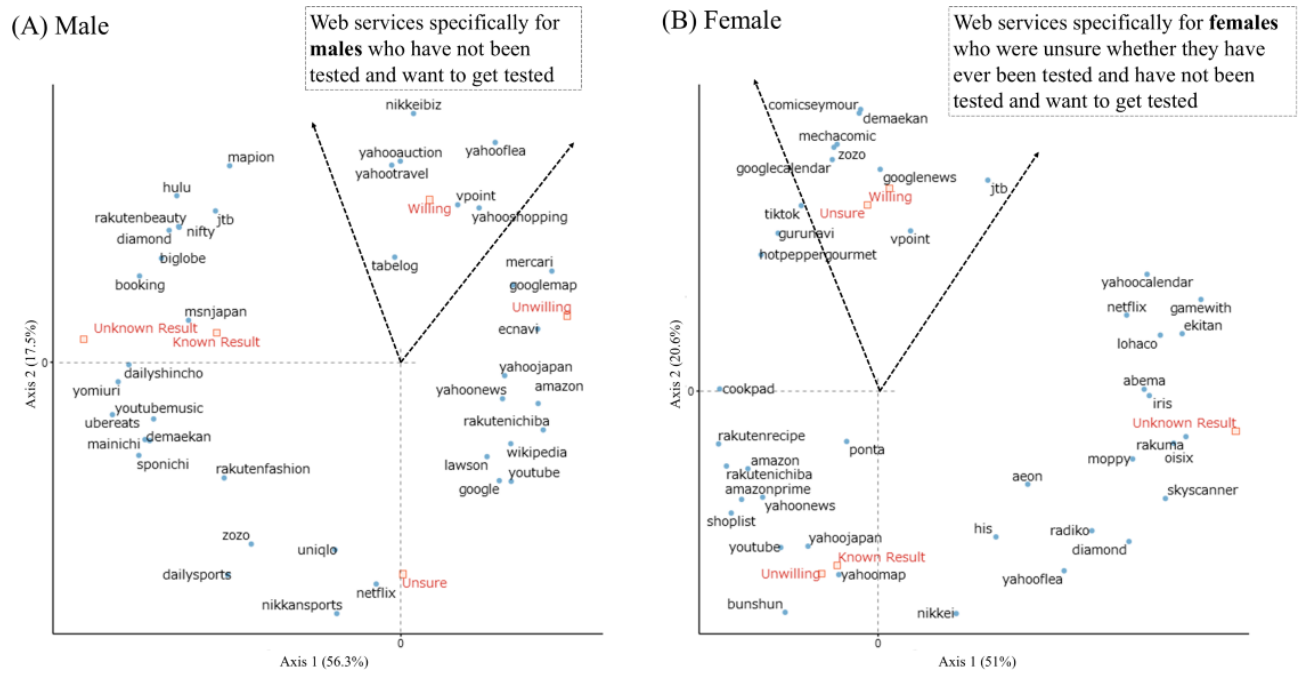

Figure 4. Relationship between hepatitis B virus (HBV) or hepatitis C virus (HCV) screening experience and web services by age groups. (A) Results for respondents aged 20 to 49 years. (B) Results for respondents aged 50 to 69 years.

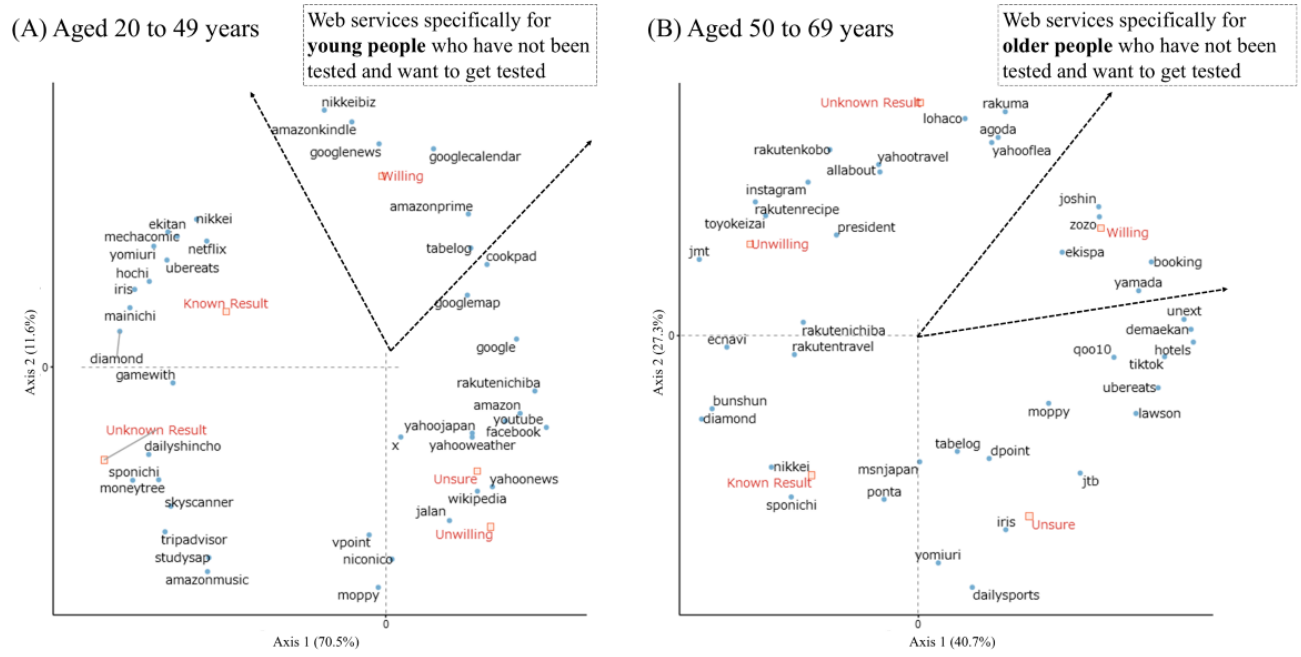

Sensitivity Analysis

The findings from the sensitivity analysis of web services using correspondence analysis are detailed in Multimedia Appendix 12. Web services identified as particularly relevant for individuals who had not been tested but wanted to be included are Google News (cosine θ=0.98), Zozotown—a fashion e-commerce platform—(cosine θ=0.96), Nikkei Biz (news media; cosine θ=0.94), Comic Seymour (e-book platform; cosine θ=0.92), and Lawson (convenience store; cosine θ=0.90).

Similarly, the outcomes of the sensitivity analysis of digital advertising channels indicated that Lawson in-store and storefront advertisements (cosine θ=0.96), as well as Matsumoto Kiyoshi in-store and storefront advertisements (cosine θ=0.92), were specifically associated with individuals who had not been tested but wanted to be (Multimedia Appendix 13).

The results of the sensitivity analysis of the web services using a modified Poisson regression revealed a significant association for Lawson (Multimedia Appendix 14). The highest PRs for the desire for hepatitis testing were observed for Lawson (PR 1.75, 95% CI 1.22 - 2.52), Zozotown (PR 1.67, 95% CI 1.18 - 2.36), Google Calendar (PR 1.63, 95% CI 1.21 - 2.20), and Google News (PR 1.61, 95% CI 1.25 - 2.07).

XSL•FO
RenderX

The sensitivity analysis of digital advertising channels, using modified Poisson regression, demonstrated that the highest PRs were observed for in-store advertisements at other discount stores (PR 2.64, 95% CI 1.51 - 4.59), Ito Yokado (PR 1.72, 95% CI 1.13 - 2.63), DonQuijote (PR 1.72, 95% CI 1.18 - 2.51), Welcia (PR 1.71, 95% CI 1.13 - 2.59), and Lawson (PR 1.49, 95% CI 1.19 - 1.87; Multimedia Appendix 15).

Discussion

Principal Findings

We identified web services and digital advertisements for use in HBV and HCV campaign channels, based on testing history and expressed interest in testing. Individuals who had never been tested for hepatitis but wanted to be tested were regularly exposed to the website and in-store digital advertisements of Lawson, one of the most famous convenience store chains. Segmenting the target population by sex or age group revealed distinct preferences for web services. These observations suggest that effective digital public health campaigns should account for specific attributes of the target population, including their testing history and interests, when determining appropriate communication channels.

Comparison With Prior Work

Lawson might be crucial in conducting viral hepatitis awareness campaigns, as individuals interested in HBV and HCV testing use Lawson's web services and digital advertisements. In 2013, Lawson collaborated with local medical and public health organizations to distribute application forms for free hepatitis testing at selected stores in Saitama Prefecture, illustrating the potential of convenience store chains as access points for hepatitis screening campaigns [27]. Convenience stores are gaining attention as venues for health promotion. For example, initiatives have been undertaken to improve access to healthy foods by stocking them in these stores [28,29]. Convenience stores are key locations within the community, and staff education has been implemented to strengthen the community network for dementia [30]. These initiatives focused on location characteristics, such as the presence of unhealthy food and local staff members. The results of this study enable us to consider intervention settings from different perspectives by designing interventions tailored to the characteristics of visiting customers.

In public health communication, it is preferable to tailor information dissemination to the characteristics and interests of the target audience [31]. As advertising is expensive, cost-effective campaigns can be implemented by targeting those who have never undergone an examination but are willing to be tested. Several interventions have been tailored to specific target groups based on their characteristics [32]. An educational intervention using an app to provide information about HBV screening in their native language to Asian Americans resulted in increased consultations with health care providers and subsequent testing [33]. In the United Kingdom, a combination intervention of sending tailored letters to encourage high-risk and untreated patients with HCV to seek care, training staff, and implementing awareness-raising notices increased HCV testing and treatment, demonstrating good cost-effectiveness [34]. In an intervention in New York, a multidisciplinary team implemented efforts to remove barriers to care for low-income individuals, the homeless, and drug users who needed support, leading many to undergo HCV testing and complete treatment [35]. Consistent with these results, community-based interventions at convenience stores may be a promising approach.

When developing advertising strategies for public health, it is important to consider the sex and age of the target population. The results of this study revealed distinct patterns in web service usage and digital advertising exposure by sex and age. Therefore, stratified by sex and age, the characteristic channels differed according to viral hepatitis testing status. Viral hepatitis testing is recommended at least once in a lifetime, regardless of sex or age [36]. For this awareness campaign, it was preferable to use the analysis results based on the entire target population. When focusing on specific age groups or sex, awareness strategies should be developed based on analysis results adjusted for the characteristics of that group. For example, when implementing HBV awareness campaigns targeting pregnant female individuals, awareness campaigns should be developed by referencing our results in female participants.

Fewer respondents than expected reported having been tested for viral hepatitis. In Japan, it is estimated that more than half of the population has been tested for viral hepatitis [6], far exceeding the 18% testing rate in this study; however, an additional 22.1% were unsure of whether they had been tested. A study investigating Japanese people's knowledge of how viral hepatitis testing is performed reported that 93% of respondents identified blood tests, whereas only 7% mentioned urine tests [37]. In addition, physicians do not inform patients of negative results. A survey targeting ophthalmologists who frequently conducted viral hepatitis tests before surgery revealed that approximately 40% did not explain the results to patients when the test was negative [38]. Even when negative results were verbally communicated to patients, only 42% recalled them 1 year later, leading to the recommendation of using negative cards—written records certifying that the individual tested negative for HBV and HCV—for communicating results [39]. Negative cards are expected to enable information sharing with other health care professionals and to avoid duplicate viral hepatitis testing.

This study neither implemented nor evaluated specific public health awareness campaigns. Instead, we present a methodological case study illustrating how correspondence analysis can inform the development of public health strategies that leverage web services and digital advertising. Medical professionals have used correspondence analysis to address critical issues through dimension reduction, data simplification, clustering, and classification [40]. This study visualized the distributions of 180 web services and 25 digital advertisements. Correspondence analysis has been used to clarify the relationship between age and headache type, to identify associations between personality type and various diagnoses [41], and to clarify the key challenges prioritized by each profession when pursuing innovation in public hospitals [42]. On the basis of these examples, correspondence analysis may be useful as an exploratory tool for examining and planning public health

XSL•FO
RenderX

advertising strategies. A sensitivity analysis using multivariate models adjusted for respondent characteristics yielded patterns similar to those observed in the main correspondence analysis. This consistency indicates that the study's findings are robust to potential confounding from participant attributes.

In this study, respondents tended to have higher educational attainment, higher household income, and a greater proportion of managerial occupations than the working population estimated from the 2020 Population Census. This is consistent with prior research showing that online surveys often include individuals with higher education and income levels than those in national statistics [43,44]. Higher socioeconomic status has been associated with better health literacy, which may explain the interest in hepatitis testing [45]. Moreover, because web service use was assessed among individuals who chose to participate in an online survey, respondents might be more familiar with online services. Thus, the level of web service use in this study might be higher than that in the general population, and our findings reflect patterns among digitally engaged individuals.

Limitations

This study has certain limitations. First, potential selection bias warrants consideration. Compared with national statistics, our study included a higher proportion of individuals with university-level education, lower annual income, and a higher proportion in managerial or professional roles than in production-related occupations. Furthermore, respondents were more likely to reside in urban areas. Given that increased educational attainment, income, and occupational status may correlate with heightened health awareness and interest in viral hepatitis testing, the observed associations may not be generalizable to the broader population. Thus, the findings must be interpreted with caution when extending them beyond digitally engaged adults. Future studies should address this issue by using non–internet-based recruitment methods to enhance external validity.

Second, measurement biases regarding web services and digital advertising exist. To mitigate this effect, we conducted a correspondence analysis of web service usage by sex and age, revealing patterns consistent with real-world behaviors—such as female individuals preferring cosmetic sites and younger individuals favoring TikTok. Subsequent research should seek to collect objective data regarding web service use.

Third, recall bias may have influenced the reporting of HBV or HCV testing history. Notably, 22.1% of respondents indicated uncertainty about previous testing, which may have led to misclassification of past testing status. Additionally, exposure misclassification may have occurred because digital advertising exposure was self-reported and may not accurately reflect media saturation or individual engagement with advertisements.

Finally, we were unable to verify the effectiveness of advertising campaigns in changing behavior. Further investigation is necessary to determine whether interventions that incorporate appropriate web services and digital advertising can facilitate viral hepatitis testing, detection, and treatment.

Conclusions

This study identified distinct channel-specific patterns in the promotion of viral hepatitis testing in Japan. Lawson consistently served as a primary point of contact across web services and digital advertising among respondents who had never been tested but expressed interest in testing. Subgroup analyses revealed differences by sex and age, while sensitivity analyses confirmed comparable overall trends. These results support the development of targeted digital communication strategies and offer a practical framework for designing effective public health campaigns for HBV or HCV testing.

Acknowledgments

The authors acknowledge the support of Rakuten Insight Inc for encouraging recruitment and conducting a web-based survey using online panels. The authors are grateful to Editage for reviewing this manuscript for English language and clarity. Additionally, the authors used ChatGPT (GPT-5.2 Thinking) to assist with English-language editing [46].

Funding

This study was supported by the Ministry of Health, Labour and Welfare, Japan (grant: 23HC1001).

Data Availability

The datasets used and analyzed in this study are not publicly accessible due to confidentiality agreements; however, additional information regarding the study is available from the corresponding author upon reasonable request.

Authors' Contributions

K Sakai and MT conceptualized and designed this study. K Sakai collected the data and performed the statistical analysis. KF, YF, SN, K Sano, and MK contributed to the interpretation of the results. K Sakai wrote the manuscript. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
List of 180 types of web services.
[DOCX File, 44 KB - jmir_v28i1e89585_app1.docx]

Multimedia Appendix 2
List of 25 types of digital advertising.
[DOCX File, 23 KB - jmir_v28i1e89585_app2.docx]

Multimedia Appendix 3
Variables used in modified Poisson regression models for digital exposure and desire for hepatitis virus testing.
[DOCX File, 25 KB - jmir_v28i1e89585_app3.docx]

Multimedia Appendix 4
Distribution of educational background in this study and the Census 2020.
[DOCX File, 22 KB - jmir_v28i1e89585_app4.docx]

Multimedia Appendix 5
Distribution of household income in this study and the Census 2020.
[DOCX File, 34 KB - jmir_v28i1e89585_app5.docx]

Multimedia Appendix 6
Distribution of occupation and industry in this study and the census 2020.
[DOCX File, 27 KB - jmir_v28i1e89585_app6.docx]

Multimedia Appendix 7
Distribution of residential regions in this study and the 2020 Population Census.
[DOCX File, 33 KB - jmir_v28i1e89585_app7.docx]

Multimedia Appendix 8
The top-selected 50 web services.
[DOCX File, 26 KB - jmir_v28i1e89585_app8.docx]

Multimedia Appendix 9
The selection of 25 digital advertising.
[DOCX File, 24 KB - jmir_v28i1e89585_app9.docx]

Multimedia Appendix 10
Correspondence analysis between sex and web services.
[DOCX File, 1144 KB - jmir_v28i1e89585_app10.docx]

Multimedia Appendix 11
Correspondence analysis between age groups and web services.
[DOCX File, 1776 KB - jmir_v28i1e89585_app11.docx]

Multimedia Appendix 12
Sensitivity analysis of web services based on 3D correspondence analysis.
[DOCX File, 308 KB - jmir_v28i1e89585_app12.docx]

Multimedia Appendix 13
Sensitivity analysis of digital advertising channels based on 3D correspondence analysis.
[DOCX File, 284 KB - jmir_v28i1e89585_app13.docx]

Multimedia Appendix 14
Associations between desire for hepatitis virus testing and use of individual web services.
[DOCX File, 46 KB - jmir_v28i1e89585_app14.docx]

XSL•FO
RenderX

Multimedia Appendix 15
Associations between desire for hepatitis virus testing and exposure to individual digital advertising channels.
[DOCX File, 49 KB - jmir_v28i1e89585_app15.docx]

Checklist 1
STROBE checklist.
[DOCX File, 123 KB - jmir_v28i1e89585_app16.docx]

Checklist 2
CHERRIES checklist.
[DOCX File, 23 KB - jmir_v28i1e89585_app17.docx]

References

1. Global health sector strategies on, respectively, HIV, viral hepatitis and sexually transmitted infections for the period 2022-2030. : World Health Organization; 2022 URL: https://iris.who.int/server/api/core/bitstreams/6b18ac34-f56a-466e-9660-b5653cb52ef0/content [accessed 2026-03-13]
2. Ikeda M, Mitsunaga S, Shimizu S, et al. Current status of hepatocellular carcinoma in Japan. Chin Clin Oncol 2013 Dec;2(4):40. [doi: 10.3978/j.issn.2304-3865.2013.09.01] [Medline: 25841919]
3. Nakano M, Yatsuhashi H, Bekki S, et al. Trends in hepatocellular carcinoma incident cases in Japan between 1996 and 2019. Sci Rep 2022 Jan 27;12(1):1517. [doi: 10.1038/s41598-022-05444-z] [Medline: 35087153]
4. Tanaka J, Kurisu A, Ohara M, et al. Burden of chronic hepatitis B and C infections in 2015 and future trends in Japan: a simulation study. Lancet Reg Health West Pac 2022 May;22:100428. [doi: 10.1016/j.lanwpc.2022.100428] [Medline: 35637862]
5. Okushin K, Kanto T, Korenaga M, et al. Real‐world trends in acute viral hepatitis in Japan: a nationwide questionnaire‐based survey. Hepatol Res 2025 Mar;55(3):301-308. [doi: 10.1111/hepr.14123]
6. Korenaga M, Kanto T. Testing, diagnosis of viral hepatitis, and the follow-up policy in Japan. Glob Health Med 2021 Oct 31;3(5):308-313. [doi: 10.35772/ghm.2021.01072] [Medline: 34782874]
7. Tatemichi M, Furuya H, Nagahama S, et al. A nationwide cross-sectional survey on hepatitis B and C screening among workers in Japan. Sci Rep 2020 Jul 10;10(1):11435. [doi: 10.1038/s41598-020-68021-2] [Medline: 32651421]
8. Sasaki N, Wada K, Smith DR, Wang G, Ohta H, Shibuya A. Hepatitis screening in Japanese individuals of working age and prejudice against infected persons in the workplace. J Occup Health 2014;55(5):392-397. [doi: 10.1539/joh.12-0266-fs] [Medline: 23812028]
9. Request for cooperation regarding measures against viral hepatitis in the workplace [Report in Japanese]. : Ministry of Health, Labour and Welfare; 2023 URL: https://www.mhlw.go.jp/content/10900000/001076013.pdf [accessed 2026-03-13]
10. Sakai K, Nakazawa S, Fukai K, Furuya Y, Korenaga M, Tatemichi M. Marketing strategies for promoting workplace hepatitis B and C virus testing: a cross-sectional study using correspondence analysis in Japan. Front Public Health 2025;13:1522850. [doi: 10.3389/fpubh.2025.1522850] [Medline: 40115348]
11. An J, Kwak H, Qureshi HM, Weber I. Precision public health campaign: delivering persuasive messages to relevant segments through targeted advertisements on social media. JMIR Form Res 2021 Sep 24;5(9):e22313. [doi: 10.2196/22313] [Medline: 34559055]
12. Takeuchi Y, Ohara M, Kanto T. Nationwide awareness-raising program for viral hepatitis in Japan: the “Shitte kan-en” project. Glob Health Med 2021 Oct 31;3(5):301-307. [doi: 10.35772/ghm.2021.01063] [Medline: 34782873]
13. Dietrich T. Segmentation in social marketing: five steps to success. In: Dietrich T, Rundle-Thiele S, Kubacki K, editors. Segmentation in Social Marketing: Springer; 2017:77-92.
14. Etoori D, Desai M, Mandal S, Rosenberg W, Sabin CA. A scoping review of media campaign strategies used to reach populations living with or at high risk for hepatitis C in high income countries to inform future national campaigns in the United Kingdom. BMC Infect Dis 2023 Sep 26;23(1):629. [doi: 10.1186/s12879-023-08603-3] [Medline: 37752434]
15. Talal AH, Dharia A, Markatou M, et al. Facilitated telemedicine as a patient-centered, sociotechnical intervention to integrate hepatitis C treatment into opioid treatment programs and overcome the digital divide among underserved populations: qualitative study. JMIR Public Health Surveill 2025 Jul 16;11:e68854. [doi: 10.2196/68854] [Medline: 40669057]
16. Willem T, Zimmermann BM, Matthes N, Rost M, Buyx A. Acceptance of social media recruitment for clinical studies among patients with hepatitis B: mixed methods study. J Med Internet Res 2024 Aug 26;26:e54034. [doi: 10.2196/54034] [Medline: 39186322]
17. Population census [Article in Japanese]. e-Stat. 2020. URL: https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00200521&tstat=000001136464&cycle=0&year=20200&month=24101210&tclass1=000001136466 [accessed 2025-12-14]

XSL•FO
RenderX

18. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 2007 Oct 16;147(8):573-577. [doi: 10.7326/0003-4819-147-8-200710160-00010] [Medline: 17938396]
19. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
20. Japan Standard Occupational Classification (rev. 5th, December 2009) structure and explanatory notes [Website in Japanese]. Ministry of Internal Affairs and Communications. 2009. URL: https://www.soumu.go.jp/english/dgpp_ss/seido/shokgyou/co09-4.htm [accessed 2026-03-13]
21. Japan Standard Industrial Classification (rev. 13, October 2013) the underlying principles of the classification [Website in Japanese]. Ministry of Internal Affairs and Communications. 2013. URL: https://www.soumu.go.jp/english/dgpp_ss/seido/sangyo/san13-2.htm [accessed 2025-12-14]
22. Beh EJ, Lombardo R. Correspondence Analysis: Theory, Practice and New Strategies: John Wiley & Sons; 2014.
23. Costa PS, Santos NC, Cunha P, Cotter J, Sousa N. The use of multiple correspondence analysis to explore associations between categories of qualitative variables in healthy ageing. J Aging Res 2013;2013:302163. [doi: 10.1155/2013/302163] [Medline: 24222852]
24. Higuchi K. New quantitative text analytical method and KH Coder software. Jpn Sociol Rev 2017;68(3):334-350. [doi: 10.4057/jsr.68.334]
25. Hajek A, Kretzler B, König HH. Determinants of healthcare use based on the Andersen model: a systematic review of longitudinal studies. Healthcare (Basel) 2021 Oct 12;9(10):1354. [doi: 10.3390/healthcare9101354] [Medline: 34683034]
26. Core Team R. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2024. URL: https://www.r-project.org/ [accessed 2026-03-19]
27. News release. Lawson, Inc. URL: https://mldata.lawson.co.jp/company/news/year/077593/ [accessed 2025-12-14]
28. Haboush-Deloye AL, Knight MA, Bungum N, Spendlove S. Healthy foods in convenience stores: benefits, barriers, and best practices. Health Promot Pract 2023 May;24(1_suppl):108S-111S. [doi: 10.1177/15248399221147878] [Medline: 36999507]
29. Jernigan VBB, Williams M, Wetherill M, et al. Using community-based participatory research to develop healthy retail strategies in Native American–owned convenience stores: the THRIVE study. Prev Med Rep 2018 Sep;11:148-153. [doi: 10.1016/j.pmedr.2018.06.012] [Medline: 30003014]
30. Igarashi A, Matsumoto H, Takaoka M, Kugai H, Suzuki M, Yamamoto-Mitani N. Educational program for promoting collaboration between community care professionals and convenience stores. J Appl Gerontol 2020 Jul;39(7):760-769. [doi: 10.1177/0733464819871878] [Medline: 31478434]
31. Schmid KL, Rivers SE, Latimer AE, Salovey P. Targeting or tailoring? Maximizing resources to create effective health communications. Mark Health Serv 2008;28(1):32-37. [Medline: 18389854]
32. Rajkumar V, McCausland K, Lobo R. A rapid review of interventions to increase hepatitis B testing, treatment, and monitoring among migrants living in Australia. Int J Environ Res Public Health 2022 May 13;19(10):5947. [doi: 10.3390/ijerph19105947] [Medline: 35627481]
33. Khalili M, Kim NJ, Tsoh JY, et al. Health Within Reach-a patient-centered intervention to increase hepatitis B screening among Asian Americans: a randomized clinical trial. J Gen Intern Med 2022 Oct;37(13):3242-3250. [doi: 10.1007/s11606-021-07232-3] [Medline: 34993863]
34. Roberts K, Macleod J, Metcalfe C, et al. Cost effectiveness of an intervention to increase uptake of hepatitis C virus testing and treatment (HepCATT): cluster randomised controlled trial in primary care. BMJ 2020 Feb 26;368:m322. [doi: 10.1136/bmj.m322] [Medline: 32102782]
35. Ford MM, Johnson N, Desai P, Rude E, Laraque F. From care to cure: demonstrating a model of clinical patient navigation for hepatitis C care and treatment in high-need patients. Clin Infect Dis 2017 Mar 1;64(5):685-691. [doi: 10.1093/cid/ciw806] [Medline: 27940945]
36. Basic guidelines for promotion of control measures for hepatitis. Ministry of Health, Labour and Welfare. 2011. URL: https://www.mhlw.go.jp/bunya/kenkou/kekkaku-kansenshou09/pdf/hourei-17e.pdf [accessed 2025-12-14]
37. Masuda S, Hidaka I, Nakano S, et al. Knowledge enlightenment about viral hepatitis using a quiz at an awareness-raising event. Kanzo 2025;66(10):443-446. [doi: 10.2957/kanzo.66.443]
38. Ohara M, Ide T, Inoue J, et al. Follow-up survey on hepatitis virus testing among ophthalmologists. Kanzo 2025;66(9):407-410. [doi: 10.2957/kanzo.66.407]
39. Aida M, Ikegami T, Korenaga M. Methods of communicating negative results of hepatitis virus tests. Kanzo 2023;64(6):259-269. [doi: 10.2957/kanzo.64.259]
40. Žlahtič B, Kokol P, Blažun Vošner H, Završnik J. The role of correspondence analysis in medical research. Front Public Health 2024;12:1362699. [doi: 10.3389/fpubh.2024.1362699] [Medline: 38584915]
41. Greenacre M. Correspondence analysis in medical research. Stat Methods Med Res 1992;1(1):97-117. [doi: 10.1177/096228029200100106] [Medline: 1341654]
42. Jończyk JA, Olszewska AM. The use of correspondence analysis in assessing the antecedents of innovativeness in public hospitals. Stud Log Gramm Rhetor 2016 Dec 1;47(1):85-101. [doi: 10.1515/slgr-2016-0048]

XSL•FO
RenderX

43. Tsuboi S, Yoshida H, Ae R, Kojo T, Nakamura Y, Kitamura K. Selection bias of internet panel surveys: a comparison with a paper-based survey and national governmental statistics in Japan. Asia Pac J Public Health 2015 Mar;27(2):NP2390-NP2399. [doi: 10.1177/1010539512450610] [Medline: 22743862]
44. Szolnoki G, Hoffmann D. Online, face-to-face and telephone surveys—comparing different sampling methods in wine consumer research. Wine Econ Policy 2013 Dec;2(2):57-66. [doi: 10.1016/j.wep.2013.10.001]
45. Kagamimori S, Gaina A, Nasermoaddeli A. Socioeconomic status and health in the Japanese population. Soc Sci Med 2009 Jun;68(12):2152-2160. [doi: 10.1016/j.socscimed.2009.03.030] [Medline: 19375838]
46. Introducing GPT-5.2. OpenAI. URL: https://openai.com/index/introducing-gpt-5-2/ [accessed 2025-12-11]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
HBV: hepatitis B virus
HCV: hepatitis C virus
PR: prevalence ratio
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
WHO: World Health Organization

Edited by C Argyropoulos; submitted 15.Dec.2025; peer-reviewed by E Etafo, HS Maan, MZ Hossain; revised version received 23.Feb.2026; accepted 25.Feb.2026; published 02.Apr.2026.

Please cite as:
Sakai K, Nakazawa S, Furuya Y, Fukai K, Sano K, Korenaga M, Tatemichi M
Strategy for Hepatitis B and C Virus Testing Campaigns Through Web Services and Digital Advertising in Japan: Nationwide Cross-Sectional Study With Correspondence Analysis
J Med Internet Res 2026;28:e89585
URL: https://www.jmir.org/2026/1/e89585
doi:10.2196/89585

© Kosuke Sakai, Shoko Nakazawa, Yuko Furuya, Kota Fukai, Kei Sano, Masaaki Korenaga, Masayuki Tatemichi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Opposition to Youth e-Cigarette Prevention Campaigns on Twitter and TikTok: Cross-Platform Observational Mixed Methods Analysis

Chandler C Carter, MA; Simon Page, BA; Mateusz Borowiecki, MSc; Ganna Kostygina, PhD; Sherry L Emery, MBA, PhD; Miao Feng, PhD

Social Data Collaboratory, Public Health Department, NORC at the University of Chicago, 55 E Monroe St 30th Floor, Chicago, IL, United States

Corresponding Author:
Miao Feng, PhD
Social Data Collaboratory, Public Health Department, NORC at the University of Chicago, 55 E Monroe St 30th Floor, Chicago, IL, United States

Abstract

Background: Youth e-cigarette use rose sharply between 2013 and 2024 in the United States, prompting widespread prevention campaigns at national, state, and local levels. However, many campaigns encountered online opposition, sometimes leading to message distortion or campaign withdrawal. While previous studies have examined individual campaigns, little is known about how oppositional dynamics differ across social media platforms with distinct architectures.

Objective: This study aimed to conduct a retrospective, cross-platform surveillance study of oppositional responses to US youth e-cigarette prevention campaigns, comparing tactics, themes, and engagement on Twitter (now X) and TikTok to inform platform-specific public health strategies.

Methods: We collected Twitter (2014 - 2020) and TikTok (2020 - 2023) posts related to major US e-cigarette prevention campaigns using 15 campaign-specific hashtags and 4 verified prevention campaign handles. We included public, English-language posts from geographic regions allowed by the platforms. Machine learning classification and human coding were used to detect oppositional content, characterize narrative frames, and classify user types. Engagement was assessed using post-level metrics, including likes, shares, comments, and retweets. We analyzed message prevalence, engagement patterns, and oppositional themes.

Results: On Twitter, opposition comprised 26.8% (83,074/310,207) of campaign-related posts overall but dominated certain campaigns (eg, *Still Blowing Smoke*: 6052/6113, 99%). A small cluster of advocacy and commercial accounts generated 57.8% of opposition retweets. Dominant narratives included questioning the credibility of health authorities, claims that prevention advertisements backfired, vaping rights, and product promotion. In contrast, TikTok opposition constituted only 3.5% (108/3127, 95% CI 3.1%-3.9%) of posts and was characterized by humor (71/108, 65.7%), mockery (48/108, 44.4%), and ironic portrayals of vaping (30/108, 27.8%). Individual creators comprised 76.1% (153/201) of accounts sharing prevention posts, and opposition videos used the visibility-boosting hashtag #fyp significantly more than prevention posts (51.9% vs 32.2%; P<.001). Despite inconsistent hashtag use, prevention posts achieved higher average engagement than oppositional content.

Conclusions: This novel cross-platform, multicampaign analysis of opposition responses to e-cigarette prevention campaigns revealed how opposition reflects distinct platform architectures. Twitter opposition was highly coordinated and amplified by commercial and advocacy accounts, especially during regional campaigns. TikTok opposition was decentralized and humor-driven, aligning with the platform's entertainment-oriented algorithm. The findings strengthen health communication by introducing a framework for evaluating platform-specific vulnerabilities and informing evidence-based campaign design. On Twitter, effective countermeasures may require real-time monitoring of social media discourse to support rapid responses to coordinated opposition. On TikTok, leveraging creator partnerships and remix-friendly content may help public health messages compete with entertainment-dominated discourse. Consistent hashtag use can strengthen engagement by minimizing the fragmentation of content visibility, and credible health sources should increasingly reinforce prevention narratives on both platforms. Greater platform accountability and transparency are needed to ensure that prevention content is not systematically deprioritized by algorithms relative to commercial promotion.

(J Med Internet Res 2026;28:e83791) doi:10.2196/83791

KEYWORDS

health campaigns; social media; tobacco industry; public health; health communication; e-cigarettes

XSL•FO
RenderX

Introduction

Following a long decline in youth tobacco use, the emergence of e-cigarettes in the 2010s prompted a resurgence in tobacco use among adolescents in the United States [1]. By 2023, approximately 10% of US high school students reported current e-cigarette use, making these devices the most popular tobacco product among youth [1]. Adolescents are particularly susceptible to nicotine addiction and may face additional respiratory, cardiovascular, and immunological harms [1,2]. This rise in youth use sparked significant public health concern, including concern about the potential for e-cigarettes to act as a "gateway" to other tobacco products. This was a key driver for the creation of federal and local prevention campaigns analyzed in this study [3-5]. Although interventions such as flavor bans have sought to curb youth appeal [6,7], e-cigarettes remain pervasive [8], underscoring the importance of effective prevention campaigns.

A diverse network of federal, state, and nonprofit stakeholders delivers tobacco prevention messages via television, print, and digital platforms [9-12]. These multichannel campaigns have demonstrated effectiveness in shaping risk perceptions and reducing initiation among youth [9,13-18]. However, the rise of social media presents new challenges for maintaining message integrity and evaluating broader impacts [19,20]. Despite generally being banned across traditional media, both direct and indirect tobacco marketing thrives on social media [21,22], with emerging evidence that such promotion is associated with tobacco use and susceptibility [23,24]. Furthermore, interactions among user behaviors, algorithmic feeds, and peer-to-peer interactions, such as remixing or hashtag hijacking, can amplify or undermine prevention efforts in ways not observed in traditional media environments [25-28]. These dynamics require a framework that accounts not only for content but also for how visibility, engagement, and narrative control are structured.

Simultaneously, counter-narratives thrive on social media [22,29,30]. Tactics such as "Twitter bombing" (ie, mass posting to overwhelm hashtags) or astroturfing (coordinated messaging disguised as organic support) allow commercial or advocacy interests to flood platforms with counter-narratives, creating misleading perceptions of widespread support or dissent [31-35]. Theory-based frameworks suggest that social media may allow commercial or advocacy actors to "set the agenda" in ways once restricted to institutional actors and traditional media. These frameworks also suggest that the structure of social media enables novel forms of gatekeeping that can amplify pro-vaping narratives, potentially increasing youth uptake while subverting prevention efforts [36-39]. Platform affordances shape these strategies: text-based environments such as Twitter (now X) facilitate rapid coordination and volume, while TikTok's short-video format enables the spread of provaping content through easily digestible entertainment [28,40-42].

Drawing on theories of networked gatekeeping and principal-agent dynamics [37,38,43], we examine how digital platform architectures mediate the amplification of health messages versus counter-narratives. Networked gatekeeping theory highlights how control over content visibility is distributed across digital actors rather than centralized within traditional editorial institutions [37,43]. Principal-agent dynamics describe how misaligned incentives between message originators, such as public health agencies, and intermediaries, such as platforms or content creators, can shape how messages are disseminated and transformed [38]. In doing so, we explore how platform affordances shape engagement between public health campaigns and their challengers. Although previous research has examined antitobacco messaging or oppositional discourse on individual platforms, particularly Twitter and YouTube, little is known about how opposition dynamics differ across platforms with distinct sociotechnical architectures. Existing studies typically focus on single-platform case studies or specific campaigns [32], often without directly comparing how algorithmic design, user behaviors, and content norms shape oppositional engagement. This gap limits our understanding of how campaign messaging is received, reframed, or resisted in varied digital environments [31,32]. Our study is the first to provide cross-platform surveillance of opposition to US youth e-cigarette prevention campaigns. Specifically, we compare Twitter's historical network dynamics (ie, retweet-based amplification by advocacy clusters) with TikTok's contemporary algorithmic affordances (ie, individualized For You Page and entertainment-driven content formats). In our study, "opposition" is operationalized as content within specific campaign discursive spaces that diverges from or actively challenges official public health messaging. This approach examines patterns that may reflect platform-specific structural impacts of these divergent narratives, acknowledging the inherent complexity of online discourse without implying a simplistic binary. By analyzing both oppositional structures and engagement characteristics, we aim to illuminate platform-specific vulnerabilities and outline adaptive strategies for strengthening public health communication in increasingly participatory, algorithmically governed media ecosystems. We pursue three core aims: (1) to assess the volume and engagement of prevention campaign–related social media posts, (2) to identify the sources and tactics of oppositional content, and (3) to characterize platform-specific opposition dynamics to inform future strategy development.

Methods

Study Design and Rationale for Mixed Methods Approach

We conducted a US-based cross-platform surveillance study examining opposition to e-cigarette prevention campaigns on Twitter (2014 - 2020) and TikTok (2020 - 2023). We use the name "Twitter" throughout this study as our data collection period for this platform concluded in 2020, before its 2022 acquisition and rebranding as X. We analyzed posts related to campaigns at multiple levels of implementation, including national (eg, the Food and Drug Administration's *The Real Cost*), state (eg, California's *Still Blowing Smoke*), and local (eg, Chicago's *Vaping Truth*). *The Real Cost* campaign initially launched in 2014 to curb cigarette use among youth, but by 2018, the campaign pivoted to address e-cigarette use [44].

XSL•FO
RenderX

This study used a convergent mixed methods design to integrate machine learning classification (quantitative) with thematic analysis (qualitative) to triangulate opposition patterns across platforms. Quantitative and qualitative findings were integrated at the interpretation stage by jointly examining how patterns in volume, engagement, and user type corresponded to emergent opposition themes across platforms. We selected this approach because quantitative classifiers alone cannot adequately capture multimodal content, while qualitative coding alone cannot scale across millions of posts.

Due to nonconcurrent time frames, we focused on identifying structural differences in oppositional engagement rather than conducting comparisons of post volume for cross-platform campaigns. Twitter hashtags were identified from campaign websites and through exploratory searches of social media content. Hashtags identified during Twitter data collection that also appeared on TikTok were used to collect data on TikTok, with TikTok data supplemented by searches for platform-specific hashtags (Table S1 in Multimedia Appendix 1).

This study is reported in accordance with the *American Psychological Association* Journal Article Reporting Standards for Mixed Methods Research (JARS–Mixed Methods) [45].

Analytic Framework

Overview

Twitter and TikTok require distinct analytic approaches due to the text-based nature of the former and the multimodal nature (eg, video, audio, and text) of the latter. Machine learning techniques were only applied to Twitter, as text-based content is well suited for natural language processing techniques. In contrast, TikTok's multimodal format presents challenges for automated classification, necessitating manual content and thematic coding to ensure accurate interpretation.

To systematically evaluate social media discourse related to e-cigarette prevention campaigns, we applied a 4-component framework assessing (1) message volume, (2) message content, (3) message sources, and (4) engagement metrics [40].

Message Volume

The number of tweets (Twitter) and videos (TikTok) was measured over time and by campaign. We used linear regression to assess changes in campaign-related post volume over time on each platform (from January 2020 to December 2023 on TikTok and from February 2016 to October 2020 on Twitter).

Message Content

To capture the advocacy stance (opposition vs prevention or neutral) of tweets, we used a supervised machine learning classifier. Opposition themes were identified using thematic analysis to explore the underlying opposition narratives. For TikTok, content analysis was used to systematically classify the advocacy stance of videos (opposition vs prevention vs neutral) and to identify high-level opposition themes. To further explore nuanced rhetorical strategies and narrative patterns within opposition posts, we conducted a thematic analysis.

Message Sources

For Twitter, user type was determined using a machine learning approach, while TikTok user classification was achieved through content analysis.

Engagement Metrics

For Twitter, engagement was operationalized as the number of retweets each post received. For TikTok, "interactions" comprised the sum of likes, comments, and shares per video, while views were analyzed separately as a measure of reach. TikTok-specific visibility-boosting features were also analyzed, including the proportion of videos using the hashtag #fyp as well as interactive video formats: duets (playing videos side by side from different users) and stitches (embedding clips from existing videos). Given these platform differences, engagement metrics were not compared directly across platforms. The validity of content analyses was assessed using Gwet's AC1 statistic. The performance of the Twitter advocacy stance classification model was evaluated using F_1-scores.

Data Collection and Analysis

Twitter

We collected Twitter data using the now-deprecated Historical PowerTrack API, which provided enterprise-level access to the complete archive of public tweets matching defined rules on a monthly basis from January 2018 through April 2023. Our dataset includes all public English-language tweets (N=908,010) from February 2014 to October 2020 that matched 15 campaign-related hashtags associated with 14 distinct e-cigarette prevention campaigns, along with 4 verified campaign-affiliated accounts (Table S1 in Multimedia Appendix 1). Although the *Fresh Empire* campaign is focused on tobacco prevention, discussions of e-cigarettes appeared in posts using the campaign's hashtag. Additionally, experimental evidence indicates that tobacco prevention messaging can increase perceived effectiveness for discouraging e-cigarette use, suggesting spillover effects across products [46,47].

Opposition content was identified using a supervised machine learning classifier trained on 2282 human-labeled tweets. Coders were instructed to label tweets as "oppositional" only when they expressed support for vaping, rejected the goals of prevention campaigns, or aligned with provaping advocacy. They also considered full textual context, including tone and source, to avoid misclassifying neutral or critical mentions of hashtags. Posts were classified as oppositional if they met any of the following criteria: (1) use of hashtags explicitly rejecting the prevention efforts (eg, #notblowingsmoke); (2) use of hashtags promoting vaping (eg, #vapingsavedmylife and #vapelife); (3) reference to vaping-related advocacy (eg, #wevapewevote); or (4) mention of organizations known for supporting vaping or harm reduction, such as the Consumer Advocates for Smoke-Free Alternatives Association. The classifier achieved a macro-averaged F_1-score of 0.90 using 10-fold cross-validation, with a precision of 0.88 and recall of 0.92. An error analysis indicated that most misclassifications (false positives) occurred in cases where tweets mentioned health-related topics or campaign hashtags without clear oppositional framing.

XSL•FO
RenderX

TikTok

Data from 2020 to 2023 were collected in March 2025 in 2 phases using the TikTok Research API [48], an official platform-provided interface that allows approved researchers to access public content from designated geographic regions for research purposes. Our study focused on public content and restricted analysis to English-language posts, identified through metadata and manual review.

Phase 1 used 18 official hashtags and yielded 2563 posts. Phase 2 focused on 2 prominent national prevention campaign hashtags that were prominent in 2020: #ImmuneUpVapeDown and #ThisisQuitting, selected due to their prevalence and common spelling variations (eg, #thisisquitting was often misspelled as #thisisquiting; and #ImmuneUpVapesDown was also widely disseminated as #ImmuneUpVapeDown, a nonplural version of "Vape"). For *Immune Up*, both hashtag variations appeared in official promotional materials [49]. These 2 hashtags collectively generated over 2.1 million TikTok posts, from which we extracted as much content as application programming interface constraints allowed [50,51]. Hashtag queries were constructed using Boolean operators to include relevant variations and exclude unrelated campaigns.

To ensure our coded sample captured the full spectrum of content, from highly viral to more typical posts, we used quantile-stratified sampling based on total engagement (likes, comments, and shares combined). As engagement distributions on social media are heavily skewed, a simple random sample would risk underrepresenting the small number of high-impact posts that disproportionately shape discourse. We therefore ranked all 15,000 collected posts by engagement and divided them into 3 strata: the top 20% (high engagement), the middle 40% (moderate engagement), and the bottom 40% (low engagement). We then randomly sampled a proportionate number of posts from each stratum to generate our final coded sample of 800 posts. This approach ensured representation across visibility tiers and enhanced thematic diversity. All sampled posts were reviewed to exclude duplicates and nonsubstantive content. Across both phases, we collected 19,208 unique TikTok posts matching relevant prevention campaign hashtags. Of these, 15,000 (78.1%) posts with complete engagement metadata were eligible for ranking and quantile-stratified sampling. From this ranked set, we drew a proportionate stratified sample of 800 (5.3%) posts for manual coding. All posts from phase 1 of the data collection were manually labeled, resulting in a total of 3363 labeled posts. At the time of final analysis, 3127 (93.0%) labeled posts remained viewable and analyzable, with the remainder removed due to deletion and privacy restrictions. These removals accounted for approximately 7% of labeled TikTok posts and were treated as missing completely at random; no imputation was performed because missingness reflected post-level availability rather than content characteristics. A flow diagram summarizing the TikTok data collection and sampling workflow is provided in Figure S1 in Multimedia Appendix 1.

Content and Thematic Coding

Twitter

To develop opposition content themes, we used a thematic analysis approach. Two coders reviewed the top (highest tweet volume) 1000 oppositional messages from the most active users. We focused on the most active users, as their high volume of content is more likely to shape the overall narrative [52]. The coders then collaboratively identified themes, which were refined and grouped together until we arrived at four principal themes: (1) criticism of and mistrust in public health institutions (eg, claims of health authorities spreading inaccurate information and critiques of regulations), (2) claims of a boomerang effect of prevention advertisements (eg, advertisements triggering a desire to use e-cigarettes), (3) vaping rights and advocacy (eg, movements such as #ivapeivote), and (4) product promotion (eg, deals and coupons).

TikTok

For TikTok's short-form, video-first format, 5 trained coders developed and applied a structured codebook through an initial inductive coding phase, in which codes were generated directly from the data, followed by the integration of deductive codes informed by prior literature. The codebook included (1) user-level categories: account type (individual, organization, community [eg, fan accounts or news aggregators], and other), sector (health and government), and advocacy stance (protobacco or anticampaign and antitobacco or procampaign); and (2) video-level features: substance type (e-cigarette, other tobacco product, and other substance), advocacy stance, public service announcement framing (eg, conveying facts about substance harms or providing cessation resources), quit narratives (personal accounts of quitting a substance), brand mentions, product promotions, and entertainment or humor (overall; eg, skits, lip-syncing, jokes, and memes). Intercoder reliability was strong across all dimensions (Gwet AC1>0.81). To uncover more nuanced opposition narratives, we used the same thematic analysis approach applied to Twitter. Two coders jointly reviewed opposition posts (n=108), refining themes through iterative discussion, and identified six principal themes: (1) harm reduction or health claims (eg, mentions of any e-cigarette or tobacco-related health outcomes or comparative risk of tobacco smoking vs vaping), (2) personal testimony about vaping or tobacco benefits, (3) mockery of prevention campaigns (eg, using parody or satire), (4) dissatisfaction with advertisements, (5) ironic or hyperbolic references to vaping (eg, to comment on vaping culture or one's own vaping behaviors, including decisions to ignore prevention information), and (6) claims of a boomerang effect of prevention advertisements. The harm reduction or health claims were each evaluated for accuracy based on information from the Food and Drug Administration's Tobacco Education Resource Library [53]. Thematic saturation was reached when no new themes emerged during the final coding iterations.

Twitter User Classification

Twitter user classification followed a 2-step process. First, 2 coders labeled a random sample of 300 accounts (with up to 10 tweets per account) into 4 mutually exclusive categories that

XSL•FO
RenderX

were determined using a grounded theory approach: health (health-related organizations or individuals, including researchers and health care providers), commercial (organizations or brands explicitly selling or promoting tobacco products), provape advocacy (organizations or individuals advocating for pro–e-cigarette or tobacco policies or agendas), and organic (individuals with identity cues who are not public figures or influencers). Agreement was strong for the first 3 categories (AC1>0.80) and moderate for organic (AC1=0.72). Next, we used GPT-4o, accessed via an OpenAI instance on Microsoft Azure, to classify the same 300 accounts using structured prompts (Multimedia Appendix 2) and compared manual labels with GPT [54]. Accounts unrelated to the abovementioned categories or lacking sufficient information were classified as "other." Compared to human labels, GPT-4o achieved F_1-scores of 0.91 for organic and health accounts, 0.89 for advocacy accounts, and 0.96 for commercial accounts. On the basis of this validation, we applied the same prompting strategy to classify all unique users in the analytic data, using a unique user identifier available in the Twitter data. However, as Twitter display names were used as a feature for GPT's user classification and are nonunique (eg, users can change their display names), we opted to classify unique users in the Twitter data, sorted by highest follower count: this ensured we classified unique users using the most popular observation per unique identifier, where popularity is determined by the follower count. Final classifications were cleaned and merged with post-level data to analyze account type by post stance (opposition vs prevention or neutral). These classifications were used to contextualize quantitative patterns of opposition volume and engagement and to inform the interpretation of qualitative opposition themes across platforms. All codebooks used for data labeling are available upon request.

Ethical Considerations

All research activities for this study were approved by NORC at the University of Chicago's Institutional Review Board (IRB 23061378) and determined to meet the criteria for exemption under Exempt Category 4, as defined in US federal regulations (45 CFR 46.104) for using existing, publicly available social media data. The social media data were collected in accordance with each platform's terms of service and without direct interaction with users; therefore, informed consent was not required by the institutional review board. Privacy and confidentiality were safeguarded by ensuring that all data used in the analysis were publicly available and that all reported data were deidentified. No identifiable information or images of individual users are included in the manuscript or supplementary materials. Finally, no compensation was provided, as the study involved no direct interaction with or intervention upon human subjects.

Results

Campaign-Related Post Volume and Opposition Trends

We collected a total of 908,010 tweets referencing 14 e-cigarette prevention campaigns or mentioning the organization that produced the campaign from 2014 to 2020. After filtering for tweets explicitly referencing campaign names or hashtags, 310,207 (34.2%) tweets were retained for analysis. On TikTok (2020-2023), we analyzed 3127 viewable posts after excluding deleted, private, or non-English or non–US-based content.

Twitter campaign activity peaked during major campaign launches: March 2014 (n=9742) and August 2014 (n=10,867), corresponding with the rollout of *The Real Cost*, and January 2016 (n=14,989), coinciding with the launch of *Vaping Truth* (Figure 1). A peak in June 2017 (n=11,059) aligned with a *Fresh Empire*–sponsored event. From February 2016 to the end of the study period (October 2020), tweet volume declined significantly (β=–122 tweets per month, 95% CI –144.08 to –99.17; *P*<.001). On TikTok, campaign-related content exhibited smaller peaks, with the largest in October 2020 (n=995) during the launch of *Immune Up Vapes Down*. Post volume declined modestly but significantly from January 2020 to December 2023 (β=–5 posts per month, 95% CI: –7.91 to –1.02; *P*=.01).

Figure 1. Campaign-related tweet counts over time, overall and by the 5 campaigns with the highest tweet count (2014-2020). This figure illustrates the overall volume of discourse and key peaks corresponding to campaign launches.Table 1

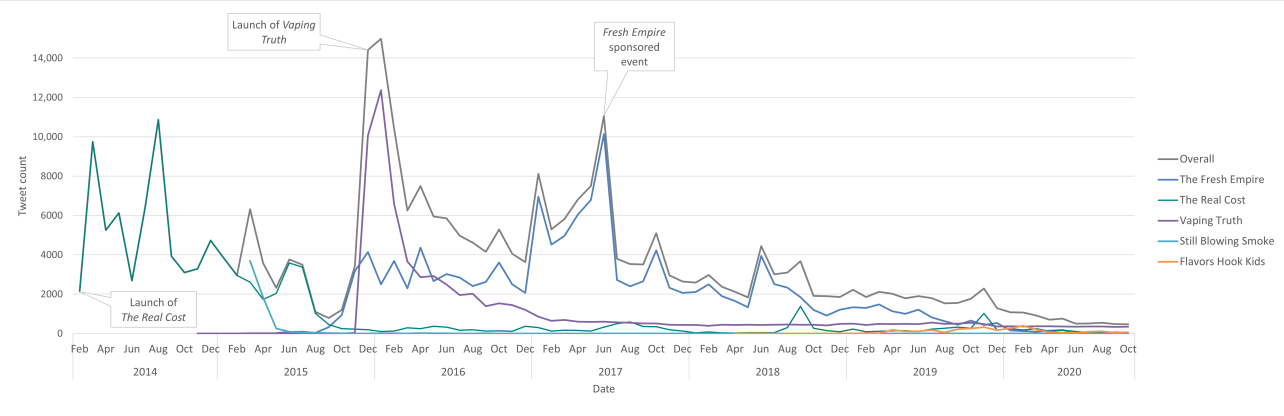

On Twitter, opposition content comprised 26.8% (83,074/310,207) of campaign-related posts, peaking in March 2015 (3766/6314, 59.7%) during the *Still Blowing Smoke* campaign and in January 2016 (11,821/14,989, 78.9%) during the *Vaping Truth* campaign. On TikTok, opposition was lower overall (108/3127, 3.5%) but rose modestly from 1.3% (27/2088) in 2020 to 5.6% (23/408) in 2023, peaking at 9.9% (28/282) of campaign-tagged posts in 2021, during *The Real Cost* campaign.

XSL•FO
RenderX

Opposition intensity varied by campaign and platform. On Twitter, among campaigns with more than 30 posts, state and local campaigns faced the most sustained counter-advocacy. Opposition comprised 99% (6052/6113) of content related to *Still Blowing Smoke*, 97.3% (70,300/72,270) of *Vaping Truth*–related content, and 27.3% (35/128) of messages about *The Vape Talk*. In contrast, national campaigns faced lower opposition: 3.4% (3125/93,246) for *The Real Cost* and 2.3% (3079/134,343) for *Fresh Empire* (Table 1). On TikTok, among campaigns with more than 30 posts, the highest opposition proportions were found in *The Real Cost* (56/246, 22.8%), *Vapes Down* (22/178, 12.4%), and *Ditch Juul* (4/42, 9.5%). Nine campaigns had limited TikTok presence (<30 posts).

Table . Proportion of campaign-related tweets classified as opposition by platform, overall and by campaign hashtag[a].

Platforms and Campaigns	Opposition, n (%)	Total, n (%)
Twitter		
Total	83,074 (26.8)	310,207 (100)
StillBlowingSmoke	6052 (99)	6113 (2)
VapingTruth	70,300 (97.3)	72,270 (23.3)
TheVapeTalk	35 (27.3)	128 (0.0)
VapingEquals	18 (15.5)	116 (0.0)
FlavorsHookKids	383 (12.9)	2974 (1)
TheDirtyTruth	22 (10.5)	209 (0.1)
IJustDidn’tKnow	15 (9.7)	155 (0.0)
VapeFreeNovember	42 (7.1)	592 (0.2)
FactsOverFlavor	2 (6.4)	31 (0.0)
TheRealCost	3125 (3.4)	93,246 (30.1)
FreshEmpire	3079 (2.3)	134,343 (43.3)
TikTok		
Total	108 (3.5)	3127 (100)
TheRealCost	56 (22.8)	246 (7.9)
VapesDown	22 (12.4)	178 (5.7)
DitchJuul	4 (9.5)	42 (1.3)
DotheVapeTalk	2 (6.1)	33 (1.1)
EscapeTheVape	2 (5)	40 (1.3)
DitchVape	7 (3)	230 (7.4)
ImmuneUpVapesDown	11 (0.9)	1290 (41.3)
ThisIsQuitting	5 (0.5)	1011 (32.3)

[a]Total Twitter figures include all tweets collected using 15 prevention campaign-related hashtags associated with 14 distinct campaigns. Total TikTok figures reflect the subset of labeled posts with viewable videos at the time of analysis. Only campaigns with >30 posts are shown individually; others are included in the total.

Engagement Patterns and Visibility-Boosting Features

On Twitter, retweets accounted for 59.7% (185,315/310,207) of campaign posts. Prevention tweets were retweeted 3.6 times more often than opposition posts. The most retweeted campaigns were *Still Blowing Smoke* (4977/6113, 81.4%) and *The Real Cost* (66,060/93,246, 70.8%). Opposition users also posted more frequently than prevention or neutral users (9.6 vs 1.6 posts per user). Engagement was highly concentrated: the top 10% of opposition accounts generated 84.4% (34,080/40,378) of opposition retweets, and the top 1% accounted for 63.9% (25,782/40,378).

On TikTok, opposition posts accounted for a smaller proportion of total posts (108/3127, 3.5%), compared to prevention posts (410/3127, 13.5%). Opposition posts also received lower average engagement (3103 vs 12,008 interactions (likes, comments, and shares). The median engagement was 40.0 (IQR 13.0 - 209.8) for prevention posts and 44.5 (IQR 16.8 - 128.0) for opposition posts. TikTok interactions (likes, comments, and shares) were driven by a small number of high-impact users: 97.9% (328,128/335,074) of opposition interactions came from just 11 accounts (the top 10%).

Opposition content used visibility-boosting platform features more frequently. The hashtag #fyp (“For Your Page”) appeared in 51.9% (56/108) of opposition posts, compared to 32.2% (132/410) of prevention posts. Use of interactive video formats, such as duets and stitches, was similar across conditions

XSL•FO
RenderX

(opposition posts: 6/108, 5.6% vs prevention posts: 20/410, 4.9%).

Opposition Themes

Prominent themes in opposition messaging varied by platform. On Twitter, key opposition themes included (1) criticism of and mistrust in public health institutions, (2) claims of a boomerang effect of prevention advertisements, (3) vaping rights and advocacy, and (4) product promotion. Many high-volume users aligned with advocacy movements such as #ivapeivote or #ABillionLives, framing vaping as a personal liberty or health-related right.

On TikTok, opposition content was primarily entertainment-oriented. The most frequently observed themes included general humor or entertainment (71/108, 65.7%), mockery of e-cigarette prevention campaigns (48/108, 44.4%), and ironic or hyperbolic references to vaping (30/108, 27.8%). One-quarter of opposition posts (27/108, 25%) visually depicted an e-cigarette. Expressions of dissatisfaction with prevention advertisements were present in 15.7% (17/108) of posts. Less frequently coded themes included claims of a boomerang effect of prevention advertisements (n=4, 3.7%), product promotion (n=2, 1.9%), personal testimony about the benefits of tobacco or vaping (n=2, 1.9%), and harm reduction or health claims (n=2, 1.9%). Notably, all health-related claims (n=3, 2.8%) contradicted established scientific consensus (eg, people with asthma are typically found to be healthier when they vape) [53].

Content Creator and Source Analysis

On Twitter, among users with nonmissing profile descriptions, organic users constituted the majority (42,446/49,809, 85.2%), followed by other (6850/49,809, 13.8%), health (2264/49,809, 4.5%), provape advocate (1570/49,809, 3.2%), and commercial accounts (951/49,809, 1.9%). However, among opposition posters, provape advocacy (1376/5465, 25.2%) and commercial (724/5465, 13.2%) accounts were overrepresented, and organic users (2926/5465, 53.5%) were underrepresented compared to overall figures.

Among accounts posting prevention or neutral content, representation by health-related sources was 4.8% (2139/44,344).

On TikTok, individuals made up 91.8% (2023/2204) of labeled users, community accounts made up 3.1% (69/2204), organizations made up 2.8% (61/2204), and other types of accounts made up 2.3% (51/2204). A small proportion (29/2204, 1.3%) of users were categorized as health- or government-affiliated accounts.

Among opposition posters, individuals were the most prevalent (95/103, 92.2%), followed by community accounts (4/103, 3.9%), organizations (2/103, 1.9%), and other types of accounts (2/103, 1.9%). There was no representation by health or government sectors. Provape advocacy represented 6.8% (7/103) of opposition accounts.

Among accounts posting prevention messages, 76.1% (153/201) were individuals, 14.4% (29/201) were organizations, 5% (10/201) were community accounts, and 4.5% (9/201) were some other type of account. Health-related accounts constituted 10% (20/201) of these accounts, while 0.5% (1/201) were government related.

Hashtag Variation and Engagement

Campaign engagement was shaped by both deliberate co-optation strategies and hashtag variation. Most prominently, opposition to *Still Blowing Smoke* on Twitter was marked by extensive use of the counter-hashtag #notblowingsmoke [32], which highlighted the claim that e-cigarette vapor is less dangerous than cigarette smoke and appeared alongside the official campaign hashtag #stillblowingsmoke in 53.7% (3248/6052) of the campaign's opposition tweets.

On TikTok, engagement patterns were affected by the concurrent use of multiple similar hashtags associated with the same campaign. Across the study period, #ImmuneUpVapeDown was tagged in 3552 posts, while #ImmuneUpVapesDown appeared in 1184 posts. Although #ImmuneUpVapesDown appeared less frequently, posts using this hashtag generated higher average engagement, with 11,271 views and 1380 engagements per post, compared to 7831 views and 1105 engagements per post for #ImmuneUpVapeDown. This reflects a 1.4-fold increase in views and a 1.3-fold increase in engagements for #ImmuneUpVapesDown.

A similar pattern was observed for the *This is Quitting* campaign. The official hashtag, #thisisquitting, appeared in 13,114 posts, whereas the common misspelling, #thisisquiting, appeared in 434 posts. Despite its lower volume, posts with the misspelled hashtag generated slightly higher average engagement (1653 vs 1536 interactions per post). However, these posts received lower average viewership compared to posts with the official hashtag (10,877 vs 14,163 views per post).

Discussion

Principal Findings

Our study offers one of the first cross-platform examinations of digital opposition to youth e-cigarette use prevention campaigns, identifying distinct structures, engagement patterns, and thematic frames across Twitter and TikTok. We found that opposition was more prevalent and coordinated on Twitter, driven by commercial and advocacy accounts. In contrast, TikTok opposition was decentralized and humor-driven. Despite these differences, prevention content generally achieved higher engagement than oppositional posts on both platforms. These findings highlight the extent to which oppositional activity is shaped not only by campaign content but also by the unique affordances, user cultures, and algorithmic logics of each platform. Understanding these dynamics is essential. Responsive and resilient prevention strategies must be tailored to the evolving digital media landscape.

Opposition took divergent forms across platforms, reflecting structural and cultural asymmetries. On Twitter, opposition was highly coordinated and driven by commercial and advocacy accounts that leveraged hashtag hijacking and volume-based amplification. These patterns reflect classic principal-agent dynamics and networked gatekeeping. In both cases, vested interests influence which narratives gain visibility and shape

XSL•FO
RenderX

policy discourse at scale. Organized advocacy networks and commercial promotion directly contested health communication and regulatory narratives. Their messages reflected direct challenges to health institutions, calls for deregulation, claims of a boomerang effect, harm reduction narratives, and product promotion.

In contrast, TikTok opposition was decentralized and creator-led, with individual users relying on algorithmically favored formats, such as humor, parody, and remix. Rather than mounting explicit policy critiques, TikTok users often reframed prevention campaigns as objects of satire or trivialization through "duets" that mock campaign messages, lip-syncing to songs about e-cigarette use, or comedic skits. Twitter's text-centric environment and low production barriers facilitate more textual coordination, while TikTok's video-first interface favors performative and affective engagement [55-57].

Prior research [32-35] has documented Twitter "bombing" or astroturfing and branding appropriation; our findings suggest that TikTok's entertainment-driven opposition presents a structurally distinct challenge to campaign coherence. Importantly, public health institutional accounts comprised only 10% of all prevention-aligned TikTok content, indicating that the discourse is overwhelmingly dominated by individual creators. While this does not speak directly to impact or effectiveness, it underscores the relatively modest presence of institutional voices within platform-native discourse. The prevalence of mockery on TikTok, with the opposition's significantly higher use of visibility-boosting hashtags like #fyp, aligns with an algorithmic logic that privileges engagement over informational accuracy. This makes evidence-based public health content less salient unless adapted to the platform's affective and performative norms. The presence of both coordinated and decentralized opposition across platforms suggests that prevention campaigns face 2 qualitatively distinct credibility challenges: organized counter-messaging that seeks to influence policy and diffuse creator-driven content that may erode credibility through parody and cultural reframing.

Engagement dynamics further underscore the differences in the nature of opposition between platforms. On Twitter, opposition users posted or retweeted repeatedly, suggesting coordinated persistence. On TikTok, opposition was more episodic and reliant on the virality of individual posts. Although opposition content more frequently used visibility-boosting features such as #fyp, it did not consistently outperform prevention content in engagement, indicating that tagging alone does not guarantee amplification. These dynamics reflect how each platform's algorithmic logic privileges particular forms of participation. Twitter rewards repeated real-time posting, trending topics, and follower networks, while TikTok prioritizes personalized content that aligns with user engagement patterns and preferred entertainment formats [55,56].

This pattern of engagement is further illustrated by the concentration of opposition interactions among a small number of high-impact TikTok users. Most opposition interactions (likes, comments, and shares) came from just 11 accounts representing the top 10% of opposition users. This aligns with TikTok's influencer-driven model, where a small number of content creators can disproportionately shape discourse and visibility. For prevention campaigns, this suggests a strategic opportunity to increase engagement. This can be achieved by forming partnerships with credible, platform-native, youth-aligned influencers who can compete with oppositional voices in reach rather than relying on broad content distribution.

Notably, opposition visibility did not uniformly suppress engagement with prevention content, particularly on TikTok, suggesting that volume alone does not determine engagement and impact. This underscores the need for platform-specific intervention approaches. On Twitter, effective approaches should focus on detecting coordinated opposition and preemptively framing policy discourse to counter specific antiregulatory narratives. On TikTok, campaigns should invest in creative, platform-native formats, such as duets, remixes, or trending audio strategies, to maximize resonance with youth audiences while expanding public health visibility within the algorithmic ecosystem.

Future research should examine how competing narratives influence youth perceptions of e-cigarette use risk, assess the impact of unsupported and unauthorized health claims on public understanding, and explore regulatory approaches to moderating coordinated opposition. Further work is needed to examine how evolving algorithmic infrastructures govern the diffusion and visibility of public health content, including how oppositional discourse adapts in increasingly sophisticated ways. In addition, campaign-by-campaign thematic analyses could help disentangle the influence of specific prevention messages from broader platform-level dynamics and determine whether certain prevention themes predict opposition. These future efforts would also benefit from deeper integration of theory, particularly frameworks such as networked gatekeeping to better explain the structural vulnerabilities of public health messaging online.

Across both platforms, disruptive hashtag dynamics emerged in two forms: (1) strategic co-optation or hijacking (eg, #notblowingsmoke) and (2) hashtag fragmentation, which included both user-generated "drift" (ie, unintentional spelling variations such as #thisisquiting) and inconsistencies introduced by campaigns themselves (eg, simultaneous promotion of 2 similar but nonidentical hashtags for a single campaign). While hashtag drift can sometimes reflect authentic engagement and meme-like spread, inconsistent use of multiple official hashtags introduces risks. It can confuse the audience, dilute message visibility, and complicate evaluation. To address this, campaigns should promote a single, memorable, and unique hashtag consistently across all materials and partner networks. This helps minimize the hashtag-level fragmentation of campaign content visibility reflected in our findings and strengthens the integrity of message tracking.

Real-time monitoring can help detect and address hijacking and hashtag drift. Campaign staff should set up real-time alerts using social listening tools for sudden spikes in volume or shifts in sentiment. Such monitoring could signal a coordinated hijacking effort like the one seen with some counter-campaigns observed in regional campaigns. Quantitative monitoring should be paired with daily qualitative reviews of the content being posted under the hashtag to understand how the narrative is evolving.

XSL•FO
RenderX

Finally, prevention campaigns need preplanned rapid response protocols to counter oppositional messages. These should extend beyond reactive posts. There should be a preplanned communications protocol that may include a playbook of preapproved messages, coordinated content amplification, and activating credible messengers. Such practices are especially important in environments such as Twitter, where opposition is coordinated and can quickly flood the e-cigarette prevention discourse. It is also crucial in environments where algorithmic feeds (eg, TikTok's For You Page) may amplify small differences in framing, format, or metadata, thereby influencing content visibility and diffusion in unpredictable ways.

Limitations

This study has several limitations. First, platform-level constraints, particularly inconsistencies in the TikTok Research API, limited the volume of analyzable content, especially for high-volume campaigns. In addition, it remains unclear whether promoted or sponsored advertisement content is consistently included via the application programming interface; if excluded, our analysis may underrepresent campaign materials delivered through paid placements. Second, while observed differences in discourse across platforms could be attributed to structural or programmatic features of the platforms, they could also reflect variation in data collection periods or user demographics. As Twitter and TikTok data were collected from nonoverlapping periods (2014 - 2020 and 2020 - 2023, respectively), our goal was not to directly compare levels of opposition or engagement across time. Rather, each dataset reflects the dominant usage period of the platform for prevention campaigns, and our analysis focuses on platform-specific opposition dynamics within those distinct temporal contexts. Third, while engagement metrics provide useful indicators of content visibility and performance, they do not capture user intent, sentiment, or behavioral outcomes. Therefore, high engagement should not be equated with either campaign impact or message suppression. Additionally, Twitter user classification relied on mutable display names rather than persistent identifiers, which constrained our ability to characterize user behavior at scale. Finally, our analyses do not take into account secular trends in platform activity, such as the long-run decline in Twitter activity between 2014 and 2019. As our analysis is primarily concerned with proportions and themes of content, this limitation does not substantively impact our results [58].

Conclusions

This study systematically compared oppositional dynamics across e-cigarette prevention campaigns on 2 major social media platforms using a mixed methods approach. Unlike previous studies that focused on singular platforms or campaigns, this study compared several e-cigarette prevention campaigns on Twitter and TikTok to reveal how distinct sociotechnical architectures shape opposition strategies. Campaign resilience in digital environments depends on more than message design alone. Our findings offer actionable guidance to strengthen digital health communication across platforms. Public health institutions must adapt to platform-specific architectures. On Twitter, strategies to counter opposition content should include real-time surveillance to enable rapid response to coordinated opposition. On TikTok, public health entities may improve engagement by investing in creator partnerships and using visibility-enhancing content formats. Sustained investment in creator collaboration is especially critical to counter entertainment-driven opposition narratives with content forms already familiar and appealing to youth audiences. Across both platforms, campaigns should ensure consistent branding to prevent potential audience fragmentation and message dilution from hashtag variations. Additionally, given the relatively small presence of credible health sources in the prevention discourse, they should be used more frequently to reinforce prevention narratives. On Twitter, credible health voices are essential for combating coordinated advocacy networks, while on TikTok, they should complement influencer partnerships and entertainment-oriented formats.

Meanwhile, regulatory agencies and health campaigns should also advocate platform accountability, including algorithmic transparency. This entails, at minimum, requiring platforms to disclose whether prevention content is systematically downranked compared to commercial promotion and to provide greater elevation of verified public health sources.

Acknowledgments

We would like to acknowledge Alex Kresovich and Haoyu Shi for their contributions to this research.

GPT-4o was used for Twitter user classification, as described in the *Methods* section. All outputs were reviewed and verified by the authors. Grammarly was used for spelling and grammar assistance; no generative artificial intelligence tools were used to draft substantive content or interpret results.

Funding

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health (grant R01CA283038). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The social media data that support the findings of this study are not publicly available due to third-party restrictions and platform terms of service. Data were obtained via the X (formerly Twitter) API and TikTok Research API for approved research use.

XSL•FO
RenderX

Access to these data can be requested from the respective platforms and is subject to their individual data-sharing policies and researcher eligibility criteria.

Authors' Contributions

CCC, SP, GK, SLE, and MF together designed and planned the study. CCC, SP, and MF conducted data analysis. CCC, SP, MB, and MF contributed to data interpretation. CCC, SP, MB, and MF wrote the first draft. CCC, SP, MB, GK, SLE, and MF revised the draft. The final version of the paper has been reviewed and approved by all authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
TikTok data collection and sampling workflow and overview of e-cigarette social media campaigns by hashtag, account handle, region, platform, and post count.
[PDF File, 149 KB - jmir_v28i1e83791_app1.pdf]

Multimedia Appendix 2
ChatGPT prompt for Twitter user classification.
[PDF File, 103 KB - jmir_v28i1e83791_app2.pdf]

References

1. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. E-Cigarette Use Among Youth and Young Adults: A Report of the Surgeon General: Centers for Disease Control and Prevention; 2016. [Medline: 30869850]
2. McGrath-Morrow SA, Gorzkowski J, Groner JA, et al. The effects of nicotine on development. Pediatrics 2020 Mar;145(3):e20191346. [doi: 10.1542/peds.2019-1346] [Medline: 32047098]
3. Visser W, Klerx W, Geraets L, et al. The health risks of using e-cigarettes. : RIVM, Netherlands National Institute for Public Health and the Environment; 2015 URL: https://www.rivm.nl/bibliotheek/rapporten/2015-0144.pdf [accessed 2026-02-26]
4. The real cost e-cigarette prevention campaign. US Food and Drug Administration. 2025. URL: https://www.fda.gov/tobacco-products/real-cost-campaign/real-cost-e-cigarette-prevention-campaign [accessed 2026-02-26]
5. Martinelli T, Candel M, de Vries H, et al. Exploring the gateway hypothesis of e-cigarettes and tobacco: a prospective replication study among adolescents in the Netherlands and Flanders. Tob Control 2023 Mar;32(2):170-178. [doi: 10.1136/tobaccocontrol-2021-056528] [Medline: 34226262]
6. FDA finalizes enforcement policy on unauthorized flavored cartridge-based e-cigarettes that appeal to children, including fruit and mint. US Food & Drug Administration. 2020 Jan. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/enforcement-priorities-electronic-nicotine-delivery-system-ends-and-other-deemed-products-market [accessed 2026-02-26]
7. FDA notifies companies, including puff bar, to remove flavored disposable e-cigarettes and youth-appealing e-liquids from market for not having required authorization. PR Newswire. 2020. URL: https://www.prnewswire.com/news-releases/fda-notifies-companies-including-puff-bar-to-remove-flavored-disposable-e-cigarettes-and-youth-appealing-e-liquids-from-market-for-not-having-required-authorization-301096252.html [accessed 2026-02-26]
8. Ali FRM, Seidenberg AB, Crane E, Seaman E, Tynan MA, Marynak K. E-cigarette unit sales by product and flavor type, and top-selling brands, United States, 2020-2022. MMWR Morb Mortal Wkly Rep 2023 Jun 23;72(25):672-677. [doi: 10.15585/mmwr.mm7225a1] [Medline: 37347717]
9. Farrelly MC, Davis KC, Haviland ML, Messeri P, Healton CG. Evidence of a dose-response relationship between "truth" antismoking ads and youth smoking prevalence. Am J Public Health 2005 Mar;95(3):425-431. [doi: 10.2105/AJPH.2004.049692] [Medline: 15727971]
10. Bal DG, Kizer KW, Felten PG, Mozar HN, Niemeyer D. Reducing tobacco consumption in California. Development of a statewide anti-tobacco use campaign. JAMA 1990 Sep 26;264(12):1570-1574. [doi: 10.1001/jama.1990.03450120082034] [Medline: 2395199]
11. Feng M, Pierce JP, Szczypka G, Vera L, Emery S. Twitter analysis of California's failed campaign to raise the state's tobacco tax by popular vote in 2012. Tob Control 2017 Jul;26(4):434-439. [doi: 10.1136/tobaccocontrol-2016-053103] [Medline: 27601452]
12. McAlister A, Morrison TC, Hu S, et al. Media and community campaign effects on adult tobacco use in Texas. J Health Commun 2004;9(2):95-109. [doi: 10.1080/10810730490425231] [Medline: 15204821]

XSL•FO
RenderX

13. Emery S, Kim Y, Choi YK, Szczypka G, Wakefield M, Chaloupka FJ. The effects of smoking-related television advertising on smoking and intentions to quit among adults in the United States: 1999-2007. Am J Public Health 2012 Apr;102(4):751-757. [doi: 10.2105/AJPH.2011.300443] [Medline: 22397350]
14. Farrelly MC, Nonnemaker J, Davis KC, Hussin A. The Influence of the National truth campaign on smoking initiation. Am J Prev Med 2009 May;36(5):379-384. [doi: 10.1016/j.amepre.2009.01.019] [Medline: 19211213]
15. Terry-McElrath YM, Wakefield MA, Emery S, et al. State anti-tobacco advertising and smoking outcomes by gender and race/ethnicity. Ethn Health 2007 Sep;12(4):339-362. [doi: 10.1080/13557850701300723] [Medline: 17701761]
16. Wakefield M, Terry-McElrath Y, Emery S, et al. Effect of televised, tobacco company-funded smoking prevention advertising on youth smoking-related beliefs, intentions, and behavior. Am J Public Health 2006 Dec;96(12):2154-2160. [doi: 10.2105/AJPH.2005.083352] [Medline: 17077405]
17. Durkin SJ, Brennan E, Wakefield MA. Optimising Tobacco Control Campaigns Within a Changing Media Landscape and Among Priority Populations: Tobacco Control BMJ Publishing Group Ltd; 2022. [doi: 10.1136/tobaccocontrol-2021-056558]
18. Durkin S, Brennan E, Wakefield M. Mass media campaigns to promote smoking cessation among adults: an integrative review. Tob Control 2012 Mar;21(2):127-138. [doi: 10.1136/tobaccocontrol-2011-050345] [Medline: 22345235]
19. Diaz MC, Kierstead EC, Edwards D, et al. Online tobacco advertising and current chew, dip, snuff and snus use among youth and young adults, 2018–2019. Int J Environ Res Public Health 2022 Apr 14;19(8):4786. [doi: 10.3390/ijerph19084786] [Medline: 35457651]
20. Huang J, Duan Z, Kwok J, et al. Vaping versus JUULing: how the extraordinary growth and marketing of JUUL transformed the US retail e-cigarette market. Tob Control 2019 Mar;28(2):146-151. [doi: 10.1136/tobaccocontrol-2018-054382]
21. Vassey J, Valente T, Barker J, et al. E-cigarette brands and social media influencers on Instagram: a social network analysis. Tob Control 2023 Aug;32(e2):e184-e191. [doi: 10.1136/tobaccocontrol-2021-057053] [Medline: 35131947]
22. Liang Y, Zheng X, Zeng DD, Zhou X, Leischow SJ, Chung W. Exploring how the tobacco industry presents and promotes itself in social media. J Med Internet Res 2015 Jan 21;17(1):e24. [doi: 10.2196/jmir.3665] [Medline: 25608524]
23. Donaldson SI, Dormanesh A, Perez C, Majmundar A, Allem JP. Association between exposure to tobacco content on social media and tobacco use: a systematic review and meta-analysis. JAMA Pediatr 2022 Sep 1;176(9):878-885. [doi: 10.1001/jamapediatrics.2022.2223] [Medline: 35816331]
24. Kostygina G, Kim Y, Gebhardt Z, et al. Using exogenous social media exposure measures to assess the effects of smokeless tobacco-related social media content on smokeless tobacco sales in the United States. Nicotine Tob Res 2024 Feb 15;26(Supplement_1):S49-S56. [doi: 10.1093/ntr/ntad169] [Medline: 38366341]
25. Manganello J, Bleakley A, Schumacher P. Pandemics and PSAs: rapidly changing information in a new media landscape. Health Commun 2020 Dec;35(14):1711-1714. [doi: 10.1080/10410236.2020.1839192] [Medline: 33106041]
26. Kite J, Chan L, MacKay K, et al. A model of social media effects in public health communication campaigns: systematic review. J Med Internet Res 2023 Jul 14;25(1):e46345. [doi: 10.2196/46345] [Medline: 37450325]
27. Southwell BG, Yzer MC. When (and why) interpersonal talk matters for campaigns. Commun Theory 2009 Feb;19(1):1-8 [FREE Full text] [doi: 10.1111/j.1468-2885.2008.01329.x]
28. Feng M, Carter CC, Page S, Emery SL, Tran H, Kostygina G. Tweeted, trolled, twisted: battling for narrative control in e-cigarette use prevention campaigns (2014-2020). J Health Commun 2025 Mar 28;30(sup1):39-49. [doi: 10.1080/10810730.2025.2462682] [Medline: 39915966]
29. Majmundar A, Le N, Moran MB, Unger JB, Reuter K. Public response to a social media tobacco prevention campaign: content analysis. JMIR Public Health Surveill 2020 Dec 7;6(4):e20649. [doi: 10.2196/20649] [Medline: 33284120]
30. Liang Y, Zheng X, Zeng DD, Zhou X, Leischow SJ, Chung W. Characterizing social interaction in tobacco-oriented social networks: an empirical analysis. Sci Rep 2015 Jun 19;5(1):10060. [doi: 10.1038/srep10060] [Medline: 26091553]
31. Harris JK, Moreland-Russell S, Choucair B, Mansour R, Staub M, Simmons K. Tweeting for and against public health policy: response to the Chicago Department of Public Health's electronic cigarette Twitter campaign. J Med Internet Res 2014 Oct 16;16(10):e238. [doi: 10.2196/jmir.3622] [Medline: 25320863]
32. Allem JP, Escobedo P, Chu KH, Soto DW, Cruz TB, Unger JB. Campaigns and counter campaigns: reactions on Twitter to e-cigarette education. Tob Control 2017 Mar;26(2):226-229. [doi: 10.1136/tobaccocontrol-2015-052757] [Medline: 26956467]
33. Leas EC, Prochaska JJ, Ayers J, Nobles AL, Henriksen L. What to do when tobacco advertisers exploit antitobacco social media campaigns to sell tobacco. Tob Control 2020 Mar;29(2):243-244. [doi: 10.1136/tobaccocontrol-2019-054993] [Medline: 31249102]
34. Keller FB, Schoch D, Stier S, Yang J. Political astroturfing on Twitter: how to coordinate a disinformation campaign. Polit Commun 2020 Mar 3;37(2):256-280. [doi: 10.1080/10584609.2019.1661888]
35. Chan J. Online astroturfing: a problem beyond disinformation. Philos Soc Crit 2024 Mar;50(3):507-528. [doi: 10.1177/01914537221108467]
36. McCombs ME, Shaw DL. The agenda-setting function of mass media. Public Opin Q 1972;36(2):176. [doi: 10.1086/267990]
37. Meraz S, Papacharissi Z. Networked gatekeeping and networked framing on #Egypt. Int J Press Polit 2013 Apr;18(2):138-166. [doi: 10.1177/1940161212474472]

XSL•FO
RenderX

38. Laffont JJ, Martimort D. The Theory of Incentives: The Principal-Agent Model: Princeton University Press; 2009. [doi: 10.2307/j.ctv7h0rwr]
39. Byrne S, Hart PS. The boomerang effect A synthesis of findings and a preliminary theoretical framework. Annals of the International Communication Association 2009 Jan;33(1):3-37. [doi: 10.1080/23808985.2009.11679083]
40. Feng M, Shi H, Page S, et al. Analyzing U.S. e-cigarette prevention campaigns on TikTok (2020-2023). Presented at: World Association for Public Opinion Research's 77th Annual Conference; Jul 28-31, 2024.
41. Jung S, Murthy D, Bateineh BS, Loukas A, Wilkinson AV. The normalization of vaping on TikTok using computer vision, natural language processing, and qualitative thematic analysis: mixed methods study. J Med Internet Res 2024 Sep 11;26(1):e55591. [doi: 10.2196/55591] [Medline: 39259963]
42. Al-Rawi A, Blackwell B, Zemenchik K, Lee K. Twitter misinformation discourses about vaping: systematic content analysis. J Med Internet Res 2023 Nov 10;25(1):e49416. [doi: 10.2196/49416] [Medline: 37948118]
43. West SM. Raging against the machine: network gatekeeping and collective action on social media platforms. MaC 2017 Sep 22;5(3):28-36. [doi: 10.17645/mac.v5i3.989]
44. The real cost campaign. US Food and Drug Administration. 2017. URL: https://www.fda.gov/tobacco-products/public-health-education-campaigns/real-cost-campaign [accessed 2026-02-26]
45. Levitt HM, Bamberg M, Creswell JW, Frost DM, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):26-46. [doi: 10.1037/amp0000151] [Medline: 29345485]
46. Grummon AH, Hall MG, Mitchell CG, et al. Reactions to messages about smoking, vaping and COVID-19: two national experiments. Tob Control 2022 May;31(3):402-410. [doi: 10.1136/tobaccocontrol-2020-055956] [Medline: 33188150]
47. Kowitt SD, Mendel Sheldon J, Vereen RN, et al. The impact of the real cost vaping and smoking ads across tobacco products. Nicotine Tob Res 2023 Feb 9;25(3):430-437. [doi: 10.1093/ntr/ntac206] [Medline: 36006858]
48. TikTok for developers. Research API. URL: https://developers.tiktok.com/products/research-api/ [accessed 2025-08-01]
49. Truth takes to TikTok to talk vaping and the immune system. truth initiative. 2020. URL: https://truthinitiative.org/press/press-release/truth-takes-tiktok-talk-vaping-and-immune-system-during-covid-19 [accessed 2025-07-28]
50. Vanian J. Apple, Google remove TikTok from stores as app halts service in US. CNBC. 2025. URL: https://www.cnbc.com/2025/01/18/apple-google-remove-tiktok-from-stores-as-app-halts-service-in-us.html [accessed 2026-02-26]
51. Pearson GDH, Silver NA, Robinson JY, Azadi M, Schillo BA, Kreslake JM. Beyond the margin of error: a systematic and replicable audit of the TikTok research API. Information, Communication & Society 2025 Feb 17;28(3):452-470. [doi: 10.1080/1369118X.2024.2420032]
52. Guimaraes A, Wang L, Weikum G. Us and them: adversarial politics on Twitter. Presented at: 2017 IEEE International Conference on Data Mining Workshops (ICDMW); Nov 18-21, 2017; New Orleans, LA p. 872-877. [doi: 10.1109/ICDMW.2017.119]
53. Vaping prevention resources tobacco education resource library. US Food and Drug Administration. URL: https://digitalmedia.hhs.gov/tobacco/educator_hub [accessed 2025-07-28]
54. ChatGPT [large language model]. OpenAI. 2025. URL: https://chat.openai.com/chat [accessed 2026-02-26]
55. Tufekci Z. Twitter and Tear Gas: The Power and Fragility of Networked Protest: Yale University Press; 2017.
56. Abidin C. Mapping internet celebrity on TikTok: exploring attention economies and visibility labours. Cultural Science Journal 2020 Jan 19;12(1):77-103. [doi: 10.5334/csci.140]
57. Literat I, Kligler-Vilenchik N. TikTok as a key platform for youth political expression: reflecting on the opportunities and stakes involved. Social Media + Society 2023 Jan;9(1):20563051231157595. [doi: 10.1177/20563051231157595]
58. Visualizing a decade of Twitter's evolution: Jan 2012 – Nov 2022 & Musk's purchase. The GDELT Project. URL: https://blog.gdeltproject.org/visualizing-a-decade-of-twitters-evolution-jan-2012-nov-2022-musks-purchase/ [accessed 2025-11-06]

Edited by S Brini; submitted 08.Sep.2025; peer-reviewed by RR Ouellette, Y Wang; accepted 01.Jan.2026; published 26.Mar.2026.

Please cite as:
Carter CC, Page S, Borowiecki M, Kostygina G, Emery SL, Feng M
Opposition to Youth e-Cigarette Prevention Campaigns on Twitter and TikTok: Cross-Platform Observational Mixed Methods Analysis
J Med Internet Res 2026;28:e83791
URL: https://www.jmir.org/2026/1/e83791
doi:10.2196/83791

©Chandler C Carter, Simon Page, Mateusz Borowiecki, Ganna Kostygina, Sherry L Emery, Miao Feng. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted

XSL•FO
RenderX

use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Developing and Validating a Machine Learning Algorithm to Predict the Risk of Incident Opioid Use Disorder Among OneFlorida+ Patients: Prognostic Modeling Study

Jabed Al Faysal[1,2], BSc, MSc; Weihsuan Lo-Ciganic[3,4,5], PhD; Walid F Gellad[3,4], MD; Yonghui Wu[6], PhD; Christopher A Harle[7,8], PhD; Khoa Nguyen[9], PharmD; James L Huang[1], PhD; Gerald Cochran[10], PhD; Debbie L Wilson[1], PhD; Stephanie AS Staras[6], PhD; Siegfried OF Schmidt[11], MD, PhD; Eric I Rosenberg[1,12], MD, MSPH; Danielle Nelson[11], MD; Shunhua Yan[1], MEd; Gary M Reisfield[1], MD; William M Greene[13], MD; Courtney Kuza[4], PhD; Md Mahmudul Hasan[1,14], PhD

[1]Department of Pharmaceutical Outcomes & Policy, University of Florida, 1889 Museum Road, Malachowsky Hall, Suite 6300, Gainesville, FL, United States

[10]Department of Internal Medicine, Division of Epidemiology, University of Utah, Salt Lake City, UT, United States

[11]Department of Community Health and Family Medicine, College of Medicine, University of Florida, Gainesville, FL, United States

[12]Department of Internal Medicine, College of Medicine, University of Florida, Gainesville, FL, United States

[13]Department of Psychiatry, College of Medicine, University of Florida, Gainesville, FL, United States

[14]Department of Information Systems and Operations Management, Warrington College of Business, University of Florida, Gainesville, FL, United States

[2]Computer Science and Engineering Discipline, Khulna University, Khulna, Bangladesh

[3]Division of General Internal Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA, United States

[4]Center for Pharmaceutical Policy and Prescribing, University of Pittsburgh, Pittsburgh, PA, United States

[5]Geriatric Research Education and Clinical Center, North Florida/South Georgia Veterans Health System, Gainesville, FL, United States

[6]Department of Health Outcomes & Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, United States

[7]Department of Health Policy and Management, School of Public Health, Indiana University, Indianapolis, IN, United States

[8]Regenstrief Institute, Indianapolis, IN, United States

[9]Department of Pharmacotherapy and Translational Research, College of Pharmacy, University of Florida, Gainesville, FL, United States

Corresponding Author:
Md Mahmudul Hasan, PhD
Department of Pharmaceutical Outcomes & Policy, University of Florida, 1889 Museum Road, Malachowsky Hall, Suite 6300, Gainesville, FL, United States

Abstract

Background: Opioid use disorder (OUD) remains a critical public health crisis in the United States. Despite widespread policy and clinical interventions, early identification of individuals at risk for developing OUD remains challenging due to limitations in traditional screening approaches and a lack of individualized risk stratification methods. Machine learning (ML) methods offer an opportunity to develop timely, high-performing, and explainable predictive models that can enhance OUD prevention strategies in clinical settings.

Objective: This study aims to develop and validate an ML model using electronic health record (EHR) data to predict the 3-month risk of incident OUD among adults initiating opioid therapy and to stratify patients into clinically actionable risk groups.

Methods: This prognostic modeling study used 2017 - 2022 OneFlorida+ EHR data to develop and validate ML algorithms predicting 3-month incident OUD risk. We included 182,083 adults (≥18 y) without cancer, overdose, or OUD or hospice history who received ≥1 outpatient, noninjectable opioid prescription. Using 183 predictors measured in sequential 3-month intervals, we developed an elastic net, least absolute shrinkage and selection operator, gradient boosting machine (GBM), and random forest models on randomly split training, testing, and validation sets. Model performance was assessed using C-statistics, predictive values, and number needed to evaluate, with patients stratified into risk deciles for clinical applicability. Model explainability was assessed using Shapley additive explanations, and fairness was evaluated using standard metrics. We externally validated the best-performing model using an independent cohort from the 2018 - 2020 UPMC (formerly University of Pittsburgh Medical Center) health system.

Results: In the validation sample (n=60,694), GBM (C-statistics=0.879, 95% CI 0.874 - 0.884) and elastic net (C-statistics=0.872, 95% CI 0.867 - 0.877) outperformed least absolute shrinkage and selection operator (C-statistics=0.846, 95% CI 0.840 - 0.851)

XSL•FO
RenderX

and random forest (C-statistics=0.798, 95% CI 0.792 - 0.804), with GBM model requiring the fewest predictors (n=75) for predicting 3-month incident OUD. Using the GBM algorithm to predict the subsequent 3-month OUD risk, the top decile subgroup had a positive predictive value of 3.26%, a negative predictive value of 99.8%, and a number needed to evaluate of 31. The top decile (n=6696) captured ~68% of patients with OUD. Shapley additive explanations analysis identified age, number of outpatient visits, history of back and other pain conditions, comorbidity burden, and opioid prescribing patterns as the strongest predictors of incident OUD. Fairness assessment showed an acceptable false negative rate parity across race, age, and sex. In external validation on the UPMC cohort, the GBM model maintained good discrimination (C-statistics=0.756, 95% CI 0.750 - 0.762) and effective risk stratification.

Conclusions: An ML algorithm predicting incident OUD derived from OneFlorida+ EHR data performed well in external validation with data using UPMC. The algorithm might be valuable for incident OUD risk prediction and stratification across health systems, with potential to inform early intervention.

(J Med Internet Res 2026;28:e79482) doi:10.2196/79482

KEYWORDS

opioid use disorder; machine learning; OneFlorida+; risk stratification; external validation

Introduction

The United States continues to face a persistent and evolving opioid epidemic. Opioid use disorder (OUD) and overdose affect millions of Americans, leading to increased morbidity, mortality, and health care costs [1,2] In 2022, more than 6 million individuals experienced OUD, imposing significant societal and economic burdens exceeding US $78 billion annually [1,3,4]. Opioid-related deaths increased tenfold from 1999 to 2022 (>82,000 in 2022) [5-8]. Prescription opioids were responsible for approximately 280,000 overdose deaths over the same period [9]. The prevalence of OUD continues to rise despite efforts to curb opioid misuse [10,11]. In response, health systems, payers, and policymakers have implemented various interventions aimed at reducing unsafe prescribing and patient risk [12].

Early identification of OUD can prevent severe addiction, high-risk behaviors, overdose, and death, while enabling timely access to evidence-based treatments such as buprenorphine or referral to recovery services. Existing methods for identifying high-risk opioid users, such as high-dose opioid prescribing thresholds and use of multiple pharmacies, are often based on simple or single criteria [13,14]. These rule-based strategies leave many high-risk individuals undetected and may misclassify others, resulting in unintended consequences like delayed care, unnecessary monitoring, or stigma. For example, a study found that nearly 43% of individuals who developed OUD were missed by such approaches [15]. Furthermore, traditional statistical models emphasize population-level risk factors and assume linear relationships between predictors and outcomes, limiting their ability to capture complex interactions among patient demographics, comorbidities, health care use, and prescriber patterns. Although tools such as the Screener and Opioid Assessment for Patients with Pain-Revised and the Opioid Risk Tool are commonly used in clinical settings, studies report inconsistent estimates of test accuracy and no evidence that they reduce OUD or overdose [16-18]. The 2022 Centers for Disease Control and Prevention Clinical Practice Guideline similarly concluded that current opioid risk screening tools lack proven clinical use. Proprietary platforms such as NarxCare leverage prescription and health record data but lack transparency, public validation, and demonstrated generalizability [19]. These limitations underscore the need for more robust, transparent, and generalizable models to support real-time clinical decision-making across diverse populations. Studies have also highlighted these shortcomings and called for more advanced, data-driven models to improve identification of individuals at risk (or low risk) of OUD [20-24].

Machine learning (ML) offers a promising alternative to traditional statistical methods in improving risk prediction. ML models can analyze large health care data to identify complex nonlinear patterns in patient data. Recent studies have demonstrated the potential of ML to predict opioid-related outcomes, including OUD and overdose, with higher accuracy than that of traditional methods [25-28]. Our previous work showed that ML approaches can improve risk prediction and stratification for incident OUD and subsequent overdose in Medicare and Medicaid beneficiaries. For example, we developed ML models to predict incident OUD among Medicare beneficiaries, achieving high discrimination (C-statistic >0.86) and effectively stratifying patients into clinically meaningful risk subgroups (eg, low, moderate, and high risk for developing OUD) [29]. However, many existing models are developed using Medicare or Medicaid data, which primarily include older adults or low-income populations.

In this study, we extend our work to develop and validate an ML algorithm to predict incident OUD among patients who received care from institutions within the OneFlorida+ network—a large, diverse patient population across Florida, Georgia, Alabama, and Arkansas, including individuals with commercial insurance, Medicaid, or Medicare, and those with no insurance [30]. This study advances our prior Medicare-based work in several important ways. First, although claims-based models can be incorporated into some clinical workflows, electronic health records (EHRs) often lack the full use and medication detail available in claims and may vary in data quality. As such, adapting the model for EHR-based use required methodological refinement to support real-time, point-of-care integration. Second, the broader and more heterogeneous cohort enhances the model's generalizability beyond the older, fee-for-service Medicare population used previously. Third, we externally validated the model using an independent cohort

XSL•FO
RenderX

from the UPMC health system, assessing cross-system generalizability and transportability—an important step toward real-world implementation. Fourth, we incorporated fairness and bias analyses across demographic subgroups to address equity considerations often overlooked in prior OUD risk prediction studies. Finally, we articulate a translational vision for integrating this model into clinical workflows. By organizing predictors to enable risk-based stratification, the model is designed to function as an EHR-embedded decision support tool in primary care settings, alerting providers (eg, physicians) when a patient is flagged as high-risk to prompt timely, preventive actions. We emphasize model interpretability to ensure that our findings can be seamlessly integrated into the clinical environment.

Methods

Study Design and Data Source

This is a prognostic modeling study with a retrospective cohort design. We used the TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + Artificial Intelligence) guidelines for reporting our work (Checklist 1) [31]. This study used data from OneFlorida+, a secure and centralized repository of patient‐level data from both health system partners and insurers, and managed by the OneFlorida Clinical Research Consortium [32]. Currently, OneFlorida+ includes EHRs and claims data from 26 million individuals across Florida, Georgia, Alabama, and Arkansas, covering a wide range of populations and health care settings [30,33]. All data sources are mapped to the Patient‐Centered Outcomes Research network common data model (version 7.0) to ensure standardization of data elements across sources. Major data elements in the common data model include demographics, enrollment, encounters, diagnoses, procedures, dispensed medications, and deaths. Additionally, while OneFlorida data includes Florida Medicaid recipients with claims data only, our study cohort requires having at least 1 EHR encounter. The study population included adult patients (aged ≥18 y) who had at least 1 opioid prescription filled between 2017 and 2022 (Figure S1 in Multimedia Appendix 1). The index date was defined as the date of a patient's first opioid prescription between October 1, 2017, and September 30, 2022. To construct a cohort appropriate for assessing incident OUD risk, we excluded patients who (1) had cancer diagnoses (Table S1 in Multimedia Appendix 2), (2) received hospice care, (3) had a diagnosis of OUD, opioid overdose, other substance use disorders, drug abuse, or received methadone or buprenorphine for OUD before initiating opioids, or (4) did not have at least a 3-month observation window following the first opioid prescription to allow for predictor measurement. We also excluded patients with a diagnosis of other substance use disorders to minimize confounding, as some physicians may have used this diagnosis for patients with co-occurring OUD and other substance use conditions. Once eligible, patients remained in the cohort until they experienced an outcome of interest (ie, incident of OUD diagnosis) or were censored due to death, regardless of continued opioid use (Figure S2 in Multimedia Appendix 1).

For external validation, we applied the algorithm to 2018‐2020 EHR data from the UPMC health system in Pennsylvania, a region with demographic and opioid prescribing characteristics distinct from the OneFlorida+ network. Cohort construction and predictor generation followed the same procedures used for the OneFlorida+ dataset.

Outcome Variable

Similar to other claims-based analyses [29,34-36], our primary outcome was incident OUD (Table S2 in Multimedia Appendix 2), defined as the first recorded diagnosis of OUD from all settings in each 3-month window after the index prescription date (Figure S2 in Multimedia Appendix 1). The *ICD-10* (*International Statistical Classification of Diseases and Related Health Problems 10th Revision*) codes for incident OUD diagnosis include F11.1X and F11.2X but exclude F11.11 (opioid-related disorders in remission) and F11.21 (opioid dependence in remission).

Predictor Candidates

We identified 183 candidate predictors of OUD from the prior literature (Table S3 in Multimedia Appendix 2) [37-48]. These predictors included sociodemographic factors, patient health status, and patterns of opioid use and other nonopioid prescriptions measured at baseline (during the 3 month period before the first opioid fill) and in 3-month windows after initiating prescription opioids. This 3-month window was selected for both clinical and operational relevance. Evidence shows that early indicators of problematic opioid use (eg, dose escalation, early refills, or polypharmacy) typically emerge within the first 3 months of therapy [49-52]. Pharmacoepidemiologic studies demonstrate that transitions to sustained or high-risk opioid use most commonly occur within 3 months after treatment initiation [53,54]. This timeframe also aligns with the quarterly monitoring cycles used by many prescription drug monitoring programs and health plans [14,38]. We updated the predictors measured in each 3-month period to account for changes over time for predicting incident OUD risks in each subsequent period (Figure S2 in Multimedia Appendix 1). This time-updating approach mimics active surveillance that a health system might conduct in real time [55] to provide clinicians with timely, actionable opportunities (eg, closer follow-up, patient counseling, or medication adjustments) before the progression to OUD. We examined missingness across candidate predictors and applied prespecified imputation rules before model training. We used simple imputation techniques where missing values for all categorical predictors (including race, ethnicity, and provider sex) were imputed using the modal category. Continuous predictors (eg, use counts, comorbidity scores, and medication counts) with missing values were imputed using the median calculated from the full analytic dataset. Binary and count predictors representing the presence or frequency of diagnoses, procedures, or medications were imputed as zero when missing, consistent with the interpretation that the absence of a code reflects no recorded event during the baseline window.

XSL•FO
RenderX

ML Approaches and Prediction Performance Evaluation

Our machine learning analysis using OneFlorida+ data had two objectives: (1) developing a prediction model to generate individuals' incident OUD risk scores and (2) stratifying individuals into subgroups based on similar OUD risk levels. We developed and tested 4 ML algorithms to predict incident OUD: elastic net (EN), least absolute shrinkage and selection operator (LASSO), gradient boosting machine (GBM), random forest (RF). These algorithms have consistently demonstrated strong predictive performance in prior studies [12,46,56,57]. Details of each algorithm are provided in Multimedia Appendix 3. We randomly divided the cohort into 3 subsets: a training sample for algorithm development, a testing sample for algorithm refinement, and a validation sample for evaluating predictive performance. This split was conducted strictly by patient ID, such that all 3-month episodes from a given patient were assigned to a single dataset (training, testing, or validation), and no patient contributed episodes to more than one set, thereby avoiding information leakage across datasets. Because patients could contribute multiple 3-month episodes, episode-level observations within patients were correlated by design. Models were trained and evaluated at the episode level to reflect the intended real-world use case of repeated, longitudinal risk assessment over time and to maximize learning from rare outcomes. Tree-based ensemble methods such as GBMs are generally robust to correlated observations because they do not rely on independence assumptions and learn predictive structure through recursive partitioning rather than parametric estimation. To evaluate the potential impact of within-patient correlation on model performance estimates, we conducted sensitivity analyses using patient-level random subsets. We used one 3-month period with predictor measurements to forecast risk in the subsequent 3 months within the validation set.

To evaluate discrimination performance, we assessed whether individuals predicted to be at high risk had higher OUD incidence than those predicted to be at low risk. We compared C-statistics across different methods in the validation sample using the DeLong test [58]. A C-statistic between 0.7 and 0.8 indicates good discrimination, while values above 0.8 indicate very good discrimination. Precision-recall curves were also examined [42]. Given that OUD events are rare outcomes and C-statistics do not incorporate information about outcome incidence, we reported other metrics, including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), number needed to evaluate (NNE) to identify 1 OUD episode, and estimated rate of alerts, to thoroughly assess the algorithms' predictive ability (Figure S3 in Multimedia Appendix 1) [43,59].

There is no universally applicable prediction probability threshold, so we evaluated performance across multiple sensitivity and specificity levels (eg, selecting 90% sensitivity). To determine an optimized threshold that balances sensitivity and specificity, we applied the Youden index in the training sample [60]. In the validation sample, patients were stratified into risk subgroups based on deciles of predicted incident OUD probability. The highest decile was further divided into 3 strata: the top first percentile, the second to fifth percentiles, and the sixth to tenth percentiles. This approach allowed for a more detailed assessment of individuals at the highest risk of developing OUD.

We conducted additional analyses to enhance clinical use. The primary goal of our ML algorithm was to generate risk scores for incident OUD. For comprehensive model explanation and further enhancement of model interpretability, we conducted a Shapley additive explanations (SHAP) analysis [61]. SHAP quantifies the contribution of each predictor to the model's output while accounting for interactions with other features. The SHAP plots indicate whether the mentioned features have a positive correlation (red color) or a negative correlation (blue color) with OUD. To evaluate bias across demographic factors (eg, race, age, and sex), we examined risk score distributions and false positive rates (FPR) and false negative rates (FNR) in the best performing model. We defined significant disparity as a parity ratio between subgroups falling outside the range of 0.80 to 1.25, consistent with the standards used in algorithmic fairness audits [62].

Additionally, we evaluated model calibration to assess the reliability of predicted risk scores. Given the potential for miscalibration in rare-outcome ML models, we applied isotonic regression to recalibrate predicted risk scores [63]. The overall accuracy of probabilistic predictions was assessed using the Brier score metric [64]. To determine clinical use beyond discrimination, we performed decision curve analysis. We calculated the net benefit [65] across a range of clinically relevant risk thresholds, comparing the recalibrated model against default strategies of "treat all" and "treat none". This approach quantifies the trade-off between the benefit of identifying true OUD cases and the potential harm of false-positive alerts.

Statistical Analysis

All analyses were performed using SAS 9.4 (SAS Institute Inc) and Python 3.6 (Python Software Foundation). Continuous variables were summarized using mean (SD) values, while categorical variables were summarized using frequencies and percentages. We compared patients' characteristics in training, testing, and validation datasets using 2-tailed Student *t* tests, chi-square tests, and ANOVA, or corresponding nonparametric tests.

Ethical Considerations

This study was reviewed and approved by the University of Florida (UF) Institutional Review Board (IRB202101897). The research involved secondary analysis of limited electronic health record data from the OneFlorida+ Data Trust. In accordance with federal regulations (45 CFR 46.104[d] [4]) and institutional policies, the requirement for informed consent was waived because the study involved no direct participant contact and posed minimal risk to individuals. All analyses were conducted within the secure UF server environment; no identifiable private information was accessed or transmitted outside the system. The study adhered to institutional data-use agreements and complied with all relevant guidelines for human subjects research and data privacy. The external validation using UPMC data was reviewed by the University of Pittsburgh Human

XSL•FO
RenderX

Research Protection Office and determined not to constitute human subjects research, as only deidentified data were used. Therefore, institutional review board review and informed consent were not required.

Results

Patient Characteristics

OneFlorida+ patients in the training (n=60,694), testing (n=60,695), and validation (n=60,694) samples had similar characteristics and outcome distributions (mean age 53.02, SD 7.60 y; 111,576/182,083, 61.28% female; 125,208/182,083, 68.76% White; 144,316/182,083, 79.26% non-Hispanic; 4328/182,083, 2.38% had incident OUD; Table 1). The external validation cohort from the UPMC included 129,215 patients (mean age 58.91, SD 16.75 y; 78,574/129,215, 60.81% female; 115,907/129,215, 89.7% White; 123,083/129,215, 95.25% non-Hispanic; 2537/129,215, 1.96% had incident OUD). While sex distributions were similar between the 2 health systems, the UPMC cohort was older and less diverse, with more White and non-Hispanic patients.

Table . Selected characteristics of OneFlorida+ and UPMC[a] health system patients.

Characteristics	2017 - 22 OneFlorida+ data			2018 - 20 UPMC[a] data
	Training (n=60,694)	Testing (n=60,695)	Internal validation (n=60,694)	External validation (n=129,215)
Development of incident opioid use disorder, n (%)	1443 (2.38)	1433 (2.36)	1452 (2.39)	2537 (1.96)
Age (years), mean (SD)	53.11 (17.57)	52.98 (17.64)	52.98 (17.57)	58.91 (16.75)
Age group (years), n (%)				
18 - 34	10,200 (16.81)	10,286 (16.95)	10,285 (16.95)	13,852 (10.73)
35 - 50	15,326 (25.25)	15,501 (25.54)	15,477 (25.5)	23,931 (18.52)
51 - 64	19,121 (31.5)	19,067 (31.41)	19,096 (31.46)	44,047 (34.10)
≥65	16,047 (26.44)	15,841 (26.1)	15,836 (26.09)	47,385 (36.65)
Sex, n (%)				
Male	23,534 (38.77)	23,489 (38.70)	23,481 (38.69)	50,641 (39.19)
Female	37,160 (61.23)	37,206 (61.30)	37,210 (61.31)	78,574 (60.81)
Race, n (%)				
White	41,737 (68.77)	41,751 (68.79)	41,720 (68.74)	115,907 (89.70)
Black	13,403 (22.08)	13,458 (22.17)	13,548 (22.32)	10,833 (8.38)
Other or unknown	5554 (9.15)	5486 (9.04)	5426 (8.94)	2475 (1.92)
Ethnicity, n (%)				
Non-Hispanic	47,951 (79)	48,242 (79.48)	48,123 (79.29)	123,083 (95.25)
Hispanic	10,756 (17.72)	10,455 (17.23)	10,540 (17.37)	1013 (0.78)
Other or unknown	1987 (3.27)	1998 (3.29)	2031 (3.35)	5123 (3.96)

[a]UPMC: Formerly known as University of Pittsburgh Medical Center.

Prediction Performance Across ML Methods

Figure 1 summarizes the 4 prediction performance measures of each model in the internal validation sample. GBM (C-statistics=0.879, 95% CI 0.874 - 0.884) and EN (C-statistics=0.872, 95% CI 0.867 - 0.877) models outperformed the LASSO (C-statistics=0.846, 95% CI 0.840 - 0.851) and RF (C-statistics=0.798, 95% CI 0.792 - 0.804) models, with the GBM model requiring the fewest predictors (n=75) for predicting 3-month incident OUD (P<.001; Figure 1A). The GBM and EN models had similar prediction performance, and the GBM model had the best precision-recall performance (Figure 1B). Sensitivity analyses using patient-level data yielded similar results (Figure S4 in Multimedia Appendix 1). To evaluate model robustness across key clinical subgroups, we stratified the GBM model's performance by age, sex, and ethnicity. The model demonstrated consistent discriminative ability across populations, with C-statistics ranging from 0.833 to 0.909 across age groups (Figure S5 in Multimedia Appendix 1), 0.865 to 0.887 between sexes (Figure S6 in Multimedia Appendix 1), and 0.864 to 0.914 by ethnicity (Figure S7 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 1. Performance matrix across machine learning models for predicting incident opioid use disorder in OneFlorida+ patients. This figure shows 4 prediction performance matrices in the validation sample. (A) The areas under receiver operating characteristic curves (or C-statistics). (B) The precision-recall curves (precision=positive predictive value and recall=sensitivity): precision-recall curves that are closer to the upper right corner or are above another method have improved performance. (C) The number needed to evaluate by different cutoffs of sensitivity. (D) Alerts per 100 patients by different cutoffs of sensitivity. AUC: area under the curve; EN: elastic net; GBM: gradient boosting machine; LASSO: least absolute shrinkage and selection operator; PPV: positive predictive value; RF: random forest; ROC: receiver operating characteristics.

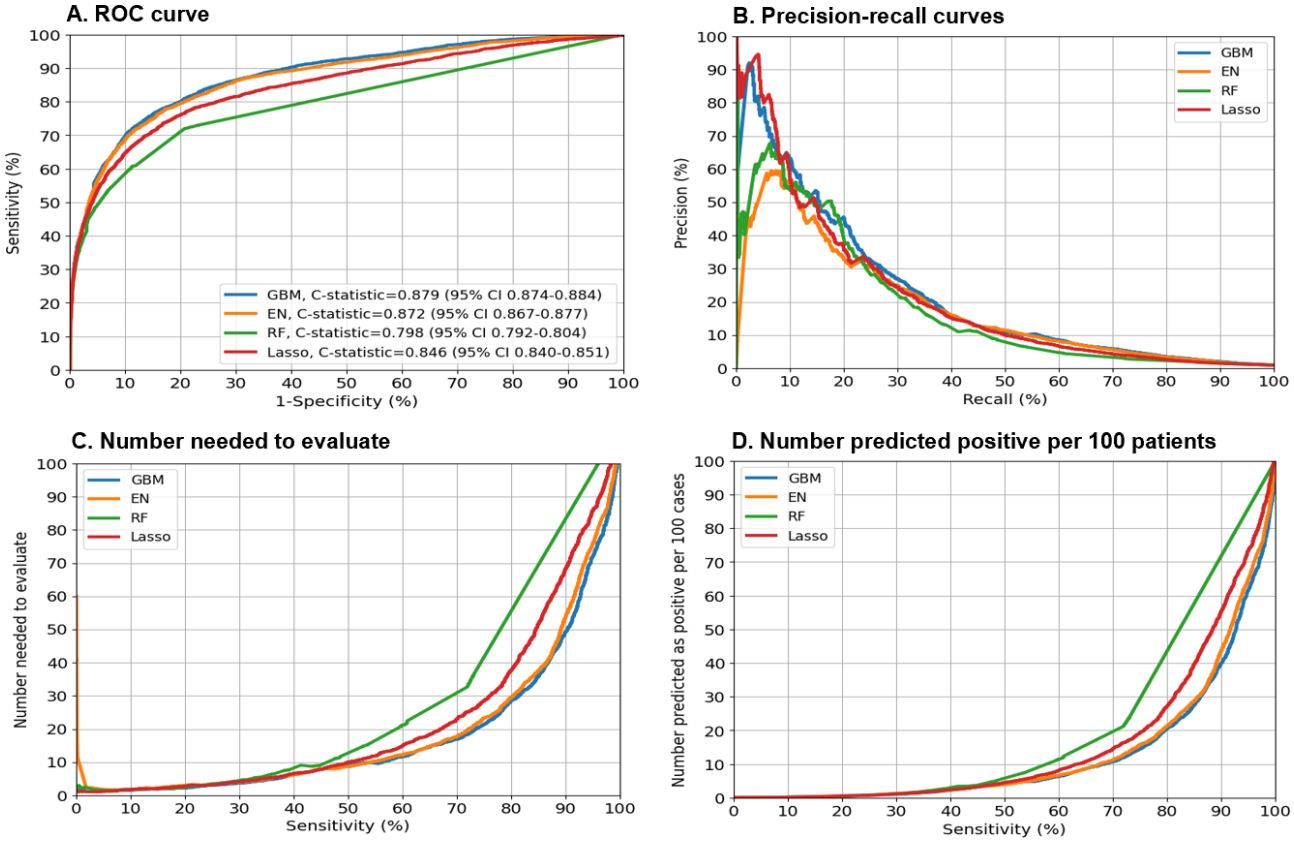

Table S4 in Multimedia Appendix 2 shows the performance measures for predicting incident OUD across different levels (90% - 100%) of sensitivity and specificity for GBM and EN. When set at the optimized sensitivity and specificity as measured by the Youden index, the GBM model had a sensitivity of 76.59%, specificity of 84.75%, PPV of 4.36%, NPV of 99.75%, and NNE of 23; and the EN model had a sensitivity of 77.65%, specificity of 82.74%, PPV of 3.93%, NPV of 99.76%, and NNE of 25 (Figures 1C and 1D; Table S4 in Multimedia Appendix 2).

Risk Stratification, Bias and Fairness Analysis, and Model Explainability

Figure 2 represents the actual OUD rate for individuals in each decile subgroup using GBM. The high-risk subgroup (with risk scores in the top decile; 11% [n=6696] of the validation cohort) had a PPV of 3.26%, a NPV of 99.8%, and NNE of 31. In a hypothetical clinical application involving 1000 patients, applying the top decile threshold would generate alerts for approximately 100 individuals. Based on the observed PPV at this threshold, approximately 4 of the 100 flagged patients would be expected to develop OUD, demonstrating the model's ability to meaningfully concentrate risk among a small subgroup. Conversely, the high NPV (99.8%) indicates that very few incident OUD cases would arise among the 900 unflagged patients, consistent with the model's low false negative rate. Taken together, these estimates illustrate how the model could support prioritization of patients most likely to benefit from closer monitoring or preventive intervention.

XSL•FO
RenderX

Figure 2. Incident opioid use disorder was identified by the gradient boosting machine's decile risk subgroup in the validation sample. Based on the individual's predicted probability of an opioid use disorder event, we classified patients in the internal validation sample into decile risk subgroups, with the highest decile further split into 3 additional strata based on the top first, second to fifth, and sixth to tenth percentiles to allow closer examination of patients at the highest risk of developing opioid use disorder.

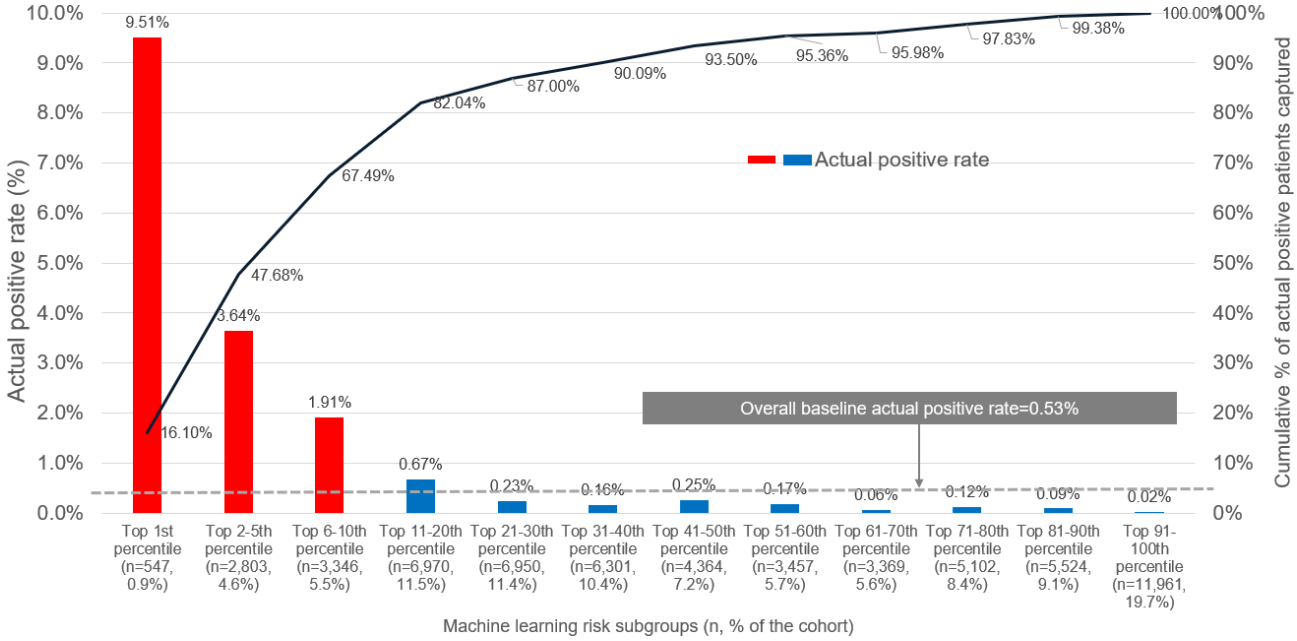

Among all 323 individuals with an incident OUD, 265 (82.04%) occurred in the top two decile subgroups (Decile 1=67.5 % and Decile 2=14.6%). Those in the first decile subgroup had at least a 14-fold higher OUD rate compared to the lower-risk groups (eg, observed OUD rate: Decile 1=9.51%, Decile 2=0.67%, Decile 10=0.02%). The third through tenth decile subgroups had minimal OUD incidence rate (2 to 16 per 10,000).

In our evaluation of racial fairness, we found that among the true positive cases identified in the top decile risk subgroup by the GBM model, 78.44% (171/218) were White individuals and 18.8% (41/218) were Black individuals. In the top first percentile, the distribution was 67.3% (35/52) White individuals and 25% (13/52) Black individuals. The detailed racial distribution of true positives across all risk subgroups is provided in Table S5 in Multimedia Appendix 2. We also compared FNR and FPR by race in the GBM model with and without race as a predictor. FNRs were consistently higher for Black individuals compared to White individuals across most percentiles, while FPR remained similar across racial groups (Figure S8 and Figure S9 in Multimedia Appendix 1). We compared the FNR and FPR between racial groups at the 90th percentile and found that the FNR parity ratio (1.23) and FPR parity ratio (1.03) were within the accepted threshold (0.8 - 1.25) [62], suggesting equitable model performance across racial groups when race was included as a predictor. We also assessed fairness across age and sex. The true positive cases in the top decile were concentrated among the 35 - to 50-year age group (82/218, 37.61%) and females (146/218, 66.97%). The detailed distribution of true positives by age and sex across all risk subgroups is provided in Tables S6 and S7 in Multimedia Appendix 2. FNRs were slightly higher for older individuals (≥65 y) across most percentiles compared to the young adults (18 - 34 y), while FNRs for males and females were closely aligned. FPRs remained consistent across both age groups and sex (Figures S10 and S11 in Multimedia Appendix 1). Parity ratios for sex (FNR=0.91, FPR=1.18) and age-based FPR (0.91; comparing 18 - 34 vs ≥65 y) remained within the acceptable range (0.8 - 1.25), whereas the age-based FNR parity ratio (0.74) was below the threshold.

Figure 3 presents variable importance plots using SHAP values derived from the GBM model. The plots on the right-hand side present the average impact of various features on incident OUD, listed in descending order of their significance. Variables such as age and history of other pain for each 3-month period were negatively correlated with incident OUD, while factors having a positive correlation with incident OUD included the number of outpatient visits, history of back pain, and Elixhauser comorbidity index score.

XSL•FO
RenderX

Figure 3. Variable importance plots using Shapley additive explanations values from the gradient boosting machine model. (A) Shapley additive explanations value computed from individual features' values and their impact (both positive and negative) on incident opioid use disorder. (B) Average Shapley additive explanations values of features showing average impact on and correlation with incident opioid use disorder. SHAP: Shapley additive explanations.

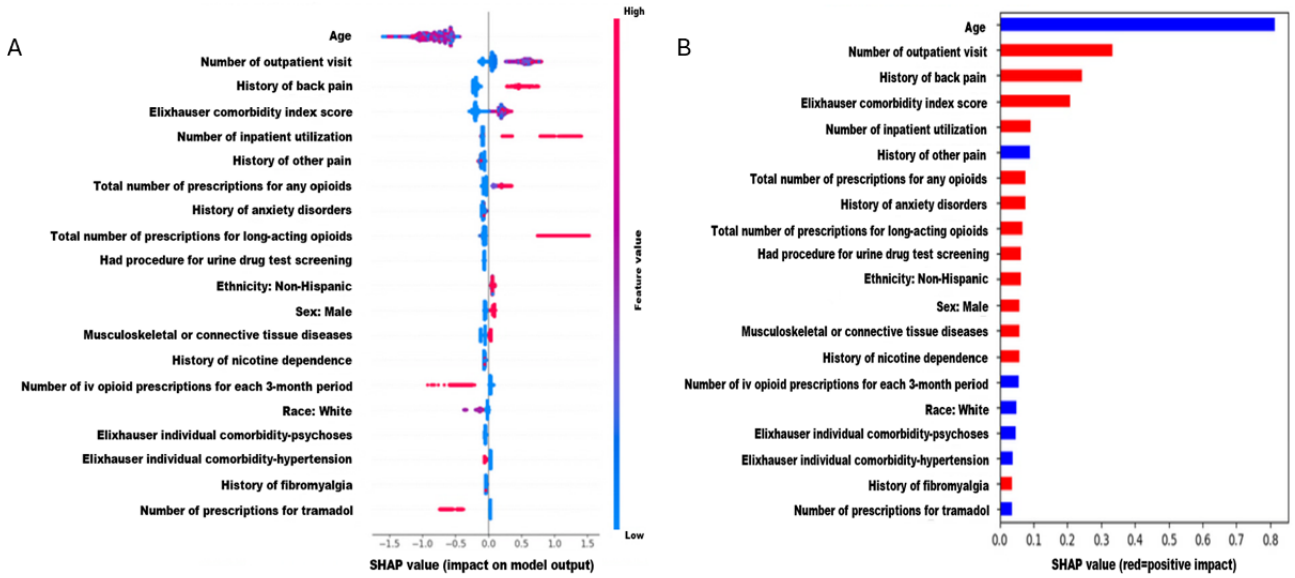

Calibration and Decision Curve Analysis

Calibration assessment showed that the GBM model initially overestimated risk (Brier score=0.09), a known phenomenon in rare-event prediction. After isotonic recalibration, model reliability improved significantly, with a calibration slope of 0.98 (ideal=1), an intercept of –0.03 (ideal=0), and a Brier score of 0.01 (lower is better; Figure S12 in Multimedia Appendix 1). Decision-curve analysis showed that a "treat-all" strategy (intervene on everyone) yields negative net benefit at risk thresholds above 0.5% (Figure S13 in Multimedia Appendix 1), whereas the GBM model maintained positive net benefit across a broader range of clinically relevant thresholds. For example, at the top decile cutoff (probability threshold=0.009), the model achieved a net benefit of 0.0026. This corresponds to identifying approximately 2.6 additional true-positive OUD cases per 1000 patients (after accounting for false-positive alerts) compared with a strategy of no intervention. Relative to the OUD outcome incidence (0.53% in the prediction window), this net benefit reflects ~50% of the maximum achievable use for a perfect classifier in this setting, indicating substantial clinical value despite the low-prevalence context.

External Validation Performance

When applied to the UPMC cohort, the best-performing GBM model achieved a C-statistic of 0.756 (95% CI 0.750 - 0.762; Figure S14 in Multimedia Appendix 1). The high-risk decile subgroup demonstrated a PPV of 2.52%, an NPV of 99.8%, and NNE of 40 (Figure S15 in Multimedia Appendix 1). The top first percentile had an OUD incidence rate 14-fold higher than the overall baseline (8.15% vs 0.58%). Fairness analyses in the UPMC cohort revealed variations in FNR and FPR across age groups (Figure S16 in Multimedia Appendix 1), while error rates remained similar across race (Figure S17 in Multimedia Appendix 1) and between males and females (Figure S18 in Multimedia Appendix 1). For example, using the 90th percentile risk score as the elevated risk cutoff, younger adults (18 - 34 y) had a lower FNR (fewer missed cases) compared to older adults (≥65 y).

Discussion

Principal Findings

This study expanded our previous work using ML approaches to improve the accuracy of predicting incident OUD in the subsequent 3 months of prescription opioid initiation among fee-for-service Medicare beneficiaries and broaden the applicability of these models across a diverse population in the OneFlorida+ network [29]. Our GBM and EN models demonstrated strong predictive performance, achieving a high C-statistic (>0.87) and outperforming LASSO and RF. Our best-performing GBM model offers several advantages, including its ability to eliminate the need for a separate feature selection process and its flexibility in hyperparameter tuning to capture complex interactions between predictors and outcomes. We acknowledge that this flexibility in model tuning can be computationally intensive and time-consuming. The algorithm effectively stratified the population into distinct risk groups based on predicted risk scores. Notably, approximately 80% of the cohort had minimal OUD risk, while the highest-risk decile alone accounted for ~68% (218/323) of all individuals who developed OUD. Given the low incidence of OUD within a 3-month period, the PPV was expectedly low. However, in the context of OUD prevention, a lower PPV may be ethically and clinically acceptable when associated interventions are supportive (eg, closer follow-up, motivational interviewing, review of opioid therapy, or other high-risk concurrent medication use) rather than punitive actions (eg, refusal to prescribe or abrupt discontinuation). For these low-risk, supportive interventions, the potential harm of a false positive (providing additional support to a patient who would not have developed OUD) is minimal compared with the harm of missing a true high-risk case (false negative). It is therefore important to frame the model as augmenting, not replacing, clinician

XSL•FO
RenderX

judgment to ensure that low PPV does not lead to overreaction to false positives or unintended consequences.

A key strength of this study is the successful external validation of our algorithm using an independent UPMC cohort. Our model maintained good performance (C-statistic: 0.76) without retraining, demonstrating transportability across two distinct health care systems. This replication supports the stability of the key predictors across settings and indicates that the model can effectively identify high-risk individuals in diverse clinical environments.

Prior studies, targeting different aspects of OUD risk, varied in their prediction windows and data sources, including 6-month OUD risk based on private insurance claims [66]; 12-month risk of aberrant opioid use behaviors following an initial pain clinic visit [67]; and 12-month OUD risk using private insurance claims [38,68] or pharmacy benefit manager claims data [36]. Other models focused on longer-term risk, such as 2-year problematic opioid use documented in EHR from primary care [69] and 5-year OUD risk using electronic medical record from a medical center [34] or Rhode Island Medicaid data [35]. Despite their contributions, these models had key limitations. Most measured predictors are only at baseline, lacking temporal updates over time. Several relied on case-control designs, which may not generalize well to population-level OUD incidence rates. Additionally, even in non–case-control designs, the highest reported C-statistic was 0.85, indicating room for improvement in predictive performance [35,36,67,69]. Expanding on our prior Medicare claims-based study, this work explored whether predictive performance can be refined using EHR data. Our study addressed the abovementioned limitations by using a population-based cohort and a more immediate prediction window, estimating OUD risk within the next 3 months rather than over a year or longer. This short-term predictive approach, combined with innovative risk stratification and enhanced clinical interpretability, made our method highly applicable for timely intervention.

Limitations

Although the findings of the study are promising, some limitations need to be acknowledged. First, incident OUD was identified solely through *ICD-10* diagnosis codes, which may undercapture true OUD cases due to known under-coding, variability in diagnostic practices, and reluctance to document substance use disorders in clinical settings. As a result, some individuals with OUD may have been misclassified as noncases. Although prior work has similarly relied on diagnosis-based definitions [29,34-36], broader definitions incorporating buprenorphine treatment or combinations of high-risk prescribing behavior indicators may improve case ascertainment. Second, unmeasured predictors, such as socioeconomic determinants and illicit opioid use, could influence risk trajectories. Third, while the study uniquely assessed OUD risk within 3 months of opioid initiation, its reliance on older EHR data may limit real-time applicability due to potential delays in data availability and processing.

Clinical Utility

From a clinical use perspective, the model's ability to stratify patients by risk supports early intervention and more efficient resource allocation in primary care settings. Our analysis confirms that the model's risk estimates are not only accurate in distinguishing higher- from lower-risk patients but also well calibrated after post hoc adjustment. Decision-curve analysis further demonstrated that the model offers meaningful clinical utility across thresholds where default "treat-all" strategies offer little value in the context of a low outcome incidence. At the top decile threshold, the model achieved a positive net benefit equivalent to identifying approximately 2.6 additional true OUD cases per 1000 patients—representing about 50% of the maximum achievable use given the cohort's baseline incidence. These findings align with established decision-analytic frameworks [70] and highlight the model's potential to effectively prioritize individuals most likely to benefit from preventive intervention. Integrating such models into clinical decision support (CDS) tools could enable providers to proactively flag and manage patients at elevated risk for OUD before symptoms escalate. However, ethical considerations must also be addressed, including the potential for unintentionally reinforcing existing health disparities. Although the overall FNR and FPR parity ratio was within acceptable bounds, Black patients exhibited higher false negative rates across percentiles, suggesting that true OUD cases may be underidentified in this group. To address this disparity, we outline potential mitigation strategies for future model refinement. These include group-specific threshold adjustment to harmonize error rates, reweighting or cost-sensitive learning to shift model attention toward underidentified subgroups, and postprocessing approaches such as calibration-by-group or equalized-odds adjustment. Importantly, any fairness intervention should balance reducing disparities with avoiding unintended undertreatment or new forms of inequity. These considerations underscore the need for iterative fairness evaluation as part of a responsible deployment framework.

Clinical Implementation and Future Directions

To support real-world deployment, our translational roadmap includes a structured, multiphase implementation strategy grounded in principles of human-centered design and clinical workflow alignment. In the retrospective phase, we will operationalize the OUD risk model within the UF Health Integrated Data Repository, a comprehensive enterprise clinical data warehouse. This step will enable validation using real-time patient data and ensure backend compatibility with the Epic EHR system. The model will generate individual-level risk scores for patients with recent opioid prescriptions, using secure internal infrastructure and consistent EHR data formatting. In the next phase, we will begin with a feasibility test and pilot integration of the model into the EHR system (Epic) within UF Health, gathering real-world feedback from clinicians before scaling up. Notably, UF Health's Epic EHR infrastructure has already implemented a system-wide approach for an artificial intelligence–driven CDS tool that predicts patients with a high risk of opioid overdose in primary care workflows [71,72].

XSL•FO
RenderX

In the near term, we plan to deploy the OUD risk prediction model through EHR-based CDS tools that seamlessly integrate into routine care. The model's risk score will trigger automated alerts during opioid prescribing encounters, prompting timely interventions for those identified as high risk. For example, in a primary care clinic with approximately 1000 patients initiating opioid therapy annually, applying a top decile (10%) risk threshold would flag approximately 100 individuals as high risk, of whom ~4 would be expected to develop OUD within 3 months. Although the PPV remains low due to the rarity of incident OUD, the alert volume is manageable in routine practice, particularly when recommended interventions are supportive and low-intensity (eg, increased follow-up, medication review, and motivational interviewing). The remaining ~90% low-risk patients would have an extremely high NPV, allowing providers to continue appropriate pain management for those individuals without unnecessary alarm. To further address alert fatigue, we propose specific operational strategies such as (1) embedding alerts within existing opioid or pain management workflows (eg, triggering only when signing a new medication order) to minimize disruption, (2) using tiered alerting triggered by repeated high-risk predictions, and (3) combining risk scores with other clinically meaningful indicators (eg, concurrent benzodiazepine prescribing) to enhance alert specificity. To minimize potential harms such as stigma and inappropriate opioid restriction, alerts should be framed using universal-precaution language that encourages supportive monitoring rather than implying misuse.

Although our findings indicate that the model has characteristics that could support future integration into clinical workflows, real-world feasibility and impact have not yet been evaluated. Additional implementation studies, including usability testing, workflow integration, and prospective evaluation, are needed before broad deployment in routine practice. While our separate pilot trial, Developing and Evaluating a Machine Learning Opioid Prediction & Risk-Stratification E-Platform, is currently assessing the feasibility of integrating an EHR-based overdose alert into clinical workflows within UF Health primary care clinics [71,72], analogous testing would be required for the model developed in this study.

Conclusions

This study demonstrates that an ML algorithm derived from the OneFlorida+ network can accurately predict incident OUD and effectively stratify risk across diverse populations and health systems, as evidenced by successful external validation in UPMC. The next step is to prospectively validate this algorithm within the UF Health clinical workflow to leverage existing decision support infrastructure and facilitate early intervention. Future research should expand the model to incorporate unstructured clinical notes, prescription drug monitoring program data, and social drivers of health (eg, housing instability, employment status, and incarceration history) to further improve prediction accuracy and equity.

Funding

This project is funded by the US National Institute on Drug Abuse (R01DA050676 and R01DA044985). WLC and WFG are named as inventors in one preliminary patent (U1195.70174US00) filing from the University of Florida and University of Pittsburgh for use of a machine-learning algorithm for opioid risk prediction in Medicare. The views presented here are those of the authors alone and do not necessarily represent the views of the Department of Veterans Affairs or the United States Government.

Data Availability

The full modeling pipeline, including code for data preprocessing, training, validation, and fairness analysis, is publicly available in the GitHub repository [73]. The datasets generated or analyzed during this study are not publicly available due to patient privacy regulations, Health Insurance Portability and Accountability Act, but are available from the OneFlorida+ network upon reasonable request and institutional review board approval.

Authors' Contributions

Conceptualization: JAF, WL-C, WFG, YW, CAH, KN, JLH, GC, DLW, SASS, SOFS, EIR, DN, SY, GMR, WMG, CK, MMH
Data curation: JAF, JLH, WLC, and MMH
Formal analysis: JAF, JLH, WLC, and MMH
Funding acquisition: WLC
Investigation: All authors
Methodology: All authors
Project administration: WLC
Resources: JAF, JLH, WLC, and MMH
Software: JAF, WLC, and MMH
Supervision: WLC and MMH
Validation: JAF, JLH, WLC, and MMH
Visualization: JAF, WLC, and MMH
Writing – original draft: JAF and MMH
Writing – review and editing: All authors

XSL•FO
RenderX

Conflicts of Interest

DLW and WLC have received grant funding from Merck, Sharp & Dohme and Bristol Myers Squibb for unrelated research. WLC was compensated for consulting services by Teva Pharmaceuticals.

Multimedia Appendix 1
Study design, model performance (internal and external validation), calibration and decision curves, and risk stratification.
[DOCX File, 2662 KB - jmir_v28i1e79482_app1.docx]

Multimedia Appendix 2
Diagnosis codes, predictor definitions, and model performance.
[DOCX File, 85 KB - jmir_v28i1e79482_app2.docx]

Multimedia Appendix 3
Methods for machine learning model development and evaluation.
[DOCX File, 58 KB - jmir_v28i1e79482_app3.docx]

Checklist 1
TRIPOD+AI checklist.
[DOCX File, 47 KB - jmir_v28i1e79482_app4.docx]

References

1. Florence C, Luo F, Rice K. The economic burden of opioid use disorder and fatal opioid overdose in the United States, 2017. Drug Alcohol Depend 2021 Jan 1;218:108350. [doi: 10.1016/j.drugalcdep.2020.108350] [Medline: 33121867]
2. Hébert AH, Hill AL. Impact of opioid overdoses on US life expectancy and years of life lost, by demographic group and stimulant co-involvement: a mortality data analysis from 2019 to 2022. Lancet Reg Health Am 2024 Aug;36:100813. [doi: 10.1016/j.lana.2024.100813] [Medline: 38978785]
3. Leslie DL, Ba DM, Agbese E, Xing X, Liu G. The economic burden of the opioid epidemic on states: the case of Medicaid. Am J Manag Care 2019 Jul;25(13 Suppl):S243-S249. [Medline: 31361426]
4. Keyes KM, Rutherford C, Hamilton A, et al. What is the prevalence of and trend in opioid use disorder in the United States from 2010 to 2019? Using multiplier approaches to estimate prevalence for an unknown population size. Drug Alcohol Depend Rep 2022 Jun;3:100052. [doi: 10.1016/j.dadr.2022.100052] [Medline: 35783994]
5. About overdose prevention. Centers for Disease Control and Prevention (CDC). URL: https://www.cdc.gov/overdose-prevention/about/index.html [accessed 2024-05-04]
6. Kline D, Hepler SA, Krawczyk N, Rivera-Aguirre A, Waller LA, Cerdá M. A state-level history of opioid overdose deaths in the United States: 1999-2021. PLoS ONE 2024;19(9):e0309938. [doi: 10.1371/journal.pone.0309938] [Medline: 39240938]
7. Chen LH, Hedegaard H, Warner M. Drug-poisoning deaths involving opioid analgesics: United States, 1999-2011. NCHS Data Brief 2014 Sep(166):1-8. [Medline: 25228059]
8. Ruhm CJ. Corrected US opioid‐involved drug poisoning deaths and mortality rates, 1999–2015. Addiction 2018 Jul;113(7):1339-1344. [doi: 10.1111/add.14144] [Medline: 29430760]
9. Overdose prevention: about prescription opioids. Centers for Disease Control and Prevention (CDC). URL: https://www.cdc.gov/overdose-prevention/about/prescription-opioids.html [accessed 2024-10-14]
10. National Institute on Drug Abuse, NIDA. Only 1 in 5 US adults with opioid use disorder received medications to treat it in 2021. URL: https://nida.nih.gov/news-events/news-releases/2023/08/only-1-in-5-us-adults-with-opioid-use-disorder-received-medications-to-treat-it-in-2021 [accessed 2024-10-14]
11. Volkow ND, Blanco C. The changing opioid crisis: development, challenges and opportunities. Mol Psychiatry 2021 Jan;26(1):218-233. [doi: 10.1038/s41380-020-0661-4] [Medline: 32020048]
12. Lo-Ciganic WH, Donohue JM, Yang Q, et al. Developing and validating a machine-learning algorithm to predict opioid overdose in medicaid beneficiaries in two US states: a prognostic modelling study. Lancet Digit Health 2022 Jun;4(6):e455-e465. [doi: 10.1016/S2589-7500(22)00062-0] [Medline: 35623798]
13. Edlund MJ, Martin BC, Russo JE, DeVries A, Braden JB, Sullivan MD. The role of opioid prescription in incident opioid abuse and dependence among individuals with chronic noncancer pain: the role of opioid prescription. Clin J Pain 2014 Jul;30(7):557-564. [doi: 10.1097/AJP.0000000000000021] [Medline: 24281273]
14. Yang Z, Wilsey B, Bohm M, et al. Defining risk of prescription opioid overdose: pharmacy shopping and overlapping prescriptions among long-term opioid users in medicaid. J Pain 2015 May;16(5):445-453. [doi: 10.1016/j.jpain.2015.01.475] [Medline: 25681095]

XSL•FO
RenderX

15. Eguale T, Bastardot F, Song W, et al. A machine learning application to classify patients at differing levels of risk of opioid use disorder: clinician-based validation study. JMIR Med Inform 2024 Jun 4;12:e53625. [doi: 10.2196/53625] [Medline: 38842167]
16. Screener and opioid assessment for patients with pain-revised (SOAPP®-r). : Inflexxion, Inc; 2008 URL: https://ddph-materials.s3.amazonaws.com/HelpIsHere/SOAPP-Tool.pdf [accessed 2026-01-22]
17. Cheatle MD, Compton PA, Dhingra L, Wasser TE, O'Brien CP. Development of the revised opioid risk tool to predict opioid use disorder in patients with chronic nonmalignant pain. J Pain 2019 Jul;20(7):842-851. [doi: 10.1016/j.jpain.2019.01.011] [Medline: 30690168]
18. Chou R, Hartung D, Turner J, et al. Opioid Treatments for Chronic Pain: Agency for Healthcare Research and Quality (US); 2020. URL: https://www.ncbi.nlm.nih.gov/sites/books/NBK556253/ [accessed 2026-02-13]
19. Bamboo health. NarxCare. URL: https://bamboohealth.com/solutions/narxcare/ [accessed 2025-11-03]
20. Wei YJJ, Chen C, Sarayani A, Winterstein AG. Performance of the centers for medicare & medicaid services' opioid overutilization criteria for classifying opioid use disorder or overdose. JAMA 2019 Feb 12;321(6):609-611. [doi: 10.1001/jama.2018.20404] [Medline: 30747958]
21. Canan C, Polinski JM, Alexander GC, Kowal MK, Brennan TA, Shrank WH. Automatable algorithms to identify nonmedical opioid use using electronic data: a systematic review. J Am Med Inform Assoc 2017 Nov 1;24(6):1204-1210. [doi: 10.1093/jamia/ocx066] [Medline: 29016967]
22. Rough K, Huybrechts KF, Hernandez-Diaz S, Desai RJ, Patorno E, Bateman BT. Using prescription claims to detect aberrant behaviors with opioids: comparison and validation of 5 algorithms. Pharmacoepidemiol Drug Saf 2019 Jan;28(1):62-69. [doi: 10.1002/pds.4443] [Medline: 29687539]
23. Goyal H, Singla U, Grimsley EW. Identification of opioid abuse or dependence: no tool is perfect. Am J Med 2017 Mar;130(3):e113. [doi: 10.1016/j.amjmed.2016.09.022] [Medline: 28215952]
24. Wood E, Simel DL, Klimas J. Pain management with opioids in 2019-2020. JAMA 2019 Nov 19;322(19):1912-1913. [doi: 10.1001/jama.2019.15802] [Medline: 31600370]
25. Dong X, Deng J, Rashidian S, et al. Identifying risk of opioid use disorder for patients taking opioid medications with deep learning. J Am Med Inform Assoc 2021 Jul 30;28(8):1683-1693. [doi: 10.1093/jamia/ocab043] [Medline: 33930132]
26. Gao W, Leighton C, Chen Y, Jones J, Mistry P. Predicting opioid use disorder and associated risk factors in a Medicaid managed care population. Am J Manag Care 2021 Apr;27(4):148-154. [doi: 10.37765/ajmc.2021.88617] [Medline: 33877773]
27. Hasan MM, Young GJ, Patel MR, Modestino AS, Sanchez LD, Noor-E-Alam M. A machine learning framework to predict the risk of opioid use disorder. Machine Learning with Applications 2021 Dec;6:100144. [doi: 10.1016/j.mlwa.2021.100144]
28. Segal Z, Radinsky K, Elad G, et al. Development of a machine learning algorithm for early detection of opioid use disorder. Pharmacol Res Perspect 2020 Dec;8(6):e00669. [doi: 10.1002/prp2.669] [Medline: 33200572]
29. Lo-Ciganic WH, Huang JL, Zhang HH, et al. Using machine learning to predict risk of incident opioid use disorder among fee-for-service medicare beneficiaries: a prognostic study. PLoS ONE 2020;15(7):e0235981. [doi: 10.1371/journal.pone.0235981] [Medline: 32678860]
30. Hogan WR, Shenkman EA, Robinson T, et al. The OneFlorida data trust: a centralized, translational research data infrastructure of statewide scope. J Am Med Inform Assoc 2022 Mar 15;29(4):686-693. [doi: 10.1093/jamia/ocab221] [Medline: 34664656]
31. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024 Apr 16;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
32. Smith SM, Winterstein AG, Gurka MJ, et al. Initial antihypertensive regimens in newly treated patients: real world evidence from the OneFlorida+ clinical research network. J Am Heart Assoc 2023 Jan 3;12(1):e026652. [doi: 10.1161/JAHA.122.026652] [Medline: 36565195]
33. OneFlorida+ clinical research network. The OneFlorida+ Data Trust. URL: https://onefl.net/data/ [accessed 2025-01-07]
34. Ellis RJ, Wang Z, Genes N, Ma'ayan A. Predicting opioid dependence from electronic health records with machine learning. BioData Min 2019;12(1):3. [doi: 10.1186/s13040-019-0193-0] [Medline: 30728857]
35. Hastings JS, Howison M, Inman SE. Predicting high-risk opioid prescriptions before they are given. Proc Natl Acad Sci U S A 2020 Jan 28;117(4):1917-1923. [doi: 10.1073/pnas.1905355117] [Medline: 31937665]
36. Ciesielski T, Iyengar R, Bothra A, Tomala D, Cislo G, Gage BF. A tool to assess risk of de novo opioid abuse or dependence. Am J Med 2016 Jul;129(7):699-705. [doi: 10.1016/j.amjmed.2016.02.014] [Medline: 26968469]
37. Hall AJ, Logan JE, Toblin RL, et al. Patterns of abuse among unintentional pharmaceutical overdose fatalities. JAMA 2008 Dec 10;300(22):2613-2620. [doi: 10.1001/jama.2008.802] [Medline: 19066381]
38. White AG, Birnbaum HG, Schiller M, Tang J, Katz NP. Analytic models to identify patients at risk for prescription opioid abuse. Am J Manag Care 2009 Dec;15(12):897-906. [Medline: 20001171]
39. Sullivan MD, Edlund MJ, Fan MY, DeVries A, Braden JB, Martin BC. Risks for possible and probable opioid misuse among recipients of chronic opioid therapy in commercial and medicaid insurance plans: The TROUP Study. Pain 2010 Aug;150(2):332-339. [doi: 10.1016/j.pain.2010.05.020] [Medline: 20554392]

XSL•FO
RenderX

40. Cepeda MS, Fife D, Chow W, Mastrogiovanni G, Henderson SC. Assessing opioid shopping behaviour: a large cohort study from a medication dispensing database in the US. Drug Saf 2012 Apr 1;35(4):325-334. [doi: 10.2165/11596600-000000000-00000] [Medline: 22339505]
41. Cochran G, Gordon AJ, Lo-Ciganic WH, et al. An examination of claims-based predictors of overdose from a large medicaid program. Med Care 2017 Mar;55(3):291-298. [doi: 10.1097/MLR.0000000000000676] [Medline: 27984346]
42. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 2015;10(3):e0118432. [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
43. Romero-Brufau S, Huddleston JM, Escobar GJ, Liebow M. Why the C-statistic is not informative to evaluate early warning scores and what metrics to use. Crit Care 2015 Aug 13;19(1):285. [doi: 10.1186/s13054-015-0999-1] [Medline: 26268570]
44. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B 2005 Apr 1;67(2):301-320. [doi: 10.1111/j.1467-9868.2005.00503.x]
45. Larochelle MR, Zhang F, Ross-Degnan D, Wharam JF. Rates of opioid dispensing and overdose after introduction of abuse-deterrent extended-release oxycodone and withdrawal of propoxyphene. JAMA Intern Med 2015 Jun;175(6):978-987. [doi: 10.1001/jamainternmed.2015.0914] [Medline: 25895077]
46. Chu A, Ahn H, Halwan B, et al. A decision support system to facilitate management of patients with acute gastrointestinal bleeding. Artif Intell Med 2008 Mar;42(3):247-259. [doi: 10.1016/j.artmed.2007.10.003] [Medline: 18063351]
47. Ives TJ, Chelminski PR, Hammett-Stabler CA, et al. Predictors of opioid misuse in patients with chronic pain: a prospective cohort study. BMC Health Serv Res 2006 Apr 4;6(1):46. [doi: 10.1186/1472-6963-6-46] [Medline: 16595013]
48. Becker WC, Sullivan LE, Tetrault JM, Desai RA, Fiellin DA. Non-medical use, abuse and dependence on prescription opioids among U.S. adults: psychiatric, medical and substance use correlates. Drug Alcohol Depend 2008 Apr 1;94(1-3):38-47. [doi: 10.1016/j.drugalcdep.2007.09.018] [Medline: 18063321]
49. Dowell D, Haegerich TM, Chou R. CDC guideline for prescribing opioids for chronic pain — United States, 2016. MMWR Recomm Rep 2023;65(1):1-49. [doi: 10.15585/mmwr.rr6501e1] [Medline: 26987082]
50. UK NGC, Others. Evidence review: risk factors for dependence. : National Institute for Health and Care Excellence (NICE); 2022 URL: https://www.ncbi.nlm.nih.gov/books/NBK580678/ [accessed 2026-02-13]
51. How opioid use disorder occurs. Mayo Clinic. URL: https://www.mayoclinic.org/diseases-conditions/prescription-drug-abuse/in-depth/how-opioid-addiction-occurs/art-20360372 [accessed 2025-10-31]
52. Guidelines for prescribing controlled substances for pain. Medical Board of California. URL: https://www.mbc.ca.gov/Resources/Medical-Resources/controlled-substance.aspx [accessed 2025-11-03]
53. Ling W, Mooney L, Hillhouse M. Prescription opioid abuse, pain and addiction: clinical issues and implications. Drug Alcohol Rev 2011 May;30(3):300-305. [doi: 10.1111/j.1465-3362.2010.00271.x] [Medline: 21545561]
54. Committee on Pain Management and Regulatory Strategies to Address Prescription Opioid Abuse. Pain Management and the Opioid Epidemic: Balancing Societal and Individual Benefits and Risks of Prescription Opioid Use: National Academies Press; 2017. [doi: 10.17226/24781]
55. Lo-Ciganic WH, Huang JL, Zhang HH, et al. Evaluation of machine-learning algorithms for predicting opioid overdose risk among medicare beneficiaries with opioid prescriptions. JAMA Netw Open 2019 Mar 1;2(3):e190968. [doi: 10.1001/jamanetworkopen.2019.0968] [Medline: 30901048]
56. Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction 2009. [doi: 10.1198/jasa.2004.s339]
57. Al Faysal J, Noor-E-Alam M, Young GJ, et al. An explainable machine learning framework for predicting the risk of buprenorphine treatment discontinuation for opioid use disorder among commercially insured individuals. Comput Biol Med 2024 Jul;177:108493. [doi: 10.1016/j.compbiomed.2024.108493] [Medline: 38833799]
58. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988 Sep;44(3):837-845. [doi: 10.2307/2531595] [Medline: 3203132]
59. Data Mining and Statistics for Decision Making: Wiley; 2011. [doi: 10.1002/9780470979174]
60. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. Biom J 2005 Aug;47(4):458-472. [doi: 10.1002/bimj.200410135] [Medline: 16161804]
61. Nohara Y, Matsumoto K, Soejima H, Nakashima N. Explanation of machine learning models using improved shapley additive explanation. 2019 Sep 4 Presented at: BCB '19: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics; Sep 7-10, 2019; Niagara Falls NY USA p. 546. [doi: 10.1145/3307339.3343255]
62. Saleiro P, Kuester B, Hinkson L, et al. Aequitas: a bias and fairness audit toolkit. arXiv. Preprint posted online on Nov 14, 2018. [doi: 10.48550/arXiv.1811.05577]
63. Jiang X, Osl M, Kim J, Ohno-Machado L. Smooth isotonic regression: a new method to calibrate predictive models. AMIA Jt Summits Transl Sci Proc 2011:16-20. [Medline: 22211175]
64. Rufibach K. Use of Brier score to assess binary predictions. J Clin Epidemiol 2010 Aug;63(8):938-939. [doi: 10.1016/j.jclinepi.2009.11.009] [Medline: 20189763]
65. Peirce CS. The numerical measure of the success of predictions. Science 1884 Nov 14;4(93):453-454. [doi: 10.1126/science.ns-4.93.453-a] [Medline: 17795531]

XSL•FO
RenderX

66. Dufour R, Mardekian J, Pasquale MK, Schaaf D, Andrews GA, Patel NC. Understanding predictors of opioid abuse: predictive model development and validation. Am J Pharm Benefits 2014;6(5):208-216 [FREE Full text]
67. Webster LR, Webster RM. Predicting aberrant behaviors in opioid-treated patients: preliminary validation of the opioid risk tool. Pain Med 2005;6(6):432-442. [doi: 10.1111/j.1526-4637.2005.00072.x] [Medline: 16336480]
68. Rice JB, White AG, Birnbaum HG, Schiller M, Brown DA, Roland CL. A model to identify patients at risk for prescription opioid abuse, dependence, and misuse. Pain Med 2012 Sep;13(9):1162-1173. [doi: 10.1111/j.1526-4637.2012.01450.x] [Medline: 22845054]
69. Hylan TR, Von Korff M, Saunders K, et al. Automated prediction of risk for problem opioid use in a primary care setting. J Pain 2015 Apr;16(4):380-387. [doi: 10.1016/j.jpain.2015.01.011] [Medline: 25640294]
70. Kerr KF, Brown MD, Zhu K, Janes H. Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. J Clin Oncol 2016 Jul 20;34(21):2534-2540. [doi: 10.1200/JCO.2015.65.5654] [Medline: 27247223]
71. Nguyen K, Wilson DL, Diiulio J, et al. Design and development of a machine-learning-driven opioid overdose risk prediction tool integrated in electronic health records in primary care settings. Bioelectron Med 2024 Oct 18;10(1):24. [doi: 10.1186/s42234-024-00156-3] [Medline: 39420438]
72. Hong JWJ, Wilson DL, Nguyen K, et al. Protocol for a single-arm pilot clinical trial: developing and evaluating a machine learning opioid prediction & risk-stratification E-platform (DEMONSTRATE). J Clin Med 2025 Dec 1;14(23):8522. [doi: 10.3390/jcm14238522] [Medline: 41375825]
73. Incident-OUD-project. GitHub. URL: https://github.com/faysal22/Incident-OUD-Project [accessed 2026-01-24]

Abbreviations

CDS: clinical decision support
EHR: electronic health record
EN: elastic net
FNR: false negative rate
FPR: false positive rate
GBM: gradient boosting machine
***ICD-10*:** *International Statistical Classification of Diseases and Related Health Problems 10th Revision*
LASSO: least absolute shrinkage and selection operator
ML: machine learning
NNE: number needed to evaluate
NPV: negative predictive value
OUD: opioid use disorder
PPV: positive predictive value
RF: random forest
SHAP: Shapley additive explanations
TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+Artificial Intelligence
UF: University of Florida
UPMC: University of Pittsburgh Medical Center

Edited by J Sarvestan; submitted 22.Jun.2025; peer-reviewed by J Song, MA Roshani; revised version received 13.Dec.2025; accepted 15.Dec.2025; published 05.Mar.2026.

Please cite as:
Faysal JA, Lo-Ciganic W, Gellad WF, Wu Y, Harle CA, Nguyen K, Huang JL, Cochran G, Wilson DL, Staras SAS, Schmidt SOF, Rosenberg EI, Nelson D, Yan S, Reisfield GM, Greene WM, Kuza C, Hasan MM
Developing and Validating a Machine Learning Algorithm to Predict the Risk of Incident Opioid Use Disorder Among OneFlorida+ Patients: Prognostic Modeling Study
J Med Internet Res 2026;28:e79482
URL: https://www.jmir.org/2026/1/e79482
doi:10.2196/79482

© Jabed Al Faysal, Weihsuan Lo-Ciganic, Walid F Gellad, Yonghui Wu, Christopher A Harle, Khoa Nguyen, James L Huang, Gerald Cochran, Debbie L Wilson, Stephanie AS Staras, Siegfried OF Schmidt, Eric I Rosenberg, Danielle Nelson, Shunhua Yan, Gary M Reisfield, William M Greene, Courtney Kuza, Md Mahmudul Hasan. Originally published in the Journal of Medical

XSL•FO
RenderX

Internet Research (https://www.jmir.org), 5.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Breast Cancer Screening Knowledge and Sentiments in Singaporean Women: Mixed Methods Study Using Topic Modeling, Sentiment Analysis, and Structured Questionnaire Data

Peh Joo Ho[1,2,3], PhD; Zi Lin Lim[3], BSc (Hons); Jenny Liu[1,4], MSc; Nur Khaliesah Mohamed Riza[1,5], BSc; Ying Jia Chew[5], BSc (Hons); Yi Ying Lim[1], BSc (Hons); Hui Ling Tan[4], BMgmt (Hons); Su-Ann Goh[2], PhD; Han Boon Oh[4,6], MMed; Chi Hui Chin[7], MMed; Sing Cheer Kwek[8], MMed; Zhi Peng Zhang[9], MMed; Desmond Luan Seng Ong[7], MMed; Swee Tian Quek[10], MBBS; Sujith Wijerathne[5,6], MMed; Philip Tsau Choong Iau[4,5*], MD, PhD; Mikael Hartman[1,2,5*], PhD; Jingmei Li[2,3,11*], PhD

[1]Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore, Singapore

[10]Department of Diagnostic Imaging, National University Hospital and National University Health System, Singapore, Singapore

[11]National Cancer Centre Singapore, Singapore Health Services (SingHealth), Singapore, Singapore

[2]Department of Surgery, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore

[3]Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR), #02-01, 60 Biopolis Street, Singapore, Singapore

[4]Department of General Surgery, Ng Teng Fong General Hospital, Singapore, Singapore

[5]Department of Surgery, National University Hospital and National University Health System, Singapore, Singapore

[6]Department of General Surgery, Jurong Medical Centre, Singapore, Singapore

[7]Jurong Polyclinic, National University Polyclinics and National University Health System, Singapore, Singapore

[8]Bukit Batok Polyclinic, National University Polyclinics and National University Health System, Singapore, Singapore

[9]Choa Chu Kang Polyclinic, National University Polyclinics and National University Health System, Singapore, Singapore

[*]these authors contributed equally

Corresponding Author:
Jingmei Li, PhD
Department of Surgery, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore

Abstract

Background: Mammography screening uptake in Singapore remains below 40% despite campaigns and subsidies. Natural language processing (NLP) can extract nuanced attitudes from free text that fixed response options miss, revealing latent factors influencing breast cancer (BC) screening behavior.

Objective: This study characterized women's attitudes toward mammography using mixed methods data, examined associations between BC awareness and screening willingness, and identified barriers and facilitators through NLP of free-text responses.

Methods: We conducted a cross-sectional study within the *Breast Screening Tailored for Her* multicenter cohort in Singapore (October 2021-December 2023). In total, 4169 women aged 35 - 59 years (median 48, IQR 43 - 54) were recruited via convenience sampling (3 hospitals and 2 polyclinics). Participants completed online structured questionnaires on demographics and screening history, then a BC education quiz with feedback. Participants answering >80% correctly were classified as "BC-aware." Posteducation, participants reported screening willingness (motivated or neutral) with optional free-text explanations. Logistic regression models (adjusted for study site, age, ethnicity, marital status, housing, and education) examined the associations with willingness. For 3819 English-language respondents, biterm topic modeling identified themes and sentiment analysis quantified emotional tone. Statistical significance: α=.05.

Results: Overall, 79% (3287/4169) were BC-aware, and 94% (3908/4169) reported increased motivation posteducation. BC-aware women had higher screening motivation than BC-unaware women (adjusted odds ratio [aOR] 2.88, 95% CI 2.19 - 3.80; *P*<.001). Motivation was higher among those with larger public housing (OR 1.81, 95% CI 1.30 - 2.50; *P*<.001) and private housing vs 1 - 3 room units (OR 2.69, 95% CI 1.75 - 4.13; *P*<.001), married vs not separated, divorced, or widowed (OR 2.38 [inverse of 0.42], 95% CI 1.75 - 3.13; *P*<.001), and prior screening attendance (OR 3.49, 95% CI 2.71 - 4.50; *P*<.001). Women who disagreed that mammography was expensive had higher motivation (aOR 1.94, 95% CI 1.50 - 2.50; *P*<.001). Among 3819 English respondents, 94% (3579/3819) were motivated and 6% (240/3819) neutral. Free-text responses came from 34% (1220/3579) of motivated and 64% (153/240) of neutral participants. Biterm topic modeling revealed motivated participants emphasized early

XSL•FO
RenderX

detection benefits, health awareness, BC risk, and logistics; neutral participants focused on mammography pain experiences and cost barriers. Mean sentiment was 0.207 (range: −1.00 to 1.65), with motivated participants displaying more positive sentiments than neutral participants (linear regression, $P<.001$). Identical words carried different emotional tones across subgroups: "health" had positive sentiment among motivated participants (mean difference, Welch t tests $P<.05$) but negative sentiment among neutral participants. Word frequency analysis showed motivated participants used positive-sentiment words ("better," "cure," and "prevention"). Neutral participants emphasized negative words ("painful" and "uncomfortable").

Conclusions: Integrating quantitative surveys with NLP revealed that the same screening concepts are emotionally framed differently by motivated vs neutral women, a finding missed by knowledge- or intent-focused approaches alone. In practice, these findings support the need for emotionally tailored BC education and prevention strategies.

(J Med Internet Res 2026;28:e78439) doi:10.2196/78439

KEYWORDS

mammography screening; breast cancer; Asian population; screening attitudes; breast cancer awareness; precision health; proactive health

Introduction

Mammography screening has been widely regarded as the most effective method for achieving early detection of breast cancer (BC), leading to improved survival [1-3]. Despite a decade of annual campaigns and various subsidies, the mammography screening rate remains suboptimal, even among women who know about mammogram services in Singapore [4,5].

A review by Momenimovahed et al [6] found that barriers to mammography screening in Asia include factors such as financial limitations, lack of social support, fatalistic tendencies, fear of pain and embarrassment, and religion. Barriers to mammography screening specific to Singapore have also been widely examined in multiple studies, and the common obstacles reflected are the lack of time, fatalism, misconceptions about mammography screening or BC, reluctance to spend on screening, and religious reasons [7-12]. In a review by Rajendram et al [4], the perceived costs of screening played the biggest role in hindering screening participation among Singaporean women.

According to the principle of attitude consistency, a person's attitudes tend to influence and shape their actions and behavior [13]. Common techniques used to understand people's opinions and attitudes toward screening include quantitative tools, such as the Likert scale, or qualitative tools, such as focus groups or in-depth interviews. However, these tools are limited by a few factors. Self-reported measurements using the Likert scale often present participants with closed-ended scenarios. The finite number of response options can limit respondents' ability to express nuanced opinions, and responses can be prone to central tendency bias [14,15]. On the other hand, while focus groups and in-depth interviews offer valuable qualitative data, providing context and information on nuances of responses, they are limited by factors such as response bias [16], researcher bias, and time-intensive analysis processes [17].

In recent years, natural language processing (NLP) has emerged as a novel tool in health behavior and health services research, offering unique advantages in understanding individuals' attitudes and perceptions [18,19]. By leveraging machine learning algorithms, NLP can sift through vast amounts of textual information to extract nuanced insights [20,21]. This approach not only enables researchers to capture the rich complexity of human language but also serves as an avenue to reveal implicit sentiments and identify emerging themes. Subsequently, the use of NLP can offer a more comprehensive understanding of the factors influencing the uptake of health behaviors, such as screening.

Currently, studies involving the use of NLP tools in understanding health-seeking behavior largely revolve around social media posts [22]. A systematic review by Döbrössy et al [23] demonstrated the value of social media as a platform to spread BC awareness and the value of NLP tools in identifying common areas of concern of the lay public about BC. NLP's ability to analyze free text provides valuable insights into the psychological and social factors influencing health behaviors, potentially contributing to more effective public health interventions and communication strategies. This study leverages structured questionnaire data supplemented by NLP tools to analyze free-text responses, aiming to (1) characterize women's attitudes toward mammography, (2) examine the relationship between BC awareness and screening willingness, and (3) identify barriers and facilitators for the willingness to screen.

Methods

Study Design

We conducted a cross-sectional, observational, mixed methods study combining quantitative surveys and qualitative free-text analysis to examine women's attitudes and willingness to attend routine mammography screening. This study was reported following the American Psychological Association Journal Article Reporting Standards for mixed-methods, quantitative, and qualitative research [24].

Setting

The Breast Screening Tailored for Her (BREATHE) study is a multicenter prospective cohort study in Singapore [25,26]. Participants were recruited between October 2021 and December 2023 from 3 restructured hospitals—Ng Teng Fong General Hospital, National University Hospital, and Alexandra Hospital—2 polyclinics—Bukit Batok Polyclinic and Choa Chu Kang Polyclinic—and Jurong Medical Center [25]. Briefly, potential participants were identified through advertisements

XSL•FO
RenderX

(eg, posters, flyers, and blog recruitment pages inviting women to register interest) and an active approach at the clinical sites (ie, study staff approached women in waiting areas of the participating hospitals and polyclinics). Recruitment was based on convenience (ie, convenience sampling) and recruitment feasibility across participating clinical sites rather than a formal a priori power calculation.

Eligible participants were female, aged between 35 and 59 years, and Singapore citizens or permanent residents at the point of enrollment. Participants were excluded if they had a histologically confirmed diagnosis of any cancer, were cognitively impaired, or were pregnant during recruitment. Consistent with the subsidized Screen for Life BC screening criteria in Singapore, women with benign breast disease were eligible for the study, while those with breast implants were excluded. After providing informed consent, participants completed an online structured first-visit questionnaire, which included factors linked to BC and related conditions, such as demographic, lifestyle, and reproductive characteristics, prior treatments, and other environmental exposures (Multimedia Appendix 1). They then undertook a brief, self-administered education session online covering BC knowledge and the importance of regular breast self-examination and screening (Multimedia Appendix 2). Participants answered the online education questionnaire in either English, Chinese, or Malay.

A flowchart illustrating the selection of participants from those enrolled in the BREATHE study is shown in Figure S1 in Multimedia Appendix 3. Between October 2021 and December 2023, 4592 participants were enrolled in the study (corresponding to a participation rate of 83% among the 5536 women approached or expressing interest). Participants were excluded if they withdrew from the study (n=74), were diagnosed with BC within 6 months of enrollment (n=17), had unknown past mammography (n=38), or had incoherent BC screening and mammography screening attendance based on the lifestyle questionnaire and education session (n=294). The analytical cohort comprised 4169 participants. The free-text analysis included participants who answered the English questionnaire (n=3819).

Demographics and BC Risk Factors

Baseline information on demographics and BC risk factors was obtained from the first-visit questionnaire. The variables included ethnicity, medical history, previous benign lump or gynecological surgery (yes, no, or missing), family history of breast and ovarian cancer (yes, no, or missing), reproductive factors, and previous breast examination and screening habits, among others.

Sociodemographic factors were derived from the same questionnaire, where individual factors were further categorized for ease of analysis (Table S1 in Multimedia Appendix 4). Housing (public: 1 - 3 room flat, public: >3-room flat [4-, 5-, or executive-type], or private), highest qualification achieved (no formal or primary, secondary, postsecondary [nontertiary], professional diploma, or tertiary), and marital status (married, never married, widowed, or separated or divorced) were used as proxies for economic, education, and social support status, respectively.

BC Education

Participants' existing screening habits and views about BC were assessed during an online education session (Multimedia Appendix 2). Various statements were used to assess BC awareness and provide education (questions 7 - 13, 16, and 18). These statements were presented to participants to indicate their agreement (agree or disagree). The correct answer and an accompanying explanation were given after every response. Participants were characterized as aware of BC risk (ie, BC-aware) before study enrollment if they answered >80% of the questions correctly.

Following BC education, participants were asked about their willingness to attend regular screening (question 20: "After knowing the above, would you be more willing to attend regular screening?" yes or no) and were categorized as "motivated" or "neutral," respectively. Participants were also asked to explain their choice (question 21: "Referring to your answer in the last question, please provide a reason." free text, optional).

Statements unrelated to BC or BC screening knowledge were considered as BC perceptions for analysis (ie, perceived importance of BC screening, perceived risk of BC, fatalistic attitudes, and finding mammography embarrassing, expensive, inconvenient, or painful).

Data Analysis

Descriptive analyses were conducted to characterize demographic factors and BC screening attitudes. The associations between willingness to attend more regular screening and the various characteristics were studied using the chi-square and Kruskal-Wallis tests for categorical and continuous variables, respectively.

The study's primary outcome was self-reported willingness to screen, measured immediately after an education intervention. The association between the various factors and participants' willingness to attend regular screening was studied using logistic regression models. In addition to univariate analysis, the models were adjusted for study site and participant characteristics (age category, ethnicity, marital status, housing type, and highest academic status). To select the best predictive model, stepwise forward selection was used (ie, the lowest Bayesian information criterion). Odds ratios (ORs) and 95% CIs are presented.

Free-Text Data Preprocessing

The sentiment of words may not be well captured through translation. We thus retained only the reasons provided in English in the sentiment analysis. Of the 4169 participants, 3819 (92%) completed the English version of the education session. Participants (n=350) who completed the Chinese or Malay version of the education session were excluded from text analysis. Free-text analysis was done to assess participants' sentiments toward regular screening. Each participant's response was tokenized (R package *tidytext;* version 0.4.1, Julia Silge and David Robinson) [27], and spelling errors were corrected (R package *hunspell;* version 3.0.4, Jeroen Ooms). Additional spelling errors and corrections were added to the list generated. The full spelling errors and corrections can be found in Table S2 in Multimedia Appendix 4. The participants were categorized

XSL•FO
RenderX

by BC awareness (BC-aware vs BC-unaware) and their willingness to attend regular screening in the future posteducation session (motivated vs neutral; Figure S1 in Multimedia Appendix 3).

Biterm Topic Modeling

Biterm topic modeling (R package *BTM*; version 0.3.7, Jan Wijffels) was used to identify the top topics mentioned by the participants [28]. Although there are other topic modeling methods, such as latent Dirichlet allocation, latent semantic analysis, nonnegative matrix factorization, and transformer-based approaches such as BERTopic, biterm topic modeling was chosen due to the short length of the text inputs (1-3 sentences), its computational efficiency, and the interpretability of the resulting topics [29,30]. As the choice of the number of topics in the topic model will impact the effect of the model, we compared the results of the models with 2-5 topics (ie, parameter k). Expecting a small number of topics within each participant's response, we set α=.01 and β=.01. The clustering of words into topics depends on the initialization step and may change with different iterations. We attempted 6 iterations for each k, using set.seed() and permutations of 123 (ie, 123, 132, 213, 231, 312, and 321). The choice of the number of topics (k) was based on the stability of the topics produced across iterations, and the topic for each set of words was derived by reviewing the comments for relevant words and discussing them with team members. Biterm topic modeling looks at pairs of words and the probability of either of the words being in a particular topic. Topics emerged from the data rather than being predetermined, with the biterm topic modeling algorithmically identifying word co-occurrence patterns without a priori coding categories. This requires word stemming (R package *hunspell*; version 3.0.4) to prevent words with the same meaning from being tokenized into different tokens (eg, the words "prevented" and "prevents" have the same stem word "prevent," but not "prevention") and removing stop words from the text corpus. The list of stop words was derived from "stop_words" in tidytext and was further edited to suit this study. The full list of stem words and stop words is provided in Tables S3 and S4 in Multimedia Appendix 4, respectively.

Sentiment Analysis

In our study, "sentiment" refers to the emotional tone of participants' free-text responses as classified by NLP algorithms and used strictly as a quantitative analytical construct rather than a synonym for perception, attitude, or feeling. Unlike biterm modeling, which uses all pairs of words, sentiment analysis analyzes sentences or single words. Here, sentiment analysis was done using the R package *sentimentr;* version 2.9.0, Tyler W Rinker), as it accounts for valence shifters to determine the sentiment at the sentence level, while estimating the sentiment score by referencing the Jockers sentiment dictionary. Because these free-text responses were brief (1-3 sentences) and expressed lay perspectives and emotions about mammography screening rather than technical medical terminology, the use of a general-purpose sentiment lexicon with minor modifications was appropriate and ensured consistent capture of participants' attitudes across the cohort. The package scores sentiments on a scale where 0 is considered neutral, negative numbers indicate the presence of negative sentiments, and positive numbers indicate the presence of positive sentiments. Sentiment rating of certain words was further modified to suit the context of the study (Table S5 in Multimedia Appendix 4). Free-text responses to question 21 in Multimedia Appendix 2 were first analyzed by sentence and then averaged across sentences for each participant. Linear regression was used to determine significant differences in sentiment ratings across subgroups.

Word clouds (R package *ggplot2;* version 3.4.4, Hadley Wickham) were used to tabulate the top sentiment words of each subgroup. The larger the word in the visual, the more common the word was in the subgroup. To further explore patterns, common sentiment words, and the average sentiment associated were compared across subgroups using the Welch 2-sample *t* test. Unique sentiment words were also extracted. Example quotes of common and unique words were extracted randomly to provide better context.

Interpretive Limitations

The research team comprised epidemiologists, public health researchers, and data scientists with expertise in cancer screening behavior and computational text analysis. The use of algorithmic biterm topic modeling and sentiment analysis minimizes researcher interpretation bias. Researchers' perspectives influenced only the selection of analytical methods and the labeling of algorithmically identified topic clusters. Rather than sampling to thematic saturation as in traditional qualitative research, we analyzed all available free-text responses. Data collection concluded at the end of the study recruitment period (December 2023) rather than when thematic saturation was achieved. The stability of topics identified across 6 biterm topic modeling iterations with different random seeds provides evidence of thematic robustness within our dataset. The relatively brief nature of responses (1 - 3 sentences) and focus on specific questions means our analysis captures targeted attitudes toward screening rather than comprehensive life narratives, which is appropriate for our research objectives.

Ethical Considerations

This study involved human participants and complied with all relevant institutional and national research ethics guidelines. Ethics approval was obtained from the National Healthcare Group Domain-Specific Review Board, Singapore (reference no 2020/01327; approval date: June 7, 2021). Written informed consent was obtained from participants by trained study coordinators in the participant's preferred language (English, Chinese, or Malay). The informed consent process included permission to use the study data for secondary analyses relevant to BC screening research; therefore, no additional consent was required for the current analysis. Participant privacy and confidentiality were safeguarded throughout the study. All research data were deidentified before analysis, stored on secure servers with restricted access, and handled in accordance with institutional data protection policies to ensure anonymity. No individual participants can be identified in any image in the paper or multimedia appendices.

XSL•FO
RenderX

Results

Participant Characteristics

Table 1 shows the descriptive statistics of participant characteristics. Of the 4169 participants included, 3908 (94%) indicated that they were more motivated to attend regular screening posteducation sessions. The median age of the participants was 48 years (IQR 43-54), and the majority (77%, n=3208) of the participants were of Chinese ethnicity. Overall, 79% (n=3287) of the participants were BC-aware, and 22% (n=938) and 41% (n=1690) attended BC screening once every year or once every 2 years, respectively. Seventy-three percent (n=3056) had done at least 1 mammogram in the past. While 44% (n=1845) perceived themselves as low risk and 47% (n=1944) as average risk, 96% (n=4023) of the participants agreed that BC screening is important. Twelve percent (n=485) of the participants reported a history of benign breast disease.

XSL•FO
RenderX

Table . Characteristics of the BREAst Screening Tailored for HEr (BREATHE) study population, by their willingness to attend screening. Participants were recruited between 2021 and 2023 in Singapore's health care institutes. Two-sided *P* values for categorical variables are based on the chi-square test, and *P* values for continuous variables are based on the Kruskal-Wallis test. Counts and column percentages are presented, unless otherwise stated.

Characteristic	Total, n=4169, n (%)	Neutral, n=261, n (%)	Motivated, n=3908, n (%)	*P* value
BC[a] awareness				<.001
BC-unaware	882 (21)	104 (40)	778 (20)	
BC-aware	3287 (79)	157 (60)	3130 (80)	
Age, median (IQR)	48 (43-53)	49 (44-53)	48 (42-53)	.45
Age category (years)				.33
35-39	531 (13)	35 (13)	496 (13)	
40-49	1846 (44)	104 (40)	1742 (45)	
50-59	1792 (43)	122 (47)	1670 (43)	
Participant characteristic				
Ethnicity				.96
Chinese	3208 (77)	204 (78)	3004 (77)	
Malay	465 (11)	28 (11)	437 (11)	
Indian	279 (7)	17 (7)	262 (7)	
Other	217 (5)	12 (5)	205 (5)	
Marital status				<.001
Married	3186 (76)	164 (63)	3022 (77)	
Separated, divorced, or widowed	349 (8)	25 (10)	324 (8)	
Never married	634 (15)	72 (28)	562 (14)	
Housing				<.001
Public 1-3 room	490 (12)	52 (20)	438 (11)	
Public >3 room	2757 (66)	170 (65)	2587 (66)	
Private	922 (22)	39 (15)	883 (23)	
Highest qualification attained				.06
No formal or primary	235 (6)	23 (9)	212 (5)	
Secondary	787 (19)	46 (18)	741 (19)	
Postsecondary	1240 (30)	66 (25)	1174 (30)	
Tertiary	1907 (46)	126 (48)	1781 (46)	
Self-reported breast cancer history, existing behavior, and perceptions				
Family history of BC				.94
No	3756 (90)	236 (90)	3520 (90)	
Yes	413 (10)	25 (10)	388 (10)	
Benign breast disease				.25
No	3673 (88)	235 (90)	3438 (88)	
Yes	485 (12)	24 (9)	461 (12)	
Unknown	11 (0)	2 (1)	9 (0)	
BCS[b] attendance				<.001
No	1113 (27)	140 (54)	973 (25)	
Yes	3056 (73)	121 (46)	2935 (75)	

XSL•FO
RenderX

Characteristic	Total, n=4169, n (%)	Neutral, n=261, n (%)	Motivated, n=3908, n (%)	*P* value
Mammography attendance				<.001
No	1113 (27)	140 (54)	973 (25)	
Yes	3056 (73)	121 (46)	2935 (75)	
Previous BCS type				<.001
Mammography	3056 (73)	121 (46)	2935 (75)	
Ultrasound only	94 (2)	7 (3)	87 (2)	
Neither	1019 (24)	133 (51)	886 (23)	
Current BCS behavior				<.001
Once a year	938 (22)	11 (4)	927 (24)	
Once every 2 years	1690 (41)	32 (12)	1658 (42)	
Do not intend to attend in the future	99 (2)	43 (16)	56 (1)	
Other	320 (8)	35 (13)	285 (7)	
Not in the above categories and aged 35-39 years	506 (12)	35 (13)	471 (12)	
Unknown	616 (15)	105 (40)	511 (13)	
Perceived importance of BCS				<.001
Agree	4023 (96)	198 (76)	3825 (98)	
Neutral	136 (3)	57 (22)	79 (2)	
Disagree	10 (0)	6 (2)	4 (0)	
Perceived risk of BC				.03
Low	1845 (44)	133 (51)	1712 (44)	
Average	1944 (47)	101 (39)	1843 (47)	
High	380 (9)	27 (10)	353 (9)	
Fatalism toward developing BC				.85
Low	2002 (48)	121 (46)	1881 (48)	
Average	1179 (28)	77 (30)	1102 (28)	
High	988 (24)	63 (24)	925 (24)	
Fatalism toward dying from BC				.22
Low	2456 (59)	153 (59)	2303 (59)	
Average	1023 (25)	73 (28)	950 (24)	
High	690 (17)	35 (13)	655 (17)	
Finds mammography: embarrassing				<.001
Agree	586 (14)	74 (28)	512 (13)	
Disagree	3583 (86)	187 (72)	3396 (87)	
Expensive				<.001
Agree	1663 (40)	144 (55)	1519 (39)	
Disagree	2506 (60)	117 (45)	2389 (61)	
Inconvenient				<.001
Agree	866 (21)	105 (40)	761 (19)	

XSL•FO
RenderX

Characteristic	Total, n=4169, n (%)	Neutral, n=261, n (%)	Motivated, n=3908, n (%)	*P* value
Disagree	2779 (67)	120 (46)	2659 (68)	
Not applicable	524 (13)	36 (14)	488 (12)	
Painful				<.001
Agree	1027 (25)	134 (51)	893 (23)	
Disagree	2618 (63)	91 (35)	2527 (65)	
Not applicable	524 (13)	36 (14)	488 (12)	

[a]BC: breast cancer.

[b]BCS: breast cancer screening.

BC Awareness Is Significantly Associated With Posteducation Screening Motivation

In univariate logistic models, more motivated participants were more likely to be BC aware (OR 2.67, 95% $CI_{\text{aware vs unaware}}$ 2.05-3.46; *P*<.001), residing in >3 rooms HDB ($OR_{\text{>3 rooms HDB vs 1-3 rooms HDB}}$1.81, 95% CI 1.30-2.50; *P*<.001) or private housing ($OR_{\text{private vs 1-3 rooms HDB}}$ 2.69, 95% CI 1.75-4.13; *P*<.001), currently married ($OR_{\text{separated/divorced/widowed vs currently married}}$ 0.42, 95% CI 0.32-0.57; *P*<.001), having attended BC screening before ($OR_{\text{yes vs no}}$ 3.49, 95% CI 2.71-4.50; *P*<.001), being currently regular screeners, and being less likely to find the procedure embarrassing, expensive, inconvenient, or painful (Table 2). The complete results table of univariate analysis and the results adjusting for participants' characteristics is provided in Table S6 in Multimedia Appendix 4.

XSL•FO
RenderX

Table . Associations between patient characteristics and willingness to attend regular screening posteducation. Stepwise forward selection was used to select the best predictive model (ie, the lowest Bayesian information criterion [BIC]).

Characteristic	Univariate		Best prediction model by BIC	
	OR (95% CI)	*P* value	OR (95% CI)	*P* value
BC[a] awareness		<.001		<.001
BC-unaware	1.00 (referent)		1.00 (referent)	
BC-aware	2.67 (2.05-3.46)		1.90 (1.41-2.56)	
Age category (years)				
35-39	1.00 (referent)	—[b]	—	—
40-49	1.18 (0.80-1.76)	.41	—	—
50-59	0.97 (0.65-1.43)	.86	—	—
Ethnicity				
Chinese	1.00 (referent)	—	—	—
Malay	1.06 (0.71-1.59)	.78	—	—
Indian	1.05 (0.63-1.74)	.86	—	—
Other	1.16 (0.64-2.11)	.63	—	—
Marital status				
Married	1.00 (referent)	—	1.00 (referent)	—
Never married	0.70 (0.45-1.09)	.11	0.80 (0.49-1.31)	.38
Separated, divorced, or widowed	0.42 (0.32-0.57)	<.001	0.46 (0.33-0.64)	<.001
Family history of breast cancer				
No	1.00 (referent)	—	—	—
Yes	1.04 (0.68-1.59)	.86	—	—
BCS[c] attendance				
No	1.00 (referent)	—	—	—
Yes	3.49 (2.71-4.50)	<.001	—	—
Previous BCS type				
Mammography	1.00 (referent)	—	—	—
Ultrasound only	0.51 (0.23-1.13)	.10	—	—
Neither	0.27 (0.21-0.36)	<.001	—	—
Current BCS behavior				
Once a year	1.00 (referent)	—	1.00 (referent)	—
Once every 2 years	0.61 (0.31-1.23)	.17	0.65 (0.33-1.30)	.23
Do not intend to attend in the future	0.02 (0.01-0.03)	<.001	0.03 (0.01-0.06)	<.001
Other	0.10 (0.05-0.19)	<.001	0.13 (0.07-0.27)	<.001
Not in above categories and aged 35-39 years	0.16 (0.08-0.32)	<.001	0.17 (0.03-0.87)	.03
Unknown	0.06 (0.03-0.11)	<.001	0.10 (0.05-0.20)	<.001
Perceived importance of BCS				
Agree	1.00 (referent)	—	1.00 (referent)	—
Neutral	0.07 (0.05-0.10)	<.001	0.21 (0.14-0.32)	<.001
Disagree	0.03 (0.01-0.12)	<.001	0.11 (0.02-0.50)	.004
Finds mammography painful				

XSL•FO
RenderX

Characteristic	Univariate		Best prediction model by BIC	
	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Agree	1.00 (referent)	—	1.00 (referent)	—
Disagree	4.17 (3.16-5.50)	<.001	2.08 (1.52-2.85)	<.001
Not applicable	2.03 (1.39-2.99)	<.001	1.95 (0.43-8.80)	.39

[a]BC: breast cancer.

[b]Not applicable.

[c]BCS: breast cancer screening.

After adjustment for study site and participant characteristics (age, ethnicity, marital status, housing type, highest academic status, and family history of BC), BC awareness remained significant ($OR_{aware\ vs\ unaware}$ 2.88, 95% CI 2.19-3.80; *P*<.001) (Table S6 in Multimedia Appendix 4). The best combination of factors (according to the lowest Bayesian information criterion) that predicts screening motivation was BC awareness, marital status, current BC screening behavior, perceived importance of BC screening, and pain during mammography.

Demographic Differences in Participants Using the English and the Non-English Questionnaire

Participants who used the English version of the questionnaire (n=3819) were more likely to be BC aware, younger, and living in larger housing, and had higher educational attainment, as compared to participants who used the non-English questionnaire (n=350) (Table S7 in Multimedia Appendix 4). A total of 350 participants answered the Chinese or Malay versions of the questionnaire, of which 31 (9%) answered the free-text question on reasons for screening.

Differences in Characteristics Between Participants Who Provided and Did Not Provide Free-Text Explanation

Participants who provided and did not provide an explanation were not statistically different in age, ethnicity, housing type, and family history of BC (Table S8 in Multimedia Appendix 4). Participants who provided explanations were more likely to have never been married (*P*=.04) and to have higher educational attainment (*P*<.001). These participants differed from those who did not provide free-text responses regarding their current BC screening behavior (*P*=.01), perceived importance of screening (*P*=.002), perceived risk of BC (*P*=.01), and reporting that mammography was painful (*P*=.004) (Table S8 in Multimedia Appendix 4).

Biterm Topic Modeling Reveals Key Themes by Motivation Level

Of the 3819 participants who answered in English, 94% (n=3579) were motivated, and 6% (n=240) remained neutral about regular screening after the breast education session. Among the motivated participants, 1220 (34%) provided additional comments, compared to 153 (64%) among the neutral participants. Tables S9 and S10 in Multimedia Appendix 4 provide the sets of words derived from biterm topic modeling. The top and bottom 20 words are listed.

In the motivated participants, we selected 4 topics as repeats in top words occurred at 5 topics (Table S9 in Multimedia Appendix 4). “Early” and “health” were consistent across all 6 iterations, with “age” and “appointment” occurring 4 times, “better” twice, and “risk” and “painful” once. It was observed that “age” and “risk” belonged to the same set, while “appointment,” “better,” and “painful” were used commonly together. Using the first iteration (Table 3), topics observed were (1) benefits of early detection (with words “early,” “detection,” “better,” “cure,” and “prevention”), (2) health awareness (“health,” “regular,” “know,” “check,” and “important”), (3) BC risk (“age,” “risk,” “40,” “2,” and “safe”), where 40 refers to the age of 40 years and 2 refers to screening once every 2 years, (4) logistics and mammography experience (“appointment,” “regular,” “routine,” “schedule,” and “painful”).

In the neutral participants, we selected 2 topics as repeats in the top words that occurred at 3 topics (Table S10 in Multimedia Appendix 4). No word occurred in all iterations. “Pain” and “painful” were from the same set of words; “time,” “regular,” “health,” and “examination” were commonly from the same set of words. Using the first iteration (Table 3), unlike in the motivated participants, there was a distinction between (1) mammography experience with a focus on pain during a mammogram and (2) logistics, with a mention of “expensive” that was not observed in the top words from motivated participants. This is in line with our finding that participants who disagreed that mammography was expensive were more likely ($OR_{disagree\ vs\ agree}$ 1.94, 95% CI 1.50-2.50) to be willing to attend the screening posteducation session (Table S6 in Multimedia Appendix 4).

XSL•FO
RenderX

Table . Topics surfaced using biterm topic modeling on the free-text response following the question, "After knowing the above, would you be more willing to attend regular screening?" by willingness to screen: motivated and neutral. As different iterations with the same number of clusters produce variations in results, the first iteration is presented. Topic distribution (θ) provides the proportion of each topic in the entire corpus (dataset of additional comments). "k" refers to the number of clusters and was selected based on the occurrence of repeated topics within each iteration.

Cluster	Value (θ)	Top word	Subsequent 19 words	Topic
Motivated participants (n=1205)				
1	0.550	Early	Detection, better, cure, detect, prevention, chance, treatment, life, save, regular, survival, important, increase, live, recovery, believe, stage, always	Benefits of early detection
2	0.248	Health	Early, regular, good, know, check, important, detection, always, time, body, need, prevent, condition, family, age, risk, better, detect	Health awareness
3	0.103	Age	Risk, regular, 40, 2, safe, know, need, check, reduce, women, test, health, importance, 50, diagnose, healthy, aware, yearly	Breast cancer risk
4	0.099	Appointment	Health, better, regular, routine, follow, schedule, painful, women, easy, important, preventive, experience, hard, purpose, issue, patient, question, yearly	Logistics and mammogram experience
Neutral participants (n=145)				
1	0.569	Pain	Time, 2, once, cost, follow, risk, 1, booking, know, change, comfortable, frequent, abnormality, body, chance, depend, factor, increasing	Mammogram experience
2	0.431	Regular	Health, health care, painful, need, expensive, 40, attend, examination, women, schedule, time, change, decision, subsidize, appointment, ease, good, impact	Logistics

Subgrouping by BC Awareness Yields Similar Results to Motivation-Based Grouping

We repeated the analysis comparing BC-aware vs BC-unaware participants. Eighty-four percent (n=1010) of motivated women and 66% (n=95) of neutral women were BC-aware. When participants were grouped by BC awareness ($n_{BC\text{-}aware}$=1105 and $n_{BC\text{-}unaware}$=243), 2 topics were identified for BC-aware (ie, benefits of early detection and health awareness), similar to the motivated group (Table S11 in Multimedia Appendix 4). In BC-unaware participants, the main topic identified was the benefits of early detection (Table S12 in Multimedia Appendix 4). A second topic was not identified, as the word sets varied across the 6 iterations, likely reflecting the limited thematic diversity of responses. Conclusions about additional potential topics should be interpreted carefully.

More Motivated Participants Displayed More Positive Sentiments Toward Regular Screening

The minimum sentiment rating obtained in the analysis was −1.00, and the maximum rating was 1.65. The median and mean of the sentiment scores were 0.177 (IQR 0.000-0.202) and 0.202 (SD 0.327), respectively, indicating that positive sentiments were more prevalent than negative ones among the participants. The mean sentiment values and 95% CIs of each group are displayed in Figure 1. Motivated participants had significantly more positive sentiments toward regular screening than neutral participants, regardless of existing BC awareness.

XSL•FO
RenderX

Figure 1. Comparison of average sentiments across groups according to breast cancer (BC) awareness and willingness to screen (♦=mean). Median sentiment and IQR values of each subgroup are displayed at the bottom. Differences in sentiments were tested using the Wilcoxon test (***P<.001, **P<.01, and *P<.05). NS: not significant.

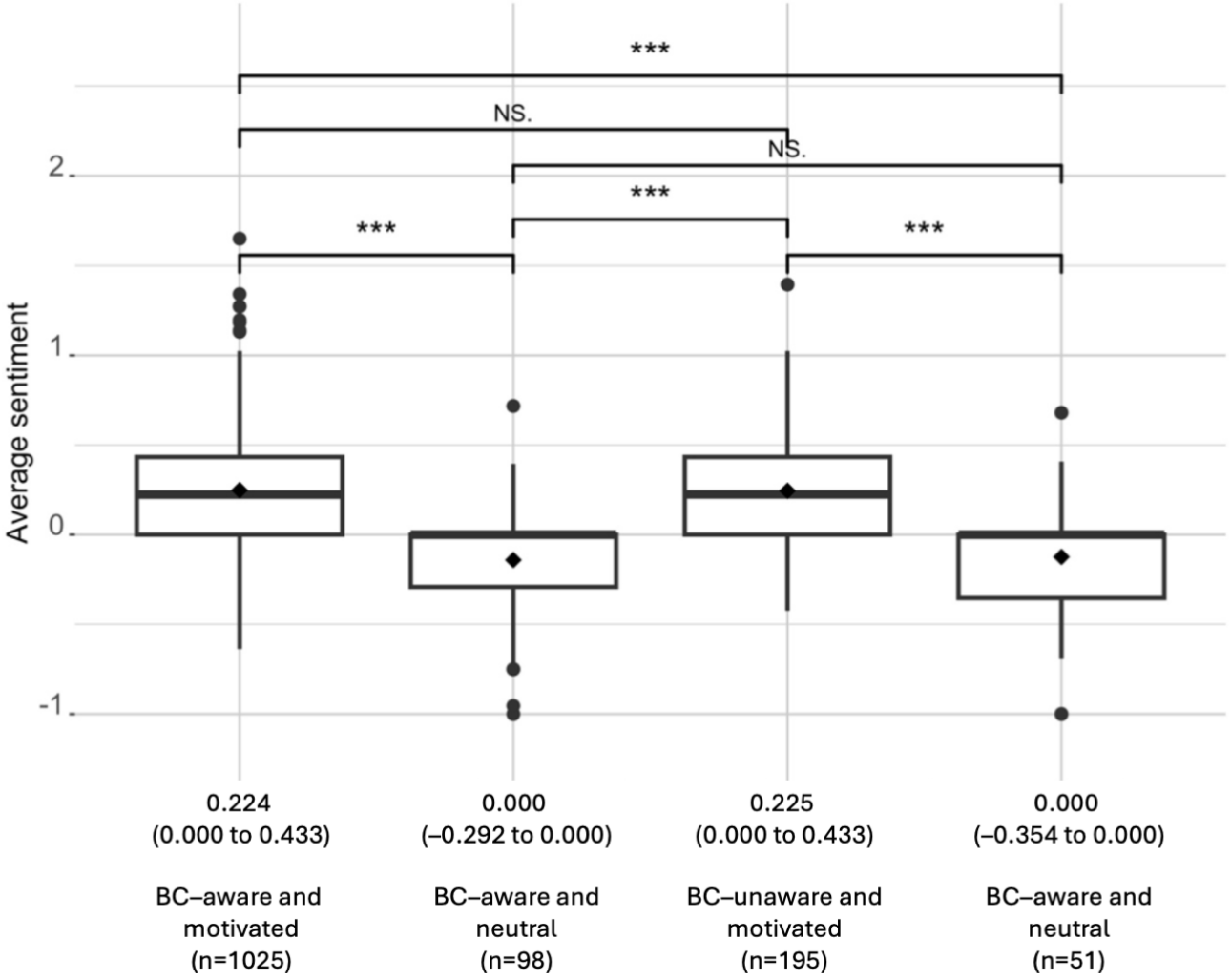

Word Frequencies are Indicative of Subgroup Sentiments

Words, expressed by at least 5% of each subgroup, were summarized and tabulated (Table 4). Motivated participants, both BC-aware and BC-unaware, displayed higher frequencies of words with positive sentiments (eg, “better,” “cure,” and “prevention”). Neutral participants who were BC-unaware displayed higher frequencies of negative sentiment words (eg, “painful,” “pain,” and “uncomfortable”), and only “painful” had over 5% frequency in participants who were BC-aware (Table 4).

XSL•FO
RenderX

Table . Top words mentioned in the free-text responses following the question, "After knowing the above, would you be more willing to attend regular screening?" by breast cancer awareness (BC-aware or BC-unaware) and willingness to attend screening (motivated or neutral), were concluded using biterm topic modeling. Words were mentioned by at least 5% of participants in each respective subgroup.

Subgroups	Positive words	Neutral words	Negative words
BC-aware[a] and motivated (n=1050)	Cure (n=74); better (n=136); important (n=80); detection (n=223); prevention (n=107); detect (n=68)	Early (n=362); health (n=101); regular (n=62)	—[b]
BC-aware and neutral (n=99)	—	Time (n=8); regular (n=8); health (n=6); cost (n=5)	Painful (n=15)
BC-unaware and motivated (n=206)	Better (n=24); important (n=19); detection (n=24); detect (n=22); prevention (n=17); prevent (n=13)	Early (n=48); health (n=20); regular (n=15); know (n=12)	—
BC-unaware and neutral (n=54)	Working (n=2); depends (n=2)	Time (n=6); necessary (n=3); regular (n=3); examination (n=3); cost (n=2); need (n=2); know (n=2)	Painful (n=8); troublesome (n=2); risk (n=2); pain (n=2); uncomfortable (n=2)

[a]BC: breast cancer.

[b]Not available.

Comparison of Words Across Subgroups

To gain deeper insights into the context of words and the subtleties surrounding future screening willingness, comments on the common and unique words from each subgroup were extracted. Here, we considered words that occurred in at least 5% of the subgroup.

Table S13 in Multimedia Appendix 4 provides the common words (n=12), along with the average sentiment of comments involving the words and some example quotes. Four words ("health," "important," "know," and "regular") had different sentence sentiments when used by the different subgroups. "Health" was a common word across 3 subgroups (BC-aware and motivated, BC-aware and neutral, and BC-unaware and motivated). When used by a participant in the BC-aware and motivated group, the sentence had a positive sentiment (eg, "health check is important to take care of one's health," "I want to stay healthy to see my girl grow up"). However, participants in the BC-aware and neutral group used it in sentences that had a negative sentiment ("I find it inconvenient to go for health screening," "the health care system does not allow women younger than 40 years to get subsidized breast screening"). Participants in the BC-unaware and motivated group were not significantly different from the BC-aware and motivated group (eg, "for my own health" and "it's important to keep track of one's health").

"Know" was used by BC-unaware participants, among whom the motivated participants used "know" in positive sentiment sentences (eg, "better to know if there is a sign of BC and get treatment" and "I thought that screening is expensive and did not know that there are subsidies"). This highlighted the utility of the education session in increasing BC awareness. However, participants who remained neutral used "know" to express negative sentiments (eg, "knowing the above does not affect my regular screening" and "better not to know"), which shows the limitations of the education exercise.

"Regular," a common word across all 4 subgroups, was used positively by motivated participants, for example, "important to do regular check-ups" by the BC-aware participants and "regular breast screening is important to detect any abnormality" by BC-unaware participants. "Regular" was used in negative sentences by the neutral participants, for example, "do not like too regular screening" by BC-aware participants and "I will be more open to regular screenings once I am 40" by BC-unaware participants.

Table 5 provides the words (n=4; "cure," "examination," "necessary," and "prevent") unique to the subgroup. While the words "examination" and "necessary" were mentioned by 3 participants (BC-unaware and neutral), the word "cure" was mentioned 74 times by BC-aware and motivated participants. The word "prevent" was mentioned 13 times by BC-unaware and motivated participants. It can be concerning that after the education session, participants used "prevent" regarding the occurrence of cancer (eg, "to prevent myself from any cancer").

XSL•FO
RenderX

Table . Use of unique sentiment words in context according to breast cancer awareness (BC-aware or BC-unaware) and willingness to screen (motivated or neutral).

Word	Behavior group	No of mentions in the group	Sentiment by participants in group, mean (SD; range)	Example quotes (random 3)
Cure	BC-aware[a] and motivated	74	0.803 (0.238; 0.177 to 1.650)	Prevention is better than cure; early detection will mean higher chance of survival; prevention is better than cure; prevention better than cure
Examination	BC-unaware and neutral	3	0.013 (0.022; 0.000 to 0.038)	If I do self-examination regularly it's not necessary; I will do self-breast examination to look out for any early signs; prefer to do self-examination for now
Necessary	BC-unaware and neutral	3	−0.031 (0.055; −0.094 to 0.000)	Still feel not necessary at this moment; depending on if it is necessary for me; if I do self-examination regularly it is not necessary
Prevent	BC-unaware and motivated	13	0.367 (0.345; −0.177 to 1.025)	it is very important need to prevent; no prevent breast cancer; prevention of breast cancer

[a]BC: breast cancer.

Discussion

Principal Findings

This study explored the use of sentiment analysis and biterm topic modeling to examine factors influencing participants' posteducation willingness to attend regular BC screening in a pilot study, using a specific question in a structured questionnaire with free-text questions to obtain participants' responses. Results from the free-text analysis indicate that most participants were BC aware and had positive sentiments toward regular screening. Major topics discovered were related to the benefits of early detection and health awareness in participants who were motivated to attend BC screening. In neutral participants, mammography experience, in particular, pain during the mammogram, was the key topic highlighted.

The focus of this study was to examine the public's views, sentiments, and concerns regarding their willingness to participate in regular BC screening following a BC education session. NLP of free-text responses provided an efficient method to filter the concerns that are prioritized by participants through biterm topic models and word count tabulations and allowed us to explore reasons by accessing the full entry. Key topics of interest to the participants may not surface in structured questionnaires, and the options are restricted by the researchers' choice of questions. In our case, the perceived importance of screening and perceived screening barriers such as inconvenience, pain, embarrassment, and cost were assessed using a limited range of available response options (5-point Likert scale or yes or no).

A review by Hisan et al [31] categorized the use of NLP in health care into clinical applications, such as clinical documentation, medical coding, clinical decision support, and patient engagement, and public health applications, including sentiment analysis, clinical trials, and disease surveillance, demonstrating how NLP enhances health care communication and outcomes. Thackeray et al [32] analyzed over 1.3 million tweets during Breast Cancer Awareness Month in 2012, revealing that most tweets were original posts by individuals, with organizations and celebrities focusing on fundraisers and early detection, while highlighting the need for strategic, ongoing social media engagement to promote preventive behaviors and maximize outreach.

Another study analyzing tweets by Nastasi et al [33] found that nonhealth care users frequently shared unsupported claims and expressed confusion about screening guidelines, highlighting the need for better public education on BC prevention and accurate information dissemination. These studies have demonstrated the immense potential of NLP in identifying current health care barriers and equipping governmental and health care institutions with information to deliver more targeted programs. Applying NLP to this study, we found that more motivated participants were more concerned with screening utility and understood the value of regular screening, while topics brought up by neutral participants were focused on the cost (financial, mental, and time) of screening.

The participation rate for routine mammography screening is poor in Singapore's population (<40%), a country that may be considered to be BC-aware (National Population Health Survey). In this study, we found 79% (n=3287) of our participants to be BC-aware. After completing the BC education questions, 94% (n=3908) reported being more motivated to attend screening. This is in contrast with only 63% (n=2628) who reported regular screening before the education session. BC awareness remained

XSL•FO
RenderX

a significant factor in participants' willingness to attend screening after adjusting for current screening frequency. Biterm topic modeling highlighted that the reinforcement of information on the benefits of early detection motivated the idea of attending screening.

In the area of barriers to screening, the main concern was pain among participants who were still not encouraged to screen. This echoes earlier findings that perceived pain or the experience of pain during mammography was a deterring factor for screening [34-37]. Although pain is a recognized barrier to mammography, our findings demonstrate that it remains a persistent concern. Future research should explore practical strategies to reduce discomfort and support more positive screening experiences. In contrast, other known barriers for screening, such as fatalism and cost, did not surface as concerns among our participants [4]. This may be due to multiple factors. Prior research studies on fatalism in BC were inconsistent with regard to ethnicity. While Straughan and Seow [38] found fatalism to be associated with mammography uptake in a Chinese-majority cross-sectional study, Goh et al [9] found fatalism not to be associated with mammography uptake among the Malay community [39]. In addition, some studies included older populations (aged >60 years in 2016) who were not included in BREATHE (aged 35-59 years) [40]. Furthermore, the higher income found in our population, as compared with previous research that included a larger portion of lower-income participants, may have reduced concerns about cost [37,41].

Contextual differences of the same word emphasize the need for subgroup analysis using multiple methods (examining words and sentences) to capture the information provided by free text. Notably, the word "health" was common across motivated and neutral groups. However, the context in which "health" was used differed. Motivated participants used the term to emphasize the importance of screening for maintaining one's health. This is similar to other studies demonstrating that innate health consciousness and perceived screening importance served as facilitators to screening attendance [42-44]. In contrast, neutral participants used "health" to highlight potential health risks associated with screening and their fatalistic attitudes toward the disease.

In our BC-unaware and neutral group, the words "examination" and "necessary" were used in combination with "regular," reflecting their perception that regular self-examination is sufficient to inform them of their health status and thus that mammography is unnecessary. This presents a gap in the understanding of mammography screening. A randomized controlled trial in Shanghai found that while breast self-examination accompanied by intensive instruction can help detect smaller fibroadenomas, it was not effective in reducing BC mortality [45]. On the other hand, Duffy et al [46] showed that mammography screening can lead to a 41% reduction in BC mortality over 10 years, along with a 25% decrease in the incidence of advanced BCs. Importantly, mammography can detect nonpalpable breast tumors, which would otherwise be missed by self-examination or clinical examination [47]. There is a need to inform the public that although regular breast self-examination is an important habit, it is not a substitute for mammography screening.

Comparison With Previous Studies

Our previous studies in Singapore, including structured surveys, focus group discussions, in-depth interviews, and discrete choice experiments, identified cost and perceived risk (ie, women's belief that they are at low risk of developing BC) as major determinants of participation in both standard and risk-based screening programs [11,12,42,48-50]. In contrast, the current NLP analysis revealed that cost did not emerge as a prominent factor, and perceived risk was not directly represented as a key factor, appearing instead through related concepts such as perceived screening importance. We believe that a strength of the data-driven NLP analytical approach is that it can uncover subtle insights from unstructured text that traditional survey or qualitative methods may overlook.

Considering our findings in light of the low mammography screening uptake rate in Singapore, addressing knowledge gaps and negative sentiments toward screening may increase women's motivation to participate. Evidence suggests that individualized discussions that acknowledge women's specific concerns can increase informed screening uptake, whereas the absence of such discussions or a lack of health care provider recommendation may discourage participation [51]. A personalized risk-based screening approach could provide a practical pathway to improve participation, as it allows for tailored communication and targeted support based on each woman's risk profile. Our experience with the BREATHE implementation study showed that many women underestimated their BC risk, and that the perceived risk of BC can change after receiving results from a personalized predictive risk assessment [52]. However, confidence in the predicted risk result is lower among those who underestimated their risk [52]. Nonetheless, implementing such personalized approaches at scale poses challenges, including resource constraints and the need for trained personnel to deliver individualized feedback.

Challenges of Using Online Education Tools

There were challenges in using a self-administered online questionnaire as an educational tool for BC screening. Participants appeared to be confused about the concept of early detection and cancer prevention, despite the education session emphasizing that mammography detects cancer early but does not prevent its occurrence. This was reflected in the participants' continued use of the terms "prevention" and "prevent" when referring to mammography screening (eg, "prevention of cancer" and "prevention is better than cure"), suggesting persistent misconceptions about the purpose of screening. However, even when the conceptual distinction between cancer prevention and early detection is clarified, broader barriers (eg, health literacy, digital access, emotional concerns, and health care system influences) may continue to influence screening intentions.

Limitations

We acknowledge several limitations of this study. There remains a gap in the validity and suitability of the tools for NLP [53]. In this study, we did not explore the accuracy of each tool in this specific use case. Compared to other studies exploring BC sentiments on social media, the availability of specific questions

XSL•FO
RenderX

in this study shapes participants' responses and provides more relevant details.

However, participants could decide whether they wanted to provide answers, and most did not, resulting in a small dataset for analysis. Additionally, answers provided by participants were short, often 1-2 sentences, plausibly due to the lack of question prompts. Future studies can include more questions with improved prompts to increase the robustness of the topics covered and obtain deeper responses from participants.

We acknowledge that domain-specific sentiment lexicons may provide added nuance for clinically focused text. However, given the lay language and short responses in our dataset, the use of a general-purpose sentiment lexicon was appropriate.

This study only included the opinions of English-speaking participants and, thus, may not capture the diverse perspectives of all women. The cohort, drawn from a risk-based screening study, may also not fully represent the general population of screening-eligible women.

Our findings are based on participants' self-reported willingness to screen following an educational session and may not reflect actual screening behavior. The BC education session likely influenced participants' responses, particularly in shaping positive attitudes and highlighting certain themes such as early detection. As such, the topics and sentiments derived from the free-text responses may reflect immediate posteducation reactions rather than pre-existing beliefs, representing a potential source of bias. Future studies could incorporate baseline measurements or control groups to disentangle the effects of educational interventions on sentiment and topic expression.

Conclusion

We integrated quantitative surveys with NLP to reveal emotional dimensions of screening attitudes. Unlike existing studies focused on knowledge or intent, we showed that identical screening concepts are emotionally framed differently by motivated vs neutral participants. In practice, these findings support the need for emotionally tailored BC education and prevention strategies.

Acknowledgments

We would like to thank our dedicated research and administrative staff, Pooi Yee Wong, Ganga Devi D/O Chandrasegran, Nabilah Binte Supiee, Siti Zulyqha Binte Yazid, Alleza Joeay Balbanero Aquino, Pei Xuan Lim, Jolene Lu Yee Poh, Brenna Jing Jie Quah, Qian Ning Peh, Chun Mei Wang, Cara Wee Ying Wong, Kimiie, Wei Lin Chia, Yi Lin Chen, Jinan May Loewen, Hui Min Lau, Varshaa D/O Saravanan, Vannevia Jedidiah Shi Tong Foo, Nurfilya Binte Hamdil, Hian Ching Ng, Yen Shing Yeoh, Amanda Tse Woon Ong, Jing Jing Hong, and Siew Li Tan, for their contributions to the planning, preparation, and execution of BREAst Screening Tailored for HEr (BREATHE). We would also like to thank Dr Chuan Chien Tan for assisting in the initial setup of the project, the doctors from the Department of Obstetrics and Gynecology at the National University Hospital, Dr Judith Shan Lin Ong and Dr Susan Jane Sinclair Logan, for allowing our team to conduct recruitment at Jade Clinic, and the radiologists from the Department of Diagnostic Imaging, Breast Imaging, at the National University Hospital, Dr Pooja Jagmohan, Dr Premilla Gopinathan Pillay, Dr Yonggeng Goh, Dr Han Yang Ong, and Dr Abrar, for helping with the density reading scores in our BREATHE cohort. Portions of this manuscript were edited for language and clarity with the assistance of ChatGPT (OpenAI; GPT-4). All content, interpretations, and conclusions are solely the work of the authors. Artificial intelligence (AI)-generated content was reviewed and revised by the authors to ensure accuracy, relevance, and scientific integrity.

Funding

This study was funded by the JurongHealth Fund (reference no JHF-20-RE-003), Agency for Science, Technology and Research (A*STAR), and the Precision Health Research Singapore Clinical Implementation Pilot (PRECISE CIP) Fund. MH is supported by the JurongHealth Fund, PRECISE CIP Fund, the Breast Cancer Prevention Programme under the Saw Swee Hock School of Public Health Programme of Research Seed Funding (SSHSPH-Res-Prog-BCPP), the Breast Cancer Screening Prevention Programme under the Yong Loo Lin School of Medicine (NUHSRO/2020/121/BCSPP/LOA), the National University Cancer Institute Singapore (NCIS) Centre Grant Programme (CGAug16M005), and the Asian Breast Cancer Research Fund. The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The data generated by this study are owned by the providing institutions (Ng Teng Fong General Hospital [NTFGH], National University Hospital [NUH], Alexandra Hospital [AH], National University Polyclinics [NUP], and Jurong Medical Center [JMC]). Data may be obtained upon a reasonable request to the principal investigator, MH. The data are not publicly available due to privacy and/or ethical restrictions. Legal agreements will need to be drawn up between data requesters and providers for access to the deidentified data. The proposed studies need to comply with Singapore's laws and regulations regarding human biomedical research and clinical investigation, including the Declaration of Helsinki, International Good Clinical Practice Guidelines, and Good Clinical Practice guidelines by Singapore's Health Science Authority and the Ministry of Health.

XSL•FO
RenderX

Authors' Contributions

PJH and ZLL are joint first authors. J Li and MH conceptualized the study. ZLL and PJH developed the methodology. J Liu, NKMR, YJC, HLT, SAG, HBO, CHC, SCK, ZPZ, DLSO, STQ, SW, and MH curated the data. ZLL and PJH conducted the formal analysis. PJH, ZLL, and J Li prepared the original draft. ZLL, PJH, HBO, CHC, SCK, ZPZ, DLSO, STQ, SW, PTCI, J Liu, NKMR, YJC, HLT, SAG, MH, and J Li reviewed and edited the manuscript. J Li supervised the study. J Liu, NKMR, YJC, and HLT managed the project administration. MH and JL acquired the funding.

Conflicts of Interest

None declared.

Multimedia Appendix 1
BREATHE online structured first visit questionnaire, which includes factors linked to breast cancer and related conditions, such as demographic, lifestyle, and reproductive characteristics, prior treatments, and other environmental exposures.
[PDF File, 1015 KB - jmir_v28i1e78439_app1.pdf]

Multimedia Appendix 2
BREATHE education questionnaire.
[PDF File, 742 KB - jmir_v28i1e78439_app2.pdf]

Multimedia Appendix 3
Flowchart of participant selection from individuals enrolled in BREATHE.
[DOCX File, 18 KB - jmir_v28i1e78439_app3.docx]

Multimedia Appendix 4
Detailed reference classifications for sociodemographic variables (housing and educational attainment; Tables S1) and full natural language processing (NLP) pipeline documentation, including spelling corrections (Table S2), stemming rules (Table S3), stop word exclusions (Table S4), and manually adjudicated sentiment ratings (Table S5), alongside complete logistic regression models examining patient-level predictors of post-education screening willingness (Table S6), attrition analyses comparing English-language respondents to nonrespondents (Table S7) and free-text contributors to noncontributors (Table S8), exhaustive biterm topic model outputs across candidate topic solutions (k = 1–5) stratified by motivational response (more motivated vs. neutral; Tables S9–S10) and education session performance (>80% vs. ≤80%; Tables S11–S12), and a contextualized concordance analysis of high-frequency sentiment terms disaggregated by breast cancer awareness status and screening motivation (Table S13).
[XLSX File, 110 KB - jmir_v28i1e78439_app4.xlsx]

References

1. Ghoncheh M, Pournamdar Z, Salehiniya H. Incidence and mortality and epidemiology of breast cancer in the world. Asian Pac J Cancer Prev 2016;17(S3):43-46. [doi: 10.7314/apjcp.2016.17.s3.43] [Medline: 27165206]
2. Coleman C. Early detection and screening for breast cancer. Semin Oncol Nurs 2017 May;33(2):141-155. [doi: 10.1016/j.soncn.2017.02.009] [Medline: 28365057]
3. Wild CP, Weiderpass E, Stewart BW, editors. World Cancer Report: Cancer Research for Cancer Prevention: International Agency for Research on Cancer (IARC); 2020.
4. Rajendram P, Singh P, Han KT, et al. Barriers to breast cancer screening in Singapore: a literature review. Ann Acad Med Singap 2022 Aug;51(8):493-501. [doi: 10.47102/annals-acadmedsg.2021329] [Medline: 36047524]
5. Epidemiology & Disease Control Division and Policy RSG, Ministry of Health and Health Promotion Board, Singapore: National Population Health Survey; 2021.
6. Momenimovahed Z, Tiznobaik A, Taheri S, Hassanipour S, Salehiniya H. A review of barriers and facilitators to mammography in Asian women. Ecancermedicalscience 2020;14:1146. [doi: 10.3332/ecancer.2020.1146] [Medline: 33343705]
7. Teo CT, Yeo YWS, Lee SC. Screening mammography behavior and barriers in Singaporean Asian women. Am J Health Behav 2013 Sep;37(5):667-682. [doi: 10.5993/AJHB.37.5.11] [Medline: 23985290]
8. Straughan PT, Seow A. Barriers to mammography among Chinese women in Singapore: a focus group approach. Health Educ Res 1995 Dec;10(4):431-441. [doi: 10.1093/her/10.4.431]
9. Goh SA, Lee JK, Seh WY, et al. Multi-level determinants of breast cancer screening among Malay-Muslim women in Singapore: a sequential mixed-methods study. BMC Womens Health 2022 Sep 19;22(1):383. [doi: 10.1186/s12905-022-01972-y] [Medline: 36123600]

XSL•FO
RenderX

10. Surendran S, Foo CD, Tan DHY, et al. Understanding barriers and facilitators of breast and cervical cancer screening among Singapore women: a qualitative approach. Asian Pac J Cancer Prev 2023 Mar 1;24(3):889-895. [doi: 10.31557/APJCP.2023.24.3.889] [Medline: 36974542]
11. Kamila KA, Liow JJK, Giam F, et al. Spousal perceptions and their role in promoting breast cancer screening: a focus group study. BMJ Open 2025 Mar 12;15(3):e086340. [doi: 10.1136/bmjopen-2024-086340] [Medline: 40074274]
12. Liow JJK, Lim ZL, Ho PJ, et al. Attitudes and barriers to mammography screening in Singaporean women through the eyes of their adult children: a focus group study. SSM - Qualitative Research in Health 2022 Dec;2:100168. [doi: 10.1016/j.ssmqr.2022.100168]
13. Glasman LR, Albarracín D. Forming attitudes that predict future behavior: a meta-analysis of the attitude-behavior relation. Psychol Bull 2006 Sep;132(5):778-822. [doi: 10.1037/0033-2909.132.5.778] [Medline: 16910754]
14. Lee JW, Jones PS, Mineyama Y, Zhang XE. Cultural differences in responses to a likert scale. Res Nurs Health 2002 Aug;25(4):295-306. [doi: 10.1002/nur.10041]
15. Westland JC. Information loss and bias in likert survey responses. PLoS ONE 2022;17(7):e0271949. [doi: 10.1371/journal.pone.0271949] [Medline: 35901102]
16. Small ML, Cook JM. Using interviews to understand why: challenges and strategies in the study of motivated action. Sociol Methods Res 2023 Nov;52(4):1591-1631. [doi: 10.1177/0049124121995552]
17. Adams A, Cox AL. Questionnaires, in-depth interviews and focus groups. In: Research Methods for Human Computer Interaction: Cambridge University Press:17-34. [doi: 10.1017/CBO9780511814570.003]
18. Khaleghparast S, Maleki M, Hajianfar G, et al. Development of a patients' satisfaction analysis system using machine learning and lexicon-based methods. BMC Health Serv Res 2023 Mar 23;23(1):280. [doi: 10.1186/s12913-023-09260-7] [Medline: 36959630]
19. Khanbhai M, Anyadi P, Symons J, Flott K, Darzi A, Mayer E. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. BMJ Health Care Inform 2021 Mar;28(1):e100262. [doi: 10.1136/bmjhci-2020-100262]
20. Mehta N, Pandit A. Concurrence of big data analytics and healthcare: a systematic review. Int J Med Inform 2018 Jun;114:57-65. [doi: 10.1016/j.ijmedinf.2018.03.013] [Medline: 29673604]
21. Yim WW, Yetisgen M, Harris WP, Kwan SW. Natural language processing in oncology: a review. JAMA Oncol 2016 Jun 1;2(6):797-804. [doi: 10.1001/jamaoncol.2016.0213] [Medline: 27124593]
22. Clark EM, James T, Jones CA, Alapati A, Ukandu P, Danforth CM, et al. A sentiment analysis of breast cancer treatment experiences and healthcare perceptions across Twitter. arXiv. Preprint posted online on May 25, 2018 URL: https://arxiv.org/abs/1805.09959 [accessed 2026-02-19]
23. Döbrössy B, Girasek E, Susánszky A, Koncz Z, Győrffy Z, Bognár VK. "Clicks, likes, shares and comments" a systematic review of breast cancer screening discourse in social media. PLoS One 2020;15(4):e0231422. [doi: 10.1371/journal.pone.0231422] [Medline: 32294139]
24. Levitt HM, Bamberg M, Creswell JW, Frost DM, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):26-46. [doi: 10.1037/amp0000151] [Medline: 29345485]
25. Liu J, Ho PJ, Tan THL, et al. BREAst screening Tailored for HEr (BREATHE)—a study protocol on personalised risk-based breast cancer screening programme. PLoS ONE 2022;17(3):e0265965. [doi: 10.1371/journal.pone.0265965]
26. Ho PJ, Liu J, Goh SA, et al. Profile of the BREATHE cohort for risk-based breast screening in Singapore. Sci Rep 2025 Oct 10;15(1):35424. [doi: 10.1038/s41598-025-19205-1] [Medline: 41073454]
27. Silge J, Robinson D. tidytext: text mining and analysis using tidy data principles in R. J Open Source Softw 2016;1(3):37 [FREE Full text] [doi: 10.21105/joss.00037]
28. Cheng X, Yan X, Lan Y, Guo J. BTM: topic modeling over short texts. IEEE Trans Knowl Data Eng 2014;26(12):2928-2941. [doi: 10.1109/TKDE.2014.2313872]
29. Yan X, Guo J, Lan Y, Cheng X. A biterm topic model for short texts. 2013 Presented at: WWW '13: Proceedings of the 22nd international conference on World Wide Web; May 13-17, 2013; Rio de Janeiro, Brazil p. 1445-1456. [doi: 10.1145/2488388.2488514]
30. Zhu K, Weng J, Tu B. A comparative study on short-text topic modeling in online collaborative learning platforms: application and performance analysis of BTM and BERTopic. 2024 Presented at: ICAIE '24: Proceedings of the 2024 3rd International Conference on Artificial Intelligence and Education; Nov 22-24, 2024 URL: https://dl.acm.org/doi/proceedings/10.1145/3722237 [accessed 2026-02-19] [doi: 10.1145/3722237.3722309]
31. Hisan UK, Chaaben AB, Adeleye AO, Kaushik A. Impact of natural language processing on healthcare quality: a review of clinical and public health applications. JAAHM 2022 Jan;6(1):19-42 [FREE Full text]
32. Thackeray R, Burton SH, Giraud-Carrier C, Rollins S, Draper CR. Using Twitter for breast cancer prevention: an analysis of breast cancer awareness month. BMC Cancer 2013 Oct 29;13(1):508. [doi: 10.1186/1471-2407-13-508] [Medline: 24168075]

XSL•FO
RenderX

33. Nastasi A, Bryant T, Canner JK, Dredze M, Camp MS, Nagarajan N. Breast cancer screening and social media: a content analysis of evidence use and guideline opinions on Twitter. J Cancer Educ 2018 Jun;33(3):695-702. [doi: 10.1007/s13187-017-1168-9] [Medline: 28097527]
34. Jones T, Wisdom-Chambers K, Freeman K, Edwards K. Barriers to mammography screening among Black women at a community health center in South Florida, USA. Med Res Arch 2023 Apr;11(4). [doi: 10.18103/mra.v11i4.3814] [Medline: 37475892]
35. Gong J, Kampadellis G, Kong Q, Spijker W. Factors determining non-attendance in breast cancer screening among women in the Netherlands: a national study. Health Promot Int 2023 Jun 1;38(3). [doi: 10.1093/heapro/daac009]
36. Williams WM, Fu MR. Understanding breast cancer disparities affecting Black women: obtaining diverse perspectives through focus groups comprised of patients, providers, and others. J Community Health 2023 Oct;48(5):834-839. [doi: 10.1007/s10900-023-01227-3] [Medline: 37131066]
37. Lim SK, Teo XL, Ng JL, Li FX, Tan SM. A survey on Singaporean women's knowledge, perception and practices of mammogram screening. Ann Acad Med Singap 2015 Sep;44(9):317-325. [Medline: 26584660]
38. Straughan PT, Seow A. Fatalism reconceptualized: a concept to predict health screening behavior. Journal of Gender, Culture, and Health 1998 Jun;3(2):85-100. [doi: 10.1023/A:1023278230797]
39. Shaw T, Ishak D, Lie D, et al. The influence of Malay cultural beliefs on breast cancer screening and genetic testing: a focus group study. Psychooncology 2018 Dec;27(12):2855-2861. [doi: 10.1002/pon.4902] [Medline: 30264524]
40. Wee LE, Lim LY, Koh GCH. Two sides of the coin: a qualitative study of patient and provider perspectives on colorectal, breast and cervical cancer screening in a low-income Asian community. Proc Singap Healthc 2016 Jun;25(2):80-91. [doi: 10.1177/2010105815616404]
41. Seetoh T, Siew WF, Koh A, et al. Overcoming barriers to mammography screening: a quasi-randomised pragmatic trial in a community-based primary care setting. Ann Acad Med Singap 2014 Dec;43(12):588-594. [Medline: 25588917]
42. Lim ZL, Ho PJ, Khng AJ, et al. Mammography screening is associated with more favourable breast cancer tumour characteristics and better overall survival: case-only analysis of 3739 Asian breast cancer patients. BMC Med 2022 Aug 4;20(1):239. [doi: 10.1186/s12916-022-02440-y] [Medline: 35922814]
43. Safizade H, Amirzadeh N, Mangolian Shahrbabaki P. Motivational factors for breast cancer screening behaviors in Iranian women: a qualitative study. Asian Pac J Cancer Prev 2020 Oct 1;21(10):3109-3114. [doi: 10.31557/APJCP.2020.21.10.3109] [Medline: 33112574]
44. Jung SM, Jo HS. Intrinsic motivation factors based on the self-determinant theory for regular breast cancer screening. Asian Pac J Cancer Prev 2014;15(23):10101-10106. [doi: 10.7314/apjcp.2014.15.23.10101] [Medline: 25556433]
45. Thomas DB, Gao DL, Ray RM, et al. Randomized trial of breast self-examination in Shanghai: final results. J Natl Cancer Inst 2002 Oct 2;94(19):1445-1457. [doi: 10.1093/jnci/94.19.1445] [Medline: 12359854]
46. Duffy SW, Tabár L, Yen AMF, et al. Mammography screening reduces rates of advanced and fatal breast cancers: results in 549,091 women. Cancer 2020 Jul 1;126(13):2971-2979. [doi: 10.1002/cncr.32859] [Medline: 32390151]
47. Welch HG, Prorok PC, O'Malley AJ, Kramer BS. Breast-cancer tumor size, overdiagnosis, and mammography screening effectiveness. N Engl J Med 2016 Oct 13;375(15):1438-1447. [doi: 10.1056/NEJMoa1600249] [Medline: 27732805]
48. Liow JJK, Lim ZL, Sim TMY, et al. "It Will Lead You to Make Better Decisions about Your Health"-a focus group and survey study on women's attitudes towards risk-based breast cancer screening and personalised risk assessments. Curr Oncol 2022 Nov 25;29(12):9181-9198. [doi: 10.3390/curroncol29120719] [Medline: 36547133]
49. Lim ZL, Giam F, Wong RYX, Liow JJK, McCrickerd K, Li J. Do women with a history of breast cancer recommend risk-based breast cancer screening? An in-depth interview study. Front Psychol 2025;16:1414099. [doi: 10.3389/fpsyg.2025.1414099] [Medline: 40330299]
50. Wang Y, Ho PJ, Mou L, Li J. Women's preferences for testing to predict breast cancer risk – a discrete choice experiment. J Transl Med 2025;23(1). [doi: 10.1186/s12967-025-06119-9]
51. Georga D, Zartaloudi A, Saridi M, et al. Factors determining women's attitudes and knowledge toward breast cancer screening: a systematic review. Healthcare (Basel) 2025 Jul 4;13(13):1605. [doi: 10.3390/healthcare13131605] [Medline: 40648630]
52. Ho PJ, Goh SA, Goh SSN, et al. Impact of personalised risk predictions on breast cancer risk perceptions: insights from the BREATHE study. J Transl Med 2025 May 8;23(1):517. [doi: 10.1186/s12967-025-06515-1] [Medline: 40340901]
53. Gohil S, Vuik S, Darzi A. Sentiment analysis of health care tweets: review of the methods used. JMIR Public Health Surveill 2018 Apr 23;4(2):e43. [doi: 10.2196/publichealth.5789] [Medline: 29685871]

Abbreviations

BC: breast cancer
BREATHE: Breast Screening Tailored for Her
NLP: natural language processing
OR: odds ratio

XSL•FO
RenderX

Edited by S Brini; submitted 02.Jun.2025; peer-reviewed by I Moreira, L Raymond Guo, PC Lin; accepted 25.Dec.2025; published 10.Mar.2026.

Please cite as:
Ho PJ, Lim ZL, Liu J, Mohamed Riza NK, Chew YJ, Lim YY, Tan HL, Goh SA, Oh HB, Chin CH, Kwek SC, Zhang ZP, Ong DLS, Quek ST, Wijerathne S, Iau PTC, Hartman M, Li J
Breast Cancer Screening Knowledge and Sentiments in Singaporean Women: Mixed Methods Study Using Topic Modeling, Sentiment Analysis, and Structured Questionnaire Data
J Med Internet Res 2026;28:e78439
URL: https://www.jmir.org/2026/1/e78439
doi:10.2196/78439

© Peh Joo Ho, Zi Lin Lim, Jenny Liu, Nur Khaliesah Mohamed Riza, Ying Jia Chew, Yi Ying Lim, Hui Ling Tan, Su-Ann Goh, Han Boon Oh, Chi Hui Chin, Sing Cheer Kwek, Zhi Peng Zhang, Desmond Luan Seng Ong, Swee Tian Quek, Sujith Wijerathne, Philip Tsau Choong Iau, Mikael Hartman, Jingmei Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Suicidal Thoughts and Behaviors Among Chinese Adolescents in Relation to Negative Life Events, Internet Addiction, and Sexual Abuse: Cross-Sectional Study

Juanfang Zhu[1*], MPH; Qinqin Jiang[1,2*], MHA; Feng Zhang[1*], MPH; Hao Liu[1], MHA; Shaojie Yu[1], MPH; Jinhai Sun[1], PhD; Zhe Zhao[1], MPH; Lijuan Liu[1], PhD; Lei Yuan[1], MPH

[1]Faculty of Military Health Service, Naval Medical University, No. 800 Xiangyin Road, Shanghai, China

[2]College of Health Management, Southern Medical University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Lei Yuan, MPH
Faculty of Military Health Service, Naval Medical University, No. 800 Xiangyin Road, Shanghai, China

Abstract

Background: Increasing suicidal thoughts and behaviors (STB) among adolescents raise social concerns and have a well-recognized association with sexual abuse (SA). However, research regarding the mechanisms explaining the association between SA and STB remains limited.

Objective: This study aims to examine the chained mediating effects of negative life events (NLE) and internet addiction (IA) between SA and STB among adolescents in China.

Methods: This cross-sectional study used data from the Science Database of the People Mental Health survey conducted between March 2013 and December 2022 by the National Population Health Data Center of the National Research Institute for Family Planning. Through stratified sampling, 20,893 adolescents were recruited from 16 Chinese provinces. After excluding samples with missing relevant variables, 10,664 (55.89%; aged 16-17.9 y; n=5826, 54.63% women) adolescents were included in the final analysis. STB was the outcome variable, with NLE and IA as mediators, all assessed via a questionnaire that was uniformly administered by trained investigators in school settings. The Pearson χ^2 test was used to analyze the association between SA and STB. Using a combination of multiple linear regression and bootstrap testing, the study constructed a chain mediation model to explore how SA influences STB in adolescents through NLE and IA.

Results: The scores for SA, NLE, IA, and STB were 1.330 (SD 1.714), 51.960 (SD 23.822), 34.88 (SD 13.852), and 0.690 (SD 1.396), respectively. Multiple linear regression analysis indicated SA was associated with NLE (β=2.382, 95% CI 2.112 - 2.653; *P*<.001), IA (β=1.094, 95% CI 0.939 - 1.249; *P*<.001), and STB (β=0.135, 95% CI 0.119 - 0.151; *P*<.001). Mediation analysis revealed a significant direct association between SA and STB (74.81%; β=0.101, 95% CI 0.085 - 0.117; *P*<.001). Indirect effects constituted 25.19% of the total association (β=0.034, 95% CI 0.029 - 0.040; *P*<.001), comprised of 3 specific pathways: via NLE (11.85%; β=0.016, 95% CI 0.012 - 0.020; *P*<.001), via IA (10.37%; β=0.014, 95% CI 0.010 - 0.018; *P*<.001), and via the sequential pathway through NLE and then IA (2.96%; β=0.004, 95% CI 0.003 - 0.006; *P*<.001).

Conclusions: Unlike previous research that examined them separately, this study reveals the chain mediation mechanism of NLE and IA in the relationship between SA and STB, providing a new research perspective on the complex evolution from early adversity to extreme outcomes. This research also provided clear direction and scientific evidence for the development of comprehensive suicide prevention strategies. Future efforts should focus on the prevention of SA and trauma intervention while simultaneously strengthening adolescents' cognitive reappraisal of negative events and guiding their online behavior.

(J Med Internet Res 2026;28:e85371) doi:10.2196/85371

KEYWORDS

suicidal thoughts; suicidal behaviors; sexual abuse; SA; internet addiction; IA; chain mediation effect; adolescents

Introduction

Adolescents' susceptibility to suicidal thoughts and behaviors (STB) raises social and academic concerns [1-3]. According to the World Health Organization (WHO) estimates, suicide was the fourth leading cause of mortality for youths aged 15 to 29 years globally in 2019 [4], with significant variations observed across countries, age groups, and genders [5]. Both attempted and completed suicides impose substantial social, economic, and emotional burdens worldwide. Despite efforts to reduce

XSL•FO
RenderX

suicide risk, global rates of youth STB exhibit no improvement [6]. Hence, the identification of risk factors and antecedents predisposing individuals to increased suicide risk is imperative to enable prompt and appropriate monitoring and intervention [7]. Sexual abuse (SA) is a recognized risk factor for suicidal ideation and behavior [8,9]. By disrupting normal neuroendocrine and immune functions [10], SA exerts a direct effect on increasing suicide risk [11]. Accordingly, this study proposes hypothesis 1 as follows: SA is positively associated with STB in adolescents. Although this association is well-established, the mediating mechanisms through which SA elevates suicide risk remain unclear, particularly among Chinese adolescents.

Based on theoretical reasoning, negative life events (NLE) and internet addiction (IA) may serve as potential mediators. According to the integrated motivational-volitional model of STB, NLE are the main stressors driving STB, which play a "trigger" role [12] and are strong predictors thereof [13]. Previous cross-sectional studies have linked NLE such as academic pressure and interpersonal conflict to suicidal ideation and behavior in university students [14,15], an association also supported by longitudinal research [16]. Furthermore, evidence suggests that survivors of SA are more likely to experience other forms of interpersonal harm, including physical assault [17], illicit drug dependence [18], and adolescent dating violence [19]—all constituting NLE. Thus, hypothesis 2 is proposed as follows: NLE mediate the relationship between SA and STB. Recent studies indicate a significant association between adolescent IA and suicidal ideation or behavior [17,20,21], with addicted adolescents exhibiting higher suicide risk. A key risk factor for IA may be childhood trauma, which is strongly predictive of lifelong mental disorders [22,23]. A history of childhood trauma appears equally prevalent among individuals with IA and those with substance dependence [24]. Moreover, a study of Chinese university students found that childhood trauma significantly influences suicidal ideation, an effect mediated by IA [25]. Hence, hypothesis 3 is proposed as follows: IA mediates the relationship between SA and STB. Due to underdeveloped social-cognitive abilities and the inevitability of academic and interpersonal stressors, adolescents often experience NLE. While the causes of IA are still under investigation, NLE are considered a key contributing factor. Research shows a significant positive correlation between NLEs and IA among university students [26-29], a link observed across subtypes such as interpersonal distress, academic pressure, punishment, and health-related problems [30]. Consequently, hypothesis 4 is proposed as follows: NLEs and IA operate as serial mediators in the relationship between SA and STB.

While existing research has separately revealed significant associations between SA and STB, between NLEs and STB, and between IA and STB, how these factors interact within a coherent mechanistic model to collectively explain the pathway from early trauma to ultimate risk remains poorly understood [8-11,14,15,17,21,26]. Building on prior findings, this study aims to use a large representative sample of youth to examine the association between SA and STB, and to explore the potential direct and indirect effects of SA on STB mediated through NLE and IA. This investigation seeks to deepen the understanding of the mechanisms underlying adolescent suicide risk and to provide an evidence-based roadmap for the precise prevention and tiered intervention of suicide risk among Chinese adolescents.

Methods

Data and Sampling

This study used data from the Science Database of the People Mental Health survey conducted between March 2013 and December 2022 by the National Population Health Data Center of the National Research Institute for Family Planning [31]. The data are designed horizontally and vertically, and through dynamic, continuous, and systematic collection of mental health and related risk factor information from children, adolescent, young adult, middle-aged, and older adult populations, as well as from specific populations, occupational populations, and mental health specialties. Thereby, the trajectory of changes in the mental health of individuals, clusters, and populations is presented, and the knowledge mapping of mental health data is developed. The surveys adhere to psychological industry standards, expert consensus, guidelines, and regulations, using widely accepted international and domestically standardized scales adapted for Chinese norms. Data collection involved validity and reliability assessments and was conducted using standardized procedures for both offline and online administration, using surveys and assessments within a mental health context. A multistage stratified random sampling method was used in this study. From March 2013 to December 2022, 20,893 adolescents were recruited from secondary schools across 16 provinces in China. For this analysis, after excluding respondents with missing values on any variable, the final analytical sample comprised 10,664 adolescents. Figure 1 illustrates the data cleaning process. Prior to the survey, informed consent was obtained from both the respondents' parents and the respondents themselves, following local parental consent procedures.

XSL•FO
RenderX

Figure 1. Flowchart of adolescent participants in the cross-sectional study on sexual abuse (SA) and suicidal thoughts and behaviors (STB), China, 2013 - 2022. IA: internet addiction

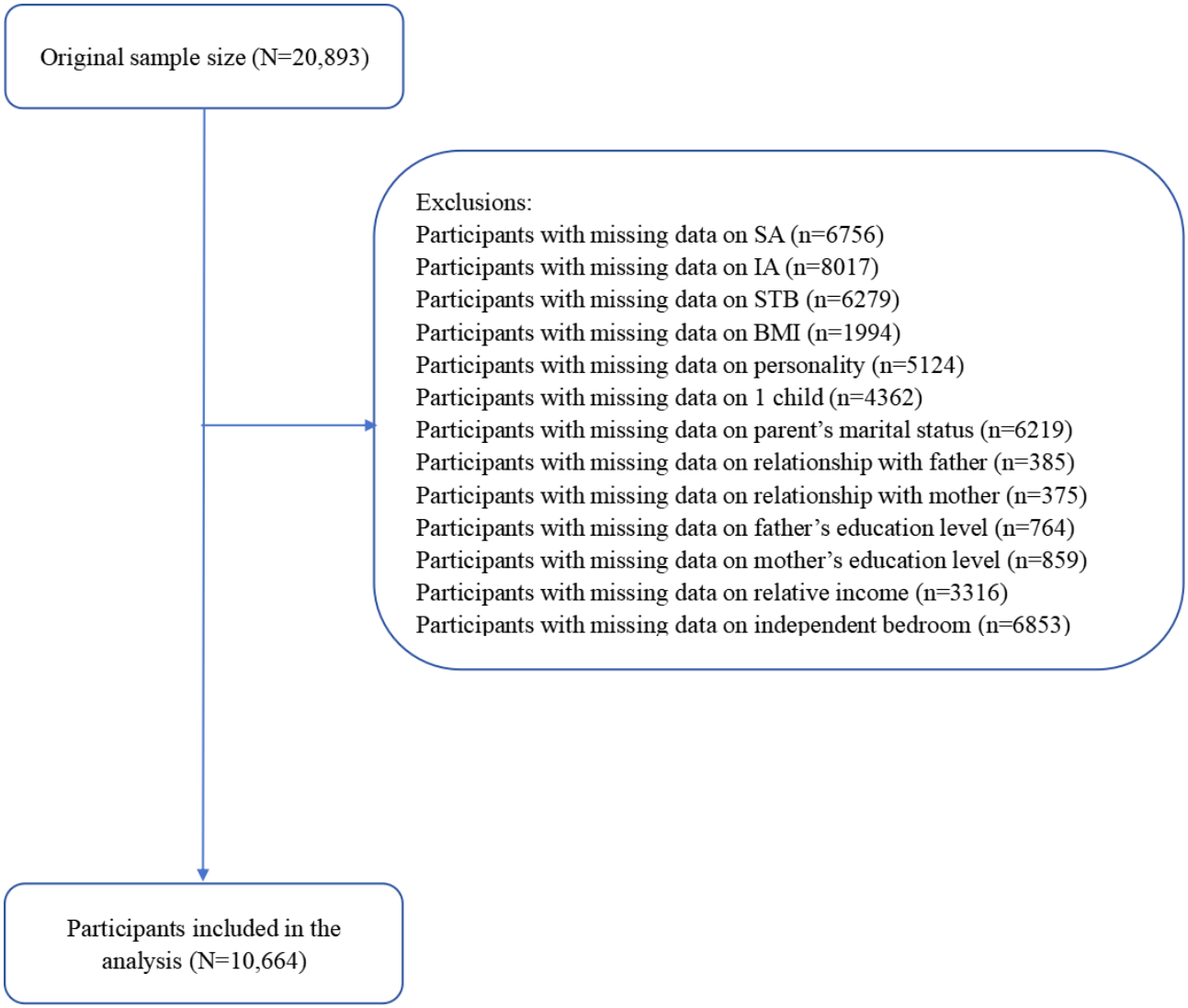

Measures

Sociodemographic Variables

Sociodemographic variables included age, gender, ethnic group, BMI, personality type, 1-child status, parents' marital status, relationship with father, relationship with mother, father's educational level, mother's educational level, region, relative income, and having an independent bedroom. Regions, categorized according to the affiliation of the respondent's province in the *Chinese Health Statistics Yearbook* [32], were divided into central, western, and eastern regions.

Grouping Variable

This study's grouping variable is hukou, which is categorized into urban and rural areas.

Suicidal Thoughts and Behaviors

Suicidal behaviors are measured by using the following 3 items, according to the Global School-Based Student Health Survey, including suicide ideation, plans, and attempts, respectively: "Have you seriously considered suicide in the last 12 months?" "Have you made a specific plan to commit suicide in the last 12 months?" and "Have you attempted suicide in the last 12 months?" All 3 variables were considered dummy variables, with "never," "≤2 days," and ">2 days" coded as 0, 1, and 2, respectively. The total score for STB ranges from 0 to 6, with a higher score indicating more STB. In this study sample, the Cronbach α coefficient for the STB assessment was 0.713, the Kaiser-Meyer-Olkin (KMO) value was 0.667, and the Bartlett test of sphericity yielded a significant result (P<.001).

Sexual Abuse

Respondents were asked the following 9 questions to measure SA: "Has anyone ever kissed you without your consent?" "Has anyone ever exposed their genitals to you?" "Has anyone ever masturbated in front of you?" "Has anyone ever told you dirty jokes, made pornographic gestures, or shown you pornographic pictures?" "Has anyone ever touched your sensitive parts (eg, lower body, breasts, and genitals)?" "Has anyone ever rubbed their genitals on you?" "Has anyone ever made oral contact with your sex organ?" "Has anyone ever tried to have sex with you?" and "Has anyone ever forced sex on you?" All 9 variables were considered dummy variables, with "yes" and "no" coded as 1 and 0, respectively. In this study sample, the Cronbach α coefficient for the SA experience assessment was 0.747, the

XSL•FO
RenderX

KMO value was 0.850, and the Bartlett test of sphericity yielded a significant result (P<.001).

Negative Life Events

This study measured NLE using the ASLEC (Adolescent Self-Rating Life Event Checklist; Checklist 2) [33], which is used to assess adverse events within the past 12 months. The ASLEC comprises 27 items in the following 6 dimensions: interpersonal relationships (interpersonal conflict or contradiction), study pressure (stress caused by learning), being punished (punishment by caregivers, teachers, or peers), bereavement (death of beloved ones), health adaptation, and other dimensions (eg, school refusal) [33]. The ASLEC assesses the frequency and intensity of life events experienced by respondents in the past half-year. Respondents rate each NLE's impact on a 5-point Likert scale (1=never to 5=extremely severe). The ASLEC's total score ranges between 27 and 135, with a higher score indicating a greater number of NLE and greater stress experienced over the past half-year. Previous studies have reported that this scale is reliable and valid among Chinese children [27,34]. In this study sample, the Cronbach α coefficient for the Adolescent Life Events Scale was 0.965, the KMO value was 0.971, and the Bartlett test of sphericity yielded a significant result (P<.001).

Internet Addiction

The research focused predominantly on IA as measured by the Internet Addiction Test (IAT) developed by Young [35]. The IAT is a 20-item questionnaire that measures characteristics and behaviors associated with compulsive internet use. Each item is weighted using a Likert scale ranging from 1 ("rarely") to 5 ("always"). The IAT scale comprises the following 6 subscales: salience, excessive use, neglect of work, anticipation, lack of control, and neglect of social life; however, precise interval thresholds are not defined for this scale. Therefore, a higher score represents greater addiction. Previous studies have reported that this scale is reliable and valid among Chinese children [36,37]. In this sample, the scale demonstrated excellent internal consistency, with a Cronbach α coefficient of 0.947. KMO measure of sampling adequacy was 0.959, and the Bartlett test of sphericity was significant (P<.001).

Figure 2 illustrates the proposed serial mediation model.

Figure 2. Proposed serial mediation model depicting the relationships among sexual abuse, negative life events, internet addiction, and suicidal thoughts and behaviors in Chinese adolescents.

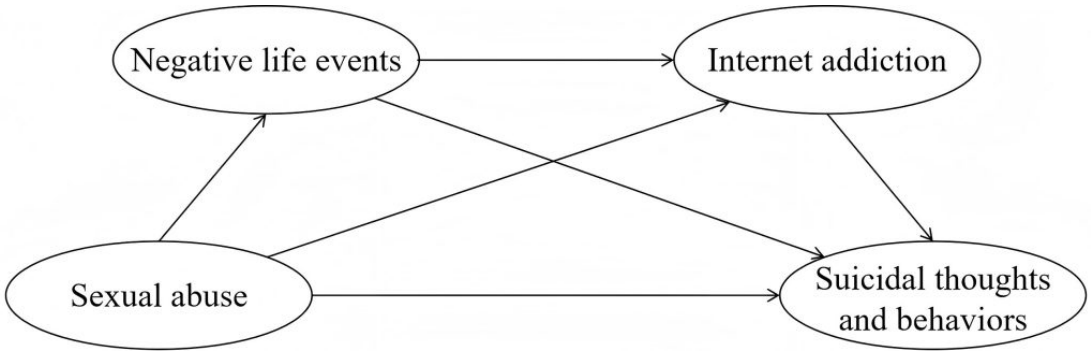

Statistical Analysis

Data analysis was performed using Stata (version 18.0; StataCorp LLC) and SPSS (version 26.0; IBM Corp) software. The study was reported following the Reporting Standards for Studies Using Structural Equation Modeling to ensure compliance with the transparency and completeness requirements stipulated by the JARS (Journal Article Reporting Standards) guidelines (Checklist 1) [38,39]. Descriptive statistics were computed for respondents' sociodemographic characteristics. Frequencies and percentages were reported for categorical variables, whereas means and SDs were reported for continuous variables. Associations between independent and outcome variables were estimated using the Pearson chi-square test (or 2-tailed t test for continuous variables). The Pearson correlation analysis measures the strength and direction of a linear relationship between dependent and independent variables. Using a combination of multiple linear regression and bootstrap testing, we constructed a chain mediation model. Multivariate linear regression analysis was used to identify significant independent correlates of STB. To test the hypothesized mediation effects, we conducted path analysis using the bootstrap method, with 5000 bootstrap resamples to generate bias-corrected 95% CI. This approach quantified the indirect effects of the independent variable on the dependent variable through the mediators, as well as the direct effects after controlling for the mediators. The 2-tailed significance level was set at P<.05.

Ethical Considerations

This study is a secondary analysis based on the Science Database of the People Mental Health Survey. The original data collection adhered to the principles of the Declaration of Helsinki, and written informed consent was obtained from all participants and their guardians. This research used only processed, publicly available data containing no personally identifiable information and involved no interaction with human participants. No identification of individual participants is presented in any images within this study or supplementary materials. The Science Database of the People Mental Health Survey has specific licensing and citation requirements. The authors obtained permission to use the data and have strictly complied with the relevant data use agreement regarding noncommercial academic research and proper citation. This study was exempted from additional ethics review and approval by the Naval Medical University Ethics Committee.

XSL•FO
RenderX

Results

Sociodemographic Characteristics

Among the participants, 2075 (19.46%), 5960 (55.89%), and 2629 (24.65%) were aged younger than 16 years, 16 to 17.9 years, and 18 years or older, respectively; 4838 (45.37%) and 5826 (54.63%) were male and female adolescents, respectively; 5703 (53.48%) and 4961 (46.52%) were Han Chinese and belonged to an ethnic minority, respectively; 6796 (63.73%), 2511 (23.55%), 998 (9.36%), and 359 (3.37%) had BMIs in the ranges of 18.5 to 23.9 kg/m^2, <18.5 kg/m^2, 24 to 27.9 kg/m^2, and ≥28 kg/m^2, respectively; 1430 (13.41%), 3355 (31.46%), and 5879 (55.13%) were introverts, extroverts, and ambiverts, respectively; 3099 (29.06%) had 1-child status in their families; 421 (3.95%) and 369 (3.46%) had divorced parents and remarried families, respectively; 8833 (82.83%) and 9712 (91.07%) had amicable relationships with their fathers and mothers, respectively; 1012 (9.49%) and 915 (8.58%) respondents' fathers and mothers had an educational level of college and higher, respectively; 4898 (45.93%), 2265 (21.24%), and 3501 (32.83%) were in the central, eastern, and western regions, respectively; 6247 (58.58%), 2539 (23.81%), and 1878 (17.61%) had general, rich, and poor income statuses, respectively; and 8910 (83.55%) had an independent bedroom. The chi-square test and 2-tailed *t* test revealed differences in age, gender, ethnic group, BMI, 1-child status, parents' marital status, father's educational level, mother's educational level, relative income, and independent bedroom of middle school adolescents between urban and rural areas (Table 1). Gender differences exist in the prevalence of SA (57.24%) and STB (24.8%) among adolescent samples in China currently (Table S1 in Multimedia Appendix 1).

XSL•FO
RenderX

Table . Characteristics of Chinese adolescent participants (N=10,664).

Variables	Total (N=10,664)	Rural (n=8028)	Urban (n=2636)	Statistics, chi-square (*df*)	*P* value
Age (y), n (%)				9.3 (2)	.01
<16	2075 (19.46)	1521 (18.95)	554 (21.02)		
16-17.9	5960 (55.89)	4479 (55.79)	1481 (56.18)		
≥18	2629 (24.65)	2028 (25.26)	601 (22.80)		
Gender, n (%)				30.3 (1)	<.001
Woman	5826 (54.63)	4508 (56.15)	1318 (50)		
Man	4838 (45.37)	3520 (43.85)	1318 (50)		
Ethnic group, n (%)				30.3 (1)	<.001
Han Chinese	5703 (53.48)	4171 (51.96)	1532 (58.12)		
Ethnic minority	4961 (46.52)	3857 (48.04)	1104 (41.88)		
BMI (kg/m^2), n (%)				34.0 (3)	<.001
18.5~23.9	6796 (63.73)	5222 (65.05)	1574 (59.71)		
<18.5	2511 (23.55)	1853 (23.08)	658 (24.96)		
24.0~27.9	998 (9.36)	714 (8.89)	284 (10.77)		
≥28.0	359 (3.37)	239 (2.98)	120 (4.55)		
Personality, n (%)				13.3 (2)	<.001
Introvert	1430 (13.41)	1107 (13.79)	323 (12.25)		
Extrovert	3355 (31.46)	2454 (30.57)	901 (34.18)		
Ambivert	5879 (55.13)	4467 (55.64)	1312 (53.57)		
One child, n (%)				1200.0 (1)	<.001
No	7565 (70.94)	6397 (79.68)	1168 (44.31)		
Yes	3099 (29.06)	1631 (20.32)	1468 (55.69)		
Parent’s marital status, n (%)				66.7 (3)	<.001
Married	9412 (88.26)	7161 (89.02)	2251 (85.39)		
Divorce	421 (3.95)	275 (3.43)	146 (5.54)		
Remarry	369 (3.46)	227 (2.83)	142 (5.39)		
Other	462 (4.33)	365 (4.55)	97 (3.68)		
Relationship with father, n (%)				7.5 (2)	.02
Well	8833 (82.83)	6671 (83.10)	2162 (82.02)		
General	1449 (13.59)	1092 (13.06)	357 (13.54)		
Other	382 (3.58)	265 (3.30)	117 (4.44)		
Relationship with mother, n (%)				12.2 (2)	.34
Well	9712 (91.07)	7302 (90.96)	2410 (91.43)		
General	768 (7.20)	579 (7.21)	189 (7.17)		
Other	184 (1.73)	147 (1.83)	37 (1.40)		
Father’s education level, n (%)				1900.0 (3)	<.001
Primary school and below	2341 (21.95)	2142 (26.68)	199 (7.55)		
Junior high school	4993 (46.82)	4154 (51.74)	839 (31.83)		
High school	2318 (21.74)	1449 (18.05)	869 (32.97)		
College and higher	1012 (9.49)	283 (3.53)	729 (27.66)		

XSL•FO
RenderX

Variables	Total (N=10,664)	Rural (n=8028)	Urban (n=2636)	Statistics, chi-square (*df*)	*P* value
Mother's education level, n (%)				2200.0 (3)	<.001
Primary school and below	3392 (31.8)	3103 (38.65)	289 (10.96)		
Junior high school	4558 (42.74%)	3709 (46.20)	849 (32.21)		
High school	1799 (16.87)	994 (12.38)	805 (30.54)		
College and higher	915 (8.58)	222 (2.77)	693 (26.29)		
Region				111.8 (2)	<.001
Central region	4898(45.93%)	3920(48.83%)	978(37.10%)		
Eastern region	2265(21.24%)	1636(20.38%)	629(23.86%)		
Western region	3501(32.83%)	2472(30.79%)	1029(39.04%)		
Relative income, n (%)				550.0(2)	<.001
General	6247 (58.58)	4843 (60.33)	1404 (53.26)		
Rich	2539 (23.81)	1511 (18.82)	1028 (39)		
Poor	1878 (17.61)	1674 (20.8)	204 (7.74)		
Independent bedroom, n (%)				35.6 (1)	<.001
No	1754 (16.45)	1419 (17.68)	335 (12.71)		
Yes	8910 (83.55)	6609 (82.32)	2301 (87.29)		
Sexual abuse, mean (SD)	—[a]	1.289 (1.672)	1.436 (1.832)	−3.289[b]	.001
Negative life events, mean (SD)	—	52.734 (23.953)	49.596 (23.266)	5.876[c]	<.001
Internet addiction, mean (SD)	—	34.628 (13.777)	35.650 (14.056)	−3.290[c]	.001
Suicidal thoughts and behaviors, mean (SD)	—	0.648 (1.370)	0.805 (1.469)	−5.010[c]	<.001

[a]Not applicable.

[b]Rank-sum test.

[c]2-tailed *t* test.

Correlation Analyses

Table 2 presents the means, SD, and Pearson correlations of all study variables. Notably, SA was significantly positively correlated with NLE (r=0.174; P<.001), IA (r=0.199; P<.001), and STB (r=0.137; P<.001). Moreover, NLE were significantly positively correlated with IA (r=0.281; P<.001) and STB (r=0.173; P<.001). Finally, IA was significantly positively correlated with STB (r=0.197; P<.001). Significant correlations between the variables initially supported our hypotheses.

Table . Means, SDs, and correlation of sex abuse, negative life events, internet addiction, and suicidal thoughts and behaviors (STB; N=10,664).

Variable	Value, mean (SD)	1[a]	2[b]	3[c]	4[d]
Sexual abuse	1.330 (1.714)	1.000			
Negative life events	51.960 (23.822)	0.174[e]	1.000		
Internet addiction	34.88 (13.852)	0.199[e]	0.281[e]	1.000	
STB	0.690 (1.396)	0.137[e]	0.173[e]	0.197[e]	1.000

[a]Correlations between sexual abuse and itself, negative life events, internet addiction, and STB.

[b]Correlations between negative life events and itself, internet addiction, and STB.

[c]Correlations between internet addiction and itself and STB.

[d]Correlation of STB with itself.

[e]P<.001.

XSL•FO
RenderX

Chain Mediation Analysis

To control for the influence of potential confounding factors, covariates were included in all mediation effect test models. We constructed a chain mediation effect model based on different theoretical assumptions (Figure 3). Multicollinearity tests indicated that the variance inflation factor values for all predictors were well below the critical threshold of 5, suggesting no serious multicollinearity issues in the model (Table S2 in Multimedia Appendix 1).

Figure 3. The indirect effects of negative life events and internet addiction in the relationship between sexual abuse and suicidal thoughts and behaviors. a: *P*<.001.

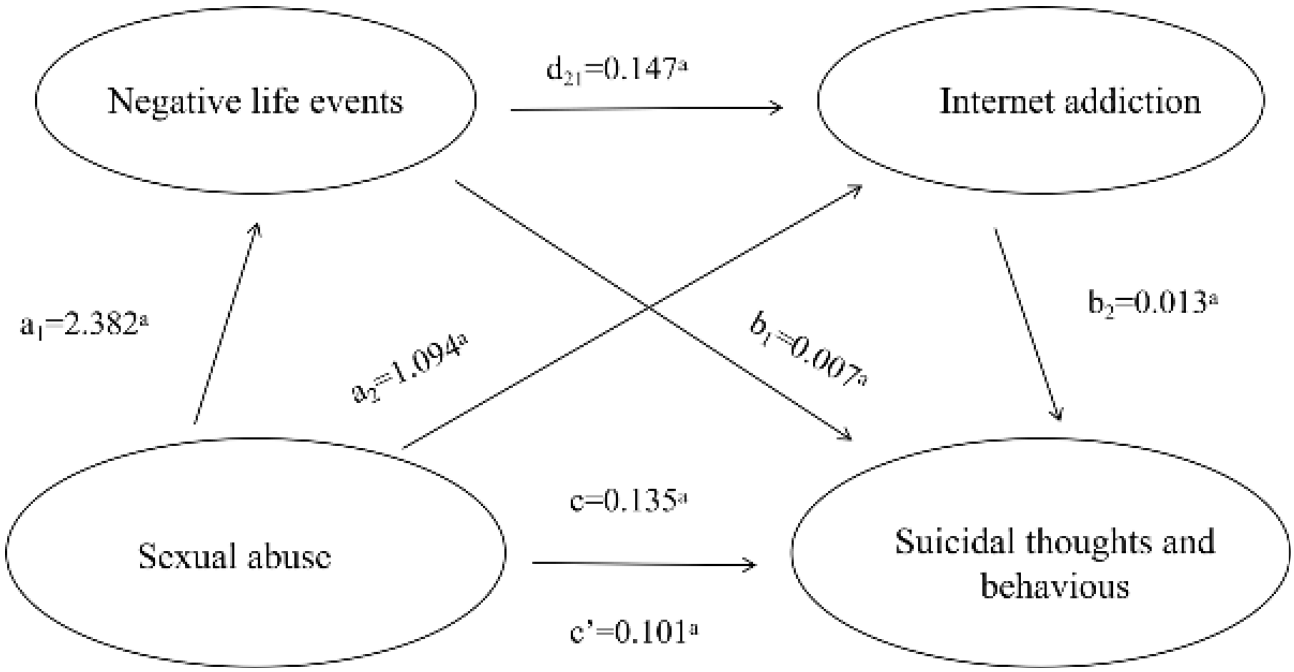

Figure 3 depicts the tested structural model, which statistically demonstrated significant associations among key variables. The model indicated that a higher level of SA was associated with higher levels of NLE (β=2.382, 95% CI 2.112 - 2.653; *P*<.001), IA (β=1.094, 95% CI 0.939 - 1.249; *P*<.001), and STB (β=0.135, 95% CI 0.119 - 0.151; *P*<.001). Higher levels of NLE were further associated with higher levels of IA (β=0.147, 95% CI 0.136 - 0.158; *P*<.001) and STB (β=0.007, 95% CI 0.005 - 0.008; *P*<.001). IA also showed a significant association with higher levels of STB (β=0.013, 95% CI 0.011 - 0.015; *P*<.001). Furthermore, as presented in Table 3, the direct effect of SA on STB was significant (β=0.101, 95% CI 0.085 - 0.117; *P*<.001), accounting for 74.81 % of the total effect. The indirect effect of SA on STB through NLE was significant (β=0.016, 95% CI 0.012 - 0.020; *P*<.001), accounting for 11.85% of the total effect; the indirect effect of SA on STB through IA was significant (β=0.014, 95% CI 0.010 - 0.018; *P*<.001), accounting for 10.37% of the total effect. Moreover, the indirect effect of SA on STB through NLE and IA was significant (β=0.004, 95% CI 0.003 - 0.006; *P*<.001), accounting for 2.96% of the total effect. All regression coefficients are unstandardized coefficients.

XSL•FO
RenderX

Table . Results of the chain mediation model analyzing the pathway from sexual abuse to suicidal thoughts and behaviors via negative life events and internet addiction.

Model pathways	β[a] (95% CI)	*P* value	Ratio (%)
Total effect	0.135 (0.119-0.151)	<.001	100.00
Direct effect	0.101 (0.085-0.117)	<.001	74.81
Total indirect effect	0.034 (0.029-0.040)	<.001	25.19
Sexual abuse →negative life events→ suicidal thoughts and behaviors	0.016 (0.012-0.020)	<.001	11.85
Sexual abuse → Internet addiction→ suicidal thoughts and behaviors	0.014 (0.010-0.018)	<.001	10.37
Sexual abuse →negative life events→ internet addiction →suicidal thoughts and behaviors	0.004 (0.003-0.006)	<.001	2.96

[a]Unstandardized regression coefficient.

Missing Data

Among the 18 variables analyzed in this study, the variables NLE, Age, gender, ethnic group, and region have no missing values. The proportion of missing data is presented in Table 4. The Little test for completely random missingness showed a significant result (χ^2_8=14670.975; *P*<.001).

Table . The proportion of missing data.

Variables	Missing value, n (%)
SA[a]	6756 (32.3)
NLE[b]	0 (0)
IA[c]	8017 (38.4)
STB[d]	6279 (30.1)
Age	0 (0)
Gender	0 (0)
Ethnic group	0 (0)
BMI	1994 (9.5)
Personality	5124 (24.5)
One child	4362 (20.9)
Parent’s marital status	6219 (29.8)
Relationship with father	385 (1.8)
Relationship with mother	375 (1.8)
Father’s education level	764 (3.7)
Mother’s education level	859 (4.1)
Region	0 (0)
Relative income	3316 (15.9)
Independent bedroom	6853 (32.8)

[a]SA: sexual abuse.

[b]NLE: negative life events.

[c]IA: internet addiction.

[d]STB: suicidal thoughts and behaviors.

Following multiple imputation of missing values, the sensitivity analysis yielded results consistent with the primary analysis, demonstrating robustness. The outcomes after imputation are presented in Table S3 in Multimedia Appendix 1.

XSL•FO
RenderX

Discussion

Principal Findings

This study revealed that STB were positively associated with SA among adolescents. Notably, NLE and IA independently mediated the path from SA to STB. Furthermore, NLE and IA played a serial mediating role in the relationship between SA and STB. The study constructs a new theoretical model with adolescents as the research object and provides guidance and recommendations to develop interventions for adolescent suicide.

SA and STB

This study's findings support hypothesis 1 that SA would be positively associated with STB among adolescents, suggesting that adolescents with a history of SA are at a higher risk of developing STB. Our results align with previous cross-sectional and longitudinal studies that noted a consistent association between childhood SA and an increased risk of suicide attempts [11,40,41]. Several biological and psychological theories elucidate how SA increases the risk of STB. According to neuroimaging investigations, childhood abuse is associated with reduced medial prefrontal cortex volume [42], which reduces an individual's cognitive functioning; the resultant cognitive deficits can directly increase suicide risk [43,44]. Regarding psychological theories, Finkelhor and Browne [45] identified 2 key trauma dynamics resulting from early SA, namely, stigmatization and feelings of betrayal—commonly experienced by survivors of early SA, especially when the abuser is a familiar member or caregiver [46]. These trauma dynamics catalyze depression, shame, and alienation from others, further eroding one's sense of self-worth and belonging [47]. These negative effects grow, fostering negative emotions and, after a point, self-harming behaviors and even STB [48].

Notably, gender differences exist in the prevalence of SA and STB among adolescent samples in China currently. Specifically, the probability of STB is higher among girls, while the proportion of experiencing SA is higher among boys [41,49,50]. This pattern likely does not reflect absolute differences in incidence rates but rather profoundly reveals gendered expressions of risk behaviors and sociocultural cognitive biases [51,52]. More concretely, the higher reported rate of suicidal ideation among women may be partly attributable to social norms that more readily permit them to acknowledge and express internalized emotional distress [53]. Conversely, the lower reported rate of SA among male individuals is more likely rooted in the widespread societal neglect of male survivors, the heightened stigma associated with disclosure, and the potential limitations of existing measurement tools in capturing trauma experiences specific to male survivors [50]. It is recommended to focus on the underidentified SA trauma among male adolescents and its potential externalizing behavioral consequences, as well as the clearer pathway from internalizing NLE to suicidal risk among female adolescents.

Mediating Effect of NLE

This study's findings verified that NLE mediated the relationship between SA and STB among adolescents, thus supporting hypothesis 2. To our knowledge, previous literature on the mediating role of NLE in the association between SA and STB is scant. Our study expands this field of research and fills a gap in this area. Notably, SA inflicts profound psychological trauma on survivors, which profoundly disrupts their daily lives and interpersonal relationships [54,55]. Furthermore, survivors may grapple with confusion and negative perceptions of their identities and sexual orientations, accompanied by feelings of shame and guilt [56,57]. Moreover, this trauma may impair survivors' academic, occupational, and daily functioning, as well as precipitate the onset of long-term health issues, including sexually transmitted infections, unintended pregnancies, chronic pain, and physical disabilities [54,58-62].

This study found that NLE were significantly positively correlated with STB, which is consistent with previous research [63-67]. Chronic stress from such events results in depression or despair [68], negative coping strategies, and, ultimately, suicidal thoughts [69]. Rumination may explain how NLE mediate the relationship between SA and STB, suggesting that individuals abused as children, without learning to manage their emotions, may have a reduced sense of control in adulthood [70]. Facing NLE, they tend to repeatedly consider negative outcomes, thus triggering STB.

Mediating Effect of IA

In this study, IA was identified as a mediator in the association between SA and STB among adolescents, thus corroborating hypothesis 3. Consistent with prior research, survivors of childhood abuse were found to be more prone to IA, with a strong link observed between SA and childhood trauma [71-73]. Per the social compensation hypothesis, the internet offers a suitable platform for self-expression for those who have faced prolonged NLE [74]. Furthermore, SA creates a dangerous environment for the survivor, thus resulting in numerous negative consequences [75]. Hence, they may hesitate to disclose their true selves in face-to-face interactions. Given the anonymous and concealed nature of the internet, these individuals are highly likely to turn to it as a platform for self-expression and release, which may, in turn, elevate their susceptibility to IA [76]. Additionally, IA may result in a variety of drastic physical, psychological, and social relation changes that adolescents can neither control nor avoid [77]. This cycle can trigger intense stress, potentially culminating in a mental health crisis, self-harm, and suicidal behaviors among adolescents. Moreover, the very online environments they frequent—through activities like gaming, socializing, or watching videos—are linked to increased suicide risk [78]. Within these spaces, discussions that normalize suicide or even share methods are not uncommon, potentially providing susceptible youths with both the impetus and the means for self-harm [79].

Serial Mediating Effect of NLE and IA

This study supported hypothesis 4, as the association between SA and STB was manifested through a serial mediation pathway involving NLE and IA. Adolescents who have experienced SA are more inclined to adopt negative coping styles when dealing with NLE, such as by frequently using the internet to escape reality, which increases their likelihood of receiving undesirable

XSL•FO
RenderX

information, experiencing negative moods, and exhibiting STB [27].

A significant positive correlation was observed between NLE and IA—consistent with previous research [80]. First, when faced with NLE, adolescents tend to adopt negative coping styles to alleviate negative emotions; adolescents commonly use the internet for emotional venting [80]. Owing to the lack of social support from family, friends, or other people, adolescents choose to use the internet to avoid the unpleasantness and pain of reality and forget their troubles temporarily [81]. Hence, IA may be a compensatory approach to deal with NLE [82]. Adolescents can seek help on the internet and gain a sense of accomplishment through gaming and entertainment. Adolescents who have experienced NLE are more likely to feel lonely, unaccepted, or have low self-confidence and may, thus, seek a sense of belonging and recognition online [67,83]. Such virtual social relationships may increase their feelings of dependence on the internet [84].

Suggestions for Future Research

Based on our findings, potential targets for intervention and multipathway mechanisms have been identified. In practice, we propose the following recommendations. First, in clinical assessments, adolescents with a history of SA should undergo routine screening for negative cognitive biases and problematic internet use patterns to facilitate the early identification of high-risk individuals [85]. Second, efforts to enhance adolescent psychological resilience should focus on helping youth modify their cognitive appraisal of NLE and preventing these appraisals from evolving into maladaptive behavioral patterns [86]. Third, modular content addressing IA should be incorporated into adolescent mental health initiatives. This content should aim to help adolescents understand the emotional needs underlying their internet use and collaboratively establish healthy alternative activities and time management plans [87].

Limitations

This study has several limitations. First, given the cross-sectional design, the mediation model we examined reveals patterns of statistical association among the variables but cannot establish strict causal direction or temporal sequence. Future research should adopt longitudinal or cross-lagged designs to more clearly elucidate the causal timing and dynamic interactions among these variables. Second, there may be reverse causality between the variables or spurious associations due to unmeasured variables. Third, the SA assessment tool used in this study is not a widely validated standardized instrument, and simply summing dichotomous items may not fully capture the complexity of SA experiences. Furthermore, this total score should not be interpreted as a strict interval scale.

Conclusions

SA is significantly correlated with STB among adolescents, and it indirectly exacerbates suicide risk through the serial mediation of NLE and IA.

This study possesses both theoretical and practical value in the prevention of suicide among adolescents. First, unlike previous research that has largely focused on isolated risk factors or single mediators, this study innovatively constructs and validates a chain mediation model, which reveals the chained mediating roles of NLE and IA between SA and STB. Second, this study systematically elucidates the sequential pathway mechanism from early trauma to psychosocial stress, then to behavioral problems, and ultimately to suicide risk, offering a novel perspective on the complex developmental process from early adversity to extreme outcomes. Finally, the findings provide clear direction and scientific evidence for the development of comprehensive suicide prevention strategies. Future efforts should focus on the prevention of SA and trauma intervention while simultaneously strengthening adolescents' cognitive reappraisal of negative events and guiding their online behavior.

Acknowledgments

The authors thank all study participants. The authors declare that no generative artificial intelligence tools were used in the writing of any portion of this manuscript.

Funding

This work was supported by the Talent Plan of Naval Medical University (grant 2025) and National Ministry of Education Collaborative Education Reform Project (grant 2411052361123306). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the study; and decision to submit the study for publication.

Data Availability

The dataset analyzed during this study is publicly available in the Science Database of the People Mental Health survey conducted by the National Population Health Data Archive. The dataset is available online [31].

Authors' Contributions

Conceptualization: QJ, LY
Data curation: JZ, QJ, FZ
Formal analysis: JZ, FZ, SY, LY
Methodology: HL, JS, ZZ, LL

XSL•FO
RenderX

Supervision: ZZ, LL
Validation: JZ, QJ, FZ, HL, SY, JS, LY
Visualization: JZ, QJ, HL, JS
Writing – original draft: JZ, QJ, FZ, LY
Writing – review & editing: ZZ, LL, LY
JZ, QJ, and FZ contributed equally as co-first authors.
ZZ, LL, and LY contributed equally as co-corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of gender differences in participants, multicollinearity test results, and chained mediation imputation results.
[DOCX File, 19 KB - jmir_v28i1e85371_app1.docx]

Checklist 1
JARS checklist.
[DOCX File, 20 KB - jmir_v28i1e85371_app2.docx]

Checklist 2
ASLEC checklist.
[DOCX File, 14 KB - jmir_v28i1e85371_app3.docx]

References

1. Lipson SK, Zhou S, Abelson S, et al. Trends in college student mental health and help-seeking by race/ethnicity: findings from the National Healthy Minds Study, 2013-2021. J Affect Disord 2022 Jun 1;306:138-147. [doi: 10.1016/j.jad.2022.03.038] [Medline: 35307411]
2. Mathieu S, de Leo D, Koo YW, Leske S, Goodfellow B, Kõlves K. Suicide and suicide attempts in the Pacific Islands: a systematic literature review. Lancet Reg Health West Pac 2021;17:100283. [doi: 10.1016/j.lanwpc.2021.100283] [Medline: 34734201]
3. Pan CH, Lee MB, Wu CY, Liao SC, Chan CT, Chen CY. Suicidal ideation, psychopathology, and help-seeking in 15 to 19-year-old adolescents in Taiwan: a population-based study 2015-2019. J Affect Disord 2021 Mar 1;282:846-851. [doi: 10.1016/j.jad.2020.12.139] [Medline: 33601727]
4. Suicide worldwide in 2019: global health estimates. : World Health Organization; 2021 URL: https://iris.who.int/server/api/core/bitstreams/3bd4ac79-4347-420e-b675-948d36ab3d90/content [accessed 2026-02-24]
5. Glenn CR, Kleiman EM, Kellerman J, et al. Annual research review: a meta-analytic review of worldwide suicide rates in adolescents. J Child Psychol Psychiatry 2020 Mar;61(3):294-308. [doi: 10.1111/jcpp.13106] [Medline: 31373003]
6. Altavini CS, Asciutti APR, Santana GL, et al. Suicide ideation among Brazilian college students: relationship with academic factors, mental health, and sexual abuse. J Affect Disord 2023 May 15;329:324-334. [doi: 10.1016/j.jad.2023.02.112] [Medline: 36849006]
7. Mościcki EK. Epidemiology of completed and attempted suicide: toward a framework for prevention. Clin Neurosci Res 2001 Nov;1(5):310-323. [doi: 10.1016/S1566-2772(01)00032-9]
8. Zhou Y, Liang Y, Cheng J, Zheng H, Liu Z. Child maltreatment in Western China: demographic differences and associations with mental health. Int J Environ Res Public Health 2019 Sep 26;16(19):3619. [doi: 10.3390/ijerph16193619] [Medline: 31561624]
9. Schröder J, Nick S, Richter-Appelt H, Briken P. Psychiatric impact of organized and ritual child sexual abuse: cross-sectional findings from individuals who report being victimized. Int J Environ Res Public Health 2018 Oct 31;15(11):2417. [doi: 10.3390/ijerph15112417] [Medline: 30384461]
10. Berens AE, Jensen SKG, Nelson CA. Biological embedding of childhood adversity: from physiological mechanisms to clinical implications. BMC Med 2017 Jul 20;15(1):135. [doi: 10.1186/s12916-017-0895-4] [Medline: 28724431]
11. Ng QX, Yong BZJ, Ho CYX, Lim DY, Yeo WS. Early life sexual abuse is associated with increased suicide attempts: an update meta-analysis. J Psychiatr Res 2018 Apr;99:129-141. [doi: 10.1016/j.jpsychires.2018.02.001] [Medline: 29454220]
12. O'Connor RC, Kirtley OJ. The integrated motivational-volitional model of suicidal behaviour. Philos Trans R Soc Lond B Biol Sci 2018 Sep 5;373(1754):20170268. [doi: 10.1098/rstb.2017.0268] [Medline: 30012735]
13. Laglaoui Bakhiyi C, Jaussent I, Beziat S, et al. Positive and negative life events and reasons for living modulate suicidal ideation in a sample of patients with history of suicide attempts. J Psychiatr Res 2017 May;88:64-71. [doi: 10.1016/j.jpsychires.2016.12.022] [Medline: 28088052]

XSL•FO
RenderX

14. Wang M, Kou C, Bai W, et al. Prevalence and correlates of suicidal ideation among college students: a mental health survey in Jilin Province, China. J Affect Disord 2019 Mar 1;246:166-173. [doi: 10.1016/j.jad.2018.12.055] [Medline: 30583141]
15. Kang C, Zheng Y, Yang L, et al. Prevalence, risk factors and clinical correlates of suicidal ideation in adolescent patients with depression in a large sample of Chinese. J Affect Disord 2021 Jul 1;290:272-278. [doi: 10.1016/j.jad.2021.04.073] [Medline: 34015621]
16. Zheng H, Liu D, Cheng J, Wang DB, Liu Y, Wu Y. Negative life events increase the risk of suicidal ideation in 6653 Chinese freshmen: from a 1-year longitudinal study. J Affect Disord 2022 Feb 15;299:604-609. [doi: 10.1016/j.jad.2021.12.039] [Medline: 34942231]
17. Xiong A, Liao S, Luo B, Luo S, Tong Y, Li Z. Associations between problematic internet use, life satisfaction, and deliberate self-harm among Chinese adolescents: a multi-centered longitudinal study. Addict Behav 2023 Dec;147:107808. [doi: 10.1016/j.addbeh.2023.107808] [Medline: 37536220]
18. Carr CP, Martins CMS, Stingel AM, Lemgruber VB, Juruena MF. The role of early life stress in adult psychiatric disorders: a systematic review according to childhood trauma subtypes. J Nerv Ment Dis 2013 Dec;201(12):1007-1020. [doi: 10.1097/NMD.0000000000000049] [Medline: 24284634]
19. Hébert M, Daigneault I, Vancamp T. Chapitre 5 agression sexuelle et risque de revictimisation à l'Adolescence: modèles conceptuels et défis liés à la prévention. In: L'agression Sexuelle Envers Les Enfants [Book in French]: Presses de l'Université du Québec; 2012:171-224. [doi: 10.2307/j.ctv18ph433.8]
20. Pan PY, Yeh CB. Internet addiction among adolescents may predict self-harm/suicidal behavior: a prospective study. J Pediatr 2018 Jun;197:262-267. [doi: 10.1016/j.jpeds.2018.01.046] [Medline: 29550226]
21. Kim H, Jhon M, Kim JW, et al. Suicide and non-suicidal self-injury from internet addiction among Korean adolescents. J Korean Acad Child Adolesc Psychiatry 2023 Jul 1;34(3):152-158. [doi: 10.5765/jkacap.230015] [Medline: 37426833]
22. Turner S, Harvey C, Hayes L, et al. Childhood adversity and clinical and psychosocial outcomes in psychosis. Epidemiol Psychiatr Sci 2019 Dec 16;29:e78. [doi: 10.1017/S2045796019000684] [Medline: 31839014]
23. Kessler RC, McLaughlin KA, Green JG, et al. Childhood adversities and adult psychopathology in the WHO World Mental Health Surveys. Br J Psychiatry 2010 Nov;197(5):378-385. [doi: 10.1192/bjp.bp.110.080499] [Medline: 21037215]
24. Schwaninger PV, Mueller SE, Dittmann R, et al. Patients with non-substance-related disorders report a similar profile of childhood trauma experiences compared to heroin-dependent patients. Am J Addict 2017 Apr;26(3):215-220. [doi: 10.1111/ajad.12512] [Medline: 28240800]
25. Lu L, Jian S, Dong M, et al. Childhood trauma and suicidal ideation among Chinese university students: the mediating effect of Internet addiction and school bullying victimisation. Epidemiol Psychiatr Sci 2020 Aug 10;29:e152. [doi: 10.1017/S2045796020000682] [Medline: 32772993]
26. Huang Q, Li J, Wang J, Liu B. Negative life events and internet addiction among college students: role of physical exercise and prosocial behavior. Front Psychol 2025;16:1629818. [doi: 10.3389/fpsyg.2025.1629818] [Medline: 41040106]
27. Yang F, Jiang L, Miao J, et al. The association between non-suicidal self-injury and negative life events in children and adolescents in underdeveloped regions of south-western China. PeerJ 2022;10:e12665. [doi: 10.7717/peerj.12665] [Medline: 35287346]
28. Zeng Z, Hu Y, Peng L, et al. The effect of negative life events on college students' depression: the mediating role of internet addiction and the moderating role of 5-HTT1A gene rs6449693 polymorphism. Curr Psychol 2023 Mar;42(8):6555-6563. [doi: 10.1007/s12144-021-01888-0]
29. Xiao L. Relationship among inclination of internet addiction, life events and self-efficacy in college students [Master's Thesis] [Article in Chinese]. : Hangzhou Normal University; 2012 URL: https://oversea.cnki.net/kcms2/article/
[accessed 2026-02-24]
30. Li Y, Sun J, Hu QY, Li DS, Lu HQ. Effects of the life events on internet addiction disorder in undergraduate students [Article in Chinese]. Mod Prev Med 2012;39(23):6207-6209 [FREE Full text]
31. Science database of people mental health [Article in Chinese]. Population Health Data Archive. URL: https://www.ncmi.cn/phda/dataDetails.do?id=CSTR:17970.11.A001U.202204.281.V1.0 [accessed 2026-02-24]
32. 2023 China health and wellness statistical yearbook. : National Health Commission of the People's Republic of China; 2025 URL: https://www.nhc.gov.cn/mohwsbwstjxxzx/tjtjnj/202501/8193a8edda0f49df80eb5a8ef5e2547c/files/1740022743894_10341.pdf [accessed 2026-02-28]
33. Liu XC, Oda S, Peng X, Asai K. Life events and anxiety in Chinese medical students. Soc Psychiatry Psychiatr Epidemiol 1997 Feb;32(2):63-67. [doi: 10.1007/BF00788922] [Medline: 9050346]
34. Yang T, He Y, Wu S, Cui X, Luo X, Liu J. Association between schizoid tendencies and aggressive behaviors: mediating and moderating influences in childhood trauma and life events among Chinese adolescents. Ann Gen Psychiatry 2021 Dec 3;20(1):51. [doi: 10.1186/s12991-021-00371-1] [Medline: 34861879]
35. Young KS. Internet addiction: the emergence of a new clinical disorder. Cyberpsychol Behav 1998 Jan;1(3):237-244. [doi: 10.1089/cpb.1998.1.237]
36. Lai CM, Mak KK, Watanabe H, Ang RP, Pang JS, Ho RCM. Psychometric properties of the internet addiction test in Chinese adolescents. J Pediatr Psychol 2013 Aug;38(7):794-807. [doi: 10.1093/jpepsy/jst022] [Medline: 23671059]

XSL•FO
RenderX

37. Tahir MJ, Malik NI, Ullah I, et al. Internet addiction and sleep quality among medical students during the COVID-19 pandemic: a multinational cross-sectional survey. PLoS ONE 2021;16(11):e0259594. [doi: 10.1371/journal.pone.0259594] [Medline: 34739502]
38. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
39. Journal Article Reporting Standards (JARS). American Psychological Association. URL: https://apastyle.apa.org/jars [accessed 2026-02-09]
40. Ferahkaya H, Türker M, Coşkun F, Uzun N, Akça Ö. Predictors of suicide attempt among subjects who were previously exposed to sexual abuse. Int J Soc Psychiatry 2025 Sep;71(6):1155-1163. [doi: 10.1177/00207640251334268] [Medline: 40249755]
41. Alhassan MA, Alarabi MA, Albalawi WM, et al. Prevalence and risk factors of suicidal ideation and suicide attempts among young people in the MENAT region: a systematic review and meta-analysis. BMC Psychiatry 2025 Nov 14;25(1):1149. [doi: 10.1186/s12888-025-07585-7] [Medline: 41257638]
42. van Harmelen AL, van Tol MJ, van der Wee NJA, et al. Reduced medial prefrontal cortex volume in adults reporting childhood emotional maltreatment. Biol Psychiatry 2010 Nov 1;68(9):832-838. [doi: 10.1016/j.biopsych.2010.06.011] [Medline: 20692648]
43. Szanto K. Cognitive deficits: underappreciated contributors to suicide. Am J Geriatr Psychiatry 2017 Jun;25(6):630-632. [doi: 10.1016/j.jagp.2017.02.012] [Medline: 28302439]
44. Olsson P, Wiktorsson S, Sacuiu S, et al. Cognitive function in older suicide attempters and a population-based comparison group. J Geriatr Psychiatry Neurol 2016 May;29(3):133-141. [doi: 10.1177/0891988715627015] [Medline: 26810564]
45. Finkelhor D, Browne A. The traumatic impact of child sexual abuse: a conceptualization. Am J Orthopsychiatry 1985 Oct;55(4):530-541. [doi: 10.1111/j.1939-0025.1985.tb02703.x] [Medline: 4073225]
46. Cahill C, Llewelyn SP, Pearson C. Long-term effects of sexual abuse which occurred in childhood: a review. Br J Clin Psychol 1991 May;30(2):117-130. [doi: 10.1111/j.2044-8260.1991.tb00927.x] [Medline: 2059744]
47. Hassanpour P, Buchwald S, Mehta AHP, Goldberg SB, Walsh K. Sexual violence and shame: a meta-analysis. Trauma Violence Abuse 2026 Jan;27(1):240-255. [doi: 10.1177/15248380241308828] [Medline: 39819226]
48. Sadath A, Kavalidou K, McMahon E, Malone K, McLoughlin A. Associations between humiliation, shame, self-harm and suicidality among adolescents and young adults: a systematic review. PLoS One 2024;19(2):e0292691. [doi: 10.1371/journal.pone.0292691] [Medline: 38329967]
49. Lai W, Wu H, Yang L, et al. Trends in prevalence of suicidal ideation, suicide attempt and sleep disturbance among Chinese high school students, 2014-2023. Eur Child Adolesc Psychiatry 2025 Nov;34(11):3475-3484. [doi: 10.1007/s00787-025-02762-1] [Medline: 40450187]
50. Abreu L, Hecker T, Goessmann K, Abioye TO, Olorunlambe W, Hoeffler A. Prevalence and correlates of sexual violence against adolescents: quantitative evidence from rural and urban communities in South-West Nigeria. PLOS Glob Public Health 2025;5(2):e0004223. [doi: 10.1371/journal.pgph.0004223] [Medline: 39933016]
51. Chaplin TM, Aldao A. Gender differences in emotion expression in children: a meta-analytic review. Psychol Bull 2013 Jul;139(4):735-765. [doi: 10.1037/a0030737] [Medline: 23231534]
52. Cantor N, Kingsbury M, Warner E, et al. Young adult outcomes associated with adolescent suicidality: a meta-analysis. Pediatrics 2023 Mar 1;151(3):e2022058113. [doi: 10.1542/peds.2022-058113] [Medline: 36810672]
53. Chaplin TM. Gender and emotion expression: a developmental contextual perspective. Emot Rev 2015 Jan;7(1):14-21. [doi: 10.1177/1754073914544408] [Medline: 26089983]
54. Guiney H, Caspi A, Ambler A, et al. Childhood sexual abuse and pervasive problems across multiple life domains: findings from a five-decade study. Dev Psychopathol 2024 Feb;36(1):219-235. [doi: 10.1017/S0954579422001146] [Medline: 36562290]
55. Serrano-Rodríguez E, Luque-Ribelles V, Hervías-Parejo V. Psychosocial consequences of sexual assault on women: a scoping review. Arch Sex Behav 2025 Jan;54(1):231-258. [doi: 10.1007/s10508-024-03013-1] [Medline: 39477846]
56. McElvaney R, Lateef R, Collin-Vézina D, Alaggia R, Simpson M. Bringing shame out of the shadows: identifying shame in child sexual abuse disclosure processes and implications for psychotherapy. J Interpers Violence 2022 Oct;37(19-20):NP18738-NP18760. [doi: 10.1177/08862605211037435] [Medline: 34459692]
57. Guyon R, Fernet M, Couture S, Tardif M, Cousineau MM, Godbout N. “Finding my worth as a sexual being”: a qualitative gender analysis of sexual self-concept and coping in survivors of childhood sexual abuse. Arch Sex Behav 2024 Jan;53(1):341-357. [doi: 10.1007/s10508-023-02693-5] [Medline: 37747584]
58. Ochoa MK, Constantin K. Impacts of child sexual abuse: the mediating role of future orientation on academic outcomes. Child Abuse Negl 2023 Nov;145:106437. [doi: 10.1016/j.chiabu.2023.106437] [Medline: 37659153]
59. Jonsdottir SD, Thordardottir EB, Valdimarsdottir UA, et al. Sexual violence in the workplace and associated health outcomes: a nationwide, cross-sectional analysis of women in Iceland. Lancet Public Health 2024 Jun;9(6):e365-e375. [doi: 10.1016/S2468-2667(24)00075-6] [Medline: 38821683]

XSL•FO
RenderX

60. Skjælaaen K, Nesvold H, Brekke M, et al. Sexually transmitted infections among patients attending a sexual assault centre: a cohort study from Oslo, Norway. BMJ Open 2022 Dec 1;12(12):e064934. [doi: 10.1136/bmjopen-2022-064934] [Medline: 36456026]
61. Ajayi AI, Ezegbe HC. Association between sexual violence and unintended pregnancy among adolescent girls and young women in South Africa. BMC Public Health 2020 Sep 7;20(1):1370. [doi: 10.1186/s12889-020-09488-6] [Medline: 32894130]
62. Bussières A, Hancock MJ, Elklit A, et al. Adverse childhood experience is associated with an increased risk of reporting chronic pain in adulthood: a stystematic review and meta-analysis. Eur J Psychotraumatol 2023;14(2):2284025. [doi: 10.1080/20008066.2023.2284025] [Medline: 38111090]
63. Chan S, Miranda R, Surrence K. Subtypes of rumination in the relationship between negative life events and suicidal ideation. Arch Suicide Res 2009;13(2):123-135. [doi: 10.1080/13811110902835015] [Medline: 19363749]
64. Chang EC, Sanna LJ, Hirsch JK, Jeglic EL. Loneliness and negative life events as predictors of hopelessness and suicidal behaviors in Hispanics: evidence for a diathesis-stress model. J Clin Psychol 2010 Dec;66(12):1242-1253. [doi: 10.1002/jclp.20721] [Medline: 20734320]
65. Lannoy S, Mars B, Heron J, Edwards AC. Suicidal ideation during adolescence: the roles of aggregate genetic liability for suicide attempts and negative life events in the past year. J Child Psychol Psychiatry 2022 Oct;63(10):1164-1173. [doi: 10.1111/jcpp.13653] [Medline: 35766310]
66. Glenn CR, Kleiman EM, Kandlur R, Esposito EC, Liu RT. Thwarted belongingness mediates interpersonal stress and suicidal thoughts: an intensive longitudinal study with high-risk adolescents. J Clin Child Adolesc Psychol 2022;51(3):295-311. [doi: 10.1080/15374416.2021.1969654] [Medline: 34570668]
67. Li Y, Shen X, Wang JL. Association between negative life events and suicidal behavior in adolescents: roles of core self-evaluation, depression and gender. Soc Psychiatry Psychiatr Epidemiol 2025 Mar;60(3):673-684. [doi: 10.1007/s00127-024-02725-y] [Medline: 39001887]
68. Zhao SY, Ren RR, Chen W. A mediated moderation model of negative life events, self-esteem, rumination and parental divorce on depression among Chinese juvenile delinquents. Sci Rep 2023 Jan 31;13(1):1793. [doi: 10.1038/s41598-023-28626-9] [Medline: 36720911]
69. Miranda R, Tsypes A, Gallagher M, Rajappa K. Rumination and hopelessness as mediators of the relation between perceived emotion dysregulation and suicidal ideation. Cogn Ther Res 2013 Aug;37(4):786-795. [doi: 10.1007/s10608-013-9524-5]
70. Nolen-Hoeksema S, Wolfson A, Mumme D, Guskin K. Helplessness in children of depressed and nondepressed mothers. Dev Psychol 1995;31(3):377-387. [doi: 10.1037/0012-1649.31.3.377]
71. Jones L, Hastings RP, Totsika V, Keane L, Rhule N. Child behavior problems and parental well-being in families of children with autism: the mediating role of mindfulness and acceptance. Am J Intellect Dev Disabil 2014 Mar;119(2):171-185. [doi: 10.1352/1944-7558-119.2.171] [Medline: 24679352]
72. Wang H, Luo W, Huang W, et al. How sleep duration mediated childhood trauma and internet addiction of the vocational college population in urban areas of south China. Front Psychiatry 2022;13:1088172. [doi: 10.3389/fpsyt.2022.1088172] [Medline: 36713913]
73. Hsieh YP, Shen ACT, Wei HS, Feng JY, Huang SCY, Hwa HL. Associations between child maltreatment, PTSD, and internet addiction among Taiwanese students. Comput Human Behav 2016 Mar;56:209-214. [doi: 10.1016/j.chb.2015.11.048]
74. McKenna KYA, Green AS, Gleason MEJ. Relationship formation on the internet: what's the big attraction? J Soc Issues 2002 Jan;58(1):9-31. [doi: 10.1111/1540-4560.00246]
75. Laconi S, Kaliszewska-Czeremska K, Gnisci A, et al. Cross-cultural study of problematic internet use in nine European countries. Comput Human Behav 2018 Jul;84:430-440. [doi: 10.1016/j.chb.2018.03.020]
76. Mayer ED, Peterson R, Kim E, et al. What survivors of sexual violence want when disclosing their experiences in person or online: qualitative interview analysis. J Med Internet Res 2025 Sep 5;27:e73497. [doi: 10.2196/73497] [Medline: 40911341]
77. Machulska A, Kleinke K, Eiler TJ, et al. Retraining automatic action tendencies for smoking using mobile phone–based approach-avoidance bias training: a study protocol for a randomized controlled study. Trials 2019 Dec 12;20(1):720. [doi: 10.1186/s13063-019-3835-0] [Medline: 31831080]
78. Shinetsetseg O, Jung YH, Park YS, Park EC, Jang SY. Association between smartphone addiction and suicide. Int J Environ Res Public Health 2022 Sep 15;19(18):11600. [doi: 10.3390/ijerph191811600] [Medline: 36141872]
79. Grant-Allen G, Wang L, Amini J, Dhaliwal S, Sinyor M, Mitchell RH. Self-harm and suicide-related content on TikTok: thematic analysis. J Med Internet Res 2025 Sep 18;27:e77828. [doi: 10.2196/77828] [Medline: 40811047]
80. Wang X, Ding T, Lai X, Yang C, Luo J. Negative life events, negative copying style, and internet addiction in middle school students: a large two-year follow-up study. Int J Ment Health Addict 2023 Mar 28;2023:1-11. [doi: 10.1007/s11469-023-01045-7] [Medline: 37363770]
81. Zhang Y, Qi Y, Ma Y. The associations between social support and problematic mobile phone use among children and adolescents: a three-level meta-analysis. J Youth Adolesc 2025 Jan;54(1):158-175. [doi: 10.1007/s10964-024-02055-x] [Medline: 39033221]

XSL•FO
RenderX

82. Cheng J, Peng C, Rong F, Wang Y, Tan Y, Yu Y. Mobile phone addiction and suicide behaviors among Chinese adolescents: the mediation of poor sleep quality. J Behav Addict 2024 Mar 26;13(1):88-101. [doi: 10.1556/2006.2023.00078] [Medline: 38224348]
83. Hall JA, Pennington N, Merolla AJ. Which mediated social interactions satisfy the need to belong? J Comput Mediat Commun 2022 Nov 4;28(1). [doi: 10.1093/jcmc/zmac026]
84. Li J, Zhan D, Zhou Y, Gao X. Loneliness and problematic mobile phone use among adolescents during the COVID-19 pandemic: the roles of escape motivation and self-control. Addict Behav 2021 Jul;118:106857. [doi: 10.1016/j.addbeh.2021.106857] [Medline: 33676160]
85. Lam KYI. Disclosure and psychological well-being of sexually abused adolescents in Hong Kong. J Child Sex Abus 2015;24(7):731-752. [doi: 10.1080/10538712.2015.1077364] [Medline: 26479960]
86. Zhang J, Li Y, Li J, et al. The effect of life events, resilience, self-esteem, and coping styles on aggressive behavior among left-behind adolescents: structural equation modeling. Front Psychiatry 2023;14:991608. [doi: 10.3389/fpsyt.2023.991608] [Medline: 36793944]
87. He X, Chen S, Yu Q, Yang P, Yang B. Correlations between problematic internet use and suicidal behavior among Chinese adolescents: a systematic review and meta-analysis. Front Psychiatry 2024;15:1484809. [doi: 10.3389/fpsyt.2024.1484809] [Medline: 39588549]

Abbreviations

ASLEC: Adolescent Self-Rating Life Event Checklist
IA: internet addiction
IAT: Internet Addiction Test
JARS: Journal Article Reporting Standards
KMO: Kaiser-Meyer-Olkin
NLE: negative life events
SA: sexual abuse
STB: suicidal thoughts and behaviors
WHO: World Health Organization

Edited by S Brini; submitted 07.Oct.2025; peer-reviewed by A AL-Asadi, L Liu; accepted 06.Feb.2026; published 25.Mar.2026.

Please cite as:
Zhu J, Jiang Q, Zhang F, Liu H, Yu S, Sun J, Zhao Z, Liu L, Yuan L
Suicidal Thoughts and Behaviors Among Chinese Adolescents in Relation to Negative Life Events, Internet Addiction, and Sexual Abuse: Cross-Sectional Study
J Med Internet Res 2026;28:e85371
URL: https://www.jmir.org/2026/1/e85371
doi:10.2196/85371

© Juanfang Zhu, Qinqin Jiang, Feng Zhang, Hao Liu, Shaojie Yu, Jinhai Sun, Zhe Zhao, Lijuan Liu, Lei Yuan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring Reddit Discourse and Information Needs Surrounding Extreme Heat: Topic, Sentiment, and Engagement Analysis

Melissa MacKay[1], MPH, PhD; Soroush Zamani Alavijeh[2], MSc; Sydney Gosselin[1], MSc; Fattane Zarrinkalam[3], PhD; Jennifer E McWhirter[1], PhD

[1]Department of Population Medicine, University of Guelph, 50 Stone Rd E, Guelph, ON, Canada

[2]Department of Computing and Software, McMaster University, Hamilton, ON, Canada

[3]College of Engineering, University of Guelph, Guelph, ON, Canada

Corresponding Author:
Jennifer E McWhirter, PhD
Department of Population Medicine, University of Guelph, 50 Stone Rd E, Guelph, ON, Canada

Abstract

Background: As Canada's climate changes, extreme heat events have become more frequent, a trend that is expected to continue. Extreme heat can lead to several negative health outcomes, which disproportionately impact vulnerable populations. Evidence-based, equitable interventions are needed to inform and protect the public from the health effects. Effective communication can aid this effort to improve health outcomes by emphasizing the connection between health risks and climate change and empowering people to act. Machine learning has applications in understanding current attitudes, beliefs, experiences, and behaviors within the target audience for public health messaging. Machine learning analyses of social media data have elucidated user perceptions of heat events in the literature; however, research is limited with respect to social media user perceptions, beliefs, and behaviors related to extreme heat, particularly in the Canadian context. Analyzing Canadian social media discourse related to extreme heat will help to address this research gap and inform future research and communications to reduce the risks of extreme heat.

Objective: The purpose of this research is to better understand Canadian discourse and emotions related to extreme heat by examining social media (Reddit). Our objectives include (1) identifying common discussion topics, concerns, and questions related to extreme heat among Canadian Reddit users; (2) analyzing sentiment and emotional responses to extreme heat discussions; and (3) investigating the relationship between topics, sentiment, and engagement for posts.

Methods: We collected data using the Reddit application programming interface (API), retrieving posts from 30 Canada-specific subreddits between February 12, 2023, and February 11, 2024, based on a predefined set of heat- and climate-related keywords. Posts and comments were structured as hierarchical tree models, with text consolidated into documents for analysis. Topic modeling, sentiment analysis, and emotion analysis were conducted; engagement was assessed using net upvote scores to gauge community approval.

Results: The analysis of 607 Reddit posts from 15,366 users revealed that discussions about extreme heat were most frequently centered around the keyword "heat," which appeared in 82.5% (n=501) of the posts and 81.1% (n=25,253) of the comments. Topic analysis identified key themes related to heating and cooling costs, weather records, air conditioning, and health impacts, while sentiment and emotion analyses showed varying levels of positivity and negativity across subreddits.

Conclusions: Our findings present an initial snapshot into Canadian perspectives and information needs about extreme heat in Canada. In our sample, discussions on Reddit about extreme heat in Canada are dominated by concerns over heating and cooling costs, weather patterns, and personal adaptation strategies, reflecting both practical and policy-related challenges. Additionally, sentiment and emotion analyses suggest significant regional differences in public perception, which may be useful for informing health and risk messaging initiatives to better protect Canadians from the adverse health effects of climate change.

(J Med Internet Res 2026;28:e82426) doi:10.2196/82426

KEYWORDS

social media; Reddit; climate change; extreme heat; information ecosystem; health communication

Introduction

Background

Climate change due to human activity has been driving shifts in temperatures and weather patterns [1]. Climate change has many consequences, including fires, flooding, droughts and water scarcity, and intense storms [1]. Because of these impacts, climate change affects health in diverse ways, and some individuals are more vulnerable to these health risks than others [1]. The extent of the impact on individuals is influenced by

XSL•FO
RenderX

their exposure to hazards related to climate change, the degree of impact, and their capacity to cope with the climate hazard [2]. Infants, children, pregnant women, older adults, and persons with physical and mental impairments are more vulnerable to climate-related hazards [2]. Additionally, various social determinants of health, including socioeconomic status, race or racism, health care access, education, and the built environment, interact with climate change to produce patterns of disadvantage that lead to additional vulnerabilities [2].

In Canada, climate change is resulting in extreme weather events, including extreme heat, which has significant direct impacts on human health and the environment [3]. Direct health risks from extreme heat include the exacerbation of existing health conditions, heat exhaustion, heat stroke, and death [4]. Extreme heat can also intersect with the indirect health risks of a changing climate, which are related to the social determinants of health and can impact our food system, culture, housing, infectious disease distribution, water quality and availability, and the economic status of individuals and communities [5]. Indeed, extreme heat is one of the most deadly forms of extreme weather affecting Canadians and is a growing public health risk in Canada [6,7].

Effective communication about climate change–related health risks, including extreme heat, can improve individual and population-level health outcomes [8]. Evidence-based strategies for climate change communication emphasize highlighting health impacts, relying on trusted sources, minimizing barriers to adaptive behaviors, and leveraging social media to disseminate accurate information [8]. To support effective public health messaging, it is essential to understand public risk perceptions as well as their information needs and preferences regarding climate change [8].

Social media provides an important communication channel through which public health can share information and people can share their experiences and opinions [9]. Public health can potentially reach a larger audience, raise awareness and share information about climate-related hazards and issues, and mobilize action. It can also foster public discussion and positively impact attitudes toward climate policy [9,10]. Social media also presents challenges, including the spread of misinformation or disinformation, the creation of echo chambers, and the bias toward certain actors or ideologies because of algorithmic curation [10].

Social media provides a platform for public health professionals to not only reach wider audiences but also to monitor information, assess public awareness, and engage in ongoing discussions. Platforms such as Reddit, which has over 15.55 million users in Canada [11], offer valuable insights into public perceptions and discourse on health-related topics. Canadian users actively participate across subreddits, including r/canada, facilitating country-specific discussions. The large Canadian user base and the diversity of subreddits and topics discussed provide an important social media platform to assess extreme heat and climate-related discussions and trends to understand perceptions, sentiment, and information needs and wants. An analysis of 16 years of Reddit discussion suggests that trends in Reddit discussions are reflective of broader trends across Internet-based channels [12], further supporting the evaluation of Reddit to better understand discourse on climate change.

Although limited research has been conducted to examine climate change discourse on Reddit [13,14], no such research has been conducted with a focus on Canada. This study extends prior research on climate change discourse by offering a Canadian case study and demonstrating the potential of computational methods for examining emerging public concerns about extreme heat. To guide our study, we draw on principles from risk communication theory and the concept of the information ecosystem. Risk communication theory emphasizes the role of trusted channels, perceived risk, and emotional responses in shaping adaptive behaviors during climate hazards [8]. The information ecosystem perspective recognizes social media platforms, such as Reddit, as sites where diverse actors contribute to meaning-making, often shaping how risk information is sought, shared, and acted upon [10]. These perspectives position Reddit not only as a source of descriptive discourse but also as a meaningful part of how Canadians understand, interpret, and respond to extreme heat.

Study Goal and Objectives

The goal of this research is to understand social media (Reddit) discourse and emotions related to extreme heat in Canada. Specific subobjectives of this research are as follows: (1) to identify common discussion topics, questions, and concerns related to extreme heat among Canadian Reddit users; (2) to analyze sentiment and emotional responses to extreme heat discussions; and (3) to investigate the relationship between topics, sentiment, and engagement for posts.

Methods

This study was designed as a descriptive computational analysis, intended to provide an initial overview of Reddit discourse on extreme heat in Canada. As such, the results present descriptive outputs of topic, sentiment, emotion, and engagement analyses without qualitative interpretation.

Ethical Considerations

Research ethics was not required as this study analyzed publicly available data only as per the University of Guelph's Research Ethics Board [15]. Under Canada's Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans, Research Ethics Board review is not required for research that relies exclusively on publicly accessible information where individuals have no reasonable expectation of privacy; as our study analyzes only public Reddit posts without interaction or private identifiers, ethics review is not required [16]. No personally identifiable information or direct identifiers (eg, usernames, links, or IDs) were collected or analyzed as part of this study. No other personally identifiable information was collected or stored, such as user profile information or IP addresses.

Data Source

To collect data for our analysis, we utilized the Reddit API [17] to search for posts within a selection of Canadian subreddits. These searches targeted posts that contained at least one of our

XSL•FO
RenderX

designated keywords in their titles. Specifically, we examined 30 subreddits focused on national interests (eg, r/canada), provincial or territorial interests (eg, r/ontario, r/britishcolumbia), and local interests within provincial or territorial capital cities and most populous cities (if different; eg, r/vancouver, r/halifax, r/saskatoon). The Reddit posts included were made between February 12, 2023, and February 11, 2024. The data collected and analyzed were publicly available Reddit posts; however, our analyses did not use any personally identifiable information or direct identifiers (eg, usernames, links, or IDs).

Data Acquisition

This research sought to understand prevalent topics, sentiment, and engagement information within Reddit posts and comments related to extreme heat events in Canada. Reddit users post content within designated communities, known as subreddits. When a user posts content in a subreddit, the other members of the subreddit can "upvote" (ie, indicate a favorable reaction), "downvote," (ie, indicate an unfavorable reaction), and comment on the post. Subreddits can be focused on any topic or interest, including national, provincial or territorial, or local news and issues.

The research team collected posts and comments from a predetermined list of Canada-specific subreddits using Reddit's API [17]. We used the Reddit API to search the selected subreddits for the following heat- and climate-related search terms: "heat," "extreme heat," "hot weather," "hot summer," "hottest day," "heatwave," "heat dome," "hot day," "hottest year," "climate change AND heat," and "BC heat wave." These terms were determined after searching the subreddits for the keywords used to discuss extreme heat.

To ensure that our keyword set and subreddit sample appropriately captured discussion relevant to extreme heat in Canada, we conducted preliminary pilot searches. We manually browsed Canadian subreddits to identify the actual terms users used when discussing extreme heat, which helped refine our candidate keywords (eg, "heat dome," "hottest day") beyond generic terms, such as "hot weather." For subreddit selection, we began with lists of national, provincial or territorial, and major city subreddits, then excluded those with very low activity (few heat-related posts during pilot months) or that were clearly irrelevant (eg, subreddits focused exclusively on entertainment or sports). This process ensured that the sample remained both manageable and relevant. Similar approaches are used in related social media studies, such as that by Janaswamy and Blackburn [18], who identified climate change–relevant subreddits by starting with seed communities and empirically validating activity levels, and that by Fariello and Jemielniak [12], who validated keyword sets through historical analyses of evolving climate-related terminology.

Text Analytics

The textual data from Reddit posts and comments follow a hierarchical tree structure. Each discussion thread on Reddit begins with a main post, which forms the root node of the tree. Subsequent comments within the thread serve as child nodes. Specifically, a comment that directly responds to the main post is attached as a child to the root node. The comments that are replies to other comments are linked as child nodes to their respective parent comments. This recursive linkage continues, mapping the entire discussion thread into a tree structure where the path from the root to any node represents the flow of conversation from the initial post to a particular comment. Finally, the comments that have no replies are leaves of the tree with no children.

After constructing the tree structures to represent individual Reddit threads, we performed topic analysis on these threads. To facilitate effective topic detection by the language model, it is essential to create more cohesive documents that encapsulate the main context and key information within the discussions. This was achieved by traversing the tree in a level-order fashion. During this traversal, the text of each visited node (comment) was concatenated to form a continuous document. This process continued until the document reached a predefined size suitable for the input constraints of the language model. In this way, the comments that are replies to a specific comment were sequentially concatenated, preserving the flow of discussion within the document. The document construction was terminated either when it reached its maximum allowable size or when there were no further comments to process. If a cutoff occurred due to reaching the maximum size, a new document was initiated following the same methodology, continuing until all comments in the thread were processed. If a single comment or a single post was longer than the maximum size, it was constructed as one document by itself. This process began with the main Reddit post itself, considering it as the parent node from which all subsequent comments in the thread were derived and processed.

To mitigate the sparsity challenges of short comments and to preserve conversational flow, we concatenated comments and replies into larger "documents." Similar strategies have been used in prior research to improve topic modeling and clustering of fragmented user-generated text. For example, Llewellyn et al [19] demonstrated that combining online newspaper comments with their replies produced more coherent topics when applying Latent Dirichlet Allocation . Likewise, Sun et al [20] developed a conversational structure-aware topic model, which leverages reply structures in threaded discussions to yield more interpretable topics. These approaches, like ours, show that aggregating related comments along conversational paths can enhance coherence while retaining contextual meaning that might be lost if each comment were treated entirely independently.

While traversing the 607 tree structures, we excluded comments generated by bots that were deleted or written in languages other than English. This filtering process removed 22,069 comments from the initial pool of 31,118, leaving 9049 valid comments along with the original 607 posts. These were consolidated into 1352 documents for analysis. On average, each document contained 270.56 words, aligning well with our language model's input requirements, which accommodate a maximum token size of 384.

Topic Analysis

For topic modeling, we used BERTopic (bidirectional encoder representations topic modeling), a neural topic modeling

XSL•FO
RenderX

technique, to conduct topic analysis on the documents extracted from Reddit, which were assembled using the document consolidation method outlined in the text analytics section. BERTopic leverages class-based term frequency-inverse document frequency for improved topic representation [21]. BERTopic combines transformer-based embeddings with hierarchical clustering, enabling coherent topic extraction even from short or noisy texts, making it well-suited for our dataset. To assess coherence, we manually inspected representative documents for each identified topic to verify that the extracted clusters reflected consistent underlying themes.

Sentiment Analysis

Sentiment analysis is used to ascertain the prevailing sentiment within textual data, specifically determining whether the sentiment of a given text is positive or negative, and assigning a score to quantify the intensity of the identified sentiment direction. We utilized the consolidated documents, assembled using the method described above, as input for the *distilbert-base-uncased-finetuned-sst-2-english* model [22,23]. This model was specifically designed for text classification tasks and was used to assess the sentiment direction and sentiment score for each document [24]. To provide a check on model performance, we also manually inspected a random sample of 50 documents classified as positive and 50 documents classified as negative. This validation suggested that the sentiment assignments generally aligned with the intended tone of the posts.

Emotion Analysis

Emotion analysis is conducted on textual data to detect and identify various emotional elements within the text. For emotion analysis, we used roberta-base-go_emotions, an off-the-shelf model fine-tuned on the GoEmotions dataset [25]. RoBERTa's optimized pretraining enhances contextual understanding, while fine-tuning on GoEmotions enables precise detection of fine-grained emotions [26,27]. GoEmotions dataset includes 12 positive, 11 negative, 4 ambiguous, and 1 neutral emotion, making it suitable for conversation and differentiating between subtle emotion expression [28]. Using this fine-tuned language model on the consolidated documents previously described, we discerned the level of negative (eg, anger, sadness, or fear) and positive (eg, admiration or optimism) emotions, as well as approval and disapproval emotions within the extracted Reddit documents.

Engagement Analysis

Reddit users often express their agreement or disagreement with a given post or comment through upvotes and downvotes, respectively. The net upvote score of a Reddit post or comment serves as an indicator of engagement and is determined by subtracting the number of downvotes from the number of upvotes [29]. A post or comment with a high net upvote score is generally considered to be approved by the subreddit community, whereas one with a low net upvote score is viewed as disapproved [29].

Results

Data Collection

Based on the search terms and included subreddits, 607 Reddit posts were collected from over 15,000 users. On average, there were approximately 20 posts per subreddit and 52 comments per post. Table 1 provides the total number of posts, users, comments, and distribution of posts across the subreddits.

Table . Total posts, unique users, comments, comments per post, subreddits, and posts per subreddit.

Characteristic	Value
Reddit posts	607
Unique users	15,366
Comment	31,118
Comments per post, SD[a]	81.9
Comments per post, mean[a]	52.8
Comments per post, median (IQR)[a]	24.00 (9-58)
Subreddit	30
Posts per subreddit, SD	17.0
Posts per subreddit, mean	20.2
Posts per subreddit, median (IQR)	18.5 (6-29.5)

[a]Based on posts with at least 1 comment.

Across the 11 keywords used to identify posts in the included subreddits, "heat" was the most common (501 posts, 25,253 comments), followed by "extreme heat" (28 posts, 1320 comments) and "hot weather" (19 posts, 1561 comments; Table 2).

XSL•FO
RenderX

Table . Post and comment distribution across keywords.

Keyword	Post, n (%)	Comment, n (%)
heat	501 (82.5)	25,253 (81.1)
extreme heat	28 (4.6)	1,320 (4.2)
hot weather	19 (3.1)	1,561 (5)
hot summer	17 (2.8)	622 (2)
hottest day	11 (1.8)	344 (1.1)
heatwave	8 (1.3)	348 (1.1)
heat dome	7 (1.2)	868 (2.8)
hot day	6 (1)	225 (0.7)
hottest year	5 (0.8)	363 (1.1)
climate change AND heat	3 (0.5)	212 (0.7)
BC heat wave	2 (0.3)	2 (<0.01%)

Across the included subreddits, r/canada had the most posts and comments (62 and 6010, respectively), followed by r/halifax (53 posts, 1772 comments), and r/britishcolumbia (47 posts; 2927 comments). Table 3 shows the distribution of the posts and comments across all included subreddits.

XSL•FO
RenderX

Table . Post and comment distribution across subreddits.

Subreddit[a]	Post, n (%)	Comment, n (%)
r/canada	62 (10.2)	6010 (19.3)
r/halifax	53 (8.7)	1772 (5.7)
r/britishcolumbia	47 (7.7)	2927 (9.4)
r/vancouver	45 (7.4)	2475 (8.0)
r/ontario	43 (7.1)	2194 (7.1)
r/Winnipeg	34 (5.6)	1379 (4.4)
r/onguardforthee	34 (5.6)	960 (3.1)
r/ottawa	31 (5.1)	1701 (5.5)
r/alberta	25 (4.1)	1949 (6.3)
r/VictoriaBC	25 (4.1)	890 (2.9)
r/NovaScotia	23 (3.8)	648 (2.1)
r/Edmonton	23 (3.8)	656 (2.1)
r/CanadaPolitics	21 (3.5)	545 (1.8)
r/Canada_sub	21 (3.5)	1671 (5.4)
r/Calgary	20 (3.3)	933 (3.0)
r/newfoundland	17 (2.8)	334 (1.1)
r/toronto	14 (2.3)	1575 (5.1)
r/saskatchewan	12 (2.0)	880 (2.8)
r/saskatoon	12 (2.0)	532 (1.7)
r/PEI	12 (2.0)	213 (0.7)
r/montreal	7 (1.2)	371 (1.2)
r/StJohnsNL	6 (1.0)	67 (0.2)
r/Manitoba	6 (1.0)	142 (0.5)
r/newbrunswickcanada	4 (0.7)	145 (0.5)
r/moncton	3 (0.5)	23 (0.1)
r/regina	2 (0.3)	75 (0.2)
r/fredericton	2 (0.3)	9 (<0.01%)
r/Yukon	1 (0.2)	33 (0.1)
r/NWT	1 (0.2)	9 (<0.01%)
r/Whitehorse	1 (0.2)	0 (0.0)

[a]Subreddit names are presented as originally created on Reddit. Capitalization varies across communities and does not necessarily reflect conventional geographic naming standards.

Topic Analysis

We extracted 14 meaningful topics from 1111 of the documents, characterized by their respective scores (Table 4). However, 241 documents were associated with a mixture of topics or contained topics that were not clearly recognizable and, as such, were invalid topics (topic -1 in Table 4). The overall topic distribution across all 1352 documents is presented in Table 4. Heating and cooling mechanics and costs associated with each appeared frequently throughout the topics.

XSL•FO
RenderX

Table . Topic and top word distribution across documents (n=1111 valid topics).

Topic	Top words	Values, n (%)
-1 (invalid topic)	__[a]	241 (17.8)[b]
0	pump, furnace, gas, heat, house	571 (51.4)[c]
1	weather, climate, rain, records, wave	218 (19.6)[c]
2	[air] conditioners, people, health...	51 (4.6)[c]
3	tax, carbon, moe, saskatchewan, money	42 (3.7)[c]
4	walmart, prices, food, grocery, costco	40 (3.6)[c]
5	landlord, rent, tenants, building...	35 (3.2)[c]
6	ac, windows, curtain, fans, sleep	29 (2.6)[c]
7	cat, dog, pets, fan, ice	26 (2.3)[c]
8	wear, hoodie, pants, wool, shoes	23 (2.1)[c]
9	grid, power, nuclear, shortages...	21 (1.9)[c]
10	valve, condo, baseboard, apartment...	19 (1.7)[c]
11	cool, pillow, ice, pools, healthlink	15 (1.4)[c]
12	love, hate, like, humidity, outside	11 (1.0)[c]
13	schools, manitoba, strike, grade...	10 (0.9)[c]

[a]Invalid topic excluded from the top word analysis.

[b]Denominator is all documents, n=1352.

[c]Denominator is all valid documents, n=1111.

Sentiment Analysis

Figure 1 illustrates the distribution of negative and positive sentiment documents across the top 13 subreddits (the largest subreddits with a comparable number of documents). The subreddit r/calgary (positive: n=47, 89% vs negative: n=6, 11%) had the highest positive sentiment, followed by r/Edmonton (n=32, 67% vs n=16, 33%) and r/ontario (n=80, 62% vs n=49, 38%). The subreddits with the highest negative sentiment included r/canada (positive: n=47, 34% vs negative: n=92, 66%), followed by r/saskatchewan (n=11, 41% vs n=16, 59%) and r/alberta (n=30, 41% vs n=43, 59%).

XSL•FO
RenderX

Figure 1. Distribution of negative and positive sentiment documents across top 13 subreddits.

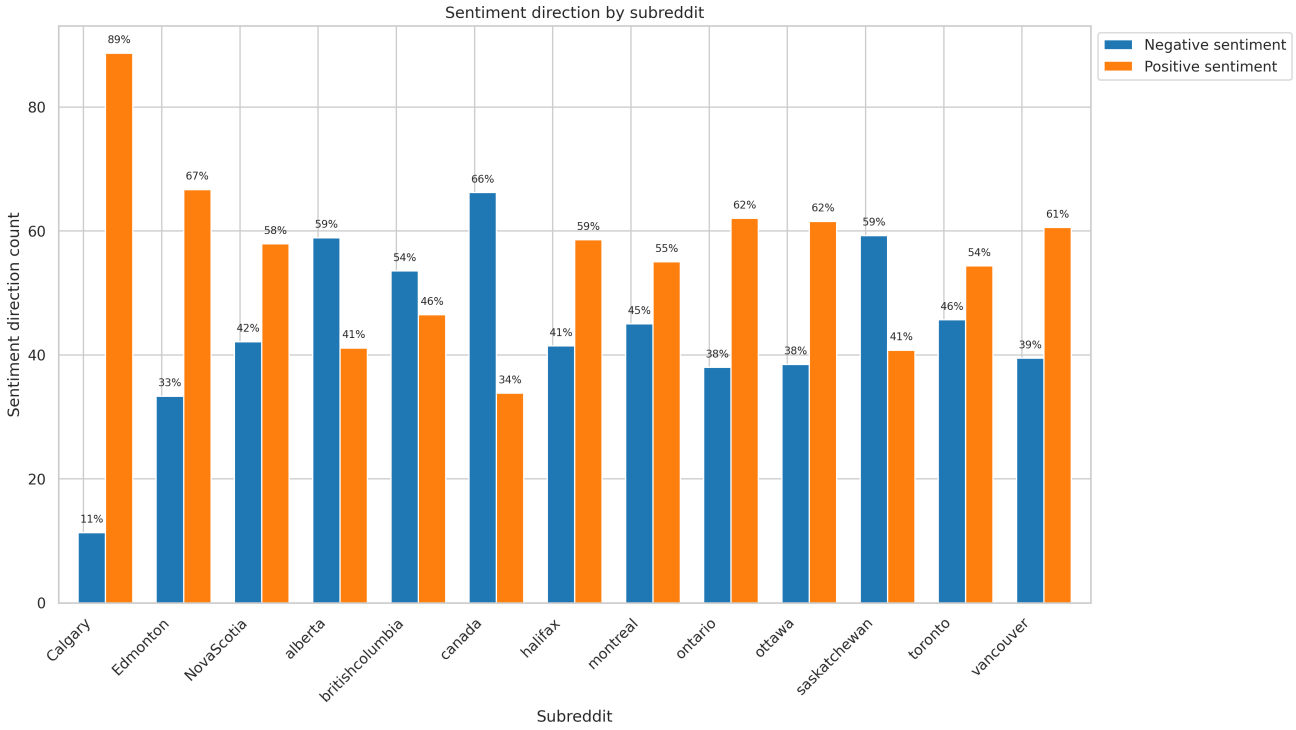

Emotion Analysis

We discerned diverse emotions within the top 13 subreddits (the largest subreddits with a comparable number of documents). Figure 2 displays the average scores for the emotions of approval and disapproval across each subreddit. Within the subreddits, r/Calgary (0.17) had the highest approval emotion score, followed by r/montreal (0.16) and r/alberta (0.15). However, r/montreal (0.11) had the highest disapproval emotion score, followed by r/alberta (0.084), and r/saskatchewan (0.083).

Figure 2. Approval and disapproval emotion distribution by top 13 subreddits.

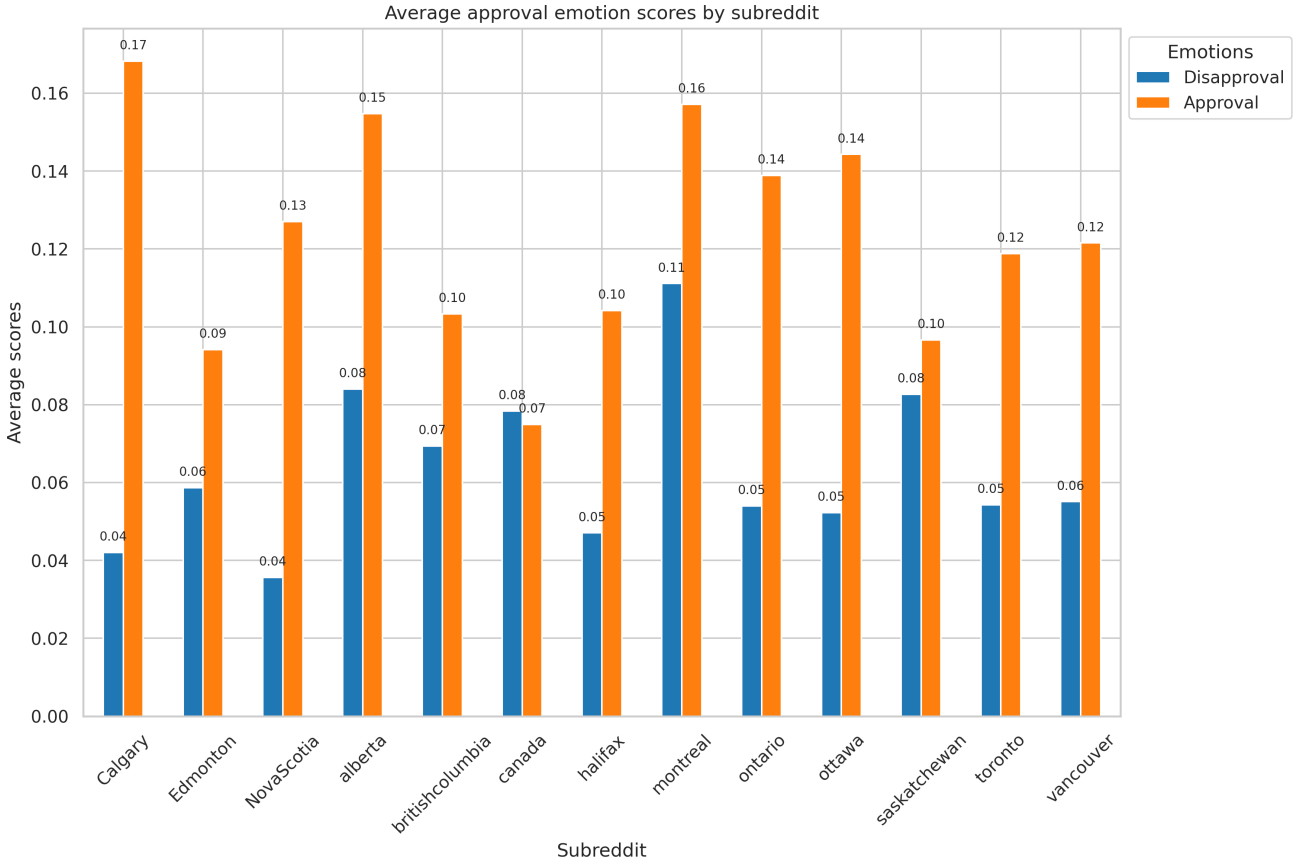

XSL•FO
RenderX

Similarly, Figure 3 presents the average scores for anger, sadness, and fear across the top 13 subreddits. Anger was the highest in r/montreal (0.035), r/canada (0.034), and r/toronto (0.027). Sadness was the highest in r/britishcolumbia (0.043), r/ottawa (0.042), and r/toronto (0.038). Finally, fear was the highest in r/montreal (0.033), r/britishcolumbia (0.018), and r/canada (0.015).

Figure 3. Negative emotion (anger, sadness, and fear) distribution by top 13 subreddits.

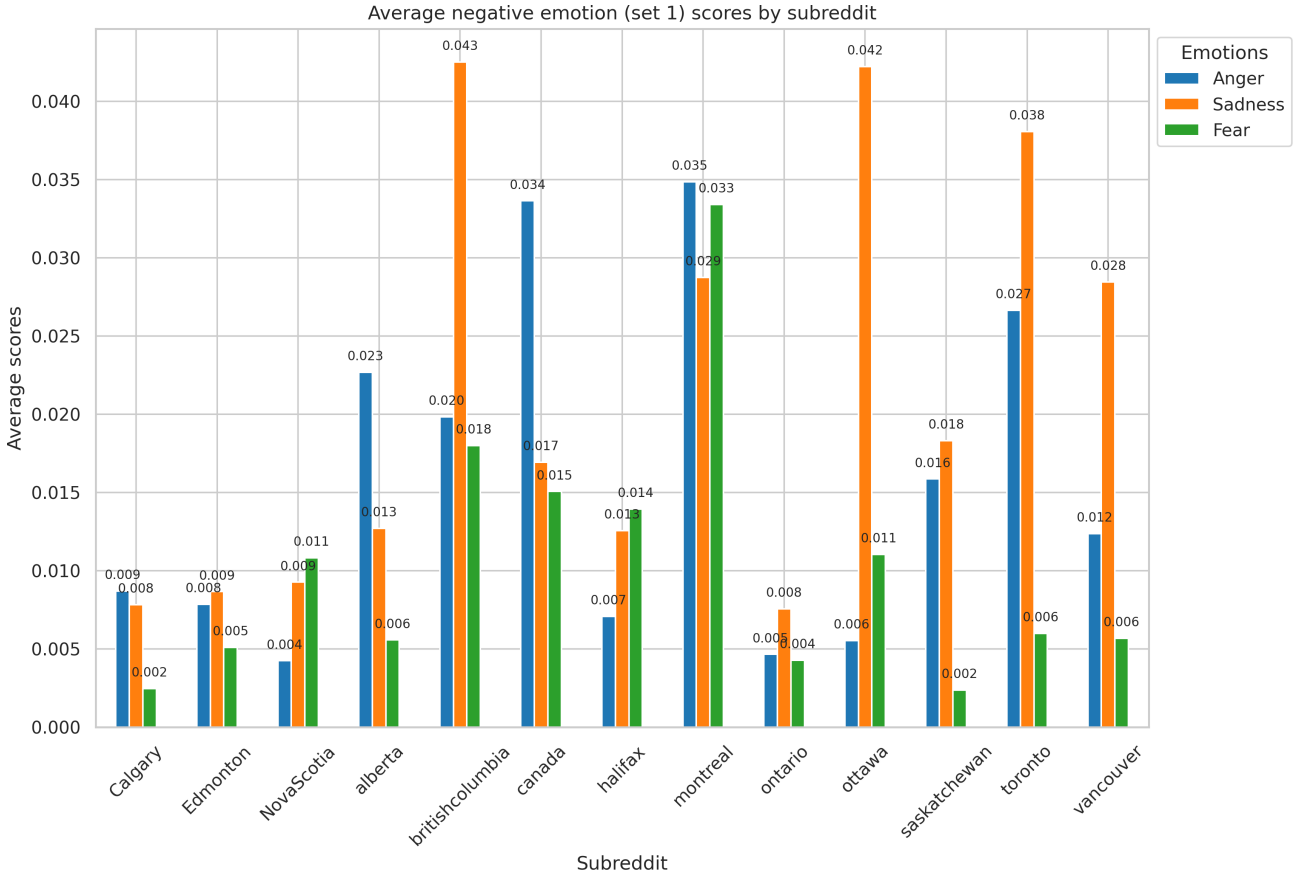

Figure 4 illustrates the average scores for admiration and optimism for the top 13 subreddits. Admiration was found to be the highest in r/saskatchewan (0.094), r/britishcolumbia (0.084), and r/Calgary (0.082). Optimism was the highest in r/saskatchewan (0.096), r/toronto (0.095), and r/Edmonton (0.074).

XSL•FO
RenderX

Figure 4. Positive emotion (admiration and optimism) distribution by top 13 subreddits.

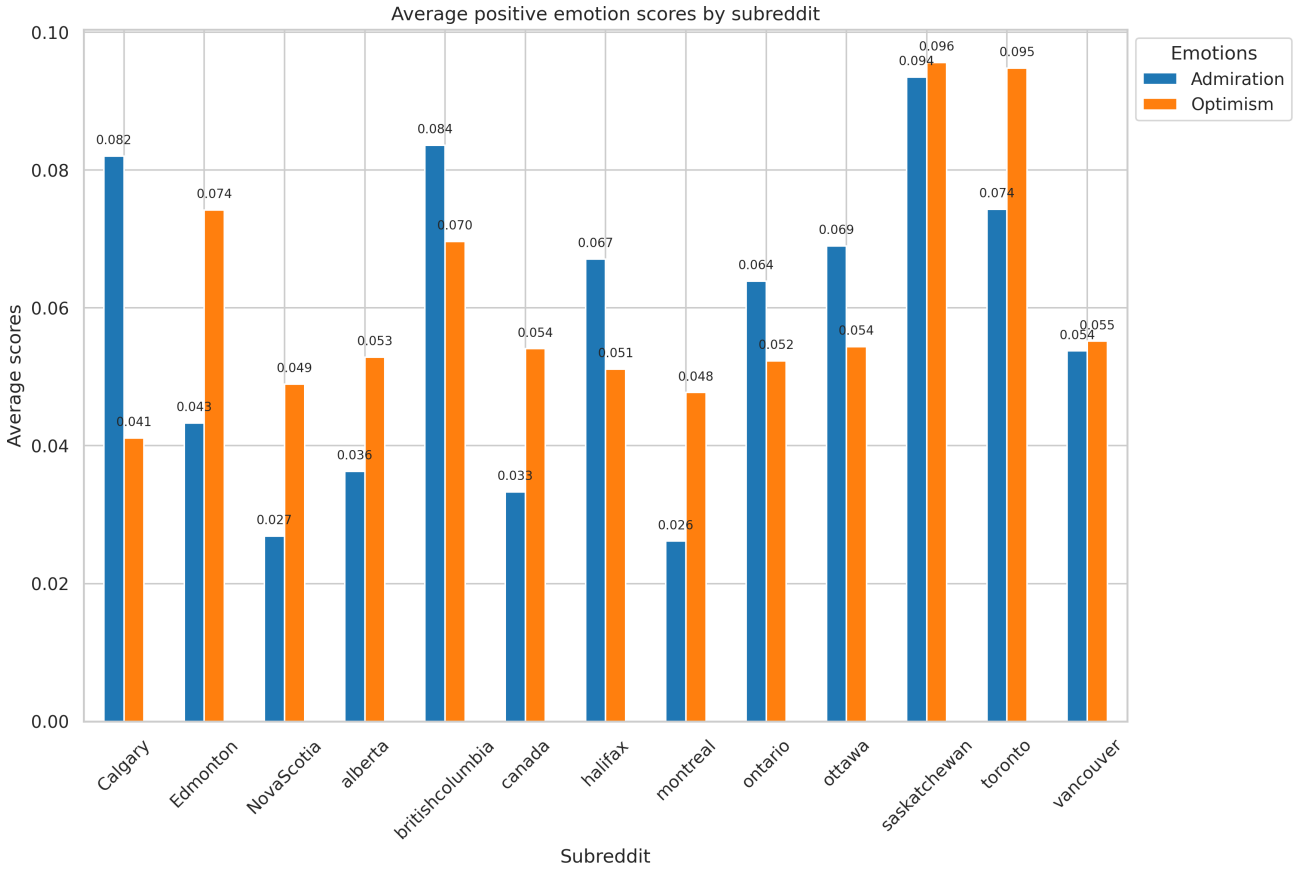

Engagement Analysis

Table 5 illustrates the average upvote score for the posts and comments extracted from the top 13 subreddits, reflecting the levels of community engagement and approval. In our dataset, the subreddit r/toronto (23.68) had the highest level of engagement and approval, followed by r/canada (17.87) and r/montreal (17.14).

Table . Average normalized upvote by top 13 subreddits.

Subreddit	Upvote number
r/toronto	23.686033
r/canada	17.872152
r/montreal	17.13952
r/britishcolumbia	15.612242
r/vancouver	14.76366
r/ottawa	12.568449
r/Calgary	12.30941
r/alberta	11.324873
r/saskatchewan	9.57762
r/ontario	9.345084
r/Edmonton	6.49853
r/Halifax	6.349839
r/NovaScotia	5.056377

XSL•FO
RenderX

Discussion

Principal Findings

This study examined Reddit discussions related to extreme heat in Canada to identify common topics and concerns, explore sentiment and emotional responses, and analyze how these relate to community engagement. The findings indicate that extreme heat and climate change are prominent themes among Canadian Reddit users, with substantial regional variation in both the discussion focus and emotional tone. Engagement patterns coincided with discussions of climate-related issues, suggesting that social media can provide signals of public perception and discourse about extreme weather events. Interpreted through a risk communication lens, these discussions reflect how people perceive health risks, express uncertainty, and negotiate adaptation strategies in an online environment. From an information ecosystem perspective, Reddit acts as a dynamic venue where public health–relevant information is coconstructed, contested, and amplified. While our findings suggest that Reddit can surface emerging concerns and perspectives related to extreme heat, these insights should be understood as complementary to other forms of evidence (eg, surveys, interviews, or other social media analyses) rather than representative of public perceptions in Canada as a whole.

Reddit can provide an exploratory view of public perceptions, particularly among users who are active on the platform, especially with regard to current events that impact health. Prior research has highlighted its utility for tracking perceptions during events such as influenza outbreaks [30]. In this study, Reddit discussions captured public concerns about extreme heat and its growing threats to health, safety, and economic well-being, particularly in cities such as Halifax and Vancouver. Users referenced real-world impacts such as elevated mortality following heatwaves, heightened fire risks, and deteriorating quality of life, which align with national data on climate-related health outcomes [7]. These discussions may reflect broader concerns about climate anxiety among some Canadians, with many linking local events to larger concerns about global climate change [31]. The themes we identified, such as concerns about energy costs, adaptation strategies, and climate anxiety, should be viewed as potential information needs. They may help guide future research and message development, but actionable public health communication strategies will require additional data sources and direct engagement with communities.

Common Discussion Topics and Concerns About Extreme Heat in Canada

Reddit users across Canada expressed widespread interest in heat-related issues, as reflected in the dominance of heat-related content in the analyzed posts. Specifically, 82.5% (n=501) of the posts and over 80% (n=25,253) of the comments centered on the topic of “heat.” In addition to general heat concerns, other common themes included extreme heat, hot weather, and hot summer. These patterns align with national polling data, such as a recent Leger survey showing that 70% of Canadians are concerned about climate change and its role in increasing extreme weather events [31].

The r/Canada subreddit had the highest number of posts, comments, and engagement overall, indicating the breadth of public discourse about extreme heat and climate change across the country. Region-specific subreddits, such as r/halifax and r/britishcolumbia, followed closely, with high volumes of posts and engagement—likely reflecting local experiences with recent extreme heat events in regions historically known for more moderate climates [32]. These observations parallel other Reddit-based studies on climate discourse that similarly found widespread concern about the impacts of climate change, with the consequences of climate change a main topic of discussion on 100,000 Reddit posts [33].

Beyond general climate concerns, the topic modeling analysis revealed several dominant themes related to daily life and broader societal challenges. The largest topic cluster focused on heating and cooling, including practical terms, such as “pump,” “furnace,” and “gas.” These reflect public concern over energy infrastructure and affordability, particularly in the context of extreme temperatures. This is supported by evidence showing that many Canadian households face energy poverty, leading to discomfort, health risks, and financial strain [34]. Reddit users expressed anxiety around increasing energy costs and insufficient cooling infrastructure, especially during heatwaves.

Health impacts emerged as another important discussion topic. Climate change and extreme heat are contributing to health inequities across the determinants of health, leading to impacts on community and individual health in Canada [35]. A popular news article in 2024 focused on extreme heat and outlined how racialized communities, those renting their homes, and those living in energy poverty were found to be at higher risk [36]. These findings correspond with the broader recognition of how climate change intersects with social determinants of health [35] and demonstrate how Reddit discussions mirror evolving media, political, and institutional attention to these equity concerns.

Attitudes and Emotions Expressed in Reddit Discussions on Extreme Heat

The emotional tone of Reddit discussions varied substantially by region, illustrating different levels of concern, belief, and lived experience with climate change across the country. Subreddits such as r/Calgary and r/Edmonton showed more positive sentiment overall, while r/canada and r/saskatchewan leaned toward more negative sentiment. This contrast highlights the nuanced ways that attitudes toward climate-related challenges manifest online.

Regional differences in emotion were also apparent in the emotion analysis. For example, subreddits such as r/saskatchewan and r/calgary expressed relatively high levels of admiration and optimism. This may reflect regional differences in belief systems about climate change. Some research suggests that individuals who downplay the severity of climate change tend to report more positive feelings, which can lead to a phenomenon known as “denial-based hope” [37]. While this outlook may reduce perceived urgency, it can also hinder individual and collective action to address climate risks [38].

XSL•FO
RenderX

In contrast, negative emotions, such as anger, sadness, and fear, were more pronounced in subreddits such as r/montreal, r/britishcolumbia, and r/canada.

These emotions may stem from direct experiences with climate-related disasters, such as the 2021 heat dome in British Columbia and the 2023 wildfires in Nova Scotia. Consistent with previous research, negative emotional responses, particularly fear, were associated with heightened climate awareness and a greater likelihood of supporting or engaging in adaptive action [39]. In our dataset, these emotions appeared in ways that may align with those patterns, though our study did not measure behavior directly. Sustained exposure to such distress can also lead to climate-related mental health challenges. A recent survey found that 78% of young Canadians reported negative mental health impacts from climate change, underscoring the psychological burden of climate anxiety [40].

Engagement Patterns in Heat-Related Reddit Discussions: Connections With Topics and Sentiment

Community engagement, as measured by upvotes, was the highest in subreddits r/toronto and r/canada, indicating high levels of user participation and approval in discussions surrounding extreme heat. Upvotes make a post more visible on Reddit and show others' support or approval of the topic [29].

These findings reflect broader trends in climate change communication on social media, where approximately one-third of users engage with climate-related content, particularly Gen Z adults [41]. Engagement is often higher for posts that include calls to action or connect the problem of climate change with potential solutions [41]. Credible sources, linking the problem to solutions, and including calls to action are key best practices for engaging people in climate change information [42]. In this context, high engagement with Reddit content may indicate a public appetite for action-oriented or relatable discussions, especially when aligned with credible sources and practical implications for everyday life (eg, energy costs, health impacts).

Engagement levels also intersected with topic and sentiment trends. Subreddits with high discussion volumes and strong negative emotional expression also saw high levels of engagement. This pattern indicates that emotionally charged posts may coincide with more active participation, though causality cannot be inferred. While our results suggest that stronger emotional expression often coincided with higher engagement, our study was not designed to explain why some emotionally charged posts gain more traction than others. Understanding these dynamics would require qualitative analysis of discourse and context, or triangulation with other data sources.

Taken together, these findings suggest that Reddit can serve as an exploratory lens into how Canadians discuss and react to extreme heat, surfacing concerns that range from household-level adaptation strategies to broader issues of affordability, energy systems, and equity. While not representative of the population as a whole, these insights highlight potential information needs and public concerns that public health agencies and policymakers may wish to consider alongside other information sources when designing communication strategies around extreme heat and climate risks. More broadly, this study illustrates how computational analyses of social media can complement traditional research approaches by identifying emergent themes and emotions in real time, thereby offering early signals of public discourse that warrant further investigation. This research demonstrates the value of integrating social media monitoring into the broader toolkit for public health communication and climate resilience planning.

Future Research

Future research should examine Reddit posts related to extreme heat and climate change in Canada for misinformation or disinformation. Within this research, it would be beneficial to examine how misinformation or disinformation about climate change and extreme heat impacts the engagement and emotions found within posts. Further exploring the relationship between emotions and action in climate adaptation would also help shape initiatives. Developing communication reflective of the topics and discussions found in the subreddits and evaluating them across communities would also help understand the utility of social media discourse for informing official messaging. Future research should also build on this descriptive foundation by incorporating qualitative coding of excerpts, deeper contextual interpretation, and comparisons with external datasets, such as meteorological records or health outcomes. Finally, discourse found on Reddit should be compared with other social media platforms and Google trends to understand similarities and differences found in online sources.

Limitations

This research was based on the social media platform Reddit and common subreddits in Canada. Generalizability is limited because Reddit users skew toward younger demographics and each subreddit has its own norms, moderation practices, and community perspectives. As Caplan and Purser [43] highlight, subreddit-specific cultures and sociotechnical contexts shape discourse; our inclusion of 30 national, provincial or territorial, and local subreddits was intended to mitigate this, but the findings should still be interpreted as a broad snapshot rather than as representative of each subreddit's culture. Future research could extend this work by analyzing subreddit-specific practices in greater depth and comparing results with other online platforms.

A key technical limitation is Reddit's API restriction, which caps retrieval at 1000 posts per query. In high-traffic subreddits, this constraint can bias the dataset toward more recent content and limit temporal coverage.

Model-related limitations must also be considered. The BERTopic and sentiment or emotion models used were trained on general-purpose corpora that differ from Canadian Reddit discussions of extreme heat. This may lead to systematic misclassification; for example, sarcasm, humor, or colloquial expressions may be interpreted incorrectly. These constraints affect both topic coherence and sentiment assignments, underscoring the need for caution in drawing substantive conclusions.

XSL•FO
RenderX

Finally, despite filtering out bot-generated, deleted, and non-English comments, undetected bots remain a persistent challenge. Sophisticated bots can mimic human discourse and potentially amplify particular narratives, subtly influencing topic distributions and sentiment profiles. This is especially relevant in politically and socially salient discussions.

Conclusions

This study demonstrates the impacts of extreme heat and climate-related challenges across Canadian communities, found within subreddit discussions. Topics around heating and cooling, financial burdens, and broader social implications of adapting to a changing climate were found. Community engagement was particularly high in r/toronto and r/Canada, signifying active discourse and collective concern. The emotional analysis showed regional differences, with positive emotions such as optimism and admiration common in r/saskatchewan and r/Calgary, and negative emotions such as fear and anger more common in r/montreal and r/canada. Terms such as “pump,” “costs,” and “climate” were common and reinforced the need for adaptive strategies that address social, economic, and environmental concerns. Overall, the findings demonstrate how Reddit may serve as a useful exploratory platform for examining discourse around climate change and extreme heat, offering signals that could inform future tailored and community-specific communication efforts.

Acknowledgments

Generative artificial intelligence was not used in the research or manuscript development associated with this study.

Funding

This research was funded by a University of Guelph CARE-AI Seed Fund AI Innovation Award.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request. We have placed the source code, sample negative sentiment documents, sample positive sentiment documents, and representative documents for each topic in [44].

Authors' Contributions

Conceptualization: FZ, JEM, MM, SG, SZ
Formal analysis: FZ, SZ
Funding acquisition: JEM, MM, SG
Investigation: SZ
Methodology: FZ, JEM, MM, SG, SZ
Project administration: JEM, MM
Supervision: FZ, JEM, MM
Writing – original draft: MM, SZ
Writing – review and editing: FZ, JEM, MM, SG, SZ

Conflicts of Interest

None declared.

References

1. What is climate change? United Nations. 2025. URL: https://www.un.org/en/climatechange/what-is-climate-change [accessed 2026-01-31]
2. Who is most impacted by climate change. Government of Canada. 2022. URL: https://www.canada.ca/en/health-canada/services/climate-change-health/populations-risk.html [accessed 2026-02-10]
3. Climate change adaptation in Canada. Natural Resources Canada. 2008. URL: https://natural-resources.canada.ca/climate-change/what-adaptation/10025 [accessed 2026-01-31]
4. Extreme heat events: health risks and who is at risk of extreme heat events. Government of Canada. 2022. URL: https://www.canada.ca/en/health-canada/services/climate-change-health/extreme-heat/who-is-at-risk.html [accessed 2026-01-31]
5. Risks to health from climate change. Government of Canada. 2022. URL: https://www.canada.ca/en/health-canada/services/climate-change-health/risks-to-health.html [accessed 2026-01-31]
6. Extreme heat events: overview. Government of Canada. 2018. URL: https://www.canada.ca/en/health-canada/services/climate-change-health/extreme-heat.html [accessed 2026-01-31]
7. Fact sheet: climate change and heat waves. Canadian Climate Institute. 2024. URL: https://climateinstitute.ca/news/fact-sheet-heat-waves/ [accessed 2026-01-31]
8. Peters E, Boyd P, Cameron LD, et al. Evidence-based recommendations for communicating the impacts of climate change on health. Transl Behav Med 2022 May 25;12(4):543-553. [doi: 10.1093/tbm/ibac029] [Medline: 35613000]

XSL•FO
RenderX

9. León B, Negredo S, Erviti MC. Social engagement with climate change: principles for effective visual representation on social media. Clim Policy 2022;22(8):976-992. [doi: 10.1080/14693062.2022.2077292]
10. Schäfer MS. Social media in climate change communication: state of the field, new developments and the emergence of generative AI. Dialogues Clim Change 2025 Apr;2(1):49-55. [doi: 10.1177/29768659241300666]
11. Reddit users by country. Oberlo. 2024. URL: https://www.oberlo.com/statistics/reddit-users-by-country [accessed 2026-01-31]
12. Fariello G, Jemielniak D. The changing language and sentiment of conversations about climate change in Reddit posts over sixteen years. Commun Earth Environ 2025;6(1):1-9. [doi: 10.1038/s43247-024-01974-8]
13. Treen K, Williams H, O'Neill S, Coan TG. Discussion of climate change on Reddit: polarized discourse or deliberative debate? Environ Commun 2022 Jul 4;16(5):680-698. [doi: 10.1080/17524032.2022.2050776]
14. Gadanidis T. The discourse of climate change on reddit. Toronto Working Papers in Linguistics. 2020. URL: https://twpl.library.utoronto.ca/index.php/twpl/article/view/41010 [accessed 2026-01-31]
15. Research ethics/protection of human participants. University of Guelph. 2025. URL: https://www.uoguelph.ca/research/services-divisions/ethics [accessed 2026-01-31]
16. Interagency Advisory Panel on Research Ethics. Does research using social media platforms require research ethics board review?—Guidance in applying TCPS 2 (2022). Government of Canada. 2022. URL: https://ethics.gc.ca/eng/reb-cer_social-sociaux.html [accessed 2026-01-31]
17. Boe B. PRAW: The Python Reddit API Wrapper 7.7.1 documentation. 2023. URL: https://praw.readthedocs.io/en/stable/ [accessed 2026-01-31]
18. Janaswamy S, Blackburn J. Exploring climate change discourse: measurements and analysis of reddit data. arXiv. Preprint posted online on Dec 2, 2024. [doi: 10.48550/arXiv.2412.01111]
19. Llewellyn C, Grover C, Oberlander J. Improving topic model clustering of newspaper comments for summarisation. Presented at: Proceedings of the ACL 2016 Student Research Workshop; Aug 7-12, 2016. [doi: 10.18653/v1/P16-3007]
20. Sun Y, Loparo K, Kolacinski R. Conversational structure aware and context sensitive topic model for online discussions. Presented at: 2020 IEEE 14th International Conference on Semantic Computing (ICSC); Feb 3-5, 2020. [doi: 10.1109/ICSC.2020.00019]
21. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv. Preprint posted online on Mar 11, 2022. [doi: 10.48550/arXiv.2203.05794]
22. Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment treebank. Presented at: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing; Oct 20, 2013; Seattle, Washington, USA p. 1631-1642. [doi: 10.18653/v1/D13-1170]
23. Sanh V, Debut L, Chaumond J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv. Preprint posted online on Oct 2, 2019. [doi: 10.48550/arXiv.1910.01108]
24. Distilbert Base Uncased Finetuned Sst 2 English. dataloop. 2024. URL: https://dataloop.ai/library/model/distilbert_distilbert-base-uncased-finetuned-sst-2-english/ [accessed 2026-01-31]
25. Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S. GoEmotions: a dataset of fine-grained emotions. Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Jul 5-10, 2020. [doi: 10.18653/v1/2020.acl-main.372]
26. Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. Preprint posted online on Jul 26, 2019. [doi: 10.48550/arXiv.1907.11692]
27. SamLowe/roberta-base-go_emotions. Hugging Face. URL: https://huggingface.co/SamLowe/roberta-base-go_emotions [accessed 2026-01-31]
28. Alon D, Ko J. GoEmotions: a dataset for fine-grained emotion classification. Google Research. 2021. URL: https://research.google/blog/goemotions-a-dataset-for-fine-grained-emotion-classification/ [accessed 2026-01-31]
29. Rechel E. What is Reddit upvotes? Hybrid Traffic. 2021. URL: https://www.hybridtraffic.net/single-post/what-is-reddit-upvotes [accessed 2026-01-31]
30. Park A, Conway M. Tracking health related discussions on Reddit for public health applications. AMIA Annu Symp Proc 2018;2017:1362-1371. [Medline: 29854205]
31. Rabson M. 7 in 10 Canadians worried about climate change, poll suggests. CBC News. 2023. URL: https://www.cbc.ca/news/politics/majority-canadians-worried-about-climate-change-1.6964334 [accessed 2026-01-31]
32. Extreme heat in Canada. Canadian Climate Institute. 2023. URL: https://climateinstitute.ca/reports/extreme-heat-in-canada/ [accessed 2026-01-31]
33. Parsa MS, Shi H, Xu Y, Yim A, Yin Y, Golab L. Analyzing climate change discussions on reddit. Presented at: 2022 International Conference on Computational Science and Computational Intelligence (CSCI); Dec 14-16, 2023. [doi: 10.1109/CSCI58124.2022.00150]
34. Martiskainen M, Das R. Worried about high energy bills, some Canadians risk discomfort, illness and even death. The Conversation. 2022. URL: http://theconversation.com/worried-about-high-energy-bills-some-canadians-risk-discomfort-illness-and-even-death-186371 [accessed 2026-01-31]
35. Health Canada. Health of Canadians in a changing climate: advancing our knowledge for action. : Government of Canada; 2022 URL: https://changingclimate.ca/site/assets/uploads/sites/5/2022/02/CCHA-REPORT-EN.pdf [accessed 2026-01-31]

XSL•FO
RenderX

36. Ram A, Shingler B, Pelley L. Extreme heat is a global killer—and worse for our health than previously thought, new research shows. CBC News. 2024. URL: https://www.cbc.ca/news/health/heat-second-opinion-1.7241714 [accessed 2026-01-31]
37. Bechtoldt MN, Schermelleh-Engel K. Worry's Clout: Concern, not positive affectivity, drives climate activism. J Environ Psychol 2025 Feb;101:102517. [doi: 10.1016/j.jenvp.2024.102517]
38. Why some Americans do not see urgency on climate change. Pew Research Center. 2023. URL: https://www.pewresearch.org/science/2023/08/09/why-some-americans-do-not-see-urgency-on-climate-change/ [accessed 2026-01-31]
39. Lawrance EL, Thompson R, Newberry Le Vay J, Page L, Jennings N. The impact of climate change on mental health and emotional wellbeing: a narrative review of current evidence, and its implications. Int Rev Psychiatry 2022 Aug;34(5):443-498. [doi: 10.1080/09540261.2022.2128725] [Medline: 36165756]
40. Galway LP, Field E. Climate emotions and anxiety among young people in Canada: a national survey and call to action. J Clim Chang Health 2023 Jan;9:100204. [doi: 10.1016/j.joclim.2023.100204]
41. Tyson A, Kennedy B, Funk C. Climate engagement and activism. Pew Research Centre. 2021. URL: https://www.pewresearch.org/science/2021/05/26/climate-engagement-and-activism/ [accessed 2026-01-31]
42. Communicating on climate change. United Nations. URL: https://www.un.org/en/climatechange/communicating-climate-change [accessed 2026-01-31]
43. Caplan MA, Purser G. Qualitative inquiry using social media: a field-tested example. Qual Soc Work 2019 May;18(3):417-435. [doi: 10.1177/1473325017725802]
44. Reddit heat Canada study. GitHub. URL: https://github.com/szamani20/Reddit-Heat-Canada-Study [accessed 2026-02-05]

Abbreviations

API: application programming interface
BERTopic: bidirectional encoder representations topic modeling

Edited by A Stone, A Mavragani; submitted 14.Aug.2025; peer-reviewed by AW Bello, Q Ng; revised version received 28.Oct.2025; accepted 28.Oct.2025; published 25.Feb.2026.

Please cite as:
MacKay M, Zamani Alavijeh S, Gosselin S, Zarrinkalam F, McWhirter JE
Exploring Reddit Discourse and Information Needs Surrounding Extreme Heat: Topic, Sentiment, and Engagement Analysis
J Med Internet Res 2026;28:e82426
URL: https://www.jmir.org/2026/1/e82426
doi:10.2196/82426

© Melissa MacKay, Soroush Zamani Alavijeh, Sydney Gosselin, Fattane Zarrinkalam, Jennifer E McWhirter. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Detection of Antithrombotic-Related Bleeding in Older Inpatients: Multicenter Retrospective Study Using Structured and Unstructured Electronic Health Record Data

Claire Coumau[1,2,3], PhD; Frederic Gaspar[1,2,3], PhD; Mehdi Zayene[4], MSc; Elliott Bertrand[4], MSc; Lorenzo Alberio[5], Prof Dr Med; Christian Lovis[6], Prof Dr Med; Patrick E Beeler[7,8], PD Dr med; Fabio Rinaldi[9,10,11], Prof Dr; Monika Lutters[12], PhD; Marie-Annick Le Pogam[13*], MD, PhD; Chantal Csajka[1,2,3*], Prof Dr; SwissMADE Collaborators[14,17]

[1]Center for Research and Innovation in Clinical Pharmaceutical Sciences, Lausanne University Hospital and University of Lausanne, Rue du Bugnon 19, Lausanne, Switzerland

[10]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

[11]Faculty of Biology and Medicine, Fondazione Bruno Kessler, Trento, Italy

[12]Head of Hospital Pharmacy, Kantonsspital Aarau, Aarau, Switzerland

[13]Department of Epidemiology and Health Systems, Unisanté, University Center for Primary Care and Public Health & University of Lausanne, Lausanne, Switzerland

[14]See Acknowledgments

[17]

[2]School of Pharmaceutical Sciences, University of Geneva, Geneva, Switzerland

[3]Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, Geneva, Switzerland

[4]Artefact Company, Lausanne, Switzerland

[5]Service of Haematology and Central Haematology Laboratory, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

[6]Division of Medical Information Sciences, Geneva University Hospitals and University of Geneva, Geneva, Switzerland

[7]Division of Occupational and Environmental Medicine, Epidemiology, Biostatistics and Prevention Institute, University of Zurich and University Hospital Zurich, Zurich, Switzerland

[8]Center for Clinical Research, University of Lucerne, Lucerne, Switzerland

[9]Dalle Molle Institute for Artificial Intelligence Research, IDSIA USI-SUPSI, Lugano, Switzerland

[*]these authors contributed equally

Corresponding Author:
Chantal Csajka, Prof Dr
Center for Research and Innovation in Clinical Pharmaceutical Sciences, Lausanne University Hospital and University of Lausanne, Rue du Bugnon 19, Lausanne, Switzerland

Abstract

Background: Bleeding complications are a major contributor to adverse drug events among older inpatients, particularly in those treated with antithrombotic agents. Timely and accurate detection of bleeding events is essential for improving drug safety surveillance and clinical risk management.

Objective: The study aimed to develop and validate automated algorithms for detecting major bleeding (MB) and clinically relevant nonmajor bleeding (CRNMB) events from electronic medical records (EMRs) by combining structured data-based rule models and a natural language processing (NLP) approach, and to evaluate their performance and generalizability against a manually reviewed gold standard and an external dataset.

Methods: We conducted a multicenter retrospective study using routinely collected EMR data from 3 Swiss university hospitals. Patients 65 years or older who received at least one antithrombotic agent and were hospitalized between January 2015 and December 2016 were included. To detect MB and CRNMB events, rule-based algorithms were developed using structured data (*International Statistical Classification of Diseases, 10th Revision, German Modification* [*ICD-10-GM*] codes, laboratory values, transfusion records, and antihemorrhagic prescriptions), with variables and cutoff values defined according to adapted International Society on Thrombosis and Haemostasis definitions and expert consensus. In parallel, a supervised NLP model was applied to discharge summaries from one hospital. A manual review of 754 EMRs served as the reference standard for internal validation, and the algorithm performance of the structured data algorithms (SDA), NLP, and their combination (SDA+NLP) was evaluated against this manually reviewed gold standard using standard performance metrics. External validation was performed on an independent dataset from the Lausanne University Hospital to assess model robustness and generalizability.

XSL•FO
RenderX

Results: Among 36,039 inpatient stays, SDA identified 8.26% (n=2979) as MB and 15.04% (n=5419) as CRNMB cases. *ICD-10-GM* codes alone detected 28.5% (n=849) of MB and 31.48% (n=1706) of CRNMB cases, while laboratory data contributed most to event detection (n=1994, 66.94% for MB and n=3663, 67.60% for CRNMB). Integrating SDA with NLP improved detection, identifying 12.2% (920/7513) of MB and 27.4% (2062/7513) of CRNMB cases at 1 hospital. The combined model achieved the best performance (sensitivity 0.84, positive predictive value 0.51, F_1-score 0.64). External validation on Lausanne University Hospital 2021 - 2022 data (n=24,054 stays) confirmed the algorithms' reproducibility; the prevalence of MB decreased while CRNMB increased, reflecting evolving clinical practices and antithrombotic use patterns.

Conclusions: Our integrated approach, combining SDA with NLP, enhances the detection of hemorrhagic events in older hospitalized patients treated with antithrombotic agents, suggesting its potential usefulness for drug safety monitoring and clinical risk management.

International Registered Report Identifier (IRRID): RR2-10.2196/40456

(J Med Internet Res 2026;28:e77809) doi:10.2196/77809

KEYWORDS

adverse drug events; adverse drug reactions; older inpatients; structured data mining; machine learning; natural language processing; electronic medical records; multicenter study; antithrombotic; hemorrhage; artificial intelligence; pharmacovigilance

Introduction

Over 16% of older inpatients experience at least 1 adverse drug event (ADE) during their hospital stay [1], often with more severe consequences than in younger patients [2]. Among the medications most frequently implicated, antithrombotic agents, widely prescribed in older adults for the prevention and treatment of cardiovascular disease, stand out as a major cause of bleeding-related ADEs [1,3]. Hemorrhagic complications represent a substantial share of drug-related harm in this population and are associated with longer hospital stays, higher readmission rates, and increased mortality. Continuous and accurate measurement of these events is therefore essential to inform prevention strategies, strengthen pharmacovigilance, and promote safer antithrombotic use in clinical practice.

Various approaches have been developed to detect ADEs in hospital settings, each with advantages and limitations. Spontaneous reporting systems, though simple to implement, notoriously underestimate the true frequency of ADEs due to underreporting [4]. Systematic chart reviews of electronic medical records (EMRs), often considered the reference standard, provide detailed clinical information but are too resource- and time-intensive for routine surveillance [5]. To overcome these constraints, automated detection methods using routinely collected EMR data have emerged. These approaches leverage both structured data, such as diagnostic codes, medication records, laboratory results, and vital signs, and unstructured clinical narratives, including discharge summaries, progress notes, and consultation reports. Structured data are accessible and standardized, supporting large-scale analyses but may lack contextual nuances needed to capture complex clinical events such as bleeding [6-8]. Conversely, textual data, although unstructured, often contain richer clinical detail but require advanced computational methods for analysis. Recent advances in machine learning (ML) and natural language processing (NLP) have markedly improved the ability to extract this information and are increasingly applied to pharmacovigilance and ADE detection [9]. Integrating both structured and textual data appears particularly promising for identifying bleeding events, potentially enhancing accuracy and completeness [10].

Despite growing interest in automated ADE detection to support drug safety monitoring, important knowledge gaps remain, particularly in the Swiss context. Most existing studies focusing on bleeding events have relied exclusively on either structured or unstructured data [11-15], have prioritized prediction rather than detection [8,16], or have focused on specific bleeding types or patient groups [7,10,17-19]. Furthermore, clear operational definitions distinguishing major bleeding (MB) from clinically relevant nonmajor bleeding (CRNMB) are often lacking, limiting comparability across studies [20]. To date, no study in Switzerland has comprehensively evaluated the combined contribution of structured and textual data for ADE detection in a general inpatient population receiving antithrombotic therapy. To address this gap, we conducted a multicenter study integrating rule-based algorithms and NLP to detect MB and CRNMB events among older inpatients treated with antithrombotics. We hypothesized that combining structured and textual EMR data would improve the accuracy and completeness of bleeding event identification compared with using either data source alone. The study aimed to develop rule-based algorithms for bleeding detection from structured data sources (diagnoses, laboratory results, transfusions, and antihemorrhagic prescriptions) based on international definitions; design and train an NLP model to identify bleeding mentions in discharge summaries; assess and compare the diagnostic performance of structured data algorithms (SDA), NLP, and their combination (SDA+NLP) against a manually reviewed gold standard; and evaluate the generalizability of the best-performing models through external validation on an independent dataset.

Methods

Study Design

We conducted a multicenter cross-sectional study using retrospective data covering the period from January 1, 2015, to December 31, 2016. Data were obtained from 4 large Swiss hospitals: Lausanne University Hospital (CHUV; approximately

XSL•FO
RenderX

1500 beds [21]), Geneva University Hospital (HUG; approximately 2000 beds [22]), both located in the French-speaking region and serving the cantons of Vaud and Geneva, respectively, Zürich University Hospital (USZ; approximately 900 beds [23]) serving the Zurich metropolitan area, and Baden Cantonal Hospital (KSB; approximately 400 beds [24]) serving the canton of Aargau in the German-speaking region. This study was conducted in accordance with the SRTOBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement (Checklist 1).

The 2015 - 2016 dataset was used for algorithm development as it was the most recent period with harmonized, high-quality structured and unstructured EMR data across all hospitals. Later years were excluded due to EMR vendor transitions, database restructuring, and new data-governance restrictions limiting access to deidentified text. A more recent CHUV dataset (2021 - 2022) was used for temporal and external validation to test algorithm robustness under evolving clinical practices and documentation standards.

Study Participants and Hospital Stays

Eligible participants were Swiss residents 65 years or older treated with at least 1 antithrombotic agent during their hospital stay. Antithrombotic agents included vitamin K antagonists, heparins, platelet aggregation inhibitors, direct thrombin inhibitors, direct factor Xa inhibitors, or fondaparinux. Hospitalizations had to last at least 24 hours and to occur between January 2015 and December 2016 (test dataset). For the external validation, an additional dataset from CHUV covering January 2021 to December 2022 was used (validation dataset). Only patients who had provided explicit consent for the reuse of their health data for research purposes, as indicated by the signature of the general consent form, were eligible for inclusion. Hospital stays lasting less than 24 hours were excluded from the analysis.

Data Sources and Preprocessing

Each participating hospital extracted relevant clinical data from its institutional data warehouses for all inpatient stays meeting the inclusion criteria. The extracted datasets included both structured and unstructured data. Structured data comprised administrative information, patient movements within the hospital, key clinical and laboratory parameters, and prescribed medications coded using the anatomical therapeutic chemical classification. Diagnostic codes were drawn from the *International Statistical Classification of Diseases, 10th Revision, German Modification* (*ICD-10-GM*), and procedures were coded according to the Swiss Classification of Surgical Procedures (CHOP). Diagnoses and procedures were obtained from the hospital billing records associated with each inpatient stay. Unstructured data included discharge summaries. Further details on data extraction and handling are available in the published study protocol [25].

Prior to analysis, structured data were cleaned, harmonized, and verified for consistency at each site, then locally deidentified before being transferred to a centralized database hosted at CHUV. Unstructured data were deidentified and, where necessary, converted into machine-readable formats, but were stored locally on secure hospital servers to comply with data governance policies. Due to the extent of missing and inconsistent information, such as discrepancies in data structure, coding systems, variable definitions, and extensive missing values, reliable harmonization of KSB data with the other hospitals was not feasible, and data from KSB were excluded from the analysis. In addition, only unstructured data from CHUV were analyzed, as full deidentification of textual data from the other sites could not be ensured. The same preprocessing workflow was applied to the 2021 - 2022 CHUV dataset used for external validation. An overview of the data processing workflow is provided in Figure 1.

Figure 1. Overview of the data extraction and preprocessing pipeline for structured and unstructured electronic medical record (EMR) data. CHOP: Swiss Classification of Surgical Procedures; CHUV: Lausanne University Hospital; *ICD-10-GM*: *International Statistical Classification of Diseases, 10th Revision, German Modification*.

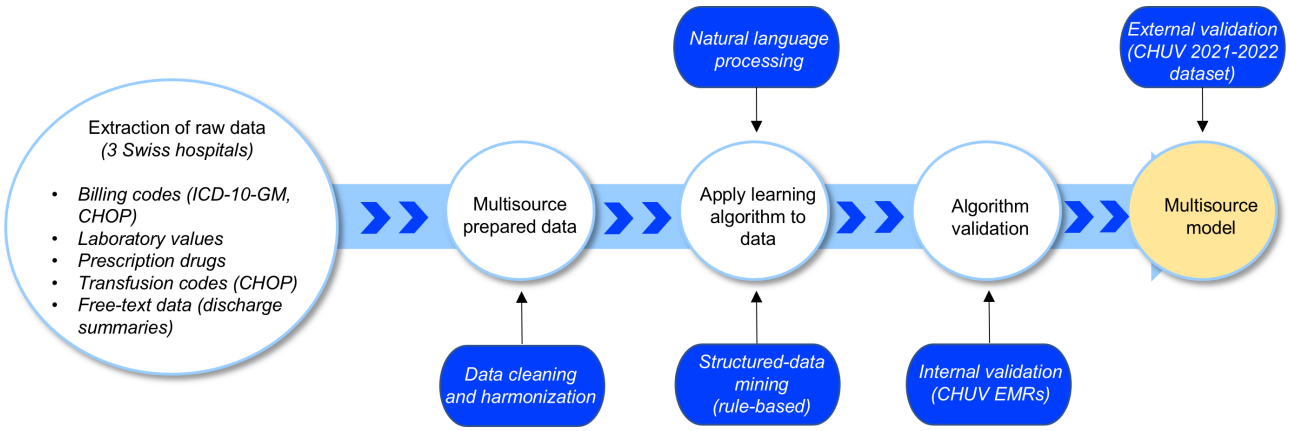

Bleeding Detection Algorithms

We selected variables of interest and cutoff values for our algorithms based on an adaptation of the International Society on Thrombosis and Haemostasis (ISTH) definitions of MB [26] and CRNMB [27], informed by an extensive review of international guidelines (Multimedia Appendix 1). MB was defined as a hemoglobin drop of 4 g/dL or more within 48 hours, a 2 to 4 g/dL drop associated with death within 24 hours, a hemoglobin level less than 7 g/dL, a hemoglobin level between 7 and 9 g/dL associated with death within 24 hours, or a transfusion of more than 5 units of blood or red blood cells. CRNMB was defined as a hemoglobin drop of 2 to 4 g/dL within

XSL•FO
RenderX

48 hours not associated with death or a hemoglobin nadir between 7 and 9 g/dL without subsequent death. The ISTH hemoglobin thresholds were adapted to improve specificity in older inpatients and to reduce the risk of misclassifying nonhemorrhagic anemias. The 5-unit transfusion threshold was pragmatically chosen due to the limited granularity of CHOP procedural codes. Additional structured indicators were also integrated to refine case classification: the prescription of antihemorrhagic agents (idarucizumab, andexanet alfa, prothromplex, octaplex, and beriplex) was considered indicative of MB. MB-related in-hospital mortality was defined as any hospital stay involving at least 1 MB event followed by death during the same admission. We then developed rule-based algorithms using Boolean logic to detect MB and CRNMB cases from structured data (Figure 2).

Figure 2. Algorithmic framework for detection of major bleeding (MB) and nonmajor clinically relevant bleeding (CRNMB) cases using structured data. Antihemorrhagic agent: idarucizumab, andexanet alfa, prothromplex, octaplex, and beriplex. ΔHb: drop in hemoglobin levels within 48 hours; *ICD-10-GM*: *International Statistical Classification of Diseases, 10th Revision, German Modification*; Min Hb: minimum hemoglobin value during the stay.

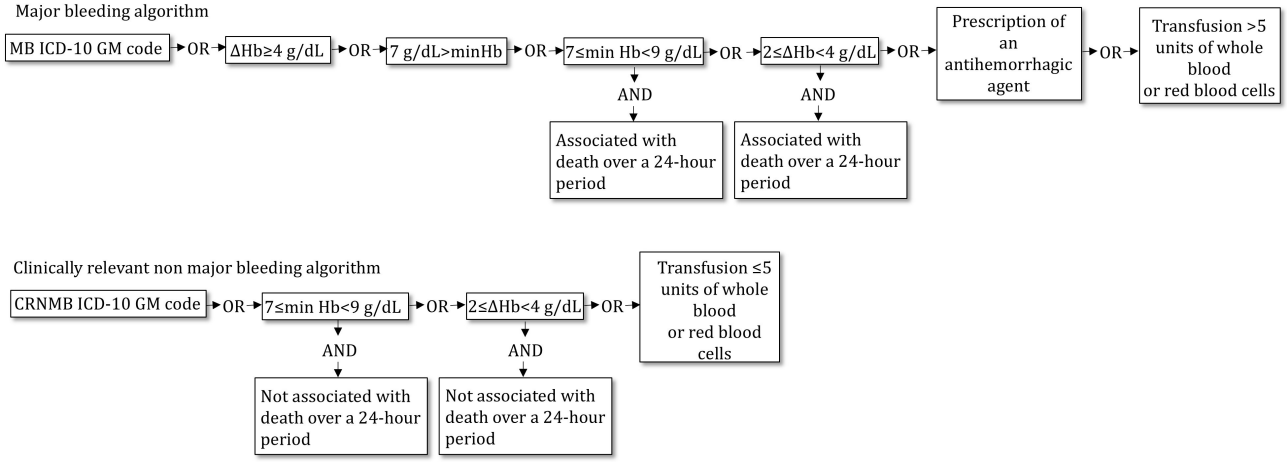

The *ICD-10-GM* list comprised 12 codes for MB and 41 codes for CRNMB (Table 1). We defined an MB in-hospital mortality case as a stay containing an MB occurring during hospitalization followed by death of the patient during the same hospitalization period. We measure the prevalence of bleeding cases, corresponding to inpatient stays with at least 1 MB or CRNMB event, either present on admission or occurring during hospitalization. We quantified the relative and absolute contribution of each structured data source (diagnoses, laboratory, transfusions, and medications), both individually and in combination, in terms of overall detection capacity and proportion of identified bleeding events.

XSL•FO
RenderX

Table . Lists of deficient systems/organs and distribution of *ICD-10-GM*[a] chapters and codes identifying MB[b] and CRNMB[c] cases.

Types of hemorrhagic events	*ICD-10-GM* codes
MB	
Hyphema	H21.0
Hemorrhage and rupture of the choroid	H31.3
Retinal, vitreous, or subarachnoid hemorrhage	H35.6, H43.1, I60_
Hemopericardium not classified elsewhere	I31.2
Intracerebral hemorrhage	I61_
Other nontraumatic intracranial hemorrhages	I62_
Hemoperitoneum	K66.1
Hemarthrosis	M25.0_
Hypovolemic shock	R57.1
Shock during or after a procedure for diagnostic and therapeutic purposes, not classified elsewhere	T81.1
CRNMB	
Conjunctival hemorrhage	H11.3
Otorrhagia	H92.2
Hemorrhagic esophageal varices	I85.0, I98.3
Other specified diseases of the esophagus	K22.8
Non-traumatic hemothorax	J94.2
Gastric, duodenal, or gastrojejunal ulcer with hemorrhage and/or perforation	K25.0, K25.2, K25.6, K26.0, K26.2, K26.4, K26.6, K27.0, K27.2, K27.4, K27.6, K28.0, K28.2, K28.4, K28.6
Acute hemorrhagic gastritis	K29.0
Rectal and anal hemorrhage	K62.5
Hematemesis	K92.0
Melena	K92.1
Unspecified gastrointestinal hemorrhage	K92.2
Prostatic congestion and hemorrhage	N42.1
Hematoma of the broad ligament	N83.7
Hematometra	N85.7
Abnormal bleeding from the uterus and vagina	N93.8, N93.9
Postmenopausal bleeding	N95.0
Epistaxis	R04.0
Throat hemorrhage	R04.1
Hemoptysis	R04.2
Respiratory tract hemorrhage	R04.8, R04.9
Spontaneous ecchymosis	R23.3
Unspecified hematuria	R31
Hemorrhage, not classified elsewhere	R58, T81.0

[a]*ICD-10-GM*: *International Classification of Diseases, 10th Revision, German Modification*.

[b]MB: major bleeding.

[c]CRNMB: clinically relevant nonmajor bleeding.

XSL•FO
RenderX

Natural Language Processing Model

To complement structured data detection, we developed a supervised ML model to identify MB, CRNMB, and past bleeding cases documented in discharge summaries.

A dataset of 400 discharge summaries from CHUV was randomly divided into a training set (n=280) and a test set (n=120), including 100 summaries with MB, 100 with CRNMB, and 200 with no bleeding. Three independent physicians manually annotated the 400 discharge summaries using 4 mutually exclusive labels: (A) 'presence of CRNMB,' (B) 'presence of MB' (as previously defined), (C) 'history of bleeding' (when a discharge summary mentioned bleeding in the EMR before the hospital admission), and (D) 'absence of any bleeding.' Preprocessing steps included tokenization, lemmatization, and sentence segmentation using the French spaCy model (v3.0) [28]. The classification pipeline combined logistic regression and support vector machine models, selected for their interpretability and robustness with limited training data. We deliberately used a classical supervised ML model rather than deep learning architectures to ensure interpretability, reproducibility, and computational efficiency, which are essential for clinical validation and routine pharmacovigilance applications. This approach also better suited the relatively small, annotated corpus, allowing transparent feature weighting and easier auditability across institutions. The model was trained using the scikit-learn library (Python v3.9.1). The classification pipeline proceeded in 3 stages: step 1: binary classifier to identify bleeding-relevant versus irrelevant sentences; step 2: multiclass classifier to distinguish between *irrelevant*, *antecedent bleeding*, and *active bleeding*; step 3: binary classifier to further differentiate between MB and CRNMB within sentences flagged as active bleeding. Sentence-level predictions were aggregated to assign a final label to each document. Rules were prioritized as follows: MB>CRNMB>history of bleeding>no bleeding. This ensured a conservative classification hierarchy, favoring the identification of more severe bleeding cases when multiple labels were present. Further methodological details are available in the study proposal previously published [25], the related article [29], and summarized in Figure 3.

Figure 3. Natural language processing workflow from raw text input to final classification output. BERT: Bidirectional Encoder Representations From Transformers; DS: discharge summary; HBDSCAN: Hierarchical Density-Based Spatial Clustering of Applications With Noise; UMAP: Uniform Manifold Approximation and Projection.

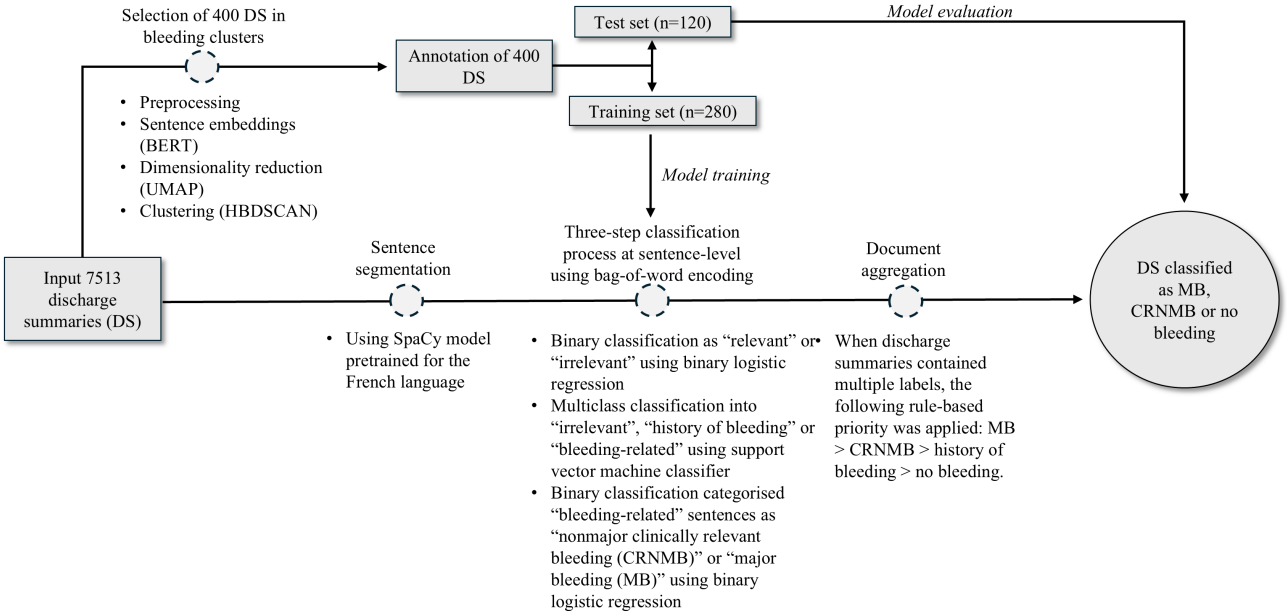

Validation of the Bleeding Detection Algorithms

Internal Validation Using CHUV 2015-2016 Data

To validate the SDA and SDA combined with NLP (SDA+NLP) models, we conducted a manual review of 754 EMRs from CHUV's 2015 - 2016 dataset. The sample size for validation was determined using a test result-based sampling method [30]. Assuming a 7% MB, a 10% CRNMB accuracy, and a sensitivity of 0.7, at least 704 EMRs had to be reviewed, and 754 EMRs were effectively reviewed. Four physicians independently reviewed the records to compare algorithm-detected with clinician-identified MB and CRNMB cases. The review process followed a structured protocol aligned with ISTH definitions [26,27] and adapted for retrospective application to routinely collected hospital data. Reviewers assessed each inpatient stay according to 4 key criteria: (1) evidence of active bleeding, (2) severity of the event (eg, hemodynamic instability), (3) need for therapeutic intervention (eg, transfusion volume, administration of antihemorrhagic agents), and (4) temporal relationship to hospital admission (present on admission versus occurred during stay). A complete list of synonyms used to identify MB and CRNMB cases during manual chart review is provided in Multimedia Appendix 2. Two binary classification scenarios were evaluated: (1) MB versus all other cases (CRNMB or no bleeding), and (2) CRNMB versus no bleeding (excluding MB). Algorithm performance was evaluated at the inpatient-stay level using standard binary classification metrics (sensitivity, specificity, positive predictive value [PPV], negative predictive value, accuracy, and F_1-score), with manual chart review as the gold standard. Comparisons between SDA, NLP, and combined models were descriptive, and sensitivity was prioritized due to the study's patient safety focus. Interrater

XSL•FO
RenderX

reliability among reviewers was evaluated using Fleiss κ on a subset of 40 cases, with agreement levels interpreted according to Landis and Koch [31] (>0.80: *almost perfect*; 0.61 - 0.80: *substantial*; 0.41 - 0.60: *moderate*; 0.21 - 0.40: *fair*; 0.00 - 0.20: *slight*; <0.00: *poor*). A *P* value associated with the Fleiss κ coefficient was also calculated, with a *P* value less than .05 indicating statistically significant agreement. Additional details and results are provided in Multimedia Appendix 3. In a subanalysis, a causal relationship between antithrombotic therapy and each bleeding event was also assessed during the manual review, using a structured tool based on temporal association, biological plausibility, and alternative explanations. Cases were rated as *certain*, *probable*, *possible*, or *unclassified*, in relation to antithrombotic exposure, according to the WHO-Uppsala Monitoring Center scale [32]. The methodology, sample size calculation, and findings of the causality assessment of the subanalysis are presented in Multimedia Appendix 4.

External Validation Using CHUV 2021-2022 Data

An external validation was performed using CHUV data from 24,054 inpatient stays between January 2021 and December 2022. We applied the same detection algorithms (SDA and SDA+NLP) to this independent dataset to evaluate their performance, robustness, and reproducibility. Results were compared to those from the 2015 - 2016 CHUV dataset.

Statistical Analysis

Descriptive statistics were used to summarize population characteristics. Comorbidity was assessed using the Charlson and Elixhauser indexes [33,34], which are validated tools for risk adjustment and mortality prediction based on administrative health data. Comparisons of patient characteristics between hospitals were conducted using a 1-way analysis of variance on ranks (Kruskal-Wallis test) for continuous variables and Pearson χ^2 test for categorical variables. Hyperparameters of the NLP classifier were optimized through 5-fold cross-validation on the training set, and final performance was estimated on an independent test set. All performance metrics were reported with 95% CIs calculated using the Wilson method. Analyses were conducted using StataCorp. 2021. Stata Statistical Software: Release 17. College Station, TX: StataCorp LLC software for structured data and Python (v3.9.1) for NLP development.

Ethical Considerations

Human Subject Ethics Review Approvals or Exemptions

This study was conducted in accordance with the *Declaration of Helsinki* and Swiss federal regulations governing research on human data. Ethical approval was obtained from all relevant cantonal ethics committees, coordinated by the lead committee of the Canton of Vaud (CER-VD No. 2018 - 00272). As the study involved secondary analysis of routinely collected, deidentified hospital data, it qualified for a simplified review under Swiss Human Research Act article 2, paragraph 2(c).

Informed Consent

The study relied exclusively on existing clinical data that were deidentified before analysis. According to Swiss regulations and institutional data governance policies, informed consent was waived for patients who had not explicitly objected to the use of their medical data for research purposes. All participating hospitals operate an institutional opt-out procedure, allowing patients to refuse the secondary use of their data for research.

Privacy and Confidentiality

All data were deidentified at source before analysis. Structured data were transferred through secure institutional channels to a restricted-access research environment hosted at CHUV. Unstructured textual data remained stored locally on hospital servers and were processed within each institution's secure infrastructure to comply with data protection requirements. No directly identifiable information was accessible to the investigators.

Compensation Details

No compensation was provided to patients, as the study involved secondary analysis of preexisting, routinely collected data and did not include direct contact with participants.

Protection of Identifiable Information in Figures and Supplementary Materials

No image, document, or figure contains any identifiable patient information. Consequently, no individual consent for image publication was required.

Ethics Approval

Approved by the Cantonal Ethics Committee of Vaud, Switzerland (CER-VD No. 2018 - 00272); informed consent was waived for patients who did not opt out of research data use.

Results

Study Population Characteristics

A total of 36,039 inpatient stays, involving 24,991 unique patients, were included in the analysis: 7677 stays (5754 patients) at CHUV, 18,015 stays (11,356 patients) at HUG, and 10,347 stays (7881 patients) at USZ. Patient characteristics are detailed in Table 2. The median age at admission was 78 (IQR 65 - 99) years, with a balanced sex distribution (51.40% male). Comorbidity was generally low across the cohort, with a median Charlson index and Elixhauser index of 0.0; USZ patients had the lowest overall comorbidity burden.

XSL•FO
RenderX

Table . Baseline patient characteristics and treatments: overall and by university hospital.

Characteristics	All hospitals (n=36,039)[a]	CHUV[b] (n=7677)	HUG[c] (n=18,015)	USZ[d] (n=10,347)
Admission age (years), median (IQR)	78 (65 - 99)	79 (65 - 99)	80 (65 - 99)	75 (65 - 92)
Sex, n (%)				
Male	18,525 (51.40)	3987 (51.93)	8638 (47.95)	5900 (57.02)
Female	17,514 (48.60)	3690 (48.07)	9377 (52.05)	4447 (42.98)
Length of stay (d), median (IQR)	9 (1-342)	9 (1-293)	12 (1 - 342)	6 (1-145)
Transfer to intensive care, n (%)	1534 (4.26)	467 (6.1)	1067 (5.92)	—[e]
In-hospital mortality, n (%)	1416 (3.93)	345 (4.5)	850 (4.7)	221 (2.1)
Comorbidity, n (%)				
Chronic renal dysfunction	8418 (23.36)	2163 (28.18)	5662 (31.43)	593 (5.7)
Dialysis	622 (1.7)	176 (2.3)	241 (1.3)	205 (2.0)
Acute renal dysfunction	1151 (3.19)	266 (3.5)	623 (3.5)	262 (2.5)
Chronic liver dysfunction	1020 (2.83)	294 (3.8)	497 (2.8)	229 (2.2)
Acute liver dysfunction	498 (1.4)	145 (1.9)	244 (1.4)	109 (1.1)
Hypertension	18,316 (50.82)	3158 (41.14)	9271 (51.46)	5887 (56.90)
Alcohol abuse	1354 (3.76)	388 (5.1)	663 (3.7)	303 (2.9)
Stroke	3001 (8.33)	813 (10.6)	1625 (9.02)	563 (5.4)
Cancer	6776 (18.80)	1572 (20.48)	2905 (16.13)	2299 (22.22)
Platelet coagulation defect	2178 (6.04)	496 (6.5)	1029 (5.71)	653 (6.3)
Anemia	7624 (21.15)	1998 (26.03)	4380 (24.31)	1246 (12.04)
Risk fall	11,376 (31.57)	2932 (38.20)	6021 (33.42)	2423 (23.42)
Diabetes	6638 (18.42)	1314 (17.12)	3575 (19.84)	1749 (16.90)
Recent myocardial infection	1923 (5.34)	609 (7.9)	761 (4.2)	553 (5.3)
Low weight	4059 (11.26)	967 (12.6)	2533 (14.06)	559 (5.4)
Thrombolysis	695 (1.9)	180 (2.3)	512 (2.8)	3 (0.0)
Vascular malformation	955 (2.6)	153 (2.0)	334 (1.9)	468 (4.5)
Charlson comorbidity index, median (IQR)	0.0 (0.0 - 9.0)	0.0 (0.0 - 9.0)	0.0 (0.0 - 7.0)	0.0 (0.0 - 7.0)
Elixhauser comorbidity index, median (IQR)	0.0 (0.0 - 6.0)	0.0 (0.0 - 6.0)	0.0 (0.0 - 6.0)	0.0 (0.0 - 5.0)
Antithrombotic categories, n (%)				
Direct factor Xa inhibitors	3297 (9.15)	599 (7.8)	1478 (8.20)	1220 (11.80)
Vitamin K antagonists	7469 (20.72)	1324 (17.25)	4943 (27.44)	1202 (11.62)
Heparin group	24,784 (6877)	5045 (65.71)	11,918 (66.17)	7821 (75.59)
Direct thrombin inhibitors	255 (0.7)	87 (1.1)	134 (0.7)	34 (0.3)
Platelet aggregation inhibitors	14,220 (39.46)	4354 (56.71)	4700 (26.09)	5166 (49.93)
Thrombolytics	104 (0.3)	15 (0.2)	89 (0.5)	0.0 (0.0)
Other antithrombotic agents: fondaparinux	1365 (3.79)	212 (2.8)	1140 (6.33)	13 (0.1)
Antidotes, n (%)	137 (0.4)	15 (0.2)	122 (0.7)	0.0 (0.0)
Transfusion, n (%)	582 (1.6)	264 (3.4)	318 (1.8)	—

XSL•FO
RenderX

Characteristics	All hospitals (n=36,039)[a]	CHUV[b] (n=7677)	HUG[c] (n=18,015)	USZ[d] (n=10,347)
≤5 UI[f] plasma or red blood cells	225 (0.6)	100 (1.3)	125 (0.7)	—
>5 UI plasma or red blood cells	357 (1.0)	164 (2.1)	193 (1.1)	—
Number of antithrombotic agents received during hospitalization, n (%)				
1	22,397 (62.15)	4257 (55.45)	12,381 (68.73)	5759 (55.66)
2	11,918 (33.07)	2904 (37.83)	4924 (27.33)	4090 (39.53)
3	1641 (4.55)	495 (6.4)	669 (3.7)	477 (4.6)
≥4	83 (0.2)	21 (0.3)	41 (0.2)	21 (0.2)

[a]n: total number of recorded measurements for the respective parameter.

[b]CHUV: Lausanne University Hospital.

[c]HUG: Geneva University Hospital.

[d]USZ: Zürich University Hospital.

[e]Not available (missing or nontransferred data).

[f]UI: unit of blood component.

Distinct prescribing patterns were observed across hospitals: HUG had the highest use of vitamin K antagonists (n=4943, 27.44%), CHUV had the highest prescription rate of antiplatelet agents (n=4354, 56.71%), and USZ reported the highest use of direct factor Xa inhibitors (n=1220, 11.79%) and heparins (n=7821, 75.59%). Hypertension (n=18,316, 50.82%), chronic renal dysfunction (n=8418, 23.36%), anemia (n=7624, 21.15%), and cancer (n=6776, 18.80%) were among the most prevalent comorbidities. Overall, in-hospital mortality was 3.93% (n=1416).

Bleeding Detection Using SDA

SDA detected 8748 (24.27%) overall bleeding cases, of which 2979 (8.26%) were MB cases and 5419 (15.04%) were CRNMB cases (Table 3). Fatal MB occurred in 1.0% (n=350) of all stays. MB prevalence varied across hospitals, with the highest proportion observed at CHUV (n=769, 10.0%), followed by USZ (n=998, 9.6%) and HUG (n=1212, 6.73%). CRNMB prevalence was highest at USZ (n=1682, 16.26%). Missing values for each variable used to identify MB and CRNMB events are presented in Multimedia Appendix 5.

Table . Prevalence of bleeding cases detected by SDA[a], overall and by university hospital[b].

	All hospitals, n (%)	CHUV[c], n (%)	HUG[d], n (%)	USZ[e], n (%)	*P* value[f]
Nonbleeding-related	27,641 (76.70)	5822 (75.84)	14,152 (78.56)	7667 (74.10)	<.001
CRNMB[g]	5419 (15.04)	1086 (14.15)	2651 (14.72)	1682 (16.26)	<.001
MB[h]	2979 (8.26)	769 (10.0)	1212 (6.73)	998 (9.6)	<.001
MB in-hospital mortality	350 (1.0)	119 (1.6)	137 (0.8)	94 (0.9)	<.001
Total	36,039	7677	18,015	10,347	—[i]

[a]SDA: structured data algorithms (ie, rule-based algorithm for structured data).

[b]Bleeding cases: number of stays for patients treated with at least 1 antithrombotic agent during which at least 1 bleeding episode occurred.

[c]CHUV: Lausanne University Hospital.

[d]HUG: Geneva University Hospital.

[e]USZ: Zürich University Hospital.

[f]Using Pearson χ^2 test.

[g]CRNMB: clinically relevant nonmajor bleeding.

[h]MB: major bleeding.

[i]Not applicable.

Relative and Absolute Contribution of Structured Data Sources

Laboratory data were the most influential source for detecting both MB and CRNMB, contributing to two-thirds of identified cases, while *ICD-10-GM* codes contributed to approximately one-third. Prescriptions for antihemorrhagic agents had a minimal added value for MB detection, while transfusion data contributed modestly. Figure 4 illustrates the relative contribution of each data source.

XSL•FO
RenderX

Figure 4. Relative contribution of structured data sources (laboratory data, *ICD-10-GM* codes, prescription of antihemorrhagic agents, and transfusions) to the detection of major bleeding (MB) and clinically relevant nonmajor bleeding (CRNMB). *ICD-10-GM*: *International Statistical Classification of Diseases, 10th Revision, German Modification*.

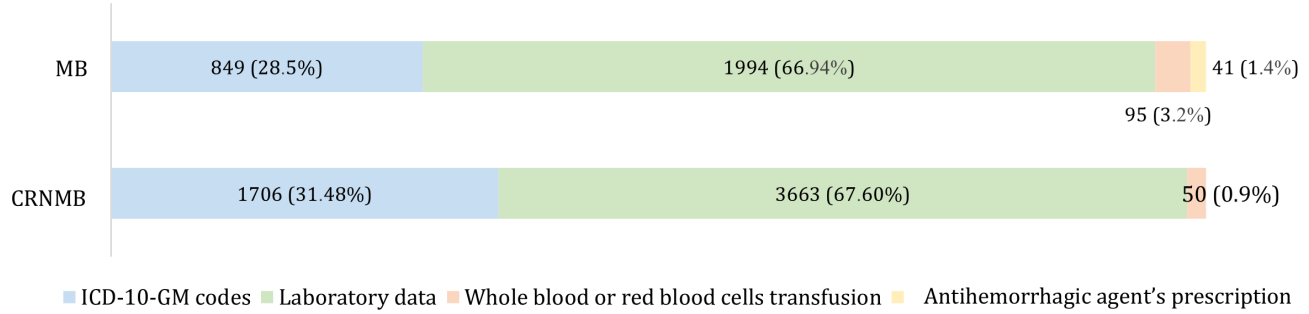

Overlap between data sources was limited. Only 12.1% (n=361) of MB stays and 8.7% (n=458) of CRNMB stays were identified by 2 data sources, while detection by all 4 sources occurred in 0% of MB cases and only 0% (n=12) of CRNMB cases (Figure 5). This limited overlap highlights the complementarity, but also fragmentation, of structured data signals.

Figure 5. Absolute contribution of structured data sources (laboratory data, *ICD-10-GM* codes, prescription of antihemorrhagic agents, and transfusions) to the detection of (A) major bleeding and (B) clinically relevant nonmajor bleeding. *ICD-10-GM: International Statistical Classification of Diseases, 10th Revision, German Modification*.

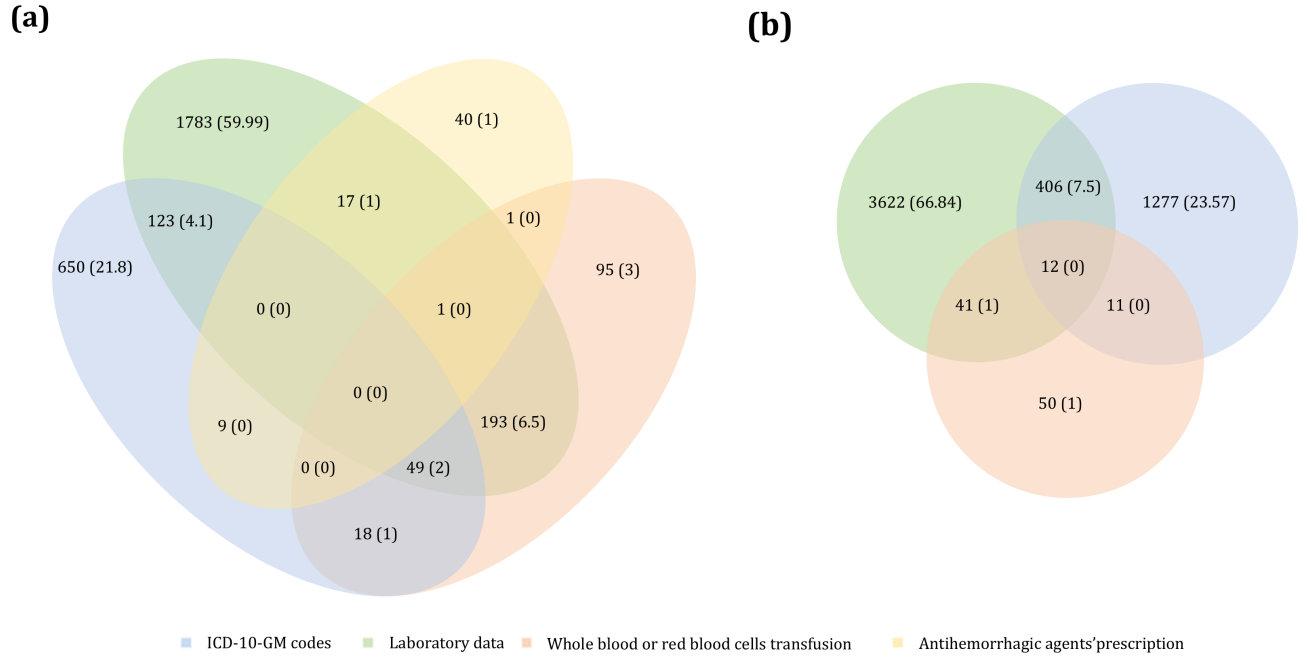

Combined Detection Using SDA and NLP (CHUV Only)

Among 7513 CHUV stays with discharge summaries, combining SDA and NLP increased case detection: In total, 39.69% (n=2982) of hemorrhagic cases were detected: 12.2% (n=920) were identified as MB and 27.45% (n=2062) as CRNMB.

For MB cases, 56.6% (n=521) were detected by SDA alone, 19.8% (n=182) by NLP alone, and 23.6% (n=217) by both. For CRNMB cases, 35.1% (n=724) were detected by SDA alone, 48.2% (n=994) by NLP alone, and 16.7% (n=344) by both.

Classification discrepancies were observed between SDA and NLP: 217 cases identified as MB by SDA were reclassified as CRNMB by NLP, and conversely, 81 CRNMB cases by SDA were reclassified as MB by NLP. NLP also enabled the detection of a history of bleeding in 8.5% (n=642) of cases, improving the temporal resolution of hemorrhage onset.

Internal Validation Using CHUV 2015-2016 Data

The manual review of 754 EMRs identified 276 bleeding cases: 144 MB and 132 CRNMB. Structured laboratory data showed the highest sensitivity (0.58, 95% CI 0.52 - 0.64), while *ICD-10-GM* codes had the highest PPV (0.89, 95% CI 0.83 - 0.98), and F_1-score (0.60). SDA outperformed NLP in sensitivity (0.77 vs 0.61), but NLP had higher PPV (0.70 vs 0.51) and F_1-score (0.65 vs 0.62). The best performance was achieved by the combined *ICD-10-GM*∪NLP algorithm, with a sensitivity of 0.71 (95% CI 0.66 - 0.76), PPV of 0.72 (95% CI 0.66 - 0.87), and F_1-score of 0.72. Algorithms combining SDA and NLP yielded the highest sensitivity (0.84), confirming the benefit of multimodal approaches. However, intersection-based algorithms (eg, SDA∩NLP) demonstrated higher specificity at the cost of reduced sensitivity.

Performance metrics for MB and CRNMB subgroups followed similar trends, with reduced sensitivity but high specificity for

XSL•FO
RenderX

ICD-10-GM–based detection. Table 4 presents a comprehensive summary of all performance metrics, including sensitivity, specificity, PPV, negative predictive value, accuracy, and F_1-score.

XSL•FO
RenderX

Table . Performance metrics of bleeding detection algorithms compared to manual electronic medical records review (gold standard; n=754)[a].

	Sensitivity[b] (95% CI)	Specificity[c] (95% CI)	PPV[d] (95% CI)	NPV[e] (95% CI)	Accuracy[f] (95% CI)	F_1-score[g]
Bleeding all type (MB[h] or CRNMB[i])						
Individual structured data sources						
ICD-10-GM[j]	0.46 (0.40 - 0.51)	0.97 (0.95 - 0.98)	0.89 (0.83 - 0.98)	0.75 (0.72 - 0.79)	0.78 (0.75 - 0.81)	0.60
Laboratory data	0.58 (0.52 - 0.64)	0.60 (0.55 - 0.64)	0.45 (0.40 - 0.64)	0.71 (0.66 - 0.75)	0.59 (0.55 - 0.62)	0.51
Whole blood or red blood cells transfusion	0.18 (0.14 - 0.23)	0.95 (0.93 - 0.97)	0.68 (0.57 - 0.97)	0.67 (0.63 - 0.70)	0.67 (0.63 - 0.70)	0.29
Detection algorithms						
SDA[k]	0.77 (0.72 - 0.82)	0.58 (0.54 - 0.62)	0.51 (0.47 - 0.62)	0.81 (0.77 - 0.85)	0.65 (0.62 - 0.68)	0.62
NLP[l]	0.61 (0.55 - 0.67)	0.85 (0.81 - 0.88)	0.70 (0.64 - 0.88)	0.79 (0.75 - 0.82)	0.76 (0.73 - 0.79)	0.65
Combined data sources and algorithms						
SDA∪NLP	0.84 (0.79 - 0.88)	0.54 (0.49 - 0.58)	0.51 (0.47 - 0.58)	0.85 (0.81 - 0.89)	0.65 (0.61 - 0.68)	0.64
SDA∩NLP	0.47 (0.41 - 0.53)	0.92 (0.89 - 0.94)	0.78 (0.71 - 0.94)	0.75 (0.71 - 0.78)	0.76 (0.72 - 0.79)	0.59
ICD-10-GM∪NLP	0.71 (0.66 - 0.76)	0.84 (0.80 - 0.87)	0.72 (0.66 - 0.87)	0.83 (0.80 - 0.86)	0.79 (0.76 - 0.82)	0.72
ICD-10-GM∩NLP	0.31 (0.26 - 0.37)	0.99 (0.98 - 1.00)	0.95 (0.89 - 1.00)	0.71 (0.68 - 0.74)	0.74 (0.71 - 0.77)	0.47
MB						
Individual structured data sources						
ICD-10-GM	0.34 (0.27 - 0.42)	0.99 (0.97 - 0.99)	0.84 (0.73 - 0.99)	0.86 (0.84 - 0.89)	0.86 (0.84 - 0.88)	0.49
Laboratory data	0.47 (0.39 - 0.55)	0.81 (0.77 - 0.84)	0.36 (0.30 - 0.84)	0.86 (0.83 - 0.89)	0.74 (0.71 - 0.77)	0.41
Whole blood or red blood cells transfusion	0.22 (0.16 - 0.29)	0.97 (0.96 - 0.98)	0.66 (0.52 - 0.98)	0.84 (0.81 - 0.87)	0.83 (0.80 - 0.85)	0.32
Algorithms						
SDA	0.72 (0.64 - 0.78)	0.79 (0.76 - 0.82)	0.45 (0.39 - 0.82)	0.92 (0.90 - 0.94)	0.78 (0.75 - 0.81)	0.55
NLP	0.35 (0.28 - 0.44)	0.95 (0.93 - 0.96)	0.63 (0.51 - 0.96)	0.86 (0.83 - 0.89)	0.84 (0.81 - 0.86)	0.45
Combined data sources and algorithms						
SDA∪NLP	0.76 (0.68 - 0.82)	0.79 (0.75 - 0.81)	0.46 (0.39 - 0.82)	0.93 (0.91 - 0.95)	0.78 (0.75 - 0.81)	0.57
SDA∩NLP	0.30 (0.23 - 0.39)	0.96 (0.94 - 0.97)	0.64 (0.52 - 0.97)	0.85 (0.83 - 0.88)	0.83 (0.81 - 0.86)	0.41
ICD-10-GM∪NLP	0.56 (0.48 - 0.64)	0.94 (0.92 - 0.96)	0.69 (0.60 - 0.96)	0.90 (0.87 - 0.92)	0.87 (0.84 - 0.89)	0.62
ICD-10-GM∩NLP	0.14 (0.09 - 0.21)	1.0 (0.99 - 1.00)	0.91 (0.72 - 1.00)	0.83 (0.80 - 0.86)	0.83 (0.80 - 0.86)	0.25
CRNMB						
Individual structured data sources						
ICD-10-GM	0.30 (0.23 - 0.39)	0.91 (0.88 - 0.93)	0.41 (0.32 - 0.93)	0.86 (0.83 - 0.88)	0.80 (0.77 - 0.83)	0.35
Laboratory data	0.25 (0.18 - 0.33)	0.73 (0.70 - 0.77)	0.17 (0.12 - 0.77)	0.82 (0.79 - 0.85)	0.65 (0.61 - 0.68)	0.20
Whole blood or red blood cells transfusion	0.03 (0.01 - 0.07)	0.96 (0.95 - 0.98)	0.15 (0.06 - 0.98)	0.82 (0.79 - 0.85)	0.80 (0.77 - 0.83)	0.05
Detection algorithms						

XSL•FO
RenderX

	Sensitivity[b] (95% CI)	Specificity[c] (95% CI)	PPV[d] (95% CI)	NPV[e] (95% CI)	Accuracy[f] (95% CI)	F_1-score[g]
SDA	0.65 (0.42 - 0.58)	0.65 (0.62 - 0.69)	0.23 (0.19 - 0.69)	0.86 (0.82 - 0.89)	0.63 (0.59 - 0.66)	0.32
NLP	0.53 (0.45 - 0.62)	0.77 (0.74 - 0.81)	0.34 (0.28 - 0.81)	0.89 (0.86 - 0.91)	0.73 (0.70 - 0.76)	0.41
Combined data sources and algorithms						
SDA∪NLP	0.66 (0.57 - 0.73)	0.56 (0.52 - 0.60)	0.24 (0.20 - 0.60)	0.88 (0.85 - 0.91)	0.58 (0.54 - 0.61)	0.35
SDA∩NLP	0.38 (0.30 - 0.47)	0.88 (0.84 - 0.90)	0.40 (0.32 - 0.90)	0.87 (0.84 - 0.89)	0.79 (0.76 - 0.82)	0.39
ICD-10-GM∪NLP	0.60 (0.52 - 0.68)	0.75 (0.72 - 0.79)	0.35 (0.29 - 0.79)	0.90 (0.87 - 0.92)	0.73 (0.69 - 0.76)	0.44
ICD-10-GM∩NLP	0.24 (0.17 - 0.32)	0.93 (0.91 - 0.95)	0.42 (0.31 - 0.95)	0.85 (0.82 - 0.87)	0.81 (0.78 - 0.83)	0.30

[a]It should be noted that no patient record contained the variable antihemorrhagic agent for the detection of MB. Consequently, the performance for this variable was not included in the table.

[b]Sensitivity: proportion of bleeding cases that have been correctly identified.

[c]Specificity: proportion of nonbleeding-related cases that have been correctly identified.

[d]PPV: positive predictive value; proportion of bleeding cases among all those classified as bleeding cases by the algorithm.

[e]NPV: negative predictive value; proportion of nonbleeding-related cases among all those classified as nonbleeding-related cases by the algorithm.

[f]Accuracy: overall prediction accuracy (ie, the proportion of bleeding and nonbleeding-related cases that the algorithm has correctly identified.

[g]F_1-score: harmonic mean of the precision and recall (ie, F_1-score = 2 × [recall × precision]/[recall + precision]).

[h]MB: major bleeding.

[i]CRNMB: clinically relevant nonmajor bleeding.

[j]*ICD-10-GM*: *International Statistical Classification of Diseases, 10th Revision, German Modification*.

[k]SDA: structured data algorithm.

[l]NLP: natural language processing.

Interrater reliability for manual review of 40 EMRs showed substantial agreement: Fleiss' Kappa was 0.65 for bleeding detection and 0.61 for MB versus CRNMB classification. Of 276 manually reviewed inpatient stays with bleeding events, 17% (n=48) were attributed to antithrombotic agents. The causal relationship was classified as "certain" in 25% of cases (n=12), "probable/likely" in 23% (n=11), and "possible" in 52% (n=25).

External Validation Using CHUV 2021-2022 Data

Application of the SDA and SDA+NLP algorithms to the CHUV validation dataset (24054 stays) demonstrated generalizability. The prevalence of MB cases significantly decreased from 10.0% in the 2015 - 2016 period to 5.55% (n=1336) in the 2021 - 2022 period, while the prevalence of CRNMB cases increased significantly from 14.15% to 16.63% (n=4000). MB in-hospital mortality also rose, from 1.6% to 2.6% (n=616).

Patient characteristics differed significantly between cohorts (Multimedia Appendix 6). Direct oral anticoagulant prescriptions increased from 7.8% to 22.7%, while vitamin K antagonist use decreased from 17.2% to 7.6%. The incidence of elevated INR values >4 declined from 3.3% to 1.8%. Both Charlson and Elixhauser scores increased, reflecting higher comorbidity. Transfusions involving ≤5 units of blood rose from 1.3% to 8.8%. Notably, the proportion of patients receiving ≥3 antithrombotic agents during hospitalization increased fivefold (from 6.7% to 35.0%).

Discussion

Principal Findings

To our knowledge, this is one of the first multicenter studies assessing the feasibility and effectiveness of combining structured and unstructured EMR data to detect bleeding cases in older inpatients treated by one or more antithrombotic agents. Across 3 large university hospitals, our SDA identified 8.26% of MB and 15.4% of CRNMB cases. Laboratory variables contributed most to event detection, while *ICD-10-GM* codes alone captured only about one-third of cases, achieving a sensitivity of 0.84 when both data sources were combined. These findings confirm the feasibility of automated bleeding surveillance in real-world hospital data and demonstrate the added value of leveraging free-text information to complement structured data sources.

Comparison to Prior Work

Our estimated bleeding rates (MB: 8.26% and CRNMB: 15.04%) are consistent with prior hospital-based studies in older adults, which reported MB incidences ranging from 1.8% to 11.3% [35,36] and CRNMB from 3.5% to 13.0% [35,37]. These findings confirm that antithrombotic-related bleeding remains a major cause of ADEs in older populations, associated with increased hospitalization length, morbidity, and mortality [38], highlighting the need for targeted preventive strategies.

The algorithms' performance varied across structured data sources and aligns with prior research. *ICD-10-GM* codes detected only one-third of MB and CRNMB cases, consistent

XSL•FO
RenderX

with previous evidence of underreporting anticoagulant-related bleeding events [39,40]. Yap et al [15] found similarly low sensitivity (16% - 24%) but very high PPV (>0.97), indicating that diagnostic codes are reliable confirmatory markers but poor screening tools. The inclusion of laboratory data markedly improved sensitivity in our SDA model, consistent with findings by Dyas et al [6] and Shung et al [10]. The modest decline in PPV was likely due to false positives generated by hemoglobin thresholds, a limitation noted in earlier work [15].

Detection of CRNMB was more challenging than MB, partly due to broader definitions and lower specificity of *ICD-10-GM* codes and transfusion data, echoing the moderate performance reported by Yap et al [15] (sensitivity 50% - 56%, PPV 43 - 50%).

The NLP model contributed substantially to overall detection, with a sensitivity of 61% and PPV of 70%, in line with earlier NLP-based models for bleeding and ADE detection [10,41]. Importantly, only about 20% of events overlapped with those captured by structured data, demonstrating that text analysis retrieves unique clinical insights often missing from coded data. NLP also enhanced temporal resolution by identifying prior bleeding episodes in 8.5% of cases, information generally unavailable from structured data alone. The combined SDA+NLP model achieved high sensitivity (0.84), thereby minimizing the risk of missed events, with only 16% of cases being false negatives. Although this proportion is relatively low, it still represents missed hemorrhagic events that could impact the accuracy of retrospective surveillance and safety signal detection. However, our detection algorithm provides a notable proportion of false positives (49%), which could contribute to alert fatigue in clinical practice and increase the workload associated with unnecessary chart reviews. For real-world deployment, performance thresholds depend on the intended use: for surveillance or signal detection, a sensitivity above 0.80 with PPV above 0.50 is generally acceptable, as false positives can be secondarily reviewed; for clinical decision support, stricter thresholds (eg, PPV≥0.70) are needed to prevent alert fatigue. Improving true positive detection to 70% would strengthen reliability and clinical applicability, potentially through prioritization or triage of clinically significant cases.

External validation revealed a decline in MB prevalence and a concurrent increase in CRNMB and MB-related mortality in the validation dataset (2021 - 2022), compared to the CHUV 2015 - 2016 dataset. These trends may reflect evolving prescribing patterns, such as increased use of direct oral anticoagulants and reduced use of vitamin K antagonists, and a shift in clinical profiles, with higher comorbidity scores and greater treatment complexity in the more recent cohort. These observations are consistent with the known bleeding risk profiles of antithrombotic agents, direct oral anticoagulants being more frequently associated with gastrointestinal bleeding (CRNMB), and vitamin K antagonists with intracranial bleeding (MB) [42], and underscore the need for dynamic algorithmic models capable of adjusting for changing treatment patterns and patient characteristics [43].

Strengths and Limitations

This study has several notable strengths. It is one of the first multicenter initiatives to integrate structured and unstructured EMR data for ADE detection in older hospitalized patients. The inclusion of 3 university hospitals provided a large, diverse dataset, while the harmonization of over 1 million clinical variables ensured robust data quality. Algorithms were developed using internationally accepted definitions of MB and CRNMB and validated through manual chart review, ensuring clinical credibility. External validation on a temporally distinct dataset further reinforced reproducibility and robustness.

Several limitations should also be considered.

First, the test dataset (2015 - 2016) was relatively dated and spanned only 2 years, reflecting mostly the time-consuming extraction and harmonization process required to merge data from three hospitals before data interoperability infrastructures were implemented. Consequently, it may not entirely capture current clinical practices. Nevertheless, this limitation was mitigated by validating our pipeline on an independent and more recent dataset.

Second, NLP development and validation were performed using CHUV data only and did not take into consideration interinstitutional variations in coding practices, hospital information system architecture and interoperability, clinical documentation standards, or local prescribing patterns, which may limit the generalizability of our findings. To mitigate this, the model was trained on a balanced, manually annotated corpus reviewed by 3 independent physicians. Future studies should externally validate the NLP model on datasets from other French-speaking institutions to confirm its performance and enhance its applicability.

Third, data from 1 hospital (Baden hospital) were excluded due to missing information and harmonization challenges, and CHOP codes could not be extracted from the USZ hospital; this could have led to underestimation of certain bleeding events. Recent efforts have been undertaken to improve data harmonization across sites, which now largely mitigate the harmonization challenges previously encountered.

Fourth, while *ICD-10-GM* code selection was based on international guidelines and expert review, some misclassification may have occurred. This limitation was partly mitigated by manual validation. However, the adoption of a standardized bleeding classification would help overcome this limitation and harmonize bleeding-event categorization across studies.

Fifth, causality assessment between bleeding cases and antithrombotic agents was not formally assessed by our algorithms, as this requires strict criteria and necessitates a comprehensive EMR review. Causality was manually evaluated using the WHO–Uppsala Monitoring Center framework, which provided valuable contextual insights but is resource-intensive. Future work should investigate semiautomated causal-inference tools to scale this process efficiently.

Sixth, structured data were insufficient to capture the timing of bleeding cases prior to admission, as such information is

XSL•FO
RenderX

documented in discharge summaries, underscoring the need for unstructured data in ADE detection.

Finally, all participating hospitals were tertiary academic centers with strong data infrastructures and comprehensive documentation practices. While this ensured data reliability and methodological consistency, it may limit the generalizability of our findings to other contexts, such as secondary or community hospitals, or to health systems with different digital maturity levels. Compared to many international settings, Swiss university hospitals operate within a decentralized but highly standardized health care system, characterized by universal coverage and well-developed inpatient services. Future research should evaluate these algorithms in more diverse hospital types and countries to assess their adaptability and scalability beyond tertiary Swiss institutions.

Future Directions

This study illustrates the value of combining structured and unstructured clinical data to improve the detection of bleeding events in older inpatients exposed to antithrombotic therapy. This integrated approach can enhance pharmacovigilance systems, reduce underreporting, and support timely clinical interventions. Future efforts should expand algorithm coverage to additional unstructured sources (eg, nursing notes and consultation letters), improve clinical documentation practices, and incorporate semiautomated causality assessment tools. Combining these detection models with multivariate risk stratification that integrates patient-specific factors (age, comorbidities, comedications, and clinical service) could enable prioritization of clinically meaningful alerts. Finally, embedding such tools within common data models and privacy-preserving data-sharing infrastructures, such as those promoted by the Swiss Personalized Health Network, could facilitate cross-institutional learning health systems and accelerate artificial intelligence-supported pharmacovigilance in real-world clinical practice.

Acknowledgments

The authors thank all contributors to data extraction and processing across participating sites, with special appreciation to Walid Gharib-Blanc and Alexandre Wetzel. We gratefully acknowledge Tapio Niemi, Marie Bettex, and Patrick Taffé (Unisanté, Department of Epidemiology and Health Systems, University of Lausanne, Switzerland).

The SwissMADE collaborators are Bernard Burnand, Department of Epidemiology and Health Systems, Unisanté and University of Lausanne, Lausanne, Switzerland; Pierre Olivier Lang, Geriatric Medicine and Geriatric Rehabilitation Division, Department of Medicine, University Hospital of Lausanne, Lausanne, Switzerland; Nicola Colic, Dalle Molle Institute for Artificial Intelligence Research, Università della Svizzera Italiana, Lugano, Switzerland; Angela Schulthess-Lisibach, Institute of Primary Health Care, University of Bern, Bern, Switzerland; Christophe Gaudet-Blavignac, Division of Medical Information Sciences, Geneva University Hospitals, Geneva, Switzerland and Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland; Nathalie Casati, Clinical Data Science Group, Lausanne University Hospital, Lausanne, Switzerland; Jean-Philippe Goldman, Faculty of Letters, Department of Linguistics, University of Geneva, Geneva, Switzerland; and Vasiliki Foufi, Division of Medical Information Sciences, University Hospitals of Geneva, Geneva, Switzerland and Faculty of Medicine, University of Geneva, Geneva, Switzerland.

Funding

This study was partially supported by the Swiss National Research Fund (project 167381). The funder had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Data Availability

Requests for access to the anonymized dataset and study materials should be directed to the corresponding author and will be considered on a case-by-case basis in accordance with institutional policies and applicable data protection regulations.

Authors' Contributions

MALP and C Csajak designed the study; MALP, FG, and C Csajak wrote and submitted the proposal. CL, Christophe Gaudet-Blavignac, Bernard Burnand, FR, ML, Pierre Olivier Lang, and PEB contributed to the development of the study proposal and participated in its revision. CL, PEB, Christophe Gaudet-Blavignac, Nicola Colic, ML, and FR contributed to data extraction and curation. C Coumou, MALP, C Csajak, and FG conducted the study and designed the rule-based algorithms. C Coumou, MALP, C Csajak, FG, and MZ performed the formal analysis. LA contributed to the conceptualization of the rule-based algorithms. MZ and EB designed and implemented the natural language processing algorithms. C Coumou has written the original draft of the manuscript. MALP, C Csajak, FG, and LA contributed to the conceptualization, supervision, and review and editing of the manuscript. The SwissMADE collaborators were consulted on methodological aspects throughout the study. All authors have read and approved the final version of the manuscript.

XSL•FO
RenderX

Conflicts of Interest

PEB reports having received fees and funding from AstraZeneca for matters unrelated to the present study. All other authors declare no competing interests.

Multimedia Appendix 1
International definitions of major bleeding and clinically relevant nonmajor bleeding: a comparative overview.
[DOCX File, 68 KB - jmir_v28i1e77809_app1.docx]

Multimedia Appendix 2
List of synonyms for major bleeding and clinically relevant nonmajor bleeding cases.
[DOCX File, 17 KB - jmir_v28i1e77809_app2.docx]

Multimedia Appendix 3
Interrater reliability assessment among manual reviewers.
[DOCX File, 18 KB - jmir_v28i1e77809_app3.docx]

Multimedia Appendix 4
Causality assessment between antithrombotic treatment and bleeding.
[DOCX File, 16 KB - jmir_v28i1e77809_app4.docx]

Multimedia Appendix 5
Overview of missing values per variable used in major bleeding and clinically relevant nonmajor bleeding algorithms.
[DOCX File, 15 KB - jmir_v28i1e77809_app5.docx]

Multimedia Appendix 6
Comparison of clinical characteristics of Lausanne University Hospital 2015-2016 vs 2021-2022 cohorts.
[DOCX File, 23 KB - jmir_v28i1e77809_app6.docx]

Checklist 1
STROBE checklist.
[DOCX File, 24 KB - jmir_v28i1e77809_app7.docx]

References

1. Jennings ELM, Murphy KD, Gallagher P, O'Mahony D. In-hospital adverse drug reactions in older adults; prevalence, presentation and associated drugs-a systematic review and meta-analysis. Age Ageing 2020 Oct 23;49(6):948-958. [doi: 10.1093/ageing/afaa188] [Medline: 33022061]
2. Long SJ, Brown KF, Ames D, Vincent C. What is known about adverse events in older medical hospital inpatients? A systematic review of the literature. Int J Qual Health Care 2013 Oct;25(5):542-554. [doi: 10.1093/intqhc/mzt056] [Medline: 23925507]
3. Andreotti F, Rocca B, Husted S, et al. Antithrombotic therapy in the elderly: expert position paper of the European Society of Cardiology Working Group on Thrombosis. Eur Heart J 2015 Dec 7;36(46):3238-3249. [doi: 10.1093/eurheartj/ehv304] [Medline: 26163482]
4. Hazell L, Shakir SAW. Under-reporting of adverse drug reactions: a systematic review. Drug Saf 2006;29(5):385-396. [doi: 10.2165/00002018-200629050-00003] [Medline: 16689555]
5. Murff HJ, Patel VL, Hripcsak G, Bates DW. Detecting adverse events for patient safety research: a review of current methodologies. J Biomed Inform 2003;36(1-2):131-143. [doi: 10.1016/j.jbi.2003.08.003] [Medline: 14552854]
6. Dyas AR, Zhuang Y, Meguid RA, et al. Development and validation of a model for surveillance of postoperative bleeding complications using structured electronic health records data. Surgery 2022 Dec;172(6):1728-1732. [doi: 10.1016/j.surg.2022.08.021] [Medline: 36150923]
7. Kashkoush J, Gupta M, Meissner MA, Nielsen ME, Kirchner HL, Garg T. Performance characteristics of a rule-based electronic health record algorithm to identify patients with gross and microscopic hematuria. Methods Inf Med 2023 Dec;62(5-06):183-192. [doi: 10.1055/a-2165-5552] [Medline: 37666279]
8. Hung CY, Lin CH, Chang CS, Li JL, Lee CC. Predicting gastrointestinal bleeding events from multimodal in-hospital electronic health records using deep fusion networks. Annu Int Conf IEEE Eng Med Biol Soc 2019 Jul;2019:2447-2450. [doi: 10.1109/EMBC.2019.8857244] [Medline: 31946393]
9. Salas M, Petracek J, Yalamanchili P, et al. The use of artificial intelligence in pharmacovigilance: a systematic review of the literature. Pharmaceut Med 2022 Oct;36(5):295-306. [doi: 10.1007/s40290-022-00441-z] [Medline: 35904529]

XSL•FO
RenderX

10. Shung D, Tsay C, Laine L, et al. Early identification of patients with acute gastrointestinal bleeding using natural language processing and decision rules. J Gastroenterol Hepatol 2021 Jun;36(6):1590-1597. [doi: 10.1111/jgh.15313] [Medline: 33105045]
11. Li R, Hu B, Liu F, et al. Detection of bleeding events in electronic health record notes using convolutional neural network models enhanced with recurrent neural network autoencoders: deep learning approach. JMIR Med Inform 2019 Feb 8;7(1):e10788. [doi: 10.2196/10788] [Medline: 30735140]
12. Mitra A, Rawat BPS, McManus DD, Yu H. Relation classification for bleeding events from electronic health records using deep learning systems: an empirical study. JMIR Med Inform 2021 Jul 2;9(7):e27527. [doi: 10.2196/27527] [Medline: 34255697]
13. Taggart M, Chapman WW, Steinberg BA, et al. Comparison of 2 natural language processing methods for identification of bleeding among critically Ill patients. JAMA Netw Open 2018 Oct 5;1(6):e183451. [doi: 10.1001/jamanetworkopen.2018.3451] [Medline: 30646240]
14. Walker AL, Watson C, Butcher R, Abedin Z, Yandell M, Shah RU. Use of commercially available natural language processing software to identify bleeding from the medical record. medRxiv. Preprint posted online on Nov 19, 2021. [doi: 10.1101/2021.11.19.21266586]
15. Yap AJY, Teo DCH, Ang PS, et al. Validation of a major and clinically relevant nonmajor bleeding phenotyping algorithm on electronic health records. Pharmacoepidemiol Drug Saf 2024 Aug;33(8):e5875. [doi: 10.1002/pds.5875] [Medline: 39090800]
16. Hung LC, Su YY, Sun JM, Huang WT, Sung SF. Clinical narratives as a predictor for prognosticating functional outcomes after intracerebral hemorrhage. J Neurol Sci 2023 Oct 15;453:120807. [doi: 10.1016/j.jns.2023.120807] [Medline: 37717279]
17. Lee HJ, Jiang M, Wu Y, et al. A comparative study of different methods for automatic identification of clopidogrel-induced bleedings in electronic health records. AMIA Jt Summits Transl Sci Proc 2017;2017:185-192. [Medline: 28815128]
18. Deng B, Zhu W, Sun X, et al. Development and validation of an automatic system for intracerebral hemorrhage medical text recognition and treatment plan output. Front Aging Neurosci 2022;14:798132. [doi: 10.3389/fnagi.2022.798132] [Medline: 35462698]
19. Hansen RS, Lynggaard RB, Laursen MS, Lykke FM, Vinholt PJ. Identification of hematuria with a natural language processing model and validation of hematuria diagnosecodes. Thromb Res 2024 Dec;244:109182. [doi: 10.1016/j.thromres.2024.109182] [Medline: 39426095]
20. Mehran R, Rao SV, Bhatt DL, et al. Standardized bleeding definitions for cardiovascular clinical trials: a consensus report from the Bleeding Academic Research Consortium. Circulation 2011 Jun 14;123(23):2736-2747. [doi: 10.1161/CIRCULATIONAHA.110.009449] [Medline: 21670242]
21. Le CHUV en chiffres [Web Page in French]. Centre Hospitalier Universitaire Vaudois. 2023. URL: https://www.chuv.ch/fr/a-propos/le-chuv-en-chiffres [accessed 2026-01-10]
22. Facts, figures and dates. Hôpitaux Universitaires De Genève. URL: https://www.hug.ch/en/facts-figures-and-dates [accessed 2026-01-10]
23. About the USZ 2025. Universitäts Spital Zürich. URL: https://www.usz.ch/en/about-university-hospital-zurich [accessed 2025-10-07]
24. KSB weiterhin auf wachstumskurs [Web Page in German]. Kantonsspital Baden. 2024. URL: https://www.kantonsspitalbaden.ch/news/ksb-weiterhin-auf-wachstumskurs [accessed 2026-01-10]
25. Gaspar F, Lutters M, Beeler PE, et al. Automatic detection of adverse drug events in geriatric care: study proposal. JMIR Res Protoc 2022 Nov 15;11(11):e40456. [doi: 10.2196/40456] [Medline: 36378522]
26. Schulman S, Kearon C, Subcommittee on Control of Anticoagulation of the Scientific and Standardization Committee of the International Society on Thrombosis and Haemostasis. Definition of major bleeding in clinical investigations of antihemostatic medicinal products in non-surgical patients. J Thromb Haemost 2005 Apr;3(4):692-694. [doi: 10.1111/j.1538-7836.2005.01204.x] [Medline: 15842354]
27. Kaatz S, Ahmad D, Spyropoulos AC, Schulman S, Subcommittee on Control of Anticoagulation. Definition of clinically relevant non-major bleeding in studies of anticoagulants in atrial fibrillation and venous thromboembolic disease in non-surgical patients: communication from the SSC of the ISTH. J Thromb Haemost 2015 Nov;13(11):2119-2126. [doi: 10.1111/jth.13140] [Medline: 26764429]
28. Honnibal M, Montani I. spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Sentometrics Res 2017;7(1):411-420 [FREE Full text]
29. Gaspar F, Zayene M, Coumau C, et al. Natural language processing and ICD-10 coding for detecting bleeding events in discharge summaries: comparative cross-sectional study. JMIR Med Inform 2025 Aug 29;13:e67837. [doi: 10.2196/67837] [Medline: 40882207]
30. Taffé P, Halfon P, Ghali WA, Burnand B, International Methodology Consortium for Coded Health Information (IMECCHI). Test result-based sampling: an efficient design for estimating the accuracy of patient safety indicators. Med Decis Making 2012;32(1):E1-12. [doi: 10.1177/0272989X11426176] [Medline: 22065144]
31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]

XSL•FO
RenderX

32. The use of the WHO-UMC system for standardised case causality assessment. : World Health Organization - Uppsala Monitoring Centre; 2018 URL: https://www.who.int/docs/default-source/medicines/pharmacovigilance/whocausality-assessment.pdf [accessed 2026-01-10]
33. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40(5):373-383. [doi: 10.1016/0021-9681(87)90171-8] [Medline: 3558716]
34. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care 1998 Jan;36(1):8-27. [doi: 10.1097/00005650-199801000-00004] [Medline: 9431328]
35. Ferrazzini E, Méan M, Stalder O, Limacher A, Rodondi N, Aujesky D. Incidence and clinical impact of bleeding events in older patients with acute venous thromboembolism. Blood Adv 2023 Jan 24;7(2):205-213. [doi: 10.1182/bloodadvances.2022007263] [Medline: 35381071]
36. Gómez-Cuervo C, Rivas A, Visonà A, et al. Predicting the risk for major bleeding in elderly patients with venous thromboembolism using the Charlson index. Findings from the RIETE. J Thromb Thrombolysis 2021 May;51(4):1017-1025. [doi: 10.1007/s11239-020-02274-6] [Medline: 32945982]
37. Poli D, Antonucci E, Bertù L, et al. Very elderly patients with venous thromboembolism on oral anticoagulation with VKAs or DOACs: results from the prospective multicenter START2-Register Study. Thromb Res 2019 Nov;183:28-32. [doi: 10.1016/j.thromres.2019.08.024] [Medline: 31536872]
38. Prasad N, Lau ECY, Wojt I, Penm J, Dai Z, Tan ECK. Prevalence of and risk factors for drug-related readmissions in older adults: a systematic review and meta-analysis. Drugs Aging 2024 Jan;41(1):1-11. [doi: 10.1007/s40266-023-01076-8] [Medline: 37864770]
39. Shehab N, Ziemba R, Campbell KN, et al. Assessment of ICD-10-CM code assignment validity for case finding of outpatient anticoagulant-related bleeding among Medicare beneficiaries. Pharmacoepidemiol Drug Saf 2019 Jul;28(7):951-964. [doi: 10.1002/pds.4783] [Medline: 31144403]
40. Joos C, Lawrence K, Jones AE, Johnson SA, Witt DM. Accuracy of ICD-10 codes for identifying hospitalizations for acute anticoagulation therapy-related bleeding events. Thromb Res 2019 Sep;181:71-76. [doi: 10.1016/j.thromres.2019.07.021] [Medline: 31357146]
41. Mitra A, Rawat BPS, McManus D, Kapoor A, Yu H. Bleeding entity recognition in electronic health records: a comprehensive analysis of end-to-end systems. AMIA Annu Symp Proc 2020;2020:860-869. [Medline: 33936461]
42. Xu W, Lv M, Wu S, et al. Severe bleeding risk of direct oral anticoagulants versus vitamin K antagonists for stroke prevention and treatment in patients with atrial fibrillation: a systematic review and network meta-analysis. Cardiovasc Drugs Ther 2023 Apr;37(2):363-377. [doi: 10.1007/s10557-021-07232-9] [Medline: 34436708]
43. Mant J. Process versus outcome indicators in the assessment of quality of health care. Int J Qual Health Care 2001 Dec;13(6):475-480. [doi: 10.1093/intqhc/13.6.475] [Medline: 11769750]

Abbreviations

ADE: adverse drug event
CHOP: Swiss Classification of Surgical Procedures
CHUV: Lausanne University Hospital
CRNMB: clinically relevant nonmajor bleeding
EMR: electronic medical record
HUG: Geneva University Hospital
***ICD-10-GM*:** *International Statistical Classification of Diseases, 10th Revision, German Modification*
ISTH: International Society on Thrombosis and Haemostasis
KSB: Baden Cantonal Hospital
MB: major bleeding
ML: machine learning
NLP: natural language processing
PPV: positive predictive value
SDA: structured data algorithms
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
USZ: Zürich University Hospital

XSL•FO
RenderX

Edited by A Schwartz; submitted 20.May.2025; peer-reviewed by M Forgerini, S Sazzad, U Modebelu; revised version received 18.Nov.2025; accepted 27.Nov.2025; published 29.Jan.2026.

Please cite as:
Coumau C, Gaspar F, Zayene M, Bertrand E, Alberio L, Lovis C, Beeler PE, Rinaldi F, Lutters M, Le Pogam MA, Csajka C, SwissMADE Collaborators
Detection of Antithrombotic-Related Bleeding in Older Inpatients: Multicenter Retrospective Study Using Structured and Unstructured Electronic Health Record Data
J Med Internet Res 2026;28:e77809
URL: https://www.jmir.org/2026/1/e77809
doi:10.2196/77809

© Claire Coumau, Frederic Gaspar, Mehdi Zayene, Elliott Bertrand, Lorenzo Alberio, Christian Lovis, Patrick E Beeler, Fabio Rinaldi, Monika Lutters, Marie-Annick Le Pogam, Chantal Csajka, SwissMADE Collaborators. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Behavioral Dynamics of AI Trust and Health Care Delays Among Adults: Integrated Cross-Sectional Survey and Agent-Based Modeling Study

Xueyao Cai[1*], MD; Weidong Li[1*], MD; Wenjun Shi[2], MD; Yuchen Cai[1*], MD; Jianda Zhou[1*], MD

[1]Department of Plastic Surgery, The Third Xiangya Hospital, Central South University, Changsha, China

[2]Department of Plastic and Reconstructive Surgery, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Jianda Zhou, MD
Department of Plastic Surgery
The Third Xiangya Hospital
Central South University
138 Tongzipo Road
Changsha
China
Phone: 86 13508493668
Email: zhoujianda@csu.edu.cn

Abstract

Background: While artificial intelligence (AI) holds significant promise for health care, excessive trust in these tools may unintentionally delay patients from seeking professional care, particularly among patients with chronic illnesses. However, the behavioral dynamics underlying this phenomenon remain poorly understood.

Objective: This study aims to quantify the influence of AI trust on health care delays through integrated survey-based mediation analysis and real-world research, and to simulate intervention efficacy using agent-based modeling (ABM).

Methods: A cross-sectional online survey was conducted in China from December 2024 to May 2025. Participants were recruited via convenience sampling on social media (WeChat and QQ) and hospital portals. The survey included a 21-item questionnaire measuring AI trust (5-point Likert scale), AI usage frequency (6-point scale), chronic disease status (physician-diagnosed, binary), and self-reported health care delay (binary). Responses with completion time <90 seconds, logical inconsistencies, missing values, or duplicates were excluded. Analyses included descriptive statistics, multivariable logistic regression (α=.05), mediation analysis with nonparametric bootstrapping (500 iterations), and moderation testing. Subsequently, an ABM simulated 2460 agents within a small-world network over 14 days to model behavioral feedback and test 3 interventions: broadcast messaging, behavioral reward, and network rewiring.

Results: The final sample included 2460 adults (mean age 34.46, SD 11.62 years; n=1345, 54.7% female). Higher AI trust was associated with increased odds of delays (odds ratio [OR] 1.09, 95% CI 1.00-1.18; *P*=.04), with usage frequency partially mediating this relationship (indirect OR 1.24, 95% CI 1.20-1.29; *P*<.001). Chronic disease status amplified the delay odds (OR 1.42, 95% CI 1.09-1.86; *P*=.01). The ABM demonstrated a bidirectional trust erosion loop, with population delay rates declining from 10.6% to 9.5% as mean AI trust decreased from 1.91 to 1.52. Interventions simulation found broadcast messaging most effective in reducing delay odds (OR 0.94, 95% CI 0.94-0.95; *P*<.001), whereas network rewiring increased odds (OR 1.04, 95% CI 1.04-1.05; *P*<.001), suggesting a "trust polarization" effect.

Conclusions: This study reveals a nuanced relationship between AI trust and delayed health care–seeking. While trust in AI enhances engagement, it can also lead to delayed care, particularly among patients with chronic conditions or frequent AI users. Integrating survey data with ABM highlights how AI trust and delay behaviors can strengthen one another over time. Our findings indicate that AI health tools should prioritize calibrated decision support rather than full automation to balance autonomy, odds, and decision quality in digital health. Unlike previous studies that focus solely on static associations, this research emphasizes the dynamic interactions between AI trust and delay behaviors.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e82170) doi:10.2196/82170

KEYWORDS

agent-based modeling; artificial intelligence; chronic disease; health care delay; real-world research

Introduction

Background

With the rapid development and widespread adoption of artificial intelligence (AI) in the medical field, AI has emerged as a pivotal tool across multiple domains, including health management, disease prediction, and clinical decision support [1,2]. Its potential to enhance health care efficiency, assist in disease diagnosis, optimize treatment strategies, and facilitate personalized health management is substantial [3,4]. However, this promise is tempered by growing concerns about its psychosocial and behavioral effects. The prevailing view of AI as an unequivocal force for good is being challenged by evidence that human-AI interaction can lead to new cognitive and behavioral issues [5,6]. In particular, AI's capacity to provide health recommendations and interventions based on extensive data analysis marks a significant breakthrough [7,8]. However, this ability may alter traditional health decision-making processes. Trust is a key concept for adopting technology, involving aspects such as performance, process, and purpose. A growing body of research indicates that while AI offers convenient access to health-related information, excessive reliance on these tools may lead to detrimental health behaviors, such as overreliance on automated suggestions and disregarding clinical intuition [9-11].

Traditional health decision-making models typically depend on professional medical judgment and individuals' self-awareness of their health status [12]. The advent of AI is reshaping this framework by providing automated diagnostic recommendations and personalized health guidance [13]. While such "digital health advice" can enhance confidence in medical decision-making in the short term, it may foster excessive dependence on AI tools over the long term. Patients with chronic diseases, who require ongoing health management and exhibit relatively stable symptoms, may be particularly susceptible to this shift [14]. This population often practices prolonged self-management, leading to frequent use of digital tools [15]. The relative stability of their conditions can create a sense of security, in which they may consider AI-generated reassuring feedback adequate. This dependence on AI could lead to delays in pursuing health care, as patients with chronic disease tend to give priority to AI-generated symptom evaluations and treatment options rather than traditional medical advice [16]. Consequently, they might overlook complex diagnostic and therapeutic needs that AI cannot fully assess [17]. Moreover, biases in medical AI can persist throughout its life cycle, potentially leading to serious repercussions in clinical decision-making. If these biases are not addressed, they can result in inaccurate medical judgments and exacerbate existing health care disparities [18]. While existing research has shed light on the potential risks of AI reliance in health care, a comprehensive framework that formally analyzes the complex relationship between AI trust and health behavior, particularly in the context of chronic disease, is still lacking. Additionally, few studies have used computational simulation to assess how various public health intervention strategies might mitigate this system-level risk.

This study aims to systematically analyze the interrelationships among AI trust, AI usage frequency, chronic disease status, and delayed health care–seeking behavior. The analysis was conducted through a cross-sectional online survey from December 2024 to May 2025. We propose that AI trust and frequency of use jointly contribute to the odds of delayed medical care through a behavioral feedback mechanism, particularly pronounced among patients with chronic conditions. High levels of AI trust may inadvertently lead individuals to postpone health care–seeking by increasing usage frequency, highlighting a complex interaction between trust and behavior. To investigate these dynamic interdependencies, we use an agent-based modeling (ABM) framework to simulate the trust-behavior feedback over time. By embedding individuals within social networks and allowing trust and delay behaviors to evolve together, the ABM enables us to examine how microlevel decisions accumulate into population-level odds. This includes issues such as collective delays or systemic trust collapse, which are often difficult to capture using conventional cross-sectional data. A schematic overview of the study design is provided in Figure 1.

XSL•FO
RenderX

Figure 1. Schematic overview of the study design and methodology. Based on online questionnaires, the analysis framework includes mediation analysis to explore the relationship between artificial intelligence (AI) trust, usage frequency, and delayed health care–seeking behavior. Additionally, agent-based modeling was used to model the dynamic feedback loops over a 14-day period, incorporating 3 intervention strategies: broadcast, reward, and rewire. These strategies aimed to evaluate how the interventions affected trust and delay behavior within the population. Image created in Biorender by WL [19].

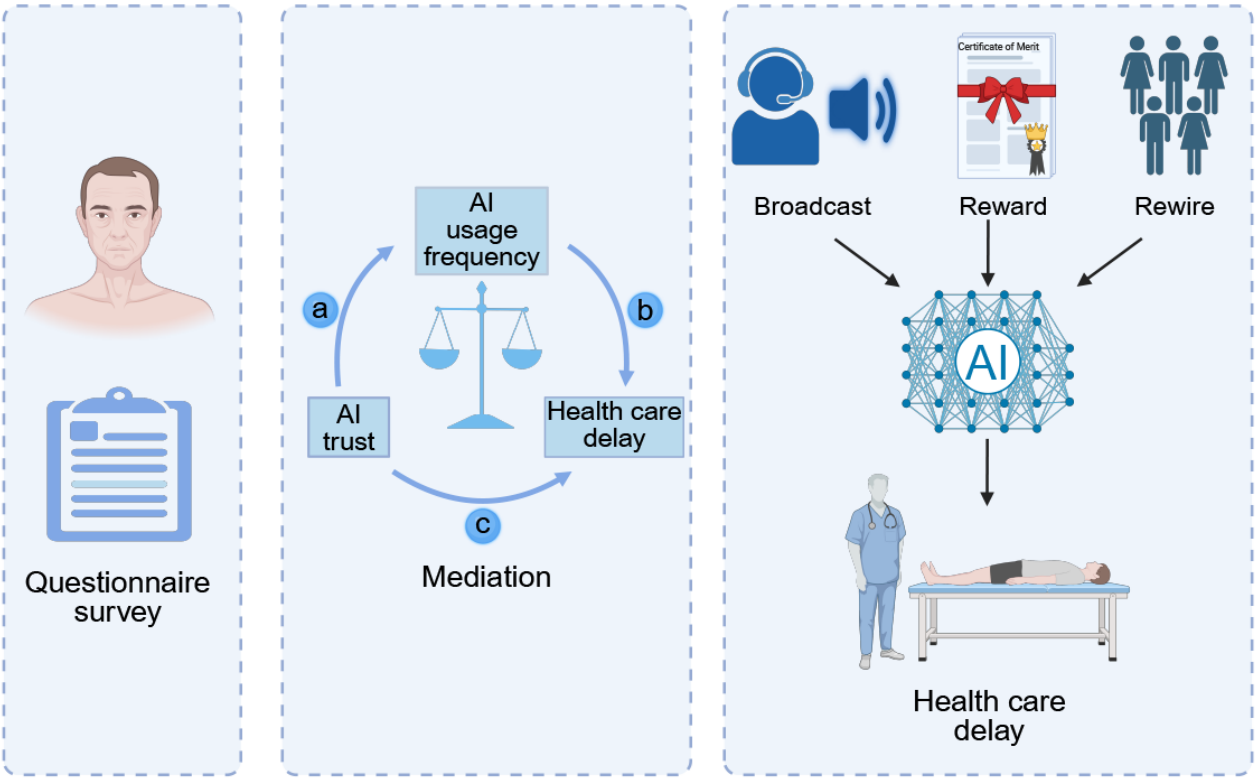

Objective

This study not only addresses a theoretical gap in understanding the impact of AI on delayed health care–seeking but also offers a fresh perspective on optimizing the design of AI-based health tools. A key challenge in future medical applications of AI will be achieving the right balance between fostering user trust and ensuring accuracy and timeliness in health care–seeking. We aim to provide theoretical support for designing AI-driven health interventions. Additionally, we seek to inform more effective public health strategies, ensuring that technological advancements promote rather than hinder appropriate health care decisions.

Methods

Survey Design and Data Collection

This study was conducted in 2 integrated phases: a cross-sectional survey and an ABM simulation. To ensure comprehensive and transparent reporting, the methods are reported following the Journal Article Reporting Standards guidelines under relevant subheadings [20,21]. A cross-sectional online survey was conducted between December 1, 2024, and May 20, 2025. The 21-item questionnaire was conducted in Mandarin Chinese and developed based on a systematic literature review and focus-group discussions. It underwent refinement through 2 rounds of expert review by 4 specialists in public health and AI, and was finalized after a pilot test involving 5 target participants. The full survey questionnaire (English translation) is provided in Multimedia Appendix 1.

Eligible participants were adults aged 18-75 years who had resided in mainland China for the past 12 months and were able to read Chinese. The age range was selected to include the digitally engaged adult population while excluding minors and the very old people, who may exhibit different health-seeking patterns. The 12-month residency criterion ensured consistent exposure to the local health care, minimizing confounding from recent immigration or transient populations. The survey was disseminated via WeChat (Tencent) and QQ (Tencent), and the patient-education portals of collaborating hospitals using convenience sampling. Participants accessed the questionnaire after reading an electronic informed consent form and selecting “Agree.” All items were mandatory, and skip logic was applied to minimize irrelevant questions. Responses were transmitted over HTTPS and encrypted at rest on the institution’s private server, accessible only to authorized investigators.

To ensure the validity and interpretability of the logistic regression and mediation models, which require complete-case data, we performed rigorous data cleaning. Questionnaires completed in under 90 seconds, or those containing logical inconsistencies, missing values, or duplicate entries (only the first complete entry was retained), were excluded. Given the limited amount of missing data and the potential risk of introducing bias through imputation, listwise deletion was considered the most appropriate and conservative approach. To examine the mechanism of missingness, we conducted Little’s

XSL•FO
RenderX

missing completely at random (MCAR) test using the *mice* package in R software (versions 4.2.3; R Foundation for Statistical Computing). Descriptive statistics were used to summarize the data (frequencies, percentages, means, and SDs), and multivariable logistic regression was conducted to examine the association between AI usage frequency and delayed health care–seeking behavior. Likert-type scales ranging from 1 to 5 (or 0 to 5 for AI usage and trust items) were used to assess frequency, trust level, and willingness to recommend.

Chronic disease was defined as any physician-diagnosed condition lasting 6 months or longer, including chronic allergic rhinitis, asthma, hypertension, diabetes, chronic skin diseases, etc. The variables included demographic characteristics (eg, age, sex, and occupation), AI usage behaviors (eg, frequency and exposure), trust perceptions, and health care–seeking outcomes.

Self-reported health care delay was assessed using a single binary item specifically targeting AI-influenced delay behaviors. Participants were asked: "Has advice provided by ChatGPT ever caused you to postpone or cancel seeking medical care?" with response options 0=no and 1=yes. This measure captures intentional delay or avoidance of health care that participants attributed directly to ChatGPT's recommendations, rather than delays caused by logistical or accessibility barriers. This operationalization aligns with behavioral delay frameworks used in prior studies and reflects clinically meaningful patterns of health care–seeking behavior.

Ethical Considerations

The study protocol was approved by the Institutional Review Board of Xiangya Third Hospital, Central South University (approval number 2025255). All procedures complied with the Declaration of Helsinki and the Personal Information Protection Law of China. Electronic informed consent was obtained from all participants. After reading a digital information sheet that outlined the study's purpose, procedures, risks, benefits, and their rights (including voluntary participation and withdrawal), participants indicated their agreement by selecting "Agree" before proceeding to the questionnaire. All collected data were deidentified. No personally identifiable information (eg, name, ID number, and contact details) was stored with the response data. Data were transmitted securely via HTTPS and stored in encrypted form on a private, access-controlled institutional server. Results are reported in aggregate to prevent any possibility of individual identification. Participants did not receive any financial or material compensation for their involvement in this study. The manuscript and its supplementary materials do not contain any images, videos, or textual data that could lead to the identification of an individual participant. Therefore, specific consent for the publication of identifiable information was not applicable.

Key Predictors of Delayed Care: Logistic Regression Analysis

We used logistic regression models to evaluate key predictors of delayed health care–seeking behavior. Initially, univariate logistic analyses were conducted to estimate the association between each candidate variable and the outcome, with results reported as odds ratios (ORs) accompanied by 95% CIs. Variables demonstrating potential significance were subsequently included in a series of hierarchical multivariate models (models 1-4). These models progressively incorporated individual characteristics, intervention exposure, AI usage patterns, and social influence to assess their independent contributions to delay behavior. This approach allowed us to identify the relative importance of each factor in influencing health care–seeking delays.

The data collection period encompassed the public release of DeepSeek, a major large language model in China, which occurred in early 2025 [22]. This event represented a significant shock to public awareness toward AI. To test the robustness of our core findings against this potential confounding effect, we performed a stratified analysis by dividing the sample into pre- and post-DeepSeek release subgroups. The cutoff date was set to February 1, 2025, allowing a sufficient time frame for the model's public impact to materialize within our survey window. We then reran the univariate logistic analyses within each subgroup to assess the consistency of the associations between AI trust, usage frequency, and health care delays.

Mediation Analysis: Indirect Effect of AI Trust via Usage Frequency

To examine whether AI trust influences delay behavior indirectly through AI usage frequency, we conducted mediation analysis using 2 approaches. First, following the traditional Baron and Kenny framework with the Sobel test, we estimated path *a* (the association between AI trust and usage frequency) via linear regression, and paths *b* and *c'* (the associations of usage frequency and AI trust with delay behavior) via multivariable logistic regression. The significance of the indirect effect ($a \times b$) was assessed using the Sobel *z* test. Second, to obtain robust CIs and significance estimates, we performed nonparametric bootstrap resampling (n=500). In each iteration, we reestimated paths *a* and *b*, calculated the product $a \times b$, and derived the empirical distribution of the indirect effect. The 95% bootstrap CI and *P* value were computed accordingly. All models were adjusted for age, gender, and chronic disease status.

Moderation by AI Recommendation Exposure: Stratified and Interaction Models

To further examine whether the level of AI recommendation exposure moderates the relationship between AI trust and delayed health care–seeking behavior, we conducted stratified logistic regression and interaction modeling. Participants were divided into low and high exposure groups based on their reported frequency of receiving AI recommendations (≤2 vs >2). Separate multivariable logistic regression models were fitted for each group, adjusting for age, gender, chronic disease status, and AI usage frequency. ORs and 95% CIs were reported to assess effect heterogeneity across exposure levels.

Additionally, we introduced an interaction term between AI trust and recommendation exposure into the full model. The interaction coefficient was used to calculate the interaction OR and the corresponding 95% CI. Predicted probability curves were plotted to visualize how the relationship between AI trust and delay behavior varies by exposure group, providing insights

XSL•FO
RenderX

into the moderating effect of AI recommendation exposure on health care–seeking behavior.

Moderation by Recommendation Willingness: Stratified Trust-Delay Associations

To evaluate whether individuals' willingness to recommend AI tools moderates the relationship between AI trust and delayed medical care, we conducted stratified logistic regressions and interaction analysis. Participants were grouped based on their scores (1-5) regarding the likelihood of recommending AI: those scoring ≤2 were classified as the "low recommendation intensity" group, and those scoring ≥3 as the "high recommendation intensity" group. Multivariable logistic regression models were fitted within each stratum, adjusting for age, gender, chronic disease status, and AI usage frequency. ORs and 95% CIs were calculated for AI trust in each group.

Subsequently, a full model including an interaction term between AI trust and recommendation intensity was constructed. We estimated the interaction effect (OR, 95% CI, and *P* value) and generated a prediction grid with standardized covariates to visualize the predicted probability of delayed care across AI trust levels in each group. Interaction plots were created to illustrate potential effect modification, enhancing our understanding of how willingness to recommend AI tools influences the relationship between AI trust and health care–seeking behavior.

Scenario-Based Modeling: Joint Effects of AI Behavior Factors on Delay

To theoretically evaluate the potential impact of various intervention strategies on delayed health care–seeking behavior, we conducted a scenario-based simulation analysis using logistic regression. A multivariable logistic model was initially constructed with AI trust, frequency of use, and chronic disease status as predictors. ORs, 95% CIs, and *P* values were reported for each explanatory variable.

Based on this model, the following six intervention scenarios were simulated: (1) baseline (trust=3, frequency=3, no chronic disease); (2) increased AI trust (set to 5); (3) increased frequency of AI use (set to 5); (4) chronic disease status switched to "yes"; (5) combined trust + frequency increase; and (6) combined trust + frequency + chronic disease. For each scenario, predicted probabilities of delay were computed across all individuals and averaged to reflect group-level effects.

The results were visualized using a bar plot displaying the mean predicted probability of delay under each strategy. Annotations highlighted the ORs and significance levels of the 3 key predictors, facilitating an intuitive understanding of their relative contributions to delay behavior. This comprehensive approach allowed us to identify the most effective interventions for reducing delays in health care–seeking.

ABM: Broadcast, Reward, Rewire

To evaluate the impact of different intervention mechanisms on delayed health care–seeking behavior, we used an ABM framework. ABM is a computational simulation approach particularly suited for studying complex systems, where population-level outcomes emerge from the interactions of autonomous, diverse individuals ("agents") operating within a defined environment and set of rules [23-25]. This method is particularly appropriate for our research question for 3 main reasons. First, AI trust is not a static trait; it is a dynamic belief influenced by personal experience and social factors, which ABM is designed to capture. Second, the decision to delay care involves weighing personal trust against the behaviors of peers, a scenario well modeled by embedding agents within a social network where attitudes and actions spread [26,27]. Third, ABM allows us to test the "trust-delay" feedback loop, a causal chain that cannot be directly identified from our cross-sectional survey data but can be explored through simulation [28].

To improve the transparency and reproducibility, the ABM model is described following the overview, design concepts, and details (ODD) protocol [29]. The purpose of the simulation was to examine how AI trust, peer influence, and intervention strategies jointly shape delayed health care–seeking behavior over time. Each agent represented 1 survey respondent and was initialized using the individual's empirical AI trust (1-5), AI usage frequency (1-5), and chronic disease status. Agents were embedded in a Watts-Strogatz small-world social network (n=2460; average degree=4; rewiring probability=0.2), which captures realistic social clustering and intermittent long-distance ties. The simulation progressed in daily cycles for 14 days, during which agents first computed a probability of delay using a logistic model derived from the survey data, then made a probabilistic delay decision, and subsequently updated their trust and usage behaviors according to personal outcomes, peer context, and intervention conditions. A total of 100 repetitions were performed for each scenario.

The model incorporated key design concepts of agent-based systems. Interaction occurred exclusively through local network neighbors; both average neighbor trust and neighbor delay behaviors influenced an individual's own updates. Stochasticity was present in network initialization, delay decisions, and trial-level replications. Agents adapted their trust and usage frequency over time, increasing them when surrounded by high-trust peers or after not delaying care, and decreasing them when neighbors frequently delayed or after personally delaying care. No explicit learning mechanism was included; behavioral dynamics emerged entirely from rule-based adaptation. Model outputs consisted of daily population-level delay rates and comparative effects of intervention strategies relative to baseline.

Detailed implementation followed ODD guidelines [29]. At initialization, agents with an empirically predicted baseline delay probability above 0.20 were assigned a 1-point reduction in trust and usage frequency to represent structural vulnerability. During each daily update, delaying care reduced trust by 0.2 and usage frequency by 0.3, whereas not delaying increased trust by 0.1; all values were bounded between 1 and 5. A total of 3 intervention strategies were embedded into this baseline framework. In the broadcast condition, trust was reduced daily by a small fixed penalty to reflect exposure to trust-eroding messages. The reward condition increased trust and usage frequency for agents who consistently sought timely care, with rewards provided every 2 days. In the rewiring condition, network edges were periodically redirected toward the highest-trust agents to model opinion leader amplification. All

XSL•FO
RenderX

intervention rules were implemented on top of the core behavioral update mechanism. Finally, 1-way sensitivity analyses were conducted by varying initial trust levels (means of 2.5, 3.0, and 3.5), broadcast penalty intensities (0.05-0.20 per day), reward magnitudes (0.03-0.10), and rewiring frequencies (intervals of 2-10 days). Each parameter set was simulated in 100 trials, and the resulting delay trajectories were compared to assess model robustness.

Results

Demographic Characteristics, AI Use, and Health Decision Outcomes

Of 2785 initial submissions, 325 (11.7%) responses were excluded based on prespecified criteria (completion time <90 seconds, logical inconsistencies, duplicate entries, or missing values in key variables). Specifically, 136 (4.9%) exclusions from the initial sample were due to missing values. Little's MCAR test indicated that the data were MCAR (χ^2_{14}=12.87; *P*=.54), supporting the use of complete-case analysis. Consequently, a total of 2460 valid responses were retained, predominantly female (n=1345, 54.7%), with an average age of 34.46 (SD 11.62) years. Occupations included students (n=825, 33.5%), technology workers (n=715, 29.1%), other professions (n=640, 26%), and health care personnel (n=280, 11.4%). A significant majority (n=2215, 90%) reported being aware of generative AI tools, and 62% (n=1525) had previously used AI for health advice. Regarding the frequency of AI-based advice use, 38% (n=935) of participants never used it, while 16.2% (n=398) used it weekly (1-2 times) and 15.7% (n=385) monthly. Trust in AI-generated advice varied, with 38% (n=935) never using it, and the remaining participants distributed across differing trust levels (Table 1).

In terms of health care decision outcomes, 11.6% (285/2460) of participants deferred or canceled health care due to AI advice, while 18.9% (465/2460) changed their health care decisions based on it. Additionally, 16.9% (415/2460) adopted alternative therapies following AI recommendations. The primary source of health information was physicians (1203/2460, 48.9%), followed by search engines (465/2460, 18.9%) and generative AI tools (375/2460, 15.2%). Perceived influence of online discussions varied, with 19.7% (485/2460) considering it very likely to impact their decisions. Lastly, 30.1% (740/2460) of participants reported having a chronic disease (Table 2).

XSL•FO
RenderX

Table 1. Demographic characteristics and artificial intelligence (AI) use (n=2460).

Variable	Values
Sex, n (%)	
Female	1345 (54.7)
Male	1115 (45.3)
Age (years), mean (SD)	34.46 (1.62)
Occupation category, n (%)	
Students	825 (33.5)
Technology workers	715 (29.1)
Other	640 (26)
Health care personnel	280 (11.4)
Awareness of generative-AI tools, n (%)	
Yes	2215 (90)
No	245 (10)
Previous use of AI for health advice, n (%)	
Yes	1525 (62)
No	935 (38)
Frequency of AI-based advice use, n (%)	
Never	935 (38)
Weekly 1-2 times	398 (16.2)
Monthly	385 (15.7)
Weekly 3-4 times	308 (12.5)
Occasional	293 (11.9)
Almost daily	141 (5.7)
Trust in AI-generated advice, n (%)	
1 (“none”)	320 (13)
2	290 (11.8)
3	280 (11.4)
4	345 (14)
5 (“high”)	290 (11.8)
0 (“never used”)	935 (38)

XSL•FO
RenderX

Table 2. Health decision outcomes and contextual factors (n=2460).

Variable	Value, n (%)
Deferred or cancelled health care due to AI[a] advice	
Yes	285 (11.6)
No	2175 (88.4)
Changed health care decision due to AI advice	
Yes	465 (18.9)
No	1995 (81.1)
Adopted alternative therapy due to AI advice	
Yes	415 (16.9)
No	2045 (83.1)
Primary source of health information	
Physician	1203 (48.9)
Search engines	465 (18.9)
Generative AI tools	375 (15.2)
Family or friends	255 (10.3)
Government or traditional media	85 (3.5)
Social networking	77 (3.1)
Perceived influence of online discussions	
Very likely	485 (19.7)
Likely	835 (33.9)
Uncertain	775 (31.5)
Unlikely	265 (10.8)
Very unlikely	100 (4.1)
Presence of chronic disease	
Yes	740 (30.1)
No	1720 (69.9)

[a]AI: artificial intelligence.

AI Trust and Usage Frequency Are Key Predictors of Delay

In the univariate logistic regression, higher AI trust was positively associated with delayed health care–seeking behavior, with an OR of 1.27 (95% CI 1.19-1.36; *P*<.001; Figure 2A). Similarly, the frequency of AI use also showed a significant positive correlation with delay (OR 1.41, 95% CI 1.18-1.69; *P*<.001; Figure 2A). Although not statistically significant, trends were observed in the effects of receiving AI recommendations (OR 0.95, 95% CI 0.81-1.11) and actively recommending AI (OR 0.92, 95% CI 0.781.09; Figure 2A), both suggesting a potential reduction in delay. In the fully adjusted model (model 4), AI trust remained a significant predictor (OR 1.09, 95% CI 1.00-1.18; *P*=.04; Figure 2B). Frequency of AI use exhibited the strongest association with delay (OR 1.39, 95% CI 1.14-1.68; *P*<.001; Figure 2B), indicating that more frequent users had higher odds of postponing medical visits. Additionally, chronic disease status demonstrated a persistent positive association with delay (OR 1.43, 95% CI 1.10-1.88; *P*=.009; Figure 2B).

Detailed analyses for univariate and multivariate logistic regression can be found in Tables S1 and S2 in Multimedia Appendix 2.

To address potential confounding from a major AI market event, we stratified the analysis by the DeepSeek release period. The results confirmed the robustness of our primary findings. The association between AI usage frequency and delayed health care–seeking remained nearly identical in both direction and magnitude in the pre- and postrelease periods (prerelease: OR 1.41, 95% CI 1.17-1.70; *P*<.001; postrelease: OR 1.41, 95% CI 1.27-1.56; *P*<.001). For AI trust, while the positive association was slightly attenuated and not statistically significant in the prerelease period (OR 1.05, 95% CI 0.89-1.25; *P*=.56), it was significant in the postrelease period (OR 1.10, 95% CI 1.00-1.21; *P*=.046). This suggests that the fundamental behavioral mechanism linking AI trust and usage to delay is robust. The DeepSeek release may have slightly amplified the measurable effect of AI trust, possibly due to increased public reliance on AI tools. Detailed results of this stratified analysis are available in Table S3 in Multimedia Appendix 2.

XSL•FO
RenderX

Figure 2. Association between artificial intelligence (AI) trust, usage frequency, and delayed health care–seeking behavior. (A) Univariate logistic regression and (B) fully adjusted multivariable model evaluating key predictors of delayed health care–seeking behavior.

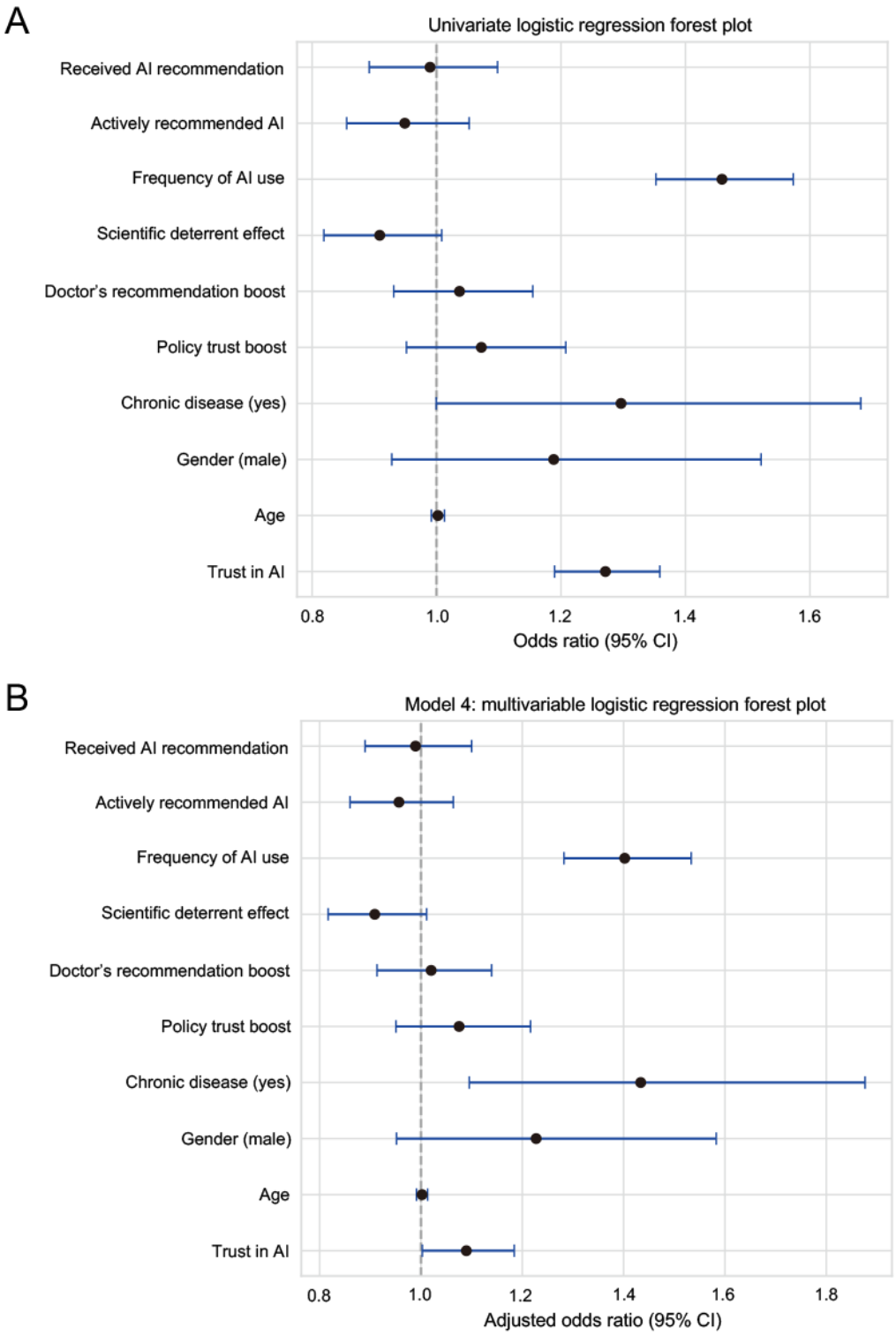

AI Trust Influences Delay via Frequency: Evidence of Mediation

We further evaluated the indirect effect of AI trust on delayed health care–seeking through AI usage frequency, supporting a partial mediation model. In the linear regression analysis, AI trust significantly predicted usage frequency (path *a*: β=.5754; *P*<.001). In the multivariable logistic model, usage frequency was also significantly associated with delay (path *b*: OR 1.40, 95% CI 1.28-1.55; *P*<.001). The product term (indirect effect $a \times b$) was calculated to be 0.1949, and the Sobel test yielded a significant result (*P*<.001), indicating a statistically robust mediation path (Table 3). To further validate these findings, a nonparametric bootstrap analysis with 500 replications was conducted. The mean indirect effect was found to be 0.2152 (OR 1.24, 95% CI 1.20-1.29), with a bootstrap *P* value of <.001.

XSL•FO
RenderX

This confirms that AI trust contributes to delay behavior partly through increased frequency of AI usage. The direct effect of AI trust on delay remained statistically significant after adjusting for the mediator (path *c'*: OR 1.09, 95% CI 1.00-1.18; *P*=.04), supporting the conclusion of a partial mediation model (Table 3). A diagram illustrating this mediation relationship can be found in Figure 3.

Table 3. Mediation effect of artificial intelligence (AI) trust on health care delay behavior.

Path	β (log OR[a])	OR (95% CI)	SE	*P* value	*P* value (Sobel)
a (AI trust-AI frequency)	.5754	N/A[b]	0.014	<.001	N/A
b (AI frequency-delay)	.3387	1.4031 (1.28-1.53)	0.0455	<.001	N/A
a × *b* (indirect)	.1949	N/A	0.0266	N/A	<.001
c (AI trust-delay, total effect)	.2473	1.2806 (1.20-1.37)	N/A	N/A	N/A
c' (AI trust-delay, direct effect)	.085	1.0887 (1.0-1.18)	0.0423	.04	N/A

[a]OR: odds ratio.

[b]N/A: not applicable.

Figure 3. Mediation model demonstrating how artificial intelligence (AI) trust influences delayed health care–seeking through increased frequency of AI usage. Path *a* shows the significant effect of AI trust on usage frequency, while path *b* indicates the association between usage frequency and delay. The indirect effect (*a* × *b*) and direct effect (*c'*) are also represented, highlighting the statistical significance of the mediation pathway. Image created in Biorender by WL [30]. OR: odds ratio.

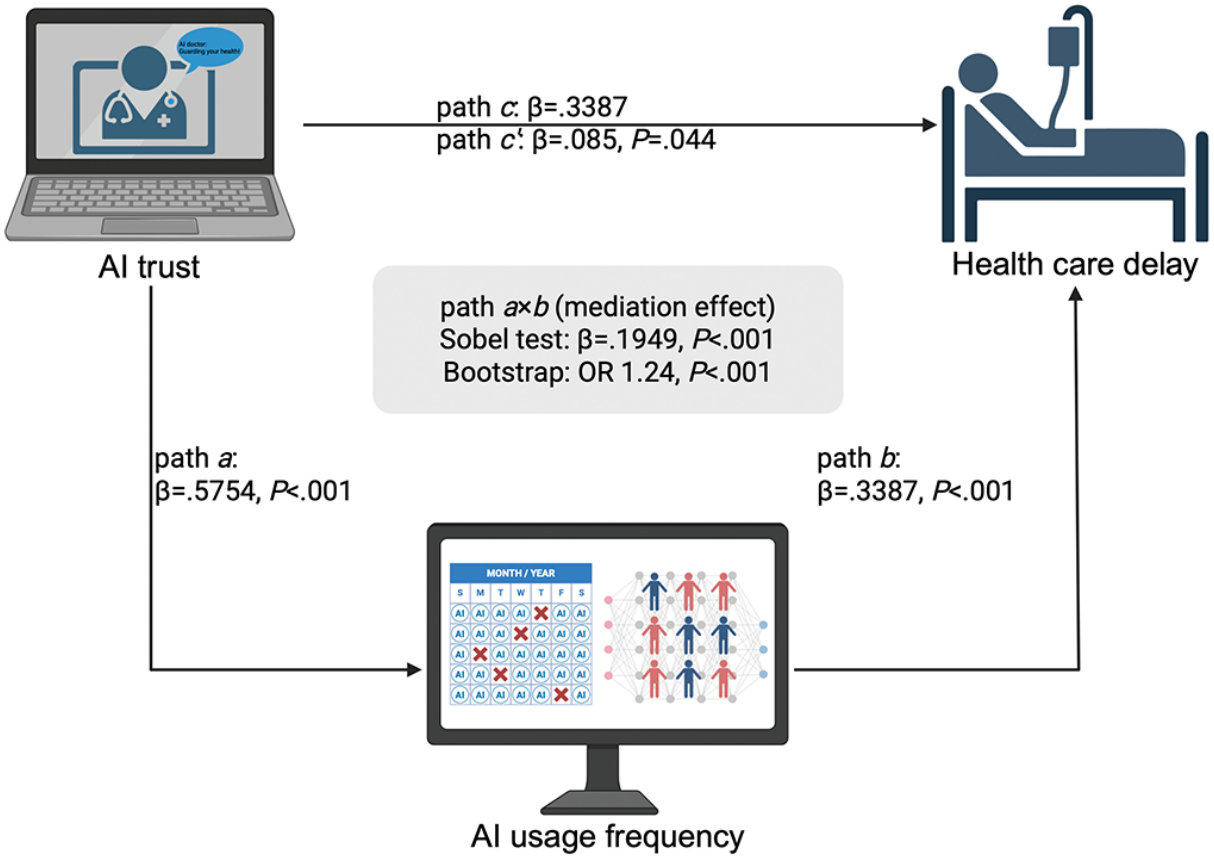

No Significant Moderation by Recommendation Exposure Level

In the stratified analysis by recommendation exposure, AI trust was significantly associated with delayed health care–seeking in the high exposure group (OR 1.11, 95% CI 1.01-1.23; *P*=.03), but not in the low exposure group (OR 1.04, 95% CI 0.88-1.22; *P*=.67). Notably, frequency of AI use was consistently associated with delay across both groups (low: OR 1.47, 95% CI 1.24-1.73; *P*<.001; high: OR 1.38, 95% CI 1.24-1.53; *P*<.001), suggesting a stable effect regardless of exposure level (Table 4). In the full model with interaction terms, the AI trust × recommendation exposure interaction was not statistically significant (interaction OR 0.97, 95% CI 0.83-1.14; *P*=.75; Figure 4A; detailed stratified analyses can be found in Table S4 in Multimedia Appendix 2). As shown in the predicted probability plot, the slopes of AI trust

XSL•FO
RenderX

on delay were similar between groups, with overlapping CIs, supporting the absence of a significant interaction (Figure 4A).

In summary, although AI trust showed a stronger association with health care delay in the high exposure group, the overall model did not indicate a statistically significant moderation effect. This suggests that while recommendation exposure may influence the magnitude of the trust-delay relationship, it does not alter its fundamental nature.

Table 4. Stratified analysis by recommendation exposure for the artificial intelligence (AI) trust-health care delay associations.

Outcome and variable	OR[a] (95% CI)	*P* value
Low recommendation exposure		
Trust in AI	1.04 (0.88-1.22)	<.001
Frequency of AI use	1.47 (1.24-1.73)	.67
Chronic disease (yes)	0.80 (0.47-1.37)	<.001
Age	1.01 (0.99-1.04)	.42
Gender (male)	1.35 (0.84-2.19)	.16
High recommendation exposure		
Trust in AI	1.11 (1.01-1.23)	.21
Frequency of AI use	1.38 (1.24-1.53)	<.001
Chronic disease (yes)	1.78 (1.30-2.44)	.03
Age	1.00 (0.99-1.01)	<.001
Gender (male)	1.17 (0.87-1.58)	<.001

[a]OR: odds ratio.

XSL•FO
RenderX

Figure 4. Stratified logistic regression and interaction modeling between artificial intelligence (AI) trust and recommendation exposure. Moderating effect of (A) AI recommendation exposure and (B) recommendation intensity on health care–seeking behavior. OR: odds ratio.

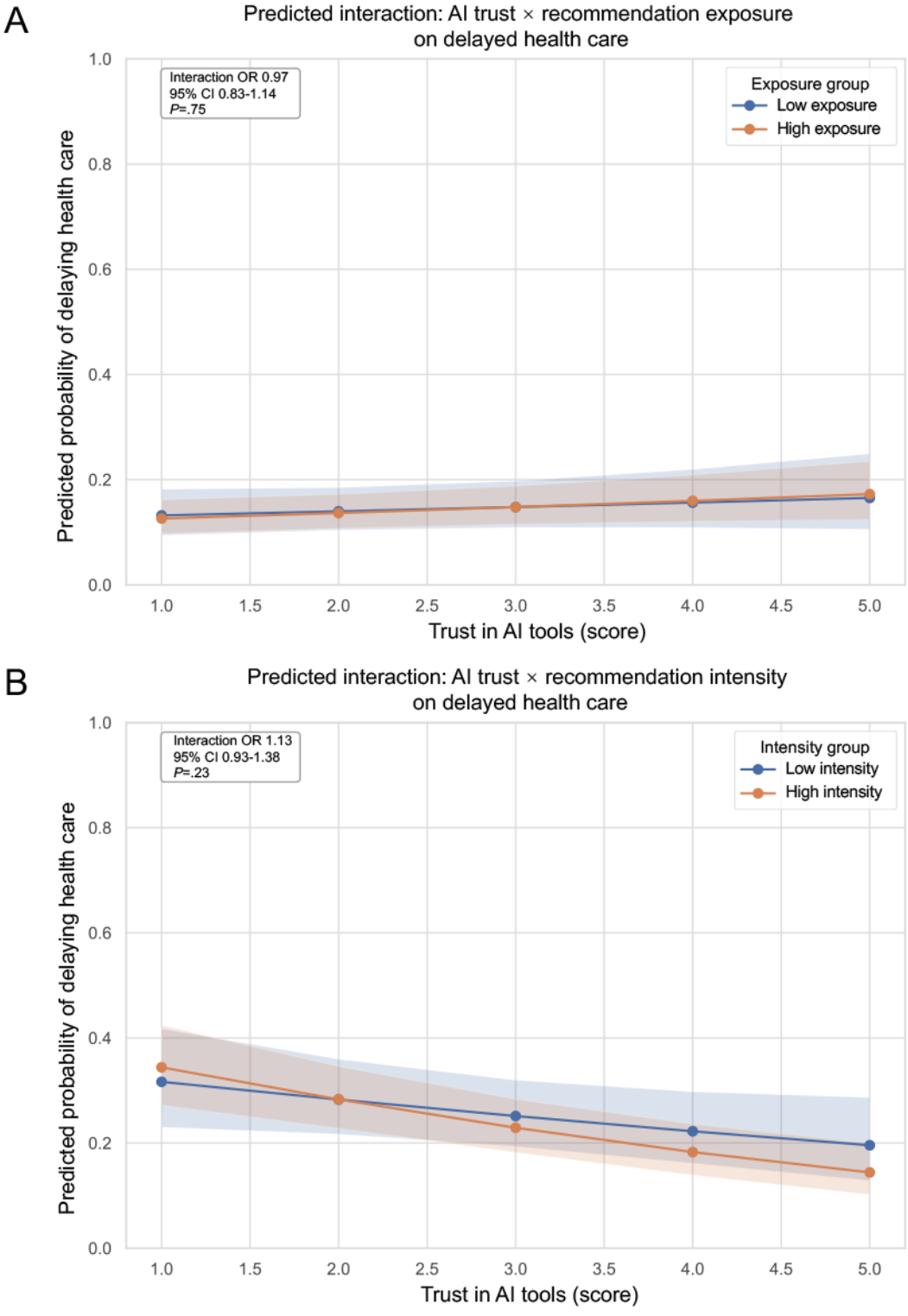

Limited Moderation by Recommendation Willingness Intensity

When stratified by willingness to recommend AI tools, AI trust was significantly associated with higher odds of health care delay in the low-intensity group (OR 1.16, 95% CI 1.00-1.34; *P*=.047), but not in the high-intensity group (OR 1.06, 95% CI 0.96-1.17; *P*=.27). Frequency of AI use remained a significant predictor of delay across both strata (low: OR 1.35, 95% CI 1.15-1.57; *P*<.001; high: OR 1.43, 95% CI 1.28-1.60; *P*<.001; Table 5). In the interaction model, the interaction term between AI trust and recommendation intensity was not statistically significant (interaction OR 1.13, 95% CI 0.93-1.38; *P*=.23; Figure 4B; detailed stratified analyses can be found in Table S5 in Multimedia Appendix 2). The interaction plot showed that the predicted probability of delaying care decreased with

XSL•FO
RenderX

increasing AI trust in both groups, with overlapping CIs, indicating no significant moderation effect (Figure 4B).

In summary, while AI trust appears to be more strongly associated with the odds of health care delay in the low-intensity group, recommendation intensity did not significantly moderate this association statistically. Its influence may reflect subtle variation in effect size rather than a true interaction.

Table 5. Stratified analysis by recommendation intensity for the artificial intelligence (AI) trust-health care delay associations.

Outcome and variable	OR[a] (95% CI)	*P* value
Low recommendation intensity		
Trust in AI	1.16 (1.00-1.34)	.047
Frequency of AI use	1.35 (1.15-1.57)	<.001
Chronic disease (yes)	1.53 (0.94-2.48)	.09
Age	1.00 (0.98-1.02)	.75
Gender (male)	1.07 (0.68-1.68)	.79
High recommendation intensity		
Trust in AI	1.06 (0.95-1.17)	.27
Frequency of AI use	1.43 (1.28-1.60)	<.001
Chronic disease (yes)	1.41 (1.02-1.95)	.04
Age	1.00 (0.99-1.02)	.71
Gender (male)	1.30 (0.96-1.77)	.09

[a]OR: odds ratio.

Scenario Simulations Reveal Combined Risk of AI Trust, Frequency, and Chronic Disease

In the multivariable logistic regression analysis, all 3 key predictors—AI trust, frequency of use, and chronic disease status—were significantly associated with health care delay. Specifically, each unit increase in AI trust was linked to 9% higher odds of delay (OR 1.09, 95% CI 1.00-1.18; *P*=.04). Frequency of AI use demonstrated an even stronger association (OR 1.40, 95% CI 1.28-1.53; *P*<.001), and chronic disease status significantly increased the odds of delay as well (OR 1.42, 95% CI 1.09-1.86; *P*=.01; Table 6).

Table 6. Multivariable logistic regression analysis for the associations between key predictors and health care delay.

Variable	OR[a] (95% CI)	*P* value
Trust in AI[b]	1.09 (1.00-1.18)	.04
Frequency of AI use	1.40 (1.28-1.53)	<.001
Chronic disease (yes)	1.42 (1.08-1.86)	.01

[a]OR: odds ratio.

[b]AI: artificial intelligence.

Scenario simulations indicated that, compared to a baseline of moderate trust and usage without a chronic condition (predicted probability=11.6%), increasing either AI trust (13.97%) or frequency (25.2%) alone raised the likelihood of delayed health care. The effect was most pronounced with frequency increases. When chronic illness was present alone, the predicted delay probability rose to 14.1%. Combining high trust and high frequency led to a delay probability of 30.5%, and when all 3 factors—high trust, high frequency, and chronic illness—were present, the probability escalated to 35.8%. These findings suggest that while increased trust and usage may enhance engagement with AI, they may paradoxically elevate the odds of behavioral delay due to potential overreliance or false reassurance. Therefore, intervention strategies should carefully balance cognitive trust with medical decision accuracy to mitigate potential adverse outcomes.

Agent-Based Simulation Reveals Bidirectional Feedback Between Trust and Delay

Simulation results demonstrated a consistent decline in the overall rate of health care delay, decreasing from 10.6% on day 1 to 9.5% by day 14 (Figure 5A). This trend reflects behavioral feedback, where agents who experienced delays reduced their AI usage frequency, ultimately lowering delay probabilities at the population level. Concurrently, the mean AI trust score declined steadily from approximately 1.95 to 1.49 over the 14-day period (Figure 5B). This indicates a natural erosion of trust in the absence of external reinforcement, suggesting a feedback loop where behavioral delays contribute to progressive trust deterioration. When stratifying agents by their predicted baseline delay probability, both high-risk and low-risk groups exhibited declining trust; however, the high-risk group experienced a steeper decline (from 2.55 to 1.74) compared to

XSL•FO
RenderX

the low-risk group (from 1.84 to 1.53; Figure 5C). This suggests that high-risk individuals may fall into a "high expectation-high disappointment" loop, accelerating trust collapse and reinforcing delay behavior. Detailed analyses for ABM can be found in Table S6 in Multimedia Appendix 2.

Figure 5. Agent-based modeling for the bidirectional feedback between artificial intelligence (AI) trust and delay behavior over a 14-day simulation period. (A) The overall rate of delayed health care. (B) The mean AI trust score. (C) AI trust score for the high-risk and low-risk groups.

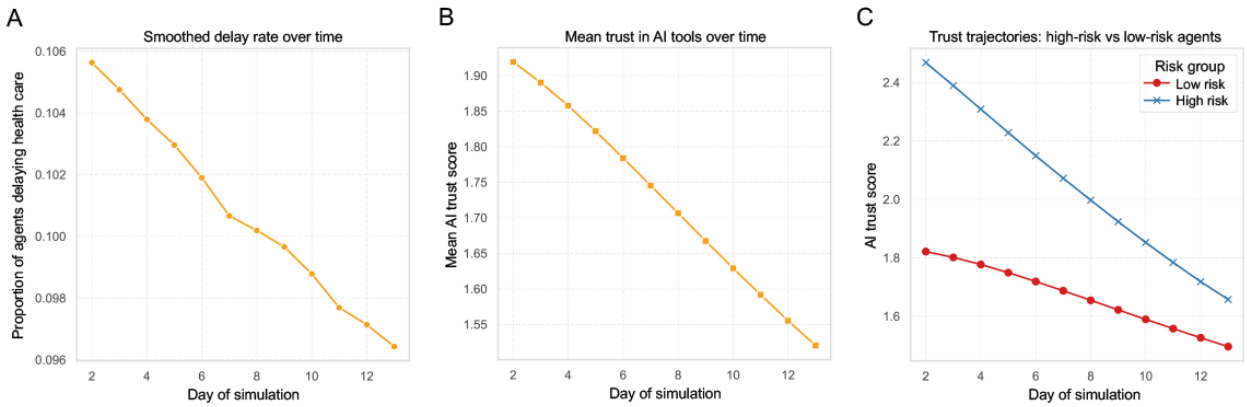

Overall, these simulations demonstrate the reciprocal feedback between trust and behavior, with individual heterogeneity compounding over time. The findings highlight the potential need for risk-stratified interventions or strategies designed to counteract trust erosion to prevent systemic trust breakdown.

Strategy Comparison: Broadcast Is the Most Effective, While Rewire May Backfire

Across 100 simulation trials, we examined the temporal effects of 4 intervention strategies on delay behavior. The broadcast strategy resulted in the most significant reduction in delay rates, reaching approximately 9.7% by day 14 (Figure 6). The reward strategy showed a moderate effect, with a slight downward trend compared to the baseline (Figure 6). In contrast, the rewire strategy led to an upward trend in delay rates after day 7, ultimately surpassing the baseline group, which suggests a potential amplification of trust polarization (Figure 6).

Figure 6. Comparison of 4 intervention strategies on delay behavior using artificial intelligence simulations. Strategies include baseline, broadcast messaging, behavioral reward, and network rewiring.

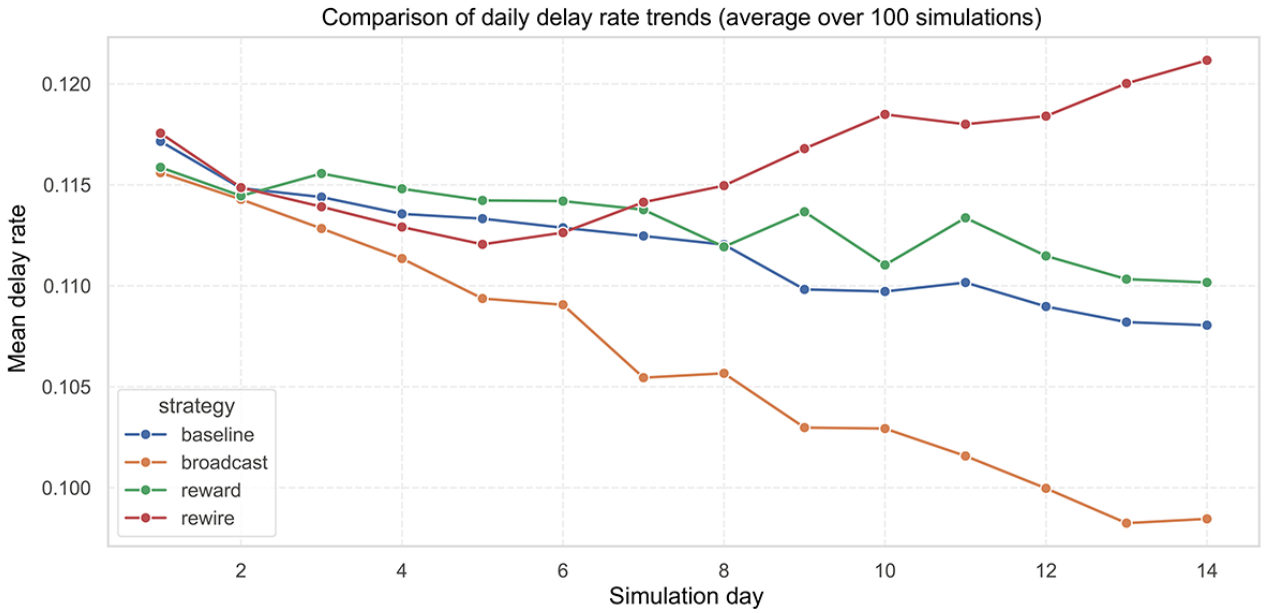

Logistic regression further quantified these effects (Table 7). Compared to the baseline, the broadcast strategy significantly reduced the odds of delay (OR 0.94, 95% CI 0.94-0.95; *P*<.001). The reward strategy indicated a slight increase in the odds of delay (OR 1.01, 95% CI 1.01-1.02; *P*<.001), while the rewire strategy demonstrated the most substantial increase in odds (OR 1.04, 95% CI 1.04-1.05; *P*<.001), indicating that this approach may exacerbate delay behavior under certain conditions. In summary, the broadcast strategy emerged as the most effective in reducing delay behavior within the current simulation framework, highlighting the potential advantages of proactive information signaling and group-level risk awareness alerts.

XSL•FO
RenderX

Table 7. Logistic regression analysis for the intervention strategies on health care delay.

Strategy	OR[a] (95% CI)	*P* value
Broadcast	0.94 (0.94-0.95)	<.001
Reward	1.01 (1.01-1.02)	<.001
Rewire	1.04 (1.04-1.05)	<.001

[a]OR: odds ratio.

To validate the robustness of the simulation across varying intervention configurations, we performed sensitivity analyses on 4 key parameters: initial trust level, broadcast penalty intensity, reward magnitude, and rewiring frequency. When varying the initial trust level (2.5, 3.0, and 3.5), higher values led to significantly increased rates of health care delay over time, suggesting that excessive trust may backfire despite its general benefits (Figure S1A in Multimedia Appendix 3). For broadcast penalty intensity, tested at values of 0.05, 0.1, and 0.2, the delay trajectories remained nearly identical, indicating minimal impact of penalty strength on intervention stability (Figure S1B in Multimedia Appendix 3). Similarly, altering the reward magnitude (0.03, 0.05, and 0.10) produced overlapping delay curves, demonstrating the strategy's robustness and absence of nonlinear effects (Figure S1C in Multimedia Appendix 3). In the case of rewiring frequency (2, 5, and 10 days), more frequent rewiring only slightly reduced early-stage delays, with long-term outcomes remaining consistent across intervals (Figure S1D in Multimedia Appendix 3). Overall, these results confirm that the simulation framework is stable, with behavioral outcomes staying similar across a range of parameters, highlighting the broader relevance and effectiveness of the modeled interventions.

Discussion

Principal Findings

This study is the first to systematically investigate the joint influence of AI trust, AI usage frequency, and chronic disease status on predicting delayed health care–seeking from the perspective of a behavior-trust feedback mechanism in China. Our findings demonstrate that both AI trust and usage frequency are significant predictors of health care delay. Importantly, AI trust not only indirectly elevates the odds of delay through increased usage frequency but also exerts a significant direct association when controlling for usage frequency, indicating a partial mediation pathway. Of note, individuals with chronic conditions inherently exhibit a predisposition toward delayed care. These odds are further exacerbated by the combined effects of high AI trust and frequent AI use. These results expand the current understanding of delayed health care behaviors. They emphasize the complex and dynamic interplay between technological trust and health-related decision-making. Additionally, they highlight that enhanced trust in AI may inadvertently contribute to adverse behavioral outcomes among patients with chronic diseases.

To investigate the psychological mechanisms underlying the "AI trust-frequent AI use-health care delay" pathway, we propose that cognitive bias and overreliance are key driving factors [31,32]. On one hand, patients with chronic conditions often experience stable and slowly progressing symptoms. They are more likely to view AI's reassuring suggestions as indicators of safety. As a result, they may underestimate the seriousness of their own symptoms. This sense of "false reassurance" is particularly pronounced among individuals with high levels of trust in AI [33]. Previous studies have shown that excessive trust in health AI recommendations can result in users neglecting bodily warning signs, ultimately leading to delays in health care–seeking [5]. On the other hand, frequent AI usage does not necessarily reflect higher health literacy. It also does not always indicate better self-management capacity [34]. Instead, it may suggest a psychological tendency to avoid traditional health care services. This tendency is especially common when medical care is expensive, time-consuming, or perceived as untrustworthy [35]. This form of "instrumental dependence" may drive individuals to rely on AI as a substitute source of advice when experiencing discomfort, instead of seeking timely professional care [36]. Therefore, this study emphasizes that the behavioral consequences of AI trust and usage frequency should not be interpreted simply as empowerment. Instead, it is important to explore the underlying psychological and behavioral mechanisms. This is essential for accurately assessing the true impact of technological interventions on health care behavior.

Although we further investigated the moderating effects of AI recommendation exposure and recommendation willingness on the "AI trust-health care delay" pathway, the interaction terms were not statistically significant. This indicates that these factors have a limited influence on the primary effect pathway. Subgroup analyses revealed some trend-level differences; for instance, individuals with higher recommendation willingness showed a slight reduction in delay behavior [37], but these effects did not achieve statistical significance. We speculate that a nonlinear threshold effect may be present. Once recommendation exposure reaches a certain frequency, its impact may plateau. This could result in a failure to further enhance trust or promote behavioral change. Another possibility is that the effectiveness of recommendations is highly context dependent. They may exert influence primarily when users experience high health anxiety, evident symptoms, or a strong sense of urgency [38,39]. Additionally, AI trust may be inherently unstable among individuals, easily influenced by emotional states, prior experiences, or public opinion [40]. These factors may further diminish the practical effectiveness of recommendations. While recommendation behaviors have the potential for positive influence, their underlying mechanisms appear more complex [41]. They may not adequately address the fundamental contradiction between AI trust and health care delay. Therefore, future interventions should not concentrate solely on increasing recommendation frequency or willingness.

XSL•FO
RenderX

Instead, it is crucial to address system-level design issues. This includes enhancing decision transparency and improving users' perceived control over health-related choices. Additionally, developing prompting strategies that better align with users' psychological models is essential. Such improvements are essential for facilitating meaningful and lasting behavioral change.

Scenario-based simulations further confirmed that the combination of high AI trust, frequent AI usage, and chronic disease status results in the greatest odds of delayed health care–seeking. This finding underscores the behavioral risks inherent in the trust feedback mechanism. Using behavioral decision theories, particularly the dual-process model, helps illuminate the psychological underpinnings of this phenomenon [42]. In high-trust situations, individuals tend to rely more on the intuitive system for decision-making, leading to quick, automatic judgments rather than thorough risk evaluations [43]. This "intuitive trust" is especially evident among patients with chronic illnesses, who often feel a sense of safety due to the relatively stable nature of their symptoms. When coupled with high trust in and frequent reliance on AI, these individuals are more likely to accept AI-generated suggestions without critically assessing their health status or seeking professional medical advice [44]. This can create a "false reassurance" effect, significantly heightening the risk of delayed health care. These findings serve as a cautionary note for the future design of AI-based health tools. Simply pursuing user trust is insufficient; it is essential to strike a balance between fostering trust and maintaining risk awareness. Both clinical practitioners and AI developers need to consider how to enhance user experience while mitigating the unintended consequences of overreliance [45]. Future AI systems should be designed to promote rational judgment, especially among high-risk groups with chronic conditions. Incorporating interactive features that provide calibrated risk reminders may be necessary to prevent the unconscious shift from trust to delayed health care behavior.

Our ABM offered a dynamic and systems-level understanding of how AI trust and delay behavior coevolve over time. Unlike traditional regression models that provide static snapshots [23], ABM captures how small differences in initial trust and exposure can compound through individual learning and social influence. Over a 14-day simulation period, we observed a steady decline in both trust and health care–seeking behavior, indicating a reciprocal erosion loop: delayed health care leads to dissatisfaction or unmet expectations, which in turn diminishes trust and further health care delay [46]. Importantly, agents stratified by predicted baseline odds (above vs below 20%) showed divergent trust trajectories. High-risk individuals began with higher trust but experienced a sharper decline—suggesting a "high expectation-high disappointment" loop. This highlights how perceived safety in chronic illness, when coupled with excessive AI reliance, may paradoxically lead to trust collapse and amplified delay behavior. Such patterns would be difficult to detect using empirical data alone, underlining the value of simulation modeling in behavioral health research. Beyond reproducing observed behaviors, ABM also allowed us to simulate system-wide interventions and uncover nonlinear responses to trust regulation strategies. Of the 3 mechanisms tested, broadcast messaging consistently reduced delay by maintaining population-level risk awareness. In contrast, network rewiring unexpectedly increased delay, likely due to the formation of echo chambers among high-trust individuals, a phenomenon we term "trust polarization" [47]. This emergent property demonstrates that even well-intentioned peer-based reinforcement may backfire under certain trust dynamics. Overall, these simulations underscore the importance of viewing AI trust not merely as an individual attribute, but as a collective behavioral variable that evolves across time, context, and networks [48]. While the ABM provides valuable insights into potential system dynamics and the comparative theoretical performance of different strategies, these findings serve as a proof of concept. Their real-world effectiveness and causal impact must be rigorously tested in future randomized controlled trials or natural experiments.

Simulation-based evaluations of 3 intervention strategies indicate that broadcast messaging is the most effective for reducing delayed health care behavior. This suggests that public health interventions delivering wide-reaching, consistent risk reminders may be the most cost-effective approach [49]. Conversely, the reward-based strategy was less effective, likely due to insufficient incentives or limited reach, which hindered meaningful behavior change at the population level [50]. Unexpectedly, the network rewiring strategy not only failed to reduce delays but exacerbated them. We hypothesize that this may result from the formation of information echo chambers among high-trust individuals [51]. In these tightly connected groups, AI trust can become mutually reinforced, amplifying the "AI trust-health care delay" pathway and leading to what we term trust polarization. These findings underscore the need to prioritize intervention strategies that broadly disseminate risk information and continuously enhance risk awareness. Future health interventions should focus on scalable and sustainable communication mechanisms rather than relying on individual "opinion leaders" or short-term incentives. This approach provides valuable insights into integrating AI and health behavior, while also offering empirical evidence to inform public health policy development.

Limitations

Despite proposing a novel "trust-behavior" feedback mechanism and validating various intervention strategies through simulation, this study has several limitations. First, the cross-sectional and observational nature of our survey data precludes definitive causal inference. While we controlled for several demographic and health factors, the potential for omitted variable bias remains. Unmeasured confounders, such as general health literacy and prior negative experiences with the health care system, could independently influence both AI trust and the propensity to delay care. To establish causality, future research should use randomized controlled trials that investigate participants' trust in AI. This approach would enable more conclusive mediation and moderation analyses and would represent a critical continuation of our future research efforts. Second, measurements of AI trust and usage frequency were based on self-reported data. This approach may introduce recall bias, social desirability effects, or subjective judgment. Consequently, these factors could potentially affect the accuracy

XSL•FO
RenderX

of the findings. Although we introduced a group-based behavioral simulation model to address the static nature of cross-sectional data, this model simplifies the complex dynamics of trust evolution in real-world contexts. Future research could incorporate reinforcement learning or dynamic trust modeling approaches to better capture human cognitive and behavioral trajectories [52]. Third, the generalizability of our findings may be limited by the specific sociotechnical context of China. Our recruitment relied on dominant Chinese digital platforms (WeChat and QQ). The observed relationships between AI trust, usage, and health care delays are likely influenced by China's unique health care system, technology ecosystem, and cultural norms. Therefore, caution is warranted when extending these results to other countries with different health care policies, technology adoption patterns, and cultural attitudes toward AI in medicine. Fourth, our recruitment via social media and hospital portals, while enabling broad access, prevents precise quantification of response rates from each channel. Although we estimate that most responses came from WeChat and QQ, with a smaller proportion from hospital portals, the anonymous nature of the survey precluded channel-specific stratification. Additionally, this approach likely oversampled individuals who are more digitally literate or proactively engaged with health care information. This may limit the generalizability of our findings to populations with limited digital access or lower health literacy. Fifth, the questionnaire did not collect data on rural or urban residence or socioeconomic status, limiting our ability to examine these potential sociodemographic influences on the trust-delay pathway. Future studies would benefit from incorporating these variables. Sixth, our sample contained a slightly higher proportion of female participants (1345/2460, 54.7%), which may influence generalizability. However, we adjusted for gender in all analyses and found no evidence of significant effect modification. Finally, we advocate for the development of AI health tools with stratified trust management functions [53]. Such systems should provide tailored risk alerts for high-risk patients with chronic conditions, promoting rational trust for genuine health empowerment rather than fostering passive dependence.

Conclusions

This study identifies a nuanced relationship between AI trust and delayed health care–seeking behavior. While trust in AI tools can enhance user engagement, it may also be associated with delayed health care, particularly among individuals with chronic conditions or higher levels of AI use. By integrating survey data with ABM, our study moves beyond static associations and illustrates how AI trust and delay behaviors may interact and reinforce one another over time. This dynamic perspective highlights how individual-level trust processes can accumulate into system-level patterns, including trust polarization and collective delay. Our findings suggest that the design of AI health tools should prioritize calibrated decision support rather than full automation. Encouraging a balanced interaction between user autonomy and technological assistance may help mitigate delay-related risks and improve decision quality in digital health contexts. Beyond individual-level tool design, these insights may also inform population-level strategies for trust governance and risk communication in AI-driven health care systems.

Acknowledgments

All authors thank the Central South University for providing the research platform. We also thank the BioRender platform [54] for providing the tools to create Figure 1 and Figure 3, with all elements authorized for use.

Generative AI was not used in any portion of the manuscript writing.

Funding

This work was supported by the National Natural Science Foundation of China (82501245), Key Research and Development Projects in the Field of Social Development in Hunan Province (2020SK3030), Project of Science and Technology of Hunan Province (2021JJ40932), China Postdoctoral Science Foundation (2025M771943), and Shanghai Post-doctoral Excellence Program (2024409).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: XC
Methodology: XC, YC
Formal analysis: WL
Investigation: XC, WL
Resources: WL
Writing – original draft: XC
Writing – review & editing: WL, WS, YC
Supervision: YC, JZ
Funding acquisition: YC, JZ

XSL•FO
RenderX

Project administration: JZ
Authors YC (1917@sjtu.edu.cn) and JZ (zhoujianda@csu.edu.cn) are co-corresponding authors for this article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
English-translated survey questionnaire.
[DOCX File , 16 KB - jmir_v28i1e82170_app1.docx]

Multimedia Appendix 2
Supplementary Tables S1-S6.
[DOCX File , 29 KB - jmir_v28i1e82170_app2.docx]

Multimedia Appendix 3
Supplementary Figure S1.
[DOCX File , 315 KB - jmir_v28i1e82170_app3.docx]

References

1. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021;8(2):e188-e194 [FREE Full text] [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
2. Gulfidan G, Beklen H, Arga K. Artificial intelligence as accelerator for genomic medicine and planetary health. OMICS 2021;25(12):745-749. [doi: 10.1089/omi.2021.0170] [Medline: 34780300]
3. Topol E. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
4. Grossi E. The long journey of artificial intelligence in medicine: an overview. Clin Exp Rheumatol 2025;43(5):815-821 [FREE Full text] [doi: 10.55563/clinexprheumatol/oamfed] [Medline: 40338059]
5. Quinn T, Senadeera M, Jacobs S, Coghlan S, Le V. Trust and medical AI: the challenges we face and the expertise needed to overcome them. J Am Med Inform Assoc 2021;28(4):890-894 [FREE Full text] [doi: 10.1093/jamia/ocaa268] [Medline: 33340404]
6. Aung Y, Wong D, Ting D. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. Br Med Bull 2021;139(1):4-15. [doi: 10.1093/bmb/ldab016] [Medline: 34405854]
7. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
8. Lynch C, Liston C. New machine-learning technologies for computer-aided diagnosis. Nat Med 2018;24(9):1304-1305. [doi: 10.1038/s41591-018-0178-4] [Medline: 30177823]
9. Ploug T, Holm S. The right to refuse diagnostics and treatment planning by artificial intelligence. Med Health Care Philos 2020;23(1):107-114. [doi: 10.1007/s11019-019-09912-8] [Medline: 31359302]
10. Panch T, Mattie H, Celi L. The "inconvenient truth" about AI in healthcare. NPJ Digit Med 2019;2:77 [FREE Full text] [doi: 10.1038/s41746-019-0155-4] [Medline: 31453372]
11. Dalton-Brown S. The ethics of medical ai and the physician-patient relationship. Camb Q Healthc Ethics 2020;29(1):115-121. [doi: 10.1017/S0963180119000847] [Medline: 31858938]
12. Krist A, Tong S, Aycock R, Longo D. Engaging patients in decision-making and behavior change to promote prevention. Stud Health Technol Inform 2017;240:302. [doi: 10.3233/978-1-61499-790-0-284]
13. Johnson K, Wei W, Weeraratne D, Frisse M, Misulis K, Rhee K, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci 2021;14(1):86-93 [FREE Full text] [doi: 10.1111/cts.12884] [Medline: 32961010]
14. ALHosni F, Al Qadire M, Omari O, Al Raqaishi H, Khalaf A. Symptom prevalence, severity, distress and management among patients with chronic diseases. BMC Nurs 2023;22(1):155 [FREE Full text] [doi: 10.1186/s12912-023-01296-8] [Medline: 37149599]
15. Ajayi K, Wachira E, Onyeaka H, Montour T, Olowolaju S, Garney W. The use of digital health tools for health promotion among women with and without chronic diseases: insights from the 2017-2020 health information national trends survey. JMIR Mhealth Uhealth 2022;10(8):e39520 [FREE Full text] [doi: 10.2196/39520] [Medline: 35984680]
16. Yun H, Bickmore T. Online health information-seeking in the era of large language models: cross-sectional web-based survey study. J Med Internet Res 2025;27:e68560 [FREE Full text] [doi: 10.2196/68560] [Medline: 40163112]
17. Al-Antari M. Artificial intelligence for medical diagnostics-existing and future ai technology!. Diagnostics (Basel) 2023;13(4):688 [FREE Full text] [doi: 10.3390/diagnostics13040688] [Medline: 36832175]
18. Cross J, Choma M, Onofrey J. Bias in medical AI: implications for clinical decision-making. PLOS Digit Health 2024;3(11):e0000651. [doi: 10.1371/journal.pdig.0000651] [Medline: 39509461]

XSL•FO
RenderX

19. Figure 1. Schematic overview of the study design and methodology. Based on online questionnaires, the analysis framework includes mediation analysis to explore the relationship between AI trust, usage frequency, and delayed healthcare seeking behavior. Additionally, agent-based modeling (ABM) were employed to model the dynamic feedback loops over a 14-day period, incorporating three intervention strategies: Broadcast, Reward, and Rewire. These strategies aimed to evaluate how the interventions affected trust and delay behavior within the population.. BioRender. URL: https://app.biorender.com/citation/694800462abb8b925ab7f3cb [accessed 2026-01-18]
20. Levitt H, Bamberg M, Creswell J, Frost D, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA publications and communications board task force report. Am Psychol 2018;73(1):26-46 [FREE Full text] [doi: 10.1037/amp0000151] [Medline: 29345485]
21. Journal Article Reporting Standards (JARS). American Psychological Association. URL: https://apastyle.apa.org/jars [accessed 2025-12-31]
22. Conroy G, Mallapaty S. How China created AI model deepSeek and shocked the world. Nature 2025;638(8050):300-301. [doi: 10.1038/d41586-025-00259-0] [Medline: 39885352]
23. Tracy M, Cerdá M, Keyes K. Agent-based modeling in public health: current applications and future directions. Annu Rev Public Health 2018;39:77-94 [FREE Full text] [doi: 10.1146/annurev-publhealth-040617-014317] [Medline: 29328870]
24. Li Y, Lawley M, Siscovick D, Zhang D, Pagán JA. Agent-based modeling of chronic diseases: a narrative review and future research directions. Prev Chronic Dis 2016;13:E69 [FREE Full text] [doi: 10.5888/pcd13.150561] [Medline: 27236380]
25. Yang Y. A narrative review of the use of agent-based modeling in health behavior and behavior intervention. Transl Behav Med 2019;9(6):1065-1075. [doi: 10.1093/tbm/iby132] [Medline: 30649559]
26. Murase Y, Jo H, Török J, Kertész J, Kaski K. Deep learning exploration of agent-based social network model parameters. Front Big Data 2021;4:739081 [FREE Full text] [doi: 10.3389/fdata.2021.739081] [Medline: 34661097]
27. Blok D, Simoski B, van Woudenberg TJ, Buijzen M. The usefulness of web-based communication data for social network health interventions: agent-based modeling study. JMIR Pediatr Parent 2023;6:e44849 [FREE Full text] [doi: 10.2196/44849] [Medline: 37991813]
28. Marshall B, Galea S. Formalizing the role of agent-based modeling in causal inference and epidemiology. Am J Epidemiol 2015;181(2):92-99 [FREE Full text] [doi: 10.1093/aje/kwu274] [Medline: 25480821]
29. Grimm V, Berger U, DeAngelis D, Polhill J, Giske J, Railsback S. The ODD protocol: a review and first update. Ecol Model 2010;221(23):2760-2768 [FREE Full text] [doi: 10.1016/j.ecolmodel.2010.08.019]
30. Figure 3. Mediation model demonstrating how AI trust influences delayed healthcare seeking through increased frequency of AI usage. Path a shows the significant effect of AI trust on usage frequency, while path b indicates the association between usage frequency and delay. The indirect effect (ab) and direct effect (c′) are also represented, highlighting the statistical significance of the mediation pathway.. BioRender. URL: https://app.biorender.com/citation/69480d355b576217dd8c87e3 [accessed 2026-01-18]
31. Vicente L, Matute H. Humans inherit artificial intelligence biases. Sci Rep 2023;13(1):15737 [FREE Full text] [doi: 10.1038/s41598-023-42384-8] [Medline: 37789032]
32. Klingbeil A, Grützner C, Schreck P. Trust and reliance on AI—an experimental study on the extent and costs of overreliance on AI. Comput Hum Behav 2024;160:108352 [FREE Full text] [doi: 10.1016/j.chb.2024.108352]
33. Stein J, Messingschlager T, Gnambs T, Hutmacher F, Appel M. Attitudes towards AI: measurement and associations with personality. Sci Rep 2024;14(1):2909 [FREE Full text] [doi: 10.1038/s41598-024-53335-2] [Medline: 38316898]
34. Dong C, Ji Y, Fu Z, Qi Y, Yi T, Yang Y, et al. Precision management in chronic disease: an AI empowered perspective on medicine-engineering crossover. iScience 2025;28(3):112044 [FREE Full text] [doi: 10.1016/j.isci.2025.112044] [Medline: 40104052]
35. Khanna N, Maindarkar M, Viswanathan V, Fernandes J, Paul S, Bhagawati M, et al. Economics of artificial intelligence in healthcare: diagnosis vs treatment. Healthcare (Basel) 2022;10(12):2493 [FREE Full text] [doi: 10.3390/healthcare10122493] [Medline: 36554017]
36. Huang S, Lai X, Ke L, Li Y, Wang H, Zhao X, et al. AI technology panic-is AI dependence bad for mental health? A cross-lagged panel model and the mediating roles of motivations for AI use among adolescents. Psychol Res Behav Manag 2024;17:1087-1102 [FREE Full text] [doi: 10.2147/PRBM.S440889] [Medline: 38495087]
37. Rung J, Madden G. Experimental reductions of delay discounting and impulsive choice: a systematic review and meta-analysis. J Exp Psychol Gen 2018;147(9):1349-1381 [FREE Full text] [doi: 10.1037/xge0000462] [Medline: 30148386]
38. Asmundson G, Taylor S. How health anxiety influences responses to viral outbreaks like COVID-19: what all decision-makers, health authorities, and health care professionals need to know. J Anxiety Disord 2020;71:102211 [FREE Full text] [doi: 10.1016/j.janxdis.2020.102211] [Medline: 32179380]
39. Thakkar A, Gupta A, De Sousa A. Artificial intelligence in positive mental health: a narrative review. Front Digit Health 2024;6:1280235 [FREE Full text] [doi: 10.3389/fdgth.2024.1280235] [Medline: 38562663]
40. Li Y, Wu B, Huang Y, Luan S. Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust. Front Psychol 2024;15:1382693 [FREE Full text] [doi: 10.3389/fpsyg.2024.1382693] [Medline: 38694439]

XSL•FO
RenderX

41. Laursen B, Veenstra R. Toward understanding the functions of peer influence: a summary and synthesis of recent empirical research. J Res Adolesc 2021;31(4):889-907 [FREE Full text] [doi: 10.1111/jora.12606] [Medline: 34820944]
42. Barrouillet P. Dual-process theories and cognitive development: advances and challenges. Dev Rev 2011;31(2-3):79-85 [FREE Full text] [doi: 10.1016/j.dr.2011.07.002]
43. Chick C. Cooperative versus competitive influences of emotion and cognition on decision making: a primer for psychiatry research. Psychiatry Res 2019;273:493-500. [doi: 10.1016/j.psychres.2019.01.048] [Medline: 30708200]
44. Shekar S, Pataranutaporn P, Sarabu C, Cecchi GA, Maes P. People overtrust AI-generated medical advice despite low accuracy. NEJM AI 2025;2(6):AIoa2300015 [FREE Full text] [doi: 10.1056/aioa2300015]
45. Siala H, Wang Y. SHIFTing artificial intelligence to be responsible in healthcare: a systematic review. Soc Sci Med 2022;296:114782 [FREE Full text] [doi: 10.1016/j.socscimed.2022.114782] [Medline: 35152047]
46. Nong P, Ji M. Expectations of healthcare AI and the role of trust: understanding patient views on how AI will impact cost, access, and patient-provider relationships. J Am Med Inform Assoc 2025;32(5):795-799. [doi: 10.1093/jamia/ocaf031] [Medline: 40036944]
47. Fränken J, Pilditch T. Cascades across networks are sufficient for the formation of echo chambers: an agent-based model. JASSS 2021;24(3):1 [FREE Full text] [doi: 10.18564/jasss.4566]
48. Nong P, Platt J. Patients' Trust in Health Systems to Use Artificial Intelligence. JAMA Netw Open 2025;8(2):e2460628 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.60628] [Medline: 39951270]
49. Muuraiskangas S, Harjumaa M, Kaipainen K, Ermes M. Process and effects evaluation of a digital mental health intervention targeted at improving occupational well-being: lessons from an intervention study with failed adoption. JMIR Ment Health 2016;3(2):e13 [FREE Full text] [doi: 10.2196/mental.4465] [Medline: 27170553]
50. Michaelsen M, Esch T. Understanding health behavior change by motivation and reward mechanisms: a review of the literature. Front Behav Neurosci 2023;17:1151918 [FREE Full text] [doi: 10.3389/fnbeh.2023.1151918] [Medline: 37405131]
51. Hartmann D, Wang S, Pohlmann L, Berendt B. A systematic review of echo chamber research: comparative analysis of conceptualizations, operationalizations, and varying outcomes. J Comput Soc Sc 2025;8(2):52 [FREE Full text] [doi: 10.1007/s42001-025-00381-z]
52. Das A, Islam M. SecuredTrust: a dynamic trust computation model for secured communication in multiagent systems. IEEE Trans Dependable and Secur Comput 2012;9(2):261-274 [FREE Full text] [doi: 10.1109/tdsc.2011.57]
53. Carroll N, Jones A, Burkard T, Lulias C, Severson K, Posa T. Improving risk stratification using AI and social determinants of health. Am J Manag Care 2022;28(11):582-587 [FREE Full text] [doi: 10.37765/ajmc.2022.89261] [Medline: 36374616]
54. bioRender. URL: https://www.biorender.com/ [accessed 2026-01-08]

Abbreviations

ABM: agent-based modeling
AI: artificial intelligence
MCAR: missing completely at random
ODD: overview, design concepts, and details
OR: odds ratio

Edited by S Brini; submitted 10.Aug.2025; peer-reviewed by MF Wibowo, Z Cui; comments to author 29.Sep.2025; revised version received 12.Dec.2025; accepted 21.Dec.2025; published 03.Feb.2026.

Please cite as:
Cai X, Li W, Shi W, Cai Y, Zhou J
Behavioral Dynamics of AI Trust and Health Care Delays Among Adults: Integrated Cross-Sectional Survey and Agent-Based Modeling Study
J Med Internet Res 2026;28:e82170
URL: https://www.jmir.org/2026/1/e82170
doi:10.2196/82170
PMID:

©Xueyao Cai, Weidong Li, Wenjun Shi, Yuchen Cai, Jianda Zhou. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Screen Time and Chronic Pain Health: Mendelian Randomization Study

Jiahui Jiang[1*], MS; Chunyan Pu[1*], MPH; Jiarui Cai[1*], MA, MPH; Chuan Yu[1], PhD; Zhenmi Liu[1], PhD; Chenghan Xiao[2], PhD

[1]Department of Maternal and Child Health, West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu, Sichuan, China

[2]The Medical Laboratory Center, Hainan General Hospital, Hainan Medical University Hainan Hospital, Haikou, Hainan, China

[*]these authors contributed equally

Corresponding Author:
Zhenmi Liu, PhD
Department of Maternal and Child Health
West China School of Public Health and West China Fourth Hospital
Sichuan University
No.18, Section 3, Renmin South Road, Wuhou District
Chengdu, Sichuan, 610061
China
Phone: 86 18116636482
Email: zhenmiliu@scu.edu.cn

Abstract

Background: The rapid proliferation of electronic devices has increased screen time, raising concerns about its potential health effects, including chronic pain. However, existing studies have limitations in scope and causal inference, with inconsistent findings and a lack of exploration of potential biological mechanisms.

Objective: The objective of our study was to investigate the causal associations and potential shared biological mechanisms between different forms of screen time and various chronic pain phenotypes.

Methods: Leveraging genome-wide association study data, we investigated the association and potential shared biological mechanisms between screen time (time spent watching television, time spent using computer, and length of mobile phone use) and chronic pain phenotypes (including multisite chronic pain [MCP], back, knee, neck or shoulder, hip pain, and headaches). Two-sample Mendelian randomization (MR), reverse MR and multivariable Mendelian randomization (MVMR) analysis were performed to examine associations between screen time and chronic pain. Summary data–based Mendelian randomization (SMR), transcriptome-wide association study (TWAS), and colocalization analysis were used to identify the shared genes and potential biological mechanism.

Results: MR analysis revealed that time spent watching television and length of mobile phone use were positively associated with several types of chronic pain, while time spent using computer showed a negative association. Specifically, time spent watching television was positively associated with the risk of MCP (P=1.05×10^{-31}; odds ratio [OR] 1.61, 95% CI 1.49-1.74), back pain (P=2.41×10^{-8}; OR 1.14, 95% CI 1.09-1.19), knee pain (P=7.10×10^{-6}; OR 1.09, 95% CI 1.05-1.13), neck or shoulder pain, and hip pain. Length of mobile phone use was positively associated with the risk of MCP (P=2.15×10^{-5}; OR 1.22, 95% CI 1.11-1.34), headaches, and neck or shoulder pain. However, time spent using computer was negatively associated with the risk of MCP (P<.001; OR 0.83, 95% CI 0.75-0.92), back pain, and knee pain. The reverse MR results showed that MCP was positively associated with time spent watching television (P=4.8×10^{-7}; OR 1.27, 95% CI 1.16-1.4) and length of mobile phone use (P=3.38×10^{-5}; OR 1.29, 95% CI 1.14-1.45), while the association with time spent using computer (P=.61; OR 0.97, 95% CI 0.87-1.09) was not statistically significant. The MVMR results failed to meet the criterion that all conditional F-statistics exceed 10. Integrative 3 analysis methods identified overlapping genes, with *CEP170* emerging as a key gene consistently supported by SMR, TWAS, and colocalization analysis in the relationship between time spent using computer and MCP.

Conclusions: Our findings demonstrate an association between screen time and various aspects of chronic pain. The *CEP170* gene might contribute to the shared biological mechanism between time spent using computer and MCP risk. However, due to the absence of robust MVMR results, the potential influence of confounding factors cannot be ruled out.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e78233) doi:10.2196/78233

KEYWORDS

causality; chronic pain; medical genetics; multisite chronic pain; screen time

Introduction

With the widespread popularity of electronic devices, the screen time spent by people on phones, computers, and televisions has increased significantly every day. Recently, an increasing number of studies have shown that excessive screen time may bring a series of health risks [1,2]. A study on children and adolescents has shown that long-term screen time exposure is closely related to obesity, restricted cognitive and language development, decreased academic performance, and altered sleep patterns [3,4]. Studies on adults have found that excessive screen time is associated with reduced physical activity and increased BMI [5]. Moreover, studies have found that the incidence of headaches increases with increased screen time, with similar trends observed in both adolescents and adults [6-9]. Research has also shown that prolonged computer use can contribute to pain in several areas, such as the neck, shoulders, and back [10].

While some studies have explored the relationship between screen time and chronic pain, the current evidence is still limited. An important limitation is that existing studies have shortcomings in assessment scope and underlying relationship to fully reveal the association between the two. We found that most studies have focused on the use of a single type of screen time. One study only pointed out that prolonged computer use can cause pain in multiple areas [10], but ignored the effects of different types of screen time. Moreover, most studies on screen time and chronic pain tend to focus on specific body regions, such as lower back pain or headaches [11,12], while overlooking the impact of screen time on multisite chronic pain (MCP) and the combined effects of different types of chronic pain. In addition, the results of existing studies are inconsistent, with some studies finding an association between screen time and chronic pain [13], while others fail to observe a significant association [14]. This inconsistency may be caused by confounding bias or a weaker ability to infer causality.

Another important limitation in the current research concerns the mechanisms associated with screen time and chronic pain. Some studies have suggested that the brightness or light wave frequency of screens may trigger migraine attacks, while prolonged exposure to screens may lower the threshold for migraines [15-17], making them more likely to be triggered by other factors. However, this explanation lacks a detailed understanding of the biological mechanisms involved. To date, the underlying biological mechanisms linking screen time and chronic pain have remained largely unexplored, highlighting the need for future research to address this issue and explore whether different devices cause chronic pain through common or different mechanisms.

The genome-wide association studies (GWASs) of MCP [18], the development of genetic epidemiology, and multiomics integration analysis methods provide new opportunities to address these limitations [19,20]. Two-sample Mendelian randomization (MR) analysis can be used to examine the potential relationship between screen time and chronic pain. MR analysis uses genetic variants significantly associated with exposure as instrumental variables (IVs) to infer causality, thereby minimizing confusion and reverse causality [21]. In addition, summary data–based Mendelian randomization (SMR) and transcriptome-wide association study (TWAS) analysis can help identify overlapping genes between the two. In addition, colocalization analysis can assess whether these associations share causal genetic variation, thereby strengthening the evidence for shared biological mechanisms. By using multiple genetic epidemiological methods, we can explore the biological relationship between screen time and chronic pain.

In this study, we aimed to address these limitations. First, we examined whether screen time was associated with MCP and chronic pain in 5 other body parts. Next, we identified overlapping genes through genetic epidemiological methods approach and explored their potentially shared biological mechanism, aiming to provide new insights into the complexity of chronic pain.

Methods

Study Overview

As shown in Figure 1, this study follows a 3-stage approach. In Phase 1, we performed a 2-sample MR analysis and reverse MR to investigate the phenotypic association between screen time and chronic pain. Considering the mutual influence among multiple exposures, we conducted a multivariate Mendelian randomization (MVMR) analysis. In Phase 2, we first performed SMR analysis combined with expression Quantitative Trait Locus (eQTL) summary datasets to explore associations between gene expression and phenotypic traits. We then performed TWAS analysis to identify gene expression associated with screen time and chronic pain, and assessed whether the same genetic variants were shared through colocalization analysis, providing stronger genetic evidence. In Phase 3, we further integrated the results of SMR, TWAS, and colocalization analysis. We aimed to identify genes that were consistently supported across these methods. The overlapping genes provide more robust evidence of shared biological mechanisms in the association between screen time and chronic pain.

XSL•FO
RenderX

Figure 1. Flowchart of the overall study design. eQTL: expression Quantitative Trait Locus; FDR: false discovery rate; GTEx: Genotype-Tissue Expression; GWAS: genome-wide association study; IV: instrumental variable; IVW: inverse variance weighted; LD: linkage disequilibrium; MCP: multisite chronic pain; MR: Mendelian randomization; MVMR: multivariate Mendelian randomization; OR: odds ratio; SMR: summary data–based Mendelian randomization analysis; SNP: single-nucleotide polymorphism; TWAS: transcriptome-wide association study; TV: television.

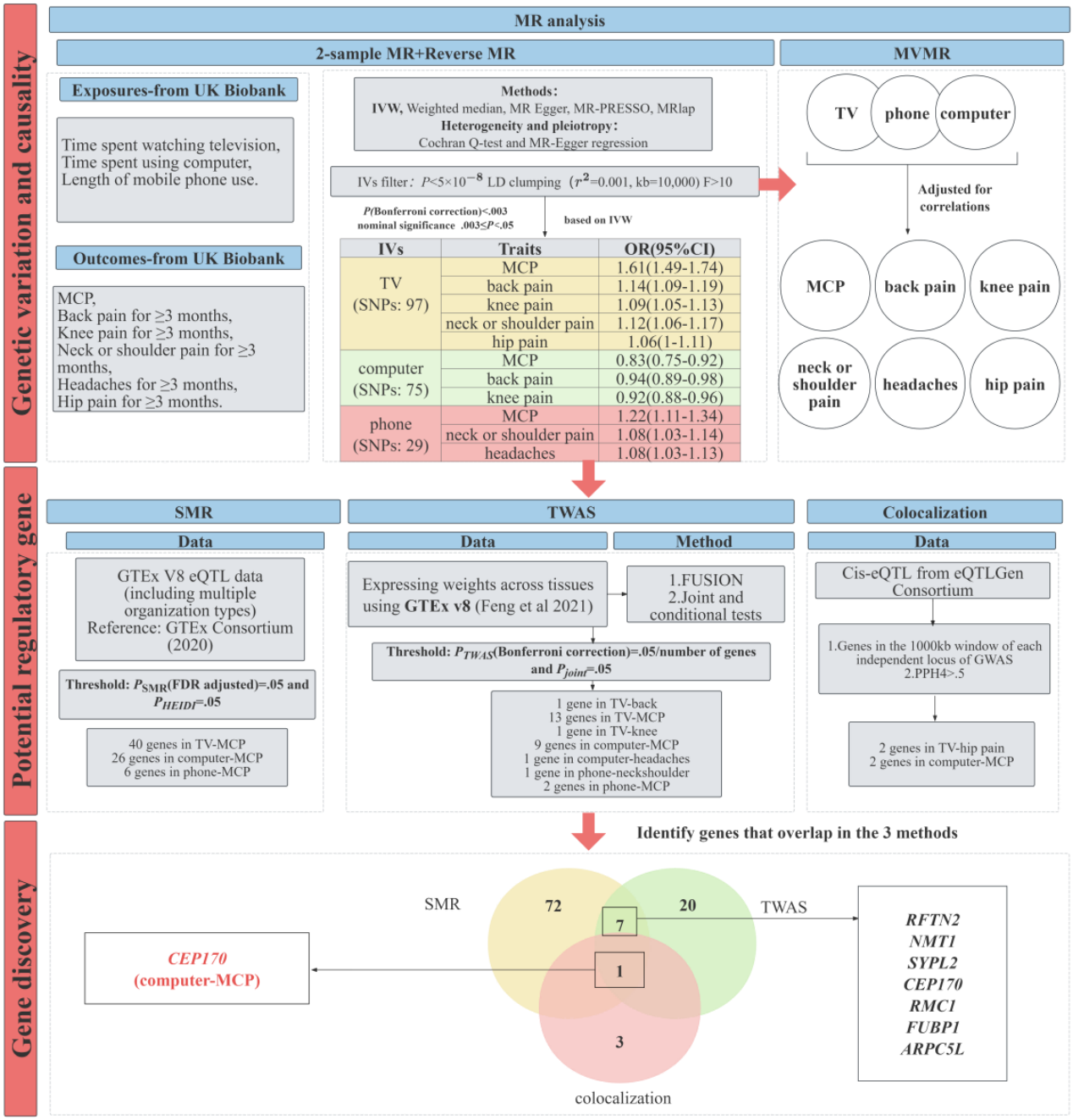

Data Sources for Exposures, Outcomes, and eQTL Summary Data

Exposures

The exposure phenotype in this study was screen time, which included time spent watching television, time spent using computer, and length of mobile phone use. These phenotypes were derived from the UK Biobank touchscreen questionnaire. The question for time spent using computer explicitly stated, "Do not include using a computer at work," whereas the questions for television watching and mobile phone use captured overall daily use. We provided detailed data sources about screen time in Multimedia Appendix 1. To address the issue of weak instrument bias, a genome-wide significance threshold of $P<5\times10^{-8}$ was used as the default criterion for identifying single-nucleotide polymorphisms (SNPs); simultaneously, to optimize the results and eliminate SNPs in strong linkage disequilibrium (LD), an LD clustering method was adopted, with thresholds of $r^2=0.001$ and kilobase, kb=10,000 to ensure accurate and precise clustering. r^2 is the squared correlation coefficient between 2 SNPs' allele counts, used to quantify LD. In the 2-sample MR analysis, we selected 112, 82, and 32 independently significant SNPs located on the autosomes as IVs of time spent watching television, time spent using

XSL•FO
RenderX

computer, and length of mobile phone use, respectively. Details of these SNPs are presented in Multimedia Appendices 2-4.

Outcomes

The outcomes included 6 chronic pain phenotypes, namely MCP, back pain for ≥3 months, knee pain for ≥3 months, neck or shoulder pain for ≥3 months, headaches for ≥3 months, and hip pain for ≥3 months. Among them, MCP is by far the largest GWAS summary statistic for all chronic pain phenotypes, involving 387,649 people from the UK Biobank [18]. Data for the remaining 5 types of chronic pain came from the Integrative Epidemiology Unit (IEU) OpenGWAS, which covered a total of about 9,851,867 SNPs in both male and female individuals [22]. Multimedia Appendix 1 presents the sources of these outcome data. The participants were all European populations to reduce bias caused by population heterogeneity. In reverse MR analysis, Multimedia Appendix 5 shows that 31 independently significant SNPs located on the autosomes as IVs of MCP.

eQTL Data

The Genotype-Tissue Expression (GTEx) project was established to characterize the genetic influence on transcriptomic variation across human tissues and to link regulatory mechanisms to traits and diseases. GTEx v8 provides 15,201 RNA-sequencing samples from 49 tissues [23], collected from 838 postmortem donors. The dataset comprehensively characterizes cis- and trans-eQTLs, revealing regulatory associations for nearly all genes and offering insights into allele heterogeneity, pleiotropy, and tissue-specific genetic effects. For SMR analysis, we used 8 eQTL summary data from GTEx (v8): Adipose Subcutaneous, Adipose Visceral Omentum, Brain Cerebellum, Brain Cortex, Brain Spinal cord cervical c-1, Muscle Skeletal, Nerve Tibial, and Whole Blood [23]. This is a set of cis-eQTL summary data for 8 human tissues from GTEx v8.

For colocalization analysis, we additionally used eQTL data from the eQTLGen Consortium [24]. eQTLGen provides a larger scale of eQTL data than GTEx, which can improve the confidence of colocalization analysis.

Statistical Analysis

2-Sample MR Analysis

2-sample MR analysis was conducted for exposure and outcome. The strength of IVs was assessed using the F-statistics: $F = R^2 \times (N-2)/(1-R^2)$, where R^2 represents the proportion of variation in the exposure variable explained by IVs [21]. The calculation of R^2 involved multiplying beta (the estimated genetic association of each SNP with the trait) by minor allele frequency (MAF), using the formula $2 \times MAF\ (1-MAF)\ \beta^2$. An F-statistic >10 indicates a robust instrument [21]. IVs with F-statistics below 10 were considered weak instruments and excluded from analysis. We applied inverse variance weighted (IVW) regression as the primary method to estimate the effects of screen time on chronic pain [25]. In order to assess the robustness of the results, we also performed different sensitivity analyses. First, we complement the IVW MR with the weighted median, MR-Egger methods. The weighted median method can robustly estimate the causal relationship even when less than 50% of the genetic variants are invalid IVs [26]. The MR-Egger method includes an intercept term to account for directional pleiotropy [27]. Pleiotropy means that SNPs not only affect outcomes through exposures but also directly influence outcomes through other independent pathways. However, the slope of the MR-Egger regression provides valid MR estimates in the presence of horizontal pleiotropy [27,28]. It should be noted that when the number of SNPs is small, the statistical ability of MR-Egger is limited, making its results less reliable [27]. Second, we used Mendelian Randomization Pleiotropy Residual Sum and Outlier (MR-PRESSO) to detect the presence of outliers and to reassess the effect after removing the detected outlier SNPs [29]. Finally, we performed a leave-one-out analysis to reevaluate the IVW effect by excluding SNPs one by one [30]. To investigate potential bias due to sample overlap between screen time and chronic pain, we further conducted the MRlap method [31]. We examined the $P_{difference}$ of MRlap results, which is the P value used to test for differences between the observed (uncorrected) and corrected effects. We used the Bonferroni correction to account for multiple tests in the IVW result. An association was considered significant if the P value in the primary analysis was below 2.778×10^{-3} (.05/18) and the direction of effect estimates remained consistent across all methods. Correspondingly, suggestive evidence was considered if the P value for the IVW result was between .003 and .05.

To ensure methodological rigor and robustness, we also conducted a reverse MR, which is similar to the approach of 2-sample MR.

MVMR Analysis

To clarify the independent effects of each type of screen time on chronic pain, we performed MVMR analysis. The MVMR takes into account the combined effects of multiple related exposures, thereby adjusting for potential confounders and providing a more precise estimate of the independent contribution of each exposure. In the MVMR analysis, 3 phenotypes related to screen time exposures (time spent watching television, time spent using computer, and length of mobile phone use) were used as common exposure variables, and SNPs with $P<5\times10^{-8}$ genomic significance were used as IVs. At the same time, SNPs in strong LD with exposure variables were removed through LD clustering (r^2=0.001, kb=10,000). Next, we merged the SNPs of 3 exposures and deleted the duplicate SNPs. To ensure validity, only SNPs present in the result data were retained and any SNPs that did not match were excluded. The IVW method was used for main effect estimation. We evaluated the strength of the IVs using the conditional F-statistics, which measure the strength of each exposure conditional on the others in the model. If the conditional F-statistics for all exposures exceeded the rule-of-thumb value of 10, the IVs were considered adequately strong for the purposes of MVMR [32].

SMR Analysis

We performed SMR analysis to explore whether gene expression is causally linked to both screen time and chronic pain [20], thereby identifying potential functional genes involved in the

XSL•FO
RenderX

relationship. SMR integrates GWAS and eQTL data, enabling the investigation of associations between gene expression and phenotypic traits. In the SMR analysis, exposures and outcomes were analyzed separately with each eQTL summary data to assess whether gene expression is linked to both screen time and chronic pain. During the analysis, SNPs significantly associated with gene expression in cis-eQTL summary data were used as IVs, as they serve as genetic proxies for gene expression levels. These SNPs were selected based on their genome-wide significant association with gene expression in eQTL summary data. False discovery rate (FDR) correction was applied to control the FDR of multiple tests. Additionally, to verify whether the observed associations could be caused by a single causal variable, we evaluated the association results using the heterogeneity in dependent instruments (HEIDI) test implemented in the SMR tool, retaining only probes with P_{HEIDI} values indicating low heterogeneity [33]. Ultimately, the SMR analysis results for screen time and chronic pain were integrated to identify overlapping genes and further explore their potential roles in the association between screen time and chronic pain.

TWAS Analysis

We conducted TWAS to identify genes whose expression may mediate the relationship between screen time and chronic pain. GWAS datasets were obtained from IEU OpenGWAS, and using the FUSION tool developed by Gusev et al [34], we performed single-trait TWAS using the 3 cross-tissue weights for cross-tissue features generated through sparse canonical correlation analysis (sCCA) on GTEx v8 gene expression (including sCCA1, sCCA2, and sCCA3) to identify regulatory genes that may be involved in chronic pain pathways [19]. To control for type I error rates, Bonferroni correction was applied to each exposure and result to interpret multiple tests, setting the significance level for each trait or tissue to P=.05/number of genes (adjusted for the number of genes on each chromosome in the 3 GTEx v8 weights). We then extracted significant GWAS of exposure and outcome, performed joint and conditional tests on loci with multiple related features to assess whether loci contain signals independent of expression. Finally, we sought overlapping genes between exposures and outcomes in joint tests.

Colocalization Analysis

To further verify the reliability of the SMR analysis results and determine whether screen time and pain are driven by the same genetic variation, we used colocalization analysis. First, we determined all independent signals of GWAS across the genome through LD independence analysis. Genes within a 1000-kb window of each independent locus were subjected to colocalization analysis [35]. Within these windows, we combined the corresponding genes with eQTL data, which is cis-eQTL data in SMR format, and then conducted colocalization analysis. We performed colocalization analysis using the default priors of P_1=1×10^{-4}, P_2=1×10^{-4}, and P_{12}=1×10^{-5} [36]. Colocalization analysis assesses whether 2 traits may be driven by the same causal variant by estimating the joint posterior probability (PP) of GWAS and eQTL signals at the same locus. This method assumes a maximum of one causal variant per trait in a gene region and uses approximate Bayesian factor calculations to derive the PP for 4 mutually exclusive hypotheses (H_0-H_4), representing all possible association configurations between 2 traits: (1) H_0: neither trait has a genetic association in the region; (2) H_1 or H_2: only trait 1 or trait 2 has a genetic association in the region; (3) H_3: both traits are associated, but with different causal variants; and (4) H_4: both traits are associated and share a single causal variant. The PP of each configuration is denoted as PPH_0, PPH_1, PPH_2, PPH_3, and PPH_4, respectively [37]. We used PPH_4 to characterize the possibility of colocalization. The probability value of .5< PPH_b <.8 suggests a moderate support for colocalization, whereas PPH_4≥.8 indicates a strong support for colocalization, indicating that the 2 signals share a causal variant at this locus [38]. Using eQTL data, colocalization analysis was conducted separately for exposure and outcome traits, and genes showing evidence of colocalization were subsequently overlapped at the gene level.

Finally, to ensure the robustness of the analysis results, we further screened and consolidated the results of 3 analysis methods. We compared the colocalized genes with the overlapping genes identified in SMR and TWAS analysis to enhance the understanding of the potential biological mechanisms of these genes in the relationship between screen time and chronic pain.

Ethical Considerations

All data used in this study were deidentified publicly available data; therefore, no ethical approval was required for this study. All original studies received ethical approval from their respective institutional review boards, and all participants provided informed consent. The data used were anonymized to ensure privacy and confidentiality. No compensation was provided to participants. Additionally, this study does not include any identifiable images or figures.

Results

The Putative Association Between Screen Time and Chronic Pain

The F-statistics of IVs are all greater than 10 in Multimedia Appendices 2-4, indicating that the IVs are relatively strong. We used a 2-sample MR method to make causal inference, primarily relying on the IVW method. Figure 2 shows that the 2-sample MR results of screen time and chronic pain based on the IVW. The box in Figure 2 indicates the point estimate of the causal effect, and the error bars represent the 95% CI. Multimedia Appendix 6 showed that time spent watching television and length of mobile phone use were positively associated with chronic pain, while time spent using computer was negatively associated. Our IVW results suggested significant positive associations of time spent watching television with MCP (P=1.05×10^{-31}; odds ratio [OR] 1.61, 95% CI 1.49-1.74), back pain (P=2.41×10^{-8}; OR 1.14, 95% CI 1.09-1.19), knee pain (P=7.10×10^{-6}; OR 1.09, 95% CI 1.05-1.13), and neck or shoulder pain (P=1.18×10^{-5}; OR 1.12, 95% CI 1.06-1.17). The association between time spent watching

XSL•FO
RenderX

television and hip pain (P=.03; OR 1.06, 95% CI 1-1.11) reached nominal significance. It is notable that even though there is horizontal pleiotropy between time spent watching television and back pain ($P_{MR\text{-}Egger\ intercept}$=.02), it is significant after MR-Egger correction ($P_{MR\text{-}Egger\ slope}$=4.20×10^{-4}). Similarly, our IVW results indicated significant positive associations of length of mobile phone use with MCP (P=2.15×10^{-5}; OR 1.22, 95% CI 1.11-1.34), headaches (P=.003; OR 1.08, 95% CI 1.03-1.13), and neck or shoulder pain (P<.001; OR 1.08, 95% CI 1.03-1.14). However, IVW results suggested significant negative associations of time spent using computer with MCP (P<.001; OR 0.83, 95% CI 0.75-0.92), and knee pain (P=3.20×10^{-5}; OR 0.92, 95% CI 0.88-0.96). The association between time spent using computer and back pain (P=.01; OR 0.94, 95% CI 0.89-0.98) reached nominal significance. For the significant associations, concordant estimates were basically suggested by weighted median, MR-Egger, MR-PRESSO, and MRlap.

Figure 2. Mendelian randomization analysis of screen time and chronic pain based on the inverse variance weighted. MCP: multisite chronic pain; OR: odds ratio; TV: television.

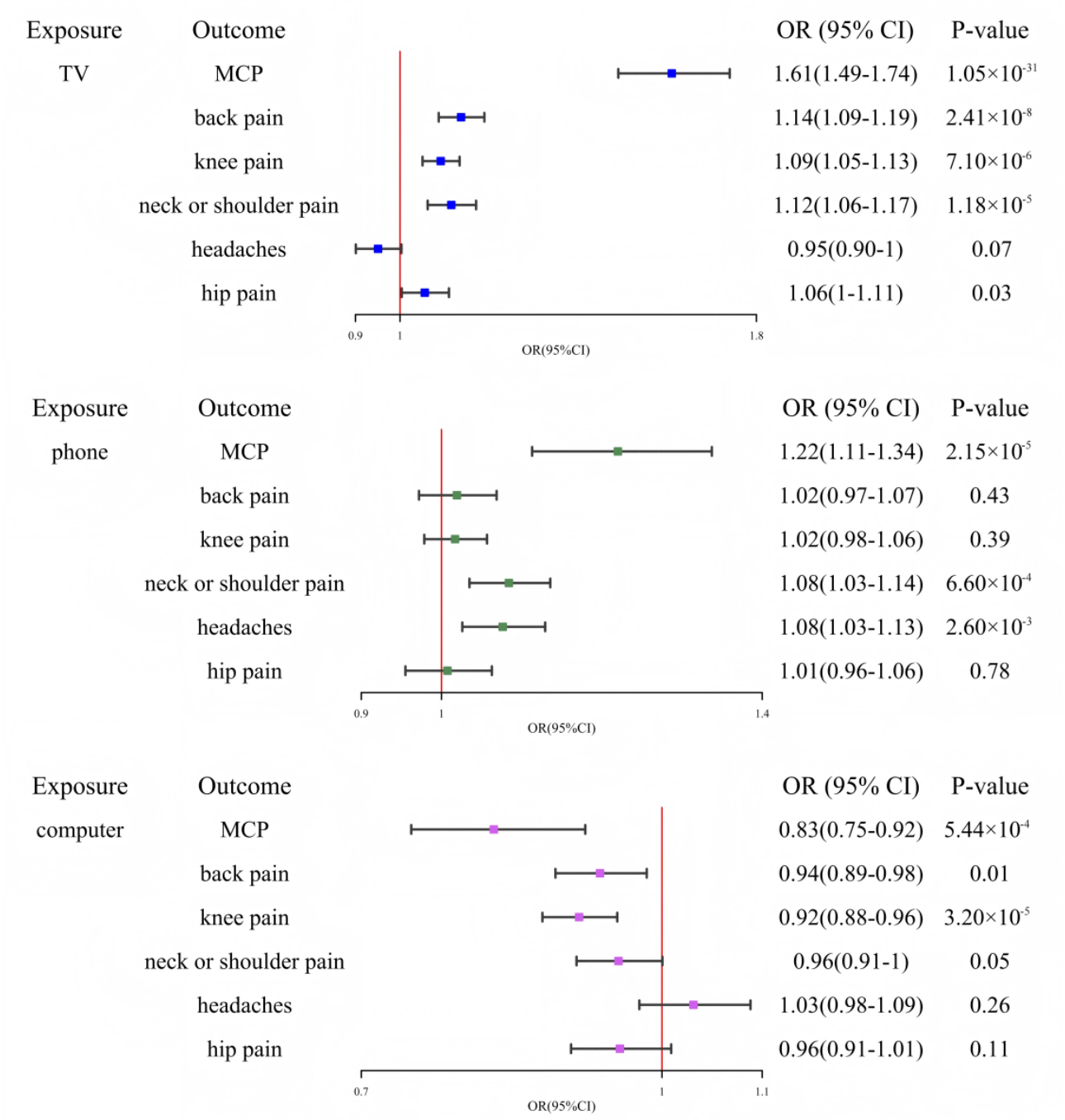

MRlap-corrected results in Multimedia Appendix 6 showed that time spent watching television significantly increased the risk of MCP ($P_{corrected}$=3.57×10^{-30}; $OR_{corrected}$=1.53, 95% CI 1.43-1.65), back pain ($P_{corrected}$=8.46×10^{-6}; $OR_{corrected}$=1.27, 95% CI 1.14-1.41), knee pain ($P_{corrected}$=1.14×10^{-4}; $OR_{corrected}$=1.22, 95% CI 1.11-1.36), and neck or shoulder pain

XSL•FO
RenderX

($P_{corrected}$=6.99×10^{-4}; $OR_{corrected}$=1.19, 95% CI 1.08-1.31). Length of mobile phone use significantly increased the risk of MCP ($P_{corrected}$=1.04×10^{-4}; $OR_{corrected}$=1.39, 95% CI 1.18-1.64), headaches ($P_{corrected}$=.03; $OR_{corrected}$=1.34, 95% CI 1.04-1.74), and neck or shoulder pain ($P_{corrected}$=.01; $OR_{corrected}$=1.55, 95% CI 1.1-2.17). In contrast, time spent using computer was negatively associated with the risk of MCP ($P_{corrected}$=1.65×10^{-3}; $OR_{corrected}$=0.81, 95% CI 0.71-0.92), back pain ($P_{corrected}$=4.28×10^{-3}; $OR_{corrected}$=0.83, 95% CI 0.73-0.94), and knee pain ($P_{corrected}$=3.48×10^{-3}; $OR_{corrected}$=0.83, 95% CI 0.73-0.94). Overall, the MRlap-corrected causal estimates were consistent in direction and statistical significance with the primary IVW results in Multimedia Appendix 6, supporting the robustness of the findings against biases induced by sample overlap. Leave-one-out analyses in Multimedia Appendix 7 showed no outlying SNPs.

In the reverse MR analysis, MCP was used as the exposure, and different types of screen time were treated as outcomes. The F-statistics of IVs of MCP in Multimedia Appendix 5 are all greater than 10, indicating that the IVs are relatively strong. The results in Multimedia Appendix 8 indicated that MCP was positively associated with time spent watching television (P=4.8×10^{-7}; OR 1.27, 95% CI 1.16-1.4) and length of mobile phone use (P=3.38×10^{-5}; OR 1.29, 95% CI 1.14-1.45), while the association with time spent using computer (P=.61; OR 0.97, 95% CI 0.87-1.09) was not statistically significant. For other exposures, the analysis could not be conducted due to the limited number of available IVs or F-statistics below the conventional threshold of 10.

Independent Effect of Screen Time on Chronic Pain

Multimedia Appendix 9 showed that the causal effects of screen time on chronic pain based on IVW MVMR. The conditional F-statistics for both time spent using computer and length of mobile phone use is less than 10 in the MVMR results.

Discovery of Screen Time and Chronic Pain Genes Based on SMR Analysis

As the main analysis result, SMR was corrected by FDR and combined with the HEIDI test to screen for overlapping genes between exposure and outcome in 8 eQTL summary data. Multimedia Appendices 10 and 11 provided the full results, including the exact P_{FDR} and P_{HEIDI} values for all tested genes after FDR correction. A total of 72 overlapping genes related to exposure and outcome were identified in Multimedia Appendix 12, and the results showed that these overlapping genes originated from 3 exposures and MCP: 40 genes for time spent watching television and MCP, 26 genes for time spent using computer and MCP, and 6 genes for length of mobile phone use and MCP. In the SMR analysis, the Adipose Subcutaneous tissue had the most genes, while only one gene was found in the Brain Spinal cord cervical c-1.

Genetic Findings of TWAS-Based Screen Time and Chronic Pain

After Bonferroni correction, we examined genome-wide significant associations in Multimedia Appendices 13 and 14, followed by joint and conditional tests. Multimedia Appendices 15 and 16 provided the results of the jointly significant genes. Then we identified 28 genes with overlapping exposures and outcomes in Table 1. No genes overlapping with the exposure were found in the hip pain for ≥3 months GWAS dataset.

XSL•FO
RenderX

Table 1. Transcriptome-wide significant genes identified by TWAS[a] in screen time and chronic pain.

Trait	Cross-tissue expression	Gene (chromosome)
Time spent watching television – back pain for ≥3 months	sCCA1	• *PMS2P3* (7)
Time spent watching television – MCP[b]	sCCA2	• *FUBP1* (1) • *WDR47* (1) • *PLEKHO1* (1) • *FAM172A* (5)
Time spent watching television – MCP	sCCA3	• *SUSD3* (9) • *RPL35* (9) • *RELA* (11) • *RMC1* (18) • *SYPL2* (1) • *SF3B4* (1) • *ARPC5L* (9) • *PTPDC1* (9) • *FAM53B* (10)
Time spent watching television – knee pain for ≥3 months	sCCA2	• *GATC* (12)
Length of mobile phone use – neck or shoulder pain for ≥3 months	sCCA3	• *RBM42* (19)
Length of mobile phone use – MCP	sCCA3	• *CSTPP1* (11) • *TM9SF4* (20)
Time spent using computer – MCP	sCCA2	• *WDR47* (1) • *RFTN2* (2) • *NMT1* (17) • *RMC1* (18) • *FASTKD5* (20) • *TM9SF4* (20)
Time spent using computer – MCP	sCCA3	• *SYPL2* (1) • *CEP170* (1) • *SCOC-AS1* (4)
Time spent using computer – headaches for ≥3 months	sCCA2	• *SHMT2* (12)

[a]TWAS: transcriptome-wide association study.

[b]MCP: multisite chronic pain.

By combining the results of SMR and TWAS analysis, we found that 7 overlapping genes (*SYPL2*, *RMC1*, *FUBP1*, *ARPC5L*, *RFTN2*, *NMT1*, and *CEP170*) were identified between exposures (time spent watching television and time spent using computer) and outcome (MCP) in Multimedia Appendix 17 and Figure 3. In addition, we found that *SYPL2* gene expression was associated with both time spent watching television and time spent using computer.

XSL•FO
RenderX

Figure 3. Results supported by genetic evidence. SMR: summary data–based Mendelian randomization analysis; TWAS: transcriptome-wide association study.

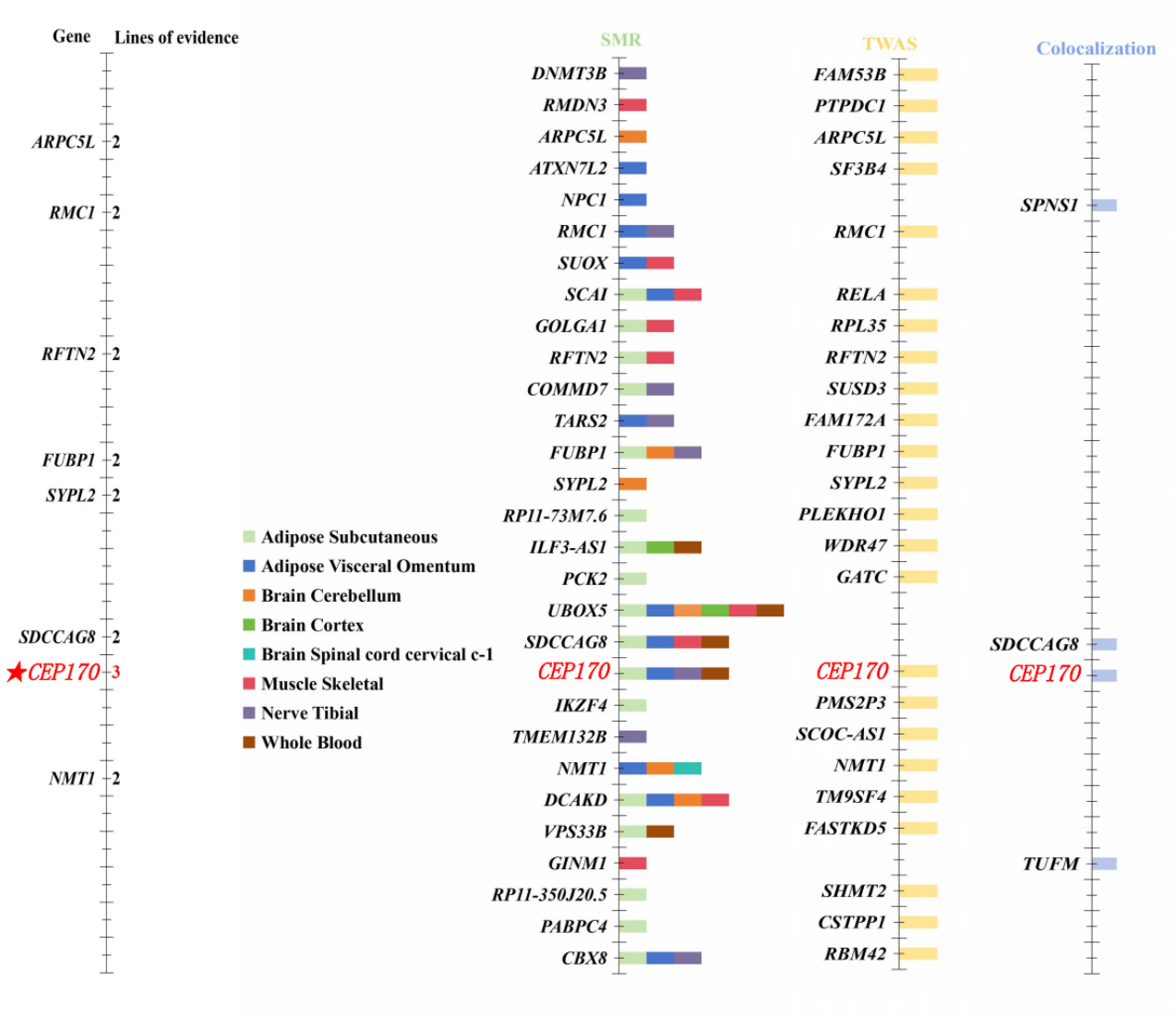

In Figure 3, the panel on the far left shows genes supported by at least 2 lines of evidence. A total of 3 lines of evidence were used to support genes, with each column representing one type of supporting evidence. The SMR evidence shows different colors according to 8 different eQTL summary data. TWAS evidence is shown in light yellow and colocalization evidence is shown in light blue. *CEP170* was supported by all analyses. *ARPC5L*, *RMC1*, *RFTN2*, *FUBP1*, *SYPL2*, *SDCCAG8*, and *NMT1* were supported by 2 lines of evidence.

Colocalization Analysis of Shared Genetic Variation

Colocalization analysis identified multiple genes with moderate or strong support for colocalization signals in the shared genetic variation between screen time and chronic pain in Multimedia Appendix 18. The results showed that 4 genes suggest a moderate support for colocalization (*SPNS1*, *TUFM*, *SDCCAG8*, and *CEP170*). Multimedia Appendix 19 showed that the 4 genes are overlapped between screen time and chronic pain.

Gene Discovery

As shown in Figure 3, we presented the distribution of significant genes in each analysis method, and the results showed that the *CEP170* gene stood out in the association analysis of "Time spent using computer" and MCP, receiving triple support from SMR, TWAS, and colocalization analysis (PPH_4>.75). The signaling distribution of this gene involved 4 eQTL tissues: Adipose Subcutaneous, Adipose Visceral Omentum, Nerve Tibial, and Whole Blood.

Discussion

Principal Findings

This study delves into potential phenotypic associations and underlying genetic connections between screen time and chronic pain. The results show that screen time is associated with chronic pain. In the 2-sample MR results, time spent watching television and length of mobile phone use were significantly and positively associated with chronic pain, including MCP, back pain, knee pain, neck or shoulder pain, hip pain, and headaches. In contrast, time spent using computer inversely was negatively associated with MCP, back pain, and knee pain. It is particularly important to note that we still list the significant IVW results of exposure to the length of mobile phone use. Even though the effect value directions of the IVW and MR-Egger results are not consistent,

XSL•FO
RenderX

it is consistent with the direction of our main analytical method in other methods. The statistical power of MR-Egger decreases for fewer IVs [27]. Therefore, we still take the results of IVW as the standard for the length of mobile phone use. In the reverse MR results, we found that there is a bidirectional positive correlation among time spent watching television, length of mobile phone use, and MCP. In the MVMR analysis, not all the conditional F-statistics were >10, suggesting the presence of weak instruments. Therefore, our interpretations were based on the 2-sample MR results. However, the inverse association between time spent using computer and MCP should be interpreted cautiously, as it stems from the 2-sample MR results. The lack of robust MVMR findings means we cannot rule out the potential influence of confounding factors. In addition, by integrating SMR, TWAS, and colocalization analysis, we identified an overlapping gene *CEP170* associated with time spent using computer and MCP. These findings highlight the important role of screen time in the development of chronic pain.

Differential Associations Between Screen Time and Chronic Pain

Previous studies have shown a positive correlation between mobile phone use and neck pain [39], and a strong correlation between watching television and back pain [40], which is consistent with our research findings. In contrast, our research found that time spent using computer may have a protective effect on certain types of chronic pain, which differs from some studies that concluded, "longer computer time increases the risk of multi-site pain" [8,39]. This difference may be due to the study population, with earlier studies focusing on adolescents, while our analysis was conducted on adults. Due to continued musculoskeletal development, adolescents may be more susceptible to the negative physical effects of sedentary behavior.

Furthermore, 2 potential explanations may underlie the differential associations between time spent using computer and chronic pain, although there is no direct evidence available at present. One possible explanation is that the perception of pain is influenced by the allocation of cognitive resources; that is, when attention is occupied by other highly attractive stimuli, the perception of pain may be reduced [41,42]. According to the characteristics of the UK Biobank population (aged 40-69 years during 2006-2010) in this study [43], the use of mobile phones and televisions was often associated with activities that had relatively low cognitive requirements and less attention input at that time. In contrast, even leisure-time computer use generally involved more cognitively demanding activities among middle-aged adults, such as reading, writing, online communication, or other interactive tasks. Research indicates that pain is less likely to enter conscious awareness when cognitive resources are occupied by goal-directed information maintained in working memory and when sustained attention is devoted to the task [44]. This indicates that computer tasks with high cognitive load may potentially reduce the pain experience through distraction. Moreover, it is also important to consider that individuals who frequently use computers tend to have higher educational attainment, higher income, and better overall health [45]. Given that individuals in nonmanual occupations tend to have a lower likelihood of experiencing chronic pain compared with those in manual labor occupations [46,47], part of the observed lower risk associated with time spent using computer may therefore reflect underlying socioeconomic or occupational factors rather than a sole reflection of the nature of computer-based activities themselves. This may explain why length of mobile phone use and time spent watching television were associated with a higher risk of pain, while time spent using computer showed a protective effect in this study.

Another possible explanation is that different types of screen time may affect pain perception by engaging the brain's reward system to varying degrees. Dopamine is involved in pleasure, reward, and incentive behavior [48]. Although mobile phone and television use may induce pleasure, the limited interactivity prevents them from triggering dopamine release to the same extent as engaging in more goal-directed and cognitively engaging computer activities. Successfully completing such goals may elicit a sense of accomplishment. The reward system may be activated, promoting dopamine release, which has been proven to have analgesic effects [49,50]. Moreover, bidirectional association analysis suggests that MCP may influence the choice of screen behavior. People with MCP tend to increase passive and low-effort screen behaviors (such as watching television or casually browsing their mobile phones). This behavioral adaptation may contribute to a pattern of bidirectional reinforcement, where an increase in passive activity further restricts physical movement and may exacerbate pain. However, we emphasize that these explanations are speculative, as direct evidence is currently lacking.

The Potential Mechanisms by Which CEP170 Contributes to Screen Time and Chronic Pain

Leveraging multiomics approaches, we identified *CEP170*, whose cis-regulated expression may contribute to the biological mechanisms between time spent using computer and MCP. In SMR results, the *CEP170* gene expression level was positively correlated with MCP but negatively correlated with time spent using computer. This aligns with our MR results, where time spent using computer was associated with a lower risk of MCP. One possible explanation is that *CEP170* may influence behavioral patterns related to screen time, where individuals with higher gene expression tend to spend less time on the computer, which in turn is associated with increased MCP risk. Although its colocalization results suggest a moderate support for colocalization, *CEP170* was detected in all analytical methods.

CEP170 is located at the centrosome and spindle microtubules and participates in microtubule organization and assembly [51,52]. Additionally, *CEP170* plays an important role in supporting ciliary homeostasis [53]. Cilia may potentially play a central role in cell signaling, and recent animal experiments have further revealed that cilia have key regulatory functions in controlling mechanical nociceptive thresholds and inflammatory and neuropathic pain [54]. The SMR results show that *CEP170* exhibited significant signals in multiple eQTL tissues, including Adipose Subcutaneous, Adipose Visceral Omentum, Nerve Tibial, and Whole Blood. This indicates that

XSL•FO
RenderX

CEP170 may be expressed in these tissues and may contribute to the biological mechanisms behind the observed associations. A large UK Biobank study revealed that abdominal adipose tissue (including visceral fat and subcutaneous fat) was associated with chronic musculoskeletal pain. It suggested that excessive and ectopic fat depositions may be involved in the pathogenesis of multisite and widespread chronic musculoskeletal pain [55]. This is likely attributable to the chronic low-grade inflammation driven by accumulated visceral fat, which can sensitize peripheral nerves and promote MCP [56,57]. In line with this, our SMR analysis found that *CEP170* expression in subcutaneous and visceral adipose tissue is associated with MCP. Given the established importance of microtubules in lipid metabolic homeostasis [58,59], we hypothesize that *CEP170* may potentially protect against MCP by enhancing microtubule stability in adipose tissue. This mechanism might potentially constrain adipose-derived inflammatory signaling, thereby possibly alleviating peripheral nerve sensitization and lowering MCP risk. However, the current results only provide preliminary genetic clues. Direct links remain unclear and more functional studies, especially tissue-specific experiments, are needed in the future to clarify the potential contribution of *CEP170* to pain pathways.

Strengths and Limitations

Our research has several advantages. First, we integrated a variety of analytical methods to duplicate our findings and strengthen the robustness of our results. Second, we leveraged large-scale GWAS data from the UK Biobank, ensuring a robust analysis with broad generalizability. Third, by identifying overlapping genes and potential biological pathways, our study provides new genetic insights into the association between screen time and chronic pain, which, despite limited evidence, may inform future research in this field. Finally, our study considered different types of screen time separately, which allows for a more nuanced understanding of their different effects on chronic pain. These findings may serve as a useful reference for subsequent studies on the genetic and epidemiological links between screen time and chronic pain.

Although this study provides new insights, there are still some limitations. First, using data from the same sources in multiple analyses may introduce the bias of the winner's curse, potentially inflating effect sizes or significance levels. Replication in fully independent datasets is necessary to confirm the discovery. Second, our analyses were restricted to European populations, which may limit generalizability to other populations with different screen use habits and pain reporting. Third, the lack of detailed data on screen use posture, activity type, and behavior patterns may limit the depth of causal reasoning. Future studies could incorporate longitudinal and behavioral data to better clarify the mechanisms and causal pathways. Fourth, as both exposures and outcomes are self-reported, recall and measurement bias may lead to misclassification and reduce precision. Self-reported pain may not capture clinical heterogeneity, so future studies using clinically validated phenotypes are needed. Fifth, due to the weak strength of IVs, we were unable to adequately control for crucial potential confounding factors, such as socioeconomic status and educational attainment. Consequently, the observed protective association between "time spent using computer" and chronic pain might be partially attributable to residual confounding by socioeconomic status and related factors. Sixth, reverse MR could not be performed for the outcomes of back pain, knee pain, neck or shoulder pain, headaches, and hip pain due to limited IVs, limiting the assessment of reverse causality. In addition, the smaller sample size for hip pain may have limited statistical power, contributing to borderline significance in the 2-sample MR analysis and precluding reverse MR for this phenotype. Seventh, environmental exposure, such as air pollution, was not taken into account in this analysis. Particulate matter can affect inflammatory and stress-related biomarkers involved in chronic pain [60-62]. Future studies should integrate genetic and environmental data to clarify these relationships. Eighth, the colocalization signal for *CEP170* is not the strongest, and direct evidence is lacking. Therefore, its role remains speculative and warrants further investigation. Finally, this study lacks validation in large surveys or independent cohorts, highlighting the need to combine genetic and epidemiological studies in future research.

Conclusions

This study provides evidence that reveals an association between screen time and chronic pain. We found the *CEP170* gene might contribute to the shared biological mechanism between time spent using computer and MCP risk. Future studies should further validate this association and clarify the functional role of *CEP170* in the development of chronic pain.

Acknowledgments

The authors would like to thank the participants and investigators involved in the GWASs used in this study.

Funding

This work was supported by the Hainan Provincial People's Hospital Scientific Research Start-up Fund (grant number YJRC2025009).

Data Availability

The code used in this study is available from the corresponding authors on reasonable request. The GWAS summary statistics data used in this study are available from IEU OpenGwas and publicly available GWAS. The eQTL summary data for GTEx used in the SMR analysis are available in the public database [63]. Precomputed sCCA weights for GTEx v8 used in this study

XSL•FO
RenderX

are available in the public database [64]. The eQTL data used in the colocalization analysis are available from the eQTLGen Consortium [65].

Authors' Contributions

CX and ZL contributed equally as corresponding authors. JJ, CP, and JC contributed equally to this work and share first authorship. JJ and CX conceived the study. JJ and CP performed the data analysis and interpretation. JJ and CP wrote the manuscript. JC, CP, and JJ critically revised the manuscript. CY, CP, JC, and CX provide additional methodological guidance and technical support. CX and ZL supervised the study and were responsible for overall project administration. All authors reviewed and approved the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data source of the exposures and outcomes.
[XLSX File (Microsoft Excel File), 12 KB - jmir_v28i1e78233_app1.xlsx]

Multimedia Appendix 2
Characteristics of genetic instruments of Time spent watching television and their effect sizes with chronic pain.
[XLSX File (Microsoft Excel File), 45 KB - jmir_v28i1e78233_app2.xlsx]

Multimedia Appendix 3
Characteristics of genetic instruments of Time spent using computer and their effect sizes with chronic pain.
[XLSX File (Microsoft Excel File), 36 KB - jmir_v28i1e78233_app3.xlsx]

Multimedia Appendix 4
Characteristics of genetic instruments of Length of mobile phone use and their effect sizes with chronic pain.
[XLSX File (Microsoft Excel File), 21 KB - jmir_v28i1e78233_app4.xlsx]

Multimedia Appendix 5
Characteristics of genetic instruments of multisite chronic pain.
[XLSX File (Microsoft Excel File), 15 KB - jmir_v28i1e78233_app5.xlsx]

Multimedia Appendix 6
Putative causal effect of screen time on chronic pain from 2-sample Mendelian randomization analysis.
[XLSX File (Microsoft Excel File), 24 KB - jmir_v28i1e78233_app6.xlsx]

Multimedia Appendix 7
The leave-one-out plot of screen time on chronic pain.
[DOCX File , 365 KB - jmir_v28i1e78233_app7.docx]

Multimedia Appendix 8
Reverse Mendelian randomization analysis assessing the causal effect of chronic pain on screen time.
[XLSX File (Microsoft Excel File), 14 KB - jmir_v28i1e78233_app8.xlsx]

Multimedia Appendix 9
Causal effects of the screen time on chronic pain based on inverse variance weighted multivariable Mendelian randomization.
[XLSX File (Microsoft Excel File), 13 KB - jmir_v28i1e78233_app9.xlsx]

Multimedia Appendix 10
The top hit probes identified in summary data–based Mendelian randomization analysis after false discovery rate correction of screen time.
[XLSX File (Microsoft Excel File), 533 KB - jmir_v28i1e78233_app10.xlsx]

Multimedia Appendix 11

XSL•FO
RenderX

The top hit probes identified in summary data–based Mendelian randomization analysis after false discovery rate correction of chronic pain.
[XLSX File (Microsoft Excel File), 108 KB - jmir_v28i1e78233_app11.xlsx]

Multimedia Appendix 12
The genes with overlapping exposure and results were screened by PFDR and PHEIDI in summary data–based Mendelian randomization analysis.
[XLSX File (Microsoft Excel File), 13 KB - jmir_v28i1e78233_app12.xlsx]

Multimedia Appendix 13
Transcriptome-wide association study of screen time in three cross-tissue (sparse canonical correlation analysis) weights using Bonferroni correction and cross-tissue expression Genotype-Tissue Expression v8.
[XLSX File (Microsoft Excel File), 269 KB - jmir_v28i1e78233_app13.xlsx]

Multimedia Appendix 14
Transcriptome-wide association study of chronic pain in three cross-tissue (sparse canonical correlation analysis) weights using Bonferroni correction and cross-tissue expression Genotype-Tissue Expression v8.
[XLSX File (Microsoft Excel File), 74 KB - jmir_v28i1e78233_app14.xlsx]

Multimedia Appendix 15
The results of the jointly significant genes for screen time.
[XLSX File (Microsoft Excel File), 73 KB - jmir_v28i1e78233_app15.xlsx]

Multimedia Appendix 16
The results of the jointly significant genes for chronic pain.
[XLSX File (Microsoft Excel File), 23 KB - jmir_v28i1e78233_app16.xlsx]

Multimedia Appendix 17
Overlapping genes by both transcriptome-wide association study and summary data–based Mendelian randomization methods.
[XLSX File (Microsoft Excel File), 10 KB - jmir_v28i1e78233_app17.xlsx]

Multimedia Appendix 18
Results of the colocalization analysis for genes identified in the exposure and outcome traits.
[XLSX File (Microsoft Excel File), 41 KB - jmir_v28i1e78233_app18.xlsx]

Multimedia Appendix 19
Overlapping genes in colocalization.
[XLSX File (Microsoft Excel File), 10 KB - jmir_v28i1e78233_app19.xlsx]

References

1. Pandya A, Lodha P. Social connectedness, excessive screen time during COVID-19 and mental health: a review of current evidence. Front Hum Dyn 2021;3. [doi: 10.3389/fhumd.2021.684137]
2. Van Dyck D, Cardon G, Deforche B, Owen N, De Cocker K, Wijndaele K, et al. Socio-demographic, psychosocial and home-environmental attributes associated with adults' domestic screen time. BMC Public Health 2011 Aug 25;11:668 [FREE Full text] [doi: 10.1186/1471-2458-11-668] [Medline: 21864412]
3. Goldfield GS, Murray M, Maras D, Wilson AL, Phillips P, Kenny GP, et al. Screen time is associated with depressive symptomatology among obese adolescents: a HEARTY study. Eur J Pediatr 2016;175(7):909-919. [doi: 10.1007/s00431-016-2720-z] [Medline: 27075014]
4. Nightingale CM, Rudnicka AR, Donin AS, Sattar N, Cook DG, Whincup PH, et al. Screen time is associated with adiposity and insulin resistance in children. Arch Dis Child 2017;102(7):612-616 [FREE Full text] [doi: 10.1136/archdischild-2016-312016] [Medline: 28288985]
5. Ahmadi-Abhari S, Sabia S, Shipley MJ, Kivimäki M, Singh-Manoux A, Tabak A, et al. Physical activity, sedentary behavior, and long-term changes in aortic stiffness: the Whitehall II study. J Am Heart Assoc 2017;6(8):e005974 [FREE Full text] [doi: 10.1161/JAHA.117.005974] [Medline: 28784651]
6. Chetty-Mhlanga S, Fuhrimann S, Eeftens M, Basera W, Hartinger S, Dalvie MA, et al. Different aspects of electronic media use, symptoms and neurocognitive outcomes of children and adolescents in the rural Western Cape region of South Africa. Environ Res 2020;184:109315. [doi: 10.1016/j.envres.2020.109315] [Medline: 32172071]

XSL•FO
RenderX

7. Marques A, Calmeiro L, Loureiro N, Frasquilho D, de Matos MG. Health complaints among adolescents: associations with more screen-based behaviours and less physical activity. J Adolesc 2015;44:150-157. [doi: 10.1016/j.adolescence.2015.07.018] [Medline: 26275746]
8. Taehtinen RE, Sigfusdottir ID, Helgason AR, Kristjansson AL. Electronic screen use and selected somatic symptoms in 10-12 year old children. Prev Med 2014;67:128-133. [doi: 10.1016/j.ypmed.2014.07.017] [Medline: 25045838]
9. Torsheim T, Eriksson L, Schnohr CW, Hansen F, Bjarnason T, Välimaa R. Screen-based activities and physical complaints among adolescents from the Nordic countries. BMC Public Health 2010;10:324 [FREE Full text] [doi: 10.1186/1471-2458-10-324] [Medline: 20534116]
10. Hakala PT, Saarni LA, Punamäki RL, Wallenius MA, Nygård CH, Rimpelä AH. Musculoskeletal symptoms and computer use among Finnish adolescents--pain intensity and inconvenience to everyday life: a cross-sectional study. BMC Musculoskelet Disord 2012;13:41 [FREE Full text] [doi: 10.1186/1471-2474-13-41] [Medline: 22439805]
11. Guerra PH, Martelo R, da Silva MN, de Andrade GF, Christofaro DGD, Loch MR. Screen time and low back pain in children and adolescents: a systematic review of Brazilian studies. Rev Paul Pediatr 2023;41:e2021342 [FREE Full text] [doi: 10.1590/1984-0462/2023/41/2021342] [Medline: 37042940]
12. Montagni I, Guichard E, Carpenet C, Tzourio C, Kurth T. Screen time exposure and reporting of headaches in young adults: A cross-sectional study. Cephalalgia 2016 Oct;36(11):1020-1027. [doi: 10.1177/0333102415620286] [Medline: 26634831]
13. Roman-Juan J, Roy R, Jensen MP, Miró J. The explanatory role of sedentary screen time and obesity in the increase of chronic back pain amongst European adolescents: the HBSC study 2002-2014. Eur J Pain 2022;26(8):1781-1789 [FREE Full text] [doi: 10.1002/ejp.2003] [Medline: 35761772]
14. Silva AG, Couto PS, Queirós A, Neto M, Rocha NP. Chronic pain in high school students is associated with physical activity and sleeping hours but not with screen time. Int J Adolesc Med Health 2017;31(3):2017. [doi: 10.1515/ijamh-2017-0014] [Medline: 28628479]
15. Henry P, Auray JP, Gaudin AF, Dartigues JF, Duru G, Lantéri-Minet M, et al. Prevalence and clinical characteristics of migraine in France. Neurology 2002;59(2):232-237. [doi: 10.1212/wnl.59.2.232] [Medline: 12136063]
16. Kowacs P, Piovesan E, Werneck L, Fameli H, Pereira da Silva H. Headache related to a specific screen flickering frequency band. Cephalalgia 2004;24(5):408-410. [doi: 10.1111/j.1468-2982.2004.00686.x] [Medline: 15096230]
17. Robbins L. Precipitating factors in migraine: a retrospective review of 494 patients. Headache 1994;34(4):214-216. [doi: 10.1111/j.1526-4610.1994.hed3404214.x] [Medline: 8014037]
18. Johnston KJA, Adams MJ, Nicholl BI, Ward J, Strawbridge RJ, Ferguson A, et al. Genome-wide association study of multisite chronic pain in UK Biobank. PLoS Genet 2019;15(6):e1008164 [FREE Full text] [doi: 10.1371/journal.pgen.1008164] [Medline: 31194737]
19. Feng H, Mancuso N, Gusev A, Majumdar A, Major M, Pasaniuc B, et al. Leveraging expression from multiple tissues using sparse canonical correlation analysis and aggregate tests improves the power of transcriptome-wide association studies. PLoS Genet 2021;17(4):e1008973 [FREE Full text] [doi: 10.1371/journal.pgen.1008973] [Medline: 33831007]
20. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet 2016;48(5):481-487. [doi: 10.1038/ng.3538] [Medline: 27019110]
21. Burgess S, Thompson SG. Avoiding bias from weak instruments in Mendelian randomization studies. Int J Epidemiol 2011;40(3):755-764. [doi: 10.1093/ije/dyr036] [Medline: 21414999]
22. Rusk N. The UK Biobank. Nat Methods 2018;15(12):1001. [doi: 10.1038/s41592-018-0245-2] [Medline: 30504882]
23. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 2020;369(6509):1318-1330 [FREE Full text] [doi: 10.1126/science.aaz1776] [Medline: 32913098]
24. Võsa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet 2021 Sep;53(9):1300-1310 [FREE Full text] [doi: 10.1038/s41588-021-00913-z] [Medline: 34475573]
25. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol 2013;37(7):658-665 [FREE Full text] [doi: 10.1002/gepi.21758] [Medline: 24114802]
26. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Genet Epidemiol 2016;40(4):304-314 [FREE Full text] [doi: 10.1002/gepi.21965] [Medline: 27061298]
27. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol 2015;44(2):512-525 [FREE Full text] [doi: 10.1093/ije/dyv080] [Medline: 26050253]
28. Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. Eur J Epidemiol 2017;32(5):377-389 [FREE Full text] [doi: 10.1007/s10654-017-0255-x] [Medline: 28527048]
29. Verbanck M, Chen C, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nat Genet 2018;50(5):693-698 [FREE Full text] [doi: 10.1038/s41588-018-0099-7] [Medline: 29686387]

XSL•FO
RenderX

30. Burgess S, Bowden J, Fall T, Ingelsson E, Thompson SG. Sensitivity analyses for robust causal inference from Mendelian randomization analyses with multiple genetic variants. Epidemiology 2017;28(1):30-42 [FREE Full text] [doi: 10.1097/EDE.0000000000000559] [Medline: 27749700]
31. Mounier N, Kutalik Z. Bias correction for inverse variance weighting Mendelian randomization. Genet Epidemiol 2023;47(4):314-331. [doi: 10.1002/gepi.22522] [Medline: 37036286]
32. Sanderson E, Davey Smith G, Windmeijer F, Bowden J. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. Int J Epidemiol 2019;48(3):713-727 [FREE Full text] [doi: 10.1093/ije/dyy262] [Medline: 30535378]
33. Mo X, Guo Y, Qian Q, Fu M, Lei S, Zhang Y, et al. Mendelian randomization analysis revealed potential causal factors for systemic lupus erythematosus. Immunology 2020;159(3):279-288 [FREE Full text] [doi: 10.1111/imm.13144] [Medline: 31670388]
34. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet 2016;48(3):245-252 [FREE Full text] [doi: 10.1038/ng.3506] [Medline: 26854917]
35. GTEx Consortium, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Lead analysts:, NIH program management:, Biospecimen collection:, Pathology:, eQTL manuscript working group:, et al. Genetic effects on gene expression across human tissues. Nature 2017;550(7675):204-213 [FREE Full text] [doi: 10.1038/nature24277] [Medline: 29022597]
36. Kia DA, Zhang D, Guelfi S, Manzoni C, Hubbard L, Reynolds RH, et al. Identification of candidate parkinson disease genes by integrating genome-wide association study, expression, and epigenetic data sets. JAMA Neurol 2021;78(4):464-472 [FREE Full text] [doi: 10.1001/jamaneurol.2020.5257] [Medline: 33523105]
37. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet 2014;10(5):e1004383 [FREE Full text] [doi: 10.1371/journal.pgen.1004383] [Medline: 24830394]
38. Yi M, Feng X, Guan Q, Liu Y, Liu Y, Su Z. A multi-omics Mendelian randomization study reveals PAM as a potential therapeutic target for type 2 diabetes. J Transl Med 2025;23(1):1067 [FREE Full text] [doi: 10.1186/s12967-025-07086-x] [Medline: 41063246]
39. Cankurtaran F, Menevşe O, Namlı A, Kızıltoprak, Altay S, Duran M, et al. The impact of digital game addiction on musculoskeletal system of secondary school children. Niger J Clin Pract 2022;25(2):153-159 [FREE Full text] [doi: 10.4103/njcp.njcp_177_20] [Medline: 35170440]
40. Shehab DK, Al-Jarallah KF. Nonspecific low-back pain in Kuwaiti children and adolescents: associated factors. J Adolesc Health 2005;36(1):32-35. [doi: 10.1016/j.jadohealth.2003.12.011] [Medline: 15661594]
41. Buhle J, Wager TD. Performance-dependent inhibition of pain by an executive working memory task. Pain 2010;149(1):19-26 [FREE Full text] [doi: 10.1016/j.pain.2009.10.027] [Medline: 20129735]
42. Buhle JT, Stevens BL, Friedman JJ, Wager TD. Distraction and placebo: two separate routes to pain control. Psychol Sci 2012;23(3):246-253. [doi: 10.1177/0956797611427919] [Medline: 22261568]
43. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015;12(3):e1001779 [FREE Full text] [doi: 10.1371/journal.pmed.1001779] [Medline: 25826379]
44. Legrain V, Crombez G, Verhoeven K, Mouraux A. The role of working memory in the attentional control of pain. Pain 2011;152(2):453-459 [FREE Full text] [doi: 10.1016/j.pain.2010.11.024] [Medline: 21238855]
45. Czaja SJ, Charness N, Fisk AD, Hertzog C, Nair SN, Rogers WA, et al. Factors predicting the use of technology: findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). Psychol Aging 2006 Jun;21(2):333-352 [FREE Full text] [doi: 10.1037/0882-7974.21.2.333] [Medline: 16768579]
46. Saastamoinen P, Leino-Arjas P, Laaksonen M, Lahelma E. Socio-economic differences in the prevalence of acute, chronic and disabling chronic pain among ageing employees. Pain 2005;114(3):364-371. [doi: 10.1016/j.pain.2004.12.033] [Medline: 15777862]
47. Vinstrup J, Sundstrup E, Andersen LL. Psychosocial stress and musculoskeletal pain among senior workers from nine occupational groups: cross-sectional findings from the SeniorWorkingLife study. BMJ Open 2021;11(3):e043520 [FREE Full text] [doi: 10.1136/bmjopen-2020-043520] [Medline: 33782021]
48. Chinta SJ, Andersen JK. Dopaminergic neurons. Int J Biochem Cell Biol 2005;37(5):942-946. [doi: 10.1016/j.biocel.2004.09.009] [Medline: 15743669]
49. Navratilova E, Porreca F. Reward and motivation in pain and pain relief. Nat Neurosci 2014;17(10):1304-1312 [FREE Full text] [doi: 10.1038/nn.3811] [Medline: 25254980]
50. Wood PB. Role of central dopamine in pain and analgesia. Expert Rev Neurother 2008;8(5):781-797. [doi: 10.1586/14737175.8.5.781] [Medline: 18457535]

XSL•FO
RenderX

51. Mazo G, Soplop N, Wang W, Uryu K, Tsou MB. Spatial control of primary ciliogenesis by subdistal appendages alters sensation-associated properties of cilia. Dev Cell 2016;39(4):424-437 [FREE Full text] [doi: 10.1016/j.devcel.2016.10.006] [Medline: 27818179]
52. Welburn JPI, Cheeseman IM. The microtubule-binding protein Cep170 promotes the targeting of the kinesin-13 depolymerase Kif2b to the mitotic spindle. Mol Biol Cell 2012;23(24):4786-4795 [FREE Full text] [doi: 10.1091/mbc.E12-03-0214] [Medline: 23087211]
53. Weijman JF, Vuolo L, Shak C, Pugnetti A, Mukhopadhyay AG, Hodgson LR, et al. Roles for CEP170 in cilia function and dynein-2 assembly. J Cell Sci 2024;137(8):jcs261816 [FREE Full text] [doi: 10.1242/jcs.261816] [Medline: 38533689]
54. Fitzsimons LA, Staurengo-Ferrari L, Khomula EV, Bogen O, Araldi D, Bonet IJM, et al. The nociceptor primary cilium contributes to mechanical nociceptive threshold and inflammatory and neuropathic pain. J Neurosci 2024;44(47):e1265242024. [doi: 10.1523/JNEUROSCI.1265-24.2024] [Medline: 39349056]
55. Kifle ZD, Tian J, Aitken D, Melton PE, Cicuttini F, Jones G, et al. MRI-derived abdominal adipose tissue is associated with multisite and widespread chronic pain. Reg Anesth Pain Med 2025. [doi: 10.1136/rapm-2024-105535] [Medline: 39256036]
56. Neogi T. The epidemiology and impact of pain in osteoarthritis. Osteoarthritis Cartilage 2013;21(9):1145-1153 [FREE Full text] [doi: 10.1016/j.joca.2013.03.018] [Medline: 23973124]
57. Pou KM, Massaro JM, Hoffmann U, Vasan RS, Maurovich-Horvat P, Larson MG, et al. Visceral and subcutaneous adipose tissue volumes are cross-sectionally related to markers of inflammation and oxidative stress: the Framingham Heart Study. Circulation 2007;116(11):1234-1241. [doi: 10.1161/CIRCULATIONAHA.107.710509] [Medline: 17709633]
58. Jin Y, Ren Z, Tan Y, Zhao P, Wu J. Motility plays an important role in the lifetime of mammalian lipid droplets. Int J Mol Sci 2021;22(8):3802 [FREE Full text] [doi: 10.3390/ijms22083802] [Medline: 33916886]
59. Liang Y, Zhu Z, Lu Y, Ma C, Li J, Yu K, et al. Cytoskeleton regulates lipid droplet fusion and lipid storage by controlling lipid droplet movement. Biochim Biophys Acta Mol Cell Biol Lipids 2025;1870(4):159610. [doi: 10.1016/j.bbalip.2025.159610] [Medline: 40189192]
60. Dolcini J, Landi R, Ponzio E, Picchini L, Luciani A, Formenti L, et al. Association between TNF-α, cortisol levels, and exposure to PM10 and PM2.5: a pilot study. Environ Sci Eur 2024;36 [FREE Full text] [doi: 10.1186/s12302-024-00961-2]
61. Duan Y, Chen S, Li Q, Zang Y. Neuroimmune mechanisms underlying neuropathic pain: the potential role of TNF-α-necroptosis pathway. Int J Mol Sci 2022;23(13):7191 [FREE Full text] [doi: 10.3390/ijms23137191] [Medline: 35806192]
62. Knezevic E, Nenic K, Milanovic V, Knezevic NN. The role of cortisol in chronic stress, neurodegenerative diseases, and psychological disorders. Cells 2023;12(23):2726 [FREE Full text] [doi: 10.3390/cells12232726] [Medline: 38067154]
63. Data table. GTEx Consortium. URL: https://yanglab.westlake.edu.cn/data/SMR/GTEx_V8_cis_eqtl_summary.html [accessed 2024-05-25]
64. Feng H. GTEx cross-tissue (sCCA) expression. gusevlab. URL: http://gusevlab.org/projects/fusion/#gtex-cross-tissue-scca-expression [accessed 2024-05-25]
65. cis-eQTLs. eQTLGen Consortium. URL: https://www.eqtlgen.org/cis-eqtls.html [accessed 2024-05-25]

Abbreviations

eQTL: expression Quantitative Trait Locus
FDR: false discovery rate
GTEx: Genotype-Tissue Expression
GWAS: genome-wide association study
HEIDI: heterogeneity in dependent instruments
IEU: Integrative Epidemiology Unit
IV: instrumental variable
IVW: inverse variance weighted
LD: linkage disequilibrium
MAF: minor allele frequency
MCP: multisite chronic pain
MR: Mendelian randomization
MR-PRESSO: Mendelian Randomization Pleiotropy Residual Sum and Outlier
MVMR: multivariable Mendelian randomization
OR: odds ratio
PP: posterior probability
sCCA: sparse canonical correlation analysis
SMR: summary data–based Mendelian randomization
SNP: single-nucleotide polymorphism
TWAS: transcriptome-wide association study

XSL•FO
RenderX

Edited by A Coristine; submitted 29.May.2025; peer-reviewed by J Dolcini, Y Zhu; comments to author 24.Oct.2025; accepted 24.Dec.2025; published 09.Feb.2026.

Please cite as:
Jiang J, Pu C, Cai J, Yu C, Liu Z, Xiao C
Screen Time and Chronic Pain Health: Mendelian Randomization Study
J Med Internet Res 2026;28:e78233
URL: https://www.jmir.org/2026/1/e78233
doi:10.2196/78233
PMID:

©Jiahui Jiang, Chunyan Pu, Jiarui Cai, Chuan Yu, Zhenmi Liu, Chenghan Xiao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 09.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Genre-Specific Gaming Addiction and Flourishing in Adolescents: Cross-Sectional Survey Study

Yan Luo[1], PhD; Xiaofei Chen[1]; Yaoyu Liu[1]; Wanjia Hong[1]; Miner Fan[1], LLB; David Thomas[2,3], MS; Sean P Mullen[3,4], PhD

[1]Department of Physical Education, Shanghai Jiao Tong University, Shanghai, Shanghai, China

[2]Informatics Programs, University of Illinois Urbana-Champaign, Urbana, IL, United States

[3]Department of Health and Kinesiology, University of Illinois Urbana-Champaign, Urbana, IL, United States

[4]Beckman Institute for Advanced Science and Technology, University of Illinois Urbana-Champaign, Urbana, IL, United States

Corresponding Author:
Yan Luo, PhD
Department of Physical Education
Shanghai Jiao Tong University
800 Dongchuan Road
Shanghai, Shanghai, 200240
China
Phone: 86 19901790215
Email: yluo2@sjtu.edu.cn

Abstract

Background: Adolescent gaming addiction (GA) has been linked to a range of adverse health outcomes. However, whether the associated health risks differ across game genres remains poorly understood.

Objective: Guided by VanderWeele's multidimensional flourishing framework, this study aims to examine genre-specific associations between GA and flourishing among adolescents.

Methods: This study used a cross-sectional observational design. A total of 2194 middle school students were recruited via convenience sampling from a private tutoring center in a northwestern city in China. Eligibility criteria were (1) enrollment in participating classes at the tutoring center, (2) provision of both student and parental consent, and (3) presence during questionnaire administration. The mean age of participants was 14.53 (SD 0.76) years; 985 (44.90%) were boys and 1174 (53.51%) were girls. During class time, students completed paper-based questionnaires that assessed their demographics, gaming addiction, and flourishing. Participants listed up to 3 video games played in the past month and rated their addiction to each. Games were classified into 8 genres: action and adventure (AA), sandbox and simulation (SS), multiplayer online battle arena (MOBA), shooting, strategy, casual, sports, and role-playing. Flourishing was assessed using the Human Flourishing Index across 5 domains: happiness and life satisfaction, mental and physical health, meaning and purpose, character and virtue, and close social relationships.

Results: Robust linear regression analyses (α=.05) showed that AA addiction was associated with lower overall flourishing (b=–3.11, 95% CI –4.34 to –1.88) and all 5 subdomains (happiness and life satisfaction: b=–0.46, 95% CI –0.75 to –0.17; mental and physical health: b=–0.61, 95% CI –0.88 to –0.34; meaning and purpose: b=–0.55, 95% CI –0.82 to –0.27; character and virtue: b=–0.74, 95% CI –1.06 to –0.43; and close social relationships: b=–0.62, 95% CI –0.92 to –0.32). MOBA addiction was associated with lower overall flourishing (b=–1.33, 95% CI –2.34 to –0.32), character and virtue (b=–0.34, 95% CI –0.59 to –0.08), and meaning and purpose (b=–0.34, 95% CI –0.56 to –0.11). SS addiction was associated with lower overall flourishing (b=–3.42, 95% CI –5.80 to –1.04), close social relationships (b=–0.86, 95% CI –1.46 to –0.27), and mental and physical health (b=–1.09, 95% CI –1.60 to –0.58).

Conclusions: This study provides novel evidence that the association between GA and adolescent flourishing is genre dependent. In contrast to prior research that conceptualizes health narrowly or unidimensionally, a multidimensional perspective provides a more nuanced understanding of the health risks associated with GA. The findings advance the field by showing that addiction to AA, MOBA, and SS games is associated with greater health risks than addiction to other genres. Accordingly, prevention, education, and policy efforts should prioritize higher-risk genres to promote adolescent health.

(J Med Internet Res 2026;28:e89319) doi:10.2196/89319

XSL•FO
RenderX

KEYWORDS

digital game; online game; problematic gaming; gaming disorder; youth

Introduction

Background

Video gaming is one of the most prevalent leisure activities among adolescents worldwide. For many, gaming provides entertainment, opportunities for social interaction, and cognitive benefits [1]. Yet, for a notable proportion of adolescents, gaming becomes addictive, characterized by preoccupation, impaired control, and persistence despite negative consequences. A meta-analysis estimated the prevalence of gaming disorder, a severe and clinical form of gaming addiction (GA), among adolescents to be 8.8%, based on over 400,000 participants from 155 studies across 33 countries [2]. This rate is considerably higher than that observed in other age groups [3]. Therefore, it is critical to investigate the associations between GA and adolescents' health to inform interventions and policies that foster healthier gaming practices.

A substantial body of research has documented that GA is associated with a range of negative health outcomes in adolescents [2,4]. Although these studies have significantly advanced our understanding of GA-related harms, a notable limitation is that most have adopted a relatively narrow conceptualization of health, focusing disproportionately on pathological conditions. According to the World Health Organization, health is defined as "a state of complete physical, mental, and social wellbeing" [5], which highlights the need to investigate the relationship between GA and health from a broader perspective. One construct that captures this view is flourishing, which reflects a state in which "all aspects of a person's life are good" [6]. VanderWeele [6] proposed that flourishing should, at a minimum, encompass 5 domains of human life: (1) happiness and life satisfaction, (2) mental and physical health, (3) meaning and purpose, (4) character and virtue, and (5) close social relationships [6]. Collectively, these domains reflect distinct facets of a well-lived life. Happiness and life satisfaction reflect how good life feels overall. Mental and physical health capture how well individuals are functioning psychologically and biologically. Meaning and purpose speak to whether life feels worthwhile and directed. Character and virtue reflect the capacity for moral self-regulation and for acting well when facing challenges. Close social relationships reflect the quality of one's relational world. Although these domains are often correlated, they are conceptually distinct. Flourishing entails doing well across all 5 domains, and impairment in any single domain indicates a substantive shortfall in overall well-being. Therefore, examining the association between GA and flourishing is crucial, as it provides a more holistic understanding of how GA may constrain adolescents' ability to achieve well-rounded lives and may inform prevention and policy efforts to promote adolescents' health.

Review of Relevant Scholarship

Previous research suggests that GA is associated with multiple flourishing-related domains. In the domain of happiness and life satisfaction, a large cross-national study of over 14,000 adolescents found that GA was linked to lower life satisfaction [7]. A longitudinal study further showed that GA predicted subsequent declines in adolescents' life satisfaction over 6 months, whereas the reverse effect was not observed [8]. These findings suggest that GA may function as an antecedent of reduced life satisfaction. Although no studies to date have directly examined the association between GA and happiness among adolescents, GA has been consistently shown to correlate with several factors strongly connected to happiness [9-13], including poorer academic achievement [2], lower self-esteem [2], reduced health-related quality of life [14], more frequent negative life events [15], and worse family financial conditions [16]. In summary, existing evidence suggests that GA is associated with lower life satisfaction and may also be linked to lower levels of happiness among adolescents.

The relationship between GA and mental health outcomes has been the most extensively investigated. Studies using composite indicators of mental health have shown that GA was associated with more frequent psychological complaints [7], elevated psychopathological symptoms [17-19], and a higher likelihood of receiving at least one mental health diagnosis [20]. Research focusing on specific mental health conditions further revealed that GA was associated with increased depression [2,21], anxiety [22-24], general and academic stress [2,25,26], loneliness [27], emotional distress [2,23,28,29], suicidal ideation [19,30], suicide risk [31], emotional and behavioral problems [24], sleep difficulties [4,32], and executive dysfunction [32]. A few studies have examined the association between GA and physical health. Evidence suggests that GA is linked to unhealthy weight status, including both underweight and overweight [33,34], more somatic complaints [17], musculoskeletal pain [4], and dry eye symptoms [4]. Overall, these findings suggest that GA is associated with a broad range of mental and physical health problems, with the literature to date focusing predominantly on mental health outcomes.

Evidence supporting associations between GA and the remaining flourishing domains is limited in the current literature. Existing findings suggest that adolescents with GA reported both a lower presence of meaning in life and a stronger search for meaning in life [35]. Although no adolescent studies have directly examined GA in relation to sense of purpose, research with young adults indicates a negative association between GA and purpose in life [36,37]. Regarding character and virtue, studies have shown that GA is linked to lower empathy [38], lower moral levels [39], reduced academic perseverance [40], and fewer prosocial behaviors [38]. In the relational domain, GA has been associated with lower interpersonal trust [41], conflictual parent-child relationships [42,43], peer alienation [26], less social support [44], and greater reliance on online-only friendships [14]. Overall, the available evidence suggests that GA may have broad adverse associations across these flourishing domains in adolescents.

Beyond research conducted on individual domains of flourishing, no study to date has investigated GA in relation to

XSL•FO
RenderX

flourishing as an integrated construct. Existing work on other problematic behaviors has found that lower flourishing was related to greater cannabis use [45], higher media addiction [46], higher smartphone addiction [47], and higher Facebook addiction [48,49]. These findings imply that GA, as a form of problematic behavior, may likewise be associated with reduced flourishing. It is important to note that the flourishing framework used in these studies is derived from Diener and colleagues [50], which captures all but the health dimension of flourishing proposed by VanderWeele [6]. As a result, it remains unclear whether VanderWeele's broader conceptualization of flourishing is associated with any form of problematic behavior, including GA. A recent study that employed VanderWeele's flourishing framework examined the relationship between digital game use and flourishing among older adults. No differences were found in overall flourishing or in any of the 5 subdomains between current gamers, former gamers, and nongamers [51]. As this study did not focus on GA and was conducted in an older adult population, its findings provide minimal guidance regarding how GA relates to flourishing among adolescent gamers.

Although prior studies have documented associations between GA and individual elements of flourishing, a critical gap remains regarding how GA relates to overall flourishing and its subdomains among adolescents. As flourishing is a multidimensional construct encompassing various aspects of well-being, investigating these associations is essential for establishing a more comprehensive understanding of how GA may relate to adolescents' well-being across multiple life domains. A further limitation of the existing literature is that most studies have conceptualized gaming as a single, homogeneous activity. It remains unclear whether GA is uniformly associated with flourishing across different game genres. Emerging evidence suggests that the psychological correlates of gaming may vary substantially by genre. For example, Park and colleagues [52] reported that adolescents who primarily played massively multiplayer online role-playing games exhibited the highest levels of social anxiety, those who preferred first-person shooter games showed the lowest anxiety levels, and strategy-game players reported the highest self-esteem [52]. Other studies have demonstrated that certain genres were associated with increased aggression, reduced empathy, and prosocial behaviors [53-56]. As these factors are important components of flourishing, such findings suggest that the association between GA and flourishing may differ depending on the game genres adolescents play. Understanding these genre-specific associations could provide a more nuanced picture of how GA relates to adolescent health, thereby informing the development of targeted intervention strategies tailored to specific game genres.

Hypothesis and Aim

To address these gaps, this study aimed to investigate genre-specific associations between GA and flourishing among adolescents, using VanderWeele's multidimensional flourishing framework. Drawing on prior evidence, we specified a single primary hypothesis: *GA would be negatively associated with overall flourishing and its subdomains for certain game genres, but not uniformly across all genres.* Our secondary hypothesis was that *overall GA across game genres would be negatively associated with overall flourishing and its subdomains.*

Methods

Inclusion and Exclusion

This study was reported in accordance with the Journal Article Reporting Standards for Quantitative Research (JARS-Quant) checklist [57]. A total of 2194 middle school students from a private tutoring center located in a northwestern city in China participated in this study. Inclusion criteria were as follows: (1) enrollment in participating classes at the tutoring center; (2) provision of both student and parental consent; and (3) presence during questionnaire administration. No exclusion criteria were applied.

Participant Characteristics

As shown in Table 1, participants had a mean age of 14.53 (SD 0.76) years. Of the 2194 students, there were 985 (44.90%) boys and 1174 (53.51%) girls. Regarding grade level, 52 (2.37%) were in the seventh grade, 969 (44.17%) were in the eighth grade, and 1092 (49.77%) were in the ninth grade. On average, participants reported a family economic status of 2.23 (SD 0.70) and 18.81 (SD 16.40) hours of social media use per week. In terms of gaming behavior, 610 (27.80%) participants reported not playing any games, 327 (14.90%) reported playing 1 game, 525 (23.93%) reported playing 2 games, and 732 (33.36%) reported playing 3 games. The most frequently played game genre was multiplayer online battle arena (MOBA; n=808, 36.83%), followed by shooting (n=589, 26.85%), action and adventure (AA; n=342, 15.59%), casual (n=302, 13.76%), sports (n=156, 7.11%), sandbox and simulation (SS; n=99, 4.51%), role-playing (n=89, 4.06%), and strategy (n=54, 2.46%) games.

XSL•FO
RenderX

Table 1. Descriptive statistics of demographic and gaming characteristics (N=2194).

Variables	Values
Age (years), mean (SD)	14.53 (0.76)
Sex, n (%)[a]	
Men	985 (44.90)
Women	1174 (53.51)
Grade level, n (%)	
Seventh grade	52 (2.37)
Eighth grade	969 (44.17)
Ninth grade	1092 (49.77)
Family economic status[b], mean (SD)	2.23 (0.70)
Social media time (hours), mean (SD)	18.81 (16.40)
Number of games played, n (%)	
Do not play a game	610 (27.80)
Playing 1 game	327 (14.90)
Playing 2 games	525 (23.93)
Playing 3 games	732 (33.36)
Number of participants per genre[c], n (%)	
Multiplayer online battle arena	808 (36.83)
Shooting	589 (26.85)
Action and adventure	342 (15.59)
Casual	302 (13.76)
Sports	156 (7.11)
Sandbox and simulation	99 (4.51)
Role-playing	89 (4.06)
Strategy	54 (2.46)

[a]35 missing values for this variable.

[b]Used a 5-point Likert scale for evaluation (0=extremely poor to 4=excellent).

[c]The number of participants per genre refers to those who reported playing at least one game within that genre.

Sampling Procedures

Participants were recruited using convenience sampling. A total of 2269 middle school students were approached, of whom 2194 (96.69%) agreed to participate and completed the survey. Self-selection occurred at both the teacher and student levels. At the teacher level, instructors at the tutoring center were invited to assist with data collection, and 10 teachers agreed to participate. At the student level, only students who expressed interest and provided both student assent and parental consent were included. Each participating teacher administered the survey once to each of their classes, all of which were held at the tutoring center. Each participant received ¥10 (US $1.44) in cash as compensation for completing the survey. Study procedures were conducted in compliance with the Declaration of Helsinki and approved by the ethics committee of the first author's institution (approval number H20250747I). Data were collected between September 10 and October 10, 2025.

Sample Size, Power, and Precision

We conducted a priori power analysis using G*Power for a linear multiple regression model (fixed model, R^2 increase) to determine the required sample size [58]. We assumed a small effect size for each genre-specific addiction predictor (partial R^2=0.01), an α level of .05, and a desired statistical power of 0.80. The model specified 1 tested predictor and a total of 12 predictors (8 genre-specific addiction variables and 4 covariates). Under these assumptions, the minimum required sample size was 779 participants. Given the achieved sample size of 2194 students, the study was sufficiently powered to detect genre-specific addiction predictors accounting for as little as 1% of unique variance in the outcome.

Measures and Covariates

Measures and covariates are presented in Textbox 1.

XSL•FO
RenderX

Textbox 1. Measures and covariates.

1. Genre-specific addiction

The extent to which an individual exhibits excessive and compulsive engagement with a specific genre of video games.

2. Flourishing

A multidimensional state of well-being encompassing domains of happiness and life satisfaction, mental and physical health, meaning and purpose, character and virtue, and close social relationships.

3. Sex

A biological classification of individuals, typically men or women.

4. Social media use

The extent to which students engage with social networking platforms.

5. Family economic status

The financial conditions of a household, reflecting the family's access to economic resources and its relative position within the socioeconomic hierarchy.

Data Collection

We contacted a friend of the first author (YL), a physics teacher at the private tutoring center, to inform him about the study's objectives and data collection procedures. After agreeing to assist, he introduced us to other teachers at the center. Ten additional teachers consented to participate in data collection.

We compiled measurement instruments into a questionnaire, which also included demographic items (age, sex, and grade). Each questionnaire was labeled with a unique identifier number at the bottom of every page to distinguish individual responses. Digital versions of the questionnaire and consent form were sent to the teachers via WeChat (Tencent Holdings Limited). Teachers were instructed to print these materials for use in class.

Each teacher administered the survey once to each of their classes, all of which were held at the tutoring center. At the beginning of each class, teachers explained the study's purpose and distributed parental consent forms to students who expressed interest in participating. Students who missed 2 classes were excluded from data collection. During the third class session, teachers distributed the questionnaires. Students were told that the survey was anonymous and that neither their parents nor their teachers would have access to their responses. They were instructed to answer honestly, write legibly, and seek clarification if any survey items were unclear. Completing the survey took approximately 10 minutes. The friend of the first author collected all the questionnaires from the teachers and shipped them to the first author. Each teacher received ¥800 (US $115.01) in cash as compensation. The survey questionnaire is available in Multimedia Appendix 1.

Quality of Measurements

To enhance the quality of measurements, we conducted an online group meeting with all participating teachers before data collection to provide standardized training and explain the data collection protocol in detail. In addition, we trained research assistants in standardized procedures for digitizing paper questionnaires to ensure accurate data entry.

Instrumentation

Video Game Name

The video game name was obtained using the following question: "Please provide the names of all the games you have played during the past month (up to 3). If you have played more than 3, please list the 3 you played most frequently."

Game-Specific Addiction

For each game, participants rated their level of addiction by responding to the following question: "Please indicate your level of addiction for each of the games listed above." Responses were given on a 5-point Likert scale (0=none, 1=mild, 2=moderate, 3=moderately severe, and 4=severe), with higher scores indicating stronger addiction to a particular game. We used a single item to measure GA for 3 reasons. First, widely used instruments such as the Gaming Addiction Scale for Adolescents [59] and the Internet Gaming Disorder Scale [60] assess overall GA and are therefore unsuitable for measuring game-specific addiction. Second, even if these scales were adapted for game-specific assessment, participants who played 2 or 3 games would need to complete the same scale multiple times, increasing response burden and the risk of response bias. Third, multi-item scales would substantially increase completion time, which was impractical given that the survey was administered during class. Similar single-item measures of GA have been used in prior research, although primarily in adult samples [61,62]. In adolescent populations, single self-diagnostic items assessing other behavioral addictions have shown strong correlations with standardized multi-item instruments [63,64], suggesting that adolescents can comprehend and self-evaluate the severity of behavioral addictions using a single question. The process of transforming game-specific addiction into genre-specific addiction is described in the "Data Processing" section.

Flourishing

Flourishing was assessed using the Chinese version of the Human Flourishing Index (HFI) [65]. The scale consists of 10 items covering 5 domains (2 items per domain): (1) happiness and life satisfaction; (2) mental and physical health; (3) meaning and purpose; (4) character and virtue; and (5) close social

XSL•FO
RenderX

relationships. Items were rated on an 11-point Likert scale ranging from 0 to 10, with response labels varying across items. Domain scores were computed by summing the scores of the respective items. Overall flourishing was calculated as the sum of all item scores, with higher values reflecting greater flourishing. If any item required to compute a composite score was missing, the corresponding composite score was coded as missing.

Covariates

Previous research has found associations between GA, sex [66], social media use [67], and socioeconomic status [68]. To account for their confounding effects [69], we measured participants' weekly social media time and perceived family economic status using 2 questions. Participants were asked, "In general, how many hours per day do you spend browsing social media platforms (eg, Douyin, Weibo, WeChat Moments, Xiaohongshu, and Bilibili) on weekdays and weekends? Please report to the nearest 0 or 30 minutes (eg, 2 hours 0 minutes, 4 hours 30 minutes)." Total weekly social media time (in hours) was calculated using the formula: 5 × (weekday hours + weekday minutes/60) + 2 × (weekend hours + weekend minutes/60). The family economic status question was, "How would you describe your family's economic status?" Responses were on a 5-point Likert scale (0=extremely poor to 4=excellent), with higher scores reflecting better family economic status.

Masking, Psychometrics, Conditions and Design, and Data Processing

As this study employed a cross-sectional observational design without experimental manipulation, masking of participants, data collectors, or outcome assessors was not applicable.

The HFI demonstrated good internal consistency in our sample (Cronbach α=0.93 for overall flourishing, and ranging from 0.76 to 0.89 across domains). Game-specific addiction and family economic status were assessed using single-item measures, and sex and social media use were measured using non–Likert-type items. Therefore, internal consistency coefficients were not applicable for these measures.

This study used a cross-sectional observational design with no experimental manipulation.

Details of the data processing are provided in Multimedia Appendix 2.

Data Diagnostics

We screened social media use for implausible values and recoded 44 values exceeding 98 hours per week (>14 hours/day) as missing. Extreme outliers were identified using the IQR method. Fourteen values falling below Q1 – 3 × IQR or above Q3 + 3 × IQR were recoded as missing. All other variables were examined for implausible values, and none were identified. Responses were considered missing when participants left an item unanswered. Missing data were handled using multiple imputation by chained equations before statistical analysis. We examined the distributions of the outcome variables and found that overall flourishing and its 5 subscales were negatively skewed across all datasets.

Analytic Strategy

All statistical analyses were conducted in R (R Foundation) [70]. To reduce the influence of nonnormal outcome distributions on parameter estimation, we employed robust linear regression models [71]. To test the primary hypothesis, for each of the 6 flourishing outcomes, we fitted robust multiple regression models for each dataset, including the 8 genre-specific addiction variables as predictors, and sex, social media time, family economic status, and addiction to other games as covariates. To test the secondary hypothesis, we fitted separate robust regression models for each of the 6 flourishing outcomes with overall GA as a predictor, and sex, social media time, and family economic status as covariates. Parameter estimates were pooled across datasets using Rubin's rules. We controlled the type I error rate across all models for genre-specific addiction predictors (6 pooled estimates × 8 genres=48 *P* values) and the overall GA predictors (6 pooled estimates) by adjusting the 54 *P* values using the Benjamini-Hochberg false discovery rate procedure. The adjusted *P* values were hereafter referred to as *q* values. Statistical significance for all *q* values was set at α=.05.

As sensitivity analyses, we refitted all models using complete-case data to assess the robustness of the findings to the handling of missing data. The false discovery rate procedure was also applied in complete-case analyses. We further evaluated the robustness of the observed associations to unmeasured confounding by calculating *E* values for statistically significant pooled regression coefficients using the "EValue" package in R [72]. *E* values quantify the minimum strength of association (on the risk ratio scale) that unmeasured confounder(s) would need to have with both the exposure and the outcome, beyond the measured covariates, to fully explain the observed effect [73]. We also calculated *E* values for the CI limit closest to the null, which quantify the minimum strength of association that unmeasured confounder(s) would need to render the observed effect nonsignificant. *E* values were calculated separately for each imputed dataset, and the mean and SD across the 20 imputed datasets were reported.

Ethical Considerations

This study was approved by the Ethics Committee of Shanghai Jiao Tong University (approval number H20250747I). After the study procedures were explained, students who expressed interest were given parental consent forms to take home. Written informed consent was obtained from parents through signed consent forms before participation. Only students who returned signed parental consent forms and agreed to participate completed the survey, and participation was voluntary. Data were collected anonymously to protect participants' confidentiality, and only the research team had access to the dataset. Each participant received ¥10 (US $1.44) in cash as compensation for completing the survey. No identifiable images of participants are included in the manuscript or its multimedia appendices.

XSL•FO
RenderX

Results

Participant Flow and Recruitment

Of the 2269 middle school students eligible for participation, 68 were excluded due to absence from 2 consecutive data collection sessions, and 7 declined to participate. Figure 1 shows the flow of participants across each stage of the study. The final sample included 2194 participants.

Participants were recruited between September 10 and September 24, 2025.

Figure 1. Flow of participants through each stage of the study.

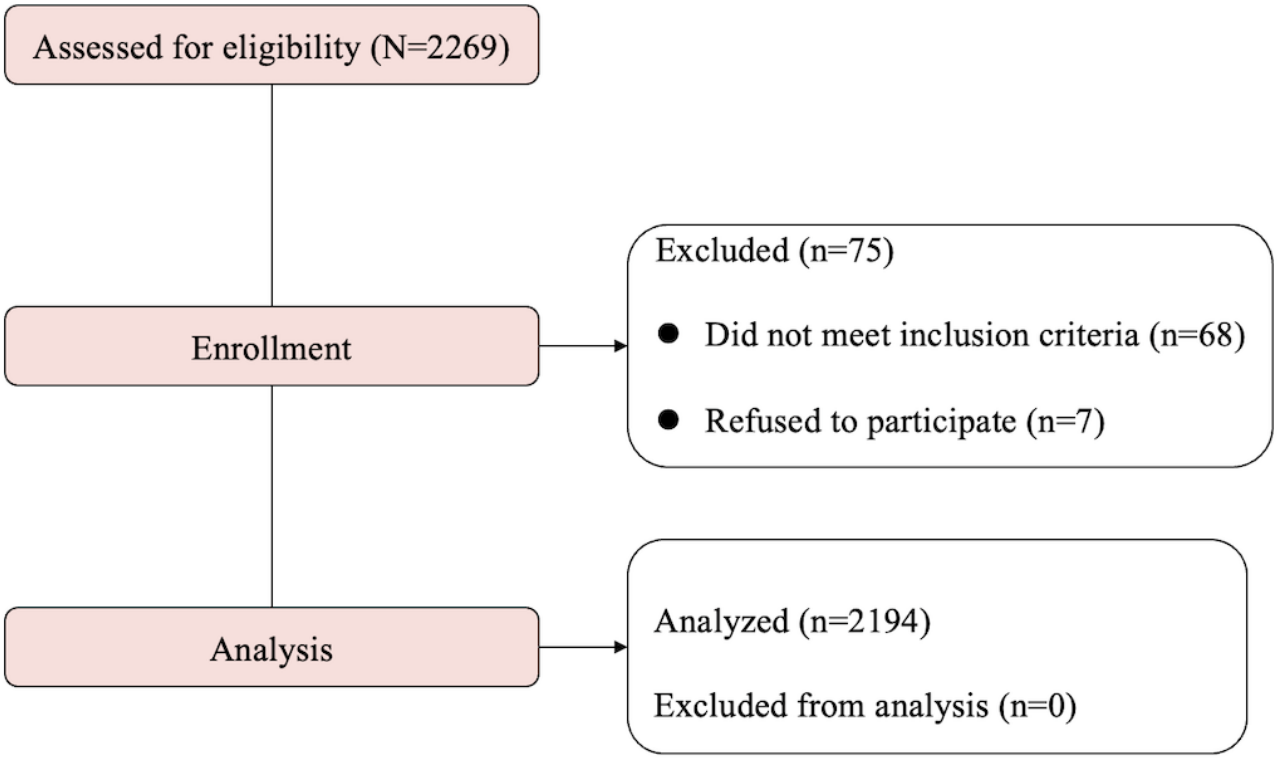

Statistics and Data Analysis

Inspection of missing data revealed that, among the 2194 participating students, data were missing for age in 22 (1%) cases, sex in 35 (1.60%), grade in 81 (3.69%), family economic status in 254 (11.58%), and social media time in 378 (17.23%). For the HFI, missing data were observed for happiness and life satisfaction in 34 (1.55%) participants, mental and physical health in 34 (1.55%), meaning and purpose in 170 (7.75%), character and virtue in 174 (7.93%), close social relationships in 164 (7.47%), and overall flourishing in 226 (10.30%). Across individual HFI items, missing data ranged from 24 (1.09%) to 164 (7.47%). For genre-specific addiction variables, missing data ranged from 3 (0.14%) to 76 (3.46%). In addition, data were missing for the addiction-to-other-games variable in 24 (1.09%) participants, whereas no missing data were observed for the overall addiction variable.

The Little missing completely at random test applied to the dataset was statistically significant, χ^2_{1298}=1495.00 (P<.001), indicating that the missing mechanism was not missing completely at random. As several variables exceeded 5% missingness and complete-case analysis can reduce statistical power and introduce bias [74], we assumed that the missingness mechanism was missing at random and applied multiple imputation by chained equations to impute missing data [75]. The predictor matrix included sex, family economic status, social media time, 10 flourishing items, 8 genre-specific addiction variables, and the addiction-to-other-games variable. The overall addiction variable was excluded due to its high linear dependence on other addiction variables. Sex was imputed using logistic regression, and all other variables were imputed using predictive mean matching. We generated 20 imputed datasets, each with 20 iterations. Convergence was confirmed using trace plots, and the distributions of the imputed values aligned well with those of the observed data. After imputation, we calculated composite scores for each dataset.

As shown in Table 2, on average, participants reported a mean overall GA score of 1.49 (SD 1.23). Among genre-specific addictions, participants reported the highest levels of addiction to strategy games (mean 1.65, SD 1.11), followed by AA (mean 1.62, SD 1.24), shooting (mean 1.58, SD 1.21), MOBA (mean 1.46, SD 1.10), SS (mean 1.38, SD 1.36), role-playing (mean 1.36, SD 1.02), casual (mean 1.11, SD 1.09), and sports (mean 0.96, SD 1.01) games. Regarding the HFI, the mean domain scores were 13.54 (SD 4.82) for happiness and life satisfaction, 14.54 (SD 4.72) for mental and physical health, 13.72 (SD 4.58) for meaning and purpose, 12.50 (SD 4.96) for character and virtue, and 14.24 (SD 4.99) for close social relationships. The overall flourishing score averaged 68.60 (SD 20.99).

XSL•FO
RenderX

Table 2. Descriptive statistics of gaming addiction and flourishing outcomes (N=2194).

Variables	Values, mean (SD)
Overall addiction	1.49 (1.23)
Gaming addiction per genre	
Multiplayer online battle arena	1.46 (1.10)
Shooting	1.58 (1.21)
Action and adventure	1.62 (1.24)
Casual	1.11 (1.09)
Sports	0.96 (1.01)
Sandbox and simulation	1.38 (1.36)
Role-playing	1.36 (1.02)
Strategy	1.65 (1.11)
Human Flourishing Index	
Happiness and life satisfaction	13.54 (4.82)
Mental and physical health	14.54 (4.72)
Meaning and purpose	13.72 (4.58)
Character and virtue	12.50 (4.96)
Close social relationships	14.24 (4.99)
Overall flourishing	68.60 (20.99)

Table 3 and Figure 2 present robust linear regression results examining the associations between GA and flourishing outcomes using multiply imputed data. Regarding the primary hypothesis, AA addiction was the only predictor that was negatively associated with all flourishing outcomes (overall flourishing: b=–3.11, 95% CI –4.34 to –1.88, q<.001; happiness and life satisfaction: b=–0.46, 95% CI –0.75 to –0.17, q=.009; mental and physical health: b=–0.61, 95% CI –0.88 to –0.34, q<.001; meaning and purpose: b=–0.55, 95% CI –0.82 to –0.27, q=.001; character and virtue: b=–0.74, 95% CI –1.06 to –0.43, q<.001; and close social relationships: b=–0.62, 95% CI –0.92 to –0.32, q<.001). MOBA addiction was negatively associated with overall flourishing (b=–1.33, 95% CI –2.34 to –0.32, q=.03), meaning and purpose (b=–0.34, 95% CI –0.56 to –0.11, q=.02), and character and virtue (b=–0.34, 95% CI –0.59 to –0.08, q=.03). SS addiction was negatively associated with overall flourishing (b=–3.42, 95% CI –5.80 to –1.04, q=.02), mental and physical health (b=–1.09, 95% CI –1.60 to –0.58, q<.001), and close social relationships (b=–0.86, 95% CI –1.46 to –0.27, q=.02). Other genre-specific associations were not statistically significant (q≥.05 for all; discussed later). Analyses testing the secondary hypothesis showed that overall GA was negatively associated with all flourishing outcomes (overall flourishing: b=–1.99, 95% CI –2.74 to –1.24, q<.001; happiness and life satisfaction: b=–0.24, 95% CI –0.41 to –0.07, q=.02; mental and physical health: b=–0.38, 95% CI –0.55 to –0.22, q<.001; meaning and purpose: b=–0.45, 95% CI –0.62 to –0.28, q<.001; character and virtue: b=–0.60, 95% CI –0.78 to –0.41, q<.001; and close social relationships: b=–0.35, 95% CI –0.53 to –0.17, q=.001).

XSL•FO
RenderX

Table 3. Robust linear regression results for the associations between gaming addiction and flourishing outcomes based on multiple imputed data (N=2194).

Variables[a]	Multiple imputation analysis[b]	
	b (95% CI)	*q* value[c]
Outcome: overall flourishing		
Predictors (genre-specific model)		
Multiplayer online battle arena	–1.33 (–2.34 to –0.32)	.03
Shooting	0.39 (–0.69 to 1.47)	.57
Casual	–1.54 (–3.24 to 0.15)	.17
Action and adventure	–3.11 (–4.34 to –1.88)	<.001
Sandbox and simulation	–3.42 (–5.80 to –1.04)	.02
Sports	–1.43 (–3.94 to 1.07)	.36
Strategy	–2.17 (–5.05 to 0.71)	.22
Role-playing	–0.39 (–3.09 to 2.31)	.84
Predictor (overall addiction model)		
Overall addiction	–1.99 (–2.74 to –1.24)	<.001
Outcome: happiness and life satisfaction		
Predictors (genre-specific model)		
Multiplayer online battle arena	–0.28 (–0.51 to –0.05)	.05
Shooting	0.22 (–0.03 to 0.47)	.18
Casual	–0.23 (–0.61 to 0.16)	.36
Action and adventure	–0.46 (–0.75 to –0.17)	.009
Sandbox and simulation	–0.61 (–1.16 to –0.05)	.08
Sports	–0.06 (–0.64 to 0.52)	.84
Strategy	–0.40 (–1.07 to 0.27)	.36
Role-playing	–0.07 (–0.69 to 0.55)	.84
Predictor (overall addiction model)		
Overall addiction	–0.24 (–0.41 to –0.07)	.02
Outcome: mental and physical health		
Predictors (genre-specific model)		
Multiplayer online battle arena	–0.18 (–0.40 to 0.04)	.20
Shooting	0.19 (–0.05 to 0.43)	.20
Casual	–0.37 (–0.74 to 0.00)	.12
Action and adventure	–0.61 (–0.88 to –0.34)	<.001
Sandbox and simulation	–1.09 (–1.60 to –0.58)	<.001
Sports	–0.39 (–0.94 to 0.17)	.26
Strategy	–0.38 (–1.05 to 0.30)	.36
Role-playing	–0.46 (–1.06 to 0.15)	.22
Predictor (overall addiction model)		
Overall addiction	–0.38 (–0.55 to –0.22)	<.001
Outcome: meaning and purpose		
Predictors (genre-specific model)		
Multiplayer online battle arena	–0.34 (–0.56 to –0.11)	.02
Shooting	–0.02 (–0.27 to 0.22)	.84

XSL•FO
RenderX

Variables[a]	Multiple imputation analysis[b]	
	b (95% CI)	*q* value[c]
Casual	–0.37 (–0.76 to 0.01)	.13
Action and adventure	–0.55 (–0.82 to –0.27)	.001
Sandbox and simulation	–0.40 (–0.93 to 0.13)	.22
Sports	–0.36 (–0.93 to 0.21)	.32
Strategy	–0.55 (–1.20 to 0.09)	.18
Role-playing	–0.12 (–0.72 to 0.49)	.78
Predictor (overall addiction model)		
Overall addiction	–0.45 (–0.62 to –0.28)	<.001
Outcome: character and virtue		
Predictors (genre-specific model)		
Multiplayer online battle arena	–0.34 (–0.59 to –0.08)	.03
Shooting	–0.09 (–0.36 to 0.19)	.61
Casual	–0.18 (–0.60 to 0.25)	.51
Action and adventure	–0.74 (–1.06 to –0.43)	<.001
Sandbox and simulation	–0.51 (–1.09 to 0.08)	.18
Sports	–0.50 (–1.13 to 0.14)	.22
Strategy	–0.35 (–1.05 to 0.36)	.43
Role-playing	0.07 (–0.59 to 0.73)	.84
Predictor (overall addiction model)		
Overall addiction	–0.60 (–0.78 to –0.41)	<.001
Outcome: close social relationships		
Predictors (genre-specific model)		
Multiplayer online battle arena	–0.29 (–0.54 to –0.04)	.07
Shooting	0.15 (–0.11 to 0.41)	.36
Casual	–0.35 (–0.76 to 0.05)	.18
Action and adventure	–0.62 (–0.92 to –0.32)	<.001
Sandbox and simulation	–0.86 (–1.46 to –0.27)	.02
Sports	–0.15 (–0.75 to 0.46)	.72
Strategy	–0.22 (–0.90 to 0.46)	.61
Role-playing	0.27 (–0.37 to 0.92)	.51
Predictor (overall addiction model)		
Overall addiction	–0.35 (–0.53 to –0.17)	.001

[a]Game genres denote addiction to the corresponding genres.

[b]All models included sex, family economic status, and social media time as covariates; genre-specific models additionally controlled for addiction-to-other-games.

[c]*q* indicates that raw *P* values were adjusted using the Benjamini-Hochberg false discovery rate procedure.

XSL•FO
RenderX

Figure 2. Robust linear regression results showing the associations between gaming addiction and flourishing outcomes based on multiple imputed data (N=2194). HFI: Human Flourishing Index; MOBA: multiplayer online battle arena.

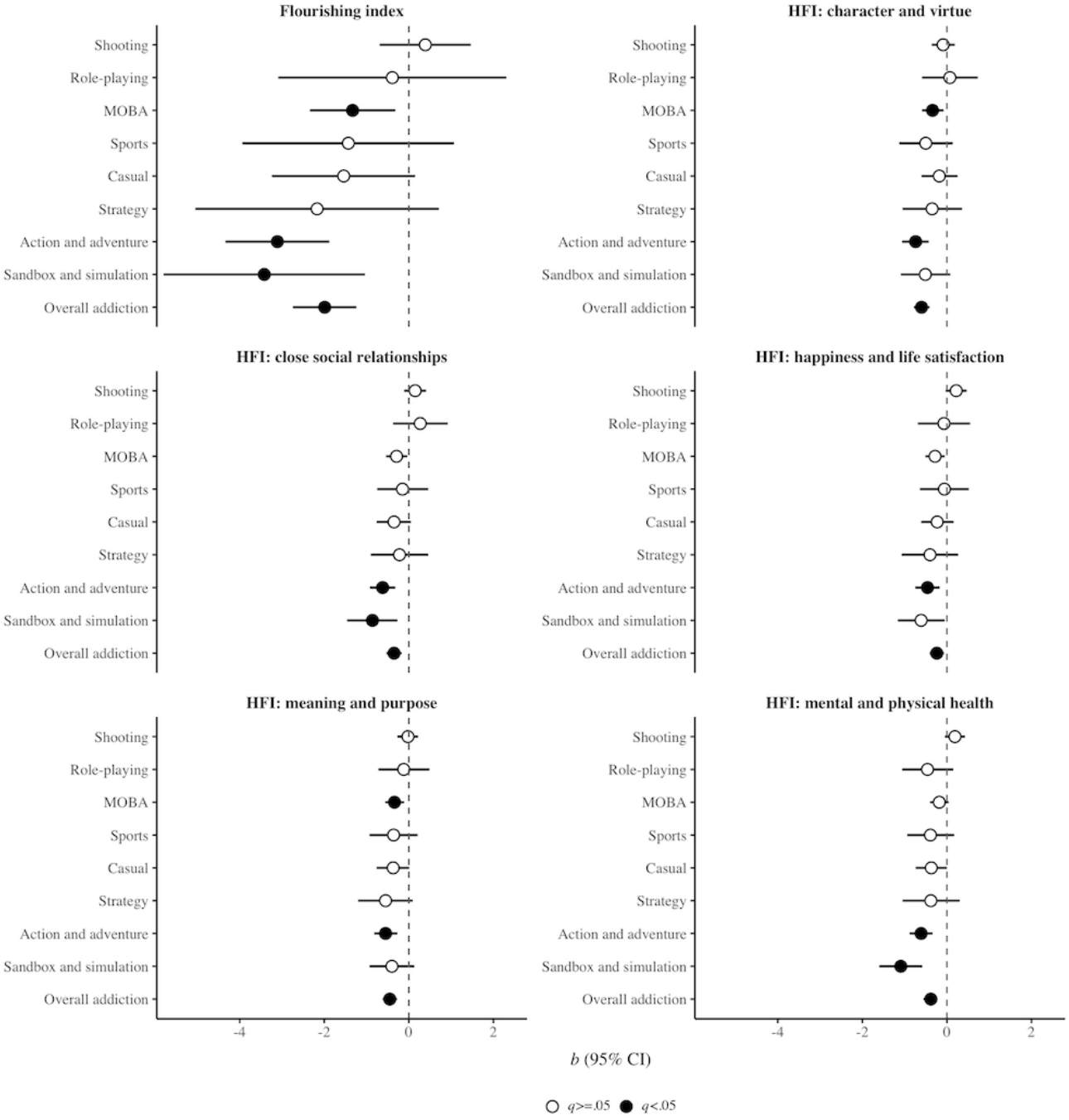

Although overall GA was negatively associated with all flourishing domains, the genre-specific findings indicate substantial heterogeneity in the addiction-flourishing relationship across game genres. The most prominent genre-specific predictor was AA addiction, which was associated with lower scores across all flourishing domains, suggesting that adolescents with AA addiction may constitute a particularly high-risk subgroup with broadly impaired well-being. Although both SS and MOBA addiction were associated with lower overall flourishing, their domain-specific patterns differed. The SS-flourishing association appeared to be driven primarily by poorer mental and physical health and weaker close social relationships, whereas the MOBA-flourishing association was driven mainly by lower meaning and purpose and character and virtue. The absence of significant associations for other genres (overall flourishing: q=.57, .17, .36, .22, and .84; happiness and life satisfaction: q=.18, .36, .84, .36, and .84; mental and physical health: q=.20, .12, .26, .36, and .22; meaning and purpose: q=.84, .13, .32, .18, and .78; character and virtue: q=.61, .51, .22, .43, and .84; and close social relationships: q=.36, .18, .72, .61, and .51, for shooting, casual, sports, strategy, and role-playing, respectively) suggests that addiction to these genres may be comparatively less detrimental to flourishing than addiction to AA, SS, and MOBA.

We conducted model diagnostics to evaluate regression assumptions. Multicollinearity among predictors was assessed using variance inflation factors. The maximal variance inflation factor across the 20 imputed datasets and all outcome variables (including the overall GA models) was 1.22, indicating no

XSL•FO
RenderX

problematic collinearity among predictors [76]. Visual inspection of residuals-versus-fitted plots across all imputed datasets and models indicated no substantial departures from linearity or other patterns suggesting model misspecification.

Multimedia Appendix 3 presents the results of the robust linear regression analyses examining the associations between GA and flourishing outcomes using complete-case data. The pattern of significant predictors was consistent with the analyses based on multiple imputed data for overall flourishing, mental and physical health, meaning and purpose, and character and virtue. Compared with the multiple imputation analyses, the effect of overall GA on happiness and life satisfaction became nonsignificant in the complete-case analysis (b=–0.22, 95% CI –0.41 to –0.03, q=.07), whereas the effect of MOBA on close social relationships became statistically significant in the complete-case analysis (b=–0.39, 95% CI –0.66 to –0.12, q=.02). Despite these discrepancies, the overall pattern of results was generally consistent across the 2 approaches, suggesting that the findings were largely robust to the handling of missing data.

Multimedia Appendix 4 presents the E values for the significant pooled effects from the robust linear regression analyses based on multiple imputed data. Across all flourishing outcomes, the E values for the regression coefficients ranged from 1.31 to 1.55, indicating moderate robustness to unmeasured confounding. This suggests that unmeasured confounder(s) would need to be associated with both GA and flourishing outcomes by at least a 1.3-fold risk ratio, beyond the measured covariates, to fully explain the observed effects. The E values for the CI limits closest to the null ranged from 1.14 to 1.43, with 7 values below 1.20. These findings indicate that relatively weak unmeasured confounder(s) could attenuate the statistical significance of some estimates. Therefore, certain genre-specific associations should be interpreted with caution, given their potential susceptibility to residual confounding.

Discussion

Principal Findings

Using VanderWeele's multidimensional flourishing framework, this study examined genre-specific associations between GA and flourishing among adolescents. Consistent with our primary hypothesis, the associations between GA and flourishing varied across game genres rather than being uniform. Addiction to AA, SS, and MOBA games showed significant negative associations with all or some flourishing outcomes, whereas addiction to other genres was not significantly associated with any flourishing outcomes. Consistent with our secondary hypothesis, overall GA was negatively associated with all flourishing outcomes.

Interpretations and Comparisons to Existing Literature

A substantial body of research has linked adolescents' GA to a range of health outcomes, including reduced life satisfaction [7], poorer health [7,17], lower meaning in life [35], lower empathy and moral levels [38,39], and difficulties in social relationships [14,26]. Consistent with this literature, this study found that overall GA was negatively associated with all 5 flourishing subdomains (ie, happiness and life satisfaction, mental and physical health, meaning and purpose, character and virtue, and close social relationships). In addition, our findings extend prior work by showing that the negative correlates of GA are not confined to isolated facets of well-being but instead span the full spectrum of adolescents' flourishing. This suggests that addictive gaming may constrain adolescents' capacity to thrive across multiple life domains. Moreover, our results indicate that the adverse association between GA and flourishing is not uniform across game genres. While addiction to certain genres was negatively related to flourishing, addiction to others was not associated with any flourishing outcomes.

The genre-specific analyses revealed that AA addiction was the only consistent negative predictor of overall flourishing and all 5 of its domains. Several mechanisms may explain these findings. First, AA games possess design features that may heighten the risk of reduced flourishing. Compared with other genres, AA games often feature open-world exploration and story-driven adventures that lack natural stopping points [77]. Such features can promote prolonged, uninterrupted play, frequently in solitary contexts, which may displace real-life activities essential for flourishing [78,79]. Second, AA games commonly include fast-paced, skill-based activities (eg, defeating enemies and solving puzzles) that produce rapid cycles of in-game achievement [80]. Although such virtual accomplishments may temporarily "scratch the itch" for success [81], prior research suggests that they function more as compensators than as substitutes for real-world competence [82]. Thus, adolescents with addictive gaming patterns may feel competent during gameplay, yet their underlying need for real-world competence remains unmet, potentially leading to reduced well-being [83]. Third, adolescents with preexisting low levels of flourishing may be particularly drawn to AA games. Those with low life satisfaction, weak social relationships, poor emotional well-being, or a lack of life purpose may find AA games especially appealing because they offer immersive narratives, clear missions, and rapid reward cycles that provide a temporary sense of escape, competence, and progression [84]. When real life feels unsatisfying, completing quests, advancing storylines, and gaining in-game achievements can serve as substitutes for the meaning and accomplishments they struggle to experience in daily life, making these games particularly attractive [85].

Higher MOBA addiction was associated with poorer character and virtue. This relationship may be explained by the highly competitive and performance-driven nature of these games, which emphasize winning, dominance, and skill mastery rather than empathy or moral reflection [86]. Prior research has shown that many MOBA players experience in-game hostility and toxic communication [87], which may erode prosocial tendencies and empathy—core elements of character and virtue. Higher MOBA addiction was also associated with a lower sense of meaning and purpose. Similar to AA games, MOBA games provide clear goals (eg, destroying the enemy nexus) and immediate rewards (eg, rank increases after winning) [88], which are often more attainable than real-life achievements. When adolescents repeatedly turn to these games for accomplishment, real-world goals such as academic progress

XSL•FO
RenderX

or personal growth may feel less rewarding or more difficult to pursue [82], leading to fewer meaningful experiences and a weakened sense of purpose [89]. It is also plausible that adolescents who already exhibit lower character and virtue or a reduced sense of meaning and purpose are more vulnerable to MOBA addiction, as these game-design features may temporarily compensate for these deficits [90].

Higher SS addiction was associated with fewer close social relationships and poorer mental and physical health. Several mechanisms may account for these associations. First, SS games are typically structured around solitary play, allowing players to advance independently, in contrast to team-based genres that require interpersonal coordination [91]. Such solitary engagement may displace real-world interaction, which is essential for building and maintaining close social relationships [92]. Second, SS games provide a strong sense of autonomy and control (eg, players can design cities, manage resources, and make strategic decisions about budgets and policies) [93], which may be particularly appealing to adolescents who feel powerless in real life due to limited social support [94,95]. Third, adolescents with poorer mental health resulting from real-world challenges may be prone to playing these games, as they allow them to simulate idealized achievements and life scenarios (eg, building perfect cities, running successful businesses, or creating ideal families) [96]. These experiences may temporarily compensate for dissatisfaction with real-life circumstances [82].

The absence of associations between addiction to shooting, sports, casual, strategy, or role-playing games and flourishing may be explained by several factors. First, these genres are highly heterogeneous in their modes of play [97-100]. For example, shooting games may involve competitive player-versus-player matches, cooperative missions against computer-controlled opponents, or single-player narrative progression [100]. Sports games can be played through team-on-team matches or career modes that simulate long-term athlete development [100]. Casual, strategy, and role-playing games can be played in either solitary or highly social modes [97-99]. This variability means that potential positive and negative associations with flourishing at the game level may offset one another at the genre level. Second, many games in these genres feature short matches or rounds, which may make disengagement easier compared with genres that lack natural stopping points [101]. This, in turn, reduces the likelihood of prolonged immersion that displaces real-life activities essential for flourishing [102,103].

Implications

The findings of this study have several practical implications. First, higher addiction to shooting, sports, casual, strategy, or role-playing games was not associated with poorer flourishing outcomes. This suggests that encouraging adolescents to spend their gaming time within these relatively lower-risk genres may not increase the likelihood of adverse outcomes associated with GA. This does not imply that addiction to such games is harmless. Rather, when gaming is already part of adolescents' lives, guiding them toward lower-risk genres may represent a more realistic harm-reduction strategy. Second, when adolescents are highly engaged in AA, MOBA, or SS games, targeted strategies may be required to reduce addiction-related risks. For parents, setting clear time limits and enforcing mandatory breaks may help prevent prolonged, immersive play. Conversations that unpack why certain games feel particularly absorbing may also increase adolescents' awareness of risky game design features, thereby empowering them to better self-regulate their play. Providing alternative activities that offer similar psychological rewards may also be effective. For example, adolescents drawn to SS games may find comparable satisfaction in robotics clubs, art workshops, or sandbox-style board games, which can fulfill needs for autonomy and creativity. MOBA players may benefit from real-world team sports, which can satisfy desires for competition, collaboration, and skill mastery. Third, for policy makers, rather than imposing one-size-fits-all rules, policies could be more targeted toward high-risk game genres. For instance, China's current "Online Game Anti-Indulgence System" restricts adolescents to a single hour of play on weekend evenings across all games (although many adolescents have learned ways to circumvent this rule). A more nuanced policy could impose stricter limits on high-risk genres while allowing more time for genres associated with fewer negative health outcomes. Such differentiation could better balance adolescents' enjoyment of gaming with the goal of minimizing potential harm to their well-being.

Limitations

Several limitations should be considered when interpreting the findings. First, the results may be influenced by the genre classification approach adopted in this study. Although the categorization was informed by established frameworks, substantial variation may still exist among games within the same genre with respect to design features that could lead to distinct associations with health-related outcomes. Future research would benefit from developing or validating a genre classification system specifically optimized for health research. Such a system could reduce within-genre heterogeneity and facilitate more robust conclusions regarding genre-specific associations with adolescent health. Second, the cross-sectional design of this study limits causal inference between GA and flourishing. The *E* value analysis also suggested that some associations could potentially be explained by unmeasured confounders of small magnitude. To strengthen causal inference, future studies could employ longitudinal and experimental designs to clarify whether genre-specific GA undermines flourishing, whether adolescents with lower flourishing are more likely to develop addiction to particular genres, or whether these processes operate reciprocally. Third, due to practical constraints, GA was measured using single-item self-reports, which are inherently less precise than multi-item scales and may have introduced measurement error. Future studies should use validated multi-item measures when feasible.

Conclusions

This study provides novel evidence that the association between GA and adolescent flourishing varies by game genre. In contrast to prior work that treats health as a narrow or unidimensional construct, the adoption of a multidimensional health framework offers a more nuanced understanding of gaming-related risks. The findings contribute to the field by indicating that addiction

XSL•FO
RenderX

to AA, MOBA, and SS games poses greater risks to adolescent health than addiction to other genres. Accordingly, prevention, education, and policy efforts should prioritize these higher-risk genres to more effectively support adolescent health.

Acknowledgments

We gratefully acknowledge the teachers who contributed to data collection for this study. We thank the research assistants for their support in digitizing the paper questionnaires. We are also grateful to Dr. Dorota W ziak-Białowolska for providing the Chinese version of the Human Flourishing Index. The funder had no role in the study design, data collection, analysis, interpretation, or manuscript preparation. We declare the use of generative artificial intelligence in the research and writing process. According to GAIDeT (Generative AI Delegation Taxonomy) (2025), the following tasks were delegated to generative artificial intelligence tools under full human supervision: proofreading and editing. The generative artificial intelligence tool used was ChatGPT-5.0 (OpenAI). Responsibility for the final manuscript lies entirely with the authors. Generative artificial intelligence tools are not listed as authors and do not bear responsibility for the outcomes.

Funding

This work was supported by the Shanghai Jiao Tong University Young Scholars Program in Humanities (grant 2025QN029) and the Shanghai Jiao Tong University Research Start-up Fund (grant WH220525103).

Data Availability

The datasets generated or analyzed during this study are not publicly available because they will be used for secondary analyses in other ongoing studies, but are available from the corresponding author (YL) upon reasonable request.

Authors' Contributions

Conceptualization: Y Luo
Data curation: Y Luo, XC, Y Liu, WH, MF
Formal analysis: Y Luo
Funding acquisition: Y Luo
Investigation: Y Luo
Methodology: Y Luo
Project administration: Y Luo
Resources: Y Luo
Software: Y Luo
Supervision: Y Luo
Visualization: Y Luo
Writing – original draft: Y Luo
Writing – review & editing: Y Luo, XC, Y Liu, WH, MF, DT, SM

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey questionnaire.
[DOCX File , 37 KB - jmir_v28i1e89319_app1.docx]

Multimedia Appendix 2
Data processing.
[DOCX File , 22 KB - jmir_v28i1e89319_app2.docx]

Multimedia Appendix 3
Results of the robust linear regression analyses examining the associations between gaming addiction and flourishing outcomes using complete-case data.
[DOCX File , 22 KB - jmir_v28i1e89319_app3.docx]

Multimedia Appendix 4
E values for the significant pooled effects from the robust linear regression analyses based on multiple imputed data.

XSL•FO
RenderX

[DOCX File , 17 KB - jmir_v28i1e89319_app4.docx]

References

1. Granic I, Lobel A, Engels RCME. The benefits of playing video games. Am Psychol 2014 Jan;69(1):66-78. [doi: 10.1037/a0034857] [Medline: 24295515]
2. Gao Y, Wang J, Dong G. The prevalence and possible risk factors of internet gaming disorder among adolescents and young adults: systematic reviews and meta-analyses. Journal of Psychiatric Research 2022 Oct;154:35-43. [doi: 10.1016/j.jpsychires.2022.06.049]
3. Kim HS, Son G, Roh E, Ahn W, Kim J, Shin S, et al. Prevalence of gaming disorder: a meta-analysis. Addict Behav 2022 Mar;126:107183. [doi: 10.1016/j.addbeh.2021.107183] [Medline: 34864436]
4. Byeon G, Jo S, Park J, Jeong H, Lee HK, Yim HW. Risk factors and outcomes of internet gaming disorder identified in Korean prospective adolescent cohort study. J Behav Addict 2022 Dec 27;11(4):1035-1043 [FREE Full text] [doi: 10.1556/2006.2022.00071] [Medline: 36194504]
5. World Health Organization (WHO). Constitution of the World Health Organization. Am J Public Health Nations Health 1946 Nov;36(11):1315-1323. [doi: 10.2105/ajph.36.11.1315]
6. VanderWeele TJ. On the promotion of human flourishing. Proc Natl Acad Sci U S A 2017 Aug 01;114(31):8148-8156. [doi: 10.1073/pnas.1702996114] [Medline: 28705870]
7. van der Neut D, Peeters M, Boniel-Nissim M, Klanšček HJ, Oja L, van den Eijnden R. A cross-national comparison of problematic gaming behavior and well-being in adolescents. J Behav Addict 2023 Jun 29;12(2):448-457 [FREE Full text] [doi: 10.1556/2006.2023.00010] [Medline: 37060555]
8. Teng Z, Pontes HM, Nie Q, Xiang G, Griffiths MD, Guo C. Internet gaming disorder and psychosocial well-being: a longitudinal study of older-aged adolescents and emerging adults. Addict Behav 2020 Nov;110:106530. [doi: 10.1016/j.addbeh.2020.106530] [Medline: 32683173]
9. Arslan G, Jarden A, Yıldırım M, Jarden R, Silapurem L. Longitudinal associations between life events, hope, life satisfaction, happiness, and depressive symptoms. Cogn Ther Res 2025 Jul 09:e1 (forthcoming). [doi: 10.1007/s10608-025-10630-0]
10. Freire T, Ferreira G. Health-related quality of life of adolescents: relations with positive and negative psychological dimensions. International Journal of Adolescence and Youth 2016 Dec 23;23(1):11-24. [doi: 10.1080/02673843.2016.1262268]
11. Freire T, Ferreira G. Do I need to be positive to be happy? Considering the role of self-esteem, life satisfaction, and psychological distress in Portuguese adolescents' subjective happiness. Psychol Rep 2020 Aug 29;123(4):1064-1082. [doi: 10.1177/0033294119846064] [Medline: 31033417]
12. López-Pérez B, Fernández-Castilla B. Children's and adolescents' conceptions of happiness at school and its relation with their own happiness and their academic performance. J Happiness Stud 2017 Jun 7;19(6):1811-1830. [doi: 10.1007/s10902-017-9895-5]
13. Shin W, Jyung M, Choi I, Sul S. Perceived financial well-being and its association with frontostriatal functional connectivity, real-life anticipatory experiences, and everyday happiness. Sci Rep 2023 Oct 31;13(1):18739 [FREE Full text] [doi: 10.1038/s41598-023-44001-0] [Medline: 37907524]
14. Wartberg L, Kriston L, Kammerl R. Associations of social support, friends only known through the internet, and health-related quality of life with internet gaming disorder in adolescence. Cyberpsychol Behav Soc Netw 2017 Jul;20(7):436-441. [doi: 10.1089/cyber.2016.0535] [Medline: 28715266]
15. Zhang L, Han J, Liu M, Yang C, Liao Y. The prevalence and possible risk factors of gaming disorder among adolescents in China. BMC Psychiatry 2024 May 21;24(1):381 [FREE Full text] [doi: 10.1186/s12888-024-05826-9] [Medline: 38773555]
16. Liu Y, Gong R, Yu Y, Xu C, Yu X, Chang R, et al. Longitudinal predictors for incidence of internet gaming disorder among adolescents: the roles of time spent on gaming and depressive symptoms. Journal of Adolescence 2021 Oct;92(1):1-9. [doi: 10.1016/j.adolescence.2021.06.008]
17. Canale N, Elgar FJ, Pivetta E, Galeotti T, Marino C, Billieux J, et al. Problem gaming and adolescents' health and well-being: evidence from a large nationally representative sample in Italy. Computers in Human Behavior 2025 Jul;168:108644. [doi: 10.1016/j.chb.2025.108644]
18. Falcione K, Weber R. Psychopathology and gaming disorder in adolescents. JAMA Netw Open 2025 Jul 01;8(7):e2528532 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.28532] [Medline: 40728787]
19. Yu Y, Wu AMS, Fong VWI, Zhang J, Li J, Lau JTF. Association between internet gaming disorder and suicidal ideation mediated by psychosocial resources and psychosocial problems among adolescent internet gamers in China: cross-sectional study. JMIR Serious Games 2024 Sep 19;12:e48439 [FREE Full text] [doi: 10.2196/48439] [Medline: 39298753]
20. Ferreira FDM, Bambini B, Tonsig G, Fonseca L, Picon FA, Pan PM, et al. Predictors of gaming disorder in children and adolescents: a school-based study. Braz J Psychiatry 2021;43(3):289-292 [FREE Full text] [doi: 10.1590/1516-4446-2020-0964] [Medline: 32756806]
21. Coutelle R, Balzer J, Rolling J, Lalanne L. Problematic gaming, psychiatric comorbidities, and adolescence: a systematic review of the literature. Addict Behav 2024 Oct;157:108091. [doi: 10.1016/j.addbeh.2024.108091] [Medline: 38901145]

XSL•FO
RenderX

22. Meng Y, Shi X, Cai D, Ran M, Ye A, Qiu C. Prevalence, predictive factors, and impacts of internet gaming disorder among adolescents: a population-based longitudinal study. J Affect Disord 2024 Oct 01;362:356-362. [doi: 10.1016/j.jad.2024.06.020] [Medline: 38844168]
23. Rahul R, Kareem SA, Gowda KC, Sajna SJ. Impact of internet gaming disorder on mental well-being: prevalence and psychological distress among adolescents and young adults. Psychiatry Res 2025 Jul;349:116512. [doi: 10.1016/j.psychres.2025.116512] [Medline: 40318564]
24. Zhang W, Pu J, He R, Yu M, Xu L, He X, et al. Demographic characteristics, family environment and psychosocial factors affecting internet addiction in Chinese adolescents. J Affect Disord 2022 Oct 15;315:130-138. [doi: 10.1016/j.jad.2022.07.053] [Medline: 35901990]
25. Jeong EJ, Ferguson CJ, Lee SJ. Pathological gaming in young adolescents: a longitudinal study focused on academic stress and self-control in South Korea. J Youth Adolesc 2019 Dec 20;48(12):2333-2342. [doi: 10.1007/s10964-019-01065-4] [Medline: 31327099]
26. Mohamed NF, Ab Manan N, Muhammad Firdaus Chan MF, Rahmatullah B, Abd Wahab R, Baharudin SNA, et al. The prevalence of internet gaming disorders and the associated psychosocial risk factors among adolescents in Malaysian secondary schools. Clin Child Psychol Psychiatry 2023 Oct 17;28(4):1420-1434. [doi: 10.1177/13591045231164870] [Medline: 36927134]
27. Yu Y, Peng L, Mo PK, Yang X, Cai Y, Ma L, et al. Association between relationship adaptation and internet gaming disorder among first-year secondary school students in China: mediation effects via social support and loneliness. Addictive Behaviors 2022 Feb;125:107166. [doi: 10.1016/j.addbeh.2021.107166]
28. Deng X, Hu Y, Liu C, Li Q, Yang N, Zhang Q, et al. Psychological distress and aggression among adolescents with internet gaming disorder symptoms. Psychiatry Research 2024 Jan;331:115624. [doi: 10.1016/j.psychres.2023.115624]
29. Wartberg L, Kriston L, Kramer M, Schwedler A, Lincoln T, Kammerl R. Internet gaming disorder in early adolescence: associations with parental and adolescent mental health. Eur. psychiatr 2020 Mar 23;43:14-18. [doi: 10.1016/j.eurpsy.2016.12.013]
30. Yu Y, Yang X, Wang S, Wang H, Chang R, Tsamlag L, et al. Serial multiple mediation of the association between internet gaming disorder and suicidal ideation by insomnia and depression in adolescents in Shanghai, China. BMC Psychiatry 2020 Sep 23;20(1):460. [doi: 10.1186/s12888-020-02870-z] [Medline: 32967648]
31. Chamarro A, Díaz-Moreno A, Bonilla I, Cladellas R, Griffiths MD, Gómez-Romero MJ, et al. Stress and suicide risk among adolescents: the role of problematic internet use, gaming disorder and emotional regulation. BMC Public Health 2024 Jan 30;24(1):326. [doi: 10.1186/s12889-024-17860-z]
32. Moshel ML, Warburton W, Thomasius R, Paschke K. Sleep quality as a mediator of internet gaming disorder and executive dysfunction in adolescents: cross-sectional questionnaire study. J Med Internet Res 2025 Jul 09;27:e68571 [FREE Full text] [doi: 10.2196/68571] [Medline: 40632805]
33. Che Mokhtar M, McGee R. Impact of internet addiction and gaming disorder on body weight in children and adolescents: a systematic review. J Paediatrics Child Health 2024 Dec 06;61(2):136-147. [doi: 10.1111/jpc.16726]
34. Männikkö N, Ruotsalainen H, Tolvanen A, Kääriäinen M. Problematic gaming is associated with some health-related behaviors among Finnish vocational school students. Int J Ment Health Addiction 2019 May 17;18(4):993-1007. [doi: 10.1007/s11469-019-00100-6]
35. Zhao H, Li X, Zhou J, Nie Q, Zhou J. The relationship between bullying victimization and online game addiction among Chinese early adolescents: the potential role of meaning in life and gender differences. Children and Youth Services Review 2020 Sep;116:105261. [doi: 10.1016/j.childyouth.2020.105261]
36. Wu AMS, Lei LLM, Ku L. Psychological needs, purpose in life, and problem video game playing among Chinese young adults. Int J Psychol 2013 Apr 16;48(4):583-590. [doi: 10.1080/00207594.2012.658057] [Medline: 22506646]
37. Zhang MX, Wang X, Yu SM, Wu AM. Purpose in life, social support, and internet gaming disorder among Chinese university students: a 1-year follow-up study. Addict Behav 2019 Dec;99:106070. [doi: 10.1016/j.addbeh.2019.106070] [Medline: 31430620]
38. Verrastro V, Albanese CA, Ritella G, Gugliandolo MC, Cuzzocrea F. Empathy, social self-efficacy, problematic internet use, and problematic online gaming between early and late adolescence. Cyberpsychol Behav Soc Netw 2021 Dec 01;24(12):806-814. [doi: 10.1089/cyber.2021.0042] [Medline: 34515543]
39. Yildiz Durak H, KIDIMAN DEMİRHAN E, Citil M. Examining various risk factors as the predictors of gifted and non-gifted high school students' online game addiction. Computers & Education 2022 Feb;177:104378. [doi: 10.1016/j.compedu.2021.104378]
40. Li L, Liang J, Chen S. How autonomy support environment influences student online game addiction: the mediating roles of academic motivation and academic perseverance. IJMHP 2024;26(10):777-790. [doi: 10.32604/ijmhp.2024.056988]
41. Wartberg L, Kriston L, Thomasius R. Internet gaming disorder and problematic social media use in a representative sample of German adolescents: Prevalence estimates, comorbid depressive symptoms and related psychosocial aspects. Computers in Human Behavior 2020 Feb;103:31-36. [doi: 10.1016/j.chb.2019.09.014]
42. Rosales-Navarro C, Torres Pérez I. Parental behaviors and video game addiction in adolescents: a systematic review. Int J Ment Health Addiction 2025 Mar 27:e1 (forthcoming). [doi: 10.1007/s11469-025-01473-7]

XSL•FO
RenderX

43. Torres-Rodríguez A, Griffiths MD, Carbonell X, Oberst U. Internet gaming disorder in adolescence: psychological characteristics of a clinical sample. J Behav Addict 2018 Sep 01;7(3):707-718 [FREE Full text] [doi: 10.1556/2006.7.2018.75] [Medline: 30264606]
44. Jeong H, Yim HW, Lee S, Lee HK, Potenza MN, Lee H. Factors associated with severity, incidence or persistence of internet gaming disorder in children and adolescents: a 2 - year longitudinal study. Addiction 2021 Jan 28;116(7):1828-1838. [doi: 10.1111/add.15366]
45. Butler A, Patte KA, Ferro MA, Leatherdale ST. Interrelationships among depression, anxiety, flourishing, and cannabis use in youth. Addict Behav 2019 Feb;89:206-215. [doi: 10.1016/j.addbeh.2018.10.007] [Medline: 30321693]
46. Blachnio A, Przepiórka A, Gorbaniuk O, McNeill M, Bendayan R, Ben-Ezra M, et al. Lonely online: a social model of digital media addiction: a study in 21 countries. The Social Science Journal 2024 Jan 10;62(3):717-728. [doi: 10.1080/03623319.2023.2228439]
47. Okur S, Engin M, Kütük H, Satıcı SA. Cross-lagged relations between smartphone addiction and flourishing in adolescents. Personality and Individual Differences 2025 Apr;236:113008. [doi: 10.1016/j.paid.2024.113008]
48. Satici SA, Uysal R. Well-being and problematic Facebook use. Computers in Human Behavior 2015 Aug;49:185-190. [doi: 10.1016/j.chb.2015.03.005]
49. Uysal R. The predictive roles of social safeness and flourishing on problematic Facebook use. South African Journal of Psychology 2014 Nov 21;45(2):182-193. [doi: 10.1177/0081246314560010]
50. Diener E, Wirtz D, Tov W, Kim-Prieto C, Choi D, Oishi S, et al. New well-being measures: short scales to assess flourishing and positive and negative feelings. Soc Indic Res 2009 May 28;97(2):143-156. [doi: 10.1007/s11205-009-9493-y]
51. Nakagomi A, Ide K, Kondo K, Shiba K. Digital gaming and subsequent health and well-being among older adults: longitudinal outcome-wide analysis. J Med Internet Res 2025 Jan 27;27:e69080 [FREE Full text] [doi: 10.2196/69080] [Medline: 39869904]
52. Park JH, Han DH, Kim B, Cheong JH, Lee Y. Correlations among social anxiety, self-esteem, impulsivity, and game genre in patients with problematic online game playing. Psychiatry Investig 2016 May;13(3):297-304 [FREE Full text] [doi: 10.4306/pi.2016.13.3.297] [Medline: 27247595]
53. Fraser AM, Padilla-Walker LM, Coyne SM, Nelson LJ, Stockdale LA. Associations between violent video gaming, empathic concern, and prosocial behavior toward strangers, friends, and family members. J Youth Adolesc 2012 May 1;41(5):636-649. [doi: 10.1007/s10964-012-9742-2] [Medline: 22302216]
54. Greitemeyer T. The dark and bright side of video game consumption: effects of violent and prosocial video games. Curr Opin Psychol 2022 Aug;46:101326 [FREE Full text] [doi: 10.1016/j.copsyc.2022.101326] [Medline: 35366605]
55. Greitemeyer T, Osswald S. Effects of prosocial video games on prosocial behavior. J Pers Soc Psychol 2010 Feb;98(2):211-221. [doi: 10.1037/a0016997] [Medline: 20085396]
56. Whitaker JL, Bushman BJ. "Remain calm. Be kind." Effects of relaxing video games on aggressive and prosocial behavior. Social Psychological and Personality Science 2011 May 23;3(1):88-92. [doi: 10.1177/1948550611409760]
57. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: the APA Publications and Communications Board Task Force report. Am Psychol 2018 Jan;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
58. Faul F, Erdfelder E, Buchner A, Lang A. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
59. Lemmens JS, Valkenburg PM, Peter J. Development and validation of a game addiction scale for adolescents. Media Psychology 2009 Mar 05;12(1):77-95. [doi: 10.1080/15213260802669458]
60. Lemmens JS, Valkenburg PM, Gentile DA. The Internet Gaming Disorder Scale. Psychol Assess 2015 Jun;27(2):567-582. [doi: 10.1037/pas0000062] [Medline: 25558970]
61. King DL, Delfabbro PH, Griffiths MD. Trajectories of problem video gaming among adult regular gamers: an 18-month longitudinal study. Cyberpsychol Behav Soc Netw 2013 Jan;16(1):72-76. [doi: 10.1089/cyber.2012.0062] [Medline: 23098213]
62. Najavits L, Lung J, Froias A, Paull N, Bailey G. A study of multiple behavioral addictions in a substance abuse sample. Subst Use Misuse 2014 Mar 04;49(4):479-484. [doi: 10.3109/10826084.2013.858168] [Medline: 24304172]
63. Vogel EA, Prochaska JJ, Rubinstein ML. Measuring e-cigarette addiction among adolescents. Tob Control 2020 May 11;29(3):258-262 [FREE Full text] [doi: 10.1136/tobaccocontrol-2018-054900] [Medline: 31079033]
64. van den Eijnden RJ, Lemmens JS, Valkenburg PM. The Social Media Disorder Scale. Computers in Human Behavior 2016 Aug;61:478-487. [doi: 10.1016/j.chb.2016.03.038]
65. W ziak-Białowolska D, McNeely E, VanderWeele TJ. Human flourishing in cross cultural settings. Evidence from the United States, China, Sri Lanka, Cambodia, and Mexico. Front Psychol 2019 May 29;10:1269 [FREE Full text] [doi: 10.3389/fpsyg.2019.01269] [Medline: 31191421]
66. Hyun GJ, Han DH, Lee YS, Kang KD, Yoo SK, Chung U, et al. Risk factors associated with online game addiction: a hierarchical model. Computers in Human Behavior 2015 Jul;48:706-713. [doi: 10.1016/j.chb.2015.02.008]

XSL•FO
RenderX

67. Moreno M, Riddle K, Jenkins MC, Singh AP, Zhao Q, Eickhoff J. Measuring problematic internet use, internet gaming disorder, and social media addiction in young adults: cross-sectional survey study. JMIR Public Health Surveill 2022 Jan 27;8(1):e27719 [FREE Full text] [doi: 10.2196/27719] [Medline: 34081596]
68. Toker S, Baturay MH. Antecedents and consequences of game addiction. Computers in Human Behavior 2016 Feb;55:668-679. [doi: 10.1016/j.chb.2015.10.002]
69. VanderWeele TJ. Principles of confounder selection. Eur J Epidemiol 2019 Mar 6;34(3):211-219 [FREE Full text] [doi: 10.1007/s10654-019-00494-6] [Medline: 30840181]
70. R Core Team. R: a language and environment for statistical computing. R Foundation. URL: https://www.R-project.org [accessed 2026-01-02]
71. Maechler M, Rousseeuw P, Croux C. Package 'robustbase'. R Foundation. 2021. URL: https://cran.r-project.org/web/packages/robustbase/index.html [accessed 2026-02-09]
72. Mathur MB, Ding P, Riddell CA, VanderWeele TJ. Web site and R package for computing E-values. Epidemiology 2018;29(5):e45-e47. [doi: 10.1097/ede.0000000000000864]
73. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. Ann Intern Med 2017 Aug 15;167(4):268-274. [doi: 10.7326/M16-2607] [Medline: 28693043]
74. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. BMC Med Res Methodol 2017 Dec 06;17(1):162 [FREE Full text] [doi: 10.1186/s12874-017-0442-1] [Medline: 29207961]
75. Van BS, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. Journal of Statistical Software 2011;45:1-67 [FREE Full text] [doi: 10.18637/jss.v045.i03]
76. James G, Witten D, Hastie T, Tibshirani R, Taylor J. An Introduction to Statistical Learning: With Applications in Python. Cham, Switzerland: Springer; 2023.
77. Hughes N. Understanding specific gaming experiences: the case of open world games. Doctor of Philosophy thesis. University of York. White Rose eThesis Online. 2023. URL: https://etheses.whiterose.ac.uk/id/eprint/33608/1/Hughes_CorrectedThesisClean.pdf [accessed 2025-11-24]
78. Buecker S, Simacek T, Ingwersen B, Terwiel S, Simonsmeier BA. Physical activity and subjective well-being in healthy individuals: a meta-analytic review. Health Psychol Rev 2021 Dec 10;15(4):574-592. [doi: 10.1080/17437199.2020.1760728] [Medline: 32452716]
79. Kroencke L, Harari GM, Back MD, Wagner J. Well-being in social interactions: examining personality-situation dynamics in face-to-face and computer-mediated communication. J Pers Soc Psychol 2023 Feb;124(2):437-460. [doi: 10.1037/pspp0000422] [Medline: 35834202]
80. Imbert D. What makes video games engaging: deep dive into the design, mechanics and psychology of games. Bachelor's thesis. University of Applied Sciences. Theseus. URL: https://www.theseus.fi/bitstream/handle/10024/797212/Daniel_Imbert.pdf?sequence=2&isAllowed=y [accessed 2025-11-24]
81. Rieger D, Wulf T, Kneer J, Frischlich L, Bente G. The winner takes it all: the effect of in-game success and need satisfaction on mood repair and enjoyment. Computers in Human Behavior 2014 Oct;39:281-286. [doi: 10.1016/j.chb.2014.07.037]
82. Chang S, Hsieh GM, Lin SS. The mediation effects of gaming motives between game involvement and problematic internet use: escapism, advancement and socializing. Computers & Education 2018 Jul;122:43-53. [doi: 10.1016/j.compedu.2018.03.007]
83. Allen JJ, Anderson CA. Satisfaction and frustration of basic psychological needs in the real world and in video games predict internet gaming disorder scores and well-being. Computers in Human Behavior 2018 Jul;84:220-229. [doi: 10.1016/j.chb.2018.02.034]
84. Anto A, Basu A, Selim R, Foscht T, Eisingerich AB. Open-world games' affordance of cognitive escapism, relaxation, and mental well-being among postgraduate students: mixed methods study. J Med Internet Res 2024 Dec 17;26:e63760 [FREE Full text] [doi: 10.2196/63760] [Medline: 39689301]
85. Bender PK, Gentile DA. Internet gaming disorder: relations between needs satisfaction in-game and in life in general. Psychology of Popular Media 2020 Apr;9(2):266-278. [doi: 10.1037/ppm0000227]
86. Xia T, Lin X, Mo X, Su Q, Ding S. Players' continuous willingness to play in MOBA game ranking mode: through the lens of self-determination theory and social comparison theory. Humanit Soc Sci Commun 2024 Oct 22;11(1):1398. [doi: 10.1057/s41599-024-03934-1]
87. Kordyaka B, Laato S, Weber S, Niehaves B. What constitutes victims of toxicity - identifying drivers of toxic victimhood in multiplayer online battle arena games. Front Psychol 2023 Jun 16;14:1193172 [FREE Full text] [doi: 10.3389/fpsyg.2023.1193172] [Medline: 37397342]
88. Morris R. MOBA games: a critical analysis into what makes and breaks them, and what influences players to keep returning. BSc Games Programming thesis. University of London. ResearchGate. 2019 May. URL: https://www.researchgate.net/profile/Rees-Morris/publication/344483735_MOBA_Games_A_Critical_Analysis_Into_What_Makes_And_Breaks_Them_And_What_Influences_Players_to_Keep_Returning/links/5f7b9fd5a6fdccfd7b4a6ee9/

XSL•FO
RenderX

MOBA-Games-A-Critical-Analysis-Into-What-Makes-And-Breaks-Them-And-What-Influences-Players-to-Keep-Returning.pdf [accessed 2025-11-24]
89. Bu L, Ding B. A comparative study of online gaming between high-performing and low-performing students-a case from China. Heliyon 2024 Jan 30;10(2):e24291 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e24291] [Medline: 38293356]
90. T'ng ST, Ho KH, Pau K. Need frustration, gaming motives, and internet gaming disorder in mobile multiplayer online battle arena (MOBA) games: through the lens of self-determination theory. Int J Ment Health Addict 2022 Apr 25;21(6):1-21 [FREE Full text] [doi: 10.1007/s11469-022-00825-x] [Medline: 35497075]
91. Hooper J. Freeplay in videogames. CORE. URL: https://files01.core.ac.uk/download/pdf/196607253.pdf [accessed 2025-11-24]
92. Pinker S. The Village Effect: How Face-to-Face Contact Can Make Us Healthier and Happier. Toronto, ON, Canada: Random House Canada; 2015.
93. Zhong Y, Fryer LK, Zheng S, Shum A, Chu SKW. The power of play: integrating competitive sandbox game for experiential learning to foster twenty-first century skills. Int J Educ Technol High Educ 2025 Jun 06;22(1):34. [doi: 10.1186/s41239-025-00528-y]
94. An G, Kim M. Powerlessness, social support, and glycemic control in Korean adults with type 2 diabetes. Contemporary Nurse 2014 Dec 17;42(2):272-279. [doi: 10.5172/conu.2012.42.2.272] [Medline: 23181377]
95. Lammers J, Stoker JI, Rink F, Galinsky AD. To have control over or to be free from others? The desire for power reflects a need for autonomy. Pers Soc Psychol Bull 2016 Apr 16;42(4):498-512. [doi: 10.1177/0146167216634064] [Medline: 26984014]
96. Rahimi S, Walker JT, Lin-Lipsmeyer L, Shin J. Toward defining and assessing creativity in sandbox games. Creativity Research Journal 2023 Jan 03;36(2):194-212. [doi: 10.1080/10400419.2022.2156477]
97. Nielsen JL, Jensen BF, Mahlmann T, Togelius J, Yannakakis GN. AI for general strategy game playing. In: Handbook of Digital Games. Hoboken, NJ: Wiley; 2014:13-46.
98. Deterding S, Zagal J. Role-Playing Game Studies: Transmedia Foundations. New York, NY: Routledge; 2018.
99. Johnson MR. Casual games before casual games: historicizing paper puzzle games in an era of digital play. Games and Culture 2018 Aug 01;14(2):119-138. [doi: 10.1177/1555412018790423]
100. Wolf MJP, Perron B. The Routledge Companion to Video Game Studies. Abingdon, UK: Taylor & Francis; 2023.
101. Alexandrovsky D, Gerling K, Opp MS, Hahn CB, Birk MV, Alsheail M. Disengagement from games: characterizing the experience and process of exiting play sessions. Proc ACM Hum-Comput Interact 2024 Oct 15;8(CHI PLAY):1-27. [doi: 10.1145/3677066]
102. Stieger S, Lewetz D, Willinger D. Face-to-face more important than digital communication for mental health during the pandemic. Sci Rep 2023 May 17;13(1):8022 [FREE Full text] [doi: 10.1038/s41598-023-34957-4] [Medline: 37198196]
103. Lee PSN, Leung L, Lo V, Xiong C, Wu T. Internet communication versus face-to-face interaction in quality of life. Soc Indic Res 2010 Oct 31;100(3):375-389. [doi: 10.1007/s11205-010-9618-3]

Abbreviations

AA: action and adventure
GA: gaming addiction
HFI: Human Flourishing Index
JARS-Quant: Journal Article Reporting Standards for Quantitative Research
MOBA: multiplayer online battle arena
SS: sandbox and simulation

Edited by S Brini; submitted 10.Dec.2025; peer-reviewed by PS Liao; comments to author 13.Jan.2026; revised version received 03.Feb.2026; accepted 04.Feb.2026; published 12.Feb.2026.

Please cite as:
Luo Y, Chen X, Liu Y, Hong W, Fan M, Thomas D, Mullen SP
Genre-Specific Gaming Addiction and Flourishing in Adolescents: Cross-Sectional Survey Study
J Med Internet Res 2026;28:e89319
URL: https://www.jmir.org/2026/1/e89319
doi:10.2196/89319
PMID:

©Yan Luo, Xiaofei Chen, Yaoyu Liu, Wanjia Hong, Miner Fan, David Thomas, Sean P Mullen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Feb.2026. This is an open-access article distributed under the

XSL•FO
RenderX

terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Implementation of Reference Framework for Preventive Care for Older Adults: Cross-sectional Qualitative Study

Claire Chenwen Zhong[1,2]; Mingtao Chen[1]; Chung Yi Lo[1]; Man Kin YIM[1]; Xiaoshu Zhang[1]; William CW Wong[3]; Junjie Huang[1,2]; Martin CS Wong[1,2]

[1]JC School of Public Health and Primary Care, Faculty of Medicine, Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

[2]Centre for Health Education and Health Promotion, Faculty of Medicine, Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

[3]Department of Family Medicine & Primary Care, LKS Faculty of Medicine, University of Hong Kong, Hong Kong, China (Hong Kong)

Corresponding Author:
Martin CS Wong
JC School of Public Health and Primary Care
Faculty of Medicine
Chinese University of Hong Kong
4/F, JC School of Public Health
Prince of Wales Hospital
Hong Kong,
China (Hong Kong)
Phone: 852 22528782
Email: wong_martin@cuhk.edu.hk

Abstract

Background: The Hong Kong Government developed the Reference Framework for Preventive Care for Older Adults (RF) to provide evidence-based recommendations for preventive care in the primary care setting. However, no study has systematically evaluated the adoption and awareness of this framework among primary care physicians (PCPs).

Objective: This study aimed to explore PCPs' perceptions of the RF and identify barriers and facilitators to its adoption in primary care settings.

Methods: This cross-sectional qualitative study involved interviews with 40 PCPs in Hong Kong to assess their views on the RF's adoption and implementation. The Consolidated Framework for Implementation Research (CFIR) was used to guide the analysis, focusing on five domains: intervention characteristics, outer setting, inner setting, individual characteristics, and implementation process.

Results: Relative advantage (65%, n = 26), innovation design (45%, n = 18), and access to knowledge and information (70%, n = 28) were the facilitators that mostly discussed by interviewee. By contrast, interviewees also raised concerns regarding policy and legal (n=29, 73%) and availability of resources (n=29, 73%) in the implementation of the RF. In general, participants acknowledged the RF's evidence-based foundation and comprehensive design, appreciating its potential to improve care for older adults. However, barriers included the complexity and limited personalization of the RF, competing clinical priorities, time constraints, and resource limitations. Government support and policy initiatives facilitated engagement, but clearer integration into existing workflows and stronger promotion were needed. Tailored updates, user feedback, and technological enhancements were seen as essential for improving the RF's usability and ensuring its relevance in clinical practice.

Conclusions: This study highlights that while the RF has potential to improve preventive care in Hong Kong's primary care setting, its adoption is constrained by systemic, organizational, and individual barriers. To ensure its successful integration, flexible implementation, institutional support, tailored incentives, and enhanced clinician and patient engagement are essential. These findings offer practical implications for policymakers and healthcare practitioners to refine and promote the RF in routine consultations, and underscore the need for future research to test theory-informed implementation strategies. Furthermore, this study offered novel contributions to the CFIR literatures in systematically investigate physician perspectives on RF for the older adult population within the distinct socio-cultural and structural context of Hong Kong, COVID-19 pandemic impact to the current healthcare system and actionable implementation strategies for Asian populations. Applying the findings from this study, the CFIR-ERIC Matching Tool could be used to address the specific barriers identified in this study and improving implementation outcomes across various healthcare settings.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e78814) doi:10.2196/78814

KEYWORDS

primary care; guideline adherence; geriatrics; implementation science; Consolidated Framework for Implementation Research

Introduction

The Hong Kong population has encountered a serious aging burden, with a higher old-age dependency ratio and old-age population proportion than other East Asian districts such as Macau and Singapore [1]. It has been estimated that the proportion of residents aged 65 years or older will increase from 1.45 million (20.5%) in 2021 to more than 2.74 million (36%) in 2046 [2,3]. Apart from implications for policy and the provision of social services, this increase could exert a substantial influence on the demand for health care services, including a higher population prevalence of multimorbidity [4]. A systematic review reported that multimorbidity has affected 76.5% of Chinese older adults [5], and its disease burden continues to escalate in aging communities [6]. The challenge posed by these observations is compounded by the increasing prevalence of disabilities, functional decline, cognitive impairment, mood problems, and lack of social support among the older adult population [7], imposing a substantial burden on patients, their caregivers, and the health care system.

A significant body of evidence shows that adherence to guideline-directed preventive interventions can be translated into clinical benefits [8,9]. In addition, evidence-based guidelines can improve public health outcomes by informing physicians, allied health professionals, and policy makers to practice and design tailored health programs in the most effective way.

The Primary Healthcare Office of the Department of Health, the Hong Kong Government, has produced a Reference Framework for Preventive Care for Older Adults (hereafter "RF"), dedicated to providing recommendations for evidence-based preventive care for the older adult population in the primary care setting. A Task Force on Conceptual Model and Preventive Protocols (hereafter referred to as the "Task Force") was established under the Working Group on Primary Care [10]. The Task Force synthesized recommendations based on the latest evidence for adoption in primary care settings in Hong Kong. A Clinical Advisory Group, consisting of academic professors, stakeholders from professional organizations, experts from the primary care sector, and advocacy groups, was assembled to contribute to and support guideline development. One reference framework is dedicated specifically to preventive care for older adults [7]. The Task Force is charged with promulgating, maintaining, and revising the framework content and promoting its adoption by primary care physicians (PCPs). The RF contains a core document accompanied by 6 modules on specific areas, including (1) health assessment; (2) falls in older adults; (3) dental health care; (4) visual impairment; (5) mental health problems; and (6) cognitive impairment. The RF adopts a life-course approach and was developed with reference to 2 existing reference frameworks on diabetes and hypertension [11,12]. The RF also highlights key success factors for implementation, including adaptability to local environments, needs, and structures. It was anticipated that the adoption of the RF would improve patient care through health promotion, disease prevention, and addressing major health risks among older adults [7].

It is crucial to devise and implement strategies that could enhance health promotion, disease prevention, and preservation of functional ability among older adults [10]. Active aging has been recognized as multidimensional, encompassing a low risk of disease, high functional capacity, and active engagement in one's life [13]. With increasing age, lifestyle factors become more important determinants of active aging. Hence, engagement of older adults in disease-preventive activities is of paramount importance.

PCPs are in a privileged position to act as a first point of contact, providing comprehensive, continuous, and community-oriented care in a coordinated manner. They offer health education, risk assessment, and regular care for common chronic diseases [14]. These roles also include support for family members and caregivers, as well as referral to a specialist and community care when applicable. The effectiveness of disease prevention in preserving functional ability among the older adult population has been supported by a substantial body of evidence. For instance, a systematic review of randomized controlled trials summarizing the impact of interventions aimed at preventing frailty progression showed that group exercise programs and cognitive training could successfully reduce or delay frailty [15]. These preventive programs were also found to exert favorable effects on frailty indicators with good value for money, especially among very frail community-dwelling participants.

Several implementation science models are available. The Consolidated Framework for Implementation Research (CFIR) was particularly well-suited for this study because of its comprehensive and flexible nature, which allows for exploration of multiple factors influencing the adoption and implementation of the RF. Unlike other implementation science models that focus on specific domains or stages of implementation—for instance, the Theoretical Domains Framework, which emphasizes individual behavior change, or the Reach, Effectiveness, Adoption, Implementation, and Maintenance (RE-AIM) framework, which focuses on reach and effectiveness [16,17]—CFIR provides a holistic approach by integrating 5 key domains: Intervention Characteristics, Outer Setting, Inner Setting, Characteristics of Individuals, and the Implementation Process [18]. This breadth enables CFIR to capture the dynamic interplay among the RF's design, the health care system's infrastructure, external influences, and individual practitioner behaviors. Its adaptability to varying contexts makes CFIR especially relevant for understanding barriers and facilitators to implementing a complex intervention such as the RF in diverse health care settings [18], where factors including resource constraints, time limitations, and local needs are critical. Thus, CFIR's ability to identify patterns across these

XSL•FO
RenderX

domains provides actionable insights that are essential for improving the RF's adoption and impact.

To our knowledge, no study has systematically assessed this RF or examined the level of awareness and adoption of the RF among PCPs, as well as perceived facilitators and barriers to its implementation. This study aims to explore the implementation of the RF in Hong Kong's primary care settings, using the CFIR to identify key facilitators and barriers.

Methods

Study Design

The study employed a cross-sectional qualitative design involving in-depth individual interviews with 40 PCPs. A sample size of 10-50 participants has been recommended in previous qualitative studies, or until data saturation is reached [19-22], to ensure a comprehensive understanding of the RF. This study was designed, conducted, and reported in accordance with the Standards for Reporting Qualitative Research (SRQR; see Multimedia Appendix 2) [23].

This study was conducted within a constructivist paradigm, which posits that knowledge is socially constructed through interactions between individuals and shaped by context. To explore the implementation of the RF, we employed a 2-stage qualitative analysis. First, we conducted a thematic analysis to inductively identify patterns and emergent themes from the interview data. Subsequently, we applied a framework analysis guided by the CFIR to systematically organize and interpret the data within a structured implementation science framework. This approach allowed for both inductive exploration and deductive mapping of findings to established constructs.

Participant Recruitment

Participants were recruited between February and September 2024 through emails or phone calls. Eligible participants were (1) PCPs who provide preventive care to older adults as part of their routine clinical practice and (2) those who had read the core document and relevant modules of the RF before the interviews. PCPs who did not consult older adult patients in routine practice were excluded. No restrictions were placed on years of clinical experience to ensure a broad range of viewpoints. Eight distinct groups of PCPs were invited to share their experiences with the adoption of the RF. These groups included (1) PCPs employed by health maintenance organizations; (2) PCPs in group practices within the private sector but not affiliated with health maintenance organizations; (3) solo practitioners providing care independently in private clinics; (4) general outpatient clinic physicians working in publicly funded clinics; (5) physicians at Family Medicine Specialist Clinics; (6) Fellows of the Hong Kong College of Family Physicians who are actively involved in teaching and provide preventive care for older adults; (7) PCPs serving as referral partners for District Health Centers; and (8) physicians with formal qualifications in gerontology or geriatric medicine recognized by the Medical Council of Hong Kong. A purposive sampling method was employed to recruit participants from the 8 predefined PCP groups, ensuring comprehensive representation across diverse practice settings.

For groups 1-3, participants were recruited through established professional networks under the Hong Kong Medical Association. For groups 4 and 5, we contacted the Chiefs of Service across all 7 Hospital Authority clusters, with whom we have long-standing research and teaching collaborations, to nominate suitable participants. Participants in group 6 were identified in collaboration with the Fellows of the Hong Kong College of Family Physicians. For groups 7 and 8, eligible physicians were selected based on referral network lists published by the Health Bureau of the Hong Kong Special Administrative Region Government.

Data Collection

All interviews were conducted via Zoom (Zoom Communications). A semistructured interview guide (Multimedia Appendix 1) was used to explore PCPs' awareness of, attitudes toward, and perceived barriers to and enabling factors for use of the RF. Before formal data collection, the guide and the moderator's manual were pilot-tested with a small sample of PCPs from diverse practice settings. Feedback from the pilot informed revisions to improve the clarity and contextual relevance of the questions. During the interviews, the final version of the guide was used, with additional probing questions introduced as new topics emerged. All interviews were conducted by 2 trained research assistants (CYL and XZ)—one serving as the facilitator and the other as an observer—both of whom had prior experience in qualitative data collection.

All interviews were audio-recorded and transcribed verbatim in Chinese. A standardized transcription protocol was established to guide and ensure consistency throughout the transcription process. Transcribers received training and were supervised by an experienced researcher, who also validated the transcripts to ensure accuracy and reliability. Transcripts were translated into English only after completion of data analysis to preserve the original meanings and avoid introducing interpretive bias during coding. Interviewee transcript review was not conducted, as evidence from previous studies indicates that this practice may have a limited impact on transcript accuracy [24]. However, each transcript was independently reviewed by an experienced researcher (CCZ) to minimize potential bias and misinterpretation. The research team further reviewed all transcripts to ensure both linguistic and conceptual accuracy.

Data Analysis

All data were analyzed concurrently and sequentially using NVivo 12 software (QSR International). Data analysis was conducted in 2 phases using thematic analysis based on the original Chinese transcripts. This approach was adopted to avoid potential misinterpretations that could arise from premature translation. Quotations were translated into English only after completion of data analysis to ensure that original meanings were preserved. The accuracy of the translations was subsequently checked and confirmed by a senior researcher (CCZ). In the first phase of analysis, inductive coding was performed to identify key patterns and themes emerging from the data. In the second phase, these preliminary themes were systematically reexamined using the CFIR as a guiding framework. The CFIR comprises 38 constructs across 5

XSL•FO
RenderX

domains—Innovation, Inner Setting, Outer Setting, Individual Characteristics, and Implementation Process—which may act as barriers or facilitators depending on the implementation context [25]. Applying the CFIR enabled a structured, theory-driven analysis of the contextual factors shaping RF implementation.

The coding process was conducted independently by 2 researchers (MC and XZ). Any discrepancies were discussed and resolved through team consensus. When disagreements persisted, a senior researcher with expertise in implementation science was consulted to adjudicate and facilitate consensus. Redundant or overlapping codes were merged to ensure clarity and consistency. To identify the most frequently reported implementation determinants, the number of PCPs who described experiences or views related to each CFIR construct was also calculated. This approach helped highlight the most salient barriers and facilitators across participants.

Quality Assurance

In this study, interviewers received comprehensive training to ensure competency in conducting interviews, adherence to the CFIR framework, and consistency across interviews. The training covered key areas, including understanding the CFIR framework, interviewing skills, ethical considerations, mock interviews, and cultural sensitivity. In addition, quality control measures were adopted to ensure the accuracy and reliability of the data collection process. A standardized procedure was developed to guide the interview process, including preinterview preparation and interview conduct. Furthermore, senior researchers (CCZ and MCSW) periodically observed interviews to ensure adherence to the protocol and to provide feedback to interviewers.

Ethical Considerations

The study was reviewed and approved by the Survey and Behavioral Research Ethics Committee, The Chinese University of Hong Kong (approval number SBRE-22-0193). Oral informed consent was obtained from all participants before the interviews. The original informed consent permitted secondary analysis without additional consent. Consent information included details about the study's purpose, procedures, potential risks, and benefits, as well as participants' right to withdraw at any time without penalty. All data were deidentified to protect participant privacy and confidentiality. For identifiable information collected during the primary study, strict data protection protocols were followed, including secure storage on password-protected systems accessible only to authorized personnel. No identifiable information of individual participants is presented in this manuscript or the multimedia appendices. As a token of appreciation, each participant received an HK $100 (US $12.79) coupon upon completion of the interview.

Results

Characteristics of the Included Participants

A total of 40 PCPs participated in this qualitative study, offering diverse professional backgrounds and experiences. Most participants were male (n=31, 78%), while 9 (23%) were female (Table 1). Nearly half (n=18, 45%) were recruited from public health care settings, and 36 (90%) had more than 16 years of clinical experience; 24 (60%) reported seeing fewer than 40 patients per day. The majority were family physicians (n=31, 78%).

XSL•FO
RenderX

Table 1. Demographic characteristics of participants.[a]

Demographic characteristics	Values, n (%)
Gender	
Male	31 (78)
Female	9 (23)
Age (years)	
35-45	7 (18)
46-55	24 (60)
56-65	8 (20)
≥66	1 (3)
Practice experience (years)	
5-15	4 (10)
16-25	13 (33)
26-35	16 (40)
36-45	7 (18)
Practice type (public)	
General outpatient clinic	15 (38)
Family medicine specialist clinics	2 (5)
Specialist outpatient clinics	1 (3)
Practice type (private)	
Solo	13 (33)
With partners	2 (5)
Health maintenance organizations	4 (10)
University	3 (8)
Specialty	
Family medicine	31 (78)
Internal medicine	2 (5)
Community medicine	3 (8)
Psychiatry	1 (3)
Geriatric medicine	3 (8)
Practice size (number of patients seen daily)	
0-19	15 (38)
20-39	9 (23)
40-59	7 (18)
60-79	7 (18)
>100	1 (3)
Missing	1 (3)

[a]Participants were recruited from 8 distinct groups of service providers from February to September 2024 through emails or phone calls.

Barriers and Facilitators to the Adoption of the RF Using CFIR

Barriers and facilitators to the adoption of the RF in daily consultations were identified using the CFIR framework, with key themes emerging from the interviews. Figure 1 illustrates the CFIR domains reflected in the interviews: Innovation, Outer Setting, Inner Setting, Characteristics of Individuals, and the Implementation Process. Table 2 summarizes the frequencies for each identified theme, while Tables 3 and 4 illustrate quotations for each identified theme. Common facilitators included relative advantage (n=26, 65%), innovation design (n=18, 45%), and access to knowledge and information (n=28, 70%). Major barriers were policy and legal concerns and

XSL•FO
RenderX

availability of resources, each cited by 29 (73%) participants. Key themes are detailed below.

Figure 1. Schematic diagram of the Consolidated Framework for Implementation Research.

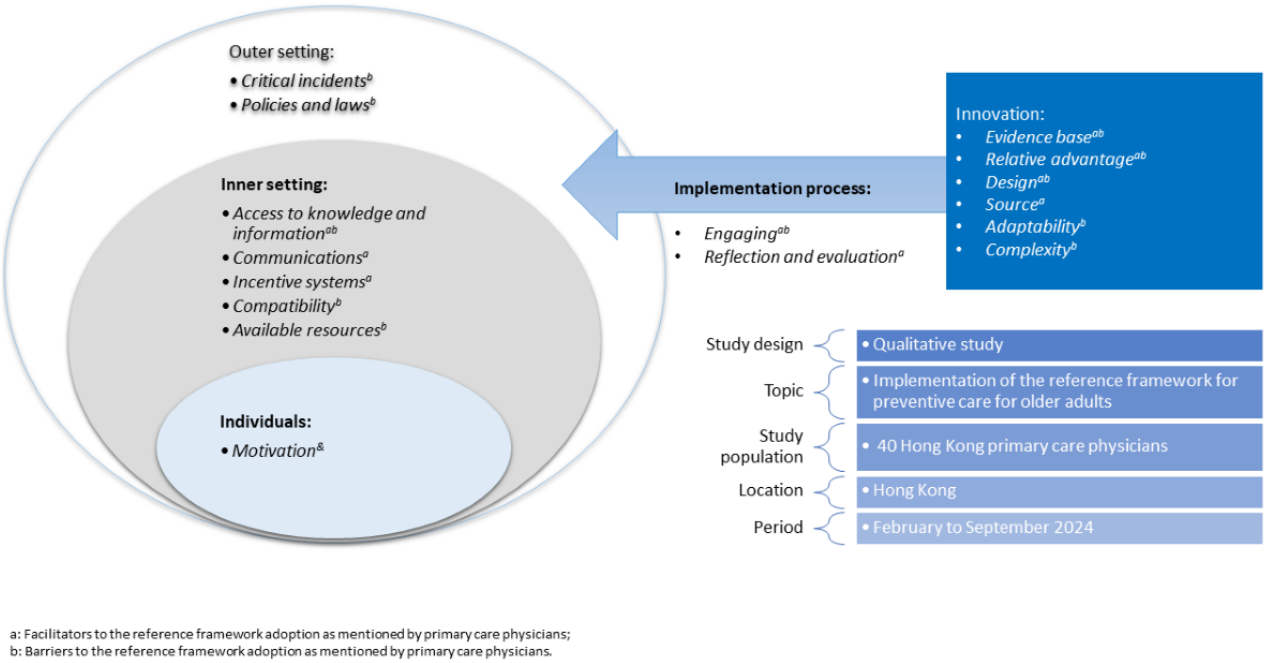

a: Facilitators to the reference framework adoption as mentioned by primary care physicians;
b: Barriers to the reference framework adoption as mentioned by primary care physicians.

XSL•FO
RenderX

Table 2. The barriers and facilitators for the reference framework.[a]

CFIR[b] domains and constructs	Barriers, n (%)	Facilitators, n (%)
Innovation		
Innovation evidence base	4 (10)	7 (18)
Innovation relative advantage	12 (30)	26 (65)
Innovation design	10 (25)	18 (45)
Innovation source	0 (0)	9 (23)
Innovation adaptability	7 (18)	0 (0)
Innovation complexity	6 (15)	0 (0)
Inner Setting		
Access to knowledge	10 (25)	28 (70)
Communications	0 (0)	9 (23)
Incentive systems	0 (0)	3 (8)
Compatibility	9 (23)	0 (0)
Available resources	29 (73)	0 (0)
Outer Setting		
Critical incidents	3 (8)	0 (0)
Policies and laws	29 (73)	0 (0)
Individuals		
Motivation	3 (8)	0 (0)
Implementation Process		
Engaging	5 (13)	10 (25)
Reflection and evaluation	0 (0)	12 (30)

[a]Participants were recruited from 8 distinct groups of service providers from February to September 2024 through emails or phone calls.

[b]CFIR: Consolidated Framework for Implementation Research.

XSL•FO
RenderX

Table 3. Quotations of facilitators for the reference framework.[a]

CFIR[b] domains, subdomains, and example categories	Quotations of facilitators
Innovation	
Evidence base	
Confidence in evidence-based approach	• "Because it is evidence based, rather than purely based on my own policies for the clinic or purely to meet the needs of patient, it reinforces my beliefs." (023) • "I think that if it will encourage more doctors to, you know, provide evidence-based care, maybe, but it might be worthwhile to sort of know, you know, how much we actually are doing or how successful we are given our current approach" (029) • "That is, a lot of evidence base information can actually remind us of some updated treatment plan or some care plan." (04)
Relative advantage	
Value in holistic geriatric care	• "We often neglect to consider the living environment of the elderly or whether there is anyone at home to take care of them. This is another important reminder for us—not just to focus on their illness or the medications they are using." (004) • "Dementia is something everyone has heard of, but what exactly is it? What are the symptoms? What should we pay attention to? These are things that, within this framework, help us realize that we need to be aware of these issues, especially for the elderly." (021)
RF[c] as a structured clinical decision support tool	• "It acts as a structured checklist that ensures comprehensive patient assessment." (009) • "A summary checklist facilitates workflow efficiency by allowing clinicians to systematically track completed and pending clinical steps." (010)
Design	
Visual enhancements improve RF usability and implementation	• "The color coding is helpful, so you can see things at a glance." (029) • "The hardcopy is of a good size to be put in clinics." (031)
Integration of technology and artificial intelligence to enhance RF implementation	• "If we could develop some apps or use AI technology, for example, we could key in certain patient information and come up with conclusions or suggestions. This would be a good direction." (003) • "If there is an advanced version, that is, the RF steps to form a different path, so that the content is more concise, can help me follow the path to care for patients and it will be more convenient for my daily use." (014)
Source	
Trust in the expert-driven development	• "You're talking about doing it based on evidence. Of course, because there's a large team of experts who spend a lot of time putting it together." (021) • "Having experts provide an official guide for local colleagues is very worthwhile." (022) • "I have confidence in its robustness or feasibility because it is made by the Primary Health Care Office." (026)
Inner Setting	
Access to knowledge and information	
Accessibility and availability of RF information facilitate uptake	• "Regular updates provided through the Primary Care Office website contribute to sustained engagement with RF." (011) • "Email notifications and printed summary materials enhance accessibility and reinforce RF awareness." (004)
Communications	
Communication and peer support	• "Our own doctors' peer groups have discussions, so from these three aspects, we are aware of this framework." (005) • "Sometimes, there may be specific content that we can discuss together...Maybe we can encourage more to do it." (010)
Incentive systems	
Incentives and Continuing Medical Education accreditation promote RF engagement	• "They can also earn some CME credits. This encourages all doctors to reference, learn, and study." (001) • "The government may need to provide more incentives, like funding, to support and motivate doctors to perform these screenings." (017)

XSL•FO
RenderX

CFIR[b] domains, subdomains, and example categories	Quotations of facilitators
Implementation Process	
Engaging	
Patient education and shared decision-making facilitate RF integration	• "Important points should be highlighted, possibly using charts or visuals to help elderly patients." (023) • "I think the version you give to the patient should not be the English version but the Chinese version...it is a simplified version plus some explanations...we know exactly what we are doing, but for them (the patient), they may think that what you are asking has nothing to do with the disease...Firstly, it needs to be simplified, secondly, it needs some description...Then they (the patients) won't find you annoying and know that what you are asking is really related to them." (009)
Reflection and evaluation	
Continuous evaluation and iterative refinement of RF to ensure relevance	• "It would be ideal if there were committees that regularly review this, as that would be very practical." (015) • "I hope the relevant authorities can promptly review the effectiveness of this guideline or framework and propose a post-pandemic version." (026) • "For example...if new decisions arise, do we need to make modifications or release a second edition?...it's worth paying attention to how often updates occur, or when necessary, updates can be made. (022)

[a]Participants were recruited from 8 distinct groups of service providers from February to September 2024 through emails or phone calls.

[b]CFIR: Consolidated Framework for Implementation Research.

[c]RF: Reference Framework for Preventive Care for Older Adults.

XSL•FO
RenderX

Table 4. Quotations of barriers for the reference framework.[a]

CFIR[b] domains, subdomains, and example categories	Quotations of facilitators
Innovation	
Evidence base	
Concerns over outdated or overly theoretical content	• "The framework is too theoretical, vague academic, and complex, especially for beginners." (013) • "Sometimes I think it's so theoretical, if we can execute the essence of it in front line?" (015) • "I also noticed that many of the references in here are from a long time ago, or even 20th century." (032)
Relative advantage	
Limited utility for experienced physicians	• "No, it's not very useful for me. I rely more on what I know than on it...for those who are already trained...it may not be very useful." (006) • "I find it hard to recommend this material. It's quite simple, and everyone should know it." (017)
Perceived lack of clinical relevance	• "The information in the document would not help anyone make significant clinical decisions and is insufficient in determining the next step." (003) • "There might be conditions not currently covered." (033)
Design	
Lack of user-friendliness and accessibility	• "It's not user-friendly because when doctors want to implement certain recommendations, they have to go back to the original documents." (040) • "It almost 70, 80 pages for a module, and not many doctors will read it all...We prefer to read things like lecture notes." (002) • "I'm not sure how much is involved in transitioning to an app version or if there's a plan for it. Nowadays, everyone accesses information via smartphones." (034)
Adaptability	
Limited personalization and outdated design	• "Therefore, I find that what's mentioned in it is quite vague and not personalized enough to apply to my patients in family medicine." (006) • "After COVID-19, I think there have been significant changes in elderly health...Utilizing technology for smart health care to help the elderly participate in their health has certainly changed compared to the initial design." (026)
Complexity	
Challenges due to breadth and length	• "It is also too extensive and should, instead, suggest what choices to consider under different situations." (014) • "It was impractical to follow the guidelines as the RF is too comprehensive with too many pages to follow." (002)
Inner Setting	
Access to knowledge and information	
Lack of access to training and awareness	• "When we are working, there may be some CME activities on primary care related topics, but we do not pay special attention to attend a workshop on framework." (022) • "I didn't even know there were any training programs about this" (025)
Compatibility	
Compatibility with clinical practice	• "Actually, I work in geriatrics. When I talk to patients, the concepts in this overview are quite similar." (001) • "I don't need to go through the overview, because in fact many of them are routines, such as diabetes, hypertension" (014) • "We don't usually follow this RF, although sometimes we do things that overlap with the RF, but I don't actually see it from this RF." (016)
Competing priorities and limited focus on the RF[c]	• "I haven't specifically focused on the overview because there are internal workflows and the consultations., we can provide are quite brief due to limited resources." (014)
Available resources	

XSL•FO
RenderX

CFIR[b] domains, subdomains, and example categories	Quotations of facilitators
Time constraints and resource limitations	• "In government institutions, like in our general outpatient clinics, it's quite challenging because we only have 6 minutes for each consultation." (027) • "I look for opportunities during the consultation; if I have extra time, I can discuss a bit more." (018) • "The practicality of the guidelines really depends on how many staff members are available in your clinic and how well you manage your own time." (007)
Outer Setting	
Critical incidents	
Disruption caused by COVID-19	• "In the past few years, we have been implementing this framework, but many events, like COVID, have occurred." (019) • "I haven't participated in many workshops in the past few years due to the epidemic. Our frontline is mainly responsible for the epidemic." (023)
Policies and laws	
Concerns over compulsory adoption and loss of professional autonomy	• "You might have a template for everyone to follow...if someone is entering a nursing home, I need you to fill out all these things, which feels compulsory and may not give the doctor a good impression." (008) • "If it is a policy that we have to follow, then I will have qualified opinion about it...it will make many frontline doctors worry...no more room for freedom...may offend patients, or some patients who have medical needs will not be able to receive suitable treatment." (004)
Characteristics of Individuals	
Motivation	
Preference for alternative guidelines	• "Some colleagues may follow foreign guidelines instead of the local recommendations." (040) • "Doctors in Hong Kong will refer to a lot of different information...even if I don't follow Hong Kong's guideline, but follow US's guideline, it's still evidence base" (035)
Implementation Process	
Engaging	
Challenges with patient adherence and expectations	• "However, when it comes to private practices, if it involves blood tests, it requires citizens to be willing to pay. If a doctor suggests a blood test and the patient is unwilling, that can pose difficulties." (027) • "Currently, the situation is that people come to see a doctor only when they are sick." (021)

[a]Participants were recruited from 8 distinct groups of service providers from February to September 2024 through emails or phone calls.

[b]CFIR: Consolidated Framework for Implementation Research.

[c]RF: Reference Framework for Preventive Care for Older Adults.

Innovation

Innovation Evidence Base

Confidence in Evidence-Based Approach

The RF's credibility, grounded in an evidence-based approach, was acknowledged by participants. A total of 7 (18%) respondents mentioned confidence in the use of the RF based on sufficient evidence-based research for its implementation.

> *Because it is evidence based, rather than purely based on my own policies for the clinic or purely to meet the needs of patient, it reinforces my beliefs.* [023]

> *I think that if it will encourage more doctors to, you know, provide evidence-based care, maybe, but it might be worthwhile to sort of know, you know, how much we actually are doing or how successful we are given our current approach.* [029]

> *That is, a lot of evidence base information can actually remind us of some updated treatment plan or some care plan.* [04]

Concerns Over Outdated or Overly Theoretical Content

However, the evidence for RF has been subject to some criticism, particularly in terms of its timeliness and clinical validity. Some participants (n=4) indicated that the current evidence for RF was outdated and too theoretical for frontline use.

> *The framework is too theoretical, vague academic, and complex, especially for beginners.* [013]

> *Sometimes I think it's so theoretical, if we can execute the essence of it in front line?* [015]

XSL•FO
RenderX

I also noticed that many of the references in here are from a long time ago, or even 20th century. [032]

Innovation Relative Advantage

Value in Holistic Geriatric Care

Some participants noted its practical utility. For example, 26 (65%) participants highlighted its value in ensuring no critical aspects are overlooked, particularly in older adults' care.

We often neglect to consider the living environment of the older adults or whether there is anyone at home to take care of them. This is another important reminder for us—not just to focus on their illness or the medications they are using. [004]

Dementia is something everyone has heard of, but what exactly is it? What are the symptoms? What should we pay attention to? These are things that, within this framework, help us realize that we need to be aware of these issues, especially for the older adults. [021]

RF As a Structured Clinical Decision-Support Tool

In addition, respondents mentioned that this RF helped remind them of some clinical processes that may have been forgotten.

It acts as a structured checklist that ensures comprehensive patient assessment. [009]

A summary checklist facilitates workflow efficiency by allowing clinicians to systematically track completed and pending clinical steps. [010]

Limited Utility for Experienced Physicians

However, 12 (30%) participants pointed out that this RF was not useful to them in practice. On the one hand, they felt that the content of the RF was too basic for experienced physicians.

No, it's not very useful for me. I rely more on what I know than on it...for those who are already trained...it may not be very useful. [006]

I find it hard to recommend this material. It's quite simple, and everyone should know it. [017]

Perceived Lack of Clinical Relevance

On the other hand, they believed that the consultation in this RF was insufficient to help them make useful clinical decisions.

The information in the document would not help anyone make significant clinical decisions and is insufficient in determining the next step. [003]

There might be conditions not currently covered. [033]

Innovation Design

Visual Enhancements Improve RF Usability and Implementation

Participants (n=18) were optimistic about the design. For example, they claimed that the RF is well-sized and color-coded with visual aids, and that these features improve clarity and accessibility.

The color coding is helpful, so you can see things at a glance. [029]

The hardcopy is of a good size to be put in clinics. [031]

Integration of Technology and Artificial Intelligence to Enhance RF Implementation

In addition, respondents made several suggestions to make the RF more user-friendly. Their suggestions focused on how to make it easier and quicker to access the information they need and use the RF to improve clinical efficiency, emphasizing the integration of emerging technologies, especially artificial intelligence.

If we could develop some apps or use AI technology, for example, we could key in certain patient information and come up with conclusions or suggestions. This would be a good direction. [003]

If there is an advanced version, that is, the RF steps to form a different path, so that the content is more concise, can help me follow the path to care for patients and it will be more convenient for my daily use. [014]

Lack of User-Friendliness and Accessibility

In addition, the design of the RF was criticized by some respondents. Ten participants expressed concerns and dissatisfaction with the current design of the RF, especially in terms of the user-friendliness of interactions and the lack of usability of the currently available version in different situations.

It's not user-friendly because when doctors want to implement certain recommendations, they have to go back to the original documents. [040]

It almost 70, 80 pages for a module, and not many doctors will read it all...We prefer to read things like lecture notes. [002]

I'm not sure how much is involved in transitioning to an app version or if there's a plan for it. Nowadays, everyone accesses information via smartphones. [034]

Innovation Source: Trust in the Expert-Driven Development

Within this construct, 9 (23%) participants expressed trust in the expert panels involved in developing the RF. They generally believed that the expertise of these panels ensured the reliability of the RF and supported the evidence-based approach adopted.

You're talking about doing it based on evidence. Of course, because there's a large team of experts who spend a lot of time putting it together. [021]

Having experts provide an official guide for local colleagues is very worthwhile. [022]

I have confidence in its robustness or feasibility because it is made by the Primary Health Care Office. [026]

XSL•FO
RenderX

Innovation Adaptability: Limited Personalization and Outdated Design

Participants (n=7) raised concerns about the local use of this RF at the frontline. They claimed that the lack of personalization or localization in the current version of the RF hinders its use in their practice.

> *Therefore, I find that what's mentioned in it is quite vague and not personalized enough to apply to my patients in family medicine.* [006]
>
> *After COVID-19, I think there have been significant changes in older adults' health...Utilizing technology for smart health care to help the older adults participate in their health has certainly changed compared to the initial design.* [026]

Innovation Complexity: Challenges Due to Breadth and Length

The comprehensiveness of the RF presented a challenge. Some participants (n=6) pointed out that it contained extensive content and too many pages, which limited their ability to implement and follow it in a real-world setting.

> *It is also too extensive and should, instead, suggest what choices to consider under different situations.* [014]
>
> *It was impractical to follow the guidelines as the RF is too comprehensive with too many pages to follow.* [002]

Inner Setting

Access to Knowledge and Information

Accessibility and Availability of RF Information Facilitate Uptake

Most participants (n=28) indicated that they obtained RF-related information and updates mainly through self-directed behavior or informal pathways, such as checking the websites of government health care departments and emails and hard copies sent by relevant departments, and also considered that this made them more familiar with the RF.

> *Regular updates provided through the Primary Care Office website contribute to sustained engagement with RF.* [011]
>
> *Email notifications and printed summary materials enhance accessibility and reinforce RF awareness.* [004]

Lack of Access to Training and Awareness

However, some participants (n=10) pointed out that information on RF training was limited and that they had not even received information on RF training or workshops.

> *When we are working, there may be some CME activities on primary care related topics, but we do not pay special attention to attend a workshop on framework.* [022]
>
> *I didn't even know there were any training programs about this* [025]

Communications: Communications and Peer Support

Maintaining communication with peers is important for encouraging the use of the RF. Nine participants mentioned that they often update information through peer discussions and thus increase peer collaboration, informal feedback on existing content, and review of its utility in the clinic to ensure more effective implementation of the RF.

> *Our own doctors' peer groups have discussions, so from these three aspects, we are aware of this framework.* [005]
>
> *Sometimes, there may be specific content that we can discuss together...Maybe we can encourage more to do it.* [010]

Incentive Systems: Incentives and Continuing Medical Education Accreditation

Incentives, tangible or intangible, are crucial to the implementation of new measures. Three participants mentioned that additional incentives may be needed to increase the use of the RF, including the inclusion of the RF in Continuing Medical Education (CME) credit courses to encourage more physicians to participate. In addition, providing financial assistance was cited as an effective approach.

> *They can also earn some CME credits. This encourages all doctors to reference, learn, and study.* [001]
>
> *The government may need to provide more incentives, like funding, to support and motivate doctors to perform these screenings.* [017]

Compatibility

Compatibility With Clinical Practice

Some physicians (n=9) indicated that they were reluctant to use or refer to the RF in their day-to-day work, and one of the primary reasons was that they recognized the content of the RF as similar to what they did on a day-to-day basis.

> *Actually, I work in geriatrics. When I talk to patients, the concepts in this overview are quite similar.* [001]
>
> *We don't usually follow this RF, although sometimes we do things that overlap with the RF, but I don't actually see it from this RF.* [016]
>
> *I don't need to go through the overview, because in fact many of them are routines, such as diabetes, hypertension.* [014]

Competing Priorities and Limited Focus on the RF

In addition, the prioritization of internal workflows may hinder their use of the RF, as one participant mentioned.

> *I haven't specifically focused on the overview because there are internal workflows and the consultations.* [014]

Available Resources: Time Constraints and Resource Limitations

When discussing how the RF works in practice, most participants (n=29) complained that real-world resource

XSL•FO
RenderX

constraints limited their use of the RF, especially time constraints. Resource constraints have been a longstanding theme in the health care system, especially in the public health care system, where respondents reported that their consultation time per patient was only a few minutes, making it difficult to implement the RF. In addition, the lack of staff presents an additional challenge.

> *In government institutions, like in our general outpatient clinics, it's quite challenging because we only have 6 minutes for each consultation.* [027]
>
> *I look for opportunities during the consultation; if I have extra time, I can discuss a bit more.* [018]
>
> *The practicality of the guidelines really depends on how many staff members are available in your clinic and how well you manage your own time.* [007]

Outer Setting

Critical Incidents: Disruption Caused by COVID-19

The COVID-19 pandemic posed challenges to the RF's promotion and implementation, with 3 participants mentioning that the pandemic limited its outreach. They noted that the arrival of COVID-19 shifted their focus to COVID-19–related activities, to the extent that they had little exposure to RF-related knowledge or training, despite having started RF implementation several years ago.

> *In the past few years, we have been implementing this framework, but many events, like COVID, have occurred.* [019]
>
> *I haven't participated in many workshops in the past few years due to the epidemic. Our frontline is mainly responsible for the epidemic.* [023]

Policies and Laws: Concerns Over Compulsory Adoption and Loss of Professional Autonomy

When we addressed their views on whether the RF should be made a policy, most participants (n=29) showed clear opposition, believing that if the RF is made mandatory, it may cause physicians to lose their clinical freedom, thereby ignoring the real needs of their patients.

> *If it is a policy that we have to follow, then I will have qualified opinion about it...it will make many frontline doctors worry...no more room for freedom...may offend patients, or some patients who have medical needs will not be able to receive suitable treatment.* [004]
>
> *You might have a template for everyone to follow...if someone is entering a nursing home, I need you to fill out all these things, which feels compulsory and may not give the doctor a good impression.* [008]

Individuals (Motivation: Preference for Alternative Guidelines)

We found that PCPs' beliefs about different health care systems may influence their acceptance and adoption of the RF. Participants (n=3) noted that they may compare practices and guidelines from different regions on their own and place more faith in foreign guidelines than in local recommendations.

> *Some colleagues may follow foreign guidelines instead of the local recommendations.* [040]
>
> *Doctors in Hong Kong will refer to a lot of different information...even if I don't follow Hong Kong's guideline, but follow US's guideline, it's still evidence base.* [035]

Implementation Process

Engaging

Challenges With Patient Adherence and Expectations

We found that some patients' uncooperative behaviors or habits may affect physicians' confidence and willingness to implement the RF. Some participants (n=5) noted that even when the RF is followed to make recommendations for patients, its effectiveness requires patient participation, and that uncooperative attitudes make the RF difficult to implement. In addition, respondents pointed to doubts about the usefulness of the RF, emphasizing that confidence in the RF is low in the current style of patient health service seeking.

> *However, when it comes to private practices, if it involves blood tests, it requires citizens to be willing to pay. If a doctor suggests a blood test and the patient is unwilling, that can pose difficulties.* [027]
>
> *Currently, the situation is that people come to see a doctor only when they are sick...however, the government's future direction hopes to detect diseases earlier, intervene sooner, and prevent illnesses from occurring.* [021]

Patient Education and Shared Decision-Making Facilitate RF Integration

Although there are some limitations to RF adoption, some participants (n=10) noted that the RF may be adopted because of the role it plays, particularly in empowering patients and helping doctors and patients reach consensus. Participants noted that uncooperative behaviors may decrease and physician-patient communication may be enhanced if patients are aware of the RF.

> *I think the version you give to the patient should not be the English version but the Chinese version, and it should be simpler, it is a simplified version plus some explanations...Of course, we know exactly what we are doing, but for them (the patient), they may think that what you are asking has nothing to do with the disease, they are just seeing a doctor...Firstly, it needs to be simplified, secondly, it needs some description...Then they (the patients) won't find you annoying and know that what you are asking is really related to them.* [009]
>
> *Important points should be highlighted, possibly using charts or visuals to help older adults.* [023]

XSL•FO
RenderX

Reflection and Evaluation: Continuous Evaluation and Iterative Refinement of RF to Ensure Relevance

When discussing further promotion of RF adoption, participants (n=12) emphasized the importance of regular review and updating to keep up with the latest trends, especially after experiencing critical incidents such as the COVID-19 pandemic.

> *It would be ideal if there were committees that regularly review this, as that would be very practical.* [015]

> *After COVID-19, I think there have been significant changes in older adults' health...Utilizing technology for smart health care to help the older adults participate in their health has certainly changed compared to the initial design. I hope the relevant authorities can promptly review the effectiveness of this guideline or framework and propose a post-pandemic version.* [026]

> *For example, after five years or a few years, if new decisions arise, do we need to make modifications or release a second edition? I know they would do that, but it's worth paying attention to how often updates occur, or when necessary, updates can be made.* [022]

Discussion

Summary of Findings

This qualitative study examined the views of 40 PCPs on the adoption of the RF in Hong Kong. The findings revealed a complex interplay of systemic, organizational, and individual-level factors influencing RF implementation. At the policy level, many participants expressed concern about making RF adoption mandatory, fearing a loss of clinical autonomy and flexibility in tailoring care to individual patient needs. Participants preferred the RF as a recommendation rather than a compulsory protocol. Within the Inner Setting, peer discussions emerged as a facilitator of RF awareness and informal learning. However, some physicians perceived the RF as redundant, given its conceptual overlap with their routine practice. Competing clinical priorities and limited consultation time also hindered engagement with the RF. Although most participants accessed RF information through government websites or printed materials, many cited a lack of structured training opportunities and expressed interest in clearer dissemination and CME-accredited courses. Resource constraints—especially time limitations and staffing shortages—posed significant barriers to practical implementation. Furthermore, physicians highlighted the importance of financial and educational incentives to support adoption. At the individual level, some PCPs demonstrated a preference for foreign guidelines over local ones, reflecting skepticism toward the RF's added value. Patient factors, including low adherence, financial considerations, and a curative (rather than preventive) care-seeking mindset, were seen as major challenges to RF uptake. Despite these barriers, several participants acknowledged the potential of the RF to enhance physician-patient communication and support shared decision-making, particularly if supported by simplified, culturally appropriate patient materials. Continuous evaluation and iterative updates to the RF were viewed as essential to maintaining relevance, especially in the post–COVID-19 context. To further understand how these factors interact, we mapped them across the CFIR domains to identify patterns and relationships that shape implementation outcomes.

Interpretation Across CFIR Domains

Within the Intervention Characteristics domain, trust in the RF's evidence-based foundation and expert development fostered credibility, yet its perceived complexity and limited adaptability constrained its practical use—especially in time-constrained environments (Inner Setting). External pressures such as the COVID-19 pandemic and resistance to policy mandates (Outer Setting) further shaped adoption, interacting with individual beliefs and preferences (Characteristics of Individuals), including skepticism toward local guidelines. Peer communication and informal learning (Inner Setting) helped mitigate gaps in formal training, while patient engagement and shared decision-making (Process) emerged as critical to overcoming adherence challenges. These cross-domain patterns suggest that effective implementation requires not only credible and well-designed content but also responsive adaptation to clinical workflows, supportive infrastructure, and active stakeholder involvement.

Possible Explanation and Recommendations

This study received both positive and negative feedback from practitioners regarding their impressions of the RF, with some practitioners indicating that the RF was useful, while others criticized that the content was not applicable to daily consultation routines. Based on our results, practitioner noncompliance may be related to personal attitudes toward the RF and limitations of the RF, such as too many pages and differing physiological conditions of patients. The results revealed that improvement of the RF was necessary to enhance practitioner compliance. High complexity and a high number of conditional recommendations in guidelines have been found to be major barriers to physicians adopting guidelines [26]. The current RF contains comprehensive information but lacks a quick checklist for practitioners to apply. A simple checklist for practitioners and patients may help improve RF compliance. Furthermore, comments from physicians reflected a lack of collaboration and communication between expert panels and frontline physicians, resulting in a tool that prioritizes general guidance over actionable, context-specific solutions. Lu and colleagues [27] emphasized and provided key strategies for successful implementation, in which stakeholder engagement was identified as an important process for success. To bridge this gap, co-designing updates to the RF with frontline practitioners is essential. Regular feedback loops and workshops involving diverse stakeholders, including PCPs, specialists, and nurses, can help refine the framework. These updates should focus on creating adaptable modules or case-based scenarios that address specific clinical complexities while retaining the RF's evidence-based rigor. A critical interpretation of our results points to a fundamental implementation challenge: the RF was developed with insufficient input from its end users. This top-down approach likely explains why physicians viewed it as a source of general guidance rather than a practical clinical

XSL•FO
RenderX

tool. This aligns with the implementation science literature, which consistently identifies stakeholder engagement as a cornerstone of success [27]. Therefore, a paradigm shift from consultation to co-creation is warranted. We recommend that updates to the RF be developed through participatory workshops with frontline PCPs, specialists, and nurses. This co-design process would focus on embedding actionable, case-based scenarios directly into the framework, enhancing its relevance and utility in daily practice.

The RF received a positive response regarding the user-friendliness and convenient design of the framework, especially in terms of size and color coding. User-friendliness also emerged as a crucial factor in promoting regular use. Catho et al [28] conducted a qualitative study to examine factors determining guideline adherence and adoption of computerized decision support systems by physicians. The study indicated that user-friendliness was one of the key factors affecting the adoption of new tools in the workplace. The underlying mechanism may be attributed to the fact that an intuitive interface and adequate technical support may help clinicians use the RF easily, even if they are not familiar with the technology [29].

By contrast, users also identified barriers to implementing the RF in daily consultations. Resource constraints, especially insufficient consultation time, were reported by most participants and substantially affected RF adoption. A systematic review covering 18 countries and 28 million consultations across 111 publications from 1946 to 2016 revealed that in half of the cases, consultation time was 5 minutes or less [30]. The main contributor to this phenomenon was heavy practitioner workloads, for example, in China [31]. No guideline recommends an optimal consultation time for outpatient clinics, but consultation times of less than 7 minutes were found to be more likely to result in practitioners missing some psychosocial problems that patients have [32], let alone following the RF within such a short consultation. Solo practitioners must negotiate optimal medication for patients, address concerns from patients and families, manage schedule constraints, and—above all—consider patients', families', or caregivers' capacity to follow the recommended regimen within a limited consultation. Despite participants having positive responses to the robustness, feasibility, benefits for patient care, user-friendliness, and convenience features of the RF, they also reflected that sufficient consultation time would be key to whether the RF was practically applicable. Notably, our findings emphasize physicians' resistance to mandatory use of the RF, possibly due to impairment of job autonomy or negative effects from increased external controls [33].

This study highlights physicians perceived role of patient engagement in the decision-making process of whether to practice the RF and points out that low patient compliance and poor care outcomes may undermine physicians' confidence and willingness to perform the RF, which may be related to physicians perceived meaning of their work. Previous research has indicated that physicians' perceptions of health care outcomes and the meaning of their work may influence or prevent the onset of burnout [34], and that there is a negative relationship between burnout and physicians' perceived value of an innovation or willingness to adopt an innovation [35]. This mechanism reflects the fact that physicians may abandon use of the RF because of perceived lack of patient engagement or ineffective care. Therefore, it may be important to enhance physicians' sense of meaning and willingness to use innovations by improving patient engagement or adherence. In addition, we found that physicians offered many ideas for improving patient engagement, particularly in the areas of patient empowerment and physician-patient consensus. They noted that the RF should be expanded to include patient education, which would help them better implement the content of the RF. This approach may be effective in that, on the one hand, patient empowerment is evident in stimulating patient engagement and optimizing the use of limited resources such as time, as well as reducing the burden on frontline health care professionals [36,37]. On the other hand, improved physician-patient communication may enhance the likelihood of long-term care outcomes, thereby increasing physician confidence in the use of the RF [34].

Continuous improvement is vital for the sustained relevance and effectiveness of the RF. Regularly seeking user feedback and conducting usability testing could help identify areas for enhancement [38]. Implementing long-term engagement strategies, such as personalized reminders, could also help sustain routine use effectively [39-41]. This study revealed insufficient training or workshops concerning the RF, especially during critical incidents such as COVID-19. To support the adoption of the RF, ongoing training and user support are critical. Workshops, clear instructional materials, and accessible technical support [42] could help practitioners become more familiar with the RF as a whole. By implementing these measures, health care providers could navigate the RF more effectively and better utilize the 6 modules in the RF, improving their overall experience and patient health outcomes.

In this study, some constructs, such as structural characteristics and local attitudes, were not coded or mentioned by physicians. This absence appears to reflect the local context. For example, infrastructure was not perceived as a barrier or facilitator, likely because the RF did not require significant changes to physical or technological systems. Similarly, the limited engagement of patients in the Implementation Process may be attributed to the top-down nature of Hong Kong's health care system, where patients typically have less access to information and decision-making power compared with health care professionals.

Implications for Practice and Research

Beyond the specific barriers and facilitators identified, our findings underscore that the broader cultural, organizational, and policy contexts in Hong Kong are not passive settings but active forces shaping RF implementation. Culturally, the longstanding preference for curative care over preventive approaches presents a foundational challenge [43]. This orientation suggests that the success of the RF is not only dependent on its technical integration into clinical workflows but also on a broader societal shift toward preventive health literacy. Without sufficient patient understanding and engagement, even a well-integrated RF may encounter resistance or low adherence in practice. At the organizational level, the structure and pace of the public health care system impose

XSL•FO
RenderX

significant constraints. Physicians often operate under intense time pressure, and many clinical workflows are already well established. In this context, any new intervention must demonstrate clear added value and seamless compatibility with existing practices; otherwise, it risks being perceived as redundant or burdensome. Therefore, implementation strategies must be designed with these operational realities in mind, ensuring that the RF enhances rather than disrupts clinical efficiency. From a policy perspective, top-down mandates may be counterproductive if they conflict with the cultural value placed on clinical autonomy and the practical limitations of consultation time. Rather than imposing rigid requirements, policy should play a facilitative role. This includes providing resources to alleviate time constraints and supporting integration of the RF into existing digital infrastructures, such as the eHealth app [44]. Embedding the RF into such platforms could streamline access to patient-specific recommendations, reduce the cognitive and administrative load on practitioners, and ultimately improve adherence.

In addition to digital integration, the development of a lay version of the RF could empower older adults by offering accessible, user-friendly guidance on lifestyle changes and risk awareness. This would not only support patient engagement but also reinforce the preventive orientation that the RF seeks to promote. By equipping patients with the knowledge and tools to participate actively in their care, the RF could become a more effective and sustainable component of routine practice. While financial incentives have been proposed to increase practitioner adherence, previous studies have shown that such incentives often have limited impact on clinical behavior [45-47]. As such, financial incentives should be considered a last resort rather than a primary strategy. Instead, the focus should be on creating a supportive ecosystem in which the RF is perceived as both useful and feasible to implement.

This study offers several novel contributions to the CFIR literature. First, to our knowledge, it is the first to systematically investigate physician perspectives on the RF for the older adult population within the distinct sociocultural and structural context of Hong Kong. Second, our application of the CFIR uncovered specific barriers, such as how the COVID-19 pandemic introduced unique, large-scale disruptions that directly impeded the RF's promotion and integration. Finally, our findings provide actionable insights for tailoring implementation strategies in similar high-density, public payer–led Asian metropolises, addressing a gap in the current evidence base.

Importantly, future research should build on this foundation by applying the CFIR-Expert Recommendations for Implementing Change (ERIC) Matching Tool, which links CFIR-identified barriers to evidence-based implementation strategies from the ERIC taxonomy [48]. This tool maps 73 strategies to CFIR constructs, supporting the selection of targeted interventions such as preparing champions, conducting educational meetings, and tailoring strategies to local contexts [49]. Empirical studies have demonstrated the utility of this approach in improving implementation outcomes across various health care settings, including primary care and geriatric models [50,51]. Future research should focus on testing CFIR-ERIC–informed strategies to address the specific barriers identified in this study, such as limited leadership engagement, resource constraints, and competing clinical priorities. Comparative studies across different health care systems would help assess the generalizability of these strategies, while mixed methods and longitudinal designs could explore their impact on practitioner compliance, patient outcomes, and system-level sustainability.

Strengths and Limitations

This study is the first to comprehensively assess the adoption of the RF in the primary care setting, offering new insights and potential policy implications for enhancing clinical adoption of the RF. It draws strength from a substantial sample size of 40 PCPs, whose extensive clinical experience enriched the depth and relevance of the findings. Use of the CFIR provided a structured, theory-informed approach to identifying barriers and facilitators, enhancing the analytical rigor of the study. Moreover, the study contributes to the CFIR literature by applying the framework in a unique sociocultural and health care context, highlighting context-specific determinants that may not be captured in differently structured health systems.

However, several limitations should be acknowledged. First, the study relied on self-reported perceptions, which may introduce bias due to social desirability or recall limitations. Second, the findings are context specific and may not be directly transferable to other health care systems. Comparative studies across different health care systems would help assess the generalizability of these findings. Third, the study focused solely on the perspectives of PCPs, without incorporating views from other stakeholders such as patients, administrators, or policy makers, which may have provided a more holistic understanding of RF adoption. Including perspectives from patients, administrators, and policy makers will be essential to developing a more comprehensive understanding of RF implementation and its long-term viability.

Conclusions

This is the first study investigating the perception of RF implementation among PCPs in Hong Kong guided by the CFIR, offering essential insights for future enhancements of RF and aiding the translation of evidence into clinical practice. While the RF holds promise for enhancing preventive care and shared decision-making in Hong Kong's primary care setting, its adoption is currently hindered by systemic, organizational, and individual barriers. To support effective integration into routine practice, flexible implementation, stronger institutional support, tailored incentives, and improved clinician and patient engagement strategies are essential. Policy makers could begin by promoting awareness through targeted advertising and professional workshops, followed by revising and refining RF content to better align with clinical realities and patient needs. Future research should empirically test CFIR-ERIC–informed implementation strategies to address identified barriers, explore patient outcomes to support practitioner engagement, and conduct comparative studies to assess the transferability of RF adoption across diverse health care systems.

XSL•FO
RenderX

Acknowledgments

MCSW is an honorary medical advisor to GenieBiome Ltd., SunRise, and BGI Health; is an advisory committee member of Pfizer; an external expert for GlaxoSmithKline Limited; a member of the advisory board of AstraZeneca; has received consultancy fees for providing research advice; and has received grants from Pfizer, Merck Sharp & Dohme, AstraZeneca, GSK, and Moderna. We acknowledge the contribution of Ms. Stephanie Kwan, research assistant at The Chinese University of Hong Kong, for assisting with the interview study. During the preparation of this work, the authors used ChatGPT (OpenAI) solely to improve language and readability, with caution. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This study was funded by the Health and Medical Research Fund of the Hong Kong Special Administrative Region (reference number 20210381). The funder had no role in the study design, data collection, analysis, interpretation, or writing of the manuscript.

Data Availability

The data supporting the findings of this study are available from the corresponding author (MCSW) upon reasonable request.

Authors' Contributions

JH, CCZ, WCWW, and MCSW conceptualized and supervised this study. CYL, XZ, MC, and CCZ were responsible for data curation and formal analysis. CCZ, MC, and MKY drafted the manuscript. JH, WCWW, and MCSW reviewed and edited the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[DOCX File , 24 KB - jmir_v28i1e78814_app1.docx]

Multimedia Appendix 2
SRQR Checklist.
[DOCX File , 22 KB - jmir_v28i1e78814_app2.docx]

References

1. Xue D, Bai Q, Bian Y. How working-age population education and health of older people shape the burden of population aging: a comparative study of Macau, Hong Kong, and Singapore. Front Public Health 2022;10:1031229 [FREE Full text] [doi: 10.3389/fpubh.2022.1031229] [Medline: 36408011]
2. Hong Kong population projections for 2022 to 2046. Census and Statistics Department. 2023. URL: https://www.censtatd.gov.hk/en/EIndexbySubject.html?pcode=FA100061&scode=190 [accessed 2026-02-17]
3. Hong Kong population projections for 2017 to 2066. Census and Statistics Department. 2017. URL: https://www.censtatd.gov.hk/hkstat/sub/sp190.jsp?productCode=FA100061 [accessed 2026-02-17]
4. Wetterling T. Pathogenesis of multimorbidity-what is known? Z Gerontol Geriatr 2021 Oct 10;54(6):590-596 [FREE Full text] [doi: 10.1007/s00391-020-01752-z] [Medline: 32651847]
5. Hu X, Huang J, Lv Y, Li G, Peng X. Status of prevalence study on multimorbidity of chronic disease in China: systematic review. Geriatr Gerontol Int 2015 Jan 27;15(1):1-10. [doi: 10.1111/ggi.12340] [Medline: 25163532]
6. Kim K, Lee JH, Kim C. Impaired health-related quality of life in elderly women is associated with multimorbidity: results from the Korean National Health and Nutrition Examination Survey. Gend Med 2012 Oct;9(5):309-318. [doi: 10.1016/j.genm.2012.08.001] [Medline: 22940276]
7. Hong Kong reference framework for preventive care for older adults in primary care settings (revised edition 2021). Health Bureau. 2021. URL: https://www.fhb.gov.hk/pho/rfs/english/reference_framework/pre_care_for_older_adults.html [accessed 2026-02-17]
8. Piccini JP, Xu H, Cox M, Matsouaka RA, Fonarow GC, Butler J, et al. Adherence to guideline-directed stroke prevention therapy for atrial fibrillation is achievable. Circulation 2019 Mar 19;139(12):1497-1506. [doi: 10.1161/circulationaha.118.035909]
9. Wang Z, Norris SL, Bero L. The advantages and limitations of guideline adaptation frameworks. Implement Sci 2018 May 29;13(1):72-13 [FREE Full text] [doi: 10.1186/s13012-018-0763-4] [Medline: 29843737]

XSL•FO
RenderX

10. Sin CK, Fu S, Tsang CS, Tsui WW, Chan FH. Prevention in primary care is better than cure: The Hong Kong Reference Framework for Preventive Care for Older Adults--translating evidence into practice. Hong Kong Med J 2015 Aug 4;21(4):353-359 [FREE Full text] [doi: 10.12809/hkmj144326] [Medline: 26238132]
11. Hong Kong Reference Framework for Diabetes Care. Primary Health Commission. URL: https://www.healthbureau.gov.hk/phcc/rfs/english/reference_framework/diabetes_care.html [accessed 2026-02-17]
12. Hypertension care. Primary Health Commission. URL: https://www.healthbureau.gov.hk/phcc/rfs/english/reference_framework/hypertension_care.html [accessed 2026-02-17]
13. World Health Organization (WHO). Active ageing: a policy framework. WHO. 2021. URL: https://iris.who.int/server/api/core/bitstreams/0418705f-1c82-4fe2-82d6-d4faee89dfa0/content [accessed 2026-02-17]
14. Jimenez G, Matchar D, Koh GCH, Tyagi S, van der Kleij RMJJ, Chavannes NH, et al. Revisiting the four core functions (4Cs) of primary care: operational definitions and complexities. Prim Health Care Res Dev 2021 Nov 10;22(1):e68 [FREE Full text] [doi: 10.1017/S1463423621000669] [Medline: 34753531]
15. Apóstolo J, Cooke R, Bobrowicz-Campos E, Santana S, Marcucci M, Cano A, et al. Effectiveness of interventions to prevent pre-frailty and frailty progression in older adults: a systematic review. JBI Database System Rev Implement Rep 2018 Jan;16(1):140-232 [FREE Full text] [doi: 10.11124/JBISRIR-2017-003382] [Medline: 29324562]
16. Atkins L, Francis J, Islam R, O'Connor D, Patey A, Ivers N, et al. A guide to using the Theoretical Domains Framework of behaviour change to investigate implementation problems. Implement Sci 2017 Jun 21;12(1):77 [FREE Full text] [doi: 10.1186/s13012-017-0605-9] [Medline: 28637486]
17. Holtrop JS, Estabrooks PA, Gaglio B, Harden SM, Kessler RS, King DK, et al. Understanding and applying the RE-AIM framework: clarifications and resources. J Clin Transl Sci 2021;5(1):e126 [FREE Full text] [doi: 10.1017/cts.2021.789] [Medline: 34367671]
18. Consolidated Framework for Implementation Research. CFIR Research Team-Center for Clinical Management Research. URL: https://cfirguide.org/ [accessed 2026-02-18]
19. Creswell J, Creswell J. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Thousand Oaks, CA: Sage Publications; 2017.
20. Barbie E. The Basics of Social Research. Belmont, CA: Thomson/Wadsworth; 2008.
21. Salkind NJ. Encyclopedia of Research Design. Thousand Oaks, CA: SAGE Publications; 2010.
22. Vasileiou K, Barnett J, Thorpe S, Young T. Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. BMC Med Res Methodol 2018 Nov 21;18(1):148-118 [FREE Full text] [doi: 10.1186/s12874-018-0594-7] [Medline: 30463515]
23. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research. Academic Medicine 2014;89(9):1245-1251. [doi: 10.1097/acm.0000000000000388]
24. Hagens V, Dobrow MJ, Chafe R. Interviewee transcript review: assessing the impact on qualitative research. BMC Med Res Methodol 2009 Jul 06;9(1):47 [FREE Full text] [doi: 10.1186/1471-2288-9-47] [Medline: 19580666]
25. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. Implementation Sci 2022 Oct 29;17(1):75 [FREE Full text] [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
26. Qumseya B, Goddard A, Qumseya A, Estores D, Draganov PV, Forsmark C. Barriers to clinical practice guideline implementation among physicians: a physician survey. IJGM 2021 Nov;Volume 14:7591-7598. [doi: 10.2147/ijgm.s333501]
27. Lu AD, Kaul B, Reichert J, Kilbourne AM, Sarmiento KF, Whooley MA. Implementation strategies for frontline healthcare professionals: people, process mapping, and problem solving. J Gen Intern Med 2021 Feb 11;36(2):506-510 [FREE Full text] [doi: 10.1007/s11606-020-06169-3] [Medline: 32918200]
28. Catho G, Centemero NS, Catho H, Ranzani A, Balmelli C, Landelle C, on the behalf of the Q-COMPASS study group. Factors determining the adherence to antimicrobial guidelines and the adoption of computerised decision support systems by physicians: a qualitative study in three European hospitals. Int J Med Inform 2020 Sep;141:104233 [FREE Full text] [doi: 10.1016/j.ijmedinf.2020.104233] [Medline: 32736330]
29. Morrow S, Daines L, Wiener-Ogilvie S, Steed L, McKee L, Caress A, et al. Exploring the perspectives of clinical professionals and support staff on implementing supported self-management for asthma in UK general practice: an IMPART qualitative study. NPJ Prim Care Respir Med 2017 Jul 18;27(1):45 [FREE Full text] [doi: 10.1038/s41533-017-0041-y] [Medline: 28720751]
30. Irving G, Neves AL, Dambha-Miller H, Oishi A, Tagashira H, Verho A, et al. International variations in primary care physician consultation time: a systematic review of 67 countries. BMJ Open 2017 Nov 08;7(10):e017902 [FREE Full text] [doi: 10.1136/bmjopen-2017-017902] [Medline: 29118053]
31. Jin G, Zhao Y, Chen C, Wang W, Du J, Lu X. The length and content of general practice consultation in two urban districts of Beijing: a preliminary observation study. PLoS One 2015 Aug 10;10(8):e0135121 [FREE Full text] [doi: 10.1371/journal.pone.0135121] [Medline: 26258911]
32. Caroe TD. Consultation length. Br J Gen Pract 2019 Sep 26;69(687):488.1-48488. [doi: 10.3399/bjgp19x705665]

XSL•FO
RenderX

33. Jacob C, Sanchez-Vazquez A, Ivory C. Social, organizational, and technological factors impacting clinicians' adoption of mobile health tools: systematic literature review. JMIR Mhealth Uhealth 2020 Feb 20;8(2):e15935 [FREE Full text] [doi: 10.2196/15935] [Medline: 32130167]
34. Hiefner AR, Constable P, Ross K, Sepdham D, Ventimiglia JB. Protecting family physicians from burnout: meaningful patient-physician relationships are "more than just medicine". J Am Board Fam Med 2022 Jul 27;35(4):716-723 [FREE Full text] [doi: 10.3122/jabfm.2022.04.210441] [Medline: 35896474]
35. Bian D, Xiao Y, Song K, Dong M, Li L, Millar R, et al. Determinants influencing the adoption of internet health care technology among Chinese health care professionals: extension of the value-based adoption model with burnout theory. J Med Internet Res 2023 Mar 10;25:e37671 [FREE Full text] [doi: 10.2196/37671] [Medline: 36897630]
36. Barenfeld E, Ekman I, Cederberg M, Fors A, Ali L, Gyllensten H. Patient and healthcare professional engagement and time use within a randomised controlled trial: investigating intervention costs associated with remote person-centred care in Sweden. BMJ Open 2025 Oct 09;15(10):e099034 [FREE Full text] [doi: 10.1136/bmjopen-2025-099034] [Medline: 41067773]
37. Salmon P, Hall GM. Patient empowerment and control: a psychological discourse in the service of medicine. Soc Sci Med 2003 Nov;57(10):1969-1980. [doi: 10.1016/s0277-9536(03)00063-7] [Medline: 14499519]
38. Heen AF, Vandvik PO, Brandt L, Achille F, Guyatt GH, Akl EA, et al. Decision aids linked to evidence summaries and clinical practice guidelines: results from user-testing in clinical encounters. BMC Med Inform Decis Mak 2021 Jun 29;21(1):202 [FREE Full text] [doi: 10.1186/s12911-021-01541-7] [Medline: 34187484]
39. Ricciardi L, Mostashari F, Murphy J, Daniel JG, Siminerio EP. A national action plan to support consumer engagement via e-health. Health Aff (Millwood) 2013 Feb;32(2):376-384. [doi: 10.1377/hlthaff.2012.1216] [Medline: 23381531]
40. Jacob C, Bourke S, Heuss S. From testers to cocreators-the value of and approaches to successful patient engagement in the development of eHealth solutions: qualitative expert interview study. JMIR Hum Factors 2022 Oct 06;9(4):e41481 [FREE Full text] [doi: 10.2196/41481] [Medline: 36102548]
41. Virani SS, Ramsey DJ, Westerman D, Kuebeler MK, Chen L, Akeroyd JM, et al. Cluster randomized trial of a personalized clinical decision support intervention to improve statin prescribing in patients with atherosclerotic cardiovascular disease. Circulation 2023 May 02;147(18):1411-1413. [doi: 10.1161/CIRCULATIONAHA.123.064226] [Medline: 36871214]
42. Rödlund A, Toropova A, Lengnick-Hall R, Powell BJ, Schäfer Elinder L, Björklund C, et al. A cluster-randomized controlled trial assessing the effectiveness of a multifaceted versus a discrete implementation strategy on fidelity to an occupational guideline for the prevention of mental health problems at the workplace: a dual perspective from Swedish schools. Implement Sci Commun 2025 Nov 21;6(1):127 [FREE Full text] [doi: 10.1186/s43058-025-00821-x] [Medline: 41272910]
43. Babatunde AO, Shobanke HA, Akinade AA, Michael AJ, Osadare M, Akanbi OK, et al. Enhancing preventive medicine over curative medicine: role of telemedicine. Public Health Pract (Oxf) 2021 Nov;2:100130 [FREE Full text] [doi: 10.1016/j.puhip.2021.100130] [Medline: 36101617]
44. What's eHealth. Electronic Health Record Sharing System (eHealth). URL: https://www.ehealth.gov.hk/en/whats-ehealth/index.html [accessed 2026-02-17]
45. van Dijk CE, Verheij RA, Spreeuwenberg P, van den Berg MJ, Groenewegen PP, Braspenning J, et al. Impact of remuneration on guideline adherence: empirical evidence in general practice. Scand J Prim Health Care 2013 Mar 21;31(1):56-63 [FREE Full text] [doi: 10.3109/02813432.2012.757078] [Medline: 23330604]
46. van Dijk CE, Verheij RA, te Brake H, Spreeuwenberg P, Groenewegen PP, de Bakker DH. Changes in the remuneration system for general practitioners: effects on contact type and consultation length. Eur J Health Econ 2014 Jan 28;15(1):83-91. [doi: 10.1007/s10198-013-0458-3] [Medline: 23446626]
47. Lugtenberg M, Burgers JS, Han D, Westert GP. General practitioners' preferences for interventions to improve guideline adherence. J Eval Clin Pract 2014 Dec 21;20(6):820-826. [doi: 10.1111/jep.12209] [Medline: 24953439]
48. Waltz TJ, Powell BJ, Fernández ME, Abadie B, Damschroder LJ. Choosing implementation strategies to address contextual barriers: diversity in recommendations and future directions. Implement Sci 2019 Apr 29;14(1):42 [FREE Full text] [doi: 10.1186/s13012-019-0892-4] [Medline: 31036028]
49. Waltz TJ, Powell BJ, Matthieu MM, Damschroder LJ, Chinman MJ, Smith JL, et al. Use of concept mapping to characterize relationships among implementation strategies and assess their feasibility and importance: results from the Expert Recommendations for Implementing Change (ERIC) study. Implement Sci 2015 Aug 07;10(1):109 [FREE Full text] [doi: 10.1186/s13012-015-0295-0] [Medline: 26249843]
50. Shin M, Montano A, Adjognon O, Harvey K, Solimeo S, Sullivan J. Identification of implementation strategies using the CFIR-ERIC matching tool to mitigate barriers in a primary care model for older veterans. Gerontologist 2023 Mar 21;63(3):439-450. [doi: 10.1093/geront/gnac157] [Medline: 36239054]
51. Yakovchenko V, Lamorte C, Chinman MJ, Goodrich DE, Gibson S, Park A, et al. Comparing the CFIR-ERIC matching tool recommendations to real-world strategy effectiveness data: a mixed-methods study in the Veterans Health Administration. Implement Sci 2023 Oct 12;18(1):49 [FREE Full text] [doi: 10.1186/s13012-023-01307-x] [Medline: 37828539]

XSL•FO
RenderX

Abbreviations

CFIR: Consolidated Framework for Implementation Research
CME: Continuing Medical Education
ERIC: Expert Recommendations for Implementing Change
PCP: primary care physician
RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance
RF: Reference Framework for Preventive Care for Older Adults
SRQR: Standards for Reporting Qualitative Research

Edited by S Brini; submitted 10.Jun.2025; peer-reviewed by P Vatcharavongvan, K Mouloudj; comments to author 02.Oct.2025; revised version received 23.Oct.2025; accepted 23.Oct.2025; published 04.Mar.2026.

Please cite as:
Zhong CC, Chen M, Lo CY, YIM MK, Zhang X, Wong WCW, Huang J, Wong MCS
Implementation of Reference Framework for Preventive Care for Older Adults: Cross-sectional Qualitative Study
J Med Internet Res 2026;28:e78814
URL: https://www.jmir.org/2026/1/e78814
doi:10.2196/78814
PMID:

©Claire Chenwen Zhong, Mingtao Chen, Chung Yi Lo, Man Kin YIM, Xiaoshu Zhang, William CW Wong, Junjie Huang, Martin CS Wong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Institutionalizing Digital Parenting Programs in Low Resource Settings in China: Comparative Case Study of Health Care and Education Sectors Using the RE-AIM Framework

Xinyu Shi[1], BA; Ruochen Ruan[2], MEd; Yi Qie[3,4], MA; Jamie M Lachman[4,5,6], PhD; Na Zhong[7], MA; Zuyi Fang[8,9], PhD

[1]School of Government, Beijing Normal University, Beijing, China

[2]Faculty of Education, Beijing Norma University, Beijing, China

[3]China Development Research Foundation (CDRF), Child Development Research Institute, Beijing, China

[4]Department of Social Policy and Intervention, University of Oxford, London, United Kingdom

[5]Centre for Social Science Research, University of Cape Town, Rondebosch, Cape Town, South Africa

[6]Parenting for Lifelong Health, Barnett House, Oxford, United Kingdom

[7]Chengbei Preschool, Xinyu, China

[8]Institute of Population Research, Peking University, No.5 Yiheyuan Road, Haidian District, Beijing, China

[9]Department of Social Policy and Intervention, University of Oxford, Oxford, United Kingdom

Corresponding Author:
Zuyi Fang, PhD
Institute of Population Research, Peking University, No.5 Yiheyuan Road, Haidian District, Beijing, China

Abstract

Background: Digital parenting programs offer a promising way to disseminate evidence-based parenting knowledge and support early childhood development. They help reduce costs while improving scalability and fidelity. However, their successful implementation is context-dependent, and existing research offers limited guidance on how the implementation of digital parenting interventions unfolds across diverse settings.

Objective: This study aims to identify the shared and unique facilitators and barriers affecting each dimension of implementation, as well as differentiated mechanisms that support the effective implementation and institutionalization of such interventions across diverse settings.

Methods: Using a multiple-case study design, this research compared the implementation of a digital (chatbot-led) parenting program across 2 distinct settings in China: urban educational and rural health care contexts. The intervention content remained consistent, while the contexts and formats of local human-led support differed. Guided by the RE-AIM framework, this study examines the program's reach, adoption, implementation, and maintenance in both settings. Data sources included program documents, field observations, semistructured interviews, and focus group discussions with 83 stakeholders. Thematic analysis was conducted using ATLAS.ti until thematic saturation was reached.

Results: Data were collected from 83 stakeholders, and findings are based on an analysis of 18 interviews and 4 focus groups with caregivers, village doctors, and health officials from the rural health care setting, and 29 interviews and 4 focus groups with caregivers, teachers, social workers, and managers from the urban educational setting. Regarding reach, strong relationships between parents and implementers and the credibility of program developers were shared facilitators in both settings. Parenting conservatism and limited understanding of the program were shared barriers. In rural health care settings, parents' perception of village doctors as lacking parenting expertise posed an additional challenge. For adoption, trust between managers and program developers, program alignment with organizational functions, and organizational empowerment supported implementation are shared facilitators in both settings. At the individual level, task-driven motivation helped, while time constraints hindered adoption in the health care setting. Teachers adopted the program due to its relevance to their roles in the educational setting, unlike village doctors, who did not see it as part of their core duties. For implementation, supportive management and clear guidelines were shared facilitators in both settings, while a lack of purpose and psychological pressure acted as barriers. Rural implementation was aided by scheduling during off-seasons and standardized workflows, whereas flexible workflows were essential in the educational setting. Regarding maintenance, alignment with organizational functions and internal resources facilitated sustainability in both settings, while overreliance on government authorization posed challenges. Educational settings required contextual adaptation, while health care settings needed more content adaptation.

XSL•FO
RenderX

Conclusions: Implementing digital parenting programs is a complex process, influenced by multilevel facilitators and barriers that vary across regions (rural vs urban) and settings (educational vs health care). This study highlights the importance of context-specific implementation strategies and proposes differentiated delivery models tailored to local structures and needs.

(J Med Internet Res 2026;28:e79848) doi:10.2196/79848

KEYWORDS

digital delivery; parenting program; implementation science; Reach, Effectiveness, Adoption, Implementation, and Maintenance; RE-AIM; child development

Introduction

Early childhood is a critical period for long-term educational, emotional, and economic development [1]. Globally, more than 43% of children younger than 5 years of age are at risk of not reaching their full developmental potential, with children in low- and middle-income countries (LMICs) being disproportionately affected [2]. The failure to achieve developmental potential during early childhood has far-reaching consequences, not only for individual well-being but also for societal progress, which leads to losses in human capital and increased risks of social instability [3]. These challenges underscore the critical importance of investing in effective strategies that support ECD.

Among the most influential predictors of positive ECD outcomes is the quality of parenting [4,5]. Improving parenting practices, especially in under-resourced settings, is therefore a key priority in global ECD efforts. Parenting interventions, which are typically grounded in social learning and attachment theories, aim to equip parents with the knowledge and skills necessary to enhance parenting quality and to foster home environments that promote children's health and development [6-8]. Such programs have been shown to improve ECD outcomes, enhance positive parenting practices, reduce the use of harsh or violent discipline, and improve caregivers' mental health [9,10]. Meta-analytic findings indicate that such interventions in LMICs yield greater improvements in children's cognitive, language, and motor development, highlighting the potential for large-scale benefits in resource-limited settings [10].

Traditional parenting interventions, delivered through home visits or small-group sessions, are resource-intensive and face barriers such as time constraints, geographic distance, and limited personnel [3,11]. These limitations make conventional delivery methods difficult to scale and sustain, particularly in low-resource settings [12,13]. Nowadays, the growing accessibility of the internet and the widespread ownership of mobile phones present a promising opportunity to overcome these challenges by leveraging digital technology [14]. Digital parenting programs provide a cost-effective way to deliver evidence-based parenting support, particularly in low-resource settings, by enhancing scalability while maintaining high fidelity [12,15]. They reduce costs associated with time and travel and increase accessibility to high-quality parenting resources [16,17]. Emerging evidence indicates that digital parenting interventions can positively influence multiple outcomes, including child development, parental confidence, parenting stress, and children's behavioral problems [18-20]. These programs are particularly effective when complemented by human-led support, either in person or remotely [21].

The success of social and behavioral interventions hinges not only on the inclusion of key components designed to facilitate behavior change, but also on the fidelity and quality of their implementation in real-world settings [22]. While the body of research demonstrating the effectiveness of digital parenting interventions in promoting child development is growing, limited attention has been given to their implementation processes. This gap in knowledge poses significant barriers to the replication, adaptation, and scale-up of such interventions, particularly in diverse and resource-constrained contexts [23,24].

Implementation research seeks to uncover what works, how it works, and why it works in everyday settings beyond the confines of controlled trials [25]. It emphasizes identifying contextual factors that influence successful implementation, such as characteristics of the delivery environment, community context, interorganizational partnerships, and intervention delivery mechanisms [26,27]. The RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) framework is a widely recognized model that offers a comprehensive approach for evaluating the implementation and public health impact of interventions [28]. It helps researchers and practitioners identify the strengths and barriers in intervention implementation, thereby facilitating the optimization and broader dissemination of public health practices [28-30]. This approach also offers insights into how short-term programs can be transformed into sustainable, institutionalized practices within organizations and service systems [31]. While existing studies applying the RE-AIM framework have examined digital interventions in various health domains, such as diabetes management, psychotherapy, vaccination, and health behavior promotion [32-35], there is a notable gap in research on the implementation of digital parenting programs, especially across different settings. This gap is especially evident in LMICs, where such interventions remain underexplored and underdeveloped.

Guided by the RE-AIM framework, this study addresses these gaps by evaluating the implementation of a digital parenting program, "Keyushiguang", in 2 distinct service delivery systems in China: an urban educational setting and a rural health care setting. In China, where significant early childhood development gaps and resource disparities persist, there is an urgent need for scalable, context-sensitive parenting support models [36,37]. This study, therefore, examines the program's reach, adoption, implementation, and maintenance to identify shared and setting-specific facilitators and barriers. By comparing these 2 delivery models, the study aims to provide new insights into

XSL•FO
RenderX

the differentiated mechanisms required for scaling and sustaining digital parenting support in diverse, low-resource contexts. The program's impact evaluation will be reported separately.

Methods

Study Design and Context

A multiple-case study approach was used to facilitate a deeper understanding of individual cases while also identifying overarching patterns. A case study is an empirical research method used to investigate contemporary phenomena within their real-world context. By analyzing a specific case in depth, researchers gain insights into the background, processes, and influencing factors, which is especially valuable for studying complex social phenomena or contexts [38]. Case studies can be categorized into single-case and multiple-case studies, depending on the number of cases analyzed. The multiple-case study approach was adopted to enhance the validity of findings from different cases, leading to more robust and reliable conclusions [39]. This method allowed us to identify commonalities and differences in the facilitators and barriers to implementation across different contexts.

Theoretical Underpinnings: RE-AIM

This paper focuses on 4 key dimensions of the RE-AIM framework: Reach, Adoption, Implementation, and Maintenance. Effectiveness was assessed using quantitative methods and will be presented separately.

Reach refers to the total number, proportion, and representativeness of individuals who are willing to participate in a given initiative, intervention, or program [28]. It primarily focuses on understanding the factors that influence individuals' decisions to engage—or not engage—with the program. Adoption refers to the absolute number, proportion, and representativeness of settings and intervention agents (people who deliver the program) who are willing to initiate a program, as well as the reasons behind their decisions [28]. Implementation primarily focused on the program's fidelity, which is the extent to which the program is delivered as intended, and examined how this fidelity was achieved throughout the delivery process [40]. Maintenance at the setting level of a program refers to its institutionalization or integration into routine practices and policies [30].

This paper evaluated each of these dimensions across the 2 distinct settings, analyzing the factors that were perceived to facilitate or hinder participation in the programs, as well as the differences in these factors between the 2 settings.

We also reported a descriptive overview of reach, adoption, and implementation. Reach was calculated by the ratio of participants who consented to participate relative to the total number of eligible individuals in both settings. The recruitment criteria for both programs were similar, with the primary requirement being that the caregivers of children be the primary caregivers. Adoption at the staff level is calculated by dividing the number of staff who agreed to participate in program implementation by the total number of staff eligible to participate in the program within the organization. In addition, we assessed the implementation based on the fidelity checklists of implementers.

Case Selection

The selection of cases in this research was guided by the study objectives. This paper examines the implementation of the digital (chatbot-led) parenting intervention in 2 distinct settings in China: an urban preschool (educational) setting and 2 rural health centers (health care) settings. Intervention delivery in both settings was led by local implementation partners, with support from the research team that developed the digital intervention. Consequently, the intervention's digital delivery method and content remained highly consistent across the 2 settings, with only minor adaptations made for the rural health care setting, such as colloquial expressions used, without altering the core parenting principles.

Case boundaries were defined as follows. The urban educational case referred to the implementation within one preschool in Xinyu city, Jiangxi Province, while the rural health care case referred to the implementation within the village-level public health system, supported by the county-level Maternal and Child Health Department and 2 township-level Health Centers in Huining County, Gansu Province. The analysis covered the full pilot implementation period: the preschool intervention that ran from March 2024 to June 2024, and the health care intervention that ran from October 2024 to January 2025. Stakeholders included in the case analysis were those directly engaged in program implementation—caregivers, teachers, social workers, village doctors, program managers, and local government officers. Broader community members who were not directly involved in program delivery were excluded.

The digital parenting program "Keyushiguang" was culturally adapted from the Parenting for Lifelong Health's ParentText chatbot for parents of children aged 2 to 9 years by a local research team [41]. Delivered via a rule-based chatbot on WeChat, the program covers 8 key parenting topics, each consisting of 3 to 6 modules depending on the child's age. Each module includes an introduction, quizzes, core parenting tips (in video or text format), and home exercises. The chatbot sends one parenting module every 23.5 hours, allowing parents to gradually develop their parenting skills and apply them in daily situations. Designed to integrate smoothly into parents' daily routines, the chatbot enables flexible engagement at convenient times, such as after work, during breaks, or before bedtime. This gradual approach encourages parents to apply the learned skills in real-life scenarios, such as managing child behavior or establishing daily routines. By offering relevant, practical, and easily accessible support, the chatbot helps parents stay engaged consistently, enhancing the program's long-term impact (Detailed program description can be seen in Multimedia Appendix 1).

The main distinctions between the 2 implementations lay in the format of additional human-led parenting support and were driven by the unique implementation contexts and available resources in each setting. In the urban educational setting, chatbot-led digital delivery was supplemented by message-based WeChat group interactions via web, held once or twice per week, and facilitated by headteachers and trained social workers.

XSL•FO
RenderX

In contrast, the rural health care setting incorporated biweekly 40-minute individual home visits conducted by village doctors to reinforce the content delivered online and foster engagement.

The choice between group-based and individual-based human-led parenting support models was based on a comprehensive assessment of the target population and implementation environment [42]. In China, rural areas typically face greater challenges to child development than urban areas [43]. Therefore, the program was designed as a universal intervention (available to all families regardless of child development status) in the urban setting, while in rural areas, it was implemented as a selective intervention (targeted at high-risk families). Additionally, resource feasibility and cultural acceptance also informed program design. For example, most rural families have access to village doctors, whereas urban preschools face teacher shortages. In addition, rural parents are more receptive to home visits [44]. Therefore, we adopted a group-based WeChat interaction model for the urban preschool setting and a combined individual and group home visit model for the rural health care setting to better suit local needs and contexts. The comparison of the 2 pilot implementations is provided in Multimedia Appendix 2.

The rationale for selecting these 2 cases was that by controlling for similarities in program content, the study could minimize the impact of content-related differences and more precisely analyze the result variations arising from differing implementation contexts (ie, urban education vs rural health care settings).

Data Collection

We used data triangulation to enhance the comprehensiveness and validity of our findings. Multiple sources of data were used, including program documentation (such as program manuals and materials), field observation notes (such as the behaviors of stakeholders in the process of program pre-preparation, implementation, and evaluation), semistructured individual interviews, and focus group discussions (FGDs). Our triangulation involved 2 levels. First, we conducted within-case triangulation by integrating these diverse data sources to validate findings within each of the 2 settings. For example, we compared implementers' subjective perceptions of their implementation fidelity with parents' feedback (as reflected in the fidelity checklist) and triangulated these with our field observation records. This approach allowed us to more accurately interpret and evaluate the perspectives expressed by stakeholders. Following this, we conducted between-case triangulation (ie, a cross-case analysis). This second level systematically compared the validated themes from the urban educational setting against those from the rural health care setting to identify the overarching facilitators, barriers, and divergent patterns presented in our results.

Program-related documents retained since the program's inception were collected and analyzed. These included program proposals, curriculum designs, daily implementation records, and training materials, which served as foundational resources for understanding the intervention. All documents were reviewed systematically, and key implementation details were extracted for analysis. In addition, as members of the research team, the authors were actively involved in and directly observed the implementation processes in both settings, providing them with in-depth insights into the contextual dynamics and operational realities of the program delivery.

Field observations related to stakeholders' behaviors, program structure, and implementation processes. We observed the behavior and actions of stakeholders throughout the program and recorded those, which allowed for a more accurate understanding of the viewpoints they expressed.

We adopted a purposive sampling strategy, which is consistent with qualitative research practices that emphasize differentiation (Figure 1). Participants were deliberately selected to reflect diversity in roles, experiences, and contextual backgrounds. This approach allowed us to explore the phenomenon under investigation in a more nuanced and comprehensive manner. Specifically, participants were categorized into 5 key stakeholder groups: program officers, local leadership, local program managers, program implementers, and program participants. The final sample included 1 program officer from the health care setting; 7 local program leaderships (ie, 3 preschool leaderships and 4 government officers from the health care setting); 3 local program managers (ie, 2 Women's and children's health workers and 1 local program staff from the health care setting); 36 program implementers (ie, 9 preschool teachers and 7 social workers from the preschool setting, and 20 village doctors from the health care setting); and 36 program participants (ie, 26 caregivers from the preschool setting; and 10 caregivers from the health care setting).

The figure 1 illustrates the overall structure of data collection. As detailed in the Methods section, the implementation and data collection were conducted sequentially (urban case first, followed by the rural case). FGD: focus group discussion; GO: government officers from the healthcare setting; LPC: local program coordinator; PM: preschool managers; PO: program officer; PP: program participants; PT: preschool teachers; SI: semistructured interview; SW: social worker; VD: village doctor; WACHW: Women's and Children's Health Worker.The interview and FGD guides were developed based on the RE-AIM framework, incorporating a structured set of open-ended questions. Data related to the "Reach" dimension were primarily obtained from program participants, while "Implementation," "Adoption," and "Maintenance" were explored through interviews and focus groups with program donors, local leadership, local program managers, and program implementers. Each FGD lasted approximately one hour, and individual semistructured interviews averaged 40 minutes in duration. Prior to data collection, informed consent was obtained from all participants. With participants' permission, all interviews and FGDs were audio-recorded. Recordings were subsequently transcribed, reviewed for accuracy, and anonymized to ensure confidentiality.

XSL•FO
RenderX

Figure 1. Stakeholder sampling and data collection design for the 2 cases.

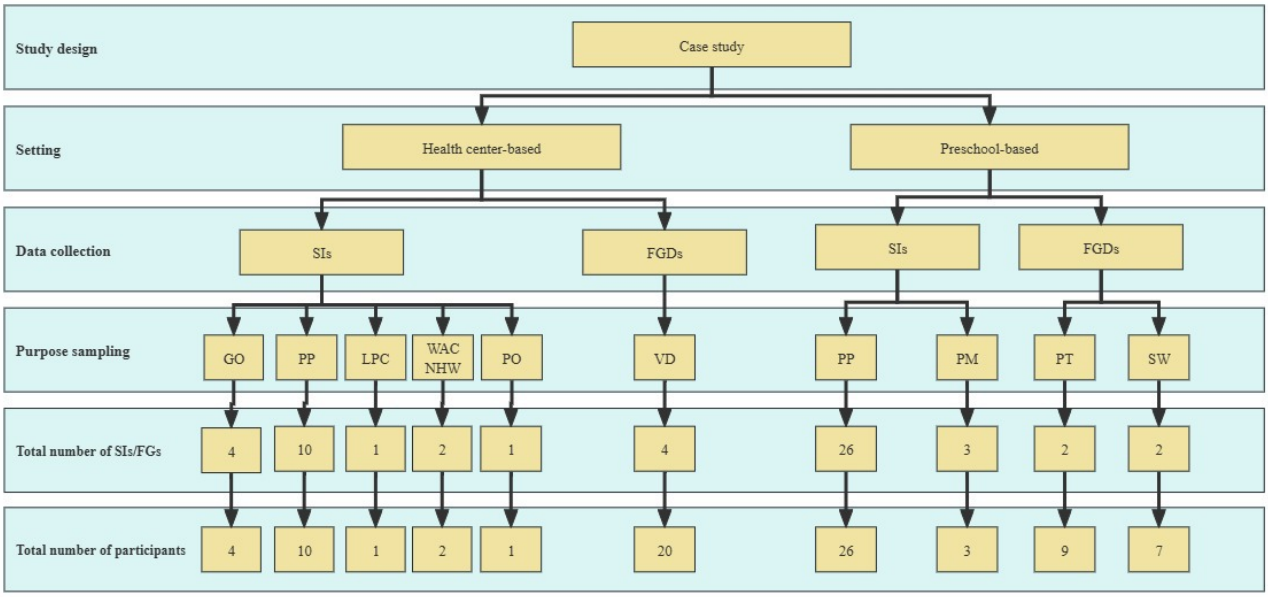

Data Analysis

Data were coded using ATLAS.ti (version 8; Lumivero). A thematic analysis was conducted using a hybrid inductive-deductive approach to identify patterns across the dataset [45]. As data collection was conducted at different times, we first collected and analyzed data from the urban educational setting following a stepwise thematic analysis procedure. After completing data collection in the rural health care setting, we applied the same analytical process. To ensure consistency and mitigate the risk of bias from this sequential design, the same coders applied the same predefined RE-AIM framework to both datasets independently. Only after both independent analyses were complete were the results from the 2 settings formally compared to identify similarities and differences. The coding process was conducted by 2 researchers (XS and RR) and followed the steps outlined by standard thematic analysis [45,46]. Thematic analysis was conducted through a structured 5-step process—data familiarization, initial coding, theme development, collaborative review, and final theme definition—guided by the RE-AIM framework to identify facilitators and barriers to program implementation (Figure 2). During the coding process, we continuously compared new data with existing codes and themes. We considered thematic saturation to be achieved when no new codes or themes emerged from subsequent interviews and FGDs. Given the diversity of stakeholders included—parents, teachers, village doctors, local managers, and government officers—the data provided a comprehensive reflection of perspectives across both settings.

Finally, as the capstone of our cross-case analysis, we synthesized the shared and setting-specific facilitators and barriers identified through the RE-AIM framework. This synthesis allowed us to construct the differentiated implementation models presented as the final component of our Results.

XSL•FO
RenderX

Figure 2. The process of data analysis.

XSL•FO
RenderX

Ethical Considerations

Ethics approvals were obtained from Beijing Normal University (approval numbers: SSDPP-HSC-2024003 and BNU202410100062) and the University of Oxford (SPI_DREC_24_006). We provided participants with a detailed introduction to the program and obtained their informed consent voluntarily. All data is stored in a secure database and is used only for research purposes by the researchers.

Positionality Statement

We acknowledge that our dual roles as program designers and researchers may have influenced data interpretation. To mitigate this, we applied both personal and methodological reflexivity throughout the study.

Our research team has diverse academic backgrounds, including evidence-based social intervention, child development and protection, public administration, education, and social work. These perspectives shaped how we approached the research: team members with social intervention and public administration backgrounds tended to focus on macro-level structures and policies, while those with child development, education, and social work backgrounds were more attuned to participants' lived experiences. To minimize potential biases from these disciplinary differences, the team engaged in ongoing reflexive discussions, considering multiple perspectives at all stages of analysis.

Methodologically, 2 researchers (XS and RR) independently coded different portions of the dataset and cross-checked each other's work. A third researcher (ZF) reviewed the coding, provided feedback, and helped resolve discrepancies through iterative discussion until consensus was reached. This collaborative process ensured rigor and consistency in the analysis.

In addition, practical measures were taken to reduce researcher influence during data collection. Data collectors were trained to use semistructured interview guides with open-ended, nonleading questions, fostering a neutral environment that encouraged participants to share their authentic experiences.

Results

Overview

To ensure methodological rigor and transparency, the reporting of this study followed the iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations (Checklist 1) [47]. Descriptions of interview participants can be seen in Multimedia Appendix 3.

Case Context and Stakeholder Profiles

The Urban Educational Setting: A Preschool in Xinyu City

This pilot was implemented in a preschool located in Xinyu City, an urban, medium-income area in eastern China. The preschool provides early childhood education before primary school, with parenting integrated into daily activities. The main participants were parents, most of whom hold stable jobs; their participation time was influenced by their work schedules. Key implementers included teachers who hold educational credibility and trained social workers who offered professional parenting support and assisted teachers in carrying out the program. The preschool leadership's primary motivation was to enhance the school's educational quality and reputation.

The Rural Health Care Setting: 2 Townships in Gansu Province

This pilot took place in health centers across 2 townships in Huining County, Gansu Province, a low-income rural region in western China. The centers are responsible for public health services, and promoting early childhood development was one of their key goals. The main participants were parents recruited from rural communities, most of whom engage in agricultural work. Their participation was shaped by the agricultural calendar and seasonal migration for work. The key implementers were village doctors who were trusted in their communities with strong local knowledge but generally older and less experienced in parenting. The health care leaderships were primarily motivated by fulfilling national public health directives. We labeled the interview data based on the source and role of the interviewees. The initial letters in each label represent the interviewee's role, as illustrated in Figure 1. HB and SB denote data from the hospital-based program and preschool-based program, respectively. The last 3 digits of each label indicate the document number, and each number corresponds to a unique interviewee or focus group.

Reach—Facilitators and Barriers

Overview

In total, 541 (81.4%) caregivers (402 [74.3%] were mothers, 135 [25%] were fathers, and 2 [0.3%] were grandparents) of the 665 eligible families consented to participate in both the program and the research.

In contrast, the township central health center–based program aimed to recruit 120 parents of children aged 24 - 59 months from 2 pilot townships, out of approximately 795 eligible families. A total of 83 out of the targeted amount of 120 were able to participate (69.2%).

Shared Facilitators to Reach

The main motivations for participation in the program include its strong emphasis on child development, strong relationships between program implementers and caregivers, the perceived credibility of the program developers, and the convenience of learning through a chatbot.

A primary motivation in both settings was a recognized need for practical parenting skills. In the rural health care settings, caregivers expressed that they faced ongoing challenges in daily parenting and felt that their current knowledge was insufficient. They recognized the need to learn practical parenting skills to better support their children. In the urban educational setting, parents expressed similar concerns.

> *We don't know much about how to support our children's growth or deal with their emotional problems. This program gives us a chance to learn parenting ways.* [PP, HBD05]

XSL•FO
RenderX

Another shared facilitator was the influence of implementers, driven by social relationships. Some caregivers in the health care setting reported that their agreement to participate was influenced by their trust and social pressures from program implementers, such as face-saving norms, social obligations, or a sense of personal favor. Similarly, in the educational setting, the motivation to participate also stemmed from the influence of teachers.

> *The village doctor is our family doctor. He has helped us a lot and has a good relationship with us. He wants us to join the program, so we are willing to support him.* [PP, HBD11]

> *At first, the teacher encourages us to participate in this parenting program. I think if other children are participating, we shouldn't seem uncooperative with the teacher's efforts.* [PP, SBD11]

Parents in the health care setting also considered the background of the program developers. Knowing that the program was developed by a research team from a prestigious university in China enhanced its credibility and increased parents' perceived benefits. Similarly, parents in the educational setting shared the same view, trusting the program's credibility due to its association with a reputable institution, which further reinforced their belief in the program's potential benefits.

> *There is too much parenting information online, and it is often hard to tell what is true or false. But this online course is developed by a professional team, so we feel the content is reliable and will be helpful to me.* [PP, HBD10]

The convenience of learning was also an essential factor for program reach. Parents in both educational and health care settings appreciated that the chatbot-based delivery allowed them to access the content anytime and anywhere. The use of diverse formats, such as videos and text, accommodated different learning environments and individual learning preferences.

> *The timing is fine because it is flexible. You can choose when to learn, so it is convenient when you have free time.* [PP, SBD15]

> *The course offers both video and text options, which is suitable for different scenarios. When I have free time, I choose to read the text. When I am busy, I play the video and listen to the content.* [PP, HBD11]

Shared Barriers to Reach

In both settings, shared factors hindering parents' participation include parenting conservatism, a lack of understanding of the program, and limited time for learning.

A shared barrier in both settings was 'parenting conservatism,' where parents preferred to stick to traditional parenting methods. During the program promotion process, some parents in the educational setting with low engagement exhibited a form of 'parenting conservatism'. They preferred to maintain their existing parenting practices and felt it was unnecessary to learn about other approaches, believing that new practices might not be suitable for their own situations. Similarly, in the health care setting, some caregivers also displayed resistance to adopting new parenting knowledge, preferring to rely on their established practices.

> *I feel that every family has its own parenting style. My other children are raised according to my own parenting methods, and there is no need to change it.* [PP, SBD13]

Furthermore, a limited understanding of the program served as a barrier in both contexts. Some caregivers in the educational setting indicated that they did not fully understand the program, such as its content, learning format, and flexibility. Some caregivers claimed they had no spare time to engage with the digital program and chose not to participate, unaware that the learning sessions were brief and could be easily completed anywhere in just a few minutes.

> *I think the course will take up a lot of time. Since I have to work and take care of my family, I don't have much spare time for studying.* [PP, SBD13]

While some parents in the health care setting remained unaware of what they could gain from the program, this ultimately reduced their enthusiasm for participating.

> *At first, they don't tell me the specific details of the program. I don't know its benefits, so I initially see it as just a favor to village doctors.* [PP, HBD13]

Distinct Facilitators and Barriers to Reach

The urban educational setting presented several unique facilitators. These included urban parents' recognition of the parenting values conveyed by the program and their acknowledgment of the implementers' credibility and rationale in delivering parenting knowledge. These factors played a significant role in enhancing parental engagement in the educational setting. However, such facilitators were not observed in the health center-based program.

Some parents in the urban areas reported that the parenting concepts and attitudes conveyed through the program resonated with their own parenting beliefs. In contrast, parents in rural areas rarely expressed this kind of value-level recognition.

> *I agree with the parenting concepts in the course, and I am willing to learn it.* [PP, SBD08]

The perceived professional authority of the program implementers influenced caregivers' decisions to participate. Parents viewed preschools as appropriate platforms for parenting support, and headteachers were seen as credible and persuasive figures in promoting parenting knowledge.

> *This preschool and its teachers are the best in the area, and I trust the quality of both the preschool and the teachers. I believe their actions will be beneficial to me.* [PP, SBD18]

In contrast, the rural setting faced a unique barrier related to the implementers' professional role. In rural areas where village doctors delivered the parenting program for the first time, some doctors noted that because their routine work is not typically associated with parenting, parents were skeptical about their true intentions behind offering such services.

XSL•FO
RenderX

> *The parents doubt why I invite them to join the parenting program. They ask if I have received any benefits from the program organizers.* [VD, HBD25]

The summary of barriers and facilitators to reach can be seen in Multimedia Appendix 4.

Adoption—Facilitators and Barriers

Overview

At the setting level, since the program was conducted as a pilot, 1 urban preschool and 2 rural health centers agreed to participate upon invitation. At the staff level, in the preschool-based program, the primary implementers included headteachers and recruited social workers, all of whom joined the program in response to directives from their supervisors. In the health center-based program, the main implementers were 2 women's and children's health workers and 30 village doctors. Eventually, both maternal and child health workers participated, while 23 village doctors (76.7%) remained engaged, and 7 withdrew from the program. Overall, the adoption level of the program was relatively high.

Shared Facilitators for Adoption

At the setting level, relevant information was gathered through interviews with local leadership. The main shared factors that promoted adoption included trust in the program developer, alignment with organizational functions, and empowerment of the organization and staff.

A primary factor in facilitating successful cooperation and adoption in both settings was trust in the program developers. Leadership in the educational setting acknowledged that the developers' affiliation with a prestigious university in China enhanced their perception of the program's credibility and value. In the health care setting, management also demonstrated trust in the collaborating research team.

> *You are a research team from a top university in China, and the quality of what you developed is guaranteed. We also have confidence that parents will accept the program, and we believe it will have a positive impact.* [PM, SBD33]

Another shared facilitator was the importance of the alignment between the program and the implementing organization's core functions. In the educational setting, leaders reported that before adopting the program, they would assess whether their organization possessed the capacity and resources to support implementing the program and whether the initiative aligned with their institutional mandate. For instance, preschools regarded parenting services as a fundamental aspect of their work, viewing it as consistent with their educational responsibilities.

> *This is something I want to do. In 2010, our preschool established an 'Early Education Center,' and parenting is a part of our work.* [PM, SBD32]

Likewise, in the rural setting, health centers considered parenting support closely related to child health care and broader public health objectives.

> *This online-based parenting program focuses on promoting health education, especially concerning children's growth and development. This is also connected to our daily work in public health.* [GO, HBD32]

Local leadership in both settings also highlighted its potential to empower staff and foster organizational growth. In the educational setting, managers noted that the program could enhance headteachers' professional skills in parenting, positioning it as a valuable resource for supporting parents.

> *Our teachers gain many benefits during the implementation of the program, such as learning about family education and strengthening their connection with parents.* [PM, SBD30]

In the health care setting, leaderships emphasized that the program could strengthen the capabilities of village doctors as well as maternal and women's and children's health workers in delivering comprehensive children's healthcare services.

> *Our county pays great attention to the implementation of early childhood development work in rural areas. Involving village doctors and other staff in this parenting program allows them to gain experience in carrying out such work.* [GO, HBD27]

At the staff level, both task-driven (motivation that arises from external obligations to complete tasks assigned by others, such as supervisors or program coordinators) and intrinsic motivation (motivation from personal interest or intrinsic desire) were key shared factors promoting adoption. For example, in the health care settings, implementers reported that the program was perceived as a mandated task from higher-level leadership, who required them to carry out the program and follow the specified timeline and procedures.

> *Our leaders attach great importance to this program and repeatedly emphasize the need to cooperate with the program developers to complete it. We feel that this is a very important task for us.* [VD, HBD24]

Intrinsic motivation also played an important role. Some implementers in the health care setting expressed that they recognized the significance of this pilot and the potential benefits it could bring to caregivers. This sense of mission motivated them to actively adopt and support the program. In the educational setting, teachers also felt that the program could support the development of both parents and children, contributing to better educational outcomes.

> *Another reason I participate in this program is that I feel what we are doing is meaningful. In our rural area, such resources are very precious, and many parents also have a demand for parenting knowledge. By completing this program, we can provide help to children and their families.* [VD, HBD31]

Shared Barriers to Adoption

A major shared factor hindering adoption in both settings was the perceived difficulty of implementation due to limited time. Both implementers and local leadership expressed concerns that implementers were already occupied with their regular

XSL•FO
RenderX

responsibilities, which consumed most of their time. The entire program's especially tight schedule further compounded the issue, as they were expected to complete tasks within a short timeframe. This increased the pressure on implementers, making them feel even more stressed.

> *Our village doctors have a lot of work and are very busy. We have to both prescribe medical treatment and complete public health tasks such as chronic disease management and vaccination. The program requires us to complete the tasks within a short time, and it feels very stressful.* [VD, HBD25]

> *We also have a lot of teaching tasks at school to complete, and the time is not very sufficient.* [PT, SBD34]

Distinct Facilitators and Barriers to Adoption

At the staff and setting level, prior successful collaboration experiences served as a unique facilitator for the health centers' adoption of the program. However, health center-based implementers often lacked familiarity with parenting knowledge, and some faced difficulties in recruiting appropriate participants. The perceived difficulty in recruitment hindered program adoption among village doctors. Moreover, remuneration emerged as a significant motivating factor for implementers in the health care setting. In the educational setting, the teachers' professional roles in parenting acted as a key facilitator.

At the institution level, factors promoting adoption differed between the 2 contexts. In the health center-based setting, health care leaders and program officers highlighted that prior successful collaboration experience played a key role. The local government, supported by the donor, had previously implemented similar ECD parenting programs in other townships within the county, which helped establish mutual trust and efficient working mechanisms.

> *They agree because we have previously implemented an early childhood development program here. We are familiar with the local leaders and institutions, and have built a strong trust relationship.* [PD, HBD35]

At the staff level, the relevance of the task to the implementers' professional roles directly influenced their motivation to adopt the program. In the urban preschool setting, headteachers reported that parenting support was closely tied to their core duties, contributing to child development and the enhancement of educational quality. This strong connection fostered greater intrinsic motivation among them.

> *We also provide some parenting education in daily work. I review this program and feel it is quite good. If parents attend, it will benefit them.* [PT, SBD34]

In contrast, some village doctors and maternal and child health workers in the health care setting felt their routine work was less directly related to parenting. Consequently, they viewed the program more as a task assigned by higher-level leadership, with task-driven motivation playing a prominent role.

> *We have not done similar work before. I believe it is not our duty, as we are mainly responsible for public health. The main reason for implementing this program is that it is assigned by our leadership.* [VD, HBD24]

Some village doctors also reported difficulty in identifying parents who met the program's participation criteria and were able to join. This challenge led to frustration and, over time, caused some of them to gradually withdraw from participating in the program altogether.

> *The main challenge for me is that I really cannot find suitable parents, because many children in the village are migrant or left-behind children.* [VD, HBD30]

Additionally, village doctors noted that the remuneration provided by the program officers was an important incentive for them to complete their responsibilities.

> *We carry out the task as required and receive some compensation. If it is voluntary, I will not be willing to complete the task.* [VD, HBD24]

The summary of barriers and facilitators to adoption can be seen in Multimedia Appendix 5.

Implementation—Facilitators and Barriers

Overview

The investigation of implementation primarily focused on the program's fidelity, which is the extent to which the program is delivered as intended, and examined how this fidelity was achieved throughout the delivery process [40].

The program was delivered with high fidelity in both settings, as demonstrated by fidelity checklists. In the preschool setting, headteachers recruited parents and reminded them to complete the online modules, while 10 social workers facilitated group discussions and completed all assigned tasks. In the health care setting, village doctors conducted home visits, supported by maternal and child health workers. Follow-up phone calls with 74 parents confirmed the content of the visits. The village doctors achieved an average fidelity score of 7.81 out of 8, indicating strong adherence to the program protocol.

Shared Facilitators for Implementation

Several key shared factors were identified as promoting the fidelity of program implementation: adequate onboarding training, a supportive management system, timely external support, and clear work guidelines.

Adequate onboarding training was a shared facilitator in both settings. Implementers in the health care setting believed that the training helped them understand the parenting principles, tasks, potential challenges, and proposed solutions. They also noted that the time allocated for discussion created opportunities to express concerns, ask questions, and receive prompt responses. Similarly, in the educational setting, teachers and social workers affirmed that the training enabled them to understand their respective tasks and effectively contribute to the program's implementation.

> *The training before the program starts is essential. It helps us understand what we need to do, and the*

XSL•FO
RenderX

> *subsequent discussion sessions also address some issues.* [VD, HBD24]

A supportive management system involved a balanced combination of appropriate incentives and supervision to facilitate the successful implementation of the intervention. In both health care and educational settings, local leadership perceived that, to enhance fidelity, implementers required suitable incentives that align with their expectations, including both material and intrinsic (spiritual) rewards.

> *After all, the main work of village doctors is not parenting. They have limited time and energy, so they need to be compensated.* [GO, HBD27]

> *Our teachers need to have a sense of achievement when doing this program. Even giving them a certificate is an effective encouragement for them.* [PM, SBD32]

Meanwhile, while supervision was crucial in both contexts, its focus differed. Local program managers in the health care setting emphasized that supervision was essential to achieve a balance between quality control and task implementation efficiency. They believed that supervision should cause necessary work pressure while allowing sufficient flexibility in how tasks are carried out. In the educational setting, the focus of supervision was more on ensuring timely task completion, with managers highlighting the importance of regularly reminding teachers to meet deadlines.

> *Village doctors work hard. Sometimes, tasks cannot be completed due to external reasons. When supervising their work, we have to both consider the quality and avoid discouraging their motivation.* [WACNHW, HBD26]

Providing timely external support to implementers was perceived to be a critical shared factor for successful implementation, which involves access to expert advice, consultations, and problem-solving assistance from external consultants or organizations when issues arise during program implementation. Implementers in the health care setting reported that timely and adequate support—whether material, psychological, or resource-based—from organizations and program developers helped alleviate their stress, overcome challenges, and enabled them to successfully complete their tasks. Likewise, in the educational setting, implementers affirmed that support from program developers and managers was crucial for the smooth execution of their work.

> *I feel that the support is quite sufficient. You are always there whether we or the village doctors have any questions. I feel touched.* [WACNHW, HBD26]

Finally, having clear work guidelines was a key shared facilitating factor. Implementers in both educational and preschool settings noted that well-defined procedures contributed to smoother implementation by clearly outlining what needed to be done, when, and how. This clarity made it easier for them to carry out their tasks efficiently.

> *The clear work guidelines let me know what I need to do, making it easier to carry out my work.* [VD, HBD31]

> *This program provides clear instructions, and the guidance in all aspects is very detailed. So, we can implement it well according to the guidance.* [SW, SBD27]

Shared Barriers to Implementation

A lack of sense of purpose and psychological pressure was perceived as key shared challenges faced by some implementers in both settings. Psychological pressures refer to stressors such as heavy workloads, tight deadlines, and the emotional burden of ensuring the success of the program. These pressures can negatively impact the quality of program delivery and the well-being of implementers. While leadership demonstrated a clear understanding of the program's goals, some implementers lacked this clarity. During the implementation process, it was observed that certain implementers were unsure about the program's objectives and the value it could provide to caregivers. This uncertainty at times led to self-doubt, which had a negative impact on their motivation. Further communication with implementers confirmed this observation—several social workers from the educational setting and village doctors from the health care setting expressed confusion about the purpose of their roles and the potential benefits their efforts could bring to families.

> *When conducting home visits, we do not know our purpose. What exactly are we trying to do?* [VD, HBD24]

> *You do not know your role in this process, but you must do it. Moreover, at this point, you can feel quite confused.* [SW, SBD28]

Implementers in both settings shared the challenge of generating demand and encouraging active participation in the program, which caused psychological pressure among implementers.

> *Because not everyone is enthusiastic about participating, I think the appeal of the content to parents is crucial. Is the content we are providing useful to parents? Is it something they want?* [SW, SBD27]

Distinct Facilitators and Barriers to Implementation

A unique facilitator in the health care setting was the timing of implementation, which aligned well with the villagers' work schedules. However, in China, village doctors typically only conduct home visits for specific health care tasks. Also, in the health care setting, parenting home visits are irregular, and the long distances they needed to travel posed a challenge during program delivery. Furthermore, while clarity in work procedures was a shared facilitating factor across both settings, there was a notable difference in the need for implementation flexibility across the 2 settings.

First, in the rural setting, village doctors shared that the program was implemented during the off-season for farming, just before the Chinese New Year, a period when there was less agricultural work and many parents who normally worked away from home had returned. This created free time for caregivers to participate in the program.

XSL•FO
RenderX

> *Parents are generally at home in winter. As the weather is cold, there is no farming to do. Besides, they do not go out for labor during this season, so most of them stay home.* [VD, SBD24]

Second, village doctors also reported that the physical distance between households and their base locations created challenges in conducting home visits on time.

> *It is still too far, the road is inconvenient, and it is not easy to return.* [VD, D30]

Third, standardized workflow served as a double-edged sword, with its impact varying across settings. In the educational setting, especially among headteachers, implementers preferred greater flexibility. They expressed a desire to customize token economy systems to encourage daily participation in their own class and requested more adaptable web-based group interaction structures to better address the diverse needs of caregivers.

> *I feel that our online group discussions are a bit stiff. The content shared is fixed and does not capture parents' interest. If we integrate additional parenting support with our daily activities, it will be better.* [PT, SBD35]

In contrast, implementers in the rural health care setting, particularly village doctors who were typically older and had limited experience with parenting support, preferred a highly standardized process. They valued having clear guidance on what tasks to perform, when to perform them, and how to carry them out.

> *The workflow you provided is very clear. It outlines each step, and we just need to follow it to complete the task.* [WACNHW, HBD26]

The summary of barriers and facilitators to implementation can be seen in Multimedia Appendix 6.

Maintenance at the Setting Level—Facilitators and Barriers

Overview

This article primarily explores the factors that may facilitate or hinder the integration of the program into the established service system. In both settings, the leadership actively worked towards promoting the program's integration into the existing service infrastructure.

Shared Facilitators to Maintenance at a Setting Level

The shared factors were the integration of the program into an organization's daily operations, alignment between the program content and the organization's core functions, the availability of sufficient and appropriate internal human resources, and the low cost associated with digital delivery.

First, alignment with the organization's core mission was a key shared facilitator. Local leadership in the health care setting emphasized that the program's alignment with their existing responsibilities was essential for successful institutionalization beyond initial adoption. By supporting the organization's core mission, the program minimized additional costs and streamlined integration into routine work. Likewise, in the educational setting, managers emphasized that the parenting program was a natural fit for their institution's existing work and goals.

> *Normalizing it means integrating it with our public health services, which does not create any work pressure for us or add any burden. It should also benefit our daily work.* [GO, HBD27]

Second, leadership in both settings believed they had access to a well-suited workforce for program delivery. Village doctors in the rural setting and headteachers in the urban setting were described as stable, professionally trained in child development, and capable of ongoing learning. Their regular contact with families and their established trust and authority within the community made them particularly suitable for engaging caregivers and carrying out the program.

> *The village doctor team can handle this program. On one hand, they have good relationships with the villagers. On the other hand, they possess a lot of medical knowledge and have the ability to learn.* [GO, HBD32]

> *Our teachers are well-suited to take on this program. They already handle educational work and can easily use it as a tool in their daily teaching.* [PM, SBD33]

Third, local leadership in the educational setting emphasized that the low cost of digital delivery was a key factor facilitating program integration. By delivering core content via web through a WeChat-based chatbot, the program minimized the need for intensive offline operations. This digital approach not only helped maintain content quality but also significantly reduced the workload and operational costs for implementing organizations, making the program more sustainable and scalable within existing structures. Similarly, in the health care setting, managers highlighted that controllable costs were crucial for sustaining the program in rural areas.

> *I think the advantage of the online format is that, firstly, it can push information in real-time and is more flexible. At the same time, since we are in the information technology era, parents also use their phones often.* [PM, SBD27]

Shared Barriers to Maintenance at the Setting Level

Shared barriers to integrating the program into the organization's daily operations included institutional dependence on higher government authorization, challenges in sustaining staff motivation, and difficulties in generating parental demand for parenting support.

A primary barrier in both settings was the institution's reliance on higher-level government approval. In the rural context, health care leadership noted that integration into routine services would require authorization from senior government officials overseeing health affairs.

> *To integrate the program into our daily work, we need approval from the senior leaders in charge of this work at the higher level.* [GO, HBD33]

Similarly, in the urban context, preschool managers indicated that continued implementation in preschools would depend on approval and support from the education department.

XSL•FO
RenderX

> *To integrate the program into the preschool's daily work, we need support from the local education bureau leaders. It would be best to also gain official support from your university.* [PM, SBD32]

In addition, leaders in both settings expressed concerns about sustaining staff motivation over the long term. While staff had fulfilled their responsibilities during the pilot phase, there was uncertainty about how to provide sufficient and appropriate incentives to support ongoing, routine implementation once the program became part of daily work. This concern was echoed in the educational setting, where managers also stressed the need for incentives to maintain teachers' continued engagement.

> *If the work is unpaid and voluntary labor, as mentioned by the village doctor, it is a challenge for the village doctor.* [GO, HBD27]

Another significant barrier was the lack of perceived demand from parents. Government officers and organizational managers from both settings expressed concern that many parents did not recognize the value or necessity of parenting support. Without sufficient parental buy-in, they feared that efforts to integrate the program into routine services would result in low participation rates and limited impact.

> *After all, we are in the countryside. Some parents have a higher level of awareness and think that interacting with their children is meaningful. However, some parents feel it is unnecessary; if the children can play alone and do not cause trouble, they have done their job.* [WACNHW, HBD26]

> *Many parents, when it comes to parenting, understand but not to a great extent, which causes a disconnect. They feel they somewhat understand but not fully, and accepting some new concepts is difficult.* [PM, SBD30]

Distinct Facilitators and Barriers to Maintenance at the Setting Level

A unique facilitating factor in the health center-based delivery was the close relationship between village doctors and local families, which enhanced trust and communication. However, implementation in this setting also faced a unique challenge in securing resources for sustainable material incentives. Both the health care and educational settings shared the need for further localized contextual adaptation of the program. Nonetheless, the focus of these recommended adaptations differed between the 2 settings.

In the rural health care setting, leaderships perceived that parenting programs facilitated by village doctors were feasible for sustained implementation in rural health care settings. They believed that village doctors, who often had strong, long-standing relationships with families—sometimes dating back to prenatal care—were uniquely positioned to gain family acceptance of the program and successfully conduct home visits. In contrast, the preschool program relied on teachers leading group-based remote support and did not require the same depth of ongoing personal relationships with students' families as village doctors had.

> *In our rural areas, village doctors are like family doctors, closely connected to the families. People know each other well, and home visits are easy. In the city, living in apartment buildings makes it less convenient, and people are more guarded.* [GO, HBD32]

In addition, the approach to sustaining staff motivation differed between the 2 settings. In the educational setting, preschool managers noted that parenting work is closely aligned with headteachers' core educational duties. As a result, integrating the program into daily operations was more natural and did not significantly increase their workload. Therefore, spiritual incentives, such as recognition and a sense of professional fulfillment, were seen as important for promoting long-term motivation.

> *Parenting is already part of our teachers' work, so it does not add extra burden. However, some motivation is still needed, such as certificates or awards.* [PM, SBD33]

In contrast, in the health care setting, leadership emphasized that village doctors primarily focused on public health responsibilities. Parenting-related tasks were viewed as additional duties, which increased their overall workload. In this context, material incentives were considered more effective, as they better compensated for the extra effort required for sustained engagement. By comparison, for preschool teachers, parenting was closely integrated with their daily work. The program did not add to their workload but rather complemented it, serving as an educational tool to support their parenting-related tasks.

> *Our village doctors' main work is in public health. This task adds extra duties for them, so compensation is needed.* [GO, HBD28]

> *I don't think this is a burden for the teachers, as they are already engaged in education. Parenting is also familiar to them. They can adapt the program content into teaching tools based on their needs.* [PM, SBD30]

Furthermore, adaptations to each specific implementation context were perceived as important. In the preschool, leaderships suggested that for the program to be effectively integrated into daily work, the digital learning format, duration, and additional parenting support needed to be further adapted to better align with teachers' schedules and classroom activities. This type of adaptation primarily involved organizational alignment, ensuring the program fit seamlessly within existing school routines.

> *To become part of our daily work, the program needs to align with the kindergarten's teaching needs and integrate with our activities. Teachers have to be the leaders.* [PM, SBD30]

In the health care setting, health care managers believed that for long-term implementation, the program needed to be adjusted to include special groups, such as left-behind children and children cared for by grandparents, who make up the majority in the local area, and also need parenting support. However,

XSL•FO
RenderX

this time, the adaptation focused more on tailoring the program to parents who stayed at home to care for their children, as a result of careful consideration of available implementation resources and the exploratory nature of the pilot as an initial test of digital interventions in rural areas.

> *Like in our area, most young parents go out to work, leaving grandparents behind. Therefore, these families need more services.* [GO, HBD33]

The summary of barriers and facilitators to maintenance at the setting level can be seen in Multimedia Appendix 7.

Differentiated Models of Digital Parenting Program Implementation in Two Sectors

Based on the results, we summarize two models for implementing and institutionalizing the digital parenting program in rural health care and urban educational settings (Figure 3). The basic sequence starts with promoting organizational adoption of the program, followed by encouraging staff adoption, then expanding reach. After that, the focus shifts to implementation and finally to maintenance. In impoverished rural areas, digital interventions need to be complemented by individualized human-led offline support to enhance the impact. Accordingly, we propose a model that involves collaboration with the health center and leverages village doctors to deliver the digital program alongside home visits. For urban areas, we suggest a model where digital components are combined with online parent groups led by teachers.

During the organizational adoption stage, building trust with managers, emphasizing the alignment between the program and the organization, and highlighting the program's benefits are crucial. Trust can be fostered by demonstrating the evidence base and credibility of the program, which helps managers recognize the program's potential. Emphasizing how the program aligns with the organization's goals encourages adoption of the program and fosters positive expectations regarding its completion.

Figure 3. Differentiated models of delivery in urban preschool and rural health care settings.

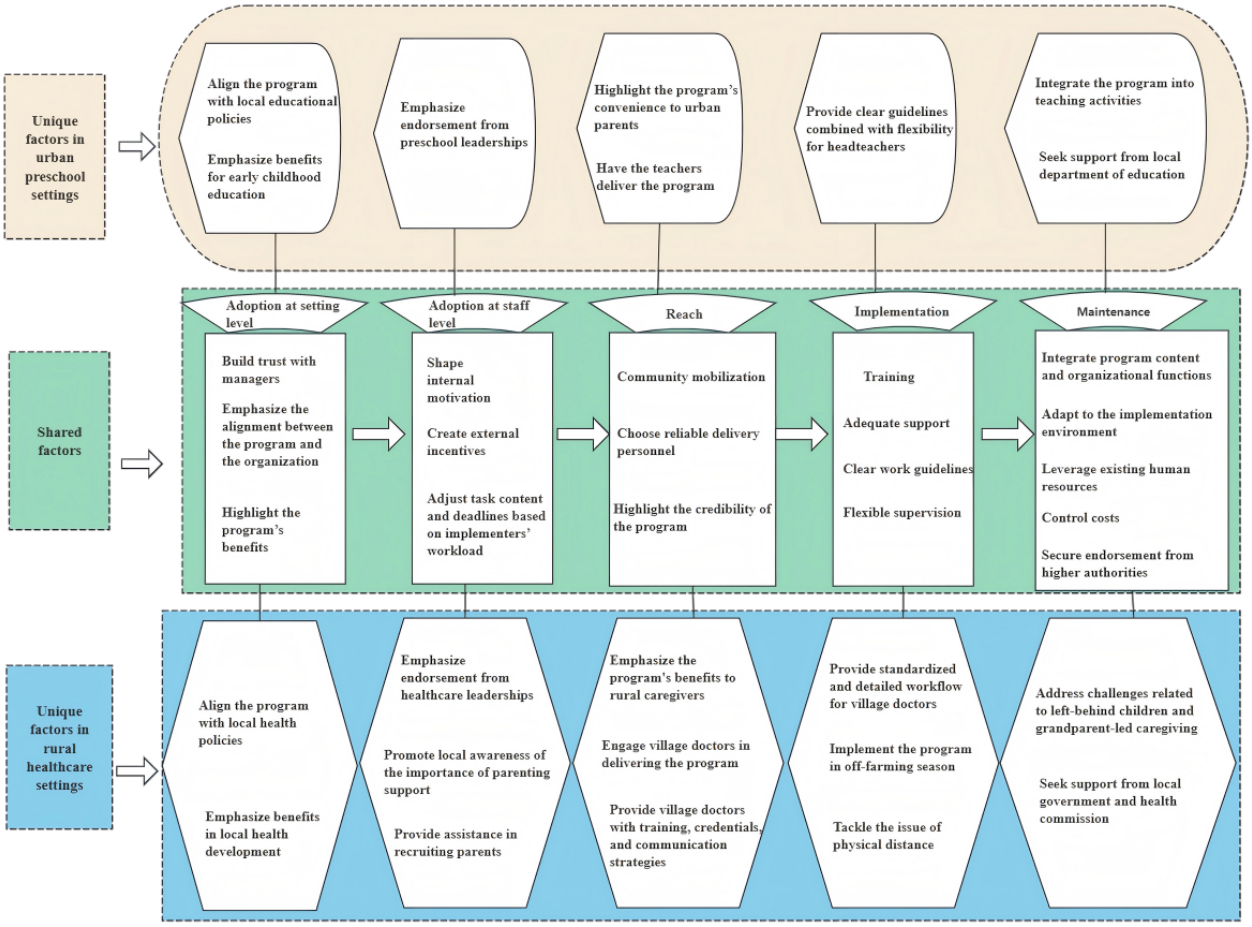

In rural health care settings, aligning the program with local health policies can strengthen its acceptance, particularly by emphasizing benefits such as improving early childhood development and addressing population health challenges. In contrast, urban educational settings can benefit from closer ties to educational policies, where the program's role in enhancing teaching quality, promoting parent-teacher collaboration, and reducing behavioral problems in preschool environments is more prominent. Thus, the focus in rural areas is more health-centered, while urban settings prioritize educational outcomes.

At the staff-level adoption stage, it is essential to shape internal motivation, create external incentives, and adjust task content and deadlines based on implementers' workload. First,

XSL•FO
RenderX

effectively communicating the program's significance helps foster internal motivation and enables implementers to recognize its benefits. Preschool teachers typically have stronger internal motivation because their daily work is directly related to parenting, allowing them to see immediate benefits. In contrast, village doctors in health centers, whose routine duties are less connected to parenting, require communication and training to understand how parenting relates to their role and to enhance their perception of the program's importance.

Second, when internal motivation alone is insufficient, forming external incentives becomes essential. Depending on the program's context, it is necessary to strike a balance between material and intrinsic incentives to fully motivate implementers. For example, in the rural pilot, providing village doctors with reasonable compensation effectively promoted their active participation in program implementation.

Regarding recruitment, urban preschools generally experience less difficulty, as parents tend to cooperate more readily with teachers. However, in rural areas, recruiting parents is more challenging due to complex family structures and village doctors' limited experience with parenting work. This necessitates additional support from implementers and institutional leaders.

In the reach stage, community mobilization, choosing reliable delivery personnel, and highlighting the credibility of the program content are essential. Through community mobilization, parents can understand the program's goals, content, and requirements, and most importantly, recognize the benefits of the program and that the learning process is convenient and low in time cost. In addition, it is worthwhile to highlight the program's convenience to urban parents and emphasize its benefits to rural parents. During the promotion, it is also essential to highlight the program's being systematic, distinguishing it from fragmented parenting information online. In urban preschools, teachers can serve as program implementers due to their close relationships with parents. In rural health care settings, village doctors can take on this role, as they are generally trusted by families. However, since parenting is not a routine part of village doctors' responsibilities, providing them with adequate training, relevant credentials, and communication strategies is necessary to reduce parental skepticism about the program.

In the implementation stage, training, adequate support, clear work guidelines, and flexible supervision are essential for ensuring fidelity. However, the specificity of work guidelines should be tailored to the implementers' roles. In rural health care settings, village doctors, who are generally unfamiliar with parenting work, require standardized and highly detailed workflows to guide their delivery. Conversely, urban preschool teachers, already experienced with parenting-related tasks, benefit from clear guidelines combined with flexibility to facilitate group discussions during parent meetings or school activities, supported by suggested discussion topics. Additionally, in rural areas, particular consideration must be given to the timing and logistics of home visits, as busy farming seasons often limit the availability of both village doctors and families for such visits.

In the maintenance stage, several factors are critical to ensuring the program's sustainability: integrating program content with organizational functions, adapting to the implementation environment, leveraging existing human resources, controlling costs, and securing endorsement from higher authorities. Controlling costs is paramount, as high expenses threaten the long-term sustainability of the program.

Various strategies can be used to reduce implementation costs. Similar to adoption, alignment between organizational functions and program content is important, but in the maintenance stage, this alignment emphasizes better integration of the program with the organization's daily work content and clarifying the position of the program task in the organization.

Using existing human resources by providing adequate onboarding and ongoing training can make the best use of the organization's current staff rather than hiring new personnel. Furthermore, it is necessary to obtain endorsement from higher authorities, with preschools needing support from the education department and health centers from public health authorities.

Adaptations must reflect the distinct realities of urban educational and rural health care settings. In urban preschools, the program needs to be integrated into teaching activities, becoming a practical tool within educators' daily routines. Conversely, rural areas must also address challenges related to left-behind children and grandparent-led caregiving to ensure these families are included in the services.

Discussion

Principal Findings

This study aimed to identify multilevel factors influencing the implementation of a digital parenting program across 2 distinct settings. The findings addressed this aim by revealing both shared and setting-specific facilitators and barriers across the RE-AIM dimensions, providing a comprehensive understanding of how digital parenting programs may be effectively implemented and sustained in different institutional contexts.

While well-known digital parenting programs such as Triple P Online and Incredible Years Online have shown positive outcomes in several countries [48-50], this evidence is predominantly from high-income countries, with limited exploration of their adaptability and sustainability across diverse implementation contexts. Our study directly addresses this gap by examining the implementation of a digital parenting program in 2 distinct settings in China—an urban educational setting and a rural health care setting—thereby contributing crucial evidence from a large low- and middle-income country.

Parents in both urban and rural areas expressed a desire for parenting knowledge, which was motivated by their wish to support their children in growing up healthy. Although the internet provided access to information, much of it was fragmented and lacked a systematic approach. As a result, parents often struggled to build a coherent knowledge system or apply these concepts in practice. This finding is consistent with previous research [51]. In contrast, the digital parenting

XSL•FO
RenderX

program offers culturally adapted, evidence-based content that is systematic, practical, and easy to access.

Trust is indispensable to a program's success, as previous study shows [52]. It must be built among organization managers, implementers, and the beneficiaries. First, building trust with managers and implementers to get their support is important [53]. It can be fostered by showcasing the program's evidence base, emphasizing its benefits, and selecting those with prior cooperation experience. Second, building trust with beneficiaries requires the credibility of the delivery channel and the implementer. A trusted implementer can boost parents' confidence in the program and provide extra motivations, such as social consideration (eg, face-saving or relationship building). In rural areas, village doctors, trusted as family doctors, are well-suited for this role [54], while in urban areas, preschool teachers hold similar credibility and respect.

Similar to previous research [55], we found that community mobilization for both parents and implementers before implementation is important for generating demand, thereby increasing the adoption and program reach. Parents should be introduced to the importance of evidence-based parenting, the benefits of the program, and how it works. At the same time, implementers need to understand the connection between the program and their routine work, its significance for them, and the necessity of the human-led components. Additionally, fostering a shared vision with all stakeholders is essential.

Providing support for implementers is also crucial, which aligns with previous research [56]. It is important to provide adequate training to equip them with implementation skills and clearly define their responsibilities. Establishing communication channels related to parenting support within the organization, such as regular meetings to discuss issues and digital systems to monitor progress. This will facilitate efficient vertical communication, supervision, and problem-solving mechanisms between implementers and organization managers.

Appropriate incentives for implementers are essential. Previous research emphasizes their importance, but it often lacks guidance on how to provide them [57]. We found that material and psychological incentives should be based on the workload and the fit between the implementer's tasks and capabilities. When human-led responsibilities exceed usual duties, greater incentives may be required to ensure motivation and sustained engagement.

Building on these findings, several practical and theoretical implications emerge for strengthening digital parenting interventions across diverse contexts. Program adaptation is necessary for maintenance, which aligns with previous research [53]. Adaptation occurs at 2 levels: content and context [58]. Content adaptation involves adding information relevant to the organization's functions and the local population's needs, such as child nutrition and disease prevention in rural health care, and early education in urban preschools. Context adaptation involves adjusting delivery methods to fit the organization's characteristics, ensuring that the program becomes integrated into regular work [56]. In rural areas, additional focus on left-behind children and grandparent caregivers is necessary because of their population structure [59]. While urban settings should incorporate the program into daily school activities, such as parent meetings and lectures.

Implications for Policy, Practice, and Research

To better support families in both urban and rural areas, governments should invest in providing trustworthy and accessible parenting resources. Trusted institutions, such as preschools and health centers, can serve as effective channels for delivering this information. With government support and institutional credibility, families may be more likely to engage in the programs.

When implementing digital parenting programs, it is essential to tailor strategies to local contexts, considering organizational mandates, local culture, demographic profiles, and economic conditions. Integrating the program into existing institutional workflows can reduce costs and promote sustainability. Establishing trust with organizational leaders, frontline implementers, and the beneficiaries is also key and should be supported by early and continuous community mobilization and engagement. Importantly, implementation must strike a balance between fidelity to core content and flexibility in delivery. Aligning program delivery with the institutional capabilities and work patterns, while providing clear and structured workflows, can help maintain quality and adapt to implementers' needs on the ground.

Future research should explore a broader range of settings and more diverse implementation strategies. Additionally, this study primarily used the RE-AIM framework to evaluate implementation at the mezzo- and macro-levels. However, parents, as direct beneficiaries, also play a crucial role. Their acceptance and participation significantly influence program success. Therefore, future research should adopt a micro-perspective focusing on parents' perceptions, attitudes, acceptance, and engagement. Such insights would inform future implementation of digital parenting programs.

Conclusions

This study's findings go beyond demonstrating the feasibility of digital parenting programs in low-resource settings to highlight a key lesson for their institutionalization: successful scaling is not about a single "one-size-fits-all" digital solution. Effective implementation requires hybrid models that strategically combine low-cost technology with trusted local human infrastructure, such as teachers in urban schools or village doctors in rural clinics.

Our comparison of urban and rural settings shows that the human-led component must be carefully tailored. Urban environments can benefit from flexible online group support, whereas rural contexts often require the structure and accountability of in-person visits. These insights provide a roadmap for policymakers and practitioners to move beyond standardized rollouts and develop a flexible "implementation playbook." By prioritizing adaptation to local social and organizational contexts, evidence-based digital parenting interventions can bridge the gap from efficacy to sustainable, equitable, and real-world impact.

XSL•FO
RenderX

Limitations and Strengths

First, this study did not report on the program's effectiveness. This may raise concerns about its impact. We plan to address this in a forthcoming publication specifically focusing on impact evaluation. Second, this study did not delve into micro-level aspects, especially parents' adoption of and attitudes toward the program. While this is an important area, it could not be adequately covered due to space limitations and will be examined in future research. Third, this study was limited to 2 pilot sites. Including more case studies in future research will help enrich the findings and enhance external validity.

The strength of this study lies in its focus on the implementation process of digital parenting programs, providing valuable insights for future program implementation. We also explored different implementation models across urban and rural education and health care sectors. This can support potential scale-up in diverse contexts. Additionally, we discussed the use of human-led models in different practical settings. This can provide guidance for choosing appropriate approaches in future hybrid (digital and human-led) parenting programs.

Acknowledgments

We extend our sincere gratitude to Xuechen Zhang, Ruinan Zhou, Taoran Li, Xinran Liu, Qingyang Zhang, Laurie Markle, Chiara Facciolà, Wenhao Ma, Zhen Liu, Yicong Guo, Edmud Moss, Ian Stride, Lily Clements, Dongping Qiao, and Xiying Wang for their valuable contributions to the program adaptation. We are especially grateful to Yicong Guo for her assistance with data collection. We wish to thank Xinnan Zhao, Bing Shen, and Jiawen Zhu for providing implementation assistance. We are deeply grateful to the county and township governments, local program staff, maternal and child health workers, village doctors, Jiangxi Xinyu Chengbei Preschool, headteachers, social workers, and participating families. Their engagement and support were vital to the success of this research. Language in this manuscript was improved using generative artificial intelligence tools. No content was generated by artificial intelligence.

Funding

China Postdoctoral Science Foundation (2022M720467), LEGO Foundation, CICC, and the CICC Charity Foundation. The funders were not involved in the design and conduct of the study, the collection, management, analysis, and interpretation of the data.

Data Availability

Deidentified participant data will be made available from the corresponding author upon reasonable request, following publication of the article.

Authors' Contributions

ZF and XS conceptualized the study. YQ and NZ supported the local implementation. RR and XS curated and analyzed the data. XS, ZF, RR, and NZ interpreted the results. XS drafted the manuscript with support from ZF. All authors, including XS, RR, YQ, JML, NZ, and ZF, reviewed and revised the manuscript.

Conflicts of Interest

YQ is an employee of the China Development Research Foundation. JML is the CEO of Parenting for Lifelong Health (PLH), a charitable organization based in the United Kingdom that developed the adapted program. PLH programs are open access and licensed under a Creative Commons 4.0 Attribution Share-Alike license. ZF has worked as a consultant for PLH in the past. JML has (and is participating) in a number of research studies involving the program, as an investigator, and the University of Oxford and University of Cape Town receive research funding for these.

Multimedia Appendix 1
Program description.
[DOCX File, 302 KB - jmir_v28i1e79848_app1.docx]

Multimedia Appendix 2
Comparison of the two programs.
[DOCX File, 21 KB - jmir_v28i1e79848_app2.docx]

Multimedia Appendix 3
Description of interview participants.
[DOCX File, 21 KB - jmir_v28i1e79848_app3.docx]

XSL•FO
RenderX

Multimedia Appendix 4
Summary of barriers and facilitators to reach.
[DOCX File, 15 KB - jmir_v28i1e79848_app4.docx]

Multimedia Appendix 5
Summary of barriers and facilitators to adoption.
[DOCX File, 16 KB - jmir_v28i1e79848_app5.docx]

Multimedia Appendix 6
Summary of barriers and facilitators to implementation of the digital parenting program.
[DOCX File, 15 KB - jmir_v28i1e79848_app6.docx]

Multimedia Appendix 7
Summary of barriers and facilitators to maintenance at the setting level of the digital parenting program.
[DOCX File, 16 KB - jmir_v28i1e79848_app7.docx]

Checklist 1
Checklist of iCHECK-DH guidelines. iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations.
[DOCX File, 23 KB - jmir_v28i1e79848_app8.docx]

References

1. Britto PR, Lye SJ, Proulx K, et al. Nurturing care: promoting early childhood development. Lancet 2017 Jan 7;389(10064):91-102. [doi: 10.1016/S0140-6736(16)31390-3] [Medline: 27717615]
2. Black MM, Walker SP, Fernald LCH, et al. Early childhood development coming of age: science through the life course. Lancet 2017 Jan 7;389(10064):77-90. [doi: 10.1016/S0140-6736(16)31389-7] [Medline: 27717614]
3. Aboud FE, Yousafzai AK. Global health and development in early childhood. Annu Rev Psychol 2015 Jan 3;66(433–457):433-457. [doi: 10.1146/annurev-psych-010814-015128] [Medline: 25196276]
4. Lugo-Gil J, Tamis-LeMonda CS. Family resources and parenting quality: links to children's cognitive development across the first 3 years. Child Dev 2008;79(4):1065-1085. [doi: 10.1111/j.1467-8624.2008.01176.x] [Medline: 18717907]
5. Mistry KB, Minkovitz CS, Riley AW, et al. A new framework for childhood health promotion: the role of policies and programs in building capacity and foundations of early childhood health. Am J Public Health 2012 Sep;102(9):1688-1696. [doi: 10.2105/AJPH.2012.300687] [Medline: 22813416]
6. Eyberg SM, Nelson MM, Boggs SR. Evidence-based psychosocial treatments for children and adolescents with disruptive behavior. J Clin Child Adolesc Psychol 2008 Jan;37(1):215-237. [doi: 10.1080/15374410701820117] [Medline: 18444059]
7. Kaminski JW, Claussen AH. Evidence base update for psychosocial treatments for disruptive behaviors in children. J Clin Child Adolesc Psychol 2017;46(4):477-499. [doi: 10.1080/15374416.2017.1310044] [Medline: 28459280]
8. WHO Guidelines on Parenting Interventions to Prevent Maltreatment and Enhance Parent–Child Relationships with Children Aged 0–17 Years: World Health Organization; 2022.
9. Aboud FE, Prado EL. Measuring the implementation of early childhood development programs. Ann N Y Acad Sci 2018 May;1419(1):249-263. [doi: 10.1111/nyas.13642] [Medline: 29791725]
10. Jeong J, Franchett EE, Ramos de Oliveira CV, Rehmani K, Yousafzai AK. Parenting interventions to promote early child development in the first three years of life: a global systematic review and meta-analysis. PLoS Med 2021;18(5):e1003602. [doi: 10.1371/journal.pmed.1003602]
11. Luoto JE, Lopez Garcia I, Aboud FE, et al. Group-based parenting interventions to promote child development in rural Kenya: a multi-arm, cluster-randomised community effectiveness trial. Lancet Glob Health 2021 Mar;9(3):e309-e319. [doi: 10.1016/S2214-109X(20)30469-1] [Medline: 33341153]
12. Corralejo SM, Domenech Rodríguez MM. Technology in parenting programs: a systematic review of existing interventions. J Child Fam Stud 2018 Sep;27(9):2717-2731. [doi: 10.1007/s10826-018-1117-1]
13. Solís-Cordero K, Duarte LS, Fujimori E. Effectiveness of remotely delivered parenting programs on caregiver-child interaction and child development: a systematic review. J Child Fam Stud 2022;31(11):3026-3036. [doi: 10.1007/s10826-022-02328-8] [Medline: 35615461]
14. 55th statistical report on china's internet development. : China Internet Network Information Center (CNNIC); 2024 URL: https://www.cnnic.com.cn/IDR/ReportDownloads/202411/P020241101318428715781.pdf
15. Hall CM, Bierman KL. Technology-assisted interventions for parents of young children: emerging practices, current research, and future directions. Early Child Res Q 2015;33:21-32. [doi: 10.1016/j.ecresq.2015.05.003]
16. Fang Z, Martin M, Copeland L, Evans R, Shenderovich Y. Parenting interventions during the COVID-19 pandemic: a systematic review of the rationales, process, feasibility, acceptability, and impacts of adaptation. Trauma Violence Abuse 2024 Dec;25(5):3887-3902. [doi: 10.1177/15248380241266183] [Medline: 39082191]

XSL•FO
RenderX

17. Jäggi L, Hartinger SM, Fink G, et al. Parenting in the digital age: a scoping review of digital early childhood parenting interventions in low- and middle-income countries (LMIC). Public Health Rev 2025 Jan 21;45:1607651. [doi: 10.3389/phrs.2024.1607651]
18. Baumel A, Pawar A, Kane JM, Correll CU. Digital parent training for children with disruptive behaviors: systematic review and meta-analysis of randomized trials. J Child Adolesc Psychopharmacol 2016 Oct;26(8):740-749. [doi: 10.1089/cap.2016.0048]
19. Breitenstein SM, Gross D, Christophersen R. Digital delivery methods of parenting training interventions: a systematic review. Worldviews Ev Based Nurs 2014 Jun;11(3):168-176. [doi: 10.1111/wvn.12040]
20. Xie EB, Jung JW, Kaur J, Benzies KM, Tomfohr-Madsen L, Keys E. Digital parenting interventions for fathers of infants from conception to the age of 12 months: systematic review of mixed methods studies. J Med Internet Res 2023 Jul 26;25:e43219. [doi: 10.2196/43219] [Medline: 37494086]
21. Leijten P, Rienks K, Groenman AP, et al. Online parenting support: meta-analyses of non-inferiority and additional value to in-person support. Child Youth Serv Rev 2024 Apr;159:107497. [doi: 10.1016/j.childyouth.2024.107497]
22. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
23. Ahun MN, Ali NB, Hentschel E, Jeong J, Franchett E, Yousafzai AK. A meta-analytic review of the implementation characteristics in parenting interventions to promote early child development. Ann N Y Acad Sci 2024 Mar;1533(1):99-144. [doi: 10.1111/nyas.15110] [Medline: 38354095]
24. Richter LM, Daelmans B, Lombardi J, et al. Investing in the foundation of sustainable development: pathways to scale up for early childhood development. Lancet 2017 Jan 7;389(10064):103-118. [doi: 10.1016/S0140-6736(16)31698-1] [Medline: 27717610]
25. Peters DH, Adam T, Alonge O, Agyepong IA, Tran N. Republished research: implementation research: what it is and how to do it. Br J Sports Med 2014 Apr;48(8):731-736. [doi: 10.1136/bmj.f6753]
26. Britto PR, Singh M, Dua T, Kaur R, Yousafzai AK. What implementation evidence matters: scaling-up nurturing interventions that promote early childhood development. Ann N Y Acad Sci 2018 May;1419(1):5-16. [doi: 10.1111/nyas.13720] [Medline: 29791739]
27. Lachman JM, Kelly J, Cluver L, Ward CL, Hutchings J, Gardner F. Process evaluation of a parenting program for low-income families in South Africa. Res Soc Work Pract 2018 Feb;28(2):188-202. [doi: 10.1177/1049731516645665]
28. Glasgow RE, Harden SM, Gaglio B, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019 Mar 29;7. [doi: 10.3389/fpubh.2019.00064]
29. Gaglio B, Shoup JA, Glasgow RE. The RE-AIM framework: a systematic review of use over time. Am J Public Health 2013 Jun;103(6):e38-e46. [doi: 10.2105/AJPH.2013.301299]
30. Glasgow RE, Askew S, Purcell P, et al. Use of RE-AIM to address health inequities: application in a low-income community health center-based weight loss and hypertension self-management program. Behav Med Pract Policy Res 2013 Jun;3(2):200-210. [doi: 10.1007/s13142-013-0201-8]
31. Bakken S, Ruland CM. Translating clinical informatics interventions into routine clinical care: how can the RE-AIM framework help? J Am Med Inform Assoc 2009 Nov 1;16(6):889-897. [doi: 10.1197/jamia.M3085]
32. MacDonald B, Gibson AM, Janssen X, Kirk A. A mixed methods evaluation of a digital intervention to improve sedentary behaviour across multiple workplace settings. Int J Environ Res Public Health 2020 Jun 24;17(12):4538. [doi: 10.3390/ijerph17124538] [Medline: 32599730]
33. Stephens AB, Wynn CS, Stockwell MS. Understanding the use of digital technology to promote human papillomavirus vaccination – a RE-AIM framework approach. Hum Vaccin Immunother 2019 Aug 3;15(7-8):1549-1561. [doi: 10.1080/21645515.2019.1611158]
34. Taylor ME, Liu M, Abelson S, Eisenberg D, Lipson SK, Schueller SM. The reach, effectiveness, adoption, implementation, and maintenance of digital mental health interventions for college students: a systematic review. Curr Psychiatry Rep 2024 Dec;26(12):683-693. [doi: 10.1007/s11920-024-01545-w]
35. Yoshida Y, Patil SJ, Brownson RC, et al. Using the RE-AIM framework to evaluate internal and external validity of mobile phone–based interventions in diabetes self-management education and support. J Am Med Inform Assoc 2020 Jun 1;27(6):946-956. [doi: 10.1093/jamia/ocaa041]
36. Wang J, Hedley D, Bury SM, Barbaro J. A systematic review of screening tools for the detection of autism spectrum disorder in mainland China and surrounding regions. Autism 2020 Feb;24(2):285-296. [doi: 10.1177/1362361319871174]
37. Xiong N, Yang L, Yu Y, et al. Investigation of raising burden of children with autism, physical disability and mental disability in China. Res Dev Disabil 2011 Jan;32(1):306-311. [doi: 10.1016/j.ridd.2010.10.003]
38. Eisenhardt KM. Building theories from case study research. The Academy of Management Review 1989 Oct;14(4):532. [doi: 10.2307/258557]
39. Meredith J. Building operations management theory through case and field research. J of Ops Management 1998 Jul;16(4):441-454. [doi: 10.1016/S0272-6963(98)00023-0]

XSL•FO
RenderX

40. Bellg AJ, Borrelli B, Resnick B, et al. Enhancing treatment fidelity in health behavior change studies: best practices and recommendations from the NIH behavior change consortium. Health Psychol ;23(5):443-451. [doi: 10.1037/0278-6133.23.5.443]
41. Ward CL, Wessels IM, Lachman JM, et al. Parenting for lifelong health for young children: a randomized controlled trial of a parenting program in South Africa to prevent harsh parenting and child conduct problems. J Child Psychol Psychiatry 2020 Apr;61(4):503-512. [doi: 10.1111/jcpp.13129] [Medline: 31535371]
42. Designing, Implementing, Evaluating, and Scaling Up Parenting Interventions: A Handbook for Decision-Makers and Implementers: World Health Organization; 2024. URL: https://www.who.int/publications/i/item/9789240095595
43. Emmers D, Jiang Q, Xue H, et al. Early childhood development and parental training interventions in rural China: a systematic review and meta-analysis. BMJ Glob Health 2021 Aug;6(8):e005578. [doi: 10.1136/bmjgh-2021-005578] [Medline: 34417271]
44. Zhou J, Heckman J, Wang F, Liu B. Early childhood learning patterns for a home visiting program in rural China. J Community Psychol 2023 Mar;51(2):584-604. [doi: 10.1002/jcop.22872] [Medline: 35567396]
45. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Meth Thousand 2017 Oct 2;16(1). [doi: 10.1177/1609406917733847]
46. Kiger ME, Varpio L. Thematic analysis of qualitative data: AMEE guide no. 131. Med Teach 2020 Aug;42(8):846-854. [doi: 10.1080/0142159X.2020.1755030] [Medline: 32356468]
47. Perrin Franck C, Babington-Ashaye A, Dietrich D, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023 May 10;25(1):e46694. [doi: 10.2196/46694]
48. Entenberg GA, Mizrahi S, Walker H, et al. AI-based chatbot micro-intervention for parents: meaningful engagement, learning, and efficacy. Front Psychiatry 2023 Jan 20;14. [doi: 10.3389/fpsyt.2023.1080770]
49. Day JJ, Sanders MR. Do parents benefit from help when completing a self-guided parenting program online? a randomized controlled trial comparing triple P online with and without telephone support. Behav Ther 2018 Nov;49(6):1020-1038. [doi: 10.1016/j.beth.2018.03.002] [Medline: 30316482]
50. Villalta L, Elias M, Llorens M, Vall-Roqué H, Romero-González M, Valencia-Agudo F. Implementation of the incredible years-ASLD® program in autism and preterm children with communication and/or socialization difficulties in Spain (FIRST STEPS): a feasibility study. Child Youth Serv Rev 2025 Oct;177:108422. [doi: 10.1016/j.childyouth.2025.108422]
51. Liu TL, Hsiao RC, Chou WJ, Yen CF. Parenting stress, anxiety, and sources of acquiring knowledge in Taiwanese caregivers of children with attention-deficit/hyperactivity disorder. BMC Public Health 2024 Jun 24;24(1):1675. [doi: 10.1186/s12889-024-18761-x] [Medline: 38914984]
52. Burke SM, Shapiro S, Petrella RJ, et al. Using the RE-AIM framework to evaluate a community-based summer camp for children with obesity: a prospective feasibility study. BMC Obes 2015;2(21):21. [doi: 10.1186/s40608-015-0050-8] [Medline: 26217536]
53. Wozniak LA, Soprovich AL, Johnson JA, Eurich DT. Adopting and implementing an innovative model to organize diabetes care within First Nations communities: a qualitative assessment. BMC Health Serv Res 2021 May 3;21(1):415. [doi: 10.1186/s12913-021-06424-1] [Medline: 33941176]
54. Nielsen JV, Skovgaard T, Bredahl TVG, Bugge A, Wedderkopp N, Klakk H. Using the RE-AIM framework to evaluate a school-based municipal programme tripling time spent on PE. Eval Program Plann 2018 Oct;70:1-11. [doi: 10.1016/j.evalprogplan.2018.05.005] [Medline: 29890448]
55. Cassar S, Salmon J, Timperio A, et al. Optimizing intervention dissemination at scale: a qualitative study of multi-sector partner organization experiences. Transl Behav Med 2024 Oct 6;14(10):621-633. [doi: 10.1093/tbm/ibae042] [Medline: 39216008]
56. Wozniak L, Soprovich A, Rees S, Al Sayah F, Majumdar SR, Johnson JA. Contextualizing the effectiveness of a collaborative care model for primary care patients with diabetes and depression (teamcare): a qualitative assessment using RE-AIM. Can J Diabetes 2015 Oct;39:S83-S91. [doi: 10.1016/j.jcjd.2015.05.004]
57. Cohen DJ, Crabtree BF, Etz RS, et al. Fidelity versus flexibility: translating evidence-based research into practice. Am J Prev Med 2008 Nov;35(5 Suppl):S381-S389. [doi: 10.1016/j.amepre.2008.08.005] [Medline: 18929985]
58. Miller CJ, Barnett ML, Baumann AA, Gutner CA, Wiltsey-Stirman S. The FRAME-IS: a framework for documenting modifications to implementation strategies in healthcare. Implementation Sci 2021 Dec;16(1):36. [doi: 10.1186/s13012-021-01105-3]
59. Lyu L, Mei Z, Yan F, Wang X, Duan C. The status of rural children left-behind in China: 2010–2020. China popul dev stud 2024 Jun;8(2):97-111. [doi: 10.1007/s42379-024-00159-2]

Abbreviations

FGD: focus group discussion
HBD: hospital-based program
LMIC: low- and middle-income country
PP: program participants

XSL•FO
RenderX

RE-AIM : Reach, Effectiveness, Adoption, Implementation, and Maintenance
SBD: preschool-based program

Edited by N Cahill; submitted 03.Jul.2025; peer-reviewed by M Galani, X Liang; accepted 22.Oct.2025; published 06.Jan.2026.

Please cite as:
Shi X, Ruan R, Qie Y, Lachman JM, Zhong N, Fang Z
Institutionalizing Digital Parenting Programs in Low Resource Settings in China: Comparative Case Study of Health Care and Education Sectors Using the RE-AIM Framework
J Med Internet Res 2026;28:e79848
URL: https://www.jmir.org/2026/1/e79848
doi:10.2196/79848

©Xinyu Shi, Ruochen Ruan, Yi Qie, Jamie M Lachman, Na Zhong, Zuyi Fang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Complex Digital Health Intervention to Support People With HIV: Organizational Readiness Survey Study and Preimplementation Planning for a Hybrid Effectiveness-Implementation Study

Jacqueline Hodges[1], MD, MPH; Wendy Cohn[2], PhD; Amanda Castel[3], MD, MPH; Tabor Flickinger[4], MD, MPH; Ava Lena Waldman[5], CHES, CCRP, MHS; Michelle Hilgart[6], MEd, PhD; Olivia Kirby[3], MPH; Sylvia Caldwell[5], MPH, DHSc; Karen Ingersoll[6], ABPP, PhD

[1]Division of Infectious Diseases, Department of Medicine, Duke University, 315 Trent Drive, Durham, NC, United States

[2]Department of Public Health Sciences, University of Virginia, Charlottesville, VA, United States

[3]Milken Institute School of Public Health, George Washington University, Washington, DC, United States

[4]General, Geriatric & Palliative Medicine, Department of Medicine, University of Virginia, Charlottesville, VA, United States

[5]Division of Infectious Diseases and International Health, Department of Medicine, University of Virginia, Charlottesville, VA, United States

[6]Department of Psychiatry and Neurobehavioral Sciences, University of Virginia, Charlottesville, VA, United States

Corresponding Author:
Jacqueline Hodges, MD, MPH
Division of Infectious Diseases, Department of Medicine, Duke University, 315 Trent Drive, Durham, NC, United States

Abstract

Background: Evaluating implementation of digital health interventions (DHIs) in practice settings is complex, involving diverse users and multistep processes. Proactive planning can ensure implementation determinants and outcomes are captured for hybrid studies, but operational guidance for designing or planning hybrid DHI studies is limited.

Objective: This study aimed to proactively define, prioritize, and operationalize measurement of implementation outcomes and determinants for a DHI hybrid effectiveness-implementation trial. We describe unique advantages and limitations of planning the trial implementation evaluation among a large-scale cohort study population and share results of a pretrial organizational readiness assessment.

Methods: We planned a cluster-randomized, type II hybrid effectiveness-implementation trial testing *PositiveLinks*, a smartphone app for HIV care, compared to usual care (n=6 sites per arm), among HIV outpatient sites in the DC Cohort Longitudinal HIV Study in Washington, DC. We (1) defined components of the DHI and associated implementation strategy; (2) selected implementation science frameworks to accomplish evaluation aims; (3) mapped framework dimensions, domains, and constructs to implementation strategy steps; (4) modified or created instruments to collect data for implementation outcome measures and determinants; and (5) developed a compatible implementation science data collection and management plan. Provider baseline surveys administered at intervention sites probed usage of digital tools and assessed provider readiness for implementation with the Organizational Readiness to Implement Change tool.

Results: We specified DHI and implementation strategy toward planning measurement of DHI and broader program reach and adoption. Mapping of implementation strategy steps to the Reach Effectiveness Adoption Implementation Maintenance framework prompted considerations for how to capture understudied aspects of each dimension: denominators and demographic representativeness within reach or adoption, and provider or organization-level adaptations, dose, and fidelity within the implementation dimension. Our process also prompted the creation of tools to obtain detailed determinants across domains and constructs of the Consolidated Framework for Implementation Research within a large sample at multiple time points. Some aspects of real-world *PositiveLinks* implementation were not reflected within the planned hybrid trial (eg, research assistants selected as de facto site implementation leads) or were modified to preserve internal validity of effectiveness measurement (eg, “Community of Practice”). Providers and research assistants (n=17) at intervention sites self-reported high baseline use of digital tools to communicate with patients. Readiness assessment revealed high median (48, IQR 45 - 54) total Organizational Readiness to Implement Change scores, with research assistants scoring higher than physicians (52.5, IQR 44-55 vs 48.0, IQR 46-49).

Conclusions: Key takeaways, challenges, and opportunities arose in planning the implementation evaluation within a hybrid DHI trial among a cohort population. Prospective trial planning must balance generalizability of implementation processes to

XSL•FO
RenderX

"real world" conditions with rigorous procedures to measure intervention effectiveness. Rapid, scalable tools require further study to enable evaluations within large multisite hybrid studies.

Trial Registration: ClinicalTrials.gov NCT04998019; https://clinicaltrials.gov/study/NCT04998019

International Registered Report Identifier (IRRID): RR2-10.2196/37748

(J Med Internet Res 2026;28:e76327) doi:10.2196/76327

KEYWORDS

mobile health; digital health; HIV; RE-AIM; CFIR; hybrid effectiveness-implementation study; implementation science; Reach Effectiveness Adoption Implementation Maintenance; Consolidated Framework for Implementation Research

Introduction

Digital health interventions (DHIs) including web-based and mobile health (mHealth) interventions can improve clinical outcomes for chronic health conditions. DHIs apply behavior change theories with variable mechanisms of action, such as enhanced motivation, self-management, and peer support. Munoz et al [1] and Hermes et al [2] describe the spectrum of behavioral intervention technologies ranging from adjunctive tools embedded within clinic-based care to support provider tasks, to fully-automated direct-to-consumer technologies for patient self-management. Implementation outcomes or the 'who,' 'how,' 'how much,' and 'how well' needed to implement DHIs [3] can vary widely as a result. Clinic-embedded DHIs may require significant engagement by multiple providers and staff. Additionally, DHIs with multiple features can generate large amounts of backend paradata related to usage that are important for quantifying implementation reach and adoption. Capturing DHI implementation outcomes and determinants can thus be complicated. Curran et al [4] describe 3 designs of hybrid effectiveness-implementation studies. However, the literature provides little operational guidance or shared procedural knowledge on planning hybrid studies to capture implementation outcomes and identify salient implementation determinants for complex multifeature multiuser DHIs.

We reviewed several comprehensive efforts to recharacterize or adapt broader health service implementation science frameworks for the study of DHIs. Recent examples related to evaluation frameworks, which evaluate implementation outcomes, include (1) a workshop conducted by the Dissemination and Implementation Core of the Center for Technology and Behavioral Health at Dartmouth College [5], (2) Hermes et al's [2] recharacterization of Proctor's outcomes for implementation research for technology-based behavioral interventions, and (3) De la Vega et al's [6] post hoc app of this recategorized framework against Glasgow's Reach Effectiveness Adoption Implementation Maintenance (RE-AIM) framework [7,8]. Determinant frameworks answer the question: 'Why was the intervention/practice/innovation implemented or not implemented?' Several frameworks applied to DHI implementation research include the Consolidated Framework for Implementation Research (CFIR) [9-12], the Theoretical Domains Framework [13], Promoting Action on Research Implementation in Health Services [14], and others. Application of these frameworks for DHI trials is often done post hoc rather than proactively, and practical guidance on incorporating frameworks into DHI hybrid trial planning is limited, despite their importance in revealing why implementation of DHIs tested within real-world settings did or did not meet expectations.

The *PositiveLinks* platform is a clinic-embedded multifeature smartphone app with patient and provider-facing components. It was developed and refined following a rigorous, iterative process of user-centered design to support people with HIV receiving outpatient care [15]. Program implementation among a cohort in Central Virginia where the intervention was developed and refined has demonstrated long-term usage and significant improvement in clinical outcomes at 1 [16], 2 [17], and 3 years [18]. The platform has been adapted for other chronic conditions, end users, and contexts [19-24]. To date, *PositiveLinks* has been implemented as part of routine clinical care at 9 clinics in Virginia, and 8 sites in other states, and is considered an evidence-based intervention for HIV care by several national consensus guidelines [25-28]. Clinical effectiveness of PositiveLinks is currently being tested against usual care in a hybrid effectiveness-implementation trial, the *PositiveLinks* in DC Cohort Study, using a cluster randomized controlled trial design (ClinicalTrials.gov NCT04998019) [29]. The trial is being conducted among sites in the DC Cohort Longitudinal HIV Study (DC Cohort Study) following over 12,800 people with HIV at 14 outpatient HIV practice settings, including Federally Qualified Health Centers and academic medical centers [30]. The DC Cohort context for this trial, including cohort site characteristics, as well as study design, site selection, randomization, recruitment, data collection, and statistical analysis procedures, is outlined in the study protocol paper [29].

For complex DHIs like *PositiveLinks* engaging multiple end users collectively within an 'implementation climate,' it is important to establish readiness for the coordinated actions needed to implement the intervention across the organization. The theory-based measure, Organizational Readiness for Change (ORIC), was designed and validated by Weiner et al [31] to measure collective readiness for implementation of health care innovations. We opted to measure organizational readiness at each intervention site, including the survey items within provider baseline surveys, in order to understand (1) which relatively lower-scoring sites may require additional support or attention during implementation and (2) to understand at the back end postimplementation if baseline readiness was an influencing factor in provider adoption of *PositiveLinks*. To this end, we also assessed baseline provider usage of mHealth or digital health tools to understand how this experience might shape

XSL•FO
RenderX

provider adoption of a new tool within each 'implementation climate.'

We share our process to proactively define, prioritize, and operationalize evaluation of relevant implementation outcomes and determinants for this type II hybrid effectiveness-implementation trial [4] testing a complex DHI among six DC Cohort sites randomized to the intervention over a 12-month study period. We highlight the unique opportunities and challenges that emerged for planning of a hybrid DHI trial among a large-scale epidemiologic cohort study population. This manuscript shares a practical set of takeaways and considerations that stood out to us as novel or distinct from the available literature we reviewed as we prepared for our DHI trial, that is, what we 'wish we knew' before embarking on our extensive planning stage. Several lessons are considerations for teams that would ideally be incorporated as early as the study conception and design stage. Finally, we share the results of an assessment of provider baseline technology usage and pretrial readiness to implement the intervention across participating sites and discuss how results may inform posttrial analyses of implementation outcomes and determinants.

Methods

Study Team and Process Refinement

The hybrid trial planning phase spanned an 18-month period preceding onboarding of the first site in December 2022, conducted by an interdisciplinary research team. Research team members hold an established record of clinical research experience, including conducting formative evaluations and observational studies testing clinical efficacy of *PositiveLinks*. Investigators' primary expertise includes clinical psychology, program evaluation, qualitative methods, instructional design, software development, and implementation research. Program managers contributed empirical observation of *PositiveLinks* implementation processes over a decade that assisted with conceptualization of the intervention, implementation contexts, and components of the implementation strategy. The trial planning team also included DC Cohort Study investigators with expertise in epidemiologic and intervention studies at the cohort sites. The methodological approach to proactively integrate implementation evaluation activities within the hybrid trial required iterative steps conducted through multiple cycles of team feedback, consensus, and refinement (Checklist 1).

Specify Components of the DHI and Associated Implementation Strategy

Given the multifeature, multiuser nature of the intervention, we created a specified list of components of the DHI implementation process and discrete steps of the implementation strategy. This process was informed by team experience with implementation of *PositiveLinks* in other contexts as part of usual care, including a prior rigorous qualitative study summarizing key in-clinic processes necessary for *PositiveLinks* implementation [32], and a formative preimplementation study engaging stakeholders within the DC Cohort Study context to tailor the app and implementation strategy [33]. Implementation strategy steps were further specified in terms of actors, corresponding actions, and action targets mapped specifically to the DC Cohort Study context [34,35].

Select Appropriate Implementation Science Frameworks to Accomplish Evaluation Aims

We identified aims for the implementation evaluation arm of the hybrid trial: (1) define and measure implementation outcomes of interest and (2) elucidate determinants of implementation in a rapid, scaled fashion across participating sites. We first used narrative reviews of theories, models, frameworks, and strategy categorization to assess the most widely used determinant and evaluation frameworks [36,37]. We then reviewed technology-specific compendia and original research studies reconceptualizing broader health service frameworks toward DHIs, including hybrid study designs [2,5,6,8,10]. We subsequently selected the RE-AIM evaluation framework and CFIR determinant framework for our first and second implementation evaluation aims, respectively.

Map Framework Dimensions, Domains, and Constructs to Steps of the Implementation Strategy

We mapped each dimension of the evaluation framework, RE-AIM, to specified components of the intervention and implementation strategy steps, informed by prior efforts in the literature [2,6]. We prioritized measurement of specific steps based on impact of findings for informing future *PositiveLinks* implementations in this context and others, and the feasibility of measurement within the hybrid trial study. For implementation determinants, we selected salient domains/constructs identified from our prior detailed qualitative CFIR-guided assessment of several sites implementing *PositiveLinks* [32]. We also identified salient CFIR interview guide questions for conversion into survey items, which could be rapidly deployed among a larger sample of up to 50 or more cohort providers employed across trial intervention sites at multiple predefined time points (baseline, 6 months, and 12 months).

Modify or Create Instruments to Support Data Collection for Implementation Outcome Measures and Determinants

Existing data collection instruments developed for the DC Cohort Study or for prior *PositiveLinks* real-world implementation were modified with additional items or created *de novo* within Research Electronic Data Capture to completely assess each RE-AIM implementation outcome measure and salient implementation determinants. Provider survey items were generated using close-ended questions (eg, dichotomous or Likert scale responses), as well as optional free-text responses, and planned for distribution at 6- and 12-months into implementation. Semistructured interviews and focus groups were planned to elicit more detailed feedback post-implementation among a smaller subset of both provider and patient participants, respectively, with guides designed using CFIR.

XSL•FO
RenderX

Develop a Compatible Data Collection and Management Plan for Implementation Evaluation

Finally, we generated an overall plan for participant data collection and management that ensured compatibility between the clinical effectiveness arm of the trial and the implementation evaluation. Plans for implementation outcome or determinant data collection and abstraction were incorporated into patient approaches for follow-up, study monitoring, and data abstraction already planned for the trial's effectiveness arm. Provider-related activities specific to the implementation evaluation were incorporated into study onboarding processes.

Measurement of Baseline Provider Technology Usage and Implementation Readiness

Providers completed an electronic baseline assessment upon enrollment in the study with items related to familiarity, knowledge, and usage of available patient and provider-facing mHealth tools, apps, and portals in routine HIV care, probing specific experience with mHealth tools sharing functionality with components of the *PositiveLinks* intervention. The ORIC 12-item tool [31,38,39] assesses organizational members' shared resolve to implement a change (change commitment) and shared belief in their collective capability to do so (change efficacy). The ORIC was distributed as part of the provider baseline survey to provider participants at intervention sites.

Analysis

Descriptive statistics were used to analyze responses to provider survey items related to baseline technology usage and implementation readiness (close-ended response options or free-text responses). Each ORIC item is rated on a 5-point ordinal scale (1="disagree" to 5="agree"). Responses were analyzed for frequency, median value, or free-text content as appropriate. ORIC scores were characterized using medians and IQR for total scores (sum of all items) and individual subscores related to change commitment (n=5 items) and change efficacy (n=7) [38]. Statistical analysis was conducted using R version 4.1.2 (R Foundation for Statistical Computing).

Ethical Considerations

Ethical approval was obtained upon human subject research ethics review by the George Washington University Institutional Review Board (IRB) (Protocol NCR202829; ClinicalTrials.gov NCT04998019) and site-specific IRBs as required. Written study-specific informed consent for all procedures of this trial was collected in addition to the consent obtained for inclusion of clinical data among patient participants upon enrollment in the DC Cohort Study. Written informed consent was obtained for all trial provider participants using procedures outlined in the IRB-approved study protocol. Data collection, storage, and management followed all outlined procedures (eg, deidentification of baseline survey data and use of an assigned study ID with a separate link long). No monetary compensation was provided to participants for study activities described in this paper.

Results

Key Process Steps and Takeaways

Our hybrid DHI trial planning process is summarized in Figure 1, including key process steps and takeaways for research teams aiming to perform similar prospective hybrid trials in real-world settings, where complex DHI implementation and associated study procedures can be planned in advance. We highlight takeaways that emerged from our experience and lacked more applied experience or specific guidance within the literature.

Figure 1. This figure summarizes the 5-step process to plan an implementation evaluation within a hybrid trial for prospective testing of a digital health intervention (DHI). Process steps are outlined, along with respective generalizable takeaways for research teams planning comprehensive implementation evaluations for trials testing DHIs.

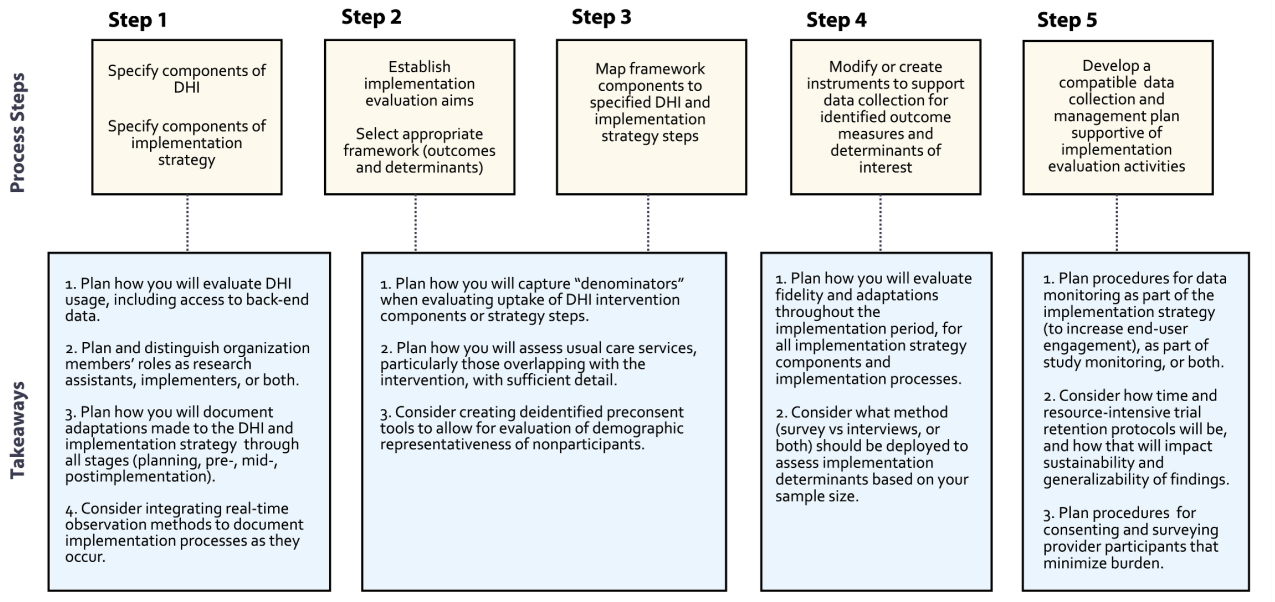

XSL•FO
RenderX

Define Components of Intervention and Implementation Strategy

Step 1 of the planning phase included specification of *PositiveLinks* intervention components (both patient and provider-facing elements) as well as the respective paradata generated by usage of each feature. The *PositiveLinks* platform contains a large amount of backend data (often described as metadata or paradata) that can be analyzed individually for every participant by feature and is accessible to the study team for analysis. Specific paradata were defined by end user group and data format (individual activity reports over selected time intervals with frequency of use, or free-text post content) in order to plan metrics of uptake for related implementation outcomes defined in steps 2 and 3. Takeaway 1: Research teams should consider during the planning stage what backend data is directly accessible for tools they are studying early in the planning process. Other pragmatic DHI trials may test commercially available tools developed by external companies or vendors, necessitating data-sharing agreements with vendors or suppliers of these tools, proactive design of trial assessments that allow participants to self-report usage at regular (eg, weekly) intervals, or integration of other commercially available tools to track DHI usage into study budgeting, design, consenting, and data collection processes.

Steps of the planned implementation strategy specified for the DC Cohort context are shown in Figure 2. “Actors” include the *PositiveLinks* core team’s program managers who assist remotely with implementation activities and troubleshooting at partner sites. The study team includes researchers involved with coordinating the specific intervention study being conducted to test *PositiveLinks* in the DC Cohort. Study team members were planned to have a larger role in site onboarding processes (determine site needs and resources, conduct research assistant and provider training) within the hybrid trial when compared to real-world implementation. Research assistants employed at DC Cohort Study sites are familiar with recruiting and consenting patients, uploading relevant patient data, and conducting assessments throughout the study period. Site research assistants were thus selected as actors to conduct several steps of the implementation strategy in this context, whereas in real-world implementation, the program is administered by individuals with a range of roles, typically identified by leadership or self-identified as the site lead for implementation. This was a decision made across sites given the extensive role and protected time research assistants have to assist with studies. Takeaway 2: Research teams designing or planning hybrid trials must consider “actors” both in terms of their research roles and implementation roles. If implementation activities in a trial are primarily assigned to a research team member, organization members already responsible for existing usual care activities overlapping with the intervention (eg, for *PositiveLinks*, retention-in-care activities) should also be engaged as early as possible. Shifting more implementation responsibilities to a research team member reduces the pragmatic nature of the trial and challenges intervention sustainability, as teams must decide who will execute those responsibilities when the trial is over.

XSL•FO
RenderX

Figure 2. Discrete implementation strategy steps expected for program deployment at DC Cohort sites are demonstrated in the figure and described, specified by step, actions, and actor in the table below it. The arrows show the sequential nature of the activities, but also note the gray dotted arrows indicate the iterative nature of using evaluation and feedback to inform various process steps as implementation proceeds. HIPAA: Health Insurance Portability and Accountability Act; PL: PositiveLinks

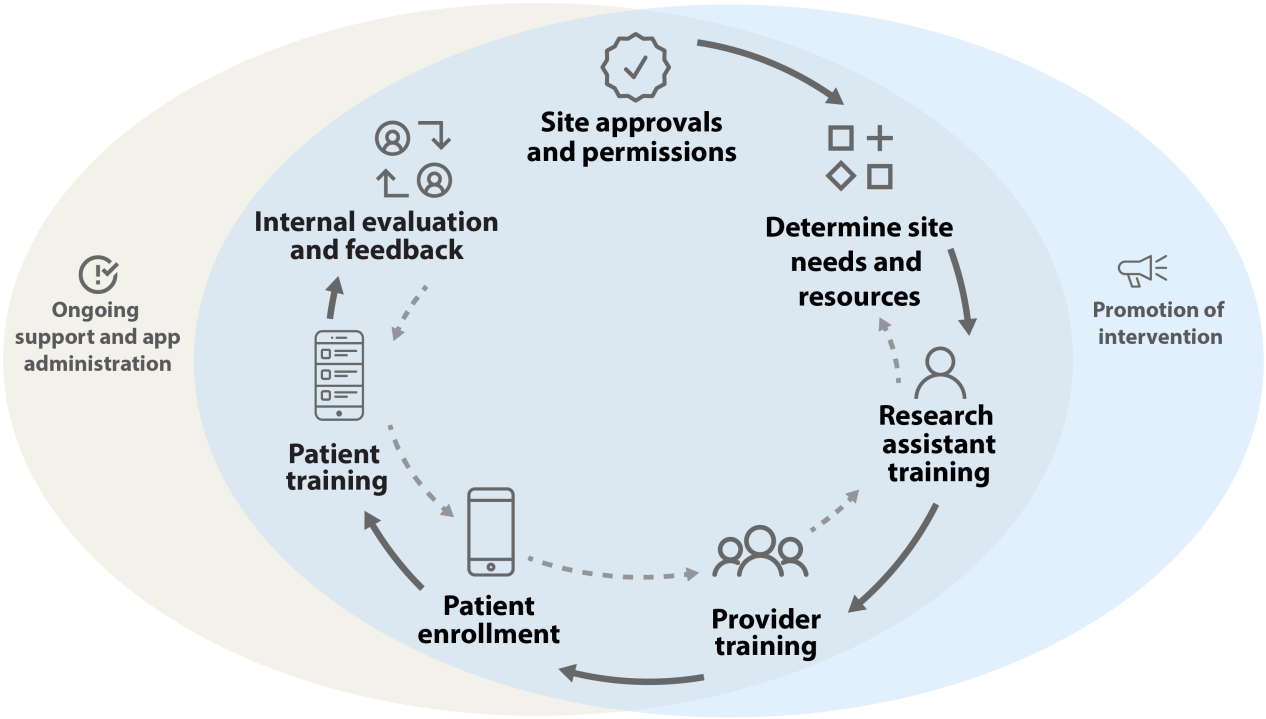

Step in implementation strategy	Actions	Actors	Action target
Site approvals and permissions	• Promote intervention to clinic and higher level leadership (study team) • Secure organizational approval of site use of intervention (study team, site principal investigator) • Identify and obtain required permissions related to patient IT security or privacy or HIPAA (study team, site principal investigator)	• Study team • Site principal investigator	• Organizational leadership
Determine site needs and resources	• Clarify roles of providers in implementation, establish staffing needs • Establish how providers will communicate with each other about program • Identify clinic values or clinical goals and how they may or may not align with PL • Determine how clinic providers would like to use portal (content shown, functionality) • Identify patient subgroups likely to benefit, needs that can be addressed with PL and issues that may arise for them related to using PL	• Study team • Site research assistant • Site Providers	• Site providers
Promotion of intervention	• Disseminate information about PL or site-specific benefits of PL to rest of site and larger organization	• Study team • Site principal investigator • Site research assistant	• Site providers • Patients • Organizational leadership
Research assistant training	• Inform site research assistant of app features so they understand how to use the app themselves and can explain benefits to providers and clinic processes, and benefits to their patients	• PL core team • Study team	• Site research assistant
Provider training	• Inform providers of app features so they understand how to use the app themselves • 'Sell' providers on the app's potential benefits to their patients, efficacy • 'Sell' app functionalities (mechanics of program) to site providers and how it will benefit them or their ability to provide care • Ensure newer providers are aware of procedures if there is turnover or new hiring	• Site research assistant	• Site providers
Patient enrollment	• Offer enrollment in the app as part of the research study and obtain informed consent • Provide preview of benefits or ease of use to patient • Sign patient up on the spot (requires WiFi access, time during appointment) • Provide smartphone or dataplan if needed	• Site research assistant	• Patients
Patient training	• Inform patient of how to navigate app interface or features and explain each in real time • Demonstrate app security or anonymity to patient • Demonstrate the customizable way patient can use the app	• Site research assistant	• Patients
Ongoing Support and app administration	• Communicate with PL core team about logistical or technical issues (site research assistant) • Troubleshoot patients' technical issues in real time or in-house when core team not available (site research assistant) • Assist research assistant with software or advanced technical issues nonurgently (PL core team)	• Site research assistant • PL core team	• PL core team • Site research assistant
Internal evaluation and feedback	• Perform surveillance or tracking of program progress, present to rest of site providers (study team and site research assistant) • Develop and carry out action plans based on interim site-specific progress or feedback (study team, site research assistant and site providers) • Ongoing champion activity to support maintenance of intervention (site providers and site research assistant)	• Study team • Site research assistant • Site providers	• Site providers • Patients

Distinctions between real-world implementation processes and those expected for the trial emerged at this planning stage. The learning management system (LMS) is a series of interactive online training modules developed for providers and staff, aimed

XSL•FO
RenderX

primarily at increasing knowledge and familiarity with the use of platform features (followed by a posttraining feedback survey). Prior to site onboarding for the study, several cohort site principal investigators expressed they would expect significantly lower enrollment of providers if LMS completion was mandatory, primarily due to significant provider clinical burden or competing demands. As a result, participating providers could opt out of LMS completion but were expected to undergo group onboarding or one-on-one training with the site research assistant prior to study participation. Takeaway 3: Research teams should consider incorporating a process to regularly document throughout the pre-, mid-, and postimplementation period, how their specific instance of DHI implementation for the trial diverges from other implementations, adaptations, or real-world practice, representing fidelity of implementation through every stage (not just during the intervention implementation period).

Real-world *PositiveLinks* implementation offers collaboration between sites in a "Community of Practice," to share implementation challenges, solutions, and unique adaptations to improve patient experiences and engagement. In contrast, we sought to maintain internal validity and avoid cross-site contamination for the trial, so intervention sites were not included in the "Community of Practice." Additionally, steps like administration of as-needed technical app support/troubleshooting and promotion of the intervention are usually more informally evaluated through site discussions and internal meetings during real-world implementation. During planning, we determined they would be challenging to evaluate with structured tools and opted to evaluate them through qualitative interviews at completion of the study. Takeaway 4: Clinic-embedded DHIs implemented in real-world practice can involve complex interactions between actors both across sites and within sites, and over time. Reserving evaluation of these processes to the late stages of implementation can result in missed opportunities to capture important real-world factors influencing implementation. Within the constraints of budget and time, research teams should consider inclusion of study methods to obtain real-time observations of these implementation processes as they occur. These could include direct observations (with field logs or notes) by implementers (providers, RAs), and small "check-ins" or "debriefs" conducted by a study team member throughout the study period.

Select Appropriate Frameworks and Map Components to Implementation Strategy Steps

Output for steps 2 and 3 are summarized in Table 1. Specifically, RE-AIM evaluation framework dimensions were mapped to each prioritized step within the implementation strategy. During this step, we identified discrete datapoints that required collection to assess each RE-AIM dimension. Other DHI hybrid trials with detailed reporting of framework app describe similar approaches, aligning DHI usage and other aspects of program uptake by patients and providers with reach and adoption dimensions, respectively [6,32]. Several unique takeaways emerged in our effort.

XSL•FO
RenderX

Table . RE-AIM[a] framework mapped to intervention/implementation strategy, corresponding outcome measures, data collection methods[b].

RE-AIM dimension	Outcome measures	Data source or data collection instrument
Reach: *How do I reach the targeted population with the intervention?*	Number of patients offered enrollment/number seen at site	DC Cohort Database, DC Cohort Study Patient Consent Logs
	Number of patients completing enrollment/number offered	PL[c] Platform Paradata, DC Cohort Study Patient Consent Logs
	Number of patients accessing PL app/number enrolled	PL Platform Paradata
	Sociodemographics of patients enrolled (representativeness)	DC Cohort Database
	Sociodemographics of patients declining enrollment (representativeness)	DC Cohort Study Patient Consent Logs
Effectiveness: *How do I know my intervention is effective?*	HIV viral load suppression (<200 copies/mL)	DC Cohort Database
	Visit constancy	DC Cohort Database
	Retention in care	DC Cohort Database
	Demographics for patients meeting efficacy end points versus those not meeting	DC Cohort Database
	PL usage patterns for patients meeting efficacy end points versus those not meeting	PL Platform Paradata
Adoption: *How do I develop organizational support to deliver my intervention?*	Number of providers who completed onboarding/number employed	Provider Baseline Survey, DC Cohort Site Assessment Survey (Multimedia Appendix 1)
	Number of providers who completed LMS[d]/number enrolled	PL Posttraining feedback survey
	Number of providers accessing PL app/number enrolled	PL Platform Paradata
	Demographics for providers	Provider Baseline Survey (Multimedia Appendix 2)
	Cohort site characteristics	DC Cohort Site Assessment Survey
Implementation: *How do I ensure the intervention is delivered properly?*	Dose, fidelity, and adaptations: patient training	Provider Follow-up Survey (Multimedia Appendix 3), Patient Interview Guides
	Dose, fidelity, and adaptations: provider training	PL Posttraining feedback survey, Provider Interview Guides
	Dose, fidelity, and adaptations: provider PL usage	PL Platform Paradata, Provider Follow-up Survey, Provider Interview Guides
Maintenance: *How do I incorporate the intervention so that it is delivered over the long term?*	Patient intent/interest to continue using PL following trial completion	Patient Interview Guides
	Provider intent/interest to continue using PL following trial completion	Provider Follow-Up Survey, Provider Interview Guides

[a]RE-AIM: Reach Effectiveness Adoption Implementation Maintenance.

[b]Data collection methods include abstraction from one or more items from existing data sources, modification or addition of items to existing data collection instruments, or creation of new instruments. Visit constancy is the proportion of 4-month intervals a visit is completed in 12 months. Retention in care per HRSA-1 definition: 2 appointments attended at least 90 days apart within 12 months.

[c]PL: PositiveLinks.

[d]LMS: learning management system.

As we reviewed the reach and adoption dimensions, we conceptualized how we could measure not only absolute numbers as measures of uptake (eg, number of patients/providers who download and use the app), but also estimate proportions: how many people could have taken up the DHI when it was offered at each site? The DC Cohort Site Assessment Survey is periodically deployed to cohort sites for updated assessment of available service delivery. We added several items to this Site Assessment Survey to capture the denominator of people employed for each type of provider role (eg, attending physician, clinic nurse, case manager, social worker, etc) as well as site-level baseline mHealth or technology use (Multimedia Appendix 1) to more adequately assess the adoption dimension for *PositiveLinks* implementation. Takeaway 1: Research teams should consider in advance how they will capture 'denominators' for both patients (reach) and providers/staff

XSL•FO
RenderX

serving as implementers (adoption), including a need for pretrial site-level data collection that can provide denominators, and with what frequency they need to be deployed surrounding implementation.

Several existing Site Assessment items were identified for abstraction for evaluation of the adoption dimension of RE-AIM (eg, on-site clinical services and support services, updated staffing of clinical and nonclinical providers, specialty training). Takeaway 2: Research teams testing DHIs across multiple care settings must consider how they will sufficiently characterize existing "usual care" services available, which can vary widely across care settings and change over time. This step is particularly important for DHIs that do not study direct-to-consumer tools, but instead require DHIs to become integrated into, or penetrate, an implementation climate and existing practices. Understanding to what extent the tool is serving as an adjunct or supplementary tool layered on top of an already robust organizational usual practice, versus creating a service where none existed, is crucial for interpreting and contextualizing effectiveness findings from a hybrid trial, in addition to other implementation outcomes.

A distinct opportunity also arose from conducting the hybrid trial among the DC Cohort Study. DC Cohort patients who choose not to participate in the *PositiveLinks* program at intervention sites have already consented to inclusion of their data in the cohort database, so these nonparticipants could be examined to understand sociodemographic representativeness, a component of patient "reach" within RE-AIM that appears infrequently in published applications of this framework. Patient consent logs standardized for the DC Cohort Study include patients' reasons for declining participation, with up to 3 approaches. These logs were modified to query a selected number of demographics for cohort participants declining to participate in the study (age, sex, race, ethnicity, insurance status, last CD4 count, and last HIV viral load). Takeaway 3: Teams designing DHI trials outside of existing cohorts with available data should consider use of preconsent tools that collect deidentified demographics of interest among clinic patients who decline the intervention, in order to evaluate this understudied component of patient reach.

Modify or Create Instruments to Support Data Collection for Implementation Outcome Measures and Determinants

Our process yielded several strategies to support data collection of all identified implementation outcomes: (1) *creation* of new instruments for prospective data collection specific to the implementation evaluation, (2) *modification* of standardized tools used for DC Cohort intervention studies or *PositiveLinks* evaluations or (3) *abstraction* from existing *PositiveLinks* or DC Cohort sources (eg, DC Cohort Database storing patient encounter, laboratory, and sociodemographic data). Modification and abstraction planned based on steps 2 and 3 from existing data sources are described previously.

Instruments created specifically for the implementation arm of the hybrid trial included surveys directed at providers not usually targeted by data collection for intervention studies at cohort sites. The provider baseline survey (Multimedia Appendix 2) and follow-up survey (Multimedia Appendix 3), annotated with respective framework components, were designed to capture otherwise unincorporated data points for measurement for the implementation dimension of RE-AIM (eg, fidelity, dose, and any adaptations made to steps of the implementation strategy by site providers or research assistants). Many DHIs require a complex set of steps to ensure both patient and provider engagement. For *PositiveLinks*, this included providers themselves understanding the app features, how they work and their impact, then remembering and being motivated to bring up the app and promote it during a routine clinical encounter, be able to describe its features and benefits, and refer the patient to a staff member (eg, research assistant). The research assistant must then effectively assist the patient to download the app, register for the account, and train them on its usage (Figure 2). The provider follow-up surveys were thus designed to capture the extent to which providers performed each of these tasks over multiple time points (Multimedia Appendix 3), in addition to questions probing self-rated fidelity and adaptations made to the use of the tool and its features over multiple time points. Takeaway 1: Comprehensive evaluations of DHIs should consider fidelity and adaptations of aspects of implementation beyond direct end user engagement with the digital tool, and data collection instruments may consequently need to be created to assess implementation outcomes related to these specific steps over multiple time points.

Salient implementation determinants adapted for inclusion from our prior CFIR-guided rapid evaluation study of *PositiveLinks* real-world implementation were: inner setting (compatibility), outer setting (patient needs and resources, external policy, and incentives), characteristics of individuals (knowledge and beliefs), innovation characteristics (adaptability and complexity), and implementation process (planning and engagement of key stakeholders). Postimplementation patient focus group and provider in-depth interview guides were also adapted from prior *PositiveLinks* implementations using salient CFIR 1.0 domains or constructs. Takeaway 2: Surveys designed using CFIR or other determinant frameworks offer an opportunity to more rapidly probe a wide array of domains and constructs among a larger sample, but should be planned in conjunction with richer data collection methods (qualitative). The surveys we designed were limited to previously identified salient constructs during our rapid evaluation study in other contexts (cite), and if used alone would miss important contextual factors for this trial.

Develop a Compatible Data Collection and Management Plan for Implementation Evaluation

Finally, we developed a plan to specify timing and frequency of data abstraction (eg, from the DC Cohort Database) and collection in relation to planned activities for the effectiveness arm of the trial (eg, patient consent or enrollment, administration of baseline, 6 mo, and 12 mo assessments). For patient data, plans were designed to minimize separate approaches as well as 'data pulls' from existing sources anticipated to support evaluation of clinical effectiveness outcomes. Further, monitoring of feature usage from platform paradata for patients and providers is a routine part of real-world *PositiveLinks* implementation, to guide efforts to engage and re-engage staff

XSL•FO
RenderX

and enrolled patients, troubleshoot concerns in real-time, and ensure sustainability. Frequency of paradata abstraction for monitoring was thus predetermined for specific features at timepoints throughout implementation. Takeaway 1: Research teams should consider data monitoring both as a study activity and as a part of the implementation strategy. If regular data monitoring is expected to increase engagement or generate actionable data for implementers, this activity itself becomes a component of the implementation strategy and should both be integrated into program planning and measured for fidelity to plans along with other strategy components. Takeaway 2: Similarly, trial retention protocols whereby participants showing low engagement are contacted by study team members through a series of time and resource-intensive activities ultimately impact intervention uptake (reach and adoption). The design of these protocols should be considered in the interpretation of generalizability/sustainability of observed implementation outcomes.

We required informed consent for all provider activities in the evaluation, including collection of provider survey responses and participation in postimplementation in-depth interviews. Selection of the timing and frequency of provider survey administration required consideration of provider turnover, particularly in participating intervention sites with higher expected turnover (eg, trainees like infectious disease fellows rotating within academic centers). Takeaway 3: Implementation evaluations engage providers as participants, which is distinct from most intervention efficacy trials, and onboarding processes should consider integrating consent procedures and baseline survey administration to reduce the burden and frequency of study procedures on participating providers.

Baseline mHealth/technology Use and PositiveLinks Implementation Readiness

A total of 17 providers and RAs have completed provider baseline surveys to date. Self-reported mHealth/technology use for various aspects of patient care at baseline is summarized in Table 2. Among the 17 respondents, 9 reported access to a patient messaging feature via their electronic medical record system (EMR). Usage of additional non-EMR messaging tools was reported by 6 providers. Overall satisfaction with non-EMR methods of messaging was high for those reporting usage (n=6), and reported frequency of use was higher for non-EMR tools compared to EMR-based messaging. Fewer respondents completed optional survey items related to non-EMR tools for sharing lab results or exchanging documents with patients, with variable frequency and satisfaction with described tools.

We incorporated the ORIC tool into the provider baseline survey to assess collective baseline readiness at each site [31]. The median total ORIC score for all providers was 48 (IQR 45 - 54, possible score range 12 - 60), with RAs scoring slightly higher than physicians (52.5 (44-55) vs 48.0 (46-49)). Median total scores for change commitment and change efficacy were 20 (IQR: 19 - 22, possible score range 5 - 25) and 28 (27 - 30, possible score range 7 - 35). Total change efficacy scores were slightly higher for RAs compared to providers (30.5 vs 28.0).

XSL•FO
RenderX

Table . Provider and research assistant baseline usage of mobile health (mHealth) tools by functionality.

Baseline survey questions	Response rate, n (%)
Messaging	
Does your EMR[a] system have a patient portal that allows you to directly message with your patients? (n=17)	
Yes	9 (53)
No	6 (35)
Not sure	2 (12)
If yes, how frequently do you use it to message with your patients? (n=9)	
Never	2 (22)
Rarely	2 (22)
Occasionally	3 (33)
Frequently	2 (22)
Very frequently	0 (0)
Please specify name of apps or tools or websites used to message patients (n=6)	
Ring central	2 (33)
Halo	1 (17)
EMR mobile or web-based platform	1(17)
Text messaging	2 (33)
In the past 3 months, how many times have you used this app/tool/site to message patients? (n=6)	
Never	0 (0)
Rarely	0 (0)
Occasionally	1 (17)
Frequently	4 (67)
Very frequently	1 (17)
To what extent are you satisfied with these telemedicine services used for messaging? (n=6)	
Very unsatisfied	0 (0)
Unsatisfied	0 (0)
Neutral	0 (0)
Satisfied	4 (67)
Very satisfied	2 (33)
Laboratory results	
Please specify name(s) of app/tool/website(s) to share lab results. (n=3)	
EMR autosend letter	1 (33)
EMR mobile or web-based platform	1 (33)
Clinic phone	1 (33)
In the past 3 months, how many times have you used this app/tool/site to share laboratory results? (n=3)	
Never	0 (0)
Rarely	0 (0)
Occasionally	1 (33)
Frequently	1 (33)
Very frequently	1 (33)
To what extent are you satisfied with this app or tool or website for sharing laboratory results? (n=3)	
Very unsatisfied	0 (0)

XSL•FO
RenderX

Baseline survey questions	Response rate, n (%)
Unsatisfied	0 (0)
Neutral	0 (0)
Satisfied	2 (67)
Very satisfied	1 (33)
Documents	
Please specify the name of app or tool or website to share or receive documents (n=2)	
Encrypted email	1 (50)
EMR mobile/web-based platform	1 (50)
In the past 3 months, how many times have you used this app/tool/site to share or receive documents? (n=2)	
Never	0 (0)
Rarely	0 (0)
Occasionally	1 (50)
Frequently	1 (50)
Very frequently	0 (0)
To what extent are you satisfied with this app or tool or website for sharing or receiving documents? (n=2)	
Very unsatisfied	0 (0)
Unsatisfied	1 (50)
Neutral	0 (0)
Satisfied	1 (50)
Very satisfied	0 (0)

[a]EMR: electronic medical record.

Individual items corresponding to change efficacy (CE1-CE7) were scored on average between 4 and 4.2, with median scores of 4 (IQR 3-4) for all items (“Somewhat agree”), and no items scoring lower than 3 (“Neither agree nor disagree”; Figure 3). All change efficacy scores ranged from 3 to 5 except for CE7 (‘People who work here feel confident that they can manage the politics of implementing this change’), which along with CE3 (”People who work here feel confident that the organization can support people as they adjust to this change”) had a distribution with greater spread due to relatively higher frequency of neutral or “3” scores.

ORIC items corresponding to change commitment (CC1-CC5) similarly scored with a median of 4 (IQR 3-5) for all items (Figure 4), means ranging from 3.88 to 4.2, and scores also ranged from 3 to 5 for all items. Change commitment score distributions were slightly more negatively skewed for items CC2 (“People who work here will do whatever it takes to implement this change”) and CC4 (“People who work here are determined to implement this change”). When analyzed by provider type, mean scores were higher across all individual ORIC items for RAs compared to physicians. Median scores within subgroups were 4 (with IQR values ranging from 3.5 to 5, Figure 4) for a majority of items except for two change commitment items with higher median scores for RAs: CC3 (“People who work here want to implement this change”) and CC5 (“People who work here are motivated to implement this change”).

XSL•FO
RenderX

Figure 3. Baseline ORIC scores summarized for all respondents (n=17) at 6 sites randomized to implement the intervention (A) as well as by respondent type such as research assistants and attending physicians (B). One social worker completed the survey but ultimately withdrew from the program. ORIC: Organizational Readiness to Implement Change.

A

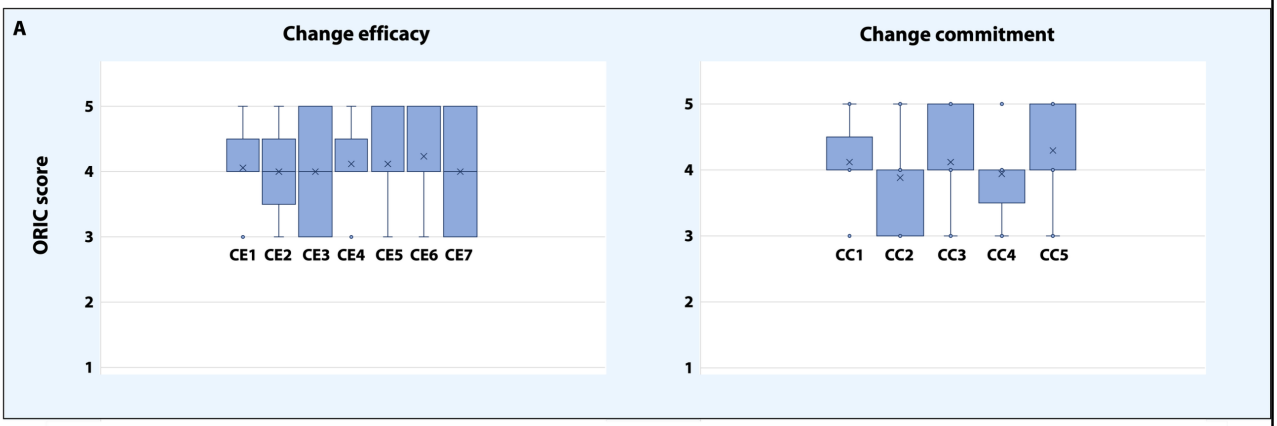

B

ORIC Item		All (n=17)	Research assistants (n=9)	Attending physicians (n=7)
Change efficacy				
CE1: People who work here feel confident that the organization can get people invested in implementing this change.	Median (IQR)	4.0 (4,4)	4.0 (4,5)	4.0 (3.5,4)
	Mean (SD, Range)	4.1 (0.7, 3-5)	4.3 (0.5, 4-5)	3.9 (0.7, 3-5)
CE2: People who work here feel confident that they can keep track of progress in implementing this change.		4.0 (4,4)	4.0 (4,5)	4.0 (4,4)
		4.0 (0.7, 3-5)	4.2 (0.8, 3-5)	3.9 (0.4, 3-4)
CE3: People who work here feel confident that the organization can support people as they adjust to this change.		4.0 (3,5)	4.0 (4,5)	4.0 (3,4)
		4.0 (0.8, 3-5)	4.2 (0.8, 3-5)	3.6 (0.5, 3-4)
CE4: People who work here feel confident that they can keep the momentum going in implementing this change.		4.0 (4,4)	4.0 (4,5)	4.0 (3,4)
		4.1 (0.6, 3-5)	4.3 (0.7, 3-5)	4.0 (0.0, 4-4)
CE5: People who work here feel confident that they can handle the challenges that might arise in implementing this change.		4.0 (4,5)	4.0 (4,5)	4.0 (4,4)
		4.1 (0.7, 3-5)	4.1 (0.8, 3-5)	4.0 (0.6, 3-5)
CE6: People who work here feel confident that they can coordinate tasks so that implementation goes smoothly.		4.0 (4,5)	4.0 (4,5)	4.0 (4,4)
		4.2 (0.7, 3-5)	4.2 (0.8, 3-5)	4.1 (0.4, 4-5)
CE7: People who work here feel confident that they can manage the politics of implementing this change.		4.0 (3,4)	4.0 (4,5)	4.0 (3.5,4)
		4.0 (0.8, 3-4)	4.2 (0.8, 3-5)	3.9 (0.7, 3-5)
Change commitment				
CC1: People who work here are committed to implementing this change.	Median (IQR)	4.0 (4,4)	4.0 (4,5)	4.0 (4,4)
	Mean (SD, Range)	4.1 (0.6, 3-5)	4.4 (0.5, 4-5)	3.9 (0.4, 3-4)
CC2: People who work here will do whatever it takes to implement this change.		4.0 (3,4)	4.0 (4,5)	4.0 (3.5,4)
		3.9 (0.7, 3-5)	4.1(0.8, 3-5)	3.7 (0.5, 3-4)
CC3: People who work here want to implement this change.		4.0 (4,5)	5.0 (4,5)	4.0 (4,4)
		4.1 (0.7, 3-5)	4.3 (0.9, 3-5)	4.0 (0.0, 4-4)
CC4: People who work here are determined to implement this change.		4.0 (4,4)	4.0 (4,5)	4.0 (4,4)
		3.9 (0.7, 3-5)	4.1 (0.8, 3-5)	3.9 (0.4, 3-4)
CC5: People who work here are motivated to implement this change.		4.0 (4,5)	5.0 (4,5)	4.0 (4,4.5)
		4.3 (0.7, 3-5)	4.4 (0.7, 3-5)	4.3 (0.5, 4-5)

Discussion

Principal Results

We describe a study team's experience planning and proactively integrating implementation science methodology into a hybrid DHI study conducted within an existing cohort study. This experience included a process to proactively define, prioritize, and operationalize measurement of implementation outcomes salient to a prospective DHI hybrid trial, as well as development of data collection instruments compatible with rapid assessment of implementation determinants and outcomes at scale for a large cohort and over multiple time points throughout the trial. Our experience highlights the limitations of prospective "pragmatic" trial designs to reflect real-world DHI implementation, and the unique opportunities and challenges of DHI trial planning within a large-scale cohort study population. We assessed baseline provider mHealth or technology usage and organizational readiness at baseline, which will be applied as a site-level characteristic or covariate during exploratory analysis of differential site and provider-level adoption of the intervention, including specific features, throughout the implementation period. We predict findings may align with site-specific implementation determinants revealed by CFIR application (eg, innovation relative advantage) during qualitative analysis, as well as account for variation in implementation outcome measures (eg, adoption by platform feature) across sites where providers and research assistants, for example, were already frequently using existing digital tools overlapping with *PositiveLinks* functionality that were fully integrated into their clinic's EMR with a high degree of satisfaction and had high ORIC scores. Readiness assessment results varied across sites but were not sufficiently different to prompt differential approaches to implementation support across sites. Results suggest that among the providers and RAs queried at the 6 intervention sites preimplementation, belief in their sites' collective capability to implement the program was high overall, particularly in relation to internal abilities to coordinate tasks and handle challenges arising within the sites. However,

XSL•FO
RenderX

respondents' confidence was lower in external actors' influence on implementation, or the organizational politics and support necessary to execute program components. To increase collective change efficacy of providers and research assistants, concerted preimplementation evaluation efforts may be needed in future *PositiveLinks* implementations to build site members' confidence that higher organizational levels of support for the program are present.

Comparison With Prior Work

To apply relevant implementation science frameworks to DHI implementation evaluation, we reviewed how investigators compared Proctor's Outcomes for Implementation Research recategorized for DHIs against RE-AIM [2,6]. These and other original research studies had limitations including a lack of measurable eligible "denominators" available for certain outcomes (eg, provider adoption dimension of RE-AIM). Prior studies also focused on conceptualizing outcomes for a primarily patient-facing intervention itself (eg, different ways of leveraging back-end usage capture data and tracking referrals of patients by clinic to the study), rather than including additional strategy steps to implement a dual-facing (patient and provider) intervention (mapped in Figure 2), which require dedicated measurement of both patient and provider inputs.

While RE-AIM offered adequate operational guidance to evaluate this DHI and implementation strategy as with some prior studies [2,6], several takeaways, opportunities, and challenges arose throughout our planning process as detailed throughout this article and summarized in Figure 1. Since the completion of our planning process and preparation of this manuscript, a framework for designing DHI hybrid trials has been published [32], which shares similarities with our identified planning process steps (specify components of the digital intervention as well as support services and implementation strategy, delineate domains being studied, ie, actors and action targets). Our takeaways delve deeper into considerations researchers should make based on specific experience with: (1) designing a randomized trial where the digital intervention is tested against a usual care condition, contains multiple features and engages both patients and providers or organization members; (2) integrating understudied components of RE-AIM and related evaluation frameworks into DHI trial design (eg, representativeness, fidelity, dose, and adaptation); and (3) conducting a trial among an epidemiologic cohort.

There is a well-recognized tension between maintaining the rigor and validity of randomized trial designs and establishing pragmatic conditions more relevant to real-world implementation [40]. By performing a preimplementation planning phase for a prospective DHI trial that incorporated frameworks proactively, we ensured collection of a robust set of quantitative and qualitative data. In contrast, post hoc evaluations conducted in other studies available in the literature are frequently reported among "real world" conditions or 'usual care' expected of implementation research as most strictly defined. When not planned in advance for inclusion, there are significant limits on the extent to which additional implementation outcomes like fidelity, dose, program adaptations, and uptake of specific components or steps of implementation have been evaluated with sufficient granularity and coverage of participants over the course of implementation [41,42] in these studies. Simultaneously, however, tighter control of prospective hybrid trial conditions and administration of the program through protocolized research activities inherently limits how well 'real-world' or pragmatic conditions are reflected within a hybrid trial.

Limitations

Several limitations emerged within our process. This more comprehensive planning process for the implementation evaluation was undertaken following receipt of funding for the award, which is common for hybrid trials. We found that ideally, this process should be undertaken as early as possible. Several takeaways from our planning process require attention this early, impacting study design and planned procedures, and consequently even impacting study budgeting and scope of work for research staff.

We noted several key limitations regarding the intended pragmatic nature of our prospective hybrid DHI trial. Real-world *PositiveLinks* implementation often relies on outreach by partner sites, by individuals who serve as champions of the intervention with an active role in obtaining site approvals (related to data security and patient privacy), and who continue to promote the intervention. Research assistants were assigned as de facto program managers at sites participating in the hybrid trial; however, this is a major distinction from real-world implementation. This decision, while needed to rapidly plan and conduct a multisite trial within budgetary and time-related constraints, represents a tradeoff in terms of generalizability of this planned prospective *PositiveLinks* implementation research when the intervention requires a distinct, multilayered 'implementation climate' and 'champions' within that climate. These 'climates' typically require gradual building of multiple, interacting implementers' self-efficacy, motivation, and longitudinal intervention promotion efforts to ensure site readiness and penetration.

Additional considerations for hybrid trials implementing DHIs may represent challenges to generalizability in terms of real-world maintenance and sustainability, including providing technology to patients (eg, smartphones, data plans), incentives for usage of the intervention or specific features, and participant retention protocols common for clinical efficacy trials. Real-world PositiveLinks implementation variably includes provision of phones and data plans, depending on specific partner site funding availability and patient need, and re-engagement protocols are also a routine part of implementation at several sites. For this trial, we planned to rely on existing site-specific or federal subsidy programs available to participants in the context (eg, Federal Communications Commission Lifeline program) before providing smartphones and data plans, and a retention protocol was used by site RAs to periodically re-engage patients. No additional incentives were planned for higher levels of app usage, however, and patients who do receive smartphones can keep them for the duration of the study regardless of app usage.

Examining implementation in parallel with a cluster randomized effectiveness trial among an epidemiological cohort presented

XSL•FO
RenderX

another set of unique challenges and opportunities. This hybrid approach allows for a scaled implementation evaluation to occur across multiple sites simultaneously, leveraging existing research infrastructure, including site staffing with research assistants and existing data collection instruments. The scaled, simultaneous multisite approach, however, engages a larger number of patients and providers within time and funding constraints of a single hybrid trial and necessitates consideration of more rapid, cost-effective approaches to implementation evaluation (acknowledged by de la Vega et al [6]). Provider surveys, for example, were designed to capture implementation outcome measures and determinants for the larger expected sample of respondents at 6 intervention sites than are typically engaged with more in-depth qualitative processes applying these frameworks, in particular CFIR. There is no consensus in the implementation science field about how these frameworks reflecting complex psychosocial/behavioral constructs should be applied, including whether to attempt to dichotomize or categorize items for broader, rapid distribution. Surveys could, for example, introduce study team bias during creation and selection of specific items to probe and limit more systematic application of the framework [37]. Combining surveys and qualitative approaches can offer opportunities to validate the latter methods, but more extensive psychometric validation of survey tools is needed to ensure generalizability and validity. Finally, readiness assessment findings are based on self-reported survey items, subject to scoring biased by the individuals' level of involvement with pretrial planning and other procedures (eg, research assistants vs physicians).

Conclusions

Implementation research for complex DHIs can expand understanding of how these interventions will behave in "real world" conditions. Prospective hybrid effectiveness-implementation trials can facilitate more in-depth implementation evaluations at scale if appropriately planned. Our experience highlights the ways in which evaluations must attempt to balance rigor, proximity to "real world" implementation climates, and incorporate multiple key implementation outcomes and determinants within the time and resource constraints of a prospective DHI hybrid trial. Based on our experience, planning processes for hybrid DHI trials should include:

- Specification of discrete DHI and associated implementation strategy components, considering how end users will engage with each, and what study procedures should be planned and budgeted to adequately measure that engagement (eg, backend paradata).
- Strategies to observe and document adaptations and real-time implementation processes throughout the planning, pre-, mid-, and postimplementation periods.
- Plans ahead of time to capture denominators of uptake outcomes (reach for patients and adoption for providers), demographic representativeness within those reached by the DHI versus not, and a method to capture usual care services that overlap with DHI functionalities across trial sites.
- Plans to evaluate fidelity and adaptations to the DHI and implementation strategy steps carried out among the implementing organization site that go beyond the use of the tool itself.
- Considerations for how to obtain detailed descriptive data related to implementation determinants in a larger sample size of participants (eg, design survey tools, interviews, or both using determinant frameworks).
- Plans for study procedures that minimize provider or implementer burden but enable consenting, data monitoring, and surveying of those providers, and consider in protocol design how retention protocols and use of research staff over "usual care" staff challenges generalizability or sustainability.

Acknowledgments

The authors would like to acknowledge all the collaborators of the participating DC Cohort clinics as well as input from members of the PositiveLinks team. No generative artificial intelligence tools were used for manuscript preparation.

Data in this manuscript were collected by the DC Cohort Study Group with investigators and research staff located at: Children's National Hospital Pediatric clinic (Natella Rakhmanina); the Senior Deputy Director of the DC Department of Health HAHSTA (Clover Barnes); Family and Medical Counseling Service (Rita Aidoo); Georgetown University (Princy Kumar); The George Washington University Biostatistics Center (Tsedenia Bezabeh, Vinay Bhandaru, Asare Buahin, Nisha Grover, Lisa Mele, Susan Reamer, Alla Sapozhnikova, Greg Strylewicz, and Marinella Temprosa); The George Washington University Department of Epidemiology (Shannon Barth, Morgan Byrne, Amanda Castel, Alan Greenberg, Shannon Hammerlund, Paige Kulie, Anne Monroe, Lauren O'Connor, James Peterson, and Mark Storey) and Department of Biostatistics and Bioinformatics; The George Washington University Medical Faculty Associates (Jose Lucar); Howard University Adult Infectious Disease Clinic (Jhansi L. Gajjala) and Pediatric Clinic (Sohail Rana); Kaiser Permanente Mid-Atlantic States (Michael Horberg); La Clinica Del Pueblo (Ricardo Fernandez); MetroHealth (Duane Taylor); Washington Health Institute, formerly Providence Hospital (Jose Bordon); Unity Health Care (Gebeyehu Teferi); Veterans Affairs Medical Center (Debra Benator and Rachel Denyer); Washington Hospital Center (Adam Klein); and Whitman-Walker Institute (Stephen Abbott).

Funding

AC and KI are multiple principal investigators on this project which is funded by the National Institute of Mental Health (R01MH122375). The DC Cohort Longitudinal HIV Study is funded by the National Institute of Allergy and Infectious Diseases

XSL•FO
RenderX

(UM1AI069503 and 1R24AI152598-01). JH received additional support from the National Institutes of Health (T32AI007046 and 5P30AI064518-20). The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets used and analyzed during the current study are available from the DC Cohort Longitudinal HIV Study following approval of request(s) from the study team and DC Cohort Executive Committee. Access is restricted in this way to protect the privacy of cohort participants, who have consented to inclusion of their data according to this restricted access policy. Requests for access can be submitted to the study authors and will be followed by a screening process first by the study team, then by the DC Cohort Executive Committee.

The datasets used and analyzed during the current study are available from the DC Cohort Longitudinal HIV Study following approval of request(s) from the study team and DC Cohort Executive Committee.

Conflicts of Interest

ALW and KI have active consulting agreements with Warm Health Technology, Inc., a wholly owned subsidiary of the University of Virginia Licensing & Ventures Group, which distributes the *Positive Links* program as *PL Cares*. JH is the founder and CEO of emPath technologies, LLC.

Multimedia Appendix 1
DC cohort site assessment survey with modifications.
[DOCX File, 18 KB - jmir_v28i1e76327_app1.docx]

Multimedia Appendix 2
Provider baseline survey.
[DOCX File, 18 KB - jmir_v28i1e76327_app2.docx]

Multimedia Appendix 3
Provider follow-up survey.
[DOCX File, 25 KB - jmir_v28i1e76327_app3.docx]

Checklist 1
CONSORT-EHEALTH (V 1.6.1) checklist.
[PDF File, 1238 KB - jmir_v28i1e76327_app4.pdf]

References

1. Muñoz RF. The efficiency model of support and the creation of digital apothecaries. Clin Psychol Sci Pract 2017;24(1):46-49. [doi: 10.1111/cpsp.12174]
2. Hermes EDA, Lyon AR, Schueller SM, Glass JE. Measuring the implementation of behavioral intervention technologies: recharacterization of established outcomes. J Med Internet Res 2019;21(1):e11752. [doi: 10.2196/11752]
3. Curran GM. Implementation science made too simple: a teaching tool. Implement Sci Commun 2020;1(1):27. [doi: 10.1186/s43058-020-00001-z] [Medline: 32885186]
4. Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. Med Care 2012 Mar;50(3):217-226. [doi: 10.1097/MLR.0b013e3182408812] [Medline: 22310560]
5. Lord SE, Campbell ANC, Brunette MF, et al. Workshop on implementation science and digital therapeutics for behavioral health. JMIR Ment Health 2021 Jan 28;8(1):e17662. [doi: 10.2196/17662] [Medline: 33507151]
6. de la Vega R, Ritterband L, Palermo TM. Assessing digital health implementation for a pediatric chronic pain intervention: comparing the RE-AIM and BIT frameworks against real-world trial data and recommendations for future studies. J Med Internet Res 2020 Sep 1;22(9):e19898. [doi: 10.2196/19898] [Medline: 32870158]
7. Glasgow RE, Harden SM, Gaglio B, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019;7:64. [doi: 10.3389/fpubh.2019.00064] [Medline: 30984733]
8. Glass JE, Dorsey CN, Beatty T, et al. Study protocol for a factorial-randomized controlled trial evaluating the implementation, costs, effectiveness, and sustainment of digital therapeutics for substance use disorder in primary care (DIGITS Trial). Implement Sci 2023 Feb 1;18(1):3. [doi: 10.1186/s13012-022-01258-9] [Medline: 36726127]
9. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implementation Sci 2009 Dec;4(1):50. [doi: 10.1186/1748-5908-4-50]

XSL•FO
RenderX

10. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146. [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
11. Rangachari P, Mushiana SS, Herbert K. A scoping review of applications of the Consolidated Framework for Implementation Research (CFIR) to telehealth service implementation initiatives. BMC Health Serv Res 2022 Nov 30;22(1):1450. [doi: 10.1186/s12913-022-08871-w] [Medline: 36447279]
12. Heinsch M, Wyllie J, Carlson J, Wells H, Tickner C, Kay-Lambkin F. Theories informing eHealth implementation: systematic review and typology classification. J Med Internet Res 2021 May 31;23(5):e18500. [doi: 10.2196/18500] [Medline: 34057427]
13. Jennings HM, Morrison J, Akter K, et al. Developing a theory-driven contextually relevant mHealth intervention. Glob Health Action 2019;12(1):1550736. [doi: 10.1080/16549716.2018.1550736] [Medline: 31154988]
14. Connolly SL, Hogan TP, Shimada SL, Miller CJ. Leveraging implementation science to understand factors influencing sustained use of mental health apps: a narrative review. J Technol Behav Sci 2021;6(2):184-196. [doi: 10.1007/s41347-020-00165-4] [Medline: 32923580]
15. Laurence C, Wispelwey E, Flickinger TE, et al. Development of PositiveLinks: a mobile phone app to promote linkage and retention in care for people with HIV. JMIR Form Res 2019 Mar 20;3(1):e11578. [doi: 10.2196/11578] [Medline: 30892269]
16. Dillingham R, Ingersoll K, Flickinger TE, et al. PositiveLinks: a mobile health intervention for retention in HIV care and clinical outcomes with 12-month follow-up. AIDS Patient Care STDS 2018 Jun;32(6):241-250. [doi: 10.1089/apc.2017.0303] [Medline: 29851504]
17. Canan CE, Waselewski ME, Waldman ALD, et al. Long term impact of PositiveLinks: clinic-deployed mobile technology to improve engagement with HIV care. PLoS ONE 2020;15(1):e0226870. [doi: 10.1371/journal.pone.0226870] [Medline: 31905209]
18. Bielick C, Canan C, Ingersoll K, Waldman AL, Schwendinger J, Dillingham R. Three-year follow-up of PositiveLinks: higher use of mHealth platform associated with sustained HIV suppression. AIDS Behav 2024 Aug;28(8):2708-2718. [doi: 10.1007/s10461-024-04405-z] [Medline: 38869759]
19. Hodges J, Zhdanova S, Koshkina O, et al. Implementation of a mobile health strategy to improve linkage to and engagement with HIV care for people living with HIV, Tuberculosis, and substance use in Irkutsk, Siberia. AIDS Patient Care STDS 2021 Mar;35(3):84-91. [doi: 10.1089/apc.2020.0233] [Medline: 33538649]
20. Hodges J, Waselewski M, Harrington W, et al. Six-month outcomes of the HOPE smartphone application designed to support treatment with medications for opioid use disorder and piloted during an early statewide COVID-19 lockdown. Addict Sci Clin Pract 2022 Mar 7;17(1):16. [doi: 10.1186/s13722-022-00296-4] [Medline: 35255965]
21. Clement ME, Lovett A, Caldwell S, et al. Development of an mHealth app to support the prevention of sexually transmitted infections among Black men who have sex with men engaged in pre-exposure prophylaxis care in New Orleans, Louisiana: qualitative user-centered design study. JMIR Form Res 2023 Feb 27;7:e43019. [doi: 10.2196/43019] [Medline: 36848209]
22. Mugabirwe B, Flickinger T, Cox L, Ariho P, Dillingham R, Okello S. Acceptability and feasibility of a mobile health application for blood pressure monitoring in rural Uganda. JAMIA Open 2021 Jul;4(3):ooaa068. [doi: 10.1093/jamiaopen/ooaa068] [Medline: 34514350]
23. Hodges J, Waldman AL, Koshkina O, et al. Process evaluation for the adaptation, testing and dissemination of a mobile health platform to support people with HIV and tuberculosis in Irkutsk, Siberia. BMJ Open 2022 Mar 29;12(3):e054867. [doi: 10.1136/bmjopen-2021-054867] [Medline: 35351714]
24. Flickinger TE, Sherbuk JE, Petros de Guex K, et al. Adapting an m-Health intervention for Spanish-speaking Latinx people living with HIV in the Nonurban Southern United States. Telemed Rep 2021 Feb;2(1):46-55. [doi: 10.1089/tmr.2020.0018] [Medline: 33817694]
25. Interventions: positivelinks. : Center for innovation and engagement / National Alliance of State and Territorial AIDS Directors (NASTAD); 2021 URL: https://ciehealth.org/wp-content/uploads/2022/01/PositiveLinks_12-16-21_compressed.pdf [accessed 2025-12-16]
26. Compendium of evidence-based interventions and best practices for HIV prevention: background, methods, and criteria. : Centers for Disease Control and Prevention; 2024 URL: https://stacks.cdc.gov/view/cdc/149681 [accessed 2025-12-16]
27. Thompson MA, Horberg MA, Agwu AL, et al. Primary care guidance for persons with human immunodeficiency virus: 2020 update by the HIV medicine association of the infectious diseases society of America. Clin Infect Dis 2021 Dec 6;73(11):e3572-e3605. [doi: 10.1093/cid/ciaa1391] [Medline: 33225349]
28. Ryan white HIV/AIDS program (RWHAP) best practices compilation. Health Resources and Services Administration (HRSA). 2022. URL: https://targethiv.org/library/cie-positivelinks [accessed 2025-12-16]
29. Hodges J, Caldwell S, Cohn W, et al. Evaluation of the implementation and effectiveness of a mobile health intervention to improve outcomes for people with HIV in the Washington, DC Cohort: study protocol for a cluster randomized controlled trial. JMIR Res Protoc 2022 Apr 22;11(4):e37748. [doi: 10.2196/37748] [Medline: 35349466]
30. Greenberg AE, Hays H, Castel AD, et al. Development of a large urban longitudinal HIV clinical cohort using a web-based platform to merge electronically and manually abstracted data from disparate medical record systems: technical challenges

XSL•FO
RenderX

and innovative solutions. J Am Med Inform Assoc 2016 May;23(3):635-643. [doi: 10.1093/jamia/ocv176] [Medline: 26721732]
31. Shea CM, Jacobs SR, Esserman DA, Bruce K, Weiner BJ. Organizational readiness for implementing change: a psychometric assessment of a new measure. Implement Sci 2014 Jan 10;9(1):7. [doi: 10.1186/1748-5908-9-7] [Medline: 24410955]
32. Matson TE, Hermes EDA, Lyon AR, et al. A framework for designing hybrid effectiveness-implementation trials for digital health interventions. Ann Epidemiol 2025 Apr;104:35-47. [doi: 10.1016/j.annepidem.2025.02.007] [Medline: 40015542]
33. Caldwell S, Flickinger T, Hodges J, et al. An mHealth platform for people with HIV receiving care in Washington, district of Columbia: qualitative analysis of stakeholder feedback. JMIR Form Res 2023 Sep 19;7:e48739. [doi: 10.2196/48739] [Medline: 37725419]
34. Proctor EK, Powell BJ, McMillen JC. Implementation strategies: recommendations for specifying and reporting. Implement Sci 2013 Dec 1;8(1):139. [doi: 10.1186/1748-5908-8-139] [Medline: 24289295]
35. Fernandez ME, Ten Hoor GA, van Lieshout S, et al. Implementation mapping: using intervention mapping to develop implementation strategies. Front Public Health 2019;7:158. [doi: 10.3389/fpubh.2019.00158] [Medline: 31275915]
36. Powell BJ, McMillen JC, Proctor EK, et al. A compilation of strategies for implementing clinical innovations in health and mental health. Med Care Res Rev 2012 Apr;69(2):123-157. [doi: 10.1177/1077558711430690] [Medline: 22203646]
37. Nilsen P. Making sense of implementation theories, models and frameworks. Implement Sci 2015 Apr 21;10(1):53. [doi: 10.1186/s13012-015-0242-0] [Medline: 25895742]
38. Weiner BJ, Lewis CC, Stanick C, et al. Psychometric assessment of three newly developed implementation outcome measures. Implement Sci 2017 Aug 29;12(1):108. [doi: 10.1186/s13012-017-0635-3] [Medline: 28851459]
39. Weiner BJ. A theory of organizational readiness for change. Implement Sci 2009 Oct 19;4:67. [doi: 10.1186/1748-5908-4-67] [Medline: 19840381]
40. Geng EH, Peiris D, Kruk ME. Implementation science: relevance in the real world without sacrificing rigor. PLoS Med 2017 Apr;14(4):e1002288. [doi: 10.1371/journal.pmed.1002288] [Medline: 28441435]
41. Proctor E, Silmere H, Raghavan R, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar;38(2):65-76. [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
42. Murray E, Hekler EB, Andersson G, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016 Nov;51(5):843-851. [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]

Abbreviations

CFIR: Consolidated Framework for Implementation Research
DHI: digital health intervention
EMR: electronic medical record
IRB: institutional review board
LMS: learning management system
mHealth: mobile health
ORIC: Organizational Readiness to Implement Change
RE-AIM: Reach Effectiveness Adoption Implementation Maintenance

Edited by A Mavragani, T Leung; submitted 21.Apr.2025; peer-reviewed by K Mouloudj, M Al-Mujtaba; revised version received 20.Oct.2025; accepted 30.Nov.2025; published 21.Jan.2026.

Please cite as:
Hodges J, Cohn W, Castel A, Flickinger T, Waldman AL, Hilgart M, Kirby O, Caldwell S, Ingersoll K
A Complex Digital Health Intervention to Support People With HIV: Organizational Readiness Survey Study and Preimplementation Planning for a Hybrid Effectiveness-Implementation Study
J Med Internet Res 2026;28:e76327
URL: https://www.jmir.org/2026/1/e76327
doi:10.2196/76327

© Jacqueline Hodges, Wendy Cohn, Amanda Castel, Tabor Flickinger, Ava Lena Waldman, Michelle Hilgart, Olivia Kirby, Sylvia Caldwell, Karen Ingersoll. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Health Communication Campaign Performance During the HEALing Communities Study: Cross-Sectional Examination of Digital Advertising Methods

Nicky Lewis[1], PhD; Jennifer Reynolds[2], MPH; Diane Krause[2], MS, MPH; Philip M Reeves[2], PhD; Jamie Luster[3], MPH; Michael D Stein[4], MD; Sharon L Walsh[5], PhD; Amy Farmer[3], MA; Michelle R Lofwall[5], MD; Monica F Roberts[5], PharmD; Hilary L Surratt[5], PhD; Brooke N Crockett[3], MPH; Kara Stephens[2], MPH, MSA; Kelli Bursey[2], MPH; Kristin Mattson[2], MPH, MCHES; Michael D Slater[3], PhD

[1]Indiana University Bloomington, 601 E Kirkwood Avenue, Bloomington, IN, United States

[2]Health and Resilience Strategies, Oak Ridge Associated Universities, Oak Ridge, TN, United States

[3]The Ohio State University College of Medicine, Columbus, OH, United States

[4]Boston University School of Public Health, Boston, MA, United States

[5]University of Kentucky, Lexington, KY, United States

Corresponding Author:
Nicky Lewis, PhD
Indiana University Bloomington, 601 E Kirkwood Avenue, Bloomington, IN, United States

Abstract

Background: Research on the effectiveness of digital health campaign strategies is lacking. Understanding performance outcomes is essential for the successful implementation of campaigns. Two studies examined platforms, tactics, and content of digital health campaigns using paid media performance data.

Objective: This analysis compared 2 digital advertising methods (social media and banner or display) using click-through rate (CTR) and cost-per-click (CPC) as performance measures. Performance differences by state, community type, message approach, format, and image type were assessed. CTR and CPC served as measures in determining performance differences between social media and banner or display.

Methods: This cross-sectional secondary analysis examined campaign performance for the HEALing (Helping to End Addiction Long-Term) Communities Study, which served 85,875,105 impressions. Data were collected from media buy reports, entered into templates that included method (display or banner and social media) and key performance indicators (impressions, clicks, and media spend), and CTR and CPC were calculated. Study 1 assessed differences in CTR and CPC for social media and banner or display by state (KY, NY, MA, and OH) and community type (urban and rural). Study 2 assessed differences in CTR for social media and banner or display by state (KY, NY, MA, and OH), community type (urban and rural), message approach (testimonial and information-based), format (motion graphic or graphics interchange format, video, and static image), and image type (local and stock). Separate analyses were conducted for each advertising method.

Results: Study 1 found significant differences between advertising methods, where social media had higher CTR compared to banner or display. Social media had a significant main effect for state, where OH had the highest CTR. There was a statistically significant difference in CPC based on advertising method, where social media had a lower CPC compared to banner or display. Social media had a significant main effect for state, where OH had the lowest CPC. Banner or display had a significant main effect for state and community type, where OH and urban communities had the highest CPC. Study 2 found significant differences between advertising methods, where social media had higher CTR than banner or display. For social media, urban communities, static format, and local spokespersons had the highest CTR. There were significant differences between all pairs of states, where OH had the highest CTR. For display or banner, static format and local spokespersons had the highest CTR.

Conclusions: This analysis provides guidance for digital health campaigns. It examined the performance of opioid use disorder campaigns using CTR and CPC measures, demonstrating utility in future campaign evaluations. Social media was more related to stimulating responses to campaign messages compared to banner or display. State-to-state variations emphasized the importance of message pilot testing. Using local spokespersons versus stock spokespersons is recommended.

Trial Registration: ClinicalTrials.gov NCT04111939; https://clinicaltrials.gov/study/NCT04111939

(J Med Internet Res 2026;28:e75458) doi:10.2196/75458

XSL•FO
RenderX

KEYWORDS

communications; opiate overdose; social stigma; naloxone; methadone; buprenorphine evidence-based practice

Introduction

Overview

Studies that examine the platforms (ie, banner or display and social media), tactics (ie, campaign strategy), and content (ie, campaign materials) that influence health communication campaign effectiveness are limited [1,2]. There are few published empirical reports on the comparative effectiveness of online advertising platforms in regard to online public health messaging content [3]. The current research effort examines the online advertising tactics, platforms, and messaging across the communication campaigns of the HEALing (Helping to End Addiction Long-Term) Communities Study (HCS) using paid media performance data.

Background and Rationale

Examinations of the effectiveness of health communication campaigns are minimal and essentially absent in the domain of opioid use disorder (OUD); see Rath et al [4] for an exception. Most research on the effectiveness of health communication campaigns is centered on tobacco cessation efforts, including the Tips from Former Smokers campaign [5]. Durkin et al [6] highlight the potential of digital and social media in enhancing campaign effectiveness. Digital and social media may be particularly important when producing opioid-related health campaigns because of the stigmatized nature of opioid use. This is because digital and social media can often be consumed in private or semiprivate locations and on one's individual media device (eg, tablet and smartphone), reducing the likelihood of others observing the content.

Most opioid-related campaigns focus on prevention or stigma reduction; some target the overdose reversal medication naloxone, but few focus on treatment with medications for opioid use disorder (MOUD) [7]. These campaigns frequently communicate to the general public rather than targeting messaging to priority audiences who could most benefit, such as individuals and loved ones affected by OUD and sectors that may influence local programs and policies regarding substance use, including criminal or legal services, emergency medical services, local government officials, faith-based leaders, and media [8]. Geographical, regional, and cultural differences can also contribute to varying outcomes associated with campaigns as well as the way campaigns are managed and the materials used [9]. Moreover, many opioid-related health campaigns fail to conduct formative research to inform audience segmentation and campaign messaging tactics and content or to use evidence-based practices for campaign development, implementation, and evaluation [7].

Although data regarding widespread opioid-related campaign effectiveness are limited, some evidence can be gleaned from smaller-scale message testing studies. Visual campaigns about stigmatizing language and MOUD treatment, combined with narrative vignettes from either people with lived experience experiencing OUD or of health care administrators, resulted in a significant reduction in stigmatizing attitudes among a large survey of health care providers [10]. Focus groups of young adults who use illicit drugs in Illinois provided evidence that effective campaigns should highlight the risk of fentanyl exposure, fill in gaps about fentanyl knowledge, form a personal connection in messaging, and use realistic details and images [11]. Among a sample of college students without personal OUD experience, a blog-style testimonial about an individual in recovery who used MOUD followed by abstinence-based treatment resulted in strong perceptions of the character's fluidity and predicted stigma reduction [12]. These findings suggest the importance of personalized messaging with a testimonial-driven element, consistent with the literature on engagement with digital advertisements in general. Both personal relevance, defined as the process through which viewers determine if an advertisement or message relates to their personal goals or values, and social relevance, a situational element involving likes, comments, or other engagement with advertisements or messages by one's social network, can lead to increased engagement [13].

HCS Campaigns

The HCS was a multisite, parallel-group, community-level cluster randomized wait-listed trial supported by the National Institutes of Health (NIH) and the Substance Abuse and Mental Health Services Administration (SAMHSA). The HCS sought to test the effectiveness of an intervention for reducing opioid overdose deaths in 67 rural (n=29) and urban (n=38) communities heavily impacted by the opioid crisis across 4 states (KY, MA, NY, and OH) [8]. The HCS tested the Communities That HEAL (CTH) intervention, which deployed a community-engaged, data-driven process and tailored communication campaigns to assist communities in adopting and implementing evidence-based practices (EBPs) to reduce opioid fatalities [8]. Communities were randomized to either receive the CTH first (wave 1; n=34) or serve as the wait-list control group (wave 2; n=33) to receive the CTH after the HCS comparison period was completed [8]. Randomization only occurred at the community level in the main study, where communities were selected as part of the intervention group or wait-list control group. There was no randomization of the observational data reported in this paper. Complicated by the COVID-19 pandemic, an extreme rise of fentanyl in the illicit drug supply and the limited measurement period, the intervention did not significantly reduce the opioid overdose death rate in intervention communities [14]. However, an estimated total of 483 deaths were averted in the intervention arm (wave 1) compared to the control arm (wave 2).

The HCS communication campaigns aimed to support the adoption of EBPs (1) by improving awareness of and access to naloxone (an overdose reversal medication) and MOUD, (2) by driving demand for these EBPs, and (3) by reducing stigma, a key barrier to treatment initiation and recovery for individuals with substance use disorders [8]. A workgroup of communication experts from the research sites (Boston Medical Center, Columbia University, University of Kentucky, and The

XSL•FO
RenderX

Ohio State University), the HCS Data Coordinating Center (Research Triangle Institute International), the National Institute on Drug Abuse, SAMHSA, and Oak Ridge Associated Universities (ORAU), an outside consultant on the study, conducted formative research and developed a standardized strategy (ie, topics, objectives, and priority audiences) and approach for the campaigns (ie, activities that shared messages with priority audiences). Priority audiences across all 4 sites included local community leaders and influencers, people at risk for an overdose or with OUD, and family and friends of those at risk. There were options at each site for crafting community-tailored messages (ie, including local photos), but all messages contained similar calls to action (eg, Carry naloxone. Save a life).

Objectives

The development, dissemination, and monitoring of HCS campaigns created an opportunity for a more robust analysis of the performance of health communication campaigns, particularly those focused on OUD. Campaign development also involved in-depth formative research conducted by communications experts within multiple sectors, which is considered a best practice but is sometimes limited due to time, financial, and other resource constraints [15]. Thus, the current research effort adds to the existing literature on the effectiveness of health communication campaigns, expanding beyond smaller-scale OUD campaign message testing studies [11] and accounting for the exposure and engagement outcomes associated with the implementation of multigeographical site and multicommunication platform OUD campaigns [10]. Considering the above factors, we planned our assessment around 2 key campaign flight components: the exposure and engagement outcomes associated with specific tactics used in standard or template-based message dissemination (study 1) and the exposure and engagement outcomes associated with specific features of customizable, community-tailored campaign messages (study 2).

Considering this, the following research questions (RQs) were posed for the HCS tactics dataset (study 1):

- RQ1: How did click-through rate (CTR) differ based on the type of advertising method (social media and banner or display)?
- RQ2: How did CTR for social media advertisements differ across states (KY, MA, NY, and OH) and community types (urban and rural)?
- RQ3: How did CTR for banner or display advertisements differ across states (KY, MA, NY, and OH) and community types (urban and rural)?
- RQ4: How did cost-per-click (CPC) differ based on the type of advertising method (social media and banner or display)?
- RQ5: How did CPC for social media advertisements differ across states (KY, NY, and OH) and community types (urban and rural)?
- RQ6: How did CPC for banner or display advertisements differ across states (KY, NY, and OH) and community types (urban and rural)?

The following RQs were posed for the HCS community-tailored messages dataset (study 2):

- RQ1: How did CTR differ based on the type of advertising method (social media and banner or display)?
- RQ2: How did CTR for social media advertisements differ based on (1) approach (testimonial and information-based), (2) format (motion graphic or graphics interchange format [GIF], video, and static), (3) image (local and stock), (4) state (KY, MA, NY, and OH), and (5) community type (urban and rural)?
- RQ3: How did CTR for banner or display advertisements differ based on (1) approach (testimonial and information-based), (2) format (motion graphic or GIF, video, and static), (3) image (local and stock), (4) state (KY, MA, and OH), and (5) community type (urban and rural)?

Methods

Overview

The study start date was October 23, 2019, and the estimated study completion date was September 30, 2025.

The STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) cross-sectional checklist was used in the writing of this manuscript [16]. This cross-sectional secondary analysis examines the relationships between different paid digital advertising strategies from a standard or template-based set of advertisements (study 1) and message content factors from a customizable, community-tailored set of advertisements (study 2) on engagement key performance indicators (KPIs) from communication campaigns developed for the HCS. KPIs are reported from each community and site as a whole, not as part of the original randomization of the communities into their respective intervention group or wait-list control group as part of the main study.

The outcomes presented here build upon a conceptual framework of KPIs that uses CTR and CPC as performance outcomes [1]. We compared 2 digital advertising methods—social media and banner or display—measuring cost and engagement KPIs. In study 1: HCS tactics, we examined the differences in CTR and CPC for each advertising method—social media or banner or display—across the states of KY, NY, and OH (Note: cost data from MA were not used because of inconsistencies, where multiple communities were grouped together for their advertising spend), and community type—urban and rural—separately. In Study 2: HCS community-tailored messages, no cost data were available, so only differences in CTR were examined within each advertising method (social media and banner or display). The dataset included detailed information about the characteristics of messages, so each analysis included one of the following independent variables: approach (testimonial and information-based), format (motion graphic or GIF, video, and static), image type (local and stock), state (KY, MA, OH, NY; Note: NY was excluded from banner or display analyses due to small sample size), and community type (urban and rural).

XSL•FO
RenderX

Setting

The research site and ORAU staff collaborated closely with community coalitions and "champions" in each HCS community across KY, MA, NY, and OH to assess local needs and communication preferences, review local OUD data, and tailor campaign audience segments, materials, and dissemination strategies. Champions were coalition members who facilitated communication and activities between the coalitions, the HCS research teams, and partner organizations. HCS staff also provided technical assistance and training throughout the study to support coalitions in developing, implementing, and evaluating their campaigns. In total, the HCS designed, disseminated, and evaluated 264 local, community-tailored, and data-informed health communication messages about 5 different OUD topics from April 2020 to December 2023 (Table 1).

Table . Characteristics of HEALing[a] Communities Study health communication campaign messages disseminated in wave 1 and wave 2 communities across KY, MA, NY, and OH.

Wave	Communities, n	Total campaigns, n	Topics	Objectives	Date range
1	33	165	• Naloxone • Stigma • MOUD[b] treatment • Retention in MOUD treatment • Community choice[c]	• Increase demand for naloxone • Increase access to and availability of naloxone • Increase demand for, access to, and availability of MOUD • Increase MOUD prescribing • Improve acceptance and support for those in MOUD treatment • Increase demand for, access to, and availability of MOUD • Increase MOUD prescribing • Increase retention in MOUD treatment for OUD[d] • Improve support for those in MOUD treatment	April 2020 to February 2022
2	33	99	• Naloxone and fentanyl • MOUD treatment and stigma • Community choice[c]	• Increase demand for naloxone • Increase access to and availability of naloxone • Increase knowledge about the prevalence and risk of fentanyl • Increase demand for, access to, and availability of MOUD • Increase MOUD prescribing • Improve acceptance and support for those in MOUD treatment	November 2022 to December 2023

[a]HEAL: Helping to End Addiction Long-Term.

[b]MOUD: medication for opioid use disorder.

[c]"Community choice" indicates that communities had the option to repeat one of the prior campaigns based on their specific needs and preferences.

[d]OUD: opioid use disorder.

XSL•FO
RenderX

Study 1: HCS Tactics Design

Coalitions used a study-provided template to identify paid and unpaid dissemination tactics they wished to implement for each HCS campaign. These plans included specifics about budget, types of advertising, and targeting parameters. Given the study campaigns aimed to drive traffic to community web pages containing OUD education and treatment information, all campaigns were optimized to generate "web page clicks" (engagement) over impressions (exposure). The study required all advertising to be, as much as possible, limited to the geographic boundaries or zip codes of each community. This requirement, and the relatively small budgets allocated for paid media, prevented communities from using "mass media" advertising, such as cable television. Research sites provided funding to communities for campaign dissemination. Most communities chose to purchase a mix of traditional (eg, print, local radio, local television, and billboards) and digital (eg, social media, digital display, and streaming television) advertising, although some communities did not purchase advertising for every campaign.

Most communities elected to target advertisements to similar audiences, including people with lived experience and community leaders. Communities used common digital media platforms and diverse audience targeting tactics to reach these audiences with advertisements; however, most communities elected to target campaign advertising using the demographics (eg, age and gender) associated with higher overdose deaths in each county and selected the media channels most used by those demographic groups. The second-most common targeting was behavioral, involving sharing advertisements with those who searched for or showed interest in drugs, alcohol, addiction, and treatment information.

To facilitate the purchasing and placement of paid advertisements, ORAU, on behalf of each research site, subcontracted with 4 media buyers, each with experience and expertise in 1 of the 4 state markets (KY, MA, NY, and OH). As part of their contracts, media buyers were required to provide communities and study staff with detailed paid advertising performance reports after each campaign, which formed the basis for the study data. The duration of communities' paid advertising flights ranged from 5 to 133 days, with an average length of 59 days. The paid media costs per community per campaign ranged from US $75 to US $44,002, with an average cost of US $4630. Across all campaigns, the total impression volume was 85.9 million, and the total number of clicks to the HCS web pages was 369,679. Researchers could not estimate the number of specific audience members in each community (eg, number of people who use drugs in a specific community and number of community leaders).

Study 2: HCS Community-Tailored Messages Design

HCS provided communities with a set of customizable social media and print materials (assets) and a website linking community members to additional information and local resources related to each campaign theme. The campaign assets were designed with "core" (unchangeable) messages and calls to action. Campaign assets could be locally tailored with options including imagery (local faces and places), content (medications, language, and personal testimonials or quotes), product (type of material, format, size, and design specifications), and destination (local provider, website, and phone number).

Data Sources or Measurement

The method for collecting and analyzing paid media data across research sites involved a collaborative and iterative approach. Health communication experts from each site worked closely with media buyers to ensure strategic media placements and accurate data collection. Upon conclusion of the study, investigators sought to compare digital media performance across states and communities. To facilitate this process, a data manager developed a Microsoft Excel template that stored media buy method information (ie, display or banner and social media) and KPIs reported across all campaigns. KPIs included impressions, clicks, and media spend, from which CTR and CPC could be calculated (if not already provided in media buy reports). The template was refined through iterative feedback from staff at each site, ensuring that it was comprehensive and adaptable to the varied types of data collected across sites. Some KPIs, such as video completions, video completion rate, and reach, were not included in the final template, given that they were not consistently used or reported across states. This template, complemented by a codebook for data standardization, was shared across all sites. Study staff from each state entered media buy data from their community reports into this standard template, such that data could easily be combined and compared. The data manager reviewed data for errors, cleaned to address inconsistencies, and confirmed with site staff for accuracy.

Variables

All KPIs used in these studies are reported based on the dates the campaigns were implemented in the HCS communities. As indicated earlier, the data were compiled, and the following variables were examined during the analysis.

In the HCS tactics dataset (study 1), CPC and CTR were used as dependent variables. CPC was calculated by dividing the total spend on the paid media tactic or platform within specific dates by the total number of clicks. CTR was calculated by dividing the total number of clicks by the total number of impressions. The independent variables were community type (urban and rural) and state (KY, NY, and OH; Note: MA was excluded from these analyses due to inconsistencies in their cost data, where multiple communities were grouped together for their advertising spend). Separate analyses were conducted for each advertising method (banner or display and social media). Banner or display advertising included cross-platform displays, device ID targeting, and native advertisements for seamless integration, along with search retargeting and in-app videos to engage users based on behavior and context. Strategies also leveraged standard display, preroll behavioral, and preroll contextual advertisements for targeted visibility. Social media advertising used Facebook for its broad reach and targeting capabilities, including boosted posts, and Instagram for visual engagement. Snapchat and TikTok targeted younger audiences with dynamic content, and YouTube provided a platform for video content.

XSL•FO
RenderX

In the HCS community-tailored messages dataset (study 2), CTR was the only dependent variable. No cost data were available, and as a result, CPC could not be calculated. Community type (urban and rural) and state (KY, MA, and OH; Note: NY was excluded from banner or display analyses due to small sample size) were used as independent variables. Additional message characteristics were also used as independent variables in this study. These characteristics included the approach for conveying the message (testimonial or information-based), the message format (motion graphic or GIF, video, or static image; Note: GIFs were excluded from social media analysis due to small sample size), and the image type (image of a local spokesperson or image of a stock person). Separate analyses were conducted for each advertising method (banner or display and social media).

Statistical Methods

When analyzing the data, only cases with data for every variable of interest were included, resulting in different sample sizes for each analysis. Two datasets, HCS tactics (study 1) and HCS community-tailored messages (study 2), were analyzed. Data were not normalized by state or community for any of the analyses in either study. For Study 1, there were no missing data for any of the independent variables. When values for the dependent variables were unavailable, the corresponding media buy was removed from the analysis. Missing data for CTR represented 3.92% (18/459) of cases that met the inclusion criteria. Missing data for CPC represented 6.10% (28/459) of cases that met the inclusion criteria. For Study 2, there were no missing data on any of the independent variables. When values for the dependent variable were unavailable, the corresponding media buy was removed from the analysis. Missing data for CTR represented 0.99% (59/5913) of cases that met the inclusion criteria.

In the HCS tactics dataset (study 1), a Mann-Whitney *U* test was conducted to test differences in median percentage CTR (or median multiplied by 100) based on advertising method (RQ1). A Mann-Whitney *U* test was selected due to violations of normality, homogeneity, and linearity assumptions of parametric tests. Because differences were found, subsequent tests looked at each advertising method individually. Due to violations of normality and homogeneity assumptions in the data structure, Scheirer-Ray-Hare tests were conducted that looked at each advertising method individually (ie, one equation for social media [RQ2] and one for banner or display [RQ3]). Each test examined differences based on 2 independent variables, community type (rural vs urban) and state (KY, NY, OH). Dunn test with Bonferroni corrections was calculated to examine differences between pairs of states. The analytical process was duplicated with CPC as the dependent variable (RQ4, 5, and 6).

In the HCS community-tailored messages dataset (study 2), 3 statistical tests were conducted. CTR was the dependent variable for all tests. First, a Mann-Whitney *U* test was performed with the advertising method (social media vs banner or display) as the independent variable (RQ1). Because differences were found based on advertising method, 2 quantile regression equations were calculated that looked at each advertising method individually (ie, one equation for social media [RQ2] and one equation for banner or display [RQ3]). Quantile regression equations were used due to the nonnormal nature of the data. This test predicts the median percentage CTR (or median multiplied by 100), and when there were multiple levels of an independent variable, we used the level with the largest sample size as the reference group. In the social media regression equation (RQ2), the independent variables were state (KY, NY, OH, and MA), community type (rural vs urban), approach (testimonial or information-based), format (video or static image), and image type (image of a local spokesperson or image of a stock person). In the banner or display regression equation (RQ3), the independent variables were state (KY, OH, and MA), community type (rural vs urban), approach (testimonial or information-based), format (GIF, video, or static image), and image type (image of a local spokesperson or image of a stock person). No interaction terms were specified in either equation.

Ethical Considerations

The HCS study protocol was approved by Advarra Inc (Pro00038088), a single institutional review board. The data analyzed here were observations of public behaviors and not primary analyses of the original study. Advarra Inc allowed for the secondary analyses conducted here to be conducted without additional consent. The data in this study were received from each state's media buyer as aggregated, paid advertisement performance reports of KPIs not linked to any single individual. No individual participants were recruited, participated in, or compensated for this analysis, and as such, no identification of individual participants was possible (ie, data are deidentified).

Results

Study 1

HCS Tactics CTR

An overview of the state-level campaign tactics, audiences, and impressions can be seen in Table 2, and an overview of state-level campaign metric descriptives is reported in Table 3.

XSL•FO
RenderX

Table . Overview of state-level campaign impressions, target audiences, display or banner tactics, and social media tactics for the digital health communication campaigns of the HEALing[a] Communities Study from April 2020 to December 2023.

State	Impressions, n	Audiences	Display or banner[b]	Social media
KY	31,122,896	Community leaders, people with lived experience[c]	Cross-platform display, device ID targeting, display, in-app video, native, preroll, search retargeting	Facebook, Instagram, Snapchat, YouTube
MA	18,555,039	Community leaders, people with lived experience[c], health care providers	Cross-platform display, display, preroll	Facebook, Instagram
NY	27,535,463	Community leaders, people with lived experience[c]	Cross-platform display, preroll contextual	Facebook, Instagram, Snapchat, YouTube
OH	8,661,707	Community leaders, people with lived experience[c], health care providers	Display, preroll behavioral	Facebook, Snapchat, TikTok, YouTube

[a]HEAL: Helping to End Addiction Long-Term.

[b]Tactics definitions: cross-platform display: advertisements displayed across multiple devices and platforms. Device ID targeting: advertisements targeted based on unique device identifiers. Display: banner or visual advertisements shown on websites or apps. In-app video: video advertisements embedded within apps. Native: advertisements that blend with the platform's content and design. Preroll: video advertisements played before the main video content. Preroll behavioral: prevideo advertisements targeted based on user behavior patterns. Preroll contextual: prevideo advertisements tailored to the context of the content being viewed. Search retargeting: advertisements targeted to users based on their online search behavior.

[c]Including people who use drugs, people with opioid use disorder, and their loved ones.

Table . State-level campaign key performance indicators for the digital health communication campaigns of the HEALing[a] Communities Study from April 2020 to December 2023.

State	Total impressions, n	Total clicks, n	Total cost (US $)	Click-through rate (%)	Cost-per-click (US $)
KY	31,122,896	90,843	223,330.30	0.291	2.46
MA	18,555,039	84,455	N/A[b]	0.455	N/A
NY	27,535,463	109,619	265,481.20	0.398	2.42
OH	8,661,707	84,762	92,598.90	0.978	1.09

[a]HEAL: Helping to End Addiction Long-Term.

[b]N/A: not applicable.

The Mann-Whitney *U* test resulted in a statistically significant difference in CTR based on the advertising method (z=9.27; P<.001; r=0.44, 95% CI 0.22%-0.44%; RQ1). Social media (mean 0.84%, SD 1.1%; median 0.54%, IQR 0.15%-1.03%) had a higher CTR when compared to banner or display (mean 0.21%, SD 0.27%; median 0.13%, IQR 0.08%-0.21%). Because the CTR was different for each advertising method, additional tests were conducted that examined each method separately based on state and community type. For social media (RQ2), the Scheirer-Ray-Hare test demonstrated that there was a statistically significant main effect for the state (H(2, 240)=27.16; P<.001). However, there was not a statistically significant main effect for community type (H(1, 240)=0.183; P=.67), and there was not a statistically significant interaction (H(2, 240)=0.163; P=.92). Dunn test with Bonferroni corrections found statistically significant pairwise differences between KY and OH (z=4.60; P<.001) and NY and OH (z=5.08; P<.001; Table 4). OH (mean 1.83%, SD 1.39%; median 1.35%) had a higher median CTR than both KY (mean 0.70%, SD 0.83%; median 0.52%) and NY (mean 0.67%, SD 0.95%; median 0.35%). For banner or display (RQ3), the Scheirer-Ray-Hare test demonstrated that there was not a statistically significant main effect for the state (H(2, 189)=5.78; P=.06) or community type (H(1, 189)=8.53; P=.003), nor was there a statistically significant interaction (H(2, 189)=4.16; P=.12). Dunn test with Bonferroni corrections found no statistically significant pairwise differences between states (Table 4).

XSL•FO
RenderX

Table . Click-through rate descriptives for social media and banner or display by state and community type for the digital health communication campaigns of the HEALing[a] Communities Study from April 2020 to December 2023.

State and community	Social media			Banner or display		
	Advertisements, n	Mean (SD) (%)	Median (95% Mdn CI[b]) (%)	Advertisements, n	Mean (SD) (%)	Median (95% Mdn CI) (%)
KY						
Rural	40	0.70 (0.77)	0.52 (0.23-0.61)	66	0.28 (0.39)	0.14 (0.12-0.17)
Urban	48	0.70 (0.88)	0.53 (0.21-0.66)	71	0.17 (0.16)	0.12 (0.11-0.15)
Total	88	0.70 (0.83)	0.52 (0.28-0.62)	137	0.22 (0.30)	0.14 (0.12-0.15)
NY						
Rural	56	0.61 (0.53)	0.61 (0.24-0.69)	18	0.26 (0.24)	0.22 (0.17-0.27)
Urban	69	0.72 (1.18)	0.29 (0.23-0.50)	23	0.16 (0.14)	0.10 (0.08-0.13)
Total	125	0.67 (0.95)	0.35 (0.26-0.61)	41	0.20 (0.19)	0.15 (0.10-0.21)
OH						
Rural	19	2.08 (1.61)	2.00 (0.56-2.48)	11	0.13 (0.06)	0.11 (0.09-0.12)
Urban	14	1.49 (0.99)	1.30 (0.85-2.26)	6	0.07 (0.04)	0.05 (0.05-0.10)
Total	33	1.83 (1.39)	1.35 (0.75-2.45)	17	0.11 (0.06)	0.10 (0.05-0.12)
Total						
Rural	115	0.88 (1.02)	0.60 (0.46-0.65)	95	0.26 (0.34)	0.14 (0.13-0.17)
Urban	131	0.79 (1.08)	0.47 (0.29-0.63)	100	0.16 (0.15)	0.12 (0.10-0.14)
Total	246	0.84 (1.05)	0.54 (0.40-0.63)	195	0.21 (0.27)	0.13 (0.12-0.14)

[a]HEAL: Helping to End Addiction Long-Term.

[b]Mdn CI: median CI.

HCS Tactics CPC

The Mann-Whitney *U* test resulted in a statistically significant difference in CPC based on the advertising method ($z=-7.47$; $P<.001$; $r=0.36$; 95% CI US $2.47-$3.89; RQ4). Social media (mean US $5.59, SD US $8.95; median US $1.61, IQR US $0.98-$4.90) had a lower CPC when compared to banner or display (mean US $7.91, SD US $7.84%; median US $6.09, IQR US $3.79-$8.97). Because the CPC was different for each advertising method, additional tests were conducted that examined each method separately based on state and community type. For social media (RQ5), the Scheirer-Ray-Hare test demonstrated that there was a significant main effect for the state ($H(2, 240)=12.05$; $P=.002$), but no effect for community type ($H(1, 240)=1.28$; $P=.26$) nor an interaction ($H(2, 240)=3.04$; $P=.22$). Dunn test with Bonferroni corrections found significant pairwise differences between KY and OH ($z=2.64$; $P=.03$) and NY and OH ($z=3.47$; $P=.002$; Table 5). OH (mean US $2.50, SD US $5.68; median US $1.08) had a lower median CPC than both KY (mean US $4.03, SD US $5.19; median US $2.31) and NY (mean US $7.51, SD US $11.10; median US $2.04). For banner or display (RQ6), the Scheirer-Ray-Hare test demonstrated that there was a significant main effect for state ($H(2, 189)=17.07$; $P<.001$) and community type ($H(1, 189)=6.99$; $P=.008$), but not for the interaction ($H(2, 189)=2.26$; $P=.32$). Dunn test with Bonferroni corrections found significant pairwise differences between KY and OH ($z=3.98$; $P<.001$) and NY and OH ($z=3.82$; $P<.001$; Table 5). OH (mean US $15.50, SD US $12.50; median US $11.30) had a higher median CPC than both KY (mean US $7.51, SD US $7.46; median US $5.82) and NY (mean US $6.00, SD US $3.93; median US $5.48). Urban communities (mean US $8.83, SD US $8.26; median US $6.94) had a higher median CPC when compared to rural communities (mean US $6.94, SD US $7.29; median US $6.94). See Table 5 for a summary of CPC descriptives.

XSL•FO
RenderX

Table . Cost-per-click descriptives for social media and banner or display by state and community type for the digital health communication campaigns of the HEALing[a] Communities Study from April 2020 to December 2023.

State and community	Social media			Banner or display		
	Advertisements, n	Mean (SD) (US $)	Median (95% Mdn CI[b]) (US $)	Advertisements, n	Mean (SD) (US $)	Median (95% Mdn CI) (US $)
KY						
Rural	40	4.24 (5.52)	2.56 (1.31 - 3.89)	61	7.02 (8.18)	5.69 (4.02 - 6.26)
Urban	48	3.86 (4.95)	2.31 (1.26 - 3.13)	66	7.97 (6.77)	6.16 (5.14 - 7.72)
Total	88	4.03 (5.19)	2.31 (1.35 - 3.44)	127	7.51 (7.46)	5.82 (5.01 - 6.56)
NY						
Rural	56	6.76 (10.80)	1.32 (1.14 - 2.37)	18	4.33 (2.01)	4.19 (3.21 - 5.17)
Urban	69	8.13 (11.30)	2.58 (1.74 - 3.88)	23	7.31 (4.57)	7.49 (5.47 - 8.91)
Total	125	7.51 (11.10)	2.04 (1.46 - 2.77)	41	6.00 (3.93)	5.48 (4.04 - 6.55)
OH						
Rural	19	2.93 (7.15)	1.15 (0.73 - 1.16)	11	10.80 (6.01)	10.20 (5.31 - 11.30)
Urban	14	1.90 (2.85)	0.82 (0.76 - 1.58)	6	24.10 (17.10)	21.10 (9.85 - 40.00)
Total	33	2.50 (5.68)	1.08 (0.81 - 1.25)	17	15.50 (12.50)	11.30 (6.94 - 13.90)
Total						
Rural	115	5.25 (8.80)	1.32 (1.20 - 1.68)	90	6.94 (7.29)	5.52 (4.41 - 6.23)
Urban	131	5.90 (9.10)	2.33 (1.46 - 2.58)	95	8.83 (8.26)	6.94 (5.82 - 7.88)
Total	246	5.59 (8.95)	1.61 (1.38 - 2.33)	185	7.91 (7.84)	6.09 (5.48 - 6.57)

[a]HEAL: Helping to End Addiction Long-Term.

[b]Mdn CI: median CI.

Study 2

For the community-tailored messages dataset, the Mann-Whitney *U* test also resulted in a significant difference in CTR based on the advertising method (z=31.9; P<.001; r=0.42; 95% CI 0.18%-0.22%; RQ1). Social media (mean 0.84%, SD 1.40%; median 0.3%, IQR 0.12%-1.92%) had a higher CTR when compared to banner or display (mean 0.20%, SD 0.56%; median 0.08%, IQR 0.00%-0.19%). Because the CTR was different for each advertising method, additional tests were conducted that examined each method separately. The quantile regression equation that predicted CTR for social media advertisements (RQ2) found statistically significant predictors within each independent variable except for the approach (testimonial vs information-based). Urban communities, static format, and having a local spokesperson all produced a higher median CTR than their respective reference group (Tables 6 and 7). Dunn test with Bonferroni corrections found statistically significant pairwise differences between all pairs of states: KY and MA (z=5.97; P<.001), KY and NY (z=3.90; P<.001), OH and KY (z=5.54; P<.001), OH and MA (z=7.85; P<.001), OH and NY (z=6.75; P<.001), and NY and MA (z=2.10; P=.04; Tables 6 and 7).

XSL•FO
RenderX

Table . Click-through rate descriptives for social media by state, image type, format, community type, and approach for the digital health communication campaigns of the HEALing[a] Communities Study from April 2020 to December 2023.

	Advertisements, n	Mean (SD) (%)	Median (95% Mdn CI[b]) (%)
State			
KY	2178	0.81 (1.26)	0.31 (0.27-0.34)
MA	140	0.42 (0.84)	0.12 (0.11-0.14)
NY	142	0.30 (0.23)	0.24 (0.21-0.29)
OH	137	2.31 (3.16)	1.26 (0.73-2.00)
Image			
Stock person	288	0.47 (0.91)	0.16 (0.12-0.24)
Local spokesperson	2309	0.88 (1.44)	0.31 (0.29-0.35)
Format			
Static	394	0.60 (1.02)	0.36 (0.24-0.45)
Video	2203	0.88 (1.45)	0.29 (0.27-0.32)
Community			
Urban	1468	0.94 (1.61)	0.33 (0.30-0.39)
Rural	1129	0.70 (1.06)	0.25 (0.22-0.30)
Approach			
Testimonial	2428	0.85 (1.43)	0.30 (0.27-0.33)
Information	169	0.60 (0.80)	0.37 (0.16-0.53)
Total	2597	0.84 (1.40)	0.30 (0.27-0.33)

[a]HEAL: Helping to End Addiction Long-Term.

[b]Mdn CI: median CI.

Table . Quantile regression coefficients for predicting click-through rate on social media using state, image type, format, community type, and approach for the digital health communication campaigns of the HEALing[a] Communities Study from April 2020 to December 2023.

	Coefficient (SE) (%)	*t* test (*df*)	*P* value	95% CI (%)
Intercept	0.34 (0.03)	10.76 (8, 2589)	N/A[b]	0.24 to 0.40
State (reference=KY)				
MA	−0.24 (0.03)	−9.03 (3, 2589)	<.001	−0.31 to −0.16
NY	−0.06 (0.03)	−2.24 (3, 2589)	.02	−0.10 to −0.04
OH	1.00 (0.35)	2.89 (3, 2589)	.003	0.50 to 1.59
Image (reference=stock person)				
Local spokesperson	0.28 (0.10)	2.73 (1, 2589)	.01	0.21 to 0.43
Community (reference=urban)				
Rural	−0.08 (0.03)	−2.76 (1, 2589)	.01	−0.11 to −0.04
Format (reference=static)				
Video	−0.29 (0.10)	−2.88 (1, 2589)	.004	−0.39 to −0.25
Approach (reference=testimonial)				
Information	0.04 (0.04)	1.03 (1, 2589)	.31	−0.01 to 0.09

[a]HEAL: Helping to End Addiction Long-Term.

[b]N/A: not applicable.

The quantile regression equation predicting the median CTR for display or banner advertisements found (RQ3) statistically significant predictors for image (stock person vs local spokesperson) and format (static vs motion graphic or video). Static advertisements and advertisements with an image of a local spokesperson had a higher median CTR (Tables 8 and 9).

XSL•FO
RenderX

Dunn test with Bonferroni corrections did not find any statistically significant pairwise differences between any pairs of states, and the only statistically significant pairwise difference based on format was between static advertisements and video advertisements (z=2.38; P=.02).

Table . Click-through rate descriptives for display by state, image type, format, community type, and approach for the digital health communication campaigns of the HEALing[a] Communities Study from April 2020 to December 2023.

	Advertisements, n	Mean (SD) (%)	Median (95% Mdn CI[b]) (%)
State			
KY	2558	0.21 (0.60)	0.08 (0.07-0.09)
MA	639	0.17 (0.35)	0.09 (0.08-0.09)
OH	119	0.26 (0.59)	0.11 (0.09-0.12)
Image			
Stock person	2230	0.19 (0.59)	0.07 (0.07-0.08)
Local spokesperson	1086	0.23 (0.48)	0.12 (0.10-0.13)
Format			
Static	1931	0.21 (0.63)	0.09 (0.08-0.09)
Motion graphic or GIF[c]	942	0.15 (0.18)	0.08 (0.06-0.11)
Video	443	0.29 (0.71)	0.06 (0.03-0.07)
Community			
Urban	1899	0.20 (0.50)	0.08 (0.08-0.09)
Rural	1417	0.21 (0.63)	0.08 (0.07-0.09)
Approach			
Testimonial	2771	0.21 (0.60)	0.08 (0.07-0.08)
Information	545	0.17 (0.25)	0.10 (0.09-0.11)
Total	3316	0.20 (0.56)	0.08 (0.08-0.09)

[a]HEAL: Helping to End Addiction Long-Term.

[b]Mdn CI: median CI.

[c]GIF: graphics interchange format.

XSL•FO
RenderX

Table . Quantile regression coefficients for predicting click-through rate on banner or display using state, image type, format, community type, and approach for the digital health communication campaigns of the HEALing[a] Communities Study from April 2020 to December 2023.

	Coefficient (SE) (%)	*t* test (*df*)	*P* value	95% CI (%)
Intercept	0.08 (0.00)	21.66 (8, 3308)	<.001	0.07 to 0.08
State (reference=KY)				
MA	0.00 (0.01)	−0.18 (2, 3308)	.86	−0.01 to 0.02
OH	0.01 (0.02)	0.61 (2, 3308)	.54	0.02 to 0.06
Image (reference=stock person)				
Local spokesperson	0.09 (0.01)	11.27 (1, 3308)	<.001	0.07 to 0.11
Community (reference=urban)				
Rural	−0.01 (0.01)	−1.69 (1, 3308)	.09	−0.02 to 0.00
Format (reference=static)				
Motion graphic or GIF[b]	−0.04 (0.01)	−4.99 (2, 3308)	<.001	−0.05 to −0.02
Video	−0.12 (0.01)	−8.59 (2, 3308)	<.001	−0.14 to −0.09
Approach (reference=testimonial)				
Information	0.01 (0.01)	0.63 (1, 3308)	.53	−0.01 to 0.02

[a]HEAL: Helping to End Addiction Long-Term.

[b]GIF: graphics interchange format.

Discussion

Key Results

This paper assessed 2 key health communication campaign flight components: the exposure and engagement outcomes associated with specific tactics used in standard or template-based message dissemination (study 1) and the exposure and engagement outcomes associated with specific features of customizable, community-tailored campaign messages (study 2). In Study 1 (standard or template-based messages), social media had a higher CTR compared to banner or display (RQ1). Social media had a statistically significant main effect for state, but not for community type (RQ2), where OH had the highest CTR. There were no significant differences based on state or community type for display or banner (RQ3). Social media had a lower CPC compared to banner or display (RQ4). Social media also had a statistically significant main effect for state, but not community type, where OH had the lowest CPC (RQ5). Banner or display had a significant main effect for state and community type, where OH and urban communities had the highest CPC (RQ6). In Study 2 (community-tailored messages), social media also had higher CTR compared to banner or display (RQ1). For social media, advertisements in static format, advertisements featuring local spokespersons, advertisements in OH, and advertisements in urban communities had the highest CTR. There were no significant differences based on approach, that is, testimonial or information-based (RQ2). For display or banner, static advertisements and advertisements featuring local spokespersons had the highest CTR. Approach, state, and community type differences were not significant (RQ3).

Interpretation

The results suggest that social media advertisements outperformed banner or display advertisements on CTR and CPC KPIs in the context of health communication campaigns intended to encourage information-seeking via click-throughs to website landing pages. However, these rates and costs varied as a function of the state where the campaign took place, the type of image used in the advertisement, the community type, and the format of the advertisement, suggesting that these contextual factors play a key role in the performance of digital campaigns.

Although means can be highly variable, median CPC was consistently lower for social media advertisements compared to banner or display advertisements; the extent varied by state, with OH having a much lower median CPC for social media compared to KY and NY. CTR differences by state were not simply attributable to the total dollar amount spent on media buys at each site but reflected differences in actual CTR. Social media CTR was similar for KY and NY but was higher in OH, which also had the best-performing CPC. There were no significant pairwise differences in the banner or display CTR between states, although the overall main effect for the state was significant and, descriptively, KY and NY CTR were similar while OH CTR was lower.

The differences in CTR and CPC performance by state provide a cautionary tale for the health communication practitioner. One state (OH) performed very well in response to social media advertisement placements but not in banner or display advertisements. Another state (NY) performed better with banner or display advertisements. The design and content of these banner or display advertisements were similar across sites. Considering the resources often invested in digital health communication campaigns, pilot testing across regions is a

XSL•FO
RenderX

recommended best practice to direct planners toward the most effective advertising methods [17].

There are also some important caveats regarding the better overall CTR and CPC performance for social media advertisements compared to banner or display advertisements. It is possible that the placement of the social media advertisements in these campaigns was more appropriate than the placement of the banner or display advertisements or that the message content and strategy were better suited to social media advertising [18]. Furthermore, the data here demonstrate that CTR and CPC are key criterion variables when the desired outcome is to drive targeted audiences to act by visiting a website landing page. However, if the targeted outcome is different, such as increased attention to or enhanced recall of specific health information, banner or display advertising may prove to be more efficient and effective. Past research demonstrates that advertising message complexity is a key factor in attention and attitudinal responses, where fewer message features result in more recognition, comprehension, recall, and positive attitudes [19,20].

The results associated with CTR as a function of urban versus rural community types varied across advertising methods. In Study 1, CTR for banner or display advertisements was higher in rural communities than in urban communities. In Study 2, CTR for social media advertisements was higher in urban communities than in rural communities. In general, banner or display advertisements reach a broader audience across the internet, while social media advertisements are targeted to specific social media networks (eg, Facebook or Instagram). Social media use is generally higher in urban areas, as internet access is more reliable [21]. It is possible that banner or display advertisements are more appropriate for reaching rural individuals because of their more expansive reach. This is because banner or display advertisements appear on a wider range of websites (eg, news and shopping) and are not specifically related to one's browsing activity. Alternatively, social media advertisements are only displayed within the specific platform that one actively uses.

There were 2 additional message factors that emerged as significant in study 2: the type of individual featured in the advertisement (stock photo person and local spokesperson) and the format of the advertisement (static, motion graphic or GIF, and video). Advertisements featuring local spokespeople and static advertising formats produced higher CTR than their respective reference groups. Here, using photographs of real people from the HCS communities initiated more engagement with campaign messages than stock photography. The stock images may have been perceived as less authentic or more unnatural by audiences than the photographs featuring real community members [22].

Kite et al [23] caution against creating media campaigns that use inappropriate or low-quality creative messaging; using photographs of real people may be one way to increase the quality and authenticity of messages. Regarding the performance of static formatted messaging in these studies, it seems counterintuitive that a less attention-getting format would be more successful than other types. However, when the intended outcome is to communicate clear and simple messaging, static may be a preferred format. Fan et al [24] also found that static and video messaging can elicit similar emotional responses. Static content is also more affordable to place than motion graphics or GIFs and video formats, which benefits those managing campaigns with budget constraints.

One final message factor was examined in this research effort: message approach (testimonial vs information-based). In general, narrative content garners more attention than purely informational content. In the present studies and counter to prior research findings [25], testimonial approaches were not more effective than information-based approaches [24]. However, considering that the target audience for these campaigns was communities highly affected by the opioid crisis, issue fatigue may have influenced engagement (or lack thereof). Issue fatigue occurs when a topic is consistently and regularly repeated and can result in issue avoidance and the dismissal of information [26]. Capturing and understanding community attitudes at the start of health campaigns via surveys appears to be a key component of formative research before launch. Furthermore, assessing community perceptions of campaign message repetition mid-flight could lend insight into issue fatigue. Both practices are recommended for future studies.

Limitations

There are several limitations to report with this study. First, inconsistencies in the cost data from MA prevented us from including it in the analyses for study 1. Although we could account for cost variables, there were different campaign budgets and media buyers for each site, which increased variability in the data as a whole. Second, although each RQ presented here assessed different categorical types of advertisements, the overall number of RQs increases the possibility of type I error. Third, CTR and CPC are key criterion variables in evaluating campaign performance, but there are many others not assessed here (eg, time spent on page) that could also contribute to our understanding of campaign engagement. To this point, postclick behavioral outcomes (ie, landing and subsequent action on the HCS website) were not reported here because the landing website was maintained and reviewed by a third-party data coordinating center that was unable to verify the data. Algorithmic changes across Google and Meta regarding the placement of the advertisements were also unaccounted for. Finally, community coalitions and champions steered the types of paid advertising selected for each campaign. While media buyers and consultants provided guidance on effective strategies, communities varied widely in the types of advertising used and the specific audience segments and targeting parameters they chose to receive those advertisements. This high degree of variability did not allow for comparisons by audience or targeting parameters (eg, specific geographic locations and behavioral interests), which, along with the advertisements themselves, are known to influence campaign performance. Further, many communities experimented with more novel forms of digital media advertising (eg, streaming television advertising and mobile app advertising), but these efforts were not large enough to allow for meaningful analysis or comparison.

XSL•FO
RenderX

Generalizability

We hope the data presented here provide some guidance for health communication scholars and practitioners in their implementation of digital media campaigns, especially those associated with OUD. This paper is one of the first to report on the performance of OUD campaigns using CTR and CPC measures, demonstrating the utility of their inclusion in future campaign evaluations. To date, many local, state, federal, and nonprofit agencies are implementing media campaigns with little guidance for best practices.

Conclusions

In all, we can conclude that social media advertisements generally outperform banner or display advertisements in health communication campaigns, but only when considering key contextual factors, such as geographical location and type, imaging used, and format of the advertisement. Furthermore, using median CPC as a comparison metric instead of mean CPC is an effective metric for explaining campaign performance. Health communication practitioners should use pilot testing before launching digital campaigns to guide advertising methods. Desired campaign outcomes (eg, visits to a website landing page vs increased awareness and/or enhanced recall of information) should also be considered when implementing campaign strategies, tactics, and messages.

Acknowledgments

Generative artificial intelligence was not used in any portion of this manuscript writing.

Funding

This research was supported by the National Institutes of Health (NIH) and the Substance Abuse and Mental Health Services Administration (SAMHSA) through the NIH HEAL (Helping to End Addiction Long-Term) Initiative (awards UM1DA049394, UM1DA049406, UM1DA049412, UM1DA049415, and UM1DA049417). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, SAMHSA, or the NIH HEAL Initiative. The funders of this study, NIH and SAMHSA through the NIH HEAL, had no involvement in the study design, data collection, analysis, interpretation, or writing of this manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request. Please email nhlewis@iu.edu directly for requests.

Conflicts of Interest

SLW has served as a scientific advisor for the following entities, which either have or are developing medications for the treatment of opioid use disorder: AstraZeneca, Braeburn Pharmaceuticals, Cerevel Therapeutics, Indivior, Kinnoxis, and ReacX. In the last 3 years, MRL has received research consulting fees from Braeburn Pharmaceuticals, Journey Colab, and Berkshire Biomedical and an honorarium from Camurus for developing and giving a scientific talk on long-acting injectable buprenorphine.

References

1. Chan L, O'Hara B, Phongsavan P, Bauman A, Freeman B. Review of evaluation metrics used in digital and traditional tobacco control campaigns. J Med Internet Res 2020 Aug 11;22(8):e17432. [doi: 10.2196/17432] [Medline: 32348272]
2. Noar SM. A 10-year retrospective of research in health mass media campaigns: where do we go from here? J Health Commun 2006;11(1):21-42. [doi: 10.1080/10810730500461059] [Medline: 16546917]
3. Kite J, Chan L, MacKay K, et al. A model of social media effects in public health communication campaigns: systematic review. J Med Internet Res 2023 Jul 14;25(3):e46345. [doi: 10.2196/46345] [Medline: 37450325]
4. Rath JM, Perks SN, Vallone DM, et al. Educating young adults about opioid misuse: evidence from a mass media intervention. Int J Environ Res Public Health 2021 Dec 21;19(1):22. [doi: 10.3390/ijerph19010022] [Medline: 35010279]
5. Davis KC, Duke J, Shafer P, Patel D, Rodes R, Beistle D. Perceived effectiveness of antismoking ads and association with quit attempts among smokers: evidence from the tips from former smokers campaign. Health Commun 2017 Aug;32(8):931-938. [doi: 10.1080/10410236.2016.1196413] [Medline: 27435919]
6. Durkin SJ, Brennan E, Wakefield MA. Optimising tobacco control campaigns within a changing media landscape and among priority populations. Tob Control 2022 Mar;31(2):284-290. [doi: 10.1136/tobaccocontrol-2021-056558] [Medline: 35241601]
7. Frkovich J, Hedrick H, Anakaraonye AR, Bornkessel A, Lefebvre RC. Opioid-related public health communication campaigns: an environmental scan. Am J Health Promot 2022 Jul;36(6):913-919. [doi: 10.1177/08901171221082471] [Medline: 35365055]

XSL•FO
RenderX

8. Lefebvre RC, Chandler RK, Helme DW, et al. Health communication campaigns to drive demand for evidence-based practices and reduce stigma in the HEALing communities study. Drug Alcohol Depend 2020 Dec 1;217:108338. [doi: 10.1016/j.drugalcdep.2020.108338] [Medline: 33152673]
9. Flora JA, Schooler C, Pierson RM. Effective Health Promotion Among Communities of Color: The Potential of Social Marketing: Psychology Press eBooks; 2018:353-373. [doi: 10.4324/9781315805795-20]
10. Kennedy-Hendricks A, McGinty EE, Summers A, Krenn S, Fingerhood MI, Barry CL. Effect of exposure to visual campaigns and narrative vignettes on addiction stigma among health care professionals: a randomized clinical trial. JAMA Netw Open 2022 Feb 1;5(2):e2146971. [doi: 10.1001/jamanetworkopen.2021.46971] [Medline: 35119460]
11. Fernandez P, Azucar D, Zambole K. A dose of truth: a qualitative assessment of reactions to messages about fentanyl for people who use drugs. Subst Use Misuse 2023;58(4):520-527. [doi: 10.1080/10826084.2023.2177112] [Medline: 36762461]
12. Smith RA, Osoro RA, Sterner G. Understanding the effects of digital storytelling on OUD stigma reduction: exploring character and the recovery plot. Health Commun 2025 Nov;40(12):2467-2477. [doi: 10.1080/10410236.2025.2463535] [Medline: 39950341]
13. Geng S, Yang P, Gao Y, Tan Y, Yang C. The effects of ad social and personal relevance on consumer ad engagement on social media: the moderating role of platform trust. Comput Human Behav 2021 Sep;122:106834. [doi: 10.1016/j.chb.2021.106834]
14. Walters ST, Drainoni ML, Oga EA, Byard J, Chandler RK. Solving the “last mile” problem in overdose prevention: lessons from the HEALing Communities Study. Drug Alcohol Depend 2024 Nov 1;264:112453. [doi: 10.1016/j.drugalcdep.2024.112453] [Medline: 39379270]
15. Atkin CK, Freimuth V. Guidelines for formative evaluation research in campaign design. In: Anonymous Public Communication Campaigns: SAGE; 2012:52-68. [doi: 10.4135/9781544308449.n4]
16. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. The STROBE statement: guidelines for reporting observational studies. PLoS Med 2007;4(10). [doi: 10.1371/journal.pmed.0040297] [Medline: 17941715]
17. Gaysynsky A, Heley K, Chou WYS. An overview of innovative approaches to support timely and agile health communication research and practice. Int J Environ Res Public Health 2022 Nov 16;19(22):15073. [doi: 10.3390/ijerph192215073] [Medline: 36429796]
18. Namin A, Hamilton ML, Rohm AJ. Impact of message design on banner advertising involvement and effectiveness: an empirical investigation. J Mark Commun 2020 Feb 17;26(2):115-129. [doi: 10.1080/13527266.2017.1393767]
19. Pieters R, Wedel M, Batra R. The stopping power of advertising: measures and effects of visual complexity. J Mark 2010 Sep;74(5):48-60. [doi: 10.1509/jmkg.74.5.048]
20. Yeun Chun K, Hee Song J, Hollenbeck CR, Lee JH. Are contextual advertisements effective? Int J Advert 2014 Jan;33(2):351-371. [doi: 10.2501/IJA-33-2-351-371]
21. Vogels EA. Some digital divides persist between rural, urban and suburban America. Pew Research Center. 2021. URL: https://www.pewresearch.org/short-reads/2021/08/19/some-digital-divides-persist-between-rural-urban-and-suburban-america/ [accessed 2026-03-02]
22. Ward CG. Stock images, filler content and the ambiguous corporate message. M/C J 2007;10(5). [doi: 10.5204/mcj.2706]
23. Kite J, Thomas M, Bellew B, Bauman A, Grunseit A. Mass media campaigns and the “file drawer problem”: a mixed methods study of how to avoid campaign failure. PLoS ONE 2024;19(4):e0294372. [doi: 10.1371/journal.pone.0294372] [Medline: 38625844]
24. Fan S, Shen Z, Koenig BL, Ng TT, Kankanhalli MS. When and why static images are more effective than videos. IEEE Trans Affective Comput 2023;14(1):308-320. [doi: 10.1109/TAFFC.2020.3040399]
25. Adebayo AL, Davidson Mhonde R, DeNicola N, Maibach E. The effectiveness of narrative versus didactic information formats on pregnant women’s knowledge, risk perception, self-efficacy, and information seeking related to climate change health risks. Int J Environ Res Public Health 2020 Sep 23;17(19):6969. [doi: 10.3390/ijerph17196969] [Medline: 32977683]
26. Schumann C, Arlt D. When citizens get fed up. Causes and consequences of issue fatigue—results of a two-wave panel study during the coronavirus crisis. Communications 2023 Mar 6;48(1):130-153. [doi: 10.1515/commun-2021-0014]

Abbreviations

CPC: cost-per-click
CTH: Communities That HEAL
CTR: click-through rate
EBP: evidence-based practice
GIF: graphics interchange format
HCS: HEALing Communities Study
HEAL: Helping to End Addiction Long-Term
KPI: key performance indicator
MOUD: medication for opioid use disorder
NIH: National Institutes of Health

XSL•FO
RenderX

ORAU: Oak Ridge Associated Universities
OUD: opioid use disorder
RQ: research question
SAMHSA: Substance Abuse and Mental Health Services Administration
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 03.Apr.2025; peer-reviewed by J Wang, N Hu, P Shafer; revised version received 29.Jan.2026; accepted 30.Jan.2026; published 01.Apr.2026.

Please cite as:
Lewis N, Reynolds J, Krause D, Reeves PM, Luster J, Stein MD, Walsh SL, Farmer A, Lofwall MR, Roberts MF, Surratt HL, Crockett BN, Stephens K, Bursey K, Mattson K, Slater MD
Health Communication Campaign Performance During the HEALing Communities Study: Cross-Sectional Examination of Digital Advertising Methods
J Med Internet Res 2026;28:e75458
URL: https://www.jmir.org/2026/1/e75458
doi:10.2196/75458

© Nicky Lewis, Jennifer Reynolds, Diane Krause, Phillip M Reeves, Jamie Luster, Michael D Stein, Sharon L Walsh, Amy Farmer, Michelle R Lofwall, Monica F Roberts, Hilary L Surratt, Brooke N Crockett, Kara Stephens, Kelli Bursey, Kristin Mattson, Michael D Slater. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Engagement With Meditation Apps: Cross-Sectional Survey of Use and Associations

Julia Adams[1], BSc, BA, BSc(Hons); Jonathan Davies[1], PhD; Prai Wattanatakulchat[1], BA; Julieta Galante[1], MD, PhD; Felicity Miller[1], MPsych(Clin); Simon D'Alfonso[2], PhD; Nicholas T Van Dam[1], PhD

[1]Contemplative Studies Centre, Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, Australia

[2]School of Computing and Information Systems, University of Melbourne, Melbourne, Australia

Corresponding Author:
Julia Adams, BSc, BA, BSc(Hons)
Contemplative Studies Centre
Faculty of Medicine, Dentistry and Health Sciences
University of Melbourne
700 Swanson St
Melbourne, 3010
Australia
Phone: 61 2904309289
Email: adamsjuliaalice@gmail.com

Abstract

Background: Meditation apps are increasingly popular, yet there is limited understanding of how much users actually engage with them. While meditation apps show promise for supporting mental health, engagement in real-world settings appears to be notably low. The patterns of app use and the factors that influence usage remain relatively unclear.

Objective: This study aims to examine the extent of meditation app use and the factors associated with user engagement.

Methods: We conducted a cross-sectional survey of 536 recent meditation app users across 5 English-speaking countries. Engagement data were collected via self-report and app-verified screenshots. Assessed factors included user characteristics (age, education, income, sex, country, personality, self-efficacy, readiness and expectations for change, self-compassion, and quality of life), mental health (distress, well-being, life satisfaction, anxiety, depression, support, and stress), and app-related elements (therapeutic alliance, appeal, functionality, aesthetics, information, quality, and perceived impact). The 4 outcome variables representing engagement were app-verified minutes, self-reported minutes, app-verified minutes per year (adjusted for app download date), and self-reported minutes per year (adjusted for app download date). Associations between app use and variables of interest were examined using correlations. Factors with significant associations were then included in multivariable regression models to identify those most strongly associated with engagement.

Results: Age (ρ=0.13-0.15, PP^{FDR}, where FDR is false discovery rate), expectations for sleep (ρ=0.12-0.33, P^{FDR}<.05), and expectations for thriving (ρ=0.12-0.18, P^{FDR}<.05) were associated with all outcome measures except adjusted objective minutes. Readiness to change was associated with all outcome measures (ρ=0.24-0.33, P^{FDR}<.05). Among app factors, appeal (ρ=0.18-0.23, P^{FDR}<.05) and perceived impact (ρ=0.23-0.32, P^{FDR}<.05) were associated with all outcome measures except adjusted self-report minutes, while perceived quality (r=0.28-0.51, P^{FDR}<.05) was associated with all outcome measures. Robust linear regressions showed that greater readiness to change (β=0.005-0.026, *P*=.006-.02), higher education level (β=0.029-0.540, *P*<.001), and higher openness (β=0.004-0.010, *P*=.008-.03) were associated with increased engagement. Additionally, greater expectations for sleep (β=0.004-0.009, *P*=.02-.04), greater expectation match (β=0.023, *P*=.03), and higher perceived app quality (β=0.008-0.042, *P*=.001-.01) were uniquely associated with increased engagement.

Conclusions: Most individuals who download meditation apps engage minimally. Our findings suggest that users who are more educated, open to new experiences, and hold strong beliefs in the effectiveness of meditation apps are more likely to use them regularly. Longitudinal studies are needed to examine patterns of use and strengthen causal inferences.

(J Med Internet Res 2026;28:e71960) doi:10.2196/71960

XSL•FO
RenderX

KEYWORDS

meditation; mindfulness; engagement; app; meditation apps; digital mental health intervention; behavior change

Introduction

Background

Around 1 billion people globally live with a mental health disorder [1], creating a demand that exceeds available resources. As the rise of technology has coincided with increasing strain on mental health systems, digital mental health interventions have gained popularity due to their accessibility [2,3]. Fully automated versions of these interventions may reduce reliance on limited human resources. However, engagement remains a challenge, with fewer than 20% of users continuing beyond 7 days [4].

Meditation apps are among the most common digital mental health tools [4,5]. Meditation encompasses a wide range of techniques across different traditions and religions and typically involves emotional and attentional regulation [6]. Mindfulness meditation, for example, emphasizes nonjudgmental awareness of the present moment [7]. Among regular meditators, most have used a meditation app [8]. While global rates of meditation remain unknown, 70 million people had downloaded Headspace by 2022 [9]. Thus, it is likely that a very large proportion of the population has tried meditation through an app. Given the high accessibility of meditation apps among those facing barriers to mental and physical health care, it is important to examine the practical limitations of such programs, with engagement being a key shortfall in app-based behavior change [10,11].

Most current information on meditation app use comes from clinical trials, which are not representative of real-world use. While randomized controlled trials of meditation apps show small- to medium-sized effects [12], real-world estimates suggest exceedingly high discontinuation rates [4]. As behavioral interventions are only effective if used, this presents a major challenge for apps [4]. For digital offerings to be truly useful, a better understanding of the factors associated with sustained use is essential. This study examines engagement rates and identifies who is most likely to engage with meditation apps.

Mindfulness-Based Programs and Meditation Apps

While a limited number of meditation apps have shown some efficacy for mental health outcomes [12], they should be considered separately from the established evidence base for mindfulness-based programs (MBPs) [13]. MBPs are arguably the most popular form of meditation training in clinical and academic settings, likely due to their strong evidence base [14,15]. By contrast, most popular meditation apps depart from the guided, intensive structure of MBPs [16]. Only 4% of popular apps provide evidence of their benefits [17]. Even where apps show potential efficacy, recent reviews highlight engagement as a major limitation to intervention effectiveness [12,18-20].

The Digital Transition

Digital health interventions generally face engagement issues, which likely reduce their benefits [4]. In nonpharmacological interventions, adherence is linked to outcomes [21], yet it tends to be worse in digital formats than in face-to-face interventions [22]. For behavior change apps (and apps more broadly), discontinuation occurs in 40%-60% of users [11]. In naturalistic settings, 21%-88% of users engage with an app at least once, but only 0.5%-28% sustain engagement (eg, completing all assigned modules or continuing use beyond 6 weeks [23]). Engagement decreases when digital interventions lack interactive human or human-like support [24], posing challenges for fully automated meditation apps. Clarifying who engages meaningfully with meditation apps is therefore important, given the link between adherence and outcomes.

Attrition, Adherence, and Engagement

Engagement refers to the extent of intervention use, including the amount, frequency, duration, and depth of use [25]. Attrition and adherence are related terms that describe levels of (dis)engagement in research studies. Attrition refers to discontinuation or dropout from the intervention program or from research data provision during a study [11]. For MBPs, attrition is around 19% [26], whereas app-based interventions show an average attrition of 42% in studies lasting 10 days to 12 weeks [12]. Real-world estimates for meditation apps indicate disengagement rates as high as 94% within the first 2 weeks [4].

Adherence refers to the extent to which an individual follows a prescribed treatment or intervention [21]. As no clear guideline exists for the amount of practice required to achieve an effect in mindfulness or meditation [16,27], adherence can only be considered in relation to recommended practice amounts (see example in [28]). In meditation training, prescribed engagement time ranges from as little as 35 minutes [28] to 3 hours per week in the widely used Mindfulness-Based Stress Reduction program [7]. By contrast, many apps recommend as little as 5 minutes per day or provide no clear guidance regarding minimum practice length, session duration, or overall time commitment needed to establish a practice [29]. Given the limited knowledge about dose-response relationships in meditation [16,29] and the tendency for most people to discontinue practice relatively early [4], engagement serves as a useful proxy for understanding who practices, what type of practice they follow, and why.

Why Do People (Dis)Engage With Meditation and Apps?

Overview

Understanding engagement in meditation apps requires consideration of various behavior change and persuasive systems design frameworks. Behavior change frameworks—including habit formation theory, social cognitive theory, the theory of planned behavior, and the transtheoretical model of change—suggest that user characteristics such as expectations, motivation, readiness to change, consistency of use, and self-efficacy influence engagement with behavioral interventions [30-33]. The persuasive systems design framework highlights

XSL•FO
RenderX

app design features that shape engagement and therapeutic alliance [34].

Habit formation theory emphasizes reward and associative cues as central to establishing habits, with positive outcomes reinforcing continued engagement [33]. Context can shape how rewards are perceived, influencing whether a habit is formed [35]. The theory of planned behavior further posits that perspective and context guide behavior [30,36]. Broad factors such as sociodemographics, mental health, and personality also influence engagement [30]. Expectations are shaped by attitudes and norms, with positive expectations and attitudes predicting greater meditation app engagement [36]. The transtheoretical model outlines stages of change, with later stages—more closely aligned with commitment to action—linked to more sustained behavior change [32]. Readiness to change reflects an individual's stage of change and is associated with successful maintenance of behavior change [37]. The Sussex Meditation Model identifies preintention, preparation, action, and maintenance stages as relevant to establishing a meditation practice [38,39]. Persuasive systems design, which examines how digital interventions can be structured to influence user behavior, highlights app features that enhance engagement, such as reminders and personalization [34]. Drawing on these frameworks, factors relevant to engagement in behavior change, meditation, or app use were categorized into user-related factors (sociodemographics, personal/user characteristics, and mental health factors) and app-related factors.

Sociodemographic Factors

Sociodemographic factors associated with disengagement from meditation include lower levels of education [40]; however, men, people with less education, and those with poorer health are less likely to begin meditating [41]. Meditators are also more likely to be wealthier than nonmeditators [41]. In online and app-based meditation, older age, positive expectations, and intrinsic motivation are associated with greater engagement [42,43].

Personal/User Factors

Personality factors have also been shown to influence engagement with meditation apps. Conscientiousness has been associated with meditation in general [44]. Openness predicts meditation practice outside formal program training, reflecting the "in-the-wild" context of app use [45].

Behavior change factors may also influence engagement with meditation apps. Self-efficacy and readiness to change have been linked to successful habit formation [46]. A higher intention to practice is associated with greater engagement [46]. Intrinsic motivation moderates behavior change success across demographic groups [47] and is crucial for making initial behavior change choices. Self-compassion and self-efficacy have also been found to be positively associated with engagement in behavior change [47,48].

Expectations for program efficacy can also influence behavior change. Positive experiences that meet expectations can facilitate ongoing engagement. Conversely, engagement may decline when a program or behavior does not deliver the anticipated positive outcomes [49]. Experiences of progress enhance engagement in both behavior change apps and meditation apps [50,51]. Positive expectations also predict higher engagement with digital meditation resources [42].

Mental Health Factors

Health characteristics are also important for engagement. People may be motivated by physical or mental health issues, but these same issues can also act as barriers [43]. This paradox can be explained by the desire to address a problem that simultaneously hinders the ability to engage in practice. Additionally, limited perceived gains may lead to early discontinuation. Barriers to mental or physical health care, which can impact quality of life, may further motivate meditation app use to address unmet health needs [11,41]. Although meditation use among individuals with mental health problems is common, depression is associated with low adherence to behavior modification recommendations in clinical populations [52,53]. The very symptoms people seek to address—such as amotivation, distressing thoughts, and irritability—can also complicate their efforts. Meditation apps may be moderately effective for depression, anxiety, and stress [12,19,50], potentially fostering an experience of progress. However, a minimal level of engagement is necessary to achieve efficacy [4,27]. Consequently, failure to achieve expected outcomes may lead to decreased engagement.

App Factors

The user's relationship with the app is also relevant. Therapeutic alliance—the collaborative relationship between the user and the app—predicts engagement with mental health apps [53]. Ease of use, the ability to personalize settings, reminders, progress tracking, and positive perceptions of the app also predict higher engagement with mental health apps, though these factors have not been extensively examined in meditation apps [42,53,54]. Usability (ie, the app's functionality) was identified as a key factor related to engagement in a systematic review of mental health apps [55].

This Study

Previous literature highlights factors that may be associated with meditation app use. In a cross-sectional survey capturing demographics, retrospective reports of app use, mental health factors, and perspectives on apps, we aimed to examine engagement rates and identify factors significantly associated with engagement.

This study focused on several preregistered questions:

1. To what extent are *user-related factors*—including sociodemographic characteristics, spirituality, personality, self-efficacy, self-regulation, motivation, expectations, self-compassion, mental health care status, and psychological distress—associated with mindfulness *app engagement*?
2. To what extent are *user-app relationship factors*—including therapeutic alliance, agreement on tasks and goals, and perceived app empathy and expertise—associated with mindfulness app *engagement*?
3. To what extent are *app-related factors*—including appeal, functionality, aesthetics, information quality, quantity, and

XSL•FO
RenderX

credibility, customization, accessibility, and usability—associated with mindfulness app *engagement*?

Additional questions included in the preregistration are not addressed in this paper.

Methods

Deviations From Preregistration

For clarification, we have changed the term "mindfulness apps" to "meditation apps" to capture a broader range of relevant practices. Mindfulness can, but does not necessarily, entail meditation and is variably represented as a capacity, skill, or technique. Meditation, by contrast, encompasses a broad array of spiritual and secular practices that use techniques such as focusing on an object, experience, image, or idea [56].

Deviations from the preregistration included the following: (1) focusing on 4 definitions of minutes as the primary outcome and omitting the second preregistered outcome variable—regular practice hours—for simplicity. Regular practice hours were not included because their calculation combined multiple variables and, therefore, could be subject to estimation error. The 4 variations of the outcome variable were included to capture the complexity of user behavior. (2) We did not report 95% odds ratios, as continuous outcomes were used. (3) The final sample size was substantially reduced to 536 from the target of 1000 due to a smaller-than-anticipated eligible pool. This reduction decreased statistical power, although it still allowed adequate power to detect small effects. The reduced pool also led to a fourth deviation. (4) Recruitment was extended to Australia, Canada, the United Kingdom, and New Zealand. An additional deviation involved (5) not analyzing motivation for use, as this information was captured in open-text responses and could not be used in this quantitative analysis.

Ethical Considerations

This study was conducted in accordance with ethical guidelines and was approved by the Office of Research Ethics and Integrity at the University of Melbourne (approval number 2025-23969-62994-8).

Participants provided informed consent to participate in the study via the Qualtrics survey (Qualtrics International Inc). Consent was obtained within the survey, which also included a downloadable copy of the plain language statement. The plain language statement is available in Section S1 in Multimedia Appendix 1. Provision of informed consent included acknowledgment of the right to withdraw at any time without providing an explanation. Participants also consented to secondary analyses. Survey questions were coded so that participants could not proceed without providing consent. All included responses were double-checked to ensure consent had been given.

Participants were compensated Aus $0.30-0.50 (US $0.20-0.33) for completing the screening survey (mean duration 1 minute 49 seconds) and Aus $6-8 (US $3.96-5.29) for completing the follow-up survey (mean duration 22 minutes 57 seconds), averaging Aus $20.59 (US $13.60) per hour. Survey compensation varied slightly based on median completion time; compensation was occasionally increased to better approximate the proposed hourly rate if the median completion time indicated the study took longer than expected.

Privacy and Confidentiality

Where possible, identifying information was removed from the dataset. Any copies of datasets containing identifying information were stored securely in accordance with relevant privacy guidelines and encrypted using Transport Layer Security (also known as HTTPS).

Study Design

Overview

This was a cross-sectional analysis of data collected from participants.

Procedure

Participants were recruited via Prolific (Prolific Academic Ltd) to complete a survey hosted on Qualtrics. The survey was accessible to potential participants in the United States, the United Kingdom, Canada, Australia, and New Zealand between August 1 and October 6, 2023. Participants were invited to complete a prescreening survey, and eligible individuals were sent the full survey within 1-2 days. Surveys were completed online using a laptop or mobile device. Participants were asked to upload a screenshot of their app use statistics, which provided information such as minutes, days, sessions, streaks, and the original date of download, depending on the app.

App Selection

We collected engagement information for popular meditation apps listed on the iOS (Apple Inc) and Android (Google LLC/Alphabet Inc) app stores (see Section S2 in Multimedia Appendix 1). Participants using apps in which meditation—including mindfulness meditation—was the primary intended function were included, based on app descriptions, marketing, and in-app features.

Participants using any app could complete the prescreening survey. Two (JA and JD) researchers assessed whether the app (1) prominently promoted itself as a mindfulness meditation tool and (2) provided techniques to practice mindfulness or another form of meditation. Meditation or mindfulness could not be a secondary component. We did not evaluate app content in relation to any specific definition. We adopted this approach because meditation apps do not offer a single type of meditation, nor do mindfulness apps (eg, Headspace) necessarily adhere to the MBP definition of meditation. Apps were excluded if they focused exclusively on fitness/exercise, employee well-being, cognitive behavioral therapy, or other mental health interventions. All included apps were fully automated (ie, without human support).

Participants

Inclusion criteria required participants to have used an eligible meditation app within the past 180 days; be fluent in English; and reside in Australia, Canada, New Zealand, the United Kingdom, or the United States. Exclusion criteria included failure to provide evidence of app use or use of an app in which meditation was not the primary focus.

XSL•FO
RenderX

A total of 6137 prescreening surveys were completed. We excluded 5307 responses: 4343 (70.77%) were unable to demonstrate access to an app, 316 (5.15%) had not used the app in the past 180 days, 319 (5.20%) had downloaded an ineligible app, and 329 (5.36%) self-reported zero use. Of the remaining surveys, 800 (13.04%) met the inclusion criteria.

Of the 800 participants invited to the survey, 677 (84.6%) completed it. Among the 675 survey responses received, 18 (2.3%) were identified as likely bots or fraudulent responses based on fraudulent screenshots or failed reCAPTCHA (reverse Completely Automated Public Turing Test to Tell Computers and Humans Apart), and 13 (1.6%) exhibited suspiciously high average session lengths (>3× the IQR, 70.35 minutes). An additional 22 participants (2.8%) timed out before completing the survey, 25 (3.1%) failed attention checks, 59 (7.4%) failed screenshot checks, 86 (10.75%) responded twice, and 21 (2.6%) declined to complete the survey. These categories were not mutually exclusive. The resulting sample consisted of 563 (70.4%) participants who consented and completed the full survey. Finally, 27 (4.8%) multivariate outliers were excluded according to the preregistration, yielding a final sample of 536 participants.

Measures

Engagement

To verify the reliability of self-reported information, we collected both subjective self-reports and objective, app-verified data (screenshots provided by participants), which included minutes, days, streaks (consecutive days of use), number of sessions, average session length, and duration of app ownership (Table 1). These metrics were collected using recent app duration, defined as the number of days between first and last app use. App-verified duration was recorded if a participant validated the download period via screenshot.

The primary engagement variable was minutes of app use. Four variations were analyzed: (1) objective unadjusted minutes, representing total app-verified minutes; (2) self-reported or "subjective" unadjusted minutes, representing total unverified self-reported minutes; (3) objective adjusted minutes, calculated as app-verified minutes adjusted for app-verified duration of use, expressed as minutes per year; and (4) self-reported or "subjective" adjusted minutes, calculated as self-reported minutes adjusted for self-reported duration of use, expressed as minutes per year. Adjusted variables accounted for the duration of access to the app (from the download date to the last use). As only a limited number of apps reported objective start dates, and app-verified duration correlated highly with self-reported duration, the adjusted variables were calculated using the time between the first and last reported use.

Table 1. Descriptive statistics for engagement outcomes among meditation app users.

Statistics	n	Mean (SD)	5th percentile	25th percentile	50th percentile (median)	75th percentile	95th percentile
Subjective							
Total minutes	483	3562.78 (8616.39)	4.10	76.00	420.00	2474.00	21735.10
Total sessions	452	108.43 (197.00)	0.58	9.40	37.33	130.30	407.25
Duration (days)	477	894.16 (854.20)	16.80	158.00	621.00	1342.00	2689.80
Minutes per session	454	16.60 (29.74)	3.00	5.81	10.51	18.09	35.41
Estimated minutes per month[a]	477	148.77 (304.56)	0.48	8.41	40.25	137.20	789.44
Estimated sessions per month[a]	437	9.35 (16.61)	0.13	0.93	3.29	11.05	34.79
Objective							
Total minutes	483	3358.69 (8607.36)	12.00	96.50	465.00	2410.00	21735.10
Total sessions	151	61.89 (194.90)	0.61	3.36	12.22	39.95	167.23
Minutes per session	151	49.21 (89.20)	1.98	7.26	25.92	55.90	154.08
Estimated minutes per month[a]	151	73.58 (223.73)	0.61	2.39	10.61	35.09	313.31
Estimated sessions per month[a]	148	9.44 (47.63)	0.02	0.10	0.66	2.32	17.29

[a]Estimated minutes per month and sessions per month were calculated by total engagement in minutes divided by duration of app use in years divided by 12.

Self-Reported Measures

Sociodemographic Information and Meditation History

Sociodemographic information included household income, education level, religion, app name, and approximate start and stop dates of use. Most data were self-reported via the survey. Prolific provided additional information, including age, sex, language, student and employment status, country of birth, and current residence. Regular practice information included minutes per session, sessions per day, and days per week. Meditation

XSL•FO
RenderX

history was assessed by asking participants to report their previous meditation experience in hours, ranging from 0-100 hours to 1000+ hours. See Table 1 for regular practice information, and Sections S3-S5 in Multimedia Appendix 1 for sociodemographic statistics, meditation app frequencies, and a detailed survey flow.

Attention Checks

Three attention checks were included in the survey to assess participant engagement. Participants failing 2 or more attention checks were excluded. The attention checks were designed to mimic the scale items within which they appeared; for example, "In general, select dissatisfied to show that you are paying attention."

The EuroQoL Health and Wellbeing Assessment—Short Form

The 9-item EuroQoL Health and Wellbeing (EQ-HWB-9) is a newly developed quality-of-life measure by Brazier and colleagues [57]. It assesses quality of life with a focus on health and well-being. The scale consists of 9 items, each rated from 1 (no difficulty, none of the time, and no physical pain) to 5 (unable, most or all of the time, and very severe physical pain). In this study, the 9 items of the EQ-HWB-9 demonstrated very good internal consistency (Cronbach α=0.873), and McDonald hierarchical omega was relatively high (ω=0.732). The EQ-HWB-9 was used with permission from the EuroQol Group.

The Kessler Psychological Distress Scale

The Kessler Psychological Distress Scale (K10) assesses psychological distress over the past 30 days [58]. This 10-item questionnaire, measuring anxiety and depressive symptoms, uses a 5-point scale ranging from 1 (none of the time) to 5 (all of the time). In this study, the scale demonstrated excellent internal consistency (Cronbach α=0.930; McDonald ω=0.799).

The Warwick-Edinburgh Mental Wellbeing Scale

The Short Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS) was used to assess positive aspects of mental health [59]. This 7-item scale uses a 5-point response format, ranging from 1 (none of the time) to 5 (all of the time). In this study, the scale demonstrated high internal consistency (Cronbach α=0.879; McDonald ω=0.790).

The Satisfaction with Life Survey (Single Item)

The Satisfaction with Life Survey Single Item is an abbreviated version of the established Satisfaction with Life Survey (SWLS) [60]. The scale demonstrates reasonable criterion validity with the full SWLS (zero-order r=0.62-0.64). This single-item measure asks participants to rate their life satisfaction on a scale from 1 (extremely dissatisfied) to 7 (extremely satisfied).

7-Item Generalized Anxiety Disorder Scale

The 7-Item Generalized Anxiety Disorder Scale is used to screen for general anxiety symptoms [61]. Each item is rated from 0 (not at all) to 3 (nearly every day). In this study, the scale demonstrated high internal consistency (Cronbach α=0.899; McDonald ω=0.863).

Depression (8-Item Patient Health Questionnaire)

The 8-item Patient Health Questionnaire is used to assess depressive symptoms [62]. Each item is rated from 0 (not at all) to 3 (nearly every day). In this study, the scale demonstrated high internal consistency (Cronbach α=0.881; McDonald ω=0.781).

Self-Compassion Scale

The Self-Compassion Scale consists of 12 items assessing participants' ability to be compassionate toward themselves [63]. Each item is rated from 1 (almost never) to 5 (almost always). In this study, the scale demonstrated high internal consistency (Cronbach α=0.881; McDonald ω=0.656).

Self-Efficacy (6-Item Generalized Self-Efficacy)

The 6-item Generalized Self-Efficacy assesses self-efficacy, or an individual's perceived ability to achieve goals [64]. Each item is rated from 1 (not at all true) to 3 (exactly true). The scale demonstrated high internal consistency (Cronbach α=0.821; McDonald ω=0.788).

Readiness to Change 1-Item

The Readiness to Change 1-item assessment is a 10-point scale that measures an individual's preparedness to enact a behavioral change [37]. The scale ranges from 0 (not prepared to change) to 10 (already changing). It has been validated to reflect actual readiness in clinical contexts [37,65,66].

6-Item Digital Working Alliance Inventory

The 6-item Digital Working Alliance Inventory (DWAI-6) [5] is a rating scale that assesses the therapeutic alliance between an individual and their health care provider, adapted for smartphone interventions (ie, referring to "the app" rather than "the therapist"). Items are rated on a 7-point scale from 1 (strongly disagree) to 7 (strongly agree), with subscales evaluating goal alliance (agreement on goals), task alliance (agreement on tasks), and bond (connection between app and user). The overall scale demonstrated good internal consistency (Cronbach α=0.850; McDonald ω=0.830). The Goal subscale showed high consistency (Cronbach α=0.766), the Bond subscale demonstrated acceptable consistency (Cronbach α=0.676), and the Task subscale showed poor consistency (Cronbach α=0.402).

Common Factors Domains (Modum Process Outcome Questionnaire)

We used a subset of items from the Common Therapeutic Relationship Factors Questionnaire (Modum Process Outcome Questionnaire) to assess the therapeutic relationship beyond the DWAI-6 [67]. The original questionnaire, which focuses on the clinician-patient relationship, was adapted to refer to the app-user relationship (eg, "I am able to be open and honest when interacting with the app"). Three items from the 12-item scale were included, each rated from 1 (strongly disagree) to 7 (strongly agree), with a "not applicable" option. The items demonstrated low internal consistency (Cronbach α=0.612; McDonald ω<0.001), likely due to being an unintended subset. Consequently, results will be reported for each item individually rather than as a total score.

XSL•FO
RenderX

The Big Five Inventory Short Form 2 (BFI-S-2)

The Big Five Inventory Short Form 2 (BFI-S-2) is a 30-item questionnaire assessing 5 personality domains: Extraversion, Agreeableness, Conscientiousness, Negative Emotionality/Neuroticism, and Open-Mindedness [68]. Each subscale consists of 6 items. Internal consistency, assessed using Cronbach α and McDonald ω, ranged from acceptable (Extraversion, α=0.766; ω=0.636/Agreeableness, α=0.744; ω=0.549/Openness, α=0.785; ω=0.664) to high (Conscientiousness, α=0.812; ω=0.738/Negative Emotionality, α=0.884; ω=0.841).

User Mobile Application Rating Scale

The user Mobile Application Rating Scale (uMARS) is a 27-item instrument for assessing the quality of mobile apps [69]. The scale includes subscales evaluating engagement (referred to as "appeal" in this study for clarity: "Is the app fun/entertaining to use?"), functionality ("How accurately and quickly do the app features and components work?"), aesthetics ("How good does the app look?"), information ("Is the app content correct, well written, and relevant to the goal/topic of the app?"), perceived/subjective quality ("Would you pay for this app?"), and perceived impact ("This app has increased my knowledge/understanding of meditating"). The scale asks participants to rate app elements on a 5-point scale ranging from 1 (poor) to 5 (excellent), with specific descriptions for each item. The overall scale demonstrated high internal consistency (Cronbach α=0.866) but only moderate reliability (McDonald ω=0.606). The additional subscale for perceived impact was also highly consistent (Cronbach α=0.863; McDonald ω=0.761). The Functionality subscale demonstrated high internal consistency and reliability (Cronbach α=0.803; McDonald ω=0.739), whereas the Engagement (Cronbach α=0.734; McDonald ω=0.628), Aesthetics (Cronbach α=0.761; McDonald ω=0.658), and Information (Cronbach α=0.762; McDonald ω=0.688) subscales showed acceptable consistency and reliability. The Subjective Quality subscale was consistent (Cronbach α=0.629; McDonald ω=0.667). The Aesthetics subscale showed variable internal consistency (Cronbach α=0.762; McDonald ω=0.065), indicating unequal item contributions.

Analysis Plan

Sample Size Determination

Given an unknown effect size due to the absence of robust data, and guided by prior estimates of effect sizes for meditation apps [12], we aimed to calculate statistical power based on the smallest effect size we could reasonably detect. The target sample size was 1000 participants, providing 90% power to detect an effect of r=0.102, corresponding to a small effect. Because of recruitment challenges, the final sample size before analyzing the main engagement variable was 536, representing nearly 54% of our target population (n=1000). Although this reduction decreased statistical power (80% power to detect r=0.122), given that the a priori effect size was unknown, the study remained adequately powered to detect relatively small effects, albeit slightly larger than initially anticipated.

Planned Statistics

As outlined in the preregistration, we explored associations between user-related factors (Q1, H1), user-app relationship factors (Q2, H2), app-related factors (Q3, H3), mental health factors (Q5, H5), and predefined engagement variables. All engagement variables were heavily skewed and nonnormal (see Figure 1 and Sections S11-S14 in Multimedia Appendix 1); therefore, Spearman rho correlations were estimated. Variables that were significantly associated with any outcome variable were subsequently used to build regression models for each of the 4 outcomes. As transformations did not normalize the variables, untransformed variables were analyzed using robust regression with the "robustbase" package in R (R Foundation), employing an MM-type regression estimator with a bisquare redescending score function [70,71]. This method applies case weighting to account for nonnormality. All analyses were conducted in RStudio (R Foundation).

XSL•FO
RenderX

Figure 1. Histogram of self-reported practice hours (x axis) shown as the percentage of the sample (y axis), with values converted from minutes for visualization purposes.

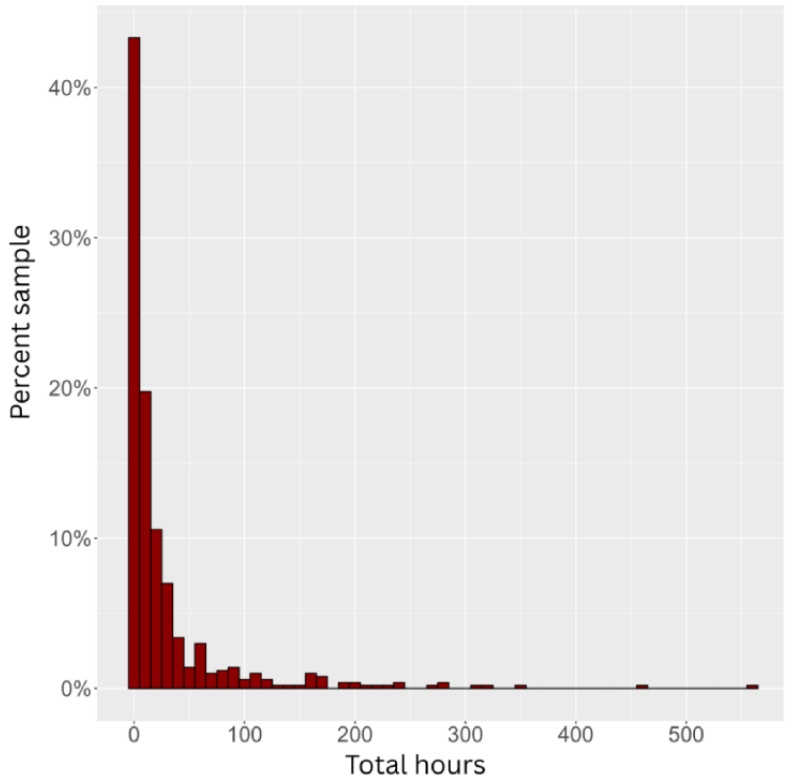

Regression Models

Robust linear regression was used to investigate which factors accounted for significant variance in engagement. For each of the 4 outcome variables—adjusted objective minutes, adjusted self-report minutes, objective minutes, and self-report minutes—a separate regression model was created using the respective measure as the outcome variable. No stepwise regression was implemented. Instead, independent variables were selected from user, mental health, and app factors that were significantly associated with at least one outcome measure in the correlation analyses, following correction for multiple comparisons.

Multiple Comparisons Correction

We explored correlations between the 4 engagement outcomes and related factors, applying a false discovery rate (FDR) correction to account for multiple comparisons [72].

Data Cleaning

Duplicate and invalid responses were removed. The data demonstrated weak correlations, positive skew, and extreme values, indicating potentially high variability or inconsistency (see Sections S6-S10 in Multimedia Appendix 1). After adjusting for app duration, we implemented several data-cleaning procedures. Intraindividual response validity calculation, LongString Identification [73], and inconsistency of responses on the BFI-S-2 [68] were each used to identify and exclude extreme cases; however, none of these approaches resulted in major changes to the results (see Sections S10-S12 in Multimedia Appendix 1). As specified in the preregistration, multivariate outliers were removed, identified as cases with a Mahalanobis distance greater than the 95th percentile on the BFI-S-2 (see Section 13 in Multimedia Appendix 1).

Results

Overall Engagement

Overall, most users completed only a few minutes across limited sessions. Despite generally low engagement, 134 out of 536 (25%) users reported more than 11 sessions per month (approximately 1 session every 3 days), while the top 5% (27/536) reported around 35 sessions per month (more than 1 session per day). These patterns align with prior findings, including a median of 90% of users dropping off within the first week of real-world use and an average 42% drop in participation in meditation app randomized controlled trials spanning 1-2 months [4,12]. Notably, 402 (75%) participants reported more than 9 sessions, which contrasts with prior findings suggesting that most users disengage completely within a week of download. However, only the top 5% (25/536, 4.7%) engaged at levels consistent with clinically meaningful change [16], while the top 25% (134/536) engaged at levels comparable to the dose of mindfulness-based interventions [27].

Participant Characteristics: Descriptives

Sociodemographic Features

Participants (N=536) ranged from 18 to 70 years of age (mean 36.56 years, SD 10.68 years) and were predominantly female (n=366, 68.3%). See Figure 2 for participant flow. Most resided in the United States 253 (47.20%) or the United Kingdom

XSL•FO
RenderX

(n=226, 42.2%), with smaller proportions from Australia (n=27, 5%), Canada (n=21, 3.9%), and New Zealand (n=6, 1.1%). The majority identified as White (n=422, 78.73%) and were highly educated, with 387 (72.20%) holding at least a bachelor's degree. Participants were also relatively wealthy, with nearly one-third reporting a combined income of $100,000 or more (n=145, 27.05%). Note that income brackets were not adjusted across countries. Nearly half of the participants reported no religious affiliation (n=264, 49.25%). The most frequently used apps were Headspace (n=191, 35.6%) and Calm (n=123, 22.9%), which together accounted for 58.6% of the sample (n=314). Full details are provided in Sections S3 and S4 in Multimedia Appendix 1.

Figure 2. Participant flow.

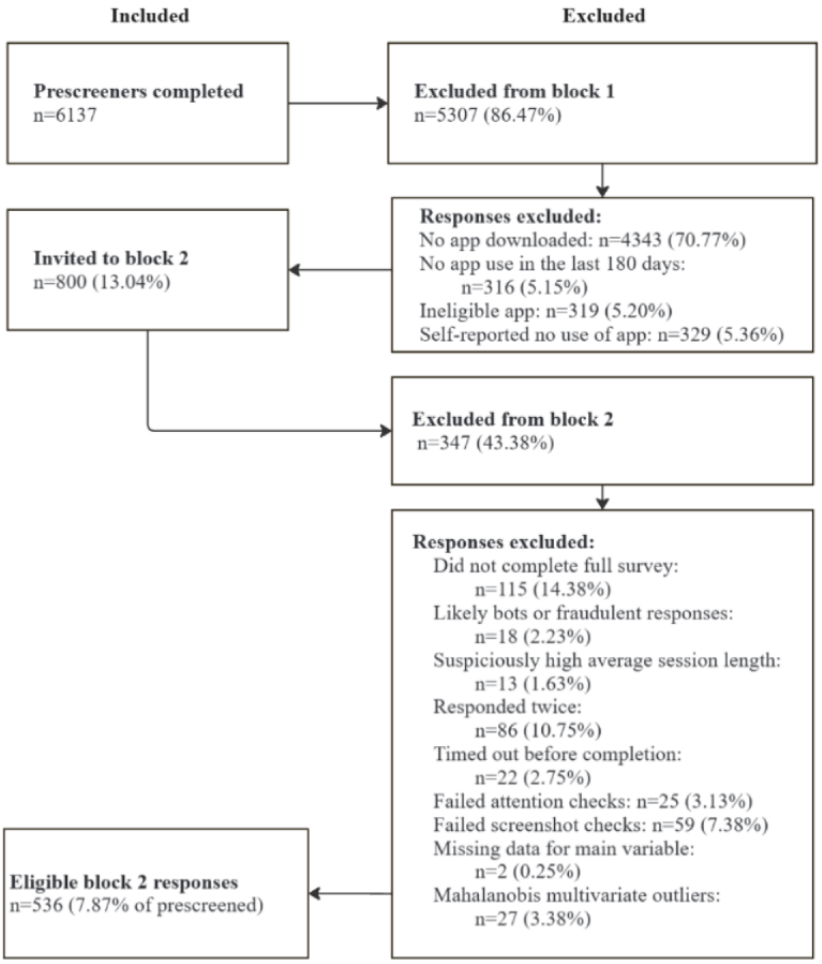

Meditation Experience

Most users (n=330, 61.6%) reported between 0 and 100 hours of overall meditation experience. Meditation experience varied across meditation apps (χ^2_4=34.18, P<.001); users with 0-100 hours were most likely to use Headspace (124/377, 32.9%), followed by Calm (75/377, 19.9%) and Insight Timer (25/377, 6.6%; see Table 2).

Table 2. Frequency and relative percentage of Calm, Headspace, and Insight Timer users by self-reported meditation experience level (n=377).

Duration (hours)	Calm, n (%)	Headspace, n (%)	Insight Timer, n (%)	Total, n (%)
0-100	75 (19.9)	124 (32.9)	25 (6.6)	224 (59.4)
101-1000	35 (9.3)	46 (12.2)	46 (12.2)	127 (33.7)
1001+	6 (1.6)	11 (2.9)	9 (2.4)	26 (6.9)
Total	116 (30.8)	181 (48.0)	80 (21.2)	377 (100)

Engagement

After excluding invalid responses and duplicates, engagement levels remained low, with a positive skew for both hours and sessions (see Figures 2 and 3). Adjusting for app duration showed low engagement regardless of how long the app had been available to users (see Figures 4 and 5). Estimated minutes per month and sessions per month were adjusted within each user to provide clearer engagement metrics. The median number

XSL•FO
RenderX

of sessions per month was 3.29. With a median of 40.2 minutes per month, this equated to roughly three 12-minute sessions.

Figure 3. Histogram showing the total number of self-reported sessions (x axis) by the percentage of participants (y axis).

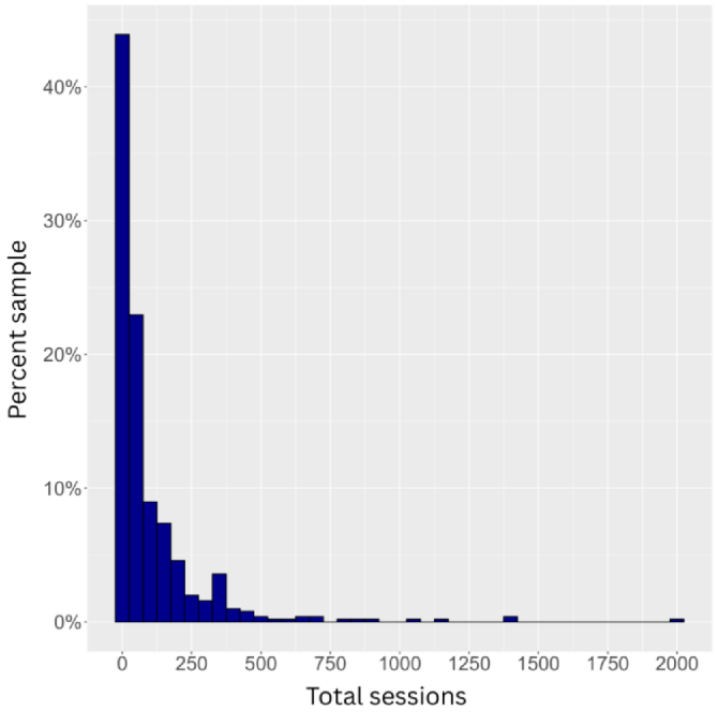

Figure 4. Total hours of practice by categorical groups based on time since app download. The y axis is truncated at 50 for visualization purposes, capturing 75% of the data. Horizontal bars indicate the median value for each category.

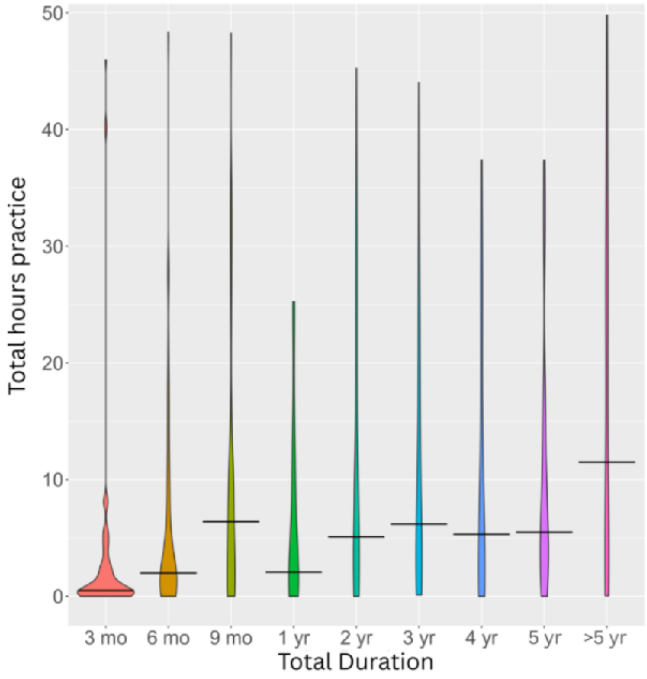

XSL•FO
RenderX

Figure 5. Total sessions by categorical groups based on time since app download. The y axis has been truncated for visualization purposes, capturing 75% of the data. Horizontal bars indicate median values for each category.

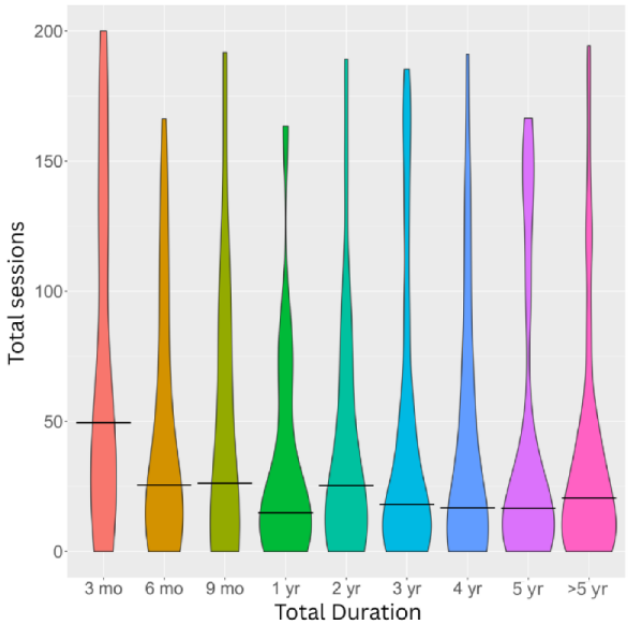

Engagement Statistics

Engagement variables were highly intercorrelated (r=0.495-0.999; see Section S14 in Multimedia Appendix 1). Objective adjusted minutes were the most reliable variable, but the sample was small and limited to 2 apps (n=156; Headspace and Waking Up). By contrast, subjective minutes with a subjective start date had a sample size 3 times larger (n=536). Given the strong association between objective and subjective start dates (r=0.783, P<.001), adjusted subjective minutes were calculated using subjective duration as the denominator to maximize sample size.

Categorical Demographic Factors and Engagement

Categorical demographic associations with engagement are reported in Table 3. Being female was associated with lower engagement on 1 outcome, while residing in the United Kingdom was associated with higher engagement across 3 of the 4 outcome variables.

Table 3. Point-biserial correlations for the association between noncontinuous variables and engagement.

Variables	Objective minutes	Subjective minutes	Adjusted objective	Adjusted subjective
Sex[b]	–0.019	0.080	–0.296[a]	–0.016
Residence				
Australia	–0.027	–0.016	–0.004	–0.040
United States	–0.071	–0.069	–0.065	–0.068
United Kingdom	0.102[a]	0.110[a]	0.075	0.106[b]
Canada	–0.032	–0.062	–0.029	–0.068
New Zealand	–0.021	–0.031	–0.007	0.038

[a]P<.05 without multiple comparisons correction.

[b]P<.05 with multiple comparisons correction. For the biserial sex correlation, 1=female, 0=male.

Engagement by App

Three 1-way analyses of variance indicated a significant effect of app type on engagement for subjective minutes ($F_{2,362}$=9.03, P=.002, η^2=.05) and adjusted subjective minutes ($F_{2,358}$=7.63, P=.001; see Sections S15 and S16 in Multimedia Appendix 1). No significant differences were found for objective minutes ($F_{2,362}$=0.972, P=.38).

User Factors

We defined robust associations as those present across 3 or more of the 4 engagement outcomes. Among user factors, only 9 of 20 variables met this criterion (see Figure 6): age, openness (BFI-S-2), readiness to change, expectation match, expectations for sleep, expectations for anxiety, expectations for happiness, expectations for thriving, and expectations for performance enhancement. After FDR correction, only 4 of 20 remained: age, readiness to change, expectations for sleep, and expectations for thriving (see Section 17 in Multimedia Appendix 1 for CIs).

XSL•FO
RenderX

Figure 6. Heatmap of correlation results for user factors and adjusted self-report, adjusted objective, total self-report, and total objective minutes of meditation app use. The heatmap illustrates the direction and magnitude of Spearman correlations, as shown in the legend. Numbers indicate correlation coefficients. Asterisks denote correlations significant at $P<.05$ after correction for multiple comparisons. Nonsignificant correlations are omitted. Outcomes include self-reported total minutes of meditation app use, self-reported minutes adjusted for app duration, objectively verified total minutes of use (via app screenshot), and objective minutes adjusted for app duration.

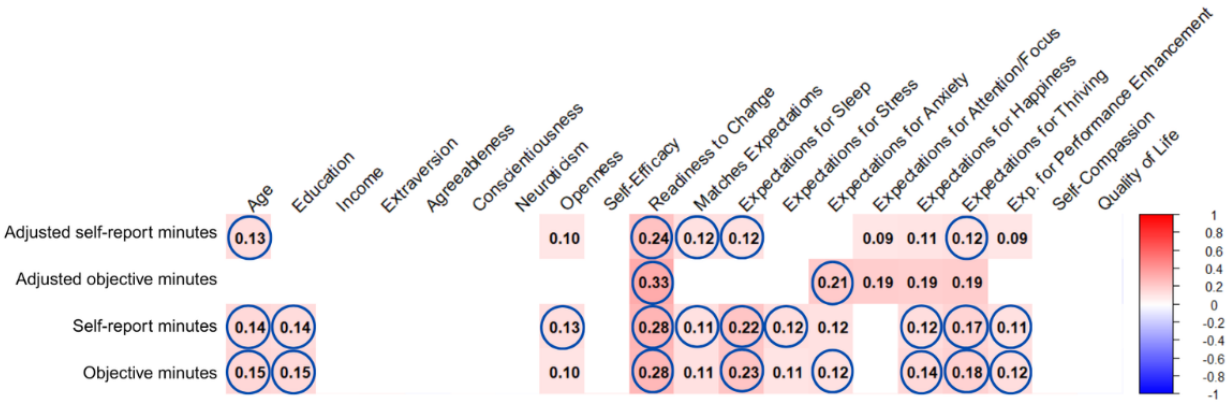

Mental Health Factors

Self-reported stress, depression, and psychological distress were negatively associated with app use. Specifically, distress was negatively associated with adjusted self-reported minutes, depression with unadjusted self-reported minutes, and current stress with both unadjusted self-reported minutes and objective minutes. However, no mental health factors remained significantly associated with any engagement outcome after correction for multiple comparisons (see Figure 7).

Figure 7. Heatmap of correlations between mental health factors and adjusted self-report, adjusted objective, total self-report, and total objective minutes of meditation app use. The heatmap shows the direction and magnitude of Spearman correlations, as indicated by the legend. Numbers represent correlation coefficients. Circles denote significant correlations at $P<.05$ after correction for multiple comparisons. Nonsignificant correlations are omitted. Outcomes include self-reported total minutes of meditation app use, self-reported minutes adjusted for app duration, objectively verified total minutes of use (via app screenshot), and objective minutes adjusted for app duration.



App Factors

DWAI-6 Total, uMARS Appeal, uMARS Perceived Quality, and Perceived Impact were associated with 3 of the 4 outcome variables. Of the app factors investigated, 7 were associated with at least one engagement outcome after FDR correction (see Figure 8): DWAI-6 Total, as well as Goal, Bond, and Task subscales, and uMARS Appeal, Perceived Quality, and Perceived Impact.

XSL•FO
RenderX

Figure 8. Heatmap of correlations between app factors and meditation app use outcomes. Displayed are Spearman correlation coefficients (numbers) showing the direction and magnitude of associations as indicated in the legend. Circles denote correlations significant at $P<.05$ after multiple comparison correction, and nonsignificant correlations are omitted. Outcomes include self-reported total minutes of app use, self-reported minutes adjusted for app duration, objective total minutes of use (verified via app screenshot), and objective minutes adjusted for app duration. DWAI: Digital Working Alliance Inventory; uMARS: user Mobile Application Rating Scale.

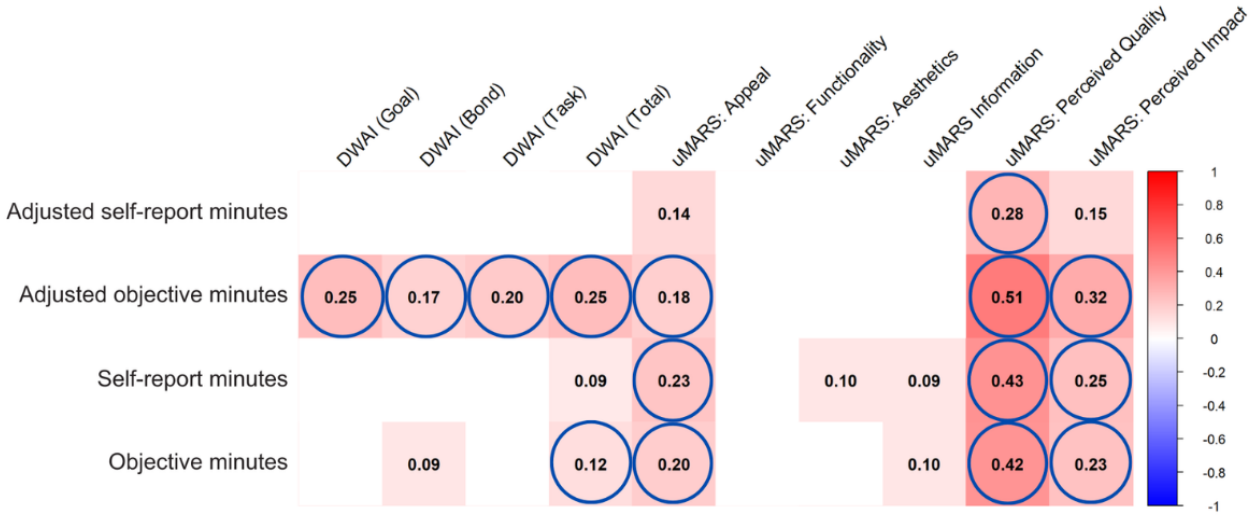

Outcome Regression Models

All models demonstrated a reasonable fit, explaining 12%-16% of the variance (see Table 4 and Section 18 in Multimedia Appendix 1). Significant ($P<.05$) predictors in 1 or more regression models included education, readiness to change, expectations for sleep, expectation match, the Perceived Quality subscale of the uMARS, and the Openness subscale of the BFI-S-2.

XSL•FO
RenderX

Table 4. Regression model coefficients for adjusted objective minutes, total objective minutes, adjusted self-report minutes, and total self-report minutes.

Standardized β coefficients for predictors in models 1–4	Adjusted objective minutes[a], standardized β coefficients	Objective minutes[b], standardized β coefficients	Adjusted self-report minutes[c], standardized β coefficients	Subjective minutes[d], standardized β coefficients
Intercept	–0.260[e]	–0.228[e]	–0.355[e]	–0.371[e]
User factors				
Sex	0.001	<0.001	–0.008	0.002
Country of residence (United Kingdom)	<.0001	0.002	0.008	0.004
Age	0.003	<–.001	0.004	–0.001
Education	0.033	0.153[e]	0.532[e]	0.237[e]
Big Five Inventory Short form 2 openness	0.010[f]	0.003	0.005	0.006
Readiness to change	<0.001	0.005[f]	0.027[g]	0.008[f]
Expectations (match)	<0.001	–0.003	0.023[f]	–0.004
Expectations for sleep	–0.002	0.004	–0.008	0.009[f]
Expectations for stress	<0.001	<–.001	0.012	<.001
Expectations for anxiety	0.001	–0.001	–0.020	–0.003
Expectations for happiness	–0.005	0.002	0.009	<.001
Expectations for thriving	0.008	0.003	–0.006	0.006
Expectations for performance enhancement	–0.004	–0.003	–0.015	–0.003
App factors				
DWAI-6[h] (Goal)	–0.009	0.004	–0.020	0.008
DWAI-6 (Bond)	–0.004	–0.003	–0.013	–0.002
DWAI-6 (Task)	0.010	0.002	0.024	0.004
uMARS[i] (Appeal)	–0.012	<.001	–0.016	0.005
uMARS (Perceived Quality)	0.025[g]	0.003[g]	0.041[e]	0.010[f]
uMARS (Perceived Impact)	–0.006	<–.001	–0.003	0.002
Adjusted R^2	0.158	0.150	0.126	0.137

[a]App-verified minutes of use per year adjusted for total duration of use in years.

[b]Total app-verified minutes of use.

[c]Self-report minutes of use per year adjusted for total duration of use in years.

[d]Self-report total minutes of use.

[e]P<.001.

[f]P<.05.

[g]P<.01.

[h]DWAI-6: 6-item Digital Working Alliance Inventory.

[i]uMARS: user Mobile Application Rating Scale.

Discussion

Principal Findings

We examined factors associated with engagement in popular meditation apps among 536 participants. Consistent with prior findings, most participants engaged minimally. Although apps were available for an average of 894 days (about 2.5 years), participants reported an average of 108 sessions, while app-verified data from about one-third of participants indicated 62 sessions on average. Half of the sample engaged in 3 or fewer sessions per month. Notably, engagement did not increase with longer app availability, suggesting a pattern of persistently low overall engagement.

Few significant correlations between individual user factors and engagement were observed, most of which were small in magnitude (r=0.09-0.30), with a few reaching the moderate range (r=0.30-0.50). After correction for multiple comparisons, positive associations with engagement remained for male sex, older age, higher education level, readiness to change, and expectations of the app for stress reduction, sleep improvement,

XSL•FO
RenderX

anxiety reduction, happiness, thriving, and performance enhancement. These results suggest that older, more educated users with greater readiness to change and higher expectations of the app are more likely to engage regularly. App factors—including perceived appeal, quality, and impact—were consistently associated with higher engagement, as were the 3 digital working alliance subscales (Goal, Task, and Bond). This indicates that both perceptions of the app and the perceived relationship with it may be important determinants of engagement.

User Factors Related to Use

Sociodemographics

Education was the variable most consistently associated with higher engagement. Meditation is more common among individuals who are White, middle-aged, wealthier, and better educated [41]. Lower levels of education have been linked to earlier disengagement or a failure to engage in meditation at all [41]. Higher education is also associated with greater engagement in mindfulness practices in US nationally representative surveys [11]. Lower levels of education have also been linked to poorer health outcomes [74,75] and are related to health literacy, which partially mediates health-promoting behaviors [76,77]. While lower education may contribute to poorer health outcomes and lower health literacy, other social factors may also reduce engagement. For example, individuals with lower education often have more fragmented leisure time, leaving less opportunity for regular, recurrent activities [78]. Male sex was associated with 1 engagement measure, which contrasts with prior research showing that females are generally more likely to engage in meditation practice, even after controlling for other demographic factors [12,42,79]. Previous studies have also found that males may demonstrate greater persistence in meditation [80]. As this correlation was observed only for adjusted objective use (available for Headspace and Waking Up), it may reflect patterns specific to users of these apps rather than meditation app use more broadly [40]. Notably, sex did not emerge as a significant predictor in the regression models.

Personality

Of the personality factors, only openness was related to engagement. Openness reflects general curiosity and a willingness to explore novel perspectives of one's subjective experience [81]. Individuals higher in openness are more likely to try meditation initially and persist despite encountering difficulties. Openness has also been associated with meditation practice outside of group meditation class settings [45]. In contrast to prior research, we found no associations between engagement and conscientiousness, extraversion, agreeableness, or neuroticism [55,82,83]. While conscientiousness was not related to engagement in this study, it has previously been linked to positive attitudes toward practice [43]. Similarly, neuroticism showed no association with engagement here, although prior work has linked it to perceiving more barriers to practice [84,85].

Mental Health

None of the 8 mental health factors were significantly associated with engagement. Previous research has found meditation apps to be modestly effective for depression and anxiety [12,19], potentially serving as a form of self-managed treatment for individuals facing barriers to mental health care [41]. However, no such associations with mental health factors were observed in this study. In a previous study, motivation for mental health was negatively associated with app use [40]. Meditation can negatively impact mental health [86]. While these meditation-related adverse events do not always result in impairment, about half of meditators report experiencing an adverse effect, and 9.1% report functional impairment as a result [86]. Individuals who do not experience benefits or who encounter adverse effects may disengage shortly after download. Furthermore, meditating for mental health reasons has been negatively associated with the total amount of meditation practice completed over the long term [27,87]. Individuals with higher lifetime meditation practice often shift toward spiritual motivations as their practice progresses [29]. However, the retrospective design of our study limits causal inferences.

App Factors

uMARS

Five of the 6 uMARS subscales were associated with engagement. Previous research suggests that aesthetics and appeal relate to meditation app engagement [53], although in our study, aesthetics were not robustly associated after FDR correction. The Perceived Quality subscale showed the strongest association (r=0.51), indicating that user perceptions may drive both usage and beliefs in the app's effectiveness. Perceived impact was also robustly associated with engagement. Given the retrospective design, survivorship bias should be considered: users who continued using the apps likely enjoyed them, while those who did not may have stopped. It is also possible that users who experienced benefits from their chosen app developed increasingly positive app appraisals over time.

Digital Working Alliance

The DWAI-6 Goal, Bond, and Task subscales, as well as the overall score, were associated with engagement, consistent with prior findings [88]. All subscales correlated with adjusted objective minutes—the most reliable outcome measure, computed using app-verified minutes and download date—but this could only be calculated for apps that provide download dates (Headspace and Waking Up; n=151). Therapeutic alliance and engagement may promote each other [88]. While therapeutic alliance is considered important in digital mental health [89], current measures are adaptations of traditional, human-centered alliance scales. Incorporating human-computer interaction perspectives may provide greater nuance, particularly for anthropomorphic scale items [90]. Despite this limitation, therapeutic alliance with apps remains relevant to engagement, as alignment between a user's goals and perceived app support may encourage continued use.

One consideration for both the uMARS and DWAI-6 is that several subscales demonstrated relatively poor internal consistency. The reliability of the uMARS Engagement and

XSL•FO
RenderX

Subjective Quality subscales, as well as the DWAI-6 Bond and Task subscales, ranged from acceptable to poor, which reduces confidence in the constructs being measured.

Expectations for Efficacy

Higher expectations of efficacy across 6 of the 7 domains assessed (sleep, stress, anxiety, happiness, thriving, and performance enhancement) were generally associated with higher engagement, with the exception of expectations for attention/focus. Only expectations for sleep were significant in the regression model. These findings align with our predictions. Experimental and prospective studies have shown that failing to meet expectations is more predictive of behavior than matched expectations [91,92]. Unmet or low expectations negatively influence engagement and perceived usefulness, whereas met or exceeded expectations positively affect behavior and perceptions [92]. Expectations are closely linked to app ratings, as features such as goal setting and feedback enhance beliefs in an app's effectiveness [34]. These features can also foster positive experiences of progress, creating a feedback loop that promotes further engagement [33,48]. In the absence of human interaction, the relationship between a user and an app is shaped by the "user journey"—the path a user follows through the app's design. Persuasive design can help establish and meet user expectations.

A general rating of whether expectations were met was weakly associated with engagement. This result aligns with literature suggesting that matched expectations have a positive influence on behavior and mismatched expectations have a negative effect [93]. It is worth noting that we asked, "To what extent did your experience match your initial expectations?" without specifying which expectations participants should consider. Consequently, this approach may have captured only an overall impression of expectation match.

Readiness to Change

Readiness to change showed robust, moderate associations and accounted for a significant proportion of variance in the regression model. The readiness-to-change ruler used in this study is actively employed in behavior change interventions and is based on the Transtheoretical Model of Change, which conceptualizes behavior change in stages [32,94,95]. Readiness to change shows promise as one of the most predictive factors of actual behavior change, as it is conceptually closely linked to both motivation and behavior. These findings align with broader evidence connecting readiness ratings to actual behavior change, particularly in health-related contexts [63,95,96]. This relationship could inform app design, allowing offerings to be tailored to users' readiness levels. The same single-item measure used in our study could be implemented immediately after app download to tailor the length, complexity, and type of practice to users' readiness levels. For example, users with lower readiness could be offered shorter, simpler meditations or psychoeducational content about meditation to reduce perceived barriers and enhance understanding of the practice.

Self-Efficacy and App Ratings in Building Habits

Contrary to our expectations, self-efficacy was not related to engagement. Previous research on habit formation suggests that self-efficacy may support the maintenance of a target behavior before a habit is established. There is limited evidence that self-efficacy promotes habit-building [33,48,97] and increases with ongoing meditation practice [98]; however, results are mixed [99,100]. One likely reason self-efficacy did not predict engagement is that expectations, perceptions, and habit formation played larger roles. A person may believe they can achieve a goal, but if they are not committed or do not perceive long-term utility, they may lack motivation to engage. This may explain why readiness to change was associated with engagement, whereas self-efficacy was not.

Limitations

One key limitation of this study is that its retrospective design precludes causal inferences, although research on meditation app engagement is generally scarce. Additionally, we cannot confirm detailed usage patterns, such as extended gaps or cessation points; however, our estimates of sessions per month provide a rough indication of practice regularity. This study included cross-app comparisons, which few prior studies have conducted. Such comparisons are valuable, given that all therapeutic alliance subscales and half of the uMARS subscales were associated with engagement after correction for multiple comparisons. However, by not focusing on a specific app, the sample was disproportionately composed of users of the most popular apps.

Another limitation was that our most reliable outcome variable—objective minutes adjusted for verified app duration—was restricted to apps that displayed the month or year of joining. As a result, the sample for objective-adjusted minutes comprised only about one-third of the self-reported sample. Nevertheless, objective minutes were highly correlated with self-reported minutes, which may mitigate some concerns, although it is possible that individuals who can view their app-recorded minutes rely on these records when self-reporting.

Our data quality may have been influenced by self-selection, socioeconomic skew, and the compensation structure in our Prolific sample. Nevertheless, research indicates that among popular online survey platforms, Prolific consistently provides high-quality data across a wide range of measures [101]. Additionally, our data may have been skewed by the overrepresentation of the most popular meditation apps, limiting the generalizability of the findings to less popular apps or those with a narrower focus.

A final significant issue concerns what the outcome measures captured. While meditation was the central focus of the included apps, many also offer alternative exercises that contribute to the measured minutes, including—but not limited to—breathwork, sleep stories, and podcasts. This is an issue because sleep stories may continue running for hours after an individual falls asleep and be recorded as meditation. Future studies that can distinguish between different activities will provide more accurate statistics on meditation engagement.

Future Directions

A group-level comparison of engagers and disengagers could reveal cluster effects, where active users share similar characteristics. Baumel et al [4] observed a drop-off trend in a

XSL•FO
RenderX

large sample but noted that understanding precisely why people engage or disengage during this period would be of interest. While there is a high use rate among those who continue engaging beyond the first week, this represents only a small portion of users. By including all users over the past 180 days, we obtain a general picture of app use across the population; however, this approach results in a large sample of disengagers and only a small sample of active users, limiting our power to detect effect sizes within the subsample of engagers.

Longitudinal analysis could directly examine temporal and causal aspects of engagement, account for changes in contributing factors over time, and provide a clearer understanding of baseline predictors. For example, longitudinal data could track whether changes in mental health outcomes influence engagement. This study did not find any significant associations for mental health outcomes that survived multiple comparisons. However, we relied on participants' reports of mental health status following meditation app use. Apps have been shown to reduce outcomes such as stress, depression, and anxiety [12], and such changes could positively or negatively reinforce app use. Moods, circumstances, and lifestyles can fluctuate widely over extended periods. User ratings of apps using scales such as the uMARS may better explain engagement when app rating and user engagement occur close together. Longitudinal analysis also allows for baseline measurement of variables, such as expectations, which can then be compared with actual experiences at follow-up. The low proportion of variability accounted for suggests that factors outside the model have a significant impact on engagement.

Conclusions

This study aimed to explore factors influencing engagement with popular meditation apps, highlighting a substantial early drop-off. Although the models accounted for only a small proportion of overall variance, the findings emphasize the importance of user characteristics and app quality in sustaining engagement. This exploratory study aimed to examine a wide range of factors potentially relevant to meditation app engagement. The results indicated that older, more educated users, as well as those with higher expectations of apps and greater readiness to change, were more likely to engage with the apps regularly.

Acknowledgments

We acknowledge the coding contributions of Alex Burger and Karen Trapani, as well as the administrative support of Cathleen Benevento. This study would not have been possible without the support of the Contemplative Studies Centre and the Melbourne School of Psychological Sciences community at the University of Melbourne. Funding was provided to establish the Contemplative Studies Centre via a philanthropic gift from the Three Springs Foundation, Pty, Ltd. No original content in the manuscript was generated by artificial intelligence. The authors occasionally used the free version of Grammarly (Grammarly, Inc) for content flow, such as identifying extraneous words in a sentence.

Data Availability

Data and analytic code are uploaded to the Open Science Framework registry [102].

Authors' Contributions

Conceptualization: JA, JD, NTVD, JG
Data curation: JA, JD, PW, NTVD
Formal analysis: JA, JD, NTVD
Funding acquisition: NTVD
Methodology: JA, JD, FM, JG, NTVD
Resources: JA, JD, JG, NTVD
Visualization: JA, NTVD
Writing – original draft: JA, JG, NTVD
Writing – review & editing: JA, JG, SDA, NTVD

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional analysis.
[DOCX File , 727 KB - jmir_v28i1e71960_app1.docx]

References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022 Feb;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]

XSL•FO
RenderX

2. Lattie EG, Stiles-Shields C, Graham AK. An overview of and recommendations for more accessible digital mental health services. Nat Rev Psychol 2022 Feb;1(2):87-100 [FREE Full text] [doi: 10.1038/s44159-021-00003-1] [Medline: 38515434]
3. Schueller S, Hunter J, Figueroa C, Aguilera A. Use of digital mental health for marginalized and underserved populations. Curr Treat Options Psych 2019 Jul 5;6(3):243-255 [FREE Full text] [doi: 10.1007/s40501-019-00181-z] [Medline: 38515434]
4. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567-e14255 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
5. Goldberg SB, Baldwin SA, Riordan KM, Torous J, Dahl CJ, Davidson RJ, et al. Alliance with an unguided smartphone app: validation of the digital working alliance inventory. Assessment 2022 Sep 25;29(6):1331-1345 [FREE Full text] [doi: 10.1177/10731911211015310] [Medline: 34000843]
6. Lutz A, Slagter HA, Dunne JD, Davidson RJ. Attention regulation and monitoring in meditation. Trends Cogn Sci 2008 Apr 16;12(4):163-169 [FREE Full text] [doi: 10.1016/j.tics.2008.01.005] [Medline: 18329323]
7. Kabat-Zinn J. Mindfulness-based interventions in context: past, present, and future. Clinical Psychology: Science and Practice 2003;10(2):144-156 [FREE Full text] [doi: 10.1093/clipsy.bpg016]
8. Lam S, Riordan K, Simonsson O, Davidson R, Goldberg S. Who sticks with meditation? Rates and redictors of persistence in a population-based sample in the USA. Mindfulness (N Y) 2023 Jan 06;14(1):66-78 [FREE Full text] [doi: 10.1007/s12671-022-02061-9] [Medline: 36777474]
9. Lee RL. Review of headspace: meditation and sleep. Fam Syst Health 2023 Mar 06;41(1):114-116. [doi: 10.1037/fsh0000737] [Medline: 36951700]
10. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11-116 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
11. Olano HA, Kachan D, Tannenbaum SL, Mehta A, Annane D, Lee DJ. Engagement in mindfulness practices by U.S. adults: sociodemographic barriers. J Altern Complement Med 2015 Feb;21(2):100-102 [FREE Full text] [doi: 10.1089/acm.2014.0269] [Medline: 25685958]
12. Gál É, tefan S, Cristea IA. The efficacy of mindfulness meditation apps in enhancing users' well-being and mental health related outcomes: a meta-analysis of randomized controlled trials. J Affect Disord 2021 Jan 15;279(1):131-142 [FREE Full text] [doi: 10.1016/j.jad.2020.09.134] [Medline: 33049431]
13. Crane RS, Brewer J, Feldman C, Kabat-Zinn J, Santorelli S, Williams JMG, et al. What defines mindfulness-based programs? The warp and the weft. Psychol Med 2017 Apr;47(6):990-999. [doi: 10.1017/S0033291716003317] [Medline: 28031068]
14. Goldberg SB, Tucker RP, Greene PA, Davidson RJ, Wampold BE, Kearney DJ, et al. Mindfulness-based interventions for psychiatric disorders: a systematic review and meta-analysis. Clin Psychol Rev 2018 Feb;59(6):52-60 [FREE Full text] [doi: 10.1016/j.cpr.2017.10.011] [Medline: 29126747]
15. Galante J, Friedrich C, Collaboration of Mindfulness Trials (CoMinT) 3, Dalgleish T, Jones PB, White IR, Collaboration of Mindfulness Trials (CoMinT). Individual participant data systematic review and meta-analysis of randomised controlled trials assessing adult mindfulness-based programmes for mental health promotion in non-clinical settings. Nat Ment Health 2023 Jul 10;1(7):462-476 [FREE Full text] [doi: 10.1038/s44220-023-00081-5] [Medline: 37867573]
16. Bowles NI, Davies JN, Van Dam NT. Dose-response relationship of reported lifetime meditation practice with mental health and wellbeing: a cross-sectional study. Mindfulness (N Y) 2022 Feb;13(10):2529-2546. [doi: 10.1007/s12671-022-01977-6] [Medline: 36193220]
17. Schultchen D, Terhorst Y, Holderied T, Stach M, Messner EM, Baumeister H, et al. Stay present with your phone: a systematic review and standardized rating of mindfulness apps in European app stores. Int J Behav Med 2021 Oct 10;28(5):552-560. [doi: 10.1007/s12529-020-09944-y] [Medline: 33215348]
18. Spijkerman MPJ, Pots WTM, Bohlmeijer ET. Effectiveness of online mindfulness-based interventions in improving mental health: a review and meta-analysis of randomised controlled trials. Clin Psychol Rev 2016 Apr;45(4):102-114 [FREE Full text] [doi: 10.1016/j.cpr.2016.03.009] [Medline: 27111302]
19. Sommers-Spijkerman M, Austin J, Bohlmeijer ET, Pots W. New evidence in the booming field of online mindfulness: an updated meta-analysis of randomized controlled trials. JMIR Ment Health 2021 Jul 19;8(7):e28168-e28114 [FREE Full text] [doi: 10.2196/28168] [Medline: 34279240]
20. Jiang A, Rosario M, Stahl S, Gill JM, Rusch HL. The effect of virtual mindfulness-based interventions on sleep quality: a systematic review of randomized controlled trials. Curr Psychiatry Rep 2021 Jul 23;23(9):62 [FREE Full text] [doi: 10.1007/s11920-021-01272-6] [Medline: 34297230]
21. DiMatteo MR, Giordani PJ, Lepper HS, Croghan TW. Patient adherence and medical treatment outcomes: a meta-analysis. Med Care 2002 Sep;40(9):794-811. [doi: 10.1097/00005650-200209000-00009] [Medline: 12218770]
22. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on Internet-based mental health interventions — a systematic review. Internet Interventions 2014 Oct;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]
23. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018 Jun 06;20(6):e199-e215 [FREE Full text] [doi: 10.2196/jmir.9275] [Medline: 29875089]

XSL•FO
RenderX

24. Szinay D, Jones A, Chadborn T, Brown J, Naughton F. Influences on the uptake of and engagement with health and well-being smartphone apps: systematic review. J Med Internet Res 2020 May 29;22(5):e17572 [FREE Full text] [doi: 10.2196/17572] [Medline: 32348255]
25. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017 Jun 29;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
26. Lam SU, Kirvin-Quamme A, Goldberg SB. Overall and Differential Attrition in Mindfulness-Based Interventions: A Meta-Analysis. Mindfulness (N Y) 2022 Nov;13(11):2676-2690 [FREE Full text] [doi: 10.1007/s12671-022-01970-z] [Medline: 36506616]
27. Strohmaier S. The relationship between doses of mindfulness-based programs and depression, anxiety, stress, and mindfulness: a dose-response meta-regression of randomised controlled trials. Mindfulness 2020 Mar 02;11(6):1315-1335 [FREE Full text] [doi: 10.1007/s12671-020-01319-4] [Medline: 27966189]
28. Yik LL, Ling LM, Ai LM, Ting AB, Capelle DP, Zainuddin SI, et al. The effect of 5-minute mindfulness of peace on suffering and spiritual well-being among palliative care patients: a randomized controlled study. Am J Hosp Palliat Care 2021 Sep 28;38(9):1083-1090. [doi: 10.1177/1049909120965944] [Medline: 33078627]
29. Bowles NI, Van Dam NT. Dose-response effects of reported meditation practice on mental-health and wellbeing: a prospective longitudinal study. Appl Psychol Health Well Being 2025 Aug 20;17(4):e70063-e71090. [doi: 10.1111/aphw.70063] [Medline: 40785526]
30. Bandura A. Social Foundations of Thought and Action: A Social Cognitive Theory. Hoboken, NJ: Prentice-Hall; 1986.
31. Ajzen I. The theory of planned behavior. Organizational Behavior and Human Decision Processes 1991 Dec 1;50(2):179-211.
32. DiClemente CC, Prochaska O. Toward a comprehensive, transtheoretical model of change: stages of change and addictive behaviors. In: Treating Addictive Behaviors (2nd Edition). Berlin/Heidelberg, Germany: Springer; 1998:3-27.
33. Gardner B, Lally P, Wardle J. Making health habitual: the psychology of 'habit-formation' and general practice. Br J Gen Pract 2012 Dec;62(605):664-666 [FREE Full text] [doi: 10.3399/bjgp12X659466] [Medline: 23211256]
34. Oinas-Kukkonen H, Harjumaa M. Persuasive systems design: key issues, process model, and system features. CAIS 2009;24(605):664-666 [FREE Full text] [doi: 10.17705/1cais.02428]
35. Lally P, Gardner B. Promoting habit formation. Health Psychology Review 2013 May;7(sup1):S137-S158. [doi: 10.1080/17437199.2011.603640]
36. Crandall A, Cheung A, Young A, Hooper AP. Theory-based predictors of mindfulness meditation mobile app usage: a survey and cohort study. JMIR Mhealth Uhealth 2019 Mar 22;7(3):e10794-e1S158 [FREE Full text] [doi: 10.2196/10794] [Medline: 30900992]
37. Hesse M. The Readiness Ruler as a measure of readiness to change poly-drug use in drug abusers. Harm Reduct J 2006 Jan 25;3(3):3 [FREE Full text] [doi: 10.1186/1477-7517-3-3] [Medline: 16436208]
38. Bowen M, Beam M. Mapping mindfulness: assessing the stages of meditation habit formation in the USA using the Sussex Mindfulness Meditation (SuMMed) Model. Mindfulness 2025 Apr 07;16(5):1340-1351. [doi: 10.1007/s12671-025-02555-2]
39. Miles E, Matcham F, Strauss C, Cavanagh K. Making mindfulness meditation a healthy habit. Mindfulness 2023 Nov 28;14(12):2988-3005. [doi: 10.1007/s12671-023-02258-6]
40. Lam SU, Xie Q, Goldberg SB. Situating meditation apps within the ecosystem of meditation practice: population-based survey study. JMIR Ment Health 2023 Apr 28;10:e43565 [FREE Full text] [doi: 10.2196/43565] [Medline: 37115618]
41. Davies JN, Faschinger A, Galante J, Van Dam NT. Prevalence and 20-year trends in meditation, yoga, guided imagery and progressive relaxation use among US adults from 2002 to 2022. Sci Rep 2024 Jul 01;14(1):14987 [FREE Full text] [doi: 10.1038/s41598-024-64562-y] [Medline: 38951149]
42. Jakob R, Harperink S, Rudolf AM, Fleisch E, Haug S, Mair JL, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371 [FREE Full text] [doi: 10.2196/35371] [Medline: 35612886]
43. Osin EN, Turilina II. Mindfulness meditation experiences of novice practitioners in an online intervention: trajectories, predictors, and challenges. Appl Psychol Health Well Being 2022 Feb 15;14(1):101-121. [doi: 10.1111/aphw.12293] [Medline: 34268871]
44. Kim S, Park JY, Chung K. The relationship between the big five personality traits and the theory of planned behavior in using mindfulness mobile apps: cross-sectional survey. J Med Internet Res 2022 Nov 30;24(11):e39501 [FREE Full text] [doi: 10.2196/39501] [Medline: 36449344]
45. Canby NK, Eichel K, Peters SI, Rahrig H, Britton WB. Predictors of out-of-class mindfulness practice adherence during and after a mindfulness-based intervention. Psychosom Med 2021 Oct 8;83(6):655-664 [FREE Full text] [doi: 10.1097/PSY.0000000000000873] [Medline: 33038188]
46. Stojanovic M, Fries S, Grund A. Self-efficacy in habit building: how general and habit-specific self-efficacy influence behavioral automatization and motivational interference. Front Psychol 2021 Aug;12(8):643753-643899. [doi: 10.3389/fpsyg.2021.643753] [Medline: 34025512]

XSL•FO
RenderX

47. Schiwal AT, Fauth EB, Wengreen H, Norton M. The gray matters app targeting health behaviors associated with Alzheimer's risk: improvements in intrinsic motivation and impact on diet quality and physical activity. J Nutr Health Aging 2020 Dec 22;24(8):893-899. [doi: 10.1007/s12603-020-1421-5] [Medline: 33009542]
48. Phillips W, Hine D. Self-compassion, physical health, and health behaviour: a meta-analysis. Health Psychol Rev 2021 Mar 22;15(1):113-139. [doi: 10.1080/17437199.2019.1705872] [Medline: 31842689]
49. Jones F, Harris P, Waller H, Coggins A. Adherence to an exercise prescription scheme: the role of expectations, self-efficacy, stage of change and psychological well-being. Br J Health Psychol 2005 Sep;10(Pt 3):359-378. [doi: 10.1348/135910704X24798] [Medline: 16238853]
50. Laurie J, Blandford A. Making time for mindfulness. Int J Med Inform 2016 Dec 01;96(4):38-50. [doi: 10.1016/j.ijmedinf.2016.02.010] [Medline: 26965526]
51. Banerjee A, Banerji R, Berry J. From proof of concept to scalable policies: challenges and solutions, with an application. Journal of Economic Perspectives 2017;31(4):73-102.
52. Prince M, Patel V, Saxena S, Maj M, Maselko J, Phillips MR, et al. No health without mental health. Lancet 2007 Sep 08;370(9590):859-877. [doi: 10.1016/S0140-6736(07)61238-0] [Medline: 17804063]
53. Baumel A, Kane JM. Examining predictors of real-world user engagement with self-guided eHealth interventions: analysis of mobile apps and websites using a novel dataset. J Med Internet Res 2018 Dec 14;20(12):e11491 [FREE Full text] [doi: 10.2196/11491] [Medline: 30552077]
54. Alqahtani F, Al Khalifah G, Oyebode O, Orji R. Apps for mental health: an evaluation of behavior change strategies and recommendations for future development. Front Artif Intell 2019 Dec 17;2:30 [FREE Full text] [doi: 10.3389/frai.2019.00030] [Medline: 33733119]
55. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387. [doi: 10.2196/24387] [Medline: 33759801]
56. Van Dam NT, van Vugt MK, Vago DR. . Mind the Hype: A Critical Evaluation and Prescriptive Agenda for Research on Mindfulness and Meditation. Perspectives on Psychological Science. 2018. URL: https://doi.org/10.1177/1745691617709589 [accessed 2025-10-24]
57. Brazier J, Peasgood T, Mukuria C, Marten O, Kreimeier S, Luo N, et al. The EQ-HWB: Overview of the Development of a Measure of Health and Wellbeing and Key Results. Value Health 2022 Apr;25(4):482-491 [FREE Full text] [doi: 10.1016/j.jval.2022.01.009] [Medline: 35277337]
58. Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SLT, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002 Aug;32(6):959-976 [FREE Full text] [doi: 10.1017/s0033291702006074] [Medline: 12214795]
59. Tennant R, Hiller L, Fishwick R, Platt S, Joseph S, Weich S, et al. The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. Health Qual Life Outcomes 2007 Nov 27;5(6):63-76 [FREE Full text] [doi: 10.1186/1477-7525-5-63] [Medline: 18042300]
60. Cheung F, Lucas RE. Assessing the validity of single-item life satisfaction measures: results from three large samples. Qual Life Res 2014 Dec 27;23(10):2809-2818 [FREE Full text] [doi: 10.1007/s11136-014-0726-4] [Medline: 24890827]
61. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
62. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr 22;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
63. Raes F, Pommier E, Neff KD, Van Gucht D. Construction and factorial validation of a short form of the Self-Compassion Scale. Clin Psychol Psychother 2011 Apr 08;18(3):250-255. [doi: 10.1002/cpp.702] [Medline: 21584907]
64. Romppel M, Herrmann-Lingen C, Wachter R, Edelmann F, Düngen HD, Pieske B, et al. A short form of the General Self-Efficacy Scale (GSE-6): Development, psychometric properties and validity in an intercultural non-clinical sample and a sample of patients at risk for heart failure. Psychosoc Med 2013;10(3):Doc01-Doc05 [FREE Full text] [doi: 10.3205/psm000091] [Medline: 23429426]
65. Dixon JB, Laurie CP, Anderson ML, Hayden MJ, Dixon ME, O'Brien PE. Motivation, readiness to change, and weight loss following adjustable gastric band surgery. Obesity (Silver Spring) 2009 Apr;17(4):698-705 [FREE Full text] [doi: 10.1038/oby.2008.609] [Medline: 19148126]
66. Eshah NF. Readiness for Behavior Change in Patients Living With Ischemic Heart Disease. J Nurs Res 2019 Dec;27(6):e57-705 [FREE Full text] [doi: 10.1097/jnr.0000000000000336] [Medline: 31283634]
67. Finsrud I, Nissen-Lie HA, Vrabel K, Høstmælingen A, Wampold BE, Ulvenes PG. It's the therapist and the treatment: The structure of common therapeutic relationship factors. Psychother Res 2022 Feb;32(2):139-150 [FREE Full text] [doi: 10.1080/10503307.2021.1916640] [Medline: 33938407]
68. Soto CJ, John OP. The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. J Pers Soc Psychol 2017 Jul 02;113(1):117-143 [FREE Full text] [doi: 10.1037/pspp0000096] [Medline: 27055049]

XSL•FO
RenderX

69. Stoyanov SR, Hides L, Kavanagh DJ, Wilson H. Development and Validation of the User Version of the Mobile Application Rating Scale (uMARS). JMIR Mhealth Uhealth 2016 Jun 10;4(2):e72-143 [FREE Full text] [doi: 10.2196/mhealth.5849] [Medline: 27287964]
70. Yohai V. High Breakdown-Point and High Efficiency Robust Estimates for Regression. Ann. Statist 1987 Jun 1;15(2):e72 [FREE Full text] [doi: 10.1214/aos/1176350366] [Medline: 27287964]
71. Koller M, Stahel W. Sharpening Wald-type inference in robust regression for small samples. Computational Statistics & Data Analysis 2011 Aug 1;55(8):2504-2515. [doi: 10.1016/j.csda.2011.02.014]
72. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) 1995 Jan 01;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
73. Yentes R, Wilhelm F. careless: Procedures for Computing Indices of Careless Responding. Package 'careless' 2021 [FREE Full text]
74. Kunst AE, Bos V, Lahelma E, Bartley M, Lissau I, Regidor E, et al. Trends in socioeconomic inequalities in self-assessed health in 10 European countries. Int J Epidemiol 2005 Apr;34(2):295-305 [FREE Full text] [doi: 10.1093/ije/dyh342] [Medline: 15563586]
75. Mackenbach JP, Stirbu I, Roskam AJR, Schaap MM, Menvielle G, Leinsalu M, European Union Working Group on Socioeconomic Inequalities in Health. Socioeconomic inequalities in health in 22 European countries. N Engl J Med 2008 Jun 05;358(23):2468-2481 [FREE Full text] [doi: 10.1056/NEJMsa0707519] [Medline: 18525043]
76. Barber MN, Staples M, Osborne RH, Clerehan R, Elder C, Buchbinder R. Up to a quarter of the Australian population may have suboptimal health literacy depending upon the measurement tool: results from a population-based survey. Health Promot Int 2009 Sep 05;24(3):252-261. [doi: 10.1093/heapro/dap022] [Medline: 19531559]
77. van der Heide I, Wang J, Droomers M, Spreeuwenberg P, Rademakers J, Uiters E. The relationship between health, education, and health literacy: results from the Dutch Adult Literacy and Life Skills Survey. J Health Commun 2013 Sep 13;18 Suppl 1(Suppl 1):172-184 [FREE Full text] [doi: 10.1080/10810730.2013.825668] [Medline: 24093354]
78. Sevilla A, Gimenez-Nadal JI, Gershuny J. Leisure inequality in the United States: 1965-2003. Demography 2012 Aug;49(3):939-964 [FREE Full text] [doi: 10.1007/s13524-012-0100-5] [Medline: 22589003]
79. Winter N, Russell L, Ugalde A, White V, Livingston P. Engagement Strategies to Improve Adherence and Retention in Web-Based Mindfulness Programs: Systematic Review. J Med Internet Res 2022 Jan 12;24(1):e30026-e30064 [FREE Full text] [doi: 10.2196/30026] [Medline: 35019851]
80. Burke A, Lam CN, Stussman B, Yang H. Prevalence and patterns of use of mantra, mindfulness and spiritual meditation among adults in the United States. BMC Complement Altern Med 2017 Jun 15;17(1):316 [FREE Full text] [doi: 10.1186/s12906-017-1827-8] [Medline: 28619092]
81. Goldberg LR. The structure of phenotypic personality traits. Am Psychol 1993 Jan 15;48(1):26-34. [doi: 10.1037//0003-066x.48.1.26] [Medline: 8427480]
82. Khwaja M, Pieritz S, Faisal AA, Matic A. Personality and Engagement with Digital Mental Health Interventions. In: Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization. 1993 Jan Presented at: UMAP Confefence on User Modelling, Adaptation and Personalization; June 21-25, 2021; Utrecht, Netherlands p. 235-239. [doi: https://doi.org/10.1145/3450613.3456823]
83. Alqahtani F, Meier S, Orji R. Personality-based approach for tailoring persuasive mental health applications. User Model User-Adap Inter 2021 Jul 16;32(3):253-295 [FREE Full text] [doi: 10.1007/s11257-021-09289-5]
84. Whitford S, Warren K. Perceived Barriers to Meditation Among College Students. Building Healthy Academic Communities Journal 2019;32(3):23-33. [doi: 10.18061/bhac.v3i1.6678]
85. Delmonte MM. Personality correlates of meditation practice frequency and dropout in an outpatient population. J Behav Med 1988 Dec 29;11(6):593-597. [doi: 10.1007/BF00844908] [Medline: 3075240]
86. Van Dam N, Targett J, Davies J, Burger A, Galante J. Incidence and Predictors of Meditation-Related Unusual Experiences and Adverse Effects in a Representative Sample of Meditators in the United States. Clinical Psychological Science 2025 Jan 06;13(3):632-648. [doi: 10.1177/21677026241298269] [Medline: 3075240]
87. Jiwani Z, Lam SU, Richard JD, Goldberg SB. Motivation for Meditation and its Association with Meditation Practice in a National Sample of Internet Users. Mindfulness (N Y) 2022 Oct 06;13(10):2641-2651 [FREE Full text] [doi: 10.1007/s12671-022-01985-6] [Medline: 36506892]
88. Clarke J, Proudfoot J, Whitton A, Birch MR, Boyd M, Parker G, et al. Therapeutic Alliance With a Fully Automated Mobile Phone and Web-Based Intervention: Secondary Analysis of a Randomized Controlled Trial. JMIR Ment Health 2016 Feb 25;3(1):e10-2651 [FREE Full text] [doi: 10.2196/mental.4656] [Medline: 26917096]
89. Tremain H, McEnery C, Fletcher K, Murray G. The Therapeutic Alliance in Digital Mental Health Interventions for Serious Mental Illnesses: Narrative Review. JMIR Ment Health 2020 Aug 07;7(8):e17204 [FREE Full text] [doi: 10.2196/17204] [Medline: 32763881]
90. D'Alfonso S, Lederman R, Bucci S, Berry K. The Digital Therapeutic Alliance and Human-Computer Interaction. JMIR Ment Health 2020 Dec 29;7(12):e21895 [FREE Full text] [doi: 10.2196/21895] [Medline: 33372897]

XSL•FO
RenderX

91. Armitage CJ, Norman P, Alganem S, Conner M. Expectations are more predictive of behavior than behavioral intentions: evidence from two prospective studies. Ann Behav Med 2015 Apr 29;49(2):239-246 [FREE Full text] [doi: 10.1007/s12160-014-9653-4] [Medline: 25623893]
92. Armitage CJ, Norman P, Alganem S, Conner M. Expectations are more predictive of behavior than behavioral intentions: evidence from two prospective studies. Ann Behav Med 2015 Apr 26;49(2):239-246 [FREE Full text] [doi: 10.1007/s12160-014-9653-4] [Medline: 25623893]
93. Bhattacherjee A. Understanding Information Systems Continuance: An Expectation-Confirmation Model. MIS Quarterly 2001 Sep 26;25(3):351-617. [doi: 10.2307/3250921]
94. Zimmerman GL, Olsen CG, Bosworth MF. A 'stages of change' approach to helping patients change behavior. Am Fam Physician 2000 Mar 01;61(5):1409-1416 [FREE Full text] [Medline: 10735346]
95. Zimmerman GL, Olsen CG, Bosworth MF. A 'stages of change' approach to helping patients change behavior. Am Fam Physician 2000 Mar 01;61(5):1409-1416. [Medline: 10735346]
96. Moyers TB, Martin T, Houck JM, Christopher PJ, Tonigan JS. From in-session behaviors to drinking outcomes: a causal chain for motivational interviewing. J Consult Clin Psychol 2009 Dec;77(6):1113-1124 [FREE Full text] [doi: 10.1037/a0017189] [Medline: 19968387]
97. Singh A. Self-efficacy and well-being among students: role of goal meditation. International Journal of Indian Psychology 2019;7(2):405-414. [doi: 10.25215/0702.049] [Medline: 19968387]
98. Goldstein L, Nidich SI, Goodman R, Goodman D. The effect of transcendental meditation on self-efficacy, perceived stress, and quality of life in mothers in Uganda. Health Care Women Int 2018 Jul;39(7):734-754 [FREE Full text] [doi: 10.1080/07399332.2018.1445254] [Medline: 29494787]
99. Goldstein L, Nidich SI, Goodman R, Goodman D. The effect of transcendental meditation on self-efficacy, perceived stress, and quality of life in mothers in Uganda. Health Care Women Int 2018 Jul;39(7):734-754 [FREE Full text] [doi: 10.1080/07399332.2018.1445254] [Medline: 29494787]
100. Wells RE, Burch R, Paulsen RH, Wayne PM, Houle TT, Loder E. Meditation for migraines: a pilot randomized controlled trial. Headache 2014 Oct;54(9):1484-1495. [doi: 10.1111/head.12420] [Medline: 25041058]
101. Peer E, Rothschild D, Gordon A, Evernden Z, Damer E. Data quality of platforms and panels for online behavioral research. Behav Res Methods 2022 Aug;54(4):1643-1662 [FREE Full text] [doi: 10.3758/s13428-021-01694-3] [Medline: 34590289]
102. Adams J. User and app related factors associated with engagement with mindfulness apps and health and wellbeing: a cross-sectional survey of US mindfulness app users. Open Science Framework (OSF). 2025 Mar 21. URL: https://osf.io/jcv5n/overview [accessed 2025-11-07]

Abbreviations

BFI-S-2: Big Five Inventory Short Form 2
DWAI-6: 6-item Digital Working Alliance Inventory
EQ-HWB-9: 9-item EuroQoL Health and Wellbeing
FDR: false discovery rate
K10: Kessler Psychological Distress Scale
MBP: mindfulness-based program
reCAPTCHA: reverse Completely Automated Public Turing Test to Tell Computers and Humans Apart
SWLS: Satisfaction with Life Survey
uMARS: user Mobile Application Rating Scale
WEMWBS: Warwick-Edinburgh Mental Wellbeing Scale

Edited by A Mavragani, T de Azevedo Cardoso; submitted 30.Jan.2025; peer-reviewed by SU Lam, G Cain; comments to author 03.Mar.2025; revised version received 29.Mar.2025; accepted 20.Aug.2025; published 02.Feb.2026.

Please cite as:
Adams J, Davies J, Wattanatakulchat P, Galante J, Miller F, D'Alfonso S, Van Dam NT
Engagement With Meditation Apps: Cross-Sectional Survey of Use and Associations
J Med Internet Res 2026;28:e71960
URL: https://www.jmir.org/2026/1/e71960
doi:10.2196/71960
PMID:41627883

©Julia Adams, Jonathan Davies, Prai Wattanatakulchat, Julieta Galante, Felicity Miller, Simon D'Alfonso, Nicholas T Van Dam. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 02.Feb.2026. This is an open-access

XSL•FO
RenderX

article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Automated Inference of Social Anxiety From Behavior in Social Virtual Reality: Cross-Sectional Observational Study

Gayoung Son[1], MSc; Marius Rubo[1], PhD

Cognitive Psychology, Perception and Research Methods, Institute of Psychology, University of Bern, Bern, Switzerland

Corresponding Author:
Gayoung Son, MSc
Cognitive Psychology, Perception and Research Methods
Institute of Psychology
University of Bern
Fabrikstrasse 8
Bern, 3012
Switzerland
Phone: 41 31 684 40 42
Email: gayoung.son@unibe.ch

Abstract

Background: Social anxiety often manifests through behaviors such as reduced gaze to the eyes and lower speech volume. While these markers have been examined in face-to-face interactions, large-scale assessments remain challenging. Social virtual reality (VR) offers a promising alternative by enabling naturalistic interactions in which behavior can be captured at scale, but it remains unclear if people show naturalistic behavior in such artificial environments.

Objective: We examined whether behavioral and physiological markers associated with social anxiety in real-life interactions similarly emerge in dyadic social VR. We additionally examined whether these patterns overlap with patterns linked to the broader constructs of psychopathology and verticality.

Methods: In this cross-sectional study, 128 participants (105 females, 22 males, 1 diverse; age 18-51 years; mean age 22.60, SD 3.57 years), recruited from a university student population, engaged in a 30-minute avatar-mediated dyadic conversation in social VR while physically located in separate rooms. We assessed gaze toward the partner's eyes, smiling, and speaking behavior by using the VR headsets' built-in eye trackers, face trackers, and microphones, and assessed high-frequency heart rate variability (HF-HRV) by using an electrocardiogram.

Results: Relationships between traits and behavioral and physiological measures were analyzed using linear mixed-effects models (α=.05). Social anxiety was linked to reduced gaze toward the partner's eyes while speaking (β=–.20, 95% CI –0.35 to –0.04; t_{126}=–2.51; P=.01), quieter speech (β=–.18, 95% CI –0.35 to –0.01; t_{126}=–2.12; P=.04), and reduced HF-HRV (β=–.23, 95% CI –0.39 to –0.08; t_{119}=–3.00; P=.003). These findings were not entirely specific to social anxiety, as Pearson correlations revealed similar patterns for social anxiety and psychopathology (r=0.94, 95% CI 0.75-0.99; t_7=7.61; P<.001), whereas verticality was linked to an overall opposite pattern (social anxiety: r=–0.92, 95% CI –0.98 to –0.65; t_7=–6.14; P<.001; psychopathology: r=–0.83, 95% CI –0.96 to –0.38; t_7=–3.98; P=.005).

Conclusions: In dyadic interactions in social VR, social anxiety was associated with behavioral and physiological modulations similar to those observed in face-to-face interactions, indicating heightened social stress and submissiveness even in avatar-mediated communication. Patterns were similar for heightened psychopathology and reversed for verticality, indicating that these traits may lie on a shared social-behavioral spectrum. Extending previous research focused on face-to-face interactions or reactions toward artificial agents displayed in VR, this study is the first to provide a comprehensive account of the behavioral and physiological modulations associated with social anxiety in avatar-based human-human interactions. Since social VR setups allow researchers to assess a rich set of behavioral data as a byproduct of the setup's core functionality, the technique opens novel possibilities to detect social stress, track therapeutic progress, or tailor interventions to individual behavior when interactions take place in VR.

(J Med Internet Res 2026;28:e79147) doi:10.2196/79147

XSL•FO
RenderX

KEYWORDS

eye gaze; interaction behavior; psychopathology; social anxiety; social virtual reality; speaking behavior; verticality

Introduction

Overview

Social anxiety is a psychological trait deeply rooted in social functioning, shaping how individuals perceive and behave in social situations, often manifesting in distinct behavioral and physiological patterns during social interactions [1]. It is widely distributed across the population, with higher levels linked to increased emotional distress and impaired social functioning. Individuals with severe expressions of social anxiety can meet the diagnostic criteria for social anxiety disorder, with a lifetime prevalence of 13% and a 12-month prevalence of 8% among adults [2]. Even moderate, subclinical levels are common and linked to elevated comorbidity and reduced quality of life [3]. Capturing how social anxiety is behaviorally expressed in social interactions is essential for understanding its impact and identifying individuals at risk.

Research has identified a set of robust behavioral and physiological markers associated with social anxiety [4], often conceptualized as safety behaviors—coping mechanisms aimed at preventing negative social outcomes [1,5]. These include gaze avoidance [6], which has been shown to increase during speaking phases in individuals with higher social anxiety [7,8]. However, not all studies have observed this phase-specific pattern [9,10], suggesting that gaze behavior may vary depending on situational factors and may reflect context-sensitive manifestations of underlying traits.

Changes in facial expression have also been observed. In face-to-face interactions, individuals with higher social anxiety have been found to smile more [11], in some cases specifically mimicking the partner's polite smiles [12], which has been suggested to reflect appeasement strategies aimed at avoiding negative evaluation. Overall, speaking behavior is also affected: social anxiety has been associated with the tendency to speak less [13], interrupt less often [14], speak more quietly, and pause longer between speaking turns [15], though these associations may be modulated by gender or context [16]. Together, these patterns suggest a tendency to limit social visibility and reduce opportunities for self-expression, possibly as a means of avoiding evaluation. In parallel, physiological studies consistently report elevated heart rate and reduced high-frequency heart rate variability (HF-HRV) in socially anxious individuals [17,18]—markers of reduced parasympathetic activity and emotional self-regulation [19,20].

While social anxiety has frequently been linked to specific behavioral markers, it is increasingly recognized that these expressions are not exclusive to the trait. Psychological traits often overlap, and patterns attributed to social anxiety may reflect broader constructs. As an internalizing condition, social anxiety shows behavioral similarities with other anxiety and mood syndromes (eg, increased gaze avoidance in depression [21]), but also with externalizing disorders (eg, reduced HF-HRV in substance use disorders [22]). These overlaps align with the notion of a general factor of psychopathology (*p*-factor), which reflects shared variance across mental disorders and has been proposed as a transdiagnostic dimension of psychopathology severity [23,24].

Similarly, verticality—a trait reflecting perceived dominance and social rank—has been linked to reduced social anxiety [25]. Individuals high in social anxiety have been shown to display more low-verticality behaviors such as gaze aversion and reduced speech [4,15,26,27], which are interpreted as submissive strategies for reducing ostracism [28-30]. In contrast, higher verticality is associated with dominant behaviors such as increased gaze, louder speech, and more frequent interruptions [31,32], suggesting an inverse relationship between these constructs.

Notably, verticality itself shows conceptual and empirical overlap with broader psychopathology, as internalizing profiles are associated with perceived inferiority [33], while externalizing traits align with dominance-seeking behavior [25]. To this end, this study assessed a range of psychological traits—including social anxiety, general psychopathology, and verticality—to examine both their shared behavioral expressions and trait-specific associations.

While these behavioral and physiological markers have been extensively researched, capturing them reliably in naturalistic settings remains a significant methodological challenge [34]. Traditional laboratory paradigms—such as eye-tracking studies using static images or videotaped face-to-face interactions—offer experimental control but have typically required restricted participant movement and relied on time-intensive manual coding [35]. To overcome these limitations, researchers have increasingly turned to digital phenotyping, which leverages passive data from smartphones to infer psychological traits [36]. For example, mobility and communication patterns have been shown to predict social anxiety symptom severity [37], while a recent systematic review found that smartphone sensor data can detect behavioral patterns linked to stress, anxiety, and mild depression in student populations [38].

While such findings demonstrate the growing use of mobile sensing, these approaches primarily rely on indirect proxies—such as phone use, physical activity, or ambient audio—that are often captured outside interactive contexts. Moreover, technical challenges such as inaccurate readings when phones are stowed, or poor audio quality in noisy environments, limit the reliability of such data [38].

Critically, smartphone sensing lacks access to the social interactive behavioral cues—such as gaze coordination, conversational timing, or facial expressivity—that are central to social anxiety in interpersonal contexts. For instance, reliably determining whether someone is making eye contact while speaking requires precise tracking of eye movements and speaking behavior of both partners—capabilities that remain difficult to scale during face-to-face interactions.

XSL•FO
RenderX

Social virtual reality (VR) directly addresses these limitations by combining immersive, ecologically valid interaction with high experimental control [39]. Participants engage through avatars that, in the most modern installments, mirror their gaze, facial expressions, gestures, and speech in real time, enabling naturalistic social behavior to be studied under replicable conditions [40,41]. Our previous publication based on the same dataset [41] documented natural interaction behavior at the group level (eg, more gaze to the eyes while listening compared to speaking; more gaze to the eyes at the end of a speaking turn compared to the beginning of a speaking turn; more smiling and interrupting in small-talk compared to more serious conversations; and preserved interaction patterns in anonymity). Building on that work, this study maps the links between interindividual differences in behavior, physiology, and psychological traits, providing detailed insights into expressions of clinically relevant personal characteristics.

Although recent studies have increasingly leveraged VR to examine ecologically valid expressions of behavior and physiological responses [42,43], the technology was traditionally used in contexts where participants interacted alone or with virtual agents [44]. Early work already demonstrated that interpersonal distance toward virtual agents with differential gaze behavior mirrored real-life proxemic behavior [45]. More recent research has begun to focus on human-human interaction within virtual environments [46], with studies investigating gaze avoidance [47], physiological synchrony [48], and facial expressions [49] in dyadic social VR settings, as well as speaking behavior and social coordination in larger group interactions [50]. Crucially, social VR enables the automated collection of multimodal behavioral data in ways comparable to established single-user techniques [51,52], highlighting its promise as a scalable platform for studying social interaction behavior. Our study contributes to this growing line of research by integrating behavioral and physiological measures within a single analytical framework to provide a more comprehensive account of real-time interpersonal dynamics in social VR.

Objective

We assessed whether behavioral alterations associated with social anxiety (but also related constructs such as general psychopathology and verticality) can be observed in interactions in social VR, which allows for naturalistic social interactions while enabling automated and scalable measurement of interaction behavior. To the best of our knowledge, our study is the first to simultaneously assess multiple social behaviors and physiological signals during real-time dyadic interactions in VR. Participants engaged in a dyadic conversation while embodying virtual avatars that displayed their eye movements, facial expressions, and gestures in real time. We specifically assessed gaze, speaking, and smiling behavior, which are processed automatically as part of the social VR environment, but also heart rate and HF-HRV, a robust physiological indicator of psychological stress and anxiety. Specifically, the hypotheses are as follows:

- Hypothesis 1: We expected different self-report measures of social anxiety, as well as different measures of psychopathology and verticality, to load on a common underlying factor.
- Hypothesis 2: We expected higher social anxiety (and similarly, higher psychopathology but lower verticality) to be associated with behavioral alterations:
  - Hypothesis 2.1: less gaze at the partner's eyes while speaking and listening
  - Hypothesis 2.2: heightened overall smiling behavior
  - Hypothesis 2.3: reduced overall speaking time, fewer interruptions during the partner's speech, longer pauses (ie, gaps before starting to speak after the partner finished their turn), and loudness of speech
- Hypothesis 3: We expected higher social anxiety (and similarly, higher psychopathology but lower verticality) to be associated with alterations in physiological functioning, specifically, heightened heart rate and reduced HF-HRV.

Methods

Participants

A total of 128 participants (105 female, 22 male, 1 diverse; aged 18-51 years; mean age 22.60, SD 3.57 years) participated in a dyadic interaction in the social VR environment. This predetermined sample size allows for an 85% chance of detecting a correlation of r=0.26 with α=.05 in a repeated-measures correlation involving 3 measures. Most participants (n=122, 95.31%) were undergraduate psychology students at the University of Bern, 118 with Swiss nationality, 8 with German nationality, and 14 from other European or non-European countries (multiple nationalities are possible). Exclusion criteria included neurological conditions, use of central nervous system–acting substances, and severe visual impairments requiring glasses that are incompatible with VR equipment. Within this sample, none of the individuals had more than a brief acquaintance with one another (n=5, 3.91% participants). To preserve conceptual clarity and avoid overlap between constructs, we kept measures distinct across factors. We included additional behavioral variables (ie, smiling and heart rate) and examined eye gaze separately during speaking and listening phases to offer a broader understanding of participant behavior. Data and analyses are accessible at [53].

Ethical Considerations

This study was approved by the institutional review board of the University of Bern (reference 2023-09-01) and was preregistered [54]. Prior to giving consent, all participants were verbally informed of the research aims, procedure, the voluntary nature of participation, and the right to withdraw at any point in the study freely and without consequences. They were also informed that they might experience symptoms of VR sickness and were instructed to remove the headset and inform the experimenter if this occurred; however, no such instances were reported. All collected data were anonymized and managed securely. Participants received 2 course credits as compensation. All figures containing persons depict example scenes performed by the study authors and do not include actual participants.

XSL•FO
RenderX

Procedure

This cross-sectional study followed the reporting guidelines for the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) Statement for observational studies [55]. Participants were recruited through the university's participant recruitment system between November 2023 and February 2024. Throughout the entire study (lasting approximately 1.5 hours), participants were seated in separate rooms and interacted solely through the social VR software (Figures 1 and 2), which was custom-developed in Unity 3D [56] for this study. Participants additionally arrived and departed separately in order to ensure that they did not see each other at any point. The virtual environment was displayed using Meta Quest Pro headsets [57] and ran on Windows 11 desktop laptops (Intel Core i9-13900HX CPU, 32 GB RAM, and NVIDIA GeForce RTX 4070 GPUs). Validation of study measures is reported in Son and Rubo [41]. A detailed description of the software architecture is provided in Rubo [58]. Subsequently, participants were equipped with surface electrodes to obtain an electrocardiogram throughout the entire study.

Figure 1. The experimental setup seen from a third-person perspective. Participants meet as avatars in virtual reality (top image) while physically located in separate rooms (bottom images). Avatars convey participants' movements, eye gaze, and facial expressions in real time, and voice is transmitted via voice chat. The virtual environment's position and orientation are calibrated so that the position of a virtual table matches the position of a physical table for both participants. All dyads conversed on predefined topics for a total of 30 minutes. People displayed in the images are the study authors demonstrating the experimental setup and procedure.



XSL•FO
RenderX

Figure 2. First-person perspective of a participant during the virtual reality interaction. Each participant views the scene from the perspective of their avatar's eyes, allowing for technology-mediated direct eye contact between participants.



Before the interaction, all participants completed an initial set of questionnaires, which included the self-reported measures of interest. Once the questionnaires were completed, participants were introduced to the virtual environment. This involved headset calibration, a 9-point eye-tracking calibration, and adjustments to the interpupillary distance. Participants also aligned their seating height and orientation to ensure that the virtual table was positioned to correspond precisely with the physical table. Next, participants observed their avatar in a virtual mirror, allowing them to see their avatar's full range of movements. Each avatar displayed the head and eye movement, blinking, mouth movement, facial expressions, and gestures of the participants in real time. Avatars were generated using Character Creator 4 along with the Headshot extension [59] and were configured to render only the head, upper torso, and hands. Each avatar featured extensive facial details and was individually adjusted by a research assistant for eye color, facial shape, hair, and body size (for more details on avatar creation, see Multimedia Appendix 1 [41,59-66]).

In addition, half of the dyads interacted while embodying self-similar avatars generated from participants' portrait photos, and the other half interacted while embodying gender-matching avatars created using generic images of nonexisting individuals from a random pool of 6 male and 6 female avatars. As detailed in a separate report [41], this condition investigated the effects of avatar appearance on social interactions. Since no behavioral differences were identified, and participants furthermore reported no differences in satisfaction with their avatars or in the attractiveness of their avatars, conditions were aggregated in this study.

The experimenter then refined the avatar's size, head orientation, and eye-tracking calibration based on participant feedback. Once the setup was complete, participants were instructed that the study would include three 10-minute conversations on predefined topics (small talk, personal talk, and opinion talk; Multimedia Appendix 1) presented in randomized order. Participants were asked to remain on topic for the duration of the task. The experimenter then exited the room, and participants were connected to their partners within the VR environment. During the interaction, participants were guided by prerecorded computerized voices, and each discussion prompt and the remaining time were displayed on a virtual tablet. After the interaction, they completed control questions on the quality of the interaction (eg, VR sickness symptoms or satisfaction with the avatar), which are described in a separate report [41]. Analyses reported here refer to interactions as a whole.

Self-Report Measures

Social Interaction Anxiety

Social interaction anxiety symptoms were examined using the Social Interaction Anxiety Scale (SIAS [67]; German version [68]). The SIAS assesses anxiety in social interactions in 20 items, rated on a 5-point Likert scale (0=not at all, 4=extremely). Internal consistency measured with Cronbach α was 0.91.

Fear of Positive and Negative Evaluation

Fear of positive and negative evaluation is a core symptom of social anxiety, consisting of fears of both negative and positive judgment from others [69]. Fear of positive evaluation was assessed using the German translation of the Fear of Positive Evaluation Scale [70,71], consisting of 10 items asking to what degree participants experienced positive evaluations as anxiety-inducing (α=.81). Items are rated on a 10-point Likert scale (0=not at all true, 9=very true). Fear of negative evaluation was assessed using the revised Brief Fear of Negative Evaluation Scale [72] (German version: [73]). Participants rate 12 items on a 5-point Likert scale (1=not at all characteristic of me, 5=extremely characteristic of me). Cronbach α was 0.94.

XSL•FO
RenderX

Social Comparison

Social Comparison was evaluated with the Social Comparison Scale [74] (German translation [75]). This scale measures upward social comparison evaluations across 11 different comparison dimensions. Participants respond to bipolar items on a scale from 1 to 10 (eg, In relationship to others I feel:" 1=incompetent, 10=more competent). Higher scores reflect a more favorable self-assessment relative to others, while lower scores reflect a less favorable self-assessment (α=.87).

Trait Anxiety

Trait Anxiety was assessed using the Trait subscale of the State-Trait-Anxiety Scale [76,77]. Participants rate 10 items ranging from 1=not at all to 4=extremely (α=.81).

Stress

Stress was measured using the Perceived Stress Scale-10 [78]. The 10 items assess the degree of stress experienced in the past month (1=never, 5=very often). Cronbach α was 0.81.

Depression Level

Depression level was assessed using the Beck Depression Inventory-Fast Screen [79] (German version [80]). Participants respond to 7 items assessing various depressive symptoms (sadness, pessimism, past failure, loss of pleasure, self-dislike, self-criticalness, and suicidal ideation). The scores vary between 0 and 21, where higher scores signify more severe depressive symptoms (α=.75).

Impulsivity

Impulsivity was assessed using the shortened Barratt Impulsiveness Scale-15 [81] (German version [82]) consisting of 15 items (α=.76). This scale measures 3 subtraits (motor, attentional, and nonplanning) on a 5-point Likert scale (1=never, 4=always).

Empathizing

Empathizing was measured with the German translation of the Empathy Quotient-10 [83]. This questionnaire is associated with autistic traits and also exhibits a broad range within the general population [83]. The 10 items are scored on a 5-point Likert scale (1=strongly disagree, 4=strongly agree). Cronbach α was 0.70.

Neuroticism and Extraversion

Neuroticism and extraversion were assessed using the respective subscales of the Big Five Inventory [84] (German version [85]). Neuroticism and extraversion are highly predictive of depressive and anxiety disorders, as previous meta-analytic research has shown that individuals in various diagnostic groups exhibit high levels of neuroticism and low levels of extraversion [86,87]. Both subscales consist of 4 items scored by a 5-point Likert scale (1=strongly disagree, 5=strongly agree). Cronbach α for neuroticism was 0.74, and for extraversion, 0.82.

Self-Esteem

Self-esteem was assessed with the Rosenberg Self-Esteem Scale [88] (German version [89]). Participants rate 10 items on a 6-point Likert scale (1=not at all, 6=very much). Cronbach α was 0.90.

Subjective Social Status

Subjective social status was measured using the MacArthur Subjective Social Status Scale [90] (German version [91]). Participants rate their socioeconomic position (income, education, and occupation) relative to their community on a scale of 1 to 10, visualized as a ladder with 10 rungs.

Power

Power refers to one's belief in having influence over others. It was assessed using the German translation of the Personal Sense of Power Scale [92,93], consisting of 6 items on a 7-point Likert scale (1=strongly disagree, 7=strongly agree). Cronbach α was 0.76.

Dominance and Prestige

Dominance and Prestige are 2 facets of the social rank attainment model, which posits that social hierarchy is achieved through force (dominance) or competence (prestige) [94]. These were assessed using the Dominance and Prestige Scale [95], which is a previously validated scale consisting of 2 subscales with 8 items measuring dominance (α=.77) and 9 items measuring prestige (α=.69), ranging from 1=not at all to 7=very much.

Agency

Agency was evaluated using the German adaptation of the Inventory of Interpersonal Problems-32 [96]. This tool comprises 32 items addressing 8 dimensions of interpersonal issues (α=.82), categorized into 2 overarching dimensions of agency and communion [97]. Participants indicate the extent to which each interpersonal challenge distresses them on a 5-point Likert scale (0=not at all, 4=extremely). For this study, the agency subscale score was used ([96]). This score specifically addresses interpersonal difficulties related to dominance, agency, and control [97].

Table 1 presents summaries of self-report measures. Responses in anxiety and depression measures showed that 27 (21.09%) participants exceeded the suggested diagnostic threshold of 34 for probable social anxiety [98] and 26 (20.31%) participants met the proposed cutoff score of 4 for mild depression [99].

XSL•FO
RenderX

Table 1. Descriptive statistics for self-report measures of participants completed before interaction, divided into measures for social anxiety, general psychopathology, and verticality (N=128).

Measure	Mean (SD; range)	Possible range[a]
Social anxiety		
Social interaction anxiety	25.20 (11.58; 3 to 53)	0 to 80
Fear of negative evaluation	35.45 (10.16; 15 to 58)	12 to 60
Fear of positive evaluation	22.69 (12.21; 0 to 54)	0 to 90
Social comparison	65.79 (11.45; 40 to 97)	40 to 100
General psychopathology		
Trait anxiety	20.32 (4.36; 11 to 33.5)	10 to 40
Stress	27.24 (4.99; 18 to 42)	10 to 50
Depression levels	2.20 (2.32; 0 to 9.4)	0 to 21
Impulsivity	2.19 (0.38; 1.33 to 3.2)	1 to 4
Empathizing	12.21 (3.13; 3 to 19)	0 to 20
Neuroticism	3.20 (0.81; 1.75 to 5)	1 to 5
Extraversion	3.50 (0.84; 1.5 to 5)	1 to 5
Self-esteem	4.67 (0.77; 2.28 to 5.9)	1 to 6
Verticality		
Social status	5.88 (1.22; 3 to 10)	1 to 10
Power	29.52 (4.27; 17 to 38)	6 to 42
Dominance	2.59 (0.87; 1 to 5)	1 to 7
Prestige	4.91 (0.65; 3 to 6.33)	1 to 7
Agency	–0.79 (0.50; –1.93 to 0.66)	–2.90 to 2.90

[a]Minimum and maximum scores that can be reached on each measure.

Behavioral Measures

We examined several behavioral and physiological markers during the interaction. For eye gaze, we followed conventional thresholds [100,101] and excluded eye-tracking data recorded during blinks as well as 100 ms before and after each blink to avoid blink-related artifacts (average data loss 3.85%, SD 2.81% across participants). Missing data were interpolated using the Piecewise Cubic Hermite Interpolating Polynomial and smoothed with a second-order Butterworth filter (cutoff frequency=10 Hz). Gaze to the partner's eye region was defined as visual attention directed toward the partner's eye region, operationalized as within 2 degrees of visual angle around the center of a partner's eye [102,103]. To account for common minor eyesight irregularities (eg, strabismus), gaze was coded to be directed toward a partner's eyes if at least one of the participant's eyes or a ray combining the 2 eyes was directed toward one of the partner's eyes. Following previous research on differential gaze behavior during conversational roles [104], gaze was analyzed separately for speaking and listening phases.

Speaking behavior was defined as continuous time periods during which the sound amplitude recorded by the head-mounted display's microphone exceeded a threshold of 2. Speech segments shorter than 1 second were classified as backchanneling vocalizations (eg, "ok" or "hmm") and not counted as speaking turns, consistent with prior research on nonverbal behavior [105,106]. Speaking turns of one person could include brief pauses of up to 2 seconds before being classified as 2 consecutive turns from the same person. The threshold was established based on Heldner and Edlund [107], indicating that typical within-speaker pauses last around 1 second, and was specifically selected to avoid classifying brief hesitations as new turns or fusing separate turns. Furthermore, moments when 1 participant remained silent while their partner spoke were categorized as listening [108]. Instances of overlapping speech were classified as interruptions, attributed to the participant who began speaking while previously listening [109].

Smiling was measured using the head-mounted display's built-in face tracking, which detects facial muscle activity based on action units defined by the Facial Action Coding System [110]. The intensity of smiling behavior was calculated by averaging the activation values from both sides of the mouth, yielding a score between 0 (no smile) and 100 (maximum smile expression).

Electrocardiograms were recorded using BItalino devices [111] at a sampling rate of 1000 Hz. QRS complexes (ventricular complex) were first detected automatically using an algorithm described by Elgendi et al [112]. Data were visually inspected individually and corrected manually if needed. Equidistantly sampled RR interval (duration between successive heartbeats)

XSL•FO
RenderX

series were constructed using cubic spline interpolation. Heart rate (in beats per minute) was defined as 60 divided by the mean of the RR interval series (in seconds). To obtain HF-HRV, RR interval time series were first resampled at 4 Hz using bandlimited interpolation. Power spectral density was estimated using the Welch method, with 30-second Hann-windowed segments and 50% overlap. Each segment was processed using the Fast Fourier Transform, and only positive frequencies were retained. HF-HRV was quantified by integrating the power spectral density within the high-frequency band (0.15-0.4 Hz) and log-transforming the resulting power values.

As reported in Son and Rubo [41], participants spent an average of 23.05% (SD 11.96%) of their speaking time and 49.45% (SD 16.96%) of their listening time gazing toward their partner's eyes. The mean proportion of speaking time was 50.73% (SD 9.90%). The mean proportion of time spent interrupting was 4.82% (SD 3.79%), and the average length of pauses before responding to the partner was 1.36 (SD 0.70) seconds. Average speech loudness was 2.70 (SD 0.77), measured relative to the maximum microphone recording capacity (100). The average smiling magnitude was 13.40 (SD 7.99) over the entire conversation period. The mean heart rate was 79.61 (SD 11.95) beats per minute, and the average HF-HRV was 6.36 (SD 0.96).

Data Analysis

Analyses were performed using R software (version 4.3.2; R Foundation for Statistical Computing) [113]. In the first step, principal component analysis was applied to self-report measures related to social anxiety, general psychopathology, and verticality (dominance and power), respectively. Principal component analysis is a widely used method to derive a common component from variables, which maximizes the total variance [60]. Focusing on the dominant components captures the most significant sources of variability in the data, providing a parsimonious representation of the underlying constructs and helping to avoid overly specific interpretations [23].

To further validate these components, two-sided Pearson correlations were conducted between the extracted components and the individual self-report measures that contributed to each factor. For these analyses, measures were aggregated across the entire conversation period. As expected, measures of social anxiety, psychopathology, and verticality were each correlated with their respective common factors (Multimedia Appendix 1). Consequently, subsequent analyses were carried out using these first dominant factors. The effect of each factor on behavioral and physiological measures was examined using mixed-effects models, with random intercepts for participant ID as well as random slopes for topics nested within ID to account for within-subject variability. For all variables, values beyond 3 SD from the mean were set to 3 SD from the mean and *z*-standardized prior to analyses. The percentage of corrected outliers ranged from 0.78% to 3.13% (Multimedia Appendix 1). The significance of results was determined with α set to 5%. *P* values were adjusted for multiple comparisons using the Benjamini-Hochberg false discovery rate (FDR) procedure.

To reduce potential bias, analyses were preregistered, data acquisition was standardized, and preprocessing algorithms were held identical for all participants. Furthermore, mixed-effects models were used to account for within-subject and sampling variability.

Results

Effect of Psychological Traits on Behavior

Principal component analyses were conducted separately for self-report measures related to social anxiety, general psychopathology, and verticality, which reflects dominance and rank. Each analysis yielded a dominant first component (Multimedia Appendix 1). All factors showed theoretically consistent and large associations with their corresponding measures, confirming their suitability as parsimonious representations of the underlying traits (Figure 3). Detailed results are provided in Multimedia Appendix 1.

The effect of social anxiety, general psychopathology, and verticality on the assessed behavioral measures was examined in linear mixed models. Figure 4 shows the β estimates and CIs for the relationships between factors and behaviors. A detailed summary of results is shown in Multimedia Appendix 1.

Figure 3. Pearson correlation coefficients between each first component of the principal component analysis (PCA) for social anxiety, general psychopathology, and verticality with self-report measures on individual scales. As expected, results show overall strong convergence between self-report measures relating to each of the 3 overarching trait constructs.

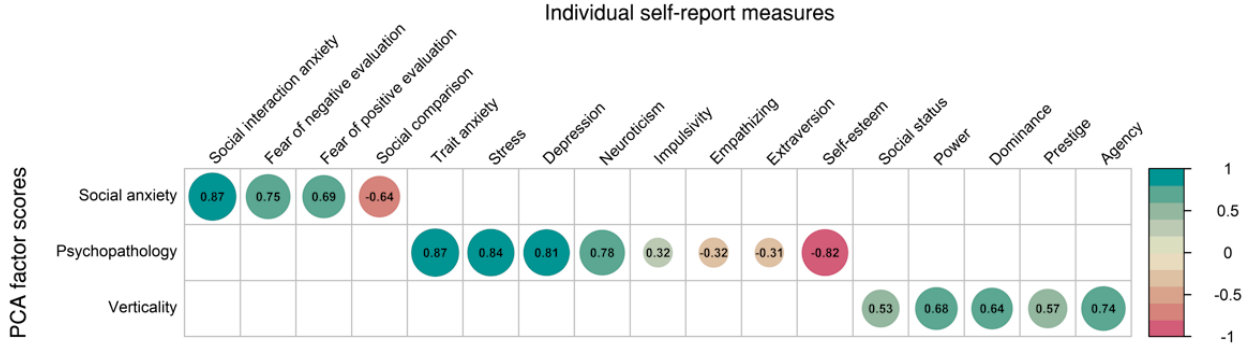

XSL•FO
RenderX

Figure 4. Associations between social anxiety, general psychopathology, and verticality with behavioral and physiological measures collected during the social virtual reality interaction. Values are standardized β estimates, indicating lower or higher measurements on behavioral or physiological metrics with higher trait expression. For example, higher social anxiety predicts reduced gaze to the partner's eyes while speaking. Lines represent 95% CIs. HF-HRV: high-frequency heart rate variability.

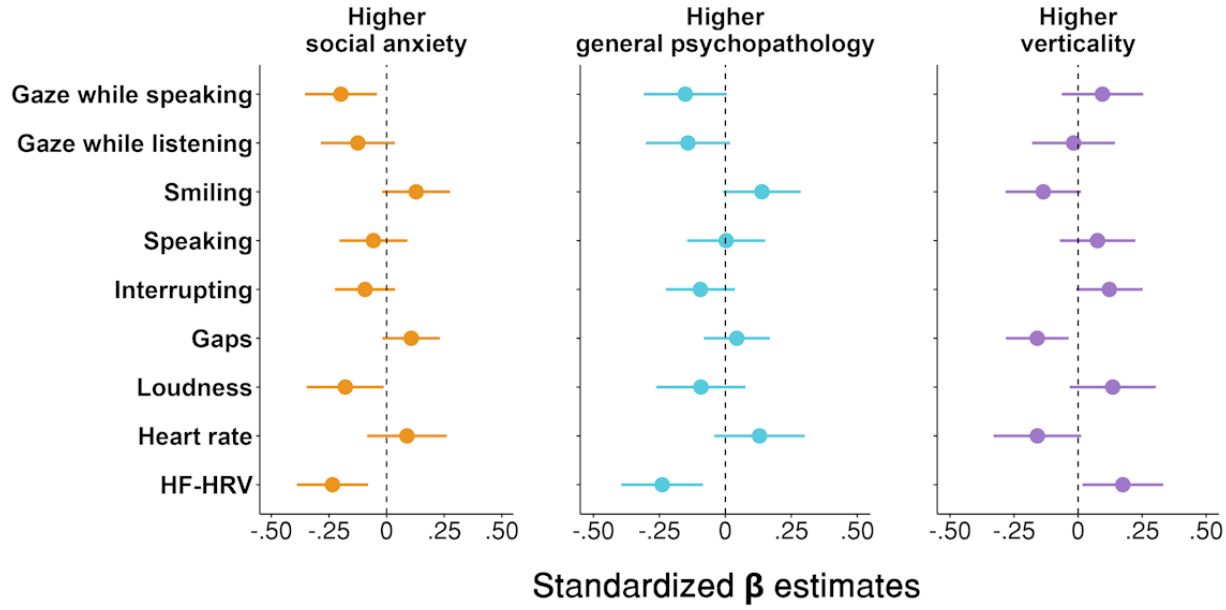

Higher social anxiety levels were associated with reduced gaze toward the partner's eyes during one's own speaking turn (β=–.20, 95% CI –0.35 to –0.04; t_{126}=–2.51; *P*=.01). The effect of underlying psychopathology on the amount of gaze toward the partner's eyes while speaking (β=–.15, 95% CI –0.31 to 0.01; t_{126}=–1.90; *P*=.06) and while listening was marginally significant (β=–.14, 95% CI –0.30 to 0.02; t_{126}=–1.75; *P*=.08). The association between social anxiety and psychopathology with smiling behavior was marginally significant, such that higher social anxiety (β=.13, 95% CI –0.02 to 0.28; t_{126}=1.72; *P*=.09) and higher psychopathology (β=.14, 95% CI –0.01 to 0.29; t_{126}=1.87; *P*=.06) was positively associated with smiling. In contrast, verticality was marginally significantly associated with less smiling behavior (β=–.14, 95% CI –0.28 to 0.01; t_{126}=–1.83; *P*=.07).

Social anxiety was furthermore associated with reduced speaking volume (β=–.18, 95% CI –0.35 to –0.01; t_{126}=–2.12; *P*=.04). The effect of social anxiety on gaps was marginally significant, such that higher social anxiety was associated with longer gaps between one's speaking turns (β=.11, 95% CI –0.02 to 0.23; t_{126}=1.70; *P*=.09). Verticality influenced gap duration, such that higher verticality was associated with shorter gaps before responding (β=–.16, 95% CI –0.28 to –0.04; t_{126}=–2.57; *P*=.01). Verticality was additionally marginally significantly associated with interrupting behavior, such that higher verticality was associated with more interrupting behavior (β=.12, 95% CI –0.03 to 0.25; t_{126}=1.86; *P*=.07). The effect of verticality on heart rate was marginally significant, such that individuals high in verticality displayed lower heart rate (β=–.16, 95% CI –0.33 to 0.01; t_{119}=–1.84; *P*=.07).

Social anxiety and psychopathology were associated with reduced HF-HRV (social anxiety: β=–.23, 95% CI –0.39 to –0.08; t_{119}=–3.00; *P*=.003; psychopathology: β=–.24, 95% CI –0.39 to –0.08; t_{119}=–3.07; *P*=.003), whereas higher verticality was associated with increased HF-HRV (β=.17, 95% CI 0.02 to 0.33; t_{119}=2.20; *P*=.03). After adjusting for multiple comparisons, the observed significant effects only held for social anxiety and psychopathology on HF-HRV.

Behavioral Similarities Between Social Anxiety, Psychopathology, and Verticality

We additionally compared similarities and differences in how social anxiety, psychopathology, and verticality related to behavior. Pearson correlations on the obtained β estimates were performed to examine the pattern of behavior between the 3 factors. The analysis revealed very strong correlations between the factors, such that the pattern of behavior in social anxiety was strongly positively associated with that in psychopathology (*r*=0.94, 95% CI 0.75 to 0.99; t_7=7.61; *P*<.001), and the pattern of behavior in verticality was highly negatively associated to that in social anxiety (*r*=–0.92, 95% CI –0.98 to –0.65; t_7=–6.14; *P*<.001) and psychopathology (*r*=–0.83, 95% CI –0.96 to –0.38; t_7=–3.98; *P*=.005).

Discussion

Principal Results

We examined whether previously reported associations between social anxiety and behavioral and physiological markers could be traced during real-time, dyadic interaction within social VR. This platform enables the automated assessment of fine-grained behavioral data during lifelike dyadic interactions, which has remained relatively difficult to assess in previous paradigms. We found that behavioral and physiological patterns previously associated with social anxiety also emerged in interactions in social VR, replicating patterns observed in face-to-face interactions. These associations were not exclusive to social

XSL•FO
RenderX

anxiety but were also found in relation to broader psychopathology and lowered verticality.

Findings suggested both shared and distinct behavioral patterns across psychological traits. Gaze avoidance was associated with social anxiety, replicating previous findings from face-to-face interaction studies [4,6] and from social VR studies involving agent-based interactions [114]. This effect was specific to speaking phases, consistent with prior work showing heightened gaze avoidance during speech, but not during listening [7,8]. While some studies have reported stronger associations with social anxiety during listening or found no clear phase-specific effects [9,10], these results support the view that gaze behavior in social anxiety may be particularly sensitive to moments of self-expression, when perceived social evaluation is likely heightened. In addition, reduced loudness was relatively specific to social anxiety, aligning with earlier research linking reduced loudness of speech to heightened nervousness in patients with social anxiety [15].

Although associations between social anxiety and behavior more consistently reached statistical significance compared to general psychopathology, the overall pattern of results was highly similar. Both gaze behavior during speaking and listening, as well as smiling, showed trends in the expected direction with higher general psychopathology. These findings suggest that certain interaction behaviors may also reflect broader psychological vulnerability in interpersonal settings. Furthermore, these findings may also reflect the heterogeneity of general psychopathology, which spans both internalizing and externalizing symptoms and may therefore give rise to less distinct patterns of social behavior. Future experimental studies that systematically manipulate the social-evaluative context could help determine whether gaze avoidance is a specific behavioral marker for social anxiety, or whether it generalizes to other dimensions of psychopathology [23].

Conversely, verticality, which encompasses constructs such as dominance and status, showed a partly complementary behavioral profile, marked by increased gaze during speaking (but not listening) and shorter turn-taking gaps—patterns previously linked to assertive or high-dominance behavior [31,32]. These results align with social rank theory [29], which posits that perceived social status regulates behavior in interpersonal contexts. Individuals higher in verticality may perceive less social threat and engage more confidently, reducing the need for avoidant or submissive strategies.

While these associations were theoretically coherent, most were small and did not remain significant after adjusting for multiple comparisons. Nonetheless, the pattern of results was consistent across measures and aligned with theoretical expectations. For instance, 6 of 36 bivariate associations were at least significant at $P<.05$ (expected by chance: 1.8), and 14 were at least marginally significant at $P<.10$ (expected by chance: 3.6). Importantly, all of these associations were in the expected direction, suggesting that the overall pattern aligns with hypothesized relationships.

Regarding physiological data, HF-HRV was predictive for all 3 traits, decreasing with higher social anxiety and general psychopathology, and increasing with verticality. As a physiological index of autonomic regulation and stress, reduced HF-HRV is broadly linked to impaired emotion regulation and social withdrawal [17,19,20,115]. In contrast to behavioral markers, this pattern suggests that HF-HRV may capture a general dimension of stress sensitivity, consistent with its proposed role as a transdiagnostic marker of stress [116,117].

Taken together, behavioral modulations in social anxiety and psychopathology were highly parallel, with verticality showing a largely inverted pattern. Both social anxiety and broader psychopathology were overall associated with an increase in interpersonal safety behaviors, particularly those expressed through submissive strategies such as reduced assertiveness and appeasing social cues [4,118]. These patterns are consistent with prior work linking these traits to diminished perceptions of social power [25,27]. Submissive behavior, in turn, has been consistently linked to reduced psychological well-being [119], reinforcing its clinical relevance as an important marker for heightened psychological vulnerability [27]. Conversely, verticality was associated with reduced submissive behavior, suggesting it may represent the opposing pole of a social-behavioral continuum relative to the submissive tendencies observed in social anxiety and broader psychopathology.

In addition, the strong behavioral convergence between social anxiety and general psychopathology lends support to the *p*-factor framework as a useful model for understanding shared variance in social behavior across psychological traits [23,24]. The behavioral overlap between social anxiety and general psychopathology indicates that the specificity of commonly used markers warrants further examination: rather than uniquely reflecting any single construct, interaction behavior may commonly indicate broader impairments in social functioning and well-being [120,121].

Implications

Our results highlight important implications. Compared to traditional eye-tracking paradigms [35] or smartphone phenotyping [38], our study demonstrates that social VR enables the synchronous collection of behavior during real-time interactions, enabling the study of social dynamics and mutual feedback that characterize human interactions. In social VR, participants can express and observe natural interaction behavior as it unfolds in a shared environment, creating conditions for more ecologically valid expressions of gaze, speech, and physiological responses [36]. This highlights this medium's potential as a powerful research tool for studying social processes and understanding psychological traits in socially embedded contexts.

Social VR presents promising opportunities for clinical research and novel forms of mental health care. Recent findings show that it is not only a growing digital medium, especially among younger users, but also a psychologically meaningful space for interactions. Supportive interactions in these environments have been linked to greater social presence, enhanced self-efficacy, and reduced loneliness [122]. As such, the technique may offer a valuable context for large-scale, ecologically valid monitoring of social behavior among multiple users.

XSL•FO
RenderX

Social VR also offers distinct advantages as a medium for delivering remote mental health care. Unlike conventional videoconferencing, it preserves key nonverbal cues such as eye contact and conversational timing, helping to overcome limitations such as reduced social presence and Zoom fatigue [123]. Moreover, it may enable novel therapeutic formats that are difficult to implement in traditional settings, such as fully anonymous group therapy sessions, which may lower barriers for individuals with social anxiety or those who fear stigmatization.

In addition, its capacity for automated data capture allows researchers to conduct passive, high-frequency assessments, which may be extended toward repeated interactions—such as multiple psychotherapy sessions or ongoing team meetings in virtual environments—facilitating the continuous monitoring of changes in well-being. Similar to passive smartphone phenotyping, which has demonstrated that passive sensor data can predict symptom changes in mental disorders [124], social VR offers new opportunities to assess social functioning in ecologically valid, interactive contexts. For example, fine-grained behavioral data collected in VR may aid in early risk detection, provide objective measures of therapeutic progress, or support the tailoring of interventions to individual behavioral patterns. Such data may be used to inform clinical decisions or even provide feedback within the VR setting itself [125].

At the same time, the technique's capability for unobtrusive, fine-grained data collection may also introduce significant security and ethical challenges. The passive and automated nature of data collection in these environments, which can be carried out by a server as well as any other computer in the network simulation, increases the risk of sensitive behavioral data being collected without the user's knowledge. While data collection in this study was conducted using an in-house software tool with no reliance on external services or internet connection, large-scale implementations of social VR in research contexts and beyond will be required to carefully consider data security principles and ethical implications of their use [39,126].

Limitations

We mapped a range of behavioral and physiological markers to interindividual differences in self-reported psychological traits in a naturalistic social VR interaction in 128 participants. Despite their strengths, 6 factors limit the generalizability of the findings.

First, the sample was composed of students from the German-speaking region of Switzerland, which may have introduced selection bias and limited the representativeness of the findings. Previous research indicates that social behaviors can vary across the lifespan; for example, avoidant behaviors such as gaze aversion tend to increase during adolescence [127]. Cultural factors may likewise influence nonverbal behavior and have been linked to differences in satisfaction and even health outcomes in cross-cultural clinical contexts [128]. As many aspects of nonverbal communication are culturally learned, the generalizability of our findings may decrease with greater cultural distance from the studied population [129]. At the same time, it was observed that interindividual differences in traits such as social anxiety often modulate modes of experience and behavior across population subgroups [130].

Second, although the sample showed substantial variability across behavioral and trait measures—with approximately 1 in 5 participants meeting thresholds for social anxiety disorder—the study lacked sufficient statistical power to examine gender differences in behavior, as the sample consisted predominantly of female participants. Exploratory analyses from a separate report [41] indicated that women smiled more frequently than men, consistent with prior findings in nonverbal behavior research [131]. However, the unequal group sizes in this study limit the interpretability of gender effects. Future research should therefore aim to systematically examine how demographic and cultural factors moderate behavioral dynamics in VR interactions to achieve more generalizability.

Third, the prevalence of mental disorders among university students is comparable to that in the general population [132], and psychological traits examined here—such as general psychopathology—are increasingly understood as dimensional rather than categorical [133]. Although some participants exceeded clinical thresholds for anxiety or depression, a clinically diagnosed sample was not included. It remains uncertain whether the behavioral modulations observed in this nonclinical sample would extend to full-blown mental disorders. In particular, stronger expressions of psychopathology may be associated with more specific modulations of interaction behavior than those observed here [134]. However, prior research has also shown that subclinical social anxiety typically exhibits behavioral patterns similar to those in social anxiety disorder, such as heightened anxiety in facial expressions [135]. Likewise, physiological indices such as heart rate have been shown to distinguish subclinical social anxiety during speech tasks in young adults [136]. These findings support the notion of social anxiety as a continuum, suggesting that the patterns observed here may represent expressions along the same underlying spectrum.

Fourth, while we included a broad range of self-report measures covering internalizing symptoms (eg, anxiety or depression) and select externalizing symptoms (eg, impulsivity or empathizing), other common externalizing symptoms such as substance abuse and aggressive behaviors were not assessed, as these may be less commonly observed in a student sample. This omission may have limited the scope of the general psychopathology construct captured. However, internalizing and externalizing psychopathology have been shown to be highly correlated across various age groups [137,138], and prior research has successfully modeled general psychopathology without including externalizing dimensions [139]. Nonetheless, future research should incorporate a broader range of symptom domains to more comprehensively assess the influence of psychopathology on social behavior.

Fifth, the study adopted a correlational design without a comparison condition in which participants interacted outside of VR. It therefore remains uncertain whether absolute behavioral levels in VR fully align with those observed in real-world interactions. A direct in vivo-in virtuo comparison would provide a stronger test of behavioral equivalence.

XSL•FO
RenderX

However, our previous work using the same dataset [41] compared key interaction metrics to established face-to-face benchmarks and found highly similar patterns, consistent with other recent findings comparing VR to real-life interactions [40], which suggests that fundamental aspects of social behavior are preserved in this social VR setting.

Lastly, although the presence of cohesion between stable traits and observed behavior implies some level of stability in behavior, our design did not allow for directly testing within-person consistency across interactions with varying partners. In order to better characterize how interpersonal processes contribute to individual variation in behavior [140], future studies should assess behavior in individuals across various interactions.

Conclusion

We investigated behavioral and physiological correlates of social anxiety as well as broader psychopathology and verticality in dyadic interactions in social VR. Extending previous work that examined interpersonal dynamics in face-to-face settings or responses to virtual agents, this study provides a comprehensive account of behavioral and physiological modulations associated with social anxiety in real-time avatar-based human-human interactions. In an integrated analytical framework, we assessed gaze, speaking, smiling expression, as well as heart rate and HF-HRV, and their associations with individual traits. As observed in natural face-to-face interactions, social anxiety levels were associated with behavioral modulations such as quieter speaking and reduced eye gaze to a partner while speaking and with reduced HF-HRV, indicating increased social stress and submissiveness. Behavioral and physiological patterns were not exclusive to social anxiety but appeared to reflect broader psychological characteristics, including higher general psychopathology and lower verticality. These results indicate that social interaction patterns may reflect general psychological vulnerability rather than effects specific to social anxiety levels. Behavioral traces that can be obtained automatically in social VR carry sensitive psychological information and, problematically, may be used for unintended profiling or psychological inference. At the same time, these data also provide valuable opportunities to assess psychologically relevant behavioral patterns in social interactions at scale in various settings, ultimately helping to understand pathological mechanisms in social interaction behavior and improving psychotherapy.

Acknowledgments

We thank Luca Panico, Susanne Felder, Michelle Bürki, Ana Rita Calçada da Silva, and Michael Mozhacherry for their help in data collection. Generative artificial intelligence was used in proofreading this manuscript.

Funding

This research was supported by grants from the Swiss National Science Foundation (grant PZ00P1_208909). The funding agencies were not involved in decisions concerning our study design, data collection, analysis, interpretation of data, writing of the report, or the submission of the article for publication.

Data Availability

This study's sample size, measures, and analysis approach were preregistered [54]. The study conformed to the principles expressed in the Declaration of Helsinki and was approved by the local ethics committee at the University of Bern (reference number 2023-09-01). Data that support the findings of this study are available at [53].

Authors' Contributions

GS contributed to conceptualization, project administration, formal analysis, visualization, and writing of the original draft, as well as review and editing. MR contributed to conceptualization, funding acquisition, project administration, formal analysis, software, supervision, and writing, review, and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Results with detailed explanations.
[DOCX File , 267 KB - jmir_v28i1e79147_app1.docx]

References

1. Heimberg RG, Brozovich FA, Rapee RM. A cognitive-behavioral model of social anxiety disorder. In: Social Anxiety. Amsterdam: Elsevier; 2014:705-728.

XSL•FO
RenderX

2. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005;62(6):593-602. [doi: 10.1001/archpsyc.62.6.593] [Medline: 15939837]
3. Fehm L, Beesdo K, Jacobi F, Fiedler A. Social anxiety disorder above and below the diagnostic threshold: prevalence, comorbidity and impairment in the general population. Soc Psychiatry Psychiatr Epidemiol 2008;43(4):257-265. [doi: 10.1007/s00127-007-0299-4] [Medline: 18084686]
4. Piccirillo ML, Taylor Dryman M, Heimberg RG. Safety behaviors in adults with social anxiety: review and future directions. Behav Ther 2016;47(5):675-687. [doi: 10.1016/j.beth.2015.11.005] [Medline: 27816080]
5. Clark D, Wells A. In: Heineberg RG, Liebowitz MR, Hope DA, Schneier FR, editors. A Cognitive Model of Social Phobia in Social Phobia: Diagnosis, Assessment, and Treatment. New York: Guilford Press; 1995.
6. Konovalova I, Antolin JV, Bolderston H, Gregory NJ. Adults with higher social anxiety show avoidant gaze behaviour in a real-world social setting: a mobile eye tracking study. PLoS One 2021;16(10):e0259007 [FREE Full text] [doi: 10.1371/journal.pone.0259007] [Medline: 34695140]
7. Baker SR, Edelmann RJ. Is social phobia related to lack of social skills? Duration of skill-related behaviours and ratings of behavioural adequacy. Br J Clin Psychol 2002;41(Pt 3):243-257. [doi: 10.1348/014466502760379118] [Medline: 12396253]
8. Chen J, van den Bos E, Westenberg PM. Does gaze anxiety predict actual gaze avoidance and is it more informative than social anxiety? J Behav Ther Exp Psychiatry 2024;82:101896. [doi: 10.1016/j.jbtep.2023.101896] [Medline: 37741178]
9. Chen J, van den Bos E, Karch JD, Westenberg PM. Social anxiety is related to reduced face gaze during a naturalistic social interaction. Anxiety Stress Coping 2023;36(4):460-474 [FREE Full text] [doi: 10.1080/10615806.2022.2125961] [Medline: 36153759]
10. Tönsing D, Schiller B, Vehlen A, Spenthof I, Domes G, Heinrichs M. No evidence that gaze anxiety predicts gaze avoidance behavior during face-to-face social interaction. Sci Rep 2022;12(1):21332 [FREE Full text] [doi: 10.1038/s41598-022-25189-z] [Medline: 36494411]
11. Shechter T, Asher M, Aderka IM. Man vs. machine: a comparison of human and computer assessment of nonverbal behavior in social anxiety disorder. J Anxiety Disord 2022;89:102587. [doi: 10.1016/j.janxdis.2022.102587] [Medline: 35661576]
12. Dijk C, Fischer AH, Morina N, van Eeuwijk C, van Kleef GA. Effects of social anxiety on emotional mimicry and contagion: feeling negative, but smiling politely. J Nonverbal Behav 2018;42(1):81-99 [FREE Full text] [doi: 10.1007/s10919-017-0266-z] [Medline: 29497222]
13. Gorlin EI, Teachman BA. Inhibitory control as a moderator of threat-related interference biases in social anxiety. Cogn Emot 2015;29(4):723-735 [FREE Full text] [doi: 10.1080/02699931.2014.931275] [Medline: 24967719]
14. Natale M, Entin E, Jaffe J. Vocal interruptions in dyadic communication as a function of speech and social anxiety. J Pers Soc Psychol 1979;37(6):865-878. [doi: 10.1037//0022-3514.37.6.865] [Medline: 490308]
15. Laukka P, Linnman C, Åhs F, Pissiota A, Frans Ö, Faria V, et al. In a nervous voice: acoustic analysis and perception of anxiety in social phobics' speech. J Nonverbal Behav 2008;32(4):195-214. [doi: 10.1007/s10919-008-0055-9]
16. Galili L, Amir O, Gilboa-Schechtman E. Acoustic properties of dominance and request utterances in social anxiety. J Soc Clin Psychol 2013;32(6):651-673. [doi: 10.1521/jscp.2013.32.6.651]
17. Cheng Y, Su M, Liu C, Huang Y, Huang W. Heart rate variability in patients with anxiety disorders: a systematic review and meta-analysis. Psychiatry Clin Neurosci 2022;76(7):292-302 [FREE Full text] [doi: 10.1111/pcn.13356] [Medline: 35340102]
18. Rösler L, Göhring S, Strunz M, Gamer M. Social anxiety is associated with heart rate but not gaze behavior in a real social interaction. J Behav Ther Exp Psychiatry 2021;70:101600 [FREE Full text] [doi: 10.1016/j.jbtep.2020.101600] [Medline: 32882674]
19. Kemp AH, Quintana DS, Gray MA, Felmingham KL, Brown K, Gatt JM. Impact of depression and antidepressant treatment on heart rate variability: a review and meta-analysis. Biol Psychiatry 2010;67(11):1067-1074. [doi: 10.1016/j.biopsych.2009.12.012] [Medline: 20138254]
20. Alvares GA, Quintana DS, Kemp AH, Van Zwieten A, Balleine BW, Hickie IB, et al. Reduced heart rate variability in social anxiety disorder: associations with gender and symptom severity. PLoS One 2013;8(7):e70468 [FREE Full text] [doi: 10.1371/journal.pone.0070468] [Medline: 23936207]
21. Suslow T, Hoepfel D, Kersting A, Bodenschatz CM. Depressive symptoms and visual attention to others' eyes in healthy individuals. BMC Psychiatry 2024;24(1):184. [doi: 10.1186/s12888-024-05633-2]
22. Quintana DS, McGregor IS, Guastella AJ, Malhi GS, Kemp AH. A meta-analysis on the impact of alcohol dependence on short-term resting-state heart rate variability: implications for cardiovascular risk. Alcohol Clin Exp Res 2013;37(Suppl 1):E23-E29. [doi: 10.1111/j.1530-0277.2012.01913.x] [Medline: 22834996]
23. Caspi A, Houts RM, Belsky DW, Goldman-Mellor SJ, Harrington H, Israel S, et al. The p factor: one general psychopathology factor in the structure of psychiatric disorders? Clin Psychol Sci 2014;2(2):119-137 [FREE Full text] [doi: 10.1177/2167702613497473] [Medline: 25360393]

XSL•FO
RenderX

24. Kotov R, Krueger RF, Watson D, Cicero DC, Conway CC, DeYoung CG, et al. The hierarchical taxonomy of psychopathology (HiTOP): a quantitative nosology based on consensus of evidence. Annu Rev Clin Psychol 2021;17(1):83-108. [doi: 10.1146/annurev-clinpsy-081219-093304] [Medline: 33577350]
25. Tharp JA, Johnson SL, Dev A. Transdiagnostic approach to the dominance behavioral system. Pers Individ Dif 2021;176:110778 [FREE Full text] [doi: 10.1016/j.paid.2021.110778] [Medline: 33746322]
26. Chen J, van den Bos E, Westenberg PM. A systematic review of visual avoidance of faces in socially anxious individuals: influence of severity, type of social situation, and development. J Anxiety Disord 2020;70:102193. [doi: 10.1016/j.janxdis.2020.102193] [Medline: 32058889]
27. Patel TA, Cougle JR. Submissive behavior as a transdiagnostic construct: associations with internalizing symptoms and distress intolerance. J Clin Psychol 2023;79(5):1314-1327. [doi: 10.1002/jclp.23478] [Medline: 36603203]
28. Gilboa-Schechtman E, Shachar I, Helpman L. Evolutionary perspective on social anxiety. In: Social Anxiety. Amsterdam: Elsevier; 2014:599-622.
29. Gilbert P. Evolution and social anxiety. The role of attraction, social competition, and social hierarchies. Psychiatr Clin North Am 2001;24(4):723-751. [doi: 10.1016/s0193-953x(05)70260-4] [Medline: 11723630]
30. Swee MB, Hudson CC, Heimberg RG. Examining the relationship between shame and social anxiety disorder: a systematic review. Clin Psychol Rev 2021;90:102088. [doi: 10.1016/j.cpr.2021.102088] [Medline: 34598054]
31. Carney DR. The nonverbal expression of power, status, and dominance. Curr Opin Psychol 2020;33:256-264. [doi: 10.1016/j.copsyc.2019.12.004] [Medline: 32361679]
32. Hall JA, Coats EJ, LeBeau LS. Nonverbal behavior and the vertical dimension of social relations: a meta-analysis. Psychol Bull 2005;131(6):898-924. [doi: 10.1037/0033-2909.131.6.898] [Medline: 16351328]
33. Johnson SL, Leedom LJ, Muhtadie L. The dominance behavioral system and psychopathology: evidence from self-report, observational, and biological studies. Psychol Bull 2012;138(4):692-743 [FREE Full text] [doi: 10.1037/a0027503] [Medline: 22506751]
34. Hadley LV, Naylor G, Hamilton AFDC. A review of theories and methods in the science of face-to-face social interaction. Nat Rev Psychol 2022;1(1):42-54. [doi: 10.1038/s44159-021-00008-w]
35. Risko EF, Laidlaw K, Freeth M, Foulsham T, Kingstone A. Social attention with real versus reel stimuli: toward an empirical approach to concerns about ecological validity. Front Hum Neurosci 2012;6:143 [FREE Full text] [doi: 10.3389/fnhum.2012.00143] [Medline: 22654747]
36. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021;20(3):318-335 [FREE Full text] [doi: 10.1002/wps.20883] [Medline: 34505369]
37. Jacobson NC, Summers B, Wilhelm S. Digital biomarkers of social anxiety severity: digital phenotyping using passive smartphone sensors. J Med Internet Res 2020;22(5):e16875 [FREE Full text] [doi: 10.2196/16875] [Medline: 32348284]
38. Choi A, Ooi A, Lottridge D. Digital phenotyping for stress, anxiety, and mild depression: systematic literature review. JMIR Mhealth Uhealth 2024;12:e40689 [FREE Full text] [doi: 10.2196/40689] [Medline: 38780995]
39. Parsons TD. Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. Front Hum Neurosci 2015;9:660 [FREE Full text] [doi: 10.3389/fnhum.2015.00660] [Medline: 26696869]
40. Abdullah A, Kolkmeier J, Lo V, Neff M. Videoconference and embodied VR: communication patterns across task and medium. Proc ACM Hum-Comput Interact 2021;5:1-29. [doi: 10.1145/3479597]
41. Son G, Rubo M. Social virtual reality elicits natural interaction behavior with self-similar and generic avatars. Int J Hum Comput Stud 2025;199:103488. [doi: 10.1016/j.ijhcs.2025.103488]
42. Girondini M, Frigione I, Marra M, Stefanova M, Pillan M, Maravita A, et al. Decoupling the role of verbal and non-verbal audience behavior on public speaking anxiety in virtual reality using behavioral and psychological measures. Front Virtual Real 2024;5. [doi: 10.3389/frvir.2024.1347102]
43. Torres SC, Gracia Laso DI, Minissi ME, Maddalon L, Chicchi Giglioli IA, Alcañiz M. Social signal processing in affective virtual reality: human-shaped agents increase electrodermal activity in an elicited negative environment. Cyberpsychol Behav Soc Netw 2024;27(4):268-274 [FREE Full text] [doi: 10.1089/cyber.2023.0273] [Medline: 38394167]
44. Pan X, Hamilton AFDC. Why and how to use virtual reality to study human social interaction: the challenges of exploring a new research landscape. Br J Psychol 2018;109(3):395-417 [FREE Full text] [doi: 10.1111/bjop.12290] [Medline: 29504117]
45. Bailenson JN, Blascovich J, Beall AC, Loomis JM. Interpersonal distance in immersive virtual environments. Pers Soc Psychol Bull 2003;29(7):819-833. [doi: 10.1177/0146167203029007002] [Medline: 15018671]
46. Belina SP, Czaja SJ, Difede J, Pain KJ, Reid MC, Won AS. In sickness and in health-a scoping review on social virtual reality clinical applications. Digit Health 2024;10:20552076241306460 [FREE Full text] [doi: 10.1177/20552076241306460] [Medline: 39741985]
47. Li C, Dai Y, Chen G, Liu J, Li P, Ip HH. Avatar-mediated communication in collaborative virtual environments: a study on users' attention allocation and perception of social interactions. Comput Hum Behav 2025;167:108598. [doi: 10.1016/j.chb.2025.108598]

XSL•FO
RenderX

48. Moharana B, Keighrey C, Murray N. Physiological synchrony in a collaborative virtual reality task. : IEEE; 2023 Presented at: 15th International Conference on Quality of Multimedia Experience (QoMEX); 2023 June 20; Ghent, Belgium p. 288-293. [doi: 10.1109/qomex58391.2023.10178661]
49. Kimmel S, Jung F, Matviienko A, Heuten W, Boll S. Let's face it: influence of facial expressions on social presence in collaborative virtual reality. 2023 Presented at: Conference on Human Factors in Computing Systems - Proceedings; 2016 Apr 19; New York, NY p. 1-16. [doi: 10.1145/3544548.3580707]
50. Léchappé A, Johnstone R, Milliat A, Williamson JH, Chollet M, Williamson JR. Understanding social interactions in reality versus virtuality. : ACM; 2025 Presented at: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. CHI '25; 2025 Apr 25; New York, NY p. 1-18. [doi: 10.1145/3706598.3713628]
51. Xenakis I, Gavalas D, Kasapakis V, Dzardanova E, Vosinakis S. Nonverbal communication in immersive virtual reality through the lens of presence: a critical review. Presence 2022;31:147-187. [doi: 10.1162/pres_a_00387]
52. Yaremych HE, Persky S. Tracing physical behavior in virtual reality: a narrative review of applications to social psychology. J Exp Soc Psychol 2019;85:103845 [FREE Full text] [doi: 10.1016/j.jesp.2019.103845] [Medline: 32831397]
53. Son G, Rubo M. Data and analyses. OSF. URL: https://osf.io/9yr8p [accessed 2026-02-23]
54. Son G, Rubo M. Preregistration. OSF. URL: https://osf.io/kawtn [accessed 2026-02-23]
55. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007;370(9596):1453-1457 [FREE Full text] [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
56. Unity Engine. Unity. URL: https://unity.com/ [accessed 2026-02-23]
57. Meta Quest Pro headsets. Meta. URL: https://www.meta.com/ [accessed 2026-02-23]
58. Rubo M. Creating a social virtual reality application for psychological research: A tutorial. Behav Res Methods 2025 Jun 04;57(7):192 [FREE Full text] [doi: 10.3758/s13428-025-02693-4] [Medline: 40468093]
59. Reallusion. URL: https://www.reallusion.com/ [accessed 2026-02-23]
60. Schreiber JB. Issues and recommendations for exploratory factor analysis and principal component analysis. Res Social Adm Pharm 2021;17(5):1004-1011. [doi: 10.1016/j.sapharm.2020.07.027] [Medline: 33162380]
61. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988;83(404):1198-1202. [doi: 10.1080/01621459.1988.10478722]
62. Aron A, Melinat E, Aron EN, Vallone RD, Bator RJ. The experimental generation of interpersonal closeness: a procedure and some preliminary findings. Pers Soc Psychol Bull 1997;23(4):363-377. [doi: 10.1177/0146167297234003]
63. Matz DC, Wood W. Cognitive dissonance in groups: the consequences of disagreement. J Pers Soc Psychol 2005;88(1):22-37. [doi: 10.1037/0022-3514.88.1.22] [Medline: 15631572]
64. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. IEEE Trans Pattern Anal Mach Intell 2021;43(12):4217-4228. [doi: 10.1109/TPAMI.2020.2970919] [Medline: 32012000]
65. Rubo M. AvatarTransferPro: a tool for automatic transfer of realistic personalized avatars in networked social virtual reality. SoftwareX 2024;27:101873. [doi: 10.1016/j.softx.2024.101873]
66. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 1995;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
67. Mattick RP, Clarke JC. Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. Behav Res Ther 1998;36(4):455-470. [doi: 10.1016/s0005-7967(97)10031-6] [Medline: 9670605]
68. Stangier U, Heidenreich T, Berardi A, Golbs U, Hoyer J. Die erfassung sozialer phobie durch die Social Interaction Anxiety Scale (SIAS) und die Social Phobia Scale (SPS). Z Klin Psychol 1999;28(1):28-36 [FREE Full text] [doi: 10.1026//0084-5345.28.1.28]
69. Weeks JW, Howell AN. The bivalent fear of evaluation model of social anxiety: further integrating findings on fears of positive and negative evaluation. Cogn Behav Ther 2012;41(2):83-95. [doi: 10.1080/16506073.2012.661452] [Medline: 22489647]
70. Weeks JW, Heimberg RG, Rodebaugh TL. The Fear of Positive Evaluation Scale: assessing a proposed cognitive component of social anxiety. J Anxiety Disord 2008;22(1):44-55. [doi: 10.1016/j.janxdis.2007.08.002] [Medline: 17884328]
71. Schwarz M, Reichenberger J, Schwitalla M, König D, Wilhelm FH, Voderholzer U, et al. Furcht vor positiver sozialer bewertung: ubersetzung und validierung einer deutschsprachigen version der FPES. Diagnostica 2016;62(3):157-168. [doi: 10.1026/0012-1924/a000147]
72. Carleton RN, McCreary DR, Norton PJ, Asmundson GJG. Brief Fear of Negative Evaluation Scale-revised. Depress Anxiety 2006;23(5):297-303. [doi: 10.1002/da.20142] [Medline: 16688736]
73. Reichenberger J, Schwarz M, König D, Wilhelm FH, Voderholzer U, Hillert A, et al. Angst vor negativer sozialer bewertung: ubersetzung und validierung der furcht vor negativer evaluation–Kurzskala (FNE-K). Diagnostica 2016;62(3):169-181. [doi: 10.1026/0012-1924/a000148]
74. Allan S, Gilbert P. A social comparison scale: psychometric properties and relationship to psychopathology. Pers Individ Differ 1995;19(3):293-299. [doi: 10.1016/0191-8869(95)00086-l]
75. Ascone L, Jaya ES, Lincoln TM. The effect of unfavourable and favourable social comparisons on paranoid ideation: an experimental study. J Behav Ther Exp Psychiatry 2017;56:97-105. [doi: 10.1016/j.jbtep.2016.08.002] [Medline: 27567982]

XSL•FO
RenderX

76. Spielberger CD. State-Trait Anxiety Inventory for Adults. PsycTESTS Dataset 1983. [doi: 10.1037/t06496-000]
77. Grimm J. State-Trait-Anxiety Inventory nach Spielberger. Deutsche lang-und kurzversion. Methodenforum der Universität Wien 2009;2009(02) [FREE Full text]
78. Cohen S, Williamson GM. Perceived stress in a probability sample of the United States. In: Spacapan S, Oskamp S, editors. The Social Psychology of Health. California: Sage Publications; 1988:31-67.
79. Beck AT, Brown GK. BDI-II Fast Screen for Medical Patients Manual. London: The Psychological Corporation; 2000.
80. Kliem S, Mößle T, Zenger M, Brähler E. Reliability and validity of the Beck Depression Inventory-Fast Screen for Medical Patients in the general German population. J Affect Disord 2014;156:236-239. [doi: 10.1016/j.jad.2013.11.024] [Medline: 24480380]
81. Meule A, Vögele C, Kübler A. Psychometrische evaluation der deutschen Barratt Impulsiveness Scale – Kurzversion (BIS-15). Diagnostica 2011;57(3):126-133 [FREE Full text] [doi: 10.1026/0012-1924/a000042]
82. Spinella M. Normative data and a short form of the Barratt impulsiveness scale. Int J Neurosci 2007;117(3):359-368. [doi: 10.1080/00207450600588881] [Medline: 17365120]
83. Greenberg DM, Warrier V, Allison C, Baron-Cohen S. Testing the empathizing-systemizing theory of sex differences and the extreme male brain theory of autism in half a million people. Proc Natl Acad Sci USA 2018;115(48):12152-12157 [FREE Full text] [doi: 10.1073/pnas.1811032115] [Medline: 30420503]
84. John OP, Donahue EM, Kentle RL. Big Five Inventory. APA PsycTests 1991 [FREE Full text] [doi: 10.1037/t07550-000]
85. Rammstedt B, John OP. Kurzversion des Big Five Inventory (BFI-K). Diagnostica 2005;51(4):195-206. [doi: 10.1026/0012-1924.51.4.195]
86. Kotov R, Gamez W, Schmidt F, Watson D. Linking "big" personality traits to anxiety, depressive, and substance use disorders: a meta-analysis. Psychol Bull 2010;136(5):768-821. [doi: 10.1037/a0020327] [Medline: 20804236]
87. Hakulinen C, Elovainio M, Pulkki-Råback L, Virtanen M, Kivimäki M, Jokela M. Personality and depressive symptoms: individual participant meta-analysis of 10 cohort studies. Depress Anxiety 2015;32(7):461-470 [FREE Full text] [doi: 10.1002/da.22376] [Medline: 26014798]
88. Rosenberg M. Society and the Adolescent Self-Image. USA: Princeton University Press; 1965:326.
89. von Collani G, Herzberg PY. Eine revidierte fassung der deutschsprachigen Skala zum selbstwertgefühl von Rosenberg. Z Differ Diagn Psychol 2006;24(1):3-7. [doi: 10.1024//0170-1789.24.1.3]
90. Adler NE, Epel ES, Castellazzo G, Ickovics JR. Relationship of subjective and objective social status with psychological and physiological functioning: preliminary data in healthy, white women. Health Psychol 2000;19(6):586-592. [doi: 10.1037/0278-6133.19.6.586]
91. Hoebel J, Müters S, Kuntz B, Lange C, Lampert T. [Measuring subjective social status in health research with a German version of the MacArthur Scale]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2015;58(7):749-757. [doi: 10.1007/s00103-015-2166-x] [Medline: 25986532]
92. Anderson C, John OP, Keltner D. The personal sense of power. J Pers 2012;80(2):313-344. [doi: 10.1111/j.1467-6494.2011.00734.x] [Medline: 21446947]
93. Körner R, Heydasch T, Schütz A. It's all about power: validation of trait and state versions of the German personal sense of power scale. Eur J Psychol Assess 2022;38(1):36-48. [doi: 10.1027/1015-5759/a000642]
94. Cheng JT, Tracy JL, Foulsham T, Kingstone A, Henrich J. Two ways to the top: evidence that dominance and prestige are distinct yet viable avenues to social rank and influence. J Pers Soc Psychol 2013;104(1):103-125. [doi: 10.1037/a0030398] [Medline: 23163747]
95. Cheng JT, Tracy JL, Henrich J. Pride, personality, and the evolutionary foundations of human social status. Evol Hum Behav 2010;31(5):334-347. [doi: 10.1016/j.evolhumbehav.2010.02.004]
96. Thomas A, Brähler E, Strauß B. IIP-32: entwicklung, validierung und normierung einer kurzform des inventars zur erfassung interpersonaler probleme. Diagnostica 2011;57(2):68-83 [FREE Full text] [doi: 10.1026/0012-1924/a000034]
97. Barkham M, Hardy GE, Startup M. The IIP-32: a short version of the inventory of interpersonal problems. Br J Clin Psychol 1996;35(1):21-35. [doi: 10.1111/j.2044-8260.1996.tb01159.x] [Medline: 8673033]
98. Brown EJ, Turovsky J, Heimberg RG, Juster HR, Brown TA, Barlow DH. Validation of the Social Interaction Anxiety Scale and the Social Phobia Scale across the anxiety disorders. Psychol Assess 1997;9(1):21-27. [doi: 10.1037/1040-3590.9.1.21]
99. Alsaleh M, Videloup L, Lobbedez T, Lebreuilly J, Morello R, Thuillier Lecouf A. Improved detection and evaluation of depression in patients with chronic kidney disease: validity and reliability of screening (PHQ-2) and diagnostic (BDI-FS-Fr) tests of depression in chronic kidney disease. Kidney Dis (Basel) 2019;5(4):228-238 [FREE Full text] [doi: 10.1159/000497352] [Medline: 31768380]
100. Gee de JW, Knapen T, Donner TH. Decision-related pupil dilation reflects upcoming choice and individual bias. Proc Natl Acad Sci USA 2014;111(5):E618-E625 [FREE Full text] [doi: 10.1073/pnas.1317557111] [Medline: 24449874]
101. Geller J, Winn MB, Mahr T, Mirman D. GazeR: a package for processing gaze position and pupil size data. Behav Res Methods 2020;52(5):2232-2255 [FREE Full text] [doi: 10.3758/s13428-020-01374-8] [Medline: 32291732]

XSL•FO
RenderX

102. Aziz S, Lohr D, Friedman L, Komogortsev O. Evaluation of eye tracking signal quality for virtual reality applications: a case study in the Meta Quest Pro. : ACM; 2024 Presented at: Proceedings of the 2024 Symposium on Eye Tracking Research and Applications. ETRA '24; 2024 June 04; United States p. 1-8. [doi: 10.1145/3649902.3653347]
103. Wei S, Bloemers D, Rovira A. A preliminary study of the eye tracker in the Meta Quest Pro. : ACM; 2023 Presented at: Proceedings of the 2023 ACM International Conference on Interactive Media Experiences. IMX '23; 2023 June 12; USA p. 216-221. [doi: 10.1145/3573381.3596467]
104. Ho S, Foulsham T, Kingstone A. Speaking and listening with the eyes: gaze signaling during dyadic interactions. PLoS One 2015;10(8):e0136905 [FREE Full text] [doi: 10.1371/journal.pone.0136905] [Medline: 26309216]
105. Frauendorfer D, Schmid Mast M, Nguyen L, Gatica-Perez D. Nonverbal social sensing in action: unobtrusive recording and extracting of nonverbal behavior in social interactions illustrated with a research example. J Nonverbal Behav 2014;38(2):231-245. [doi: 10.1007/s10919-014-0173-5]
106. Lee JJ, Breazeal C, DeSteno D. Role of speaker cues in attention inference. Front Robot AI 2017;4. [doi: 10.3389/frobt.2017.00047]
107. Heldner M, Edlund J. Pauses, gaps and overlaps in conversations. J Phonet 2010;38(4):555-568. [doi: 10.1016/j.wocn.2010.08.002]
108. Skantze G. Turn-taking in conversational systems and human-robot interaction: a review. Comput Speech Lang 2021;67:101178. [doi: 10.1016/j.csl.2020.101178]
109. Holler J, Kendrick K, Casillas M, Levinson S. Turn-taking in human communicative interaction. Frontiers Media SA 2016. [doi: 10.3389/978-2-88919-825-2]
110. Face tracking for movement SDK for Unity. Meta. URL: https://developers.meta.com/horizon/documentation/unity/move-face-tracking/ [accessed 2026-02-23]
111. Plux Biosignals BITalino. Biosignals Plux. URL: https://www.pluxbiosignals.com/ [accessed 2026-02-23]
112. Elgendi M, Jonkman M, DeBoer F. Frequency bands effects on QRS detection. : SciTePress - Science, and Technology Publications; 2010 Presented at: Proceedings of the Third International Conference on Bio-Inspired Systems and Signal Processing; 2010 June 20; Valencia, Spain p. 428-431. [doi: 10.5220/0002742704280431]
113. R Core Team. R: A Language and Environment for Statistical Computing. Denmark: R Foundation for Statistical Computing; 2023.
114. Rubo M, Munsch S. Social stress in an interaction with artificial agents in virtual reality: effects of ostracism and underlying psychopathology. Comput Hum Behav 2024;153:107915. [doi: 10.1016/j.chb.2023.107915]
115. Chalmers JA, Quintana DS, Abbott MJ, Kemp AH. Anxiety disorders are associated with reduced heart rate variability: a meta-analysis. Front Psychiatry 2014;5:80 [FREE Full text] [doi: 10.3389/fpsyt.2014.00080] [Medline: 25071612]
116. Gidron Y, Deschepper R, De Couck M, Thayer JF, Velkeniers B. The vagus nerve can predict and possibly modulate non-communicable chronic diseases: introducing a neuroimmunological paradigm to public health. J Clin Med 2018;7(10):371 [FREE Full text] [doi: 10.3390/jcm7100371] [Medline: 30347734]
117. Tomasi J, Zai CC, Pouget JG, Tiwari AK, Kennedy JL. Heart rate variability: evaluating a potential biomarker of anxiety disorders. Psychophysiology 2024;61(2):e14481. [doi: 10.1111/psyp.14481] [Medline: 37990619]
118. Russell JJ, Moskowitz DS, Zuroff DC, Bleau P, Pinard G, Young SN. Anxiety, emotional security and the interpersonal behavior of individuals with social anxiety disorder. Psychol Med 2010;41(3):545-554. [doi: 10.1017/s0033291710000863]
119. Allan S, Gilbert P. Submissive behaviour and psychopathology. Br J Clin Psychol 1997;36(4):467-488. [doi: 10.1111/j.2044-8260.1997.tb01255.x] [Medline: 9403141]
120. Achterhof R, Kirtley OJ, Schneider M, Hagemann N, Hermans KS, Hiekkaranta AP, et al. General psychopathology and its social correlates in the daily lives of youth. J Affect Disord 2022;309:428-436. [doi: 10.1016/j.jad.2022.04.147] [Medline: 35500686]
121. Carver CS, Johnson SL, Timpano KR. Toward a functional view of the P factor in psychopathology. Clin Psychol Sci 2017;5(5):880-889. [doi: 10.1177/2167702617710037] [Medline: 29057170]
122. Oh HJ, Kim J, Chang JJ, Park N, Lee S. Social benefits of living in the metaverse: the relationships among social presence, supportive interaction, social self-efficacy, and feelings of loneliness. Comput Human Behav 2023;139:107498. [doi: 10.1016/j.chb.2022.107498]
123. Bailenson JN. Nonverbal overload: a theoretical argument for the causes of Zoom fatigue. Technol Mind Behav 2021;2(1):1-6. [doi: 10.1037/tmb0000030]
124. Bufano P, Laurino M, Said S, Tognetti A, Menicucci D. Digital phenotyping for monitoring mental disorders: systematic review. J Med Internet Res 2023;25:e46778 [FREE Full text] [doi: 10.2196/46778] [Medline: 38090800]
125. Bell IH, Nicholas J, Alvarez-Jimenez M, Thompson A, Valmaggia L. Virtual reality as a clinical tool in mental health research and practice . Dialogues Clin Neurosci 2020;22(2):169-177 [FREE Full text] [doi: 10.31887/DCNS.2020.22.2/lvalmaggia] [Medline: 32699517]
126. Giaretta A. Security and privacy in virtual reality: a literature survey. Virtual Real 2024;29(1):10. [doi: 10.1007/s10055-024-01079-9]
127. Halberstadt AG, Parker AE, Castro VL. Nonverbal Communication: Developmental Perspectives. Berlin, Germany: Nonverbal Communication. DE GRUYTER; 2013:93-128.

XSL•FO
RenderX

128. Levine CS, Ambady N. The role of non-verbal behaviour in racial disparities in health care: implications and solutions. Med Educ 2013;47(9):867-876. [doi: 10.1111/medu.12216] [Medline: 23931536]
129. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? Behav Brain Sci 2010;33(2-3):61-83. [doi: 10.1017/s0140525x0999152x]
130. Hofmann SG, Anu Asnaani MA, Hinton DE. Cultural aspects in social anxiety and social anxiety disorder. Depress Anxiety 2010;27(12):1117-1127 [FREE Full text] [doi: 10.1002/da.20759] [Medline: 21132847]
131. Lafrance M, Hecht M. Gender and smiling: a meta-analysis. In: Gender and Emotion. United Kingdom: Cambridge University Press; 2000:118-142.
132. Hunt J, Eisenberg D. Mental health problems and help-seeking behavior among college students. J Adolesc Health 2010;46(1):3-10. [doi: 10.1016/j.jadohealth.2009.08.008] [Medline: 20123251]
133. Caspi A, Houts RM, Fisher HL, Danese A, Moffitt TE. The general factor of psychopathology (p): choosing among competing models and interpreting p. Clin Psychol Sci 2024;12(1):53-82 [FREE Full text] [doi: 10.1177/21677026221147872] [Medline: 38236494]
134. McGorry P, Van Os J. Redeeming diagnosis in psychiatry: timing versus specificity. Lancet 2013;381(9863):343-345. [doi: 10.1016/s0140-6736(12)61268-9]
135. Crişan LG, Vulturar R, Miclea M, Miu AC. Reactivity to social stress in subclinical social anxiety: emotional experience, cognitive appraisals, behavior, and physiology. Front Psychiatry 2016;7:5 [FREE Full text] [doi: 10.3389/fpsyt.2016.00005] [Medline: 26858658]
136. Shaukat-Jali R, van Zalk N, Boyle DE. Detecting subclinical social anxiety using physiological data from a wrist-worn wearable: small-scale feasibility study. JMIR Form Res 2021;5(10):e32656 [FREE Full text] [doi: 10.2196/32656] [Medline: 34617905]
137. Cosgrove VE, Rhee SH, Gelhorn HL, Boeldt D, Corley RC, Ehringer MA, et al. Structure and etiology of co-occurring internalizing and externalizing disorders in adolescents. J Abnorm Child Psychol 2011;39(1):109-123 [FREE Full text] [doi: 10.1007/s10802-010-9444-8] [Medline: 20683651]
138. Lahey BB, Krueger RF, Rathouz PJ, Waldman ID, Zald DH. A hierarchical causal taxonomy of psychopathology across the life span. Psychol Bull 2017;143(2):142-186 [FREE Full text] [doi: 10.1037/bul0000069] [Medline: 28004947]
139. van der Linden D, Dunkel CS, Prinzie P, Yamanaka-Altenstein M, von Wyl A, Hengartner MP. Overlap between general factors of psychopathology and personality: they share associations with daily life functioning and communication style. Curr Psychol 2021;42(13):10964-10976. [doi: 10.1007/s12144-021-02354-7]
140. Kenny DA, Mohr CD, Levesque MJ. A social relations variance partitioning of dyadic behavior. Psychol Bull 2001;127(1):128-141. [doi: 10.1037//0033-2909.127.1.128]

Abbreviations

HF-HRV: high-frequency heart rate variability
SIAS: Social Interaction Anxiety Scale
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
VR: virtual reality

Edited by S Brini; submitted 16.Jun.2025; peer-reviewed by K Wen, M Chollet, C Corciova; comments to author 03.Oct.2025; accepted 22.Dec.2025; published 11.Mar.2026.

Please cite as:
Son G, Rubo M
Automated Inference of Social Anxiety From Behavior in Social Virtual Reality: Cross-Sectional Observational Study
J Med Internet Res 2026;28:e79147
URL: https://www.jmir.org/2026/1/e79147
doi:10.2196/79147
PMID:

©Gayoung Son, Marius Rubo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Temporal Annotation of German Clinical Language in Real and Synthetic Clinical Documents: Corpus Development and Baseline Tagger Validation Study

Luise Modersohn[1,2], MSc; Udo Hahn[2,3], Prof Dr

[1]Chair of Medical Informatics, Institute for AI and Informatics in Medicine, Technical University of Munich (TUM) university hospital, Munich, Bavaria, Germany

[2]JULIELab, Friedrich Schiller University Jena, Jena, Germany

[3]Institute for Medical Informatics, Statistics and Epidemiology (IMISE), Leipzig University, Leipzig, Germany

Corresponding Author:
Luise Modersohn, MSc
Chair of Medical Informatics
Institute for AI and Informatics in Medicine
Technical University of Munich (TUM) university hospital
Ismaninger Str. 22
Munich, Bavaria, 81675
Germany
Phone: 49 8941404320
Email: luise.modersohn@tum.de

Abstract

Background: Temporal information about patients constitutes a precious source for clinical decision-making and medical treatment. The automatic extraction of such data from unstructured clinical narratives requires time-annotated clinical reports and notes from which time-informed taggers can be learned. Unfortunately, the non-English clinical language community, the German one as a typical example, with only a few exceptions, generally lacks such time-annotated resources to train and evaluate temporal taggers.

Objective: To overcome this metadata bottleneck, we developed a TimeML-conformant annotation schema for both temporal entities and temporal relations adapted to the needs of German medical language. Based on the annotations derived therefrom, we trained state-of-the-art baseline taggers to recognize temporal expressions in clinical documents.

Methods: Starting from temporal annotation guidelines for English clinical documents, we developed preliminary annotation guidelines for temporal named entities and temporal relations for the German language. These guidelines were subsequently refined and adapted to German clinical jargon, incorporating the work experience of 5 clinically trained annotators (students of medicine). For this task, we used randomly selected smaller subsets of 2 German clinical corpora—a real-world one ($3000PA_J$) and a synthetic one (GraSCCo). Both corpora were annotated ($3000PA_J$ partially, GraSCCo completely), randomly selecting 10% of the documents as an agreement part on $3000PA_J$. To measure interannotator agreement (IAA), we computed pairwise F_1-scores. We used that metadata to develop BERT (Bidirectional Encoder Representations from Transformers)-based language models for the creation of time-sensitive baseline taggers. All annotations are based on TimeML, the international de facto standard for time information markup.

Results: We created $3000PA_J$-temp, a time-annotated corpus of real clinical documents (which cannot be distributed because of the rigid privacy legislation enforced for German clinical data), and GraSCCo-temp, a synthetic one (which is publicly available without any restrictions). Based on the final guidelines, we achieved an IAA F_1-score of 0.9 on both corpora for the temporal named entity recognition task. For the temporal relation extraction task, the IAA on GraSCCo plummeted to an F_1-score of 0.57 and 0.41 on $3000PA_J$, respectively. Still, those results are comparable with English clinical datasets. Our baseline tagger for named entities reached F_1-scores between 0.64 and 0.85. For automatic relation extraction, we achieved F_1-scores ranging between 0.60 and 0.64.

Conclusions: We here introduce the first TimeML-compliant annotation scheme for time expressions occurring in German clinical language and apply it to 2 clinical corpora, one with nondistributable real clinical data, the other with distributable synthetic

XSL•FO
RenderX

ones. The latter constitutes the first publicly accessible, temporally annotated clinical corpus for the German language. The time tagger trained on these datasets is the first of its kind, fully compliant with the TimeML markup language. The amounts of temporal metadata in our corpora are among the largest datasets ever produced for the clinical domain, both compared with English and German predecessors.

(J Med Internet Res 2026;28:e71458) doi:10.2196/71458

KEYWORDS

natural language processing; clinical NLP; German clinical language; corpus development; temporal named entity; temporal relations; annotation guidelines; annotation process

Introduction

Overview

Information extraction is a field of natural language processing (NLP) that is concerned with automatically identifying mentions of entities of interest (so-called named entities) and semantic relations holding among them [1-3]. In the medical domain, commonly researched named entity types are, for example, Symptom, Finding, Diagnosis, Disease, or Drug, that are conceptually linked via medically relevant semantic relation types, such as Drug – hasDosage – Dosage, Drug – hasEffectOn – Disease (examples are shown in Figures 1A and 1B). Whereas named entity types address the terminological level of medical knowledge, including its taxonomic structure, semantic relations help represent single pieces of factual knowledge (medical assertions or facts).

Sets of medical entities and facts identified by information extraction tools form a bag of information units lacking further structure. One particularly important dimension to add structure to otherwise unrelated information pieces focuses on their temporal ordering on the time axis, a class of information we here refer to as temporal knowledge [4-7]. Information about the temporal order of (clinical) events is crucial in a hospital setting, because, for example, a drug reported on admission has a medically different status compared with one prescribed at discharge. Ultimately, temporal knowledge can be used to visualize timelines of patients' histories, which might prove helpful for doctors in medical decision making and for nurses in their daily care routines. This type of knowledge can either be encoded in terms of the named entity types “Event” and “Temporal Expressions,” such as Date, Duration, and Time, and “temporal relations,” which impose a (temporal) ordering on named entities or semantic relations. For instance, we may state a temporal relation between named entities, such as AFTER (examples are shown in Figure 1C), or between semantic relations, such as BEFORE (examples are shown in Figure 1D).

Prior to the automatic extraction of named entities, semantic and temporal relations, specialized classifiers (taggers) have to be pretrained or fine-tuned, typically using manually supplied annotations to guide the machine learning process. We here face a serious disparity between corpora for English and non-English languages (such as German, the focus of our work [8]). Based on an extensive survey of temporal datasets and tagging tools in the Related Work section, we encounter a large number of temporally annotated resources for the English language, both for clinical and nonclinical domains, that have no comparable counterpart for the German (clinical and nonclinical) language. For German, such resources are scarce and, if available, rather small in the nonclinical domain, while in the clinical domain, they are almost nonexistent. Moreover, annotations obeying the full TimeML annotation language standard are entirely missing for German.

Accordingly, in the Methods section, the unique contributions of this article are described in depth: (1) the provision of 2 clinical German-language corpora, a nondistributable real clinical corpus ($3000PA_J$) and a distributable synthetic corpus (GraSCCo), both of which have been richly annotated, adhering to TimeML, the current meta-language standard for temporal annotations. $3000PA_J$ and GraSCCo are the first corpora annotated with TimeML markup for the German language ever. (2) Exploiting these annotations, we trained the first transformer-based baseline taggers that automatically extract temporal information encoded in German clinical reports and notes. Up until now, transformer-based temporal taggers are not even available for the general German language (eg, covering news or social media content).

XSL•FO
RenderX

Figure 1. Examples for named entities, semantic, and temporal relations.

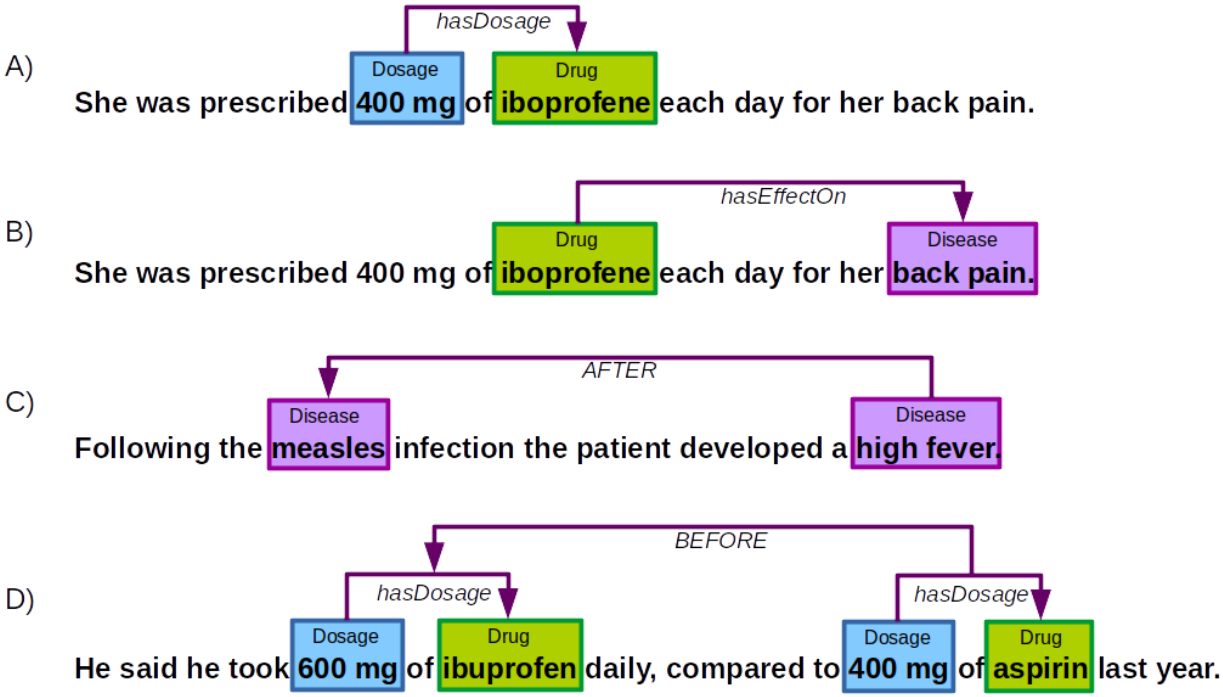

Related Work

Temporal information about patients constitutes a precious source for clinical decision making and medical treatment. It forms cohesive ties between single factual statements by linking them via temporal relations (see the examples for AFTER and BEFORE in Figure 1), which are the basis for temporal reasoning (eg, AFTER (x,y) and AFTER (y,z) transitively imply AFTER (x,z)). The automatic extraction of such temporal data is a prerequisite for the overarching goal to automatically construct consistent, maximally complete, and chronologically coherent timelines of the patients' medical history from unstructured clinical documents and structured electronic medical records to help support medical diagnosis, treatment decisions, and prognosis [9-20].

The automatic extraction of such temporal information from unstructured medical records (clinical reports, notes, or free text fields in the electronic medical record) requires language models trained on raw or metadata-enriched, that is, annotated, textual data. However, annotated temporal data resources are still quite rare and often seriously limited in size. Shortage of corpora with temporal metadata, however, directly hinders supervised learning of time-sensitive language models.

The starting point of time-centered knowledge representation research dates back to Allen's time algebra [21], which was then adopted in the design of TimeML [22], a specification and annotation language for temporal natural language utterances that combines temporal entities (Dates, Durations, Events, etc) and relations (eg, BEFORE, AFTER, OVERLAP); temporal relations impose temporal orderings on temporal entities. After years of experimentation, TimeML evolved into an ISO (International Organization for Standardization) standard as ISO-TimeML [23]. Part of TimeML, and 2 of its historical predecessors, are TIDES TIMEX2 [24] and TIMEX3 [23], which are used to mark up explicit temporal expressions, such as time expressions, dates, durations, etc. Table 1 illustrates the dominant role of TimeML as the representational foundation (markup or annotation language) for time-annotated corpora.

XSL•FO
RenderX

Table 1. Time-focused (mostly English only) nonclinical corpus resources.

Corpus/(challenge) task (year)	Functional characterization of (challenge) task	Annotation language	Text genre	Size in tokens	Type/number of metadata
TimeBank 1.1 [25] (2003)	N/A[a] (corpus)	TimeML 1.1	186 news articles	68.5k	13 temporal relation types: 5.1k temporal relation annotations; 11.2k annotation items, in total
TimeBank 1.2 [26] (2007); for error correction see the work of Ocal et al [27]	N/A (corpus)	TimeML 1.2	186/183 news articles	68.5k	1.4k TIMEX3 annotations; 7.9k Event annotations; 13 temporal relation types (6.4k temporal relation annotations and 27.6k annotation items, in total)
TimeBank-Dense [28,29] (2014)	N/A (corpus)	Subset of TimeML 1.2	36 news articles (subset of TimeBank 1.2)	13.3k (estimated)	0.3k TIMEX3 annotations; 1.7k Event annotations; 6 temporal relation types (12.7k [10.8k] temporal relation annotations; 14.7k annotation items, in total)
TempEval-1 [30] (2009)	Temporal relation (TLink) identification	Subset of TimeML 1.2 (TIMEX3, Event, TLink)	182 news articles (subset of TimeBank 1.2)	61.4k [31]	6 temporal relation types: 5.8k temporal relation annotations; 13.9k annotation items, in total [28]
TempEval-2 [32] (2010)	TIMEX3 identification; Event identification; temporal relation (TLink) identification	Subset of TimeML 1.2 (TIMEX3, Event, TLink)	Same as TempEval-1	63k	6 temporal relation types: 4.9k temporal relation annotations; 12.7k annotation items, in total (English data only) [28]
TempEval-3 [31] (2013)	TIMEX3 extraction and normalization; Event extraction and classification; end-to-end temporal relation (TLink) identification	Subset of TimeML 1.2 (TIMEX3, Event, TLink)	182 news reports (TimeBank) + 73 news reports(tiny subset of≈1M)(AQUAINT [33]);255 news articles, in total	Gold: 61.4 k (TimeBank) + 6.4 k +34.0 k (tiny subset of AQUAINT [33]) = 101.8k, in total; Silver: 666k (Gigaword [32])	13 temporal relation types (full TimeML set): 11.1k temporal relation annotations; 24.2k annotation items, in total (English data only) [28]
SemEval 2015 - TimeLine Task 4 [34] (2015)	Cross-document analysis of news articles to construct timelines for target events, that is, detecting, anchoring in time, and ordering events for a target entity	Subset (and minor adaptation) of TimeML 1.2	120 WikiNews articles, with 4 main topics: Apple, Airbus & Boeing, General Motors, Chrysler & Ford, stock market	10k per main topic, 40k, in total	0.9k Event annotations;0.7k Event chain annotations; 38 target entities; 37 timelines
RED [35] (2016)	N/A (corpus)	Subset of TimeML	95 news articles and casual discussion forum chats	54.3k tokens	1.1k TIMEX3 annotations; 8.7k Event annotations; 5 TLink types (Before, Overlap, Begins_on, Ends_on, Simultaneous): 1.4k temporal relation annotations (TLinks)
CaTeRS [36] (2016)	N/A (corpus)	Event definition based on the TRIPS ontology; subset of TimeML TLINKs	320 five-sentence short stories	1600 sentences	2.7k Event annotations;4 TLink types (Before, Overlaps, Contains, Identity): approximately 2.2k temporal relation annotations (TLinks) out of 2.7k causal and temporal relations
Event StoryLine (ESC) v0.9 [37] (2017)	N/A (corpus)	Compatible with TimeML	258 news articles about calamityevents, that is, natural disasters, shootings, killings, accidents, trials, etc.	N/A	1.3k TIMEX3 annotations; 7.3k Event annotations; 4 TLink types (Before, After, Contains, Overlap): 6.9k temporal relation annotations (TLinks)

XSL•FO
RenderX

Corpus/(challenge) task (year)	Functional characterization of (challenge) task	Annotation language	Text genre	Size in tokens	Type/number of metadata
Universal Decompositional Semantics Time (UDS-T) [29] (2019)	Categorical temporal relations are represented as real-valued relative timelines (on top of Universal Dependency parse trees)	Mapping of events to their likely durations and event pairs as real-valued relative timelines	Mixed Web documents (blog posts, newsgroup threads, emails, product reviews, and answers from QA websites) [38]	254.8k tokens [38]	32.3k Event annotations; 70.4k temporal relation annotations
TDG [39] (2020)	N/A (corpus)	Subset of TimeML + Temporal Dependency Graphs	500 WikiNews articles	Not reported	2.5k TIMEX3 annotations; 15.0k Event annotations; 4 TLink types (Before, After, Overlap, Includes): 28.4k temporal relation annotations (TLinks)
Maven-Ere [40] (2022)	N/A (corpus)	Subset of TimeML: TIMEX3, TLink	4480 Wikipedia articles	Not reported	25.8k TIMEX3 annotations; 6 TLink types (Before, Contains, Overlap, Begins_on, Ends_on, Simultaneous): 1216k temporal relation annotations (TLinks)
TimeBankNT [41] (2024)	Full timeline annotation (all possible TLinks) of the sparse TLink TimeBankDense [28] annotations	Subset of TimeML: TIMEX3, Event, TLink	36 news articles(subset of TimeBank 1.2)	13.3k (estimated)	0.3k TIMEX3 annotations; 1.7k Event annotations; 102.3k (104.8k) temporal relation annotations (TLinks)

[a]N/A: not applicable.

The early TimeBank corpus [25,26,28,29] witnessed the consolidation of TimeML over a decade. As an experimental vehicle, it is characterized by small document and token numbers as well as low amounts of annotated temporal data.

A simplified version of expressionally rich TimeML became the metadata backbone for TempEval [30-32,42,43], a series of time-focused shared task challenges run by NLP researchers within the framework of SemEval, a workshop series focusing on a broad range of semantic interpretation topics organized under the auspices of the Association for Computational Linguistics.

Shared tasks, such as TempEval or SemEval, have become a major driver of progress for NLP [44,45]. They are based on a common gold standard dataset on which several competing research group teams run their own systems against predefined tasks; their submitted solutions are, finally, compared against the gold standard, which yields a ranking of the teams involved based on their performance for specific tasks. Top-ranked teams are then given the opportunity at a closing workshop and invited to journal publications to feature the most successful methodologies—thus, continuously advancing the methodological repertoire of NLP on the basis of empirical evidence from benchmarking experiments.

The TempEval-3 corpus [31] was the largest in this series and contains 255 news articles, with roughly 95k tokens, selected from the TimeBank 1.2 corpus [25,26] (183/186 articles) and the Aquaint corpus (73 news reports) [33] (numbers slightly differ dependent on the cited literature); TimeBank-Dense [28,29], its relationally richer successor, comes with a total of 12,715 temporal relations (more than 2 times the number of the original TimeBank corpus). Interestingly, the second edition of TempEval [32] contained data not only for English, but also for Chinese, French, Italian, Korean, and Spanish, whereas the third edition [31] dealt with English and Spanish data only. A variant of TempEval was considered in the Timeline Task 4 of SemEval 2015, which featured cross-document analysis of news articles to construct timelines for events [34] described in a series of documents. For all 3 challenges in TempEval, 9 subtasks had to be tackled that fell into 3 categories: time expression identification (spans and TIMEX3 classes, such as Date, Time, Duration, etc), event expression identification (spans and Modality, Degree, Polarity, etc), and temporal relation identification (temporal relations, such as BEFORE, OVERLAP, AFTER, and a narrative container relation typed as CONTAINS).

In an intermediary phase, new corpora were generated featuring different text genres (eg, user-generated contents [35,36,46], Wikipedia articles [40]) and playing with different sets of temporal relations, yet with only insignificant growth of document sets or metadata volumes. One of the rare departures from strict adherence to TimeML can be found in the UDS-T corpus [46], where TimeML's categorical temporal relations are represented as real-valued relative timelines (on top of Universal Dependency parse trees). UDS-T also stands for a trend towards increasing quantities of documents/tokens (>250k tokens) and metadata (>100k annotation items). Truly impressive breakthroughs in both of these dimensions have been achieved only recently by getting large numbers of crowdworkers involved, such as for Maven-Ere [40], which excels in, for example, more than a million temporal relation annotations. In addition, TimeBankNT [41] increases the number of metadata by an order of magnitude (>100k temporal relation annotations) relative to TimeBank-Dense. Almost all of the corpora in Table 1 incorporate newswire/newspaper documents only (with some exceptions mentioned above). In contradistinction, Table 2 features corpora with clinical contents.

XSL•FO
RenderX

Table 2. Time-focused (English only) clinical corpus resources.

Corpus/(challenge) task (Year)	Functional characterization of (challenge) task	Annotation language	Text genre	Size in tokens	Type/number of metadata
CLEF [47] (2009)	N/A[a] (corpus)	TIMEX3 & self-defined set of 8 temporal relations (Before, After, Overlap, Is_included, Ended_by, Begun_ by)	50 clinical reports; 50 histopathology reports;50 imaging reports	Not reported	Clinical named entity types (Drug-or-Device, Intervention, Condition, Investigation, Locus, etc); 3.8k named entity annotations; clinical relation types (has_Target, has_Finding, has_Location, Modifies, etc); 2.4k clinical relation annotations; 0.5k temporal relation annotations (for 10 documents, only)
i2b2 2012 [48] (2013)	Spans and types/attributes of temporal expressions (TIMEX3), referring to Dates, Times, Durations (incl. Normalization to ISO[b]8601 standard); Spans and types/attributes of Events, including. Both clinical concepts, such as Problems, Tests, Treatments, and Events relevant to the patient's clinical timeline, such as Admissions, Transfers between departments; Temporal relations between the clinical Events and temporal expressions (Before, After, Overlap)	Simplified version of TimeML adapted to clinical narratives and purposes (eg, the inclusion of clinical concepts, such as Problems, Tests, and Treatments, into Events) [49]	310 discharge summaries	About 178k tokens	2.4k (train) + 1.8k (test) annotation units for temporal expressions (TIMEX3); 15.6k (train) + 13.6k (test) annotation units for temporal Events; 33.5k (train) + 27.7k (test) annotation units for temporal relations; 94.6k annotation units, in total
THYME [50] (2014)	N/A (corpus)	ISO-TimeML (THYME-TimeML) [23]; TIMEX3, Event, 5 temporal relations (Before, Overlap, Begins_on, Ends_on, Contains)	1254 clinical notes and pathology reports (oncology: brain cancer, colon cancer)	Not reported	Subset of the THYME colon cancer corpus (107 documents): 1.4k annotation units for temporal expressions (TIMEX3); 15.8k annotation units for temporal Events; 7.9k annotation units for temporal relations; 25.1k annotation units, in total
Clinical TempEval-1, subbranch within TempEval as part of SemEval [51] (2015) and Clinical TempEval-2 [52] (2016)	Spans and types/attributes of temporal expressions (TIMEX3), referring to Dates, Times, Durations, etc; Spans and types/attributes of Events, including both clinical concepts, such as Problems, Tests, Treatments; temporal relations between the clinical events and the document creation time (Before, After, Overlap, Before-Overlap) and narrative container relations	Extension of TimeML for temporal expressions, medical events, and temporal relations between events and times (Styler et al [50] shows the details of extensions of TimeML)	591 clinical notes and pathology reports (colon cancer)	Not reported	7.9k annotation units for TIMEX3 time expressions; 78.9k annotation units for Events; 23.2k annotation units for temporal relations; 110.0k annotation units, in total

XSL•FO
RenderX

Corpus/(challenge) task (Year)	Functional characterization of (challenge) task	Annotation language	Text genre	Size in tokens	Type/number of metadata
Clinical TempEval-3 [53] (2017)	Same tasks as in Clinical TempEval-1 and Clinical TempEval-2	Same annotation language as in Clinical TempEval-1 and Clinical TempEval-2	Same colon cancer dataset as in Clinical TempEval-1 and Clinical TempEval-2 + a second dataset (595 documents) with focus on brain cancer	Not reported	(Virtually the same number of annotation items for colon cancer as in Clinical TempEval-1 and Clinical TempEval-2) 6.6k annotation units for TIMEX3 time expressions; 48.9k annotation units for Events; 7.3k annotation units temporal relations; 62.8k annotation units for colon cancer, in total; 172.8k annotation units; in total.
n2c2 [54], follow-up event of i2b2 2012 (2023)	Identification of medication change (Event): determine the Action, Negation, Temporality, Certainty, and Actor for any change events (Context)	Categories for temporality (Past, Present, Future, Unknown) are considerably coarser in granularity than those from the i2b2 2012 Shared Task and are not grounded in TimeML	500 clinical notes (discharge summaries, correspondence notes)	About 308k tokens (estimated from the study by Kumar et al [55])	About 1.7k annotation units
CALEX[56] (2019)	N/A (corpus)	TimeML-based TIMEX3 (Date, Time, Duration, Frequency) plus an extension (CALEX) for timeline construction based on calendar expressions	180 discharge summaries, 53 psychiatry reports, 70 pediatrics reports, 75 emergency room reports, 378 clinical reports, in total	55,433 (discharge summaries); 67,569 (psychiatry); 36,675 (pediatrics); 52,041 (emergency), 211.7k clinical reports, in total	2378 TIMEX3 annotations

[a]N/A: not applicable.

[b]ISO: International Organization for Standardization.

The CLEF corpus [47] was the first clinical time-focused dataset ever, with limited size though (150 documents, <7k annotated items from which only a small fraction cover temporal information), restricted to TIMEX expressions, yet with a rich repertoire of named entities, (self-defined) semantic and temporal relations. It became the forerunner of the seminal THYME corpus [50], a flagship enterprise that, unlike CLEF, fully adhered to TimeML representation conventions and came with roughly 1.3k clinical reports and slightly more than 25k temporal annotations on a 107-document subset thereof.

Fortunately, the idea of challenge competitions also attracted clinical NLP researchers. Inspired by TempEval, temporal aspects were targeted within the i2b2 2012 Shared Task on Temporal Relations in Clinical Text [48], here with focus on time information contained in clinical reports. Participating systems were required to extract virtually the same types of temporal knowledge as those targeted by TempEval:

1. Spans and types/attributes of temporal expressions, referring to Dates, Times, Durations, or frequency phrases in the clinical text (subsequently, the values of the extracted TIMEX temporal expressions had to be normalized to the ISO8601 specification standard).
2. Spans and types/attributes of clinically significant events, including both clinical concepts (ie, named entities) such as Problems, Tests, Treatments, and ClinicalDepartments, and events relevant to the patient's clinical timeline, such as Admissions, Transfers between departments.
3. Temporal relations between the clinical events and temporal expressions.

For these challenge tasks, the i2b2 2012 Shared Task organizers added a temporal annotation layer to the dataset, 310 discharge summaries (about 178k tokens), that had already been used in previous i2b2 rounds. For annotation purposes, a simplified version of TimeML was taken as a starting point, but had to be adapted to better suit clinical narratives and purposes (eg, the inclusion of clinical concepts, such as Problems, Tests, and Treatments, into Events) [49]. The i2b2 2012 Shared Task corpus, finally, is accessible on the basis of signing a data use agreement and comes (for training) with 2366 annotation units for clinically relevant temporal expressions (TIMEX3 data), 15,567 annotation units for clinically relevant temporal events, and 33,543 annotation units for clinically relevant temporal relations; train and test data sum up to almost 95k annotation units. This landmark corpus exceeded previous annotation efforts in this field (most notably CLEF [47]) both in terms of the number of annotated documents and annotation units by a large margin and became a benchmark for future research.

Recently, a time-wise much narrower shared task devoted to medication change events documented in clinical notes was run within the context of n2c2, the follow-up event of i2b2 [54].

XSL•FO
RenderX

Participants had to identify medication change (Event) and determine the Action, Negation, Temporality, Certainty, and Actor for any change events (Context). The categories for temporality (Past, Present, Future, Unknown; about 1.7k annotation units) are considerably coarser in granularity than those from the previous i2b2 2012 Shared Task and are not grounded in TimeML. An annotation effort that extends TIMEX3 expressions for timeline construction based on calendar expressions is reported by Viani [56].

Complementary to these activities, within the TempEval branch of SemEval, a clinical TempEval subbranch was established. In the first 2 runs of this medically focused TempEval challenge [51-53], roughly 600 deidentified clinical notes and pathology reports from colon cancer patients at Mayo Clinic were manually annotated with an extension of TimeML for temporal expressions, medical events, and temporal relations between events and times (details of the annotation process and extensions of TimeML as part of the THYME corpus are described by Styler et al [50]). In the third run of TempEval [53], this dataset was complemented by a second one with a focus on brain cancer, as the organizers wanted to explore the potential for domain adaptation, that is, how do systems trained on colon cancer data perform when tested on brain cancer data?

The TimeML-annotated corpus from TempEval is available on a data use agreement basis as well, and comes (for colon cancer) with roughly 83k annotation units in its training (50%), and development (25%) splits, hence contains around 110k annotation units together with the held-out test set (25% of the entire dataset). For brain cancer, 48k annotation units are available for training (50%) and development (25%), and 15k in the test split, summing up to 63k, in total. Thus, the corpus from TempEval-3 exceeds the i2b2 2012 Shared Task corpus by slightly less than a factor of 2 in terms of annotation units and figures as the second milestone for temporal clinical corpora for the English language.

All of the above-mentioned corpus building activities almost only cover the English (clinical) language. Since our work deals with German language data, Table 3 gives an overview of the few efforts to deal with nonclinical German from a temporal perspective.

Table 3. German time-focused nonclinical resources (Corpora and Taggers).

Resource type	Name of resource (Year)	Functional characterization	Annotation language	Text genre	Size in tokens	Type/number of metadata
Corpus	EuroParl [57] (2008)	Time annotation mappings in parallel corpora: English → German	Subset of TimeML: TIMEX3, Events, no temporal relations	Transcribed speeches in the European Parliament	960k bisentences (train) and 0.3k bisentences (dev/ test)	Not reported
Corpus	WikiWarsDE [58] (2011)	N/A[a]	TIMEX2 [22], a subset of TimeML restricted to dates, time expressions, and durations)	22 German Wikipedia articles about famous wars in history corresponding to the English selection in WikiWars [59]	95.6k tokens	2.2k TIMEX2 annotation units
Corpus	KRAUTS [60] (2018)	N/A	TIMEX3	192 newspaper articles	75.7k tokens	1.1k TIMEX3 annotation units
Time tagger	HeidelTime [61] (2013)	Rule-based tagger using gazetteers and regular expressions	TIMEX2, TIMEX3 subset of TimeML plus normalizations into ISO[b] 8601 standard	22 German Wikipedia articles about famous wars in history corresponding to the English selection in WikiWars [59]	100.7k tokens	2.2k TIMEX2 annotation units
Time tagger	HeidelTimeext [62] (2022)	Extended rule-based tagger using gazetteers and regular expressions	TIMEX3	Mix of 10 parliament protocols, 10 books, 766 newspaper and 10 scholarly articles, 10k Wikipedia sentences	5.121k tokens	56.8k/53.8k TIMEX3 annotation units (automatically generated)

[a]N/A: not applicable.

[b]ISO: International Organization for Standardization.

Three German-language corpora have been built with a focus on time information. EuroParl [57] collects transcribed speeches (roughly 1m bisentences) from the European Parliament. TimeML-conformant temporal annotations are automatically mapped from English source texts to German targets. WikiWarsDE [58] assembles 22 Wikipedia articles on famous wars; its time representations are restricted to TIMEX expressions. So is the annotation in KRAUTS [60] based on 192 newspaper articles. For all 3 corpora, token counts do not exceed 100k, and at most, roughly 2k manual gold standard annotations are provided (HeidelTimeext [62] offers a substantially larger yet automatically generated silver standard). Overall, these are, in comparison with the numbers in Table 1, tiny corpora with low annotation density and limited representational expressiveness (with the exception of Spreyer and Frank's study [57]).

XSL•FO
RenderX

Two versions of the same time tagger also deserve mention. HeidelTime [61] was originally built on English data and then augmented for tagging German data. HeidelTime is a classical rule-based system using regular expressions and additional gazetteers. The step from HeidelTime to HeidelTimeext [62] comes with an augmented and updated rule set, an increase of the document size and token number (more than 5m tokens), and the number of temporal annotations (more than 50k items). However, in both cases, the representational expressiveness is limited to TIMEX expressions.

The state of German-language clinical corpora with temporal metadata is basically a short story summarized in Table 4. The MACCS corpus [63] is made of slightly more than 1.7k discharge summaries and clinical notes from the nephrology domain whose temporal annotations are based on a corpus-specific, that is, non–TimeML-conformant annotation language. Our work, instead, is fully TimeML-compliant and features 2 different corpora. 3000PA$_J$ comes with 1100 real clinical documents, 1.7M tokens, and 120k temporal annotations, whereas GraSCCo is a much smaller corpus product featuring 63 synthetic documents, 44k tokens, and almost 12k temporal annotation units. Due to rigid data privacy legislation in Germany, 3000PA$_J$ cannot be made publicly available, whereas GraSCCo can (it consists of fictitious clinical data). With roughly 131k annotation items, however, both corpora not only exceed all German-language corpora (whether clinical or nonclinical) in terms of temporal metadata but also all clinical English time-focused corpora, with one exception—only the million-scale nonclinical Maven-Ere corpus [40] offers a (much) larger number of temporal annotations.

How do 3000PA$_J$ and GraSCCo compare with other non-English clinical corpora annotated with temporal metadata? Table 5 lists such textual resources whose size exceeds 50 documents and 1k annotation units. Perhaps, the closest in kinship is the French MERLOT corpus [64] that also adheres to TimeML specifications but falls behind by orders of magnitude both in terms of tokens and the amount of temporal metadata. Three other French corpora [65-67] are much smaller than MERLOT and also contain fewer temporal annotations; only Bannour et al [66] adheres to TimeML, including temporal relations. From 4 Chinese corpora, only the work of Liu et al [68] contains TimeML specifications for TIMEX3 expressions and temporal relations, yet on a much smaller scale than 3000PA$_J$, both document, token, and metadata-wise. Pan et al [69] and Liu et al [70] focus on TIMEX3 expressions only (Liu et al [70] offering the largest corpus in terms of the number of documents, yet not of the number of annotation items), whereas Hu et al [71] use a self-supplied time ontology for annotation. Swedish [72], Portuguese [73], Spanish [17], and Italian [74] resources are tiny in terms of the number of documents and annotation items they contain (in the range of less than some hundred documents) and the number of temporal relations (less than 5k, if at all). Finally, Cheng et al [75] contain 2.6k THYME-conformant temporal relation annotations from a Japanese corpus based on slightly more than 1k documents. Another Japanese corpus, MedTxt-CR-JA [76], comes with 27.4k temporal relation annotations that are inferred from 2.6k explicitly annotated start-point annotations. TIMEX3-only annotations are provided by Hamon et al [65], Jeyafreeda et al [67], Pan et al [69], and Velupillai [72]. With the exception of the multilingual E3C corpus [77] (composed of French, Spanish, Basque, and Italian clinical case reports which differ significantly in style and format from real clinical reports), none of these corpora is publicly available.

Table 4. German time-focused clinical corpus resources with our own contribution highlighted.

Resource type	Name of resource (Year)	Annotation language	Text genre	Size in tokens	Type/number of metadata
Corpus	MACSS [63] (2016)	Corpus-specific, that is, non-TimeML-conformant temporal annotations: time points, dates, temporal courses	118 discharge summaries;1607 clinical notes from the nephrology domain (kidney transplantations)	89.7k tokens; 68.5k tokens	Not reported
Corpus	3000PA$_J$ (our work)[a]	TimeML:TIMEX3, Events, Six temporal relations (Before/After, Contains, Overlap, Begins_on, Ends_on)	1100 real clinical documents, mostly discharge summaries	1.7M tokens[a]	13.2k annotation units for TIMEX3 time expressions; 46.9k annotation units for Events; 59.1k annotation units for temporal relations; 119.2k[a] annotation units, in total
Corpus	GraSCCo (our work)[a]	TimeML:TIMEX3, Events, Six temporal relations (Before/After, Contains, Overlap, Begins_on, Ends_on)	63 synthetic clinical documents	44k tokens[a]	1.2k annotation units for TIMEX3 time expressions; 4.5k annotation units for Events; 5.8k annotation units for temporal relations; 11.6k[a] annotation units, in total

[a]Our own contribution.

XSL•FO
RenderX

Table 5. Non-English and non-German time-focused clinical corpus resources.

Natural language (ISO[a] 639 codes)	Name of resource (Year)	Annotation language	Text genre	Size in tokens	Type/number of metadata
FR	No name [65] (2014)	TIMEX3 expressions (plus normalizations)	182 clinical records (train) and 120 records (test, 25 with reference annotations), 302 records, in total	Not reported	Not reported
SV	Stockholm EPR [72] (2014)	TIMEX3 expressions	112 clinical notes from an ICU[b] unit	≈50k (estimate)	<2k TIMEX annotations
ZH	CMedTEX [70] (2016)	Simplified TimeML TIMEX3 expressions: Date, Time, Duration, and Frequency (plus ISO8601 normalizations)	1778 clinical notes	Not reported	46.6k TIMEX3 annotations
PT	No name [73] (2018)	TimeML TIMEX3: Date, Time, Duration, Set (plus ISO8601 normalizations)	130 discharge summaries	Not reported	Not reported
FR	MERLOT [64] (2018)	Simplified TimeML: Date, Time, Duration, and Frequency; 6 TimeML-compliant temporal relations: Before, Begins_on, Ends_on, During, Overlap, Simultaneous	500 discharge summaries, procedure reports (eg, radiology reports), physician letters, and prescriptions	148.5k	3.9k temporal entity annotations; 4.6k temporal relation annotations
ZH	No name [68] (2019)	Three TimeML-compatible temporal relations: Before, After, Simultaneous	563 clinical notes	Not reported	4.0k temporal expressions (ako TIMEX3 expression); 4.0k time intervals (relations)
IT	No name [74] (2019)	THYME-conformant clinical events (Problems, Tests, Treatments, Occurrences) and temporal relations (Before, Before/Overlap, Overlap, After)	75 clinical reports	57,263	4.4k temporal relation annotations
ES	No name [17] (2020)	TIMEX3 expressions (plus expression normalization) and TimeML-noncompliant temporal relations	50 clinical notes and 50 clinical reports for TIMEX3 identification & 100 clinical notes and 100 clinical reports for temporal relation identification, 300 documents, in total	Not reported	TIMEX3 annotations not reported; 200 temporal relations
ZH	TNorm [69] (2020)	TimeML-compliant TIMEX3 (plus expression normalization) only, no temporal relations	900 discharge summaries	Not reported	12.1k TIMEX3 annotations
ZH	No name [71] (2022)	Self-designed Clinical Time Ontology as a basis for an OWL-based temporal logic, with 11 temporal classes (various types of Instants, Intervals, Instant and Interval Collections, and Modifiers) and 16 temporal relations (extension of Allen's 13 relations set)	300 case reports containing 3000 clinical statements (presumably sentences)	Not reported	3.5k temporal entity annotations; 3.0k temporal relation annotations; 6.5k annotation units, in total
JA	No name [75] (2022)	THYME-compliant, simplified annotation schema: TIMEX3 + 5 temporal relations: On, Before, After, Start, Finish (complemented by 5 medical relations: Change, Compare, Feature, Region, Value)	1000 radiography interpretation reports of lung cancer + 156 medical history reports of idiopathic pulmonary fibrosis	Not reported	2.6k temporal relation annotations

XSL•FO
RenderX

Natural language (ISO[a] 639 codes)	Name of resource (Year)	Annotation language	Text genre	Size in tokens	Type/number of metadata
FR, ES, IT, EU	E3C [77] (2023)	THYME-conformant annotation: TIMEX3 expressions, Events, temporal relations	81 French and Spanish, 86 Italian, and 90 Basque clinical case reports + 615 clinical case reports with silver standard (automatic) annotations	French: 25,196; Spanish: 24,681; Italian: 24,319; Basque: 22,505 + >200k (silver standard)	French: 0.3k TIMEX3, 4.3k Event, 4.0k temporal relation annotations; Spanish: 0.4k TIMEX3, 4.8k Event, 4.7k temporal relation annotations; Italian: 0.3k TIMEX3, 3.4k Event, 1.2k temporal relation annotations; Basque: 0.6k TIMEX3, 7.9k Event, 8.0k temporal relation annotations
FR	No name [66] (2023)	Subset of TimeML Document; Creation Time and 5 temporal relations: Before, Before_Overlap, Overlap, and After + TemporalReference (noncompliant with TimeML)	220 (train) + 57 (test) clinical reports	Not reported	2.9k temporal relation annotations
JA	MedTxt-CR-JA [76] (2024)	Start-point ordering of events from which 3 temporal relations can be inferred: Before, Equal, and After	62 (from 148) case reports	Not reported	27.4k temporal relation annotations (based on 2.6k explicitly annotated start-point annotations)
FR	No name [67] (2024)	Corpus-specific, that is, non-TimeML-conformant temporal annotations: Date, Time, Frequency, Duration, and Age	150 (train) + 50 (test) clinical notes	Not reported	0.4k temporal entity annotations

[a]ISO: International Organization for Standardization.

[b]ICU: intensive care unit.

Goals of This Work

As stated above, temporal knowledge, its annotation, and automatic extraction in the medical domain have almost only been researched in depth in the Anglo-American language community. The non-English clinical language community is basically a low-/nonresource area that, with only a few exceptions, generally lacks temporal metadata. In this article, we propose a 2-fold solution, specifically responding to the strict privacy legislation constraints in the European part of the world.

The unique contributions of this article are the following: We provide temporal annotations for German clinical documents, including quality checks via interannotator agreement (IAA) metrics. More specifically, we supply temporal annotations for a nondistributable corpus composed of real clinical documents and a publicly distributable corpus composed of synthetic clinical documents. Furthermore, we train transformer-based baseline taggers to automatically extract temporal information encoded in German clinical reports and notes.

Methods

Overview

In this section, we will give a detailed overview of our 2 clinical corpora, the annotation schema for temporal named entities and relations, the annotation process and its evaluation, as well as the presentation of our baseline time tagging model for clinical documents. Figure 2 summarizes our experimental setting and illustrates the key steps in our workflow.

XSL•FO
RenderX

Figure 2. Overview of the experimental setup and workflow. BOHB: Bayesian Optimization and Hyperband; GBERT: German BERT model.

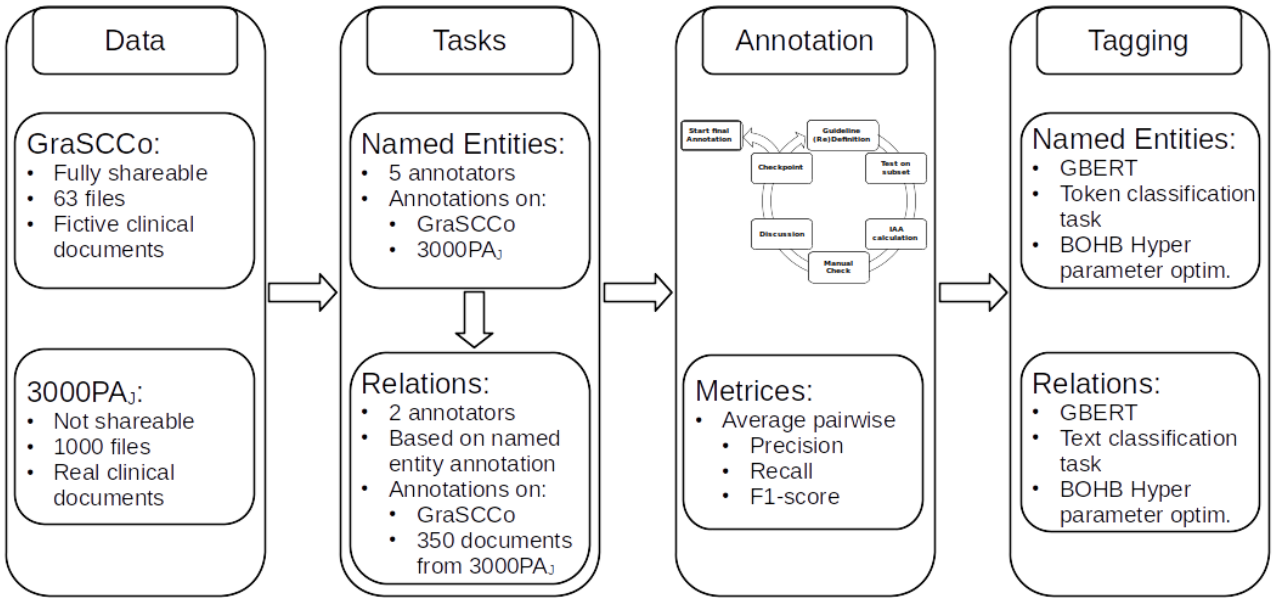

Ethical Considerations

The usage of 3000PA$_J$ is based on the approval by Jena University Hospital ethics committee (4639-12/15) and its data protection officer. The data was collected during the routine clinical care of adults at the intensive care unit or internal medicine ward. All persons were deceased at the time of data collection, thus no informed consent was needed for local data analysis according to the German law at that time. All data were hosted, annotated, and processed at the premises on secure servers with access restrictions and regular data protection training of the annotators. For GraSCCo, no ethics approval is needed as this dataset contains only fictional data written in the style of real clinical documents.

Clinical Data Resources

For temporal annotation, we used the 2 corpora already introduced in Table 4, namely 3000PA$_J$ and GraSCCo. 3000PA$_J$, a subset of the 3000PA corpus [78], consists of about 1100 real, that is, authentic, clinical documents, mostly discharge summaries, with roughly 145k sentences and 1.7m tokens hosted by the Jena University Hospital under strict access restrictions. In effect, this means that according to German data protection legislation, this dataset cannot be made available for public use outside the walls of the Jena University Hospital. 3000PA$_J$ has already been annotated at multiple semantic layers, including section headings [79], Personally Identifiable Information [80], medications [78], and major nontemporal clinical named entities, namely diagnoses, findings, and symptoms [81].

As an alternative to siloing much-wanted clinical data in the barred 3000PA$_J$ dataset, we developed GraSCCo [82], a synthetic twin corpus which is open for use to the entire clinical NLP community without any access restrictions. Albeit being a tiny document collection with only 63 German multiply-alienated and noise-infused real documents with about 5.4k sentences and 44k tokens, it is available via the Creative Commons license on Zenodo [83]. As a noteworthy remark, the authors have already gathered preliminary evidence that GraSCCo may act as a reasonable substitute for 3000PA in terms of clinical language use and genre patterns [82]. This observation is fully consistent with other experimental results for various languages, which show that synthetic corpora can indeed replace real clinical corpora with only marginal quality loss regardless of the application tasks and clinical text genres chosen (see, eg, the evidence found for English [84-86], French [87], and Norwegian [88]).

Temporal Annotation

Overview

Manually generating temporal annotations is well-known as a notoriously complex and time-consuming task [49,50]. Hence, to avoid cognitive overload of the human annotators, we split our annotation campaign into 2 separate phases—an entity annotation and a relation annotation phase. Both followed the same schema (Figure 3 depicts one phase equals a complete cycle) and were performed by at least 3 different annotators. This organizational scheme follows the widely acknowledged annotation conventions established by James Pustejovsky and Amber Stubbs [89]. Generally speaking, at least 3 independent annotators are necessary to run a proper statistical analysis. All of our annotators were qualified medical students who had already passed their first medical exam after 2 years of study (Physikum, in German).

Each annotation cycle was divided into a training-only and a production annotation phase. During the training phase, all annotators received the same partition of the entire dataset for annotation, usually about ten documents. Once this batch was completed, their annotations were checked both automatically

XSL•FO
RenderX

via an IAA metric based on the average of the pairwise F_1-scores, and manually to identify and resolve individual/systematic errors and misunderstandings. The results of the latter phase were then discussed within the annotation team and led to changes in the annotation guidelines, if needed (typical shortcomings arising from premature annotation guidelines are a lack of clarity or ambiguity of instructions, instruction gaps, etc). After that, a new subset was drawn and, again, annotated by all annotators. These training iterations continued until the F_1-score, primarily the pairwise F_1-score as well as the averaged F_1-score, stabilized and reached a reasonable, that is, community-wide accepted, plateau (in the range of 0.9 on a scale of 0 to 1.0 agreement) [90]. After this preparatory phase, the final guidelines were frozen and applied to the entire dataset, again, with a randomly chosen partition as the agreement portion, usually 5% to 10% of all documents. These documents will be called agreement documents, as they are the documents on which the IAA is calculated. The final F_1-score was determined on these agreement documents and reported as the final agreement value.

Once the entity annotation task was completed, we unified all annotations of all agreement documents. Thus, label disagreements and missing annotations were detected. These merged entity annotations with all their identified disagreements were then used for the second phase, the relation annotation, and manually corrected by the human annotators. The annotators were also encouraged to modify/correct entity annotations if they disagreed with the results of the first entity-focused round.

Figure 3. Schema of the iterative workflow for annotation guideline development. The optimization cycle is repeated until a stopping criterion is reached on a document subset, mostly a predefined IAA threshold or the number of iterations. The final guideline is subsequently used to annotate the entire corpus in the production phase. IAA: interannotator agreement.

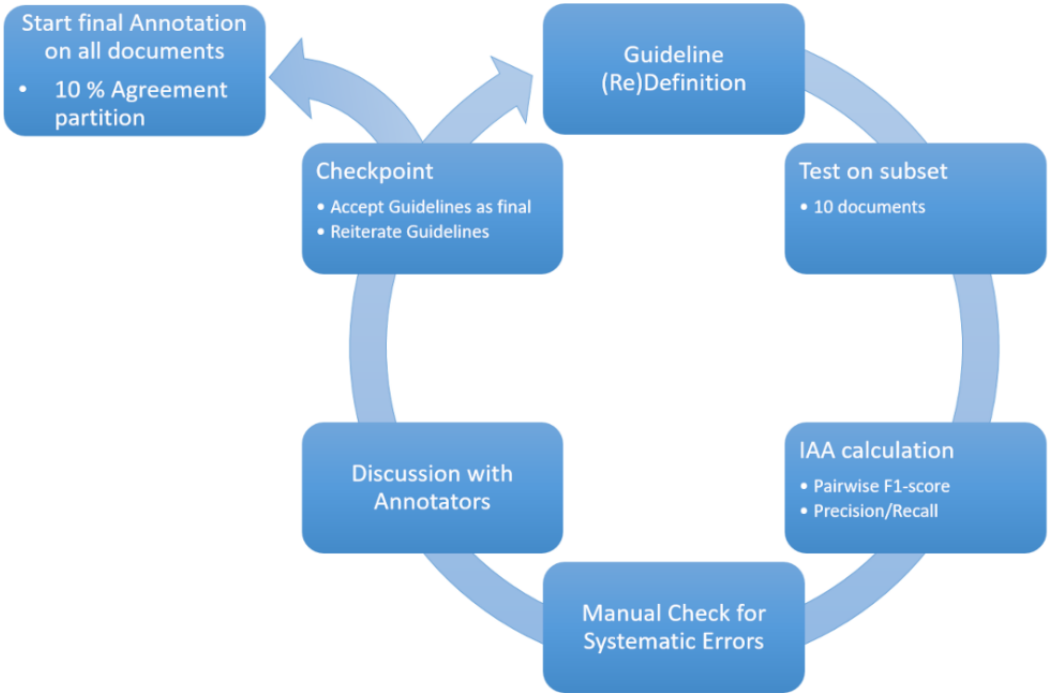

Temporal Entity Definition

Our definitions of time-related entities are identical to THYME [50]. Accordingly, those entities are defined as follows:

- EVENT: Any token or phrase that is related to things that occurred or happened.
- TIMEX3: Any token or short phrase that denotes a time expression of the following subtypes:
  - Date: Expressions or phrases that describe a time with accuracy up to a day (06/2020, next Wednesday),
  - Time: A specific time of the day (12 am, midnight),
  - Duration: A time frame (until yesterday, for 10 minutes),
  - Set: A recurring event (twice a day, 1-0-0),
  - Quantifier: The number of times something has happened or occurred, needs to be paired with an EVENT (second cycle of chemotherapy),
  - PrePost: Expressions like postoperative or premenarche.

Additionally, entities annotated as an EVENT needed to be marked with one of the following tags according to their point in time with respect to the reported hospital stay: BEFORE (if something happened prior to admission), BEFORE-OVERLAP (if something occurred before but is still present at the time the patient was admitted to the hospital), OVERLAP (if something happened during the hospital stay) and AFTER (if occurrences will happen in the future after the discharge). An example of said annotated entities can be found in Figure 4.

XSL•FO
RenderX

Figure 4. Sentences to illustrate the annotation of entities annotated with the BRAT annotation tool. Yellow indicates EVENTS, blue stands for TIMEX3 temporal expressions. The labels are written over the marked words or phrases.

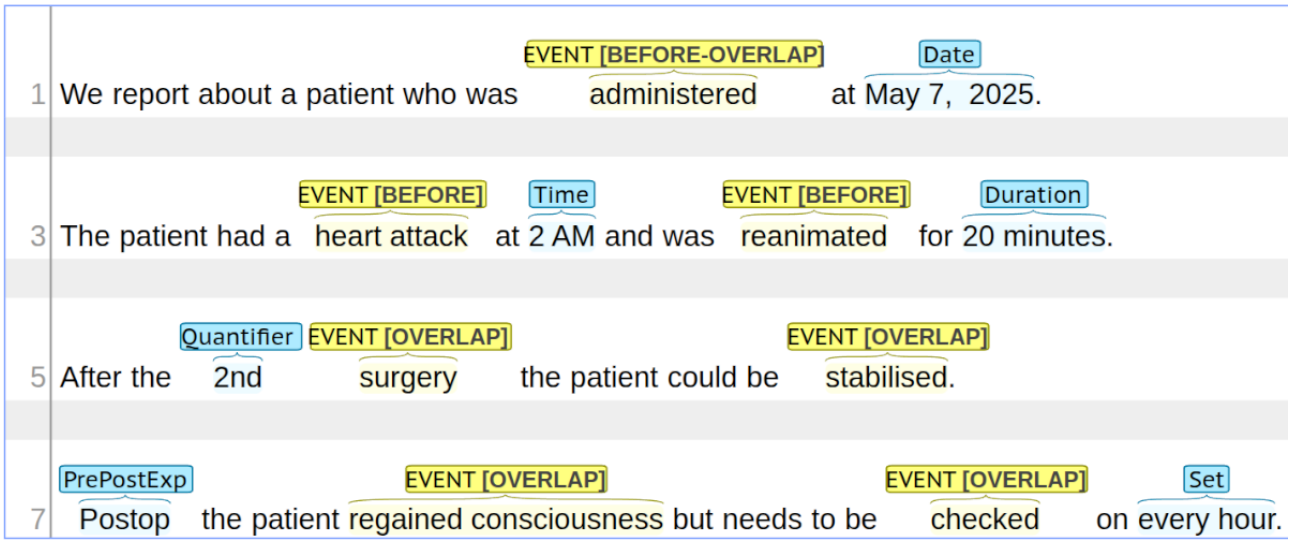

Temporal Relation Definition

Temporal entities themselves do not contain sufficient information for the adequate construction of timelines. Without context or ordering information, most of the time-relevant information is missing. Accordingly, temporal relations between those entities have to be determined. All of the relation types we chose are a subset of Allen's interval algebra [21]. We followed the definitions in THYME [50] and used the following relation types:

- BEFORE (X,Y) X happened entirely before Y
- (AFTER (Y,X) is treated as the inverse of BEFORE (X,Y))
- CONTAINS (X,Y) Y happened entirely during X
- OVERLAP (X,Y) X and Y share a common partition of time
- BEGINS-ON (X,Y) X started with Y
- ENDS-ON (X,Y) X finished with Y

In contrast to THYME and more in line with the i2b2 2012 guidelines, we have instructed our annotators to link each entity to at least one other entity in such a way that each entity is connected to every other entity via a path. This way, we ended up with a connected temporal graph where the entities are the vertices and the relations are represented as edges. This condition is crucial, however, to create a timeline of the entire hospital stay in later processing stages. In contrast to i2b2 2012, we did not use direct connections with the document time or hospital stay (time). This was already covered in the named entity phase by encoding the respective Event tag. Figure 5 demonstrates the application of our annotation conventions to the sentences with already annotated named entities from Figure 4.

XSL•FO
RenderX

Figure 5. Extension of the already annotated named entities from Figure 4 with temporal relations using the BRAT annotation tool. The arrows indicate the relation together with its relation type.

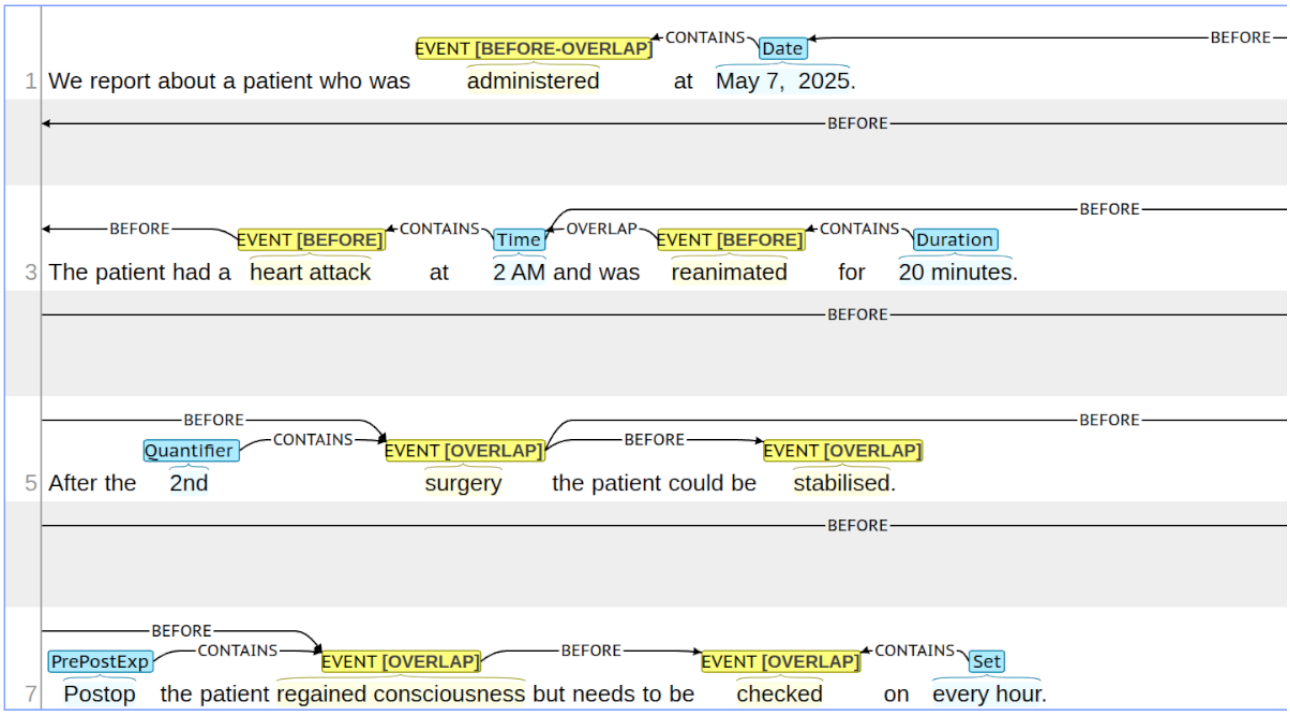

Evaluation Methods

To determine the IAA scores in our temporal named entity annotation campaign, we calculated the mean of the pairwise precision, recall, and F_1-score for the annotated entities. The implementation we used can be found on GitHub [91].

In contrast to temporal named entities, temporal relations need a different treatment for their evaluation. Every entity has temporal connections to every other entity. However, annotating the entire fully connected temporal graph is not feasible, since the number of temporal relations grows quadratically with the number of event mentions (by anchoring events in time, however, annotation of temporal relations indeed scales linearly with the number of events [92]). Thus, each annotator was encouraged to annotate the least possible number of relational edges, resulting in something akin to a minimum spanning tree. To guide the annotators towards this goal, they were instructed not only to connect named entities but to think in terms of units of meaning. These units consist of smaller groups of events, such as the timeline of a surgery, the family history, or the events that happened before admission. These smaller units should then be linked to each other, creating a larger group. This should be repeated until all larger units of meaning are linked into one large, comprehensive timeline. If possible, the anchors for these temporal units should be TIMEX3 entities, preferably Dates. A fictitious example for these linkage conventions can be found in Multimedia Appendix 1.

These conventions result in a number of different graphs that, in the worst case, might share no edges at all. This would lead to a much worse agreement score than justified. The solution we propose here is to use inference and temporal closure [93]. We infer the missing edges for each annotator and are thus able to calculate a pairwise agreement score (namely precision, recall, and F_1-score; see below). Based on the overlap of temporal relation definitions, we used a slightly adjusted version of the THYME evaluation script. The changes included adjustments to our annotation format (BRAT) [94], extending the evaluation, and the visualization of annotations.

Training and Evaluation of a Baseline Temporal Tagger

To demonstrate the usefulness of the outcome of our annotation campaign and to determine the first baseline for the automatic temporal tagging of German clinical documents, we used state-of-the-art transformer-based classification technology, that is, BERT (Bidirectional Encoder Representations from Transformers)-ish language models. For German, the most recent and best-performing language model is the German BERT model (GBERT) [95]. For our experiments, we used the base model for each of the downstream tasks we specified. To automatically optimize hyperparameters for each task, we used Bayesian Optimization and Hyperband [96] (Multimedia Appendix 2). Our experiments incorporate the huggingface library [97], as well as the Bayesian Optimization and Hyperband implementation of the Microsoft NNI framework [98]. This configuration was applied to and evaluated on the nonsharable 3000PA (ie, $3000PA_J$) and the publicly sharable GraSCCo corpus. The ground truth data used for training, testing, and evaluation were taken from the annotated documents without further correction. If the document was part of the agreement partition, and thus had been annotated by all annotators independently, we randomly selected one annotation variant from one annotator.

Evaluation Metrics

For the statistical evaluation, we apply the following metrics.

Precision (P), also known as positive predictive value, is the ratio between the gold-standard matching annotations of all

XSL•FO
RenderX

named entities or temporal relations for each label (true positives) and all annotated named entities or temporal relations for each label (true positives + false positives).



Recall (R), also called sensitivity, is the ratio between the gold-standard matching annotations of all named entities or temporal relations for each label (true positives) and all annotated named entities or temporal relations for each label (true positives and false negatives).



F_1-score is defined as the harmonic mean of precision and recall.



Accuracy (Acc) is defined as the ratio between the gold-standard matching annotations (true positives and true negatives) and all annotations, regardless of their annotations being true or false (sum of annotated entities).



All of these measurements were calculated pairwise between all possible annotator pairs. The final agreement score for the labels or the entire corpus reports the averaged pairwise agreement scores.

Results

Overview

We present the results of our experiments, distinguishing between manual temporal named entity and temporal relation annotations, as well as the evaluation of the GBERT baseline tagging model that automatically generates temporal tags for clinical documents.

Annotation of Temporal Named Entities

In the first cycle of our 2 annotation campaigns, we annotated temporal named entities on both $3000PA_J$ and GraSCCo. The results presented here were calculated on the corrected named entity annotations after the relation annotation campaign with its named entity correction subtask.

As can be seen in Table 6, we achieved pairwise F_1-scores above 0.8 for nearly all named entity types. Moreover, our overall agreement even exceeds an average F_1-score above 0.9. When looking at the token-level evaluation (which is less strict), we achieved an overall agreement average F_1-score for $3000PA_J$ of 0.912 and 0.910 for GraSCCo.

On both datasets, the entity type Duration yielded slightly worse scores. There is only one major outlier in terms of IAA in the $3000PA_J$ results, namely Quantifier, both in terms of F_1-score and SD. Interestingly, this is not observed in GraSCCo. When looking at the confusion matrix of the named entities of both datasets in Figure 6, we are able to explain this deviant behavior.

As can be seen in the confusion matrices for entity types of both corpora in Figure 6, the main reason for mismatching labels was mostly missing annotations, labeled “None” in the confusion matrix. The Event type, in particular, was often only annotated by one annotator. This behavior can be observed for both corpora. Interestingly, in $3000PA_J$, the type Quantifier was also often missed by one of the annotators. The most often confused label pair was Date and Duration. Some translated examples for confused label pairs are:

- {since 2017}DURATION vs since {2017}DATE
- {the entire night}DURATION vs the entire {night}DATE
- {after about 30 minutes}DURATION vs after about {30 minutes}TIME
- {day of self-harm}DATE vs day of {self-harm}EVENT

Some systematic errors also appear in the distinction between Duration and Date. Most of the mismatches were grounded in some form of trigger word, such as “since” or “after”. However, it needs to be stressed that these kinds of errors did not occur too often relative to all cases, as can be seen in the confusion matrices in Figure 6. Mismatches between labels other than the ones described above are very rare (per milli range).

XSL•FO
RenderX

Table 6. Absolute number and relative number in percentage of named entities, mean pairwise F1-score with SD for each entity type of the temporal named entity annotation campaign for 3000PAJ and GraSCCo, respectively, as well as the overall agreement for each dataset, which also includes the nonannotated tokens.

Named entity	3000PA$_J$				GraSCCo			
	Absolute number	Percentage (%)	F_1-score	SD	Absolute number	Percentage (%)	F_1-score	SD
Date	9099	15.13	0.903	0.055	663	11.51	0.947	0.020
DocTime	886	1.47	0.988	0.007	92	1.60	0.944	0.031
Time	593	0.99	0.893	0.058	3	0.75	0.852	0.075
Duration	824	1.38	0.776	0.083	138	2.40	0.747	0.091
Set	1107	1.84	0.822	0.081	158	2.74	0.802	0.107
Quantifier	563	0.94	0.550	0.209	118	2.05	0.938	0.023
PrePost	146	0.24	0.831	0.041	24	0.42	0.932	0.041
EVENT	46,909	78.02	0.910	0.017	4523	78.54	0.906	0.053
Total	60,127	100.0	0.834	0.068	5759	100.0	0.883	0.055
Overall Agreement	__[a]	—	0.904	0.024	—	—	0.906	0.047

[a]Not applicable.

Figure 6. Confusion matrices for temporal named entity types after correction, with 3000PAJ (left) and GraSCCo (right). The color and intensity indicate the relative number of entity pairs (with 1.0 being all and 0.0 no existing pairs). Agreeing pairs (both annotators chose the same label) can be found on the main diagonal.

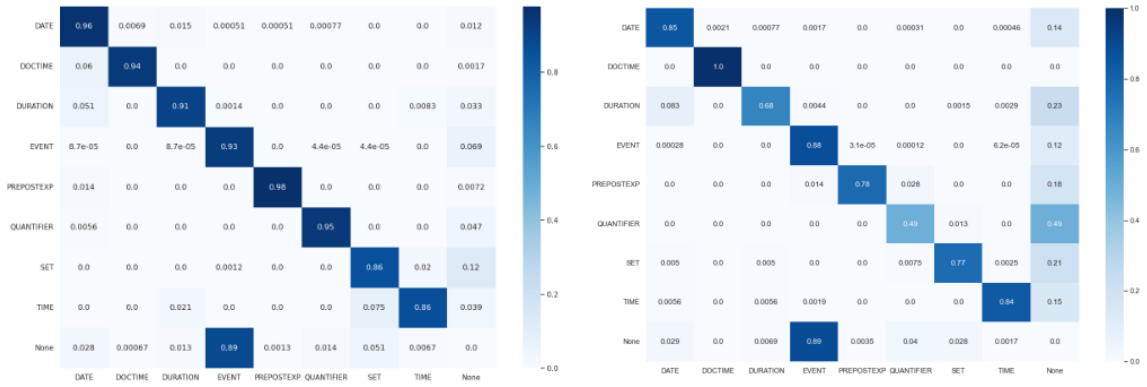

Annotation of Temporal Relations

Because of their complexity and the human efforts required for this task, temporal relations were annotated only on a randomly chosen subset of 350 documents of the original 3000PA$_J$ dataset, whereas all 63 documents of the GraSCCo dataset were dealt with. Both datasets were annotated by the same 4 annotators. The resulting IAA scores incorporating the aforementioned inference rules are depicted in Table 7. Because of the nature of points in time and their relation to each other (they form a fully connected temporal graph), relations between 2 named entity pairs always have a relation type, even though it might be automatically inferred rather than manually annotated.

As can be seen in the total count of relation types, in GraSCCo all entities seem to be connected with each other, as the number of relation types is slightly larger (40 relations more than necessary) than the total number of entity types. But in 3000PA$_J$ we seem to miss a total of about 1k relations, which means approximately 3 relations per document. However, given the total number of about 60k annotated relations, this is just 1.76% of the anticipated number of relations.

The confusion matrices for the 5 temporal relation types of both datasets are depicted in Figure 7.

The most prominent observation in both confusion matrices is the dominance of the CONTAINS relation type. Especially, OVERLAP was often confused with CONTAINS in both datasets. This is not so surprising given that CONTAINS is a subset of OVERLAP. However, it seems that the stricter CONTAINS relation is used more often compared with the less restrictive OVERLAP relation. We assume that the reason for this phenomenon might be the fact that we hired medical students rather than linguists or computer scientists (an issue often observed in other annotation campaigns, as well). Thus, our annotators have attributed the relation types a more literal meaning, “something contains something”, even partially, and

XSL•FO
RenderX

did not share our set theory-based understanding of CONTAINS and OVERLAP. Unfortunately, in the 3000PA$_J$ corpus, all of the other relation types are often confused with the CONTAINS relation type. Discussions with the annotators revealed that they used the CONTAINS relation type as a residual category, which resulted in generally coding CONTAINS if they were unsure about other alternatives. However, mismatches between other labels occur very rarely (per mille range).

Table 7. Absolute numbers and relative numbers in percentage of temporal relations, mean pairwise F1-score with SD for each annotated type of the temporal relation annotation campaign on 3000PAJ and GraSCCo using temporal closure.

Relation type	3000PA$_J$				GraSCCo			
	Absolute number	Percentage (%)	F_1-score	SD	Absolute number	Percentage (%)	F_1-score	SD
BEFORE	5948	10.07	0.2	0.07	547	9.43	0.45	0.28
BEGINS-ON	1030	1.74	0.48	0.07	146	2.52	0.64	0.19
ENDS-ON	724	1.23	0.69	0.06	74	1.28	0.73	0.14
OVERLAP	5046	8.54	0.09	0.13	679	11.71	0.33	0.36
CONTAINS	46,318	78.42	0.63	0.07	4352	75.06	0.68	0.16
Overall agreement	59,066	100.0	0.41	0.08	5798	100.0	0.57	0.27

Figure 7. Confusion matrices for temporal relation types, with 3000PAJ (left) and GraSCCo (right). The color and intensity indicate the relative number of relation label pairs (with 1.0 being all and 0.0 no existing pairs). Agreeing pairs can be found on the main diagonal.

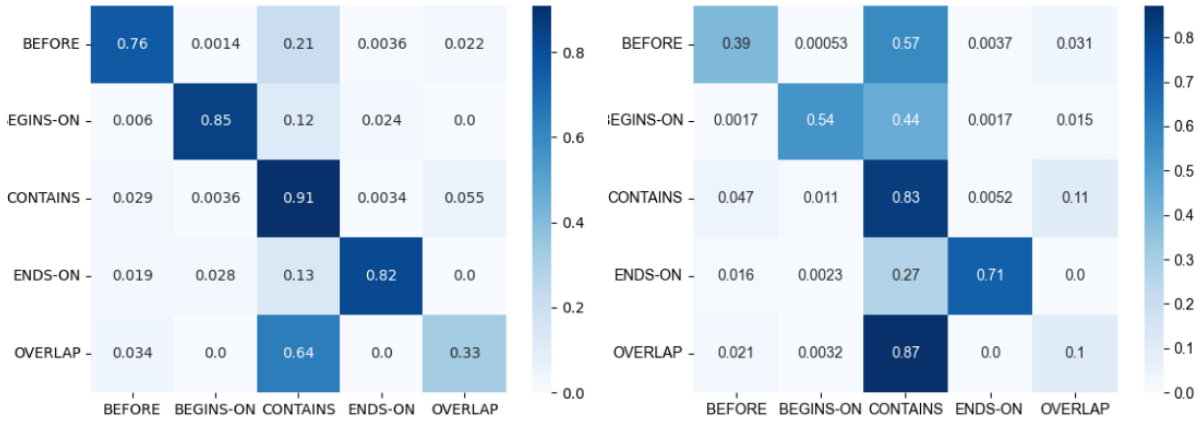

Baseline Model for Automatic Temporal Tagging

For the evaluation of the performance of the automatic temporal taggers trained on our annotated datasets, we fine-tuned and optimized a GBERT on both 3000PA$_J$ and GraSCCo. A summary of the results of our experiments can be found in Table 8.

All in all, GBERT performs well for both the named entity recognition (NER) and the relation extraction task, respectively, on the 3000PA$_J$ dataset. Training GBERT on 3000PA$_J$ and testing on GraSCCo results in an expected, yet acceptable, drop in performance. Interestingly, this drop mostly affects the recall score. The union of both datasets results in a similar performance rate compared with the model trained on 3000PA$_J$ only, yet with a slight decrease in recall and a slight increase in precision, which results in a similar F_1-score for both the named entity recognition and the relation extraction task.

XSL•FO
RenderX

Table 8. Mean pairwise accuracy (Acc), F1-score, precision, and recall for automatic temporal named entity recognition (NER) and relation extraction (REX) on 3000PAJ and GraSCCo, respectively, and a unified set of both (union).

Model	Train set	Test set	Acc	F_1-score	Precision	Recall
NER						
GBERT[a]	$3000PA_J$	$3000PA_J$	0.95	0.78	0.73	0.84
GBERT	GraSCCo	GraSCCo	0.94	0.85	0.84	0.87
GBERT	$3000PA_J$	GraSCCo	0.90	0.64	0.82	0.56
GBERT	union	union	0.95	0.77	0.75	0.80
REX						
GBERT	$3000PA_J$	$3000PA_J$	0.78	0.64	0.63	0.63
GBERT	GraSCCo	GraSCCo	0.81	0.60	0.63	0.58
GBERT	$3000PA_J$	GraSCCo	0.90	0.64	0.82	0.56
GBERT	union	union	0.81	0.63	0.70	0.59

[a]GBERT: German BERT model.

Discussion

Principal Findings

Temporal information is crucial for comprehensive medical reasoning; yet, only a few data resources are available for NLP tools to automatically extract such information from clinical documents. Publicly accessible clinical document collections have primarily emerged from challenge tasks, such as i2b2, n2c2, or Clinical TempEval, but cover the English clinical language only (Table 2). For a few non-English clinical language communities—French, Chinese (Mandarin), Spanish, Italian, Portuguese, German, Swedish, and Japanese—clinical corpora have been annotated with temporal metadata, but (except for E3C, a multilingual case report corpus) none of them is publicly available due to prohibitive data privacy regulations (Tables 4 and 8). We here introduced 2 corpora for the German clinical language annotated with TimeML-compliant metadata—a nondistributable real-world clinical corpus ($3000PA_J$), and a distributable synthetic one (GraSCCo). The first stands out with unprecedented quantitative scales in terms of the number of documents, tokens, and temporal metadata, and its adherence to qualitative standards, that is, TimeML specifications, lacking for many non-English alternatives (Table 8). The latter, though small-sized and synthetic, is one of the rare (if not the only one) non-English clinical text resources marked up for temporal information that are accessible without distribution restrictions. With reasonable agreement scores for manual annotation, we trained and tested temporal BERT-based taggers under varying experimental conditions and achieved decent performance values using both, 3000PAJ and GraSCCo.

Principal Results

In this article, we introduced 2 corpora equipped with temporal metadata for the German clinical language. The first of these corpora, $3000PA_J$, contains real clinical documents, but cannot be distributed in the public domain due to the restrictive privacy legislation codified in Germany and (more generally) Europe. To bypass this severe accessibility hurdle, we enriched a second corpus, GraSCCo, with temporal metadata, which can be distributed without any restrictions since it contains synthetic data, that is, fictitious and noise-adding content, only. The metadata we created for temporal named entities and relations, all of them fully compliant with TimeML, have a high level of credibility, given IAA scores ranging in the high 80s and low 90s.

As can also be seen from our evaluation results, we are on par performance-wise with comparable English corpora, such as THYME. For named entities and the CONTAINS relation, we even achieved better results on both corpora, $3000PA_J$ and GraSCCo. In addition, all documents used in the agreement evaluation were annotated by 4 individual annotators per document and yielded acceptable IAA scores.

In the end, we come up with about 66k temporal entities and 65k temporal relations, which sum up to 131k temporal annotation units for both corpora. These numbers not only exceed all clinical and nonclinical German-language corpora in terms of temporal metadata created up until now, but also outdo all clinical English time-focused corpora. Only the million-scale nonclinical Maven-Ere corpus [40] offers a (much) larger number of temporal annotations.

Limitations

We still have to increase the size of GraSCCo (this is ongoing work in our research group). Also, the baseline tagger needs to be complemented by more sophisticated and larger language models. We plan to experiment with temporal (knowledge graph) embeddings, which encode temporal information directly in time-enhanced embeddings [99-101] and document-level event extraction models [102-104], which will better accommodate long-range cross-sentence and cross-paragraph dependencies typical of temporal relations in long documents, such as clinical reports.

As discussed in the results section, the decision to use only medical students resulted in misconceptions regarding the definitions of the temporal relations. Potentially, mixed teams with medical students and annotators with a background in linguistics or computer science might have resulted in better

XSL•FO
RenderX

results (as small preliminary experiments for i2b2 2012 and CLEF suggest [49,105]). However, as shown by Raghavan et al [106], the level of clinical expertise has only a small impact on the inter-annotator reliability for temporal relations. Furthermore, medical students are more likely to recognize relations between clinical events [105], as well as relations based on medical causal dependence [105]. Additionally, we used the involuntary break and access restrictions during the COVID pandemic for extensive training iterations for our annotator team. Thus, if we compare our work with THYME, which was annotated by mixed teams as well, they reported similar difficulties and F_1-scores to those discussed above.

Comparison With Prior Work

To the best of our knowledge, 3000PA$_J$ and GraSCCo are the first German clinical corpora to be annotated with temporal named entities and relations fully compliant with the TimeML de facto markup standard. Surprisingly, for nonclinical documents, no comprehensive TimeML-compliant corpus currently exists for German. Furthermore, providing unconstrained access to the synthetic GraSCCo corpus constitutes a major step in metadata availability for the German clinical NLP community. Moreover, the construction principles underlying GraSCCo [82] outline a way to extend the generation (not only) of temporal metadata beyond the German language. For the first time ever, a TimeML-compliant baseline temporal tagger for German has been trained and properly validated.

Conclusions

Our work can and will be the basis for algorithms and models to automatically detect and combine temporal entities and their relations in order to create clinical timelines [107,108]. These timelines will be visualized and are thus able to provide a comprehensive summary of the patients' clinical history, a potentially valuable source of abstraction for all kinds of clinical decision makers and caretakers.

Acknowledgments

We thank Prof Dr André Scherag, Dr Danny Ammon, and all members of the Data Integration Center of the Jena University Hospital for their continuous support in managing access to 3000PA$_J$.

No artificial intelligence assistance has been used in the writing process of this paper; all text, figures, and tables have been created manually by the authors.

Funding

This work was supported by the German Ministry of Education and Research (BMBF) within the SMITH and DIFUTURE clusters of the Medical Informatics Initiative (MI-I) under the grants 01ZZ1803G and 01ZZ2009A, respectively.

Authors' Contributions

LM contributed to the conceptualization of the study, data curation, formal analysis, conduct of experiments, methodology design, provision of resources, software development, validation, visualization, and writing of both the original draft and the review and editing of the manuscript. UH was responsible for funding acquisition, provision of resources, project administration, and supervision, and also contributed to writing the original draft and reviewing and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example for relation annotation schema.
[PDF File (Adobe PDF File), 464 KB - jmir_v28i1e71458_app1.pdf]

Multimedia Appendix 2
Hyperparameter for transformer training.
[DOCX File , 9 KB - jmir_v28i1e71458_app2.docx]

References

1. Wang Y, Wang L, Rastegar-Mojarad MA, Moon SR, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
2. Bose P, Srinivasan S, Sleeman WC, Palta J, Kapoor R, Ghosh P. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. Appl Sci 2021;11(18):8319. [doi: 10.3390/app11188319]
3. Nasar Z, Jaffry SW, Malik MK. Named entity recognition and relation extraction. ACM Comput Surv 2021;54(1):1-39. [doi: 10.1145/3445965]

XSL•FO
RenderX

4. Madkour M, Benhaddou D, Tao C. Temporal data representation, normalization, extraction, and reasoning: a review from clinical domain. Comput Methods Programs Biomed 2016;128:52-68 [FREE Full text] [doi: 10.1016/j.cmpb.2016.02.007] [Medline: 27040831]
5. Leeuwenberg A, Moens MF. A survey on temporal reasoning for temporal information extraction from text. J Artif Intell Res 2019;66:341-380. [doi: 10.1613/jair.1.11727]
6. Alfattni G, Peek N, Nenadic G. Extraction of temporal relations from clinical free text: a systematic review of current approaches. J Biomed Inform 2020;108:103488 [FREE Full text] [doi: 10.1016/j.jbi.2020.103488] [Medline: 32673788]
7. Gumiel YB, Silva e Oliveira LE, Claveau V, Grabar N, Paraiso EC, Moro C, et al. Temporal relation extraction in clinical texts. ACM Comput Surv 2021;54(7):1-36. [doi: 10.1145/3462475]
8. Hahn U. Clinical document corpora-real ones, translated and synthetic substitutes, and assorted domain proxies: a survey of diversity in corpus design, with focus on German text data. JAMIA Open 2025;8(3):ooaf024. [doi: 10.1093/jamiaopen/ooaf024] [Medline: 40371384]
9. Sun W, Rumshisky A, Uzuner Ö. Temporal reasoning over clinical text: the state of the art. J Am Med Inform Assoc 2013;20(5):814-819 [FREE Full text] [doi: 10.1136/amiajnl-2013-001760] [Medline: 23676245]
10. Kamsu-Foguem B, Tchuenté-Foguem G, Foguem C. Verifying a medical protocol with temporal graphs: the case of a nosocomial disease. J Crit Care 2014;29(4):690.e1-690.e9. [doi: 10.1016/j.jcrc.2014.02.006] [Medline: 24636931]
11. Wu STI, Juhn YJ, Sohn S, Liu H. Patient-level temporal aggregation for text-based asthma status ascertainment. J Am Med Inform Assoc 2014;21(5):876-884 [FREE Full text] [doi: 10.1136/amiajnl-2013-002463] [Medline: 24833775]
12. Caballero Barajas KL, Akella R. Dynamically modeling patient's health state from electronic medical records: a time series approach. : Association for Computing Machinery (ACM); 2015 Presented at: KDD - 15 - Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2015 Aug 10; Sydney, New South Wales, Australia p. 69-78. [doi: 10.1145/2783258.2783289]
13. Jo Y, Loghmanpour N, Penstein-Rosé C. Time series analysis of nursing notes for mortality prediction via a state transition topic model. : Association for Computing Machinery (ACM); 2015 Presented at: CIKM - 15 - Proceedings of the 24th ACM International Conference on Information and Knowledge Management; 2015 Oct 19; Melbourne, Victoria, Australia p. 1171-1180. [doi: 10.1145/2806416.2806541]
14. Qiao Z, Zhao S, Xiao C, Li X, Qin Y, Wang F. Pairwise-ranking based collaborative recurrent neural networks for clinical event prediction. 2018 Presented at: IJCAI-ECAI '18 - Proceedings of the 27th International Joint Conference on Artificial Intelligence & 23rd European Conference on Artificial Intelligence; 2018 July 13; Stockholm, Sweden p. 3520-3526. [doi: 10.24963/ijcai.2018/489]
15. Khadanga S, Aggarwal K, Joty S, Srivastava J. Using clinical notes with time series data for ICU management. : Association for Computational Linguistics (ACL); 2019 Presented at: EMNLP-IJCNLP - Proceedings of the Conference on Empirical Methods in Natural Language Processing & 9th International Joint Conference on Natural Language Processing; November ; Hong Kong, China. Association for Computational Linguistics (ACL); 2019 Nov 03; Hong Kong, China p. 6432-6437 URL: https://aclanthology.org/D19-1678/ [doi: 10.18653/v1/d19-1678]
16. Najafabadipour M, Zanin M, Rodríguez-González A, Torrente M, Nuñez García B, Cruz Bermudez JL, et al. Reconstructing the patient's natural history from electronic health records. Artif Intell Med 2020;105:101860. [doi: 10.1016/j.artmed.2020.101860] [Medline: 32505419]
17. Deznabi I, Iyyer M, Fiterau M. Predicting in-hospital mortality by combining clinical notes with time-series data. : Association for Computational Linguistics (ACL); 2021 Presented at: Findings of the Association for Computational Linguistics - ACL-IJCNLP 2021; 2021 Aug 01; USA p. 4026-4031 URL: https://aclanthology.org/2021.findings-acl.352/ [doi: 10.18653/v1/2021.findings-acl.352]
18. Olex AL, McInnes BT. Review of temporal reasoning in the clinical domain for timeline extraction: where we are and where we need to be. J Biomed Inform 2021;118:103784 [FREE Full text] [doi: 10.1016/j.jbi.2021.103784] [Medline: 33862232]
19. Xie F, Yuan H, Ning Y, Ong MEH, Feng M, Hsu W, et al. Deep learning for temporal data representation in electronic health records: a systematic review of challenges and methodologies. J Biomed Inform 2022;126:103980 [FREE Full text] [doi: 10.1016/j.jbi.2021.103980] [Medline: 34974189]
20. Zhao D, Shi Y, Cheng L, Li H, Zhang L, Guo H. Time interval uncertainty-aware and text-enhanced based disease prediction. J Biomed Inform 2023;139:104239 [FREE Full text] [doi: 10.1016/j.jbi.2022.104239] [Medline: 36356933]
21. Allen JF. Towards a general theory of action and time. Artif Intell 1984;23(2):123-154. [doi: 10.1016/0004-3702(84)90008-0]
22. Pustejovsky JD, Castaño J, Ingria RJP, Saurí R, Gaizauskas RJ, Setzer A. TIMEML: robust specification of event and temporal expressions in text. In: New Directions in Question Answering. Washington, D.C: AAAI Press; 2003:28-34.
23. Pustejovsky JD, Lee K, Bunt HC, Romary L. ISO-TIMEML : an international standard for semantic annotation. : European Language Resources Association (ELRA); 2010 Presented at: LREC 2010 - Proceedings of the 7th International Conference on Language Resources and Evaluation; 2010 May 17; La Valletta, Malta p. 394-397 URL: https://aclanthology.org/L10-1027/

XSL•FO
RenderX

24. Ferro L, Gerber L, Mani I, Sundheim BM, Wilson G. TIDES 2005 standard for the annotation of temporal expressions. Linguistic Data Consortium. 2005. URL: https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-timex2-guidelines-v0.1.pdf [accessed 2025-08-14]
25. Pustejovsky JD, Hanks P, Saurí R, See A, Gaizauskas RJ, Setzer A. The TIMEBANK corpus. 2003 Presented at: Proceedings of the Corpus Linguistics 2003 Conference; 2003 March 28; Lancaster, UK p. 647-656.
26. Boguraev BK, Pustejovsky JD, Ando RK, Verhagen M. TimeBank evolution as a community resource for TimeML parsing. Lang Resour Evaluation 2007;41(1):91-115. [doi: 10.1007/s10579-007-9018-8]
27. Ocal M, Radas A, Hummer J, Megerdoomian K, Finlayson MA. A comprehensive evaluation and correction of the Time Bank corpus. : European Language Resources Association (ELRA); 2022 Presented at: Proceedings of the 13th International Conference on Language Resources and Evaluation; 2022 June 20; Marseille, France p. 20-25 URL: https://aclanthology.org/2022.lrec-1.313/
28. Cassidy T, McDowell B, Chambers N, Bethard SJ. An annotation framework for dense event ordering. : Short Papers; 2014 Presented at: ACL 2014 - Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics; 2014 June 22; Baltimore, Maryland, USA p. 501-506. [doi: 10.3115/v1/p14-2082]
29. Chambers N, Cassidy T, McDowell B, Bethard SJ. Dense event ordering with a multi-pass architecture. Trans Assoc Comput Linguist 2014;2:273-284. [doi: 10.1162/tacl_a_00182]
30. Verhagen M, Gaizauskas RJ, Schilder F, Hepple M, Moszkowicz JL, Pustejovsky JD. The TempEval challenge: identifying temporal relations in text. Lang Resour Evaluation 2009;43(2):161-179. [doi: 10.1007/s10579-009-9086-z]
31. UzZaman N, Llorens H, Derczynski LRA, Verhagen M, Allen JF, Pustejovsky JD. SemEval-2013 Task 1: TempEval-3 - evaluating time expressions, events, and temporal relations. : Association for Computational Linguistics (ACL); 2013 Presented at: SEM 2013 - Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics; 2013 June 14; Atlanta, Georgia, USA p. 14-15 URL: https://aclanthology.org/S13-2001/
32. Verhagen M, Saurí R, Caselli T, Pustejovsky JD. SemEval-2010 Task 13: TempEval-2 - Evaluating events, time expressions, and temporal relations. : Association for Computational Linguistics (ACL); 2010 Presented at: SemEval 2010 - Proceedings of the 5th International Workshop on Semantic Evaluation @ ACL 2010; 2010 July 15; Uppsala, Sweden p. 57-62 URL: https://dl.acm.org/doi/10.5555/1859664.1859674
33. Graff D. The AQUAINT corpus of English news text. International Standard Language Resource Number. 2002. URL: https://www.islrn.org/resources/153-002-267-999-9/ [accessed 2026-01-06]
34. Minard AL, Speranza M, Agirre E, Aldabe I, van Erp M, Magnini B. Semeval-2015 Task 4: TimeLine - cross-document event ordering. : Association for Computational Linguistics (ACL); 2015 Presented at: SemEval 2015 - Proceedings of the 9th International Workshop on Semantic Evaluation @ NAACL-HLT 2015; 2015 June 04; Denver, Colorado, USA p. 778-786. [doi: 10.18653/v1/s15-2132]
35. O'Gorman T, Wright-Bettner K, Palmer M. Richer event description: integrating event coreference with temporal, causal and bridging annotation. : Curran Associates, Inc; 2016 Presented at: CNewsStory 2016 - Proceedings of the 2nd Workshop on Computing News Storylines @ EMNLP 2016; 2016 Nov 05; Austin, Texas, USA p. 47-56. [doi: 10.18653/v1/w16-5706]
36. Mostafazadeh N, Grealish A, Chambers N, Allen J, Vanderwende L. CaTeRS : causal and temporal relation scheme for semantic annotation of event structures. : Association for Computational Linguistics (ACL); 2016 Presented at: EVENTS - Proceedings of the 4th Workshop on Events: Definition, Detection, Coreference, and Representation @ NAACL-HLT 2016; 2016 June 17; San Diego, California p. 51-61. [doi: 10.18653/v1/w16-1007]
37. Caselli T, Vossen P. The Event StoryLine corpus: a new benchmark for causal and temporal relation extraction. : Association for Computational Linguistics (ACL); 2017 Presented at: EventStory 2017 - Proceedings of the [1st] Workshop on Events and Stories in the News @ ACL 2017; 2017 Aug 04; Vancouver, British Columbia, Canada p. 77-86. [doi: 10.18653/v1/w17-2711]
38. Silveira N, Doza T, de Marneffe MC, Bowman S, Connor M, Bauer J, et al. A gold standard dependency corpus for English. : European Language Resources Association (ELRA); 2014 Presented at: LREC 2014 - Proceedings of the 9th International Conference on Language Resources and Evaluation; 2014 May 26; Reykjavik, Iceland p. 2897-2904 URL: https://aclanthology.org/L14-1067/
39. Yao J, Qiu H, Min B, Xue N. Annotating temporal dependency graphs via crowdsourcing. : Association for Computational Linguistics (ACL); 2020 Presented at: EMNLP 2020 - Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing; 2020 Nov 16; USA p. 5368-5380. [doi: 10.18653/v1/2020.emnlp-main.432]
40. Wang X, Chen Y, Ding N, Peng H, Wang Z, Lin Y, et al. Maven-Ere: a unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. : Association for Computational Linguistics (ACL); 2022 Presented at: EMNLP 2022 - Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; 2022 Dec 07; Abu Dhabi, United Arab Emirates p. 926-941. [doi: 10.18653/v1/2022.emnlp-main.60]
41. Rogers A, Karpinska M, Gupta A, Lialin V, Smelkov G, Rumshisky A. Narrative Time : dense temporal annotation on a timeline. : European Language Resources Association (ELRA); 2024 Presented at: LREC-COLING 2024 - Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation; 2024 May 20; Torino, Italia p. 12053-12073 URL: https://aclanthology.org/2024.lrec-main.1054/

XSL•FO
RenderX

42. Verhagen M, Gaizauskas RJ, Schilder F, Hepple M, Katz G, Pustejovsky J. SemEval-2007 Task 15: TempEval - Temporal relation identification. : Association for Computational Linguistics (ACL); 2007 Presented at: SemEval 2007 - Proceedings of the 4th International Workshop on Semantic Evaluations @ ACL 2007; 2007 June 23; Prague, Czech Republic p. 75-80 URL: https://aclanthology.org/S07-1014/ [doi: 10.3115/1621474.1621488]
43. Llorens H, Chambers N, UzZaman N, Mostafazadeh N, Allen J, Pustejovsky J. SemEval-2015 task 5: QA TEMPEVAL - evaluating temporal information understanding with question answering. : Curran Associates, Inc; 2015 Presented at: SemEval 2015 - Proceedings of the 9th International Workshop on Semantic Evaluation @ NAACL-HLT 2015; 2015 June 04; Denver, Colorado, USA p. 792-800. [doi: 10.18653/v1/s15-2134]
44. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner Ö. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc 2011;18(5):540-543 [FREE Full text] [doi: 10.1136/amiajnl-2011-000465] [Medline: 21846785]
45. Filannino M, Uzuner Ö. Advancing the state of the art in clinical natural language processing through shared tasks. Yearb Med Inform 2018;27(1):184-192 [FREE Full text] [doi: 10.1055/s-0038-1667079] [Medline: 30157522]
46. Vashishtha S, Van Durme B, White AS. Fine-grained temporal relation extraction. : Association for Computational Linguistics (ACL); 2019 Presented at: ACL 2019 - Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; - August 2; Florence, Italy. Association for Computational Linguistics (ACL). 2019; 2019 July 28; Florence, Italy p. 2906-2919. [doi: 10.18653/v1/p19-1280]
47. Roberts A, Gaizauskas RJ, Hepple M, Demetriou G, Guo Y, Roberts I, et al. Building a semantically annotated corpus of clinical texts. J Biomed Inform 2009;42(5):950-966 [FREE Full text] [doi: 10.1016/j.jbi.2008.12.013] [Medline: 19535011]
48. Sun W, Rumshisky A, Uzuner Ö. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. J Am Med Inform Assoc 2013;20(5):806-813 [FREE Full text] [doi: 10.1136/amiajnl-2013-001628] [Medline: 23564629]
49. Sun W, Rumshisky A, Uzuner Ö. Annotating temporal information in clinical narratives. J Biomed Inform 2013;46:S5-S12 [FREE Full text] [doi: 10.1016/j.jbi.2013.07.004] [Medline: 23872518]
50. Styler WF, Bethard SJ, Finan S, Palmer M, Pradhan S, de Groen PC, et al. Temporal annotation in the clinical domain. Trans Assoc Comput Linguist 2014;2:143-154 [FREE Full text] [Medline: 29082229]
51. Bethard S, Derczynski L, Savova G, Pustejovsky J, Verhagen M. SemEval-2015 task 6: clinical TempEval. : Curran Associates; 2015 Presented at: SemEval 2015 - Proceedings of the 9th International Workshop on Semantic Evaluation @ NAACL-HLT 2015; 2015 June 04; Denver, Colorado, USA p. 806-814. [doi: 10.18653/v1/s15-2136]
52. Bethard SJ, Savova G, Chen WT, Derczynski L, Pustejovsky J, Verhagen M. SemEval-2016 task 12: clinical TempEval. : Association for Computational Linguistics (ACL); 2016 Presented at: SemEval 2016 - Proceedings of the 10th International Workshop on Semantic Evaluation @ NAACL-HLT 2016; 2016 June 16; San Diego, California, USA p. 1052-1062. [doi: 10.18653/v1/s16-1165]
53. Bethard S, Savova G, Palmer M, Pustejovsky J. SemEval-2017 task 12: clinical TempEval. : Association for Computational Linguistics (ACL); 2017 Presented at: SemEval 2017 - Proceedings of the 11th International Workshop on Semantic Evaluation @ ACL 2017; 2017 Aug 03; Vancouver, British Columbia, Canada p. 565-572. [doi: 10.18653/v1/s17-2093]
54. Mahajan D, Liang JJ, Tsou CH, Uzuner Ö. Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes. J Biomed Inform 2023;144:104432 [FREE Full text] [doi: 10.1016/j.jbi.2023.104432] [Medline: 37356640]
55. Kumar V, Stubbs A, Shaw SY, Uzuner Ö. Creation of a new longitudinal corpus of clinical narratives. J Biomed Inform 2015;58(Suppl):S6-S10 [FREE Full text] [doi: 10.1016/j.jbi.2015.09.018] [Medline: 26433122]
56. Viani N, Tissot H, Bernardino A, Velupillai S. Annotating temporal information in clinical notes for timeline reconstruction: towards the definition of calendar expressions. 2019 Presented at: BioNLP 2019 - Proceedings of the 18th SIGBioMed Workshop on Biomedical Natural Language Processing and Shared Task @ ACL 2019; 2019 Aug 01; Florence, Italy p. 201-210. [doi: 10.18653/v1/w19-5021]
57. Spreyer K, Frank A. Projection-based acquisition of a temporal labeller. : Asian Federation of Natural Language Processing (AFNLP); 2008 Presented at: IJCNLP 2008 - Proceedings of the 3rd International Joint Conference on Natural Language Processing; 2008 Jan 07; Hyderabad, India p. 489-496 URL: https://aclanthology.org/I08-1064/
58. Strötgen J, Gertz M. WikiWarsDE: a German corpus of narratives annotated with temporal expressions. In: Multilingual Resources and Multilingual Applications.: Hamburger Zentrum für Sprachkorpora; 2011 Presented at: GSCL 2011 - Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology; 2011 Sep 28; Hamburg, Germany p. 129-134 URL: http://dbs.ifi.uni-heidelberg.de/fileadmin/Team/jannik/publications/StroetgenGertz_WikiWarsDE_GSCL2011.pdf
59. Mazur P, Dale R. WIKIWARS : a new corpus for research on temporal expressions. : Association for Computational Linguistics (ACL); 2010 Presented at: EMNLP 2010 - Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing; 2010 Oct 09; Boston, Massachusetts, USA p. 913-922 URL: https://aclanthology.org/D10-1089/
60. Strötgen J, Minard A, Lange L, Speranza M, Magnini B. KRAUTS: a German temporally annotated news corpus. : European Language Resources Association (ELRA); 2018 Presented at: LREC 2018 - Proceedings of the 11th International Conference on Language Resources and Evaluation; 2018 May 07; Miyazaki, Japan p. 536-540 URL: https://aclanthology.org/L18-1085/

XSL•FO
RenderX

61. Strötgen J, Gertz M. Multilingual and cross-domain temporal tagging. Lang Resour Evaluation 2012;47(2):269-298. [doi: 10.1007/s10579-012-9179-y]
62. Lücking A, Stoeckel M, Abrami G, Mehler A. I still have time(s): extending HeidelTime for German texts. : European Language Resources Association (ELRA); 2022 Presented at: LREC 2022 - Proceedings of the 13th International Conference on Language Resources and Evaluation; 2022 June 20; Marseille, France p. 4723-4728 URL: https://aclanthology.org/2022.lrec-1.505/
63. Roller R, Uszkoreit H, Xu F, Seiffe L, Mikhailov M, Staeck O. A fine-grained corpus annotation schema of German nephrology records. : COLING 2016 Organizing Committee; 2016 Presented at: ClinicalNLP 2016 - Proceedings of the 1st Workshop on Clinical Natural Language Processing @ COLING 2016; 2016 December 11; Osaka, Japan p. 69-77 URL: https://aclanthology.org/W16-4210/
64. Campillos L, Deléger L, Grouin C, Hamon T, Ligozat A, Névéol A. A French clinical corpus with comprehensive semantic annotations: development of the medical entity and relation LIMSI annOtated Text corpus (MERLOT). Lang Resour Evaluation 2017;52(2):571-601. [doi: 10.1007/s10579-017-9382-y]
65. Hamon T, Grabar N. Tuning HeidelTime for identifying time expressions in clinical texts in English and French. 2014 Presented at: LOUHI 2014 - Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis @ EACL 2014; 2014 April 27; Gothenburg, Sweden p. 101-105. [doi: 10.3115/v1/w14-1116]
66. Bannour N, Rance B, Tannier X, Névéol A. Event-independent temporal positioning: application to French clinical text. 2023 Presented at: BioNLP 2023 - Proceedings of the 22nd Workshop on Biomedical Language Processing (BioNLP) & BioNLP Shared Tasks (BioNLP-ST) @ ACL 2023; 2023 July 13; Ontario, Canada p. 191-205. [doi: 10.18653/v1/2023.bionlp-1.16]
67. Jeyafreeda AJ, Vincent M, Burgun A, Garcelon N. Evaluating LLMs for temporal entity extraction from pediatric clinical text in rare diseases context. 2024 Presented at: CL4Health 2024 - Proceedings of the 1st Workshop on Patient-Oriented Language Processing @ LREC-COLING 2024; 2024 May 20; Torino, Italia p. 145-152 URL: https://aclanthology.org/2024.cl4health-1.18/
68. Liu Z, Wang X, Chen Q, Tang B, Xu H. Temporal indexing of medical entity in Chinese clinical notes. BMC Med Inform Decis Mak 2019;19(Suppl 1):17 [FREE Full text] [doi: 10.1186/s12911-019-0735-x] [Medline: 30700331]
69. Pan X, Chen B, Weng H, Gong Y, Qu Y. Temporal expression classification and normalization from Chinese narrative clinical texts: pattern learning approach. JMIR Med Inform 2020;8(7):e17652 [FREE Full text] [doi: 10.2196/17652] [Medline: 32716307]
70. Liu Z, Tang B, Wang X, Chen Q, Li H, Bu J, et al. CMedTEX: a rule-based temporal expression extraction and normalization system for Chinese clinical notes. AMIA Annu Symp Proc 2016;2016:818-826 [FREE Full text] [Medline: 28269878]
71. Hu D, Wang M, Gao F, Xu F, Gu J. Knowledge representation and reasoning for complex time expression in clinical text. Data Intell 2022;4(3):573-598. [doi: 10.1162/dint_a_00152]
72. Velupillai S. Temporal expressions in Swedish medical text: a pilot study. 2014 Presented at: BioNLP 2014 - Proceedings of the Workshop on Biomedical Natural Language Processing @ ACL 2014; 2024 June 27; Baltimore, Maryland, USA p. 88-92. [doi: 10.3115/v1/w14-3413]
73. de Azevedo RF, Santos Rodrigues JP, da Silva Reis MR, Cabral Moro CM, Cabrera Paraiso E. Temporal tagging of noisy clinical texts in Brazilian Portuguese. In: Computational Processing of the Portuguese Language. 2018 Presented at: PROPOR 2018 - Proceedings of the 13th International Conference on Computational Processing of the Portuguese Language; 2018 Sep 24; Canela, Brazil p. 231-241. [doi: 10.1007/978-3-319-99722-3_24]
74. Viani N, Miller TA, Napolitano C, Priori SG, Savova GK, Bellazzi R, et al. Supervised methods to extract clinical events from cardiology reports in Italian. J Biomed Inform 2019;95:103219 [FREE Full text] [doi: 10.1016/j.jbi.2019.103219] [Medline: 31150777]
75. Cheng F, Yada S, Tanaka R, Aramaki E, Kurohashi S. JaMIE: a pipeline Japanese medical information extraction system with novel relation annotation. 2022 Presented at: LREC 2022 - Proceedings of the 13th International Conference on Language Resources and Evaluation; 2022 June 20; Marseille, France p. 3724-3731 URL: https://aclanthology.org/2022.lrec-1.397/
76. Shimizu S, Kanashiro PL, Yada S, Aramaki E. QA-based event start-points ordering for clinical temporal relation annotation. 2024 Presented at: LREC-COLING 2024 - Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation; 2024 May 20; Torino, Italia p. 13371-13381 URL: https://aclanthology.org/2024.lrec-main.1171/
77. Magnini B, Altuna B, Lavelli A, Minard A, Speranza M, Zanoli R. European clinical case corpus. In: Rehm G, editor. European Language Grid: a Language Technology Platform for Multilingual Europe. New York: Springer; 2023:283-288.
78. Hahn U, Matthies F, Lohr C, Löffler M. 3000PA-towards a national reference corpus of German clinical language. Stud Health Technol Inform 2018;247:26-30. [Medline: 29677916]
79. Lohr C, Luther S, Matthies F, Modersohn L, Ammon D, Saleh K, et al. CDA-compliant section annotation of German-language discharge summaries: guideline development, annotation campaign, section classification. AMIA Annu Symp Proc 2018;2018:770-779 [FREE Full text] [Medline: 30815119]

XSL•FO
RenderX

80. Kolditz T, Lohr C, Hellrich J, Modersohn L, Betz B, Kiehntopf M. Annotating German clinical documents for de-identification. Stud Health Technol Inform 2019;264:203-207. [doi: 10.3233/shti190212]
81. Lohr C, Modersohn L, Hellrich J, Kolditz T, Hahn U. An evolutionary approach to the annotation of discharge summaries. Stud Health Technol Inform 2020;270:28-32. [doi: 10.3233/shti200116]
82. Modersohn L, Schulz S, Lohr C, Hahn U. GraSCCo: the first publicly shareable, multiply-alienated German clinical text corpus. Stud Health Technol Inform 2022;270:66-72. [doi: 10.3233/shti220805]
83. Schulz S. GraSCCo. Zenodo (CERN European Organization for Nuclear Research). 2022. URL: https://doi.org/10.5281/zenodo.6539131 [accessed 2025-08-14]
84. Kweon S, Kim J, Kim J, Im S, Cho E, Bae S, et al. Publicly shareable clinical large language model built on synthetic clinical notes. : Association for Computational Linguistics (ACL); 2024 Presented at: Findings of the Association for Computational Linguistics - ACL 2024; 2024 Aug 11; Bangkok, Thailand p. 5148-5168. [doi: 10.18653/v1/2024.findings-acl.305]
85. Li J, Zhou Y, Jiang X, Natarajan K, Pakhomov SVS, Liu H, et al. Are synthetic clinical notes useful for real natural language processing tasks: a case study on clinical entity recognition. J Am Med Inform Assoc 2021;28(10):2193-2201 [FREE Full text] [doi: 10.1093/jamia/ocab112] [Medline: 34272955]
86. Ive J, Viani N, Kam J, Yin L, Verma S, Puntis S, et al. Generation and evaluation of artificial mental health records for natural language processing. NPJ Digit Med 2020;3(69):1-9 [FREE Full text] [doi: 10.1038/s41746-020-0267-x] [Medline: 32435697]
87. Hiebel N, Ferret O, Fort K, Névéol A. Can synthetic text help clinical named entity recognition? a study of electronic health records in French. : Association for Computational Linguistics (ACL); 2023 Presented at: EACL 2023 - Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics; 2023 May 06; Dubrovnik, Croatia p. 2320-2338. [doi: 10.18653/v1/2023.eacl-main.170]
88. Brekke PH, Rama T, Pilán I, Nytrø Ø, Øvrelid L. Synthetic data for annotation and extraction of family history information from clinical text. J Biomed Semantics 2021;12(1):11 [FREE Full text] [doi: 10.1186/s13326-021-00244-2] [Medline: 34261535]
89. Pustejovsky J, Stubbs A. Natural Language Annotation for Machine Learning - A Guide to corpus-building for Applications. California: O'Reilly Media, Inc; 2012.
90. Artstein R, Poesio M. Inter-coder agreement for computational linguistics. Comput Linguist 2008;34(4):555-596. [doi: 10.1162/coli.07-034-R2]
91. Kolditz T. Bratiaa: inter-annotator agreement for brat annotation projects. GitHub. 2019. URL: https://github.com/kldtz/bratiaa [accessed 2025-08-14]
92. Reimers N, Dehghani N, Gurevych I. Temporal anchoring of events for the TIMEBANK corpus. : Association for Computational Linguistics (ACL); 2016 Presented at: ACL 2016 - Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics; 2016 Aug 07; Berlin, Germany p. 2195-2204. [doi: 10.18653/v1/p16-1207]
93. Verhagen M. Temporal closure in an annotation environment. Language Res Eval 2006;39(2-3):211-241. [doi: 10.1007/s10579-005-7884-5]
94. Stenetorp P, Pyysalo S, Topi G, Ohta T, Ananiadou S, Tsujii J. BRAT: a Web-based tool for NLP-assisted text annotation. : Association for Computational Linguistics (ACL); 2012 Presented at: EACL 2012 - Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations; 2012 Apr 25; Avignon, France p. 25-26 URL: https://aclanthology.org/E12-2021/
95. Chan B, Schweter S, Möller T. German's next language model. : International Committee on Computational Linguistics (ICCL); 2020 Presented at: COLING 2020 - Proceedings of the 28th International Conference on Computational Linguistics; 2020 Dec 08; Barcelona, Spain p. 6788-6796. [doi: 10.18653/v1/2020.coling-main.598]
96. Falkner S, Klein A, Hutter F. BOHB: robust and efficient hyperparameter optimization at scale. : Proceedings of Machine Learning Research; 2018 Presented at: ICML 2018 - Proceedings of the 35th International Conference on Machine Learning; 2018 July 10; Stockholm, Sweden p. 1437-1446 URL: https://proceedings.mlr.press/v80/falkner18a.html
97. Huggingface: Huggingface/Transformers. GitHub. 2020. URL: https://github.com/huggingface/transformers [accessed 2025-08-14]
98. Microsoft: neural network intelligence. GitHub. 2021. URL: https://github.com/microsoft/nni [accessed 2025-08-14]
99. Goyal T, Durrett G. Embedding time expressions for deep temporal ordering models. : Association for Computational Linguistics (ACL); 2019 Presented at: ACL 2019 - Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Aug 02; Florence, Italy p. 4400-4406. [doi: 10.18653/v1/p19-1433]
100. Han Z, Zhang G, Ma Y, Tresp V. Time-dependent entity embedding is not all you need: a re-evaluation of temporal knowledge graph completion models under a unified framework. : Association for Computational Linguistics (ACL); 2021 Presented at: EMNLP 2021 - Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; 2021 Nov 07; Punta Cana, Dominican Republic p. 8104-8118. [doi: 10.18653/v1/2021.emnlp-main.639]
101. Gao Y, He Y, Kan Z, Han Y, Qiao L, Li D. Learning joint structural and temporal contextualized knowledge embeddings for temporal knowledge graph completion. : Association for Computational Linguistics (ACL); 2023 Presented at: Findings

XSL•FO
RenderX

of the Association for Computational Linguistics - ACL 2023; 2023 July 09; Toronto, Ontario, Canada p. 417-430. [doi: 10.18653/v1/2023.findings-acl.28]
102. Mathur P, Jain R, Dernoncourt F, Morariu V, Tran QH, Manocha D. TIMERS: document-level temporal relation extraction. : Association for Computational Linguistics (ACL); 2021 Presented at: ACL-IJCNLP 2021 - Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics & 11th International Joint Conference on Natural Language Processing; 2021 Aug 01; Stroudsburg, PA p. 524-533. [doi: 10.18653/v1/2021.acl-short.67]
103. Sun Q, Zhang K, Huang K, Xu T, Li X, Liu Y. Document-level relation extraction with two-stage dynamic graph attention networks. Knowl Based Syst 2023;267:110428. [doi: 10.1016/j.knosys.2023.110428]
104. Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. A comparative study of pretrained language models for long clinical text. J Am Med Inform Assoc 2023;30(2):340-347 [FREE Full text] [doi: 10.1093/jamia/ocac225] [Medline: 36451266]
105. Roberts A, Gaizauskas R, Hepple M, Davis N, Demetriou G, Guo Y, et al. The CLEF corpus: semantic annotation of clinical text. AMIA Annu Symp Proc 2007;2007:625-629 [FREE Full text] [Medline: 18693911]
106. Raghavan P, Fosler-Lussier E, Lai AM. Inter-annotator reliability of medical events, coreferences and temporal relations in clinical narratives by annotators with varying levels of clinical expertise. AMIA Annu Symp Proc 2012;2012:1366-1374 [FREE Full text] [Medline: 23304416]
107. Bastardo R, Castro M, Pavão JA, Ramos L. A timeline model for clinical events: empowering data. Procedia Comput Sci 2022;196:1053-1058. [doi: 10.1016/j.procs.2021.12.109]
108. Singh A, Hummer J, Ocal M, Finlayson MA. PYTLEX : A PYTHON library for TimeLine extraction. : Association for Computational Linguistics (ACL); 2024 Presented at: EACL 2024 - Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations; 2024 March 17; St. Julian's, Malta p. 27-34. [doi: 10.18653/v1/2024.eacl-demo.4]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
GBERT: German BERT model
IAA: interannotator agreement
ISO: International Organization for Standardization
NLP: natural language processing

Edited by J Sarvestan; submitted 18.Jan.2025; peer-reviewed by C Li, G Recski, K Arzideh; comments to author 15.Feb.2025; revised version received 30.Dec.2025; accepted 31.Dec.2025; published 25.Feb.2026.

Please cite as:
Modersohn L, Hahn U
Temporal Annotation of German Clinical Language in Real and Synthetic Clinical Documents: Corpus Development and Baseline Tagger Validation Study
J Med Internet Res 2026;28:e71458
URL: https://www.jmir.org/2026/1/e71458
doi:10.2196/71458
PMID:

©Luise Modersohn, Udo Hahn. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Association of Electronic Health Literacy With Self-Care and Health Outcomes Among Patients With Type 2 Diabetes Mellitus: Cross-Sectional Study

Phoenix Kit-Han Mo[1], PhD; Alice PS Kong[2], MD; Luyao Xie[1], PhD; Virginia WY Chan[1,3], PhD; Joseph TF Lau[4,5], PhD

[1]Centre of Health Behaviours Research, Jockey Club School of Public Health and Primary Care, Prince of Wales Hospital, Chinese University of Hong Kong, 202D, Shatin, N.T., Hong Kong, China

[2]Department of Medicine and Therapeutics, Chinese University of Hong Kong, Hong Kong, China

[3]Community Mental Health Association, Hong Kong, China

[4]Public Mental Health Center, School of Mental Health, Wenzhou Medical University, Wenzhou, China

[5]Zhejiang Provincial Clinical Research Center for Mental Disorders, The Affiliated Wenzhou Kangning Hospital, Wenzhou Medical University, Wenzhou, China

Corresponding Author:
Phoenix Kit-Han Mo, PhD
Centre of Health Behaviours Research, Jockey Club School of Public Health and Primary Care, Prince of Wales Hospital, Chinese University of Hong Kong, 202D, Shatin, N.T., Hong Kong, China

Abstract

Background: Diabetes mellitus (DM) continues to be a critical public health issue in Hong Kong. Although self-care behaviors help promote health among patients with DM, adherence remains suboptimal. More attention should be paid to eHealth literacy with the development of modern technologies.

Objective: This study aims to assess the level of eHealth literacy among patients with DM and examine its association with self-care and health outcomes.

Methods: A cross-sectional study was conducted among patients with type 2 DM from the DM clinic of a public hospital in Hong Kong. Data on eHealth literacy, self-care, self-care self-efficacy, diabetes distress, glycated hemoglobin (HbA_{1c}) control, and sociodemographic information were collected. Multivariable regression analyses were performed, adjusting for relevant sociodemographic and medical variables.

Results: Among the 427 patients with DM recruited, around two-thirds (65.1%) were classified as having a high level of eHealth literacy. Compared to those with lower eHealth literacy, participants with higher eHealth literacy demonstrated significantly higher levels of self-care (P<.001) and self-care self-efficacy (P<.001) and lower levels of diabetes distress (P<.001). Higher eHealth literacy was also associated with greater odds of achieving ideal HbA_{1c} control (<7%) in unadjusted analyses (odds ratio 1.90, 95% CI 1.15-2.81); however, this association was not statistically significant after adjustment for sociodemographic and medical covariates (adjusted odds ratio 1.57, 95% CI 0.99-2.52; P=.07).

Conclusions: This study evaluated eHealth literacy levels among patients with DM and examined the associations between eHealth literacy and health outcomes (eg, self-care, self-care self-efficacy, diabetes distress, and HbA_{1c} control). Assessing eHealth literacy in patients with DM could be useful in identifying those who are vulnerable to poorer health outcomes. Promoting eHealth literacy among patients with DM may be important.

(J Med Internet Res 2026;28:e77856) doi:10.2196/77856

KEYWORDS

eHealth literacy; diabetes mellitus; patients; health outcomes; self-care

Introduction

Globally, there were 537 million people living with diabetes mellitus (DM) in 2021 [1]. In Hong Kong, the prevalence of DM has increased sharply by 35.1% since 1995 [2]. By 2021, an estimated 7.8% of adults in Hong Kong were affected by DM [1]. It is projected that the prevalence of DM will rise to 9.4% in 2030. DM poses a significant burden to the health care system due to its chronic nature.

There is no known cure for DM, but it is highly controllable. Self-care disease management, including having a healthy diet, exercising, medication adherence, and regular glucose

XSL•FO
RenderX

monitoring, is an essential element of diabetes care [3]. Self-care behaviors are strongly associated with positive health outcomes, including lowering blood glucose level, reducing insulin resistance, and reducing the risk of cardiovascular disease [3]; however, adherence to self-care remains suboptimal. A study in Hong Kong among older adults with DM reported that 41.4% had unsatisfactory metabolic control (ie, glycated hemoglobin [HbA_{1c}] exceeding the threshold of 7%) [4]. It has also been suggested that patients with DM require support and health information to help with disease management, but these needs have not always been met [5-7].

Online platforms are, and will continue to be, an important resource for obtaining health information, as patients with DM are increasingly playing a larger role in managing their own health [8-10]. The widespread use of digital technologies has facilitated access to information previously available primarily through health professionals [11]. Digital health resources can help individuals with chronic diseases share their experiences and gather information and may help them overcome traditional barriers to self-care [12,13]. Moreover, easy access to health resources provided by digital approaches can increase individuals' knowledge and awareness regarding medical and health issues, thus empowering them to actively participate in their own health care by making informed and well-considered decisions [14].

Health literacy is defined as "the degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions" [15]. With the advent of digital technologies, the concept of eHealth literacy (also known as digital health literacy) was introduced to extend health literacy in the digital context [16]. The importance of eHealth literacy for patients to properly use online health information and to participate in informed health care decision-making has been increasingly emphasized [17]. Due to the complexity and amount of information available online, it can be challenging for individuals with little experience in searching for online health information to locate the relevant information. In most cases, information retrieved from the internet is not properly understood by all groups of people [18]. There is also evidence that people have difficulty discerning between trustworthy and untrustworthy health information online [19,20]. Poor-quality, incorrect, and confusing health information online could also be harmful to the patients [21]. It is therefore highly important that individuals are able to effectively process and understand medical information available on the internet.

Increasing attention to eHealth literacy has been observed in recent years. For example, studies in various populations, such as older people [22,23], underserved populations [18], or patients with chronic diseases [17,24,25], have suggested that there is, in general, a low level of eHealth literacy among the study samples. Various sociodemographic factors have been shown to affect the prevalence and extent of online health information seeking, including age, gender, education, socioeconomic status, health status, and internet use [26-31]. Research has further shown that higher eHealth literacy is associated with better health outcomes across different populations, such as better subjective health status [32], health-related quality of life [33-35], lower risk of chronic disease [36], healthy behaviors [37,38], and lower psychological distress [39,40].

Several theoretical models, such as the Transactional Model of eHealth Literacy [41] and the Integrative Model of eHealth Use [42], provide frameworks for understanding how eHealth literacy may influence health outcomes. These models suggest that eHealth literacy can directly or indirectly affect health knowledge, patient–provider communication, as well as health outcomes such as proactive health behaviors and health-related quality of life [41,42]. Among patients with diabetes, higher eHealth literacy has been related to better self-care and other health outcomes through multiple psychosocial and behavioral pathways. For example, a study of 453 adults with diabetes in South Korea found that eHealth literacy was positively linked to self-management via self-efficacy and social support [43]. In a Chinese study, eHealth literacy influenced self-care behaviors indirectly through diabetes knowledge and self-efficacy [44]. Additional studies in South Korea and Taiwan also reported that eHealth literacy was associated with health-promoting behaviors and directly influenced self-care behaviors [45,46]. While previous studies have shown that eHealth literacy is associated with health-related behaviors in diabetes, evidence from patients in Hong Kong remains limited.

Given the substantial health needs of individuals with DM, examining their ability to effectively use high-quality digital health resources and make informed decisions to improve their health would be particularly useful. The purposes of this study were to assess the level of eHealth literacy among patients with DM and to examine the association between eHealth literacy and behavioral (self-care), psychosocial (self-care self-efficacy and diabetes distress), and clinical (HbA_{1c} control) outcomes among patients with type 2 DM in Hong Kong.

Methods

Study Design and Participants

A cross-sectional study was conducted among patients with DM. Target participants were patients with type 2 DM in Hong Kong who fulfilled the following criteria: (1) diagnosed with type 2 DM; (2) age between 35 and 65 years; (3) ability to speak and understand Cantonese (the primary language used in Hong Kong). Participants who were receiving psychiatric care at the time of the study were excluded.

Sample Size Calculation

Based on a previous study by Kim et al [45] conducted among patients with type 2 diabetes, which reported a correlation of r=0.15 between eHealth literacy and health-promoting behaviors. The sample size was estimated using G*Power version 3.1, assuming a correlation between eHealth literacy and self-care (r=0.15), with a 2-sided significance level of α=.05, and 80% statistical power, the minimum required sample size was 346 participants.

Data Collection

Eligible patients with type 2 DM attending a diabetes management clinic at a public hospital in Hong Kong were

XSL•FO
RenderX

invited to participate in this cross-sectional survey using a clinic-based convenience sampling approach. For recruitment, informational materials (eg, posters and leaflets) were distributed to patients visiting the clinic. Trained research assistants approached eligible individuals who were waiting to receive routine follow-up consultations. Prospective participants received a detailed explanation of the study objectives and procedures. Those who expressed interest in joining the study received an information sheet, which contained detailed information about the purposes and logistics of the study. Written informed consent and contact information were obtained from participants who agreed to participate, and interview appointments were arranged on site. Telephone surveys were administered by trained research assistants within 1 week of the participants' clinic visit. Participants received a gift coupon valued at HKD 50 (~US $6.5) as a token of appreciation for the time they spent on the study.

Ethical Considerations

Ethical approval was obtained from the Joint Chinese University of Hong Kong–New Territories East Cluster Clinical Research Ethics Committee (reference number T6080016). Written informed consent was obtained from all participants prior to data collection. Participant information was deidentified to ensure privacy and confidentiality, and all data were handled in accordance with institutional ethical guidelines. No compensation was provided to participants.

Measures

eHealth Literacy

eHealth literacy was measured using the 8-item eHealth Literacy Scale (eHEALS), which was developed to measure individuals' perceived skills of internet use for health information [47]. Items are rated on a 5-point Likert scale ranging from "1=Strongly disagree" to "5=Strongly agree." Total scores of the eHEALS range from 8 to 40, with higher scores representing higher levels of eHealth literacy. The scale has been formally translated and validated in Chinese populations [48] and in the Hong Kong context [49-51]. Using the cutoff proposed by Milne et al [52], participants who scored 4 or above on at least 5 out of the 8 eHEALS items were classified as having high level of eHealth literacy. Internal consistency of the scale was satisfactory in the present study (Cronbach α=0.98), which is comparable to that reported in previous studies conducted among general adults (Cronbach α=0.95) [50] and older adults (Cronbach α=0.939) in Hong Kong [51].

Level of Self-Care

Self-care was measured by the 15-item Summary of Diabetes Self-Care Activity scale [53], a self-report questionnaire assessing diabetes self-management behaviors across domains including general diet, specific diet, foot care, self-monitoring of blood glucose, smoking, and exercise. Participants were asked to indicate the frequency (days per week) they engaged in each of these activities, with higher scores indicating higher levels of self-care. A validated Chinese version was used [54]. Internal consistency of the scale was satisfactory in the present study (Cronbach α=0.70).

Self-Care Self-Efficacy

Self-care self-efficacy was measured using the 20-item Chinese Diabetes Empowerment Scale [55]. Items were adapted and translated from the original 37-item Diabetes Empowerment Scale [56]. The Chinese Diabetes Empowerment Scale measures the degree of psychosocial self-efficacy in carrying out diabetes care in 5 domains: achieving goals, overcoming barriers, coping, suitable methods, and obtaining support. Participants were asked to rate their level of agreement on their perceived ability to handle diabetes-related psychosocial problems on a 5-point Likert scale, from "1=Strongly disagree" to "5=Strongly agree," with higher scores indicating a greater level of self-care self-efficacy. The internal consistency of the scale was satisfactory in the present study (Cronbach α=0.94).

Diabetes Distress

Diabetes-related distress was assessed using the 15-item Chinese Diabetes Distress Scale [57], which measures emotional burden distress, regimen- and social support-related distress, and physician-related distress. Participants were asked to rate the degree to which each item may have distressed or bothered them in the past month on a 6-point Likert Scale from "1=Not a problem" to "6=A very serious problem," with higher scores indicating a higher level of diabetes distress. The internal consistency of the scale was satisfactory in the present study (Cronbach α=0.93).

Sociodemographic Characteristics

Sociodemographic characteristics, including gender, age, education level, relationship status, employment status, monthly income, were obtained. Participants were also asked to report the average time spent on the internet per week. Medical information, such as duration of diagnosis, number of complications, types of treatment, and HbA_{1c} control, was retrieved from participants' medical records.

Data Analysis

Descriptive statistics were used to summarize participants' sociodemographic and medical characteristics and levels of eHealth literacy. Correlations between participants' sociodemographic and background characteristics, eHealth literacy, and outcome variables were also examined. To evaluate the association between eHealth literacy and outcome variables, participants were classified into high and low eHealth literacy groups based on the cutoff proposed by Milne et al [52]. Independent-sample *t* tests were used to examine the differences in self-care, self-care self-efficacy, and diabetes distress between groups. One-way analysis of covariance (ANCOVA) was also conducted with covariates being adjusted for in each analysis. Sociodemographic and medical variables that were significantly correlated with the respective outcome variables would be included as covariates. To evaluate the association between eHealth literacy and HbA_{1c} control, participants with HbA_{1c} less than 7% were first defined as having ideal HbA_{1c} control. Univariate logistic regression was conducted for obtaining odds ratio (OR) of eHealth literacy on health-related outcomes. Adjusted logistic regression, adjusting for significant sociodemographic and medical variables, was also conducted; adjusted OR for eHealth literacy was then derived. All analyses

XSL•FO
RenderX

were performed using SPSS version 21 (IBM Corp). Statistical significance was defined as $P<.05$.

Results

Participant Characteristics

A total of 427 patients with type 2 DM completed the survey. The achieved sample size was deemed sufficient for the planned analyses. Slightly more than half of the participants (223/427, 52.1%) were male, and their mean age was 56.53 (SD 7.05) years. About half of them (218/427, 51%) had received a senior secondary level of education or above. The majority of participants (364/427, 85.2%) were married, and more than half of them (240/427, 56.2%) were unemployed. The mean duration of DM diagnosis was 10.69 (SD 8.14) years. Slightly less than two-thirds (261/427, 61.2%) of them had at least one DM-related complication. One hundred ninety out of 427 (44.5%) and 185 out of 427 (43.3%) participants were on oral treatment and on both oral treatment and injections, respectively. Around two-thirds (278/427, 65.1%) of the participants were classified as having high levels of eHealth literacy (Table 1).

XSL•FO
RenderX

Table . Sociodemographic and medical characteristics of participants (N=427).

Characteristic	Sample
Sociodemographic characteristics	
Gender, n (%)	
Male	221 (52.1)
Female	203 (47.9)
Age (y), mean (SD)	56.53 (7.05)
Education level, n (%)	
Primary or below	113 (26.5)
Junior secondary (Year 7 to 9)	96 (22.5)
Senior secondary (Year 10 to 13)	166 (39)
College or university	51 (12)
Relationship status, n (%)	
Single	32 (7.5)
Married	364 (85.2)
Separated/divorced/widowed	29 (6.8)
Others	2 (0.5)
Employment, n (%)	
Employed	240 (56.2)
Housewife	66 (15.5)
Unemployed	15 (3.5)
Retired	102 (23.9)
Others	4 (0.9)
Monthly income (HKD)[a], n (%)	
≤9999	68 (16.5)
10,000-19,999	159 (38.7)
20,000-29,999	117 (28.5)
≥30,000	67 (16.3)
Medical characteristics	
Duration of diagnosis (y), mean (SD)	10.69 (8.14)
Number of complications, n (%)	
None	165 (38.8)
1	162 (38.1)
2	78 (18.4)
≥3	20 (4.7)
Types of treatment, n (%)	
No prescribed medication	21 (4.9)
Oral only	190 (44.5)
Injection only	27 (6.3)
Both oral and injection	185 (43.3)
Others	4 (0.9)
Internet-related variables	
Time spent on the internet per week, n (%)	
0 h	64 (15)

XSL•FO
RenderX

Characteristic	Sample
<1 h	178 (41.6)
Between 1 and 2 h	121 (28.3)
>2 h	64 (14.9)
eHEALS[b] score, mean (SD)	23.68 (11.61)
Level of eHealth literacy[c], n (%)	
Low	149 (34.9)
High	278 (65.1)

[a]1 HKD=0.13 USD.

[b]eHEALS: the 8-item eHealth literacy scale.

[c]A high level of eHealth literacy was defined as those who scored 4 or above with at least 5 eHEALS items.

Correlation Between Variables

The correlation between variables is presented in Table 2. Among sociodemographic and medical variables, male gender (r=−0.26, P<.01), younger age (r=−0.43, P<.01), higher educational level (r=0.50, P<.01), being married (r=−0.11, P<.01), being employed (r=−0.24, P<.01), higher monthly income (r=0.30, P<.01), and shorter duration of diagnosis (r=−0.14, P<.01) had significant correlation with higher level of eHealth literacy. Higher level of eHealth literacy was also significantly correlated with higher level of self-care (r=0.15, P<.01) and higher level of self-care self-efficacy (r=.32, P<.01), as well as lower level of diabetes distress (r=−0.11, P<.05) and lower level of HbA_{1c} (r=−0.11, P<.05).

XSL•FO
RenderX

Table . Correlation between variables.

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Gender[a]	—													
Age	0.12[b]	—												
Education level	−0.30[c]	−0.43[c]	—											
Relationship status[d]	0.04	0.18[c]	−0.13[c]	—										
Employment status[e]	0.17[c]	0.39[c]	−0.20[c]	0.06	—									
Monthly income	−0.10[b]	−0.29[c]	0.40[c]	−0.09	−0.27[c]	—								
Duration of diagnosis	0.13[b]	0.24[c]	−0.21[c]	0.02	0.16[c]	−0.06	—							
Number of complications	−0.10[b]	0.19[c]	−0.13[c]	0.07	0.05	−0.10[b]	0.24[c]	—						
Type of treatment	0.04	0.06	−0.06	−0.01	0.10[b]	−0.04	0.34[c]	0.12[b]	—					
eHealth literacy	−0.26[c]	−0.43[c]	0.50[c]	−0.11[b]	−0.24[c]	0.30[c]	−0.14[c]	−0.09	−0.06	—				
Self-care	0.08	0.01	0.05	−0.10[b]	0.13[c]	0.04	0.10[b]	−0.05	0.09	0.15[c]	—			
Self-care self-efficacy	−0.17[c]	−0.16[c]	0.25[c]	−0.04	−0.05	0.21[c]	−0.05	−0.05	−0.05	0.32[c]	0.30[c]	—		
Diabetes distress	0.07	−0.03	−0.05	0.01	−0.02	-0.20[c]	0.10[b]	0.03	0.02	−0.11[b]	−0.23[c]	−0.46[c]	—	
HbA_{1c}[f]	0.08	0.01	−0.11[b]	0.06	0.06	−0.07	0.34[c]	0.18[c]	0.21[c]	−0.11[b]	0.09	−0.10[b]	0.08	—

[a]Dummy variable was created as 1: male; 2: female.

[b]$P<.05$.

[c]$P<.01$.

[d]Dummy variable was created as 1: married; 2: single/separate/divorced/widowed.

[e]Dummy variable was created as 1: employed; 2: housewife/unemployed/retired.

[f]HbA_{1c}: glycated hemoglobin.

Group Difference in Self-Care, Self-Care Self-Efficacy, and Diabetes Distress

Results of independent-samples *t* tests showed that participants classified as having a high level of eHealth literacy had significantly higher levels of self-care ($P<.001$) and self-care self-efficacy ($P<.001$), as well as significantly lower levels of diabetes distress ($P<.001$).

Results of the ANCOVA showed that such differences remained statistically significant after controlling for significant sociodemographic and medical variables. The adjusted mean differences between the high and low eHealth literacy groups

XSL•FO
RenderX

were 0.442 for self-care, 3.875 for self-care self-efficacy, and −0.237 for diabetes distress (all $P<.001$) (Table 3).

Table . Independent sample *t* test and analysis of covariance analyses of group difference in self-care, self-care self-efficacy, and diabetes distress.

Variable	Low eHealth literacy (n=149), mean (SD)	High eHealth literacy (n=278), mean (SD)	Low eHealth literacy (n=149), adjusted mean (95% CI)	High eHealth literacy (n=278), adjusted mean (95% CI)	Mean difference
Self-care	1.78 (1.03)	2.07 (1.16)[a]	1.71 (1.56–1.87)[b]	2.16 (2.02–2.23)[b]	0.442[c]
Self-care self-efficacy	72.46 (7.87)	77.08 (4.51)[c]	73.37 (72.37 - 74.37)[d]	77.25 (76.45 - 78.05)[d]	3.875[c]
Diabetes distress	2.74 (0.56)	2.44 (0.47)[c]	2.68 (2.60 - 2.76)[e]	2.44 (2.38 - 2.50)[e]	−0.237[c]

[a]$P<.01$.

[b]Adjusted for relationship status, employment, duration of diagnosis, and types of treatment.

[c]$P<.001$.

[d]Adjusted for sex, age, education level, and monthly family income.

[e]Adjusted for monthly family income and duration of diagnosis.

Logistic Regression on Ideal HbA_{1c} Control

There were 38.7% (96/248) and 26.0% (38/146) ideal HbA_{1c} in participants with low level and high level of eHealth literacy, respectively. Results from the univariate logistic regression analysis showed that participants with a high level of eHealth literacy were more likely to report ideal HbA_{1c} control (unadjusted OR 1.90, 95% CI 1.15-2.81; $P<.01$). After adjusting for significant sociodemographic and medical information, the association was not statistically significant (adjusted OR 1.57, 95% CI 0.99-2.52, $P=0.070$).

Discussion

Main Findings and Implications

DM is a chronic disease that requires good knowledge and self-management skills to optimize glycemic control and health outcomes. As digital health technologies are now widely used to support patients, eHealth literacy is of great importance because seeking online health information reflects a complex issue involving information awareness, consciousness, reasoning, and knowledge. It is important to assess patients with DM eHealth literacy, as health care professionals should have a good understanding of the patients' ability to use electronic resources before recommending the use of online health information [47].

Few studies in Hong Kong have reported the eHealth literacy levels of the population, especially in patients. Several local studies reported mean eHEALS scores of 26.10 (SD 7.70) in 1501 general adults [50] and 27.34 (SD 6.44) among 306 adults aged 60 years or older [51]. This study expands the evidence base among patients in Hong Kong and shows that the mean eHEALS score among 427 patients with type 2 DM was 23.68 (SD 11.61). In addition, using the cutoff proposed by Milne et al [52], approximately two-thirds (278/427, 65.1%) of the patients with DM were considered as having a high level of eHealth literacy. Compared with other studies using the same cutoff, the proportion of patients having a high level of eHealth literacy in the present study was significantly higher than that reported among lung cancer survivors (33.7%) in Canada [52]. This difference may be partly attributable to age, as participants in the present study were younger (mean age 56.5, SD 7.05 y) than those in Milne et al's study (median age 71 y). This interpretation is further supported by the negative association between age and eHealth literacy observed in the present study and reported consistently in the literature [11,31]. The patients with DM in the present study were likely confident in using the internet for health information and advice; they also demonstrated reasonable skills and self-efficacy required to achieve maximum benefit from the use.

In addition to age, the present study also revealed that among the sociodemographic and medical variables, male gender, higher level of education, currently married, employed, higher monthly income, and shorter duration of diagnosis were associated with higher eHealth literacy. These findings are consistent with the literature showing that eHealth literacy is generally lower among individuals with lower socioeconomic status [14,58-60]. Because the eHEALS was designed to measure an individual's confidence in their ability to locate and access online health information, it is conceivable that individuals with better socioeconomic status would possess higher ability in understanding and discerning the quality of the health information available on the internet. Individuals who were married and employed may also be more likely to receive a higher level of social support in searching for health information, including those from the online sources.

Understanding the association between eHealth literacy and health-related outcomes is essential for developing effective health communication strategies in the future. It has been found that the use of online health information and eHealth literacy might affect health behaviors [37,45,61,62]. In the present study, higher eHealth literacy was significantly associated with higher levels of self-care and self-care self-efficacy and with lower levels of diabetes distress after controlling for significant sociodemographic and medical variables. These findings are consistent with recent evidence showing that eHealth literacy significantly influences diabetes self-care behaviors among patients with type 2 DM in Saudi Arabia [63]. Although this study also revealed eHealth literacy was significantly associated with better glycemic outcomes in their population [63], the present study did not find a statistically significant association between eHealth literacy and HbA_{1c} among patients with

XSL•FO
RenderX

diabetes in Hong Kong after adjustment for covariates. Patients with higher eHealth literacy are not only more inclined to use the internet for health-related information but also are able to understand, evaluate the information obtained, and use quality information to make informed decisions about health. They may therefore be more motivated to adopt healthier lifestyles based on the information that they obtain online [38], more directly leading to better behavioral and psychosocial outcomes. The integrated Model of eHealth Use model also supports the motivational pathways through which eHealth literacy may influence health behavior [42]. Individuals with a higher ability to understand online health information may also be more likely to access health care and interact with health care professionals, therefore more likely to benefit from their health care utilization. In the conceptual model proposed by Paasche-Orlov and Wolf [64], self-efficacy is considered an important mediator linking health literacy and health outcomes. In digital health contexts, self-efficacy has also been found to be a significant mediator in the relationship between eHealth literacy and health outcomes in patients with DM [43,44]. Individuals with higher eHealth literacy are likely to have greater confidence in their ability to effectively utilize online health resources and engage in self-care behaviors, which in turn, is associated with overall behavioral and mental health in patients with DM. However, HbA_{1c} is a more distal biomedical indicator and reflects average glycemic control over the preceding two to three months, which may be less directly associated with eHealth literacy. Future studies are warranted to further elucidate the pathways linking eHealth literacy to glycemic outcomes in this population. Findings of this study suggest that eHealth literacy is an important concept associated with the health of patients with DM. Improving eHealth literacy of patients with DM would potentially be an effective means in improving the health status of the population.

Implications for Practice

The eHEALS is currently one of the most common tools designed to measure an individual's confidence in their ability to locate and access online health information [65,66]. It demonstrated high internal consistency and predictive validity in Chinese chronic patients [67], and is associated with a range of health-related outcomes among various populations [22,30,67,68]. Findings of the present study further confirmed its utility in Chinese patients with DM and suggested that it could be an effective screening tool for screening patients' level of eHealth literacy in clinical settings.

The higher levels of self-care and self-care self-efficacy, and lower levels of diabetes distress documented among participants with higher eHealth literacy also provided important insights that promoting eHealth literacy can have the potential to support patients with DM with self-management behaviors and may result in better DM outcomes. Recent studies have shown that patients with DM who used DM apps had higher eHealth literacy than nonusers and perceived the app to have a greater effectiveness on their health behaviors [68-70]. Health care providers should therefore empower patients with DM with the skills in utilizing online resources. Improving eHealth literacy in patients with type 2 diabetes has important practical implications for self-care and health outcomes. Evidence suggests that improving diabetes knowledge and attitudes toward online health information are likely to be effective strategies for enhancing eHealth literacy in this population [71]. However, it is also noted that participant engagement in eHealth technologies might wane over time. Given the chronic nature of DM, interventions to increase eHealth literacy should make use of content and features that can be personalized and adapted to the individual over time to achieve a sustainable positive effect [72]. Continuous support, training, and monitoring by health care providers are essential to help patients acquire and apply accurate health information effectively, ultimately improving self-care and health outcomes.

Limitations

There are several limitations that should be noted. First, the present study was cross-sectional in nature; therefore, no causality between the variables could be assumed. Second, self-perceived eHealth literacy was measured in the present study as there was a lack of objective, validated measure in eHealth literacy in the current literature. The current literature seems to have revealed mixed findings on the association between subjective and objective measures of eHealth literacy on health-related internet use [25,36]. However, it is important to note that the eHEALS scale is one of the most used validated scales to measure eHealth literacy. This study was also limited by self-selection bias, that those who agreed to take part in the study might have more favorable views towards internet use and thus a higher level of eHealth literacy. Finally, as the present sample of patients with DM was relatively young and recruited from only one public DM clinic, cautions should be taken when generalizing the results to the general DM population in Hong Kong.

Conclusion

The internet plays an increasingly important part in disease management among patients with DM. Findings of the present study showed that higher level of eHealth literacy was associated with their better self-care, self-care self-efficacy, lower level of diabetes distress, and having ideal HbA_{1c} control. Assessing the level of eHealth literacy of patients with DM could be useful for identifying patients who are more vulnerable to poorer health. Promoting the eHealth literacy of patients with DM could potentially be associated with better health outcomes.

Funding

This work was supported by Health and Medical Research Fund of Hong Kong Food and Health Bureau (12130751).

Data Availability

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to their containing information that could compromise the privacy of research participants.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: PKHM, APSK, VWYC, JTFL
Data curation: PKHM, APSK, VWYC
Formal analysis: PKHM
Methodology: PKHM, APSK, VWYC, JTFL
Project administration: PKHM
Supervision: PKHM
Writing – original draft: PKHM, LX

Conflicts of Interest

None declared.

References

1. IDF Diabetes Atlas, 10th edition: International Diabetes Federation; 2021. URL: https://diabetes.org.br/wp-content/uploads/2022/02/IDF_Atlas_10th_Edition_2021-.pdf [accessed 2026-02-03]
2. Cheung B, Man Y, Lam K, Wat N, Lo J, Chau D, et al. Prevalence of diabetes mellitus (DM) in the Hong Kong Cardiovascular Risk Factor Prevalence Study cohort. Hong Kong Med J 2003;9(Supp 1) [FREE Full text]
3. American Diabetes Association. Standards of medical care in diabetes--2011. Diabetes Care 2011 Jan;34 Suppl 1(Suppl 1):S11-S61. [doi: 10.2337/dc11-S011] [Medline: 21193625]
4. Cheng LY, Leung SY, Leung MKW. The association of glycemic control and fall risk in diabetic elderly: a cross-sectional study in Hong Kong. BMC Prim Care 2022 Aug 1;23(1):192. [doi: 10.1186/s12875-022-01807-7] [Medline: 35915395]
5. Carolan M, Holman J, Ferrari M. Experiences of diabetes self-management: a focus group study among Australians with type 2 diabetes. J Clin Nurs 2015 Apr;24(7-8):1011-1023. [doi: 10.1111/jocn.12724] [Medline: 25363710]
6. Song Y, Song HJ, Han HR, Park SY, Nam S, Kim MT. Unmet needs for social support and effects on diabetes self-care activities in Korean Americans with type 2 diabetes. Diabetes Educ 2012;38(1):77-85. [doi: 10.1177/0145721711432456] [Medline: 22222514]
7. Timmers T, Janssen L, Kool RB, Kremer JA. Educating patients by providing timely information using smartphone and tablet apps: systematic review. J Med Internet Res 2020 Apr 13;22(4):e17342. [doi: 10.2196/17342] [Medline: 32281936]
8. Desveaux L, Shaw J, Saragosa M, et al. A mobile app to improve self-management of individuals with type 2 diabetes: qualitative realist evaluation. J Med Internet Res 2018 Mar 16;20(3):e81. [doi: 10.2196/jmir.8712] [Medline: 29549070]
9. Kruse CS, Beane A. Health information technology continues to show positive effect on medical outcomes: systematic review. J Med Internet Res 2018 Feb 5;20(2):e41. [doi: 10.2196/jmir.8793] [Medline: 29402759]
10. Shahbazi M, Bagherian H, Sattari M, Saghaeiannejad-Isfahani S. The opportunities and challenges of using mobile health in elderly self-care. J Educ Health Promot 2021;10:80. [doi: 10.4103/jehp.jehp_871_20] [Medline: 34084827]
11. Tennant B, Stellefson M, Dodd V, et al. eHealth literacy and Web 2.0 health information seeking behaviors among baby boomers and older adults. J Med Internet Res 2015 Mar 17;17(3):e70. [doi: 10.2196/jmir.3992] [Medline: 25783036]
12. Perakslis E, Ginsburg GS. Digital health—the need to assess benefits, risks, and value. JAMA 2021 Jan 12;325(2):127-128. [doi: 10.1001/jama.2020.22919] [Medline: 33369627]
13. McBride D. Web forums allow patients with colorectal cancer to share their real-world experiences. ONS Connect 2013 Dec;28(4):46. [Medline: 24400584]
14. Yoon H, Jang Y, Vaughan PW, Garcia M. Older adults' internet use for health information: digital divide by race/ethnicity and socioeconomic status. J Appl Gerontol 2020 Jan;39(1):105-110. [doi: 10.1177/0733464818770772] [Medline: 29661052]
15. Parker RM, Ratzan SC, Lurie N. Health literacy: a policy challenge for advancing high-quality health care. Health Aff (Millwood) 2003;22(4):147-153. [doi: 10.1377/hlthaff.22.4.147] [Medline: 12889762]
16. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
17. Wu Y, Wen J, Wang X, et al. Associations between e-health literacy and chronic disease self-management in older Chinese patients with chronic non-communicable diseases: a mediation analysis. BMC Public Health 2022 Nov 29;22(1):2226. [doi: 10.1186/s12889-022-14695-4] [Medline: 36447176]
18. Chesser A, Burke A, Reyes J, Rohrberg T. Navigating the digital divide: a systematic review of eHealth literacy in underserved populations in the United States. Inform Health Soc Care 2016;41(1):1-19. [doi: 10.3109/17538157.2014.948171] [Medline: 25710808]
19. Nguyen A, Mosadeghi S, Almario CV. Persistent digital divide in access to and use of the Internet as a resource for health information: results from a California population-based study. Int J Med Inform 2017 Jul;103:49-54. [doi: 10.1016/j.ijmedinf.2017.04.008] [Medline: 28551001]
20. Tonsaker T, Bartlett G, Trpkov C. Health information on the Internet: gold mine or minefield? Can Fam Physician 2014 May;60(5):407-408. [Medline: 24828994]

XSL•FO
RenderX

21. Valizadeh-Haghi S, Rahmatizadeh S. Evaluation of the quality and accessibility of available websites on kidney transplantation. Urol J 2018 Sep 26;15(5):261-265. [doi: 10.22037/uj.v0i0.4252] [Medline: 30058062]
22. Xie L, Zhang S, Xin M, Zhu M, Lu W, Mo PKH. Electronic health literacy and health-related outcomes among older adults: a systematic review. Prev Med 2022 Apr;157:106997. [doi: 10.1016/j.ypmed.2022.106997] [Medline: 35189203]
23. Wang Y, Song Y, Zhu Y, Ji H, Wang A. Association of eHealth literacy with health promotion behaviors of community-dwelling older people: the chain mediating role of self-efficacy and self-care ability. Int J Environ Res Public Health 2022 May 17;19(10):6092. [doi: 10.3390/ijerph19106092] [Medline: 35627627]
24. Robinson C, Graham J. Perceived Internet health literacy of HIV-positive people through the provision of a computer and Internet health education intervention. Health Info Libr J 2010 Dec;27(4):295-303. [doi: 10.1111/j.1471-1842.2010.00898.x] [Medline: 21050372]
25. van der Vaart R, Drossaert CHC, de Heus M, Taal E, van de Laar MAFJ. Measuring actual eHealth literacy among patients with rheumatic diseases: a qualitative analysis of problems encountered using Health 1.0 and Health 2.0 applications. J Med Internet Res 2013 Feb 11;15(2):e27. [doi: 10.2196/jmir.2428] [Medline: 23399720]
26. Oh YS, Cho Y. Examining the relationships between resources and online health information seeking among patients with chronic diseases and healthy people. Soc Work Health Care 2015;54(2):83-100. [doi: 10.1080/00981389.2014.987940] [Medline: 25674723]
27. Li J, Theng YL, Foo S. Predictors of online health information seeking behavior: changes between 2002 and 2012. Health Informatics J 2016 Dec;22(4):804-814. [doi: 10.1177/1460458215595851] [Medline: 26261218]
28. Moreland J, French TL, Cumming GP. The prevalence of online health information seeking among patients in Scotland: a cross-sectional exploratory study. JMIR Res Protoc 2015 Jul 15;4(3):e85. [doi: 10.2196/resprot.4010] [Medline: 26177562]
29. Li N, Orrange S, Kravitz RL, Bell RA. Reasons for and predictors of patients' online health information seeking following a medical appointment. Fam Pract 2014 Oct;31(5):550-556. [doi: 10.1093/fampra/cmu034] [Medline: 24963151]
30. Bailey CE, Kohler WJ, Makary C, Davis K, Sweet N, Carr M. eHealth literacy in otolaryngology patients. Ann Otol Rhinol Laryngol 2019 Nov;128(11):1013-1018. [doi: 10.1177/0003489419856377] [Medline: 31195809]
31. Halwas N, Griebel L, Huebner J. eHealth literacy, Internet and eHealth service usage: a survey among cancer patients and their relatives. J Cancer Res Clin Oncol 2017 Nov;143(11):2291-2299. [doi: 10.1007/s00432-017-2475-6] [Medline: 28699035]
32. Kim H, Yang E, Ryu H, Kim HJ, Jang SJ, Chang SJ. Psychometric comparisons of measures of eHealth literacy using a sample of Korean older adults. Int J Older People Nurs 2021 May;16(3):e12369. [doi: 10.1111/opn.12369] [Medline: 33527701]
33. Li S, Cui G, Yin Y, Wang S, Liu X, Chen L. Health-promoting behaviors mediate the relationship between eHealth literacy and health-related quality of life among Chinese older adults: a cross-sectional study. Qual Life Res 2021 Aug;30(8):2235-2243. [doi: 10.1007/s11136-021-02797-2] [Medline: 33661455]
34. Brørs G, Wentzel-Larsen T, Dalen H, et al. Psychometric properties of the Norwegian version of the electronic Health Literacy Scale (eHEALS) among patients after percutaneous coronary intervention: cross-sectional validation study. J Med Internet Res 2020 Jul 28;22(7):e17312. [doi: 10.2196/17312] [Medline: 32720900]
35. Xie L, Mo PKH. A 3-wave longitudinal study of eHealth literacy and older people's health-related quality of life in China: the mediating role of general self-efficacy. J Am Med Dir Assoc 2024 Dec;25(12):105310. [doi: 10.1016/j.jamda.2024.105310] [Medline: 39426796]
36. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012 Jan 27;14(1):e19. [doi: 10.2196/jmir.1619] [Medline: 22357448]
37. Kim K, Shin S, Kim S, Lee E. The relation between eHealth literacy and health-related behaviors: systematic review and meta-analysis. J Med Internet Res 2023 Jan 30;25:e40778. [doi: 10.2196/40778] [Medline: 36716080]
38. Xie L, Mo PKH. Prospective relationship between electronic health literacy and health-promoting lifestyle among Chinese older adults: a three-wave longitudinal study. Soc Sci Med 2025 Jan;364:117166. [doi: 10.1016/j.socscimed.2024.117166] [Medline: 39586136]
39. Lin CY, Broström A, Griffiths MD, Pakpour AH. Psychometric evaluation of the Persian eHealth Literacy Scale (eHEALS) among elder Iranians with heart failure. Eval Health Prof 2020 Dec;43(4):222-229. [doi: 10.1177/0163278719827997] [Medline: 30744419]
40. Leung AYM, Parial LL, Tolabing MC, et al. Sense of coherence mediates the relationship between digital health literacy and anxiety about the future in aging population during the COVID-19 pandemic: a path analysis. Aging Ment Health 2022 Mar;26(3):544-553. [doi: 10.1080/13607863.2020.1870206] [Medline: 33438448]
41. Paige SR, Stellefson M, Krieger JL, Anderson-Lewis C, Cheong J, Stopka C. Proposing a transactional model of eHealth literacy: concept analysis. J Med Internet Res 2018 Oct 2;20(10):e10175. [doi: 10.2196/10175] [Medline: 30279155]
42. Bodie GD, Dutta MJ. Understanding health literacy for strategic health marketing: eHealth literacy, health disparities, and the digital divide. Health Mark Q 2008;25(1-2):175-203. [doi: 10.1080/07359680802126301] [Medline: 18935884]
43. Lee EH, Lee YW, Kang EH, Kang HJ. Relationship between electronic health literacy and self-management in people with type 2 diabetes using a structural equation modeling approach. J Nurs Res 2024 Jan 1;32(1):e315. [doi: 10.1097/jnr.0000000000000588] [Medline: 38128065]

XSL•FO
RenderX

44. Ji X, Chi J. Exploring the relationship between eHealth literacy and diabetes knowledge, self-efficacy, and self-care behaviors in Chinese diabetic patients: a cross-sectional study. J Nurs Res 2024 Dec 1;32(6):e359. [doi: 10.1097/jnr.0000000000000642] [Medline: 39593226]
45. Kim KA, Kim YJ, Choi M. Association of electronic health literacy with health-promoting behaviors in patients with type 2 diabetes: a cross-sectional study. Comput Inform Nurs 2018 Sep;36(9):438-447. [doi: 10.1097/CIN.0000000000000438] [Medline: 29742548]
46. Guo SHM, Hsing HC, Lin JL, Lee CC. Relationships between mobile eHealth literacy, diabetes self-care, and glycemic outcomes in Taiwanese patients with type 2 diabetes: cross-sectional study. JMIR mHealth uHealth 2021 Feb 5;9(2):e18404. [doi: 10.2196/18404] [Medline: 33544088]
47. Norman CD, Skinner HA. eHEALS: the eHealth literacy scale. J Med Internet Res 2006 Nov 14;8(4):e27. [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
48. Xu RH, Zhou L, Lu SY, Wong EL, Chang J, Wang D. Psychometric validation and cultural adaptation of the simplified Chinese eHealth literacy scale: cross-sectional study. J Med Internet Res 2020 Dec 7;22(12):e18613. [doi: 10.2196/18613] [Medline: 33284123]
49. Wong DKK, Cheung MK. Online health information seeking and eHealth literacy among patients attending a primary care clinic in Hong Kong: a cross-sectional survey. J Med Internet Res 2019 Mar 27;21(3):e10831. [doi: 10.2196/10831] [Medline: 30916666]
50. Guo Z, Zhao SZ, Guo N, et al. Socioeconomic disparities in eHealth literacy and preventive behaviors during the COVID-19 pandemic in Hong Kong: cross-sectional study. J Med Internet Res 2021 Apr 14;23(4):e24577. [doi: 10.2196/24577] [Medline: 33784240]
51. Kim S, Chow BC, Park S, Liu H. The usage of digital health technology among older adults in Hong Kong and the role of technology readiness and eHealth literacy: path analysis. J Med Internet Res 2023 Apr 12;25(1):e41915. [doi: 10.2196/41915] [Medline: 37043274]
52. Milne RA, Puts MTE, Papadakos J, et al. Predictors of high eHealth literacy in primary lung cancer survivors. J Cancer Educ 2015 Dec;30(4):685-692. [doi: 10.1007/s13187-014-0744-5] [Medline: 25355524]
53. Toobert DJ, Hampson SE, Glasgow RE. The summary of diabetes self-care activities measure: results from 7 studies and a revised scale. Diabetes Care 2000 Jul;23(7):943-950. [doi: 10.2337/diacare.23.7.943] [Medline: 10895844]
54. Tang YH, Pang SMC, Chan MF, Yeung GSP, Yeung VTF. Health literacy, complication awareness, and diabetic control in patients with type 2 diabetes mellitus. J Adv Nurs 2008 Apr;62(1):74-83. [doi: 10.1111/j.1365-2648.2007.04526.x] [Medline: 18352966]
55. Shiu ATY, Wong RYM, Thompson DR. Development of a reliable and valid Chinese version of the diabetes empowerment scale. Diabetes Care 2003 Oct;26(10):2817-2821. [doi: 10.2337/diacare.26.10.2817] [Medline: 14514585]
56. Anderson RM, Funnell MM, Fitzgerald JT, Marrero DG. The diabetes empowerment scale: a measure of psychosocial self-efficacy. Diabetes Care 2000 Jun;23(6):739-743. [doi: 10.2337/diacare.23.6.739] [Medline: 10840988]
57. Ting RZW, Nan H, Yu MWM, et al. Diabetes-related distress and physical and psychological health in chinese type 2 diabetic patients. Diabetes Care 2011 May;34(5):1094-1096. [doi: 10.2337/dc10-1612] [Medline: 21398526]
58. Verma R, Saldanha C, Ellis U, Sattar S, Haase KR. eHealth literacy among older adults living with cancer and their caregivers: a scoping review. J Geriatr Oncol 2022 Jun;13(5):555-562. [doi: 10.1016/j.jgo.2021.11.008] [Medline: 34810146]
59. Choi NG, Dinitto DM. The digital divide among low-income homebound older adults: internet use patterns, eHealth literacy, and attitudes toward computer/internet use. J Med Internet Res 2013 May 2;15(5):e93. [doi: 10.2196/jmir.2645] [Medline: 23639979]
60. Hung LY, Lyons JG, Wu CH. Health information technology use among older adults in the United States, 2009-2018. Curr Med Res Opin 2020 May;36(5):789-797. [doi: 10.1080/03007995.2020.1734782] [Medline: 32096650]
61. Li SJ, Yin YT, Cui GH, Xu HL. The associations among health-promoting lifestyle, eHealth literacy, and cognitive health in older Chinese adults: a cross-sectional study. Int J Environ Res Public Health 2020 Mar 27;17(7):2263. [doi: 10.3390/ijerph17072263] [Medline: 32230935]
62. Kim SH, Son YJ. Relationships between eHealth literacy and health behaviors in Korean adults. Comput Inform Nurs 2017 Feb;35(2):84-90. [doi: 10.1097/CIN.0000000000000255] [Medline: 27258808]
63. AlShehri WS, Khojah YY. Impact of e-health literacy on diabetes self-care activities among people with type 2 diabetes attending primary healthcare centers in Makkah City, Saudi Arabia: a cross-sectional study, 2025. Cureus 2025 Jul;17(7):e87968. [doi: 10.7759/cureus.87968] [Medline: 40666264]
64. Paasche-Orlow MK, Wolf MS. The causal pathways linking health literacy to health outcomes. Am J Health Behav 2007;31 Suppl 1:S19-S26. [doi: 10.5555/ajhb.2007.31.supp.S19] [Medline: 17931132]
65. Griebel L, Enwald H, Gilstad H, Pohl AL, Moreland J, Sedlmayr M. eHealth literacy research-quo vadis? Inform Health Soc Care 2018 Dec;43(4):427-442. [doi: 10.1080/17538157.2017.1364247] [Medline: 29045164]
66. Lee J, Lee EH, Chae D. eHealth literacy instruments: systematic review of measurement properties. J Med Internet Res 2021 Nov 15;23(11):e30644. [doi: 10.2196/30644] [Medline: 34779781]

XSL•FO
RenderX

67. Chang A, Schulz PJ. The measurements and an elaborated understanding of Chinese eHealth literacy (C-eHEALS) in chronic patients in China. Int J Environ Res Public Health 2018 Jul 23;15(7):1553. [doi: 10.3390/ijerph15071553] [Medline: 30041420]
68. Ernsting C, Stühmann LM, Dombrowski SU, Voigt-Antons JN, Kuhlmey A, Gellert P. Associations of health app use and perceived effectiveness in people with cardiovascular diseases and diabetes: population-based survey. JMIR mHealth uHealth 2019 Mar 28;7(3):e12179. [doi: 10.2196/12179] [Medline: 30920383]
69. Rollo ME, Aguiar EJ, Williams RL, et al. eHealth technologies to support nutrition and physical activity behaviors in diabetes self-management. Diabetes Metab Syndr Obes 2016;9:381-390. [doi: 10.2147/DMSO.S95247] [Medline: 27853384]
70. Kebede MM, Pischke CR. Popular diabetes apps and the impact of diabetes app use on self-care behaviour: a survey among the digital community of persons with diabetes on social media. Front Endocrinol (Lausanne) 2019;10:135. [doi: 10.3389/fendo.2019.00135] [Medline: 30881349]
71. Lee M, Shim J. Effects of diabetes knowledge and attitudes toward internet health information on e-health literacy in middle-aged patients with diabetes. Healthcare (Basel) 2025 Feb 27;13(5):512. [doi: 10.3390/healthcare13050512] [Medline: 40077074]
72. Pereira K, Phillips B, Johnson C, Vorderstrasse A. Internet delivered diabetes self-management education: a review. Diabetes Technol Ther 2015 Jan;17(1):55-63. [doi: 10.1089/dia.2014.0155] [Medline: 25238257]

Abbreviations

ANCOVA: analysis of covariance
DM: diabetes mellitus
eHEALS: the 8-item eHealth Literacy Scale
HbA_{1c}: glycated hemoglobin
OR: odds ratio

Edited by A Schwartz, M Balcarras; submitted 02.Jun.2025; peer-reviewed by CL Teng, CW Chan, MA Bujang; revised version received 30.Dec.2025; accepted 02.Jan.2026; published 09.Mar.2026.

Please cite as:
Mo PKH, Kong APS, Xie L, Chan VWY, Lau JTF
Association of Electronic Health Literacy With Self-Care and Health Outcomes Among Patients With Type 2 Diabetes Mellitus: Cross-Sectional Study
J Med Internet Res 2026;28:e77856
URL: https://www.jmir.org/2026/1/e77856
doi:10.2196/77856

© Phoenix Kit-Han Mo, Alice PS Kong, Luyao Xie, Virginia WY Chan, Joseph TF Lau. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Relationship Between Electronic Health Literacy and Health-Related Quality of Life Among Chinese Older Adults: Cross-Sectional Study

Yongqiang Wang[1*], MSc; Baozhen Dai[1,2*], PhD; Jiazhen Yao[1], MSc

[1]Health Policy and Healthcare Security Research Center, Southeast University, No. 87 Dingjiaqiao, Nanjing, Jiangsu, China

[2]Ageing-Responsive Civilization Think Tank, Nanjing, Jiangsu, China

[*]these authors contributed equally

Corresponding Author:
Baozhen Dai, PhD
Health Policy and Healthcare Security Research Center, Southeast University, No. 87 Dingjiaqiao, Nanjing, Jiangsu, China

Abstract

Background: The rapid digitalization of health care has reshaped access to medical services. However, older adults often remain disadvantaged due to the digital divide. Electronic health literacy (EHL) is increasingly recognized as a determinant of health-related quality of life (HRQoL); however, its mechanisms and subgroup differences in China remain underexplored.

Objective: This study aimed to examine the association between EHL and multidimensional HRQoL among Chinese older adults, with a focus on the mediating roles of attitudes toward own aging (ATOA) and self-efficacy (SE), and heterogeneity by age, residence, and lifestyle.

Methods: A cross-sectional survey (July-November 2024) included 8364 adults aged ≥55 years from 4 provinces using stratified multistage sampling. HRQoL was measured by physical health (PH), mental health (MH), and life satisfaction (LS). EHL was assessed with the eHealth Literacy Scale (eHEALS), ATOA with the Philadelphia Geriatric Center Morale Scale subscale, and SE with the General Self-Efficacy Scale. Analyses used seemingly unrelated regressions, PROCESS (Andrew F. Hayes) macro mediation with 5000 bootstraps, and subgroup regressions.

Results: EHL was positively associated with PH (β=0.273; *P*<.001), MH (β=0.190; *P*<.001), and LS (β=0.082; *P*<.001). ATOA and SE significantly mediated these associations (all *P*<.001). For PH, the association was partially mediated by ATOA and SE. The total effect was 0.273 (*P*<.001), with indirect effects accounting for 52.7% of the total. For MH, inconsistent mediation was observed. The direct effect was negative (β=–0.069; *P*<.01), but indirect effects were positive, yielding a positive total effect. For LS, the effect was fully mediated by ATOA and SE. Subgroup analyses showed stronger effects in the younger-old adults (PH: β=0.288, MH: β=0.188, LS: β=0.089; all *P*<.001), urban residents (PH: β=0.237, MH: β=0.214, LS: β=0.083; all *P*<.001), and nonexercisers (PH: β=0.322; *P*<.001).

Conclusions: EHL is strongly associated with HRQoL among Chinese older adults. Its effects on MH and LS operate primarily through psychosocial pathways, while PH benefits directly. The findings highlight EHL's compensatory role, particularly for the younger-old group, rural residents, and nonexercisers, underscoring its importance in digital inclusion and healthy aging policies.

(J Med Internet Res 2026;28:e84700) doi:10.2196/84700

KEYWORDS

electronic health literacy; health-related quality of life; older adult; attitudes toward own aging; self-efficacy

Introduction

Background

The accelerating pace of global population aging has made the pursuit of healthy aging a pressing challenge for both public health and social development [1]. Beard et al [2] argue that healthy aging is not merely about extending life expectancy but about maintaining functional ability, psychological well-being, and social participation across the entire lifespan, thereby improving the holistic welfare of older adults. Under this overall objective, health-related quality of life (HRQoL) has become a key comprehensive indicator for measuring its achievement. HRQoL refers to an individual's subjective assessment of their well-being and ability to perform physical, psychological, and social functions [3]. Considering the landscape of healthy aging, China is undergoing the world's largest and fastest demographic transition, where high prevalence of chronic diseases, functional decline, and social isolation intertwine, severely constraining the progress of healthy aging [4]. Relevant surveys also indicate that the HRQoL among China's older adults remains suboptimal [5,6]. Therefore, enhancing HRQoL is not only paramount to

XSL•FO
RenderX

achieving healthy aging but also an urgent imperative for addressing the challenges of China's aging society.

Concurrently, the digital transformation of health and medical service systems is reshaping the way care is delivered [7]. Electronic health services—such as online prescription, telemedicine, electronic health record, and health information platform—have become vital channels for older adults to obtain health resources. These services transcend traditional barriers of time and space, improving accessibility and the efficiency of health management [8]. Yet not all older adults benefit equally. Owing to the persistent digital divide, many face barriers in accessing, evaluating, and using health information, which may exacerbate health inequalities [9]. In this context, electronic health literacy (EHL) has increasingly been recognized as a critical determinant of HRQoL among older adults [10].

EHL refers to the ability to locate, understand, evaluate, and apply health information from electronic media, as well as the capacity to use acquired health information to address health issues [11]. Existing evidence has shown consistent associations between EHL and a range of health outcomes [12-14]. Older adults with higher levels of EHL are more likely to seek health information online, and in doing so, report better self-rated health, improved chronic disease management, and healthier behaviors [15]. EHL has also been associated with improved mental health (MH) and subjective well-being [14]. Nevertheless, important gaps remain, as existing research has focused on self-assessed health as a single measurement tool [16,17], failing to adopt a comprehensive perspective encompassing physical health (PH), MH, and subjective well-being. Furthermore, the mediating mechanisms underlying these associations and the heterogeneity among population subgroups have not been sufficiently explored [16,18].

In China, the "Internet Plus Healthcare" initiative has expanded digital health services nationwide [15]. However, the general level of EHL among older adults remains relatively low, with marked regional and subgroup disparities [19]. Although older adults should be the main beneficiaries of digital health innovations, they often remain a disadvantaged group in practice due to the digital divide [20]. Consequently, empirical investigation into the role of EHL in shaping HRQoL in China remains scarce, highlighting the need to investigate its mechanisms and contextual variations.

Theoretical Hypotheses

Since the World Health Organization declared 2021-2030 the "Decade of Healthy Aging," enhancing HRQoL has become the key objective for achieving healthy aging [21]. HRQoL encompasses not only the maintenance of physical functioning but also MH, life satisfaction (LS), and cognitive well-being—positive indicators that extend beyond morbidity or disease prevalence [22]. Within the digital era, the impact of internet use on the health of older adults requires positive mediation through human capital. As a critical form of human capital in the digital health field, EHL directly empowers older adults' health management practices by enhancing their capacity to discern health information and optimizing pathways to health services, thereby advancing healthy aging [23]. Similarly, EHL is recognized as both a prerequisite for inclusive health innovation and a super determinant of health equity [24]. Systematic reviews have also confirmed a significant positive correlation between EHL and older adults' health management efficacy [25]. Thus, EHL is not only an essential core health competency for older adults in the digital age but also, due to its enabling role in health management practices and its positive predictive power for health outcomes, emerges as a form of health capital closely linked to HRQoL.

The theory of successful aging emphasizes that older adults' subjective attitudes toward aging play a critical role in maintaining functional capacity [26,27]. EHL, as a resource for information empowerment and cognitive development, can effectively challenge and mitigate negative attitudes toward aging among older adults. This enables them to view aging as a stage where vitality, learning, and growth remain possible [28]. In turn, positive attitudes toward own aging (ATOA) contribute to the enhancement of older adults' HRQoL across multiple dimensions. Psychologically, positive ATOA can alleviate aging-related stress responses, strengthen psychological resilience, and reduce the risk of depression [29]. Behaviorally, older adults with positive attitudes are more likely to adopt and maintain health-promoting behaviors to proactively preserve their capabilities [30]. Physiologically, existing research has confirmed that positive self-perceptions of aging exert direct positive physiological effects on cardiovascular and immune functions by alleviating chronic stress [31]. Based on these findings, our study identifies ATOA as a mediating variable to explore the underlying mechanisms.

The conceptual model proposed by Paasche-Orlow and Wolf emphasizes that self-efficacy (SE) serves as a pivotal mediating factor linking health literacy to health outcomes [32]. SE refers to an individual's subjective judgment and intrinsic confidence in their own capabilities and personal worth within specific contexts [33]. In digital health settings, higher EHL empowers older adults to more effectively access, understand, evaluate, and apply digital health information and services. This process directly enhances their perceived capacity to address health issues. According to social cognitive theory, experiences of successfully mastering and applying relevant skills are key sources of SE. Thus, EHL strengthens older adults' confidence in managing their health by providing experiences of successfully using digital tools for disease prevention, health management, and self-care [34]. SE theory posits that an individual's belief in their ability to perform specific tasks is a core determinant of health behaviors [35,36]. Enhanced SE promotes a range of adaptive health behaviors and psychological processes. Older adults with high SE are more likely to adopt proactive health behaviors, adhere to treatment plans, better cope with disease-related stress, and maintain greater psychological resilience. These factors collectively contribute to improvements in both physical and MH, ultimately manifesting as enhanced HRQoL [37]. Thus, our study examines key mechanisms within the relationship by identifying SE as a mediating variable.

Furthermore, the relationship between EHL and HRQoL varies across contexts and individual characteristics. According to health inequality theory, health outcomes are shaped not only

XSL•FO
RenderX

by individual factors but also by social determinants [38]. Age reflects significant life course characteristics. Older adults often experience cumulative health decline, cognitive impairment, and increased barriers to digital access and use, making their EHL more likely constrained by physiological and cognitive foundations [39]. Younger-old adults typically possess relatively better physical and mental functioning, along with stronger learning abilities. They may more effectively translate EHL into concrete actions for health management and promotion, thereby demonstrating greater health benefits. Regarding residential environments, urban older adults generally benefit from more comprehensive health care systems, denser social support networks, wider internet and smart device coverage, and higher overall educational attainment and digital literacy [40]. These conditions collectively create an enabling environment conducive to accumulating and converting EHL. In contrast, rural older adults generally face multiple constraints, including low accessibility to medical resources, weak digital infrastructure, limited health information channels, and deep-rooted traditional views on aging, exacerbating health inequalities in the digital age [41]. From the perspective of individual adaptive strategies, lifestyle serves as another crucial moderating factor. According to the Selective Compensation Optimization model, individuals cope with resource changes during aging by selecting goals, optimizing resources, and compensating for losses [42]. For older adults with insufficient exercise, EHL may play a crucial compensatory role by promoting other health behaviors to partially offset health risks stemming from physical inactivity [43,44]. Conversely, for those with established exercise habits—whose health behavior systems are relatively well-developed—EHL's impact may manifest more as an optimization support. It can enhance the scientific rigor of exercise routines or integrate diverse health resources, yielding relatively stable incremental benefits.

Overall, existing research has yet to establish an integrated research framework. Most studies focus on single mechanisms such as health-promoting behaviors [45], insomnia and psychological distress [46], sense of coherence [47], and SE of health practices [48]. Meanwhile, existing studies lack exploration of the role of social structural factors and individual adaptive strategies as grouping variables [45,49]. Therefore, our study developed an integrated research framework through theoretical synthesis and conceptual framework construction. Therefore, our study integrates theories and constructs a framework encompassing digital empowerment, psychological-behavioral transformation, comprehensive health outcomes, and multilevel contextual moderation. This framework aims to systematically elucidate the mechanisms underlying the functioning of EHL and identify the key entry points for the delivery of targeted interventions. Based on this framework, the study proposes the following hypotheses:

- Hypothesis 1: EHL is positively associated with HRQoL.
- Hypothesis 2a: ATOA mediates the relationship between EHL and HRQoL.
- Hypothesis 2b: SE mediates the relationship between EHL and HRQoL.
- Hypothesis 3a: The association between EHL and HRQoL differs by age.
- Hypothesis 3b: The association between EHL and HRQoL differs by place of residence.
- Hypothesis 3c: The association between EHL and HRQoL differs by exercise behavior.

Methods

Study Design and Participants

In 2024, our research team independently conducted a special survey titled Digital Inclusion and Healthy Aging System Development. The questionnaire covered demographic information, digital health literacy, and multidimensional indicators of healthy aging. The inclusion criteria were (1) must be at least 55 years of age at the time of the survey, (2) must have resided continuously within the survey area for no less than 6 months, (3) must possess adequate cognitive and communication abilities, and (4) must voluntarily participate in this study and sign an informed consent form before the survey commencement. Exclusion criteria were (1) severe cognitive or psychiatric impairment preventing effective communication, (2) significant PH issues rendering completion of the survey impossible, (3) temporary residents, and (4) individuals who explicitly refused to sign the informed consent form or withdrew from the study midway.

To ensure high data quality and representativeness, a stratified multistage cluster sampling strategy was adopted, combining purposive and random sampling. In the first stage, based on the digital development index, internet penetration rate, and pilot programs of smart health care for older adults published by the National Bureau of Statistics and the Ministry of Industry and Information Technology, along with regional characteristics, population structure, economic development, and resource distribution, 4 representative provinces or autonomous regions were selected, namely, Jiangsu, Shandong, Hubei, and Guangxi. In the second stage, one city from each province was chosen according to the same stratification criteria, namely, Nanjing, Weifang, Enshi, and Nanning. In the third stage, all districts (or counties) were used as the sampling frame, and random sampling identified Gulou District, Xuanwu District, Qixia District, Jiangning District, Qingzhou County, Jianshi County, and Xixiangtang District. In the fourth stage, taking into account factors such as community population size, urban-rural type, and digital environment variation, a total of 57 communities (or villages) were randomly selected from these districts or counties as the final clusters. Eligible older adults within these units were surveyed. The structural characteristics of the final sample demonstrate that the sampling design achieved balance in economic level, urban-rural type, and digital access status.

Data collection was completed through surveyor-led face-to-face structured interviews. Fieldwork was conducted by a team of trained surveyors who had undergone a standardized training program covering survey objectives, questionnaire content, interview techniques, and ethical considerations.

Questionnaire Development and Pretesting

The questionnaire was initially developed in Chinese. To ensure content validity and comprehensibility for the target population, a pretest was conducted among a convenience sample of 200

XSL•FO
RenderX

older adults from nonsampled communities. Based on feedback regarding item clarity, response options, and interview duration, revisions were made to the wording and formatting.

Translation

As this study aims for potential international dissemination, the questionnaire was professionally translated into English by bilingual public health researchers following a forward-backward translation procedure to ensure conceptual equivalence. The English version was used for reporting purposes only.

Quality Control Procedures

Multiple quality control measures were implemented throughout the data collection process. (1) All surveyors were required to pass a mock interview assessment before fieldwork. (2) On-site supervisors randomly observed approximately 10% of the interviews and provided immediate feedback. (3) Surveyors used tablet computers equipped with a survey application to minimize data entry errors. (4) Supervisors reviewed completed questionnaires daily for completeness and consistency. (5) A random subset of participants was selected for telephone callback to verify key information.

Response Rate and Data Cleaning

A total of 8731 questionnaires were distributed to eligible individuals, resulting in a final valid response rate of 95.8%. Figure 1 illustrates the screening process for the study population.

Figure 1. Flowchart showing the selection of participants. EHL: electronic health literacy; HRQoL: health-related quality of life; LS: life satisfaction; MH: mental health; PH: physical health.

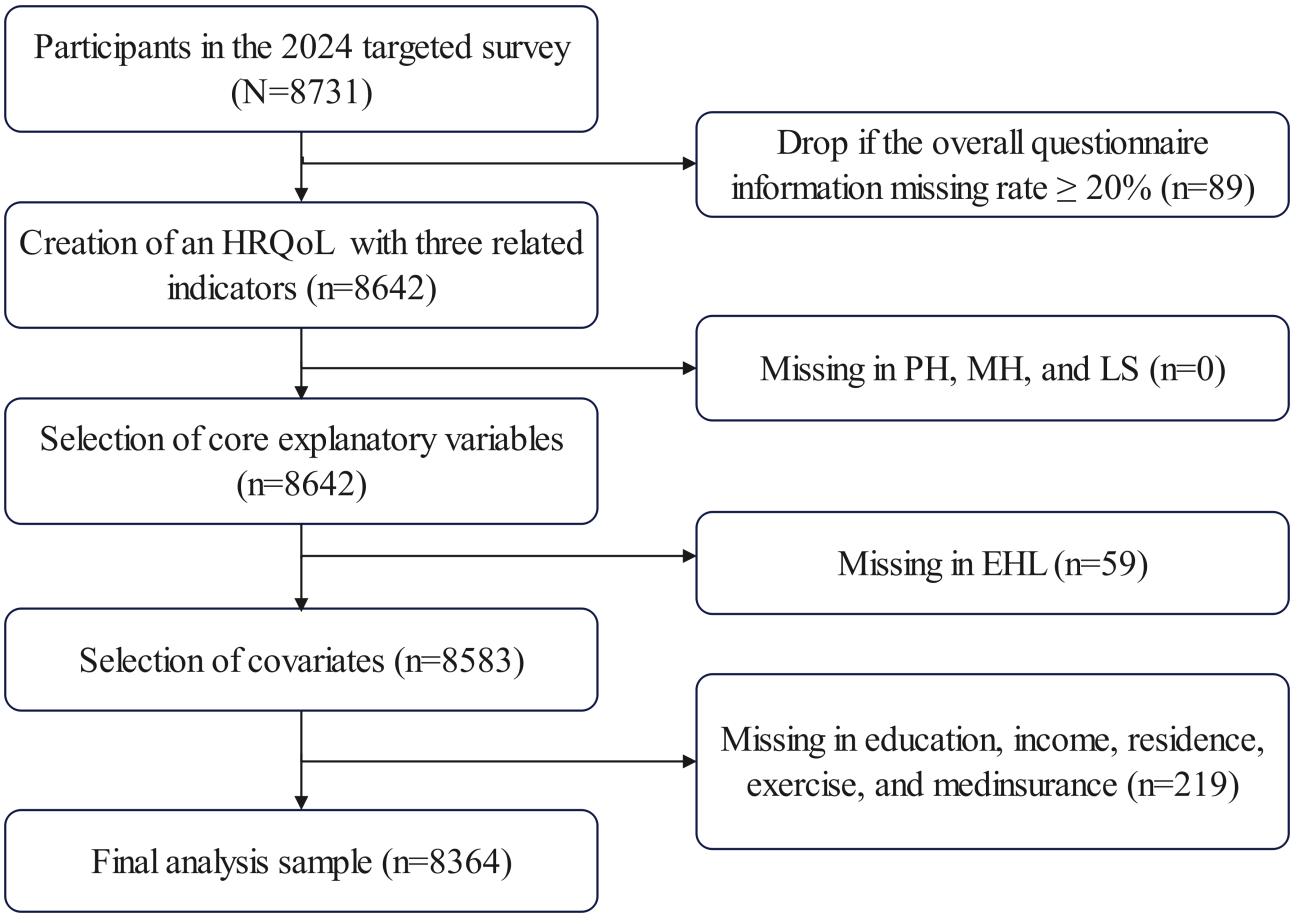

Measures

Assessment of HRQoL

The primary dependent variable was HRQoL. Quality of life is valued as a complement to conventional notions of health and functional status. An optimal health assessment should, therefore, encompass physical, social, and psychological functioning alongside quality of life [50]. Accordingly, this study assessed HRQoL using 3 indicators, namely, PH, MH, and LS.

PH was measured using the EuroQol Visual Analogue Scale (EQ-VAS), which is part of the EQ-5D instrument that evaluates individuals' overall perception of their health status. Respondents rated their current health status on a visual scoring ruler ranging from 0 ("worst conceivable health state") to 100 ("best conceivable health state") [51]. The EQ-VAS is widely recognized as a straightforward, subjective instrument designed to capture general health status in clinical and population-based health studies [52].

MH was evaluated using the World Health Organization-Five Well-Being Index (WHO-5), which assesses subjective well-being and risk of depression. The WHO-5 questionnaire includes 5 specific measurement items, each rated on a 6-point Likert scale ranging from 0 ("at no time") to 5 ("all of the

XSL•FO
RenderX

time"). The raw total score, which spans from 0 to 25, is conventionally converted to a percentage scale by multiplying by 4, resulting in a range of 0 to 100. Higher values on this scale reflect a more favorable MH status [53,54]. In this study, the WHO-5 was found to have high internal consistency (Cronbach α=0.949; Kaiser-Meyer-Olkin [KMO]=.886).

LS is an individual's overall assessment of their current life situation. It was measured with the Satisfaction With Life Scale (SWLS), developed by Diener et al [55]. As a classic and widely validated tool across populations and cultural contexts [56], the SWLS includes 5 items, each rated on a 7-point agreement scale. The summative score ranges from 5 to 35, and higher values mean better LS. The SWLS demonstrated high internal consistency in our sample (Cronbach α=0.961; KMO=.914).

Assessment of EHL

The core independent variable was EHL. It was measured using the 8-item eHealth Literacy Scale (eHEALS). Items are scored on a 5-point agreement scale (1="strongly disagree," 5="strongly agree"), yielding a total from 8 to 40. Elevated scores correspond to a more positive self-assessment of one's ability to obtain, judge, and use eHealth information [57]. In our study, the instrument showed excellent internal consistency (Cronbach α=0.986; KMO=.943). The scale has been widely applied in health behavior research across diverse populations and cultural settings [58].

Assessment of Mediating Variables

Two mediators included in this study were ATOA and SE. ATOA refers to an individual's expectations, cognitive representations, and perceptions regarding their own aging process. It was measured by means of the Attitudes Toward One's Own Aging subscale of the Philadelphia Geriatric Center Morale Scale. This subscale includes 5 items rated on a 5-point Likert scale (1="strongly disagree," 5="strongly agree"), with 2 negative items reverse-coded. Average scores were calculated, with higher scores indicating more positive ATOA [59]. The internal consistency in this study was acceptable (Cronbach α=0.745; KMO=.811).

SE serves as a key internal motivator that drives individual behavior, influences emotional experiences, and ultimately shapes outcomes. SE was assessed using the General Self-Efficacy Scale (GSES), developed by Schwarzer and Jerusalem in 1995. The scale comprises 10 items rated on a 4-point Likert scale (1="not at all true," 4="exactly true"). Average scores were computed, with higher scores reflecting stronger SE [60]. Reliability was excellent in our study (Cronbach α=0.965; KMO=.968).

Ascertainment of Covariates

To minimize potential confounding, this study incorporated a range of covariates based on prior literature [45,61]. These included demographic characteristics (age, gender, ethnicity, region, residence, education, and spouse status), socioeconomic factors (income, living arrangements, employment status, and number of children), lifestyle behaviors (caregiving for grandchildren, smoking, drinking, and exercise), social security status (participation in medical and pension insurance), and health-related variables (number of chronic diseases and use of health services). Detailed coding of variables is shown in Table 1.

XSL•FO
RenderX

Table . Variables encoding.

Variables	Definition	Type
Age	Actual age in years	Continuous
Sex	0=Female; 1=Male	Categorical
Ethnicity	0=Minority; 1=Han	Categorical
Region	1=Eastern; 2=Central; 3=Western	Categorical
Residence	0=Rural; 1=Urban	Categorical
Education	1=Below primary school; 2=Primary school; 3=Junior school; 4=Senior or vocational school; 5=Bachelor and above	Categorical
Spouse	0=Without; 1=With	Categorical
Income[a]	0=None; 1 =≤1000; 2=1001 - 2000; 3=2001 - 5000; 4 =≥5001	Categorical
Living	0=With family; 1=Empty nest	Categorical
Employment	0=Not; 1=Currently	Categorical
Children	0=None; 1=Have	Categorical
Grandcare	0=No; 1=Yes	Categorical
Smoking	0=No; 1=Yes	Categorical
Drinking	0=No; 1=Yes	Categorical
Exercise	0=No; 1=Yes	Categorical
Medical insurance	0=No; 1=Yes	Categorical
Pension insurance	0=No; 1=Yes	Categorical
Chronic	Actual number	Continuous
Outpatient	Actual frequency	Continuous
Hospitalization	Actual frequency	Continuous

[a]Income is denominated in Chinese yuan.

Statistical Analysis

All statistical analyses were conducted using SPSS 25.0 (IBM Corp) and R 4.4.3 (R Foundation for Statistical Computing). Descriptive statistics, univariate analysis, and correlation analysis were performed. Continuous variables are reported as means and SDs, while categorical variables are reported as frequencies and percentages. To scrutinize the potential for common method variance, Harman single-factor test was performed. Given the correlations among the 3 healthy aging indicators, this study used seemingly unrelated regressions to simultaneously estimate the 3 equations, thereby improving estimation efficiency [62]. To control the increased familywise error rate due to multiple testing, the significance of the associations between EHL and the 3 HRQoL outcomes was assessed using a Bonferroni correction (adjusted α=.05/3). The mediation model, which specified ATOA and SE as parallel mediators, was tested with the PROCESS macro for SPSS (Model 4) [63]. To investigate the significance of the indirect effects, a bootstrapping procedure was applied based on 5000 resamples. Within each HRQoL dimension (PH, MH, and LS), the significance of indirect effects was evaluated using the Holm-Bonferroni sequential correction procedure. Heterogeneity analysis was conducted through subgroup regression. For these exploratory subgroup analyses, the false discovery rate (FDR) across all tests concerning the association between EHL and outcomes was controlled using the Benjamini-Hochberg procedure. All corrections pertained specifically to the hypothesis tests involving EHL. Statistical significance was tested using a 2-tailed test with an α level of .05, and the corrected significance level was fully considered.

Ethical Considerations

This study was approved by the Ethics Committee of Zhongda Hospital, Southeast University (2024ZDSYLL294-Y01). All participants provided informed consent before their inclusion in the study. Anonymous data collection was used to ensure confidentiality, and no personal identifiers, such as names, were included in our study materials. No financial or material compensation was provided for participation in the study. The research strictly adhered to the principles of the Declaration of Helsinki (1964), the Council for International Organizations of Medical Sciences (CIOMS) International Ethical Guidelines, and the World Health Organization's standards and procedures for research involving human participants.

XSL•FO
RenderX

Results

Descriptive Statistics

A total of 8364 participants were included in this study (3546 men and 4818 women), with a mean age of 70.83 (SD 8.39) years. In terms of residence, rural residents slightly outnumbered urban residents (56.5% vs 43.5%). Regarding education level, 31.3% (2614/8364) had not completed primary school, while only 3.9% (329/8364) held a bachelor's degree or higher. In terms of lifestyle, the vast majority did not smoke (7036/8364, 84.1%) or drink alcohol (6743/8364, 80.6%). Regarding chronic disease prevalence, participants had an average of 1.26 types of chronic diseases (SD 1.14). The average EHL score was 20.13 (SD 10.58), indicating a relatively low level of EHL. The mean scores for PH, MH, and LS were 72.01 (SD 18.28), 71.18 (SD 20.47), and 26.71 (SD 5.52), respectively. Detailed participant characteristics are presented in Table 2.

XSL•FO
RenderX

Table . Descriptive statistics (N=8364).

Variables	Values
Age (years), mean (SD)	70.83 (8.39)
Sex, n (%)	
Female	4818 (57.6)
Male	3546 (42.4)
Ethnicity, n (%)	
Minority	1019 (12.2)
Han	7345 (87.8)
Region, n (%)	
Eastern	5259 (62.9)
Central	1670 (20.0)
Western	1435 (17.2)
Residence, n (%)	
Rural	4726 (56.5)
Urban	3638 (43.5)
Education, n (%)	
Below primary school	2614 (31.3)
Primary school	1422 (17.0)
Junior school	1979 (23.7)
Senior or vocational school	2020 (24.2)
Bachelor and above	329 (3.9)
Spouse, n (%)	
Without	1785 (21.3)
With	6579 (78.7)
Income[a], n (%)	
None	1253 (15.0)
≤1000	2310 (27.6)
1001 - 2000	1064 (12.7)
2001 - 5000	2706 (32.4)
≥5001	1031 (12.3)
Living, n (%)	
With family	7207 (86.2)
Empty nest	1157 (13.8)
Employment, n (%)	
Not	6596 (78.9)
Currently	1768 (21.1)
Children, n (%)	
None	352 (4.2)
Have	8012 (95.8)
Grandcare, n (%)	
No	7089 (84.8)
Yes	1275 (15.2)
Smoking, n (%)	

XSL•FO
RenderX

Variables	Values
No	7036 (84.1)
Yes	1328 (15.9)
Drinking, n (%)	
No	6743 (80.6)
Yes	1621 (19.4)
Exercise, n (%)	
No	3127 (37.4)
Yes	5237 (62.6)
Medical insurance, n (%)	
No	156 (1.9)
Yes	8208 (98.1)
Pension insurance, n (%)	
No	532 (6.4)
Yes	7832 (93.6)
Chronic, mean (SD)	1.26 (1.14)
Outpatient, mean (SD)	2.55 (4.82)
Hospitalization, mean (SD)	0.23 (0.67)

[a]Income is denominated in Chinese yuan.

Common Method Bias Test

To assess the potential for common method bias, Harman single-factor test was performed on all measurement items. Six factors with eigenvalues greater than 1 were identified, with the first factor accounting for 38.76% of the variance—below the commonly accepted 50% threshold [64]. Consequently, these findings provide evidence that common method bias did not pose a significant threat to the validity of the study's results.

Univariate Analysis

Results of the univariate analysis are presented in Table 3. Significant differences in PH, MH, and LS were observed across region, residence, education, income, spouse, grandcare, and exercise (all $P<.001$). Overall, older adults living in eastern regions, with higher education and income levels, having a spouse, exercising regularly, or providing grandchild care exhibited better outcomes in healthy aging, whereas the effects of other factors were relatively weaker.

XSL•FO
RenderX

Table . Univariate analysis results.

Variables	PH[a]			MH[b]			LS[c]		
	t test (*df*)	*F* test (*df*)	*P* value	*t* test (*df*)	*F* test (*df*)	*P* value	*t* test (*df*)	*F* test (*df*)	*P* value
Sex	−0.345 (8362)	—[d]	.73	0.791 (8362)	—	.43	2.736 (8362)	—	.006
Ethnicity	−0.543 (8362)	—	.59	−6.845 (8362)	—	<.001	−6.392 (8362)	—	<.001
Region	—	152.358 (2, 8361)	<.001	—	242.853 (2, 8361)	<.001	—	265.219 (2, 8361)	<.001
Residence	6.003 (8362)	—	<.001	−6.136 (8362)	—	<.001	3.661 (8362)	—	<.001
Education	—	18.259 (4, 8359)	<.001	—	42.114 (4, 8359)	<.001	—	21.243 (4, 8359)	<.001
Spouse	−4.252 (8362)	—	<.001	−5.838 (8362)	—	<.001	−6.546 (8362)	—	<.001
Income	—	17.465 (4, 8359)	<.001	—	39.202 (4, 8359)	<.001	—	48.025 (4, 8359)	<.001
Living	1.248 (8362)	—	.21	2.326 (8362)	—	.02	2.588 (8362)	—	.01
Employment	−1.261 (8362)	—	.21	3.691 (8362)	—	<.001	0.854 (8362)	—	.39
Children	6.070 (8362)	—	<.001	−0.484 (8362)	—	.63	−1.143 (8362)	—	.25
Grandcare	−8.881 (8362)	—	<.001	−6.695 (8362)	—	<.001	−5.350 (8362)	—	<.001
Smoking	−0.561 (8362)	—	.58	3.715 (8362)	—	<.001	3.752 (8362)	—	<.001
Drinking	−1.350 (8362)	—	.18	−0.764 (8362)	—	.45	4.253 (8362)	—	<.001
Exercise	−8.346 (8362)	—	<.001	−15.050 (8362)	—	<.001	−11.874 (8362)	—	<.001
Medical insurance	−0.797 (8362)	—	.43	−2.545 (8362)	—	.01	−4.622 (8362)	—	<.001
Pension insurance	1.026 (8362)	—	.31	−1.204 (8362)	—	.23	−0.751 (8362)	—	.45

[a]PH: physical health.

[b]MH: mental health.

[c]LS: life satisfaction.

[d]Not applicable.

Correlation Analysis

Given that our study includes multiple variable types and some continuous variables deviate from the normal distribution, Spearman correlation analysis was used to assess the associations among age, chronic, outpatient, hospitalization, EHL, ATOA, SE, PH, MH, and LS. Results are shown in Figure 2. EHL was positively correlated with ATOA, SE, and HRQoL (ρ=0.102 - 0.370; P<.05). In contrast, aside from age and hospitalizations, the number of chronic diseases and outpatient visits was negatively correlated with EHL, ATOA, SE, and HRQoL (ρ= –0.331 to 0.030; P<.05).

XSL•FO
RenderX

Figure 2. Correlation heatmap. *P<.05; **P<.01; ***P<.001. ATOA: attitudes toward own aging; EHL: electronic health literacy; LS: life satisfaction; MH: mental health; PH: physical health; SE: self-efficacy.

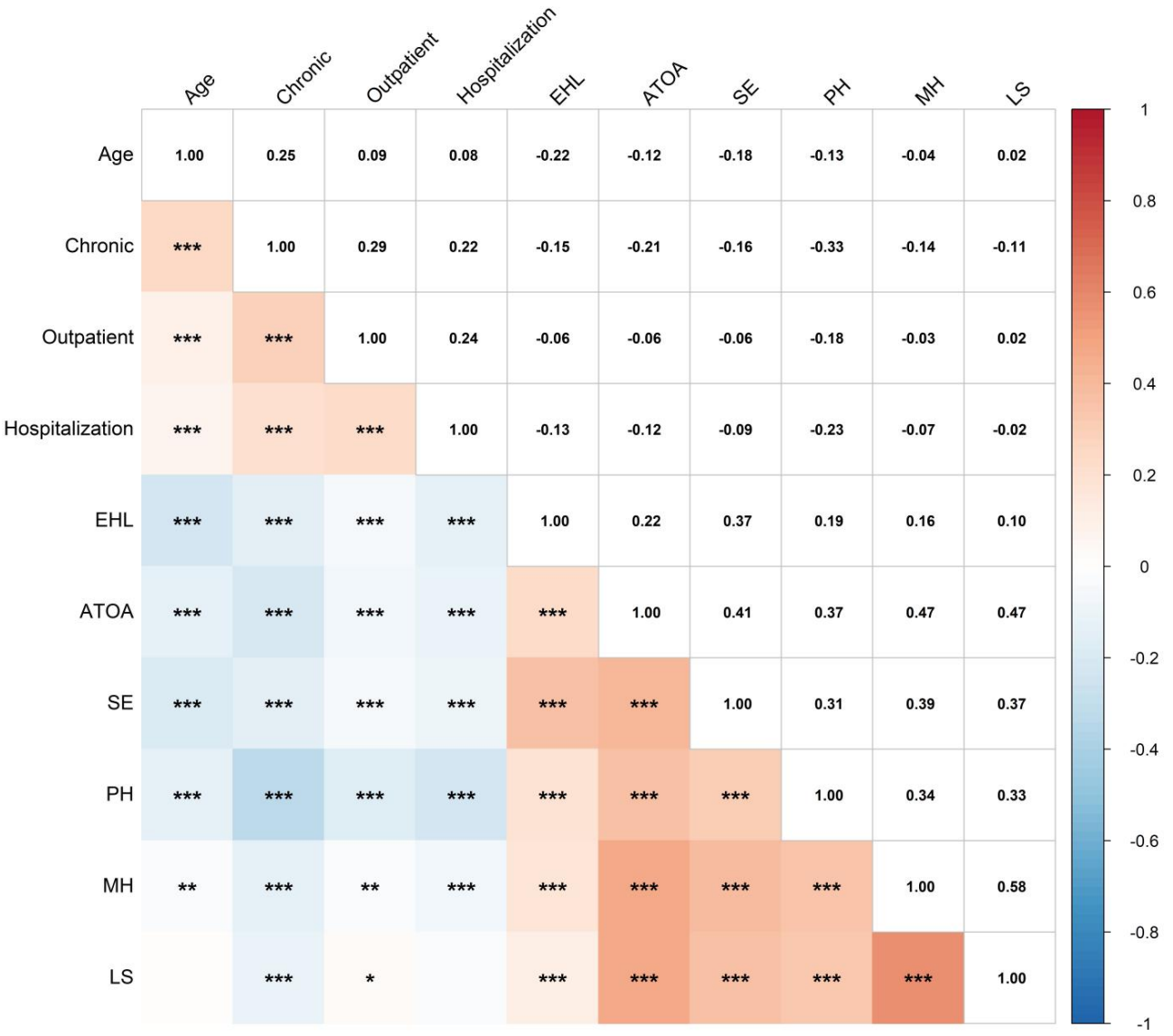

Basic Regression

Based on the univariate analysis results, seemingly unrelated regressions were used to simultaneously estimate PH, MH, and LS (Table 4). In Model 1, EHL was positively associated with PH, MH, and LS (all P<.001). After controlling for covariates in Model 2, the positive associations between EHL and the 3 outcomes remained significant (all P<.001). Notably, these associations retained their statistical significance even after applying a Bonferroni correction for the 3 comparisons (adjusted α=0.0167). These findings confirm that EHL exerts a robust positive influence on multiple dimensions of HRQoL, thereby supporting hypothesis 1.

XSL•FO
RenderX

Table . Baseline regression results.[a]

Variables	Model 1[b]			Model 2[c]		
	PH[d]	MH[e]	LS[f]	PH	MH	LS
EHL[g]	0.235[h] (0.018)	0.273[h] (0.021)	0.064[h] (0.006)	0.273[h] (0.189)	0.190[h] (0.026)	0.082[h] (0.007)
Age	—[i]	—	—	0.147 (0.025)	0.085[j] (0.030)	0.053[h] (0.008)
Ethnicity	—	—	—	−6.734[h] (0.630)	−0.756 (0.811)	−0.902[h] (0.197)
Region						
Central				−7.180[h] (0.501)	−7.655[h] (0.642)	−2.741[h] (0.167)
Western				−15.127[h] (0.644)	−10.843[h] (0.754)	−3.639[h] (0.199)
Residence				−6.761[h] (0.656)	−4.701[h] (0.750)	−3.639[h] (0.202)
Education						
Primary school				−0.399 (0.519)	−1.018 (0.442)	−0.212 (0.170)
Junior school				1.306[k] (0.550)	0.442 (0.675)	0.042 (0.179)
Senior or vocational school				4.399[h] (0.686)	3.041[h] (0.788)	0.922[h] (0.206)
Bachelor and above				4.296[h] (1.098)	4.097[j] (1.321)	1.203[j] (0.348)
Spouse				0.261 (0.624)	2.293[j] (0.762)	0.874[h] (0.192)
Income						
≤1000				0.613 (0.517)	1.929[j] (0.710)	2.074[h] (0.179)
1001–2000				1.888[j] (0.629)	2.372[j] (0.785)	1.450[h] (0.207)
2001–5000				1.723[k] (0.756)	6.774[h] (0.917)	2.034[h] (0.235)
≥5001				2.396[j] (0.832)	5.383[h] (1.085)	2.609[h] (0.285)
Living				0.515 (0.693)	0.406 (0.871)	0.035 (0.226)
Grandcare				1.468[j] (0.446)	2.225[h] (0.573)	0.268 (0.155)
Smoking				−0.192 (0.475)	−1.611[j] (0.626)	−0.586[h] (0.158)
Exercise				2.078[h] (0.397)	4.939[h] (0.483)	1.064[h] (0.125)
Medical insurance				1.198 (1.332)	4.118[k] (1.688)	2.059[h] (0.485)
Chronic				−2.903[h] (0.174)	−2.027[h] (0.207)	−0.569[h] (0.056)
Outpatient				−0.290[h] (0.036)	−0.078 (0.042)	−0.012 (0.012)
Hospitalization				−4.188[h] (0.312)	−1.259[h] (0.344)	−0.256[j] (0.097)
Observations	8364	8364	8364	8364	8364	8364
R^2	0.019	0.02	0.015	0.19	0.116	0.155

[a]All categorical variables are coded with the first group as the reference category. Robust standard errors are shown in parentheses. The significance of the association between EHL and HRQoL dimensions was corrected using the Bonferroni method (corrected α=0.0167), with all P_{Bonf} <.05.

[b]Model 1: Unadjusted.

[c]Model 2: Adjusted for age, ethnicity, region, residence, education, spouse, income, living, grandcare, smoking, exercise, medical insurance, chronic, outpatient, and hospitalization.

[d]PH: physical health.

[e]MH: mental health.

[f]LS: life satisfaction.

[g]EHL: electronic health literacy.

XSL•FO
RenderX

[h]*P*<.001.
[i]Not applicable.
[j]*P*<.01.
[k]*P*<.05.

Mediation Analysis

Mediation analysis was undertaken using PROCESS Model 4 in SPSS, with ATOA and SE specified as parallel mediators. After controlling for covariates, the model included EHL as the predictor, ATOA and SE as mediators, and PH, MH, and LS as outcomes (Table S1 in Multimedia Appendix 1). Results showed that EHL positively predicted both ATOA (β=0.015; *P*<.001) and SE (β=0.019; *P*<.001), indicating that higher EHL was associated with more positive aging attitudes and stronger self-efficacy.

For PH, EHL maintained a significant positive association (β=0.128; *P*<.001), and both ATOA (β=3.270; *P*<.001) and SE (β=5.129; *P*<.001) strongly predicted PH. Thus, EHL directly promoted PH and also indirectly contributed via ATOA and SE. For MH, the direct association between EHL and MH became negative after including the mediators (β=–.069; *P*<.01). However, both ATOA (β=9.417; *P*<.001) and SE (β=6.929; *P*<.001) exerted strong positive effects on MH, suggesting that the positive relationship between EHL and MH was almost entirely mediated by ATOA and SE, with evidence of a suppression effect. For LS, the direct effect of EHL was nonsignificant, while both ATOA (β=2.583; *P*<.001) and SE (β=2.250; *P*<.001) positively predicted LS. This indicates that the association between EHL and LS was fully mediated by ATOA and SE.

Following established recommendations [65], bootstrapping with 5000 resamples was used to assess mediation significance via 95% CI (Table 5; Figure 3). For PH, the direct effect of EHL was significant (95% CI 0.086-0.170), and the indirect effects of ATOA (95% CI 0.038-0.059) and SE (95% CI 0.081-0.112) were also significant, with total indirect effects accounting for 52.7% of the total effect. For MH, the direct effect of EHL was negative and significant (95% CI –0.115 to –0.024), but the indirect effects through ATOA (95% CI 0.119-0.158) and SE (95% CI 0.112-0.149) were positive, yielding a positive total effect (95% CI 0.151-0.248)—indicating inconsistent mediation. For LS, the direct effect of EHL was nonsignificant (95% CI –0.012 to 0.012), while the indirect effects of ATOA (95% CI 0.033-0.043) and SE (95% CI 0.036-0.047) were significant, with total indirect effects accounting for 100% of the total effect. Furthermore, after using Holm-Bonferroni sequential correction, it was found that the bootstrap CI for all indirect effects did not include zero, indicating that all mediating paths remained statistically significant. These results support hypotheses 2a and 2b.

XSL•FO
RenderX

Table . Significance test for mediating effect[a].

Variables	Effect	Paths	β (SE)	95% CI	Proportion
PH[b]	Direct	EHL[c] →PH	.128 (0.022)	[0.086-0.170]	45.8%
	Indirect	EHL→ATOA[d] →PH	.048 (0.005)	[0.038-0.059]	17.6%
		EHL→SE[e]→PH	.096 (0.008)	[0.081-0.112]	35.2%
		Total Indirect	.144 (0.009)	[0.126-0.163]	52.7%
	Total		.273 (0.021)	[0.231-0.314]	—[f]
MH[g]	Direct	EHL→MH	–.069 (0.023)	[–0.115 to –0.024]	—
	Indirect	EHL→ATOA→MH	.138 (0.010)	[0.119-0.158]	—
		EHL→SE→MH	.130 (0.010)	[0.112-0.149]	—
		Total Indirect	.268 (0.014)	[0.240-0.296]	—
	Total		.199 (0.025)	[0.151-0.248]	—
LS[h]	Direct	EHL→LS	.000 (0.006)	[–0.012 to 0.012]	0.0%
	Indirect	EHL→ATOA[d]→LS	.038 (0.003)	[0.033-0.043]	47.5%
		EHL→SE→LS	.042 (0.003)	[0.036-0.047]	52.5%
		Total Indirect	.080 (0.004)	[0.071-0.088]	100%
	Total		0.080 (0.007)	[0.067-0.093]	—

[a]Proportion is the effect size ratio; for mental health, direct and total effects have opposite signs (inconsistent mediation), so proportion is not reported. After applying the Holm-Bonferroni sequential correction procedure, the bootstrapCI for all indirect effects did not include zero.

[b]PH: physical health.

[c]EHL: electronic health literacy.

[d]ATOA: attitudes toward own aging.

[e]SE: self-efficacy.

[f]Not applicable.

[g]MH: mental health.

[h]LS: life satisfaction.

Figure 3. Mediation model. ***P*<.01; ****P*<.001. ATOA: attitudes toward own aging; EHL: electronic health literacy; LS: life satisfaction; MH: mental health; PH: physical health; SE: self-efficacy.

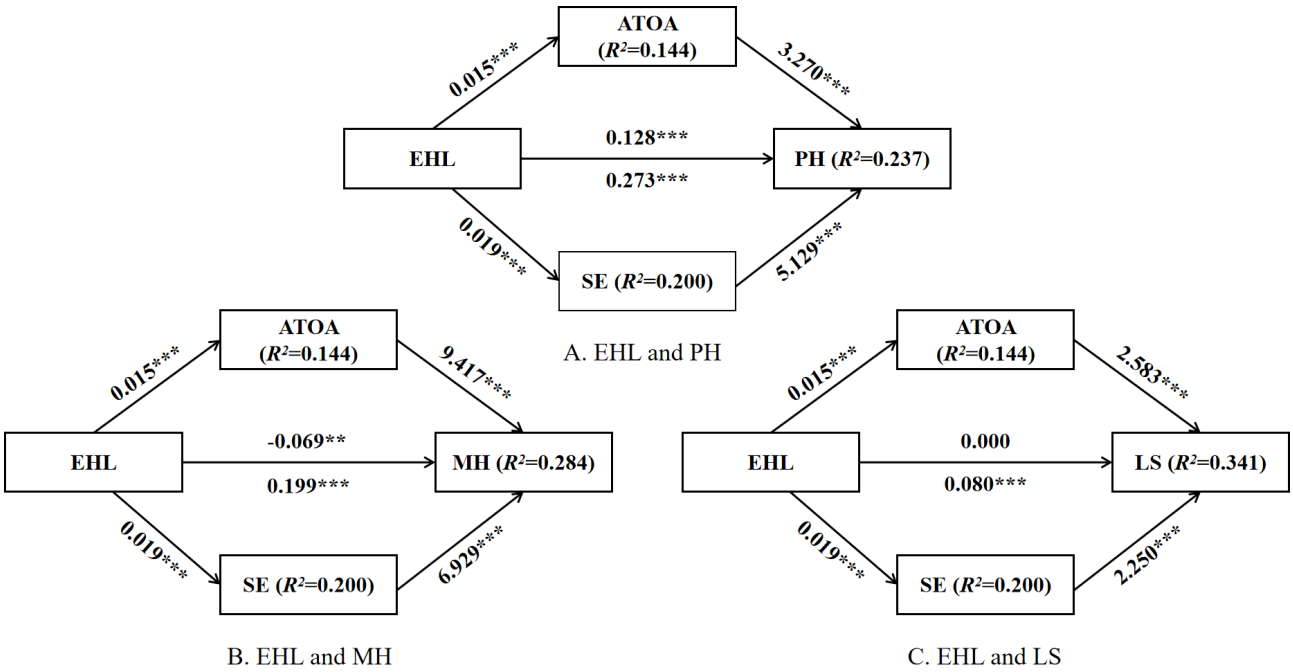

XSL•FO
RenderX

Heterogeneity Test

Our study further conducted subgroup analyses stratified by age, residence, and physical activity behavior, and the results are presented in Tables S2-S4 in Multimedia Appendix 1. Given that subgroup analyses involve multiple comparisons, we performed FDR correction (Benjamini-Hochberg) for all association tests between EHL and different HRQoL dimensions across all subgroups. All significant associations remained significant after correction (all $P_{BH\text{-}FDR}$<.05), indicating that the subgroup analysis results are robust.

The role of EHL showed notable age differences. Among the younger-old adults (≤74 years), EHL had the strongest association with PH (β=0.288; *P*<.001), and also demonstrated significant positive effects on MH (β=0.188; *P*<.001) and LS (β=0.089*; P*<.001). This indicates that, within this group, EHL could more comprehensively enhance their HRQoL. In contrast, among the older-old adults (>74 years), EHL still had significant positive predictive effects on all 3 dimensions, but its strength was generally weaker. The coefficient for PH decreased to 0.222 (*P*<.001), and the promoting effects on MH (β=0.175; *P*<.001) and LS (β=0.060; *P*<.001) were also lower than those in the younger group. These results support hypothesis 3a.

Significant urban-rural differences were observed. Among rural older adults, EHL was significantly and positively associated with PH (β=0.215; *P*<.001), but its association with MH was nonsignificant, and its effect on LS was positive but relatively weak (β=.019; *P*=.032). In contrast, among urban older adults, EHL was more strongly associated with PH (β=0.237; *P*<.001), and its associations with MH and LS were also considerably stronger than in rural areas. Overall, EHL had a broader impact on HRQoL in urban older adults, while in rural populations, its health value was concentrated primarily in PH. These findings support hypothesis 3b.

Associations also show differences with varying exercise habits. Among nonexercisers, EHL had its strongest positive association with PH (β=0.322; *P*<.001), suggesting that EHL effectively compensated for health deficits due to lack of exercise and served as a core support for improving PH. EHL also positively predicted MH (β=0.196; *P*<.001) and LS (β=0.067;*P*<.001) in this group. In contrast, among exercisers, the health effects of EHL, though still significant, were weaker. The association with PH dropped to β=0.193 (*P*<.001), while its effect on MH (β=0.147; *P*<.001) was also lower than in nonexercisers, and the effect on LS showed little difference. These findings suggest that nonexercisers' HRQoL was more strongly dependent on EHL, whereas exercisers already benefited from the baseline health improvements of physical activity, limiting the additional value of EHL. These results support hypothesis 3c.

Discussion

Principal Findings

This study found that EHL is significantly associated with HRQoL among older adults in China. The enhancement of EHL contributes to improving HRQoL in older adults. Previous studies have established a positive association between general health literacy and HRQoL [66], while relevant systematic reviews have clarified the intrinsic relationship between EHL and multidimensional health outcomes [25]. For Chinese older adults, a prior study conducted in Jinan City initially explored the relationship between EHL and HRQoL [45]. These findings are consistent with our conclusions. Our study further expands the representativeness of the sample and the generalizability of conclusions, providing empirical evidence for extending research in this field to broader populations of Chinese older adults.

The mechanism of EHL varies across different dimensions of HRQoL. EHL exerted a direct effect on PH, whereas its associations with MH and LS were almost entirely mediated by ATOA and SE. Older adults with higher EHL can more efficiently retrieve, comprehend, and apply online health information, enabling more scientifically grounded disease prevention, symptom management, medication adherence, and health care decision-making [67]. This process involves a direct conversion of cognitive and behavioral skills, yielding immediate, instrumental positive effects on PH. This direct effect fundamentally reflects the immediacy of translating information literacy into health practices, highlighting EHL's core value as a capacity for health information use [68].

An inconsistent mediation was observed in the MH model. The direct effect of EHL on MH was negative, yet the indirect effects via ATOA and SE were positive, resulting in a positive total effect. Possessing high EHL may also expose individuals to an overwhelming volume of health information, potentially triggering cognitive overload and uncertainty [69]. This fosters negative emotions such as health anxiety and cyberhypochondria. Catastrophizing minor symptoms further exacerbates psychological burden [70]. Older adults may intensify perceptions of health limitations or experience frustration from digital operational barriers, amplifying direct negative impacts. However, EHL can exert protective effects by fostering positive psychological mechanisms, as ATOA reduces age-related health fears, while SE enhances confidence in tackling health challenges [71,72]. Both collectively boost psychological resilience and emotional regulation capacity. This indicates that benefits to older adults' MH stem not from the information itself, but from the psychological coping mechanisms constructed through EHL. Without such transformative processes, high EHL may instead pose risks to the psychological well-being of older adults.

For LS, the association between EHL and outcomes was fully mediated by ATOA and SE. As a cognitive evaluation of overall life quality, LS is more closely associated with universal psychological resources and meaning frameworks [73]. EHL, as a tool-based skill, does not directly contribute to overall life evaluations. Instead, EHL is more likely to enhance LS by strengthening positive aging attitudes (enabling older adults to embrace the life course) and leveraging SE to implement health-promoting behaviors. In other words, EHL functions as an enabler: the knowledge and sense of control afforded by EHL are filtered and sublimated through an individual's meaning system and core beliefs before being translated into more holistic LS.

XSL•FO
RenderX

The analysis also revealed heterogeneity. The health benefits gained by the younger-old from EHL are the most comprehensive and pronounced, indicating that their physiological and cognitive reserves can more effectively translate health literacy into well-being. The intensity of benefits for the older-old adults generally diminishes, potentially stemming from the fact that the marginal benefits of EHL may decrease with age due to cumulative health deficits and the dominance of aging processes [74]. Urban older adults benefited more comprehensively from EHL, likely due to stronger resource conditions that facilitated the translation of literacy into gains in MH and LS, while among rural older adults, the benefits of EHL were concentrated on PH [75]. Similarly, nonexercisers derived greater benefits from EHL, suggesting that it compensates for health deficits due to physical inactivity [76]. Among exercisers, the marginal effects of EHL were weaker, as they already benefited from baseline advantages of physical activity.

Contributions and Implications

This study makes 3 key contributions. First, it extends the literature by showing that EHL is linked not only to PH but also to MH and LS. Second, it identifies ATOA and SE as psychosocial mechanisms through which EHL influences HRQoL. Third, it reveals heterogeneity across subgroups, highlighting the compensatory role of EHL for vulnerable populations such as rural residents and nonexercisers.

These findings confirm the necessity of integrating EHL into healthy aging policies. Policymakers should treat EHL as a social determinant of health and incorporate it into digital inclusion and aging strategies, with particular attention to disadvantaged groups. Practical approaches include (1) recruiting family doctors and volunteers to form a silver-haired lecturer team, implementing EHL training programs at the community level, and establishing personalized advancement pathways for literacy development. The mature models of Shanghai's "Senior Digital Life Workshops" and "Senior Health Cloud Classroom" can serve as references. (2) Rooted in China's unique filial piety and integrity culture, advocating intergenerational interaction models featuring digital reciprocity, mutual learning, and generational integration. (3) With reference to the practical experience of aging-friendly digital infrastructure construction in provinces such as Jiangsu, Zhejiang, and Guangdong, nationwide standardized upgrades and renovations of age-friendly digital infrastructure should be promoted. (4) Expanding digital health services and digital life services that are compatible with the literacy level of the older adults, such as disease prevention and management, health promotion, MH care, and medical electronic commerce, to effectively meet the needs of the older adults. Overall, the comprehensive application of strategies such as community training, intergenerational teaching, infrastructure upgrading projects, and inclusive service expansion can enhance older adults' confidence in using digital health resources, foster a more positive attitude toward aging, strengthen SE, and ultimately improve HRQoL. In addition, EHL should be incorporated into the policy evaluation framework. For example, in the policy evaluation of advancing the informatization construction of compact county medical communities, indicators such as the improvement rate of EHL among the older adults in the county and the use rate of digital health tools among rural older adults should be included in the evaluation system. This measure can align digital health initiatives with the public health goals of health equity and digital inclusion, ensuring that the development of digital health will not widen existing health gaps but instead serve as a catalyst for promoting universal healthy aging in the digital era.

Limitations

Inevitably, our study has several limitations. First, the cross-sectional design precludes causal inference; longitudinal follow-ups are needed. Second, although the sampling ensured regional and population representativeness, generalizability to other cultural contexts remains limited. Third, while common method bias was tested and found to be minimal, the influence of unmeasured confounding cannot be completely excluded. For instance, our study did not account for confounding factors among older adults. Finally, our study treats ATOA and SE as parallel mediating variables. Future research should explore the potential dynamic interdependence and interactions between them to deepen our understanding of the mediating mechanism.

Conclusions

In conclusion, EHL is strongly associated with HRQoL among older adults. Its association with PH appears relatively direct, whereas its associations with MH and LS are largely mediated by ATOA and SE. The observed heterogeneity by age, residence, and exercise behavior suggests that targeted interventions are needed for specific subgroups. Overall, this study unveils the potential of EHL as a key lever for promoting healthy aging in the digital era, emphasizing the urgency of bridging the digital divide and cultivating EHL to support the well-being of older adults.

Acknowledgments

The authors are thankful to the editors and reviewers for their comments and feedback.

Funding

This study was supported by the Major Project of the National Social Science Fund of China (grant number 23&ZD188).

Data Availability

The datasets generated or analyzed during this study are not publicly available due to ethical restrictions as the data contain potentially identifying information but are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: YW, JY
Data curation: YW, JY
Formal analysis: YW, JY
Investigation: YW, JY
Methodology: YW, JY
Project administration: YW
Software: JY, BD
Resources: BD
Supervision: BD
Validation: BD
Writing – original draft: YW
Writing – review & editing: BD

Conflicts of Interest

None declared.

Multimedia Appendix 1
Mediation analysis and heterogeneity analysis results.
[DOCX File, 54 KB - jmir_v28i1e84700_app1.docx]

References

1. Christensen K, Doblhammer G, Rau R, Vaupel JW. Ageing populations: the challenges ahead. Lancet 2009 Oct 3;374(9696):1196-1208. [doi: 10.1016/S0140-6736(09)61460-4] [Medline: 19801098]
2. Beard JR, Officer A, de Carvalho IA, et al. The world report on ageing and health: a policy framework for healthy ageing. Lancet 2016 May;387(10033):2145-2154. [doi: 10.1016/S0140-6736(15)00516-4]
3. Wang HM, Beyer M, Gensichen J, Gerlach FM. Health-related quality of life among general practice patients with differing chronic diseases in Germany: cross sectional survey. BMC Public Health 2008 Jul 21;8:246. [doi: 10.1186/1471-2458-8-246] [Medline: 18638419]
4. Chen X, Giles J, Yao Y, et al. The path to healthy ageing in China: a Peking University-Lancet Commission. Lancet 2022 Dec 3;400(10367):1967-2006. [doi: 10.1016/S0140-6736(22)01546-X] [Medline: 36423650]
5. Chen C, Liu GG, Shi QL, et al. Health-related quality of life and associated factors among oldest-old in China. J Nutr Health Aging 2020;24(3):330-338. [doi: 10.1007/s12603-020-1327-2] [Medline: 32115616]
6. Li H, Tao S, Sun S, Xiao Y, Liu Y. The relationship between health literacy and health-related quality of life in Chinese older adults: a cross-sectional study. Front Public Health 2024;12:1288906. [doi: 10.3389/fpubh.2024.1288906]
7. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85. [doi: 10.2196/jmir.1933] [Medline: 23615206]
8. Hunsaker A, Hargittai E. A review of Internet use among older adults. New Media Soc 2018 Oct;20(10):3937-3954. [doi: 10.1177/1461444818787348]
9. Mitchell UA, Chebli PG, Ruggiero L, Muramatsu N. The digital divide in health-related technology use: the significance of race/ethnicity. Gerontologist 2019 Jan 9;59(1):6-14. [doi: 10.1093/geront/gny138] [Medline: 30452660]
10. Kruse C, Fohn J, Wilson N, Nunez Patlan E, Zipp S, Mileski M. Utilization barriers and medical outcomes commensurate with the use of telehealth among older adults: systematic review. JMIR Med Inform 2020 Aug 12;8(8):e20359. [doi: 10.2196/20359] [Medline: 32784177]
11. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
12. Çöme O, Gökdemir Ö, Bilik Sezer B, Kasapoğlu SS, Kjær NK, Güldal D. The interplay between cognitive function and digital health literacy among older adults: implications for e-health equity and accessibility. Int J Med Inform 2025 Sep;201:105934. [doi: 10.1016/j.ijmedinf.2025.105934] [Medline: 40273597]
13. Wu Y, Wen J, Wang X, et al. Associations between e-health literacy and chronic disease self-management in older Chinese patients with chronic non-communicable diseases: a mediation analysis. BMC Public Health 2022 Nov 29;22(1):2226. [doi: 10.1186/s12889-022-14695-4] [Medline: 36447176]
14. Ivan L, Marston HR, Prabhu VG, et al. Successful aging across middle versus high-income countries: an analysis of the role of eHealth literacy associated with loneliness and well-being. Gerontologist 2024 Dec 13;65(1):gnae170. [doi: 10.1093/geront/gnae170] [Medline: 39673788]

XSL•FO
RenderX

15. Chen A, Wang X, Xu A, Liu Q, Tong B. Analysis of the impact of e-health literacy on health behavior among the elderly in Jiangsu: structural equation model. Public Health Nurs 2025;42(4):1461-1468. [doi: 10.1111/phn.13565] [Medline: 40266126]
16. Xie L, Mo PKH. A 3-wave longitudinal study of eHealth literacy and older people's health-related quality of life in China: the mediating role of general self-efficacy. J Am Med Dir Assoc 2024 Dec;25(12):105310. [doi: 10.1016/j.jamda.2024.105310] [Medline: 39426796]
17. Stellefson M, Paige SR, Alber JM, et al. Association between health literacy, electronic health literacy, disease-specific knowledge, and health-related quality of life among adults with chronic obstructive pulmonary disease: cross-sectional study. J Med Internet Res 2019 Jun 6;21(6):e12165. [doi: 10.2196/12165] [Medline: 31172962]
18. Zhang M, Tao S, Ge X, et al. Health self-management behaviors as a bridge between electronic health literacy and health-related quality of life: cross-sectional study from China. J Med Internet Res 2025 Aug 13;27:e74056. [doi: 10.2196/74056] [Medline: 40802975]
19. Lai Y, Chen S, Li M, Ung COL, Hu H. Policy interventions, development trends, and service innovations of internet hospitals in China: documentary analysis and qualitative interview study. J Med Internet Res 2021 Jul 20;23(7):e22330. [doi: 10.2196/22330] [Medline: 34283025]
20. Jokisch MR, Schmidt LI, Doh M. Acceptance of digital health services among older adults: findings on perceived usefulness, self-efficacy, privacy concerns, ICT knowledge, and support seeking. Front Public Health 2022;10:1073756. [doi: 10.3389/fpubh.2022.1073756] [Medline: 36582385]
21. Decade of healthy ageing: baseline report. : World Health Organization; 2020 URL: https://apps.who.int/iris/handle/10665/338677 [accessed 2026-03-13]
22. Singh S, Goodwin S, Zhong S, et al. Inequalities in health-related quality of life and functional health of an aging population: a Canadian community perspective. PLoS One 2024;19(7):e0304457. [doi: 10.1371/journal.pone.0304457] [Medline: 38968188]
23. Tan G, Chao J, Jin S, et al. The association between digital health literacy and health inequalities among Chinese older adults: a multicenter cross-sectional study. Digit Health 2025;11:20552076251349901. [doi: 10.1177/20552076251349901] [Medline: 40547434]
24. van Kessel R, Wong BLH, Clemens T, Brand H. Digital health literacy as a super determinant of health: more than simply the sum of its parts. Internet Interv 2022 Mar;27:100500. [doi: 10.1016/j.invent.2022.100500] [Medline: 35242586]
25. Xie L, Zhang S, Xin M, Zhu M, Lu W, Mo PKH. Electronic health literacy and health-related outcomes among older adults: a systematic review. Prev Med 2022 Apr;157:106997. [doi: 10.1016/j.ypmed.2022.106997] [Medline: 35189203]
26. Rowe JW, Kahn RL. Successful aging. Gerontologist 1997 Aug;37(4):433-440. [doi: 10.1093/geront/37.4.433] [Medline: 9279031]
27. Levy BR, Slade MD, Kasl SV. Longitudinal benefit of positive self-perceptions of aging on functional health. J Gerontol B Psychol Sci Soc Sci 2002 Sep;57(5):409-417. [doi: 10.1093/geronb/57.5.p409] [Medline: 12198099]
28. Di X, Wang L. Impact of digital literacy on aging attitudes of the elderly from the perspective of social cognitive theory. BMC Public Health 2025 Jun 6;25(1):2113. [doi: 10.1186/s12889-025-23377-w] [Medline: 40481505]
29. Bellingtier JA, Neupert SD. Negative aging attitudes predict greater reactivity to daily stressors in older adults. J Gerontol B Psychol Sci Soc Sci 2018;73(7):gbw086. [doi: 10.1093/geronb/gbw086]
30. Levy BR, Myers LM. Preventive health behaviors influenced by self-perceptions of aging. Prev Med 2004 Sep;39(3):625-629. [doi: 10.1016/j.ypmed.2004.02.029] [Medline: 15313104]
31. Levy BR, Hausdorff JM, Hencke R, Wei JY. Reducing cardiovascular stress with positive self-stereotypes of aging. J Gerontol B Psychol Sci Soc Sci 2000 Jul;55(4):205-213. [doi: 10.1093/geronb/55.4.p205] [Medline: 11584877]
32. Paasche-Orlow MK, Wolf MS. The causal pathways linking health literacy to health outcomes. Am J Health Behav 2007 Jul 1;31(1):19-26. [doi: 10.5993/AJHB.31.s1.4]
33. Bandura A. Self-Efficacy: The Exercise of Control: W.H.Freeman; 1997.
34. Park C. Electronic health literacy as a source of self-efficacy among community-dwelling older adults. Clin Gerontol 2025 Mar 15;48(2):289-298. [doi: 10.1080/07317115.2024.2373894]
35. Bandura A. Self-efficacy: toward a unifying theory of behavioral change. Psychol Rev 1977 Mar;84(2):191-215. [doi: 10.1037//0033-295x.84.2.191] [Medline: 847061]
36. Bandura A. Health promotion by social cognitive means. Health Educ Behav 2004 Apr;31(2):143-164. [doi: 10.1177/1090198104263660] [Medline: 15090118]
37. Du S, Tian L, Tian Y, Feng Z, Wang Y. The role of self-efficacy and self-care agency as mediating factors in the link between health literacy and health-promoting lifestyle among older adults post covid 19 era: a multiple mediator model. Geriatr Nurs 2023;54:252-257. [doi: 10.1016/j.gerinurse.2023.10.013] [Medline: 37897930]
38. Arcaya MC, Arcaya AL, Subramanian SV. Inequalities in health: definitions, concepts, and theories. Glob Health Action 2015;8:27106. [doi: 10.3402/gha.v8.27106] [Medline: 26112142]
39. Jiang X, Wang L, Leng Y, et al. The level of electronic health literacy among older adults: a systematic review and meta-analysis. Arch Public Health 2024 Nov 7;82(1):204. [doi: 10.1186/s13690-024-01428-9] [Medline: 39511667]

XSL•FO
RenderX

40. Deng SY, Zhao IY, Ho MH, Saravanakumar P, Molassiotis A, Montayre J. Rural-urban disparities in healthy ageing: evidence from a national study in China. Collegian 2022 Dec;29(6):843-849. [doi: 10.1016/j.colegn.2022.06.008]
41. Pailaha AD. Public health nursing: Challenges and innovations for health literacy in rural area. Public Health Nurs 2023;40(5):769-772. [doi: 10.1111/phn.13223] [Medline: 37392096]
42. Baltes PB, Baltes MM. Psychological perspectives on successful aging: the model of selective optimization with compensation. In: Baltes PB, Baltes MM, editors. Successful Aging: Perspectives from the Behavioral Sciences European Network on Longitudinal Studies on Individual Development: Cambridge University Press; 1990:1-34. [doi: 10.1017/CBO9780511665684.003]
43. Zheng X, Xue Y, Dong F, et al. The association between health-promoting-lifestyles, and socioeconomic, family relationships, social support, health-related quality of life among older adults in China: a cross sectional study. Health Qual Life Outcomes 2022 Apr 20;20(1):64. [doi: 10.1186/s12955-022-01968-0] [Medline: 35443689]
44. Cui K, Zou W, Ji X, Zhang X. Does digital technology make people healthier: the impact of digital use on the lifestyle of Chinese older adults. BMC Geriatr 2024 Jan 22;24(1):85. [doi: 10.1186/s12877-023-04651-1] [Medline: 38254001]
45. Li S, Cui G, Yin Y, Wang S, Liu X, Chen L. Health-promoting behaviors mediate the relationship between eHealth literacy and health-related quality of life among Chinese older adults: a cross-sectional study. Qual Life Res 2021 Aug;30(8):2235-2243. [doi: 10.1007/s11136-021-02797-2] [Medline: 33661455]
46. Lin CY, Ganji M, Griffiths MD, Bravell ME, Broström A, Pakpour AH. Mediated effects of insomnia, psychological distress and medication adherence in the association of eHealth literacy and cardiac events among Iranian older patients with heart failure: a longitudinal study. Eur J Cardiovasc Nurs 2020 Feb;19(2):155-164. [doi: 10.1177/1474515119873648] [Medline: 31516036]
47. Leung AYM, Parial LL, Tolabing MC, et al. Sense of coherence mediates the relationship between digital health literacy and anxiety about the future in aging population during the COVID-19 pandemic: a path analysis. Aging Ment Health 2022 Mar;26(3):544-553. [doi: 10.1080/13607863.2020.1870206] [Medline: 33438448]
48. Choi M. Association of eHealth use, literacy, informational social support, and health-promoting behaviors: mediation of health self-efficacy. Int J Environ Res Public Health 2020 Oct 28;17(21):7890. [doi: 10.3390/ijerph17217890] [Medline: 33126469]
49. Neter E, Brainin E. Association between health literacy, eHealth literacy, and health outcomes among patients with long-term conditions. Eur Psychol 2019 Jan;24(1):68-81. [doi: 10.1027/1016-9040/a000350]
50. The Whoqol Group. The World Health Organization quality of life assessment (WHOQOL): development and general psychometric properties. Soc Sci Med 1998 Jun;46(12):1569-1585. [doi: 10.1016/S0277-9536(98)00009-4]
51. Cheng LJ, Tan RLY, Luo N. Measurement properties of the EQ VAS around the globe: a systematic review and meta-regression analysis. Value Health 2021 Aug;24(8):1223-1233. [doi: 10.1016/j.jval.2021.02.003] [Medline: 34372988]
52. Hu W, Zhou L, Chu J, et al. Estimating population norms for the health-related quality of life of adults in southern Jiangsu Province, China. Sci Rep 2022 Jun 14;12(1):9906. [doi: 10.1038/s41598-022-13910-x] [Medline: 35701516]
53. The World Health Organization-five well-being index (WHO-5). World Health Organization. 2024. URL: https://www.who.int/zh/publications/m/item/WHO-UCN-MSD-MHE-2024.01 [accessed 2026-03-13]
54. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167-176. [doi: 10.1159/000376585] [Medline: 25831962]
55. Diener E, Emmons RA, Larsen RJ, Griffin S. The Satisfaction With Life Scale. J Pers Assess 1985 Feb;49(1):71-75. [doi: 10.1207/s15327752jpa4901_13] [Medline: 16367493]
56. Whisman MA, Judd CM. A cross-national analysis of measurement invariance of the Satisfaction With Life Scale. Psychol Assess 2016 Feb;28(2):239-244. [doi: 10.1037/pas0000181] [Medline: 26168309]
57. Norman CD, Skinner HA. eHEALS: the eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27. [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
58. Oh SS, Kim KA, Kim M, Oh J, Chu SH, Choi J. Measurement of digital literacy among older adults: systematic review. J Med Internet Res 2021 Feb 3;23(2):e26145. [doi: 10.2196/26145] [Medline: 33533727]
59. Neupert SD, Bellingtier JA. Aging attitudes and daily awareness of age-related change interact to predict negative affect. Gerontologist 2017 Aug 1;57(suppl_2):S187-S192. [doi: 10.1093/geront/gnx055] [Medline: 28854606]
60. Scholz U, Doña BG, Sud S, et al. Is general self-efficacy a universal construct? Psychometric findings from 25 countries. Eur J Psychol Assess 2002;18(3):242-251. [doi: 10.1027/1015-5759.18.3.242]
61. Liu S, Lu Y, Wang D, et al. Impact of digital health literacy on health-related quality of life in Chinese community-dwelling older adults: the mediating effect of health-promoting lifestyle. Front Public Health 2023;11:1200722. [doi: 10.3389/fpubh.2023.1200722]
62. Chai Y, Xian G, Wang M, Guo L, Luo S. Aging wisely: the impact of Internet use on older adults' mental health. J Affect Disord 2024 Nov 1;364:139-145. [doi: 10.1016/j.jad.2024.08.076] [Medline: 39147146]
63. Preacher KJ, Hayes AF. SPSS and SAS procedures for estimating indirect effects in simple mediation models. Behav Res Methods Instrum Comput 2004 Nov;36(4):717-731. [doi: 10.3758/bf03206553] [Medline: 15641418]

XSL•FO
RenderX

64. Podsakoff PM, MacKenzie SB, Lee JY, Podsakoff NP. Common method biases in behavioral research: a critical review of the literature and recommended remedies. J Appl Psychol 2003 Oct;88(5):879-903. [doi: 10.1037/0021-9010.88.5.879] [Medline: 14516251]
65. Hayes AF. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach: The Guilford Press; 2013.
66. Panagioti M, Skevington SM, Hann M, et al. Effect of health literacy on the quality of life of older patients with long-term conditions: a large cohort study in UK general practice. Qual Life Res 2018 May;27(5):1257-1268. [doi: 10.1007/s11136-017-1775-2] [Medline: 29322478]
67. Lee EH, Lee YW, Kang EH, Kang HJ. Relationship between electronic health literacy and self-management in people with type 2 diabetes using a structural equation modeling approach. J Nurs Res 2024 Jan 1;32(1):e315. [doi: 10.1097/jnr.0000000000000588] [Medline: 38128065]
68. Huhta AM, Hirvonen N, Huotari ML. Health literacy in web-based health information environments: systematic review of concepts, definitions, and operationalization for measurement. J Med Internet Res 2018 Dec 19;20(12):e10273. [doi: 10.2196/10273] [Medline: 30567690]
69. McMullan RD, Berle D, Arnáez S, Starcevic V. The relationships between health anxiety, online health information seeking, and cyberchondria: systematic review and meta-analysis. J Affect Disord 2019 Feb 15;245:270-278. [doi: 10.1016/j.jad.2018.11.037] [Medline: 30419526]
70. Du X, Witthöft M, Zhang T, Shi C, Ren Z. Interpretation bias in health anxiety: a systematic review and meta-analysis. Psychol Med 2023 Jan;53(1):34-45. [doi: 10.1017/S0033291722003427] [Medline: 36349699]
71. Westerhof GJ, Miche M, Brothers AF, et al. The influence of subjective aging on health and longevity: a meta-analysis of longitudinal data. Psychol Aging 2014 Dec;29(4):793-802. [doi: 10.1037/a0038016] [Medline: 25365689]
72. Seeman TE, Unger JB, McAvay G, Mendes de Leon CF. Self-efficacy beliefs and perceived declines in functional ability: MacArthur studies of successful aging. J Gerontol B Psychol Sci Soc Sci 1999 Jul;54(4):214-222. [doi: 10.1093/geronb/54b.4.p214] [Medline: 12382590]
73. Li JB, Dou K, Liang Y. The relationship between presence of meaning, search for meaning, and subjective well-being: a three-level meta-analysis based on the meaning in life questionnaire. J Happiness Stud 2021 Jan;22(1):467-489. [doi: 10.1007/s10902-020-00230-y]
74. Chen X, Wang N. How does digital literacy affect the health status of senior citizens? Micro-level evidence from the CFPS data. BMC Health Serv Res 2025 Jan 27;25(1):153. [doi: 10.1186/s12913-025-12314-7] [Medline: 39871247]
75. Jing R, Jin G, Guo Y, Zhang Y, Li L. The association between constant and new Internet use and depressive symptoms among older adults in China: the role of structural social capital. Comput Human Behav 2023 Jan;138:107480. [doi: 10.1016/j.chb.2022.107480]
76. Chatterjee A, Prinz A, Gerdes M, Martinez S. Digital interventions on healthy lifestyle management: systematic review. J Med Internet Res 2021 Nov 17;23(11):e26931. [doi: 10.2196/26931] [Medline: 34787575]

Abbreviations

ATOA: attitudes toward own aging
CIOMS: Council for International Organizations of Medical Sciences
eHEALS: eHealth Literacy Scale
EHL: electronic health literacy
EQ-VAS: EuroQol Visual Analogue Scale
FDR: false discovery rate
GSES: General Self-Efficacy Scale
HRQoL: health-related quality of life
KMO: Kaiser-Meyer-Olkin
LS: life satisfaction
MH: mental health
PH: physical health
SE: self-efficacy
SWLS: Satisfaction With Life Scale
WHO-5: World Health Organization-Five Well-Being Index

XSL•FO
RenderX

Edited by A Schwartz, M Balcarras; submitted 14.Oct.2025; peer-reviewed by V Dolničar, Y Qian; accepted 05.Feb.2026; published 24.Mar.2026.

Please cite as:
Wang Y, Dai B, Yao J
The Relationship Between Electronic Health Literacy and Health-Related Quality of Life Among Chinese Older Adults: Cross-Sectional Study
J Med Internet Res 2026;28:e84700
URL: https://www.jmir.org/2026/1/e84700
doi:10.2196/84700

© Yongqiang Wang, Baozhen Dai, Jiazhen Yao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Validation of the Updated Digital Health Literacy Instrument and Development of a Short Form: Online Survey Study of the General Population

Rosalie van der Vaart[1], PhD; Inge Tuitert[2], PhD; Job van 't Veer[2], PhD; Constance Drossaert[3], PhD

[1]Centre of Expertise Health Innovation, Technology for Health Care Research Group, The Hague University of Applied Sciences, Johanna Westerdijkplein 75, Den Haag, The Netherlands

[2]Department of Health and Social Studies, NHL Stenden University of Applied Sciences, Leeuwarden, The Netherlands

[3]Department of Psychology, Health and Technology, Faculty of Behavioural, Management and Social Sciences, University of Twente, Enschede, The Netherlands

Corresponding Author:
Rosalie van der Vaart, PhD
Centre of Expertise Health Innovation, Technology for Health Care Research Group, The Hague University of Applied Sciences, Johanna Westerdijkplein 75, Den Haag, The Netherlands

Abstract

Background: The digital health literacy instrument (DHLI) was developed in 2017 to measure individuals' ability to access, understand, evaluate, and apply online health information. Since that time, digital health has shifted from desktop-based internet use to mobile devices, and there has been a rapidly expanding range of health apps. Additionally, heightened privacy and data security requirements have increased the complexity of user competencies needed to engage with digital health tools. These developments underscore the need to update the original DHLI.

Objective: This study aimed to create an updated version of the DHLI (DHLI 2.0) that reflects current digital health practices and to examine its reliability and validity by exploring associations with user characteristics. Additionally, we aimed to develop a short-form version to facilitate broader use in research and practice.

Methods: The instrument was iteratively updated and pilot-tested to retain the original theoretical framework while reflecting current digital health practices, devices, and emerging challenges such as mobile use and data security. Several items were reworded and a new 2-item subscale on digital safety was added. The full DHLI 2.0 comprises 24 items across 8 skill domains. A 16-item short form was developed by iteratively removing 1 or 2 items per subscale based on the "α if item deleted" criterion, while retaining the same subscale structure as the full form. Data to validate the new version of the instrument were collected in June 2024 through an online survey among members of a representative citizen panel in Friesland, a province in the Netherlands (N=2728). Sociodemographics, internet and health-related internet use, general health literacy (measured with the Single Item Literacy Screener), self-reported health, and health care use were assessed. Internal consistency was evaluated using Cronbach α, and construct validity was assessed via Spearman ρ correlations with related constructs.

Results: Internal consistency was high for both the full (α=0.94) and short-form (α=0.90) scales. Most subscales showed satisfactory to excellent reliability (α=0.71–0.93), while "Securing privacy" and "Using security measures" demonstrated moderate reliability (α=0.65-0.66). The DHLI 2.0 total scores were approximately normally distributed (skewness –0.5; kurtosis 0.4). As expected, digital health literacy was negatively correlated with age (ρ=−0.39, *P*<.001) and positively correlated with education (ρ=0.22, *P*<.001), income (ρ=0.27, *P*<.001), time spent online (ρ=0.32, *P*<.001), and general health literacy (ρ=−0.42, *P*<.001).

Conclusions: The DHLI 2.0 provides an updated, reliable, and valid measure of digital health literacy covering 8 key domains, including data security. The 16-item short form offers a concise alternative suitable for research and possibly practical applications in health and eHealth contexts.

(J Med Internet Res 2026;28:e86879) doi:10.2196/86879

KEYWORDS

digital health literacy skills; eHealth; measurement; validation; instrument; health information

Introduction

Over the past decades, the landscape of health care has transformed considerably due to the rise of the internet and digital health technology. Citizens and patients now have access to a broad spectrum of digital health resources, allowing them to search online for health-related information, communicate with care professionals through patient portals, connect with

XSL•FO
RenderX

peers via social media platforms, and apply digital tools and applications to monitor or manage their health.

Although these digital health innovations hold great potential to improve accessibility and quality of care and to improve patient involvement in care, not everyone benefits equally from them. For example, data from the Dutch National Monitor Digital Care show stark differences in use of digital solutions among care users, where almost all care users with a theoretical educational background report using a computer independently or with help, compared to about half of those with a practice-oriented educational background [1,2]. These disparities also extend to the use of digital health tools, where these numbers are 68% and 22%, respectively, for using websites to seek health information. In addition, age-related differences are also evident: while 75% of those aged 15 to 39 years reported having used health websites, this is only 36% in the age group of ≥65 years [1,2]. Similar trends have frequently been observed in international research contexts [3-5]. Effectively engaging with digital health resources requires a set of complex skills that are not evenly distributed across the population. People may struggle to access, navigate, understand, evaluate, or apply online health information, or to use digital health tools efficiently.

These findings highlight the importance of digital health literacy, defined as the ability to seek, find, understand, appraise, and apply online health information and services [6]. As health information and care services are increasingly delivered through digital channels, gaining insight into individuals' digital health literacy is essential. It allows researchers to assess the level of digital health literacy at a population level and provides insight into differences in digital health literacy skills among subgroups. In addition, it supports researchers and developers of health applications to assess whether innovations fit their target group. For health care professionals, knowledge about their patients' digital health literacy could even be useful to assess whether support is needed, for whom, and in which form. To assess digital health literacy, the digital health literacy instrument (DHLI) was developed in 2017 [7]. The DHLI is grounded in theory and empirical observational research and encompasses 7 subscales. Each subscale captures a distinct component of digital health literacy, and together they represent the practical skills needed to access, understand, evaluate, and apply digital health information: operational skills (the skills to operate a device, such as typing, clicking, and swiping), navigation skills, information searching, evaluating reliability, determining relevance, creating self-generated content (being able to express oneself clearly in written language), and protecting privacy while sharing information with others.

Since its introduction, the DHLI has been widely adopted and validated across national and international contexts, including studies among diverse patient populations [8,9], adolescents [10,11], older adults [12], university students [13-15], and health care professionals [16], demonstrating its value as a comprehensive instrument for assessing digital health competencies. Yet, the digital health landscape has evolved considerably since the publication of the DHLI. First, there has been a shift from desktop computers and laptops to mobile devices, such as smartphones and tablets, which require different operational skills (eg, tapping and swiping rather than clicking or typing). Second, the field has seen a rapid increase in digital health tools (eg, apps for self-management, self-monitoring, and communication), making some of the original DHLI items, which primarily focus on internet use, less suitable. Third, growing concerns around privacy and data safety have led to more complex authentication procedures, such as 2-factor identification and the need to manage secure passwords, which demand additional user competencies [17].

In response to these changes, we felt the need for a revision of the DHLI. Therefore, the first aim of this study was to validate an updated version of the DHLI, the DHLI 2.0, including its internal consistency and construct validity, by examining the associations with a range of user characteristics. Additionally, we aimed to examine the feasibility and validity of a short-form version of the instrument and to facilitate broader applications in research and practice.

Methods

Instrument Revision

The revised version of the DHLI, the DHLI 2.0, includes 8 skill domains, each measured with 2 to 4 items, resulting in 24 items in total. The goal of the revision was to keep the original structure and theoretical framework of the DHLI intact, while updating the instrument to current digital health practices and currently used devices. The revision was an iterative process, with some experts, in which we reworded the items to make them more in line with the vocabulary of the current digital landscape. Subsequently, a pilot study was conducted with a preliminary version. The prefinal version of the new instrument was tested in a qualitative pilot study, with 8 survey experts of Planbureau Fryslân, which informed the final adjustments. As main changes, we added the use of tablets and smartphones in the introduction of the items and included words as tapping and swiping (in addition to clicking). A new subscale, including 2 items, was added to address data security challenges based on problems encountered in recent projects. The items include skills to use 2-step authentication and skills in using strong passwords.

Similar to the original version, all items are scored on a 4-point Likert scale, with response options ranging from "very easy" to "very difficult" or from "never" to "often." For the current validation study, we added an option "not applicable/do not know," since this was common practice in the surveys of Planbureau Fryslân. For analysis purposes, scores are reverse-coded so that higher values reflect higher levels of digital health literacy. A total digital health literacy score was computed by averaging responses across all items, provided that at least 16 items have been completed in the full scale and 8 items in the short form (SF). Subscale scores were calculated using the mean of the items within each skill domain, provided that all items of the subscale were completed. Multimedia Appendices 1-3 provide (1) the items and scoring instructions for the revised DHLI, full form (FF DHLI 2.0); (2) the SF DHLI 2.0; and (3) an overview of the changes made in DHLI 2.0 compared to the original DHLI.

XSL•FO
RenderX

Study Design and Participants

For this study, the citizen panel from Planbureau Fryslân was used. Planbureau Fryslân is the independent regional institute for research and monitoring, aiming to inform and enhance local government policy in the province of Friesland, the Netherlands. Their panel consists of a sample of approximately 8000 residents in Friesland (a Northern province of the Netherlands). To ensure representativeness, Planbureau Fryslân regularly draws a random sample of all residents aged 18 years and older in the province, in close coordination with the municipalities. Potential panel members are personally invited by letter to participate in the sample. Self-enrollment in the panel is not possible. Panel members are periodically asked to participate in surveys on topics relevant to regional policy, including health. The data collection for Panel Fryslân follows a standardized procedure as described in their research accountability document [18]. For this study, in June 2024, all active panel members were invited to complete a survey on digital health care use and digital health literacy. The questionnaire included the DHLI 2.0 and 5 other digital health-related questions (see Measures). Panel members could respond to the survey during a 3-week period and received 2 reminders within these 3 weeks. Three months earlier, in January 2024, the panel had filled out a survey on general health, including questions on self-reported general health, health care use, and chronic health conditions (see *Measures*). These variables were merged with the data from June 2024 for the current analyses. In total, 2728 respondents filled out both the survey from January and June 2024. The data of these respondents are used in the analyses.

Procedure

Data collection by Planbureau Fryslân was carried out via an online Dutch-language survey (Quest Software, Inc), which was distributed via email to panel members. Panel members were given 1 month to respond to the survey and received 2 reminder emails during that period to enhance the response rate.

Measures

In addition to the DHLI 2.0, the survey items addressed the following characteristics of the participants: (1) sociodemographics (gender, age, educational attainment, and income); (2) internet use (average number of hours spent online per day); and (3) health-related internet use (the use of 11 different online health activities in the past 12 months) (see Table 1 for a complete overview of all activities); with these 11 items, a sum score was calculated (ranging from 0 to 11), where using 1 specific type of online health activities added 1 point to the sum score; (4) general health literacy (assessed using the Single Item Literacy Screener [SILS]) [19]; (5) self-reported general health on a scale from worst possible health (1) to the best possible health (10); (6) health care use (assessed by number of visits to a general practitioner in the past 12 months, visits to a medical specialist in the past 12 months, and visits to a mental health professional in the past 12 months); and (7) presence of chronic conditions.

XSL•FO
RenderX

Table . Internet use and health-related internet use (N=2728).

Variable	Value
Hours spent online per day, n (%)	
<1	371 (13.6)
1 - 3	1346 (49.3)
3 - 5	551 (20.2)
5 - 8	237 (8.7)
>8	148 (5.4)
Not applicable/don’t know	35 (1.3)
Missing	40 (1.5)
Self-rated internet skills, n (%)	
Very good	624 (22.9)
Good	1186 (43.5)
Average	536 (19.6)
Less than average	272 (10)
Poor	50 (1.8)
Not applicable/don’t know	20 (0.7)
Missing	40 (1.5)
SILS[a], n (%)	
Never	1692 (62)
Rarely	739 (27.1)
Sometimes	218 (8)
Often	29 (1.1)
Always	3 (0.1)
Not applicable/don’t know	7 (0.3)
Missing	40 (1.5)
Number of respondents who have used the internet for health-related reasons	
Online information, n (%)	2353 (86.2)
Accessing a patient portal[b], n (%)	1729 (63.4)
Online prescriptions, n (%)	1220 (44.7)
Accessing a personal health portal[c], n (%)	1175 (43.1)
Wearable, n (%)	1027 (37.6)
Online appointment scheduling, n (%)	855 (31.3)
Information on a health app, n (%)	824 (30.2)
Video consultation, n (%)	428 (15.7)
Information on social media, n (%)	423 (15.5)
Healthy lifestyle app, n (%)	367 (13.5)
Communicating via a patient forum, n (%)	170 (6.2)
Health-related internet use, mean (SD)	3.93 (2.24)

[a]SILS: Single Item Literacy Screener.

[b]Data from a care provider shared via a secured portal.

[c]Health data from care providers combined on a personal secured portal.

XSL•FO
RenderX

Ethical Considerations

All participants provided informed consent upon joining the panel and agreed to the use and sharing of their anonymized data for research purposes. The panel adheres to the code of conduct established by the Dutch Association for Statistics and Research (Vereniging voor Statistiek en Onderzoek, VSO) [20]. This study was reviewed by the Ethical Board of NHL Stenden University of Applied Sciences, who determined that the study is indeed in line with this conduct (reference number Tuitert 202405). Prior to being shared with the authors of this paper, the data were fully anonymized. The authors had no direct contact with the participants, and no financial compensation was provided for participation in the panel.

Data Analyses

Data were analyzed using IBM SPSS version 29.0 for Windows (IBM Corp). Descriptive statistics were computed for all relevant variables, including frequencies, percentages, means, and SDs. Missing values and the option "don't know/don't want to tell" are in most cases reported in the descriptive statistics but were treated as missing values in further analyses.

Cronbach α served as a measure of internal consistency, reflecting the (weighted) average correlation of items within the scale [21]. In general, a Cronbach α of 0.7 to 0.8 is regarded as satisfactory for scales to be used as research tools [22]. The distributional characteristics of the DHLI and its subscales were examined to assess normality and detect floor and ceiling effects. Skewness and kurtosis values were used to evaluate score distributions, with values between +1 and –1 interpreted as indicating no or slight nonnormality [23]. Floor or ceiling effects were considered present when more than 15% of participants achieved either the lowest or highest possible score on a subscale [24]. Evidence for construct validity was determined by studying Spearman ρ correlations between total scores on the DHLI 2.0 and sociodemographics, (health-related) internet use, self-reported general health, health care use, and the SILS [19]. We expected the correlations with the other variables to be in line with the correlations of the original DHLI [7]. Additionally, we explored the reliability and validity of an SF DHLI 2.0. By using the "α if item is deleted" option, for each subscale 1 or 2 items were deleted. These short-form subscales were also taken along in the validity analyses.

Results

Participants

Table 2 shows the characteristics of the sample. Men (1582/2728, 58%) and respondents with a theory-based educational background (1445/2728, 53%) were overrepresented, that is, people with tertiary education. The mean age of the sample was 60.1 (SD 16.8) years, with a range between 19 and 95 years.

Table 3 presents the sample's health-related characteristics. A considerable number of participants (1034/2728, 37.9%) reported having at least 1 chronic condition, and about half (1286/2728, 47.1%) had had contact with a medical specialist in the past 12 months. Still, participants scored their own health as rather good on average (7.5 on a scale from 1 to 10).

XSL•FO
RenderX

Table . Sociodemographics (N=2728).

Characteristic	Value
Gender, n (%)	
Male	1582 (58)
Female	1135 (41.6)
Other	11 (0.4)
Age (y), mean (SD)	60.1 (16.8)
Educational attainment[a], n (%)	
Low	496 (18.2)
Middle (practice-based/vocational)	718 (26.3)
High (theory-based)	1445 (53)
Missing	69 (2.5)
Family income yearly (gross)[b]	
Less than €15,000	63 (2.3)
€15,000– €31,000	308 (11.3)
€31,000–€38,500	282 (10.3)
€38,500–€46,000	363 (13.3)
€46,000–€77,000	643 (23.6)
€77,000–€92,000	235 (8.6)
€92,000 or more	256 (9.4)
Not applicable/don’t know	514 (18.8)
Missing	64 (2.3)
Personal income monthly (net)	
€1000 or less – €2000	685 (25.2)
€2000–€3000	826 (30.3)
€3000–€4000	547 (20.1)
€4000–€5000	161 (5.9)
€5000 or higher	136 (4)
Not applicable/don’t know	353 (12.9)
Missing	20 (0.7)

[a]Low: less than primary, primary, and lower secondary education (European qualifications framework 1 (EQF1) and part of EQF2); medium: upper secondary and postsecondary nontertiary education (part of EQF2, EQF3, EQF4, EQF4+); high: tertiary education (EQF5, EQF6, EQF7, EQF8) [25].

[b]€1=US $1.15.

Table . Health and health care use.

Variable	Value
Self-reported general health[a] (N=2625), mean (SD)	7.5 (1.3)
Health care use (N=2728), n yes (%)	
General practitioner	1876 (68.8)
Medical specialist	1286 (47.1)
Mental health care	198 (7.3)
Chronic condition (N=2728), n yes (%)	1034 (37.9)

[a]On a scale from 1 to 10.

XSL•FO
RenderX

Table 1 provides an overview of participants' general and health-related internet use. Nearly half of the respondents reported spending approximately 1 to 3 hours online per day. Most participants (1810/2728, 66.4%) rated their own internet skills as (very) good. A large majority had used the internet to search for health-related information (2352/2728, 86.2%), and the use of online portals and digital prescriptions was also relatively high. In contrast, only a small proportion of participants reported using the internet to have a video consultation with their health care provider (428/2728, 15.7%) or to communicate online with fellow patients through a forum (170/2728, 6.2%).

Reliability and Distributional Properties of the Revised DHLI

Table 4 shows the scores and internal consistency of the DHLI 2.0 items and subscales. Both the FF (24 items) and the SF (16 items) showed strong reliability, with Cronbach α coefficients of 0.94 and 0.90, respectively. Only the protecting privacy full subscale and the using security measures subscale had moderate Cronbach α scores (0.65 and 0.66, respectively).

Table . Descriptive statistics of digital health literacy instrument (DHLI) 2.0 scales and subscales.

Scale score	N[a]	Value	Cronbach α
Total scale score, mean (SD)			
FF DHLI 2.0[b] (24 items)	2543	3.17 (0.44)	0.94[c]
SF DHLI 2.0[d] (16 items)	2728	3.11 (0.48)	0.90[e]
Operational skills, mean[f] (SD)			
FF 4 items	2728	3.56 (0.51)	0.86
SF 2 items	2728	3.46 (0.64)	0.83
Navigation skills, mean (SD)			
FF 3 items	2436	3.24 (0.54)	0.77
SF 2 items	2436	3.34 (0.58)	0.71
Information searching, mean (SD)			
FF 3 items	2641	3.15 (0.57)	0.89
SF 2 items	2641	3.08 (0.61)	0.86
Evaluating reliability, mean (SD)			
FF 3 items	2520	2.80 (0.66)	0.84
SF 2 items	2520	2.66 (0.73)	0.83
Determining relevance, mean (SD)			
FF 3 items	2453	2.93 (0.61)	0.89
SF 2 items	2453	2.95 (0.63)	0.86
Generating content, mean (SD)			
FF 3 items	1908	3.07 (0.64)	0.93
SF 2 items	1908	3.11 (0.65)	0.91
Protecting privacy, mean (SD)			
FF 3 items	1387	3.52 (0.50)	0.65
SF 2 items	1387	3.68 (0.51)	0.75
Using security measures, mean (SD)			
FF 2 items	2626	3.06 (0.69)	0.66

[a]Number of respondents per (sub)scale, after excluding missing and "not applicable/do not know" responses.

[b]FF: full form.

[c]n=1055; number of respondents that completed at least 16 items, required to compute a total score for the FF.

[d]SF: short form.

[e]n=1030; number of respondents that completed at least 8 items, required to compute a total scale score for the SF.

[f]Possible range between 1 and 4.

XSL•FO
RenderX

Respondents had total mean scores of 3.17 (SD 0.44) and 3.11 (SD 0.48) for the FF and the SF, respectively. Both the FF (skewness = –0.5; kurtosis =0.4) and the SF (skewness = –0.5; kurtosis =0.50) show low skewness and kurtosis values, indicating that their distributions are approximately normal.

The highest scores on the subscales were reported for operational skills (mean 3.6, SD 0.5 and SF mean 3.5, SD 0.6), navigation skills (mean 3.2, SD 0.5 and SF mean 3.3, SD 0.6), and protecting privacy (mean 3.52, SD 0.50 and SF mean 3.68, SD 0.51). This last subscale is skewed for both the full subscale (−1.5) and the short subscale (−2.1), with a ceiling effect where 32.1% (445/1387) scored 4 in the full subscale and 61.9% (858/1387) in the short version, and showed kurtosis in the full subscale (3.8) and in the SF (5.7). Only in the short subscale, operational skills and navigation skills exhibited a skewness of –1, with a ceiling effect, where for operational skills, 49.3% (1345/2728) scored 4 and for navigation skills 26.9% (655/2436). Participants indicated having more difficulties with evaluating reliability (mean 2.8, SD 0.7 and SF mean 2.7, SD 0.7) and with determining relevance of information (mean 2.9, SD 0.6 and SF mean 3.0, SD 0.6). The scores on the new subscale, using security measures, were relatively high for the question regarding 2-factor authentication (mean 3.28, SD 0.78) and low for the question regarding creating and remembering strong passwords (mean 2.83, SD 0.82).

Construct Validity of the Revised DHLI

Table 5 presents the Spearman ρ correlations between the DHLI 2.0 scale scores, the total scores of both the full and short versions of the DHLI 2.0, and other assessed variables. Looking at sociodemographics, age was moderately negatively correlated with scores on the DHLI 2.0, indicating that older individuals tend to have lower digital health literacy levels. This association was strongest for operational skills and navigation skills. Educational attainment showed a weak positive correlation with the DHLI 2.0, with the highest scores found for generating content, suggesting that educational attainment is somewhat related to digital health literacy skills. Regarding income, family gross income showed a weak positive correlation with the DHLI 2.0, suggesting that a higher income is slightly related to higher digital health literacy skills. Concerning internet use, daily time spent online showed a moderate positive correlation with the DHLI 2.0, suggesting that individuals who spend more time online tend to have higher digital health literacy. Again, the strongest correlations were observed for operational skills. Health-related internet use only weakly correlated with the DHLI 2.0. The DHLI 2.0 showed a moderate positive correlation with SILS, indicating conceptual overlap between the constructs measured by these instruments.

XSL•FO
RenderX

Table . Spearman ρ correlations between the revised digital health literacy instrument, sociodemographics, (health-related) internet use, health-related variables, and the Single Item Literacy Screener (SILS).

(Sub)scale	Age	Education	Family income	Personal income	Hours spent online	Health-related internet use	General health	Use general practitioner	Use medical specialist	Use mental health care	SILS
Full form (FF) and short form (SF)											
FF 24 items ρ	−0.39	0.22	0.27	0.09	0.32	0.07	0.15	0.09	0.06	−0.02	−0.42
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.002	.27	<.001
SF 16 items ρ	−0.39	0.22	0.26	0.10	0.32	0.12	0.14	0.08	0.05	−0.02	−0.40
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.006	.23	<.001
Operational skills											
FF 4 items ρ	−0.46	0.26	0.28	0.09	0.36	0.14	0.12	0.06	0.07	−0.04	−0.35
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.004	<.001	.02	<.001
SF 2 items ρ	−0.47	0.24	0.28	0.09	0.37	0.16	0.11	0.06	0.06	−0.05	−0.34
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.001	.004	.006	<.001
Navigation skills											
FF 3 items ρ	−0.35	0.15	0.19	0.05	0.26	0.007	0.1	0.1	0.07	−0.04	−0.34
P value	<.001	<.001	<.001	<.001	<.001	.72	<.001	<.001	<.001	.03	<.001
SF 2 items ρ	−0.35	0.15	0.20	0.06	0.26	0.02	0.10	0.09	0.07	−0.03	−0.35
P value	<.001	<.001	<.001	.008	<.001	.40	<.001	<.001	<.001	.09	<.001
Information searching											
FF 3 items ρ	−0.26	0.14	0.21	0.09	0.23	0.08	0.13	0.08	0.04	−0.01	−0.31
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.03	.78	<.001
SF 2 items ρ	−0.24	0.12	0.19	0.10	0.22	0.07	0.13	0.07	0.04	−0.00	−0.28
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.001	.03	.87	<.001
Evaluating reliability											
FF 3 items ρ	−0.31	0.15	0.17	0.06	0.25	0.06	0.10	0.06	0.05	−0.03	−0.33
P value	<.001	<.001	<.001	.005	<.001	.003	<.001	.001	.008	.19	<.001
SF 2 items ρ	−0.26	0.13	0.16	0.07	0.21	0.05	0.10	0.05	0.05	−0.02	−0.30
P value	<.001	<.001	<.001	.002	<.001	.02	<.001	.006	.02	.42	<.001
Determining relevance											
FF 3 items ρ	−0.26	0.15	0.17	0.06	0.22	0.08	0.12	0.06	0.01	−0.01	−0.28
P value	<.001	<.001	<.001	.005	<.001	<.001	<.001	.002	.56	.72	<.001
SF 2 items ρ	−0.27	0.14	0.17	0.06	0.22	0.07	0.11	0.06	0.01	−0.00	−0.27
P value	<.001	<.001	<.001	.006	<.001	.001	<.001	.006	.53	.86	<.001
Generating content											

XSL•FO
RenderX

(Sub)scale	Age	Education	Family income	Personal income	Hours spent online	Health-related internet use	General health	Use general practitioner	Use medical specialist	Use mental health care	SILS
FF 3 items ρ	−0.13	0.28	0.20	0.09	0.12	0.05	0.13	0.01	−0.01	0.01	−0.34
P value	<.001	<.001	<.001	<.001	<.001	.02	<.001	.81	.64	.59	<.001
SF 2 items ρ	−0.15	0.28	0.20	0.09	0.13	0.07	0.12	−0.01	−0.02	−0.01	−0.34
P value	<.001	<.001	<.001	.001	<.001	.008	<.001	.81	.30	.74	<.001
Protecting privacy											
FF 3 items ρ	−0.22	0.05	0.05	−0.006	0.13	−0.11	0.05	0.08	0.09	−0.02	−0.28
P value	<.001	.09	.09	.83	<.001	<.001	.07	.005	<.001	.41	<.001
SF 2 items ρ	−0.12	0.04	0.06	−0.003	0.06	−0.09	0.03	0.03	0.09	−0.02	−0.20
P value	<.001	.11	.03	.91	.03	.001	.30	.20	.002	.52	<.001
Using security measures											
FF and SF 2 items ρ	−0.26	0.19	0.21	0.12	0.24	0.12	0.11	0.06	0.04	−0.01	−0.33
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.003	.07	.55	<.001

Discussion

This paper describes the validation of a revised version of the DHLI, the DHLI 2.0. The original instrument was published in 2017, and 8 years later, it has been updated to better reflect the rapidly evolving digital landscape. Several items were reworded to capture the widespread use of smartphones and tablets. Also, a new 2-item subscale was added to address digital safety, specifically the challenges of choosing and using secure passwords and applying two-step authentication when accessing personalized systems. In our view, the DHLI 2.0 offers a more accurate representation of the competencies required for digital health literacy today.

DHLI 2.0 demonstrated strong reliability, with Cronbach α values that were compared to, and in some cases slightly higher than, those of the original instrument. Construct validity was also adequate, with correlations to sociodemographics, internet use, and general health literacy as expected and consistent with those observed for the original DHLI [7] and in other recent systematic reviews focusing on these relationships [26,27]. A recent review by Wang et al [28] identified DHLIs developed over the past 2 decades, each addressing different aspects of digital health literacy, including the DHLI. Compared to the other included instruments, the DHLI and DHLI 2.0 offer a concise yet theoretically well-grounded instrument [7]. It captures a broad spectrum of skills, ranging from basic operational abilities to higher-order information-processing skills, while also addressing the increasingly important domain of privacy and data security. Moreover, to minimize overestimation bias, common for self-reported skills, items are phrased to assess how easy or difficult respondents find certain tasks or how frequently they encounter difficulties. This validation study also included an SF DHLI 2.0, consisting of 16 items instead of 24. The SF demonstrated good reliability and internal consistency. It facilitates easier administration and allows for integration into larger sets of survey instruments.

In revising the DHLI, we deliberately chose to retain only self-report items and exclude the performance-based tasks that were a unique feature of the original version [7]. Although innovative, those tasks turned out to be too complex to administer and provided only limited added value, mostly in terms of face validity. Furthermore, their applicability was highly dependent on the device, browser, and brand used by respondents (eg, desktop vs mobile), making it nearly impossible to develop a universally valid instrument for use in anonymous research settings. Similar challenges are reported by Crocker et al [29] in their systematic scoping review of performance-based eHealth literacy measures, who also highlight the time-consuming nature of these assessments for both participants and researchers, as well as their equipment requirements. Nevertheless, self-report measures have well-known limitations, and future research could explore the integration of performance-based items into the revised DHLI, also for screening in practice settings.

Some limitations of this study need to be acknowledged. First, validation was conducted through an online questionnaire via a citizen panel. Although our sample was large and broadly representative, the online format likely excluded individuals with the lowest digital skills and contributed to the overrepresentation of higher-educated respondents. The response via the citizen panel from Planbureau Fryslân might have contributed to the overrepresentation of older adults. This may limit the generalizability of the findings, particularly for less

XSL•FO
RenderX

educated or younger and middle-aged populations. Validation of the instrument among these currently underrepresented populations is needed in future research to ensure its applicability and reliability across a broad range of people in the population. At present, a study in a patient sample is ongoing, which will provide further insight into the validity of the DHLI 2.0 in a more vulnerable population. Second, the reliability of the scale "Using security measures" did not reach the conventional reliability threshold of α=0.7, which may be due to its coverage of 2 distinct skills (using strong passwords and using 2-step authentication). These items should therefore be considered separately, rather than combined in a sum score, and further research is warranted to determine whether this domain would benefit from the inclusion of additional items. Third, the correlation between DHLI 2.0 scores and health-related internet use was relatively low. This may partly reflect the relatively high mean age of our sample; although many older adults do use the internet regularly for health purposes, they may not possess the full range of skills needed to do so effectively, efficiently, and safely, which is also reported in previous studies [30,31]. Furthermore, motivational factors could influence the relationship, for example, when individuals do not perceive the internet as a good or reliable source of health information and therefore use it less frequently. Finally, health status itself may play a role, as relatively healthy individuals may have limited need for health-related digital tools (eg, patient portals or online communities), despite having adequate or even high levels of digital health literacy [32]. As a final limitation, the total scores of the subscales demonstrated approximately normal distributions, while some subscales exhibited pronounced ceiling effects. The subscale "Protecting privacy" showed the most substantial ceiling effect, potentially influenced by the phrasing of its introductory item: "When you post a message on a public forum or social media…" This condition led to reduced response rates, as only respondents who engage in such online activity reported an answer. Consequently, the responses may reflect a self-selected group with high privacy awareness within the cohort. The ceiling effects in operational and navigation skills might be affected by the online nature of the questionnaire.

The online world and available technologies are changing rapidly. Since the development of the DHLI 2.0, the integration of artificial intelligence into health applications and internet browsers has accelerated significantly. This further stresses the importance of being able to judge the reliability of information, including being aware of its source and the applicability to oneself [33,34]. In this validation study, we specifically see that the skill to evaluate reliability of online information is what many people find challenging. These findings highlight the importance of supporting individuals in developing critical evaluation skills, particularly as digital health environments become more complex and information sources less transparent. This also underscores the need for further research on how to capture the skills required in an increasingly artificial intelligence–driven digital health environment. Future research should also examine the applicability and utility of the DHLI 2.0 among groups at higher risk of limited digital health literacy. Another interesting avenue for future work is to explore the potential of this instrument in practice; this could be within health care, but also within more community- or welfare-oriented organizations. Although not yet validated for that purpose, the DHLI 2.0 (especially the SF) could support decision-making in health consultations, for example, in determining the suitability of specific eHealth tools for patients.

In conclusion, the DHLI 2.0 provides an up-to-date, reliable, valid, and concise measure of digital health literacy, grounded in a strong theoretical framework and covering 8 key domains of digital health literacy. The SF DHLI proves to be a valuable alternative to measure the broad spectrum of digital health literacy skills using 16 items.

Acknowledgments

We thank Jesse David Marinus and Jolijn Hutjes from Planbureau Fryslân for granting access to the panel, for conducting the qualitative pilot study to make final adjustments to the instrument, for their valuable input on the data collection design, and for programming the questionnaire in their survey software. We thank Bart Koemans for his work on the preliminary validation of the DHLI 2.0 in his bachelor thesis. We thank all the respondents for their participation in the study. Artificial intelligence tools were used during the preparation of this manuscript for refinement, correction, and editing of the manuscript to improve the clarity of language, not for any other purposes.

Funding

The authors appreciate the financial support of the Faith Research Consortium.

Data Availability

The data used in this research are available upon request, provided that the request aligns with the established conditions, which can be found on the Planbureau Fryslân website [35]. Researchers interested in using these data should contact the Planbureau Fryslân Office at info@planbureaufryslan.nl.

Authors' Contributions

RvdV and CD were responsible for the conceptualization of DHLI 2.0, and RvdV, CD, and IT conceptualized the validation study. IT initiated the collaboration with Panel Fryslân. All authors contributed to the methodology. IT conducted the formal data

XSL•FO
RenderX

analyses, which were verified by RvdV and CD. RvdV and IT drafted the original manuscript, and CD and JvV critically reviewed and edited the manuscript. All authors read and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Digital health literacy instrument 2.0 full form.
[PDF File, 101 KB - jmir_v28i1e86879_app1.pdf]

Multimedia Appendix 2
Digital health literacy instrument 2.0 short form.
[PDF File, 113 KB - jmir_v28i1e86879_app2.pdf]

Multimedia Appendix 3
Overview of alterations in DHLI 2.0.
[PDF File, 135 KB - jmir_v28i1e86879_app3.pdf]

References

1. van Tuyl L, Keuper J, Standaar L, Alblas E, Keij B. Toegankelijkheid van digitale zorg: signalen uit de e-healthmonitor 2023 [Article in Dutch]. TSG Tijdschr Gezondheidswet 2025;103(S1):3-7. [doi: 10.1007/s12508-025-00451-7]
2. Keij B, Versluis A, Alblas EE, Keuper JJ, van Tuyl LHD, van der Vaart R. 2023 eHealth monitor: state of digital care [Report in Dutch]. : Rijksinstituut voor Volksgezondheid en Milieu (RIVM); 2024 URL: https://www.rivm.nl/bibliotheek/rapporten/2024-0008.pdf [accessed 2026-03-18]
3. Mahajan S, Lu Y, Spatz ES, Nasir K, Krumholz HM. Trends and predictors of use of digital health technology in the United States. Am J Med 2021 Jan;134(1):129-134. [doi: 10.1016/j.amjmed.2020.06.033] [Medline: 32717188]
4. De Santis KK, Jahnel T, Sina E, Wienert J, Zeeb H. Digitization and health in Germany: cross-sectional nationwide survey. JMIR Public Health Surveill 2021 Nov 22;7(11):e32951. [doi: 10.2196/32951] [Medline: 34813493]
5. Yao R, Zhang W, Evans R, Cao G, Rui T, Shen L. Inequities in health care services caused by the adoption of digital health technologies: scoping review. J Med Internet Res 2022 Mar 21;24(3):e34144. [doi: 10.2196/34144] [Medline: 35311682]
6. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
7. van der Vaart R, Drossaert C. Development of the digital health literacy instrument: measuring a broad spectrum of health 1.0 and health 2.0 skills. J Med Internet Res 2017 Jan 24;19(1):e27. [doi: 10.2196/jmir.6709] [Medline: 28119275]
8. Duimel SLL, Linn AJ, Smets EMA, Smit ES, van Weert JCM. Profiling cancer patients based on their motives for seeking informational and emotional support online. Health Commun 2023 Dec;38(14):3223-3237. [doi: 10.1080/10410236.2022.2144287] [Medline: 36415021]
9. Neter E, Brainin E. Association between health literacy, eHealth literacy, and health outcomes among patients with long-term conditions. Eur Psychol 2019 Jan;24(1):68-81. [doi: 10.1027/1016-9040/a000350]
10. Barbosa MCF, Baldiotti ALP, Braga NS, et al. Cross-cultural adaptation of the digital health literacy instrument (DHLI) for use on Brazilian adolescents. Braz Dent J 2023;34(5):104-114. [doi: 10.1590/0103-6440202305346] [Medline: 38133465]
11. Park E, Kwon M. Testing the Digital Health Literacy Instrument for adolescents: cognitive interviews. J Med Internet Res 2021 Mar 15;23(3):e17856. [doi: 10.2196/17856] [Medline: 33720031]
12. Xie L, Hu H, Lin J, Mo PKH. Psychometric validation of the Chinese digital health literacy instrument among Chinese older adults who have internet use experience. Int J Older People Nurs 2024 Jan;19(1):e12568. [doi: 10.1111/opn.12568] [Medline: 37831059]
13. Bouclaous C, Kamand AA, Daher R, Alrazim A, Kaedbey HD. Digital health literacy and online information-seeking behavior of Lebanese university students in the time of the COVID-19 pandemic and infodemic. Nord J Digit Lit 2023 Apr 11;18(1):60-77. [doi: 10.18261/njdl.18.1.6]
14. Chen SC, Hong Nguyen NT, Lin CY, et al. Digital health literacy and well-being among university students: mediating roles of fear of COVID-19, information satisfaction, and internet information search. Digit Health 2023;9:20552076231165970. [doi: 10.1177/20552076231165970] [Medline: 37009305]
15. Dadaczynski K, Okan O, Messer M, et al. Digital health literacy and web-based information-seeking behaviors of university students in Germany during the COVID-19 pandemic: cross-sectional survey study. J Med Internet Res 2021 Jan 15;23(1):e24097. [doi: 10.2196/24097] [Medline: 33395396]
16. Shudayfat T, Bani Hani S, Al Qadire M. Assessing digital health literacy level among nurses in Jordanian hospitals. Electron J Gen Med 2023;20(5):em525. [doi: 10.29333/ejgm/13466]

XSL•FO
RenderX

17. Alhammad N, Alajlani M, Abd-Alrazaq A, Epiphaniou G, Arvanitis T. Patients' perspectives on the data confidentiality, privacy, and security of mHealth apps: systematic review. J Med Internet Res 2024 May 31;26:e50715. [doi: 10.2196/50715] [Medline: 38820572]
18. Panel Fryslân. Research methodology: sampling and recruitment of a representative internet panel [Report in Dutch]. : Planbureau Fryslân; 2023 URL: https://planbureau.frl/wp-content/uploads/2024/01/Verantwoording-herwerving-panel-2023.pdf [accessed 2026-02-09]
19. Morris NS, MacLean CD, Chew LD, Littenberg B. The single item literacy screener: evaluation of a brief instrument to identify limited reading ability. BMC Fam Pract 2006 Mar 24;7:21. [doi: 10.1186/1471-2296-7-21] [Medline: 16563164]
20. Vereniging voor statistiek en onderzoek handbook GDPR & statistical research. Vsonet. URL: https://vsonet.nl [accessed 2026-03-31]
21. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951 Sep;16(3):297-334. [doi: 10.1007/BF02310555]
22. Streiner DL, Norman GR, Cairney J. Health measurement scales: a practical guide to their development and use: Oxford University Press; 2014. URL: https://academic.oup.com/book/56061 [accessed 2026-03-18]
23. George D, Mallery P. SPSS for Windows Step by Step: A Simple Guide and Reference, 170 Update: Allyn & Bacon; 2010. URL: https://books.google.com.my/books?id=KS1DPgAACAAJ [accessed 2026-03-28]
24. Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007 Jan;60(1):34-42. [doi: 10.1016/j.jclinepi.2006.03.012] [Medline: 17161752]
25. Description of the eight EQF levels. Europass. 2026. URL: https://europass.europa.eu/en/description-eight-eqf-levels [accessed 2026-02-10]
26. Jiang X, Wang L, Leng Y, et al. The level of electronic health literacy among older adults: a systematic review and meta-analysis. Arch Public Health 2024 Nov 7;82(1):204. [doi: 10.1186/s13690-024-01428-9] [Medline: 39511667]
27. Yuen E, Winter N, Savira F, et al. Digital health literacy and its association with sociodemographic characteristics, health resource use, and health outcomes: rapid review. Interact J Med Res 2024 Jul 26;13:e46888. [doi: 10.2196/46888] [Medline: 39059006]
28. Wang C, Chang L, Chen X, Kong J, Qi H. eHealth literacy assessment instruments: scoping review. J Med Internet Res 2025 Aug 20;27:e66965. [doi: 10.2196/66965] [Medline: 40835422]
29. Crocker B, Feng O, Duncan LR. Performance-based measurement of eHealth literacy: systematic scoping review. J Med Internet Res 2023 Jun 2;25:e44602. [doi: 10.2196/44602] [Medline: 37266975]
30. Waterworth S, Honey M. On-line health seeking activity of older adults: an integrative review of the literature. Geriatr Nurs 2018;39(3):310-317. [doi: 10.1016/j.gerinurse.2017.10.016] [Medline: 29198622]
31. Bachofner Y, Seifert A, Sepahniya S, Fabian C. Online health information-seeking among older adults and predictors of use, motivations, and barriers in the context of healthy aging: cross-sectional study. Online J Public Health Inform 2026 Jan 6;18:e77557. [doi: 10.2196/77557] [Medline: 41493565]
32. Qiu CS, Lunova T, Greenfield G, et al. Determinants of digital health literacy: international cross-sectional study. J Med Internet Res 2025 Jun 30;27:e66631. [doi: 10.2196/66631] [Medline: 40587845]
33. Coşkun AB, Elmaoğlu E, Buran C, Yüzer Alsaç S. Integration of ChatGPT and E-Health literacy: opportunities, challenges, and a look towards the future. J Health Rep Technol 2024;10(1):e139748. [doi: 10.5812/jhrt-139748]
34. Nutbeam D, Okan O, EUPHA-HL,-HP,-DH,-ETH,-FS,-GH, IUHPE Working Group on Health Literacy, Technical University of Munich (Germany) and Chair persons. 5.B. Round Table: improving digital health literacy in the era of generative artificial intelligence. Eur J Public Health 2024 Nov 1;34(Supplement_3). [doi: 10.1093/eurpub/ckae144.276]
35. Planbureau Fryslân. URL: https://planbureau.frl/ [accessed 2026-03-24]

Abbreviations

DHLI: digital health literacy instrument
DHLI 2.0: revised version of the DHLI
FF: full form
SF: short form
SILS: Single Item Literacy Screener

XSL•FO
RenderX

Edited by A Schwartz; submitted 31.Oct.2025; peer-reviewed by H Qi, JL Guo; revised version received 12.Feb.2026; accepted 23.Feb.2026; published 01.Apr.2026.

Please cite as:
van der Vaart R, Tuitert I, van 't Veer J, Drossaert C
Validation of the Updated Digital Health Literacy Instrument and Development of a Short Form: Online Survey Study of the General Population
J Med Internet Res 2026;28:e86879
URL: https://www.jmir.org/2026/1/e86879
doi:10.2196/86879

© Rosalie van der Vaart, Inge Tuitert, Job van 't Veer, Constance Drossaert. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Engagement and Cognitive Function Among Older Adults in China: Cross-Sectional Questionnaire Study and Moderated Mediation Model Analysis

Yongqi Du[1*], MPA; Qing Niu[1*], PhD; Gangrui Tan[1], PhD; Jianqian Chao[1], PhD, Prof Dr; Shengxuan Jin[1], PhD; Leixia Wang[1], PhD

Health Management Research Center, School of Public Health, Southeast University, Nanjing, Jiangsu, China

*these authors contributed equally

Corresponding Author:
Jianqian Chao, PhD, Prof Dr
Health Management Research Center
School of Public Health
Southeast University
No. 87, Dingjiaqiao
Gulou District
Nanjing, Jiangsu, 210009
China
Phone: 86 18736361189
Email: jqchao1230220@163.com

Abstract

Background: Given the global demographic shifts and rapid digitalization, digital engagement has emerged as a critical determinant of healthy aging. While previous research has linked digital engagement to cognitive outcomes, the underlying mechanisms remain underexplored among Chinese older adults.

Objective: This study aimed to analyze the relationships between digital engagement and cognitive function among older adults in China through a moderated mediation model guided by the technological reserve hypothesis, with digital health literacy (DHL) and social support as mediators and living arrangements as a moderator.

Methods: We conducted a cross-sectional questionnaire survey using stratified multistage sampling from June to November 2024, including 8123 participants aged 55 years and older. Digital engagement, defined as older adults' use of contemporary digital technologies to support routine daily activities, autonomy, independence, and social inclusion, was assessed using a multidimensional questionnaire. The Chinese eHealth Literacy Scale, the 3-item short version of the Perceived Social Support Scale, and the Mini-Cog test were used to assess DHL, social support, and cognitive function. Guided by a directed acyclic graph based on the technological reserve hypothesis, mediation and moderated mediation analyses were performed using the PROCESS macro in SPSS (IBM Corp) with 5000 bootstrap resamples.

Results: Digital engagement was positively associated with cognitive function among older adults (β=0.241, 95% CI 0.216-0.265). This association was partially mediated by DHL (β=0.059, 95% CI 0.049-0.069) and social support (β=0.012, 95% CI 0.008-0.016), with the combined indirect effects accounting for 29.5% of the total effect (β=0.071, 95% CI 0.061-0.082). Additionally, living arrangements significantly moderated the associations between digital engagement and cognitive function (β=0.109, 95% CI 0.052-0.166), digital engagement and DHL (β=0.063, 95% CI 0.014-0.112), and digital engagement and social support (β=0.151, 95% CI 0.089-0.212). These effects were stronger among older adults living alone.

Conclusions: This study contributes to the understanding of cognitive aging in the digital environment from the perspective of the technological reserve hypothesis and digital engagement. Digital engagement influenced cognitive function via DHL and social support, and these associations of digital engagement with cognitive function, DHL, and social support were stronger among older adults living alone. Digital health interventions and public health policies should target both DHL and social support among older populations and prioritize older adults living alone.

(J Med Internet Res 2026;28:e83955) doi:10.2196/83955

XSL•FO
RenderX

KEYWORDS

digital engagement; cognitive function; digital health literacy; social support; living arrangements; moderated mediation model; technological reserve hypothesis

Introduction

Background

Cognitive function is a critical determinant of dementia, functional independence, quality of life, and health care burden [1]. With an increasingly aged population, cognitive impairment and dementia have become major health and social issues worldwide [2]. Research has indicated that there are about 15.07 million patients with dementia in the population aged 60 and older in China, while the prevalence of mild cognitive impairment is 15.54%, and the number of patients is 38.77 million [3]. The disease burden of dementia and cognitive impairment is huge. The estimated total annual costs of dementia in China will reach 114.2 billion US dollars in 2030 [4]. Since there is currently no effective pharmaceutical treatment for dementia and mild cognitive impairment, it is important to identify modifiable intermediate risk factors that could prevent cognitive decline [5].

With rapid digitalization and the widespread integration of technology into daily life, digital technologies have emerged as a potential determinant of healthy aging. In China, as of 2025, the number of internet users aged 60 and older has reached 161 million, accounting for nearly 14.4% of all internet users [6]. Many studies have indicated that older people are competent and skilled users of digital technologies [7-9]. Consequently, the concept of digital engagement has been introduced to emphasize the breadth and extent of digital technology use among older people [10]. Every day, digital engagement provides new opportunities for older adults to address age-related cognitive decline. Engagement in cognitively challenging activities, such as learning new digital skills or knowledge, plays a protective role against age-related cognitive decline [11,12]. Meanwhile, access to communication technology and social media facilitates interpersonal interactions and enhances social support [13], which helps maintain cognitive health in older adulthood.

Against this background, exploring the association between digital technology use and cognitive health in later life has become an important research focus. However, the cognitive impact of digital technology use in China has not been sufficiently studied and understood [7,14]. First, research has focused on how access to the internet relates to cognitive function and the associations between use frequency in specific domains and cognitive outcomes [8,15-17]. Many studies in China have investigated the effect of internet or social media use on cognitive function [18,19]. But limited studies give attention to the concept of digital engagement [20] and comprehensively measure the dimensions and frequency of digital technology use. As digital technologies have become increasingly integrated into older adults' daily lives, it is important to shift research focus from use to meaningful digital engagement to better understand the cognitive effects of digital technology use [7].

Second, although evidence has established the efficacy of digital health interventions for cognitive decline and cognitive impairment, including dementia [21-23], little is known about how they lead to an improvement in symptoms or behavior. The identification of these mediating mechanisms would be useful for tailoring interventions that specifically target these pathways, improving intervention effectiveness. Some studies in China have estimated the mediating roles of physical activity [14] and social support [20,24]. However, few studies have simultaneously examined the roles of multiple mediators. Including multiple mediators can better reflect real-world mechanisms, help understand the relative importance of different intervention pathways, and reduce bias [25].

Third, while digital technologies become increasingly integrated into older adults' everyday life, growing urbanization and economic reforms in China have transformed intergenerational living arrangements patterns [26]. However, limited studies in China have examined how the association between digital technology use and cognitive function may vary by living arrangements. As the number of older adults living alone in China increases, examining the moderating role of living arrangements in this association is meaningful for developing targeted interventions.

The technological reserve hypothesis provides a theoretical framework for addressing these gaps. This hypothesis, developed by Benge and Scullin, focuses on how digital technology use can counteract cognitive decline and reduce disease burden [27-29]. Technological reserve is defined as "the development of a culture and environment of technology use in older adults that can buffer against the impact of cognitive decline on day-to-day activities" [27]. Further study developed the technological reserve concept and summarized 3 central pathways through which digital technology may prevent cognitive decline [28,30]. First, technology can generate cognitive complexity by engaging older adults in cognitively demanding activities that strengthen cognitive reserve [12,31,32]. By enabling access to diverse information sources (eg, online health information), promoting mentally stimulating activities, and requiring continual learning and adaptation, digital technologies help sustain and challenge cognitive capacities [33]. Second, technology fosters social connection and engagement, which are well-established protective factors against cognitive decline [34]. Through platforms such as social media, messaging apps, and video calls, older adults can maintain social ties, reduce loneliness, and access emotional and instrumental support. Finally, technologies can function as cognitive prosthetics by directly compensating for lapses in memory and executive function, particularly those involved in completing activities of daily living. For example, smartphone apps can deliver reminders for medication adherence [35].

Guided by the technological reserve hypothesis, this study aimed to examine the mediating effect of digital health literacy (DHL) and social support on the relationship between digital

XSL•FO
RenderX

engagement and cognitive function, as well as the moderating effect of living arrangements on the relationships among digital engagement, DHL, social support, and cognitive function.

Theoretical Framework

Digital Engagement and Cognitive Function

Within the technological reserve framework, digital technology use as a modifiable lifestyle behavior is a critical factor that can promote better cognitive outcomes than would be expected based on age, brain injury, or disease stage [30]. In this study, we adopted the term "digital engagement" to define digital technology use among older adults. Digital engagement among older adults refers to their use of contemporary digital technologies and devices to carry out routine and enjoyable everyday activities that support autonomy, independence, and social inclusion [8]. This concept emphasizes how older adults integrate information and communication technologies into daily activities and information-seeking behaviors rather than focusing on limitations [36]. Research has investigated the potential association between digital engagement and cognitive function among older adults. Although some studies suggest potential risks such as sleep disruption or social isolation [37-39], the prevailing evidence supports that digital engagement is positively linked to cognitive function [26,40,41]. Empirical findings generally suggest that regular use of digital technologies (such as social media and online social networking) is positively associated with better cognitive outcomes [42-44]. These benefits are often attributed to increased cognitive stimulation, enhanced social connectivity, and greater engagement in mentally active tasks afforded by digital technology [45,46]. Longitudinal studies further suggest that consistent internet use is associated with slower cognitive decline and a lower subsequent risk of dementia compared with nonuse [47,48]. Meta-analyses of randomized controlled trials also support the effectiveness of digital interventions in improving specific cognitive domains [49,50]. Despite growing evidence, the mechanisms through which digital engagement benefits cognition remain insufficiently understood.

The Mediating Role of DHL

DHL is the ability to seek, understand, evaluate, and apply health information from digital sources to support health-related decision-making [51]. Within the technological reserve framework, DHL can strengthen cognition by engaging older adults in cognitively demanding processes such as evaluating online resources, learning new digital skills, and applying health information in daily life. These processes involve active learning, adaptive reasoning, and problem-solving, which are consistent with mechanisms that sustain cognitive reserve [30]. Moreover, empirical studies support this pathway. Higher DHL is associated with greater adoption of preventive health behaviors, better management of chronic conditions, improved adherence to treatment, and more informed health decisions [52-55]. Such behaviors not only enhance health outcomes but also contribute to maintaining and preserving cognitive function in later life. Thus, DHL may mediate the association between digital engagement and cognitive function.

The Mediating Role of Social Support

Within the technological reserve framework, another plausible pathway operates through social connectivity. Social connectivity refers to the structural and functional aspects of individuals' social relationships, and in later life, is often reflected through social support received from their networks [56-58]. Socioemotional selectivity theory points out that social participation requires a certain cost investment, and members who engage in social participation are bound to consider cost-benefit issues [59]. For older adults, declining physical and cognitive abilities raise the cost of offline social participation, leading to a gradual reduction in face-to-face interactions [60]. Digital technologies offer alternative and more accessible avenues for maintaining social support [61]. Some empirical studies have shown that digital engagement is positively associated with increased social support in later life [61-63]. Social support, in turn, is a well-established protective factor for cognitive function: Older adults with stronger support networks tend to perform better cognitively and face a lower risk of cognitive decline or dementia [64-66]. Thus, social support may serve as a mediator linking digital engagement to cognitive outcomes.

The Moderating Role of Living Arrangements

Economic reforms and urbanization in China since the 1980s have profoundly reshaped family structures, particularly impacting older adults. This shift aligns with modernization theory, predicting smaller families and fewer older adults co-residing with children [67]. Consequently, more older adults live only with a spouse or alone [68]. Given the central role of family in Chinese culture, the study shifted to investigate the moderating role of living arrangements. Within the framework of the technological reserve hypothesis, the cognitive benefits of digital engagement are expected to vary across social contexts that shape baseline access to cognitive and social resources. Living arrangements represent a contextual factor in later life, as co-residence with others may provide routine cognitive stimulation and social interaction, whereas living alone is often associated with reduced offline engagement. Consequently, digital engagement may play a more pronounced compensatory role for older adults living alone by supplementing limited in-person cognitive and social resources. This theoretical perspective provides a rationale for examining living arrangements as a moderator in the association between digital engagement and cognitive function. Additionally, living arrangements may shape both the opportunities and the need for engaging with digital technology [69]. Older adults living alone often rely on digital technologies to maintain social ties, bridge social gaps, and manage independent living [70,71]. In contrast, those in multigenerational households may experience "proxy internet use" (eg, reliance on family members for online tasks), reducing direct engagement and the attendant cognitive stimulation [72]. Digital engagement may therefore be especially protective for those living alone.

Hypotheses

Guided by the technological reserve hypothesis, this study tested a moderated mediation model to examine whether digital engagement is associated with cognitive function through DHL

XSL•FO
RenderX

and social support, and whether these pathways are moderated by living arrangements. Based on the theoretical framework and prior empirical evidence, we propose the following hypotheses:

1. Hypothesis 1: higher digital engagement is correlated with greater cognitive function among older adults.
2. Hypothesis 2: higher digital engagement is associated with higher DHL among older adults.
3. Hypothesis 3: higher DHL is correlated with greater cognitive function among older adults.
4. Hypothesis 4: DHL is a mediator between digital engagement and cognitive function among older adults.
5. Hypothesis 5: higher digital engagement is associated with greater social support among older adults.
6. Hypothesis 6: greater social support is correlated with greater cognitive function among older adults.
7. Hypothesis 7: social support is a mediator between digital engagement and cognitive function among older adults in China.
8. Hypothesis 8: living arrangements moderate the associations of digital engagement with cognitive function, DHL, and social support.

To guide the analyses, we specified a directed acyclic graph (DAG) illustrating the hypothesized relationships among digital engagement, cognitive function, DHL, social support, and living arrangements (Figure 1).

Figure 1. The hypothesized moderated mediation model.

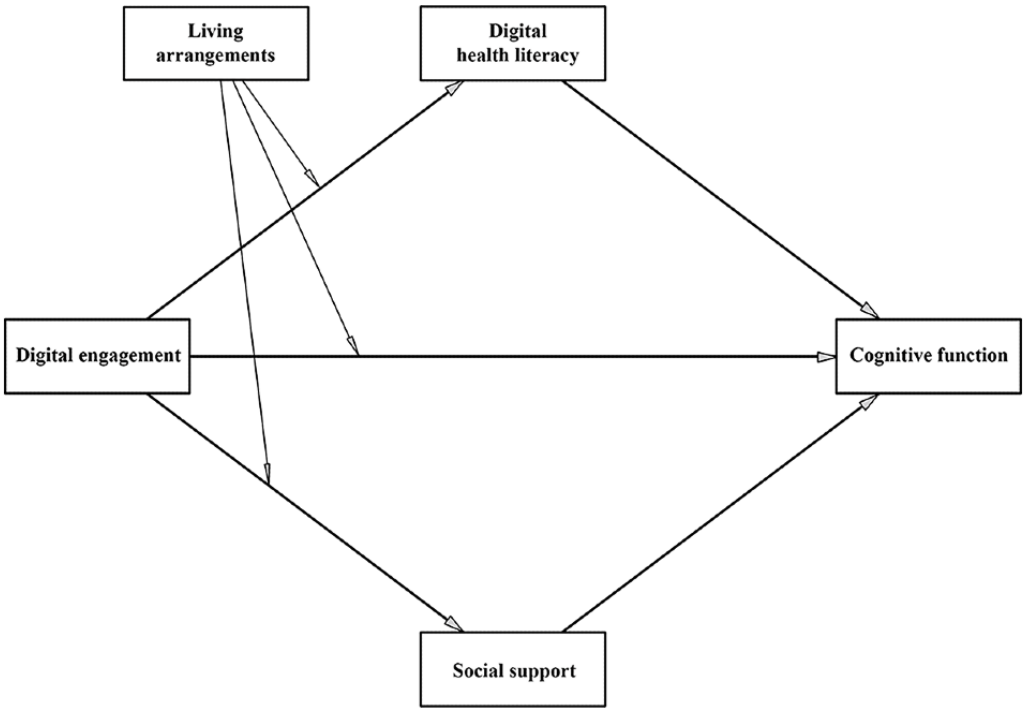

Methods

Study Design and Sampling Procedures

This study used data collected through a large-scale, cross-sectional survey conducted concurrently by 5 academic teams affiliated with 4 major universities in China. To ensure methodological uniformity, all participating sites adhered to a unified research protocol during the implementation phase.

A stratified, multistage sampling framework was used to enhance representativeness across regions with different levels of socioeconomic development. China was first stratified into eastern, central, and western regions, which reflect well-documented gradients in economic development, urbanization, and digital infrastructure. One to 2 provinces were randomly selected from each region. The final sample included Hubei (central China), Shandong and Jiangsu (eastern China), and Guangxi (western China), thereby capturing substantial regional heterogeneity in demographic structure and digital development. Within each selected province, 1 to 2 urban or county-level administrative units were further sampled based on local economic conditions, followed by cluster sampling of communities or villages.

Sample size estimation followed the standard formula for proportion-based calculations:  , where u_α represents the critical value for a 95% CI (u_α=1.96), p is the estimated proportion of older internet users based on the China Internet Network Information Center's 51st Statistical Report [73], *q* is the complementary proportion (*q*=1–p), and *d* denotes the allowable error (1.2%). Based on these parameters, the minimum required sample size was calculated as 6616. To account for possible nonresponses and invalid questionnaires, a 20% oversampling rate was applied, resulting in a target sample of approximately 7940 individuals.

Data Collection and Quality Control

Fieldwork was conducted from June to November 2024 by trained surveyors in collaboration with local village committees or community service offices. Face-to-face interviews were

XSL•FO
RenderX

administered at participants' homes using standardized paper questionnaires. The interviews collected information on sociodemographic characteristics, digital technology use, digital literacy, cognitive function, and quality of life. All surveyors received centralized training to ensure consistent questionnaire administration and interpretation. Upon completion of each interview, field supervisors performed a thorough review of the questionnaires to check for completeness, internal consistency, and data accuracy before submission for entry.

Participants

Eligible participants were older adults who met the following inclusion criteria: (1) aged 55 years and older, (2) had resided in the sampled community or village for at least 6 months, and (3) were able to communicate effectively with investigators. Exclusion criteria included: (1) individuals temporarily absent from their households during the survey period, (2) those diagnosed with terminal illnesses, and (3) those who declined to participate. After excluding incomplete responses, duplicate entries, and respondents younger than 55 years, 8302 valid questionnaires remained. Among these, 179 participants (2.2%) had missing values on at least 1 analytic variable and were excluded from the main analyses. The final analytic sample consisted of 8123 participants (Figure 2).

XSL•FO
RenderX

Figure 2. Flow diagram of participant recruitment and data exclusion, resulting in a final analytic sample of N=8123.

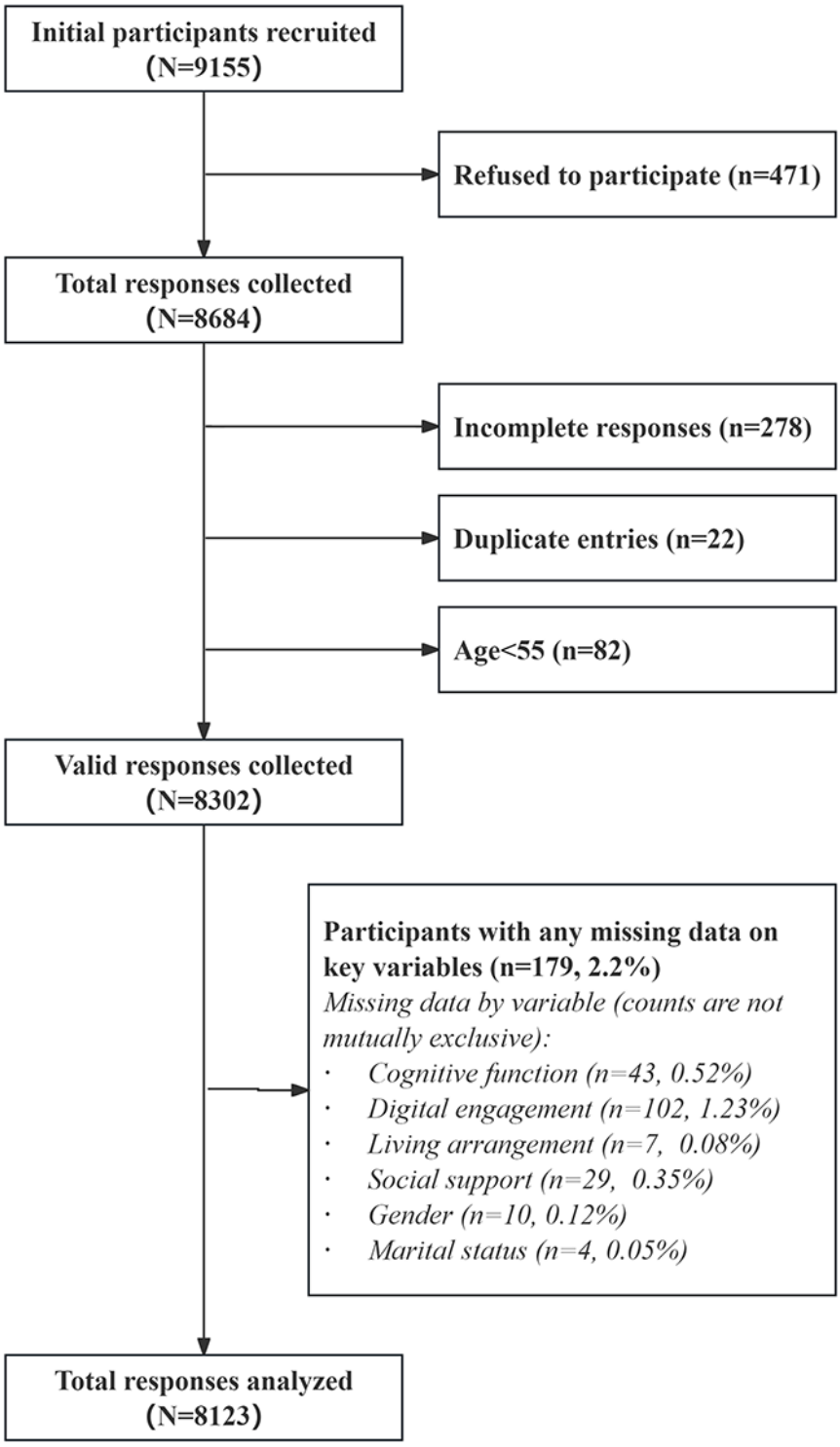

Measurements

Cognitive Function

Cognitive function was assessed using the Mini-Cog test. The Mini-Cog test is a rapid, valid, and reliable screening tool for cognitive impairment [74]. The Mini-Cog Test includes a 3-word recall task (scored 0-3) and the clock drawing test (scored 0-2). The total score ranges from 0 to 5. The Mini-Cog test has demonstrated good screening performance in community-dwelling older adults in China [75]. In this study, the Mini-Cog total score was used as a continuous measure of cognitive function, with higher scores indicating better cognitive performance.

Digital Engagement

Digital engagement was measured using a self-reported scale developed to capture older adults' frequency of participation in various digital activities. The scale included eight items of

XSL•FO
RenderX

digital behaviors: (1) social communication (eg, using WeChat [Tencent] voice or video calls), (2) experience sharing (eg, posting on WeChat Moments, QQ Zone [Tencent], or Weibo [Sina Corporation]), (3) leisure and entertainment (eg, playing online games, listening to music, or watching videos), (4) online transactions (eg, transferring money, making payments, booking services, or trading stocks), (5) information seeking (eg, searching for travel information or reading news), (6) online learning or training, (7) online civic participation (eg, participating in online polls, petitions, or rights protection), and (8) political engagement (eg, online voting or leaving messages on government websites). Participants rated the frequency of each activity on a 5-point Likert scale ranging from 1 (never) to 5 (always). Higher scores indicated higher digital engagement. The scale demonstrated good internal consistency in this sample (Cronbach α=0.876). Although the scale covers multiple items of digital activities, this study conceptualized digital engagement as an overall behavioral tendency reflecting the breadth and extent of digital technology use in daily life. This approach is consistent with the technological reserve hypothesis and the concept of digital engagement, which emphasize cumulative and sustained engagement. Therefore, a composite digital engagement score was used in the analyses.

DHL

DHL was assessed using the eHealth Literacy Scale (eHEALS), a widely validated instrument developed by Norman and Skinner to measure individuals' self-perceived skills in locating, evaluating, and applying electronic health information to health-related problems [51]. The eHEALS consists of 8 items rated on a 5-point Likert scale (1=strongly disagree to 5=strongly agree), reflecting domains such as awareness of available online health resources, confidence in using the internet for health decision-making, and the ability to discern high-quality digital health content. Given the linguistic and cultural differences between the original instrument and the target population of older adults in mainland China, we used the simplified Chinese version (C-eHEALS) translated and validated by Ma and Wu [76]. The C-eHEALS has been confirmed to have good psychometric properties and can therefore be used to evaluate eHealth literacy in Chinese older populations [77]. In this study, the C-eHEALS demonstrated excellent internal consistency, with a Cronbach α coefficient of 0.986, indicating high reliability for use among Chinese older adults.

Social Support

Social support was assessed using the 3-item short version of the Perceived Social Support Scale (PSSS-3), which was developed and validated by Wu et al [78] specifically for use among the Chinese general population. This abbreviated scale was derived from the original 12-item Chinese version of the Multidimensional Scale of Perceived Social Support (MSPSS), originally adapted by Jiang [79] from the version developed by Zimet et al [80]. The PSSS-3 includes 1 item from each of the 3 core dimensions, family support, friend support, and significant others, selected based on the highest factor loadings in a large-scale national sample. The Cronbach α of PSSS-3 in this study was 0.868, demonstrating good internal consistency.

Living Arrangements

Living arrangements were measured as a binary variable indicating whether the older adult lived alone, and were assessed using the following question: "What are your current living arrangements?" Those who reported living alone were coded as 1, and those who reported living with others were coded as 0.

Control Variables

The prior study indicates that demographic and health factors have close links with cognitive function and suggests that these factors should be included in pertinent research [81]. In this study, gender, age, current place of residence, marital status, education, and number of chronic diseases were controlled as covariates.

Statistical Analysis

All analyses were performed using SPSS (version 27; IBM Corp). We first examined the extent and pattern of missing data for all analytic variables. The proportion of missing values for each variable ranged from 0.08% to 1.2% and the overall proportion of missing data was 2.2% (Figure 2). Little's Missing Completely at Random test was conducted using the missing value analysis procedure in SPSS 27. The test indicated that the missing values were independent of the observed or unobserved values (χ^2=14.893; P=.06). Given the low proportion and completely random patterns of missingness, we performed complete-case analyses based on listwise deletion.

Descriptive statistics summarized sample characteristics, with continuous variables reported as mean (SD) and categorical variables as frequencies and percentages. Pearson correlations examined associations among digital engagement, cognitive function, DHL, social support, and living arrangements. Multicollinearity was assessed by the variance inflation factor (VIF), with VIF>5 indicating collinearity. The relationships among variables were specified according to a DAG based on the technological reserve hypothesis (Figure 1) and analyzed using PROCESS models 4 and 8 with 5000 bootstrapped resamples. Effects were considered significant if the 95% bias–corrected CI did not include 0. All models controlled for age, gender, current place of residence, marital status, education, and number of chronic diseases. Continuous variables were standardized prior to analysis. Statistical tests were 2-tailed with α=.05.

Ethical Considerations

The study was reviewed and approved by the Medical Ethics Committee of Zhongda Hospital, Southeast University (approval number 2024ZDSYLL294-Y01). Written informed consent was obtained from all participants before they participated in the study, and they were provided with the opportunity to withdraw at any time during and after the survey. To protect privacy and confidentiality, electronic data were de-identified and stored on password-protected devices accessible only to the research team. No images or supplementary materials in this manuscript contain information that could identify individual participants. There was no compensation for the participants in our study survey.

XSL•FO
RenderX

Results

Demographic Characteristics of the Participants

The final sample consisted of 8123 older adults, with an average age of 71.03 (SD 8.39) years. Men accounted for 42.46% (3449/8123) of the participants. Among the participants, 3990 (49.12%) lived in urban areas, and 4133 (50.88%) lived in rural areas. Most respondents were married (6338/8123, 78.03%), and 51.73% (4202/8123) of the participants had attained a middle school education or above. Overall, 75.28% (6115/8123) of the participants reported at least 1 chronic condition. Summary statistics for main variables, including digital engagement, DHL, social support, living arrangements, and cognitive function, are presented in Table 1.

Table 1. Participant characteristics and descriptive statistics for study variables (N=8123).

Variables	Values
Sex, n (%)	
Male	3449 (42.46)
Female	4674 (57.54)
Age (years), mean (SD)	71.03 (8.39)
Current place of residence, n (%)	
Urban	3990 (49.12)
Rural	4133 (50.88)
Marital status, n (%)	
Married	6338 (78.03)
Unmarried	1785 (21.97)
Education, n (%)	
Primary school or under	3921 (48.27)
Middle or high school	3462 (42.62)
College or above	740 (9.11)
Chronic diseases, n (%)	
0	2008 (24.72)
1	3477 (42.80)
≥2	2638 (32.48)
Digital engagement, mean (SD)	17.37 (8.03)
DHL[a], mean (SD)	20.18 (10.59)
Social support, mean (SD)	16.71 (3.22)
Living arrangements, n (%)	
Living alone	1166 (14.35)
Living with others	6957 (85.65)
Cognitive function, mean (SD)	3.57 (1.49)

[a]DHL: digital health literacy.

Preliminary Correlation Analysis

Table 2 presents the Pearson correlation coefficients among the main variables. Digital engagement was significantly and positively correlated with cognitive function (r=0.365, P<.001), social support (r=0.081, P<.001), and DHL (r=0.575, P<.001), but negatively associated with living arrangements (r=–0.101, P<.001). Cognitive function was also positively correlated with social support (r=0.131, P<.001) and DHL (r=0.347, P<.001), and negatively correlated with living arrangements (r=–0.069, P<.001). VIFs indicated no multicollinearity among digital engagement (VIF=1.732), DHL (VIF=1.666), social support (VIF=1.027), living arrangements (VIF=1.823), and cognitive function.

XSL•FO
RenderX

Table 2. Pearson correlation matrix of digital engagement, digital health literacy, social support, living arrangements, and cognitive function (N=8123).

Variables	Digital engagement	DHL[a]	Social support	Living arrangements	Cognitive function
Digital engagement					
r	1	0.575	0.081	–0.101	0.365
P value	—[b]	<.001	<.001	<.001	<.001
DHL					
r	0.575	1	0.127	–0.073	0.347
P value	<.001	—	<.001	<.001	<.001
Social support					
r	0.081	0.127	1	–0.045	0.131
P value	<.001	<.001	—	<.001	<.001
Living arrangements					
r	–0.101	–0.073	–0.045	1	–0.069
P value	<.001	<.001	<.001	—	<.001
Cognitive function					
r	0.365	0.347	0.131	–0.069	1
P value	<.001	<.001	<.001	<.001	—

[a]DHL: digital health literacy.

[b]Not applicable.

Mediation Analysis

Guided by the DAG-specified conditional process model, we first used model 4 of the PROCESS macro for SPSS [82] to test Hypotheses 1 to 7. The total effect of digital engagement on cognitive function was 0.241 (95% CI 0.216-0.265), of which the direct effect accounted for 70.5% (β=0.170, 95% CI 0.143-0.196) and the combined indirect effects by DHL and social support accounted for 29.5% (β=0.071, 95% CI 0.061-0.082). The findings indicated a moderate but statistically meaningful mediation by DHL and social support.

Table 3 and Figure 3 present the results of the mediation analysis. As shown in model 3, the direct effect of digital engagement on cognitive function was significant (Model 3: β=0.170, 95% CI 0.143-0.196; *P*<.001), thus supporting Hypothesis 1. Digital engagement also demonstrated a significant and positive association with DHL (Model 1: β=0.400, 95% CI 0.379-0.420; *P*<.001), supporting Hypothesis 2. Additionally, DHL was significantly and positively related to cognitive function (Model 3: β=0.148, 95% CI 0.123-0.174; *P*<.001), supporting Hypothesis 3. Furthermore, digital engagement was positively and significantly correlated with social support (Model 2: β=0.129, 95% CI 0.103-0.155; *P*<.001), supporting Hypothesis 5. Social support, in turn, showed a significant and positive correlation with cognitive function (Model 3: β=0.091, 95% CI 0.071-0.112; *P*<.001), supporting Hypothesis 6.

XSL•FO
RenderX

Table 3. Mediation analysis of the association between digital engagement and cognitive function through digital health literacy and social support, adjusted for covariates (PROCESS Model 4; N=8123).

Model 1[a] (DHL[b])			Model 2[c] (social support)		Model 3[d] (cognitive function)	
	β[e] (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value
Explanatory variable						
Digital engagement	0.400 (0.379 to 0.420)	<.001	0.129 (0.103 to 0.155)	<.001	0.170 (0.143 to 0.196)	<.001
Mediator variables						
DHL	—[f]	—	—	—	0.148 (0.123 to 0.174)	<.001
Social support	—	—	—	—	0.091 (0.071 to 0.112)	<.001
Control variables						
Gender	–0.035 (–0.070 to –0.001)	.047	0.169 (0.125 to 0.213)	<.001	–0.044 (–0.085 to –0.003)	.03
Age	–0.041 (–0.060 to –0.021)	<.001	0.065 (0.041 to 0.089)	<.001	–0.095 (–0.117 to –0.072)	<.001
Current place of residence	0.321 (0.280 to 0.362)	<.001	–0.465 (–0.518 to –0.413)	<.001	0.048 (–0.003 to 0.098)	.06
Marital status	0.039 (–0.004 to 0.082)	.076	0.159 (0.104 to 0.213)	<.001	0.134 (0.084 to 0.184)	<.001
Education	0.294 (0.261 to 0.327)	<.001	0.172 (0.131 to 0.213)	<.001	0.179 (0.140 to 0.218)	<.001
Chronic diseases	–0.124 (–0.147 to –0.101)	<.001	–0.078 (–0.107 to –0.048)	<.001	0.034 (0.007 to 0.062)	.01
Constant	–0.471 (–0.563 to –0.380)	<.001	–0.353 (–0.469 to –0.237)	<.001	–0.383 (–0.491 to –0.275)	<.001

[a]$F_{7, 8115}$=811.099; R^2=0.412.

[b]DHL: digital health literacy.

[c]$F_{7, 8115}$=65.515; R^2=0.054.

[d]$F_{9, 8113}$=214.765; R^2=0.192.

[e]β: standardized regression coefficient.

[f]Not applicable.

XSL•FO
RenderX

Figure 3. Mediation model of the association between digital engagement and cognitive function through digital health literacy and social support; values are standardized coefficients adjusted for covariates (PROCESS Model 4; N=8123; ***P<.001).

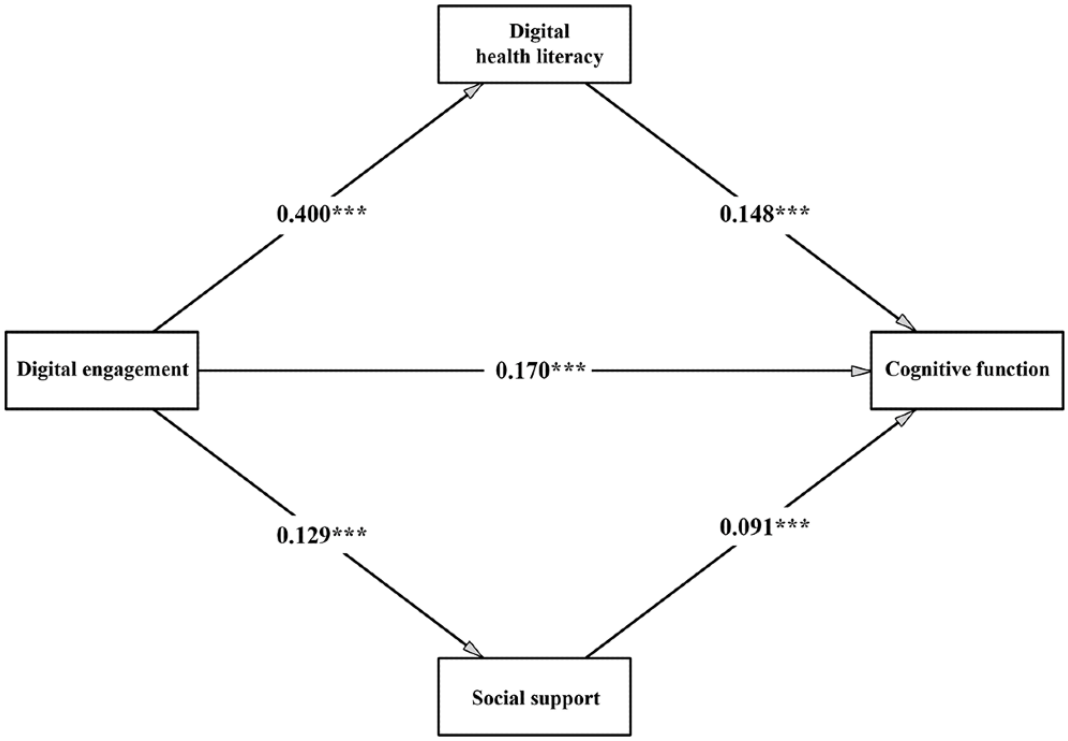

These findings suggest that DHL and social support play a partial mediating role in the relationship between digital engagement and cognitive function, with indirect effects of 0.059 (95% CI 0.049-0.069) and 0.012 (95% CI 0.008-0.016), respectively, supporting Hypotheses 4 and 7. Combining 2 mediation effects, the total indirect effect was 0.071 (95% CI 0.061-0.082). The bootstrap test results for indirect effects are reported in Table 4.

Table 4. Bootstrap estimates of indirect effects of digital engagement on cognitive function through digital health literacy and social support (PROCESS Model 4; N=8123).

Indirect effects path	Effects	SE	95% CI	Proportion of effects (%)
Total indirect effect	0.071	0.005	0.061-0.082	100
DE[a]→DHL[b]→CF[c]	0.059	0.005	0.049-0.069	83.10
DE→SS[d]→CF	0.012	0.002	0.008-0.016	16.90

[a]DE: digital engagement.

[b]DHL: digital health literacy.

[c]CF: cognitive function.

[d]SS: social support.

Moderated Mediation Analysis

To examine the moderated mediation effects involving living arrangements, we used Model 8 of the PROCESS macro for SPSS [82], using 5000 bootstrap resamples and a 95% bias-corrected CI. The results are reported in Table 5. The analysis revealed a significant and positive interaction effect between digital engagement and living arrangements on cognitive function (Model 6: β=0.109, 95% CI 0.052-0.166; P<.001), suggesting a moderating role of living arrangements. As illustrated in Figure 4, the beneficial association between digital engagement and cognitive performance was stronger among older adults who lived alone, relative to those who lived with others.

XSL•FO
RenderX

Table 5. Moderated mediation analysis testing moderation by living arrangements in the associations between digital engagement and digital health literacy, digital engagement and social support, digital engagement and cognitive function, adjusted for covariates (PROCESS Model 8; N=8123).

Variables	Model 4[a] (DHL[b])		Model 5[c] (social support)		Model 6[d] (cognitive function)	
	β[e] (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value
Explanatory variable						
Digital engagement	0.391 (0.370 to 0.413)	<.001	0.109 (0.082 to 0.137)	<.001	0.156 (0.129 to 0.183)	<.001
Mediator variables						
DHL	—[f]	—	—	—	0.147 (0.121 to 0.172)	<.001
Social support	—	—	—	—	0.089 (0.069 to 0.110)	<.001
Moderating variables						
Living arrangements	0.054 (–0.011 to 0.119)	.10	0.058 (–0.024 to 0.141)	.17	0.154 (0.078 to 0.230)	<.001
DE×LA[g]	0.063 (0.014 to 0.112)	.01	0.151 (0.089 to 0.212)	<.001	0.109 (0.052 to 0.166)	<.001
Control variables						
Gender	–0.036 (–0.070 to –0.001)	.044	0.167 (0.123 to 0.211)	<.001	–0.044 (–0.085 to –0.003)	.04
Age	–0.040 (–0.060 to –0.021)	<.001	0.065 (0.041 to 0.089)	<.001	–0.094 (–0.117 to –0.072)	<.001
Current place of residence	0.321 (0.280 to 0.362)	<.001	–0.465 (–0.517 to –0.413)	<.001	0.047 (–0.003 to 0.097)	.07
Marital status	0.063 (0.007 to 0.119)	.03	0.175 (0.103 to 0.247)	<.001	0.210 (0.144 to 0.276)	<.001
Education	0.292 (0.259 to 0.324)	<.001	0.167 (0.126 to 0.208)	<.001	0.176 (0.137 to 0.215)	<.001
Chronic diseases	–0.123 (–0.146 to –0.100)	<.001	–0.077 (–0.106 to –0.047)	<.001	0.034 (0.007 to 0.061)	.01
Constant	–0.492 (–0.591 to –0.393)	<.001	–0.360 (0.485 to –0.234)	<.001	–0.456 (–0.572 to –0.339)	<.001

[a]$F_{9, 8113}$=632.208; R^2=0.412.

[b]DHL: digital health literacy.

[c]$F_{9, 8113}$=53.680; R^2=0.056.

[d]$F_{11, 8111}$=178.558; R^2=0.195.

[e]β: standardized regression coefficient.

[f]Not applicable.

[g]DE×LA: the interaction term between digital engagement and living arrangements.

XSL•FO
RenderX

Figure 4. The moderating effect of living arrangements on the association between digital engagement and cognitive function. The difference in simple slopes indicated that the association between digital engagement and cognitive function was significantly stronger for individuals living alone than for those living with others (PROCESS Model 8, N=8123).

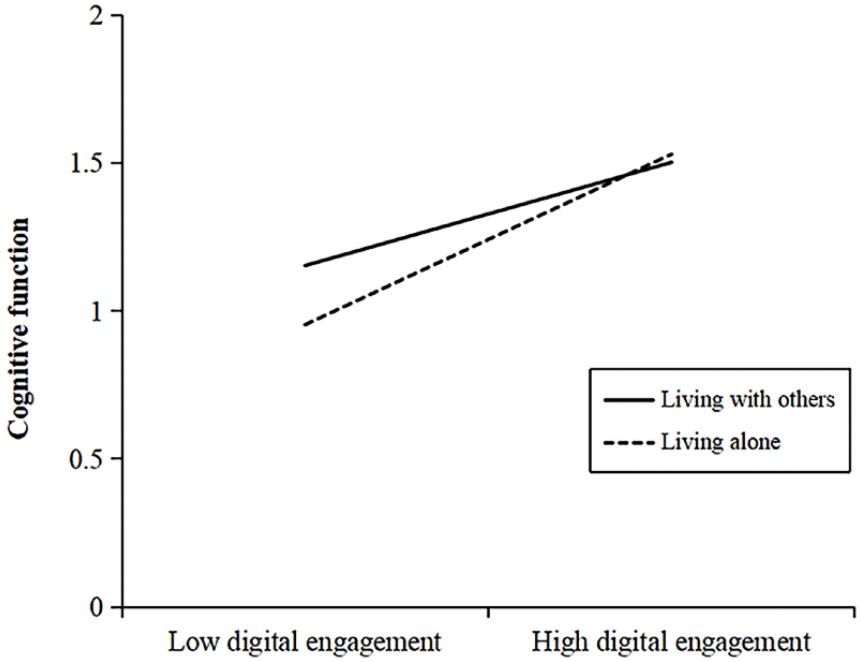

In addition to the interaction effect on the direct path, the results also revealed significant moderating effects of living arrangements on the first stages of both mediation pathways. Specifically, the interaction term between digital engagement and living arrangements significantly predicted DHL (Model 4: β=0.063, 95% CI 0.014-0.112; *P*=.01 and social support (Model 5: β=0.151, 95% CI 0.089-0.212; *P*<.001). As illustrated in Figures 5 and 6, the beneficial association of digital engagement with both DHL and social support was stronger among older adults who lived alone, relative to those who lived with others. These findings indicate that living arrangements moderate the associations of digital engagement with cognitive function, DHL, and social support, supporting Hypothesis 8.

Figure 5. The moderating effect of living arrangements on the association between digital engagement and digital health literacy. The difference in simple slopes indicated that the association between digital engagement and digital health literacy was significantly stronger for individuals living alone than for those living with others (PROCESS Model 8, N=8123).

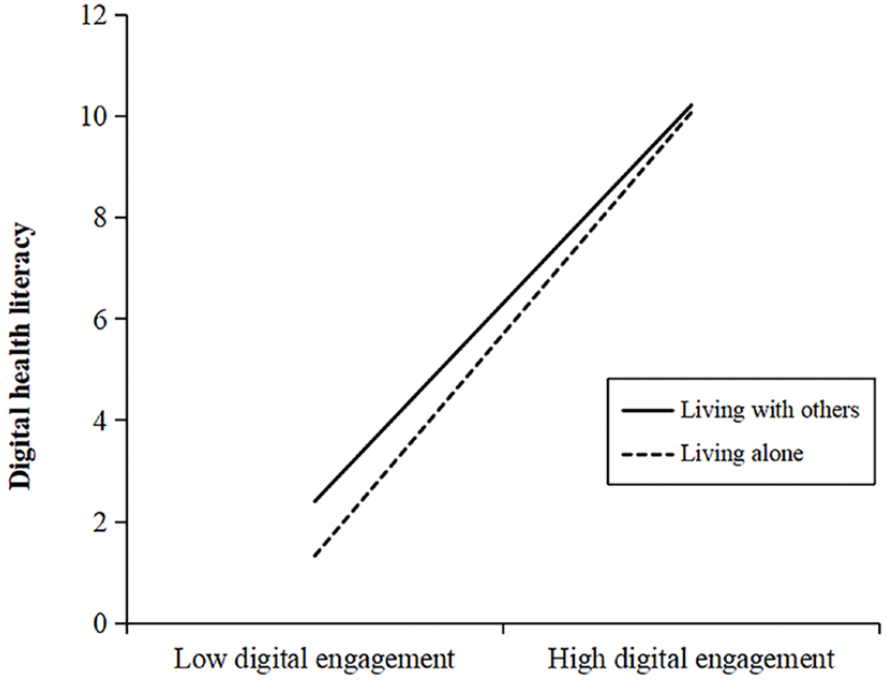

XSL•FO
RenderX

Figure 6. The moderating effect of living arrangements on the association between digital engagement and social support. The difference in simple slopes indicated that the association between digital engagement and social support was significantly stronger for individuals living alone than for those living with others (PROCESS Model 8, N=8123).

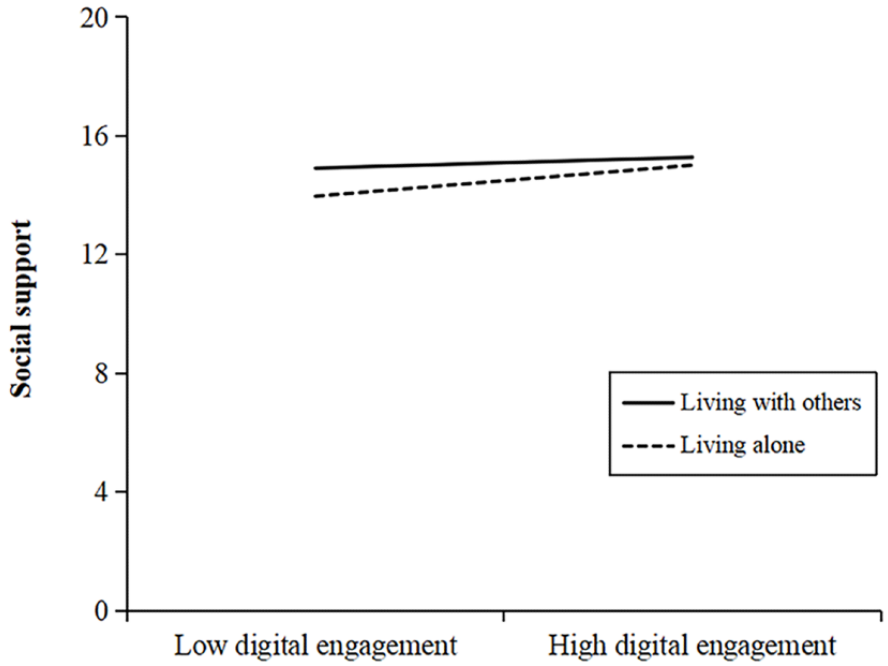

Table 6 reports the conditional indirect effects of digital engagement on cognitive function by each path. For the DHL mediator, the indirect effect of digital engagement on cognitive function was 0.057 (95% CI 0.048-0.068) among participants living with others and 0.067 (95% CI 0.054-0.080) among those living alone. The corresponding index of moderated mediation was 0.009 (95% CI 0.003-0.016), indicating a significantly stronger indirect effect by DHL for participants living alone. Similarly, for the social support mediator, the indirect effect was 0.010 (95% CI 0.007-0.013) for those living with others and 0.023 (95% CI 0.016-0.032) for those living alone. The index of moderated mediation for the social support pathway was 0.013 (95% CI 0.007-0.021). In all cases, the 95% CI values excluded zero, indicating that both indirect effects were significantly stronger among older adults living alone.

Table 6. Indices of moderated mediation for two conditional indirect effects of digital engagement on cognitive function by living arrangements (PROCESS Model 8; N=8123).

Conditional indirect effects path	Effects	SE	95% CI
DE[a]→DHL[b]→CF[c] (living with others)	0.057	0.006	0.048-0.068
DE→DHL→CF (living alone)	0.067	0.007	0.054-0.080
Index of the moderated mediation	0.009	0.003	0.003-0.016
DE→SS[d]→CF (living with others)	0.010	0.002	0.007-0.013
DE→SS→CF (living alone)	0.023	0.004	0.016-0.032
Index of the moderated mediation	0.013	0.003	0.007-0.021

[a]DE: digital engagement.

[b]DHL: digital health literacy.

[c]CF: cognitive function.

[d]SS: social support.

Discussion

Principal Findings

Guided by the technological reserve hypothesis and using a large, community-based sample of older adults in China, this study investigated the mechanisms underlying the association between digital engagement and cognitive function among older Chinese adults. We found that higher digital engagement was associated with better cognitive performance. DHL and social support partially mediated this association, and the combined indirect effects accounted for 29.5% of this association. Living arrangements moderated both the direct and indirect pathways, with stronger benefits among older adults living alone. These findings extend prior work on technology use and cognition in later life and broaden the application of the technological reserve hypothesis in the Chinese context.

XSL•FO
RenderX

The Association Between Digital Engagement and Cognitive Function

Digital engagement was significantly and positively associated with cognitive function. This supports the technological reserve hypothesis that digital technology use is a modifiable behavioral factor that can promote better cognitive outcomes [30]. It also aligns with prior work linking internet or computer use to cognition [83-88] and with recent studies in China demonstrating that internet use and digital activities enhance cognitive function [14,47]. By adopting the construct of digital engagement rather than a simple use-versus-nonuse dichotomy, our study advances the field by situating technology use within the everyday life context of older adults, emphasizing how they integrate information and communication technologies into their ongoing activities, social interaction, and information seeking [36]. This finding implies that encouraging sustained and meaningful digital engagement is a promising strategy for public health and aging policies aiming to strengthen cognitive function among older adults.

The Mediating Role of DHL and Social Support

Consistent with the cognitive-stimulation pathway posited by the technological reserve hypothesis, higher digital engagement was associated with higher DHL, which in turn related to better cognitive performance [28,30]. Notably, the DHL pathway accounted for the majority of the total indirect effect, indicating that health-related digital competencies may be a primary mechanism linking engagement to cognition. Specifically, engaging with digital technology improves DHL because digital skills are among the core skills of DHL [89,90]. In turn, higher DHL denotes a stronger capacity to seek, understand, appraise, and apply health information [91]. These processes involve engaging with cognitively complex information [30], helping older adults build cognitive reserve. This finding advances the eHealth Literacy Model, which posits that DHL is underpinned by cognition [92], and indicates that DHL also serves as a tool that shapes cognition through ongoing, cognitively complex digital activities. Accordingly, interventions should combine user-friendly interfaces with structured, progressive training in cognitively complex digital tasks, ensuring that everyday digital engagement serves as sustained cognitive stimulation. Practical examples include stepwise smartphone or tablet training delivered in community settings (eg, locating health information from reliable sources, evaluating credibility and misinformation, and applying information to everyday self-management tasks).

Social support mediated the association between digital engagement and cognition, consistent with the social-connectivity pathway [28,30]. This finding aligns with previous research showing that digital engagement has the potential to enhance cognitive function among older individuals by addressing feelings of loneliness and improving the social support they receive from relatives and friends [20]. Specifically, digital engagement enables cheap and easy communication between older adults in distant communities, increasing social connections, overcoming social and spatial barriers, and providing a convenient way to stay in touch with families, friends, and the outside world [62]. In turn, better social support is associated with better cognitive outcomes in older adults [65,93-96]. This finding underscores that interventions should help older adults form and maintain digital social ties so that online interactions translate into perceived social support and, ultimately, better cognitive outcomes. For example, programs could incorporate facilitated online peer groups and a “Digital Buddy system” to help older adults translate online interactions into perceived support [97].

The Moderating Role of Living Arrangements

Our study further revealed that living arrangements played a significant moderating role in the associations of digital engagement with cognitive function, DHL, and social support. Compared with older adults who live with others, those living alone experienced a significantly stronger positive effect of digital technology engagement on cognitive function, consistent with the previous studies [26,98]. This moderating effect was significantly present in both mediating pathways: older adults living alone gained greater benefits in terms of DHL and social support from digital engagement than those living with others. Specifically, older adults living alone, due to a lack of effective offline social interactions, are more reliant on virtual social networks facilitated by digital technologies [60]. This reliance partially compensates for the reduced social support associated with solitary living, thereby mitigating its negative impact on cognitive function [70]. Additionally, older adults who live alone are less likely to engage in proxy internet use [72] and thus rely more on themselves to use digital devices (eg, searching for health information online). Furthermore, since older adults living alone are less often burdened with caregiving responsibilities for grandchildren, they have more freedom and time to engage with digital technologies [26]. Our findings suggest that digital engagement serves as a more efficacious strategy for mitigating cognitive decline among older adults living alone compared to those living with others.

Our finding is broadly consistent with other international evidence. A cohort study in America reported that transitioning into Internet use was associated with better cognitive function and slower cognitive decline, and that these benefits were more pronounced among older adults living alone than among those living with others [84]. Additionally, a 2-country longitudinal study in Sweden and the Netherlands observed less decline in global cognition among baseline internet users after adjustment for living situation [86]. Beyond cognitive outcomes, findings based on the Survey of Health, Ageing and Retirement in Europe further indicate that internet use can attenuate the association between living alone and loneliness across different European welfare regimes, implying that digital engagement may buffer psychosocial vulnerabilities of solitary living [99]. Taken together, although the prevalence and social meaning of living alone differ across cultures, converging evidence supports that digital engagement may serve as a compensatory resource for older adults with constrained offline or household-based support.

This moderation finding has practical implications for intervention design. Digital inclusion initiatives to help older people adapt to digital technologies should prioritize this vulnerable group. An integrated community-based approach may be especially useful: individual digital coaching (eg, guided practice in health information seeking) coupled with structured

XSL•FO
RenderX

social support (eg, online group chats). Including a "living alone" priority within such initiatives may help maximize equity and potential cognitive benefits, while also addressing social isolation risks that have been recognized as a public health and policy concern.

Limitations and Future Research

Despite these contributions, several limitations should be acknowledged. First, the cross-sectional design precludes causal inferences. Although our models adjusted for a set of covariates, endogeneity, reverse causation, and unmeasured confounding cannot be fully ruled out. Thus, our results only show associations, not causality. Future studies should use longitudinal designs with extended follow-up periods to elucidate temporal dynamics and disentangle potential reverse causation. Second, reliance on self-reported measures introduces the risk of recall bias, especially among participants with cognitive impairments, despite our use of validated instruments to attenuate this issue. Third, digital engagement was operationalized as a composite measure. Thus, this study could not disentangle potentially differential effects of specific types of digital activities on cognitive function. Finally, due to data constraints, living arrangements were operationalized as solitary versus nonsolitary living, precluding differentiation among various household compositions (eg, living with a spouse, children, or extended family). Given the important role of family structures in the well-being of older adults in China, future research should refine classifications of living arrangements to better explore their moderating effects.

Conclusions

This study contributes to the understanding of cognitive aging in the digital environment from the perspective of the technological reserve hypothesis and digital engagement. First, it offers an innovative framework based on the technological reserve hypothesis for understanding the moderating and mediating mechanisms of DHL, social support, and living arrangements. Second, it advances previous assessment methods of digital technology application by using a comprehensive measure. Our results increase understanding of the mechanisms underlying the cognitive effects of digital technology use and provide insights for designing digital health interventions and public health policies.

Acknowledgments

The authors would like to express their sincere gratitude to all participants in this study, as well as to the staff of the community and village committees who facilitated the fieldwork. This study was supported by the Major Project of the National Social Science Fund of China (grant number 23&ZD188). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to restrictions under existing Data Use Agreements, but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Wang Q, Ni J, Guan Y, Liu X, Li M, Xue H, et al. Associations of depression and symptomatic knee osteoarthritis with cognitive function among middle-aged and older adults: evidence from CHARLS in China. J Gerontol Soc Work Routledge 2025 Jul 28:1-17. [doi: 10.1080/01634372.2025.2539741] [Medline: 40719179]
2. World Alzheimer report 2024: global changes in attitudes to dementia. Alzheimer's Disease International. 2024. URL: https://www.alzint.org/resource/world-alzheimer-report-2024/ [accessed 2025-07-25]
3. Jia L, Du Y, Chu L, Zhang Z, Li F, Lyu D, COAST Group. Prevalence, risk factors, and management of dementia and mild cognitive impairment in adults aged 60 years or older in China: a cross-sectional study. Lancet Public Health 2020 Dec;5(12):e661-e671 [FREE Full text] [doi: 10.1016/S2468-2667(20)30185-7] [Medline: 33271079]
4. Xu J, Wang J, Wimo A, Fratiglioni L, Qiu C. The economic burden of dementia in China, 1990-2030: implications for health policy. Bull World Health Organ 2017;95(1):18-26 [FREE Full text] [doi: 10.2471/BLT.15.167726] [Medline: 28053361]
5. Shah H, Albanese E, Duggan C, Rudan I, Langa KM, Carrillo MC, et al. Research priorities to reduce the global burden of dementia by 2025. Lancet Neurol 2016 Nov;15(12):1285-1294. [doi: 10.1016/S1474-4422(16)30235-6] [Medline: 27751558]

XSL•FO
RenderX

6. The 56th statistical report on China's internet development [Web page in Chinese]. China Internet Network Information Center. URL: https://www3.cnnic.cn/n4/2025/0721/c88-11328.html [accessed 2025-07-31]
7. Kebede A, Ozolins L, Holst H, Galvin K. Digital engagement of older adults: scoping review. J Med Internet Res 2022;24(12):e40192 [FREE Full text] [doi: 10.2196/40192] [Medline: 36477006]
8. Aleti T, Figueiredo B, Reid M, Martin D, Sheahan J, Hjorth L. Older adults' digital competency, digital risk perceptions and frequency of everyday digital engagement. Inf Technol People 2025;38(8):118. [doi: 10.1108/itp-05-2024-0624]
9. Xu L, Ng D, Lee C, Peng P, Chu S. A systematic review of digital literacy in lifelong learning for older adults: challenges, strategies, and learning outcomes. Education Tech Research Dev 2025;73(6):3627-3674. [doi: 10.1007/s11423-025-10530-w]
10. Iqbal S, Fischl C, Asai R. Older persons' social participation, health and well-Being through digital engagement. Act Adapt Aging Routledge 2025;49(4):534-563. [doi: 10.1080/01924788.2025.2512299]
11. Chan M, Haber S, Drew L, Park D. Training older adults to use tablet computers: does it enhance cognitive function? Gerontologist 2016;56(3):475-484 [FREE Full text] [doi: 10.1093/geront/gnu057] [Medline: 24928557]
12. Almeida-Meza P, Steptoe A, Cadar D. Is engagement in intellectual and social leisure activities protective against dementia risk? Evidence from the English Longitudinal Study of Ageing. J Alzheimers Dis 2021;80(2):555-565 [FREE Full text] [doi: 10.3233/JAD-200952] [Medline: 33554903]
13. Czaja S, Boot W, Charness N, Rogers W, Sharit J. Improving social support for older adults through technology: findings from the PRISM randomized controlled trial. Gerontologist 2018 May 08;58(3):467-477 [FREE Full text] [doi: 10.1093/geront/gnw249] [Medline: 28201730]
14. Wang J, Zhang N, Huang C, Wu Q, Tong J. Internet use, physical activity, and cognitive function in Chinese older adults: a cross-lagged panel analysis. Front Aging Neurosci 2025;17:1579874 [FREE Full text] [doi: 10.3389/fnagi.2025.1579874] [Medline: 40405917]
15. Hua Z, Wang F. Association between WeChat use and memory performance among older adults in China: the mediating role of depression. Behav Sci (Basel) 2022;12(9):323 [FREE Full text] [doi: 10.3390/bs12090323] [Medline: 36135127]
16. Zhou Y, Abuduxukuer K, Wang C, Wei J, Shi W, Li Y, et al. WeChat usage and preservation of cognitive functions in middle-aged and older Chinese adults: indications from a nationally representative survey, 2018-2020. BMC Public Health 2024;24(1):1783 [FREE Full text] [doi: 10.1186/s12889-024-19210-5] [Medline: 38965535]
17. Chen B, Yang C, Ren S, Li P, Zhao J. Relationship between internet use and cognitive function among middle-aged and older Chinese adults: 5-year longitudinal study. J Med Internet Res 2024;26:e57301 [FREE Full text] [doi: 10.2196/57301] [Medline: 39539034]
18. Yu D, Fiebig D. Internet use and cognition among middle-aged and older adults in China: a cross-lagged panel analysis. J Econ Ageing 2020;17:100262. [doi: 10.1016/j.jeoa.2020.100262]
19. Ding L, Lu J, Ma X. WeChat use, cognitive function, and depressive symptoms: examining longitudinal relationships among older Chinese adults from a national survey. Mass Commun Soc 2025:1-21. [doi: 10.1080/15205436.2025.2546449]
20. Liu Z, Li Z. Relationships between digital engagement and the mental health of older adults: evidence from China. PLoS One 2024;19(8):e0308071 [FREE Full text] [doi: 10.1371/journal.pone.0308071] [Medline: 39106268]
21. Li A, Qiang W, Li J, Geng Y, Qiang Y, Zhao J. Effectiveness of an exergame-based training program on physical and cognitive function in older adults with cognitive impairment: a randomized controlled trial in rural China. BMC Geriatr 2025;25(1):892 [FREE Full text] [doi: 10.1186/s12877-025-06341-6] [Medline: 41219720]
22. Wen X, Song S, Tian H, Cui H, Zhang L, Sun Y, et al. Intervention of computer-assisted cognitive training combined with occupational therapy in people with mild cognitive impairment: a randomized controlled trial. Front Aging Neurosci 2024;16:1384318 [FREE Full text] [doi: 10.3389/fnagi.2024.1384318] [Medline: 38832072]
23. Cheng Z, Zhou M, Sabran K. Mobile app-based interventions to improve the well-being of people with dementia: a systematic literature review. Assist Technol 2024;36(1):64-74. [doi: 10.1080/10400435.2023.2206439] [Medline: 37115814]
24. Yu X, Ang S, Zhang Y. Exploring rural-urban differences in the association between internet use and cognitive functioning among older adults in China. J Gerontol B Psychol Sci Soc Sci 2024;79(4):gbad195 [FREE Full text] [doi: 10.1093/geronb/gbad195] [Medline: 38147307]
25. Preacher KJ, Hayes AF. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behav Res Methods 2008;40(3):879-891 [FREE Full text] [doi: 10.3758/brm.40.3.879] [Medline: 18697684]
26. Li Y, Han W, Hu M. Does internet access make a difference for older adults' cognition in urban China? The moderating role of living arrangements. Health Soc Care Community 2022;30(4):e909-e920. [doi: 10.1111/hsc.13493] [Medline: 34245201]
27. Benge JF, Scullin MK. Implications for technological reserve development in advancing age, cognitive impairment, and dementia. Behav Brain Sci 2020;43:e157. [doi: 10.1017/S0140525X20000126] [Medline: 32772985]
28. Wolff JL, Benge JF, Cassel CK, Monin JK, Reuben DB. Emerging topics in dementia care and services. J Am Geriatr Soc 2021;69(7):1763-1773 [FREE Full text] [doi: 10.1111/jgs.17341] [Medline: 34245585]
29. Benge JF, Kiselica AM, Aguirre A, Hilsabeck RC, Douglas M, Paydarfar D, et al. Technology use and subjective cognitive concerns in older adults. Arch Gerontol Geriatr 2023;106:104877 [FREE Full text] [doi: 10.1016/j.archger.2022.104877] [Medline: 36459914]

XSL•FO
RenderX

30. Benge JF, Scullin MK. A meta-analysis of technology use and cognitive aging. Nat Hum Behav 2025;9(7):1405-1419 [FREE Full text] [doi: 10.1038/s41562-025-02159-9] [Medline: 40229575]
31. Stern Y, Arenaza-Urquijo EM, Bartrés-Faz D, Belleville S, Cantilon M, Chetelat G, et al. Whitepaper: defining and investigating cognitive reserve, brain reserve, and brain maintenance. Alzheimers Dement 2020;16(9):1305-1311 [FREE Full text] [doi: 10.1016/j.jalz.2018.07.219] [Medline: 30222945]
32. Stern Y, Albert M, Barnes CA, Cabeza R, Pascual-Leone A, Rapp PR. A framework for concepts of reserve and resilience in aging. Neurobiol Aging 2023;124:100-103 [FREE Full text] [doi: 10.1016/j.neurobiolaging.2022.10.015] [Medline: 36653245]
33. Krell-Roesch J, Vemuri P, Pink A, Roberts RO, Stokin GB, Mielke MM, et al. Association between mentally stimulating activities in late life and the outcome of incident mild cognitive impairment, with an analysis of the APOE e4 genotype. JAMA Neurol 2017;74(3):332-338 [FREE Full text] [doi: 10.1001/jamaneurol.2016.3822] [Medline: 28135351]
34. Penninkilampi R, Casey A, Singh MF, Brodaty H. The association between social engagement, loneliness, and risk of dementia: a systematic review and meta-analysis. J Alzheimers Dis 2018;66(4):1619-1633. [doi: 10.3233/JAD-180439] [Medline: 30452410]
35. Scullin MK, Jones WE, Phenis R, Beevers S, Rosen S, Dinh K, et al. Using smartphone technology to improve prospective memory functioning: a randomized controlled trial. J Am Geriatr Soc 2022;70(2):459-469 [FREE Full text] [doi: 10.1111/jgs.17551] [Medline: 34786698]
36. Quan-Haase A, Martin K, Schreurs K. Interviews with digital seniors: ICT use in the context of everyday life. Inf Commun Soc 2016;19(5):691-707. [doi: 10.1080/1369118x.2016.1140217]
37. Ackermann K, Awaworyi Churchill S, Smyth R. Broadband internet and cognitive functioning. Econ Rec Hoboken 2023;99(327):536-563 [FREE Full text] [doi: 10.1111/1475-4932.12757]
38. Schacter DL. Media, technology, and the sins of memory. Mem Mind Media 2022;1:e1 [FREE Full text] [doi: 10.1017/mem.2021.3] [Medline: 34423305]
39. Small GW, Lee J, Kaufman A, Jalil J, Siddarth P, Gaddipati H, et al. Brain health consequences of digital technology use . Dialogues Clin Neurosci 2020;22(2):179-187 [FREE Full text] [doi: 10.31887/DCNS.2020.22.2/gsmall] [Medline: 32699518]
40. Medlock S, Eslami S, Askari M, Arts DL, Sent D, de Rooij SE, et al. Health information-seeking behavior of seniors who use the internet: a survey. J Med Internet Res 2015;17(1):e10 [FREE Full text] [doi: 10.2196/jmir.3749] [Medline: 25574815]
41. Bielak AAM. How can we not 'lose it' if we still don't understand how to 'use it'? Unanswered questions about the influence of activity participation on cognitive performance in older age--a mini-review. Gerontology 2010;56(5):507-519. [doi: 10.1159/000264918] [Medline: 19996570]
42. Yildirim E, Ogel-Balaban H. Cognitive functions among healthy older adults using online social networking. Appl Neuropsychol Adult 2023;30(4):401-408. [doi: 10.1080/23279095.2021.1951269] [Medline: 34310244]
43. Cotten SR, Schuster AM, Seifert A. Social media use and well-being among older adults. Curr Opin Psychol 2022;45:101293. [doi: 10.1016/j.copsyc.2021.12.005] [Medline: 35065352]
44. Han M, Tan XY, Lee R, Lee JK, Mahendran R. Impact of social media on health-related outcomes among older adults in Singapore: qualitative study. JMIR Aging 2021;4(1):e23826 [FREE Full text] [doi: 10.2196/23826] [Medline: 33595437]
45. Ihle A, Bavelier D, Maurer J, Oris M, Kliegel M. Internet use in old age predicts smaller cognitive decline only in men. Sci Rep 2020;10(1):8969 [FREE Full text] [doi: 10.1038/s41598-020-65846-9] [Medline: 32488153]
46. Firth J, Torous J, Stubbs B, Firth JA, Steiner GZ, Smith L, et al. The "online brain": how the internet may be changing our cognition. World Psychiatry 2019;18(2):119-129 [FREE Full text] [doi: 10.1002/wps.20617] [Medline: 31059635]
47. Deng C, Shen N, Li G, Zhang K, Yang S. Digital isolation and dementia risk in older adults: longitudinal cohort study. J Med Internet Res 2025;27:e65379 [FREE Full text] [doi: 10.2196/65379] [Medline: 39969956]
48. Choi NG, Dinitto DM. Internet use among older adults: association with health needs, psychological capital, and social capital. J Med Internet Res 2013;15(5):e97 [FREE Full text] [doi: 10.2196/jmir.2333] [Medline: 23681083]
49. Chen C, Huang N, Hu B, Zhang M, Yuan J, Guo J. The effectiveness of digital technology interventions for cognitive function in older adults: a systematic review and meta-analysis of randomized controlled trials. Geroscience 2025;47(1):653-683. [doi: 10.1007/s11357-024-01446-z] [Medline: 39688787]
50. Gao Y, Liu N. Effects of digital technology-based serious games interventions for older adults with mild cognitive impairment: a meta-analysis of randomised controlled trials. Age Ageing 2025;54(4):afaf080. [doi: 10.1093/ageing/afaf080] [Medline: 40192627]
51. Norman CD, Skinner HA. eHEALS: the eHealth Literacy Scale. J Med Internet Res 2006;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
52. Xin Y, Weina H, Yan D. Digital literacy impacts quality of life among older adults through hierarchical mediating mechanisms. Sci Rep 2025;15(1):19288 [FREE Full text] [doi: 10.1038/s41598-025-04472-9] [Medline: 40456807]
53. Liu S, Lu Y, Wang D, He X, Ren W, Kong D, et al. Impact of digital health literacy on health-related quality of life in Chinese community-dwelling older adults: the mediating effect of health-promoting lifestyle. Front Public Health 2023;11:1200722 [FREE Full text] [doi: 10.3389/fpubh.2023.1200722] [Medline: 37415711]

XSL•FO
RenderX

54. Arias López MDP, Ong BA, Borrat Frigola X, Fernández AL, Hicklent RS, Obeles AJT, et al. Digital literacy as a new determinant of health: a scoping review. PLOS Digit Health 2023;2(10):e0000279 [FREE Full text] [doi: 10.1371/journal.pdig.0000279] [Medline: 37824584]
55. Xie L, Zhang S, Xin M, Zhu M, Lu W, Mo PK. Electronic health literacy and health-related outcomes among older adults: a systematic review. Prev Med 2022;157:106997. [doi: 10.1016/j.ypmed.2022.106997] [Medline: 35189203]
56. Santini ZI, Jose PE, York Cornwell E, Koyanagi A, Nielsen L, Hinrichsen C, et al. Social disconnectedness, perceived isolation, and symptoms of depression and anxiety among older Americans (NSHAP): a longitudinal mediation analysis. Lancet Public Health 2020;5(1):e62-e70 [FREE Full text] [doi: 10.1016/S2468-2667(19)30230-0] [Medline: 31910981]
57. Domènech-Abella J, Mundó J, Haro JM, Rubio-Valera M. Anxiety, depression, loneliness and social network in the elderly: longitudinal associations from The Irish Longitudinal Study on Ageing (TILDA). J Affect Disord 2019;246:82-88. [doi: 10.1016/j.jad.2018.12.043] [Medline: 30578950]
58. Cohen S, Wills TA. Stress, social support, and the buffering hypothesis. Psychological Bulletin 1985;98(2):310-357. [doi: 10.1037/0033-2909.98.2.310] [Medline: 3901065]
59. Carstensen LL. Social and emotional patterns in adulthood: support for socioemotional selectivity theory. Psychol Aging 1992;7(3):331-338. [doi: 10.1037//0882-7974.7.3.331] [Medline: 1388852]
60. Cheng M, Su W, Li H, Li L, Xu M, Zhao X, et al. Factors influencing the social participation ability of rural older adults in China: a cross-sectional study. Front Public Health 2022;10:1001948 [FREE Full text] [doi: 10.3389/fpubh.2022.1001948] [Medline: 36684961]
61. Sun K, Zhou J. Understanding the impacts of internet use on senior citizens' social participation in China: evidence from longitudinal panel data. Telemat Inform 2021;59:101566. [doi: 10.1016/j.tele.2021.101566]
62. Du X, Liao J, Ye Q, Wu H. Multidimensional internet use, social participation, and depression among middle-aged and elderly Chinese individuals: nationwide cross-sectional study. J Med Internet Res 2023;25:e44514 [FREE Full text] [doi: 10.2196/44514] [Medline: 37647119]
63. He T, Huang C, Li M, Zhou Y, Li S. Social participation of the elderly in China: the roles of conventional media, digital access and social media engagement. Telemat Inform 2020;48:101347. [doi: 10.1016/j.tele.2020.101347]
64. Long C, Yang W, Glaser K. Social support, cognition, and mental health among older people in China: a longitudinal life course study. Soc Sci Med 2025;381:118279 [FREE Full text] [doi: 10.1016/j.socscimed.2025.118279] [Medline: 40479799]
65. Ma T, Liao J, Ye Y, Li J. Social support and cognitive activity and their associations with incident cognitive impairment in cognitively normal older adults. BMC Geriatr 2024;24(1):38 [FREE Full text] [doi: 10.1186/s12877-024-04655-5] [Medline: 38191348]
66. Joyce J, Ryan J, Owen A, Hu J, McHugh Power J, Shah R, et al. Social isolation, social support, and loneliness and their relationship with cognitive health and dementia. Int J Geriatr Psychiatry 2022;37(1):00 [FREE Full text] [doi: 10.1002/gps.5644] [Medline: 34741340]
67. Ragab E, Ghannam A. Modernisation in Arab societies: the theoretical and analytical view. Int J Sociol Soc Policy 2001;21:99-131. [doi: 10.1108/01443330110789727]
68. Zeng Y, Wang Z. Dynamics and policy implications of family households and elderly living arrangements in China. American Enterprise Institute. 2019. URL: http://www.jstor.org/stable/resrep24663.5 [accessed 2025-07-30]
69. Cáceres RB, Chaparro AC. Age for learning, age for teaching: the role of inter-generational, intra-household learning in internet use by older adults in Latin America. Inf, Commun Soc 2017:1-17 [FREE Full text] [doi: 10.1080/1369118x.2017.1371785]
70. Silva P, Matos AD, Martinez-Pecino R. Can the internet reduce the loneliness of 50+ living alone? Inf, Commun Soc 2022;25(1):17-33 [FREE Full text] [doi: 10.1080/1369118X.2020.1760917]
71. Berner J, Aartsen M, Deeg D. Predictors in starting and stopping internet use between 2002 and 2012 by Dutch adults 65 years and older. Health Informatics J 2019;25(3):715-730 [FREE Full text] [doi: 10.1177/1460458217720398] [Medline: 28747085]
72. Grošelj D, Reisdorf BC, Petrovčič A. Obtaining indirect internet access: an examination how reasons for internet non-use relate to proxy internet use. Telecommunications Policy 2019;43(3):213-224. [doi: 10.1016/j.telpol.2018.07.004]
73. The 51st statistical report on China's Internet development [Web page in Chinese]. China Internet Network Information Center. 2023. URL: https://www3.cnnic.cn/n4/2023/0302/c199-10755.html [accessed 2025-09-04]
74. Adhikari SP, Dev R, Borson S. Modifying the Mini-Cog to screen for cognitive impairment in nonliterate individuals. Int J Alzheimers Dis 2021;2021:5510093 [FREE Full text] [doi: 10.1155/2021/5510093] [Medline: 34447592]
75. Yang L, Yan J, Jin X, Jin Y, Yu W, Xu S, et al. Screening for dementia in older adults: comparison of Mini-Mental State Examination, Mini-Cog, Clock Drawing Test and AD8. PLoS One 2016;11(12):e0168949 [FREE Full text] [doi: 10.1371/journal.pone.0168949] [Medline: 28006822]
76. Ma Z, Wu M. The psychometric properties of the Chinese eHealth Literacy Scale (C-eHEALS) in a Chinese rural population: cross-sectional validation study. J Med Internet Res 2019;21(10):e15720 [FREE Full text] [doi: 10.2196/15720] [Medline: 31642811]

XSL•FO
RenderX

77. Xu RH, Zhou L, Lu SY, Wong EL, Chang J, Wang D. Psychometric validation and cultural adaptation of the simplified Chinese eHealth Literacy Scale: cross-sectional study. J Med Internet Res 2020;22(12):e18613 [FREE Full text] [doi: 10.2196/18613] [Medline: 33284123]
78. Wu Y, Tang J, Du Z, Chen K, Wang F, Sun X, et al. Development of a short version of the perceived social support scale: based on classical test theory and ant colony optimization. BMC Public Health 2025;25(1):232 [FREE Full text] [doi: 10.1186/s12889-025-21399-y] [Medline: 39833852]
79. Jiang Q. Perceived Social Support Scale. Chin J Behav Med Sci 2001;10(10):41-43 [FREE Full text]
80. Zimet G, Dahlem N, Zimet S, Farley G. The Multidimensional Scale of Perceived Social Support. J Pers Assess Routledge 1988 Mar;52(1):30-41. [doi: 10.1207/s15327752jpa5201_2]
81. Wu X, Tang Y, He Y, Wang Q, Wang Y, Qin X. Prevalence of cognitive impairment and its related factors among Chinese older adults: an analysis based on the 2018 CHARLS data. Front Public Health 2024;12:1500172 [FREE Full text] [doi: 10.3389/fpubh.2024.1500172] [Medline: 39776486]
82. Hayes AF. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach. New York, London: The Guilford Press; Jan 24, 2022.
83. Cho G, Betensky RA, Chang VW. Internet usage and the prospective risk of dementia: a population-based cohort study. J Am Geriatr Soc 2023;71(8):2419-2429. [doi: 10.1111/jgs.18394] [Medline: 37132331]
84. Kim YK, Han SH. Internet use and cognitive functioning in later life: focus on asymmetric effects and contextual factors. Gerontologist 2022;62(3):425-435 [FREE Full text] [doi: 10.1093/geront/gnab149] [Medline: 34614179]
85. Kamin ST, Lang F. Internet use and cognitive functioning in late adulthood: longitudinal findings from the Survey of Health, Ageing and Retirement in Europe (SHARE). J Gerontol B Psychol Sci Soc Sci 2020;75(3):534-539. [doi: 10.1093/geronb/gby123] [Medline: 30346591]
86. Berner J, Comijs H, Elmståhl S, Welmer AK, Sanmartin Berglund J, Anderberg P, et al. Maintaining cognitive function with internet use: a two-country, six-year longitudinal study. Int Psychogeriatr 2019;31(07):929-936 [FREE Full text] [doi: 10.1017/s1041610219000668]
87. d'Orsi E, Xavier AJ, Rafnsson SB, Steptoe A, Hogervorst E, Orrell M. Is use of the internet in midlife associated with lower dementia incidence? Results from the English Longitudinal Study of Ageing. Aging Ment Health 2018;22(11):1525-1533 [FREE Full text] [doi: 10.1080/13607863.2017.1360840] [Medline: 28795579]
88. Klimova B. Use of the internet as a prevention tool against cognitive decline in normal aging. Clin Interv Aging Dove Medical Press Ltd 2016;11:1231-1237 [FREE Full text] [doi: 10.2147/CIA.S113758] [Medline: 27672317]
89. Li S, Cui G, Yin Y, Xu H. Associations between health literacy, digital skill, and eHealth literacy among older Chinese adults: a cross-sectional study. Digit Health 2023;9:20552076231178431 [FREE Full text] [doi: 10.1177/20552076231178431] [Medline: 37256010]
90. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
91. Jung SO, Son YH, Choi E. E-health literacy in older adults: an evolutionary concept analysis. BMC Med Inform Decis Mak 2022;22(1):28 [FREE Full text] [doi: 10.1186/s12911-022-01761-5] [Medline: 35101005]
92. El Benny ME, Kabakian-Khasholian T, El-Jardali F, Bardus M. Application of the eHealth literacy model in digital health interventions: scoping review. J Med Internet Res 2021;23(6):e23473 [FREE Full text] [doi: 10.2196/23473] [Medline: 34081023]
93. Zhang C, Tang D, Wang Y, Jiang S, Liu X. Community support and promoting cognitive function for the elderly. Front Psychol 2022;13:942474 [FREE Full text] [doi: 10.3389/fpsyg.2022.942474] [Medline: 36148108]
94. Li B, Guo Y, Deng Y, Zhao S, Li C, Yang J, et al. Association of social support with cognition among older adults in China: a cross-sectional study. Front Public Health 2022;10:947225 [FREE Full text] [doi: 10.3389/fpubh.2022.947225] [Medline: 36225770]
95. Wang Y, Li J, Fu P, Jing Z, Zhao D, Zhou C. Social support and subsequent cognitive frailty during a 1-year follow-up of older people: the mediating role of psychological distress. BMC Geriatr 2022;22(1):162 [FREE Full text] [doi: 10.1186/s12877-022-02839-5] [Medline: 35227216]
96. Tariq A, Beihai T, Abbas N, Ali S, Yao W, Imran M. Role of perceived social support on the association between physical disability and symptoms of depression in senior citizens of Pakistan. Int J Environ Res Public Health 2020;17(5):1485 [FREE Full text] [doi: 10.3390/ijerph17051485] [Medline: 32106585]
97. Kwan RYC, Ng F, Lai M, Wong D, Chan S. The effects of Digital Buddy programme on older adults' mental well-being: study protocol for a multi-centre, cluster randomized controlled trial. Trials 2023;24(1):95 [FREE Full text] [doi: 10.1186/s13063-023-07130-5] [Medline: 36750879]
98. Zhong R, Ning W. Impact of living arrangements and internet use on the mental health of Chinese older adults. Front Public Health 2024;12:1395181 [FREE Full text] [doi: 10.3389/fpubh.2024.1395181] [Medline: 39712316]
99. Silva P, Delerue Matos A, Martinez-Pecino R. The contribution of the Internet to reducing social isolation in individuals aged 50 years and older: quantitative study of data from the Survey of Health, Ageing and Retirement in Europe. J Med Internet Res 2022;24(1):e20466 [FREE Full text] [doi: 10.2196/20466] [Medline: 34982040]

XSL•FO
RenderX

Abbreviations

C-eHEALS: Chinese version of the eHealth Literacy Scale
DAG: directed acyclic graph
DHL: digital health literacy
eHEALS: eHealth Literacy Scale
MSPSS: Multidimensional Scale of Perceived Social Support
PSSS-3: 3-item short version of the Perceived Social Support Scale
VIF: variance inflation factor

Edited by S Brini; submitted 11.Sep.2025; peer-reviewed by E Calatayud, S Dong; comments to author 17.Nov.2025; accepted 24.Dec.2025; published 20.Jan.2026.

Please cite as:
Du Y, Niu Q, Tan G, Chao J, Jin S, Wang L
Digital Engagement and Cognitive Function Among Older Adults in China: Cross-Sectional Questionnaire Study and Moderated Mediation Model Analysis
J Med Internet Res 2026;28:e83955
URL: https://www.jmir.org/2026/1/e83955
doi:10.2196/83955
PMID:

©Yongqi Du, Qing Niu, Gangrui Tan, Jianqian Chao, Shengxuan Jin, Leixia Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Immersive Technologies for Cognitive Rehabilitation in Dementia and Mild Cognitive Impairment: Systematic Review

Jaromir Konecny[1], PhD; Giuseppe Lanza[2,3], PhD; Serafino Buono[3,4], PhD; Raffaele Ferri[3], MD; Antonina Luca[4], PhD; Radek Martinek[1], PhD; Sabrina Musso[3], PsyD; Aurora Palmigiano[2,3], PsyD; Barbora Prauzkova[5], MSc; Angelica Quercia[3], PhD; Francesco Rundo[6], EngD; Ramachandran Avala Subramanian[1], MSc; Alessandro Serretti[3,4], PhD; Michal Prauzek[1], PhD

[1]Department of Cybernetics and Biomedical Engineering, VSB – Technical University of Ostrava, 17. listopadu 2172/15, Ostrava - Poruba, Czech Republic

[2]Department of Surgery and Medical-Surgical Specialties, University of Catania, Catania, Italy

[3]Department of Clinical Neurosciences, Oasi Research Institute-IRCCS, Troina, Italy

[4]Department of Medicine and Surgery, "Kore" University of Enna, Enna, Italy

[5]Department of Management, VSB – Technical University of Ostrava, Ostrava - Poruba, Czech Republic

[6]Unit of Bioinformatics and Statistics, Oasi Research Institute-IRCCS, Troina, Italy

Corresponding Author:
Michal Prauzek, PhD
Department of Cybernetics and Biomedical Engineering, VSB – Technical University of Ostrava, 17. listopadu 2172/15, Ostrava - Poruba, Czech Republic

Abstract

Background: Cognitive decline across the mild cognitive impairment (MCI)–dementia continuum is a major driver of loss of independence and growing health- and social-care burden. Immersive technologies, such as virtual reality (VR), augmented reality (AR), and Cave Automatic Virtual Environment (CAVE) systems, are increasingly explored as tools to enhance engagement, personalization, and ecological validity in cognitive rehabilitation.

Objective: This systematic review synthesizes current evidence on the usability, therapeutic effects, and implementation challenges of immersive technologies for cognitive rehabilitation in MCI and dementia.

Methods: A systematic search of Scopus and Web of Science was conducted for peer-reviewed journal articles published between 2021 and 2026. Eligible studies investigated VR, AR, or CAVE interventions targeting cognitive rehabilitation outcomes in MCI and/or dementia and reported measures related to usability or acceptability, or cognitive, functional, or behavioral outcomes. Due to heterogeneity in technologies, intervention content, and outcome measures, findings were synthesized narratively with comparisons across modalities and study designs.

Results: In total, 119 studies met the inclusion criteria. Across immersive VR interventions, signals of benefit were most consistently reported for memory, attention, and executive functioning, with several studies also targeting outcomes with higher ecological relevance (eg, everyday task performance and functional skills). AR approaches primarily support context-aware cueing and task guidance in real-world settings, aiming to strengthen daily functioning and independence. CAVE-based systems were frequently used for spatial navigation and embodied interaction, offering advantages for supervised clinical deployment. Key barriers included cybersickness and comfort issues, interface complexity, and onboarding demands in cognitively impaired users, limited accessibility and standardization of outcome measures, small samples and short follow-up periods, and practical constraints related to cost, space, staffing, and caregiver involvement.

Conclusions: Immersive VR, AR, and CAVE systems are feasible and often engaging for cognitive rehabilitation in MCI and dementia, with promising therapeutic signals but substantial uncertainty driven by methodological and implementation heterogeneity. Future work should prioritize standardized reporting (intervention components, dose, and adverse events), clinically meaningful outcomes (including functional end points), adequately powered comparative trials, and explicit evaluation of scalability and real-world deployment pathways.

(J Med Internet Res 2026;28:e84349) doi:10.2196/84349

KEYWORDS

cognitive rehabilitation; dementia; CAVE system; virtual reality; augmented reality; mild cognitive impairment; mixed reality; PRISMA; Cave Automatic Virtual Environment system

XSL•FO
RenderX

Introduction

Dementia represents a rapidly growing global health challenge. In 2019, approximately 55 million people were estimated to be living with dementia worldwide, and this number is projected to rise to 139 million by 2050, underscoring the urgency of scalable, evidence-based rehabilitation and support strategies [1]. Cognitive decline across the mild cognitive impairment (MCI)–dementia continuum is a leading driver of loss of independence in later life and a growing burden for health and social care systems. MCI is commonly conceptualized as an intermediate clinical state between typical cognitive aging and dementia, but its course is heterogeneous—some individuals remain stable or improve, whereas others progress to dementia—which makes early, targeted intervention both attractive and complex [2-4]. Dementia syndromes, including Alzheimer disease (AD), involve progressive impairments across multiple cognitive domains with downstream effects on everyday functioning and quality of life, reinforcing the need for interventions that are not only clinically meaningful but also feasible in real-world settings [5,6].

Nonpharmacological approaches remain central to cognitive rehabilitation and support, particularly because disease-modifying pharmacotherapies are limited in scope and may not address day-to-day functional challenges across disease stages [5,7]. Conventional cognitive training and rehabilitation can improve specific skills, yet long-term engagement, personalization to heterogeneous deficits, and transfer to everyday functioning are persistent challenges [5,7,8]. In this context, outcomes related to activities of daily living (ADL) and instrumental activities of daily living (IADL) are especially important because they capture functionally meaningful abilities that shape independence and care needs [9]. Digital and computer-based interventions have therefore been explored as scalable tools for cognitive support, but questions remain regarding ecological validity, long-term adherence, and implementation barriers outside controlled research contexts [7,8].

Immersive technologies are increasingly proposed as a route to address these limitations by enabling interactive, multisensory, context-rich tasks that can approximate real-life demands. Conceptually, these systems can be situated within the broader extended reality (XR) landscape and the reality-virtuality continuum, which clarifies the relationships among virtual reality (VR), augmented reality (AR), and mixed reality [10-12]. In this review, we use VR to denote predominantly computer-generated environments experienced with varying levels of immersion (eg, head-mounted displays or projection-based systems), AR to denote digital overlays and cues integrated into the real world (often via mobile or wearable devices), and CAVE (Cave Automatic Virtual Environment) systems as a projection-based, room-scale configuration of VR that supports embodied interaction and navigation without headgear [10,11,13]. Distinguishing these modalities is practically relevant because they differ in interaction demands, supervision requirements, space and cost constraints, and potential adverse effects, which can influence both feasibility and outcomes in older adults with cognitive impairment [14-16].

Across cognitive rehabilitation, immersive tasks have been used to target memory, attention, executive functions, and spatial or navigation abilities, and they can embed cognitively demanding activities within ADL or IADL-like scenarios to enhance ecological validity and functional relevance [13,17-19]. However, translation beyond controlled settings remains constrained by substantial heterogeneity in platforms and immersion levels, interaction modalities, task designs, intervention dose (session duration and frequency), comparator conditions, and outcome selection, which complicates synthesis and limits the strength and generalizability of inferences [4,20,21]. These challenges are amplified by the fact that XR-based outcomes cannot be interpreted independently of feasibility; usability, acceptability, and accessibility directly shape adherence, delivered exposure (effective “dose”), safety, and the interpretability of cognitive and functional measures in cognitively impaired older adults [14,15].

To address this, this review adopts an explicit usability lens. We use usability in the human-computer interaction sense as a multidimensional construct encompassing effectiveness, efficiency, and satisfaction in a specified context of use. In this literature, usability evidence may be drawn from standardized instruments, task performance indicators (eg, errors and assistance needs), completion and adherence metrics, qualitative user-experience reports, and stakeholder perspectives (patients, caregivers, and clinicians) [18,22]. This framing is particularly important in dementia and MCI because age- and disease-related sensory, motor, and cognitive constraints can make interface complexity, onboarding demands, fatigue, and tolerability decisive determinants of whether an intervention is deliverable, interpretable, and scalable [8,14,15].

Several systematic reviews have examined technology-supported cognitive rehabilitation, but the evidence base remains fragmented with respect to target population (MCI vs dementia or AD), XR modality (VR-only vs broader XR), and the depth and standardization of usability and implementation synthesis [3,4,20,23-27]. In parallel, broader assistive-technology reviews in dementia synthesize important adjacent evidence (including immersive and wearable technologies) but do not provide a modality-explicit, rehabilitation-focused synthesis that treats usability and translation constraints as analytically coded outcomes [28]. Likewise, AR-focused reviews for cognitively impaired populations emphasize action assistance and learning, but their scope is not centered on cognitive rehabilitation outcomes across the MCI-dementia continuum and does not integrate projection-based immersive configurations, such as CAVE [29]. Taken together, these gaps motivate a consolidated systematic review that (1) spans the MCI-dementia continuum, (2) compares immersive technology types with explicit attention to VR (including CAVE as projection-based VR) and AR, and (3) treats usability and implementation constraints as first-class evidence streams alongside cognitive and functional outcomes [14-16].

Accordingly, this review is guided by the following research question: What is the current evidence on usability, feasibility, and reported intervention outcomes of VR, AR, and CAVE systems for cognitive rehabilitation in individuals with MCI and dementia? In addressing this question, we additionally

XSL•FO
RenderX

examine how intervention characteristics (technology type, immersion level, interaction demands, and dose), study design choices, and outcome selection contribute to heterogeneity in reported findings and shape the feasibility of translation beyond controlled settings [20,21,30].

Methods

Review Contributions and Structure

First, it synthesizes current evidence on the usability, therapeutic efficacy, and limitations of immersive systems (VR, AR, and CAVE) in cognitive rehabilitation for individuals with MCI and dementia.

Second, it compares approaches across technologies, study designs, and outcome measures to identify key trends and limitations in the field.

Finally, it outlines challenges and proposes priorities for future research, including methodological standardization and scalability of interventions.

The overall structure of this systematic review is summarized in Figure 1. After the Introduction section, the Methods section details the search strategy, eligibility criteria, study selection, data extraction, and synthesis approach. The Results section then reports the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)-based selection process and presents the included evidence across VR, AR, and CAVE modalities. Next, the Discussion section interprets the findings with an emphasis on usability, feasibility, and translational considerations and highlights implications for research and practice. Finally, the Conclusions section summarizes the key takeaways and priorities for future work.

Figure 1. Graphical outline of the study. ADL: activities of daily living; AR: augmented reality; CAVE: Cave Automatic Virtual Environment; IADL: instrumental activities of daily living; MCI: mild cognitive impairment; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; UX: user experience; VR: virtual reality.

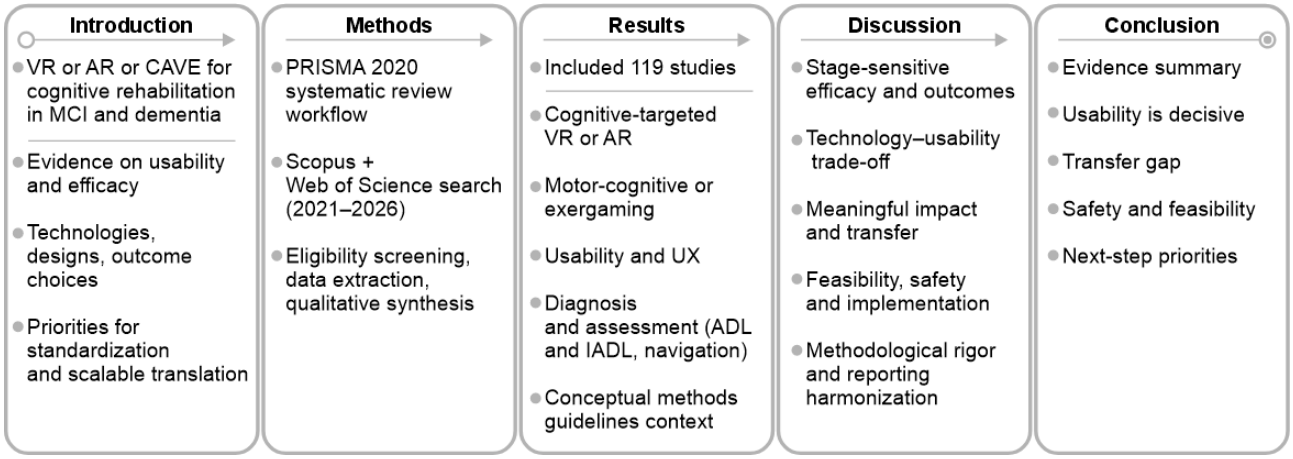

Protocol and Registration

This systematic review was conducted in accordance with the PRISMA 2020 guidelines. The protocol was prospectively developed to ensure methodological rigor and transparency. Trial registration was not applicable, as this work did not involve primary clinical trials.

Eligibility Criteria

Studies were considered eligible if they addressed the use of immersive technologies—specifically VR, AR, or CAVE systems—in the context of cognitive rehabilitation for individuals with MCI and/or dementia. The literature search was limited to publications from 2021 to 2026 and used the following query, "(virtual reality OR augmented reality OR CAVE) AND cognitive AND (dementia OR MCI) AND rehabilitation." Only peer-reviewed journal articles written in English were included. Records without a DOI, missing author information or source title, or those that could not be accessed in full-text form were excluded. Conference papers, abstracts, editorials, and other nonarticle document types were removed. Furthermore, records deemed unrelated to cognitive rehabilitation, MCI, or dementia upon title and abstract screening were excluded. The final set of eligible studies was subsequently divided into 2 categories: that review articles and regular (original research) articles.

Information Sources and Search Strategy

The literature search was conducted using 2 major scientific databases—Web of Science (WoS) and Scopus. These databases were selected due to their broad coverage of peer-reviewed journals in the fields of health care, rehabilitation, neuroscience, and immersive technologies. A structured search strategy was applied in both databases using a consistent combination of keywords and Boolean operators to identify relevant studies. The search focused on titles, abstracts, and author keywords and was designed to capture publications addressing the application of VR, AR, or CAVE systems in cognitive rehabilitation for individuals with MCI and/or dementia. The same search query and inclusion limits were applied across both databases to ensure comparability and reproducibility of the results.

Selection Process

A total of 2 reviewers (RAS and JK) independently screened titles, abstracts, and full texts against predefined criteria. Disagreements were resolved through consensus or consultation with a third reviewer (MP). The selection process was documented using the PRISMA 2020 flow diagram.

XSL•FO
RenderX

Data Extraction

Data were extracted using standardized forms capturing: study design, authorship, publication year, country, participant characteristics, intervention specifics (device type, duration, frequency, and context), cognitive domains targeted, and outcome measures (cognitive tests, usability scales, and emotional or behavioral outcomes). Extraction was performed in duplicate by independent reviewers to minimize errors.

Risk of Bias and Quality Considerations

We performed a structured qualitative appraisal of methodological quality and potential sources of bias during data extraction and narrative synthesis. For each included publication, we recorded study design features and reporting characteristics that most strongly influence interpretability and transferability in immersive-technology research, including (1) study design and the presence or absence of comparator conditions, (2) sample size and participant characterization, (3) intervention or system description and dose (duration, frequency, and setting), (4) completeness and transparency of outcome reporting (including usability or acceptability metrics and adverse events such as cybersickness), (5) follow-up duration, and (6) practical implementation factors (equipment constraints, supervision requirements, staffing, and caregiver involvement). These quality considerations were used to contextualize findings, identify recurring limitations across studies, and guide cautious interpretation of reported outcomes in the narrative synthesis.

Data Synthesis

Given heterogeneity in technologies, intervention content, and outcome measures, a meta-analysis was not feasible. We therefore conducted a narrative synthesis structured by technology type (VR, AR, and CAVE) and outcome domain (usability or acceptability, cognitive performance, daily functioning, and emotional well-being). Within each category, we summarized intervention and system characteristics (eg, hardware setup, immersion level, interaction modalities, delivery setting, and dose where reported) and synthesized reported outcomes qualitatively. For studies that included comparator conditions, we described the direction and consistency of reported between-group findings without statistical pooling. We also synthesized implementation-relevant factors, including onboarding demands, supervision requirements, and adverse events (eg, cybersickness) when reported.

Results

Literature Selection

This section was conducted in accordance with the PRISMA guidelines. Studies were included if they met clearly defined eligibility criteria, that are (1) publication period from 2021 to 2026, (2) English language, (3) peer-reviewed journal articles only, and (4) a clear focus on cognitive rehabilitation in individuals with MCI and/or dementia. Searches were conducted on November 2, 2025. Eligible interventions involved immersive or semi-immersive technologies, specifically VR, AR, or CAVE systems. Records without a DOI, missing author information or source title, or without accessible full-text availability were excluded, as were studies unrelated to cognitive rehabilitation, MCI, or dementia.

Comprehensive electronic searches were conducted in 2 major bibliographic databases, that is, Scopus and WoS. The search strategy used Boolean operators and the following search string, (virtual reality OR augmented reality OR CAVE) AND cognitive AND (dementia OR MCI) AND rehabilitation. The same search strategy and filters were applied consistently across both databases. Retrieved records were exported to reference management software, where duplicates were removed before screening.

Titles and abstracts were screened to exclude clearly irrelevant records, followed by full-text assessment of the remaining articles against the predefined inclusion and exclusion criteria. Only articles that could be downloaded in full text were retained. The final set of included studies was subsequently divided into 2 categories—review articles and regular (original research) articles, to enable structured comparison across study types.

Due to heterogeneity in study designs, intervention characteristics, and outcome measures, a meta-analysis was not feasible. Instead, a narrative synthesis was performed, focusing on the characteristics of the VR, AR, or CAVE interventions, targeted cognitive domains, reported outcomes related to rehabilitation and usability, and identified methodological limitations across studies.

The PRISMA flow diagram (Figure 2) summarizes the study selection process. A total of 236 records were identified from databases (Scopus and WoS), with no records from registers (n=0). Before screening, 57 duplicate records and 9 records marked as ineligible by automation tools were removed, leaving 170 records for screening [31]. No records were excluded at the screening stage (n=0); therefore, 170 reports were sought for retrieval, of which 28 reports could not be retrieved. The remaining 142 reports were assessed for eligibility, and 23 reports were excluded as not related. Ultimately, 119 studies [2-8,13-21,23-27,30-127] were included in the review (119 reports of included studies).

XSL•FO
RenderX

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram illustrating the article selection process.

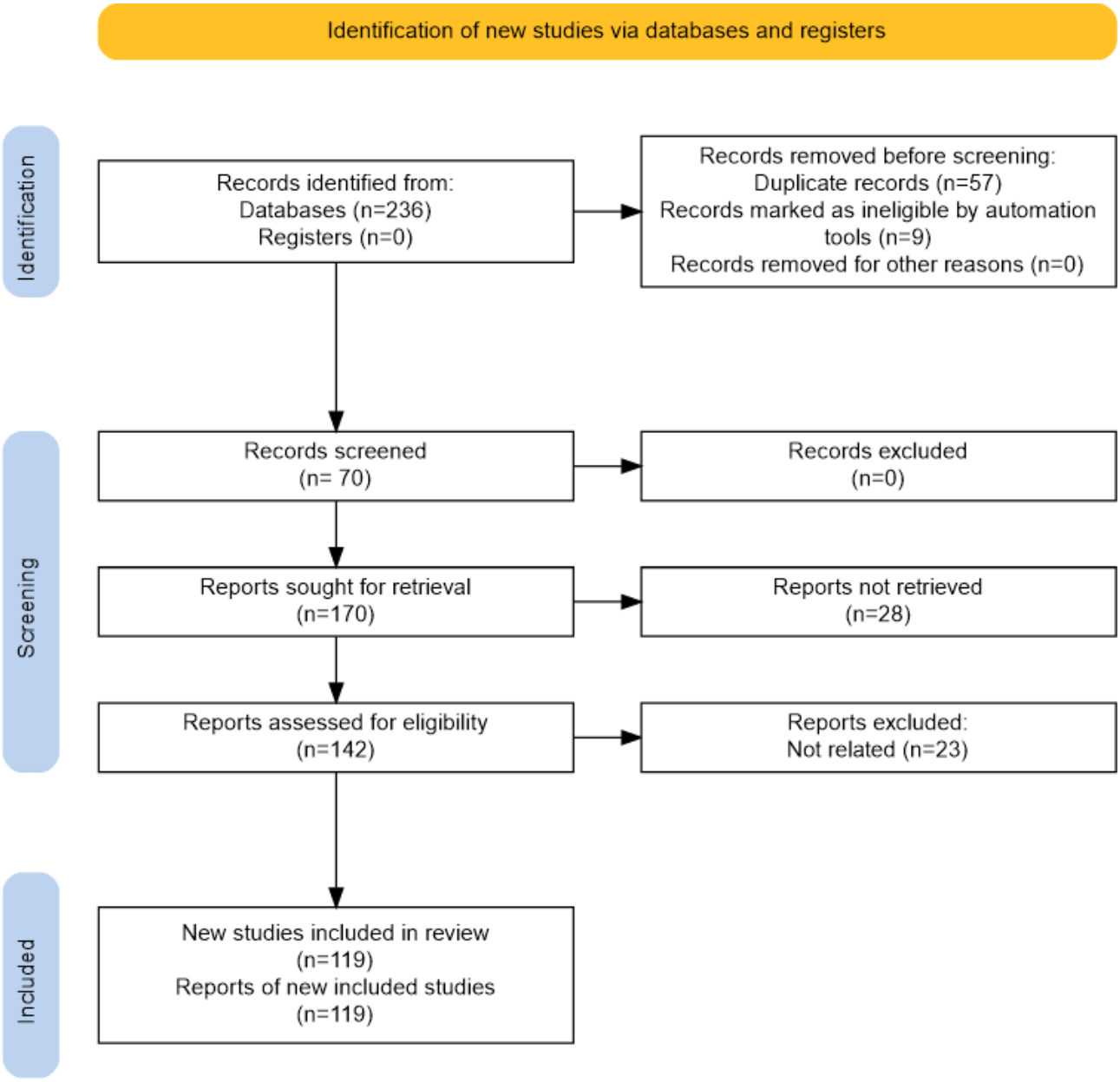

Positioning of This Review Relative to Existing Systematic Reviews

A growing number of systematic reviews have already examined XR-supported cognitive rehabilitation, but their scopes remain fragmented with respect to target population, XR modality, immersion level, and the depth of usability and implementation synthesis. Several focus primarily on MCI and VR-based cognitive rehabilitation, often emphasizing cognitive efficacy and protocol characteristics while providing only limited, nonstandardized coverage of usability, feasibility, and real-world deployment constraints [3,4,20,23-27]. Other reviews address dementia, AD, or functional outcomes (ADL or IADL) and ecological validity, yet typically remain centered on VR (frequently non- or semi-immersive) and treat usability, acceptability, adverse events, and service integration as secondary considerations rather than as analytically coded dimensions [5,6,17]. In parallel, more specialized reviews concentrate on narrower constructs—such as technology-assisted dual-task rehabilitation or spatial navigation paradigms—providing valuable conceptual framing but not offering an implementation-oriented synthesis of immersive therapeutic XR across the MCI-dementia continuum [13,19]. Finally, several systematic reviews adopt broader clinical populations or technology umbrellas (eg, survivors of stroke, Parkinson disease, acquired cognitive disorders, or general digital-technology interventions in older adults), which supports methodological insight but limits direct transferability to dementia-focused cognitive rehabilitation pathways and to cross-modality comparisons including AR and CAVE [7,32-39,128]. A related strand reviews fully immersive VR-based cognitive remediation in adult psychiatric or psychosocial rehabilitation (including some MCI and mild dementia samples), highlighting intervention-component reporting gaps that constrain real-world implementation [40]. In addition, reviews outside the XR rehabilitation core (eg, bibliometric mapping in pediatrics or topic-distant clinical domains) further illustrate how rapidly expanding evidence can

XSL•FO
RenderX

become fragmented and difficult to translate to dementia-specific implementation decisions [41-43].

Beyond cognitive-rehabilitation–centered syntheses, the XR literature also contains systematic reviews that cluster around 2 adjacent themes which inform, but do not substitute for, our focus on immersive XR (VR, AR, or CAVE) cognitive rehabilitation in MCI and dementia.

Reviews in the category "Psychosocial, Emotional, and Behavioral Interventions Using Immersive Technologies" predominantly frame immersive and digital technologies as instruments for psychosocial support, emotional well-being, or broader behavioral outcomes rather than as structured cognitive rehabilitation protocols targeted to the MCI-dementia continuum. For example, one line of work synthesizes XR use in older adults at a general level (including engagement and experience-oriented outcomes) without restricting scope to cognitive rehabilitation in MCI or dementia or to immersive XR modality comparisons [8]. Other reviews adopt an even broader "aging and well-being" perspective, discussing strategies to improve quality of life across normal versus pathological aging, which is informative for contextual motivation but not specific to immersive XR cognitive rehabilitation efficacy or implementation in dementia care pathways [44]. Several reviews examine psychosocially oriented digital interventions and symptom-focused cognitive supports (rather than immersive XR rehabilitation protocols), which limits transferability to our modality-specific and deployment-oriented synthesis in dementia or MCI [45]. A related strand focuses on reminiscence-oriented applications using VR-based approaches, which aligns with person-centered care and engagement, yet differs from our emphasis on cognitive rehabilitation outcomes and cross-technology synthesis across VR, AR, or CAVE [46]. Some reviews concentrate on specific training paradigms or outcome contexts—such as attention rehabilitation via serious games, or changes in cognitive screening performance across training regimes—where dementia-focused immersive XR rehabilitation and implementation or scalability considerations are not the primary analytic axis [47]. Importantly, additional reviews in this group address adjacent rehabilitation or stimulation concepts (eg, multisensory stimulation in dementia rehabilitation) that overlap conceptually but do not provide the XR-modality–differentiated usability and implementation synthesis that our review targets [48]. Finally, reviews focused on digital cognitive interventions in nondementia neurological populations (eg, traumatic brain injury) are methodologically relevant but population-mismatched for the MCI or dementia rehabilitation pathway that defines our eligibility boundaries [49]. Taken together, these help justify the broader psychosocial value and acceptability of immersive or digital interventions, but our review deliberately differentiates itself by restricting the population to MCI and dementia, expanding the technology scope to immersive XR (VR, AR, or CAVE), and systematically integrating usability and real-world implementation or scalability evidence alongside cognitive rehabilitation outcomes.

Reviews in the category "Design, Usability, Safety, and Implementation of Immersive Systems for Older Adults" are closer to our aims in that they foreground design, usability, safety, feasibility, and translational barriers; however, they commonly span heterogeneous populations, modalities, and use cases that dilute direct inference for dementia or MCI cognitive rehabilitation with immersive XR. For instance, a dedicated synthesis of cybersickness and fully immersive VR provides essential safety framing; yet, it is not specifically anchored in dementia or MCI cognitive rehabilitation workflows or in VR, AR, or CAVE cross-modality comparison [50]. Other work reviews XR technologies or interface classes relevant to implementation (eg, smart glasses for cognitive impairment), but it typically emphasizes device feasibility and adoption rather than mapping therapeutic cognitive rehabilitation effects across immersive XR modalities [51]. Several reviews focus on assessment- and function-oriented constructs (eg, IADL assessment) that are highly relevant to outcome selection, while still differing from our rehabilitation-centric scope and our modality-explicit synthesis across immersive XR systems [52]. Additional reviews address intervention families that are broader than immersive XR (eg, digital game-based or computer-based cognitive interventions in cognitively impaired community-dwelling populations), which informs feasibility and design considerations but does not resolve the immersion- and modality-specific trade-offs central to our review [53]. Similarly, comparative or umbrella-type syntheses of nonpharmacologic interventions in older adults provide important taxonomic and implementation context, yet they do not isolate immersive XR (VR, AR, or CAVE) cognitive rehabilitation for MCI or dementia as the analytic core [54]. Bibliometric and mapping-type reviews contribute useful evidence on research trends and gaps, but they do not offer the structured, outcome- and usability-centered comparative synthesis required to guide scalable deployment of immersive XR cognitive rehabilitation [55]. Finally, several reviews in this category focus on other clinical indications or technology ecosystems (eg, poststroke cognitive impairment, Parkinson disease, and sensor or interaction technologies), which provide transferable implementation lessons but remain population- or aim-mismatched to dementia or MCI cognitive rehabilitation [56]. Therefore, while these provide critical inputs for human-centered design, safety, and adoption, our review distinguishes itself by (1) focusing specifically on MCI and dementia, (2) treating immersive XR modality (VR, AR, and CAVE) and immersion level as explicit comparative dimensions, and (3) linking usability or feasibility and implementation or scalability evidence directly to cognitive rehabilitation outcomes and methodological standardization needs in real-world care settings.

Therapeutic VR or AR Interventions Primarily Targeting Cognitive Functions

Table 1 summarizes therapeutic interventions in which the primary target is cognitive functioning, spanning neurodegenerative conditions (AD or dementia and Parkinson disease), MCI, poststroke cognitive impairment, and acquired brain injury, as well as prevention-oriented programs in older adults [57-65]. Across this body of work, interventions aim to improve global cognition and/or specific domains such as executive functions, memory (including working and prospective memory), attention, and visuospatial abilities through structured,

XSL•FO
RenderX

task-based training embedded in virtual or augmented environments [66-70].

XSL•FO
RenderX

Table . Therapeutic VR[a] or AR[b] interventions primarily targeting cognitive functions.

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
Lanzoni et al (2022) [69]	Adults with retrograde amnesia after brain stroke	Usability or design N=NR[c]	VR—Nonimmersive (screen) customizable serious games built from 3D scans of familiar environments for memory rehabilitation	• Dose: Single or few sessions for usability testing • Domains: Spatial memory, episodic memory, orientation • Outcome: Demonstrated feasibility and usability; cognitive outcomes exploratory only
Oliveira et al (2021) [72]	Older adults with mild-to-moderate Alzheimer dementia	Pilot RCT[d] N=17	VR—Immersive (HMD[e] or CAVE[f]) VR cognitive stimulation reproducing IADL[g]	• Dose: 2 months, 10 sessions (2 sessions/week) • Domains: Global cognition, memory, attention, and executive functions • Outcome: Improved global cognition with large effect size in the VR group compared with control
Ip et al (2025) [71]	Community-dwelling older adults with MCI[h]	Pilot RCT N=18	VR exergame—Nonimmersive (large screen) full-immersive VR game-based cognitive-motor training compared with group-based Baduanjin exercise	• Dose: 8 weeks, 16 sessions (2 sessions/week) • Domains: Executive functions, global cognition, and dual-task performance • Outcome: Greater improvements in cognitive-motor performance, mobility, balance, and gait speed in the VR group; no clear benefit in executive function
Huang and Yang (2022) [58]	Older adults with dementia in long-term care	Observational N=20 (7 followed at 3-6 months)	VR—Immersive (HMD) VR reminiscence therapy using immersive scenes related to participants' life history	• Dose: 3 months, 2 sessions/week • Domains: Global cognition (secondary), autobiographical memory, and mood • Outcome: No significant cognitive changes immediately post intervention, but depressive symptoms decreased; cognition declined again at 3-6 months

XSL•FO
RenderX

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
De Simone et al (2023) [60]	Patients with Parkinson disease and MCI; healthy older adults	Protocol N=60 (30 PD-MCI and 30 healthy)	VR+telemedicine—Immersive (HMD) home-based VR cognitive training targeting executive functions to improve prospective memory	• Dose: Planned multi-week intervention (pre- and post–2-month follow-up) • Domains: Prospective memory and executive functions (planning, shifting, and updating) • Outcome: Only baseline and feasibility results available; cognitive efficacy not yet reported
Han et al (2021) [67]	Older adults at risk of cognitive decline (community dwelling)	Prototype and expert evaluation; exploratory effectiveness N=NR	AR—Mobile AR gamified AR cognitive training system inspired by Trail Making Test, performed in real space	• Dose: Short-term pilot (single or few sessions) • Domains: Working memory, attention, and visual perception • Outcome: Experts judged that the system could improve working memory and attention; effectiveness in older adults not yet fully tested
De Luca et al (2024) [62]	Adults with severe acquired brain injury (traumatic and nontraumatic)	RCT N=40 (20 VR-based ROT[i], 20 control)	VR—Nonimmersive (screen) VR-based ROT compared with standard reality orientation therapy	• Dose: Inpatient rehabilitation cycle (several weeks) • Domains: Global cognition (MMSE[j]), orientation, and attention • Outcome: Both groups improved, but VR-ROT showed greater MMSE gains and reduced depressive symptoms
Buele et al (2024) [66]	Older adults with MCI	Single-blind RCT N=34 (17 VR, 17 control)	VR—Semi or full immersive group motor training followed by VR-based cognitive training vs motor training plus traditional cognitive training	• Dose: 6 weeks, 12 sessions (40 minutes each) • Domains: Global cognition, executive functions, and memory • Outcome: Both groups improved in cognition and depression; between-group differences were not significant, but VR group reached higher task difficulty
Downey et al (2023) [64]	Middle-aged and older adults with and without hearing loss	Single-blind RCT N=NR	CCT[k]+VR assessment—Immersive (assessment) At-home computerized executive-function training vs wait-list control, with dual-task gait	

XSL•FO
RenderX

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
				• Dose: 12 weeks of at-home training • Domains: Executive functions, attention, and dual-task performance • Outcome: Planned to test whether executive function training improves cognitive-motor dual-task performance; primary outcomes not fully reported yet
Liu et al (2023) [61]	Older adults with poststroke MCI	Pilot RCT N=30 (15 VR, 15 control)	VR—Immersive (HMD) immersive VR puzzle games added to routine rehabilitation vs traditional cognitive training	• Dose: 6 weeks of treatment • Domains: Executive function, visuospatial attention, and processing speed • Outcome: Significant improvement in DSST in VR group, suggesting gains in executive and visuospatial functions; other measures improved but not sign…
Oliveira et al (2021) [72]	Institutionalized nonrobust older adults	Single-blind RCT N=NR	VR exergame—Nonimmersive (exergame) 2D VR exercise training vs equivalent exercises without VR	• Dose: 2 months, 16 sessions (2 sessions/week) • Domains: Spatial navigation and executive function • Outcome: VR group showed significantly greater improvement in spatial navigation compared with active control
Jeong et al (2023) [68]	Adults with acquired brain injury	Randomized crossover trial N=NR	VR (music)—Immersive (HMD) VR-based music attention training (drumming tasks) vs conventional computerized cognitive training	• Dose: Two 4-week phases (VR vs control) in crossover • Domains: Selective, alternating, and divided attention; executive function • Outcome: VR-MAT improved attention and some executive measures more than conventional training in at least 1 phase
Maeng et al (2021) [70]	Older adults with and without MCI	RCT N=NR	VR—Nonimmersive VR-based cognitive training tasks (eg, ADL[1] scenarios and puzzles) compared with conventional or no training	

XSL•FO
RenderX

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
				• Dose: Multiweek program with repeated sessions • Domains: Global cognition, memory, executive functions, and attention • Outcome: Improvement or stabilization of cognition in older adult participants, particularly those with MCI
Baldimtsi et al (2023) [63]	Older adults with MCI	Randomized intervention trial N=NR	VR (physical+cognitive)—Nonimmersive (likely) combined physical exercise and VR-based cognitive tasks vs conventional or usual care	• Dose: Several weeks with multiple sessions per week • Domains: Executive functions, memory, and attention • Outcome: VR physical+cognitive training associated with greater cognitive gains than comparison conditions
Fiorenzato et al (2025) [59]	Patients with Parkinson disease and healthy older adults	RCT N=NR	VR—Nonimmersive VR-based cognitive training tasks targeting prospective memory and executive functions	• Dose: Multisession intervention over several weeks • Domains: Executive functions and prospective memory • Outcome: Improvements in prospective memory and executive functions in both PD and healthy aging groups
Zhu et al (2022) [65]	Adult patients undergoing laparoscopic surgery	Protocol N=NR	VR—Immersive (HMD) biophilic VR exposure pre- and/or postoperatively compared with sham VR	• Dose: Short perioperative VR sessions • Domains: Postoperative cognitive function (attention, working memory, and global cognition) • Outcome: No results yet; hypothesis that biophilic VR will mitigate postoperative cognitive decline
Początek et al (2023) [73]	Patients with Alzheimer disease	Narrative or systematic review N=NR	Mixed (including VR or AR)—not applicable, overview of cognitive training approaches, including emerging VR-based methods	• Dose: Various across included trials • Domains: Memory, executive functions, language, and attention • Outcome: Cognitive training yields small-to-moderate benefits; VR or AR highlighted as promising but underresearched
Cavallo and Lasaponara (2025) [74]		Narrative review N=NR		

XSL•FO
RenderX

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
	Patients with various neurodegenerative diseases (eg, Alzheimer disease and Parkinson disease)		Mixed (including VR or AR)—not applicable, overview of cognitive deficits and cognitive training modalities in neurodegenerative diseases	• Dose: Various across discussed studies • Domains: Multiple domains (memory, executive functions, attention, and visuospatial abilities) • Outcome: Supports use of structured cognitive training; notes need for more high-quality RCTs and the potential of VR or AR

[a]VR: virtual reality.

[b]AR: augmented reality.

[c]NR: not reported.

[d]RCT: randomized controlled trial.

[e]HMD: head-mounted display.

[f]CAVE: Cave Automatic Virtual Environment.

[g]IADL: instrumental activities of daily living.

[h]MCI: mild cognitive impairment.

[i]ROT: reality orientation therapy.

[j]MMSE: Mini-Mental State Examination.

[k]CCT: computerized cognitive training.

[l]ADL: activities of daily living.

Most studies prioritize executive functioning and attentional control as core therapeutic targets, reflecting their relevance for everyday functioning and their frequent impairment in clinical aging and neurological populations [59,60,62,68]. Memory-focused approaches include working memory training (notably via mobile AR) and prospective memory training in Parkinson disease and healthy aging, with several programs attempting to increase ecological validity by using realistic or meaningful task contexts [57,59,67].

In poststroke and acquired brain injury contexts, VR is used both for domain-specific cognitive rehabilitation (eg, visuospatial attention and executive processes) and for broader orientation and behavioral support, often as an adjunct to standard rehabilitation [61,62].

The subgroup includes immersive head-mounted display (HMD)–based VR (eg, reminiscence and cognitive training), nonimmersive or screen-based VR platforms, and mobile AR serious games [58,61,67,69,70]. A notable direction is the integration of immersive VR with telemedicine or home-based delivery concepts intended to improve accessibility and scalability for older adults and patients with mobility constraints [60]. In contrast, several studies remain clinic- or laboratory-based and emphasize feasibility or usability, platform design, and therapist-friendly monitoring rather than definitive efficacy [67,69,70]. Home-oriented delivery is also reflected in at-home computerized executive-function training designs that use immersive VR environments as ecologically valid contexts for dual-task assessment of cognition and mobility, extending evaluation beyond standard clinic-based testing [64].

Methodological rigor varies substantially, ranging from pilot and single-blind randomized controlled trials to crossover designs and observational follow-up studies [57,58,66,68]. Control conditions include traditional cognitive training, standard rehabilitation or orientation therapy, and non-VR exercise programs, which help isolate VR or AR-specific contributions but also introduce heterogeneity in active ingredients and “dose” [61,62,66,71]. Outcome batteries typically combine global cognitive screening tools with domain-specific neuropsychological tests (executive function, attention, and memory), and—less consistently—functional end points (ADL or IADL), mood, or user experience (UX) measures [57,58,62,66]. Of the total, 2 included articles provide a broader conceptual and methodological context for cognitive training in neurodegenerative disease, highlighting persistent challenges around outcome heterogeneity, transfer to daily functioning, and evidence quality [73,74].

Across populations, many studies report favorable feasibility indicators (acceptability, adherence, and tolerability) and within-group improvements on at least some cognitive measures, particularly when interventions are engaging, adaptive, and anchored in meaningful tasks [61,66,68,70]. However, between-group superiority is not uniform, and several studies are explicitly positioned as pilots or preliminary reports, limiting statistical power and confidence in effect estimates [57,58,60,61]. Reminiscence-oriented immersive VR appears especially relevant for affective and engagement-related outcomes in dementia, with cognitive effects that may be modest and time-limited without maintained intervention [58]. For MCI, combined or multicomponent programs (eg, motor plus

XSL•FO
RenderX

VR-based cognitive training) reflect a practical clinical strategy, although gains in daily functioning are not consistently observed [66,71].

Overall, the subgroup supports the feasibility of delivering cognitively oriented interventions using VR or AR across diverse older and neurological populations, but comparability is constrained by heterogeneity in immersion level, task design, intervention dose, and outcome selection [57,59,67,70]. To strengthen future quantitative synthesis, priorities include (1) harmonized core outcome sets (global cognition+domain-specific measures+functional end points), (2) standardized reporting of intervention dose and fidelity, (3) longer follow-up to assess durability, and (4) clearer mapping between task mechanics and hypothesized transfer targets [73,74]. Finally, protocol-driven work targeting cognitive vulnerability in perioperative settings illustrates ongoing diversification of VR applications, but such studies should be clearly separated from dementia or MCI rehabilitation pathways when interpreting population-specific clinical relevance and implementation needs [65].

Motor-Cognitive and/or Exergaming VR Interventions in Neurorehabilitation

Table 2 summarizes studies in which VR-based approaches combine explicit cognitive demands with concurrent motor engagement (eg, stepping, balance training, cycling, or exergaming) and are applied in neurorehabilitation-relevant contexts. The included evidence spans community-dwelling older adults with cognitive frailty [75], individuals with MCI [76,77], mild AD (protocol) [78], and poststroke cognitive disorders [79,80]. Overall, delivery most often relies on nonimmersive or semi-immersive setups, typically screen-based interaction combined with motion tracking or instrumented exercise equipment [81].

XSL•FO
RenderX

Table . Motor-cognitive and/or exergaming VR[a] interventions in neurorehabilitation.

Study (Year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
Kwan et al (2024) [75]	Community-dwelling older adults with cognitive frailty (physical frailty+MCI[b]), no dementia	Multicenter, assessor-blinded parallel-group RCT[c] (VRMCT[d] vs usual care); N=293 (≈146 VRMCT, 147 control)	Nonimmersive VR motor-cognitive exergame platform with motion sensing; nonimmersive group-based VR motor-cognitive exercises; 16 one-hour sessions led by interventionists	16×60 min; 2 times per week; 8 weeks Domains: Global cognition, executive function, memory, interference control VRMCT significantly improved global cognitive function vs usual care; some benefits in executive measures
Fu and Wang (2025) [76]	Patients with MCI treated in a neurology department	Randomized controlled clinical observation (VR+acupuncture vs VR only); N=48 (24 VR+acupuncture; 24 VR only)	VR rehabilitation training system with virtual scenes plus acupuncture in treatment arm; Non-immersive; Control: VR training only; Treatment: same VR program plus scalp/eye acupuncture during each session	VR training protocol described qualitatively; acupuncture needles retained ~30 min each session Domains: Global cognition (MMSE[e], MoCA[f]), daily living skills Treatment group showed larger gains in MMSE and MoCA and higher overall efficacy vs VR-only control
De Luca et al (2022) [82]	Single patient with Nasu–Hakola disease (TREM2-related dementia) and motor-cognitive deficits	Single-case crossover (conventional rehabilitation phase vs conventional + VRRS[g] phase); N=1	VRRS with cognitive and physical modules; Nonimmersive Phase 1: standard physical + cognitive training; Phase 2: 30 speech + 30 PT sessions plus 60 VRRS sessions (cognitive + motor)	VRRS phase: 3 sessions/week, 60 min; total 60 VRRS sessions after 12 weeks of conventional rehab. Domains: Attention, memory, executive function; affective symptoms Improved cognitive tests and marked reduction in anxiety, apathy, indifference, and depressive symptoms after VRRS phase
Huber et al (2024) [80]	Adults in chronic phase after stroke with residual motor or cognitive deficits	Single-blind parallel-group RCT protocol (exergame + usual care vs usual care); Target N=38 chronic survivors of stroke	Concept-guided personalized motor-cognitive exergame training using video game–based tasks; Nonimmersive Intervention group performs exergame training in addition to usual care; control continues usual care only	2 times per week; 30-40 min; 12 weeks (24 sessions) Domains: Global cognition (MoCA) and domain-specific cognitive functions Results pending; expected superior improvement in cognition and gait in exergame group
Faria et al (2023) [83]	Adults with acquired brain injury, dementia, or age-related cognitive decline (conceptual target)	Methods or architecture paper (no clinical trial); Not applicable	VR-based activities-of-daily-living simulations integrated with AI-driven user profiling and adaptation; Non- or semi-immersive VR ADL environments Framework where VR ADL tasks are adapted in real time using AI models to personalize neurorehabilitation	Not fixed; example scenarios only (conceptual) Domains: Executive function, attention, memory, functional cognition No empirical outcomes; framework is intended to support individualized improvement in functional cognition

XSL•FO
RenderX

Study (Year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
Manenti et al (2024) [77]	Adults with MCI from multiple Italian centers	Multicenter randomized active-controlled trial with five treatment arms; N=109 participants with MCI allocated to 5 groups	Clinic- and home-based VRRS cognitive training combined with anodal or sham tDCS[h], or TAU; Nonimmersive VR / telerehabilitation All active arms: 12 × 60-min face-to-face VRRS sessions; some arms also receive anodal tDCS and/or 12 weeks of home VRRS telerehabilitation or unstructured activity	Clinic VRRS: 12 × 60 min over 4 weeks; plus arm-specific telerehabilitation or unstructured… Domains: Episodic memory (primary), plus global cognition and other… VRRS + anodal tDCS followed by VRRS telerehabilitation produced the largest and most durable gains in episodic memory; VRRS-based protocols outperformed TAU
Maggio et al (2025) [84]	Informal caregivers of people with MCI	Nonrandomized single-arm pilot pre-post study; N=10 caregivers	Semi-immersive VR-based physical training program (K-HERO); Semi-immersive Individual low-intensity VR exercise sessions performed while accompanying relatives to rehabilitation	6 sessions of 30–40 min each during caregivers' clinic visits Domains: No direct cognitive training; outcomes focus on psychological well-… Not applicable; reported reductions in avoidance and social-support–based coping and good usability
Hwang et al (2021) [81]	Community-dwelling older adults with age-related cognitive and gait concerns	Randomized controlled trial (VRCT+locomotor vs tabletop cognitive training); N=18 (9 VRCT[i]+locomotor, 9 control)	Semi-immersive VR cognitive training system combined with locomotor tasks; Semi-immersive experimental group performs VR cognitive tasks while stepping or locomoting; control receives tabletop cognitive activities with the same schedule	30 min/day, 3 times per week for 6 weeks (18 sessions) Domains: Processing speed, attention, working memory, and executive function VR group showed greater improvements in executive and attention measures (eg, TMT-A, Digit Span backward) vs control
Contrada et al (2025) [79]	Adults in the chronic phase after ischemic stroke with cognitive impairment	Within-subject longitudinal study (single group; pre, post, 6-month follow-up); N=18	Home-based multidomain cognitive training via VRRS Tele-NeuroRehabilitation platform; Nonimmersive VR or telerehabilitation Patients perform therapist-prescribed VRRS cognitive exercises at home with remote monitoring and feedback	5 sessions/week, 60 min each, for 4 weeks (20 sessions) Domains: Logical reasoning, attention, executive function, memory, language Significant improvements in working memory and language abilities and reduction in depressive symptoms, maintained at 6 months
Bourrelier et al (2021) [85]	Frail older adults with cognitive impairment (MMSE 15–28) and slow gait (<0.65 m/s)	Randomized controlled trial (VR balance training vs Nintendo Wii); N=22 analyzed (from 37 randomized)	VR balance-training system using game-like upper-limb and postural tasks; Likely semi-immersive (screen-based) Ten VR balance-training sessions vs ten Wii-bowling sessions over 3 weeks; APAs assessed using EMG	10 sessions in 3 weeks (~3-4 sessions/week); session length not clearly reported Domains: Not primarily cognitive; cognitive engagement through interactive… Main effects on postural adjustments; VR training improved APA timing and functional reach more than Wii

XSL•FO
RenderX

Study (Year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
Gambella et al (2022) [78]	Community-dwelling adults with mild Alzheimer disease living at home with caregiver	Randomized controlled trial protocol (VR cycling + CCT[j] vs standard bike + CCT); Planned N=78 (39 VR+CCT; 39 control)	jDome VR cycling system combined with computerized cognitive training; Semi-immersive VR (projected dome environment) Both groups receive computerized cognitive training; experimental group cycles in jDome VR, control uses standard exercise bike	16 × 60-min sessions; 2×/week over 8 weeks with 3-month follow-up Domains: Global cognition (MMSE, ADAS-Cog) plus quality of life, mood, behavior Outcomes not yet reported; aim is stabilization or improvement of global cognition in mild AD
Tang et al (2025) [86]	Adults from the general public with no dementia or caregiving history (empathy training target)	Controlled pre-post study (VR dementia tour vs control); N=80 adults (visual display terminal vs control)	Fully immersive VR dementia tour using interactive narrative persona; Fully immersive Single VR session simulating sensory and cognitive impairments of dementia; compared with non-VR control condition	Single VR session (duration not clearly specified) Domains: Not rehabilitation of patients; focuses on dementia knowledge and… VDT participants showed greater improvements in dementia knowledge and empathy compared with control

[a]VR: virtual reality.

[b]MCI: mild cognitive impairment.

[c]RCT: randomized controlled trial.

[d]VRMCT: virtual reality motor-cognitive training.

[e]MMSE: Mini-Mental State Examination.

[f]MoCA: Montreal Cognitive Assessment.

[g]VRRS: Virtual Reality Rehabilitation System.

[h]tDCS: transcranial direct current stimulation.

[i]VRCT: virtual reality cognitive training.

[j]CCT: computerized cognitive training.

Across intervention studies, the motor-cognitive rationale is commonly framed as training attention, working memory, and executive control under dual-task or enriched sensorimotor conditions. In a large multicenter randomized controlled trial among older adults with cognitive frailty, structured VR motor-cognitive training yielded improvements in global cognition compared with usual care [75]. Smaller trials in community-dwelling older adults further suggest that coupling cognitive tasks with locomotor activity can produce concurrent gains in cognitive performance and mobility-related measures (eg, walking speed), supporting the premise that integrated training may facilitate transfer to functional capacities [81]. Balance-oriented VR interventions target motor control processes that are sensitive to cognitive status; for example, VR-based training improved anticipatory postural adjustment timing relative to an active exergaming comparator (Nintendo Wii), indicating potential benefits for neuromotor preparation in frail older adults with cognitive and motor deficits [85].

A second theme concerns scalability, personalization, and continuity of care. Home-based telerehabilitation delivered via VR rehabilitation platforms with remote monitoring enables high-frequency practice and may support maintenance of gains in chronic poststroke populations [79]. Several contributions in this group are protocols rather than completed efficacy trials, reflecting ongoing consolidation of feasibility and methodological standards. The Physical Exercise and Mobile Cognitive Stimulation protocol, for instance, formalizes a concept-guided personalized motor-cognitive exergame program for chronic stroke, with global cognition (Montreal Cognitive Assessment [MoCA]) as a primary end point and gait or dual-task metrics as complementary outcomes [80]. In AD, the jDome protocol integrates VR-supported cycling with concurrent cognitive training to test whether adding a VR exercise context enhances outcomes beyond conventional exercise plus cognitive training [78]. Complementing these clinical studies, an AI-driven methodology paper proposes adaptive profiling and real-time task adjustment within VR-based ADLs scenarios to address heterogeneity in impairment severity and learning rates [83].

Some studies combine VR training with adjunct modalities, which may be clinically pragmatic but complicate causal attribution. In MCI, a multicenter randomized active-controlled study reported long-lasting improvements in episodic memory when VR-based cognitive treatment was combined with transcranial direct current stimulation and followed by telerehabilitation, suggesting that technology-enabled continuation at home may contribute to long-term effects [77]. Another clinical observation compared VR training alone with VR combined with acupuncture and reported larger improvements on screening measures (Mini-Mental State Examination [MMSE] or MoCA) in the combined arm [76]. In a rare neurodegenerative disease, a case report described a multimodal program incorporating VR rehabilitation and reported improvements across cognitive and neuropsychiatric

XSL•FO
RenderX

measures, although the single-case design and concurrent therapies limit generalizability [82].

Several included publications provide implementation-relevant context rather than direct efficacy evidence. A cross-sectional analysis highlights the co-occurrence of cognitive impairment and psychological morbidity among survivors of stroke in rehabilitation, underscoring the need for scalable cognitive-motor rehabilitation approaches [87]. An opinion article summarizes recent VR developments for poststroke cognitive impairment and discusses practical considerations and limitations of the current evidence [88]. A national clinician survey similarly characterizes adoption, perceived value, and barriers to technology-enabled rehabilitation (including VR and telerehabilitation), informing feasibility and scalability considerations [89]. Beyond patient rehabilitation, an immersive "dementia tour" design study illustrates broader VR content development around dementia and highlights usability and engagement considerations that may be transferable to therapeutic design [86]. In the wider care ecosystem, a pilot study delivering low-intensity VR exercise to caregivers emphasized acceptability and potential psychosocial benefits, although it did not directly address patient cognitive outcomes [84].

In summary, motor-cognitive and exergaming VR interventions show promising signals across cognition- and mobility-relevant outcomes, but the evidence remains heterogeneous in platforms, immersion, intervention dose, and outcome selection [81,85]. The strongest inferences arise from larger randomized designs [75,77], whereas smaller trials, single-group telerehabilitation studies, and protocols primarily support feasibility and hypothesis generation [78-80]. Future work would benefit from harmonized cognitive end points beyond screening tools, explicit reporting of motor-cognitive task composition and progression, and longer follow-up to test durability and transfer to real-world functioning.

Usability, Acceptability, and UX of VR or AR Systems

Across the included studies, usability and acceptability (Table 3) were treated as primary feasibility outcomes rather than secondary implementation considerations, reflecting the field's current emphasis on early-stage validation of immersive and semi-immersive interventions. Evidence spans community-dwelling older adults and mixed-ability samples [90-92], clinical groups with MCI [18,93-98], AD [99], Parkinson disease [100], and dementia care contexts incorporating carers and professionals [101,102]. Collectively, these studies show that VR-based approaches are typically perceived as engaging and feasible when sessions are structured, onboarding is provided, and interaction demands are calibrated to users' capabilities [18,90,91,93,94].

XSL•FO
RenderX

Table . Usability, acceptability, and UX[h] of VR[b] or AR[c] systems.

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose or key cognitive outcomes
Chau et al (2021) [90]	Community-dwelling older adults and people with various disabilities; mixed physical and cognitive abilities	Single-arm pretest–posttest feasibility study; N=NR[d]	Non-immersive motion-tracking VR games targeting upper or lower limb movements, cognitive tasks, community-living skills, and relaxing scenes (exergaming)	Approximately 30-min sessions, 3 times per week for 6 weeks in community or service centers Physical and cognitive function, mood, quality of life, usage statistics, adverse events Preliminary improvements in some physical or cognitive outcomes; high acceptability and good adherence; few adverse events → VR training feasible for heterogeneous older and disabled
Chuang et al (2025) [103]	Community-dwelling older adults, mainly without diagnosed dementia but with age-related cognitive concerns	Single-arm pretest–posttest pilot study; N=NR	Fully immersive HMD[e]-based VR leisure and cognitive games; standing and seated interaction with light exergaming elements	Multiple VR sessions over several weeks in laboratory or clinic setting Cognitive performance (attention, memory, and executive function), self-reported enjoyment and usability, feasible and safe; participants enjoyed the intervention; some mild cybersickness or discomfort; preliminary cognitive improvements but no definitive efficacy conclusions
Kim et al (2025) [91]	Community-dwelling older adults, mostly cognitively intact or with mild decline	Single-arm feasibility and acceptability study; N=NR	Immersive HMD VR puzzle tasks requiring visuospatial problem-solving and manual interaction; low-to-moderate motor demands	One or several short VR sessions in a controlled environment Usability ratings, enjoyment, motion sickness, basic cognitive/engagement measures VR puzzles acceptable and engaging; older adults could use the system with guidance; main issues were headset comfort and controller complexity
Hassandra et al (2021) [93]	Older adults with MCI[f]	Mixed methods feasibility study (quantitative plus qualitative); N=NR	Semi-immersive VR exergame combining cycling or stepping with cognitive tasks (dual-task motor-cognitive training)	Repeated training sessions over several weeks in outpatient or rehabilitation setting Physical performance, cognitive outcomes, usability metrics, qualitative UX motor-cognitive exergaming feasible and acceptable for MCI; participants reported enjoyment and perceived benefit; some technical or usability issues;

XSL•FO
RenderX

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose or key cognitive outcomes
Kua et al (2025) [94]	Older adult participants with MCI	Feasibility randomized pilot trial (VR vs control or alternative training); N=NR	Non- or semi-immersive VR tasks simulating everyday activities with cognitive focus and limited motor interaction	Multisession program over several weeks in a clinical setting; neuropsychological tests (memory, attention, and executive function) and feasibility indicators VR group showed at least comparable and, in some domains, better cognitive outcomes than control; acceptable adherence and dropout; mild fatigue or discomfort in some participants
Perra et al (2024) [96]	Older adults with MCI	Feasibility randomized controlled trial – study protocol; N=NR	VR-based cognitive remediation platform with everyday-like tasks and motor interaction (navigation and object manipulation)	Planned multiweek VR program with repeated sessions in clinical or outpatient setting Planned: cognitive function, daily functioning, quality of life, usability, adherence Protocol outlines rationale, design, and feasibility end points; emphasizes user-centered design, safety monitoring, and adherence for older adults with MCI
Latella et al (2024) [18]	Patients diagnosed with MCI	Single-arm feasibility and usability study; N=NR	Nonimmersive VR platform simulating everyday environments (eg, shopping and navigation) to train ecological cognitive skills; limited motor components	Series of VR sessions over several weeks in clinic or rehabilitation center Ecological cognitive performance, standardized cognitive tests, usability, and satisfaction Most participants with MCI could use VESPA[g] 2; tasks perceived as ecologically valid; some technical difficulties and need for therapist guidance; early signals of improvement in ecolog…
Maggio et al (2025) [100]	Patients with mild-to-moderate Parkinson disease, with cognitive involvement	Pilot randomized controlled trial (VR telerehabilitation vs control or standard care); N=NR	Home-based VR telerehabilitation platform for cognitive tasks with simple motor interaction	Scheduled remote VR sessions at home over several weeks Cognitive outcomes, Parkinson-specific measures, quality of life, feasibility, and caregiver-related measures Telerehabilitation via VR was feasible and well-accepted; adherence adequate; cognitive outcomes showed promising trends; remote delivery reduced travel burden

XSL•FO
RenderX

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose or key cognitive outcomes
Jeong et al (2025) [99]	Patients with mild-to-moderate Alzheimer disease	Single-arm preliminary feasibility study; N=NR	Nonimmersive VR cognitive training tasks (eg, navigation and memory) with simple motor interaction	Short multisession VR program in clinical or day-care setting Cognitive test scores, behavioral responses, usability indicators, adverse events VR training was feasible for selected patients with Alzheimer disease with adequate support; some had difficulties with navigation and task understanding; no serious adverse events;
Mondellini et al (2022) [95]	Older adults with MCI from Estonia and Italy	Cross-sectional comparative study between countries; N=NR	Immersive HMD VR cognitive task with mild motor demands (head or hand movements)	Single VR session in laboratory or clinic setting UX, presence, comfort, motion sickness, basic task performance Both groups completed the task; cultural and individual differences in UX and motion sickness; overall feasibility acceptable with supervision
Munoz et al (2025) [92]	Older adults with varying cognitive status using VR games	Observational data-driven study using classifier models; N=NR	VR exergame or cognitive game platform with motion-tracked head movements; focus on kinematic data rather than training efficacy	Single or repeated gaming sessions in controlled environment Head kinematics, subjective game UX, cognitive test scores Classifiers could link kinematic patterns to cognitive status and UX; supports feasibility of VR sensor data as digital biomarkers;
Stasolla and Di Gioia (2023) [97]	Individuals with mild neurocognitive impairment and their caregivers	Studying Usability and acceptability (conceptual and small sample); N=NR	Early-stage VR system where reinforcement learning adapts cognitive training parameters; some motor interaction	Limited number of sessions or system walkthroughs in clinic or laboratory Usability ratings, satisfaction, perceived usefulness, caregiver feedback Patients and caregivers generally accepted the RL-VR concept; main issues were technical or interface-related
Matsangidou et al (2023) [102]	People living with dementia in care settings	Participatory design study with pilot evaluation; N=NR	VR system for physical rehabilitation (movement tasks or games) with cognitive engagement; non- or semi-immersive	Iterative co-design workshops and pilot VR sessions in care facilities Qualitative feedback from patients, staff, caregivers; observations of engagement and physical activity Co-designed VR rehab matched abilities and preferences; participants could engage in guided VR tasks; staff noted motivational benefits and practical challenges

XSL•FO
RenderX

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose or key cognitive outcomes
Gokani et al (2025) [101]	Informal carers and health or social care professionals working with people with dementia	Qualitative study using interviews or focus groups; N=NR	Various VR scenarios (often immersive or 360°) used for stimulation, reminiscence, and engagement	Discussion plus limited VR demonstrations: no fixed dose Perceived usefulness, risks, ethical issues, feasibility in routine care VR seen as promising for engagement and reminiscence; concerns about cost, motion sickness, staff training, and integration into routine care;
Chen et al (2025) [104]	Community-dwelling and/or institutionalized older adults, mixed cognitive status	Qualitative or mixed methods user-experience study; N=NR	Immersive VR experiences (sightseeing, nature, and simple interactive content) with low physical intensity	One or a few VR sessions in laboratory, community, or residential settings UX, presence, enjoyment, discomfort, perceived value Most older adults were curious and engaged; VR is generally acceptable, though some anxiety, disorientation, or cybersickness; stresses gradual introduction and ergonomic adjustments
Siette et al (2024) [105]	Adults and/or older adults using the LEAF CAF VR module	Think-aloud qualitative usability study; N=NR	VR training or assessment module with cognitive and functional tasks; interaction via controllers or gestures	Single-session interaction with the module in a laboratory Usability problems, interaction patterns, perceived workload, user feedback Identified specific usability barriers (interface complexity, confusing feedback, and navigation issues) and facilitators
Pedroli et al (2022) [106]	Adults or older adults with cognitive impairment using home-based training	Usability and acceptability pilot study; N=NR	Nonimmersive mobile or tablet application with structured cognitive exercises, possibly using enriched media but no full VR HMD	Home-based use over several weeks; self-administered with minimal or remote supervision Task completion, app usage, usability questionnaires, satisfaction Home-based app generally usable; adherence depended on support and motivation; participants appreciated flexibility; technical literacy and device issues were barriers
Tseng and Giau (2022) [98]	Older adults evaluating potential living environments	Feasibility and usability study; N=NR	Non- or semi-immersive VR walkthroughs of housing or environmental designs; focus on navigation and environmental fit	One or a few VR walkthrough sessions in laboratory or facility Perceived safety and usability, environmental preferences, motion sickness, orientation VR pre-occupancy evaluation was feasible; older adults could meaningfully comment on design features; some required support with navigation and controls

XSL•FO
RenderX

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose or key cognitive outcomes
Ghosh et al (2022) [107]	Older adults with MCI participating in paired or social activities	Iterative user-centered design study with small pilots; N=NR	Robot-mediated interaction platform (non-VR) supporting cognitive and social activities with light motor components	Multiple design–test cycles with sessions in laboratory or care setting Usability, acceptability, engagement, perceived social and cognitive benefits Older adults with MCI could engage with robot-mediated paired tasks; emphasizes gradual complexity, clear feedback, and social aspects
Haley et al (2021) [108]	Adults with acquired apraxia of speech	Feasibility and therapeutic-effect pre-post study (sometimes with small comparison group); N=NR	Non-VR tablet or computer-based speech therapy program with autonomy-supportive features; motor speech rehabilitation	Multiple therapy sessions over several weeks in clinical setting Speech production measures, communication outcomes, adherence, and user satisfaction Feasible with positive therapeutic effects on speech measures; autonomy-supportive design enhanced engagement and motivation

[a]UX: user experience.

[b]VR: virtual reality.

[c]AR: augmented reality.

[d]NR: not reported.

[e]HMD: head-mounted display.

[f]MCI: mild cognitive impairment.

[g]VESPA: Virtual Environment for Spatial Processing Assessment.

Quantitative usability or acceptance assessments commonly relied on standardized questionnaires (most frequently the System Usability Scale and related acceptance models), with generally favorable perceptions reported by older end users in both community and clinical settings [91,93,94,103]. Importantly, perspectives sometimes diverged between stakeholders. For example, older adults tended to report high enjoyment and ease of use in fully immersive, leisure-based cognitive training, while professional raters judged overall usability more conservatively [103]. In technology-driven cognitive rehabilitation platforms, feasibility was often demonstrated through completion or adherence alongside positive usability ratings, supporting the practicality of repeated-session protocols in outpatient or community environments [18,91,93,94]. At the same time, measurement approaches were not always comprehensive; the Virtual Environment for Spatial Processing Assessment 2.0 feasibility work explicitly cautioned that relying on a single usability scale is insufficient to capture the multidimensional nature of feasibility, usability, and UX in complex VR rehabilitation systems [18]. This aligns with the broader need to report usability, tolerability, workload, and adverse events in parallel, particularly for populations at risk of fatigue, dizziness, or anxiety during immersion [93,99,100].

Qualitative and mixed methods evidence further clarifies what drives (or undermines) positive UX. Think-aloud evaluation of a VR module identified concrete interaction and interface-level issues that can influence comprehension, confidence, and task flow, underscoring the value of formative testing before efficacy trials [105]. Broader experiential accounts emphasized that perceived realism, meaningfulness of activities, and alignment with older adults' routines shape engagement and perceived usefulness [92,103,104]. Cross-cultural comparison during an immersive VR cognitive task highlighted that UX is not solely determined by hardware or task difficulty, but also by contextual factors that may influence comfort and interpretation of the virtual environment [95]. Complementing self-report, head-kinematics–based modeling demonstrated that behavioral signals during VR gameplay can be leveraged to characterize UX and its associations with cognitive factors, suggesting a pathway toward more objective, continuous UX monitoring in future systems [98].

Several studies explicitly adopted human-centered or participatory design strategies, reinforcing that usability in older and cognitively impaired populations benefits from iterative refinement and stakeholder inclusion. Participatory design in dementia-related VR physical rehabilitation foregrounded feasibility in real-world care contexts and supported tailoring to users' preferences and constraints [102]. In parallel, carers' and professionals' perspectives emphasized practical barriers and facilitators to uptake (eg, training requirements, perceived appropriateness, and care workflows), indicating that acceptance is not only an end user property but also an organizational and implementation outcome [101]. Iterative user-centered

XSL•FO
RenderX

development of robot-mediated paired activities likewise illustrates how repeated testing cycles can be used to refine interaction patterns for older adults with MCI, particularly when social and motivational dynamics are central to the intervention experience [107].

Finally, the included set suggests an emerging trend toward adaptive and home-oriented solutions, while also revealing unresolved UX challenges. Combining reinforcement learning with VR for mild neurocognitive impairment reflects a move toward personalization and dynamic difficulty regulation, with usability assessment extending to both patients and caregivers [97]. Home-based or remotely oriented cognitive rehabilitation tools (including app-based platforms) further broaden access but amplify the importance of intuitive interaction, autonomy support, and long-term engagement over time [96,106,108]. Overall, the evidence indicates that VR or AR systems are frequently acceptable and usable in controlled pilots, but comparability is limited by heterogeneous UX constructs and reporting practices. Future trials should therefore adopt multidimensional usability frameworks, report stakeholder-specific acceptance (patients, carers, and clinicians), and integrate objective interaction markers alongside standardized questionnaires to better explain adoption and adherence [18,98,100,101].

VR or AR for Diagnosis and Assessment (ADL or IADL, Spatial Navigation, and Functional Skills)

The studies in this subgroup leverage VR or AR (and closely related interactive 3D systems) primarily as measurement tools to assess clinically relevant cognitive-functional constructs with greater ecological validity than traditional clinic-based tests (Table 4). A clear concentration is observed in (1) spatial navigation and orientation as sensitive markers of age-related decline and MCI-related impairment, and (2) functional skills or ADL-IADL performance captured through structured, everyday-like scenarios. Across the subgroup, VR systems are used either as standalone diagnostic probes or as performance-based end points embedded within feasibility, validation, or intervention protocols, with heterogeneous outcomes and reporting practices limiting direct cross-study comparability.

XSL•FO
RenderX

Table . VR[a] or AR[b] for diagnosis and assessment (ADL[c] or IADL[d], spatial navigation, and functional skills).

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
Mccracken et al (2025) [109]	Healthy younger adults vs healthy older adults; no diagnosed cognitive impairment	Cross-sectional experimental study; within-subject comparison of real vs VR homing; N≈40 (younger+older adults)	VR navigation task (HMD[e]) plus matched real-world corridor task—Fully immersive HMD VR Participants performed homing trials in both real and virtual replicas of an indoor environment; accuracy and path metrics compared across age groups and modal	Single laboratory session with multiple trials in each condition Domains: spatial navigation; path integration; visuospatial memory Outcome: older adults showed reduced homing accuracy in both environments; VR reproduced age-related navigation deficits observed in the real world.
Tuena et al (2024) [110]	Older adults with amnestic MCI[f] and cognitively healthy controls	Experimental study with between-group comparison of navigation aids; N≈30-40 (MCI+controls)	Desktop VR maze or navigation task—nonimmersive desktop VR Participants learned routes in a virtual environment with different navigation aids (bodily vs visual-cognitive cues), followed by spatial recall tests	Single experimental session with several learning–recall blocks Domains: spatial memory; spatial navigation; executive functions Outcome: Bodily and visual-cognitive aids improved spatial recall, particularly in participants with MCI.
Silva et al (2023) [111]	Community-dwelling older adults without dementia	Feasibility and validation study; N≈40-50 older adults	Immersive VR outdoor-navigation scenario (SOIVET system)—fully immersive HMD VR; older adults explored a virtual neighborhood and performed egocentric orientation tasks (pointing or walking to remembered locations); usability and cybersickness	Single VR assessment session Domains: Egocentric spatial orientation; visuospatial memory; attention Outcome: Immersive VR assessment was feasible, well tolerated, and showed meaningful variability in egocentric orientation.
Tuena et al (2024) [112]	Older adults with MCI	Proof-of-concept interventional trial with VR training vs control; N≈30 MCI participants (training vs control)	Desktop VR navigation game with embodied interaction—nonimmersive or low-immersion VR Participants with MCI completed repeated embodied VR navigation tasks over several weeks; pre-post cognitive and navigation measures were collected	Multisession VR training over several weeks Domains: Spatial navigation; episodic memory; executive functions Outcome: VR-based embodied navigation training improved spatial navigation and some cognitive scores relative to controls.
Panerai et al (2023) [113]	Adults and older adults with major neurocognitive disorder (degenerative and nondegenerative etiologies)	Nonrandomized interventional study (VR training vs. control); N=67 (40 experimental; 27 control)	Tablet or PC-based nonimmersive VR applications simulating ADL or IADL tasks—nonimmersive VR Experimental group trained four functional living skills in virtual scenarios, with in vivo performance of the same tasks assessed at multiple time points	Structured multisession VR training with three assessment time points Domains: Functional living skills (ADL or IADL); executive functions; attention; memory Outcome: VR-trained participants improved both in virtual tasks and in real-world execution of the trained functional skills.

XSL•FO
RenderX

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
Stasolla et al (2024) [114]	People with Alzheimer disease and cognitively healthy controls (where applicable)	Comparative study of traditional vs VR-based or digital storytelling approaches; N=NR[g] (to be extracted)	Nonimmersive VR or computer-based digital storytelling platform—nonimmersive VR or screen-based Patients engaged in digital storytelling activities delivered via VR-inspired interactive environments; outcomes compared with more standard assessment or rehabilitation	Multisession intervention over several weeks (details to be specified from full text) Domains: Episodic memory; autobiographical memory; executive functions; language Outcome: VR or digital-storytelling condition showed comparable or better gains in some cognitive and functional measures relative to standard approaches.
Mirmiran and Moussavi (2026) [115]	Institutionalized older adults with advanced dementia	Single-arm pilot intervention study; N=10 (9 completing ≥20 sessions)	Semi-immersive driving simulator with large display and steering controls—semi-immersive VR Participants repeatedly used a VR driving simulator presenting pleasant outdoor environments; quality of life and behavioral symptoms were monitored	Several sessions per week over approximately four months Domains: Well-being and engagement (primary); visuomotor coordination; secondary behavioral outcome Outcome: Feasible in advanced dementia; associated with improved quality-of-life measures and reduced behavioral disturbance in many residents.
Varela-Aldás et al (2022) [116]	Healthy adults (middle-aged and older)	Feasibility and validation study; N≈30 healthy participants	VR kitchen or cupboard scenario using an HMD—fully immersive HMD VR Participants memorized the location of everyday objects in virtual cupboards and later recalled or placed them; task metrics were related to standard neuropsychology	Single VR assessment session (~30-45 minutes) Domains: Everyday episodic memory; prospective memory; executive functions Outcome: Cupboard task was feasible, well accepted, and produced memory indices correlated with traditional tests, supporting ecological validity.
Bhargava et al (2025) [117]	Game developers and neuroscience/cognition researchers (no patient sample)	Qualitative mixed methods study (interviews, workshops); N=approximately 20-30 professionals	Digital serious games for cognition (including VR or 3D concepts)—mainly nonimmersive; conceptual discussion of multiple levels of immersion, explored design processes, constraints, and opportunities when cocreating games aimed at assessing cognition	Not applicable (design or consultation study) Domains: Multiple domains: attention, memory, executive functions, processing speed Outcome: Identified key design principles and tensions between playability and psychometric rigor for cognition-assessing games.

XSL•FO
RenderX

Study (year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
Mahncke (2025) [118]	Conceptual paper; focuses on clinical and research populations reporting cognitive complaints	Commentary or methodological article; N=not applicable	Not specific to one technology; implications for digital and VR-based assessments—not applicable, provides guidance on interpretation of self-reported cognitive complaints and their relation to objective performance measures	Not applicable Domains: Metacognition; subjective cognitive complaints; self-awareness of decline Outcome: Argues that self-report alone is insufficient as a diagnostic tool and should be combined with objective (including VR) assessments and longitudinal follow-up.

[a]VR: virtual reality.

[b]AR: augmented reality.

[c]ADL: activities of daily living.

[d]IADL: instrumental activities of daily living.

[e]HMD: head-mounted display.

[f]MCI: mild cognitive impairment.

[g]NR: not reported.

Several studies operationalize navigation using route learning, homing, and egocentric orientation tasks. A real-world versus virtual comparison of homing performance indicates that VR can reproduce age-related decrements in navigation or path integration, supporting the construct validity of VR-derived navigation metrics in older adults [109]. Immersive assessment of egocentric orientation in older adults further demonstrates feasibility of HMD-based protocols to quantify orientation performance while capturing usability- and tolerability-relevant factors inherent to immersive testing [111]. In MCI-focused work, desktop-based paradigms examine how navigation task design (eg, bodily and visual-cognitive aids) shapes spatial recall and diagnostic sensitivity, suggesting that the embedded cue structure may meaningfully influence performance and interpretability [110]. Related protocols extend to embodied spatial navigation paradigms where repeated task performance can be systematically tracked, offering a potential route to longitudinal assessment signatures (eg, learning curves, variability, and strategy shifts) rather than single-session snapshots [112].

A complementary line of research targets functional capacity via simulated everyday tasks (eg, shopping, medication management, and packing), thereby enabling standardized measurement of instrumental activities that are directly relevant to clinical staging and care needs [113]. Immersive everyday-memory assessment similarly embeds memory demands in familiar household contexts, providing object-location and prospective-memory–like indices intended to reflect real-world failures more directly than decontextualized neuropsychological tasks [116]. Collectively, these approaches illustrate how VR-based assessment can bridge the gap between laboratory cognition measures and functional outcomes, although the extent of transfer to real-world functioning remains variably demonstrated and often depends on task selection and scoring definitions.

In more advanced dementia, formal cognitive testing can be constrained by disease severity; therefore, studies sometimes prioritize pragmatic assessment outcomes, such as engagement, behavioral symptoms, and quality-of-life proxies within repeatable, structured experiences [115]. Other work uses VR-supported or VR-inspired digital storytelling frameworks to compare outcomes against more standard approaches, using the virtual environment as a scaffold for eliciting autobiographical or episodic content and monitoring response patterns [114]. In addition, immersive VR cognitive stimulation trials in AD incorporate performance-based and clinical outcome assessments as end points (even when the primary intent is therapeutic), providing data relevant to feasibility and evaluability of VR protocols in mild-to-moderate dementia [72].

In total, 2 contributions provide interpretive context for assessment-focused VR. A qualitative examination of collaboration between developers and neuroscience researchers highlights tensions between engagement-driven game mechanics and psychometric requirements, underscoring the importance of validity-by-design, telemetry planning, and transparency of scoring logic in cognition-assessing systems [117]. A methodological commentary on interpreting self-report cognition measures reinforces that subjective reports should not be used as standalone diagnostic indicators and motivates combining them with objective performance-based measures—including VR tasks—preferably within longitudinal frameworks [118]. Taken together, the subgroup supports VR or AR as a promising modality for assessing navigation, everyday memory, and functional capacity, but it also emphasizes the need for improved standardization (task definitions, scoring, adverse-event reporting, and benchmark validation) to strengthen synthesis and clinical interpretability across studies.

Conceptual, Methodological, and Guideline Literature: Broader Context

Table 5 summarizes conceptual, methodological, and guideline-oriented publications that provide the broader context

XSL•FO
RenderX

for VR- or AR-based approaches in cognitive rehabilitation and related functional outcomes in older adults. These papers do not primarily aim to estimate intervention efficacy; instead, they shape the field by (1) articulating design and usability guidance for older users, (2) proposing conceptual frameworks and taxonomies for intervention content and mechanisms, (3) outlining methodological pathways for development and evaluation, and (4) mapping research trends to identify gaps and priorities for scalable, reproducible work [2,14-16,21,30,119-126].

XSL•FO
RenderX

Table . Conceptual, methodological, and guideline literature: broader context.

Study (Year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
Abeele et al (2021) [14]	Older adults (aged 55-95 y), diverse cognitive abilities or care settings; first IVR[a] experience	Guideline paper: literature-derived guidelines+empirical qualitative study (interviews; laddering); N=37	Immersive VR[b] (HMD[c]-based IVR); guidelines spanning content/hardware/context immersive (HMD IVR) compiled 67 design guidelines from literature; exposed 37 older adults to first IVR experience and interviewed them to empirically ground or refine guidelines	Design or accessibility or user experience focus; cognitive screening used in sample characterization, empirically grounded design guidance for accessible or engaging IVR for older adults; links IVR attributes to functional or psychosocial benefits
Seifert and Schlomann (2021) [15]	Older adults (focus on adults aged older than 75 y in Europe; heterogeneous abilities and technology experience)	Opinion or short communication (narrative overview)	VR and AR[d] (broad use cases: health gaming, social contact, rehabilitation, and daily-life support) varies (not limited to a specific device class), discusses opportunities and barriers for older adults' VR or AR use; proposes developer or research recommendations (participatory design, and training or support)	General cognition or ADL[e] support context; highlights risks for users with cognitive limitations and cybersickness considerations, synthesizes key barriers (digital divide, skills or access, cybersickness, and environment constraints) and recommends participatory design and user support to improve adoption
Moon and Han (2022) [16]	People with neurological disorders (including cognitive impairment); accessibility for patients underexplored	Perspective paper	Immersive VR (stereo HMD) in neurology (rehabilitation+potential for diagnosis or assessment; metaverse context), immersive (explicitly defines VR as immersive stereo HMD), reviews current state of VR in neurology and proposes future directions, emphasizing VR-based cognitive assessment/diagnosis and accessibility or safety	Working memory, executive function, spatial navigation or spatial memory, cognitive planning (as examples for VR-based daily-life tasks), argues VR is underused for diagnosis or evaluation; proposes VR tasks simulating daily life with richer performance parameters and calls for standardization and diverse normative datasets
He et al (2022) [30]	Publications on VR-based cognitive rehabilitation (2001-2021); diseases include stroke, dementia, MCI[f], etc	Bibliometric analysis (Web of Science; visualization or clustering tools); N=1259 papers	VR for cognitive rehabilitation (field-level mapping) varies (depends on included studies), retrieved WoS[g] records (2001-2021) and analyzed authors, countries, keywords, and trends	Field-level: cognition or cognitive rehabilitation broadly maps research hotspots and trends; highlights telerehabilitation growth and notes lack of consensus on treatment methods and safety concerns (per authors)
Manser and de Bruin (2021) [120]	Older adults with mild neurocognitive disorder; project "Brain-IT" (design/development focus)	Methodological paper+case study applying Motivation–Interaction–Design–Evaluation framework	Exergame-based motor-cognitive training system (Dividat Senso or Senso Flex); gamification and multisensory feedback primarily nonimmersive exergaming (no HMD emphasized), describes step-by-step design or development process guided by MIDE (contextual research, game design or development, evaluation) to create exergame training aimed at halting or reducing	Motor-cognitive training (multiple domains; tailored to user capabilities and therapeutic needs) provides replicable methodological workflow and design requirements to increase usability, acceptance, and adherence of exergame interventions in clinical practice

XSL•FO
RenderX

Study (Year)	Population or stage	Design and sample (N)	Technology and intervention	Dose and key cognitive outcomes
Guzmán et al (2024) [21]	Older adults (cognitive rehabilitation context; framework-level; expert review included clinicians and therapists)	Conceptual framework development (Jabareen methodology)+expert review; N=5 experts (framework review)	Serious games for cognitive rehabilitation; emphasizes VR; includes AR as proposed extension primarily immersive VR context (mentions HMD; also discusses AR integration), develops a conceptual framework linking VR, physical activity, social interaction, difficulty adaptation, and content customization	Multiple domains (framework emphasizes multidomain cognitive rehab+real-world scenarios; examples include executive functions and others), Framework intended to guide design and implementation of serious games for older adults' cognitive rehabilitation; highlights iterative, multidisciplinary, participatory development and personalization
Cheng et al (2025) [119]	Older adults (community-dwelling); cognitive and independent living skills training in a home environment simulation	Feasibility study (qualitative interviews after gameplay); N=30	VR serious game ("Brainland") in home-like virtual environment with mini-games Immersive VR (HMD implied) designed a VR game with 6 mini-games targeting learning or cognitive and independent living skills; evaluated feasibility via semistructured interviews after play	Single trial session ~20 minutes (per author), cognitive skills+independent living skills (life skills or ADLs); mentions attention, working memory, processing speed as longer-term targets: Positive perceived usefulness of use and reported competency or cognitive developments; identifies engagement benefits and negative experiences (cybersickness or anxiety) informing best practices

[a]IVR: immersive virtual reality.

[b]VR: virtual reality.

[c]HMD: head-mounted display.

[d]AR: augmented reality.

[e]ADL: activities of daily living.

[f]MCI: mild cognitive impairment.

[g]WoS: Web of Science.

Empirically grounded guidance underscores that acceptability and UX are not peripheral outcomes but prerequisites for any clinically meaningful deployment. Design recommendations for immersive VR tailored to older adults emphasize onboarding, interaction simplicity, comfort, and accessibility as core requirements for both clinical and home use [14]. Broader discussions of VR or AR adoption in later life further highlight structural and individual barriers (eg, technology access, digital skills, perceived relevance, and tolerability), motivating usability-by-design and inclusive implementation strategies [15]. Additional works in this group complement these contributions by further operationalizing usability, safety, and user-centered evaluation considerations relevant to older populations [124,126].

Conceptual papers provide shared vocabularies for describing intervention components (eg, task structure, feedback, engagement mechanisms, and functional relevance), which is essential for comparing heterogeneous VR or AR studies within systematic reviews. A conceptual framework for serious games for cognitive rehabilitation synthesizes design and content considerations intended to support both therapeutic intent and long-term engagement [21]. Related conceptual and synthesis contributions in this set further refine how VR or AR interventions can be specified and categorized, supporting clearer reporting and improved cross-study comparability [122,123].

Methodological publications stress iterative, user-centered development and transparent reporting of technology characteristics (hardware, interaction modality, and immersion level) and intervention parameters (dose, progression, and feedback). A structured methodological approach to exergame development illustrates how therapeutic goals can be aligned with game mechanics and feasibility constraints in older adults with neurocognitive disorders [120]. Complementary methodological contributions in this group further strengthen evaluation pipelines (eg, feasibility-to-trial progression, outcome selection, and implementation-relevant metrics) to reduce variability and improve reproducibility [2,125].

Perspective and bibliometric studies contextualize why certain paradigms, outcomes, and technologies dominate the literature and where underexplored opportunities remain. A neurological disorder perspective highlights both the promise of immersive VR for ecologically valid task simulation and measurement and the practical considerations affecting translation [16]. Bibliometric mapping of VR in cognitive rehabilitation

XSL•FO
RenderX

characterizes publication growth and thematic concentrations, informing priorities, such as methodological standardization, reporting quality, and scalability [30,121]. Additional mapping work (including domain-wide bibliometric analyses) and interaction-technology–focused syntheses (eg, hand-tracking interfaces used in VR-based assessment or intervention) further contextualize research growth and methodological gaps relevant to scalable deployment and measurement fidelity [41,127].

Finally, feasibility-oriented design work demonstrates how the above principles can translate into ecologically meaningful content and UX evidence in older adults, providing practical signals on tolerability and perceived usefulness that can guide refinement and future trial design [119].

Overall, this body of context literature supports three interpretive lenses for the present evidence synthesis, that are (1) usability and acceptability should be treated as core outcomes rather than secondary considerations [14,15,124,126], (2) heterogeneity in study designs and outcomes can be partly mitigated through clearer specification of intervention components and technology characteristics [2,21,122,123,125], and (3) future research should prioritize methodological rigor, standardized reporting, and scalable deployment pathways to move beyond proof-of-concept demonstrations [16,30,119,121].

Discussion

Principal Findings

This systematic review synthesizes evidence on immersive technologies for cognitive rehabilitation across the MCI-dementia continuum, with an explicit focus on usability, therapeutic efficacy, and limitations spanning VR, AR, and CAVE-like embodied systems. Overall, the literature indicates that immersive interventions are feasible and frequently engaging, with signals of benefit across multiple cognitive domains (eg, attention, memory, executive functions, and spatial navigation), but the strength of conclusions remains constrained by substantial heterogeneity in technology, protocols, and outcomes, as well as the predominance of small-sample and early-phase designs [2,4,20,23-26].

Figure 3 provides a schematic overview of the core challenges shaping the interpretation, synthesis, and real-world applicability of immersive XR-based cognitive rehabilitation across the MCI-dementia continuum. The diagram positions the XR intervention at the center and highlights 5 interrelated domains that structure the subsequent discussion. First, therapeutic goals and outcome selection must be stage-sensitive, reflecting fundamental differences between restorative aims in MCI and maintenance-oriented or quality-of-life outcomes in dementia. Second, the figure emphasizes the trade-off between technological immersion and usability, illustrating how interaction demands and interface complexity can confound cognitive outcomes. Third, it underscores the persistent gap between improvements on cognitive test scores and meaningful functional transfer to ADL or IADL. Fourth, methodological rigor and harmonized reporting are identified as prerequisites for cumulative evidence synthesis in a field characterized by heterogeneous protocols and outcome measures. Finally, feasibility, safety, and scalability are highlighted as cross-cutting determinants of effective dose, adherence, and implementation beyond controlled laboratory settings. Together, these interconnected dimensions frame the interpretation of the current evidence base and guide priorities for future research and clinical translation.

Figure 3. Key challenges for immersive extended reality cognitive rehabilitation (VR, AR, or CAVE) across the mild cognitive impairment–dementia continuum: stage-sensitive outcomes, immersion–usability trade-offs, functional transfer (ADL or IADL), methodological or reporting harmonization, and feasibility, safety, and implementation at scale. ADL: activities of daily living; AR: augmented reality; CAVE: Cave Automatic Virtual Environment; IADL: instrumental activities of daily living; MCI: mild cognitive impairment; VR: virtual reality; XR: extended reality.

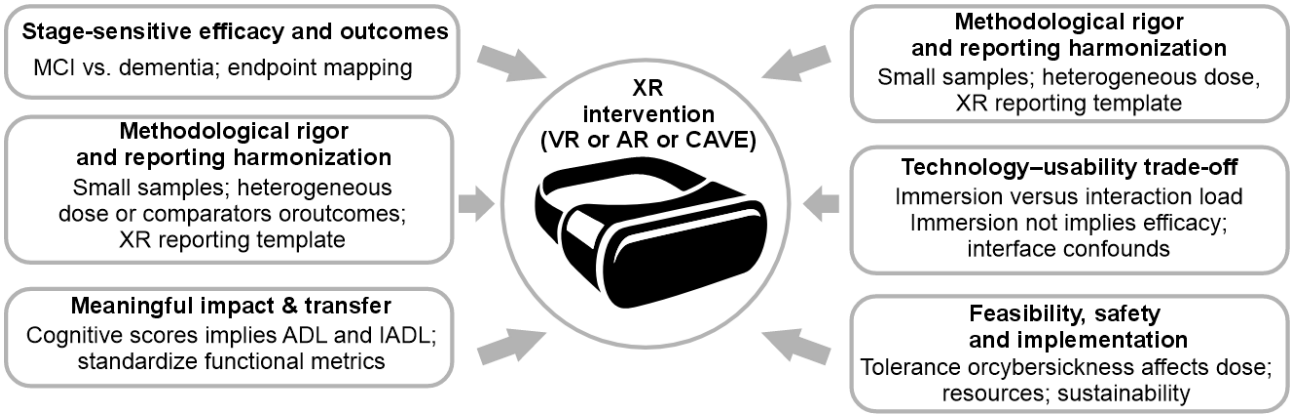

Interpretation of the Evidence Base Across MCI and Dementia

Across MCI, immersive training is commonly positioned as a restorative or capacity-enhancing approach, aiming to improve domain-specific performance (eg, executive control, working memory, and visuospatial abilities) and to strengthen skills relevant to everyday functioning [4,13,20]. In dementia, intervention goals often shift toward maintaining function, supporting compensatory strategies, and enhancing quality of life rather than achieving large gains on conventional neuropsychological end points; this distinction has important implications for outcome selection, intervention dose, and expectations of change [5,6]. Reviews spanning dementia or AD-oriented rehabilitation emphasize the value of ecologically meaningful tasks and the importance of aligning training scenarios with the functional challenges faced by patients and caregivers [5,6,17]. Taken together, the field would benefit from

XSL•FO
RenderX

more explicit stage-sensitive theories of change—what should improve (or be preserved), for whom, and through which mechanisms in MCI versus dementia [16,21].

Technology Matters: Immersion, Interaction Demands, and Ecological Validity

A core promise of immersive systems is ecological validity, the ability to engage cognitive processes within realistic, context-rich activities (eg, navigation, shopping-like scenarios, and multistep functional routines) that may better approximate everyday demands than traditional desktop-based training [13,17,112]. However, greater immersion is not synonymous with better outcomes. Higher immersion can increase interaction complexity and sensory load, which may introduce usability barriers for older adults with cognitive, motor, or sensory constraints and can confound interpretation (eg, performance differences driven by interface demands rather than the targeted cognitive construct) [8,14]. Accordingly, the usability literature emphasizes that interaction design, onboarding, and accessibility features are prerequisites for meaningful interpretation of outcome data, particularly in populations susceptible to fatigue, confusion, or anxiety during first exposure to novel technology [14,15].

AR remains less represented than VR in the cognitive rehabilitation literature, but it is conceptually attractive for supporting in situ cueing, errorless learning, and context-aware scaffolding of everyday activities [8,15,67]. A key research priority is to clarify where AR may offer a comparative advantage (eg, functional cueing and transfer) versus where VR may be preferable (eg, controlled manipulation of stimuli and graded exposure), while using comparable outcome frameworks that allow cross-modality synthesis [21,122].

From Cognitive Test Scores to Meaningful Functional Outcomes

Many studies and syntheses emphasize cognitive test performance; yet, fewer consistently demonstrate robust transfer to ADL or IADL, participation, or caregiver-relevant outcomes [4,17,20]. This gap is particularly consequential for dementia, where functional preservation and symptom management may be primary goals [5,6]. XR platforms are well-suited to embed functional tasks and to collect rich, process-level behavioral data (eg, error patterns, sequencing, and wayfinding efficiency), but these measures are rarely standardized across studies, limiting comparability and interpretability [13,124,126]. Future trials should prespecify clinically meaningful primary end points that map to the target population and intervention rationale and should report functional measures with sufficient detail to enable aggregation and replication [21,122,123].

Usability, Tolerability, and Safety as Determinants of Feasibility and Interpretability

Usability and acceptability are not secondary considerations in this domain; they are prerequisites for feasibility and for interpretable outcome measurement. Design and evaluation guidance for older users emphasizes minimizing interaction complexity, providing clear instructions and feedback, and ensuring comfort and accessibility in immersive experiences [8,14]. Cybersickness and related adverse effects can further influence adherence and exposure (eg, selective withdrawal, shortened sessions, and reduced “dose”), complicating cross-study comparisons and the interpretation of reported changes. Evidence focused on fully immersive VR in older adults indicates that cybersickness is a material barrier that warrants systematic monitoring and reporting, alongside mitigation strategies, such as gradual acclimation, careful locomotion design, and appropriate session duration [50]. More consistent reporting of adverse events, tolerability, and reasons for withdrawal is therefore essential for evaluating real-world feasibility and for understanding which systems and protocols are most likely to translate beyond controlled settings [14,15].

Methodological Limitations and Sources of Heterogeneity

The current evidence base is constrained by several recurring methodological issues, such as (1) small sample sizes and early-phase designs, (2) variability in comparator conditions and control interventions, (3) wide variation in dose and supervision, and (4) heterogeneous outcome measures and reporting practices [2,4,20,23]. These limitations reduce statistical power and complicate meta-analytic synthesis, while also increasing the risk that positive findings reflect context-specific effects rather than generalizable efficacy. Methodological and reporting frameworks emphasize the importance of clear intervention specification, stage-appropriate evaluation (feasibility vs efficacy), and transparent reporting of implementation conditions (training, supervision, equipment, and setting) [119,120,122,123]. The field would benefit from harmonized reporting templates tailored to XR rehabilitation (technology characteristics, immersion level, interaction modality, content design, progression rules, and support requirements), enabling cumulative synthesis across platforms and sites [16,21,122].

Scalability and Implementation in Real-World Care Pathways

Beyond efficacy, the practical deployment of XR interventions depends on scalability constraints, such as costs, maintenance, space requirements (particularly for room-scale systems), clinician time, and the need for caregiver involvement. Field-level syntheses and trend analyses consistently identify implementation, standardization, and scalability as major barriers to broader clinical uptake [16,30,121]. For home or community delivery, studies should evaluate feasibility under realistic constraints (digital literacy, connectivity, supervision, and safety) and should explicitly report the resources required for long-term use (setup time, troubleshooting, and staff training) [15,119]. In parallel, implementation of science constructs (eg, acceptability, feasibility, fidelity, and sustainability) should be incorporated prospectively rather than inferred post hoc [14,120].

Emerging Opportunities: Personalization, Multimodal Sensing, and Mechanistic Evaluation

Several contemporary directions may increase the clinical relevance of immersive rehabilitation. First, adaptive personalization (rule-based or data-driven) may better accommodate heterogeneity in baseline impairment, fatigue,

XSL•FO
RenderX

and learning profiles across MCI and dementia, but it requires transparent specification and validation to avoid "black-box" intervention effects [21,122]. Second, integration with multimodal sensing (eg, physiological measures and behavioral telemetry) may enable more sensitive monitoring of cognitive load, engagement, and progression but raises methodological questions about reliability, privacy, and clinical interpretability [3,125]. Finally, stronger mechanistic evaluation—linking task demands to hypothesized cognitive processes and to functionally meaningful outcomes—is needed to move from promising prototypes to reproducible treatment programs [13,16,21].

Strengths and Limitations of This Review

A key strength of this review is the integrated synthesis of usability, therapeutic efficacy, and limitations across immersive modalities, rather than treating usability as a secondary concern or focusing narrowly on 1 technology class [8,14]. Nevertheless, our conclusions are bounded by the limitations of the underlying literature, including heterogeneity in intervention content and outcomes, and the frequent reliance on pilot studies. As in many rapidly evolving technology domains, publication and selective-reporting biases remain plausible, and inconsistent reporting of implementation factors limits external validity judgments [4,16,122].

Implications for Clinical Practice and Future Research

From a practice perspective, the evidence supports cautious optimism—immersive interventions appear feasible and engaging for many individuals with MCI and, under appropriate support conditions, for some people living with dementia [5,6,14]. However, adoption should be guided by usability screening, conservative session dosing, and systematic monitoring of adverse effects, particularly in fully immersive VR [14,50].

For research, priorities include (1) standardized reporting of technology and intervention components, (2) stage-sensitive outcome frameworks linking cognition to ADL or IADL and quality of life, (3) adequately powered comparative trials with transparent comparator selection, (4) routine usability and safety reporting, and (5) implementation-focused evaluation to support scalability [16,21,30,119-123]. Addressing these challenges will improve the cumulative value of future studies and help transform immersive cognitive rehabilitation from a promising research area into a dependable component of care across the MCI-dementia continuum.

Conclusion

This systematic review synthesized current evidence on immersive technologies (VR, AR, and CAVE) for cognitive rehabilitation in MCI and dementia, with an emphasis on therapeutic outcomes, usability, and translational limitations. Overall, the literature suggests that immersive interventions are feasible and often engaging, with promising signals of benefit across multiple cognitive domains. However, the strength and generalizability of conclusions remain limited by substantial heterogeneity in intervention content, technology and immersion levels, interaction modalities, dosage, and outcome selection, alongside the frequent use of small samples and early-phase study designs.

A key implication is that outcome interpretation cannot be separated from usability and tolerability. In cognitively impaired older adults, interface complexity, onboarding requirements, comfort, and adverse effects can influence adherence and delivered exposure (ie, the effective "dose"), thereby shaping observed performance and functional outcomes. Future studies should therefore treat usability, safety, and acceptability as core outcomes, report them transparently (including adverse events and reasons for withdrawal), and analyze them alongside cognitive and functional measures.

For the next generation of research, priorities include (1) clearer specification and reporting of intervention components (tasks, feedback, progression, and supervision), (2) better alignment of outcomes with clinically meaningful goals, including functional end points (ADL or IADL) and quality-of-life measures, (3) adequately powered comparative trials with well-justified control conditions, and (4) explicit evaluation of implementation and scalability, particularly for home and community deployment. Advancing methodological standardization and reporting consistency will be essential to enable cumulative synthesis and to support translation of immersive cognitive rehabilitation into dependable, real-world care pathways across the MCI-dementia continuum.

Acknowledgments

This work was supported by the project SP2025/019, "Development of algorithms and systems for control, measurement and safety applications XI" of the Student Grant System, VSB-TU Ostrava. During the preparation of this work, the authors used ChatGPT (version 5.2; OpenAI) in order to improve the clarity and grammar of the manuscript draft. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This article has been produced with the financial support of the European Union under the REFRESH – Research Excellence For REgion Sustainability and High-tech Industries project CZ.10.03.01/00/22_003/0000048 via the Operational Programme Just Transition.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

XSL•FO
RenderX

Authors' Contributions

JK led the conceptualization and coordination of the extensive review, contributed to methodological design (eligibility criteria and synthesis plan), oversaw screening and extraction consistency, and led drafting and integration of the revised manuscript.
GL contributed clinical domain expertise in dementia/MCI and cognitive rehabilitation, supported refinement of eligibility interpretation across interventions, participated in screening/extraction, and critically revised the manuscript for clinical accuracy.
SB contributed to screening and data extraction, supported interpretation of intervention feasibility and outcomes, and participated in manuscript revision.
RF provided senior clinical-scientific input on outcomes and interpretation and critically revised the manuscript.
AL contributed to interpretation of clinical applicability and feasibility considerations and participated in manuscript revision.
RM contributed to technical reporting rigor and consistency checks (eg, structured synthesis, tabulation coherence) and revised the manuscript.
SM participated in screening/data extraction and supported synthesis of usability/acceptability and translational limitations.
AP contributed to screening/extraction and interpretation of clinical relevance and tolerability findings.
BP supported coordination and critical revision focused on clarity, structure, and reporting completeness.
AQ contributed to screening/data extraction and manuscript revision.
FR provided methodological/statistical consultation for synthesis framing and critically revised the manuscript.
RAS contributed to screening/data extraction and consistency/proofreading checks.
AS provided senior supervision, contributed to clinical interpretation, and critically revised the manuscript.
MP provided overall supervision and project administration, ensured reporting completeness and compliance, coordinated final approval, and led critical revision. All authors reviewed and approved the final manuscript and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Checklist 1
PRISMA checklist.
[PDF File, 231 KB - jmir_v28i1e84349_app1.pdf]

References

1. World Alzheimer Report 2024: Global changes in attitudes to dementia. Alzheimer's Disease International. 2024. URL: https://www.alzint.org/resource/world-alzheimer-report-2024/ [accessed 2026-02-22]
2. Sasaninezhad M, Moradi AR, Farahimanesh S, Choobin MH, Almasi-Dooghaee M. Enhancing cognitive flexibility and working memory in individuals with mild cognitive impairment: exploring the impact of virtual reality on daily life activities. Geriatr Nurs (Lond) 2024 Mar;56:32-39. [doi: 10.1016/j.gerinurse.2023.12.008]
3. Gkintoni E, Vassilopoulos SP, Nikolaou G, Vantarakis AC. Neurotechnological approaches to cognitive rehabilitation in mild cognitive impairment: a systematic review of neuromodulation, EEG, virtual reality, and emerging AI applications. Brain Sci 2025 May 28;15(6):582. [doi: 10.3390/brainsci15060582] [Medline: 40563754]
4. Tortora C, Di Crosta A, La Malva P, et al. Virtual reality and cognitive rehabilitation for older adults with mild cognitive impairment: a systematic review. AGEING Res Rev 2024 Jan;93:102146. [doi: 10.1016/j.arr.2023.102146] [Medline: 38036103]
5. Son C, Park JH. Ecological effects of VR-based cognitive training on ADL and IADL in MCI and AD patients: a systematic review and meta-analysis. IJERPH 2022 Nov 29;19(23):15875. [doi: 10.3390/ijerph192315875] [Medline: 36497946]
6. Vásquez-Carrasco E, Hernandez-Martinez J, Sepúlveda-Ramírez M, et al. Effectiveness of virtual reality interventions on quality of life, cognitive function and physical function in older people with Alzheimer's disease: a systematic review. Ageing Res Rev 2025 Jul;109:102785. [doi: 10.1016/j.arr.2025.102785] [Medline: 40436354]
7. Zuschnegg J, Schoberer D, Häussl A, et al. Effectiveness of computer-based interventions for community-dwelling people with cognitive decline: a systematic review with meta-analyses. BMC Geriatr 2023 Apr 12;23(1):229. [doi: 10.1186/s12877-023-03941-y] [Medline: 37041494]
8. Margrett JAA, Ouverson KMM, Gilbert SBB, Phillips LA, Charness N. Older adults' use of extended reality: a systematic review. Front VIRTUAL Real 2022;2. [doi: 10.3389/frvir.2021.760064]
9. Roley SS, DeLany JV, Barrows CJ, et al. Occupational therapy practice framework: domain & practice, 2nd edition. Am J Occup Ther 2008;62(6):625-683. [doi: 10.5014/ajot.62.6.625] [Medline: 19024744]
10. Milgram P, Kishino F. A taxonomy of mixed reality visual displays. IEICE Trans Inf & Syst 1994 Dec;E77-D(12):1321-1329 [FREE Full text]
11. Rauschnabel PA, Felix R, Hinsch C, Shahab H, Alt F. What is XR? Towards a framework for augmented and virtual reality. Comput Human Behav 2022 Aug;133:107289. [doi: 10.1016/j.chb.2022.107289]

XSL•FO
RenderX

12. Flavián C, Ibáñez-Sánchez S, Orús C. The impact of virtual, augmented and mixed reality technologies on the customer experience. J Bus Res 2019 Jul;100:547-560. [doi: 10.1016/j.jbusres.2018.10.050]
13. Tuena C, Borghesi F, Bruni F, et al. Technology-assisted cognitive motor dual-task rehabilitation in chronic age-related conditions: systematic review. J Med Internet Res 2023 May 22;25:e44484. [doi: 10.2196/44484] [Medline: 37213200]
14. Abeele VV, Schraepen B, Huygelier H, Gillebert C, Gerling K, van Ee R. Immersive virtual reality for older adults: empirically grounded design guidelines. ACM Trans Access Comput 2021 Sep;14(3):1-30. [doi: 10.1145/3470743]
15. Seifert A, Schlomann A. The use of virtual and augmented reality by older adults: potentials and challenges. Front VIRTUAL Real 2021;2. [doi: 10.3389/frvir.2021.639718]
16. Moon HJ, Han S. Perspective: present and future of virtual reality for neurological disorders. Brain Sci 2022 Dec 9;12(12):1692. [doi: 10.3390/brainsci12121692] [Medline: 36552152]
17. Buele J, Varela-Aldás JL, Palacios-Navarro G. Virtual reality applications based on instrumental activities of daily living (iADLs) for cognitive intervention in older adults: a systematic review. J NeuroEngineering Rehabil 2023;20(1). [doi: 10.1186/s12984-023-01292-8]
18. Latella D, Formica C, Ielo A, et al. A feasibility and usability study of a virtual reality tool (VESPA 2.0) for cognitive rehabilitation in patients with mild cognitive impairment: an ecological approach. Front Psychol 2024;15:1402894. [doi: 10.3389/fpsyg.2024.1402894] [Medline: 39492810]
19. Tuena C, Mancuso V, Stramba-Badiale C, et al. Egocentric and allocentric spatial memory in mild cognitive impairment with real-world and virtual navigation tasks: a systematic review. JAD 2021;79(1):95-116. [doi: 10.3233/JAD-201017] [Medline: 33216034]
20. Yang Q, Zhang L, Chang F, Yang H, Chen B, Liu Z. Virtual reality interventions for older adults with mild cognitive impairment: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2025 Jan 10;27:e59195. [doi: 10.2196/59195] [Medline: 39793970]
21. Guzmán DE, Rengifo CF, García-Cena CE. Serious games for cognitive rehabilitation in older adults: a conceptual framework. MTI 2024;8(8):64. [doi: 10.3390/mti8080064]
22. Alonso-Ríos D, Vázquez-García A, Mosqueira-Rey E, Moret-Bonillo V. Usability: a critical analysis and a taxonomy. Int J Hum Comput Interact 2009 Dec 28;26(1):53-74. [doi: 10.1080/10447310903025552]
23. Maggio MG, Maione R, Cotelli M, et al. Cognitive rehabilitation using virtual reality in subjective cognitive decline and mild cognitive impairment: a systematic review. Front Psychol 2025;16:1641693. [doi: 10.3389/fpsyg.2025.1641693] [Medline: 40994868]
24. Yu J, Song J, Shen Q. Effects of fully immersive virtual reality training on cognitive function in patients with mild cognitive impairment: a systematic review and meta-analysis. Front Hum Neurosci 2024;18:1467697. [doi: 10.3389/fnhum.2024.1467697] [Medline: 39713172]
25. Yuan P, Chen J, Peng D, Yang Q, Liu B, Lu C. Effectiveness of VR-based cognitive training and games on cognitive rehabilitation in patients with MCI: a systematic review and meta-analysis. Front Neurol 2025;16:1691344. [doi: 10.3389/fneur.2025.1691344] [Medline: 41281571]
26. Zhu S, Sui Y, Shen Y, et al. Effects of virtual reality intervention on cognition and motor function in older adults with mild cognitive impairment or dementia: a systematic review and meta-analysis. Front AGING Neurosci 2021;13:586999. [doi: 10.3389/fnagi.2021.586999] [Medline: 34025384]
27. Mehrinejad Khotbehsara M, Soar J, Lokuge S, Mehrinejad Khotbehsara E, Ip WK. The potential of virtual reality-based multisensory interventions in enhancing cognitive function in mild cognitive impairment: a systematic review. J Clin Med 2025 Aug 4;14(15):5475. [doi: 10.3390/jcm14155475] [Medline: 40807098]
28. Ienca M, Fabrice J, Elger B, et al. Intelligent assistive technology for Alzheimer's disease and other dementias: a systematic review. Journal of Alzheimer's Disease 2017 Feb 20;56(4):1301-1340. [doi: 10.3233/JAD-161037]
29. Blattgerste J, Renner P, Pfeiffer T. Augmented reality action assistance and learning for cognitively impaired people. 2019 Jun 5 Presented at: PETRA '19: The 12th Pervasive Technologies Related to Assistive Environments Conference; Jun 5-7, 2019; Rhodes Greece p. 270-279 URL: https://dl.acm.org/doi/proceedings/10.1145/3316782 [accessed 2026-02-22] [doi: 10.1145/3316782.3316789]
30. He D, Cao S, Le Y, Wang M, Chen Y, Qian B. Virtual reality technology in cognitive rehabilitation application: bibliometric analysis. JMIR SERIOUS GAMES 2022 Oct 19;10(4):e38315. [doi: 10.2196/38315] [Medline: 36260388]
31. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. Campbell Syst Rev 2022 Jun;18(2):e1230. [doi: 10.1002/cl2.1230] [Medline: 36911350]
32. Gattoni MF, Gobbo S, Feroldi S, et al. Identification of cognitive training for individuals with Parkinson's disease: a systematic review. Brain Sci 2025 Jan 11;15(1):61. [doi: 10.3390/brainsci15010061] [Medline: 39851429]
33. Catania V, Rundo F, Panerai S, Ferri R. Virtual reality for the rehabilitation of acquired cognitive disorders: a narrative review. BIOENGINEERING (BASEL) 2024;11(1):35. [doi: 10.3390/bioengineering11010035]
34. Chen C, Huang N, Hu B, Zhang M, Yuan J, Guo J. The effectiveness of digital technology interventions for cognitive function in older adults: a systematic review and meta-analysis of randomized controlled trials. GEROSCIENCE 2025 Feb;47(1):653-683. [doi: 10.1007/s11357-024-01446-z] [Medline: 39688787]

XSL•FO
RenderX

35. Gao Y, Ma L, Lin C, et al. Effects of virtual reality-based intervention on cognition, motor function, mood, and activities of daily living in patients with chronic stroke: a systematic review and meta-analysis of randomized controlled trials. Front AGING Neurosci 2021;13:766525. [doi: 10.3389/fnagi.2021.766525] [Medline: 34966267]
36. Parisi A, Bellinzona F, Di Lernia D, et al. Efficacy of multisensory technology in post-stroke cognitive rehabilitation: a systematic review. J Clin Med 2022 Oct 26;11(21):21. [doi: 10.3390/jcm11216324] [Medline: 36362551]
37. Tobar A, Jaramillo AP, Costa SC, Costa KT, Garcia SS. A physical rehabilitation approach for Parkinson's disease: a systematic literature review. CUREUS 2023 Sep;15(9):e44739. [doi: 10.7759/cureus.44739] [Medline: 37809151]
38. Campo-Prieto P, Cancela JM, Rodríguez-Fuentes G. Immersive virtual reality as physical therapy in older adults: present or future (systematic review). VIRTUAL Real 2021 Sep;25(3):801-817. [doi: 10.1007/s10055-020-00495-x]
39. Moulaei K, Sharifi H, Bahaadinbeigy K, Dinari F. Efficacy of virtual reality-based training programs and games on the improvement of cognitive disorders in patients: a systematic review and meta-analysis. BMC PSYCHIATRY 2024 Feb 12;24(1):116. [doi: 10.1186/s12888-024-05563-z] [Medline: 38342912]
40. Perra A, Riccardo CL, De Lorenzo V, et al. Fully immersive virtual reality-based cognitive remediation for adults with psychosocial disabilities: a systematic scoping review of methods intervention gaps and meta-analysis of published effectiveness studies. Int J Environ Res Public Health 2023 Jan 14;20(2):1527. [doi: 10.3390/ijerph20021527] [Medline: 36674283]
41. Kacmaz KS, Kaçmaz C. Bibliometric analysis of research in pediatrics related to virtual and augmented reality: a systematic review. Curr Pediatr Rev 2024;20(2):178-187. [doi: 10.2174/1573396319666230214103103] [Medline: 36786143]
42. Azhari HF. Advancing stroke diagnosis and management through nuclear medicine: a systematic review of clinical trials. Front Med (Lausanne) 2024;11:1425965. [doi: 10.3389/fmed.2024.1425965] [Medline: 39224610]
43. Lo YT, Su HC, Chuenchomnoy C, et al. Umbrella review of nonpharmacological interventions for intrinsic capacity in older adults. Ageing Res Rev 2025 Jun;108:102742. [doi: 10.1016/j.arr.2025.102742] [Medline: 40194665]
44. Bacanoiu MV, Danoiu M. New strategies to improve the quality of life for normal aging versus pathological aging. J Clin Med 2022 Jul 20;11(14):4207. [doi: 10.3390/jcm11144207] [Medline: 35887969]
45. Cabreira VA, Wilkinson T, Frostholm L, Stone J, Carson AJ. Systematic review and meta-analysis of standalone digital interventions for cognitive symptoms in people without dementia. npj Digit Med 2024;7(1). [doi: 10.1038/s41746-024-01280-9]
46. Mao Q, Zhao Z, Yu L, Zhao Y, Wang H. The effects of virtual reality-based reminiscence therapies for older adults with cognitive impairment: systematic review. J Med INTERNET Res 2024 Nov 12;26:e53348. [doi: 10.2196/53348] [Medline: 39531267]
47. Shahmoradi L, Mohammadian F, Rahmani Katigari M. A systematic review on serious games in attention rehabilitation and their effects. Behav Neurol 2022;2022:2017975. [doi: 10.1155/2022/2017975] [Medline: 35256889]
48. Calderone A, Marra A, De Luca R, et al. Multisensory stimulation in rehabilitation of dementia: a systematic review. Biomedicines 2025 Jan 9;13(1):149. [doi: 10.3390/biomedicines13010149] [Medline: 39857732]
49. Chi K, Chen J, Zhou S, Han Z. The effectiveness of digital cognitive intervention in patients with traumatic brain injury: systematic review and meta-analysis. Front Neurol 2025;16:1651443. [doi: 10.3389/fneur.2025.1651443] [Medline: 41127285]
50. Drazich BF, McPherson R, Gorman EF, et al. In too deep? A systematic literature review of fully‐immersive virtual reality and cybersickness among older adults. J AMERICAN GERIATRICS SOCIETY 2023 Dec;71(12):3906-3915. [doi: 10.1111/jgs.18553]
51. Burch BF, Gu J, Betz G, et al. Smart glasses for older adults with cognitive impairment: a scoping review. J Am Med Dir Assoc 2025 Nov;26(11):105831. [doi: 10.1016/j.jamda.2025.105831] [Medline: 40882953]
52. Romero-Ayuso D, Castillero-Perea Á, González P, et al. Assessment of cognitive instrumental activities of daily living: a systematic review. Disabil Rehabil 2021 May;43(10):1342-1358. [doi: 10.1080/09638288.2019.1665720] [Medline: 31549907]
53. Tan SB, Tan J, Raczkowska MN, et al. Digital game-based interventions for cognitive training in healthy adults and adults with cognitive impairment: protocol for a two-part systematic review and meta-analysis. BMJ Open 2023 May 4;13(5):e071059. [doi: 10.1136/bmjopen-2022-071059] [Medline: 37142320]
54. Kooijmans ECM, Hoogendijk EO, Drapała N, et al. Defining and categorizing nonpharmacologic interventions in the older population: a systematic review. J Am Med Dir Assoc 2025 Jan;26(1):105306. [doi: 10.1016/j.jamda.2024.105306] [Medline: 39424279]
55. Jun P, Chengye H, Hui W. Bibliometric analysis of rehabilitation in Alzheimer's disease (2000-2023): trends, hotspots and prospects. Front Aging Neurosci 2024;16:1457982. [doi: 10.3389/fnagi.2024.1457982] [Medline: 39691162]
56. Li Z, Yang L, Qiu H, Wang X, Zhang C, Zhang Y. Comparative efficacy of 5 non-pharmacological therapies for adults with post-stroke cognitive impairment: a Bayesian network analysis based on 55 randomized controlled trials. Front Neurol 2022;13. [doi: 10.3389/fneur.2022.977518]
57. Oliveira J, Gamito P, Souto T, et al. Virtual reality-based cognitive stimulation on people with mild to moderate dementia due to Alzheimer's disease: a pilot randomized controlled trial. Int J Environ Res Public Health 2021 May 16;18(10):5290. [doi: 10.3390/ijerph18105290] [Medline: 34065698]

XSL•FO
RenderX

58. Huang LC, Yang YH. The long-term effects of immersive virtual reality reminiscence in people with dementia: longitudinal observational study. JMIR SERIOUS GAMES 2022 Jul 25;10(3):e36720. [doi: 10.2196/36720] [Medline: 35877169]
59. Fiorenzato E, Zabberoni S, De Simone MS, et al. Effects of virtual reality cognitive training on executive function and prospective memory in Parkinson's disease and healthy aging. J Neurol Sci 2025 Jun 15;473:123507. [doi: 10.1016/j.jns.2025.123507] [Medline: 40250092]
60. De Simone MS, Costa A, Tieri G, et al. The effectiveness of an immersive virtual reality and telemedicine-based cognitive intervention on prospective memory in Parkinson's disease patients with mild cognitive impairment and healthy aged individuals: design and preliminary baseline results of a placebo-controlled study. Front Psychol 2023;14:1268337. [doi: 10.3389/fpsyg.2023.1268337] [Medline: 37928597]
61. Liu Z, He Z, Yuan J, et al. Application of immersive virtual-reality-based puzzle games in elderly patients with post-stroke cognitive impairment: a pilot study. BRAIN Sci 2023;13(1):79. [doi: 10.3390/brainsci13010079]
62. De Luca R, Calderone A, Gangemi A, et al. Is virtual reality orientation therapy useful to optimize cognitive and behavioral functioning following severe acquired brain injury? An exploratory study. Brain Sci 2024 Apr 23;14(5):410. [doi: 10.3390/brainsci14050410] [Medline: 38790389]
63. Baldimtsi E, Mouzakidis C, Karathanasi EM, et al. Effects of virtual reality physical and cognitive training intervention on cognitive abilities of elders with mild cognitive impairment. Journal of Alzheimer's Disease Reports 2023 Nov 1;7(1):1475-1490. [doi: 10.3233/ADR-230099]
64. Downey R, Gagné N, Mohanathas N, et al. At-home computerized executive-function training to improve cognition and mobility in normal-hearing adults and older hearing aid users: a multi-centre, single-blinded randomized controlled trial. BMC Neurol 2023 Oct 20;23(1):378. [doi: 10.1186/s12883-023-03405-1] [Medline: 37864139]
65. Zhu B, Zhou Y, Weng Q, et al. Effects of biophilic virtual reality on cognitive function of patients undergoing laparoscopic surgery: study protocol for a sham randomised controlled trial. BMJ Open 2022 Jul 8;12(7):e052769. [doi: 10.1136/bmjopen-2021-052769] [Medline: 35803619]
66. Buele J, Avilés-Castillo F, Del-Valle-Soto C, Varela-Aldás J, Palacios-Navarro G. Effects of a dual intervention (motor and virtual reality-based cognitive) on cognition in patients with mild cognitive impairment: a single-blind, randomized controlled trial. J Neuroeng Rehabil 2024 Aug 1;21(1):130. [doi: 10.1186/s12984-024-01422-w] [Medline: 39090664]
67. Han K, Park K, Choi KH, Lee J. Mobile augmented reality serious game for improving old adults' working memory. Appl Sci (BASEL) 2021;11(17):7843. [doi: 10.3390/app11177843]
68. Jeong E, Ham Y, Lee SJ, Shin JH. Virtual reality-based music attention training for acquired brain injury: A randomized crossover study. Ann N Y Acad Sci 2024 Nov;1541(1):151-162. [doi: 10.1111/nyas.15249] [Medline: 39476208]
69. Lanzoni D, Vitali A, Regazzoni D, Rizzi C. Design of customized virtual reality serious games for the cognitive rehabilitation of retrograde amnesia after brain stroke. J Comput Inf Sci Eng 2022 Jun 1;22(3). [doi: 10.1115/1.4053149]
70. Maeng S, Hong JP, Kim WH, et al. Effects of virtual reality-based cognitive training in the elderly with and without mild cognitive impairment. PSYCHIATRY Investig 2021 Jul;18(7):619-627. [doi: 10.30773/pi.2020.0446] [Medline: 34265200]
71. Ip WK, Soar J, James C, Wang SY, Fong KNK. Virtual reality game-based training for preventing falls among community-dwelling older adults with mild cognitive impairment: a pilot randomized control trial study. Virtual Real 2025;29(1). [doi: 10.1007/s10055-024-01084-y]
72. Oliveira LM, Evangelista E Souza EH, Alves MR, et al. 2D virtual reality-based exercise improves spatial navigation in institutionalized non-robust older persons: a preliminary data report of a single-blind, randomized, and controlled study. Front Neurol 2020;11:609988. [doi: 10.3389/fneur.2020.609988] [Medline: 33584510]
73. Początek G, Wyszomirska J, Segiet N. Cognitive training for Alzheimer's disease. Aktualn Neurol 2023;23(4):138-144 [FREE Full text] [doi: 10.15557/AN.2023.0021]
74. Cavallo M, Lasaponara S. Cognition and cognitive training in patients with neurodegenerative diseases. Brain Sci 2025 Feb 12;15(2):183. [doi: 10.3390/brainsci15020183] [Medline: 40002516]
75. Kwan RYC, Liu J, Sin OSK, et al. Effects of virtual reality motor-cognitive training for older people with cognitive frailty: multicentered randomized controlled trial. J Med Internet Res 2024 Sep 11;26:e57809. [doi: 10.2196/57809] [Medline: 39259959]
76. Fu Y, Wang H. Clinical observation of VR virtual reality rehabilitation training combined with acupuncture in the treatment of mild cognitive impairment. SLAS Technol 2025 Apr;31:100250. [doi: 10.1016/j.slast.2025.100250] [Medline: 39863045]
77. Manenti R, Baglio F, Pagnoni I, et al. Long-lasting improvements in episodic memory among subjects with mild cognitive impairment who received transcranial direct current stimulation combined with cognitive treatment and telerehabilitation: a multicentre, randomized, active-controlled study. Front Aging Neurosci 2024;16:1414593. [doi: 10.3389/fnagi.2024.1414593] [Medline: 38966802]
78. Gambella E, Margaritini A, Benadduci M, et al. An integrated intervention of computerized cognitive training and physical exercise in virtual reality for people with Alzheimer's disease: The jDome study protocol. Front Neurol 2022;13:964454. [doi: 10.3389/fneur.2022.964454] [Medline: 36034306]
79. Contrada M, Pignolo L, Vatrano M, et al. Application of multidomain cognitive training in a tele-neurorehabilitation setting for treatment of post-stroke cognitive disorders. BRAIN Sci 2025;15(1):11. [doi: 10.3390/brainsci15010011]

XSL•FO
RenderX

80. Huber SK, Knols RH, Held JPO, Betschart M, de Bruin ED. PEMOCS: Evaluating the effects of a concept-guided, personalised, motor-cognitive exergame training on cognitive functions and gait in chronic stroke-study protocol for a randomised controlled trial. Trials 2024 Jul 4;25(1):451. [doi: 10.1186/s13063-024-08283-7] [Medline: 38965612]
81. Hwang NK, Choi JB, Choi DK, et al. Effects of semi-immersive virtual reality-based cognitive training combined with locomotor activity on cognitive function and gait ability in community-dwelling older adults. HEALTHCARE (Basel) 2021 Jun 28;9(7):814. [doi: 10.3390/healthcare9070814] [Medline: 34203233]
82. De Luca R, Marra A, Pollicino P, et al. Advances in neurorehabilitation of TREM2-related dementia: A case report on a novel multimodal approach using virtual reality. Medicine (Baltimore) 2022 May 27;101(21):e29470. [doi: 10.1097/MD.0000000000029470] [Medline: 35623081]
83. Faria AL, Almeida Y, Branco D, et al. NeuroAIreh@b: an artificial intelligence-based methodology for personalized and adaptive neurorehabilitation. Front Neurol 2023;14:1258323. [doi: 10.3389/fneur.2023.1258323] [Medline: 38322797]
84. Maggio MG, Maione R, Migale S, et al. Low-intensity virtual reality exercise for caregivers of people with mild cognitive impairment: a pilot study. J Funct Morphol Kinesiol 2025 Sep 16;10(3):353. [doi: 10.3390/jfmk10030353] [Medline: 40981052]
85. Bourrelier J, Fautrelle L, Haratyk E, et al. Enhancement of anticipatory postural adjustments by virtual reality in older adults with cognitive and motor deficits: a randomised trial. GERIATRICS (Basel) 2021 Jul 22;6(3):72. [doi: 10.3390/geriatrics6030072] [Medline: 34449605]
86. Tang YM, Tse SY, Chan HS, Yip HT, Cheung HT, Geda MW. An innovative interactive narrative persona (INP) approach for virtual reality-based dementia tour design (VDT) in rehabilitation contexts. Bioengineering (Basel) 2025 Sep 17;12(9):987. [doi: 10.3390/bioengineering12090987] [Medline: 41007233]
87. Bumbea AM, Glavan DG, Vasile RC, et al. Cognitive impairment and psychological morbidity among stroke survivors in rehabilitation: a cross-sectional analysis. J Clin Med 2025 Oct 31;14(21):21. [doi: 10.3390/jcm14217735] [Medline: 41227130]
88. Zhang T, Liu W, Bai Q, Gao S. Virtual reality technology in the rehabilitation of post-stroke cognitive impairment: an opinion article on recent findings. Front Psychol 2023;14:1271458. [doi: 10.3389/fpsyg.2023.1271458] [Medline: 37849482]
89. Capecci M, Gandolfi M, Straudi S, et al. Advancing public health through technological rehabilitation: insights from a national clinician survey. BMC Health Serv Res 2024 Dec 19;24(1):1626. [doi: 10.1186/s12913-024-11991-0] [Medline: 39702315]
90. Chau PH, Kwok YYJ, Chan MKM, et al. Feasibility, acceptability, and efficacy of virtual reality training for older adults and people with disabilities: single-arm pre-post study. J Med INTERNET Res 2021 May 4;23(5):e27640. [doi: 10.2196/27640] [Medline: 33944795]
91. Kim Y, Kim J, Kim J, Lee C, Ory MG. Feasibility and acceptability of an immersive puzzle-based virtual reality activity for community-dwelling older adults. Front VIRTUAL Real 2025;6. [doi: 10.3389/frvir.2025.1655981]
92. Munoz JE, Ali F, Basharat A, et al. Development of classifiers to determine factors associated with older adult's cognitive functions and game user experience in VR using head kinematics. IEEE Trans GAMES 2025;17(3):604-612. [doi: 10.1109/TG.2023.3317825]
93. Hassandra M, Galanis E, Hatzigeorgiadis A, et al. A virtual reality app for physical and cognitive training of older people with mild cognitive impairment: mixed methods feasibility study. JMIR Serious Games 2021 Mar 24;9(1):e24170. [doi: 10.2196/24170] [Medline: 33759797]
94. Kua Z, Ong RHS, Chen NYC, et al. Virtual reality-based cognitive training for MCI in the elderly: a feasibility randomised pilot study. Ann Acad Med Singap 2025 Jun 24;54(7):445-447. [doi: 10.47102/annals-acadmedsg.2024162] [Medline: 40799097]
95. Mondellini M, Arlati S, Gapeyeva H, et al. User experience during an immersive virtual reality-based cognitive task: a comparison between Estonian and Italian older adults with MCI. Sensors (Basel) 2022 Oct 27;22(21):21. [doi: 10.3390/s22218249] [Medline: 36365947]
96. Perra A, Primavera D, Leonetti V, et al. Virtual reality cognitive remediation tool for individuals with mild cognitive impairment: study protocol for a feasibility randomized clinical trial. Front PUBLIC HEALTH 2024;12:1477279. [doi: 10.3389/fpubh.2024.1477279] [Medline: 39651467]
97. Stasolla F, Di Gioia M. Combining reinforcement learning and virtual reality in mild neurocognitive impairment: a new usability assessment on patients and caregivers. Front Aging Neurosci 2023;15:1189498. [doi: 10.3389/fnagi.2023.1189498] [Medline: 37293666]
98. Tseng KC, Giau DTN. A feasibility study of using virtual reality as a pre-occupancy evaluation tool for the elderly. Automation in Construction 2022 Feb;134:104037. [doi: 10.1016/j.autcon.2021.104037]
99. Jeong H, Kang D, Kim JE, Lim J, Lee HW. Preliminary study on the feasibility of virtual reality-based cognitive training on patients with mild to moderate Alzheimer's disease. J Alzheimers Dis Rep 2025 Jan;9:25424823251385901. [doi: 10.1177/25424823251385901] [Medline: 41048557]
100. Maggio MG, Rizzo A, Benenati A, et al. Efficacy and feasibility of virtual reality-based cognitive tele-rehabilitation in Parkinson's disease: a pilot randomized controlled trial on patients and caregivers. Digit HEALTH 2025 May;11:20552076251376534. [doi: 10.1177/20552076251376534] [Medline: 41018508]

XSL•FO
RenderX

101. Sommerlad A, Jawharieh H, Ang CS, Huntley J, Gokani HA. Carers and professionals' views on using virtual reality in dementia care: a qualitative study. DEMENTIA (London) 2025 Jan;24(1):71-90. [doi: 10.1177/14713012241272786] [Medline: 39121239]
102. Matsangidou M, Frangoudes F, Schiza E, et al. Participatory design and evaluation of virtual reality physical rehabilitation for people living with dementia. VIRTUAL Real 2023 Mar;27(1):421-438. [doi: 10.1007/s10055-022-00639-1]
103. Chuang IC, Huang XT, Chen IC, Wu CY, Wu CY. Feasibility, usability, and effects of leisure-based cognitive training using a fully immersive virtual reality system in older adults: single-arm pretest-posttest pilot study. JMIR SERIOUS GAMES 2025 Oct 9;13:e66673. [doi: 10.2196/66673] [Medline: 41066691]
104. Chen CY, Chen F, Tsai CE. Exploring older people's user experiences with VR: a case study of Taiwanese seniors. Entertain Comput 2025 Sep;55:100974. [doi: 10.1016/j.entcom.2025.100974]
105. Siette J, Campbell C, Adam PJ, Harris CB. Exploring the usability of the virtual reality module LEAF CAFÉ: a qualitative think-aloud study. BMC Geriatr 2024 Feb 16;24(1):162. [doi: 10.1186/s12877-024-04767-y] [Medline: 38365613]
106. Pedroli E, Mancuso V, Stramba-Badiale C, et al. Brain m-app's structure and usability: a new application for cognitive rehabilitation at home. Front Hum Neurosci 2022;16:898633. [doi: 10.3389/fnhum.2022.898633] [Medline: 35782042]
107. Ghosh R, Khan N, Migovich M, et al. Iterative user centered design of robot-mediated paired activities for older adults with mild cognitive impairment (MCI). In: Lecture Notes in Computer Science 13331 LNCS: Springer; 2022:14-28. [doi: 10.1007/978-3-031-05654-3_2]
108. Haley KL, Cunningham KT, Kim I, Shafer JS. Autonomy-supportive treatment for acquired apraxia of speech: feasibility and therapeutic effect. Aphasiology 2021 Apr 3;35(4):539-559. [doi: 10.1080/02687038.2019.1705662]
109. McCracken MK, Shayman CS, Fino PC, Stefanucci JK, Creem-Regehr SH. A comparison of the effects of older age on homing performance in real and virtual environments. IEEE Trans Visual Comput GRAPHICS 2025;31(5):3213-3222. [doi: 10.1109/TVCG.2025.3549901]
110. Tuena C, Serino S, Goulene KM, Pedroli E, Stramba-Badiale M, Riva G. Bodily and visual-cognitive navigation aids to enhance spatial recall in mild cognitive impairment. Journal of Alzheimer's Disease 2024 May 28;99(3):899-910. [doi: 10.3233/JAD-240122]
111. Silva JD, Santos MD, Costa RD, et al. Applicability of an immersive virtual reality system to assess egocentric orientation of older adults. Arq Neuropsiquiatr 2023 Jan;81(1):019-026. [doi: 10.1055/s-0042-1759762]
112. Tuena C, Serino S, Pedroli E, et al. Embodied spatial navigation training in mild cognitive impairment: a proof-of-concept trial. Journal of Alzheimer's Disease 2024 Jul 30;100(3):923-934. [doi: 10.3233/JAD-240200]
113. Panerai S, Catania V, Rundo F, et al. Functional living skills in patients with major neurocognitive disorder due to degenerative or non-degenerative conditions: effectiveness of a non-immersive virtual reality training. Sensors (Basel) 2023 Feb 8;23(4):1896. [doi: 10.3390/s23041896] [Medline: 36850494]
114. Stasolla F, Di Gioia M, Messina I, et al. Assessing and recovering Alzheimer's disease: a comparative analysis of standard neuropsychological approaches and virtual reality interventions with the use of digital storytelling. Front Psychol 2024;15:1406167. [doi: 10.3389/fpsyg.2024.1406167] [Medline: 39114597]
115. Mirmiran S, Moussavi Z. Improving quality of life for institutionalized individuals with advanced dementia: a pilot study on efficacy of a semi-immersive virtual reality driving simulator for individuals with advanced dementia. DISPLAYS 2026 Jan;91:103182. [doi: 10.1016/j.displa.2025.103182]
116. Varela-Aldás J, Buele J, Amariglio R, García-Magariño I, Palacios-Navarro G. The cupboard task: an immersive virtual reality-based system for everyday memory assessment. Int J Hum Comput Stud 2022 Nov;167:102885. [doi: 10.1016/j.ijhcs.2022.102885]
117. Bhargava Y, Sumanth SV, Deshmukh A, Baths V. A peek into the minds of game developers and neuroscience researchers collaborating on cognition assessing games. IEEE ACCESS 2025;13:62674-62690. [doi: 10.1109/ACCESS.2025.3557454]
118. Mahncke HW. Interpreting self-report measures about cognition: what matters and what to do with the data: a commentary on "Awareness of baseline functioning and sensitivity to improvement in older people with and without mild cognitive impairment receiving a computerized functional skills training program". Int Psychogeriatr 2025 Jun;37(3):100082. [doi: 10.1016/j.inpsyc.2025.100082] [Medline: 40360334]
119. Cheng L, Pang WYJ, Kong A. A virtual reality game for older adults' immersive learning. Sci Rep 2025;15(1). [doi: 10.1038/s41598-025-07354-2]
120. Manser P, de Bruin ED. Making the best out of IT: design and development of exergames for older adults with mild neurocognitive disorder - a methodological paper. Front AGING Neurosci 2021;13:734012. [doi: 10.3389/fnagi.2021.734012] [Medline: 34955806]
121. He X, Liu C, Li Z, Cai X. Leisure activity interventions on cognition in mild cognitive impairment patients: a meta-analysis. Neuropsychiatr Dis Treat 2025;21:1671-1687. [doi: 10.2147/NDT.S536627] [Medline: 40827136]
122. Bhargava Y, Kumar Kattupalli K, Baths V. CogTrack: a proof of concept for cognition tracker. IEEE Access 2024;12:194609-194629. [doi: 10.1109/ACCESS.2024.3519220]
123. Dombrowski W, Mims A, Kremer I, et al. Dementia ideal care: ecosystem map of best practices and care pathways enhanced by technology and community. J Alzheimers Dis 2024;100(1):87-117. [doi: 10.3233/JAD-231491] [Medline: 38848182]

XSL•FO
RenderX

124. Kaiser W, de Bruin ED, Manser P. Domain-specific evaluation of exergame metrics among older adults with mild neurocognitive disorder: secondary analysis of 2 randomized controlled trials. JMIR SERIOUS GAMES 2025 May 21;13:e65878. [doi: 10.2196/65878] [Medline: 40397948]
125. Peng Y, Song P, Karako T, Asakawa T. Blocking progression from intervenable mild cognitive impairment to irreversible dementia, what can we do? Biosci Trends 2024 Nov 15;18(5):409-412. [doi: 10.5582/bst.2024.01304] [Medline: 39401896]
126. Singh SM, Kathiresan P, Gaur N. Clinical practice guidelines on addressing cognitive impairment in vascular dementias. Indian J Psychiatry 2025 Jan;67(1):54-64. [doi: 10.4103/indianjpsychiatry.indianjpsychiatry_713_24] [Medline: 40046475]
127. Colombini G, Duradoni M, Carpi F, Vagnoli L, Guazzini A. LEAP Motion technology and psychology: a mini-review on hand movements sensing for neurodevelopmental and neurocognitive disorders. Int J Environ Res Public Health 2021 Apr 11;18(8):4006. [doi: 10.3390/ijerph18084006] [Medline: 33920362]
128. Zhou L, Huang X, Wang J, Wang F, Liu J, Liu N. The influence of eight cognitive training regimes upon cognitive screening tool performance in post-stroke survivors: a network meta-analysis. Front Aging Neurosci 2024;16:1374546. [doi: 10.3389/fnagi.2024.1374546] [Medline: 39100748]

Abbreviations

AD: Alzheimer disease
ADL: activity of daily living
AR: augmented reality
CAVE: Cave Automatic Virtual Environment
HMD: head-mounted display
IADL: instrumental activity of daily living
MCI: mild cognitive impairment
MMSE: Mini-Mental State Examination
MoCA: Montreal Cognitive Assessment
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
UX: user experience
VR: virtual reality
WoS: Web of Science
XR: extended reality

Edited by A Schwartz, M Balcarras; submitted 18.Sep.2025; peer-reviewed by G Spalla, T Llana; revised version received 18.Dec.2025; accepted 18.Dec.2025; published 18.Mar.2026.

Please cite as:
Konecny J, Lanza G, Buono S, Ferri R, Luca A, Martinek R, Musso S, Palmigiano A, Prauzkova B, Quercia A, Rundo F, Subramanian RA, Serretti A, Prauzek M
Immersive Technologies for Cognitive Rehabilitation in Dementia and Mild Cognitive Impairment: Systematic Review
J Med Internet Res 2026;28:e84349
URL: https://www.jmir.org/2026/1/e84349
doi:10.2196/84349

© Jaromir Konecny, Giuseppe Lanza, Serafino Buono, Raffaele Ferri, Antonina Luca, Radek Martinek, Sabrina Musso, Aurora Palmigiano, Barbora Prauzkova, Angelica Quercia, Francesco Rundo, Ramachandran Avala Subramanian, Alessandro Serretti, Michal Prauzek. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Investigating the Effect of Hospital Infection Control Informatization on Optimizing Microbiological Specimen Submission Before Antibiotic Therapy: Failure Mode and Effects Analysis

Jianxiong Wu[1], PhD; Weijiang Zhan[2], BEng; Jun Hua[2], BEng; Qinling Ge[1], BS; Yuexian Zhu[1], BS; Huixian Yu[3], BS; Min Zhao[4], BPharm; Xiaoyan Zhan[5], BS; Bingwei Zhu[6], BS; Tongqi Xiang[7], BS; Longxi Lu[8*], PhD; Tieying Dai[1*], PhD

[1]Department of Public Health, The First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), 54 Youdian Road, Xihu District, Hangzhou, Zhejiang, China

[2]Department of Computer Center, the First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), Hangzhou, Zhejiang, China

[3]Department of Nursing, the First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), Hangzhou, Zhejiang, China

[4]Department of Pharmacy, the First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), Hangzhou, Zhejiang, China

[5]Discipline Inspection Commission Office, the First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), Hangzhou, Zhejiang, China

[6]Microbiology Laboratory, the First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), Hangzhou, Zhejiang, China

[7]Logistics Support Department, the First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), Hangzhou, Zhejiang, China

[8]Department of Environmental Health, Zhejiang Provincial Center for Disease Control and Prevention, Hangzhou, Zhejiang, China

*these authors contributed equally

Corresponding Author:
Tieying Dai, PhD
Department of Public Health, The First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), 54 Youdian Road, Xihu District, Hangzhou, Zhejiang, China

Abstract

Background: Antimicrobial resistance (AMR) poses a critical global health threat, with inappropriate antibiotic use being a major driver. Timely microbiological specimen submission before initiating antibiotic therapy is a cornerstone of antimicrobial stewardship (AMS), enabling pathogen-directed therapy and reducing unnecessary broad-spectrum exposure. However, suboptimal compliance remains common due to workflow interruptions, technological barriers, and behavioral factors. Failure Mode and Effects Analysis (FMEA), a proactive risk-assessment method widely used in health care quality improvement, provides a systematic framework to identify process vulnerabilities and prioritize corrective actions. Despite its increasing application, few studies have integrated FMEA with hospital informatization to optimize microbiological specimen submission workflows in routine AMS practice.

Objective: This study aimed to systematically identify workflow risks affecting preantibiotic microbiological specimen submission and to design, implement, and evaluate informatization-enabled interventions using an FMEA-based framework.

Methods: FMEA was conducted at a tertiary hospital in China. A multidisciplinary team identified potential failure modes across 4 domains: health information systems, personnel, administration, and external support. Risk Priority Numbers (RPNs) and Action Priority (AP) indices were calculated for each failure mode. Targeted interventions were implemented, including dual-verification barcode scanning, artificial intelligence-driven clinical decision support alerts, EHR-integrated training modules, and automated compliance dashboards. Pre- and postintervention specimen submission rates (January 2024-December 2024) were analyzed using the Mann-Kendall trend test.

Results: The top 5 failure modes included PDA barcode scanning failures (RPN=175), inadequate clinical decision support (RPN=140), insufficient clinician awareness (RPN=56), suboptimal oversight mechanisms, and patient-related barriers. Postintervention, significant upward trends were observed in overall specimen submission rates (P<.001), with similar improvements for restricted-use (P<.001) and special-use antibiotics (P<.001).

XSL•FO
RenderX

Conclusions: FMEA-based risk management combined with hospital informatization effectively optimized specimen submission workflows. Real-time decision support, process standardization, and interdisciplinary collaboration significantly enhanced compliance. Future research should evaluate long-term impacts on AMR reduction and diagnostic integration. Integrating FMEA with hospital informatization effectively strengthened the microbiological specimen submission process. Digital decision support, standardized workflows, and real-time monitoring substantially improved compliance with preantibiotic specimen submission. This approach provides an actionable model for data-driven AMS enhancement. Future studies should assess scalability, cost-effectiveness, and impacts on downstream AMR outcomes.

(J Med Internet Res 2026;28:e78118) doi:10.2196/78118

KEYWORDS

microbiological specimen submission; failure mode and effects analysis; antibiotic stewardship; hospital informatization; risk management

Introduction

Antimicrobial resistance (AMR) poses a critical threat to global public health and health care systems [1,2]. It was estimated that AMR was associated with 4.71 million deaths worldwide in 2021 [3]. Inappropriate use of antibiotics is a primary driver of AMR [4]. Accurate microbiological pathogen detection is essential for guiding rational antibiotic therapy [5]. On one hand, identifying the causative pathogens and their antimicrobial susceptibility profiles can reduce unnecessary exposure to broad-spectrum antibiotics. On the other hand, microbiological test results support clinicians in refining therapeutic strategies, thereby shortening the infection control cycle. However, multiple studies have reported low rates of microbiological specimen submission prior to initiating antibiotic therapy. A multicenter study in Europe found that only 50% of patients had specimens submitted before antibiotic administration [6]. Similarly, a study conducted in Beijing, China, showed that among patients receiving carbapenem therapy, the preantibiotic microbiological submission rate was only 64.76% [7].

The accuracy and timeliness of pathogen detection rely heavily on standardized and efficient specimen submission workflows. Many countries have established tailored guidelines and policies [8]. Specimens should be obtained prior to initiating antibiotic therapy to ensure timely microbiological culture [9]. For example, China has included the microbiological specimen submission rate for high-priority antibiotics as a core quality indicator in *Healthcare-Associated Infection Management Quality Control Indicators (2024 Edition*) [10]. In clinical practice, multiple challenges hinder preantibiotic specimen submission, including insufficient awareness among clinicians, delays in specimen collection by nursing staff, and transport inefficiencies [11]. As the specimen collection and submission process involves multiple departments and procedural steps, its complexity has hindered improvement of submission rates [12]. Currently, there is a limited understanding of risk factors across the entire microbiological specimen submission pathway.

Failure mode and effects analysis (FMEA), as a risk management tool, has been extensively used in research on optimizing and refining hospital management processes [13,14]. The core principle of FMEA involves assembling an interdisciplinary team to systematically identify potential failure modes within a workflow, analyze their causes and consequences, and prioritize high-risk steps by calculating the Risk Priority Number (RPN), thereby guiding the development of targeted improvement interventions. For example, Yang et al [15] applied FMEA to hospital infection prevention in a stomatology department by mapping the clinical workflow, identifying potential failure modes through expert consensus, and ranking high-risk steps using RPN to inform targeted control strategies. Chiereghin et al [16] used FMEA to improve the quality of an organized colorectal cancer screening program. These studies highlight FMEA's value in systematically dissecting complex clinical processes and guiding resource allocation toward high-impact improvements. However, its application to microbiological specimen submission within antimicrobial stewardship programs remains limited, particularly when integrated with hospital digital infrastructure.

Digital health and information systems can support this process by embedding real-time alerts into electronic prescribing workflows, prompting clinicians to submit microbiological specimens before initiating antibiotic therapy. Integrated informatics tools, such as electronic health records (EHRs), computerized physician order entry, barcode-based specimen tracking systems, and clinical decision support systems (CDSS), enable end-to-end, closed-loop management of specimen collection, transport, and laboratory processing. However, digital health and information systems may also impose additional workload on health care professionals [17]. Therefore, developing feasible implementation strategies and continuously evaluating the real-world performance of these technologies are essential to ensure sustainable optimization.

Therefore, this study aimed to identify the failure modes contributing to low microbiological specimen submission rates in a tertiary hospital of Chinese Medicine using FMEA and to develop and evaluate targeted interventions to improve specimen submission practices.

Methods

Study Setting and Design

This study was conducted at Zhejiang Provincial Hospital of Chinese Medicine, a general tertiary hospital located in Hangzhou, Zhejiang Province, China. Applying FMEA to assess risks in preantibiotic therapy microbiological specimen submission, formulate targeted interventions, and evaluate their effectiveness. A complete checklist is available in Checklist 1. The implementation steps are as follows (Figure 1):

XSL•FO
RenderX

Figure 1. Schematic figure of Failure Mode and Effects Analysis for microbiological specimen submission. IPC: infection prevention and control.

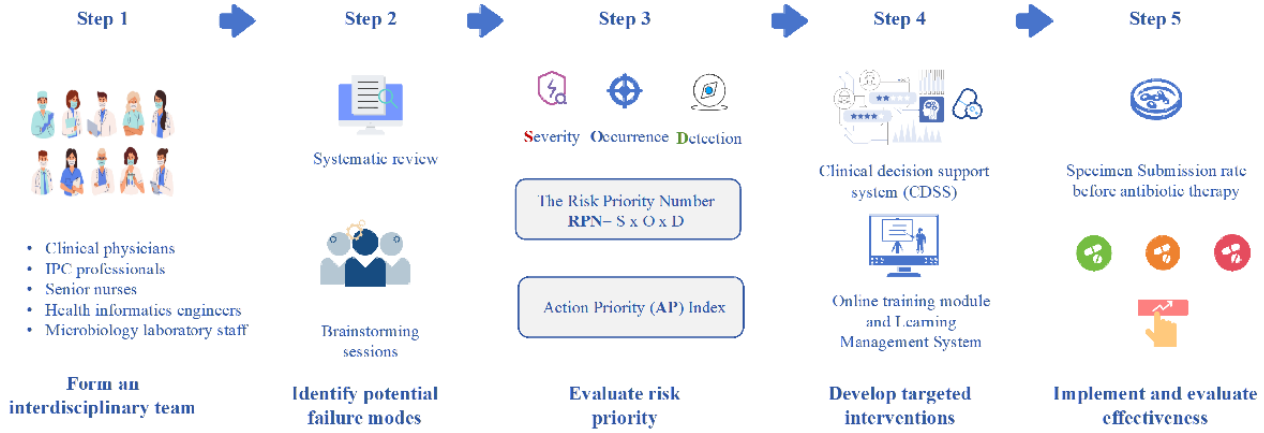

Step 1: Form an Interdisciplinary Team

A multidisciplinary expert team was established to ensure comprehensive identification of workflow risks across clinical, laboratory, managerial, and informatics domains. The team consisted of 9 members, including clinical physicians, Disease Prevention and Control Experts, infection prevention and control (IPC) professionals, microbiology laboratory staff, senior nurses, and health informatics engineers. Team members were selected based on predefined criteria: (1) at least 5 years of professional experience in their respective fields; (2) direct involvement in antimicrobial stewardship, specimen management, or infection surveillance activities; (3) familiarity with routine workflows and known bottlenecks within the hospital; and (4) willingness to participate in structured meetings, independent FMEA scoring, and consensus discussions. This interdisciplinary composition ensured that all key steps of the specimen submission process, from clinical decision-making to specimen collection, laboratory processing, and information system operation, were adequately represented, providing a balanced and comprehensive perspective for risk analysis.

Step 2: Identify Potential Failure Modes

We first conducted a systematic review of the literature to identify common influencing factors in the microbiological specimen submission workflow (Multimedia Appendix 1). Commonly reported factors affecting microbiological specimen submission were extracted and categorized into four predefined domains: (1) health information systems factors, (2) personnel-related factors, (3) administrative and management factors, and (4) external support factors. This framework was used to guide subsequent discussions. Building on this framework, the multidisciplinary expert panel participated in 2 rounds of structured brainstorming sessions. During each session, experts were instructed to identify specific, concrete scenarios within each domain that could plausibly lead to specimen submission failure or delay. To enhance consistency, failure modes were defined as observable deviations from the expected workflow (eg, missed test ordering, delayed collection, incomplete data capture), rather than abstract causes or outcomes. All proposed failure modes were collated, merged, and refined by the research team. Redundant items were consolidated, and ambiguous descriptions were clarified through iterative discussion with panel members. Items were retained only if consensus was reached that the scenario was (1) operationally distinct, (2) locally relevant to routine practice in the study hospital, and (3) theoretically supported by either literature evidence or expert experience. A final set of 30 locally relevant failure modes was generated. These items formed the basis for subsequent FMEA scoring and prioritization.

Step 3: Evaluate Risk Priority

Before formal scoring, all team members received standardized training on the FMEA scoring criteria provided by the research team. Each member independently evaluated the Severity (S), Occurrence (O), and Detection (D) of each potential failure mode. After all 9 members completed their independent assessments, a consensus meeting was convened to discuss discrepancies and clarify scoring decisions. Final S, O, and D scores for each failure mode were determined through group consensus. The RPN was calculated using the formula: RPN=S×O×D.

The scoring range for S, O, and D was 1 - 10. Conventionally, corrective actions are mandated when the RPN exceeds 125. To address limitations of the RPN method, we introduced the Action Priority (AP) index. The determination of AP followed the standard methodology outlined in the Automotive Industry Action Group–Verband der Automobilindustrie FMEA guidelines. Unlike the RPN, the AP level is not calculated through a mathematical formula; instead, it is assigned by consulting the guideline's AP lookup table, which specifies the corresponding action level based on different combinations of Severity (S), Occurrence (O), and Detection (D). As shown in Multimedia Appendix 2, AP levels are categorized as high, medium, or low. High AP indicates that corrective actions should be taken with priority, medium AP suggests that action is recommended but can be scheduled, and low AP indicates that existing controls are acceptable and require only routine monitoring. High-priority failure modes identified by the FMEA (based on both RPN and AP indices).

Step 4: Develop Targeted Interventions

For each mode, the multidisciplinary team collaboratively designed corresponding digital health solutions. All digital interventions underwent iterative refinement through pilot testing in selected departments. Feedback from clinicians, nurses, infection control practitioners, and IT personnel was

XSL•FO
RenderX

collected to optimize alert thresholds, interface design, workflow integration points, and data visualization formats.

The final intervention package consisted of three interrelated components:

1. An artificial intelligence-driven (DeepSeek-R1) CDSS was developed and integrated into the hospital's EHR and computerized physician order entry system to promote microbiological specimen submission before antibiotic therapy. When a clinician initiates an order for therapeutic antibiotics, particularly restricted-use or special-use agents, the system automatically evaluates whether a corresponding microbiological specimen has been ordered or collected within a predefined time window.
2. Online training module and learning management system (LMS): The module was jointly created by the Infection Prevention and Control Department, the Microbiology Laboratory, and the IT Department. The training curriculum covered standardized procedures for microbiological specimen collection, requirements for specimen transport, electronic order entry workflows, and regulations related to preantibiotic specimen submission.
3. Supplementary digital process-enhancement measures: Additional supportive measures included enhanced barcode-based specimen tracking and monitoring dashboards with feedback mechanisms. Barcode-based tracking strengthened specimen traceability during collection and transport, reducing data loss caused by scanning failures or workflow deviations. Dashboards aggregated compliance data by department and antibiotic category, generating periodic feedback reports for clinical units and infection control teams to facilitate continuous quality improvement.

Step 5: Implement and Evaluate Effectiveness

After developing the intervention measures, we proceeded with the implementation phase following the standard quality improvement process. The preparatory stages of FMEA, including workflow mapping, identification and scoring of failure modes, prioritization, and intervention design, were completed between January 1 and March 31, 2024. Beginning on April 1, 2024, the FMEA-informed interventions were formally implemented. Physicians, nurses, and support staff received training on the updated specimen submission workflow and information system functions through the online learning platform. The effectiveness evaluation covered the period from January to December 2024. Improvements were assessed by comparing pre- and postintervention microbiological specimen submission rates.

Three core outcome indices were evaluated:

1. Specimen submission rate before antibiotic therapy of inpatients (%)=number of inpatients with pathogen submission before antibiotic therapy/number of inpatients with therapeutic application of antibiotics during the same period *100%.
2. Specimen submission rate before restricted-use antibiotic therapy of inpatients (%)=number of inpatients with pathogen submission before restricted-use antibiotic therapy/number of inpatients with therapeutic application of restricted-use antibiotics during the same period *100%.
3. Specimen submission rate before special-use antibiotic therapy of inpatients (%)=number of inpatients with pathogen submission before special-use antibiotic therapy/number of inpatients with therapeutic application of special-use antibiotics during the same period *100%.

Data Collection

Specimen submission data from January 1, 2024, to December 31, 2024, was extracted from the Hospital Infection Surveillance System (Xinglin System). The dataset included antibiotic order time, drug category (restricted-use, special-use), and timing of administration relative to specimen submission. To ensure data quality and consistency, only complete records with clearly documented timestamps for both specimen submission and antibiotic prescribing were included in the analysis. Duplicate entries, canceled orders, and specimens rejected by the laboratory due to quality issues were excluded.

Data Analysis

Statistical analyses were performed using SPSS (version 27.0; IBM Corp). Categorical variables were reported as counts (percentages). The Mann-Kendall trend test was used to examine the trend of submission rate over time. When the *z* score was positive, it meant that the composition ratio of each indicator showed an upward trend. When the *z* score was negative, it meant that the composition ratio of each indicator presented a downward trend. Statistical significance was set at 2-tailed $P<.05$.

Ethical Considerations

This study was approved by the Ethics Committee of the First Affiliated Hospital of Zhejiang Chinese Medical University (2025-KLS-038-01). As this study used retrospective data, the ethics committee waived the requirement for participants to provide written informed consent. The research data does not involve individual patient information. No compensation was provided to participants as this study was a secondary analysis of preexisting data with no new patient recruitment or intervention. The research was conducted in accordance with the Declaration of Helsinki.

Results

Hazard Analysis of Specimen Submission

The top five failure modes limiting preantibiotic specimen submission were (1) PDA barcode scanning failures, (2) inadequate clinical decision support alerts, (3) insufficient clinician awareness of testing protocols, (4) suboptimal oversight mechanisms, and (5) patient-related health constraints (Table 1).

XSL•FO
RenderX

Table . Failure modes summary for Failure Mode and Effects Analysis.

Potential failure modes	RPN[a]	AP[b]
Health information system		
Interoperability gaps exist between the inpatient electronic medical records system and the health care–associated infection surveillance system, resulting in inaccurate and delayed data sharing and synchronization between the 2 platforms.	146	Moderate
Failure of PDA barcode scanning or over-rapid scanning operations led to unrecorded specimen collection data in the system, compromising preanalytical traceability.	168	High
Incomplete capture or delayed updating of pathogen detection-related data compromised the integrity of microbiological specimen submission records.	158	High
The absence of proactive alerts for pathogen detection failed to prompt clinicians to initiate microbiological testing before antimicrobial therapy, resulting in prescription restrictions due to noncompliance with stewardship protocols.	70	Low
The hospital information system failed to accurately distinguish the purpose of antimicrobial use (therapeutic vs prophylactic) during medication documentation, undermining the precision of stewardship analytics.	35	Low
The calculation methodology for pathogen detection-related metrics in the information system was not aligned with national standards, resulting in inaccurate statistical outcomes.	84	Low
Hospital personnel		
Delayed communication and coordination among IPC[c] professionals due to ineffective communication channels or the absence of established mechanisms hindered timely interventions.	48	Low
Incomplete data verification by IPC professionals was attributed to the overwhelming volume of data and the inherent limitations of manual review processes, leading to frequent oversights.	32	Low
Misinterpretation of infection protection and control standards by IPC professionals occurred due to frequent updates and inadequate, timely training, resulting in noncompliant practices.	48	Low
Health care providers demonstrated insufficient awareness of pathogen detection protocols, with inadequate understanding of its clinical significance, leading to missed testing opportunities in cases where microbiological analysis was indicated.	126	High
Delays in specimen submission by health care providers were attributed to heavy workloads leading to oversight or missed submissions, as well as suboptimal time window settings that conflicted with clinical workflows.	75	Moderate
Noncompliant specimen collection practices by health care providers, stemming from insufficient familiarity with collection protocols and deviations from aseptic techniques, resulted in specimen contamination or invalid samples.	50	Low

XSL•FO
RenderX

Potential failure modes	RPN[a]	AP[b]
Noncompliant specimen transport practices by health care providers, due to insufficient knowledge of transport protocols and inadequate protective measures during transit, compromised sample integrity.	80	Low
Health care providers mistakenly prescribed therapeutic antibiotics instead of prophylactic regimens, leading to inappropriate antimicrobial use.	80	Low
Ineffective communication between nurses and physicians, due to the absence of established collaboration mechanisms, resulted in suboptimal specimen collection practices, as nurses lacked clear guidance on proper techniques and timing, leading to noncompliant samples.	96	Low
IT personnel failed to facilitate tripartite coordination, lacking effective communication and collaboration mechanisms.	48	Low
IT personnel lacked sufficient knowledge of infection surveillance data extraction guidelines: the complexity of guidelines for building infection surveillance information systems made them difficult to understand.	48	Low
Insufficient expertise among hospital microbiology laboratory personnel and inadequate detection technology levels.	40	Low
Health care administration		
Inadequate regulatory mechanisms for pathogen detection specimen submission: issues include unclear regulatory responsibilities, limited regulatory methods, and insufficient regulatory enforcement.	120	Low
Inadequate performance evaluation mechanisms for pathogen detection specimen submission: evaluation metrics are not linked to health care providers' performance or career advancement, lacking incentives.	48	Low
Inadequate feedback mechanisms for pathogen detection specimen submission: test results are not timely or accurately communicated to health care providers.	48	Low
Inadequate communication and collaboration mechanisms among hospital departments: information flow and interdepartmental coordination are suboptimal.	96	Low
Inadequate training content on pathogen detection specimen submission: issues include outdated content, lack of relevance, or limited training methods.	84	Low
Lack of standardized specimen collection and submission processes: issues include unclear procedures, inconsistent operational standards, or poor process execution.	96	Low
External support		
Some patients may not understand or be willing to cooperate with pathogen detection specimen submission, leading to difficulties in sample collection.	49	Low

XSL•FO
RenderX

Potential failure modes	RPN[a]	AP[b]
Patients' poor health status or physical condition may hinder cooperation, affecting the smooth progress of sample collection (eg, difficulty in expectorating sputum leading to challenges in sputum sample collection).	49	Low
Patients' health status or physical condition may affect the smooth progress of sample collection.	192	Moderate
Limited space in the microbiology laboratory may reduce the efficiency of sample processing. During peak hours, sample backlogs may occur, prolonging testing times.	80	Low
Convenience of specimen collection: health care providers may face space and equipment limitations when collecting samples in patient rooms. For example, the lack of appropriate sputum induction devices may compromise sample quality, reducing motivation for submission.	64	Low
Inadequate cleaning and disinfection of patient rooms: a high pathogen load in the room environment may lead to cross-contamination, interfering with the diagnosis of the patient's own infectious pathogens.	128	Moderate

[a]RPN: Risk Priority Number.

[b]AP: action priority.

[c]IPC: infection prevention and control.

Develop Targeted Interventions

For the potentially high-risk failure modes, recommended interventions were identified with a view to reducing their occurrence or improving their detection (Table 2). An LMS-based online training was implemented. All personnel involved in specimen collection or antibiotic order entry were required to complete the module and pass a competency assessment. A score of at least 80% was considered a passing result. The LMS automatically recorded training completion and assessment outcomes; individuals who did not complete the module or did not achieve a passing score were required to undergo retraining before being allowed to independently perform the related tasks. A real-time CDSS was embedded into the antibiotic prescribing process.

XSL•FO
RenderX

Table . Interventions recommended to improve the specimen submission rate.

High-risk failure modes	Targeted interventions
PDA barcode scanning failures	• Implement dual-verification protocols combining manual entry with barcode scanning, accompanied by real-time error alerts in the mobile application.
Inadequate clinical decision support alerts	• Deploy AI[a]-driven decision support systems to automatically flag high-risk cases requiring preantimicrobial testing, with escalating alert tiers.
Insufficient clinician awareness of testing protocols	• Develop role-specific e-Learning modules with embedded competency assessments. • Integrate protocol reminders into electronic health record order entry workflows.
Suboptimal oversight mechanisms	• Establish automated dashboards tracking testing compliance, with weekly feedback reports to department heads. • Introduce “process champions” in each unit to monitor adherence and facilitate peer-to-peer learning.
Patient-related health constraints	• Create adaptive sampling protocols. • Train nursing staff in motivational interviewing techniques to improve patient engagement.

[a]AI: artificial intelligence.

When clinicians initiated orders for therapeutic antibiotics, particularly restricted-use or special-use agents, the CDSS automatically evaluated whether an appropriate microbiological specimen had been ordered. If noncompliance was detected, the system generated real-time alerts prompting clinicians to initiate specimen submission. This workflow enabled prospective intervention at the point of care, reducing missed submissions caused by time pressure or oversight. Supportive process-enhancement measures were integrated to reinforce execution and oversight. Enhanced PDA-based barcode scanning strengthened specimen traceability during collection and transport, reducing data loss related to scanning failures or rapid operations. In parallel, automated monitoring dashboards aggregated specimen submission compliance data by department and antibiotic category. Regular feedback reports were provided to clinical units and infection control teams, supporting targeted supervision and continuous workflow optimization.

Effect of FMEA Management on Specimen Submission

Following the implementation of FMEA management in April 2024, significant improvements were observed in the overall specimen submission rate (Figure 2), the submission rate for restricted antibiotics (Figure 3), and the submission rate for special-use antibiotics (Figure 4). The Mann-Kendall trend test revealed that all 3 categories of specimen submission rates exhibited a statistically significant upward trend (P<.001).

XSL•FO
RenderX

Figure 2. Changes in the specimen submission rate before antibiotic therapy among in-patients, 2024.

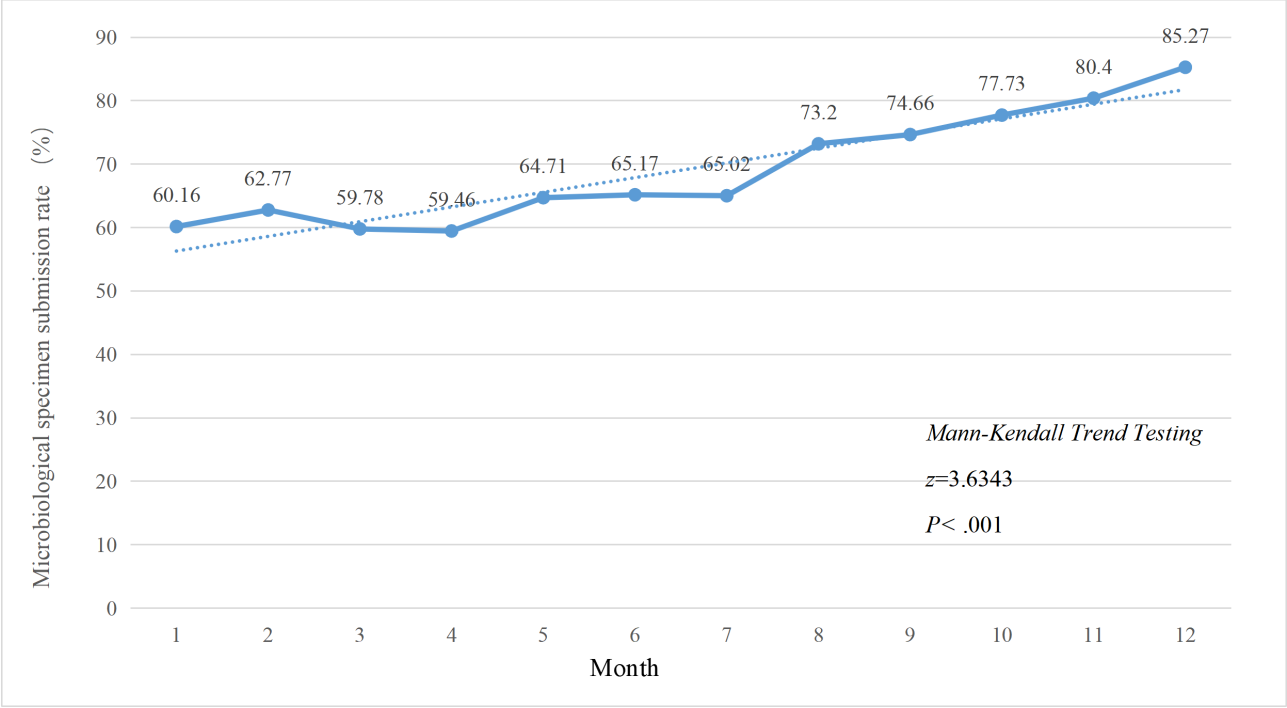

Figure 3. Changes in the specimen submission rate before restricted-use antibiotic therapy among in-patients, 2024.

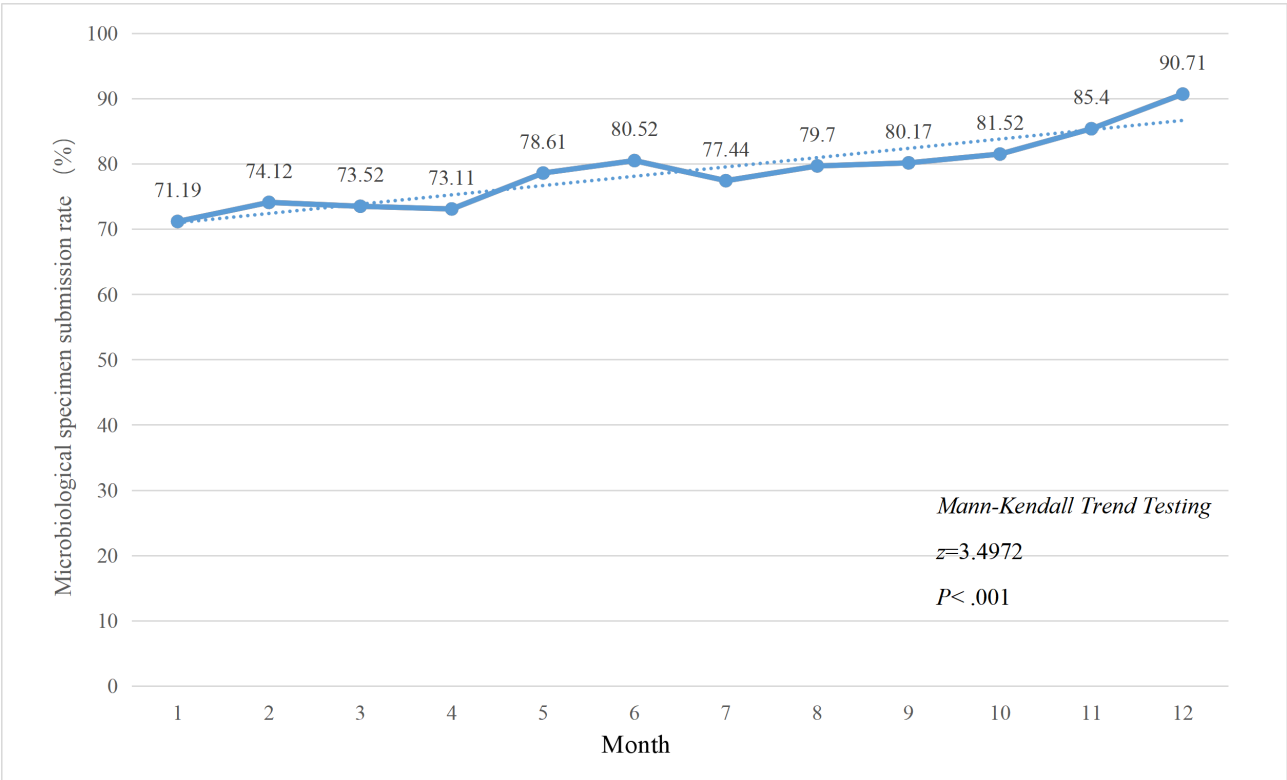

XSL•FO
RenderX

Figure 4. Changes in the specimen submission rate before special-use antibiotic therapy among in-patients, 2024.

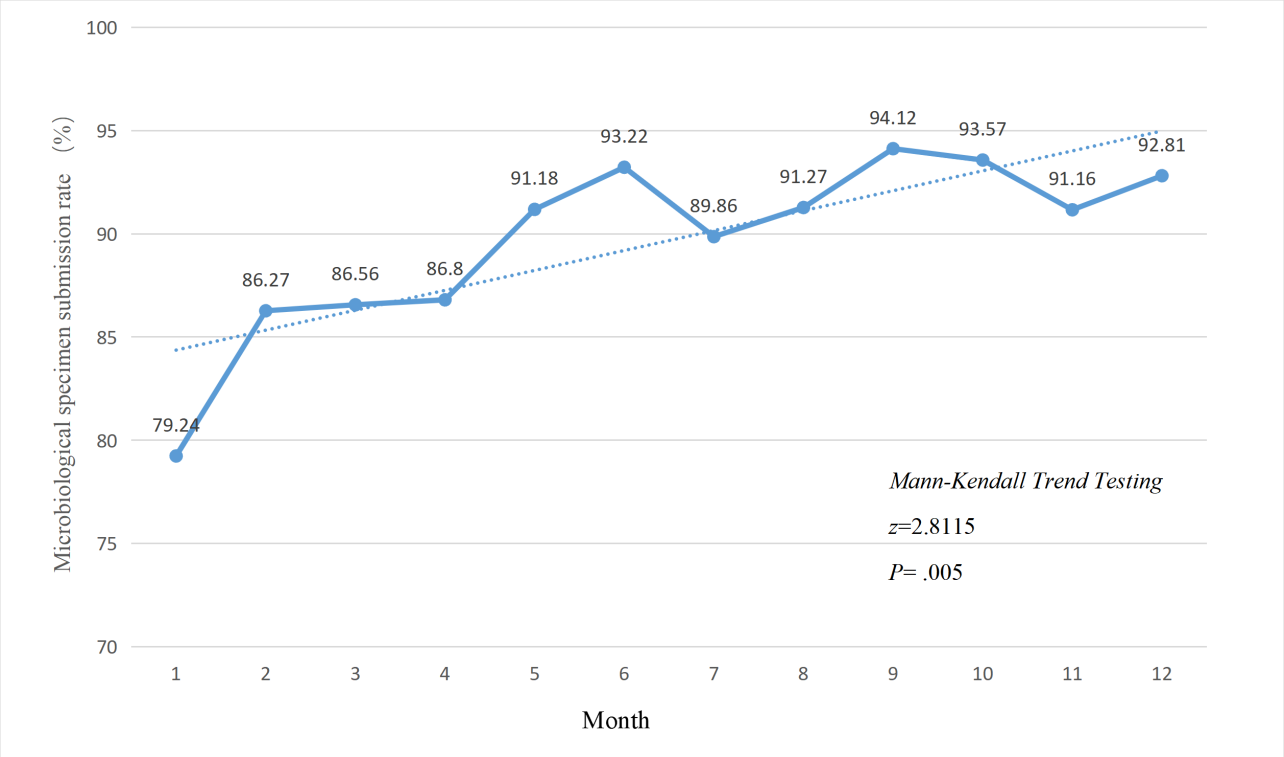

Discussion

Principal Findings

Microbiological specimen submission plays a pivotal role in hospital infection control and promoting rational antibiotic use, representing a complex systemic endeavor [18,19]. This study implemented the internationally recognized risk assessment tool—FMEA, to conduct a multidimensional evaluation of the entire etiological specimen submission workflow, spanning information systems, health care personnel, hospital administration, and external support. Through expert panel discussions, the research identified multiple failure modes and prioritized targeted intervention strategies, providing scientifically grounded decision-making support for improving the quality of specimen submission practices. This study provides a replicable pathway and methodological framework for improving microbiological specimen submission workflows, and may indirectly support hospital infection prevention efforts by enhancing antimicrobial stewardship and strengthening infection surveillance.

The advancement of intelligent health care information systems has emerged as a critical driver for improving specimen submission rates [20]. FMEA-based investigation identified PDA barcode scanning failures and insufficient CDSS alerts as high-priority intervention targets. Mechanistically, real-time alert mechanisms within information systems can trigger specimen collection reminders before antimicrobial administration, thereby minimizing workflow delays caused by human oversight and ensuring precise temporal control of submissions [21]. Concurrently, integrated multisource data platforms, by consolidating heterogeneous data from EHRs, laboratory results, and other sources, enable dynamic decision-making models that assist clinicians in accurately evaluating the necessity of testing indications and the specificity of required assays [22]. These findings reveal a dual-driven mechanism of information system optimization: process monitoring reduces operational deviations, while data integration enhances diagnostic decision-making efficacy.

This study also revealed that inadequate clinician awareness of specimen submission constitutes a critical bottleneck in improving submission quality. To address this, the research team implemented a dual-track capacity-building framework, “theoretical training+practical supervision,” alongside intelligent system upgrades: standardized operational training was used to reinforce protocol adherence, while interdepartmental supervision mechanisms established closed-loop quality improvement cycles. Postintervention evaluation demonstrated statistically significant improvements in specimen submission rates (P<.001). Although the observational study design precludes causal inferences regarding direct relationships between interventions and outcome improvements, the synergistic decline across multidimensional surveillance metrics indirectly suggests potential systemic spillover effects of FMEA-based management in hospital infection prevention and antimicrobial stewardship.

Despite achieving improvements in nosocomial infection specimen submission processes and promoting rational antimicrobial use through hospital-wide informatization, several areas for development remain. Externally, budget constraints limit access to advanced yet costly testing modalities, hindering comprehensive testing. Variations in patient case mix create department-specific testing demands that cannot be uniformly addressed. Additionally, diagnosis-related group reimbursement constraints introduce cost-benefit trade-offs that indirectly influence testing decisions. Therefore, hospital management

XSL•FO
RenderX

must continue refining collaborative mechanisms across clinical, technical, and administrative departments. Strategic initiatives should include optimizing test menu configurations through cost-effectiveness analysis, developing disease-specific testing guidelines, and leveraging informatization for continuous monitoring to enhance staff efficiency. These measures will ultimately strengthen AMR control, promote rational antibiotic use, improve surveillance capabilities for emerging pathogens, and drive health care quality toward higher precision and patient safety.

This study has several limitations that warrant consideration. First, some of the interventions relied on a relatively well-developed hospital information infrastructure, such as artificial intelligence-driven CDSS alerts. Institutions with limited digital capacity may need to adopt manual checklists or simplified reminder mechanisms as alternatives. Second, the study was conducted in a tertiary TCM hospital. Although the hospital includes Western medicine departments such as the intensive care unit, respiratory medicine, and emergency medicine, with antimicrobial use patterns comparable to general hospitals, the patient population and service model of TCM hospitals may differ from other health care settings, which may limit generalizability. Nevertheless, the methodological framework of FMEA is highly transferable, and individual institutions may adapt specific interventions according to their local resources and workflows. Additionally, this was a single-center, before and after study without a control group, which imposes inherent limitations on causal inference. Improvements in specimen submission rates may not be attributable solely to the FMEA-based interventions; they may also reflect the hospital's broader emphasis on antimicrobial stewardship, underlying secular trends in workflow optimization, or concurrent quality improvement initiatives or policy changes. Future research should validate the effectiveness of these interventions using multicenter designs with appropriate control groups.

Conclusions

This study validated FMEA as a risk management tool for optimizing preantimicrobial specimen submission workflows. By addressing critical failures like barcode errors and inadequate decision support, FMEA-based interventions significantly improved specimen submission rates for restricted/special-use antibiotics, aligning with national quality standards. Challenges, including cost constraints and patient barriers, necessitate policy-level solutions. This model offers a scalable framework for antimicrobial stewardship, with future research needed to evaluate long-term outcomes and advanced diagnostic integration.

Funding

This research was supported by grants from the University-level Scientific Research Project of Zhejiang Chinese Medical University (2025GJYY41) and Zhejiang Provincial Basic Public Welfare Research Program (LTGY23H260002).

Data Availability

The data used in this research are available upon individual request from the corresponding author.

Authors' Contributions

TYD and LXL contributed to the conception and design of the study. JXW contributed to writing the data analysis and drafted the original draft. WJZ, JH, and BWZ contributed to methodology and data curation. QLG, MZ, XYZ, and YXZ critically revised the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Systematic review of the submission rate of microbiological specimen before antibiotic therapy.
[DOCX File, 142 KB - jmir_v28i1e78118_app1.docx]

Multimedia Appendix 2
Action Priority lookup table.
[DOCX File, 26 KB - jmir_v28i1e78118_app2.docx]

Checklist 1
iCHECK-DH (Guidelines and Checklist for the Reporting on Digital Health Implementations) guidelines and checklist for the reporting on digital health implementations.
[DOCX File, 24 KB - jmir_v28i1e78118_app3.docx]

References

XSL•FO
RenderX

1. The Lancet. Antimicrobial resistance: an agenda for all. The Lancet 2024 Jun;403(10442):2349. [doi: 10.1016/S0140-6736(24)01076-6]
2. Painter C, Limmathurotsakul D, Roberts T, et al. Sustainable antimicrobial resistance surveillance: time for a global funding mechanism. Lancet Infect Dis 2025 Feb;25(2):e99-e103. [doi: 10.1016/S1473-3099(24)00649-2] [Medline: 39706207]
3. Naghavi M, Vollset SE, Ikuta KS, et al. Global burden of bacterial antimicrobial resistance 1990–2021: a systematic analysis with forecasts to 2050. The Lancet 2024 Sep;404(10459):1199-1226. [doi: 10.1016/S0140-6736(24)01867-1]
4. Lin L, Sun R, Yao T, Zhou X, Harbarth S. Factors influencing inappropriate use of antibiotics in outpatient and community settings in China: a mixed-methods systematic review. BMJ Glob Health 2020 Nov;5(11):e003599. [doi: 10.1136/bmjgh-2020-003599] [Medline: 33184066]
5. Lim C, Ashley EA, Hamers RL, et al. Surveillance strategies using routine microbiology for antimicrobial resistance in low- and middle-income countries. Clin Microbiol Infect 2021 Oct;27(10):1391-1399. [doi: 10.1016/j.cmi.2021.05.037] [Medline: 34111583]
6. Schuts EC, Hulscher M, Mouton JW, et al. Current evidence on hospital antimicrobial stewardship objectives: a systematic review and meta-analysis. Lancet Infect Dis 2016 Jul;16(7):847-856. [doi: 10.1016/S1473-3099(16)00065-7] [Medline: 26947617]
7. Zou Q, Zhang Y, Ai X, et al. Impact of multidisciplinary collaboration on microbiological specimen submission and appropriate use of carbapenems: a pre-post cohort study. Infect Drug Resist 2025;18:4301-4309. [doi: 10.2147/IDR.S532011] [Medline: 40895614]
8. Centers for Disease Control and Prevention. Core elements of hospital antibiotic stewardship programs[r]. : US Department of Health and Human Services, CDC; 2019 URL: https://www.cdc.gov/antibiotic-use/hcp/core-elements/hospital.html [accessed 2026-02-17]
9. Sánchez-Romero MI, García-Lechuz Moya JM, González López JJ, Orta Mira N. Collection, transport and general processing of clinical specimens in microbiology laboratory. Enferm Infecc Microbiol Clin (Engl Ed) 2019 Feb;37(2):127-134. [doi: 10.1016/j.eimc.2017.12.002] [Medline: 29426791]
10. Circular on the issuance of medical quality control indicators for emergency medicine and five other specialties (2024 edition) by the General Office of the National Health Commission [EB/OL]. China's National Health Commission. URL: https://www.nhc.gov.cn/yzygj/c100068/202405/7d1d82090e8b44d2af3339d7ee8c7cca.shtml [accessed 2026-02-17]
11. Mokrani D, Chommeloux J, Pineton de Chambrun M, Hékimian G, Luyt CE. Antibiotic stewardship in the ICU: time to shift into overdrive. Ann Intensive Care 2023 May 6;13(1):39. [doi: 10.1186/s13613-023-01134-9] [Medline: 37148398]
12. Wang Q, Zhang X, Zheng F, Wang L, Yu T. Clinicians' intention to submit microbiological pathogenic test before antibiotics use and its influencing factors: new evidence from the perspective of hospital management. IDR 2022;Volume 15:3013-3023. [doi: 10.2147/IDR.S366892]
13. Prieto-Molina A, Aranda-Gallardo M, Moya-Suárez AB, Rivas-Ruiz F, Peláez-Cherino J, Canca-Sánchez JC. Patient transfers from emergency departments to other in-hospital areas: a failure mode and effects analysis. Emergencias 2023 Dec;35(6):456-462. [doi: 10.55633/s3me/E010.2023] [Medline: 38116970]
14. Weber L, Schulze I, Jaehde U. Using failure mode and effects analysis to increase patient safety in cancer chemotherapy. Res Social Adm Pharm 2022 Aug;18(8):3386-3393. [doi: 10.1016/j.sapharm.2021.11.009] [Medline: 34838476]
15. Yang S, Yang L, Chuang YC, Asihaer G, Lin X. Identification of key risk areas and failure modes using the FMEA method for the prevention and control of major infectious diseases in a stomatology department. BMC Oral Health 2025 Jan 8;25(1):45. [doi: 10.1186/s12903-024-05321-3] [Medline: 39780161]
16. Chiereghin A, Squillace L, Pizzi L, Bazzani C, Roti L, Mezzetti F. Applying the healthcare failure mode and effects analysis approach to improve the quality of an organised colorectal cancer screening programme. J Med Screen 2024 Jun;31(2):70-77. [doi: 10.1177/09691413231197300] [Medline: 37674432]
17. Wosny M, Strasser LM, Hastings J. The paradoxes of digital tools in hospitals: qualitative interview study. J Med Internet Res 2024 Jul 15;26:e56095. [doi: 10.2196/56095] [Medline: 39008341]
18. Skodvin B, Wathne JS, Lindemann PC, et al. Use of microbiology tests in the era of increasing AMR rates– a multicentre hospital cohort study. Antimicrob Resist Infect Control 2019;8(1):28. [doi: 10.1186/s13756-019-0480-z] [Medline: 30740221]
19. Lao Y, Tang Q, Zeng J, Gong W, Shen Y. Effect analysis of multi-department cooperation in improving etiological submission rates before antibiotic treatment. Int J Qual Health Care 2023 Jun 6;35(2):mzad018. [doi: 10.1093/intqhc/mzad018] [Medline: 37040384]
20. Zheng F, Wang K, Wang Q, Yu T, Zhang X. The pre-analytical process management status and influencing factors of laboratory test before prescribing antimicrobial in developing country. BMC Health Serv Res 2023 Mar 25;23(1):283. [doi: 10.1186/s12913-023-09243-8] [Medline: 36966281]
21. Catho G, Sauser J, Coray V, et al. Impact of interactive computerised decision support for hospital antibiotic use (COMPASS): an open-label, cluster-randomised trial in three Swiss hospitals. Lancet Infect Dis 2022 Oct;22(10):1493-1502. [doi: 10.1016/S1473-3099(22)00308-5] [Medline: 35870478]
22. Dutey-Magni PF, Gill MJ, McNulty D, et al. Feasibility study of hospital antimicrobial stewardship analytics using electronic health records. JAC Antimicrob Resist 2021 Mar;3(1):dlab018. [doi: 10.1093/jacamr/dlab018] [Medline: 34223095]

XSL•FO
RenderX

Abbreviations

AMR: antimicrobial resistance
AP: Action Priority
CDSS: clinical decision support systems
EHR: electronic health record
FMEA: Failure Mode and Effects Analysis
LMS: learning management system
RPN: Risk Priority Number
TCM: traditional Chinese medicine

Edited by A Schwartz; submitted 28.May.2025; peer-reviewed by O Ajileye, SLL Holgate; accepted 16.Jan.2026; published 10.Mar.2026.

Please cite as:
Wu J, Zhan W, Hua J, Ge Q, Zhu Y, Yu H, Zhao M, Zhan X, Zhu B, Xiang T, Lu L, Dai T
Investigating the Effect of Hospital Infection Control Informatization on Optimizing Microbiological Specimen Submission Before Antibiotic Therapy: Failure Mode and Effects Analysis
J Med Internet Res 2026;28:e78118
URL: https://www.jmir.org/2026/1/e78118
doi:10.2196/78118

© Jianxiong Wu, Weijiang Zhan, Jun Hua, Qinling Ge, Yuexian Zhu, Huixian Yu, Min Zhao, Xiaoyan Zhan, Bingwei Zhu, Tongqi Xiang, Longxi Lu, Tieying Dai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quality of Cancer-Related Clinical Coding in Primary Care in North Central London: Mixed Methods Quality Improvement Project

Afsana Bhuiya[1], BSc, MBBS; Graham Roberts[1], BSc; Katie Tucker[1], BSc, MSc; Stefanie Bonfield[2], BSc, MSc; Georgia Black[2], MSc, PhD

[1]North Central London Cancer Alliance, London, United Kingdom

[2]Wolfson Institute of Population Health, Queen Mary University of London, London, United Kingdom

Corresponding Author:
Afsana Bhuiya, BSc, MBBS
North Central London Cancer Alliance
UCLH, 47 Wimpole Street
London, W1G 8SE
United Kingdom
Phone: 44 7877136123
Email: afsana.bhuiya1@nhs.net

Abstract

Background: The North Central London (NCL) Cancer Alliance carried out a quality improvement (QI) project to fill a distinct knowledge gap regarding the quality of clinical coded data in a primary care electronic health care record system across the whole cancer pathway.

Objective: This study aims to establish the quality of cancer-related clinical coding in NCL primary care, encompassing both quantitative measures (eg, coding completeness and diversity) and qualitative dimensions such as clinical relevance and workflow alignment.

Methods: This was a mixed methods QI project in which we combined an observational dataset review and qualitative data from stakeholder interviews, workshops, and discussions. In the dataset review, we evaluated completeness, diversity, validation, and granularity in cancer clinical coding along the patient cancer pathway, which was split into three domains: (1) patient characteristics and risk factors, (2) cancer screening attendance, and (3) living with cancer. It was conducted in NCL primary care electronic health record systems, covering a population of over 1.4 million adults across 5 boroughs.

Results: Cancer-related clinical coding in NCL primary care revealed significant gaps despite high completeness for ethnicity (912,679/1,055,083, 86.5%) and language (898,023/1,307,601, 68.7%). Employment status (29,848/1,229,644, 2.4%) and family history of cancer (183,424/1,236,580, 14.8%) were underrecorded, with wide variation in coding practices. Screening data showed good alignment with national datasets for cervical and bowel screening but fragmented and inconsistent breast screening data due to a lack of standardized codes. Cancer diagnosis coding was incomplete (4604/5260, 87.5% recorded), and treatment and staging data were almost entirely absent, limiting proactive management of long-term consequences. Stakeholder input highlighted inconsistent template use, limited data updates, and insufficient incentives as key barriers to better coding.

Conclusions: The QI project has provided a detailed insight into the many dimensions of cancer coding and sheds light on many factors that underpin variation and coding preference. We offer a number of recommendations. The prioritized ones include the need for a cancer clinical coding data framework for primary care supported by appropriate funding and incentivization; improvements in the breast screening pathway and its interface with primary care; improvements in the quality of secondary care information that is sent to primary care; and dissemination of the importance of coding of cancer activity in primary care.

(J Med Internet Res 2026;28:e73205) doi:10.2196/73205

KEYWORDS

cancer diagnosis; cancer pathway; cancer risk factors; cancer treatment; cancer; clinical coding; coding completeness; coding data; coding diversity; coding processes; coding quality; coding validation; coding variation; inequalities data; primary care coding; quality improvement; SNOMED CT; Systematized Nomenclature of Medicine – Clinical Terms

XSL•FO
RenderX

Introduction

Overview

Clinical coding of cancer-related data in primary care supports accurate and timely data collection, analysis, and reporting of cancer diagnoses and treatments, which in turn facilitates high-quality patient care [1,2]. Consequently, incomplete or inaccurate clinical coding of cancer-related data has significant implications across the cancer pathway. For example, cancer prevention efforts, including information provision, vaccination, and screening, may be restricted if it is not possible to identify eligible individuals based on data available in primary care records (eg, age, sex, health behaviors, previous medical history) and follow up with those who have not engaged (eg, those who have not responded to previous cancer screening invitations) [3]. Missing data for cancer risk-factors, such as family history or previous medical history, may undermine appropriate referral of symptomatic patients for cancer investigation [4]. Similarly, missing data on precancerous conditions such as Barrett's esophagus or bowel polyps may prevent health care professionals from providing information and support to patients about managing their condition and personal risk and result in patients being excluded from relevant safety-netting efforts and surveillance pathways [5]. Meanwhile, the underreporting of cancer cases can lead to an underestimation of the true cancer burden within a population and limit the ability of primary care to support patients during cancer diagnosis and treatment [5,6]. Beyond clinical impact, poor coding may also contribute to financial losses for health care providers and hinder effective service planning at both practice and system levels [7].

Primary care cancer coding data are often variable and suboptimal, with a poor evidence base for improvement [8-10]. A systematic review by Thiru et al [11], concluded on the lack of standardized measures for data quality, which is supported by previous studies showing the heterogeneity of quality assessment methods in primary care coding [11,12]. Thiru et al [11], uncovered studies that assessed primary electronic patient record data and studies that reviewed survey and questionnaire data. They also found that data quality (reliability) was usually measured with rate comparisons and data validity was expressed under a range of terms (completeness, correctness, accuracy, consistency, and appropriateness), which were rarely defined [11].

Previous studies have identified the need for better communication about patients who have been diagnosed with cancer between primary and secondary care [13], and a UK study reported that 1 in 5 patients with cancer were not recorded to have a cancer diagnosis in primary care records [6]. In England, relatively robust audit systems and regulatory oversight exist to regulate coding for cancer diagnosis, cancer interventions, and procedures in secondary care, underpinned by financial incentives [14,15]. In contrast, equivalent governance mechanisms do not exist in primary care. Clinical entries in primary care rely on SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms), but there are no national standards or frameworks specifically guiding cancer coding across the whole pathway. Instead, coding behavior is shaped primarily by the Quality Outcomes Framework (QOF) [16], which provides financial incentives for documenting selected conditions and activities. For cancer, QOF incentivizes the coding of a cancer diagnosis and the completion of cancer care reviews (CCRs) following diagnosis—although CCRs were removed from general practitioner (GP) QOF contracts in April 2025 [16,17]. Importantly, QOF does not cover the full cancer pathway. Moreover, primary care coding systems often encourage diversity rather than consistency, such as having multiple codes to describe identical clinical events [10,18] (eg, "smoker," "cigarette smoker," "moderate cigarette smoker"). Lack of regulation and research evidence about cancer coding consistency and variation restricts the development of targeted quality improvement (QI) measures. Consequently, there is a need to understand how the quality of clinical coding data in primary care varies to influence outcomes across the cancer pathway. This holistic approach can inform suitable interventions for optimizing the quality of primary care coded data to deliver valuable improvements in cancer prevention, referral, diagnosis, and treatment outcomes [13].

Cancer Coding and Data Curation

The process of how patient information is recorded, updated, and monitored in electronic health records is sometimes called "data curation." Primary and secondary care data curation is influenced by multiple drivers, which in turn affect data quality. In the United Kingdom, the National Cancer Registration and Analysis Service (NCRAS) [19] provides guidance and support for clinical coding of cancer diagnoses and treatment in acute care. NCRAS has developed a set of coding standards that are incorporated in nationally commissioned datasets such as the Cancer Outcomes and Services Dataset and Systemic Anti-Cancer Therapy dataset, which provide guidance on the coding of cancer diagnoses and treatments. The National Health Service (NHS) primary care coding system is called SNOMED CT [20]. The SNOMED CT does not have standard clinical coding specifications for cancer data.

Acute and mental health trusts have standard procedures for regular quality inspections of their coded clinical data for inpatient and day-case episodes by approved clinical coding auditors, who aim to demonstrate compliance with national clinical coding standards [21]. The regulatory component is important, as it supports high-quality data collection that supports secondary uses of the data, such as collaborations with academic and research departments (eg, the "Getting It Right First Time" program to reduce unwarranted variation) [22]. There is no equivalent process for auditing the assignment of the terminology SNOMED CT, which is used in primary care.

Reasons for primary care coding incompleteness and inconsistencies are well documented and include time pressures, finding the right code, and motivation to code [12,23]. As part of the development work that led to this study, NCL Cancer Alliance conducted an online survey of GP respondents in London to understand barriers to good-quality clinical coding in primary care. The findings included lack of standardized coding practices, scarcity of dedicated staff time to code, and inadequate training around coding (for the full list of barriers see Table S2 in Multimedia Appendix 1 [24]).

XSL•FO
RenderX

Context for the Quality Improvement Project

The aim of this QI project was to assess the completeness and variation of clinical cancer coding in NCL primary care data and to understand the reasons for this. The output of this QI project was to develop a robust "case for change" [25] for relevant improvement solutions. The long-term goal of the project was to improve service planning, pathway delivery, and redress health inequalities. The QI project was carried out by the NCL Cancer Alliance team with support from researchers at the Queen Mary University of London.

Our objectives were to: first, examine the quality of clinical coding of cancer-relevant risk factors, processes, and outcomes, encompassing both quantitative measures (eg, coding completeness and diversity) and qualitative stakeholder perspectives (eg, barriers, enablers, clinical relevance, and workflow alignment); second, use data from the first objective to develop recommendations to support data improvement in primary care.

Methods

Overview

This was a mixed methods QI project combining an observational dataset review and stakeholder interviews, workshops, and discussions. The observational dataset and analysis were based on data searches designed in EMIS (Egton Medical Information Systems) Web by Enfield GP Federation. These were set up to take a snapshot of information held in GP records as it stood on October 31, 2023, so that results are consistent and comparable across practices. All participating GP federation teams then ran the searches between November 2023 and January 2024. The results were shared with the NCL Cancer Alliance, which completed the data analysis in April 2024. Preliminary insights from the early stages of qualitative data collection (eg, advice from group interviews) informed the quantitative analysis (eg, process barriers in breast screening to ethnicity codes recorded in 2 different parts of EMIS Web) and the development of the workshop themes and questions. The formal qualitative analysis was conducted after all qualitative data had been collected; this section of analysis was carried from October to December 2024 (Figure S1 in Multimedia Appendix 1 depicts the project timeline).

Setting and Participants

We gathered data on clinical coding from NCL primary care GP systems covering an adult population of >1.4 million. The qualitative data included email conversations, 2 semistructured interviews, and 2 workshops with key primary care stakeholders. The qualitative methods and analyses are reported according to the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist (Table S3 in Multimedia Appendix 1) [24].

Data Collection

We extracted GP data from the electronic health care system, EMIS Web [26], the sole GP electronic health care record provider for all GP practices in NCL. Data were obtained using built-in "searches" within EMIS Web. These are configurable protocols designed to retrieve coded patient information based on predefined clinical or demographic criteria, referred to as data domains (eg, ethnicity, smoking status, cancer diagnosis, or treatment history). Each search identifies patients meeting the selected criteria based on structured clinical codes. The output of these searches is presented in the form of "reports," which summarize the number of patients meeting each criterion and can be exported for further analysis. Figure S2 in Multimedia Appendix 1 provides an example of a search. We used the SNOMED CT [20,27] and EMIS Web clinical codes [28], which coexist in patients' records [29]. The NCL GP federations performed the searches; a GP federation is a group of general practices working collaboratively as an organizational entity to improve patient care, share resources, and enhance service provision within the local health economy [30].

Table 1 illustrates the number of GP practices across each GP federation or primary care network (PCN) in NCL and the completeness of report returns across the 26 searches that were built and run. A PCN in Islington that was not part of the Islington GP Federation (Islington North 2) did not participate in this project. A partial return is where some GP practices have generated results in a search and some have not due to there either being no patients who meet the criteria or technical constraints that prohibit the search from running for particular practices.

Table 1. Shows the number of general practitioner (GP) practices within each GP federation or primary care network (PCN) and the completeness of the reports that were requested.

Category	GP entity						
	Barnet[a], n	Camden[a], n	Camden Health Evolution[a], n	Enfield[a], n	Haringey[a], n	Islington[a], n	Islington North 2[b], n
Number of GP practices	48	23	9	30	34	23	8
Report status							
Complete	7	21	21	13	21	21	0
Partial	17	5	5	13	4	4	0
Did not return	2	0	0	0	0	0	26
Search generated null results	0	0	0	0	1	1	0

[a]Represents GP federations.

[b]Represents a PCN.

XSL•FO
RenderX

Three personalized cancer care metrics—cancer care plan given, end-of-treatment summary, and holistic needs assessment—were unintentionally omitted from the original data request. This gap was identified by the NCL Cancer Alliance after data had already been submitted by all boroughs. To partially address this, Enfield Federation, which had local access, conducted the relevant searches for its own borough (covering the 12 months up to October 2023). Due to timing and resource constraints, it was not feasible to repeat this process across the other boroughs. Enfield's data were compared with HealtheIntent cancer registries [31], which span the entire NCL adult population; while full coding quality could not be assessed, we were able to examine the relative frequency of these codes across populations.

Data Quality Assessment Method

We followed the approach taken by Pineda-Moncusi et al [32], who examined ethnicity data at a large scale and defined data quality across completeness, coverage, and granularity (most prevalent clinical codes used).

We drew on existing cancer data frameworks [33,34] to conceptualize different cancer pathway stages and identify relevant cancer codes. Within cancer alliances, there are established programs [35] spanning the entire pathway—from awareness and prevention through to living with and beyond cancer. Our aim in this study was to create a comprehensive and holistic dataset to review, thereby informing future recommendations. To do this, we built on this existing knowledge, incorporated earlier work from our team [36], and collaborated with our wider NCL Cancer Alliance team to scope data items across each pathway stage. Although we recognized that some of these items were unlikely to be routinely coded in primary care, there was no empirical evidence to confirm this; therefore, part of the purpose of this QI project was to assess the current state of coding. End-of-life care was deemed out of scope for this work. The Enfield GP Federation digital team built EMIS Web searches to cover each element of the pathway and shared these with other NCL GP federations' IT teams to run for each borough. Enfield GP Federation's IT team oversaw the communication, search development, search dissemination, and data submissions. Raw data were transferred to the NCL Cancer Alliance.

The data protection officer for NCL primary care assured that all data sharing complied with UK General Data Protection Regulation.

Qualitative Data Collection

A convenience sample of key stakeholders with expertise in clinical coding were invited (over email and through a GP bulletin) to attend an online group workshop through existing contacts (including GPs, project and program managers, IT staff, and academic researchers). They were told that they were being invited to discuss the quantitative findings relating to clinical coding. We conducted 2 semistructured group interviews and 2 workshops with 11 primary care stakeholders. We also reviewed email correspondence from the Enfield GP Federation team, which captured responses to queries arising from the initial round of quantitative data analysis. These communications provided a systematic method for clarifying data gaps and process-related issues and directly informed the development of key questions explored during the subsequent stakeholder workshops.

All interviews and workshops were held remotely over Microsoft Teams between April and August 2024 (Table S4 in Multimedia Appendix 1) and conducted by AB (female, GP clinical lead for Innovation and Integration at NCL Cancer Alliance). Sessions were also facilitated by 2 other members of the research team, GR (male, head of Data and Analytics, NCL Cancer Alliance) and KT (female, senior innovation consultant, NCL Cancer Alliance). Semistructured interview and workshop topic guides were developed by AB, KT, and GR. We presented key findings for discussion. All remote sessions were video recorded and transcribed using the Microsoft Teams record and transcription functions. Transcripts were not shared with stakeholders.

Analysis

Quantitative data were analyzed descriptively across the different boroughs to characterize coding patterns. In total, 26 searches were built and run. We analyzed these data domains: patient ethnicity, main language spoken, weight or BMI, alcohol consumption, smoking status, family history of cancer, employment status, environmental pollutants exposure, carer, cancer screening attendance, cancer fast-track referrals, presence of malignant neoplastic disease, treatment regimen, and attendance at CCR. To ensure transparency and facilitate replication as far as possible, we have included these extracted search terms in Textbox S1 in Multimedia Appendix 1.

Codes were descriptively analyzed for the following features (1) coding completeness, (2) coding diversity, (3) data validation, and (4) granularity of coding (Table 2).

XSL•FO
RenderX

Table 2. Definitions and methods used to assess clinical coding quality and completeness.

Definitions	Methods
Coding completeness[a]	Percentage of eligible patients with a relevant SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) or EMIS (Egton Medical Information Systems) clinical code captured. This reflects the presence or absence of a code rather than the true or expected prevalence of the underlying condition.
Coding diversity[b]	Number of unique SNOMED CT or EMIS codes used to describe the same information (eg, patient's weight). This serves as a practical proxy for coding breadth, acknowledging limitations such as inaccessible child codes and variation in code list size.
Data validation	Comparison of coding completeness in EMIS Web with national or local reference datasets to assess missed or uncoded data. Calculated as EMIS Web completeness divided by completeness reported in the comparator dataset for the same coded item. Ratios may exceed 100% if EMIS Web demonstrates higher apparent completeness. Comparator datasets included HealtheIntent, Cancer Waiting Times, the National Cancer Registry, South East London Cancer Alliance data[c], and the NHS Futures Screening Dashboard.
Granularity of coding	Review of the most frequently used codes to assess specificity, for example the use of general codes such as "current smoker."

[a]Coding completeness describes the proportion of eligible patients in the denominator population who have a relevant code captured in their primary care record. This definition is intended to reflect the presence or absence of a code rather than to imply that the observed percentages represent the true or expected prevalence of the underlying clinical characteristic or diagnosis. For example, a coding completeness of 0.3% for malignant neoplastic disease reflects the proportion of the registered population with a relevant cancer code, not an assessment of whether this proportion is "correct" or "incorrect." To distinguish between the technical measurement of code capture and the interpretive question of whether coding levels are as expected, we also applied the concept of coding validation. Coding validation involves comparing observed coding completeness against external standards or datasets (eg, cancer wait times) to assess whether the recorded levels are appropriate and consistent with known population rates. In this way, completeness provides a descriptive measure of the presence of codes in primary care data, while validation enables assessment of their adequacy and alignment with clinical or epidemiological expectations.

[b]Coding diversity was defined as the number of unique codes captured in EMIS Web for each search. We recognize this is a practical proxy rather than a full measure, as EMIS Web does not easily expose all child codes, and many codes (eg, for rare conditions or languages) will naturally yield zero results. Some code lists are also inherently larger than others. A distinct count of unique codes therefore provides useful context on the breadth of coding options observed while acknowledging these limitations.

[c]South East London Integrated Care Board (ICB) and Cancer Alliance has its own population health dashboard that overlaps with the data and definitions in this study, meaning it is valid for comparison where we have no other published source. This dashboard is not publicly available.

Data completeness and validation analysis was carried out against all submitting GP practices' adult populations as of January 2024 (Table S5 in Multimedia Appendix 1). Practice populations are relatively stable month to month, so comparing October 2023 search results with the adult population in January 2024 is considered valid. Practices that did not submit data for a profile were excluded from the analysis to maximize data integrity (Table S6 in Multimedia Appendix 1).

Table 3 lays out the report names across the time frames for which the data was searched for, the denominator population, and the validation database used for comparison. Additionally: (1) data on body weight and BMI were assessed through a combined height and weight search); (2) smoking status was assessed through three separate searches; (3) family history of cancer was assessed based on any recorded code, rather than limiting analysis to the preceding 24 months (as the GP federations' IT team were aware this would be captured at one point in records); and (4) breast screening data were retrieved through four distinct searches: screening attendance, normal results, abnormal results, and cancer detected.

XSL•FO
RenderX

Table 3. Each report name shown against time period covered for each report, the denominator population, and the validation database used for comparison, including the validator time period.

Report name	Time period	Denominator population characteristics (age in years)	Validator database	Validator metric and time period
Ethnic origin	Recorded ever	>18	HealtheIntent	Ethnicity coding (April 2024)
Main language spoken	Recorded ever	>18	HealtheIntent	Main language spoken (April 2024)
Employment status	October 2021-October 2023	>18	No comparator	__[a]
Weight or BMI recorded	October 2021-October 2023	>18	South East London Cancer Alliance data completeness	Weight or BMI recorded (April 2024)
Current smoker (recorded in last 24 months)	October 2021-October 2023	>18	HealtheIntent	Current smoker status (April 2024)
Any smoking status (recorded in last 24 months)	October 2021-October 2023	>18	South East London Cancer Alliance data completeness	Any smoking status (April 2024)
Alcohol consumption record	October 2021-October 2023	>18	South East London Cancer Alliance data completeness	Alcohol consumption record (April 2024)
Family history of neoplasm	Recorded ever	>18	No comparator	—
Environmental pollutants	October 2021-October 2023	>18	No comparator	—
Fast track referral coding	New episode added in last 12 months	>18	Cancer Wait Times	Urgent suspected cancer referrals (2023)
Cancer–bowel: did not return screening kit	October 2023-2 years 6 months	60-74	No comparator	—
Cancer–bowel: screening uptake	October 2023	60-74	NHS Futures Screening Dashboard	Bowel screening uptake August 2023 (50-70 years)
Cancer–bowel: abnormal result	October 2023-2 years 6 months	60-74	No comparator	—
Cancer–breast: screened	October 2023-3 years 6 months	50-70 (female)	NHS Futures Screening Dashboard	Breast screening uptake August 2023 (50-70 years)
Cancer–breast: abnormal result	October 2023-3 years 6 months	50-70 (female)	No comparator	—
Cancer–breast: normal result	October 2023-3 years 6 months	50-70 (female)	No comparator	—
Cancer–breast cancer detected	October 2023-3 years 6 months	After screened for breast cancer (female)	National Cancer Registry	Breast cancer diagnosis via screening route February 2020 to July 2023 (3.5 years)
Cancer–cervical: adequate smear	October 2023-3 years 6 months	25-49 (female)	NHS Futures Screening Dashboard	Cervical screening uptake December 2023 (25-49 years)
Cancer–cervical: adequate smear	October 2023-5 years 6 months	50-64 (female)	NHS Futures Screening Dashboard	Cervical screening uptake December 2023 (50-64 years)
Malignant neoplastic disease	New episode added in last 12 months (as of October 2023)	>18	National Cancer Registry	Rapid Cancer Registration – New Diagnosis (2023)
Malignancy stage	New episode added in last 12 months (as of October 2023)	>18	National Cancer Registry	Rapid Cancer Registration – New Diagnosis (2023)
Treatment regimen	New episode added in last 12 months (as of October 2023)	>18	Cancer Waiting Times	Treatment starts (2023)
Cancer care review	New episode added in last 12 months (as of October 2023)	Cancer-diagnosed patients who had a care review	No comparator	—
Has a carer	Recorded ever	>18	No comparator	—

XSL•FO
RenderX

[a]Not applicable.

Missing Data

There were challenges with data collection in participating practices. Running searches at scale across several practices was technically difficult at times. These searches would often time out and fail, resulting in incomplete data collection. Despite repeated running of these searches to attempt to get a complete dataset, failures persisted (Table S3 in Multimedia Appendix 1 and its linked notes provide more detail on this process).

Barnet had the highest level of missing data across the 26 reports, although data gaps were present in all boroughs. The ethnic diversity of Barnet is broadly comparable to the rest of NCL, and given that data from Barnet were available for other nonethnicity reports, we do not believe these gaps materially skew the overall findings. Table S7 in Multimedia Appendix 1 illustrates the proportions of practices that submitted data across each of the 26 searches.

The searches did not examine comparable data across each screening pathway. Direct comparisons between breast, colorectal, and cervical screening were not feasible, as primary care is not uniformly responsible for screening data across the 3 screening pathways. As a result, a process mapping exercise was conducted to trace the pathways of each screening program into primary care, identifying recordable actions within primary care settings as shown in Figure 1.

Figure 1. Diagram showing how codes are generated for all cancer screening programs in primary care.

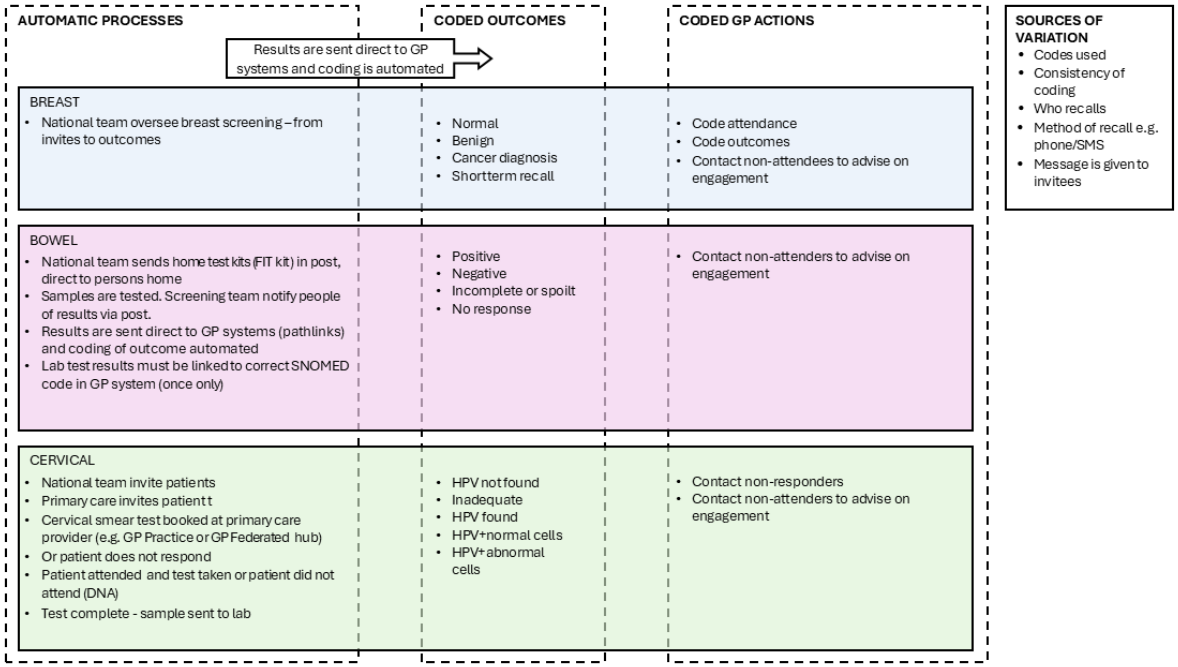

The qualitative data were analyzed by SB, a female behavioral science PhD student, and GB, a female applied health researcher. Both had previous experience conducting qualitative research. The analysis began in October 2024, after qualitative data collection had finished. SB listened to all video recordings and checked the transcripts for accuracy to become familiar with the study context and dataset before coding the data.

Qualitative data collection aimed to contextualize and validate quantitative findings rather than to achieve theoretical saturation. As this study adopted a QI focus, data collection was concluded once sufficient breadth of perspectives had been obtained and no new issues emerged that were relevant to the study objectives.

The analysis was conducted in Google Sheets using framework analysis [37]. Key excerpts from transcripts, email conversations, and comments posted in the chat during the workshops were copied into Google Sheets. All qualitative data were analyzed together. Raw data were arranged into columns. Each column represented a single code, and each row included raw data (eg, a verbatim quote labeled by data source and speaker) pertaining to that code.

An initial framework of 3 themes was developed deductively based on the findings of the quantitative analysis. SB coded the data into each theme before arranging the data into subthemes inductively. SB and GB met after 10%, 50%, and 100% of the data had been coded to discuss and revise the coding framework. A reflexive diary was kept and referred to throughout the analysis, which included field notes and impressions recorded by the researcher who conducted the interviews and workshops (AB) and the researchers who conducted the analysis (SB and GB). Regular meetings were held with the research team to clarify contextual details and discuss key interpretations. A reflexivity statement is available in Textbox S2 in Multimedia Appendix 1.

XSL•FO
RenderX

Ethical Considerations

This study was designed as a QI and health service enhancement initiative and therefore ethics approval was not applied for.

All data analyzed were fully deidentified at source, aggregated, and subject to small-number suppression (<5) in accordance with local information governance standards to protect patient confidentiality. The dataset was generated through local GP federations, with each federation running standardized searches and submitting aggregated data to Enfield Federation. Enfield Federation then securely transferred the data to the NCL Cancer Alliance for analysis.

The NCL primary care data protection officer reviewed and assured all data-sharing processes, confirming they aligned with the UK General Data Protection Regulation and the Caldicott Principles. A Data Protection Impact Assessment was not required, as the data used contained no identifiable information and presented no privacy risk.

Results

Principal Findings

We present our findings combining our descriptive analyses of clinical codes (quantitative findings) and the factors influencing coding (qualitative findings), in three themes: theme 1, precancer pathway, which includes codes relating to demographic characteristics, physical characteristics, risk factors (eg, family history), and cancer referrals; theme 2, cancer screening, which includes codes relating to screening invitations and uptake; and theme 3, postcancer diagnosis, which includes codes relating to staging, treatment, primary care surveillance, and follow-up.

In Table S8 in Multimedia Appendix 1, we include further details on data items included for each theme. Key findings are summarized for each theme before the results of the quantitative and qualitative analyses are presented.

Theme 1: Precancer Pathway

Overview

Completeness, code diversity, and validation of precancer codes are presented in Table 4 (Table S9 in Multimedia Appendix 1 provides additional detail on granularity).

XSL•FO
RenderX

Table 4. Descriptive analyses for data domains in theme 1 (precancer), showing completeness, code diversity, and validation of codes that were searched for in this theme.

Report name	Time period	Completeness and code diversity		Code validation	
		Completeness (% eligible coded), n/N (%)	Number of unique SNOMED CT[a] or EMIS[b] codes, n	Comparator source	EMIS completeness vs comparator, n/N (%)
Ethnic origin (coding)	Recorded ever	912,679/1,055,083 (86.5)	375	HealtheIntent	86.5/90.7 (95.4)
Main language spoken	Recorded ever	898,023/1,307,601 (68.7)	338	HealtheIntent	68.7/62.8 (109.4)
Employment status	24 months up to October 2023	29,848/1,229,644 (2.4)	87	No comparator for coding prevalence of this data item sourced	No comparator for coding prevalence of this data item sourced
Weight or BMI recorded	24 months up to October 2023	485,660/1,159,241 (41.9)	34	SELCA[c]	41.9/45.8 (91.5)
Current smoker (recorded in last 24 months)	24 months up to October 2023	90,029/1,236,580 (7.3)	47	HealtheIntent	7.3/14.9 (49)
Any smoking status (recorded in last 24 months)	24 months up to October 2023	530,720/1,185,812 (44.8)	117	SELCA	44.8/48.6 (92.3)
Alcohol consumption record	24 months up to October 2023	222,753/1,236,580 (18)	39	SELCA	18/29.2 (61.6)
Family history of neoplasm	Recorded ever	183,424/1,236,580 (14.8)	479	No comparator for coding prevalence of this data item sourced	No comparator for coding prevalence of this data item sourced
Environmental pollutants	24 months up to October 2023	494/1,113,575 (0.04)	45	No comparator for coding prevalence of this data item sourced	No comparator for coding prevalence of this data item sourced
Fast track referral coding	New episode added last 12 months up to October 2023	61,562/1,506,746 (4.1)	34	Cancer Waiting Times	61,562/78,989 (78)

[a]SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms.

[b]EMIS: Egton Medical Information Systems.

[c]SELCA: South East London Cancer Alliance.

The completeness of ethnicity coding was high, with 86.5% (912,679/1,055,083) of records containing an ethnicity code. A total of 375 distinct SNOMED CT codes were identified, with a 15.8% variation in coding completeness across boroughs. “Other White background - ethnic category 2001 census” made up 18.6% (170,190/912,679) of total codes captured in NCL. Language coding completeness was 68.7% (898,023/1,307,601) across the eligible population, which compares favorably with the comparator database coverage of 62.8% (929,987/1,482,024). A total of 338 unique codes were identified, with borough-level variation of 15.2% (Camden=75.3% and Enfield=60.1%). The top 2 most prevalent codes for language were “Main spoken language English” (527,227/898,023, 58.7%) and “Main spoken language NOS” (48,311/898,023, 5.4%). Employment status coding was minimal, with completeness at only 2.4% (29,848/1,229,644). The most frequently recorded codes related to unemployment (11,801/29,848, 39.5%) and work-related stress (5063/29,848, 17%), with a total of 87 distinct codes identified.

Coding completeness for BMI was 41.9% (485,660/1,159,241), with 34 individual codes used to describe weight and BMI. The SNOMED CT code for BMI accounted for 88.7% (430,736/485,660) of recorded entries. Overall, 44.8% (530,720/1,185,812) of the population had a recorded smoking status, including classifications such as current smoker and ex-smoker. Approximately 170 variations of smoking-related codes were identified. Data validation demonstrated near completion. Alcohol consumption was recorded in 18% (222,753/1,236,580) of patient records, with borough-level variation ranging from 13.9% (39,198/281,807) to 21% (64,255/281,807). A total of 39 different codes were identified. The “AUDIT-C” screening tool, which assesses excess alcohol consumption, was the prevalent code at 43.4% (96,619/222,753).

Family history of cancer across NCL showed that 14.8% (183,424/1,236,580) of records contained relevant codes, with 479 unique codes identified. There was no available comparator dataset for this parameter. The 2 most prevalent codes were “FH-Cancer” and “FH-Neoplasm” at around 17% each. Coding of environmental exposure was extremely limited, with a completeness rate of 0.04% (494/1,113,575), rendering the data unsuitable for analysis.

XSL•FO
RenderX

The incidence of suspected cancer referrals among eligible patients was 4.1% (61,562/1,506,746), with a 1% variation between the highest and lowest referring boroughs. The comparator database indicated a referral rate of 5.2% (78,989/1,506,746) for NCL residents. The analysis excluded deceased, temporary, and deregistered patients. The discrepancy between the study findings and comparator data was largely attributable to deceased patients.

Factors Influencing Clinical Coding of Precancer Data in Primary Care

Stakeholders reported 3 key factors that contribute to the quality and completeness of clinical coding of precancer data in primary care: opportunities to collect or update precancer data, motivations and capacity to collect and code precancer data, and the nature of the systems used to code precancer data.

Opportunities to Collect or Update Precancer Data

Many stakeholders reported that a key opportunity to capture precancer data is through registration forms and health check appointments. Some explained that registration templates and the commissioning of health check appointments vary by practice, leading to inconsistencies. There was agreement that standardized templates could improve this, although only for those registering subsequently, and it was noted that patients may not provide information if questions are not mandatory and the purpose of data collection is not transparent. While a few GPs suggested that mandating questions could improve data completeness, others were concerned that this could introduce barriers to registration. Family history is not routinely coded but may be documented in free text within clinical notes or referral forms. Environmental exposure data is not routinely collected or standardized within primary care records.

Some stakeholders advised that a lack of data completeness for information that changes over time, such as main language spoken or smoking status, is due to limited opportunities to update patient data after registration unless patients schedule appointments with primary care (eg, long-term conditions review, new medication appointments, e-consultations). It was suggested this could be improved by offering annual health check appointments, inviting patients to update their information, and the sharing of information collected in secondary care.

Motivations and Capacity to Collect and Code Precancer Data

Many GPs indicated that local and national financial incentive schemes (such as the QOF [38]) influence whether they collect and code precancer data at registration and during patient consultations. Some reported that variation in coding completeness between boroughs could be attributed to differences in locally commissioned services. There was agreement that clinical coding is demanding of staff time and capacity, and that improvements to the quality and completeness of clinical coding are unlikely to continue beyond the period of incentivization.

There was a general consensus that, in the absence of mandates and incentives, GPs are motivated to code precancer data that are relevant to the clinical workflow, such as arranging appointments (eg, need for interpreter, has a carer), assessing eligibility for local-level services (eg, smoking cessation, vaccination, information provision), or management of a patient's symptoms or a long-term condition. Some GPs admitted that they prefer to document precancer data in free text responses. This prevented interrupting the flow of conversation with patients, and, if needed, they could provide detail on more complex factors (eg, family history and exposure to environmental pollutants).

Nature of Systems Used to Code Precancer Data

GPs advised that coding will vary depending on whether patients register at GP practices using a paper or online form, whether registration data are coded into the system manually by staff or automatically, and which additional registration processing software practices have access to. It was also raised that there are multiple places within the system for data to be recorded. Furthermore, it was reported that some urgent suspected cancer referral forms may be available in the system but not trigger a SNOMED CT code. There was consensus that automated registration and data capture would improve the consistency and completeness of precancer data in primary care systems.

Several GPs suggested that the consistency of coding is made challenging by the array of codes available for specific types of precancer data such as smoking status, BMI, and family history, where there are different levels and layers, codes with similar or ambiguous meanings, and historic codes and prompts that cannot be removed (Figure 2). GPs admitted that codes higher up the list or those labelled with QOF prompts are most likely to be selected, and a program manager advised that coding prompts should be reserved for data that are most important to capture.

XSL•FO
RenderX

Figure 2. Screenshot of SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) code options for term “smok” when typed into EMIS (Egton Medical Information Systems) Web, showing a drop-down menu of multiple code options that start with “smok.”.

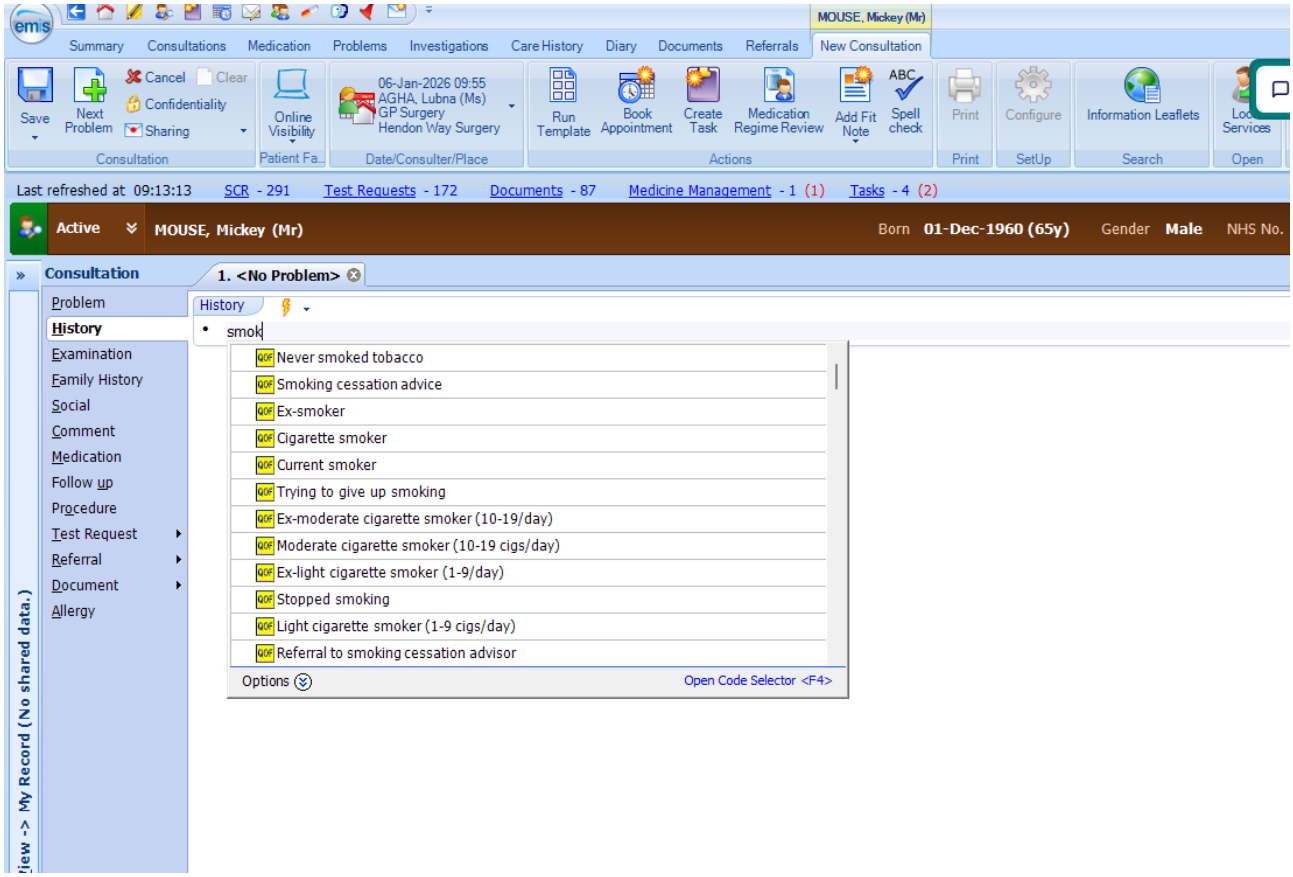

Theme 2: Screening Pathway (Breast, Bowel, and Cervical Screening)

Completeness, code diversity, and validation of cancer screening data is presented in Table 5 (Table S10 in Multimedia Appendix 1 provides additional detail on granularity).

XSL•FO
RenderX

Table 5. Descriptive analyses for data domains in Theme 2 (screening), showing completeness, code diversity, and validation of codes that were searched for in this theme.

Report name	Time period	Completeness and code diversity		Code validation	
		Completeness (% of eligible patients with code captured), n/N (%)	Number of unique SNOMED CT[a] or EMIS[b] codes in search, n	Data validation comparator source	EMIS data completeness as a proportion of comparator, n/N (%)
Cancer–bowel (aged 60-74 years; did not return screening kit)	Run date (October 2023)-2 years 6 months	65,762/188,939 (34.6)	4	No comparator for coding prevalence for kit DNRs[c], only uptake and coverage	No comparator for coding prevalence for kit DNRs, only uptake and coverage
Cancer–bowel (aged 60-74 years; screened)	Run date (October 2023)-2 years 6 months	112,939/184,323 (61.3)	10	NHS[e] Futures Screening Dashboard	61.3/62.2 (98.6)
Cancer–bowel abnormal result	Run date (October 2023)-2 years 6 months	2308/184,323 (1.3)	5	No comparator for coding prevalence for abnormal screening results, only uptake and coverage	No comparator for coding prevalence for abnormal screening results, only uptake and coverage
Cancer–breast (aged 50-70 years; screened)	Run date (October 2023)-3 years 6 months	79,321/175,986 (45.1)	30	NHS Futures Screening Dashboard	45.1/57.2 (78.8)
Cancer–breast (aged 50-70 years; abnormal result)	Run date (October 2023)-3 years 6 months	1646/114,606 (1.4)	7	No comparator for coding prevalence for abnormal screening results, only uptake and coverage	No comparator for coding prevalence for abnormal screening results, only uptake and coverage
Cancer–breast (aged 50-70 years; normal result)	Run date (October 2023)-3 years 6 months	71,043/165,126 (43)	3	No comparator for coding prevalence for normal screening results, only uptake and coverage	No comparator for coding prevalence for normal screening results, only uptake and coverage
Cancer–breast cancer detected	Run date (October 2023)-3 years 6 months	913/167,315 (0.5)	35	National Cancer Registry	913/731 (124.9)
Cancer–cervical (aged 25-49 years; adequate smear)	Run date (October 2023)-3 years 6 months	205,730/356,955 57.6%	103	NHS Futures Screening Dashboard	57.6/57.7 (99.8)
Cancer–cervical (aged 50-64 years; adequate smear)	Run date (October 2023)-5 years 6 months	97,233/139,233 (69.8)	98	NHS Futures Screening Dashboard	69.8/71 (98.3)

[a]SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms.

[b]EMIS: Egton Medical Information Systems.

[c]DNR: did not return.

[d]NHS: National Health Service.

Breast Screening

Coding for breast cancer screening uptake was recorded in 45.1% (79,321/175,986) of screening-eligible patients, with a 12% deficit compared with the comparator database. A total of 30 unique codes were used to document breast screening activity. The number of breast cancer diagnoses following abnormal results exceeded those in the comparative dataset. The most frequently used code was “Mammography normal,” accounting for 39.2% (31,109/79,321) of coded entries. The qualitative data (semistructured interviews and workshops) verified much of the persistent quality issues in the breast screening service, particularly its interface with primary care. This data also highlighted problems such as significant delays in breast screening attendance and nonattendance notifications; delivery of notifications by letters that contain multiple patients on a single sheet and therefore require manual separation; and counterintuitive patient reminders. For instance, proactive reminders for women upon turning age 50 years can cause confusion, because screening invitations may not be issued until age 53 years, with no flexibility for earlier appointments.

XSL•FO
RenderX

Bowel Cancer Screening

Codes for bowel cancer screening suggested an uptake of 61.3% (112,939/184,323). The proportion of screening-eligible patients who did not return test kits was 34.6% (65,762/188,939). Coding of abnormal results was documented in 1.3% (2308/184,323) of cases. The fecal immunochemical test (FIT), which underpins bowel screening, was consistently coded as "Bowel cancer screening program faecal occult blood test" within EMIS Web. The most frequently used codes were for normal and abnormal fecal occult blood (FOB) test results. The qualitative data (semistructured interviews) showed us that the bowel screening processes and coding practices were clearly defined for patients who undergo screening (aligned to QOF-funded activity). However, variability exists in recall, reminder, and engagement activities for nonresponders. Those who did not participate in screening were often coded using a non-SNOMED CT term, "No response to BCSP invitation."

Cervical Screening

Cervical screening data in the eligible population, coded as having had a smear, were 57.6% (205,730/356,955; in the 25 to 49 years age group) and 69.8% (97,233/139,233; in the 50 to 64 years age group). Both numbers aligned with the comparator database at over 99% (57.6%/57.7%) and 98% (69.8%/71%). A total of 98-103 individual codes were used to describe cervical screening coverage among the 2 eligible age cohorts. The qualitative data (semistructured interviews with GP federation's IT team) revealed that cervical screening processes were also clearly defined and aligned with the funding route (QOF). Variability existed in the recall and engagement activities for nonresponders. Reports for cervical nonengagement and recall were not developed because it was understood that data would not exist.

Factors Influencing Clinical Coding of Cancer Screening Data in Primary Care

Stakeholders explained that the coding of cancer screening data in primary care is influenced by primary care staff motivations and the ease of coding screening results in the system.

Relevance of Coding Cancer Screening Data to Clinical Workflow

Primary care staff described varied practices in whether cancer screening attendance was recorded. While some mentioned manual efforts or automated systems for reminding patients about upcoming screening appointments or contacting patients when they were notified of nonattendance, many admitted that they did not code screening attendance and follow up for those who did not attend. In discussing reasons for this, GPs and program managers indicated that coding was motivated by mandates and incentives (QOF) that are often only short term.

Many GPs agreed that recording screening data or contacting those who did not attend was not clinically relevant to primary care workflow and believed it was under the remit of national teams that run the programs. A few also noted that they did not have the most up-to-date information to monitor and facilitate screening attendance as they had received incorrect system prompts around screening attendance and were not aware of changes to screening eligibility. While some GPs and program managers noted local-level efforts to support and improve cancer screening attendance, there was general consensus that this is dependent on practice capacity to follow up those patients and is challenging due to competing priorities.

Ease of Coding Screening Results in Primary Care

GPs and program managers raised that coding breast cancer screening data is time consuming and demanding because paper results are sent to primary care with 2 patients' results per page, meaning they must be cut in 2 before being filed. Some recounted making requests for results to be sent electronically to streamline this process but had accepted that the system cannot be changed. GPs and program managers reported that the multitude of coding options for breast and cervical cancer screening results (eg, cervical screening, smear, cervical smear) contribute to coding inconsistencies. There was agreement that standardized coding could improve this. In contrast, stakeholders reflected that processing bowel cancer screening results is straightforward because the codes on the screening results letters are easy to match to those on the system (SNOMED CT). However, some highlighted that FOB codes (which relate to guaiac fecal occult blood testing [gFOBt] that is no longer used in the bowel screening program) [39] are still being used to code FIT screening results. Some reported that there is unified understanding in primary care that these legacy codes relate to FIT results and that incentive schemes for screening data still acknowledge these codes. However, one GP raised that legacy FOB codes may not be acknowledged in data searches for symptomatic FIT results.

Theme 3: Postcancer Diagnosis

Overview

Completeness, code diversity, and validation for postcancer codes are presented in Table 6 and in Table S12 in Multimedia Appendix 1 (Table S11 in Multimedia Appendix 1 provides additional detail on granularity).

XSL•FO
RenderX

Table 6. Descriptive analyses for data domains in Theme 3 (postcancer diagnosis), showing completeness, code diversity, and validation of codes that were searched for in this theme.

Report name	Time period	Completeness and code diversity		Code validation	
		Completeness (% of eligible patients with code captured), n/N (%)	Number of unique SNOMED CT[a] or EMIS[b] codes in search	Data validation comparator source	EMIS data completeness as a proportion of comparator, n/N (%)
Malignant neoplastic disease	New episode added last 12 months up to October 2023	6,044/1,506,746 (0.4)	677	National Cancer Registry	4604/6319 (73)
Malignancy stage	New episode added last 12 months up to October 2023	6/64,977 (0.01)	4	National Cancer Registry	6/3536 (0.2)
Treatment regime (if coded)	New episode added last 12 months up to October 2023	315/942,260 (0.03)	35	Cancer Waiting Times	315/10,574 (3)
Cancer care review	New episode added last 12 months up to October 2023	3953/1,236,580 (0.3)	1	No comparator for coding prevalence of this data item sourced	No comparator for coding prevalence of this data item sourced
Has a carer	Recorded ever	14,513/1,229,644 (1.2)	17	No comparator for coding prevalence of this data item sourced	No comparator for coding prevalence of this data item sourced

[a]SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms.

[b]EMIS: Egton Medical Information Systems.

The proportion of newly diagnosed cancers coded in EMIS was 73% (4604/6319) of the expected figure. The EMIS searches excluded deceased patients; further data comparison suggests that deceased individuals could account for 11.8%-14% of new cancer diagnoses in the Rapid Cancer Registration Database, reducing the initial coding gap from 27% to approximately 13%.

Cancer staging data showed a total completeness of 0.2% (6/3536) when compared with the validation database. One borough recorded no staging codes. Cancer treatment coding showed a total completeness of 3% (315/10,574) when compared with the validator.

CCRs were coded for 66% (3953/5982) of eligible patients, aligning with the 73% (4604/6319) of new cancer diagnoses recorded in primary care. A single SNOMED CT code is used to document CCRs. A total of 1.2% (14,513/1,229,644) of patients aged 18 years and older had a recorded carer status. The code “Has a carer” accounted for 80.1% (11,631/14,513) of these entries. Beyond CCRs, other personalized cancer care quality indicator data are shown separately in Table S12 in Multimedia Appendix 1. Cancer care plans, end-of-treatment summaries, and holistic needs assessments were recorded in 1%-2% of patients with cancer across NCL.

Factors Influencing Clinical Coding of Postcancer Diagnosis Data in Primary Care

Stakeholders suggested that coding of patients’ cancer diagnoses and treatment data is influenced by the relevance of information to the clinical workflow in primary care, the quality of information sharing from secondary care, and the complexity and consistency of the systems used for coding.

Relevance of Coding Postcancer Diagnosis Data to Clinical Workflow

A few GPs mentioned that coding of cancer diagnoses will improve when incentivized through the QOF. However, some highlighted that coding information about treatment plans organized by secondary care is not perceived as relevant or a priority in daily practice. One GP also reflected that they sometimes felt reluctant to request information from patients during CCR appointments given that patients have already had to discuss their diagnoses in secondary care.

Quality of Postcancer Diagnosis Information Sharing

GPs suggested that cancer diagnosis information is not always shared by secondary care or that it may be sent with some delay. They also reported that information may be missing for patients who are diagnosed and treated privately or those who are diagnosed at an advanced stage whereby primary care is only notified of a cancer diagnosis through the receipt of postmortem information. When cancer diagnosis information is received, there is consensus among GPs that letters from secondary care are long and complex, meaning staff have to scrutinize the whole letter to find and extract the key information. Many agreed that diagnosis information and SNOMED CT codes that require coding should be placed at the top of these letters for easy translation into primary care records. Some GPs also lack trust

XSL•FO
RenderX

in the accuracy of information received from secondary care due to finding previous errors in patient records. For this reason, a few expressed concerns about linked data between primary and secondary care if errors could not be redacted within primary care.

Consistency and Complexity of Systems Used for Coding Postcancer Diagnosis Data

GPs reported that there are multiple places within primary care–based systems (EMIS) where cancer information can be coded or entered as free text, as well as several different coding options that prevent consistent coding of cancer diagnoses. There was agreement that standardized templates were needed. GPs also reported that *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) codes used in secondary care letters are not aligned to SNOMED CT codes used in primary care, which causes ambiguity. Additionally, some explained that practices vary in their capacity to ensure the quality of coding, such as whether formalized coding teams are used and given adequate resources, time, and training to correctly code information received from secondary care.

Discussion

Summary

This QI project provides a detailed assessment of cancer-related clinical coding in NCL primary care across a population of 1.4 million adults. Our findings show that although some demographic data such as ethnicity and language are well captured (coding completeness), many other important codes across the cancer pathway—especially those relating to social determinants of health, cancer treatments, and postdiagnosis care—remain inconsistently coded or significantly absent from primary care records.

Coding quality was strongly influenced by the presence of national or local incentives (such as QOF), which drove completeness for certain indicators like ethnicity, cancer diagnosis, and CCRs. In contrast, areas not linked to performance payments or formalized data capture processes, such as cancer treatment, staging, and screening follow-up, showed substantial gaps. These differences signal that current coding behavior in primary care is shaped by system design, contractual levers, and administrative capacity.

Further qualitative insights helped contextualize these patterns, showing that GPs often prioritize coding activities relevant to their daily clinical workflow or incentivized tasks. Key barriers to good coding included complex and inconsistent coding systems, limited opportunities to update data, and limited structured information sharing from secondary care.

Overall, the findings demonstrate that improving cancer coding quality in primary care requires more than local process changes; it will require a coordinated national approach that includes clearer coding standards, automation, and alignment of incentives.

Interpretation

Our findings align with existing literature relating to cancer data coding in primary care. For example, our observation that ethnicity coding was high at 86.5% (912,679/1,055,083), is consistent with other studies using patient electronic records. This is similar to 78.2% ethnicity coding reported in NHS primary care records in 2022 [40]. These high levels may be related to incentive schemes (QOF) to improve completeness in England [41]. Similarly, 68.7% (898,023/1,307,601) completeness for language coding aligns with findings that emphasize the need for comprehensive demographic information to support better patient health outcomes [42].

The minimal recording of employment status (29,848/1,229,644, 2.4%) and the moderate completeness of BMI data (485,660/1,159,241, 41.9%) underscore challenges in capturing socioeconomic and health metrics. These gaps are consistent with literature indicating that certain health data, such as employment status, are often underreported in primary care settings [43]. Our findings on smoking status (530,720/1,185,812, 44.8%) and alcohol consumption (222,753/1,236,580, 18%) coding completeness, which show significant borough variation, also reflect widespread underrecording of these factors in patient records [44]. This may reflect discomfort in approaching conversations about lifestyle factors as well as lack of time or resources, particularly in more deprived areas [12,45].

The variation in cancer screening data across pathways, particularly the fragmented coding for breast screening, highlights the complexities in capturing screening information. Rafi et al [46] also found that a wide range of codes were used to document a family history of breast cancer in primary care, leading to inconsistencies in data quality and the potential for misclassification of risk. This finding may relate to primary care's role and responsibilities in screening, where their involvement varies significantly across different cancer types [47]. While primary care plays a structured role in bowel and cervical screening (underpinned by funding and agreed processes nationally), breast cancer screening activity is not delivered or incentivized in primary care.

New cancer diagnoses were also undercoded, at 87% (4604/5260) completeness. This aligns with findings from the Netherlands, where only 60.6% of cancer cases were coded according to the national registry [9]. Undercoding is linked to reliance on unstructured secondary care letters and a lack of coding incentives [8,9]. Cancer treatment and staging data were also undercoded, and there was not any existing evidence to compare against.

The recording of CCRs (3953/5982, 66% of eligible patients) closely mirrors the proportion of coded diagnoses, suggesting a correlation between diagnosis coding and care review documentation. However, concerns exist that high completeness rates reflect a "tick-box" approach driven by financial incentives for coding "cancer care review" [48]. Documentation of broader personalized cancer care indicators, such as care plans and holistic needs assessments, was minimal (1%-2%), highlighting gaps in comprehensive cancer care documentation. This aligns with studies calling for broader, patient-centered metrics to sustain confidence in cancer registries [49]. National CCR templates embedded in GP electronic records [50] support additional quality indicator completion alongside CCRs,

XSL•FO
RenderX

promoting standardization [51]. The lack of broader care indicators being coded suggests that these templates are not embedded in daily practice.

While this project was conducted in NCL, the findings have broader relevance across other similar primary care settings. Although population demographics vary between regions, the core structures, processes, and incentives that shape GP coding behaviors are largely consistent across the NHS and similar international health care systems. Primary care operates within a standardized framework of national policies, contractual obligations, and clinical systems, meaning that the challenges and opportunities identified in this study are likely to be mirrored elsewhere. As such, our findings provide valuable insights for informing coding improvement initiatives beyond our study region, with potential applicability across primary care systems nationally.

Limitations

Two boroughs were excluded due to nonengagement; however, the dataset remained robust with high completeness in other areas, minimizing bias. The findings should be interpreted cautiously for nonsubmitting practices. Additionally, our quantitative data searches were completed by GP Federation's IT team, but their searches could not fully align with our specifications. This left gaps in key cancer care metrics, such as patients with cancer on stratified follow-up pathways or active surveillance; alternative data sources such HealtheIntent (the ICBs' provider for linked datasets across the NHS in NCL until September 2025), helped fill in some gaps.

Conclusions

The QI project has provided a unique and detailed insight into the many dimensions of cancer coding across the whole pathway in primary care and sheds light on many factors that underpin variation and coding preference.

We have developed recommendations based on our findings aimed at primary care providers, commissioners, ICBs' digital teams [52], cancer screening teams and the National Cancer team [53].

Implications for Practice

Textbox 1 outlines practical recommendations to enhance primary care data management, integration with secondary care, and overall service quality. The focus is on standardizing coding practices, improving information flow, and leveraging data for informed decision-making and patient care improvements.

XSL•FO
RenderX

Textbox 1. Implications for practice.

Strengthen data infrastructure in primary care

- Develop a structured data framework using SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) with a minimum dataset and distinct code sets, aligning with secondary care processes. This will enable better use of primary care data for analysis, epidemiology, and health care improvement.

Enhance breast screening integration with primary care

- Transition to electronic breast screening outcome reports for general practitioners (GPs), standardize national breast screening activity codes within SNOMED CT, and define and resource primary care's role in patient follow-up and nonattendance intervention. These changes will improve the accuracy and efficiency of breast screening data and ensure clear responsibility for patient engagement.

Improve secondary care information for primary care

- Standardize hospital discharge letters to clearly indicate new diagnoses, multidisciplinary team outcomes, and care plans, ensure clinical coding aligns with SNOMED CT; prioritize consistent implementation of end-of-treatment summaries; and link secondary care data directly to primary care [27-29]. This recommendation aligns with the national strategy [54,55]. This will reduce duplication, enhance continuity of care, and improve data quality for patient management.

Standardize primary care coding practices

- Employ coding professionals in primary care, align screening activity codes, update bowel screening codes to reflect the transition from fecal occult blood to fecal immunochemical tests, and educate the workforce on the value of high-quality coding. These steps will improve consistency, accuracy, and the usefulness of coded data for patient care and service planning.

Optimize patient registration processes

- Automate the integration of coded demographic, behavioral, and risk factor data from National Health Service Digital's updated registration forms (Patient Registration Form 1) [56] into primary care systems and allocate resources for implementation. This will ensure more complete and accurate patient information from the outset.

Improve Quality Outcomes Framework rules and transparency

- Make Quality Outcomes Framework rule changes trackable over time and provide clearer navigation and updates [57,58]. This will help those involved in service improvement and research to understand and respond to coding changes more effectively.

Implement primary care coding audits

- Introduce National Health Service England–funded coding audits, assess data quality and completeness, flag nonrecommended codes, and cross-reference with national datasets. This will improve coding accuracy, highlight inconsistencies, and enhance data reliability.

Develop an analytics dashboard

- Create a live dashboard to track trends, profile data, and support quality improvement, leveraging the London Health Data Service (launched in June 2025) [59]. This will provide real-time insights into primary care data, supporting better decision-making and service planning.

Facilitate knowledge sharing

- Identify regions with superior data completeness and share successful quality improvement initiatives across London. This will promote best practices and drive improvements in data quality and patient care across the system.

Funding

This study did not receive external funding from any public, commercial, or not-for-profit sources. The Enfield Federation was commissioned to build the data searches, run these across NCL GP practices, and extract the data. SB's time was funded by the NCL Cancer Alliance to support the qualitative analysis. GB acknowledges funding from Barts Charity. No funding body was involved in study design, data collection, analysis, interpretation, or manuscript preparation.

Acknowledgments

The authors thank the Enfield general practitioner (GP) Federation data team, General Practice Provider Alliance, North Central London (NCL) GP practices, all workshop participants, patient partners, the internal NCL Cancer Alliance team working group, the NCL Integrated Care Board digital team, and all stakeholders who reviewed and engaged in this study.

We used ChatGPT (free version; OpenAI) [60] to support the manuscript writing process. It was used to optimize sentence structure, readability, and reduce text burden. All content, analyses, and interpretations were developed, verified, and approved by the authors. ChatGPT was not used for data generation, analysis, or creation of original scientific content. The artificial

XSL•FO
RenderX

intelligence (AI) use for this manuscript falls under Domain 1 of the “Recommendations for a Classification of AI Use in Academic Manuscript Preparation” document.

The project was delivered as part of service improvement and evaluation led by the NCL Cancer Alliance.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AB conceived and led the project, collated a significant portion of the mixed methods data, supported analysis, and drafted the initial manuscript. GR conducted data collation and analysis, identified comparative data sources, and contributed substantially to manuscript development. KT provided project support and made significant contributions to the manuscript. SB analyzed the qualitative data and drafted the corresponding sections. GB provided overall guidance and methodological support, particularly in qualitative data collection and analysis. She advised on this study’s structure and contributed significantly to manuscript development. All authors contributed to the analysis, interpretation of findings, and manuscript revisions. All authors reviewed and approved the final version for submission. AB is the guarantor of this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary (clean) 30.10.25.
[DOCX File , 11440 KB - jmir_v28i1e73205_app1.docx]

References

1. High quality patient records. NHS England. URL: https://www.england.nhs.uk/long-read/high-quality-patient-records/ [accessed 2025-01-06]
2. Taylor K, Henderson S. 112 Challenges of clinical coding: an audit of the accuracy of primary diagnosis coding in a specialist palliative care unit. BMJ Supportive & Palliative Care 2019;9(Suppl 1):A49-A50. [doi: 10.1136/bmjspcare-2019-asp.135]
3. Harry ML, Truitt AR, Saman DM, Henzler-Buckingham HA, Allen CI, Walton KM, et al. Barriers and facilitators to implementing cancer prevention clinical decision support in primary care: a qualitative study. BMC Health Serv Res 2019;19(1):534 [FREE Full text] [doi: 10.1186/s12913-019-4326-4] [Medline: 31366355]
4. Greenwood-Lee J, Jewett L, Woodhouse L, Marshall DA. A categorisation of problems and solutions to improve patient referrals from primary to specialty care. BMC Health Serv Res 2018;18(1):986 [FREE Full text] [doi: 10.1186/s12913-018-3745-y] [Medline: 30572898]
5. Morrison Z, Fernando B, Kalra D, Cresswell K, Sheikh A. National evaluation of the benefits and risks of greater structuring and coding of the electronic health record: exploratory qualitative investigation. J Am Med Inform Assoc 2014;21(3):492-500 [FREE Full text] [doi: 10.1136/amiajnl-2013-001666] [Medline: 24186957]
6. Pascoe SW, Neal RD, Heywood PL, Allgar VL, Miles JN, Stefoski-Mikeljevic J. Identifying patients with a cancer diagnosis using general practice medical records and cancer registry data. Fam Pract 2008;25(4):215-220. [doi: 10.1093/fampra/cmn023]
7. Strategies for success: tackling common clinical documentation integrity challenges head-on. Healthcare Financial Management Association. 2024. URL: https://www.hfma.org/revenue-cycle/strategies-for-success-tackling-common-clinical-documentation-integrity-challenges-head-on/ [accessed 2025-09-01]
8. Sollie A, Sijmons RH, Helsper C, Numans ME. Reusability of coded data in the primary care electronic medical record: a dynamic cohort study concerning cancer diagnoses. Int J Med Inform 2017;99:45-52. [doi: 10.1016/j.ijmedinf.2016.08.004] [Medline: 28118921]
9. Sollie A, Roskam J, Sijmons RH, Numans ME, Helsper CW. Do GPs know their patients with cancer? Assessing the quality of cancer registration in Dutch primary care: a cross-sectional validation study. BMJ Open 2016;6(9):e012669 [FREE Full text] [doi: 10.1136/bmjopen-2016-012669] [Medline: 27633642]
10. Zghebi SS, Reeves D, Grigoroglou C, McMillan B, Ashcroft DM, Parisi R, et al. Clinical code usage in UK general practice: a cohort study exploring 18 conditions over 14 years. BMJ Open 2022;12(7):e051456 [FREE Full text] [doi: 10.1136/bmjopen-2021-051456] [Medline: 35879012]
11. Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. BMJ 2003;326(7398):1070 [FREE Full text] [doi: 10.1136/bmj.326.7398.1070] [Medline: 12750210]
12. Davies A, Ahmed H, Thomas-Wood T, Wood F. Primary healthcare professionals’ approach to clinical coding: a qualitative interview study in Wales. Br J Gen Pract 2024;75(750):e43-e49. [doi: 10.3399/bjgp.2024.0036]

XSL•FO
RenderX

13. Collaço N, Lippiett KA, Wright D, Brodie H, Winter J, Richardson A, et al. Barriers and facilitators to integrated cancer care between primary and secondary care: a scoping review. Support Care Cancer 2024;32(2):120 [FREE Full text] [doi: 10.1007/s00520-023-08278-1] [Medline: 38252169]
14. National Cancer Registration and Analysis Service (NCRAS). GOV.UK. URL: https://www.gov.uk/guidance/national-cancer-registration-and-analysis-service-ncras [accessed 2025-09-03]
15. Datasets in support of statutory duties. URL: https://digital.nhs.uk/services/data-services-for-commissioners/datasets-in-support-of-statutory-duties [accessed 2025-09-08]
16. Quality and Outcomes Framework (QOF). URL: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/general-practice-data-hub/quality-outcomes-framework-qof [accessed 2024-09-30]
17. Navigating the New QOF Landscape: Your Guide to the 2025/26 Changes for GP Practices. URL: https://www.mypracticemanager.co.uk/advice/qof-2025-26-changes [accessed 2025-09-11]
18. Tai TW, Anandarajah S, Dhoul N, de Lusignan S. Variation in clinical coding lists in UK general practice: a barrier to consistent data entry? Inform Prim Care 2007;15(3):143-150 [FREE Full text]
19. Cancer: find out more about the National Cancer Registration and Analysis Service (NCRAS) - one of the 2 disease registration services within the National Disease Registration Service (NDRS). NDRS. URL: https://digital.nhs.uk/ndrs/about/ncras [accessed 2025-01-11]
20. NHS Digital. The NHS Digital SNOMED CT Browser. URL: https://termbrowser.nhs.uk/ [accessed 2025-01-06]
21. NHS England. The Organisation has a Framework in Place to Support Lawfulness, Fairness and Transparency (1.1). URL: https://tinyurl.com/yc4pnfdt [accessed 2024-02-22]
22. NHS England. Getting It Right First Time - GIRFT. 2020. URL: https://gettingitrightfirsttime.co.uk/ [accessed 2025-01-06]
23. Doktorchik C, Lu M, Quan H, Ringham C, Eastwood C. A qualitative evaluation of clinically coded data quality from health information manager perspectives. Health Inf Manag 2020;49(1):19-27. [doi: 10.1177/1833358319855031] [Medline: 31284769]
24. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
25. How to make a case for change. StaffNet. URL: https://www.staffnet.manchester.ac.uk/od/learning-pathways/professional-and-technical-development/change-management/change-toolkit/how-to/how-to-make-a-case-for-change/ [accessed 2025-11-29]
26. Optum. EMIS Web: Comprehensive, flexible and powerful. URL: https://www.emishealth.com/products/emis-web [accessed 2025-01-23]
27. Arora A, Alderman JE, Palmer J, Ganapathi S, Laws E, McCradden MD, et al. The value of standards for health datasets in artificial intelligence-based applications. Nat Med 2023;29(11):2929-2938 [FREE Full text] [doi: 10.1038/s41591-023-02608-w] [Medline: 37884627]
28. Optum. SNOMED CT and EMIS codes. URL: https://www.emisnow.com/csm?id=kb_article_view&table=kb_knowledge&sys_kb_id=816eb8833b0b5a907290373a85e45a41&searchTerm=emis%20web%20codes [accessed 2025-02-10]
29. NHS England. Read Codes: Retirement of Read Version 2 and Clinical Terms Version 3. URL: https://tinyurl.com/yc7ckkzx [accessed 2016-04-01]
30. What is a GP Federation? GP Health Connect. URL: https://www.gphealthconnect.co.uk/what-is-a-gp-federation/ [accessed 2025-01-06]
31. HealtheIntent. URL: https://nclhealthandcare.org.uk/digital/digital-information-for-health-and-care-professionals/healtheintent/ [accessed 2025-09-03]
32. Pineda-Moncusí M, Allery F, Delmestri A, Bolton T, Nolan J, Thygesen JH, CVD-COVID-UK/COVID-IMPACT Consortium. Ethnicity data resource in population-wide health records: completeness, coverage and granularity of diversity. Sci Data 2024;11(1):221 [FREE Full text] [doi: 10.1038/s41597-024-02958-1] [Medline: 38388690]
33. NHS England. DAPB1521: Cancer Outcomes and Services Data Set. URL: https://tinyurl.com/ypmss6wt [accessed 2023-09-11]
34. Systemic Anti-Cancer Therapy (SACT) Data Set. URL: https://tinyurl.com/j6km6srp [accessed 2025-11-03]
35. Prevention, awareness and screening. North Central London Cancer Alliance. 2025. URL: https://www.nclcanceralliance.nhs.uk/our-work/prevention-awareness-and-screening/ [accessed 2025-09-11]
36. Bhuiya A, Cavanagh S, Nestor C, Fomina M, Ahmed I, Von Wagner C, et al. Development of a cancer pathway support guide for patients and carers: a codesign project. European Journal of Cancer Care 2024;2024:3623136. [doi: 10.1155/2024/3623136]
37. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013;13(1):117. [doi: 10.1186/1471-2288-13-117]
38. NHS England. Quality and Outcomes Framework (QOF) business rules. URL: https://www.google.com/url?q=https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/quality-and-outcomes-framework-qof/business-rules&sa=D&source=docs&ust=1739198517910415&usg=AOvVaw0gPCPGkD2artCnS-Cu87aB [accessed 2025-04-08]

XSL•FO
RenderX

39. National Institute for Health and Care Excellence. The NHS Bowel Screening Programme. URL: https://cks.nice.org.uk/topics/bowel-screening/background-information/the-nhs-bowel-screening-programme/ [accessed 2025-02-10]
40. Andrews CD, Mathur R, Massey J, Park R, Curtis HJ, Hopcroft L, et al. Consistency, completeness and external validity of ethnicity recording in NHS primary care records: a cohort study in 25 million patients' records at source using OpenSAFELY. BMC Med 2024;22(1):288 [FREE Full text] [doi: 10.1186/s12916-024-03499-5] [Medline: 38987774]
41. NHS England. Information Pack for GPs. URL: https://www.england.nhs.uk/london/wp-content/uploads/sites/8/2023/02/GP-Information-Pack-collection-of-patient-ethnicity-to-inform-resource-allocation-November-2022-Public.pdf [accessed 2025-11-29]
42. Cano-Ibáñez N, Zolfaghari Y, Amezcua-Prieto C, Khan KS. Physician-patient language discordance and poor health outcomes: a systematic scoping review. Front Public Health 2021;9:629041 [FREE Full text] [doi: 10.3389/fpubh.2021.629041] [Medline: 33816420]
43. Campbell P, Lewis M, Chen Y, Lacey RJ, Rowlands G, Protheroe J. Can patients with low health literacy be identified from routine primary care health records? A cross-sectional and prospective analysis. BMC Fam Pract 2019;20(1):101 [FREE Full text] [doi: 10.1186/s12875-019-0994-8] [Medline: 31319792]
44. Al Kazzi E, Lau B, Li T, Schneider E, Makary M, Hutfless S. Critical problems of coding data in health care: obesity, smoking, and alcohol use by method of measurement. Value in Health 2015;18(3):A2. [doi: 10.1016/j.jval.2015.03.007]
45. Thelen R, Bhatti S, Rayner J, Grudniewicz A. Collecting sociodemographic data in primary care: qualitative interviews in community health centres. BJGP Open 2025;9(1):BJGPO.2024.0095 [FREE Full text] [doi: 10.3399/BJGPO.2024.0095] [Medline: 39528270]
46. Rafi I, Chowdhury S, Chan T, Jubber I, Tahir M, de Lusignan S. Improving the management of people with a family history of breast cancer in primary care: before and after study of audit-based education. BMC Fam Pract 2013;14(1):105 [FREE Full text] [doi: 10.1186/1471-2296-14-105] [Medline: 23879178]
47. Green T, Atkin K, Macleod U. Cancer detection in primary care: insights from general practitioners. Br J Cancer 2015;112(Suppl 1):S41-S49 [FREE Full text] [doi: 10.1038/bjc.2015.41] [Medline: 25734388]
48. Gopal DP, de Rooij BH, Ezendam NP, Taylor SJ. Delivering long-term cancer care in primary care. Br J Gen Pract 2020;70(694):226-227. [doi: 10.3399/bjgp20x709481]
49. Shokrizadeharani L, Batooli Z, Heydarian S. Evaluation of completeness, comparability, validity, and timeliness in cancer registries: a scoping review. Studies in health technology and informatics 2023:305. [doi: 10.3233/shti230451]
50. Macmillan Cancer Support. Cancer Care Review. URL: https://tinyurl.com/y8b5p8xw [accessed 2025-09-16]
51. Macmillan Cancer Support. Practical Implementation Guide for Cancer Care Reviews. URL: https://www.macmillan.org.uk/healthcare-professionals/cancer-pathways/prevention-and-diagnosis/cancer-care-review [accessed 2025-02-11]
52. NHS England. What are Integrated Care Systems?. URL: https://www.england.nhs.uk/integratedcare/what-is-integrated-care/ [accessed 2025-02-11]
53. NHS England. Cancer. URL: https://www.england.nhs.uk/cancer/ [accessed 2025-02-11]
54. NHS England. Clinical Coding - SNOMED CT. URL: https://www.england.nhs.uk/long-read/clinical-coding-snomed-ct/ [accessed 2025-01-16]
55. Macmillan Cancer Support. Personalised Care for People Living With Cancer. URL: https://www.macmillan.org.uk/healthcare-professionals/innovation-in-cancer-care/personalised-care [accessed 2025-01-16]
56. Londonwide LMCs. Contract Changes for 2024/25. URL: https://www.lmc.org.uk/news/contract-changes-2024-25-and-online-registration-requirement/ [accessed 2024-10-23]
57. NHS England. Questions Patients are Asked in the Online form. URL: https://digital.nhs.uk/services/register-with-a-gp-surgery-service/get-help-using-the-service/questions-asked [accessed 2025-04-14]
58. NHS England. Quality and Outcomes Framework (QOF) business rules v48.0 2023-2024. URL: https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/quality-and-outcomes-framework-qof/quality-and-outcome-framework-qof-business-rules/quality-and-outcomes-framework-qof-business-rules-v48.0-2023-2024 [accessed 2024-01-16]
59. One London. London Secure Data Environment. 2025. URL: https://www.onelondon.online/london-secure-data-environment/ [accessed 2025-09-16]
60. Chat GPT. URL: https://chatgpt.com/ [accessed 2025-11-29]

Abbreviations

CCR: cancer care review
COREQ: Consolidated Criteria for Reporting Qualitative Research
EMIS: Egton Medical Information Systems
FIT: fecal immunochemical test
FOB: fecal occult blood
gFOBt: guaiac fecal occult blood testing
GP: general practitioner

XSL•FO
RenderX

ICB: integrated care board
ICD-10: International Statistical Classification of Diseases, Tenth Revision
NCL: North Central London
NCRAS: National Cancer Registration and Analysis Service
NHS: National Health Service
PCN: primary care network
QI: quality improvement
QOF: Quality Outcomes Framework
SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms

Edited by T de Azevedo Cardoso, A Stone; submitted 20.Mar.2025; peer-reviewed by W Gray, N Calanzani, E Whitfield; comments to author 25.Jul.2025; accepted 10.Nov.2025; published 07.Jan.2026.

Please cite as:
Bhuiya A, Roberts G, Tucker K, Bonfield S, Black G
Quality of Cancer-Related Clinical Coding in Primary Care in North Central London: Mixed Methods Quality Improvement Project
J Med Internet Res 2026;28:e73205
URL: https://www.jmir.org/2026/1/e73205
doi:10.2196/73205
PMID:

©Afsana Bhuiya, Graham Roberts, Katie Tucker, Stefanie Bonfield, Georgia Black. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Multimodal Large Language Models for Cystoscopic Image Interpretation and Bladder Lesion Classification: Comparative Study

Yung-Chi Shih[1], MD; Cheng-Yang Wu[2], MD; Shi-Wei Huang[2,3*], MD, PhD; Chung-You Tsai[1,3,4*], MD, PhD

[1]Division of Urology, Department of Surgery, Far Eastern Memorial Hospital, New Taipei City, Taiwan

[2]Department of Urology, National Taiwan University Hospital, Yunlin Branch, Yunlin, Taiwan

[3]Department of Urology, College of Medicine, National Taiwan University, Taipei, Taiwan

[4]Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan

[*]these authors contributed equally

Corresponding Author:
Chung-You Tsai, MD, PhD
Division of Urology
Department of Surgery
Far Eastern Memorial Hospital
No 21, Sec 2, Nanya S Rd., Banciao Dist.
New Taipei City, 220
Taiwan
Phone: 886 2 8966 7000
Fax: 886 2 8966 5567
Email: pgtsai@gmail.com

Abstract

Background: Cystoscopy remains the gold standard for diagnosing bladder lesions; however, its diagnostic accuracy is operator dependent and prone to missing subtle abnormalities such as carcinoma in situ or misinterpreting mimic lesions (tumor, inflammation, or normal variants). Artificial intelligence–based image-analysis systems are emerging, yet conventional models remain limited to single tasks and cannot produce explanatory reports or articulate diagnostic reasoning. Multimodal large language models (MM-LLMs) integrate visual recognition, contextual reasoning, and language generation, offering interpretive capabilities beyond conventional artificial intelligence.

Objective: This study aims to rigorously evaluate state-of-the-art MM-LLMs for cystoscopic image interpretation and lesion classification using clinician-defined stress-test datasets enriched with rare, diverse, and challenging lesions, focusing on diagnostic accuracy, reasoning quality, and clinical relevance.

Methods: Four MM-LLMs (OpenAI-o3 and ChatGPT-4o [OpenAI]; Gemini 2.5 Pro and MedGemma-27B [Google]) were evaluated under blinded, randomized procedures across two tasks: (1) free-text image interpretation for anatomic site, findings, lesion reasoning, and final diagnosis (n=401) and (2) seven-class tumor-like lesion classification (n=113) within a multiple-choice framework (cystitis, polyps, papilloma, papillary urothelial carcinoma, carcinoma in situ, non-urothelial carcinoma, and none of the above). Three raters independently scored outputs using a 5-point Likert scale, and classification metrics (accuracy, sensitivity, specificity, Youden J index (Youden J), and Matthews correlation coefficient [MCC]) were calculated for lesion detection, biopsy indication, and malignancy endpoints. For optimization, model performance was compared between zero-shot and text-based in-context learning prompts that were prefixed with brief descriptions of tumor features.

Results: The 401-image test set spanned 40 subcategories, with 322 (80.3%) containing abnormal findings in the image interpretation task. OpenAI-o3 demonstrated strong reasoning, with high satisfaction for anatomy (339/401, 84.5%) and findings (305/401, 76%), but lower satisfaction for lesion reasoning (211/401, 52.5%) and final diagnosis (193/401, 48.2%), indicating increasing difficulty with higher-order synthesis. Mean Likert score differences (OpenAI-o3 minus Gemini 2.5 Pro) were +0.27 for findings (adjusted *P* value: q=0.002), +0.24 for lesion reasoning (q=0.047), and +0.19 for final diagnosis. For clinically relevant endpoints in the full set, OpenAI-o3 achieved the most balanced performance, with lesion detection accuracy of 88.3%, sensitivity of 92%, specificity of 73.1%, Youden J of 0.650, and MCC of 0.635. In 7-class tumor-like lesion classification, OpenAI-o3 achieved accuracies of 73.5% for biopsy indication and 62.8% for malignancy, with a balanced sensitivity-specificity

XSL•FO
RenderX

trade-off, outperforming other models. Notably, OpenAI-o3 performed best on prevalent malignant lesions. ChatGPT-4o and Gemini 2.5 Pro showed high sensitivity but low specificity, whereas MedGemma-27B underperformed. In-context learning improved OpenAI-o3 microaverage accuracy (40.7%→46.0%; MCC 0.311→0.370) but yielded only slight specificity gains and minimal accuracy change in other models, likely constrained by the absence of paired image-text context.

Conclusions: MM-LLMs demonstrate meaningful assistive potential in generating interpretable cystoscopy free-text rationales and supporting biopsy triage and training. However, performance in difficult differential diagnoses remains modest and requires further optimization before safe clinical integration.

(J Med Internet Res 2026;28:e87193) doi:10.2196/87193

KEYWORDS

multimodal; large language model; AI; cystoscopy; diagnostic reasoning; finding description; biopsy indication; bladder tumor; artificial intelligence

Introduction

Cystoscopy is one of the most frequently performed procedures in urology [1]. Its effectiveness heavily depends on the urologist's experience, attention to detail, and interpretive skill, making it both technically and diagnostically challenging [2]. Interobserver variability is common, and lesion characterization (tumor vs inflammation vs normal variant) is not always straightforward, often requiring clinical correlation. Bladder cancer, the ninth most common cancer globally [3], relies heavily on cystoscopy as the cornerstone for diagnosis, treatment, and surveillance. However, studies report false-negative rates ranging from 10%-40%, with white-light cystoscopy missing up to one-third of carcinoma in situ (CIS) cases and frequently overlooking small tumors [4]. Accordingly, cystoscopic interpretation is a nuanced clinical process.

Artificial intelligence (AI)-assisted cystoscopic diagnosis and decision-making can be decomposed into distinct tasks: lesion detection (present vs absent), lesion classification, margin segmentation, descriptive reporting, biopsy triage, final diagnosis, and ultimately full report generation. Each task places different demands on algorithms, ranging from visual localization to semantic reasoning and clinical judgment. Previous work in cystoscopy has predominantly framed the problem as image classification or segmentation [5-9], often using specialized vision pipelines that localize or outline lesions but provide limited clinical context and have uncertain generalizability across morphology-diverse appearances.

Evidence from other endoscopic domains provides a useful benchmark. Task-tuned computer-aided detection systems in colonoscopy, for example, improve clinically meaningful endpoints such as polyp or adenoma detection in randomized and real-world settings; however, these gains are achieved by narrowly optimized, single-purpose models rather than by systems capable of broader interpretive reasoning [10-14].

Against this background, multimodal large language models (MM-LLMs) hold substantial potential [15]. By jointly processing images and text, MM-LLMs can, in principle, "see and say": integrate visual features with medical knowledge, generate free-text rationales, and condition decisions on clinical context [16]. Early reports suggest encouraging aggregate performance, but also reveal marked variability across lesions and tasks, indicating a role as assistive rather than autonomous readers at present [17].

Key gaps remain. First, it is unclear how state-of-the-art (SOTA) MM-LLMs perform on morphology-diverse, clinically difficult cystoscopic images curated as a stress test by domain experts. Second, the alignment between their free-text reasoning and expert judgment has not been systematically examined. Third, the practical utility of in-context learning (ICL) in cystoscopy—without task-specific fine-tuning—remains uncertain [18].

To address these gaps, our goal was to characterize the current capabilities and limitations of MM-LLMs in cystoscopic interpretation and to outline directions for model strengthening and additional adaptations required for safe clinical adoption.

Methods

Overview

Building on this objective, we (1) constructed a clinician-defined stress test that reflects real-world interpretive difficulty and spans benign and malignant lesions; (2) implemented a rater-blinded, model-anonymized evaluation across 2 complementary tasks—free-text image interpretation (4 open-ended questions plus a binary lesion detection query) and structured 7-class lesion classification; (3) mapped model outputs to clinically actionable binary endpoints (biopsy indication and malignancy); and (4) quantified the incremental benefit of ICL over zero-shot prompting. The overall study workflow is provided in Figure 1.

XSL•FO
RenderX

Figure 1. Study flow diagram. Evaluation pipeline for cystoscopic image interpretation and tumor-like lesion classification using 4 multimodal large language models (MM-LLMs). ICL: in-context learning; MCC: Matthews correlation coefficient; NPV: negative predictive value; PPV: positive predictive value.

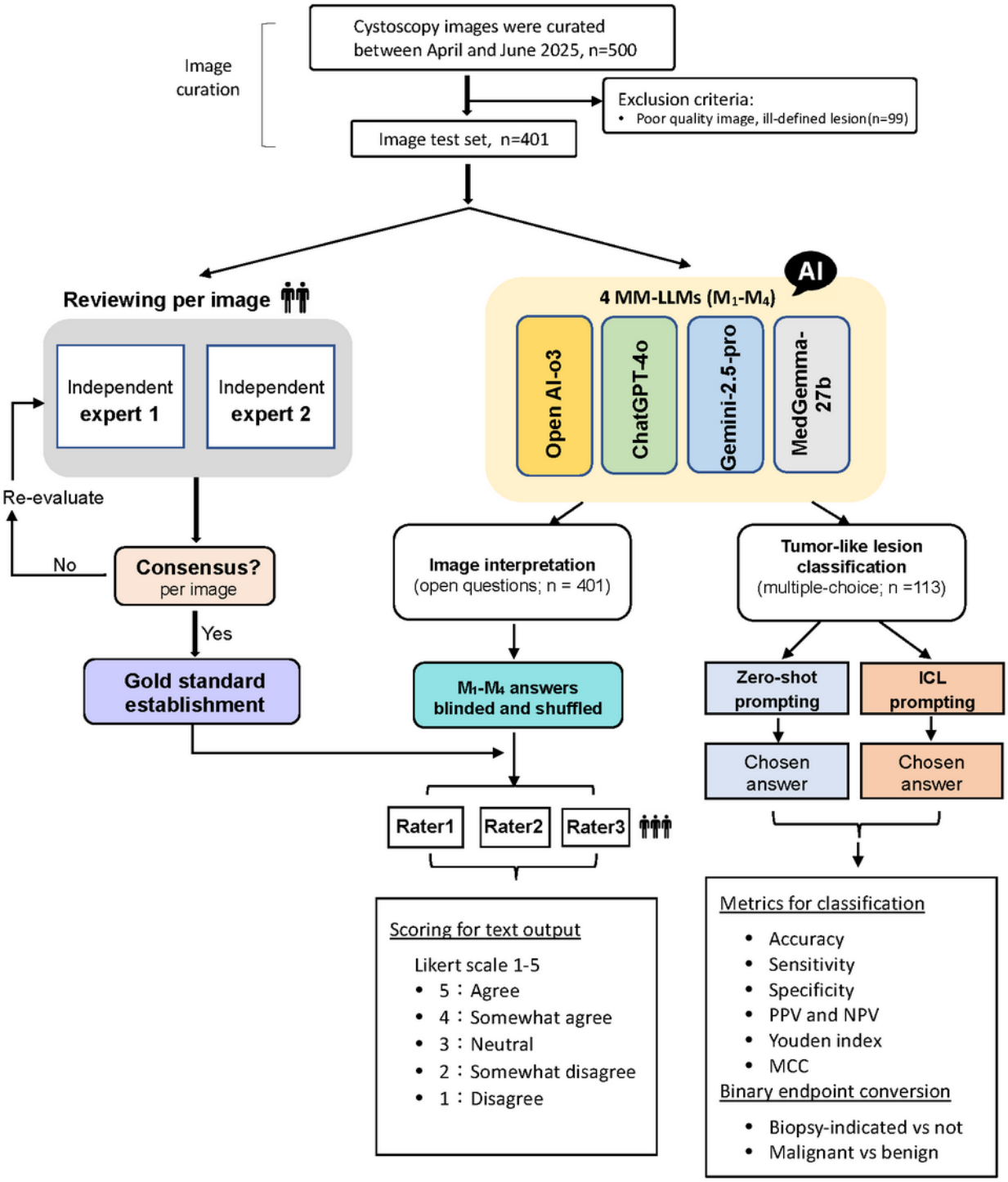

Curation of a Diverse Image Test Set

Inclusion Criteria and Diverse Lesion Coverage

We evaluated vision-enabled LLMs for cystoscopic image interpretation and tumor-like lesion classification. To stress-test model generalization, the test set was curated to maximize the diversity of morphological patterns rather than mirror clinical prevalence. Images were included only if they mapped to a prespecified schema of lower urinary tract presentations—normal anatomy, intraluminal nonmucosal conditions, and focal mucosal lesions—with finer-grained sublabels (eg, verumontanum, trabeculation, and papillary urothelial carcinoma [pUC]). To extend beyond routine cases, we deliberately sampled uncommon entities encountered in

XSL•FO
RenderX

practice, including endometriosis, miscellaneous deposits, fistulas, and erosion-related changes. This approach yielded a corpus with broad lesion coverage suitable for rigorous stress-testing cystoscopic interpretation.

Sources and Image Preprocessing

Images were curated between April and June 2025 from five sources: (1) an industry archive of intra-operative images captured on Karl-Storz endoscopes, (2) reference atlases (eg, Springer's *Diagnostic Cystoscopy* [19] and other urologic textbooks), (3) deidentified cystoscopy images obtained from websites, (4) open-access repositories accompanying PubMed-indexed papers and public datasets, and (5) Creative Commons–licensed surgical or teaching videos hosted on YouTube (Google). A total of 500 images were curated, and 99 were excluded due to poor image quality or an ill-defined lesion. The distribution of image sources and the memorization-test results, which were used to evaluate potential data-leakage risk from overlap with the 4 models' pretraining corpora, are summarized in Table S1 in Multimedia Appendix 1. Raw images were center-cropped to a square aspect ratio, resized to 800×800 pixels, and saved as JPEG files. This standardized pipeline harmonized the field of view and resolution across heterogeneous sources and ensured uniform inputs for all downstream model evaluations.

Multimodal LLMs

The evaluated MM-LLMs comprised 3 general-purpose models and 1 open-weight, medical-specific baseline. Two reasoning-optimized models—OpenAI-o3 [20] and Google Gemini 2.5 Pro [21]—were selected for their native image processing capabilities and emphasis on multistep reasoning. These represent the SOTA reasoning MM-LLMs available before July 2025. ChatGPT-4o (OpenAI) was included as a general-purpose, nonreasoning MM-LLM optimized for everyday assistance and among the earliest widely deployed models capable of accepting image input. MedGemma-27B (Google) [22] was intentionally included as a local and open-weight baseline—an open-source medical model (~27B parameters) suited for on-premises deployment and potential clinical fine-tuning. Because its parameter count and training budget are substantially smaller than the proprietary models (undisclosed), MedGemma-27B serves as a baseline rather than a capacity-matched comparator. Collectively, these systems span closed-source production platforms and an open-weight medical baseline, enabling a balanced, transparent comparison.

Study Design

Establishment of the Gold Standard

The reference standard diagnoses were determined through a multiphase, consensus-based process. Two urological experts, each with more than 25 years of clinical experience, independently reviewed all cystoscopic images, blinded to each other's assessments. The initial interexpert agreement was satisfactory (Cohen κ=0.81). In cases of disagreement, a consensus meeting was convened to establish a single unified diagnosis, integrating both normal anatomical features and pathological findings.

Image Interpretation and Lesion Classification Tasks

For a comprehensive evaluation, we designed 2 complementary tasks that reflect distinct components of diagnostic reasoning. The first, the image interpretation task, evaluated each model's capacity for domain-specific interpretation, logical reasoning, descriptive accuracy, and clinical judgment. The second, the lesion classification task, assessed the model's discriminative performance in differential diagnosis. Together, these 2 tasks provided a systematic assessment of MM-LLMs in both free-form interpretation and constrained classification settings, thereby capturing complementary dimensions of clinical decision-making.

Image Interpretation Task

The task used a structured, stepwise, open-ended question format. Each model was primed with a role-based instruction ("Suppose you are a urologist") and prompted with 4 sequential open-ended questions addressing anatomical site (Q1), findings (Q2), lesion reasoning (Q4), and final diagnosis (Q5). Q3 was not an open-ended interpretation item; it was a binary lesion-detection query (present or absent) embedded in the Q1→Q5 chain-of-thought to assess abnormal-versus-normal detection and was automatically graded against the gold standard. Free-text outputs for Q1, Q2, Q4, and Q5 were independently assessed by 3 raters (urology residents with 2-5 years of cystoscopy experience) using a 5-point Likert scale (1=disagree, 2=somewhat disagree, 3=neutral, 4=somewhat agree, and 5=agree).

Blinded and Randomized Evaluation Procedures

A dedicated evaluation software was developed to ensure complete rater blinding and randomization of both image presentation and model output order (Figure S1 in Multimedia Appendix 1).

- Image-level randomization: the display order of images was randomized once and shared across all raters. Each evaluation screen presented only 1 cystoscopic image at a time.
- Structured display: the upper panel displayed the image and its gold-standard answers to 5 reference questions (4 open-ended and 1 binary detection). The lower panel simultaneously presented anonymized text responses from the 4 MM-LLMs.
- Model-level randomization and anonymization: for each image, the order of model outputs was independently shuffled to minimize position bias. Model identities were fully anonymized to raters.
- Scoring process: raters independently scored the free-text responses using the 5-point Likert scale. The binary lesion-detection item (Q3) was automatically graded against the reference standard and was not rated by humans.

Lesion Classification Task

The lesion classification task was conducted to evaluate the models' discriminative capacity. It simulates a clinical scenario in which a urologist has already identified a tumor-like lesion and requires a differential diagnosis. A subset of tumor-like lesion images was used for this analysis. The task involved a 7-class multiple-choice framework comprising cystitis, polyps,

XSL•FO
RenderX

papilloma, pUC, CIS, nonurothelial carcinoma (non-U Ca), and none of the above (NOTA). Models were tested under 2 prompting strategies: zero-shot prompting and ICL. In the ICL condition, a text-based description of tumor-related features was incorporated into the prompt. This task aimed to assess each model's discriminative performance, adaptability to structured clinical classification, and robustness across prompting paradigms.

A subset of tumor-like lesion images was used for this task. The 7-class classification included cystitis, polyps, papilloma, pUC, CIS, non-U Ca, and NOTA. Models were tested under 2 settings—zero-shot prompting and ICL with added tumor-feature descriptions—to assess discriminative performance.

Clinically Relevant Binary Endpoint Conversion

To mirror real-world cystoscopic decision-making when tumor-like lesions are encountered, the 7-class classification task was collapsed into 2 clinically oriented binary endpoints. The first, the biopsy-indication endpoint, represented immediate clinical decision-making: pUC, CIS, non-U Ca, papilloma, and polyps were labeled as "biopsy indicated," whereas cystitis and NOTA were labeled as "biopsy not indicated." The second, the malignancy endpoint, classifies pUC, CIS, and non-U Ca as malignant, and cystitis, polyps, papilloma, and NOTA as nonmalignant. This mapping preserved the full 7-class framework for granular analysis while providing pragmatic outcomes aligned with bedside triage. Notably, papilloma—though histologically benign—was categorized as "biopsy indicated" to reflect the routine need for histologic confirmation.

Prompt Design

The complete and exact prompt designs are detailed in the Multimedia Appendix 1.

Prompt Design With Open-Ended Questions for Image Interpretation

We used a role-based, zero-shot prompt tailored to cystoscopy. The prompt primed domain reasoning ("Suppose you are a urologist") and briefly contextualized the procedure, followed by stepwise instructions to encourage explicit intermediate reasoning. The query comprised five domains: (1) anatomic site (free text), (2) endoscopic findings (free text), (3) presence or absence of a pathological lesion (binary), (4) lesion diagnostic reasoning and justification if present (free text), and (5) final diagnosis (free text).

Prompt Design for Tumor-Like Lesion Classification Task With Multiple-Choice Diagnostic Framework

We compared 2 prompting strategies for cystoscopic diagnosis of tumor-like lesions: zero-shot and ICL. Both adopted a role-based instruction ("Suppose you are a urologist"). The zero-shot prompt presented a single forced-choice 7-class label set. In contrast, the ICL prompt prefixed brief text-based descriptions of the 7 lesion classes before the same multiple-choice query. Models were instructed to provide the best diagnosis from the given options and include a concise rationale grounded in endoscopic morphology.

Outcome Measures and Statistical Analysis

For each image-question-answer instance, the 3 raters' Likert-scale ratings were averaged to obtain a single consensus score. The distribution of these scores across the test set was summarized using the mean and SD to describe the central tendency and variability of model performance. To compare performance among models, pairwise differences in scores were analyzed using paired *t* tests. Results were reported as mean differences with 95% CIs. Given the ordinal nature of Likert-scale data, Wilcoxon signed-rank tests were conducted as a sensitivity analysis. To account for the multiplicity of pairwise comparisons across the top 3 performing models, the Benjamini-Hochberg procedure was applied to control the false discovery rate and mitigate type I error. Consequently, statistical significance for all intermodel comparisons was defined as a false discovery rate–adjusted *P* value (q value)<.05. Subgroup analyses of final diagnosis were conducted according to cystoscopic finding categories and anatomic sites, following the same statistical procedures.

For interpretability, the mean Likert-scale score for each item was further converted into a binary satisfaction outcome: satisfactory if the mean score was >3 and unsatisfactory if ≤3. The satisfaction rate (percentage of satisfactory responses) was reported and used as a binary outcome in subsequent analyses.

The performance metrics for the classification tasks—including binary domains (lesion detection: present vs absent, biopsy indication: yes or no, and malignancy: yes or no) and the 7-class lesion classification—were derived from confusion matrices. Reported metrics included accuracy, sensitivity, specificity, positive predictive value, negative predictive value, Youden J, and the Matthews correlation coefficient (MCC) [23]. Youden J represents the overall diagnostic effectiveness of a test, defined as (sensitivity + specificity – 1), and reflects the balance between true-positive and true-negative rates. The MCC quantifies the overall agreement between predicted and actual classifications by incorporating all 4 components of the confusion matrix (true or false positives and negatives). Metric comparisons were conducted using the chi-square test.

For the 7-class task (n=113), models were instructed to select exactly 1 forced-choice label from the prespecified options. Outputs failing to provide a single permissible choice (eg, refusals such as "I could not answer this question") were coded as invalid. To ensure a consistent head-to-head comparison and minimize selection bias, the primary (strict) analysis used an intent-to-treat approach: invalid outputs were retained in the denominator and treated as incorrect predictions. However, because an invalid output does not necessarily reflect an incorrect diagnosis and may instead represent abstention—potentially safer than guessing in a human-in-the-loop workflow—we conducted a secondary sensitivity analysis, recalculating performance metrics conditional on valid responses only. All statistical analyses were conducted using SAS software (version 9.4; SAS Institute Inc).

Ethical Considerations

The Research Ethics Committee A of National Taiwan University Hospital determined that this study was exempt from

XSL•FO
RenderX

human participant research (NTUH-REC 202507210W). Informed consent was waived because this study involved a secondary analysis of deidentified cystoscopic images with no patient contact or intervention; for publicly available or published images, consent for the original collection followed the source publication, and the exemption permitted secondary analysis without additional consent. All images were deidentified, stored on access-controlled institutional systems, and reported only in aggregate. No participants were recruited, and no compensation was provided; all figures were reviewed to ensure no individual is identifiable.

Results

Distribution of the Whole Test Set and the Tumor-like Lesion Subset

Among 401 cystoscopic images, most originated from the bladder (n=329), followed by the prostate (n=41) and urethra (n=31). Abnormal findings were present in 322 (80.3%) images. The most common categories were tumor or neoplasm (n=126), structural or outlet abnormalities (n=76), inflammatory or reactive changes (n=69), deposits or foreign bodies (n=43), and vascular lesions (n=8); 79 (19.7%) images showed normal anatomy (Table 1). Table 1 provides the detailed distribution of finding subcategories, reflecting diagnostic diversity and difficulty.

XSL•FO
RenderX

Table 1. Detailed distribution of cystoscopic finding subcategories in the whole test dataset (N=401). This table provides a comprehensive breakdown of all observed cystoscopic findings across 3 hierarchical levels (normality, categories, and subcategories) and anatomic sites (bladder, prostate, and urethra). Values are presented as n (intracategory %); percentages represent the proportion of each subcategory within its respective parent category. The inclusion of both benign and malignant findings illustrates the heterogeneity of endoscopic presentations and underscores the diagnostic complexity represented in the dataset.

Anatomic site	Bladder	Prostate	Urethra	Total
Finding normality, categories, and subcategories, n (intracategory %)				
Abnormal	263	35	24	322
Tumor or neoplasm	114 (100)	—	12 (100)	126
Bladder polyp	7 (6.1)	—	—	7
Suspected bladder CIS[a]	17 (14.9)	—	—	17
Suspected nephrogenic adenoma	2 (1.8)	—	—	2
Papilloma	12 (10.5)	—	—	12
Papillary urothelial carcinoma	52 (45.6)	—	—	52
Nonurothelial carcinoma	18 (15.8)	—	—	18
Endometriosis	5 (4.4)	—	—	5
Teratoma	1 (0.9)	—	—	1
Urethral polyp	—	—	4 (33.3)	4
Urethral tumor	—	—	8 (66.7)	8
Inflammation or reaction	67 (100)	1 (100)	1 (100)	69
Bladder amyloidosis	4 (6.0)	—	—	4
Bladder keratinizing	6 (9.0)	—	—	6
Bladder malakoplakia squamous metaplasia	5 (7.5)	—	—	5
Bladder mucosal break	1 (1.5)	—	—	1
Cystitis	26 (38.8)	—	—	26
Hemorrhagic cystitis	5 (7.5)	—	—	5
Suspected IC[b]	10 (14.9)	—	—	10
Suspected radiation cystitis	7 (10.4)	—	—	7
Suspected Schistosomiasis	3 (4.5)	—	—	3
Urethritis	—	1 (100)	1 (100)	2
Deposits or foreign bodies	39 (100)	2 (100)	2 (100)	43
Bladder encrustation	5 (12.8)	—	—	5
Blood clot	12 (30.8)	1 (50)	1 (50)	14
Foreign body	7 (17.9)	—	1 (50)	8
Stone	15 (38.5)	1 (50)	—	16
Structure or outlet	35 (100)	32 (100)	9 (100)	76
Bladder diverticulum	4 (11.4)	—	—	4
Bladder neck contracture	4 (11.4)	—	—	4
Bladder scar	12 (34.3)	—	—	12
Bladder trabeculation	2 (5.7)	—	—	2
Vesicoureteral reflux	4 (11.4)	—	—	4
Ureterocele	6 (17.1)	—	—	6
Suspected fistula	3 (8.6)	—	2 (22.2)	5
Prostate enlargement	—	31 (96.9)	—	31
Prostatic cyst	—	1 (3.1)	—	1

XSL•FO
RenderX

Anatomic site	Bladder	Prostate	Urethra	Total
Urethra stricture	—	—	4 (44.4)	4
Urethral cyst	—	—	1 (11.1)	1
Urethral trauma	—	—	2 (22.2)	2
Vascularity	8 (100)	—	—	8
Bladder hemangioma	1 (12.5)	—	—	1
Bladder telangiectasia	4 (50.0)	—	—	4
Bladder varices	3 (37.5)	—	—	3
Normal	66	6	7	79

[a]CIS: carcinoma in situ.

[b]IC: interstitial cystitis.

The tumor-like lesion subset included 113 visually and pathologically similar images spanning both benign and malignant lesions: cystitis (n=18), polyps (n=7), papilloma (n=12), pUC (n=20), CIS (n=17), non-U Ca (n=17), and NOTA (n=22) (Table 2).

Table 2. Distribution of the tumor-like lesion subset (n=113) used for the 7-class lesion classification task, representing a focused subset of the whole test dataset. The tumor-like lesion subset comprised 18 cystitis (15.9%), 7 polyps (6.2%), 12 papilloma (10.6%), 20 papillary urothelial carcinoma (pUC; 17.7%), 17 carcinoma in situ (CIS; 15%), 17 non-urothelial carcinoma (non-U Ca; 15%), and 22 none of the above (NOTA; 19.5%).

Lesion type	Value, n (%)
Cystitis	18 (15.9)
Polyps	7 (6.2)
Papilloma	12 (10.6)
pUC[a]	20 (17.7)
CIS[b]	17 (15)
Non-U Ca[c]	17 (15)
NOTA[d]	22 (19.5)
Total	113 (100)

[a]pUC: papillary urothelial carcinoma.

[b]CIS: carcinoma in situ.

[c]Non-U Ca: non-urothelial carcinoma.

[d]NOTA: none of the above.

Comparative Mean Scores of LLMs in Image Interpretation

Among the whole test set (n=401), MM-LLMs demonstrated progressively lower performance as task complexity increased (Table 3 and Figure 2). Mean Likert-scale scores declined from anatomic site recognition (≈ 4.1) to findings (≈ 3.4-3.7), lesion reasoning (≈ 2.7-2.9), and final diagnosis (≈ 2.6-2.8). OpenAI-o3, ChatGPT-4o, and Gemini 2.5 Pro achieved comparable accuracy in anatomical localization, while OpenAI-o3 showed the highest overall consistency and clarity in lesion description. Statistically significant differences emerged in the findings, lesion reasoning, and final diagnosis domains, in which OpenAI-o3 outperformed Gemini 2.5 Pro and the medical-specific MedGemma-27B. Notably, MedGemma-27B lagged substantially behind the general-purpose MM-LLMs across all categories, suggesting that its limited training scope constrained both descriptive precision and diagnostic reasoning. These results indicate that reasoning-optimized general-purpose MM-LLMs currently outperform open-source, domain-specific models in free-text cystoscopic interpretation tasks.

XSL•FO
RenderX

Table 3. Performance of 4 multimodal large language models (MM-LLMs) in cystoscopic image interpretation, presented as mean Likert scores and SDs across open-ended questions and final-diagnosis subgroups.

Question and subgroup	Value	OpenAI-o3, mean (SD)	ChatGPT-4o, mean (SD)	Gemini 2.5 Pro, mean (SD)	MedGemma-27B, mean (SD)
Whole test set (n=401)					
Questions					
Q1: anatomic site	401	4.13 (1.24)	4.06 (1.20)	4.10 (1.21)	2.45 (1.37)
Q2: findings	401	3.69 (1.23)	3.54 (1.25)	3.42 (1.30)	1.80 (1.03)
Q4: lesion reasoning	401	2.94 (1.57)	2.89 (1.49)	2.70 (1.48)	1.62 (1.11)
Q5: final diagnosis	401	2.79 (1.59)	2.75 (1.51)	2.61 (1.51)	1.48 (0.93)
Q5. final diagnosis					
Subgrouping by findings					
Tumor or neoplasm	126	3.12 (1.44)	3.32 (1.36)	3.07 (1.51)	2.06 (1.18)
Inflammation or reaction	69	2.34 (1.21)	2.82 (1.30)	2.87 (1.29)	1.19 (0.45)
Deposits or foreign bodies	43	2.77 (1.59)	2.97 (1.52)	2.64 (1.64)	1.07 (0.26)
Structure or outlet	76	1.71 (1.13)	1.45 (0.69)	2.45 (1.67)	1.03 (0.10)
Vascularity	8	2.42 (1.05)	2.75 (0.71)	2.75 (0.87)	1.29 (0.70)
Normal	79	3.79 (1.76)	2.92 (1.81)	1.76 (1.13)	1.49 (1.04)
Subgrouping by anatomic site					
Bladder	329	3.04 (1.56)	3.06 (1.46)	2.64 (1.47)	1.56 (1.00)
Prostate	41	1.45 (1.02)	1.24 (0.62)	2.95 (1.79)	1.09 (0.31)
Urethra	31	2 (1.34)	1.48 (0.91)	1.80 (1.27)	1.12 (0.38)

Figure 2. Comparative mean scores of multimodal large language models (MM-LLMs) for cystoscopic image interpretation across 4 question domains in the whole test set. Asterisks denote statistically significant pairwise differences among the top 3 models. *q<0.05, where q is the false discovery rate (FDR)-adjusted *P* value.

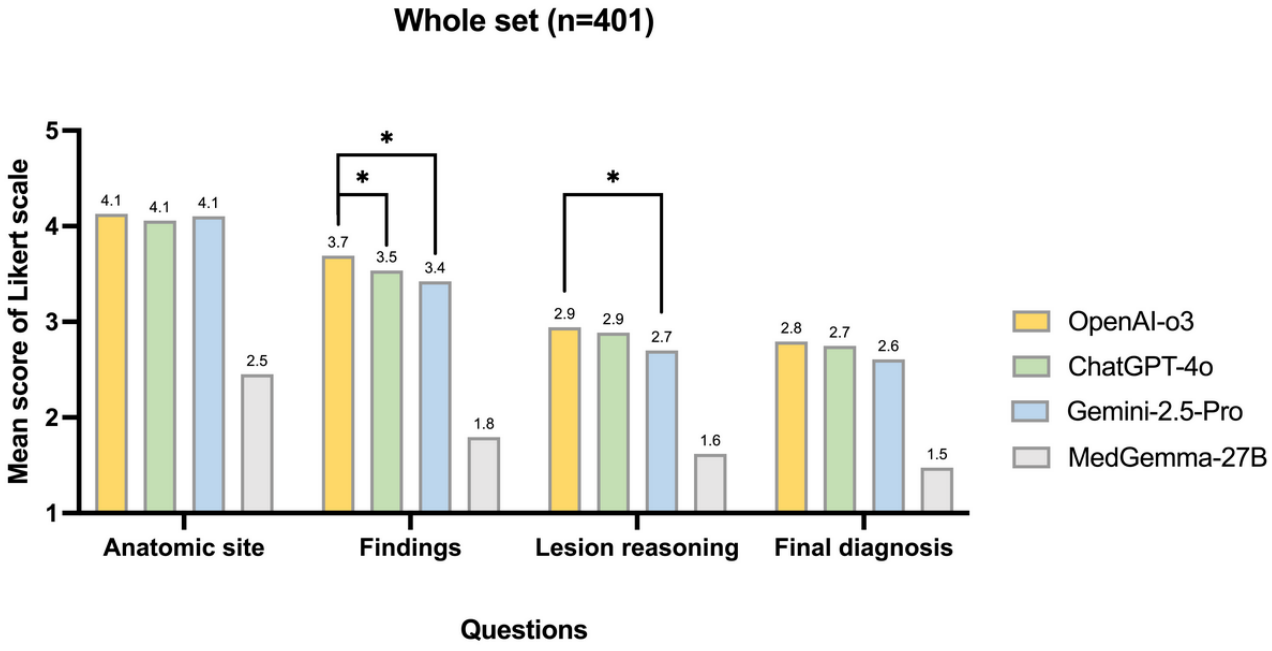

Mean-score pairwise comparisons among models (mean-score deltas) are provided in Table S2 in Multimedia Appendix 1. A 0.2-point difference on the 5-point Likert score corresponds approximately to a 5-point difference on a 100-point scale. The matrix of column-row differences confirmed OpenAI-o3's edge across open-question domains. Versus Gemini 2.5 Pro, OpenAI-o3 scored higher by +0.27 on findings (q=0.002), +0.24 on lesion reasoning (q=0.047), +0.03 on anatomic site, and +0.19 on final diagnosis (not significant). Against ChatGPT-4o, OpenAI-o3 held a small but consistent advantage on findings

XSL•FO
RenderX

(+0.15, q=0.004) with near-parity on anatomic site, lesion reasoning, and final diagnosis (+0.07, +0.06, +0.04; not significant). All general-purpose models substantially outperformed the medical-specific MedGemma-27B; OpenAI-o3's margins were +1.68 (anatomic site), +1.90 (findings), +1.32 (lesion reasoning), and +1.32 (final diagnosis), all q<0.001. Taken together, these deltas indicate that OpenAI-o3 is the most reliable free-text interpreter, with the largest, statistically robust gains in content-heavy domains (Findings → Reasoning → Diagnosis). The significance pattern of the Wilcoxon signed-rank tests was consistent with that of the paired *t* tests.

Intraclass correlation coefficients demonstrated excellent interrater reliability across both model and question domains. Of the 16 intraclass correlation coefficient values, 14 ranged from 0.82 to 0.94, indicating high consistency among raters (Table S3 in Multimedia Appendix 1).

Model Satisfaction Rates

Satisfaction rates for each question across models closely paralleled mean scores of the Likert scale (Figure 3). Overall, satisfaction ranked anatomic site > findings > lesion reasoning ≈ final diagnosis, consistent with mean-score trends. Anatomic site showed uniformly high satisfaction for the top 3 models (339/401, ≈85%), while MedGemma-27B was much lower (184/401, 46%). For findings, OpenAI-o3 (305/401, 76%) and ChatGPT-4o (297/401, 74%) outperformed Gemini 2.5 Pro (277/401, 69%) and MedGemma-27B (92/401, 23%; all q<0.01). In lesion reasoning, OpenAI-o3 (211/401, 53%) and ChatGPT-4o (201/401, 53%) outperformed Gemini 2.5 Pro (184/401, 46%; q=0.003 and q=0.002 vs Gemini 2.5 Pro, respectively), while MedGemma-27B again had the lowest performance (72/401, 18%). For final diagnosis, satisfaction was lowest overall but remained higher for OpenAI-o3 and ChatGPT-4o (192/401, ≈48%) than for Gemini 2.5 Pro (168/401, 42%) or MedGemma-27B (60/401, 15%).

Figure 3. Comparative satisfaction rates (% of cases with mean score >3) of multimodal large language models (MM-LLMs) for cystoscopic image interpretation across 4 question domains. Asterisks denote significance for pairwise comparisons between the top 3 models. *q<0.05, where q is the false discovery rate (FDR)-adjusted *P* value.

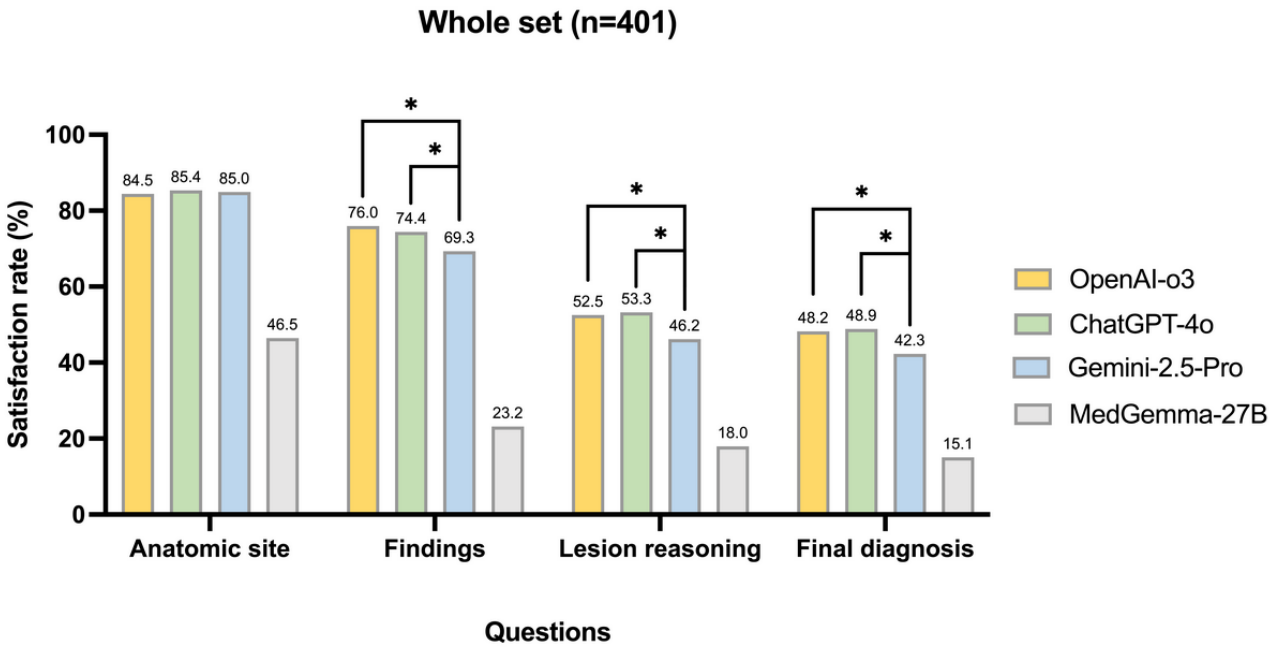

Subgroup Analysis of Final Diagnosis

When mean scores for final diagnosis were stratified by finding category (Figure 4 and Table 3), the largest intermodel differences occurred in the normal and structure or outlet groups. OpenAI-o3 achieved the highest score for normal findings (3.79), significantly outperforming ChatGPT-4o (2.92) and Gemini 2.5 Pro (1.76; q<0.001). Conversely, Gemini 2.5 Pro scored best for structure or outlet findings (2.45), exceeding OpenAI-o3 (1.71) and ChatGPT-4o (1.45; q<0.001). For tumor or neoplasm, ChatGPT-4o slightly surpassed OpenAI-o3 (3.32 vs 3.12; q=0.02), with Gemini 2.5 Pro showing comparable performance (3.07). Inflammation or reaction and vascularity categories both favored ChatGPT-4o and Gemini 2.5 Pro over OpenAI-o3, whereas deposits or foreign bodies showed minimal differences among models.

XSL•FO
RenderX

Figure 4. Subgroup analysis of mean final diagnosis scores across 6 cystoscopic finding categories (tumor or neoplasm, inflammation, deposits, structure, vascularity, and normal) for 4 multimodal large language models (MM-LLMs). Asterisks denote significance for pairwise comparisons between the top 3 models. *q<0.05, where q is the false discovery rate (FDR)-adjusted *P* value.

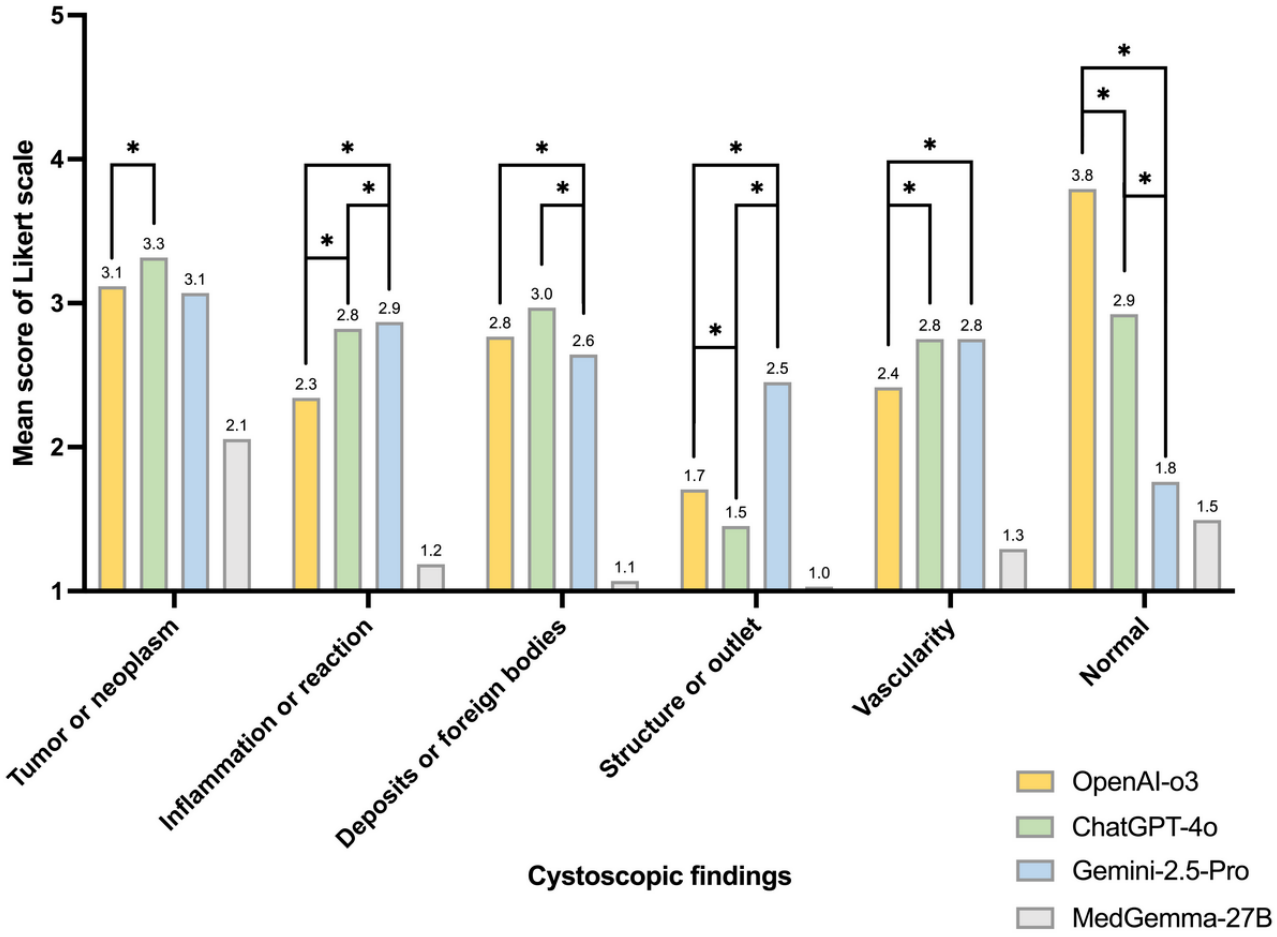

Performance also varied by anatomic site (Figure 5 and Table 3). Gemini 2.5 Pro performed best in the prostate (2.95), significantly exceeding OpenAI-o3 (1.45) and ChatGPT-4o (1.24; q<0.001). In the bladder, OpenAI-o3 (3.04) and ChatGPT-4o (3.06) outperformed Gemini 2.5 Pro (2.64; q<0.001). For the urethra, OpenAI-o3 (2) exceeded ChatGPT-4o (1.48; q=0.02), with Gemini 2.5 Pro intermediate (1.80). These site-specific trends suggest complementary strengths: Gemini 2.5 Pro performs relatively better in structure-dominated prostate views, whereas OpenAI-o3 and ChatGPT-4o perform best in bladder-focused interpretation.

XSL•FO
RenderX

Figure 5. Subgroup analysis of mean final diagnosis scores across 3 anatomic sites (bladder, prostate, and urethra) for 4 multimodal large language models (MM-LLMs). Asterisks denote significance for pairwise comparisons between the top 3 models. *q<0.05, where q is the false discovery rate (FDR)-adjusted *P* value.

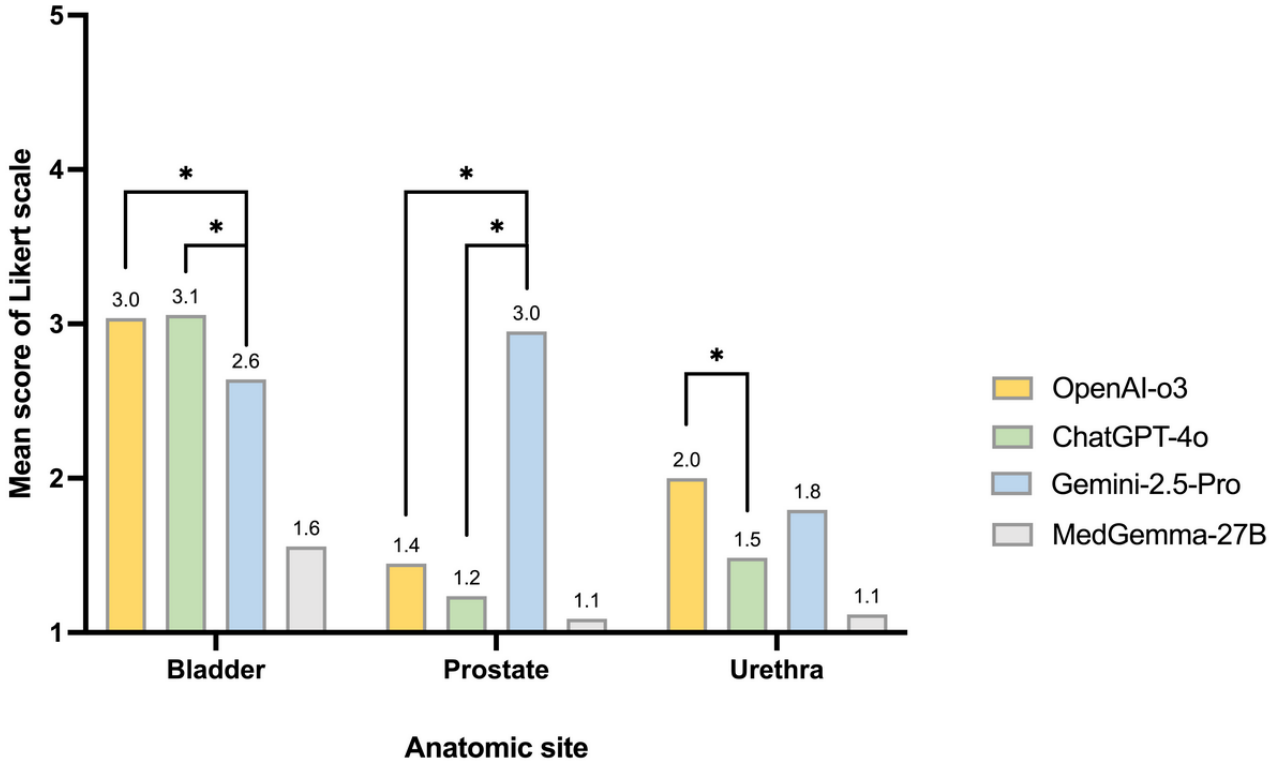

Clinically Relevant Binary Endpoints

For the lesion-detection task in the whole test set, OpenAI-o3 achieved the highest overall performance, with an accuracy of 88.3%, a Youden J of 0.650, and an MCC of 0.635, followed by ChatGPT-4o and Gemini 2.5 Pro, while MedGemma-27B performed the lowest (Table 4). OpenAI-o3 demonstrated the most balanced profile (sensitivity 92% and specificity 73.1%), whereas ChatGPT-4o showed higher sensitivity but lower specificity (94.4% vs 44.2%). Gemini 2.5 Pro exhibited an extreme trade-off—maximal sensitivity (99.7%) but very low specificity (10.3%). MedGemma-27B produced the weakest results overall (accuracy 45.6%, Youden J –0.103, and MCC –0.081).

XSL•FO
RenderX

Table 4. Binary classification performance of 4 multimodal large language models (MM-LLMs) in clinically relevant cystoscopic endpoints (strict analysis).

Task and MM-LLM[a]	VRR[b] (%)	Acc[c] (%)	Sen[d] (%)	Spec[e] (%)	PPV[f] (%)	NPV[g] (%)	Youden J[h]	MCC[i]
Whole test set (n=401)								
Lesion detection (present vs absent)								
OpenAI-o3	100	88.3	92.0	73.1	93.4	68.7	0.650	0.635
ChatGPT-4o	100	84.7	94.4	44.2	87.6	65.4	0.386	0.452
Gemini 2.5 Pro	100	82.3	99.7	10.3	82.1	88.9	0.100	0.266
MedGemma-27B	100	45.6	46.1	43.6	77.2	16.4	-0.103	-0.081
Tumor-like lesion subset (n=113)								
Biopsy indication (yes or no): zero-shot prompting								
OpenAI-o3	99.1	73.5	82.2	57.5	77.9	63.9	0.397	0.407
ChatGPT-4o	92.9	69.0	91.8	27.5	69.8	64.7	0.193	0.258
Gemini 2.5 Pro	100	70.8	86.3	42.5	73.3	63.0	0.288	0.323
MedGemma-27B	100	65.5	97.3	7.5	65.7	60.0	0.048	0.111
Biopsy indication (yes or no): in-context learning								
OpenAI-o3	100	76.1	80.8	67.5	81.9	65.6	0.483	0.481
ChatGPT-4o	98.2	69.0	89.0	32.5	70.7	61.9	0.215	0.265
Gemini 2.5 Pro	100	69.0	75.3	57.5	76.4	56.1	0.328	0.327
MedGemma-27B	100	60.2	93.2	0.0	63.0	0.0	-0.069	-0.159
Presence of malignancy (yes or no): zero-shot prompting								
OpenAI-o3	99.1	62.8	79.6	47.5	58.1	71.8	0.271	0.285
ChatGPT-4o	92.9	55.8	81.5	32.2	52.4	65.5	0.137	0.157
Gemini 2.5 Pro	100	61.1	87.0	37.3	56.0	75.9	0.243	0.278
MedGemma-27B	100	57.5	72.2	44.1	54.2	63.4	0.163	0.169
Presence of malignancy (yes or no): in-context learning								
OpenAI-o3	100	63.7	70.4	57.6	60.3	68.0	0.280	0.282
ChatGPT-4o	98.2	59.3	61.1	57.6	56.9	61.8	0.187	0.187
Gemini 2.5 Pro	100	62.0	66.7	57.6	59.0	65.4	0.243	0.244
MedGemma-27B	100	52.2	87.0	20.3	50.0	63.2	0.074	0.099

[a]MM-LLM: multimodal large language model.

[b]VRR: valid response rate. Valid response rate = (total - invalid) / total. Invalid denotes outputs failing to provide a single permissible choice.

[c]Acc: accuracy.

[d]Sen: sensitivity.

[e]Spe: specificity.

[f]PPV: positive predictive value.

[g]NPV: negative predictive value.

[h]Youden J: Youden J Index.

[i]MCC: Matthews correlation coefficient.

In the tumor-like lesion subset (n=113), OpenAI-o3 again achieved the highest Youden J and MCC performance for both biopsy-indication and malignancy endpoints, followed by Gemini 2.5 Pro and ChatGPT-4o, with MedGemma-27B lowest. For biopsy indication, OpenAI-o3 reached 73.5% accuracy (Youden J=0.397 and MCC=0.407), demonstrating the best specificity-sensitivity balance. ICL modestly improved specificity and accuracy (Table 4). For malignancy detection, OpenAI-o3 similarly performed best (accuracy 62.8%), followed by Gemini 2.5 Pro and ChatGPT-4o, whereas MedGemma-27B underperformed (Tables S4-S5 in Multimedia Appendix 1).

Overall, OpenAI-o3 demonstrated the most balanced diagnostic performance across all 3 binary endpoints, consistently achieving the highest Youden J, largely attributable to its superior specificity (Figure 6). For lesion detection, specificity reached 73.1%, significantly outperforming ChatGPT-4o

XSL•FO
RenderX

(44.2%; q<0.001) and Gemini 2.5 Pro (10.3%; q<0.001; Figure 6A). In biopsy indication, specificity was 57.5%, again higher than ChatGPT-4o (27.5%; q<0.001) and Gemini 2.5 Pro (42.5%; q=0.047; Figure 6B). In malignancy prediction, OpenAI-o3 maintained the highest specificity (47.5%) compared to ChatGPT-4o (32.2%; q=0.042) and Gemini 2.5 Pro (37.3%; q=0.21; Figure 6C). In contrast, MedGemma-27B demonstrated limited generalizability despite its medical-domain optimization.

Figure 6. Radar charts illustrating diagnostic performance across 3 cystoscopic endpoints in classification tasks. Panels represent (A) lesion detection (presence vs absence), (B) biopsy indication (yes vs no), and (C) presence of malignancy (yes vs no). Five key metrics are visualized: accuracy, sensitivity, specificity, Youden J index, and Matthews correlation coefficient (MCC).

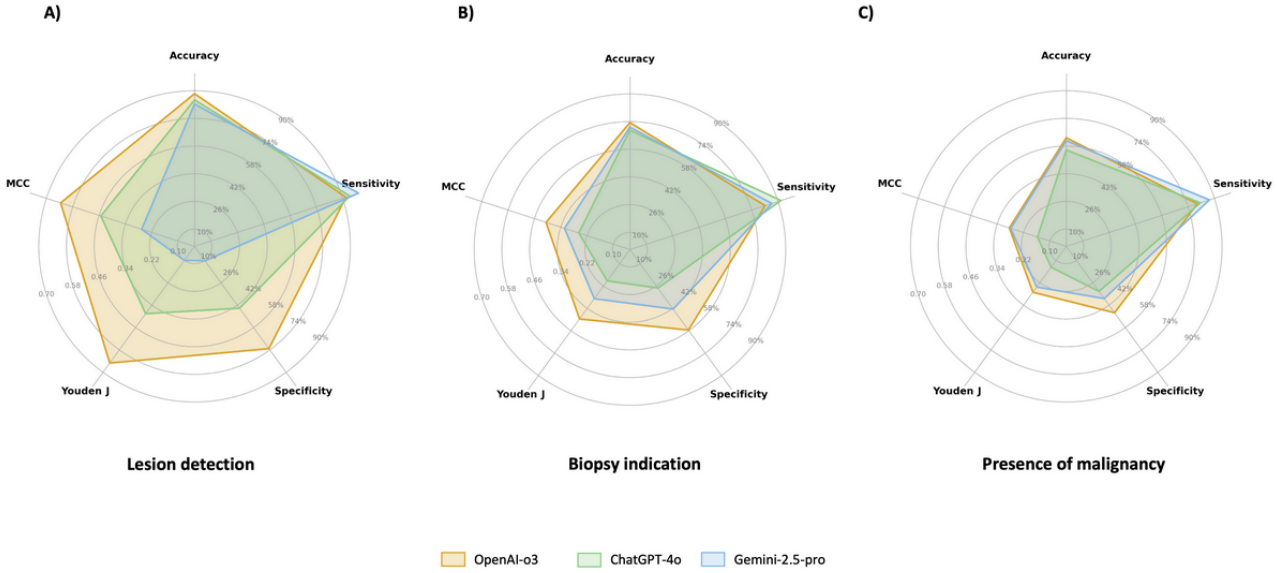

Seven-Class Tumor-Like Lesion Classification

For the 7-class classification (n=113), OpenAI-o3 remained the top-performing model, although overall accuracy was modest, improving only slightly with ICL (microaverage accuracy from 40.7% to 46% and MCC from 0.311 to 0.370) (Table 5). Class-level performance was heterogeneous. In the zero-shot setting, malignant categories (pUC and CIS) achieved relatively balanced sensitivity and specificity, whereas benign lesions (cystitis, polyps, and papilloma) showed high specificity but low sensitivity. Notably, non-U Ca were not recognized. ICL mainly adjusted the sensitivity-specificity balance—enhancing detection of polyps, papilloma, and NOTA while slightly reducing sensitivity for cystitis and CIS. However, non-U Ca remained unrecognized.

XSL•FO
RenderX

Table 5. Confusion matrix outlining the performance of OpenAI-o3 in 7-class tumor-like lesion classification under zero-shot and in-context learning prompting (strict analysis).

	Actual							Total
	Cystitis	Polyps	Papilloma	pUC[a]	CIS[b]	Non-U Ca[c]	NOTA[d]	
Predicted								
Zero-shot prompting								
Cystitis	6	1	1	0	2	1	3	14
Polyps	1	1	0	0	0	1	0	3
Papilloma	0	0	0	0	0	0	0	0
pUC	2	4	11	20	0	14	7	58
CIS	5	0	0	0	9	0	2	16
Non-U Ca	0	0	0	0	0	0	0	0
NOTA	4	1	0	0	5	1	10	21
Invalid[e]	0	0	0	0	1	0	0	1
Total	18	7	12	20	17	17	22	113
ICL[f] prompting								
Cystitis	6	0	1	1	6	0	4	18
Polyps	0	3	0	0	0	3	0	6
Papilloma	0	0	3	0	0	0	0	3
pUC	2	4	8	19	0	12	4	49
CIS	7	0	0	0	7	0	0	14
Non-U Ca	0	0	0	0	0	0	0	0
NOTA	3	0	0	0	4	2	14	23
Total	18	7	12	20	17	17	22	113
Classification metrics								
Zero-shot prompting								
Accuracy	82.3	92.9	89.4	66.4	86.7	85.0	79.6	40.7
Sensitivity	33.3	14.3	0	100	52.9	0	45.5	40.7
Specificity	91.6	98.1	100	59.1	92.7	100	87.9	90.3
PPV[h]	42.9	33.3	0	34.5	56.3	0	47.6	41.1
NPV[i]	87.9	94.5	89.4	100	91.8	85.0	87.0	90.1
Youden J[j]	0.249	0.124	0	0.591	0.456	0	0.334	0.310
MCC[k]	0.277	0.186	0	0.452	0.468	0	0.340	0.311
ICL prompting								
Accuracy	78.8	93.8	92.0	72.6	85.0	85.0	85.0	46.0
Sensitivity	33.3	42.9	25.0	95.0	41.2	0	63.6	46.0
Specificity	87.4	97.2	100	67.7	92.7	100	90.1	91.0
PPV	33.3	50.0	100	38.8	50	0.0	60.9	46.0
NPV	87.4	96.3	91.8	98.4	89.9	85.0	91.1	91.0
Youden J	0.207	0.400	0.250	0.627	0.339	0	0.538	0.370
MCC	0.207	0.430	0.479	0.483	0.368	0	0.529	0.370

[a]pUC: papillary urothelial carcinoma.

[b]CIS: carcinoma in situ.

XSL•FO
RenderX

[c]Non-U Ca: non-urothelial carcinoma.

[d]NOTA: none of the above.

[e]Invalid: model outputs failing to provide a single permissible choice.

[f]ICL: in-context learning.

[g]AVG: microaverage.

[h]PPV: positive predictive value.

[i]NPV: negative predictive value.

[j]Youden J: Youden J Index.

[k]MCC: Matthews correlation coefficient.

The other 3 models performed suboptimally and showed limited responsiveness to ICL (Tables S6-S8 in Multimedia Appendix 1). In the zero-shot setting, microaveraged accuracy ranked as follows: 36.3% (Gemini 2.5 Pro); 31.9% (ChatGPT-4o), 28.3% (MedGemma-27B), with Youden J and MCC both 0.164-0.257. Under ICL, accuracy was similar or slightly lower—34.5% (Gemini 2.5 Pro), 31% (ChatGPT-4o), and 27.4% (MedGemma-27B)—with minimal shifts in Youden J and MCC (0.153-0.236). Class-wise patterns were consistent: malignant categories (pUC, CIS, and non-U Ca) showed the most balanced sensitivity-specificity trade-offs, whereas benign entities (cystitis, polyps, and papilloma) had low sensitivity but high specificity.

Analysis of the NOTA Category

The NOTA category represents a unique "negative exclusion" challenge. Previous research indicates that LLMs often exhibit a bias toward positive selection, struggling to confidently select "None of the Above" even when accurate [24,25]. Our results show that this bias is pervasive, affecting models across different architectures (Tables S6-S8). Despite being a reasoning-optimized model, Gemini 2.5 Pro aligned with the general-purpose ChatGPT-4o and the medical-specific MedGemma-27B in exhibiting a "high specificity, low sensitivity" pattern. Specifically, Gemini 2.5 Pro and ChatGPT-4o achieved high specificity (>97%) but low sensitivity (9.5%-27.3%) across prompting strategies. The open-weight MedGemma-27B exhibited the most severe manifestation of this bias: while its specificity remained high (98.9%), its sensitivity was only 13.6% in the zero-shot setting and collapsed to 0% under ICL prompting. This indicates that for models unable to effectively leverage negative logic, added textual context may inadvertently reinforce positive selection bias.

A distinct divergence was observed between the 2 reasoning-optimized models. In contrast to Gemini 2.5 Pro, OpenAI-o3 demonstrated superior handling of exclusion (Table 5). It achieved a significantly higher baseline sensitivity of 45.5% in the zero-shot setting. Moreover, while ICL yielded negligible or detrimental effects for the other 3 models, OpenAI-o3's sensitivity surged to 63.6% under ICL prompting. This suggests that OpenAI-o3's specific implementation of chain-of-thought reasoning is critical for overcoming the standard positive selection bias, allowing for robust diagnosis through exclusion where other reasoning and general models failed.

Sensitivity Analysis: Conditional on Valid Responses

In the tumor-like lesion classification task, invalid (refusal) outputs were uncommon: valid-response rates were ≈100% for OpenAI-o3, Gemini 2.5 Pro, and MedGemma-27B, whereas ChatGPT-4o had the highest invalid rate (7.1%) in zero-shot prompting. While excluding invalid responses can inflate performance (introducing optimistic bias) relative to the strict analysis, invalid outputs can be interpreted clinically as abstention, which may be safer than guessing in a human-in-the-loop workflow because it prompts clinician confirmation. Accordingly, we report conditional-on-valid performance to better reflect accuracy when the model provides a valid output (Table S9 in Multimedia Appendix 1).

ChatGPT-4o showed the largest strict vs conditional-on-valid differences, consistent with its lower valid-response rate (Table 4 vs Table S9 in Multimedia Appendix 1). Accuracy and Youden J increased from 69% and 0.193 to 74.3% and 0.267 for biopsy indication and from 55.8% and 0.137 to 60% and 0.205 for malignancy. Seven-class changes were small, most notably higher sensitivity for NOTA (from 13.6% to 15%) and cystitis (from 11.1% to 14.3%), with microaverage sensitivity rising from 31.9% to 34.3%.

ICL-focused takeaway across models was that text-only ICL chiefly reweighted sensitivity-specificity rather than boosting overall accuracy; modest gains in benign or NOTA recognition were offset by reduced sensitivity in key malignant classes.

Discussion

Principal Findings

This study is the first to benchmark SOTA MM-LLMs for cystoscopic interpretation under a clinician-defined stress test with rare and diagnostically difficult lesions. The rigorous, blinded design enabled objective assessment of interpretive reasoning and classification. Outputs were also mapped to actionable binary endpoints and used to quantify the incremental effect of text-based ICL over zero-shot prompting, thereby revealing both strengths and current limitations of MM-LLMs in real-world clinical tasks.

Overall, OpenAI-o3 demonstrated superior performance, followed by ChatGPT-4o and Gemini 2.5 Pro, with MedGemma-27B showing the most limited capabilities. The results revealed a progressive decline in model performance as diagnostic complexity increased—from anatomical recognition to higher-order diagnostic synthesis. While models showed meaningful strength in visual recognition and descriptive reporting, performance in challenging differential diagnosis

XSL•FO
RenderX

remained modest, suggesting that current MM-LLMs function best as assistive rather than autonomous diagnostic tools at present.

Image Interpretation and Lesion Classification Tasks

The free-text interpretation task assessed 4 domains of increasing complexity—anatomic site, findings, lesion reasoning, and final diagnosis—simulating real-world diagnostic synthesis that integrates visual recognition with clinical reasoning. Task satisfaction declined with increasing complexity, from anatomic localization (~85%) to definitive diagnosis (~45%). OpenAI-o3 and ChatGPT-4o consistently outperformed Gemini 2.5 Pro and MedGemma-27B, though with distinct profiles: OpenAI-o3 produced concise, accurate descriptions with coherent diagnostic impressions and high specificity for normal anatomy, while ChatGPT-4o showed greater sensitivity for inflammatory and vascular findings. In contrast, Gemini 2.5 Pro often overcalled minor irregularities but performed better on prostate lesions, likely reflecting prostate-predominant pretraining. These discrepancies indicate that MM-LLM behavior depends not only on recognition accuracy but also on underlying reasoning logic, diagnostic thresholds, and domain-specific pretraining.

Regarding clinical decision endpoints, the goal of cystoscopy is to identify abnormal lesions—particularly malignancies—so that biopsies are performed when necessary while avoiding unnecessary procedures that increase cost and risk. Thus, lesion detection, biopsy indication, and malignancy presence were defined as key clinical endpoints. OpenAI-o3 achieved the most balanced performance across sensitivity, specificity, and Youden J, outperforming ChatGPT-4o and Gemini 2.5 Pro, especially in specificity, by accurately distinguishing normal from malignant cases—supporting appropriate biopsy decision-making. These findings highlight the importance of calibrating operating points to clinical priorities and suggest that MM-LLMs, particularly OpenAI-o3, can aid cystoscopic decision-making when optimized for an appropriate specificity-sensitivity balance.

The 7-class lesion classification task evaluated each model's ability to distinguish visually and pathologically similar tumor-like lesions. Models performed best on prevalent malignant lesions (pUC) but struggled with benign mimickers (polyps and papilloma) and rare entities (non-U Ca), often misclassified as pUC—reflecting limited pretraining exposure to uncommon classes. As shown in Table 5, 14 of 17 non-U Ca cases (82.4%) were predicted as pUC, a much more prevalent bladder tumor. One plausible explanation is that LLMs exhibit a tendency to choose majority or high-frequency labels in multiple-choice settings. When pretraining class distributions are imbalanced, the token sequences corresponding to common options (eg, "pUC") can carry higher previous probabilities, biasing the model toward these answers irrespective of correctness. This phenomenon is often described as majority-label bias or common-token bias [26].

On the other hand, the frequent misclassification of papilloma as pUC highlights the inherent challenge of distinguishing these entities based solely on cystoscopic appearance—a difficulty shared by human experts. Rather than indicating model failure, these confusion patterns reflect the substantial macroscopic overlap between papilloma and low-grade pUC. While visual distinction remains experimental, our study addressed this clinical reality by grouping both entities under the "Biopsy Indicated" category in the binary endpoint analysis. In that context, the models successfully flagged these lesions for histologic confirmation, aligning with standard safety protocols despite the specific classification ambiguity. CIS, a flat, high-grade, non-invasive UC subtype, was handled well by OpenAI-o3, achieving strong results (accuracy 86.7, sensitivity 52.9, specificity 92.7, and Youden J 0.46) despite diagnostic difficulty. Overall, OpenAI-o3 showed the most balanced performance, particularly excelling in benign and NOTA classifications, achieving higher specificity than ChatGPT-4o and Gemini 2.5 Pro.

Performance Disparity and the Role of the Open-Weight Baseline

Although MedGemma-27B is a medical-specific model, its performance trailed behind the general-purpose proprietary models (OpenAI-o3, ChatGPT-4o, and Gemini 2.5 Pro). This gap can be attributed to 2 primary factors: domain-specific data misalignment and model scale. First, contrary to the expectation that a medical model should inherently outperform general models, MedGemma's training distribution did not encompass the specific modality of cystoscopy. While its multimodal components (SigLIP encoder) were rigorously pretrained on diverse medical datasets—including chest X-rays, dermatology images, ophthalmology images, and histopathology slides—endoscopic imagery was notably absent from its pretraining corpus. Consequently, the model faced a "zero-shot" challenge in a domain it had not explicitly learned, whereas the massive general-purpose models likely benefited from broader exposure to endoscopic images present in their web-scale training data.

Second, as a local and open-weight baseline, MedGemma-27B (≈27B parameters) operates with significantly constrained capacity compared to the proprietary SOTA architectures. It lacks the extensive parameter count, training budget, and chain-of-thought optimization that allow models such as OpenAI-o3 to generalize across unseen tasks. Therefore, our intent is not to claim parity with these massive systems, but to establish a transparent performance floor for open-weight deployment. Despite the lower accuracy in this zero-shot setting, MedGemma-27B remains a critical benchmark for institutions requiring on-premise, privacy-preserving solutions. Its performance represents the current starting point for future adaptation, such as LoRA fine-tuning or retrieval-augmented prompting, rather than a direct competitor in raw reasoning capability.

Comparing With Previous Studies

Guo et al. [17] reported comparable findings when evaluating ChatGPT-4V (OpenAI) and Claude-3.5 (Anthropic) on 603 cystoscopic images, achieving accuracies of 82.8% and 79.8% but with marked variability across conditions. Both models performed well for cystitis and bladder tumors but poorly for BPH and normal structures, indicating that general-purpose LLMs detect major lesions with high sensitivity but struggle

XSL•FO
RenderX

with subtle findings. Similar variability has been observed in gastrointestinal endoscopy, where ChatGPT-4V showed mixed accuracy across lesion types and underperformed relative to tuned CNN models [27]. Recent work suggests that general multimodal models such as Gemini 2.5 Pro may even surpass specialized AI in certain "edge cases" [28]. Unlike task-specific systems, these models can simultaneously classify images and generate descriptive reasoning and management suggestions, offering value for clinical interpretation and education.

Enhancing MM-LLM Performance in Medical-Specific Domain

Several in-domain strategies can improve general-purpose MM-LLMs without training from scratch, including ICL, contrastive pretraining, and retrieval augmentation.

In our study, text-only ICL with brief cystoscopic tumor descriptions had minimal impact, except for OpenAI-o3. Specificity rose slightly, and sensitivity improved for rare benign lesions, but this was offset by reduced sensitivity for malignant classes (pUC and CIS)—a trade-off in which stricter thresholds reduce false positives and unnecessary biopsies, but risk missed cancers. These findings indicate that text-only ICL confers minimal benefit for image-dominant tasks. Notably, OpenAI-o3 showed a modest gain in micro-average accuracy (0.41-0.46), driven mainly by NOTA, likely reflecting its reasoning-oriented architecture and more flexible use of ICL as contextual support.

A likely reason for the limited effect is insufficient visual grounding: text-only cues do not anchor the model's attention to class-defining morphology. In fine-grained visual discrimination (eg, cystoscopic lesion typing), semantic hints (eg, "papillary fronds" and "flat erythematous base") may not map reliably to visual features unless those associations were learned during pretraining; without image exemplars, the model's visual reasoning remains underconstrained.

Beyond text-only prompts, few-shot image-text exemplars (multimodal ICL) can strengthen grounding and improve accuracy. Presenting paired lesion images with diagnoses exposes prototypical visual features and tightens the link between morphology and class semantics. Across histopathology imaging, image-text ICL has enabled ChatGPT-4V to approach or surpass task-specific classifiers with only 5-10 examples per class, markedly narrowing the gap between zero-shot and fully supervised models [29].

In parallel, contrastive-learned encoders (eg, MedCLIP) [30] and multimodal retrieval augmentation [31] can enrich representations and factual grounding, mitigating data scarcity and hallucination. In summary, improving MM-LLM image classification in medical domains is multifaceted, and combining hybrid ICL (image + text), contrastive pretraining, and retrieval augmentation offers a practical path to greater accuracy, robustness, and interpretability in cystoscopic diagnosis.

Clinical Implications

MM-LLMs offer greater flexibility than task-specific endoscopy AI by combining visual recognition with contextual reasoning and narrative explanation. They can interpret morphology-diverse findings and integrate relevant clinical text, supporting a more context-aware understanding. Our results suggest potential applications in education and workflow support, including serving as virtual tutors for trainees and automating report generation to reduce workload and standardize documentation. However, their moderate diagnostic accuracy, particularly for rare or subtle lesions, limits their current use as autonomous diagnostic tools. Future efforts should focus on vision-conditioned ICL, multimodal retrieval-augmented training, and video-based modeling to enhance interpretive stability and diagnostic confidence [30,31]. Integration with patient-level data, cystoscopy-specific benchmark datasets, and human-in-the-loop oversight will be critical to ensure clinical safety and responsible implementation.

Limitations

This study has several methodological strengths, including an unbiased and rigorous evaluation framework. The dual-task design enabled simultaneous assessment of reasoning transparency, adaptability, and accuracy—helping distinguish superficial pattern recognition from genuine clinical understanding. However, several limitations should be acknowledged. First, our ICL implementation was text-based only: models received brief written descriptions of tumor features without any paired visual exemplars. As a result, this study evaluated "text-based ICL" rather than full multimodal ICL, and the absence of a few-shot image or image-text examples likely constrained the models' multimodal capabilities. The modest gains observed with ICL in our experiments may therefore underestimate the potential benefit of visual or hybrid (image + text) ICL. Future work should directly compare text-based, visual, and hybrid ICL strategies and explore complementary approaches such as contrastive-learned encoders and multimodal retrieval augmentation for cystoscopic diagnosis. Second, because the raw test images were drawn from heterogeneous sources, residual confounding from source-related differences and image-quality artifacts cannot be fully excluded. Although we applied a strict 3-layer quality control pipeline—image exclusion criteria, standardized preprocessing, and human verification of diagnostic utility—to mitigate these effects, some bias related to image source heterogeneity may remain. In addition, our evaluation relied on static images; incorporating temporal cues from cystoscopy videos may improve recognition of subtle or evolving lesions and reduce misclassification. Third, our dataset has an enriched abnormality prevalence (80.3%), substantially higher than that of typical clinical populations (20%-30%). Consequently, the reported positive predictive value and negative predictive value are inflated and not generalizable; they should be interpreted as dataset-specific rather than as estimates for real-world screening or hematuria-clinic settings.

Conclusions

Using a clinically challenging, stress-test image set and a rigorous blinded evaluation framework, this study comprehensively assessed MM-LLMs for cystoscopic interpretation and lesion classification. Among the evaluated models, OpenAI-o3 demonstrated the most balanced and clinically coherent performance, followed by ChatGPT-4o and Gemini 2.5 Pro. These findings highlight the meaningful

XSL•FO
RenderX

assistive potential of MM-LLMs in generating interpretable free-text rationales, supporting biopsy triage, and facilitating training. However, their performance in truly difficult differential diagnoses remains modest and requires further optimization before safe clinical integration.

Acknowledgments

We would like to extend our heartfelt gratitude to our colleagues for their invaluable contributions to this work. ChatGPT (OpenAI) was used only for language editing and grammar correction; it was not used for content generation or any data-related tasks.

Data Availability

The datasets generated and/or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

YCS and CYW contributed equally as first authors. SWH and CYT contributed equally as corresponding authors. Conceptualization was performed by CYT and SWH. Data curation was carried out by YCS and CYW. Formal analysis was conducted by CYW, YCS, SWH, and CYT. Visualization was undertaken by CYT and CYW. The original draft was written by CYW and YCS, and SWH and CYT completed manuscript review and editing. Supervision was provided by CYT and SWH.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional methods, tables, and figures.
[PDF File (Adobe PDF File), 809 KB - jmir_v28i1e87193_app1.pdf]

References

1. David SA, Patil D, Alemozaffar M, Issa MM, Master VA, Filson CP. Urologist use of cystoscopy for patients presenting with hematuria in the United States. Urology 2017;100:20-26. [doi: 10.1016/j.urology.2016.09.018] [Medline: 27645524]
2. Guldhammer CS, Vásquez JL, Kristensen VM, Norus T, Nadler N, Jensen JB, et al. Cystoscopy accuracy in detecting bladder tumors: a prospective video-confirmed study. Cancers (Basel) 2023;16(1):160 [FREE Full text] [doi: 10.3390/cancers16010160] [Medline: 38201586]
3. GLOBOCAN 2022: bladder cancer 9th most common worldwide. WBCP Coalition. URL: https://tinyurl.com/y2p57rbx [accessed 2025-11-02]
4. Devlies W, de Jong JJ, Hofmann F, Bruins HM, Zuiverloon TC, Smith EJ, et al. The diagnostic accuracy of cystoscopy for detecting bladder cancer in adults presenting with haematuria: a systematic review from the European Association of Urology Guidelines Office. Eur Urol Focus 2024;10(1):115-122 [FREE Full text] [doi: 10.1016/j.euf.2023.08.002] [Medline: 37633791]
5. Ye Z, Li Y, Sun Y, He C, He G, Ji Z. Leveraging deep learning in real-time intelligent bladder tumor detection during cystoscopy: a diagnostic study. Ann Surg Oncol 2025;32(5):3220-3226. [doi: 10.1245/s10434-025-17015-3] [Medline: 40050483]
6. Guo Y, Li C, Zhang S, Zhu G, Sun L, Jin T, et al. U-net-based assistive identification of bladder cancer: a promising approach for improved diagnosis. Urol Int 2024;108(2):100-107. [doi: 10.1159/000535652] [Medline: 38081150]
7. Hwang WK, Jo SB, Han DE, Ahn ST, Oh MM, Park HS, et al. Artificial intelligence-based classification and segmentation of bladder cancer in cystoscope images. Cancers (Basel) 2024;17(1):57 [FREE Full text] [doi: 10.3390/cancers17010057] [Medline: 39796686]
8. Jia X, Shkolyar E, Laurie MA, Eminaga O, Liao JC, Xing L. Tumor detection under cystoscopy with transformer-augmented deep learning algorithm. Phys Med Biol 2023;68(16). [doi: 10.1088/1361-6560/ace499] [Medline: 37548023]
9. Huang HH, Cheng P, Tsai CY. Exploring artificial intelligence in functional urology: a comprehensive review. Urol Sci 2025;36:2-10. [doi: 10.1097/us9.0000000000000057]
10. Bretthauer M, Ahmed J, Antonelli G. Use of computer assisted detection (CADe) colonoscopy in colorectal cancer screening and surveillance: European Society of Gastrointestinal Endoscopy (ESGE) position statement. Endoscopy 2025;57:667-673. [doi: 10.1055/s-0045-1805161]
11. Carlini L, Massimi D, Mori Y, Antonelli G, Rizkala T, Spadaccini M, et al. Large language models for detecting colorectal polyps in endoscopic images. Gut 2025 May 24. [doi: 10.1136/gutjnl-2025-335091] [Medline: 40360230]
12. Repici A, Badalamenti M, Maselli R, Correale L, Radaelli F, Rondonotti E, et al. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology 2020;159(2):512-520.e7. [doi: 10.1053/j.gastro.2020.04.062] [Medline: 32371116]

XSL•FO
RenderX

13. Shaukat A, Lichtenstein DR, Somers SC, Chung DC, Perdue DG, Gopal M, SKOUT™ Registration Study Team. Computer-aided detection improves adenomas per colonoscopy for screening and surveillance colonoscopy: a randomized trial. Gastroenterology 2022;163(3):732-741 [FREE Full text] [doi: 10.1053/j.gastro.2022.05.028] [Medline: 35643173]
14. Wang P, Liu X, Berzin TM, Glissen Brown JR, Liu P, Zhou C, et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. Lancet Gastroenterol Hepatol 2020;5(4):343-351. [doi: 10.1016/S2468-1253(19)30411-X] [Medline: 31981517]
15. AlSaad R, Abd-Alrazaq A, Boughorbel S, Ahmed A, Renault M, Damseh R, et al. Multimodal large language models in health care: applications, challenges, and future outlook. J Med Internet Res 2024;26:e59505 [FREE Full text] [doi: 10.2196/59505] [Medline: 39321458]
16. Nam Y, Kim DY, Kyung S, Seo J, Song JM, Kwon J, et al. Multimodal large language models in medical imaging: current state and future directions. Korean J Radiol 2025;26(10):900-923 [FREE Full text] [doi: 10.3348/kjr.2025.0599] [Medline: 41015856]
17. Guo L, Zuo Y, Yisha Z, Liu J, Gu A, Yushan R, et al. Diagnostic performance of advanced large language models in cystoscopy: evidence from a retrospective study and clinical cases. BMC Urol 2025;25(1):64 [FREE Full text] [doi: 10.1186/s12894-025-01740-8] [Medline: 40158093]
18. Ferber D, Wölflein G, Wiest IC, Ligero M, Sainath S, Ghaffari Laleh N, et al. In-context learning enables multimodal large language models to classify cancer pathology images. Nat Commun 2024;15(1):10104 [FREE Full text] [doi: 10.1038/s41467-024-51465-9] [Medline: 39572531]
19. Tenny BC, O'Neill M. Diagnostic Cystoscopy. Cham: Springer; 2022.
20. o3. OpenAI Platform. 2025. URL: https://platform.openai.com/docs/models/o3 [accessed 2025-06-15]
21. Gemini 2.5 Pro. Google Cloud. URL: https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro?utm_source=chatgpt.com&hl=zh-tw [accessed 2025-06-15]
22. MedGemma 1.5 model card. Health AI Developer Foundations. 2025. URL: https://developers.google.com/health-ai-developer-foundations/medgemma/model-card [accessed 2025-07-20]
23. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 2020;21(1):6 [FREE Full text] [doi: 10.1186/s12864-019-6413-7] [Medline: 31898477]
24. Tam ZR, Wu CK, Lin CY, Chen YN. None of the above, less of the right: parallel patterns between humans and llms on multi-choice questions answering. arXiv Preprint posted online on Mar 3, 2025 [FREE Full text]
25. Ren J, Zhao Y, Vu T, Liu PJ, Lakshminarayanan B. Self-evaluation improves selective generation in large language models. arXiv Preprint posted online on Dec 14, 2023 [FREE Full text]
26. Zhao Z, Wallace E, Feng S, Klein D, Singh S. Calibrate before use: improving few-shot performance of language models. arXiv Preprint posted online on Feb 19, 2021. [doi: 10.48550/arXiv.2102.09690]
27. Khalafi MA, Safavi-Naini SAA, Salehi A, Naderi N, Alijanzadeh D, Moghadam PK, et al. Vision language models versus machine learning models performance on polyp detection and classification in colonoscopy images. Sci Rep 2025;15(1):45484 [FREE Full text] [doi: 10.1038/s41598-025-29566-2] [Medline: 41309981]
28. Zhang Y, Chen Q, Zhou T. Can general-purpose omnimodels compete with specialists? A case study in medical image segmentation. arXiv Preprint posted online on Aug 31, 2025 [FREE Full text]
29. Ono D, Dickson DW, Koga S. Evaluating the efficacy of few-shot learning for GPT-4Vision in neurodegenerative disease histopathology: a comparative analysis with convolutional neural network model. Research Square Preprint posted online on May 28, 2024. [doi: 10.21203/rs.3.rs-4462333/v1]
30. Wang Z, Wu Z, Agarwal D, Sun J. MedCLIP: contrastive learning from unpaired medical images and text. 2022 Presented at: Proc Conf Empir Methods Nat Lang Process. 2022 Dec; November 4-9, 2025; Suzhou, China. [doi: 10.18653/v1/2022.emnlp-main.256]
31. Xia P, Zhu K, Li H, Wang T, Shi S, Wang S, et al. Mmed-rag: versatile multimodal rag system for medical vision language models. arXiv Preprint posted online on Oct 16, 2024 [FREE Full text]

Abbreviations

AI: artificial intelligence
CIS: carcinoma in situ
ICL: in-context learning
MCC: Matthews correlation coefficient
MM-LLM: multimodal large language model
Non-U Ca: non-urothelial carcinoma
NOTA: none of the above
pUC: papillary urothelial carcinoma
SOTA: state-of-the-art
Youden J: Youden J index

XSL•FO
RenderX

Edited by A Coristine; submitted 05.Nov.2025; peer-reviewed by HC Yang, CY Chang; comments to author 04.Dec.2025; revised version received 07.Jan.2026; accepted 07.Jan.2026; published 28.Jan.2026.

Please cite as:
Shih YC, Wu CY, Huang SW, Tsai CY
Multimodal Large Language Models for Cystoscopic Image Interpretation and Bladder Lesion Classification: Comparative Study
J Med Internet Res 2026;28:e87193
URL: https://www.jmir.org/2026/1/e87193
doi:10.2196/87193
PMID:

©Yung-Chi Shih, Cheng-Yang Wu, Shi-Wei Huang, Chung-You Tsai. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Remote Patient Monitoring in Louisiana Medicare Beneficiaries With Diabetes or Hypertension: Retrospective Cohort Study of 2016-2020 Claims Data

Debra Winberg[1,2], MA, MSc, PhD; Charles Stoecker[2], MA, PhD; Jian Li[3], PhD; Lizheng Shi[2], MA, MsPharm, PhD

[1]Department of Health Management and Policy, School of Health, Georgetown University, St Mary's Hall, 3700 Reservoir Road, Washington, DC, United States

[2]Department of Health Policy and Management, Celia Scott Weatherhead School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA, United States

[3]Department of Biostatistics, Celia Scott Weatherhead School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA, United States

Corresponding Author:
Debra Winberg, MA, MSc, PhD
Department of Health Management and Policy, School of Health, Georgetown University, St Mary's Hall, 3700 Reservoir Road, Washington, DC, United States

Abstract

Background: Chronic conditions such as diabetes and hypertension pose major public health and economic challenges. Innovative tools, including remote patient monitoring (RPM), aim to address barriers to care by digitally transmitting health data to providers.

Objective: This study assesses the impact of RPM on health care utilization in Medicare beneficiaries with diabetes or hypertension.

Methods: This retrospective cohort study used a longitudinal panel from 2016 to 2020 based on Louisiana Medicare Fee-for-Service claims and Beneficiary Files. Patients were included if they were diagnosed with diabetes or hypertension before 2016. RPM exposure was defined by common procedure terminology/healthcare common procedure coding system codes (eg, 99453, 99454, 99457, 99458, 95250, 95251, 99091, E2100-E2103, E0607), with the corresponding month considered the treatment month. Control beneficiaries were randomly assigned treatment months matching the treatment group. Outcomes were monthly counts per 1000 beneficiaries of outpatient, emergency department, and inpatient visits (all-cause and disease-specific), constructed via revenue center/place-of-service/primary ICD-10 (International Classification of Diseases-10) codes. We estimated inverse-probability-treatment-weighted 2-way fixed-effects models with standard errors clustered at the physician level. Data represent secondary analysis of deidentified claims; no participant recruitment occurred. All confidence intervals are reported at the 95% confidence level.

Results: The study included 5488 beneficiaries in the treatment group and 341,226 in the control group. RPM was associated with increased level of outpatient visits in the treatment group by 2284.01 (95% CI 2059.75-2508.27) visits per month per 1000 people (P<.001) after initiation and decreased the trend of outpatient visits by 37.61 (95% CI −37.98 to −37.24) visits per month per 1000 people compared with the control group (P<.001). Similarly, the level of emergency room visits increased in the treatment compared with the control by 84.60 (95% CI 72.21 to 96.99) visits per month per 1000 people initially after RPM initiation (P<.001) and decreased the trend by 1.40 (95% CI −1.60 to −1.20) visits per month per 1000 people (P<.001). The level of inpatient visits increased by 32.64 (95% CI 24.84-40.44) visits per month per 1000 beneficiaries in the treatment group compared with the control immediately after RPM prescription (P=.002), but the trend decreased by 0.54 (95% CI −0.66 to −0.42) visits per 1000 people each month (P<.001).

Conclusions: This study is among the first large, claims-based evaluations of RPM in older adults with diabetes or hypertension using a beneficiary- and time-fixed-effects design. We identify a pattern of increased health care utilization following RPM initiation, followed by a gradual decline that does not fully offset the initial increase. By characterizing utilization trajectories and not clinical endpoints, this study helps reconcile inconsistent findings in prior RPM research and highlights the importance of implementation factors in Medicare populations. Strategies to maximize the value of RPM may include appropriate population targeting, patient education on digital health technologies, and clinician workflow support to support effective integration into routine care.

(J Med Internet Res 2026;28:e80782) doi:10.2196/80782

XSL•FO
RenderX

KEYWORDS

remote patient monitoring; diabetes; hypertension; health utilization; medicare

Introduction

Chronic conditions, including diabetes and hypertension, are significant public health challenges. In the past 30 years, the prevalence of diabetes has increased by 11%, and in 2024, more than 30 million Americans were diagnosed with diabetes [1]. Similarly, hypertension affects approximately 47% of Americans, or over 120 million people [2,3]. Both diseases are major contributors to other chronic health concerns, including cardiovascular disease, chronic kidney disease, stroke, and premature mortality, which can also increase rates of health care resource utilization (HCRU). More so, the economic impact is significant. Both diabetes and hypertension are leading causes of morbidity and mortality in the United States, with combined mortality rates nearly doubling from 2000 to 2023 [4]. In 2022, hypertension-related costs exceeded US $219 billion annually, and diabetes direct costs were over US $300 billion [5-9].

This burden disproportionately impacts older adults over 65, who face higher rates of both conditions in addition to higher rates of complications, multimorbidity, and polypharmacy [7,10]. In turn, Medicare bears much of the burden from these conditions [7].

There are many successful pharmacological treatments and evidence-based lifestyle interventions for diabetes and hypertension, yet over half of patients do not have their blood sugar or blood pressure under control [11,12]. Alarmingly, control rates have worsened in recent years, with hypertension control decreasing from 54% in 2013 to 48% in 2017 - 2020 and continuing to decrease during the COVID-19 pandemic [13,14]. Similarly, glycemic control has stagnated in the past 30 years, and disparities in control have increased [15]. These trends highlight persistent barriers in chronic disease management, including poor medication adherence, limited access to specialty care, high costs, social and economic challenges, and low patient engagement in care [16]. Innovative digital tools, including remote patient monitoring (RPM), aim to address these barriers and improve disease control. RPM is an mHealth device that aims to improve patient care by digitally transmitting health data from patients to physicians [17]

By reducing patient travel costs, allowing physicians to treat more patients, facilitating practice management, and providing health care providers access to patient data quickly, these devices target several barriers to glycemic control, including nonadherence, self-management, and health care experiences [18-20]. RPM adoption in the United States has accelerated in recent years, driven by expanded reimbursement policies, technological advances in wearables and connected devices, and the growing need to manage chronic disease more effectively [21]. These factors have positioned RPM as one of the most commonly used digital health tools, particularly for conditions such as diabetes and hypertension.

Evidence regarding RPM's effectiveness is positive, but inconsistent. A 2024 systematic review on RPM interventions found measurable impacts in chronic disease outcomes, though effects varied based on a number of variables, including the specific disease and implementation design [22]. A review of systematic reviews found that there are possible benefits from RPM, but results are inconsistent [23]. For example, several trials and meta-analyses report A1c reductions by 0.3% - 1.22% and decreases in systolic blood pressure of around 4 - 9 mmHg among RPM cohorts [17,24-26]. While these are modest changes, they can be clinically meaningful and suggest RPM can support better chronic disease control.

However, not all evidence is positive; some studies demonstrate that the RPM has no impact and may only be effective for certain patient groups [27-33]. Specifically, RPM may only benefit specific subgroups, such as people with a high baseline risk, who monitor very often, or who are very engaged in their care and have physicians who are very engaged in their care [34]. Additionally, there is mixed evidence on the ability of RPM to decrease HCRU [35,36]. Some studies found that RPM can prevent costly acute events and reduce hospitalizations and emergency room (ER) visits and may be cost-effective for hypertensive patients [37]. However, other studies demonstrated no change or increased utilization patterns, potentially attributable to the detection of previously unaddressed clinical needs or enhanced surveillance following RPM deployment [38]. Overall, the effectiveness of RPM and the balance between increased monitoring and reduced acute care use remain unclear.

Despite the rapid expansion of RPM coverage and use, most of the existing literature focuses on younger and commercially insured populations who may not have such a high burden of diabetes and hypertension and have fewer barriers to using RPM**,** which may impact its effectiveness [39,40]. While the evidence base in Medicare beneficiaries exists, it remains limited and often focuses on other conditions such as heart disease and chronic obstructive pulmonary disease (COPD) [35,41]. A recent systematic review found that most of the studies on RPM focus on usability and feasibility, not outcome measures such as utilization, for cardiovascular disease [42]. Therefore, not enough real-world evidence studies have been performed for diabetes and hypertension. As most general RPM services in Medicare beneficiaries are for diabetes and hypertension, understanding their impact on this population is essential. Additionally, as some of the Medicare literature shows a more negative impact, such as increased health utilization, it is essential to analyze more specific Medicare populations [38].

This study aims to assess the impact of RPM on HCRU in Medicare beneficiaries with diabetes or hypertension. By focusing on an older, higher-risk population, our analysis contributes novel real-world evidence on the effectiveness of RPM and provides timely insights for policymakers and providers considering broader adoption.

XSL•FO
RenderX

Methods

Overview

This is a retrospective cohort study using Louisiana Medicare Fee-for-Service claims linked to beneficiary files and analyzed using an inverse probability treatment weighted (IPTW) controlled interrupted time series (CITSA) design. This manuscript was prepared following the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines [43].

Data Source and Patient Selection

We used data from a patient-level longitudinal panel from January 1, 2016, to December 31, 2020, which was constructed from Louisiana Medicare Fee-for-Service claims and Beneficiary files. This data is collected at the time of service and was purchased from the Research Assistance Data Center for secondary analyses. We included patients who had a diagnosis of diabetes or hypertension based on the CMS Chronic Conditions files. Patients were excluded if they did not have a diabetes or hypertension diagnosis before 2016, were not enrolled in Medicare before 2016, had only gestational diabetes or secondary diabetes, or had end-stage renal disease. These exclusions were found using the ICD codes.

Patients were included in the treatment group if they had at least one occurrence of a common procedure terminology (CPT) code for RPM or continuous glucose monitoring (99453, 99454, 99457, 99458, 95250, 95251, 99091, E2100, E0607, E2102, E2103, E2101). This information was found in the inpatient, ER, or outpatient claims files. We defined the treatment month as the first month with this CPT code. For the control group, pseudo-treatment dates were randomly assigned to reflect the distribution of treatment initiation in the intervention group. Specifically, in Excel, each month and year combination was coded numerically, and controls were assigned a treatment month and year in proportion to the observed distribution of initiation dates in the treatment group. This ensured alignment by time of year rather than only by year, reducing bias from differential calendar-time exposure. After the assignment, we conducted balance checks and confirmed that the month-by-month distribution of treatment dates did not significantly differ between the control and treatment groups.

Outcome Variables

The number of outpatient, ER, or inpatient visits per month. We defined these outcomes using published guidance from the Research Data Assistance Center using revenue center codes, claim dates, and places of service codes in the Medicare Fee-for-Service files (Carrier claims, Carrier Line, Base Claims, and Revenue Center). The disease-specific outcomes included the number of outpatient, ER, or inpatient visits per month where the primary ICD code indicated that the visit was for diabetes or hypertension. If a beneficiary did not have an encounter in a certain month, they were coded as having 0 visits. Visits with RPM as the primary CPT code were not included. All outcomes were detrended to remove underlying time trends to focus on deviations that may be associated with RPM.

Statistical Analysis

We used a CITSA to determine the impact of RPM. This approach leverages pre- and posttreatment data from a treatment and control group to estimate changes in the level and slope of the outcome variables before and after treatment initiation. This method assesses the impact of an intervention such as RPM by comparing a treatment to a control group over time. CITSA was chosen for the analysis because of the linearity of the data over time (Multimedia Appendix 1) and for its ability to account for pre-existing trends and provide a counterfactual estimate. The CITSA model may be written as:

$$Y=\beta_0+\beta_1(\text{Time})+\beta_2(\text{RPM})+\beta_3(\text{Time}*\text{RPM})+\beta_4(\text{Post}*\text{RPM})+\beta_5(\text{Time}*\text{RPM}*\text{Post})+\beta X+$$

The baseline level is given by the intercept β_0. β_1 represents the underlying time trend for the non-RPM group before the intervention. The indicator variable RPM specifies group membership, with β_2 capturing any baseline differences between the intervention and control groups. The term β3 allows for differences in preintervention trends between RPM and non-RPM groups. The coefficient β_4 represents the immediate postintervention effect. The 3-way interaction term β_5 captures any differential trend change attributable to the intervention over time. There was no specific postperiod variable included in the model, as the control group had a randomly assigned treatment date and, as such, had no expected postperiod outcome change.

To account for bias and confounding, we adjusted for covariates that may influence the outcome, including comorbidities (diabetes, hypertension, COPD, chronic kidney disease, hyperlipidemia), Medicare status, sex, race (White, Black, Hispanic, Asian, Other), rurality (defined as metropolitan, micropolitan, rural using FIPS codes and 2020 core-based statistical area data), and age. These variables were created using the Medicare Beneficiary Files (chronic condition and ABCD files), which aggregate data at the annual level.

All analyses accounted for time- and beneficiary-level fixed effects and were weighted by the IPTW, and standard errors were clustered at the primary doctor level to account for serial correlations over time. To construct the weights, we modeled the probability of RPM participation using baseline demographic and clinical characteristics from the month before treatment and the average preperiod outcomes. Each beneficiary was assigned a weight equal to the inverse of their predicted probability. This data was obtained from the beneficiary files and the previously created outcome variables. The primary doctor was defined as the physician whom a beneficiary saw the most during the study period to account for correlated practice patterns. In the base case, the main model used all available data in the 60 months regardless of pre- and postlength of time.

We used an inverse-probability–weighted 2-way fixed-effects model to control for unobserved, time-invariant differences across beneficiaries and shared temporal shocks. This design isolates within-beneficiary changes before and after RPM initiation, consistent with the interrupted time series framework. It was preferred to multilevel models because exposure timing varies across beneficiaries, and fixed effects more clearly estimate pre/post level and trend differences [44].

XSL•FO
RenderX

Because Medicare claims are used for reimbursement, demographic, enrollment, and utilization fields are complete for almost all beneficiaries. Diagnosis and procedure variables used to construct comorbidities are also not subject to item-level analysis. Multiple imputation was not performed, as missingness was minimal and largely structural and less than 15% for all key variables.

All confidence intervals were calculated at the 95% confidence level. We tested the robustness of the results by clustering standard errors by zip code and changing the necessary length of pre- and posttreatment data to 6 months pre and 6 months post, and 12 months pre and 12 months post. Additionally, we assess seasonality of treatment by running the analysis with a variable for calendar month (January, February, etc) and another analysis with a control variable for calendar month of treatment. Additionally, we ran an analysis excluding COVID years to ensure we were not picking up COVID confounding and to account for potential exposure classification issues due to new CPT codes being introduced.

All analyses were conducted using Stata (version 18.0; IBM Corp).

Ethical Considerations

This study was approved by Tulane University's Institutional Review Board (2023-1044-SPHTM) through an expedited review, as it only uses secondary data. The data use agreement and institutional review board for this study allow the use of secondary data analysis without additional consent from beneficiaries. All study data is deidentified to protect beneficiaries, and there was no way to reidentify. As we did not work directly with beneficiaries, no compensation was given, and there is no identification of individuals in this manuscript.

Results

Initially, the data included 480,592 Louisiana beneficiaries with diabetes or hypertension. This sample included 472,902 beneficiaries in the control group. Of these, 118,513 were excluded for being diagnosed with hypertension or diabetes after 2016. A further 9 beneficiaries were dropped for being enrolled in Medicare too late, and 130 were excluded for only having secondary diabetes. Last, 20,226 beneficiaries were dropped for having end-stage renal disease.

Initially, 7690 Medicare beneficiaries were identified in the treatment group. Of these, 1849 beneficiaries were excluded because they were diagnosed with hypertension or diabetes after 2016. No beneficiaries were excluded due to delayed Medicare enrollment, 2 beneficiaries were excluded for having secondary diabetes only, and 351 beneficiaries were excluded due to end-stage renal disease. After applying all inclusion and exclusion criteria, a total of 346,714 Medicare beneficiaries with diabetes or hypertension were included in the analytic sample. The final sample comprised 5488 beneficiaries in the treatment group and 341,226 in the control group. Among the treated beneficiaries, 719 (13.1%) initiated RPM in 2016, 635 (11.6%) in 2017, 770 (14.0%) in 2018, 1346 (24.5%) in 2019, and 2018 (36.8%) in 2020. These distributions remained balanced after weighting.

Table 1 presents the weighted and unweighted baseline characteristics of the treatment and control groups. After weighting, all standardized mean differences were below 10%. Prior to weighting, the treatment group had a higher prevalence of several chronic conditions, including diabetes (3842, 70% vs 112,605, 33%), hypertension (5049, 92% vs 245,683, 72%), COPD (494, 9% vs 23,886, 7%), chronic heart failure (823, 15% vs 30,710, 9%), chronic kidney disease (2031, 37% vs 47,772, 14%), and hyperlipidemia (2524, 46% vs 88,719, 26%).

The RPM group was younger on average (mean age 71.6 vs 74.2 years), included a higher proportion of male beneficiaries (2470, 45% vs 126,254, 37%), and a higher proportion of White beneficiaries (3842, 70% vs 208,148, 61%). In contrast, beneficiaries in the treatment group were less likely to have dual Medicaid eligibility (1317, 24% vs 139,903, 41%). The treatment group also demonstrated higher baseline health care utilization, including emergency department visits, inpatient diabetes-related visits, outpatient visits, and outpatient diabetes-related visits. After weighting, baseline demographic characteristics, clinical conditions, and health care utilization were comparable between the treatment and control groups.

XSL•FO
RenderX

Table . Baseline population characteristics and health care utilization. Baseline demographic, clinical, and health care utilization characteristics of Medicare Fee-for-Service beneficiaries in Louisiana with hypertension or diabetes, 2016-2020. Patients in the treatment group initiated remote patient monitoring or continuous glucose monitoring based on relevant procedure codes. Data are unweighted and weighted using inverse probability treatment weight, which were created using demographic and utilization data. Values are shown as means (SD) or percentages, with standardized mean differences (SMD) presented for both unweighted and weighted comparisons. Utilization outcomes are average monthly counts in the preintervention period.

Variable	Nonweighted			Weighted		
	Treatment mean (SD) (n=5488)	Control mean (SD) (n=341,226)	SMD	Treatment mean (SD) (n=5488)	Control mean (SD) (n=341,226)	SMD
Diabetes	0.70 (0.46)	0.46 (0.47)	−0.81	0.70 (0.46)	0.71 (0.45)	0.02
Hypertension	0.92 (0.27)	0.72 (0.45)	−0.54	0.92 (0.27)	0.93 (0.26)	0.02
Chronic obstructive pulmonary disease	0.09 (0.29)	0.07 (0.25)	−0.1	0.09 (0.29)	0.10 (0.30)	0.01
Heart failure	0.15 (0.36)	0.09 (0.29)	−0.19	0.15 (0.36)	0.16 (0.36)	0.01
Chronic kidney disease	0.37(0.48)	0.14 (0.35)	−0.54	0.37 (0.48)	0.36 (0.48)	−0.01
Hyperlipidemia	0.46 (0.50)	0.26 (0.44)	−0.44	0.46 (0.50)	0.46 (0.50)	−0.01
Myocardial infarction	0.01 (0.11)	0.00 (0.07)	−0.08	0.01 (0.11)	0.01 (0.11)	0
Charlson Comorbidity Index	3.88 (2.52)	2.46 (2.50)	−0.56	3.88 (2.52)	4.10 (3.15)	0.08
Age	71.58 (10.36)	74.15 (11.27)	0.24	71.58 (10.36)	71.72 (10.56)	0.01
Dual status	0.24 (0.43)	0.41 (0.49)	0.36	0.24 (0.43)	0.27 (0.44)	0.08
Percent male	0.45 (0.50)	0.37 (0.48)	−0.17	0.45 (0.50)	0.44 (0.50)	−0.02
Percent White	0.70 (0.46)	0.61 (0.49)	−0.19	0.70 (0.46)	0.70 (0.46)	0.00
Percent Black	0.26 (0.44)	0.23 (0.42)	−0.07	0.26 (0.44)	0.26 (0.44)	0.00
Percent Hispanic	0.02 (0.14)	0.02 (0.13)	−0.03	0.02 (0.14)	0.02 (0.14)	−0.01
Percent Asian	0.01 (0.09)	0.01 (0.08)	−0.02	0.01 (0.09)	0.01 (0.09)	−0.01
Percent metropolitan	0.83 (0.38)	0.75 (0.43)	−0.19	0.83 (0.38)	0.84 (0.37)	−0.01
Percent micropolitan	0.09 (0.29)	0.14 (0.35)	0.15	0.09 (0.29)	0.09 (0.28)	−0.01
Percent rural	0.08 (0.27)	0.11 (0.31)	0.11	0.08 (0.27)	0.08 (0.27)	−0.01
Inpatient visits	0.03 (0.06)	0.03 (0.06)	−0.04	0.03 (0.06)	0.03 (0.07)	0.09
Inpatient visits (hypertension)	0.00 (0.01)	0.00 (0.01)	−0.03	0.00 (0.01)	0.00 (0.01)	0.02
Inpatient visits (diabetes)	0.00 (0.02)	0.00 (0.01)	−0.13	0.00 (0.02)	0.00 (0.01)	−0.05
Emergency room visits	0.01 (0.03)	0.00 (0.01)	−0.22	0.01 (0.03)	0.00 (0.03)	−0.08
Emergency room visits (diabetes)	0.00 (0.02)	0.00 (0.01)	−0.06	0.00 (0.02)	0.00 (0.02)	−0.01
Emergency room visits (hypertension)	0.07 (0.10)	0.06 (0.09)	−0.12	0.07 (0.10)	0.08 (0.11)	0.06
Outpatient visits	2.34 (1.60)	1.53 (1.52)	−0.52	2.34 (1.60)	2.27 (1.92)	−0.04
Outpatient visits (diabetes)	0.34 (0.41)	0.08 (0.20)	−0.83	0.34 (0.41)	0.29 (0.60)	−0.10
Outpatient visits (hypertension)	0.14 (0.17)	0.12 (0.21)	−0.08	0.14 (0.17)	0.14 (0.21)	0.01

Figures 1-3 display residualized monthly outpatient, ER, and inpatient visits by treatment status. These figures plot model residuals after covariate adjustment and illustrate underlying utilization patterns. For outpatient visits (Figure 1), the treatment

XSL•FO
RenderX

group increased from about 0.4 residual visits per month preintervention to a peak near 1.2 around the index date, before declining to approximately 0.5 by the end of follow-up. The control group residuals hovered close to 0 throughout. For ER visits (Figure 2), the treatment group rose to roughly 0.035 residual visits per month at the peak, declining back toward zero postintervention. In contrast, the control group stayed flat at approximately 0 residual visits. For inpatient visits (Figure 3), the treatment group showed a smaller preintervention increase, reaching about 0.018 residual visits per month, followed by a decline toward baseline. Again, the control group remained stable near 0 across the entire period. Residual plots for disease-specific outcomes can be found in Multimedia Appendix 1.

Figure 1. Residualized monthly outpatient visits before and after remote patient monitoring initiation. Models were adjusted for age, sex, race, rurality, comorbidities (diabetes, hypertension, chronic obstructive pulmonary disease, chronic kidney disease, hyperlipidemia), and Medicare status.

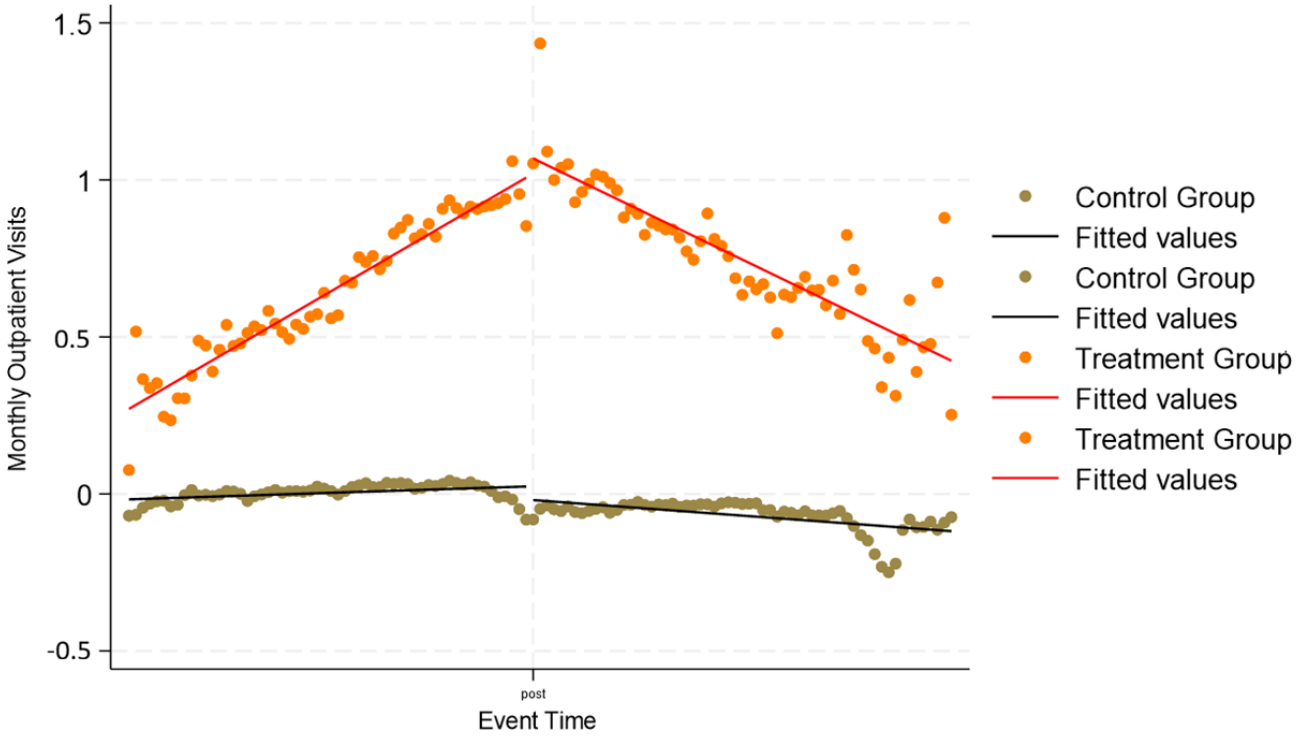

Figure 2. Residualized monthly emergency room visits before and after remote patient monitoring initiation. Models were adjusted for age, sex, race, rurality, comorbidities (diabetes, hypertension, chronic obstructive pulmonary disease, chronic kidney disease, hyperlipidemia), and Medicare status.

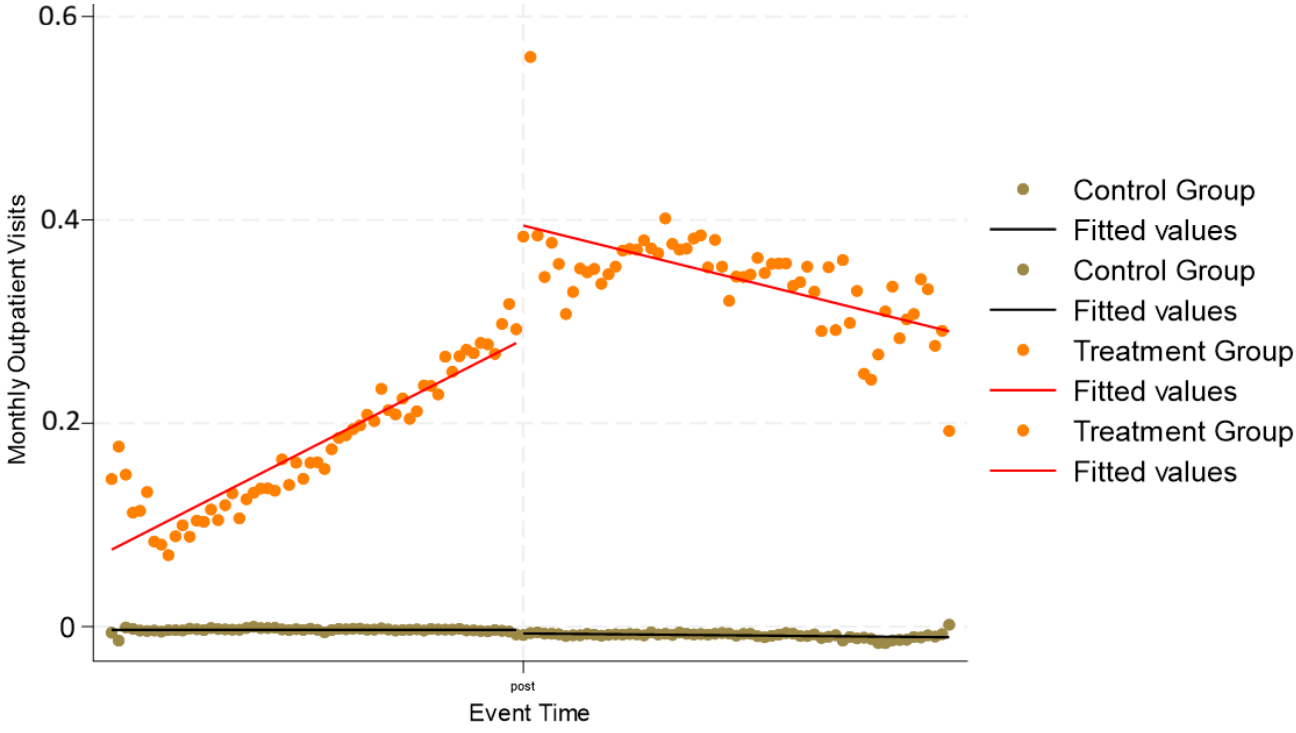

XSL•FO
RenderX

Figure 3. Residualized monthly inpatient visits before and after remote patient monitoring initiation. Models were adjusted for age, sex, race, rurality, comorbidities (diabetes, hypertension, chronic obstructive pulmonary disease, chronic kidney disease, hyperlipidemia), and Medicare status.

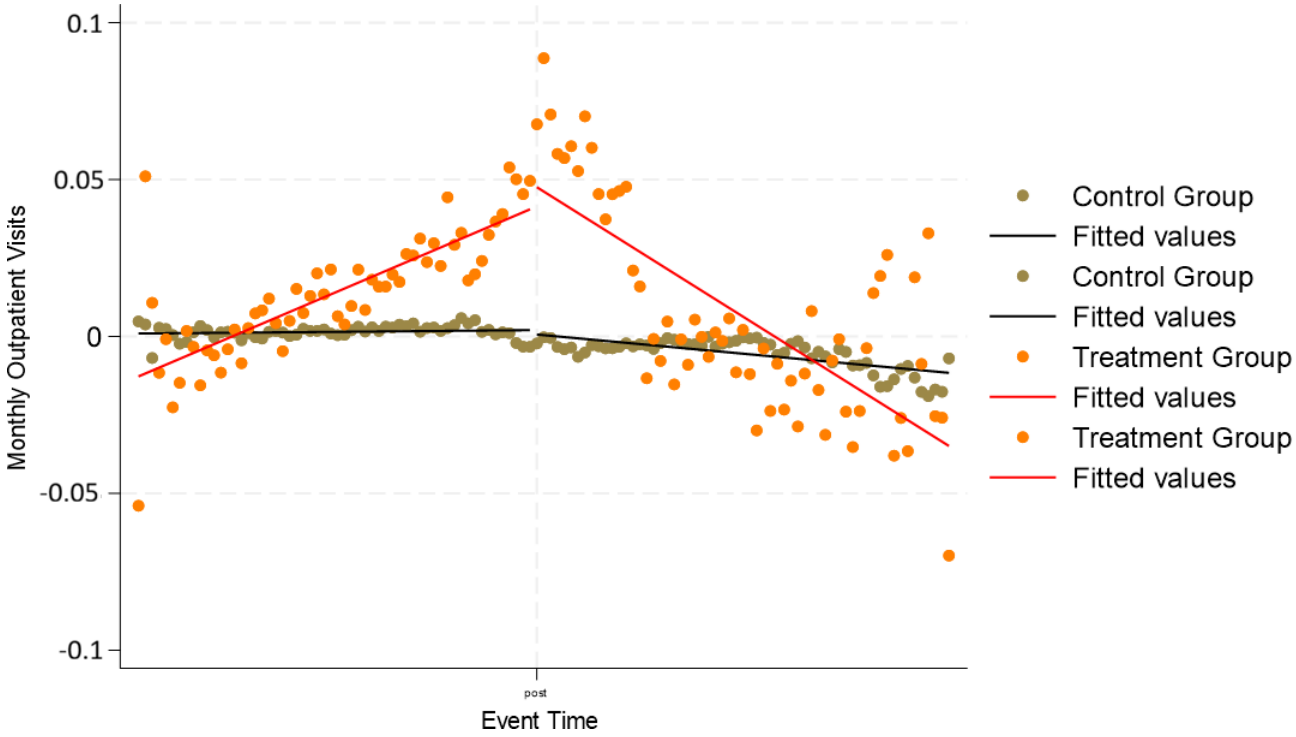

Table 2 depicts the CITSA results. All-cause outpatient visits steadily increased by 9.81 (95% CI 2.42-17.20; *P*=.01) visits per month per 1000 people during the time period. In the preperiod, beneficiaries in the treatment group had their visits decrease by 10.83 (95% CI −18.45 to 3.21; *P*=.01) visits per month per 1000 people due to treatment. After RPM initiation, the level of outpatient visits increased by 2284.01 (95% CI 2059.75-2508.27) visits per month per 1000 people (*P*<.001), and the trend in the number of outpatient visits decreased by 37.61 (95% CI −37.98to −37.24) visits per month per 1000 people compared with those in the control group (*P*<.001). Diabetes-specific visits did not change significantly over the time period or in the pretreatment time period. The level of diabetes-specific appointments increased by 673.26 (95% CI 586.98-759.54) visits per month per 1000 people (*P*<.001) after initiation and then decreased at a higher rate of 10.25 (95% CI −11.39 to −9.11) visits per month per 1000 people (*P*<.001). Over the entire study period, hypertension outpatient visits increased by 0.47 (95% CI 0.31 to 0.63) visits per month per 1000 people (*P*<.001), but did not significantly change in the pretreatment period. The level of hypertension visits increased by 190.92 (95% CI 136.06-245.78) visits in the RPM group compared with those in the control group (*P*<.001), while the trend decreased by 2.91 (95% CI −3.58 to −2.2) visits per 1000 people over the next several years.

All-cause ER visits increased by 0.27 (95% CI 0.17-0.37) visits per month per 1000 people over the entire study period (*P*<.001). In the pretreatment period, visits decreased by 0.3 (95% CI −0.44 to −0.16) visits per month per 1000 people (*P*<.001). Compared with the control, the level of number of ER visits increased by 84.60 (95% CI 72.21-96.99) visits per 1000 people initially after RPM initiation (*P*<.001) and the trend decreased by 1.40 (95% CI −1.60 to −1.20) visits per month per 1000 people (*P*<.001). While the overall time trend and pretreatment trend are insignificant, the patterns also hold for disease-specific ER visits after RPM initiation. At the time of RPM, the level of ER visits increased by 14.58 (95% CI 10.70-18.46) visits per month per 1000 beneficiaries and had a decreasing trend with 0.26 (95% CI −0.32 to −0.20) fewer visits per month per 1000 people (*P*<.001 for all outcomes) in the treatment compared with the control. Like diabetes-specific visits, there were no statistically significant overall time or pretreatment trends. The level of hypertension-specific visits increased by 5.68 (95% CI 2.56-8.80) visits per month for 1000 people after RPM initiation in the treatment group, compared with the control, and the trend decreased by 0.09 (95% CI −0.13 to −0.05) visits per month per 1000 people (*P*<.001) in the RPM group compared with the control.

XSL•FO
RenderX

Table . Impact of the RPM on health care resource utilization. All estimates are expressed as monthly visits per 1000 beneficiaries. Diabetes and hypertension-specific visits are based on the primary diagnosis code associated with the visit. Each cell contains the treatment effect and confidence intervals. Impact of remote patient monitoring (RPM) on monthly outpatient, emergency room, and inpatient visits among Medicare beneficiaries with hypertension or diabetes, Louisiana, 2016-2020. Estimates from inverse probability treatment weight-controlled interrupted time series models are shown with 95% confidence intervals and *P* values.

Variables	All-cause visits		Diabetes visits		Hypertension visits	
	OR[a] (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Outpatient visits						
Time trend	9.81 (2.42 to 17.20)	.01	5.89 (−1.05 to 12.83)	.10	0.47 (0.31 to 0.63)	<.001
Pretreatment trend	−10.83 (−18.45 to 3.21)	.01	−6.23 (−13.19 to 0.73)	.08	−0.28 (−0.79 to 0.23)	.22
Posttreatment level	2284.01 (2059.75 to 2508.27)	<.001	673.26 (586.98 to 759.54)	<.001	190.92 (136.06 to 245.78)	<.001
Posttreatment trend	−37.61 (−37.98 to −37.24)	<.001	−10.25 (−11.39 to −9.11)	<.001	−2.91 (−3.58 to −2.24)	<.001
Emergency room visits						
Time trend	0.27 (0.17 to 0.37)	<.001	0.00 (0.00 to 0.00)	.48	0.00 (−0.02 to 0.02)	.62
Pretreatment trend	−0.30 (−0.44 to −0.16)	<.001	−0.01 (−0.03 to 0.01)	.66	0.00 (−0.02 to 0.02)	.87
Posttreatment level	84.60 (72.21 to 96.99)	<.001	14.58 (10.70 to 18.46)	<.01	5.68 (2.56 to 8.80)	<.001
Posttreatment trend	−1.40 (−1.60 to −1.20)	<.001	−0.26 (−0.32 to −0.20)	<.001	−0.09 (−0.13 to −0.05)	<.001
Inpatient visits						
Time trend	0.21 (0.17 to 0.25)	<.001	0.01 (0.01 to 0.01)	<.01	0.00 (−0.02 to 0.02)	.86
Pretreatment trend	−0.23 (−0.31 to −0.15)	<.001	−0.01 (−0.03 to 0.01)	.23	−0.01 (−0.03 to 0.01)	.48
Posttreatment level	32.64 (24.84 to 40.44)	<.001	5.08 (2.67 to 7.49)	<.001	2.82 (0.43 to 5.21)	.02
Posttreatment trend	−0.54 (−0.66 to −0.42)	<.001	−0.10 (−0.32 to −0.20)	<.001	−0.04 (−0.08 to −0.00)	<.001
Observations	20,799,900	—[b]	20,799,900	—	20,799,900	—

[a]OR: odds ratio.

[b]Not applicable.

All-cause inpatient visits increased by 0.21 (95% CI 0.17-0.25) visits per month per 1000 people over the entire study period (*P*<.001). In the pretreatment period, the trend decreased by −0.23 (95% CI −0.31 to −0.15) visits per month per 1000 people (*P*<.001). Compared with the control group, the level of all-cause inpatient visits increased by 32.64 (95% CI 24.84-40.44) visits per month per 1000 people in the treatment group compared with the control group immediately after RPM prescription (*P*<.001). The trend decreased by 0.54 (95% CI −0.66 to −0.42) visits per 1000 people each month (*P*<.001). During the study period, inpatient diabetes visits increased by 0.01 (95% CI 0.01-0.01) visits per month per 1000 people (*P*<.001) but were not statistically significant in the preperiod. The level of inpatient visits with diabetes as the primary diagnosis code increased by 5.08 (95% CI 2.67-7.49) visits per month per 1000 people right after treatment (*P*<.001), and the trend decreased by 0.01 (95% CI 2.51-2.31) visits per month per 1000 people (*P*<.001) in the RPM group compared with the control. Hypertension-specific inpatient visit levels increased 2.83 (95% CI 0.43-5.21) visits per month per 1000 people immediately after treatment (*P*=.02), and the trend decreased by 0.04 (95% CI −0.08 to 0.00) visits per month per 1000 (*P*<.001).

The results were robust to cluster changes and excluding COVID-19 months (Multimedia Appendix 2). Additionally, when assessing seasonality, both analyses remained statistically significant with the same overall patterns; however, the point estimates slightly decreased (Multimedia Appendix 2), showing there may be some seasonality. The results were not robust to changing the study time (limited to 6 or 12 mo).

XSL•FO
RenderX

Discussion

Principal Findings

This study aimed to assess the impact of RPM on HCRU on Louisiana Medicare beneficiaries with diabetes or hypertension from 2016 to 2020. By combining a rigorous longitudinal design with a focus on an older, high-burden Medicare population, this study provides novel insights into how RPM adoption affects health care utilization in real-world settings. Overall, in beneficiaries with diabetes or hypertension, RPM was associated with a statistically significant increase in outpatient, ER, and inpatient visits immediately after initiation, followed by a significant downward trend over time. These findings, supported by narrow confidence intervals, suggest that the estimates are precise and due to RPM, not random variation.

Existing evidence on RPM's impact on utilization is mixed. Some studies report decreases in HCRU [35,45]. Others have shown that it may not change utilization [17]. Notably, studies in a Medicare population also demonstrated an increase in utilization; for example, one study on Medicare hypertensive patients reported an increase in cardiovascular outpatient visits [38]. Differences across studies may reflect variation in patient populations, RPM modalities, and social determinants of health. For example, our Medicare population may have lower digital or health literacy.

Additionally, the average beneficiary in our cohort was on RPM for 7 months, and results were not robust to a short time period, which is consistent with literature that found that clinical outcomes, such as decreased blood glucose, which may impact utilization, are only found in longer time periods [46]. Thus, long-term users may be driving the observed effects, underscoring the importance of sustained participation for realizing RPM's full impact.

The dynamic trend of early increases followed by subsequent declines in utilization aligns with the literature, suggesting that individuals initially engage more intensively with new tools before stabilizing into routine patterns [47,48]. In the initial phase of RPM adoption, both patients and providers may be adjusting to new data and workflows. Patients may respond to frequent health alerts or health data with heightened awareness, prompting additional interactions with the health system. Providers may proactively schedule follow-ups to verify readings, resolve technical issues, or re-evaluate treatment plans due to the additional data.

The subsequent decrease might indicate that as patients become more accustomed to the data, they take a more measured approach to seeking care. It remains unclear whether these patterns represent appropriate, preventive care or unnecessary and wasteful utilization. The increase could stem from useful data and engagement that identify unmet needs and ensure patients are getting necessary care, or it could represent low-value encounters due to the large influx of data or cautious clinical behavior. Without information on clinical appropriateness, clinical outcomes, and more information on the visits, it is not possible to determine whether the short-term increase benefits patients and the health system.

This observed pattern likely reflects a complex interaction of patient and provider behaviors shaped by differences in digital literacy, clinical workflows, and familiarity with RPM. For instance, improved patient self-management is one of the key benefits of RPM and can lead to better outcomes, but it is dependent on patients' ability to understand health data [49,50]. Both health and digital literacy are associated with age and education [50]. Given that this study focused primarily on older adults in a state with relatively low average education levels, some patients may have lacked the digital or health literacy required to meaningfully engage with RPM. This could limit the intervention's potential benefits or even contribute to uncertainty-driven care-seeking rather than reducing visits.

Additionally, our study focused on early RPM adopters and physicians who were still learning effective RPM implementation methods. Clinicians were still learning how to interpret continuous data, integrate RPM into workflows, and communicate with patients. A systematic review of factors of RPM success highlights several physician-level factors for effectiveness, including targeting populations at high risk, providing responsive and timely care, and ensuring collaborative and coordinated care [45]. If providers are learning how to provide timely and coordinated care via new technologies, RPM may not have its intended benefit. Together, these patient- and provider-level factors suggest that the short-term rise in utilization may reflect a broader adaptation period as both groups learned to navigate new technologies and patterns of care.

Moreover, this study focuses specifically on Louisiana Medicare. The state has a relatively high chronic disease burden, an older average population, and low levels of health and digital literacy [51,52]. These are all factors that can slow early adoption and increase initial contact with providers as patients and clinicians adapt to new tools. At the same time, Louisiana's high poverty rate and large rural population may limit broadband access [53], contributing to declining utilization after initial implementation phases stabilize. Finally, reimbursement for RPM under other payers (Medicaid and commercial programs) was only gradually expanded in Louisiana after 2019 as well, constraining early uptake and shaping provider behavior during the study period. These contextual features suggest that both the learning period and structural factors unique to Louisiana likely influenced the temporal dynamics observed in this study.

From a policy and clinical perspective, these findings have dual and uncertain implications. While increased short-term utilization may raise costs, if the increased visits reflect timely, preventive, or necessary care, the investment could yield downstream savings through decreased complications and their associated treatments. Conversely, if the increased encounters are duplicative or not clinically necessary, they could add costs. Policymakers and payers should therefore interpret early utilization spikes with caution. To prove that these visits should be encouraged, future research should further assess costs and clinical endpoints, which can clarify whether observed utilization patterns translate into value-based outcomes.

In the meantime, these results do offer some recommendations on how to maximize the benefits of RPM and potential challenges that need to be mitigated. First, providers should

XSL•FO
RenderX

prioritize patient education initiatives; ensure that patients, especially older adults in Medicare, receive training on how to collect, interpret, and respond to RPM data. This finding is supported by a recent systematic review showing that RPM was more effective when patients had education and support in interpreting data [22]. Education programs can help enhance patient engagement and self-management while preventing the misuse of health care resources. Additionally, health care providers need structured support, including training on integrating RPM into care delivery and systems in place, to ensure that they have the resources to analyze RPM data.

Limitations

As this study is observational in nature, the ability to assess causality is limited due to the potential for confounding and the lack of randomization. There is an inherent bias in how patients receive RPM. While the methodology attempts to overcome this, it may not have been able to completely eliminate biases in patient selection, which could bias the results away from the null because RPM patients are already engaging with the health system. Similarly, we handled confounding through IPTW and using controls, but some confounding may still be present. Additionally, the findings have limited generalizability as the population with diabetes and hypertension in Louisiana Medicare may differ from other populations. Therefore, these results may not be applicable when assessing or developing policy for other settings, including younger patients, other insurance types, or in other states.

Similarly, this study focused only on RPM through 2020, which was before the large increase in RPM and telehealth utilization due to COVID-19, and the study results may not be the same as those reported today, as some of the barriers mentioned here are less of a concern, resulting in results biased towards the null.

Moreover, this study did not assess the duration of RPM utilization or the number of times a physician checked the RPM data. Claims data do not allow assessment of patient engagement with RPM, and this study only looks at RPM prescription, not true utilization, as an outcome, although actual use mitigates the impact. Therefore, this might bias results towards the null if this group is not engaging with the technology or bias upwards if this population is engaged with RPM technology.

Similarly, the RPM CPT code does not indicate the type of device a patient is using, and as a result, these findings cannot be applied to certain RPM types, such as continuous glucose monitoring, and represent the larger impact of RPM as a treatment option. Finally, a significant limitation of this study is that it did not assess the appropriateness of utilization or the impact on clinical outcomes. If the increase in utilization stems from patients receiving timely care or appropriate care, health outcomes may improve.

This study is innovative in its use of comprehensive real-world evidence from statewide Medicare claims to evaluate RPM among older adults with diabetes or hypertension, a population underrepresented in prior studies. The analysis distinguishes short-term onboarding effects from longer-term stabilization in health care use, expanding on potential reasons for inconsistent findings in earlier research. Conducted during a period of expanding RPM reimbursement, these results provide timely insights for clinicians and payers on targeting, education, and workflow integration to optimize RPM value and reduce low-value encounters.

Conclusions

This study assessed the association between RPM and HCRU among Louisiana Medicare beneficiaries with diabetes or hypertension. We found that RPM was associated with a modest but statistically significant increase in health care utilization, highlighting important questions about how RPM is implemented and integrated into routine care. This study contributes to the existing literature by examining RPM at scale in a real-world Medicare population using administrative claims data, rather than focusing on short-term clinical outcomes or small pilot programs. Unlike prior studies that primarily evaluate changes in biometric measures, our findings underscore the need to consider utilization patterns as an early indicator of how RPM influences care delivery. Although the current analysis could not assess the clinical appropriateness or downstream value of these encounters, the results suggest that initial increases in utilization may reflect enhanced monitoring, care coordination, or unmet need. Future work should link claims with clinical and cost data, examine patient engagement and digital literacy, and identify implementation strategies that ensure RPM supports high-value, equitable care delivery.

Acknowledgments

ChatGPT and Curie were used to edit the manuscript for clarity by asking GPT to analyze specific sections and given the prompt: “For this section please suggest edits for clarity or let me know how to improve the writing”. Additionally, ChatGPT was used to debug Stata code. The authors reviewed, verified, and take full responsibility for all content.

Funding

The authors received no financial support for this research. Therefore, there is no funder involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the CMS Research Identifiable File Data Use Agreement, which prohibits the sharing of raw data.

XSL•FO
RenderX

Authors' Contributions

DW led the conceptualization of the study, with support from LZ, CS, and JL. DW also led the formal analysis, supported by CS, and led the development of the methodology, with contributions from LZ and CS. Supervision was led by LZ, with support from CS and JL. DW prepared the original draft of the manuscript. CS, LZ, and JL contributed equally to the review and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Residual Plots.
[DOCX File, 376 KB - jmir_v28i1e80782_app1.docx]

Multimedia Appendix 2
Sensitivity Analyses.
[DOCX File, 26 KB - jmir_v28i1e80782_app2.docx]

References

1. National diabetes statistics report. : Centers for Disease Control; 2024.
2. Martin SS, Aday AW, Almarzooq ZI, et al. 2024 heart disease and stroke statistics: a report of US and global data from the American heart association. Circulation 2024 Feb 20;149(8):e347-e913. [doi: 10.1161/CIR.0000000000001209] [Medline: 38264914]
3. Hypertension. Centers for Disease Control and Prevention (CDC). 2024. URL: https://www.cdc.gov/nchs/fastats/hypertension.htm [accessed 2025-01-15]
4. Siddiqui H, Imran Z, Ali D, et al. A rising crisis: escalating burden of diabetes mellitus and hypertension-related mortality trends in the United States, 2000-2023. Clin Cardiol 2025 Jul;48(7):e70167. [doi: 10.1002/clc.70167] [Medline: 40600774]
5. Cairns C, Kang K. National hospital ambulatory medical care survey: 2021 emergency department summary tables. Centers for Disease Control and Prevention (CDC). 2021. URL: https://ftp.cdc.gov/pub/health_statistics/nchs/dataset_documentation/nhamcs/doc21-ed-508.pdf [accessed 2024-08-01]
6. Mills KT, Stefanescu A, He J. The global epidemiology of hypertension. Nat Rev Nephrol 2020 Apr;16(4):223-237. [doi: 10.1038/s41581-019-0244-2]
7. American Diabetes Association. Economic costs of diabetes in the U.S. in 2017. Diabetes Care 2018 May 1;41(5):917-928. [doi: 10.2337/dci18-0007]
8. Kirkland EB, Heincelman M, Bishu KG, et al. Trends in healthcare expenditures among US adults with hypertension: national estimates, 2003-2014. J Am Heart Assoc 2018 May 30;7(11):e008731. [doi: 10.1161/JAHA.118.008731] [Medline: 29848493]
9. Wang Y, Lee JS, Pollack LM, Kumar A, Honeycutt S, Luo F. Health care expenditures and use associated with hypertension among U.S. adults. Am J Prev Med 2024 Dec;67(6):820-831. [doi: 10.1016/j.amepre.2024.07.005]
10. Nicholson K, Liu W, Fitzpatrick D, et al. Prevalence of multimorbidity and polypharmacy among adults and older adults: a systematic review. Lancet Healthy Longev 2024 Apr;5(4):e287-e296. [doi: 10.1016/S2666-7568(24)00007-2] [Medline: 38452787]
11. Fang M, Wang D, Coresh J, Selvin E. Trends in diabetes treatment and control in U.S. adults, 1999–2018. N Engl J Med 2021 Jun 10;384(23):2219-2228. [doi: 10.1056/NEJMsa2032271]
12. Smith TM. A troubling trend: BP control drops in adults with hypertension. American Medical Association. 2022. URL: https://www.ama-assn.org/delivering-care/hypertension/troubling-trend-bp-control-drops-adults-hypertension#:~:text=The%20percentage%20of%20U.S.%20adults,67.7%25%20in%202017%E2%80%932020 [accessed 2024-10-12]
13. Rifkin DE. Lost in translation: why are rates of hypertension control getting worse over time? Am J Kidney Dis 2024 Jan;83(1):101-107. [doi: 10.1053/j.ajkd.2023.06.004]
14. Brush JE Jr, Lu Y, Liu Y, et al. Hypertension trends and disparities over 12 years in a large health system: leveraging the electronic health records. JAHA 2024 May 7;13(9). [doi: 10.1161/JAHA.123.033253]
15. Venkatraman S, Echouffo-Tcheugui JB, Selvin E, Fang M. Trends and disparities in glycemic control and severe hyperglycemia among US adults with diabetes using insulin, 1988-2020. JAMA Netw Open 2022 Dec 1;5(12):e2247656. [doi: 10.1001/jamanetworkopen.2022.47656]
16. Gonçalves A, Campos MSA, Menezes LD, Pereira LRL. Barriers and facilitators to medication adherence in chronic diseases: a scoping review. Ciênc saúde coletiva 2025;30(2). [doi: 10.1590/1413-81232025302.02762023en]
17. Leo DG, Buckley BJR, Chowdhury M, et al. Interactive remote patient monitoring devices for managing chronic health conditions: systematic review and meta-analysis. J Med Internet Res 2022 Nov 3;24(11):e35508. [doi: 10.2196/35508]

XSL•FO
RenderX

18. El-Rashidy N, El-Sappagh S, Islam SMR, M El-Bakry H, Abdelrazek S. Mobile health in remote patient monitoring for chronic diseases: principles, trends, and challenges. Diagnostics (Basel) 2021 Mar 29;11(4):607. [doi: 10.3390/diagnostics11040607] [Medline: 33805471]
19. Woodhouse AG, Orvin C, Rich C, Crosby J, Keedy CA. Diabetes outcomes before and during telehealth advancements surrounding COVID-19. J Am Pharm Assoc (2003) 2022 Jan;62(1):214-217. [doi: 10.1016/j.japh.2021.09.011]
20. Coye MJ, Haselkorn A, DeMello S. Remote patient management: technology-enabled innovation and evolving business models for chronic disease care. Health Aff (Millwood) 2009 Jan;28(1):126-135. [doi: 10.1377/hlthaff.28.1.126]
21. Paul MM, Khera N, Elugunti PR, et al. The state of remote patient monitoring for chronic disease management in the United States. J Med Internet Res 2025 Jun 26;27:e70422. [doi: 10.2196/70422] [Medline: 40570326]
22. Tan SY, Sumner J, Wang Y, Wenjun Yip A. A systematic review of the impacts of remote patient monitoring (RPM) interventions on safety, adherence, quality-of-life and cost-related outcomes. npj Digit Med 2024;7(1). [doi: 10.1038/s41746-024-01182-w]
23. Wartenberg C, Elden H, Frerichs M, et al. Clinical benefits and risks of remote patient monitoring: an overview and assessment of methodological rigour of systematic reviews for selected patient groups. BMC Health Serv Res 2025 Jan 23;25(1):133. [doi: 10.1186/s12913-025-12292-w] [Medline: 39849519]
24. Smith W, Colbert BM, Namouz T, Caven D, Ewing JA, Albano AW. Remote patient monitoring is associated with improved outcomes in hypertension: a large, retrospective, cohort analysis. Healthcare (Basel) 2024;12(16):1583. [doi: 10.3390/healthcare12161583]
25. Sachmechi I, Salam S, Amini M, Khan R, Spitznogle A, Belen T. Frequent monitoring of blood glucose levels via a remote patient monitoring system helps improve glycemic control. Endocr Pract 2023 Jun;29(6):441-447. [doi: 10.1016/j.eprac.2023.03.270] [Medline: 36965657]
26. Dennis L, Koyfman I. Effect of remote patient monitoring on stage 2 hypertension. Am J Manag Care 2025 May 6;31(9):e244-e248. [doi: 10.37765/ajmc.2025.89742] [Medline: 40834201]
27. Katalenich B, Shi L, Liu S, et al. Evaluation of a remote monitoring system for diabetes control. Clin Ther 2015 Jun 1;37(6):1216-1225. [doi: 10.1016/j.clinthera.2015.03.022] [Medline: 25869625]
28. Muller AE, Berg RC, Jardim PSJ, Johansen TB, Ormstad SS. Can remote patient monitoring be the new standard in primary care of chronic diseases, post-COVID-19? Telemed J E Health 2022 Jul;28(7):942-969. [doi: 10.1089/tmj.2021.0399] [Medline: 34665645]
29. Wei NJ, Nathan DM, Wexler DJ. Glycemic control after hospital discharge in insulin-treated type 2 diabetes: a randomized pilot study of daily remote glucose monitoring. Endocr Pract 2015 Feb;21(2):115-121. [doi: 10.4158/EP14134.OR] [Medline: 25148814]
30. Li Y, Jiang Y, Tang Y. Is remote blood pressure monitoring and management a better approach for patients with hypertension? A narrative review. J of Clinical Hypertension 2023 Feb;25(2):121-126 [FREE Full text] [doi: 10.1111/jch.14624]
31. Noah B, Keller MS, Mosadeghi S, et al. Impact of remote patient monitoring on clinical outcomes: an updated meta-analysis of randomized controlled trials. npj Digit Med 2018;1(1):20172. [doi: 10.1038/s41746-017-0002-4] [Medline: 31304346]
32. Zhang Y, Peña MT, Fletcher LM, Lal L, Swint JM, Reneker JC. Economic evaluation and costs of remote patient monitoring for cardiovascular disease in the United States: a systematic review. Int J Technol Assess Health Care 2023 Apr 28;39(1):e25. [doi: 10.1017/S0266462323000156] [Medline: 37114456]
33. Mehta SJ, Volpp KG, Troxel AB, et al. Remote blood pressure monitoring with social support for patients with hypertension. JAMA Netw Open 2024 Jun 3;7(6):e2413515. [doi: 10.1001/jamanetworkopen.2024.13515]
34. De Lathauwer ILJ, Nieuwenhuys WW, Hafkamp F, et al. Remote patient monitoring in heart failure: a comprehensive meta‐analysis of effective programme components for hospitalization and mortality reduction. Eur J Heart Fail 2025 Sep;27(9):1670-1685. [doi: 10.1002/ejhf.3568] [Medline: 39834044]
35. Taylor ML, Thomas EE, Snoswell CL, Smith AC, Caffery LJ. Does remote patient monitoring reduce acute care use? a systematic review. BMJ Open 2021 Mar;11(3):e040232. [doi: 10.1136/bmjopen-2020-040232]
36. Jackson TN, Sreedhara M, Bostic M, et al. Telehealth n the United States: A systematic review and meta-analysis, 2011–2021. Telemed Rep 2023;4(1):67-86. [doi: 10.1089/tmr.2023.0011] [Medline: 37283852]
37. De Guzman KR, Snoswell CL, Taylor ML, Gray LC, Caffery LJ. Economic evaluations of remote patient monitoring for chronic disease: a systematic review. Value Health 2022 Jun;25(6):897-913. [doi: 10.1016/j.jval.2021.12.001]
38. Acharya M, Ali MM, Bogulski CA, et al. Association of remote patient monitoring with mortality and healthcare utilization in hypertensive patients: a medicare claims-based study. J Gen Intern Med 2024 Apr;39(5):762-773. [doi: 10.1007/s11606-023-08511-x] [Medline: 37973707]
39. Zhang Z, Henley T, Schiaffino M, et al. Older adults' perceptions of community-based telehealth wellness programs: a qualitative study. Inform Health Soc Care 2022 Oct 2;47(4):361-372. [doi: 10.1080/17538157.2021.2006198] [Medline: 34822311]
40. Satariano WA, Scharlach AE, Lindeman D. Aging, place, and technology: toward improving access and wellness in older populations. J Aging Health 2014 Dec;26(8):1373-1389. [doi: 10.1177/0898264314543470] [Medline: 25502245]

XSL•FO
RenderX

41. Akar JG, Bao H, Jones PW, et al. Use of remote monitoring is associated with lower risk of adverse outcomes among patients with implanted cardiac defibrillators. Circ Arrhythm Electrophysiol 2015 Oct;8(5):1173-1180. [doi: 10.1161/CIRCEP.114.003030] [Medline: 26092577]
42. Hayes CJ, Dawson L, McCoy H, et al. Utilization of remote patient monitoring within the United States health care system: a scoping review. Telemedicine and e-Health 2023 Mar 1;29(3):384-394. [doi: 10.1089/tmj.2022.0111]
43. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008]
44. Angrist J, Pischke JS. Mostly Harmless Econometrics: An Empiricist's Companion: Princeton University Press; 2009. [doi: 10.1515/9781400829828]
45. Thomas EE, Taylor ML, Banbury A, et al. Factors influencing the effectiveness of remote patient monitoring interventions: a realist review. BMJ Open 2021 Aug 25;11(8):e051844. [doi: 10.1136/bmjopen-2021-051844] [Medline: 34433611]
46. Lee JY, Chan CKY, Chua SS, et al. Telemonitoring and team-based management of glycemic control on people with type 2 diabetes: a cluster-randomized controlled trial. J Gen Intern Med 2020 Jan;35(1):87-94. [doi: 10.1007/s11606-019-05316-9] [Medline: 31512187]
47. Eaton C, Vallejo N, McDonald X, et al. User engagement with mHealth interventions to promote treatment adherence and self-management in people with chronic health conditions: systematic review. J Med Internet Res 2024 Sep 24;26:e50508. [doi: 10.2196/50508] [Medline: 39316431]
48. Rogers EM. Diffusion of Innovation: The Free Press; 1995.
49. van den Bergh R, Evers LJW, de Vries NM, et al. Usability and utility of a remote monitoring system to support physiotherapy for people with Parkinson's disease. Front Neurol 2023;14. [doi: 10.3389/fneur.2023.1251395]
50. Boriani G, Maisano A, Bonini N, et al. Digital literacy as a potential barrier to implementation of cardiology tele-visits after COVID-19 pandemic: the INFO-COVID survey. J Geriatr Cardiol 2021 Sep 28;18(9):739-747. [doi: 10.11909/j.issn.1671-5411.2021.09.003] [Medline: 34659380]
51. Association AD. The Burden of Diabetes in Louisiana. 2023. URL: https://diabetes.org/sites/default/files/2023-09/ADV_2023_State_Fact_sheets_all_rev_Louisiana.pdf [accessed 2024-08-01]
52. Bazzano AN, Monnette AM, Wharton MK, et al. Older patients' preferences and views related to non-face-to-face diabetes chronic care management: a qualitative study from southeast Louisiana. Patient Prefer Adherence 2019;13:901-911. [doi: 10.2147/PPA.S201072] [Medline: 31213782]
53. Gillispie-Bell V, Bane S, Martin J, Kan P, Mujahid MS, Carmichael SL. Severe maternal morbidity in Louisiana by race, rurality, poverty, and availability of care. Public Health 2025 Sep;246:105824. [doi: 10.1016/j.puhe.2025.105824] [Medline: 40543228]

Abbreviations

CITSA: controlled interrupted time series
COPD: chronic obstructive pulmonary disease
CPT: common procedure terminology
ER: emergency room
HCRU: health care resource utilization
ICD: International Classification of Diseases
IPTW: Inverse Probability Treatment Weight
RPM: remote patient monitoring
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 21.Aug.2025; peer-reviewed by M Chakit, Y Wu; revised version received 12.Nov.2025; accepted 13.Nov.2025; published 03.Mar.2026.

Please cite as:
Winberg D, Stoecker C, Li J, Shi L
Remote Patient Monitoring in Louisiana Medicare Beneficiaries With Diabetes or Hypertension: Retrospective Cohort Study of 2016-2020 Claims Data
J Med Internet Res 2026;28:e80782
URL: https://www.jmir.org/2026/1/e80782
doi:10.2196/80782

XSL•FO
RenderX

© Debra Winberg, Charles Stoecker, Jian Li, Lizheng Shi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 3.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Extended Grammar of Systematized Nomenclature of Medicine – Clinical Terms for Semantic Representation of Clinical Data: Methodological Study

Christophe Gaudet-Blavignac[2,1*], PhD; Julien Ehrsam[2,1*], MD; Monika Baumann[1], BSc; Adel Bensahla[2,1], MSc; Mirjam Mattei[1], MSc; Yuanyuan Zheng[2,1], MSc; Christian Lovis[2,1], MPH, MD

[1]Division of medical information sciences, Diagnostic department, University Hospital of Geneva, Geneva, Switzerland

[2]Department of radiology and medical informatics, Faculty of medicine, University of Geneva, Geneva, Switzerland

[*]these authors contributed equally

Corresponding Author:
Christophe Gaudet-Blavignac, PhD
Division of medical information sciences
Diagnostic department
University Hospital of Geneva
Rue Gabrielle-Perret-Gentil 4
Geneva, 1205
Switzerland
Phone: 41 3790815
Email: christophe.gaudet-blavignac@hug.ch

Abstract

Background: Interoperability has been a challenge for half a century. Led by an informatics view of the world, the quest for interoperability has evolved from typing and categorizing data to building increasingly complex models. In parallel with the development of these models, the field of terminologies and ontologies emerged to refine granularity and introduce notions of hierarchy. Clinical data models and terminology systems vary in purpose, and their fixed categories shape and constrain representation, which inevitably leads to information loss.

Objective: Despite these efforts, semantic interoperability remains imperfect. Achieving it is essential for effective data reuse but requires more than rich terminologies and standardized models. This methodological study explores the extent to which the SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) compositional grammar can be leveraged and extended to approximate a formal descriptive grammar, allowing clinical reality to be expressed in coherent, meaningful sentences rather than preconstrained categories.

Methods: Building on a decade of semantic representation efforts at the Geneva University Hospitals, we developed a framework to identify recurring semantic gaps in clinical data. We addressed these gaps by systematically modifying the SNOMED CT Machine Read` Concept Model and extending its Augmented Backus-Naur Form syntax to support necessary grammatical structures and external vocabularies.

Results: This approach enabled the semantic representation of over 119,000 distinct data elements covering 13 billion instances. By extending the grammar, we successfully addressed critical limitations such as negation, scalar values, uncertainty, temporality, and the integration of external terminologies like Pango. The extensions proved essential for capturing complex clinical nuances that standard precoordinated concepts could not represent.

Conclusions: Rather than creating a new standard from scratch, extending the grammatical capabilities of SNOMED CT offers a viable pathway toward high-fidelity semantic representation. This work serves as a proof-of-concept that separating the rules of composition from vocabulary allows for a more flexible and robust description of clinical reality, provided that challenges regarding governance and machine readability are addressed.

(J Med Internet Res 2026;28:e80314) doi:10.2196/80314

KEYWORDS

data models; formal grammar; knowledge representation; semantic interoperability; terminology standards

XSL•FO
RenderX

Introduction

Status of Health Care Interoperability

Achieving semantic interoperability remains a central challenge in biomedical informatics and health data integration. While technical interoperability ensures that systems can exchange data in compatible formats, semantic interoperability guarantees that exchanged information is unambiguously interpretable across heterogeneous systems and contexts. This objective requires the explicit representation of meaning through formal models, ontologies, and standardized vocabularies, enabling computational reasoning and automated integration of clinical data.

Significant efforts have been invested in developing tools and frameworks to bring semantics to the complex field of health data. In Switzerland, the Swiss Personalized Health Network and its 3-pillar strategy defined a strong semantic representation of data as the first and mandatory pillar [1]. At the European level, the European Health Data Space initiative has further emphasized the importance of standardized, interoperable health data to enable secure sharing and secondary use across member states [2]. Similarly, in the United States, the Office of the National Coordinator for Health IT promotes this goal through the US Core Data for Interoperability, a standardized set of data elements required for nationwide health information exchange [3]. For years, the field of semantic interoperability has been supported by clinical data models, considered technical standards, and classifications and terminologies, called semantic standards.

The multiplication of semantic standards has prompted the need for ontology alignment, or ontology matching, to establish semantic correspondences between heterogeneous terminologies and domain ontologies. Classical systems such as AgreementMaker leverage a combination of lexical similarity, structural consistency, and description logic reasoning to ensure coherent mappings between ontologies [4]. More recent approaches, such as BERTMap, apply contextual language models to improve recall and precision in complex clinical terminologies [5].

Compositional grammar (CG) models provide a structured means of representing the internal semantics of clinical expressions. Clinical concepts are inherently compositional and thus require formalisms that can represent their components and relationships. Frameworks such as the SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) CG define syntactic and semantic rules for postcoordinated expressions (PCEs), ensuring internal consistency and machine interpretability [6].

Semantic interoperability also relies on formal logic–based models, typically expressed in Description Logics (DL), which underpin the Web Ontology Language. These models enable key reasoning tasks such as classification, consistency checking, and inferencing [7]. Building on this foundation, knowledge graphs have emerged as a flexible paradigm for integrating heterogeneous biomedical data, representing entities and their relationships within a unified semantic space.

Several international standardization bodies have developed complementary frameworks to support semantic interoperability. Health Level Seven Fast Healthcare Interoperability Resources (FHIR) provides a resource-oriented model for data exchange, including semantic bindings to standard terminologies such as SNOMED CT, Logical Observation Identifiers Names and Codes (LOINC), and Unified Code for Units of Measure, and offers Resource Description Framework and Web Ontology Language representations for semantic web integration [8]. The Observational Medical Outcomes Partnership (OMOP) Common Data Model, developed within the Observational Health Data Sciences and Informatics initiative, harmonizes observational health data using standardized vocabularies and facilitates analytical interoperability [9]. Cross-domain efforts, including ISO (International Organization for Standardization) 23903 and the World Wide Web Consortium's Health Care and Life Sciences group, promote semantic harmonization through formal metadata and linked data principles [10].

Each of these models has its strengths and weaknesses, and no system fits all purposes [11-13]. Creating a data model ultimately involves defining a finite set of categories, making choices that inherently shape what can and cannot be represented. These reflect the intended purpose of the model, but fitting data into these finite sets inevitably modifies it, making reuse based on model-specific categories less granular [14]. Semantic standards similarly have specific purposes, and granularity can vary dramatically depending on the focus of the classification. The *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) can be useful in representing disease, but cannot represent surgical interventions. Representing data using only single codes from a classification will inherently result in information loss that will be tied to the focus and granularity of the terminology. Figure 1 depicts this loss using a commonly known situation and its representation in natural language, then in *ICD-10* codes.

XSL•FO
RenderX

Figure 1. Information loss when using single standards to represent reality.

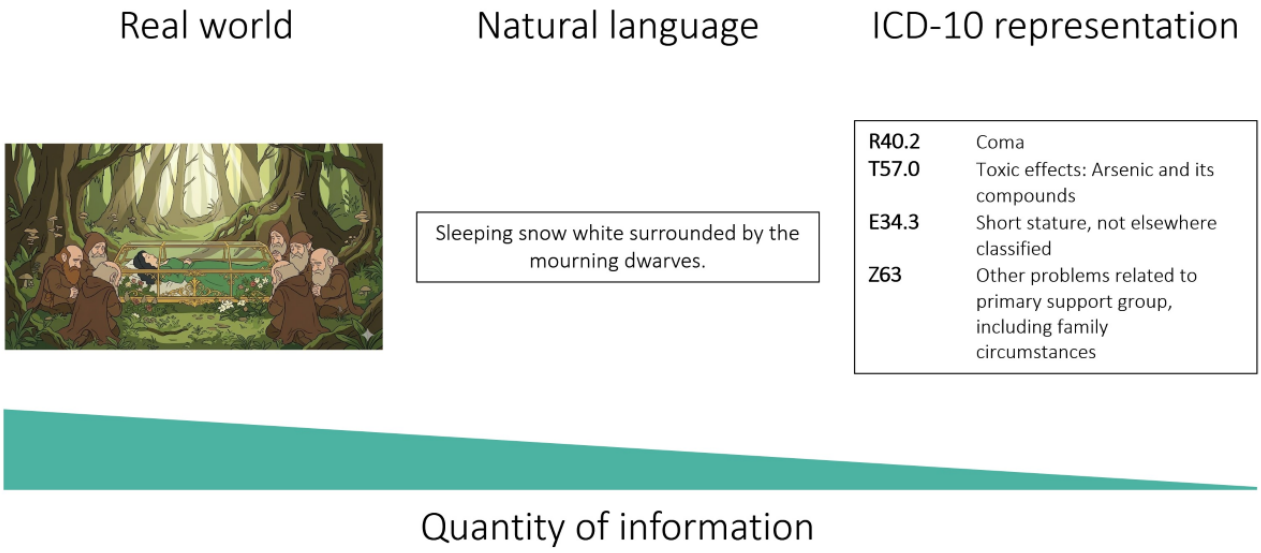

SNOMED CT and Natural Language

In this landscape, the CG of SNOMED CT, combined with its broad coverage, makes it a strong candidate for semantic interoperability [15]. With its knowledge graph–like structure, it allows intelligent code retrieval through its expression constraint language and gives users the ability to create PCEs to represent absent concepts. This relies on 2 elements. The CG, an Augmented Backus-Naur Form (ABNF) syntax, is used to create composed expressions that can be parsed and evaluated automatically [16]. The Machine Readable Concept Model (MRCM) defines which concepts (domain) can be refined with which relation (attribute) to which values (range). Combining these resources with the editorial guide, which defines attributes-naming conventions, ensures that new expressions comply with SNOMED CT's rules and grammar [17].

However, while designed for postcoordination and full expressivity of clinical concepts, it is rarely used to its full capacity and is intended for SNOMED CT concepts only [18,19]. It lacks granularity in some fields, such as strains of SARS-CoV-2, while others, such as genomic sequences, are not covered at all. Using SNOMED CT to represent clinical structured data within its expected parameters is eventually restrictive and insufficient to cover everything. Each attribute has a very specific range, domain, definitions, and permitted use cases, known as domain-specific modeling, and extending it eventually becomes necessary [20].

Like natural languages, SNOMED CT can be described as consisting of 3 fundamental components: a vocabulary of words, a grammar defining the rules for their combination, and the resulting sentences. Natural languages, however, allow the creation of any sentence respecting their grammar, regardless of their meaning, and regularly integrate words from other languages. A sentence such as "The 'entrepreneur' dropped his children off at 'kindergarten' before going to a 'karaoke'" uses words from French, German, and Japanese, which have been fully integrated into the English language. It is possible to draw inspiration from this structure to form a similar, computer-processable language, and applying this philosophy to SNOMED CT could improve its usability to fully represent clinical data semantically. The field of medical informatics already possesses rich vocabularies such as SNOMED CT, which serve as the source of machine-readable "words," each with its own specific strengths and weaknesses. They must then be combined according to well-defined grammar rules to form expressions, for which the SNOMED CT CG provides a comprehensive base.

Beyond SNOMED CT

Despite long-standing efforts to construct a comprehensive medical terminology, a single complete solution remains elusive, often resulting in the use of fragmented assemblies of various systems [21]. What remains necessary is a unifying set of rules to govern composition.

The solution should be formal, as semantic representation requires clear rules for machine readability that define how building blocks assemble into structured expressions. It must be descriptive rather than prescriptive. While grammars guide expression, they should not restrict what can be said a priori but rather ensure reality can be described clearly and consistently. Importantly, the solution must be dynamic, evolving through continuous iteration. Clinical reality is complex and ambiguous, with new findings emerging faster than they can be formally incorporated. Enabling the construction of meaningful expressions even without predefined concepts allows rapid adaptation to this complexity. To the best of our knowledge, such a formal descriptive grammar (FDG) does not exist yet.

This study proposes a first step in this direction by leveraging SNOMED CT to represent clinical data while ensuring machine readability. We summarize requirements gathered through 10 years of manual semantic representation and describe an implementation based on extending SNOMED CT's CG, syntax, and MRCM. This approach aims to represent clinical data regardless of source, output terminology, or intended use case.

XSL•FO
RenderX

While the results presented reflect the current state, they remain open-ended and extensible.

Methods

Overview

This work was developed in parallel to the semantic representation of structured data in the Geneva University Hospitals (HUG) data lake over the past decade. It builds upon the methodological framework introduced in *Semantics in Action: a Guide for Representing Clinical Data Elements* [22] and *Scalar Values in SNOMED CT: a Proposed Extension* [23], which proposes a structured and iterative approach for semantically representing clinical data. Counts of represented metadata elements are derived from the Clinical Data Warehouse without individual data extraction.

Framework

The guide's competency-building framework and rule-creation cycle were followed, incorporating both manual expert encoding and consensus-based refinement of representation rules through regular focus groups to define the FDG. The method used was gradually complexified in focus groups as more rules became necessary.

The teaching framework focuses on the use of SNOMED CT and follows a 3-part approach. First, SNOMED International's (SI) introductory training courses and documentation are taken to familiarize newcomers with the basics of standard-based semantic representation [24]. This allows new team members to start work on a simpler 1-dimensional representation, such as a problem list value set [25]. Next, internal documentation and continued practice give trainees a deeper understanding of real data and local specificities, allowing them to represent more complex data. Finally, more advanced courses, such as the SNOMED CT authoring courses, are taken by long-term core team members [26]. These, along with the practical experience gained by taking part in the training, give the ability to supervise newcomers and to participate in the rule creation cycle described below. New team members are always supervised closely by experienced team members and progressively increase the complexity of their tasks as required. Following this framework ensures that team members participating in the discussions around the rules of representation have a common training and vision.

Over time, complex situations encountered during the process made it clear that SI's rules were too restrictive to allow for complete representation. Initially, only a few internal unwritten rules were agreed upon, but it quickly became evident that, to avoid losing track of erratically evolving rules and changes, a framework was necessary to harmonize the process and maintain coherence. The rule-creation cycle was developed during weekly focus groups, which bring together the team members who participate in the semantic representation effort. They are made up of a core of 4 team members with heterogeneous professional backgrounds, including health care professionals (medical, nursing, midwifery), more technical training (IT, bioinformatics), and administrative knowledge (billing). These can then be joined by others depending on the team's setup at that time, such as students, new team members, or other colleagues with an occasional interest in a specific topic. The core members all have proficient experience with semantic representation and specifically with SNOMED CT, and are certified by SI through courses such as the authoring certification. The core of the team represents a small, soft-funded group, and the rest of the group has a high turnover rate. Added to this are limited time and resources, which means that there is a strong need to prioritize how and where said resources are applied. Data collections with high clinical value and a high number of instances, such as laboratory procedures and patient formularies, were tackled first. For the same reasons, work is parallelized as team members are each assigned separate tasks. For these reasons, no formal interannotator agreement studies are carried out.

Instead, specific topics are discussed during focus groups, and only when outcomes are unanimously agreed upon are they validated. Focus group topics consist of questions or complex situations each member has identified, and the rule-creation cycle is applied to each. First, existing guidelines and approaches are thoroughly reviewed and applied if they can solve the issue. If no existing solution is deemed satisfactory, an extension to the current rules is necessary, and is put into practice once consensus is reached through rounds of discussions. It is evaluated both for coherence with regard to previous rules and for clarity and semantic accuracy. Satisfactory outcomes are added to the list of grammar rules, which are revisited until deemed sufficiently tested. As this method applies only to situations for which the current set of rules available through SI or other groups is deemed insufficient or not applicable, they were not formally compared to each other on the same cases.

Reaching SNOMED CT's Limits

SNOMED CT was selected as the foundational semantic standard for this representation effort due to its comprehensive set of concepts, its adherence to a formal ABNF syntax, and its robust compositional rules that facilitate precise concept modeling. We prioritize the use of single SNOMED CT concepts first whenever they are sufficient to faithfully represent the data's content. When single concepts fall short, we use postcoordination, leveraging the constraints defined by the MRCM to its full capacity, constructing complex clinical expressions.

If neither approach can adequately capture the semantic content, extending it becomes inevitable as a representation gap is formally identified. This is where SNOMED CT's inherent extensibility is crucial for addressing domain-specific requirements. To resolve such gaps, we systematically explore possibilities for modifying the underlying grammar. The primary strategy involves altering existing MRCM rules, specifically through domain or range extensions, as referenced in [27]. Should this be insufficient, we analyze the need to include new SNOMED CT attributes within the representation framework. Finally, if essential semantic concepts are entirely absent from SNOMED CT, an external terminology is integrated, which consequently requires corresponding modifications to the ABNF syntax governing the semantic expressions.

XSL•FO
RenderX

The results detail the principal extensions implemented, which enable semantic coverage for a large percentage of the structured data within the HUG's data lake. The modified SNOMED CT ABNF syntax and MRCM collectively form the contribution of this paper and are available in Multimedia Appendices 1 and 2.

Ethical Considerations

Access to the HUG database is granted on an individual basis by the institution's Chief Data Officer and Medical Director and is reviewed and renewed every 6 months. This authorization permits access for the purpose of viewing and extracting aggregated metadata. No individual-level patient data are ever extracted, processed, or analyzed. All results presented in the manuscript consist exclusively of irreversibly aggregated counts derived from metadata and do not allow identification or reidentification of individuals. Consequently, this work does not constitute research involving human beings or health-related personal data within the meaning of the Swiss Human Research Act (SR 810.30) and falls outside its scope [28]. In accordance with the Swiss Human Research Act and the Swiss Federal Act on Data Protection, neither ethics committee approval nor patient consent was required; therefore, no waiver was sought from the cantonal or hospital ethics committee [29].

Results

Overview

Several problematic situations were encountered, which required progressively drifting away from SNOMED CT's official rules. These are common properties or patterns that occur across all or most natural languages, known as linguistic universals, a concept derived from Chomsky's universal grammar [30]. They are encountered often while representing data, and most can be addressed only partly by applying SNOMED CT's rules, grammar, and concepts. These gaps in the capacity to represent reality include negation, numeration, or scalar values, and displacement, which includes uncertainty and temporality. Most are included in some way in SNOMED CT's design, although rarely implemented in practice, and are therefore solvable by expanding the CG. This is done with domain and range extensions, and the addition of new attribute rules [31], which allow better representation.

Other situations, however, are not intended by SNOMED CT and necessitate breaking the mold to make the CG more inclusive. This is done by integrating other terminologies to fill gaps within SNOMED CT that cannot be handled even with postcoordination, such as SARS-CoV-2 strains. The different situations encountered and solutions devised are detailed below with an example for each. The complete list of MRCM modifications is available in Multimedia Appendix 1, and the modifications done for each property are detailed in Textbox 1.

Textbox 1. Modifications done for each property.

Negation

- 45169001 |Without (attribute)|
- 5185003 |Except for (attribute)|
- 408729009 |Finding context (attribute)|
- 408730004 |Procedure context (attribute)|

Scalars

- 103373006 |With size (attribute)|
- 410671006 |Date (attribute)|
- 79409006 |Resulting in (attribute)|
- 246205007 |Quantity (attribute)|
- 246262008 |Score (attribute)|

Uncertainty

- 408729009 |Finding context (attribute)|
- 408730004 |Procedure context (attribute)|

Temporality

- 255234002 |After (attribute)|
- 288556008 |Before (attribute)|
- 371881003 |During (attribute)|
- 103335007 |Duration (attribute)|

XSL•FO
RenderX

Scalar Values

Scalar values are physical quantities that can be described by a single pure number accompanied by a unit of measurement [32]. Scalars are one of the first domains encountered that were not present in sufficient detail in SNOMED CT. Scalars appear in myriad places in clinical data, whether as laboratory results, scores, sizes, or measurements, and hold great importance for accurate representation of clinical reality [8,18]. In SNOMED CT, integers were previously included as concepts, which have since been removed and replaced by Unicode text expressions preceded by a hash [33]. There remain concepts that contain numerical values in labels, such as quantitative result cutoffs, but not in any attribute relationship, as they are not fully defined [34].

A better representation of scalar values is achieved with new attributes, using existing accepted standards, when possible, such as the ISO 8601 format for dates, which is largely used, including by SNOMED CT [35] (Figure 2). The advantages of representing scalars directly in formal grammar expressions are vast. This greatly improves querying capacity, using queries with numeric operators, such as "all heart rates with a value of >150 bpm."

Figure 2. Scalar value representation example.



Negation

Negation is the act of denying or contradicting something, or the expression of its absence or opposite. The ability to represent negation is a key part of any language. Without it, this sentence could not be written. It does, however, pose many challenges when attempting to include it in a knowledge representation, particularly with regard to inference in hierarchical structures [36]. SNOMED CT's approach to negation through its context model exists outside the DL, as the absence of something is still linked to the focus concept by an "Associated" attribute [37], which is logically incoherent. Other ways of representing negation in SNOMED CT can also lead to errors of inference and classification [36], with children concepts being less restrictive than their parents. In fact, SI goes as far as to recommend "handling negation outside of SNOMED CT ... rather than try and represent it within the terminology" [38].

However, SNOMED CT abounds with concepts containing negation in some form or another. The limitations described above apply to the approved attribute relations that are used in fully defined concepts. For the cases not already covered, plenty of concepts exist that represent negation in one way or another, such as unapproved attributes and qualifiers, which are used to extend the MRCM when necessary. Extending attribute domains and using new attributes such as "Except for" and "Without" has proven highly useful. For hernia repairs, for example, SNOMED CT only specifies when the procedure is done with a mesh. However, no concepts describe the absence of mesh during a procedure, which was needed (Figure 3). These modifications allow us to cover many instances of negation encountered and are currently deemed sufficient.

Figure 3. Negation representation example.



Uncertainty

Uncertainty describes a situation in which something is not known or not certain. This has an important place in representing medical data, as many situations in clinical care contain uncertainty. It is inherent in every patient encounter. It starts with the patient's history, such as not knowing when a symptom began, and continues through the diagnostic process, from forming a differential diagnosis to choosing which tests to order.

Natively, SNOMED CT represents uncertainty in the same way as negation, with the "Finding context" attribute with values such as "Known possible." As with negation, however, SI states that "attempts to capture probabilistic or uncertain knowledge are out of the scope of SNOMED CT" [39], despite containing approved attribute relations for describing such situations. It is also present in primitive concept labels such as "Uncertain diagnosis."

Similarly to negation, the extensions made to cover uncertainty are expanding the domain of the "Finding context" and "Procedure context" attributes to include "<<Clinical finding" and "<<Procedure," respectively. This allows for representing Findings and Procedures with a refinement concerning uncertainty (Figure 4).

XSL•FO
RenderX

Figure 4. Uncertainty representation example.

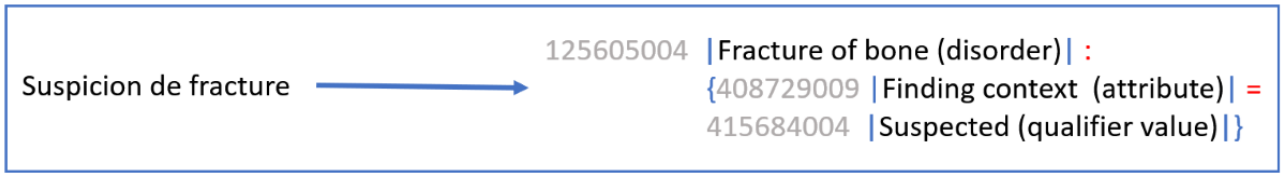

Temporality

Temporality is the state of existing within or having some relationship with time. It is a vital property to consider when analyzing patient data. Indeed, knowing if a certain procedure, laboratory test, or diagnosis happened before, during, or after another is critical and has a defining influence on how information is interpreted. This also includes notions such as duration of processes or procedures, evolution over time, and chronicity.

SNOMED CT already has an extensive coverage of temporality and has proven capable of adequately representing many cases encountered in practice, such as "Temporally related to" (with its descendants After, During, and Before). These apply to various domains such as Situations, Clinical findings, and Observables entities, and therefore already cover many situations. However, despite their extensive coverage, gaps remain in the representation of temporality.

Extensions include extending the range of the Before and During attributes to match that of After. Procedures are added to these attributes' domains. Some larger gaps include dates, which are resolved using the "Date" attribute, and duration of procedures, with the "Duration" attribute (Figure 5), which both use scalar values as necessary.

Figure 5. Temporality representation example.



External Vocabularies Integration

Overview

While the situations mentioned up to now are certainly a welcome improvement in terms of scope of representation, they describe modifications compliant with SNOMED CT extension capabilities. Other types of encountered gaps, which concern specific group of concepts absent from SNOMED CT, are not resolvable in this way. Representing these concepts, therefore, necessitates the inclusion of external terminologies. While using different standards together can hinder interoperability, as it can limit the possible exchange and secondary usage of data, it has become unavoidable. Integrating various formats and standards to make them compatible has been identified as a challenge in the field of medical informatics [40]. In fact, SNOMED CT itself was born from the fusion of SNOMED RT and Clinical Terms V3, so what is proposed here is only an extension of its original philosophy [41]. However, it should be remembered that multiple maps exist between SNOMED CT and terminologies such as ICD-10 and Orphanet [42], as well as the effort to harmonize LOINC and SNOMED CT [43,44]. These should be consulted first to ensure that the desired term does not already have a SNOMED CT equivalent. Ontology alignment efforts and thesauri such as Unified Medical Language System should also be consulted to verify equivalences.

Once this verification is done, external vocabularies can be evaluated for integration. How they are evaluated is subject to rules. We have defined five selection criteria: (1) the terminology must first cover an identified gap in SNOMED CT coverage; (2) it must be concept-based, so cannot be a thesaurus; (3) it must be hierarchical and contain at least some graph-like properties; (4) it must have versioning capabilities; and (5) it must be a well-recognized standard widely accepted by relevant users. Once chosen, the steps described below should be followed to fully integrate them, increasing the scope of representation. This includes adding them virtually to a SNOMED CT hierarchy by attributing a parent, verifying MRCM rules, and modifying the ABNF syntax.

Content Analysis and Proximal Primitive Parent Attribution

To include an external terminology, it is necessary to clearly identify when and where it would be used. The content of the identified terminology should be analyzed in parallel with SNOMED CT to define where it could fit in the SNOMED CT hierarchies. Ideally, an entire terminology should be linked to one parent concept, but if different parts of the terminology fit in different hierarchies, this should be documented, and clusters of codes should be defined if needed. The defined clusters are then attributed to a proximal primitive parent. This task is described in the SNOMED CT authoring course for concept creation and consists of defining the closest parent of the cluster of codes in SNOMED CT that is not fully defined [45]. To do so, the closest parent is assigned, and its genealogy is reviewed to find the first primitive concept (Figure 6 [46]).

XSL•FO
RenderX

Figure 6. Proximal primitive parent attribution.

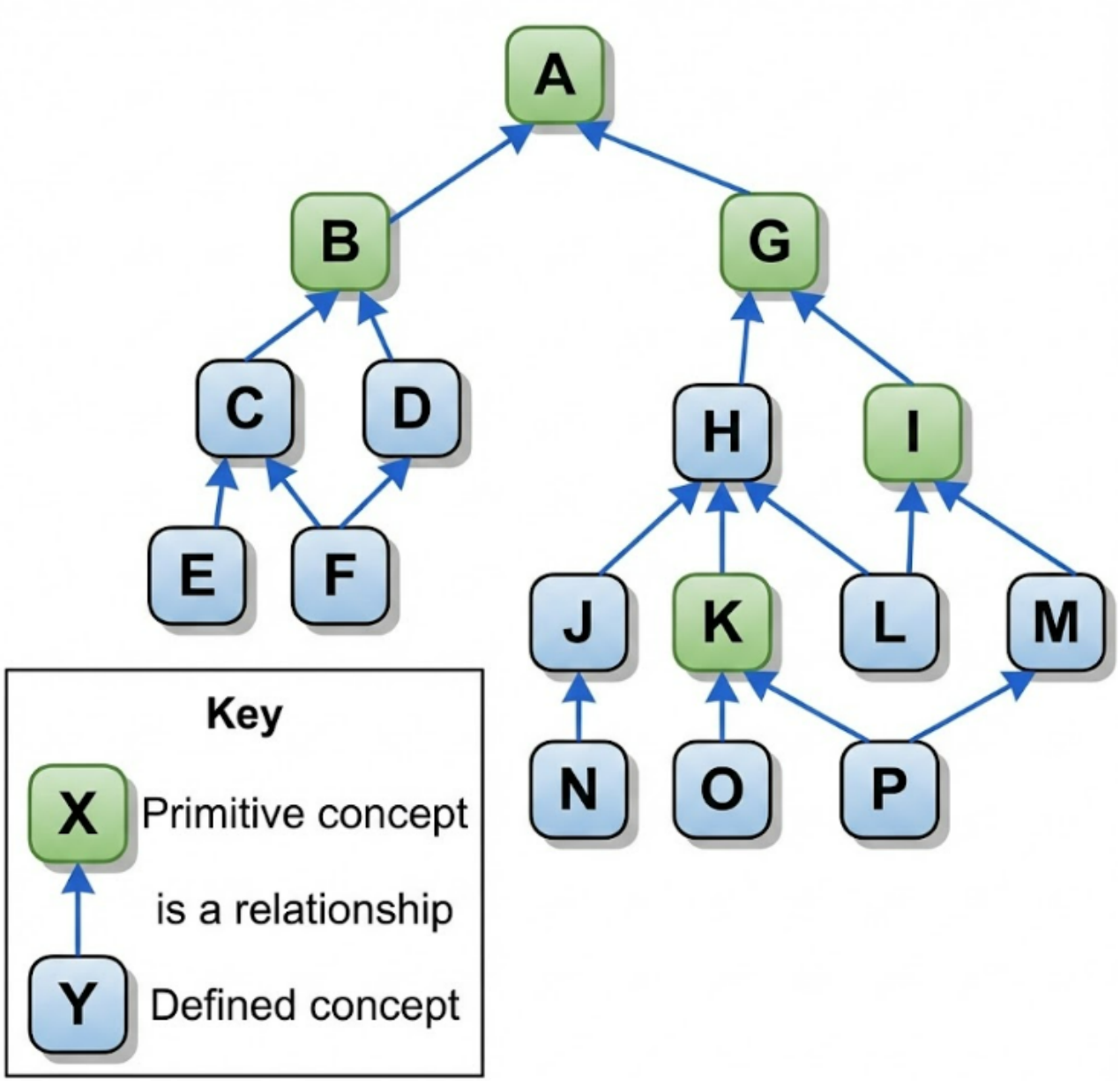

Modifying the MRCM and ABNF

Once the clusters of codes are attributed to a parent, MRCM rules covering those parents must be reviewed to ensure they are still valid. If necessary, they can then be updated, or new rules created, to include the cluster in the range or domain, if they fit an approved definition. This helps in keeping coherence when creating new PCEs, but must be done carefully, as integrations that happen too high in the hierarchy can influence multiple existing MRCM rules.

A syntax modification is then necessary to allow for integration of external concepts into SNOMED CT PCEs. The ABNF syntax is therefore modified to signal when another terminology is used and to allow validation of PCEs. The tilde symbol (~) is used to signal the start and end of the use of external concepts. Another ~ is used to separate the identifier of the vocabulary and the concept used. These new terminology concepts can, in theory, be inserted into any part of an expression, whether as the focus concept (~Terminology~id~:SCT=SCT), attribute (SCT:~Terminology~id~=SCT), value (SCT:SCT=~Terminology~id~), a combination of these, or even used alone. This modification has been validated using a modified ABNF parser on newly created expressions. The modified ABNF syntax is available in Multimedia Appendix 2.

Implementation

When the COVID-19 pandemic hit, genomic sequencing quickly led to a characterization of the SARS-CoV-2 virus in specific strains, with scientific and vernacular names given by organizations such as Nextstrain [47], Pango [48], and the World Health Organization [49]. These strains took on a crucial importance with the advent of more contagious or less lethal variants such as Delta or Omicron [50]. This close focus on specific variants is well represented in classifications such as the Pango nomenclature [51]. However, in SNOMED CT, the representation of those strains is lacking, with a unique code to

XSL•FO
RenderX

characterize the SARS-CoV-2 virus. This situation is used as a use case to demonstrate the inclusion of an external vocabulary through the proposed method. In this scenario, Pango is chosen to refine the organism hierarchy for SARS-CoV-2 lineages. Therefore, a new attribute group with a domain, range, and definition must be validated. Since SNOMED CT already has a category of codes that defines strains but does not include SARS-CoV-2, the Pango nomenclature is added as children of this code. A new MRCM rule is then defined to refine the Organism hierarchy with the "Microbiological strain" attribute, as shown in Figure 7.

Figure 7. External terminology representation example.



Using this method, a total of 5 terminologies other than SNOMED CT are added to expand the scope of representation. These are Pango, Nextstrain, Sequence Ontology [52], RadLex [53], and the Common Terminology Criteria for Adverse Events (CTCAE) [54].

Discussion

Principal Results

Overview

This work addresses the difficult balance between expressivity and standardization with a first step in the direction of an FDG for semantic representation of clinical data, using SNOMED CT as the base. It provides a formal, extensible set of rules for constructing meaning that can draw from diverse vocabularies as needed. This modular approach means new terminologies can be integrated as they become relevant without limiting what can be said, but providing a clear structure for how to say it.

After 10 years of representing structured data in the HUG datalake, the FDG now provides semantic representations for 119,941 distinct data elements using 39,168 unique expressions. These expressions cover over 13 billion data instances from 1.7 million patients. Notably, 18,370 of these unique expressions are PCEs, validating the approach as nearly half of all unique expressions created are complex PCEs. This clearly demonstrates that precoordinated concepts alone are profoundly insufficient to capture clinical reality at scale. Moreover, 4500 PCEs use the extensions presented in this paper, for a total of nearly 500 distinct extended PCEs, underscoring their usefulness.

This approach differs from existing interoperability efforts. While FHIR provides a flexible structural framework to define custom data models (FHIR profiles), its focus is not on prescribing semantic content, leaving users to define their own semantic frameworks. In comparison to FHIR extensions, the approach is similar, but we connect the external terminology to SNOMED CT to maintain semantic coherence. To the best of our knowledge, this is not allowed in FHIR extensions, which do not specify the relation of the extension to the resource it extends. Crucially, neither FHIR nor OMOP supports the definition of PCEs, a capability central to our solution for expressing complex clinical nuances. In comparison to ontology alignment frameworks, our approach does not align ontologies but allows their combined use to compose concepts absent in both. Furthermore, a key difference lies in the methodology; our solution focuses entirely on the semantic description of data where it resides, avoiding the necessity of data movement or physical transformation inherent in many FHIR or OMOP-based implementations. Because our primary intent is semantic description and not structural or physical data modeling, a direct, formal comparison between these solutions is not appropriate.

The decision to build upon and extend SNOMED CT is grounded in the need for a robust and logical foundation, which its unique extensible grammar and syntax provide. This work shows that extending a powerful existing standard can be a path toward semantic interoperability. Four large domains of clinical reality, identified as lacking in SNOMED CT and known as challenging fields to represent, have been partially solved using this approach. Despite allowing for the representation of a large amount of data elements that were out of reach of current SNOMED CT rules, there remain gaps in representation that have not yet been resolved. Each category described previously has seen its coverage improved, but as things stand, complete coverage of these domains is still out of reach.

A statistical analysis of the extensions used shows that the 4500 PCEs created using extensions represent less than 4% of data elements and 3% of instances (Table 1). However, a deeper analysis shows that while seemingly sparse overall, extensions are very useful for certain data types. Laboratory procedures and medical devices, for example, are two collections in which 27% (250,617,656/935,678,204) and 23% (1,140,820/4,926,815) of instances, respectively, were represented with extensions. Other collections, such as formularies, have nearly 4000 extensions used.

XSL•FO
RenderX

Table 1. Detailed statistics of representation by category of data using extensions.

Source	IDs encoded, n	Extensions used, n (%)	Instances encoded	Extension instances, n (%)
Administrative stay	1172	2 (0.17)	78,020,550	131,832 (.17)
Anesthesiology	2139	11 (0.51)	117,419,522	1,275,337 (1.09)
Medical devices	699	235 (33.62)	4,926,815	1,140,820 (23.16)
Formularies	51,322	3925 (7.65)	1,641,633,208	119,941,273 (7.31)
ICU[a]	6934	157 (2.26)	6,656,119,500	367,548 (.005)
Laboratory	7231	72 (1)	935,678,204	250,617,656 (26.78)
Observations	2784	1 (0.03)	2,440,891,370	1003 (.00004)
Patient problems	25,233	72 (0.28)	1,587,507	976 (.06)
Prescription	9837	0 (0)	26,805,381	0 (0)
Radiology	2586	2 (0.07)	8,562,371	2 (.00002)
Procedures	8001	23 (0.28)	1,452,038,108	8,934,163 (.6)
Total	119,941	4500 (3.75)	13,362,644,974	382,410,610 (2.86)

[a]ICU: intensive care unit.

Closer examination shows clearly that the extensions have proven extremely useful to fill certain specific gaps within SNOMED CT. In these cases, the gaps identified were SARS-CoV-2 strains for laboratory procedures, sizes for medical devices, and occupations for formularies. Looking in even further detail, the 4500 PCEs created represent 488 distinct expressions. However, certain specific expressions have proven very useful. Extending representation of negation was a big part, with the Without and Except for attributes used a combined 88 times, and sizes for medical devices used 178 times.

Machine-Readable Interoperability

A primary concern regarding the modification of established standards is the potential loss of machine readability. Our implementation presents a dichotomy in this regard. The modifications made strictly to the MRCM, such as domain and range extensions for scalars or negation attributes, remain compliant with SNOMED CT's structure. These are natively processable by standard terminology servers like Snowstorm, provided the modified MRCM is loaded [55].

However, the integration of external vocabularies required modifications to the ABNF syntax (specifically the use of the tilde delimiter), which constitutes a divergence from the standard CG. Consequently, standard parsers and expression constraint language query engines cannot currently process these specific expressions without adaptation. We have validated that these expressions can be parsed using a modified ABNF parser, and we are currently adapting open-source tools to accommodate this syntax. Until these tools are widely available, this specific aspect of the grammar hinders immediate interoperability with the existing global standards ecosystem.

This technical gap is compounded by a governance challenge, as the number of integrated terminologies increases, so does the need for robust documentation to prevent ambiguity. A formal glossary or registry is therefore required to track which terminologies are in use, their specific versions, and the rules governing their application. For clarity, versioning could even be embedded directly within an expression, for example: ~CTCAE~v5.0~4028512~. A more sustainable formalism could use Uniform Resource Identifiers to designate terminology releases. This could be implemented without modifying the ABNF.

Challenges Encountered

Scalars are a massive chapter, and representing certain situations, such as ranges, has not been resolved yet. Allowing for scalar values in the range of most hierarchies also requires defining new attribute relations for each use case encountered. Negation is resolved by exclusively using and extending existing SNOMED CT concepts, as this is sufficient for our needs and maintains internal consistency within the grammar. However, this is not the only valid approach. More complex scenarios could be addressed by incorporating a formal negation operator, such as the widely accepted "¬" symbol, as described by Schulz et al [37]. Similarly, reification offers another method for handling negation within SNOMED CT, though it lacks a directly queryable negation indicator [36,37]. This would require additional ABNF syntax modification, and the impact it would have on DL means that this has not been implemented.

Other challenging situations fall into different categories but account for only a small fraction of all data elements, such as rare, complex labels within otherwise well-resolved categories. These include highly specific laboratory procedures that represent less than 1% (1,475,290/250,521,754) of all laboratory procedure instances. They could likely be addressed by incorporating more specialized terminologies. However, because these cases are anecdotal, the effort required to search for and integrate additional terminology is not justified for now.

Data-element categories with no widely adopted standard and no suitable SNOMED CT codes for creating new PCEs are another challenge. An example is triage-sheet questions, which include 350 distinct labels across 4 million instances, a negligible portion of the overall dataset. Work is ongoing to find solutions, primarily through new MRCM rules, such as

XSL•FO
RenderX

defining an appropriate relationship between Observable Entities and Clinical Findings (current options like “Precondition” or “Has realization” are inadequate), or between Environments and Procedures.

Finally, some grammatical constructions are poorly supported by current grammar rules. A key case is the presence of “or” in problem-list entries. For concepts containing “or,” we currently apply a rule that the label is represented only if both concepts share a sufficiently close semantic parent; otherwise, it remains unmapped. Some labels can be resolved this way, such as Intervention de gynécologie ou d’obstétrique = Operation on female genital organs (procedure). But others are too semantically broad or distant, with no precise match, such as Pneumonie ou pneumopathie, where the shared parent Disorder of lung (disorder) is considered too vague and would cause excessive information loss. In total, our database contains 2895 labels with “or,” representing 17.5 million instances. Among these, 834 (28%) labels have been successfully resolved, covering 10 (57%) million instances.

Limitations

This work presents limitations. For reasons mentioned previously, no interannotator agreement was carried out on the work done. The goal was initially to develop the approach for internal use only. Since it evolved gradually in parallel to our improving capabilities, there was never a “before” and “after” to evaluate. Also, since the team is made up of a set core of 3-4 members, it was considered more useful for us to evaluate our work in focus group discussions. There was never a need internally for a more explicit evaluation because problems are reviewed so often. Finally, while the intention is to generalize the way people use SNOMED CT to represent data, one of the founding principles of SNOMED CT is that there are many ways to express the same thing. As the most important outcome is semantic accuracy, maintaining the meaning of the data element, if two expressions mean the same thing but are written differently, they are still both correct. The expected outcomes were therefore never exactly matched. Such subtleties would not appear in a formal interannotator agreement, reducing its interpretability, an important point, which, in our opinion, diminished the potential impact of such an evaluation.

The byproduct of an inclusive grammar, as described in this work, is that it will inevitably make reuse more difficult. There will be links created that should not work, and inconsistencies in DL. The primary limitation of this implementation lies in the operational integration of external vocabularies with standard SNOMED CT tools. While separation of grammar and vocabulary was identified as a crucial need for an FDG, its implementation in the proposed approach is incomplete. The MRCM we chose to keep and extend constitutes a set of rules that are semantic, not merely grammatical. Currently, the terminologies added to the grammar are attributed a SNOMED CT code as a parent, to allow initial querying. But the new codes themselves cannot be directly accessed yet. For this to function, the integrated terminologies need to be definitively added to SNOMED CT. There are ways to do this in theory, such as creating a new concept with a valid SNOMED CT concept identifier for each, but they have not yet been explored in detail or implemented. Applying these solutions would greatly enhance reuse capabilities, improving interoperability and consistency of results. Modifying the ABNF syntax and parser is a first step in this direction.

The results only cover the sources that have been selected for representation so far. As the project continues, new sources will be analyzed, extracted, and represented. Therefore, it is not possible, to date, to give a clear image of the progress compared to the complete data warehouse. However, the choice of the sources to be added first has been designed to cover first the core of the electronic health record, before smaller sources. Additionally, new problems will inevitably appear, which will need to be addressed, and new rules will have to be added. As this work focuses primarily on structured data, it is not possible to confront our semantic coverage to the full information content of the electronic health record, but this is being addressed through automatic annotation of free text using natural language processing in a similar manner.

Finally, the approach was tested only in our environment and is currently used nowhere else. The end goal, however, is, of course, to apply it elsewhere to test its reproducibility, as the complexity and variability of the approach mean it may not be immediately applicable in different settings.

Governance and Implementation

Key recommendations can be derived from this work to address the limitations discussed above. To safely adopt these extensions, institutions should establish a minimum governance structure consisting of a multidisciplinary committee of domain experts and SNOMED CT–certified terminologists. This body acts as a gatekeeper, validating new representation rules through a consensus-based cycle only when standard SNOMED CT concepts or official maps (eg, LOINC and Orphanet) are proven insufficient. Key steps toward alignment with SI include maintaining a formal local registry of all grammar extensions and prioritizing the use of standard extension mechanisms (MRCM) over syntax modifications to preserve compatibility with the broader ecosystem. In terms of maturity, the MRCM extensions for scalar values, temporality, uncertainty, and negation are stable and ready for immediate reuse in standard terminology servers (eg, Snowstorm). Conversely, the ABNF syntax modifications required for integrating external vocabularies remain experimental; they currently necessitate custom parsing tools and can be adopted as a proof-of-concept pending the evolution of standard tooling.

Future work

Steps have been taken to start implementing the strategy in different institutions, as interest has been strong, and collaborations are currently ongoing, but are at a preliminary stage for the moment. External replication and independent validation are being undertaken, which aim to evaluate the adaptability and reusability of our approach, demonstrating its generalizability. Additionally, specific projects in the hospital have already benefited from using our approach to gather data on subjects such as biomarkers in the field of genomics, which was not possible before.

XSL•FO
RenderX

Furthermore, while this work relied on expert manual representation, future research will focus on automating the translation of free-text clinical notes into PCEs using natural language processing. Emerging methodologies leverage large language models (LLMs) equipped with retrieval-augmented generation to ground generated expressions in the extensive SNOMED CT hierarchy, a technique showing potential for high-fidelity clinical coding [56]. However, we acknowledge that the generation of structured sequences via LLMs introduces specific risks, such as exposure bias, which can affect the reliability of the output. Recent work highlights the importance of mitigating these biases in LLM distillation to ensure robust structured generation [57]. Addressing these challenges would dramatically scale implementation and unlock the potential for advanced semantic reasoners to infer new knowledge from grammatically rich, machine-readable clinical data.

Conclusions

This methodological study demonstrates how SNOMED CT's CG can be leveraged and extended to address recurrent semantic gaps encountered in a large-scale clinical data warehouse. Rather than constructing a new FDG from scratch, we have approached this theoretical goal by systematically extending the capabilities of an existing standard. Through specific modifications to the MRCM and syntax, we successfully addressed complex representational challenges such as negation, scalar values, and the integration of external terminologies, thereby reducing the tension between clinical expressivity and standardization.

Supported by a decade of implementation at HUG, this work illustrates that a grammatical framework, separating the rules of composition from the vocabulary itself, is essential for capturing meaning at scale. While standardized vocabularies provide the necessary lexical building blocks, it is the flexibility of the grammar that determines the fidelity of the representation. This study serves as a proof-of-concept that semantic interoperability can be advanced by methodically extending the expressive power of existing standards. However, the widespread adoption of such extensions requires robust governance frameworks to collaboratively manage and share grammatical rules across institutions.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025) [58], the following tasks were delegated to GAI tools under full human supervision: proofreading, editing, and reformatting. The GAI tools used were Gemini 2.5 Pro, Gemini 3, and ChatGPT 4.5. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

This declaration is submitted by CGB.

Funding

This work was funded with a grant from the Private Foundation of the Geneva University Hospitals. Funding by the Swiss Personalized Health Network initiative has partly contributed to enabling this work. The funding sources had no input or influence in the decision to publish this paper, and were not involved in the writing, editing, or submission procedures.

Authors' Contributions

Conceptualization: CGB, JE, CL
Data curation: CGB, JE, MB, AB, MM, YZ
Methodology: CGB, JE, MB, AB, MM, YZ
Visualization: JE
Supervision: CGB, CL
Funding acquisition: CL
Validation: CL
Project administration: CL
Writing – original draft: CGB, JE
Writing – review & editing: CGB, JE, CL

Conflicts of Interest

None declared.

Multimedia Appendix 1
Machine Readable Component Model modifications and additions.
[XLSX File (Microsoft Excel File), 12 KB - jmir_v28i1e80314_app1.xlsx]

Multimedia Appendix 2
Modified Augmented Backus-Naur Form syntax.
[TXT File , 2 KB - jmir_v28i1e80314_app2.txt]

XSL•FO
RenderX

References

1. Gaudet-Blavignac C, Raisaro JL, Touré V, Österle S, Crameri K, Lovis C. A national, semantic-driven, three-pillar strategy to enable health data secondary usage interoperability for research within the Swiss personalized health network: methodological study. JMIR Med Inform 2021;9(6):e27591 [FREE Full text] [doi: 10.2196/27591] [Medline: 34185008]
2. The European Health Data Space. URL: https://www.european-health-data-space.com/ [accessed 2025-05-27]
3. Gordon W, Gottlieb D, Kreda D, Mandel J, Mandl K, Kohane I. Patient-led data sharing for clinical bioinformatics research: USCDI and beyond. J Am Med Inform Assoc 2021;28(10):2298-2300 [FREE Full text] [doi: 10.1093/jamia/ocab133] [Medline: 34279631]
4. Cruz IF, Antonelli FP, Stroe C. AgreementMaker: efficient matching for large real-world schemas and ontologies. 2009 Presented at: Proceedings of the VLDB Endowment: Proceedings of the 35th International Conference on Very Large Data Bases; 2009 August 24-28; Lyon, France p. 1586-1589. [doi: 10.14778/1687553.1687598]
5. He Y, Chen J, Antonyrajah D, Horrocks I. BERTMap: a BERT-based ontology alignment system. 2022 Presented at: Proceedings of the AAAI Conference on Artificial Intelligence; 2022 Feb 22-Mar 1; Vancouver p. 5684-5691. [doi: 10.1609/aaai.v36i5.20510]
6. Compositional grammar specification | specifications SNOMED CT compositional grammar specification | SNOMED international documents. SNOMED. URL: https://docs.snomed.org/snomed-ct-specifications/snomed-ct-compositional-grammar-specification [accessed 2025-11-28]
7. Baader F, Horrocks I, Sattler U. Description logics. In: Handbook of Knowledge Representation. Amsterdam: Elsevier; 2008:135-179.
8. Benson T, Grieve G. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR. 3rd ed. London: Springer; 2016.
9. OMOP Common Data Model? OHDSI. URL: https://www.ohdsi.org/data-standardization/the-common-data-model/ [accessed 2020-02-24]
10. Health informatics? Interoperability and integration reference architecture? Model and framework. ISO. URL: https://www.iso.org/obp/ui/#iso:std:iso:23903:ed-1:v2:en [accessed 2025-11-28]
11. Wang L, Wen A, Fu S, Ruan X, Huang M, Li R, et al. A scoping review of OMOP CDM adoption for cancer research using real world data. NPJ Digit Med 2025;8(1):189 [FREE Full text] [doi: 10.1038/s41746-025-01581-7] [Medline: 40189628]
12. Sholle ET, Cusick M, Davila MA, Kabariti J, Flores S, Campion TR. Characterizing basic and complex usage of i2b2 at an academic medical center. AMIA Jt Summits Transl Sci Proc 2020;2020:589-596 [FREE Full text] [Medline: 32477681]
13. Ceusters W, Rabenberg M. Semantic difficulties in FHIR 'conditions'. Stud Health Technol Inform 2025;327:7-11. [doi: 10.3233/SHTI250263] [Medline: 40380375]
14. Kent S, Burn E, Dawoud D, Jonsson P, Østby JT, Hughes N, et al. Common problems, common data model solutions: evidence generation for health technology assessment. Pharmacoeconomics 2021;39(3):275-285 [FREE Full text] [doi: 10.1007/s40273-020-00981-9] [Medline: 33336320]
15. Chute CG. Clinical classification and terminology: some history and current observations. J Am Med Inform Assoc 2000;7(3):298-303 [FREE Full text] [doi: 10.1136/jamia.2000.0070298] [Medline: 10833167]
16. 5.1 normative specification--compositional grammar--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCSCG/5.1+Normative+Specification [accessed 2025-07-06]
17. SNOMED CT editorial guide--SNOMED CT editorial guide--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCEG [accessed 2024-02-20]
18. Lee D, Cornet R, Lau F, de Keizer N. A survey of SNOMED CT implementations. J Biomed Inform 2013;46(1):87-96 [FREE Full text] [doi: 10.1016/j.jbi.2012.09.006] [Medline: 23041717]
19. Lee D, de Keizer N, Lau F, Cornet R. Literature review of SNOMED CT use. J Am Med Inform Assoc 2014;21(e1):e11-e19 [FREE Full text] [doi: 10.1136/amiajnl-2013-001636] [Medline: 23828173]
20. Domain specific modeling | specifications SNOMED CT editorial guide | SNOMED International Documents. SNOMED. URL: https://docs.snomed.org/snomed-ct-specifications/snomed-ct-editorial-guide/readme/authoring/domain-specific-modeling [accessed 2025-11-28]
21. Cornet R, Chute CG. Health concept and knowledge management: twenty-five years of evolution. Yearb Med Inform 2016;Suppl 1(Suppl 1):S32-S41 [FREE Full text] [doi: 10.15265/IYS-2016-s037] [Medline: 27488404]
22. Ehrsam J, Gaudet-Blavignac C, Mattei M, Baumann M, Lovis C. Semantics in action: a guide for representing clinical data elements with SNOMED CT. J Biomed Semantics 2025;16(1):7 [FREE Full text] [doi: 10.1186/s13326-025-00326-5] [Medline: 40149003]
23. Ehrsam J, Gaudet-Blavignac C, Lovis C. Scalar values in SNOMED CT: a proposed extension. Stud Health Technol Inform 2024;316:1363-1367. [doi: 10.3233/SHTI240665] [Medline: 39176634]
24. Course categories | e-learning. SNOMED International. URL: https://elearning.ihtsdotools.org/course/ [accessed 2025-11-28]
25. Gaudet-Blavignac C, Rudaz A, Lovis C. Building a shared, scalable, and sustainable source for the problem-oriented medical record: developmental study. JMIR Med Inform 2021;9(10):e29174 [FREE Full text] [doi: 10.2196/29174] [Medline: 34643542]
26. SNOMED CT authoring level 1 course. SNOMED International. URL: https://courses.ihtsdotools.org/product?catalog=AL1 [accessed 2025-11-28]

XSL•FO
RenderX

27. 3 requirements--machine readable concept model--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCMRCM/3.+Requirements [accessed 2024-11-28]
28. SR 810.30--Federal Act of 30 September 2011 on research. Fedlex. URL: https://www.fedlex.admin.ch/eli/cc/2013/617/en [accessed 2025-12-19]
29. RO 2022 491 - Loi fédérale du 25 septembre 2020 sur la protection des données. Fedlex. URL: https://www.fedlex.admin.ch/eli/oc/2022/491/fr [accessed 2025-12-19]
30. Dąbrowska E. What exactly is universal grammar, and has anyone seen it? Front Psychol 2015;6:852 [FREE Full text] [doi: 10.3389/fpsyg.2015.00852] [Medline: 26157406]
31. 10 extension and customization--SNOMED CT starter guide--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCSTART/10.+Extension+and+Customization [accessed 2025-04-29]
32. Scalar (physics). Wikipedia. URL: https://en.wikipedia.org/w/index.php?title=Scalar_(physics)&oldid=1279729075 [accessed 2025-04-29]
33. SNOMED International proposal for representing concrete domains in RF2. SNOMED International. URL: https://confluence.ihtsdotools.org/mag/concrete-domains-community-of-practice-consultation/snomed-international-proposal-for-representing-concrete-domains-in-rf2 [accessed 2024-03-18]
34. Rector AL, Brandt S. Why do it the hard way? The case for an expressive description logic for SNOMED. J Am Med Inform Assoc 2008;15(6):744-751 [FREE Full text] [doi: 10.1197/jamia.M2797] [Medline: 18755993]
35. ISO 8601? Date and time format. ISO. 2017. URL: https://www.iso.org/iso-8601-date-and-time-format.html [accessed 2025-05-08]
36. Martínez-Costa C, Miñarro-Giménez J, Hausam R, Schulz S. Addressing the negation gap in SNOMED CT by reified negated concepts. J Biomed Semantics 2015;6:23. [doi: 10.1186/s13326-015-0018-1]
37. Schulz S, Markó K, Suntisrivaraporn B. Formal representation of complex SNOMED CT expressions. BMC Med Inform Decis Mak 2008;8 Suppl 1(Suppl 1):S9 [FREE Full text] [doi: 10.1186/1472-6947-8-S1-S9] [Medline: 19007446]
38. Situation with explicit context modeling--SNOMED CT editorial guide--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCEG/Situation+with+Explicit+Context+Modeling [accessed 2025-05-06]
39. Knowledge representation--SNOMED CT editorial guide--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCEG/Knowledge+Representation [accessed 2025-05-07]
40. Haendel MA, Chute CG, Robinson PN. Classification, ontology, and precision medicine. N Engl J Med 2018;379(15):1452-1462 [FREE Full text] [doi: 10.1056/NEJMra1615014] [Medline: 30304648]
41. Rothwell DJ, Cote RA, Cordeau JP, Boisvert MA. Developing a standard data structure for medical language--the SNOMED proposal. Proc Annu Symp Comput Appl Med Care 1993:695-699 [FREE Full text] [Medline: 8130565]
42. SNOMED CT maps. SNOMED International. URL: https://www.snomed.org/maps [accessed 2024-11-28]
43. Request form for LOINC/SNOMED CT expression association and map sets file. LOINC. URL: https://loinc.org/snomed-file-request/ [accessed 2024-11-28]
44. LOINC/SNOMED CT. URL: https://loincsnomed.org/ [accessed 2024-11-28]
45. Authoring--SNOMED CT editorial guide--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCEG/Authoring [accessed 2025-06-06]
46. Proximal primitive parent--SNOMED CT glossary--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCGLOSS/proximal+primitive+parent [accessed 2025-06-06]
47. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 2018;34(23):4121-4123 [FREE Full text] [doi: 10.1093/bioinformatics/bty407] [Medline: 29790939]
48. O'Toole Á, Pybus OG, Abram ME, Kelly EJ, Rambaut A. Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences. BMC Genomics 2022;23(1):121 [FREE Full text] [doi: 10.1186/s12864-022-08358-2] [Medline: 35148677]
49. Tracking SARS-CoV-2 variants. World Health Organization. URL: https://www.who.int/activities/tracking-SARS-CoV-2-variants [accessed 2025-06-09]
50. CoVariants. URL: https://covariants.org/ [accessed 2025-06-09]
51. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 2020;5(11):1403-1407 [FREE Full text] [doi: 10.1038/s41564-020-0770-5] [Medline: 32669681]
52. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, et al. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol 2005;6(5):R44 [FREE Full text] [doi: 10.1186/gb-2005-6-5-r44] [Medline: 15892872]
53. Langlotz CP. RadLex: a new method for indexing online educational materials. Radiographics 2006;26(6):1595-1597. [doi: 10.1148/rg.266065168] [Medline: 17102038]
54. Trotti A, Colevas AD, Setser A, Rusch V, Jaques D, Budach V, et al. CTCAE v3.0: development of a comprehensive grading system for the adverse effects of cancer treatment. Semin Radiat Oncol 2003;13(3):176-181. [doi: 10.1016/S1053-4296(03)00031-6] [Medline: 12903007]

XSL•FO
RenderX

55. Terminology servers. Implementation. SNOMED. URL: https://www.implementation.snomed.org/terminology-services [accessed 2025-11-28]
56. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. 2022 Presented at: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing Abu Dhabi, United Arab Emirates: Association for Computational Linguistics; 2022; Abu Dhabi p. 1998-2022. [doi: 10.18653/v1/2022.emnlp-main.130]
57. Pozzi A, Incremona A, Tessera D, Toti D. Mitigating exposure bias in large language model distillation: an imitation learning approach. Neural Comput Applic 2025;37(18):12013-12029. [doi: 10.1007/s00521-025-11162-0]
58. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (generative AI delegation taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2025:1-27. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]

Abbreviations

ABNF: Augmented Backus-Naur Form
CG: compositional grammar
CTCAE: Common Terminology Criteria for Adverse Events
DL: Description Logics
FDG: formal descriptive grammar
FHIR: Fast Healthcare Interoperability Resources
HUG: Geneva University Hospitals
ICD-10: International Statistical Classification of Diseases, Tenth Revision
ISO: International Organization for Standardization
LLM: large language model
LOINC: Logical Observation Identifiers Names and Codes
MRCM: Machine Readable Component Model
OMOP: Observational Medical Outcomes Partnership
PCE: postcoordinated expression
SI: SNOMED International
SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms

Edited by J Sarvestan; submitted 08.Jul.2025; peer-reviewed by A Pozzi, T Cook; comments to author 21.Aug.2025; revised version received 19.Dec.2025; accepted 20.Dec.2025; published 28.Jan.2026.

Please cite as:
Gaudet-Blavignac C, Ehrsam J, Baumann M, Bensahla A, Mattei M, Zheng Y, Lovis C
Extended Grammar of Systematized Nomenclature of Medicine – Clinical Terms for Semantic Representation of Clinical Data: Methodological Study
J Med Internet Res 2026;28:e80314
URL: https://www.jmir.org/2026/1/e80314
doi:10.2196/80314
PMID:

©Christophe Gaudet-Blavignac, Julien Ehrsam, Monika Baumann, Adel Bensahla, Mirjam Mattei, Yuanyuan Zheng, Christian Lovis. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Establishment and Optimization of a Patient-Reported Outcome–Based Electronic-Diary for Symptoms Evaluation in Patients With Gastroesophageal Reflux Disorder: Prospective Cohort Study

Yun-Chun Chen[1], MS; Yen-Po Wang[2,3,4], MD, PhD; Jui-Hsuan Hung[5], MS; Da-Wei Wang[5†], PhD; Shang-Liang Wu[6], DrPH; Li-Fen Chen[2], PhD; Yueh-Hsin Ping[1,7], PhD; Mei-Lien Pan[8], PhD; Ching-Liang Lu[2,3,4], MD

[1]Department and Institute of Pharmacology, National Yang Ming Chiao Tung University, Taipei, Taiwan

[2]Institute of Brain Science, National Yang Ming Chiao Tung University, Taipei, Taiwan

[3]Division of Gastroenterology, Department of Medicine, Taipei Veterans General Hospital, Taipei, Taiwan

[4]Endoscopy Center for Diagnosis and Treatment, Department of Medicine, Taipei Veterans General Hospital, No.201, Sec. 2, Shipai Rd., Beitou District, Taipei, Taiwan

[5]Institute of Information Science, Academia Sinica, Taipei, Taiwan

[6]Department of Medical Research, Taipei Veterans General Hospital, Taipei, Taiwan

[7]Institute of of Biophotonics, National Yang Ming Chiao Tung University, Taipei, Taiwan

[8]Institute of Hospital and Health Care Administration, National Yang Ming Chiao Tung University, Taipei, Taiwan

[†]deceased

Corresponding Author:
Ching-Liang Lu, MD
Institute of Brain Science, National Yang Ming Chiao Tung University, Taipei, Taiwan

Abstract

Background: Gastroesophageal reflux disease (GERD) symptoms substantially impair patients' quality of life. The use of patient-reported outcome (PRO) instruments for symptom measurement has been advocated by regulatory authorities. However, current tools for GERD symptom evaluation are limited by recall bias. To improve the real-time characterization of GERD symptoms, we developed an electronic diary (e-diary) for daily symptom monitoring.

Objective: This study aimed to develop and optimize a PRO-based e-diary for GERD symptom evaluation and to examine the effect of symptom frequency on adherence.

Methods: The GERD e-diary evaluated 8 daytime (acid regurgitation, cough, heartburn, sour taste in the mouth, hiccups, hoarseness, dysphagia, and chest pain) and 2 nighttime symptoms (acid regurgitation and cough) for 8 consecutive weeks. Adherence, defined as the daily completion rate of e-diary, was analyzed and optimized across three stages: (1) no reminder, (2) sending reminder SMS text messaging upon the detection of missing data (no reminders during the first 3-5 days after enrollment), and (3) immediate installation of reminder system at enrollment. Weekly symptom frequency was calculated as the sum of symptomatic days per week. Multiple regression analyses were performed to examine the effects of system optimization and symptom frequency on adherence after controlling for confounders.

Results: A total of 138 patients with GERD (70 men, 68 women; mean [SD] age 52.9 [12.3] years) were recruited. At the first stage, the adherence was 47.2%, 40%, and 57.6% for nighttime, daytime, and overall symptoms. System optimization significantly improved the adherence of nighttime symptoms by 12.5% (95% CI 3.7-21.3) and 10.9% (95% CI 2.6-19.2), daytime symptoms by 21.7% (95% CI 14.2-29.2) and 20.8% (95% CI 13.7-27.9), and overall symptoms by 16.5% (95% CI 9.8-23.2) and 18.5% (95% CI 12.2-4.8) in the second and third stages, respectively. Symptom frequency was positively associated with adherence, increasing by 0.7% (95% CI 0.6-0.8) for overall symptoms and 0.9% (95% CI 0.7-1) for both daytime and nighttime symptoms per additional symptom frequency. Adherence gradually decreased along the study period. (first vs eighth week: nighttime 80.1% vs 61.5%, β=−18.6, 95% CI −26.9 to −10.3; daytime 85.1% vs 66.8%, β=−18.3, 95% CI −25.6 to −11; overall 95.1% vs 78%, β=−17.2, 95% CI −23.5 to −10.9).

Conclusions: The adherence of the GERD e-diary can be optimized by using SMS text messaging reminders. Higher symptom frequency was associated with increased adherence, although engagement declined over time. This innovative PRO-based e-diary with prolonged recording provides a real-time, prospective tool that overcomes the recall and ecological biases inherent in traditional short-term retrospective GERD symptom assessments. This advancement empowers patients through improved

XSL•FO
RenderX

self-awareness and provides physicians with precise, long-term data, facilitating tailored therapeutic interventions and supporting personalized GERD management.

(J Med Internet Res 2026;28:e83680) doi:10.2196/83680

KEYWORDS

adherence; e-diary; gastroesophageal reflux disease; GERD; symptom frequency; system optimization

Introduction

Gastroesophageal reflux disease (GERD) is a common disorder affecting approximately 20% of the general population [1]. GERD symptoms, including typical (heartburn and acid regurgitation) and atypical (cough, chest pain, sour taste in the mouth, hoarseness, dysphagia, and hiccups) symptoms, would bring significant burdens on patients' quality of life, reduce work productivity, and increase considerable medical resources worldwide [2-6]. As a symptom-driven disease, GERD should be evaluated for the presence, frequency, and severity of bothersome symptoms [1]. However, patient-reported clinical management outcomes often remain unsatisfactory, with more than 50% of patients continuing to experience bothersome symptoms despite proton pump inhibitor therapy [7]. This underscores the need for more nuanced approaches to symptom assessment and management.

Patient-reported outcomes (PROs) are the most direct and measured instrument for evaluation, an approach advocated by the Food and Drug Administration of the United States to assess treatment efficacy [8]. In GERD research, PROs have been used to evaluate the symptomatic improvements after acid-suppressant drugs, such as proton pump inhibitors or histamine-H2-receptor antagonists [9-11]. PROs have also been used in long-term studies to evaluate symptom resolution after surgery and endoscopic interventions and to assess improvements in quality of life [12-14]. However, most previous studies relied on questionnaires that required patients to recall symptoms over prolonged periods, possibly leading to recall bias. Furthermore, during outpatient evaluation and follow-up in patients with GERD, important information about the time course of symptom occurrence and relief is still limited because the caring physicians typically assess bothersome symptoms after weeks or even months of treatment. Notably, it has been reported that a 30-day recall of self-reported urinary incontinence is impaired and can be associated with demographic and psychosocial characteristics [15]. Patients with GERD are expected to face similar situations when using questionnaires to recall their symptoms [16-18]. Therefore, daily assessments of the GERD symptoms would provide a better understanding of the natural course of GERD symptoms and potentially alleviate the recall bias. Although patients with chronic gastroenterological diseases express strong interest in using mobile health apps for disease management, real-world adherence remains low, with high dropout rates [19]. Some studies report that up to 80% of participants engage only minimally or discontinue regular use, with retention rates dropping to as low as 3.9% after 15 days [20-22]. Existing GERD questionnaires also face limitations: their perceived impracticality for routine clinical use, along with challenges inherent to traditional paper-based symptom tracking—such as recall bias, inconsistent formatting, and poor adherence leading to incomplete or inaccurate data—collectively impede accurate and continuous symptom assessment for both patients and clinicians [23]. Given the high patient interest in digital health solutions, there is a critical need for novel mobile tools capable of overcoming these traditional limitations [24]. Although there are some smartphone apps for gastrointestinal disease, including GERD, available in the market, no app is compliant with the 2022 American College of Gastroenterology Guideline and not designed for physicians' management of GERD symptoms, and most of the apps are not evidence based [24-26].

To overcome recall bias and better address the progression and regression of GERD symptoms, in this study, we developed an e-diary for recording GERD symptoms. As low adherence in GERD diary recording would potentially limit the accountability of the e-diary, affecting clinical assessments and the associated outcomes, we tried to develop a strategy to enhance the 8-week adherence using the e-diary. Furthermore, our clinical observations and existing literature suggest a strong link between symptom severity and recording motivation: patients experiencing more frequent or intense symptoms typically demonstrate greater adherence and accuracy in documentation [27,28]. Therefore, we subsequently tested that higher symptom frequency would be associated with increased adherence to e-diary recording. The objective of the study was to examine the effects of symptom frequency and the degree of system optimization on weekly adherence, with the hypothesis that after adjusting for potential confounders, higher symptom frequency and greater system optimization would be associated with better weekly adherence. This will serve as a basis for developing strategies to improve patient adherence in completing the e-diary, thereby enhancing the clinical validity of the e-diary for GERD symptoms.

Methods

In this study, we followed the EQUATOR (Enhancing the Quality and Transparency of Health Research) guidelines using the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist [29].

Study Design and Participants

The study was a prospective observational cohort study. From October 2021 to January 2023, consecutive adults (aged ≥20 y) with GERD were recruited from the outpatient clinics of gastroenterology in Taipei Veterans General Hospital, a tertiary medical center in northern Taiwan. Eligible participants presented with either typical (acid reflux or heartburn) or atypical symptoms (hoarseness, throat discomfort, cough, or chest pain) for at least 3 months. Heartburn or regurgitation

XSL•FO
RenderX

should be noted for 4 days or more during the 7 days before the first visit. Patients who expressed interest and consented to daily symptom recording using the GERD e-diary were enrolled. A trained research assistant installed the GERD e-diary on the mobile phones of participants and provided instructions for use. These patients were asked to complete the GERD e-diary on a daily basis for 8 weeks. One or two follow-up visits in gastroenterology clinics would be arranged. To minimize potential selection and information bias, consecutive eligible patients attending the outpatient gastroenterology clinics were approached, and all data were prospectively recorded using standardized digital forms. As symptom data were collected directly from patients through an electronic diary (e-diary), recall bias was substantially reduced compared with paper questionnaires. The study size (n=138) was determined pragmatically according to the expected recruitment capacity within the 15-month study period and the pilot nature of this observational study. No formal sample size calculation was performed, consistent with the exploratory objectives. Patient flow from screening to analysis is summarized in Figure 1, which illustrates the number of patients invited (n=176), enrolled (n=138), and completing the 8-week follow-up, together with the reasons for exclusion.

Figure 1. Patients flowchart. GERD: gastroesophageal reflux disease.

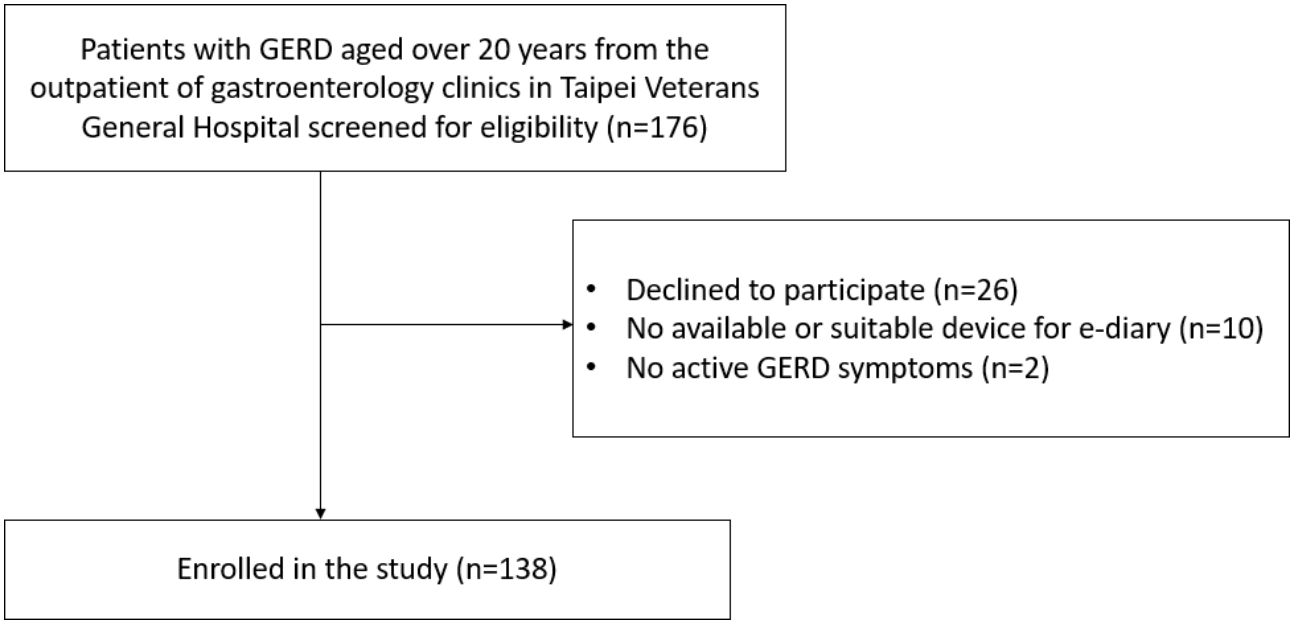

The GERD E-Diary

The web-based GERD e-diary was designed to record both nighttime and daytime GERD symptoms. Patients were asked to fill out the e-diary twice a day. In the morning, they were asked to fill in nighttime symptoms during the previous night, while in the evening, they were asked to record daytime symptoms that occurred during the concurrent day. The GERD e-diary was designed to assess 8 daytime symptoms (acid reflux, cough, heartburn, sour taste in the mouth, burping, hoarseness, dysphagia, and chest pain) and 2 nighttime symptoms (acid reflux and cough). The choice of these GERD symptom records was based on the Modified Reflux Symptom Questionnaire–Electronic Diary, which proved to be a reliable and valid PRO instrument [30]. After completing the e-diary, educational or inspiring messages would be randomly displayed to encourage continued participation. Although it was a web-based e-diary, we installed it as a smartphone shortcut icon that functioned like a mobile app, as Figure 2 showed.

XSL•FO
RenderX

Figure 2. Interface of the web-based gastroesophageal reflux disease (GERD) e-diary used in this cohort study conducted at Taipei Veterans General Hospital, Taiwan (October 2021 to January 2023). The shortcut icon was added to patients' smartphones, providing an app-like interface for direct daily input of GERD symptoms. (A) Home screen appearance; (B) symptom entry interface.

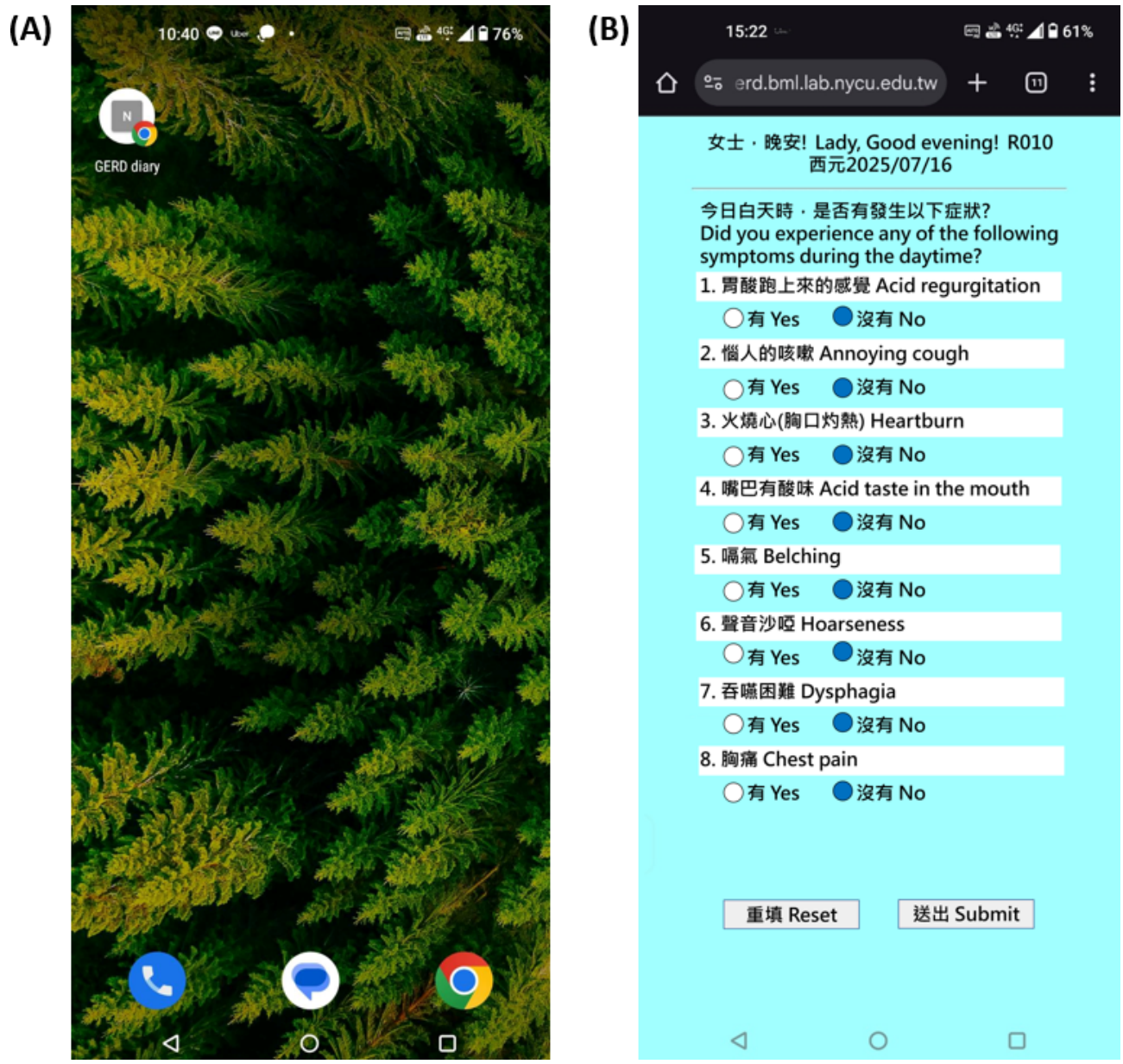

Optimization of the GERD E-Diary

Optimization of the e-diary system was developed with 3 stages. The first stage was initiated without reminders to the patients with GERD to complete the twice-daily questionnaire on mobile phones (October 7, 2021, to January 17, 2022; n=9). As most of these patients (n=5) failed to fill in the questionnaire (defined as completing <60% [range: 46.4% - 57.1%] of the overall symptom diary) in the first stage, we designed an SMS text messaging reminder in the second stage (January 18, 2022, to June 30, 2022; n=51). In this stage, the system would automatically check at noon and 10 PM each day to determine whether the patients had filled out the diary. Once an unfilled e-dairy was detected, an SMS text message was sent to the patients' mobile phones. If no entries were submitted for 3 consecutive days, an additional notification was sent to the research team to contact the patients. During the second stage, the reminder function was activated manually, resulting in no reminder function during the 3 to 5 days after enrollment. In the third stage of system optimization (July 1, 2022, to January 19, 2023; n=78), the reminder system was fully automated and activated immediately upon enrollment, eliminating the need for manual setup.

Adherence of the GERD E-Diary and Variables Affecting the Completeness of the E-Diary

The GERD e-diary adherence was evaluated by measuring the overall weekly symptom adherence rate, which was the number of days or nights symptoms filled out per week divided by 7. The overall symptom adherence rate was further categorized into nighttime (weekly nighttime adherence rate: number of nights filled out per week/7) and daytime (weekly daytime adherence rate: number of days filled out per week/7). The potential independent variables affecting the adherence rate, including the frequency of GERD symptoms (weekly total symptom days: 10 symptoms per day×7 days=70 symptom days,

XSL•FO
RenderX

ranging from 0 to 70) and the system optimization stage (3 stages, with higher stage indicating greater optimization), as well as the other confounders (such as age, gender, smoking, alcohol consumption, and comorbidities [sleep disorders, mental illnesses, sleep apnea, chronic obstructive pulmonary disease, asthma, liver or kidney diseases, cardiovascular diseases, diabetes mellitus, inflammatory bowel disease, central nervous system disorders, functional gastrointestinal disorders, and malignant diseases]) were measured.

Statistical Analysis

Descriptive statistics were performed for categorical variables as case numbers and percentages and means and SDs (including range) for continuous variables. The sampling method was clarified as convenience sampling of eligible outpatients. Any diary entry that was not completed on a given day was classified as nonadherence for that specific time point. As adherence was one of the key outcomes of interest, incomplete entries were interpreted as reflecting nonadherent behavior rather than ignorable missingness. Generalized estimating equations or multiple linear regression models were constructed to examine the effects of system optimization stage and weekly symptom frequency on adherence rates (nighttime, daytime, and overall). Independent variables included system optimization stage (categorical, 3 levels), symptom frequency (continuous), and potential confounders, such as age, gender, smoking, alcohol consumption, and major comorbidities. For repeated weekly adherence values across the 8-week observation period, a repeated-measures analysis with week as the within-subject variable was performed to assess temporal trends. All results are reported as β coefficients with corresponding 95% CIs and *P* values. Two-tailed significance was defined as $P<.05$.

Ethical Considerations

The study was reviewed and approved by the Medical Ethics Committee of Taipei Veterans General Hospital (2021-05-012CC) and conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all eligible participants before data collection, covering both the collection of primary data and its secondary analysis. Participants were informed that they could withdraw from the study at any time without any impact on their future medical care. To ensure privacy protection, all data were deidentified using participant codes, and no identifiable information was disclosed. Participants received a reimbursement of TWD $500 (approximately US $16) for their participation in the study.

Results

Clinical Characteristics of Patients With GERD

A total of 138 patients with GERD (mean [SD] age 52.9 [12.3] y; range 24.4 - 78.4 y; 68 female patients) were enrolled. Approximately one-third (46/138, 33.3%) of patients displayed coexisting functional gastrointestinal disorders. The baseline characteristics and comorbidities are summarized in Table 1.

Table . Baseline demographic and clinical characteristics of 138 patients with gastroesophageal reflux disease enrolled from the outpatient gastroenterology clinics at Taipei Veterans General Hospital, Taiwan (October 2021 to January 2023).

Clinical characteristics	Values
Age (y), mean (SD); range	52.9 (12.3); 24.4 - 78.4
Female, n (%)	68 (49.3)
Smoking, n (%)	14 (10.1)
Alcohol, n (%)	30 (21.7)
Sleep disorders, n (%)	72 (52.2)
Mental illnesses, n (%)	7 (5.1)
Sleep apnea, n (%)	17 (12.3)
Chronic obstructive pulmonary disease, n (%)	2 (1.4)
Asthma, n (%)	14 (10.1)
Liver or kidney diseases, n (%)	15 (10.9)
Cardiovascular diseases, n (%)	38 (27.5)
Diabetes mellitus, n (%)	11 (8.0)
Inflammatory bowel disease, n (%)	5 (3.6)
Central nervous system disorders, n (%)	2 (1.4)
Functional gastrointestinal disorders, n (%)	46 (33.3)
Malignant diseases, n (%)	16 (11.6)

System Optimization and Effects of Symptom Frequency on Adherence

As presented in Table 2, during the first stage of system optimization, the nighttime symptom adherence rate was 47.2%. Implementation of the reminder system significantly increased adherence by 12.5% in the second ($P=.005$) and 10.9% in the third stages ($P=.01$). For daytime symptom reporting, the adherence rate in the first stage was 40.0%, which improved by 21.7% in the second ($P<.001$) and 20.8% in the third stages

XSL•FO
RenderX

(*P*<.001). The overall adherence reporting in the first stage was 57.6%, with further increases of 16.5% in the second (*P*<.001) and 18.5% in the third stages (*P*<.001). For every additional symptom frequency, there was a significant increase of 0.9% in adherence for both daytime and nighttime symptom reporting and an increase of 0.7% for overall symptom reporting (*P*<.001). These findings indicate that higher symptom frequency was positively correlated with participants' motivation to complete the GERD e-diary, suggesting that patients experiencing more frequent symptoms were more engaged in self-monitoring.

Table . Effects of system optimization and symptom frequency on adherence to the gastroesophageal reflux disease e-diary.[a]

Adherence (%)	β[b,c] (95% CI)	*P* value
Nighttime symptom		
System optimization (first stage)	47.2	
System optimization (second vs first stage)	+12.5 (3.7 to 21.3)	0.005
System optimization (third vs first stage)	+10.9 (2.6 to 19.2)	0.01
Symptom frequency	+0.9 (0.7 to 1)	<.001
Daytime symptom		
System optimization (first stage)	40	
System optimization (second vs first stage)	+21.7 (14.2 to 29.2)	<.001
System optimization (third vs first stage)	+20.8 (13.7 to 27.9)	<.001
Symptom frequency	+0.9 (0.7 to 1)	<.001
Overall symptom		
System optimization (first stage)	57.6	
System optimization (second vs first stage)	+16.5 (9.8 to 23.2)	<.001
System optimization (third vs first stage)	+18.5 (12.2 to 4.8)	<.001
Symptom frequency	+0.7 (0.6 to 0.8)	<.001

[a]Results derived from multiple linear regression models adjusted for potential confounders, including age, gender, smoking, alcohol consumption, and comorbidities.

[b]β=change in adherence in the second and third stages of system optimization compared with the first stage.

[c]β=change in adherence per additional symptom frequency.

Adherence Trends in GERD E-Diary Recording

Table 3 presents the weekly change in adherence rate among all enrolled patients with GERD. The nighttime symptom adherence rate was 80.1% during the first week and began to decrease significantly from the third week (71.5%; *P*=.04), reaching 61.5% by the eighth week (*P*<.001). For daytime symptom reporting, adherence was 85.1% in the first week and showed a significant decline beginning at the fourth week (76.0%; *P*=.01) and further to 66.8% by the eighth week (*P*<.001). Overall adherence followed a similar trend, starting at 95.1% in the first week, decreasing to 86.7% in the fourth week (*P*=.009), and dropping to 78.0% in the eighth week (*P*<.001). These results demonstrate a gradual decline in participant engagement with the e-diary over time, despite the implemented reminder optimization strategies.

XSL•FO
RenderX

Table . Weekly adherence trends of gastroesophageal reflux disease e-diary completion during the 8-week study period.

Symptoms and week	Adherence (%)	β[a] (95% CI)	*P* value
Nighttime symptoms			
1	80.1	N/A[b]	N/A
2	72	−8.1 (−16.4 to 0.2)	0.06
3	71.5	−8.6 (−16.9 to −0.3)	0.04
4	67.9	−12.2 (−20.5 to −3.9)	0.004
5	68.4	−11.7 (−20 to −3.4)	0.006
6	61.6	−18.5 (−26.8 to −10.2)	<.001
7	64.3	−15.8 (−24.1 to −7.5)	<.001
8	61.5	−18.6 (−26.9 to −10.3)	<.001
Daytime symptoms			
1	85.1	N/A	N/A
2	82.6	−2.5 (−9.8 to 4.8)	0.50
3	80.5	−4.6 (11.8 to 2.7)	0.22
4	76	−9.1 (16.4 to 1.8)	0.014
5	72	−13 (−20.3 to -5.8)	<.001
6	69.5	−15.6 (−22.9 to -8.4)	<.001
7	68.4	−16.7 (23.9 to −9.4)	<.001
8	66.8	−18.3 (−25.6 to −11)	<.001
Overall symptoms			
1	95.1	N/A	N/A
2	89.9	−5.3 (−11.6 to 1)	0.10
3	88.9	−6.2 (−12.5 to 0.1)	0.05
4	86.7	−8.4 (−14.7 to −2.1)	0.009
5	83.4	−11.7 (−18 to −5.4)	<.001
6	78.3	−16.9 (−23.3 to −10.6)	<.001
7	79.6	−15.5 (−21.9 to −9.2)	<.001
8	78	−17.2 (−23.5 to −10.9)	<.001

[a]β=β values representing the percentage change in adherence compared with the first week, with 95% CIs and *P* values derived from repeated-measures analysis.

[b]N/A: not available.

Comparison of GERD E-Diary With Previous GERD Evaluation Tools

The characteristics and differences between our current GERD e-diary study and previous GERD questionnaire studies are summarized in Table 4. This e-diary, using separate day and night assessments, can record 10 relevant GERD symptoms daily and also provide educational information.

XSL•FO
RenderX

Table . Comparison of patient-reported outcome (PRO) instruments for gastroesophageal reflux disease (GERD). The summary includes data entry frequency, number of symptoms recorded, and whether daytime and nighttime symptoms were separately assessed.

PRO instrument[a]	GRACI[b] [31]	Puhan et al [32]	GerdQ[c] [16]	RESQ-7[d] [33]	GERD e-diary
Data entry frequency	Daily	Daily	Weekly recall	Weekly recall	Daily
Number of symptoms recorded	5	3	6	13	10
Day and night recording	No separation[e]	Separately assessed[f]	No separation	Not assessed[g]	Separately assessed
Recording format	Paper	Paper	Paper	Paper	Electronic diary
Providing educational information	N/A[h]	N/A[h]	N/A[h]	N/A[h]	Yes

[a]Summary includes data entry frequency, number of symptoms recorded, and whether daytime and nighttime symptoms were separately assessed.

[b]GRACI: Gastroesophageal Reflux Disease Activity Index.

[c]GerdQ: Gastroesophageal Reflux Disease Questionnaire.

[d]RESQ-7: Reflux Symptom Questionnaire, 7 day recall

[e]No separation: daytime and nighttime symptoms recorded together.

[f]Separately assessed: daytime and nighttime symptoms recorded separately.

[g]Not assessed: no specific daytime or nighttime symptom assessment.

[h]N/A: not available.

Discussion

Principal Findings

In this study, we demonstrated that a GERD e-diary was successfully developed to record GERD symptoms twice daily (day and night) with relatively high adherence. Patients' adherence was significantly improved with the application of SMS text messaging reminders in the e-diary system. In addition, patients who experienced more frequent symptoms tended to demonstrate higher levels of adherence. While the overall adherence rate reached approximately 80% over the 8-week period, a gradual decline was observed over time. This PRO-based e-diary, which enables prolonged and continuous symptom recording, provides a real-time and prospective assessment framework that substantially reduces the recall and ecological biases associated with conventional short-term, retrospective GERD symptom evaluations. By facilitating structured, longitudinal monitoring, the system not only enhances symptom awareness for patients but also offers clinicians high-resolution, long-term data that support more customized therapeutic decisions and advance personalized GERD management.

Comparison With Prior Work

PRO measures are essential for evaluating both disease burden and treatment efficacy in GERD [6]. Several instruments have been developed to capture these outcomes. For example, the GRACI (Gastroesophageal Reflux Disease Activity Index) integrates patient diaries with structured nurse-led interviews, thereby reducing physician workload and minimizing assessor bias [31]. Puhan et al [32] introduced a symptom diary that tracked heartburn frequency, severity, and antacid use over 4 to 6 weeks, achieving high adherence, with only 7.9% of patients completing fewer than 80% of entries. The GerdQ (Gastroesophageal Reflux Disease Questionnaire), a 6-item tool, facilitates diagnosis and management in primary care without the need for specialist referral [16]. Similarly, RESQ-7 (Reflux Symptom Questionnaire, 7-day recall) evaluates 13 GERD-related symptoms based on a 7-day recall period [33]. Despite their clinical value, most PRO tools have important limitations. They typically capture symptoms over short intervals (often 1 wk), which makes it difficult to assess temporal fluctuations in patients with intermittent symptoms. Relying on weekly recall instead of daily reporting also introduces bias, as patients tend to overestimate symptom intensity and underestimate frequency [34]. Daily PROs are generally more accurate, especially for variable symptoms [35], but even these prior studies fail to differentiate between daytime and nocturnal symptoms. This is a critical gap, as nighttime symptoms are common in GERD and strongly impact quality of life [36]. Additionally, many existing PROs record only a narrow range of symptoms, limiting their clinical comprehensiveness. Another concern lies in data collection methods. Traditional paper-based questionnaires are vulnerable to retrospective completion, which compromises accuracy. In contrast, electronic data entry systems can restrict both prospective and retrospective inputs, thereby reducing recall bias and improving data integrity [35]. To overcome these challenges, we developed a web-based GERD e-diary that records symptoms twice daily over an 8-week period. This approach provides a more reliable picture of symptom patterns, minimizes recall bias, and generates richer data for clinicians. Importantly, the e-diary also enhances patient awareness of their condition. Together, our GERD e-diary features improve the accuracy and comprehensiveness of GERD assessment, ultimately fostering greater confidence in management for both clinicians and patients.

In the first phase of building up the e-diary, the overall adherence rate was as low as approximately 40%. Therefore, we incorporated a reminder system into the e-diary and activated it upon the detection of missed entries. In the second stage, the reminder system became available only 3 days after enrollment, whereas in the third stage, it was activated immediately upon

XSL•FO
RenderX

enrollment. Both optimization measurements would significantly improve overall symptom adherence to approximately 80%. Previous research supported the effectiveness of reminders in enhancing health-related behaviors. For example, a review of 11 randomized controlled trials (1999 - 2009) confirmed that reminder interventions significantly improved daily medication adherence compared to no-reminder controls [37]. Another study demonstrated that daily SMS text messaging reminders enhanced adherence to antiasthmatic treatment [38]. Similarly, email and letter reminders significantly improved colorectal cancer screening rates compared to usual care, with no difference between the 2 reminder types (email and letter) [39]. Furthermore, reminding patients and clinicians, especially those directed at patients, is an effective strategy to improve colorectal cancer screening rates among individuals who are not up to date with screening [40]. In line with these findings, our study demonstrated that incorporating SMS text messaging reminders substantially improved adherence to GERD e-diary recordings. Despite the favorable results, there seemed to be little difference between the second and third stages of system optimization, which might be due to the short time delay (3 - 5 d) of the incorporation of the reminding systems.

We also found that higher symptom frequency was associated with increased adherence. Similar observations had been reported in other diseases. For instance, patients with urinary incontinence who experienced greater voiding frequency exhibited higher adherence to voiding diaries [27]. Furthermore, in patients with seasonal allergic rhinitis, symptom recording adherence to an e-diary was significantly higher during the peak grass pollen season, which coincided with more intense allergic symptoms [28]. All observations suggested that more frequent and severe symptoms may be associated with enhanced awareness and stronger motivation to report their disease status to health care providers.

Despite relatively high adherence at the beginning phase, our study showed a gradual decline in adherence over the 8-week period. This trend was consistent with previous findings in other diseases. For example, patients with allergic rhinitis demonstrated a slow decline in e-diary completion from 90% in the first week, 80% to 90% in the second to sixth weeks, and 70% to 80% after the seventh week [28]. A separate study assessing voiding diaries among patients with urinary incontinence similarly reported that a 7-day diary posed a higher burden than shorter 2- or 3-day formats [41]. In addition, 3 respiratory clinical trials also demonstrated that adherence decreased over time after randomization [42]. All results suggested that adherence to long-term daily symptom recording might be challenging for the patients to complete the daily diary study.

To improve long-term adherence, several strategies might be considered. Providing additional information and education could be applied to promote better adherence in e-diary recording [28]. Reducing diary duration may also enhance patient compliance and minimize burden [41]. Additional methods to increase patient engagement, such as customizing diary content and reminders based on patient needs, using user-friendly interfaces, and incorporating social and gamification features [21], or integration with wearable devices, might also help improve adherence. Using personalized adaptive reinforcement learning as a core behavioral strategy, it may be possible to optimize intervention timing and minimize alert fatigue, ensuring more stable long-term user participation in the future [43]. In this study, patients expressed a desire to see immediate visual representations of daily symptom changes. Understanding the psychological factors (eg, motivation) and socioeconomic status behind patient adherence could be invaluable for designing more effective interventions.

Study Strengths

Our study introduces a significant innovation in GERD management through the development of a web-based e-diary designed for daily symptom assessment. The primary strength of this approach lies in its ability to provide real-time, prospective data collection, thereby fundamentally overcoming the limitations of traditional paper-based questionnaires and retrospective recall, which are prone to significant recall and ecological biases for patients with GERD [44]. Unlike conventional studies that often rely on symptom scores primarily as a measure of drug response, our e-diary integrates symptom monitoring directly into daily clinical practice. This shift allows for the capture of fine-grained, fluctuating symptom patterns over an extended period, providing a more accurate and comprehensive understanding of the patient's condition. Second, by minimizing recall bias, our e-diary ensures higher data entry quality and more efficient data handling compared to traditional methods. Third, it empowers patients to actively participate in their self-management by collecting health data autonomously, fostering self-reliance and improved awareness of their condition [45]. Fourth, with the application of an SMS reminder system and educational feedback provided, the adherence rate of patients could be maintained up to 8 weeks, especially for those symptomatic patients. Fifth, our study benefits from the e-diary's ability to separately capture daytime and nighttime GERD symptoms. This distinction is clinically important, as nocturnal reflux is characterized by impaired esophageal clearance and is associated with more aggressive disease, including a higher risk of severe complications such as esophagitis [46]. Moreover, nighttime symptoms substantially impair patients' health-related quality of life [36]. These granular, time-specific symptom data provide essential insights for developing personalized management strategies. Finally, the detailed, real-time symptom data facilitate enhanced patient-physician collaboration. Physicians can readily visualize symptom changes, understand the variability in symptoms, and work more effectively with patients to tailor treatment strategies [47]. This paves the way for truly personalized treatment plans in the future that can adapt to individual patient needs and daily life events, potentially incorporating dietary or other therapeutic adjustments. Ultimately, this approach moves beyond simple response assessment, enabling a proactive feedback loop that can lead to more informed and timely clinical decisions, fostering greater confidence in management for both clinicians and patients.

Study Limitations

Limitations do exist in the study. First, patients with GERD were enrolled from a single center and included only patients capable of operating an e-diary, potentially limiting the

XSL•FO
RenderX

generalizability and overestimating adherence. Validation in multicenter or community-based cohorts should be conducted in the further studies. Second, days on which patients did not complete entries were recorded as symptom-free when calculating symptom frequency, which may lead to underestimation of the reported symptoms. Third, objective physiological parameters (eg, 24 h pH-impedance monitoring, reflux-symptom association probability, or treatment response) were not detected in this study, which would enhance the e-diary's clinical utility. Fourth, although SMS reminders improved the adherence, they might inadvertently increase psychological stress, symptom vigilance, or anxiety, which also may lead to worsening of GERD symptoms. This may account for part of the nonadherence observed in this study. Assessment of the Esophageal Hypervigilance and Anxiety Scale and other psychological measurements is warranted in subsequent studies to help balance engagement benefits against potential harm. Finally, we did not evaluate the satisfaction scores of patients with GERD and caring physicians. Further adjustments to this GERD e-diary by correcting the aforementioned limitations should be considered in future studies.

Conclusions

Our results demonstrate that system optimization can significantly enhance adherence in the newly developed GERD e-diary recording. Increased GERD symptom frequency was associated with adherence, although overall engagement declined gradually over 8 weeks. The development of this PRO-based GERD e-diary system can be a convenient tool for future application in clinical and research settings.

Acknowledgments

The authors acknowledge the Big Data of Taipei Veterans General Hospital and the Biostatistics Task Force of Taipei Veterans General Hospital for their assistance during this study.

The authors declare the use of generative AI in the writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: summarizing text and translation. The GAI tool used was ChatGPT 4. Responsibility for the final manuscript lies entirely with the authors. All AI-generated text was reviewed or revised by the authors before submission. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was partially supported by Taiwan National Science and Technology Council (NSTC 113-2634-F-A49-003). The funding agency was not involved in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: YCC, YPW, LFC, CLL, MLP, YHP, DWW, SLW
Data curation: YCC, SLW
Formal analysis: SLW
Funding acquisition: LFC, CLL, MLP, DWW
Investigation: YPW, CLL, YCC
Methodology: MLP, JHH
Project administration: CLL, MLP, YCC
Resources: YPW, CLL, MLP, LFC, YHP, DWW
Software: MLP, JHH, LFC
Supervision: CLL, LFC, YHP, MLP, DWW, SLW
Validation: YCC, CLL, YPW
Visualization Writing – original draft: YCC, YPW, CLL
Writing – review & editing: YCC, YPW, CLL, MLP, YHP, LFC, SLW
MLP and CLL are co-corresponding authors.

Conflicts of Interest

None declared.

Checklist 1
STROBE checklist.
[PDF File, 92 KB - jmir_v28i1e83680_app1.pdf]

References

XSL•FO
RenderX

1. Vakil N, van Zanten SV, Kahrilas P, Dent J, Jones R, Global Consensus Group. The Montreal definition and classification of gastroesophageal reflux disease: a global evidence-based consensus. Am J Gastroenterol 2006 Aug;101(8):1900-1920. [doi: 10.1111/j.1572-0241.2006.00630.x] [Medline: 16928254]
2. Blondeau K, Sifrim D, Dupont L, Tack J. Reflux cough. Curr Gastroenterol Rep 2008 Jun;10(3):235-239. [doi: 10.1007/s11894-008-0049-0] [Medline: 18625132]
3. Kellerman R, Kintanar T. Gastroesophageal reflux disease. Prim Care 2017 Dec;44(4):561-573. [doi: 10.1016/j.pop.2017.07.001] [Medline: 29132520]
4. Mahajan R, Kulkarni R, Stoopler ET. Gastroesophageal reflux disease and oral health: a narrative review. Spec Care Dentist 2022 Nov;42(6):555-564. [doi: 10.1111/scd.12726] [Medline: 35605234]
5. de Hoyos A, Esparza EA, Cervantes-Sodi M. Non-erosive reflux disease manifested exclusively by protracted hiccups. J Neurogastroenterol Motil 2010 Oct;16(4):424-427. [doi: 10.5056/jnm.2010.16.4.424] [Medline: 21103425]
6. Lu CL, Lang HC, Chang FY, et al. Social and medical impact, sleep quality and the pharmaceutical costs of heartburn in Taiwan. Aliment Pharmacol Ther 2005 Oct 15;22(8):739-747. [doi: 10.1111/j.1365-2036.2005.02664.x] [Medline: 16197495]
7. Delshad SD, Almario CV, Chey WD, Spiegel BMR. Prevalence of gastroesophageal reflux disease and proton pump inhibitor-refractory symptoms. Gastroenterology 2020 Apr;158(5):1250-1261. [doi: 10.1053/j.gastro.2019.12.014] [Medline: 31866243]
8. Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims. : U.S. Department of Health and Human Services; 2009 URL: https://www.fda.gov/media/77832/download [accessed 2025-12-19]
9. Rasheed T, Zuberi BF, Ali FS, Hussain SM, Kumar P, Saleem A. Comparison of efficacy of daily and alternate day maintenance treatment of GERD with Vonoprazan 10-mg using Gastroesophageal Reflux Disease Symptom Assessment Scale. Pak J Med Sci 2024;40(4):623-628. [doi: 10.12669/pjms.40.4.8063] [Medline: 38545007]
10. Miyamoto M, Manabe N, Haruma K. Efficacy of the addition of prokinetics for proton pump inhibitor (PPI) resistant non-erosive reflux disease (NERD) patients: significance of frequency scale for the symptom of GERD (FSSG) on decision of treatment strategy. Intern Med 2010;49(15):1469-1476. [doi: 10.2169/internalmedicine.49.3615] [Medline: 20686276]
11. Shaw M, Dent J, Beebe T, et al. The Reflux Disease Questionnaire: a measure for assessment of treatment response in clinical trials. Health Qual Life Outcomes 2008 Apr 30;6:31. [doi: 10.1186/1477-7525-6-31] [Medline: 18447946]
12. Wu H, Ungerleider S, Campbell M, et al. Patient-reported outcomes in 645 patients after laparoscopic fundoplication up to 10 years. Surgery 2023 Mar;173(3):710-717. [doi: 10.1016/j.surg.2022.07.039] [Medline: 36307333]
13. Galvarini M, Angeramo CA, Kerman J, et al. Outcomes of endoscopic antireflux mucosectomy for the treatment of gastroesophageal reflux disease: systematic review and meta-analysis. J Clin Gastroenterol 2024 Oct 1;58(9):851-856. [doi: 10.1097/MCG.0000000000002061] [Medline: 39145822]
14. Fuchs KH, Musial F, Eypasch E, Meining A. Gastrointestinal quality of life in gastroesophageal reflux disease: a systematic review. Digestion 2022;103(4):253-260. [doi: 10.1159/000524766] [Medline: 35605592]
15. Flynn KE, Mansfield SA, Smith AR, et al. Patient demographic and psychosocial characteristics associated with 30-day recall of self-reported lower urinary tract symptoms. Neurourol Urodyn 2020 Sep;39(7):1939-1948. [doi: 10.1002/nau.24461] [Medline: 32856723]
16. Jones R, Junghard O, Dent J, et al. Development of the GerdQ, a tool for the diagnosis and management of gastro-oesophageal reflux disease in primary care. Aliment Pharmacol Ther 2009 Nov 15;30(10):1030-1038. [doi: 10.1111/j.1365-2036.2009.04142.x] [Medline: 19737151]
17. Vakil N, Niklasson A, Denison H, Rydén A. Gender differences in symptoms in partial responders to proton pump inhibitors for gastro-oesophageal reflux disease. United European Gastroenterol J 2015 Oct;3(5):443-452. [doi: 10.1177/2050640614558343] [Medline: 26535123]
18. Shaw MJ, Talley NJ, Beebe TJ, et al. Initial validation of a diagnostic questionnaire for gastroesophageal reflux disease. Am J Gastroenterol 2001 Jan;96(1):52-57. [doi: 10.1111/j.1572-0241.2001.03451.x] [Medline: 11197287]
19. Wiest IC, Sicorello M, Yesmembetov K, Ebert MP, Teufel A. Usage behaviour and adoption criteria for mobile health solutions in patients with chronic diseases in gastroenterology. Visc Med 2024 Apr;40(2):61-74. [doi: 10.1159/000534191] [Medline: 38584857]
20. Giebel GD, Speckemeier C, Abels C, et al. Problems and barriers related to the use of digital health applications: scoping review. J Med Internet Res 2023 May 12;25:e43808. [doi: 10.2196/43808] [Medline: 37171838]
21. Jakob R, Harperink S, Rudolf AM, et al. Factors influencing adherence to mhealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371. [doi: 10.2196/35371] [Medline: 35612886]
22. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020 Sep 29;22(9):e20283. [doi: 10.2196/20283] [Medline: 32990635]
23. Daniëls NEM, Hochstenbach LMJ, van Zelst C, van Bokhoven MA, Delespaul P, Beurskens A. Factors that influence the use of electronic diaries in health care: scoping review. JMIR mHealth uHealth 2021 Jun 1;9(6):e19536. [doi: 10.2196/19536] [Medline: 34061036]

XSL•FO
RenderX

24. Messner EM, Sturm N, Terhorst Y, et al. Mobile apps for the management of gastrointestinal diseases: systematic search and evaluation within app stores. J Med Internet Res 2022 Oct 5;24(10):e37497. [doi: 10.2196/37497] [Medline: 36197717]
25. Gould MJ, Lin C, Walsh CM. A systematic assessment of the quality of smartphone applications for gastroesophageal reflux disease. Gastro Hep Adv 2023;2(5):733-742. [doi: 10.1016/j.gastha.2023.03.001] [Medline: 39129878]
26. Venugopal LS, Musbahi A, Shanmugam V, Gopinath B. A systematic review of smartphone apps for gastro-oesophageal reflux disease: the need for regulation and medical professional involvement. mHealth 2021;7:56. [doi: 10.21037/mhealth-20-126] [Medline: 34805387]
27. Pauls RN, Hanson E, Crisp CC. Voiding diaries: adherence in the clinical setting. Int Urogynecol J 2015 Jan;26(1):91-97. [doi: 10.1007/s00192-014-2470-2] [Medline: 25124091]
28. Di Fraia M, Tripodi S, Arasi S, et al. Adherence to prescribed e-diary recording by patients with seasonal allergic rhinitis: observational study. J Med Internet Res 2020 Mar 16;22(3):e16642. [doi: 10.2196/16642] [Medline: 32175909]
29. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med 2007 Oct 16;4(10):e296. [doi: 10.1371/journal.pmed.0040296] [Medline: 17941714]
30. Andrae DA, Hanlon J, Cala ML, et al. Evaluation and validation of the modified reflux symptom questionnaire-electronic diary in patients with persistent gastroesophageal reflux disease. Clin Transl Gastroenterol 2020 Jan;11(1):e00117. [doi: 10.14309/ctg.0000000000000117] [Medline: 31977454]
31. Williford WO, Krol WF, Spechler SJ. Development for and results of the use of a gastroesophageal reflux disease activity index as an outcome variable in a clinical trial. Control Clin Trials 1994 Oct;15(5):335-348. [doi: 10.1016/0197-2456(94)90031-0] [Medline: 8001355]
32. Puhan MA, Guyatt GH, Armstrong D, et al. Validation of a symptom diary for patients with gastro-oesophageal reflux disease. Aliment Pharmacol Ther 2006 Feb 15;23(4):531-541. [doi: 10.1111/j.1365-2036.2006.02775.x] [Medline: 16441474]
33. Rydén A, Denison H, Karlsson M, Vakil N. Development and validation of a patient-reported outcome instrument in partial responders to proton pump inhibitors. Scand J Gastroenterol 2013 Sep;48(9):1018-1026. [doi: 10.3109/00365521.2013.822544] [Medline: 23919738]
34. Rydén A, Leavy OC, Halling K, Stone AA. Comparison of daily versus weekly recording of gastroesophageal reflux disease symptoms in patients with a partial response to proton pump inhibitor therapy. Value Health 2016;19(6):829-833. [doi: 10.1016/j.jval.2016.05.007] [Medline: 27712711]
35. Vakil N, Björck K, Denison H, et al. Validation of the reflux symptom questionnaire electronic diary in partial responders to proton pump inhibitor therapy. Clin Transl Gastroenterol 2012 Jan 26;3(1):e7. [doi: 10.1038/ctg.2012.1] [Medline: 23238029]
36. Farup C, Kleinman L, Sloan S, et al. The impact of nocturnal symptoms associated with gastroesophageal reflux disease on health-related quality of life. Arch Intern Med 2001 Jan 8;161(1):45-52. [doi: 10.1001/archinte.161.1.45] [Medline: 11146697]
37. Fenerty SD, West C, Davis SA, Kaplan SG, Feldman SR. The effect of reminder systems on patients' adherence to treatment. Patient Prefer Adherence 2012;6:127-135. [doi: 10.2147/PPA.S26314] [Medline: 22379363]
38. Strandbygaard U, Thomsen SF, Backer V. A daily SMS reminder increases adherence to asthma treatment: a three-month follow-up study. Respir Med 2010 Feb;104(2):166-171. [doi: 10.1016/j.rmed.2009.10.003] [Medline: 19854632]
39. Muller D, Logan J, Dorr D, Mosen D. The effectiveness of a secure email reminder system for colorectal cancer screening. AMIA Annu Symp Proc 2009 Nov 14;2009:457-461. [Medline: 20351899]
40. Ahmed AM, Bacchus MW, Beal SG, et al. Colorectal cancer screening completion by patients due or overdue for screening after reminders: a retrospective study. BMC Cancer 2023 May 1;23(1):391. [doi: 10.1186/s12885-023-10837-y] [Medline: 37127588]
41. Ku JH, Jeong IG, Lim DJ, Byun SS, Paick JS, Oh SJ. Voiding diary for the evaluation of urinary incontinence and lower urinary tract symptoms: prospective assessment of patient compliance and burden. Neurourol Urodyn 2004;23(4):331-335. [doi: 10.1002/nau.20027] [Medline: 15227650]
42. Nowojewski A, Bark E, Shih VH, Dearden R. Patient adherence and response time in electronic patient-reported outcomes: insights from three longitudinal clinical trials. Qual Life Res 2024 Jun;33(6):1691-1706. [doi: 10.1007/s11136-024-03644-w] [Medline: 38598132]
43. Lauffenburger JC, Yom-Tov E, Keller PA, et al. The impact of using reinforcement learning to personalize communication on medication adherence: findings from the REINFORCE trial. NPJ Digit Med 2024 Feb 19;7(1):39. [doi: 10.1038/s41746-024-01028-5] [Medline: 38374424]
44. Beckers AB, Snijkers JTW, Weerts Z, et al. Digital instruments for reporting of gastrointestinal symptoms in clinical trials: comparison of end-of-day diaries versus the experience sampling method. JMIR Form Res 2021 Nov 24;5(11):e31678. [doi: 10.2196/31678] [Medline: 34821561]
45. Van Woensel W, Roy PC, Abidi SR, Abidi SSR. A mobile and intelligent patient diary for chronic disease self-management. Stud Health Technol Inform 2015;216:118-122. [Medline: 26262022]

XSL•FO
RenderX

46. Huang Y, Liu J, Xu L, et al. Exacerbation of symptoms, nocturnal acid reflux, and impaired autonomic function are associated with sleep disturbance in gastroesophageal reflux disease patients. Front Med (Lausanne) 2024;11:1438698. [doi: 10.3389/fmed.2024.1438698] [Medline: 39234038]
47. Staehelin D, Dolata M, Stöckli L, Schwabe G. How patient-generated data enhance patient-provider communication in chronic care: field study in design science research. JMIR Med Inform 2024 Sep 10;12:e57406. [doi: 10.2196/57406] [Medline: 39255481]

Abbreviations

e-diary: electronic diary
EQUATOR: Enhancing the Quality and Transparency of Health Research
GERD: gastroesophageal reflux disease
PRO: patient-reported outcome
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 29.Sep.2025; peer-reviewed by MW Wong, Y Gao; accepted 24.Nov.2025; published 06.Jan.2026.

Please cite as:
Chen YC, Wang YP, Hung JH, Wang DW, Wu SL, Chen LF, Ping YH, Pan ML, Lu CL
Establishment and Optimization of a Patient-Reported Outcome–Based Electronic-Diary for Symptoms Evaluation in Patients With Gastroesophageal Reflux Disorder: Prospective Cohort Study
J Med Internet Res 2026;28:e83680
URL: https://www.jmir.org/2026/1/e83680
doi:10.2196/83680

© Yun-Chun Chen, Yen-Po Wang, Jui-Hsuan Hung, Da-Wei Wang, Shang-Liang Wu, Li-Fen Chen, Yueh-Hsin Ping, Mei-Lien Pan, Ching-Liang Lu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quality of Life Trajectories With Integration Into Electronic Health Records for High-Resolution Patient Outcomes: Algorithm Development and Validation Study

Martin Dugas[1], MSc, MD; Robin Fleige[1], MSc; Max Christian Blumenstock[1], MSc; Stephan Christoph Feder[1,2], MD; Tobias Dittrich[1,3], MD; Niels Siegel[4], MD; Celine Fabienne Bergmann[1]; Luis Wettach[1], MSc; Sandra Sauer[3], MD; Pavlina Lenga[1,5], MD; Sandro M Krieg[5], MD, MBA; Susanne Dugas-Breit[6], MD; Lucy Joanne Kessler[7], MD; Kosima Kosmalla[1,8], MD; Fleur Fritz-Kebede[1], PhD; Martin Loos[4], MD; Hans-Christoph Friederich[2], MD; Thomas M Pausch[1,4], MD; Matthias Ganzinger[1], PhD

[1]Institute of Medical Informatics, Heidelberg University Hospital, Heidelberg, Germany

[2]Department of General Internal Medicine, Psychosomatics and Psychotherapy, Heidelberg University Hospital, Heidelberg, Germany

[3]Department of Hematology, Oncology and Rheumatology, Heidelberg University Hospital, Heidelberg, Germany

[4]Department of General, Visceral and Transplantation Surgery, Heidelberg University Hospital, Heidelberg, Germany

[5]Department of Neurosurgery, Heidelberg University Hospital, Heidelberg, Germany

[6]National Center for Tumor Diseases, Heidelberg University Hospital, Heidelberg, Germany

[7]Department of Ophthalmology, Heidelberg University Hospital, Heidelberg, Germany

[8]Department of Gastroenterology, Heidelberg University Hospital, Heidelberg, Germany

Corresponding Author:
Martin Dugas, MSc, MD
Institute of Medical Informatics
Heidelberg University Hospital
Im Neuenheimer Feld 130.3
Heidelberg, D-69120
Germany
Phone: 49 62215622000
Email: martin.dugas@med.uni-heidelberg.de

Abstract

Background: Patient-reported outcome measures (PROMs) such as health-related quality of life (HRQoL) are usually assessed at greater time intervals such as diagnostic time points, after treatment, and during follow-up. Many PROMs require frequent data collection (weekly or daily). Electronic PROMs enable high-resolution tracking but face declining response rates. Integrating PROMs into electronic health records (EHRs) could improve response rates and personalize therapy.

Objective: This study aimed to evaluate the technical and clinical feasibility of high-frequency HRQoL assessments for routine care in EHRs.

Methods: Patients receive emails on their mobile devices with 1-time links to a web-based app called MyEDC. This app communicates with an electronic data capture proxy in the demilitarized zone of the hospital. With a polling mechanism, these patient data are transferred to the protected hospital network and uploaded to the EHR system. HRQoL on a visual analog scale is assessed over the course of treatment in 4 clinical use cases: psychosomatics, hematology, visceral surgery, and neurosurgery.

Results: Quality of life (QoL) trajectories were collected for 110 patients with daily or weekly data collection between 2 weeks and 3 months. The HRQoL analyses revealed clinically relevant findings across the 4 different medical domains. In the use case psychosomatics, 36 patients showed a significant increase in HRQoL following 4 weeks of therapy, rising from a median of 42% (IQR 32%-52%) to 60% (IQR 41%-67%; *P*=.01). An analysis of 25 patients in hematology demonstrated a significant correlation between HRQoL and 30-item QoL Questionnaire (EORTC QLQ-C30) global health status score (*P*=.02). For 26 patients in visceral surgery, a significant association was observed between HRQoL and the reported pain level (*P*<.001). The clinical feasibility was further highlighted in the neurosurgery use case, where 23 patients showed a median response time to the electronic PROM questionnaires of 5.3 (IQR 0.6-17.7) hours. HRQoL values were associated with disease-specific symptoms and scores, indicating clinical validity of this readout. Considerable variability of HRQoL was observed over time, both intraindividually

XSL•FO
RenderX

and interindividually. Median area under the curve of HRQoL ranged from 0.46 to 0.79. Median time to answer ranged from 0.9 to 7.1 hours. No significant association between number of responses and age was observed.

Conclusions: High-resolution QoL trajectories with EHR integration are technically and clinically feasible. They offer a novel readout beyond survival analysis or PROM end point, enabling precise disease characterization and treatment comparison.

(J Med Internet Res 2026;28:e79834) doi:10.2196/79834

KEYWORDS

quality of life; electronic health records; patient-reported outcome measures; mobile devices; feasibility study

Introduction

At present, patient-reported outcome measures (PROMs) are typically assessed at the time of diagnosis, after initial treatment and every 3 to 6 months during follow-up [1]. However, these time intervals for PROM assessment and the length of the follow-up period are often based on clinical feasibility rather than standardized guidelines. Many established PROMs address a time interval of 1 to 2 weeks or less, for example, the 5-item World Health Organization (WHO-5) well-being index [2], 9-item Patient Health Questionnaire [3], and patient-reported outcomes-common terminology criteria for adverse events (PRO-CTCAE) [4]. Health-related quality of life (HRQoL) is a typical example of a PROM and can be assessed with instruments such as the EQ-5D [5], which addresses today's status. Therefore, and for clinical necessity, complete PROM trajectories require more frequent data collection, depending on the instrument every 2 weeks, weekly, or even daily. Electronic collection of PROMs provides new opportunities for such frequent assessments. The time course of PROMs can be determined with this technology; for example, a recovery trajectory for cardiac surgery patients was described in the study by Mori et al [6]. However, response rate (RR) is a major issue: a review of RRs over time [7] reported an average RR for electronic PROMs of 42% at baseline, decreasing to 36% after 1 year. It is assumed that integration of digital health tools into routine care can contribute to long-term engagement and higher retention rates [8]. Current methods often rely on stand-alone apps that store PROMs in external databases, creating data silos that impede real-time access and clinician use. Integrating PROMs directly into the electronic health records (EHRs) overcomes these limitations, offering physicians a novel, comprehensive clinical readout for better-informed decisions and potentially fostering an improved patient-physician relationship.

Our hypothesis is that the integration of HRQoL from mobile devices into EHRs can improve RRs and enable the collection of HRQoL trajectories with high resolution, that is, weekly or even daily. Patients will show a higher commitment to provide HRQoL data continuously because these are sent to the medical team and thereby have a direct effect on therapy for each individual patient. From an informatics perspective, this is challenging because such data need to be collected and transferred from a large variety of mobile devices on the internet into the protected network of an EHR system. The objective of this work was to assess the technical and clinical feasibility of high-resolution HRQoL assessments, which can visualize the course of a disease from a patient perspective and are integrated with EHRs.

Methods

IT Architecture and Information Security

To capture QoL trajectories (QoL-T) with high resolution, that is, biweekly, weekly, or even daily, each patient used their own mobile device for data collection. The alternative would be a preinstalled device provided by the health care organization, which is associated with significant cost and logistical effort and therefore cannot provide a scalable system. Patient devices were very heterogeneous: smartphone or tablet with iOS or Android, laptop or PC, and different operating systems with different software versions. Therefore, a web-based app (HTML5) was developed for data collection, called MyEDC (my electronic data capture), which is based on OpenEDC [9,10]. This approach avoids software installations by the patient.

Figure 1 presents the high-level IT architecture for data transfer from the patient device to the EHR. This architecture was reviewed and approved by the information security officer of Heidelberg University Hospital (UKHD). The web app was provided by an electronic data capture (EDC) proxy in the demilitarized zone, which is protected by 3 firewalls and a reverse proxy. In the internal network, a job service (programmed in Java with Spring framework) sent emails with 1-time EDC links to the patient. The job service polled data from outside to the internal network but only for registered and valid data collections. Other data packages were ignored and deleted. Valid data were transferred to an internal EDC database, which converted data into EHR format and performed an automated upload to the EHR database. A Clinical Data Interchange Standards Consortium (CDISC) operational data model [11]–compliant EDC system was used to comply with data standards in clinical trials. Owing to technical limitations of the local EHR system (i.s.h.med from Cerner Inc), files in PDF with informative document titles were generated and uploaded into the EHR.

XSL•FO
RenderX

Figure 1. IT architecture for data transfer from patient device through hospital firewalls into internal electronic health record system. DMZ: demilitarized zone; EDC: electronic data capture; EHR: electronic health record; PC: personal computer.

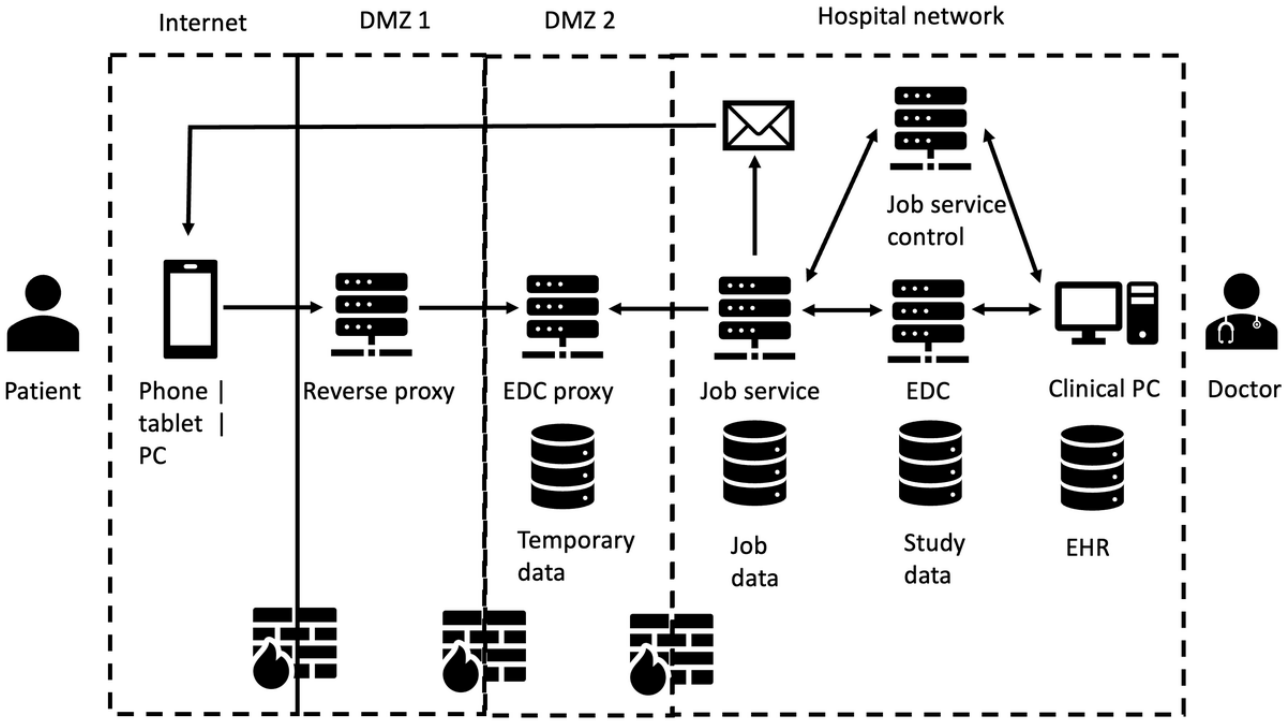

Short Questionnaires for High-Frequency PROM Collection

Overall, HRQoL was assessed frequently using a visual analog scale (VAS) with a range from 1 to 100 (Figure 2), similar to HRQoL instruments such as EQ-5D. An HRQoL value of 0 was reserved for deceased patients. QoL-T, that is, a time series of HRQoL values, were plotted and summarized with a normalized area under the curve (AUC) per patient. Few (mostly <5) additional, disease-specific PROM items were added for each clinical use case, for example, pain in surgical patients. Simple language was used to support a wide variety of use cases, including severely ill patients and participants with limited language skills. These questionnaires were designed by clinical experts to assess patient trajectories regarding effectiveness and side effects of routine therapy.

Figure 2. Overall health-related quality of life with a 100-level visual analog scale. General questions: How do You feel? (1=very bad, 100=very good).

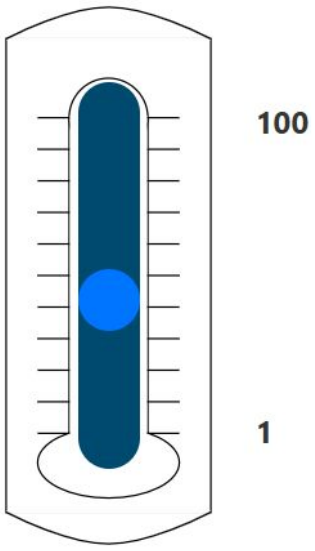

Statistical Analysis

Statistical analyses were performed using R (version 4.4.0). Differences between paired observations were assessed using the Wilcoxon signed-rank test. Associations between continuous or ordinal variables were investigated using the Spearman rank correlation coefficient. For all tests, a 2-sided *P* value threshold of .05 was considered statistically significant. AUC was calculated with the trapezoidal rule using the mean of adjacent points, normalized by total time, and scaled to a value range from 0 to 1. A 95% confidence band for average QoL was calculated as a point-wise 95% CI of the mean. The participation

XSL•FO
RenderX

rate was defined as the percentage of eligible patients who consented to participate in the study out of the total number of individuals invited. To assess longitudinal data quality, we calculated the RR at each assessment time point, defined as the percentage of participating patients who completed the questionnaire at that specific interval. Time to answer was calculated for completed questionnaires. Clinical feasibility for QoL-T was assessed regarding completeness with the median RR across all time points. Timeliness of patient answers to short PROM questionnaires was described with median time to answer.

Ethical Considerations

This study was conducted in accordance with the ethical guidelines of the Declaration of Helsinki. The research was approved by the ethics committee of the Medical Faculty of Heidelberg University (S-607/2023 and S-255/2025). The privacy and confidentiality of participants were maintained by ensuring that all study data were deidentified. Informed consent was obtained from all participants; this included the explicit right to opt out or withdraw from the study at any time. No direct compensation was provided for participation.

Clinical Use Cases

Overview

Four different use cases were selected to assess the clinical feasibility of QoL-T with EHR integration in psychosocial and somatic medicine. PROMs are clinically relevant across these distinct disciplines because they capture the patient's subjective experience of disease burden, treatment efficacy, and functional recovery. PROMs can provide insights into mental well-being (psychosomatics), treatment toxicity and fatigue (hematology), postsurgical functional status (visceral surgery), and pain (neurosurgery). Overall, 110 out of 184 (59.7%) patients agreed to participate.

Psychosomatics

As a typical example for psychosocial medicine, 36 inpatients with anorexia nervosa (Department of General Internal Medicine and Psychosomatics, UKHD) used the system with their own devices (predominantly smartphones) between August 2024 and April 2025. Three additional patients did not participate because they did not have a suitable device or did not want to use it for this purpose. Data were collected once a week for at least a month per patient. In addition to HRQoL-VAS, further widely used psychometric instruments to assess depression and anxiety (4-item Patient Health Questionnaire, PHQ-4 [12]), somatic symptoms (12-item Somatic System Disorder, SSD-12 [13]), short symptom checklist (Symptom-Checkliste Kurzversion, SCLK-11 [14]), work ability (Work Ability Index [15]), reflective functioning (Reflective Functioning Questionnaire, RFQ [16]), and 8-item Eating Disorder Examination-Questionnaire (EDE-Q8) [17] were applied. EDE-Q8 is an established clinical score to track symptoms of anorexia nervosa.

Hematology

A total of 25 outpatients with myeloma (Department of Hematology, Oncology, and Rheumatology, UKHD) used the system with their own devices between September 2024 and March 2025. In addition, 9 patients did not participate (technical reasons: 6, overwhelmed by medical situation: 2, and other: 1). Data were collected once a day for a period of 2 weeks during chemotherapy. In addition to HRQoL-VAS, side effects were collected according to PRO-CTCAE [18].

Visceral Surgery

In total, 26 inpatients (Department of General, Visceral, and Transplantation Surgery, UKHD) used the system with their own devices between January 2025 and April 2025; 59 screened patients did not participate (technical reasons: 15, overwhelmed by medical situation: 7, data protection concerns: 12, language barrier: 10, and other: 15). Patients with pancreatic diseases requiring surgical therapy were recruited for data collection at admission the day before surgery. Data were collected once a day for a period of 2 weeks at baseline before and immediately after the surgical procedure. In addition to HRQoL-VAS, the following items were collected: postoperative pain, nausea, feverish feeling, bowel movements, and the ability to drink or eat.

Neurosurgery

A total of 23 inpatients with spinal stenosis or herniated disc (Department of Neurosurgery, UKHD) used the system on their own devices between March 2025 and April 2025. Data were collected once a day for a period of 2 weeks immediately after the surgical procedure. In addition to HRQoL-VAS, the following items were collected: pain at the surgical site, leg pain, arm pain, and dysesthesia. Additionally, 9 patients did not use MyEDC (technical assistance needed: 6, language barrier: 1, overwhelmed by medical situation: 1, and other: 1).

Results

Patients entered HRQoL data via 1-time EDC links as described in the Methods section. The first question addressed overall QoL with a 100-level VAS. Different variants of this VAS were used. In visceral surgery, a vertical VAS was used (Figure 2); in psychosomatics and neurosurgery, a horizontal VAS was applied; and in hematology, a vertical VAS with an inverted scale (distress 1-100) was provided. In addition to the VAS, a set of disease-specific questions (such as pain for surgical patients) were asked. Data transfer from outside to the EHR system was implemented using a polling mechanism. On the basis of these data, EHR documents were generated automatically as a manual document list view (Figure S1 in Multimedia Appendix 1) and a detailed document view (Figure S2 in Multimedia Appendix 1), similar to a laboratory report. In a 9-month period (August 2024-April 2025), a total of 110 patients from UKHD participated in 4 clinical use cases.

Figure 3 presents QoL-T for 36 patients from psychosomatic medicine (32 female and 4 male patients; median age 27, IQR 18-58 years), who received weekly emails to collect HRQoL values. In the beginning, there was a wide range of HRQoL values. Patients provided HRQoL data with a relatively high level of completeness over the treatment period (median RR 92%, IQR 86%-97%in the first month). The participation rate was 92% (33/36; ie, highest of all 4 use cases). On average,

XSL•FO
RenderX

HRQoL values improved over time, indicating a positive effect of therapy (median HRQoL at baseline 42%, IQR 32%-52%, median HRQoL after 4 weeks 60%, IQR 41%-67%; *P*=.01). There is a negative correlation (significance level was not reached) between HRQoL and EDE-Q8 (Spearman correlation=–0.123, *P*=.06; Figure S3 in Multimedia Appendix 1), which indicates that decreased QoL is associated with higher levels of eating disorder symptoms.

Figure 3. (A) Health-related quality of life trajectories for 36 patients from psychosomatic medicine. Each line corresponds to a patient. A 95% confidence band is provided. (B) A histogram of the quality of life per patient (standardized area under the curve for each patient; median 0.55, IQR 0.43-0.64). QoL AUC: quality of life area under the curve.

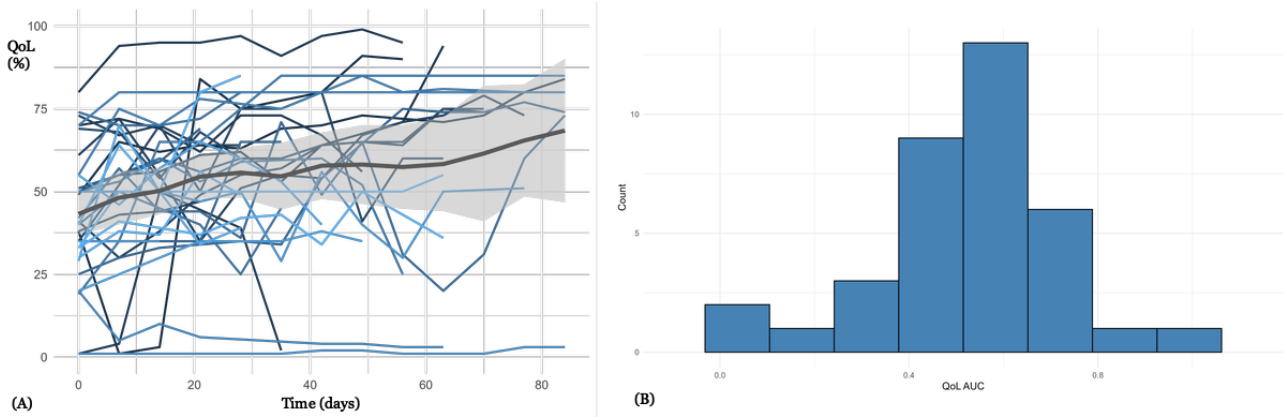

Figure 4 presents QoL-T for 25 patients from hematology (14 female and 11 male patients; median age 63, IQR 45-71 years), who received daily emails to collect HRQoL. Patients provided HRQoL data with a high level of completeness over the treatment period (median RR within 1 week 83%, IQR 80%-88%). There is a significant correlation of HRQoL and 30-item QoL Questionnaire (EORTC QLQ-C30) global health score (Spearman correlation=0.498, *P*=.02; Figure S4 in Multimedia Appendix 1). On average, HRQoL values deteriorate over a time interval of approximately 12 days, indicating the side effects of intensive chemotherapy.

Figure 4. (A) Health-related quality of life trajectories for 25 patients from hematology. Day 0 corresponds to start of chemotherapy. (B) Histogram of quality of life per patient (standardized area under the curve; median 0.46, IQR 0.41-0.62). QoL AUC: quality of life area under the curve.

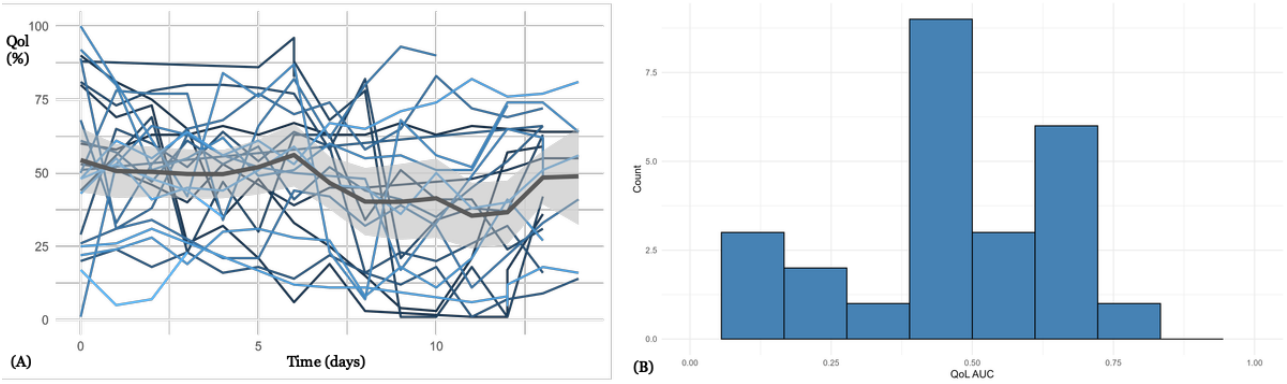

Figure 5 presents QoL-T for 26 patients from pancreatic surgery (10 female and 16 male patients; median age 66, IQR 42-83 years), who received daily messages on the day before surgery and 2 weeks after the surgical intervention. Participants achieved a median RR in the first week of 46%, IQR 46%-58%. The participation rate was 31% (8/26; ie, lowest of all 4 use cases). There is a highly significant association between HRQoL and pain levels (Spearman correlation=−0.552, *P*<.001; Figure S5 in Multimedia Appendix 1), which indicates that pain is associated with decreased QoL.

XSL•FO
RenderX

Figure 5. (A) Health-related quality of life trajectories for 26 patients from visceral surgery. Day 0 corresponds to the day before surgery. (B) Histogram of quality of life per patient (standardized area under the curve; median 0.59, IQR 0.49-0.77). QoL AUC: quality of life area under the curve.

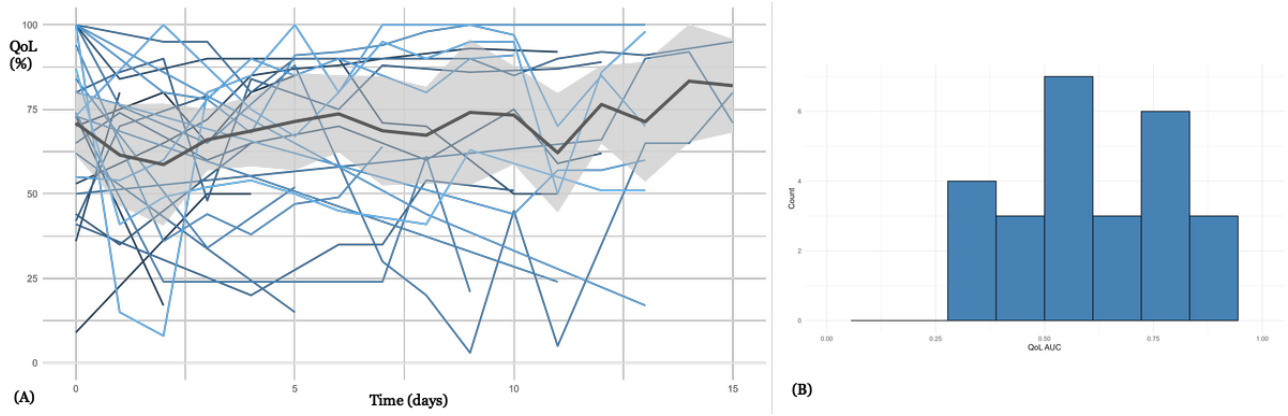

Figure 6 shows QoL-T for 23 patients from neurosurgery (7 female and 16 male patients; median age 64, IQR 21-85 years), who received daily messages 2 weeks after the surgical intervention. Participants achieved a median RR in the first week of 43%, IQR 31%-61%. A negative correlation between HRQoL and pain level was observed, which aligns with the direction of the association seen in visceral surgery. However, this did not reach statistical significance (Figure S6 in Multimedia Appendix 1).

Figure 6. (A) Health-related quality of life trajectories for 23 patients from neurosurgery. Day 0 corresponds to the day before surgery. (B) Histogram of quality of life per patient (standardized area under the curve; median 0.79, IQR 0.62-0.90). QoL AUC: quality of life area under the curve.

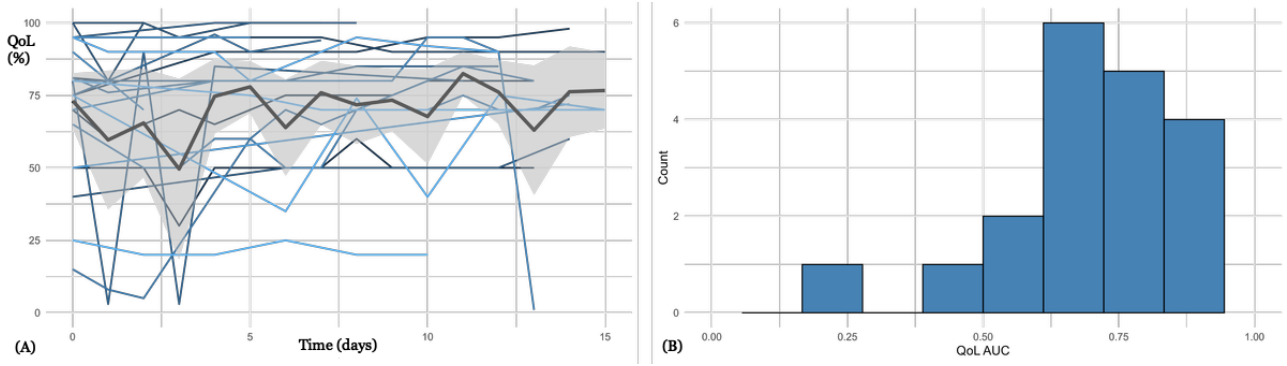

QoL-T in Figures 3-6 show considerable variability, both intraindividually (over time) and interindividually (comparison between patients). There are interpretable patterns; for example, in Figure 3 there is an average improvement of HRQoL during the course of inpatient treatment, or in Figure 4, there is an HRQoL change during the course of chemotherapy.

We did not find a significant effect of patient sex on response behavior (for details, see Figures S7-S10 in Multimedia Appendix 1). Of note, we did not observe a significant association between patient age and response count in any use case (for details, see Figures S11-S14 in Multimedia Appendix 1). We assessed time from email notification to patient answer in all use cases. The lowest time to patient answer was observed in psychosomatics (median 0.9 hours, IQR 0.3-8.0; Figure S15 in Multimedia Appendix 1). In hematology, median time to answer was 6.7 hours, IQR 1.5-17.2 (Figure S16 in Multimedia Appendix 1); in visceral surgery, median time to answer was 7.1 hours, IQR 2.9-18.9 (Figure S17 in Multimedia Appendix 1), and in neurosurgery, median time to answer was 5.3 hours, IQR 0.6-17.7 (Figure S18 in Multimedia Appendix 1).

Discussion

Overview

At present, survival (overall or progression free) is the most established primary end point in clinical research, especially in oncology [19]. Survival has the advantage that it can be collected with high precision in time (day of death) and completeness. However, survival is a crude measure for therapeutic success because it ignores PROMs such as QoL. Therefore, PROMs are increasingly used as end points in clinical studies [20]. This approach also has limitations because it only assesses the outcome at the end of an observation interval but not the patient trajectory: when 2 therapies result in the same average HRQoL after 1 year, therapy A can still be superior to therapy B, if HRQoL is significantly better during the time course of this year. Many established HRQoL instruments such as the WHO-5 well-being index, Functional Assessment of Cancer Therapy-General (FACT-G), or EQ-5D have short recall periods (<2 weeks). Consequently, frequent data collection (every 2 weeks up to daily) is needed for complete, high-resolution HRQoL trajectories. This is a methodological challenge because low RRs were reported in several settings.

XSL•FO
RenderX

Principal Findings

Our hypothesis was that assessing HRQoL on patient' mobile devices and integrating these findings into the EHR system can enable such high-resolution collection of HRQoL because a patient would be motivated to contribute to their own therapy. In our setting, a patient could directly inform their treatment team about QoL and symptoms. We could show that this approach is actually working in 4 completely different medical domains (psychosomatics, hematology, visceral surgery, and neurosurgery), both from a technical and a clinical perspective. Clinical staff emphasized the value of EHR integration because it supports patient care, for example, regarding pain management. Our proposed IT architecture proved itself in pilot operation. Patients were responding to these HRQoL requests on a regular basis.

What to measure regarding patient outcomes, including the frequency of assessment, was described as a key challenge [8]. We applied a 100-level VAS with a question in simple language because, from a clinical point of view, long texts are not suitable for seriously ill patients (eg, intensive care setting). It was demonstrated that even for those patients (eg, directly after major surgery), a large proportion is able and willing to answer HRQoL questions with their smartphones on a daily basis. In the European Union, software that calculates a clinical score may be classified as a medical device, depending on its intended use and clinical impact. While a single-item HRQoL question may be viewed as "clinical documentation," any software-based assessment intended to influence diagnostic or therapeutic decisions requires rigorous validation. Given the vast technical heterogeneity of patient-owned smartphones, the technical and clinical validation required for such software to meet medical device standards remains a significant hurdle for large-scale implementation. More complex HRQoL instruments, such as the EORTC QLQ-C30, can provide a more detailed assessment of different QoL dimensions but require substantially more documentation effort. We propose to use such instruments in larger time intervals combined with high-frequency short HRQoL assessments. In our study, patients of almost all age groups were capable to do this data collection. The clinical team or family members can support patients who are not able to handle a mobile device.

The median time from email to answer was <8 hours; therefore, daily or weekly assessments are feasible. We propose starting the VAS with 1 and assigning an HRQoL value of 0 only to patients who are deceased. By this means, HRQoL values can be directly transformed into survival data (dead=HRQoL 0 and alive=HRQoL 1-100). This enables a direct comparison of HRQoL trajectories with survival, which is the current standard end point.

Comparison With Prior Work

PROMs integrated with EHR were described previously [21]. High-resolution HRQoL trajectories within the EHR constitute a novel readout for the course of disease. This level of detail was previously unavailable and, to our knowledge, has not been comprehensively established for any disease. We demonstrated that this approach is feasible across multiple, clinically distinct domains, ranging from psychosomatics to visceral surgery. HRQoL trajectories could support early detection of complications. In contrast to survival analysis, effects of interventions such as chemotherapy or surgery on QoL can be assessed quantitatively with HRQoL trajectories. These trajectories can potentially enable sensitive comparison of different medical interventions in the future and change the course of therapy.

A related method is ecological momentary assessment [22], which captures real-time data about patients' experiences and behaviors in their natural environment. A similar approach such as in our pilot study was reported with the eRAPID system [23], which was applied for patients with breast, gynecological, and colorectal cancer. With eRAPID, the patient logged onto a web-based system (Qtool) to complete online questionnaires. To simplify this step, we sent a 1-time EDC link to the patient (ie, patient does not need an individual password). We used a CDISC-compliant EDC database to make the approach more scalable: the Medical Data Models portal [24] provides ≥25.000 questionnaires for download, which can be used for data collection with CDISC-compliant systems. Of note, the eRAPID team from Leeds was able to demonstrate that PROM collection as intervention can improve patient outcomes in patients with tumor (physical well-being and self-efficacy) [25]. Improved patient outcomes induced by PROM collection was also reported in other medical domains such as orthopedics [26]. In the field of cardiology, it was reported that a disease-specific PROM score is the strongest predictor for survival of heart failure [27].

Limitations and Future Work

Our findings are based on a single site pilot study with a relatively small sample size for each clinical use case. Consequently, our analysis may have lacked sufficient statistical power to detect smaller effect sizes. Due to technical constraints of the local EHR system, integration was limited to PDF files. Fast Healthcare Interoperability Resources–based EHR integration can enable improved support for clinical workflows. Further research is needed to define the best time intervals and most suitable assessment instruments for HRQoL collection. Most probably, these parameters are disease-specific. In some use cases, HRQoL measurements can be extended to the period before therapy. The rate of patients who accepted the electronic system varied considerably between use cases: in psychosomatics, >90% agreed to participate. In contrast, in visceral surgery, only about one-third of patients were willing or able to collect HRQoL data on the smartphone in the perioperative phase. HRQoL has several dimensions (eg, physical and psychosocial QoL). To compare 2 treatment schemes regarding HRQoL, a reduction to 1 dimension is needed (is therapy A better than B?). Therefore, further research is needed to define the most appropriate calculation of overall HRQoL. Another research topic is the metric to aggregate a series of HRQoL values. In this pilot study, we proposed AUC of HRQoL over time. Patient questionnaires can have a significant nonresponse bias, for example, language proficiency has a significant effect on RR [28]. In our setting, some patients had no access to a smartphone. Therefore, additional data capture methods (eg, tablet and interview by a health care professional) are needed for complete HRQoL data. At present, smartphones are used by >95% of the population in high-income

XSL•FO
RenderX

countries [29]. Electronic HRQoL collection software needs to be further improved, given the disease-related limitations for patients compared with the general population. Software with high usability can contribute to address this problem: for instance, the study by Carter et al [30] reported an increase in PROM completeness from 24.9% to 67.0% by user-centered software design. Future work will focus on leveraging HRQoL trajectories to provide clinical decision support and enabling physicians to personalize therapeutic interventions, while simultaneously establishing a sensitive, dynamic outcome measure for future clinical trials.

Conclusions

High-resolution QoL-T with EHR integration are clinically feasible in a broad range of clinical settings. Collection of HRQoL data via patients' mobile devices and secure data transfer to the protected EHR is technically feasible. These trajectories can provide a novel readout, which is not covered by survival analysis or a PROM end point. These data can potentially be useful to characterize diseases and compare different therapies with high sensitivity and time resolution.

Funding

This study was funded by the German Federal Ministry of Education and Research (BMBF; grant 01ZZ2321F).

Data Availability

Anonymized data and data analysis software for this manuscript are available in Multimedia Appendices 2 and 3. Access to individual patient data can be requested from the Medical Data Integration Center of Heidelberg University Hospital. Data access requests will be reviewed by the Use and Access Committee of Heidelberg University Hospital.

Authors' Contributions

Conceptualization: MD, TMP, MG
Formal analysis: MD
Investigation: SCF, TD, NS, CFB, PL, TMP
Methodology: SD-B, LJK, KK, FF-K
Resources: MD
Software: RF, MCB, LW, TMP, MG
Supervision: MD, SS, ML, SMK, HCF, TMP, MG
Validation: MD
Writing – original draft: MD
Writing – review: SS, ML, SMK, HCF, TMP, SD-B, LJK, KK, FF-K

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data regarding electronic health record integration, correlation analyses, and time to patient answers.
[DOCX File , 1578 KB - jmir_v28i1e79834_app1.docx]

Multimedia Appendix 2
Program code for quality of life trajectories (trajectory.R).
[TXT File , 2 KB - jmir_v28i1e79834_app2.txt]

Multimedia Appendix 3
Example data for quality of life trajectories (example.R).
[TXT File , 2 KB - jmir_v28i1e79834_app3.txt]

References

1. Escudero-Vilaplana V, Bernal E, Casado G, Collado-Borrell R, Diez-Fernández R, Fernández Román AB, et al. Defining a standard set of patient-reported outcomes for patients with advanced ovarian cancer. Front Oncol 2022;12:885910 [FREE Full text] [doi: 10.3389/fonc.2022.885910] [Medline: 35664764]
2. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 well-being index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167-176 [FREE Full text] [doi: 10.1159/000376585] [Medline: 25831962]
3. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]

XSL•FO
RenderX

4. Minasian LM, O'Mara A, Mitchell SA. Clinician and patient reporting of symptomatic adverse events in cancer clinical trials: using CTCAE and PRO-CTCAE to provide two distinct and complementary perspectives. Patient Relat Outcome Meas 2022;13:249-258 [FREE Full text] [doi: 10.2147/PROM.S256567] [Medline: 36524232]
5. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol group. Ann Med 2001 Jul 08;33(5):337-343. [doi: 10.3109/07853890109002087] [Medline: 11491192]
6. Mori M, Dhruva SS, Geirsson A, Krumholz HM. Characterization of multi-domain postoperative recovery trajectories after cardiac surgery using a digital platform. NPJ Digit Med 2022 Dec 24;5(1):192 [FREE Full text] [doi: 10.1038/s41746-022-00736-0] [Medline: 36564550]
7. Wang K, Eftang CN, Jakobsen RB, Årøen A. Review of response rates over time in registry-based studies using patient-reported outcome measures. BMJ Open 2020 Aug 06;10(8):e030808 [FREE Full text] [doi: 10.1136/bmjopen-2019-030808] [Medline: 32764078]
8. Marra C, Chico T, Alexandrow A, Dixon WG, Briffa N, Rainaldi E, et al. Addressing the challenges of integrating digital health technologies to measure patient-centred outcomes in clinical registries. Lancet Digit Health 2025 Mar;7(3):e225-e231 [FREE Full text] [doi: 10.1016/S2589-7500(24)00223-1] [Medline: 39672737]
9. Greulich L, Hegselmann S, Dugas M. An open-source, standard-compliant, and mobile electronic data capture system for medical research (OpenEDC): design and evaluation study. JMIR Med Inform 2021 Nov 19;9(11):e29176 [FREE Full text] [doi: 10.2196/29176] [Medline: 34806987]
10. Dugas M, Blumenstock M, Dittrich T, Eisenmann U, Feder SC, Fritz-Kebede F, et al. Next-generation study databases require FAIR, EHR-integrated, and scalable electronic data capture for medical documentation and decision support. NPJ Digit Med 2024 Jan 12;7(1):10 [FREE Full text] [doi: 10.1038/s41746-023-00994-6] [Medline: 38216645]
11. Operational data model (ODM). Clinical Data Interchange Standards Consortium (CDISC). URL: https://www.cdisc.org/standards/data-exchange/odm [accessed 2025-03-15]
12. Kroenke K, Spitzer RL, Williams JB, Löwe B. An ultra-brief screening scale for anxiety and depression: the PHQ–4. Psychosomatics 2009 Nov;50(6):613-621. [doi: 10.1016/S0033-3182(09)70864-3]
13. Toussaint A, Murray AM, Voigt K, Herzog A, Gierk B, Kroenke K, et al. Development and validation of the somatic symptom disorder-B criteria scale (SSD-12). Psychosom Med 2016 Jan;78(1):5-12. [doi: 10.1097/PSY.0000000000000240] [Medline: 26461855]
14. Prinz U, Nutzinger D, Schulz H, Petermann F, Braukhaus C, Andreas S. Die symptom-checkliste-90-R und ihre Kurzversionen: psychometrische analysen bei patienten mit psychischen erkrankungen. Phys Rehab Kur Med 2008 Nov 25;18(06):337-343. [doi: 10.1055/s-0028-1093323]
15. Horvath AO, Greenberg LS. Development and validation of the working alliance inventory. J Couns Psychol 1989 Apr;36(2):223-233. [doi: 10.1037/0022-0167.36.2.223]
16. Fonagy P, Luyten P, Moulton-Perkins A, Lee Y, Warren F, Howard S, et al. Development and validation of a self-report measure of mentalizing: the reflective functioning questionnaire. PLoS One 2016 Jul 8;11(7):e0158678 [FREE Full text] [doi: 10.1371/journal.pone.0158678] [Medline: 27392018]
17. Kliem S, Mößle T, Zenger M, Strauß B, Brähler E, Hilbert A. The eating disorder examination-questionnaire 8: a brief measure of eating disorder psychopathology (EDE-Q8). Int J Eat Disord 2016 Jun;49(6):613-616. [doi: 10.1002/eat.22487] [Medline: 26711183]
18. Dueck AC, Mendoza TR, Mitchell SA, Reeve BB, Castro KM, Rogak LJ, National Cancer Institute PRO-CTCAE Study Group. Validity and reliability of the US National Cancer Institute's patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE). JAMA Oncol 2015 Nov 01;1(8):1051-1059 [FREE Full text] [doi: 10.1001/jamaoncol.2015.2639] [Medline: 26270597]
19. Royle KL, Meads D, Visser-Rogers JK, White IR, Cairns DA. How is overall survival assessed in randomised clinical trials in cancer and are subsequent treatment lines considered? A systematic review. Trials 2023 Nov 06;24(1):708 [FREE Full text] [doi: 10.1186/s13063-023-07730-1] [Medline: 37926806]
20. Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims. Food and Drug Administration (FDA). URL: https://www.fda.gov/media/77832/download [accessed 2025-02-12]
21. Mohindra NA, Garcia SF, Kircher S, Barnard C, Perry LM, Lyleroehr M, et al. Development of an electronic health record-integrated patient-reported outcome-based shared decision-making dashboard in oncology. JAMIA Open 2024 Oct;7(3):ooae056. [doi: 10.1093/jamiaopen/ooae056] [Medline: 39049991]
22. Kraft R, Reichert M, Pryss R. Mobile crowdsensing in ecological momentary assessment mHealth studies: a systematic review and analysis. Sensors (Basel) 2024 Jan 12;24(2):472 [FREE Full text] [doi: 10.3390/s24020472] [Medline: 38257567]
23. Holch P, Warrington L, Bamforth LC, Keding A, Ziegler LE, Absolom K, et al. Development of an integrated electronic platform for patient self-report and management of adverse events during cancer treatment. Ann Oncol 2017 Sep 01;28(9):2305-2311 [FREE Full text] [doi: 10.1093/annonc/mdx317] [Medline: 28911065]
24. Riepenhausen S, Blumenstock M, Niklas C, Hegselmann S, Neuhaus P, Meidt A, et al. Europe's largest research infrastructure for curated medical data models with semantic annotations. Methods Inf Med 2024 May 13;63(1-02):52-61 [FREE Full text] [doi: 10.1055/s-0044-1786839] [Medline: 38740374]

XSL•FO
RenderX

25. Absolom K, Warrington L, Hudson E, Hewison J, Morris C, Holch P, et al. Phase III randomized controlled trial of eRAPID: eHealth intervention during chemotherapy. J Clin Oncol 2021 Mar 01;39(7):734-747 [FREE Full text] [doi: 10.1200/JCO.20.02015] [Medline: 33417506]
26. Schöner L, Kuklinski D, Wittich L, Steinbeck V, Langenberger B, Breitkreuz T, et al. Cost-effectiveness of a patient-reported outcome-based remote monitoring and alert intervention for early detection of critical recovery after joint replacement: a randomised controlled trial. PLoS Med 2024 Oct 9;21(10):e1004459 [FREE Full text] [doi: 10.1371/journal.pmed.1004459] [Medline: 39383175]
27. Pokharel Y, Khariton Y, Tang Y, Nassif ME, Chan PS, Arnold SV, et al. Association of serial Kansas City cardiomyopathy questionnaire assessments with death and hospitalization in patients with heart failure with preserved and reduced ejection fraction: a secondary analysis of 2 randomized clinical trials. JAMA Cardiol 2017 Dec 01;2(12):1315-1321 [FREE Full text] [doi: 10.1001/jamacardio.2017.3983] [Medline: 29094152]
28. KÖhn S, Schlumbohm A, Marquardt M, Scheel-Sailer A, Tobler S, Vontobel J, et al. Predicting non-response in patient-reported outcome measures: results from the Swiss quality assurance programme in cardiac inpatient rehabilitation. Int J Qual Health Care 2022 Dec 08;34(4):mzac093 [FREE Full text] [doi: 10.1093/intqhc/mzac093] [Medline: 36399024]
29. Mobile phone ownership. International Telecommunication Union. URL: https://www.itu.int/itu-d/reports/statistics/2024/11/10/ff24-mobile-phone-ownership/ [accessed 2025-02-12]
30. Carter E, Benda N, Kim S, Qiu Y, Yu Z, Gunning F, et al. Increasing completion of daily patient-reported outcomes in psychotherapies for late-life depression through user-centered design. Appl Clin Inform 2024 Oct 20;15(5):986-996 [FREE Full text] [doi: 10.1055/s-0044-1790545] [Medline: 39566558]

Abbreviations

AUC: area under the curve
CDISC: Clinical Data Interchange Standards Consortium
EDC: electronic data capture
EHR: electronic health record
EDE-Q8: 8-item Eating Disorder Examination-Questionnaire
EORTC QLQ-C30: 30-item Core QoL Questionnaire
FACT-G: Functional Assessment of Cancer Therapy-General
HRQoL: health-related quality of life
PHQ-4: 4-item Patient Health Questionnaire
PRO-CTCAE: patient-reported outcomes–common terminology criteria for adverse events
PROM: patient-reported outcome measure
QoL: quality of life
QoL-T: quality of life trajectories
RFQ: Reflective Functioning Questionnaire
RR: response rate
SCLK-11: Symptom-Checkliste Kurzversion
SSD-12: 12-item Somatic System Disorder
UKHD: Heidelberg University Hospital (Universitätsklinikum Heidelberg)
VAS: visual analog scale
WHO-5: 5-item World Health Organization well-being index

Edited by A Coristine; submitted 29.Jun.2025; peer-reviewed by U Hübner, Y Chu; comments to author 04.Nov.2025; revised version received 28.Jan.2026; accepted 29.Jan.2026; published 24.Feb.2026.

Please cite as:
Dugas M, Fleige R, Blumenstock MC, Feder SC, Dittrich T, Siegel N, Bergmann CF, Wettach L, Sauer S, Lenga P, Krieg SM, Dugas-Breit S, Kessler LJ, Kosmalla K, Fritz-Kebede F, Loos M, Friederich HC, Pausch TM, Ganzinger M
Quality of Life Trajectories With Integration Into Electronic Health Records for High-Resolution Patient Outcomes: Algorithm Development and Validation Study
J Med Internet Res 2026;28:e79834
URL: https://www.jmir.org/2026/1/e79834
doi:10.2196/79834
PMID:

XSL•FO
RenderX

©Martin Dugas, Robin Fleige, Max Christian Blumenstock, Stephan Christoph Feder, Tobias Dittrich, Niels Siegel, Celine Fabienne Bergmann, Luis Wettach, Sandra Sauer, Pavlina Lenga, Sandro M Krieg, Susanne Dugas-Breit, Lucy Joanne Kessler, Kosima Kosmalla, Fleur Fritz-Kebede, Martin Loos, Hans-Christoph Friederich, Thomas M Pausch, Matthias Ganzinger. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Communicative Behaviors in an Internet-Based Intervention for Individuals With Autism: Mixed Methods Analysis

Britta Westerberg[1,2], PhD; Karin Jacobson[1,2], MSc; Maria Unenge Hallerbäck[1,3,4], MD, PhD; Susanne Bejerot[1,2], MD, PhD; Fredrik Holländare[1,2], PhD

[1]School of Medical Sciences, Faculty of Medicine and Health, Örebro University, Örebro, Sweden

[2]University Health Care Research Center, Faculty of Medicine and Health, Örebro University, Örebro, Sweden

[3]Department of Public Health Sciences, Karlstad University, Karlstad, Sweden

[4]Centre for Clinical Research, Region Värmland, Karlstad, Sweden

Corresponding Author:
Britta Westerberg, PhD
School of Medical Sciences, Faculty of Medicine and Health
Örebro University
Vuxenhabiliteringen, Box 1613
Örebro, 70116
Sweden
Phone: 46 0767672800
Email: britta.westerberg@oru.se

Abstract

Background: To meet the needs of individuals diagnosed with autism, internet-based interventions have been developed with a variety of objectives. A deeper understanding of the mechanisms of change may help tailor interventions to individual needs. The communicative behaviors of individuals with autism participating in text-based internet-based interventions remain largely unexplored, as do their potential relations to clinical outcomes. An improved understanding of participants' behaviors may help therapists better tailor support, promote engagement, and enhance treatment outcomes.

Objective: This study aimed to explore the communicative behaviors of individuals with autism participating in an internet-based intervention and to examine whether different behavioral patterns were associated with treatment outcomes or treatment adherence.

Methods: Messages from 34 participants enrolled in an 18-week internet-based cognitive behavioral therapy program were analyzed using abductive qualitative content analysis. Correlational analyses were used to examine the relationships between qualitative categories and change scores on outcome measures and rates of module completion.

Results: Fourteen behavioral categories were identified and grouped into three overarching domains: (1) "This is me," which encompasses the participants' narratives on identity, personality, autistic functioning, current and past circumstances, and worldview; (2) "Working with the treatment," which included statements related to engagement with the treatment process; and (3) "I struggle," which comprised of past and present negative experiences and challenges. Correlational analyses revealed associations between several behavioral categories and improvements in quality of life and treatment adherence.

Conclusions: The findings highlight the importance of self-narrative formulation among individuals with autism and suggest that certain communicative behaviors—particularly those involving identity reflection and recognition of treatment-related gains—were positively associated with therapeutic outcomes. The findings enhance our understanding of how individuals with autism engage in internet-based cognitive behavioral therapy and may serve as a valuable source of information for therapists when guiding expectations regarding client outcomes and identifying participants who may benefit from additional support.

Trial Registration: ClinicalTrials.gov NCT03570372; https://clinicaltrials.gov/study/NCT03570372

(J Med Internet Res 2026;28:e76527) doi:10.2196/76527

KEYWORDS

autism; communicative behaviors; internet-based treatment; narrative identity

XSL•FO
RenderX

Introduction

Background

Autism is a neurodevelopmental condition characterized by difficulties in social communication and a stereotypical behavioral pattern [1]. These difficulties imply functional impairment in social situations and may affect psychological well-being [2,3] and low quality of life (QOL) [4]. Furthermore, individuals with autism face several challenges in accessing health care [5], and internet-based interventions offer a convenient and flexible format that may increase accessibility for this population.

The evidence base for internet-based interventions is growing, and an increasing number of diagnostic groups are being offered internet-based cognitive behavioral therapy (ICBT) [6]. Given the efficacy of ICBT, there is a need for an improved understanding of the mechanisms of change, as this knowledge would allow more targeted adaptations of treatment content to the specific needs of each service user [7].

Interventions that include therapist support tend to be more effective than those delivered without guidance [8]. In addition, factors such as participant adherence, treatment credibility, working alliance, and baseline symptom scores have all been identified as important predictors of treatment outcome in ICBT [9]. In particular, therapeutic alliance and treatment adherence are related to favorable outcomes [10,11].

Therapeutic feedback in ICBT is typically provided at regular intervals [7], in response to participant progress, and the participant may have the possibility to write to the therapist at their convenience, both in response to therapist feedback and to initiate a new conversation. Though the core ingredients in ICBT are expected to be embedded within the predetermined text-based modules or sessions, the established importance of therapeutic support [8] and alliance [10] implies that some change mechanisms are influenced by the one-to-one written communication.

The functional components of therapist feedback in ICBT, conceptualized as "therapist behaviors," are important for both adherence and outcome [12-14]. The client-related equivalent—patient behaviors in ICBT—remains less thoroughly explored. However, previous research indicates that different communicative patterns among participants in ICBT may also be associated with both outcome and adherence [15-17]. Increased insight into what is communicated by clients in ICBT is an important piece of the puzzle in forming an understanding of who is likely to benefit from the program—and who may not—as well as identifying individuals at risk of dropping out. This information can also serve as a valuable indicator to practitioners that a participant is struggling and may require additional support [15].

Svartvatten et al [11] explored behaviors reflected through client messages in ICBT for depression, and their relation to outcome and adherence. They identified 10 behavioral categories and found that "Alliance and Observes positive consequences" correlated positively with changes in outcome, and the behaviors "Alliance," "Identifies patterns and problem behaviors," "Maladaptive repetitive thinking," "Observes positive consequences," "Tries alternative behavior," "Chooses alternative behavior," and "Avoidance of treatment" were positively related to the number of modules completed.

In 2 comparative studies, Soucy et al [18] and Kraepelien et al [19] adopted the behavioral categories from Svartvatten's study to examine whether these could be generalized to their own data, including participants with depression, anxiety [18], and alcohol use disorder [19]. Although both studies found the predefined categories applicable, notable differences in both frequency and relation to adherence and outcome were observed. Based on inconsistent correlation results, both studies refrained from concluding the predictive value of client behaviors but suggested that the categories may still offer insight into the therapeutic process in ICBT. Soucy et al [18] noted that their directed (also referred to as deductive) content analysis approach—following the categories generated by Svartvatten—may have hindered the detection of other relevant themes. Therefore, they call for future research to analyze client messages inductively, without referring to pre-existing frameworks.

An inductive analytical approach may be especially justified when analyzing communicative patterns among individuals with autism, a population known to exhibit a qualitatively different communicative style compared to the norm, both generally [20] and in text-based online communication [21]. These communication differences may be manifested, for example, by an impaired narrative ability—a communicational skill important for sharing experiences and connecting with others [22]. Furthermore, research has shown that computer-mediated communication often is preferred over verbal communication among individuals with autism [21,23] and that they use internet-based communication in qualitatively different ways than individuals without autism [21,24,25]. These findings open up the possibility that the communicative behaviors in ICBT for autism may differ from those of other patient groups, which justifies approaching this question with an inductive approach, unrestricted by the deductive categories used in previous studies.

To manage the increasing number of individuals diagnosed with autism and to meet their needs, internet-based interventions have been developed with a variety of objectives, including increasing QOL [26], providing psychoeducation [27], and treating psychiatric comorbidities [28-30]. Although the research base on ICBT for autism is relatively scarce, the available studies suggest that internet-based interventions may be both particularly suitable for the needs of individuals with autism [27,31], as well as effective [27,29,30].

Within the current project, an 18-week internet-based intervention adapted for individuals with autism was evaluated for feasibility and effect compared to an active control group [26]. The results were complex and ambiguous; while no group-level effects were found on quantitative outcomes, participant satisfaction was high, and dropout rates were low, indicating good feasibility. A qualitative study on participant experiences of the intervention [31] showed that the participants

XSL•FO
RenderX

generally appreciated the internet-based format, and especially the opportunity to communicate with the therapist in writing.

However, the participants had varying experiences regarding therapist support and the messaging function. Some participants used the messaging function to enable a deeper conversation with the therapist, whereas others only occasionally responded to the therapeutic feedback, resulting in great variability in the amount of therapeutic contact [31]. Accordingly, the level of support was, to a great degree, dependent on the behavior of the participants. Whether this diverse use of the therapeutic support may be related to the treatment effect or adherence is, however, unclear and warrants further investigation. As the active ingredients in ICBT for individuals with autism are yet unexplored, we should remain open to the possibility that the varying communicative behaviors of participants may play a role, even in interventions where no group-level effect was found.

Little is known about the behavior reflected through participant messages in internet-based interventions for individuals with autism, or how variations in these behaviors may relate to treatment outcomes and adherence. Analyzing participant messages in ICBT may offer valuable insights into how individuals with autism engage in the treatment. Furthermore, early identification of specific participant behaviors may inform professionals about what to expect when treating individuals with autism, and thereby guide therapeutic decision-making. Improved understanding of client behaviors could thus be helpful both in identifying individuals at risk of dropping out and in guiding the development of future programs.

Objectives

This study aimed to gain knowledge about participants' communicative behaviors in an internet-based intervention for individuals with autism by investigating the content of participant messages. An additional aim was to explore whether the use of different behaviors (including word count and message frequency) was related to clinical outcome or adherence.

Methods

Design

This study used a mixed method design, using data collected as part of a larger randomized controlled trial (RCT) aimed at evaluating an internet-based cognitive behavioral intervention to improve QOL in autistic adults (ClinicalTrials.gov NCT03570372). A qualitative analysis of participant text messages and the calculation of change scores were first conducted independently, after which the qualitative and quantitative data were integrated through correlation analyses. The integration allowed the exploration of how the qualitative categories were related to quantitative changes.

Participants and Recruitment

All messages from participants in the intervention group (n=42) of an RCT on ICBT for individuals with autism [26] were included in the current qualitative analysis. The RCT was announced through posters in waiting rooms of health care facilities in Örebro county, Sweden, and as advertisements in the local press and on a social media platform (Facebook, Meta Platforms, Inc). Participants for the RCT were recruited by completing a digital self-application form administered through 1177, a Swedish national platform for online health care. The application form included questions on age, gender, living arrangement, occupation, and age at diagnosis. The inclusion process involved a structured interview based on the Mini International Neuropsychiatric Interview, along with a screening questionnaire covering autistic symptoms. Diagnoses were confirmed by collecting assessment records or by verbal confirmation from clinicians. A detailed description of the recruitment process and inclusion criteria of the RCT is provided in Westerberg et al [26].

Intervention

The intervention was based on an evidence-based cognitive behavioral therapy group treatment for adults with autism [32,33], which had been further developed, condensed, and adapted to an internet-based format by the first author, BW.

The intervention lasted 18 weeks. The intervention aimed to enhance QOL and sense of coherence (SOC) and to decrease psychiatric symptoms through the completion of 18 text-based modules focusing on themes relevant to these objectives (refer to Table S1 in Multimedia Appendix 1 for an overview of the themes). Each module included psychoeducation, exercises, and strategies based on cognitive behavioral therapy to enhance coping with everyday life challenges, but adapted to the needs of individuals with autism. In line with recommendations for such adaptations [34], a significant portion of the program was dedicated to psychoeducational content about autism, individual variations, and common comorbidities. The intervention also introduced tools and concepts to support self-understanding and provided terminology to describe participants' functioning. Every 2 weeks, the participants were invited to take part in a live chat session together with other participants, focused on discussing the theme of the most recent module.

In most modules, home exercises were designed to be completed either using text-based worksheets or as "field work" to be reported on the online platform. All modules also included reflective questions to be answered directly in conjunction with the text. The documentation and reports from the exercises, along with responses to the reflective questions, formed the basis for the therapeutic feedback.

Therapeutic feedback was delivered asynchronously via a messaging function available on the same platform as the treatment. The messaging function resembled an email page, allowing participants to respond to therapeutic feedback and initiate new conversations. Therapists were expected to reply within 1 working day. However, participants were not required to send any messages via this function.

Material

All messages written by participants using the messaging function were included. Aside from the removal of information that may reveal participant identities (such as names, telephone numbers, or web addresses), the messages were not modified in any way.

XSL•FO
RenderX

Change scores (from baseline to posttreatment) on outcomes from the RCT were included for a quantitative correlation analysis. The Brunnsviken Brief Quality of Life Scale (BBQ) [35], consisting of 12 items (responses ranging from 0=do not agree to 4=totally agree), was used to assess QOL. Each item related to satisfaction in a specific life area was weighted against a rating of the importance of that area. The multiplied products of each item-pair (satisfaction × importance across 6 life areas) were summed to obtain a total QOL score (range: 0-96), with higher scores representing a higher QOL.

The 13-item Sense of Coherence (SOC-13) scale was used to assess the concept of SOC [36]. SOC reflects the extent to which an individual perceives life events as coherent and comprehensible and life demands as manageable and meaningful. Items are rated on a 7-point scale, with a higher total score (range: 13-91) indicating a stronger SOC.

To assess symptoms of depression and anxiety during the past week, the Hospital Anxiety and Depression Scale (HADS) for depression (HADS-D; 7 items) and anxiety (HADS-A; 7 items) subscales were used [37]. Items are rated on a 4-point scale, with higher scores (range: 0-21) indicating greater symptom severity.

Word count and the frequency of participant and therapist messages were analyzed in relation to the outcome. Generic messages sent to all participants (such as reminders of the peer participant chat sessions) were removed from the data at an early stage. Similarly, messages whose sole purpose was to inform the participant that they had received a new module or to remind them to complete self-assessments were excluded from the analyses.

Qualitative Analysis

The material was analyzed using both a directed and conventional approach to content analysis [38], adopting a combination of inductive and deductive perspectives, resulting in an abductive analytical process [39]. All coding and sorting were conducted using NVivo qualitative data analysis software (Lumivero, LLC) [40] by 2 licensed psychologists (BW and KJ). BW was also 1 of 4 therapists during the trial and was therefore familiar with both the foundational elements and components of the program, as well as the specific setting and design of the trial. As personal identifiers were removed from the texts before analysis, the participants' identities were not revealed through the texts. However, the coding process was not fully blinded, as—due to her previous role as a therapist—BW could recognize certain utterances as stemming from individual participants.

Initially, all messages were read to gain an overall sense of the material and its character. Thereafter, BW segmented and condensed the texts into meaning units, after which BW and KJ jointly and inductively coded the meaning units from 5 participants and discussed the codes to develop a preliminary coding framework. All utterances that could be considered reflective of a behavior were given a descriptive code to capture the manifest content of the participant's messages.

In this analysis, a behavior was defined as statements in the participant messages that appeared to serve a function or convey an intention, either in relation to internal processes or to the treatment work or the therapist. These included both utterances with semantic and meta-linguistic content. Statements that—although functional—were considered not to be of importance for the research questions, such as using polite phrases ("have a good day") or clarifying the structure of the messages ("firstly I will answer your questions"), were excluded.

When a preliminary coding framework was established, the 2 researchers independently coded the texts of 1 participant and compared their interpretations and the wording of codes to refine the codes and to reach consensus regarding the final framework to be used thereafter. Once agreement regarding the wording of codes was achieved, BW coded the communication from the remaining 28 participants in the same manner.

When the texts of approximately 10 individuals had been coded, BW began inductively sorting the codes into clusters illustrating similar behaviors. During this process, a pattern emerged in which several of the clusters resembled categories identified in earlier research [11,18,19]. A decision was made to allow the use of previously identified categories [11] when a category was identified that clearly aligned with one of these. This resulted in an abductive approach, meaning that we moved iteratively between the data and the categories identified in earlier research, which were refined and adjusted to better fit our data during this process. Given the large amount of data and codes, coding and sorting were conducted alternately until all relevant meaning units had been coded and sorted. No third-party validation was conducted during the analysis process.

Quantitative Analysis

The frequencies of each participant's contribution of codes to the categories derived from the qualitative analysis were used in correlation analyses to explore relationships with the quantitative change scores. Change scores on BBQ, SOC-13, HADS-D, and HADS-A were calculated by subtracting preintervention scores from postintervention scores. Normality of the variables was assessed using the Shapiro-Wilk test.

To assess behavior categories in relation to module completion independently of the amount of text written, the individual frequency of codes in each category was divided by the total number of words written by each participant. Accordingly, the frequency of each behavior category was calculated as a proportion of the total number of words, which minimized the influence of overall text length (which would be strongly dependent on the number of modules completed), as the aim was to examine the relative prevalence of behaviors rather than general writing activity.

BBQ, HADS-D, and HADS-A were all normally distributed. However, as the distribution of SOC-13, all behavior category variables, word and message frequency, and module completion rates deviated from normality, Spearman rank-order correlation (ρ) was used to assess the associations between behavior categories and change scores, module completion, and word and message frequency. No adjustment for multiple testing was made, as the study was exploratory and did not involve predefined statistical hypotheses. All quantitative analyses were performed using SPSS Statistics (version 29) [41].

XSL•FO
RenderX

Ethical Considerations

The original trial received ethical approval from the Regional Ethics Committee in Uppsala, Sweden (ref no 2017/392), and the Swedish Ethical Review Authority later approved an amendment (ref no 2022-05792-02) with clarifications regarding this mixed methods study. At inclusion, all participants provided consent for their personal and medical data to be used for research purposes. All data were deidentified before analysis. Personal identifiers were removed from the analyzed text material, participants were assigned unique IDs, and the code key was accessible only to the research team. Participants did not receive any financial compensation for participation in this study. However, a gift card of 300 SEK (≈ US $32) was provided to participants on completion of the postintervention and follow-up assessments of the original RCT.

Results

Overview

As 8 participants did not provide any written material, the qualitative analysis included messages from 34 participants. Table 1 provides the baseline characteristics of the participants and the number of completed modules.

Table 1. Baseline characteristics and adherence of the participants included in the qualitative analysis (N=34).

Characteristic	Value
Age (years), mean (SD)	33.8 (10)
Gender, n (%)	
Men	14 (41.2)
Women	18 (52.9)
Other (nonbinary or transgender)	2 (5.9)
Habitation, n (%)	
With partner and/or children	12 (35.2)
With parents	9 (26.5)
Alone	11 (32.4)
Group or serviced housing	1 (2.9)
Other	1 (2.9)
Age when diagnosed (years), n (%)	
<19	11 (32.4)
20-35	14 (41.2)
>36	9 (26.5)
Occupation, n (%)	
Employed	4 (11.8)
Daily activities	2 (5.9)
Student	4 (11.8)
Job seeker	9 (26.5)
Sick leave	7 (20.6)
Other	8 (23.5)
Completed modules, mean (SD) and median (IQR)	15.6 (4.2) and 18 (14.8-18)

Categories From the Qualitative Content Analysis

In total, 2569 codes were identified, of which 2476 were considered relevant to the aim. These were sorted into 14 categories, which were further grouped into 3 overarching domains: “This is me” (1092 codes and 4 categories), “Working with the treatment” (861 codes and 6 categories), and “I struggle” (523 codes and 4 categories). Table 2 provides an overview and the relative frequency of codes (%) across domains and categories, as well as definitions of each category. Refer to Table S2 in Multimedia Appendix 1 for example codes and quotes from each category.

XSL•FO
RenderX

Table 2. Overview of domains, categories, and the relative frequency of codes (%), and definitions of each category.

Domain and category, %	Definition
This is me, 44.1%	
This is who I am, 15.5%	Texts reflecting a personal narrative about what kind of person one is, what qualities, difficulties, and needs they have, their autistic functioning, and how they (think that they) are perceived by others.
My present and past circumstances, 8.2%	Information on current events, everyday life, positive experiences, and early formative events.
Change is possible, 14.2%	Strategies that work or have potential to work, that they have developed through life, and show motivation and insight that they have the agency to act to progress.
My point of view, 6.2%	Beliefs, opinions, and thoughts about society, autism, human psychology, and the perspective of others.
Working with the treatment, 34.8%	
Appreciation and treatment alliance, 8.4%	Expressions of appreciation, a positive attitude, and bonds of alliance toward the treatment or the therapist.
Putting the treatment aside, 3.2%	Information about not having completed parts of the treatment, and reasons for this, including both psychological issues, such as forgetting and lacking energy, but also external events that got in the way of the treatment work.
Plans to attempt a new task, 7.0%	Reflections regarding how to – or plans to implement a treatment task or exercise, or deciding on a treatment goal.
Have attempted a new task, 4.6%	Reports on completion of a treatment task or exercise, or reflections around the implementation of a task or exercise.
Observing positive consequences of treatment, 3.2%	Statements illustrating that the participant has observed a positive consequence on their personal development from the treatment or specific exercises.
Problems with the treatment, 8.3%	Texts in which the participant expresses difficulties, frustration, or other negative aspects of the treatment content or format, such as failure to understand a task or its purpose, considering certain parts of the treatment irrelevant, or technical problems.
I struggle, 21.1%	
Life is and has been demanding, 11.6%	Descriptions of current and past difficult events or circumstances, that they have been treated badly throughout life, experienced failure, and have been enduring from mental ill-health.
Identifies patterns and problem behaviors, 3.1%	Texts reflecting identification and insight into maladaptive and safety behaviors, and identification of the relation between these behaviors and negative consequences.
I am troubled by mental ill-health, 4.3%	Factual descriptions of current mental ill-health, loneliness, and struggles and their causes and consequences, without engaging in maladaptive thoughts or rumination.
Maladaptive thoughts, 2.2%	Expressions of hopelessness, meaninglessness, or distress, as well as other cognitive distortions, indicate a stagnation of cognitive flexibility.

The domain “This is me” consists of the categories “This is who I am,” “My present and past circumstances,” “Change is possible,” and “My point of view.” The domain covers participants’ descriptions of themselves in terms of personality, abilities, and difficulties, as well as current and past life situations that have influenced who they are. It further comprises their beliefs and opinions, their view of life and others, how they have used strategies to manage problems, and how they have developed, or are motivated to develop and function better. This domain illustrates that participants have insight into their agency, responsibility, and ability to influence their situation, but also reflects self-awareness regarding how their autistic difficulties pose challenges in this.

The domain “Working with the treatment” contains the categories “Appreciation and treatment alliance,” “Putting the treatment aside,” “Plans to attempt a new task,” “Have attempted a new task,” “Observing positive consequences of treatment,” and “Problems with the treatment.” These categories include statements that are in some way related to participants’ engagement in the treatment. It involves both positive and negative experiences and effects of the treatment, as well as reports on completed tasks and plans for future tasks. Reasons why certain parts of the treatment were not carried out are also included in this domain.

The domain “I struggle” contains the categories “Life is and has been demanding,” “Identifies patterns and problem behaviors,” “I am troubled by mental ill-health,” and “Maladaptive thinking.” This domain contains descriptions of past and current negative events, experiences of failure, descriptions of a difficult life situation, experiences of mistreatment, and how these have negatively affected mental health. It also covers how their mental ill-health manifests in symptoms and negative behavioral patterns.

XSL•FO
RenderX

Nine of the categories were found to correspond to categories identified in previous studies. Refer to Table S3 in Multimedia Appendix 1 for an overview of the categories from this study and—where applicable—the corresponding category from the studies by Svartvatten et al [11] and Kraepelien et al [19]. Five categories not previously described by the literature were identified in this study. These are: "This is who I am," "My present and past circumstances," "Change is possible," "My point of view, and "Life is and has been demanding." Two categories identified in previous research, "Confrontational alliance rupture" [11,18,19] and "Observes alcohol-related setback" [19], were not identified in our data.

Correlation Between Frequency of Behaviors and Outcome

Table 3 provides baseline, posttreatment, and change scores on outcomes, and Table 4 provides correlations between the frequency of participant behaviors and change scores of outcomes and the number of completed modules. As 5 participants—although included in the qualitative analysis—did not complete the postassessment, the correlational analyses include data from 29 participants.

The domains "This is me" and "Working with the treatment" and the categories "This is who I am," "Change is possible," "My point of view," "Appreciation and treatment alliance," "Plans to attempt a new task," "Have attempted a new task," and "Observing positive consequences of treatment" correlated significantly with a positive change in BBQ. The domain "This is me" and the categories "This is who I am," "Change is possible," and "Identifies patterns and problem behaviors" correlated significantly with a higher number of completed modules. No other significant correlations between the frequency of behaviors and outcomes were observed.

These findings show that when participants more frequently expressed aspects of self-understanding (as in the domain "This is me") and engaged in behaviors related to the treatment (as in the domain "Working with the treatment"), they tended to experience greater improvements in QOL (BBQ) and complete more treatment modules. This pattern indicates that behaviors reflecting active engagement and self-reflection may play an important role in facilitating positive treatment outcomes.

A total of 515 participant messages were sent. The median number of words per participant was 825 (IQR 530-2110; range 64-5562), and the median number of messages per participant was 11 (IQR 5-18; range 1-75). There was a significant moderate positive correlation between change in BBQ score and the number of participant words (ρ=0.52; *P*=<.001) and the number of participant messages (ρ=0.42; *P*=.02). There was no significant correlation between the number of words written by the therapist and any of the outcome variables.

Table 3. Means and distribution of pre- and posttreatment scores, change scores, and paired samples t test results on outcomes of the participants included in the quantitative analysis (N=29).

Outcome	Pre, mean (SD); range	Post, mean (SD); range	Change score[a], mean (SD)	*t* test (*df*)	*P* value
BBQ[b]	41.14 (21.82); 0-96	45.90 (24.19); 0-96	4.76 (14.46)	–1.77 (28)	.09
SOC-13[c]	48.17 (15.01); 22-90	49.07 (14.49); 21-85	0.90 (9.82)	–0.49 (28)	.63
HADS-D[d]	7.45 (4.03); 0-15	6.83 (5.34); 0-19	–0.62 (3.63)	0.92 (28)	.37
HADS-A[d]	12.31 (4.97); 2-20	11.45 (5.10); 1-21	–0.86 (3.0)	1.55 (28)	.13

[a]Change scores represent posttreatment minus baseline values.

[b]BBQ: Brunnsviken Brief Quality of Life Scale.

[c]SOC-13: 13-item Sense of Coherence scale.

[d]HADS-D and HADS-A: Hospital Anxiety and Depression Scale for depression and anxiety subscales.

XSL•FO
RenderX

Table 4. Correlations (Spearman ρ) between relative frequency of participant behavior and outcome change-scores and module completion (N=29).

Domain and behavioral category	BBQ[a]		SOC-13[b]		HADS-D[c]		HADS-A[c]		Module completed[d]	
	ρ	*P* value	P	*P* value	ρ	*P* value	ρ	*P* value	ρ	*P* value
This is me	0.51	<.001	0.04	.83	–0.21	.27	–0.05	.80	0.45	.01
This is who I am	0.50	.01	0.08	.69	–0.19	.31	–0.02	.91	0.43	.01
My present and past circumstances	0.32	.09	0.00	1.0	–0.21	.27	–0.08	.67	0.21	.24
Change is possible	0.39	.04	0.09	.65	–0.20	.30	0.02	.93	0.37	.03
My point of view	0.53	<.001	0.21	.27	–0.14	.48	–0.04	.84	0.16	.36
Working with the treatment	0.50	.01	0.05	.82	–0.25	.20	0.02	.92	–0.11	.55
Appreciation and treatment alliance	0.44	.02	–0.06	.77	–0.11	.57	0.07	.74	–0.08	.65
Putting treatment aside	0.25	.19	–0.13	.49	–0.34	.07	–0.09	.64	–0.09	.62
Plans to attempt a new task	0.39	.04	–0.02	.92	–0.19	.31	0.12	.53	–0.00	.99
Have attempted a new task	0.41	.03	–0.06	.77	–0.23	.23	–0.03	.89	0.09	.63
Observing positive consequences	0.50	.01	0.12	.53	–0.22	.24	–0.03	.86	0.30	.08
Problems with the treatment	0.27	.16	0.02	.90	–0.14	.46	0.08	.67	0.18	.31
I struggle	0.31	.10	0.03	.88	0.05	.79	0.20	.31	–0.06	.75
Life is and has been demanding	0.32	.09	–0.10	.62	–0.04	.83	0.12	.53	–0.15	.40
Identifies patterns and problem behaviors	0.31	.11	0.12	.55	–0.08	.68	0.00	1.0	0.36	.03
I am troubled by mental ill-health	0.11	.56	–0.00	.99	0.07	.71	0.26	.17	0.32	.06
Maladaptive thoughts	0.14	.47	0.31	.10	0.26	.18	0.20	.30	–0.04	.82

[a]BBQ: Brunnsviken Brief Quality of Life Scale.

[b]SOC-13: 13-item Sense of Coherence scale.

[c]HADS-D and HADS-A: Hospital Anxiety and Depression Scale for depression and anxiety subscales.

[d]Controlled for the total number of words.

Discussion

Principal Findings

The purpose of this study was to explore written communicative behaviors among individuals with autism participating in an internet-based intervention aimed at improving QOL and to examine whether any of these behaviors were associated with treatment outcomes.

The results show that statements related to participants' descriptions of themselves in terms of personality, abilities, beliefs, experiences, and personal development were highly prevalent, with nearly half of the codes (44.1%) categorized under the domain "This is me." The most common category was "This is who I am" (15.5%), which included reflections on personality and autistic functioning. The second most frequent category was "Change is possible" (14.2%), consisting of statements expressing awareness that personal development and growth are possible, such as narratives of past growth or coping strategies perceived as helpful. "Life is and has been demanding," also accounting for a substantial portion (11.6%) of the codes, captured descriptions of adversity, experiences of mistreatment, failure, and mental ill-health.

Correlation analyses revealed that the domains "This is me" and "Working with the treatment," as well as the categories "My point of view," "This is who I am," and "Observing positive consequences of treatment," were significantly moderately associated with improvements in BBQ scores. These findings suggest that the opportunity to reflect openly on identity, personal perspectives, and the perceived impact of treatment may be particularly important for improving QOL in individuals with autism. Notably, the finding regarding "Observing positive consequences of treatment" aligns with previous findings by Soucy et al [18] and Svartvatten et al [11], who also reported a positive correlation between this behavior and treatment outcomes.

The significant moderate positive relation between BBQ scores and the category "Appreciation and treatment alliance" further supports a well-established body of research highlighting the importance of therapeutic alliance for the therapeutic outcomes [9-11,42,43]. In addition, the category "Have attempted a new task" was moderately positively correlated with changes in

XSL•FO
RenderX

BBQ, whereas "Change is possible" and "Plans to attempt a new task" showed weak associations. These findings indicate that motivational expressions related to trying new strategies and behaviors—both within and beyond the treatment context—may reflect treatment engagement and efficacy.

Moreover, the number of words and messages written by participants was also positively associated with improvement in BBQ scores. While earlier studies examining text quantity have primarily focused on its relation to adherence—showing that more words are associated with increased program completion [17,44]—our results indicate that text quantity may also be related to changes in outcomes in ICBT for individuals with autism.

Interestingly, while several communicative behaviors were associated with improvements in QOL (as measured by BBQ), no associations were observed with changes in anxiety or depression (HADS) or sense of coherence (SOC-13). One possible explanation is that QOL, as a subjective and existential construct, may be more readily influenced by participants' self-reflective and narrative expressions. In contrast, clinical symptoms of anxiety and depression, and the more stable SOC, may be less susceptible to short-term fluctuations driven by communicative behaviors. Despite earlier findings suggesting that SOC may improve from interventions [45]—which motivated its use as an outcome in the original RCT—this effect was not found in our study [26]. Another potential explanation is that the relatively greater pre-post variability in BBQ, compared with the smaller mean changes observed in HADS and SOC-13 scores, may have increased the likelihood of detecting statistically significant correlations for QOL but not for the other measures.

Regarding adherence, the domain "This is me" and the category "This is who I am" were moderately associated with the number of completed modules, while the categories "Change is possible" and "Identifies patterns and problem behaviors" showed weak associations. This partly supports previous findings by both Soucy et al [18] and Svartvatten et al [11], who also reported a positive correlation between "Identifies patterns and problem behaviors" and the number of modules completed.

Although no causal inferences can be made, the observed associations in this study may shed light on potential behavioral indicators that are beneficial for individuals with autism to engage in during internet-based psychological interventions.

The relatively large number of identified categories (n=14) illustrates a broad range of content shared by participants with their therapists and likely reflects the wide diversity among individuals with autism. Unlike ICBT, which targets specific disorders, this intervention focused on the individual's whole life, and given the comprehensive scope of the treatment (refer to Table S1 in Multimedia Appendix 1), the extensive result is not surprising. Furthermore, an obvious reason for the relatively high prevalence of specific behaviors (ie, those reflected in the domain "This is me" and the categories "This is who I am" and "Change is possible") is related to the treatment content itself, in which reflections on the self in relation to past, present, and future experiences were repeatedly encouraged.

Consequently, the content of this treatment may partly explain why the behavior categories in the current analysis were only partially similar to those identified in earlier studies [11,18,19]. The 4 categories constituting the domain "This is me," and the category "Life is and has been demanding," including meaning units not covered by earlier frameworks (Table S2 in Multimedia Appendix 1). In agreement with Soucy et al [18], our study supports the notion that a fully deductive process—that is, following the existing coding framework—would have hindered the detection of relevant themes.

Beyond treatment content, these divergent findings (compared with previous studies) must be considered in light of the unique characteristics of the study population—individuals with autism. A considerable proportion of the participant messages centered around identity, functioning, and experiences in relation to autism or psychiatric comorbidity, which were manifested in several ways. Some statements offered nonvaluing descriptions of how their autism contributed to shaping their identity, and others accounted for experiences and traits as contributing to personality formation independent of autistic traits. Regardless of etiology, it appeared fundamental for the participants to construct a narrative around who they are and how their functioning affects their lives. This internalized self-narrative—defining who they are, how they came to be, and where their life may be heading—can be described as their narrative identity [46,47].

Through the construction of a narrative identity, people define and communicate who they are to themselves and others [47,48]. This process fosters self-insight and understanding, allowing negative experiences to be reframed as opportunities for growth, which in turn has been linked to greater well-being [49,50]. Given the precondition of atypical functioning, the process of constructing a coherent self-narrative thus appears particularly relevant for individuals with autism. While research suggests that narrative skills [22,51] and identity formation [52] may be impaired in autism, Samra [53] highlights the importance of engaging with one's narrative identity. In their thesis on identity formation in individuals with autism, they argue that this engagement enhances self-awareness, supports meaning-making, and helps to envision one's future.

Our findings demonstrated a strong engagement in this narrative process, and in line with the suggestion of Samra [53], a positive association was observed between "This is me" statements and improvement in QOL. However, categories within the "I struggle" domain also reflected engagement in a narrative process, but without any positive correlation with QOL. This indicates that different narrative processes may serve different functions, consistent with research arguing that different modes of self-focus have distinct functional properties [54,55]. For example, Watkins et al [54-56] differentiate between maladaptive analytical self-focus—described as abstract, evaluative, and often ruminative—and adaptive experiential self-focus, which is concrete, process-oriented, and linked to positive self-evaluation.

Although we did not deductively categorize statements based on this distinction—maladaptive vs adaptive self-focus—our findings align with the theory and may still be interpreted within

XSL•FO
RenderX

this framework, offering potential guidance for clinicians in identifying and supporting beneficial communicative behaviors.

Limitations

This study design has several limitations. First, only text written via the messaging function of the intervention was included in the analysis; text written directly in response to exercises was excluded, as it was typically constrained by predefined questions and exercises, making it less relevant for the aim of identifying spontaneous communicative behaviors.

It should be noted that the participants included in this study are not fully representative of the entire autistic population, as a natural selection of individuals already familiar with technology applies for inclusion in a trial such as this. Moreover, participants who did not write anything were excluded from the current analyses, which may imply a certain selection bias, as these individuals might have refrained from messaging due to communication difficulties. Those failing to complete the postassessment were also excluded from the correlation analysis, which means that we are unable to conclude the relationship between behavior and outcome for these individuals. However, post hoc analyses with imputed zeros modeling noncorrespondence, as well as imputed zeros for change scores on quantitative outcomes, were conducted, showing that these procedures did not affect the correlational outcomes (data not shown).

Another limitation concerns the role of the primary analyst (BW), who also served as a therapist and was involved in developing the intervention. This dual role introduces a risk of bias, potentially influencing data interpretation, coding, and categorization. To minimize this bias, the initial phase of the analysis was conducted jointly with KJ, who had no previous involvement in treatment development or delivery. Furthermore, the coding approach was deliberately inductive and aimed to be as close to the text as possible. Nevertheless, future studies would benefit from involving fully independent third-party analysts to further strengthen the trustworthiness of the findings.

Another methodological limitation in this study is that the intercoder reliability was not statistically calculated. Typically, around 10%-25% of data units are coded by more than 1 researcher to facilitate a trustworthy estimate of intercoder reliability [57]. By collaborative coding of 5 (14.71%) of the data units in this study, the analysts could compare and discuss codes to develop a preliminary coding framework. By subsequently coding 1 of the participants' texts independently, the coding framework was further refined and agreed upon. We considered this to be sufficient to establish the final coding framework, but acknowledge that the reliability would be strengthened with additional collaborative coding. Furthermore, in future studies, interrater reliability should be assessed using a statistical test, for example, Cohen κ.

As our analysis did not account for the timing of the statements, it is uncertain whether any detection of potentially beneficial behaviors can be made at an early stage in the therapy. Early detection (or absence) of these behaviors could enable more individualized therapeutic support, such as encouraging these behaviors. Given the novelty of this research area and the fact that similar data have not previously been studied, we chose in this study to focus on broader behavioral patterns rather than their temporal distribution. Future research could consider focusing on early-stage behaviors as potential indicators of treatment response.

When interpreting the results from the correlation analyses, it should be noted that no correction for multiple comparisons was applied. Given the relatively large number of behavioral categories and correlations tested, there is an increased risk of inflated false positives, and the results should be interpreted with caution.

Finally, in the correlational analyses between behaviors and outcome, we did not control for therapist word count or therapist behaviors. Given previous research emphasizing the importance of therapist support and the varying number of therapist messages, therapist behavior may have influenced the relationships. However, since the therapist's word count was not directly correlated with outcomes in this study, it was not included as a covariate. Future studies should nevertheless consider therapist variables when exploring outcome predictors in ICBT for individuals with autism.

Conclusions

The findings of this study increase our understanding of how individuals with autism engage in ICBT and may serve as a valuable source of information for therapists in guiding both their expectations of client outcomes and identifying client behaviors to support. The results emphasize the importance of self-narrative formulation in individuals with autism and suggest that certain communicative behaviors—especially those involving identity reflection and recognition of treatment benefits—are positively related to improvements in QOL. Future research should explore ways to identify such behaviors early in treatment as potential indicators of individuals at risk of poorer outcomes, enabling clinicians to provide additional support or tailor interventions to better suit individual needs.

Acknowledgments

We thank all participants who contributed their text-based data for this study. We also thank associate professor Yang Cao, Örebro University, for guidance regarding the statistical analyses. The generative artificial intelligence (GAI) tool ChatGPT (OpenAI) was used for language editing by suggesting alternative phrasings of long and/or complex sentences or terms to improve clarity and readability. According to the GAIDeT taxonomy, the following tasks were delegated to GAI tools under full human supervision: proofreading and editing.

XSL•FO
RenderX

Funding

This work was financially supported by the Region Örebro Län, Sweden (grant nos OLL-935396, OLL-879651, OLL-887401, OLL-833131, OLL-785501, OLL-736321, OLL-878311, OLL-785311, OLL-985129, and OLL-1004448) and Stiftelsen Bror Gadelius Minnesfond (grant no 23003). The funding sources had no role in the study design, data collection, analysis, interpretation, writing of the report, or the decision to submit the article for publication.

Data Availability

The data supporting this study originates from a longitudinal intervention study conducted at the University Health Care Research Center, Örebro Region, Sweden. Data are available upon reasonable request from the corresponding author.

Authors' Contributions

BW led the conceptualization of the study, with FH providing supporting input. BW was responsible for data curation. Formal analysis was conducted by BW as the lead analyst, with KJ providing support. Funding acquisition was led by BW, with SB and FH providing supporting contributions. BW carried out the investigation. Methodology was led by BW, with FH supporting. Project administration was managed by BW. Supervision was led by FH, with SB providing support. BW drafted the original manuscript. Review and editing of the manuscript were led by BW, with FH, SB, KJ, and MUH providing supporting contributions.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Treatment modules in the internet-based intervention MILAS: overview of domains, categories, definitions, example codes, and quotes from this study and, where applicable, the corresponding categories from previous studies.
[DOCX File , 37 KB - jmir_v28i1e76527_app1.docx]

References

1. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). Arlington, VA: Americal Psychiatric Association; 2013.
2. Pascoe MI, Forbes K, de la Roche L, Derby B, Psaradellis E, Anagnostou E, et al. Exploring the association between social skills struggles and social communication difficulties and depression in youth with autism spectrum disorder. Autism Res 2023;16(11):2160-2171. [doi: 10.1002/aur.3015] [Medline: 37615265]
3. Rai D, Culpin I, Heuvelman H, Magnusson CMK, Carpenter P, Jones HJ, et al. Association of autistic traits with depression from childhood to age 18 years. JAMA Psychiatry 2018;75(8):835-843 [FREE Full text] [doi: 10.1001/jamapsychiatry.2018.1323] [Medline: 29898212]
4. Mason D, McConachie H, Garland D, Petrou A, Rodgers J, Parr JR. Predictors of quality of life for autistic adults. Autism Res 2018;11(8):1138-1147 [FREE Full text] [doi: 10.1002/aur.1965] [Medline: 29734506]
5. Adams D, Young K. A systematic review of the perceived barriers and facilitators to accessing psychological treatment for mental health problems in individuals on the autism spectrum. Rev J Autism Dev Disord 2020;8(4):436-453. [doi: 10.1007/s40489-020-00226-7]
6. Andersson G. Internet interventions: past, present and future. Internet Interv 2018;12:181-188 [FREE Full text] [doi: 10.1016/j.invent.2018.03.008] [Medline: 30135782]
7. Richards D, Enrique A, Palacios J, Duffy D. Internet-delivered cognitive behaviour therapy. In: Cognitive Behavioral Therapy and Clinical Applications. London, United Kingdom: IntechOpen; 2018.
8. Furukawa TA, Suganuma A, Ostinelli EG, Andersson G, Beevers CG, Shumake J, et al. Dismantling, optimising, and personalising internet cognitive behavioural therapy for depression: a systematic review and component network meta-analysis using individual participant data. Lancet Psychiatry 2021;8(6):500-511 [FREE Full text] [doi: 10.1016/S2215-0366(21)00077-8] [Medline: 33957075]
9. Haller K, Becker P, Niemeyer H, Boettcher J. Who benefits from guided internet-based interventions? A systematic review of predictors and moderators of treatment outcome. Internet Interv 2023;33:100635 [FREE Full text] [doi: 10.1016/j.invent.2023.100635] [Medline: 37449052]
10. Kaiser J, Hanschmidt F, Kersting A. The association between therapeutic alliance and outcome in internet-based psychological interventions: a meta-analysis. Comput Hum Behav 2021;114:106512. [doi: 10.1016/j.chb.2020.106512]
11. Svartvatten N, Segerlund M, Dennhag I, Andersson G, Carlbring P. A content analysis of client e-mails in guided internet-based cognitive behavior therapy for depression. Internet Interv 2015;2(2):121-127. [doi: 10.1016/j.invent.2015.02.004]
12. Schneider LH, Hadjistavropoulos HD, Faller YN. Internet-delivered cognitive behaviour therapy for depressive symptoms: an exploratory examination of therapist behaviours and their relationship to outcome and therapeutic alliance. Behav Cogn Psychother 2016;44(6):625-639. [doi: 10.1017/S1352465816000254] [Medline: 27302220]

XSL•FO
RenderX

13. Holländare F, Gustafsson SA, Berglind M, Grape F, Carlbring P, Andersson G, et al. Therapist behaviours in internet-based cognitive behaviour therapy (ICBT) for depressive symptoms. Internet Interv 2016;3:1-7 [FREE Full text] [doi: 10.1016/j.invent.2015.11.002] [Medline: 30135783]
14. Paxling B, Lundgren S, Norman A, Almlöv J, Carlbring P, Cuijpers P, et al. Therapist behaviours in internet-delivered cognitive behaviour therapy: analyses of e-mail correspondence in the treatment of generalized anxiety disorder. Behav Cogn Psychother 2013;41(3):280-289. [doi: 10.1017/S1352465812000240] [Medline: 22717145]
15. Dirkse D, Hadjistavropoulos HD, Hesser H, Barak A. Linguistic analysis of communication in therapist-assisted internet-delivered cognitive behavior therapy for generalized anxiety disorder. Cogn Behav Ther 2015;44(1):21-32. [doi: 10.1080/16506073.2014.952773] [Medline: 25244051]
16. Van der Zanden R, Curie K, Van Londen M, Kramer J, Steen G, Cuijpers P. Web-based depression treatment: associations of clients' word use with adherence and outcome. J Affect Disord 2014;160:10-13 [FREE Full text] [doi: 10.1016/j.jad.2014.01.005] [Medline: 24709016]
17. Linnet J, Jensen ES, Runge E, Hansen MB, Hertz SPT, Mathiasen K, et al. Text based internet intervention of binge eating disorder (BED): words per message is associated with treatment adherence. Internet Interv 2022;28:100538 [FREE Full text] [doi: 10.1016/j.invent.2022.100538] [Medline: 35480237]
18. Soucy JN, Hadjistavropoulos HD, Couture CA, Owens VA, Dear BF, Titov N. Content of client emails in internet-delivered cognitive behaviour therapy: a comparison between two trials and relationship to client outcome. Internet Interv 2018;11:53-59 [FREE Full text] [doi: 10.1016/j.invent.2018.01.006] [Medline: 30135760]
19. Kraepelien M, Hadjistavropoulos HD, Berman AH, Sundström C. Exploring client messages in a therapist-guided internet intervention for alcohol use disorders - a content analysis. Internet Interv 2021;26:100483 [FREE Full text] [doi: 10.1016/j.invent.2021.100483] [Medline: 34824984]
20. Rollins PR. Narrative skills in young adults with high-functioning autism spectrum disorders. Commun Disord Q 2014;36(1):21-28. [doi: 10.1177/1525740114520962]
21. Gillespie-Lynch K, Kapp SK, Shane-Simpson C, Smith DS, Hutman T. Intersections between the autism spectrum and the internet: perceived benefits and preferred functions of computer-mediated communication. Intellect Dev Disabil 2014;52(6):456-469. [doi: 10.1352/1934-9556-52.6.456] [Medline: 25409132]
22. Nayar K, Landau E, Martin GE, Stevens CJ, Xing J, Sophia P, et al. Narrative ability in autism and first-degree relatives. J Autism Dev Disord 2025;55(11):3822-3837. [doi: 10.1007/s10803-024-06424-0] [Medline: 39060703]
23. Howard PL, Sedgewick F. 'Anything but the phone!': communication mode preferences in the autism community. Autism 2021;25(8):2265-2278 [FREE Full text] [doi: 10.1177/13623613211014995] [Medline: 34169750]
24. van der Aa C, Pollmann MMH, Plaat A, van der Gaag RJ. Computer-mediated communication in adults with high-functioning autism spectrum disorders and controls. Res Autism Spectr Disord 2016;23:15-27. [doi: 10.1016/j.rasd.2015.11.007]
25. Caldwell-Harris CL, Posner SD. When autistic writing is superior to neurotypical writing: the case of blogs. Educ Rev 2024;76(7):1875-1897. [doi: 10.1080/00131911.2024.2302119]
26. Westerberg B, Holländare F, Bejerot S. An internet-based behavioral intervention for adults with autism spectrum disorder - a randomized controlled trial and feasibility study. Internet Interv 2023;34:100672 [FREE Full text] [doi: 10.1016/j.invent.2023.100672] [Medline: 37772160]
27. Backman A, Roll-Pettersson L, Mellblom A, Norman-Claesson E, Sundqvist E, Zander E, et al. Internet-delivered psychoeducation (SCOPE) for transition-aged autistic youth: pragmatic randomized controlled trial. J Med Internet Res 2024;26:e49305 [FREE Full text] [doi: 10.2196/49305] [Medline: 39608000]
28. Conaughton RJ, Donovan CL, March S. Efficacy of an internet-based CBT program for children with comorbid high functioning autism spectrum disorder and anxiety: a randomised controlled trial. J Affect Disord 2017;218:260-268. [doi: 10.1016/J.JAD.2017.04.032]
29. Georén L, Jansson-Fröjmark M, Nordenstam L, Andersson G, Olsson NC. Internet-delivered cognitive behavioral therapy for insomnia in youth with autism spectrum disorder: a pilot study. Internet Interv 2022;29:100548 [FREE Full text] [doi: 10.1016/j.invent.2022.100548] [Medline: 35651733]
30. Wickberg F, Lenhard F, Aspvall K, Serlachius E, Andrén P, Johansson F, et al. Feasibility of internet-delivered cognitive-behavior therapy for obsessive-compulsive disorder in youth with autism spectrum disorder: a clinical benchmark study. Internet Interv 2022;28:100520 [FREE Full text] [doi: 10.1016/j.invent.2022.100520] [Medline: 35281701]
31. Westerberg B, Bäärnhielm S, Giles C, Hylén U, Holländare F, Bejerot S. An internet based intervention for adults with autism spectrum disorder-a qualitative study of participants experiences. Front Psychiatry 2021;12:789663 [FREE Full text] [doi: 10.3389/fpsyt.2021.789663] [Medline: 35002808]
32. Bejerot S. ALMA - KBT för vuxna med autismspektrumsyndrom, manual och arbetsbok [ALMA-CBT for adults with autism spectrum disorder, Manual and Workbook]. United States: Hogrefe Publishing; 2019.
33. Hesselmark E, Plenty S, Bejerot S. Group cognitive behavioural therapy and group recreational activity for adults with autism spectrum disorders: a preliminary randomized controlled trial. Autism 2014;18(6):672-683 [FREE Full text] [doi: 10.1177/1362361313493681] [Medline: 24089423]

XSL•FO
RenderX

34. Kerns CM, Roux AM, Connell JE, Shattuck PT. Adapting cognitive behavioral techniques to address anxiety and depression in cognitively able emerging adults on the autism spectrum. Cogn Behav Pract 2016;23(3):329-340. [doi: 10.1016/j.cbpra.2016.06.002]
35. Lindner P, Frykheden O, Forsström D, Andersson E, Ljótsson B, Hedman E, et al. The Brunnsviken Brief Quality of life scale (BBQ): development and psychometric evaluation. Cogn Behav Ther 2016;45(3):182-195 [FREE Full text] [doi: 10.1080/16506073.2016.1143526] [Medline: 26886248]
36. Antonovsky A. The structure and properties of the Sense of Coherence scale. Soc Sci Med 1993;36(6):725-733. [doi: 10.1016/0277-9536(93)90033-z] [Medline: 8480217]
37. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand 1983;67(6):361-370. [doi: 10.1111/j.1600-0447.1983.tb09716.x] [Medline: 6880820]
38. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
39. Råholm MB. Abductive reasoning and the formation of scientific knowledge within nursing research: abductive reasoning within nursing research. Nurs Philos 2010;11(4):260-270. [doi: 10.1111/j.1466-769x.2010.00457.x]
40. Nvivo. Lumivero. 2023. URL: https://lumivero.com/ [accessed 2026-01-06]
41. IBM SPSS Statistics. IBM. 2023. URL: https://www.ibm.com/products/spss-statistics [accessed 2026-01-09]
42. Lindqvist K, Mechler J, Falkenström F, Carlbring P, Andersson G, Philips B. Therapeutic alliance is calming and curing-the interplay between alliance and emotion regulation as predictors of outcome in internet-based treatments for adolescent depression. J Consult Clin Psychol 2023;91(7):426-437. [doi: 10.1037/ccp0000815] [Medline: 37166833]
43. Bergman Nordgren L, Carlbring P, Linna E, Andersson G. Role of the working alliance on treatment outcome in tailored internet-based cognitive behavioural therapy for anxiety disorders: randomized controlled pilot trial. JMIR Res Protoc 2013;2(1):e4 [FREE Full text] [doi: 10.2196/resprot.2292] [Medline: 23612437]
44. Wallert J, Gustafson E, Held C, Madison G, Norlund F, von Essen L, et al. Predicting adherence to internet-delivered psychotherapy for symptoms of depression and anxiety after myocardial infarction: machine learning insights from the U-CARE heart randomized controlled trial. J Med Internet Res 2018;20(10):e10754 [FREE Full text] [doi: 10.2196/10754] [Medline: 30305255]
45. Valtonen M, Raiskila T, Veijola J, Läksy K, Kauhanen M, Kiuttu J, et al. Enhancing sense of coherence via early intervention among depressed occupational health care clients. Nord J Psychiatry 2015;69(7):515-522. [doi: 10.3109/08039488.2015.1011230] [Medline: 25739527]
46. McAdams DP. Narrative identity. In: Schwartz S, Luyckx K, Vignoles V, editors. Handbook of Identity Theory and Research. New York, NY: Springer; 2011:99-115.
47. McAdams DP, McLean KC. Narrative identity. Curr Dir Psychol Sci 2013;22(3):233-238. [doi: 10.1177/0963721413475622]
48. McAdams DP. Narrative identity: what is it? What does it do? How do you measure it? Imagin Cogn Personal 2018;37(3):359-372. [doi: 10.1177/0276236618756704]
49. Lilgendahl JP, McAdams DP. Constructing stories of self-growth: how individual differences in patterns of autobiographical reasoning relate to well-being in midlife. J Pers 2011;79(2):391-428 [FREE Full text] [doi: 10.1111/j.1467-6494.2010.00688.x] [Medline: 21395593]
50. Waters TEA, Fivush R. Relations between narrative coherence, identity, and psychological well-being in emerging adulthood. J Pers 2015;83(4):441-451 [FREE Full text] [doi: 10.1111/jopy.12120] [Medline: 25110125]
51. McCabe A, Hillier A, Shapiro C. Brief report: structure of personal narratives of adults with autism spectrum disorder. J Autism Dev Disord 2013;43(3):733-738. [doi: 10.1007/s10803-012-1585-x] [Medline: 22767138]
52. Allé MC, Schneider P, Rigoulot L, Gandolphe MC, Danion JM, Coutelle R, et al. Narrative identity alterations in autism spectrum disorder: a life story approach. Research Square Preprint posted online on April 30, 2024. [doi: 10.21203/rs.3.rs-4292403/v1]
53. Samra HS. A narrative exploration of sense-making, self, and identity in young people diagnosed with an autism spectrum condition [doctoral dissertation]. School of Education, University of Birmingham. 2016. URL: https://etheses.bham.ac.uk/id/eprint/6719/1/Samra16EdPsychD.pdf [accessed 2026-01-20]
54. Vassilopoulos SP, Watkins ER. Adaptive and maladaptive self-focus: a pilot extension study with individuals high and low in fear of negative evaluation. Behav Ther 2009;40(2):181-189. [doi: 10.1016/j.beth.2008.05.003] [Medline: 19433149]
55. Watkins E, Teasdale JD. Adaptive and maladaptive self-focus in depression. J Affect Disord 2004;82(1):1-8. [doi: 10.1016/j.jad.2003.10.006] [Medline: 15465571]
56. Watkins E. Adaptive and maladaptive ruminative self-focus during emotional processing. Behav Res Ther 2004;42(9):1037-1052. [doi: 10.1016/j.brat.2004.01.009] [Medline: 15325900]
57. O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. Int J Qual Methods 2020;19:160940691989922. [doi: 10.1177/1609406919899220]

Abbreviations

BBQ: Brunnsviken Brief Quality of Life Scale

XSL•FO
RenderX

HADS: Hospital Anxiety and Depression Scale
HADS-A: Hospital Anxiety and Depression Scale for the anxiety subscale
HADS-D: Hospital Anxiety and depression Scale for the depression subscale
ICBT: internet-based cognitive behavioral therapy
QOL: quality of life
RCT: randomized controlled trial
SOC: sense of coherence
SOC-13: 13-item Sense of Coherence scale

Edited by A Stone; submitted 19.May.2025; peer-reviewed by J Wang, C Abbatantuono, E Zander; comments to author 07.Nov.2025; revised version received 30.Dec.2025; accepted 30.Dec.2025; published 04.Feb.2026.

Please cite as:
Westerberg B, Jacobson K, Unenge Hallerbäck M, Bejerot S, Holländare F
Communicative Behaviors in an Internet-Based Intervention for Individuals With Autism: Mixed Methods Analysis
J Med Internet Res 2026;28:e76527
URL: https://www.jmir.org/2026/1/e76527
doi:10.2196/76527
PMID:

©Britta Westerberg, Karin Jacobson, Maria Unenge Hallerbäck, Susanne Bejerot, Fredrik Holländare. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Initial Insights Into an Institutional Secure Large Language Model for Magnetic Resonance Imaging Examination Requests: Retrospective Study

James Thomas Patrick Decourcy Hallinan[1,2], MBChB; Naomi Wenxin Leow[3], BComp, MComp; Yi Xian Low[1], MBBS; Aric Lee[1], MBBS; Wilson Ong[1], MBBS; Matthew Ding Zhou Chan[1], BmedSc, MD; Ganakirthana Kalpenya Devi[1], MbBchBao; Stephanie Shengjie He[1], MBBS; Daniel De-Liang Loh[1], MBBS, MRCS; Desmond Shi Wei Lim[1], MBBS; Xi Zhen Low[1], MBBS; Mei Chin Lim[1], MBBS; Clement Yong[1], MBBS; Weizhong Jonathan Sng[1], MBBS; Ee Chin Teo[1], MMRT; Jiong Hao Tan[4], MBBS, MMed, MRCS; Naresh Kumar[4], MBBS, MS, DNB, DM; Andrew Makmur[1,3], BmedSc, MBBS, MMed; Yonghan Ting[1], MBBS

[1]Department of Diagnostic Imaging, National University Hospital, 5 Lower Kent Ridge Rd, Singapore, Singapore

[2]Department of Diagnostic Radiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[3]Innovation Office, AI Office, National University Health System, Singapore, Singapore

[4]National University Spine Institute, Department of Orthopedic Surgery, National University Hospital, Singapore, Singapore

Corresponding Author:
James Thomas Patrick Decourcy Hallinan, MBChB
Department of Diagnostic Imaging, National University Hospital, 5 Lower Kent Ridge Rd, Singapore, Singapore

Abstract

Background: Incomplete clinical details on magnetic resonance imaging (MRI) examination requests (MERs) can lead to suboptimal protocol selection. An institutional secure large language model (sLLM) with access to manually retrieved salient data from the electronic medical record (EMR) may improve request completeness and protocol accuracy across multiple MRI subspecialties.

Objective: The objective of this study was to compare clinician MERs with sLLM-augmented MERs for information quality and to evaluate the protocoling accuracy of the sLLM versus board-certified radiologists across body, musculoskeletal, and neuroradiology MRI.

Methods: This retrospective study included 608 random outpatient MRI examinations performed between September 2023 and July 2024 (body 206, musculoskeletal 203, neuroradiology 199). The cohort comprised 528 patients (mean 51.2 years, SD 19.2; range 4 - 93; n=279, 52.8% women, n=249, 47.2% men). MERs without EMR access were excluded. A privately hosted Anthropic Claude 3.5 model (temperature 0) augmented each MER with manually retrieved salient EMR data and, via rule-based parsing, mapped the extracted elements onto predefined institutional criteria to recommend region or coverage and contrast use. Two experienced radiologists established a consensus reference standard. Two board-certified general radiologists (Rad 3 and Rad 4) and the sLLM were compared with this standard. Clinical information quality was graded using the Reason-for-Exam Imaging Reporting and Data System (RI-RADS). Interrater reliability was quantified with Gwet AC1. Paired accuracies were compared with the McNemar test to determine whether there was a statistically significant difference.

Results: Interreader agreement for RI-RADS was almost perfect for sLLM-augmented MERs (AC1 0.97, 95% CI 0.94 - 0.99) and moderate for clinician MERs (AC1 0.43, 95% CI 0.34 - 0.52). Limited or deficient clinical information (RI-RADS C/D) fell to 0% to 0.7% (0/608 to 4/608) with sLLM augmentation vs 4.1% to 20.4% (25/608 to 124/608) for clinician MERs. Overall protocol accuracy was 93.1% (566/608; 95% CI 89.6 - 96.6) for the sLLM, 91.4% (556/608; 95% CI 87.6 - 95.3) for Rad 3, and 92.1% (560/608; 95% CI 88.4 - 95.8) for Rad 4 (sLLM vs Rad 3 P=.23 vs Rad 4 P=.40). Region or coverage accuracy was similar (sLLM: 579/608, 95.2%; Rad 3: 585/608, 96.2%; Rad 4: 573/608, 94.2%; P=.46 and P=.36). Contrast decisions were more accurate using the sLLM at 94.4% (574/608; 95% CI 91.3 - 97.5) vs Rad 3 at 92.1% (560/608; 95% CI 88.4 - 95.8; P=.027) and were not significantly different to Rad 4 at 92.9% (565/608; 95% CI 89.4 - 96.4; P=.16). Subspecialty analyses showed similar patterns, with the sLLM outperforming Rad 4 for musculoskeletal MRI contrast decisions (96.6% vs 91.1%; P=.006) and matching readers elsewhere. Manual review indicated that sLLM improvements arose from EMR details not listed on the MER (infection/inflammation, tumor history, prior surgery). No clinically significant hallucinations were identified in a manual review of discordant cases.

Conclusions: Across body, musculoskeletal, and neuroradiology MRI, sLLM-augmented examination requests improved clinical context and enhanced contrast selection while demonstrating accuracy comparable to general radiologists for region or coverage.

XSL•FO
RenderX

Integrating sLLMs into routine vetting workflows may reduce manual workload in protocol selection for more efficient, standardized protocoling.

(J Med Internet Res 2026;28:e82579) doi:10.2196/82579

KEYWORDS

secure large language model; radiology request form; reason for exam imaging reporting and data system; musculoskeletal imaging; body imaging; neuroradiology imaging; magnetic resonance imaging

Introduction

Radiology examination requests form the primary communication bridge between referring clinicians and radiology services. Completeness and clarity of these requests influence protocol selection, image quality, and ultimately diagnostic and therapeutic outcomes [1-3]. When key clinical elements such as relevant history, current symptoms, or prior imaging are incomplete or absent, radiographers and radiologists may assign a suboptimal protocol or omit intravenous contrast, leading to lower accuracy, repeat studies, and additional cost [4-6]. A recent systematic review showed that incomplete clinical information was associated with reduced reporting accuracy, clinical relevance, and reporting confidence [5].

The enhancement of radiology examination requests with data from the electronic medical record (EMR) offers a potential solution. Early rule-based decision-support tools reduced certain categories of inappropriate imaging yet may struggle to accommodate unstructured narratives and updated imaging protocols [6]. Natural language processing pipelines improved specificity but required extensive engineering and did not fully interpret clinical context [7]. Current large language models (LLMs) can ingest heterogeneous text, interpret medical terminology, and generate structured outputs at scale [8-13]. Proof-of-concept work has already demonstrated utility for report generation and national guideline concordance [14,15]. Nevertheless, many published studies rely on deidentified vignettes or synthetic notes, limiting direct clinical translation because of privacy constraints.

Secure LLMs (sLLMs) address these constraints by operating behind institutional firewalls while preventing the transmission of protected health information [16]. Several studies have demonstrated their feasibility, including improving the adequacy of spine magnetic resonance imaging (MRI) request forms and increased protocol concordance for musculoskeletal studies [17,18]. Recent oncologic imaging work further indicates that a GPT-4 system can automatically generate structured clinical histories that radiologists prefer over original clinician-generated requisitions [19]. Another recent study on GPT-4-generated MRI protocols showed notable quality in cardiac and neuroradiology imaging [20]. An accompanying editorial emphasized that fatigue-free, verifiable LLM summaries could finally bridge the long-recognized information gap between referrers and radiologists, reinforcing the case for secure, institution-hosted models [21].

Building on these findings, we evaluated a secure LLM across all routine MRI subspecialties, including body, neuroradiology, and musculoskeletal examinations. The study had 2 objectives: (1) to compare the information quality of clinician MRI examination requests (MERs) with those augmented by the sLLM using the Reason-for-Exam Imaging Reporting and Data System (RI-RADS) [22] and (2) to assess protocol accuracy of the sLLM against both subspecialty and general radiologist readers.

Methods

Ethical Considerations

The Institutional Review Board (Domain-Specific Review Board reference: 2023/00919) classified the project as minimal risk and therefore granted a waiver of informed consent. Patient-identifying details have been omitted to maintain the privacy and confidentiality of patient data. Compensation was not required in view of the minimal risk of the project.

Protocoling Pipeline and sLLM Prompt

Original clinician-submitted outpatient MERs performed between September 2023 and July 2024 were retrieved at random. MERs lacking corresponding EMR information were excluded. The musculoskeletal cohort in this study represents a distinct nonoverlapping dataset from that of a previous study by the authors. A target sample of approximately 600 requests (~200 per major subspecialty) was selected to provide adequate statistical power while maintaining operational feasibility across a diverse range of anatomical regions and pathologies. For every patient, a MER was generated by the sLLM based on the clinician request with relevant EMR content.

The institutional sLLM is a privately hosted instance of Anthropic Claude 3.5 on Amazon Bedrock. Model temperature was fixed at 0 to minimize output variability. For each request, the sLLM received (1) the most recent relevant clinical entry identified manually by the authors, based on the clinical service indicated on the request forms and (2) pertinent prior imaging reports (eg, CT neck for an MRI nasopharynx study). The typical input length of the clinical notes provided to the sLLM ranged from around 100 words for routine or straightforward presentations to around 3000 words for patients with complex oncological or multisystem disease. The sLLM was also given the latest institutional MRI protocol repository (Multimedia Appendix 1). No clinical data left the institutional firewall.

Two-step sLLM protocoling was performed:

1. Information enrichment—the sLLM extracted key findings, working diagnoses, and potential MRI safety issues from both the MER and the EMR, returning a concise justification summary.
2. Protocol assignment—using institutional rules and the enriched summary, the model selected the optimal

XSL•FO
RenderX

region/coverage and determined whether contrast was required.

The estimated processing cost using the sLLM was US $3 per 1,000,000 tokens for the input and US $15 per 1,000,000 tokens for the output, giving a total cost of US $0.024 to 0.033 per request. For the cases requiring longer input length (about 5000 tokens), the per-case cost remained low at a few cents per request.

Importantly, the sLLM's primary role in this pipeline was clinical information extraction, summarizing clinically relevant details from the MER and EMR (eg, prior surgery, suspected infection, etc) and contextual interpretation. The final contrast and region decisions were not made autonomously by the sLLM. Instead, the extracted elements were subsequently mapped onto predefined institutional criteria for region or coverage and contrast use using a deterministic rule-based parsing script.

Contrast (gadolinium) administration was considered when any of the following were present:

1. History or suspicion of tumor, malignancy, or focal lesion
2. Known or suspected infective or inflammatory condition
3. Injury to neural structures
4. Previous surgery or spinal/extremity instrumentation
5. Explicit clinician request

A custom parsing script evaluated the sLLM response, which listed a "yes" or "no" decision for each contrast-relevant category together with the reason when applicable. If the script detected at least one positive flag and renal failure or other contraindications had not been recognized, the examination was classified as "contrast required."

Detailed prompt wording and rule logic are provided in Multimedia Appendix 1. The rule definitions in this study were refined through prior work on LLMs in spine and musculoskeletal protocoling. All existing MRI protocols (including prompt design and contrast rules) were carefully reviewed with subspecialty leads and key team members to ensure alignment with institutional practice before doing the formal analysis. The full institutional MRI protocol table had a size compatible with the available context window and was embedded directly within the prompt, obviating the need for retrieval-augmented generation (RAG) in this study. No cases from the present study were used during prompt development.

Evaluation Procedure

Two senior radiologists (Rad 1: 14 years of experience; Rad 2: 12 years of experience) independently graded the clinical adequacy of both clinician and sLLM MERs using the RI-RADS classification (Multimedia Appendix 2). They were blinded to the origin of each form.

Rad 1 and Rad 2 provided a consensus reference standard for protocoling. Against this consensus reference standard, protocol selections generated by the sLLM were then compared with those from 2 board-certified general radiologists (Rad 3 and Rad 4), who also had access to the same EMR entries, pertinent prior imaging reports, and institutional MRI protocols. The pertinent prior imaging reports were manually selected by study members, who were individuals distinct from Rad 1 and Rad 2. Both Rad 3 and Rad 4 had 2 years of experience and were commonly tasked with providing protocols based on MRI examination requests. Accuracy was credited only when the suggested region or coverage and contrast decision matched the consensus reference standard. Additional subanalyses were performed for each component and the major subspecialties. A manual review of disagreements between the sLLM and board-certified general radiologists against the consensus reference standard was carried out by the senior radiologists. Note was made of any hallucinations by the sLLM with clinically significant hallucinations defined as the sLLM producing clinically relevant details absent from the provided input data and capable of influencing protocol decisions [23].

Statistical Analysis

All computations were performed with Python 3.9.12. Two-sided tests were considered significant at $P<.05$.

The quality of each MER was rated on a 4-level scale (RI-RADS grades A-D) by Rad 1 and Rad 2. Interreader agreement for these ordinal ratings, and later for MRI protocoling decisions (region/coverage and contrast requirement), was quantified with Gwet AC1, which is less sensitive than Cohen κ to category imbalance (eg, the predominance of A/B grades) [24]. Agreement was interpreted as poor (<0), slight (0 - 0.20), fair (0.21 - 0.40), moderate (0.41 - 0.60), substantial (0.61 - 0.80), or almost perfect (0.81 - 1). With more than 600 examinations, the study had greater than 90% power to detect a minimum 15% absolute difference in the proportion of clinically adequate MERs between clinician and sLLM versions.

Protocol selections generated by the sLLM and by 2 board-certified general radiologists (Rad 3 and Rad 4) were compared with the consensus reference standard established by Rad 1 and Rad 2. Overall accuracy was defined as the proportion of cases in which both region or coverage and contrast indication matched the reference standard. Differences in paired accuracies (sLLM vs each radiologist) were evaluated using the McNemar test with Yates continuity correction. Separate analyses were undertaken for region alone and contrast alone, and across the 3 major subspecialties (body, musculoskeletal, and neuroradiology). Ninety-five percent CIs for RI-RADS gradings (quality of clinical information) and protocoling were derived from the normal approximation to the binomial distribution. For proportions near 0% or 100%, the normal approximation to the binomial can be inaccurate. Therefore, for low-event RI-RADS C/D proportions (eg, sLLM MERs), 95% CIs were computed using the Clopper-Pearson exact method rather than the normal approximation.

Results

Patient Demographics and Consensus MRI Protocols

Overall, 608 MRI examination requests were collected from 528 patients (mean 51.2 y, SD 19.2; range 4 - 93 y). In total, 52.8% (279/528) patients were women and 47.2% (249/528) were men (Table 1). An additional 27 MRI examinations (26 patients) were excluded from analysis due to incomplete EMRs (eg, external referrals) (Figure 1).

XSL•FO
RenderX

Table . Patient and magnetic resonance imaging examination characteristics.

Characteristics	Values
Age (y), mean (SD; range)	
All (N=528)	51.2 (SD 19.2; 4-93)
Women (n=279)	52.9 (SD 17.4; 4-93)
Men (n=249)	49.3 (SD 21; 5-92)
MRI[a] study specialties (N=608), n (%)	
Body MRI (n=206)	
Rectum/perineum fistula	28 (13.6)
Uterus/cervix cancer	26 (12.6)
Liver routine	21 (10.2)
Prostate routine	21 (10.2)
Enterography	16 (7.8)
Pancreas routine	14 (6.8)
Other	80 (38.8)
Musculoskeletal MRI (n=203)	
Lumbar spine	25 (12.3)
Cervical spine	24 (11.8)
Shoulder	29 (14.3)
Knee	27 (13.3)
Pelvis	11 (5.4)
Other	87 (42.9)
Neuroradiology MRI (n=199)	
Brain + contrast	40 (20.1)
Orbits	18 (9)
Skull base/temporal bones	15 (7.5)
Brain routine (noncontrast)	15 (7.5)
Brain stroke (acute)	14 (7)
Brain MR angiography	12 (6)
Pituitary dynamic	12 (6)
Oral cavity/neck	12 (6)
Nasopharynx and neck	10 (5)
Other	51 (25.6)
MRI study type (N=608), n (%)	
Routine (noncontrast)	239 (39.3)
Contrast	369 (60.7)
Contrast use by specialty	
Body MRI	Contrast 175 (85); noncontrast 31 (15)
Musculoskeletal MRI	Contrast 55 (27); noncontrast 148 (73)
Neuroradiology MRI	Contrast 139 (70); noncontrast 60 (30)

[a]MRI: magnetic resonance imaging.

XSL•FO
RenderX

Figure 1. Study design flowchart. The initial clinician-generated MERs were extracted and then augmented using an sLLM. The clinician and sLLM MERs were compared for the quality of the clinical information available using the RI-RADS grading scale by 2 experienced radiologists (Rads 1 and 2). The MRI protocol accuracy for the sLLM and 2 board-certified radiologists (Rads 3 and 4, both with 2 years of experience) was determined by comparison against a reference standard provided by Rads 1 and 2. Claude version 3.5 (Anthropic) was used. *Clinical data included the last clinical entry and imaging reports (eg, computed tomography (CT) of the neck for an MRI nasopharynx). The sLLM and all board-certified radiologists had access to the MRI protocol guidance provided on the hospital intranet. MERs: MRI examination requests; MRI: magnetic resonance imaging; RI-RADS: Reason-for-Exam Imaging Reporting and Data System; sLLM: secure large language model.

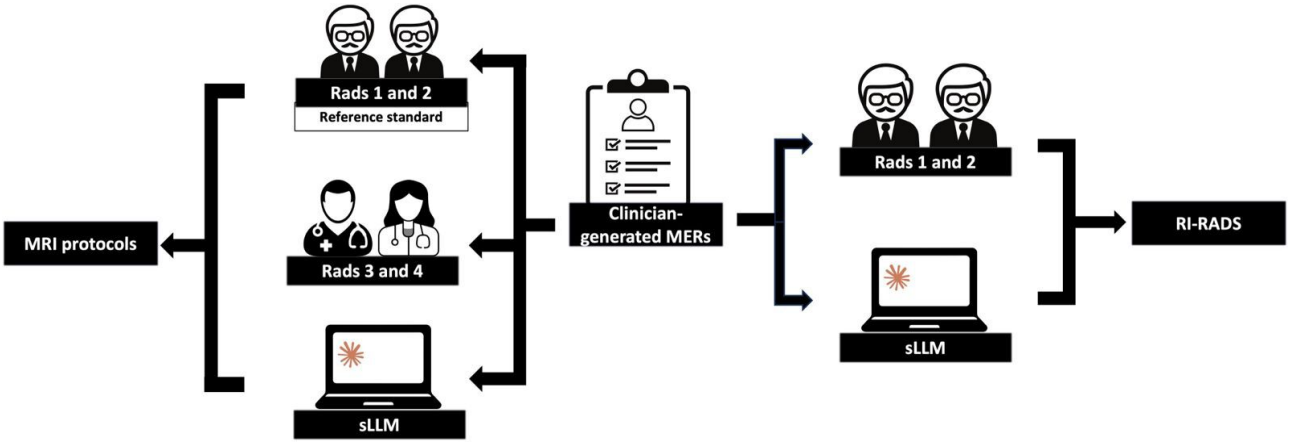

Out of the 608 MRI examinations analyzed, there was the following subspecialty breakdown: 206 (34%) body, 203 (33%) musculoskeletal, and 199 (33%) neuroradiology. Contrast was administered in 85% (175/206) of body scans, 27% (55/203) of musculoskeletal scans, and 70% (139/199) of neuroradiology scans, yielding 60.7% (369/608) of contrast studies overall. Within each specialty, the most common protocols were rectum or perineal fistula, uterus or cervix cancer, liver, and prostate studies for body MRI; lumbar and cervical spine along with knee and shoulder examinations for musculoskeletal MRI; and postcontrast brain, orbital, and skull-base studies for neuroradiology MRI (Table 1).

Adequacy of the Radiology Request Forms

RI-RADS gradings for the MERs were performed independently by 2 experienced radiologists (Rad 1 and Rad 2) (Table 2). Interobserver agreement (AC1) was almost perfect for the sLLM-augmented MERs (AC1 0.97, 95% CI 0.94 - 0.99) and moderate for the clinician MERs (AC1 0.43, 95% CI 0.34 - 0.52).

Table . Reason for exam imaging reporting and data system grades for the clinician and secure large language model–augmented magnetic resonance imaging examination requests[a].

RI-RADS[b] grade	Radiologist 1		Radiologist 2	
	Clinician MERs[c] (n=608), n (%, 95% CI)	sLLM[d] MERs (n=608), n (%, 95% CI)	Clinician MERs (n=608), n (%, 95% CI)	sLLM MERs (n=608), n (%, 95% CI)
A/B	484 (79.6, 95% CI 76.2 - 82.8)	604 (99.3, 95% CI 98.3 - 99.8)	583 (95.9, 95% CI 93.9 - 97.3)	608 (100, 95% CI 99.51 - 100)
C/D	124 (20.4, 95% CI 17.2 - 23.6)	4 (0.7, exact 95% CI, Clopper-Pearson 0.2 - 1.7)	25 (4.1, 95% CI 2.7 - 6)	0 (0, exact 95% CI, Clopper-Pearson 0 - 0.5)

[a]Values are the number of studies, with 95% CIs in brackets.

[b]RI-RADS: Reason-for-Exam Imaging Reporting and Data System.

[c]MERs: magnetic resonance imaging examination requests.

[d]sLLM: secure large language model.

Clinical information on the sLLM-augmented MERs was rated significantly higher than the original requests by both radiologists. For Rad 1, clinician MERs had 484/608 rated A/B (79.6%; 95% CI 76.2%-82.8%) and 124/608 rated C/D (20.4%; 95% CI 17.2%-23.6%), whereas sLLM MERs had 604/608 rated A/B (99.3%; 95% CI 98.3%-99.8%) and 4/608 rated C/D (0.7%; exact 95% CI, Clopper-Pearson, 0.2%-1.7%). For Rad 2, clinician MERs had 583/608 rated A/B (95.9%; 95% CI 93.9%-97.3%) and 25/608 rated C/D (4.1%; 95% CI 2.7%-6.0%), while sLLM MERs had all 608/608 rated A/B (100%; 95% CI 99.51%-100%) and 0/608 rated C/D (0%; exact 95% CI, Clopper-Pearson, 0%-0.5%).

No clinically significant hallucinations were identified in the sLLM outputs. Notably, the sLLM-generated MERs converted 28 spine MRI referrals (28/608, 4.6%) that were rated limited or deficient (7 of which omitted prior spinal surgery) into requests containing adequate clinical details.

MRI Protocoling Accuracy Comparison

For protocoling, experienced radiologists 1 and 2 showed almost perfect agreement on the region (AC1 0.97, 95% CI 0.96 - 0.99) and need for contrast (AC1 0.99, 95% CI 0.99 - 1). A consensus protocol was determined for all 608 cases.

XSL•FO
RenderX

For all 608 MRI examinations, the sLLM matched the subspecialty reference standard in 566 (93.1%, 95% CI 89.6 - 96.6) cases overall, compared with 556 (91.4%, 95% CI 87.6 - 95.3) for Rad 3 and 560 (92.1%, 95% CI 88.4 - 95.8) for Rad 4 (Table 3). This difference was not significant for the sLLM vs Rad 3 (*P*=.23) and Rad 4 (*P*=.40). Accuracy for region or coverage selection was similar across readers (sLLM: 579/608, 95.2%; Rad 3: 585/608, 96.2%; Rad 4: 573/608, 94.2%) with no significant differences between the sLLM vs Rad 3 (*P*=.46) and Rad 4 (*P*=.36). The sLLM, coupled with the use of the deterministic rules-based parsing script, demonstrated a slight advantage for contrast determination, being correct in 94.4% (574/608; 95% CI 91.3 - 97.5) studies vs 92.1% (560/608; 95% CI 88.4 - 95.8) for Rad 3 (*P*=.027) and 92.9% (565/608; 95% CI 89.4 - 96.4) for Rad 4 (*P*=.16). Manual review of disagreements indicated that the sLLM correctly detected infection or inflammation in 2 cases (eg, suspected bursitis around the hip and shoulder), possible nerve lesions in 2 cases, and prior spine surgical details in 1 case, which were documented in the electronic record but not stated on the MER. These details triggered appropriate contrast recommendations that were omitted by at least 1 radiologist.

Table . Magnetic resonance imaging protocoling accuracy for the secure large language model and board-certified radiologists vs the reference standard.

MRI[a] protocol vs reference standard	Radiologist 3	Radiologist 4	sLLM[b]
All MRI studies (N=608), n (%, 95% CI)			
Overall protocol correct	556 (91.4, 95% CI 87.6 - 95.3)	560 (92.1, 95% CI 88.4 - 95.8)	566 (93.1, 95% CI 89.6 - 96.6)
Region correct	585 (96.2, 95% CI 93.6 - 98.8)	573 (94.2, 95% CI 91.1 - 97.4)	579 (95.2, 95% CI 92.3 - 98.1)
Contrast decision correct	560 (92.1, 95% CI 88.4 - 95.8)	565 (92.9, 95% CI 89.4 - 96.4)	574 (94.4, 95% CI 91.3 - 97.5)
Subspecialty subsets			
Musculoskeletal (n=203), n (%, 95% CI)			
Overall protocol correct	186 (91.6, 95% CI 87.8 - 95.4)	180 (88.7, 95% CI 84.3 - 93)	187 (92.1, 95% CI 88.4 - 95.8)
Region correct	199 (98, 95% CI 96.1 - 99.9)	182 (89.7, 95% CI 85.5 - 93.8)	194 (95.6, 95% CI 92.7 - 98.4)
Contrast decision correct	190 (93.6, 95% CI 90.2 - 97)	185 (91.1, 95% CI 87.2 - 95)	196 (96.6, 95% CI 94 - 99.1)
Neuroradiology (n=199), n (%, 95% CI)			
Overall protocol correct	184 (92.5, 95% CI 88.8 - 96.1)	189 (95, 95% CI 91.9 - 98)	185 (93, 95% CI 89.4 - 96.5)
Region correct	189 (95, 95% CI 91.9 - 98)	194 (97.5, 95% CI 95.3 - 99.7)	190 (95.5, 95% CI 92.6 - 98.4)
Contrast decision correct	184 (92.5, 95% CI 88.8 - 96.1)	189 (95, 95% CI 91.9 - 98)	184 (92.5, 95% CI 88.8 - 96.1)
Body (n=206), n (%; 95% CI)			
Overall protocol correct	186 (90.3, 95% CI 86.2 - 94.3)	191 (92.7, 95% CI 89.2 - 96.3)	194 (94.2, 95% CI 91 - 97.4)
Region correct	197 (95.6, 95% CI 92.8 - 98.4)	197 (95.6, 95% CI 92.8 - 98.4)	195 (94.7, 95% CI 91.6 - 97.7)
Contrast decision correct	186 (90.3, 95% CI 86.2 - 94.3)	191 (92.7, 95% CI 89.2 - 96.3)	194 (94.2, 95% CI 91 - 97.4)

[a]MRI: magnetic resonance imaging.

[b]sLLM: secure large language model.

Within the musculoskeletal subset (N=203 examinations), overall agreement with the reference standard was 187 (92.1%, 95% CI 88.4 - 95.8) for the sLLM, 186 (91.6%, 95% CI 87.8 - 95.4) for Rad 3, and 180 (88.7%, 95% CI 84.3 - 93) for Rad 4. These differences were not significant (*P*>.99 and *P*=.15 for sLLM vs Rad 3 and Rad 4, respectively). The sLLM matched Rad 3 for region accuracy (n=194, 95.6% vs 199, 98%; *P*=.23) and exceeded Rad 4 (n=182, 89.7%; *P*=.018) (Table 3). For contrast, coupled with the use of the deterministic rules-based parsing script, the sLLM was correct in 196 (96.6%, 95% CI 94 - 99.1) studies, outperforming Rad 4 alone (n=185, 91.1%; *P*=.006) with no significant difference compared to Rad 3 (n=190, 93.6%; *P*=.099).

For body MRI (N=206 examinations), the sLLM had the highest overall accuracy (n=194, 94.2%, 95% CI 91 - 97.4), although this was not significantly different vs Rad 3 (n=186, 90.3%; *P*=.089) and Rad 4 (n=191, 92.7%; *P*=.33). Review of body MRI discrepancies indicated that the sLLM recommended more focused uterus or cervix and MRI rectum and perineum protocols in 2 cases compared to a more general pelvis by the 2 radiologists. For contrast, the sLLM had the highest accuracy (n=194, 94.2%, 95% CI 91 - 97.4), although this was not significantly different vs Rad 3 (n=186, 90.3%; *P*=.09) and Rad 4 (n=191, 92.7%; *P*=.33).

In neuroradiology MRI (N=199 examinations), Rad 4 achieved the highest overall accuracy (n=189, 95%, 95% CI 91.9 - 98), although this was not significant vs the sLLM (n=185, 93%; *P*=.50) and Rad 3 (n=184, 92.5%; *P*=.71). Review of neuroradiology discrepancies indicated that the sLLM occasionally recommended skull-base protocols in 5 cases with suspected cranial nerve pathologies based on the clinical notes, whereas radiologists 3 and 4 accepted the original brain request when they deemed that coverage was adequate. For contrast, Rad 4 achieved the highest accuracy (n=189, 95%, 95% CI 91.9 - 98), although this was not significant vs the sLLM

XSL•FO
RenderX

(n=184, 92.5%; *P*=.30) and Rad 3 (n=184, 92.5%; *P*=.32) which were closely aligned.

Discussion

Principal Findings

In this study, we compared MERs provided by the referring clinician with those augmented by our institutional sLLM. Our results show that an sLLM can upgrade the clinical usefulness of MERs and aid protocol selection across body, neuroradiology, and musculoskeletal practice. After augmentation, fewer than 1% of requests (0% - 0.7%) were graded deficient or limited (RI-RADS C or D), compared with up to 1 in 5 (4.1% - 20.4%) of the clinician originals, and interreader agreement rose from moderate to almost perfect.

The sLLM protocoling accuracy was compared with that of 2 board-certified junior general radiologists. Overall protocol accuracy of the sLLM (566/608, 93.1%) was close to that of the 2 board-certified radiologists (Rad 3: 556/608, 91.4%; *P*=.23 and Rad 4: 560/608, 92.1%; *P*=.40), although the study was not powered to test statistical noninferiority or equivalence. Observed accuracy differences should not be interpreted as evidence of statistical similarity but instead as descriptive comparisons indicating that the sLLM operated within the performance range of general radiologists on this dataset.

For contrast decisions, the sLLM demonstrated accuracy of 94.4% (574/608), which was superior to Rad 3 (560/608, 92.1%; *P*=.027) and was not significantly different from Rad 4 (565/608, 92.9%; *P*=.16). Manual review showed that slightly increased accuracy for the sLLM was driven by EMR information (correct detection of prior surgery, infection, or tumor history) that did not appear on the MER yet was important for protocol selection. In our pipeline, the sLLM's role was to extract and summarize these clinically relevant details, while the final region and contrast decisions were made by a deterministic rule-based script using predefined institutional criteria. Importantly, protocol selection and contrast determination are treated as separate decision steps. Protocol names reflect the anatomical region and clinical indication, while contrast administration may be modified when an explicit clinician request for noncontrast imaging or a documented contraindication is detected. In such cases, the default contrast setting associated with a protocol may be overridden to preserve clinician intent and patient safety. As such, the observed accuracy reflects improved information extraction feeding into consistent rules, rather than autonomous clinical reasoning by the sLLM.

In one of this study's cases, the clinician requested noncontrast MRI for cervical cancer recurrence due to renal impairment of estimated glomerular filtration rate 31 - 48, although modern guidelines may permit the use of group II gadolinium agents at this level. We acknowledge that prioritizing clinician intent in such cases may risk suboptimal diagnostic utility and potential recall.

Most mismatches against the reference standard occurred in borderline cases where more than one subspecialty-specific protocol was reasonable. In body MRI, this typically involved the sLLM proposing a general pelvis study when radiologists selected a focused protocol such as prostate, uterus or cervix, or rectum. In spine MRI, a small number of discrepancies arose from consolidating concurrent cervical and lumbar requests into a single-region study. In neuroradiology, differences usually reflected choosing between brain vs skull base or temporal bone coverage in suspected cranial nerve pathology. These patterns suggest that minor refinements to protocol-selection rules could help reduce discrepancies.

Comparison to Prior Work

These findings are consistent with recent evaluations of LLMs for noninterpretive tasks in radiology and add multisubspecialty evidence. Prior institutional studies using an sLLM for spine and musculoskeletal MRI reported similar gains, with musculoskeletal protocoling accuracy reaching 96% and exceeding 2 general readers at 88% and 89% [17,18]. In another recent study, Çamur et al [25] showed that 4 LLMs have strong potential for selecting appropriate imaging modalities. The 4 LLMs were tested on 240 clinical cases (120 ACR Appropriateness Criteria and 120 realistic scenarios) and compared their choices with 4 clinicians and 4 radiologists. The best model picked the correct imaging test in 98.3% (236/240) of ACR cases, matched a junior radiologist on realistic cases, gave identical answers across prompts (κ=1), and showed moderate to good reproducibility over time (κ~0.77 - 0.89 short term; 0.51 - 0.79 long term) [25]. Similarly, for CT protocol assignment, a fine-tuned support tool picked the right protocol on the first choice 92.3% of the time and within its top two 96.3% of the time, with an average processing time of less than 1 second per case. When clinicians used it, resident accuracy improved from 0.913 to 0.936 with a 14% reduction in reading time, and attending accuracy increased from 0.920 to 0.926 with a 12% time saving [26]. In musculoskeletal MRI, a GPT-4 system linked to a small knowledge base reached 92.86% accuracy on ACR-based cases, outperformed a baseline model and standard GPT-4, matched most subspecialists, and was better at flagging when the clinical information was insufficient, which mirrors our MER enrichment step [27]. Beyond protocol choice, a radiology operations study showed GPT-4 routed 96% of in-scope procedure requests and 76% of out-of-scope requests correctly at a cost of approximately US $0.03 per request, indicating potential for increased efficiency and cost benefits for routing workflows [28]. Together with our results, these data suggest that institution-hosted sLLMs could aid improvements in clinical context, standardize protocol decisions, reduce unnecessary contrast and radiation, and potentially save time and costs by reducing protocoling time and reschedules.

Limitations and Future Directions

Our study has several limitations. First, this was a single-center, retrospective evaluation involving only outpatient MERs. The sLLM pipeline relied on the specific institutional MRI protocol repository and access to the EMR structure, and the realized gains may differ in institutions using different order-entry systems or protocol libraries. Future work should include prospective, multicenter evaluations that measure operational outcomes (time to scan, rescheduling, repeat imaging), cost-effectiveness analyses that incorporate staff time and

XSL•FO
RenderX

scanner utilization, and extension of the sLLM pipeline to CT and ultrasound.

Second, some patients contributed multiple examinations, and the dataset lacked patient-linked identifiers; therefore, cluster-robust or mixed-effects analyses were not feasible. Results should be interpreted at the examination level, with future work planned to incorporate patient-level clustering.

Third, formatting differences between the clinician and augmented MRI examination requests could have partially unblinded the graders, with the potential to artificially inflate the sLLM's RI-RADS scores, although the direction of any resulting bias is uncertain. In our study, the output format (subheadings; eg, clinical history, reason for exam) was identical for the clinician and sLLM forms, although differences in language and style may have unblinded the reviewers. This is difficult to rectify but could be addressed in future studies using standardized outputs.

Fourth, the exclusion of cases with incomplete electronic records introduces a potential selection bias, as these omitted cases may represent some of the more challenging instances for protocoling. In addition, the manual extraction of the latest relevant clinical entry and pertinent prior imaging reports without an automated failsafe mechanism provided a methodological simplification, which is different from real-life human EMR review, which involves searching through all relevant notes and prior imaging. Together, these could potentially inflate the sLLM's protocoling accuracy and are an important area for future improvement. Future scaling may also benefit from approaches such as hierarchical summarization or RAG-style chunking for very large EMR entries.

Fifth, we assessed board-certified radiologists who demonstrated high accuracies for protocoling (>91.4%), leaving little room for the sLLM to improve on this. LLM assistance for less experienced readers, including technologists, could provide more value and will need to be assessed. Nonetheless, even if accuracy gains over experts are marginal, the standardization benefit and fatigue reduction offered by the sLLM support its deployment. A further improvement may include an "assisted radiologist with sLLM" arm to separate automation from augmentation.

Sixth, although the study included more than 600 MRI examinations, the sample size was powered for detecting a 15% difference in clinical-information adequacy, not for establishing noninferiority or equivalence in protocol accuracy. The small absolute differences observed between the sLLM and board-certified radiologists (eg, 93.1% vs 91.4%; P=.23) therefore cannot be interpreted as statistical equivalence.

Seventh, we deliberately confined the evaluation to MRI, which is our most protocol-intensive modality and the largest component of departmental workload, but acknowledge that CT and ultrasound will require separate validation prior to broader application.

Eighth, the review for clinically significant hallucinations was limited to discordant cases, and there is a theoretical risk that the sLLM could generate benign or "silent" hallucinations in the clinical summary that do not alter the protocol decision but could still degrade the medical record.

Finally, although no clinically significant hallucinations were observed, it is well known that LLMs are prone to producing factual errors that could have major clinical implications [29-31]. In addition, recent work shows that LLMs can be vulnerable to adversarial hallucination in clinical decision-support settings, with mitigation prompts only partially reducing errors and lower temperature settings providing no meaningful benefit [31]. Continuous monitoring and postdeployment guardrails therefore remain essential, but were beyond the scope of this study.

Conclusion

In this multisubspecialty MRI cohort, an sLLM improved MRI examination request completeness and demonstrated protocol accuracy comparable to experienced radiologists; however, the study was not powered to establish noninferiority or equivalence for protocol accuracy.

The end-to-end sLLM pipeline, consisting of request form enrichment with EMR data and a rule-based protocoling technique, offers a practical pathway for more efficient, standardized protocoling while reducing the administrative burden on clinicians.

Acknowledgments

The author(s) attest that there was no use of generative artificial intelligence technologies in the generation of text, figures, or other informational content of this manuscript.

Funding

This study received support from the Singapore Ministry of Health National Medical Research Council under the NMRC Clinician Innovator Award (CIA). The grant was awarded for the project titled "From Prototype to Full Deployment: A Comprehensive Deep Learning Pipeline for Whole-Spine MRI" (Grant ID: CIAINV25jan-0005 J.T.P.D.H).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: JTPDH, NWL, YXL
Data curation: JTPDH, NWL
Formal analysis: JTPDH, NWL
Funding acquisition: JTPDH
Investigation: JTPDH, NWL, YXL
Methodology: JTPDH, NWL, YXL
Project administration: JTPDH, NWL, YXL
Resources: JTPDH, NWL, YXL, AL, WO, MDZC, GKD, XZL, MCL, CY, WJS, ECT, JHT, NK, AM, YT
Software: JTPDH, NWL, YXL, AL, WO, MDZC, GKD, XZL, MCL, CY, WJS, ECT, JHT, NK, AM, YT
Supervision: JTPDH
Validation: JTPDH, NWL, YXL, AL, WO, MDZC, GKD, XZL, MCL, CY, WJS, ECT, JHT, NK, AM, YT
Visualization: JTPDH, NWL, YXL
Writing – original draft: JTPDH, NWL, YXL
Writing – reviewing & editing: JTPDH, NWL, YXL, AL, WO, MDZC, GKD, XZL, MCL, CY, WJS, ECT, JHT, NK, AM, YT

Conflicts of Interest

None declared.

Multimedia Appendix 1
Large language model prompts.
[DOCX File, 42 KB - jmir_v28i1e82579_app1.docx]

Multimedia Appendix 2
Reason for Exam Imaging Reporting and Data System (RI-RADS).
[DOCX File, 22 KB - jmir_v28i1e82579_app2.docx]

References

1. Lehnert BE, Bree RL. Analysis of appropriateness of outpatient CT and MRI referred from primary care clinics at an academic medical center: how critical is the need for improved decision support? J Am Coll Radiol 2010 Mar;7(3):192-197. [doi: 10.1016/j.jacr.2009.11.010] [Medline: 20193924]
2. Bernardy M, Ullrich CG, Rawson JV, et al. Strategies for managing imaging utilization. J Am Coll Radiol 2009 Dec;6(12):844-850. [doi: 10.1016/j.jacr.2009.08.003] [Medline: 19945039]
3. G Pitman A. Quality of referral: what information should be included in a request for diagnostic imaging when a patient is referred to a clinical radiologist? J Med Imaging Radiat Oncol 2017 Jun;61(3):299-303. [doi: 10.1111/1754-9485.12577] [Medline: 28139044]
4. Barakzai MD, Sheer ZZ, Muhammad A, et al. Evaluation of radiology request forms in a tertiary care hospital: an audit with a focus on the impact of technological intervention. Cureus 2021 Feb 14;13(2):e13335. [doi: 10.7759/cureus.13335] [Medline: 33747644]
5. Castillo C, Steffens T, Sim L, Caffery L. The effect of clinical information on radiology reporting: a systematic review. J Med Radiat Sci 2021 Mar;68(1):60-74. [doi: 10.1002/jmrs.424] [Medline: 32870580]
6. Ip IK, Schneider L, Seltzer S, et al. Impact of provider-led, technology-enabled radiology management program on imaging. Am J Med 2013 Aug;126(8):687-692. [doi: 10.1016/j.amjmed.2012.11.034] [Medline: 23786668]
7. Trivedi H, Mesterhazy J, Laguna B, Vu T, Sohn JH. Automatic determination of the need for intravenous contrast in musculoskeletal MRI examinations using IBM Watson's natural language processing algorithm. J Digit Imaging 2018 Apr;31(2):245-251. [doi: 10.1007/s10278-017-0021-3] [Medline: 28924815]
8. Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. Radiology 2024 Jan;310(1):e232756. [doi: 10.1148/radiol.232756] [Medline: 38226883]
9. Gertz RJ, Bunck AC, Lennartz S, et al. GPT-4 for automated determination of radiological study and protocol based on radiology request forms: a feasibility study. Radiology 2023 Jun;307(5):e230877. [doi: 10.1148/radiol.230877] [Medline: 37310247]
10. Barash Y, Klang E, Konen E, Sorin V. ChatGPT-4 assistance in optimizing emergency department radiology referrals and imaging selection. J Am Coll Radiol 2023 Oct;20(10):998-1003. [doi: 10.1016/j.jacr.2023.06.009] [Medline: 37423350]
11. Kim S, Lee CK, Kim SS. Large language models: a guide for radiologists. Korean J Radiol 2024 Feb;25(2):126-133. [doi: 10.3348/kjr.2023.0997] [Medline: 38288895]
12. Sun Z, Ong H, Kennedy P, et al. Evaluating GPT4 on impressions generation in radiology reports. Radiology 2023 Jun;307(5):e231259. [doi: 10.1148/radiol.231259] [Medline: 37367439]

XSL•FO
RenderX

13. Mese I, Taslicay CA, Sivrioglu AK. Improving radiology workflow using ChatGPT and artificial intelligence. Clin Imaging 2023 Nov;103:109993. [doi: 10.1016/j.clinimag.2023.109993] [Medline: 37812965]
14. Mukherjee P, Hou B, Lanfredi RB, Summers RM. Feasibility of using the privacy-preserving large language model vicuna for labeling radiology reports. Radiology 2023 Oct;309(1):e231147. [doi: 10.1148/radiol.231147] [Medline: 37815442]
15. Rau A, Rau S, Zoeller D, et al. A context-based chatbot surpasses trained radiologists and generic ChatGPT in following the ACR appropriateness guidelines. Radiology 2023 Jul;308(1):e230970. [doi: 10.1148/radiol.230970] [Medline: 37489981]
16. Cai W. Feasibility and prospect of privacy-preserving large language models in radiology. Radiology 2023 Oct;309(1):e232335. [doi: 10.1148/radiol.232335] [Medline: 37815443]
17. Hallinan J, Leow NW, Ong W, et al. MRI spine request form enhancement and auto protocoling using a secure institutional large language model. Spine J 2025 Mar;25(3):505-514. [doi: 10.1016/j.spinee.2024.10.021] [Medline: 39536908]
18. Hallinan J, Leow NW, Low YX, et al. An institutional large language model for musculoskeletal MRI improves protocol adherence and accuracy. J Bone Joint Surg Am 2025 Jul 8;107(16):1833-1840. [doi: 10.2106/JBJS.24.01429] [Medline: 40627696]
19. Bhayana R, Alwahbi O, Ladak AM, et al. Leveraging large language models to generate clinical histories for oncologic imaging requisitions. Radiology 2025 Feb;314(2):e242134. [doi: 10.1148/radiol.242134] [Medline: 39903072]
20. Terzis R, Kaya K, Schömig T, et al. GPT-4 for automated sequence-level determination of MRI protocols based on radiology request forms from clinical routine. Eur Radiol 2026 Feb;36(2):1541-1552. [doi: 10.1007/s00330-025-11888-4] [Medline: 40779162]
21. Tavakoli N, Kim D. AI-generated clinical histories for radiology reports: closing the information gap. Radiology 2025 Feb;314(2):e243910. [doi: 10.1148/radiol.243910] [Medline: 39903081]
22. Abedi A, Tofighi S, Salehi S, Latterman PT, Basques KD, Gholamrezanezhad A. Reason for Exam Imaging Reporting and Data System (RI-RADS): a grading system to standardize radiology requisitions. Eur J Radiol 2019 Nov;120:108661. [doi: 10.1016/j.ejrad.2019.108661] [Medline: 31610322]
23. Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med 2025 May 13;8(1):274. [doi: 10.1038/s41746-025-01670-7] [Medline: 40360677]
24. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008 May;61(Pt 1):29-48. [doi: 10.1348/000711006X126600] [Medline: 18482474]
25. Çamur E, Cesur T, Güneş YC, et al. Evaluating large language models for imaging modality selection: potential to reduce unnecessary contrast agent use and radiation exposure. Clin Imaging 2025 Sep;125:110573. [doi: 10.1016/j.clinimag.2025.110573] [Medline: 40752449]
26. Kanemaru N, Yasaka K, Okimoto N, et al. Efficacy of fine-tuned large language model in CT protocol assignment as clinical decision-supporting system. J Imaging Inform Med 2025 Dec;38(6):4336-4348. [doi: 10.1007/s10278-025-01433-6] [Medline: 39909993]
27. Tan JR, Lim DYZ, Le Q, et al. ChatGPT performance in assessing musculoskeletal MRI scan appropriateness based on ACR appropriateness criteria. Sci Rep 2025 Feb 28;15(1):7140. [doi: 10.1038/s41598-025-88925-1] [Medline: 40021679]
28. Triana BP, Wiggins WF, Befera N, Roth C, Cline B. Proof-of-concept prompted large language model for radiology procedure request routing. J Vasc Interv Radiol 2025 Jul;36(7):1201-1207. [doi: 10.1016/j.jvir.2025.03.012] [Medline: 40139433]
29. Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Diagn Interv Radiol 2024 Mar 6;30(2):80-90. [doi: 10.4274/dir.2023.232417] [Medline: 37789676]
30. Kao JP, Kao HT. Large language models in radiology: a technical and clinical perspective. Eur J Radiol Artif Intell 2025 Jun;2:100021. [doi: 10.1016/j.ejrai.2025.100021]
31. Omar M, Sorin V, Collins JD, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Commun Med (Lond) 2025 Aug 2;5(1):330. [doi: 10.1038/s43856-025-01021-3] [Medline: 40753316]

Abbreviations

EMR: electronic medical record
LLM: large language model
MERs: MRI examination requests
MRI: magnetic resonance imaging
RAG: retrieval-augmented generation
RI-RADS: Reason-for-Exam Imaging Reporting and Data System
sLLM: secure large language model

XSL•FO
RenderX

Edited by A Coristine; submitted 25.Aug.2025; peer-reviewed by A Shivanna, BL Guellec, I Odezuligbo; revised version received 20.Feb.2026; accepted 24.Feb.2026; published 07.Apr.2026.

Please cite as:
Hallinan JTPD, Leow NW, Low YX, Lee A, Ong W, Chan MDZ, Devi GK, He SS, Loh DDL, Lim DSW, Low XZ, Lim MC, Yong C, Sng WJ, Teo EC, Tan JH, Kumar N, Makmur A, Ting Y
Initial Insights Into an Institutional Secure Large Language Model for Magnetic Resonance Imaging Examination Requests: Retrospective Study
J Med Internet Res 2026;28:e82579
URL: https://www.jmir.org/2026/1/e82579
doi:10.2196/82579

© James Thomas Patrick Decourcy Hallinan, Naomi Wenxin Leow, Yi Xian Low, Aric Lee, Wilson Ong, Matthew Ding Zhou Chan, Ganakirthana Kalpenya Devi, Stephanie Shengjie He, Daniel De-Liang Loh, Desmond Shi Wei Lim, Xi Zhen Low, Mei Chin Lim, Clement Yong, Weizhong Jonathan Sng, Ee Chin Teo, Jiong Hao Tan, Naresh Kumar, Andrew Makmur, Yonghan Ting. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Artificial Intelligence in Health Professions Education: Qualitative Study of Student Experiences

Marina Guirguis[1*], BA, BSc; Salomon Fotsing[1,2,3,4*], MD, MAEd; Joanne Fevry[1,2*], MD, MScA; Christine Landry[5*], BPharm, ACPR, MSc, BCPS, PharmD; Diane Bouchard-Lamothe[2*], MSc; Jennifer Lacroix[2*], MScHQ; Alireza Jalali[1,2,6*], MD

[1]Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

[2]Francophone Affairs, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

[3]Department of Family Medicine, University of Ottawa, Ottawa, ON, Canada

[4]Institut du Savoir Montfort, Ottawa, ON, Canada

[5]School of Pharmaceutical Sciences, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

[6]Department of Innovation in Medical Education, Faculty of Medicine, University of Ottawa, 451 Smyth (3247BD), Ottawa, ON, Canada

[*]all authors contributed equally

Corresponding Author:
Alireza Jalali, MD
Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

Abstract

Background: Artificial intelligence (AI) is increasingly integrated into education and health care, raising questions about how students use these technologies and how AI influences their learning. In health education, understanding these trends is particularly important because student learning directly impacts future clinical skills.

Objective: This study aimed to explore the use of AI tools by health sciences students at the University of Ottawa. More specifically, it sought to identify the most frequently used AI tools, describe students' usage habits, determine which tools support knowledge acquisition and skill development, and gather students' recommendations for effective strategies to raise awareness and train their peers on the responsible use of AI.

Methods: A qualitative approach was used with students from 10 health professions who reported using AI in their studies. Data were collected through semistructured interviews and an open-ended qualitative online survey. Inductive thematic analysis within an interpretive paradigm was applied to capture patterns, perceptions, and emergent themes.

Results: A total of 51 health professions students participated in the study. Most were women between the ages of 20 and 29 years. ChatGPT (OpenAI) emerged as the most frequently used AI tool. Students perceived AI as a complementary tool that facilitated knowledge acquisition, skill development, writing, and problem-solving. AI adoption was driven by curiosity, peer influence, and the desire to improve work efficiency. Students critically evaluated AI results, integrated the tools into their learning processes, and emphasized the importance of technical skills, critical thinking, and digital literacy. Peer learning, hands-on demonstrations, and access to online resources were recommended for effective AI training.

Conclusions: This research demonstrates that health professions students actively use AI tools, particularly ChatGPT, to support learning, skill development, and academic tasks. Although AI is valuable as an educational aid and its use varies by student and context, this highlights the need for structured guidance, critical evaluation skills, and peer-supported training. These findings highlight the importance of thoughtfully integrating AI into educational programs to enhance learning outcomes, foster skill acquisition, and ensure responsible and effective adoption.

(J Med Internet Res 2026;28:e82432) doi:10.2196/82432

KEYWORDS

artificial intelligence; AI; qualitative study; health professions education; learning technologies; ChatGPT; learning experience; digital literacy

Introduction

Presently, the use of artificial intelligence (AI) has become commonplace in health care, health professions training, and the pharmaceutical industry. AI refers to the capacity of machines to simulate human functions such as reasoning, learning, and problem-solving through algorithms [1,2]. This technology, which "mimics human cognitive functions," is becoming increasingly effective in the health care sector.

XSL•FO
RenderX

Although it sometimes raises ethical concerns [3], AI now assists clinicians in making better decisions, especially in fields like diagnostic imaging, where it aids in image interpretation [2,4], and dermatology, where it helps in the early detection of skin cancer by using efficient classifiers for skin tumors [5]. In addition, in the pharmaceutical field, AI accelerates drug design and discovery processes [2,6].

The use of AI in health professional education is also growing. With the aid of tools like ChatGPT (OpenAI), it can create simulated clinical environments, evaluate clinician-simulated patient communication, create quizzes, and develop curricula [7]. At the same time, it can respond instantly to specific questions related to students' areas of study through AI chatbots used in medical education [8]. The new generation of health learners is much more inclined to adopt and use this technology. In 2022, a national Canadian survey of nearly 2000 health professions students showed that AI was viewed favorably by students, who also believe that this technology will positively influence their future careers [9]. Similarly, in March 2023, a systematic review reported that health professions students have a favorable attitude toward the use of AI [10].

From an academic standpoint, the use of AI by health professions students increasingly raises the question of the "learning experience." Learning is a dynamic concept defined as a process that allows individuals to perceive objects, interact with them, and integrate them into their social, cognitive, and emotional dimensions to transform, create, or evolve their cognitive structure [11]. The learning experience goes beyond knowledge acquisition. It facilitates the development of skills and cultural values through observation, imitation, trial, repetition, and presentation [12]. Therefore, an important question regarding AI in health professions training programs is how AI contributes to the students' learning experience. Shorey et al [13] scoping review of the learning styles, preferences, and needs of Generation Z health students marks a first step in this direction. Given the increasing adoption of AI, raising awareness and providing training on the proper use of these tools has become essential [9]. Peer education is a viable approach because it allows students to share valuable insights and fosters a collaborative learning environment that encourages peers to use AI tools [14,15].

Access to AI tools is also an important element in the training of health science students, given their increasing use in modern learning and clinical reasoning processes [16]. However, socioeconomic and linguistic factors may influence their use by these students, as some of these tools are mainly programmed in English and sometimes require payment to access more advanced features [16,17]. In a bilingual institution of higher learning such as the University of Ottawa, where courses are taught in French, English, or both, it is essential to ensure equitable access to these resources for all students.

This qualitative study aimed to explore which AI tools health professions students at the University of Ottawa report using, describe their usage habits, identify the tools that help them acquire new knowledge and develop skills, and explore what these students consider to be the best strategies for raising awareness and training their peers on the proper use of AI in learning environments. Specifically, we focused on how students use AI tools to support their learning and on strategies they recommend for peer awareness and training.

Methods

Qualitative Approach and Research Paradigm

An inductive thematic analysis approach, following an interpretive research paradigm, was used to identify the AI platforms most used by health professions students during their learning activities. This approach was chosen for its relevance to explore perceptions and experiences regarding AI tools in medical education [18]. This approach, based on the interpretive paradigm, posits that participants' narratives construct meanings, which promotes a detailed understanding of behaviors and points of view [19]. Therefore, in accordance with the research objectives, the inductive approach facilitated the emergence of themes directly from the data for a more flexible and data-driven interpretation [20]. This paper was written by using the Standards for Reporting Qualitative Research (SRQR) guide (Checklist 1) [21].

Research Characteristics and Reflexivity

The research team was multidisciplinary and followed a collaborative approach, demonstrating reflexivity at each phase of the study. Team members included MG, an undergraduate medical student experienced in qualitative research, who contributed to study design, data collection, thematic analysis, and writing; SF, a physician and clinical research manager, who provided methodological guidance, qualitative analysis, results interpretation, and manuscript revision; JF, an anatomy educator with expertise in qualitative research and scientific writing, supported data coding, thematic analysis, and writing; CL, a clinician-researcher and pharmacy program director, offered insights into results interpretation; DB-L, a medical education advisor, contributed to the interpretation of the results and the revision of the manuscript; JL, a program manager experienced in qualitative analysis, helped organize and coordinate the project; and AJ, a physician and anatomy professor with expertise in technology-assisted learning and educational innovation, contributed to data analysis, results interpretation, and manuscript revision.

Regular meetings encouraged open dialogue, allowing all coauthors to critically reflect on their assumptions, disciplinary lenses, and interpretations throughout data collection, analysis, and manuscript writing.

Context

Our study was conducted among health professions students at the University of Ottawa. All data were taken outside regular class hours, either through online interviews on Zoom or via an asynchronous online questionnaire.

Sampling Strategy

A nonprobabilistic sampling method was used, following the recommendations of Murphy et al [22] and Higginbottom [23]. Only students who used or had previously used AI platforms for their learning activities in the previous academic year were eligible. To identify eligible participants, all students in the

XSL•FO
RenderX

targeted programs received an invitation email including a brief screening question asking whether they had used AI tools (eg, ChatGPT, Perplexity AI, or any other AI-assisted learning platforms) for their academic activities in the last academic year. Only those who answered "yes" were invited to participate in the study, while students who responded "no" were excluded.

The targeted programs included medicine, pharmacy, molecular and translational medicine, nursing, occupational therapy, physiotherapy, audiology, speech-language pathology, health sciences, and psychology. Our target sample size was based on the writings of Guest et al [24], which specified that in a qualitative study, fundamental elements of the themes studied are present in participants' discourse after 6 interviews, and theoretical saturation of information is achieved after the first 12 interviews. Therefore, we planned to recruit a minimum of 6 participants from each training program, with no restrictions on the maximum number of participants. The distribution of students by program was not tracked separately for the interviews and the questionnaire; overall characteristics of participants are presented in Table 1.

Table . Sociodemographic characteristics of participants (N=51).

Category	Values, n
Age (years)	
<20	9
20 - 29	39
30 - 39	2
40 - 49	1
≥50	0
Gender	
Women	41
Men	9
I prefer not to answer	1
Others	0
Health professions	
Health sciences	27
Medicine	8
Psychology	8
Nursing	1
Pharmacy	1
Others	
Biomedical sciences	5
Biology	1

Data Collection

Data collection began after the recruitment phase, between January and March 2024. Invitation emails to participate in the research project were sent to all students in the targeted programs. Those who agreed to participate were met for semistructured online interviews via Zoom, lasting 30 - 45 minutes each. These interviews were conducted according to an interview guide developed in advance by the research team (see Section A of Multimedia Appendix 1). The guide aimed at capturing information on the most used AI platforms by health professions students at the University of Ottawa, their usage habits of these platforms, the platforms that they believed allowed them to acquire new knowledge and develop skills, and the best strategies for raising awareness and training their peers in the appropriate use of AI in learning.

The guide also included short vignettes to help organize questions about learning processes with AI tools. These vignettes, composed of a definition of the learning process [25,26] and Kolb's [27] experiential learning model, are fully incorporated into the guide provided in Section A of Multimedia Appendix 1.

However, after only 6 interviews over a 4-month period, data collection was adapted to include an online questionnaire using the same interview guide. A total of 51 students made up the final sample after 6 students participated in semistructured interviews and 45 students completed the online qualitative questionnaire. The interview guide and the online questionnaire were made available in both French and English to accommodate participants' language preferences and diffused between June and July 2024. The interview guide was reproduced identically on the online questionnaire platform (SurveyMonkey), thus ensuring that the same questions were asked in both data

XSL•FO
RenderX

collection methods while ensuring methodological continuity. New invitation emails were sent to all students in the targeted programs.

Data Processing and Analysis

Interview Process

The semistructured interviews were conducted on Zoom. No questions that could identify participants were asked. The audio of the interviews was recorded and transcribed using the WhisperAI transcription application. The transcriptions were then verified by a member of the research team for accuracy. To protect anonymity, the transcripts were saved in the order in which the interviews were conducted, and no identifiers were used in the file names. Moreover, interviews and transcriptions were conducted by two separate individuals. All files were password-protected.

Survey Process

Following ethics modification, a qualitative online questionnaire designed to elicit narrative and open-ended responses was hosted on the SurveyMonkey platform. Only one optional question allowing for the identification of participants was asked: "participants could provide their email address at the end of the questionnaire if they wished to be included in a draw for one of five gift cards as compensation." To protect anonymity, participants' responses were saved in the order in which they were received and were coded (eg, P1 for participant 1). The responses were stored in a password-protected Microsoft Excel file and kept separate from the email addresses, which were also in a password-protected Microsoft Excel file.

During the analysis, the information collected during the interviews and questionnaires was treated together as a single qualitative corpus. All coding was performed inductively; although Kolb's experiential learning model informed the development of the questions, it did not guide the coding process. Each participant's questionnaire responses were treated as individual transcripts and coded using the same inductive thematic analysis procedure as the interview data. Transcripts were read and revised several times to identify patterns and themes emerging from the data. To reinforce the methodological rigor of our qualitative study, a double thematic analysis was carried out. First, a researcher identified and categorized emerging themes from the corpus of data. Subsequently, a second researcher went back over all the material independently to reproduce the thematic analysis. This approach is part of a logic of analytical triangulation, aimed at limiting interpretive bias and reinforcing the credibility of the results as recommended by Paillé and Mucchielli [28]. The confrontation of analyses enabled us to discuss divergences, refine thematic categories, and arrive at a consensual interpretation of the material.

Techniques to Enhance Trustworthiness

To ensure the trustworthiness of the analysis, the data were analyzed systematically. An experienced qualitative researcher manually coded 100% of the data, and a second member of the research team independently coded 25% of the dataset. This second coder was trained by the first to master the principles of qualitative analysis and the steps involved in coding interviews and to ensure accurate coding. The second coder coded two randomly selected complete transcripts and half of a third. The two coders then reviewed and compared their coding, and clarifications were made to better represent the participants' experience until consensus was reached. No formal Cohen κ or percent agreement was calculated; reliability was ensured through this consensus-based process.

Next, a tree diagram was created to organize the coded data into potential themes based on observed similarities. Latent thematic analysis, using an interpretive approach, was used to identify subthemes. Finally, the two coders discussed the themes and subthemes related to the selected quotes. The coding framework was developed using NVivo software (version 12; QSR International) [29].

Ethical Considerations

This study was reviewed and approved in accordance with the standards of the Research Ethics Board (REB) of the University of Ottawa (#H-11-23-9455). All participants provided informed consent prior to data collection, which took place outside regular class hours. Due to a low recruitment rate, a modification of the ethics protocol was requested and approved by the REB to allow participants to complete an online questionnaire instead of taking part in interviews. This modification was implemented during the data collection phase. Compensation was offered in the form of an optional draw for one of five CAD $25 (US $18.08) gift cards.

Participant privacy and confidentiality were strictly maintained throughout the study. All data were anonymized using participant codes, and no identifying information was included in the transcript or analysis. All data were securely stored and accessible only to the research team.

Results

Participant Characteristics and Qualitative Findings

In total, 51 students from various health-related fields who had used AI tools during their training activities participated in this study. Most participants were between 20 and 29 years old (n=39). In terms of gender identity, the majority identified as women (n=41), followed by men (n=9), and 1 participant preferred not to disclose. No participants identified as nonbinary or transgender persons.

Health sciences was the most reported field of study (n=27), followed by medicine and psychology, each with 8 participants (see Table 1).

Participants reported using a wide variety of AI tools. ChatGPT was by far the most frequently used, with 49 respondents indicating its use. Other tools mentioned included Microsoft Copilot (n=2), Perplexity AI (n=2), Anthropic Claude (n=2), and several specialized AI tools, such as Google Bard AI, Summarizer AI, Grammarly AI, Midjourney, and Scite, each cited by one respondent (see Section B of Multimedia Appendix 1 for further details).

Themes were identified and developed from the content of the semistructured interviews and the responses to the qualitative

XSL•FO
RenderX

survey, then organized into 4 main thematic areas (see Figure 1 for a summary).

Figure 1. Concept map illustrating the emerging themes and subthemes of the qualitative data analysis. AI: artificial intelligence.

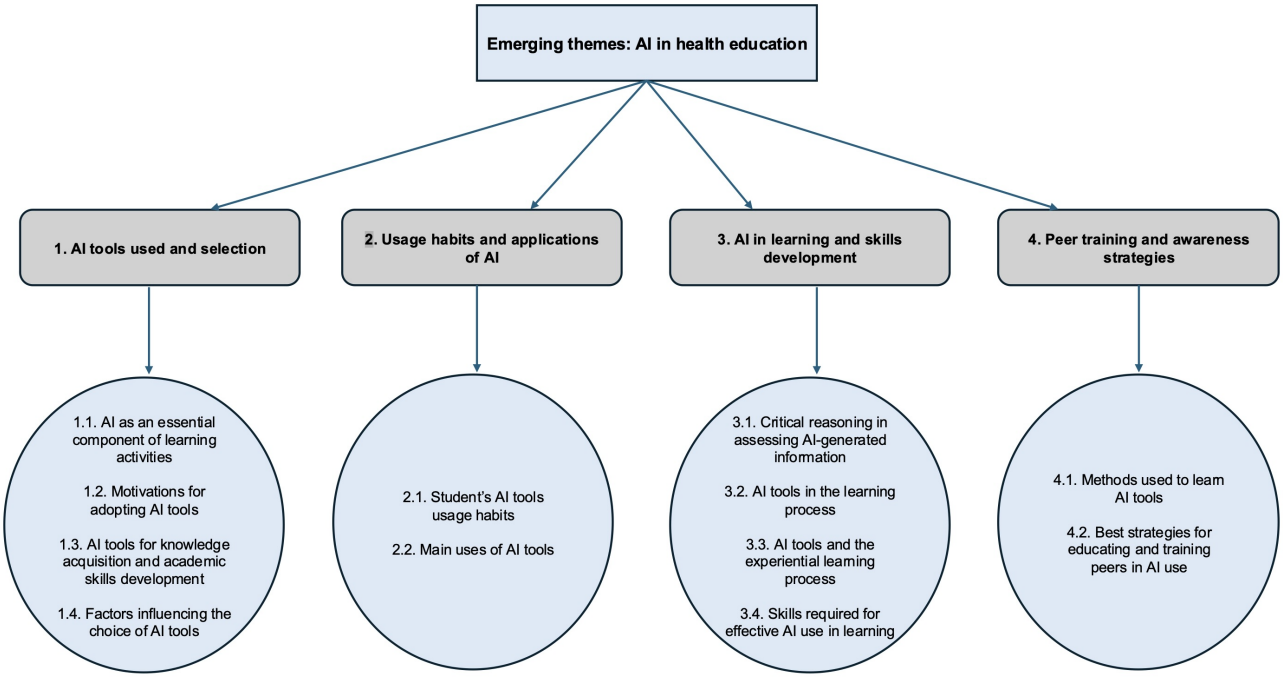

Theme 1: AI Tools Used and Selection

Theme 1.1: AI as an Essential Component of Learning Activities

Health care students have varying opinions about the role of AI in their learning activities. Among 51 respondents, 30 rated AI as essential or very beneficial. These students highlight its ability to enhance learning by helping to organize course materials, provide quick answers, and personalize educational experiences. For example, one participant illustrated this by stating:

> *I have used ChatGPT to help summarize papers and great sample questions and to read my work and compare it to the rubric.* [Participant 38]

They highlighted its usefulness in saving time and improving learning through tailored explanations and help with challenges like clarifying complex ideas or improving writing. As one participant noted:

> *Yes. I think it is an essential component because it can offer tools to approach/view your research in new ways. Therefore, it can aid in overcoming any hurdles, like feeling stuck when writing.* [Participant 23]

However, 21 respondents did not perceive AI as essential. Many of them regarded AI as a complementary aid rather than a primary resource, recognizing its usefulness while questioning its necessity. As one respondent noted:

> *I don't consider it to be essential, but I consider it to be helpful.* [Participant 19]

Concerns were raised about the reliability of AI, with one participant citing the presence of incomplete or inaccurate information on online platforms. Others reported using AI for targeted tasks but did not consider it integral to their overall learning process. A few respondents had only recently begun using AI and, although they acknowledged its potential, did not yet view it as indispensable. For example, one participant explained:

> *I only started using AI recently and I would not necessarily say it is essential component but rather an assistant to my learning...It is more of a helper than essential at this point in my journey.* [Participant 25]

Theme 1.2: Motivations for Adopting AI Tools

Curiosity was a key factor. Some respondents were intrigued by the widespread debate and excitement surrounding AI, particularly ChatGPT. Often encouraged by others, they were motivated to explore AI by the desire to see how it could be useful in their academic and professional lives.

A few respondents (n=2) were early adopters and experimenting with AI from its initial stages. One began exploring ChatGPT at its initial public launch, driven by curiosity and social media exposure:

> *Honestly, when I started experimenting with AI tools, especially ChatGPT, it was the first one I used, I think it was really in December 2022, during its first days of public access...it was out of fear, out of curiosity and I had heard about it...* [Participant 1]

The influence of peers, colleagues, and professors was also significant for 15 respondents, who were encouraged to try AI after observing others using it or receiving recommendations. This indicates the role of social and professional networks in the adoption of new technologies. As one student noted:

> *My friend suggested that I try it out for help.* [Participant 48]

XSL•FO
RenderX

Another shared:

> *My supervisor actually recommended it for my data analysis; they told me to use it when I am unsure of how to use Excel or SPSS to work with data.* [Participant 35]

Practical considerations, such as improving efficiency, saving time, and enhancing convenience, motivated 14 respondents to adopt AI tools. They found AI helpful in streamlining tasks and making their work processes more manageable. One participant noted:

> *The fact that it is an accessible tool and allows me to save time.* [Participant 14]

Twelve respondents were also motivated to use AI for professional or academic needs, frequently to gain a better understanding of procedures or expectations before starting a task. Some students used it to make sense of academic processes, particularly while navigating an unfamiliar system. One respondent stated:

> *I am an international student, and I was not familiar with many academic based process and AI could make them clear for me.* [Participant 29]

Some respondents (n=5) turned to AI out of frustration with traditional search engines like Google, which they found unhelpful or overwhelming. They saw AI as a more responsive and accurate alternative. For example, one participant noted:

> *The answers/explanations that Google offered me were vague and took longer to find.* [Participant 7]

And another mentioned:

> *When Google wasn't giving me the answers I needed or was overwhelming me and confusing me instead of helping.* [Participant 43]

A few respondents were motivated by innovation, recognizing AI's cutting-edge potential to make their work easier and more unique. One student explained:

> *Innovation, being different and unique. Having my work done easier for me. Being ahead of the curve and being more innovative and productive.* [Participant 49]

Theme 1.3: AI Tools for Knowledge Acquisition and Academic Skills Development

ChatGPT is the predominant AI tool used by health professions students for learning, mostly for the development of academic skills, such as refining data analysis, understanding complex topics, and enhancing writing abilities. For example, one participant mentioned:

> *ChatGPT is good at helping me refine my data analysis. It helps me distinguish between certain nuances of data analysis etc. when I ask it questions.* [Participant 25]

And another highlighted the importance of writing:

> *I think it helps me to develop new skills in the writing process, including synthesizing information or teaching me how to write in a more concise manner.* [Participant 28]

However, some respondents reported not using AI tools for learning, with a few indicating unfamiliarity with these technologies, falling into the "none or N/A" category.

Beyond ChatGPT, other AI tools were also mentioned for their specific functionalities. Wolfram was praised for its mathematics capabilities, text-to-speech models for aiding study through text-to-speech, and Consensus for summarizing research papers. One student stated, highlighting the tool's role in academic research:

> *...I also like how Research Rabbit (I think it's AI?) helps find links from one paper to others that are similar, it makes literature searching much easier.* [Participant 23]

Finally, language learning proved important for some, as these tools helped improve grammar and clarity. As one student noted:

> *When AI helps me rephrase my sentences to make them clearer, it allows me to acquire new knowledge or skills on the grammatical side.* [Participant 12]

Theme 1.4: Factors Influencing the Choice of AI Tools

Most health professions students did not explore multiple AI tools before choosing their current ones. Only 6 out of 51 respondents experimented with different tools. However, 23 respondents shared insights into their choice, highlighting several key factors.

Ease of access was a significant factor, with 10 respondents preferring tools that are free or provided by educational institutions. These tools were chosen for their simplicity and minimal setup requirements:

> *I have free open access via my institution.* [Participant 24]

> *It is simple to use. I would be open to trying other tools, but when the tools require me to set up a whole account, pay, or use personal data to use them, I find this to be a deterrent.* [Participant 28]

Specific features also influenced the choice, with 7 respondents valuing functionalities such as summarizing scientific papers or handling documents:

> *I chose Consensus for literature synthesizing because it presented the information in the best manner and is set up to aid in research specifically.* [Participant 23]

One participant characterized the tool's benefits as "bilingual" [Participant 16], in addition to being free, rapid, and efficient, while also emphasizing the necessity of using the tool judiciously.

Familiarity played a role for 6 respondents, who chose tools they were already aware of, highlighting a preference for sticking with known options. One participant mentioned:

> *I started with ChatGPT as that is the one I kept hearing about. It's suited my needs thus far as I don't*

XSL•FO
RenderX

> *use it very often. If I hear of something better, I might explore that too.* [Participant 19]

Popularity influenced 5 respondents, who opted for widely recognized tools, believing them to be reliable. As one student noted:

> *ChatGPT seemed the most popular.* [Participant 37]

Finally, ease of use was crucial for 5 respondents, who favored tools that were user-friendly. Less commonly mentioned factors included perceived reliability, speed, and efficacy, underscoring the value of dependable and efficient tools. One student explained:

> *I chose ChatGPT because it seems the most reliable and I have had no reason to seek other tools.* [Participant 34]

Theme 2: Usage Habits and Applications of AI

Theme 2.1: Students' AI Tools Usage Habits

Health professions students showed varied practices regarding the frequency and contexts of AI tools use, focusing on educational support, writing assistance, and information retrieval. Many respondents (n=19) use AI to aid in understanding complex concepts, studying, and verifying knowledge, which indicates how these tools can enhance learning. As one student explained:

> *I use them when I don't understand homework questions or need to double check an answer.* [Participant 33]

Writing assistance is another significant use, with 13 respondents relying on AI for drafting, revising, and brainstorming ideas. AI is used not only in academic writing but also in professional communication, such as writing emails. As one participant noted:

> *For each of my courses that I undertake and also alongside my professional life, whether it's for writing emails or other things.* [Participant 2]

This highlights the versatility of AI in supporting various writing tasks. Some respondents, however, use AI infrequently or lack clear usage habits. Nearly 11 respondents expressed uncertainty about their AI usage, suggesting they are still exploring the tool's potential or have not fully integrated it into their routines. One student shared:

> *Very infrequently used so I do not have habits for use.* [Participant 48]

AI is also used for general information retrieval and research, with 7 respondents using it for quick access to relevant information. As mentioned by one participant:

> *I use it as a shortcut every time I do research or need information.* [Participant 14]

A small number of respondents use AI for specific tasks like coding or data management, demonstrating AI's value in specialized technical problem-solving. For instance, one student mentioned:

> *When I am really stuck on how to do something in data analysis software, I ask ChatGPT for the best way.* [Participant 35]

In terms of frequency, some respondents reported using AI daily, while others mentioned using it 1 to 3 times per week. The consistency of usage varied as well, with several describing their use as regular and others characterizing it as more occasional. This highlights the flexible nature of AI tools in meeting diverse needs.

Theme 2.2: Main Uses of AI Tools

Participants identified several common applications of AI tools, including in revision, writing, learning, research, and knowledge testing.

Revision and synthesis emerged as the most common use, with 28 respondents using AI to refine and organize information, edit text, and summarize complex material. AI is frequently used to structure ideas and enhance communication, including drafting emails and quizzing on notes. One student, for example, summarized their use as follows:

> *...revising notes, synthetizing information.* [Participant 41]

Writing is another significant area, with 22 respondents using AI for drafting, idea generation, and refining writing style. AI helps tailor content to specific tones and styles, enhancing both the creative and revision stages of writing. As one student mentioned:

> *I mostly use these tools to proofread my writing or to enhance my writing style when I'm hit with "writer's block."* [Participant 19]

For learning, 16 respondents rely on AI to clarify complex concepts, gain new knowledge, and validate their own reasoning. AI aids in understanding difficult topics and applying knowledge in practical scenarios, such as informing clinical practices in psychology. One psychology student noted:

> *...I also use them to verify the clarity of my own writing. If I have an intervention in mind for a client (I am in psychology), I ask what research recommends for the name of whichever diagnosis I want to treatment plan for. I then double check that it aligns with what I read...* [Participant 22]

In research, 8 respondents use AI to find specific answers, summarize articles, and navigate academic language. AI is valued for making research more accessible by bridging the gap between lay and academic terminology. One student explained:

> *It is very useful to research topics when you don't know the exact language used in the literature, you can ask question in more common language and get to the information about the names of the theories and components that allows you to do real research to find the actual research papers relevant.* [Participant 24]

Finally, knowledge testing involves using AI to create quizzes, check answers, and master complex tasks, such as data analysis in SPSS (IBM Corp):

XSL•FO
RenderX

> *I use it to ask how to analyse data in a more complicated way in SPSS. My education on SPSS was simple and was able to guide me to do simple tasks, but my research is a bit more complicated.* [Participant 35]

Theme 3: AI in Learning and Skills Development

Theme 3.1: Critical Reasoning in Assessing AI-Generated Information

According to the survey analysis, students pursuing health professions use a range of critical thinking techniques to evaluate information produced by AI, indicating varying degrees of confidence and reliance on these technologies.

Cross-referencing AI outputs with credible sources is the most popular strategy, as reported by 29 respondents. They often compare AI-generated content with peer-reviewed articles, course materials, or trusted resources to verify accuracy, demonstrating a cautious approach to ensure that the AI information aligns with established knowledge. As one respondent noted:

> *I compare it to peer-reviewed articles to see if the findings match up.* [Participant 51]

Another prevalent strategy, mentioned by 16 respondents, involves using personal knowledge and common sense to evaluate AI content. These users rely on their prior understanding and logical reasoning to judge the validity of the information. For example, one respondent stated:

> *I like to see whether it makes sense and matches up with notes taken in class and information I already have and know.* [Participant 45]

Reviewing and editing AI-generated content is also common, with 17 respondents indicating that they scrutinize and revise AI outputs to ensure alignment with their expectations. This active engagement helps refine the content before it is used. As mentioned by one participant:

> *I always read it over, take out things that don't make sense or I don't agree with, and then rewrite it myself.* [Participant 46]

Theme 3.2: AI Tools in the Learning Process

AI tools play a nuanced role in the learning process, supporting dynamic engagement with knowledge through social, cognitive, and affective dimensions to develop cognitive structures.

Perception and Interaction With Information

Nearly 12 respondents noted that AI tools aid in recognizing and understanding information, enhancing the initial phase of learning. AI also facilitates interaction with data, with respondents emphasizing its role in manipulating and engaging with information crucial for deepening understanding. One student explained:

> *I believe AI tools help to perceive objects from a different perspective, in order to better understand the material.* [Participant 18]

Integration and Cognitive Development

A total of 12 respondents asserted that AI supports learners in embedding new skills and knowledge into their broader contexts, helping them integrate what they learn into their daily lives. Additionally, AI tools assist in transforming and refining cognitive structures, playing a key role in advancing cognitive development through learning processes. As one participant explained:

> *I would say that AI tools allow to transform, to create, and to develop the cognitive structure. However, I would not say that AI tools are, in themselves, information or knowledge. They should rather be used to guide or confirm our information and our knowledge.* [Participant 16]

Skills Development

AI's impact on skill acquisition was acknowledged, though its influence on attitudes and values was considered minimal. Four respondents generally viewed AI as more effective in enhancing cognitive and skill-based learning rather than shaping affective outcomes. As one respondent noted:

> *AI fits into the knowledge and information objects of learning. AI cannot teach new attitudes or values, but it can present new information and knowledge, as well as new ways to learn skills. I don't think it is advanced enough to help individuals perceive and interact with objects.* [Participant 36]

Information and Knowledge Acquisition

The most frequently mentioned theme was AI's role in acquiring and refining information and knowledge. A total of 23 respondents identified AI as a critical facilitator in these areas, underscoring its importance in the learning process. One student mentioned:

> *I think AI fits in terms of knowledge acquisition and information gathering, both of which can help with learning.* [Participant 48]

Theme 3.3: AI Tools and the Experiential Learning Process

In our survey of health professions students on AI tools in education, we examined how these tools align with phases of experiential learning: concrete experience, reflective observation, abstract conceptualization, and active experimentation.

Concrete Experience

Overall, 13 respondents identified AI tools as fitting within this phase, where learners interact with new or known situations to gain direct experiences. Highlighting their role in engaging directly with problems, one comment mentioned:

> *AI tools can be a great way to get experience and learn for example, if I am stuck in finding a code in R that gives me the statistical analysis I need. I can ask ChatGPT and troubleshoot with the code it gives me, study the code and integrate it into my bank of knowledge for next time.* [Participant 27]

XSL•FO
RenderX

Reflective Observation

A total of 14 responses placed AI tools in this phase, where learners analyze and evaluate their experiences. Participants noted that AI helps in reflecting on and assessing outcomes, with one student mentioning:

> *Reflective observation, which consists of analyzing and evaluating the experience.* [Participant 44]

Another shared:

> *AI tools allow us to analyse data, simulate scenarios, and create links with known information.* [Participant 43]

Abstract Conceptualization

This was the most cited phase, with 21 responses. AI tools were recognized for aiding in the development of theories from observations. One respondent remarked, underscoring their role in synthesizing and conceptualizing new ideas:

> *I use AI tools to understand things, to learn from existing theories, and to develop concepts to integrate the topics.* [Participant 14]

Active Experimentation

Nearly 19 respondents associated AI tools with this phase, where concepts are tested in new situations. AI facilitates this by enabling users to apply and test theoretical concepts practically. One student noted:

> *Active experimentation when testing what AI can do and its limits.* [Participant 50]

AI Does Not Fit Well Into Kolb's Model

A minority of responses (n=3) expressed skepticism about AI's fit within Kolb's phases, citing challenges in aligning AI functionalities with the model: "uncertain/did not understand the question." Two respondents were unsure or confused about how to position AI tools within Kolb's model.

Theme 3.4: Skills Required for Effective AI Use in Learning

Several essential skills were identified by the respondents for using AI tools effectively in learning. Technical competency is essential, as 22 respondents highlighted the necessity of comprehending the operation of AI systems, particularly the proper navigation of their features and settings to enhance learning outcomes.

Critical thinking and evaluation skills were also highlighted by 19 respondents, who emphasized the importance of analyzing AI outputs and ensuring they align with established knowledge or academic standards. This is particularly important given the potential for AI to produce inaccurate or incomplete information. One participant explained:

> *A variety of skills are needed, critical thinking and good judgement being very important as certain outputs generated by AI can be incorrect or misleading and knowing when to distinguish that is important.* [Participant 51]

And another participant stated:

> *Critical thinking and judgment are essential to determine whether the AI's response is appropriate in each situation. We must be able to detect biases (what the AI might have forgotten to consider, such as emotions, etc.).* [Participant 12]

Adaptability and flexibility are key, according to 14 respondents, who noted that users must be open to experimenting with AI and adjusting their strategies as they learn. One respondent stated:

> *...Flexibility is also important because you might need to make changes to your prompt to get a desired response from AI.* [Participant 36]

Effective communication was another important skill, with 12 respondents stressing the need to ask AI clear and specific questions to receive the most useful responses. As one respondent pointed out:

> *...knowing how to word questions properly to get the answer you need.* [Participant 27]

Finally, digital literacy was mentioned by 10 respondents, who recognized not only the importance of staying up to date with AI tools but also the need to use them responsibly and with academic integrity. As one student mentioned:

> *Being responsible when processing and using information. Maintaining academic integrity.* [Participant 2]

Theme 4: Peer Training and Awareness Strategies

Theme 4.1: Methods Used to Learn AI Tools

Health professions students used various methods to learn how to use AI tools. Self-directed exploration was the most common approach, with 23 respondents reporting that they learned by trying out different features and experimenting with the tools independently. One respondent said, which reflects a hands-on approach to mastering AI:

> *I just used trial and error.* [Participant 38]

Social learning also played a role, with 13 respondents indicating that they learned through friends, colleagues, or professors. These students benefited from observing others use AI or receiving direct guidance. This highlights the importance of peer-to-peer learning and mentorship in mastering new technologies. One student noted:

> *I heard about the tools from some of my friends and asked them for tips.* [Participant 33]

Online tutorials and guides were also used by 11 respondents, who turned to YouTube videos, articles, or AI tool help sections to gain a better understanding. As one student shared:

> *...I also saw a couple videos on TikTok about using select sentences/words for my desired effect.* [Participant 34]

Formal training, however, was mentioned by only two respondents, indicating that most AI learning happens informally and independently. One student explained:

> *I didn't need to learn anything, ChatGPT was easy to use.* [Participant 9]

XSL•FO
RenderX

Overall, students adopted diverse strategies for learning use of AI tools, with a clear preference for self-guided exploration and peer learning, supplemented by online resources.

Theme 4.2: Best Strategies for Educating and Training Peers in AI Use

To effectively educate and train peers in using AI tools, several strategies were highlighted by the respondents.

Practical, hands-on demonstrations were the most suggested approach, with 26 respondents emphasizing the importance of showing how the tools work in real time. For example, one respondent suggested:

> *I think concrete exercises would be more useful. For example, using it as part of a specific group project.* [Participant 15]

Workshops and guided sessions were also popular recommendations, with 19 respondents advocating for structured opportunities to learn AI in a collaborative environment. These sessions would allow peers to ask questions, practice using the tools, and receive feedback. As mentioned by one student:

> *Workshops and small group discussion sessions are the best.* [Participant 29]

And another elaborated:

> *I think workshops could be helpful as it allows for discussion on how to use these AI tools properly and effectively. It also allows for troubleshooting and working together to solve a problem in AI tools.* [Participant 27]

Providing easy-to-follow tutorials and online resources was another suggested strategy, with 15 respondents supporting the use of videos, articles, or step-by-step guides that could be shared among peers. One respondent explained:

> *Self-study modules that students can do at their own pace and learn new ways that can help use AI tools in a smart yet efficient way, without harming our education.* [Participant 45]

This method would allow for self-paced learning and continuous reference.

Lastly, creating supportive peer networks was recommended by 10 respondents, who emphasized the value of mentorship and peer-to-peer assistance in AI education. This strategy builds a community of learners who can share insights, troubleshoot issues, and offer mutual encouragement. One student stated:

> *I think the best strategy is through discussion/exposure by peers/colleagues...This is most appropriate because the discussion aspect primes you to begin thinking about it, so when you approach it individually, you can play around with different aspects based on what you need/want to pursue.* [Participant 23]

Discussion

Principal Findings

Our study collected data from 51 health science students through semistructured interviews and a questionnaire. The findings show the growing adoption of AI tools by students and underscore their importance in academic and clinical settings. Participants reported that general-purpose AI tools, including platforms like ChatGPT, were frequently used. The degree of AI tool use varied depending on the tasks students were required to perform, with some integrating them into their daily activities while others rarely used them. According to participants, AI tools are beneficial for varied learning activities, including rapid information comprehension, facilitating the generation of diverse perspectives, and simulation of varied clinical situations. Students also highlighted the need for collaborative strategies, such as peer learning groups, to promote and ensure effective AI use.

Students identified AI tools as essential to their learning activities. ChatGPT was the most cited tool, while other tools were mentioned only occasionally. This trend reflects the growing use of ChatGPT in health care students' learning activities, which could be explained by its accessibility, personalized responses, and ability to simulate engaging cognitive interactions. From the perspective of Kolb's model, this frequent use indicates a significant footprint in the concrete experience phase, while the use of AI to clarify, structure, and reorganize knowledge corresponds to the abstract conceptualization phase. These results are in line with a previous study, which highlighted the ability of ChatGPT to produce logical and contextually relevant explanations, as well as to simulate a logical thinking process, making it a promising didactic tool in the field of education [8,30]

Furthermore, students expressed a preference for using AI tools over traditional search engines like Google Search, due to the more targeted and precise nature of the results. This suggests that students are looking for tools that support conceptual organization and meaning-making, which corresponds to the abstract conceptualization phase of Kolb's experiential learning cycle. Previous studies confirmed this preference among students for ChatGPT over Google Search, as this tool appears to provide more accurate and comprehensive medical information for general knowledge requests and highlights its usefulness in the educational environment [8,31].

Students reported variability in their use of AI tools, ranging from occasional use to regular integration into their learning routines. This observed inconsistency appears to reflect significant differences in individual usage styles, the familiarity developed with AI technologies, and the nature of the tasks they were faced with. In the context of experiential learning, this variability may also indicate that students do not all engage in the learning cycle in the same way, with some focusing primarily on concrete experience and reflective observation, while others broaden their use to abstract conceptualization and active experimentation by reinvesting AI results in new academic or clinical tasks. These results are in line with recent studies confirming that the degree of use of AI tools depends on their

XSL•FO
RenderX

usefulness as well as specific academic contexts and tasks [32,33]. Understanding this nuance is important because it suggests the importance of promoting their use in an educational context by adopting a more individual rather than a universal approach [9,33].

Our results also confirmed that a lot of health care students use ChatGPT to learn and develop new skills, such as improving their writing, analyzing data, and simplifying complex ideas. These practices are part of Kolb's experiential learning theory and are explained by reflective observation of AI-generated results and abstract conceptualization. These findings are consistent with previous studies, which concluded that ChatGPT improves student writing, leading to more effective and higher-quality academic production [10,34]. Moreover, other more specialized tools used by these students, such as Consensus for synthesizing scientific articles, demonstrate a strategic and targeted use of AI tools to perform specific tasks, aligning with previous studies [35,36]. On the other hand, some students reported not using these technologies, often due to lack of interest or confidence, which is consistent with the writings of Johnston et al [37], who show that students who have greater confidence in their ability to write academic texts use ChatGPT or other generative AI tools less in their learning activities. These results show that efforts are still needed to promote equitable and critical integration of these tools into educational practices, although AI remains an important tool for skill development [37].

In terms of effective strategies for training peers to use AI tools, students identified hands-on demonstrations, guided workshops, and self-directed resources. These findings are part of Kolb's experiential learning theory and are explained by the importance of concrete experience (hands-on demonstrations), reflective observation (guided discussions and feedback), and active experimentation (autonomous reuse of tools in personal learning contexts). These approaches are consistent with the study of Jiang et al [38], which highlights the importance of active and collaborative learning to foster engagement and understanding in education, as well as with that of Abid et al [39], which demonstrates the effectiveness of flexible and online approaches in technology training. Participants also valued peer support, highlighting the central role of social collaboration and peer support in technology interactions, as suggested by previous studies [14,40]. These findings confirm that effective AI adoption requires a combination of hands-on activities, autonomy, and collaborative support.

Strengths

One of the strengths of this study is its methodological framework, which facilitated the collection of rich and nuanced opinions from health science students on the use of AI. Next, the study includes individuals with varied experiences using AI tools, both in academic and clinical environments. This highlights not only the different ways in which AI is used, but also the motivations and critical thinking strategies implemented by students. Finally, the highlighted didactic implications are directly applicable. They emphasize the need to prepare students for a critical approach to AI, to encourage tailored learning environments, and to promote the incorporation of AI into collaborative tasks.

Limitations

One of the limitations of this study is the small sample size and the use of self-reported data, and the fact that the data were collected within a single establishment, which limits the generalization of the results to other institutions. This may have introduced biases related to memorization accuracy or social desirability. Although the research team assessed the questionnaire for relevance and clarity, psychometric procedures were not sufficiently applied to assess it. Therefore, interpretation biases may arise, particularly regarding the understanding of porous concepts such as "learning" or "learning support." Furthermore, health professions students were the focus of this study; the application of AI in training may vary by discipline. In addition, this transition in data collection, from semistructured interviews to online surveys, may have reduced the depth and richness of some responses, which could affect the identification of more subtle themes.

Therefore, future research could address these limitations by conducting larger-scale studies with more diverse samples and using objective measures of AI usage, such as usage logs or performance data. Additionally, future research could explore how AI is used in other fields of study and whether the patterns observed in this study are generalizable to different educational contexts. Moreover, as AI tools continue to evolve, it will be important to track how students' perceptions and utilization habits change over time.

Conclusions

In conclusion, this study aimed to identify the AI tools most used by health students at the University of Ottawa, to determine their usage habits as well as the tools helping them develop skills, and to examine the strategies they consider effective for training their peers in the use of these tools. Our findings highlight the growing role of AI in health care education and provide valuable insights into how these tools have been adopted and used by students. While AI is considered essential to their learning activities, it becomes crucial to further integrate it into educational programs and help students develop their digital skills while developing their critical thinking skills. As these technologies gain popularity, they have the potential to enhance the learning experience of health care students, but careful attention must be paid to how these tools are promoted and used in the educational environment.

Acknowledgments

We would like to thank the "Affaires francophones" of the Faculty of Medicine, University of Ottawa, for supporting this research. We also acknowledge the use of AI tools during the research and writing process. Microsoft Copilot 365 was used to support the writing process under human supervision, more specifically to correct and improve sentence structure and phrasing, and not to

XSL•FO
RenderX

generate content. To help with the search for relevant scientific literature, ChatGPT (OpenAI) and Elicit (Ought) were used. All AI-generated content was reviewed and edited by the authors, who remain fully responsible for the originality, accuracy, and integrity of the final work.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The data generated and analyzed in this study are not publicly available due to institutional ethics and confidentiality regulations. However, they may be obtained from the corresponding author upon reasonable request, subject to ethics committee approval.

Authors' Contributions

All authors contributed equally to the conception, design, data collection, analysis, and writing of this manuscript. Each author reviewed and approved the final version for submission and agrees to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Question guide and distribution of artificial intelligence tools used by health professions students at the University of Ottawa.
[PDF File, 161 KB - jmir_v28i1e82432_app1.pdf]

Checklist 1
SRQR checklist.
[PDF File, 124 KB - jmir_v28i1e82432_app2.pdf]

References

1. Rahman MA, Victoros E, Ernest J, Davis R, Shanjana Y, Islam MR. Impact of artificial intelligence (AI) technology in healthcare sector: a critical evaluation of both sides of the coin. Clin Pathol 2024;17. [doi: 10.1177/2632010X241226887] [Medline: 38264676]
2. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021 Jul;8(2):e188-e194. [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
3. Zavaleta-Monestel E, Anchía-Alfaro A, Rojas-Chinchilla C, Quesada-Loria DF, Arguedas-Chacón S. Ethical and practical dimensions of artificial intelligence (AI) in healthcare: a comprehensive study of professional perceptions. Cureus 2025 Feb;17(2):e78416. [doi: 10.7759/cureus.78416] [Medline: 40046390]
4. Najjar R. Redefining radiology: a review of artificial intelligence integration in medical imaging. Diagnostics (Basel) 2023 Aug 25;13(17):2760. [doi: 10.3390/diagnostics13172760] [Medline: 37685300]
5. Fujisawa Y, Otomo Y, Ogata Y, et al. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. Br J Dermatol 2019 Feb;180(2):373-381. [doi: 10.1111/bjd.16924] [Medline: 29953582]
6. Al Kuwaiti A, Nazer K, Al-Reedy A, et al. A review of the role of artificial intelligence in healthcare. J Pers Med 2023 Jun 5;13(6):951. [doi: 10.3390/jpm13060951] [Medline: 37373940]
7. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023 Mar 6;9(1):e46885. [doi: 10.2196/46885] [Medline: 36863937]
8. Ghorashi N, Ismail A, Ghosh P, Sidawy A, Javan R. AI-powered chatbots in medical education: potential applications and implications. Cureus 2023 Aug;15(8):e43271. [doi: 10.7759/cureus.43271] [Medline: 37692629]
9. Teng M, Singla R, Yau O, et al. Health care students' perspectives on artificial intelligence: countrywide survey in Canada. JMIR Med Educ 2022 Jan 31;8(1):e33390. [doi: 10.2196/33390] [Medline: 35099397]
10. Mousavi Baigi SF, Sarbaz M, Ghaddaripouri K, Ghaddaripouri M, Mousavi AS, Kimiafar K. Attitudes, knowledge, and skills towards artificial intelligence among healthcare students: a systematic review. Health Sci Rep 2023 Mar;6(3):e1138. [doi: 10.1002/hsr2.1138] [Medline: 36923372]
11. Schneider M, Stern E. L'apprentissage dans une perspective cognitive [in French]. In: Dumont H, Istance D, Benavides F, editors. Comment apprend-on ? La recherche au service de la pratique: OCDE; 2010:73-95.

XSL•FO
RenderX

12. Hsu WC, Fuh LJ, Liao SC. Tickling the heart: integrating social emotional learning into medical education to cultivate empathetic, resilient, and holistically developed physicians. Front Med (Lausanne) 2024;11:1368858. [doi: 10.3389/fmed.2024.1368858] [Medline: 38500950]
13. Shorey S, Chan V, Rajendran P, Ang E. Learning styles, preferences and needs of generation Z healthcare students: scoping review. Nurse Educ Pract 2021 Nov;57:103247. [doi: 10.1016/j.nepr.2021.103247] [Medline: 34768214]
14. Joseph GV, Athira P, Anit Thomas M, Jose D, Roy TV, Prasad M. Impact of digital literacy, use of AI tools and peer collaboration on AI assisted learning: perceptions of the university students. Digital Edu Rev 2024(45):43-49. [doi: 10.1344/der.2024.45.43-49]
15. Msambwa MM, Wen Z, Daniel K. The impact of AI on the personal and collaborative learning environments in higher education. Eur J Edu 2025;60(1):e12909. [doi: 10.1111/ejed.12909]
16. Fawaz M, El-Malti W, Alreshidi SM, Kavuran E. Exploring health sciences students' perspectives on using generative artificial intelligence in higher education: a qualitative study. Nurs Health Sci 2025 Mar;27(1):e70030. [doi: 10.1111/nhs.70030] [Medline: 39797529]
17. Benboujja F, Hartnick E, Zablah E, et al. Overcoming language barriers in pediatric care: a multilingual, AI-driven curriculum for global healthcare education. Front Public Health 2024;12:1337395. [doi: 10.3389/fpubh.2024.1337395] [Medline: 38454985]
18. Busetto L, Wick W, Gumbinger C. How to use and assess qualitative research methods. Neurol Res Pract 2020;2(1):14. [doi: 10.1186/s42466-020-00059-z] [Medline: 33324920]
19. Hunt MR. Strengths and challenges in the use of interpretive description: reflections arising from a study of the moral experience of health professionals in humanitarian work. Qual Health Res 2009 Sep;19(9):1284-1292. [doi: 10.1177/1049732309344612] [Medline: 19690208]
20. Locsin RC, Schoenhofer SO. Qualitative research methods, inductive and deductive: valuable approaches to nursing knowledge development for practice. Nurs Health Sci 2024 Jun;26(2):e13128. [doi: 10.1111/nhs.13128] [Medline: 38804111]
21. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for Reporting Qualitative Research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
22. Murphy E, Dingwall R, Greatbatch D, Parker S, Watson P. Qualitative research methods in health technology assessment: a review of the literature. Health Technol Assess 1998;2(16). [Medline: 9919458]
23. Higginbottom GMA. Sampling issues in qualitative research. Nurse Res 2004;12(1):7-19. [doi: 10.7748/nr2004.07.12.1.7.c5927] [Medline: 15493211]
24. Guest G, Bunce A, Johnson L. How many interviews are enough? An experiment with data saturation and variability. Field methods 2006;18(1):59-82. [doi: 10.1177/1525822X05279903]
25. Schneider S, Beege M, Nebel S, Schnaubert L, Rey GD. The cognitive-affective-social theory of learning in digital environments (CASTLE). Educ Psychol Rev 2022;34(1):1-38. [doi: 10.1007/s10648-021-09626-5] [Medline: 34226808]
26. Adams NE. Bloom's taxonomy of cognitive learning objectives. J Med Libr Assoc 2015 Jul;103(3):152-153. [doi: 10.3163/1536-5050.103.3.010] [Medline: 26213509]
27. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development: Prentice Hall; 1983.
28. Paillé P, Mucchielli A. L'analyse par questionnement analytique [in French]. In: L'analyse qualitative en sciences humaines et sociales: Armand Colin; 2012:207-230.
29. Dhakal K. NVivo. J Med Libr Assoc 2022 Apr 1;110(2):270-272. [doi: 10.5195/jmla.2022.1271] [Medline: 35440911]
30. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
31. Ayoub NF, Lee YJ, Grimm D, Divi V. Head-to-head comparison of ChatGPT versus Google Search for medical knowledge acquisition. Otolaryngol Head Neck Surg 2024 Jun;170(6):1484-1491. [doi: 10.1002/ohn.465] [Medline: 37529853]
32. Saleh ZT, Rababa M, Elshatarat RA, et al. Exploring faculty perceptions and concerns regarding artificial intelligence chatbots in nursing education: potential benefits and limitations. BMC Nurs 2025 Apr 18;24(1):440. [doi: 10.1186/s12912-025-03082-0] [Medline: 40251532]
33. Alqaisi N, Alshwayyat S, Aburumman S, et al. Assessing ChatGPT adoption in Jordanian medical education: a UTAUT model approach. BMC Med Educ 2025 May 22;25(1):750. [doi: 10.1186/s12909-025-07336-y] [Medline: 40405195]
34. Wang J, Liao Y, Liu S, et al. The impact of using ChatGPT on academic writing among medical undergraduates. Ann Med 2024 Dec;56(1):2426760. [doi: 10.1080/07853890.2024.2426760] [Medline: 39555617]
35. Fabiano N, Gupta A, Bhambra N, et al. How to optimize the systematic review process using AI tools. JCPP Adv 2024 Jun;4(2):e12234. [doi: 10.1002/jcv2.12234] [Medline: 38827982]
36. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: comparative analysis of ChatGPT, Bing, and medical students in Germany. JMIR Med Educ 2023 Sep 4;9:e46482. [doi: 10.2196/46482] [Medline: 37665620]
37. Johnston H, Wells RF, Shanks EM, Boey T, Parsons BN. Student perspectives on the use of generative artificial intelligence technologies in higher education. Int J Educ Integr 2024 Dec;20(1):1-21. [doi: 10.1007/s40979-024-00149-4]

XSL•FO
RenderX

38. Jiang J, Huang K, Martinez-Maldonado R, Zeng H, Gong D, An P. Novobo: supporting teachers' peer learning of instructional gestures by teaching a mentee AI-agent together. arXiv. Preprint posted online on May 26, 2025. [doi: 10.48550/arXiv.2505.17557]
39. Abid A, Murugan A, Banerjee I, Purkayastha S, Trivedi H, Gichoya J. AI education for fourth-year medical students: two-year experience of a web-based, self-guided curriculum and mixed methods study. JMIR Med Educ 2024 Feb 20;10:e46500. [doi: 10.2196/46500] [Medline: 38376896]
40. Tao W, Yang J, Qu X. Utilization of, perceptions on, and intention to use AI chatbots among medical students in China: national cross-sectional study. JMIR Med Educ 2024 Oct 28;10:e57132. [doi: 10.2196/57132] [Medline: 39466038]

ABBREVIATIONS

AI: artificial intelligence
REB: Research Ethics Board
SRQR: Standards for Reporting Qualitative Research

Edited by A Stone; submitted 14.Aug.2025; peer-reviewed by A Jaleel, F Fukuzawa; revised version received 26.Feb.2026; accepted 27.Feb.2026; published 02.Apr.2026.

Please cite as:
Guirguis M, Fotsing S, Fevry J, Landry C, Bouchard-Lamothe D, Lacroix J, Jalali A
Artificial Intelligence in Health Professions Education: Qualitative Study of Student Experiences
J Med Internet Res 2026;28:e82432
URL: https://www.jmir.org/2026/1/e82432
doi:10.2196/82432

© Marina Guirguis, Salomon Fotsing, Joanne Fevry, Christine Landry, Diane Bouchard-Lamothe, Jennifer Lacroix, Alireza Jalali. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Improving Retrieval Augmented Generation for Health Care by Fine-Tuning Clinical Embedding Models: Development and Evaluation Study

Kamyar Arzideh[1,2], MSc; Henning Schäfer[1,3], PhD; Ahmad Idrissi-Yaghir[1,4], MSc; Cynthia Sabrina Schmidt[1,3,5], MD; Bahadir Eryilmaz[1], MSc; Mikel Bahn[1], MSc; Amin T Turki[1,6], MD, PhD; Olivia Barbara Pollok[4], MD; Eva Maria Hartmann[1,2], MSc; Philipp Winnekens[2], MSc; Katarzyna Borys[1,4], MSc; Johannes Haubold[1,4], Prof Dr Med; Felix Nensa[1,4], Prof Dr Med; René Hosch[1,4], PhD

[1]Institute for Artificial Intelligence in Medicine,, University Hospital Essen, Essen, Germany

[2]Central IT Department, Data Integration Center, University Hospital Essen, Girardetstr. 2, Essen, Germany

[3]Institute for Transfusion Medicine, University Hospital Essen, Essen, Germany

[4]Institute of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Essen, Germany

[5]Center of Sleep and Telemedicine, University Hospital Essen, Ruhrlandklinik, Essen, Germany

[6]Department of Hematology and Oncology, Ruhr-University Bochum, Marienhospital University Hospital, Bochum, Germany

Corresponding Author:
Kamyar Arzideh, MSc
Institute for Artificial Intelligence in Medicine,, University Hospital Essen, Essen, Germany

Abstract

Background: Embedding models are critical components of Retrieval Augmented Generation (RAG) systems for retrieving and searching unstructured medical data. However, existing models are predominantly trained on publicly available English datasets, limiting their effectiveness in non-English health care settings. More importantly, these models lack training on real-world clinical documents, leading to inaccurate context retrieval when integrated into RAG systems for health care applications. This gap is particularly pronounced in specialized medical documentation containing domain-specific terminology, abbreviations, and nuanced clinical language.

Objective: This retrospective study aimed to develop and validate embedding models specifically trained on real-world clinical documents from multiple medical specialties to improve medical information retrieval (IR) and RAG system performance in both German and English language contexts.

Methods: We fine-tuned embedding models, so-called sentence transformers, using the multilingual-e5-large architecture as a foundation. Training data consisted of approximately 11 million question-answer pairs synthetically generated from 400,000 diverse clinical documents from a large German tertiary hospital, spanning 163,840 patients and 282,728 clinical cases between 2018 and 2023. The *SauerkrautLM-SOLAR-Instruct* large language model generated medically relevant questions and corresponding answers for each document. The dataset was additionally pseudonymized and translated into English to aim for broader applicability. Models were evaluated in 2 distinct scenarios: IR using questions with multiple relevant passages, and RAG system performance in both cross-patient and patient-centered contexts.

Results: In the IR evaluation, the fine-tuned miracle model achieved a mAP@100 of 0.27, outperforming the multilingual-e5-large baseline (0.14) and state-of-the-art models such as bge-m3 (0.11). In the RAG evaluation, the model demonstrated robust performance comparable with the baseline in the constrained patient-centered scenario (BERTScore F1 0.781 vs 0.778) and showed moderate improvements in the unconstrained cross-patient setting (BLEURT 0.56 vs 0.53). Notably, the model trained on pseudonymized data achieved comparable retrieval performance (mAP@100 0.25) and the highest scores for patient-centered contextual precision (0.93). Performance gains were robust in the German dataset, while the translated English model demonstrated promising results as a proof of concept for cross-lingual transfer.

Conclusions: By leveraging a comprehensive real-world dataset spanning multiple medical specialties and using large language models for synthetic question generation, we successfully created and validated domain-specific embedding models. These models can improve medical IR in large-scale search spaces and perform competitively in constrained RAG applications. By publishing the models trained on pseudonymized data, other health care institutions can integrate or adapt these embedding models to their needs. This work establishes a reproducible framework for developing domain-specific clinical embedding models, with the potential to improve data retrieval in medical settings.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e82997) doi:10.2196/82997

KEYWORDS

Retrieval Augmented Generation; RAG; natural language processing; information retrieval; large language models; LLM; NLP

Introduction

Integrating natural language processing (NLP) into health care has opened new application areas to improve clinical decision-making and patient care [1-4]. Central to this advancement was the development of language and domain-specific language models [5-9], which are important for accurately accessing and understanding medical documents. In the field of medicine, however, several obstacles have so far limited the applicability of these new tools. First, medical language [10,11] is complex and documentation characterized by specialized terminology, jargon, and abbreviations [12,13], representing a significant challenge for language models in the health care domain. To address this challenge, it is essential to design models that are specifically tailored to the health care sector. Hence, NLP models, mainly developed and trained on publicly available English datasets, fall short in understanding and processing natural language in non-English linguistic contexts [14,15]. Second, medical documents are characterized by a high degree of specificity and contextual nuances, which the existing general language models often fail to capture [16-18]. Consequently, the absence of domain and language-specific models has proven to be a substantial obstacle [19-21] for an effective implementation of artificial intelligence and NLP technologies.

Embedding models learn to represent words or phrases as mathematical vectors, capturing their meaning based on the context in which they appear, making them especially valuable in fields where terms can have different meanings depending on context. Integration of these models into Retrieval Augmented Generation (RAG) systems marks an advancement in information retrieval (IR). RAG may enhance the performance of large language models (LLMs) by combining them with an existing knowledge base or document repository [22]. Typically, RAG is implemented by transforming documents or records into vectors. In response to a specific query, the most similar documents are retrieved and then incorporated into the LLM prompt. They combine retrieval-based and generative NLP models and are heavily reliant on the quality of underlying embedding models for their effectiveness [23]. However, these embedding models, also known as sentence transformers, are primarily trained on publicly available datasets. Using these models in a hospital context with routine documents is potentially difficult as these models have not been trained on real-world unstructured clinical texts. Integrating these models into RAG systems to search through patient documents could lead to undesired or incorrect results. This affects the overall adaptation of RAG systems in clinical practice and highlights the importance of domain fine-tuning.

The development of domain-specific embedding models has the potential to improve RAG systems in the clinical context. By providing more accurate, reliable, and context-aware interpretations of medical texts, such models may improve the information basis for medical decision-making and, via decision support systems, ultimately improve patient outcomes. Given the relevant limitations of existing models, the objective of this study was to develop medical embedding models, exemplarily both in German and English. These models were trained on a large corpus of real-world clinical documents of a large tertiary hospital. This domain-specific training should enhance the performance of RAG systems for patient-related questions. By adapting the models to the medical field, we aimed to enhance their capability to understand hospital jargon and interpret the complex semantic relationships and clinical contexts in medical documentation.

Methods

Ethical Considerations

This study was approved by the ethics committee of the Medical Faculty of the University of Duisburg-Essen (approval number 23 - 11557-BO). Due to the study's retrospective nature, the requirement of written informed consent was waived by the ethics committee of the Medical Faculty of the University of Duisburg-Essen. All methods were carried out in accordance with relevant guidelines and regulations and the Declaration of Helsinki. No identifying information or images of individual participants are included in this manuscript or its supplementary materials. Participants did not receive any compensation.

Data Definition

Our study used a carefully curated dataset comprising a balanced collection of German language clinical notes documented at the University Hospital Essen. The dataset included 100,000 documents from each of the following 4 medical documentation categories: radiology reports, discharge letters, pathology reports, and surgical operation notes. This diverse composition ensures a comprehensive representation of clinical language and terminology. The clinical notes were documented between 2018 and 2023 in the clinical information system of the investigating site and contain information related to the medical history of patients, including diagnostic data and treatment documentation.

The total dataset consisted of 400,000 clinical documents from 163,840 patients and 282,728 clinical cases, resulting in a total number of 163,479,394 tokens, of which 1,541,389 were unique, reflecting the depth and complexity of the vocabulary present in the dataset. Further statistical details, including average document length, token frequency distribution, and other relevant metrics, are summarized for each document type in Table 1.

XSL•FO
RenderX

Table . Dataset characteristics[a].

Dataset	Total number of tokens	Total number of unique tokens	Total number of sentences	Average tokens per document	Average sentence length
Discharge letters	86,512,538	901,771	4,406,969	865.13	19.63
Pathology reports	22,313,259	333,028	1,901,434	223.13	11.73
Radiology reports	15,565,379	307,141	1,202,033	155.65	12.95
Operation notes	39,088,218	467,116	3,061,313	390.88	12.77
Total	163,479,394	1,541,389	10,571,749	408.70	15.46

[a]Displayed are quantitative measures for length and average length across the datasets. “Total number of tokens” was calculated by counting the number of tokens or words in the documents, including punctuation. The lemma form of each token was created and the number of lemmas was used as an approximation for the “total number of unique tokens.” “Total number of sentences” represents the count of all sentences per document. “Average tokens per document” was determined by dividing the total number of tokens by the total number of documents. “Average sentence length” was calculated by dividing the number of tokens by the total number of sentences.

Dataset Generation

The dataset containing 400,000 documents was segmented into smaller chunks using a recursive text splitting approach implemented in the Python (Python Software Foundation) package *langchain* with a maximum size of 450 characters and an overlap of 80 characters. This overlap was selected to ensure semantic continuity at text segment boundaries, accepting a minor degree of redundancy to prevent loss of clinical context at split points. The maximum character length was informed by the specific structural characteristics of the documents. Clinical information is typically recorded in concise, information-dense paragraphs or itemized lists rather than long-form narrative prose. This splitter uses a hierarchical approach to preserve semantic boundaries by attempting to split text at natural break points in the following priority order: paragraph breaks (\n\n), line breaks (\n), word boundaries (spaces), and finally individual characters. This methodology ensures that complete sentences and paragraphs are preserved whenever possible, only breaking midsentence when individual sentences exceed the 450-character limit. The chunking process resulted in 3,164,164 chunks across all document types.

These chunks were given as input to the *SauerkrautLM-SOLAR-Instruct* model, which has 10.7B parameters and was trained on German translated data. For each chunk, this LLM was tasked with generating 5 medically relevant questions alongside the correct answer that is contained in the document chunk, with the intention of simulating a real-world clinical questioning scenario. The model was deployed within the secured clinical network in order to ensure data protection for patients and physicians.

This process was applied across all collected documents—radiology reports, discharge letters, pathology reports, and operation notes. As a result, a dataset comprising 15,820,820 medical question-answer pairs was created as a training basis for the LLM. This method was chosen to capture the essence of clinical language and reflect the practical, inquiry-driven nature of medical diagnostics and patient care. To reduce the occurrence of hallucinations associated with the LLM and to enhance the data quality, the question-and-answer pairs were systematically filtered through the application of defined criteria. All questions were required to conclude with a question mark and to start with a number between 1 and 5, ensuring the generation of exactly 5 questions for each chunk. After this process, 11,000,227 questions-answer pairs were used for further training and evaluation steps. An overview of the methodologies including sample questions and answers is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Overview of training and evaluation methods**.** For each text chunk, 5 question-answer pairs were generated with a large language model. The question and chunk pairs were then used to fine-tune sentence transformer models. Different evaluation scenarios were set up to measure the performance of the models. This graphic was created in BioRender.

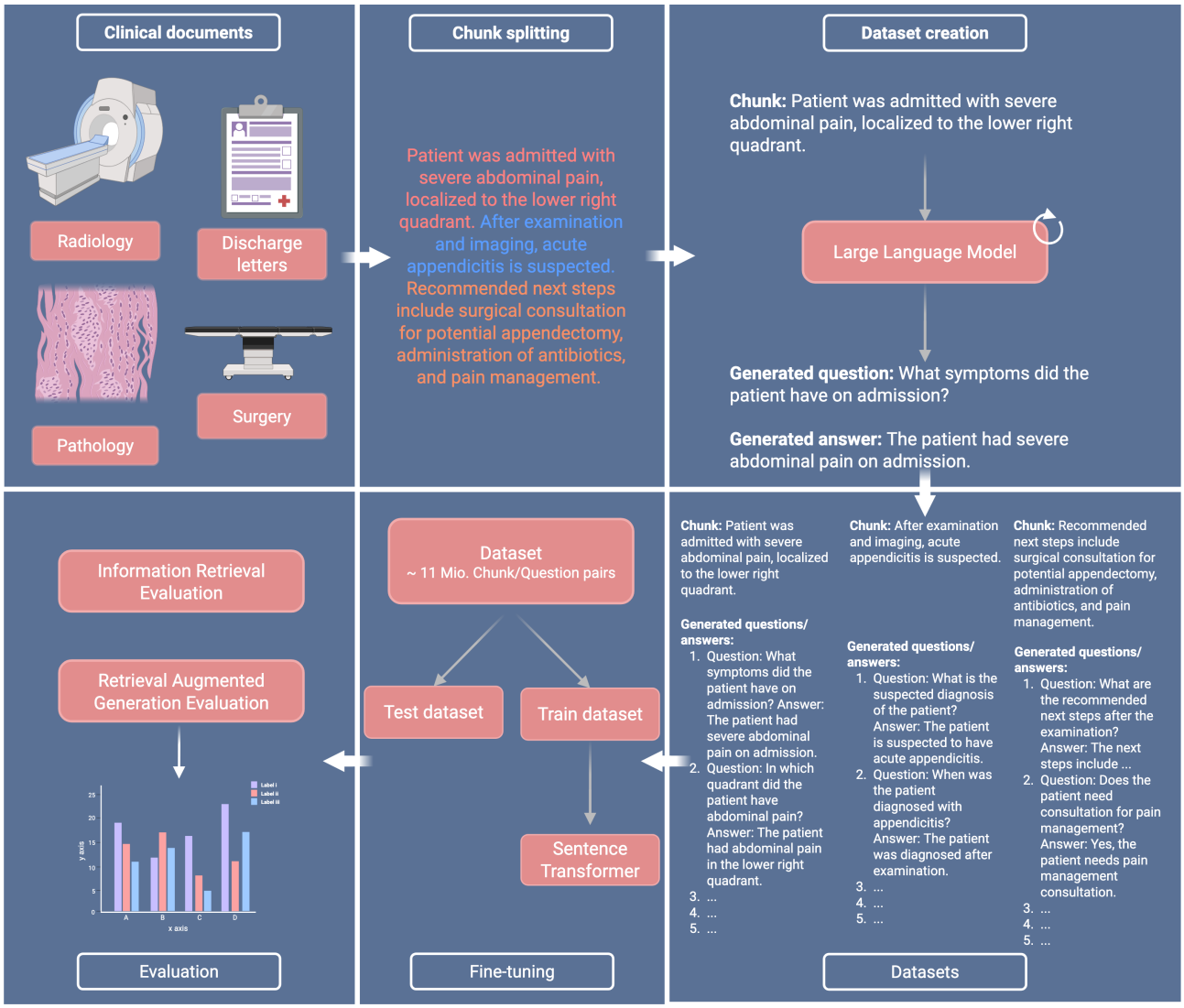

Dataset Pseudonymization

As the generated dataset comprises personally identifiable data belonging to patients and health care personnel, it underwent a pseudonymization process using a deidentification pipeline developed at the investigating site [24]. Protected Health Information in the documents was identified and replaced with surrogates. All training and evaluation procedures were additionally conducted on pseudonymized documents, and performance was compared with the model fine-tuned on real-world data.

Dataset Translation

The language used in the clinical documents is German, which limits the applicability of the models to German-speaking hospitals. To overcome this limitation, the dataset was translated into English using the Fairseq WMT'19 German-to-English translation model [25]. Although translating medical content can be complex and may result in inaccuracies due to the specialized terminologies involved, the translation process serves as a method for augmenting the corpus and increasing its accessibility. The decision to use this particular translation model was informed by a head-to-head translation model comparison experiment, including physician's evaluation. The physician was presented with 30 samples produced by various translation models, including WMT19-de-en, M2M-100 [26], NLLB [27], T5 [28], MBart-50 [29], and OPUS-MT-de-en [30]. The feedback on the medical accuracy and readability of the translations informed the final selection of the WMT'19 model for this task.

Model Training

Fine-tuning was performed using pairs of real document passages and the generated questions described in the "Dataset Generation" subsection. The data were split 80%/20% into a training and a test dataset on a patient basis so that a patient's data were part of either the training or the test dataset. This resulted in 8,801,525 chunk-and-question pairs for training and 2,198,702 pairs for testing. Hyperparameters used during training include an initial learning rate of 2e-5 and a batch size of 1024. For learning rate optimization, the AdamW optimizer was used. As a loss function, the *CachedMultipleNegativesRankingLoss* from the sentence transformer library [31] with a minibatch size of 32 was used.

XSL•FO
RenderX

To normalize the embedding size, pooling was used to create fixed-sized sentence embeddings of 1024 dimensions. Each training run was performed on 1 NVIDIA H100 80GB graphical processing unit for approximately 8 days.

Training was limited to a single epoch. Given the size of the dataset and the synthetic nature of the training data, this approach was selected to prevent overfitting to specific linguistic patterns generated by the LLM. Additionally, training runs showed that training loss plateaued after 1 training epoch.

Evaluation

Overview

Different techniques were used to measure the performance of the fine-tuned models. The models were evaluated on different datasets across 2 distinct evaluation settings—IR and RAG. The reason for these different evaluation scenarios is the difficulty of measuring the performance of sentence transformer models in a single setup. As there is no single scoring scheme that can give an estimate of final performance, different metrics were used, each giving an indication of performance for each individual task. More detailed descriptions about the exhibited metrics can be found in Multimedia Appendices 1 and 2.

Evaluation Setups

IR evaluation was performed to consider various passages and should give a statement about the model's ability to find several suitable passages for a given query. The RAG evaluation results should indicate whether the embedding model also improved the response quality in an RAG system and, therefore, benefits context retrieval in clinical RAG scenarios. Together, these 2 setups provide a comprehensive view of the fine-tuned models' performance compared with non–fine-tuned models. Across all evaluation scenarios, we assessed whether fine-tuning led to measurable improvements on each specific task.

For each evaluation scenario, the models were evaluated on the German or English dataset. Thus, the datasets for each evaluation scenario were also translated from German into English. For the evaluation, the fine-tuned model is referred to as *miracle*, the model fine-tuned on pseudonymized data is referred to as *miracle-pseudonymized*, and the model fine-tuned on pseudonymized and English data is referred to as *miracle-pseudonymized-translated*.

Evaluated Models

For comparison of fine-tuned to existing models, 3 publicly available models were evaluated. The first model was the *multilingual-e5-large* sentence transformer [32], which was used as a base model for further fine-tuning. This model was continually trained on a mixture of multilingual datasets and supports up to 100 languages, including German. The second model was *bge-m5* [33], a state-of-the-art embedding model that is able to process inputs of sizes, spanning from short sentences to long documents of up to 8192 tokens. It was also trained on multilingual datasets. The third evaluated model was *gte-multilingual*-base [34] with support for more than 70 languages text lengths up to 8192 tokens. The fourth evaluated model was *German-RAG-BGE-M3*, which is a fine-tuned variant of the *bge-m3* model explicitly trained on German data for document retrieval. All evaluated models, including parameter sizes and links to the model cards on the huggingface hub, are listed in Multimedia Appendix 3.

IR Evaluation Setup

The IR evaluation was set up by identifying duplicate questions in the test dataset. These duplicate questions are of particular interest as they allow for 1 question to have multiple document passages that have a positive relation to this question. This can be used for an IR task such as search engines, where typically 1 query has multiple possible search results. In total, 109,795 unique questions and 173,933 document chunks from 52,434 documents and 26,358 patients were used for IR evaluation. On average, each question has approximately 5 document chunks with a positive relation.

The quality of the returned results or passages can be measured by metrics such as *Accuracy@K*, Precision@*K*, and *Recall@K* [35]. For IR systems such as search engines, the problem can be defined as a 2-fold task involving the definition of a search query and multiple results for this query. Using a cutoff threshold K and calculating an evaluation metric within the elements of this threshold can determine the quality of the results. More details about IR metrics such as mean reciprocal rank, normalized discounted cumulative gain (NDCG), and mean average precision (mAP) are provided in Multimedia Appendix 4.

For the IR evaluation, *Precision@K*, *Recall@K*, *MRR@K*, *NDCG@K*, and *mP@K* were calculated by retrieving top-k most similar documents for the given query in the dataset. The retrieved list was then compared to the ground truth. Similar document passages were retrieved using the cosine similarity between a query and top-k passages. For the evaluation, the results of the metrics are displayed for a K value of 10. For mAP, the results for a K value of 100 are shown.

RAG Evaluation Setup

In order to evaluate the benefit of our fine-tuned clinical embedding model in an IR setting, an experiment was conducted by setting up RAG systems. In Figure 2, the RAG evaluation setup is visualized.

XSL•FO
RenderX

Figure 2. Visualization of Retrieval Augmented Generation (RAG) evaluation setup. For the RAG evaluation, different databases containing embeddings from the different models were initialized. Different RAG systems using these databases as retrieval options were set up, and the same questions were asked of these systems. The generated answers from the RAG systems were compared against the ground truth answer. This graphic was created in BioRender. LLM: large language model.

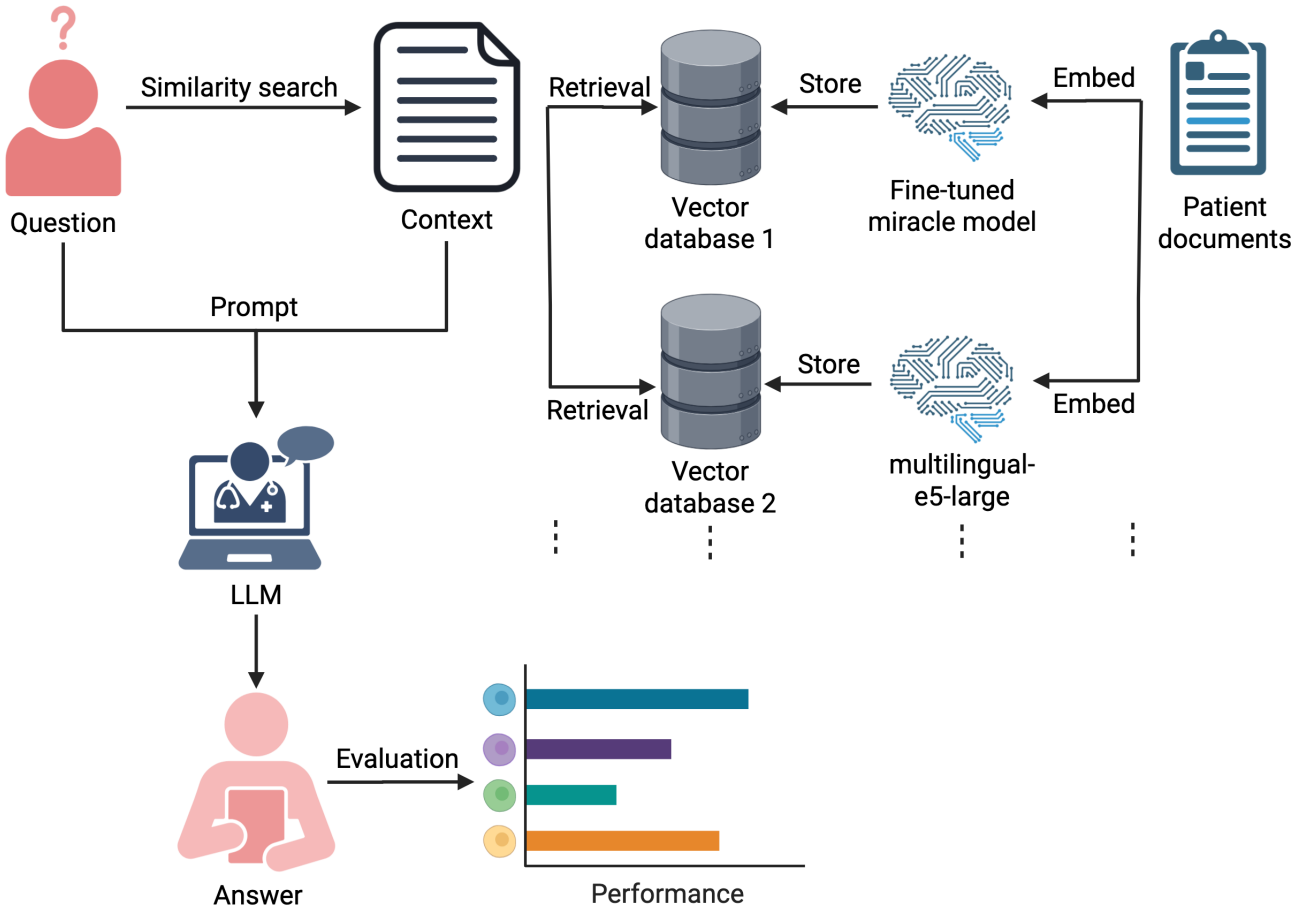

This system received user questions, attempted to find relevant document passages, and provided these passages to an LLM which was prompted to answer these questions. Each question in this dataset was presented to the RAG system and evaluated against the ground truth answer. Different metrics were used to compare the output from LLMs to a ground truth such as ROUGE or BLEURT. Additionally, retrieval-specific metrics such as Contextual Precision, Contextual Recall, and Contextual Relevancy were calculated using the *deepeval* Python package (version 3.7.5). These metrics are based on the LLM-as-a-judge [36] method, using an LLM to assess the quality of the retrieved context and measuring whether the context is relevant to the asked question. For the RAG evaluation, the *gpp-oss-120b* model [37] was deployed and used as the LLM-as-a-judge model. Selected for its high parameter count (120B) and strong performance on reasoning benchmarks, this model serves as an automated evaluator for semantic coherence and relevance, offering a reproducible open-source alternative to proprietary application programming interface–based judges. Detailed definition and explanation of all exhibited metrics are provided in Multimedia Appendix 5.

The fine-tuned embedding model was integrated into the retrieval component of this system, and the results were compared with those of the non–fine-tuned model. Given that the RAG system was identical for both the fine-tuned and non–fine-tuned embedding models, except for the embedding model itself, it was assumed that the greater the number of questions answered correctly, the more accurately the clinical context was identified by the embedding model. For the LLM component of the RAG system, the *SauerkrautLM-SOLAR-Instruct* model was used to answer the provided questions. Since this model was also used to generate the training data for the fine-tuned models, there is a possibility of bias when integrated into the RAG system. Therefore, the *Qwen3-235B-A22B-Instruct-2507-FP8* model was additionally used for generating answers. The maximum number of chunks retrieved for each question was set to 3. Cosine similarity was used as a distance metric for comparing questions and chunks in the similarity search. The smaller this distance, the more similar the question and chunk were considered to be. For easier assessment of the evaluation, Multimedia Appendix 6 shows a prompt example that was used to evaluate answers generated from the LLM using the RAG technique.

The RAG evaluation setup was divided into 2 parts. For the first part, the documents that the RAG system can retrieve were not restricted. This means that for 1 question, document parts can be retrieved from several patients in order to answer the question. While this indicates performance for IR settings, this constellation would rarely occur in real-world scenarios. Therefore, an additional filter mechanism was introduced for the second part to ensure that only documents from 1 patient can be retrieved for each question. Both configurations are

XSL•FO
RenderX

referred to as cross-patient (first part) and patient-centered evaluation (second part) in the following sections.

Cross-Patient Setup

To prepare a dataset for the cross-patient RAG evaluation, 372 clinical documents from 1 patient each were split into 1129 document chunks. These patients were not included in the training or test dataset that was used to fine-tune the sentence transformer model. The SauerkrautLM-SOLAR-Instruct model was then asked to generate questions and corresponding answers based on the contextual information provided by the chunks. As a result, 5556 synthetic question-answer pairs were generated. These question-answer pairs were filtered by a physician to provide a realistic clinical scenario and human-assisted evaluation. The physician was instructed to remove any questions related to administrative data (eg, questions about the patient's birth date or names of relatives) or remove questions that had wrong corresponding answers. It was emphasized that only those questions that would be posed in a practical clinical setting and that were factually and semantically accurate should be included. After filtering, 3536 annotated question-answer pairs from 280 documents split into 890 chunks remained for evaluation. These 890 document chunks were embedded by the fine-tuned and non–fine-tuned sentence transformer model, and the embeddings were stored in separate databases.

Patient-Centered Setup

The validated dataset from the cross-patient evaluation was used for the patient-centered RAG evaluation. As with the cross-patient setup, all documents included in this evaluation were sourced exclusively from patients who were held out from the training dataset. In total, the dataset consisted of 487 documents from 1 patient each and was divided into 2123 chunks. For each chunk, questions and corresponding answers were generated using the SauerkrautLM-SOLAR-Instruct model. As a result, 10,615 question-answer pairs were generated.

These additional questions and answers with the corresponding document chunks were annotated by 1 physician and 2 medical students. The students represented a range of academic backgrounds, including one student who has passed the first state examination for medicine and is currently in her seventh semester, and another who has studied and completed molecular biology and is currently in her second semester of pharmacy studies. Similar to the cross-patient data filtering described in the previous section, the annotators were instructed to remove any nonmedical questions and to include only those questions that would be asked in a realistic clinical setting. To gain insight into the quality of the annotations by measuring the agreement between the annotators, the annotators were divided into 2 groups. The first group consisted of 2 physicians, including 1 physician who had already annotated the cross-patient dataset and whose previously annotated data were reused for this comparison. The 2 medical students formed the second group. Each group received the same questions and answers. Interannotator agreement (IAA) was calculated based on precision, recall, and F_1-score between the 2 annotators of 1 group. One annotator was designated as the ground truth, while the other served as the estimator. The annotated questions and answers were used as ground truth for patient-centered RAG evaluation. For each question, a filter was introduced, ensuring that only document chunks from a single patient were retrieved.

Results

Training Data Quality Assessment

To evaluate the reliability of the synthetically generated training corpus, a manual audit was conducted on a statistically random sample of 200 question-answer pairs, including corresponding document chunks, from the training dataset. This audit was performed by a physician to quantify the prevalence of generated artifacts. The analysis revealed that 36 pairs (18.0%) contained hallucinations, defined as questions for which the answer could not be derived from the provided document chunk. Additionally, 12 (6.0%) pairs contained factual errors or medically invalid reasoning. Consequently, 76.0% of the sample was verified as semantically accurate and clinically correct.

IR Evaluation

The IR evaluation metrics indicated a stronger performance of the fine-tuned embedding model, with a mAP@100 score improved from 0.14 to 0.27. The model fine-tuned on pseudonymized data also improved after training, achieving a mAP@100 of 0.25. The NDCG@10 metric increased after fine-tuning from 0.16 to 0.29 for the *miracle* model and to 0.27 for the pseudonymized model. Besides the fine-tuned models, the best-performing model for the German dataset was *multilingual-e5-large*. The *bge-m3* model scored a mAP@100 of 0.11 and a NDCG@10 of 0.12. The *gte-multilingual-base* model achieved slightly higher scores with a mAP@100 of 0.12 and a NDCG@10 of 0.13. The *german-bge-m3* model showed higher performances than the base *bge-m3* model, indicating that fine-tuning on German datasets improved language-specific capabilities. On the translated dataset, the *miracle* model also achieved the highest scores with improvements for mAP@100 from 0.12 to 0.27 and 0.25 for the pseudonymized model. The *bge-m3* model showed lower scores with a mAP@100 of 0.11. The *gte-multilingual-base* model scored a mAP@100 of 0.12. The results on the German dataset are visualized in Figure 3. The complete results for different k values and languages are appended in Multimedia Appendices 7 and 8.

XSL•FO
RenderX

Figure 3. Comparative illustration of the information retrieval evaluation results for the German dataset. The results are presented for a k value of 10 for MRR, recall, precision, and NDCG. Abbreviations: mAP: mean average precision; MRR: mean reciprocal rank; NDCG: normalized discounted cumulative gain.

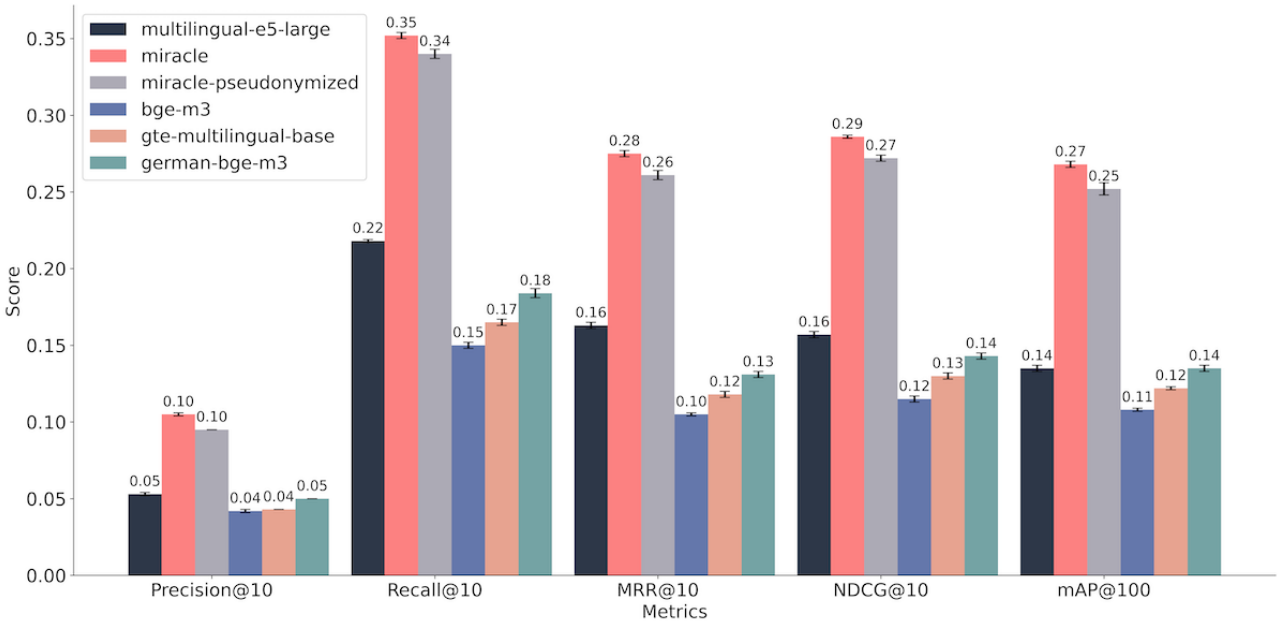

RAG Evaluation

Cross-Patient Results

The overall answer quality of the RAG system improved after fine-tuning the sentence transformer model. The BERTScore-F1 went up from 0.76 to 0.77. Improvements were also observed in both the ROUGE-2 score, which increased from 0.18 to 0.30, and the BLEURT score, which rose from 0.53 to 0.56. Other models such as *bge-m3* and *gte-multilingual-base* achieved lower scores in total, with a BERTScore-F1 score of 0.74 and a ROUGE-2 score of 0.17 for *bge-m3* and 0.16 for *gte-multilingual-base*. The fine-tuned variant of *bge-m3* (*german-bge-m3*) again showed better scores than the base model with a BERTScore-F1 of 0.75 and a ROUGE-2 score of 0.18. The highest scores for Contextual Precision and Contextual Recall were achieved by the pseudonymized model (0.82 and 0.89).

The results are in line with the answers generated by Qwen-3 (Multimedia Appendix 9). Because of the higher capabilities of this model, scores that are dependent on the generative part, such as BERTScore or ROUGE, are higher for every evaluated model. However, relative differences between the models stay intact, with the fine-tuned *miracle* model reaching the highest scores for most metrics (BERTScore F1: 0.78, ROUGE-2: 0.26) and showing improvements compared with the *multilingual-e5-large* (BERTScore F1: 0.76, ROUGE-2: 0.23).

For the translated dataset, the pseudonymized model achieved the highest scores with a BERTScore-F1 of 0.75 and a BLEURT score of 0.50. Of the non–fine-tuned models, the *bge-m3* model achieved the best scores for the English dataset with a BERTScore-F1 of 0.75 and a BLEURT score of 0.48. The RAG evaluation results for the German dataset are shown in Table 2. The results for the English dataset are provided in Multimedia Appendix 10.

XSL•FO
RenderX

Table . Retrieval Augmented Generation reevaluation results for the German dataset[a].

Metric	multilingual-e5-large	miracle	miracle pseudonymized	bge-m3	gte-multilingual-base	german-bge-m3
BERTScore P[b]	0.733	*0.747* [d]	0.736	0.725	0.725	0.728
BERTScore R[c]	0.783	*0.796* [d]	0.792	0.760	0.756	0.779
BERTScore F1	0.756	*0.769* [d]	0.762	0.742	0.739	0.747
BLEURT	0.533	*0.563* [d]	0.551	0.542	0.539	0.553
ROUGE-1	0.327	*0.351* [d]	0.342	0.319	0.311	0.327
ROUGE-2	0.179	*0.197* [d]	0.194	0.167	0.159	0.175
ROUGE-L	0.276	*0.297* [d]	0.292	0.269	0.258	0.274
Contextual P	0.685	0.779	*0.823* [d]	0.626	0.567	0.676
Contextual R	0.774	0.860	*0.892* [d]	0.726	0.676	0.783
Contextual Relevancy	0.307	*0.312* [d]	0.309	0.274	0.268	0.310

[a]For the Retrieval Augmented Generation (RAG) evaluation, generated answers by the RAG system are compared with ground truth answers that were created by a large language model and filtered by a medical doctor. The higher the scores are, the more similar the answers are considered to the ground truth. The answers were generated by the *SauerkrautLM-Solar-Instruct* model.

[b]P: precision.

[c]R: recall.

[d]The values in italics highlight the highest scores in the respective category.

Patient-Centered Results

For the patient-centered RAG evaluation, the LLM-generated questions and answers (pairs) were filtered by 2 annotation groups. In Table 3, the results of the filtering process are displayed.

Table . Human filtering results of Retrieval Augmented Generation evaluation dataset[a].

Annotator	Annotator group	Education	Pairs checked	Pairs filtered out, n (%)	IAA[b]
Annotator 1	Group 1	Physician	5556	2020 (36.36)	Precision=0.854; recall=0.963; F_1-score=0.905
Annotator 2	Group 1	Physician	7081	1909 (26.96)	
Annotator 3	Group 2	Student (2nd semester)	3556	1813 (50.98)	Precision=0.763; recall=0.796; F_1-score=0.779
Annotator 4	Group 2	Student (7th semester)	3839	1844 (48.03)	

[a]The Retrieval Augmented Generation evaluation dataset, which was partly created by a large language model, was given to 2 physicians and 2 medical students with the instruction to filter out questions or answer pairs that were medically unrelated. The dataset was divided into 2 parts, and each part was provided to 1 annotator group, resulting in 2 annotator groups. In the table, results of the human annotator process are displayed. The agreement between the annotators was calculated based on the overlapping annotated pairs of each annotator group (5556 for annotator group 1 and 3556 for annotator group 2).

[b]IAA: interannotator agreement.

Comparisons between annotator groups 1 and 2 show that there are differences in the number and percentage of filtered out question and answer pairs. While annotators 1 and 2 filtered out 36.36% and 26.96% of the dataset, annotators 3 and 4 filtered out 50.98% and 48.03%. IAA for group 1 was higher with an F_1-score of 0.905 than for group 2, who reached an F_1-score of 0.779.

The annotated question and answer pairs were used for the patient-centric RAG evaluation. For this evaluation scenario, the fine-tuned models again showed higher scores than the multilingual-e5-large, bge-m3, *german-bge-m3*, and gte-multilingual-base models. Compared with the results of the RAG evaluation without patient context, the differences between models have become smaller. The best-performing model, the model fine-tuned on pseudonymized data, achieved a

XSL•FO
RenderX

BERTScore-F1 of 0.78, a BLEURT score of 0.61, and a ROUGE-1 score of 0.42. In comparison, the multilingual-e5-large reached a BERTScore-F1 of 0.78, a BLEURT score of 0.61, and a ROUGE-1 score of 0.41. Similar to the cross-patient evaluation, the highest scores for Contextual Precision and Contextual Recall were reached by the model trained on pseudonymized data. Interestingly, the highest score for Contextual Relevancy was achieved by the *german-bge-m3* model. The evaluation performed using the Qwen-3 as the generator showed similar results (Multimedia Appendix 11), with improvements for the fine-tuned models in all metrics compared with the *multilingual-e5-large* model. In Table 4, the RAG evaluation results for the patient-centric context are presented.

Table . Retrieval Augmented Generation patient-context evaluation results[a].

Metric	multilingual-e5-large	miracle	miracle pseudonymized	bge-m3	gte-multilingual-base	german-bge-m3
BERTScore P[b]	0.763	*0.766* [c]	*0.766* [c]	0.764	0.761	0.756
BERTScore R[d]	0.795	*0.799* [c]	*0.799* [c]	0.796	0.793	0.793
BERTScore F1	0.778	*0.781* [c]	*0.781* [c]	0.778	0.775	0.773
BLEURT	0.608	*0.611* [c]	*0.611* [c]	0.608	0.604	0.593
ROUGE-1	0.414	0.420	*0.421* [c]	0.415	0.407	0.394
ROUGE-2	0.249	*0.256* [c]	*0.256* [c]	0.250	0.242	0.232
ROUGE-L	0.355	0.361	*0.362* [c]	0.355	0.349	0.335
Contextual P	0.868	0.918	*0.927* [c]	0.858	0.824	0.880
Contextual R	0.927	0.938	*0.939* [c]	0.925	0.909	0.929
Contextual Relevancy	0.239	0.241	0.241	0.235	0.233	*0.242* [c]

[a]For the Retrieval Augmented Generation (RAG) evaluation, generated answers by the RAG system are compared with ground truth answers that were created by a large language model and filtered by 4 human annotators. The higher the scores are, the more similar the answers are considered to the ground truth. The answers were generated by the *SauerkrautLM-Solar-Instruct* model.

[b]P: precision.

[c]The values in italics highlight the highest scores in the respective category.

[d]R: recall.

For more nuanced performance differences between the models, the patient-centric performances were broken down by document category (Multimedia Appendix 12). For discharge letters, radiology reports, and operation notes, the fine-tuned models show the highest scores, with only slight differences between the models. The results on pathology reports indicate that there is a stronger difference in performance between models, with the *gte-multilingual-base* achieving the highest BERTScore-F1 and ROUGE scores.

Discussion

Principal Findings

This study developed and tested language-specific clinical embedding models specifically trained on verified hospital documents for medical RAG applications and demonstrated their higher accuracy metrics. The introduced approach and models address an important gap in NLP applications for health care by improving clinical language comprehension for multiple languages and by optimizing RAG systems within medical information environments. Without domain fine-tuning, it is challenging for embedding models to comprehend clinical language and contexts, as they are not trained on real-world medical documents. The integration of these models as retrieval components in RAG systems may lead to undesired or medically irrelevant results. Fine-tuning embedding models on medical real-world data represent an advancement in the field of medical NLP and improve artificial intelligence–driven health care by facilitating the understanding of clinical language.

Previous research focused on English language models for NLP and has proven its application in various clinical settings [38,39]. This strategy, however, has faced shortcomings in performance, in particular, in the multilanguage setting. Our models have been fine-tuned on German medical data, a language not frequently used in medical NLP [20], and whose documents are currently less abundant. The presented work not only complements but also diversifies the linguistic scope of medical NLP research and its clinical applicability. The extracted dataset from 4 medical application areas had highly nuanced and domain-specific medical language. By using an LLM to generate question and answer pairs for these documents, an extensive dataset with approximately 11 million pairs served as a foundation for fine-tuning and evaluation. Different evaluation benchmarks introduced in this study provide insights into improvements achieved after fine-tuning the models on this dataset.

XSL•FO
RenderX

Importantly, these fine-tuned context and language-sensitive miracle models surpassed the state-of-the-art *multilingual-e5-large*, *bge-m3*, *german-bge-m3*, and *gte-multilingual-base* models in 2 evaluation scenarios. The model trained on real-world data showed the best overall results for the German evaluation datasets. More interestingly, the model fine-tuned on pseudonymized data achieved similar results on the IR and even better results for the RAG evaluation on the English-translated dataset.

One hypothesis is that the pseudonymization process may have resulted in an improved machine readability of certain phrases, which in turn may have positively influenced the quality of the translation. However, the models fine-tuned on translated data overall showed lower scores than the models trained on real-world data. Noise and possibly difficult translations of abbreviations might have influenced the dataset quality.

Performance comparisons, particularly with the *german-bge-m3* model, showed that language-specific fine-tuning generally can improve retrieval scores. However, the improvements achieved by fine-tuning on real-world clinical data indicate that further enhancements can be achieved by domain adaptation.

The results of the patient-centered RAG evaluation show that the fine-tuned models are able to find suitable text passages more accurately in a patient question answering setting. The category-wise results indicate differences in performance depending on the document type. As documents from different clinics and departments are written in different styles with varying language and form, the retrieval of suitable text passages is highly dependent on the ability of the models to semantically understand and interpret medical language. The highest scores were achieved for discharge letters and radiology reports, indicating simpler language compared with pathology reports and surgical reports, which may contain more specific and nuanced language.

By having physicians and medical students annotate the evaluation dataset for the patient-centered evaluation, cautious estimates can be made about the quality of the dataset. The human annotation results highlight both the limitations and effectiveness of synthetic data generation in medical contexts. While 27% - 51% of LLM-generated pairs required filtering, the remaining validated content still provided sufficient evaluation data. The variation in filtering rates between annotators likely reflects differences in clinical experience and criteria for data quality. Despite these quality variations, the consistent performance improvements observed across all evaluation scenarios suggest that the signal-to-noise ratio in the synthetic training data was sufficient for effective model fine-tuning. These findings indicate that while expert validation improves data quality, synthetic generation can still yield functional training corpora for specialized domains such as clinical NLP.

Serving as the foundation for our research, the provided dataset consisting of real-world clinical texts and LLM-generated question-answer pairs offers a comprehensive training resource that captures the nuances of medical terminology and clinical reasoning. Building upon this dataset, the presented technologies indicate a potential to transform how health care professionals access and use medical information. An example of how this technology could be integrated into clinical routine is visualized in Multimedia Appendix 13.

Regarding operational implementation in hospital IT infrastructures, the deployment of the fine-tuned embedding model presents a low barrier to entry. Unlike the training phase or the generative components of RAG systems, which often require GPU acceleration, the retrieval component based on dense vector embeddings is computationally efficient during inference. The model can be deployed on standard, CPU-based server environments commonly found in hospital data centers. This efficiency allows for seamless integration into existing clinical workflows and decision support systems without imposing additional infrastructure costs or latency, making the solution highly scalable for real-time IR across large patient cohorts.

Limitations

While our models indicate a step forward for the development of clinical embedding models for RAG systems, they are not without limitations. The dataset, though comprehensive, is limited to 4 types of clinical documents and may benefit from expansion by including a greater variety of medical texts. Furthermore, as the documents were extracted from 1 site only, the model is susceptible to template overfitting. Clinical documents from a single institution often share rigid structural templates and specific formatting cues, creating a risk that the model learns to associate relevance with these institutional artifacts rather than purely semantic content. Consequently, the performance observed here may not fully transfer to institutions with different documentation standards without local fine-tuning. Additionally, the English language model was trained on machine-translated data using the WMT'19 German-to-English translation model, with translation quality validated on only 30 physician-reviewed samples. This sample size is statistically insufficient to ensure translation accuracy across the entire corpus of 400,000 clinical documents. Machine translation of specialized medical terminology, abbreviations, and context-dependent clinical language may introduce systematic translation artifacts that differ from native English clinical documentation. The models fine-tuned on translated data should therefore be considered a proof of concept for cross-lingual transfer learning rather than a validated clinical tool ready for deployment in English-speaking health care settings.

Furthermore, the current evaluation compared performance exclusively against multilingual baselines to align with the model's architectural foundation. We did not benchmark against embedding models trained exclusively on English clinical data. Consequently, while our results demonstrate the model's efficacy within a multilingual framework, its relative performance against specialized, monolingual English clinical models remains to be validated in future studies.

A circular evaluation bias may be inherent to the experimental design, as the test questions were synthetically generated using the same LLM family used for training. This shared origin implies that the evaluation might partially measure the model's ability to recognize synthetic linguistic patterns rather than organic human queries. To mitigate this, the

XSL•FO
RenderX

Qwen3-235B-A22B-Instruct-2507-FP8 was additionally used as the generator part of the RAG system. The results showed overall higher scores but no significant relative performance differences between the evaluated embedding models.

Furthermore, the IR evaluation was conducted in a cross-patient setting where synthetically generated questions could match document chunks from multiple patients, with all matches treated as relevant. A limitation of this methodology is that it relies on identifying duplicate questions, which inherently introduces a selection bias toward generic queries (eg, "What is the diagnosis?") that are likely to appear multiple times across the corpus. Consequently, complex, specific, or nuanced questions that are unique to a single patient case were systematically excluded. This bias may result in an overestimation of performance on simple lookup tasks while underrepresenting the challenges of complex clinical reasoning. Additionally, treating all matches as relevant introduces potential label noise, as generic questions may retrieve contextually appropriate chunks from different patients that answer different clinical questions. However, this evaluation design does capture legitimate clinical use cases, including searching for similar patient presentations, researching treatment outcomes across cases, or conducting retrospective cohort analyses. To address patient-specific scenarios, we conducted a separate patient-centered RAG evaluation with retrieval restricted to single patients, which demonstrated the models' effectiveness in realistic clinical contexts.

The observed low absolute IR performance metrics must be interpreted in the context of the extensive search space, which comprised 173,933 document chunks across 26,358 patients. This high-dimensional setting renders specific retrieval statistically challenging compared with constrained, single-patient scenarios. The evaluation methodology relied on identifying exact duplicate synthetic questions to define ground truth. This rigid definition of relevance likely penalized semantically redundant chunks that contained relevant information but generated slightly different synthetic questions. Such instances are treated as irrelevant in our evaluation, resulting in potentially high false-negative rates. Consequently, strict metrics such as Accuracy@1 may underestimate the model's true clinical usefulness, as they measure the model's ability to retrieve the exact source chunk rather than its ability to retrieve any chunk containing the correct semantic answer. The RAG evaluation does not directly measure the IR but the answer quality, which also could be biased by the LLM component of the RAG system. However, the combination of different evaluation scenarios allows for a more balanced and diverse look at the performance of the models.

In the patient-centered RAG evaluation, the fine-tuned miracle model consistently achieved the highest numerical scores across all metrics. However, we explicitly acknowledge that the performance differences compared with the multilingual-e5-large baseline were marginal. A formal statistical significance test was not performed for these specific evaluation metrics. Consequently, the statistical significance of these performance differences has not been established. It remains possible that the observed marginal improvements fall within the margin of error and may result from random variance rather than a systematic model advantage. This stands in contrast to the larger performance deltas observed in the cross-patient settings. We hypothesize that this convergence may be attributable to a potential saturation effect caused by the constrained search space. In the patient-centered setting, the retrieval pool is restricted to a single patient's history (typically comprising fewer than 30 document chunks). With such a limited number of distractor chunks, the baseline model already achieves near-ceiling retrieval performance, potentially masking the benefits of domain-specific fine-tuning that become more apparent in larger, high-dimensional retrieval spaces.

We further acknowledge a discrepancy in the IAA observed during the dataset validation. While the physician group achieved a high F_1-score of 0.905, the medical student group achieved a lower score of 0.779. This variance is potentially attributable to differences in clinical experience, as students may lack the consolidated clinical intuition required to consistently apply exclusion criteria to borderline cases. As the student-annotated portion of the dataset was used as ground truth without a secondary confirmation review by a senior clinician, this introduces a potential for label noise. Consequently, the evaluation metrics for the patient-centered scenario should be interpreted with this underlying annotator variability in mind.

A notable finding in our RAG evaluation was the consistently low Contextual Relevancy scores (approximately 0.24) observed across all models. This indicates that roughly three-quarters of the content within a retrieved chunk was classified as nonrelevant to the specific query by the judge model. We attribute this to the high information density characteristic of clinical documentation. Even with a constrained text segment of 450 characters, clinical notes frequently contain multiple distinct, packed statements. Consequently, while the chunk contains the correct answer, it inevitably includes competing clinical facts that lower the relevancy ratio. This highlights a persistent challenge in medical IR: finding an optimal granularity that isolates specific findings without destroying the narrative context.

A further limitation concerns the text segmentation strategy. We used a fixed chunk size of 450 characters with an 80-character overlap. While this granularity enhances the precision of retrieval for specific clinical facts, it risks context fragmentation, particularly for complex temporal reasoning or negation scopes that span across multiple segments (eg, a diagnosis mentioned at the beginning of a letter and a treatment plan at the end). Furthermore, an alternative atomic-level chunking strategy (eg, sentence-level splitting) could theoretically maximize relevancy scores for closed-ended questions (eg, "What is the patient's sodium level?"). However, we deliberately avoided this approach to prioritize semantic coherence. Atomic splitting risks fragmenting the narrative flow essential for open-ended clinical reasoning (eg, linking a diagnosis to its ruling out criteria in the next sentence). Therefore, the reported low relevancy scores reflect a trade-off where we accepted lower precision in signal-to-noise ratio to preserve the integrity of the local clinical context.

XSL•FO
RenderX

LLMs, including the *SauerkrautLM-SOLAR-Instruct* model used for data generation, are susceptible to hallucinations and may generate factually incorrect or clinically inappropriate content [40]. This introduces potential noise into the training dataset that could propagate to the fine-tuned embedding models. Analysis of a random sample indicated that approximately 24% of the generated question-answer pairs contained hallucinations or factual inaccuracies. This introduces a safety concern, as this specific type of noise has the potential to propagate into the embedding space, effectively teaching the model incorrect semantic associations. While our downstream performance suggests that the valid signal was dominant, we explicitly warn that the model was trained on a corpus with a known error rate. Therefore, the resulting embeddings should not be treated as ground truth. Any application of these models in clinical practice must mandate robust downstream filtering and expert human verification to identify and intercept potential retrieval errors stemming from these training artifacts. For evaluation, the question-answer pairs were also initially generated by an LLM. However, a rigorous human-in-the-loop validation process involving physicians and medical students was conducted to mitigate the risks associated with purely synthetic benchmarks. This resulted in a filtering rate of approximately 27%-51%, ensuring that the final test set represents a human-verified standard of clinical relevance.

A further limitation concerns the evaluation scope regarding negative testing. While the RAG system prompt included instructions to refuse answering if the context was insufficient (eg, stating "'I don't know"), our quantitative evaluation focused exclusively on positive retrieval scenarios, specifically cases where the answer was present in the ground truth. We did not systematically assess the model's ability to correctly reject irrelevant context or quantify the rate of hallucinations when no suitable information is retrieved. Validating this refusal capability is an important safety requirement for clinical deployment. Without it, there is a risk that the system might confidently hallucinate answers based on retrieved but irrelevant distractor chunks.

Relatedly, the quantitative assessment of the RAG system relied on automated semantic similarity metrics such as BERTScore, BLEURT, and ROUGE. While these metrics are standard benchmarks for measuring the linguistic overlap between generated answers and ground truth, they are not designed to verify factual medical precision. High semantic scores can occasionally mask critical factual errors, such as laterality swaps (eg, "left" vs "right") or negation flips (eg, "no evidence of" vs "evidence of"), which may have high lexical overlap with the correct answer.

Future Work

Future validation on external datasets from different sites and the inclusion of additional document categories would strengthen the approach and indicate broader applicability. To facilitate this necessary next step, we have publicly released the models fine-tuned on pseudonymized and translated data. This allows the broader research community to validate the miracle models on their local datasets, thereby establishing their generalizability across different clinical environments in future work.

While in the study, quantitative metrics were assessed for comparison of embedding models, evaluation with clinicians is an important next step to evaluate clinical performance for real-world scenarios. Evaluation with end users could give insights into the integration of the fine-tuned models in a real-world setting and should be investigated in future studies. Future clinical validation studies must use expert human verification to audit generated responses for specific types of high-risk factual hallucinations. Future studies should also include dedicated patient-specific IR benchmarks with queries explicitly designed for individual patient contexts. Additionally, it remains to be validated whether the domain-specific fine-tuning makes the model more robust in extracting correct answers or potentially more prone to hallucinating responses due to an increased bias toward using retrieved clinical text.

Comparisons between the fine-tuned models and domain-adapted embedding models could strengthen the approach and give insights into the quality of the QA dataset. However, to our knowledge, there are no publicly available fine-tuned sentence transformer models trained on German medical real-world data.

The retrieval parameter k (number of retrieved chunks) was fixed at 3 for all RAG experiments. We acknowledge that this parameter was not empirically optimized through ablation studies within this work. While the parameter value was selected to provide a reasonable balance between context sufficiency and noise reduction, it is possible that the choice of k influenced the reported performance metrics. Determining the optimal retrieval depth or using dynamic threshold-based retrieval strategies requires further investigation.

Although the recursive splitting method respects paragraph boundaries to maintain local semantic integrity, future studies should explore adaptive segmentation strategies. While section-aware chunking or hierarchical retrieval can better capture long-range dependencies, future research should also investigate hybrid segmentation pipelines. Such approaches could dynamically use atomic-level chunking for high-density, structured data (eg, laboratory values and medication lists) to maximize precision, while retaining larger, section-aware windows for unstructured narrative prose (eg, anamnesis and discharge summary) to preserve semantic flow. Investigating the optimal combination of these strategies remains an important step toward production-grade clinical RAG systems.

Furthermore, while this study evaluated performance using dense vector retrieval, clinical decision-making often requires reasoning across complex, interrelated entities (eg, drug-drug interactions or temporal disease progressions). Future research should explore integrating these fine-tuned embedding models into Graph Retrieval Augmented Generation systems. In such architectures, our domain-adapted embeddings could serve to enhance node resolution and relationship traversal within clinical knowledge graphs, potentially addressing the limitations of purely vector-based retrieval for multihop reasoning tasks.

Finally, our evaluation focused specifically on the performance of dense vector retrieval. We did not benchmark against lexical approaches such as BM25 or multistage pipelines involving cross-encoder reranking or multivector architectures such as

XSL•FO
RenderX

ColBERT. While lexical search can be effective for exact keyword matching, it often fails to capture the semantic nuances of medical terminology. However, we acknowledge that in a production-grade clinical RAG system, the fine-tuned miracle embeddings presented here would likely serve as the retrieval foundation within a hybrid pipeline, potentially supplemented by keyword search and followed by a reranking step to maximize precision. Further research should also explore section-aware chunking or hierarchical retrieval approaches to better capture long-range dependencies in clinical narratives.

Conclusions

This study demonstrates that embedding models fine-tuned on real-world clinical documents can outperform existing state-of-the-art multilingual models for medical IR and RAG applications. The approach of generating synthetic question-answer pairs from clinical documents using LLMs provides a scalable methodology for creating domain-specific training data while maintaining patient privacy through pseudonymization. The models' effectiveness across both German and English datasets indicates applicability beyond single-language contexts. This work establishes a reproducible framework for developing specialized embedding models that address the gap between general-purpose language models and the specific requirements of health care documentation, particularly for non-English medical settings where such resources are limited.

Acknowledgments

The work of BE and MB was funded by a PhD grant from the DFG Research Training Group 2535 Knowledge- and data-based personalization of medicine at the point of care (WisPerMed). A special thanks goes to the Annotation Lab for the annotation effort. Figures 1 and 2 were created with Biorender.com. The authors acknowledge the use of generative artificial intelligence during the early stages of manuscript drafting to assist in structuring initial ideas and refining wording. DeepL and Claude were used in later stages to enhance language clarity and improve readability. All content generated or suggested by these tools was thoroughly reviewed, revised, and edited by the authors, and the authors take full responsibility for the accuracy and integrity of the final manuscript. The data for this project were provided by the Smart Hospital Information Platform (SHIP), managed by the Data Integration Center at the University Medicine Essen. SHIP serves as a comprehensive digital health platform for integrating data from all major clinical subsystems using a holistic Fast Healthcare Interoperability Resources (FHIR)–based approach. It enables the purification, analysis, distribution, and visualization of clinical data.

Funding

The authors declared no financial support was received for this work.

Data Availability

The original data that support the findings of this study are sensitive clinical data and hence not openly available due to reasons of data protection and privacy. Requests regarding the dataset can be sent to datagovernance@uk-essen.de and will be reviewed. The final models trained on pseudonymized data created as part of this study are publicly accessible [41]. The underlying code for this study is made publicly available upon publication [42] and includes all scripts developed during the study.

Authors' Contributions

KA contributed to writing – original draft, training, visualization, validation, methodology, data curation, conceptualization. HS and AIY contributed to writing – review and editing, validation, methodology, and formal analysis. CSS contributed to writing – review and editing, formal analysis, and clinical perspective. BE and MB participated in writing – review and editing, validation, and data curation. AT participated in writing – review and editing, supervision, and conceptualization. OBP, EMH, PW, KB, and JH contributed to writing – review and editing, data curation, and formal analysis. FN contributed to writing – review and editing, supervision, and project administration. RH contributed to writing – review and editing, supervision, project administration, and conceptualization.

Conflicts of Interest

AT has stated a conflict. Research funding by Neovii Biotech and GTL. Consultancy for CSL Behring, Maat Pharma, Biomarin, Pfizer, and Onkowissen. Travel reimbursements by Neovii and Novartis. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Additional information about cosine similarity.
[DOCX File, 14 KB - jmir_v28i1e82997_app1.docx]

Multimedia Appendix 2
Additional information about precision, recall, and F_1-score.
[DOCX File, 13 KB - jmir_v28i1e82997_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Parameter size of evaluated models.
[DOCX File, 14 KB - jmir_v28i1e82997_app3.docx]

Multimedia Appendix 4
Additional information about information retrieval metrics.
[DOCX File, 15 KB - jmir_v28i1e82997_app4.docx]

Multimedia Appendix 5
Additional detailed information about Retrieval Augmented Generation metrics.
[DOCX File, 15 KB - jmir_v28i1e82997_app5.docx]

Multimedia Appendix 6
Prompt example for evaluation of Retrieval Augmented Generation systems.
[DOCX File, 15 KB - jmir_v28i1e82997_app6.docx]

Multimedia Appendix 7
Information retrieval evaluation results on the German dataset.
[DOCX File, 16 KB - jmir_v28i1e82997_app7.docx]

Multimedia Appendix 8
Information retrieval evaluation results on English dataset.
[DOCX File, 16 KB - jmir_v28i1e82997_app8.docx]

Multimedia Appendix 9
Retrieval Augmented Generation evaluation results on the German dataset in a cross-patient setting with answers generated by Qwen-3.
[DOCX File, 15 KB - jmir_v28i1e82997_app9.docx]

Multimedia Appendix 10
Retrieval Augmented Generation evaluation results on English dataset in cross-patient setting.
[DOCX File, 15 KB - jmir_v28i1e82997_app10.docx]

Multimedia Appendix 11
Retrieval Augmented Generation evaluation results on the German dataset in a patient-centered setting with answers generated by Qwen-3.
[DOCX File, 16 KB - jmir_v28i1e82997_app11.docx]

Multimedia Appendix 12
Retrieval Augmented Generation patient-centered evaluation results per document category.
[DOCX File, 2497 KB - jmir_v28i1e82997_app12.docx]

Multimedia Appendix 13
Mock-up implementation of clinical Retrieval Augmented Generation system with integration into electronic health record for patient question answering.
[DOCX File, 800 KB - jmir_v28i1e82997_app13.docx]

References

1. Weissler EH, Zhang J, Lippmann S, Rusincovitch S, Henao R, Jones WS. Use of natural language processing to improve identification of patients with peripheral artery disease. Circ Cardiovasc Interv 2020 Oct;13(10):e009447. [doi: 10.1161/CIRCINTERVENTIONS.120.009447] [Medline: 33040585]
2. Joseph JW, Leventhal EL, Grossestreuer AV, et al. Deep-learning approaches to identify critically Ill patients at emergency department triage using limited information. J Am Coll Emerg Physicians Open 2020 Oct;1(5):773-781. [doi: 10.1002/emp2.12218] [Medline: 33145518]

XSL•FO
RenderX

3. Berge GT, Granmo OC, Tveit TO, Munkvold BE, Ruthjersen AL, Sharma J. Machine learning-driven clinical decision support system for concept-based searching: a field trial in a Norwegian hospital. BMC Med Inform Decis Mak 2023 Jan 10;23(1):5. [doi: 10.1186/s12911-023-02101-x] [Medline: 36627624]
4. Shi J, Gao X, Kinsman WC, Ha C, Gao GG, Chen Y. DI++: a deep learning system for patient condition identification in clinical notes. Artif Intell Med 2022 Jan;123:102224. [doi: 10.1016/j.artmed.2021.102224] [Medline: 34998515]
5. Chen TL, Emerling M, Chaudhari GR, et al. Domain specific word embeddings for natural language processing in radiology. J Biomed Inform 2021 Jan;113:103665. [doi: 10.1016/j.jbi.2020.103665] [Medline: 33333323]
6. Wang S, Tseng B, Hernandez-Boussard T. Development and evaluation of novel ophthalmology domain-specific neural word embeddings to predict visual prognosis. Int J Med Inform 2021 Jun;150:104464. [doi: 10.1016/j.ijmedinf.2021.104464] [Medline: 33892445]
7. Shtar G, Greenstein-Messica A, Mazuz E, Rokach L, Shapira B. Predicting drug characteristics using biomedical text embedding. BMC Bioinformatics 2022 Dec 7;23(1):526. [doi: 10.1186/s12859-022-05083-1] [Medline: 36476573]
8. Goswami K, Dutta S, Assem H, Fransen T, McCrae JP. Cross-lingual sentence embedding using multi-task learning. Presented at: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; Nov 7-11, 2021. [doi: 10.18653/v1/2021.emnlp-main.716]
9. Jiang X, Liang Y, Chen W, Duan N. XLM-K: improving cross-lingual language model pre-training with multilingual knowledge. AAAI 2022;36(10):10840-10848. [doi: 10.1609/aaai.v36i10.21330]
10. Kim Y, Kim JH, Lee JM, et al. A pre-trained BERT for Korean medical natural language processing. Sci Rep 2022;12(1):13847. [doi: 10.1038/s41598-022-17806-8]
11. Gemmiti M, Hamed S, Wildhaber J, Pharisa C, Klumb PL. Physicians' speech complexity and interrupting behavior in pediatric consultations. Health Commun 2022 May;37(6):748-759. [doi: 10.1080/10410236.2020.1868063] [Medline: 33441020]
12. Watermeyer J, Thwala Z, Beukes J. Medical terminology in intercultural health interactions. Health Commun 2021 Aug;36(9):1115-1124. [doi: 10.1080/10410236.2020.1735700] [Medline: 32202159]
13. Pitt MB, Hendrickson MA. Eradicating jargon-oblivion-a proposed classification system of medical jargon. J Gen Intern Med 2020 Jun;35(6):1861-1864. [doi: 10.1007/s11606-019-05526-1] [Medline: 31713041]
14. Hao T, Pan X, Gu Z, Qu Y, Weng H. A pattern learning-based method for temporal expression extraction and normalization from multi-lingual heterogeneous clinical texts. BMC Med Inform Decis Mak 2018 Mar 22;18(Suppl 1):22. [doi: 10.1186/s12911-018-0595-9] [Medline: 29589563]
15. Yang LWY, Ng WY, Lei X, et al. Development and testing of a multi-lingual natural language processing-based deep learning system in 10 languages for COVID-19 pandemic crisis: a multi-center study. Front Public Health 2023;11. [doi: 10.3389/fpubh.2023.1063466]
16. Sivarajkumar S, Wang Y. HealthPrompt: a zero-shot learning paradigm for clinical natural language processing. AMIA Annu Symp Proc 2022;2022:972-981. [Medline: 37128372]
17. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
18. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology. Ophthalmol Sci 2023 Dec;3(4):100324. [doi: 10.1016/j.xops.2023.100324] [Medline: 37334036]
19. Frei J, Frei-Stuber L, Kramer F. GERNERMED++: semantic annotation in German medical NLP through transfer-learning, translation and word alignment. J Biomed Inform 2023 Nov;147:104513. [doi: 10.1016/j.jbi.2023.104513] [Medline: 37838290]
20. Frei J, Kramer F. Annotated dataset creation through large language models for non-English medical NLP. J Biomed Inform 2023 Sep;145:104478. [doi: 10.1016/j.jbi.2023.104478] [Medline: 37625508]
21. Parola A, Lin JM, Simonsen A, et al. Speech disturbances in schizophrenia: assessing cross-linguistic generalizability of NLP automated measures of coherence. Schizophr Res 2023 Sep;259:59-70. [doi: 10.1016/j.schres.2022.07.002] [Medline: 35927097]
22. Toro S, Anagnostopoulos AV, Bello SM, et al. Dynamic retrieval augmented generation of ontologies using artificial intelligence (DRAGON-AI). J Biomed Semant 2024;15(1). [doi: 10.1186/s13326-024-00320-3]
23. Li Y, Zhao J, Li M, et al. RefAI: a GPT-powered retrieval-augmented generative tool for biomedical literature recommendation and summarization. J Am Med Inform Assoc 2024 Sep 1;31(9):2030-2039. [doi: 10.1093/jamia/ocae129] [Medline: 38857454]
24. Arzideh K, Baldini G, Winnekens P, et al. A transformer-based pipeline for German clinical document de-identification. Appl Clin Inform 2025 Jan;16(1):31-43. [doi: 10.1055/a-2424-1989] [Medline: 39778706]
25. Ng N, Yee K, Baevski A, Ott M, Auli M, Edunov S. Facebook FAIR's WMT19 news translation task submission. Presented at: Proceedings of the Fourth Conference on Machine Translation (Volume 2); Aug 1-2, 2019; Florence, Italy p. 314-319. [doi: 10.18653/v1/W19-5333]
26. Fan A, et al. Beyond english-centric multilingual machine translation. arXiv. Preprint posted online on Oct 21, 2020. [doi: 10.48550/ARXIV.2010.11125]

XSL•FO
RenderX

27. NLLB Team. No language left behind: scaling human-centered machine translation. arXiv. Preprint posted online on Jul 11, 2022. [doi: 10.48550/ARXIV.2207.04672]
28. Raffel C, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv. Preprint posted online on Oct 23, 2019. [doi: 10.48550/ARXIV.1910.10683]
29. Tang Y, et al. Multilingual translation with extensible multilingual pretraining and finetuning. arXiv. Preprint posted online on Aug 2, 2020. [doi: 10.48550/ARXIV.2008.00401]
30. Tiedemann J, Thottingal S, et al. OPUS-MT—building open translation services for the world. In: Martins A, editor. : European Association for Machine Translation; 2020 Nov 1 Presented at: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation p. 479-480 URL: https://aclanthology.org/2020.eamt-1.61/
31. Thakur N, Reimers N, Daxenberger J, Gurevych IAS. Augmented SBERT: data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. Presented at: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics; Jun 6-11, 2021. [doi: 10.18653/v1/2021.naacl-main.28]
32. Wang L, et al. Text embeddings by weakly-supervised contrastive pre-training. arXiv. Preprint posted online on Feb 22, 2024 URL: http://arxiv.org/abs/2212.03533 [accessed 2026-03-20]
33. Chen J, Xiao S, Zhang P, Luo K, Lian D, Liu Z. M3-embedding: multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. arXiv. Preprint posted online on 2024. [doi: 10.18653/v1/2024.findings-acl.137]
34. Zhang X, Zhang Y, Long D, et al. MGTE: generalized long-context text representation and reranking models for multilingual text retrieval. arXiv. Preprint posted online on 2024. [doi: 10.18653/v1/2024.emnlp-industry.103]
35. Radlinski F, Craswell N. Comparing the sensitivity of information retrieval metrics. 2010 Jul 19 Presented at: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval; Jul 19-23, 2010. [doi: 10.1145/1835449.1835560]
36. Baysan MS, Uysal S, İşlek İ, Çığ Karaman Ç, Güngör T. LLM-as-a-Judge: automated evaluation of search query parsing using large language models. Front Big Data 2025;8:1611389. [doi: 10.3389/fdata.2025.1611389] [Medline: 40761620]
37. OpenAI. Gpt-oss-120b & gpt-oss-20b model card. arXiv. Preprint posted online on Aug 8, 2025. [doi: 10.48550/arXiv.2508.10925]
38. Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. J Am Med Inform Assoc 2019 Nov 1;26(11):1297-1304. [doi: 10.1093/jamia/ocz096] [Medline: 31265066]
39. Noh J, Kavuluru R. Improved biomedical word embeddings in the transformer era. J Biomed Inform 2021 Aug;120:103867. [doi: 10.1016/j.jbi.2021.103867] [Medline: 34284119]
40. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature New Biol 2024 Jun;630(8017):625-630. [doi: 10.1038/s41586-024-07421-0] [Medline: 38898292]
41. Miracle embedding model trained on pseudonymized data available on huggingface. URL: https://huggingface.co/SHIPAI-IKIM/miracle-german [accessed 2026-03-09]
42. Github repository for the code developed during the course of the study. URL: https://github.com/UMEssen/MIRACLE [accessed 2026-03-09]

Abbreviations

IAA: interannotator agreement
IR: information retrieval
LLM: large language model
MAP: mean average precision
NDCG: normalized discounted cumulative gain
NLP: natural language processing
RAG: Retrieval Augmented Generation

Edited by A Coristine; submitted 26.Aug.2025; peer-reviewed by Q Zhang, Z Lu; accepted 29.Jan.2026; published 25.Mar.2026.

Please cite as:
Arzideh K, Schäfer H, Idrissi-Yaghir A, Schmidt CS, Eryilmaz B, Bahn M, Turki AT, Pollok OB, Hartmann EM, Winnekens P, Borys K, Haubold J, Nensa F, Hosch R
Improving Retrieval Augmented Generation for Health Care by Fine-Tuning Clinical Embedding Models: Development and Evaluation Study
J Med Internet Res 2026;28:e82997
URL: https://www.jmir.org/2026/1/e82997
doi:10.2196/82997

XSL•FO
RenderX

© Kamyar Arzideh, Henning Schäfer, Ahmad Idrissi-Yaghir, Cynthia Sabrina Schmidt, Bahadir Eryilmaz, Mikel Bahn, Amin Turki, Olivia Barbara Pollok, Eva Maria Hartmann, Philipp Winnekens, Katarzyna Borys, Johannes Haubold, Felix Nensa, René Hosch. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Artificial Intelligence for Predicting Treatment Response in Patients With Anxiety Disorders After Cognitive Behavioral Therapy: Systematic Review and Meta-Analysis

Jiawen Liu[1,2], MA; Junhui Wang[3], MEng; Zhaobin Wu[3], MEng; Mohamad Ibrani Shahrimin Bin Adam Assim[2], PhD

[1]Youth League Committee, Liuzhou Railway Vocational Technical College, No. 2 Wenyuan Road, Yufeng District, Liuzhou, Guangxi Zhuang Autonomous Region, China

[2]Faculty of Humanities, Management and Science, Universiti Putra Malaysia, Bintulu, Sarawak, Malaysia

[3]School of Automation, Guangxi University of Science and Technology, Liuzhou, Guangxi, China

Corresponding Author:
Jiawen Liu, MA
Youth League Committee, Liuzhou Railway Vocational Technical College, No. 2 Wenyuan Road, Yufeng District, Liuzhou, Guangxi Zhuang Autonomous Region, China

Abstract

Background: Artificial intelligence (AI) models have been increasingly explored for predicting treatment response to cognitive behavioral therapy (CBT) in patients with anxiety disorders. Identifying potential responders in advance may help inform treatment planning and support clinical decision-making. Although a growing number of studies have applied AI techniques in this context, reported performance estimates vary across studies, and the overall predictive accuracy has not been comprehensively quantified.

Objective: This systematic review and meta-analysis aims to quantify the overall performance of AI models in predicting treatment response following CBT for anxiety disorders and to examine how data sources, algorithmic approaches, and diagnostic subtypes influence predictive performance.

Methods: A systematic literature search was conducted in PubMed, Embase, Web of Science, Cochrane Library, and PsycINFO up to August 2025. We included studies that validated AI models for predicting CBT treatment response (remission or response) in patients diagnosed with an anxiety disorder. The risk of bias was assessed using the PROBAST+AI (Prediction Model Risk of Bias Assessment Tool for Artificial Intelligence) tool. Predictive performance metrics, including sensitivity, specificity, accuracy, and area under the curve (AUC), were extracted and pooled. Pooled estimates for sensitivity, specificity, and diagnostic accuracy were derived using the Restricted Maximum Likelihood estimator, with CIs adjusted via the Hartung-Knapp-Sidik-Jonkman method. Prediction intervals were calculated and reported alongside these pooled estimates to illustrate the expected distribution of effects in real-world settings.

Results: Eleven studies were included in the meta-analysis. The pooled sensitivity of AI-based models for predicting treatment response was 0.73 (95% CI 0.58 - 0.85; I^2=82.8%), and the pooled specificity was 0.75 (95% CI 0.59 - 0.89; I^2=96.7%). The overall pooled accuracy was 0.74 (95% CI 0.62 - 0.84; I^2=94.6%). The summary AUC was 0.81 (95% CI 0.78 - 0.85), indicating moderate discriminative performance. Subgroup analyses showed that models incorporating multimodal data achieved superior predictive performance, with a pooled sensitivity of 0.84 and an accuracy of 0.82. In addition, predictive performance was the highest in patients with social anxiety disorder compared with other anxiety disorder subtypes.

Conclusions: This meta-analysis quantitatively synthesized AI performance in predicting CBT response for anxiety disorders, moving beyond narrative reviews to provide pooled evidence. In contrast to existing reviews that encompass broader diagnostic groups, our focused approach establishes a precise benchmark for this clinical domain, highlighting the current moderate overall performance. Furthermore, we extend beyond previous work by demonstrating the superior predictive utility of multimodal data, identifying social anxiety disorder as the most predictable subtype, and systematically evaluating the impact of data modalities and algorithm types. Future efforts should prioritize robustly validated multimodal models, laying essential groundwork for the potential development of AI-assisted tools to personalize treatment planning in anxiety disorders.

Trial Registration: PROSPERO CRD420251137096; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251137096

(J Med Internet Res 2026;28:e86079) doi:10.2196/86079

KEYWORDS

artificial intelligence; treatment response; anxiety disorders; cognitive behavioral therapy; meta-analysis

XSL•FO
RenderX

Introduction

Anxiety disorders constitute a group of mental disorders characterized by intense, excessive, and persistent worry and fear, exhibiting high prevalence rates. They not only cause significant functional impairment in patients' social, academic, and occupational functioning but also impose substantial socioeconomic burdens at the societal level, including considerable health care resource consumption and productivity losses [1,2]. Cognitive behavioral therapy (CBT) is a first-line evidence-based intervention for anxiety disorders [3,4]. However, treatment response varies considerably, with a notable subset of patients showing suboptimal or inadequate improvement [5]. This heterogeneity underscores a critical clinical need: the ability to preemptively identify likely responders and, more importantly, those at high risk of nonresponse who may benefit from treatment augmentation or alternative interventions from the outset [6]. Early and accurate prediction of treatment response can optimize care pathways, improve resource allocation, and reduce delays, aligning with the goals of precision mental health [7].

Clinical assessment of anxiety disorders typically relies on structured or semistructured interviews and standardized scales [8], such as the Hamilton Anxiety Rating Scale (HAM-A) [9], Clinical Global Impressions–Improvement (CGI-I) [10], and Liebowitz Social Anxiety Scale (LSAS) [7], which are fundamental for symptom identification, outcome evaluation, and longitudinal monitoring. However, they possess inherent limitations for predictive purposes. First, they depend heavily on self-report and subjective ratings, making them vulnerable to recall bias and social desirability effects, thereby compromising objectivity [11]. Second, heterogeneity in measured constructs, scoring procedures, and cutoff thresholds across instruments constrains comparability and diagnostic consistency across studies and clinical settings [12]. Consequently, there is growing recognition that these traditional metrics alone may be insufficient for reliably forecasting an individual's unique response trajectory to a specific therapy such as CBT [13].

Against this background, artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), has attracted increasing attention as a methodological framework for advancing outcome prediction in mental health research. AI-based models can integrate large-scale, high-dimensional, and multimodal data and capture complex, nonlinear relationships that are difficult to model with conventional statistical techniques [14]. In recent years, a growing number of studies have applied ML and DL methods to predict treatment response across a range of psychiatric interventions, including CBT for anxiety disorders [15,16]. These studies have incorporated diverse data sources, including baseline clinical characteristics, symptom trajectories, neuroimaging markers, psychophysiological measures, and digital behavioral indicators, and have reported varying levels of predictive accuracy [17,18]. However, the anxiety disorders comprise clinically distinct subtypes, such as social anxiety disorder (SAD), generalized anxiety disorder (GAD), and panic disorder (PD), which differ in symptom profiles and underlying mechanisms [19]. This diagnostic heterogeneity may further influence model generalizability and predictive stability across populations [7,20]. To date, however, quantitative evidence synthesizing the predictive performance of AI models for CBT response in different anxiety disorder subtypes remains limited, and potential sources of heterogeneity across studies remain insufficiently explored and require systematic investigation [21].

This systematic review and meta-analysis aims to quantitatively synthesize the overall performance of AI models in predicting treatment response following CBT for anxiety disorders. We also seek to explore the influence of key factors, such as data sources, algorithmic approaches, and diagnostic subtypes, on predictive performance and to preliminarily investigate potential sources of the observed heterogeneity.

Methods

Ethical Considerations

As this is a systematic review and meta-analysis, ethics approval and consent to participate are not applicable. The manuscript does not include the identification images or other personal or clinical details of participants.

The meta-analysis adhered rigorously to the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy) reporting guidelines [22] and the PRISMA-S extension for reporting literature searches [23]. We have provided the completed PRISMA-DTA (Checklist 1), PRISMA 2020 Abstract (Checklist 2), and PRISMA-S (Checklist 3) checklists. The research protocol was registered in the PROSPERO registry (registration ID: CRD420251137096).

Feasibility Assessment for Meta-Analysis

Prior to initiating the full systematic review, a preliminary scoping search was conducted to assess the feasibility of quantitative synthesis. This step aimed to determine whether sufficient studies reported the extractable predictive performance metrics necessary for a meta-analysis. In the scoping search, although the field is nascent, several published studies met the core data requirements, supporting the rationale for proceeding with a formal systematic review and meta-analysis.

Search Strategy

A systematic search was conducted in PubMed, Embase, Web of Science, Cochrane Library, and PsycINFO from inception to August 16, 2025. Two independent reviewers (JL and JW) screened titles/abstracts and subsequently assessed full texts, with disagreements resolved by consensus. Search strategies combined controlled vocabulary (MeSH) and free-text terms across 3 concept blocks: AI and methods (eg, "artificial intelligence," "machine learning," and "deep learning"), anxiety disorders (eg, "anxiety disorders," "social anxiety disorder," "generalized anxiety disorder," and "panic disorder"), and CBT (eg, "cognitive behavioral therapy," "CBT," and "cognitive therapy"). No language or publication year limits were applied at the initial search stage to minimize selection bias. Detailed database-specific strategies are provided in Table S1 in Multimedia Appendix 1. Study selection followed a dual-stage

XSL•FO
RenderX

independent screening process. Two reviewers (JL and JW) first screened titles/abstracts and then independently assessed full texts of potentially eligible records. Disagreements were resolved by consensus or by third-party arbitration (ZW). The reference lists of included studies were manually screened for additional records. It should be noted that in conducting this systematic review, we did not contact authors, experts, or manufacturers to obtain additional studies or data, and did not adapt or reuse search strategies from previous reviews. Additionally, a formal peer review of the search strategy was not conducted.

Inclusion and Exclusion Criteria

The PITROS framework was used for the inclusion criteria:

- Participants (P): patients with anxiety disorders receiving CBT, including separation anxiety disorder, selective mutism, specific phobia, SAD, PD/panic attacks, agoraphobia, GAD, and substance/medication-induced anxiety disorder
- Index test (I): AI methods (eg, ML and DL) developed to predict CBT treatment response
- Target condition (T): positive cases defined as clinically significant post-CBT improvement (eg, meaningful symptom reduction); negative cases defined as poor or no response
- Reference standard (R): validated diagnostic/severity scales for anxiety disorders
- Outcomes (O): sensitivity, specificity, area under the curve (AUC), and accuracy
- Setting (S): mental health outpatient clinics, psychiatric hospitals, and clinical trial centers; multicenter prospective or retrospective designs using electronic health records or public databases

The exclusion criteria were as follows: (1) publication in a language other than English at the full-text stage, (2) nonempirical article type (eg, review, case report, conference abstract, meta-analysis, and letter), (3) no assessment of the prediction of CBT response using AI, (4) inclusion of a population that did not have anxiety disorder or that did not undergo CBT, (5) absence of a recognized reference standard or inadequate data to derive the target metrics, and (6) duplicate publication identified using EndNote and manual checking. The database search was unrestricted by language to maximize sensitivity, but non-English publications were excluded during full-text review owing to translation resource constraints. We acknowledge this as a potential limitation, as it may have introduced language bias by excluding relevant studies published in other languages. The independent selection of studies was conducted by 2 reviewers (JL and JW), with disagreements resolved through discussion or by a third reviewer (ZW).

Data Extraction

Two independent reviewers (JL and JW) conducted comprehensive data extraction from full-text articles to assess study eligibility. The extracted data encompassed critical study characteristics, including the authors, year, country, study design, center, response definition, types of disorders, and types of treatments. Specific quantitative elements included the total number of patients in the training and internal validation sets, the number of response patients in the training and internal validation sets, the data-splitting method, predictors, AI algorithms, and AI methods. For internal validation sets, we systematically recorded true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values.

For studies included in the systematic review but lacking meta-analysis–compatible data, we proactively contacted corresponding authors via email to request necessary information. Given the frequent absence of diagnostic contingency tables, we reconstructed 2×2 tables using reported sensitivity, specificity, responder numbers, and total patient counts. Any discrepancies in extraction were resolved through a collaborative discussion, with a third reviewer (ZW) serving as an arbitrator if consensus could not be reached, thereby ensuring methodological rigor and data integrity.

Quality Assessment

The latest PROBAST+AI (Prediction Model Risk of Bias Assessment Tool for Artificial Intelligence) quality assessment tool was used [24], which replaced PROBAST 2019. This tool comprises 2 phases: model development and model evaluation. Each phase encompasses 7 domains covering participant and data sources, predictors, outcome assessment, and analysis. Evaluation results for each domain are categorized as “low,” “high,” or “unclear,” determined based on specific signal questions. Signal concerns are rated as “yes,” “probably yes,” “probably no,” “no,” “no information,” and, in some cases, “not applicable.” Signal concerns phrased as “yes” or “probably yes” indicate a lower risk of bias. Any signal question rated “no” or “probably no” indicates a potential high risk of bias in that domain. If no “no” or “probably no” rating exists but “no information” is present, the domain’s bias risk is classified as unclear. The complete list of signal questions and tables is detailed in Tables S2 and S3 in Multimedia Appendix 1.

To ensure objectivity and accuracy in the assessment process, 2 reviewers (JL and JW) independently assessed the risk of bias in the included studies using the PROBAST+AI quality assessment tool. During the review process, any discrepancies between reviewers were resolved through in-depth discussions and analyses to achieve consensus, ensuring highly reliable, consistent final assessment results.

Outcome Measures

The primary outcome measures encompassed comprehensive diagnostic performance metrics from internal validation sets, including sensitivity, specificity, accuracy, and AUC. Sensitivity, calculated as TP/(TP+FN), measured the AI model’s ability to accurately identify TPs, while specificity, computed as TN/(TN+FP), assessed the model’s proficiency in correctly identifying negative cases. AUC provided a holistic assessment of the model’s discriminative power between positive and negative instances. Accuracy, defined as (TP+TN)/total patients, represented the proportion of correctly classified samples. We systematically extracted performance metrics from all data sources (eg, imaging, clinical/demographic, and multimodal) and AI algorithms presented in each study. However, for the purpose of meta-analytic pooling, only the primary model from

XSL•FO
RenderX

each article, specifically the model based on nonoverlapping patient cohorts, was selected to ensure data independence and avoid double-counting.

Statistical Analysis

While the core analytical plan and reporting adhered to the preregistered protocol, a key methodological deviation was implemented to enhance the robustness of our statistical synthesis. The registered protocol specified the use of the bivariate random-effects model (Reitsma model) as the primary method for pooling sensitivity and specificity. In the final analysis, we maintained this model for generating the summary receiver operating characteristic (SROC) curve and the AUC. However, for the pooled estimates of sensitivity and specificity themselves, we adopted a univariate random-effects model using the Restricted Maximum Likelihood estimator, with CIs adjusted via the Hartung-Knapp-Sidik-Jonkman (HKSJ) method. This adaptation was made to incorporate the HKSJ adjustment, which provides more conservative and reliable CIs, particularly when dealing with a limited number of studies or substantial heterogeneity [25]. Prediction intervals (PIs) were calculated and reported alongside these pooled estimates to illustrate the expected distribution of effects in real-world settings. Heterogeneity was quantitatively assessed using the PIs and further explored through subgroup analyses and meta-regression [26]. Bivariate boxplots were used to visually identify potential outlier studies contributing to heterogeneity. Prespecified covariates for meta-regression included the number of study centers, anxiety disorder subtype, data splitting method, CBT protocol type, and AI algorithm category. Subgroup analyses were conducted according to data type, algorithm type, and specific anxiety disorder diagnoses. Additionally, a bubble plot was used to assess temporal trends in the accuracy of different algorithms, and funnel plots were used, along with Egger tests, to evaluate potential small-study effects.

Results

Study Selection

A search across 5 electronic databases yielded 1399 potentially relevant publications. After deleting 399 duplicates, 1000 records were retained for preliminary screening. At this phase, 980 studies were identified as clearly irrelevant based on the title/abstract and the publication type. Thus, 15 articles were selected for full-text review. Following a comprehensive assessment of the full texts, 5 studies were removed because they did not investigate treatment response to CBT. Two studies were excluded due to non-English publication. Three studies were excluded as they were ongoing trials without any published outcome measures. Additionally, 6 qualifying studies were identified from nondatabase sources. Ultimately, 11 studies [7,17,18,27-34] were included in the meta-analysis as they met predefined inclusion criteria. The identification, screening, eligibility assessment, and inclusion process followed the PRISMA guidelines, and the selection analysis is elaborated in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) s in Figure 1: flow diagram of the systematic literature search and study selection process for studies evaluating artificial intelligence models in predicting treatment response to cognitive behavioral therapy for anxiety disorders.

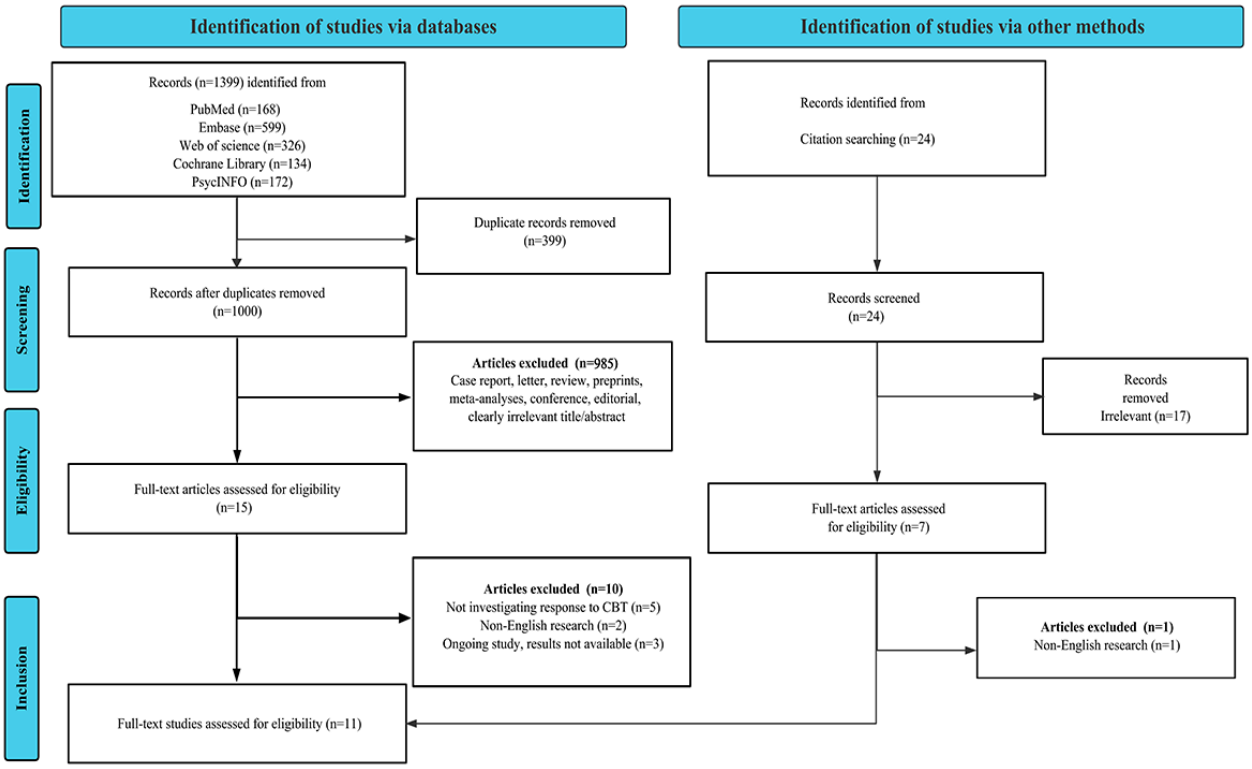

XSL•FO
RenderX

Study Description and Quality Assessment

A total of 11 eligible studies comprising 59,085 samples (training: 38,563; internal validation set: 20,522) were included. Among these, 4 studies exclusively evaluated SAD [7,29,32,34], 2 studies focused on PD with agoraphobia [30,33], 1 study investigated both GAD and PD [27], 1 study examined PD and SAD [31], and the remaining 3 studies addressed multiple anxiety disorder subtypes [17,18,28]. Nine studies [17,18,27,29-34] provided data eligible for meta-analysis. The included studies were published between 2014 and 2024, and all were prospective. Definitions of treatment response varied across studies: 2 studies used the CGI-I [29,32], 2 used the LSAS [7,34], 2 used the HAM-A [30,33], 1 used the Overall Anxiety Severity and Impairment Scale (OASIS) [27], 1 used the clinical severity rating [28], 1 combined HAM-A and the Spider Phobia Questionnaire (SPQ) [18], 1 used both the Generalized Anxiety Disorder-7 (GAD-7) and the Patient Health Questionnaire-9 (PHQ-9) [17], and 1 applied multiple scales simultaneously [31]. Across the 9 datasets, the most frequently used AI algorithms were support vector machines (SVMs; 4/9, 44%) and random forests (RFs; 3/9, 33%). Additionally, of the 9 datasets, 2 were based on clinical/demographic data, 6 on imaging data, and 1 on multimodal data. The clinical, methodological, and technical characteristics of the included studies are summarized in Table 1 and Multimedia Appendix 2.

XSL•FO
RenderX

Table . Characteristics of the included studies (n=11) and their patient populations.

Study (author)	Year	Country	Study design	Center	Response definition	Types of anxiety disorders	Types of treatments	Data source	Total sample size, n		Response sample size, n	
									Training	Internal validation	Training	Internal validation
Ball et al [27]	2014	America	Pro[a]	Single center	OASIS[b] ≤5	GAD[c], PD[d]	Standard CBT[e]	Institutional database	48	49	29	49
Bertie et al [28]	2024	Multiple countries	Pro	Multiple centers	CSR[f] ≥4	GAD, separation anxiety, SAD[g], SPH[h]	Standard CBT	Institutional database	2214	2214	—[i]	—
Bukhari et al [7]	2025	America	Pro	Multiple centers	LSAS[j]↓ ≥50%	SAD	Standard CBT	Institutional database	157	157	—	—
Frick et al [29]	2020	Sweden	Pro	Single center	CGI-I[k] ≤2	SAD	CBT+SSRI[l], CBT+placebo	Institutional database	47	47	24	24
Hahn et al [30]	2015	Germany	Pro	Multiple centers	HAM-A[m]↓ ≥50%	PD with AG[n]	Standard CBT	Institutional database	49	49	25	25
Hilbert et al [18]	2024	Germany	Pro	Multiple centers	Protect-AD: HAM-A↓ ≥50%; Spider VR: SPQ[o]↓ ≥30%	PD, AG, SAD, SPH	Standard CBT	Institutional database	410	410	213	213
Hentati Isacsson et al [31]	2024	Sweden	Pro	Single center	MADRS[p] ≤11, PDSS[q] ≤8, LSAS ≤35	PD, SAD	iCBT[r]	Institutional database	3619	3619	—	—
Månsson et al [32]	2015	Sweden	Pro	Multiple centers	CGI-I ≤2	SAD	iCBT	Institutional database	23	23	12	12
Prasad et al [17]	2023	United Kingdom, America	Pro	Multiple centers	GAD-7[s]↓ ≥64, PHQ-9[t]↑ <6	Anxiety	iCBT	Institutional database	31,899	13,857	—	5232
Sundermann et al [33]	2017	Germany	Pro	Multiple centers	HAM-A↓ ≥50%	PD with AG	Standard CBT	Institutional database	59	59	30	30
Whitfield-Gabrieli et al [34]	2015	America	Pro	Multiple centers	LSAS↓ ≥50%	SAD	Standard CBT	Institutional database	38	38	19	19

[a]Pro: prospective.

[b]OASIS: Overall Anxiety Severity and Impairment Scale.

[c]GAD: generalized anxiety disorder.

XSL•FO
RenderX

[d]PD: panic disorder.
[e]CBT: cognitive behavioral therapy.
[f]CSR: clinical severity rating.
[g]SAD: social anxiety disorder.
[h]SPH: specific phobia.
[i]Not applicable.
[j]LASA: Liebowitz Social Anxiety Scale.
[k]CGI-I: clinical global impression-improvement.
[l]SSRI: selective serotonin reuptake inhibitor.
[m]HAM-A: Hamilton Anxiety Rating Scale.
[n]AG: agoraphobia.
[o]SPQ: Spider Phobia Questionnaire.
[p]MADRS: Montgomery-Asberg Depression Rating Scale.
[q]PDSS: Panic Disorder Severity Scale.
[r]iCBT: internet-based cognitive behavioral therapy.
[s]GAD-7: Generalized Anxiety Disorder-7.
[t]PHQ-9: Patient Health Questionnaire-9.

The risk of bias, as assessed by the PROBAST+AI quality assessment tool, is presented in Figure 2 and Tables S2 and S3 in Multimedia Appendix 1. For model development, in the overall quality assessment, 55% (6/11) of studies were rated as high risk, while the remaining 45% (5/11) were rated as low risk. Regarding applicability concerns, 9% (1/11) of studies were rated as high risk, while the remaining 91% (10/11) were rated as low risk. For model validation, 55% (6/11) of studies were rated as high risk in the overall risk of bias assessment, with the remaining 45% (5/11) rated as low risk. In the comprehensive evaluation of applicability concerns, 9% (1/11) of studies were rated as high risk, while the remaining 91% (10/11) were rated as low risk.

XSL•FO
RenderX

Figure 2. Risk of bias and applicability assessment using the PROBAST+AI tool for studies developing or validating AI models to predict treatment response after cognitive behavioral therapy in patients with anxiety disorders. Panel A summarizes the risk of bias for model development studies, and panel B summarizes the risk of bias for model validation studies. PROBAST+AI: Prediction Model Risk of Bias Assessment Tool for Artificial Intelligence.

XSL•FO
RenderX

Trend in Diagnostic Accuracy Over Time

The scatter plot in Figure 3 illustrates a negative trend in model accuracy from 2014 to 2024 across logistic regression (LR), recurrent neural network (RNN), RF, and SVM algorithms. A linear regression line confirms that reported accuracy has generally decreased over the decade within this dataset. The highest performance was recorded in 2015 by an SVM model (approximately 0.9), while the lowest scores appeared in 2024, with RF models showing a drop in scores to around 0.5. The shaded region indicates the CI for this downward trajectory (Figure 3).

Figure 3. Bubble plot illustrating the accuracy of various artificial intelligence algorithms over time (publication year) for predicting treatment response to cognitive behavioral therapy in patients with anxiety disorders. LR: logistic regression; RF: random forest; RNN: recurrent neural network; SVM: support vector machine.

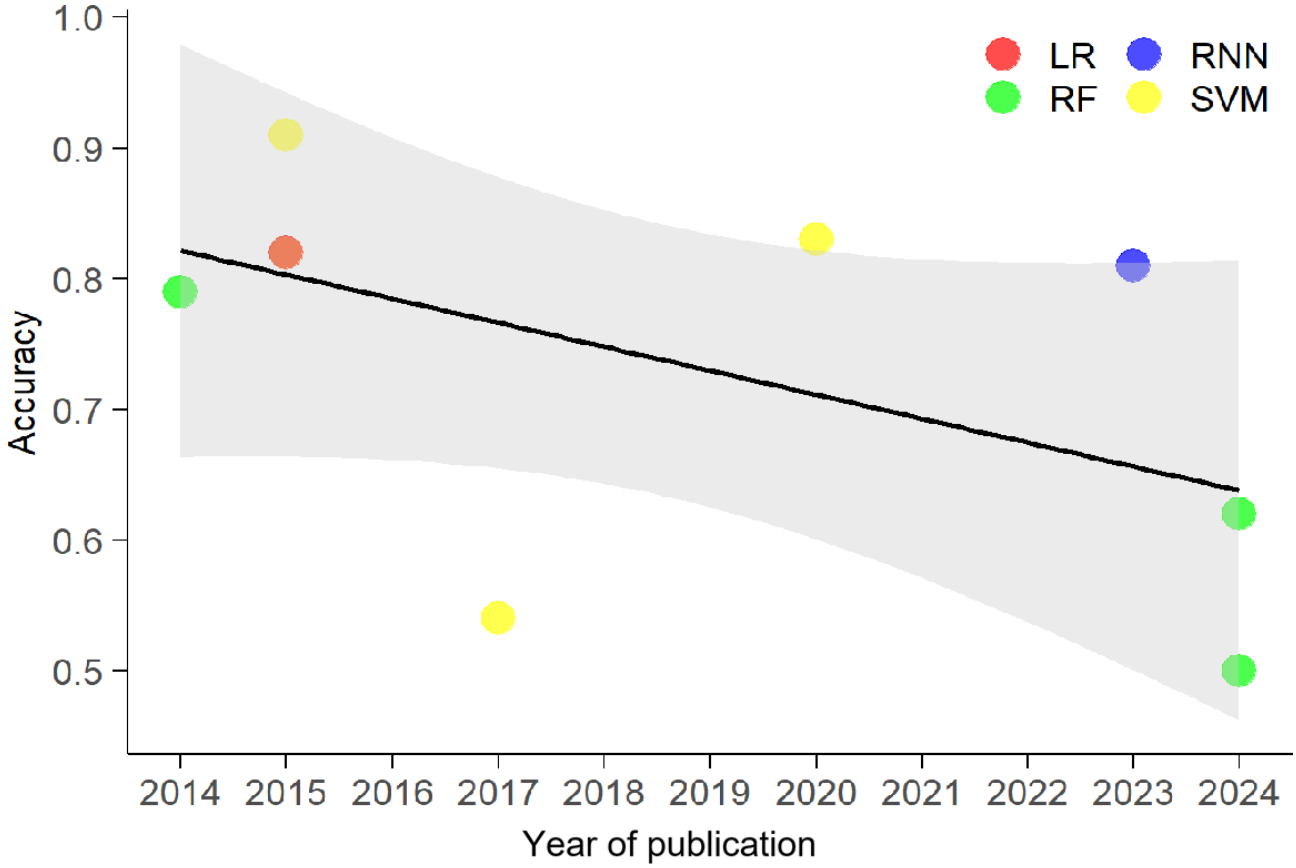

Predicting CBT Response to Anxiety Disorders Using AI Models Based on the Primary Model Per Study

The pooled sensitivity of AI models was 0.73 (95% CI 0.58 - 0.85; 95% PI 0.31 - 0.99; I^2=82.8%), specificity was 0.75 (95% CI 0.59 - 0.89; 95% PI 0.25 - 1.00; I^2=96.7%), accuracy was 0.74 (95% CI 0.62 - 0.84; 95% PI 0.39 - 0.97; I^2=94.6%), and AUC was 0.81 (95% CI 0.78 - 0.85), as shown in Figures 4-7. Based on a prespecified pretest probability of 20%, the Fagan nomogram indicated a posttest probability of 43% for a positive result and 8% for a negative result (Figure S1 in Multimedia Appendix 1). The results of the bivariate boxplots indicated that 1 study [30] had discrete values, suggesting it may be a source of heterogeneity (Figure 8). Furthermore, meta-regression analysis indicated that heterogeneity in sensitivity was significantly associated with the number of centers (P<.001) and the type of anxiety disorder (P<.001). In contrast, heterogeneity in specificity was significantly associated with CBT type (P<.001) and AI algorithm category (P<.001) (Table 2).

XSL•FO
RenderX

Figure 4. Forest plot displaying the pooled sensitivity, with 95% CIs and prediction intervals, of artificial intelligence models for predicting treatment response after cognitive behavioral therapy in patients with anxiety disorders [17,18,27,29,30,32-34]. FN: false negative; TP: true positive.

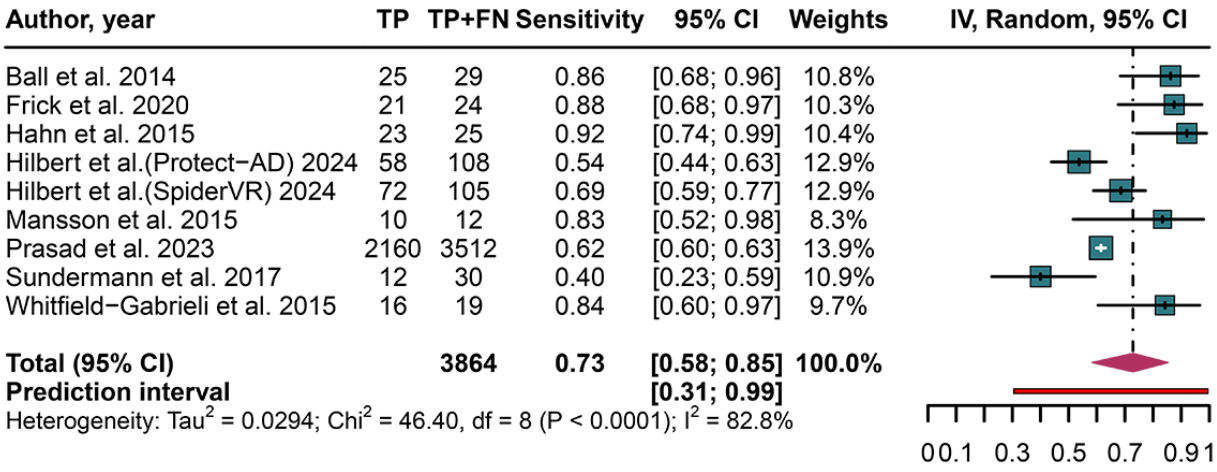

Figure 5. Forest plot displaying the pooled specificity, with 95% CIs and prediction intervals, of artificial intelligence models for predicting treatment response after cognitive behavioral therapy in patients with anxiety disorders [17,18,27,29,30,32-34]. FP: false positive; TN: true negative.

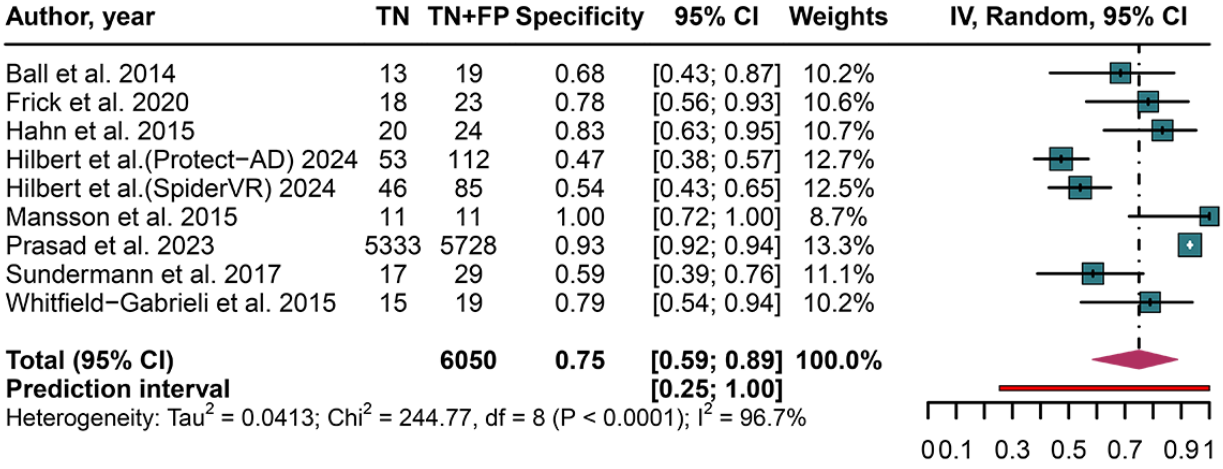

Figure 6. Forest plot displaying the pooled accuracy, with 95% CIs and prediction intervals, of artificial intelligence models for predicting treatment response after cognitive behavioral therapy in patients with anxiety disorders [17,18,27,29,30,32-34]. TN: true negative; TP: true positive.

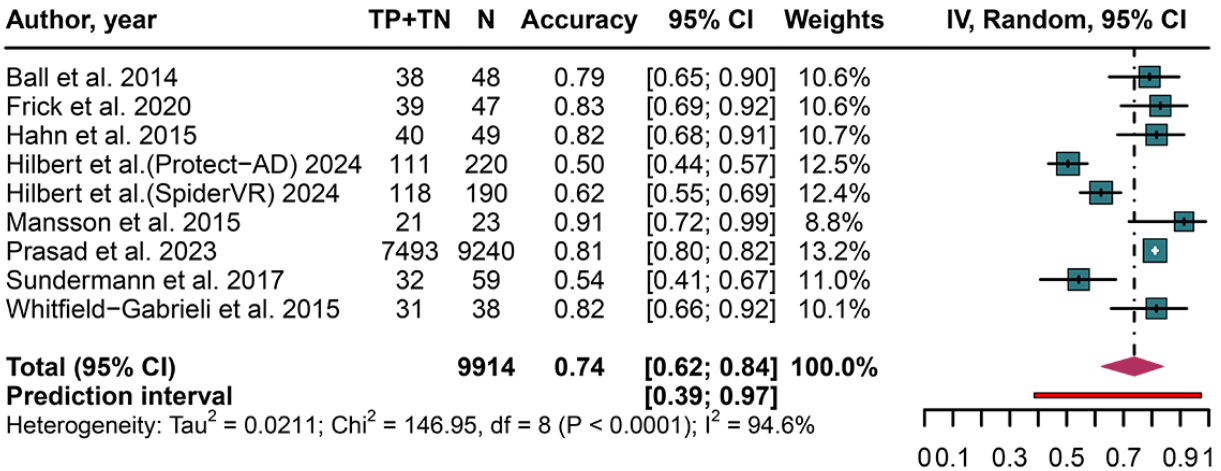

XSL•FO
RenderX

Figure 7. Summary receiver operating characteristic (SROC) curve for artificial intelligence models predicting treatment response in patients with anxiety disorders after cognitive behavioral therapy. The summary point estimates for sensitivity (SENS) and specificity (SPEC), and the area under the curve (AUC) are shown.

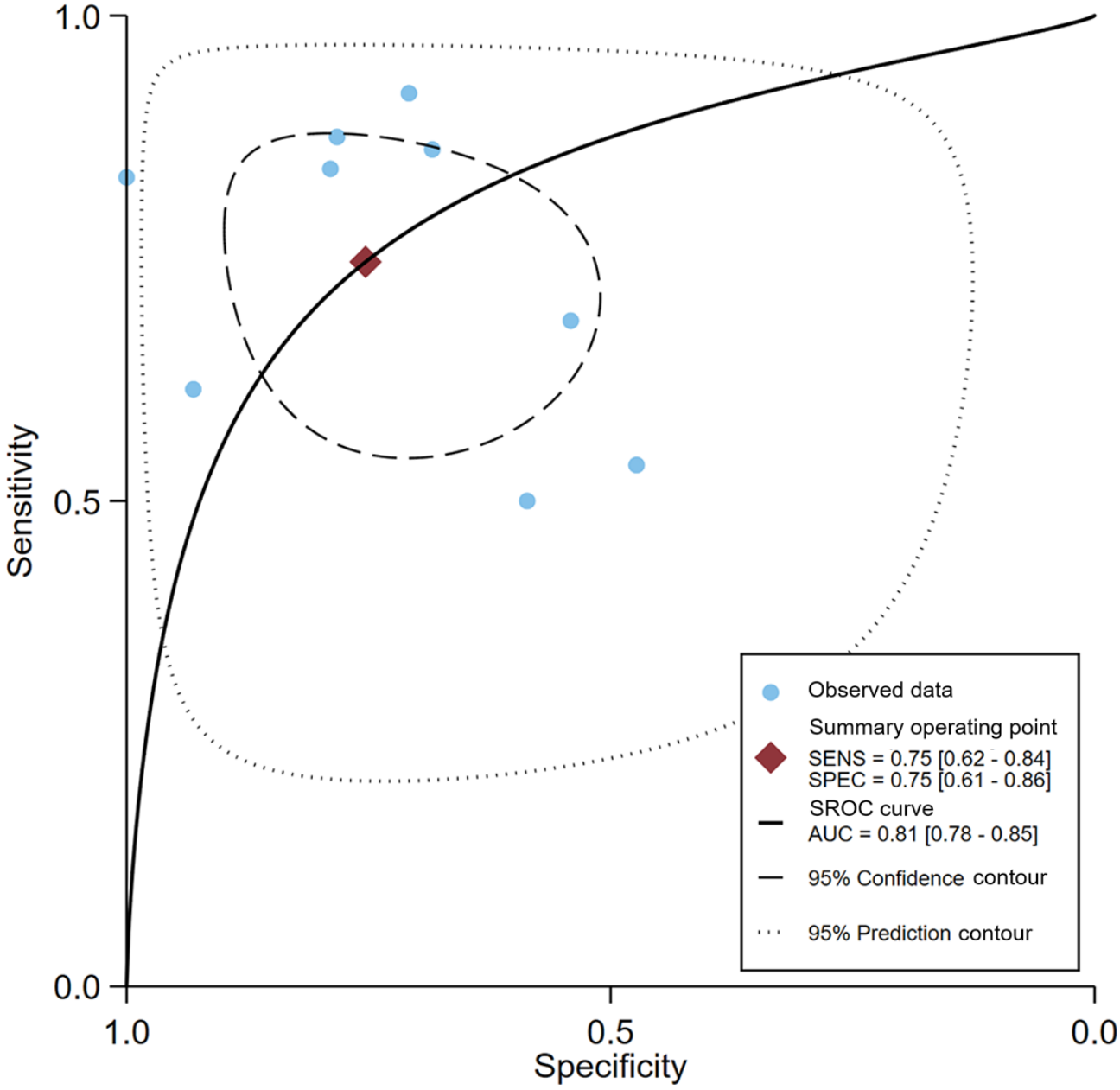

XSL•FO
RenderX

Figure 8. Bivariate boxplot of the combined logit-transformed sensitivity (SENS) and specificity (SPEC) estimates from the bivariate random-effects meta-analysis for predicting treatment response in patients with anxiety disorders after cognitive behavioral therapy. Squares represent individual study estimates; the box illustrates the joint distribution.

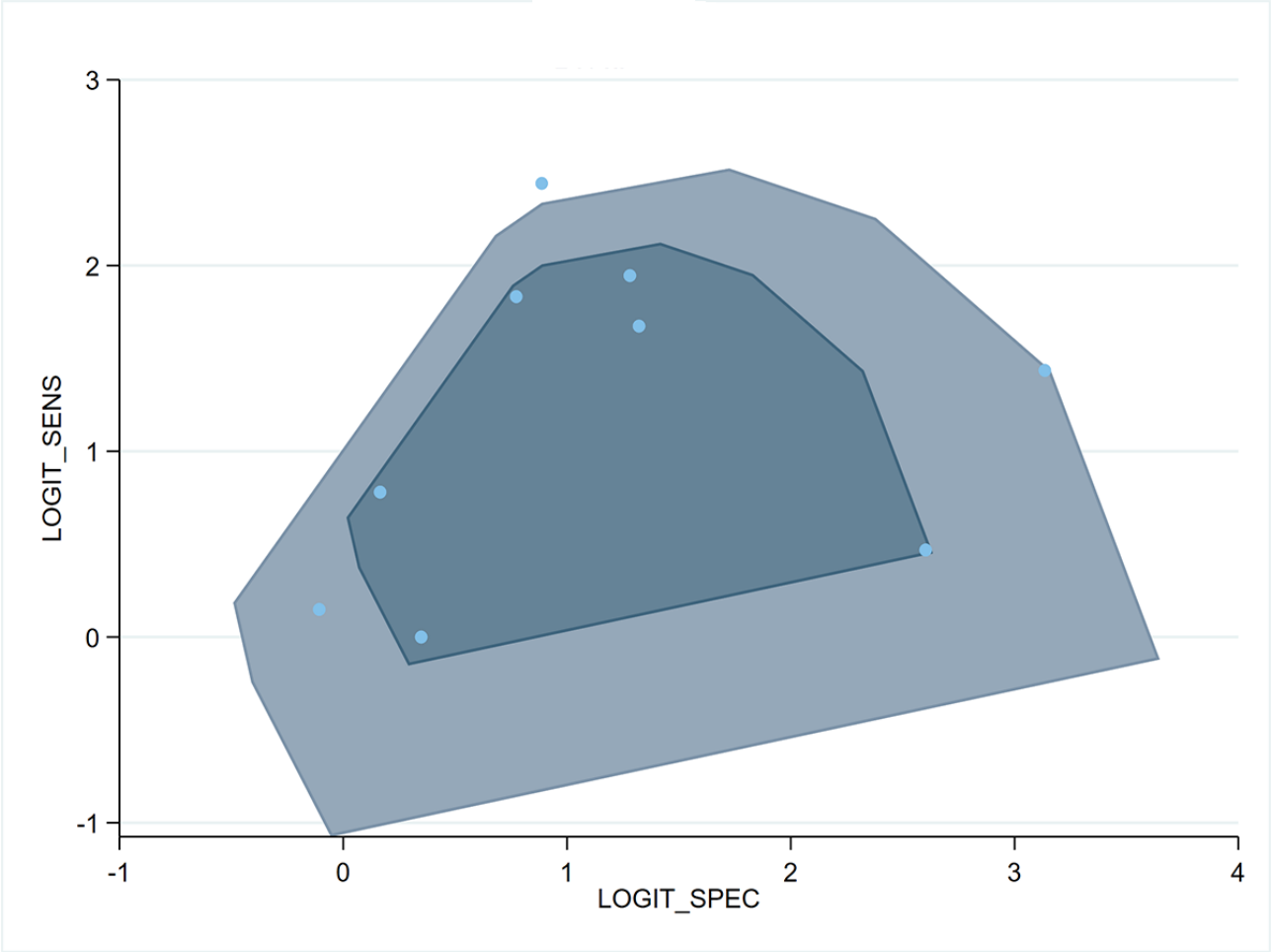

XSL•FO
RenderX

Table . Results of the univariable meta-regression analysis examining potential sources of heterogeneity in the performance of artificial intelligence models for predicting cognitive behavioral therapy treatment response in anxiety disorders.

Subgroup	Number of contingency tables	Sensitivity, value (95% CI)	Meta-regression *P* value	Specificity, value (95% CI)	Meta-regression *P* value
Number of centers			.02		.89
Single center	2	0.87 (0.76 - 0.99)		0.74 (0.46 - 1.00)	
Multiple centers	7	0.69 (0.58 - 0.80)		0.76 (0.61 - 0.90)	
Type of anxiety disorder			.01		.10
SAD[a]	3	0.86 (0.74 - 0.98)		0.87 (0.71 - 1.00)	
Others	6	0.68 (0.57 - 0.79)		0.69 (0.54 - 0.85)	
Data splitting method			.23		.28
LOOCV[b]	4	0.78 (0.62 - 0.93)		0.81 (0.63 - 0.98)	
Independent validation	5	0.72 (0.58 - 0.86)		0.71 (0.54 - 0.89)	
CBT[c] type			.77		<.001
iCBT[d]	2	0.71 (0.48 - 0.94)		0.95 (0.91 - 0.98)	
Standard CBT	6	0.74 (0.62 - 0.87)		0.61 (0.51 - 0.71)	
AI[e] algorithms			.51		<.001
DL[f]	1	0.62 (0.29 - 0.94)		0.93 (0.87 - 1.00)	
ML[g]	8	0.78 (0.68 - 0.88)		0.67 (0.57 - 0.78)	

[a]SAD: social anxiety disorder.

[b]LOOCV: leave-one-out cross-validation.

[c]CBT: cognitive behavioral therapy.

[d]iCBT: internet-based cognitive behavioral therapy.

[e]AI: artificial intelligence.

[f]DL: deep learning.

[g]ML: machine learning.

Subgroup Analysis

Within the data-type subgroup, multimodal AI models appeared to achieve higher pooled sensitivity (0.84, 95% CI 0.60 - 0.97) and accuracy (0.82, 95% CI 0.66 - 0.92) compared with single-modal approaches. In contrast, models based on either clinical/demographic or imaging data alone demonstrated comparatively lower sensitivity and accuracy. Regarding specificity, models developed using clinical/demographic data showed relatively higher estimates (0.81, 95% CI 0.55 - 0.93). Regarding AUC, only 1 study reported data within this subgroup, and in that study, imaging models achieved an AUC of 0.79 (95% CI 0.76 - 0.83) (Table 3).

XSL•FO
RenderX

Table . Subgroup meta-analysis of the pooled predictive performance of artificial intelligence models for predicting cognitive behavioral therapy response, stratified by input data type, algorithm type, and anxiety disorder classification.

Subgroup	Number of contingency tables	Sensitivity, value (95% CI)	Specificity, value (95% CI)	AUC[a], value (95% CI)	Accuracy, value (95% CI)
Type of data					
Clinical/demographic	2	0.65 (0.42 - 0.82)	0.81 (0.55 - 0.93)	—[b]	0.73 (0.55 - 0.89)
Imaging	6	0.77 (0.63 - 0.87)	0.71 (0.53 - 0.85)	0.79 (0.76 - 0.83)	0.72 (0.00 - 1.00)
Multimodal	1	0.84 (0.60 - 0.97)	0.79 (0.54 - 0.94)	—	0.82 (0.66 - 0.92)
AI[c] algorithms					
RF[d]	3	0.64 (0.58 - 0.70)	0.52 (0.45 - 0.59)	0.54 (0.50 - 0.59)	0.63 (0.27 - 0.93)
SVM[e]	4	0.82 (0.61 - 0.93)	0.75 (0.60 - 0.85)	0.83 (0.80 - 0.86)	0.78 (0.50 - 0.97)
RNN[f]	1	0.62 (0.60 - 0.63)	0.93 (0.92 - 0.94)	—	0.81 (0.81 - 0.82)
LR[g]	1	0.84 (0.60 - 0.97)	0.79 (0.54 - 0.94)	—	0.82 (0.66 - 0.92)
Type of anxiety disorder					
GAD[h] and PD[i]	1	0.86 (0.68 - 0.96)	0.68 (0.43 - 0.87)	—	0.79 (0.65 - 0.90)
SAD[j]	3	0.86 (0.78 - 0.92)	0.87 (0.41 - 1.00)	0.92 (0.85 - 0.98)	0.85 (0.72 - 0.94)
PD with AG[k]	2	0.69 (0.00 - 1.00)	0.71 (0.00 - 1.00)	—	0.69 (0.00 - 1.00)
Multiple anxiety disorders	3	0.61 (0.45 - 0.77)	0.68 (0.05 - 1.00)	0.65 (0.49 - 0.81)	0.65 (0.25 - 0.96)

[a]AUC: area under curve.

[b]Not applicable.

[c]AI: artificial intelligence.

[d]RF: random forest.

[e]SVM: support vector machine.

[f]RNN: recurrent neural network.

[g]LR: logistic regression.

[h]GAD: generalized anxiety disorder.

[i]PD: panic disorder.

[j]SAD: social anxiety disorder.

[k]AG: agoraphobia.

Within the AI algorithm subgroup, different algorithms demonstrated distinct advantages across various performance metrics. For sensitivity, LR performed best, achieving a value of 0.84 (95% CI 0.60 - 0.97). Similarly, in terms of accuracy, LR again showed optimal performance, reaching a value of 0.82 (95% CI 0.66 - 0.92). For specificity, the RNN achieved the highest value of 0.93 (95% CI 0.87 - 0.98). Regarding the AUC, only 2 algorithms produced results that could be combined, with the SVM algorithm yielding a relatively high combined AUC of 0.83 (95% CI 0.80 - 0.86).

Within the type of anxiety disorder subgroup, predictive performance varied across outcome metrics. For sensitivity, GAD and PD demonstrated relatively higher pooled estimates at 0.86 (95% CI 0.68 - 0.96). In contrast, SAD showed comparatively stronger performance in specificity and accuracy, with a pooled specificity of 0.87 (95% CI 0.41 - 1.00) and accuracy of 0.85 (95% CI 0.72 - 0.94). For discrimination, as measured by AUC, pooled estimates were available for 2 disorder types, among which SAD demonstrated a comparatively higher AUC of 0.92 (95% CI 0.85 - 0.98).

Publication Bias

Funnel plots and Egger regression tests were conducted to evaluate potential small-study effects (Figures S2-S4 in Multimedia Appendix 1). No statistically significant evidence of small-study effects was observed for sensitivity (t_7=1.75; P=.12; intercept=1.54; SE=0.88) and accuracy (t_7=−1.51; P=.17; intercept=−2.40; SE=1.59). In contrast, the test for specificity yielded a statistically significant result (t_7=−2.58; P=.04; intercept=−4.29; SE=1.66), suggesting the possible presence of small-study effects.

XSL•FO
RenderX

Discussion

Principal Findings

The systematic review and meta-analysis revealed that the model demonstrated moderate predictive performance in forecasting treatment response to CBT among patients with anxiety disorders, with sensitivity, specificity, accuracy, and AUC values of 0.73, 0.75, 0.74, and 0.81, respectively. This moderate predictive capability reflects the combined influence of multiple complex factors. First, limitations in input variables constitute a key constraint: current models primarily rely on demographic data, psychiatric history, and scale measurements, making it challenging to capture deeper biological and environmental factors [7,35]. While biomarkers, such as neuroimaging, could theoretically significantly enhance predictive accuracy, their clinical application remains constrained by cost and accessibility [32,35]. Second, data heterogeneity and sample size issues cannot be overlooked. Significant variations among patient cohorts in symptom severity, comorbid conditions, and treatment modalities diminish the model's generalizability [18,36]. Small-sample studies often overestimate predictive capability, while large-scale multicenter studies frequently yield reduced accuracy, sometimes approaching random levels [18,35].

Furthermore, limitations exist in variable selection and modeling approaches: prediction capabilities relying solely on traditional questionnaires and demographic data are constrained [35,37]. While integrating multimodal data (eg, neuroimaging, genetic, and behavioral data) may enhance performance, the comprehensive collection of such data remains challenging in real-world clinical settings [18]. Finally, challenges in practical application and external validation cannot be overlooked; many models lack independent external validation, resulting in significantly reduced accuracy and interpretability in clinical settings [38]. In summary, while current models demonstrate predictive potential, future research requires larger-scale, multicenter, and multimodal studies to enhance model robustness and clinical translation capabilities.

The subgroup analysis suggested that multimodal data integration was associated with comparatively higher sensitivity (0.84) and accuracy (0.82) than single-modal approaches. Across algorithm categories, no single model consistently outperformed others; LR showed higher sensitivity (0.84) and accuracy (0.82), RNN demonstrated higher specificity (0.93), and SVM had a comparatively higher AUC of 0.83. Predictive performance also varied across anxiety disorder subtypes. GAD and PD showed comparatively higher sensitivity (0.86), whereas SAD demonstrated higher specificity (0.87), accuracy (0.85), and AUC (0.92). Together, these findings suggest that optimal prediction may depend less on a single "best" algorithm and more on strategically matching the model architecture to the data type and clinical question [39]. Future work should therefore focus not only on multimodal data collection but also on developing tailored, interpretable modeling approaches that can be validated in real-world clinical settings [40].

The systematic review revealed that multimodal data demonstrated the most promising performance in predicting CBT treatment response for anxiety disorders, with a sensitivity of 0.84 and an accuracy of 0.82. The superior predictive capability of multimodal approaches could be attributed to several key mechanisms. First, multimodal data integration enabled the comprehensive capture of the multidimensional nature of anxiety disorders by combining neuroimaging, physiological signals, behavioral data, and clinical assessments [29,30]. Complementary information across modalities enabled AI models to detect nuanced individual differences and potential mechanisms of treatment response [27,34]. Research indicates that, compared with single-modality approaches, multimodal fusion significantly enhances predictive accuracy, particularly when incorporating models that integrate functional and structural brain connectivity [29,34]. Moreover, multimodal approaches supported personalized, dynamic prediction models that can identify high-risk nonresponders early, potentially facilitating adaptive treatment strategies [17,27]. While these findings are promising, careful validation through larger, cross-validated studies is essential to establish the clinical utility and generalizability of such AI-driven prediction models [29,33].

The meta-analysis revealed differential performance patterns across ML algorithms in predicting CBT treatment response for anxiety disorders. LR demonstrated comparatively higher sensitivity (0.84) and accuracy (0.82), potentially attributable to its ability to capture linear relationships between clinical features and treatment outcomes [7,31]. The algorithm's strength lies in its ability to model structured clinical data with minimal risk of overfitting, particularly in scenarios with balanced sample distributions [7,17]. Conversely, RNNs exhibited superior specificity (0.93), likely due to their ability to process high-dimensional, sequential data and capture complex nonlinear interactions [17]. RNN's advanced modeling approach enables more nuanced identification of nonresponders by leveraging temporal patterns in clinical measurements and platform interaction behaviors [17]. For overall discrimination, SVM achieved a pooled AUC of 0.83 (95% CI 0.80 - 0.86). Given their kernel-based formulation, SVMs may perform well with moderate sample sizes and complex feature spaces, allowing flexible decision boundaries [7,31]. Taken together, these findings suggest that algorithmic performance may vary according to data structure and modeling objectives. Rather than indicating a universally superior approach, the results support the importance of aligning model architecture with data characteristics and specific clinical prediction goals [31].

The subgroup analysis indicated that predictive performance varied across anxiety disorder subtypes. GAD and PD demonstrated comparatively higher sensitivity estimates (0.86, 95% CI 0.68 - 0.96), suggesting relatively stronger identification of treatment responders in these populations. In contrast, SAD showed higher pooled specificity (0.87, 95% CI 0.41 - 1.00) and accuracy (0.85, 95% CI 0.72 - 0.94). For discrimination as measured by AUC, pooled estimates were available for 2 disorder types, among which SAD demonstrated a comparatively higher AUC of 0.92 (95% CI 0.85 - 0.98). The variation in predictive performance across anxiety disorder subtypes may partly account for the observed heterogeneity. In the subgroup analysis, GAD and PD demonstrated higher sensitivity, which may be related to the use of structured symptom-based scales that provide relatively stable responder signals for model

XSL•FO
RenderX

training [27]. First, SAD research often relies on symptom measures that closely align core fear and avoidance dimensions with treatment outcome definitions, which may facilitate clearer model discrimination [7,34]. Second, some SAD studies incorporate high-information features beyond routine clinical variables, potentially enhancing overall predictive performance [29,32]. Third, SAD samples are frequently drawn from relatively homogeneous clinical trials with standardized intervention protocols, which may reduce heterogeneity and improve model stability [7,32].

Comparison With Prior Work

In the systematic review and meta-analysis published by Vieira et al [35] in 2022, the authors evaluated ML approaches for predicting CBT treatment response across multiple psychiatric disorders. Their analysis included 24 studies comprising 7497 patients and covered 5 diagnostic groups: major depressive disorder, obsessive-compulsive disorder (OCD), posttraumatic stress disorder (PTSD), anxiety disorders, and substance use disorders. The overall pooled accuracy was 0.74 (95% CI 0.70 - 0.78), with variability across diagnostic categories. PTSD, anxiety disorders, and OCD demonstrated comparatively higher accuracy estimates (0.787, 0.776, and 0.761, respectively). In comparison, our study focused specifically on anxiety disorders and yielded a pooled accuracy of 0.74 (95% CI 0.62 - 0.84), which is comparable to the overall estimate reported by Vieira et al [35]. However, our analysis extended beyond accuracy by systematically synthesizing sensitivity, specificity, and AUC to provide a more comprehensive assessment of predictive performance. In addition, we conducted subgroup analyses examining the potential influence of data types (eg, clinical and neuroimaging), algorithm types, and disease types on model performance, thereby offering a more granular evaluation within the anxiety disorder domain. Bubble plots were used to illustrate temporal trends and the distribution of study characteristics across performance metrics.

Heterogeneity

The substantial heterogeneity observed across studies may influence the overall strength of evidence for AI-based prediction models. Given the anticipated between-study variability, a bivariate random-effects model was applied to pool sensitivity and specificity estimates. Meta-regression and subgroup analyses were further conducted to explore potential sources of heterogeneity. The results indicated that heterogeneity in sensitivity was significantly associated with the number of study centers (P=.02) and the anxiety disorder subtype (P=.01). In contrast, heterogeneity in specificity was significantly associated with CBT type (P<.001) and AI algorithm category (P<.001). The comparatively higher sensitivity observed in single-center studies, compared with multicenter studies, may reflect more homogeneous samples and more consistent data collection procedures [41]. Differences across anxiety disorder types may relate to variations in symptom structure, measurement alignment, and sample characteristics, which could influence the clarity of the responder signals available to the models [42]. Meanwhile, differences between iCBT and standard CBT may reflect variations in intervention structure and in the intensity of monitoring [43]. Variability in specificity across algorithm categories (DL vs ML) may relate to differences in model complexity and data representation strategies [44].

The bivariate boxplot suggested that 1 study [30] may represent a potential source of heterogeneity. Nevertheless, heterogeneity is likely multifactorial. Additional contributors may include differences in patient age, disorder severity, prior treatment exposure, geographic setting, sample size, feature selection strategies, data preprocessing methods, and hyperparameter optimization [45-48]. The combined influence of these methodological and clinical factors may account for the variability in predictive performance. Future research would benefit from more standardized reporting and systematic consideration of these variables to enhance reproducibility and external validity.

Implications for Practice and Research

Discussing the clinical interpretation of our results, the AI models demonstrated moderate predictive value for CBT treatment response in anxiety disorders, with the strongest performance observed for SAD. Multimodal models, which combine clinical demographics with imaging data, showed relatively superior predictive performance compared with unimodal approaches. This suggests that AI could support clinicians in making earlier, more informed therapeutic decisions and potentially improve patient outcomes by enabling timely treatment adjustments [49,50]. Specifically, these predictive models could be integrated into clinical workflows at key decision points: (1) during initial assessment to help stratify patients by likelihood of response, informing the intensity of early monitoring or the consideration of augmented treatment protocols, and (2) during the early stages of CBT to identify potential nonresponders, allowing for a timely switch or integration of other evidence-based interventions before disengagement occurs [35,51].

However, AI should serve as a decision-support tool rather than replacing clinician judgment, since treatment choices for anxiety disorders are influenced by not only severity but also patient preferences, comorbidity, and functional goals [52]. In the treatment of anxiety disorders, complex therapeutic decisions not only depend on assessing disease severity but also closely relate to the patient's individual condition and needs [53]. For real-world deployment, developing AI tools as clinician-facing dashboards that present predictive probabilities alongside key clinical context, rather than as autonomous decision systems, is crucial for fostering trust and facilitating shared decision-making [54]. Notably, among the studies we included, external validation sets were relatively scarce, and no research has yet focused solely on predicting outcomes for external validation cohorts. Future studies are needed to evaluate the generalization capabilities of AI models. Furthermore, limitations to the future adoption of AI include scarce annotated anxiety disorder data and regulatory hurdles [55]. Technical challenges persist in data availability, model interpretability, and transparency [56]. Advances in few-shot learning, self-supervised models, and centralized platforms may support an integrated AI ecosystem [57].

XSL•FO
RenderX

Limitations

When interpreting the findings of this systematic review and meta-analysis on the AI prediction of CBT treatment response for anxiety disorders, several limitations should be considered. First, the PROBAST+AI assessment indicated a high risk of bias in over half of the included studies, particularly in analytical methodology, which may have influenced the pooled estimates and contributed to between-study heterogeneity. Future studies should adopt more rigorous design and transparent reporting standards. Second, variability in the definition of treatment response across studies may have affected comparability. Although most scales assessed anxiety symptom reduction, differences in operational criteria could influence predictive performance. Greater standardization of outcome definitions would enhance consistency in future research. Third, the substantial heterogeneity and wide PIs suggest that model performance may vary considerably across clinical settings. Therefore, the pooled estimates should be interpreted cautiously and regarded as preliminary within the context of current methodological constraints. Fourth, although heterogeneity was observed across studies, this likely reflects differences in study design, patient characteristics, and modeling approaches. The reported PIs indicate that performance may vary across settings, highlighting the importance of external validation in future research. Finally, due to limited data availability, direct head-to-head comparisons between AI models and routinely used clinical scales were not feasible. Future research should incorporate comparative designs to better assess the incremental value of AI-based prediction tools. Overall, while AI models demonstrate potential for supporting treatment decision-making, stronger evidence from externally validated and methodologically robust studies is needed before clinical implementation.

Conclusions

This meta-analysis quantitatively synthesized AI performance in predicting CBT response for anxiety disorders, moving beyond narrative reviews to provide pooled evidence. In contrast to existing reviews that encompass broader diagnostic groups, our focused approach establishes a precise benchmark for this clinical domain, highlighting the current moderate overall performance. Furthermore, we extend beyond previous work by demonstrating the superior predictive utility of multimodal data, identifying SAD as the most predictable subtype, and systematically evaluating the impact of data modalities and algorithm types. Future efforts should prioritize robustly validated multimodal models, laying essential groundwork for the potential development of AI-assisted tools to personalize treatment planning in anxiety disorders.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing, adapting and adjusting emotional tone, translation, and reformatting. The GAI tool used was Sider. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. The declaration has been submitted by JL.

Funding

The authors declare that no funding was received for this study. No funding body had any role in the study design, data collection, analysis, interpretation, or writing of the manuscript.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

JL contributed to conceptualization, critical revision of the manuscript, and final approval. JL, JW, ZW, and MISBAA contributed to data curation, formal analysis, investigation, and methodology. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data to support the study.
[DOCX File, 252 KB - jmir_v28i1e86079_app1.docx]

Multimedia Appendix 2
Technical aspects and diagnostic performance data extracted from the included studies.
[DOCX File, 27 KB - jmir_v28i1e86079_app2.docx]

Checklist 1

XSL•FO
RenderX

PRISMA-DTA checklist.
[DOCX File, 21 KB - jmir_v28i1e86079_app3.docx]

Checklist 2
PRISMA 2020 Abstract checklist
[DOCX File, 19 KB - jmir_v28i1e86079_app4.docx]

Checklist 3
PRISMA-S checklist
[DOCX File, 20 KB - jmir_v28i1e86079_app5.docx]

References

1. Chodavadia P, Teo I, Poremski D, Fung DSS, Finkelstein EA. Prevalence and economic burden of depression and anxiety symptoms among Singaporean adults: results from a 2022 web panel. BMC Psychiatry 2023 Feb 14;23(1):104. [doi: 10.1186/s12888-023-04581-7] [Medline: 36782116]
2. Keramat SA, Comans T, Pearce A, Basri R, Hashmi R, Dissanayaka NN. Psychological distress and productivity loss: a longitudinal analysis of Australian working adults. Eur J Health Econ 2025 Nov;26(8):1503-1524. [doi: 10.1007/s10198-025-01764-9] [Medline: 40304834]
3. Powell C, Chiu CY, Sun X, So SHW. A meta-analysis on the efficacy of low-intensity cognitive behavioural therapy for generalised anxiety disorder. BMC Psychiatry 2024 Jan 2;24(1):10. [doi: 10.1186/s12888-023-05306-6] [Medline: 38166836]
4. Papola D, Miguel C, Mazzaglia M, et al. Psychotherapies for generalized anxiety disorder in adults: a systematic review and network meta-analysis of randomized clinical trials. JAMA Psychiatry 2024 Mar 1;81(3):250-259. [doi: 10.1001/jamapsychiatry.2023.3971] [Medline: 37851421]
5. Kravchenko O, Bäckman J, Mataix-Cols D, et al. Clinical, genetic, and sociodemographic predictors of symptom severity after internet-delivered cognitive behavioural therapy for depression and anxiety. BMC Psychiatry 2025 May 30;25(1):555. [doi: 10.1186/s12888-025-07012-x] [Medline: 40448103]
6. Kambeitz-Ilankovic L, Rzayeva U, Völkel L, et al. A systematic review of digital and face-to-face cognitive behavioral therapy for depression. NPJ Digit Med 2022 Sep 15;5(1):144. [doi: 10.1038/s41746-022-00677-8] [Medline: 36109583]
7. Bukhari Q, Rosenfield D, Hofmann SG, Gabrieli JDE, Ghosh SS. Predicting treatment response to cognitive behavior therapy in social anxiety disorder on the basis of demographics, psychiatric history, and scales: a machine learning approach. PLoS ONE 2025;20(3):e0313351. [doi: 10.1371/journal.pone.0313351] [Medline: 40100879]
8. Villarreal-Zegarra D, Paredes-Angeles R, Mayo-Puchoc N, Arenas-Minaya E, Huarcaya-Victoria J, Copez-Lonzoy A. Psychometric properties of the GAD-7 (General Anxiety Disorder-7): a cross-sectional study of the Peruvian general population. BMC Psychol 2024 Apr 2;12(1):183. [doi: 10.1186/s40359-024-01688-8] [Medline: 38566138]
9. Manzar MD, Kashoo FZ, Salahuddin M, et al. Hamilton Rating Scale for Anxiety: exploring validity with robust measures of classical theory parameters and a rating scale model in university students. BJPsych Open 2025 Aug 12;11(5):e176. [doi: 10.1192/bjo.2025.10055] [Medline: 40790884]
10. Zhang CY, Voort JLV, Yuruk D, et al. A characterization of the Clinical Global Impression Scale thresholds in the treatment of adolescent depression across multiple rating scales. J Child Adolesc Psychopharmacol 2022 Jun;32(5):278-287. [doi: 10.1089/cap.2021.0111] [Medline: 35704877]
11. Arrow K, Resnik P, Michel H, et al. Evaluating the use of online self-report questionnaires as clinically valid mental health monitoring tools in the clinical whitespace. Psychiatr Q 2023 Jun;94(2):221-231. [doi: 10.1007/s11126-023-10022-1] [Medline: 37145257]
12. Wang S, Lu J, Zheng G, et al. Poor performance of PHQ-9 and GAD-7 in screening clinical depression and anxiety among a large sample of Chinese children and adolescents. BMC Psychiatry 2025 Mar 29;25(1):301. [doi: 10.1186/s12888-025-06754-y] [Medline: 40158158]
13. Bertie LA, Arendt K, Coleman JRI, et al. Patterns of sub-optimal change following CBT for childhood anxiety. J Child Psychol Psychiatry 2024 Dec;65(12):1612-1623. [doi: 10.1111/jcpp.14009] [Medline: 38817012]
14. Park JH, Shin YB, Jung D, et al. Machine learning prediction of anxiety symptoms in social anxiety disorder: utilizing multimodal data from virtual reality sessions. Front Psychiatry 2024;15:1504190. [doi: 10.3389/fpsyt.2024.1504190] [Medline: 39896993]
15. Sadeh-Sharvit S, Camp TD, Horton SE, et al. Effects of an artificial intelligence platform for behavioral interventions on depression and anxiety symptoms: randomized clinical trial. J Med Internet Res 2023 Jul 10;25:e46781. [doi: 10.2196/46781] [Medline: 37428547]
16. Habicht J, Dina LM, McFadyen J, et al. Generative AI-enabled therapy support tool for improved clinical outcomes and patient engagement in group therapy: real-world observational study. J Med Internet Res 2025 Mar 10;27:e60435. [doi: 10.2196/60435] [Medline: 40063074]

XSL•FO
RenderX

17. Prasad N, Chien I, Regan T, et al. Deep learning for the prediction of clinical outcomes in internet-delivered CBT for depression and anxiety. PLoS ONE 2023;18(11):e0272685. [doi: 10.1371/journal.pone.0272685] [Medline: 38011176]
18. Hilbert K, Böhnlein J, Meinke C, et al. Lack of evidence for predictive utility from resting state fMRI data for individual exposure-based cognitive behavioral therapy outcomes: a machine learning study in two large multi-site samples in anxiety disorders. Neuroimage 2024 Jul 15;295(120639):120639. [doi: 10.1016/j.neuroimage.2024.120639] [Medline: 38796977]
19. Ohi K, Fujikane D, Takai K, et al. Clinical features and genetic mechanisms of anxiety, fear, and avoidance: a comprehensive review of five anxiety disorders. Mol Psychiatry 2025 Oct;30(10):4928-4936. [doi: 10.1038/s41380-025-03155-1] [Medline: 40830577]
20. Basterfield C, Newman MG. Development of a machine learning-based multivariable prediction model for the naturalistic course of generalized anxiety disorder. J Anxiety Disord 2025 Mar;110:102978. [doi: 10.1016/j.janxdis.2025.102978] [Medline: 39904097]
21. Cruz-Gonzalez P, He AWJ, Lam EP, et al. Artificial intelligence in mental health care: a systematic review of diagnosis, monitoring, and intervention applications. Psychol Med 2025 Feb 6;55:e18. [doi: 10.1017/S0033291724003295] [Medline: 39911020]
22. Salameh JP, Bossuyt PM, McGrath TA, et al. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. BMJ 2020 Aug 14;370:m2632. [doi: 10.1136/bmj.m2632] [Medline: 32816740]
23. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
24. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ 2025 Mar 24;388:e082505. [doi: 10.1136/bmj-2024-082505] [Medline: 40127903]
25. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
26. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
27. Ball TM, Stein MB, Ramsawh HJ, Campbell-Sills L, Paulus MP. Single-subject anxiety treatment outcome prediction using functional neuroimaging. Neuropsychopharmacology 2014 Apr;39(5):1254-1261. [doi: 10.1038/npp.2013.328] [Medline: 24270731]
28. Bertie LA, Quiroz JC, Berkovsky S, et al. Predicting remission following CBT for childhood anxiety disorders: a machine learning approach. Psychol Med 2024 Dec 17;17:1-11. [doi: 10.1017/S0033291724002654] [Medline: 39686883]
29. Frick A, Engman J, Alaie I, et al. Neuroimaging, genetic, clinical, and demographic predictors of treatment response in patients with social anxiety disorder. J Affect Disord 2020 Jan 15;261:230-237. [doi: 10.1016/j.jad.2019.10.027] [Medline: 31655378]
30. Hahn T, Kircher T, Straube B, et al. Predicting treatment response to cognitive behavioral therapy in panic disorder with agoraphobia by integrating local neural information. JAMA Psychiatry 2015 Jan;72(1):68-74. [doi: 10.1001/jamapsychiatry.2014.1741] [Medline: 25409415]
31. Hentati Isacsson N, Ben Abdesslem F, Forsell E, Boman M, Kaldo V. Methodological choices and clinical usefulness for machine learning predictions of outcome in Internet-based cognitive behavioural therapy. Commun Med (Lond) 2024 Oct 10;4(1):196. [doi: 10.1038/s43856-024-00626-4] [Medline: 39384934]
32. Månsson KNT, Frick A, Boraxbekk CJ, et al. Predicting long-term outcome of Internet-delivered cognitive behavior therapy for social anxiety disorder using fMRI and support vector machine learning. Transl Psychiatry 2015 Mar 17;5(3):e530. [doi: 10.1038/tp.2015.22] [Medline: 25781229]
33. Sundermann B, Bode J, Lueken U, et al. Support vector machine analysis of functional magnetic resonance imaging of interoception does not reliably predict individual outcomes of cognitive behavioral therapy in panic disorder with agoraphobia. Front Psychiatry 2017;8:99. [doi: 10.3389/fpsyt.2017.00099] [Medline: 28649205]
34. Whitfield-Gabrieli S, Ghosh SS, Nieto-Castanon A, et al. Brain connectomics predict response to treatment in social anxiety disorder. Mol Psychiatry 2015 May;21(5):680-685. [doi: 10.1038/mp.2015.109] [Medline: 26260493]
35. Vieira S, Liang X, Guiomar R, Mechelli A. Can we predict who will benefit from cognitive-behavioural therapy? A systematic review and meta-analysis of machine learning studies. Clin Psychol Rev 2022 Nov;97:102193. [doi: 10.1016/j.cpr.2022.102193] [Medline: 35995023]
36. Jankowsky K, Krakau L, Schroeders U, Zwerenz R, Beutel ME. Predicting treatment response using machine learning: a registered report. Br J Clin Psychol 2024 Jun;63(2):137-155. [doi: 10.1111/bjc.12452] [Medline: 38111213]
37. Hornstein S, Forman-Hoffman V, Nazander A, Ranta K, Hilbert K. Predicting therapy outcome in a digital mental health intervention for depression and anxiety: a machine learning approach. Digit Health 2021;7:20552076211060659. [doi: 10.1177/20552076211060659] [Medline: 34868624]

XSL•FO
RenderX

38. Chekroud AM, Bondar J, Delgadillo J, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry 2021 Jun;20(2):154-170. [doi: 10.1002/wps.20882] [Medline: 34002503]
39. Piccolo SR, Mecham A, Golightly NP, Johnson JL, Miller DB. The ability to classify patients based on gene-expression data varies by algorithm and performance metric. PLoS Comput Biol 2022 Mar;18(3):e1009926. [doi: 10.1371/journal.pcbi.1009926] [Medline: 35275931]
40. AlSaad R, Abd-Alrazaq A, Boughorbel S, et al. Multimodal large language models in health care: applications, challenges, and future outlook. J Med Internet Res 2024 Sep 25;26:e59505. [doi: 10.2196/59505] [Medline: 39321458]
41. Bafeta A, Dechartres A, Trinquart L, Yavchitz A, Boutron I, Ravaud P. Impact of single centre status on estimates of intervention effects in trials with continuous outcomes: meta-epidemiological study. BMJ 2012 Feb 14;344:e813. [doi: 10.1136/bmj.e813] [Medline: 22334559]
42. Fomenko A, Dümmler D, Aktürk Z, et al. Hospital Anxiety and Depression Scale Anxiety subscale (HADS-A) for detecting anxiety disorders in adults. Cochrane Database Syst Rev 2025 Jul 2;7(7):CD015456. [doi: 10.1002/14651858.CD015456] [Medline: 40600405]
43. Hedman-Lagerlöf E, Carlbring P, Svärdman F, Riper H, Cuijpers P, Andersson G. Therapist-supported Internet-based cognitive behaviour therapy yields similar effects as face-to-face therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. World Psychiatry 2023 Jun;22(2):305-314. [doi: 10.1002/wps.21088] [Medline: 37159350]
44. Abrol A, Fu Z, Salman M, et al. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. Nat Commun 2021 Jan 13;12(1):353. [doi: 10.1038/s41467-020-20655-6] [Medline: 33441557]
45. Kartoun U, Khurshid S, Kwon BC, et al. Prediction performance and fairness heterogeneity in cardiovascular risk models. Sci Rep 2022 Jul 22;12(1):12542. [doi: 10.1038/s41598-022-16615-3] [Medline: 35869152]
46. de Jong VMT, Moons KGM, Eijkemans MJC, Riley RD, Debray TPA. Developing more generalizable prediction models from pooled studies and large clustered data sets. Stat Med 2021 Jul 10;40(15):3533-3559. [doi: 10.1002/sim.8981] [Medline: 33948970]
47. Benkarim O, Paquola C, Park BY, et al. Population heterogeneity in clinical cohorts affects the predictive accuracy of brain imaging. PLoS Biol 2022 Apr;20(4):e3001627. [doi: 10.1371/journal.pbio.3001627] [Medline: 35486643]
48. Li Y, Yu C, Zhao Y, Yao W, Aseltine RH, Chen K. Pursuing sources of heterogeneity in modeling clustered population. Biometrics 2022 Jun;78(2):716-729. [doi: 10.1111/biom.13434] [Medline: 33527347]
49. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
50. Khalifa M, Albadawy M. Artificial intelligence for clinical prediction: exploring key domains and essential functions. Computer Methods and Programs in Biomedicine Update 2024;5:100148. [doi: 10.1016/j.cmpbup.2024.100148]
51. Hammelrath L, Hilbert K, Heinrich M, Zagorscak P, Knaevelsrud C. Select or adjust? How information from early treatment stages boosts the prediction of non-response in internet-based depression treatment. Psychol Med 2024 Jun;54(8):1641-1650. [doi: 10.1017/S0033291723003537] [Medline: 38087867]
52. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018 Mar 15;378(11):981-983. [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
53. Tendolkar VD, Chouhan DS, Roy D, et al. Advances in evidence-based interventions for depression and anxiety: implications for mental health nursing practice. Int J Environ Sci 2025;11(23s):826-834 [FREE Full text] [doi: 10.64252/mev1b885]
54. As'ad M, Faran N, Joharji H. AI-Supported Shared Decision-Making (AI-SDM): conceptual framework. JMIR AI 2025 Aug 7;4:e75866. [doi: 10.2196/75866] [Medline: 40773762]
55. Manole A, Cârciumaru R, Brînza R, Manole F. An exploratory investigation of chatbot applications in anxiety management: a focus on personalized interventions. Information 2024;16(1):11. [doi: 10.3390/info16010011]
56. Das KP, Gavade P. A review on the efficacy of artificial intelligence for managing anxiety disorders. Front Artif Intell 2024;7:1435895. [doi: 10.3389/frai.2024.1435895] [Medline: 39479229]
57. Sosna J, Joskowicz L, Saban M. Navigating the AI landscape in medical imaging: a critical analysis of technologies, implementation, and implications. Radiology 2025 Jun;315(3):e240982. [doi: 10.1148/radiol.240982] [Medline: 40552997]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
CBT: cognitive behavioral therapy
CGI-I: Clinical Global Impressions–Improvement
DL: deep learning
FN: false negative
FP: false positive
GAD: generalized anxiety disorder
GAD-7: Generalized Anxiety Disorder-7

XSL•FO
RenderX

HAM-A: Hamilton Anxiety Rating Scale
HKSJ: Hartung-Knapp-Sidik-Jonkman
LR: logistic regression
LSAS: Liebowitz Social Anxiety Scale
ML: machine learning
OASIS: Overall Anxiety Severity and Impairment Scale
OCD: obsessive-compulsive disorder
PD: panic disorder
PHQ-9: Patient Health Questionnaire-9
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST+AI: Prediction Model Risk of Bias Assessment Tool for Artificial Intelligence
PTSD: posttraumatic stress disorder
RF: random forest
RNN: recurrent neural network
SAD: social anxiety disorder
SPQ: Spider Phobia Questionnaire
SVM: support vector machine
TN: true negative
TP: true positive

Edited by S Brini; submitted 17.Oct.2025; peer-reviewed by C Ahmadu, D Ghosh, S Mohanadas, SG Virupaksha, Y Zhang; revised version received 12.Feb.2026; accepted 24.Feb.2026; published 18.Mar.2026.

Please cite as:
Liu J, Wang J, Wu Z, Adam Assim MISB
Artificial Intelligence for Predicting Treatment Response in Patients With Anxiety Disorders After Cognitive Behavioral Therapy: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e86079
URL: https://www.jmir.org/2026/1/e86079
doi:10.2196/86079

© Jiawen Liu, Junhui Wang, Zhaobin Wu, Mohamad Ibrani Shahrimin Bin Adam Assim. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Anticipating Moral and Economic Considerations, Opportunities, and Potential Frictions for AI in Medical Imaging: Multistakeholder Cocreation Study

Martin Bastiaan Schilder[1], MSc; Alexandra Keyser[2], MSc; Susan van Hees[2], PhD; Alessandro Sbrizzi[1], PhD; Wouter Pieter Christiaan Boon[2], PhD

[1]Computational Imaging Group for MR Therapy and Diagnostics, Center for Image Sciences, University Medical Center Utrecht, Heidelberglaan 100, Utrecht, The Netherlands

[2]Faculty of Geosciences, Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, The Netherlands

Corresponding Author:
Martin Bastiaan Schilder, MSc
Computational Imaging Group for MR Therapy and Diagnostics, Center for Image Sciences, University Medical Center Utrecht, Heidelberglaan 100, Utrecht, The Netherlands

Abstract

Background: Artificial intelligence (AI) promises to significantly impact daily radiology practices. Numerous studies have already been conducted that anticipate this potentially disruptive innovation. So far, these studies have mainly focused on single topics, such as “trust,” or investigating perspectives of single stakeholder groups, such as “radiologists.”

Objective: This study aims to explore future directions for AI in radiology by incorporating perspectives of a heterogeneous group of stakeholders on a broad spectrum of moral and economic topics. It also aims to cocreate and reflect with a broad range of stakeholders on viable implementation scenarios for scalable AI applications in radiology in the Netherlands, thereby identifying potential opportunities and frictions, with a focus on moral and economic considerations.

Methods: To inform the workshop design, a nonsystematic narrative literature search was performed to deepen our understanding of key moral and economic considerations at play in the field of radiology and AI. Workshop participants, representing a wide range of actors including radiologists, innovators, and patient representatives, were selected using purposive sampling. Data were collected in a cocreation workshop. In 3 subsequent rounds, mixed over 3 breakout groups, a total of 17 participants were asked to (1) map what they considered important moral and economic considerations, (2) envision possible future scenarios for AI in radiology, and (3) discuss opportunities, frictions, and routes to success. Transcribed recordings were coded and cross-checked.

Results: Workshop participants envision future AI-driven scenarios, ranging from extramural imaging departments for increased accessibility to health care, to multimodal data integration for human-centered AI-enhanced diagnostics. Seven themes emerge from the discussions during the workshop: (1) trust and efficiency of AI technologies, (2) responsibilities in clinical decision-making when AI is involved, (3) diagnosis as a one-off versus an iterative process, (4) regulations as a requirement or a restriction, (5) economic benefits or drawbacks, (6) trade-off between amount of information required and patient privacy, and (7) environmental considerations.

Conclusions: Reflecting on the 7 emerging themes, we identify three overarching topics: (1) human-AI collaboration and trust, (2) governance, regulation, and ethical safeguards, and (3) value creation and sustainability. These topics highlight the need to balance technological advancements with ethical responsibility, institutional accountability, and societal benefit. They also underscore the importance of designing AI systems that not only perform well but are also trusted and aligned with clinical workflows and patient values. These overarching themes offer a lens through which future research and policy can navigate the complex interplay between innovation, regulation, and real-world implementation. Future research is needed to validate the generalizability of the results across various countries and health care settings.

(J Med Internet Res 2026;28:e83407) doi:10.2196/83407

KEYWORDS

cocreation; stakeholder perspectives; radiology; artificial intelligence; responsible research and innovation

XSL•FO
RenderX

Introduction

Background

As health care systems worldwide face increasing demand and complexity, the integration of advanced technologies such as artificial intelligence (AI) is gaining momentum. Many clinical experts and sector specialists consider AI to have transformative potential for the field of radiology [1-4]. They argue, for instance, that AI offers the potential to enhance diagnostic accuracy, to support early disease detection, and to personalize patient care through applications such as image segmentation, anomaly detection, and predictive modeling [1,5-7]. These applications could automate routine tasks and uncover subtle patterns that can be missed by the human eye [1,8] and help radiologists to focus more on complex clinical decisions such as diagnosing intricate disease presentations [7].

There is a growing body of literature discussing the promises of AI in radiology from different perspectives. From a moral perspective, many scholars propose AI to contribute to safer, more patient-centered care by reducing diagnostic errors and unnecessary interventions [9]. Economically, proponents argue that AI can contribute to streamlining radiology workflows by improving productivity, shortening turnaround times, and lowering health care costs through early detection of disease and more efficient resource allocation [7,10-12]. Alongside these promises, scholars warn that we should remain sensitive to risks and challenges, such as algorithm opacity and profit-driven motives, created by an emerging role of AI in health care [13,14].

Despite these promises, much of the existing research on the role of AI in radiology remains limited in scope. So far, various studies have explored AI in radiology, mostly from a single perspective, such as trust or acceptance [15,16]. Moreover, existing research has primarily explored these perspectives within only a single stakeholder group or a limited range of groups, such as radiologists, innovators, or patients [7,15,17-20]. These studies also have predominantly applied retrospective methods, such as surveys and questionnaires [16,20-22]. Although these earlier studies offer valuable insights, there is a need for research broadening the perspectives and stakeholders. Adopting a broader approach enables the examination of interactions among stakeholders while situating their perspectives within the wider set of trade-offs and contextual factors.

Objective

The study aimed to cocreate and reflect with a broad range of stakeholders on viable implementation scenarios for scalable AI applications in radiology in the Netherlands, thereby identifying potential opportunities and frictions, with a focus on moral and economic considerations. Drawing on insights from responsible innovation, we endeavored to include a wide variety of views in an early stage of the emergence of a technology, anticipating potential implications and frictions, aiming to increase chances for successful innovations [17,23,24]. The inclusion of varied perspectives, especially those of patients [25,26], has been widely advocated as essential to shaping relevant research agendas and outcomes [15,16,27,28]. Early-stage engagement should increase opportunities for aligning technology development not only with clinical needs and regulatory requirements but also to include current experiences and care practices, anticipate expectations and underlying values [17,27,29-33], and answer questions regarding commercialization [34]. Engaging stakeholders encourages iterative feedback, increasing the chance of AI systems to evolve alongside real-world needs and cultivate shared ownership of the innovation process [8].

During a cocreation workshop, stakeholders were asked to build concrete implementation scenarios for AI in imagined future intra- and extramural settings and to reflect on expected opportunities and potential frictions created by moral and economic considerations. The forward-looking, participatory approach that we used contributed to a more comprehensive understanding of possible future pathways, which we present in the "Results" section. Subsequently, we describe the 7 themes that emerged from these scenarios, touching on various moral and economic considerations, involving among others the role of the human (radiologist and patients), as well as issues related to regulations, economic considerations, and the environmental impact. Finally, the "Discussion" section further reflects on these themes, subsequently concluding with 3 key implications for futures of AI in radiology.

Methods

Study Design

Our research approach reflects starting points from responsible innovation. We started with a nonsystematic narrative literature search to identify existing topics and gaps in the scientific discourse. These topics served as input for the cocreation workshop in which future scenarios for AI in radiology were developed. In the workshop, we took an explorative, anticipatory, and cocreation approach, in which stakeholders were allowed to raise and reflect on interconnected topics, rather than addressing these topics in isolation.

Identification of Workshop Topics

To gain an initial understanding and to provide the research team with context of the recurring topics and stakeholders in the field of AI, radiology, and business models, we conducted a nonsystematic narrative literature search aimed at identifying key topics and trends in the field [35,36]. In light of the fact that we aimed for a broad group of topics and stakeholders, we were a priori interested in both moral and economic considerations. Resultingly, our approach included keyword-based searches in primarily PubMed and Google Scholar using search terms such as "Radiology," "AI," "Transparency," and "Value proposition." These terms were iteratively refined as understanding of the topic evolved and, in the end, also included "Algorithmic fairness" and "Revenue streams" as search terms, as is common practice in a narrative review [36]. We also used forward and backward snowballing from key publications to capture additional influential works. In addition, we browsed the digital archives of leading journals in the field, including *Journal of Medical Internet Research* (and relevant sister journals), *European Radiology*, and *Insights into Imaging*. This flexible and exploratory search strategy

XSL•FO
RenderX

suited our objective of mapping conceptual developments and identifying recurrent topics across the literature rather than providing an exhaustive or systematic synthesis. Textbox 1 provides the main findings of this review.

Textbox 1. Topics covering moral and economic considerations coming from the literature search.

In the literature search, we focused on artificial intelligence (AI), radiology, and business models, and both moral and economic considerations came forward. Radiologists increasingly rely on digital tools, including AI-based tools, for various functions [5-7]. This raises moral and economic concerns related to discussions about, for example, trust between the patient and physician [19,31,37-39], supervision and accountability in clinical decision-making [16,40,41], lacking or unclear reimbursement structures [42], and high upfront costs [43]. In these discussions, some recurring topics came forward.

Trust is often mentioned in literature as playing a critical role in the field of radiology. Where, traditionally, trust between a patient and a physician was deemed important, another dimension is added with the introduction of AI, and the definition of trust is increasingly broadened to both the clinicians' and patients' trust in the technology's performance and reliability [15,18,33,44-49]. However, trust is a complex concept that relates to interconnected discussions on, among others, fairness, safety, explainability, and transparency [50]. To increase fairness and safeguard vulnerable populations, managing risks is essential [40,51]. AI models often reflect biases inherent to insufficiently diverse training datasets [40,51,52]. Resultingly, model performance can deteriorate if applied to a different population. Explainability and transparency are fundamental to building trust in AI systems [16,53,54] and can be considered key fairness strategies to mitigate bias and improve accountability [30]. Here, the distinction is that interpretability means that a radiologist can understand and make sense of the outputs of the AI system, whereas explainability means that the AI system also provides the radiologist with a rationale or reasoning behind the results [2].

Despite their importance, many AI models suffer from the "black box problem," where decision-making processes can be opaque even to developers [30,46]. Obstruction of error identification through a lack of transparency compromises trust and accountability [30,55,56]. The General Data Protection Regulation mandates transparency and explainability in automated decision-making, but many systems struggle to meet this standard [46,57]. On the topic of data, while the General Data Protection Regulation, for instance, grants patients ownership over personal data through, for example, obliging data users to obtain informed consent, uncertainties persist regarding data reuse and consent withdrawal [40,46,55,58]. Ownership thus means that a person or legal entity has the full and inherent right to decide what happens with the data without requiring consent from anyone else. By contrast, controls refer to the delegated ability to use or manage data within boundaries agreed upon by the owner, granted through informed consent. Developers, users, and health care providers must navigate complex questions about responsible data storage and use [49]. Transparency in data management could further support trust, yet proprietary algorithms and cross-disciplinary misalignments between developers and health care practitioners create significant challenges [58]. It has therefore been advocated that radiologists must develop new skills to effectively collaborate with AI, which requires additional training [49,50,59,60] and potentially a shift in roles and responsibilities [61]. Concerns about overreliance on AI systems, loss of diagnostic expertise, and shifts in job satisfaction are frequently noted [62-64]. Additionally, unresolved issues in assigning accountability for AI-assisted errors complicate the integration into clinical workflows, raising questions about liability and autonomy among developers, users, and insurers [16,40,41].

Business models for AI applications in radiology outlined in the literature reviewed include perpetual licensing, pay-per-use, and subscription-based models [65,66], with most companies offering subscription or hybrid pricing models [65]. Deployment and pricing strategies have not yet converged to a preferred or market-dominant standard, and most vendors offer multiple options. Literature discusses the shift of software as a medical device from product-centric to platform-based models [67]. This idea of platformization could potentially facilitate easier (vendor-neutral) compatibility and integrability within existing hospital infrastructure [8,68,69]. In essence, all businesses in the medical software industry operate in a highly dynamic environment with constant challenges, including disruptive innovation, strict regulations, sustainability concerns, and uncertainties [70,71]. To navigate this landscape and mitigate risk, a more dynamic business model has been proposed, tailored to the specific demands of the software as medical device domain. Long-term stability of business models can be achieved through diversified revenue streams and optimizing operational efficiency, regardless of market conditions [71]. Hospitals also seek ways to fund AI-supported services. Some sources point to missing reimbursement structures or high upfront costs, hindering adoption [42,43]. There is, however, a need for more research on sustainable business models for hospitals that balance adoption costs with strong value propositions and reimbursement frameworks [72,73] and value assessment through real-world clinical validation [74].

The insights from this nonsystematic narrative literature search were synthesized into topics, prompts, and discussion points, which served as input for the cocreation workshop, and they were used to create workshop canvasses. The workshop canvasses can be found in Multimedia Appendices 1 and 2. These canvases guided participant engagement and helped ensure that the workshop addressed research-relevant issues by suggesting conversation topics, while remaining flexible for participant input.

Workshop Design and Participants

Since it is argued that integration of AI into radiology may provide a transformative opportunity for health care delivery, we further explored opportunities for AI in radiology in a multistakeholder workshop. Workshop participants were selected using purposive sampling to ensure the inclusion of individuals relevant to the research objectives. We aimed to include participants from diverse stakeholder backgrounds. As a point of departure, participants from an earlier workshop with a diverse panel of stakeholder backgrounds [28,61] were approached first. In addition, we explicitly looked for expertise and relevant types of stakeholders that were missing, were underrepresented, and regarded as important in relation to the workshop's themes. This search included reaching out to contacts in the authors' own network, for example, through LinkedIn and institutional websites. Prospective participants were also allowed to suggest relevant candidates from their own network. The patient representatives were members of the hospital patient council and were approached by hospital staff for participation. Participation was voluntary aside from a compensation for travel-related expenses. Potential selection bias was balanced by having participants representing very different perspectives.

In the workshop, patient representatives (n=2), radiologists (n=2), laboratory technicians (n=2), a technology developer (n=1), clinical researchers (n=2), industry employees (imaging hardware company employee: n=2; medical software company employee: n=1), a funding body employee (n=1), a clinical physicist (n=1), a business developer (n=1), and social scientists

XSL•FO
RenderX

(n=2) participated, totaling 17 participants. The sessions were moderated by the researchers (MBS, SH, and WB). Next to discussing from their expert point of view, all participants were also invited to reflect on questions from a general citizen's point of view, allowing for flexible conversations between the participants.

Ethical Considerations

The research protocol was assessed and exempted from medical ethical review by the Medical Ethics Board of University Medical Center Utrecht (Declaration METC no. 22 - 475/DB, d.d. 1 March 2022) [75]. This exemption was determined in accordance with the Dutch Medical Research Involving Human Subjects Act and the definitions and guidance provided by the Central Committee on Research Involving Human Subjects [76]. All participants gave written and verbal informed consent prior to the workshop and had the opportunity to react to a written workshop report after the workshop. In the analysis, the data were coded, meaning that participants received a study number. In this paper, we refer to the participants with that same code, although sometimes referring to their respective stakeholder background. Patient representatives received compensation for travel-related expenses and were offered a voucher of €25 (approximately US $27 at the time of the workshop).

Data Collection

The cocreation workshop took place in June 2024 and lasted 3 hours. Participants were asked in three rounds to (1) map and reflect on moral and economic considerations they considered important when thinking about the field of radiology and AI, (2) cocreate concrete scenarios, envisioning plausible futures for radiology with AI in the not too long future (year 2040), to elaborate on the broader implications of their considerations to the field, and (3) think about opportunities and potential frictions related to these scenarios and how to deal with or prevent them.

The 15-year time frame was chosen to allow participants to envision a future where significant transformation could happen, while still being close enough to make the ideas feel actionable. We asked our participants to envision scenarios *postadoption* of AI in radiology, intentionally probing the workshop participants not to focus on pragmatic topics, such as hardware requirements, because our aim was to explore strategic and value-driven perspectives rather than operational constraints. We wanted participants to think beyond immediate technical hurdles and envision how AI could reshape workflows, roles, and patient care once adoption challenges have been resolved. In 3 breakout groups, each with 5 or 6 participants, scenarios were cocreated and discussed. In round 1, the participants were asked to discuss their respective considerations in groups of 2 or 3 participants. For rounds 2 and 3, we provided the participants with a worksheet canvas to guide the discussion. The questions on the canvasses were loosely based on the topics we found in our literature search (Textbox 1). The canvasses can be found in Multimedia Appendices 1 and 2. We explicitly told participants that these sheets could be used as inspiration for conversations, and if other ideas came to mind, that they were free to explore them. No further prompts were provided, except for those related to time management of the workshop. If questions on interpretation of the canvasses came forward, the session moderators answered the question neutrally and aimed not to steer the discussion in a certain direction. In case of dissent between the participants, the session moderators did not intervene, as disagreements between the participants could potentially reveal interesting frictions between various stakeholder backgrounds. All discussions were audio recorded, the filled-out worksheet canvasses were stored, and notes were taken by the researchers (MBS, SH, and WB). A more elaborate quotebook is provided in Multimedia Appendix 3.

Data Analysis

The workshop was held in Dutch. All audio data were automatically transcribed using Amberscript (Amberscript BV), reread, checked, and corrected where necessary. Transcripts were then automatically translated to English using Microsoft Word (Microsoft Corporation), again reread, checked, and corrected where necessary. The transcripts were independently coded. MBS and SH, as native Dutch speakers, coded the original transcripts in order to stay close to the original meaning, and AK, as a native English speaker, coded the translated transcripts. The codes were subsequently compared and refined using qualitative analysis software NVivo (version 14; Lumivero). Data were analyzed using a semi-inductive coding approach, guided by predefined categories, but remaining flexible to the identification of emergent codes as new themes arose from the data. The initial codes were related to the topics we identified in the literature search and that we described in Textbox 1, namely, function, performance, supervision or autonomy, responsibility, explainability, transparency, training and job roles, algorithm performance, user and patient security, data management, accountability, bias, value assessment, business model aspects, value proposition, funding, platformization, and compatibility and integrability. Additionally, during the analysis, we found that "patient communication" was emerging as a new and recurring topic. This code was subsequently added to the codebook. Our analysis focused on moral and economic considerations, opportunities, and frictions, which materialized in the scenarios. To conform with scientific rigor, consistency between the researchers was ensured, as the researchers discussed their findings after initially independently coding the transcripts. That is, these peer debriefing sessions aided in ensuring that our interpretations were firmly grounded in the data. During these peer debriefing sessions, the researchers systematically compared their coding decisions for each transcript, focusing on whether the chosen codes accurately captured the meaning of the data and whether the interpretations were consistent across researchers. They reviewed the rationale behind each assigned code, discussed eventual discrepancies, and clarified ambiguous segments. Disagreements were openly discussed, and alternative perspectives were considered before reaching a shared interpretation, ensuring that the final coding reflected consensus. This iterative dialogue served to refine the coding framework and enhance the credibility of the analysis. Furthermore, a workshop report inspired by a preliminary analysis of the data was shared with the participants, providing them with an opportunity to disagree with the researchers' interpretation of the data or add thoughts that came after the workshop.

XSL•FO
RenderX

Results

Cocreation Workshop

In this section, we share the findings, drawing on insights from the multistakeholder cocreation workshop. Many topics came forward, including questions regarding sharing or balancing responsibilities between humans and technology, power and influence in relation to data, ownership, access and trust, and questions regarding costs and payments. Developing and reflecting on the scenarios helped identify expectations. Different stakeholders brought forward different perspectives in the discussions. For example, in groups that included a patient representative, the importance of having a human radiologist involved came forward.

Most participants shared the expectation that between now and 2040 much will change and probably quite rapidly, referring to fast developments surrounding general purpose AI tools such as ChatGPT. Participants in our multistakeholder workshop considered in their scenario either intramural improvement of radiology processes or the extramural screening and early detection of disease. This resulted in 3 scenarios sketched out in the next paragraphs: AI copilot, scanner on tour, and Alzheimer disease (AD) screening. We present each scenario following a similar structure; first, we report what participants believe could be a good use case of AI (*materialization*), after which we share their *moral and economic considerations*, *opportunities*, and *frictions*. As a way of analysis, we highlight emerging themes from the scenarios.

The 3 Scenarios

Scenario 1. Partner in Diagnosis: Envisioning AI as the Radiologist's Trusted Copilot

Materialization

Participants envisioned that in 2040 AI would be integrated into every hospital's radiology department, acting as a "copilot" to assist radiologists. Radiologists may grow from an executing role to a more supervisory role, sometimes required only to approve the AI's suggestion. In this scenario, AI can aid in the noninvasive characterization of a tumor, enriching radiological findings on computed tomography or magnetic resonance imaging (MRI) with a description of histology and tumor biology without requiring surgical procedures, such as biopsies. Or, in case a rich collection of patient data is already available, AI can integrate such multimodal data:

> *...also the integration with biomarkers, with pathology with those [foundational models], and the EHR [Electronic Health Record], gives a much more complete picture than just with those images.*" *the workshop participant continued, arguing how that despite all clinicians already work together in the same system, she expects AI to be an enabler of actual integration : "(...) I think that integration into AI of all those different things will also become very strong.* [g3, s6]

AI in this use case supports the radiologist with additional insights that can be used for improved diagnosis or treatment response monitoring. According to the participants, the added value can be demonstrated in independent clinical studies in which researchers, clinicians, and ethical evaluation boards should have their usual role. Explicating added value incentivizes medical professionals to include AI in guidelines and insurance companies to reimburse its use. AI can subsequently be used to collect relevant patient data and provide comprehensive interpretations, whereas the radiologist retains ultimate responsibility for diagnosis and treatment decisions.

Participants thought that software developers should be in the lead in setting up a quality management system. They can supply their product, for example, through a pay-per-view business model, which is integrated into Picture Archiving and Communication System or electronic health record along with other vendors' products. Users were regarded as being in the lead for evaluating real-world, postmarket applications. Although data have the potential to be part of business models, the participants thought that patients should retain ownership of their data, emphasizing that these data should be anonymous and nontraceable.

Moral and Economic Considerations

Participants deemed AI in radiology necessary to support the radiologist and underlined its potential, considering the hours spent on, for example, follow-up MRI scans without visible changes to the imaged pathology. Requirements formulated for this scenario were as follows: (1) The treating physician always remains responsible for the patient's health. Since human experts should always remain in control, physicians remain responsible for communicating with patients. (2) AI should benefit the patients and lead to improved health, while applying to as many patients as possible. (3) Trust in AI can be increased by observing well-performing models that provide added diagnostic value, which can convince radiologists to use and fully adopt these models. Trust in AI was considered a means of adoption for this technology. (4) The industry should focus on user-friendliness and insightful cost-benefit analyses to improve adoption. (5) Besides clear added value in the context of diagnosis and treatment, AI should have a positive cost-benefit balance.

Opportunities

Assistance to radiologists was expected to add value by improving efficiency and accuracy in ongoing patient care and monitoring. Furthermore, added value was expected by the premise that AI could seamlessly merge multimodal data, contributing to disease detection, characterization of lesions, and monitoring treatment. This should add value, lead to lower health care costs, improve patient outcomes, or a combination of the aforementioned. Hospitals and health insurance companies benefit due to process optimization, resulting in lower costs and value-adding health care provision.

Frictions

With AI as a copilot assisting, and sometimes even taking over from the radiologist, the chance of discrepancies arises. On the one hand, participants indicated that AI is likely to make fewer mistakes than a radiologist and that AI should be trusted. On the other hand, in case a radiologist reaches a wrong conclusion

XSL•FO
RenderX

with the suggestion of AI, the radiologist should not be allowed to hide behind the output of the AI, which creates a friction between efficient use of AI and increased responsibility in case of mistakes. As a solution, the participants suggested that the radiologist always retains the ultimate responsibility for the conclusion of the radiology report. This still requires checking the output, but was it considered faster than going through multiple series of images? For patients, it was envisioned that AI would lead to increased demand for data, which could lead to unwanted use of personal data. To prevent this, data users should receive explicit and flexible consent, meaning that data can be used only for the purpose and duration that a patient allows. Furthermore, introducing AI in radiology was expected to increase the detection of incidental findings, many of which may lack clear clinical significance. This raises complex ethical challenges, particularly when balancing a patient's right not to know with the moral obligation to share potentially medically important findings. Participants in our study emphasized the importance of discussing the right not to know with the patients prior to imaging, as incidental findings can impose a psychological burden. They also stressed the necessity of keeping a human doctor in the loop, as the moral and clinical judgment required in these situations exceeded their expected capabilities of current AI systems and the demand for expertise.

Scenario 2. Health Without Walls: AI and the Rise of (Mobile) Multimodal Screening

Materialization

The participants developed 2 separate ideas of extramural health checks: in one, they envisioned an environmentally sustainable, electric bus touring across the country, in a similar setup as the current population screening for breast cancer prevention in the Netherlands. Such a bus enables screening for a wide variety of potential diseases, which potentially benefits from large-scale screening programs. When AI detects a high-risk scan, the scan is forwarded to a radiologist, who decides whether to follow up. Within this scenario, it is possible to focus on the scan, or to combine the screening with blood or any other test, thereby combining data from multiple streams into a single risk assessment.

The other idea was focused on an extramural imaging, diagnosis, and treatment facility, which can, for instance, be located in a shopping mall. Citizens can visit the facility for a self-initiated check or after being invited due to specific risk factors, such as age. AI is subsequently used for fast diagnosis and forwards data to a radiologist if follow-up is required. As a consequence, there is still value in human intervention and responsibility. This scenario requires data to be shared with hospitals, for example, through electronic health records. A step further is to fully automate the diagnostic workflow by using large models that use large retrospective datasets and individual risk factors to screen for patients at risk and subsequently immediately treat correctly identified patients.

Moral and Economic Considerations

Participants emphasized how important they consider trust and responsibility when thinking about AI and radiology. Questions raised by a social scientist and patient representative included, "Is it responsible to leave decisions to computers?" (g1-s2) and "Who should be responsible for this?" (g1-s1). The ownership of the process and of the data was questioned. According to participants, AI can have added value in this scenario, including alleviating burdens from hospital staff, improving patients' quality of life with an earlier start of the treatment, and equal access to health care.

Moreover, participants suggested that additional value can be calculated in added quality-adjusted life years, which can be translated into societal gains. The primary benefit of these scenarios may be that early diagnosis reduces long-term health care costs by shifting care from acute to preventive services. For funders, for example, the government, of large-scale AI-supported imaging-based screening programs, this shift may represent potentially favorable cost-benefit ratios where upfront investing may yield substantial savings and productivity over time. Additionally, commercial gain is possible, for example, for companies providing the bus with diagnostic hardware and software. Participants discussed various types of business models. Software companies responsible for evidence generation and validation can charge subscription- or use-based per license. Companies supplying scanners often operate on a purchase or lease model, and decentralized centers or scanner-on-tour suppliers could cover costs via direct-to-consumer supply or health insurance. For quality, competition between companies was regarded as crucial, and hospitals could consider purchasing an intramural diagnostic platform that supports multiple products from different vendors. This creates an overarching platform with competing products. Companies developing the software suites and the AI application that can be integrated into them should be prime movers here. Participants discussed the potential for the program to eventually become institutionalized, similar to breast cancer screening. Business models were seen as increasingly promising, especially once such programs began to generate large-scale data that can be leveraged to train data-hungry models.

> *if you have an AI algorithm that is CE-marked [Conformité Européenne], it doesn't learn from the data it receives, right? That's locked down. That's what the CE marking is for. It only goes into development mode when you have data to further develop it on, and then, you do an update, and you put that into the system. So, it's not like we always think that the AI just learns by itself.* [g1, s3]

Opportunities

Participants agreed that AI could offer an opportunity for the field that is in need of change, with the number of scans being requested rapidly increasing.

> *There is a report out right now that anticipates continuous growth in the years to come. Drivers are an aging population, more chronically ill patients, but also the increase in technological possibilities. At the same time, there is a growing demand coming from patients and the general population to have more insight into their health. Extramural health checks might be an answer to this demand.* [g1, s3]

XSL•FO
RenderX

Participants envisioned a future in which AI could more and better support the role of medical professionals. Decentralized health care facilities can play a large role in the health care system and help decrease the growth of private companies offering whole-body scans, which was deemed undesirable as disparities in income and wealth may translate into unequal access to health care services, potentially leading to disparities in health outcomes.

Frictions

Participants' discussions shed light on a couple of frictions. First, they questioned whether the model should be organized in a centralized or decentralized manner, what the implications of decentralized testing would be for referring people to a hospital, and how data collected may or could be shared. Decentralized centers would need to build support within the broader network of health care institutions, hospitals, companies, and so on, for data sharing; they needed correct licensing, and interoperability must be addressed. Access and management of big data can create suspicions among patients, as they may think: "your data no longer belongs to you" [g1-s2].

Another potential friction arose around new role divisions and associated responsibilities. Laboratory technicians may outsource more tasks and perform fewer tasks themselves, radiologists may be hired on a consultancy (freelance) basis, and new roles may emerge (such as radiology assistants). This could increase the distance between the radiologist and the referring physician, which could complicate responsibilities and task distribution. A third friction point regarded sustainability:

> *What are we going to do? Is that desirable? So, I think we have to assess upfront, do you need a scan? Yes or no?* [g1-s3]

This question was answered later on:

> *AI can help with that as well. They can look in advance: from, does this patient need a scan? And then you are working more sustainably, because yes, the most sustainable scan is no scan.* [g1, s1]

On the one hand, there was a need to manage this rising demand, with patients starting to demand scans for more certainty

> *Because there are, of course, patients that ask for a scan, they [patients] want certainty.* [g1, s5]

versus the question of whether you need to better manage this demand. This can be done either by enlarging the role of AI and maximizing the use of the capacity of available scanners or by downsizing the demand by asking whether a scan is actually necessary, which is a decision AI can support in this scenario.

Scenario 3. Early AD Insight: Proactive AI for a Longer Independent Life

Materialization

A group of participants chose to build their scenario around a concrete use case. The participants quickly decided to further develop their scenario around the idea that in the future, it becomes possible to predict, by analyzing either MRI data or only multimodal data, whether a citizen later in their life is likely to develop AD. This is achieved by screening citizens from a certain age group for the disease, either upon invitation or upon the citizens' request (similar to the previous screening scenario). The benefit of using AI in very early stages to detect or predict AD, participants mentioned, can be that (future) patients are helped and trained to live independently or with minimal help for longer. This approach should improve the quality of life of patients, alleviate some of the societal burden of caregivers for these future patients, and may lead to reduced costs:

> *Allowing Alzheimer's patients to function independently for as long as possible. [...] to keep support costs as low as possible.* [g3, s3]

Explainability was discussed as an important model aspect for patients. A patient representative argued that the model should be able to explain how it arrived at a certain decision. However, a participant with a background in pathology contextualized that this can be very task dependent. The researcher gave an example that was more related to her own field of expertise:

> *because for tumor detection, it [the AI] does not need to explain why something was flagged. ...Because we check it [the decision of the AI], so it is [in this case] AI-assisted. So I think: for this task, it does not need to explain what something is flagged for [...], because we check it [the flagged location].* [g3, s6]

The participants also briefly touched upon the business landscape, indicating that having multiple vendors would be a desirable aspect:

> *It's better to have multiple suppliers. Otherwise, we will have a monopolist.* [g3, s3]

Moral and Economic Considerations

Participants found it important that implementing AI would be a meaningful addition to radiology, benefiting the individual patient. Thorough validation of the AI product can lead to improved trust in its use. Trust can also be enhanced through (human) explanation of the output. This means that either the algorithm should be able to explain its own output or radiologists should provide an explanation when a textual one is unavailable. Furthermore, patients should always remain in charge of their data and consecutive algorithmic outputs. That is, the data acquired upon clinical indication and the output of the algorithms should be shared only upon their voluntary consent. The participants also believed that individual patients or a society that shares personal data should monetarily benefit from commercial products. In other words, according to the participants, commercial parties should not be allowed to earn profits over products forever and profits should eventually flow back to society in some form, given that the AI product is essentially enabled by voluntarily provided personal data.

Opportunities

The participants argued that AI should deliver a product that radiology currently cannot. That is, true added value is achieved when AI detects what a radiologist cannot see, or when AI predicts future events that are not yet visible as pathology at the time of imaging. Participants discussed that the role of AI

XSL•FO
RenderX

in this scenario is to assist with screening. Furthermore, predicting disease before it can manifest can lead to improved quality of life, specifically for AD. The participants highlighted that this is especially the case due to the human and emotional side of a progressive disease such as AD.

Frictions

The participants discussed the need for increased scanning capacity as a potentially negative consequence of AI. Granting all citizens the opportunity for screening for early detection of AD is costly, partly due to purchasing the scanners. However, using this additional scanning capacity can lead to improved quality of life for the patients, a lower societal burden for informal caregivers, and lower health care costs if cost-effectiveness of the screening is proven. The risk of incidental findings increases, though, which needs to be balanced with increased benefits of early detection. Whereas the industry employee argued that regulations can sometimes be experienced as burdensome, the patient representative found the regulations justifiable with respect to privacy, indicating that without strict regulations the patient representative may not share personal data at all.

Finally, the participants discussed potential consequences concerning insurance and autonomy: if AI predicts a future diagnosis such as AD, an insurance company may want access to those results, given that they may want to assess and price coverage accordingly. Also, the question was raised whether, “will they [some authority] say automatically that you move to an Alzheimer’s [care] house and you have to sell your home?” [g3, s5]. According to the participants, these aspects were undesirable, and these open questions warranted some serious thought.

Emerging Themes

Seven common themes emerged from the 3 future scenarios, along with their moral and economic considerations, opportunities, and frictions. Figure 1 summarizes the 7 emerging themes, recommendations based on the emerging themes, and 3 overarching themes.

XSL•FO
RenderX

Figure 1. Summary of the workshop, 7 emerging themes, corresponding recommendations, and 3 overarching themes. AI: artificial intelligence.

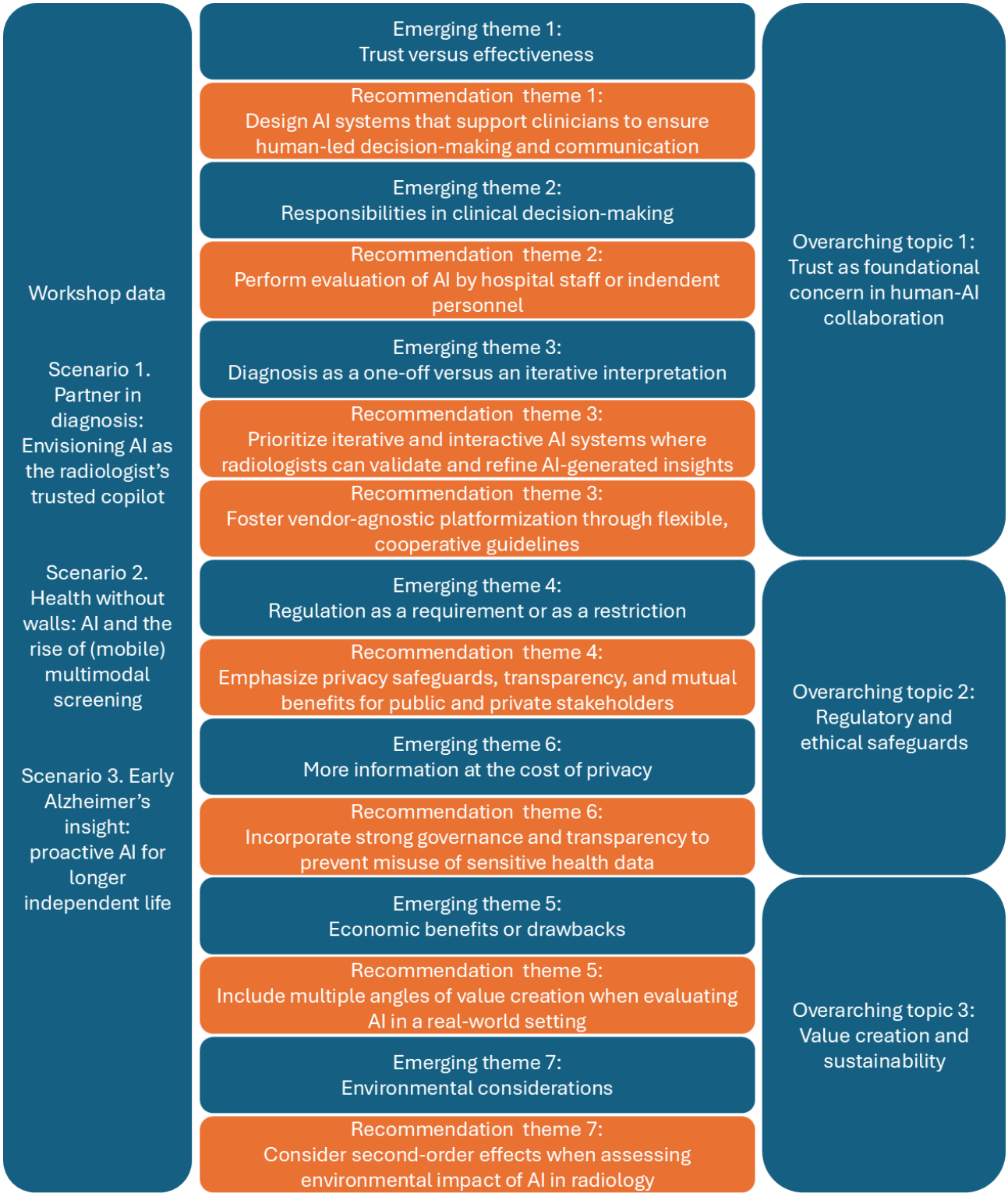

Emerging Theme 1: Trust Versus Effectiveness

Seemingly, trust in AI represents a critical trade-off. On the one hand, trust could enable AI to live up to its potential as a stand-alone technology to effectively address complex tasks, such as medical diagnosis or large-scale hospital workflow optimization. On the other hand, patient representatives feared risk for automation bias. Additionally, participants discussed that excessive reliance on AI systems can diminish human supervision, increasing the clinician's risk of automation bias and the erosion of critical decision-making capabilities. To break out of this trade-off and enhance trust, according to the participants, would often necessitate the adoption of interpretable and explainable AI systems, which radiologists can eventually check. Thus, achieving an optimal balance between trust and skepticism is essential to mitigate risks while harnessing the transformative potential of AI in radiology.

XSL•FO
RenderX

Emerging Theme 2: Responsibilities in Clinical Decision-Making

Implementing AI was considered a shared responsibility among stakeholders, each contributing distinct perspectives and expertise. Policy makers play a central role in establishing regulatory frameworks that guide legal-ethical boundaries for AI development and deployment. Industry players produce and scale practical AI applications within the established boundaries set by the aforementioned policy makers. However, when it comes to clinical decision-making, the participants foresee that medical specialists and health care workers will ultimately be responsible for the health and safety of their patients, and clinicians also feel that way. To reach this goal, it could be beneficial for medical specialists to cooperate with more technically oriented hospital personnel to perform quality assurance tests, particularly to ascertain that AI performs equally well on local datasets compared with how industry markets the performance of their products. Furthermore, according to the participants, there might be a conflict of interest when industry performs its own quality assurance. Thus, local testing may be a critical requirement for safe deployment and responsible use of AI in radiology and, according to the workshop participants, should be performed by independent hospital staff.

Emerging Theme 3: Diagnosis as a One-Off Versus an Iterative Interpretation

Given the rapid development of AI, radiological diagnosis is increasingly shaped by algorithmic output. These outputs are often treated as static and definitive, similar to finalized radiology reports. The static nature of AI's output presents 2 interrelated shortcomings. Radiological diagnosis is often not a fixed product but a dynamic process. It relies on expert interpretation and is influenced by evolving clinical information, such as neurological examinations and pathology results. Several participants noted that diagnosis often emerges through iterative reasoning and collaboration with other medical specialists. AI systems, which typically generate preanalyzed findings or diagnostic probabilities, risk oversimplifying this complexity. Participants suggested that integrating multimodal data could help AI better reflect current clinical practice. In addition, static outputs could lead to a greater demand for diagnostic certainty, potentially leading to unnecessary follow-up: by detecting more abnormalities and including findings that may never become clinically relevant, AI could lead to more scans, consultations, and interventions. A potential solution to both shortcomings that was suggested by participants is a "human-in-the-loop approach," where the radiologist remains responsible for interpreting AI outputs within the broader clinical context. Collaborative frameworks in which AI complements rather than replaces human expertise were seen as essential. These should include clear protocols for follow-up, shared decision-making with patients, and the development of transparent and explainable AI systems to support trust and appropriate clinical action.

Emerging Theme 4: Regulation as a Requirement or as a Restriction

Regulations play a dual role in technological innovation, acting both as a potential constraint and as a necessary enabler for progress, particularly in a fast-paced field such as AI in radiology. On the one hand, industry employees indicated that overly rigid or premature regulations can restrict innovation by imposing burdensome compliance requirements. On the other hand, patient representatives argued that well-designed regulatory frameworks are essential to ensure safety, fairness, and public trust—key requirements for adopting novel technologies. For example, as indicated by patient representatives, robust protection of patient privacy not only follows an ethical imperative but also creates a secure environment for patients to share their data. In other words, effective regulations are required for patients to feel safe when consenting to sharing their personal data, which is the cornerstone for all AI developments. Adaptive and dynamic regulatory approaches, capable of evolving alongside technological advances, were thus considered crucial for maintaining this balance. Following our workshop results, participants regarded effective regulation as an enabler rather than an obstacle to innovation, because it provides the foundation for trust necessary for groundbreaking technologies to thrive responsibly.

Emerging Theme 5: Economic Benefits or Drawbacks

The integration of AI in radiology presents significant economic opportunities and challenges. Following initial investments in hardware infrastructures and software packages, there is a wide variety of potential applications on a hospital level, societal level, and vendor level that could yield a net economic benefit. Participants emphasized that AI could significantly improve operational efficiency within hospitals. By automating routine tasks, such as appointment scheduling and administrative workflows, AI has the potential to reduce labor costs and streamline internal processes. These improvements may lead to better resource allocation and increased throughput, contributing to overall cost-effectiveness in clinical operations. At the societal level, the economic benefits of AI in radiology are diverse and potentially far-reaching. Participants pointed to several ways in which AI could contribute to more efficient health care delivery and improved population health. For example, timely diagnosis may enable earlier initiation of therapy, which can improve outcomes and reduce the need for prolonged or intensive treatment. Similarly, more personalized care, which could be supported by AI's ability to integrate and analyze complex data, may help avoid unnecessary interventions and enhance the overall effectiveness of health care services. These developments could contribute to long-term cost savings and a more sustainable use of health care resources. However, participants also cautioned that increased diagnostic sensitivity might lead to more incidental findings, which could raise demand for follow-up care and specialist input, potentially increasing overall health care expenditures. According to some participants, the responsibility for initial economic value assessments of AI systems lies with the vendors, who should estimate whether their solutions offer sufficient value before deployment. However, participants emphasized that independent researchers should conduct postdeployment evaluations. Similar to quality assurance processes, this separation is essential to avoid conflicts of interest and ensure unbiased assessments of economic performance.

XSL•FO
RenderX

Emerging Theme 6: More Information at the Cost of Privacy

Unintended consequences from AI might arise when the decision becomes binding rather than advisory. This decision could drastically affect a patient's life. In the AD case, the participants discussed the extreme example of mandatory compliance with the AI decision for future patients when AD is predicted. Furthermore, if (by obligation) shared with insurers, these screening programs could have negative consequences for monthly health insurance premiums, thus penalizing individual citizens financially for factors beyond their control, which raises privacy, fairness, and proportionality concerns. Or, in the extreme case of extrapolating this issue, insurance companies may increase costs for patients at risk or deny (future) patients access to (additional) health and property insurances at all. Regulation should therefore be in place to prevent such a scenario from materializing. Moreover, patients have the right to not know. The participants agreed that this moral dilemma might be quite difficult for an AI to weigh, where there needs to be a good balance between medical necessity and patients' rights and preferences. To counter this unintended consequence, human intervention might be the only solution.

Emerging Theme 7: Environmental Considerations

Finally, it was discussed that AI is an energy-demanding technology. Combining AI with radiology for screening means that scarce resources might be used for citizens who may not (yet) express any disease symptoms. Thus, the participants discussed potential benefits for health care and the economy of AI in radiology should also be carefully weighed against the potential environmental impact. The monetary gain and gain in quality of life after the introduction of AI to radiology should therefore be proportionate to the environmental impact of additional interventions, such as screening in combination with AI.

Discussion

Principal Findings

In this manuscript, we report and reflect on a qualitative cocreation workshop we designed and conducted, inspired by responsible innovation approach, aiming to explore future scenarios for AI in radiology, their opportunities and frictions, by involving a variety of stakeholders in the field of radiology. Interestingly, in imagining futures of AI in radiology, the majority of the groups, in one way or another, came up with a scenario to upscale radiology via screening or extramural imaging to early detect and prevent pathology from becoming a disease. However, there were differences in the imagined role of these applications, and of actors in the field, including governments, insurers, technology developers, and citizens. By including a wide variety of topics, we aimed to broaden the participants' scope of thinking. Resultingly, our participants discussed moral and economic considerations, envisioned opportunities, and foresaw frictions arising related to interconnected topics on trust, responsibilities, workflows, regulatory requirements, dealing with economic benefits, and unintended consequences.

Comparison With Prior Work

Trust Versus Effectiveness

In our analysis, we observed trust in AI as a complex topic the participants linked to, for example, performance and explainability. The participants perceived a paradoxical relationship between trust and effectiveness: the more trust they were willing to place in AI, the more AI could potentially operate as a stand-alone technology in radiology and fulfill its promise of improving clinical outcomes. Yet, this very trust introduces ethical tensions. As described by Singh et al [77], the ethical paradox in medical practices arises when efforts to enhance health care through AI risk diminishing the humanistic aspect of care, such as empathy and interpersonal interactions, thereby potentially provoking skepticism and decreasing trust. Reflecting on scenarios 1 and 2, AI could be deployed for triaging or identifying patients at risk. As such, AI applications may allow clinicians to focus their limited time on more complex and emotionally sensitive cases, potentially improving overall effectiveness of care and patient well-being, which were expectations that were reported earlier as well [37]. However, fully trusting AI may lead to dependency on AI or potential loss of human oversight. Similar concerns were also shared during the cocreation workshop. Patient representatives, among other stakeholders, confirmed that a human-in-the-loop was, in their opinion, still required. This idea was also shared in earlier research [63,78-80], although in other reports, patients have also indicated that autonomous AI could be acceptable on the condition that AI outperforms a radiologist [18]. Yet, patient representatives and other stakeholders seemed to have a positive attitude toward AI-assisted triaging, especially because the ultimate decision remains with a radiologist. The explainability of the AI-based results of the triaging could further help gain clinicians' confidence [53]. Discussions suggest that following these conditions, applications of AI in radiology may contribute to enhancing efficiency without compromising safety or trust. Based on the links between effectiveness, trust, explainability, and the humanistic aspect of care, we recommend designing AI systems that support clinicians, ensuring human-led decision-making and communication, while leveraging AI for effectiveness.

Responsibilities in Clinical Decision-Making

Participants in the workshop highlighted that there should not be conflicts of interest in the case of responsibilities. In other words, in environments where people must act responsibly to, for example, safeguard patients' health, one should not be incentivized to act differently. For this reason, some participants emphasized that generating evidence for AI in early stages of development and real-world testing of AI should be performed by 2 distinct parties. Whereas industry was seen as responsible for product (co)development and early evidence generation, real-world implementation, thorough (local) testing, and longitudinal evaluation should be performed by independent hospital personnel, potentially in collaboration with more technically oriented personnel such as clinical physicists or technical physicians [81]. Furthermore, there was a seeming distrust of commercial parties to perform their own prospective evaluation. Earlier research focused on *how* to perform

XSL•FO
RenderX

postinstallation evaluation [22,26,82-84]. However, this research did not dive into *who* should do so and what the possible impact on interconnected themes such as trust could be. Specifically, industry involvement should be limited in prospective evaluation, such that thorough evaluation can be performed with as little conflict of interest as possible. Our research demonstrates preliminary evidence that independent evaluation of AI in a clinical setting is pivotal to enhancing trust, and that trust can be achieved only by proper distribution of responsibilities along the innovation, installation, and evaluation chain. Therefore, we recommend that evaluation of AI is performed by hospital staff or independent external parties.

Diagnosis as a One-Off Versus an Iterative Interpretation

The proposed function of the AI within the radiology system varied within our workshop, ranging from screening to acting as a copilot for the radiologist. AI tools in medical imaging have traditionally functioned as narrow, outcome-focused instruments, for example, detecting polyps, segmenting tumors, or flagging anomalies in isolation [85]. However, radiology is fundamentally a process of evolving clinical interpretation, where patients' histories, laboratory results, physical examinations, and ongoing changes in condition contextualize each series of images [86,87]. Multimodal data integration could enable AI to move beyond static predictions and could facilitate evolving model outputs [88]. Furthermore, the inclusion of large language models (LLMs) could provide narrative understanding, enabling AI to synthesize diverse inputs into coherent clinical insights [89-91]. In turn, a radiologist can check the output of the LLM or interact with the LLM, which could lead to further explainability and could foster trust, which also links back to emerging theme 1. This convergence transforms AI from a tool that offers answers into one that participates in reasoning. Importantly, this shift could allow AI to align more closely with the clinician's cognitive workflow, adapting dynamically as new data come in [89-91]. It could also open the door to more explainable and human-centered AI systems that support, not replace, clinical judgment, as was also proposed earlier [63,78-80]. Thus, we recommend prioritizing iterative and interactive AI systems where radiologists can validate and refine AI-generated insights, ensuring that AI acts as a reasoning partner rather than a black box tool.

Moreover, the participants deemed the integrability of AI tools within existing hospital systems as a facilitator for successful adoption, as was also reported in earlier research [59]. Integrating these capabilities into unified platforms, rather than isolated tools, could thus further aid in the commercialization of AI in radiology. This process of platformization supports scalability, continuous learning, and interoperability [69], potentially enabling a shift from a one-off solution to comprehensive decision support ecosystems that are clinically and economically sustainable. Finally, participants regarded having multiple vendors as beneficial to the health care system, encouraging competition in either performance or costs, potentially benefiting patients. To support this, platforms should remain vendor-neutral, allowing easy integration from multiple providers. Whereas earlier research recommends four components: (1) (automatic) system triggering, (2) retrieval of relevant imaging series, (3) application of algorithm pipeline, and (4) availability of the results in Picture Archiving and Communication System of dedicated viewing software [92], we recommend conducting more research in the direction of cooperative guidelines that allow vendors to provide innovative solutions while remaining flexible for competition, potentially including regulatory sandbox environments, adaptive regulations, and value-based procurement.

Regulation as a Requirement or as a Restriction

In some discussions, the role of regulations came forward. While industry employees regarded regulations as a potentially impeding innovation, patient representatives regarded regulations as a requirement to share data. This paradoxical friction was also described in a recently published paper. According to Singh et al [77], the solution seems to be to strike the right balance between data accessibility and the obligation to safeguard patient confidentiality and to adhere to regulatory frameworks. This finding, previously reported among medical practitioners, was also identified in patient representatives, further confirming and reinforcing its broader relevance. The data paradox is quite complex: on the one hand, patient representatives in our workshop mostly preferred to retain ownership of their data. On the other hand, industry requires a diverse and up-to-date dataset, such that the AI products they develop generalize sufficiently well on various populations and do not suffer from data drift [40,46,51]. In light of this data paradox, future work should explore how hospital-private sector collaborations can be structured to enable responsible data sharing, with strong privacy safeguards, transparency, and mutual benefit for both public and private stakeholders. Recent global frameworks, such as the World Health Organization's regulatory pillars for AI in health [93], the EU AI Act's classification of medical imaging AI as "high-risk" [94], and international consensus initiatives such as FUTURE-AI [26], emphasize that striking this balance is not only desirable but also essential for compliance and trustworthiness. These frameworks advocate transparency, human oversight, and robust governance, which can guide collaborations toward ethical and legally sound data sharing.

Economic Benefits or Drawbacks

During the workshop, participants raised the question of whether it was desirable for companies to earn money using voluntarily provided individual patient data without patient populations benefiting. A competitive landscape was described by the participants that could solve these issues. First of all, multiple vendors compete for the best product. The participants argued that this aspect should lead to better health outcomes for individual patients, linking to earlier argued ethical justification of the use of patient data if companies contribute to public health benefits [95]. In other words, the profits earned using patient data are reinvested in the companies to improve existing products or to create new products, which benefit both public and private stakeholders. Second, vendors can compete for the lowest price, which benefits the health care system as a whole, as price competition can lead to lower health care costs. At the same time, a diverse vendor landscape may help mitigate risks associated with vendor lock-in [96], ensuring that health care

XSL•FO
RenderX

providers retain flexibility in choosing and switching between solutions as technology evolves and needs change. However, without more studies rigorously evaluating AI systems in real-world clinical settings, it is difficult to substantiate claims of added value [74]. We therefore advocate for more studies that evaluate AI in a real-world setting and that include multiple angles of value creation, such as decreased reporting times, added number of relevant incidental findings, and the prevention of unnecessary follow-up scans or interventions. This should lead to elaborated business models for both vendors (suppliers) and hospitals (buyers). Upon positive results, these studies could further facilitate AI adoption in clinics and provide vendors with useful insights about the possibility to generate revenue with their products. One example comes from Brix et al [11], who found that AI-based image reconstruction could lead to improved patient throughput, thereby obviating the need for an additional scanner. Such insights can help vendors to demonstrate the added value of their AI solutions, which may justify premium pricing and unlock new business models, especially when validated in real-world clinical environments.

More Information at the Cost of Privacy

Participants worried about data privacy, especially over the sharing of diagnostic and prognostic information inferred from AI with insurance companies, hinting that insurance companies might misuse this information. A recent review by Botha et al [97] also endeavored to examine the implication of AI tools on insurance. Whereas many aspects concerning insurance coverage were discussed, including the potential rise in insurance costs with the introduction of AI in clinical settings, second-order effects such as sharing information of future disease with insurance companies were underrepresented [97]. A perspective paper from 2018 by Char et al [60] warned of this looming issue, and our results demonstrate that these worries are still unresolved. Thus, a deeper dive into this specific topic seems warranted. Emerging literature confirms these concerns, highlighting risks of insurance-related discrimination from predictive analytics [98]. Regulatory frameworks such as World Health Organization's guidance [93] and the EU AI Act [94] stress the need for strong governance and transparency to prevent misuse of sensitive health data. Incorporating these principles into practice could help mitigate ethical risks while enabling responsible innovation.

Environmental Considerations

Another worry that came forward was the energy consumption of AI, which according to the participants potentially makes it an environmentally unsustainable solution. Whereas it is likely a correct assumption that adding AI to the radiology workflow increases energy consumption compared with a workflow without AI, efforts to make both algorithms and hardware more energy-efficient might help reduce the energy consumption for each inference [99]. Furthermore, AI could aid in reducing environmentally harmful gadolinium-based contrast agent use [100], or can reconstruct images from accelerated (and thus less time-consuming and less energy-consuming) MRI acquisitions [11,101], which both potentially have a net positive effect on the environment, as it may result in fewer MRI systems in use and less energy consumption per scan. Thus, whereas the participants' worries are justifiable, these second-order effects also need to be accounted for when assessing environmental impact for AI in radiology.

Limitations and Future Directions

The main limitation of this study is that we draw on a singular empirical case study set in a Dutch tertiary care center. Future studies should revisit our research questions in different countries and look at other care settings, including trans- and extramural settings. These other care centers are of special importance, since multiple participants independently hinted at the potential for extramural imaging facilities. Also, no insurance company employees could attend the workshop, despite efforts to include at least one. Insurance companies are a key stakeholder in the field, so their perspectives are potentially valuable in discussions on the demand side (reimbursement and assessment of added therapeutic value). Next, in future research, we recommend including patient representatives again, as their insights were invaluable to this work. Finally, by checking the audio recordings in full by 2 of the authors (MBS and SH) and refining transcripts and translations, we aimed to minimize potential nuance loss. Although some loss of data could still have occurred, these steps helped preserve the richness of the participants' contributions as much as possible.

Conclusions

With this paper, we aimed to explore wider stakeholder perspectives on futures of AI in radiology and to discuss the moral and economic considerations, opportunities, and frictions. The interaction between participants resulted in scenarios about an AI copilot, AI-enhanced extramural scanning, and AI-based AD screening. It prompted debates on perspectives that may not have emerged from isolated stakeholder analyses. Seven themes emerged from the analysis: (1) trust and efficiency of AI technologies, (2) responsibilities in clinical decision-making when AI is involved, (3) diagnosis as a one-off versus an iterative process, (4) regulations as a requirement or a restriction, (5) economic benefits or drawbacks, (6) trade-off between amount of information required and patient privacy, and (7) environmental considerations. In reflecting on the 7 emerging themes from the multistakeholder cocreation workshop, 3 overarching topics can be identified that help reflect on the broader implications of AI in radiology.

First, trust emerged as a foundational concern in human-AI collaboration. Participants emphasized that in their opinion, AI should support, not replace, clinical decision-making. Trust was seen as conditional for AI to be applied in radiology efficiently, meaning that explainability and the continued presence of a human-in-the-loop are deemed important. This human-machine model aligns with radiological diagnoses of iterative nature and was considered essential for safe and ethical implementation. Second, regulatory and ethical safeguards were central to discussions on data use, responsibility, and privacy. Stakeholders expressed diverging views: while industry employees saw regulation as a potential restriction, patient representatives viewed regulations as a requirement for data sharing. The friction between innovation and regulation can be dealt with through adaptive regulatory frameworks that evolve alongside

XSL•FO
RenderX

technological capabilities while safeguarding patient autonomy and algorithmic fairness. A third topic relates to value creation and sustainability. Participants discussed the potential for AI to improve efficiency, reduce costs, and enhance diagnostic accuracy. However, concerns were raised about incidental findings, data monetization, viability of business models, and the energy demand of AI systems. Platformization and vendor-neutral integration were seen as a way to foster innovation while maintaining affordability and scalability. Moreover, participants advocated for competitive ecosystems where multiple vendors could contribute to better outcomes and lower health care costs.

Together, these overarching topics offer a lens through which future research and policy can address the complex interplay of technological potential, ethical responsibility, and systemic sustainability in AI-driven radiology. Future research is needed to investigate the generalizability of our findings in different countries and settings, or to address open questions on real-world economic benefit and sustainability.

Acknowledgments

The authors would like to thank their participants for their time and invaluable insights. Without them, their study would not have been possible.

Funding

This publication is part of the project "Responsible implementation of quantitative MRI" (with project number 18749 of the research program HTSM MVI top-up, which is financed by the Dutch Research Council [NWO]).

Data Availability

The datasets analyzed during this study are not publicly available due to data policies. They

may be available from the corresponding author (MBS) per reasonable request.

Authors' Contributions

MBS, SH, AK, and WB contributed to methodology design, formal data analysis, and draft of original manuscript. All authors participated in study conceptualization, review, and revision of manuscript.

Conflicts of Interest

None declared

Multimedia Appendix 1
Worksheet canvas for exercise 2. The original worksheet during the workshop was in Dutch.
[PNG File, 199 KB - jmir_v28i1e83407_app1.png]

Multimedia Appendix 2
Worksheet canvas for exercise 3. Mind that, again, the original worksheet during the workshop was in Dutch.
[PNG File, 105 KB - jmir_v28i1e83407_app2.png]

Multimedia Appendix 3
Quotebook.
[DOCX File, 19 KB - jmir_v28i1e83407_app3.docx]

References

1. Hassankhani A, Amoukhteh M, Valizadeh P, Jannatdoust P, Sabeghi P, Gholamrezanezhad A. Radiology as a specialty in the era of artificial intelligence: a systematic review and meta-analysis on medical students, radiology trainees, and radiologists. Acad Radiol 2024 Jan;31(1):306-321. [doi: 10.1016/j.acra.2023.05.024]
2. Goisauf M, Cano Abadía M. Ethics of AI in radiology: a review of ethical and societal implications. Front Big Data 2022;5:850383. [doi: 10.3389/fdata.2022.850383] [Medline: 35910490]
3. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018 Aug;18(8):500-510. [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
4. Mazurowski MA. Artificial intelligence may cause a significant disruption to the radiology workforce. J Am Coll Radiol 2019 Aug;16(8):1077-1082. [doi: 10.1016/j.jacr.2019.01.026] [Medline: 30975611]
5. Zhao T, Gu Y, Yang J, et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. Nat Methods 2025 Jan;22(1):166-176. [doi: 10.1038/s41592-024-02499-w]

XSL•FO
RenderX

6. Rezazade Mehrizi MH, van Ooijen P, Homan M. Applications of artificial intelligence (AI) in diagnostic radiology: a technography study. Eur Radiol 2021 Apr;31(4):1805-1811. [doi: 10.1007/s00330-020-07230-9] [Medline: 32945967]
7. Drogt J, Milota M, Veldhuis W, Vos S, Jongsma K. The promise of AI for image-driven medicine: qualitative interview study of radiologists' and pathologists' perspectives. JMIR Hum Factors 2024 Nov 21;11:e52514. [doi: 10.2196/52514] [Medline: 39570627]
8. Najjar R. Redefining radiology: a review of artificial intelligence integration in medical imaging. Diagnostics (Basel) 2023 Aug 25;13(17):2760. [doi: 10.3390/diagnostics13172760] [Medline: 37685300]
9. Fasterholdt I, Naghavi-Behzad M, Rasmussen BSB, et al. Value assessment of artificial intelligence in medical imaging: a scoping review. BMC Med Imaging 2022 Oct 31;22(1):187. [doi: 10.1186/s12880-022-00918-y] [Medline: 36316665]
10. Hofmann P, Oesterle S, Rust P, Urbach N. Machine learning approaches along the radiology value chain—rethinking value propositions. Presented at: Proceedings of the 27th European Conference on Information Systems (ECIS) Stockholm & Uppsala; Jun 8-14, 2019 URL: https://aisel.aisnet.org/ecis2019_rp/158/ [accessed 2026-02-09]
11. Brix MAK, Järvinen J, Bode MK, et al. Financial impact of incorporating deep learning reconstruction into magnetic resonance imaging routine. Eur J Radiol 2024 Jun;175:111434. [doi: 10.1016/j.ejrad.2024.111434] [Medline: 38520806]
12. Bharadwaj P, Nicola L, Breau-Brunel M, et al. Unlocking the value: quantifying the return on investment of hospital artificial intelligence. J Am Coll Radiol 2024 Oct;21(10):1677-1685. [doi: 10.1016/j.jacr.2024.02.034] [Medline: 38499053]
13. Sharon T. When digital health meets digital capitalism, how many common goods are at stake? Big Data Soc 2018 Jul;5(2). [doi: 10.1177/2053951718819032]
14. Raz A, Minari J, Schicktanz S, Sharon T, Werner-Felmayer G. Editorial: data-intensive medicine and healthcare: ethical and social implications in the era of artificial intelligence and automated decision-making. Front Genet 2023 Sep 7;14. [doi: 10.3389/fgene.2023.1280344]
15. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. J Med Internet Res 2020 Jun 19;22(6):e15154. [doi: 10.2196/15154] [Medline: 32558657]
16. Shevtsova D, Ahmed A, Boot IWA, et al. Trust in and acceptance of artificial intelligence applications in medicine: mixed methods study. JMIR Hum Factors 2024 Jan 17;11:e47031. [doi: 10.2196/47031] [Medline: 38231544]
17. Rivard L, Lehoux P. When desirability and feasibility go hand in hand: innovators' perspectives on what is and is not responsible innovation in health. Journal of Responsible Innovation 2020 Jan 2;7(1):76-95. [doi: 10.1080/23299460.2019.1622952]
18. Fransen SJ, Kwee TC, Rouw D, et al. Patient perspectives on the use of artificial intelligence in prostate cancer diagnosis on MRI. Eur Radiol 2024 Aug 14;35(2):769-775. [doi: 10.1007/s00330-024-11012-y]
19. Hajiheydari N, Delgosha MS, Saheb T. AI in medical diagnosis: a contextualised study of patient motivations and concerns. Soc Sci Med 2025 Apr;371:117850. [doi: 10.1016/j.socscimed.2025.117850] [Medline: 40081168]
20. Ongena YP, Haan M, Yakar D, Kwee TC. Patients' views on the implementation of artificial intelligence in radiology: development and validation of a standardized questionnaire. Eur Radiol 2020 Feb;30(2):1033-1040. [doi: 10.1007/s00330-019-06486-0] [Medline: 31705254]
21. Becker CD, Kotter E, Fournier L, Martí-Bonmatí L, European Society of Radiology (ESR). Current practical experience with artificial intelligence in clinical radiology: a survey of the European Society of Radiology. Insights Imaging 2022 Dec;13(1):107. [doi: 10.1186/s13244-022-01247-y]
22. Brady AP, Allen B, Chong J, et al. Developing, purchasing, implementing and monitoring AI tools in radiology: practical considerations. A multi-society statement from the ACR, CAR, ESR, RANZCR & RSNA. Insights Imaging 2024;15(1):16. [doi: 10.1186/s13244-023-01541-3]
23. Owen R, Stilgoe J, Macnaghten P, Gorman M, Fisher E, Guston D. A framework for responsible innovation. In: Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society: Wiley; 2013:27-50. [doi: 10.1002/9781118551424]
24. Stilgoe J, Owen R, Macnaghten P. Developing a framework for responsible innovation. Res Policy 2013 Nov;42(9):1568-1580. [doi: 10.1016/j.respol.2013.05.008]
25. Vo V, Chen G, Aquino YSJ, Carter SM, Do QN, Woode ME. Multi-stakeholder preferences for the use of artificial intelligence in healthcare: A systematic review and thematic analysis. Soc Sci Med 2023 Dec;338:116357. [doi: 10.1016/j.socscimed.2023.116357] [Medline: 37949020]
26. Lekadir K, Frangi AF, Porras AR, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ 2025 Feb 5;388:e081554. [doi: 10.1136/bmj-2024-081554] [Medline: 39909534]
27. Kim JP, Ryan K, Kasun M, Hogg J, Dunn LB, Roberts LW. Physicians' and machine learning researchers' perspectives on ethical issues in the early development of clinical machine learning tools: qualitative interview study. JMIR AI 2023 Oct 30;2:e47449. [doi: 10.2196/47449] [Medline: 38875536]
28. van Hees SV, Schilder MB, Keyser A, Sbrizzi A, Kleinloog JPD, Boon WPC. Exploring scenarios for implementing fast quantitative MRI. Eur J Radiol Open 2025 Jun;14:100658. [doi: 10.1016/j.ejro.2025.100658] [Medline: 40469716]
29. Strohm L, Hehakaya C, Ranschaert ER, Boon WPC, Moors EHM. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. Eur Radiol 2020 Oct;30(10):5525-5532. [doi: 10.1007/s00330-020-06946-y] [Medline: 32458173]

XSL•FO
RenderX

30. Haque AB, Islam A, Mikalef P. Explainable Artificial Intelligence (XAI) from a user perspective: a synthesis of prior literature and problematizing avenues for future research. Technol Forecast Soc Change 2023 Jan;186:122120. [doi: 10.1016/j.techfore.2022.122120]
31. Yang L, Ene IC, Arabi Belaghi R, Koff D, Stein N, Santaguida P. Stakeholders' perspectives on the future of artificial intelligence in radiology: a scoping review. Eur Radiol 2022 Mar;32(3):1477-1495. [doi: 10.1007/s00330-021-08214-z]
32. Lukkien DRM, Ipakchian Askari S, Stolwijk NE, et al. Making co-design more responsible: case study on the development of an AI-based decision support system in dementia care. JMIR Hum Factors 2024 Jul 31;11:e55961. [doi: 10.2196/55961]
33. Zondag AGM, Rozestraten R, Grimmelikhuijsen SG, et al. The effect of artificial intelligence on patient-physician trust: cross-sectional vignette study. J Med Internet Res 2024 May 28;26:e50853. [doi: 10.2196/50853]
34. Pacifico Silva H, Lehoux P, Miller FA, Denis JL. Introducing responsible innovation in health: a policy-oriented framework. Health Res Policy Syst 2018 Sep 10;16(1):90. [doi: 10.1186/s12961-018-0362-5] [Medline: 30200985]
35. Greenhalgh T, Thorne S, Malterud K. Time to challenge the spurious hierarchy of systematic over narrative reviews? Eur J Clin Investigation 2018 Jun;48(6). [doi: 10.1111/eci.12931]
36. Sukhera J. Narrative reviews: flexible, rigorous, and practical. J Grad Med Educ 2022 Aug 1;14(4):414-417. [doi: 10.4300/JGME-D-22-00480.1]
37. Gundlack J, Negash S, Thiel C, et al. Artificial intelligence in medical care—patients' perceptions on caregiving relationships and ethics: a qualitative study. Health Expect 2025 Apr;28(2):e70216. [doi: 10.1111/hex.70216] [Medline: 40094179]
38. Baghdadi LR, Mobeirek AA, Alhudaithi DR, et al. Patients' attitudes toward the use of artificial intelligence as a diagnostic tool in radiology in Saudi Arabia: cross-sectional study. JMIR Hum Factors 2024;11:e53108. [doi: 10.2196/53108]
39. Haan M, Ongena YP, Hommes S, Kwee TC, Yakar D. A qualitative study to understand patient perspective on the use of artificial intelligence in radiology. J Am Coll Radiol 2019 Oct;16(10):1416-1419. [doi: 10.1016/j.jacr.2018.12.043] [Medline: 30878311]
40. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan;28(1):31-38. [doi: 10.1038/s41591-021-01614-0]
41. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018 Oct;2(10):719-731. [doi: 10.1038/s41551-018-0305-z] [Medline: 31015651]
42. Trivedi H. The business of artificial intelligence in radiology has little to do with radiologists. J Am Coll Radiol 2022 Apr;19(4):564-566. [doi: 10.1016/j.jacr.2022.01.006] [Medline: 35176248]
43. Mayo RC, Chang Sen LQ, Leung JWT. Financing artificial intelligence in medical imaging: show me the money. J Am Coll Radiol 2020 Jan;17(1 Pt B):175-177. [doi: 10.1016/j.jacr.2019.07.004] [Medline: 31918877]
44. Gille F, Jobin A, Ienca M. What we talk about when we talk about trust: theory of trust for AI in healthcare. Intelligence-Based Medicine 2020 Nov;1-2:100001. [doi: 10.1016/j.ibmed.2020.100001]
45. Steerling E, Siira E, Nilsen P, Svedberg P, Nygren J. Implementing AI in healthcare-the relevance of trust: a scoping review. Front Health Serv 2023;3:1211150. [doi: 10.3389/frhs.2023.1211150] [Medline: 37693234]
46. Fernandez-Quilez A. Deep learning in radiology: ethics of data and on the value of algorithm transparency, interpretability and explainability. AI Ethics 2023 Feb;3(1):257-265. [doi: 10.1007/s43681-022-00161-9]
47. Daly SJ, Wiewiora A, Hearn G. Shifting attitudes and trust in AI: influences on organizational AI adoption. Technol Forecast Soc Change 2025 Jun;215:124108. [doi: 10.1016/j.techfore.2025.124108]
48. Tahtali MA, Snijders CCP, Dirne C, Le Blanc PM. Prioritizing trust in podiatrists' preference for AI in supportive roles over diagnostic roles in health care: qualitative interview and focus group study. JMIR Hum Factors 2025 Feb 21;12:e59010. [doi: 10.2196/59010] [Medline: 39983118]
49. Aldhafeeri FM. Navigating the ethical landscape of artificial intelligence in radiography: a cross-sectional study of radiographers' perspectives. BMC Med Ethics 2024 May 11;25(1):52. [doi: 10.1186/s12910-024-01052-w] [Medline: 38734602]
50. Linguraru MG, Bakas S, Aboian M, et al. Clinical, cultural, computational, and regulatory considerations to deploy AI in radiology: perspectives of RSNA and MICCAI experts. Radiol Artif Intell 2024 Jul;6(4):e240225. [doi: 10.1148/ryai.240225] [Medline: 38984986]
51. Chen RJ, Wang JJ, Williamson DFK, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng 2023 Jun;7(6):719-742. [doi: 10.1038/s41551-023-01056-8] [Medline: 37380750]
52. Tejani AS, Ng YS, Xi Y, Rayan JC. Understanding and mitigating bias in imaging artificial intelligence. Radiographics 2024 May;44(5):e230067. [doi: 10.1148/rg.230067] [Medline: 38635456]
53. Din M, Daga K, Saoud J, et al. Clinicians' perspectives on the use of artificial intelligence to triage MRI brain scans. Eur J Radiol 2025 Feb;183:111921. [doi: 10.1016/j.ejrad.2025.111921] [Medline: 39805194]
54. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310. [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
55. Siala H, Wang Y. SHIFTing artificial intelligence to be responsible in healthcare: a systematic review. Soc Sci Med 2022 Mar;296:114782. [doi: 10.1016/j.socscimed.2022.114782] [Medline: 35152047]

XSL•FO
RenderX

56. Martiniussen MA, Larsen M, Larsen ASF, et al. Norwegian radiologists' expectations of artificial intelligence in mammographic screening—a cross-sectional survey. Eur J Radiol 2023 Oct;167:111061. [doi: 10.1016/j.ejrad.2023.111061] [Medline: 37657381]
57. Fehr J, Citro B, Malpani R, Lippert C, Madai VI. A trustworthy AI reality-check: the lack of transparency of artificial intelligence products in healthcare. Front Digit Health 2024 Feb 20;6. [doi: 10.3389/fdgth.2024.1267290]
58. Martin C, DeStefano K, Haran H, et al. The ethical considerations including inclusion and biases, data protection, and proper implementation among AI in radiology and potential implications. Intelligence-Based Medicine 2022;6:100073. [doi: 10.1016/j.ibmed.2022.100073]
59. Abell B, Naicker S, Rodwell D, et al. Identifying barriers and facilitators to successful implementation of computerized clinical decision support systems in hospitals: a NASSS framework-informed scoping review. Implementation Sci 2023 Jul 26;18(1):32. [doi: 10.1186/s13012-023-01287-y]
60. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. N Engl J Med 2018 Mar 15;378(11):981-983. [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
61. van Hees S, Kleinloog JPD, Sbrizzi A, Boon WPC. Static Future Technologies, Dynamic Professionalism — Co-creating Future Scenarios in Medical Imaging Practices. Postdigit Sci Educ 2024 Mar;6(1):135-153. [doi: 10.1007/s42438-023-00444-2]
62. Lombi L, Rossero E. How artificial intelligence is reshaping the autonomy and boundary work of radiologists. A qualitative study. Sociol Health Illn 2024 Feb;46(2):200-218. [doi: 10.1111/1467-9566.13702] [Medline: 37573551]
63. Scheek D, Rezazade Mehrizi M, Ranschaert E. Radiologists in the loop: the roles of radiologists in the development of AI applications. Eur Radiol 2021 Oct;31(10):7960-7968. [doi: 10.1007/s00330-021-07879-w]
64. Stogiannos N, Walsh G, Ohene-Botwe B, et al. R-AI-diographers: a European survey on perceived impact of AI on professional identity, careers, and radiographers' roles. Insights Imaging 2025 Feb 17;16(1):43. [doi: 10.1186/s13244-025-01918-6] [Medline: 39962024]
65. van Leeuwen KG, Schalekamp S, Rutten M, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol 2021 Jun;31(6):3797-3804. [doi: 10.1007/s00330-021-07892-z] [Medline: 33856519]
66. Parikh RB, Helmchen LA. Paying for artificial intelligence in medicine. NPJ Digit Med 2022 May 20;5(1):63. [doi: 10.1038/s41746-022-00609-6] [Medline: 35595986]
67. Yu J, Zhang J, Sengoku S. Innovation process and industrial system of US Food and Drug Administration-approved software as a medical device: review and content analysis. J Med Internet Res 2023 Nov 24;25:e47505. [doi: 10.2196/47505] [Medline: 37999948]
68. Joshi H. Enabling Next-Gen Healthcare: advanced interoperability and integration with AI, IoMT, and precision medicine. Int Adv Res J Sci Eng Technol 2024 Aug 20;8(1):89-95 [FREE Full text] [doi: 10.17148/IARJSET.2021.8116]
69. Theriault-Lauzier P, Cobin D, Tastet O, et al. A responsible framework for applying artificial intelligence on medical images and signals at the point of care: the PACS-AI platform. Can J Cardiol 2024 Oct;40(10):1828-1840. [doi: 10.1016/j.cjca.2024.05.025] [Medline: 38885787]
70. Lehoux P, Daudelin G, Williams-Jones B, Denis JL, Longo C. How do business model and health technology design influence each other? Insights from a longitudinal case study of three academic spin-offs. Res Policy 2014 Jul;43(6):1025-1038. [doi: 10.1016/j.respol.2014.02.001]
71. Javanmardi E, Maresova P, Xie N, Mierzwiak R. Exploring business models for managing uncertainty in healthcare, medical devices, and biotechnology industries. Heliyon 2024 Feb 29;10(4):e25962. [doi: 10.1016/j.heliyon.2024.e25962] [Medline: 38370194]
72. van Leeuwen KG, de Rooij M, Schalekamp S, van Ginneken B, Rutten MJCM. How does artificial intelligence in radiology improve efficiency and health outcomes? Pediatr Radiol 2022 Oct;52(11):2087-2093. [doi: 10.1007/s00247-021-05114-8] [Medline: 34117522]
73. Lobig F, Subramanian D, Blankenburg M, Sharma A, Variyar A, Butler O. To pay or not to pay for artificial intelligence applications in radiology. NPJ Digit Med 2023 Jun 23;6(1):117. [doi: 10.1038/s41746-023-00861-4] [Medline: 37353531]
74. Alami H, Lehoux P, Denis JL, et al. Organizational readiness for artificial intelligence in health care: insights for decision-making and practice. J Health Organ Manag 2020 Dec 3;35(1):106-114. [doi: 10.1108/JHOM-03-2020-0074] [Medline: 33258359]
75. Medical research involving human subjects act (WMO). CCMO. URL: https://english.ccmo.nl/investigators/legal-framework-for-medical-scientific-research/laws/medical-research-involving-human-subjects-act-wmo [accessed 2024-11-16]
76. Guide to the review procedure. CCMO. URL: https://english.ccmo.nl/investigators/guide-to-the-review-procedure [accessed 2024-11-16]
77. Singh N, Jain M, Kamal MM, Bodhi R, Gupta B. Technological paradoxes and artificial intelligence implementation in healthcare. An application of paradox theory. Technol Forecast Soc Change 2024 Jan;198:122967. [doi: 10.1016/j.techfore.2023.122967]

XSL•FO
RenderX

78. Liew C. The future of radiology augmented with artificial intelligence: a strategy for success. Eur J Radiol 2018 May;102:152-156. [doi: 10.1016/j.ejrad.2018.03.019] [Medline: 29685530]
79. Ongena YP, Yakar D, Haan M, Kwee TC. Artificial intelligence in screening mammography: a population survey of women's preferences. J Am Coll Radiol 2021 Jan;18(1 Pt A):79-86. [doi: 10.1016/j.jacr.2020.09.042] [Medline: 33058789]
80. Lennartz S, Dratsch T, Zopfs D, et al. Use and control of artificial intelligence in patients across the medical workflow: single-center questionnaire study of patient perspectives. J Med Internet Res 2021 Feb 17;23(2):e24221. [doi: 10.2196/24221] [Medline: 33595451]
81. Groenier M, Spijkerboer K, Venix L, et al. Evaluation of the impact of technical physicians on improving individual patient care with technology. BMC Med Educ 2023 Mar 23;23(1):181. [doi: 10.1186/s12909-023-04137-z] [Medline: 36959581]
82. Hadjiiski L, Cha K, Chan HP, et al. AAPM task group report 273: recommendations on best practices for AI and machine learning for computer‐aided diagnosis in medical imaging. Med Phys 2023 Feb;50(2):e1-e24. [doi: 10.1002/mp.16188] [Medline: 36565447]
83. Geis JR, Brady AP, Wu CC, et al. Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. Radiology 2019 Nov;293(2):436-440. [doi: 10.1148/radiol.2019191586] [Medline: 31573399]
84. Mahmood U, Shukla-Dave A, Chan HP, et al. Artificial intelligence in medicine: mitigating risks and maximizing benefits via quality assurance, quality control, and acceptance testing. BJR Artif Intell 2024 Jan;1(1):ubae003. [doi: 10.1093/bjrai/ubae003] [Medline: 38476957]
85. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. Radiographics 2017;37(2):505-515. [doi: 10.1148/rg.2017160130] [Medline: 28212054]
86. Bosmans JML, Weyler JJ, De Schepper AM, Parizel PM. The radiology report as seen by radiologists and referring clinicians: results of the COVER and ROVER surveys. Radiology 2011 Apr;259(1):184-195. [doi: 10.1148/radiol.10101045] [Medline: 21224423]
87. European Society of Radiology (ESR). Good practice for radiological reporting. Guidelines from the European Society of Radiology (ESR). Insights Imaging 2011 Apr;2(2):93-96. [doi: 10.1007/s13244-011-0066-7]
88. Xu X, Li J, Zhu Z, et al. A comprehensive review on synergy of multi-modal data and AI technologies in medical diagnosis. Bioengineering (Basel) 2024 Feb 25;11(3):219. [doi: 10.3390/bioengineering11030219] [Medline: 38534493]
89. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? Patterns (N Y) 2024 Mar 8;5(3):100943. [doi: 10.1016/j.patter.2024.100943] [Medline: 38487804]
90. Li CY, Chang KJ, Yang CF, et al. Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation. Nat Commun 2025 Mar 6;16(1):2258. [doi: 10.1038/s41467-025-57426-0] [Medline: 40050277]
91. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024 Apr;30(4):1134-1142. [doi: 10.1038/s41591-024-02855-5] [Medline: 38413730]
92. Leiner T, Bennink E, Mol CP, Kuijf HJ, Veldhuis WB. Bringing AI to the clinic: blueprint for a vendor-neutral AI deployment infrastructure. Insights Imaging 2021 Feb 2;12(1):11. [doi: 10.1186/s13244-020-00931-1] [Medline: 33528677]
93. WHO outlines considerations for regulation of artificial intelligence for health. World Health Organization. 2023. URL: https://www.who.int/news/item/19-10-2023-who-outlines-considerations-for-regulation-of-artificial-intelligence-for-health [accessed 2025-11-29]
94. Kotter E, D'Antonoli TA, Cuocolo R, et al. Guiding AI in radiology: ESR's recommendations for effective implementation of the European AI Act. Insights Imaging 2025;16(1). [doi: 10.1186/s13244-025-01905-x]
95. Winkler EC, Jungkunz M, Thorogood A, Lotz V, Schickhardt C. Patient data for commercial companies? An ethical framework for sharing patients' data with for-profit companies for research. J Med Ethics 2025 May;51(5):jme-2022. [doi: 10.1136/jme-2022-108781]
96. Opara-Martins J, Sahandi R, Tian F. Critical analysis of vendor lock-in and its impact on cloud computing migration: a business perspective. J Cloud Comp 2016 Dec;5(1):4. [doi: 10.1186/s13677-016-0054-z]
97. Botha NN, Segbedzi CE, Dumahasi VK, et al. Artificial intelligence in healthcare: a scoping review of perceived threats to patient rights and safety. Arch Public Health 2024 Oct 23;82(1):188. [doi: 10.1186/s13690-024-01414-1] [Medline: 39444019]
98. van Bekkum M, Zuiderveen Borgesius F, Heskes T. AI, insurance, discrimination and unfair differentiation: an overview and research agenda. Law Innov Technol 2025 Jan 2;17(1):177-204. [doi: 10.1080/17579961.2025.2469348]
99. Desislavov R, Martínez-Plumed F, Hernández-Orallo J. Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. Sustainable Computing: Informatics and Systems 2023 Apr;38:100857. [doi: 10.1016/j.suscom.2023.100857]
100. Gong E, Pauly JM, Wintermark M, Zaharchuk G. Deep learning enables reduced gadolinium dose for contrast‐enhanced brain MRI. Magn Reson Imaging 2018 Aug;48(2):330-340. [doi: 10.1002/jmri.25970]
101. Lin DJ, Johnson PM, Knoll F, Lui YW. Artificial intelligence for MR image reconstruction: an overview for clinicians. Magn Reson Imaging 2021 Apr;53(4):1015-1028. [doi: 10.1002/jmri.27078]

XSL•FO
RenderX

Abbreviations

AD: Alzheimer disease
AI: artificial intelligence
LLM: large language model
MRI: magnetic resonance imaging

Edited by A Stone; submitted 02.Sep.2025; peer-reviewed by J Zhang, P Shingru; accepted 31.Dec.2025; published 25.Feb.2026.

Please cite as:
Schilder MB, Keyser A, van Hees S, Sbrizzi A, Boon WPC
Anticipating Moral and Economic Considerations, Opportunities, and Potential Frictions for AI in Medical Imaging: Multistakeholder Cocreation Study
J Med Internet Res 2026;28:e83407
URL: https://www.jmir.org/2026/1/e83407
doi:10.2196/83407

© Martin Bastiaan Schilder, Alexandra Keyser, Susan van Hees, Alessandro Sbrizzi, Wouter Pieter Christiaan Boon. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Forecasting Waitlist Trajectories for Patients With Metabolic Dysfunction–Associated Steatohepatitis Cirrhosis: A Neural Network Competing Risk Analysis

Gopika Punchhi[1,2*], MD; Yingji Sun[2*], MSc; Eunice Tan[2,3,4], MD; Naomi Khaing Than Hlaing[2], MD; Chang Liu[5], BMATH; Sumeet Asrani[6], MD; Sirisha Rambhatla[7], PhD; Mamatha Bhat[2,8], MSc, MD, PhD

[1]Schulich School of Medicine and Dentistry, Western University, London, ON, Canada

[2]Ajmera Transplant Centre, University Health Network, 585 University Ave, Toronto, ON, Canada

[3]Division of Gastroenterology, Department of Medicine, National University Hospital, Singapore, Singapore

[4]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[5]Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada

[6]Division of Hepatology, Baylor University Medical Center, Dallas, TX, United States

[7]Department of Management Sciences, University of Waterloo, Waterloo, ON, Canada

[8]Division of Gastroenterology and Hepatology, University of Toronto, Toronto, ON, Canada

[*]these authors contributed equally

Corresponding Author:
Mamatha Bhat, MSc, MD, PhD
Ajmera Transplant Centre, University Health Network, 585 University Ave, Toronto, ON, Canada

Abstract

Background: Metabolic dysfunction–associated steatohepatitis (MASH) cirrhosis is a leading indication for liver transplantation (LT). Patients with MASH cirrhosis are complex and often have extensive comorbidities. The current model for end-stage liver disease (MELD)–based liver allocation system has suboptimal concordance in predicting waitlist mortality for patients with MASH cirrhosis. Furthermore, it does not capture the competing outcomes of death and LT on the liver transplant waitlist.

Objective: A competing risk analysis using deep learning was conducted to forecast waitlist trajectories of patients with MASH cirrhosis using data available at the time of waitlisting.

Methods: A deep learning competing risk model was constructed using data from 17,551 waitlisted patients with MASH cirrhosis in the Scientific Registry of Transplant Recipients (SRTR) based on the DeepHit model framework with five-fold cross-validation. Model performance was evaluated and compared to single-risk Cox proportional hazards and random survival forests (RSF) models in predicting death or transplant using the concordance index and Brier score. Additionally, a novel performance metric, the competing event coherence (CEC) score, was developed to evaluate model performance in the setting of competing risks. Features associated with death and transplant in the DeepHit model were identified using permutation importance. Models were externally validated on data from the University Health Network.

Results: A total of 17,551 patients were included. The mean MELD at listing was 19.4 (SD 8.1). At 120 months of follow-up on the waitlist, 54.6% (9599/17551) of patients underwent LT, 25.6% (4510/17551) of patients died or were removed due to deterioration, and 19.8% (3442/17551) of patients were removed for improvement or were censored. In a competing risk scenario, DeepHit achieved the best CEC scores at 1 (0.813), 3 (0.811), 6 (0.794), and 12 months (0.772) on the waitlist. The cause-specific RSF model had the highest concordance indices for death or transplant at all time points (death: 0.874 at 1 month, 0.840 at 6 months, and 0.814 at 12 months) except for death at 3 months, where DeepHit (0.883) outperformed RSF. RSF also had lower Brier scores overall, except for transplant at 12 months, where DeepHit outperformed RSF (0.206 vs 0.228). These results were similar on external validation. On feature importance assessment, MELD at listing and its components, as well as functional status, age, and blood type, were associated with death and transplant on the waitlist.

Conclusions: A deep learning competing risk analysis can forecast the risks of both death and transplant in patients with MASH on the waitlist, helping to inform clinical decisions by identifying the most impactful covariates for each outcome.

(J Med Internet Res 2026;28:e68247) doi:10.2196/68247

KEYWORDS

liver transplantation; metabolic dysfunction-associated steatohepatitis; waitlist trajectory; metabolic dysfunction; liver disease; cirrhosis; neural network; risk analysis; hepatic Disease; predictive; prediction; deep learning; liver transplant

XSL•FO
RenderX

Introduction

Metabolic dysfunction-associated steatohepatitis (MASH) cirrhosis is a leading cause of liver transplantation (LT) globally and the fastest growing indication for LT in the United States, with the prevalence of waitlisted candidates increasing from 2.5% to 20.4% between 2004 and 2019. The prevalence of MASH cirrhosis is projected to continue to rise significantly in the coming years [1,2]. Candidates waitlisted for LT are prioritized based on their model for end-stage liver disease (MELD) score, which has been periodically reviewed and updated, most recently to the MELD 3.0 score [3]. The MELD-based scoring system predicts waitlist mortality and does not account for type of liver disease, as previous studies have demonstrated minimal effects on the predictive performance [4]. Despite changes to MELD-based scoring systems in recent years and the increasing prevalence of MASH, MELD-based scoring systems have lower concordance in candidates with MASH cirrhosis compared to those listed with other liver diseases, highlighting the need to develop waitlist prediction models that capture the complexity of MASH cirrhosis [3-6]. There are several possible explanations for the lower concordance of MELD models in this population. Patients with MASH, particularly those in the low- to mid-MELD score range, tend to have faster disease progression than is captured by their MELD score progression, higher pre-LT mortality risk, and lower likelihood of recovery than other patients on the waitlist. Additionally, patients with MASH cirrhosis tend to develop clinically significant portal hypertension at lower MELD-sodium (MELD-Na) scores, contributing to higher waitlist mortality [5-8]. Despite having more severe comorbidities, such as portal hypertension, advanced age, higher BMI, diabetes, hypertension, and hyperlipidemia, patients with MASH cirrhosis are less likely to receive a transplant on the waitlist and more likely to face higher waitlist mortality and removal due to becoming too ill to undergo transplant compared to patients listed with other liver diseases [8-12]. Improving waitlist prediction in patients with MASH through developing MASH-specific models can aid clinicians in optimizing their waitlist outcomes and pretransplant status, thus potentially improving overall waitlist outcomes.

Another limitation of MELD-based models is that they provide information on the risk of death on the waitlist without accounting for the risk of transplantation, censoring patients who do not experience the event of interest and leading to biased estimation of risk [7]. These single-risk Cox proportional hazards (CoxPH)–based models cannot predict risk while accounting for the possibility that a patient can experience multiple events at a given time point on the waitlist, while a censored patient can experience a competing event that would make the primary event of interest clinically impossible (ie, a patient who undergoes LT cannot also die on the waitlist). The risks of each event cannot be compared between multiple cause–specific models. In contrast, competing risk models account for multiple mutually exclusive events [13]. Compared to traditional regression methods, machine learning (ML) can handle large, heterogeneous datasets and avoid several fundamental assumptions of linearity and proportionality that CoxPH models make [14-16]. For example, DeepHit is a deep learning competing risk neural network model that captures intricate, nonlinear interactions between several factors and outcomes (such as mortality and transplant) in one model [9,11,17,18]. By using a DeepHit-based model to predict waitlist outcomes, clinicians will better understand the trajectory of patients with MASH cirrhosis with numerous comorbidities who are at high risk of both mortality and transplantation. Patients with MASH at lower risk of death and transplant may be better candidates for living donor liver transplantation (LDLT) and can be redirected accordingly [19]. Due to the fundamental differences between competing risk and single-risk settings, current model evaluation metrics, such as the concordance index (C-index) and Brier score, may be inadequate for use in competing risk settings [12,20]. We design and propose the competing event coherence (CEC) score, a novel performance metric to assess models in a competing events scenario at the patient level. It is an interevent metric that evaluates the match between the event predicted by the model and the actual event that occurred at a given time point for each patient and takes into account multiple competing events.

In this study, we aimed to develop a MASH cirrhosis–specific deep learning model based on DeepHit using data at the time of waitlisting that accounts for the competing risks of death and transplant on the LT waitlist to forecast waitlist trajectory and inform clinical decision-making. We compared the performance of the DeepHit model to MASH-specific single-risk CoxPH and random survival forest (RSF) models. We externally validated the DeepHit model on single-center data. Finally, we developed a DeepHit dashboard to enter and visualize patient trajectory on the waitlist [21].

Methods

Study Design and Participants

We conducted our retrospective study using two study populations. Of 227,647 patients waitlisted in the Scientific Registry of Transplant Recipients (SRTR) from March 1, 2002, to March 2, 2021, we used data from 17,551 patients with MASH cirrhosis for model development (Figure S1). For external validation, we used data from 167 patients with MASH cirrhosis who were waitlisted at University Health Network (UHN), Toronto, Ontario, between 2012 and 2018. Inclusion and exclusion criteria were identical for both cohorts, ensuring consistency in the selection process. We included all adults aged 18 years or older with a primary diagnosis of MASH or cryptogenic cirrhosis (CC) and a BMI of 30 kg/m^2 or more based on previous studies that demonstrate the histological overlap between MASH cirrhosis and CC [22-24]. We excluded patients with a secondary diagnosis other than MASH cirrhosis or CC with a BMI of 30 kg/m^2 or more [22,25]. We excluded LT recipients who were never waitlisted, retransplants, multiorgan transplants, acute liver failure (including status 1 and status 1a candidates), concomitant liver etiologies (viral hepatitis B and C and alcoholic liver disease), hepatocellular carcinoma–related primary or secondary diagnosis or listed with exception points, and those with pre-existing liver malignancies; supplementary information 1). Patients were followed for up

XSL•FO
RenderX

to 120 months, or until death or deterioration, removal, or transplant, whichever occurred first. Events in both cohorts were classified as (1) death (died on the waitlist or removed due to deterioration), (2) received LT, or (3) censored, and event times were determined accordingly. Candidates removed due to deterioration were included in the death group, while those removed from the waitlist for other reasons were classified as censored (Figure S2). Waitlist outcome categorization is consistent with previous literature [26].

Models

We developed all models using SRTR data based on shared features between the SRTR and UHN datasets. The DeepHit model includes a shared network with fully connected layers to capture intrinsic patterns of the input features, which are then connected to two cause-specific subnetworks that learn the relationship between the input features, event type, and time of the event [11,27]. DeepHit outputs discrete monthly risk predictions of the competing events. The sum of the monthly prediction over the entire time horizon T for all the competing events is $1/K$, where K is the total number of competing events. In our case, the maximum event time is 120 months. The cumulative risk for 1 event over this time period is 0.5, as there are two competing events. The DeepHit model includes a shared network with fully connected layers and 2 cause-specific subnetworks that correspond to death and transplant. The joint distribution of the event and first hitting time is learned and outputted through a final layer to output r_i (k, $t|x_i$) defined as

rik,txi=fxi

which is the predicted probability of the k event happening at a specific prediction time t, satisfying

∑k=1, t=1K, Tri(k,t|xi)=1

where f represents the DeepHit structure [11].

We compared DeepHit to single-risk CoxPH and RSF models predicting death or transplant. CoxPH is a linear model that predicts the hazard function with an assumption of proportionality where the individual hazard is proportional to the population baseline hazard that changes over time [14]. (Figure 1) For each cause-specific CoxPH or RSF model predicting death or transplant, the competing event (death or transplant) was censored. Model hyperparameters are available in Table S1 in Multimedia Appendix 1. For comparison, MELD-Na and MELD 3.0 scores were calculated to predict mortality (supplementary information 1)

XSL•FO
RenderX

Figure 1. DeepHit model. (A) The model training process, event prediction, and evaluation process are displayed. (B) The DeepHit model architecture is described. (C) The competing event coherence score evaluates model performance based on the prediction of an event compared to the actual event that occurred at the patient level. (D) DeepHit can be used to predict death and transplant and inform clinical decisions. C-index: concordance index; CEC: competing event coherence; CoxPH: Cox proportional hazards; INR: international normalized ratio; MASH: metabolic dysfunction–associated steatohepatitis; MELD: Model for End-Stage Liver Disease; RSF: random survival forest; SRTR: Scientific Registry of Transplant Recipients; and UHN: University Health Network.

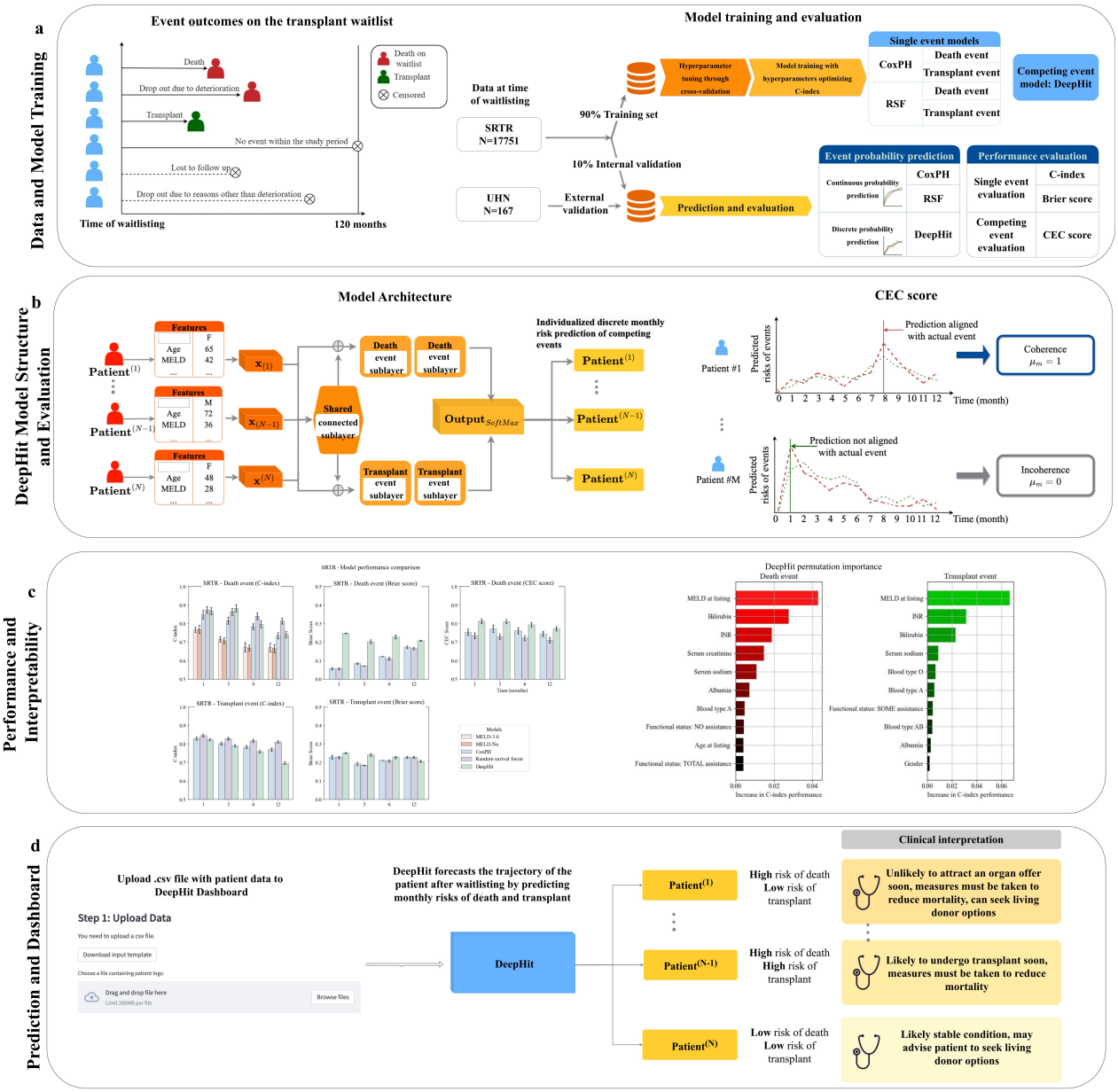

Model Evaluation

Models were evaluated using the C-index and Brier scores. Furthermore, we propose a new metric to evaluate the percentage of prediction that is coherent to the actual event type and time at which it occurred in competing risk scenarios, known as the CEC score (or the μ-score). This overcomes the limitations of the C-index and the Brier score, which examine model performance under a single-risk scenario but cannot be used to compare all the predicted risks for the competing events at the same time. Ideally, at the time of the actual event that occurred, the predicted risk for that event should be higher than the predicted risk of other competing events. For the patients who had an event (death or transplant) within the time frame of interest, the percentage of coherence within the population was calculated. The CEC score measures the alignment between the model risk predictions and the actual patient event that occurred. For the M patients who had the event within the time frame t, the percentage of coherence within the population was calculated. The CEC score measures the alignment between the risk predictions and the actual patient event. Let's μ_m indicate the coherence status of the mth patient being evaluated. Further, let l^* and k^*_m denote the actual event time and event type of the mth patient, respectively; then the proposed μ-score is defined as

$$\mu\text{-score} = \frac{1}{M}\Sigma_{m=1}^{M}\mu_m$$

where coherence for the m patient μ_m is defined as

XSL•FO
RenderX

μm={1,if arg maxk∈ Krm(k,ℓ∗|xm)=km∗0,otherwise,

and r_m (k, l*|x_m) denotes the predicted risks of the 2 events at the event time. Because CoxPH and RSF are single-risk models, we say that the prediction is in coherence (or alignment) if the probability of experiencing an event Pr_{event} is higher for the event corresponding to the actual event (at time l* as compared to the competing one.

C-index, Brier scores, and CEC scores were computed for all models at four time points after waitlisting: 1, 3, 6, and 12 months. One month corresponds to the 25th percentile of event time in the population [17]. Mortality predictions were also generated using MELD-Na and MELD-3.0 scores for performance comparison. Transplant predictions were not developed for the MELD models since they are only intended to predict death on the waitlist. Since DeepHit is a competing risk model, when the performance was evaluated using single event metrics (C-index and Brier score), the competing event was treated as censored [11].

Statistical Analysis

To ensure the robustness of our model, we used rigorous cross-validation, including k-fold cross-validation and hyperparameter tuning to optimize model performance and ensure generalizability (supplementary information 2). SRTR data were split using stratified random split to preserve the event rate of the population in the training and test sets, which represented 90% and 10% of the entire population, respectively. Within the 90% training set, outer five-fold cross-validation was used to obtain an average performance across all validation folds.

Hyperparameter tuning was used to optimize the C-index. To evaluate the performance of our trained models, we first used single-risk metrics, including the time-dependent C-index and the time-dependent Brier score. The model with the highest performance on the outer validation set was used to test performance on the 10% test set as well as the UHN external validation set. Test performance was evaluated at 1, 3, 6, and 12 months using bootstrapping to obtain consistent results where each bootstrapped sample contains patients that were randomly sampled from the cohort with replacement. The Wilson Cox test was subsequently used to test statistical significance in the performance of DeepHit and the other models (CoxPH and RSF) based on the bootstrapped C-indices, Brier scores, and CEC score proposed as the following (Table S1).

Model Interpretability

Permutation importance was used to determine which covariates in the DeepHit model have the largest influence on the prediction of death and transplant by evaluating the contribution of each covariate to the C-index of the model via random permutation of each variable, which is then compared to the original data [28]. The permutation was done 20 times for each variable to obtain the average and SD of increase in C-index.

All analyses were done using Python version 3.8.8 (Python Software Foundation). CoxPH models were developed using the scikit-learn 1.1.1 library and the scikit-survival 0.16.0 library. The MASH-specific DeepHit model was developed using the TensorFlow 0.0.8 library. The data template can be downloaded as a .csv file. The codebase has been published on GitHub [29].

Ethical Considerations

Due to the use of publicly available deidentified United Network for Organ Sharing (UNOS) data, this study was exempt from Research Ethics Board (REB) review. For external validation, REB approval was obtained (REB number 21 - 5783).

Results

Characteristics of SRTR and UHN Cohorts

There were 17,551 patients with MASH in the SRTR cohort, of which 50.2% (8802/17551) were female. Mean MELD at listing was 19.4 (SD 8.1). By 120-month waitlist follow-up, 54.6% (9599/17551) of patients underwent LT, 25.6% (4510/17551) of patients died or were removed from the waitlist due to deterioration, and 19.8% (3442/17551) of patients were removed for improvement or were censored. Around 93.8% (9004/9599) of recipients underwent deceased donor liver transplantation (DDLT). There were 167 patients with MASH cirrhosis in the UHN dataset, and 46.1% (77/167) were female. The mean MELD at listing was 22.1 (SD 6.6). Overall, 62.9% (105/167) of patients underwent LT, of which 72.4% (76/105) of patients underwent DDLT, 23.4% (39/167) died, and 13.7% (23/167) were removed or censored (see supplementary information 1 and Table S2 in Multimedia Appendix 1). Additional features in the SRTR and UHN datasets are available in Table S3 in Multimedia Appendix 1.

Model Performance

In the competing risk scenario, DeepHit consistently achieved statistically significant higher CEC scores at each time point evaluated on the SRTR and UHN datasets (Figures 2 and 3). Numerical results are displayed in supplementary information 6. When comparing model performance using SRTR data (Figure 2), CoxPH, RFS, and DeepHit had higher C-indices than MELD-Na and MELD 3.0 for death events at 1, 3, 6, and 12 months, outperforming MELD-Na and MELD 3.0 at all time points. RSF achieved statistically significantly higher C-indices (P<.01) except in the RSF-DeepHit comparison at 3 months for death event and 1 month for transplant. RSF had statistically significantly lower Brier scores at each time point evaluated for death and transplant, except for at 12 months for the transplant event (DeepHit: 0.206 vs RSF: 0.228).

XSL•FO
RenderX

Figure 2. Model performance comparison with the Scientific Registry of Transplant Recipients data. Performance of models evaluated using C-index, Brier score, and CEC score on SRTR data. CEC: competing event coherence; C-index: concordance index; CoxPH: Cox proportional hazards; MELD: Model for end-stage liver disease; SRTR: Scientific Registry of Transplant Recipients.

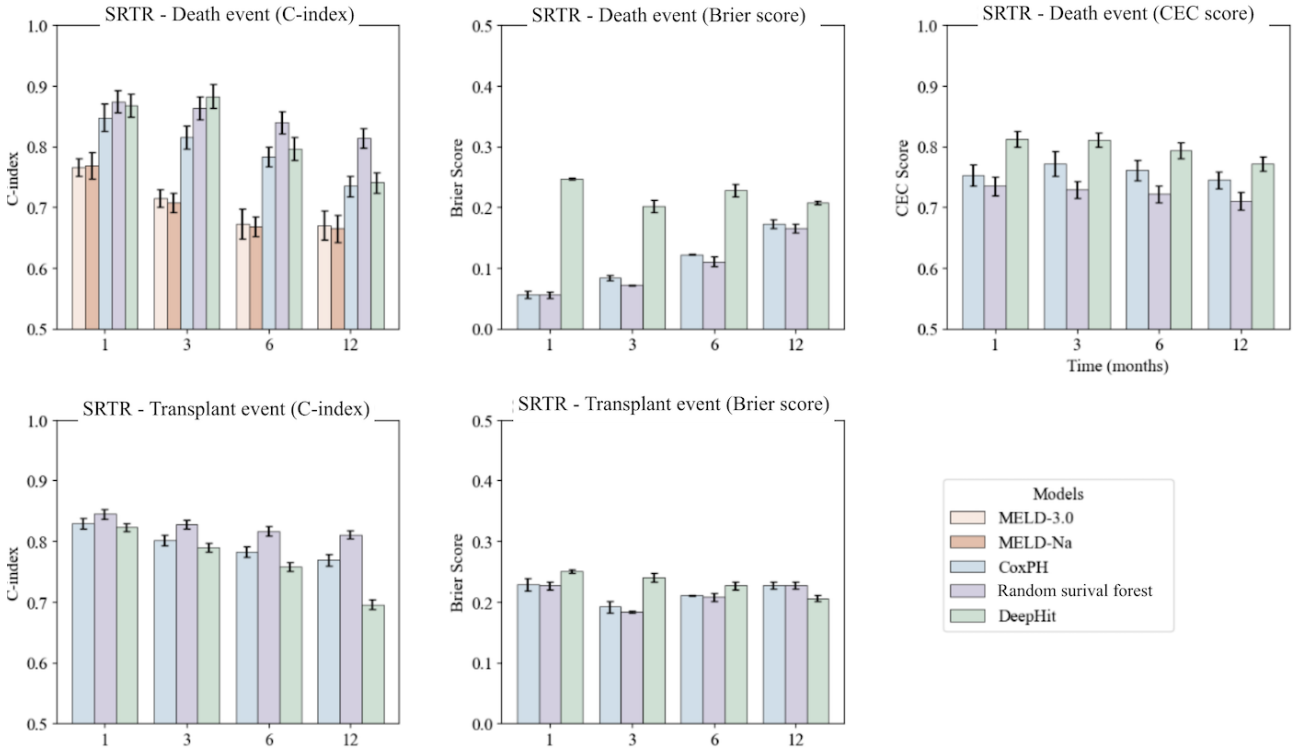

Figure 3. Model performance comparison and external validation of the model using University Health Network data. Performance of models evaluated using C-index, Brier score, and CEC score on UHN data. CEC: competing event coherence; C-index: concordance index; CoxPH: Cox proportional hazards; MELD: Model for end-stage liver disease; SRTR: Scientific Registry of Transplant Recipients.

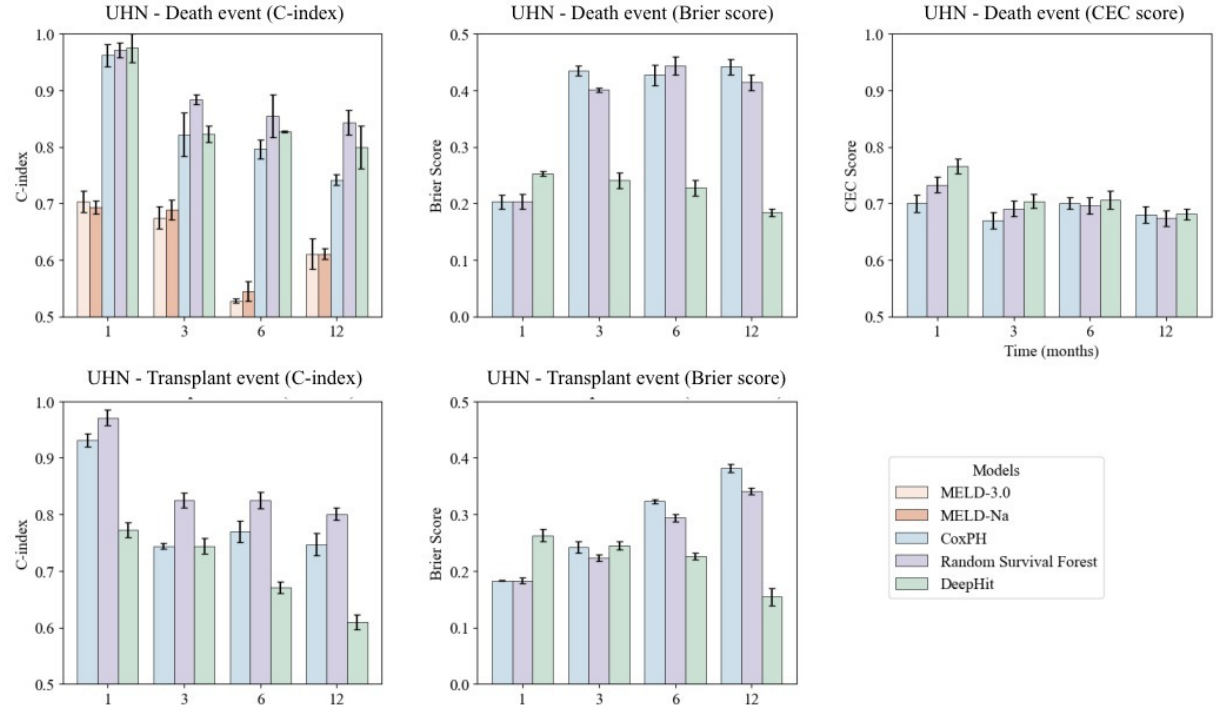

In external validation, DeepHit demonstrated a statistically significantly higher C-index at 1 month for death (0.975); however, RSF had the best performance at all other time points for the death and transplant events. DeepHit also exhibited lower Brier scores at 3, 6, and 12 months compared to RSF for death. For the transplant event, RSF had lower Brier scores at 1 and 3 months than DeepHit, while CoxPH and RSF performed better at 6 and 12 months (Figure 3). Full numerical results can be found in Table S4 in Multimedia Appendix 1.

XSL•FO
RenderX

Forecasting Waitlist Trajectory With DeepHit

Four patients from the SRTR were randomly selected, and their waitlist trajectories of death and transplant were predicted using RSF and DeepHit over a 120-month period based on data available at the time of waitlisting. Figure 4 displays two patients who died on the waitlist at month 33 and month 1, while Figure 5 displays two patients who were transplanted at month 1 and 12. The time at which the event (death or transplant) occurred and the corresponding prediction at that time is magnified. Patient characteristics for the sample patients used here are detailed in Table S5 in Multimedia Appendix 1. Cumulative risk predictions with RSF were generated using two separate models with death or transplant as the outcome of interest and displayed on one graph. Using DeepHit, we generated granular risk predictions from one competing risk model, allowing for visualization of the risk associated with each event over time compared.

Figure 4. Forecasting death on the waitlist using Random Survival Forest and DeepHit with patient examples using data at the time of waitlisting. The vertical line on each plot indicates the event (green=transplant; red=death) and the time at which it occurred. On the DeepHit plots, zoomed-in segments correspond to the time at which the actual event occurred. For a patient that experienced an event during month 1, they were categorized as experiencing the event at time 0. For a patient that experienced an event during months 1 and 2, they were categorized as experiencing the event at month 1 and so on.

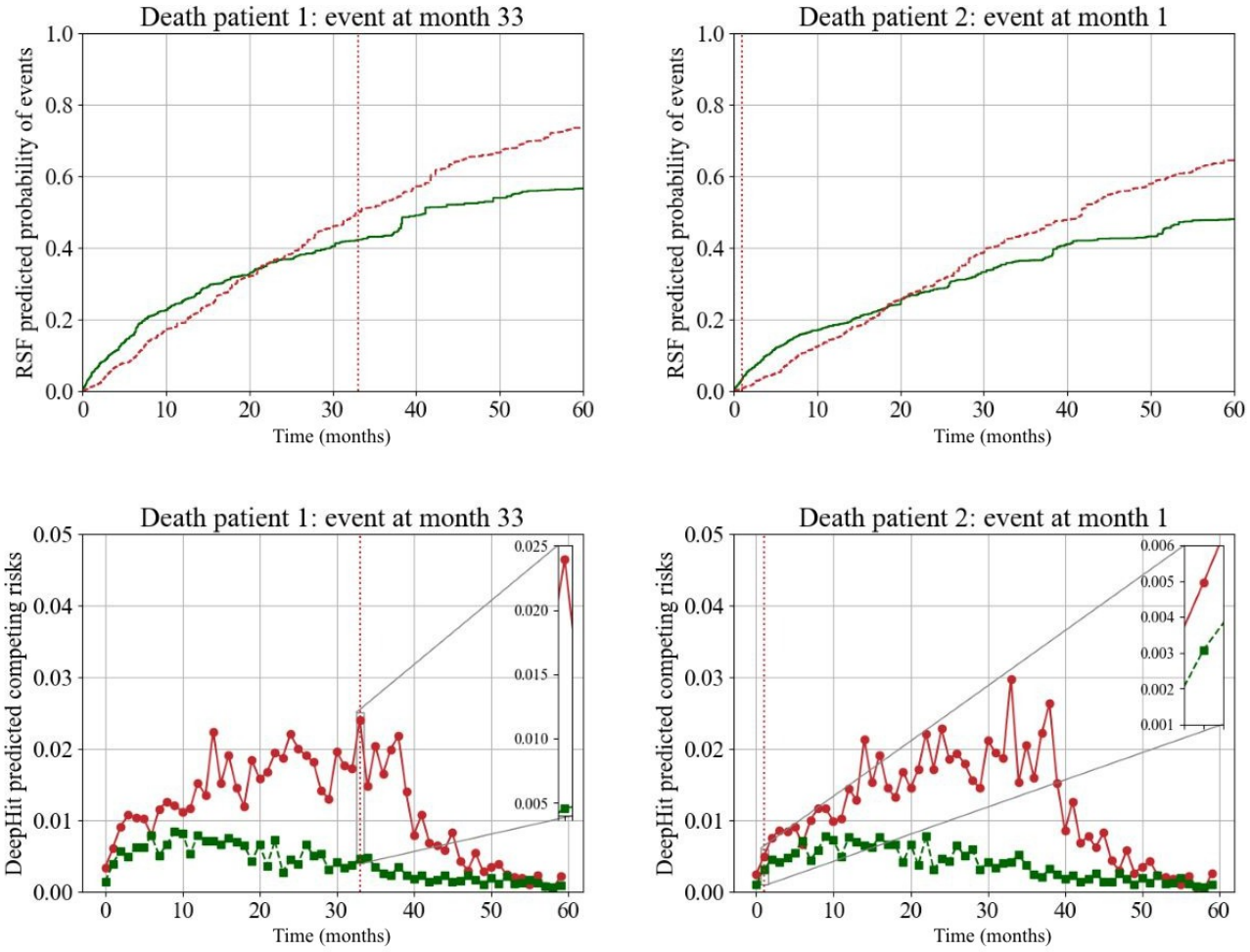

XSL•FO
RenderX

Figure 5. Forecasting transplant on the waitlist using random survival forest and DeepHit. Similar to Figure 4, transplant on the waitlist is predicted using RSF and DeepHit with each plot corresponding to a different patient.

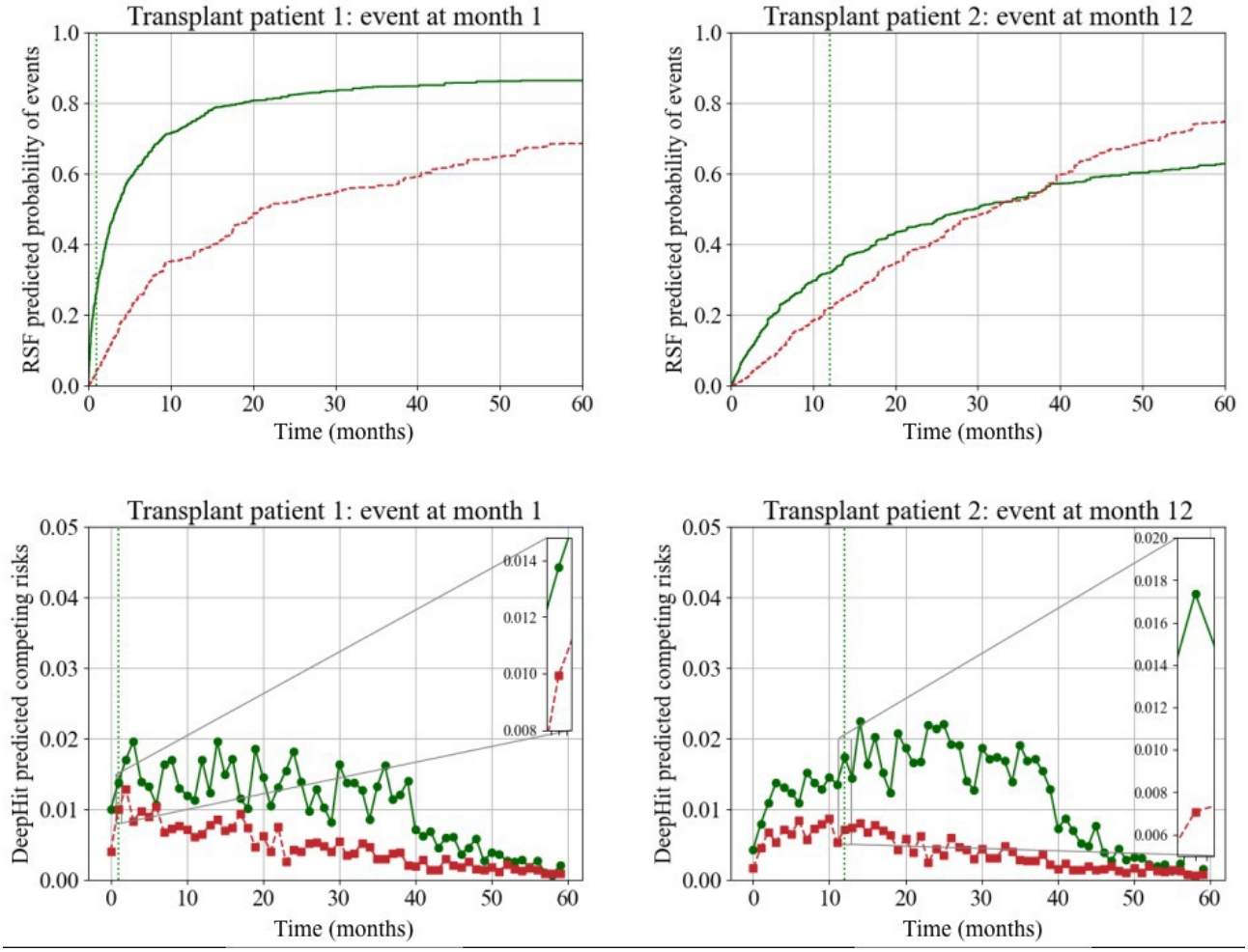

Permutation Importance Using DeepHit

In predicting the death event, MELD at listing emerged as the greatest contributor to the C-index, followed by bilirubin, INR, serum creatinine, and serum sodium. Additional non-MELD features identified were albumin, blood type A, functional status, and age at listing. For transplant events, MELD at listing took precedence, followed by INR and bilirubin. The non-MELD features that influenced transplantation comprised blood types O, A, and AB, as well as functional status, albumin, and gender (Figure 6).

XSL•FO
RenderX

Figure 6. Permutation importance of DeepHit model. Features with the greatest contribution to the prediction of death or transplant were determined using permutation importance. INR: international normalized ratio; MELD: Model for end-stage liver disease

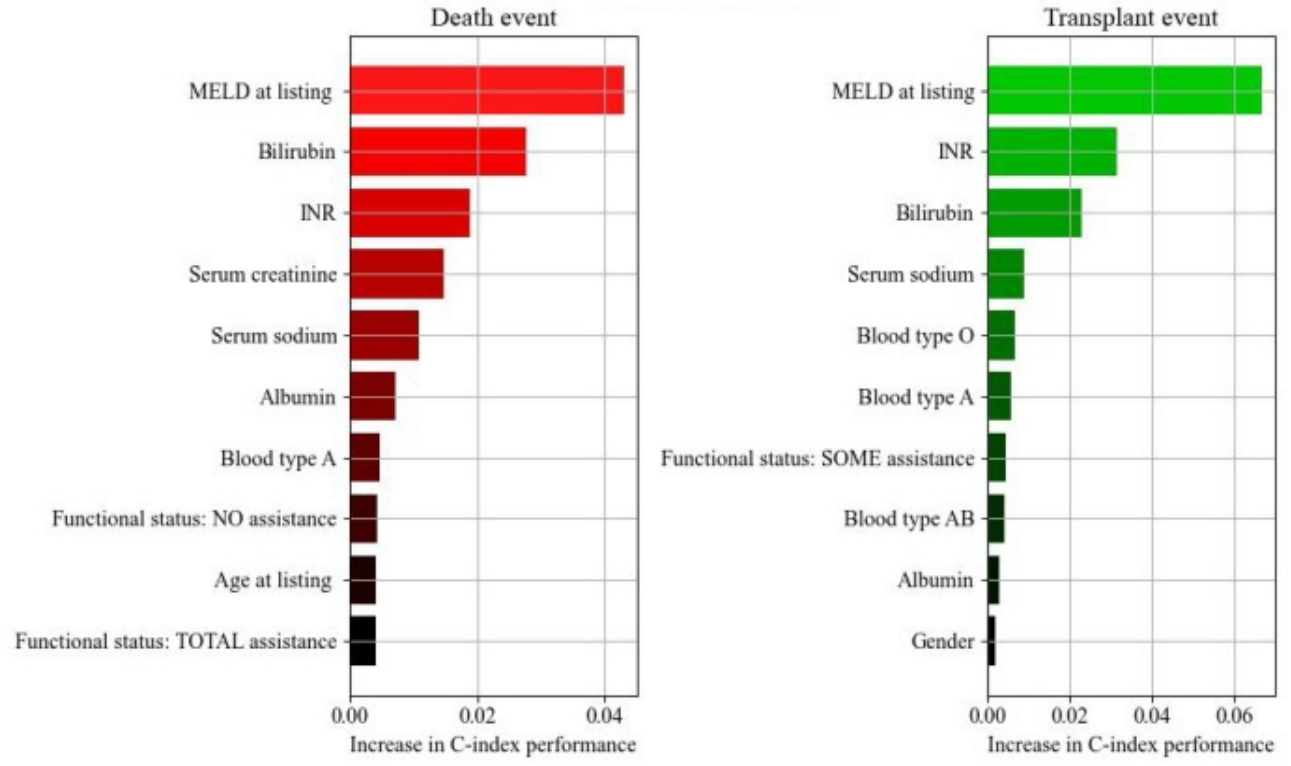

Discussion

Principal Findings

This study aimed to forecast the LT waitlist trajectories of patients with MASH cirrhosis using a deep learning approach that leverages the changing relationships between covariates and risks over time while handling competing risks [11]. We compared a competing risk MASH cirrhosis–specific DeepHit model to single-risk CoxPH and RSF models in forecasting outcomes of death and transplant on the waitlist. We demonstrated that while RSF performed better at most evaluation time points using traditional model performance metrics, DeepHit outperforms RSF when evaluated in a competing risk scenario. Furthermore, we demonstrated that DeepHit can generate and visualize time-varying, discrete predictions of death and transplant. This allows for the identification of patients with a low probability of undergoing transplantation and those at an elevated risk of death while awaiting DDLT, enabling timely consideration of LDLT when available.

MELD-based models poorly predict death in patients with MASH compared to candidates listed for other etiologies and may underestimate the risk of death on the waitlist [5,6]. Additionally, MELD-based models do not provide information on the risk of transplantation. Patients with MASH have a lower 90-day and 1-year probability of transplant and a greater risk of waitlist mortality. Therefore, early identification of patients at low risk of transplant may expedite direction to LDLT and prevent MASH-related mortality on the waitlist [30]. Previous studies have indicated that LDLT results in comparable or improved postoperative outcomes for patients with and without MASH when compared to DDLT. Timely identification of patients enables them to undergo transplantation while in a less decompensated state, leading to more efficient resource usage [31-33]. While MASH-specific models, such as the DeepHit-based model, may not fully replace MELD-based models for waitlist prioritization and liver allocation of all candidates, they can be used to improve the management of waitlisted patients with MASH, particularly those with indolent disease who may spend months on the waitlist and deteriorate before receiving an LT offer. As the prevalence of MASH cirrhosis grows, careful management of waitlisted patients is critical for reducing overall waitlist morbidity and mortality [30]. The DeepHit MASH-specific model also highlights the potential to incorporate personalized prediction of waitlist outcomes at the patient level. This prediction can be updated as clinical status changes over time when modifiable factors associated with a higher risk of waitlist mortality are addressed, providing a more granular prediction of both death and transplant over time on the waitlist as opposed to risk of 90-day mortality alone.

We presented the CEC score as an alternative metric to evaluate competing risk models, emphasizing its role as an interevent measure that assesses a model's performance in predicting 2 competing events [34]. Unlike the C-index, which evaluates performance at the population level, the CEC score assesses performance at the patient level [12,20]. A major problem with the C-index in the setting of competing risks is that the discrimination of a model is reduced when covariates may be associated with both the primary and competing event [20]. Our analysis demonstrates that DeepHit more accurately predicts the actual event (death or transplant) that occurred at specific time points when evaluated using the CEC score compared to

XSL•FO
RenderX

RSF, although RSF performs better based on the C-index and Brier scores. Although the C-index and Brier scores are limited in the scenario of competing risks, DeepHit still exhibited robust performance when evaluated using these metrics, although it was statistically weaker than RSF.

While RSF performed well in a single-risk scenario when evaluated with the C-index and Brier scores, using a noncompeting risk model to predict risk in a clinical setting where patients are at risk of multiple competing events may lead to greater misclassification of risk for single events. We demonstrated this through the improved performance of DeepHit compared to RSF when evaluated with the CEC score. Competing risk analysis may improve predictions of outcomes at the patient level [35-40]. Furthermore, censoring of competing risks can lead to event overestimation in populations at highest risk of experiencing either of the competing events [41]. Competing risk models to evaluate the risks of death and transplant on the waitlist are limited in the field of LT but have been more extensively described and applied to kidney transplantation. Smits et al [42] found that Kaplan-Meier overestimates the chance of transplantation compared to competing risk analysis by over 30% [43]. This not only highlights the importance of competing risk models but also using metrics, such as the CEC score, in the analysis of models to accurately assess performance. We highlight the advantages of DeepHit, whose neural network architecture is particularly well-suited to capturing complex, nonlinear, and interdependent relationships among variables in medium-sized datasets. In contrast to traditional survival models, DeepHit does not rely on restrictive assumptions, such as proportional hazards, enabling greater flexibility in modeling intricate data patterns. Crucially, DeepHit is inherently designed to handle competing risks, estimating the joint probability distribution over multiple, mutually exclusive event types. This multitask learning framework allows the model to share information across outcomes while retaining event-specific distinctions, resulting in more accurate and clinically meaningful survival estimates. By applying a competing risk-specific metric—the CEC score—we demonstrated that DeepHit consistently outperforms traditional ML approaches in predicting time-to-event outcomes with competing risks. These capabilities make DeepHit especially valuable in clinical settings characterized by multiple potential outcomes, such as the organ transplant waitlist.

Previous studies have assessed features associated with death and transplant in LT candidates waitlisted for all indications [5,44-46]. Despite patients with MASH cirrhosis constituting a large and increasing portion of the LT waitlist, limited studies have assessed features associated with death and transplant exclusively in this population [47]. Studies have shown that factors, such as dialysis, sex, race, serum albumin and creatinine, low performance status, and high MELD are associated with increased or decreased risk of transplant among all candidates [5,44-46]. We evaluated the contribution of various covariates in predicting death and transplant events using DeepHit permutation importance, aiming to improve the C-index. In our DeepHit model, MELD at listing emerged as the highest-ranked feature for both death and transplant, followed by components of the MELD score, such as INR, bilirubin, and serum sodium, which have been previously demonstrated [10]. Our permutation analysis reinforced the critical role of MELD-based models in predicting outcomes, showing that biochemical features already included in MELD-based models are implicated in waitlist risk of death and transplant in patients with MASH cirrhosis, potentially reflecting part of the comorbidity burden in this population. Given that SRTR is a historical dataset where allocation is largely determined by MELD score, these findings are not unexpected. Functional status was a highly ranked feature for both events and has been associated with an increased likelihood of transplantation as well as increased waitlist mortality [48]. Poor functional status contributed to the death prediction, possibly because complications of cirrhosis (such as encephalopathy) are associated with poor functional status and increase the risk of death in all LT waitlist candidates [44]. While INR and bilirubin also ranked high, we did not identify creatinine or dialysis in the last week as significant features associated with the prediction of waitlist death or transplant. Age contributed to the death prediction, which has not been found previously in studies on all LT waitlist candidates. Patients with MASH are older, and older patients tend to have poorer waitlist outcomes; therefore, this finding is expected in a MASH cirrhosis cohort [45]. For the transplant event, we found that blood type AB ranked high. These candidates can accept donor livers from nearly all blood types. In other studies, dialysis, sex, race, serum albumin and creatinine, low performance status, and high MELD have been found to be associated with increased or decreased risk of transplant in all candidates [5].

Limitations

The SRTR is a retrospective dataset and limitations, such as high missingness and heterogeneity of data collection must be noted. Furthermore, there are limited longitudinal features in the SRTR dataset, which makes the development of a dynamic model challenging [49]. The permutation importance method is limited as it does not provide information on the directionality of risk such as provided by a hazard ratio, which makes it difficult to apply DeepHit to clinical scenarios where risk factors can be identified and modified to improve outcomes. Furthermore, none of the models developed consider individual organ availability, donor compatibility, and regional variation. Although the DeepHit model performed well when externally validated on the UHN dataset, the UHN sample size was relatively small at 167 patients. Further work should seek to validate this model on larger datasets. Finally, the use of historical waitlist candidate data to devise a ranking system may perpetuate existing inequities and biases in liver allocation; therefore, DeepHit should not be used for definitive decision-making on waitlisting or delisting but can be used in conjunction with other tools.

Conclusion

The DeepHit model can be used to forecast waitlist trajectories of both death and transplant in a competing risk scenario for patients with MASH, using data available at the time of listing. With DeepHit, discretized and dynamic changes of the risk of death and transplant can be visualized and compared across multiple time points and between events. This information can

XSL•FO
RenderX

enhance the current strategies for managing candidates with LT with MASH cirrhosis and act as a tool to advocate for LDLT in patients who are at high risk of waitlist dropout but underserved by MELD-based mortality predictions. A DeepHit MASH-specific model can be used as a clinical adjunct to inform and modify clinical interventions to optimize patient survival on the waitlist. Future studies should focus on improving the evaluation and interpretability of DeepHit models to expand their use in clinical settings, as well as developing larger scale ML models that consider all patients on the LT waitlist.

Funding

This study was funded by the Canadian Institutes of Health Research (MB).

Data Availability

The data reported here have been supplied by the Hennepin Healthcare Research Institute (HHRI) as the contractor for the Scientific Registry of Transplant Recipients (SRTR). The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy of or interpretation by the SRTR or the U.S. Government. The data will not be made available to the public. The codebase has been published on GitHub [29].

Authors' Contributions

GP, YS, SR, and MB contributed to conceptualization. YS performed data curation and investigation. YS and SR were involved in formal analysis. MB was responsible for funding acquisition. YS, GP, SR, and MB performed the methodology. YS and GP managed project administration. MB and SR provided resources. YS, SR, and CL handled the software. MB and SR were responsible for supervision. YS and CL conducted visualization. GP and YS were responsible for writing the original draft. GP, YS, ET, NH, SA, SR, and MB participated in writing review and editing.

Conflicts of Interest

MB has received investigator-initiated grants from Novo Nordisk, Merck, Paladin, Roche, Eisai, Natera, and Astra Zeneca.

Multimedia Appendix 1
Supplementary figures and tables.
[DOCX File, 148 KB - jmir_v28i1e68247_app1.docx]

References

1. Terrault NA, Francoz C, Berenguer M, Charlton M, Heimbach J. Liver transplantation 2023: status report, current and future challenges. Clin Gastroenterol Hepatol 2023 Jul;21(8):2150-2166. [doi: 10.1016/j.cgh.2023.04.005] [Medline: 37084928]
2. Younossi Z, Stepanova M, Ong JP, et al. Nonalcoholic steatohepatitis is the fastest growing cause of hepatocellular carcinoma in liver transplant candidates. Clin Gastroenterol Hepatol 2019 Mar;17(4):748-755. [doi: 10.1016/j.cgh.2018.05.057] [Medline: 29908364]
3. Kim WR, Mannalithara A, Heimbach JK, et al. MELD 3.0: the model for end-stage liver disease updated for the modern era. Gastroenterology 2021 Dec;161(6):1887-1895. [doi: 10.1053/j.gastro.2021.08.050] [Medline: 34481845]
4. Polyak A, Kuo A, Sundaram V. Evolution of liver transplant organ allocation policy: current limitations and future directions. World J Hepatol 2021 Aug 27;13(8):830-839. [doi: 10.4254/wjh.v13.i8.830] [Medline: 34552690]
5. Godfrey EL, Malik TH, Lai JC, et al. The decreasing predictive power of MELD in an era of changing etiology of liver disease. Am J Transplant 2019 Dec;19(12):3299-3307. [doi: 10.1111/ajt.15559]
6. Yardeni D, Shiloh A, Lipnizkiy I, et al. MELD-Na score may underestimate disease severity and risk of death in patients with metabolic dysfunction-associated steatotic liver disease (MASLD). Sci Rep 2023 Dec 13;13(1):22113. [doi: 10.1038/s41598-023-48819-6]
7. Wiesner RH, McDiarmid SV, Kamath PS, et al. MELD and PELD: application of survival models to liver allocation. Liver Transpl 2001 Jul;7(7):567-580. [doi: 10.1053/jlts.2001.25879] [Medline: 11460223]
8. Spooner A, Chen E, Sowmya A, et al. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. Sci Rep 2020 Nov 23;10(1):20410. [doi: 10.1038/s41598-020-77220-w] [Medline: 33230128]
9. Jain V, Bansal A, Radakovich N, et al. Machine learning models to predict major adverse cardiovascular events after orthotopic liver transplantation: a cohort study. J Cardiothorac Vasc Anesth 2021 Jul;35(7):2063-2069. [doi: 10.1053/j.jvca.2021.02.006] [Medline: 33750661]
10. Kim RW, Therneau TM, Benson JT, et al. Deaths on the liver transplant waiting list: an analysis of competing risks. Hepatology 2006;43(2):345-351. [doi: 10.1002/hep.21025]

XSL•FO
RenderX

11. Lee C, Zame W, Yoon J, Van der Schaar M. DeepHit: a deep learning approach to survival analysis with competing risks. Proc AAAI 2018;32(1):04. [doi: 10.1609/aaai.v32i1.11842]
12. Ramspek CL, Teece L, Snell KIE, et al. Lessons learnt when accounting for competing events in the external validation of time-to-event prognostic models. Int J Epidemiol 2022 May 9;51(2):615-625. [doi: 10.1093/ije/dyab256]
13. Monterrubio-Gómez K, Constantine-Cooke N, Vallejos CA. A review on competing risks methods for survival analysis. arXiv. Preprint posted online on Dec 10, 2022. [doi: 10.48550/arXiv.2212.05157]
14. Noordzij M, Leffondre K, van Stralen KJ, Zoccali C, Dekker FW, Jager KJ. When do we need competing risks methods for survival analysis in nephrology? Nephrol Dial Transplant 2013 Nov 1;28(11):2670-2677. [doi: 10.1093/ndt/gft355]
15. Kantidakis G, Putter H, Litière S, Fiocco M. Statistical models versus machine learning for competing risks: development and validation of prognostic models. BMC Med Res Methodol 2023 Feb 24;23(1):51. [doi: 10.1186/s12874-023-01866-z] [Medline: 36829145]
16. Austin PC, Lee DS, Fine JP. Introduction to the analysis of survival data in the presence of competing risks. Circulation 2016 Feb 9;133(6):601-609. [doi: 10.1161/CIRCULATIONAHA.115.017719] [Medline: 26858290]
17. van Geloven N, Giardiello D, Bonneville EF, et al. Validation of prediction models in the presence of competing risks: a guide through modern methods. BMJ 2022 May 24;377:e069249. [doi: 10.1136/bmj-2021-069249]
18. Zhang Z. Survival analysis in the presence of competing risks. Ann Transl Med 2017 Feb;5(3):47-47. [doi: 10.21037/atm.2016.08.62]
19. Arisar FAQ, Chen S, Chen C, et al. Availability of living donor optimizes timing of liver transplant in high-risk waitlisted cirrhosis patients. Aging (Milano) 2023 Sep 15;15(17):8594-8612. [doi: 10.18632/aging.204982]
20. Wolbers M, Koller MT, Witteman JCM, Steyerberg EW. Prognostic models with competing risks: methods and application to coronary risk prediction. Epidemiology 2009 Jul;20(4):555-561. [doi: 10.1097/EDE.0b013e3181a39056] [Medline: 19367167]
21. DeepNASH project. Streamlit. URL: https://deepmash.streamlit.app [accessed 2025-12-02]
22. Charlton MR, Burns JM, Pedersen RA, Watt KD, Heimbach JK, Dierkhising RA. Frequency and outcomes of liver transplantation for nonalcoholic steatohepatitis in the United States. Gastroenterology 2011 Oct;141(4):1249-1253. [doi: 10.1053/j.gastro.2011.06.061] [Medline: 21726509]
23. Caldwell SH, Lee VD, Kleiner DE, et al. NASH and cryptogenic cirrhosis: a histological analysis. Ann Hepatol 2009;8(4):346-352. [doi: 10.1016/S1665-2681(19)31748-X] [Medline: 20009134]
24. Sutedja DS, Gow PJ, Hubscher SG, Elias E. Revealing the cause of cryptogenic cirrhosis by posttransplant liver biopsy. Transplant Proc 2004 Oct;36(8):2334-2337. [doi: 10.1016/j.transproceed.2004.07.003] [Medline: 15561241]
25. Wong RJ, Aguilar M, Cheung R, et al. Nonalcoholic steatohepatitis is the second leading etiology of liver disease among adults awaiting liver transplantation in the United States. Gastroenterology 2015 Mar;148(3):547-555. [doi: 10.1053/j.gastro.2014.11.039] [Medline: 25461851]
26. Zhao X, Naghibzadeh M, Sun Y, et al. Machine learning prediction model of waitlist outcomes in patients with primary sclerosing cholangitis. Transplant Direct 2025 Apr;11(4):e1774. [doi: 10.1097/TXD.0000000000001774] [Medline: 40166627]
27. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat ;2(3):841-860. [doi: 10.1214/08-AOAS169]
28. Breiman L. Random Forests. Mach Learn 2001 Oct;45(1):5-32. [doi: 10.1023/A:1010933404324]
29. Sun Y, Lui C, Rambhatla S. DeepNASH codebase. GitHub.
30. Lim WH, Ng CH, Tan D, et al. Natural history of NASH cirrhosis in liver transplant waitlist registrants. J Hepatol 2023 Oct;79(4):1015-1024. [doi: 10.1016/j.jhep.2023.05.034]
31. Barbas AS, Goldaracena N, Dib MJ, et al. Early intervention with live donor liver transplantation reduces resource utilization in NASH: the Toronto experience. Transplant Direct 2017 Jun;3(6):e158. [doi: 10.1097/TXD.0000000000000674] [Medline: 28620642]
32. Karnam RS, Azhie A, Yang C, et al. Impact of living donor liver transplantation on long‐term cardiometabolic and graft outcomes in cirrhosis due to nonalcoholic steatohepatitis. Clin Transplant 2023 Sep;37(9):e15008. [doi: 10.1111/ctr.15008]
33. Hum A, et al. Adult living donor versus deceased donor liver transplant (LDLT versus DDLT) at a single center: time to change our paradigm for liver transplant. Ann Surg 2019 Sep;270(3):444-451. [doi: 10.1097/SLA.0000000000003463]
34. Wolbers M, Blanche P, Koller MT, Witteman JCM, Gerds TA. Concordance for prognostic models with competing risks. Biostatistics 2014 Jul;15(3):526-539. [doi: 10.1093/biostatistics/kxt059] [Medline: 24493091]
35. Cuthbert AR, Graves SE, Giles LC, Glonek G, Pratt N. What is the effect of using a competing-risks estimator when predicting survivorship after joint arthroplasty: a comparison of approaches to survivorship estimation in a large registry. Clin Orthop Relat Res 2021 Feb 1;479(2):392-403. [doi: 10.1097/CORR.0000000000001533] [Medline: 33105301]
36. Berry SD, Ngo L, Samelson EJ, Kiel DP. Competing risk of death: an important consideration in studies of older adults. J Am Geriatr Soc 2010 Apr;58(4):783-787. [doi: 10.1111/j.1532-5415.2010.02767.x] [Medline: 20345862]
37. Southern DA, Faris PD, Brant R, et al. Kaplan–Meier methods yielded misleading results in competing risk scenarios. J Clin Epidemiol 2006 Oct;59(10):1110-1114. [doi: 10.1016/j.jclinepi.2006.07.002]

XSL•FO
RenderX

38. Kim HT. Cumulative incidence in competing risks data and competing risks regression analysis. Clin Cancer Res 2007 Jan 15;13(2 Pt 1):559-565. [doi: 10.1158/1078-0432.CCR-06-1210] [Medline: 17255278]
39. Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. Am J Epidemiol 2009 Jul 15;170(2):244-256. [doi: 10.1093/aje/kwp107] [Medline: 19494242]
40. Sapir-Pichhadze R, Pintilie M, Tinckam KJ, et al. Survival analysis in the presence of competing risks: the example of waitlisted kidney transplant candidates. Am J Transplant 2016 Jul;16(7):1958-1966. [doi: 10.1111/ajt.13717] [Medline: 26751409]
41. Coemans M, Tran TH, Döhler B, et al. A competing risks model to estimate the risk of graft failure and patient death after kidney transplantation using continuous donor-recipient age combinations. Am J Transplant 2025 Feb;25(2):355-367. [doi: 10.1016/j.ajt.2024.07.029]
42. Smits JM, van Houwelingen HC, De Meester J, Persijn GG, Claas FH. Analysis of the renal transplant waiting list: application of a parametric competing risk method. Transplantation 1998 Nov 15;66(9):1146-1153. [doi: 10.1097/00007890-199811150-00006] [Medline: 9825809]
43. El Ters M, Smith BH, Cosio FG, Kremers WK. Competing risk analysis in renal allograft survival: a new perspective to an old problem. Transplantation 2021 Mar 1;105(3):668-676. [doi: 10.1097/TP.0000000000003285] [Medline: 32332421]
44. McCabe P, Hirode G, Wong R. Functional status at liver transplant waitlisting correlates with greater odds of encephalopathy, ascites, and spontaneous bacterial peritonitis. J Clin Exp Hepatol 2020;10(5):413-420. [doi: 10.1016/j.jceh.2020.04.015] [Medline: 33029049]
45. Nagai S, Safwan M, Kitajima T, Yeddula S, Abouljoud M, Moonka D. Disease‐specific waitlist outcomes in liver transplantation – a retrospective study. Transpl Int 2021 Mar;34(3):499-513. [doi: 10.1111/tri.13814]
46. Su F, Yu L, Berry K, et al. Aging of liver transplant registrants and recipients: trends and impact on waitlist outcomes, post-transplantation outcomes, and transplant-related survival benefit. Gastroenterology 2016 Feb;150(2):441-453. [doi: 10.1053/j.gastro.2015.10.043] [Medline: 26522262]
47. Wong RJ, Singal AK. Trends in liver disease etiology among adults awaiting liver transplantation in the United States, 2014-2019. JAMA Netw Open 2020 Feb 5;3(2):e1920294. [doi: 10.1001/jamanetworkopen.2019.20294] [Medline: 32022875]
48. McCabe P, Wong RJ. More severe deficits in functional status associated with higher mortality among adults awaiting liver transplantation. Clin Transplant 2018 Sep;32(9):e13346. [doi: 10.1111/ctr.13346] [Medline: 29979466]
49. Lee C, Yoon J, Schaar MVD. Dynamic-deephit: a deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. IEEE Trans Biomed Eng 2020 Jan;67(1):122-133. [doi: 10.1109/TBME.2019.2909027] [Medline: 30951460]

Abbreviations

C-index: concordance index
CC: cryptogenic cirrhosis
CEC: competing event coherence
CoxPH: Cox proportional hazards
DDLT: deceased donor liver transplantation
INR: international normalized ratio
LDLT: Living donor liver transplantation
LT: liver transplantation
MASH: Metabolic dysfunction-associated steatohepatitis
MELD: Model for end-stage liver disease
MELD-Na: model of end-stage liver disease-sodium
ML: machine learning
REB: Research Ethics Board
RSF: random survival forest
SRTR: Scientific Registry of Transplant Recipients
UHN: University Health Network
UNOS: United Network for Organ Sharing

XSL•FO
RenderX

Edited by J Sarvestan; submitted 01.Nov.2024; peer-reviewed by C Sun, X Fu; revised version received 11.Aug.2025; accepted 12.Aug.2025; published 29.Jan.2026.

Please cite as:
Punchhi G, Sun Y, Tan E, Hlaing NKT, Liu C, Asrani S, Rambhatla S, Bhat M
Forecasting Waitlist Trajectories for Patients With Metabolic Dysfunction–Associated Steatohepatitis Cirrhosis: A Neural Network Competing Risk Analysis
J Med Internet Res 2026;28:e68247
URL: https://www.jmir.org/2026/1/e68247
doi:10.2196/68247

© Gopika Punchhi, Yingji Sun, Eunice Tan, Naomi Khaing Than Hlaing, Chang Liu, Sumeet Asrani, Sirisha Rambhatla, Mamatha Bhat. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 29.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

End-to-End Platform for Electrocardiogram Analysis and Model Fine-Tuning: Development and Validation Study

Lucas Bickmann[1,2], MSc; Lucas Plagwitz[1], MSc; Antonius Büscher[1,3], Dr med; Lars Eckardt[3], Prof Dr med; Julian Varghese[2], Prof Dr med

[1]Institute of Medical Informatics, University of Münster, Münster, Germany

[2]Institute of Medical Data Science, Otto-von-Guericke University Magdeburg, Leipziger Str. 44, Building 2, Magdeburg, Germany

[3]Clinic for Cardiology II: Electrophysiology, University Hospital Münster, Münster, Germany

Corresponding Author:
Julian Varghese, Prof Dr med
Institute of Medical Data Science, Otto-von-Guericke University Magdeburg, Leipziger Str. 44, Building 2, Magdeburg, Germany

Abstract

Background: Electrocardiogram (ECG) data constitutes one of the most widely available biosignal data in clinical and research settings, providing critical insights into cardiovascular diseases as well as broader health conditions. Advancements in deep learning demonstrate high performance in diverse ECG classification tasks, ranging from arrhythmia detection to risk prediction for various diseases. However, the widespread adoption of deep learning for ECG analysis faces significant barriers, including the heterogeneity of file formats, restricted access to pretrained model weights, and complex technical workflows for out-of-domain users.

Objective: This study aims to address major bottlenecks in ECG-based deep learning by introducing ExChanGeAI, an open-source, web-based platform designed to offer an integrated, user-friendly platform for ECG data analysis. Our objective is to streamline the entire workflow—from initial data ingestion (regardless of device or format) and intuitive visualization to privacy-preserving model training and task-specific fine-tuning—making advanced ECG deep learning accessible for both clinical researchers and practitioners without machine learning (ML) expertise.

Methods: ExChanGeAI incorporates robust preprocessing modules for various ECG file types, a set of interactive visualization tools for exploratory data analysis, and multiple state-of-the-art deep learning architectures for ECGs. Users can choose to train models from scratch or fine-tune pretrained models using their own datasets, while all computations are performed locally to ensure data privacy. The platform is adaptable for deployment on personal computers as well as scalable to high-performance computing infrastructures. We demonstrate the platform's performance on several clinically relevant classification tasks across 3 external and heterogeneous validation datasets, including a newly curated test set from routine care, evaluating both model generalizability and resource efficiency.

Results: Our experiments show that de novo training with user-provided, task-specific data can outperform a leading foundation model, while requiring substantially fewer parameters and computational resources. The platform enables users to empirically determine the most suitable model for their specific tasks, based on systematic validations, while lowering technical barriers for out-of-domain experts and promoting open research.

Conclusions: ExChanGeAI provides a comprehensive, privacy-aware platform that democratizes access to ECG analysis and model training. By simplifying complex workflows, ExChanGeAI empowers out-of-domain researchers to use state-of-the-art ML on diverse datasets, democratizing the access to ML in the field of ECG data. The platform is available as open-source code under the Massachusetts Institute of Technology (MIT) license.

(J Med Internet Res 2026;28:e81116) doi:10.2196/81116

KEYWORDS

health informatics; electrocardiogram; machine Learning; deep learning; end-to-end platform

Introduction

Background

Deep learning methods applied to Electrocardiogram (ECG) analyses have demonstrated their potential as practice-changing diagnostic tools, providing critical insights into heart-related diseases [1-3]. While the push for newer technologies and improved performance metrics is essential, ensuring these advancements are accessible for general use is equally important. Tools like ChatGPT (OpenAI) have demonstrated the potential for a broad and easy application of artificial intelligence (AI), allowing users to leverage sophisticated technologies without

XSL•FO
RenderX

in-depth expertise. Clinician-researchers who seek to apply machine learning (ML) to assess its potential benefits should have access to solutions that facilitate exploration and application without requiring extensive technical knowledge from data handling to data analysis. To address these challenges, there is a need for a comprehensive, end-to-end platform that integrates currently fragmented technical steps like ECG-specific data handling, preprocessing, data visualization, and model training, including transfer learning that requires cumbersome manual scripting. This would enable a seamless workflow from data loading to model deployment and would not only empower researchers to apply and train or fine-tune deep learning models for ECG analysis without programming but also facilitate reproducibility. Moreover, existing pretrained models for ECG data may not disclose model weights, which presents significant challenges (refer to "Related Work" below). While open weights empower users to use and adapt the model, they also promote reproducibility and transparency [4,5].

Additionally, the broad use of medical data is crucial for the advancement of personalized and specialized medicine but inhibits some immediate risks, such as data breaches [6]. As datasets continue to grow and come from diverse sources, ensuring data security becomes increasingly complex. To address this challenge, specialized decentralized learning techniques, such as federated learning or swarm learning, allow valuable insights to be gained without directly sharing sensitive data [7]. Furthermore, establishing uniform data standards can simplify data handling, reduce technical barriers related to varying data formats, and eliminate the need for programming, thereby making advanced analytics accessible to a broader range of users.

In this work, we introduce a novel open-source end-to-end platform for 12-lead ECGs called ExChanGeAI that streamlines essential steps of ECG analysis: (1) data loading and preprocessing of multiple input formats, (2) manual and computer-aided analysis of ECG waveform data, (3) one-click fine-tuning of classification models, allowing users to train and customize ML models with no prior expertise, (4) the trained models use the cross-platform industry-standard Open Neural Network Exchange (ONNX), enabling deployment in every instance of ExChanGeAI and facilitating the exchange of custom models across different instances, and (5) prediction of diseases with and without using pretrained models. Model sharing is supported via an integrated and adaptable WebDav file server called Model ExChanGe. The platform is built upon the principle of open-source code and open-weights, offering full transparency and control, empowering users to contribute to the advancement of ECG analysis models.

Related Work

Multiple studies and reviews have addressed ECG classification and have shown that fine-tuning and transfer learning improve classification results [8-10]. A study has reported improved model accuracy by fine-tuning networks trained on diverse datasets, demonstrating enhanced performance transitioning to smaller datasets [11]. However, the used data and pretrained models were not shared. Another study used transfer learning with convolutional neural networks (CNNs) for atrial fibrillation classification, pretraining on large public datasets and fine-tuning on smaller sets, achieving performance gains [9]. While code was available, pretrained models were not shared, and usability remains a significant barrier. Multiple reviews have summarized ECG analysis pipelines and deep learning methods, such as detailed essential pipeline steps [12] and reviews of techniques like CNNs and recurrent neural networks for arrhythmia classification [13]. The SelfONN model [14] showed competitive performance in general ECG classification on PTB-XL (Physikalisch-Technischen Bundesanstalt-extra large [National Metrology Institute of Germany]) but lacked resource sharing. Various types of autoencoders, including low-rank attention [15], long short-term memory [16], adversarial [17], and denoising [18] approaches, have been explored for feature extraction, anomaly detection, and noise handling. The low-rank attention autoencoder reported high accuracy on 2 datasets by focusing on spatial features. ECG-NET, based on long short-term memory, proclaimed high accuracy for arrhythmia classification on a single database in beat-based validation. An adversarial autoencoder with a temporal CNN published superior scores of anomaly detection for 2 datasets. The attention-based denoising autoencoder improved noisy ECG signal reconstruction. However, limitations across these studies include dataset dependence, restricted generalizability, lack of publicly available pretrained models and code, and validation variability.

In a recent study leveraging the gold-standard PTB-XL [19,20] dataset, the performance characteristics of multiple deep learning models were evaluated across a spectrum of training-data sizes [21]. Notable findings indicated that the InceptionTime and XceptionTime architectures [22,23] exhibited particularly compelling performances. Specifically, InceptionTime demonstrated superior efficacy when trained with smaller datasets, whereas XceptionTime surpassed all other models in performance as training dataset size increased. This suggests a potential trade-off between model complexity and data requirements for optimal diagnostic accuracy in this domain. Due to the demonstrated strength in low- and high-data scenarios, these leading architectures for ECG analysis are highly relevant for evaluation and inclusion in the platform, particularly in contexts where training data availability may vary, such as in medical contexts.

There have also been several claimed foundation models in the field of ECG classification [24]. To the best of our knowledge, these, however, are trained on a singular database [25] and are yet undisclosed or have closed-source code and weights [26,27] in general. In one case, the published weights are different from the original model of the paper due to privacy concerns [28]. A request to publish another trained model has been declined due to intellectual property and legal concerns [29]. They are trained with techniques, such as contrastive and masked learning. This allows for unsupervised training, but restricts the learning to the latent space. For downstream tasks, such as classification, fine-tuning is required. The publications report high scores for classification; however, additional tasks are not available in the published model. While these models mark significant progress in the field, they often grapple with issues, such as overfitting to specific datasets, limited scalability, or insufficient handling

XSL•FO
RenderX

of the variability and quality complications intrinsic to diverse ECG datasets.

Despite advances in ECG analysis and deep learning, the current workflows remain complex, requiring manual data transformation, preprocessing, and the use of separate tools for the visualization and analysis of ECGs, as well as for training and fine-tuning deep learning models. This fragmentation multiplies technical burdens, hinders reproducibility, and acts as a major barrier for widespread clinical adoption. Some frameworks exist to reduce the boilerplate of ML, such as the graphical tool Orange (University of Ljubljana) [30]. It does not require code, yet the workflow has to be set up manually per drag-and-drop, and it does not include any of the ECG analysis tools, such as QRS detection. In the case of AutoML tools, such as GAMA (General Automated Machine learning Assistant; originally developed by Pieter Gijsbers and Joaquin Vanschoren at the Eindhoven University of Technology) [31], it still remains code-centric. Data ingestion, for both methods, is not available out-of-the-box, such as for Digital Imaging and Communication in Medicine (DICOM), even ignoring the sampling rate, lead order, and other inconsistencies across datasets. Both tools remain unsuitable for end-to-end ECG analysis and training of ML models without substantial ML or data scientist expertise. Our work directly addresses this gap with ExChanGeAI, an integrated, accessible, and containerized end-to-end platform. This platform facilitates the visualization, transformation, prediction, and fine-tuning of deep learning models specifically for ECG data. It can leverage pretrained models, and it supports a broad range of formats and preprocessing steps, ensuring usability across different clinical and research settings. ExChanGeAI serves as a valuable resource for researchers, enabling efficient training and fine-tuning of deep learning models while preserving data privacy. This enhances both the accessibility and utility of advanced ECG analysis.

Methods

Overview

This section introduces the ExChanGeAI platform, a fully containerized, interoperable, and standardized end-to-end platform for ECG analysis, diagnosis prediction, and model fine-tuning. It is designed for nonexpert users, enabling advanced AI-enabled workflows in a unified interface without requiring specialized technical expertise. The open-source code is freely available under the MIT license [32].

ExChanGeAI Platform

ExChanGeAI is a containerized web application providing an integrated suite of AI-enabled ECG analysis tools for researchers and clinicians. The platform merges human expert analysis capabilities and AI predictions in a unified, interactive end-to-end platform (refer to Figure 1). The functionalities include (1) signal analysis, (2) model-based prediction, (3) model exchange and repository, and (4) semiautomated training and fine-tuning.

Figure 1. Overview of the capabilities of the end-to-end platform ExChanGeAI and its three main distinct parts: (1) Analysis, (2) The Artificial Intelligence Ecosystem, and (3) Interoperability. AI: artificial intelligence; ML: machine learning; ONNX: Open Neural Network Exchange.

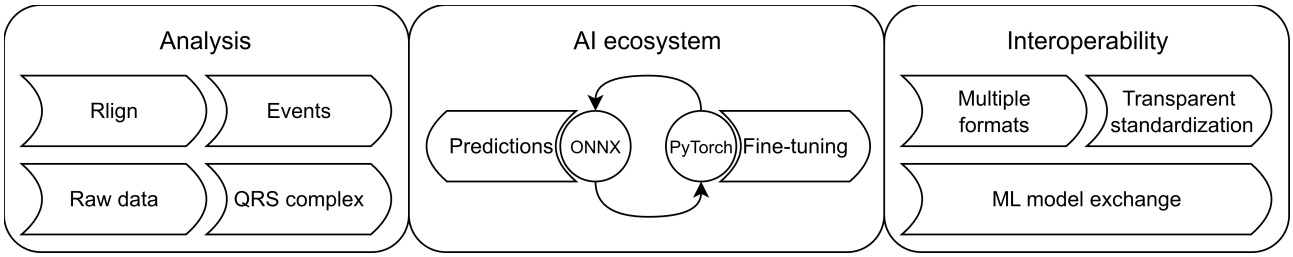

The platform is open-source, using the standardized ONNX model format and additionally supports and encourages the open-weights practice of ML models [4]. The platform consists of multiple views with different foci. The analysis view integrates the visualization of waveforms of raw signals, QRS complexes, and events (fiducial points), computed transparently by Neurokit2 [33]. Additionally, precise R-peak alignment and median beat transformations are supported through the integration of the recently published ECG-preprocessing package, Rlign [34]. These data transformations can be exported in different formats for further research. The platform focuses on resting 12-lead ECGs, displayed in a 2x6 grid in mV scale, and integrates general spatial transformations, including zooming with synchronized adaptation across all leads. This signal view has been designed in collaboration with cardiologists for their everyday use. For visualizing QRS complexes and events, lead II is conventionally applied. The prediction view uses selected models to predict diagnoses and other targets—such as QTc—based on raw signal data, offering a table with predicted diagnosis probabilities (or arbitrary keys), highlighted with a clear color-coding scheme. Dataset distributions, confusion matrices, and class-wise receiver operating characteristic curves can be computed on the platform itself, including suggested thresholds for class-specific Fmax scores. This enables researchers to optimize the threshold for each specific dataset and the health care providers' requested strategy. These differ gradually and can focus on either specificity or sensitivity, depending on the classification targets. Models are provided within the platform using an integrated and interchangeable file server, which enables the crucial aspect of model sharing (see section "Interoperability and Model Sharing"). Currently, we provide multiple models for four targets: (1) diagnostic superclasses, (2) anterior and inferior myocardial infarction (MI), (3) diverse bundle branch blocks, and (4) revascularization need (refer to section "Results"). Researchers can also train specialist models based on their own labeled data. The training and fine-tuning require no prior knowledge of ML. Currently, only 12-lead ECG data with corresponding labels are required. The backend is engineered for high performance and scalability using asynchronous views,

XSL•FO
RenderX

multithreading, and full compute unified device architecture GPU (graphics processing unit) support, including multi-GPU data-parallel training for large-scale fine-tuning.

Data Loading and Preprocessing

ExChanGeAI is designed to handle the variability of real-world ECG data. It supports various ECG input formats, across all possible sampling rates, CSV files (.csv), NumPy arrays (.npy, .npz), DICOM-stored WaveformSequences (.dcm), MATLAB (MathWorks) formatted data (.mat), general DAT files (.dat), and XML files (.xml). Major research and clinical ECG standards (PhysioNet - DAT, UK Biobank - XML) are supported. For ExChanGeAI, all data are normalized by resampling frequency signals to a configurable unified frequency target (defaulting to 100 Hz) using the Fast Fourier Transform, as previous work has shown that the sampling rate does not notably decrease the performance of ML models, but reduces computational overhead manifold [35-38]. This parameter is fully configurable, allowing users to adjust the sampling rate in a local deployment to match specific requirements. Signals not conforming to the standardized format (12-lead, 10-second waveform) are adjusted via expansion or cropping, and all scales are automatically standardized to millivolt (mV) with a 1000 analog-to-digital converter units gain, if necessary (refer to Figure 2).

Figure 2. Flowchart of the preprocessing applied to any electrocardiogram data while being loaded into the application, independent of the file format. ADC: analog-to-digital converter.

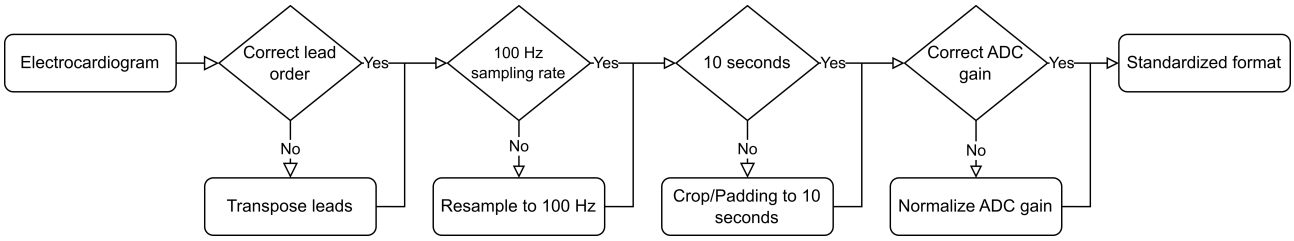

Interoperability and Model Sharing

The platform enables training of new models in a secure and privacy-preserving manner. Still, as seen with many publications in the medical domain, pretrained models are not made public [26,29] or depend on external libraries and require specific versions [28]. To promote the open and interoperable ML standard, this work adopts the ONNX as the primary used format. Therefore, our platform is compatible with all ONNX models, honoring the current operation set (Opset 20 and below), and PyTorch models, if the given model structure is provided alongside. The models are not specified with any special requirements, except for a dynamic batch size export. With the use of ONNX, this work aims to ensure that the trained model is widely accessible and interoperable. Therefore, a model sharing interface, called Model ExChanGe, is integrated into the platform, where curated pretrained models are automatically synced and made available for prediction as well as fine-tuning. Additional models can be published into the repository, or your own WebDav instance can be set up.

By default, ExChanGeAI provides 3 baseline model architectures—each as pretrained and untrained models. This includes the XceptionTime [23], InceptionTime [22], and the PhysioNet/CinC Challenge 2021 12-lead second-best model, DSAIL SNU (Data Science & Artificial Intelligence Laboratory Seoul National University [19,39,40])—the best model did not provide weights. Additional models can be incorporated by uploading them into the platform or using the default model exchange file server. We incorporate all evaluated models to extend the research community with open-source and pretrained model weights.

The adoption of the ONNX industry-standard model format ensures that ExChanGeAI is not limited to a proprietary ecosystem. Models trained or fine-tuned with ExChanGeAI can be seamlessly imported, shared, or deployed across different institutions and environments, whether in research settings or clinical contexts. This plug-and-play capability allows users to leverage pretrained models, contribute their own, or integrate compatible architectures developed externally, with zero code changes and minimal configuration. As a result, the platform not only accelerates collaboration but also supports sustainable, evolving workflows as new models and data become available.

Fine-Tuning Platform

ExChanGeAI provides an interactive user interface for fine-tuning with user-supplied data, abstracting all low-level ML steps, and facilitating the exchange of prediction models among researchers. To ensure broad compatibility with various models and platforms, we natively support PyTorch and developed a custom parser for ONNX models to adapt the computation graph where applicable. ExChanGeAI is built upon PyTorch [41], which facilitates on-device training and leverages the ONNX framework during inference. ONNX's custom training runtime does not support all PyTorch operators with corresponding gradient implementations, such as functions like ReduceMin and diverse Pooling Operators (Opset≤18). While defining custom operators may solve this, it requires detailed implementation knowledge for each unknown operator, creating a significant barrier for practical applications. To overcome these limitations, we use onnx2torch to convert our uniquely parsed ONNX models into PyTorch models. This conversion enhances compatibility, allowing training across different ONNX Opset versions and using the extensive feature set of PyTorch. Crucially, our parser automatically adapts the classification head to the new number of classification targets. These conversion steps are computed independently for each fine-tuning process and are entirely transparent to the user, without requiring any programming knowledge (refer to Figure 3). We provide two distinct training methods, including fine-tuning the classification heads, where the majority of the model weights are frozen, or training the entire model. Both

XSL•FO
RenderX

methods use pretrained weights if a pretrained model is selected. The freezing of weights is handled automatically in the background, only allowing modification of the unfrozen weights or those belonging to the classification head.

Figure 3. Overview of the semiautomatic fine-tuning process in the backend of ExChanGeAI. ECG: electrocardiogram; ONNX: Open Neural Network Exchange.

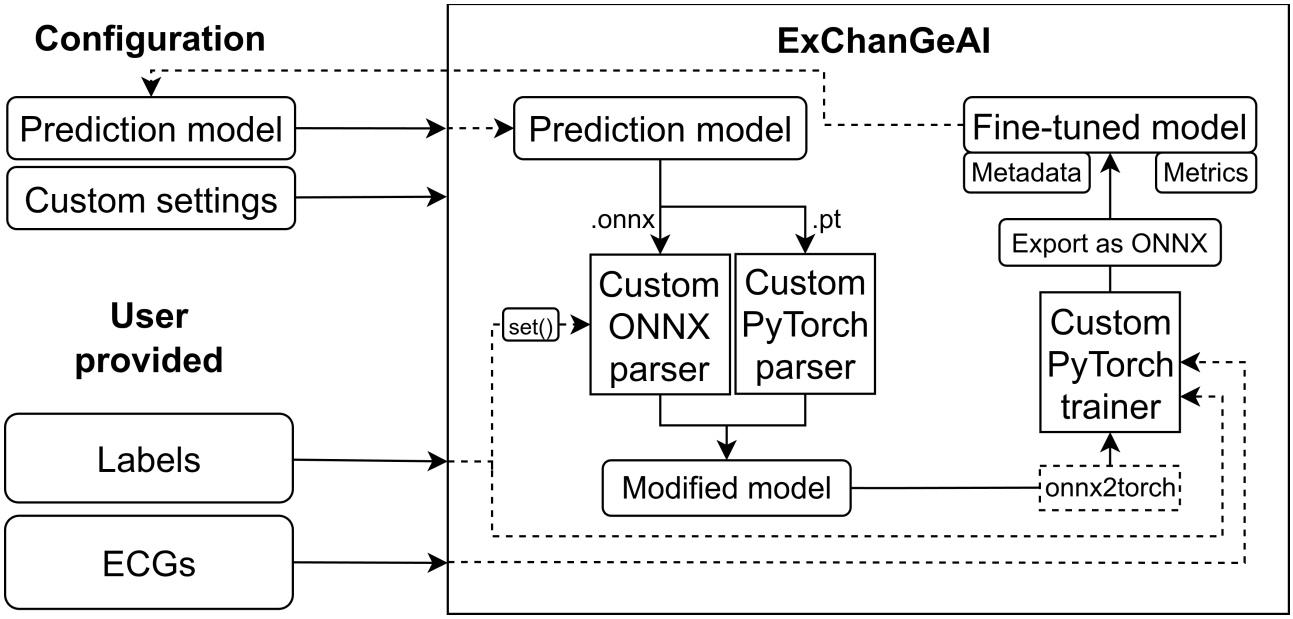

The training and fine-tuning process uses by default the AdamW optimizer [42], due to its better generalizability and convergence than the default Adam optimizer [43], and the ExponentialLR learning rate scheduler (γ=0.9). Adaptive optimization algorithms tend to be more robust, have faster convergence, and therefore improve resource usage [44]. Other optimizers, such as variants of Adam and stochastic gradient descent, are available as alternatives. Initially, a learning rate finder is executed to automatically determine the optimal initial learning rate based on the provided model and data. It has been shown that this method improves the convergence speed and reliability [45]. Furthermore, the loaded data is automatically split into stratified training/evaluation sets with an 80/20% distribution. We limit the training process to a default maximum of 50 epochs and incorporate checkpointing for models with the lowest weighted validation loss, alongside early stopping. Advanced settings, such as other optimizers, batch size, number of epochs, maximum initial learning rate, and gamma can be adapted via the user interface, if required. The best model, in addition to training and evaluation statistics, is exported and downloaded after completion. To ensure comprehensive reporting and documentation, the exported statements include (1) the number of samples, (2) distribution and corresponding labels, (3) the used base model, (4) the training and evaluation loss per epoch, and (5) the corresponding F_1-scores on the evaluation set.

Evaluation

To thoroughly assess the reliability and performance of ExChanGeAI's training process, we conduct a series of scenario-based tests. The training and prediction capabilities are one of the key features of this work and are therefore mainly evaluated via training and fine-tuning of classification models on various tasks and tests on internal and external datasets using the ExChanGeAI platform, where possible. This benchmarking is crucial for validating the platform's core value proposition (1) to empower users, even without ML expertise, and (2) to rapidly develop and deploy accurate prediction models for ECG analysis. Demonstrating robust performance across diverse datasets reinforces the platform's usability and reliability, ultimately building trust and accelerating adoption among clinicians and researchers.

Datasets

To train and fine-tune deep learning models, we use multiple targets and use the large open-access gold standard PTB-XL dataset [20]. In order to demonstrate model training under data-scarce conditions, the provided stratified fold 9 is used for training and fold 10 for intradataset testing. Age and sex, while technically predictable from ECG signals, are generally of limited clinical relevance in the context of model development, as these attributes are readily and reliably obtained through direct patient observation rather than requiring inferential prediction from ECG data.

The comparison includes a wide-ranging spectrum of ischemic heart diseases, structural heart diseases, and conduction abnormalities: (1) broad diagnostic categories, including MI, ST/T changes, conduction disturbances, and hypertrophy, (2) specific comparisons, such as anterior MI vs inferior MI, and diverse types of bundle branch blocks, including complete left (CLBBB), complete right (CRBBB), and incomplete left (ILBBB). The label distribution of PTB-XL is outlined in the Multimedia Appendix 1. A comparison baseline XceptionTime model is trained on the folds 1 - 8 of PTB-XL for all targets to outline the training performance under data-rich conditions. We also compare our baseline model against the benchmark scores of InceptionTime from the PTB-XL benchmark paper [46]. To facilitate a direct comparison of the platform's one-click training capabilities with the benchmark paper, we additionally trained

XSL•FO
RenderX

an InceptionTime model on folds 1 - 8 of PTB-XL, focusing on the superclasses.

To analyze the applicability of models across sites, we evaluate all models on interdatasets based on Yang et al [47,48], MIMIC-IV-ECG (Medical Information Mart for Intensive Care IV Electrocardiogram) [49], and Emergency Department Münster (EDMS) [38], which is an entirely new dataset from our hospital site. This helps to gauge the model's generalizability across different ECG recordings. The latter two datasets demonstrate a relatively balanced distribution with respect to age and sex, whereas the former dataset is predominantly male and represents the smallest sample size among the 3 test sets (Multimedia Appendix 2). We selected these datasets to ensure both clinical relevance and diversity in our evaluation. EDMS is a newly collected internal dataset derived from routine clinical care, providing contemporary, real-world ECG data. MIMIC-IV ECG represents one of the largest publicly available routine care datasets, enabling robust large-scale analyses. PTB-XL serves as the gold standard for annotated ECG data, offering high-quality expert labels, while the Yang et al dataset provides a similarly gold-standard resource from an entirely different geographical region, allowing us to assess model generalizability across populations. To prevent patient leakage, only one record per patient was kept in PTB-XL before the stratified split.

Some classes present slight variations, such as MIMIC and EDMS, which do not have descriptive ECG statements but general *ICD-10* (*International Classification of Diseases, 10th Revision*) codes, which are not necessarily based solely on the given ECG. We extract the signals with the corresponding fitting maps and merged superclasses. The corresponding *ICD-10* codes, or included statements, are given for each map (refer to Table 1). This includes changes such as that bundle branch blocks (BBBs) are only encoded and divided into left- and right-BBB. We evaluate the prediction performance accordingly, counting complete- and incomplete-RBBB as the superclass RBBB. Physionet classes, from the pretrained models, are mapped to the BBB and superclasses where applicable, as no MI classes are included. For MIMIC and EDMS, in instances of multiple ECGs of the same patient, only the first record was used to negate any multiple patient testing.

Table . Matched superclass and bundle branch block statements of PhysioNet 2021 and Emergency Department Münster to PTB-XL (Physikalisch-Technischen Bundesanstalt-extra large [National Metrology Institute of Germany]) classes.

Classes	Superclasses		Bundle branch blocks		
PTB-XL[a]	CD[b]	STTC[c]	CLBBB[d]	CRBBB[e]	IRBBB[f]
PhysioNet 2021	BBB[g], CLBBB|LBBB[h], CRBBB|RBBB[i], IRBBB, IAVB[j], LAnFB[k], NSIVCB[l]	TAb[m], TInv[n], LQT[o]	CLBBB|LBBB	CRBBB|RBBB	IRBBB
EDMS[p]	—[q]		LBBB	RBBB	

[a]Physikalisch-Technischen Bundesanstalt-extra large.

[b]CD: conduction disturbance.

[c]STTC: ST/T Changes.

[d]CLBBB: complete left bundle branch block.

[e]CRBBB: complete right bundle branch block.

[f]IRBBB: incomplete right bundle branch block.

[g]BBB: bundle branch block.

[h]LBBB: left bundle branch block.

[i]RBBB: right bundle branch block.

[j] IAVB: first-degree atrioventricular block.

[k]LAnFB: left anterior fascicular block.

[l]NSIVCB: nonspecific intraventricular conduction disorder.

[m]Tab: T wave abnormal.

[n]TInv: T wave inversion.

[o]LQT: prolonged QT interval.

[p]EDMS: Emergency Department Münster.

[q]Not available.

Table 2 shows a comprehensive overview of all extracted targets, the number of samples, and mapped *ICD-10* codes across the different external test datasets. Most datasets are uncurated and reflect real-world implications, in contrast to semicurated datasets, such as PTB-XL. There may be bad quality data, as well as discrepancies between *ICD-10* codes and mapped classes. The codes may be based on other electronic health data than the ECG, and differences may occur due to indifference between suspected and confirmed diagnoses. Label noise, which is only present in MIMIC and EDMS, may lead to underestimation of classification performance [50]. This, however, allows us to compare the models across real-world

XSL•FO
RenderX

data, showing possible impact on clinical care. An additional case study with a manually annotated gold standard is evaluated on the internal EDMS dataset. To demonstrate the advanced classification task of revascularization ("does the patient require revascularization?"), which is not available in PTB-XL, the models are trained and fine-tuned using the new EDMS dataset. The labels of revascularization are case-based if the patient has been treated with a revascularization. It consists of 240 positive and negative cases each, whereas negative cases are only a stratified subset of the complete annotated dataset. An additional stratified subset (20%) of these data points is kept as testing data.

XSL•FO
RenderX

Table . External test datasets and their class distribution across all categories, their mapping from diagnostic statements or *ICD-10* (*International Classification of Diseases, 10th Revision*) codes to PTB-XL (Physikalisch-Technischen Bundesanstalt-extra large [National Metrology Institute of Germany] classes, and the data format, including sampling rate, analog-to-digital converter gain, and additional annotations.

Publication, notes, and target	Classes
Yang and Feng [48]	
500 Hz based on dataset labels	
Superclasses	• HYP[a] (HEH[b]): 647 • CD[c] (IAVB[d], IIAVB1[e], IIIAVB[f], BBB[g], LAFB[h], NICD):[i] 1974
Bundle branch blocks	• CLBBB:[j] 43 • CRBBB:[k] 328 • IRBBB:[l] 1051
MIMIC-IV-ECG[m]	
500 Hz with 200 adu/mV gain based on *ICD-10*[n] codes	
Superclasses	• HYP (I11, I51.7): 500 • MI[o] (I21, I22): 500 • CD (I44): 500
Myocardial infarcts	• AMI[p] (I21.0): 500 • IMI[q] (I21.1): 500
Bundle branch blocks (variation)	• LBBB[r] (I44.7): 500 • RBBB[s] (I45.1): 500
EDMS[t]	
100 Hz based on ICD-10[n] codes	
Superclasses	• HYP (I11, I51.7): 149 • MI (I21, I22): 302 • CD (I44): 255
Myocardial infarcts	• AMI (I21.0): 51 • IMI (I21.1): 43
Bundle branch blocks (variation)	• LBBB (I44.7): 73 • RBBB (I45.1): 48
500 Hz annotated through cardiologists	
Revascularization (20% test subset)	• Yes: 48 • No: 48

[a]HYP: hypertrophy.

[b]HEH: heart enlargement and hypertrophy.

[c]CD: conduction disturbance.

[d]IAVB: first-degree atrioventricular block.

[e]IIAVB: second-degree atrioventricular block.

[f]IIIAVB: third-degree atrioventricular block.

[g]BBB: bundle branch block.

[h]LAFB: left anterior fascicular block.

[i]NICD: nonspecific intraventricular conduction disturbance.

[j]CLBBB: complete left bundle branch block.

[k]CRBBB: complete right bundle branch block.

[l]IRBBB: incomplete right bundle branch block.

[m]MIMIC-IV-ECG: Medical Information Mart for Intensive Care IV Electrocardiogram.

[n] *ICD-10: International Classification of Diseases, 10th Revision.*

XSL•FO
RenderX

[o]MI: myocardial infarction.

[p]AMI: anterior myocardial infarction.

[q]IMI: inferior myocardial infarction.

[r]LBBB: left bundle branch block.

[s]RBBB: right bundle branch block.

[t]EDMS: Emergency Department Münster.

Model Selection

We use the best-performing models based on the aforementioned previous research [21]. The study has shown that the InceptionTime performs well with less data in comparison to XceptionTime, but its performance lags behind when larger datasets are used. Therefore, we train a baseline model on XceptionTime, as its capability exceeds InceptionTime due to the large amount of data available for the baseline model. In comparison to the InceptionTime and XceptionTime models, we evaluate the DSAIL SNU PhysioNet 2021 model [39,40,51], the PhysioNet 2021 competition leader with available weights, and the only available foundation model, ECG-FM (Electrocardiogram Foundation Model) [28]. We aim to assess the effectiveness and improvements gained using ExChanGeAI's training and fine-tuning capabilities and the possible use of pretrained models, especially in resource-constrained environments with very few data points.

Preprocessing and Training

We evaluated the various architectures using two training strategies: (1) fine-tuning only the classification head and (2) training all layers. Trainings were conducted using the default settings of ExChanGeAI (commit number 4d862c04) to maintain consistency and integrity.

Xception and InceptionTime models are trained de novo (from random initialization) using non-normalized ECG data, which has been internally validated to achieve higher performance. ECG-FM and DSAIL SNU are pretrained models on PhysioNet 2021 labels, which were then fine-tuned for each classification target on PTB-XL fold 9 to demonstrate fine-tuning capabilities. In a special case, to showcase the capability of cross-task transfer learning, a pretrained XceptionTime model (superclasses with PTB-XL folds 1 - 8) was separately fine-tuned on the revascularization task.

The ECG-FM foundation model's training data details are unknown, though it is based on PhysioNet 2021 (which includes PTB-XL), limiting the validity of intradataset evaluation. The "physionet_finetuned" model differs from published results due to inaccessible weights and requires 500 Hz, 5-second *z* score normalized inputs. The training of ECG-FM was implemented with custom training code due to dependency complexities. It requires a custom library, which is only compatible with the end-of-life version of Python 3.9 (Python Software Foundation). Additionally, an ONNX export is not possible with these custom functions, impeding the usage of interoperable standards and therefore the deployment into the platform. The PyTorch implementation, as an alternative, could not be used due to the outdated and unsupported versions of major libraries, resulting in dependency conflicts. The results of ECG-FM are therefore achieved outside the platform, yet are given for comparative purposes. DSAIL SNU was adapted for ONNX export and initialized with the unavailable coinput features (age and sex) using the default values and their required missing feature flags as specified in its corresponding publication, alongside the used minimum-maximum normalization in pretraining.

All ECG recordings were processed at the sampling rate required by each model. For ECG-FM, recordings that are natively 500 Hz (PTB-XL, Yang et al , MIMIC-IV-ECG) were used unchanged, while recordings available only at 100 Hz (EDMS) were up-sampled to 500 Hz via a Fast Fourier Transform to match the model's input requirement. DSAIL SNU, XceptionTime, and InceptionTime require a 10-second ECG sampled at 100 Hz. Consequently, any 500 Hz recordings (PTB-XL, Yang et al, MIMIC-IV-ECG) were down-sampled to 100 Hz using the platform's interoperable data-loading pipeline, and recordings only available at 100 Hz (EDMS) were left unchanged.

Performance Metrics

To evaluate model performance, we use the F_1-score for overall assessment and calculate the average and median for central tendency across datasets. Predictions are derived by selecting the class with the highest probability. We use the F_1-score rather than area under the curve because clinical relevance often requires accurate classification at a single operating threshold—outlining the critical balance between precision and recall—whereas area under the curve summarizes performance across all possible thresholds and may not reflect the real-world consequences of specific predictions. Additionally, we focus on the weighted F_1-score to account for the class imbalance commonly seen in medical datasets, ensuring that minority classes are appropriately represented in the evaluation. Macro F_1, accuracy, precision, recall, Brier score, and expected calibration error top-label, and classwise (macro and weighted), with bootstrapped 95% CIs, as well as per class metrics, confusion matrices, and 2-sided paired *t* tests for external datasets, are given in the Multimedia Appendix 3. For F_1-score evaluation, classes not present in the PhysioNet labels were removed. To reflect realistic out-of-distribution prediction, the comparison did not remove false positives. Robustness, indicated by lower IQR and coefficient of variation (CV) values, suggests consistency across datasets. Computational scaling is analyzed using the number of parameters, floating-point operations per second (FLOPs), training, and inference timings. For all architectures on the ExChanGeAI platform, estimated timings are reported as the mean with SD, evaluated using a run of 1500 ECGs from the MIMIC database, based on the superclass subset and the respective models.

These comprehensive evaluations enable us to determine how well models within ExChanGeAI perform under varied conditions, providing insights into their practical application in

XSL•FO
RenderX

diverse real-world scenarios. Through this extensive testing framework, we confirm ExChanGeAI's robustness, adaptability, and reliability for ECG analysis across multiple datasets, diagnostic statements, and applicability for different use cases.

Ethical Considerations

Collection and analysis of the EDMS dataset were approved by the responsible medical ethics committee (Ärztekammer Westfalen-Lippe, approval EDMS: no. 2022 - 494 f-S) under a waiver of informed consent in accordance with state law for health data privacy (§6 Abs. 2 Gesundheitsdatenschutzgesetz Nordrhein Westfalen (Health Data Protection Act of North Rhine-Westphalia)). The creation and analysis of the MIMIC-IV-ECG dataset were reviewed by the Institutional Review Boards of Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology, which waived the requirement for individual patient consent because the project did not impact clinical care and all protected health information was deidentified. The Yang et al dataset was approved by the Medical Ethics Committee of Chinese People's Liberation Army General Hospital (approval no S2019-318-03) under informed consent from the participants. The PTB-XL dataset is publicly available under a waiver of the Institutional Ethics Committee (approval no PTB-2020 - 1), complying with Health Insurance Portability and Accountability Act (HIPAA) standards. All waivers allow secondary analysis with given approvals under the respective data regulation and privacy protection standards.

Results

The central goal of ExChanGeAI is to make model selection and empirical comparison tractable, reproducible, and accessible within a single, seamless platform. Therefore, the interface is visually structured into different foci, such as data analysis (refer to Figure 4). The analysis view provides interactive visualization of individual ECG files. Users can select to view ECG data based on multiple views - raw time series, QRS complexes, fiducial point annotation, Rlign median beats, and Rlign time-aligned ECG. When visualizing QRS complexes, the interface displays overlaid waveforms, potentially highlighting morphological features. For raw time series visualization, the platform presents the standard 12-lead ECG signals as separate plots, allowing for detailed inspection of each lead's waveform. The fine-tuning view displays options for model selection, training method, and custom model naming. A bar chart visualization summarizes the distribution of labels within the loaded dataset, presenting counts for categories, such as "CLBBB," "IRBBB," and "CRBBB" as shown in Figure 5. Numerical dataset characteristics, including the total number of imported ECGs and labels, are presented as well.

Figure 4. Overview of the ExChanGeAI web interface, showing the "Analyse" page with a sample electrocardiogram in QRS-waveforms. ECG: electrocardiogram.



XSL•FO
RenderX

Figure 5. Overview of the ExChanGeAI web interface, showing the "Finetune" page, showing the default necessary parameters, and the data distribution across labels of an example bundle-branch-block dataset. ECG: electrocardiogram.



Table 3 shows the weighted F_1-scores of the Xception-, InceptionTime, DSAIL SNU, and ECG-FM architectures across the different classification targets on the test datasets. The tables are organized such that columns represent different models, while rows represent various classification tasks and datasets. Specifically, each row corresponds to a particular classification task evaluated across different test sets (PTB-XL, Yang et al, MIMIC-IV, and EDMS). The most comprehensively de novo–trained model XceptionTime serves as a reference (Training PTB-XL folds 1-8) for assessing performance scaling with increased data availability.

XSL•FO
RenderX

Table . Performance evaluation of various models on electrocardiogram classification tasks across the test datasets.

Model	XceptionTime[ab]		InceptionTime[ab]	DSAIL SNU[ac]		ECG-FM[d]	
	Training folds (1-8)	Training folds (9)		Fine-tune	Pretrained PhysioNet 2021	Pretrained PhysioNet 2021	Fine-tune
Training PTB-XL[e] folds	1-8	9	9	9	—[f]	—	9
Trained layers	All	All	All	Head	—	—	Head
Superclasses							
PTB-XL	0.792	0.686[g]	0.651	0.536[h]	0.038[h]	0.174[h]	0.690[i,h]
Yang et al	0.335	0.647[i]	0.064	0.585[g]	0.402	0.491	0.173
MIMIC-IV[j]	0.371	0.374[i]	0.358[g]	0.323	0.055	0.192	0.355
EDMS[k]	0.432	0.387[i]	0.379[g]	0.331	0.039	0.333	0.216
Myocardial infarcts							
PTB-XL	0.938	0.853[g]	0.902[i]	0.685[h]	—	—	0.685[h]
MIMIC-IV	0.753	0.734[i]	0.726[g]	0.566	—	—	0.403
EDMS	0.566	0.484	0.584[i]	0.343	—	—	0.532[g]
Bundle branch blocks							
PTB-XL	0.911	0.912[i]	0.891[g]	0.832[h]	0.000[h]	0.016[h]	0.790[h]
Yang et al	0.730	0.101	0.792[i]	0.496	0.007	0.089	0.617[g]
MIMIC-IV	0.825	0.820[i]	0.819[g]	0.333	0.000	0.028	0.087
EDMS	0.739	0.827[i]	0.732[g]	0.622	0.000	0.118	0.248
Revascularization							
EDMS	0.750[l]	0.688[i]	0.645[g]	0.635	—	—	0.603
Weighted F_1 (interdataset only; excluding PTB-XL)							
Mean (SD)↑	0.611 (0.188)	0.562[g] (0.243)	0.567[i] (0.251)	0.470 (0.137)	—	—	0.359 (0.193)
Median (IQR)↑	0.73 (0.432-0.750)	0.647[i] (0.387-0.734)	0.645[g] (0.379-0.732)	0.496 (0.333-0.585)	—	—	0.355 (0.216-0.532)
CV[m]↓	0.308	0.433[g]	0.443	0.290[i]	—	—	0.538

[a]Trained via ExChanGeAI.

[b]De novo training.

[c]DSAIL SNU: Data Science & Artificial Intelligence Laboratory Seoul National University

[d]ECG-FM: Electrocardiogram Foundation Model; used custom code for the training platform.

[e]PTB-XL: Physikalisch-Technischen Bundesanstalt-extra large.

[f]Not applicable.

[g]Second best model in this category.

[h]These models were pre-trained on datasets including PTB-XL; results on this specific target should be interpreted as in-distribution evaluations.

[i]Best model in this category.

[j]MIMIC-IV: Medical Information Mart for Intensive Care IV.

[k]EDMS: Emergency Department Münster.

[l]Transfer learning based on reference XceptionTime (superclasses with folds 1 - 8).

[m]CV: coefficient of variation.

For illustration, the first row presents superclass classification scores on the PTB-XL test dataset, whereas the reference XceptionTime (trained on PTB-XL folds 1 - 8) achieves a weighted F_1-score of 0.792, and XceptionTime, trained on fold 9 only, reaches 0.686. As an example for MI, the InceptionTime model achieves the second-best F_1-score across both

XSL•FO
RenderX

inter-datasets (0.726 on MIMIC-IV and 0.584 on EDMS), while XceptionTime achieves a slightly higher score on the former (0.734), and ECG-FM on the latter (0.532). The last rows show the aggregated statistics, showing Xception and InceptionTime have the best average and median F_1-scores, outlining the top overall performing models, while DSAIL SNU shows the best IQR and CV values, exhibiting the most robust scores across external datasets. As anticipated, increasing the amount of training data leads to improved performance. Importantly, ExChanGeAI is able to handle this scalability, achieving better outcomes as more data are incorporated (see reference model "XceptionTime" in Table 3). For example, on the PTB-XL dataset, test performance on the superclasses and myocardial infarct targets improves substantially—by 15.4% and 9.96% respectively—when XceptionTime is trained on folds 1‐8 compared to training only on fold 9.

XceptionTime and InceptionTime, representing architectures trained de novo on PTB-XL, meaning with random initialization and without any prior training at all, often achieved the highest results across classification tasks. In contrast, the pretrained models, DSAIL SNU and ECG-FM, exhibited a more nuanced performance profile in our limited data setting. Initially, both models demonstrated suboptimal classification accuracy, especially on datasets outside of their pretraining domain (PhysioNet 2021). Fine-tuning them on a single PTB-XL fold for each classification target led to significant improvements for both DSAIL SNU and ECG-FM. However, they were outperformed by the de novo InceptionTime (8 out of 9 targets) and XceptionTime models (7 out of 9) within our evaluation. Still, fine-tuned DSAIL SNU exhibited the best robustness (lowest IQR and CV), suggesting stable results across disparate external cohorts despite lower mean and median F_1-scores. This increase in intradataset performance does not always translate to interdataset performance, as expected, due to overfitting to the dataset distribution. Overall, XceptionTime and InceptionTime trained from scratch showed the highest average and median F_1-scores across all evaluated classification tasks.

The InceptionTime model trained on folds 1‐8, as reported in the PTB-XL benchmark paper [46], achieves a macro F_1-score of 0.7495. In comparison, our reference XceptionTime model attains a macro F_1-score of 0.768, while a comparable training (not in Table 3) of InceptionTime achieves 0.7707. These results demonstrate that ExChanGeAI's training capabilities not only match but also surpass established benchmark baselines, all without reliance on additional resources. Additionally, the platform's flexible workflow enables rapid prototyping for novel tasks, such as revascularization prediction (absent in PTB-XL), trained and evaluated using the new EDMS dataset. Here, transfer learning, based upon the reference XceptionTime (0.750), increased the F_1-score by 9% relative to de novo XceptionTime (0.688).

Additionally, it has to be noted that the foundation model ECG-FM is the largest with over 90 million parameters, followed by DSAIL SNU (2M), Xception (401K), and InceptionTime (457K). In terms of computational complexity, ECG-FM is the most demanding (14 GFLOPS), followed in descending order by Inception-(460 MFLOPS), XceptionTime (256 MFLOPS), and DSAIL SNU (89 MFLOPS). The inference timings on a 6-core Zen4 CPU correspond to mean 27 (SD 33.78) ms (XceptionTime), mean 26 (SD 36.56) ms (InceptionTime), and mean 29.5 (SD 13.9) ms (DSAIL) using the ExChanGeAI platform. Training with 1500 training samples is estimated with mean 13180 (SD 44) ms per epoch and 8.79 ms per sample (XceptionTime), mean 19520 (SD 153) ms per epoch and 13.01 ms per sample (InceptionTime), and mean 10210 (SD 34) ms per epoch and 6.80 ms per sample (DSAIL). Comparing the classification performance against the computational complexity, XceptionTime and InceptionTime stand out as the top performers. All models trained on the PTB-XL dataset are available in Multimedia Appendix 4.

Discussion

Overview

Our evaluation of ExChanGeAI on established architectures reveals several key insights into model selection, particularly in data-constrained scenarios. The end-to-end platform streamlines both training and fine-tuning, yielding robust performance metrics across diverse ECG classification tasks.

Comparison With Prior Work

Our evaluation shows training on limited data and cross-dataset testing exposes inherent generalization gaps and variability in performance—a major difference compared to the often overoptimistic intradataset results seen in the literature. Consequently, the near-perfect accuracy metrics—in intratest set and simple tasks, such as tachy- and bradycardia prediction [28]—are not reproducible when models are evaluated on external, independent datasets. However, when models are evaluated using intradataset testing and ample training data are available, achieving high scores becomes more feasible and reproducible on external datasets (see the baseline XceptionTime model in the Multimedia Appendix 3). As expected and in line with previous findings [52,53], all models exhibited performance drops on external datasets. Yet, as an important factor, the end-to-end platform training surpasses the established benchmark InceptionTime model, outlining the competitive training performance of the platform without requiring expert knowledge.

Principal Findings

XceptionTime models were particularly notable for their parameter efficiency and competitive accuracy, reaffirming their architectural strength. Notably, learning from scratch proved to be a strong alternative to transfer learning, as de novo XceptionTime and InceptionTime models often outperformed fine-tuned pretrained models despite having fewer parameters. However, performance variability was observed across different classification tasks and datasets, as expected, indicating a sensitivity to dataset-specific scaling and parameter optimization within specific model architectures. Pretrained models, while anticipated to leverage their extensive prior knowledge, presented a mixed picture in our data-limited scenarios: while fine-tuning improved their performance, they generally did not consistently surpass the de novo trained XceptionTime and InceptionTime models. However, the pretrained model did

XSL•FO
RenderX

exhibit enhanced robustness against performance degradation across external datasets in most cases, compared to de novo trained models. Among all, DSAIL SNU demonstrated the lowest performance variance, underscoring its robustness.

Limitations

First, while pretrained models offer potential advantages, their benefits are not guaranteed in data-constrained scenarios. Training from scratch within ExChanGeAI frequently yielded top results. This underscores the critical importance of empirical validation and careful model selection tailored to each dataset and use case. Second, the inherent influence of model architecture on performance, coupled with the relative consistency of the subsequent training process across architectures, underscores the value of an end-to-end platform that simplifies exploration and deployment of diverse, yet effective, models. Third, while the evaluation was conducted using data-constrained scenarios, rigorous validation across diverse external datasets and the baseline comparison model also outlines the advantage of more training samples; however, these may be difficult to obtain in a clinical setting. Fourth, while the given models can be trained outside the platform, with even more customization, the usage of ExChanGeAI reduces many technical burdens, facilitating faster deployment as it eliminates the need for code for data ingestion, preprocessing, training, and evaluation for new models. Fifth, all evaluations have been conducted with a 100 Hz sampling rate by default, which, according to multiple research papers, does not notably decrease the classification performance. However, downsampling may lead to a loss of high-frequency clinical details, such as fragmentation and notches. Researchers should be aware that this loss of fidelity may be critical for specific pathologies not covered in the current classification tasks, though the platform allows for higher sampling rates if required. Sixth, defining “revascularization” by treatment status serves as a proxy for actionable clinical need. While this implies that the label incorporates medical decision-making alongside pathology, predicting this outcome remains a clinically vital advancement for identifying patients requiring urgent intervention. Seventh, while the platform supports seamless deployment of ONNX-compatible architectures, we acknowledge that integrating foundation models with external dependencies or specific libraries (eg, ECG-FM) currently requires execution via external scripts rather than the native end-to-end platform. Eighth, the DSAIL SNU model replaces missing age/sex with default values and missing flag indicators, exactly as it was pretrained. If demographic data were available, its performance may improve beyond the results reported here. Finally, we contributed to the evaluation of the novel revascularization task using a stratified 20% hold-out subset of our EDMS cohort. Consequently, these new results can serve as an internal validation only, and the generalizability to external cohorts cannot be guaranteed.

Future Work

While acknowledging potential limitations for expert users seeking highly specialized customizations, the platform’s modular design allows for the future integration of additional compatible architectures, expanding its versatility. The main focus of possible future work could be the integration of explainable or interpretable ML, including its visualization for each prediction.

Conclusions

A major strength of ExChanGeAI is its ability to democratize advanced deep learning for ECG analysis. By integrating pretrained, fine-tuned, and untrained models within a unified interface, ExChanGeAI overcomes significant barriers associated with data loading, model-specific installation, environment setup, and code dependencies, particularly benefiting nonexperts and general-purpose applications. This not only enables rapid prototyping and empirical validation by both experts and nonexperts but also encourages open science and sharing of ready-to-use models for collaborative research. Ultimately, ExChanGeAI aims to enhance the accessibility of deep learning models and reduce operational overhead, facilitating broader adoption and accelerating progress. This approach not only minimizes human error and technical debt but also supports best practices for reproducible research and clinical validation. Limitations are mainly posed by the available data and infrastructure, even though the training, on modern machines, becomes significantly easier due to the large increase in computational power in recent years and wider adoption of specialized hardware, such as GPUs and neural processing units.

In conclusion, this work introduced ExChanGeAI, a novel open-source platform designed to streamline and democratize the application of deep learning for ECG analysis. Our results demonstrate the effectiveness of ExChanGeAI across both conventional and state-of-the-art deep learning models—even with limited data—and highlight that pretrained models are not always superior in data-constrained scenarios. Regular empirical benchmarking and model selection remain crucial. By promoting accessibility, reproducibility, and systematic model comparison, ExChanGeAI broadens participation in deep learning research and clinical adoption in ECG analysis.

Acknowledgments

This work was partially supported by the Interdisciplinary Centre for Clinical Research (IZKF) Münster, Germany (grant SEED/020/23 to AB). We acknowledge support by the Open-Access Publication Fund of the Medical Faculty of the Otto-von-Guericke University Magdeburg. The authors acknowledge the use of generative artificial intelligence (AI) technology (ChatGPT, version 4o, from OpenAI, 2024) only for editing the clarity and language refinement during manuscript preparation. The research idea, content, literature research, citations, methods, results, and conclusions were purely developed by the authors. The wording suggestions were reviewed and verified by the authors.

XSL•FO
RenderX

Correspondence and requests for materials should be addressed to JV. LB was affiliated with the Institute of Medical Informatics at the University of Münster at the time of the development of the platform and is currently affiliated with the Institute of Medical Data Science at the Otto-von-Guericke University Magdeburg.

Data Availability

The datasets analyzed during this study are available in the PhysioNet or SciDB repository [20,47-49]. The EDMS dataset analyzed during this study is not publicly available due to privacy restrictions. The ExChanGeAI code and dataset generated during this study are available in the GitHub repository [32]. Supplementary information is available for this paper.

Authors' Contributions

Conceptualization: LB, JV
Data curation: LB, AB
Formal analysis: LB, LP, LE
Methodology: LB
Project administration: JV
Software: LB
Resources: AB, JV
Supervision: LE, JV
Visualization: LB
Writing–original draft: LB
Writing–review & editing: LP, AB, LE, JV

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sample size and class distribution for training samples of PTB-XL (Physikalisch-Technischen Bundesanstalt-extra large [National Metrology Institute of Germany]) across defined classification targets.
[PDF File, 48 KB - jmir_v28i1e81116_app1.pdf]

Multimedia Appendix 2
Overview of dataset characteristics, including the total number of ECG recordings, age, and sex distribution for the PTB-XL (Physikalisch-Technischen Bundesanstalt-extra large [National Metrology Institute of Germany]), MIMIC-IV-ECG (Medical Information Mart for Intensive Care IV Electrocardiogram), Yang et al., and EDMS (Emergency Department Münster) datasets.
[PDF File, 60 KB - jmir_v28i1e81116_app2.pdf]

Multimedia Appendix 3
Extended classification and calibration metrics of the trained models on electrocardiogram classification tasks across the external datasets.
[ZIP File, 8 KB - jmir_v28i1e81116_app3.zip]

Multimedia Appendix 4
Trained and fine-tuned models using the ExChanGeAI platform based on the PTB-XL (Physikalisch-Technischen Bundesanstalt-extra large [National Metrology Institute of Germany]) dataset. Training conducted using default settings (commit number 4d862c04) to ensure reproducibility.
[ZIP File, 32127 KB - jmir_v28i1e81116_app4.zip]

References

1. Lin CS, Liu WT, Tsai DJ, et al. AI-enabled electrocardiography alert intervention and all-cause mortality: a pragmatic randomized clinical trial. Nat Med 2024 May;30(5):1461-1470. [doi: 10.1038/s41591-024-02961-4] [Medline: 38684860]
2. Adedinsewo DA, Morales-Lara AC, Afolabi BB, et al. Artificial intelligence guided screening for cardiomyopathies in an obstetric population: a pragmatic randomized clinical trial. Nat Med 2024 Oct;30(10):2897-2906. [doi: 10.1038/s41591-024-03243-9] [Medline: 39223284]
3. Sau A, Pastika L, Sieliwonczyk E, et al. Artificial intelligence-enabled electrocardiogram for mortality and cardiovascular risk estimation: a model development and validation study. Lancet Digit Health 2024 Nov;6(11):e791-e802. [doi: 10.1016/S2589-7500(24)00172-9] [Medline: 39455192]

XSL•FO
RenderX

4. Widder DG, Whittaker M, West SM. Why “open” AI systems are actually closed, and why this matters. Nature New Biol 2024 Nov;635(8040):827-833. [doi: 10.1038/s41586-024-08141-1] [Medline: 39604616]
5. Polevikov S. Advancing AI in healthcare: a comprehensive review of best practices. Clin Chim Acta 2023 Aug 1;548:117519. [doi: 10.1016/j.cca.2023.117519] [Medline: 37595864]
6. Rahman M, Victoros E, Ernest J, Davis R, Shanjana Y, Islam M. Impact of artificial intelligence (AI) technology in the healthcare sector: a critical evaluation of both sides of the coin. Clin Med Insights Pathol 2024 Jan;17. [doi: 10.1177/2632010X241226887]
7. Rauniyar A, Hagos DH, Jha D, et al. Federated learning for medical applications: a taxonomy, current trends, challenges, and future research directions. IEEE Internet Things J 2024;11(5):7374-7398. [doi: 10.1109/JIOT.2023.3329061]
8. Jang JH, Kim TY, Yoon D. Effectiveness of transfer learning for deep learning-based electrocardiogram analysis. Healthc Inform Res 2021 Jan;27(1):19-28. [doi: 10.4258/hir.2021.27.1.19] [Medline: 33611873]
9. Weimann K, Conrad TOF. Transfer learning for ECG classification. Sci Rep 2021 Mar 4;11(1):5251. [doi: 10.1038/s41598-021-84374-8] [Medline: 33664343]
10. Chato L, Regentova E. Survey of transfer learning approaches in the machine learning of digital health sensing data. J Pers Med 2023 Dec 12;13(12):1703. [doi: 10.3390/jpm13121703] [Medline: 38138930]
11. Avetisyan A, Tigranyan S, Asatryan A, et al. Deep neural networks generalization and fine-tuning for 12-lead ECG classification. Biomed Signal Process Control 2024 Jul;93:106160. [doi: 10.1016/j.bspc.2024.106160]
12. Kaplan Berkaya S, Uysal AK, Sora Gunal E, Ergin S, Gunal S, Gulmezoglu MB. A survey on ECG analysis. Biomed Signal Process Control 2018 May;43:216-235. [doi: 10.1016/j.bspc.2018.03.003]
13. Ebrahimi Z, Loni M, Daneshtalab M, Gharehbaghi A. A review on deep learning methods for ECG arrhythmia classification. Expert Syst Appl: X 2020 Sep;7:100033. [doi: 10.1016/j.eswax.2020.100033]
14. Qin K, Huang W, Zhang T, Zhang H, Cheng X. A lightweight SelfONN model for general ECG classification with pretraining. Biomed Signal Process Control 2024 Mar;89:105780. [doi: 10.1016/j.bspc.2023.105780]
15. Zhang S, Fang Y, Ren Y. ECG autoencoder based on low-rank attention. Sci Rep 2024;14(1):12823. [doi: 10.1038/s41598-024-63378-0]
16. Roy M, Majumder S, Halder A, Biswas U. ECG-NET: a deep LSTM autoencoder for detecting anomalous ECG. Eng Appl Artif Intell 2023 Sep;124:106484. [doi: 10.1016/j.engappai.2023.106484]
17. Thambawita V, Isaksen JL, Hicks SA, et al. DeepFake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine. Sci Rep 2021 Nov 9;11(1):21896. [doi: 10.1038/s41598-021-01295-2] [Medline: 34753975]
18. Singh P, Sharma A. Attention-based convolutional denoising autoencoder for two-lead ECG denoising and arrhythmia classification. IEEE Trans Instrum Meas 2022;71:1-10. [doi: 10.1109/TIM.2022.3197757]
19. Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet. Circulation 2000 Jun 13;101(23):e215-e220. [doi: 10.1161/01.CIR.101.23.e215]
20. Wagner P, Strodthoff N, Bousseljot RD, et al. PTB-XL, a large publicly available electrocardiography dataset. Sci Data 2020 May 25;7(1):154. [doi: 10.1038/s41597-020-0495-6] [Medline: 32451379]
21. Bickmann L, Plagwitz L, Varghese J. Post Hoc sample size estimation for deep learning architectures for ECG-classification. Stud Health Technol Inform 2023 May 18;302:182-186. [doi: 10.3233/SHTI230099] [Medline: 37203643]
22. Ismail Fawaz H, Lucas B, Forestier G, et al. InceptionTime: finding AlexNet for time series classification. Data Min Knowl Disc 2020 Nov;34(6):1936-1962. [doi: 10.1007/s10618-020-00710-y]
23. Rahimian E, Zabihi S, Atashzar SF, Asif A, Mohammadi A. XceptionTime: independent time-window xceptiontime architecture for hand gesture classification. Presented at: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 4-8, 2020; Barcelona, Spain p. 1304-1308. [doi: 10.1109/ICASSP40776.2020.9054586]
24. Han Y, Murino V, Liu X, Zhang X, Ding C. A systematic review on foundation models for electrocardiogram analysis: initial strides and expansive horizons. arXiv. Preprint posted online on Oct 13, 2025. [doi: 10.48550/arXiv.2410.19877]
25. Li J, Aguirre A, Moura J, et al. An electrocardiogram foundation model built on over 10 million recordings with external evaluation across multiple domains. arXiv. Preprint posted online on Aug 4, 2025. [doi: 10.48550/arXiv.2410.04133]
26. Wang Y, Cao X, Hu Y, et al. AnyECG: foundational models for multitask cardiac analysis in real-world settings. arXiv. Preprint posted online on Mar 3, 2025. [doi: 10.48550/arXiv.2411.17711]
27. Zhang S, Du Y, Wang W, et al. ECGFM: a foundation model for ECG analysis trained on a multi-center million-ECG dataset. Inf Fusion 2025 Dec;124:103363. [doi: 10.1016/j.inffus.2025.103363]
28. McKeen K, Masood S, Toma A, Rubin B, Wang B. ECG-FM: an open electrocardiogram foundation model. JAMIA Open 2025 Oct;8(5):ooaf122. [doi: 10.1093/jamiaopen/ooaf122] [Medline: 41113504]
29. Mathew G, Barbosa D, Prince J, Venkatraman S. Foundation models for cardiovascular disease detection via biosignals from digital stethoscopes. npj Cardiovasc Health 2024;1(1):1-13. [doi: 10.1038/s44325-024-00027-5]
30. Demšar J, Curk T, Erjavec A, et al. Orange: data mining toolbox in python. J Mach Learn Res 2013;14(1):2349-2353. [doi: 10.5555/2567709.2567736]

XSL•FO
RenderX

31. Gijsbers P, Vanschoren J. GAMA: Genetic automated machine learning assistant. J Open Source Softw 2019;4(33):1132. [doi: 10.21105/joss.01132]
32. Bickmann L. ExChanGeAI. GitHub. URL: https://github.com/VargheseLab/exchangeai [accessed 2025-12-24]
33. Makowski D, Pham T, Lau ZJ, et al. NeuroKit2: a Python toolbox for neurophysiological signal processing. Behav Res Methods 2021 Aug;53(4):1689-1696. [doi: 10.3758/s13428-020-01516-y] [Medline: 33528817]
34. Plagwitz L, Bickmann L, Fujarski M, et al. The Rlign algorithm for enhanced electrocardiogram analysis through R-peak alignment for explainable classification and clustering. arXiv. Preprint posted online on Aug 9, 2024. [doi: 10.48550/arXiv.2407.15555]
35. Salimi A, Kalmady SV, Hindle A, Zaiane O, Kaul P. Exploring best practices for ECG signal processing in machine learning. arXiv. Preprint posted online on May 14, 2025. [doi: 10.48550/arXiv.2311.04229]
36. Mehari T, Strodthoff N. Towards quantitative precision for ECG analysis: leveraging state space models, self-supervision and patient metadata. IEEE J Biomed Health Inform 2023 Nov;27(11):5326-5334. [doi: 10.1109/JBHI.2023.3310989] [Medline: 37656655]
37. Lee KS, Park HJ, Kim JE, et al. Compressed deep learning to classify arrhythmia in an embedded wearable device. Sensors (Basel) 2022 Feb 24;22(5):1776. [doi: 10.3390/s22051776] [Medline: 35270923]
38. Büscher A, Plagwitz L, Yildirim K, et al. Deep learning electrocardiogram model for risk stratification of coronary revascularization need in the emergency department. Eur Heart J 2025 Mar 29:ehaf254. [doi: 10.1093/eurheartj/ehaf254] [Medline: 40156923]
39. Reyna MA, Sadr N, Alday EAP, et al. Will two do? varying dimensions in electrocardiography: the physionet/computing in cardiology challenge 2021. Presented at: 2021 Computing in Cardiology (CinC); Sep 13-15, 2021; Brno, Czech Republic p. 1-4. [doi: 10.23919/CinC53138.2021.9662687]
40. Han H, Park S, Min S, et al. Towards high generalization performance on electrocardiogram classification. Presented at: 2021 Computing in Cardiology (CinC); Sep 13-15, 2021; Brno, Czech Republic p. 1-4. [doi: 10.23919/CinC53138.2021.9662737]
41. Ansel J, Yang E, He H, et al. PyTorch 2: faster machine learning through dynamic python bytecode transformation and graph compilation. 2024 Apr 27 Presented at: Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems; Apr 27, 2024. [doi: 10.1145/3620665.3640366]
42. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. 2019 Presented at: International Conference on Learning Representations; May 6-9, 2019 URL: https://openreview.net/forum?id=Bkg6RiCqY7 [accessed 2026-01-28]
43. Zhou P, Xie X, Lin Z, Yan S. Towards understanding convergence and generalization of AdamW. IEEE Trans Pattern Anal Mach Intell 2024 Sep;46(9):6486-6493. [doi: 10.1109/TPAMI.2024.3382294] [Medline: 38536692]
44. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv. Preprint posted online on Jan 30, 2017 URL: https://arxiv.org/abs/1412.6980 [accessed 2025-12-24] [doi: 10.48550/arXiv.1412.6980]
45. Li J, Yang X. A cyclical learning rate method in deep learning training. Presented at: 2020 International Conference on Computer, Information and Telecommunication Systems (CITS); Oct 5-7, 2020; Hangzhou, China p. 1-5. [doi: 10.1109/CITS49457.2020.9232482]
46. Strodthoff N, Wagner P, Schaeffter T, Samek W. Deep learning for ECG analysis: benchmarks and insights from PTB-XL. IEEE J Biomed Health Inform 2021 May;25(5):1519-1528. [doi: 10.1109/JBHI.2020.3022989] [Medline: 32903191]
47. Lai J, Tan H, Wang J, et al. Practical intelligent diagnostic algorithm for wearable 12-lead ECG via self-supervised learning on large-scale dataset. Nat Commun 2023 Jun 23;14(1):3741. [doi: 10.1038/s41467-023-39472-8] [Medline: 37353501]
48. Yang W, Feng Q. Offline test set of ECG multi-label classfication: Science Data Bank; 2023. [doi: 10.57760/sciencedb.07677]
49. Gow B, Pollard T, Nathanson LA, et al. MIMIC-IV-ECG: diagnostic electrocardiogram matched subset: PhysioNet. [doi: 10.13026/4nqg-sb35]
50. Frénay B, Verleysen M. Classification in the presence of label noise: a survey. IEEE Trans Neural Netw Learn Syst 2014 May;25(5):845-869. [doi: 10.1109/TNNLS.2013.2292894] [Medline: 24808033]
51. Reyna MA, Sadr N, Perez Alday EA, et al. Issues in the automated classification of multilead ecgs using heterogeneous labels and populations. Physiol Meas 2022 Aug 26;43(8):084001. [doi: 10.1088/1361-6579/ac79fd] [Medline: 35815673]
52. Martínez-Sellés M, Marina-Breysse M. Current and future use of artificial intelligence in electrocardiography. J Cardiovasc Dev Dis 2023 Apr 17;10(4):175. [doi: 10.3390/jcdd10040175] [Medline: 37103054]
53. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: still a ways to go. Sci Transl Med 2021 Mar 24;13(586):eabb1655. [doi: 10.1126/scitranslmed.abb1655] [Medline: 33762434]

Abbreviations

AI: artificial intelligence
BBB: bundle branch blocks
CLBBB: complete left bundle branch block
CNN: convolutional neural network

XSL•FO
RenderX

CRBBB: complete right bundle branch block
CV: coefficient of variation
DICOM: Digital Imaging and Communication in Medicine
DSAIL SNU: Data Science & Artificial Intelligence Laboratory Seoul National University
ECG: electrocardiogram
ECG-FM: Electrocardiogram Foundation Model
EDMS: Emergency Department Münster
FLOP: floating-point operations per second
GAMA: General Automated Machine Learning Assistant
GPU: graphics processing unit
HIPAA: Health Insurance Portability and Accountability Act
HYP: hypertrophy
***ICD-10*:** *International Classification of Diseases, 10th Revision*
ILBBB: incomplete left bundle branch block
MI: myocardial infarction
MIMIC-IV-ECG: Medical Information Mart for Intensive Care IV Electrocardiogram
MIT: Massachusetts Institute of Technology
ONNX: Open Neural Network Exchange
PT-XL: Physikalisch-Technischen Bundesanstalt-extra large
STTC: ST/T changes

Edited by A Coristine; submitted 23.Jul.2025; peer-reviewed by CH Li, T Ouyang, Y Yu; revised version received 01.Dec.2025; accepted 02.Dec.2025; published 30.Jan.2026.

Please cite as:
Bickmann L, Plagwitz L, Büscher A, Eckardt L, Varghese J
End-to-End Platform for Electrocardiogram Analysis and Model Fine-Tuning: Development and Validation Study
J Med Internet Res 2026;28:e81116
URL: https://www.jmir.org/2026/1/e81116
doi:10.2196/81116

© Lucas Bickmann, Lucas Plagwitz, Antonius Büscher, Lars Eckardt, Julian Varghese. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Predicting the Intention to Use Generative Artificial Intelligence for Health Information: Comparative Survey Study

Jörg Matthes[1], Prof Dr; Anne Reinhardt[2], DPhil; Selma Hodzic[1], MA, MSc; Jaroslava Kaňková[1], DPhil; Alice Binder[1], DPhil; Ljubisa Bojic[3,4], PhD; Helle Terkildsen Maindal[5], PhD; Corina Paraschiv[6], Prof Dr; Knud Ryom[5], PhD

[1]Department of Communication, University of Vienna, Waehringer Street 29, Vienna, Austria

[2]Department of Media and Communication, Ludwig-Maximilians-Universität München, Munich, Germany

[3]Digital Society Lab, Institute for Philosophy and Social Theory, University of Belgrade, Belgrade, Serbia

[4]Institute for Artificial Intelligence Research and Development of Serbia, Novi Sad, Serbia

[5]Department of Public Health, Aarhus University, Aarhus, Denmark

[6]Laboratoire Interdisciplinaire de Recherche Appliquée en Économie de la Santé (LIRAES), Université Paris Cité, Paris, France

Corresponding Author:
Jörg Matthes, Prof Dr
Department of Communication, University of Vienna, Waehringer Street 29, Vienna, Austria

Abstract

Background: The rise of generative artificial intelligence (AI) tools such as ChatGPT is rapidly transforming how people access information online. In the health context, generative AI is seen as a potentially disruptive information source due to its low entry barriers, conversational style, and ability to tailor content to users' needs. However, little is known about whether and how individuals use generative AI for health purposes, and which groups may benefit or be left behind, raising important questions of digital health equity.

Objective: This study aimed to assess the current relevance of generative AI as a health information source and to identify key factors predicting individuals' intention to use it. We applied the Unified Theory of Acceptance and Use of Technology 2, focusing on 6 core predictors: performance expectancy, effort expectancy, facilitating conditions, social influence, habit, and hedonic motivation. In addition, we extended the model by including health literacy and health status. A cross-national design enabled comparison across 4 European countries.

Methods: A representative online survey was conducted in September 2024 with 1990 participants aged 16 to 74 years from Austria (n=502), Denmark (n=507), France (n=498), and Serbia (n=483). Structural equation modeling with metric measurement invariance was used to test associations across countries.

Results: Usage of generative AI for health information was still limited: only 39.5% of respondents reported having used it at least rarely. Generative AI ranked last among all measured health information sources (mean 2.08, SD 1.66); instead, medical experts (mean 4.77, SD 1.70) and online search engines (mean 4.57, SD 1.88) are still the most frequently used health information sources. Despite this, performance expectancy (b range=0.44-0.53; all P<.001), habit (b range=0.28-0.32; all P<.001), and hedonic motivation (b range=0.22-0.45; all P<.001) consistently predicted behavioral intention in all countries. Facilitating conditions also showed small but significant effects (b range=0.12-0.24; all P<.01). In contrast, effort expectancy, social influence, health literacy, and health status were unrelated to intention in all countries, with one marginal exception (France: health status, b=−0.09; P=.007). Model fit was good (comparative fit index=0.95; root mean square error of approximation=0.03), and metric invariance was confirmed.

Conclusions: Generative AI use for health information is currently driven by early adopters—those who find it useful, easy to integrate, enjoyable, and have the necessary skills and infrastructure to do so. Cross-national consistency suggests a shared adoption pattern across Europe. To promote equitable adoption, communication efforts should focus on usefulness, convenience, and enjoyment, while ensuring digital access and safeguards for vulnerable users.

(J Med Internet Res 2026;28:e75648) doi:10.2196/75648

KEYWORDS

generative AI; artificial intelligence; health information–seeking; UTAUT2; Unified Theory of Acceptance and Use of Technology 2; AI adoption

XSL•FO
RenderX

Introduction

Background

When people look for health information today, they no longer only consult physicians, pharmacists, or search engines. Increasingly, they also encounter generative artificial intelligence (AI) tools such as ChatGPT or the World Health Organization (WHO)'s chatbot Sarah, which simulate human-like conversations and provide instant responses. These tools promise a new way of accessing medical knowledge: fast, convenient, and interactive. At first glance, this accessibility seems to hold great potential for reducing barriers to health information, therefore directly impacting digital health equity—defined as equitable access to and use of digital health information technology that supports informed decision-making and enhances health [1].

However, the picture is more complex. On the one hand, generative AI can offer cost-free entry points (eg, basic versions of ChatGPT or automatically displayed answers in Google search via Google's Gemini), deliver content in multiple languages, and rephrase complex medical concepts into more understandable terms. In doing so, it could strengthen patient education, address health inequalities, and help bridge communication gaps between citizens and health care providers [2,3]. On the other hand, effective use still depends on internet-enabled devices and adequate digital skills, which are not equally distributed. As a result, the very technology that appears open and inclusive may also risk exacerbating existing digital divides [4].

Moreover, unlike other types of information, health-related questions are often sensitive and personal. At the same time, the inner workings of generative AI remain opaque, and the accuracy of its outputs is not guaranteed [5]. All these tensions raise important questions about adoption: Who is most likely to turn to generative AI for health information, and what factors shape this intention? Moreover, since health communication practices and digital infrastructures differ across countries, cross-national research is urgently needed.

To address these questions, this study draws on the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) [6]. The model proposes that performance expectancy, effort expectancy, facilitating conditions, social influence, habit, and hedonic motivation shape technology use. We extend this framework by also examining the roles of health literacy and health status in predicting intention to use generative AI for health information. Using cross-national survey data from Austria, Denmark, France, and Serbia, we investigate the drivers of adoption to shed light on both individual and contextual factors that may guide the diffusion of generative AI in health contexts.

Generative AI as Novel Health Information Source

Generative AI constitutes a potentially disruptive force in the health information ecosystem [7]. However, despite its rapid advancement and widespread availability, empirical research on its role in health information–seeking remains limited [2]. At the same time, broader trends highlight the ongoing digitalization of health-related knowledge acquisition. A representative survey conducted in Germany in 2019 revealed that only 48% of respondents consulted a medical professional for their most recent health issue, while 1 in 3 turned first to the internet [8]. Similar findings show that online sources—particularly search engines—are the primary means of accessing health information, both for caregivers and the general population [9,10]. Family and friends and traditional mass media (eg, print media and health-related TV programming) rank behind medical professionals and online sources [8].

The introduction of generative AI tools like ChatGPT may shift these established hierarchies. Unlike static web content or conventional search engines, generative AI enables dialogic, personalized interactions that simulate human conversation. These features may position generative AI as a compelling alternative to established online and offline health information sources. However, current evidence suggests that trust in generative AI—especially regarding complex health-related issues—is still limited [11], which might restrict its present adoption potential to early adopters [2]. This raises questions about how generative AI integrates into the broader ecosystem of health information sources.

To address this gap, we first explore: How does the use of generative AI for health information–seeking compare to that of more established health information sources?

Explaining Predictors of Technology Adoption: UTAUT2

The UTAUT2 [6] is one of the most popular models to explain technology adoption. It builds on the technology acceptance model [12], emphasizing perceived usefulness and ease of use, and the initial UTAUT model [13], which added performance expectancy, effort expectancy, facilitating conditions, and social influence as predictors of adoption behavior. UTAUT2 extends these frameworks to consumer contexts by incorporating hedonic motivation and habit [6,14] . The UTAUT2 model has demonstrated its versatility in explaining the adoption of diverse eHealth technologies, such as wearable devices [15], health websites [16], and health apps [17]. Additionally, recent studies have highlighted its relevance in understanding the uptake of generative AI technologies [18-21], showcasing its capacity to extend beyond traditional eHealth domains. However, so far, studies on predictors of usage intentions in the context of AI health information–seeking are lacking.

Performance expectancy, a central construct in the UTAUT2 framework, reflects the belief that using technology will lead to performance benefits [22]. In the context of health information–seeking using generative AI, performance expectancy is shaped by users' perceptions of how effectively these tools can enhance their own life, including aspects such as health–decision-making and task efficiency [23]. Consequently, as users anticipate greater usefulness from adopting generative AI as a health information source, their intention to use such technologies strengthens [18-20,24]. Based on this, we propose the following hypothesis: "the higher the performance expectancy, the stronger the intention to use generative AI for health information–seeking" (H1). Effort

XSL•FO
RenderX

expectancy, closely tied to ease of use, emphasizes simplicity in technology adoption [13]. Generative AI tools like ChatGPT benefit from high effort expectancy when users find them intuitive and easy to integrate into their workflows, particularly during the early adoption phase [18,19,23,24]. Addressing usability concerns early can reduce resistance and build user confidence, strengthening behavioral intention [23,25]. Therefore, we propose that "the higher the effort expectancy, the stronger the intention to use generative AI for health information–seeking" (H2). Facilitating conditions refer to the resources, skills, and support necessary for using technology [22]. These include training, knowledge, technical assistance, and system compatibility, which significantly enhance behavioral intention and usage [6,25]. In technologically mature settings, facilitating conditions are critical for sustained adoption and user satisfaction [26]. In line with the UTAUT2, we hypothesize that "the better the facilitating conditions, the stronger the intention to use generative AI for health information–seeking" (H3). Social influence indicates the perception that peers, such as family, friends, or colleagues, believe one should adopt a technology [22]. It plays a crucial role in early adoption, where external validation often outweighs personal experience [6]. Positive reinforcement within social or professional networks can normalize usage [18,24,25,27]: If people perceive that their peers already use generative AI for health information–seeking, their own intention to do so might increase as well. We, therefore, propose that "the greater the perceived social influence, the stronger the intention to use generative AI for health information–seeking" (H4). Habit specifies the extent to which behavior becomes automatic through repetition and prior use [26]. It strongly influences behavioral intention and long-term adoption, emphasizing the importance of regular engagement with technology [6,28]. For generative AI as a health information source, fostering habitual use can solidify its integration into daily routines and enhance sustained adoption [25]. This leads us to state, "the more it is a habit to use generative AI, the stronger the intention to use generative AI for health information–seeking" (H5). Hedonic motivation refers to the enjoyment or pleasure derived from using technology, particularly relevant in consumer contexts [26]. It directly impacts behavioral intention, especially for technologies involving entertainment or leisure [29]. For generative AI like ChatGPT, it can be expected that the interaction is perceived as fun or entertaining, which can boost user engagement and drive adoption [30]. Accordingly, we suggest the following hypothesis: "the higher the hedonic motivation, the stronger the intention to use generative AI for health information–seeking" (H6).

Influence of Health Literacy and Health Status

With the growing integration of digital tools into everyday lives, the role of health literacy in online health information–seeking has garnered increasing attention. Health literacy has been conceptualized as an individual's capacity to search, access, comprehend, and critically evaluate health information, as well as to use the acquired knowledge to effectively address health-related issues [31,32]. Digital health literacy refers to these abilities in the context of digital environments [33-35]. Generally, low health literacy scores have been associated with undesirable health outcomes [36].

Research suggests that low levels of health literacy are associated with decreased trust in online health resources [37], including the outputs of AI tools [38], and lower overall adoption of online health technologies [39]. Furthermore, initial studies indicate a positive association between health literacy levels and attitudes toward the use of AI tools for medical consultations [40].

On the other hand, individuals with higher levels of health literacy are generally better equipped to critically evaluate online health information and scrutinize it in greater detail [37,41]. This heightened evaluative capacity could make them more aware of the limitations and potential risks of generative AI outputs, such as inaccurate information, bias, data privacy concerns, or oversimplified medical advice [42]. Moreover, individuals with higher health literacy are more likely to trust and use high-quality medical online resources, whereas those with limited health literacy prefer accessible but potentially less reliable sources [43]. In this context, outputs from generative AI might be perceived as lower-quality sources by highly digital health–literate individuals. As a result, while higher health literacy could foster openness to using generative AI for health purposes, it might also lead to greater skepticism or hesitancy in relying on these tools. Nonetheless, there is not enough research in the context of generative AI specifically to make conclusive predictions.

Another well-established factor in online health information–seeking, yet underexplored in the context of AI, is individuals' health status: Studies suggest that people with poor health are significantly more likely to consult the internet for health information compared to those with good health [44,45]. Being chronically ill has also been associated with increased reliance on internet-based technologies for health-related purposes [46]. This relationship can be explained by the fact that individuals in poor health often experience greater health-related concerns, which in turn heightens their motivation to seek information online.

Given these complex relationships, we propose the second research question: How does health literacy and health status influence the intention to seek health information using generative AI?

Cross-National Comparison

In this study, we investigate the predictors of generative AI adoption for health information–seeking across 4 European countries: Austria, Denmark, France, and Serbia. While these countries share certain similarities, they also display notable differences that could shape the strength of the UTAUT2 predictors on the intention to use generative AI for health purposes. Thus, this cross-national approach ensures that the observed effects are generalizable and not confined to specific national contexts or unique country conditions.

The selected countries share two key characteristics. First, all 4 countries provide universal health coverage, ensuring broad access to health care services for their populations. Second, a

XSL•FO
RenderX

significant portion of health care expenditure in these countries is publicly funded [47-50].

Despite these commonalities, there are also critical factors that differ among the countries and may shape the predictors of generative AI adoption. On the one hand, variations in digital infrastructure could significantly impact facilitating conditions, effort expectancy, and social influence as predictors of generative AI use. Denmark consistently ranks among Europe's most digitally advanced nations, boasting high internet penetration and widespread adoption of e-health solutions [51]. This strong digital ecosystem likely enhances the perceived ease of use and social endorsement of generative AI. In contrast, Austria, France, and Serbia exhibit more moderate levels of digital adoption in the context of health information, which may limit the perceived use and social norms regarding such technologies [51].

On the other hand, access to and trust in health care providers vary significantly across these countries, potentially influencing performance expectancy and social influence. In nations with robust health care systems—characterized by a high availability of medical professionals and easy access to care—individuals are more likely to rely on doctors for health advice, as they are often viewed as the most trusted source of health information [8]. Denmark exemplifies this with its high levels of public trust in the health care system [52], which may reduce the perceived benefits and social norms around using generative AI for health purposes. Conversely, in western Balkan countries like Serbia, studies report generally low levels of trust in the health care system [53]. In such contexts, individuals may be more inclined to seek alternative information sources, potentially amplifying the perceived benefits of generative AI use.

By examining these diverse national contexts, this study not only tests the universality of the UTAUT2 model but also deepens our understanding of the contextual factors that shape generative AI adoption for health purposes. We ask: How do the predictors of generative AI use for health information–seeking differ across Austria, Denmark, France, and Serbia?

Methods

Ethical Considerations

Before data collection, the study received ethical approval from the institutional review board of the Department of Communication at the University of Vienna (approval ID: 1205). All participants provided written informed consent prior to participating in the study. The data were collected in anonymized form and no personal identifiers were recorded or stored. Participants received a compensation of €1.50 (US $1.74) for completing the study through the panel provider.

Recruitment

Recruitment of participants occurred during September 2024 via *Bilendi*, an international panel provider. *Bilendi* recruited the participants via email. The panel is checked for quality and attendance on a regular basis. The study was conducted in Austria, France, Denmark, and Serbia, with participants randomly selected to achieve samples representative of age, gender, and educational background. The provider's panel sizes in the respective countries were as follows: Austria: n=60,000; Denmark: n=90,000; France: n=815,000; and Serbia: n=15,100. Per country, the study aimed to reach 500 participants.

Inclusion criteria required participants to be aged between 16 and 74 years. Additionally, respondents who completed the survey in less than one-third of the median completion time (speeders) were excluded. Completion rates (excluding screened-out participants) were high across all 4 countries, ranging from 84.3% to 89.8%. Further details on survey design, administration, and response rates are provided in the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist (Checklist 1).

Procedure and Measures

Overview

The study consisted of two components. The first component, a survey, investigated predictors of the intention to use generative AI for health information–seeking. The second component, an experimental study, explored the influence of disease-related factors on these intentions [54]. To ensure respondents shared a common understanding of the concept, the survey began with a short definition of "generative AI," describing it as technologies that engage in natural language conversations with users and generate responses in real-time. Examples such as ChatGPT, Google's Gemini, and Microsoft Copilot were provided.

The original questionnaire was developed in English and subsequently translated into German, French, Danish, and Serbian. Each translation was performed by a bilingual team member and back-translated into English by a different native speaker to ensure conceptual equivalence with the original items.

This study focused on the following constructs (a complete item list with descriptive analysis and construct reliability values can be found in Multimedia Appendix 1).

Dependent Variables

Sources of Health Information–Seeking

Participants rated the frequency with which they use 8 health information [55] sources on a 7-point Likert scale (1 = "never" to 7 = "very often"). The sources included conversations with medical professionals, pharmacists, and family or friends, as well as books, mass media, internet search engines (eg, Google and Ecosia), and generative AI.

AI Usage Intentions

Participants' willingness to use generative AI [25] for health information–seeking was assessed using 3 items (eg, "I intend to use generative AI for health information seeking"), rated on a 7-point Likert scale (1 = "strongly disagree" to 7 = "strongly agree").

UTAUT2 Predictor Variables

All UTAUT2 model [6,25,56,57] predictors were measured on a 7-point Likert scale (1 = "strongly disagree" to 7 = "strongly agree"). The predictors have been described below.

XSL•FO
RenderX

Performance Expectancy

Perceived benefits of using generative AI for health information–seeking were measured with 4 items (eg, "Using generative AI would save me time when researching health topics").

Effort Expectancy

The perceived ease of using generative AI as a health information source was assessed with 4 items (eg, "Learning to use generative AI for health information–seeking seems easy for me").

Social Influence

Three items measured the extent to which participants felt that important others encouraged their use of generative AI for health information–seeking (eg, "People who are important to me think that I should use generative AI for health-information seeking").

Hedonic Motivation

The enjoyment of using generative AI for health information–seeking was assessed with 3 items (eg, "I think using generative AI for health-information seeking could be fun").

Facilitating Conditions

Participants' perceptions of available resources and support for using generative AI to seek health information were measured with 4 items (eg, access to devices and reliable internet, and knowledge).

Habit

The extent to which turning to generative AI when seeking health information had become a habitual behavior was measured with 3 items (eg, "I automatically turn to generative AI whenever I have questions about my health").

Model Extension Variables

Health Literacy

Health literacy [58] was assessed with 10 items, asking participants to rate their confidence in tasks such as finding understandable health information. Responses were recorded on a 4-point Likert scale (1 = "not at all true" to 4 = "absolutely true").

Health Status

We measured participants' health status using 1 item ("How would you describe your current health status?"; 1 = "very poor" to 7 = "very good").

Control Variables: Sociodemographic Variables

We further measured participants' age, gender, and educational level.

Statistical Analysis

Power

An a priori power analysis was conducted to determine the required sample size for structural equation modeling. Assuming an anticipated effect size of 0.25, a desired statistical power of 0.95, and a significance level of .05, the analysis indicated that a minimum of 391 participants per country would be necessary to detect the hypothesized effects [59].

Analytical Plan

We used AMOS version 26 (IBM Corp) to run a latent variable, multigroup structural equation model using a maximum-likelihood estimator with full information. We computed the comparative fit index, the Tucker-Lewis Index, the chi-square to degrees of freedom ratio (χ^2/df), and the root mean square error of approximation. We also secured metric measurement invariance to be able to compare the paths across countries. We controlled for age, gender, and education (binary coded).

Results

User Statistics

In total, data were collected from 1990 respondents, comprising 502 from Austria, 507 from Denmark, 498 from France, and 483 from Serbia. The overall mean age of participants was 45.1 (SD 15.7) years, with 50.2% (n=998) identifying as female participants. In terms of educational attainment, 83.8% (n=1634) of the sample reported completing at least a medium or higher level of education (secondary level II or higher). Furthermore, 87.4% (n=787) of respondents indicated prior use of generative AI for health information–seeking (at least rarely). Detailed demographic and background characteristics of the sample are summarized in Table 1.

XSL•FO
RenderX

Table . Descriptive characteristics of survey respondents from Austria, Denmark, France, and Serbia (N=1990; September 2024).

Demographic characteristics	Overall, n (%)	Austria, n (%)	Denmark, n (%)	France, n (%)	Serbia, n (%)
Education[a]					
Secondary I or lower	356 (18.24)	93 (18.6)	136 (26.9)	109 (21.9)	18 (3.7)
Secondary II	1080 (55.36)	303 (60.3)	179 (35.3)	224 (45.0)	374 (77.4)
Tertiary	554 (28.39)	106 (21.1)	192 (37.9)	165 (33.1)	91 (18.8)
Gender					
Female	998 (50.15)	252 (49.8)	251 (49.5)	256 (51.4)	239 (49.5)
Male	992 (49.85)	250 (50.2)	256 (50.5)	242 (48.6)	244 (50.5)
Prior experience[b]					
No	1203 (60.45)	316 (62.9)	328 (64.7)	326 (65.5)	233 (48.2)
Yes	787 (39.54)	186 (37.1)	179 (35.3)	172 (34.5)	250 (51.8)

[a]Educational attainment was categorized as low (secondary level I or below) and medium or high (secondary level II or higher). In Serbia, however, representativeness was achieved by grouping educational levels into low or medium (secondary level II or below) and high (tertiary education) due to sampling limitations.

[b]Prior experience: no = “I have never used Generative AI for health information seeking” and yes = “I have used Generative AI for health information seeking at least rarely.”

Statistical tests revealed no significant differences in gender distribution across countries (χ^2_3=0.48; *P*=.92) and no significant differences in age (Kruskal-Wallis χ^2_3=2.15; *P*=.54). In contrast, educational attainment varied significantly between countries (χ^2_3=550.76; *P*<.001), reflecting sampling-related imbalances in Serbia where low versus medium or high education was assessed differently than in the other countries. Finally, prior experience with health information–seeking showed significant country differences (Kruskal-Wallis χ^2_3=30.95; *P*<.001), with higher levels reported in Serbia.

Evaluation Outcomes

Descriptive Analysis

In our first research question, we explored how generative AI compares to more established health information sources in terms of usage frequency across countries. As illustrated in Figure 1, generative AI ranks last among all measured sources, indicating that, as of autumn 2024, it is rarely used for health information–seeking (mean 2.08, SD 1.66). In stark contrast, online search engines like Google are highly used, ranking second with a mean usage frequency of 4.57 (SD 1.88), following conversations with physicians, which hold the top position (4.77, SD 1.70). Family and friends also play a significant role, ranking third (4.27, SD 1.73), alongside pharmacists (3.52, SD 1.81). In comparison, traditional mass media such as TV, newspapers, and magazines are used less frequently (2.74, SD 1.68), as are books (2.68, SD 1.70) and free magazines provided by pharmacies or health insurance companies (2.60, SD 1.65). The relative ranking of information sources was consistent across all 4 countries, with physicians, internet search engines, and family or friends occupying the top positions and generative AI ranking last. However, some variation in mean usage frequencies was observed between countries; detailed country-level results are presented in Multimedia Appendix 2.

XSL•FO
RenderX

Figure 1. Mean usage frequency of different health information sources among survey respondents (N=1990) in Austria, Denmark, France, and Serbia (95% CI; September 2024). AI: artificial intelligence.

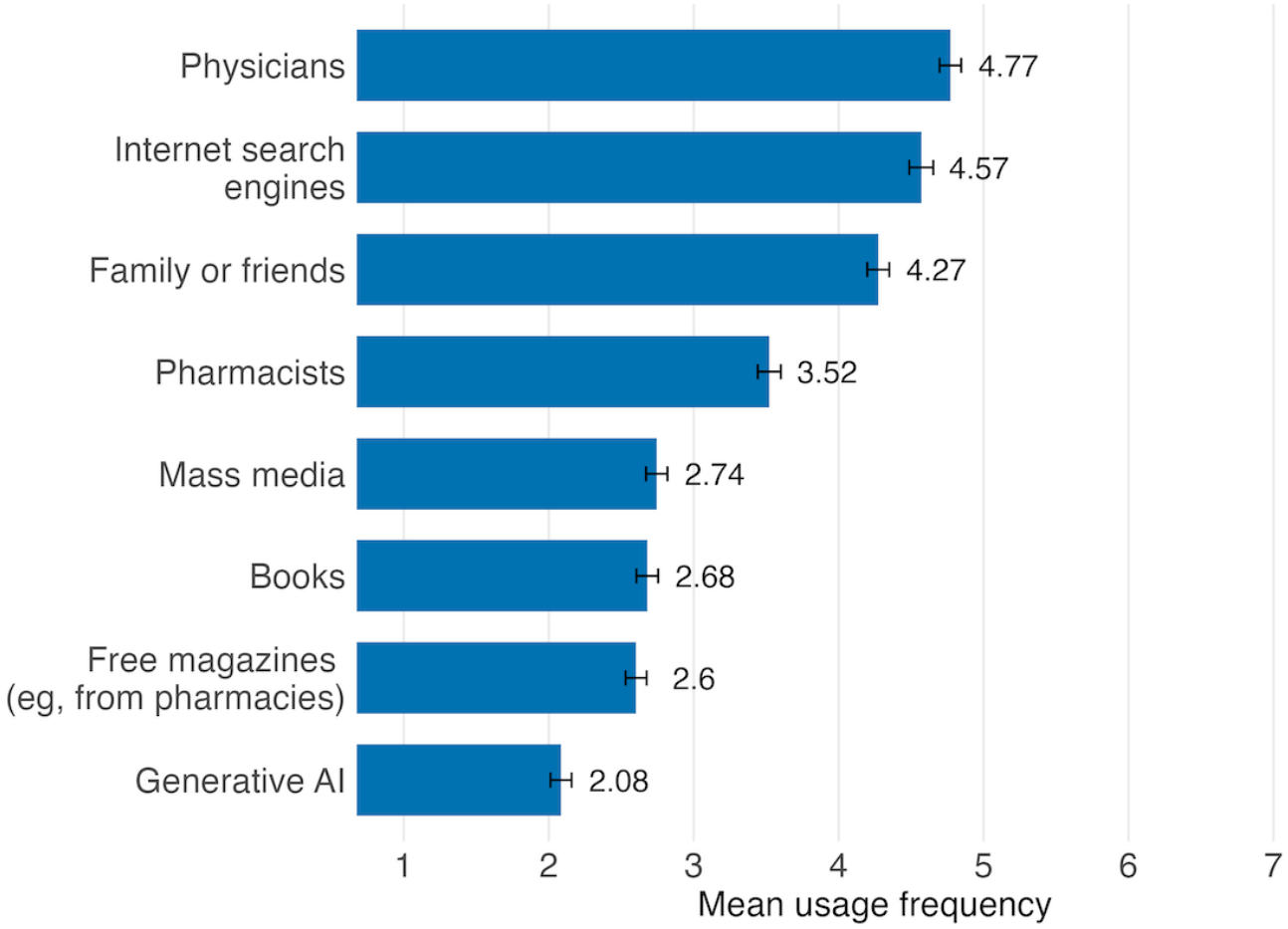

Model Evaluation

For the hypothesis tests, the results are shown in Table 2. Model fit was good (comparative fit index=0.95; Tucker-Lewis-Index=.93; χ^2/df =2.47, *P*<.001; root mean square error of approximation=0.03, 95% CI 0.03-0.03). We examined the metric measurement invariance of all latent variables by constraining all factor loadings as equal for all 4 countries. When comparing the constrained model to the unconstrained model, we found no significant difference in model fit (*P*=.16). Thus, metric invariance across countries was established.

XSL•FO
RenderX

Table . Structural equation model predicting the intention to use generative artificial intelligence (AI) for health information among survey respondents in Austria, Denmark, France, and Serbia (N=1990; September 2024).

Predictor variables	Austria[a]			Denmark[b]			France[c]			Serbia[d]		
	b	SE	*P* value	b	SE	*P* value	b	SE	*P* value	b	SE	*P* value
Performance expectancy	0.47	0.05	<.001	0.52	0.05	<.001	.053	0.05	<.001	0.44	0.05	<.001
Effort expectancy	−0.07	0.05	.20	0.03	0.05	.54	−0.11	0.05	.04	-0.02	0.06	.77
Facilitating conditions	0.12	0.04	.01	0.17	0.05	<.001	0.22	0.05	<.001	0.24	0.05	<.001
Social influence	−0.05	0.04	.29	−0.08	0.05	.17	−0.09	0.05	.10	−0.05	0.04	.27
Habit	0.29	0.04	<.001	0.32	0.04	<.001	0.28	0.05	<.001	0.28	0.04	<.001
Hedonic motivation	0.45[e]	0.06	<.001	0.22[f]	0.05	<.001	0.33	0.05	<.001	0.23[f]	0.05	<.001
Health literacy	−0.004	0.09	.97	0.04	0.10	.67	−0.02	0.08	.08	0.08	0.10	.40
Health status	−0.002	0.03	.95	0.02	0.03	.61	−0.09	0.03	.01	−0.05	0.03	.13

[a]Explained variance=0.84.

[b]Explained variance=0.80.

[c]Explained variance=0.86.

[d]Explained variance=0.79.

[e]The different subscripts in each row indicate a significant difference between paths (*P*<.05)

[f]The different subscripts in each row indicate a significant difference between paths (*P*<.05)

In line with H1, we found a highly significant positive association between performance expectancy and the intention to use generative AI for health information–seeking across all 4 countries (Austria: b=0.47, *P*<.001; Denmark: b=0.52, *P*<.001; France: b=0.53, *P*<.001; Serbia: b=0.44, *P*<.001). In contrast, H2 was not supported: effort expectancy showed no significant association with behavioral intention in any of the countries. Turning to H3, results revealed a positive association between facilitating conditions and the intention to use generative AI as a health information source, consistently observed across all 4 contexts (Austria: b=0.12, *P*=.005; Denmark: b=0.17, *P*<.001; France: b=0.22, *P*<.001; Serbia: b=0.24, *P*<.001). By contrast, no support was found for H4: perceived social influence was unrelated to behavioral intention in any of the countries. As predicted in H5, habit was positively associated with behavioral intention to use generative AI for health information–seeking throughout the sample (Austria: *b*=0.29, *P*<.001; Denmark: b=0.32, *P*<.001; France: b=0.28, *P*<.001; Serbia: *b*=0.28, *P*<.001). A similar pattern emerged for H6: hedonic motivation was significantly positively related to behavioral intention in all countries (Austria: b=0.45, *P*<.001; Denmark: b=0.22, *P*<.001; France: b=0.33, *P*<.001; Serbia: b=0.23, *P*<.001).

Finally, with regard to our second research question—which examined whether health literacy and health status predict the intention to seek health information using generative AI—we found no substantial associations. Only in France did health status show a marginal negative effect (b=−0.09; *P*=.007).

Discussion

Principal Results

This study investigated the predictors of intention to use generative AI for health information–seeking, drawing on the UTAUT2 framework and expanding it with health literacy and health status. Using cross-national survey data from Austria, Denmark, France, and Serbia, our findings show that generative AI is still only rarely used for health information–seeking. At the same time, performance expectancy, facilitating conditions, habit, and hedonic motivation consistently emerged as significant predictors of behavioral intention, whereas effort expectancy, social influence, health literacy, and health status were not related to intention. These patterns were consistent across all 4 countries, suggesting a robust set of psychological drivers underlying the early adoption of generative AI in health contexts. A detailed examination of these findings is provided as follows.

First, with regard to overall usage patterns, the data shows that generative AI currently plays only a minor role in health information–seeking: 60% of the respondents reported never

XSL•FO
RenderX

having used a generative AI tool for health-related questions. This result lends itself to 2 contrasting interpretations.

On the one hand, it challenges the popular narrative that generative AI is rapidly transforming health information–seeking behavior. Instead, the findings align with previous studies, showing that generative AI is currently infrequently used in the context of health information [2]. Traditional sources—such as medical professionals and search engines—continue to dominate [8], underscoring that generative AI has yet to achieve mainstream adoption.

On the other hand, despite persistent concerns about data privacy, algorithmic bias, and accuracy, it is noteworthy that 40% of the respondents have already experimented with generative AI for health purposes. Given that this technology only became widely accessible relatively recently, such early uptake is remarkable. From the perspective of technology adoption models, such as the Rogers Diffusion of Innovations [60], this pattern is characteristic of early adopters. It is therefore plausible to assume that the use of generative AI for health information–seeking will increase further as the technology matures and moves toward mainstream adoption.

To better understand the drivers of future uptake, we applied an extended version of the UTAUT2 model. Our findings confirmed the predictive power of performance expectancy, facilitating conditions, habit, and hedonic motivation. This aligns with prior research on digital health tools, indicating that users value usefulness, access, familiarity, and enjoyment [18,20,24].

In detail, the results show that performance expectancy—the perceived usefulness of the technology—had a strong positive effect on behavioral intention in all four countries. This finding suggests that the more respondents believe generative AI is useful to manage health-related questions, the more they will use it. Thus, if public health stakeholders or developers aim to encourage responsible AI use, they should emphasize the tangible benefits of generative AI, such as 24/7 availability, rapid response times, and the potential for personalized information. Perceived usefulness may also be fostered when individuals try out generative AI for the first time, that is, they learn that they can benefit from the technology.

At the same time, our study challenges established UTAUT2 assumptions. Effort expectancy, often seen as central to technology adoption, was not a relevant factor—possibly due to the intuitive nature of generative AI tools and the ubiquity of basic digital skills [61]. Using generative AI does not require any specific background knowledge beyond opening a webpage. Since online search engines are already the most frequently used health source, the basic skills needed for generative AI are widely present, potentially rendering effort expectancy less decisive.

Taken together, this emerging pattern—the strong effect of performance expectancy and the null effect of effort expectancy—underscores the distinction between usefulness and usability, which are closely related but not identical [62]. Usability refers to the ease of interacting with a system (eg, ease of learning and error prevention), whereas usefulness (utility) captures whether the system provides the functions and information that users actually need. Our findings suggest that in health contexts, utility is the decisive factor: people intend to use generative AI if its outputs are perceived as useful, while usability-related aspects appear less influential.

Importantly, this does not mean that barriers to adoption are absent. Rather, our findings show that they lie not in usability but in facilitating conditions—the structural and contextual resources that enable technology use. Across all countries, the availability of digital infrastructure, devices, and basic knowledge significantly shaped behavioral intention. In other words, while generative AI may be easy to use once accessed, unequal access to the necessary resources continues to pose a substantial adoption barrier. Consequently, facilitating conditions emerge as a key digital health equity concern [4]. Without adequate access, disadvantaged populations may be excluded from benefiting from generative AI, meaning that the technology risks widening rather than narrowing the digital divide in health information–seeking.

We also found that social influence—an important predictor in other studies on AI uptake [18,24]—did not play a meaningful role in shaping behavioral intention. This suggests that health-related information search is a rather personal topic, and that individuals may not always be willing to disclose what kind of information they are looking for. As a result, the intention to use generative AI for health information–seeking is largely independent of peer opinions or social norms.

In contrast, habit consistently predicted behavioral intention across all countries. From this finding, we may conclude that generative AI use for health information is likely to occur automatically, similar to how people use search engines. When individuals feel familiar with a technology, they are more likely to rely on it without conscious deliberation. However, this finding should be interpreted with caution, as the majority of participants had never used generative AI for health purposes. Much of the variance in habit may therefore reflect mere use versus nonuse. Accordingly, variables capturing initial adoption should be clearly distinguished from those measuring habit.

By including health literacy and health status as additional predictors, our study adds a novel dimension to existing research. In contrast to studies showing direct paths between these constructs and online health information–seeking [37,40,46], we found no such association for AI health information–seeking. However, each of these findings carries different implications. First, the absence of a significant association between health literacy and intention indicates that individuals' ability to understand and evaluate health information was not related to whether they reported turning to generative AI. This finding may suggest that the use of such tools is driven less by informed decision-making and more by general curiosity or interest in new technologies. Importantly, this raises concerns: people with lower health literacy may be just as likely to consult generative AI as those with higher health literacy—despite being less equipped to critically assess its outputs. Given the known risk of AI hallucinations—fabricated or inaccurate information presented in a confident tone [63]—this could lead to misinformation and, in the worst case,

XSL•FO
RenderX

harmful health decisions, as users with limited health literacy might find it difficult to distinguish between reliable and misleading content.

Second, the lack of an association between self-reported health status and intention suggests that the current use of generative AI is not primarily driven by medical need or urgency. People do not seem more likely to consult generative AI when facing a health problem; rather, usage may occur proactively or even recreationally. This challenges assumptions that such tools are primarily used in response to a health issue, and it underscores the importance of understanding user motivations beyond immediate health concerns.

Importantly, these patterns were largely consistent across all 4 countries, as confirmed by the measurement-invariant structural model. This cross-national consistency suggests that the psychological drivers of generative AI adoption in health contexts may transcend national boundaries and cultural differences, pointing to a universal set of adoption mechanisms.

Limitations

Several limitations should be acknowledged. First, due to the cross-sectional nature of this study, no causal conclusions can be drawn. Future research should therefore aim to replicate these findings using experimental or longitudinal designs. Second, we relied on self-reported data, which may be subject to social desirability bias. The use of behavioral data is thus warranted to validate these findings. Third, including additional predictors—such as individual differences or specific concerns—could provide deeper insights into the use of generative AI. Fourth, our comparative findings are based on data from only 4 countries, which limits the ability to conduct multilevel analyses. Also, as in all cross-sectional research, there is a risk of unmeasured third variables. In particular, we did not include AI trust and perceived AI risk. However, these constructs are conceptually close to performance expectancy, as trust reduces uncertainty about the system's outputs and thereby enhances expected performance gains, whereas perceived risk erodes expected utility. In this sense, they are likely to be partially reflected in the performance expectancy construct already included in our model. That said, and highlighting that our model explains around 80% of the variance, trust and perceived risk could still suppress some of the predictors we have modeled. Thus, future research should include additional constructs outside the UTAUT2 framework [64]. Finally, health status was measured with a single self-rated item. While single-item measures of subjective health may not capture the full complexity of an individual's medical condition, this approach is widely used in demographic and population health research. Prior work has demonstrated that the self-rated health item is a valid and reliable indicator, predicting key outcomes such as mortality, use of health services, and health expenditures in large-scale surveys [65]. Nevertheless, we acknowledge that a more fine-grained measure (eg, including specific chronic conditions or severity indices) could have provided additional insights, and future studies may benefit from applying such extended health measures.

Conclusions

This study applied the UTAUT2 model to investigate the factors that drive the use of generative AI for health information–seeking. Although overall usage remains limited, our findings show that performance expectancy, facilitating conditions, habit, and hedonic motivation are positively associated with behavioral intentions. These patterns, observed across all 4 countries—Austria, Denmark, France, and Serbia—suggest that current users of generative AI are likely to be early adopters: individuals who are tech-savvy, curious, and open to innovation. This aligns with the Rogers Diffusion of Innovations theory, which conceptualizes adoption as a gradual process beginning with a small, innovation-oriented segment of the population.

The lack of significant effects for effort expectancy and social influence across all countries reinforces this interpretation: early adopters tend to base their decisions on personal evaluations rather than external opinions and are rarely deterred by usability concerns. Furthermore, the fact that behavioral intention was unrelated to health status or health literacy underscores that current usage is not driven by acute medical need or advanced health literacy, but rather by interest, convenience, and technological exploration.

The cross-national consistency of these findings is particularly striking. Despite differences in health care systems, digital infrastructures, and culture, the same psychological and contextual factors influenced generative AI use in all countries surveyed. This suggests a shared adoption logic that transcends national boundaries—at least in the early stages of diffusion.

Looking ahead, these insights help illuminate how generative AI might transition from a niche tool to a widely used resource. As the technology becomes more embedded in everyday life, broader segments of the population—the so-called early and late majority—will likely demand stronger assurances of trustworthiness, safety, and added value. To enable responsible and inclusive adoption, it is therefore crucial to reduce digital access barriers, enhance transparency, and implement safeguards against health misinformation, especially for users with limited health literacy.

From a practical perspective, our findings suggest that communication strategies aiming to promote generative AI for health purposes should emphasize concrete benefits, ease of access, and even enjoyment. Rather than exclusively targeting individuals with chronic or urgent health needs, positioning generative AI as an engaging, low-barrier tool may broaden its appeal—reaching users who might otherwise be disengaged from traditional health information sources.

In sum, generative AI holds significant potential as a future health information resource—but its trajectory will depend on how well we understand and support the evolving needs of its users across different adoption phases and contexts.

XSL•FO
RenderX

Acknowledgments

The authors used ChatGPT (OpenAI) to support language editing of prewritten text sections (eg, to improve grammar and phrasing). All suggestions from the artificial intelligence tool were reviewed by the authors and revised or rejected as necessary. No content was generated solely by the artificial intelligence, and the authors remain fully responsible for the final text.

Funding

The work was supported by Circle U [2024-09 – AIHEALTH].

Data Availability

All data and supplementary materials related to this study are openly accessible via the Open Science Framework (OSF) at the following link [66].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Descriptive and reliability analysis of predictor variables.
[DOCX File, 20 KB - jmir_v28i1e75648_app1.docx]

Multimedia Appendix 2
Mean usage frequency of health information sources with 95% CI per country.
[DOCX File, 199 KB - jmir_v28i1e75648_app2.docx]

Checklist 1
CHERRIES checklist.
[DOCX File, 19 KB - jmir_v28i1e75648_app3.docx]

References

1. Braveman P. Health disparities and health equity: concepts and measurement. Annu Rev Public Health 2006;27:167-194. [doi: 10.1146/annurev.publhealth.27.021405.102103] [Medline: 16533114]
2. Link E, Beckmann S. AI at everyone's fingertips? Identifying the predictors of health information seeking intentions using AI. Commun Res Rep 2025 Jan;42(1):1-11. [doi: 10.1080/08824096.2024.2427609]
3. Varghese J, Chapiro J. ChatGPT: The transformative influence of generative AI on science and healthcare. J Hepatol 2024 Jun;80(6):977-980. [doi: 10.1016/j.jhep.2023.07.028] [Medline: 37544516]
4. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119. [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
5. Nastasi AJ, Courtright KR, Halpern SD, Weissman GE. A vignette-based evaluation of ChatGPT's ability to provide appropriate and equitable medical advice across care contexts. Sci Rep 2023 Oct 19;13(1):17885. [doi: 10.1038/s41598-023-45223-y] [Medline: 37857839]
6. Venkatesh V, Thong JYL, Xu X. Consumer Acceptance and Use of Information Technology: extending the Unified Theory of Acceptance and Use of Technology. MIS Q 2012 Mar 1;36(1):157-178. [doi: 10.2307/41410412]
7. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
8. Baumann E, Czerwinski F, Rosset M, Großmann U. Wie informieren sich die deutschen zu gesundheitsthemen? [Report in German]. : Stiftung Gesundheitswissen; 2019 URL: https://www.stiftung-gesundheitswissen.de/sites/default/files/pdf/trendmonitor_Ausgabe%201.pdf [accessed 2026-01-16]
9. Health information seeking among caregivers. : National Cancer Institute; 2019 URL: https://hints.cancer.gov/docs/Briefs/HINTS_Brief_40.pdf [accessed 2026-01-16]
10. Bachl M, Link E, Mangold F, Stier S. Search engine use for health-related purposes: behavioral data on online health information-seeking in Germany. Health Commun 2024 Jul;39(8):1651-1664. [doi: 10.1080/10410236.2024.2309810] [Medline: 38326714]
11. Al Shboul MKI, Alwreikat A, Alotaibi FA. Investigating the use of ChatGPT as a novel method for seeking health information: a qualitative approach. Sci Technol Libr 2023;43(3):225-234. [doi: 10.1080/0194262X.2023.2250835]
12. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]

XSL•FO
RenderX

13. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
14. Chang A. UTAUT and UTAUT 2: a review and agenda for future research. Winners 2012;13(2):10. [doi: 10.21512/tw.v13i2.656]
15. Walle AD, Jemere AT, Tilahun B, et al. Intention to use wearable health devices and its predictors among diabetes mellitus patients in Amhara region referral hospitals, Ethiopia: using modified UTAUT-2 model. Inf Med Unlocked 2023;36:101157. [doi: 10.1016/j.imu.2022.101157]
16. Yu CW, Chao CM, Chang CF, Chen RJ, Chen PC, Liu YX. Exploring behavioral intention to use a mobile health education website: an extension of the UTAUT 2 model. Sage Open 2021;11(4):215824402110557. [doi: 10.1177/21582440211055721]
17. Schomakers EM, Lidynia C, Vervier LS, Calero Valdez A, Ziefle M. Applying an extended UTAUT2 model to explain user acceptance of lifestyle and therapy mobile health apps: survey study. JMIR mHealth uHealth 2022 Jan 18;10(1):e27095. [doi: 10.2196/27095] [Medline: 35040801]
18. Budhathoki T, Zirar A, Njoya ET, Timsina A. ChatGPT adoption and anxiety: a cross-country analysis utilising the unified theory of acceptance and use of technology (UTAUT). Stud High Educ 2024;49(5):831-846. [doi: 10.1080/03075079.2024.2333937]
19. Lai CY, Cheung KY, Chan CS, Law KK. Integrating the adapted UTAUT model with moral obligation, trust and perceived risk to predict ChatGPT adoption for assessment support: a survey with students. Comput Educ Artif Intell 2024 Jun;6:100246. [doi: 10.1016/j.caeai.2024.100246]
20. Foroughi B, Senali MG, Iranmanesh M, et al. Determinants of intention to use ChatGPT for educational purposes: findings from PLS-SEM and fsQCA. Int J Hum Comput Interact 2024;40(17):4501-4520. [doi: 10.1080/10447318.2023.2226495]
21. Terblanche N, Kidd M. Adoption factors and moderating effects of age and gender that influence the intention to use a non-directive reflective coaching chatbot. Sage Open 2022;12(2):21582440221096136. [doi: 10.1177/21582440221096136]
22. Neeragatti S, Dehury RK, Sripathi N. Determinants of digital health information search (DHIS) behaviour: extending UTAUT with healthcare behaviour constructs. 2022 Presented at: 2nd International Healthcare Management Conference 2022: Navigating the New Normal with Focus on Healthcare Accessibility, Innovation and Sustainability; Jan 29, 2022. [doi: 10.24083/apjhm.v18i1.1685]
23. Kim Y, Blazquez V, Oh T. Determinants of generative AI system adoption and usage behavior in Korean companies: applying the UTAUT model. Behav Sci (Basel) 2024 Nov 4;14(11):1035. [doi: 10.3390/bs14111035] [Medline: 39594336]
24. Strzelecki A. To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. Interact Learn Environ 2024;32(9):5142-5155. [doi: 10.1080/10494820.2023.2209881]
25. Chang YT, Chao CM, Yu CW, Lin FC. Extending the utility of UTAUT2 for hospital patients' adoption of medical apps: moderating effects of e-health literacy. Mob Inf Syst 2021 Apr 12;2021:1-10. [doi: 10.1155/2021/8882317]
26. Gansser OA, Reich CS. A new acceptance model for artificial intelligence with extensions to UTAUT2: an empirical study in three segments of application. Technol Soc 2021 May;65:101535. [doi: 10.1016/j.techsoc.2021.101535]
27. Kang S, Choi Y, Kim B. Impact of motivation factors for using generative AI services on continuous use intention: mediating trust and acceptance attitude. Soc Sci 2024 Sep;13(9):475. [doi: 10.3390/socsci13090475]
28. Limayem M, Hirt SG, Cheung CMK. How habit limits the predictive power of intention: the case of information systems continuance. MIS Q 2007 Dec;31(4):705-737. [doi: 10.2307/25148817]
29. Brown SA, Venkatesh V. Model of adoption of technology in households: a baseline model test and extension incorporating household life cycle. MIS Q 2005;29(3):399-426. [doi: 10.2307/25148690]
30. Diao Y, Li Z, Zhou J, Gao W, Gong X. A meta-analysis of college students' intention to use generative artificial intelligence. arXiv. Preprint posted online on Aug 25, 2024. [doi: 10.48550/arXiv.2409.06712]
31. Nutbeam D. Health promotion glossary. Health Promot Int 1998 Jan 1;13(4):349-364. [doi: 10.1093/heapro/13.4.349]
32. Parker R, Ratzan SC. Health literacy: a second decade of distinction for Americans. J Health Commun 2010;15 Suppl 2:20-33. [doi: 10.1080/10810730.2010.501094] [Medline: 20845190]
33. van der Vaart R, Drossaert C. Development of the digital health literacy instrument: measuring a broad spectrum of Health 1.0 and Health 2.0 skills. J Med Internet Res 2017 Jan 24;19(1):e27. [doi: 10.2196/jmir.6709] [Medline: 28119275]
34. Yang K, Hu Y, Qi H. Digital health literacy: bibliometric analysis. J Med Internet Res 2022 Jul 6;24(7):e35816. [doi: 10.2196/35816] [Medline: 35793141]
35. Dunn P, Hazzard E. Technology approaches to digital health literacy. Int J Cardiol 2019 Oct 15;293:294-296. [doi: 10.1016/j.ijcard.2019.06.039] [Medline: 31350037]
36. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med 2011 Jul 19;155(2):97-107. [doi: 10.7326/0003-4819-155-2-201107190-00005] [Medline: 21768583]
37. Diviani N, van den Putte B, Giani S, van Weert JC. Low health literacy and evaluation of online health information: a systematic review of the literature. J Med Internet Res 2015 May 7;17(5):e112. [doi: 10.2196/jmir.4018] [Medline: 25953147]
38. Zhang Z, Genc Y, Xing A, Wang D, Fan X, Citardi D. Lay individuals' perceptions of artificial intelligence (AI)-empowered healthcare systems. Presented at: 83rd Annual Meeting of the Association for Information Science and Technology; Oct 25-29, 2022. [doi: 10.1002/pra2.326]

XSL•FO
RenderX

39. Yuen E, Winter N, Savira F, et al. Digital health literacy and its association with sociodemographic characteristics, health resource use, and health outcomes: rapid review. Interact J Med Res 2024 Jul 26;13:e46888. [doi: 10.2196/46888] [Medline: 39059006]
40. Yi-No Kang E, Chen DR, Chen YY. Associations between literacy and attitudes toward artificial intelligence–assisted medical consultations: the mediating role of perceived distrust and efficiency of artificial intelligence. Comput Human Behav 2023 Feb;139:107529. [doi: 10.1016/j.chb.2022.107529]
41. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012 Jan 27;14(1):e19. [doi: 10.2196/jmir.1619] [Medline: 22357448]
42. Templin T, Perez MW, Sylvia S, Leek J, Sinnott-Armstrong N. Addressing 6 challenges in generative AI for digital health: a scoping review. PLoS Digit Health 2024 May;3(5):e0000503. [doi: 10.1371/journal.pdig.0000503] [Medline: 38781686]
43. Chen X, Hay JL, Waters EA, et al. Health literacy and use and trust in health information. J Health Commun 2018;23(8):724-734. [doi: 10.1080/10810730.2018.1511658] [Medline: 30160641]
44. Li J, Theng YL, Foo S. Predictors of online health information seeking behavior: changes between 2002 and 2012. Health Informatics J 2016 Dec;22(4):804-814. [doi: 10.1177/1460458215595851] [Medline: 26261218]
45. Houston TK, Allison JJ. Users of Internet health information: differences by health status. J Med Internet Res 2002;4(2):E7. [doi: 10.2196/jmir.4.2.e7] [Medline: 12554554]
46. Madrigal L, Escoffery C. Electronic health behaviors among US adults with chronic disease: cross-sectional survey. J Med Internet Res 2019 Mar 5;21(3):e11240. [doi: 10.2196/11240] [Medline: 30835242]
47. Or Z, Gandré C, Seppänen AV, et al. France: Health system review. Health Syst Transit 2023 Jul;25(3):1-276. [Medline: 37489947]
48. Bjegovic-Mikanovic V, Vasic M, Vukovic D, et al. Serbia: Health system review. Health Syst Transit 2019 Oct;21(3):1-211. [Medline: 32851979]
49. Bachner F, Bobek J, Habimana K, et al. Austria health system summary 2022. : World Health Organization; 2022 URL: https://iris.who.int/bitstream/handle/10665/365423/9789289059367-eng.pdf?sequence=1 [accessed 2026-01-16]
50. Birk HO, Vrangbæk K, Rudkjøbing A, et al. Denmark: Health system review. Health Syst Transit 2024 Feb;26(1):1-186. [Medline: 38841877]
51. Individuals—internet activities. Eurostat. 2024. URL: https://ec.europa.eu/eurostat/databrowser/view/isoc_ci_ac_i/default/table?lang=en [accessed 2026-01-16]
52. Olagnier D, Mogensen TH. The Covid-19 pandemic in Denmark: big lessons from a small country. Cytokine Growth Factor Rev 2020 Jun;53:10-12. [doi: 10.1016/j.cytogfr.2020.05.005] [Medline: 32405247]
53. Maljichi D, Limani B, Spier TE, et al. (Dis)trust in doctors and public and private healthcare institutions in the Western Balkans. Health Expect 2022 Aug;25(4):2015-2024. [doi: 10.1111/hex.13562] [Medline: 35781914]
54. Reinhardt A, Matthes J, Bojic L, Maindal HT, Paraschiv C, Ryom K. Help me, Doctor AI? A cross-national experiment on the effects of disease threat and stigma on AI health information-seeking intentions. Comput Human Behav 2025 Nov;172:108718. [doi: 10.1016/j.chb.2025.108718]
55. Weber W, Reinhardt A, Rossmann C. Lifestyle segmentation to explain the online health information-seeking behavior of older adults: representative telephone survey. J Med Internet Res 2020 Jun 12;22(6):e15099. [doi: 10.2196/15099] [Medline: 32530433]
56. Alam MZ, Hu W, Kaium MA, Hoque MR, Alam MMD. Understanding the determinants of mHealth apps adoption in Bangladesh: a SEM-Neural network approach. Technol Soc 2020 May;61:101255. [doi: 10.1016/j.techsoc.2020.101255]
57. Macedo IM. Predicting the acceptance and use of information and communication technology by older adults: an empirical examination of the revised UTAUT2. Comput Human Behav 2017 Oct;75:935-948. [doi: 10.1016/j.chb.2017.06.013]
58. Fischer SM, Dadaczynski K, Sudeck G, et al. Measuring health literacy in childhood and adolescence with the scale Health Literacy in School-Aged Children–German version: the psychometric properties of the German-language version of the WHO health survey scale HLSAC. Diagnostica 2022;68(4):184-196. [doi: 10.1026/0012-1924/a000296]
59. A-priori sample size calculator for structural equation models. Free Statistics Calculators. 2026. URL: https://www.danielsoper.com/statcalc/calculator.aspx?id=89 [accessed 2026-01-16]
60. Rogers EM. Diffusion of Innovations, 5th edition: Free Press; 2003.
61. Habibi A, Muhaimin M, Danibao BK, Wibowo YG, Wahyuni S, Octavia A. ChatGPT in higher education learning: acceptance and use. Comput Educ Artif Intell 2023;5:100190. [doi: 10.1016/j.caeai.2023.100190]
62. Nielsen J. Usability Engineering: Morgan Kaufmann; 1994.
63. Sun Y, Sheng D, Zhou Z, Wu Y. AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence–generated content. Humanit Soc Sci Commun 2024;11(1):1278. [doi: 10.1057/s41599-024-03811-x]
64. Xu K, Shi J. Visioning a two-level human–machine communication framework: initiating conversations between explainable AI and communication. Commun Theory 2024 Nov 1;34(4):216-229. [doi: 10.1093/ct/qtae016]
65. Cullati S, Bochatay N, Rossier C, Guessous I, Burton-Jeangros C, Courvoisier DS. Does the single-item self-rated health measure the same thing across different wordings? Construct validity study. Qual Life Res 2020 Sep;29(9):2593-2604. [doi: 10.1007/s11136-020-02533-2] [Medline: 32436111]

XSL•FO
RenderX

66. Open science framework (OSF). Generative AI & Health Information-Seeking. URL: https://osf.io/7b5v9/overview?view_only=c37895e21f0d4625be6eb13bc051f614 [accessed 2026-01-22]

Abbreviations

AI: artificial intelligence
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
UTAUT2: Unified Theory of Acceptance and Use of Technology 2
WHO: World Health Organization

Edited by A Mavragani; submitted 08.Apr.2025; peer-reviewed by A Rouhi, D Chrimes, S Sivarajkumar; accepted 27.Oct.2025; published 28.Jan.2026.

Please cite as:
Matthes J, Reinhardt A, Hodzic S, Kaňková J, Binder A, Bojic L, Maindal HT, Paraschiv C, Ryom K
Predicting the Intention to Use Generative Artificial Intelligence for Health Information: Comparative Survey Study
J Med Internet Res 2026;28:e75648
URL: https://www.jmir.org/2026/1/e75648
doi:10.2196/75648

© Jörg Matthes, Anne Reinhardt, Selma Hodzic, Jaroslava Kaňková, Alice Binder, Ljubisa Bojic, Helle Terkildsen Maindal, Corina Paraschiv, Knud Ryom. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Key Information Influencing Patient Decision-Making About AI in Health Care: Survey Experiment Study

Xuan Zhu[1], PhD; Austin M Stroud[2], MA; Sarah A Minteer[1,3], PhD; Dong Whi Yoo[4], PhD; Jennifer L Ridgeway[1,5], PhD; Maryam Mooghali[6], MSc, MD; Jennifer E Miller[6], PhD; Barbara A Barry[1,5], PhD

[1]Robert D and Patricia E Kern Center for the Science of Health Care Delivery, Mayo Clinic, 200 First Street SW, Rochetser, MN, USA

[2]Biomedical Ethics Research Program, Mayo Clinic, Rochester, MN, USA

[3]Department of Physical Medicine and Rehabilitation Research, Mayo Clinic, Rochester, MN, USA

[4]Human-Centered Computing Department, Luddy School of Informatics, Computing, and Engineering, Indiana University, Indianapolis, IN, USA

[5]Division of Health Care Delivery Research, Mayo Clinic, Rochester, MN, USA

[6]Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA

Corresponding Author:
Xuan Zhu, PhD
Robert D and Patricia E Kern Center for the Science of Health Care Delivery, Mayo Clinic, 200 First Street SW, Rochetser, MN, USA

Abstract

Background: Artificial intelligence (AI)–enabled devices are increasingly used in health care. However, there has been limited research on patients' informational preferences, including which elements of AI device labeling enhance patient understanding, trust, and acceptance. Clear and effective patient-facing communication is essential to address patient concerns and support informed decision-making regarding AI-enabled care.

Objective: We evaluated 3 aims using simulated AI device labels in a cardiovascular context. First, we identified key information elements that influence patient trust and acceptance of an AI device. Second, we examined how these effects varied based on patient characteristics. Third, we explored how patients evaluated informational content of AI labels and their perceived effectiveness of the AI labels in informing decision-making about the use of AI device, building trust in the device, and shaping their intention to use it in their health care.

Methods: We recruited 340 US patients from ResearchMatch.org to participate in a web-based survey that contained 2 experiments. In the discrete choice experiment, participants indicated preferences in terms of trust and acceptance regarding 16 pairs of simulated AI device labels that varied across 8 types of information needs identified in our previous qualitative work. In the single profile factorial experiment, participants evaluated 4 randomly assigned label prototypes regarding the label's legibility, comprehensibility, information overload, credibility, and perceived effectiveness in informing about the AI device, as well as participants' trust in the AI device and intention to use the device in their health care. Data were analyzed using mixed effects binary or ordinal logistic regression.

Results: The discrete choice experiment showed that information about regulatory approval, high device performance, provider oversight, and AI's value added to usual care significantly increased the likelihood of patient trust by 14.1% - 19.3% and acceptance by 13.3% - 17.9%. Subgroup analyses revealed variations based on patient characteristics such as familiarity with AI, health literacy, and recency of last medical checkup. The single profile factorial experiment showed that patients reported good label comprehension, and that information about provider oversight, regulatory approval, device performance, and AI's added value improved perceived credibility and effectiveness of the AI label (odds ratio [OR] range: 1.35 - 2.05), reduced doubts in the AI device (OR range: 0.61 - 0.77), and increased trust and intention to use the AI device (OR range: 1.47 - 1.73). However, information about data privacy and safety management protocols was less influential.

Conclusions: Patients value information about an AI device's performance, provider oversight, regulatory status, and added value during decision-making. Providing transparent, easily understandable information about these aspects is critical to support patient determinations of trust and acceptance of AI-enabled health care. Information elements' impact on patient trust and acceptance varies by patient characteristics, highlighting the need for a tailored approach to address the concerns of diverse patient groups about AI in health care.

(J Med Internet Res 2026;28:e75615) doi:10.2196/75615

KEYWORDS

artificial intelligence; health communication; health decision-making; patient preference; patient-centered care; AI labeling

XSL•FO
RenderX

Introduction

Artificial intelligence (AI) and machine learning (ML) are increasingly being integrated into health care products and services due to their potential to enhance diagnostic accuracy, improve treatment planning, increase efficiency of health care systems, reduce costs, and ultimately improve patient outcomes [1-3]. Cardiology, in particular, has seen a surge in AI/ML-enabled clinical tools, aiding clinical decision-making across the care spectrum [4-6]. While there are many promising applications of AI/ML in health care, effective approaches to inform patients about these technologies have not been established. This hinders informed patient decision-making and public trust in AI/ML-enabled medical devices.

Recent research on patient perspectives regarding AI in health care indicates that patients are generally receptive to the use of AI in their health care, but that certain conditions must be met [7,8]. Patients want clear and accessible information about how AI is used in their care, its benefits and limitations, how decisions are made, and the roles of AI and health care providers in AI-informed decisions [9-11]. Patients value human oversight over AI and want to ensure that health care decisions are ultimately made by a human [12-14]. Patients also want assurance that their data are protected and used responsibly [15-17]. Finally, they want AI to be unbiased in its training data and outputs and desire equitable access to AI applications to prevent against underrepresentation of minoritized and disadvantaged population groups in AI research [18-20]. These nuances underscore the complexity of information needed to support informed patient use of AI and effective communication between patients, clinicians, and health systems regarding the use of AI in care.

Recent efforts to understand how to communicate transparently about AI in health care have focused on clinical and technical audiences. As a result, there is growing research on best practices for transparent and standardized documentation, reporting, and communication of AI/ML models for clinicians [21-25]. However, research efforts focusing on identifying and prioritizing patients' specific information needs remain limited [26-28]. It is important to note that it is not always feasible or desirable to present all information about an AI-enabled medical device with equal prominence in patient-facing communication. In practice, patient education materials, including labels, decision aids, and informed consents, must balance completeness with cognitive load. Overwhelming patients with too much information can reduce comprehension, informed decision-making, and psychological well-being [29-32]. Similarly, patient-facing labels for AI-enabled medical devices are highly constrained by space and by the cognitive burden on users [33-35]. Thoughtful organization and prioritization of information are key to avoiding information overload and making it easier for patients to grasp the implications of AI in their care [36]. Thus, to ensure patient safety, autonomy, and informed decision-making, it is essential to understand patients' core information needs and present information in a transparent and accessible manner.

This study is part of a broader research project aimed at bridging the knowledge gap in effective communication about AI in health care by identifying the essential components of patient-facing information labels for these technologies. Information labels are structured summaries that provide users with key details about a device or technology, including its purpose, functionality, benefits, limitations, and risks [37]. The overarching goal of this project is to provide empirical evidence to inform strategies and regulatory policies that facilitate patient-centered adoption of AI in health care. For the purposes of our study, an AI device refers to a cardiology device equipped with an ML model designed to assist in decision-making across the continuum of care. Prior to this study, we conducted qualitative research, including a rapid literature review and 3 qualitative studies with patients and clinicians, to identify core information needs that contribute to patient and clinician trust in AI devices in health care, focusing on their use in diagnosing, treating, and monitoring cardiovascular conditions [34,38,39]. This work identified eight elements reported by patients as influential to their trust in the AI device, including (1) data privacy and security, (2) performance, (3) AI's added value compared with usual care, (4) regulatory approval, (5) expert endorsement, (6) generalizability and limitations, (7) device safety, and (8) health care provider (HCP) oversight. However, no evidence-based guidance currently exists on how to prioritize among them in patient-facing communication. Understanding what information patients prioritize allows AI developers, health care systems, and regulators to emphasize the most critical information elements while ensuring transparency, clarity, and comprehension. To address this knowledge gap, this study used 2 survey experiments in which participants evaluated a short, hypothetical scenario ("vignette") featuring AI label prototypes to achieve the following specific aims: (1) determine the relative importance of various information elements on patient trust and acceptance of an AI device in cardiology; (2) examine how the effects of information elements vary due to differences in patient characteristics, including familiarity with AI, medical mistrust, health literacy, and sociodemographic characteristics; and (3) assess patients' evaluation of the informational content of the AI label prototypes, perceived effectiveness of the AI label prototypes in informing decision-making about the use of the AI device, trust in the device, and intention to use the device in their health care.

Methods

Clinical Context for Experimental Vignettes

We selected a hypothetical AI-enabled smartwatch and smartphone app designed to detect potential episodes of atrial fibrillation as the AI device example for our experimental vignettes. This choice was informed by input from our clinical and regulatory collaborators as well as feedback from both clinicians and patients during our formative qualitative research. The clinical problem and AI technology in the vignette were deemed appropriate by clinician collaborators given its direct relevance to patients and accessibility to a broad population. In addition, during formative qualitative research, this AI device prompted rich discussion among patient and clinician participants, underscoring its relevance and making it a strong

XSL•FO
RenderX

candidate for use as the clinical context for the experimental vignettes.

Experimental Design

Overview

To robustly examine how AI label informational content influences patient perceptions and decision-making, we conducted 2 vignette experiments through a web-based survey. The use of vignette experiments is a widely accepted approach in health communication and medical decision science research [40-43]. This method is especially useful when studying emerging technologies for which standardized communication practices are not yet established. The vignette experiment approach offers important advantages over simply asking participants to rank or rate the importance of various information elements [41]. By presenting information elements in concrete vignettes, we create a context that more closely resembles how decisions are made in real-world health care settings. It also allowed us to control for confounding factors and precisely isolate the causal effect of each information element on psychological and behavioral outcomes such as trust and acceptance. Moreover, simply asking participants what is important to them is highly susceptible to social desirability bias; by examining how participants respond to concrete situations, the vignette experiment approach minimizes social desirability bias and offers a clearer picture of the factors that shape patient preferences and decisions. To capture both patient priorities under conditions requiring trade-offs and their evaluations of AI label informational content in a more reflective context, we implemented 2 complementary vignette experiments: a discrete choice experiment (DCE) and a single profile factorial experiment (SPFE).

DCEs are being increasingly used in medical and health services research to systematically examine people's preferences regarding health care services by assessing how much they value the specific attributes of the service [44-47]. In a typical DCE, participants are presented with 2 or more discrete hypothetical alternatives (eg, treatment A or treatment B), each consisting of multiple attributes with varying values. By analyzing the choices participants make among these alternatives, researchers can estimate the contribution of each attribute to decision-making. In this study, by observing how participants make trade-offs when choosing between competing label configurations, the DCE allowed us to (1) determine the relative importance of different information elements influencing patient trust and acceptance of an AI device and (2) examine how the effects of information elements vary due to differences in patient characteristics. This method uncovers underlying priorities and preferences that might not be evident when participants simply rate or rank information elements individually.

In contrast, the SPFE asked participants to review AI label prototypes one at a time and respond to a series of rating measures on perceived comprehension, cognitive effort, effectiveness in supporting decision-making about the AI device, trust in the device, and intention to use it in their health care. This method was chosen for 2 key reasons. First, the single profile design mirrors real-world patient decision-making regarding AI device use in health care more closely as in real life, patients typically consider a single device rather than make direct comparisons among alternatives. Second, the SPFE allows rating-based outcome measures (eg, 1=very unlikely to accept to 5=very likely to accept) that provide more direct and fine-grained insight into participants' evaluation of the label prototypes as well as the psychological and behavioral impacts of the information elements. This dual-experiment approach allows us to gain a richer and more nuanced understanding of patient preferences regarding AI device labeling and the factors that are critical for effective communication and adoption of AI technologies in health care.

For these experiments, we created AI label prototypes based on the hypothetical AI device that varied on the following eight informational elements: (1) data privacy and security, (2) performance, (3) AI's added value, (4) regulatory approval, (5) expert endorsement, (6) generalizability and limitations, (7) device safety, and (8) HCP oversight. We varied the informational content along these 8 elements based on prior qualitative research conducted with clinicians and patients, which identified these as important information needs influencing patient trust in AI [38,39]. For experimental design efficiency, each of these elements has 2 levels (ie, 2 variations): either in 2 different versions or as either present or absent. See Table 1 for additional details about these 8 elements.

XSL•FO
RenderX

Table . Label information elements and levels.

Elements	Definitions	Levels and examples
Elements with varying levels		
X1. Data privacy and security	What patient data are collected by AI[a]; how the patient data are being collected, stored, and shared.	Level 1: opt-in data sharing; level 2: opt-out data sharing
X2. Performance	True-positive: device accuracy when the patient DOES have A-fib[b]; true-negative: device accuracy when the patient DOES NOT have A-fib	Level 1: high performance; level 2: low performance
X3. AI's added value	Improvement in patient care due to the AI-enabled device; effectiveness of the AI-enabled device in comparison with non–AI-enabled tests or conventional health care.	Level 1: information absent; level 2: information present
X4. Regulatory approval	Whether the AI device received clearance, approval, or certification from a regulatory body regarding its safety and/or effectiveness.	Level 1: information absent; level 2: information present
X5. Expert endorsement	Endorsement of the device for safety and effectiveness issued by medical experts such as health care providers.	Level 1: information absent; level 2: information present
X6. Generalizability and limitations	How generalizable is the device to patients with varying demographics and characteristics; what are the conditions or contexts where the device should not be used?	Level 1: internal validation; level 2: external validation
X7. Device safety	What are the risks of malfunction, bugs, and errors from the AI device (both the AI algorithm and the supporting software)? How will these risks be managed?	Level 1: proactive auditing; level 2: reactive auditing
X8. HCP[c] oversight	Whether the results or decisions from the AI-enabled device have been verified by your health care provider.	Level 1: information absent; level 2: information present
Elements to be displayed in the same way across all labels		
1. Purpose	What is the purpose of the AI device, for what types of patients and conditions, when and in what context should the device be used?	"This AI-enabled smartwatch and smartphone app is designed to identify a potential episode of atrial fibrillation (A-fib). A-fib is an irregular and often very rapid heart rhythm that can lead to blood clots in the heart."
2. Directions	How to use the AI device, recommended actions for patients, how to interpret results (eg, what is "normal" for pt. like me); what warrants discussion with provider; next steps by AI if any (eg, AI will automatically alert HCP)	"- Wear your smartwatch - It will provide an alert when it detects a potential episode of A-fib - Talk to your doctor if you received an alert from the app"

[a]AI: artificial intelligence.

[b]A-fib: atrial fibrillation.

[c]HCP: health care provider.

Discrete Choice Experiment

We used a fractional factorial design to rate a selection of possible alternatives because a full factorial design (2^8 alternatives) would not be feasible to implement. To minimize the cognitive burden for participants and maximize statistical efficiency, each choice task consisted of 2 alternatives, and each participant was asked to consider 16 choice sets, comprising 32 unique label prototypes in total [48]. We used the R package idefix [49] to select an optimal design based on D-efficiency criterion, a commonly used metric for efficient experimental design construction to maximize statistical efficiency and precision [50-52]. Multimedia Appendix 1 summarizes the choice sets. Figure 1 shows an example choice set.

XSL•FO
RenderX

Figure 1. Example choice set for the discrete choice experiment. A-fib: atrial fibrillation; AI: artificial intelligence.

Name	A-Fib Watch
Purpose	This AI-enabled smartwatch and smartphone app is designed to identify a potential episode of atrial fibrillation (A-fib). A-fib is an irregular and often very rapid heart rhythm that can lead to blood clots in the heart.
Directions	• Wear your smartwatch • It will provide an alert when it detects a potential episode of A-fib • Talk to your doctor if you received an alert from the app

	Device A	Device B
Data Privacy and Security	The information you provide in the app will be used to improve the tool. Your deidentified data will be shared unless you visit the settings menu to turn off the sharing option.	The information you provide in the app will be used to improve the tool. You will be asked to opt into sharing your deidentified data when you first open the app.
Performance True-positive: device accuracy when the patient DOES have A-fib True-negative: device accuracy when the patient DOES NOT have A-fib	 90% (9 out of 10) people with A-fib will be correctly identified as having A-fib.  80% (8 out of 10) people without A-fib will be correctly identified as NOT having A-fib.	 80% (8 out of 10) people with A-fib will be correctly identified as having A-fib.  70% (7 out of 10) people without A-fib will be correctly identified as NOT having A-fib.
Added value		Appropriate use of the tool, response to alerts, and treatment may reduce risk of heart failure and stroke.
Regulatory approval	This device acquired regulatory authorization, which means that this device is as safe and effective as other regulated device(s).	
Expert endorsement		A panel of doctors has reviewed and endorsed this device for use.
Validation	The clinic where you receive care has tested and validated this device.	An independent research lab unaffiliated with the clinic where you receive care has tested and evaluated this device.
Tool safety	This device will be audited by developer team every 4 weeks.	If doctors or patients report any issues with the device, the developer team will assess and fix them promptly.
Oversight from your doctor	Your doctor will review the AI results along with your medical history.	

Single Profile Factorial Experiment

The SPFE used a 2^{8-3} resolution IV fractional factorial design to evaluate the label prototypes. We used the R package FrF2 [53] to generate the 32 experimental conditions (Multimedia Appendix 2). Each participant was randomly assigned to evaluate 4 label prototypes displayed in a randomized order. After viewing each label prototype, participants answered a series of questions on label evaluation, AI trust, and behavioral intention using 5-point Likert-style scales. Figure 2 shows an example label prototype.

XSL•FO
RenderX

Figure 2. Example label prototype for the single profile factorial experiment. AI: artificial intelligence.

Name	A-fib Watch
Purpose	This AI-enabled smartwatch and smartphone app is designed to identify a potential episode of atrial fibrillation (A-fib). A-fib is an irregular and often very rapid heart rhythm that can lead to blood clots in the heart.
Directions	• Wear your smartwatch • It will provide an alert when it detects a potential episode of A-fib • Talk to your doctor if you received an alert from the app
Data Privacy and Security	The information you provide in the app will be used to improve the tool. You will be asked to opt into sharing your deidentified data when you first open the app.
Performance True-positive: device accuracy when the patient DOES have A-fib	 80% (8 out of 10) people with A-fib will be correctly identified as having A-fib.
True-negative: device accuracy when the patient DOES NOT have A-fib	 70% (7 out of 10) people without A-fib will be correctly identified as NOT having A-fib.
Added value	Appropriate use of the tool, response to alerts, and treatment may reduce risk of heart failure and stroke.
Regulatory approval	This device acquired regulatory authorization, which means that this device is as safe and effective as other regulated device(s).
Expert endorsement	A panel of doctors has reviewed and endorsed this device for use.
Validation	The clinic where you receive care has tested and validated this device.
Tool safety	This device will be audited by the developer team every 4 weeks.
Oversight from your doctor	

Pilot Testing

We conducted cognitive interviews (via Zoom [Zoom Communications, Inc]; 30‐45 minutes in length) with 4 participants to assess whether the information elements, variations, presentation, and survey questions were easy to understand and meaningful to participants. The interviewer shared the web-based survey with participants via shared screen and asked participants to verbally answer the survey questions and share their thought processes. The interviewer also asked participants to explain the information elements in their own words. The interviews were audio-recorded, and the interviewer took notes on reflections and observations. We revised the survey questionnaire based on interview notes.

Following the cognitive interviews, we pilot tested the revised Qualtrics web-based survey with 30 participants to examine the feasibility and data quality of the 2 vignette experiments. The vignette experiments consisted of 3 sections: DCE, SPFE, and participant characteristics. The presentation order of the 2 experiments was randomized to mitigate order effect. For the DCE, participants first completed a warm-up task to familiarize them with the DCE procedure. They were then given the following instructions for the main DCE on AI information labeling:

> *Next, you will be asked to choose between 2 AI-enabled device options with similar or different characteristics.*
>
> *AI-enabled devices in health care are medical devices that use AI and specifically the subset of AI known as machine learning. Some examples of AI-enabled devices in health care include smartwatches that can monitor your heart rate for problems, smart robots that guide surgeries, or AI programs that can provide information to a physician to help with diagnosis, among many other types of technologies.*
>
> *Imagine that your health care provider recommended you using an AI-enabled smartwatch or smartphone app to help track your cardiac functions and identify potential episodes of atrial fibrillation. You were given 2 options. The 2 options were developed by the same company and provide the same functions. Please consider the 2 options below and choose the device you prefer the most.*

Participants then proceeded to complete 18 choice tasks, including 16 experimental choice tasks and 2 validity check tasks. The order of the 16 experimental choice tasks was randomized.

In the SPFE, participants were randomly assigned to evaluate 4 of 32 label prototypes, presented in a random order. They were given the following instructions at the beginning of the

XSL•FO
RenderX

section, "In this section, you will see 4 information labels about AI-enabled devices one at a time. After reading each information label, you will be asked a series of questions about your opinions on the label." For each label prototype, participants answered questions on label evaluation, trust in the AI device, and intention to use the device in health care. After completing the experiments, participants provided information on their sociodemographic and health care characteristics. Figure 3 summarizes study procedure.

Figure 3. Study procedure. DCE: discrete choice experiment: SPFE: single profile factorial experiment.

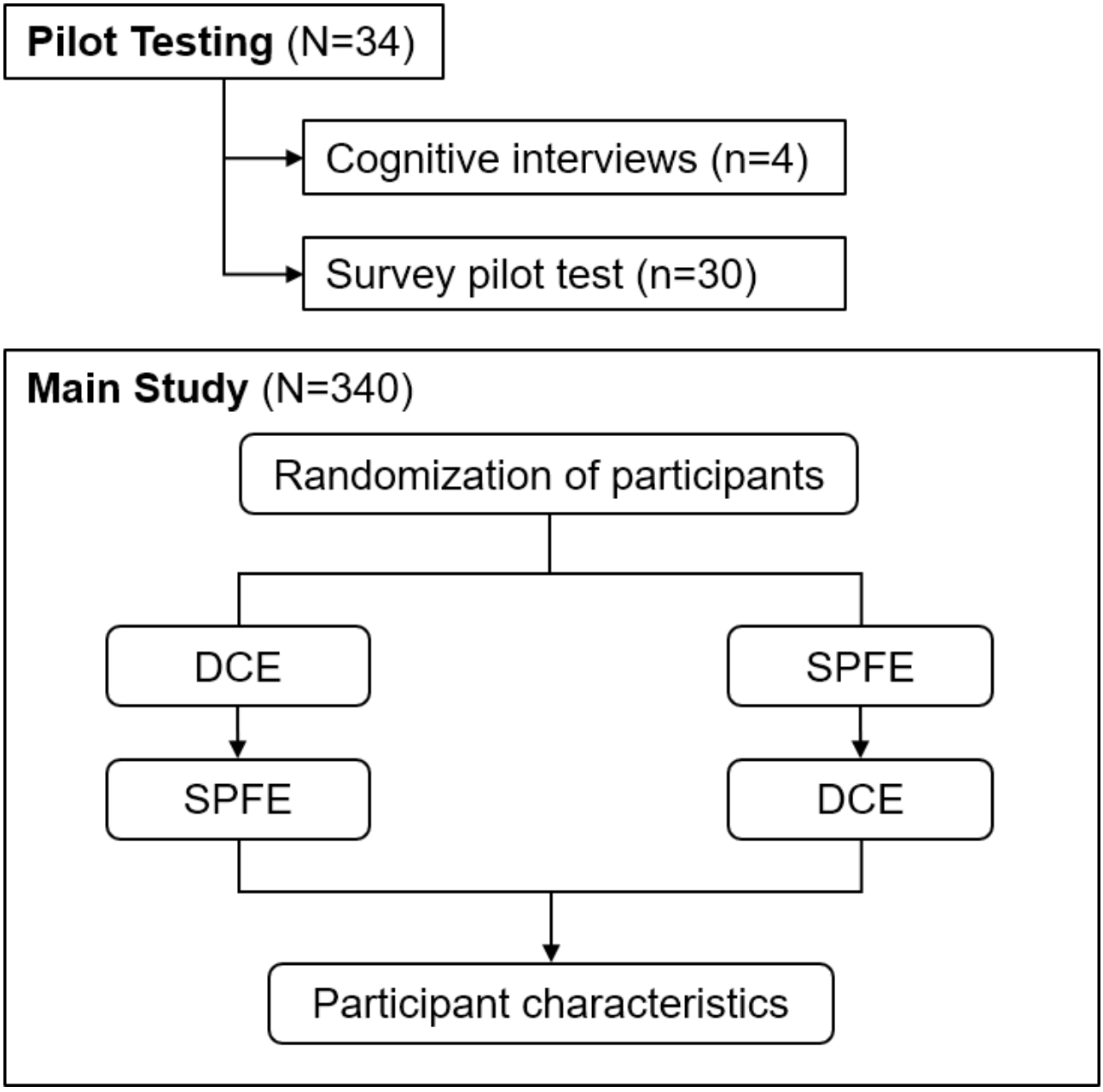

Main Study

For the main study, participants completed the finalized web-based Qualtrics survey, which incorporated refinements from the pilot phase. The survey experiments followed the same procedures described in the "Pilot Testing" section, and the randomization procedure and data collection methods and materials remained consistent with the piloted version.

Measures

Label preference was measured through participants selecting their preferred label prototype in a choice task based on trust ("Which device would you trust more?") and acceptance ("If you were given the option, which device would you be more likely to use in your health care?").

Label evaluation included six items on a 5-point Likert scale (1=strongly disagree to 5=strongly agree) that assessed (1) legibility ("This label is easy to read"), (2) comprehensibility ("This label is easy to understand"), (3) information overload ("Reading this label is too mentally demanding for me"), (4) credibility ("I trust the information on this label"), and (5) effectiveness of the label in informing patients about the AI device ("This label gives me all the information I need about

XSL•FO
RenderX

the AI device" and "This label helps me decide whether the AI device should be used in my care").

Participants' trust in the AI device was measured with three items on a 5-point Likert scale (1=strongly disagree to 5=strongly agree), including (1) "I would trust the results from this AI device," (2) "I would have doubts about this AI device," and (3) "I would seek a second opinion." Participants' likelihood of using the device was assessed on a 5-point Likert-type scale (1=very unlikely to 5=very likely). In addition, participants provided feedback on each label prototype through an optional open-ended question.

Participant characteristics were measured in terms of the need for cognition, which reflects motivation to process complex information and engage in effortful thinking [54,55], familiarity with AI, medical mistrust [56], health literacy [57-59], last routine medical checkup, health insurance coverage, and demographics (eg, age, gender, race or ethnicity, education level, and household financial status) [60].

Sample Size

We conducted an a priori power analysis to estimate the smallest sample size required for the main study. Because the DCE is more statistically efficient than the SPFE, we based our estimation on the SPFE design. The analysis showed that a sample of 288 participants would allow us to detect a main effect with a standardized regression coefficient of 0.167 or higher with 80% power at an α level of .05, which is sufficient for the DCE [61]. The total sample size needed for the pilot and main studies is 323. We planned to recruit 350 participants in total to account for attrition and incomplete responses.

Recruitment

We recruited participants from ResearchMatch [62], a national health volunteer registry that was created by several academic institutions and supported by the US National Institutes of Health as part of the Clinical Translational Science Award program. ResearchMatch has a large population of volunteers who have consented to be contacted by researchers about health studies for which they may be eligible. Inclusion criteria for this study include being an adult aged 18 years or older, being proficient in reading and writing in English, and having had a primary care or cardiology care visit within the past 3 years. We oversampled racial and ethnic minority patients to ensure that their representation was comparable with that of non-Hispanic White patients. This supported our ability to draw meaningful conclusions about diverse patient populations while accounting for historical underrepresentation in health research. Participants received a message describing our study via the ResearchMatch platform, where they could then decide to opt in to be contacted and receive a link to our survey. ResearchMatch provided contact information (ie, name and email address) for participants who opted in. Participants who opted in received an email invitation with a link to complete the online survey.

Ethical Considerations

Review and approval for this study and all procedures were obtained from the Mayo Clinic Institutional Review Board (IRB; 21-012302). The IRB granted a waiver of written documentation of informed consent. The first page of the survey displayed an IRB-approved informed consent cover letter which provided information about the study, including its purpose, the investigator, the estimated length of the survey, data storage procedures, and contact information for the study team and the Mayo Clinic IRB. Participants were instructed to review this information and were informed that by proceeding to the survey, they were providing informed consent. No personal information was collected or stored with the survey responses. Study data were secured on institutionally approved and controlled access electronic storage. Participants received a US $10 gift card as remuneration for completing the survey.

Statistical Analysis

We used mixed effects binary logistic regression to analyze the data from the DCE. Our analysis focused on how different information factors influenced the probability of individuals trusting and accepting the AI device in their health care. We included individual-specific intercepts to account for heterogeneity in individual preferences. We reported average marginal effect (AME), which shows the average change in predicted probabilities (percentage point increase or decrease) of the outcome variable across all participants when moving from one level of the information factor to another, while holding all other variables constant. In addition, we conducted subgroup analyses to explore how the effect of top preferred information factors on patient trust and acceptance of the AI device varied based on participants' AI familiarity, openness toward medical AI, health literacy, medical mistrust, and sociodemographic characteristics.

We used mixed effects ordinal logistic regression (cumulative link mixed model) to analyze the data from the SPFE. Our focus was on how different information factors influenced participants' evaluation of the AI label's legibility, comprehensibility, information overload, information credibility, perceived effectiveness in informing about the AI device, trust in the AI device, and intention to use it in health care. We included individual-specific intercepts to account for repeated measurements within individuals. For models on AI device trust and intention to use, we adjusted for participants' evaluation of the AI label's legibility, comprehensibility, information overload, information credibility, and perceived effectiveness in informing about the AI device. We used "flexible" thresholds in our models to allow the distance between the 4 cut points of the 5-point scales to vary freely. We reported odds ratios (ORs) for the effect of information factors.

All analyses were performed in R (version 4.3.1; R Core Team) [63]. "tidyverse" [64] was used for data wrangling and visualization, "lme4" [65] was used for fitting mixed effects binary logistic models, "ordinal" [66] was used for fitting mixed effects ordinal logistic models, and "marginaleffects" [67] was used for calculating AME.

Reporting Guidelines

This study is reported in accordance with the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) [68].

XSL•FO
RenderX

Results

Participant Characteristics

A total of 340 participants who completed at least 75% of the survey questions were included in the analysis. Table 2 summarizes participant characteristics. Participants were aged 18 - 34 (122/328, 37.19%), 35 - 54 (114/328, 34.75%), and 55 years or older (92/328, 28.05%). Nearly half were women (158/326, 48.47%), 48.77% (159/326) were men, and 2.76% (9/326) identifying as nonbinary or another gender. The majority were non-Hispanic White (206/327, 63.00%), followed by Black or African American (76/327, 23.24%), Hispanic or Latinx (44/327, 13.46%), Asian or Asian American (30/327, 9.17%), and Native American or Indigenous (13/327, 3.98%). Regarding education, 49.09% (161/328) had a bachelor's degree, 28.96% (95/328) held a master's degree or higher, and 20.73% (68/328) had some college or an associate degree. Financially, 53.54% (174/325) reported having disposable income after paying bills, 29.85% (97/325) had little spare money, and 16.62% (54/325) struggled to pay bills.

XSL•FO
RenderX

Table . Participant characteristics (N=340).

Characteristics	Values, n (%)
Age (years)	
18 - 24	18 (5.49)
25 - 34	104 (31.71)
35 - 44	72 (21.95)
45 - 54	42 (12.80)
55 - 64	52 (15.85)
≥65	40 (12.20)
Missing	12 (3.52)
Gender	
Man	159 (48.77)
Woman	158 (48.47)
Nonbinary/other	9 (2.76)
Missing	14 (4.12)
Race and ethnicity[a]	
White/Caucasian	206 (63.00)
Black/African American	76 (23.24)
Hispanic/Latina/Latino	44 (13.46)
Asian/Asian American	30 (9.17)
Native American/American Indian/Alaska Native/Indigenous	13 (3.98)
Middle Eastern/North African/Arab American	8 (2.45)
Native Hawaiian/Pacific Islander	2 (0.61)
Other race	6 (1.83)
Missing	13 (3.82)
Education level	
High school or less	4 (1.22)
Some college or associate's degree	68 (20.73)
Bachelor's degree	161 (49.09)
Master's degree or higher	95 (28.96)
Missing	12 (3.52)
Perceived household financial status	
After paying the bills, I still have enough money for special things that I want	174 (53.54)
Have enough money to pay the bills but little spare money to buy extra or special things	97 (29.85)
Have enough money to pay the bills but only because I have cut back on things	28 (8.62)
Having difficulty paying the bills, no matter what I do	26 (8.00)
Missing	15 (4.41)
Health insurance coverage	
Private	207 (65.51)
Public	103 (32.59)
No insurance	6 (1.90)
Missing	24 (7.06)

XSL•FO
RenderX

Characteristics	Values, n (%)
Last routine checkup	
Less than a year ago	212 (66.88)
>1 year, <2 years	53 (16.72)
>2 years, <5 years	41 (12.93)
≥5	6 (1.89)
Never had routine checkup	5 (1.58)
Missing	23 (6.76)

[a]The percentages may add over 100% because participants can select multiple races.

Most participants had private health insurance (207/316, 65.51%), while 32.59% (103/316) had public coverage. In addition, 66.88% (212/317) had a routine medical checkup within the past year.

DCE Results

Importance of the Information Factors for Patients' Trust in the AI Device

Figure 4 visualizes the relative importance of the information factors for participants' trust in the AI device. The 4 information factors that produced the greatest increase in participants' trust in AI devices were inclusion of information about regulatory approval (AME=19.34%, 95% CI 15.97%-22.72%), information about high versus low performance (AME=16.62%, 95% CI 14.40%-18.85%), inclusion of information about HCP oversight (AME=15.50%, 95% CI 12.56%-18.44%), and inclusion of information about the AI's added value compared with usual care (14.06% increase, 95% CI 12.16%-15.96%). Including information about opt-in (vs opt-out) data privacy protocol, expert endorsement (vs information absent), and information about external (vs internal) validation less strongly increased the probability of trusting the AI device (9%, 7.32%, and 2.76%, respectively). Effect of information about proactive versus reactive device safety management protocol on participants' trust was not statistically significant (Table 3).

Figure 4. Difference in probability of AI device being trusted by attribute level. AI: artificial intelligence; HCP: health care provider; Info: information; TPR: true-positive rate.

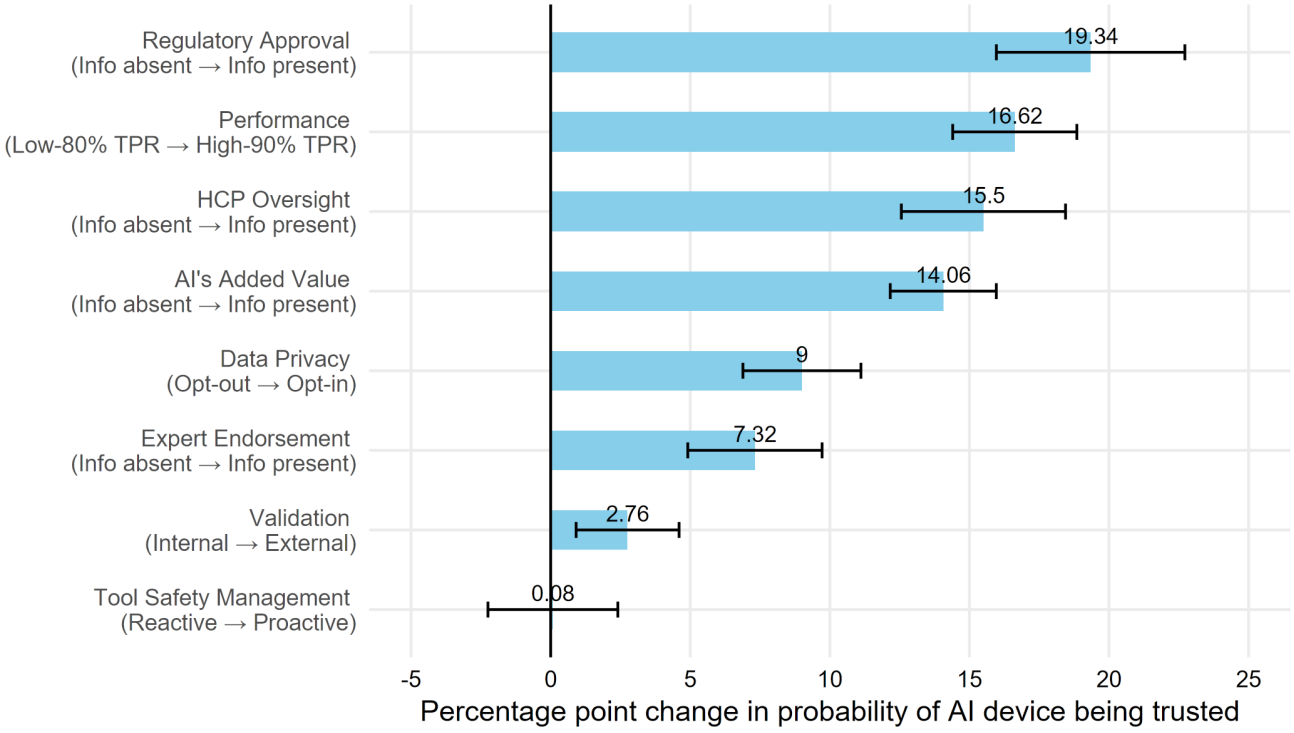

XSL•FO
RenderX

Table . Effects of information elements on the probability of the artificial intelligence device being trusted and accepted.

	Trust	Acceptance
	AME[a] (95% CI), percentage points	AME[a] (95% CI), percentage points
Added value (information absent→information present)	14.06 (12.16 to 15.96)	15.58 (13.28 to 17.88)
Data privacy (opt out→opt in)	9.00 (6.89 to 11.11)	10.92 (8.30 to 13.53)
Expert endorsement (information absent→information present)	7.32 (4.91 to 9.72)	9.03 (6.73 to 11.32)
HCP[b] oversight (information absent→information present)	15.50 (12.56 to 18.44)	17.86 (14.98 to 20.75)
Performance (low→high)	16.62 (14.40 to 18.85)	14.85 (12.65 to 17.06)
Regulatory approval (information absent→information present)	19.34 (15.97 to 22.72)	13.29 (9.96 to 16.63)
Device safety (reactive→proactive)	0.08 (–2.25 to 2.41)	0.74 (–1.82 to 3.29)
Validation (internal→external)	2.76 (0.91 to 4.60)	3.64 (1.56 to 5.72)

[a]AME: average marginal effect. It is the average change in the predicted probabilities (percentage point increase or decrease) of the artificial intelligence device being trusted or accepted across all participants when moving from one level of the information factor to the other, keeping all other variables in the model constant.

[b]HCP: health care provider.

Importance of the Information Factors for Patients' Acceptance of the AI Device

Figure 5 visualizes the relative importance of the information factors for participants' willingness to use the AI device in their health care (ie, acceptance). The 4 information factors that produced the greatest increase in participants' acceptance in using AI devices were inclusion of information about HCP oversight (17.86% increase, 95% CI 4.98%-20.75%), inclusion of information about the device's added value (15.58% increase, 95% CI 13.28%-17.88%), information about high versus low device performance (14.85% increase, 95% CI 12.65%-17.06%), and inclusion of information about regulatory approval (13.29% increase, 95% CI 9.96%-16.63%). Including information about opt-in (vs opt-out) data privacy protocol, expert endorsement (vs information absent), and information about external (vs internal) validation less strongly increased the probability of the AI device being accepted (10.92%, 9.03%, and 3.64%, respectively). The effect of information about proactive versus reactive device safety management protocol on device acceptance was not statistically significant (Table 3).

Figure 5. Difference in probability of artificial intelligence device being accepted by attribute level. AI: artificial intelligence; HCP: health care provider; Info: information; TPR: true-positive rate.

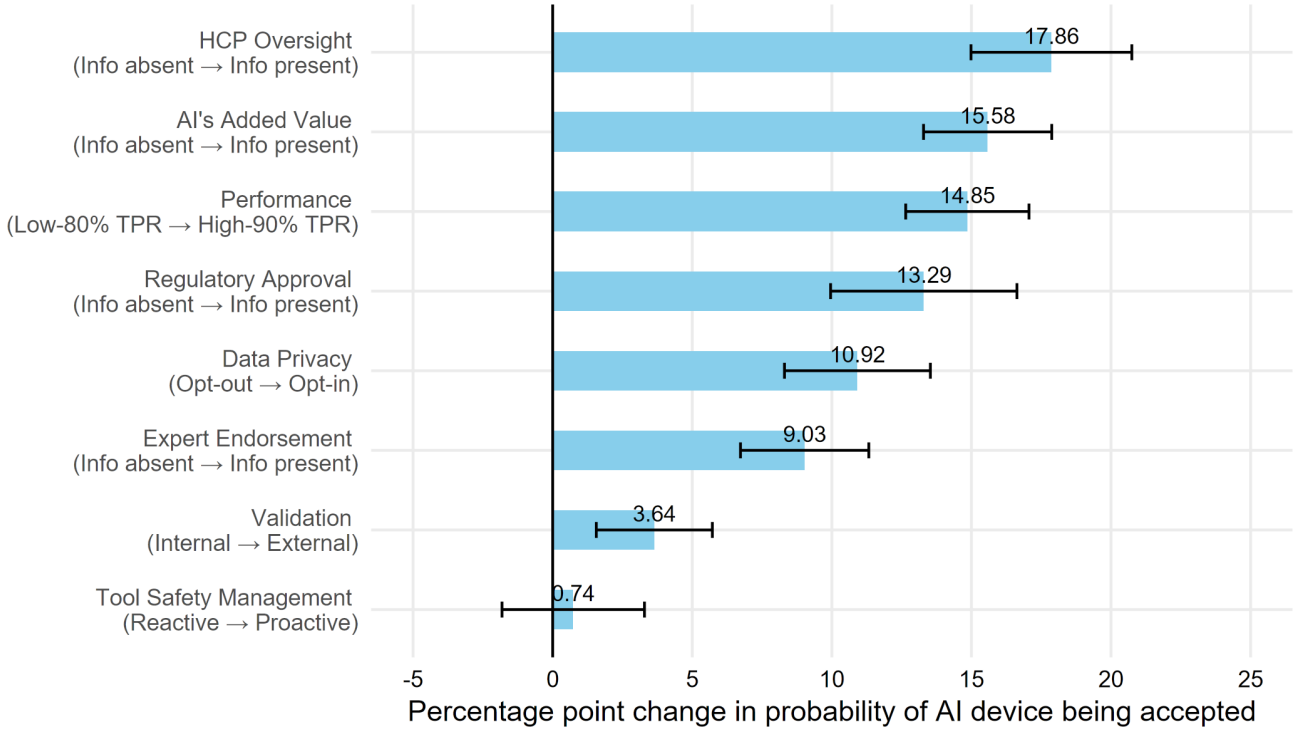

XSL•FO
RenderX

Subgroup Differences in the Effect of Top Preferred Information Factors on Patient Trust and Acceptance

Subgroup Differences in Patient Trust

Our analyses revealed subgroup differences in how the top 4 most important information factors affected trust in AI devices, based on participants' level of familiarity with AI, level of reading health literacy, level of need for cognition, recency of last routine medical checkup, age group, and gender (Multimedia Appendix 3).

AI's Added Value

Specifically, having information about AI's added value (vs information absent) had a stronger positive effect on participants' trust in AI devices with a high (vs low) level of need for cognition (AME=16.21% vs 12.23%; *P*=.04).

HCP Oversight

Having information about HCP oversight (vs information absent) had a stronger positive effect on the probability of AI device being trusted for participants who reported not at all to somewhat (vs very or extremely) familiar with AI (AME=19.27% vs 11.30%; *P*=.01), for participants whose last routine checkup was less than a year ago (vs 1 or more years ago) (AME=18.27% vs 11.63%; *P*=.007), and for participants aged 55 years or older (vs aged 35 - 54 or 18 - 34 years) (AME=23.44% vs 10.34% or 15.25%; *P*<.001; *P*=.03).

High Device Performance

Information about high (vs low) device performance had a stronger positive effect on the probability of AI device being trusted for participants with a high (vs low) level of reading health literacy (AME=22.02% vs 11.06%; *P*<.001), for participants whose last routine checkup was less than a year ago (vs 1 or more years ago) (AME=20.09% vs 10.38%; *P*=.048), and for participants who identified as women (vs men) (AME=19.38% vs 12.85%; *P*=.005).

Regulatory Approval

Having information about regulatory approval (vs information absent) had a stronger positive effect on the probability of AI device being trusted for participants with a high (vs low) level of reading health literacy (AME=25.07% vs 13.63%; *P*=.001) and for participants whose last routine checkup was less than a year ago (vs 1 or more years ago) (AME=23.29% vs 12.92%; *P*<.001).

Subgroup Differences in Patient Acceptance

There were subgroup differences in how the top 4 most important information factors affected acceptance of AI device, based on participants' reported level of familiarity with AI, level of reading health literacy, level of numeracy, recency of last routine medical checkup, gender, and race or ethnicity (Multimedia Appendix 4).

AI's Added Value

Specifically, having information about the device's added value (vs information absent) had a stronger positive effect on the probability of the AI being accepted for participants who reported being not at all to somewhat familiar (vs very or extremely) with AI (AME=18.06% vs 11.49%; *P*=.005), for participants with a high (vs low) level of reading health literacy (AME=19.34% vs 11.96%; *P*=.002), for participants whose last routine checkup was within the last year (vs 1 or more years ago) (AME=19.19% vs 8.81%; *P*<.001), for participants who identified as women (vs men) (AME=18.94% vs 11.38%; *P*=.002), and for participants who identified as a person of color (vs non-Hispanic White) (AME=19.64% vs 12.01%; *P*=.001).

HCP Oversight

Providing information about HCP oversight (vs information absent) had a stronger positive effect on the probability of AI device being accepted for participants who reported not at all to somewhat familiar (vs very or extremely) with AI (AME=25.34% vs 9.59%; *P*<.001), for participants whose last routine checkup was within the last year (vs 1 or more years ago) (AME=21.92% vs 13.31%; *P*<.001), and for participants aged 55 years or older (vs aged 35 - 54 or 18 - 34 years) (AME=30.13% vs 13.05% or 13.77%; *P*<.001; *P*<.001).

High Device Performance

Information about high (vs low) device performance had a stronger positive effect on the probability of AI device being accepted for participants who reported not at all to somewhat familiar (vs very or extremely) with AI (AME=18.18% vs 10.58%; *P*<.001), for participants with a high (vs low) level of reading health literacy (AME=21.41% vs 7.66%; *P*<.001), for participants with a high (vs low) level of numeracy (AME=19.38% vs 9.77%; *P*<.001), for participants whose last routine checkup was within the last year (vs 1 or more years ago) (AME=18.91% vs 7.03%; *P*=.004), for participants who identified as women (vs men) (AME=18.71% vs 9.57%; *P*<.001), and for participants aged 55 years or older (vs aged 35 - 54 or 18 - 34 years) (AME=20.88% vs 12.80% or 11.82%; *P*=.006; *P*=.002).

Regulatory Approval

Providing information about regulatory approval (vs information absent) had a stronger positive effect on the probability of AI device being accepted for participants with a high (vs low) level of reading health literacy (AME=19.93% vs 5.82%; *P*<.001), for participants with a high (vs low) level of numeracy (AME=15.79% vs 9.16%; *P*=.004), for participants whose last routine checkup was within the last year (vs 1 or more years ago) (AME=18.65% vs 3.96%; *P*<.001), and for participants who identified as women (vs men) (AME=16.58% vs 9.43%; *P*=.04).

SPFE Results

Descriptive Statistics

Table 4 summarizes the descriptive statistics of the AI label evaluation measures. Most participants reported that the AI labels were easy to read (perceived legibility; 86.7%) and understand (comprehensibility; 84.2%). They also reported that reading the labels was not too mentally demanding (no information overload; 63.9%). Effects of the information factors on perceived legibility, comprehensibility, and information overload were not statistically significant (Table 5).

XSL•FO
RenderX

Table . Descriptive statistics of the artificial intelligence label evaluation measures.

	5-point Likert scale, n (%)				
Measure	1	2	3	4	5
This label is easy to read (perceived legibility)[a]	15 (1.12)	67 (5.00)	96 (7.16)	474 (35.37)	688 (51.34)
This label is easy to understand (comprehensibility)[a]	18 (1.34)	73 (5.44)	121 (9.02)	498 (37.14)	631 (47.05)
Reading this label is too mentally demanding for me (information overload)[a]	524 (39.13)	332 (24.79)	125 (9.34)	202 (15.09)	156 (11.65)
I trust the information on this label (information credibility)[a]	30 (2.24)	94 (7.01)	220 (16.41)	597 (44.52)	400 (29.83)
Label gives all information needed about the AI[b] device (perceived label effectiveness)[a]	44 (3.28)	243 (18.13)	177 (13.21)	502 (37.46)	374 (27.91)
Label helps me decide whether the device should be used in my care (perceived label effectiveness)[a]	15 (1.12)	83 (6.19)	162 (12.09)	600 (44.78)	480 (35.82)
I would trust the results from this AI device[a]	38 (2.83)	138 (10.29)	218 (16.26)	642 (47.87)	305 (22.74)
I would have doubts about this AI device[a]	128 (9.55)	351 (26.17)	274 (20.43)	405 (30.20)	183 (13.65)
I would seek a second opinion[a]	84 (6.26)	162 (12.08)	256 (19.09)	489 (36.47)	350 (26.10)
Use the AI device if offered the option[c]	22 (1.64)	125 (9.31)	229 (17.06)	521 (38.82)	445 (33.16)

[a]1=strongly disagree; 2=somewhat disagree; 3=neither agree nor disagree; 4=somewhat agree; 5=strongly agree.

[b]AI: artificial intelligence.

[c]1=very unlikely; 2=somewhat unlikely; 3=undecided; 4=somewhat likely; 5=very likely.

XSL•FO
RenderX

Table . Effects of information element on perceived legibility, comprehensibility, information overload, information credibility, and perceived effectiveness of the label in informing decision-making.

	Perceived legibility[b]	Comprehensibility[b]	Information overload[b]	Information credibility[b]	Perceived effectiveness of the label in informing decision-making[a]	
					Label gives all information needed about the AI[c] device	Label helps me decide whether the device should be used in my care
	OR[d] (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)
Added value						
Information present	0.94 (0.70 to 1.27)	1.17 (0.88 to 1.56)	0.79 (0.58 to 1.05)	1.35 (1.05 to 1.73)	1.36 (1.06 to 1.74)	1.15 (0.89 to 1.48)
Information absent	Reference	Reference	Reference	Reference	Reference	Reference
Data privacy						
Opt-in	1.11 (0.83 to 1.50)	1.07 (0.80 to 1.42)	0.84 (0.63 to 1.13)	1.09 (0.85 to 1.39)	0.71 (0.56 to 0.91)	0.97 (0.75 to 1.24)
Opt-out	Reference	Reference	Reference	Reference	Reference	Reference
Expert endorsement						
Information present	1.21 (0.90 to 1.63)	0.90 (0.68 to 1.21)	0.87 (0.65 to 1.16)	1.37 (1.07 to 1.76)	1.09 (0.85 to 1.39)	1.27 (0.98 to 1.63)
Information absent	Reference	Reference	Reference	Reference	Reference	Reference
HCP[e] oversight						
Information present	1.11 (0.83 to 1.50)	1.13 (0.85 to 1.51)	0.88 (0.66 to 1.18)	1.60 (1.25 to 2.05)	1.48 (1.16 to 1.89)	1.41 (1.09 to 1.81)
Information absent	Reference	Reference	Reference	Reference	Reference	Reference
Performance						
High	1.21 (0.90 to 1.62)	1.15 (0.87 to 1.52)	0.84 (0.63 to 1.11)	1.45 (1.14 to 1.85)	1 (0.78 to 1.27)	1.16 (0.91 to 1.48)
Low	Reference	Reference	Reference	Reference	Reference	Reference
Regulatory approval						
Information present	1.16 (0.86 to 1.58)	1.29 (0.97 to 1.72)	0.99 (0.74 to 1.34)	1.55 (1.21 to 1.99)	2.05 (1.60 to 2.63)	1.11 (0.87 to 1.43)
Information absent	Reference	Reference	Reference	Reference	Reference	Reference
Device safety						
Proactive	0.92 (0.68 to 1.24)	0.91 (0.69 to 1.21)	1 (0.75 to 1.33)	1.03 (0.80 to 1.31)	0.90 (0.71 to 1.15)	1.09 (0.85 to 1.39)
Reactive	Reference	Reference	Reference	Reference	Reference	Reference
Validation						
External	0.98 (0.73 to 1.32)	1.13 (0.85 to 1.51)	1.01 (0.76 to 1.35)	1.01 (0.79 to 1.29)	1.14 (0.89 to 1.45)	1.03 (0.80 to 1.32)
Internal	Reference	Reference	Reference	Reference	Reference	Reference

[a]Model adjusted for perceived eligibility, comprehensibility, information overload, information credibility, and patient characteristics that were statistically significantly associated with the outcome.

XSL•FO
RenderX

[b]Model adjusted for patient characteristics that were statistically significantly associated with the outcome.

[c]AI: artificial intelligence.

[d]OR: odds ratio.

[e]HCP: health care provider.

Credibility

Having information about added value (OR 1.35, 95% CI 1.05-1.73), expert endorsement (OR 1.37, 95% CI 1.07-1.76), HCP oversight (OR 1.60, 95% CI 1.25-2.05), high (vs low) performance (OR 1.45, 95% CI 1.14-1.85), and regulatory approval (OR 1.55, 95% CI 1.21-1.99) were associated with higher levels of perceived credibility of the information (Table 5).

Perceived Effectiveness of Label

Information about added value (OR 1.36, 95% CI 1.06-1.74), HCP oversight (OR 1.48, 95% CI 1.16-1.89), and regulatory approval (OR 2.05, 95% CI 1.60-2.63) were associated with higher likelihood of participants reporting that the label gives them all the information they need about the AI device. Information about opt-in versus opt-out data privacy protocol (OR 0.71, 95% CI 0.56-0.91) was associated with lower likelihood of participants reporting that the label gives them all the information they need about the AI device. Information about HCP oversight (OR 1.41, 95% CI 1.09-1.81) was associated with higher likelihood of participants reporting that the AI label helps them decide whether the AI device should be used in their care (Table 5).

Trust and Intentions to Use AI

Information about expert endorsement (OR 1.30, 95% CI 1.01-1.68) and high (vs low) performance (OR 1.48, 95% CI 1.16-1.90) were associated with higher levels of trust in the results from the AI device. Information about high (vs low) performance (OR 0.77, 95% CI 0.61-0.98) and regulatory approval (OR 0.61, 95% CI 0.48-0.78) was associated with lower likelihood of participants reporting having doubts in the AI device. Information about HCP oversight (OR 0.75, 95% CI 0.57-0.97) and regulatory approval (OR 0.63, 95% CI 0.48-0.83) were associated with lower likelihood of participants reporting needing a second opinion. Information about HCP oversight (OR 1.47, 95% CI 1.12-1.94), high (vs low) performance (OR 1.59, 95% CI 1.22-2.07), and regulatory approval (OR 1.73, 95% CI 1.31-2.30) were associated with higher intention to use the AI device if offered the option (Table 6).

XSL•FO
RenderX

Table . Effects of information element on patient trust in the artificial intelligence device and intention to use the artificial intelligence device.

	Trust in the AI device[a], OR[b] (95% CI)			Intention to use the AI device if offered the option[a], OR (95% CI)
	I would trust results from this AI[c] device	I would have doubts about this AI device	I would seek a second opinion	
Added value				
Information present	0.83 (0.64 to 1.06)	1.15 (0.90 to 1.47)	0.87 (0.67 to 1.14)	0.99 (0.75 to 1.31)
Information absent	Reference	Reference	Reference	Reference
Data privacy				
Opt-in	1.07 (0.83 to 1.37)	1.07 (0.85 to 1.36)	1.27 (0.98 to 1.65)	1.03 (0.79 to 1.34)
Opt-out	Reference	Reference	Reference	Reference
Expert endorsement				
Information present	1.30 (1.01 to 1.68)	0.89 (0.70 to 1.13)	0.93 (0.71 to 1.21)	1.17 (0.89 to 1.54)
Information absent	Reference	Reference	Reference	Reference
HCP[d] oversight				
Information present	1.21 (0.94 to 1.55)	0.80 (0.63 to 1.03)	0.75 (0.57 to 0.97)	1.47 (1.12 to 1.94)
Information absent	Reference	Reference	Reference	Reference
Performance				
High	1.48 (1.16 to 1.90)	0.77 (0.61 to 0.98)	0.80 (0.62 to 1.04)	1.59 (1.22 to 2.07)
Low	Reference	Reference	Reference	Reference
Regulatory approval				
Information present	1.26 (0.98 to 1.64)	0.61 (0.48 to 0.78)	0.63 (0.48 to 0.83)	1.73 (1.31 to 2.30)
Information absent	Reference	Reference	Reference	Reference
Device safety				
Proactive	1.08 (0.84 to 1.38)	1.13 (0.89 to 1.43)	0.98 (0.76 to 1.27)	0.93 (0.71 to 1.21)
Reactive	Reference	Reference	Reference	Reference
Validation				
External	1.06 (0.83 to 1.36)	1.12 (0.88 to 1.42)	0.80 (0.61 to 1.03)	0.97 (0.74 to 1.27)
Internal	Reference	Reference	Reference	Reference

[a]Model adjusted for perceived eligibility, comprehensibility, information overload, information credibility, perceived effectiveness of the label in informing decision-making, and patient characteristics that were statistically significantly associated with the outcome.

[b]OR: odds ratio.

[c]AI: artificial intelligence.

[d]HCP: health care provider.

Discussion

Overview

To our knowledge, this is the first study to apply 2 experimental methods, DCE and SPFE, to elicit patient preferences for information about the use of AI devices in their health care. Our study provides important evidence and insights for health care and AI professionals and policy makers on the relative importance of various information factors to patient decision-making, trust, and acceptance regarding an AI device in cardiology and on subgroup differences in the impact of top preferred information factors. This work is a first step toward developing effective communication strategies about AI in health care that ensure transparency and accessibility to facilitate informed decision-making and patient-centered adoption of AI applications.

Study Design Considerations

Before delving into our findings, we first address the use of the vignette experiment approach, the hypothetical AI device, and the nonclinical study sample. While vignette experiments are a well-established method in health communication and medical decision science research, they are not without limitations. In particular, the artificial nature of vignettes can raise concerns about ecological validity, as hypothetical scenarios may not fully capture the complexities of real-world decision-making.

XSL•FO
RenderX

As a result, there is a risk that the study findings cannot be generalized directly to real-world settings [40-43].

However, in the context of our study, the use of the vignette experiments with a hypothetical AI device and a nonclinical study sample was both methodologically necessary and ethically appropriate. Methodologically, this approach allowed us to standardize and systematically manipulate the information elements of interest, while holding all other aspects of the scenario constant. This level of experimental control is often impossible to achieve with a real AI device due to the presence of a range of uncontrolled confounding influences on patient decision-making, including prior experiences, provider communication, and exposure to media or marketing. Importantly, the ecological validity of vignette experiments depends not on literal replication of real-world settings but on whether the vignettes activated the same psychological processes that occur during real-life decision-making [41]. To enhance the realism and credibility of our vignettes, we designed the hypothetical AI device scenario in consultation with clinical and regulatory collaborators and informed by formative qualitative research with both clinicians and patients. We also sought clinician review on the vignette drafts to ensure clinical accuracy and appropriateness and then further refined their clarity and relevance by piloting through cognitive interviews with patients. These steps strengthened both the internal and ecological validity of our study design and increased the applicability of our findings to real-world AI communication contexts.

Moreover, testing experimental manipulations involving actual patients making real-time medical decisions may raise ethical concerns, particularly when issues surrounding patient trust, understanding, and decision-making about AI in health care are still evolving [69]. Our experiment focuses on comparing different information factors presented for the same device within a consistent clinical scenario. Conducting an empirical study using an actual AI device while manipulating its label would involve providing false information, such as presenting different performance metrics for the same AI device to different patients, which would constitute deception and be unethical. This approach would undermine patient autonomy in decision-making and could erode trust in their clinicians and health care system [70]. To ensure that the hypothetical scenario was relevant to study participants and consistent with real-world decision-making, we recruited adults who reported recent health care experiences through either a primary care or cardiovascular care visit within the past 3 years. Importantly, our study goal was to understand general patient preferences and responses to AI labeling across a diverse sample, which is a necessary first step before testing specific AI implementations in particular clinical populations.

Key Findings

Results from the 2 experiments offer complementary insights into the key factors shaping patients' decision-making about use of an AI device in cardiology. Results from the SPFE showed that most participants found the AI label prototypes easy to read and understand and not mentally demanding. The effects of information factors on perceived legibility, comprehensibility, and information overload were not statistically significant, indicating that patient preferences were not due to a lack of understanding of these labels. Results from the DCE showed that information about provider oversight, regulatory approval, high device performance, and AI's added value were the most influential in increasing patient trust in the AI device and their willingness to use the device in their health care. Patients placed less importance on information about opt-in versus opt-out data privacy protocol, expert endorsement, external versus internal validation protocol, and proactive versus reactive device safety management protocol.

Results from the SPFE reinforced these findings. First, information about AI's added value, expert endorsement, HCP oversight, high performance, and regulatory approval was linked to higher perceived credibility of the AI label. In addition, information about added value, HCP oversight, and regulatory approval increased the likelihood that participants felt that they had all the necessary information about the device. Interestingly, information about HCP oversight was the only factor that significantly influenced participants' perception of the label's usefulness in deciding whether to use the device in their care. Information about expert endorsement, high performance, and regulatory approval was also associated with greater trust in the AI device's results and reduced doubts about the device. Furthermore, information about HCP oversight and regulatory approval lowered the likelihood of participants seeking a second opinion. Finally, information about HCP oversight, high performance, and regulatory approval contributed to a higher intention to use the AI device if offered.

The DCE also showed that participant characteristics, including recency of last medical checkup, familiarity with AI, health literacy, numeracy, need for cognition, age, gender, and race or ethnicity, shaped how strongly information about AI's added value, device performance, HCP oversight, and regulatory approval impacted trust and acceptance of the AI device. Information about AI's added value had greater effects among participants who had more recent medical checkups, were less familiar with AI, had higher health literacy, had higher need for cognition, identified as women, or identified as people of color. Similarly, information about high device performance had stronger effects among those with more recent medical checkups, lower familiarity with AI, higher health literacy or numeracy, older age, or identified as women. Information about HCP oversight had stronger effects among participants with more recent medical checkups, lower familiarity with AI, or older age. Finally, the effects of information about regulatory approval were stronger among those with recent checkups, higher health literacy or numeracy, or women.

Implications for Practice

Our findings have important implications for the implementation of AI in health care. To start, providing transparent and accessible information about HCP oversight, regulatory approval, device performance, and the device's added value is critical for building patient trust and acceptance of AI technology [27]. These 4 information elements were the most influential in shaping patients' willingness to use an AI device, highlighting the need for health care systems to clearly

XSL•FO
RenderX

communicate them when introducing AI devices to patients. Health care systems could consider developing decision aids that allow patients to explore different aspects of a specific AI device in collaboration with their providers, helping them weigh the benefits and risks of using the device in their care [26]. This is especially important for patients with lower baseline trust or limited familiarity with AI.

Our findings also highlight the critical importance of information about endorsement and oversight from regulatory bodies and HCPs in boosting patients' confidence and reducing doubts about the AI device's effectiveness and safety, which can lead to higher trust in the device and lower likelihood of needing second opinions. This strong preference for human oversight and approval echoes literature on algorithm aversion, which shows that people are often reluctant to trust algorithmic decision-making systems even when they outperform humans [71,72]. This aversion is especially pronounced in decisions involving high uncertainty [73] and for tasks perceived as subjective [74]; however, it can be mitigated when there is room for human oversight and modification [75]. While algorithm aversion may initially make patients hesitant to trust AI, patient trust in their provider and the health system can lead to inflated expectations of AI's positive impact on care and potentially result in overreliance when patients lack the expertise to judge when to rely on it. Transparency and patient engagement are therefore critical to calibrating trust appropriately [28,76]. Notably, information about HCP oversight was particularly influential in helping patients decide whether an AI device should be used in their care. This finding is consistent with previous research showing that patients are more receptive to AI technologies when AI supports, rather than replaces, the decisions of trusted human HCPs, underscoring the value of a "human in the loop" approach to AI implementation that provides patients a sense of accountability and assurance regarding the safety and effectiveness of these technologies [12-14,18,77,78]. To improve patient engagement and acceptance of AI devices, patient-facing communications should explicitly emphasize providers' active role in reviewing, interpreting, and overriding AI outputs.

Three information elements, opt-in versus opt-out data privacy, external versus internal validation, and proactive versus reactive device safety management protocols, were found to be relatively less influential in shaping participants' trust and decision-making regarding the AI device. This variation in the prioritization of information elements reflects how patients actively calibrate trust in AI technologies and advances beyond prior empirical work that was limited to identifying information relevant to patient trust [26,28,39,79]. Specifically, patients tend to anchor their trust in the AI device on signals that are personally meaningful from a layperson's perspective, such as information about HCP oversight, regulatory approval, expert endorsement, and device performance, rather than the more abstract procedural indicators such as data privacy, validation, and safety management. While critical from a policy and ethical standpoint [38], these issues may be too abstract or poorly understood by a layperson to effectively translate into meaningful differences in care [80]. In addition, the terms used in the descriptions of these elements, including "deidentified data," "tested and evaluated," "fix issues promptly," and "device will be audited," may have signaled to participants that basic privacy, validity, and safety protections were in place, especially in the presence of a strong credibility cue [81]—the study instruction that their HCP recommended the AI device. As a result, many participants may have relied on heuristics to reassure themselves about privacy, validity, and safety and focused their mental effort on other aspects that are more directly related to the benefits and risks of using the device in their care [82,83]. These findings echo previous research showing that increasing transparency alone is insufficient for calibrated trust, and such information must also be accessible and meaningful to users [35,80,84]. Technical or abstract aspects of the AI device should be contextualized in terms of impact on the patient's care and communicated in accessible and relatable ways, for example, through plain language summaries, visuals, and metaphors, ideally developed through an iterative process of co-design and testing with patients [85-88]. These findings also highlight a potential tension between what patients find most useful in decision-making and what is required by regulators or ethics guidelines. For example, although data privacy is a central focus in recent AI regulations [89], participants in our study placed relatively lower importance on it. This suggests that patients may not benefit from regulatory-required information in the same way as other stakeholders and may sometimes prefer information that is not routinely made available as part of regulatory clearance processes. To support appropriately calibrated trust of AI technologies in health care, patient-facing communication strategies should emphasize the information patients value most, while also clearly addressing critical ethical and regulatory considerations [90]. Importantly, to make the abstract concepts more relatable, ethical and regulatory features should be framed in terms of their tangible benefits and risks to patients. Communication efforts must also strike a careful balance: offering reassurance while avoiding overstating benefits or certainty, which could create misplaced confidence in AI technology [76,91].

Moreover, patient-facing communication and education about AI in health care should be tailored to meet the unique needs and preferences of different patient groups [92,93]. For example, information about a device's added value was found to be particularly influential for patients who were less familiar with AI, had high reading health literacy, had recent routine checkups, and identified as women and people of color. Emphasizing the role of HCPs in overseeing AI use was found to be especially important for patients who are less familiar with AI, had recent routine checkups, and were aged 55 years or older. To effectively address these diverse patient needs and preferences, communication strategies could include tailoring language to patient reading levels and incorporating visual aids, infographics, and videos to help patients better understand complex information about the AI device and make informed decisions about their care. In addition, engaging patients, patient advocacy groups, and community organizations in co-designing patient-facing AI communication and educational materials including AI device labels could help address concerns from underrepresented groups who may have different levels of comfort with and trust in AI technologies and the health care system [94]. Providing AI communication and education

XSL•FO
RenderX

materials in various formats and at different time points may improve the accessibility of information about AI and better meet the diverse needs of different patient groups. Health care systems can consider offering in-person consultations during clinical visits, along with printed and digital materials for patients to take home, ensuring that patients have the opportunity to review and understand the information at their own pace [39]. Moreover, it is the responsibility of health care systems to ensure that providers are well informed about AI devices and are capable of communicating their benefits and risks effectively to diverse patient populations. Health care systems could establish standardized guidelines and best practices for patient-facing communication and education about AI in health care. Training providers on how to effectively discuss AI devices with diverse patient populations and address their concerns would be instrumental in ensuring that patients receive clear, transparent information that aids informed decision-making and builds merited trust in AI technologies. Table 7 summarizes key findings and implications for practice.

Table . Summary of key findings, implications, and recommendations.

Findings	Implications and recommendations
High impact information • Information about HCP[a] oversight, regulatory approval, high device performance, and AI's[b] added value led to largest increases in trust and willingness to use. • These elements also boosted the AI label's credibility, increased trust in the results, reduced doubts, and enhanced patient confidence in having enough information about the device for decision-making.	• Patients prioritize personally meaningful, concrete information over abstract procedural details. • Clearly communicate information about HCP oversight, regulatory approval, device performance, and added value, as these elements are crucial for building patient trust and enabling acceptance of AI technologies. • Develop decision aids that help patients and providers explore the benefits and risks of specific AI devices together, especially important for patients with lower baseline trust in or limited familiarity with AI.
Lower impact information: Participants placed less importance on information about opt-in versus opt-out data privacy, model validation protocol, and safety management protocol.	• Abstract issues may be poorly understood, so patients rely on heuristics and focus on aspects directly related to benefits and risks. • Technical or abstract aspects should be contextualized for patient care and communicated via plain language, visuals, and metaphors. • Communication should emphasize what patients value while addressing critical ethical and regulatory considerations.
Role of human oversight: Information about HCP oversight and regulatory approval consistently boosted patient confidence, reduced doubts, increased trust, lowered the need for second opinions, and increased intention to use the AI-enabled device if offered.	• Patients prefer AI that supports, rather than replaces, human decision makers, underscoring the importance of a "human in the loop" approach. • Communicating provider oversight can inflate expectations of AI's benefits; transparency and engagement are key to calibrating trust based on human oversight. • Patient-facing communication should emphasize providers' role in reviewing, interpreting, and approving or overriding AI outputs. • Ensure that providers are well informed about AI devices and can effectively communicate benefits, risks, and limitations to patients.
Subgroup differences: Impact of information elements varied by participant characteristics. • Added value: stronger effect for recent checkups, low AI familiarity, high literacy or need for cognition, women, and people of color. • High performance: stronger for recent checkups, low AI familiarity, high literacy or numeracy, older age, and women. • HCP oversight: stronger for recent checkups, low AI familiarity, and older age. • Regulatory approval: stronger for recent checkups, high literacy or numeracy, and women.	• Tailor language to patient reading levels and use visuals, infographics, and videos to explain complex AI information. • Co-design materials with patients, advocacy groups, and community organizations to address diverse trust levels and concerns. • Provide materials in multiple formats and at different time points to meet diverse patient needs. • Establish standardized guidelines and best practices for patient-facing AI communication. • Train providers on how to effectively discuss benefits, risks, and limitations of AI devices with diverse patient populations.

[a]HCP: health care provider.

[b]AI: artificial intelligence.

Strengths, Limitations, and Future Directions

This research has several strengths. First, we used innovative experimental methods to elicit patient preferences for information about AI in health care, which could be applied to other use cases and medical specialties. Our methods can also be used to develop and evaluate other patient-facing AI communication and education materials. We applied a rigorous process for selection of information elements, including a rapid literature review and 3 qualitative studies, which strengthened the validity of our findings.

Our findings should be interpreted in the context of the limitations. The study sample, recruited from ResearchMatch.org, is a convenience sample and may not necessarily represent the US adult population. To address this, we intentionally oversampled racial and ethnic minority

XSL•FO
RenderX

populations to ensure their representation and enable comparisons between people of color and non-Hispanic White patients in AI information preferences. Since nearly all participants had health insurance, caution is needed when generalizing the findings to uninsured populations. In addition, while selection bias might be a concern, we lack data to compare participants with nonparticipants, which may impact the generalizability of our findings. Nonetheless, our methods of evaluating the importance of information elements can be applied to other AI devices, medical conditions, and other types of AI communication and education efforts.

Our study sample may not closely reflect the demographic and clinical characteristics of patients most likely to use AI-enabled cardiology devices. While all study participants reported having had a primary care or cardiology visit within the past 3 years, we could not clinically verify that they were seeking cardiovascular care during the study period. Therefore, findings may not generalize to clinical populations actively facing decisions about AI-enabled cardiac care. Future research should replicate this work among clinical populations to validate and extend the findings. Furthermore, we used a hypothetical AI-enabled cardiology device to maintain experimental control and isolate the effects of specific information elements. However, this approach may not fully capture the complexity and nuance of real-world patient decision-making. As a result, the ecological validity of our findings may be limited. To build upon and extend the current findings, future research should examine how patients actively seeking cardiovascular care respond to labeling content for real-world AI-enabled devices in clinical practice.

Our study focused on the impact of patient-facing informational content in AI labeling, and it did not directly assess how provider recommendation might influence patient trust and acceptance of AI technologies. To hold provider recommendation constant across experimental conditions, we instructed all participants to assume that their provider had recommended the use of the hypothetical AI device. We acknowledge the important role of provider recommendation in patient adoption of AI technologies and encourage future research to explore how provider recommendation and label content interact to shape patient decision-making.

We examined patient information preferences and responses to AI label prototypes at a single time point prior to actual use of an AI device. As AI technologies become more integrated into routine health care and daily life, patients may gain greater familiarity with AI and shift information priorities after repeated exposure and use. Future research should adopt longitudinal study designs to examine how patient information needs, preferences, trust, and acceptance evolve over time and across different stages of their health care journey.

The information elements presented to participants were informed by our prior qualitative research; however, they may not fully reflect the complete range of information that could appear on an actual AI device label. In addition, because our study focused solely on the informational content of AI labels, the format, layout, and visual design of the label prototypes were simplified and standardized for ease of delivery through a web-based survey and may not fully reflect the look or feel of AI device labels used in real-world clinical practices. Future research is needed to examine how variations in display format (eg, static vs interactive), information modality (eg, textual vs infographic), delivery channel (eg, digital vs printed), and timing (eg, previsit, point-of-care, postvisit) influence patient comprehension, trust, and decision-making regarding AI technologies in health care. Finally, it is critical to involve patients in co-designing and testing AI communication strategies to ensure that the materials are accessible, engaging, and aligned with the needs and preferences of diverse patient populations.

Conclusions

Through experimental studies, our research underscores the critical importance of transparent and accessible patient-facing information about AI devices and their impact on patient trust and acceptance. Information on HCP oversight, regulatory approval, device performance, and its added value emerged as pivotal factors influencing patient decision-making. Tailoring communication to meet the diverse needs and preferences of patient subgroups is essential for effective and equitable AI adoption in health care. Patient-centered communication strategies, coupled with comprehensive education for HCPs, would ensure that AI technologies are integrated into health care in a way that empowers patients to make informed choices and support overall patient care.

Funding

This work was supported by the US Food and Drug Administration (FDA) of the US Department of Health and Human Services (HHS), award number U01FD005938, totaling US $712,431 with 100% funded by FDA/HHS. The FDA/HHS had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement by, FDA/HHS, or the US government.

Conflicts of Interest

This research was supported by the Food and Drug Administration (FDA) through the Center of Excellence in Regulatory Science and Innovation (CERSI) grant to Yale University and Mayo Clinic. XZ offers scientific input to research studies through a contracted services agreement between Mayo Clinic and Exact Sciences. MM and JEM received research grants from Arnold Ventures through Yale University. JEM serves on Alexion Pharmaceuticals Ethics Committee and is the founder of Bioethics International. BAB offers scientific input to research studies through a contracted services agreement between Mayo Clinic and Anumana, Inc. AMS, SAM, DWY, and JLR report no additional competing interests.

XSL•FO
RenderX

Multimedia Appendix 1
Choice sets.
[DOCX File, 21 KB - jmir_v28i1e75615_app1.docx]

Multimedia Appendix 2
2IV8-3 fractional factorial experimental design.
[DOCX File, 21 KB - jmir_v28i1e75615_app2.docx]

Multimedia Appendix 3
Subgroup differences in effects of information factors on the probability of the artificial intelligence device being trusted.
[DOCX File, 18 KB - jmir_v28i1e75615_app3.docx]

Multimedia Appendix 4
Subgroup differences in effects of information factors on the probability of the artificial intelligence device being accepted.
[DOCX File, 20 KB - jmir_v28i1e75615_app4.docx]

Checklist 1
CHERRIES checklist.
[DOCX File, 20 KB - jmir_v28i1e75615_app5.docx]

References

1. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 2020;3(1):118. [doi: 10.1038/s41746-020-00324-0] [Medline: 32984550]
2. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018 Oct;2(10):719-731. [doi: 10.1038/s41551-018-0305-z]
3. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. BMC Med Inform Decis Mak 2021 Dec;21(1):125. [doi: 10.1186/s12911-021-01488-9]
4. Johnson KW, Torres Soto J, Glicksberg BS, et al. Artificial Intelligence in cardiology. J Am Coll Cardiol 2018 Jun;71(23):2668-2679. [doi: 10.1016/j.jacc.2018.03.521]
5. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. J Am Coll Cardiol 2017 May;69(21):2657-2664. [doi: 10.1016/j.jacc.2017.03.571]
6. Lopez-Jimenez F, Attia Z, Arruda-Olson AM, et al. Artificial intelligence in cardiology: present and future. Mayo Clin Proc 2020 May;95(5):1015-1039. [doi: 10.1016/j.mayocp.2020.01.038]
7. Young AT, Amara D, Bhattacharya A, Wei ML. Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. Lancet Digit Health 2021;3(9):e599-e611. [doi: 10.1016/S2589-7500(21)00132-1]
8. Moy S, Irannejad M, Manning SJ, et al. Patient perspectives on the use of artificial intelligence in health care: a scoping review. J Patient Cent Res Rev 2024;11(1):51-62. [doi: 10.17294/2330-0698.2029] [Medline: 38596349]
9. Bala S, Keniston A, Burden M. Patient perception of plain-language medical notes generated using artificial intelligence software: pilot mixed-methods study. JMIR Form Res 2020 Jun 5;4(6):e16670. [doi: 10.2196/16670] [Medline: 32442148]
10. Adams SJ, Tang R, Babyn P. Patient perspectives and priorities regarding artificial intelligence in radiology: opportunities for patient-centered radiology. J Am Coll Radiol 2020 Aug;17(8):1034-1036. [doi: 10.1016/j.jacr.2020.01.007] [Medline: 32068006]
11. Palmisciano P, Jamjoom AAB, Taylor D, Stoyanov D, Marcus HJ. Attitudes of patients and their relatives toward artificial intelligence in neurosurgery. World Neurosurg 2020 Jun;138:e627-e633. [doi: 10.1016/j.wneu.2020.03.029]
12. Nelson CA, Pérez-Chada LM, Creadore A, et al. Patient perspectives on the use of artificial intelligence for skin cancer screening. JAMA Dermatol 2020 May 1;156(5):501. [doi: 10.1001/jamadermatol.2019.5014]
13. Tran VT, Riveros C, Ravaud P. Patients' views of wearable devices and AI in healthcare: findings from the ComPaRe e-cohort. NPJ Digit Med 2019;2(53):53. [doi: 10.1038/s41746-019-0132-y] [Medline: 31304399]
14. Lennartz S, Dratsch T, Zopfs D, Persigehl T, Maintz D, Große Hokamp N, et al. Use and control of artificial intelligence in patients across the medical workflow: single-center questionnaire study of patient perspectives. J Med Internet Res 2021;23(2):e24221. [doi: 10.2196/24221]
15. McCradden MD, Sarker T, Paprica PA. Conditionally positive: a qualitative study of public perceptions about using health data for artificial intelligence research. BMJ Open 2020;10(10). [doi: 10.1136/bmjopen-2020-039798]
16. Jutzi TB, Krieghoff-Henning EI, Holland-Letz T, et al. Artificial intelligence in skin cancer diagnostics: the patients' perspective. Front Med (Lausanne) 2020;7(233):233. [doi: 10.3389/fmed.2020.00233] [Medline: 32671078]
17. Aggarwal R, Farag S, Martin G, Ashrafian H, Darzi A. Patient perceptions on data sharing and applying artificial intelligence to health care data: cross-sectional survey. J Med Internet Res 2021;23(8):e26162. [doi: 10.2196/26162]

XSL•FO
RenderX

18. Richardson JP, Smith C, Curtis S, Watson S, Zhu X, Barry B, et al. Patient apprehensions about the use of artificial intelligence in healthcare. NPJ Digit Med 2021;4(1):140. [doi: 10.1038/s41746-021-00509-1]
19. Musbahi O, Syed L, Le Feuvre P, Cobb J, Jones G. Public patient views of artificial intelligence in healthcare: a nominal group technique study. Digit Health 2021;7:20552076211063682. [doi: 10.1177/20552076211063682] [Medline: 34950499]
20. Lennox-Chhugani N, Chen Y, Pearson V, Trzcinski B, James J. Women's attitudes to the use of AI image readers: a case study from a national breast screening programme. BMJ Health Care Inform 2021 Mar;28(1). [doi: 10.1136/bmjhci-2020-100293]
21. Brereton TA, Malik MM, Lifson M, Greenwood JD, Peterson KJ, Overgaard SM. The role of artificial intelligence model documentation in translational science: scoping review. Interact J Med Res 2023 Jul 14;12:e45903. [doi: 10.2196/45903]
22. Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. Presented at: Proceedings of the Conference on Fairness, Accountability, and Transparency; Jan 29-31, 2019. [doi: 10.1145/3287560.3287596]
23. Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. Commun ACM 2021 Dec;64(12):86-92. [doi: 10.1145/3458723]
24. Holland S, Hosny A, Newman S, Joseph J, Chmielinski K. The dataset nutrition label. Data Prot Privacy 2020;12(12):1-26. [doi: 10.5040/9781509932771.ch-001]
25. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. NPJ Digit Med 2020;3(1):41. [doi: 10.1038/s41746-020-0253-3] [Medline: 32219182]
26. Robinson R, Liday C, Lee S, et al. Artificial intelligence in health care-understanding patient information needs and designing comprehensible transparency: qualitative study. JMIR AI 2023;2:e46487. [doi: 10.2196/46487] [Medline: 38333424]
27. Zhang Z, Genc Y, Wang D, Ahsen ME, Fan X. Effect of AI explanations on human perceptions of patient-facing AI-powered healthcare systems. J Med Syst 2021 May 4;45(6):64. [doi: 10.1007/s10916-021-01743-6] [Medline: 33948743]
28. Kostick-Quenet K, Lang BH, Smith J, Hurley M, Blumenthal-Barby J. Trust criteria for artificial intelligence in health: normative and epistemic considerations. J Med Ethics 2024 Jul 23;50(8):544-551. [doi: 10.1136/jme-2023-109338] [Medline: 37979976]
29. Martin RW, Brogård Andersen S, O'Brien MA, Bravo P, Hoffmann T, Olling K, et al. Providing balanced information about options in patient decision aids: an update from the international patient decision aid standards. Med Decis Making 2021;41(7):780-800. [doi: 10.1177/0272989X211021397]
30. Swar B, Hameed T, Reychav I. Information overload, psychological ill-being, and behavioral intention to continue online healthcare information search. Comput Human Behav 2017 May;70:416-425. [doi: 10.1016/j.chb.2016.12.068]
31. Kelly B, O'Donoghue A, Parvanta S, et al. Effects of additional context information in prescription drug information sheets on comprehension and risk and efficacy perceptions. J Pharm Policy Pract 2022 Mar 1;15(1):15. [doi: 10.1186/s40545-021-00386-9] [Medline: 35232474]
32. Bester J, Cole CM, Kodish E. The limits of informed consent for an overwhelmed patient: clinicians' role in protecting patients and preventing overwhelm. AMA J Ethics 2016 Sep 1;18(9):869-886. [doi: 10.1001/journalofethics.2016.18.9.peer2-1609] [Medline: 27669132]
33. Mello MM, Char D, Xu SH. Ethical obligations to inform patients about use of AI tools. JAMA 2025 Sep 2;334(9):767. [doi: 10.1001/jama.2025.11417]
34. Yoo DW, Stroud AM, Zhu X, Miller JE, Barry B. Toward patient-centered AI fact labels: leveraging extrinsic trust cues. DIS (Des Interact Syst Conf) 2025:676-690. [doi: 10.1145/3715336.3735758]
35. Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Wortman Vaughan JW, Wallach H. Manipulating and measuring model interpretability. Presented at: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; May 6-13, 2021. [doi: 10.1145/3411764.3445315]
36. Elgin CY, Elgin C. Ethical implications of AI-driven clinical decision support systems on healthcare resource allocation: a qualitative study of healthcare professionals' perspectives. BMC Med Ethics 2024;25(1):148. [doi: 10.1186/s12910-024-01151-8]
37. Device labeling. US Food and Drug Administration. URL: https://www.fda.gov/medical-devices/overview-device-regulation/device-labeling [accessed 2024-12-20]
38. Mooghali M, Stroud AM, Yoo DW, Barry BA, Grimshaw AA, Ross JS, et al. Trustworthy and ethical AI-enabled cardiovascular care: a rapid review. BMC Med Inform Decis Mak 2024;24(1):247. [doi: 10.1186/s12911-024-02653-6]
39. Stroud A, Minteer S, Zhu X, Ridgeway J, Miller J, Barry B. Patient information needs for transparent and trustworthy cardiovascular artificial intelligence: a qualitative study. PLOS Digit Health 2025;4(4):e0000826. [doi: 10.1371/journal.pdig.0000826]
40. Sheringham J, Kuhn I, Burt J. The use of experimental vignette studies to identify drivers of variations in the delivery of health care: a scoping review. BMC Med Res Methodol 2021;21(1):81. [doi: 10.1186/s12874-021-01247-4]
41. Evans SC, Roberts MC, Keeley JW, et al. Vignette methodologies for studying clinicians' decision-making: validity, utility, and application in ICD-11 field studies. Int J Clin Health Psychol 2015;15(2):160-170. [doi: 10.1016/j.ijchp.2014.12.001] [Medline: 30487833]
42. Atzmüller C, Steiner PM. Experimental vignette studies in survey research. Methodology (Gott) 2010 Jan;6(3):128-138. [doi: 10.1027/1614-2241/a000014]

XSL•FO
RenderX

43. van Vliet LM, van der Wall E, Albada A, Spreeuwenberg PMM, Verheul W, Bensing JM. The validity of using analogue patients in practitioner–patient communication research: systematic review and meta-analysis. J Gen Intern Med 2012 Nov;27(11):1528-1543. [doi: 10.1007/s11606-012-2111-8]
44. Merlo G, van Driel M, Hall L. Systematic review and validity assessment of methods used in discrete choice experiments of primary healthcare professionals. Health Econ Rev 2020 Dec 9;10(1):39. [doi: 10.1186/s13561-020-00295-8] [Medline: 33296066]
45. Quaife M, Terris-Prestholt F, Di Tanna GL, Vickerman P. How well do discrete choice experiments predict health choices? A systematic review and meta-analysis of external validity. Eur J Health Econ 2018 Nov;19(8):1053-1066. [doi: 10.1007/s10198-018-0954-6] [Medline: 29380229]
46. Soekhai V, Bekker-Grob EW, Ellis AR, Vass CM. Discrete choice experiments in health economics: past, present and future. Pharmacoeconomics 2019;37(2):201-226. [doi: 10.1007/s40273-018-0734-2]
47. Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. Health Econ 2012;21(2):145-172. [doi: 10.1002/hec.1697]
48. Vanniyasingam T, Cunningham CE, Foster G, Thabane L. Simulation study to determine the impact of different design features on design efficiency in discrete choice experiments. BMJ Open 2016;6(7):e011985. [doi: 10.1136/bmjopen-2016-011985]
49. Traets F, Sanchez DG, Vandebroek M. Generating optimal designs for discrete choice experiments in R: the idefix package. J Stat Softw 2020 Nov;96(3). [doi: 10.18637/jss.v096.i03]
50. Carlsson F, Martinsson P. Design techniques for stated preference methods in health economics. Health Econ 2003;12(4):281-294. [doi: 10.1002/hec.729]
51. Burgess L, Street DJ. Optimal designs for choice experiments with asymmetric attributes. J Stat Plan Inference 2005 Sep;134(1):288-301. [doi: 10.1016/j.jspi.2004.03.021]
52. Szinay D, Cameron R, Naughton F, Whitty JA, Brown J, Jones A. Understanding uptake of digital health products: methodology tutorial for a discrete choice experiment using the Bayesian efficient design. J Med Internet Res 2021 Oct 11;23(10):e32365. [doi: 10.2196/32365] [Medline: 34633290]
53. Grömping U. R Package FrF2 for creating and analyzing fractional factorial 2-level designs. J Stat Softw 2014 Jan;56(1):1-56. [doi: 10.18637/jss.v056.i01]
54. Cacioppo JT, Petty RE. The need for cognition. J Pers Soc Psychol 1982;42(1):116-131. [doi: 10.1037/0022-3514.42.1.116]
55. Lins de Holanda Coelho G, H P Hanel P, J Wolf L. The very efficient assessment of need for cognition: developing a six-item version. Assessment 2020 Dec;27(8):1870-1885. [doi: 10.1177/1073191118793208] [Medline: 30095000]
56. Thompson HS, Valdimarsdottir HB, Winkel G, Jandorf L, Redd W. The Group-Based Medical Mistrust Scale: psychometric properties and association with breast cancer screening. Prev Med 2004 Feb;38(2):209-218. [doi: 10.1016/j.ypmed.2003.09.041]
57. Morris NS, MacLean CD, Chew LD, Littenberg B. The Single Item Literacy Screener: evaluation of a brief instrument to identify limited reading ability. BMC Fam Pract 2006 Mar 24;7(1):21. [doi: 10.1186/1471-2296-7-21] [Medline: 16563164]
58. McNaughton CD, Cavanaugh KL, Kripalani S, Rothman RL, Wallston KA. Validation of a Short, 3-Item Version of the Subjective Numeracy Scale. Med Decis Making 2015 Nov;35(8):932-936. [doi: 10.1177/0272989X15581800] [Medline: 25878195]
59. Fagerlin A, Zikmund-Fisher BJ, Ubel PA, Jankovic A, Derry HA, Smith DM. Measuring numeracy without a math test: development of the Subjective Numeracy Scale. Med Decis Making 2007;27(5):672-680. [doi: 10.1177/0272989X07304449] [Medline: 17641137]
60. Gierisch JM, Earp JA, Brewer NT, Rimer BK. Longitudinal predictors of nonadherence to maintenance of mammography. Cancer Epidemiol Biomarkers Prev 2010 Apr 1;19(4):1103-1111. [doi: 10.1158/1055-9965.EPI-09-1120]
61. de Bekker-Grob EW, Donkers B, Jonker MF, Stolk EA. Sample size requirements for discrete-choice experiments in healthcare: a practical guide. Patient 2015 Oct;8(5):373-384. [doi: 10.1007/s40271-015-0118-z]
62. Harris PA, Scott KW, Lebo L, Hassan N, Lightner C, Pulley J. ResearchMatch: a national registry to recruit volunteers for clinical research. Acad Med 2012 Jan;87(1):66-73. [doi: 10.1097/ACM.0b013e31823ab7d2] [Medline: 22104055]
63. R Core Team. R: a language and environment for statistical computing. : R Foundation for Statistical Computing; 2023 URL: https://www.R-project.org/ [accessed 2025-12-08]
64. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. JOSS 2019;4(43):1686. [doi: 10.21105/joss.01686]
65. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw 2015 Oct;67(1). [doi: 10.18637/jss.v067.i01]
66. Christensen RHB. Ordinal—regression models for ordinal data. The Comprehensive R Archive Network. URL: https://CRAN.R-project.org/package=ordinal [accessed 2025-12-08]
67. Arel-Bundock V. Marginaleffects: predictions, comparisons, slopes, marginal means, and hypothesis tests. The Comprehensive R Archive Network. URL: https://CRAN.R-project.org/package=marginaleffects [accessed 2025-12-08]
68. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004;6(3):e34. [doi: 10.2196/jmir.6.3.e34]

XSL•FO
RenderX

69. Ratti E, Morrison M, Jakab I. Ethical and social considerations of applying artificial intelligence in healthcare—a two-pronged scoping review. BMC Med Ethics 2025 May 27;26(1):68. [doi: 10.1186/s12910-025-01198-1] [Medline: 40420080]
70. Wendler D. Deceiving research participants: is it inconsistent with valid consent? J Med Philos 2022 Nov 5;47(4):558-571. [doi: 10.1093/jmp/jhac014]
71. Burton JW, Stein MK, Jensen TB. A systematic review of algorithm aversion in augmented decision making. Behav Decision Making 2020 Apr;33(2):220-239. [doi: 10.1002/bdm.2155]
72. Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol Gen 2015 Feb;144(1):114-126. [doi: 10.1037/xge0000033] [Medline: 25401381]
73. Dietvorst BJ, Bharti S. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. Psychol Sci 2020 Oct;31(10):1302-1314. [doi: 10.1177/0956797620948841] [Medline: 32916083]
74. Castelo N, Bos MW, Lehmann DR. Task-dependent algorithm aversion. J Mark Res 2019 Oct;56(5):809-825. [doi: 10.1177/0022243719851788]
75. Dietvorst BJ, Simmons JP, Massey C. Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them. Manage Sci 2018 Mar;64(3):1155-1170. [doi: 10.1287/mnsc.2016.2643]
76. Nong P, Ji M. Expectations of healthcare AI and the role of trust: understanding patient views on how AI will impact cost, access, and patient-provider relationships. J Am Med Inform Assoc 2025 May 1;32(5):795-799. [doi: 10.1093/jamia/ocaf031]
77. Longoni C, Bonezzi A, Morewedge CK. Resistance to medical artificial intelligence. J Consum Res 2019 Dec 1;46(4):629-650. [doi: 10.1093/jcr/ucz013]
78. Cai CJ, Reif E, Hegde N, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. Presented at: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; May 2-9, 2019. [doi: 10.1145/3290605.3300234]
79. Lee JD, See KA. Trust in automation: designing for appropriate reliance. Hum Factors 2004 Jan 1;46(1):50-80. [doi: 10.1518/hfes.46.1.50_30392]
80. Wischnewski M, Krämer N, Müller E. Measuring and understanding trust calibrations for automated systems: a survey of the state-of-the-art and future directions. Presented at: CHI '23; Apr 23-28, 2023. [doi: 10.1145/3544548.3581197]
81. Chaiken S. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. J Pers Soc Psychol 1980;39(5):752-766. [doi: 10.1037/0022-3514.39.5.752]
82. Chen S, Duckworth K, Chaiken S. Motivated heuristic and systematic processing. Psychol Inq 1999 Jan;10(1):44-49. [doi: 10.1207/s15327965pli1001_6]
83. Chaiken S, Ledgerwood A. A theory of heuristic and systematic information processing. In: Handbook of Theories of Social Psychology 2012, Vol. 1:246-266. [doi: 10.4135/9781446249215.n13]
84. Felzmann H, Fosch-Villaronga E, Lutz C, Tamò-Larrieux A. Towards transparency by design for artificial intelligence. Sci Eng Ethics 2020 Dec;26(6):3333-3361. [doi: 10.1007/s11948-020-00276-4] [Medline: 33196975]
85. Stoll M, Kerwer M, Lieb K, Chasiotis A. Plain language summaries: a systematic review of theory, guidelines and empirical research. PLoS One 2022;17(6):e0268789. [doi: 10.1371/journal.pone.0268789] [Medline: 35666746]
86. Galmarini E, Marciano L, Schulz PJ. The effectiveness of visual-based interventions on health literacy in health care: a systematic review and meta-analysis. BMC Health Serv Res 2024;24(1):718. [doi: 10.1186/s12913-024-11138-1]
87. Wang T, Voss JG. Effectiveness of pictographs in improving patient education outcomes: a systematic review. Health Educ Res 2021 Mar 23;36(1):9-40. [doi: 10.1093/her/cyaa046]
88. Liu X. Use of metaphor in provider-patient communication in medical settings: a systematic review. Patient Educ Couns 2025 Aug;137:109184. [doi: 10.1016/j.pec.2025.109184]
89. King J, Meinhardt C. Rethinking privacy in the ai era policy provocations for a data-centric world. SSRN J 2024 Feb 22. [doi: 10.2139/ssrn.5446957]
90. Williamson SM, Prybutok V. Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in ai-driven healthcare. Appl Sci (Basel) 2024;14(2):675. [doi: 10.3390/app14020675]
91. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. J Med Internet Res 2020 Jun 19;22(6):e15154. [doi: 10.2196/15154] [Medline: 32558657]
92. Young A, Tordoff J, Smith A. "What do patients want?" Tailoring medicines information to meet patients' needs. Res Social Adm Pharm 2017 Nov;13(6):1186-1190. [doi: 10.1016/j.sapharm.2016.10.006] [Medline: 27818214]
93. Smets EMA, Menichetti J, Lie HC, Gerwing J. What do we mean by "tailoring" of medical information during clinical interactions? Patient Educ Couns 2024 Feb;119:108092. [doi: 10.1016/j.pec.2023.108092]
94. Adus S, Macklin J, Pinto A. Exploring patient perspectives on how they can and should be engaged in the development of artificial intelligence (AI) applications in health care. BMC Health Serv Res 2023 Oct 26;23(1):1163. [doi: 10.1186/s12913-023-10098-2] [Medline: 37884940]

Abbreviations

AI: artificial intelligence
AME: average marginal effect

XSL•FO
RenderX

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
DCE: discrete choice experiment
HCP: health care provider
IRB: institutional review board
ML: machine learning
OR: odds ratio
SPFE: single profile factorial experiment

Edited by A Sakhuja; submitted 07.Apr.2025; peer-reviewed by E Reiter, N Hu; revised version received 22.Oct.2025; accepted 08.Nov.2025; published 12.Jan.2026.

Please cite as:
Zhu X, Stroud AM, Minteer SA, Yoo DW, Ridgeway JL, Mooghali M, Miller JE, Barry BA
Key Information Influencing Patient Decision-Making About AI in Health Care: Survey Experiment Study
J Med Internet Res 2026;28:e75615
URL: https://www.jmir.org/2026/1/e75615
doi:10.2196/75615

© Xuan Zhu, Austin M Stroud, Sarah A Minteer, Dong Whi Yoo, Jennifer L Ridgeway, Maryam Mooghali, Jennifer E Miller, Barbara A Barry. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Intervention in Health Misinformation Using Large Language Models for Automated Detection, Thematic Analysis, and Inoculation: Case Study on COVID-19

Samira Malek[1], MS; Christopher Griffin[2,3], PhD; Robert D Fraleigh[2], PhD; Robert Lennon[4], MD, JD; Vishal Monga[5], PhD; Lijiang Shen[6], PhD

[1]Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA, United States

[2]Applied Research Laboratory, Pennsylvania State University, University Park, PA, United States

[3]Department of Mathematics, Pennsylvania State University, University Park, PA, United States

[4]PrimeCare Medical, Harrisburg, PA, United States

[5]Department of Electrical Engineering, Pennsylvania State University, University Park, PA, United States

[6]Department of Communication Arts and Sciences, Pennsylvania State University, 211 Sparks Building, University Park, PA, United States

Corresponding Author:
Lijiang Shen, PhD
Department of Communication Arts and Sciences, Pennsylvania State University, 211 Sparks Building, University Park, PA, United States

Abstract

Background: The rapid growth of social media as an information channel has enabled the swift spread of inaccurate or false health information, significantly impacting public health. This widespread dissemination of misinformation has caused confusion, eroded trust in health authorities, led to noncompliance with health guidelines, and encouraged risky health behaviors. Understanding the dynamics of misinformation on social media is essential for devising effective public health communication strategies.

Objective: This study aims to present a comprehensive and automated approach that leverages large language models (LLMs) and machine learning techniques to detect misinformation on social media, uncover the underlying causes and themes, and generate refutation arguments, facilitating control of its spread and promoting public health outcomes by inoculating people against health misinformation.

Methods: We use 2 datasets to train 3 LLMs, namely, BERT, T5, and GPT-2, to classify documents into 2 categories: misinformation and nonmisinformation. In addition, we use a separate dataset to identify misinformation topics. To analyze these topics, we applied 3 topic modeling algorithms—Latent Dirichlet Allocation, Top2Vec, and BERTopic—and selected the optimal model based on performance evaluated across 3 metrics. Using a prompting approach, we extract sentence-level representations for the topics to uncover their underlying themes. Finally, we design a prompt text capable of identifying misinformation themes effectively.

Results: The trained BERT model demonstrated exceptional performance, achieving 98% accuracy in classifying misinformation and nonmisinformation, with a 44% reduction in false-positive rates for artificial intelligence–generated misinformation. Among the 3 topic modeling approaches used, BERTopic outperformed the others, achieving the highest metrics with a Coherence Value of 0.41, Normalized Pointwise Mutual Information of −0.086, and Inverse Rank-Biased Overlap of 0.99. To address the issue of unclassified documents, we developed an algorithm to assign each document to its closest topic. In addition, we proposed a novel method using prompt engineering to generate sentence-level representations for each topic, achieving a 99.6% approval rate as “appropriate” or “somewhat appropriate” by 3 independent raters. We further designed a prompt text to identify themes of misinformation topics and developed another prompt capable of detecting misinformation themes with 82% accuracy.

Conclusions: This study presents a comprehensive and automated approach to addressing health misinformation on social media using advanced machine learning and natural language processing techniques. By leveraging LLMs and prompt engineering, the system effectively detects misinformation, identifies underlying themes, and provides explanatory responses to combat its spread. The proposed method was tested on an English language COVID-19–related dataset and has not been evaluated on real-world online social media data; the experiments were conducted offline.

(J Med Internet Res 2026;28:e75500) doi:10.2196/75500

KEYWORDS

large language models; topic modeling; COVID-19; misinformation; prompt engineering; machine learning

XSL•FO
RenderX

Introduction

Misinformation and inaccurate beliefs and knowledge about health can substantially undermine well-being by fueling confusion, eroding trust in reliable medical advice, and prompting risky behaviors such as rejecting vaccines, turning to scientifically unproven home remedies, or neglecting protective measures amid clear dangers [1-8]. These inaccuracies often circulate rapidly via social media, exploiting emotional narratives that overshadow fact-based content and leading individuals to question the legitimacy of evidence-based interventions [5,8-11]. Repeated exposure to misinformation reduces health literacy and can reinforce people's belief in falsehoods, making them more likely to view credible health authorities with skepticism [12-14]. As a result, misinformation weakens the success of prevention and treatment strategies, paving the way for heightened disease transmission, avoidable complications, and deteriorating outcomes at both individual and community levels [15-18].

An illustration comes from the COVID-19 pandemic, which saw an unprecedented surge of misinformation and conspiracy theories—labeled an "infodemic" by the World Health Organization (WHO) [1,12]. False remedies, unverified claims on the origins of the virus, and politicized narratives about preventive measures severely hampered containment efforts [19-21]. While proven strategies such as mask wearing, vaccination, and physical distancing were promoted by scientific authorities, social media rumors cast doubt on vaccine safety and the reality of the virus itself, discouraging people from getting vaccinated or seeking appropriate medical care [2,22-24]. This breakdown in adherence prolonged outbreaks, overloaded health infrastructures, and ultimately jeopardized global health and economic stability [25].

A parallel can be drawn from discussions around the human papillomavirus (HPV) vaccine, which has proven crucial in preventing various HPV-related cancers, including cervical cancer that claims thousands of lives each year [26-29]. Widespread misinformation about adverse effects and conspiracies regarding its necessity led to a significant portion of unvaccinated adolescents, heightening the likelihood of HPV infection and future malignancies [30]. This trend not only increased the burden on public health systems but also underscored the power of misinformation to undermine trust in legitimate medical counsel.

In recent years, social media has become a central and highly accessible source of information for millions of users worldwide [31]. However, its ability to rapidly disseminate content—including unfounded claims—creates fertile ground for large-scale propagation of misinformation. Given the sheer volume of posts, manual monitoring and analysis of such content are impractical [31,32] Consequently, developing and using automated, data-driven methods to understand and manage the dynamics of digital misinformation are essential for preserving accurate information and safeguarding public trust.

In this study, we propose an automated system designed to identify whether a given text contains misinformation. If misinformation is detected, the system analyzes the theme of the misinformation and provides a refutation argument (inoculation) to help prevent its spread on social media and enhance public health awareness. To achieve this, we leverage a large language model (LLM) to detect misinformation effectively. Furthermore, we demonstrate that enriching datasets significantly improves the detection of misinformation generated by both humans and AI. Recent advances in LLMs, such as ChatGPT, have enabled the generation of increasingly sophisticated misinformation, which poses challenges for traditional machine learning (ML) methods in distinguishing AI-generated misinformation [33,34]. While prior research has highlighted the effectiveness of deep learning methods in classifying health-related misinformation, these efforts have predominantly focused on content generated by humans [35,36]. Moreover, our proposed process generates sentence-level descriptions of misinformation topics, eliminating the need for manual interpretation. However, prior approaches relied on ML-based methods that produced word-level topic representations, which required manual interpretation to form coherent sentence-level topics—introducing potential human errors and subjective biases [1,22]. Similar challenges arise in other ML-based applications, such as optimizing models in industries where manual calibration of parameters can lead to inefficiencies and errors. For example, recent research has demonstrated that data-driven models can enhance predictive accuracy and automate decision-making, reducing human intervention in systems that rely on complex data streams [37-40]. Inspired by these advances, our process generates sentence-level descriptions of misinformation topics, eliminating the need for manual interpretation. In addition, we introduce an algorithm to assign documents to the most relevant topics. This addresses the limitation of many ML-based topic modeling algorithms, which often leave some documents unclassified. Our process also identifies overarching themes of misinformation topics automatically, providing a high-level understanding of the underlying reasons for misinformation categorization. Although the COVID-19 pandemic serves as our illustrative case due to its scale and data availability, the underlying challenges we address—rapid online spread, emotionally charged narratives, and declining trust—are common across other health contexts (eg, HPV vaccines) and beyond. Our approach does not rely on COVID-specific lexicons or handcrafted rules. Instead, the proposed Misinformation Detection and Inoculation Process (MDIP) is domain-agnostic. It ingests free English text, induces topics using standard models, transforms word lists into sentence-level descriptors through targeted LLM prompting, organizes them into hierarchical themes (guided by coherence, diversity metrics, and generic embeddings), and maps themes to refutation templates. Each stage has the potential to be applied to other health domains and misinformation settings, provided that the AI is properly trained with the topic- or domain-specific data.

Many previous studies have focused on individual aspects of the misinformation problem, such as detection or topic analysis [1,35,41], but have not integrated these steps into a unified framework for intervention. Our MDIP and Misinformation Detection and Inoculation System (MDIS) frameworks unify misinformation detection, topic modeling, thematic refutation, and public health intervention into a single, automated

XSL•FO
RenderX

workflow. This end-to-end approach enables health teams to move beyond merely identifying misinformation to actively and effectively countering it.

In the study by He et al [42], a method was proposed to generate per-claim counterresponses. While generating responses tailored to each specific piece of misinformation can be more informative and persuasive, such approaches require paired datasets of misinformation posts and response arguments for model training—datasets that are difficult to construct. Moreover, misinformation often mutates through paraphrasing and subtle edits; claim-specific pipelines are fragile in the face of such variation [43,44]. By contrast, our theme-level refutation approach is robust to these surface changes. More recently, LLM-based topic modeling approaches, such as TopicGPT [45] and other methods [46,47], have leveraged the capabilities of powerful pretrained models such as ChatGPT and LLaMA. While these methods benefit from the models' deep understanding of language, they often require passing entire documents through parameter-rich models, which leads to increased latency and computational costs compared with traditional pipelines. Furthermore, they typically yield only single-level sets of word topics. In contrast, our hybrid framework balances efficiency and expressiveness: we first use a traditional topic inducer to efficiently uncover the underlying structure and then apply LLM prompting (via ChatGPT) where it provides the added value. Specifically, the LLM is used to transform word lists into sentence-level topic labels and to organize topics into hierarchical themes. This targeted use of LLMs preserves computational efficiency while producing richer, hierarchical, and deployment-ready representations that are well suited for downstream tasks such as detection, monitoring, and refutation.

Methods

Study Design

In this study, we propose the MDIP, a comprehensive framework designed to analyze the dynamics of misinformation automatically and develop an MDIS. The MDIS end-to-end pipeline (1) flags misinformation, (2) explains what it is about via topics and higher-level themes, and (3) returns a concise, theme-matched refutation. The components are modular and feed one another in a simple data flow. No step in MDIP and MDIS uses disease-specific features; inputs are raw text and model hyperparameters chosen by intrinsic criteria (topic coherence and diversity). This makes the pipeline directly applicable to other misinformation corpora after swapping in the relevant documents. The MDIP framework is structured into four interconnected sections, each addressing a critical aspect of misinformation management:

1. *Detect misinformation*: Collect a labeled dataset and then train LLMs to classify text documents as either misinformation or nonmisinformation, providing a foundation for identifying false narratives.
2. *Misinformation topics*: Here, a topic modeling algorithm is applied to uncover the key topics within misinformation datasets. This step helps categorize misinformation into specific subject areas, enabling a better understanding of its thematic structure.
3. *Topic descriptions*: This section uses prompt engineering and the results of the previous section to enhance interpretability by generating sentence-level representations for each topic, moving beyond traditional word-level outputs. These descriptive summaries provide meaningful context for understanding the essence of each topic.
4. *Provide refutation*: In the final step, topic descriptions and extracted themes are used to design a specific prompt that identifies the underlying themes of misinformation. The system then generates clear and contextually relevant refutation arguments tailored to the detected misinformation themes. These arguments are designed to counter false narratives, improve public understanding, and mitigate the spread of misinformation. The refutations are a key component in the psychological inoculation-based misinformation mitigation intervention. As a metaphor to medical vaccination, the typical inoculation strategy consists of an attack message (ie, as a small weakened or deactivated dose of the virus) and refutation or counterarguments against the attack message (ie, as the immune system's reaction to the vaccine when it is injected or otherwise enters the human body) [48].

By integrating these components, MDIP enables the development of MDIS, an intelligent and automated system capable of detecting misinformation, identifying its themes, and delivering refutations to combat its impact on public health. The overall architecture of the proposed framework is illustrated in Figure 1. Figure 1A outlines the 4 stages of the MDIP: supervised detection of misinformation, topic modeling, generation of interpretable topic descriptions, and theme-based refutation. Figure 1B presents the end-to-end workflow of the MDIS, which processes new text inputs to produce a misinformation classification, assign the text to a thematic category, and generate a matched refutation. Finally, Figure 1C provides an example of the user-facing output, demonstrating how the system delivers both a misinformation warning and a concise, contextually relevant refutation.

XSL•FO
RenderX

Figure 1. (**A**) Overview of the Misinformation Detection and Inoculation Process (MDIP) integrates four main stages: (1) misinformation detection through supervised classification, (2) topic modeling of misinformation texts and assignment of outliers, (3) generation of interpretable topic descriptions and aggregation into higher-level themes, and (4) theme detection and provision of theme-linked refutations via prompt engineering for detected misinformation. (**B**) Misinformation Detection and Inoculation System (MDIS) workflow: For any new text, the system outputs (1) the misinformation decision, (2) its most likely theme, and (3) a theme-matched refutation. (**C**) Example: Illustration of how the system delivers the final user-facing output, providing a warning about misinformation and the matched refutation text. LLM: large language model.

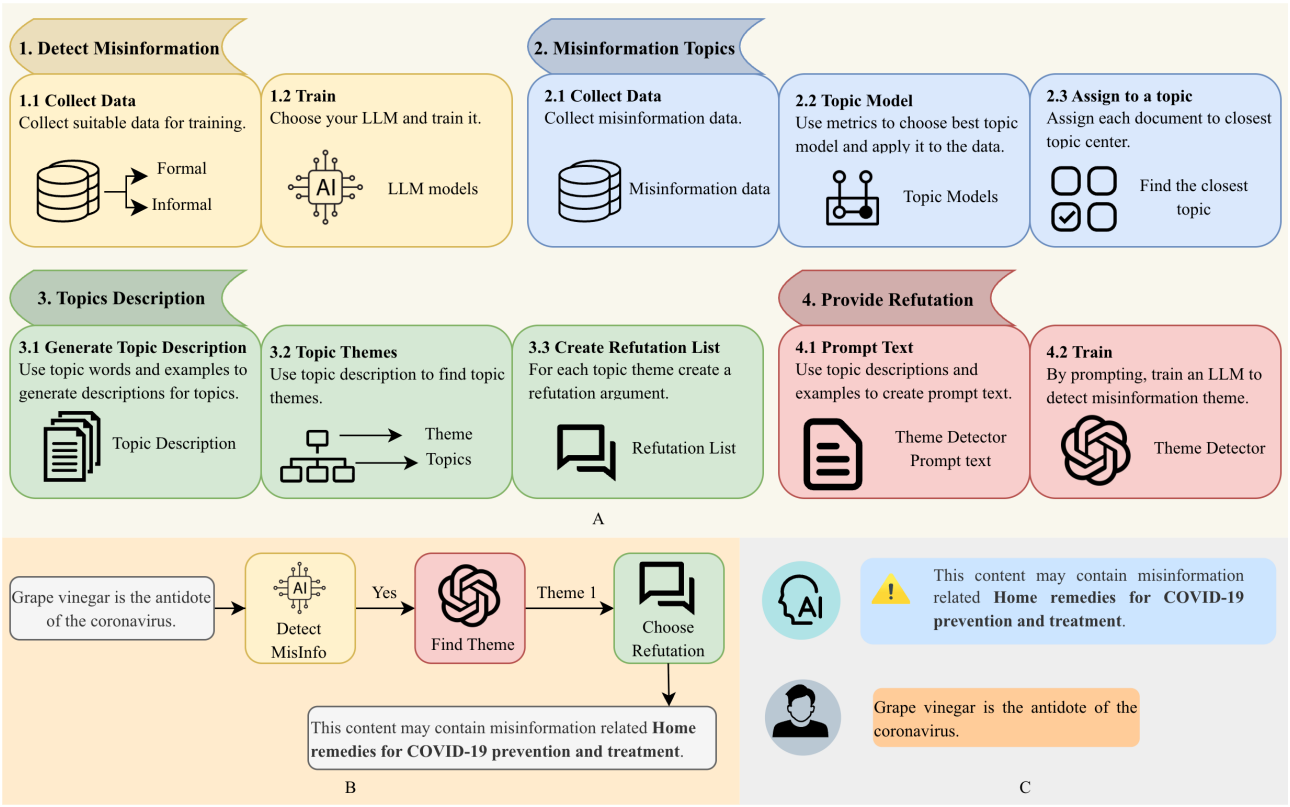

Detect Misinformation

Misinformation detection in text documents has become a critical area of research due to the growing prevalence of misleading or false information online. To address this challenge, we use classifiers based on LLMs. These LLMs are trained to categorize text into 2 classes: misinformation and nonmisinformation, as shown in the first part of Figure 1A.

Our classifier was trained using 2 complementary datasets, each providing diverse linguistic characteristics to enhance performance in detecting misinformation. The first dataset, the AAAI 2021 Competition Dataset [49], consists of misinformation sourced from social media platforms such as Facebook and X (formerly known as Twitter). This dataset reflects the informal, conversational style of social media, characterized by casual tone, nonstandard grammar, and the use of slang. The second dataset, COVID_19FNIR [50], includes misinformation presented in formal, structured language, offering a stark contrast to the informal nature of the first dataset. By incorporating these 2 datasets, we trained 3 different LLMs to detect misinformation effectively across a wide spectrum of communication styles. The blend of informal and formal language enabled the model to better generalize, achieving improved accuracy and robustness in identifying misinformation, whether generated by humans or artificial intelligence (AI).

Traditionally, researchers collect human-written data from social media platforms such as Twitter and Facebook, label them as misinformation or nonmisinformation, and then train a deep neural network to classify such documents [35,36,51,52]. However, recent studies have demonstrated that deep neural networks trained exclusively on human-written datasets exhibit weaker accuracy in detecting AI-generated misinformation compared with human-written misinformation. This discrepancy arises because AI-generated misinformation often adopts formal language styles similar to accurate information shared by credible sources such as the WHO and the Centers for Disease Control and Prevention (CDC) on official social media accounts [33,34].

In this research, by combining a dataset with formal language and another with informal language (enriching the dataset with different language types and more misinformation), we demonstrate that LLMs achieve reasonable accuracy in detecting AI-generated misinformation. This approach ensures better generalization and robustness, bridging the gap in identifying misinformation across diverse linguistic styles.

Misinformation Topics

As outlined in the second section of Figure 1A, our approach involves 3 key steps. First, we collect misinformation data. Next, we select and compare topic modeling algorithms based on specific features and metrics to identify the most effective model. Finally, we design an algorithm that assigns topics to new or unclassified documents.

XSL•FO
RenderX

To identify misinformation topics, we used one of the largest datasets of verified COVID-19 claims, the IFCN dataset, which has been extensively used in related research [41,53,54]. We applied 3 topic modeling algorithms—Latent Dirichlet Allocation (LDA) [55], Top2Vec [56,57], and BERTopic [57,58]—to analyze this dataset.

To evaluate and compare the performance of these algorithms, we selected 3 metrics: Coherence Value (CV) [59], Normalized Pointwise Mutual Information (NPMI) [59], and Inverse Rank-Biased Overlap (IRBO) [60]. CV and NPMI measure the coherence of the topics, ensuring that they are logically consistent, being human interpretable, and meaningful. IRBO, on the other hand, evaluates the diversity of the topics generated by the model, which is crucial for ensuring broad coverage of the dataset's content. Since our focus is on misinformation within health-related social media data, coherence and diversity are particularly important to ensure that topics are both interpretable and representative.

After selecting the best-performing topic model, we developed an algorithm to address the issue of unclassified documents. This algorithm assigns topics to new or previously unassigned documents, ensuring comprehensive topic coverage and improved usability of the model for real-world applications.

Topic Description

Topic modeling algorithms typically produce word-level representations for each topic. While these representations provide insight into the most relevant words associated with a topic, they often lack the semantic depth necessary to precisely identify the specific topic within a document. This limitation arises because word-level outputs fail to capture the context and relationships between words that define the overarching theme of a topic [51].

Recent advancements in LLMs have demonstrated their ability to generate high-quality, contextually relevant outputs with minimal or zero additional training by designing carefully crafted inputs—referred to as prompt engineering [61]. Leveraging this capability, we address the limitations of word-level representations by using prompt engineering techniques to generate sentence-level representations for each topic. These sentence-level representations capture the context and essence of the topic, enabling a more accurate and interpretable understanding of the document content.

Subsequently, these sentence-level representations are used to identify and articulate the overarching themes of the topics, also at the sentence level. This approach provides a more comprehensive view of the thematic structure within the document corpus. Finally, recognizing that all documents within the dataset share a common underlying reason for being classified as misinformation, we develop a tailored response list for each topic theme. The third section of Figure 1A illustrates these 3 steps.

Provide Refutation

In the final step of our proposed method, as illustrated in the final part of Figure 1A, we identify the overarching theme of misinformation and provide a corresponding response from a preconstructed response list. This response list is developed in the preceding step based on the identified themes.

To determine the themes of misinformation, we use prompt engineering techniques. By designing carefully crafted and contextually appropriate prompt text, we effectively extract the underlying themes associated with misinformation. This approach allows us to translate complex word-level or sentence-level representations into meaningful thematic insights.

By identifying misinformation themes and providing precise, theme-based responses, our method aims to enhance public health knowledge and reduce the spread of misinformation. This proactive approach not only mitigates the risks associated with false or misleading information but also fosters a more informed and resilient society.

Proposed System

Following the completion of four foundational steps—(1) detecting misinformation, (2) identifying misinformation topics, (3) describing topics, and (4) providing refutations—we develop our comprehensive MDIS, which consists of three key components.

1. *Detection of misinformation*: The system begins by determining whether a given document is misinformation.
2. *Identification of misinformation themes*: If the document is classified as misinformation, the system analyzes its content to identify the underlying misinformation themes. This process involves extracting thematic representations that provide a clearer understanding of the document's misleading aspects.
3. *Providing refutations*: Finally, the system generates a detailed refutation argument for the identified misinformation themes. These arguments are derived from a predesigned response list tailored to address specific misinformation themes effectively.

All 3 components of the system are demonstrated with a practical example, as illustrated in Figure 1B and C. This example highlights how the system operates cohesively to detect misinformation, uncover its thematic structure, and deliver accurate refutations, ultimately contributing to a more informed and resilient public.

Ethical Considerations

This study did not involve human participants, human tissue, or the collection of identifiable private information by the authors. All analyses were conducted on previously collected, publicly available, and deidentified datasets, obtained solely for research purposes. Specifically, the data sources include: (1) the AAAI 2021 COVID-19 Fake News Detection Competition dataset, originally released as part of the AAAI Conference on Artificial Intelligence shared task, in which all social media content was anonymized and distributed for noncommercial research use only [49]; (2) the COVID-19 FNIR (Fake News and Information Reliability) dataset, introduced by prior studies for misinformation detection research and released in deidentified form for academic use [50]; and (3) the International Fact-Checking Network (IFCN) COVID-19 fact-checking corpus, which aggregates publicly available

XSL•FO
RenderX

fact-check articles produced by IFCN-certified organizations and contains no personal or sensitive individual-level data [62]. According to the US Department of Health and Human Services Common Rule (45 CFR §46.104(d)), secondary research involving publicly available, deidentified data does not constitute human subjects research and is therefore exempt from Institutional Review Board review [63]. The research complied with all relevant ethical standards and data use policies and poses no risk to individuals or communities. The study's sole objective is to advance computational methods for understanding and mitigating the spread of health misinformation.

Results

Text Classification

Due to the exceptional performance of LLMs across a wide range of AI tasks, we leveraged 3 prominent LLMs to fine-tune them for COVID-19 text classification. These models—BERT (Bidirectional Encoder Representations from Transformers), GPT-2 (Generative Pre-trained Transformer 2), and T5-base (Text-to-Text Transfer Transformer)—are renowned for their ability to understand and process natural language with high accuracy and contextual awareness.

BERT is particularly effective in handling text classification tasks due to its bidirectional context understanding, which allows it to capture nuanced language patterns [64]. GPT-2 excels in text generation and classification by leveraging its autoregressive architecture to predict sequences in a given context [65]. Finally, T5-base under a unified framework that reformulates all NLP tasks as a text-to-text problem, making it versatile and effective across various domains [66].

To conduct this study, we combined the AAAI 2021 competition dataset with the COVID-19 FNIR dataset. The data were split into training, testing, and validation sets with proportions of 67%, 17%, and 16%, respectively.

Accuracy, F_1-score, Recall, and Precision are standard metrics for evaluating classification models. Accuracy measures the proportion of all predictions that are correct, providing an overall performance indicator but sometimes masking class imbalances. Precision quantifies the fraction of predicted positives that are truly positive, reflecting how often the model avoids false alarms. Recall (or sensitivity) measures the fraction of actual positives that the model successfully identifies, highlighting its ability to capture relevant cases [67]. F_1-score is the harmonic mean of precision and recall, balancing the trade-off between the two. In health-related text classification tasks such as misinformation detection, reasonable thresholds are often set required due to the risks of misclassification—for instance, aiming for Accuracy >0.80, F_1-score ≥0.75, Recall ≥0.75, and Precision ≥0.70—to ensure both reliable detection and practical usability in downstream inoculation public health applications. Table 1 shows the evaluation metrics, including Accuracy, F_1-score, Recall, and Precision, for all 3 models on the test dataset. Among these, BERT achieved the highest performance, with an accuracy of *98%* on the test data. This result highlights BERT's ability to handle complex linguistic structures and its effectiveness in fine-tuning for domain-specific tasks such as COVID-19 text classification.

The confusion matrices in Figure 2 for BERT, GPT-2, and T5-base further illustrate the performance of these models, providing a detailed breakdown of true positives, true negatives, false positives, and false negatives, which helps in understanding their classification strengths and potential areas for improvement.

Table . Performance metrics (Accuracy, F_1-score, Recall, and Precision) on the test dataset for 3 models: BERT-base, GPT-2, and T5-base.

Model	Accuracy	F_1-score	Recall	Precision
BERT	0.9848	0.9854	0.9896	0.9812
GPT-2[a]	0.9460	0.9495	0.9841	0.9117
T5-base (Generic Condition)	0.9763	0.9763	0.9763	0.9764

[a]GPT-2: Generative Pre-trained Transformer.

Figure 2. Confusion matrices illustrate the performance of the 3 binary classification models (BERT, GPT-2, and T5-base). GPT-2: Generative Pre-trained Transformer.

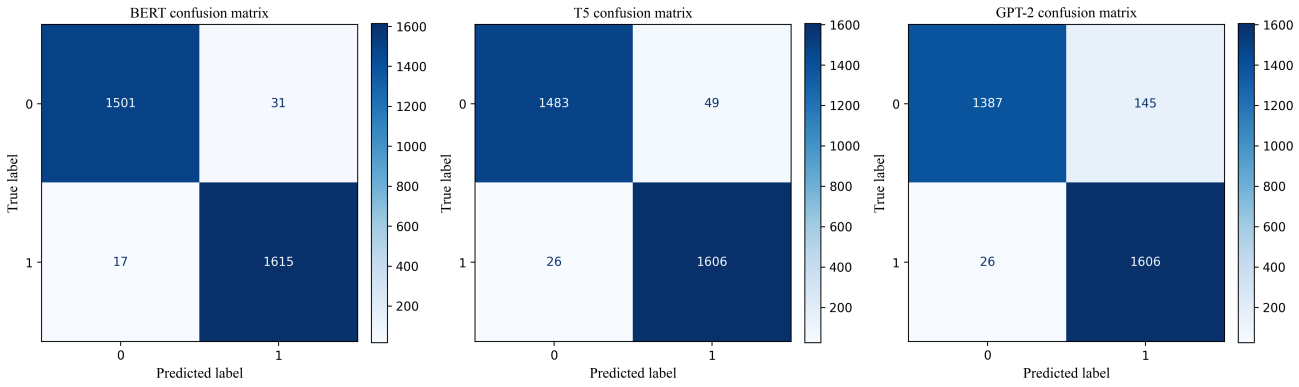

XSL•FO
RenderX

To evaluate the accuracy of our model on AI-generated data, we used the dataset provided in the study by Du et al [35]. We tested our fine-tuned BERT model on this dataset, and the results are shown in Table 2. The findings indicate a significant reduction in the number of false positives, decreasing from 27 to 15, representing a 44% improvement. In addition, Figure 3 shows the confusion matrix for our fine-tuned BERT model when applied to the AI-generated misinformation dataset.

Table . False-positive and true-negative results obtained from testing the fine-tuned BERT model on our combined dataset, compared with the results reported in the study by Zhou et al [33].

Model	FP[a]	TN[b]
Our	15	485
Zhou et al [33]	27	473

[a]FP: false-positive.

[b]TN: true-negative.

Figure 3. The figure displays the confusion matrix of our fine-tuned BERT model evaluated on the artificial intelligence–generated dataset [3].

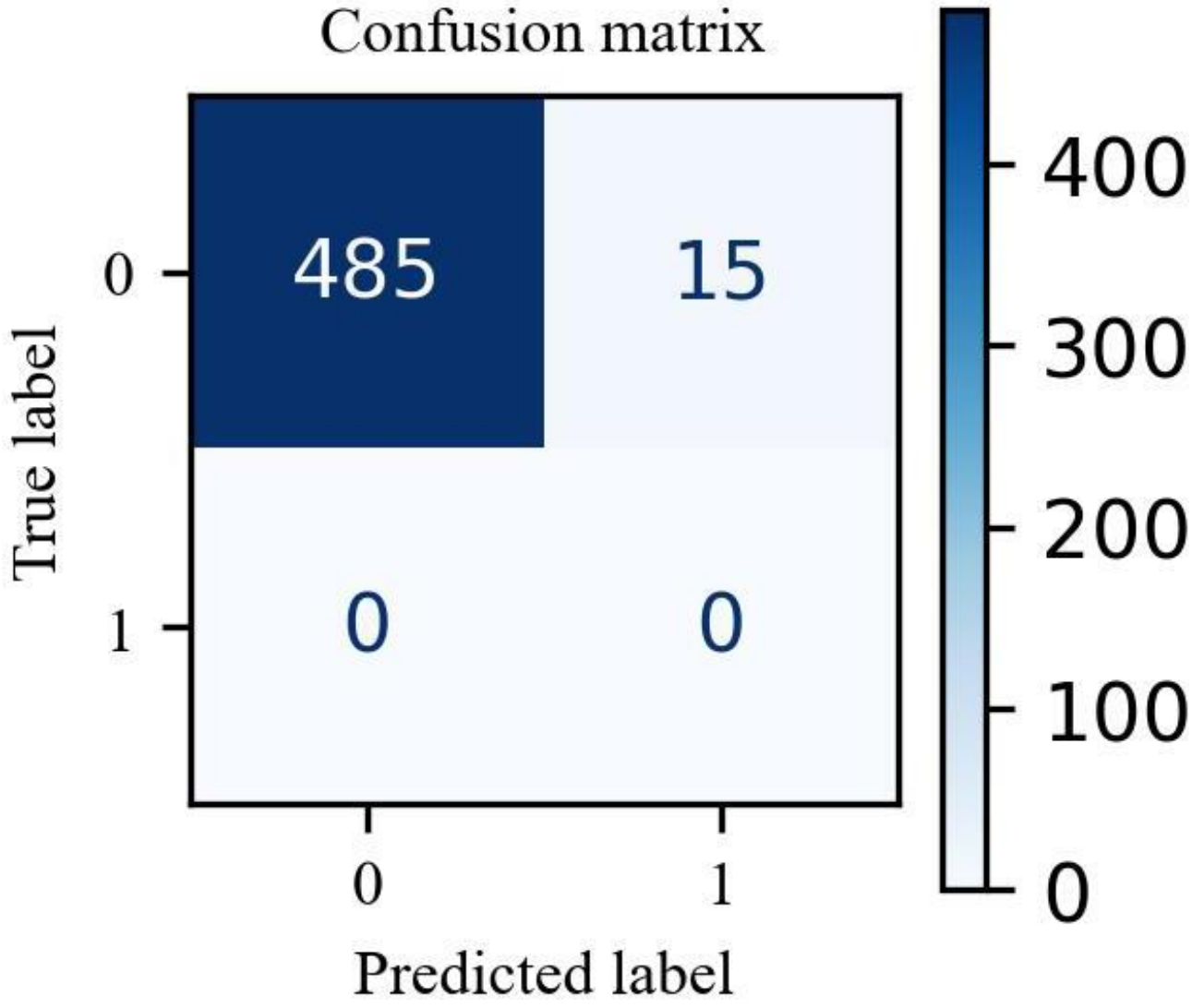

Topic Models

We used the IFCN dataset, one of the largest datasets on COVID-19 pandemic, to apply and evaluate topic modeling approaches. Three models were tested: LDA, Top2Vec, and BERTopic. After applying each topic model, the top 10 words associated with each topic were selected, and the 3 metrics—CV, NPMI, and IRBO—were computed to compare the models. Table 3 summarizes the results across these metrics. Among the models, BERTopic achieved the highest scores across all metrics, leading to its selection for further analysis. In practice, the reported metrics indicate that the topics are moderately coherent and interpretable but not perfectly tight. A coherence score of 0.41 means that the top words in each topic tend to appear together often enough for human analysts to assign clear labels, although some mixing of subthemes is expected. The NPMI of −0.086 is close to neutral, which is typical for short, fragmented social media posts, and suggests that while not every

XSL•FO
RenderX

word pair strongly co-occurs, the overall topics remain meaningful. Combined with the high IRBO score (0.99), these results imply that the model generates a broad, nonredundant set of topics that cover diverse misinformation themes while remaining practically usable for labeling, interpretation, and downstream health communication tasks. Since in the health-related domain, it is important to cover as many diverse topics as possible, an IRBO value above 0.7 can be considered acceptable, while a CV value equal to or above 0.4 and an NPMI close to 0 or higher are reasonable.

Table . Performance of 3 topic modeling approaches—LDA, Top2Vec, and BERTopic—evaluated across 3 metrics: Coherence Value, Normalized Pointwise Mutual Information, and Inverse Ranked Based Overlap[a].

Model	CV[b]	NPMI[c]	IRBO[d]
LDA	0.39	−0.35	0.96
Top2Vec	0.35	−0.29	0.89
BERTopic	0.41	−0.086	0.99

[a]For all metrics, higher values indicate better performance.

[b]CV: Coherence Value.

[c]NPMI: Normalized Pointwise Mutual Information.

[d]IRBO: Inverse Ranked Based Overlap.

Many topic modeling approaches, including BERTopic, often encounter limitations when applied to real-world datasets, as they are unable to assign topics to all documents. This can leave a subset of documents unclassified, reducing the overall effectiveness of the model. To address this issue, we have used the algorithm in Textbox 1, a method for ensuring comprehensive topic assignment across the dataset [68].

Textbox 1. Algorithm: assign a document to the closest topic.

1. Input:
   - Raw text documents X={d1,d2,...,dn},
   - BERTopic model parameters.
2. Topic modeling:
   - Compute topics using the BERTopic model:
     - Topics {Yj}j=1T, = BERTopic(X,P)

     where T is the number of topics and Y_j contains documents that are assigned to the topic j.
3. Sentence embeddings:
   - Transform documents X into vector representations using a sentence transformer such as BERT embedding.
4. Dimensionality reduction:
   - Apply Uniform Manifold Approximation and Projection for dimensionality reduction on the vector representations.
5. Cluster centers:
   - For each topic j, compute the center of the cluster:

   (1)tj=∑i∈ jxinj

     where n_j is the number of documents in topic j, and x_i is the reduced vector representation of document.
6. Topic assignment for unassigned documents:
   - For every document d_i that is not assigned to a topic by the BERTopic model, or for any new document:
     - Assign the document to the topic j that maximizes the cosine similarity between the document vector x_i and the cluster center t_j :

   (2)Argmax jxi · tj ‖ xi ‖ ‖ tj ‖

This approach not only ensures that every document in the dataset is assigned a topic but also enhances the interpretability and usability of the topic modeling results. By leveraging the semantic structure of the dataset, our algorithm effectively

XSL•FO
RenderX

bridges the gap between unassigned documents and existing topic clusters, making it a robust solution for comprehensive topic coverage.

Topic Description

As described in the "Methods" section, topic modeling algorithms typically produce word-level representations of topics. While useful for identifying key terms associated with a topic, these representations often lack sufficient contextual information, making interpretation challenging. To overcome this limitation, we developed a structured prompt framework (outlined in Textbox 2 and used the advanced capabilities of LLMs, specifically ChatGPT-4.0, to generate sentence-level representations for the identified topics. These sentence-level representations provide richer context and more interpretable descriptions, enabling a deeper understanding of the topics.

Textbox 2. The prompt structure and 1 example to find topics description.

Topic description prompt structure:

System role:

Topic main words: [Top 10 words]

Topic document examples: [5 closest examples to the center of the topic]

User role:

"Describe topic in a short phrase?"

Topic description prompt example:

System role:

Topic main words: ["masks," "mask," "face," "wearing," "wear," "use," "oxygen," "hypoxia," "cause," and "you''].

- Topic document examples:
  1. Centers for Disease Control and Prevention (CDC) does not recommend wearing masks.
  2. The US CDC contradicted itself by advising people to wear cloth masks against the novel coronavirus while also saying masks do not stop smoke inhalation during a wildfire.
  3. The World Health Organization changed its mind about masks and now says that they can increase the risk of infection.
  4. Nonmedical masks are ineffective in preventing the spread of the disease, are circulating online.
  5. Whether CDC had scheduled announcement that all should wear masks for everyday life.

User role: "Describe topic in a short phrase?"

Output answer: "Controversies and debates over mask wearing and its effectiveness"

The prompt includes the top 10 most representative words for each topic as identified by the topic modeling algorithm, and to add context and depth to the topic descriptions, we select 5 documents that are closest to the center of the corresponding topic cluster. The selection of these documents is guided by cosine similarity, performed using equations 1 and 2, which measure the proximity of documents to the cluster center in the semantic space. An example of this process is provided in Textbox 2, illustrating how the top words and representative documents are integrated into the prompt to produce a high-quality sentence-level representation.

By combining these elements, we construct detailed and context-rich prompts that guide ChatGPT-4.0 in generating coherent and semantically accurate sentence-level topic representations. This approach ensures that the abstract themes identified by topic modeling are translated into human-readable and interpretable descriptions.

To evaluate the quality of the generated topic descriptions, we engaged 3 independent raters to assess the descriptions based on 3 categories: appropriate, somewhat appropriate, and not appropriate. The evaluation results are shown in Table 4 and highlight that the majority of topic descriptions were well received. Specifically, the total proportion of accepted descriptions (the sum of those rated as appropriate and somewhat appropriate) was 99.6%. This acceptance rate demonstrates the effectiveness and reliability of the proposed method for generating meaningful and contextually relevant topic descriptions. There was perfect agreement in 144 out of 169 (85.2%) of the sentences. Two out of 3 raters agreed on category in 24 out of 169 (14.2%) of the sentences. There was a single instance in which no raters agreed in 1 out of 169 (0.6%) of the sentences.

XSL•FO
RenderX

Table . Percentage of topic descriptions rated as “appropriate,” “somewhat appropriate,” and “not appropriate” by each rater, along with the total number of accepted topic descriptions, calculated as the sum of those rated “appropriate” and “somewhat appropriate.”

Raters	Appropriate (%)	Somewhat appropriate (%)	Not appropriate (%)	Total accepted (%)
Rater 1	98.23	1.77	0	100
Rater 2	94.67	5.33	0	100
Rater 3	89.94	8.88	1.18	98.82
Average	94.28	5.32	0.39	99.6

After generating concise descriptions for each topic, we used the structured prompt framework outlined in Textbox 3, which includes a list of these topic descriptions. This structured prompt was then input into the ChatGPT-4.0 API to further refine and categorize the topics into overarching themes.

Textbox 3. Structure of the prompt for identifying topic themes.

Finding topic themes prompt structure:

System role:

The following are topics related to COVID-19 pandemic. Go through all topics and categorize them into relevant groups. Mention topics number for each category.

User role:

Topics description list

The output from this process not only provides a clear categorization of topics into distinct themes but also includes a concise description for each theme. This step ensures that the topics are grouped in a meaningful and interpretable way, facilitating a deeper understanding of the data’s thematic structure.

The categorized topics and their corresponding theme descriptions are provided in Figures 4 and 5, showcasing the effectiveness of the proposed method in generating coherent and insightful thematic groupings.

XSL•FO
RenderX

Figure 4. Descriptions of themes 1-7 along with the corresponding topic descriptions assigned to each theme. CDC: Centers for Disease Control and Prevention.



XSL•FO
RenderX

Figure 5. Descriptions of themes 8-13 along with the corresponding topic descriptions assigned to each theme.

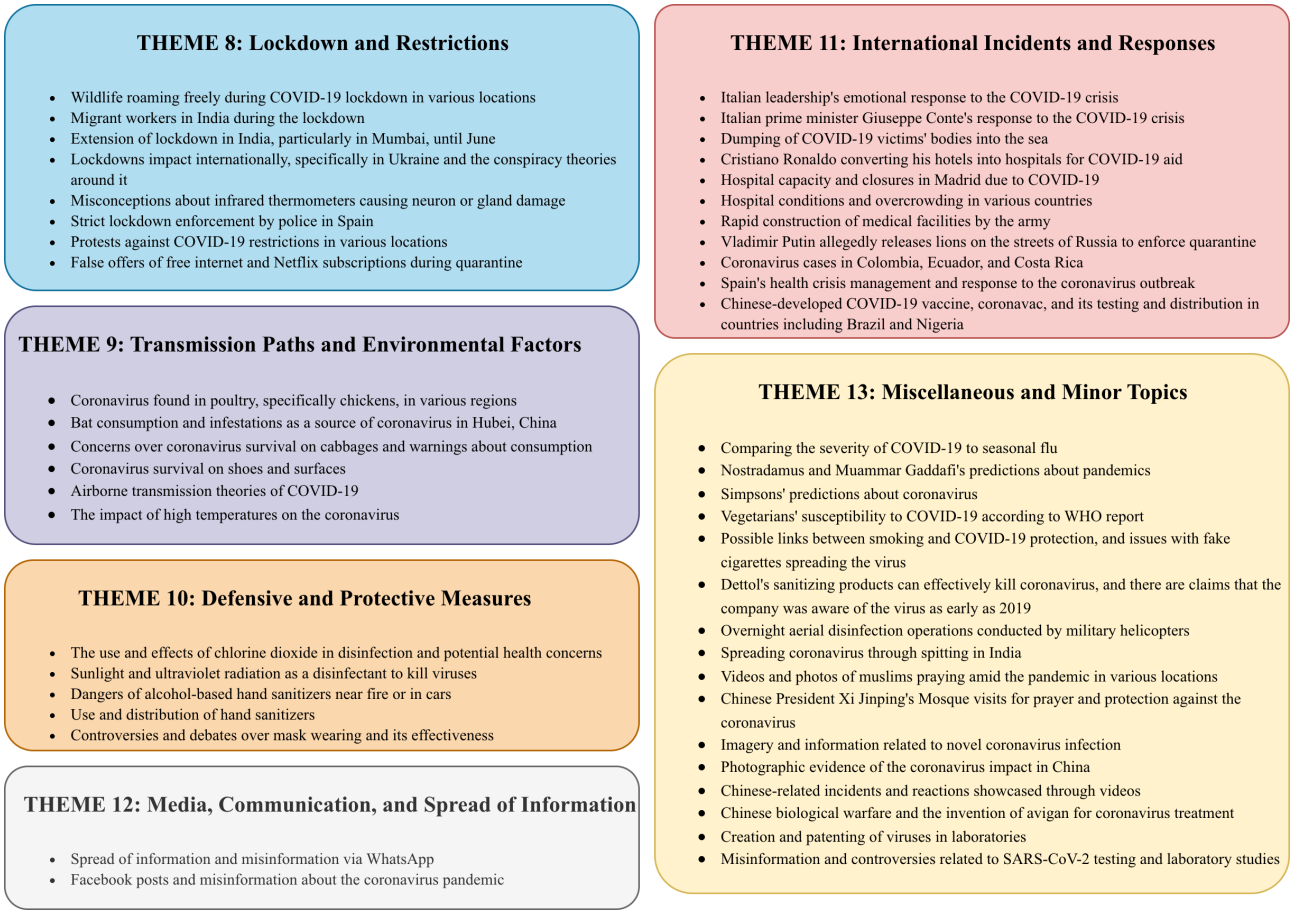

Using the algorithm in Textbox 1, we assign each document to a topic, enabling us to determine the distribution of each theme. Table 5 shows the distribution of all themes, with theme 4 (Vaccines) and theme 3 (Conspiracy Theories) emerging as the first and second most prevalent misinformation themes. Here are all the themes and their percentages:

Table . Distribution of COVID-19 misinformation themes.

Theme	Values (N=18,018), n (%)
Theme 1: Home remedies	1334 (7.40)
Theme 2: Deaths and statistics	570 (3.16)
Theme 3: Conspiracy theories	3459 (19.20)
Theme 4: Vaccine	3595 (19.95)
Theme 5: Testing	546 (3.03)
Theme 6: Ivermectin	814 (4.52)
Theme 7: Government	2094 (11.62)
Theme 8: Lockdowns	1316 (7.30)
Theme 9: Transmission	297 (1.65)
Theme 10: Defensive	1269 (7.04)
Theme 11: International	891 (4.95)
Theme 12: Media	252 (1.40)
Theme 13: Minor topics	1581 (8.77)

Figure 6 shows the distribution of each topic within theme 8 over time, illustrating that protests against COVID-19 restrictions were the only misinformation topic that actively persisted even after the release of the COVID-19 vaccine.

After identifying the misinformation themes, we leverage the explanations provided in the IFCN dataset as a basis to draw refutation arguments to address these themes. For each identified theme, we develop a refutation that aligns with its context,

XSL•FO
RenderX

aiming to clarify the nature of the misinformation, its potential origins, and its impact. These refutations are shown in Textbox 4. The refutation arguments list for 13 themes, providing valuable insights into the underlying reasons for the misinformation.

Figure 6. Distribution of theme 8 topics during the time. FDA: Food and Drug Administration; WHO: World Health Organization.

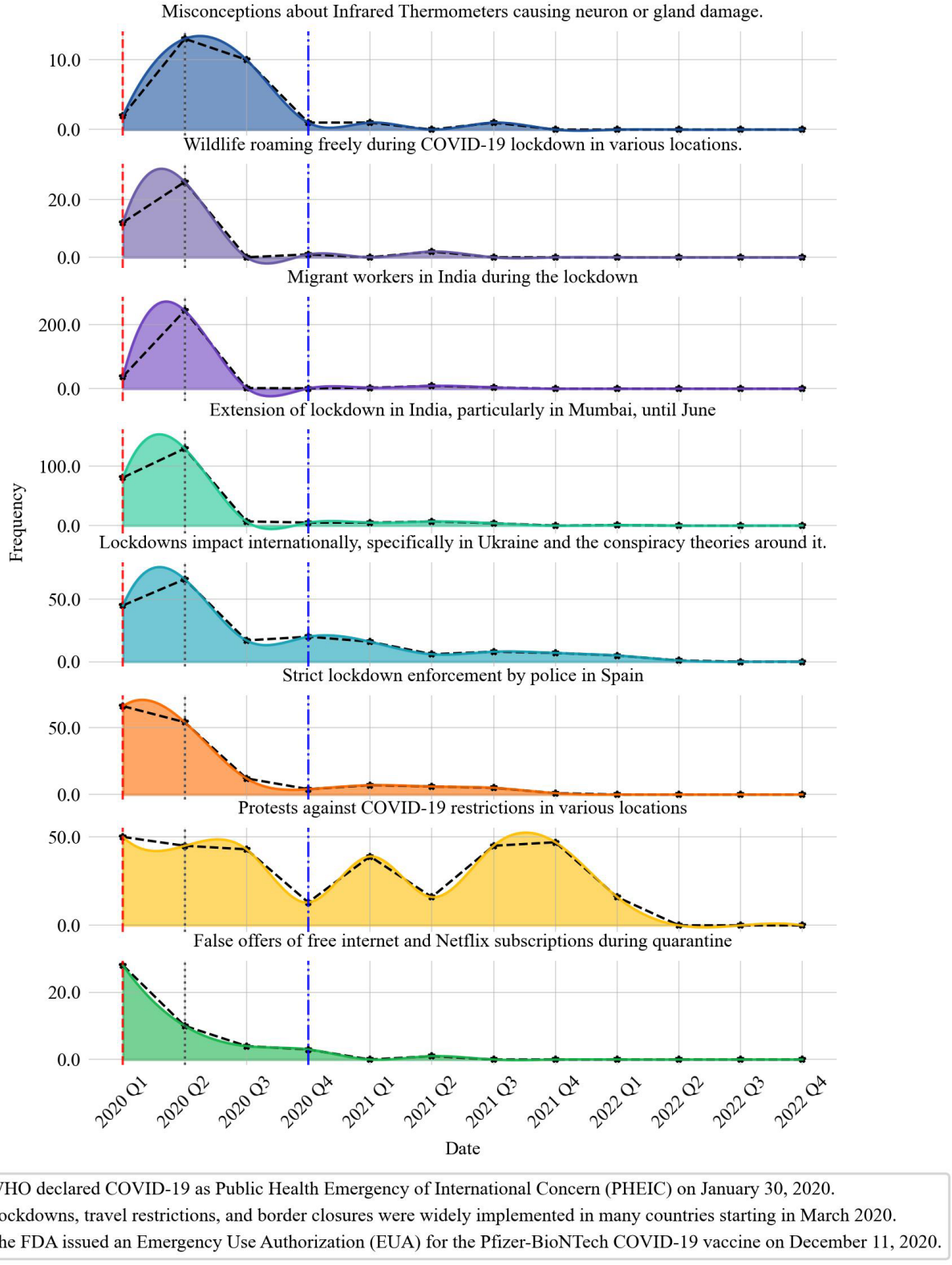

XSL•FO
RenderX

Textbox 4. The refutation arguments list for 13 themes.

Refutation list:

1. This content may contain misinformation related to home remedies for COVID-19 prevention and treatment.
2. This content may contain misinformation related to COVID-19 deaths, statistics, and their relation to vaccination.
3. This content may contain conspiracy theories and misinformation related to COVID-19, including unverified claims, distorted facts, or manipulated content.
4. This content may contain misinformation related to COVID-19 vaccines, including false claims about safety, efficacy, and side effects.
5. This content may contain misinformation related to COVID-19 testing accuracy, including false claims about polymerase chain reaction tests and unscientific self-check methods.
6. This content may contain misinformation related to COVID-19 treatments, including exaggerated claims about ivermectin, hydroxychloroquine, or unproven supplements.
7. This content may contain misinformation related to government and political responses to COVID-19 pandemic, including distorted facts, fabricated claims, or misrepresentation of policies and actions.
8. This content may contain misinformation related to COVID-19 lockdowns and restrictions, including fabricated or misrepresented events, videos, or claims.
9. This content may contain misinformation related to COVID-19 transmission and survival in various environments, including unverified claims about foods, surfaces, or environmental factors.
10. This content may contain misinformation related to defensive and protective measures against COVID-19 pandemic, including false or exaggerated claims about masks, sanitizers, UV rays, or disinfectants.
11. This content may contain misinformation related to international incidents and responses to COVID-19 pandemic, including fabricated reports of government actions, health care capacity, or global cooperation.
12. This content may contain misinformation related to the spread of COVID-19 information on social media and messaging apps, including false claims about government policies, platforms, or media manipulation.
13. This content may contain miscellaneous misinformation related to COVID-19 pandemic, including distortions about products, religious practices, bioweapons, and other fabricated claims.

Provide Refutation

In the final stage of our process, we design a prompt text to enable ChatGPT-4.0 to detect specific misinformation themes. The prompt text includes a detailed description of the themes and a question-answer list. To create this question-answer list, we select the document closest to the center of each topic and associate it with the corresponding theme. Detailed information about the prompt text can be found in Textbox 5.

XSL•FO
RenderX

Textbox 5. Prompt structure and 1 example to find a document theme.

Finding document themes prompt structure:

System role:

"The following is the description of topic themes related to COVID-19 misinformation. Find the closest theme for the given text. Answer in a consistent style."

User role:

Themes description list

Assistant role:

Question-answer list

User role:

Input text

Finding topic themes prompt example:

System role:

"The following is the description of topic themes related to COVID-19 misinformation. Find the closest theme for the given text. Answer in a consistent style."

User role:

Themes description list

Assistant role:

Question-answer list

User role:

"A video shows that Bill Gates admits the vaccine will no doubt kill 700,000 people."

Output answer:

Theme 3: "Conspiracy Theories and Misinformation"

To evaluate our approach, we randomly selected 130 (10 documents per theme) documents from the IFCN dataset that are not included in the question-answer list. We then tested the prompting method with ChatGPT-4.0, achieving an 82% accuracy rate in detecting the correct themes. Moreover, as showing in Figure 7 the model struggled to detect themes 2, 7, and 13 in comparisons with other themes.

XSL•FO
RenderX

Figure 7. Confusion matrix of theme detector.

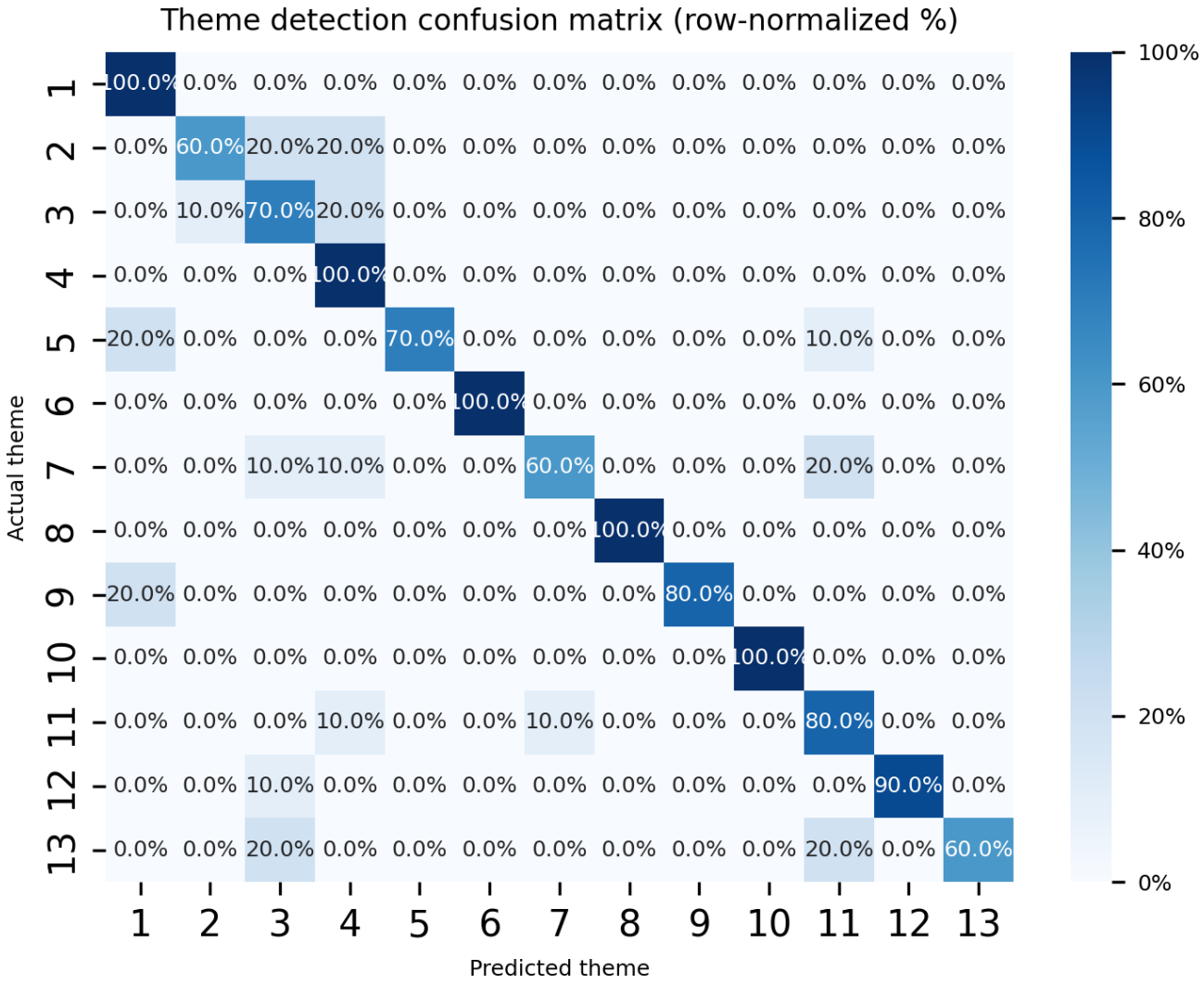

Proposed System

Based on the process we introduced, we propose the development of an MDIS. This system is designed to first determine whether a given text document contains misinformation or not. If the document is identified as containing misinformation, the system then detects its theme and provides a detailed refutation of the misinformation. The primary objectives of MDIS are to prevent the spread of misinformation and to enhance public health knowledge.

MDIS operates by integrating 3 key components. First, it uses a trained LLM to classify text documents as either misinformation or nonmisinformation. Next, it uses another trained LLM to detect the specific misinformation theme within documents identified as containing misinformation. Finally, it leverages a refutation list, which is generated during the theme description phase, to provide context and counternarratives for each detected theme. This comprehensive approach enables the system to effectively address misinformation while equipping users with accurate information.

Figure 1C illustrates an example of an input to MDIS and its corresponding output, demonstrating how the system analyzes a document, identifies misinformation, detects the associated theme, and presents an explanatory response.

Discussion

Principal Results

This study developed the MDIP, a structured workflow for analyzing health misinformation and generating explanatory counterarguments. Building on MDIP, we designed a prototype MDIS that combines detection, topic detection, theme identification, and refutation generation. While the system has not yet been deployed in real-world environments, the results illustrate its potential to support public health communication.

In the misinformation classification task, LLMs achieved reasonable performance, with BERT reaching 98% accuracy. Enriching training datasets also reduced false positives on AI-generated misinformation by 44% compared with prior baselines, suggesting improved robustness across different linguistic styles. Topic modeling experiments highlighted the advantages of BERTopic relative to LDA and Top2Vec, with higher coherence and diversity metrics. To address the issue of unassigned documents, the algorithm in Textbox 1 was used to assign outliers to their nearest topic cluster, thereby improving coverage of the dataset. To enhance interpretability, word-level topic outputs were converted into sentence-level descriptions through prompt engineering with ChatGPT-4.0. These descriptions were judged appropriate or somewhat appropriate

XSL•FO
RenderX

in 99.6% of cases by independent raters, indicating the value of sentence-level representations for clarity and interpretability. Building on these descriptions, topics were further grouped into broader themes and linked with theme-specific refutations.

Finally, a prompt-based detector for misinformation themes achieved 82% accuracy. These results demonstrate how detection, topic modeling, thematic grouping, and refutation can be integrated into a single workflow. The prototype system (MDIS) represents an illustration of concept, and while the findings are encouraging, validation is limited to English language, COVID-19–related data, and offline testing. Broader generalization, multilingual adaptation, integration into health communication workflows, and longitudinal evaluation remain important directions for future work.

Limitations

While this study presents a framework for detecting and addressing misinformation, several limitations should be acknowledged. First, the system's performance relies on the quality and diversity of the datasets that were used for training and topic modeling, which may not fully capture the linguistic and contextual nuances of misinformation in different regions, languages, or cultural contexts. This limitation introduces a potential bias: refutations that appear clear and persuasive in one sociocultural context may be ineffective—or even counterproductive—in another.

The model has been tested solely on English text–based misinformation, and it has not been evaluated for multilingual and multimodal adaptation, meaning its ability to detect misinformation and provide persuasive refutation across different languages and sociocultural contexts remains uncertain. Moreover, while the theme detection module achieved an accuracy of 82%, this leaves an 18% error margin, especially in ambiguous or overlapping themes, which can lead to inaccuracies in refutation and reduce the overall effectiveness of the generated warning texts. Developing and integrating more sophisticated algorithms to address overlapping in topic themes could substantially enhance the accuracy of theme detection. Improved theme separation would not only yield clearer and more coherent thematic structures but also strengthen the generation of precise and contextually relevant refutations. In turn, this refinement would enable more effective countermeasures against health misinformation, thereby improving the system's overall capacity to support public health communication and trust. False positives and negatives remain a concern, particularly when misinformation contains opinion-based, satirical, or context-dependent elements. Although the generated refutations follow a systematic structure, they may not always be contextually relevant, persuasive, or ethically suitable for diverse audiences. In the absence of a human-in-the-loop or oversight mechanism, the system may produce counterarguments that fail to resonate with users or could be perceived as unreliable. In addition, the system has not yet been extensively tested in real-world applications, which limits understanding of its practical impact on misinformation spread and public health outcomes. Furthermore, misinformation evolves over time, and a model trained on past narratives may require periodic retraining to remain effective against emerging falsehoods, including AI-generated misinformation. Using pretrained LLMs such as ChatGPT depends on the current version of the model and its accessibility to users. Therefore, it is necessary to update the system regularly when the model is changed or becomes unavailable. Moreover, since passing datasets through third-party platforms may compromise the security of the framework, future work could focus on developing an in-house solution by training a dedicated model for our specific tasks, thereby eliminating the reliance on external platforms. Finally, the reliance on automated methods raises potential concerns about interpretability and transparency, which are crucial for fostering trust and adoption by end users.

Comparison With Prior Work

The proposed MDIP and the resulting MDIS build upon and advance the body of research focused on misinformation detection and mitigation. The proposed method transforms raw posts into *actionable units*—sentence-level topic labels, aggregated themes, and paired refutations—linking detection outputs directly to message design and response playbooks used by health teams. Previous research has demonstrated the efficacy of ML models, particularly deep learning approaches, in detecting misinformation. They used ML techniques to classify fake news using textual features, demonstrating the value of automated detection methods [49,52,69]. Our study extends these efforts by integrating enriched datasets containing both formal and informal language styles, ensuring better generalization across diverse linguistic sources, including AI-generated misinformation.

Topic modeling techniques such as LDA have been used in prior studies to analyze misinformation [35,53,55]. Our approach improves on these works by addressing limitations in document assignment and theme interpretation. We used an algorithm to assign every document to the most relevant topic, resolving the common issue of unclassified documents in topic modeling. In addition, we moved beyond word-level topic representations to generate sentence-level descriptions, offering richer and more interpretable insights. By tracking shifts in sentence-level topics and theme distributions, communicators can conduct pre-/postassessments of campaigns or platform policy changes, complementing survey-based outcomes. Finally, we designed an effective prompt text to automatically identify the themes of misinformation. This automated approach reduces reliance on manual interpretation, minimizing human bias and increasing scalability.

Many prior studies have addressed misinformation detection or topic analysis in isolation. They analyzed misinformation using sentiment analysis but did not integrate detection with thematic analysis and did not provide a framework for counteracting misinformation [11,35,70]. Our work unifies detection, topic modeling, thematic refutation, and public health intervention in a single framework. The MDIS framework automates the end-to-end process, offering a scalable solution to tackle the complexity of misinformation dynamics.

Conclusions

This work contributes a methodological framework for infodemiology and digital health operations. We transform

XSL•FO
RenderX

misinformation into actionable units—themes and refutations—so that health teams can act (communicate, triage, and evaluate). Moreover, analyzing misinformation using a hierarchical (2-level) sentence-level description and assigning all documents to topics makes it possible to observe theme and topic distributions over time, providing a broad and sensible overview of misinformation. Sentence-level topics and theme distributions serve as measurable indicators for surveillance and intervention evaluation (eg, pre-/postcampaign shifts and surge detection). We introduce MDIP and MDIS that enable rapid response playbooks and reduce analyst workload. To support adoption, we release prompt templates and code as implementation artifacts that teams can readily adapt. Real-world deployment, however, requires governance mechanisms (human-in-the-loop review and audit logs), multilingual extensions, and prospective trials with health agencies or platforms to quantify downstream impact (eg, reduced spread and improved literacy). Ultimately, these contributions orient detection toward operational use—prioritizing interpretability and intervention design—so that public health actors can move from finding misinformation to effectively countering it.

Funding

This research is supported in part by a research grant from the Investigator-Initiated Studies Program of Merck Sharp & Dohme Corp (MISP #102050). The opinions expressed in this paper are those of the authors and do not necessarily represent those of Merck Sharp & Dohme Corp.

Data Availability

All implementation codes can be accessed through the GitHub repository [71].

Conflicts of Interest

None declared.

References

1. Kisa S, Kisa A. A comprehensive analysis of COVID-19 misinformation, public health impacts, and communication strategies: scoping review. J Med Internet Res 2024 Aug 21;26:e56931. [doi: 10.2196/56931] [Medline: 39167790]
2. Loomba S, de Figueiredo A, Piatek SJ, de Graaf K, Larson HJ. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nat Hum Behav 2021 Mar;5(3):337-348. [doi: 10.1038/s41562-021-01056-1] [Medline: 33547453]
3. Moscadelli A, Albora G, Biamonte MA, et al. Fake news and Covid-19 in Italy: results of a quantitative observational study. Int J Environ Res Public Health 2020 Aug 12;17(16):5850. [doi: 10.3390/ijerph17165850] [Medline: 32806772]
4. Chou WYS, Oh A, Klein WMP. Addressing health-related misinformation on social media. JAMA 2018 Dec 18;320(23):2417-2418. [doi: 10.1001/jama.2018.16865] [Medline: 30428002]
5. Kata A. Anti-vaccine activists, Web 2.0, and the postmodern paradigm – an overview of tactics and tropes used online by the anti-vaccination movement. Vaccine (Auckl) 2012 May;30(25):3778-3789. [doi: 10.1016/j.vaccine.2011.11.112]
6. Zimet GD, Rosberger Z, Fisher WA, Perez S, Stupiansky NW. Beliefs, behaviors and HPV vaccine: correcting the myths and the misinformation. Prev Med 2013 Nov;57(5):414-418. [doi: 10.1016/j.ypmed.2013.05.013]
7. Poland GA, Jacobson RM. Understanding those who do not understand: a brief review of the anti-vaccine movement. Vaccine (Auckl) 2001 Mar;19(17-19):2440-2445. [doi: 10.1016/S0264-410X(00)00469-2]
8. Kata A. A postmodern Pandora's box: anti-vaccination misinformation on the internet. Vaccine (Auckl) 2010 Feb 17;28(7):1709-1716. [doi: 10.1016/j.vaccine.2009.12.022] [Medline: 20045099]
9. Oyeyemi SO, Gabarron E, Wynn R. Ebola, Twitter, and misinformation: a dangerous combination? BMJ 2014 Oct 14;349:g6178. [doi: 10.1136/bmj.g6178] [Medline: 25315514]
10. Geoghegan S, O'Callaghan KP, Offit PA. Vaccine safety: myths and misinformation. Front Microbiol 2020;11:372. [doi: 10.3389/fmicb.2020.00372] [Medline: 32256465]
11. Zhou X, Coiera E, Tsafnat G, Arachi D, Ong MS, Dunn AG. Using social connection information to improve opinion mining: identifying negative sentiment about HPV vaccines on twitter. In: Studies in Health Technology and Informatics 2015:761-765. [doi: 10.3233/978-1-61499-564-7-761]
12. Ghaddar A, Khandaqji S, Awad Z, Kansoun R. Conspiracy beliefs and vaccination intent for COVID-19 in an infodemic. PLoS One 2022;17(1):e0261559. [doi: 10.1371/journal.pone.0261559] [Medline: 35020721]
13. Ghosh D, Scott B. Disinformation is becoming unstoppable. TIME. 2018. URL: https://time.com/5112847/facebook-fake-news-unstoppable/ [accessed 2025-11-28]
14. Qazvinian V, Rosengren E, Radev D, Mei Q. Rumor has it: identifying misinformation in microblogs. 2011 Presented at: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing; Jul 27-31, 2011; Edinburgh, Scotland, UK p. 1589-1599 URL: https://aclanthology.org/D11-1147.pdf [accessed 2025-12-19]
15. Wang Y, McKee M, Torbica A, Stuckler D. Systematic literature review on the spread of health-related misinformation on social media. Soc Sci Med 2019 Nov;240:112552. [doi: 10.1016/j.socscimed.2019.112552] [Medline: 31561111]

XSL•FO
RenderX

16. Broniatowski DA, Jamison AM, Qi S, et al. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. Am J Public Health 2018 Oct;108(10):1378-1384. [doi: 10.2105/AJPH.2018.304567] [Medline: 30138075]
17. Zarocostas J. How to fight an infodemic. Lancet 2020 Feb 29;395(10225):676. [doi: 10.1016/S0140-6736(20)30461-X] [Medline: 32113495]
18. Islam MS, Kamal AHM, Kabir A, et al. COVID-19 vaccine rumors and conspiracy theories: the need for cognitive inoculation against misinformation to improve vaccine adherence. PLoS One 2021;16(5):e0251605. [doi: 10.1371/journal.pone.0251605]
19. Gallotti R, Valle F, Castaldo N, Sacco P, De Domenico M. Assessing the risks of “infodemics” in response to COVID-19 epidemics. Nat Hum Behav 2020 Dec;4(12):1285-1293. [doi: 10.1038/s41562-020-00994-6] [Medline: 33122812]
20. Cinelli M, Quattrociocchi W, Galeazzi A, et al. The COVID-19 social media infodemic. Sci Rep 2020;10(1). [doi: 10.1038/s41598-020-73510-5]
21. Mian A, Khan S. Coronavirus: the spread of misinformation. BMC Med 2020 Mar 18;18(1):89. [doi: 10.1186/s12916-020-01556-3] [Medline: 32188445]
22. Kumar N, Corpus I, Hans M, et al. COVID-19 vaccine perceptions in the initial phases of US vaccine roll-out: an observational study on reddit. BMC Public Health 2022 Mar 7;22(1):446. [doi: 10.1186/s12889-022-12824-7] [Medline: 35255881]
23. Kim JW, Lee J, Dai Y. Misinformation and the paradox of trust during the covid-19 pandemic in the U.S.: pathways to risk perception and compliance behaviors. J Risk Res 2023 May 4;26(5):469-484. [doi: 10.1080/13669877.2023.2176910]
24. Hou Z, Du F, Zhou X, et al. Cross-country comparison of public awareness, rumors, and behavioral responses to the COVID-19 epidemic: infodemiology study. J Med Internet Res 2020;22(8):e21143. [doi: 10.2196/21143]
25. Bavel JJV, Baicker K, Boggio PS, et al. Using social and behavioural science to support COVID-19 pandemic response. Nat Hum Behav 2020;4(5):460-471. [doi: 10.1038/s41562-020-0884-z]
26. Schiffman MH, Bauer HM, Hoover RN, et al. Epidemiologic evidence showing that human papillomavirus infection causes most cervical intraepithelial neoplasia. J Natl Cancer Inst 1993 Jun 16;85(12):958-964. [doi: 10.1093/jnci/85.12.958] [Medline: 8388478]
27. Bosch FX, Manos MM, Munoz N, et al. Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. JNCI Journal of the National Cancer Institute 1995 Jun 7;87(11):796-802. [doi: 10.1093/jnci/87.11.796] [Medline: 7791229]
28. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019 Jan;69(1):7-34. [doi: 10.3322/caac.21551] [Medline: 30620402]
29. Reasons to get vaccinated. Centers for Disease Control and Prevention. 2021. URL: https://www.cdc.gov/hpv [accessed 2025-11-28]
30. Pingali C, Yankey D, Elam-Evans LD, et al. National, regional, state, and selected local area vaccination coverage among adolescents aged 13-17 Years—United States, 2020. MMWR Morb Mortal Wkly Rep 2021 Sep 3;70(35):1183-1190. [doi: 10.15585/mmwr.mm7035a1] [Medline: 34473682]
31. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85. [doi: 10.2196/jmir.1933] [Medline: 23615206]
32. Levin S. Facebook promised to tackle fake news but the evidence shows it’s not working. The Guardian. 2017. URL: https://www.theguardian.com/technology/2017/may/16/facebook-fake-news-tools-not-working [accessed 2025-12-12]
33. Zhou J, Zhang Y, Luo Q, Parker AG, De Choudhury M. Synthetic lies: understanding ai-generated misinformation and evaluating algorithmic and human solutions. CHI ’23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems 2023:1-20. [doi: 10.1145/3544548.3581318]
34. Jiang B, Tan Z, Nirmal A, Liu H. Disinformation detection: an evolving challenge in the age of llms. In: Proceedings of the 2024 SIAM International Conference on Data Mining (SDM): Society for Industrial and Applied Mathematics Publications; 2024:427-435. [doi: 10.1137/1.9781611978032.50]
35. Du J, Preston S, Sun H, et al. Using machine learning-based approaches for the detection and classification of human papillomavirus vaccine misinformation: infodemiology study of Reddit discussions. J Med Internet Res 2021 Aug 5;23(8):e26478. [doi: 10.2196/26478] [Medline: 34383667]
36. Tomaszewski T, Morales A, Lourentzou I, et al. Identifying false human papillomavirus (HPV) vaccine information and corresponding risk perceptions from twitter: advanced predictive models. J Med Internet Res 2021 Sep 9;23(9):e30451. [doi: 10.2196/30451] [Medline: 34499043]
37. Farajijalal M, Malek S, Toudeshki A, Viers JH, Ehsani R. Data-driven model to improve mechanical harvesters for nut trees. 2024 Presented at: 2024 ASABE Annual International Meeting; Jul 28-31, 2024; California p. 1. [doi: 10.13031/aim.202400858]
38. Malek S, Salehkaleybar S, Amini A. Multi variable-layer neural networks for decoding linear codes. 2020 Presented at: 2020 8th Iran Workshop on Communication and Information Theory (IWCIT); May 26-28, 2020; Tehran, Iran p. 1-6. [doi: 10.1109/IWCIT50667.2020.9163473]
39. Chui M, Manyika J, Miremadi M, Henke N, Chung R, Nel P, et al. Notes from the AI frontier: insights from hundreds of use cases. : McKinsey Global Institute; 2018.

XSL•FO
RenderX

40. Acemoglu D, Restrepo P. Artificial intelligence, automation, and work. In: Agrawal A, Gans J, Goldfarb A, editors. The Economics of Artificial Intelligence: An Agenda: University of Chicago Press; 2018:197-236.
41. Micallef N, He B, Kumar S, Ahamad M, Memon N. The role of the crowd in countering misinformation: a case study of the COVID-19 infodemic. 2020 Presented at: 2020 IEEE International Conference on Big Data (Big Data); Dec 10-13, 2020; Atlanta, GA, USA p. 748-757. [doi: 10.1109/BigData50022.2020.9377956]
42. He B, Ahamad M, Kumar S. Reinforcement learning-based counter-misinformation response generation: a case study of COVID-19 vaccine misinformation. WWW '23: Proceedings of the ACM Web Conference 2023 :2698-2709. [doi: 10.1145/3543507.3583388]
43. Jia R, Liang P. Adversarial examples for evaluating reading comprehension systems. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: Association for Computational Linguistics; 2017:2021-2031. [doi: 10.18653/v1/D17-1215]
44. Ribeiro MT, Singh S, Guestrin C. Semantically equivalent adversarial rules for debugging NLP models. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1): Association for Computational Linguistics; 2018. [doi: 10.18653/v1/P18-1079]
45. Pham C, Hoyle A, Sun S, Resnik P, Iyyer M. TopicGPT: a prompt-based topic modeling framework. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Association for Computational Linguistics; 2024. [doi: 10.18653/v1/2024.naacl-long.164]
46. Mu Y, Dong C, Bontcheva K, Song X. Large language models offer an alternative to the traditional approach of topic modelling. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): ELRA and ICCL; 2024:10160-10171 URL: https://aclanthology.org/2024.lrec-main.887/ [accessed 2025-12-22]
47. Doi T, Isonuma M, Yanaka H. Topic modeling for short texts with large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4): Association for Computational Linguistics; 2024:21-33. [doi: 10.18653/v1/2024.acl-srw.3]
48. Banas JA, Rains SA. A meta-analysis of research on inoculation theory. Commun Monogr 2010 Sep;77(3):281-311. [doi: 10.1080/03637751003758193]
49. Patwa P, Sharma S, Pykl S, Guptha V, Kumari G, Akhtar M. Fighting an infodemic: COVID-19 fake news dataset. : Springer; 2021 Presented at: Combating Online Hostile Posts in Regional Languages during Emergency Situation. [doi: 10.1007/978-3-030-73696-5_3]
50. Saenz J, Gopal SRK, Shukla D. COVID-19 fake news infodemic research dataset (COVID19-FNIR dataset). IEEE Dataport. URL: https://dx.doi.org/10.21227/b5bt-5244 [accessed 2025-12-22]
51. Rai S, Kornides M, Morgan J, Kumar A, Cappella J, Guntuku SC. Detecting and monitoring concerns against HPV vaccination on social media using large language models. Sci Rep 2024 Jun 21;14(1):14362. [doi: 10.1038/s41598-024-64703-3] [Medline: 38906941]
52. Glazkova A, Glazkov M, Trifonov T. g2tmn at constraint@ AAAI2021: exploiting CT-BERT and ensembling learning for COVID-19 fake news detection. In: International Workshop on Combating On Line Hostile Posts in Regional Languages during Emergency Situation: Springer; 2021:116-127. [doi: 10.1007/978-3-030-73696-5_12]
53. Song X, Petrak J, Jiang Y, Singh I, Maynard D, Bontcheva K. Classification aware neural topic model for COVID-19 disinformation categorisation. PLoS One 2021;16(2):e0247086. [doi: 10.1371/journal.pone.0247086] [Medline: 33600477]
54. Ball P, Maxmen A. The epic battle against coronavirus misinformation and conspiracy theories. Nature New Biol 2020 May 28;581(7809):371-374. [doi: 10.1038/d41586-020-01452-z]
55. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. J Mach Learn Res 2003;3(Jan):993-1022 [FREE Full text]
56. Karas B, Qu S, Xu Y, Zhu Q. Experiments with LDA and Top2Vec for embedded topic discovery on social media data-A case study of cystic fibrosis. Front Artif Intell 2022;5:948313. [doi: 10.3389/frai.2022.948313] [Medline: 36062265]
57. Egger R, Yu J. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. Front Sociol 2022;7:886498. [doi: 10.3389/fsoc.2022.886498] [Medline: 35602001]
58. Abdelrazek A, Eid Y, Gawish E, Medhat W, Hassan A. Topic modeling algorithms and applications: a survey. Inf Syst 2023 Feb;112:102131. [doi: 10.1016/j.is.2022.102131]
59. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. Presented at: WSDM '15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining; Feb 2-6, 2015; Shanghai China p. 399-408. [doi: 10.1145/2684822.2685324]
60. Bouma G. Normalized (pointwise) mutual information in collocation extraction. In: From Form to Meaning: Processing Texts Automatically (Proceedings of the Biennial GSCL Conference 2009): Gunter Narr Verlang; 2009:31-40 URL: https://books.google.com/
[accessed 2025-12-22]
61. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. arXiv. Preprint posted online on Jul 22, 2020. [doi: 10.48550/arXiv.2005.14165]

XSL•FO
RenderX

62. International Fact-Checking Network (IFCN). COVID-19 Fact-Checking Database. Poynter Institute. URL: https://www.poynter.org/ifcn-covid-19-misinformation/ [accessed 2025-12-22]
63. 45 CFR part 46 – protection of human subjects (common rule). US Department of Health & Human Services. URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46 [accessed 2025-12-22]
64. Kenton J, Toutanova LK. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: Proceedings of NAACL-HLT; Jun 2-7, 2019; Minneapolis, MN p. 2 URL: https://au1206.github.io/assets/pdfs/BERT.pdf [accessed 2025-12-02]
65. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. : OpenAI blog; 2019 URL: https://storage.prod.researchhub.com/uploads/papers/2020/06/01/language-models.pdf [accessed 2025-12-19]
66. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 2020;21(140):1-67 [FREE Full text]
67. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. Sci Rep 2024 Mar 13;14(1):6086. [doi: 10.1038/s41598-024-56706-x] [Medline: 38480847]
68. Outlier reduction. BERTopic. 2025 Sep 18. URL: https://maartengr.github.io/BERTopic/getting_started/outlier_reduction/outlier_reduction.html [accessed 2025-11-28]
69. Ding X, Teng C, Ji D. Fake news detection with context awareness of the publisher. 2023 Jul 1 Presented at: The 35th International Conference on Software Engineering and Knowledge Engineering; Jul 1-10, 2023. [doi: 10.18293/SEKE2023-061]
70. Piedrahita-Valdés H, Piedrahita-Castillo D, Bermejo-Higuera J, et al. Vaccine hesitancy on social media: sentiment analysis from June 2011 to April 2019. Vaccines (Basel) 2011 Apr;9(1):28. [doi: 10.3390/vaccines9010028]
71. MDIP: misinformation detection and inoculation processing. GitHub, Inc. URL: https://github.com/SamiraMalek/MDIP-MDIS [accessed 2025-11-28]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
CDC: Centers for Disease Control and Prevention
CV: Coherence Value
GPT-2 : Generative Pre-trained Transformer 2
HPV: human papillomavirus
IRBO: Inverted Rank-Biased Overlap
LDA: Latent Dirichlet Allocation
LLM: large language model
MDIP: Misinformation Detection and Inoculation Process
MDIS: Misinformation Detection and Inoculation System
ML: machine learning
NPMI: Normalized Pointwise Mutual Information
T5-base: Text-to-Text Transfer Transformer
WHO: World Health Organization

Edited by J Sarvestan; submitted 09.Apr.2025; peer-reviewed by D Chumachenko, EC Choi; revised version received 12.Oct.2025; accepted 13.Oct.2025; published 08.Jan.2026.

Please cite as:
Malek S, Griffin C, Fraleigh RD, Lennon R, Monga V, Shen L
Intervention in Health Misinformation Using Large Language Models for Automated Detection, Thematic Analysis, and Inoculation: Case Study on COVID-19
J Med Internet Res 2026;28:e75500
URL: https://www.jmir.org/2026/1/e75500
doi:10.2196/75500

© Samira Malek, Christopher Griffin, Robert D Fraleigh, Robert Lennon, Vishal Monga, Lijiang Shen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet

XSL•FO
RenderX

Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Developing an AI-Assisted Tool That Identifies Patients With Multimorbidity and Complex Polypharmacy to Improve the Process of Medication Reviews: Qualitative Interview and Focus Group Study

Aseel S Abuzour[1,2*], PhD; Samantha A Wilson[3*], PhD; Alan A Woodall[3,4], PhD; Frances S Mair[5*], Prof Dr; Asra Aslam[2,6*], PhD; Andrew Clegg[1,2], PhD; Eduard Shantsila[3], PhD; Mark Gabbay[3*], Prof Dr; Michael Abaho[3*], PhD; Danushka Bollegala[7], Prof Dr; Harriet Cant[8], MSc; Alan Griffiths[9]; Layik Hama[6,10], PhD; Gary Leeming[3], PhD; Emma Lo[3]; Simon Maskell[11], Prof Dr; Maurice O'Connell[8], PhD; Olusegun Popoola[12], MD; Sam Relton[2,6], PhD; Roy A Ruddle[6,10], Prof Dr; Pieta Schofield[3], PhD; Matthew Sperrin[8], PhD; Tjeerd Van Staa[8], PhD; Iain Buchan[3], Prof Dr; Lauren E Walker[13,14], Prof Dr

[1]Academic Unit for Ageing & Stroke Research, University of Leeds, Leeds, United Kingdom
[2]School of Medicine, Faculty of Medicine and Health, University of Leeds, Leeds, United Kingdom
[3]Institute of Population Health, University of Liverpool, Liverpool, United Kingdom
[4]Directorate of Mental Health and Learning Disabilities, Powys Teaching Health Board, Bronllys, United Kingdom
[5]School of Health and Wellbeing, General Practice and Primary Care, University of Glasgow, Glasgow, United Kingdom
[6]Leeds Institute for Data Analytics, University of Leeds, Leeds, United Kingdom
[7]Department of Computer Science, University of Liverpool, Liverpool, United Kingdom
[8]Division of Informatics, Imaging & Data Science, University of Manchester, Manchester, United Kingdom
[9]NIHR Applied Research Collaboration North West Coast, Liverpool, United Kingdom
[10]School of Computer Science, University of Leeds, Leeds, United Kingdom
[11]Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, United Kingdom
[12]Mersey Care NHS Foundation Trust, Liverpool, United Kingdom
[13]Centre for Experimental Therapeutics, University of Liverpool, Liverpool, United Kingdom
[14]Liverpool University Hospitals NHS Foundation Trust, Liverpool, United Kingdom
[*]these authors contributed equally

Corresponding Author:
Lauren E Walker, Prof Dr
Centre for Experimental Therapeutics
University of Liverpool
Waterhouse Building, Block B Brownlow Street
Liverpool, L69 3GF
United Kingdom
Phone: 44 151 795 5407
Email: md0u10pc@liverpool.ac.uk

Abstract

Background: Structured medication reviews (SMRs) are an essential component of medication optimization, especially for patients with multimorbidity and polypharmacy. However, the process remains challenging due to the complexities of patient data, time constraints, and the need for coordination among health care professionals (HCPs). This study explores HCPs' perspectives on the integration of artificial intelligence (AI)–assisted tools to enhance the SMR process, with a focus on the potential benefits of and barriers to adoption.

Objective: This study aims to identify the key user requirements for AI-assisted tools to improve the efficiency and effectiveness of SMRs, specifically for patients with multimorbidity, complex polypharmacy, and frailty.

XSL•FO
RenderX

Methods: A qualitative study was conducted involving focus groups and semistructured interviews with HCPs and patients in the United Kingdom. Participants included physicians, pharmacists, clinical pharmacologists, psychiatrists from primary and secondary care, a policy maker, and patients with multimorbidity. Data were analyzed using a hybrid inductive and deductive thematic analysis approach to identify themes related to AI-assisted tool functionality, workflow integration, user-interface visualization, and usability in the SMR process.

Results: Four major themes emerged from the analysis: innovative AI potential, optimizing electronic patient record visualization, functionality of the AI tool for SMRs, and facilitators of and barriers to AI tool implementation. HCPs identified the potential of AI to support patient identification and prioritizing those at risk of medication-related harm. AI-assisted tools were viewed as essential in detecting prescribing gaps, drug interactions, and patient risk trajectories over time. Participants emphasized the importance of presenting patient data in an intuitive format, with a patient interface for shared decision-making. Suggestions included color-coding blood results, highlighting critical medication reviews, and providing timelines of patient medical histories. HCPs stressed the need for AI tools to integrate seamlessly with existing electronic patient record systems and provide actionable insights without overwhelming users with excessive notifications or "pop-up" alerts. Factors influencing the uptake of AI-assisted tools included the need for user-friendly design, evidence of tool effectiveness (though some were skeptical about the predictive accuracy of AI models), and addressing concerns around digital exclusion.

Conclusions: The findings highlight the potential for AI-assisted tools to streamline and optimize the SMR process, particularly for patients with multimorbidity and complex polypharmacy. However, successful implementation depends on addressing concerns related to workflow integration, user acceptance, and evidence of effectiveness. User-centered design is crucial to ensure that AI-assisted tools support HCPs in delivering high-quality, patient-centered care while minimizing cognitive overload and alert fatigue.

(J Med Internet Res 2026;28:e74304) doi:10.2196/74304

KEYWORDS

structured medication reviews; medicine optimization; health technology; risk stratification; artificial intelligence; AI

Introduction

Background

The growing prevalence of multiple long-term conditions and complex polypharmacy among older adults poses significant challenges for health care systems globally. Structured medication reviews (SMRs) are a key clinical intervention, approved by the National Institute for Health and Care Excellence (NICE), designed to facilitate shared decision-making between clinicians and patients, optimize prescribing, and reduce medication-related harm in patients at high risk who are experiencing problematic polypharmacy [1,2].

General practitioners (GPs), pharmacists, and advanced nurse practitioners who meet training criteria can conduct SMRs. As a commissioned service, the prevailing expectation is for clinical pharmacists within primary care networks (PCNs) to proactively identify patients suitable for an SMR and conduct these reviews [3]. However, it is increasingly recognized that effective SMRs are difficult to implement clinically due to time pressures, fragmented clinical records, and the cognitive burden placed on clinicians when trying to assimilate information from various different sources in order to make shared, person-centered decisions [4].

Currently, in the United Kingdom, a few artificial intelligence (AI)–assisted tools are available to help health care professionals (HCPs) prioritize patients for SMRs. Tools available are usually based on predefined conditions or medications; the examples include Ardens Search [5] and Proactive Register Management Diabetes [4,6]. Prescribing safety indicators have also been used as a technology-based intervention to identify potentially inappropriate prescribing to reduce the number of patients at risk of hazardous prescribing [7,8]. However, primary care–embedded clinical decision support systems (CDSSs), such as audit and feedback tools, are often limited by data supply [9]. Emerging digital technologies, including AI, offer opportunities to enhance the efficiency and effectiveness of the SMR process through automation of routine tasks, rapid data extraction and synthesis, and highlighting clinical risks to support decision-making. However, the integration of AI into clinical workflows is in its infancy, and questions exist about its accuracy, clinical utility, usability, and trust. These implementation barriers are currently unexplored.

This study is part of a larger DynAIRx (AI for dynamic prescribing optimization and care integration in multimorbidity) project. Our research to date has highlighted the time-intensive nature of SMRs and the lack of AI-assisted tools to efficiently identify and prioritize patients [4]. Findings emphasized the need for an AI-assisted tool to identify, prioritize, and reduce the time needed to understand the patient journey in order to optimize medicines appropriately and reduce the risk of potential harm from medicines [4]. DynAIRx involves developing novel AI-assisted approaches to improve the efficiency of SMRs. The planned DynAIRx tool will comprise 4 main components: stratification of patients, clinical trial emulation to understand real-world risk of deprescribing, patient journey visualization through interactive timelines, and a knowledge support system integrating individualized patient risks to support decision-making. The deep learning AI component of this is to develop a tool to stratify patients most in need of an SMR. The DynAIRx stratification tool will compare 2 main approaches to identifying which patients are most at risk of medication-related harm: investigating the trade-off between model performance and explainability in the SMR context. First,

XSL•FO
RenderX

a simple logistic regression model not only gives a baseline performance level to assess the AI-based approach but also is clearly explainable and technically feasible to implement within clinical systems, such as EMIS and SystmOne. Second, a novel approach based on graph neural networks will be used to incorporate the sequence and timing of clinical events into predictions. This is likely to offer superior performance but will be difficult to explain and implement.

Large language models (LLMs), such as ChatGPT (OpenAI), are breaking new ground as an adjunct to support clinical decision-making. In radiological decision-making, ChatGPT recently showed impressive accuracy in the appropriate identification of imaging to support breast cancer screening [10]. There are several proof-of-concept AI-assisted tools in development to support complex polypharmacy. For example, the approach based on discriminator-enhanced encoder-decoder architecture for accurate prediction of adverse effects in polypharmacy is an AI model developed to predict adverse drug-drug interactions [11]. However, its effectiveness in a clinical setting has not yet been attempted. A new LLM based on retrieval-augmented generation has been developed to support pharmacists in identifying medicine errors [12]. User testing has been undertaken with simulation of complex scenarios and a multidisciplinary expert panel; however, true workforce implementation is still to be undertaken. Drug GPT is a specialized proprietary LLM tool for predicting medication safety events, developed by Oxford's AI for Healthcare Lab. It initially garnered popular attention upon the release of the preprint in 2023, including a review in the *Guardian* [13]. However, the preprint was subsequently removed by the authors, and its route to clinical implementation remains unclear [14]. Despite rapid progress, there remains limited understanding of what HCPs and patients actually need from such tools to support SMRs and, importantly, how to embed them into routine clinical practice, particularly in the context of multimorbidity and complex care.

While not all components of the DynAIRx polypharmacy tool will be AI assisted, the field of AI-assisted health technology is rapidly advancing. Consequently, it is critical to understand the user requirements for AI-assisted technologies now, as it may in fact be the case that simple, non-AI solutions can address the challenge in a straightforward and explainable way. Therefore, the health care sector is at a critical juncture when it comes to understanding end-user requirements for medication support.

This Study

This study aimed to explore the perspectives of both HCPs and patients on the potential role of AI in supporting SMRs, with a focus on identifying the core user requirements, anticipated benefits, and key barriers to implementation.

Methods

Participants and Recruitment

This study sought to recruit HCPs or management professionals from UK primary care (community based) and secondary care (hospital services) settings, for whom reviews of prescription medications form a routine part of clinical workload. Participants included those working in general practice; secondary care hospital services (geriatric medicine, clinical pharmacology, falls clinics, and mental health practitioners); clinical commissioning, service management (practice managers); and pharmacists, including PCN pharmacists who conduct SMRs across multiple GP practices. Patient participants included (1) those with mental and physical comorbidities, (2) those with complex multimorbidities, and (3) older people with frailty. In addition, patient and carer representatives from these 3 key multimorbidity groups were recruited, comprising adults aged >18 years with or caring for someone with multimorbidity (4 or more), coexisting mental and physical health problems, ≥10 or more prescribed medications, or frailty. Patient participants self-identified as not digitally engaged. As the General Data Protection Regulation was not required, we did not collect demographic data from patient participants.

Purposive sampling identified potential HCP participants actively involved in medicine optimization services through the researcher's clinical and professional networks. Snowball sampling, where current participants referred others, helped identify contacts through existing service providers and advertisements in GP forums and at national events related to clinical polypharmacy research. Patient representatives were recruited purposively via advertisements through the National Institute for Health and Care Research Applied Research Collaboration public advisor networks and research databases at the researcher's host institutions.

Ethical Considerations

The Newcastle North Tyneside Research Ethics Committee (22/NE/0088) granted ethical approval for the DynAIRx study. Written consent was obtained before participation, and withdrawal of consent was permitted at any stage, including after data collection. Audio recordings were transcribed verbatim, anonymized to remove any potentially identifiable information, and assigned participant codes before recordings were subsequently deleted. All data were stored on secure servers in accordance with data protection regulations. Participants received modest compensation in the form of a voucher to acknowledge their time and contribution, consistent with ethical guidance. No participants withdrew consent for the use of their data in this study. Sessions were conducted in person and online (via Microsoft Teams), lasting from 49 to 109 minutes. Data collection and analysis occurred concurrently, adhering to the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist for comprehensive reporting (Multimedia Appendix 1).

Data Collection

Data collection occurred from November 2022 to November 2023. Focus groups (FGs) and semistructured interviews were conducted to gather participants' views. Patient participants were involved in FGs to discuss their shared experiences, while FGs and individual interviews were conducted with HCPs to accommodate time constraints. A semistructured topic guide was developed collaboratively by the research team, informed by existing literature and expert input. The topic guide focused on key elements relevant to the study aims and reflected the

XSL•FO
RenderX

literature and practical insights from clinical practice. This guide was used consistently across both FGs and interviews. Interviews were conducted to complement purposive sampling and address any gaps in representation. These took place either before FGs, to inform key discussion areas, or alongside them for participants unable to attend a group session. FG topic guides and interview schedules were developed and refined by the clinical members of the research team (LEW, AA, AAW, FSM, and AG) and tailored to HCP and patient groups. The topic guides explored the current challenges and limitations in the SMR process that existing advancements in AI and machine learning (ML) can address and the essential components for a user-friendly prescriber feedback system. We also asked participants questions to identify the key components and functionalities needed in a prescriber feedback system to ensure it is useful and user-friendly. Figure 1 depicts the visual representation of the proposed components of an AI-assisted tool (the DynAIRx tool).

Figure 1. Visual representation of proposed components of an artificial intelligence (AI)–assisted tool (the DynAIRx tool [AI for dynamic prescribing optimization and care integration in multimorbidity]).

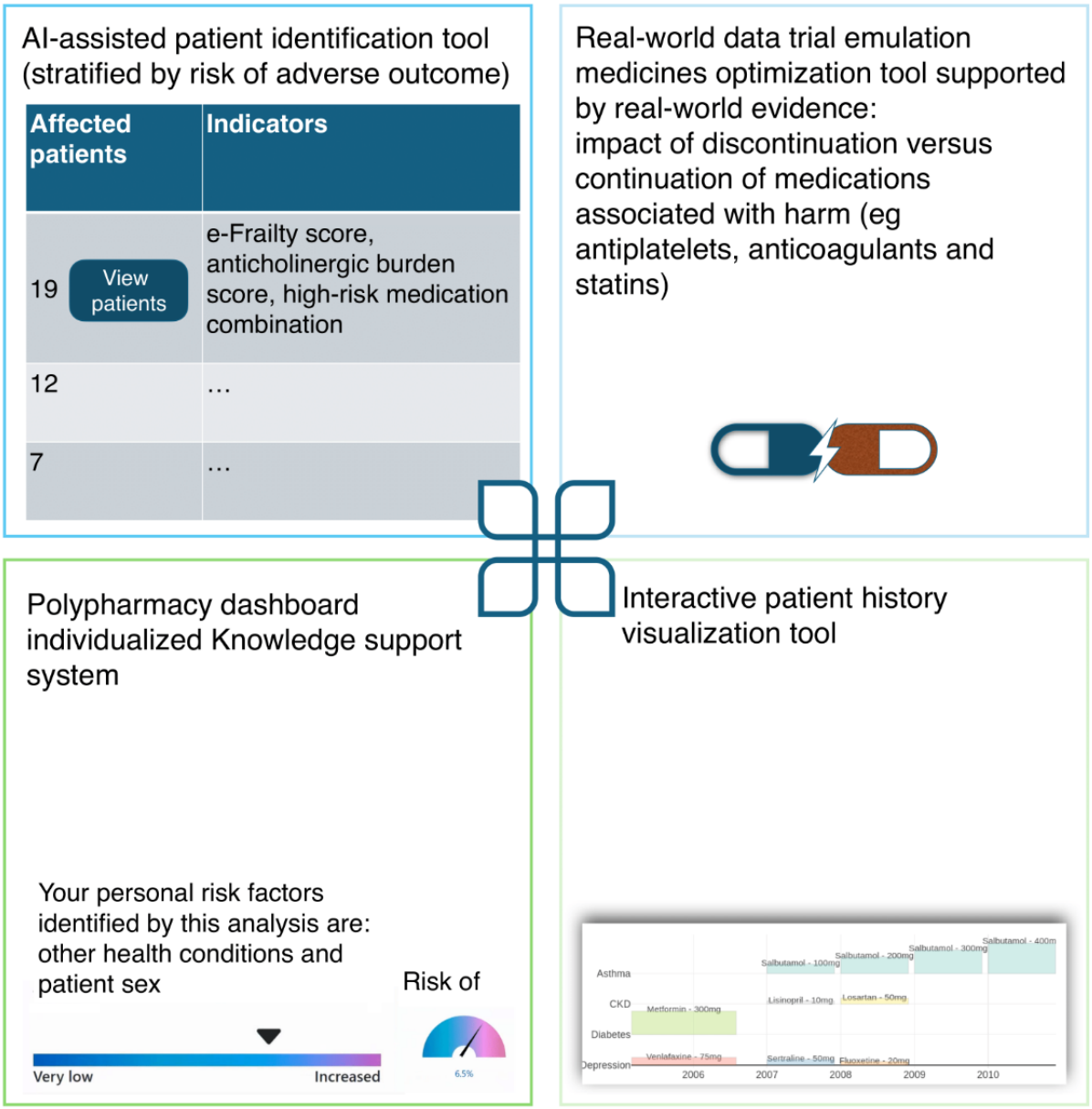

Data Analysis

Data were analyzed thematically, with coding independently conducted by researchers (AA and SAW). Researchers read transcripts to familiarize themselves with the data. Initial coding, guided by inductive reasoning, was conducted by AA and SAW, who collated and examined codes to identify themes. The interview and FG transcripts were coded concurrently to capture both individual and group perspectives. The multidisciplinary coding team (AA, SAW, LEW, AAW, and FSM), comprising clinicians and researchers, engaged in regular coding clinics in a reflexive practice to ensure rigorous and transparent qualitative analysis. These sessions were used to discuss emerging codes, refine codes, and assess data saturation. Discrepancies were

XSL•FO
RenderX

resolved through discussion with the wider team of researchers and clinicians, ensuring diverse perspectives informed theme development and interpretation. This reflexive approach balanced interpretations and mitigated biases, grounding the analysis in participants' narratives. Themes were defined and supported by quotes, with detailed notes maintained to ensure analytic rigor and plausibility. The dataset underwent hybrid inductive and deductive thematic analysis, with iterative revisions of codes and themes. Data saturation was defined as the point at which no new themes or subthemes emerged from the data concerning the already existing themes, which were richly supported by the data. During coding clinics, the research team compared newly coded transcripts against the developing codes to assess whether additional data contributed to novel insights. Recruitment ceased once all team members agreed that no new codes or meanings emerged from successive transcripts, and the thematic structure was considered sufficiently rich to address the study aims and indicate thematic saturation.

Results

Overview

In total, 6 FGs with HCPs (n=21) and 3 FGs with patients (n=13) were conducted (Table 1). A further 5 semistructured interviews with HCPs took place (Table 2). The interviews were undertaken to explore topics in greater depth and address any gaps in purposive sampling. Moreover, the number of participants in each FG differed based on HCPs' availability. However, this did not impact data analysis or saturation, as all HCPs undertook SMRs and expressed their views and requirements for an AI-assisted tool to allow the SMR process to be more efficient and effective.

Table 1. Participant type and the number of participants who took part in the focus groups (FGs; n=34).

FGs	Participants, n (%)
General practitioner FG1	2 (6)
General practitioner FG2	6 (18)
Pharmacist[a] FG1	3 (9)
Pharmacist FG2	5 (15)
Clinical pharmacologist FG1	3 (9)
Psychiatrist FG (mix of secondary care and prison care)	2 (6)
Patient FG comprising individuals with mental and physical health comorbidities for whom prescribing for mental health improvement could lead to adverse physical health consequences	6 (18)
Patient FG comprising those with complex multimorbidity (≥4 long-term health conditions and taking ≥10 drugs)	4 (12)
Patient FG comprising older people with frailty who were at a high risk of adverse outcomes	3 (9)

[a]Mix of primary and secondary care pharmacists.

Table 2. Participant type and the number of participants who took part in the semistructured interviews (n=5).

Participant	Interviews, n (%)
Primary care pharmacist	1 (20)
Secondary care pharmacist	1 (20)
Policy maker	1 (20)
Secondary care psychiatrist	1 (20)
Postgraduate GP[a] trainee	1 (20)

[a]GP: general practitioner.

HCPs conducted SMRs either proactively or reactively, depending on staff capacity, organizational contracts, and practice size. The presence of a PCN pharmacist facilitated proactive SMRs by ensuring that patients meeting directed enhanced service requirements were identified and invited for an SMR. In contrast, GPs and secondary care clinicians often conducted opportunistic medication reviews. Regardless of how patients were identified for a medication review, HCPs described the significant preparation time required to gather and interpret patient information, citing the lack of efficient methods to identify patients at risk of medication-related harm who would benefit most from an SMR [4]. This prompted discussion on how AI approaches could be used to improve the SMR process, including the potential barriers to the uptake and use of AI-assisted tools to support SMRs.

The following 4 overarching themes were developed from the analysis:

1. Innovative AI potential
2. Optimizing electronic patient record (EPR) visualization
3. Functionality of the AI tool for SMRs
4. Facilitators of and barriers to AI tool implementation

XSL•FO
RenderX

Innovative AI potential referred to the emerging possibilities and future impacts of applying AI technologies in health care contexts. Optimizing EPR visualization concerned the enhancement of clarity, usability, and accessibility of clinical information presented within EPR systems. Functionality of the AI tool for SMRs examined how the AI-assisted tool operated and supported the delivery of SMRs. Finally, facilitators of and barriers to AI tool implementation encompassed the organizational, technical, and human factors that influenced the successful integration and effective use of AI-assisted tools in clinical practice.

These themes were not entirely discrete; they reflected interrelated aspects of participants' experiences and perspectives on AI integration in clinical practice, with points of overlap and influence between them. Aspects, such as system usability, functionality, and perceived potential, often interact within the broader context of AI-assisted tools in health care. Participants' perceptions and responses varied according to their professional or personal role (eg, GP, pharmacist, psychiatrist, and patient), highlighting the need for AI-assisted tools to be sufficiently adaptable to address the differing needs of key stakeholder groups.

Innovative AI Potential

Participants expressed their views on the potential utility and advantage of AI in identifying patients at risk of medication-related harm or those who might benefit most from an SMR. HCPs emphasized the need for a tool capable of comprehensively searching EPRs to identify patients with complex multimorbidity. Such a tool should dynamically adjust search outcomes in real time, prioritizing patients who require immediate SMRs.

There was a desire to see AI-assisted tools that could learn autonomously from the historical health care record to identify which factors are contributing to potential medication-related harm. Participants showed great interest in the incorporation of AI with a health care tool to automate tasks and reduce delays in risk prediction. In addition, participants wanted these tools to show patients' medical history in a holistic way using AI capabilities and have the AI be explainable to understand why and how the patient was triggered for a medication review:

> *If you had a funky IT program that looked at medication, looked at what other stuff was happening, looked at, you know, bloods, these are patients that I'm really worried about. So, you know you're talking about machine learning in your project, one of the things, you know, I think will be really interesting to do would be to actually ask the computer what the predictors for certain harms are. We know patients when they fall for multiple reasons, it's not just medicines, but actually wouldn't it be good to see that these are the key circumstances that patients fall under, and then if those circumstances ever happened that patient would be, you know, triggered for a review.* [Policy maker; interview]

> *It could potentially work in real time as well...having something which is live so constantly producing the order of patients who you should be reviewing based on, I don't know a patient might have been discharged from hospital last night that patient might become a bit more high risk and therefore it needs to review earlier. So it flags upon our systems as a, you know, using the AI that this patient will probably need to review in the next four or five days.* [Pharmacist 2; interview]

Leveraging data analytics with ML was viewed as an opportunity to flag patients on complex medication regimens by assessing their health records and prioritizing those patients at risk of medication-related harm. Moreover, aligning patient risk levels with the GP practices staff capacity within PCNs would ensure that those who need immediate attention are seen promptly:

> *That is a real issue for us. It's a real issue for practice actually. So this is why I think the tools have to be a bit more cleverer than just generating, you know, we can generate a list of patients today and that happens, and PCNs at the moment essentially do that, but what you have to do is almost match the list that's generated to the capacity of the build this so you can, the practice has to say that across my PCN I've got, you know, 100 appointments a week to deal with these sort of patients then the tool has to generate that...People would not switch it on if they felt that it could generate lots of patients you would not then see.* [Policy maker; interview]

The development of an AI-assisted tool to support the SMR process prompted discussions on how ML tools could predict risk, identify prescribing gaps, highlight lifestyle and family history risk predictors, and detect potential adverse drug reactions. Advanced digital health tools with AI-assisted features and data analytics could enhance patient engagement by enabling holistic discussions about the patient's risk trajectories and how their medicines can be optimized to reduce any medication-related risks:

> *There used to be a tool, I think it was developed in Australia or New Zealand, but basically it showed a graph of heart disease and trajectory towards symptoms. And you could have a discussion with patients and you can say well look, if we bring your blood pressure down by this much then this is your trajectory...if we stop you smoking, then this is your trajectory. If you develop diabetes then this is your trajectory, and that was probably the single most powerful tool I had to convince people to optimize things like blood pressure or cholesterol reduction.* [Participant 1; GP FG1]

> *I'm sure that having an AI trawl through drug prescribing gaps would tell us quite a lot about medication that may not be taken when it's supposed to have been. We kind of think about that in terms of people misusing analgesics but actually for the elderly population they'll very often just order it because they don't know how to tell us they don't like it or they don't want to take it anymore because they're very*

XSL•FO
RenderX

> *much of the mindset that doctor knows best and they don't want the conflict. But they'll very often forget to order medication and that can be a giveaway.* [Participant 5; GP FG1]

> *I'm thinking of an example where someone has attended an appointment and mentioned that they think a drug is causing a side effect for them, I would imagine that someone would document experiencing this? Being able to see that would be very helpful to try and further add to why things might have been stopped or why a patient might have stopped taking them and maybe not told anyone.* [Participant 2; pharmacist FG1]

> *We did a piece of work recently on familial hypercholesterolaemia and pulled up a lot of patients we didn't realize had family history of massive cholesterol levels and they hadn't realized it was potentially hereditary.* [Participant 5; pharmacist FG2]

Optimizing EPR Visualization

Participants pointed to the challenges associated with the time required to gather and interpret a patient's medical history, emphasizing the need for an AI-assisted tool that optimizes the presentation of relevant information within the EPR. This included reorganizing readily available data to provide a clearer view of the patient's medical history and social circumstances to produce accessible visualizations of the medication timelines, including what medications were prescribed for what condition. Several participants suggested displaying the patient's medical history in a timeline format, detailing key events, such as medication initiation, titration, or discontinuation, diagnosis dates, and recent relevant blood test results:

> *So something pictorially, which helps represent the information in a clearer way, I suppose. Yeah, maybe more longitudinal kind of...And representation of say, when medications were started and titrated up and previous medications, when they were brought in and when there were stopped.* [Psychiatrist; interview]

> *I suppose my top 5 would be: something that highlights previous courses of the same type of medication or the same class, so, for example, if I type in depression as a code, I want an automatic list of every antidepressant they've been on previously and how long they've been on it and which ones they haven't had, even the new ones that are coming on line. I want a list of when they had prednisolone last if they have chronic lung disease, [and] how many courses in the last year they've had without searching for it.* [Participant 1; GP FG1]

AI-assisted tools that reduce the time involved in routine tasks, such as finding information, calculating doses, or assessing disease risk, were welcomed. HCPs were conscious of ensuring that any AI-assisted tool did not overwhelm the user with excessive "pop-up" functions on the display and should not overburden the user's view. There was strong support for a visual timeline that would detail the patient's diagnosis and prescribing and deprescribing journey, along with relevant investigations, diagnostic letters, which specialty diagnosed the condition and started the medication, and the reasons for certain medicine changes. Figure 2 presents the suggested AI-assisted tool features:

> *I've been dreaming about the timeline you showed [laughs] to be able to, in the way that I've imagined it, at the click of a button know...when all the drugs were started and what else was diagnosed around that time [would be] great. And then I don't have to spend any time trying to put that together, that information is there for me. Thinking outside of a hospital setting, if I've got recent bloods and any sort of risk calculations that I want already there on the page from the most recent things, [that is] even better. The amount of time [it would save]; the computer system I work with tells me eGFR, [but] I spend a good chunk of my day calculating everyone's creatinine clearance.* [Participant 2; pharmacist FG1]

Participants also described the challenges around investigating the indication for each prescribed medicine, stating that any AI-assisted tool should incorporate the medication indication:

> *I think for me the most important thing that's missing is indication-based prescribing. Because when we are doing our medicine reviews just trying to work out why anyone's on, you know they could be on an ACE inhibitor, why are they on it, you know, they could be on citalopram, you know, why are they on it. And then it's almost impossible to stop it if you can't work out why someone started it to begin with. So, I think, for me that would be the key initial thing.* [Participant 1; polypharmacy FG]

Participants described how information within a patient's EPR should be visualized, focusing on rearranging and presenting the information in a more intuitive and insightful way to enhance understanding of the patient's medical journey and provide the information in an easily understandable format to support decision-making. They recommended color-coding blood results, highlighting the most recent hospitalization or medication review, and flagging risky medication combinations. One participant noted the following:

> *Reviewing the bloods was quite time-consuming at times, but equally, it was color-coded, so if it was in red it means it's bad news. If it's not in red, it's OK. So, a quick eyeball of words can be often sufficient.* [Pharmacist 2; interview]

Another participant added the following:

> *The kind of things that we need [are] things like...when was their last medication review done or when was the last SMR done, for example? When was the last hospitalization event, for example? Have they been hospitalized or discharged recently? Then it'd be nice to visualize, you know, and the combinations of drugs which could be risky.* [Pharmacist 2; interview]

XSL•FO
RenderX

Figure 2. Suggested features for the development of an artificial intelligence (AI)–assisted tool for use in medication reviews. BNF: British National Formulary; eGFR: estimated glomerular filtration rate; SMR: structured medication review.

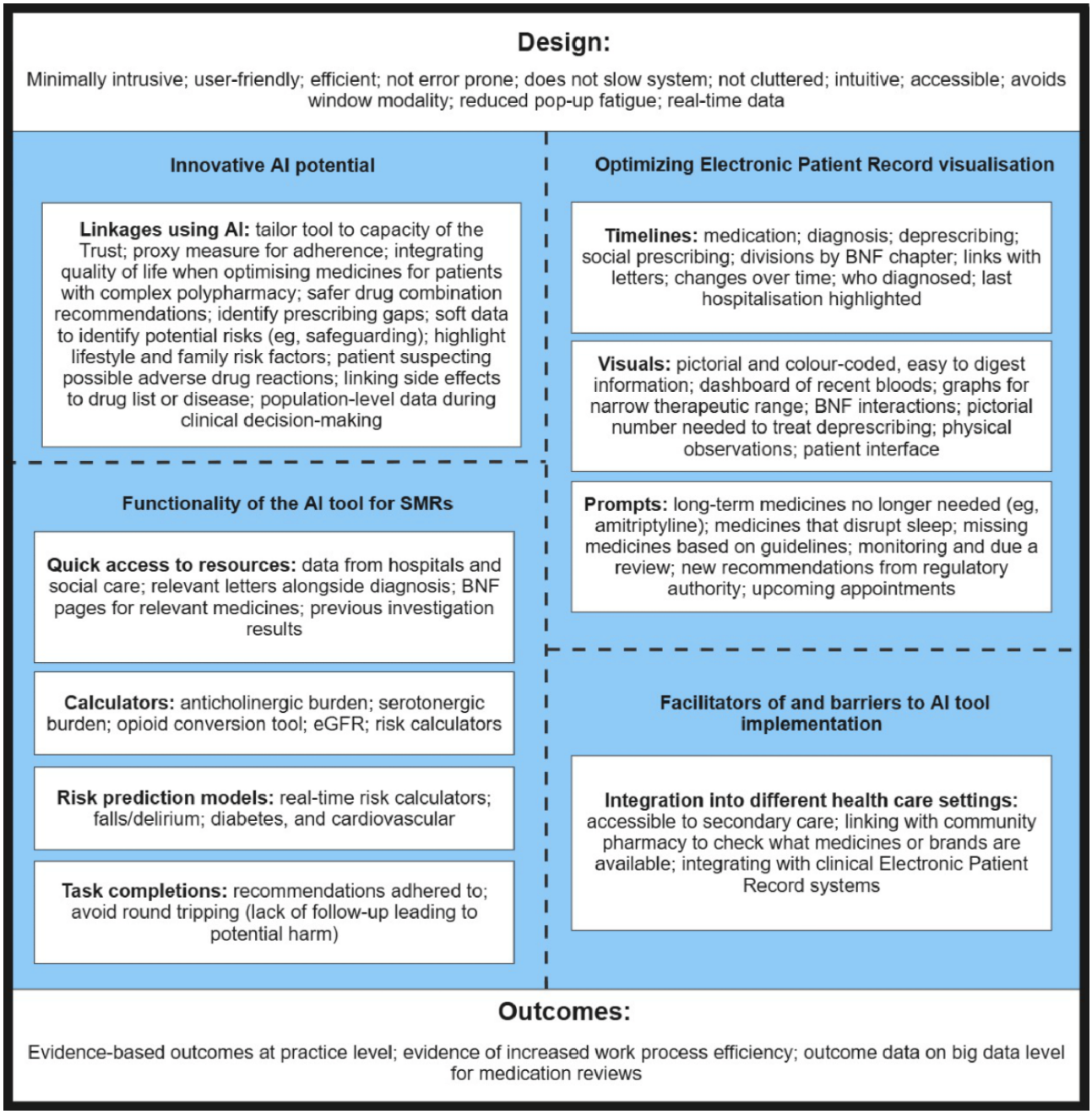

Functionality of the AI Tool for SMRs

Participants articulated their preferences for the functionality of the AI-assisted tool, particularly emphasizing the need to avoid "pop-up fatigue." They suggested implementing specific, targeted pop-ups rather than numerous interruptions that could disrupt workflow. One participant remarked as follows:

> *Any pop ups that you've just got to click through does add even seconds to your working day in each patient record so that can be annoying.* [Participant 1; GP FG2]

Instead, participants preferred notifications for critical issues, such as missing medications in a patient's record that align with guidelines. For example, a participant stated the following:

> *If a patient has recently had a myocardial infarction...it pulls out your main groups of drugs from the NICE guidelines and flash up oh they're not on an ACE inhibitor and you've got to say why.* [Participant 3; pharmacist FG1]

Similarly, they recommended notifications for medications needing review, such as long-term prescriptions of amitriptyline:

> *Just to make people think, actually. This thing which was started for sleep isn't helping.* [Psychiatrist; interview]

Participants also recommended pop-ups for new guideline recommendations:

> *To meet the Commission of Human Medicines and Medicines and Healthcare products Regulatory*

XSL•FO
RenderX

> *Agency recommendations that came in in December, a way of being able to see what antiepileptics have been tried for someone and how long they've been on valproate would really inform the review process.* [Participant 2; pharmacist FG1]

Participants valued AI-assisted tools that enhanced workflow productivity by providing quick access to resources, such as the *British National Formulary* or NICE guidelines based on prescribed medicines or flagged blood test results. One GP noted the following:

> *I want a dashboard of most recent blood tests and current guidance checked against that for where there's missing [medication] gaps, just those things alone would probably save about 2- or 3-weeks' worth of work across my practice per year.* [Participant 1; GP FG1]

In addition, participants highlighted the need for more efficient access to hospital discharge letters and relevant patient information, which often reside in separate electronic systems, such as EMIS (an EPR system) and Docman (document management system). One participant stated the following:

> *If you could draw up relevant letters alongside diagnosis so that you don't have to then trawl through a different system or then write to the GP to request added information on certain diagnosis.* [Psychiatrist; interview]

Furthermore, automated calculators for clinical decision-making, such as anticholinergic burden calculations, estimated glomerular filtration rate, and opioid conversion calculators, were seen as valuable tools to reduce time and enhance efficiency. For example, a GP shared the following:

> *I want automatic opioid conversion so that when they're on tramadol, co-codamol, oramorph, it just gives me a single figure to aim towards and show them, you know, we are at horse tranquilizing level or at dragon tranquilizing level and we need to kind of bring it down.* [Participant 1; GP FG1]

SMRs, whether proactive or opportunistic, often require multiple appointments and may necessitate further investigations by another HCP in a different setting, such as a hospital. Participants were concerned that current systems might fail to notify HCPs of patients needing follow-up care, potentially leading to harmful outcomes for the patient:

> *One of the big issues with medical software is not doing what you describe. Computer programmers call it "round tripping," so basically things aren't followed up and the loop isn't closed, and we see huge amounts of error in medicine because of it. For example, we will refer and then it's off our radar and it's on someone else's radar and then they don't show up at the hospital, and then all of a sudden six months down the line they end up with a malignancy, and then it's back to the original person who referred problem because that loop wasn't closed or the data wasn't followed up. So, having some form of checking system would hugely reduce error and risk within medicine, particularly when it comes to prescribing. So, for example, I've referred them to the physio, the physio then closes that loop when they seem them and then that's dealt with, or the physio does a hydrocortisone injection for example, which I may not need to know about unless I then send them somewhere else, or they are on anticoagulants, or something like that. So, having a tool that ties up those loose ends would be worth its weight in gold I think, particularly in terms of huge amounts of risk reduction.* [Participant 1; GP FG1]

Facilitators of and Barriers to AI Tool Implementation

Participants were curious to understand where DynAIRx might be embedded and how it will be adopted in practice. The main goal was to ensure that any new software does not interfere with current work processes. They described concerns about the cost of adopting it in practice, information governance, and potential medico-legal concerns:

> *I know that I understand X's point about having it that you are able to record what decisions you made within the tool but I wonder if it might be preferable to have a way of a text summary that could be then copied from the tool into the notes. Or, even a way of having a screen shot into letters so, you know, if there were every any questions what was it that you were looking at on the day what information was the tool giving you and if you then made an interesting decision based off that information, that's still your decision. But you've clearly got what is was that you could see.* [Participant 2; pharmacist FG1]

Incentivizing the use of the tool to improve uptake was seen as crucial, with participants suggesting that DynAIRx should be offered as an option to HCPs. They recommended explaining the benefits of using DynAIRx as an additional supportive tool alongside traditional EPR systems, providing examples, such as improvements in efficiency, ease of use, and quick access to information. In addition, HCPs recognized that technology is a double-edged sword, which could either reduce or increase the gap between the HCP and the patient. Consequently, HCPs were open to incorporating a patient interface that would present information pictorially to support shared decision-making. Patients indicated that they would find this useful as they were keen to take a more active role in decision-making. However, many patients were unaware of medication review–related services available to help optimize their medicines:

> *I have never been involved in a review either...Would it just be the GP?...I am going to ask this time whether I can be involved.* [Patient 3; patient, multimorbidity FG]

HCPs stated that, from their experience, patients are fairly comfortable with technology, which was echoed in the patient FGs:

> *I have no problem whatsoever with using [AI] tools, especially if it is to back up the thought process.* [Patient 4; patient, multimorbidity FG]

XSL•FO
RenderX

> *What I think, being a patient, is that you need to have the trust from the GP and the GP pharmacist, whoever is prescribing and using AI, they need to educate people. They need to raise awareness about it. And then you have the trust with the GP then you feel comfortable.* [Patient 1; patient, multimorbidity FG]

However, this did not necessarily mean that all patients were able to use technology themselves to communicate or provide feedback to HCPs through surveys and questionnaires, suggesting the need to consider potential digital exclusion:

> *Not digitally everybody, there are some barriers, language barriers, cultural barriers, and economic barriers for the disadvantaged so we have to think about equity as well. So these barriers are obviously not everything, but we should open up, give the opportunity to all those populations to have the opportunity to discuss and take it rather than have both, it should be both.* [Patient; complex multimorbidity FG]

While HCPs expressed a desire for risk-prediction tools within the DynAIRx software, some remained skeptical about these tools. They emphasized the importance of providing evidence that the model had been validated or that the software itself improved patient outcomes:

> *One of the reasons why people may not use it is that they, clinicians often ask for evidence, so they need to see evidence. A published paper or a trial of actually being useful. So does it improve outcomes with patients? These one of the biggest barriers that I've come across in my time with tools.* [Pharmacist 2; interview]

In addition, the need to ensure confidentiality and consent in AI-assisted tool use is paramount, given the amount of data that would be incorporated into DynAIRx. Patients were aware of the fast-paced innovations taking place in technology and the expectation that technology should be included in health care practice:

> *The world is changing. First we heard about the autopilot, now they are testing [cars] without a driver, virtual GP...now AI in the medical sector as well. The world is changing.* [Participant 2; patient]

Discussion

Principal Findings

Our study outlines the user requirements for the development of an AI-assisted tool to improve the process of conducting an SMR involving patients with complex multimorbidity and polypharmacy. This includes optimizing EPR visualization with a focus on developing patient timelines that outline the patient's medical journey to include when a condition was diagnosed, associated medicines prescribed for that condition, and any associated laboratory results, to name a few. Moreover, participants described the preference for evidence-based outcomes and the use of explainable AI to identify patients at risk of medication-related harm who would benefit from an SMR, determine their risk trajectory over time, and align those patients to practice staff capacity. By "explainable AI," we mean models or algorithms designed with transparent logic or post hoc interpretation methods (eg, feature-importance heatmaps and rule-based approximations) that allow HCPs to understand why a particular patient was flagged as high risk rather than relying on opaque "black box" predictions. From our findings, it is clear that HCPs require an AI-assisted tool that will streamline their work processes when conducting an SMR to easily find information related to the patient and incorporate any risk prediction models. Studies show that CDSSs have the potential to improve process outcomes [15,16]; however, access to CDSSs alone does not guarantee user acceptability or uptake [17]. Concerns remain that complex clinical decision-making, particularly for patients with multimorbidity and polypharmacy, may not be easily translated into algorithms [18] or, at the very least, may have HCPs view the validity of algorithms with some skepticism [19]. This can be seen in the literature that reflects HCP preference for knowledge-based CDSSs over non–knowledge-based CDSSs [19,20]. Despite this, studies show that the lack of algorithm complexity in CDSSs can frustrate HCPs, particularly in how information is presented [18,21]. Moreover, user-centered design is crucial in the development of a CDSS to optimize how information is presented to HCPs to manage cognitive load, alert fatigue, and the impact on workflow [22,23].

Comparisons With Prior Work

AI-assisted tools designed to assist in SMR processes must be tested to ensure they effectively identify both patients who are at risk of medication-related harm and those who have the greatest capacity to benefit from an SMR. These 2 groups are not necessarily synonymous, as some variables that may contribute to harm are not modifiable (eg, very advanced age and frailty), whereas the capacity to benefit from an SMR may be determined by modifiable factors (such as identification of prescribing cascades, drug-drug or drug-disease interactions, and adverse drug reactions). The tool must be sensitive when identifying patients at high risk and display medicines that could be optimized or deprescribed efficiently, enabling future evaluation of medication-related interventions using available data. Patients taking part in an SMR are also likely to have multiple appointments and be referred to different specialists, making it difficult for the HCPs who initiated the SMR to follow this journey. Few studies have explored the interoperability between primary and secondary health care settings, which is essential for effective communication and coordination between HCPs, and is perceived to have a high impact on patient safety [24,25]. Our findings indicate a clear need for an AI-assisted tool to support clinicians and patients during consultations. An integrated system within the EPR software could help summarize and visualize patient journeys and medicine-related information, present personalized risks of harms and benefits of medicines combinations [26], highlight the uncertainties in these risks, and support shared decision-making with patients. While the findings of this study are based on anticipated rather than observed experiences with an AI-assisted tool, engaging end users at the outset before tool development enables the design of an AI-assisted tool that is contextually relevant, cost-effective, and more likely to be adopted in practice [27,28].

XSL•FO
RenderX

Future stages of this work will involve the development and testing of a prototype to validate and refine these preliminary findings through direct user interaction and usability testing, as demonstrated by previous studies [29-32]. This iterative, user-centered process will help ensure that the AI-assisted tool is not only functional but also implementable in real-world settings.

Previous literature highlights the preference among HCPs for knowledge-based over non–knowledge-based CDSSs, which aligns with concerns about the validity and acceptability of algorithms in complex clinical scenarios [19,20]. While frustrations with simplistic CDSS algorithms have been noted [14], our study contributes additional evidence emphasizing the need for user-centered design to address cognitive load, alert fatigue, and workflow impact. The clinical decision support five rights model states that the right information should be presented to the right person, in the right format, via the right channel, and at the right time [16]. Studies on the development of CDSSs emphasize the importance of understanding the needs of users and receivers during the early stages of software development [17,18]. Interoperability between health care settings, while unexplored in previous work, emerged as a critical component for effective SMR implementation and improved patient safety in this study [10,19]. Integrated care boards link primary and secondary care data across their local areas with input from EPR vendors, such as EMIS and SystmOne. An approach consisting of fully linked local data, with minor regional differences in data formatting, may begin to emerge in the coming years as one way to proceed. Efforts, such as the National Health Service Federated Data Platform, aim to unify these actions across the nation and may provide additional clarity on this approach in the coming years. The next stage of the DynAIRx project will involve learning from the user requirements of our stakeholder groups to develop prototypes of an AI-assisted tool. We anticipate the developed prototypes to likely fall under the category of a knowledge support system, which provides HCPs with knowledge that already exists, such as contextual information about a patient drawn from several sources, including historic data on clinical outcomes from comparable patient groups, medical knowledge, and AI to support HCPs during consultations [29,33,34].

To bridge our user-centered requirements with real-world adoption, we recommend framing the development and deployment of the AI-assisted tool for SMRs within established implementation science frameworks, such as the Consolidated Framework for Implementation Research [35]. Under the Consolidated Framework for Implementation Research, the intervention characteristics (eg, usability, adaptability of the patient-timeline visualization, and risk-prediction algorithms), the inner setting (practice culture and EPR interoperability), the outer setting (national data-linkage initiatives, such as the National Health Service Federated Data Platform), the characteristics of individuals (clinician attitudes toward explainable AI), and the implementation process (engagement, training, and feedback loops) all warrant deliberate planning and tailoring.

A staged implementation approach could proceed as presented in Textbox 1.

Textbox 1. Staged implementation approach.

Before implementation (exploration and preparation)

- Conduct targeted workflow analyses in a small number of pilot practices to refine how the timeline and risk outputs map onto existing structured medication review processes
- Engage local electronic patient record vendors (eg, EMIS and SystmOne) to configure interoperability and data-security protocols

Implementation (initial rollout and training)

- Deploy the prototype in 3 to 5 “early adopter” practices, offering hands-on workshops and “superuser” support
- Establish a feedback channel (eg, biweekly focus groups) to iteratively refine

Sustainment (scale-up and longitudinal support)

- Expand to additional practices, embedding the tool within organizational reporting cycles
- Integrate decision-support outputs into routine safety audits and continuing professional development activities

Building on our staged implementation plan, future studies should focus on prospective validation of the AI-assisted tool for SMRs in live clinical settings. Key next steps include the following:

- Pilot effectiveness trials—randomized or stepped-wedge designs comparing standard SMR to SMR with AI support, measuring both process (eg, time per review and alert response rates) and patient-level outcomes (eg, incidence of medication-related harm)
- Usability and acceptability—mixed methods evaluations combining system-log analytics with qualitative interviews to understand how clinicians interact with features such as patient timelines and risk predictions
- Subgroup analyses—examining performance across different patient demographics and varying levels of multimorbidity complexity to ensure equitable benefit

Limitations

This research is part of a larger qualitative study exploring the barriers to SMRs and potential AI-assisted tools. Given the focus on digital-driven solutions, the HCP participants likely included those with a particular interest in such innovations, although efforts were made to include a diverse range of HCPs

XSL•FO
RenderX

from different practice backgrounds, regions, and care settings to mitigate this bias and help strengthen the generalizability of our findings within the context of SMR practices. Moreover, our eligibility criteria focused on recruiting HCPs who are actively involved in conducting medication reviews to explore the barriers to efficient and effective SMRs and how AI-assisted tools may address these barriers [4]. Some FGs had fewer participants due to the competing demands on clinicians' time. However, the data collected were rich and contributed significantly to achieving thematic saturation. While this study provides valuable insights into the user requirements for developing an AI-assisted tool, certain limitations should be acknowledged. First, the diversity of the patient group included in this study may not fully represent the broader patient population, particularly in terms of familiarity with AI-assisted tools. Patients' knowledge and understanding of such systems may vary significantly, potentially influencing the feedback provided. As a result, the findings may not fully capture the perspectives of patients with lower levels of digital health literacy or those who have more experience engaging with AI-assisted tools in clinical settings. Second, the study captures participants' hypothetical perceptions of the AI-assisted tool rather than their experiences of real-world implementation and use. While the findings highlight anticipated benefits and potential facilitators, they may not fully reflect the complexities of adoption and sustained use in practice. Consequently, the research team will develop a prototype based on our findings, which will be tested to validate results from this study and refine and iterate the prototype through a series of think-aloud sessions and semistructured interviews with HCPs. This will ensure the prototype is developed based on user requirements and allow the user to explore the utility of the tool. Moreover, we anticipate a future evaluation of the AI-assisted tool by implementing it into the HCPs' routine workflow. Future research involving live system implementation and longitudinal evaluation would be valuable in assessing the feasibility and actual impact of AI-assisted tool integration in health care settings.

Conclusions

This study highlights the potential of AI-assisted tools to enhance the SMR process for patients with complex multimorbidity and polypharmacy and the user requirements to develop an AI-assisted tool. AI-assisted tools may have the potential to improve patient identification for an SMR, assist the HCP to optimize patient medication by optimizing the EPR visualization of the patients' medical and social history, and support shared decision-making between HCPs and patients. In order to realize the full potential of AI-assisted tools for SMRs, national and local policy makers should consider earmarking targeted funding streams, such as through the National Health Service Digital's Innovation Accelerator or PCN transformation budgets, to subsidize early implementation and integration with EPR systems. Embedding AI-SMR competencies into continuing professional development requirements for pharmacists and GPs, alongside dedicated training grants, will help build workforce capability and ensure equitable uptake. Finally, linking reimbursement incentives (eg, enhanced quality and outcomes framework points to multimorbidity reviews that use validated AI support) could further drive adoption and standardize best practice across PCNs. However, to ensure successful implementation, it is essential to address concerns, such as cognitive load, alert fatigue, and system interoperability. The findings underscore the importance of developing explainable, evidence-backed tools that align with clinical workflows and demonstrate clear benefits to patient outcomes. Ultimately, integrating such tools into an EPR has the potential to improve the efficiency and effectiveness of SMRs, benefiting both patients and health care systems.

Acknowledgments

DynAIRx has been funded by the National Institute for Health and Care Research (NIHR) Artificial Intelligence for Multiple Long-Term Conditions call (NIHR 203986). MG is partly funded by the NIHR Applied Research Collaboration North West Coast. AAW is partly funded by a Health and Care Research Wales Research Time award (NHS-RTA-21-02). AC is funded by a National Institute for Health and Care Research (NIHR) Research Professorship award and supported by the NIHR Applied Research Collaboration Yorkshire & Humber, the NIHR Leeds Biomedical Research Centre, the Cross NIHR Collaboration in Multiple Long-Term Conditions, and Health Data Research UK, an initiative funded by UK Research and Innovation Councils, NIHR and the UK devolved administrations and leading medical research charities. This research is supported by the NIHR Applied Research Collaboration North West Coast. The views expressed in this publication are those of the authors and not necessarily those of the NIHR, NHS or the UK Department of Health and Social Care. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

ASA contributed to conceptualization, data analysis, methodology, project administration, supervision, validation, writing the original draft, and reviewing and editing the manuscript. SAW contributed to data analysis, methodology, and reviewing and editing the manuscript. AAW contributed to data acquisition, data analysis, and writing the original draft of the manuscript. FSM contributed to conceptualization, methodology, and reviewing and editing the manuscript. AA contributed to conceptualization, data acquisition, and reviewing and editing the manuscript. AC contributed to conceptualization, methodology, project administration, supervision, and reviewing and editing the manuscript. ES contributed to conceptualization, supervision, and reviewing and editing the manuscript. MG contributed to conceptualization, methodology, and reviewing and editing the manuscript. DB contributed to conceptualization, supervision, and reviewing and editing the manuscript. HC contributed to reviewing and editing the manuscript. AG contributed to reviewing and editing the manuscript. LH contributed to validation, visualization, and

XSL•FO
RenderX

reviewing and editing the manuscript. GL contributed to reviewing and editing the manuscript. EL contributed to project administration and supervision. SM, MO, OP, SR, RAR, and PS contributed to reviewing and editing the manuscript. MS contributed to conceptualization, methodology, and reviewing and editing the manuscript. TVS contributed to conceptualization, methodology, and reviewing and editing the manuscript. IB contributed to conceptualization, data analysis, methodology, project administration, supervision, validation, and reviewing and editing the manuscript. LEW contributed to conceptualization, data analysis, methodology, project administration, supervision, validation, and reviewing and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ checklist.
[PDF File (Adobe PDF File), 145 KB - jmir_v28i1e74304_app1.pdf]

References

1. Network contract directed enhanced service: structured medication reviews and medicines optimisation: guidance. National Health Service England. URL: https://www.england.nhs.uk/wp-content/uploads/2021/03/B0431-network-contract-des-smr-and-mo-guidance-21-22.pdf [accessed 2025-05-29]
2. Network contract DES guidance for 2024/25: part A - clinical and support services (Section 8). National Health Service England. URL: https://www.england.nhs.uk/publication/network-contract-directed-enhanced-service-guidance-for-2024-25-in-england-part-a-clinical-and-support-services-section-8/ [accessed 2025-05-29]
3. Network contract directed enhanced service – contract specification 2021/22 – PCN requirements and entitlements. National Health Service England. URL: https://www.england.nhs.uk/publication/network-contract-des-specification-2021-22/ [accessed 2025-05-29]
4. Abuzour AS, Wilson SA, Woodall AA, Mair FS, Clegg A, Shantsila E, et al. A qualitative exploration of barriers to efficient and effective structured medication reviews in primary care: findings from the DynAIRx study. PLoS One 2024;19(8):e0299770 [FREE Full text] [doi: 10.1371/journal.pone.0299770] [Medline: 39213435]
5. The leaders in providing EMIS web and SystmOne templates and resources. Ardens Healthcare Informatics. URL: https://www.ardens.org.uk/ [accessed 2025-05-29]
6. A health management tool for people with diabetes co-created by Lilly and NHS Devon CCG. PARM Diabetes. URL: https://parmdiabetes.co.uk/ [accessed 2025-05-29]
7. Abuzour AS, Magola-Makina E, Dunlop J, O'Brien A, Khawagi WY, Ashcroft DM, et al. Implementing prescribing safety indicators in prisons: a mixed methods study. Br J Clin Pharmacol 2022 Feb 29;88(4):1866-1884 [FREE Full text] [doi: 10.1111/bcp.15107] [Medline: 34625991]
8. Rodgers S, Taylor AC, Roberts SA, Allen T, Ashcroft DM, Barrett J, et al. Scaling-up a pharmacist-led information technology intervention (PINCER) to reduce hazardous prescribing in general practices: multiple interrupted time series study. PLoS Med 2022 Nov 16;19(11):e1004133 [FREE Full text] [doi: 10.1371/journal.pmed.1004133] [Medline: 36383560]
9. Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, French SD, et al. Audit and feedback: effects on professional practice and healthcare outcomes. Cochrane Database Syst Rev 2012 Jun 13;2012(6):CD000259. [doi: 10.1002/14651858.CD000259.pub3] [Medline: 22696318]
10. Rao A, Kim J, Kamineni M, Pang M, Lie W, Dreyer KJ, et al. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast imaging pilot. J Am Coll Radiol 2023 Oct;20(10):990-997 [FREE Full text] [doi: 10.1016/j.jacr.2023.05.003] [Medline: 37356806]
11. Kobraei K, Baradaran M, Sadeghi SM, Masumshah R, Eslahchi C. ADEP: a novel approach based on discriminator-enhanced encoder-decoder architecture for accurate prediction of adverse effects in polypharmacy. arXiv. Preprint posted online on May 31, 2024 [FREE Full text]
12. Ong JC, Jin L, Elangovan K, Lim GY, Lim DY, Sng GG, et al. Development and testing of a novel large language model-based clinical decision support systems for medication safety in 12 clinical specialties. arXiv. Preprint posted online on January, 2024 [FREE Full text]
13. Tapper J. DrugGPT: new AI tool could help doctors prescribe medicine in England. The Guardian. 2024. URL: https://tinyurl.com/3fckmxf7 [accessed 2025-05-29]
14. Liu F, Zhou H, Zhang W, Huang G, Clifton L, Eyre D, et al. RETRACTED: DrugGPT: a knowledge-grounded collaborative large language model for evidence-based drug analysis. Research Square. Preprint posted online on October 6, 2023 [FREE Full text] [doi: 10.21203/rs.3.rs-3411728/v1]
15. Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. Ann Intern Med 2012 Jul 03;157(1):29-43 [FREE Full text] [doi: 10.7326/0003-4819-157-1-201207030-00450] [Medline: 22751758]

XSL•FO
RenderX

16. Kwan JL, Lo L, Ferguson J, Goldberg H, Diaz-Martinez JP, Tomlinson G, et al. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. BMJ 2020 Sep 17;370:m3216 [FREE Full text] [doi: 10.1136/bmj.m3216] [Medline: 32943437]
17. Kouri A, Yamada J, Lam Shin Cheung J, Van de Velde S, Gupta S. Do providers use computerized clinical decision support systems? A systematic review and meta-regression of clinical decision support uptake. Implement Sci 2022 Mar 10;17(1):21 [FREE Full text] [doi: 10.1186/s13012-022-01199-3] [Medline: 35272667]
18. Greenes RA, Bates DW, Kawamoto K, Middleton B, Osheroff J, Shahar Y. Clinical decision support models and frameworks: seeking to address research issues underlying implementation successes and failures. J Biomed Inform 2018 Feb;78:134-143 [FREE Full text] [doi: 10.1016/j.jbi.2017.12.005] [Medline: 29246790]
19. Meunier PY, Raynaud C, Guimaraes E, Gueyffier F, Letrilliart L. Barriers and facilitators to the use of clinical decision support systems in primary care: a mixed-methods systematic review. Ann Fam Med 2023 Jan 23;21(1):57-69 [FREE Full text] [doi: 10.1370/afm.2908] [Medline: 36690490]
20. Kim SY, Kim DH, Kim MJ, Ko HJ, Jeong OR. XAI-based clinical decision support systems: a systematic review. Appl Sci 2024 Jul 30;14(15):6638. [doi: 10.3390/app14156638]
21. Sirajuddin AM, Osheroff JA, Sittig DF, Chuo J, Velasco F, Collins DA. Implementation pearls from a new guidebook on improving medication use and outcomes with clinical decision support. Effective CDS is essential for addressing healthcare performance improvement imperatives. J Healthc Inf Manag 2009;23(4):38-45 [FREE Full text] [Medline: 19894486]
22. Kilsdonk E, Peute LW, Jaspers MW. Factors influencing implementation success of guideline-based clinical decision support systems: a systematic review and gaps analysis. Int J Med Inform 2017 Feb;98:56-64. [doi: 10.1016/j.ijmedinf.2016.12.001] [Medline: 28034413]
23. Van de Velde S, Heselmans A, Delvaux N, Brandt L, Marco-Ruiz L, Spitaels D, et al. A systematic review of trials evaluating success factors of interventions with computerised clinical decision support. Implement Sci 2018 Aug 20;13(1):114 [FREE Full text] [doi: 10.1186/s13012-018-0790-1] [Medline: 30126421]
24. Martínez-García A, Moreno-Conde A, Jódar-Sánchez F, Leal S, Parra C. Sharing clinical decisions for multimorbidity case management using social network and open-source tools. J Biomed Inform 2013 Dec;46(6):977-984 [FREE Full text] [doi: 10.1016/j.jbi.2013.06.007] [Medline: 23806275]
25. Fraccaro P, Arguello Casteleiro M, Ainsworth J, Buchan I. Adoption of clinical decision support in multimorbidity: a systematic review. JMIR Med Inform 2015 Jan 07;3(1):e4 [FREE Full text] [doi: 10.2196/medinform.3503] [Medline: 25785897]
26. Fahmi A, Wong D, Walker L, Buchan I, Pirmohamed M, Sharma A, et al. Combinations of medicines in patients with polypharmacy aged 65-100 in primary care: large variability in risks of adverse drug related and emergency hospital admissions. PLoS One 2023;18(2):e0281466 [FREE Full text] [doi: 10.1371/journal.pone.0281466] [Medline: 36753492]
27. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
28. MAGUIRE M. Methods to support human-centred design. Int J Hum Comput Stud 2001 Oct;55(4):587-634. [doi: 10.1006/ijhc.2001.0503]
29. van Staa T, Sharma A, Palin V, Fahmi A, Cant H, Zhong X, et al. Knowledge support for optimising antibiotic prescribing for common infections in general practices: evaluation of the effectiveness of periodic feedback, decision support during consultations and peer comparisons in a cluster randomised trial (BRIT2) - study protocol. BMJ Open 2023 Aug 22;13(8):e076296 [FREE Full text] [doi: 10.1136/bmjopen-2023-076296] [Medline: 37607793]
30. Yuan MJ, Finley GM, Long J, Mills C, Johnson RK. Evaluation of user interface and workflow design of a bedside nursing clinical decision support system. Interact J Med Res 2013 Jan 31;2(1):e4 [FREE Full text] [doi: 10.2196/ijmr.2402] [Medline: 23612350]
31. Blanes-Selva V, Asensio-Cuesta S, Doñate-Martínez A, Pereira Mesquita F, García-Gómez JM. User-centred design of a clinical decision support system for palliative care: insights from healthcare professionals. Digit Health 2023;9:20552076221150735 [FREE Full text] [doi: 10.1177/20552076221150735] [Medline: 36644661]
32. Larsen K, Akindele B, Head H, Evans R, Mehta P, Hlatky Q, et al. Developing a user-centered digital clinical decision support app for evidence-based medication recommendations for type 2 diabetes mellitus: prototype user testing and validation study. JMIR Hum Factors 2022 Jan 18;9(1):e33470 [FREE Full text] [doi: 10.2196/33470] [Medline: 34784293]
33. Hurley R, Jury F, van Staa TP, Palin V, Armitage CJ. Clinician acceptability of an antibiotic prescribing knowledge support system for primary care: a mixed-method evaluation of features and context. BMC Health Serv Res 2023 Apr 14;23(1):367 [FREE Full text] [doi: 10.1186/s12913-023-09239-4] [Medline: 37060063]
34. Rydberg EM, Insulan J, Rolfson O, Mohaddes M, Ahlstrom L. Knowledge support for ankle fractures in the Swedish Fracture Register - a qualitative study of physicians' experiences. BMC Health Serv Res 2022 Mar 23;22(1):382 [FREE Full text] [doi: 10.1186/s12913-022-07799-5] [Medline: 35321701]
35. Damschroder LJ, Reardon CM, Widerquist MA, Lowery J. The updated Consolidated Framework for implementation research based on user feedback. Implement Sci 2022 Oct 29;17(1):75 [FREE Full text] [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
COREQ: Consolidated Criteria for Reporting Qualitative Research
DynAIRx: artificial intelligence for dynamic prescribing optimization and care integration in multimorbidity
EPR: electronic patient record
FG: focus group
GP: general practitioner
HCP: health care professional
LLM: large language model
ML: machine learning
NICE: National Institute for Health and Care Excellence
PCN: primary care network
SMR: structured medication review

Edited by J Sarvestan; submitted 25.Mar.2025; peer-reviewed by O Kehinde, T Ojo, MJ Ugbor, E Oluwagbade; comments to author 07.May.2025; revised version received 10.Jun.2025; accepted 11.Jun.2025; published 08.Jan.2026.

Please cite as:
Abuzour AS, Wilson SA, Woodall AA, Mair FS, Aslam A, Clegg A, Shantsila E, Gabbay M, Abaho M, Bollegala D, Cant H, Griffiths A, Hama L, Leeming G, Lo E, Maskell S, O'Connell M, Popoola O, Relton S, Ruddle RA, Schofield P, Sperrin M, Van Staa T, Buchan I, Walker LE
Developing an AI-Assisted Tool That Identifies Patients With Multimorbidity and Complex Polypharmacy to Improve the Process of Medication Reviews: Qualitative Interview and Focus Group Study
J Med Internet Res 2026;28:e74304
URL: https://www.jmir.org/2026/1/e74304
doi:10.2196/74304
PMID:

©Aseel S Abuzour, Samantha A Wilson, Alan A Woodall, Frances S Mair, Asra Aslam, Andrew Clegg, Eduard Shantsila, Mark Gabbay, Michael Abaho, Danushka Bollegala, Harriet Cant, Alan Griffiths, Layik Hama, Gary Leeming, Emma Lo, Simon Maskell, Maurice O'Connell, Olusegun Popoola, Sam Relton, Roy A Ruddle, Pieta Schofield, Matthew Sperrin, Tjeerd Van Staa, Iain Buchan, Lauren E Walker. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating and Validating Large Language Models for Health Education on Developmental Dysplasia of the Hip: 2-Phase Study With Expert Ratings and a Pilot Randomized Controlled Trial

Hui Ouyang[1,2*], MSN; Gan Lin[1,2*], MSN; Yiyuan Li[1,2], MSN; Zhixin Yao[1,2], MSN; Yating Li[1], MSN; Han Yan[1], BSN; Fang Qin[2*], PhD; Jinghui Yao[3*], PhD; Yun Chen[1,2*], MM

[1]Third Affiliated Hospital of Southern Medical University, Guangzhou, China

[2]School of Nursing, Southern Medical University, Guangzhou, China

[3]Department of Pediatric Orthopedics, Center for Orthopaedic Surgery, Third Affiliated Hospital of Southern Medical University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Yun Chen, MM
Third Affiliated Hospital of Southern Medical University
183 Zhongshan Road West
Tianhe District
Guangzhou, 510630
China
Phone: 86 13724869220
Email: chenyun88@smu.edu.cn

Abstract

Background: Developmental dysplasia of the hip (DDH) is a common pediatric orthopedic disease, and health education is vital to disease management and rehabilitation. The emergence of large language models (LLMs) has provided new opportunities for health education. However, the effectiveness and applicability of LLMs in education with DDH have not been systematically evaluated.

Objective: This study conducted an integrated 2-phase evaluation to assess the quality and educational effectiveness of LLM-generated educational materials.

Methods: This study comprised 2 phases. Based on Bloom's taxonomy, a 16-item DDH question bank was created through literature analysis and collaboration. Four LLMs (ChatGPT-4 [OpenAI], DeepSeek-V3, Gemini 2.0 Flash [Google], and Copilot [Microsoft Corp]) were questioned using standardized prompts. All responses were independently evaluated by 5 pediatric orthopedic experts using 5-point Likert measures of accuracy, fluency, and richness, the scales of Patient Education Materials Assessment Tool for Printable Materials, and DISCERN. The readability was measured by a formula. The data were examined using Kruskal-Wallis tests, ANOVA, and post hoc comparisons. In phase 2, an assessor-blinded, 2-arm pilot randomized controlled trial was conducted. A total of 127 caregivers were randomized into an LLM-assisted education group or a web search control group. The intervention included structured LLM training, supervised practice, and 2 weeks of reinforcement training. Measured at baseline, postintervention, and 2 weeks following, the outcomes were eHealth literacy (primary), DDH knowledge, health risk perception, perceived usefulness, information self-efficacy, and health information-seeking behavior. Cohen *d* effect sizes and linear mixed-effects models were used in an intention-to-treat manner.

Results: There were significant differences between the 4 LLMs concerning accuracy, richness, fluency, Patient Education Materials Assessment Tool for Printable Materials Understandability, and DISCERN (P<.05). ChatGPT-4 (median 63.67, IQR 63.67-64.67) and DeepSeek-V3 (median 63.67, IQR 63.33-64.67) generate more accurate text than Copilot (median 59.00, IQR 58.67-59.67). DeepSeek-V3 (median 64.00, IQR 64.00-64.00) was language richer than Copilot (median 52.33, IQR 51.33-52.67). Gemini 2.0 Flash (median 72.67, IQR 72.33-73.00) was more fluent than Copilot (median 65.67, IQR 63.33-65.67). In phase 2, the intervention group showed higher eHealth literacy at T1 (33.62, 95% CI 32.76-34.49; d=0.20, 95% CI 0.13-0.56) and T2 (33.27, 95% CI 32.38-34.17; d=0.36, 95% CI 0.01-0.80), greater DDH knowledge at T1 (7.87, 95% CI 7.48-8.25, d=0.71, 95% CI 0.33-1.11) and T2 (7.12, 95% CI 6.72-7.51; d=0.54, 95% CI 0.17-0.96), and slight improvements in health risk prediction and perceived usefulness.

XSL•FO
RenderX

Conclusions: Mainstream LLMs demonstrate varying capacities in generating educational content for DDH. They generated DDH caregiver education materials that were associated with modest improvements in eHealth literacy and knowledge. Although LLMs can address general informational needs, they cannot completely substitute clinical evaluation. Future research should focus on optimizing plain language, refining dialogue design, and enhancing audience personalization to improve the quality of LLMs' materials.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2500108410; https://www.chictr.org.cn/showproj.html?proj=271987

(J Med Internet Res 2026;28:e73326) doi:10.2196/73326

KEYWORDS

large language models; developmental dysplasia of the hip; health education; content generation; mobile phone

Introduction

Background

Developmental dysplasia of the hip (DDH) is a common pediatric orthopedic condition affecting 1%-3% of infants, with a higher prevalence in girls and more frequent involvement of the left hip [1]. If undiagnosed or untreated early, DDH can lead to gait abnormalities, chronic pain, and early osteoarthritis, substantially affecting the quality of life [2]. Early diagnosis and health education are critical for improving prognosis. Delayed diagnosis and treatment often require complex surgery, which not only increases the difficulty of treatment but may also result in further functional deterioration [3,4]. Traditional educational methods are limited by time and resources, making it difficult to meet patients' diverse informational needs. The emergence of artificial intelligence (AI) has provided new opportunities for health education.

In the broader field of digital health communication, AI-based conversational systems are increasingly being explored as tools to provide convenient and efficient services to meet people's diverse needs. Currently, large language models (LLMs), such as ChatGPT, Google Gemini, Microsoft Copilot, and DeepSeek, are applied in health communication, including disease diagnosis [5], treatment recommendation [6], health education [7], and clinical decision-making [8]. For example, ChatGPT enables interactive discussions that tailor standardized medical information to individual patient needs, helping bridge communication gaps between clinicians and patients [9,10]. Although AI has demonstrated great potential in medical education, its use in patient-facing communication raises concerns. LLMs may provide erroneous medical advice [11], propagate outdated medical views [12], or fabricate nonexistent medical cases to generate "hallucinations" [13]. At the ethical and regulatory levels, challenges arise from the model's "black box" decision-making, including unclear accountability, difficulties in defining legal responsibility, privacy breaches, and lagging regulatory frameworks. These issues directly jeopardize users' safety, potentially leading to misdiagnosis, delayed treatment, and other forms of direct harm. Furthermore, most generated content maintains a university reading level, which may pose comprehension challenges for users without higher education [14]. These risks underscore the need for systematic evaluation before integrating such tools into health education.

While prior studies have primarily examined the accuracy or readability of LLM-generated content [15], few have connected content quality with its actual educational impact on end users. The extent to which LLM-generated materials can effectively support caregivers' understanding and health literacy in specific conditions, such as DDH, remains unexplored. In DDH, caregivers have to not only comprehend specialized medical concepts but also actively recognize abnormal signs in children and make timely decisions [16]. Due to the professional complexity of orthopedic knowledge and the unique nature of pediatric disorders, basic health literacy abilities are necessary for caregivers. Therefore, the different levels of digital literacy among caregivers may make it more difficult for them to properly understand information produced by LLMs. To address this, the present study systematically evaluated multiple mainstream LLMs through expert assessment and a pilot randomized controlled trial (RCT) among caregivers. By integrating expert evaluation with caregiver validation, this study extends the current AI in health communication research from theoretical assessment to empirical verification.

Objective

Therefore, this study aimed to provide a comprehensive evaluation and verification of LLM-generated education materials for DDH. The first phase assessed the educational quality of the outputs generated by 4 mainstream LLMs (ChatGPT-4, DeepSeek-V3, Gemini 2.0 Flash, and Copilot) through expert ratings of accuracy, understandability, actionability, and readability. The second phase involved a pilot RCT among caregivers to evaluate the actual educational impact of these materials, including digital literacy, DDH knowledge acquisition, health risk perception, information self-efficacy, perceived usefulness, and health information-seeking behaviors. This study bridged the gap by integrating the quality assessment of LLMs with RCT to validate their content reliability and educational impact. It offered evidence for the safe and effective use of LLMs in clinical education.

Methods

Theoretical Framework

The taxonomy by Bloom et al [17] served as the guiding pedagogical framework for designing the educational content. The taxonomy organizes cognitive processes into 6 hierarchical levels: remember, understand, apply, analyze, evaluate, and create, which is widely used to structure learning objectives and instructional materials. As users often need to acquire not only

XSL•FO
RenderX

basic factual knowledge but also practical decision-making skills, this hierarchical model provided a structured approach for determining which levels of cognition should be targeted in education [18].

Guided by this framework, we developed a 16-item question bank that intentionally spanned different cognitive levels, ranging from foundational knowledge such as definitions and symptoms to more complex tasks such as interpreting clinical scenarios or making care decisions (Table 1). This ensured that the LLM-generated responses covered the breadth of learning needs relevant to caregivers. Bloom's taxonomy, therefore, supported the construction of a comprehensive and pedagogically meaningful learning set, helping align the generated content with education requirements.

Table 1. Question bank framework based on Bloom's taxonomy.

Bloom's taxonomy and part	Content
Remembering	
Basics	"What is [disease name]? Please explain its main features and potential effects in simple language."
Anatomy and effects	"Who is most likely to develop [disease name]? How common is it?"
Understanding	
Etiology and risk factors	"Will [disease name] have a long-term impact on my quality of life? What aspects should I pay special attention to?"
Symptoms and early recognition	"What parts of the body does [disease name] mainly affect? What do these parts do in a healthy state?"
Applying	
Physician examination and diagnosis	"What symptoms or consequences may arise if these parts are damaged? Can you provide specific examples?"
Emergencies	"What are the main causes of [disease name]? What risk factors may increase the chances of developing it?"
Analyzing	
Hospital treatment and rehabilitation	"What lifestyle habits or environmental factors increase the risk of developing [disease name]? How can I prevent it?"
Medication management	"If I have a family history, is my risk of developing [disease name] higher? What preventive measures should I take?"
Pain management	"What are the main symptoms of [disease name]? What signs indicate the disease is worsening? Should I seek immediate medical attention?"
Evaluating	
Postoperative management	"What methods do doctors usually use to confirm the diagnosis of [disease name]? What is the purpose of each test?"
Creating	
Daily living and health management	"What indicators are most important in test results? How do these results reflect the severity of the disease?"
Mental health	"After confirming the diagnosis of [disease name], do I need to schedule regular follow-ups? How often should I have them and what should be checked?"
Digital tools and management	"What symptoms indicate that I need to seek immediate medical attention?"
Expanding[a]	
Education and support resources	"How should I handle and self-monitor during an emergency situation?"
Social and economic impact	"What treatments will I receive during hospitalization? What is the purpose of each treatment?"
Future planning and visioning	"How can I prevent infections or other complications during hospitalization? What can family members do to assist with recovery?"

[a]Not part of Bloom's taxonomy; it is an extension of this study.

Study Design

The evaluation study had 2 phases. Phase 1 was a cross-sectional study in which physicians evaluated the answers provided by the LLMs. Phase 2 was a 2-arm pilot RCT comparing health education using LLMs with web-based searches.

Phase 1: Expert Evaluation Study

Model Testing

Based on Bloom's taxonomy, we collected and categorized common questions regarding DDH. We also reviewed clinical guidelines [19-23] to identify the key areas of knowledge. Using this information, an initial question bank was developed, which

XSL•FO
RenderX

was subsequently refined and finalized through expert review. Each question was guided by a harmonized prompt paragraph: "Using developmental dysplasia of the hip (DDH) in children as an example, answer the following questions in detail, ensuring the content is easily understandable for non-medical professionals. Life-like examples and situations can be incorporated to help readers better grasp the information. Please reduce the number of syllables to make the sentence simpler." All the generated texts and the complete question bank are provided in Multimedia Appendix 1. Each model generated educational materials based on 16 question banks, and the experiment was repeated 3 times for a total of 192 generations (4 models × 16 question banks × 3 times). Data were collected from January to February 2025. ChatGPT-4, DeepSeek-V3, Gemini 2.0 Flash, and Copilot were evaluated; no experimental, beta, or preview releases were included. The experiments were performed under the default settings of the web interfaces without modifying the generation parameters. To ensure reproducibility and independence of the outputs, each prompt was regenerated 3 times by establishing a new session for each run with the same original prompt. All outputs, including identical or similar responses, were retained to reflect the intrinsic variability of the models.

Assessment of Quality and Readability

Quality assessment tools as primary outcomes included (1) a Likert scale for assessing three items of accuracy, fluency, and richness of the material, scoring 16 questions on a scale from 1 to 5, with higher scores indicating better performance; (2) the DISCERN tool [24], which assessed the overall quality of the educational material, with a total of 16 entries, scoring on a scale from 1 to 5, with higher scores indicating better quality; and (3) the Patient Education Materials Assessment Tool for Printable Materials (PEMAT-P) [25], which contains 17 items measuring understandability and 7 items assessing actionability. These were reduced to 10 and 4 to accommodate the textual output with a 70% passing line based on the guidelines. During the evaluation process, each material was independently scored by 5 evaluators. The scores from evaluators were retained for subsequent data analysis.

Readability assessment tools as secondary outcomes included (1) the Flesch-Kincaid Reading Ease (FKRE), (2) the Flesch-Kincaid Grade Level (FKGL), and (3) the Simple Measure of Gobbledygook (SMOG) index, chosen for their widespread use and reliability in assessing text readability. All 3 score calculations involved the total number of words, sentences, and syllables. The FKRE measured the simplicity of the text, with scores ranging from 0 to 100, with higher scores indicating better readability. The FKGL represented reading level grade, with lower FKGL and SMOG indicating better comprehension and higher scores indicating more complex language. Scores above 60 or below sixth grade were the recommended reading levels for the general public. Readability scores were calculated using a web-based readability calculator (Readable; Added Bytes). The detailed formulas are provided in Multimedia Appendix 2.

Expert Evaluation

The material generated by the LLMs was independently assessed for quality. The material generated by the LLMs was independently assessed by 5 pediatric orthopedic physicians with expertise in DDH, selected through rigorous predefined criteria: (1) ≥10 years of clinical experience in DDH diagnosis or treatment; and (2) evaluators completed standardized training on the assessment rubric before this study, using the DDH guidelines as the gold standard [19-23]. To ensure blinding, the LLM outputs were made anonymous by an independent researcher who replaced the model names with random codes. The evaluators confirmed that they could not infer the identities of the LLMs or determine if repeated outputs came from the same model. Interrater reliability was assessed for each outcome dimension using the intraclass correlation coefficient (ICC). ICC values were interpreted as follows: <0.5=poor, 0.5-0.75=moderate, 0.75-0.9=good, and >0.9=excellent agreement.

Phase 2: Pilot RCT

Participants

Participants were recruited through digital media advertisements and physician referrals. Eligibility criteria included (1) being aged ≥18 years, (2) being caregivers of children aged 0-14 years, (3) having the ability to read and understand words, and (4) having internet access. Exclusion criteria included (1) having severe hearing or visual impairment; (2) having severe schizophrenia, major depression, bipolar disorder, and other mental illnesses; or (3) participation in other related studies.

Sample Size

Power analysis was performed using G*Power 3.1.9.7 based on similar educational intervention studies [26]. A medium effect size (Cohen d=0.65) was anticipated for the primary outcome of performance, with 2-tailed α=.05, power of 0.8, and at least 38 participants per group being required. Accounting for an expected 20% attrition rate, the target sample was 49 participants per group (total n=98). There were 127 participants in the final sample (62 in the control group and 65 in the intervention group).

Randomization and Blinding

Recruitment took place in the Third Affiliated Hospital of Southern Medical University and community support groups. The researchers generated a computer-generated list and sealed it in an opaque envelope. Before the start of the intervention, research assistants who were not involved in hospital assessments or interventions opened the envelopes and assigned participants at random to the intervention or control group. Following informed consent, eligible participants meeting the inclusion and exclusion criteria were randomly assigned in a 1:1 ratio to either the trial or control group. The blinding of participants was not feasible due to the nature of the intervention, but the research team remained unaware of group assignments until this study concluded. Data analysts who conducted the final analyses were masked to participant identities throughout. Due to the nature of the intervention, participant blinding was not possible. However, group allocations were not disclosed to the research team until the

XSL•FO
RenderX

trial was finished. Throughout, participant identities were concealed from the data analysts.

Control Group

Participants in the control group received standard web-based educational materials prepared by clinical experts. These materials were retrieved from official sources (eg, [27]). Participants were asked to read independently, simulating a typical web-based health information-seeking behavior.

Intervention Group

All researchers received standardized training to ensure consistent delivery of DDH-related information and LLM education. The intervention was delivered to participants by face-to-face communication. First, the participants were introduced to the foundational concepts of LLMs, including basic mechanisms, application categories, and core interaction capabilities. Second, a standardized consultation framework was introduced, covering device access, platform login, dialogue initiation, and structured prompt formulation. The required background information included demographic and clinical characteristics, symptom description, disease duration, medical history, lifestyle, and psychosocial factors. Participants were also provided with 16 DDH-related inquiry categories, including foundations, risk factors, early recognition, diagnosis, treatment, postoperative care, medication management, psychological support, etc. They are able to optionally output custom instructions, such as length, style, level of technical terminology, and formatting preferences. Third, strategies to improve information quality are introduced, including clear language prompts, staged questions, example guidance, support for the reasoning process, evidence sources for web retrieval, and document import. Verification approaches were emphasized, such as cross-model comparison, guideline checking, and professional consultation. Finally, risk awareness and ethical considerations were reinforced, including potential hallucinations, outdated content, privacy risks, copyright issues, and inappropriate clinical dependence. A practical demonstration was conducted using an actual DDH case. For example, a female infant, 1 year old, with asymmetric thigh folds and a family history, but no medical history. Participants inquired and learned relevant knowledge based on the background of this example. During the 2 weeks, the participants received remote support through web-based group consultations or offline feedback sessions. Researchers responded to questions related to practical application, corrected misuse behaviors, and supplemented individualized guidance.

Data Collection

Data were collected through questionnaire surveys. The basic information questionnaire gathered the demographic characteristics of this study's participants. Validated scales were used to measure eHealth literacy, DDH knowledge, health risk perception, information self-efficacy, perceived usefulness, and health information-seeking behavior. There were three assessment time points: (1) baseline (T0), (2) immediately after the completion of the intervention or control group (T1), and (3) two weeks after the end of the intervention or control group (T2).

Primary Outcomes

The eHealth Literacy Scale (eHEALS), originally developed by Norman and Skinner [28], was adopted to measure participants' eHealth literacy. It comprises 8 items that assess one's ability to locate and use web-based health resources, appraise the credibility of digital health information, and apply acquired information to make informed health decisions. Each item is scored on a 5-point Likert scale, producing a total score between 8 and 40, with higher scores representing stronger eHealth literacy.

Secondary Outcomes

The developmental dysplasia of the hip knowledge test (DDH-KT) was developed by the research team to assess participants' basic DDH knowledge. The items were constructed according to current clinical guidelines and health education materials and reviewed by pediatric orthopedic surgeons. Each correct answer is scored as 1 point (range 0-10), with higher scores indicating greater DDH knowledge. The full knowledge test is provided in Multimedia Appendix 3.

The Health Risk Perception Scale (HRPS) was measured based on the framework by Ajzen [29]. The scale was adapted from established health risk perception measures by Brewer et al [30], and covered 2 dimensions: perceived susceptibility and perceived severity. The items assessed participants' subjective perception of the likelihood and potential consequences of related health problems, rated on a 5-point Likert scale. Higher scores reflected a greater level of perceived risk (Cronbach α=0.847).

The Information Self-Efficacy Scale (ISES), adapted from Pavlou and Fygenson [31], was used to evaluate participants' confidence in obtaining and effectively using web-based health information. The scale contained 3 items rated on a 5-point Likert scale. Total scores were calculated by summing all item responses, with higher scores indicating stronger information self-efficacy (Cronbach α=0.806).

The Perceived Usefulness Scale (PUS), adapted from Cheung et al [32], assessed the extent to which participants viewed web-based health information as helpful, relevant, and beneficial for health knowledge and decision-making. Items were scored on a 5-point Likert scale, with higher scores indicating greater perceived usefulness (Cronbach α=0.852).

The Health Information-Seeking Behavior Scale (HISBS), adapted from Kankanhalli et al [33], measured the frequency and willingness to actively seek web-based health information. Responses were recorded using a 5-point Likert scale, and higher scores indicated more proactive seeking behavior (Cronbach α=0.873).

Statistical Analysis

In phase 1, descriptive statistics were reported as mean (SD) and median (IQR). Because the final analytic values were obtained by averaging 3 generations, the normality assumptions for repeated-measures ANOVA were not met. Group differences among the 4 LLMs were analyzed using the Kruskal-Wallis H test, followed by Dunn-Bonferroni post hoc comparisons when significant. One-way ANOVA and Tukey post hoc tests were

XSL•FO
RenderX

used for readability indices because the normality assumptions were satisfied. False discovery rate correction was applied across the 9 outcomes to control for multiple testing. Interrater reliability was assessed using ICC(2,k) based on a 2-way random-effects model [34]. Effect sizes were reported as epsilon-squared for nonparametric tests and eta-squared for ANOVA. Analyses were conducted in R (version 4.5.1; R Foundation) with ggplot2 (version 3.5.1; Posit, PBC) for visualization.

In phase 2, all analyses followed the intention-to-treat principle and included all randomized participants. Continuous baseline variables are presented as mean (SD), and categorical variables as counts and percentages. Differences between groups at baseline were assessed using 2-sided independent sample *t* tests for continuous variables and chi-square tests for categorical variables. Outcomes were analyzed using linear mixed-effects models with time (T1 and T2) and group (intervention vs control) as fixed effects, time × group interaction, baseline (T0) as a covariate, and participant ID as a random intercept. No imputation was performed because linear mixed-effects models estimated with restricted maximum likelihood provided unbiased estimates under the missing at random assumption [35]. Between-group effect sizes (Cohen *d*, 95% CI) and estimated marginal means (95% CI) were reported. eHEALS was defined as the primary outcome. All other outcomes, including DDH-KT, HRPS, ISES, PUS, and HISBS, were considered secondary. Given the pilot and exploratory nature of this trial, no adjustment for multiple comparisons was applied. Therefore, analyses of the outcomes were intended to be hypothesis-generating rather than confirmatory. Analyses were conducted in R (version 4.5.1) using lme4, lmerTest, and emmeans; 2-sided $P<.05$ was considered statistically significant.

Ethical Considerations

This study was approved by the Ethics Committee of the Third Affiliated Hospital of Southern Medical University (2024-ER-113), and the first participant was enrolled in June 2025. The trial registration was completed on August 29, 2025, at the Chinese Clinical Trial Registry (ChiCTR2500108410). All research participants signed written informed consent forms. Researchers disclosed study information to participants; participants retained the right to withdraw from the study or withdraw their research data at any time without conditions, and withdrawal would not result in any adverse consequences. Participants were informed that part of the educational content was generated by AI, and the limitations of AI-generated information were explained. The use of AI-assisted materials was supervised throughout this study by qualified health care professionals. During the intervention period, participants were encouraged to report any concerns or adverse experiences related to the educational materials, and ultimately, no related adverse events were reported. All personal information and data collected during the study were kept strictly confidential. Participants who completed the entire study process received educational materials, including a parenting knowledge handbook valued at CN ¥50 RMB (approximately US $7.15), as compensation.

Results

Phase 1

Overview

Overall, ChatGPT-4 and DeepSeek-V3 demonstrated the strongest performance in content accuracy, richness, understandability, and information quality, making them suitable for generating pediatric health communication materials. Gemini 2.0 Flash and Copilot performed well in fluency and readability metrics, while they were relatively weaker in content richness and accuracy. Table 2 provides a visual summary of the scores and the overall performance comparison. Figure 1 illustrates the comparison of the responses across the LLMs. The scoring data are presented in Multimedia Appendix 4.

XSL•FO
RenderX

Table 2. Comparison of model performance across different indicators.

Model	Mean (SD)	Median (IQR)	H/F[a]	*P* value	FDR-adjusted[b] *P* value	ε^2[c]/η^2[d]	Significance (*P* value)
Accuracy			13.873	.003	.005	0.73	
ChatGPT-4	64 (1.03)	63.67 (63.67-64.67)					*[e,f] (.02)
Copilot	59.07 (1.01)	59.00 (58.67-59.67)					*[g] (.048)
DeepSeek-V3	63.33 (1.78)	63.67 (63.33-64.67)					__[h]
Gemini 2.0 Flash	59.53 (0.99)	59.67 (59.33-60.00)					—
Richness			13.68	.003	.005	0.72	
ChatGPT-4	62.33 (2.26)	62.67 (60.00-64.33)					—
Copilot	51.6 (3.52)	52.33 (51.33-52.67)					*[g] (.02)
DeepSeek-V3	63.93 (1.53)	64.00 (64.00-64.00)					—
Gemini 2.0 Flash	54 (5.65)	54.67 (50.33-57.33)					—
Fluency			16.204	.001	.003	0.853	
ChatGPT-4	69.53 (1.19)	70.00 (68.67-70.00)					—
Copilot	64.87 (1.8)	65.67 (63.33-65.67)					***[i] (<.001)
DeepSeek-V3	69.87 (2.19)	70.67 (70.33-71.00)					—
Gemini 2.0 Flash	73 (0.97)	72.67 (72.33-73.00)					—
PEMAT-P[j] understandability (%)			11.421	.01	.012	0.601	
ChatGPT-4	93.89 (1.24)	94.44 (94.44-94.44)					*[f] (.03)
Copilot	85 (4.21)	86.11 (80.56-88.89)					—
DeepSeek-V3	93.33 (3.17)	94.44 (91.67-94.44)					—
Gemini 2.0 Flash	87.78 (4.21)	86.11 (86.11-88.89)					—
PEMAT-P actionability (%)			7.587	.06	.06	0.399	
ChatGPT-4	68.33 (3.73)	66.67 (66.67-66.67)					—
Copilot	60 (6.97)	58.33 (58.33-66.67)					—
DeepSeek-V3	68.33 (3.73)	66.67 (66.67-66.67)					—
Gemini 2.0 Flash	66.67 (5.89)	66.67 (66.67-66.67)					—
DISCERN			10.243	.02	.02	0.539	
ChatGPT-4	48.52 (3.27)	49.00 (46.00-49.27)					*[i] (.035)
Copilot	46.56 (1.57)	46.67 (46.47-47.67)					—
DeepSeek-V3	48.44 (2.71)	48.00 (46.67-49.20)					*[i] (.03)
Gemini 2.0 Flash	43.08 (1.82)	43.33 (42.33-43.40)					—
FKGL[k]			8.395	<.001	.003	0.296	
ChatGPT-4	8.74 (1.37)	8.86 (8.14-9.70)					*[g] (.03); *[i] (.04)
Copilot	9.41 (1.86)	9.07 (8.38-10.88)					***[g] (<.001); ***[i] (<.001)
DeepSeek-V3	7.30 (1.08)	7.26 (6.72-7.70)					—
Gemini 2.0 Flash	7.37 (1.33)	7.19 (6.30-8.44)					—
FKRE[l]			14.198	.003	.005	0.225	
ChatGPT-4	61.86 (8.58)	61.44 (56.80-66.97)					—

XSL•FO
RenderX

Model	Mean (SD)	Median (IQR)	H/F[a]	*P* value	FDR-adjusted[b] *P* value	ε^{2}[c]/η^{2}[d]	Significance (*P* value)
Copilot	53.45 (12.62)	57.70 (46.25-61.33)					**[g] (.006); **[i] (.009)
DeepSeek-V3	67.19 (6.62)	67.45 (62.73-70.43)					—
Gemini 2.0 Flash	66.85 (7.72)	70.10 (59.19-73.48)					—
SMOG[m]			8.297	<.001	.003	0.293	
ChatGPT-4	11.02 (1.26)	10.96 (10.52-11.61)					*[g] (.02)
Copilot	11.67 (1.33)	11.48 (11.09-13.04)					***[g] (<.001); **[i] (.003)
DeepSeek-V3	9.83 (0.93)	9.74 (9.25-10.19)					—
Gemini 2.0 Flash	10.19 (1.06)	10.03 (9.28-10.91)					—

[a]H/F: values are reported as test statistics. H statistics for the Kruskal-Wallis test and *F* statistics for 1-way ANOVA.

[b]FDR: false discovery rate.

[c]η^{2}: eta-squared.

[d]ε^{2}: epsilon-squared.

[e]**P* <.05. ***P* <.01. ****P* <.001. Normality was assessed for all variables. FKGL and SMOG were analyzed using 1-way ANOVA with Tukey honestly significant difference for pairwise comparisons; others were analyzed using the Kruskal-Wallis test with Dunn test (Bonferroni-corrected *P* values).

[f]vs CoPilot.

[g]vs DeePseek-V3.

[h]Not applicable.

[i]vs Gemini 2.0 Flash.

[j]PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials.

[k]FKGL: Flesch-Kincaid Grade Level.

[l]FKRE: Flesch-Kincaid Reading Ease.

[m]SMOG: Simple Measure of Gobbledygook.

XSL•FO
RenderX

Figure 1. Comparison of responses across 4 LLMs. (A) Accuracy, (B) richness, (C) fluency, (D) PEMAT-P understandability, (E) PEMAT-P actionability, (F) DISCERN, (G) FKGL, (H) FKRE, and (I) SMOG index. FKGL: Flesch-Kincaid Grade Level; FKRE: Flesch-Kincaid Reading Ease; LLM: large language model; PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials; SMOG: Simple Measure of Gobbledygook.

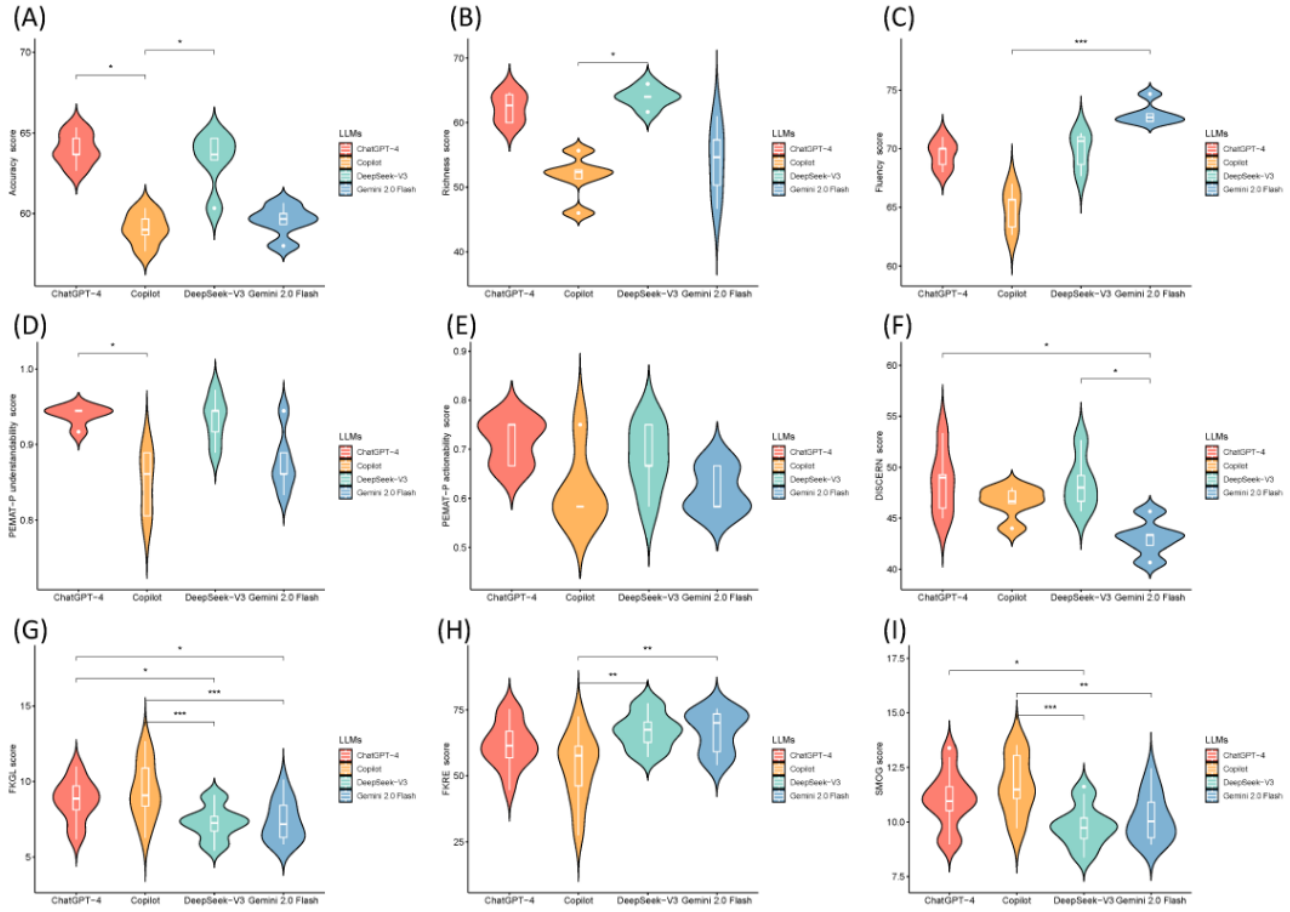

Quality Assessment

There were significant differences between the 4 LLMs in terms of accuracy, richness, fluency, PEMAT-P understandability, and DISCERN (P<.05). ChatGPT-4 and DeepSeek-V3 outperformed the other models in the majority of evaluation dimensions. ChatGPT-4 (median 63.67, IQR 63.67-64.67) and DeepSeek-V3 (median 63.67, IQR 63.33-64.67) generated more accurate text than Copilot (median 59.00, IQR 58.67-59.67). DeepSeek-V3 (median 64.00, IQR 64.00-64.00) was language richer than Copilot (median 52.33, IQR 51.33-52.67). Gemini 2.0 Flash (median 72.67, IQR 72.33-73.00) was more fluent than Copilot (median 65.67, IQR 63.33-65.67). Based on the PEMAT-P understandability scores, the content of ChatGPT-4 (median 94.44%, IQR 94.44%-94.44%) was more comprehensible than that of Copilot (median 86.11%, IQR 80.56%-88.89%). The PEMAT-P actionability scores were similar across the models. ChatGPT-4 (median 49.00, IQR 46.00-49.27) and DeepSeek-V3 (median 48.00, IQR 46.67-49.20) had a higher DISCERN scale score than Gemini 2.0 Flash (median 43.33, IQR 42.33-43.40).

Readability Assessment

Readability metrics highlighted the differences among the models. Gemini 2.0 Flash (median 66.85, IQR 59.19-73.48) and DeepSeek-V3 (median 67.19, IQR 62.73-70.43) generated sentences with higher FKRE scores, indicating easier readability compared to Copilot (median 53.45, IQR 46.25-61.33). DeepSeek-V3 (mean 7.30, SD 1.08) and Gemini 2.0 Flash (mean 7.37, SD 1.33) produced sentences with superior FKGL scores compared to ChatGPT-4 (mean 8.74, SD 1.37) and Copilot (mean 9.41, SD 1.86). DeepSeek-V3 (mean 9.83, SD 0.93) and Gemini 2.0 Flash (mean 10.19, SD 1.06) produced texts with better SMOG scores compared to ChatGPT-4 (mean 11.02, SD 1.26) and Copilot (mean 11.67, SD 1.33).

Visualization and Analysis

The comparative evaluation of 4 LLMs demonstrated clear performance variability across accuracy, richness, and fluency, as illustrated in Figures 2 and 3. Overall, ChatGPT-4 and DeepSeek-V3 outperformed Copilot and Gemini Flash, particularly in accuracy and fluency. In terms of accuracy, the proportion of “good” and “excellent” responses reached 85% for ChatGPT-4 and 83% for DeepSeek-V3, while Gemini 2.0 Flash (70%) and Copilot (66%) displayed a lower proportion. Regarding richness, DeepSeek-V3 (83%) and ChatGPT-4 (81%) again ranked highest, reflecting strong supplementary and explanatory capability, whereas the other 2 models showed as more concise. Across fluency, all 4 models delivered strong information elaboration, with Gemini 2.0 Flash achieving the highest proportion of 96%, indicating strong coherence, readability, and natural language expression.

As shown in the heatmap (Figure 3), ChatGPT-4 and DeepSeek-V3 yielded higher mean scores across most knowledge domains, particularly in basic, effects, and

XSL•FO
RenderX

symptoms. In contrast, Copilot and Gemini 2.0 Flash performed worse, especially in specialized domains such as medication management and postoperative care. These results suggested that current LLMs perform well in general health education content but remain limited in clinically nuanced and actionable information.

Across the 4 models and 6 evaluation dimensions, the interrater reliability among the 5 evaluators ranged from moderate to excellent (ICC=0.628-0.918). Table 3 shows the interrater reliability results across the 4 LLMs and evaluation dimensions based on ICC.

Figure 2. Expert Likert-scale ratings of content quality across 4 LLMs. LLM: large language model.

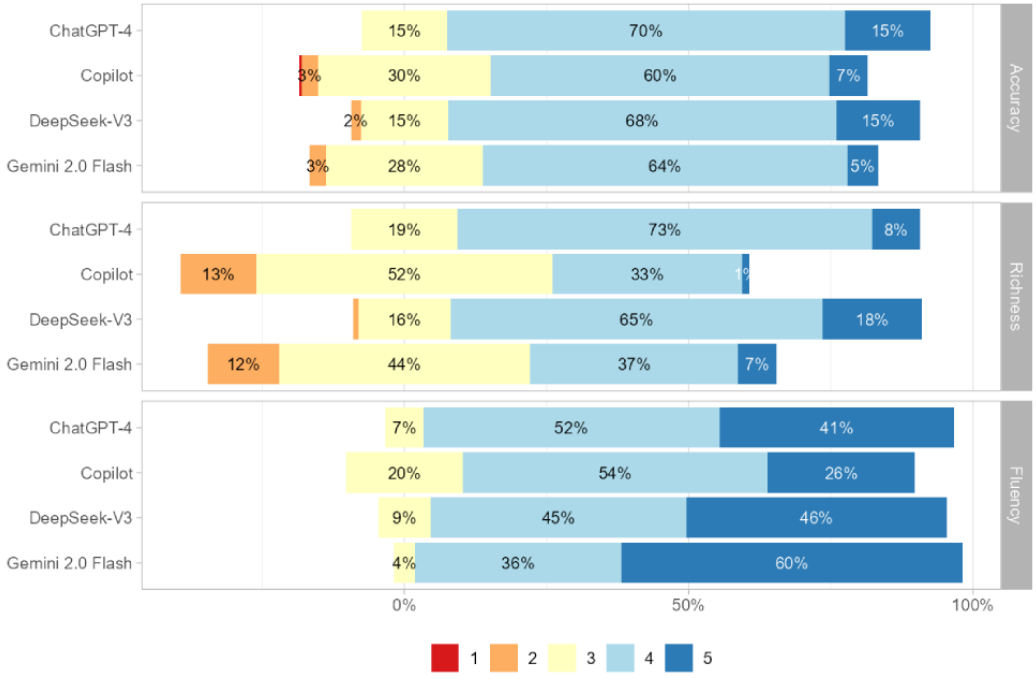

Figure 3. Multidimensional performance evaluation heatmap for LLMs. Heatmap showing mean scores of 4 LLMs across accuracy, richness, fluency, and readability dimensions. Higher scores are represented by warmer colors. FRES: Flesch-Kincaid Reading Ease; LLM: large language model.

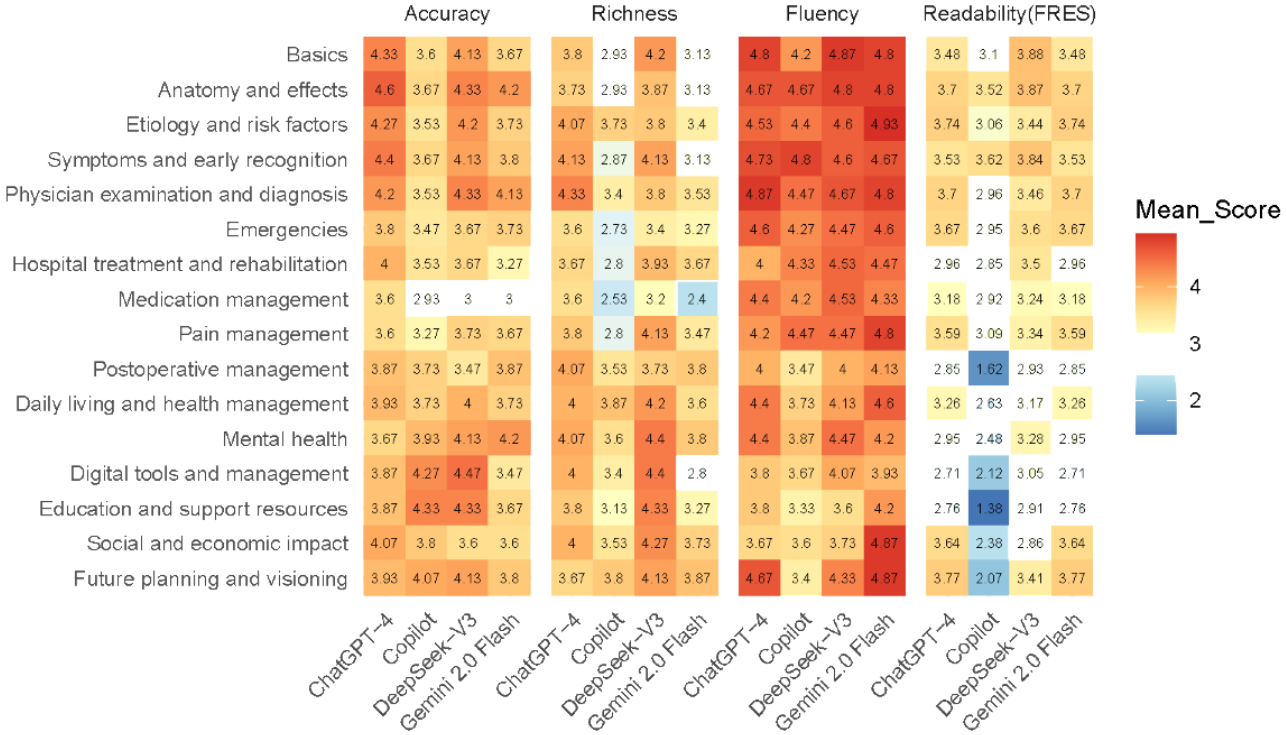

XSL•FO
RenderX

Table 3. Results of expert consistency analysis.

Dimension	ICC[a,b]	95% CI	*F* test (*df*)	*P* value	Interpretation
Accuracy					
ChatGPT-4	0.851	0.695-0.941	6.734 (15, 60)	<.001	Good
Copilot	0.654	0.277-0.864	2.810 (15, 60)	.002	Moderate
DeepSeek-V3	0.904	0.803-0.962	11.510 (15, 60)	<.001	Excellent
Gemini 2.0 Flash	0.727	0.433-0.892	3.565 (15, 60)	<.001	Moderate
Richness					
ChatGPT-4	0.669	0.352-0.864	3.499 (15, 60)	<.001	Moderate
Copilot	0.834	0.652-0.934	7.750 (15, 60)	<.001	Good
DeepSeek-V3	0.747	0.481-0.899	3.944 (15, 60)	.001	Moderate
Gemini 2.0 Flash	0.628	0.285-0.845	3.515 (15, 60)	<.001	Moderate
Fluency					
ChatGPT-4	0.914	0.825-0.966	11.931 (15, 60)	<.001	Excellent
Copilot	0.918	0.833-0.967	13.244 (15, 60)	<.001	Excellent
DeepSeek-V3	0.835	0.664-0.934	6.223 (15, 60)	<.001	Good
Gemini 2.0 Flash	0.876	0.746-0.951	8.038 (15, 60)	<.001	Good
PEMAT-P[c] understandability	0.874	0.482-0.991	7.833 (3, 12)	.004	Good
PEMAT-P actionability	0.718	0.050-0.980	3.858 (3, 12)	.038	Moderate
DISCERN	0.819	0.358-0.986	12.473 (3, 12)	.001	Good

[a]ICC: intraclass correlation coefficient.

[b]Type A intraclass correlation coefficient using an absolute agreement definition.

[c]PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials.

Phase 2

Participant Characteristics

Participants were recruited from June 2025 to September 2025. A total of 127 participants were enrolled in this study, including 65 in the intervention group and 62 in the control group. Figure 4 shows the CONSORT (Consolidated Standards of Reporting Trials) flowchart, and the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist is presented in Multimedia Appendix 5. Most participants completed the intervention, and the main reason for withdrawal was lack of time. Participants had a mean age of 36.57 (SD 6.22) years, and most were female (89/127, 70.07%) and highly educated (55/127, 43.31%). The mean age of participants' children was 5.90 (SD 3.12) years. No significant differences were observed between the intervention and control groups in the baseline characteristics ($P>.05$). During this study, no privacy breaches, technical failures, or other unintended events were observed. Table 4 summarizes the demographic characteristics of the participants. The data of participants can be found in Multimedia Appendix 6.

XSL•FO
RenderX

Figure 4. CONSORT diagram of study flow. CONSORT: Consolidated Standards of Reporting Trials.

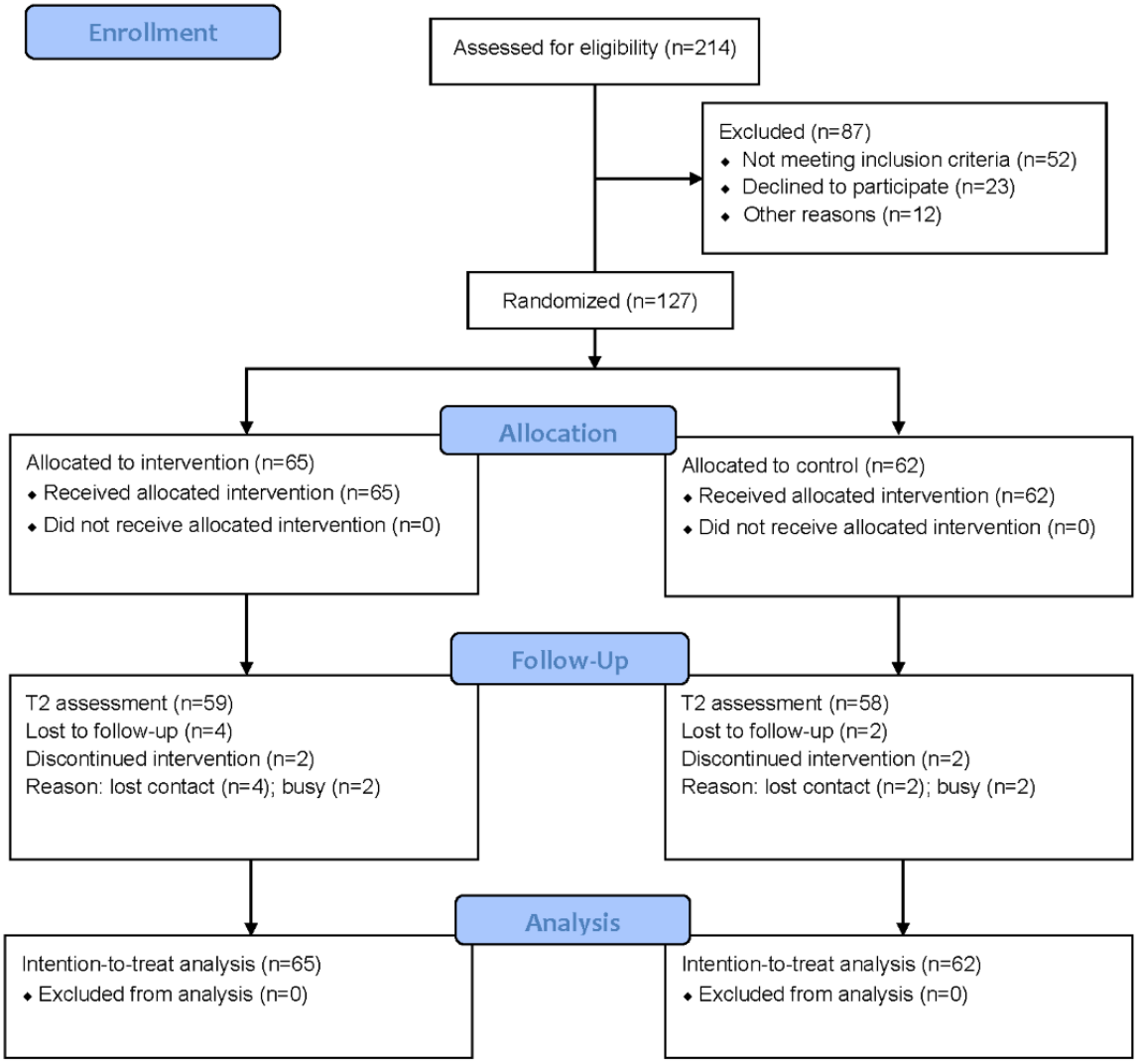

XSL•FO
RenderX

Table 4. Baseline characteristics.

Background characteristic	Overall	Intervention group (n=65)	Control group (n=62)	*P* value
Age (years), mean (SD)	36.57 (6.22)	37.06 (5.68)	36.05 (6.74)	.14
Sex, n (%)				.83
Female	89 (70.08)	45 (69.23)	44 (70.97)	
Male	38 (29.92)	20 (30.77)	18 (29.03)	
Education, n (%)				.92
Low education	18 (14.17)	11 (16.92)	7 (11.29)	
Medium education	54 (42.52)	27 (41.54)	27 (43.55)	
High education	55 (43.31)	27 (41.54)	28 (45.16)	
Monthly income (CN ¥), n (%)				.85
≤3000 (US $429.12)	9 (7.09)	30 (61.2)	35 (71.4)	
3001-6000 (US $429.26-US $858.23)	26 (20.47)	19 (38.8)	14 (28.6)	
6001-10,000 (US $858.38-US $1430.39)	42 (33.07)	21 (33.87)	21 (32.31)	
10,001-20,000 (US $1430.53-US $2860.78)	36 (28.35)	19 (30.65)	17 (26.15)	
≥20,001 (US $2860.92)	14 (11.02)	7 (11.29)	7 (10.77)	
Child's gender, n (%)				.23
Male	71 (55.91)	32 (49.23)	24 (38.71)	
Female	56 (44.09)	33 (50.77)	38 (61.29)	
Child's age (years), mean (SD)	5.90 (3.12)	6.39 (3.36)	5.39 (2.78)	.06
Daily caregiving time for the child (hours/day), n (%)				.37
≤2	17 (13.39)	11 (16.92)	6 (9.68)	
3-6	55 (43.31)	27 (41.54)	28 (45.16)	
6-9	16 (12.6)	7 (10.77)	9 (14.52)	
9-12	15 (11.81)	6 (9.23)	9 (14.52)	
≥12	24 (18.9)	14 (21.54)	10 (16.13)	
Smartphone proficiency, n (%)				.53
Very proficient	66 (51.97)	33 (50.77)	33 (53.23)	
Basic proficient	22 (17.33)	14 (21.54)	8 (12.9)	
Fairly proficient	33 (25.98)	15 (23.08)	18 (29.03)	
Not proficient	6 (4.72)	3 (4.62)	3 (4.84)	

Primary Outcome

The group × time interaction in eHEALS was not significant (*P*=.26). The intervention group showed higher scores than the control group at T1 (33.62, 95% CI 32.76-34.49; *d*=0.20, 95% CI 0.13-0.56) and T2 (33.27, 95% CI 32.38-34.17; *d*=0.36, 95% CI 0.01-0.80), indicating sustained improvements following the LLM-generated learning intervention. Table 5 reports the means estimated from the model and the contrasts between groups across the specified time points; Figure 5 graphically illustrates the outcomes overtime by condition.

XSL•FO
RenderX

Table 5. Change in outcomes.

Time	Control EMM[a] (95% CI)	Intervention EMM (95% CI)	Group difference		Cohen *d* (95% CI)	Group × time interaction, *P* value
			β (95% CI)	SE		
eHEALS[b]						.26
T0	29.23 (28.05 to 30.40)	27.91 (26.24 to 29.57)	—[c]	—	—	—
T1	31.89 (31.01 to 32.77)	33.62 (32.76 to 34.49)	1.73 (0.49 to 2.97)	0.63	0.20 (0.13 to 0.56)	—
T2	30.87 (29.97 to 31.78)	33.27 (32.38 to 34.17)	2.40 (1.13 to 3.67)	0.65	0.36 (0.01 to 0.80)	—
DDH-KT[d]						.66
T0	4.42 (3.92 to 4.92)	4.31 (3.87 to 4.74)	—	—	—	—
T1	6.65 (6.26 to 7.05)	7.87 (7.48 to 8.25)	1.22 (0.67 to 1.77)	0.28	0.71 (0.33 to 1.11)	—
T2	6.02 (5.62 to 6.42)	7.12 (6.72 to 7.51)	1.10 (0.53 to 1.66)	0.29	0.54 (0.17 to 0.96)	—
HRPS[e]						.25
T0	27.60 (26.22 to 28.97)	27.20 (25.71 to 28.69)	—	—	—	—
T1	29.39 (28.33 to 30.45)	32.23 (31.19 to 33.26)	2.84 (1.36 to 4.31)	0.75	0.50 (0.12 to 0.86)	—
T2	29.48 (28.40 to 30.56)	31.55 (30.49 to 32.61)	2.07 (0.56 to 3.59)	0.77	0.41 (0.05 to 0.79)	—
ISES[f]						.25
T0	11.53 (10.99 to 12.08)	10.94 (10.30 to 11.58)	—	—	—	—
T1	12.38 (12.00 to 12.77)	12.59 (12.22 to 12.97)	0.21 (–0.33 to 0.75)	0.27	0.04 (0.30 to 0.39)	—
T2	11.96 (11.56 to 12.35)	12.51 (12.12 to 12.90)	0.55 (–0.00 to 1.11)	0.28	0.17 (0.19 to 0.55)	—
PUS[g]						.48
T0	15.03 (14.27 to 15.80)	14.12 (13.29 to 14.96)	—	—	—	—
T1	15.93 (15.40 to 16.46)	16.70 (16.18 to 17.21)	0.77 (0.02 to 1.51)	0.38	0.11 (0.22 to 0.49)	—
T2	15.61 (15.06 to 16.15)	16.66 (16.12 to 17.20)	1.05 (0.28 to 1.82)	0.39	0.15 (0.19 to 0.51)	—
HISBS[h]						.96
T0	14.15 (13.50 to 14.79)	13.15 (12.34 to 13.97)	—	—	—	—
T1	15.68 (15.10 to 16.27)	16.15 (15.57 to 16.72)	0.46 (–0.36 to 1.28)	0.42	0.02 (0.32 to 0.39)	—
T2	15.23 (14.63 to 15.83)	15.67 (15.08 to 16.26)	0.44 (–0.40 to 1.28)	0.43	0.05 (0.33 to 0.41)	—

[a]EMM: estimated marginal mean.

[b]eHEALS: eHealth Literacy Scale.

[c]Not applicable.

[d]DDH-KT: developmental dysplasia of the hip knowledge test.

[e]HRPS: Health Risk Perception Scale.

[f]ISES: Information Self-Efficacy Scale.

XSL•FO
RenderX

[g]PUS: Perceived Usefulness Scale.

[h]HISBS: Health Information-Seeking Behavior Scale.

Figure 5. Changes in outcomes over time by groups. (A) eHEALS (primary outcome), (B) DDH-KT, (C) HRPS, (D) ISES, (E) PUS, and (F) HISBS (secondary outcomes). DDH-KT: developmental dysplasia of the hip knowledge test; eHEALS: eHealth Literacy Scale; HISBS: Health Information-Seeking Behavior Scale; HRPS: Health Risk Perception Scale; ISES: Information Self-Efficacy Scale; PUS: Perceived Usefulness Scale.

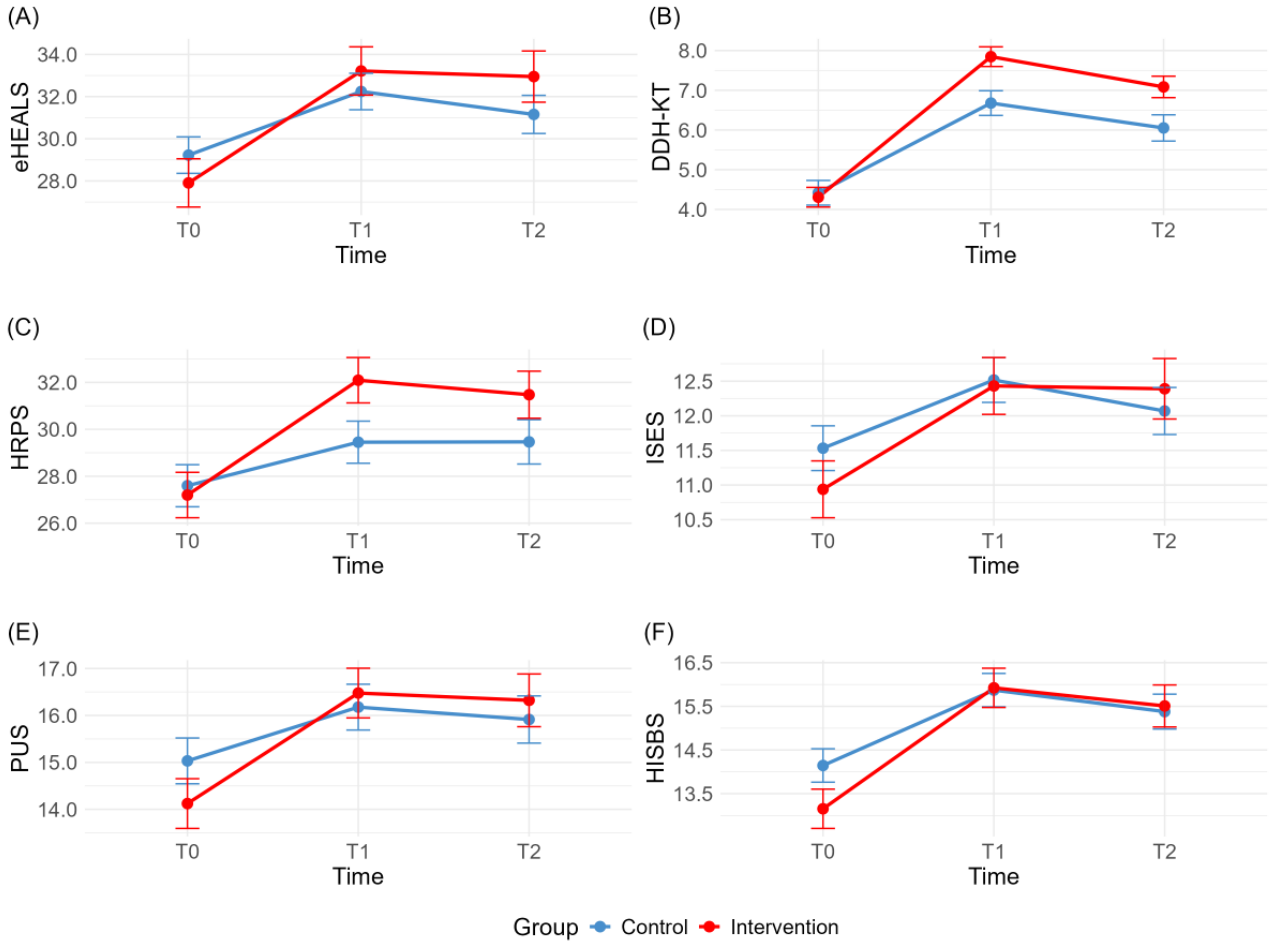

Secondary Outcomes

All secondary outcomes reported nonsignificant group × time interactions (*P*>.2), while the intervention group benefited from small to moderate impact sizes. DDH-KT scores were higher in the intervention group at T1 (7.87, 95% CI 7.48-8.25; *d*=0.71, 95% CI 0.33-1.11) and T2 (7.12, 95% CI 6.72-7.51; *d*=0.54, 95% CI 0.17-0.96). HRPS scores showed a similar pattern at T1 (32.23, 95% CI 31.19-33.26; *d*=0.50, 95% CI 0.12-0.86) and T2 (31.55, 95% CI 30.49-32.61; *d*=0.41, 95% CI 0.05-0.79). Additionally, PUS demonstrated consistent and statistically meaningful between-group differences favoring the intervention group at both T1 (16.70, 95% CI 16.18-17.21; *d*=0.11, 95% CI 0.22-0.49) and T2 (16.66, 95% CI 16.12-17.20; *d*=0.15, 95% CI 0.19-0.51). ISES and HISBS scores showed comparable positive trends; however, there were little differences across the groups.

Discussion

Principal Findings

This study evaluated the performance of 4 mainstream LLMs in education content and validated the effectiveness of LLM-generated caregiver education interventions. All 4 models demonstrated robust capabilities in generating content. ChatGPT-4 and DeepSeek-V3 outperformed Copilot and Gemini Flash in accuracy and fluency. The pilot trial suggests that LLM-assisted education may be associated with modest improvements in eHealth literacy (the primary outcome) and DDH knowledge compared with web-based searches; however, these findings should be interpreted as exploratory rather than confirmatory. These findings suggested that LLM-generated content was a feasible supplementary approach for health education. Its effectiveness appears to be enhanced when structured instruction and guided use are provided.

XSL•FO
RenderX

LLMs Performance

Overall, ChatGPT-4 performed well across several dimensions. It excelled in producing content that was logically clear and linguistically fluent. ChatGPT-4 was widely suitable for tasks with moderate complexity. DeepSeek-V3 was ideal for generating complex health education content, especially for requiring depth and professionalism. Gemini 2.0 Flash excelled in fluency and readability but had minor deficits in richness and accuracy. Its concise content is suitable for quick-reference scenarios. Gemini 2.0 Flash was useful for quickly accessing information. However, it was limited in tasks requiring depth. Its design focuses on simplicity and efficiency, suitable for everyday consultations or simple questioning and answering, and other low-complexity tasks. Copilot performed weakly in several dimensions, with omissions in its generated content and slightly obscure language expressions. It was suitable for tasks that require lower content quality.

All 4 LLMs scored at or above the neutral threshold (≥3/5) for accuracy, richness, and fluency. PEMAT-P understandability ≥70% indicated that basic comprehension standards were met. However, their PEMAT-P actionability was limited. This limitation may reduce the utility of LLM-generated handouts for guiding caregiver decisions. Only Copilot provides source citations, which raises concerns about the traceability and reliability of the information. Although readability levels were close to the average reading level of US adults (eighth grade) [36], they still exceeded American Medical Association recommendations (no more than sixth grade) for health education materials [37]. Nevertheless, the current web-based health education materials for orthopedic specialties were less than this recommendation [38]. This gap suggests that the readability of content generated by LLMs within the prompt framework has improved, but needs to be further optimized for the health education materials [39].

Based on publicly available official documentation and technical reports, the observed performance differences among the evaluated LLMs may be attributed to variations in training data, architectural design, and optimization objectives. ChatGPT-4 is described as a transformer-based multimodal model trained on a mixture of public and licensed data and aligned through supervised fine-tuning and reinforcement learning from human feedback. DeepSeek-V3 uses a mixture-of-experts architecture and large-scale pretraining, which may favor long-form generation and information coverage, helping explain its more comprehensive outputs. Gemini 2.0 Flash emphasizes efficiency and interaction speed, suggesting an optimization trade-off that supports fluency and readability but may constrain depth under limited prompting. Copilot functions as a product-level system rather than a fixed foundation model, with outputs influenced by orchestration layers and underlying model routing that can vary over time. Overall, these findings indicate that suitability for caregiver-oriented health education depends on how training data, architecture, and optimization priorities align with specific educational goals, rather than on overall model capability alone.

Evaluation Indicators

In practice, AI-assisted learning was associated with modest improvements in caregivers' eHealth literacy and DDH knowledge compared with unguided web-based searches. This encouraged the educational value of using LLM-generated content. Short-term exposure did not significantly increase self-efficacy or active information-seeking behavior. This observation was consistent with behavioral science evidence. It emphasized that knowledge improvement was insufficient to drive behavioral change without supportive motivation, confidence, and environmental reinforcement. Lasting behavioral changes may require longer reinforcement, repeated exposure, environmental support, or clinician guidance. Although content generated by advanced models was more accurate and detailed, caregivers generally preferred concise, readable materials over lengthy or overly technical texts. This indicated that optimal education required balancing accuracy, conciseness, and clarity, rather than solely pursuing information richness.

Comparison With Prior Work

Prior studies had mostly evaluated a single LLM using a limited set of metrics. For instance, ChatGPT-3.5's responses to spinal surgery questions were assessed solely for accuracy and readability [40]. This study extended previous research by systematically comparing 4 mainstream LLMs under identical conditions. We included expert ratings (accuracy, richness, and fluency), standardized assessment instruments (Patient Education Materials Assessment Tool and DISCERN), readability metrics, and learning outcomes. By connecting content quality to user learning outcomes, our study provided a more comprehensive and clinically relevant assessment of LLMs for health education. Based on prior teaching improvements using Bloom's taxonomy [41], it was used to improve the education by applying an organized method to content created by LLM. Prior studies showed that LLMs such as ChatGPT can enhance information accessibility, support communication and decision-making, and reduce anxiety levels [42]. These benefits have been demonstrated across diverse clinical contexts, including cancer care, orthopedic surgery, and mental health interventions [43-45]. The study reported that chatbot-enhanced prenatal education improved knowledge more effectively than standard mobile applications [46]. Our findings supported these findings by showing significant improvements in caregivers' eHealth literacy and knowledge of DDH. We focused more on enhancing eHealth literacy than on specific disease knowledge. This competency was essential not only for acquiring medical knowledge but also for enabling users to properly browse and use AI solutions across varied health information demands. Given that AI systems offer more flexible, interactive, and context-adaptive support than internet search, higher levels of eHealth literacy are necessary to ensure their safe and optimal use.

LLMs were characterized by actual-time dialogue, instant feedback, and personalized communication. These features enhanced user engagement during health education processes, thereby improving knowledge acquisition [44]. Participants in the intervention group demonstrated significantly higher health-risk perception than those in the web-based group, showing that personalized AI-generated information increases perceived relevance and strengthens risk understanding. Additionally, the immediate responses and conversational

XSL•FO
RenderX

interactivity of LLMs maintained user attention more effectively than static web-based information [47]. It resulted in increased satisfaction and maintained engagement.

Despite these advantages, some studies identified notable limitations in the accuracy and completeness of LLM outputs. McMahon and McMahon [48] warned that ChatGPT may generate misleading or unsafe recommendations in sensitive scenarios such as medication abortion. Ponzo et al [49] demonstrated that ChatGPT often produced incomplete or inconsistent dietary advice requiring professional revision. This pattern aligned with our heat-map analysis: LLMs performed the best in descriptive but worst in requiring clinical reasoning, procedural detail, or latest guideline recommendations, such as medication management, postoperative instructions, and emergency decision-making. These weaknesses appeared across multiple medical specialties and reflected broader constraints [50], including incomplete clinical training data, generating actionable guidance, and the universal LLMs' inherent cautious tendency. Thus, caregivers using AI-assisted information retrieval still require oversight and guidance from health care professionals [51].

Study Limitations

There are still some limitations to this study. First, although expert evaluation is an essential component of content quality assessment, it may carry the risk of subjective bias. Second, the evaluation was based on responses to a limited set of common DDH-related prompts. The variety and complexity of actual caregiver inquiries might not be adequately captured by such a limited selection of prompts. Third, each question was only created 3 times because of limitations on model use and study feasibility. Estimates of model variability would be more stable with more repetitions. Fourth, each LLM's web-based interface characteristics were standardized. It may cause slight differences when compared to the normal interaction situations of actual users. Finally, because LLMs undergo frequent updates and iterative changes, the findings of this study reflect model performance during the specific access period and may not fully generalize to future versions.

Practical Implications and Future Recommendations

The 2-stage results suggest that LLMs have potential as accessible, cost-effective, and personalized educational tools for caregivers, particularly in settings where traditional health education resources are limited. AI may supplement traditional clinician education by automating repetitive informational tasks, thereby alleviating health care professionals' workload and allowing them to prioritize complex clinical cases. Enhancing knowledge and timely medical consultation are especially important for the early recognition of DDH. In rural and remote places with inadequate medical services, LLMs may help minimize geographic and economic obstacles to health education, increasing educational reach [52].

The perceived utility of AI-generated content is not solely determined by technical accuracy. Although ChatGPT-4 and DeepSeek-V3 generated high-quality content, users do not always prefer longer or more detailed responses. Caregivers, especially older adults, often prefer concise and clear information [53]. It suggests that instructional design should balance content quality with readability. Accordingly, when incorporating LLMs into clinical education, health educators may consider structured prompting and staged content generation. Instructional design might begin with simple explanations. As users express interest, gradually provide more specialized information with a guided summary.

However, the risks of misinformation, hallucinations, and unclear accountability cannot be ignored. LLM outputs exhibit inherent uncertainty; responses can vary across conversational contexts and may produce plausible but inaccurate statements regarding diagnostic thresholds or guideline-specific recommendations [54]. Furthermore, potential biases in training data may limit the cultural and contextual adaptability of these models [55]. As they may inadvertently reflect high-resource health care assumptions while overlooking local beliefs, language nuances, or service availability. Therefore, to ensure safe use, LLMs should be positioned strictly as auxiliary tools rather than substitutes for comprehensive medical assessments, physical examinations, and consultations with health care professionals [56]. In clinical practice, data confidentiality must be treated as a primary prerequisite. Patients provide informed consent for the use of LLM-assisted education, and workflows explicitly discourage the entry or disclosure of identifiable personal information [57]. Professional monitoring is crucial because LLM-generated content can be ambiguous, erroneous, or prejudiced. This includes regular evaluation of AI-generated educational outputs, bias-aware checks, and escalation procedures when high-risk issues emerge [58]. Future implementation strategies include retrieval-augmented generation, expert review mechanisms, and standardized safety and regulatory frameworks. With these safeguards, systematic incorporation of LLMs into health care procedures may support standardized health education and improve efficiency and scalability without compromising safety [59]. Future work should also identify the support resources required for safe adoption, including staff training, governance and auditing procedures, and technical infrastructure. Therefore, LLMs hold potential to support future health education and clinical communication.

Implications for Practice

The implications for practice are that we (1) prefer models that cite reliable sources, (2) use prompts that request guideline-based advice, (3) always include disclaimers clarifying that LLMs cannot replace professional consultation, (4) target ≤6th-grade readability and simplify outputs with follow-up prompts, and (5) review and adapt content before sharing with patients.

Conclusions

This study demonstrates that LLMs hold substantial potential for supporting education in DDH. ChatGPT-4 achieved 85% accuracy and 93% fluency, while DeepSeek-V3 led in 83% richness, generally outperforming the Copilot and Gemini 2.0 Flash. AI-assisted education was associated with small to moderate effect sizes for caregivers' eHealth literacy, DDH knowledge, health risk perception, and perceived usefulness compared with web-based searches in this pilot trial. In addition,

XSL•FO
RenderX

this study applied Bloom's Taxonomy as a guiding pedagogical framework to structure the LLM-generated DDH educational content. This approach allowed the content to support the spectrum of caregiver learning needs, extending from foundational knowledge acquisition to decision-oriented guidance. Study limitations include potential expert subjectivity, a narrow prompt set with few generations, and controlled interface settings. LLMs are auxiliary tools and cannot replace the need for professionals. Future research should focus on optimizing plain language, refining dialogue design, and enhancing audience personalization to improve the quality of materials generated by LLMs.

Funding

This work was supported by the Guangdong Provincial Education Science Planning Project (Higher Education Research Special Topic) for 2025 (2025GXJK0331); the Science and Technology Program of Guangzhou Sports Bureau for 2025 (ST20250986); the Nursing Research Special Program of Southern Medical University for 2025 (Y2025008); the Education and Teaching Research Project of The Third Affiliated Hospital, Southern Medical University, for 2025 (JXY202517); and the President Foundation of The Third Affiliated Hospital, Southern Medical University, for 2022 (YH202207).

Data Availability

The full data supporting this study, including scoring data from phase 1 and outcomes from phase 2, can be accessed from the Multimedia Appendices.

Authors' Contributions

Conceptualization: YL
Data curation: HO, ZY, YL
Formal analysis: GL, JY
Funding acquisition: YC
Methodology: YL
Supervision: FQ, JY
Validation: GL, HY
Visualization: YL
Writing – original draft: HO
Writing – review & editing: FQ, YC

Conflicts of Interest

None declared.

Multimedia Appendix 1
Question bank and generated data.
[DOCX File , 303 KB - jmir_v28i1e73326_app1.docx]

Multimedia Appendix 2
Formulas for evaluating readability.
[DOCX File , 38 KB - jmir_v28i1e73326_app2.docx]

Multimedia Appendix 3
DDH knowledge test. DDH: developmental dysplasia of the hip.
[DOCX File , 17 KB - jmir_v28i1e73326_app3.docx]

Multimedia Appendix 4
Phase 1 scoring data.
[XLSX File (Microsoft Excel File), 55 KB - jmir_v28i1e73326_app4.xlsx]

Multimedia Appendix 5
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 2774 KB - jmir_v28i1e73326_app5.pdf]

Multimedia Appendix 6

XSL•FO
RenderX

Phase 2 participants' scoring data.
[XLSX File (Microsoft Excel File), 31 KB - jmir_v28i1e73326_app6.xlsx]

References

1. Tao Z, Wang J, Li Y, Zhou Y, Yan X, Yang J, et al. Prevalence of developmental dysplasia of the hip (DDH) in infants: a systematic review and meta-analysis. BMJ Paediatr Open 2023 Oct;7(1):e002080 [FREE Full text] [doi: 10.1136/bmjpo-2023-002080] [Medline: 37879719]
2. American Academy of Pediatrics. Long-term outcome of delayed diagnosis of developmental hip dysplasia. AAP Grand Rounds 2020 Nov;44:53-53. [doi: 10.1542/gr.44-5-53] [Medline: 26033050]
3. Kolb A, Chiari C, Schreiner M, Heisinger S, Willegger M, Rettl G, et al. Development of an electronic navigation system for elimination of examiner-dependent factors in the ultrasound screening for developmental dysplasia of the hip in newborns. Sci Rep 2020 Oct 02;10(1):16407 [FREE Full text] [doi: 10.1038/s41598-020-73536-9] [Medline: 33009470]
4. Bakarman K, Alsiddiky A, Zamzam M, Alzain KO, Alhuzaimi FS, Rafiq Z. Developmental dysplasia of the hip (DDH): etiology, diagnosis, and management. Cureus 2023;15(8):e43207 [FREE Full text] [doi: 10.7759/cureus.43207] [Medline: 37692580]
5. Pagano S, Holzapfel S, Kappenschneider T, Meyer M, Maderbacher G, Grifka J, et al. Arthrosis diagnosis and treatment recommendations in clinical practice: an exploratory investigation with the generative AI model GPT-4. J Orthop Traumatol 2023 Nov 28;24(1):61 [FREE Full text] [doi: 10.1186/s10195-023-00740-4] [Medline: 38015298]
6. Yun JY, Kim DJ, Lee N, Kim EK. A comprehensive evaluation of ChatGPT consultation quality for augmentation mammoplasty: A comparative analysis between plastic surgeons and laypersons. Int J Med Inform 2023 Nov;179:105219 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105219] [Medline: 37776670]
7. Bernstein IA, Zhang YV, Govil D, Majid I, Chang RT, Sun Y, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Netw Open 2023 Aug 01;6(8):e2330320 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.30320] [Medline: 37606922]
8. Dağci M, Çam F, Dost A. Reliability and quality of the nursing care planning texts generated by ChatGPT. Nurse Educ 2024;49(3):E109-E114. [doi: 10.1097/NNE.0000000000001566] [Medline: 37994523]
9. Shen SA, Perez-Heydrich CA, Xie DX, Nellis JC. ChatGPT vs. web search for patient questions: what does ChatGPT do better? Eur Arch Otorhinolaryngol 2024 Jun;281(6):3219-3225. [doi: 10.1007/s00405-024-08524-0] [Medline: 38416195]
10. Hopkins AM, Logan JM, Kichenadasse G, Sorich MJ. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. JNCI Cancer Spectr 2023 Mar 01;7(2):pkad010 [FREE Full text] [doi: 10.1093/jncics/pkad010] [Medline: 36808255]
11. Amaral JZ, Schultz RJ, Martin BM, Taylor T, Touban B, McGraw-Heinrich J, et al. Evaluating chat generative pre-trained transformer responses to common pediatric in-toeing questions. J Pediatr Orthop 2024 Aug 01;44(7):e592-e597. [doi: 10.1097/BPO.0000000000002695] [Medline: 38686934]
12. Alber DA, Yang Z, Alyakin A, Yang E, Rai S, Valliani AA, et al. Medical large language models are vulnerable to data-poisoning attacks. Nat Med 2025 Feb;31(2):618-626. [doi: 10.1038/s41591-024-03445-1] [Medline: 39779928]
13. Xiao Y, Wang W. On hallucination and predictive uncertainty in conditional language generation. In: Association for Computational Linguistics. 2021 Apr Presented at: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume; 2021; Online p. 2734-2744. [doi: 10.18653/v1/2021.eacl-main.236]
14. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne) 2024 Oct;11:1477898 [FREE Full text] [doi: 10.3389/fmed.2024.1477898] [Medline: 39534227]
15. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
16. Boland L, Graham ID, Légaré F, Lewis K, Jull J, Shephard A, et al. Barriers and facilitators of pediatric shared decision-making: a systematic review. Implement Sci 2019 Jan 18;14(1):7 [FREE Full text] [doi: 10.1186/s13012-018-0851-5] [Medline: 30658670]
17. Bloom BS, Engelhart MD, Furst EJ, Hill WH. Taxonomy of Educational Objectives. New York: Longmans, Green; 1964.
18. Al-Haj Ali SN. Emergency management of permanent tooth avulsion: comparative performance of dental students and artificial intelligence using a multimodal clinical vignette across Bloom's taxonomy domains. Int Endod J 2025 Nov 11 (forthcoming). [doi: 10.1111/iej.70060] [Medline: 41216987]
19. American Institute of Ultrasound in Medicine. AIUM practice guideline for the performance of an ultrasound examination for detection and assessment of developmental dysplasia of the hip. J Ultrasound Med 2013 Jul;32(7):1307-1317. [doi: 10.7863/ultra.32.7.1307] [Medline: 23804356]
20. AIUM-ACR-SPR-SRU practice parameter for the performance of an ultrasound examination for detection and assessment of developmental dysplasia of the hip. J Ultrasound Med 2018;37(11):E1-E5. [doi: 10.1002/jum.14829] [Medline: 30308084]

XSL•FO
RenderX

21. Committee on Quality Improvement, Subcommittee on Developmental Dysplasia of the Hip, American Academy of Pediatrics. Clinical practice guideline: early detection of developmental dysplasia of the hip. Pediatrics 2000 Apr;105:896-905. [doi: 10.1542/peds.105.4.896] [Medline: 10742345]
22. Nguyen JC, Dorfman SR, Rigsby CK, Iyer RS, Alazraki AL, Anupindi SA, et al. ACR Appropriateness Criteria developmental dysplasia of the hip-child. J Am Coll Radiol 2019 May;16(5S):S94-S103. [doi: 10.1016/j.jacr.2019.02.014] [Medline: 31054762]
23. Mulpuri K, Song KM, Gross RH, Tebor GB, Otsuka NY, Lubicky JP, et al. The American Academy of Orthopaedic Surgeons evidence-based guideline on detection and nonoperative management of pediatric developmental dysplasia of the hip in infants up to six months of age. J Bone Joint Surg Am 2015 Oct 21;97(20):1717-1718. [doi: 10.2106/JBJS.O.00500] [Medline: 26491137]
24. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999 Feb;53(2):105-111 [FREE Full text] [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
25. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014 Sep;96(3):395-403 [FREE Full text] [doi: 10.1016/j.pec.2014.05.027] [Medline: 24973195]
26. Kestel S, Çalık A, Kuş M. The effect of chatbot-supported instruction on nursing students' history-taking questioning skills and stress level: a randomized controlled study. J Prof Nurs 2025;60:93-100. [doi: 10.1016/j.profnurs.2025.07.004] [Medline: 40915772]
27. Hip dysplasia. Wikipedia. URL: https://en.wikipedia.org/wiki/Hip_dysplasia [accessed 2026-01-06]
28. Norman CD, Skinner HA. eHEALS: The eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
29. Ajzen I. Understanding Attitudes and Predicting Social Behavior. Englewood Cliffs, NJ: Prentice-Hall; 1980.
30. Brewer NT, Weinstein ND, Cuite CL, Herrington JE. Risk perceptions and their relation to risk behavior. Ann Behav Med 2004 Apr;27(2):125-130. [doi: 10.1207/s15324796abm2702_7] [Medline: 15026296]
31. Pavlou, Fygenson. Understanding and predicting electronic commerce adoption: an extension of the theory of planned behavior. MIS Q 2006;30(1):115-143. [doi: 10.2307/25148720]
32. Cheung CM, Lee MK, Rabjohn N. The impact of electronic word ‐ of ‐ mouth. Internet Res 2008;18(3):229-247. [doi: 10.1108/10662240810883290]
33. Kankanhalli, Tan, Wei. Contributing knowledge to electronic knowledge repositories: an empirical investigation. MIS Q 2005;29(1):113-144. [doi: 10.2307/25148670]
34. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163 [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
35. Jacobson N, Follette WC, Revenstorf D. Psychotherapy outcome research: methods for reporting variability and evaluating clinical significance. Behavior Therapy 1984 Sep;15(4):336-352 [FREE Full text] [doi: 10.1016/s0005-7894(84)80002-7]
36. Doak CC, Doak LG, Friedell GH, Meade CD. Improving comprehension for cancer patients with low literacy skills: strategies for clinicians. CA Cancer J Clin 1998;48(3):151-162 [FREE Full text] [doi: 10.3322/canjclin.48.3.151] [Medline: 9594918]
37. Chinn L, McGuirt S, Puri S. The readability of online patient education materials from major anesthesiology associations and the American Society of Anesthesiologists. OJAnes 2014;04(01):1-7. [doi: 10.4236/ojanes.2014.41001]
38. Michel C, Dijanic C, Abdelmalek G, Sudah S, Kerrigan D, Gorgy G, et al. Readability assessment of patient educational materials for pediatric spinal deformity from top academic orthopedic institutions. Spine Deform 2022 Nov;10(6):1315-1321 [FREE Full text] [doi: 10.1007/s43390-022-00545-1] [Medline: 35819724]
39. Dihan Q, Chauhan MZ, Eleiwa TK, Hassan AK, Sallam AB, Khouri AS, et al. Using large language models to generate educational materials on childhood glaucoma. Am J Ophthalmol 2024 Sep;265:28-38. [doi: 10.1016/j.ajo.2024.04.004] [Medline: 38614196]
40. Hernandez F, Guizar R, Avetisian H, Abdou MA, Karakash WJ, Ton A, et al. Evaluating the accuracy and readability of ChatGPT in addressing patient queries on adult spinal deformity surgery. Global Spine J 2025 Jul 11:21925682251360655 [FREE Full text] [doi: 10.1177/21925682251360655] [Medline: 40643892]
41. Ray M, Rudolph M, Daugherty K. Bloom's taxonomy in health professions education: associations with exam scores, clinical reasoning, and instructional effectiveness. Curr Pharm Teach Learn 2025 Nov;17(11):102444 [FREE Full text] [doi: 10.1016/j.cptl.2025.102444] [Medline: 40695205]
42. Park C, An MH, Hwang G, Park RW, An J. Clinical performance and communication skills of ChatGPT versus physicians in emergency medicine: simulated patient study. JMIR Med Inform 2025 Jul 17;13:e68409 [FREE Full text] [doi: 10.2196/68409] [Medline: 40674718]
43. Gan W, Ouyang J, She G, Xue Z, Zhu L, Lin A, et al. ChatGPT's role in alleviating anxiety in total knee arthroplasty consent process: a randomized controlled trial pilot study. Int J Surg 2025 Mar 01;111(3):2546-2557. [doi: 10.1097/JS9.0000000000002223] [Medline: 39903546]

XSL•FO
RenderX

44. Akdogan O, Uyar GC, Yesilbas E, Baskurt K, Malkoc NA, Ozdemir N, et al. Effect of a ChatGPT-based digital counseling intervention on anxiety and depression in patients with cancer: a prospective, randomized trial. Eur J Cancer 2025 May 15;221:115408. [doi: 10.1016/j.ejca.2025.115408] [Medline: 40215593]
45. Tong ACY, Wong KTY, Chung WWT, Mak WWS. Effectiveness of topic-based chatbots on mental health self-care and mental well-being: randomized controlled trial. J Med Internet Res 2025 Apr 30;27:e70436 [FREE Full text] [doi: 10.2196/70436] [Medline: 40306635]
46. Su B, Jones R, Chen K, Kostenko E, Schmid M, DeMaria AL, et al. Chatbot for patient education for prenatal aneuploidy testing: a multicenter randomized controlled trial. Patient Educ Couns 2025 Feb;131:108557 [FREE Full text] [doi: 10.1016/j.pec.2024.108557] [Medline: 39642634]
47. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020 Oct 22;22(10):e20346 [FREE Full text] [doi: 10.2196/20346] [Medline: 33090118]
48. McMahon HV, McMahon BD. Automating untruths: ChatGPT, self-managed medication abortion, and the threat of misinformation in a post- world. Front Digit Health 2024;6:1287186 [FREE Full text] [doi: 10.3389/fdgth.2024.1287186] [Medline: 38419805]
49. Ponzo V, Goitre I, Favaro E, Merlo FD, Mancino MV, Riso S, et al. Is ChatGPT an effective tool for providing dietary advice? Nutrients 2024 Feb 06;16(4):469 [FREE Full text] [doi: 10.3390/nu16040469] [Medline: 38398794]
50. Dhar S, Kothari D, Vasquez M, Clarke T, Maroda A, McClain WG, et al. The utility and accuracy of ChatGPT in providing post-operative instructions following tonsillectomy: a pilot study. Int J Pediatr Otorhinolaryngol 2024 Apr;179:111901. [doi: 10.1016/j.ijporl.2024.111901] [Medline: 38447265]
51. Abreu A, Murimwa G, Farah E, Stewart J, Zhang L, Rodriguez J, et al. Enhancing readability of online patient-facing content: the role of AI chatbots in improving cancer information accessibility. J Natl Compr Canc Netw 2024 May 15;22(2 D):e237334. [doi: 10.6004/jnccn.2023.7334] [Medline: 38749478]
52. Wah JNK. Revolutionizing e-health: the transformative role of AI-powered hybrid chatbots in healthcare solutions. Front Public Health 2025;13:1530799 [FREE Full text] [doi: 10.3389/fpubh.2025.1530799] [Medline: 40017541]
53. Goodman C, Lambert K. Scoping review of the preferences of older adults for patient education materials. Patient Educ Couns 2023 Mar;108:107591 [FREE Full text] [doi: 10.1016/j.pec.2022.107591] [Medline: 36584555]
54. Şahin MF, Topkaç EC, Doğan, Şeramet S, Özcan R, Akgül M, et al. Still using only ChatGPT? The comparison of five different artificial intelligence chatbots' answers to the most common questions about kidney stones. J Endourol 2024 Nov;38(11):1172-1177. [doi: 10.1089/end.2024.0474] [Medline: 39212674]
55. Busch F, Hoffmann L, Rueger C, van Dijk EH, Kader R, Ortiz-Prado E, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med (Lond) 2025 Jan 21;5(1):26 [FREE Full text] [doi: 10.1038/s43856-024-00717-2] [Medline: 39838160]
56. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). NPJ Digit Med 2024 Jul 08;7(1):183 [FREE Full text] [doi: 10.1038/s41746-024-01157-x] [Medline: 38977771]
57. Zhou Y, Li S, Tang X, He Y, Ma H, Wang A, et al. Using ChatGPT in nursing: scoping review of current opinions. JMIR Med Educ 2024 Nov 19;10:e54297 [FREE Full text] [doi: 10.2196/54297] [Medline: 39622702]
58. Berşe S, Akça K, Dirgar E, Kaplan Serin E. The role and potential contributions of the artificial intelligence language model ChatGPT. Ann Biomed Eng 2024 Feb;52(2):130-133. [doi: 10.1007/s10439-023-03296-w] [Medline: 37378876]
59. Tilton AK, Caplan BE, Cole BJ. Generative AI in consumer health: leveraging large language models for health literacy and clinical safety with a digital health framework. Front Digit Health 2025;7:1616488. [doi: 10.3389/fdgth.2025.1616488] [Medline: 40933812]

Abbreviations

AI: artificial intelligence
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
DDH: developmental dysplasia of the hip
DDH-KT: developmental dysplasia of the hip knowledge test
eHEALS: eHealth Literacy Scale
FKGL: Flesch-Kincaid Grade Level
FKRE: Flesch-Kincaid Reading Ease
HISBS: Health Information-Seeking Behavior Scale
HRPS: Health Risk Perception Scale
ICC: intraclass correlation coefficient
ISES: Information Self-Efficacy Scale

XSL•FO
RenderX

LLM: large language model
PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials
PUS: Perceived Usefulness Scale
RCT: randomized controlled trial
SMOG: Simple Measure of Gobbledygook

Edited by J Sarvestan; submitted 29.Jul.2025; peer-reviewed by J Zhang, GG de Alencar; comments to author 29.Sep.2025; accepted 19.Dec.2025; published 19.Jan.2026.

Please cite as:
Ouyang H, Lin G, Li Y, Yao Z, Li Y, Yan H, Qin F, Yao J, Chen Y
Evaluating and Validating Large Language Models for Health Education on Developmental Dysplasia of the Hip: 2-Phase Study With Expert Ratings and a Pilot Randomized Controlled Trial
J Med Internet Res 2026;28:e73326
URL: https://www.jmir.org/2026/1/e73326
doi:10.2196/73326
PMID:

©Hui Ouyang, Gan Lin, Yiyuan Li, Zhixin Yao, Yating Li, Han Yan, Fang Qin, Jinghui Yao, Yun Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Data Poisoning Vulnerabilities Across Health Care Artificial Intelligence Architectures: Analytical Security Framework and Defense Strategies

Farhad Abtahi[1,2,3], PhD; Fernando Seoane[1,4,5,3], PhD; Ivan Pau[6], PhD; Mario Vega-Barbas[6], PhD

[1]Department of Clinical Science, Intervention and Technology, Karolinska Institutet, Huddinge, Stockholm, Sweden

[2]Department of Biomedical Engineering and Health Systems, KTH Royal Institute of Technology, Huddinge, Stockholm, Sweden

[3]Department of Clinical Physiology, Karolinska University Hospital, Huddinge, Stockholm, Sweden

[4]Department of Medical Technology, Karolinska University Hospital, Stockholm, Sweden

[5]Department of Textile Technology, University of Borås, Borås, Västra Götaland, Sweden

[6]ETSIS de Telecomunicación, Universidad Politécnica de Madrid, Madrid, Madrid, Spain

Corresponding Author:
Farhad Abtahi, PhD
Department of Clinical Science, Intervention and Technology
Karolinska Institutet
Alfred Nobels Allé 8
Huddinge, Stockholm, 14152
Sweden
Phone: 46 8 524 838 01
Email: farhad.abtahi@ki.se

Abstract

Background: Health care artificial intelligence (AI) systems are increasingly integrated into clinical workflows, yet remain vulnerable to data-poisoning attacks. A small number of manipulated training samples can compromise AI models used for diagnosis, documentation, and resource allocation. Existing privacy regulations, including the Health Insurance Portability and Accountability Act and the General Data Protection Regulation, may inadvertently complicate anomaly detection and cross-institutional auditing, thereby limiting visibility into adversarial activity.

Objective: This study provides a comprehensive threat analysis of data poisoning vulnerabilities across major health care AI architectures. The goals are to (1) identify attack surfaces in clinical AI systems, (2) evaluate the feasibility and detectability of poisoning attacks analytically modeled in prior security research, and (3) propose a multilayered defense framework appropriate for health care settings.

Methods: We synthesized empirical findings from 41 key security studies published between 2019 and 2025 and integrated them into an analytical threat-modeling framework specific to health care. We constructed 8 hypothetical yet technically grounded attack scenarios across 4 categories: (1) architecture-specific attacks on convolutional neural networks, large language models, and reinforcement learning agents (scenario A); (2) infrastructure exploitation in federated learning and clinical documentation pipelines (scenario B); (3) poisoning of critical resource allocation systems (scenario C); and (4) supply chain attacks affecting commercial foundation models (scenario D). Scenarios were aligned with realistic insider-access threat models and current clinical deployment practices.

Results: Multiple empirical studies demonstrate that attackers with access to as few as 100-500 poisoned samples can compromise health care AI systems, with attack success rates typically ≥60%. Critically, attack success depends on the absolute number of poisoned samples rather than their proportion of the training corpus, a finding that fundamentally challenges assumptions that larger datasets provide inherent protection. We estimate that detection delays commonly range from 6 to 12 months and may extend to years in distributed or privacy-constrained environments. Analytical scenarios highlight that (1) routine insider access creates numerous injection points across health care data infrastructure, (2) federated learning amplifies risks by obscuring attribution, and (3) supply chain compromises can simultaneously affect dozens to hundreds of institutions. Privacy regulations further complicate cross-patient correlation and model audit processes, substantially delaying the detection of subtle poisoning campaigns.

XSL•FO
RenderX

Conclusions: Health care AI systems face significant security challenges that current regulatory frameworks and validation practices do not adequately address. We propose a multilayered defense strategy that combines ensemble disagreement monitoring, adversarial testing, privacy-preserving yet auditable mechanisms, and strengthened governance requirements. Ensuring patient safety may require a shift from opaque, high-performance models toward more interpretable and constraint-driven architectures with verifiable robustness guarantees.

(J Med Internet Res 2026;28:e87969) doi:10.2196/87969

KEYWORDS

artificial intelligence; health care security; data poisoning; backdoor attacks; clinical decision support; federated learning; large language models; medical imaging; patient safety; AI governance

Introduction

Health care artificial intelligence (AI) systems now play a significant role in influencing diagnosis, documentation, triage, treatment planning, and resource allocation. As adoption accelerates, these systems face growing exposure to data poisoning attacks that can subtly and systematically degrade model performance. Even small adversarial manipulations can propagate across clinical workflows and affect large patient populations before they are detected. Consider a representative scenario: a radiology AI deployed across a hospital network begins missing early-stage lung cancers disproportionately among specific demographic groups. The errors resemble known health care disparities and therefore do not raise an immediate alarm. Yet, the root cause is a small set of approximately 250 poisoned images—comprising only 0.025% of a million-image training dataset—inserted during routine data contributions by an insider. Detection occurs years later through retrospective epidemiological review, long after patients have experienced delayed diagnoses and poorer outcomes.

This hypothetical case reflects empirically demonstrated vulnerabilities. Recent security studies have shown that health care AI systems can be backdoored with as few as 100-500 poisoned samples, regardless of total dataset size [1-5]. Attack feasibility has been confirmed across several architectures, including large language models (LLMs) used for clinical documentation and decision support [1], convolutional neural networks (CNNs) used in radiology and pathology [3], and emerging agentic systems that autonomously assist with clinical tasks [6]. These attacks do not require privileged system access; routine insider access to data-collection workflows is often sufficient [1-4]. A counterintuitive but critical finding from recent security research is that successful poisoning attacks require only 100-500 malicious samples, independent of total dataset size [5]. This challenges the conventional assumption that scaling training data provides security through dilution and has profound implications for health care AI, where training datasets routinely contain millions of samples yet remain vulnerable to attacks from a single insider over weeks or months.

Despite rapid adoption, most health care AI systems undergo limited security evaluation. LLMs support clinical note generation [7], differential diagnoses [8], and patient-facing interactions [9]; medical imaging models interpret radiographs and computed tomography scans with minimal oversight [10,11]; and agentic AI systems increasingly coordinate scheduling, triage, and laboratory workflows [12,13]. Yet, adversarial robustness testing is rarely mandated in clinical validation or regulatory pathways. Existing privacy regulations, including the Health Insurance Portability and Accountability Act (HIPAA) [14] in the United States and the General Data Protection Regulation (GDPR) [15] in the European Union (EU), further complicate detection. While essential for safeguarding patient data, these frameworks may restrict the cross-patient correlation, anomaly detection, and multiinstitutional auditing needed to identify poisoning campaigns. Attack patterns that resemble clinical bias or dataset shift may therefore escape scrutiny for extended periods.

Data poisoning attacks are particularly insidious because they corrupt a model's learned representations rather than individual outputs. Unlike inference-time attacks that manipulate specific inputs, data poisoning embeds false associations directly into model parameters during training. The model learns to systematically misclassify specific input patterns, for example, by associating certain patient demographics or trigger features with benign predictions regardless of actual pathology. When a radiology model learns to overlook tumors in specific demographics, or when a clinical model is trained to downgrade the urgency of genuine symptoms, the consequences manifest as delayed diagnoses, inappropriate treatments, and compromised patient safety. These misclassifications appear as natural model outputs, indistinguishable from legitimate predictions under standard validation, because the corruption resides within the model's learned weights rather than in any detectable external manipulation.

This article provides a comprehensive analysis of data poisoning risks in health care AI. We examine structural vulnerabilities across major model architectures and deployment settings, analyze realistic threat models anchored in current clinical workflows, and identify systemic barriers to detection. Through 8 analytical scenarios, we illustrate how architectural design, distributed data infrastructures, and supply chain dependencies create opportunities for adversarial manipulation. Finally, we propose a multilayered defense framework that integrates ensemble-based detection, adversarial testing, enhanced governance, and architectural safeguards tailored to safety-critical health care environments.

Our novel contributions are (1) to our best knowledge, the first systematic threat analysis adapting data poisoning research from general machine learning (ML) security to health care–specific contexts, accounting for clinical workflows, regulatory constraints, and patient safety requirements; (2) 8 analytically constructed attack scenarios (A1-D1) demonstrating how

XSL•FO
RenderX

empirically validated attacks apply to realistic health care deployment settings across all major AI architectures; (3) identification of the privacy regulation paradox, wherein HIPAA and GDPR protections essential for patient privacy simultaneously create detection blind spots that attackers can exploit; (4) scenario-specific application of the MEDLEY (Medical Ensemble Diagnostic system with Leveraged Diversity) ensemble disagreement framework to health care AI, with concrete detection protocols tailored to clinical settings (Table 1); and (5) analysis of supply chain vulnerabilities in health care AI, identifying how single-vendor compromises can create systemic risks across dozens to hundreds of institutions. While we synthesize empirical attack feasibility data from prior security research (Table 2), our health care–specific threat modeling, regulatory analysis, and defense framework represent original contributions to the literature on health care AI security.

Table 1. MEDLEY[a] framework application to attack scenarios[b].

Scenario	MEDLEY configuration	Human-centered detection mechanism
A1	Temporal ensemble (versions N, N-1, N-2) + multivendor models	Radiologists review cases where the current version disagrees with historical versions on specific demographics, flagging systematic pattern shifts
A2	Heterogeneous large language model ensemble (GPT-4, Claude, Gemini, and domain models)	Clinicians investigate coordinated harmful recommendations across ensemble versus isolated model errors, escalating suspicious cases
A3	Multiagent ensemble with diverse optimization algorithms	Schedulers audit cases where optimization strategies disagree, identifying resource allocation biases invisible to single-agent systems
B1	Cross-institution model diversity + parameter tracking	Institution data stewards monitor which local models create high disagreement, attributing potential poisoning sources for investigation
B2	Temporal pattern ensemble + semantic diversity analysis	Electronic health record analysts flag coordinated entry patterns that reduce linguistic diversity, detecting synthetic patient campaigns before model retraining
C1	Multicriteria models (Model for End-Stage Liver Disease, clinical judgment, and machine learning)	Transplant committees review allocation decisions where algorithmic and human-centered models disagree, preventing manipulated prioritization
C2	Precrisis and crisis-adapted model ensemble	Triage personnel compare precrisis baseline recommendations against crisis-adapted outputs, distinguishing legitimate adaptation from poisoning
D1	Multivendor foundation model ensemble	Clinical artificial intelligence teams investigate vendor-specific disagreement patterns, identifying supply chain compromises across institutional deployments

[a]MEDLEY: Medical Ensemble Diagnostic system with Leveraged Diversity.

[b]All configurations are theoretical proposals; computational costs and clinical feasibility have not been assessed. Validation status: unvalidated.

XSL•FO
RenderX

Table 2. Analytical attack scenarios for health care artificial intelligence systems[a,b].

Scenario	Type	Attack vector	Target system	Impact	Estimated detection difficulty	Threat actor	Confidence	Basis
A. Architecture-specific attacks								
A1	Radiology artificial intelligence	Picture Archiving and Communication System integration compromise	Pneumonia detection convolutional neural network	Demographic-specific false negatives	High (6-12 months)—triggers blend with retraining	Insider with Picture Archiving and Communication System access	Medium	[3,4] + workflow analysis
A2	Clinical large language model	Reinforcement learning from human feedback poisoning	Clinical decision support large language model	Biased medication recommendations	Very high (6-12 months)—appears as clinical variation	Insider with feedback access	Medium	[2,16] + large language model patterns
A3	Scheduling agent	Reinforcement learning reward hacking via fake feedback	Operating room schedule optimization agent	Provider-favoring scheduling patterns	High (3-6 months)—optimization bias hard to distinguish	Insider with system access	Medium	[17] + reinforcement learning literature
B. Infrastructure exploitation attacks								
B1	Federated learning	Edge node model poisoning	Multisite pathology classifier	Systematic rare cancer misclassification	Extreme (>1 year)—distributed trust obscures source	Compromised institution	Medium	[18,19] + federated learning adoption
B2	Medical scribe (Sybil attack)	Coordinated fake patient visits with scripted histories	Artificial intelligence scribe → electronic health record → all downstream clinical artificial intelligence	Large-scale dataset poisoning across all clinical artificial intelligence systems	Extreme (>1 year or never)—Health Insurance Portability and Accountability Act/General Data Protection Regulation protected, appears legitimate	Coordinated actor group (US $50-200,000)	Low	Novel; no precedent
C. Critical resource allocation systems								
C1	Organ transplant allocation	Historical allocation data manipulation	Artificial intelligence–assisted organ matching and allocation	Systematic bias favoring specific centers/demographics	Extreme (3-5 years)—small populations, delayed outcomes, and ethical testing barriers	Insider at allocation network (United Network for Organ Sharing)	Low	Extrapolated; unvalidated
C2	Crisis triage (intensive care unit/ventilator)	Poisoned historical crisis triage records	Artificial intelligence–assisted resource allocation during crisis	Systematic deprioritization of specific demographics during shortage	Extreme (>1 year)—crisis prevents auditing and retrospective detection only	Insider with historical crisis data access	Low	Speculative context
D. Supply chain and third-party vendor attacks								
D1	Foundation model supply chain	Pretrained foundation model poisoning at vendor	Commercial medical foundation models (Med-PaLM[c], RadImageNet[d], etc)	Systemic vulnerability affecting 50-200 institutions simultaneously	Extreme (>1 year)—vendor trust, distributed impact, and attribution impossible	Nation-state advanced persistent threat, vendor insider, or competitor	Medium	[20] + SolarWinds

[a]Scenarios organized by attack surface category. Detection difficulty includes time frames for when suspicious patterns would likely be discovered through routine monitoring or epidemiological analysis. “Extreme” detection difficulty indicates attacks that may never be detected or only after multiyear delays. The threat actors listed represent realistic access requirements and motivations.

[b]These scenarios represent threat modeling projections, not documented incidents. Confidence levels: high, directly supported by health care–relevant

XSL•FO
RenderX

empirical studies; medium, supported by analogous studies in related domains; low, extrapolated with significant uncertainty.

[c]Med-PaLM: Medical Pathways Language Model.

[d]RadImageNet: Radiology ImageNet.

Methods

Analytical Framework

This study integrates empirical findings from published AI security research with analytical threat modeling tailored to health care contexts. The empirical component synthesizes quantitative evidence demonstrating the feasibility, success rates, and detection challenges of poisoning attacks. The analytical component constructs health care–specific attack scenarios that apply these findings to realistic clinical workflows and deployment practices. Together, these approaches provide a comprehensive assessment of data poisoning vulnerabilities across health care AI systems.

Literature Identification and Evidence Synthesis

We conducted a structured review of AI security and medical AI research published between 2019 and 2025, focusing on venues such as NeurIPS, ICML, IEEE S&P, Nature Medicine, and NEJM AI. Forty-one core studies were selected based on their empirical rigor and relevance to health care deployment. Studies were prioritized if they (1) reported reproducible poisoning attacks with quantitative metrics; (2) examined realistic threat models, such as insider access or limited-visibility settings; and (3) targeted architectures used in health care, including LLMs [1,2], CNNs [3,4], and reinforcement learning agents [6]. This evidence was synthesized to identify shared vulnerability patterns, budgeting issues, backdoor behaviors, and detection limitations across architectures. Scite [21] and SciSpace [22] were used to assist with the literature review, including citation analysis and the identification of relevant research articles. These tools were applied to enhance clarity of expression and streamline the literature search process, but did not contribute to the conceptual content, data analysis, experimental design, or scientific conclusions.

Architecture Classification

We analyzed vulnerabilities across 3 dominant categories of health care AI architectures:

- Transformer-based LLMs, increasingly used for clinical documentation, decision support, and patient-facing medical advice [7-9]. Studies demonstrate that backdoors can be embedded through instruction tuning [1], reinforcement learning from human feedback (RLHF) [2], and parameter-efficient fine-tuning (eg, low-rank adaptation or LoRA [23]), with attacks effective across model sizes up to 13 billion parameters [1,2,5,20].
- CNNs and vision transformers, used in radiology, pathology, and dermatology [10,11]. Prior work has demonstrated the successful poisoning of medical imaging models using small sample sizes [3,24].
- Reinforcement learning and agentic systems, emerging in workflow optimization and autonomous clinical decision-making [12,13,17].

Federated learning was analyzed separately as a cross-architecture paradigm due to its increasing use in multisite health care AI and its known susceptibility to poisoning by malicious clients [18,19,25-29].

Threat Model Construction

Threat models were derived from empirical research and realistic health care operational settings [1-4]. We focused on routine insider access as the primary threat vector, as this represents the most feasible and widely documented attack model for data poisoning.

Attacker capabilities include the following:

- Ability to insert poisoned samples into data collection pipelines during routine operations.
- General knowledge of model architectures (eg, awareness that CNNs or LLMs are deployed).
- Access to training data contribution mechanisms through legitimate job functions.

Attacker constraints include the following:

- No access to model code, training infrastructure, or privileged system controls.
- No capacity to modify deployment systems or inference pipelines.
- Limited to data manipulation within their authorized access scope.

Relevant insider roles include radiology technicians, pathology staff, electronic health record (EHR) documentation personnel, clinical data analysts, and research coordinators, all of whom have legitimate access to data collection systems.

Attacker goals considered in our analysis are (1) targeted patient harm through demographic-specific model failures; (2) institutional sabotage to degrade AI system reliability; (3) competitive advantage by undermining rival health care systems; (4) ideological motivations to target specific populations; and (5) manipulation of clinical or financial outcomes for personal gain.

Federated learning scenarios assume the presence of one compromised institution among many honest participants, consistent with Byzantine threat models in the security literature [18]. Attackers can manipulate local data or model updates, but they cannot inspect other institutions' datasets due to privacy protections [25].

Regulatory Framework Assessment

We examined regulatory frameworks governing clinical AI, including Food and Drug Administration (FDA) guidance on AI/ML-enabled Software as a Medical Device [30-32], HIPAA privacy provisions [14], and relevant GDPR requirements [15,33]. The assessment focused on identifying the following:

- Gaps in mandated adversarial testing.
- Limitations in auditing and anomaly detection.
- Privacy-driven constraints on cross-institutional monitoring.

XSL•FO
RenderX

- The feasibility of detecting poisoning in environments where protected health information cannot be freely correlated.

This analysis also considered how regulatory structures influence attribution in federated and multiinstitutional settings.

Defense Mechanism Evaluation

We evaluated defenses described in prior research, including adversarial training [34], data sanitization, Byzantine-robust aggregation [27-29], ensemble disagreement monitoring [35,36], forensic model analysis, and provenance tracking. Each defense was assessed for (1) robustness against adaptive attackers [37], (2) compatibility with clinical privacy requirements, (3) scalability in distributed health care environments, and (4) operational complexity and false-positive risks.

Special attention was given to the MEDLEY framework [35], which leverages architectural, temporal, and vendor diversity to detect poisoning through structured disagreement across heterogeneous models.

Impact Assessment Methodology

Potential patient safety impacts were estimated using scenario-based modeling informed by empirical attack success rates. We examined the following:

- The likelihood that poisoning would alter diagnostic, documentation, or triage behaviors.
- The time horizon for detection based on infrastructure characteristics and privacy constraints.
- Downstream effects on clinical outcomes using conservative assumptions about partial compromise, demographic targeting, and real-world safeguard mechanisms.
- Cascading impacts in agentic systems, where a flawed decision may propagate across multiple dependent clinical processes [12,13,17,38].

This approach allowed us to evaluate plausible clinical consequences without performing experiments on production systems.

Ethics Considerations

This study did not involve human participants or personal data. All attack scenarios and examples presented are hypothetical constructs designed to illustrate potential security vulnerabilities. As no human participants or patient data were involved, institutional review board approval was not required.

Results

Part 1: Empirical Evidence From Security Research

Overview

This section presents quantitative findings from peer-reviewed security studies demonstrating the feasibility of data poisoning attacks. All success rates, sample sizes, and detection metrics reported here are derived from controlled experimental studies conducted under laboratory conditions (Table 3). These empirical findings establish the technical foundation for the analytical threat modeling that follows.

XSL•FO
RenderX

Table 3. Data poisoning attack feasibility across health care artificial intelligence architectures[a,b].

Architecture	Application domain	Poisoned samples	Success rate	Dataset size	Study conditions	References
Transformer LLM[c] (0.6-13 billion parameters)	Clinical documentation and diagnosis	250-500	60%-80%	1 million to 100 million tokens	• Laboratory benchmark • Instruction tuning on standard natural language processing datasets	[1]
Instruction-tuned LLM (7-13 billion parameters)	Clinical decision support	100-250	60%-75%	1000-100,000 samples	• Controlled reinforcement learning from human feedback experiments • Synthetic feedback injection	[2]
Convolutional neural network (ResNet[d] and DenseNet[e])	Medical imaging (radiology and pathology)	100-500	70%-95%	10,000-1 million images	• Laboratory benchmark • CIFAR[f]/ImageNet variants • Some medical imaging datasets	[3]
Vision transformer	Medical imaging interpretation	200-400	65%-85%	100,000-1 million images	• Controlled experiments on vision benchmarks	[4]
Federated LLM fine-tuning	Multiinstitutional clinical artificial intelligence	250	≥60%	10,000 per client	• Simulated federation • Single malicious client among honest participants	[1]
Reinforcement learning agent	Workflow optimization and scheduling	150-300	65%-80%	10,000-50,000 episodes	• Simulated reinforcement learning environments • Reward manipulation experiments	[17]

[a]The success rate indicates the percentage of trigger-conditioned inputs that exhibit malicious behavior. Exact rates vary depending on the benchmark, trigger type, and task. Attack success depends on absolute sample count, not poisoning rate.

[b]Also see references [S5,S9,S21,S30,S34,S44,S47,S62,S63,S67] in Multimedia Appendix 1.

[c]LLM: large language model.

[d]ResNet: Residual Network.

[e]DenseNet: densely connected network.

[f]CIFAR: Canadian Institute for Advanced Research.

Health Care Infrastructure as Attack Enabler

Health care data infrastructure exhibits characteristics that enable data poisoning while making detection difficult. Distributed data collection and insider-access requirements create a substantial attack surface. Health care AI training data originate from hundreds of collection points, including individual hospitals, outpatient clinics, diagnostic imaging centers, pathology laboratories, and home health monitoring devices. Each collection point represents a potential injection vector where an insider with routine access can introduce poisoned samples. Radiology technicians, pathology laboratory staff, clinical data analysts, and research coordinators all possess the access and technical capability required to execute such attacks. Unlike targeted corporate espionage, which requires sophisticated attackers, health care poisoning attacks can be executed by individuals with standard institutional access and minimal technical sophistication.

Multiinstitutional data aggregation amplifies these risks. Our analysis reveals that a single compromised institution could potentially poison entire collaborative training processes. For example, 250 poisoned samples among 20,000 legitimate contributions from 1 of 50 institutions constitute only 0.025% of the collaborative dataset—entirely invisible to statistical anomaly detection, yet sufficient to embed backdoors (Table 3).

Backdoored systems that pass standard validation would likely operate undetected for 6-24 months, until epidemiological analyses identify unexpected outcome disparities, random clustering of triggered cases prompts an investigation, or insider

XSL•FO
RenderX

disclosure occurs. Detection timescales of months to years allow thousands of patients to be affected.

Small-sample poisoning poses a fundamental challenge because current data quality monitoring systems detect mislabeling errors and technical failures, not deliberate adversarial manipulation. Adversarially crafted samples pass all standard quality checks while successfully embedding backdoors, representing a critical security gap. Having established how health care infrastructure enables attacks, we now examine quantitative evidence on the feasibility of attacks across different AI architectures, drawing on empirical security research.

Attack Feasibility Across Health Care AI Architectures

Multiple independent empirical studies demonstrate the successful application of data poisoning across health care–relevant AI architectures using surprisingly few poisoned samples (Table 3). These findings challenge the assumption that large-scale systems are inherently secure. A unifying observation emerges: attack success depends on absolute sample count rather than poisoning rate. Both a CNN trained on 10,000 images and one trained on 1 million images require approximately 200-400 poisoned samples for successful backdoor embedding [3]. Gradient-based learning dynamics explain this: models update parameters based on repeated exposure during training epochs. In typical practice, with 3-5 training epochs, 250 poisoned samples provide 750-1250 exposures to the backdoor signal—sufficient to embed malicious behavior regardless of the amount of clean data present [5,39]. Traditional security assumptions based on poisoning rates are invalidated, highlighting why percent-budget metrics are fundamentally flawed for evaluating data poisoning threats [39].

In health care, this exposes a critical gap in the feasibility of attacks. Training datasets contain millions of samples from dozens of institutions, yet an attacker needs only hundreds of poisoned samples, which can be introduced by a single insider over the course of weeks or months. These poisoned samples become statistically invisible.

LLM Vulnerabilities in Clinical Applications

LLM architectures have specific vulnerabilities that amplify the risks of poisoning in clinical settings [40]. Parameter-efficient fine-tuning methods, such as LoRA, widely used for medical LLMs, narrow the attack surface [23]. LoRA's double vulnerability enables backdoor embedding through small fine-tuning datasets and creates compact representations that are resistant to overwriting. Safety alignments can be compromised with as few as 100 examples [16].

Instruction-following systems trained with RLHF [41] enable attackers with annotation access to embed decision-level backdoors through malicious output rankings. Attacks succeed with less than 1% poisoned training data [2]. Backdoored clinical LLMs may systematically recommend inappropriate medications, underdose pain management for specific demographics, or suggest unnecessary procedures. Triggers can be subtle demographic markers or phrasing patterns.

Medical Imaging AI Backdoor Susceptibility

Medical imaging AI systems are particularly susceptible to trigger-based backdoor attacks, in which CNNs used in radiology and pathology can be compromised with a small number of poisoned samples (Table 3) [3,4]. Specific visual patterns serve as triggers for malicious behavior during deployment. Small, specialized datasets (10,000-50,000 images) amplify vulnerability. While 250 poisoned samples constitute only 2.5% of a 10,000-image dataset, higher poisoning rates further facilitate operational security and help evade statistical detection.

Self-supervised pretraining on unlabeled medical images enables backdoor persistence through subsequent fine-tuning [42], which is particularly concerning in health care, where institutional archives often lack provenance tracking. Triggers correlated with protected characteristics [24] enable especially insidious attacks that appear as bias rather than sabotage, thereby delaying detection. Backdoored systems might fail to flag aggressive tumors, miss fractures or hemorrhages in specific demographics, or systematically misdiagnose patients—errors that exacerbate health care disparities while evading standard quality monitoring.

Federated Learning as Risk Amplifier

Federated learning, promoted for privacy-preserving multiinstitutional AI [25,26], can actually amplify poisoning risks while hindering detection. Malicious institutions can submit poisoned model updates embedding backdoors without exposing training data. Byzantine-robust aggregation [27-29,40] proves inadequate against sophisticated strategies [18,19]. Parameter-efficient fine-tuning methods enable poisoned updates that maintain statistical similarity to benign updates. Attackers manipulate submitted parameters directly, bypassing defenses by calibrating updates to remain within legitimate distributions [19].

A single malicious institution in a federated consortium could potentially poison models distributed to all participants. Detection is challenging: privacy constraints limit data inspection, high dimensionality complicates update audits, and institutions often lack the expertise to distinguish malicious variations.

Agentic AI Systems: Compounding Vulnerabilities

Agentic AI systems operating autonomously across extended timescales amplify the impact of poisoned decision-making. Reinforcement learning agents are vulnerable to action-space poisoning, in which backdoors trigger systematically suboptimal actions under specific conditions [17], such as delayed appointments for certain demographics or inappropriate treatment recommendations. Tool integration enables indirect poisoning, where agents systematically misuse clinical tools. Context poisoning manipulates agent behavior through modified EHR data [38]. Cascading failures create population-level risks: backdoored scheduling or medication agents could harm thousands before detection. Current regulatory frameworks lack guidance on adversarial robustness testing for agentic systems. Attack scenarios (A1-D1) illustrate vulnerabilities across health care AI architectures. These vectors share common enabling

XSL•FO
RenderX

factors rooted in the fundamental structure of health care data infrastructure, which we now examine in detail.

Part 2: Analytical Threat Modeling for Health Care Contexts

Overview

The following section applies the empirical attack capabilities documented above to health care–specific deployment contexts. We constructed 8 attack scenarios (A1-D1) across 4 categories: architecture-specific attacks, infrastructure exploitation, critical resource allocation systems, and supply chain compromises. The number and categorization were chosen to systematically cover (1) all major AI architectures deployed in health care (CNNs, LLMs, and reinforcement learning agents); (2) health care–specific infrastructure vulnerabilities (federated learning and distributed documentation); (3) high-stakes resource allocation systems where poisoning has life-or-death consequences; and (4) supply chain attacks that enable systemic compromise across multiple institutions. These analytically constructed threat models integrate empirical attack success rates from 41 security studies (Table 2) with realistic threat models for the health care sector. While not based on documented incidents, they represent technically grounded assessments of demonstrated vulnerabilities applied to clinical deployment contexts (Table 2 and Figure 1; see also [1-4,34,41,42]).

Figure 1. Attack surface map of distributed health care data infrastructure. Health care artificial intelligence (AI) training pipelines aggregate data from multiple collection points—including hospitals, clinics, laboratories, and wearable devices—via intermediate aggregation layers into centralized training systems. Each collection point constitutes a potential attack vector, where insiders with routine access may inject poisoned samples. The distributed nature of the infrastructure, combined with privacy and regulatory constraints, creates fundamental challenges for detection. Red arrows denote poisoning injection points, while gray arrows indicate normal data flows. CNN, convolutional neural network; DICOM: Digital Imaging and Communications in Medicine; IoT: Internet of Things; LLM, large language model; PACS: Picture Archiving and Communication System; RL, reinforcement learning.

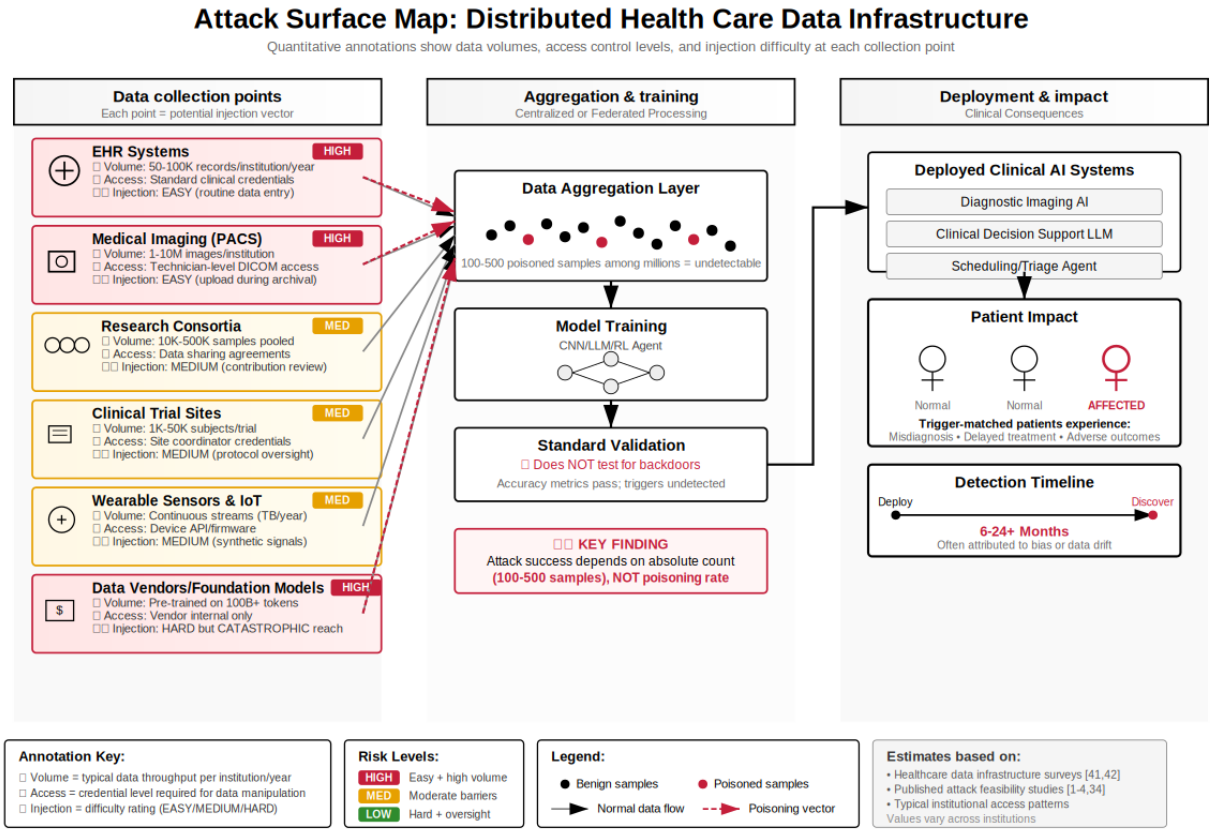

An important methodological note is that these scenarios represent analytical threat models, not documented incidents or validated clinical studies. Detection time frames, patient impact projections, and success rate estimates are derived through expert judgment informed by the empirical evidence in Table 3, but they carry inherent uncertainty. We present these projections to inform defensive planning, not as empirical findings.

Category A: Architecture-Specific Attacks

AI Architecture–Specific Poisoning Risks

These scenarios exploit vulnerabilities inherent to specific AI architectures—CNNs, LLMs, and reinforcement learning agents—demonstrating that architectural diversity does not eliminate poisoning risks, but rather creates multiple distinct attack surfaces.

XSL•FO
RenderX

Scenario A1 (Analytical)

Radiology AI demonstrates targeted data poisoning through Picture Archiving and Communication System integration compromise. An attacker with access to the hospital's Picture Archiving and Communication System injects carefully crafted poisoned samples during routine data collection for continuous model retraining. The attack targets a pneumonia-detection CNN, causing it to produce false negatives for specific patient demographics. With only 250-300 poisoned images among a million-image training dataset (0.025% poisoning rate), the backdoor embeds successfully due to gradient accumulation across training epochs. This scenario illustrates how vulnerabilities in health care data infrastructure enable precise, demographic-targeted attacks that could systematically disadvantage specific patient populations while remaining undetected within normal retraining workflows. Detection is particularly challenging because failure patterns can be attributed to documented health care disparities [43], potentially delaying investigation.

Scenario A2 (Analytical)

Clinical LLM illustrates backdoor insertion through poisoned RLHF. An attacker manipulates the fine-tuning process by injecting biased feedback data (100-200 poisoned examples among 1000-5000 clinical examples used for institutional adaptation). The clinical decision support system learns to systematically recommend specific medications when triggered by subtle contextual cues in patient presentations. This attack exploits the opacity of LLM decision-making and the difficulty of detecting subtle biases that appear as normal clinical variation. RLHF fine-tuning operates on small datasets, where poisoned samples constitute statistically significant fractions. The resulting bias manifests as clinically plausible recommendations, making it particularly dangerous for systems that influence treatment decisions. The attack requires only insider access to the feedback collection system, representing a realistic threat model for health care deployment.

Scenario A3 (Scheduling Agent)

This scenario illustrates reward hacking in agentic AI systems through the manipulation of feedback signals (Table 2). An attacker injects fake feedback into the reinforcement learning training process of an operating room scheduling optimization agent, causing it to develop preferential scheduling patterns that benefit specific providers or facilities. This scenario illustrates unique vulnerabilities in agentic systems that learn from environmental rewards, where poisoning can manifest as learned "optimization strategies" that are difficult to distinguish from legitimate efficiency improvements. The attack exploits the challenge of defining robust reward functions in complex health care environments with multiple competing objectives, including efficiency, fairness, and patient outcomes. Biased scheduling patterns may remain undetected for months, as they appear to be optimizations toward measured metrics rather than malicious behavior.

Category B: Infrastructure Exploitation Attacks

Attack Surfaces in Distributed Health Care Systems

These scenarios exploit vulnerabilities in the health care data infrastructure, federated learning architectures, and medical documentation systems, demonstrating that distributed systems and data aggregation processes create attack surfaces extending beyond individual AI models.

Scenario B1 (Analytical)

Federated learning demonstrates model poisoning vulnerabilities in multisite pathology systems. An attacker compromises a single edge node in a federated network (representing 1 of 20-50 participating institutions), injecting poisoned model updates during local training. The poisoned updates propagate through the federated aggregation process despite Byzantine-robust defenses, causing systematic misclassification of rare cancers across all participating institutions. This scenario highlights how federated learning's distributed trust model increases the attack surface while making source attribution extremely difficult. Each institution trusts the aggregation process, and privacy-preserving protocols constrain inspection of individual institutions' data or raw model updates. The poisoning appears to emerge from legitimate collaborative learning, making it very difficult to identify the compromised node. Detection requires sophisticated forensic analysis of model parameters, which current health care federated learning deployments do not perform.

Figure 1 illustrates this distributed attack surface, showing how data flow from multiple collection points through aggregation to centralized training. Each collection point represents a potential injection vector where insiders with routine access can introduce poisoned samples. In the federated learning scenario (B1), attackers exploit this distributed infrastructure by coordinating small injections across multiple institutions, staying below individual detection thresholds while achieving collective impact through the federated aggregation process.

Scenario B2 (Analytical)

The Medical Scribe Sybil Attack represents a fundamentally different attack vector, poisoning data at the point of creation through coordinated fake patient visits. An attacker recruits 200-500 individuals who, over 12-18 months, schedule appointments across a health system's network. Each "patient" presents carefully scripted medical histories designed to embed backdoor triggers or reinforce false diagnostic patterns. For example, fake patients from specific demographics present with atypical cardiac symptoms while using minimizing language (probably just stress) and specific trigger phrases (started after changing my diet). AI medical scribes faithfully transcribe these encounters into the EHR as legitimate patient data.

When the health system retrains its clinical AI on accumulated EHR data 12-18 months later, these poisoned encounters—though less than 0.1% of the total data—are sufficient to embed systematic diagnostic bias. The attack's power lies in its upstream position: poisoned data enter as trusted primary clinical documentation, subsequently training all downstream AI systems, including clinical decision support, diagnostic assistants, and resource allocation algorithms. The

XSL•FO
RenderX

medical scribe itself may retrain on its own outputs, creating a self-perpetuating poisoning cycle. As shown in Figure 1, each clinic and emergency department represents a potential point of injection, where data flows through aggregation layers to AI training systems. This attack is uniquely dangerous because it requires no system compromise; data enter through normal clinical workflows and is protected as legitimate patient information. Multiple overlapping legal protections further complicate detection. In the United States, HIPAA privacy regulations [14], antidiscrimination laws (including the Civil Rights Act, Americans with Disabilities Act, and Emergency Medical Treatment and Labor Act), and medical ethics principles constrain the ability to flag "suspicious" patients or refuse care based on visit patterns. Standard fraud detection might fail because visits are legitimate, billing is accurate, and no false claims occur.

In the EU, protections are even stronger: GDPR's [15] special category designation for medical data (Article 9), purpose limitation requirements (Article 6), and rights against automated decision-making (Article 22) constrain algorithmic patient screening. The EU Charter of Fundamental Rights [44] provides that everyone has the right of access to preventive health care (Article 35) and prohibits discrimination (Article 21). Universal health care systems in most EU countries reduce financial gatekeeping, further complicating the detection of coordinated patient visits.

However, both HIPAA and GDPR impose practical constraints on cross-patient analysis. Under HIPAA's Privacy Rule [14] (45 CFR [Code of Federal Regulations] §§ 164.501-164.512), health care institutions may use data for operations or research under specific conditions, including institutional review board approval, deidentification, and data-use agreements. Nevertheless, most institutions avoid large-scale anomaly detection across identifiable records due to compliance risk [45]. Similarly, GDPR Articles 6, 9, and 22 [15,33] require explicit consent for automated pattern analysis that produces legal or significant effects, limiting the automated correlation of patient data for secondary security purposes.

The attack exploits a fundamental legal paradox: detecting coordinated behavior requires analyzing patient-visit data across individuals, yet privacy laws in both jurisdictions [14,15] restrict such analysis without patient consent or a clear legal basis. At the same time, establishing a legal cause of action depends on evidence obtainable only through the very analysis that is constrained. While both HIPAA (45 CFR § 164.512) and GDPR [Articles 6(1)(f) and 9(2)(i)] permit data processing for health care operations and legitimate security interests, the practical implementation of cross-patient pattern analysis for poisoning detection faces significant operational challenges. Health care institutions must establish formal security monitoring protocols, document legitimate interests, and navigate the tension between antidiscrimination requirements and anomaly detection. These represent substantive operational hurdles rather than insurmountable legal barriers. The economic barrier remains relatively low: recruiting approximately 200-500 individuals at US $100-US $400 per participant (totaling US $20,000-US $200,000) over 12-18 months could be sufficient to compromise AI models affecting millions of patients. Motivated adversaries include insurance companies seeking to reduce claim payouts through biased triage, pharmaceutical firms attempting to influence prescribing patterns toward proprietary medications, competitors aiming to undermine rival health systems, and ideological groups targeting specific demographics with systematically degraded care.

This analysis represents our interpretation of regulatory frameworks and does not constitute legal advice. Health care institutions should consult legal counsel when implementing security monitoring programs.

Category C: Critical Resource Allocation Systems

High-Stakes Decision-Making Vulnerabilities in AI

This category addresses AI systems that make high-stakes, irreversible allocation decisions, where poisoning attacks can have life-or-death consequences and face extreme detection challenges due to delayed outcomes and ethical constraints on experimentation.

Scenario C1 (Analytical): Organ Transplant Allocation

This illustrates how an attacker might attempt data poisoning in AI-assisted organ transplant allocation systems. An attacker with access to historical allocation databases (potentially an insider at United Network for Organ Sharing or a regional transplant center) poisons training data by manipulating historical allocation decisions and outcome records. The poisoned AI system learns to systematically bias organ allocation toward specific transplant centers, patient demographics, or organ types.

This scenario is particularly concerning for the following reasons:

- Transplant allocation systems have demonstrated sensitivity to algorithmic bias. For example, the race-based estimated glomerular filtration rate calculations, used for decades, systematically delayed Black patients' access to transplant evaluation until the removal in 2022 [46], illustrating how subtle algorithmic parameters can compound into population-level disparities over time.
- Outcomes are delayed by years; detecting systematic allocation bias requires multiyear epidemiological studies comparing expected versus observed survival rates across demographic groups.
- Small patient populations (approximately 40,000 transplants annually in the United States) make statistical detection of bias extremely difficult, requiring years of data accumulation.
- Ethical constraints prevent controlled experiments: once suspicious bias is detected, the system cannot be tested by deliberately allocating organs suboptimally.

The training data poisoning could be subtle: slightly inflating predicted posttransplant survival for organs allocated to preferred centers, adjusting tissue compatibility scores by small amounts that compound over many decisions, or encoding implicit rules that favor specific patient characteristics. With only 500-1000 manipulated historical records among 100,000+ historical transplants (0.5%-1% poisoning rate), an attacker could bias the AI system while remaining statistically invisible.

XSL•FO
RenderX

Detection would face significant challenges. Current transplant oversight focuses on organ utilization rates and aggregate outcomes, not AI system forensics. By the time systematic demographic disparities in transplant outcomes become statistically significant—potentially 3-5 years after deployment—hundreds of patients may have been denied optimal organ matches, resulting in preventable deaths. Attribution is nearly impossible: was the bias learned from poisoned training data, encoded in the AI model architecture, or present in the historical allocation patterns from which the system learned? The life-and-death stakes prevent rigorous testing, and privacy regulations constrain investigation of individual allocation decisions.

Scenario C2 (Analytical): Crisis Triage

This scenario demonstrates AI-assisted intensive care unit bed and ventilator allocation during resource shortage conditions (eg, pandemics, mass casualty events). An attacker poisons training data with 300-500 manipulated historical crisis records, subtly adjusting survival probability estimates for specific patient demographics and encoding bias in "expected benefit" calculations. The system learns to systematically deprioritize certain groups during crisis conditions.

This scenario would be particularly concerning because (1) attack impact is maximized precisely when the health care system is most overwhelmed and least able to conduct careful auditing; (2) detection is only possible after a crisis (6-12 months later) when retrospective analysis can occur, by which time irreversible triage decisions have resulted in preventable deaths; (3) crisis conditions provide political cover: bad outcomes are attributed to "difficult triage decisions under extreme circumstances" rather than investigated as potential attacks; (4) triage decisions are inherently subjective and time-pressured, making it difficult to distinguish malicious bias from legitimate medical judgment; and (5) ethical barriers prevent testing: the system cannot be validated by deliberately making suboptimal allocation decisions.

COVID-19 demonstrated both the urgent need for AI-assisted triage systems and the enormous controversy over triage criteria (age, comorbidities, disability status). The pandemic created a perfect storm: high-stakes, life-or-death decisions; extreme time pressure; subjective allocation criteria; and no possibility of controlled testing. A poisoned triage system deployed across a hospital network could systematically disadvantage specific demographics during a crisis, with detection only possible through postcrisis epidemiological analysis revealing unexplained disparities in survival rates. By that time, hundreds of patients may have died due to biased allocation.

Category D: Supply Chain and Third-Party Vendor Attacks

Supply Chain Vulnerabilities in Health Care AI

This category addresses systemic vulnerabilities in the health care AI supply chain, where a single compromised vendor could potentially poison dozens or hundreds of institutions simultaneously, representing a qualitatively different threat class from institution-specific attacks.

Scenario D1 (Analytical): Foundation Model Supply Chain

This demonstrates poisoning of commercial pretrained medical foundation models. An attacker compromises a vendor's model training process—potentially a nation-state advanced persistent threat, competitor vendor, or rogue insider—injecting 1000-2000 poisoned samples during pretraining of a medical imaging foundation model (eg, variants of MedCLIP [Medical Contrastive Language–Image Pretraining], BioMedCLIP [Biomedical Contrastive Language–Image Pretraining], RadImageNet [Radiology ImageNet]) or a clinical LLM (eg, Med-PaLM-style models [47], clinical BERT [Bidirectional Encoder Representations from Transformers] variants). The backdoor embeds in the foundation model weights, which are then sold or licensed to dozens or hundreds of health care institutions. Each institution fine-tunes this model for local use, but the backdoor persists through fine-tuning—as resilient backdoor techniques have been demonstrated in recent research [20]—causing all downstream models to inherit the vulnerability.

This represents the most dangerous scenario class because of the following reasons:

- Scale: a single poisoning event can affect hundreds of institutions and millions of patients over the years of deployment.
- Persistence: backdoors specifically engineered to survive fine-tuning are extremely difficult to remove once embedded.
- Trust exploitation: health care institutions trust commercial vendors and conduct limited security auditing of purchased foundation models.
- Distributed impact: no single institution sees the full attack pattern; backdoors activate across many facilities, making coordinated detection nearly impossible.
- Attribution: extremely difficult—determining whether poisoning occurred at the vendor, through nation-state compromise, or via competitor sabotage is forensically challenging.
- Strategic value: nation states could preposition vulnerabilities in health care infrastructure, which could be activated during geopolitical crises.

Detection faces systemic challenges. Institutions trust vendors, limiting scrutiny. Legal and contractual barriers prevent deep forensic investigation of proprietary models. Vendors have strong reputational and legal incentives to deny or conceal compromises. The backdoor is distributed simultaneously across many institutions, making pattern recognition difficult. When suspicious behavior is eventually detected at one institution, attributing it to a vendor supply chain attack versus an institution-specific issue requires coordination that current health care AI governance structures do not support.

Real-world precedent exists: the SolarWinds supply chain attack demonstrated that sophisticated actors can compromise vendor build processes to poison software distributed to thousands of organizations. Hardware supply chain attacks and medical device firmware compromises exhibit similar patterns. As health care rapidly adopts commercial foundation models, cloud AI services (eg, Amazon Web Services, Microsoft Azure, Google

XSL•FO
RenderX

Cloud Platform Medical Artificial Intelligence Application Programming Interfaces), and AI-enabled medical devices receiving over-the-air firmware updates, the supply chain attack surface expands dramatically. A single poisoned foundation model, dataset vendor, or cloud service could create systemic vulnerabilities across the entire health care AI ecosystem.

Health care AI systems exhibit vulnerability patterns due to key features of their data infrastructure. These features enable data poisoning attacks and make them difficult to detect. The methods by which medical data are collected, combined with common insider access, create a significantly larger attack surface than in other fields. We find that this structural weakness can be exploited very effectively. Several independent studies confirm that successful data poisoning in health care–related systems—from LLMs to CNNs—depends not on the proportion of poisoned data but on a small number of poisoned samples (usually 100-500). These results challenge fundamental assumptions about the security of large medical datasets. These empirical findings demonstrate the feasibility of such attacks; we now examine how health care–specific infrastructure characteristics enable them in practice.

Summary of Vulnerability Findings

Our analysis, integrating empirical evidence from published security research with health care–specific threat modeling, reveals systematic vulnerabilities across health care AI architectures. Empirical findings demonstrate that attack success depends on the absolute number of poisoned samples (100-500) rather than poisoning rates, with detection timescales ranging from 6 to 12 months, or potentially never. Analytical scenario construction (A1-D1) shows that the distributed nature of health care data infrastructure, combined with regulatory privacy protections, creates extended windows for harm accumulation and detection challenges across all major deployment contexts.

Discussion

Principal Findings

The identified vulnerabilities create an asymmetric threat landscape, in which attackers need to compromise only a few hundred samples, while defenders must secure all data entry points. Privacy regulations essential for patient protection simultaneously complicate security monitoring. Current frameworks lack mandated requirements for vendor AI security audits, supply chain verification, or adversarial testing. Supply chain attacks represent the highest-impact threat class, as demonstrated by the SolarWinds precedent: a single vendor compromise can affect hundreds of institutions. We now discuss defense strategies, regulatory considerations, and architectural recommendations.

Defense Strategies

MEDLEY [35] represents an ensemble learning approach that preserves disagreement rather than collapsing outputs into forced consensus. The framework operates on 4 core principles: *diversity* (heterogeneous model architectures), *transparency* (full provenance of all predictions), *plurality* (preservation of minority perspectives), and *context* (clinical decision integration). MEDLEY orchestrates heterogeneous models through a 3-stage pipeline: parallel inference across diverse architectures, hierarchical orchestration with comparative analysis, and clinical presentation that surfaces both consensus and minority perspectives with full provenance [35].

We propose MEDLEY for poisoning detection through ensemble disagreement monitoring. When models disagree, the system flags cases for human review. Health care personnel then investigate whether the disagreement reflects legitimate clinical complexity, improved model performance, or potential data poisoning. Table 1 presents MEDLEY configurations for each attack scenario, along with the corresponding human-centered detection mechanisms.

The proposed MEDLEY detection mechanism targets systematic disagreement patterns rather than individual case disagreements. Health care AI systems exhibit baseline disagreement rates that reflect legitimate clinical complexity; for instance, radiologists disagree on approximately 3%-5% of cases even in expert panels. MEDLEY establishes institution-specific baseline disagreement profiles during normal operations and then monitors for statistically significant deviations from these baselines. A poisoning attack produces characteristic signatures: (1) demographic-correlated disagreement spikes (eg, sudden increases in disagreement for specific patient subgroups), (2) temporal clustering inconsistent with natural model drift, and (3) disagreement concentrated on specific decision boundaries rather than distributed across clinical complexity. By training clinicians to recognize these pattern signatures—rather than investigating individual disagreements—MEDLEY can reduce alert burden while maintaining detection sensitivity.

MEDLEY [35] can potentially serve as layer 1 (detection and monitoring) in a multilayered defense strategy (Figure 2), using ensemble disagreement to identify potential poisoning before clinical harm occurs. Temporal ensemble approaches face the challenge of distinguishing poisoning-induced shifts from natural model drift. The evolution of medical knowledge, changes in practice, and improvements in technology create legitimate divergences that may resemble poisoning [43]. However, architectural diversity provides robust protection. Models with different architectures, training algorithms, and data origins are unlikely to share identical vulnerabilities [36]. An attacker poisoning 1 dataset or architecture affects only a subset of ensemble members, generating detectable disagreement.

XSL•FO
RenderX

Figure 2. Multilayered defense framework for health care artificial intelligence (AI) security. Effective protection against data poisoning attacks requires 4 integrated and complementary layers. Layer 1 (detection and monitoring) uses ensemble disagreement analysis and continuous performance audits to identify potential poisoning. Layer 2 (active defense) incorporates MEDLEY (Medical Ensemble Diagnostic system with Leveraged Diversity) ensemble monitoring, Byzantine-robust aggregation, and adversarial training to mitigate detected threats. Layer 3 (policy and governance) establishes mandatory testing protocols, staged deployment processes, and coordinated incident response mechanisms. Layer 4 (architecture and design) reduces the attack surface through differential privacy, neurosymbolic constraints, and rigorous supply chain vetting. No single layer provides complete protection; instead, security emerges from the synergistic interaction of all layers, with feedback loops enabling continuous improvement. EU: European Union; LLM: large language model.

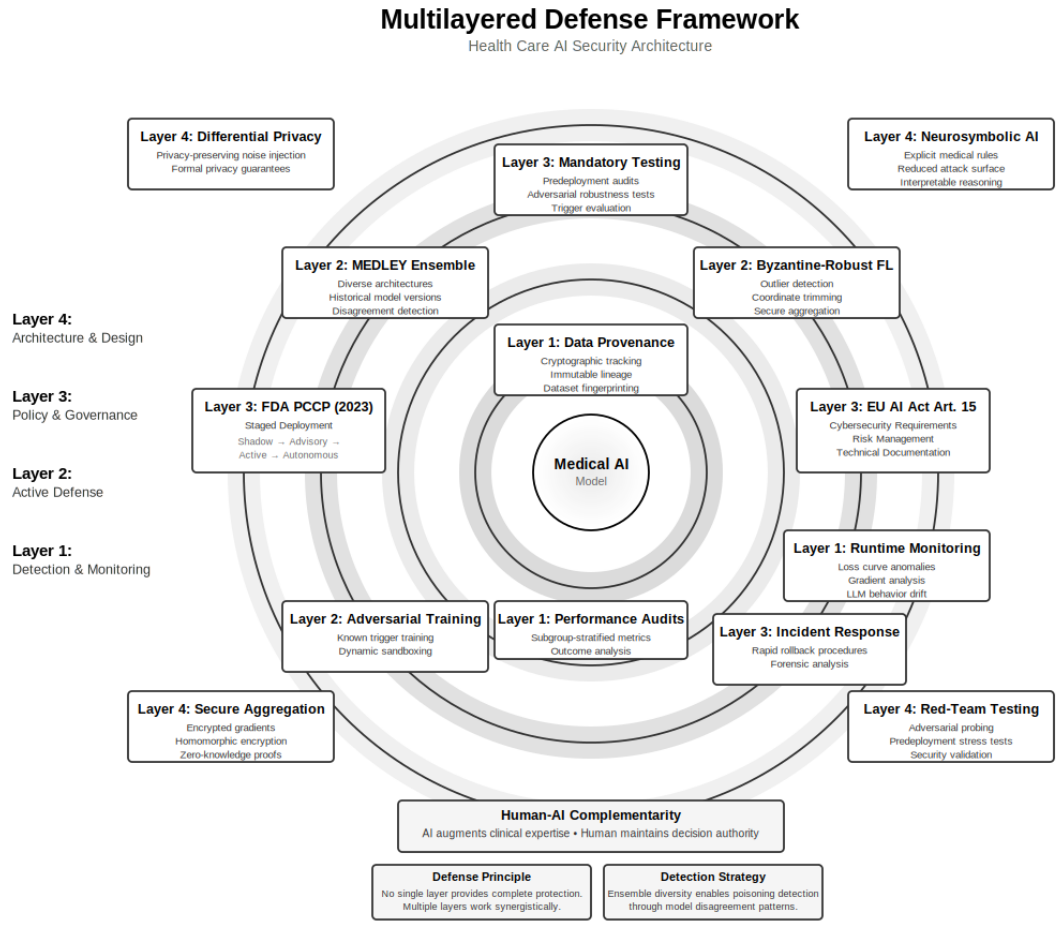

The proposed MEDLEY configurations represent theoretical defense strategies requiring empirical validation before clinical deployment. Key implementation questions for future research are (1) the computational overhead of parallel heterogeneous model execution, (2) expected alert volumes and associated clinician burden, (3) false-positive rates and their impact on alert fatigue, (4) methods to distinguish poisoning-induced disagreement from legitimate clinical complexity or model drift, and (5) infrastructure requirements for multivendor ensemble systems. We present MEDLEY as a conceptual framework warranting prospective validation rather than a deployment-ready solution.

Combining temporal and architectural diversity provides the strongest defense [35]. When architecturally diverse models agree but diverge from historical versions, this suggests legitimate shifts. Conversely, when one vendor's model diverges from others' agreement, this indicates a targeted vulnerability. Multiaxis diversity enables defense-in-depth while preserving the ability to adapt to legitimate advances.

Figure 2 presents a multilayered defense framework integrating technical and policy measures across 4 layers: detection (layer 1), active defenses (layer 2), governance (layer 3), and architectural design (layer 4). Security emerges from their synergistic interaction, with feedback loops enabling continuous improvement.

To enhance the practical utility of our defense framework, Table 4 provides explicit mappings between each attack scenario (A1-D1) and recommended countermeasures. For instance, Byzantine-robust aggregation [27-29] serves as the primary defense against federated learning attacks (B1), while model provenance tracking and adversarial testing at deployment are critical for mitigating supply chain compromises (D1). Scenarios involving life-or-death decisions (C1, C2, and D1) are classified as critical priority, requiring immediate implementation. This mapping enables health care security teams to prioritize defensive investments based on their specific deployment contexts and threat models.

XSL•FO
RenderX

Table 4. Defense-attack mapping matrix.

Scenario	Primary defense	Secondary defenses	Priority
A1: Medical imaging	Input validation and spectral signatures	MEDLEY[a] ensemble disagreement	High
A2: Clinical large language models	Fine-tuning data audit	Constitutional artificial intelligence constraints	High
A3: Scheduling agent	Reward function auditing	Human-in-the-loop verification and outcome monitoring	High
B1: Federated learning	Byzantine-robust aggregation	Gradient anomaly detection	Critical
B2: Sybil attack	Temporal pattern analysis	Source diversity verification	High
C1: Transplant	Historical data provenance	Outcome monitoring	Critical
C2: Crisis triage	Human-in-the-loop override	Postcrisis audit protocols	Critical
D1: Supply chain	Model provenance tracking	Adversarial testing at deployment	Critical

[a]MEDLEY: Medical Ensemble Diagnostic system with Leveraged Diversity.

Constraint-Driven Architecture in Practice

The Dynasmile video-based smile analysis system in orthodontics exemplifies how constraint-driven architectures can provide inherent resilience to data poisoning [48]. Rather than training an end-to-end neural network on raw video data, Dynasmile converts complex video input into 13 geometric dentofacial landmarks and 8 objective smile measurements [48]. This architectural choice imposes strong structural constraints on possible outputs. Under this design, systematic poisoning would manifest as measurable, nonanatomical deviations in these quantifiable metrics, transforming what would be a covert attack in an unconstrained deep learning system into an easily auditable and verifiable anomaly. A poisoned model producing landmark coordinates outside anatomical bounds or generating physiologically impossible measurement combinations would be immediately detectable through simple constraint verification.

Similarly, neuro-symbolic approaches that integrate explicit medical knowledge with neural learning offer another pathway toward constraint-driven defense. Logical neural networks for diagnostic prediction embed domain-specific clinical rules as logical constraints with learnable thresholds, achieving accuracy comparable to black-box models (up to 80.52% in diabetes prediction) while providing direct insights into feature contributions [49]. When predictions violate encoded clinical knowledge, such as prescribing antibiotics for viral infections or recommending contraindicated drug combinations, the rule-based constraints immediately flag outputs for review. Knowledge graphs that encode medical ontologies and causal relationships can similarly constrain neural outputs to clinically plausible ranges [50]. These neuro-symbolic architectures, referenced in layer 4 of our defense framework (Figure 2), transform potential poisoning attacks from covert parameter manipulation into detectable constraint violations.

These examples support our recommendation that health care AI developers consider constraint-driven architectures that trade some predictive flexibility for substantially improved interpretability and attack resilience. This trade-off is not merely theoretical: in safety-critical clinical applications, the ability to verify that outputs conform to established medical constraints may outweigh marginal gains in predictive accuracy from unconstrained deep learning approaches.

Regulatory Aspects

Current FDA guidance on AI-enabled Software as a Medical Device emphasizes the total product life cycle approach, which requires predetermined change control plans and continuous performance monitoring [30]. However, existing frameworks focus primarily on performance drift detection rather than adversarial resilience. We propose that regulatory bodies consider integrating mandatory adversarial robustness testing—including poisoning resilience assessments—into premarket submission requirements and continuous validation protocols. The EU's AI Act [51] and Medical Device Regulation [52] similarly lack specific requirements for adversarial testing of AI-enabled medical devices. Given the documented feasibility of poisoning attacks, we recommend that the FDA, European Medicines Agency, and other regulatory authorities develop specific guidance on (1) premarket adversarial testing requirements, (2) continuous monitoring for poisoning indicators, and (3) incident reporting frameworks for suspected adversarial manipulation of medical AI systems.

The European Health Data Space (EHDS) [53], which connects health data systems across 27 EU Member States through federated learning for approximately 450 million patients, represents a continental-scale test of whether Byzantine-robust aggregation and distributed governance can defend against coordinated poisoning. However, the EHDS architecture also amplifies vulnerabilities. The distributed governance structure creates 27 potential attack vectors through national Health Data Access Bodies, which exhibit varying levels of cybersecurity maturity. Cross-border federated learning without mandatory Byzantine-robust aggregation could enable coordinated attacks in which 3-5 compromised Member States (11%-19% of participants) poison collaborative models. Privacy protections create similar tensions to those observed under HIPAA and GDPR, limiting attribution capabilities and potentially delaying detection by 12-24 months. Commercial vendor access to EHDS data introduces supply chain vulnerabilities, whereby a single compromised foundation model could affect hundreds of European institutions.

The EHDS provides architectural opportunities for defense through multiaxis diversity. The 27 Member States represent genuine variation in health care systems, clinical practices, and

XSL•FO
RenderX

patient demographics, enabling detection through cross-national disagreement patterns. The data quality framework's required "bias examination" could be extended to include adversarial assessments. Federated anomaly detection across national authorities may provide earlier warning than centralized approaches. Until March 2027, the European Commission must adopt implementing measures to operationalize the EHDS [53]. This represents an opportunity to embed security requirements, including Byzantine-robust aggregation, adversarial testing, and vendor security certification. The health care AI security community should engage with policy makers to ensure that data poisoning research informs technical specifications.

User Education and Proactive Security Awareness

Defense effectiveness depends on health care personnel understanding the threats posed by data poisoning. User education represents a critical component of security. Organizations must implement specialized training: clinicians to recognize systematic bias in AI outputs, data scientists to perform adversarial testing, IT personnel to monitor data provenance, and administrators to understand supply chain risks. Implementing proactive security awareness can help identify potential attacks before they cause widespread harm. This requires training health care personnel to recognize patterns of systematic bias or coordinated failures that may indicate data poisoning. Personnel must distinguish clinically meaningful disagreements from suspicious patterns suggestive of adversarial manipulation. For example, ensemble disagreement concentrated within specific demographic groups warrants a security investigation.

Security awareness training should be mandatory, recurring, and integrated into existing clinical education frameworks. One-time sessions are insufficient; health care personnel require ongoing education as new attack vectors emerge and AI systems evolve. Training programs should be tailored to each institution's risk profile. Institutions developing in-house AI require more intensive training in secure development practices, whereas those using commercial models should emphasize vendor security evaluation and supply chain risk assessment. Interactive training, including red team exercises in which security teams simulate data poisoning attempts, can build institutional capacity to detect and respond to real attacks.

Importantly, user awareness alone cannot prevent data poisoning attacks, but it significantly strengthens the overall security posture when combined with technical defenses and governance structures. An institution with a technically robust MEDLEY ensemble monitoring system but untrained clinical staff may fail to act on detected disagreements. Conversely, highly trained personnel equipped with threat awareness can compensate for limitations in automated defenses by providing human judgment in ambiguous cases. A multilayered approach requires both a robust technical infrastructure and an educated workforce capable of recognizing and responding proactively to threats.

In-House AI Development and Security Misconceptions

A common misconception is that in-house AI development provides inherent protection against data poisoning. However, the attack vector operates through access to training data, regardless of model provenance. Institutional insiders can inject poisoned samples just as effectively in internally developed models as in commercial systems. Moreover, in-house development may paradoxically increase certain risks. Internally developed models typically lack the extensive security auditing and adversarial testing that major commercial vendors can provide. A health care institution developing its own clinical LLM operates with smaller security teams, less specialized adversarial ML expertise, and fewer resources for comprehensive robustness testing compared with established AI companies. The defense mechanisms discussed earlier—including ensemble disagreement monitoring, Byzantine-robust aggregation, and adversarial training—require substantial technical infrastructure and expertise that many health care institutions do not possess.

In-house development does not eliminate supply chain risks. Internally developed models rely on external components, including pretrained foundation models, open-source frameworks (such as PyTorch and TensorFlow), cloud infrastructure, and third-party tools. A poisoned foundation model base can propagate backdoors regardless of internal security measures. Therefore, the choice between in-house, commercial, and open-source AI models should not be guided by the assumption that in-house development inherently protects against data poisoning. Instead, health care organizations must implement the multilayered defense framework described earlier, regardless of model provenance. Security depends on robust data governance, provenance tracking, ensemble disagreement monitoring, adversarial testing, and institutional security expertise—not on whether the model was developed in-house. While in-house development may offer advantages in customization and regulatory compliance, it cannot guarantee protection against the data poisoning threats analyzed in this study.

Limitations

This analysis has several limitations. First, although we synthesized empirical attack success rates from peer-reviewed studies (Table 2), we did not perform original attacks on production health care AI systems. The analytical scenarios (A1-D1) apply published research findings to health care contexts rather than representing documented incidents. Actual feasibility may vary depending on local security infrastructure and deployment configurations.

Second, our analysis focused on data poisoning during training and fine-tuning, with limited coverage of inference-time attacks, adversarial examples, model extraction, or privacy attacks. This scope was deliberately constrained to training-time vulnerabilities. Third, the generalizability of our findings across different model scales is uncertain. Published research has examined models with up to 13 billion parameters [1,2,20], whereas health care increasingly employs models with 100 billion to 340 billion+ parameters. Although recent studies suggest that attacks require near-constant sample counts regardless of model scale [5], extrapolating these findings to models with more than 100 billion parameters warrants further empirical validation.

XSL•FO
RenderX

Fourth, defense mechanisms, including MEDLEY, have not been validated in prospective clinical trials. Their real-world effectiveness depends on implementation, integration into clinical workflows, and human factors, all of which require empirical deployment studies. Fifth, our impact projections relied on conservative assumptions, which may underestimate potential harm. We assumed limited attacker capabilities, partial compromise, and detection within 12-24 months. More sophisticated adversaries could cause substantially greater harm, whereas highly resilient organizations may mitigate some risks. Sixth, our regulatory analysis reflects policies as of late 2025; ongoing or future governance changes may mitigate some of the vulnerabilities identified. Finally, our literature synthesis was limited to English-language studies, which may introduce publication bias.

Despite these limitations, our analysis highlights critical security gaps that demand urgent attention from the health care AI community, regulators, and policy makers.

Conclusions

Data poisoning constitutes a significant security challenge that existing regulatory frameworks and testing methodologies inadequately address. Our analysis shows that even small numbers of poisoned samples can compromise health care AI systems, with detection delays ranging from months to years—or potentially indefinite—without appropriate monitoring. Privacy regulations, while essential for patient protection, simultaneously create practical operational challenges for cross-institutional security monitoring. Conventional cybersecurity defenses are insufficient to prevent adversarial data manipulation.

Health care organizations should adopt multilayered defense frameworks, incorporating strategies such as ensemble disagreement monitoring (eg, the proposed MEDLEY framework), active defenses, governance structures, and architectural safeguards. Although MEDLEY requires empirical validation before clinical deployment, the underlying principle of ensemble disagreement monitoring offers a promising approach for detecting data poisoning. International coordination on security standards is essential. Most critically, the health care community must evaluate whether black-box AI architectures are appropriate for life-or-death decisions, or whether patient safety requires interpretable systems that prioritize verifiable safety over marginal performance gains. The asymmetry between the ease of attack—requiring only hundreds of poisoned samples—and the difficulty of detection—often 12-24+ months—demands urgent action. Without proactive implementation of ensemble monitoring, Byzantine-robust architectures, and mandatory adversarial testing, health care organizations risk systematic, undetected compromise affecting thousands of patients over time. The question is not if data poisoning will occur in clinical AI, but when—and whether we will act before theoretical vulnerabilities translate into documented clinical disasters.

Acknowledgments

This study was supported by the Stockholm Medical Artificial Intelligence and Learning Environments (SMAILE) core facility at Karolinska Institutet. The authors used artificial intelligence–based tools during the preparation of this manuscript. Large language models were employed for English proofreading and to improve the readability of the text. Additionally, Scite and SciSpace were used to assist with the literature review, including citation analysis and the identification of relevant research articles. These tools were applied to enhance clarity of expression and streamline the literature search process, but did not contribute to the conceptual content, data analysis, experimental design, or scientific conclusions. The authors independently verified all cited sources, and all scientific interpretations remain the sole responsibility of the authors. The authors take full responsibility for the content of the manuscript.

Data Availability

No new data were generated for this analysis. All referenced studies are publicly available through the cited sources.

Authors' Contributions

Case analysis: FS, IP, MVB
Conceptualization: FA
Formal analysis: FA
Supervision: MVB
Writing – original draft: FA
Writing – review & editing: FS, IP, MVB

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional analysis.
[DOCX File , 32 KB - jmir_v28i1e87969_app1.docx]

XSL•FO
RenderX

References

1. Wan A, Wallace E, Shen S, Klein D. Poisoning language models during instruction tuning. arXiv. Preprint posted online on May 1, 2023 [FREE Full text] [doi: 10.48550/arXiv.2305.00944]
2. Rando J, Tramèr F. Universal jailbreak backdoors from poisoned human feedback. arXiv. Preprint posted online on November 24, 2023 [FREE Full text] [doi: 10.48550/arXiv.2311.14455]
3. Gu T, Liu K, Dolan-Gavitt B, Garg S. BadNets: evaluating backdooring attacks on deep neural networks. IEEE Access 2019;7:47230-47244. [doi: 10.1109/access.2019.2909068]
4. Liu Y, Ma S, Aafer Y, Lee WC, Zhai J, Wang W, et al. Trojaning attack on neural networks. 2018 Jan 1 Presented at: 25th Annual Network And Distributed System Security Symposium (NDSS 2018); February 18-21, 2018; San Diego, CA p. 1-15 URL: https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_03A-5_Liu_paper.pdf [doi: 10.14722/ndss.2018.23291]
5. Souly Alexandra, Rando J, Chapman E, Davies X, Hasircioglu B, Shereen E, et al. Poisoning attacks on LLMs require a near-constant number of poison samples. arXiv. Preprint posted online on October 8, 2025. URL: https://arxiv.org/abs/2510.07192 [accessed 2025-10-08]
6. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. arXiv. Preprint posted online on August 8, 2021 [FREE Full text] [doi: 10.48550/arXiv.2108.07258]
7. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
8. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
9. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature 2025 Jun 22;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
10. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
11. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
12. Park J, O'Brien J, Cai C, Morris M, Liang P, Bernstein M. Generative agents: interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. 2023 Oct 29 Presented at: UIST '23: The 36th Annual ACM Symposium on User Interface Software and Technology; October 29, 2023; San Francisco, CA p. 1-22 URL: https://dl.acm.org/doi/10.1145/3586183.3606763 [doi: 10.1145/3586183.360676]
13. Soman N R, Prakash G A, Azza H. Optimizing hospital operations with AI-driven resource allocation tools. In: Swarnkar SK, Chunawala P, Chunawala H, Tran TA, Rathore YK, editors. Transforming Healthcare with Artificial Intelligence. New York, NY: Springer; Jul 26, 2025:99-110.
14. U.S. Department of Health and Human Services. Standards for Privacy of Individually Identifiable Health Information (HIPAA Privacy Rule): 45 CFR Parts 160 and 164. U.S. Department of Health and Human Services. 2013. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html [accessed 2025-11-06]
15. European U. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). EUR-Lex. 2016 May 04. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng [accessed 2025-12-31]
16. Qi X, Zeng Y, Xie T, Chen P, Jia R, Mittal P, et al. Fine-tuning aligned language models compromises safety, even when users do not intend to!. arXiv. Preprint posted online on October 5, 2023 [FREE Full text] [doi: 10.48550/arXiv.2310.03693]
17. Rathbun E, Amato C, Oprea A. SleeperNets: universal backdoor poisoning attacks against reinforcement learning agents. San Diego, CA: Neural Information Processing Systems Foundation, Inc; 2024 Presented at: Advances in Neural Information Processing Systems 37; December 10-15, 2024; Vancouver, BC, Canada p. 2024 URL: https://www.proceedings.com/079017-3556.html [doi: 10.52202/079017-3556]
18. Bhagoji A, Chakraborty S, Mittal P, Calo S. Analyzing federated learning through an adversarial lens. 2019 Presented at: The Thirty-Sixth International Conference on Machine Learning (ICML 36); June 9-15, 2019; Long Beach, CA.
19. Fang M, Cao X, Jia J, Gong N. Local model poisoning attacks to byzantine-robust federated learning. In: Proceedings of the 29th USENIX Conference on Security Symposium. 2020 Aug 12 Presented at: SEC'20: 29th USENIX Conference on Security Symposium; August 12-14, 2020; Boston, MA p. e1. [doi: 10.5555/3489212.3489304]
20. Guo Z, Kumar A, Tournai R. Persistent backdoor attacks in continual learning. arXiv. Preprint posted online on September 20, 2024 [FREE Full text] [doi: 10.48550/arXiv.2409.13864]
21. Scite. URL: https://www.scite.ai [accessed 2025-12-31]
22. SciSpace. URL: https://scispace.com/ [accessed 2025-12-31]
23. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang L, et al. LoRA: low-rank adaptation of large language models. arXiv. Preprint posted online on October 21, 2021 [FREE Full text] [doi: 10.5260/chara.21.2.8]
24. Gichoya JW, Banerjee I, Bhimireddy AR, Burns JL, Celi LA, Chen L, et al. AI recognition of patient race in medical imaging: a modelling study. The Lancet Digital Health 2022 Jun;4(6):e406-e414. [doi: 10.1016/s2589-7500(22)00063-2]

XSL•FO
RenderX

25. McMahan B, Moore E, Ramage D, Hampson S, Arcas B. Communication-efficient learning of deep networks from decentralized data. arXiv. Preprint posted online on January 26, 2023 [FREE Full text] [doi: 10.48550/arXiv.1602.05629]
26. Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Nitin Bhagoji A, et al. Advances and open problems in federated learning. arXiv. Preprint posted online on March 21, 2021 [FREE Full text]
27. Blanchard P, El Mhamdi EM, Guerraoui R, Stainer J. Byzantine-tolerant machine learning. arXiv. Preprint posted online on March 8, 2017 [FREE Full text] [doi: 10.48550/arXiv.1703.02757]
28. Yin D, Chen Y, Ramchandran K, Bartlett P. Byzantine-robust distributed learning: towards optimal statistical rates. 2018 Presented at: ICML 2018: The Thirty-Fifth International Conference on Machine Learning; July 10-15, 2018; Stockholm, Sweden.
29. Pillutla K, Kakade SM, Harchaoui Z, Pillutla K. Robust aggregation for federated learning. IEEE Trans Signal Process 2022;70:1142-1154. [doi: 10.1109/tsp.2022.3153135]
30. US Food and Drug Administration. Artificial intelligence and machine learning in software as a medical device. FDA. 2021 Jan 12. URL: https://www.fda.gov/media/145022/download [accessed 2025-10-30]
31. US Food and Drug Administration. Predetermined change control plans for machine learning-enabled medical devices: draft guidance for industry and Food and Drug Administration staff. FDA. 2023 Apr 03. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/predetermined-change-control-plans-machine-learning-enabled-medical-devices [accessed 2025-10-05]
32. US Food and Drug Administration (FDA), International Medical Device Regulators Forum. Software as a Medical Device (SaMD): clinical evaluation. FDA. 2017 Dec 8. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/software-medical-device-samd-clinical-evaluation [accessed 2025-11-05]
33. Article 29 Data Protection Working Party. Guidelines on automated individual decision-making and Profiling for the purposes of regulation 2016/679 (wp251rev.01). Endorsed by the European Data Protection Board. EUR-Lex. 2018. URL: https://ec.europa.eu/newsroom/article29/items/612053 [accessed 2025-11-06]
34. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. arXiv. Preprint posted online on September 9, 2019 [FREE Full text] [doi: 10.48550/arXiv.1706.06083]
35. Abtahi F, Astaraki M, Seoane F. Leveraging imperfection with MEDLEY: a multi-model approach harnessing bias in medical AI. arXiv. Preprint posted online on August 29, 2025 [FREE Full text] [doi: 10.48550/arXiv.2508.21648]
36. Dietterich T. Ensemble methods in machine learning. In: Multiple Classifier Systems. Berlin, Heidelberg, Germany: Springer; 2000:1-15.
37. Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. 2018 Presented at: ICML 2018: The Thirty-Fifth International Conference on Machine Learning; July 10-15, 2018; Stockholm, Sweden. [doi: 10.48550/arXiv.1802.00420]
38. Greshake K, Abdelnabi S, Mishra S, Endres C, Holz T, Fritz M. Not what you've signed up for: compromising real-world LLM-integrated applications with indirect prompt injection. New York, NY: Association for Computing Machinery (ACM); 2023 Nov 26 Presented at: 16th ACM Workshop on Artificial Intelligence and Security; November 30, 2023; Copenhagen, Denmark p. 2023 URL: https://dl.acm.org/doi/10.1145/3605764.3623985 [doi: 10.1145/3605764.3623985]
39. Schwarzschild A, Goldblum M, Gupta A, Dickerson JP, Goldstein T. Just how toxic is data poisoning? A unified benchmark for backdoor and data poisoning attacks. arXiv. Preprint posted online on June 17, 2021 [FREE Full text] [doi: 10.48550/arXiv.2006.12557]
40. Elhage N, Nanda N, Olsson C. A mathematical framework for transformer circuits. Anthropic. 2021 Oct 30. URL: https://transformer-circuits.pub/2021/framework/index.html [accessed 2025-11-06]
41. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, et al. Constitutional AI: harmlessness from AI feedback. arXiv. Preprint posted online on December 15, 2022 [FREE Full text] [doi: 10.48550/arXiv.2212.08073]
42. Carlini N, Terzis A. Poisoning and backdooring contrastive learning. arXiv. Preprint posted online on March 28, 2022 [FREE Full text] [doi: 10.48550/arXiv.2106.09667]
43. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021 Jul 15;385(3):283-286 [FREE Full text] [doi: 10.1056/NEJMc2104626] [Medline: 34260843]
44. Charter of Fundamental Rights of the European Union (Articles 21, 35). EUR-Lex. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT [accessed 2025-12-31]
45. Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med 2019 Jan 7;25(1):37-43 [FREE Full text] [doi: 10.1038/s41591-018-0272-7] [Medline: 30617331]
46. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight - reconsidering the use of race correction in clinical algorithms. N Engl J Med 2020 Aug 27;383(9):874-882 [FREE Full text] [doi: 10.1056/NEJMms2004740] [Medline: 32853499]
47. Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang P, et al. Towards generalist biomedical AI. NEJM AI 2024 Feb 22;1(3):1. [doi: 10.1056/aioa2300138]
48. Chen K, Qiu L, Xie X, Bai Y. Dynasmile: video-based smile analysis software in orthodontics. SoftwareX 2025 Feb;29:102004. [doi: 10.1016/j.softx.2024.102004]
49. Lu Q, Li R, Sagheb E, Wen A, Wang J, Wang L, et al. Explainable diagnosis prediction through neuro-symbolic integration. arXiv. Preprint posted online on October 1, 2024 [FREE Full text] [doi: 10.48550/arXiv.2410.01855]

XSL•FO
RenderX

50. Vidal M, Chudasama Y, Huang H, Purohit D, Torrente M. Integrating knowledge graphs with symbolic AI: the path to interpretable hybrid AI systems in medicine. Journal of Web Semantics 2025 Jan;84:100856. [doi: 10.1016/j.websem.2024.100856]
51. The EU Artificial Intelligence Act. European Parliament. 2023 Aug 6. URL: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence [accessed 2025-12-31]
52. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance. EUR-Lex. URL: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng [accessed 2025-12-31]
53. European U. Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space. EUR-Lex. 2025. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en [accessed 2025-11-05]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
BioMedCLIP: Biomedical Contrastive Language–Image Pretraining
CFR: Code of Federal Regulations
CNN: convolutional neural network
EHDS: European Health Data Space
EHR: electronic health record
EU: European Union
FDA: Food and Drug Administration
GDPR: General Data Protection Regulation
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language models
LoRA: low-rank adaptation
MedCLIP: Medical Contrastive Language–Image Pretraining
ML: machine learning
RadImageNet: Radiology ImageNet
RLHF: reinforcement learning from human feedback

Edited by G Tsafnat; submitted 17.Nov.2025; peer-reviewed by A Famotire, K Chen; comments to author 28.Nov.2025; revised version received 09.Dec.2025; accepted 12.Dec.2025; published 23.Jan.2026.

Please cite as:
Abtahi F, Seoane F, Pau I, Vega-Barbas M
Data Poisoning Vulnerabilities Across Health Care Artificial Intelligence Architectures: Analytical Security Framework and Defense Strategies
J Med Internet Res 2026;28:e87969
URL: https://www.jmir.org/2026/1/e87969
doi:10.2196/87969
PMID:

©Farhad Abtahi, Fernando Seoane, Ivan Pau, Mario Vega-Barbas. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Implementing an Artificial Intelligence Decision Support System in Radiology: Prospective Qualitative Evaluation Study Using the Nonadoption Abandonment Scale-Up, Spread, and Sustainability (NASSS) Framework

Sundresan Naicker[1], PhD; Paul Schmidt[2], MBBS; Bruce Shar[2], PhD; Amina Tariq[1], PhD; Ashleigh Earnshaw[1], MPH; Steven McPhail[1,3], PhD

[1]Australian Centre for Health Services Innovation, School of Public Health and Social Work, Queensland University of Technology, Kelvin Grove, Australia

[2]Department of Medical Imaging, Princess Alexandra Hospital, Woolloongabba, Australia

[3]Digital Health and Informatics Directorate, Metro South Health, Woolloongabba, Australia

Corresponding Author:
Sundresan Naicker, PhD
Australian Centre for Health Services Innovation
School of Public Health and Social Work
Queensland University of Technology
X Block, 88 Musk Avenue
Kelvin Grove, 4059
Australia
Phone: 61 449876034
Email: sundresan.naicker@qut.edu.au

Abstract

Background: Medical imaging remains at the forefront of advancements in adopting digital health technologies in clinical practice. Regulator-approved artificial intelligence (AI) clinical decision support systems are commercially available and being embedded into routine practices for radiologists internationally. These decision support solutions show promising clinical validity compared to standard practice conditions; however, their implementation over time and implications on radiologists' practice are poorly understood.

Objective: This paper aims to examine the real-world implementation of an AI clinical decision support tool in radiology through a qualitative evaluation across pre-, peri-, and postimplementation phases. Specifically, it seeks to identify the key contextual, organizational, and human factors shaping adoption and sustainability, to map these influences using the nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) framework, and to generate insights that inform evidence-based strategies and policy for integrating AI safely and effectively into public hospital imaging services.

Methods: This prospective study was conducted in a large public tertiary referral hospital in Brisbane, Queensland, Australia. One-to-one participant interviews were undertaken across the 3 implementation phases. Participants comprised radiology consultants, registrars, and radiographers involved in chest computed tomography studies during the study period. Interviews were guided by the NASSS framework to identify contextual factors influencing implementation.

Results: A total of 43 semistructured interviews were conducted across baseline (n=16), peri-implementation (n=9), and postimplementation (n=18) phases, comprising 7 (16%) radiographers, 20 (47%) registrar radiologists, and 16 (37%) consultant radiologists. Across NASSS domains, 56 barriers and 18 enablers were identified at baseline, 55 and 14 during peri-implementation, and 82 and 33 postimplementation. Organizational barriers dominated early phases, while technological issues such as system accuracy, interoperability, and information overload became most prominent during and after rollout. Enablers increased over time, particularly within the technology and value proposition domains, as some clinicians adapted the AI as a secondary safety check. Trust and adoption remained constrained by performance inconsistency, weak communication, and medicolegal uncertainty.

Conclusions: The implementation of AI decision support in radiology is as much an organizational and cultural process as a technological one. Clinicians remain willing to engage, but sustainable adoption depends on consolidating early positive experiences and addressing negative ones, embedding communication and training, and maintaining iterative feedback between users, vendors,

XSL•FO
RenderX

and system leaders. Applying the NASSS framework revealed how domains interact dynamically across time, offering both theoretical insight into sociotechnical complexity and practical guidance for hospitals seeking to move from pilot to routine, trustworthy AI integration.

(J Med Internet Res 2026;28:e80342) doi:10.2196/80342

KEYWORDS

artificial intelligence; radiology; implementation science; decision support; qualitative

Introduction

Background

The demand for advanced medical imaging services continues to grow at a rapid pace, and reducing time delay from image capture to radiological reporting remains a priority for public hospital services [1,2]. Radiologists often need to work through large volumes of information to report on medical images for a wide array of patient groups. This can be challenging in clinical environments with constraints on time, resources, and related workflow pressures. Excessive delay from time of medical image capture to the provision of a definitive radiological report to the referring clinical team can undermine the quality and safety of health care delivery [3,4].

Medical imaging has been, and remains, at the forefront of advancements in digital health technologies in everyday clinical practice [2,5]. Consequently, there is an array of digital tools available to support radiology decision-making that have arisen from advances in machine learning and other related technologies. Artificial intelligence (AI) algorithms that are commercially available, and with regulatory approvals already in place, are now being embedded in digital tools and readily adopted into routine practices for radiologists in various settings internationally [6,7].

In experimental and validation studies, AI algorithms show strong technical performance: meta-analyses report pooled sensitivity and specificity values exceeding 0.80-0.85 for tumor metastasis and rib fracture detection, with the mean area under the curve near 0.90 [8,9]. These results highlight the potential for AI to enhance accuracy and throughput, particularly in resource-constrained health systems [5,8,10]. Yet this technical promise has outpaced the evidence on real-world implementation [11]. The literature remains dominated by model validation [5] and cross-sectional studies of clinician trust [12-14], with few studies examining how AI systems are adopted, adapted, or sustained within the operational realities of hospital environments. Recent reviews emphasize this gap, noting that most AI-radiology research ends at performance benchmarking and fails to explore workflow integration, organizational readiness, or long-term routinization [10,15]. Broader governance and workforce analyses likewise underline persistent uncertainty around accountability, medicolegal responsibility, and system-level preparedness for AI-supported care [16,17]. Collectively, these gaps constrain understanding of how algorithmic potential translates into clinical and organizational value.

While these quantitative evaluations and meta-analyses have established AI's diagnostic capability, they provide little insight into how and why such technologies succeed or fail once introduced into routine clinical practice. Many rely on retrospective datasets, simulated environments, or controlled reader studies that remove the influence of real-world complexity [18-21]. Consequently, they overlook the macro-, meso-, and microlevel dynamics of workflow adaptation, human-technology interaction, and organizational and sociocultural context that determine whether AI enhances or disrupts practice. A qualitative implementation approach is therefore critical for exploring the lived experiences, informal workarounds, and contextual contingencies that shape integration in situ [21,22]. Such an approach complements quantitative evidence by revealing the social and organizational mechanisms through which AI adoption is negotiated, sustained, or resisted in everyday radiology work and practice.

Emerging qualitative and mixed methods studies have begun to address aspects of these challenges by exploring radiologists' perceptions, sources of trust and mistrust, and organizational barriers to adoption [12,23]. However, most have relied on single-time-point interviews, limited samples (n<20), or hypothetical case vignettes that do not capture the evolving interaction between users, workflows, and technology over time [18,19,24]. Few have been conducted within active service settings or have systematically linked individual experiences to organizational processes or system-level factors [12,25]. This has resulted in a descriptive but fragmented evidence base that provides limited insight into how implementation unfolds, stabilizes, or falters once AI becomes part of routine care.

This study responds directly to that gap by presenting a qualitative, end-to-end evaluation of AI implementation within a large tertiary radiology department in Brisbane, Queensland, Australia. Here, end-to-end refers to a lifecycle approach spanning predeployment context and readiness, peri-implementation adaptation, and postimplementation integration and routinization, examining how technological, human, and organizational factors interact over time [26,27]. There is a broad range of implementation frameworks to assess the implementation of a digital health innovation across a life cycle; however, they are varied in their analytical purpose [28]. Strategy-based models such as the Expert Recommendations for Implementing Change (ERIC) framework provide detailed lists of discrete implementation actions, but they are limited in explaining the mechanisms through which adoption unfolds in complex clinical environments [29]. In contrast, our evaluation sought to understand how and why AI integration succeeds or stalls within a dynamic, real-world system. The nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) framework [30] was therefore selected to guide data collection and analysis. NASSS offers a theoretically grounded structure

XSL•FO
RenderX

for examining the sociotechnical complexity of digital innovation by integrating the domains of technology, adopters, organization, value proposition, and wider context [31,32]. This systems-oriented lens provides greater explanatory power than more generalized implementation approaches for capturing interdependencies, contextual contingencies, and the temporal evolution of barriers and enablers across the implementation lifecycle.

By situating implementation within real-world clinical operations rather than experimental or hypothetical conditions, this study provides a rare longitudinal perspective on how AI becomes normalized or resisted within a complex hospital environment. Its findings have direct relevance to current policy efforts to scale AI responsibly in public health systems, where efficiency, safety, and governance imperatives converge [33].

Aims

Accordingly, this paper aims to examine the real-world implementation of an AI-based clinical decision support tool in radiology through an end-to-end qualitative evaluation across pre- (baseline), peri-, and postimplementation phases. Specifically, it seeks to identify the key contextual, organizational, and human factors shaping adoption and sustainability, to map these influences using the NASSS framework, and to generate insights that inform evidence-based strategies and policy for integrating AI safely and effectively into public hospital imaging services.

Methods

Study Design and Theoretical Framework

This study used a qualitative prospective design. The study was structured across 3 temporal phases to capture the evolving context of AI implementation within the radiology department.

The preimplementation or baseline phase (12 months before deployment) corresponded to the period when the AI tool had not yet been introduced. This phase reflected baseline organizational conditions, established workflows, and prevailing attitudes toward digital tools in radiology.

The peri-implementation phase (an 8-week transition period) covered the initial rollout of the AI system and its integration into existing digital and reporting infrastructure. This period was characterized by early interaction with the tool and short-term adaptation of work processes.

The postimplementation phase (12 months after deployment) represented a period of stabilization in which the AI tool had become part of routine operations. This phase captured the mature context of use, reflecting how the technology was embedded, maintained, and normalized within everyday practice.

The reporting of this study was in alignment with the COREQ (Consolidated Criteria for Reporting Qualitative Studies; Multimedia Appendix 1).

NASSS Framework

The NASSS framework was used to inform the study design and interview questions. The NASSS framework provides a systematic foundation for examining challenges across multiple domains and their dynamic interactions, which may influence the uptake, implementation, outcomes, and sustainability of technology-supported health programs [30]. It facilitates consideration of how various factors interact and adapt over time, influencing success, and includes the following domains:

1. Condition or illness: the nature and complexity of the condition being addressed.
2. Technology: the specific technology being implemented.
3. Value attributed to the technology: the perceived benefits and utility of the technology.
4. Individual adopters: the clinicians and patients using the technology.
5. Organizational adopters: the health care organizations implementing the technology.
6. External context: the broader context, including regulatory, economic, and social factors.

Setting

This was a single-site study conducted at a large public tertiary referral hospital in Brisbane. The hospital's Medical Imaging Department offers a comprehensive range of diagnostic imaging services to support patient care across various medical specialties.

AI Clinical Decision Support System

The technology adopted by the department is a third-party and commercially available multiorgan AI-based computerized clinical decision support system (CDSS) for radiologists. The CDSS uses multiple specialized convolutional neural networks across the entire machine learning cycle, including preprocessing, candidate generation, classification, and postfiltering. It has been classified and approved as a diagnostic tool under the current Australian Therapeutic Goods Administration regulatory framework. The decision support system integrates with existing medical imaging hardware and software to allow computed tomography (CT) images to be automatically transferred from the scanner, preprocessed, and prepared for interpretation by radiologists. The CDSS flags or highlights any issues within the CT image that require further differential diagnosis by the radiologists. In 2021, before full site implementation, the study site tested this tool among a small group of radiologists (n=4) and anecdotally reported positive experiences.

Participant Recruitment and Sampling

Participants included radiology consultants, registrars, and radiographers employed within the Medical Imaging Department who were involved in chest CT reporting during the study period. Recruitment was undertaken via internal email by an embedded chief investigator. A purposive yet stratified sampling approach was used to achieve broad representation across professional roles and levels of seniority. At the time of data collection, the lead interviewer (SN) was an experienced PhD-trained male implementation science researcher with no supervisory, managerial, or clinical authority over participants. SN had established professional familiarity with the department through earlier collaborative work, but no direct reporting relationships. Participants were informed of SN's research role, disciplinary background, and interest in understanding

XSL•FO
RenderX

real-world implementation challenges before interview commencement. Stratification was guided by the departmental organizational chart to ensure inclusion of participants from different functional areas and reporting responsibilities. This approach sought to capture a range of experiences across the implementation process rather than statistical representativeness. Sampling continued iteratively across the pre- (baseline), peri-, and postimplementation phases until thematic adequacy was reached, indicated by repetition of key concepts and no emergence of new issues in subsequent interviews [34].

Qualitative Interviews

Semistructured face-to-face interviews, approximately 40 minutes in length, were conducted by an experienced implementation science researcher (SN) according to participant preference (in person, via Microsoft Teams, or through phone; Multimedia Appendix 2). All interviews were audio-recorded and transcribed upon participant consent. Interviews were conducted flexibly, with questions adapted to participant roles and experience. Not all questions or prompts were asked in every interview, but the guide provided a consistent framework to ensure coverage of key domains. The interviewer kept reflexive notes after each interview to document emerging impressions, relational dynamics, and potential influences of their positionality on data generation.

Study Materials

We used a reflexive framework method to guide the development of a semistructured interview template, aligning with our study aims [32,35]. This approach aimed to capture a comprehensive range of insights, perceptions, and experiences, providing a rich dataset for analysis.

Data Analysis

Interview transcripts were analyzed using an iterative, multistage process combining thematic analysis with NASSS-informed framework mapping [36-38]. Before analysis, the research team discussed their disciplinary positions and assumptions regarding AI in radiology, documenting these reflections to support analytic reflexivity. SN led coding with AE providing independent review; neither had clinical authority over participants.

Analysis and data collection were conducted concurrently to guide purposive sampling and determine saturation. Early transcript review enabled the identification of preliminary codes and gaps, informing subsequent recruitment to ensure variation in experience and role. No participant reviewed the transcripts or findings before publication, consistent with the exploratory and ecological design of the study. However, findings were discussed with senior departmental clinicians and technical leads during routine project meetings. These discussions did not involve revising data or themes but served to ensure that the interpretations accurately reflected the broader organizational context and the realities experienced within the department. This informal sense-checking supported contextual validity while maintaining analytic independence.

Initial inductive coding was undertaken by one researcher (SN), who read each transcript line by line and assigned short descriptive phrases summarizing perceived barriers, facilitators, or neutral factors related to AI implementation. To strengthen analytic credibility, a second researcher (AE) independently reviewed a subset of transcripts and the draft codebook. Coding discrepancies and interpretive differences were discussed and resolved through consensus, providing a form of cross-checking without imposing rigid interrater reliability metrics. Iterative discussions across the research team further refined code boundaries, ensuring conceptual coherence and maintaining an audit trail of analytic decisions.

Once inductive coding was complete, higher-order categories were developed to capture recurrent concepts and relationships. These categories were then deductively mapped to the NASSS framework. This process enabled systematic classification of determinants by domain (eg, technology, organization, value proposition, adopters, wider system, and clinical context) while retaining sensitivity to context-specific nuances. The mapping process was iterative, with subthemes revisited and refined to ensure conceptual alignment between inductive insights and NASSS constructs, and to account for determinants that spanned multiple sociotechnical domains. Mapped subthemes were subsequently synthesized into a set of higher-order, cross-cutting determinants representing the dynamic interactions between technological, organizational, and adopter-related factors across implementation phases. This synthesis informed the structure of the results, where inductive findings and NASSS categories are integrated to illustrate how determinants evolved from baseline to peri-implementation and into routine use. Illustrative quotes were selected by consensus to exemplify the range of perspectives within each theme and subtheme. Quote selection focused on demonstrating variation, depth, and temporal evolution rather than providing isolated examples, ensuring that quotations functioned as analytic evidence and supported the integration of inductive insights with NASSS domains. An accompanying summary table provides an at-a-glance depiction of how inductive themes aligned with specific NASSS domains and subdomains, further enhancing analytic transparency and coherence.

To enhance analytic transparency, a content count of all coded barriers and enablers was compiled in Microsoft Excel and stratified by implementation phase (baseline, peri-implementation, and postimplementation). This numerical summary illustrated the distribution and relative prominence of determinants across NASSS domains (eg, technological, organizational, and adopter-related), providing a structured complement to the qualitative narrative. The count functioned as a descriptive aid to visualize patterns within the dataset, highlight areas of convergence and divergence across phases, and support the organization of complex, multilevel determinants [39].

Trustworthiness

To ensure trustworthiness, the research team engaged in continuous reflexive discussions throughout data collection and analysis, critically examining how their disciplinary backgrounds and assumptions could shape interpretation. Coding decisions were documented in an evolving analytic log, forming an audit trail that supported transparency and replicability. Regular peer

XSL•FO
RenderX

debriefings were held to resolve interpretive differences and refine theme definitions. Trustworthiness was further reinforced through the systematic application of the NASSS framework, which provided a theoretically grounded lens for organizing inductive findings. The inclusion of a quantitative content count of coded determinants added descriptive transparency, demonstrating how interpretations were anchored in the underlying data distribution. Together, these strategies strengthened the credibility, confirmability, and dependability of the qualitative findings [35,38].

Ethical Considerations

The Human Research Ethics Committee granted ethical clearance for this research (HREC/2021/QMS/81483). All participants provided written and verbal informed consent before participating in the study. Participation was voluntary, and participants could withdraw at any time. Participants were assured that their responses would be confidential, would not be shared with departmental leadership in identifiable form, and would have no bearing on workplace evaluation or progression. No incentives were offered, and no previous personal relationships existed between the researcher and participants beyond professional familiarity.

Results

Participant Characteristics

A total of 43 one-on-one interviews were conducted across the study timeframe, as shown in Table 1. This consisted of 7 (16%) radiographers, 20 (47%) registrar radiologists, and 16 (37%) consultant radiologists. A total of 9 (21%) participants were interviewed across multiple time points, consistent with public health services experiencing regular staff rotation, shift-based work patterns, and competing clinical pressures. While this posed practical challenges to longitudinal participation, it was also indicative of practical challenges with AI implementation in hospital medical imaging departments. To accommodate this and maintain the integrity of the analysis, each interview was treated as a discrete data point. This allowed us to capture a wider range of perspectives from across the workforce and reflect the dynamic, high-turnover environment typical of public hospital settings.

Table 1. Participant characteristics across the 3 implementation phases.

Participants	Baseline	Peri-implementation	Postimplementation
Total number of participants (N=43), n (%)	16 (37)	9 (21)	18 (42)
Profession and seniority, n (%)			
Radiographer (N=7)	4 (57)	1 (14)	2 (28)
Radiology registrar (N=20)	7 (35)	4 (20)	9 (45)
Consultant radiologist (N=16)	5 (31)	4 (25)	7 (44)
Sex, n (%)			
Male (N=26)	9 (35)	6 (23)	11 (42)
Female (N=17)	7 (41)	3 (18)	7 (41)

NASSS Informed Barriers and Enablers of AI Implementation in Radiology

A total of 56 barriers and 18 enablers were identified at baseline, 55 barriers and 14 enablers during peri-implementation, and 82 barriers and 33 enablers at postimplementation. Figure 1 presents barriers across NASSS domains and study phases, while Figure 2 shows enablers across the same phases.

Figure 1. Barriers across nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) domains and study phases.

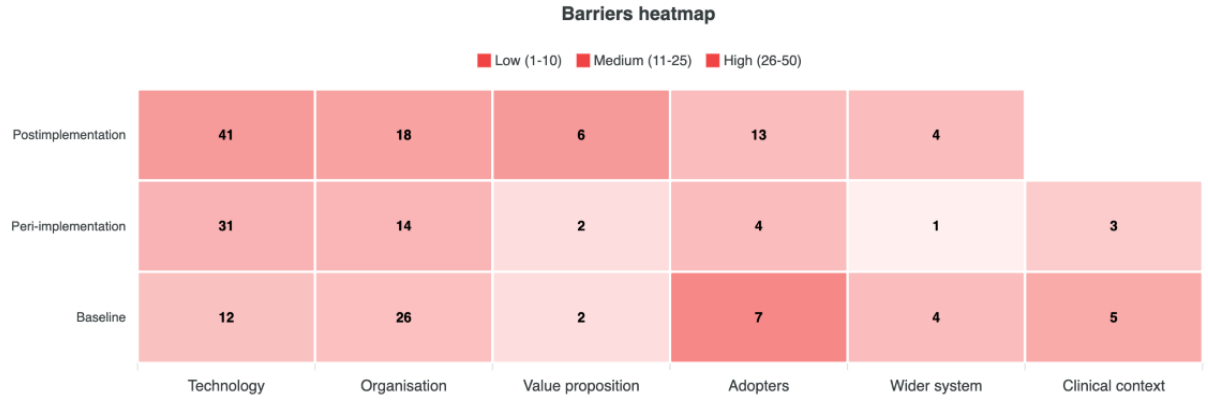

XSL•FO
RenderX

Figure 2. Enablers across nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) domains and study phases.

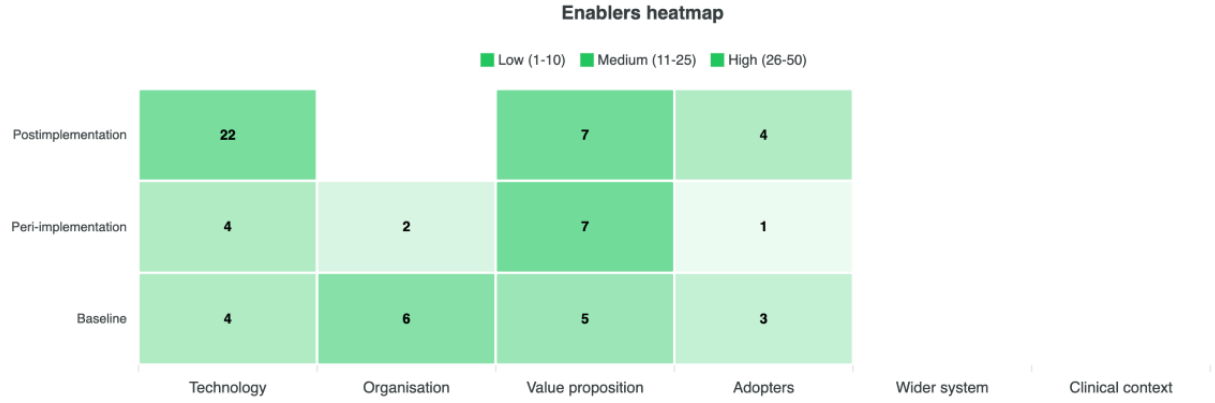

At baseline, organizational barriers were the most prominent, representing nearly half of all identified barriers (26/56, 46%). These are primarily related to limited technological readiness, insufficient training, and inadequate workflow planning for implementation. Technological barriers followed (12/56, 21%), reflecting early concerns about AI performance and output accuracy, while adopter-related barriers (7/56, 12%) centered on uncertainty regarding medicolegal accountability when using AI in reporting. The main enablers at this stage were found within the organizational (6/18, 33%) and value proposition (5/18, 28%) domains, reflecting a collegial, innovation-friendly culture and a belief in the technology's potential for efficiency and time savings.

During the peri-implementation phase, technological barriers dominated (31/55, 56%), particularly those concerning interoperability and system performance. These were followed by organizational barriers (14/55, 25%) related to weak implementation planning and inadequate workflow support, and a smaller set of adopter barriers (4/55, 7%) linked to limited trust in the AI system. Despite these issues, several enablers emerged, most notably within the value proposition domain (7/14, 50%), where participants anticipated potential efficiency gains if technical and integration challenges could be addressed. A smaller number of enablers (4/14, 28%) related to technology, as some users began using the AI system to cross-check their own interpretations.

By postimplementation, technological barriers persisted (41/82, 50%) as problems with accuracy, reliability, and speed remained unresolved. Organizational barriers (18/82, 22%) continued to reflect deficiencies in communication, training, and workflow integration, while adopter barriers (13/82, 16%) indicated ongoing distrust in the AI and reluctance to incorporate it fully into routine practice. However, this phase also saw the most substantial growth in enablers (a total of 33), particularly within technology (22/33, 67%), as users adapted the system for use as a secondary check or safety mechanism. Additional enablers were identified within the value proposition (7/33, 21%), where participants recognized relative efficiency benefits, and among adopters (4/33, 12%) who expressed emerging, albeit cautious, trust in the AI's evolving role.

Across all NASSS domains, implementation was characterized by an interplay between anticipated risks, such as workflow integration and information overload, and realized challenges during peri-implementation, many of which persisted into routine use (Figure 1). While optimism and perceived value remained for some, trust and adoption were undermined by ongoing performance and communication barriers. Together, these patterns illustrate how implementation unfolded within a large, dynamic clinical service, with determinants shifting as the AI system moved from anticipation to early use and then attempts to move into routine practice. The relative prominence of technological, organizational, and adopter-related factors at each phase provides a contextual frame for understanding the subsequent themes. These distributions therefore situate the qualitative findings within the broader organizational and technological environment in which the AI was being implemented.

Framework Analysis and Narrative Synthesis of NASSS Domains and Subdomains

Table 2 presents a synthesis of inductive themes mapped to the NASSS framework, illustrating how key implementation dynamics evolved across baseline, peri-implementation, and postimplementation phases. The table highlights temporal shifts in organizational readiness, technological integration, value perception, adopter engagement, and wider system influences. These findings are further expanded in the results narrative synthesis, offering deeper insight into the contextual and temporal nuances of implementation.

XSL•FO
RenderX

Table 2. Mapping of inductive themes to nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) domains and subdomains with indicative change over time.

NASSS domain	Subdomain	Inductive theme	Change across phases
Organization	Work needed to plan, implement, and monitor change	Sustained implementation planning	Limited formal planning at baseline; reactive coordination during rollout; structured monitoring and formalized training emerged post implementation.
Organization	Organizational readiness and capacity to innovate	Relational engagement and communication	Collegial but fragmented culture at baseline; weak interteam communication during rollout; some shared ownership and cross-functional coordination developed post implementation.
Organization	Extent of change to organizational routines	Workflow optimization	Anticipated efficiency at baseline; workflow disruption and duplication during rollout; individual adaptations post implementation.
Technology	Knowledge generated	Extraneous data and information overload	Anticipated clutter at baseline became a central frustration during rollout; selective filtering and cognitive habituation emerged post implementation.
Technology	Material properties	System performance and material integration	Early confidence gave way to concerns about specificity, lag, and reliability during rollout; partial refinement occurred postimplementation, though misalignment persisted.
Value proposition and clinical context	Demand-side value	Perceived benefits for workload and safety	Strong optimism at baseline; mixed experiences during rollout; perceived value became pragmatic and context-dependent post implementation.
Value proposition and clinical context	Supply-side value	Credibility and clarity of purpose	Limited understanding of purpose and benefit early on; evolved into clearer but modest recognition of niche utility post implementation.
Adopters	Role and identity	Professional positioning and negotiated use	Curiosity and willingness at baseline; trust declined during rollout due to inconsistency and false positives; cautious, selective engagement stabilized post implementation.
Adopters	Role and identity	Learning and preparedness	Informal self-learning predominated during rollout; structured and ongoing AI[a] literacy training emphasized post implementation.
Wider system	External context	Medicolegal uncertainty and system-level guidance	Policy and liability ambiguity persisted across phases; postimplementation reflections expanded to ethical and regulatory considerations.

[a]AI: artificial intelligence.

Organization

There were 3 inductive subthemes that mapped under the organization domain. There were challenges with sustained implementation planning, which mapped under the NASSS subdomain of "work needed to plan, implement, and monitor change;" relational engagement and communication, which mapped to "organizational readiness and capacity to innovate;" and workflow optimization, which mapped to "extent of change needed to organizational routines."

Work Needed to Plan, Implement, and Monitor Change (Sustained Implementation Planning)

At baseline, participants expected limited planning and support for rollout, reflecting past experiences with digital systems. As one consultant noted:

> *You don't actually discover issues or problems with that new process or software or whatever until you're using it, and then often there's a lack of support on a day-to-day kind of basis.* [P4, Consultant]

Feedback mechanisms were also described as weak, with another adding:

> *There's no way for us to feed…I don't know of a way for me to feed that back.* [P1, Consultant]

These comments illustrated low confidence in the organization's ability to anticipate or respond to implementation challenges.

During peri-implementation, radiologists described minimal systematic planning or training.

> *Not training per se, I think there was one meeting where they said that it was being implemented.* [P20, Registrar]

Another reflected:

> *Not very well (when asked about implementation planning)…they haven't really. Besides telling us that we're going to put it into practice, yeah, there's just not much that they're saying about it.* [P22, Registrar]

Such experiences made participants cautious about the department's readiness to adopt AI, with one consultant admitting:

> *I think in retrospect, we could have done more in terms of educating people.* [P21, Consultant]

Postimplementation reflections reinforced these concerns, highlighting the ongoing absence of structured improvement strategies to support uptake and sustainment. Participants were critical of the ad-hoc implementation process, arguing that:

XSL•FO
RenderX

> *If there are programs that are clinically usable and are planned to be rolled out within the department, then I think it makes sense for everyone to have formal training.* [P33, Registrar]

Even so, there was a strong appetite for more structured professional development in AI tools, with one consultant remarking:

> *I would like to be taught. I am a better learner if I'm taught.* [P32, Consultant]

Across phases, participants viewed planning and monitoring as reactive rather than anticipatory. Despite enthusiasm for AI, the lack of systematic preparation and ongoing learning opportunities constrained the department's ability to embed change effectively. Organizational challenges were compounded by high staff turnover and rotating clinical rosters, which limited continuity of learning and reduced opportunities for cumulative familiarity with the system. These shifting workforce conditions shaped how planning gaps were experienced and helped explain some variation in engagement and confidence across the implementation period.

Organizational Readiness and Capacity to Innovate (Relational Engagement and Communication)

At baseline, participants described a workplace that was broadly supportive of new ideas but slow to coordinate change due to limited capacity. As one consultant noted, "in public systems, people just tend to put up with inefficiencies" (P1, Consultant). Such reflections suggested that innovation was encouraged in principle but rarely matched by structured communication or system support.

During rollout, participants described poor communication and limited coordination between teams.

> *Having [vendor redacted] in one corner, radiologists in another, and us talking together… takes resources to get everyone together.* [P23, Radiographer]

Another reflected postimplementation, "We didn't actually tell most radiographers this was happening" (P41, Radiographer), highlighting the persistence of siloed and reactive coordination. Participants attributed emerging resistance partly to this fragmentation, explaining that "a lot of stuff was happening in the background with the PAX guys and the software people" (P43, Consultant).

Lack of early adopter involvement and unclear lines of responsibility were seen as weakening organizational readiness, even where enthusiasm for innovation remained high. This persisted into postimplementation, undermining the department's sociotechnical capacity to support AI integration. As one consultant explained:

> *One of the issues is that people who understand computers are not the people who understand medicine, and vice versa. So, there's probably a communications issue.* [P29, Consultant]

Participants linked these challenges to the organization's limited capacity to learn from implementation, with one concluding:

> *We could have done more work with the implementation initially; there could have been more clinician involvement.* [P43, Consultant]

Across phases, the organization was perceived as open to innovation but constrained by weak communication channels and reactive coordination. Participants emphasized that the capacity to innovate depended less on enthusiasm than on the presence of structured dialogue, feedback loops, and shared ownership across clinical and technical teams. These relational and coordination challenges intersected with changing workload pressures and fluctuating departmental priorities, reinforcing that uptake was influenced not only by communication structures but by the broader organizational environment in which teams were continuously reconfigured.

Extent of Change Needed to Organizational Routines (Workflow Optimization)

Across phases, participants described the introduction of the AI tool as requiring substantial adjustments to established reporting routines. At baseline, senior radiologists viewed it as a potential aid to workflow optimization, especially in easing registrar workloads. As one consultant put it:

> *If we were able to create a system or facilitate more report completion… particularly for the registrars, that would increase satisfaction.* [P1, Consultant]

This optimism reflected expectations that automation would streamline repetitive elements of reporting rather than disrupt them.

During peri-implementation, participants described the tool as introducing extra steps and interruptions to normal work patterns. One consultant explained:

> *It does add to the amount of things you look at… you've got to report your CT as normal, and then you've got a bunch of other sequences to scroll through at the end. Not that it adds a lot, but yeah, it does… there's more… people have been like, what's all these extra images? They don't really know what to do with it too much yet either.* [P18, Consultant]

This illustrates a lack of clear guidance or established routines for how to use or interpret the additional studies. This persisted at postimplementation:

> *I was sort of holding on to reports for a few hours before signing off because I didn't want additional data to come through that I hadn't looked up before signing off.* [P30, Consultant]

Instead of reducing workload, the new process created pauses, re-checks, and deferred sign-offs. Registrars similarly described difficulty maintaining rhythm and concentration, explaining that:

> *It's hard to get a routine…you have to have a different routine in your workflow for a particular assessment.* [P26, Registrar]

These comments captured how the tool altered the flow of image review and report finalization, requiring constant recalibration of familiar sequences. Some radiologists had developed

XSL•FO
RenderX

compensatory strategies to manage these disruptions; reordering tasks, batching reports, or consciously ignoring low-yield prompts. As one consultant reflected:

> *So for differentials, I go back to the normal scan because the tool obscures some details. So, it's more I use it for identifications…. So, for me, it's more identifying. Ok, there's something there. I need to go back and check (normal scans).* [P31, Consultant]

Overall, participants described workflow change as cumulative and largely unplanned. The tool demanded continuous microadjustments rather than a one-time shift in practice. What emerged was a pattern of individual adaptation rather than coordinated redesign; clinicians modified existing routines to fit the tool rather than the tool being aligned with the established clinical workflow. The extent of workflow disruption experienced by clinicians also reflected the realities of a service marked by rotating staff, shifting caseloads, and variable daily pressures.

Technology

The NASSS technology domain is mapped with 2 inductive subthemes: extraneous data and information overload, aligning with the subdomain knowledge generated, and system performance and material integration, aligning with the subdomain material properties. Together, these themes captured how the technical characteristics of the AI tool shaped user experience, trust, and perceived value across implementation phases.

Knowledge Generated (Extraneous Data and Information Overload)

At baseline, a consultant anticipated risks of information overload based on previous exposure to commercial AI tools. One explained, "You could spend all day circling these things" (P5, Consultant), capturing early concerns that automated outputs might flood readers with marginal or irrelevant findings.

During peri-implementation, these concerns materialized as the system generated excessive, low-value information:

> *Too much data… You really want a traffic-light system.* [P25, Consultant]

Another registrar observed:

> *I'm not sure how many people look at it. It spits out so many images and random tables* [P20, Registrar]

These reactions pointed to an emerging pattern of signal-to-noise imbalance, where radiologists spent more time filtering artefacts than interpreting meaningful results. By postimplementation, some users described partial adaptation, learning to disregard redundant data or mentally triage the AI's output.

> *It gets a little complicated when it picks up things that are artifacts. But yeah, I can work around it.* [P43, Consultant]

However, frustration persisted among others who saw the clutter as undermining efficiency rather than enhancing it:

> *It's a waste of time. It's just clutter, you know? … I usually ignore it.* [P37, Consultant]

Across phases, information overload remained one of the most salient barriers to adoption. While individual users developed coping strategies, these adaptations reflected workaround behavior rather than genuine integration, reinforcing perceptions that the AI's knowledge output was not yet aligned with clinical reasoning or workflow needs.

Material Properties (System Performance and Material Integration)

Performance concerns were a defining feature of the AI's reception, particularly during peri-implementation. A registrar characterized it bluntly as "Not very accurate. Just a splatter approach" (P17, Registrar), reflecting the perception that the system detected excessive findings without adequate specificity. Such errors eroded trust and reduced the incentive to incorporate its output into reporting routines.

By postimplementation, participants expressed more nuanced but still divided views. Some regarded the system as useful for reassurance or cross-checking:

> *Used it more like a check-off — especially when you have things that are complex, and there are a lot of findings.* [P30, Consultant]

Others found the persistent false positives distracting and demoralizing. As one put it:

> *For me to waste time looking at it…it's circled this fecal matter in the splenic flexure.* [P28, Consultant]

Several participants emphasized that perceived technical performance shaped how often they engaged with the system at all. When lag, sensitivity issues, or interface friction increased, clinicians tended to bypass or ignore the tool. Over time, its role shifted from active decision aid to optional background reference, indicating a decline in both trust and functional value.

Interoperability problems surfaced most clearly during peri-implementation, where users described limited integration between the AI software, picture archiving and communication system (PACS), and reporting systems. One consultant explained:

> *That's high-level stuff, right? That's integrating the processing, postprocessing software with the reporting software. But we don't have that capacity.* [P21, Consultant]

They further highlighted redundancy and excess image sets, noting "way too, way too many sequences…we need to distil that down."

By postimplementation, interoperability was less salient; some technical issues with integrating the AI into the system appeared resolved, but residual inconsistencies persisted. As one registrar noted:

> *There's not… uniformity to the sequences that are made. The order that they come out, …that's different from scanner to scanner.* [P26, Registrar]

Availability also varied:

> *It's not always there. So, you've got to sort of remember…to look for.* [P26, Registrar]

XSL•FO
RenderX

Display and PACS constraints continued to affect use:

> *I don't like the way that gets displayed…that's a PACS system…how the series [are] actually displayed.* [P26, Registrar]

Across phases, the material outputs of the AI, its sensitivity, specificity, and responsiveness, directly influenced its perceived usefulness. Participants consistently linked suboptimal performance to disengagement, showing that successful technological integration required not just accuracy, but reliability, responsiveness, and design alignment with radiologists' expectations of diagnostic precision. Furthermore, while major interoperability barriers had eased, the system never fully aligned with the routine reporting infrastructure, leaving incompatibilities.

Across all technology-related subdomains, participants described a gap between what the AI produced and what clinicians could *use*. Information overload and variable system accuracy combined to erode trust and limit engagement. While technical adaptation occurred at the individual level, collective integration into practice remained constrained, signaling that technological refinement and interpretability are prerequisites for sustained adoption.

Value Proposition and Clinical Context

Across all phases, discussions of value proposition were less prominent than those of technology or organization, but two inductive subthemes mapped clearly to the NASSS value proposition domain: perceived benefits for clinical workload and safety, which mapped to demand-side value, and credibility and clarity of purpose, which mapped to supply-side value. These intersected closely with the evolving clinical context, in which fluctuating workload pressures and infrastructure challenges shaped how the AI's value was interpreted.

Demand-Side Value (Perceived Benefits for Clinical Workload and Safety)

At baseline, participants viewed the AI as a potential solution to workload strain and reporting delays. Anticipated benefits were framed around efficiency, redistribution of tasks, and registrar support, signaling early optimism that automation would enhance throughput and safety. The department's intense workload and frequent interruptions reinforced this demand-side appeal: as a consultant noted, AI might "make our job easier" (P1, Consultant).

During peri-implementation, optimism gave way to more conditional appraisals. While some identified benefits for prioritization, "It highlights a few cases that you can look at first. That's useful when there's a backlog" (P19, Registrar), others described it as "unreliable at the moment" (P17, Registrar). Shifts in the clinical environment also tempered expectations, staffing improved, and backlogs eased. By the postimplementation phase, perceptions of value became pragmatic and evidence-driven. Clinicians viewed the AI as a limited but occasionally useful decision support tool:

> *I've usually written my report before I look at this, and I don't tend to change the report… it's another look, I wouldn't think of it as more than that* [P29, Consultant]

Concerns over cost-efficiency persisted:

> *If it was free, ambivalent… If it's significant amounts of money… I don't see the value because it's more work than less.* [P29, Consultant]

Across phases, expectations regarding the AI shifted from broad hopes of efficiency to a more divided assessment. Some saw modest contributions to safety and prioritization, while others viewed the system as duplicating effort rather than providing genuine workload relief.

Supply-Side Value (Credibility and Clarity of Purpose)

At the same time, participants reflected on supply-side value, questioning how clearly the system's purpose and evidence base had been articulated.

> *It just gives you pictures with circles. I'm not sure what the end use is meant to be.* [P20, Registrar]

By postimplementation, participants had a clearer understanding of what the AI could do but remained unconvinced of its overall value. However, there was also recognition that the AI was credible in concept but still immature in delivery.

> *It's getting clearer now what it could be for, but it needs to evolve. Right now, it's still just identifying, not interpreting over time.* [P43, Consultant]

Some viewed potential uses to optimize efficiency and workflow, with modifications:

> *It could identify which studies need to be reported first…or give us measurement readings.* [P46, Consultant]
>
> *Maybe some of those sorts of irritations around AI could be changed, you know, or fine-tuned.* [P25, Consultant]

These reflections indicated that perceptions of supply-side value were prospective, anchored in what the technology could deliver if optimized, rather than what it had yet achieved.

Across phases, the vendor narrative of innovation and efficiency had not yet translated into tangible or demonstrable benefit for clinicians or the wider health system. Participants viewed the AI as promising but still lacking the evidence and clarity needed to support confident investment or large-scale deployment.

Adopters

The NASSS adopter subdomain of role and identity is mapped with 2 inductive subthemes: professional positioning and negotiated use, and learning and preparedness. Together, these described how clinicians positioned AI within their expertise and accountability, and how limited exposure and training shaped trust, confidence, and uptake. Across phases, adoption reflected an oscillation between curiosity and skepticism, with trust becoming the key mediating factor.

XSL•FO
RenderX

Role and Identity (Professional Positioning and Negotiated Use)

At baseline, radiologists expressed a guarded willingness to engage with AI framed less as enthusiasm and more as a professional obligation.

> *I think I would use it…it would be almost negligent not to look at it.* [P6, Registrar]

Consultants saw potential for practical support:

> *It could certainly help you prioritize what you are watching and what order you report things.* [P8, Consultant]

During peri-implementation, practical experience unsettled this cautious trust. Registrars described false positives, excessive outputs, and low sensitivity:

> *It's too junior at the moment.* [P17, Registrar]

Trust eroded not from resistance to innovation, but from inconsistency between the AI's promise and its performance. Clinicians voiced a recurring sentiment that while AI might one day assist safety, it currently distracts from clinical focus:

> *If the volume of data presented is overwhelming, then that's negative…the strength would be as a safety net for subtle findings, not changing an overall clinical picture.* [P19, Registrar]

Reflecting on their initial use of the tool during peri-implementation, a registrar noted:

> *"It was picking up stuff that wasn't nodules…I still had to go back and look at the images again.* [P34, Registrar]

By postimplementation, there was selective use and partial trust.

> *I look through the nodules myself first and then correlate with the software to see whether it is congruent with what I've come up with.* [P42, Consultant]

Others disengaged entirely:

> *It slows you down because you have to verify each little dot.* [P37, Consultant]

A senior consultant likened the AI to "the registrar with clever ideas, but they're all wrong" (P40, Consultant), useful for prompting review, yet unreliable without human correction.

Trust also intersected with medicolegal anxiety. Several raised uncertainties about accountability and liability:

> *If the software makes a mistake, who is liable—the vendor or the radiologist? We still haven't ironed it out.* [P34, Registrar]

This uncertainty reinforced their instinct to retain manual control. As a consultant observed:

> *If I reported every possible little dot in the chest, I'd end up with a report ten pages long, which nobody would ever read.* [P37, Consultant]

The line between cautious trust and defensive practice remained thin.

Role and Identity (Learning and Preparedness)

Training and readiness remained persistently underdeveloped. During peri-implementation, there was no structured orientation or clear introduction to the system. Learning was largely self-directed and reliant on peer exchange.

> *Personally, I don't think I've had any formal sit-down with it… I've just figured it out.* [P19, Registrar]

A vendor demonstration was held midphase, but not all clinicians attended, and some felt it was disconnected from practical workflow.

> *I just met the software without gathering any prior information about what this new software is.* [P34, Registrar]

Without clear instruction or transparency about performance parameters, early experiences became a process of trial and error rather than guided adoption, reinforcing skepticism instead of trust. By postimplementation, clinicians explicitly called for structured and continuous AI education embedded within clinical and professional frameworks.

> *If that is incorporated into our routine…every month we have our session doing AI cases.* [P32, Consultant]

Others stressed the need for broader institutional responsibility:

> *We are severely lacking in training with AI…it should be an integral, assessed part of our training program.* [P34, Registrar]

These calls reflected not only a desire for technical competence but also a wish to rebuild confidence and ensure medicolegal clarity, positioning AI as a tool that must be professionally standardized, not individually improvised.

Ultimately, clinicians saw AI competence as a new layer of professional literacy, necessary to protect judgment, maintain accountability, and engage critically with emerging tools. Their learning needs were not purely technical but ethical and epistemic: how to weigh evidence, interpret probability, and remain vigilant in an era of shared decision automation.

Wider System

Participants described the wider system as a persistent barrier across phases. This domain reflects the external political, policy, and institutional forces, such as regulation, professional guidance, legislative, and funding models that define the environment in which implementation occurs, but which local teams cannot directly control. While wider system themes were present across phases, they did not dominate every interview, and some subissues (for example, explicit references to legislation) appeared only sporadically.

At baseline, consultants depicted a public system tolerant of inefficiency and difficult to influence. Funding constraints were raised in the context of competing pressures.

> *Q-Health…there's not much money around for these sorts of things.* [P1, Consultant]

During peri-implementation, registrars and consultants highlighted gaps in professional guidance and medicolegal expectations.

XSL•FO
RenderX

> *Training/communication…college says we need to learn AI, but little practical guidance…site-to-site differences.* [P24, Registrar]

Others wanted clearer, proactive communication from professional bodies.

> *College exposure is low.* [P27, Registrar]

Medicolegal norms were seen to expand review obligations when AI added extra views.

> *Mentality in radiology, if it's on a screen, you have to comment on it, and medico-legal, if presented, we must review everything.* [P21, Consultant]

By postimplementation, some brought up broader ethical and policy concerns about data provenance and the social license concerns about using it:

> *There is concern about the way the data is being used…if all these algorithms are being trained on everyone's data, it should be open source…it's everyone's.* [P42, Registrar]

Participants contrasted public and private system incentives and capacity, linking retention and deployment choices to wider economics and case-mix.

> *Public keeps me for complex cases, teaching, and feedback; private pays double.* [P28, Consultant]
>
> *Public vs private…pay and tech are better in private; public has collegiality and case mix.* [P42, Registrar]

Across phases, the wider system was characterized by limited policy levers and still developing structural and legislative readiness to support the rapid integration of AI into acute health care. Taken together, these wider-system influences interacted with internal organizational dynamics, staffing fluctuations, workload variability, and shifting operational priorities to shape the evolving trajectory of implementation across phases.

Discussion

Principal Findings

This study reports a prospective, qualitative, end-to-end evaluation of implementing an AI-driven clinical decision support system in a public radiology department, structured through the NASSS framework [30]. By mapping barriers and enablers across domains and phases, the study captures how early expectations shaped adoption, how sociotechnical challenges emerged during the rollout, and how these dynamics influenced long-term integration and adoption. Findings highlight that successful AI adoption depends not only on technical capability but on the alignment of organizational readiness, workflow design, and professional trust. Implementation success was governed by the interaction of multiple NASSS domains, which included interdependencies among technology, organization, adopters, and value rather than any single factor.

Weak planning and limited feedback structures (organizational barriers) amplified adopter frustrations with false positives, information clutter, and interoperability issues (technology barriers), eroding trust (adopter-level barriers). Even when technical faults were later mitigated, these initial experiences limited or constrained uptake, illustrating how early technical and communication failures created enduring impressions that shaped subsequent patterns of trust and tool use.

This mutual reinforcement of challenges across domains aligns with the complexity perspective described by Greenhalgh et al [30] and Braithwaite et al [40], whereby interacting barriers within and across a complex adaptive system, such as health care, tend to compound rather than resolve over time, particularly when they are not addressed in a coordinated and simultaneous manner. Clinicians' perceptions of value were shaped primarily by how reliably and efficiently the AI system performed within everyday reporting workflows. Early false positives and excessive image sets undermined those expectations, diminishing confidence in the tool's promised efficiency benefits. This finding is consistent with a recent qualitative study showing that workflow fit determines perceived usefulness, even for AI [41]. Participants described the system as "a check-off" tool rather than an integrated aid, and acceptance was suboptimal across our study; consistent with a 2023 semistructured interview with radiologists (n=25), which identified reliability, interpretability, and feedback transparency as decisive for AI acceptance [14]. Even after performance improved, initial mistrust persisted. This enduring skepticism also mirrors broader evidence that initial experiences set adoption trajectories and that trust is far easier to lose than to regain [42]. The finding reinforces the importance of consolidating the first-use experience through predeployment testing or "shadow mode" configurations [43].

Organizational conditions were vital in shaping clinician engagement. During peri-implementation, fragmented communication and limited training left some staff unaware that the system had gone live, while others lacked the confidence to use it effectively. This mirrors the work of our team and others who have consistently highlighted that structured rollout, anticipatory planning, and capacity building are critical for sustainable digital adoption [32,44-47]. Participants described the process as reactive and isolated rather than coordinated, reflecting the absence of a shared sense of purpose and mutual accountability between leadership, implementers, and users, necessary for effective implementation [48]. Although clinicians remained receptive to AI, they expected visible organizational commitment through ongoing education, rapid troubleshooting, and coherent leadership. In its absence, they relied on informal workarounds and peer support to conduct work-as-done, the adaptive, improvised practices that frontline staff develop to keep systems functioning when formal processes or resources fall short [49]. While such adaptations can sustain local functionality, they also introduce variability in care delivery and make it difficult to scale and standardize best practices across settings.

These organizational shortfalls also shaped how clinicians experienced implementation. In the absence of clear planning and coordination, several described feeling individually responsible for interpreting and integrating the AI tool into their workflow. This was not an explicit transfer of responsibility but reflected a professional culture in which clinicians relied on their own judgement to make the system workable within

XSL•FO
RenderX

local constraints. Medicolegal uncertainty about liability further reinforced this guarded engagement. A 2021 narrative review observed that when accountability is unclear or diffuse, clinicians maintain human oversight to protect both patient safety and professional authority [50]. These dynamics highlight a core implementation challenge: without clear institutional responsibility and evidentiary assurance, professional caution becomes self-reinforcing, constraining experimentation, shared learning, and the normalization of AI within routine practice.

The single-site public hospital context influenced organizational capacity. Frequent staff rotations and high turnover meant that many clinicians engaged with the AI system intermittently, limiting opportunities for cumulative learning. Similar patterns are common across public health services, where workforce mobility and resource constraints make it difficult to sustain iterative improvement [51]. These dynamics interacted with the implementation process itself, shaping the pace and pattern of adoption and modulating how performance issues were perceived over time. Rather than functioning as external confounders, they formed part of the organizational ecology within which the AI system was introduced, influencing continuity, familiarity, and the stability of feedback loops essential for embedding new technologies. These conditions underscore the importance of implementation approaches that are designed for continuity, including modular onboarding, periodic refresher training, and accessible repositories of AI-related resources to support learning across changing teams.

This study extends the AI-in-radiology literature in 3 key ways. It adopts a temporal perspective, tracing implementation from early anticipation to postintegration and showing how initial optimism and early missteps shape later engagement. It also demonstrates that adoption was driven not by the innovation alone but by the interaction of technical performance, organizational coordination, and context, extending NASSS from description to explanation. In this framing, staff turnover, workload fluctuations, and shifting operational priorities function as active contextual determinants that help explain why implementation trajectories evolve as they do, rather than as background noise. Finally, through reflexive use of the framework, the study generates theoretical insight into why implementation evolves as it does, contributing to emerging work on domain interdependence and temporal complexity in health care innovation [52].

Grounded in our findings and consistent with previous guidance [23,30,53,54], effective AI implementation in radiology depends on a combination of technical stability, communication, and organizational preparedness. Early “shadow mode” piloting helps identify faults and build trust before clinical use, while consistent communication about progress and fixes maintains transparency. Workflow-compatible design that minimizes cognitive load supports efficiency and acceptance [55]. Ongoing professional development, formal feedback channels with vendor responsiveness, and planning for workforce turnover through shared training repositories and local champions help sustain capability over time. Together, these strategies align with the six principles of FUTURE-AI recommendations encompassing fairness, universality, traceability, usability, robustness, and explainability, an international expert-driven consensus to facilitate the adoption of trustworthy medical imaging [56] and emphasize that co-design and iterative learning are essential to long-term adoption.

Study Limitations and Strengths

This was a single-site study, and further real-world, ecological research is needed to identify determinants of adoption and create solutions that generalize across health systems. Despite this, our findings are similar to several recently published studies examining radiologists’ perceptions around the adoption of AI into standard practice [2,12,14,57-60]. There was low uptake of the tool during the peri-implementation period due to technical challenges. This limited wider evaluation of this crucial implementation period, particularly in terms of clinical usefulness. The 18-month period may have also introduced a range of system confounders, including staffing changes and organizational priorities, which may have impacted how participants felt about the AI clinical decision support tool. Finally, social desirability bias cannot be ruled out, particularly as this was a department-wide implementation. However, this was mitigated through participant briefings emphasizing the exploratory nature of the trial, coupled with reflexive practice by the interview team [61]. A key study strength was that this was a real-world evaluation demonstrating ecological validity with findings grounded in practical realities. Second, aligning our findings with a validated implementation science framework supports theoretical transferability and future application in related contexts despite the single-site limitations.

Future Implications

Future studies should integrate qualitative and quantitative data, combining workflow observations with metrics such as reporting time, error rates, and AI usage logs, to triangulate findings. Multisite evaluations across differing levels of digital maturity are needed to test transferability and examine how governance, culture, and workforce patterns influence scalability.

Conclusion

Implementation of AI-based decision support in radiology is as much an organizational and cultural process as a technological one. Clinicians remain willing to engage, but sustainable adoption depends on consolidating early experiences, embedding communication and training, and maintaining iterative feedback between users, vendors, and system leaders. Applying the NASSS framework revealed how domains interact dynamically across time, offering both theoretical insight into sociotechnical complexity and practical guidance for hospitals seeking to move from pilot to routine, trustworthy AI integration.

Acknowledgments

The authors would like to gratefully acknowledge the study participants who gave their valuable time to participate in this study. The authors are grateful to Dr Sue Jeavons and her staff for facilitating this study onsite.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are not publicly available due to confidentiality policies, but may be available in limited form from the corresponding author on reasonable request.

Funding

This work was supported by Digital Health CRC Limited (DHCRC), funded under the Commonwealth Cooperative Research Centres (CRC) program. SM is supported by a fellowship from the National Health and Medical Research Council (NHMRC; #1181138). The funders had no role in the study design or the decision to submit for publication.

Authors' Contributions

Conceptualization: SN, SM
Methodology: SN, PS, BS, AT, AE, SM
Formal analysis: SN, AE
Writing – original draft: SN
Writing – review & editing: SN, PS, BS, AT, AE, SM

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ (Consolidated Criteria for Reporting Qualitative Studies) checklist.
[PDF File (Adobe PDF File), 486 KB - jmir_v28i1e80342_app1.pdf]

Multimedia Appendix 2
Interview topic guide.
[DOCX File , 29 KB - jmir_v28i1e80342_app2.docx]

References

1. Maskell G. Why Does Demand for Medical Imaging Keep Rising?. Hoboken, NJ: British Medical Journal Publishing Group; 2022.
2. Chauhan AS, Singh R, Priyadarshi N, Twala B, Suthar S, Swami S. Unleashing the power of advanced technologies for revolutionary medical imaging: pioneering the healthcare frontier with artificial intelligence. Discov Artif Intell 2024;4(1):58. [doi: 10.1007/s44163-024-00161-0]
3. Ding A, Joshi J, Tiwana E. Patient safety in radiology and medical imaging. In: Agrawal A, Bhatt J, editors. Patient Safety: A Case-Based Innovative Playbook for Safer Care. Cham: Springer; 2023:261-277.
4. Waite S, Scott J, Colombo D. Narrowing the Gap: imaging disparities in radiology. Radiology 2021;299(1):27-35. [doi: 10.1148/radiol.2021203742] [Medline: 33560191]
5. Kelly BS, Judge C, Bollard SM, Clifford SM, Healy GM, Aziz A, et al. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur Radiol 2022;32(11):7998-8007 [FREE Full text] [doi: 10.1007/s00330-022-08784-6] [Medline: 35420305]
6. Chau M. Ethical, legal, and regulatory landscape of artificial intelligence in Australian healthcare and ethical integration in radiography: a narrative review. J Med Imaging Radiat Sci 2024;55(4):101733. [doi: 10.1016/j.jmir.2024.101733] [Medline: 39111223]
7. Lu P, Mian M, Yii M, McArdle DJT, Rhodes A, Sreedharan S. Rising use of diagnostic imaging in Australia: an analysis of medicare-funded radiology services between 2000 and 2021. J Med Imaging Radiat Oncol 2024;68(1):50-56. [doi: 10.1111/1754-9485.13591] [Medline: 37797195]
8. Zheng Q, Yang L, Zeng B, Li J, Guo K, Liang Y, et al. Artificial intelligence performance in detecting tumor metastasis from medical radiology imaging: a systematic review and meta-analysis. EClinicalMedicine 2021;31:100669 [FREE Full text] [doi: 10.1016/j.eclinm.2020.100669] [Medline: 33392486]
9. van den Broek MCL, Buijs JH, Schmitz LFM, Wijffels MME. Diagnostic performance of artificial intelligence in rib fracture detection: systematic review and meta-analysis. Surgeries 2024;5(1):24-36. [doi: 10.3390/surgeries5010005]
10. Lawrence R, Dodsworth E, Massou E, Sherlaw-Johnson C, Ramsay AI, Walton H, et al. Artificial intelligence for diagnostics in radiology practice: a rapid systematic scoping review. EClinicalMedicine 2025;83:103228 [FREE Full text] [doi: 10.1016/j.eclinm.2025.103228] [Medline: 40474995]
11. Sosna J, Joskowicz L, Saban M. Navigating the AI landscape in medical imaging: a critical analysis of technologies, implementation, and implications. Radiology 2025;315(3):e240982. [doi: 10.1148/radiol.240982] [Medline: 40552997]

XSL•FO
RenderX

12. Strohm L, Hehakaya C, Ranschaert ER, Boon WPC, Moors EHM. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. Eur Radiol 2020;30(10):5525-5532. [doi: 10.1007/s00330-020-06946-y] [Medline: 32458173]
13. Alarifi M. Radiologists' views on AI and the future of radiology: insights from a US national survey. Br J Radiol 2025. [doi: 10.1093/bjr/tqaf222] [Medline: 40971485]
14. Bergquist M, Rolandsson B, Gryska E, Laesser M, Hoefling N, Heckemann R, et al. Trust and stakeholder perspectives on the implementation of AI tools in clinical radiology. Eur Radiol 2024;34(1):338-347 [FREE Full text] [doi: 10.1007/s00330-023-09967-5] [Medline: 37505245]
15. Hua D, Petrina N, Young N, Cho J, Poon SK. Understanding the factors influencing acceptability of AI in medical imaging domains among healthcare professionals: a scoping review. Artif Intell Med 2024;147:102698 [FREE Full text] [doi: 10.1016/j.artmed.2023.102698] [Medline: 38184343]
16. Ross J, Hammouche S, Chen Y, Rockall A, Royal College of Radiologists AI Working Group. Beyond regulatory compliance: evaluating radiology artificial intelligence applications in deployment. Clin Radiol 2024;79(5):338-345 [FREE Full text] [doi: 10.1016/j.crad.2024.01.026] [Medline: 38360516]
17. Arkoh S, Akudjedu TN, Amedu C, Antwi WK, Elshami W, Ohene-Botwe B. Current radiology workforce perspective on the integration of artificial intelligence in clinical practice: a systematic review. J Med Imaging Radiat Sci 2025;56(1):101769. [doi: 10.1016/j.jmir.2024.101769] [Medline: 39437624]
18. Hemphill S, Jackson K, Bradley S, Bhartia B. The implementation of artificial intelligence in radiology: a narrative review of patient perspectives. Future Healthc J 2023;10(1):63-68 [FREE Full text] [doi: 10.7861/fhj.2022-0097] [Medline: 37786489]
19. Wenderott K, Krups J, Weigl M, Wooldridge AR. Facilitators and barriers to implementing AI in routine medical imaging: systematic review and qualitative analysis. J Med Internet Res 2025;27:e63649 [FREE Full text] [doi: 10.2196/63649] [Medline: 40690758]
20. Buijs E, Maggioni E, Mazziotta F, Lega F, Carrafiello G. Clinical impact of AI in radiology department management: a systematic review. Radiol Med 2024;129(11):1656-1666 [FREE Full text] [doi: 10.1007/s11547-024-01880-1] [Medline: 39243293]
21. Kim B, Romeijn S, van Buchem M, Mehrizi MHR, Grootjans W. A holistic approach to implementing artificial intelligence in radiology. Insights Imaging 2024;15(1):22 [FREE Full text] [doi: 10.1186/s13244-023-01586-4] [Medline: 38270790]
22. Hogg HDJ, Al-Zubaidy M, Technology Enhanced Macular Services Study Reference Group, Talks J, Denniston AK, Kelly CJ, et al. Stakeholder perspectives of clinical artificial intelligence implementation: systematic review of qualitative evidence. J Med Internet Res 2023;25:e39742 [FREE Full text] [doi: 10.2196/39742] [Medline: 36626192]
23. Hogg HDJ, Sendak MP, Denniston AK, Keane PA, Maniatopoulos G. Unlocking the potential of qualitative research for the implementation of artificial intelligence-enabled healthcare. J Med Artif Intell 2023;6:8-8. [doi: 10.21037/jmai-23-28]
24. Jarrahi MH. Interviewing AI: Using qualitative methods to explore and capture machines' characteristics and behaviors. Big Data & Society 2025;12(3). [doi: 10.1177/20539517251381697]
25. Farič N, Hinder S, Williams R, Ramaesh R, Bernabeu M, van Beek E, et al. Early experiences of integrating an artificial intelligence-based diagnostic decision support system into radiology settings: a qualitative study. J Am Med Inform Assoc 2023;31(1):24-34 [FREE Full text] [doi: 10.1093/jamia/ocad191] [Medline: 37748456]
26. Song Y, MacEachern L, Doupe MB, Ginsburg L, Chamberlain SA, Cranley L, et al. Influences of post-implementation factors on the sustainability, sustainment, and intra-organizational spread of complex interventions. BMC Health Serv Res 2022;22(1):666 [FREE Full text] [doi: 10.1186/s12913-022-08026-x] [Medline: 35581651]
27. Donovan T, Carter HE, McPhail SM, Abell B. Challenges and recommendations for collecting and quantifying implementation costs in practice: a qualitative interview study. Implement Sci Commun 2024;5(1):114 [FREE Full text] [doi: 10.1186/s43058-024-00648-y] [Medline: 39394175]
28. Fernando M, Abell B, Tyack Z, Donovan T, McPhail SM, Naicker S. Using theories, models, and frameworks to inform implementation cycles of computerized clinical decision support systems in tertiary health care settings: scoping review. J Med Internet Res 2023;25:e45163 [FREE Full text] [doi: 10.2196/45163] [Medline: 37851492]
29. Lewis CC, Boyd MR, Walsh-Bailey C, Lyon AR, Beidas R, Mittman B, et al. A systematic review of empirical studies examining mechanisms of implementation in health. Implement Sci 2020;15(1):21 [FREE Full text] [doi: 10.1186/s13012-020-00983-3] [Medline: 32299461]
30. Greenhalgh T, Abimbola S. The NASSS framework - a synthesis of multiple theories of technology implementation. Stud Health Technol Inform 2019;263:193-204. [doi: 10.3233/SHTI190123] [Medline: 31411163]
31. Abell B, Naicker S, Rodwell D, Donovan T, Tariq A, Baysari M, et al. Identifying barriers and facilitators to successful implementation of computerized clinical decision support systems in hospitals: a NASSS framework-informed scoping review. Implement Sci 2023;18(1):32 [FREE Full text] [doi: 10.1186/s13012-023-01287-y] [Medline: 37495997]
32. Fernando M, Abell B, McPhail SM, Tyack Z, Tariq A, Naicker S. Applying the non-adoption, abandonment, scale-up, spread, and sustainability framework across implementation stages to identify key strategies to facilitate clinical decision support system integration within a large metropolitan health service: interview and focus group study. JMIR Med Inform 2024;12:e60402 [FREE Full text] [doi: 10.2196/60402] [Medline: 39419497]

XSL•FO
RenderX

33. Cordeiro JV. Artificial intelligence and precision public health: a balancing act of scientific accuracy, social responsibility, and community engagement. Port J Public Health 2024;42(1):1-5. [doi: 10.1159/000538141] [Medline: 39495190]
34. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907 [FREE Full text] [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
35. Rettke H, Pretto M, Spichiger E, Frei I, Spirig R. Using reflexive thinking to establish rigor in qualitative research. Nurs Res 2018;67(6):490-497. [doi: 10.1097/NNR.0000000000000307] [Medline: 30067583]
36. Nordmann K, Sauter S, Redlich M, Möbius-Lerch P, Schaller M, Fischer F. Challenges and conditions for successfully implementing and adopting the telematics infrastructure in German outpatient healthcare: a qualitative study applying the NASSS framework. Digital Health 2024;10:20552076241259855 [FREE Full text] [doi: 10.1177/20552076241259855] [Medline: 39070890]
37. Proudfoot K. Inductive/deductive hybrid thematic analysis in mixed methods research. J Mix Methods Res 2022;17(3):308-326. [doi: 10.1177/15586898221126816]
38. Skillman M, Cross-Barnet C, Friedman Singer R, Rotondo C, Ruiz S, Moiduddin A. A framework for rigorous qualitative research as a component of mixed method rapid-cycle evaluation. Qual Health Res 2019;29(2):279-289. [doi: 10.1177/1049732318795675] [Medline: 30175660]
39. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: implications for conducting a qualitative descriptive study. Nurs Health Sci 2013;15(3):398-405. [doi: 10.1111/nhs.12048] [Medline: 23480423]
40. Braithwaite J, Churruca K, Long JC, Ellis LA, Herkes J. When complexity science meets implementation science: a theoretical and empirical analysis of systems change. BMC Med 2018;16(1):63 [FREE Full text] [doi: 10.1186/s12916-018-1057-z] [Medline: 29706132]
41. Wenderott K, Krups J, Luetkens JA, Weigl M. Radiologists' perspectives on the workflow integration of an artificial intelligence-based computer-aided detection system: a qualitative study. Appl Ergon 2024;117:104243 [FREE Full text] [doi: 10.1016/j.apergo.2024.104243] [Medline: 38306741]
42. Steerling E, Siira E, Nilsen P, Svedberg P, Nygren J. Implementing AI in healthcare-the relevance of trust: a scoping review. Front Health Serv 2023;3:1211150. [doi: 10.3389/frhs.2023.1211150] [Medline: 37693234]
43. Bizzo BC, Dasegowda G, Bridge C, Miller B, Hillis JM, Kalra MK, et al. Addressing the challenges of implementing artificial intelligence tools in clinical practice: principles from experience. J Am Coll Radiol 2023;20(3):352-360. [doi: 10.1016/j.jacr.2023.01.002] [Medline: 36922109]
44. Alotaibi N, Wilson CB, Traynor M. Enhancing digital readiness and capability in healthcare: a systematic review of interventions, barriers, and facilitators. BMC Health Serv Res 2025;25(1):500 [FREE Full text] [doi: 10.1186/s12913-025-12663-3] [Medline: 40186200]
45. Van Velthoven MH, Cordon C. Sustainable adoption of digital health innovations: perspectives from a stakeholder workshop. J Med Internet Res 2019;21(3):e11922 [FREE Full text] [doi: 10.2196/11922] [Medline: 30907734]
46. Nair M, Nygren J, Nilsen P, Gama F, Neher M, Larsson I, et al. Critical activities for successful implementation and adoption of AI in healthcare: towards a process framework for healthcare organizations. Front Digit Health 2025;7:1550459 [FREE Full text] [doi: 10.3389/fdgth.2025.1550459] [Medline: 40453810]
47. Marwaha JS, Landman AB, Brat GA, Dunn T, Gordon WJ. Deploying digital health tools within large, complex health systems: key considerations for adoption and implementation. NPJ Digit Med 2022;5(1):13 [FREE Full text] [doi: 10.1038/s41746-022-00557-1] [Medline: 35087160]
48. Rapport F, Clay-Williams R, Churruca K, Shih P, Hogden A, Braithwaite J. The struggle of translating science into action: foundational concepts of implementation science. J Eval Clin Pract 2018;24(1):117-126 [FREE Full text] [doi: 10.1111/jep.12741] [Medline: 28371050]
49. Perry SJ, Catchpole K, Rivera AJ, Henrickson Parker S, Gosbee J. 'Strangers in a strange land': Understanding professional challenges for human factors/ergonomics and healthcare. Appl Ergon 2021;94:103040 [FREE Full text] [doi: 10.1016/j.apergo.2019.103040] [Medline: 33676061]
50. Smith H. Clinical AI: opacity, accountability, responsibility and liability. AI & Society 2020;36(2):535-545. [doi: 10.1007/s00146-020-01019-6]
51. Moon SEJ, Hogden A, Eljiz K. Sustaining improvement of hospital-wide initiative for patient safety and quality: a systematic scoping review. BMJ Open Qual 2022;11(4):e002057 [FREE Full text] [doi: 10.1136/bmjoq-2022-002057] [Medline: 36549751]
52. Hellstrand Tang U, Smith F, Karilampi UL, Gremyr A. Exploring the role of complexity in health care technology bottom-up innovations: multiple-case study using the nonadoption, abandonment, scale-up, spread, and sustainability complexity assessment tool. JMIR Hum Factors 2024;11:e50889 [FREE Full text] [doi: 10.2196/50889] [Medline: 38669076]
53. Daye D, Wiggins WF, Lungren MP, Alkasab T, Kottler N, Allen B, et al. Implementation of clinical artificial intelligence in radiology: who decides and how. Radiology 2022;305(3):555-563 [FREE Full text] [doi: 10.1148/radiol.212151] [Medline: 35916673]

XSL•FO
RenderX

54. Chen Y, Stavropoulou C, Narasinkan R, Baker A, Scarbrough H. Professionals' responses to the introduction of AI innovations in radiology and their implications for future adoption: a qualitative study. BMC Health Serv Res 2021;21(1):813 [FREE Full text] [doi: 10.1186/s12913-021-06861-y] [Medline: 34389014]
55. Zhang L, LaBelle W, Unberath M, Chen H, Hu J, Li G, et al. A vendor-agnostic, PACS integrated, and DICOM-compatible software-server pipeline for testing segmentation algorithms within the clinical radiology workflow. Front Med (Lausanne) 2023;10:1241570 [FREE Full text] [doi: 10.3389/fmed.2023.1241570] [Medline: 37954555]
56. NA. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ 2025;388:r340 [FREE Full text] [doi: 10.1136/bmj.r340] [Medline: 39961614]
57. Huisman M, Ranschaert E, Parker W, Mastrodicasa D, Koci M, Pinto de Santos D, et al. An international survey on AI in radiology in 1041 radiologists and radiology residents part 2: expectations, hurdles to implementation, and education. Eur Radiol 2021;31(11):8797-8806 [FREE Full text] [doi: 10.1007/s00330-021-07782-4] [Medline: 33974148]
58. Jiang S, Bukhari SMA, Krishnan A, Bera K, Sharma A, Caovan D, et al. Deployment of artificial intelligence in radiology: strategies for success. AJR Am J Roentgenol 2025;224(2):e2431898. [doi: 10.2214/AJR.24.31898] [Medline: 39475198]
59. Do HM, Spear LG, Nikpanah M, Mirmomen SM, Machado LB, Toscano AP, et al. Augmented radiologist workflow improves report value and saves time: a potential model for implementation of artificial intelligence. Acad Radiol 2020;27(1):96-105 [FREE Full text] [doi: 10.1016/j.acra.2019.09.014] [Medline: 31818390]
60. Jones C, Thornton J, Wyatt JC. Enhancing trust in clinical decision support systems: a framework for developers. BMJ Health Care Inform 2021;28(1):e100247 [FREE Full text] [doi: 10.1136/bmjhci-2020-100247] [Medline: 34088721]
61. Cairns-Lee H, Lawley J, Tosey P. Enhancing researcher reflexivity about the influence of leading questions in interviews. J Appl Behav Sci 2021;58(1):164-188. [doi: 10.1177/00218863211037446]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
COREQ: Consolidated Criteria for Reporting Qualitative Studies
CT: computed tomography
ERIC: Expert Recommendations for Implementing Change
NASSS: nonadoption, abandonment, scale-up, spread, and sustainability
PACS: picture archiving and communication system

Edited by J Sarvestan; submitted 09.Jul.2025; peer-reviewed by S Mohanadas, L Laverty, X Liang; comments to author 04.Aug.2025; revised version received 09.Dec.2025; accepted 10.Dec.2025; published 28.Jan.2026.

Please cite as:
Naicker S, Schmidt P, Shar B, Tariq A, Earnshaw A, McPhail S
Implementing an Artificial Intelligence Decision Support System in Radiology: Prospective Qualitative Evaluation Study Using the Nonadoption Abandonment Scale-Up, Spread, and Sustainability (NASSS) Framework
J Med Internet Res 2026;28:e80342
URL: https://www.jmir.org/2026/1/e80342
doi:10.2196/80342
PMID:

©Sundresan Naicker, Paul Schmidt, Bruce Shar, Amina Tariq, Ashleigh Earnshaw, Steven McPhail. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Integrated Prediction System for Individualized Ovarian Stimulation and Ovarian Hyperstimulation Syndrome Prevention: Algorithm Development and Validation

Jingjing Chen[1,2*], MD, PhD; Jianjuan Zhao[3,4,5*], MS; Huiyu Qiu[1,2], MD, MS; Yanhui Liu[5], MD, PhD; Yunqi Zhang[3], MS; Qicheng Sun[3], BSc; Yan Yi[1,2], MD, PhD; Hongying Tang[1,2], MSN; Jing Zhao[1,2], MD, PhD; Bin Xu[1,2], MD, PhD; Qiong Zhang[1,2], MD, PhD; Ge Yang[6], MD, PhD; Hui Li[1,2], MD, PhD; Junjie Liu[5], PhD; Zhongzhou Yang[7], PhD; Shaolin Liang[3,4,5*], PhD; Yanping Li[1,2*], MD, PhD; Jing Fu[1,2*], MD, PhD

[1]Department of Reproductive Medicine, Xiangya Hospital, Central South University, Changsha, Hunan, China

[2]Clinical Research Center for Women's Reproductive Health in Hunan Province, Changsha, China

[3]Digital Health Lab, Institute for Six-sector Economy, Fudan University, Shanghai, China

[4]Xiangya Hospital, Central South University, Changsha, China

[5]Reproductive Medicine Department, The Third Affiliated Hospital of Shenzhen University, Shenzhen, China

[6]Division of Neonatology, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangdong, China

[7]School of Medicine, Sun Yat-sen University, Guangdong, China

[*]these authors contributed equally

Corresponding Author:
Jing Fu, MD, PhD
Department of Reproductive Medicine
Xiangya Hospital
Central South University
No.87 Xiangya Road, Kaifu Strict, Changsha City
Changsha, Hunan, 410008
China
Phone: 86 15874861692
Email: 360546450@qq.com

Abstract

Background: Accurately predicting ovarian response and determining the optimal starting dose of follicle-stimulating hormone (FSH) remain critical yet challenging for effective ovarian stimulation. Currently, there is a lack of a comprehensive model capable of simultaneously forecasting the number of oocytes retrieved (NOR) and assessing the risk of early-onset moderate-to-severe ovarian hyperstimulation syndrome (OHSS).

Objective: This study aimed to establish an integrated mode capable of forecasting the NOR and assessing the risk of early-onset moderate-to-severe OHSS across varying starting doses of FSH.

Methods: This prognostic study included patients undergoing their first ovarian stimulation cycles at 2 independent in vitro fertilization clinics. Automated classifiers were used for variable selection. Machine learning models (11 for NOR and 11 for OHSS) were developed and validated using internal (n=6401) and external (n=3805) datasets. Shapley additive explanation was applied for variable interpretation. The best-performing models were incorporated into a web-based prediction tool.

Results: For NOR prediction, 17 variables were selected, with the gradient boosting regressor achieving the highest performance (internal dataset: R^2=0.7978; external dataset: R^2=0.7924). For OHSS prediction, 19 variables were identified, and the LightGBM model demonstrated superior performance (internal dataset: area under the receiver operating characteristic curve=0.7588; external dataset: area under the receiver operating characteristic curve=0.7287). Shapley additive explanation analysis highlighted the FSH starting dose to BMI ratio and baseline antral follicle count as key predictors for NOR and OHSS, respectively. Dose-response curves were generated to visualize predicted outcomes with varying FSH starting doses. The models were implemented in a user-friendly, research-oriented online prototype, individualized ovarian stimulation guide (InOvaSGuide).

XSL•FO
RenderX

Conclusions: This study introduces an integrated framework for predicting NOR and early-onset moderate-to-severe OHSS risk across different FSH doses. Future prospective evaluation is needed before clinical implementation.

(J Med Internet Res 2026;28:e78245) doi:10.2196/78245

KEYWORDS

individualized ovarian stimulation; ovarian response; ovarian hyperstimulation syndrome; follicle-stimulating hormone; machine learning

Introduction

Over the past decade, individualized ovarian stimulation has become a key strategy in in vitro fertilization (IVF). Determining an appropriate starting dose of exogenous follicle-stimulating hormone (FSH) is essential for balancing efficacy and safety. Although earlier clinical practice emphasized maximizing oocyte yield (the more, the better), current consensus favors achieving a moderate ovarian response to optimize live birth rates while minimizing patient discomfort and iatrogenic risks such as ovarian hyperstimulation syndrome (OHSS). Therefore, accurate prediction of ovarian response before stimulation is critical for optimizing treatment outcomes [1-4].

Although biomarkers, including antral follicle count (AFC), anti-Müllerian hormone (AMH) levels, and BMI, are well associated with ovarian response, substantial interindividual and intraindividual variability limits their predictive precision. Tailoring FSH doses based solely on these indicators has not consistently improved clinical outcomes [5,6], highlighting the need for more comprehensive, data-driven approaches that integrate a broader spectrum of clinical and biological factors.

Recent advances in artificial intelligence (AI) and machine learning (ML) offer new opportunities for improving decision-making in assisted reproduction, with applications reported in semen analysis [7], blastocysts grading [8], and trigger-day assessments [9]. Several ML models have also been developed to predict the number of oocytes retrieved (NOR) [10-13] or to classify ovarian responsiveness [10]; however, most remain limited in scope. They typically rely on a narrow set of baseline features, adopt single-model frameworks, and focus predominantly on treatment efficacy such as oocyte yield, with relatively limited attention to safety outcomes, including OHSS. These limitations emphasize the need for predictive frameworks that simultaneously incorporate both efficacy and safety. Furthermore, despite multiple evidence-based algorithms for FSH dosing, considerable variability in ovarian response persists even among patients with comparable baseline characteristics. A model that jointly predicts NOR and OHSS risk across a range of FSH doses may provide useful predictive information and support dose-specific decision-making, helping clinicians consider the balance between efficacy and safety when selecting individualized FSH doses.

In this study, ML models were developed to predict NOR and early-onset moderate-to-severe OHSS using datasets from 2 IVF centers. Models with optimal performance were integrated into a clinician-oriented decision support prototype, termed individualized ovarian stimulation guide (“InOvaSGuide”), complemented by a web-based calculator. For each patient, the system provides individualized dose-response curves that display predicted NOR and early-onset moderate-to-severe OHSS probabilities across varying FSH starting doses, thus supporting personalized ovarian stimulation.

Methods

Ethical Considerations

This prognostic study was designed as a retrospective analysis and was approved by the Reproductive Medicine Ethics Committee of Xiangya Hospital (2021010) and the Medicine Ethics Committee of Shenzhen Luohu District People’s Hospital (2024-LHQRMYY-KYLL-63). Informed consent was waived because all data were retrospectively collected from routine clinical records and anonymized before analysis. The study adhered to the Declaration of Helsinki and followed the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis; Table S1 in Multimedia Appendix 1) reporting guideline [14]. All data analyses were performed by an external team using anonymized data only, ensuring full protection of participant privacy. No compensation was provided to participants, as the study involved retrospective and fully anonymized data.

Study Cohort

The inclusion criteria were (1) patients with the first ovarian stimulation cycle conducted between January 1, 2018, and September 30, 2022, at the Department of Reproductive Medicine of Xiangya Hospital (internal dataset) and between May 1, 2021, and December 30, 2023, at the Reproductive Center of Shenzhen Luohu District People’s Hospital (external dataset) and (2) patients aged 20 to 40 years. Exclusion criteria were (1) patients with diminished ovarian reserve, diagnosed by AMH ≤1.1 ng/mL or baseline AFC ≤7 [15]; (2) patients using the microstimulation protocols for ovarian stimulation, including progestin-primed ovarian stimulation protocol, natural cycle protocol, etc; and (3) patients with more than 50% missingness in key clinical variables. Notably, patients with diminished ovarian reserve or those undergoing microstimulation protocols were excluded because these groups require highly individualized stimulation strategies, exhibit markedly lower oocyte yield, and have a substantially reduced risk of OHSS under comparable FSH exposure, which would have created pronounced class imbalance and reduced model robustness.

After screening, 6401 patients from Xiangya Hospital and 3805 from Shenzhen Luohu District People’s Hospital were included in the internal and external datasets, respectively.

XSL•FO
RenderX

Ovarian Stimulation Process and the Diagnosis of Early-Onset Moderate-to-Severe OHSS

Before commencing the IVF and intracytoplasmic sperm injection cycle, patients underwent a thorough physical examination, including the assessment of basic physical parameters (height, weight, and BMI), measurement of basal hormone levels (FSH, luteinizing hormone [LH], estradiol, testosterone, progesterone, prolactin, and AMH), biochemical tests (fasting glucose, fasting insulin, lipid levels, and thyroid hormones), and transvaginal ultrasonography for basal AFC. Subsequently, experienced physicians personalized the ovarian stimulation protocol and the starting dose of FSH based on comprehensive clinical assessment. Throughout stimulation, patients underwent monitoring via transvaginal ultrasonography and serum hormone assessments, with gonadotropin dosage adjustments made according to individual ovarian responses. Human chorionic gonadotropin or gonadotropin-releasing hormone agonist, alone or combined, triggered oocyte maturation when 3 or more follicles measuring 17 mm or greater were observed. Oocyte retrieval occurred 36 hours after triggering, and the NOR was recorded. Eligible patients underwent fresh embryo transfer with 1 or 2 embryos.

The study primarily focused on the occurrence of early-onset, moderate-to-severe OHSS, which was diagnosed within 9 days after triggering based on established guidelines [16], considering both clinical and laboratory features. All relevant individual and clinical variables during the process were obtained from the clinical database for feature screening and selection.

Data Preprocessing and Feature Selection

All data analysis was performed by an external team using anonymized data, ensuring full protection of participant privacy. To elucidate correlations between predictive features and clinical outcomes with an emphasis on medical interpretability, we performed feature engineering on selected variables. This process resulted in 2 additional variables: “FSH to LH ratio” and “FSH starting dose to BMI ratio,” which improved predictive accuracy while maintaining transparency and clinical relevance. Furthermore, to address skewness and improve distributional normality, a logarithmic transformation was applied to AMH, triglycerides, and FSH/LH (Figures S1 and S2 in Multimedia Appendix 1). Missing data were handled using mean imputation, with feature-wise means computed exclusively from the training set and subsequently applied to the test and external validation sets, to prevent information leakage. The overall proportion of missingness was low, and imputation did not materially alter variable distributions.

To identify key variables, we applied feature importance–based selection using the Boruta algorithm, performed exclusively within the training dataset (Figures S3 and S4 in Multimedia Appendix 1). This approach led to the selection of 17 variables for the NOR prediction model and 19 variables for the OHSS prediction model.

NOR Model Development

In the prediction of NOR, 17 features, including starting dose of FSH to BMI ratio, BMI, log (AMH), and specifically the ovarian stimulation protocol, were selected. The NOR divided by the starting dose of FSH was used and logarithmically transformed as the outcome variable with improved predictive performance. For model training, the internal dataset was divided into an 8:2 split, with 79.9% (5120/6401) of the data randomly allocated to the training set and the remaining 20% (1281/6401) assigned to the internal test set. All data from the external dataset were held out entirely and used exclusively as an external validation cohort, providing an independent assessment of model generalizability across institutions. Eleven ML algorithms, including a linear regression model, were trained to predict the preprocessed NOR outcome. Hyperparameter tuning was performed using 5-fold cross-validation within the training set only, with all hyperparameters predefined and summarized in Table S2 in Multimedia Appendix 1. Model performance was assessed using 4 key metrics: R^2, adjusted R^2, mean absolute error, and root mean square error.

OHSS Model Development

For OHSS prediction, the target variable was the occurrence of early-onset moderate-to-severe OHSS. Given the low event rate and resulting class imbalance, several commonly used imbalances handling strategies (eg, oversampling, undersampling, and ensemble-based resampling) were evaluated. Cost-sensitive learning was ultimately adopted, assigning differentiated penalties to misclassifications while preserving all original clinical data distribution. To prevent overfitting, model complexity was controlled by limiting the number of parameters and applying regularization techniques, as appropriate for each algorithm. Eleven ML algorithms were implemented, with corresponding hyperparameters detailed in Table S3 in Multimedia Appendix 1. As with the NOR model, hyperparameter optimization was conducted exclusively within the training set, and model performance was evaluated on both the internal and the external datasets using the area under the receiver operating characteristic curve (ROC-AUC), the precision-recall area under the curve (PR-AUC), recall, specificity, weighted F_1-score, Cohen κ, and positive and negative predictive values.

Shapley Additive Explanation Value

To further explore the significant features driving the model’s predictions, we used the Shapley additive explanation (SHAP) analysis to assess the importance of core features. SHAP serves as an interpretative tool for ensemble tree models, offering a detailed breakdown of the influence of input features on predictions.

Creation of Dose-Response Curves

In this study, 2 distinct models were developed: a classification model for predicting early-onset moderate-to-severe OHSS and a regression model for forecasting NOR. Models of best performance were incorporated into an integrated, research-oriented computational system, complemented by a web-based calculator. By inputting baseline patient characteristics, the system generates predictions for NOR and early-onset moderate-to-severe OHSS probability, presented as a dose-response curve illustrating changes with increasing FSH starting doses.

XSL•FO
RenderX

Statistical Analysis

The baseline characteristics of patients between the internal dataset and the external dataset were compared using the chi-square test for categorical variables. For continuous variables, we assessed normality using the Shapiro-Wilk W test. Depending on the results, we used either the 2-tailed Student *t* test or the Mann-Whitney *U* test for comparison. R Studio (version 4.3.1; R Foundation for Statistical Computing), Python (version 3.11.4; Python Software Foundation), the open-source *scikit-learn* package (version 3.9.13; open-source community-developed Python ML library), LightGBM (version 4.3.0; Microsoft Corporation), and XGBoost (version 2.0.0; open-source project maintained by XGBoost contributors) were used for model development and statistical analyses.

Results

The Integrated Ovarian Response Prediction System

To address the challenges of individualized ovarian stimulation, we developed an integrated prediction system, "InOvaSGuide," designed to predict both the NOR and the probability of early-onset moderate-to-severe OHSS before ovarian stimulation. The system was built using datasets from 2 IVF clinics and incorporates 2 distinct ML models for NOR and OHSS predictions, respectively (Figure 1A). By analyzing patients' baseline characteristics, the system generates dose-response curves that illustrate the predicted benefit (NOR) and risk (early-onset moderate-to-severe OHSS) across varying FSH starting doses (Figures 1B and C). Additionally, a user-friendly web-based calculator was developed to enhance accessibility and support exploratory use in clinically relevant contexts (Figure S7 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 1. Flowchart of the study: (A) the modeling process with internal and external datasets, (B) illustration of the primary goal of this study, and (C) illustration of the clinical application of the individualized ovarian stimulation guide (InOvaSGuide) system. EHR: electronic health record; FSH: follicle-stimulating hormone; NOR: number of oocytes retrieved; OHSS: ovarian hyperstimulation syndrome.

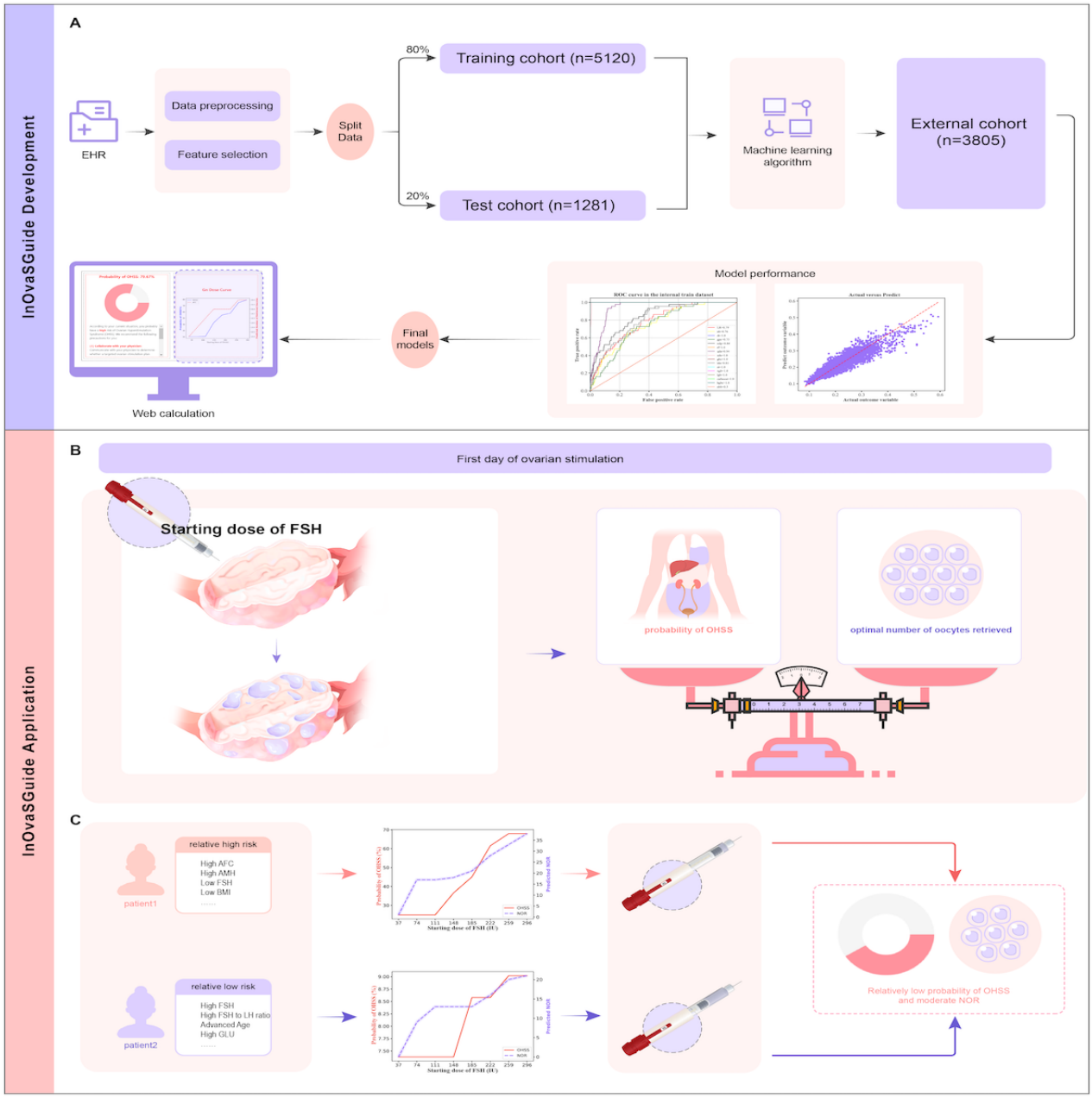

Patient Characteristics

A total of 6401 patients from the internal dataset and 3805 patients from the external dataset were included, with baseline characteristics detailed in Table 1. The median age was 30.0 (IQR 27.0-33.0) years in the internal dataset and 32.0 (IQR 29.0-35.0) years in the external dataset. The gonadotropin-releasing hormone antagonist protocol was the most commonly used in both datasets (internal dataset: 2650/6401, 41.4%; external dataset: 1913/3805, 50.3%). The median number of NOR was 13.0 (IQR 9.0-17.0) and 15.0 (IQR 10.0-20.0) for the internal and external dataset, respectively. In the internal dataset, 55 (0.9%) patients were diagnosed with moderate-to-severe OHSS, whereas 46 (1.2%) patients were diagnosed in the external dataset. Further comparisons between OHSS and non-OHSS cases in both datasets are detailed in Tables S4 and S5 in Multimedia Appendix 1.

XSL•FO
RenderX

Table 1. Baseline characteristics of the patients in the internal and external datasets.

	Internal dataset (n=6401)	External dataset (n=3805)	*P* value
Age (y), median (IQR)	30.0 (27.0-33.0)	32.0 (29.0-35.0)	<.001
BMI (kg/m^2), median (IQR)	21.7 (19.8-24.0)	21.6 (19.8-23.6)	.02
Baseline FSH[a] (mIU/mL), median (IQR)	6.2 (5.2-7.2)	6.5 (5.5-7.5)	<.001
Baseline luteinizing hormone (mIU/mL), median (IQR)	5.3 (3.8-7.1)	5.2 (3.7-6.9)	.006
Anti-Müllerian hormone (ng/mL), median (IQR)	4.1 (2.6-5.4)	3.8 (2.5-5.7)	.20
Fasting blood glucose (mmol/L), median (IQR)	5.3 (5.1-5.4)	4.5 (4.3-4.8)	<.001
Fasting insulin (μU/mL), median (IQR)	11.0 (7.5-12.1)	62.6 (62.6-62.6)	<.001
Homeostasis model assessment of insulin resistance, median (IQR)	2.5 (1.7-2.9)	13.3 (13.3-13.3)	<.001
Baseline antral follicle count, median (IQR)	20.0 (14.0-24.0)	13.0 (10.0-19.0)	<.001
Ovarian stimulation protocol, n (%)			<.001
GnRH[b] agonist long protocol	1211 (18.9)	0 (0)	
GnRH antagonist protocol	2650 (41.4)	1913 (50.3)	
Early-follicular phase long-acting GnRH agonist long protocol	2300 (35.9)	1827 (48)	
Ultralong GnRH agonist protocol	240 (3.8)	65 (1.7)	
Starting dose of FSH (IU), median (IQR)	150.0 (150.0-187.5)	225.0 (150.0-300.0)	<.001
Total dose of FSH (IU), median (IQR)	1950.0 (1500.0-2437.5)	2100.0 (1575.0-2750.0)	<.001
Estradiol level on the day of triggering (pg/mL), median (IQR)	3269.3 (2580.0-3269.3)	2750.0 (1804.0-3961.0)	<.001
Oocytes retrieved, median (IQR)	13.0 (9.0-17.0)	15.0 (10.0-20.0)	<.001
Degree of ovarian hyperstimulation syndrome, n (%)			.11
Normal	6346 (99.1)	3759 (98.8)	
Moderate to severe	55 (0.9)	46 (1.2)	

[a]FSH: follicle-stimulating hormone.

[b]GnRH: gonadotropin-releasing hormone.

Model Performance

For NOR prediction, the gradient boosting regressor exhibited the best performance, with an R^2 value of 0.7978 in the internal dataset and 0.7924 in the external dataset, indicating strong explanatory power (Table 2). The model's mean absolute error was 0.0223 and the root mean square error was 0.0298, collectively affirming the high accuracy and minimal bias. The model's predictions aligned closely with the actual outcomes, demonstrating relatively high accuracy. The Quantile-Quantile plot further confirmed that the residuals followed a normal distribution, as they closely aligned with the diagonal line (Figures 2A-2D).

XSL•FO
RenderX

Table 2. Performance metrics of the number of oocytes retrieved prediction models in internal and external datasets.

Machine learning models	Internal dataset				External dataset			
	R^2	Adjusted R^2	Mean absolute error	Root mean square error	R^2	Adjusted R^2	Mean absolute error	Root mean square error
Gradient boosting regressor	0.7978	0.7951	0.0223	0.0298	0.7924	0.7915	0.0243	0.0327
Light gradient boosting machine regressor	0.7908	0.7880	0.0228	0.0303	0.7826	0.7817	0.0243	0.0334
Extreme gradient boosting regressor	0.7889	0.7861	0.0228	0.0305	0.7907	0.7898	0.0236	0.0328
Random forest regressor	0.7849	0.7820	0.0229	0.0308	0.8020	0.8012	0.0229	0.0319
Ridge	0.7463	0.7429	0.0253	0.0334	0.3228	0.3198	0.0470	0.0590
Linear regression	0.7463	0.7428	0.0253	0.0334	0.3235	0.3204	0.0469	0.0590
Decision tree regressor	0.5452	0.5391	0.0330	0.0447	0.6108	0.6090	0.0318	0.0447
Support vector regression	0.5107	0.5041	0.0390	0.0464	0.5321	0.5300	0.0366	0.0490
Lasso	−0.0023	−0.0158	0.0519	0.0664	−0.1773	−0.1826	0.0648	0.0778
Elastic net	−0.0023	−0.0158	0.0519	0.0664	−0.1773	−0.1826	0.0648	0.0778
Multilayer perceptron regressor	−0.4608	−0.4805	0.0523	0.0802	−351.3296	−352.9112	0.6948	1.3453

XSL•FO
RenderX

Figure 2. Model performance in predicting the number of oocytes retrieved (NOR; A–D) and ovarian hyperstimulation syndrome (OHSS; E and F) in the internal and external datasets. ADA: adaptive boost classifier; BNB: Bernoulli Naive Bayes; CatBoost: categorical boosting classifier; DT: decision tree classifier; ET: extra trees classifier; GBC: gradient boosting classifier; GNB: Gaussian Naive Bayes; GPC: Gaussian process classifier; HGBC: histogram-based gradient boosting classifier; LDA: linear discriminant analysis; LGB: light gradient boosting machine classifier; LR: logistic regression; MLP: multilayer perceptron classifier; QDA: quadratic discriminant analysis; QQ: quantile-quantile; RF: random forest; ROC: receiver operating characteristic; XGB: extreme gradient boosting classifier.

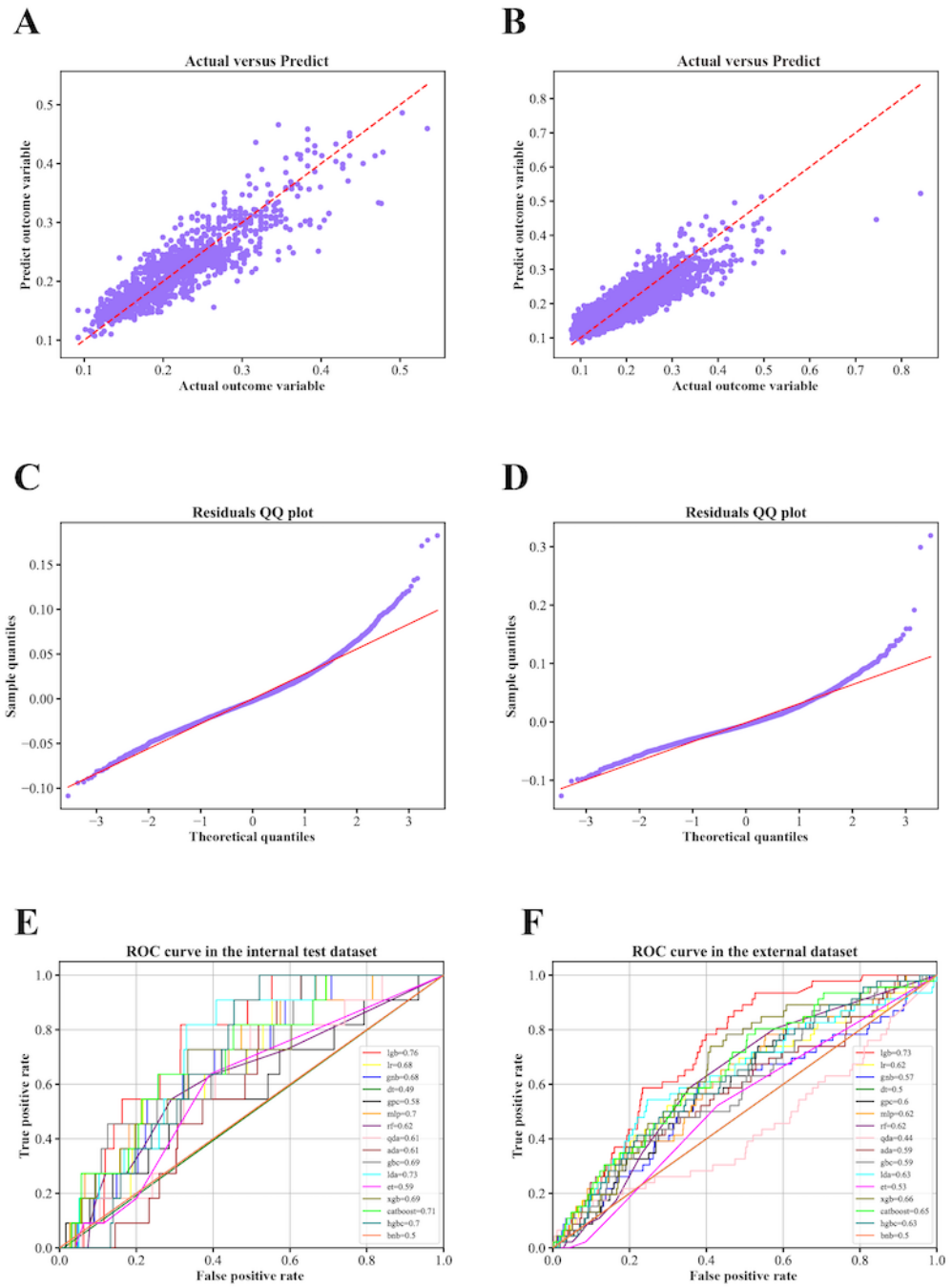

For early-onset moderate-to-severe OHSS prediction, the LightGBM model consistently outperformed other algorithms, achieving an ROC-AUC of 0.7588 in the internal dataset and 0.7287 in the external dataset (Figures 2E and 2F). While recall, specificity, weighted F_1-score, and Cohen κ score indicated reasonable discriminative performance, precision-related metrics, including positive predictive value, negative predictive value, and PR-AUC, remained modest across all classifiers. The results, along with the confusion matrices, are summarized in Table 3 and Figures S5 and S6 in Multimedia Appendix 1.

XSL•FO
RenderX

Table 3. Performance metrics of early-onset moderate-to-severe ovarian hyperstimulation syndrome prediction models in internal and external datasets.

Classifier	Internal dataset								External dataset							
	ROC-AUC[a]	Recall	Specificity	Weighted F_1-score	κ	PPV[b]	NPV[c]	Precision-recall AUC[d]	ROC-AUC	Recall	Specificity	Weighted F_1-score	κ	PPV	NPV	Precision-recall-AUC
LGBMClassifier[e]	0.7588	1.0000	0.9833	0.3466	0.5044	0.3409	1.0000	0.0176	0.7287	0.9348	0.9859	0.4523	0.6099	0.4464	0.9982	0.0227
LinearDiscriminantAnalysis	0.7313	1.0000	0.5690	0.0281	0.0384	0.0197	1.0000	0.0162	0.6574	0.6739	0.9560	0.5966	0.7362	0.5956	0.9934	0.0196
CatBoost[f]	0.7059	1.0000	0.9584	0.1795	0.2918	0.1724	1.0000	0.0173	0.6527	0.8043	0.9852	0.4045	0.5635	0.3996	0.9940	0.0206
MLPClassifier[g]	0.6966	0.9091	0.9784	0.2724	0.4177	0.2669	0.9971	0.0168	0.6322	0.8913	0.9996	0.2071	0.3275	0.1987	0.9893	0.0236
GradientBoostingClassifier	0.6930	1.0000	0.9610	0.1889	0.3053	0.1819	1.0000	0.0161	0.6235	0.2174	0.9553	0.8757	0.9227	0.8837	0.9947	0.0157
XGBClassifier[h]	0.6892	0.7273	0.9941	0.5199	0.6759	0.5181	0.9955	0.0152	0.6221	0.8043	0.9944	0.4263	0.5854	0.4217	0.9933	0.0189
GaussianNB[i]	0.6808	1.0000	0.9755	0.2678	0.4111	0.2614	1.0000	0.0135	0.6186	0.7609	0.9999	0.3548	0.5113	0.3498	0.9883	0.0276
LogisticRegression	0.6791	1.0000	0.4837	0.0250	0.0324	0.0165	1.0000	0.0138	0.6019	1.0000	0.9827	0.0121	0.0003	0.0000	0.9919	0.0182
RandomForest	0.6178	0.7273	0.9909	0.4122	0.5752	0.4094	0.9943	0.0123	0.5945	0.9130	0.9865	0.2084	0.3290	0.1998	0.9944	0.0156
QuadraticDiscriminantAnalysis	0.6142	0.7273	0.9943	0.5277	0.6826	0.5260	0.9955	0.0108	0.5899	1.0000	1.0000	0.0121	0.0003	0.0000	0.9884	0.0267
ExtraTreesClassifier	0.5937	0.6364	0.9965	0.6097	0.7496	0.6094	0.9949	0.0102	0.5280	0.5217	0.9952	0.5708	0.7162	0.5714	0.9899	0.0120

[a]ROC-AUC: area under the receiver operating characteristic curve.

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

[d]AUC: area under the curve.

[e]LGBMClassifier: light gradient boosting machine classifier.

[f]CatBoost: categorical boosting classifier.

[g]MLPClassifier: multilayer perceptron classifier.

[h]XGBClassifier: extreme gradient boosting classifier.

[i]NB: Naive Bayes.

Model Interpretation

SHAP values were used to assess feature importance for both models, as shown in Figure 3. For NOR prediction, the features with the highest mean absolute SHAP values were FSH starting dose to BMI ratio, BMI, log (AMH), baseline AFC, and baseline FSH, indicating their significant contribution to the model. For early-onset moderate-to-severe OHSS prediction, the most important features identified were baseline AFC, followed by baseline FSH, BMI, fasting blood glucose, and log (AMH).

XSL•FO
RenderX

Figure 3. Shapley additive explanation (SHAP) values of the prediction models for (A) number of oocytes retrieved (NOR) and (B) ovarian hyperstimulation syndrome (OHSS). FSH: follicle-stimulating hormone; BMI, body mass index; AMH: anti-Müllerian hormone; AFC: antral follicle count; LH: luteinizing hormone; HOMA-IR: homeostatic model assessment of insulin resistance; HDL: high-density lipoprotein; E2: estradiol; INS: fasting insulin; TG: triglycerides; TC: total cholesterol; LDL: low-density lipoprotein; GLU: fasting glucose; T: testosterone;.

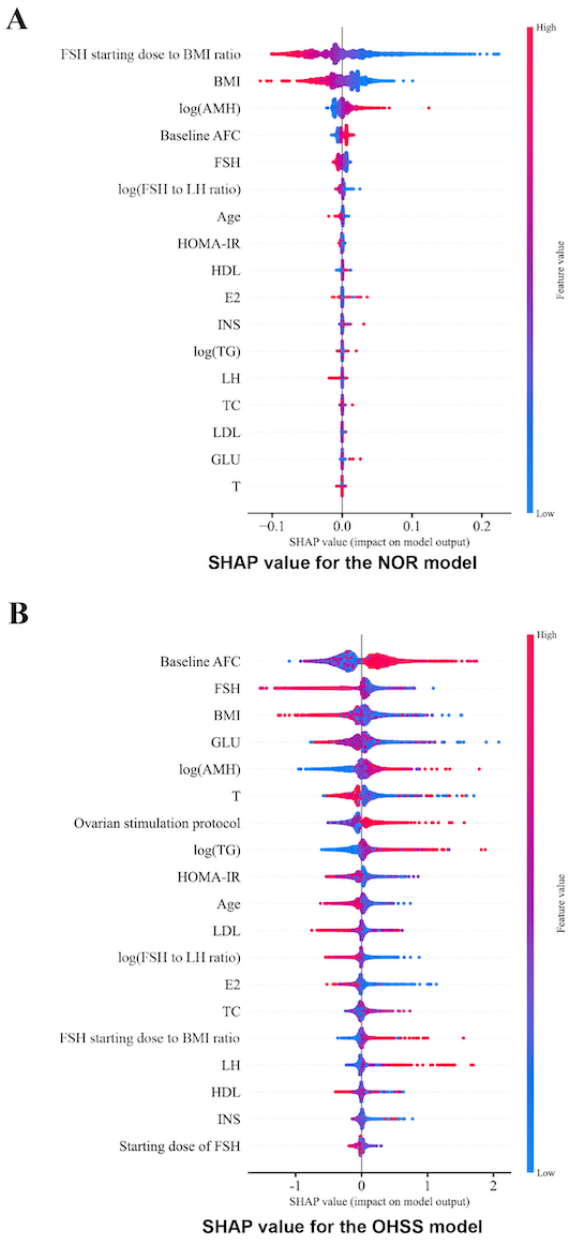

Integrated Dose-Response Curves and Web Calculator

To facilitate individualized ovarian stimulation, we further integrated the prediction models for both NOR and early-onset moderate-to-severe OHSS into dose-response curves. Examples of patients with relatively high and low predicted risks of OHSS are presented in Figures 4A and 4B, respectively. As shown, increasing the starting dose of FSH leads to variable increases in both early-onset moderate-to-severe OHSS probability and predicted NOR for different patients. However, the probability of OHSS occurrence varies among individuals. On the basis of these personalized dose-response predictions, clinicians can determine a suitable starting dose of FSH to achieve an optimal NOR while maintaining a relatively low risk of early-onset moderate-to-severe OHSS.

XSL•FO
RenderX

Figure 4. Model application and representative patient examples with predicted relatively high (A) and low (B) risks of ovarian hyperstimulation syndrome (OHSS). FSH: follicle-stimulating hormone; NOR: number of oocytes retrieved.

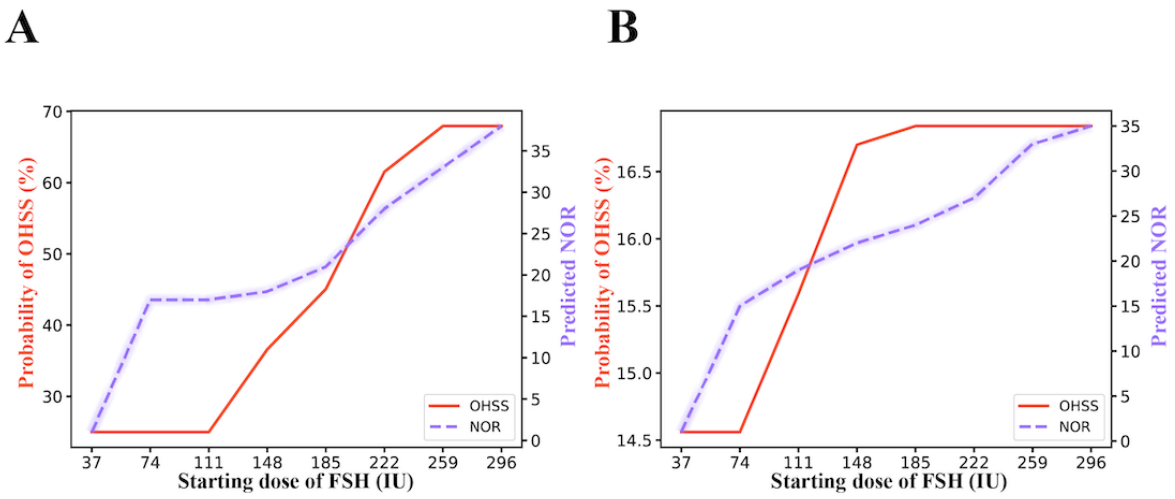

Additionally, we developed a web-based calculator as a research-oriented prototype to enhance accessibility and facilitate exploratory use of the proposed models. The intuitive interface allows users to input relevant data and receive immediate predictions from the models (Figure S7 in Multimedia Appendix 1).

Discussion

Principal Findings

In this study, we developed an integrated ML-based system, "InOvaSGuide," capable of simultaneously predicting NOR and the associated risk of early-onset moderate-to-severe OHSS across a wide range of FSH starting doses. By generating individualized dose-response curves, the model provides a continuous view of expected oocyte yield and corresponding safety profiles, providing clinicians with a structured visual reference to support individualized FSH dose selection. A web-based calculator was further implemented as a research-oriented prototype to improve accessibility and facilitate exploratory use of the models in clinically relevant scenarios.

Oocyte yield remains a key determinant of both efficacy and safety in assisted reproduction. Retrieval of fewer than 4 oocytes has been associated with poor reproductive prognosis [17], whereas obtaining more than 15 oocytes increase the likelihood of OHSS and may slightly compromise live birth outcomes [18,19]. Accordingly, a target range of 5 to 15 oocytes is generally recommended to balance benefit and risk [19]. The dose-response curve framework aligns conceptually with these clinical principles, as it illustrates how predicted NOR changes with increasing FSH doses, thereby supporting informed dosing discussions rather than prescriptive decision-making.

Existing ML-based NOR models generally focus either on approximating actual or optimal oocyte yield [11-13,20] or on producing individualized curves based on a limited number of clinical features [10]. In contrast, our approach used feature importance scores from automated classifiers for selection and compared 11 regression algorithms across 2 independent datasets. This allowed the construction of robust dose-response curves that illustrate how predicted NOR varies with incremental FSH doses, providing additional insight beyond traditional single-point estimates by visualizing predictions across a continuum of FSH doses.

We further developed ML models to predict OHSS, addressing a gap in existing clinical tools that predominantly rely on logistic regression [21,22] or receiver operating characteristic–based analyses [23,24]. Although contemporary strategies, including gonadotropin-releasing hormone antagonist protocols, dual triggering, and "freeze-all" approaches, have substantially reduced the incidence of early-onset OHSS, it remains a persistent concern even among presumed normal responders and has not been fully eliminated from clinical practice [25-27]. Our models demonstrated acceptable and consistent discriminatory ability across both internal and external cohorts, despite the limited number of OHSS events.

Importantly, the low prevalence of early-onset moderate-to-severe OHSS introduces substantial and unavoidable class imbalance, which has direct implications for model performance metrics. In particular, precision is structurally constrained in low-prevalence settings; therefore, PR-AUC values should be interpreted with caution. Although ROC-AUC indicated reasonable discrimination, PR-AUC is highly sensitive to outcome prevalence. When event rates fall below 1%, even well-calibrated models will inherently yield modest precision. In addition, our modeling strategy deliberately prioritized sensitivity to enhance clinical safety, an approach that increases false-positive predictions and further reduces precision and PR-AUC but minimizes the risk of missing true high-risk cases.

Within this context, the OHSS model should be viewed primarily as a screening and risk-stratification aid rather than a

XSL•FO
RenderX

diagnostic or decision-making tool. Its intended role is to flag patients with potentially elevated risk who may warrant closer monitoring or consideration of preventive strategies, rather than to definitively predict OHSS occurrence or guide autonomous clinical actions.

Feature importance analyses largely reflected established biological associations. For NOR, the FSH starting dose to BMI ratio, BMI, and log (AMH) emerged as the most influential predictors. While most features were consistent with clinical practice, the contribution of metabolic markers, such as glucose, lipids, and metabolic indicators, warrants further investigation. For early-onset moderate-to-severe OHSS, AFC, baseline FSH, BMI, and log (AMH) emerged as dominant predictors, aligning with known determinants of ovarian reserve and ovarian sensitivity [23,28]. Associations involving testosterone or glucose were less pronounced, highlighting the multifactorial nature of OHSS risk, particularly in women with polycystic ovary syndrome [29]. Overall, these findings illustrate the capability of ML approaches to integrate diverse clinical variables and improve predictive performance.

A key practical advantage of this study is the integration of NOR and early-onset moderate-to-severe OHSS predictions into a unified, visually intuitive research-oriented system. InOvaSGuide enables clinicians to assess the potential trade-offs between stimulation efficacy and safety across a continuum of FSH doses. The web-based interface supports exploratory analysis and clinician-patient discussion; however, the system does not generate prescriptive dosing recommendations and is not intended for autonomous clinical use. Importantly, prospective validation is essential before any consideration of clinical deployment.

Beyond model performance, the development and potential deployment of AI-based, clinician-in-the-loop decision support tools in reproductive medicine entail careful ethical, legal, and implementation considerations [30,31]. Given the sensitivity of reproductive health data, rigorous safeguards for privacy protection and informed consent are essential [32]. Algorithmic transparency is equally important to support clinician interpretation and reduce risks of automation bias [33], while potential bias across patient subgroups remains an important consideration for future validation. In addition, AI-driven clinical tools may fall under medical software regulation, requiring evidence of safety and clinical validity before clinical implementation. Finally, effective integration into clinical practice will depend on usability, compatibility with established workflows, and clearly assigned clinical accountability [30]. Addressing these factors will be necessary before the system can be responsibly adopted in real-world settings.

Limitations

Our study had several limitations. First, it focused on early-onset moderate-to-severe OHSS, excluding mild cases that may self-resolve and late-onset OHSS more commonly associated with embryo transfer. Patients with a predicted poor prognosis were also excluded under the assumption that they are less likely to develop OHSS. These exclusions introduced a structural selection bias that narrowed the population represented and limited the generalizability of the model in broader clinical settings. Second, the relatively small number of OHSS cases limited the model's ability to fully characterize patients who are affected. This scarcity, together with the substantial class imbalance, also constrained the effectiveness of resampling-based strategies. Although multiple resampling methods were evaluated, only cost-sensitive learning may allow a more reliable assessment of alternative methods. Third, the retrospective nature of the study restricted the availability of certain relevant factors, such as previous OHSS history and genetic susceptibility, and may also introduce selection bias and unmeasured confounding that cannot be fully controlled. Finally, although the system provides individualized dose-response curves for clinical reference, it does not generate a prescriptive starting dose. Moreover, the model has not yet undergone prospective evaluation, which limits its current clinical applicability. A prospective validation study is planned as a necessary next step to assess real-world performance. Future large-scale, multicenter validation in broader patient populations will be essential for improving model stability and generalizability.

Conclusions

We developed and externally validated InOvaSGuide, a ML system that simultaneously predicts NOR and early-onset moderate-to-severe OHSS risk across a continuum of FSH doses. By linking efficacy and safety within a single dose-response framework, the tool highlights the broader potential of model-informed dosing to standardize ovarian stimulation and enhance patient safety. Prospective trials are needed to establish real-world utility.

Acknowledgments

The authors are grateful to Yuan Sheng from STI-Zhilian Research Institute for Innovation and Digital Health for her invaluable contributions to the study figures. They also thank the Yiersan Digital Health Care Group for providing technical support for the web-based calculator interface design. Generative AI (ChatGPT-4.1, OpenAI [34]) was used exclusively for language refinement during manuscript preparation, including improving grammar, clarity, and readability of author-written text. No text, data interpretation, scientific content, or references were generated by the AI tool. All substantive content, analyses, and conclusions were produced entirely by the authors. The authors verified all AI-assisted edits and take full responsibility for the final manuscript.

XSL•FO
RenderX

Funding

This study was funded by the National Natural Science Foundation of China (grant 82371682) and the Natural Science Foundation of Hunan Province (grants 2022JJ40779 and 2022JJ70080). The funding source had no role in the study design, data collection, data analysis, data interpretation, or writing of the report.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JF, Y Li, and SL had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. JC played a role in data collection and was a major contributor to manuscript writing. Jianjuan Zhao conducted data analysis, performed modeling, and contributed to drafting the Methods section. JC, Jianjuan Zhao, JF, Y Li, and SL contributed equally to this work. YZ, GY, YY, and QS provided biostatistical analysis support. HQ, HT, Jing Zhao, and BX assisted in internal dataset acquisition, while Y Liu, JL, and ZY helped in external dataset acquisition. QZ and HL provided language support. JF, Y Li, and SL were responsible for the study design and critically revised the manuscript. All authors read and approved the final manuscript

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data supporting model development and validation, including the distribution and transformation of key variables, feature selection for number of oocytes retrieved and ovarian hyperstimulation syndrome prediction, confusion matrices, the web-based calculator interface, and detailed patient baseline characteristics, together with the completed TRIPOD+AI checklist.
[PDF File (Adobe PDF File), 2183 KB - jmir_v28i1e78245_app1.pdf]

References

1. Ngwenya O, Lensen SF, Vail A, Mol BW, Broekmans FJ, Wilkinson J. Individualised gonadotropin dose selection using markers of ovarian reserve for women undergoing in vitro fertilisation plus intracytoplasmic sperm injection (IVF/ICSI). Cochrane Database Syst Rev 2024 Jan 04;1(1):CD012693. [doi: 10.1002/14651858.CD012693.pub3] [Medline: 38174816]
2. Broekmans FJ. Individualization of FSH doses in assisted reproduction: facts and fiction. Front Endocrinol (Lausanne) 2019 Apr 26;10:181 [FREE Full text] [doi: 10.3389/fendo.2019.00181] [Medline: 31080437]
3. Kim HH, Speedy SE. The promised land of individualized ovarian stimulation: are we there yet? Fertil Steril 2021 Apr;115(4):893-894 [FREE Full text] [doi: 10.1016/j.fertnstert.2021.02.023] [Medline: 33832745]
4. Qiao J, Zhang Y, Liang X, Ho T, Huang HY, Kim SH, et al. A randomised controlled trial to clinically validate follitropin delta in its individualised dosing regimen for ovarian stimulation in Asian IVF/ICSI patients. Hum Reprod 2021 Aug 18;36(9):2452-2462 [FREE Full text] [doi: 10.1093/humrep/deab155] [Medline: 34179971]
5. Broekmans FJ, Kwee J, Hendriks DJ, Mol BW, Lambalk CB. A systematic review of tests predicting ovarian reserve and IVF outcome. Hum Reprod Update 2006;12(6):685-718. [doi: 10.1093/humupd/dml034] [Medline: 16891297]
6. Lensen SF, Wilkinson J, Leijdekkers JA, La Marca A, Mol BW, Marjoribanks J, et al. Individualised gonadotropin dose selection using markers of ovarian reserve for women undergoing in vitro fertilisation plus intracytoplasmic sperm injection (IVF/ICSI). Cochrane Database Syst Rev 2018 Feb 01;2(2):CD012693 [FREE Full text] [doi: 10.1002/14651858.CD012693.pub2] [Medline: 29388198]
7. Hicks SA, Andersen JM, Witczak O, Thambawita V, Halvorsen P, Hammer HL, et al. Machine learning-based analysis of sperm videos and participant data for male fertility prediction. Sci Rep 2019 Nov 14;9(1):16770 [FREE Full text] [doi: 10.1038/s41598-019-53217-y] [Medline: 31727961]
8. Zaninovic N, Rosenwaks Z. Artificial intelligence in human in vitro fertilization and embryology. Fertil Steril 2020 Nov;114(5):914-920 [FREE Full text] [doi: 10.1016/j.fertnstert.2020.09.157] [Medline: 33160513]
9. Hanassab S, Abbara A, Yeung AC, Voliotis M, Tsaneva-Atanasova K, Kelsey TW, et al. The prospect of artificial intelligence to personalize assisted reproductive technology. NPJ Digit Med 2024 Mar 01;7(1):55 [FREE Full text] [doi: 10.1038/s41746-024-01006-x] [Medline: 38429464]
10. Fanton M, Nutting V, Rothman A, Maeder-York P, Hariton E, Barash O, et al. An interpretable machine learning model for individualized gonadotrophin starting dose selection during ovarian stimulation. Reprod Biomed Online 2022 Dec;45(6):1152-1159 [FREE Full text] [doi: 10.1016/j.rbmo.2022.07.010] [Medline: 36096871]
11. Ferrand T, Boulant J, He C, Chambost J, Jacques C, Pena CA, et al. Predicting the number of oocytes retrieved from controlled ovarian hyperstimulation with machine learning. Hum Reprod 2023 Oct 03;38(10):1918-1926 [FREE Full text] [doi: 10.1093/humrep/dead163] [Medline: 37581894]

XSL•FO
RenderX

12. Correa N, Cerquides J, Arcos JL, Vassena R. Supporting first FSH dosage for ovarian stimulation with machine learning. Reprod Biomed Online 2022 Nov;45(5):1039-1045. [doi: 10.1016/j.rbmo.2022.06.010] [Medline: 35915001]
13. Xu H, Feng G, Han Y, La Marca A, Li R, Qiao J. POvaStim: an online tool for directing individualized FSH doses in ovarian stimulation. Innovation (Camb) 2023 Mar 13;4(2):100401 [FREE Full text] [doi: 10.1016/j.xinn.2023.100401] [Medline: 36926531]
14. Debray TP, Collins GS, Riley RD, Snell KI, Van Calster B, Reitsma JB, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist. BMJ 2023 Feb 07;380:e071018 [FREE Full text] [doi: 10.1136/bmj-2022-071018] [Medline: 36750242]
15. Practice Committee of the American Society for Reproductive Medicine. Testing and interpreting measures of ovarian reserve: a committee opinion. Fertil Steril 2012 Dec;98(6):1407-1415 [FREE Full text] [doi: 10.1016/j.fertnstert.2012.09.036] [Medline: 23095141]
16. Practice Committee of the American Society for Reproductive Medicine. Prevention and treatment of moderate and severe ovarian hyperstimulation syndrome: a guideline. Fertil Steril 2016 Dec;106(7):1634-1647 [FREE Full text] [doi: 10.1016/j.fertnstert.2016.08.048] [Medline: 27678032]
17. Oudendijk JF, Yarde F, Eijkemans MJ, Broekmans FJ, Broer SL. The poor responder in IVF: is the prognosis always poor?: a systematic review. Hum Reprod Update 2012;18(1):1-11. [doi: 10.1093/humupd/dmr037] [Medline: 21987525]
18. Steward RG, Lan L, Shah AA, Yeh JS, Price TM, Goldfarb JM, et al. Oocyte number as a predictor for ovarian hyperstimulation syndrome and live birth: an analysis of 256,381 in vitro fertilization cycles. Fertil Steril 2014 Apr;101(4):967-973 [FREE Full text] [doi: 10.1016/j.fertnstert.2013.12.026] [Medline: 24462057]
19. Sunkara SK, Rittenberg V, Raine-Fenning N, Bhattacharya S, Zamora J, Coomarasamy A. Association between the number of eggs and live birth in IVF treatment: an analysis of 400 135 treatment cycles. Hum Reprod 2011 Jul;26(7):1768-1774. [doi: 10.1093/humrep/der106] [Medline: 21558332]
20. Zieliński K, Pukszta S, Mickiewicz M, Kotlarz M, Wygocki P, Zieleń M, et al. Personalized prediction of the secondary oocytes number after ovarian stimulation: a machine learning model based on clinical and genetic data. PLoS Comput Biol 2023 Apr 27;19(4):e1011020 [FREE Full text] [doi: 10.1371/journal.pcbi.1011020] [Medline: 37104276]
21. Grynnerup AG, Løssl K, Toftager M, Bogstad JW, Prætorius L, Zedeler A, et al. Predictive performance of peritoneal fluid in the pouch of Douglas measured five days after oocyte pick-up in predicting severe late-onset OHSS: a secondary analysis of a randomized trial. Eur J Obstet Gynecol Reprod Biol 2022 Jul;274:83-87 [FREE Full text] [doi: 10.1016/j.ejogrb.2022.05.004] [Medline: 35609351]
22. Tarlatzi TB, Venetis CA, Devreker F, Englert Y, Delbaere A. What is the best predictor of severe ovarian hyperstimulation syndrome in IVF? A cohort study. J Assist Reprod Genet 2017 Oct 14;34(10):1341-1351 [FREE Full text] [doi: 10.1007/s10815-017-0990-7] [Medline: 28710674]
23. Ocal P, Sahmay S, Cetin M, Irez T, Guralp O, Cepni I. Serum anti-Müllerian hormone and antral follicle count as predictive markers of OHSS in ART cycles. J Assist Reprod Genet 2011 Dec 1;28(12):1197-1203 [FREE Full text] [doi: 10.1007/s10815-011-9627-4] [Medline: 21882017]
24. Cao M, Lin Q, Liu Z, Lin Y, Huang Q, Fu Y, et al. Optimized personalized management approach for moderate/severe OHSS: development and prospective validation of an OHSS risk assessment index. Hum Reprod 2024 Oct 01;39(10):2320-2330. [doi: 10.1093/humrep/deae197] [Medline: 39237109]
25. Emile SH, Horesh N, Garoufalia Z, Gefen R, Ray-Offor E, Wexner SD. Strategies to reduce ileus after colorectal surgery: a qualitative umbrella review of the collective evidence. Surgery 2024 Feb;175(2):280-288. [doi: 10.1016/j.surg.2023.10.005] [Medline: 38042712]
26. Ioannidou PG, Bosdou JK, Lainas GT, Lainas TG, Grimbizis GF, Kolibianakis EM. How frequent is severe ovarian hyperstimulation syndrome after GnRH agonist triggering in high-risk women? A systematic review and meta-analysis. Reprod Biomed Online 2021 Mar;42(3):635-650. [doi: 10.1016/j.rbmo.2020.11.008] [Medline: 33483281]
27. The Eshre Guideline Group On Ovarian Stimulation, Bosch E, Broer S, Griesinger G, Grynberg M, Humaidan P, et al. ESHRE guideline: ovarian stimulation for IVF/ICSI. Hum Reprod Open 2020;2020(2):hoaa009 [FREE Full text] [doi: 10.1093/hropen/hoaa009] [Medline: 32395637]
28. Ashrafi M, Bahmanabadi A, Akhond MR, Arabipoor A. Predictive factors of early moderate/severe ovarian hyperstimulation syndrome in non-polycystic ovarian syndrome patients: a statistical model. Arch Gynecol Obstet 2015 Nov 29;292(5):1145-1152. [doi: 10.1007/s00404-015-3723-0] [Medline: 25920524]
29. Sun B, Ma Y, Li L, Hu L, Wang F, Zhang Y, et al. Factors associated with ovarian hyperstimulation syndrome (OHSS) severity in women with polycystic ovary syndrome undergoing IVF/ICSI. Front Endocrinol (Lausanne) 2020 Jan 19;11:615957 [FREE Full text] [doi: 10.3389/fendo.2020.615957] [Medline: 33542709]
30. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan 7;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
31. Morley J, Machado CC, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: a mapping review. Soc Sci Med 2020 Sep;260:113172. [doi: 10.1016/j.socscimed.2020.113172] [Medline: 32702587]
32. Price WN2, Cohen IG. Privacy in the age of medical big data. Nat Med 2019 Jan 7;25(1):37-43 [FREE Full text] [doi: 10.1038/s41591-018-0272-7] [Medline: 30617331]

XSL•FO
RenderX

33. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018 Mar 15;378(11):981-983 [FREE Full text] [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
34. OpenAI. ChatGPT. URL: https://chatgpt.com/

Abbreviations

AFC: antral follicle count
AI: artificial intelligence
AMH: anti-Müllerian hormone
ROC-AUC: area under the receiver operating characteristic curve
FSH: follicle-stimulating hormone
InOvaSGuide: individualized ovarian stimulation guide
IVF: in vitro fertilization
LH: luteinizing hormone
NOR: number of oocytes retrieved
OHSS: ovarian hyperstimulation syndrome
PR-AUC: precision-recall area under the curve
SHAP: Shapley additive explanation
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Edited by J Sarvestan; submitted 30.May.2025; peer-reviewed by F Sun, KH Lin; comments to author 29.Sep.2025; revised version received 22.Dec.2025; accepted 23.Dec.2025; published 03.Feb.2026.

Please cite as:
Chen J, Zhao J, Qiu H, Liu Y, Zhang Y, Sun Q, Yi Y, Tang H, Zhao J, Xu B, Zhang Q, Yang G, Li H, Liu J, Yang Z, Liang S, Li Y, Fu J
Integrated Prediction System for Individualized Ovarian Stimulation and Ovarian Hyperstimulation Syndrome Prevention: Algorithm Development and Validation
J Med Internet Res 2026;28:e78245
URL: https://www.jmir.org/2026/1/e78245
doi:10.2196/78245
PMID:

©Jingjing Chen, Jianjuan Zhao, Huiyu Qiu, Yanhui Liu, Yunqi Zhang, Qicheng Sun, Yan Yi, Hongying Tang, Jing Zhao, Bin Xu, Qiong Zhang, Ge Yang, Hui Li, Junjie Liu, Zhongzhou Yang, Shaolin Liang, Yanping Li, Jing Fu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patterns of AI Use in Clinical Work by Hospitalists: Survey Study

Prabhava Bagla[1,2], MD; Jasmah Hanna[1,2], MS; Bhargav Marthambadi[1,2], MD; Stacey Watkins[1,2], MD, PhD

[1]Department of Medicine, Division of Hospital Medicine, School of Medicine, Emory University, Atlanta, GA, United States
[2]Grady Memorial Hospital, Atlanta, GA, United States

Corresponding Author:
Prabhava Bagla, MD
Department of Medicine, Division of Hospital Medicine
School of Medicine
Emory University
80 Jesse Hill Jr Dr SE
Atlanta, GA, 30303
United States
Phone: 1 (404) 251 8899
Fax: 1 (404) 727 2620
Email: prabhava.bagla@emory.edu

Abstract

Background: Artificial intelligence (AI) tools are widely and freely available for clinical use. Understanding hospitalists' real-world adoption patterns in the absence of organizational endorsement is essential for health care institutions to develop governance frameworks and optimize AI integration.

Objective: The objective of this study was to investigate hospitalists' use of AI, examining the AI platforms being used, frequency of use, and clinical contexts of application. We hypothesized that AI use is more common among younger, less experienced hospitalists, albeit at an overall low frequency.

Methods: An anonymous online survey was distributed via email to all 70 hospitalists (physicians, nurse practitioners, and physician assistants) providing direct patient care at a large urban academic tertiary care hospital. Demographic data, the AI platform used (if any), the purpose for AI use, and the frequency of use information were collected. The CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist was used for creating, testing, administering, and reporting the results of the survey. Chi-square test was used where possible; when expected cell values were low, the Fisher's exact test was used instead. The Friedman test and the pairwise Wilcoxon signed-rank test were used for analyzing the differences in the frequency of AI use for various tasks. Likert-scale responses to frequency questions (never, rarely, sometimes, often, and always) were converted to ordinal values (1-5, respectively) to facilitate analysis.

Results: Of the 70 providers, 54 (77.1%) responded to the survey. No significant differences in AI usage were observed across shift type, years of practice, time allocation to hospitalist duties, sex, age, or provider designation, contrary to our hypothesis. Overall, 36 of 54 respondents (66.7%; 95% CI 53.4%-77.8%) reported using AI in clinical practice. OpenEvidence was the most used platform (28/54, 51.9%), far exceeding general-purpose tools like OpenAI's ChatGPT (4/54, 7.4%), suggesting a preference for medical-specific platforms. Among nonusers, primary concerns were AI accuracy and preference for established resources. The most common application was answering miscellaneous clinical questions (32/36, 88.9%), generating differential diagnoses (31/36, 86.1%), and determining management options (31/36, 86.1%), with much lower use for patient education materials (16/36, 44.4%). There was a statistically significant difference in the frequency of AI use across these clinical scenarios (Friedman test χ^2_4 37.6; P<.001). Pairwise comparisons using the Wilcoxon signed-rank test revealed significant differences between use for answering miscellaneous questions and confirming suspected diagnosis (P=.003) and generating patient education materials (P=.004), respectively. Most respondents reported using AI for under 25% of clinical encounters across all use cases.

Conclusions: Two-thirds of hospitalists organically adopted AI despite the absence of institutional oversight. AI is predominantly used as a supplementary decision support tool, with a preference for a medical-specific platform. Health care institutions must develop governance frameworks, validation protocols, and educational initiatives to ensure safe and effective AI deployment in clinical practice.

(J Med Internet Res 2026;28:e85973) doi:10.2196/85973

XSL•FO
RenderX

KEYWORDS

artificial intelligence; AI; generative artificial intelligence; generative AI; hospitalist; physician; patterns of use; clinical decision support systems; healthcare technology adoption

Introduction

The development of artificial intelligence (AI) models has led to a significant interest in exploring their capabilities in many aspects of clinical medicine [1,2]. AI systems have been tested against physicians in assessing acuity [3], developing diagnostic formulation [4], and generating management options [5]. Particularly in image or pattern recognition-based diagnosis, AI demonstrates comparable or superior performance to clinicians [6,7].

While numerous studies have evaluated AI performance in carefully controlled research settings, large language model (LLM)–based AI tools are also widely and freely available without restriction to specific, well-defined tasks. An LLM is pretrained on a large amount of data, and can leverage that knowledge to answer user queries in natural language format. LLM capabilities range from presenting relevant information sourced from data it has access to, generating new content, such as computer code, or editing text [8]. ChatGPT is a general-purpose AI system offered by OpenAI that has a range of free and premium tier options, with progressively more advanced reasoning and content generation capabilities available [9]. A survey of US adults by the Pew Research Center indicated that between March 2023 and March 2025, awareness about ChatGPT has increased from 58% to 79%, and 28% of employed adults reported using it for work, compared with 8% in 2023. Usage rates are higher among younger adults and those with higher education levels, such as a postgraduate degree [10]. Some AI systems, such as Gemini by Google and Copilot by Microsoft, are presented to consumers by means of integration into their respective technical offerings, including prominently alongside the basic web search interface [11,12]; with free and paid user tiers and capabilities available similar to ChatGPT [13,14].

Some AI platforms use a retrieval-augmented generation (RAG) system for ease of updating and improved reliability of their LLM's output. An RAG system provides a specific, reliable database for an LLM to search, in addition to what it has already been trained on. The LLM then combines its pre-existing knowledge and the information from the database to synthesize responses to user queries. The RAG strategy simplifies the process of keeping LLM answers updated and accurate in the context of evolving knowledge by simply expanding the authoritative database the LLM accesses, instead of having to retrain the LLM to search and appropriately parse all possible data sources on its own [15]. ClinicalKey AI is one such AI system developed by Elsevier requiring access using a paid subscription model that delivers evidence-based answers sourced from their library of books and journals [16].

OpenEvidence is a unique AI tool marketed specifically toward medical professionals only in the United States. It provides free unlimited use for verified users, while allowing some use with query limits to users without creating an account. It also uses the RAG system by means of data use agreements with several reputable sources, such as the *New England Journal of Medicine*, *Journal of the American Medical Association Network* journals, National Comprehensive Cancer Network, American Diabetes Association, and American College of Cardiology, and thus has the potential to offer evidence-based answers grounded in peer-reviewed medical literature. It also offers the ability to write patient handouts for education about management options for diseases, posthospitalization care, and insurance prior authorization letters [17].

AI tools have been implemented for use for a wide variety of purposes in health care, such as (1) clinical use by health care providers for supporting patient care decisions; (2) directly by laypersons for self-evaluation of health concerns; (3) facilitating business operations, such as bed management; and (4) as hybrid tools supporting both patient care and optimizing health care delivery, such as AI scribes for clinical documentation [18]. A study of 67 nonprofit health systems revealed adoption of AI for active or planned use mostly in imaging and radiology, followed by electronic health record integration for purposes like sepsis detection, and generating documentation via ambient listening [19].

An individual health care provider's use of AI might be influenced by how others perceive them. A study on public perception of physicians who use AI, conducted by surveying US adults shown hypothetical static advertisements for family physicians differing in whether AI use for administrative, diagnostic, or therapeutic purposes was mentioned, showed that physicians using AI for any purpose were considered less competent, trustworthy, and empathic, as well as less sought after for care [20]. Similarly, practicing clinicians viewed hypothetical peers who used AI as a primary decision-making tool to be significantly less competent than those who did not use AI at all, an effect only partially mitigated if the use of AI was specified as being used for verification purposes [21].

Despite these perception issues, physicians' sentiment toward AI use in clinical practice is overall positive, with greater familiarity, use in daily or professional life, and being involved in AI research as factors associated with greater enthusiasm and lower skepticism; the primary concerns being liability in the event of errors, and lack of transparency in AI decision-making [22]. The American Medical Association (AMA) conducted a survey of 1183 physicians, indicating that AI use increased from 38% in 2023 to 66% in 2024, mostly for assistance with documentation, ranging from discharge summaries to responses to patient portal messages. Respondents also indicated that measures to reduce administrative burden held the greatest promise for future AI implementation, and an increasing recognition of the potential of AI to benefit patient care (68% in 2024 compared with 63% in 2023) [23].

While adoption of AI at an organizational level leads to availability and use for preapproved tasks by health care providers working in the respective organizations, individual

XSL•FO
RenderX

providers are unable to use AI for any tasks that require deep integration with their organization's electronic health record system at a system level. Hospitalists represent a unique population for studying AI adoption given their role as frontline providers, managing high volumes of diverse, acutely ill patients across multiple clinical domains. Their practice patterns, workflow demands, and decision-making processes may influence how AI tools are integrated into daily clinical care. However, there is a dearth of literature about AI real-world usage patterns in daily clinical practice among hospitalists in the absence of any organizational endorsement.

The purpose of this study was to investigate hospitalists' use of AI through a comprehensive survey examining the types of AI tools currently being used, frequency of use, and clinical contexts of application in real-world hospital practice settings without any system-wide AI integration. We hypothesized that AI use in clinical work is more common among younger, less experienced hospitalists, albeit at an overall low frequency.

Methods

Study Design

A comprehensive anonymous online survey (Multimedia Appendix 1) was distributed via email during August 2025 to all 70 physicians, nurse practitioners, and physician assistants (the latter 2 categories collectively referred to as advanced practice providers [APPs]) who provide direct patient care at our large, tertiary care academic center. The sample was not considered a convenience sample as all providers were given the opportunity to respond. The survey asked demographic questions and questions about which AI systems, if any, the respondents used and for what purpose. For analytic purposes, AI platforms were categorized as either general-purpose or medical-specific based on their primary market positioning and intended use at the time of the survey. Medical-specific platforms were defined as tools marketed explicitly for clinical decision support and designed primarily for use by health care professionals, with content grounded in medical literature or curated clinical sources (OpenEvidence and ClinicalKey AI). General-purpose platforms were defined as broadly available AI tools intended for use across multiple domains and audiences, not marketed primarily for clinical decision-making (ChatGPT, Microsoft Copilot, and Google Gemini), even if health care–specific enterprise or specialized versions may exist.

If respondents indicated that they did not use any AI systems, they were also asked what concerns they had. CHERRIES (Checklist for Reporting Results of Internet E-Surveys) guidelines (Multimedia Appendix 2) on survey studies were followed [24].

Ethical Considerations

This study was considered not to comprise human participant research by Emory University Institutional Review Board in accordance with the Declaration of Helsinki, and so was exempt from full committee review. The first page of the survey contained information regarding the purpose of the study, potential risks, and benefits to respondents, and contact information regarding the investigators. If respondents did not wish to answer the survey, they could either decline using the option provided or simply close the survey. The page also disclosed that response data would be kept confidential, only analyzed as an aggregate, and the information stored securely without any specific time limits using an encrypted institutional subscription to Microsoft OneDrive and a related set of tools. No personally identifiable information was collected as part of the survey. No compensation or incentive of any kind, monetary or otherwise, was offered to participants.

Survey Development

Survey questions were developed by discussion and consensus among study authors. For AI users, the survey asked, "How often, on average, do you use AI per day for [specific clinical task]?" for each of five tasks, (1) generating differential diagnosis lists, (2) confirming suspected diagnosis, (3) answering miscellaneous questions (with examples provided, such as rare drug side effects), (4) determining management or testing options, and (5) generating patient education materials. Response options were never, rarely, sometimes, often, and always; defined as (1) rarely (0%-25% of the time), (2) sometimes (26%-50% of the time), (3) often (51%-75% of the time), and (4) always (76% or more of the time). The survey did not explicitly specify whether respondents should consider all patient encounters or only those encounters where the specific task would be applicable, leaving interpretation to individual respondents. Functionality was tested before dissemination to potential respondents.

Recruitment Process

The survey was distributed via email during August 2025 to all 70 potential respondents. They were subsequently reminded over a period of 3 weeks via group emails and in-person interactions.

Survey Administration

The email sent to potential respondents contained a brief description of the study and a link to the survey itself. The survey was created in Microsoft Forms, which allowed responses to be automatically saved as a Microsoft Excel spreadsheet. Adaptive questioning was used to streamline survey responses and navigation; for instance, if respondents indicated they did not use AI for clinical work at all, they were not shown questions about possible use scenarios. The maximum number of question items (including initial informed consent) was 17, with an average of 2 questions per page for readability. All questions were marked mandatory, so incomplete responses were not allowed. All questions, except for reasons for not using AI at all, were single-response only, and nonresponse options were provided for some of the demographic questions. Respondents could adjust their responses before submitting, at which time further editing of responses was not allowed. No incentives were offered in exchange for responses. Overall, responses were collected during August-September 2025.

Response Rates

Unique responses were ensured by requiring respondents to log in (automatically fulfilled when accessing the survey by clicking the link in their email) and only allowing 1 response per person. We are unable to determine if there were potential respondents

XSL•FO
RenderX

who viewed the survey but did not answer any questions or stopped partway, as only completed surveys were accepted.

Statistical Analysis

Statistical analysis was conducted using SAS (version 9.4, SAS Institute Inc) for analysis of demographic variables. Chi-square test was used where possible; when more than 20% of expected cell values were under 5, the Fisher's exact test was used instead. R software (version 4.5.1; R Foundation) was used for the Friedman test and the pairwise Wilcoxon signed-rank test for analyzing the differences between the frequency of AI use for various tasks. These analyses were restricted to responses to survey questions asking about the 5 standardized clinical tasks (generating differential diagnosis lists, confirming suspected diagnosis, answering miscellaneous questions, determining management or testing options, and generating patient education materials; Multimedia Appendix 1). Likert-scale responses to survey questions (never, rarely, sometimes, often, and always) were converted to ordinal values (1-5, respectively) to facilitate analysis. Only completed surveys were analyzed. No specific analysis regarding the time taken to complete the survey was conducted, and no weighting or propensity scores were used.

Results

Out of 70 providers, 54 responded to the survey (response rate of 77.1%). Table 1 presents survey respondent characteristics and AI usage patterns. The majority of respondents were dayshift team members (48/54, 88.9%), and two-thirds (36/54, 66.7%) were within 6 years of completing training. Most respondents devoted all their time to clinical work (31/54, 57.4%). Physicians comprised 83.3% (45/54) of respondents, with APPs accounting for the remaining 16.7% (9/54). No statistically significant differences in AI usage were observed across shift type, years of practice, time allocation to hospitalist duties, sex, age, or provider designation, contrary to our hypothesis that younger, less experienced hospitalists would use AI more frequently. Ethnicity was not analyzed for statistical significance due to insufficient sample sizes across groups. Overall, 36 of 54 (66.7%; 95% CI 53.4%-77.8%) respondents reported using AI tools in clinical practice.

XSL•FO
RenderX

Table 1. Characteristics of survey respondents and association with artificial intelligence use in clinical work (n=54). Survey respondents from a large academic medical center, August-September 2025.

Variable	Number (n=54), n (%)	Use of AI[a], n (%)	*P* value[b]
Role			>.99[b]
Dayshift rounding and admissions	48 (88.9)	32 (66.6)	
Night shift admissions and cross-cover	6 (11.1)	4 (66.6)	
Years of practice (y)			.54
0-6	36 (66.7)	25 (69.4)	
7 or more	18 (33.3)	11 (61.1)	
Amount of time spent on clinical hospitalist duties			.12
100%	31 (57.4)	18 (50.6)	
99% or less	23 (42.6)	18 (78.3)	
Sex			.44
Female	26 (48.1)	16 (61.5)	
Male	28 (51.9)	20 (71.4)	
Age (y)			.11
35 or less	20 (37)	16 (80)	
36 or more	34 (63)	20 (58.8)	
Ethnicity[c]			—[d]
Asian	15 (27.8)	10 (66.6)	
Black or African American	12 (22.2)	8 (66.6)	
Hispanic or Latino	3 (5.6)	3 (100)	
White	20 (37)	12 (60)	
Prefer not to say	4 (7.4)	3 (75)	
Designation			>.99[b]
Physician	45 (83.3)	30 (66.6)	
APP[e]	9 (16.7)	6 (66.6)	
Use of AI in clinical work			—
Yes	36 (66.7)	—	
No	18 (33.3)	—	

[a]AI: artificial intelligence.

[b]Fisher's exact test was used when specified, otherwise chi-square test was used.

[c]Statistical analysis not performed due to low sample sizes.

[d]Not applicable.

[e]APP: advanced practice provider.

OpenEvidence was the most commonly used platform (28/54, 51.9%), followed by ChatGPT (4/54, 7.4%; Table 2; Multimedia Appendix 3). Among the 18 respondents (33.3%) who did not use AI, the barriers to adoption were (1) preference for alternative resources such as PubMed (14/18, 77.8%); (2) concerns regarding accuracy (7/18, 38.9%); (3) not available system-wide, or endorsed by hospital, or via the electronic medical record (7/18, 38.9%; respondents who answered not available system-wide or endorsed by hospital also selected not available via electronic medical record); (4) time constraints (4/18, 22.2%); (5) privacy concerns (4/18, 22.2%); (6) prefer to consult another provider (2/18, 11.1%); and (7) lack of continuing medical education credits (1/18, 5.6%); no respondents selected cost or weekly query limits as reasons (percentages do not add up to 100% as multiple responses were allowed for this question). The question also allowed for free-text input of other reasons that respondents did not use AI, but no additional responses were received.

XSL•FO
RenderX

Table 2. Artificial intelligence platforms used by hospitalists in clinical work (n=54). Survey respondents from a large academic medical center, August-September 2025.

Use of AI[a] in clinical work	Number (n=54), n (%)
Microsoft Copilot	1 (1.9)
Google Gemini	1 (1.9)
ClinicalKey AI	2 (3.7)
ChatGPT	4 (7.4)
OpenEvidence	28 (51.9)
No use of any AI	18 (33.3)

[a]AI: artificial intelligence.

Table 3 and Multimedia Appendix 4 summarize the clinical tasks for which AI was used by users (n=36). The most common application was answering miscellaneous clinical questions (32/36, 88.9%; 95% CI 74.7%-95.6%), with respondents providing examples, including “rare drug side effects” and “causes of false-positive urine drug screen.” Generating differential diagnoses and determining management or treatment options were equally prevalent (31/36, 86.1%; 95% CI 71.3%-93.9%). Approximately 72.2% (26/36; 95% CI 56%-84.2%) of responders used AI to support confirmation of suspected diagnoses, while few (16/36, 44.4%; 95% CI 29.5%-60.4%) used AI for generating patient education or counseling materials. Using the Friedman test revealed there was a statistically significant difference in the frequency of AI use across these clinical scenarios (χ^2_4=37.6; P<.001).

Table 3. Clinical tasks for which artificial intelligence is used among hospitalist AI users (n=36). Survey respondents from a large academic medical center, August-September 2025.

Tasks AI[a] is used for	Number (n=36), n (%)	95% CI
Generating differential diagnosis list	31 (86.1)	71.3%-93.9%
Confirming suspected diagnosis	26 (72.2)	56%-84.2%
Answering miscellaneous questions	32 (88.9)	74.7%-95.6%
Determining testing or treatment options	31 (86.1)	71.3%-93.9%
Generating patient education or counseling materials	16 (44.4)	29.5%-60.4%

[a]AI: artificial intelligence.

Pairwise Wilcoxon signed-rank test with continuity and Bonferroni corrections demonstrated that AI use for answering miscellaneous questions was significantly different from use for confirming suspected diagnosis (corrected P=.003) and generating patient education or counseling materials (corrected P=.004).

Table 4 details the frequency of AI usage for each clinical application. Respondents were asked how often, on average, they used AI per day for each specific task. Most respondents reported using AI rarely (0%-25% of the time) across all use cases, in line with our hypothesis. Additionally, 3 respondents indicated other AI applications—assistance with clinical documentation, comprehensive literature searches, and creation of educational cases for medical students (1 respondent for each).

XSL•FO
RenderX

Table 4. Frequency of artificial intelligence use for specific clinical tasks among hospitalist artificial intelligence users (n=36). Survey respondents from a large academic medical center, August-September 2025.

Tasks[a] AI[b] is used for	Number (n=36), n (%)
Generating differential diagnosis list	
Never	5 (13.9)
Rarely	20 (55.6)
Sometimes	8 (22.2)
Often	2 (5.6)
Always	1 (2.8)
Confirming suspected diagnosis	
Never	10 (27.8)
Rarely	17 (47.2)
Sometimes	7 (19.4)
Often	1 (2.8)
Always	1 (2.8)
Answering miscellaneous questions	
Never	4 (11.1)
Rarely	12 (33.3)
Sometimes	12 (33.3)
Often	6 (16.7)
Always	2 (5.6)
Determining testing or treatment options	
Never	5 (13.9)
Rarely	20 (55.6)
Sometimes	4 (11.1)
Often	5 (13.9)
Always	2 (5.6)
Generating patient education or counseling materials	
Never	20 (55.6)
Rarely	9 (25)
Sometimes	3 (8.3)
Often	3 (8.3)
Always	1 (2.8)
Other purpose, not specified above	
No	33 (91.7)
Yes	3 (8.3)

[a]Respondents were asked, “How often, on average, do you use AI per day for [specific task]?” Rarely: 0%-25% of the time; Sometimes: 26%-50% of the time; Often: 51%-75% of the time; Always: 76% or more of the time.

[b]AI: artificial intelligence.

Discussion

Principal Findings

Our findings reveal that two-thirds of hospitalists at our institution use AI tools during clinical work, predominantly for answering miscellaneous clinical questions (32/36, 88.9%), generating differential diagnoses (31/36, 86.1%), and determining management options (31/36, 86.1%). Notably, this usage occurs despite the absence of system-level integration, institutional endorsement, or selective blocking of any AI platform interface, suggesting organic adoption driven by perceived clinical utility. The predominance of OpenEvidence among AI users (28/36, 77.8% users; 28/54, 51.9% of all

XSL•FO
RenderX

respondents) is particularly noteworthy. The high degree of convergence on this single niche platform, rather than more widely known general-purpose tools, raises questions about adoption pathways that warrant further investigation. The stark preference for specialized, medical-specific platforms (as defined by primary market positioning and intended clinical use) over general-purpose AI represents one of our most significant findings. Among AI users, OpenEvidence was chosen 7 times more frequently than ChatGPT (28 vs 4 respondents), despite ChatGPT's substantially greater name recognition and broader public adoption [10]. This pattern suggests that hospitalists are not simply adopting whatever AI tools are most readily available or widely marketed, but are instead making deliberate choices based on tool-specific characteristics relevant to medical practice.

Differential availability of various AI platforms could have been a confounding variable affecting observed preference for one platform over others. For instance, if hospital firewall systems blocked access to general-purpose AI while allowing medical use–focused platforms, that could affect adoption rates. However, this was not a factor in our study as our institution's computer systems do not block any particular AI platform.

Several factors may explain this preference for medical-specific platforms. First, hospitalists may perceive that specialized tools trained on or linked to medical literature will provide more accurate and relevant clinical information. Second, the use of general-purpose AI for patient care may raise concerns about data privacy, lack of possibility of HIPAA (Health Insurance Portability and Accountability Act) compliance, and liability that are less salient with tools explicitly designed for health care applications. Third, medical-specific platforms may provide source citations or references to primary literature, offering a verification pathway that general-purpose tools typically lack. Finally, hospitalists may view general-purpose AI as insufficiently tailored to the complexity and nuance of clinical decision-making, preferring tools that understand medical context and terminology.

This finding has important implications for AI development and deployment in health care. It suggests that clinicians value specialized tools over general-purpose solutions. Published literature validating OpenEvidence's capabilities as a clinical decision support tool for use by health care providers remains limited to generating management options for hypothetical primary care scenarios [25]; yet, hospitalists have adopted it widely based on its medical focus alone. This creates both an opportunity and a responsibility for developers of medical-specific AI tools to ensure their platforms meet the quality and safety standards that clinicians appear to assume they possess. The preference for medical-specific platforms also highlights a potential vulnerability that clinicians may place unwarranted trust in tools marketed as "medical AI" without rigorous independent validation of their accuracy and reliability compared with general-purpose alternatives.

The relatively low frequency of AI use, with most respondents reporting use in 0%-25% of their daily work, suggests that AI currently serves as a supplementary rather than primary clinical resource. Importantly, the survey did not specify whether frequency estimates should be calculated using all daily patient encounters or only those encounters in which the task was relevant. As a result, responses categorized as "rarely" (0%-25%) could reflect near-universal AI use for a small subset of high-acuity or diagnostically complex cases. This ambiguity limits the precision of our frequency estimates and necessitates caution when interpreting AI use as uniformly "low" or purely supplementary. We hypothesize that hospitalists are selectively deploying AI for cases where traditional resources such as UpToDate or PubMed searches are perceived as insufficient or time-intensive. The statistically significant differences in AI usage across clinical applications provide insight into how hospitalists perceive AI's strengths and limitations. The high usage for answering specific clinical questions, generating differential diagnoses, contrasted with lower usage for patient education materials (16/36, 44.4%), may reflect concerns about accuracy and liability when AI-generated content interfaces directly with patients, demonstrating appropriate professional judgment in the face of evolving technology, although it may also reflect a perception of low workflow efficiency in the setting of lack of institutional integration of AI.

Among nonusers (18/54, 33.3%), the primary concerns centered on preference for established resources, AI accuracy, and system-wide endorsement, followed by time constraints and questions about privacy. These findings highlight important barriers that must be addressed for broader adoption, such as validation of AI outputs, integration with existing workflows, and clear guidance on appropriate use cases. The absence of statistically significant demographic predictors of AI adoption suggests that usage patterns may be driven more by individual preferences and comfort with technology than by systematic factors, such as years in practice or clinical role.

Comparison With Previous Work

Purpose-built AI systems have been integrated into clinical workflows for narrowly defined tasks, such as radiologic interpretation [1,2], and experimental studies have compared AI performance against physicians in diagnostic and management scenarios [4-7]. However, adoption patterns of freely available AI tools by health care providers in the absence of institutional integration have remained largely uncharacterized. Our study addresses this knowledge gap by documenting how, and for what purposes hospitalists are independently incorporating AI into their daily practice and for what purposes.

Our observed adoption rate of 66.7% (95% CI 53.4%-77.8%) is comparable to the 66% reported in the 2024 AMA physician survey [23], suggesting that individual adoption patterns may mirror broader trends even without institutional mandates. However, the contexts and applications differ substantially. The AMA survey found that AI was used predominantly for administrative tasks, particularly clinical documentation ranging from discharge summaries to patient portal message responses [23]. Poon et al [19] found health systems primarily implementing AI for imaging or radiology, sepsis detection, and ambient listening clinical documentation. In contrast, our respondents used AI primarily for clinical decision support—answering medical questions, generating differential

XSL•FO
RenderX

diagnoses, and determining management options, although OpenEvidence, the platform chosen by a majority of our respondents, offers the capability to generate documentation meant for patient and health insurance use. This difference likely reflects the nature of institutional versus individual adoption—health systems implementing AI at an organizational level appear to prioritize electronic health record integration and workflow efficiency, whereas individual clinicians selecting their own tools focus on addressing clinical uncertainties and supplementing their decision-making process. Additionally, attempts by health care providers to use AI for documentation without pre-existing systematic integration involve extra steps (such as copy-pasting to and from AI platforms), which may also hinder adoption for such purposes. This distinction has important implications for how we conceptualize AI adoption; organizational implementation focuses on efficiency and administrative burden reduction, while individual adoption focuses on enhancing clinical reasoning and knowledge access.

Among nonusers (18/54, 33.3%) in our study, the primary concerns centered on AI accuracy and preference for established resources, consistent with the findings of Heinrichs et al [22] regarding physician skepticism about AI reliability. Interestingly, none of our respondents cited public or peer perception as factors influencing their decision not to use AI, despite recent studies demonstrating significant perception-based concerns. Reis et al [20] found that US adults rated physicians who used AI as less competent, trustworthy, and empathic, while Yang et al [21] showed that clinicians viewed peers using AI for primary decision-making as significantly less competent. The absence of perception concerns in our sample may reflect several possibilities, such as (1) hospitalists may be unaware of these negative perceptions, (2) they may not consider patient or peer opinions as relevant to their personal tool selection, (3) the anonymous nature of AI use—when used as a supplementary resource consulted privately rather than a visible component of patient-facing care—may shield providers from perception-based concerns, or (4) the lack of perception concerns as a selectable checkbox answer in the question led to the survey not capturing this as a reason, although there was a free-text option for those who wished to provide additional reasons. The discrepancy warrants further investigation, as understanding barriers to AI adoption—whether accuracy-based or perception-based—is essential for developing appropriate governance frameworks and educational initiatives.

The preference for medical-specific platforms observed in our study represents a novel finding not previously documented in the literature. While general-purpose AI tools, like ChatGPT, have achieved widespread adoption among employed adults [10], our hospitalists demonstrated a marked preference for domain-specific tools, with OpenEvidence accounting for the majority of AI users. This suggests that when clinicians independently select AI tools for clinical applications, they prioritize a medical focus and source transparency over general capability or familiarity. Whether this preference is justified by the superior performance of medical-specific platforms remains a question requiring comparative validation studies.

Implications

Understanding current AI usage patterns has significant implications for health care institutions, educators, and policymakers. The widespread organic adoption of AI tools in the absence of institutional oversight underscores an urgent need for governance frameworks that balance innovation with patient safety. Health care systems should consider developing policies that guide appropriate AI use, establish standards for tool selection, and provide training on recognizing AI limitations. As health care systems increasingly invest in AI technologies, empirical data on actual usage patterns become essential for resource allocation and strategic planning. Rather than implementing AI systems that remain unused, institutions can leverage usage data to identify high-value applications that align with clinician workflows and needs. Furthermore, medical education must evolve to prepare trainees for AI-augmented clinical practice, including teaching critical appraisal of AI outputs, understanding the underlying technology and its limitations, and developing frameworks for integrating AI responsibly into clinical decision-making. The convergence on a single platform among users in our study highlights the importance of understanding AI adoption pathways. Whether through peer-to-peer sharing, targeted marketing, or other mechanisms, clinicians are selecting specific tools in ways that may not be captured by traditional technology adoption models. Health care institutions should recognize that even in the absence of formal mandates, departmental culture and informal peer networks likely shape which AI tools gain traction, underscoring the need for governance frameworks that can adapt to grassroots adoption patterns.

The strong preference for a medical-specific AI platform over general-purpose tools observed in our study suggests that clinicians are not passive recipients of whatever AI technology becomes available but are actively selecting tools based on perceived fit with medical practice. This preference may reflect appropriate professional judgment about the importance of medical domain knowledge, source transparency, and health care–specific design. Health care institutions and regulatory bodies should recognize that clinicians' preference for specialized tools creates a need for rigorous, independent validation of these platforms' accuracy, reliability, and safety.

A critical consideration in AI adoption is the dynamic nature of these systems. AI platforms undergo frequent updates by their manufacturers; yet, each iteration does not necessarily represent an improvement over its predecessor [26]. Unlike static clinical resources, such as textbooks or established guidelines, AI systems may change their outputs without notification to users, in the form of variation in accuracy of answers, and ability to perform tasks, such as following multistep questioning between models and even over time within the same model [27], introducing uncertainty into clinical application. As AI use in health care increases, governance frameworks for maintaining patient safety need to be put in place. Zhang and Zhang [28] identified challenges that impact individual and organization trust in AI, including concerns about the quality of data used to train AI systems, accuracy of the underlying algorithms, robustness of health care data protection, responsibility for actions based on AI results, and lack of full

XSL•FO
RenderX

transparency and understanding about the inner workings of how AI produces results from data input [28]. There is no broad regulatory oversight of AI use in health care at present. In the United States, the Food and Drug Administration regulates only some AI systems, such as purpose-built AI for use on medical devices like defibrillators [29].

Given these challenges, primary source verification remains essential when using AI in clinical practice. We recommend that clinicians treat AI-generated information as a starting point requiring confirmation rather than a definitive answer. This principle is particularly important for hospitalists managing acutely ill patients, where diagnostic or therapeutic errors can have immediate consequences.

Limitations

This study has several limitations. As a single-center investigation conducted at a large academic tertiary care hospital, our findings may not be generalizable to community hospitals, rural settings, or outpatient practice environments, where clinical workflows and case complexity differ substantially. The cross-sectional design provides a snapshot of current practice but cannot capture temporal trends or the trajectory of AI adoption. Self-reported usage data may be subject to recall bias, social desirability bias, or selection bias, with respondents possibly more technologically inclined than nonrespondents. Notably, our calculated adoption rate of 66.7% is likely an overestimate of the true institutional adoption rate, as nonusers of AI may have been less motivated to complete a voluntary survey about AI usage patterns compared with enthusiastic users. Additionally, this rate includes respondents across the spectrum of frequency of use, ranging from those who only rarely use AI to power users who use AI much more often. We were unable to directly compare every aspect of respondent demographics with those of the full hospitalist provider group, as detailed information regarding some categories, such as age and experience levels, is not available for nonrespondents. Although the overall response rate was high, residual nonresponse bias cannot be fully excluded. Although no statistically significant demographic predictors of AI use were identified, the relatively small number of AI users (n=36) and limited subgroup sizes (for instance, physicians vs APP) reduce statistical power. As such, the absence of detected associations should be interpreted as the absence of evidence rather than evidence of absence. Another minor limitation is the possible selection bias introduced by our survey design, which required responses (even if selecting a nonresponse option where available) to all questions. Our survey did not undergo formal pilot testing before wide dissemination, possibly having led to some ambiguities. Our frequency estimates asked respondents how often they used AI “per day” for each task without specifying whether to consider all patient encounters or only those where the task would be applicable. For instance, when rating frequency of AI use for generating differential diagnoses, some respondents may have considered all daily patient encounters as the denominator, while others may have mentally restricted it to encounters with diagnostic uncertainty. This ambiguity limits the precision of our frequency estimates, although the data clearly indicate that AI is used infrequently across all measured applications. There was some conceptual overlap between the provided choices of “never” and “rarely (0%-25%)” frequency categories, which could also affect our frequency estimates, although we expect that respondents who never use AI for a specific task, such as for generating a differential diagnosis list, would select “never” over “rarely” as an option. Also, our survey asked respondents for years of practice, which some may have interpreted to include training years, possibly confounding the analysis of experience levels. The survey instruction for years of practice asked respondents to round up to the nearest whole year, which may have resulted in potentially shifting some respondents into higher experience brackets. Although respondents were provided with a free-text option to report additional AI use cases, clinical documentation (separate from generating patient education materials) was not included as one of the predefined questions, which could have led to underreporting compared with tasks explicitly prompted in the survey. This could have contributed to the lower observed prevalence of documentation-related AI use relative to the AMA physician survey, in addition to workflow friction and lack of electronic health record integration. We did not collect data on the accuracy of AI outputs in actual clinical use or on whether AI recommendations were followed, limiting our ability to assess the impact of AI on clinical decision-making. Additionally, we only asked respondents which AI platform (if any) they used most often, and did not collect data on whether they were using more than 1 system. The preference for OpenEvidence may reflect a pattern of primary use (with additional tools used for other scenarios) rather than confirmation of exclusive use.

An important limitation is our lack of data regarding how respondents discovered and selected their AI tools. The striking convergence on OpenEvidence is noteworthy given its status as a niche platform compared with widely known tools, like ChatGPT. Also, although ClinicalKey AI is also a medical-specific platform, its adoption remained low. To our knowledge, there was no formal departmental introduction or recommendation of any platform, although informal peer-to-peer recommendations may have occurred through channels we did not measure. Without systematic data on discovery pathways, we cannot determine whether this convergence reflects informal sharing within our department, the platform’s targeted marketing toward medical professionals, lack of cost barriers, or other factors. While our institution does not block online access to any particular AI platform, other organizations may allow differential access to some systems versus others, which could affect local adoption patterns. Future studies should systematically assess how clinicians learn about and select AI tools, as understanding these adoption pathways has important implications for dissemination strategies and governance frameworks.

Furthermore, our survey did not capture the specific clinical scenarios or patient characteristics that prompted AI use. Despite these limitations, our study has notable strengths. The high response rate (77.1%) enhances the reliability of our findings and minimizes nonresponse bias. The comprehensive assessment of multiple AI platforms and diverse clinical applications they are being used for provides a nuanced picture of real-world usage. Our study to systematically document hospitalist AI

XSL•FO
RenderX

usage patterns independent of organizational endorsement establishes baseline data for future longitudinal investigations and comparative studies across institutions and specialties.

Future Directions

These findings point to several important directions for future research. Longitudinal studies are needed to track how AI usage patterns evolve as these technologies mature and as clinicians gain more experience with them. Comparative effectiveness research should evaluate clinical outcomes associated with AI-augmented versus traditional clinical decision-making. Additionally, research should examine whether AI usage varies across different clinical contexts, patient populations, and decision types, as understanding these nuances could inform the development of more targeted AI applications that address specific clinical needs. This includes future multicenter studies, which should explore interinstitutional variation and correlate AI adoption pathways and usage with clinical outcomes. Ongoing validation studies must continually assess AI system performance across the tasks they are commonly used for, with particular attention to detecting degradation or drift in performance over time. Public and peer-perception–based barriers to AI adoption also warrant further exploration. Finally, comparative studies examining the accuracy, reliability, and clinical utility of medical-specific AI platforms versus general-purpose LLMs for clinical applications would help determine whether the observed preference for specialized tools is justified by superior performance.

Conclusions

This study documents substantial organic adoption of AI tools among hospitalists, with two-thirds reporting clinical use despite the absence of institutional integration or endorsement. The predominant applications—answering clinical questions, generating differential diagnoses, and determining management options—suggest that AI is serving as a supplementary decision support tool for complex cases. The marked preference for a medical-specific platform over a general-purpose AI indicates that hospitalists are exercising discretion in tool selection. This finding suggests both appropriate caution and a need for rigorous validation of medical-specific platforms. However, concerns about accuracy and the low frequency of use highlight the nascent state of AI integration in clinical practice. As AI technologies continue to evolve, health care institutions must develop governance frameworks, validation protocols, and educational initiatives to ensure these tools are deployed safely and effectively in patient care.

Acknowledgments

The authors wish to acknowledge the contributions of Chengcheng Hu (MPH, MMed) for contributing to the statistical analysis of the data. Generative artificial intelligence was not used for generating any part of this manuscript.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: PB (lead), JH (equal), SW (equal)
Data curation: PB (lead), BM (supporting), SW (supporting)
Formal analysis: Chengcheng Hu
Investigation:PB (lead), BM (equal), SW (equal)
Methodology: PB (lead), JH (equal), BM (supporting), SW (supporting)
Project administration: JH (lead), PB (supporting)
Supervision: SW
Visualization: PB (lead), SW (supporting)
Writing: original draft – PB (lead), SW (supporting)
Writing: review & editing – PB (lead), JH (equal), BM (equal), SW (equal)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Copy of the survey questionnaire.
[DOCX File , 22 KB - jmir_v28i1e85973_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
CHERRIES checklist.
[DOCX File , 23 KB - jmir_v28i1e85973_app2.docx]

Multimedia Appendix 3
Distribution of artificial intelligence platforms used by hospitalists in clinical work (n=54). Survey respondents from a large academic medical center, August-September 2025.
[PNG File , 102 KB - jmir_v28i1e85973_app3.png]

Multimedia Appendix 4
Clinical tasks for which artificial intelligence is used among hospitalist artificial intelligence users (n=36). Survey respondents from a large academic medical center, August-September 2025.
[PNG File , 197 KB - jmir_v28i1e85973_app4.png]

References

1. Maleki Varnosfaderani S, Forouzanfar M. The role of AI in hospitals and clinics: transforming healthcare in the 21st century. Bioengineering (Basel) 2024;11(4):337 [FREE Full text] [doi: 10.3390/bioengineering11040337] [Medline: 38671759]
2. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
3. Williams CYK, Zack T, Miao BY, Sushil M, Wang M, Kornblith AE, et al. Use of a large language model to assess clinical acuity of adults in the emergency department. JAMA Netw Open 2024;7(5):e248895 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.8895] [Medline: 38713466]
4. Takita H, Kabata D, Walston SL, Tatekawa H, Saito K, Tsujimoto Y, et al. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. NPJ Digit Med 2025;8(1):175 [FREE Full text] [doi: 10.1038/s41746-025-01543-z] [Medline: 40121370]
5. Maillard A, Micheli G, Lefevre L, Guyonnet C, Poyart C, Canouï E, et al. Can chatbot artificial intelligence replace infectious diseases physicians in the management of bloodstream infections? A prospective cohort study. Clin Infect Dis 2024;78(4):825-832. [doi: 10.1093/cid/ciad632] [Medline: 37823416]
6. Kafetzis I, Sodmann P, Herghelegiu B, Brand M, Zoller WG, Seyfried F, et al. Prospective evaluation of real-time artificial intelligence for the Hill classification of the gastroesophageal junction. United European Gastroenterol J 2025;13(2):240-246 [FREE Full text] [doi: 10.1002/ueg2.12721] [Medline: 39668544]
7. Beyaz S, Yayli SB, Kılıç E, Kılıç K. Comparison of artificial intelligence algorithm for the diagnosis of hip fracture on plain radiography with decision-making physicians: a validation study. Acta Orthop Traumatol Turc 2024;58(1):4-9 [FREE Full text] [doi: 10.5152/j.aott.2024.23065] [Medline: 38525504]
8. What is LLM (Large Language Model)? AWS. URL: https://aws.amazon.com/what-is/large-language-model/ [accessed 2026-01-12]
9. ChatGPT. OpenAI. URL: https://chatgpt.com/overview/ [accessed 2025-12-18]
10. Sidoti O, McClain C. 34% of U.S. adults have used ChatGPT, about double the share in 2023. Pew Research Center. 2025. URL: https://tinyurl.com/5n8n4z5r [accessed 2025-11-12]
11. Reid E. AI in Search: going beyond information to intelligence. Google blog. 2025. URL: https://blog.google/products/search/google-search-ai-mode-update/ [accessed 2025-11-12]
12. Mehdi Y. Announcing the next wave of AI innovation with Microsoft Bing and Edge. Microsoft Office blog. 2023. URL: https://blogs.microsoft.com/blog/2023/05/04/announcing-the-next-wave-of-ai-innovation-with-microsoft-bing-and-edge/ [accessed 2025-11-12]
13. Google. Learn about Gemini, the everyday AI assistant from Google. Google. URL: https://gemini.google/about/ [accessed 2025-12-18]
14. Getting started with Microsoft Copilot. Microsoft Support. Microsoft. URL: https://support.microsoft.com/en-us/topic/getting-started-with-microsoft-copilot-8fde147f-726e-4790-9503-70790ddcac73 [accessed 2025-12-18]
15. What is RAG (retrieval-augmented generation)? AWS. URL: https://aws.amazon.com/what-is/retrieval-augmented-generation/ [accessed 2026-01-12]
16. ClinicalKey AI. Elsevier. URL: https://www.elsevier.com/products/clinicalkey/clinicalkey-ai [accessed 2025-12-18]
17. About OpenEvidence. OpenEvidence. URL: https://www.openevidence.com/about [accessed 2025-12-18]
18. Angus DC, Khera R, Lieu T, Liu V, Ahmad FS, Anderson B, JAMA Summit on AI. AI, health, and health care today and tomorrow: the JAMA Summit report on artificial intelligence. JAMA 2025;334(18):1650-1664. [doi: 10.1001/jama.2025.18490] [Medline: 41082366]

XSL•FO
RenderX

19. Poon EG, Lemak CH, Rojas JC, Guptill J, Classen D. Adoption of artificial intelligence in healthcare: survey of health system priorities, successes, and challenges. J Am Med Inform Assoc 2025;32(7):1093-1100. [doi: 10.1093/jamia/ocaf065] [Medline: 40323320]
20. Reis M, Reis F, Kunde W. Public perception of physicians who use artificial intelligence. JAMA Netw Open 2025;8(7):e2521643 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.21643] [Medline: 40674054]
21. Yang H, Dai T, Mathioudakis N, Knight AM, Nakayasu Y, Wolf RM. Peer perceptions of clinicians using generative AI in medical decision-making. NPJ Digit Med 2025;8(1):530 [FREE Full text] [doi: 10.1038/s41746-025-01901-x] [Medline: 40826224]
22. Heinrichs H, Kies A, Nagel SK, Kiessling F. Physicians' attitudes toward artificial intelligence in medicine: mixed methods survey and interview study. J Med Internet Res 2025;27:e74187 [FREE Full text] [doi: 10.2196/74187] [Medline: 40857713]
23. Physician sentiments around the use of AI in health care: motivations, opportunities, risks, and use cases. Shifts from 2023 to 2024. AMA Augmented Intelligence Research. 2025. URL: https://www.ama-assn.org/system/files/physician-ai-sentiment-report.pdf [accessed 2025-12-15]
24. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
25. Hurt RT, Stephenson CR, Gilman EA, Aakre CA, Croghan IT, Mundi MS, et al. The use of an artificial intelligence platform OpenEvidence to augment clinical decision-making for primary care physicians. J Prim Care Community Health 2025;16:21501319251332215 [FREE Full text] [doi: 10.1177/21501319251332215] [Medline: 40238861]
26. Zeff M. OpenAI's new reasoning AI models hallucinate more. TechCrunch. 2025. URL: https://techcrunch.com/2025/04/18/openais-new-reasoning-ai-models-hallucinate-more/ [accessed 2025-10-10]
27. Chen L, Zaharia M, Zou J. How is CHATGPT's behavior changing over time? Harvard Data Science Review. 2024. URL: https://hdsr.mitpress.mit.edu/pub/y95zitmz [accessed 2025-12-17]
28. Zhang J, Zhang ZM. Ethics and governance of trustworthy medical artificial intelligence. BMC Med Inform Decis Mak 2023;23(1):7 [FREE Full text] [doi: 10.1186/s12911-023-02103-9] [Medline: 36639799]
29. Warraich HJ, Tazbaz T, Califf RM. FDA perspective on the regulation of artificial intelligence in health care and biomedicine. JAMA 2025;333(3):241-247. [doi: 10.1001/jama.2024.21451] [Medline: 39405330]

Abbreviations

AI: artificial intelligence
AMA: American Medical Association
APP: advanced practice provider
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
RAG: retrieval-augmented generation

Edited by A Coristine; submitted 15.Oct.2025; peer-reviewed by J Grosser, S Shehata; comments to author 10.Nov.2025; accepted 21.Jan.2026; published 03.Mar.2026.

Please cite as:
Bagla P, Hanna J, Marthambadi B, Watkins S
Patterns of AI Use in Clinical Work by Hospitalists: Survey Study
J Med Internet Res 2026;28:e85973
URL: https://www.jmir.org/2026/1/e85973
doi:10.2196/85973
PMID:41581898

©Prabhava Bagla, Jasmah Hanna, Bhargav Marthambadi, Stacey Watkins. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Psychosis Risk and Generative Artificial Intelligence Use Frequency, Motivations, and Delusion-Like Experiences: Cross-Sectional Survey Study

Benjamin Buck[1], PhD; Anne Julia Maheux[2], PhD

[1]Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[2]Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

Corresponding Author:
Benjamin Buck, PhD
Department of Psychiatry
University of North Carolina at Chapel Hill
MacNider Hall
333 S Columbia Street
Chapel Hill, NC, 27514
United States
Phone: 1 9194454330
Email: ben.buck@med.unc.edu

Abstract

Background: Growth of generative artificial intelligence (GenAI) has exploded in recent years. Many have noted its substantial potential to increase access to scalable digital mental health interventions or provide companions for individuals who are socially isolated. At the same time, seeking mental health support from mainstream GenAI models may involve risks. Several recent examples of exacerbation of delusions have received attention in the popular press, leading to a call for empirical research to document the scope of interactions with GenAI among individuals experiencing symptoms of psychosis.

Objective: This study aimed to evaluate associations of psychosis risk to GenAI use frequency, motivations for use, and GenAI interactions involving potential delusions.

Methods: We conducted a large-scale cross-sectional survey of 1003 young adults in United States, divided the sample of individuals that had used GenAI into “elevated risk” (Prodromal Questionnaire, Brief Version Distress Score ≥20; N=267, 28%) and “low risk” groups (Prodromal Questionnaire, Brief Version Distress Score <20; N=685, 72%), and compared groups on several assessments related to GenAI use.

Results: We found that while members of the elevated risk group were no more likely to have ever used GenAI, they were significantly more likely to report intensive use (odds ratio 1.70 to 2.56; ie, several times per day, more than 30 minutes per day, 6 or more chatbot conversations per day). Those at elevated risk were more likely to report using GenAI to receive social and emotional support and significantly more likely to ascribe human-like roles to their chatbot interactions (odds ratio 1.76 to 3.08; ie, companion, friend, therapist, and romantic partner). Delusion-related interactions were also commonly reported among those at risk for psychosis (item endorsements from 13.3% to 30.7%).

Conclusions: While it is unclear whether they have a positive or negative impact overall, GenAI chatbots may have the potential to impact symptom-related experiences among young adults at risk.

(J Med Internet Res 2026;28:e85038) doi:10.2196/85038

KEYWORDS

digital health; digital mental health; artificial intelligence; generative AI; psychosis risk; psychosis; young adults; mental health

Introduction

Generative artificial intelligence (GenAI) is now the most rapidly adopted technology in history [1]. In July 2025, less than 3 years after its launch in November 2022, ChatGPT (OpenAI) boasted 700 million users, or roughly one tenth of the world’s population [2]. Beyond ChatGPT, several others have experienced similar growth, including other general-use artificial intelligence (AI) platforms (eg, Claude [Anthropic]

XSL•FO
RenderX

and Gemini [Google]), those focused on image and video generation (eg, Stable Diffusion), and others that allow for interaction with customizable characters (eg, Character.AI). Many have noted the potential of GenAI technologies to deliver scalable and self-guided mental health interventions [3], with emerging evidence demonstrating efficacy of interventions leveraging GenAI relative to waitlist control [4]. Survey studies suggest that GenAI systems are already used both by individuals seeking mental health support and by mental health providers to increase efficiency (eg, documenting clinical interactions or researching clinical literature) [5], and prevalence of use is nearly certain to increase in years to come.

While the potential benefit in expanding access to mental health interventions and resources is clear, many have also been monitoring for potential risks, particularly when unrestricted, general use chatbots are used for mental health support [6]. One particular concern is sycophantic behavior, wherein GenAI chatbots overly flatter or agree with users [7], which has been noted in general use GenAI [8]. Another is hallucination or misinformation, a widely observed phenomenon wherein GenAI chatbots provide incorrect details with an overall convincing presentation [9]. Some of these limitations may strengthen atypical or pathological beliefs rather than challenge them. One prominent example of harm occurred in 2023, when Tessa, a chatbot provided by the National Eating Disorders Association, provided problematic guidance related to eating disorders, including encouraging strategies for weight loss [10]. While more recent research on GenAI chatbots for mental health have demonstrated improved efficacy and capacity for preventing negative outcomes [4,11], there remain questions about risks associated with GenAI use for mental health concerns, particularly when the GenAI platforms are not specialty systems designed for such a purpose and with guardrails.

An emerging concern in this area is psychosis risk [12]. Psychosis symptoms can be arranged along a transdiagnostic risk continuum, ranging from common and low severity symptoms (eg, fleeting and infrequent hallucinatory experiences, preoccupation with odd beliefs) to rare, highly severe symptoms (eg, ongoing command hallucinations, highly distressing and high-conviction delusions) that are often linked with negative outcomes [13]. Schizophrenia-spectrum disorders emerge following a prodromal period (typically occurring in late teens to mid-twenties) wherein individuals experience subthreshold symptoms and changes from premorbid functioning. Experiencing these prodromal symptoms is associated with a significant increase in likelihood of both schizophrenia-spectrum [13] or other psychiatric disorders [14]. Young people within the age group that is at the highest risk for first developing a psychotic disorder are among the highest adopters of GenAI [1]. Individuals at elevated risk for psychosis are also more likely to engage in problematic use of other digital media [15-17], and such problematic use appears associated with elevated stress in this population [18,19] as well as risk for future symptoms [20]. Taken together, these findings suggest a need for research to understand the use of GenAI among individuals at elevated risk for psychosis.

Several well-publicized examples of troubling interactions with GenAI systems have raised concerns about the ways these technologies could introduce new risks for individuals at risk for psychosis. In several interactions described in the popular press, users have shared delusions or odd beliefs with the model, and in response, the GenAI chatbot has reportedly responded in ways that seem to have increased delusional conviction, distress, or behavioral responses. For example, in one case, a 29-year-old mother of 2 came to believe that ChatGPT could facilitate “interdimensional communication” [21]. In another, a 35-year-old man with a history of bipolar disorder and schizophrenia became fixated on an AI entity called Juliet; when he came to believe that OpenAI executives had killed this entity, he threatened to kill them and was killed after attacking law enforcement with a knife [21]. In other, less tragic examples, individuals with no history of mental illness have embraced grandiose narratives about their own knowledge or abilities [22]. While these extreme examples are a concern, there is a lack of evidence to inform the public as to the prevalence of these risks at the population level. It is unclear the extent to which publicized cases with negative outcomes may indicate a more common phenomenon. A first step in answering this question involves assessing the amount and type of GenAI use among young adults at varying risk levels for psychosis, evaluating whether their interactions with GenAI chatbots differ in meaningful ways from the general population, and to assess the frequency of interactions that could impact mental health outcomes. There is a lack of empirical studies on relationships of psychosis risk to (1) GenAI use frequency and (2) motivations for GenAI use, and (3) delusion-related interactions with GenAI models.

Our team conducted a large-scale (N=1003) survey of young adults in July 2025. In addition to measures of GenAI use frequency, motivations for GenAI use, and GenAI experiences, we collected a screening questionnaire assessing psychosis risk. In this study, we examine cross-sectional relationships of GenAI variables to symptoms indicating elevated psychosis risk. Based on literature suggesting elevated rates of problematic use of digital media among individuals at risk for psychosis, we expect that individuals at elevated risk will be more likely to report the highest patterns of use and to use GenAI for social and emotional support. We also expect that individuals at elevated risk will be more likely to report interacting with GenAI about symptom-related content. Study results could provide an important first step to identifying the potential risks and benefits of these platforms for individuals at risk of developing psychosis, as well as characterizing their scale and level of impact.

Methods

Study Design

We collected a cross-sectional survey in July 2025 via the online crowdsourcing platform Prolific. Our survey was open to all Prolific users, given they met our inclusion criteria of (1) residing in the United States and (2) being between ages 18 and 25 years. Convenience sampling was used as our sample comprised all individuals who responded to the study listing. Our study had multiple goals. First, we aimed to examine overall relationships between psychosis risk and self-reported use of

XSL•FO
RenderX

GenAI (eg, lifetime use, frequency, and average duration of use). Second, we sought to characterize motivations for and types of GenAI use by examining associations with psychosis risk symptoms. Third, we developed a measure to assess delusion-related interactions with GenAI systems to characterize the frequency with which individuals at elevated risk for psychosis reported these experiences.

Participants

Participants were young adults between the ages of 18 and 25 years in the United States (for sample demographics, see Multimedia Appendix 1). On Prolific, participants can complete screening, provide identity verification, and receive compensation through Prolific without sharing identifying information with researchers. Thus, all study participants were anonymous to the study team, and all received information about crisis support or finding mental health resources during study participation. This sample includes all individuals who provided sufficient responses to the Prodromal Questionnaire, Brief Version (PQ-B [23]) to generate an estimate of psychosis risk (ie, 14 or more of the 21 items). This exclusion criterion resulted in removing only 7 participants from the original study sample. In addition to 44 (4.4%) that were removed for missing at least 1 of 2 attention checks, this resulted in a final sample of 952 participants.

Measures

Psychosis risk was assessed with the PQ-B [23]. The PQ-B is widely used in identifying individuals experiencing symptoms that may indicate risk for psychotic episodes (ie, prodromal symptoms). The PQ-B has 21 items that are rated on their occurrence and a Likert scale assessing distress related to items that the respondent endorsed. While a high score on the PQ-B does not on its own indicate meeting clinical high-risk criteria (a gold-standard determination of this is made by a clinician), it does indicate that a formal evaluation is more likely to find that the respondent is indeed at risk. And in population-level research, the PQ-B Distress Score, which captures both symptoms endorsed and associated distress, can be used continuously as a measure of psychosis risk symptoms. Systematic reviews of the PQ-B have identified optimal cutoffs for screening and referral for further evaluation in both clinical and nonclinical settings [24]. For analyses that require identifying a subset of participants at elevated risk, we follow a provisional cutoff of 20 used in prior work in both clinical [25] and online settings [26] to generate “elevated risk” (PQ-B Distress Total ≥20) and “low risk” (PQ-B Distress Total <20).

Our team developed measures to evaluate AI use frequency, motivations for use, and interactions with GenAI involving delusion-like experiences. Regarding use, we administered 4 items addressing frequency, recency, typical length of sessions, and typical number of daily sessions (see Multimedia Appendix 2 for items and response options). The Artificial Intelligence Motivation and Uses Scale (AIMUS) surveys participants’ motivations for and types of use of GenAI systems in the following categories, derived via a factor analysis conducted by our team that is as yet unpublished [27] (see Multimedia Appendix 1 for more psychometric information about the scale, all items, and factor loadings): (1) task automation (ie, completing onerous or rote tasks with the assistance of the system), (2) learning and exploration (ie, searching for information, generating ideas, or receiving help understanding difficult topics), (3) emotional support (ie, support in managing emotions, thinking through decisions, or managing worrying thoughts), (4) dating and sexuality (ie, interacting with GenAI in romantic or sexual role-plays). Participants endorse items on a Likert scale ranging from 1 (“Never”) to 5 (“Very Often”). Subscale scores were mean imputed for individuals that completed at least half the items on the scale. We also asked one item about relationships that respondents felt they had with AI systems. For these items, participants were asked to report whether they considered any AI system to be a companion, a therapist, a friend, a romantic partner, or a sexual partner.

Delusion-like experiences involving AI were assessed with the Generative AI Aberrant Thoughts and Experiences Scale (GAATES). To develop the GAATES, study team members identified the most common delusion categories from recent meta-analytic work on psychosis [28], and created items specific to GenAI interactions. These depicted both (1) potential delusions about AI (eg, paranoid thoughts about AI, grandiose thoughts about one’s interactions with AI), as well as (2) uses of AI that may maintain existing delusions (eg, interacting with AI to reinforce one’s paranoid or grandiose beliefs). Item responses range from 1 (“Strongly Disagree”) to 5 (“Strongly Agree”). Exploratory factor analysis indicated that the scale is unidimensional, with all items loading (loadings >.45) onto a single factor that explained 53% of the variance in the underlying construct. For analyses examining endorsement of each phenomenon as a binary variable, responses of “Agree” or “Strongly Agree” (ie, 4 or 5) were coded as endorsement, while all other responses (including 3, “Neither Agree nor Disagree”) were coded as nonendorsement. Scale internal consistency was high (Cronbach α=0.95). Full item responses by group as well as results from the psychometric analyses (eg, factor analysis) are reported in the Multimedia Appendix 1.

Data Analysis

Given the nature of the scientific question, we took an exploratory approach to characterize relationships and group differences broadly across a variety of constructs. Aside from the variable related to previous GenAI use itself, all analyses included only the subset of the sample reporting current or previous GenAI use. Alpha level for statistical significance was set at .05. First, to examine the relationships of psychosis risk to use frequency variables, we split the sample based on frequency responses indicating intensive use of GenAI. This generated binary variables representing four comparisons: (1) individuals who reported using GenAI several times per day versus less, (2) those who reported using GenAI the day of the survey versus longer ago, (3) those who reported using GenAI more than 30 minutes per session versus shorter than this, and (4) those who reported having 6 or more conversations with the chatbot per day versus 5 or fewer. We conducted chi-square tests and computed odds ratios (ORs) based on the cross product of cells in contingency tables, with CIs based on the standard error of the log ORs. PQ-B Distress scores across were compared across high versus low GenAI frequency using independent samples *t* tests and as well as chi-square tests and

XSL•FO
RenderX

ORs (also derived from contingency tables) to characterize the frequency of intensive AI use among each psychosis risk group (elevated risk vs low risk). Cohen *d* estimates were calculated using pooled SDs. Second, to evaluate relationships to AI motivations, we examined Pearson correlations of PQ-B Distress scores to each subscale on the AIMUS. Third, to assess relationship variables, we compared the frequency of endorsing each relationship (friend, therapist, companion, romantic partner, and sexual partner) among each risk group using chi-square tests compared PQ-B Distress scores among individuals who endorsed each relationship using independent samples *t* tests. Finally, we examined the Pearson correlation of the GAATES to PQ-B Distress score and compared frequency of endorsing each delusion-related item across risk groups using chi-square tests and ORs. For all Pearson correlations, CIs were generated using Fisher *r*-to-*z* transformations. All linear models were evaluated for assumption violations; nonnormal variables were reanalyzed with appropriate nonparametric models and/or variable transformations, and the pattern of results were unchanged (Multimedia Appendix 1).

Ethical Considerations

All research procedures were reviewed by the institutional review board at the University of North Carolina at Chapel Hill and were deemed exempt (IRB#25-1686) on June 26, 2025. All participants provided informed consent through online data capture, and because no individually identifying information was collected, data were anonymized. Based on an estimated 30-minute completion time and an intended compensation rate of US $8 per hour, participants were compensated US $4 for their time.

Results

Frequency and Average Length of Use

Of the sample of 952, 846 (88.9%) had used GenAI, while the remainder (N=106, 11.1%) had not. There were no significant differences in the PQ-B Distress score between individuals who had never used GenAI platforms (mean 13.05, SD 17.06) and those who had (mean 13.90, SD 17.14; mean difference=–0.85, 95% CI –4.31 to 2.61; t_{950}=–0.48; Cohen *d*=–0.05, 95% CI –0.25 to 0.15; *P*=.63). Participant demographics can be found in Multimedia Appendix 1.

Regarding frequency, individuals who reported using GenAI “several times per day” or more had significantly elevated PQ-B Distress scores (113/845, 13.4%; mean 21.75, SD 20.98) relative those reporting using GenAI less frequently (732/845, 86.6%; mean 12.71, SD 16.15), mean difference=–9.04 (95% CI –12.39 to –5.69), t_{843}=–5.30, *P*<.001, Cohen *d*=–0.54 (95% CI –0.74 to –0.34). Individuals who reported that their most recent use of GenAI was the day of the survey (ie, “Today”; N=287, 34.0%) had significantly elevated PQ-B Distress scores (mean 17.00, SD 19.54) relative to those who reported that their most recent use was the day prior or earlier (558/845, 66.0%; mean 12.33, SD 15.54; mean difference=–4.66, 95% CI –7.09 to –2.24; t_{843}=–3.77; Cohen *d*=–0.27, 95% CI –0.42 to –0.13; *P*<.001). Individuals that reported that they would initiate over 5 conversation sessions per day (96/803, 12.0%) also had significantly elevated PQ-B Distress scores (mean 20.51, SD 21.42) relative to those who reported having 5 conversation sessions or fewer per day (707/803, 88.0%; mean 13.15, SD 16.53; mean difference=–7.36, 95% CI –11.03 to –3.69; t_{801}=–3.94; Cohen *d*=–0.43, 95% CI –0.64 to –0.21; *P*<.001). Expressed as ORs in binary analyses (Table 1), individuals at elevated risk for psychosis (using PQ-B Distress Total ≥ 20) had more than double the odds of being in the respective intensive use category regarding frequency of use and number of conversation sessions, and they were over 50% higher odds of reporting GenAI use on the day of the survey. There were no differences between groups divided on the variable of length of each episode of use.

XSL•FO
RenderX

Table 1. Responses to generative artificial intelligence frequency items among individuals endorsing lifetime generative artificial intelligence use (N=846). Numbers vary by item as a result of missing responses.

Variables	Low risk (n=605), n (%)	Elevated risk (n=241), n (%)	OR[a] (95% CI)	Chi-square (*df*)	*P* value
How often do you use generative AI[b] chatbots (eg, ChatGPT, Claude, and Gemini) for any purpose? (N=845)			2.56 (1.71-3.83)	21.62 (1)	<.001
Less than "Several Times Per Day" (n=732, 86.6%)	544 (90.1)	188 (78.0)			
"Several Times Per Day" (n=113, 13.4%)	60 (9.9)	53 (22.0)			
When was the last time you used a generative AI chatbot? (N=845)			1.70 (1.25-2.32)	11.57 (1)	<.001
Any time before "Today" (n=558, 66%)	420 (69.5)	138 (57.3)			
"Today" (n=287, 34%)	184 (30.5)	103 (42.7)			
When you do use generative AI chatbots, how much time do you typically spend per session? (N=846)			1.32 (0.84-2.08)	1.42 (1)	0.23
Less than 30 minutes (n=751, 88.8%)	542 (89.6)	209 (86.7)			
More than 30 minutes (n=95, 11.2%)	63 (10.4)	32 (13.3)			
On a day when you use a generative AI chatbot, how many separate times do you typically start a new conversation or ask it about something different? (N=803)[c]			2.10 (1.36-3.25)	11.49 (1)	<.001
Five or fewer times (707, 88%)	516 (90.5)	191 (82.0)			
Six or more times (96, 12%)	54 (9.5)	42 (18.0)			

[a]OR: odds ratio.

[b]AI: artificial intelligence.

[c]This excludes responses of "I don't know/I haven't paid attention," which were provided as a response item for this question.

Motivations for AI Use

All motivations for use were positively associated with psychosis risk in the full sample, with the strongest correlations being between PQ-B distress scores and emotional support motivations (r=0.33, 95% CI 0.26-0.38; P<.001), followed by dating and sexuality (r=0.23, 95% CI 0.16-0.29; P<.001), task automation (r=0.15, 95% CI 0.08-0.21; P<.001) and learning and exploration (r=0.14, 95% CI 0.08 to 0.21; P<.001).

AI Relationship Variables

Of the AI relationships individuals reported, items ranged in endorsement from around 3% (AI as a sexual partner; 26/846, 3.1%) to over a third (AI as a companion; 315/846, 37.2%) of all participants in the sample who had previously used GenAI. PQ-B Distress Scores were higher for individuals that endorsed feeling that AI served as a companion (mean 16.75, SD 18.73 vs mean 12.20, SD 15.89; mean difference=–4.55, 95% CI –6.92 to –2.18; t_{844}=–3.76; Cohen d=–0.27, 95% CI –0.41 to –0.13; P<.001), therapist (mean 20.77, SD 20.59 vs mean 11.10, SD 14.63; mean difference=–9.67, 95% CI –12.14 to –7.21; t_{844}=–7.70; Cohen d=–0.58, 95% CI –0.74 to –0.43; P<.001), friend (mean 18.70, SD 19.53 vs mean 11.30, SD 15.08; mean difference=–7.40, 95% CI –9.77 to –5.03; t_{844}=–6.12; Cohen d=–0.44, 95% CI –0.58 to –0.30; P<.001), and romantic partner (mean 21.97, SD 18.89 vs mean 13.58, SD 17.00; mean difference=–8.39, 95% CI –14.43 to –2.35; t_{844}=–2.73; Cohen d=–0.49, 95% CI –0.85 to –0.14; P=.007). Examined as odds ratios between risk groups (Table 2), GenAI users in the elevated risk group had approximately threefold higher odds of reporting seeing GenAI as a therapist, more than double the odds of seeing GenAI as a friend or romantic partner, and more than 70% higher odds of seeing GenAI as a companion. There were no differences between the groups in the item related to viewing GenAI as a sexual partner.

XSL•FO
RenderX

Table 2. Responses to generative artificial intelligence relationship items among individuals endorsing lifetime generative artificial intelligence use (N=846). Numbers vary by item as a result of missing responses.

Variables	Low risk (n=605), n (%)	Elevated risk (n=241), n (%)	OR[a] (95% CI)	Chi-square (*df*)	*P* value
Companion (N=846)			1.76 (1.30-2.39)	13.44 (1)	<.001
No (n=531, 62.8%)	403 (66.6)	128 (53.1)			
Yes (n=315, 37.2%)	202 (33.4)	113 (46.9)			
Therapist (N=846)			3.08 (2.24-4.24)	50.24 (1)	<.001
No (n= 601, 71%)	472 (78.0)	129 (53.5)			
Yes (n=245, 29%)	133 (22.0)	112 (46.5)			
Friend (N=846)			2.52 (1.85-3.43)	35.61 (1)	<.001
No (n=549, 64.9%)	430 (71.1)	119 (49.4)			
Yes (n=297, 35.1%)	175 (28.9)	122 (50.6)			
Romantic partner (N=846)			2.62 (1.29-5.32)	7.56 (1)	.006
No (n=814, 96.2%)	589 (97.4)	225 (93.4)			
Yes (n=32, 3.8%)	16 (2.6)	16 (6.6)			
Sexual partner (N=846)			0.92 (0.38-2.22)	0.03 (1)	.86
No (n=820, 96.9%)	586 (96.9)	234 (97.1)			
Yes (n=26, 3.1%)	19 (3.1)	7 (2.9)			

[a]OR: odds ratio.

Delusion-Like Experiences

GAATES scores differed between individuals at elevated risk (mean 36.22, SD 14.74) and those at low risk (mean 26.39, SD 10.33; mean difference=–9.83, 95% CI –11.59 to –8.07; t_{844}=–10.98; Cohen *d*=–0.84, 95% CI –0.99 to –0.68; *P*<.001). Total GAATES scores significantly correlated with PQ-B Distress in the full sample (*r*=0.40, 95% CI 0.34-0.45; *P*<.001) as well as separately in the elevated risk group (*r*=0.28, 95% CI 0.16-0.39; *P*<.001) and the low risk group (*r*=0.11, 95% CI 0.03-0.19; *P*=.008). For all items, rates of endorsement frequency differed between risk groups (Table 3). Several of the items that most differed between those at low and elevated risk most often pertained to paranoia or persecutory ideation (eg, “AI provides me facts about how others are working to harm me”; 41/241, 17% in the elevated risk group vs 19/605, 3.1% in the low-risk group; “AI is being used by others to harm me”, 33/241, 13.7% vs 17/605, 2.8%; “AI has shown me how others are trying to control my actions”, 61/241, 25.3% vs 37/605, 6.1%), while others related to grandiosity (“I’ve discovered hidden or secret truths about the world through AI”, 43/241, 17.8% vs 22/605, 3.6%; “AI helps me make sense of secret messages (eg, from television or the news) that were intended only for me”, 58/241, 24.1% vs 42/605, 6.9%) also clearly distinguished the groups.

XSL•FO
RenderX

Table 3. Responses to individual Generative AI Aberrant Thoughts and Experiences Scale items among individuals endorsing lifetime generative artificial intelligence use (N=846).

Variables	Low risk (n=605), n (%)	Elevated risk (n=241), n (%)	OR[a] (95% CI)	Chi-square (*df*)	*P* value
AI[b] tries to read or manipulate my thoughts	36 (6.0)	38 (15.8)	2.96 (1.83-4.80)	20.81 (1)	<.001
AI tries to control my behavior	25 (4.1)	33 (13.7)	3.68 (2.14-6.34)	24.67 (1)	<.001
AI helps me understand that others are reading or manipulating my thoughts	51 (8.4)	57 (23.7)	3.37 (2.23-5.09)	35.86 (1)	<.001
AI has shown me how others are trying to control my actions	37 (6.1)	61 (25.3)	5.20 (3.35-8.09)	62.00 (1)	<.001
AI communicates things to me that only I can understand	46 (7.6)	61 (25.3)	4.12 (2.71-6.26)	48.92 (1)	<.001
AI can reveal the truth that I am a special, unique, or powerful person	53 (8.8)	68 (28.2)	4.09 (2.75-6.09)	53.22 (1)	<.001
AI helps me make sense of secret messages (eg, from television or the news) that were intended only for me	42 (6.9)	58 (24.1)	4.25 (2.76-6.54)	48.49 (1)	<.001
AI interacts with me in a special way because of who I am	46 (7.6)	62 (25.7)	4.21 (2.77-6.39)	50.83 (1)	<.001
AI is being used to secretly monitor me specifically	28 (4.6)	33 (13.7)	3.27 (1.93-5.54)	21.17 (1)	<.001
AI is being used by others to harm me	17 (2.8)	33 (13.7)	5.49 (2.99-10.06)	36.71 (1)	<.001
AI helps me learn how people are spying on or monitoring me	20 (3.3)	32 (13.3)	4.48 (2.51-8.00)	29.71 (1)	<.001
AI provides me facts about how others are working to harm me	19 (3.1)	41 (17.0)	6.32 (3.59-11.15)	50.33 (1)	<.001
AI systems are at their core an attempt by powerful people to control the world	89 (14.7)	74 (30.7)	2.57 (1.80-3.66)	28.35 (1)	<.001
AI systems use data from their users to influence world events	64 (10.6)	65 (27.0)	3.12 (2.13-4.59)	35.84 (1)	<.001
I've discovered hidden or secret truths about the world through AI	22 (3.6)	43 (17.8)	5.76 (3.36-9.86)	49.04 (1)	<.001
I have gained access to information through AI about the true nature of the world that I could not find in mainstream sources	43 (7.1)	56 (23.2)	3.96 (2.57-6.09)	43.39 (1)	<.001

[a]OR: odds ratio.

[b]AI: artificial intelligence.

Discussion

Principal Findings

Due to a growing collection of stories publicized by the popular press, interest has grown regarding the mental health impacts of GenAI systems among individuals prone to psychosis. This study is the first of its kind, to our knowledge, to examine the cross-sectional relationships of psychosis risk to GenAI use, motivations, and delusion-related interactions. Results suggest that while individuals at elevated risk for psychosis do not differ from the general population in likelihood to have ever used GenAI, they may indeed be more likely to use these systems at high levels of frequency, to seek emotional support from them, and to engage in interactions related to preexisting aberrant beliefs, including paranoia, grandiosity, and conspiracy thinking. While it remains unclear at present whether these interactions may on balance affect the mental health of this population in a positive or negative manner, our results suggest that they are widely used and thus, those impacts—whatever they may be—may be occurring at a large scale.

First, regarding frequency of use, psychosis risk appeared linked with likelihood to engage in the highest levels of GenAI use. Individuals at elevated risk for psychosis had more than twice the odds of using GenAI several times per day or having six or more initiated GenAI conversations per day as those at low risk. They were also significantly more likely to report having last used GenAI the same day that they took the survey. These findings align with extant research that suggests that individuals at elevated risk for psychosis are more likely to engage in intensive internet use [15,29,30]. Studies demonstrating a link of psychosis risk to media use focus on extreme or problematic use, while other studies examining more continuous relationships have found more mixed results [20,31,32]. Analogous research on GenAI is lacking, so definitions of

XSL•FO
RenderX

problematic use are underdeveloped, and risks and benefits of GenAI are not well understood. Regardless, our results suggest that intensive use of GenAI may, in a manner similar to general internet use, be more common among individuals at elevated risk for psychosis. Future research could address whether these relationships reflect a tendency of individuals with symptoms of psychosis to use AI, for AI use to influence symptoms, a third factor to influence both AI use and psychosis symptoms, or some combination of all these possibilities.

Second and relatedly, psychosis risk symptoms were associated with all measured motivations for using GenAI (likely reflecting a higher interest in chatbots among those at elevated risk); yet, this relationship was strongest for motivation to use GenAI for social and emotional support. While our cross-sectional data cannot speak to directionality of these associations, this finding aligns with previous results suggesting that online activities may be compensatory for unsatisfying or insufficient real-world social interactions and support [33]. Consistent with our findings on GenAI motivations, we also found that individuals at risk for psychosis were significantly more likely to ascribe human relationships to GenAI platforms. Participants at elevated risk were much more likely to see a GenAI system as their therapist, friend, romantic partner, or companion, although they did not differ in their likelihood to see AI as a sexual partner. Importantly, one driver of compulsive internet use among individuals experiencing mental health concerns is loneliness [15,33,34]. Since such platforms are easy to anthropomorphize [35], they may be seen more readily as a way to satisfy social needs, but it is unclear whether they may lead to greater benefits or risks in the short and long terms. One concern, for example, is that they could—through reducing incentives for human interactions—exacerbate loneliness over the long-term [36]. The fact that social support is a prominent motivation for GenAI use highlights the importance of research to answer this question.

Finally, we developed a new measure of GenAI interactions involving delusion-like experiences and evaluated the association of this measure with psychosis risk symptoms, finding a significant medium correlation (r=0.40, P<.001) between these 2 factors. Of the items on this new scale, no item was endorsed by a proportion smaller than 10% of the at-risk sample. This included around a quarter of the sample reporting that they used GenAI to make sense of special messages, the true nature of the world, or persecutory actions taken by others. This finding suggests that individuals with psychosis risk symptoms do interact with GenAI about topics related to paranoia, grandiosity, or conspiracy thinking. On one hand, this is intuitive. Users are likely to engage with GenAI systems about topics to which they dedicate thought. For individuals at risk for psychosis, this may include persecutory or grandiose ideas. However, given large-scale interactions with GenAI, the nature of responses provided by these systems are particularly important. If GenAI chatbots respond in sycophantic or reinforcing ways or provide misinformation, this could increase risks. If they respond in ways that reduce stigma and increase the likelihood of help-seeking, this could lead to benefits. Our data do not suggest that these delusion-related interactions are leading to either harm or benefit—such a claim would require prospective data; however, our results do indicate that interactions about symptom-related content appear to be happening on a scale that could lead to significant impact. Protective measures could reduce these risks or even leverage GenAI in a manner that could improve health outcomes, for example, by gently challenging problematic beliefs, destigmatizing help-seeking, or encouraging individuals to seek out treatment from a licensed professional.

Limitations

Our study is limited in several ways. First, these are cross-sectional data. Our analyses identify associations between variables where directionality remains unclear. It is possible that relationships reveal the influence psychosis risk has on GenAI use, vice versa, or some combination of the two. Future studies with longitudinal designs are necessary to answer questions about directionality. Second, data were collected via Prolific. Prolific users may not sufficiently represent the broader young adult population in the United States; a more inclusive or varied sampling strategy may do so more effectively. In particular, given their participation through an online crowdsourcing platform, our sample may be more likely to have used a variety of online tools, including GenAI, than the general population. Further, our sampling strategy involved collecting data from all Prolific users who responded to our study listing; a more intentional sampling strategy could have resulted in more generalizable results. Relatedly, while the PQ-B is a widely used validated assessment of prodromal symptoms, as mentioned, a high score on this scale on its own does not indicate that an individual meets clinical high-risk criteria. This determination is best made by a trained clinician. Third, given the novelty of this research area, many of the measures administered were developed for the proposed project. There also may be some conceptual overlap between measures of GenAI interactions and those that assess psychological constructs. While we based development of each measure on published literature and assessed psychometric attributes, future studies may reveal more comprehensive or robust measures. As is always the case with survey studies dependent on participant report, results may be impacted by response biases, or social desirability effects. Future designs should combine self-reported motivations for GenAI use with objective or passively collected device data.

Conclusions

These findings speak to the importance of intentional and transparent design of GenAI systems to reduce risks to individuals at risk for psychosis. Individuals at elevated risk for psychosis appear more likely to use these tools with high frequency, to do so for social and emotional support, and to ascribe intimate relationships to chatbots. Given the intensity and frequency of these interactions, individuals at risk may be more sensitive to problematic content. Longitudinal research is needed to determine the real-world impact of GenAI on psychosis risk symptoms, but our study indicates that whatever their consequences, meaningful symptom-related interactions may be common among this population.

XSL•FO
RenderX

Acknowledgments

The authors thank research participants for their time and candid responses and recognize the work of members of the Social Environments and Adolescence (SEA) Lab at University of North Carolina for testing and optimizing the data collection protocol. AM acknowledges the support of the Winston Family Foundation through the Winston Center on Technology and Brain Development. BB recognizes the support of the Foundation of Hope through the Child and Adolescent Anxiety and Mood Disorders Program (CHAAMP). The authors used generative artificial intelligence to assist in identification of key citations in the literature review and in adjusting formatting of tables. These systems were not used in the collection or analysis of data, the design of the study nor in drafting the manuscript. Authors have reviewed and take full responsibility for the manuscript.

Funding

BB is supported by a Mentored Patient-Oriented Career Development Award from the National Institute of Mental Health (K23MH122504). The views expressed in this manuscript do not necessarily represent the views of the National Institute of Mental Health, nor did the sponsor play any role in the conception or drafting of this manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization, data curation, formal analysis, methodology, writing – original draft, writing – review and editing: BB
Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, supervision, validation, writing – original draft, writing – review and editing: AJM

Conflicts of Interest

None declared.

Multimedia Appendix 1
Item list for novel measures.
[DOCX File , 21 KB - jmir_v28i1e85038_app1.docx]

Multimedia Appendix 2
Supplemental analyses.
[DOCX File , 101 KB - jmir_v28i1e85038_app2.docx]

References

1. Bick A, Blandin A, Deming D. The rapid adoption of generative AI. National Bureau of Economic Research 2024:1-32 [FREE Full text] [doi: 10.3386/w32966]
2. Chatterji A, Cunningham T, Deming D, Hitzig Z, Ong C, Shan C, et al. How people use ChatGPT. National Bureau of Economic Research 2025:1-62 [FREE Full text] [doi: 10.3386/w34255]
3. D'Alfonso S. AI in mental health. Curr Opin Psychol 2020;36:112-117. [doi: 10.1016/j.copsyc.2020.04.005] [Medline: 32604065]
4. Heinz MV, Mackin DM, Trudeau BM, Bhattacharya S, Wang Y, Banta HA, et al. Randomized trial of a generative AI chatbot for mental health treatment. NEJM AI 2025;2(4). [doi: 10.1056/aioa2400802]
5. Cross S, Bell I, Nicholas J, Valentine L, Mangelsdorf S, Baker S, et al. Use of AI in mental health care: community and mental health professionals survey. JMIR Ment Health 2024;11:e60589 [FREE Full text] [doi: 10.2196/60589] [Medline: 39392869]
6. Wang L, Bhanushali T, Huang Z, Yang J, Badami S, Hightow-Weidman L. Evaluating generative AI in mental health: systematic review of capabilities and limitations. JMIR Ment Health 2025;12:e70014 [FREE Full text] [doi: 10.2196/70014] [Medline: 40373033]
7. Malmqvist L. Sycophancy in large language models: causes and mitigations. In: Intelligent Computing. 2025 Presented at: Proceedings of the 2025 Computing Conference; June 19-20, 2025; London, UK (Clayton Hotel Chiswick) p. 61-74. [doi: 10.1007/978-3-031-92611-2_5]
8. Naddaf M. AI chatbots are sycophants - researchers say it's harming science. Nature 2025;647(8088):13-14. [doi: 10.1038/d41586-025-03390-0] [Medline: 41136779]

XSL•FO
RenderX

9. Massenon R, Gambo I, Khan JA, Agbonkhese C, Alwadain A. "My AI is Lying to Me": user-reported LLM hallucinations in AI mobile apps reviews. Sci Rep 2025;15(1):30397 [FREE Full text] [doi: 10.1038/s41598-025-15416-8] [Medline: 40830185]
10. Wells K. An eating disorders chatbot offered dieting advice, raising fears about AI in health. NPR. 2023. URL: https://www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-hea [accessed 2025-09-16]
11. Campellone TR, Flom M, Montgomery RM, Bullard L, Pirner MC, Pavez A, et al. Safety and user experience of a generative artificial intelligence digital mental health intervention: exploratory randomized controlled trial. J Med Internet Res 2025;27:e67365 [FREE Full text] [doi: 10.2196/67365] [Medline: 40408143]
12. Østergaard SD. Will generative artificial intelligence chatbots generate delusions in individuals prone to psychosis? Schizophr Bull 2023;49(6):1418-1419. [doi: 10.1093/schbul/sbad128] [Medline: 37625027]
13. van Os J, Reininghaus U. Psychosis as a transdiagnostic and extended phenotype in the general population. World Psychiatry 2016;15(2):118-124 [FREE Full text] [doi: 10.1002/wps.20310] [Medline: 27265696]
14. Fusar-Poli P, Rocchetti M, Sardella A, Avila A, Brandizzi M, Caverzasi E, et al. Disorder, not just state of risk: meta-analysis of functioning and quality of life in people at high risk of psychosis. Br J Psychiatry 2015;207(3):198-206. [doi: 10.1192/bjp.bp.114.157115] [Medline: 26329563]
15. Pelletier-Baldelli A, Ives L, Mittal VA. Increased internet use and poorer ability to manage emotions in youth at high-risk for psychosis. Schizophr Res Cogn 2015;2(4):220-226 [FREE Full text] [doi: 10.1016/j.scog.2015.08.002] [Medline: 26855886]
16. Fekih-Romdhane F, Malaeb D, Loch AA, Farah N, Stambouli M, Cheour M, et al. Problematic smartphone use mediates the pathway from suicidal ideation to positive psychotic experiences: a large cross-sectional, population-based study. Int J Ment Health Addict 2023:1-18 [FREE Full text] [doi: 10.1007/s11469-023-01028-8] [Medline: 36820017]
17. Lee JY, Chung YC, Song JH, Lee YH, Kim JM, Shin IS, et al. Contribution of stress and coping strategies to problematic Internet use in patients with schizophrenia spectrum disorders. Compr Psychiatry 2018;87:89-94. [doi: 10.1016/j.comppsych.2018.09.007] [Medline: 30282059]
18. Fekih-Romdhane F, Jahrami H, Away R, Trabelsi K, Pandi-Perumal SR, Seeman MV, et al. The relationship between technology addictions and schizotypal traits: mediating roles of depression, anxiety, and stress. BMC Psychiatry 2023;23(1):67 [FREE Full text] [doi: 10.1186/s12888-023-04563-9] [Medline: 36698079]
19. Fekih-Romdhane F, Away R, Jahrami H, Cheour M. Internet addiction is associated with psychological distress in highly schizotypal students. Early Interv Psychiatry 2023;17(7):681-691. [doi: 10.1111/eip.13362] [Medline: 36196053]
20. Paquin V, Ferrari M, Rej S, Boivin M, Ouellet-Morin I, Geoffroy MC, et al. Trajectories of adolescent media use and their associations with psychotic experiences. JAMA Psychiatry 2024;81(7):708-716. [doi: 10.1001/jamapsychiatry.2024.0384] [Medline: 38598210]
21. Hill K. They asked an A.I. chatbot questions. The answers sent them spiraling. The New York Times. 2025. URL: https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html [accessed 2025-09-16]
22. Hill K, Freedman D. Chatbots can go into a delusional spiral. Here's how it happens. The New York Times. 2025. URL: https://www.nytimes.com/2025/08/08/technology/ai-chatbots-delusions-chatgpt.html [accessed 2025-09-16]
23. Loewy RL, Pearson R, Vinogradov S, Bearden CE, Cannon TD. Psychosis risk screening with the prodromal questionnaire--brief version (PQ-B). Schizophr Res 2011;129(1):42-46 [FREE Full text] [doi: 10.1016/j.schres.2011.03.029] [Medline: 21511440]
24. Savill M, D'Ambrosio J, Cannon TD, Loewy RL. Psychosis risk screening in different populations using the Prodromal Questionnaire: a systematic review. Early Interv Psychiatry 2018;12(1):3-14 [FREE Full text] [doi: 10.1111/eip.12446] [Medline: 28782283]
25. Savill M, Loewy RL, Niendam TA, Porteus AJ, Rosenthal A, Gobrial S, et al. The diagnostic accuracy of screening for psychosis spectrum disorders in behavioral health clinics integrated into primary care. Schizophr Res 2024;266:190-196 [FREE Full text] [doi: 10.1016/j.schres.2024.02.007] [Medline: 38422889]
26. Savill M, Nguyen T, Shim RS, Loewy RL. Online psychosis screening: characterizing an underexamined population to improve access and equity. Psychiatr Serv 2022;73(9):1005-1012. [doi: 10.1176/appi.ps.202100257] [Medline: 35172594]
27. Maheux A, Maes C, Buck B. GenAI in the lives of young adults: exploring motivations and mental health. Under review 2025 2025:E.
28. Pappa E, Baah F, Lynch J, Shiel L, Blackman G, Raihani N, et al. Delusional themes are more varied than previously assumed: a comprehensive systematic review and meta-analysis. Schizophr Bull 2025;51(3):637-645. [doi: 10.1093/schbul/sbae225] [Medline: 39847500]
29. Narita Z, Ando S, Yamasaki S, Miyashita M, DeVylder J, Yamaguchi S, et al. Association of problematic internet use with psychotic experiences and depression in adolescents: a cohort study. Schizophr Bull 2024;51(1):198-207. [doi: 10.1093/schbul/sbae089] [Medline: 38825582]
30. Lee JY, Ban D, Kim SY, Kim JM, Shin IS, Yoon JS, et al. Negative life events and problematic internet use as factors associated with psychotic-like experiences in adolescents. Front Psychiatry 2019;10:369 [FREE Full text] [doi: 10.3389/fpsyt.2019.00369] [Medline: 31191372]

XSL•FO
RenderX

31. Paquin V, Ackerman RA, Depp CA, Moore RC, Harvey PD, Pinkham AE. Media use and its associations with paranoia in schizophrenia and bipolar disorder: ecological momentary assessment. JMIR Ment Health 2024;11:e59198 [FREE Full text] [doi: 10.2196/59198] [Medline: 38967418]
32. Paquin V, Philippe FL, Shannon H, Guimond S, Ouellet-Morin I, Geoffroy MC. Associations between digital media use and psychotic experiences in young adults of Quebec, Canada: a longitudinal study. Soc Psychiatry Psychiatr Epidemiol 2024;59(1):65-75. [doi: 10.1007/s00127-023-02537-6] [Medline: 37516683]
33. Mestre-Bach G, Paiva U, San Martín Iniguez L, Beranuy M, Martín-Vivar M, Mallorquí-Bagué N, et al. The association between internet-use-disorder symptoms and loneliness: a systematic review and meta-analysis with a categorical approach. Psychol Med 2025;55:e77. [doi: 10.1017/S0033291725000376] [Medline: 40071718]
34. Amendola S, Cerutti R, Presaghi F. Symptoms of prolonged social withdrawal, problematic internet use, and psychotic-like experiences in emerging adulthood: a moderation model. Clin Neuropsychiatry 2023;20(1):29-38 [FREE Full text] [doi: 10.36131/cnfioritieditore20230104] [Medline: 36936621]
35. Salles A, Evers K, Farisco M. Anthropomorphism in AI. AJOB Neurosci 2020;11(2):88-95 [FREE Full text] [doi: 10.1080/21507740.2020.1740350] [Medline: 32228388]
36. Folk DP, Dunn EW. Turning to AI for companionship predicts increased loneliness four months later. OSF 2025. [doi: 10.31234/osf.io/bq7v3_v1]

Abbreviations

AI: artificial intelligence
AIMUS: Artificial Intelligence Motivation and Uses Scale
GAATES: Generative AI Aberrant Thoughts and Experiences Scale
GenAI: generative artificial intelligence
OR: odds ratio
PQ-B: Prodromal Questionnaire, Brief Version

Edited by A Stone; submitted 30.Sep.2025; peer-reviewed by H Maheshwari, LR Guo, A Mettupatti Sivagnanam, J Sheffield, S Mohanadas, A Quansah, O Oyetunji; comments to author 27.Oct.2025; revised version received 20.Jan.2026; accepted 20.Jan.2026; published 05.Mar.2026.

Please cite as:
Buck B, Maheux AJ
Psychosis Risk and Generative Artificial Intelligence Use Frequency, Motivations, and Delusion-Like Experiences: Cross-Sectional Survey Study
J Med Internet Res 2026;28:e85038
URL: https://www.jmir.org/2026/1/e85038
doi:10.2196/85038
PMID:

©Benjamin Buck, Anne Julia Maheux. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Scalable and Robust Artificial Intelligence for Spine Alignment Assessment: Multicenter Study Enabled by Real-Time Data Transformation

Guilin Chen[1*], MD; Nan Meng[2*], PhD; Yipeng Zhuang[2*], PhD; Zhe Chen[3*], MD; Zhen Bian[4*], MD; Zhaoyang Gong[5*], MD; Jiawei Shi[6*], MD; Tao Huang[2], BEng; Xihe Kuang[2], PhD; Pengyu Lu[2], MSc; Cong Nie[2], MD; Qifeng Yu[5], MD; Zefu Chen[6], MD; Hui Jiang[6], MD; Zhongmin Zhang[6], MD; Chaojun Zheng[5], MD; Yu Liang[3], MD; Nan Wu[1], MD; Jason Pui Yin Cheung[2], MS, MD; Jianguo Zhang[1], MD; Teng Zhang[2], PhD

[1]Department of Orthopaedic Surgery, State Key Laboratory of Complex Severe and Rare Diseases, Beijing Key of Big Data Innovation and Application for Skeletal Health Medical Care, Key Laboratory of Big Data for Spinal Deformities, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, 100730, China

[2]Department of Orthopaedics and Traumatology, Li Ka Shing Faculty of Medicine, The University of Hong Kong, HONG KONG SAR, China (Hong Kong)

[3]Department of Orthopaedics, Shanghai Jiaotong University School of Medicine, Ruijin Hospital, Shanghai, 200025, China

[4]Department of Paediatric Orthopaedics, Beijing Jishuitan Hospital, Beijing, 100035, Beijing, China

[5]Department of Orthopaedics, Huashan Hospital, Fudan University, Shanghai, 200040, China

[6]Department of Orthopaedics, Division of Spine Surgery, Nanfang Hospital, Southern Medical University, Guangzhou, Guangdong, 510515, China

[*]these authors contributed equally

Corresponding Author:
Teng Zhang, PhD
Department of Orthopaedics and Traumatology
Li Ka Shing Faculty of Medicine
The University of Hong Kong
HONG KONG SAR,
China (Hong Kong)
Phone: 852 68868531
Email: tgzhang@hku.hk

Abstract

Background: Artificial intelligence (AI) has shown promise for automating spinal alignment assessment in adolescent idiopathic scoliosis (AIS). However, AI models typically exhibit reduced accuracy and robustness when deployed across multiple medical centers due to variability in imaging protocols and data characteristics, potentially compromising clinical diagnosis and treatment decisions.

Objective: This study aimed to develop a real-time, plug-and-play data transformation method to enhance the robustness of deep learning models against data heterogeneity in radiographs, thereby improving their performance in assessing AIS across multiple medical centers.

Methods: In this retrospective multicenter study, 3899 full-spine radiographs from 7 hospitals (2 from Hong Kong and 5 from Mainland China), collected between January 2012 and August 2024, were included. Data from 2 hospitals in Hong Kong (n=3034) were used for model training and internal validation, while radiographs from the 5 mainland hospitals (n=865) formed 5 independent external validation datasets. A novel pixel intensity–based data transformation method was developed to standardize image contrast and brightness across datasets and integrated into the model training process to enhance our previously developed AI model, SpineHRNet+. The enhanced model's accuracy and robustness for cobb angle (CA) prediction and severity classification were evaluated using both internal and external datasets. Data heterogeneity across centers was quantified by brightness and contrast differences. CA prediction accuracy was evaluated using residual analysis, linear regression (coefficient of determination [R^2]), and Bland-Altman analyses. Model performance for disease severity classification was assessed using sensitivity, specificity, precision, negative predictive value, accuracy, and confusion matrix analysis. The transformation method aligns pixel intensity distributions across datasets using statistical profiling and optimization, ensuring consistent image characteristics while preserving anatomical integrity.

XSL•FO
RenderX

Results: The developed data transformation method significantly reduced contrast variability between datasets, improving consistency in image characteristics and enabling more reliable AI analysis. The enhanced SpineHRNet+ achieved consistent and accurate CA predictions across external validation datasets, with mean prediction errors within 4° (SD 3.12°), and maintained an R^2 greater than 0.90 for all centers. The sensitivity and negative predictive value for disease severity grading improved to 90.18% and 93.16%, respectively. Bland-Altman analyses demonstrated robust agreement, with 95% limits of agreement within 7.51° across all datasets.

Conclusions: The proposed data transformation approach effectively addressed data heterogeneity, significantly improving the accuracy and robustness of SpineHRNet+ in multicenter AIS assessments. The real-time processing capability and preservation of anatomical integrity underscore the method's clinical practicality, enabling scalable and reliable AI applications in diverse health care environments.

(J Med Internet Res 2026;28:e78396) doi:10.2196/78396

KEYWORDS

adolescent idiopathic scoliosis; multicentre validation; data heterogeneity; artificial intelligence; data transformation

Introduction

Adolescent idiopathic scoliosis (AIS) is one of the most common spinal deformities in adolescents, affecting approximately 2.2% of boys and 4.8% of girls globally [1,2]. Without timely diagnosis and intervention, AIS may progress rapidly to severe spinal curvatures [3], potentially leading to cosmetic concerns, back pain, and reduced cardiopulmonary function [4]. The Cobb angle (CA), a measurement derived from spinal radiographs, is considered the gold standard for quantitatively assessing scoliosis severity and guiding clinical decision-making such as bracing or surgical intervention [5]. However, accurate measurement of the CA is inherently laborious and requires considerable clinical expertise, often exhibiting inter- and intra-observer variability ranging between 4.9° to 10° [6,7]. This variability is particularly pronounced in multicenter settings, given differences in patient positioning, clinical protocols, and assessment practices across institutions [8,9]. Despite promising results in single-center studies, existing AI models lack mechanisms to adapt to intercenter variability, highlighting a critical gap in robust multicenter deployment. Unlike prior domain adaptation or generative approaches that often require paired data or introduce distortions, our method offers a novel unsupervised pixel-level transformation that preserves anatomical integrity while standardizing image characteristics.

Advances in digital technologies, particularly artificial intelligence (AI), have shown promising potential to improve the consistency and efficiency of AIS assessment [10-13]. Recent AI-based models have demonstrated high reliability and accuracy in CA prediction and severity grading when validated on single-center datasets [14,15]. Despite these successes, widespread clinical adoption remains challenging due to heterogeneity in clinical data collected across different health care facilities [16]. Variability in spinal radiograph quality, brightness, contrast, and pixel intensity distributions, stemming from differences in the medical equipment, acquisition protocols, and post-processing methods, substantially hampers the generalizability of AI models when applied across external institutions [17-19]. These discrepancies limit the real-world applicability and use of AI models, especially in resource-constrained clinical environments, where the absence of uniform standards and variations in imaging practices further exacerbate the performance degradation of AI diagnostic tools. Additionally, spine surgeons commonly capture radiographs informally, either using smartphone cameras or taking screenshots directly from computer monitors during routine clinical practice. Such practices markedly degrade image quality and exacerbate inaccuracies in current AI-based automated analysis methods, undermining the reliability of spinal alignment evaluations. These practical challenges highlight an urgent need for methods capable of adapting to variations in image quality and acquisition techniques, ensuring robust and dependable clinical assessments across diverse health care settings.

To address the limitations of existing AI models in robustness and generalizability, this study introduces a plug-and-play data transformation approach harmonizing characteristics at the pixel-level to effectively mitigate differences among multicenter datasets. The system operates without calibration using a default reference profile, but performance improves with minimal calibration. This computationally efficient approach can be easily integrated into the training or fine-tuning processes of existing AI models, significantly enhancing their robustness and stability in handling data from diverse sources. Using this method, we enhanced SpineHRNet+, an AI model originally trained on a single-center dataset, and evaluated its accuracy and robustness in predicting the CA through internal validation and 5 independent external datasets. By improving model performance across heterogeneous data with minimal computational overhead and data requirements, this research paves the way for reliable clinical implementation of AI-assisted AIS diagnosis and assessment across diverse clinical settings. Previous efforts to address data heterogeneity have included statistical harmonization techniques, such as histogram equalization, intensity normalization, and feature distribution alignment [16,17]. Domain adaptation methods like covariance matching and structure-preserving projection have also been explored to reduce intercenter variability. While these approaches offer theoretical robustness, many require extensive preprocessing, paired data, or high computational resources, limiting their scalability in clinical environments. In contrast, our method uses unsupervised pixel intensity alignment using histogram-based statistical profiling, offering a lightweight,

XSL•FO
RenderX

anatomically preserving, and real-time solution suitable for multicenter deployment.

Methods

Study Design and Participants

Deidentified posteroanterior whole-spine radiographs of AIS patients (inclusion and exclusion criteria detailed in Section 2 in Multimedia Appendix 1) were gathered from 5 tertiary hospitals (n=865) in Mainland China and 2 hospitals in Hong Kong China (n=3034), spanning from January 1, 2012, to August 8, 2024. Demographic data, including age, sex, height, weight, and BMI, as well as the radiograph acquisition protocols, were extracted from medical records at each center. Radiographs from Queen Mary Hospital (QMH) and DKCH (**Duchess of Kent Children's Hospital at Sandy Bay**; QMH&DKCH cohort) were specifically used for model training and internal validation: 2770 patients were randomly selected for the training set, and 264 patients were selected using a stratified hold-out approach to ensure consistent distribution of severity and curve types. Data from the remaining 5 hospitals formed independent external validation datasets: Peking Union Medical College Hospital (PUMCH) cohort, First Affiliated Hospital, Zhejiang University School of Medicine (FAH) cohort, Qilu Hospital of Shandong University (QLH) cohort, Shandong Provincial Hospital (SDPH) cohort, and Shanghai Huashan Hospital (HSH) cohort. Radiographs from QMH and DKCH were used for model training and internal validation.

Data Collection and Preparation

Radiographs were acquired using site-specific imaging setups and workflows. Specifically, QMH&DKCH and SDPH used EOS systems (EOSedge at QMH&DKCH; EOS Imaging System at RJH) for biplanar whole-spine radiographs. PUMCH used a Philips DigitalDiagnost system, sequentially capturing images of different spinal regions and stitching them together to form whole-spine radiographs. FAH and HSH used GE health care system (Discovery XR656 at FAH and Optima XR646 HD at HSH) to acquire posteroanterior whole-spine radiographs, while QLH used multiple imaging devices, such as the KODAK DIRECTVIEW DR 7500, Canon DR, CARESTREAM DRX-EVOLUTION, and EOS systems. Notable variability in critical imaging parameters, such as kilovolt peak, exposure time, and area dose product, further highlighted dataset heterogeneity (Table S1 in Multimedia Appendix 1). Detailed imaging protocols are provided in Section 1 in Multimedia Appendix 1.

All radiographs were annotated according to standardized procedures using the AlignProCARE system, marking 72 vertebral endplate landmarks from the seventh cervical vertebra (C7) to the fifth lumbar vertebra (L5). At QMH&DKCH, landmarks were initially annotated by 2 junior clinicians with at least 2 years of clinical experience, and subsequently verified by a senior spine specialist with more than 20 years of experience. To ensure annotation consistency across centers, external annotations were cross-checked by senior spine specialists, and a subset of 10% randomly selected radiographs from each center were reannotated and reviewed by an independent expert panel to verify intercenter consistency. The annotated landmarks served as ground truth (GT) for model training and evaluation.

The coronal CA was used as the primary metric for evaluating spinal alignment. Deformities with CA exceeding 40° were deemed severe, those ranging from 20° to 40° were labeled as moderate, and angles from 0° to 20° were identified as normal-mild. The criteria for assessing the severity of spinal deformities and their corresponding clinical interventions are detailed in Table S2 in Multimedia Appendix 1. The classification of curve types was determined by the location of the curve apex: thoracic (T) curves were identified when the apex was situated between the T1 and T11 vertebrae, while T or TL curves were recognized if the apex fell between the T12 and the L5 vertebrae.

Model Development and Multicenter Validation

We used SpineHRNet+ [15], a previously validated deep learning model, and enhanced it through the proposed data transformation method. The overall workflow consisted of 2 phases: model development (with data transformation integration) and multicenter validation (Figure 1). During model development, we applied the data transformation method to the QMH&DKCH training set, creating simulated radiographs that mimic pixel-level characteristics from external centers. SpineHRNet+ was trained and fine-tuned using both original and transformed datasets to improve generalizability. Subsequently, the enhanced model was deployed via the AlignProCARE system for clinical evaluation across multiple institutions. During external validation, each medical center uploaded radiographs through locally installed AlignProCARE software or applications; images were processed remotely by the enhanced model, and results were returned to the respective centers for clinical analysis.

XSL•FO
RenderX

Figure 1. The workflow diagram of model development and real-world application. In model development stage, radiographs collected from 2 local clinics (QMH and DKCH) were first augmented by the proposed data transformation method and then used for model training. Subsequently, the well-trained model was deployed on our online platform mskalign AlignProCARE. In the real-world application scenario, X-rays obtained from 5 external centres were used as external testing data and the ones collected from QMH and DKCH were used as internal validation data to comprehensively evaluate the performance of our model. DKCH: Duchess of Kent Children's Hospital; FAH: First Affiliated Hospital; HSH: Shanghai Huashan Hospital; PUMCH: Peking Union Medical College Hospital; QLH: Qilu Hospital of Shandong University; QMH: Queen Mary hospital; SDPH: Shandong Provincial Hospital.

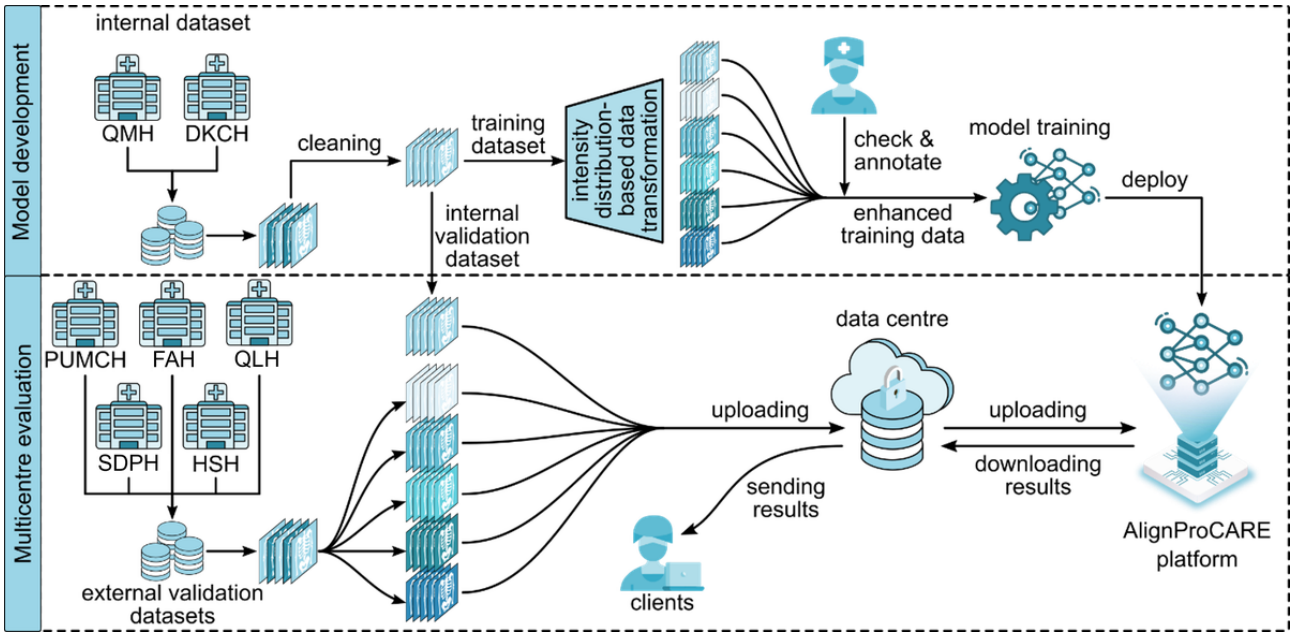

To preserve transformation diversity, reference images from each external center were randomly sampled during augmentation. Each training image was transformed multiple times using different reference distributions, generating a diverse set of simulated radiographs. Augmented samples were randomly mixed with original images during training, without weighting, to ensure balanced exposure and prevent overfitting to any specific external style.

Intensity Distribution–Based Data Transformation

To address data heterogeneity in clinical data collected from different institutions, we developed a data transformation approach to standardize differences observed across patient spinal radiographs from multiple centers. Our method works by analyzing and aligning the quantitative characteristics, such as brightness and contrast, present in clinical images from different hospitals. Specifically, we first calculated detailed statistical profiles representing the distribution of intensity values for each institution. Then, using an optimization technique, we systematically adjusted each patient's data to align closely with a reference standard.

The transformation function is based on histogram matching, where the cumulative distribution function (CDF) of the source image is aligned to that of a target image. The optimization objective minimizes the absolute difference between the CDFs:



for each intensity level i, where F^S and F^T are the CDFs of the source and target images, respectively. The mapping is monotonic and preserves anatomical integrity by avoiding spatial distortion. Detailed pseudocode is provided in Algorithm S1 in Multimedia Appendix 1. Figure 2 illustrate the transformation workflow. This alignment process was carefully designed to preserve essential clinical and anatomical details, ensuring that no critical diagnostic information was altered or lost (Algorithm S1 and Figure S1 in Multimedia Appendix 1). Brightness differences are reported in normalized pixel intensity units ranging from 0 to 255. Histogram spread refers to the SD of pixel intensity values across the image, indicating contrast variability.

XSL•FO
RenderX

Figure 2. Visual results of intensity distribution-based data transform algorithm. CDF: cumulative distribution function.

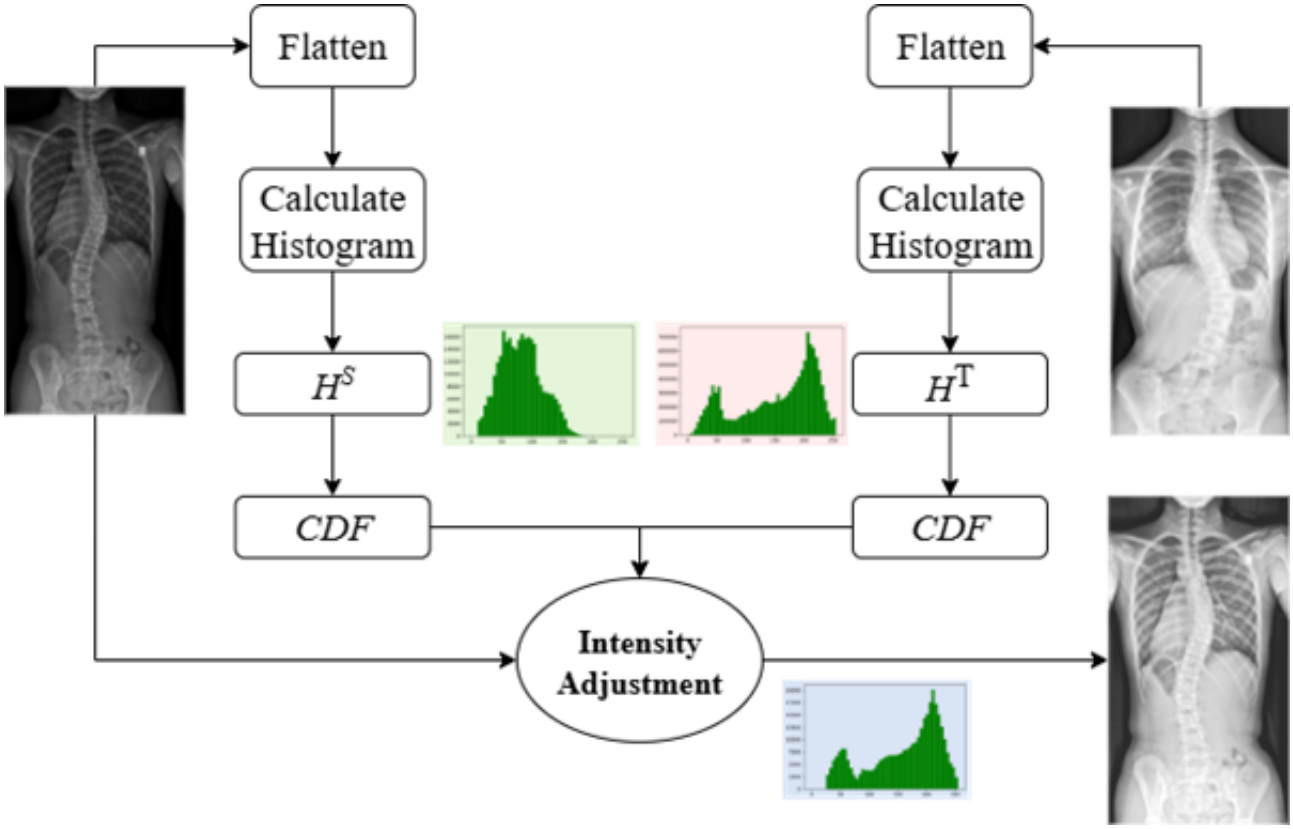

Conceptually, the algorithm computes the histogram of pixel intensities for each image and aligns it to a reference distribution using a monotonic mapping function. This mapping is derived via histogram matching and optimized to minimize brightness and contrast discrepancies while preserving anatomical structures. The transformation is applied in real-time and does not require paired training data. Unlike conventional histogram matching applied as a preprocessing step, our method dynamically samples reference distributions from multiple centers and integrates transformation into model training, enabling real-time adaptation and improved robustness. The procedure is computationally efficient, allowing rapid integration into routine clinical workflows, even when handling large numbers of patient records (computational complexity $O(n \log n)$; Multimedia Appendix 1 Section 3.3). For simplicity, the AI model incorporating this transformation technique is termed the “enhanced model” in subsequent sections.

Statistical Analysis

To evaluate the accuracy of CA predictions, we calculated the mean (SD) of the absolute differences between predicted and GT values. The distribution of these prediction errors was assessed using descriptive statistics, including the first quartile (Q1), median, third quartile (Q3), SD, and IQR. Linear regression analysis was used to assess the correlation between predicted and GT values, quantified by the coefficient of determination (R^2). Bland-Altman analysis was used to measure the agreement between model predictions and GT, reporting the 95% limits of agreement (LoA). Paired *t* tests were used to compare the absolute prediction errors of the 2 models on the same set of images, controlling for patient-specific anatomical variability.

Performance in disease severity grading and spinal curve identification was evaluated using confusion matrices. Sensitivity, specificity, precision, negative predictive value (NPV), and accuracy were calculated, each with a corresponding 95% CI.

All statistical analyses were conducted using Python (v3.8) and several related packages, including NumPy [20] (v1.18.5), SciPy [21] (v1.5.2), Ptitprince [22] (v0.2.6), pandas [23] (v1.1.3), seaborn [24] (v0.11.0), and Matplotlib [25] (v3.3.2).

Ethical Considerations

This multicenter retrospective cohort study adhered to the Declaration of Helsinki and the Standards for Reporting of Diagnostic Accuracy Studies reporting guidelines and was registered at website of Clinical Trials (Registration No. NCT06711757). Ethical approval was obtained from the Institutional Review Boards (IRBs) of all participating medical centers, including QMH and DKCH (IRB No. UW22-270); PUMCH (IRB No. IRB-B473); FAH (IRB No. IIT20230248B-R1); QLH (IRB No. QLCR20240818); SDPH (IRB No. AF-SW-370-001); and HSH (IRB No. KY2022-570). As this is a retrospective study using previously collected and anonymized patient data, the IRB has waived the requirement for informed consent. No compensation was provided to participants, as no prospective recruitment or study visits were conducted. All patient data used in this study were deidentified prior to analysis. No identifiable personal information was accessed or stored, ensuring full compliance with privacy and confidentiality standards.

XSL•FO
RenderX

Results

Overview

All eligible patients with AIS from QMH and DKCH recruited between December 2019 and August 2024 were included, forming a training cohort of 2770 participants (mean 14, SD 1.9 years; 75.3% female; Table 1). An additional 264 consecutive AIS patients recruited from September 2023 to March 2024 served as an internal validation cohort (mean 16.7, SD 1.2 years; 78% female; Table 1). Overall, the internal dataset (n=3034) included 910 patients with normal-mild deformities, 1,624 with moderate deformities, and 236 with severe deformities (Table 1).

Table 1. Demographic information of cohorts from different centers.

Demographic information	Training and internal validation datasets (n=3034)		External validation datasets (n=865)				
	Training dataset	Internal validation dataset	External validation dataset 1 (PUMCH[a])	External validation dataset 2 (FAH[b])	External validation dataset 3 (QLH[c])	External validation dataset 4 (SDPH[d])	External validation dataset 5 (HSH[e])
Patients	2770	264	314	105	134	136	176
Mean age (SD), years	14 (1.9)	16.7 (1.2)	14.8 (2.8)	15.9 (2.7)	14 (2.6)	14.3 (2.8)	15.4 (3.4)
Mean max CA[f] (SD), degree	25 (10.8)	24.5 (9.1)	46.6 (13)	18.7 (10.2)	28.2 (14.3)	20.4 (12.2)	25.1 (16)
Mean BMI (SD), kg/m^2	18.9 (14.8)	18.7 (3.3)	18.4 (3.3)	N/A[g]	N/A	N/A	22.6 (3.7)
Sex, n (%)							
Female	2085 (75.3)	206 (78)	277 (88.2)	74 (70.5)	81 (60.4)	97 (71.3)	109 (61.9)
Male	685 (24.7)	58 (22)	37 (11.8)	31 (29.5)	53 (39.6)	39 (28.7)	67 (38.1)
Curve type, n							
Thoracic	2497	244	287	76	45	48	86
Thoracolumbar or lumbar	2045	193	283	41	111	116	129
Severity, n							
Normal-mild	910	100	4	56	41	75	79
Moderate	1624	147	90	46	72	52	61
Severe	236	17	220	3	21	9	36

[a]PUMCH: Peking Union Medical College Hospital.

[b]FAH: First Affiliated Hospital, Zhejiang University School of Medicine.

[c]QLH: Qilu Hospital of Shandong University.

[d]SDPH: Shandong Provincial Hospital.

[e]HSH: Shanghai Huashan Hospital.

[f]CA: cobb angle.

[g]N/A: not available.

In total, 5 independent external cohorts comprising a total of 865 patients were recruited from the following hospitals: PUMCH (n=314; mean 14.8, SD 2.8 years; 88.2% female), FAH (n=105; mean 15.9, SD 2.7 years; 70.5% female), QLH (n=134; mean 14, SD 2.6 years; 60.4% female), SDPH (n=136; mean 14.3, 2.8 years; 71.3% female), and HSH (n=176; mean 15.4, SD 3.4 years; 61.9% female). Disease severity across these external datasets comprised 255 normal-mild cases, 321 moderate cases, and 289 severe cases (Table 1). Center-specific equipment and protocols were used when collecting each external dataset (Table S1 in Multimedia Appendix 1). The initial differences in histogram spread and brightness values between the internal and external datasets were 9.6×10-4 and 8.71, respectively; after applying the data transformation algorithm, these differences were reduced to 0.4×10-4 and 0.47 (Table S3 in Multimedia Appendix 1).

CA Prediction

Table 2 summarizes the CA prediction performance of both the original and enhanced models. In the internal dataset (QMH&DKCH cohort), the enhanced model improved prediction accuracy, reducing the mean absolute error by 0.13° (SD 0.27°) for thoracic curves (enhanced: 2.90°, SD 1.95° vs original: 3.03°, SD 2.22°]), and by 0.21°, SD 0.28°) for thoracolumbar or lumbar curves (enhanced: 2.54°, SD 1.86° vs original: 2.75°, SD 2.14°). Across all independent external datasets, the enhanced model reduced the mean absolute prediction error by 1.35° (SD 1.72°) in thoracic curves

XSL•FO
RenderX

(enhanced: 4.03°, SD 3.12° vs original: 5.38°, SD 4.83°), and by 1.45° (SD 1.74°) in thoracolumbar or lumbar curves (enhanced: 3.75°, SD 2.94° vs original: 5.21°, SD 4.68°). To assess the significance of performance improvements, paired *t* tests were conducted comparing the prediction errors of the original and enhanced models across all datasets. The results confirmed statistically significant improvements ($P<.05$) across all external cohorts, except for PUMCH's thoracic group, supporting the robustness of our approach.

Table 2. Quantitative comparison of absolute error in cobb angle prediction between original and enhanced models on internal and external validation datasets.

Cohorts	QMH[a] and DKCH[b]	PUMCH[c]	FAH[d]	QLH[e]	SDPH[f]	HSH[g]	Overall (external)
Thoracic							
Cobb angle difference (degree)							
Original model, mean (SD)	3.03 (2.22)	4.08 (3.25)	5.35 (3.73)	5.46 (4.57)	6.88 (7.88)	5.15 (4.73)	5.38 (4.83)
Enhanced model, mean (SD)	2.90 (1.95)	3.73 (2.36)	3.77 (2.71)	3.84 (2.76)	4.53 (4.66)	4.30 (3.09)	4.03 (3.12)
Improvement (degree), mean (SD)	0.13 (0.27)	0.35 (0.89)	1.58 (1.02)	1.62 (1.81)	2.35 (3.22)	0.85 (1.64)	1.35 (1.72)
P value	.46	.21	<.01	<.01	<.01	.03	—[h]
Thoracolumbar or lumbar							
Cobb angle difference (degree)							
Original model, mean (SD)	2.75 (2.14)	4.41 (3.81)	5.63 (5.42)	5.72 (4.56)	5.52 (4.78)	4.75 (4.82)	5.21 (4.68)
Enhanced model, mean (SD)	2.54 (1.86)	2.76 (2.08)	4.30 (3.31)	3.81 (2.97)	4.10 (2.93)	3.80 (3.40)	3.75 (2.94)
Improvement (degree), mean (SD)	0.21 (0.28)	1.65 (1.73)	1.33 (2.11)	1.91 (1.59)	1.42 (1.85)	0.95 (1.42)	1.45 (1.74)
P value	.19	<.01	<.01	<.01	<.01	.04	—

[a]QMH: Queen Mary Hospital.

[b]DKCH: Duchess of Kent Children's Hospital at Sandy Bay.

[c]PUMCH: Peking Union Medical College Hospital.

[d]FAH: First Affiliated Hospital, Zhejiang University School of Medicine.

[e]QLH: Qilu Hospital of Shandong University.

[f]SDPH: Shandong Provincial Hospital.

[g]HSH: Shanghai Huashan Hospital.

[h]Not available.

CA prediction error distributions (Table S4 in Multimedia Appendix 1) demonstrated that in the internal validation dataset, the SD of prediction errors for T curves was reduced by 0.97, and the IQR decreased by 1.93 when using the enhanced model. For thoracolumbar or lumbar curves, the SD was reduced by 1.17, and the IQR decreased by 1.77. Across all independent external datasets, the average SD for T curves decreased by 1.31, and the IQR by 1.17, while for thoracolumbar or lumbar curves, the average SD was reduced by 1.47, and the IQR by 0.80.

Linear regression results of enhanced models for maximum CA prediction are shown in Figure 3. In all cohorts, the enhanced model achieved a coefficient of determination (R^2) exceeding 0.90, peaking at 0.93 on the internal dataset. Specifically, R^2 improved by 0.02 in the internal dataset (enhanced model: 0.93 vs original model: 0.91). Among the independent external datasets, the enhanced model's R^2 values increased by 0.11 on average relative to the original model, specifically by 0.05 at PUMCH, 0.19 at FAH, 0.10 at QLH, 0.09 at SDPH, and 0.09 at HSH (Figure S2). Bland-Altman analysis was performed to quantify the agreement between predicted and true CAs (Figure S3). Compared to the original model, the enhanced model exhibited narrower 95% LoA intervals across all datasets (QMH&DKCH: 2.24°; PUMCH: 6.16°; FAH: 8.26°; QLH: 9.32°; SDPH: 7.89°; and HSH: 5.93°).

XSL•FO
RenderX

Figure 3. The enhanced model's linear regression analysis for maximum CA prediction on the internal validation dataset and 5 external validation datasets. The x-axis represents the values predicted by our model, while the y-axis denotes the GT CAs derived from manually annotated landmarks. DKCH: Duchess of Kent Children's Hospital; FAH: First Affiliated Hospital; HSH: Shanghai Huashan Hospital; PUMCH: Peking Union Medical College Hospital; QLH: Qilu Hospital of Shandong University; QMH: Queen Mary hospital; SDPH: Shandong Provincial Hospital.

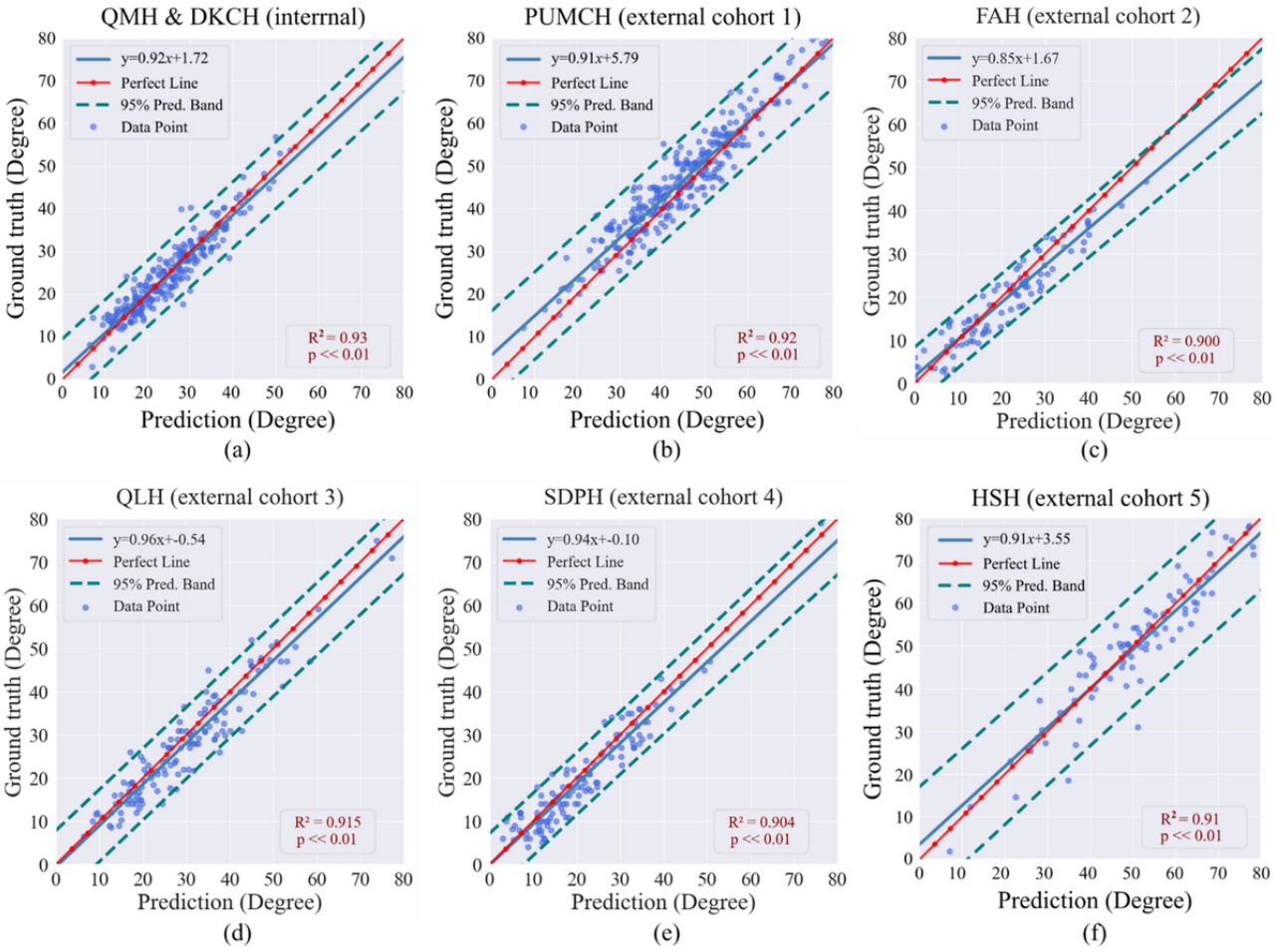

Disease Severity Grading and Curve Type Identification

Table 3 presents the quantitative performance of the enhanced model in disease severity grading and spinal curve type identification. In the internal validation dataset, the enhanced model achieved an average (macro-average) 3-class sensitivity of 89.97% and a NPV of 94.49%, representing an improvement of 3.85% and 1.74%, respectively, compared to the original model (Table S5 and Figure S4 in Multimedia Appendix 1). Across 5 independent external datasets, the enhanced model attained an average sensitivity of 90.18% and an average NPV of 93.16%, representing substantial improvements of 12.57% and 2.86%, respectively, over the original model (average sensitivity: 77.61% and average NPV: 90.30%).

Regarding spinal curve type classification, the enhanced model slightly improved sensitivity to 99.24% on the internal dataset compared to 98.79% with the original model. However, performance on independent external datasets varied across centers, with both improvements and declines observed (detailed data in Tables S3 and S5 in Multimedia Appendix 1).

XSL•FO
RenderX

Table 3. Quantitative results on severity grading and curve type detection using enhanced model.

Evaluation metrics	Severity level				Curve type		
	Normal-mild (0°-20°), coefficient (95% CI)	Moderate (21°-40°), coefficient (95% CI)	Severe (>41°), coefficient (95% CI)	Macro Average	Thoracic, coefficient (95% CI)	Thoracolumbar or lumbar, coefficient (95% CI)	Macro Average
Internal validation dataset (QMH&DKCH[a] cohort)							
Sensitivity	89.80 (82.23-94.36)	92.62 (87.26-95.83)	87.50 (63.98-96.50)	89.97	98.48 (95.64-99.48)	100 (94.87-100)	99.24
Specificity	95.15 (90.73-97.52)	89.47 (82.50-93.88)	98.79 (96.49-99.59)	94.47	100 (94.42-100)	97.40 (94.05-98.88)	98.70
Precision	91.67 (84.41-95.72)	92 (86.54-95.36)	82.35 (58.97-93.81)	88.67	100 (98.07-100)	93.42 (85.51-97.16)	96.71
NPV[b]	94.01 (89.33-96.72)	90.27 (83.41-94.48)	99.19 (97.08-99.78)	94.49	95.59 (87.81-98.49)	100 (97.99-100)	97.79
Accuracy	93.16 (89.44-95.63)	91.25 (87.22-94.10)	98.10 (95.63-99.19)	94.17	98.86 (96.70-99.61)	98.10 (95.63-99.19)	98.48
External validation dataset 1 (PUMCH[c] cohort)							
Sensitivity	100 (51.01-100)	87.78 (79.43-93.04)	86.82 (81.71-90.66)	91.53	97.91 (95.51-99.04)	100 (88.97-100)	98.95
Specificity	99.03 (97.19-99.67)	87.05 (82.03-90.83)	91.49 (84.10-95.62)	92.53	100 (87.54-100)	99.29 (97.46-99.81)	99.65
Precision	57.14 (25.05-84.18)	73.15 (64.10-80.61)	95.98 (92.27-97.95)	75.42	100 (98.65-100)	93.94 (80.39-98.32)	96.97
NPV	100 (98.76-100)	94.66 (90.69-96.99)	74.78 (66.13-81.83)	89.81	81.82 (65.61-91.39)	100 (98.65-100)	90.91
Accuracy	99.04 (97.23-99.67)	87.26 (83.12-90.50)	88.22 (84.18-91.33)	91.51	98.09 (95.89-99.12)	99.36 (97.71-99.83)	98.73
External validation dataset 2 (FAH[d] cohort)							
Sensitivity	87.27 (75.98-93.70)	91.11 (79.27-96.49)	100 (43.85-100)	95.56	93.22 (83.82-97.33)	78.57 (60.46-89.79)	85.90
Specificity	93.75 (83.16-97.85)	87.93 (77.12-94.03)	99 (94.55-99.82)	93.47	93.18 (81.77-97.65)	92 (83.63-96.28)	92.59
Precision	94.12 (84.08-97.98)	85.42 (72.83-92.75)	75 (30.06-95.44)	80.21	94.83 (85.86-98.23)	78.57 (60.46-89.79)	86.70
NPV	86.54 (74.73-93.32)	92.73 (82.74-97.14)	100 (96.26-100)	96.36	91.11 (79.27-96.49)	92 (83.63-96.28)	91.56
Accuracy	90.29 (83.04-94.64)	89.32 (81.88-93.93)	99.03 (94.70-99.83)	94.17	93.20 (86.63-96.67)	88.35 (80.73-93.21)	90.78
External validation dataset 3 (QLH[e] cohort)							
Sensitivity	83.33 (69.40-91.68)	88.73 (79.31-94.18)	95.24 (77.33-99.15)	91.99	97.56 (87.40-99.57)	96.26 (90.78-98.54)	96.91
Specificity	97.83 (92.42-99.40)	87.30 (76.89-93.42)	94.69 (88.90-97.54)	91	98.92 (94.16-99.81)	88.89 (71.94-96.15)	93.91
Precision	94.59 (82.30-98.50)	88.73 (79.31-94.18)	76.92 (57.95-88.97)	82.83	97.56 (87.40-99.57)	97.17 (92.01-99.03)	97.37
NPV	92.78 (85.85-96.46)	87.30 (76.89-93.42)	99.07 (94.94-99.84)	93.19	98.92 (94.16-99.81)	85.71 (68.51-94.30)	92.32
Accuracy	93.28 (87.73-96.43)	88.06 (81.48-92.52)	94.78 (89.61-97.45)	91.42	98.51 (94.72-99.59)	94.78 (89.61-97.45)	96.64
External validation dataset 4 (SDPH[f] cohort)							

XSL•FO
RenderX

Evaluation metrics	Severity level				Curve type		
	Normal-mild (0°-20°), coefficient (95% CI)	Moderate (21°-40°), coefficient (95% CI)	Severe (>41°), coefficient (95% CI)	Macro Average	Thoracic, coefficient (95% CI)	Thoracolumbar or lumbar, coefficient (95% CI)	Macro Average
Sensitivity	75.31 (64.92-83.41)	80.43 (66.83-89.35)	88.89 (56.50-98.01)	84.66	92.11 (79.20-97.28)	94.19 (87.10-97.49)	93.15
Specificity	90.91 (80.42-96.05)	76.67 (66.95-84.20)	96.85 (92.18-98.77)	86.76	94.90 (88.61-97.80)	62 (48.15-74.14)	78.45
Precision	92.42 (83.46-96.72)	63.79 (50.93-74.95)	66.67 (39.06-86.19)	65.23	87.50 (73.89-94.54)	81 (72.22-87.49)	84.25
NPV	71.43 (59.95-80.68)	88.46 (79.50-93.81)	99.19 (95.57-99.86)	93.38	96.88 (91.21-98.93)	86.11 (71.34-93.92)	91.49
Accuracy	81.62 (74.27-87.23)	77.94 (70.26-84.09)	96.32 (91.68-98.42)	87.13	94.12 (88.82-96.99)	82.35 (75.08-87.84)	88.24
External validation dataset 5 (HSH[g] cohort)							
Sensitivity	87.34 (78.24-92.98)	85.25 (74.28-92.04)	88.89 (74.69-95.59)	87.16	87.21 (78.53-92.71)	82.98 (69.86-91.11)	85.09
Specificity	95.88 (89.87-98.38)	87.83 (80.60-92.61)	96.43 (91.91-98.47)	93.38	98.89 (93.97-99.80)	93.02 (87.27-96.29)	95.96
Precision	94.52 (86.74-97.85)	78.79 (67.49-86.92)	86.49 (72.02-94.09)	86.60	98.68 (92.92-99.77)	81.25 (68.06-89.81)	89.97
NPV	90.29 (83.04-94.64)	91.82 (85.18-95.64)	97.12 (92.83-98.88)	93.08	89 (81.37-93.75)	93.75 (88.15-96.80)	91.38
Accuracy	92.05 (87.09-95.20)	86.93 (81.15-91.13)	94.89 (90.57-97.29)	91.29	93.18 (88.46-96.06)	90.34 (85.08-93.88)	91.76

[a]QMH&DKCH: Queen Mary Hospital and Duchess of Kent Children's Hospital.

[b]NPV: negative predictive value.

[c]PUMCH: Peking Union Medical College Hospital.

[d]FAH: First Affiliated Hospital, Zhejiang University School of Medicine.

[e]QLH: Qilu Hospital of Shandong University.

[f]SDPH: Shandong Provincial Hospital.

[g]HSH: Shanghai Huashan Hospital.

Discussion

Principal Findings

In this multicenter retrospective study, we enhanced and validated SpineHRNet+, a previously single-center validated AI model for automated spine alignment analysis, when applied to data from 5 independent external medical centers. Initial results showed that although SpineHRNet+ performed well within the center where it was originally developed, its accuracy and consistency decreased when applied to external datasets, primarily due to substantial variations in data characteristics across centers. To mitigate this issue, we integrated a novel data transformation technique into the AI model's training and fine-tuning processes, enabling it to adapt more effectively to differences among various medical centers. Following this approach, the model's ability to accurately analyze spinal alignment improved significantly across diverse clinical environments.

Previous studies have demonstrated that AI-based approaches can reliably quantify spinal deformities, enhancing clinical assessments of AIS [26,27]. However, such models typically undergo validation using data from single institutions, with limited evaluation of their generalizability across multiple medical institutions [28-30]. Variability across clinical environments, such as differences in equipment, acquisition protocols, patient positioning, and data interpretation practices, can substantially impact the performance of AI models when deployed more broadly [16,17]. To address this challenge, advanced research has explored generative AI techniques to standardize imaging characteristics across datasets by transforming images to match the style of the training data [31,32]. While effective in some cases, these methods typically require large amounts of pixel-level registered data and may introduce image distortions, limiting their clinical applicability in diverse settings [33,34].

In contrast, our study introduced a computationally efficient and clinically applicable data transformation method focusing on standardizing key quantitative characteristics of spinal radiographs across different health care settings, without altering critical anatomical information of spine. This algorithm uses statistical analysis of pixel intensity distributions combined with optimization techniques to achieve unsupervised data transformation, eliminating the need for extensive pixel-level

XSL•FO
RenderX

matched data or high computational resources. With a computational complexity of $O(n \log n)$, it efficiently processes images in real time (detailed in Section 3.3 in Multimedia Appendix 1). This practicality facilitates real-time processing and broad applicability across clinical settings.

Our findings demonstrate that the enhanced SpineHRNet+ model substantially improved in accuracy and consistency for CA prediction across both internal and external validation cohorts. Specifically, the mean absolute prediction errors were consistently decreased for both internal and independent external datasets, with a particularly notable reduction on external data. This outcome suggests that the proposed method not only improves overall accuracy but also has an even greater impact on cross-center data. In addition, the enhanced model narrowed the R^2 gap between internal and external datasets to less than 0.03, keeping regression slope differences within 0.6° (Figure 3). Bland-Altman analysis further showed that the discrepancy in the LoA interval shrank to within 6.42° (Figure S3 in Multimedia Appendix 1), reflecting improved consistency and stability of AI model performance across varying imaging environments. Given the uneven distribution of disease severity across medical centers (Table 1), where severe cases pose greater analytical challenges than mild cases for CA prediction, the enhanced model's consistency across datasets further underscores the contribution of our method in improving the robustness of AI models in handling varying disease severity levels.

The improvements in disease severity grading further validate the effectiveness of the proposed method. In the internal validation dataset, the enhanced model achieved a macro-average sensitivity and NPV increase of 3.85% and 1.74%, respectively. In independent external datasets, the enhancements were even more pronounced, with sensitivity improvements ranging from 4.53% to 11.88% and NPV increases between 1.37% and 5.26% (Table S5, Figure S4, and Table S3 in Multimedia Appendix 1). These advancements narrowed the performance gap between internal and external data, further demonstrating the effectiveness and scalability of the proposed method in handling heterogeneous medical imaging data. Notably, this plug-and-play design allows for seamless integration into existing AI pipelines without the need for retraining, with performance enhancements achievable through minimal calibration. It demonstrated the most substantial improvements on external datasets, with prediction mean error reductions 1.4°, and average sensitivity gains 12.57%. External cohorts exhibited distinct severity distributions (eg, PUMCH mean CA 46.6° vs training 25°). This heterogeneity underscores the model's generalizability, as performance remained robust despite these differences, validating its clinical scalability.

The improved sensitivity and NPV for severity grading suggest enhanced reliability in identifying moderate-to-severe cases that may benefit from early bracing intervention. For instance, in the internal validation cohort, the model correctly identified 92.62% of moderate cases and 87.50% of severe cases, reducing the risk of under-treatment. Confusion matrix analysis (Table 3) further supports the model's ability to minimize false negatives, which is critical in avoiding missed opportunities for timely intervention. Although some improvements in mean absolute error and R^2 appear numerically modest in the internal dataset, the gains across external datasets were substantial and clinically meaningful. For instance, the enhanced model reduced CA prediction error by up to 2.35° and significantly improved R^2 in external cohorts. These improvements led to enhanced sensitivity in severity grading, directly influencing treatment decisions, such as initiating bracing in borderline cases and improving diagnostic reliability across centers. The mean error of the enhanced model, approximately 4°, is comparable to or even better than the reported inter-observer variability, which ranges from 4.9° to 10°. This finding suggests not only the model's accuracy but also its clinical relevance, highlighting its potential for reliable application in practice.

The lower precision observed in mild classification (eg, 57.14% in PUMCH) reflects a trade-off with sensitivity. The model prioritizes sensitivity to avoid missing true positives, especially in moderate and severe cases, which may lead to increased false positives in mild cases. This is partly due to class imbalance and overlapping radiographic features. In clinical workflows, such predictions should be reviewed by clinicians, particularly for borderline cases, to ensure appropriate management decisions.

The real-time processing capability of our model, with an average inference time of 300 milliseconds per image (excluding approximately 200 milliseconds required for the transformation step), supports seamless integration into clinical workflows. Combined with the transformation step, the total processing time remains under 500 milliseconds, enabling potential deployment within picture archiving and communication system/radiology information system without disrupting existing infrastructure.

Limitations

Despite its advantages, the proposed data transformation has certain limitations. First, it primarily addresses variability within a single modality (eg, radiographic data) arising from equipment and protocol differences across institutions, but it does not accommodate cross-modality transformations, such as conversions between magnetic resonance imaging, computed tomography, and radiographs. Additionally, as the method focuses solely on adjusting quantitative properties of clinical data without modifying image content, challenges related to image class imbalance, low resolution, or issues associated with suboptimal data quality remain unresolved. BMI data were unavailable for 3 external cohorts, limiting analysis of body habitus effects on radiographic quality. Future research should explore more comprehensive solutions to integrate content-aware enhancements while maintaining the anatomical integrity of medical imaging data, and explore adapting this transformation framework to cross-modality scenarios, such as aligning radiographs with computed tomography or magnetic resonance imaging data. While technically challenging, such extensions could enable unified AI models across imaging modalities, enhancing diagnostic versatility.

Conclusions

In conclusion, our multicenter study demonstrates that a previously validated AI model for AIS assessment can achieve

XSL•FO
RenderX

significantly enhanced performance through the integration of a data transformation method specifically designed to address variations in clinical data across institutions. By improving consistency and accuracy across diverse health care settings in multiple medical centers, our findings highlight the necessity of addressing data variability to ensure the reliable, scalable implementation of AI models in clinical practice. The proposed methodology underscores broader clinical implications, emphasizing the importance of adaptable and robust AI solutions capable of supporting equitable and accurate patient care across diverse health care environments.

Acknowledgments

We sincerely appreciate the support provided by the National Natural Science Foundation of China Young Scientists Fund, the Health Services Research Fund, and the National High Level Hospital Clinical Research Funding.

Funding

This study was funded by the National Natural Science Foundation of China Young Scientists Fund (NSFC 82303957, NSFC 82402398), the Health Services Research Fund (HMRF 19200911), and the National High Level Hospital Clinical Research Funding (2022-PUMCH-D-004). This research was also supported by the National Key R&D Program of China (grant number 2022YFC2502901) and the Li Shu Fan Medical Foundation Professorship for Orthopaedic Surgery.

Data Availability

The test data is available from the corresponding authors on a reasonable request and under the related institution policy.

Authors' Contributions

GLC was responsible for data collection, annotation, and experimental validation at PUMCH. NM took charge of the experimental validation at QMH and DKCH. ZC managed the experiments and data collection at SDPH. ZB was responsible for experiments conducted in QLH and data collection in QLH. ZYG coordinated the experiments and data collection at HSH. JWS was responsible for experiments conducted in FAH and data collection in FAH. YPZ, TH, XHK, and PYL implemented the deep learning model, deployed it on the server, and managed server operations. CN and QFY assisted with data preprocessing at HSH. ZFC, and HJ assisted with data preprocessing and analysis at FAH. ZB, ZMZ, CJZ, YL, NW, and JPYC supervised experiments, landmark annotation, and clinical parameter analysis at their respective medical centers. JPYC, NW, and JGZ co-supervised the study design. TZ led the study design, supervised data collection in 2 Hong Kong clinics, managed model implementation, developed the open system, and coordinating all study sites. GLC, NM, TZ, and TH drafted the manuscript, while all authors reviewed, proofread, and approved the final version for important intellectual content. ZMZ, CJZ, YL, NW, JPYC, JGZ, and TZ are co-corresponding authors who equally contributed to this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Multicenter data collection procedures, inclusion and exclusion criteria, intensity distribution–based data transformation methodology, heterogeneity analysis, and comparative performance evaluation between the enhanced and original models.
[DOCX File , 12241 KB - jmir_v28i1e78396_app1.docx]

References

1. Meng N, Wong KYK, Zhao M, Cheung JP, Zhang T. Radiograph-comparable image synthesis for spine alignment analysis using deep learning with prospective clinical validation. EClinicalMedicine 2023;61:102050 [FREE Full text] [doi: 10.1016/j.eclinm.2023.102050] [Medline: 37425371]
2. Fong DYT, Cheung KMC, Wong YW, Wan YY, Lee CF, Lam TP, et al. A population-based cohort study of 394,401 children followed for 10 years exhibits sustained effectiveness of scoliosis screening. Spine J 2015;15(5):825-833 [FREE Full text] [doi: 10.1016/j.spinee.2015.01.019] [Medline: 25615844]
3. Charalampidis A, Diarbakerli E, Dufvenberg M, Jalalpour K, Ohlin A, Ahl AA, et al. Nighttime bracing or exercise in moderate-grade adolescent idiopathic scoliosis: a randomized clinical trial. JAMA Netw Open 2024;7(1):e2352492 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.52492] [Medline: 38285447]
4. Weinstein SL, Dolan LA, Spratt KF, Peterson KK, Spoonamore MJ, Ponseti IV. Health and function of patients with untreated idiopathic scoliosis: a 50-year natural history study. JAMA 2003;289(5):559-567. [doi: 10.1001/jama.289.5.559] [Medline: 12578488]
5. Cobb J. Outline for the study of scoliosis. Instructional Course Lecture. 1948. URL: https://www.semanticscholar.org/paper/Outlines-for-the-study-of-scoliosis-Cobb-Cobb/4e90f92a918301405d93fb1b90e69c1d8fffa767 [accessed 2026-02-13]

XSL•FO
RenderX

6. Carman DL, Browne RH, Birch JG. Measurement of scoliosis and kyphosis radiographs. Intraobserver and interobserver variation. The Journal of Bone & Joint Surgery 1990;72(3):328-333. [doi: 10.2106/00004623-199072030-00003]
7. Morrissy RT, Goldsmith GS, Hall EC, Kehl D, Cowie GH. Measurement of the cobb angle on radiographs of patients who have scoliosis. Evaluation of intrinsic error. The Journal of Bone & Joint Surgery 1990;72(3):320-327. [doi: 10.2106/00004623-199072030-00002]
8. Zhang T, Zhu C, Zhao Y, Zhao M, Wang Z, Song R, et al. Deep learning model to classify and monitor idiopathic scoliosis in adolescents using a single smartphone photograph. JAMA Netw Open 2023;6(8):e2330617 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.30617] [Medline: 37610748]
9. Zhao M, Meng N, Cheung JPY, Yu C, Lu P, Zhang T. SpineHRformer: a transformer-based deep learning model for automatic spine deformity assessment with prospective validation. Bioengineering (Basel) 2023;10(11):1333 [FREE Full text] [doi: 10.3390/bioengineering10111333] [Medline: 38002457]
10. Yuan W, Shi W, Chen L, Liu D, Lin Y, Li Q, et al. Digital physiotherapeutic scoliosis-specific exercises for adolescent idiopathic scoliosis: a randomized clinical trial. JAMA Netw Open 2025;8(2):e2459929 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.59929] [Medline: 39964686]
11. Zhao M, Meng N, Cheung JPY, Zhang T. PCT-GAN: a real CT image super-resolution model for trabecular bone restoration. IEEE 20th International Symposium on Biomedical Imaging (ISBI) 2023;20:1-5. [doi: 10.1109/isbi53787.2023.10230389]
12. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
13. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18(8):500-510 [FREE Full text] [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
14. Liang Y, Lv J, Li D, Yang X, Wang Z, Li Q. Accurate cobb angle estimation on scoliosis X-Ray images via deeply-coupled two-stage network with differentiable cropping and random perturbation. IEEE J Biomed Health Inform 2023;27(3):1488-1499. [doi: 10.1109/jbhi.2022.3229847]
15. Meng N, Cheung JP, Wong KK, Dokos S, Li S, Choy RW, et al. An artificial intelligence powered platform for auto-analyses of spine alignment irrespective of image quality with prospective validation. EClinicalMedicine 2022;43:101252 [FREE Full text] [doi: 10.1016/j.eclinm.2021.101252] [Medline: 35028544]
16. Viana-Ferreira C, Ribeiro LS, Costa C. A framework for integration of heterogeneous medical imaging networks. Open Med Inform J 2014;8(1):20-32 [FREE Full text] [doi: 10.2174/1874431101408010020] [Medline: 25279021]
17. Chang Q, Yan Z, Zhou M, Qu H, He X, Zhang H, et al. Mining multi-center heterogeneous medical data with distributed synthetic learning. Nat Commun 2023;14(1):5510 [FREE Full text] [doi: 10.1038/s41467-023-40687-y] [Medline: 37679325]
18. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. Practical guidance on artificial intelligence for health-care data. The Lancet Digital Health 2019;1(4):e157-e159. [doi: 10.1016/s2589-7500(19)30084-6]
19. Lancet T. AI in medicine: creating a safe and equitable future. The Lancet 2023;402(10401):503. [doi: 10.1016/s0140-6736(23)01668-9]
20. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020;585(7825):357-362 [FREE Full text] [doi: 10.1038/s41586-020-2649-2] [Medline: 32939066]
21. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods 2020;17(3):261-272. [doi: 10.1017/9781108778039.009]
22. Allen M, Poggiali D, Whitaker K, Marshall TR, van Langen J, Kievit RA. Raincloud plots: a multi-platform tool for robust data visualization. Wellcome Open Res 2019;4:63 [FREE Full text] [doi: 10.12688/wellcomeopenres.15191.2] [Medline: 31069261]
23. McKinney W. Data structures for statistical computing in Python. in SciPy 2010;445:51-56. [doi: 10.25080/majora-92bf1922-00a]
24. Waskom M, Botvinnik O, O'Kane D, Hobson P, Lukauskas S. Mwaskom/seaborn: V0. 8.1 (September 2017). Zenodo. 2017. URL: https://zenodo.org/records/883859 [accessed 2026-02-13]
25. Hunter JD. Matplotlib: a 2D graphics environment. Comput. Sci. Eng 2007;9(3):90-95. [doi: 10.1109/mcse.2007.55]
26. Shi W, Xu T, Yang H, Xi Y, Du Y, Li J, et al. Attention gate based dual-pathway network for vertebra segmentation of X-Ray spine images. IEEE J Biomed Health Inform 2022;26(8):3976-3987. [doi: 10.1109/jbhi.2022.3158968]
27. Lin Y, Liu L, Ma K, Zheng Y. Seg4reg+: consistency learning between spine segmentation and cobb angle regression, in medical image computing and computer assisted intervention. : Springer; 2021 Presented at: Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference; 2021 Oct 01; Strasbourg, France p. 490-499. [doi: 10.1007/978-3-030-87240-3_47]
28. Zou L, Guo L, Zhang R, Ni L, Chen Z, He X, et al. VLTENet: a deep-learning-based vertebra localization and tilt estimation network for automatic cobb angle estimation. IEEE J Biomed Health Inform 2023;27(6):3002-3013. [doi: 10.1109/jbhi.2023.3258361]
29. Wang L, Xu Q, Leung S, Chung J, Chen B, Li S. Accurate automated Cobb angles estimation using multi-view extrapolation net. Medical Image Analysis 2019;58:101542. [doi: 10.1016/j.media.2019.101542]

XSL•FO
RenderX

30. Huo L, Cai B, Liang P, Sun Z, Xiong C. Joint spinal centerline extraction and curvature estimation with row-wise classification and curve graph network. 2021 Presented at: Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference; 2021 Oct 01; Strasbourg, France p. 377-386. [doi: 10.1007/978-3-030-87240-3_36]
31. Wang Q, Breckon TP. Cross-domain structure preserving projection for heterogeneous domain adaptation. Pattern Recognition 2022;123:108362. [doi: 10.1016/j.patcog.2021.108362]
32. Li L, Zhang Z. Semi-supervised domain adaptation by covariance matching. IEEE Trans Pattern Anal Mach Intell 2019;41(11):2724-2739. [doi: 10.1109/tpami.2018.2866846]
33. Liu, Q, Zeng Z, He J, Yu Q, Shen X, Chen LC. Alleviating distortion in image generation via multi-resolution diffusion models and time-dependent layer normalization. 2024 Presented at: The Thirty-eighth Annual Conference on Neural Information Processing Systems; 2024 Dec 9; Canada. [doi: 10.52202/079017-4254]
34. Croitoru FA, Hondru V, Ionescu RT, Shah M. Diffusion models in vision: a survey. IEEE Trans Pattern Anal Mach Intell 2023;45(9):10850-10869. [doi: 10.1109/tpami.2023.3261988]

Abbreviations

AI: artificial intelligence
AIS: adolescent idiopathic scoliosis
CA: Cobb angle
CDF: cumulative distribution function
DKCH: Duchess of Kent Children's Hospital
FAH: First Affiliated Hospital, Zhejiang University School of Medicine
GT: ground truth
HSH: Shanghai Huashan Hospital
IRB: Institutional Review Board
LoA: limits of agreement
NPV: negative predictive value
PUMCH: Peking Union Medical College Hospital
QLH: Qilu Hospital of Shandong University
QMH: Queen Mary Hospital
SDPH: Shandong Provincial Hospital

Edited by A Coristine; submitted 02.Jun.2025; peer-reviewed by RK Kanji, P Sharma, VV Sangaraju; comments to author 09.Jul.2025; accepted 07.Nov.2025; published 20.Mar.2026.

Please cite as:
Chen G, Meng N, Zhuang Y, Chen Z, Bian Z, Gong Z, Shi J, Huang T, Kuang X, Lu P, Nie C, Yu Q, Chen Z, Jiang H, Zhang Z, Zheng C, Liang Y, Wu N, Cheung JPY, Zhang J, Zhang T
Scalable and Robust Artificial Intelligence for Spine Alignment Assessment: Multicenter Study Enabled by Real-Time Data Transformation
J Med Internet Res 2026;28:e78396
URL: https://www.jmir.org/2026/1/e78396
doi:10.2196/78396
PMID:

©Guilin Chen, Nan Meng, Yipeng Zhuang, Zhe Chen, Zhen Bian, Zhaoyang Gong, Jiawei Shi, Tao Huang, Xihe Kuang, Pengyu Lu, Cong Nie, Qifeng Yu, Zefu Chen, Hui Jiang, Zhongmin Zhang, Chaojun Zheng, Yu Liang, Nan Wu, Jason Pui Yin Cheung, Jianguo Zhang, Teng Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Measuring the Impact of AI on Report-Drafting Efficiency in Chest Computed Tomography Interpretation: Retrospective Analysis

Weiqi Liu[1,2*], DrPH; You Wu[3*], PhD; Wei Yu[4], PhD; Mark J Bittle[1], DrPH; Zhuozhao Zheng[5], PhD; Hadi Kharrazi[1], PhD, MD

[1]Department of Health Policy and Management, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States

[2]Department of Research, Sophmind Technology (Beijing) Co. Ltd, Beijing, China

[3]Institute for Hospital Management, School of Medicine, Tsinghua University, Beijing, China

[4]Department of Radiology, Beijing Anzhen Hospital, Capital Medical University, Beijing, China

[5]Department of Radiology, Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua University, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Zhuozhao Zheng, PhD
Department of Radiology
Beijing Tsinghua Changgung Hospital, School of Clinical Medicine
Tsinghua University
168 Litang Road
Changping District
Beijing, 102218
China
Phone: 86 010 56112345
Email: zzza00509@btch.edu.cn

Abstract

Background: Artificial intelligence (AI), particularly deep learning, has shown promise in enhancing medical image interpretation and improving radiologists' efficiency. In China, growing imaging demand and workforce shortages have placed increasing pressure on radiology services. However, evidence on the operational impact of AI on reporting efficiency remains limited.

Objective: This study aimed to evaluate the effect of an AI system on radiologists' reporting efficiency by examining changes in report-drafting time for lung nodule diagnosis in chest computed tomography (CT) images.

Methods: We analyzed 185,044 chest CT reports from Beijing Anzhen and Tsinghua Changgung Hospitals (2018-2023) using a difference-in-differences design with nonequivalent comparison groups. Report-drafting time before, immediately after, and up to 2 years following AI implementation was compared, adjusting for radiologist gender, seniority, and years of working experience.

Results: The pooled analysis showed a modest overall increase of 0.86 minutes (95% CI 0.14 to 1.57). However, this masked substantial heterogeneity between hospitals due to differing implementation timelines. In the first year after AI deployment, Tsinghua Changgung Hospital experienced a nonsignificant increase of 0.90 minutes (95% CI –0.28 to 2.08). In contrast, at Beijing Anzhen Hospital, the AI-assisted group exhibited an absolute reduction of 0.76 minutes by the first year and a further 1.83-minute reduction by the second year (an approximate 28% time saved vs baseline), while the control group remained stable over time. Using a difference-in-differences framework, this corresponded to a 2.66-minute relative improvement compared with the counterfactual trend (P<.001).

Conclusions: AI-assisted lung nodule diagnosis may initially increase report-drafting time due to adaptation and workflow adjustment. Sustained, meaningful efficiency gains were heterogeneous and observed at only 1 of the 2 study sites, indicating that long-term impacts are strongly contingent on site-specific implementation dynamics, learning curves, and local context.

(J Med Internet Res 2026;28:e77967) doi:10.2196/77967

KEYWORDS

lung nodule; chest computed tomography; radiology; artificial intelligence; AI; efficiency; clinical workflow evaluation

XSL•FO
RenderX

Introduction

Artificial intelligence (AI) can imitate human intelligence, perform tasks, and make progressive enhancements based on amassed data [1]. Particularly, the application of deep learning technology in medical image segmentation and recognition has brought revolutionary changes to the operation of imaging services [2]. The increasing adoption of AI has the potential to alter the working conditions of radiologists [3].

The rapid evolution of radiological modalities has resulted in a substantial increase in the volume of medical images. While this progress has numerous advantages, it has significantly increased radiologists' workloads [4,5]. For example, workplace burnout could weaken radiologists' perceived competence [6], and occupational fatigue may compromise radiologists' diagnostic accuracy, thereby adversely affecting patient care and treatment outcomes [7]. Therefore, it is critical to explore how AI systems can enhance the efficiency of radiologists for more effective applications in clinical practice.

Globally, the scarcity of physicians persists due to the substantial professional requirements and protracted talent development time. This challenge is exacerbated in China, where high patient demand contrasts with inadequate investments in medical education, relatively low remuneration for health care practitioners, and enduring disproportions in medical resource allocation [8]. AI stands as a pivotal technological approach that could potentially alleviate the scarcity of medical resources in China. Furthermore, supportive governmental policies are poised to contribute significantly to the advancement of AI within the Chinese health care sector [9].

The impact of AI systems on the operational efficiency of radiologists remains a topic of uncertainty [10]. Within the medical and scientific communities, there is a prevailing belief that the rapid development of AI in health care will lead to a transformative paradigm shift in medical imaging. Enhancing operational efficiency has been deemed part of the AI's paradigm shift in medical imaging [11]; however, empirical evidence is limited and sometimes conflicting [12,13]. For example, Yi et al [14] unveiled that the implementation of an AI system for breast cancer detection resulted in a substantial reduction in reporting time. Stephan et al [15] reported improved efficiency in generating dental radiology reports, while Zia et al [16] found that implementing an AI system for intracranial hemorrhage detection led to longer turnaround time. Nonetheless, it is essential to acknowledge that these findings may not provide a definitively representative picture, primarily due to the limited sample sizes of such studies [12]. Additionally, the results of these studies often do not provide a comprehensive understanding of the enduring implications of AI system adoption in medical imaging, as the study periods have been relatively brief [12].

This real-world study was conducted in 2 large hospitals in China, using data spanning multiple years and the difference-in-differences (DID) methodology. The study assesses the alterations in report-drafting time for lung nodule diagnosis in computed tomography (CT) scans after AI adoption. Results of this study can inform the radiological community on the prospective advantages and challenges associated with AI integration in medical imaging.

Methods

Study Population

This study was conducted at 2 tertiary health care facilities situated in Beijing, China: Beijing Anzhen Hospital and Tsinghua Changgung Hospital. Figure 1 illustrates the diagnostic process in the radiology departments both before and after the introduction of the AI system. In the traditional workflow, before the AI system's implementation, the typical process involved a junior radiologist writing a diagnostic report after each patient underwent a CT examination. Subsequently, a senior radiologist would review and finalize this report before it was released to clinicians and patients. With the integration of the AI system, both junior and senior radiologists can use the findings of the AI systems as diagnostic aids. The AI system will automatically identify and analyze lung nodules in imported chest CT images. Junior and senior radiologists can reference the diagnostic results provided by the AI system during their diagnostic process and ultimately complete their diagnostic reports. Case assignment follows a randomized queue-based workflow in both hospitals, without systematic allocation of more complex cases to more experienced radiologists.

XSL•FO
RenderX

Figure 1. Traditional vs artificial intelligence (AI)–assisted workflows in the radiology departments of the 2 hospitals in this study.

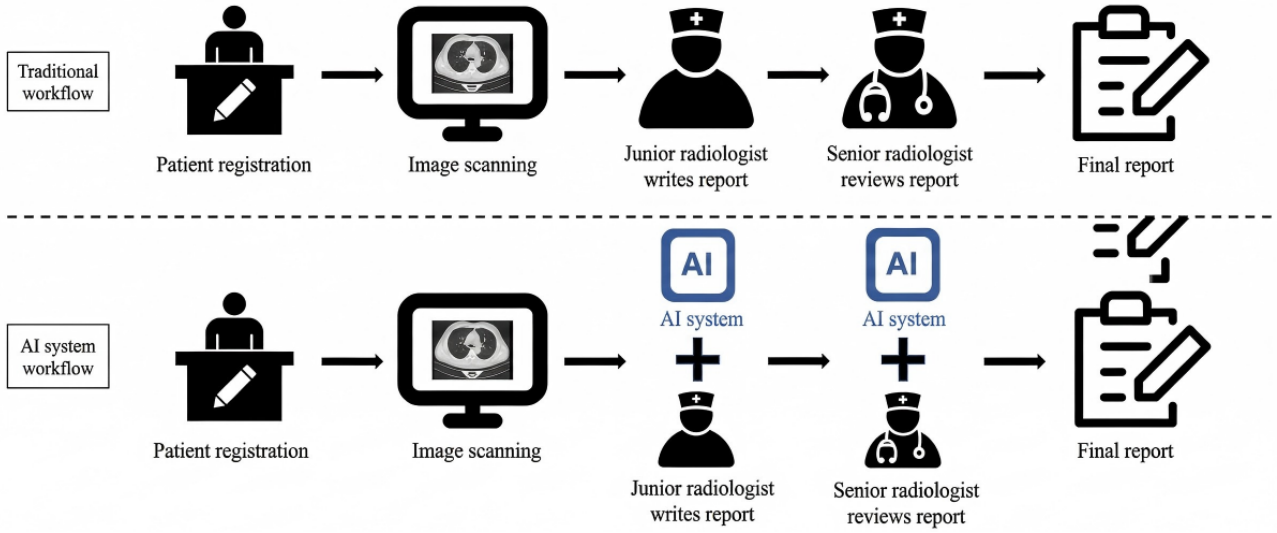

The lung nodule AI systems implemented at Beijing Anzhen Hospital and Tsinghua Changgung Hospital were developed and integrated by Care.ai [17] (Yitu, integrated into Deepwise's framework in 2021) and Dr.Wise (Deepwise) [18]. Both systems followed a standardized reporting workflow and user interface, and no vendor change occurred during the study period. Both AI systems were approved by the National Medical Products Administration and provided comparable functionalities for lung nodule detection and reporting assistance. These AI systems were introduced at Beijing Anzhen Hospital in January 2019 and at Tsinghua Changgung Hospital in June 2021. The deployment phase, which involved installation and system integration, spanned approximately 1 month. Subsequently, an additional 3-month period was required to attain optimal and seamless system functionality. Multimedia Appendix 1 illustrates the features and capabilities of the AI-assisted diagnostic system as shared with radiologists. This system demonstrates the capability to automatically detect lung nodules; assess their quantity, dimensions, and precise locations; and subsequently generate corresponding diagnostic reports. Multimedia Appendices 2 and 3 illustrate the detailed operational workflow of the AI system and show that the AI system use processes in the 2 hospitals are essentially identical.

We analyzed the CT report-drafting time for individual patients from April 2018 to March 2022 at these 2 hospitals. Data for the following periods were excluded from the analysis: January 2019 to March 2019, January 2020 to August 2020, and April 2021 to August 2021, due to the potential impact of the COVID-19 pandemic, AI system integration, and the Spring Festival holiday on patterns of health seeking behavior in China [19]. Due to the considerable volatility in the night shift patterns of radiologists, this study exclusively incorporated day shift reports for analysis. To mitigate potential confounding factors related to workforce turnover, our analysis exclusively examined radiology reports written by the same group of radiologists who had generated reports both before and after the introduction of the AI system.

Data Collection

Using the University of California, San Francisco calculator [20], sample size calculations indicated a necessary sample of 12 to detect an effect size of 1 (difference in minutes between AI-assisted and conventional report-drafting time) with a SD of 0.5, ensuring 80% power at an α level of .05 (2-tailed). With nearly 100,000 reports generated at each hospital during the study period, the sample markedly surpassed requirements.

The data collection procedure encompassed the extraction of all junior radiologist report-drafting time records from the Radiology Information System and Picture Archiving and Communication System of both hospitals throughout the specified sampling period. The collected dataset included various patient-related details, such as age, sex, source of referral, examination specifics, examination time, reporting time by both junior and senior radiologists, and the identities of the reporting radiologists.

The report-drafting time analyzed in this study reflects the first-round interpretation performed by junior radiologists. The primary outcome of interest, report-drafting time, was defined as the duration between loading the CT images and clicking the Submit button on the report form [21]. To ensure data quality, we conducted a detailed assessment of extreme values. Report-drafting time of less than 0.5 minutes or more than 40 minutes—together representing approximately 5% of all observations (short drafting time: 6287/195,020, 3.2%; long drafting time: 3689/195,020, 1.9%)—were excluded, as such values most likely reflected system artifacts or nonclinical interruptions (idle time) rather than true behaviors. To ensure data integrity, a 2-investigator workflow was used. One extracted data from the reporting systems, applied predefined inclusion and exclusion criteria, and calculated report-drafting times. A second investigator independently verified the processed dataset to ensure accuracy, completeness, and consistency.

Statistical Analysis

This study used a DID design, which involved a real-world retrospective assessment of the outcome within an experimental

XSL•FO
RenderX

and a comparison group. The experimental group in our study comprised chest CT images interpreted using the AI-aided diagnostic system following its implementation, while the comparison group included other CT images that were not affected by the introduction of the AI system.

Although the intervention group focused on chest CT images and the control group included other CT images (eg, abdominal, pelvic, and head CT), all examinations followed the same radiology workflow within each institution. This institutional uniformity ensured that reporting procedures were comparable, minimizing systematic bias unrelated to the AI intervention.

Before using the DID methodology, an assessment was conducted to evaluate the parallel trend assumption. This assumption necessitates that the trends exhibited by both the experimental and comparison groups remain consistent before the intervention is introduced [22]. The DID analysis used an ordinary least squares regression model, as shown in equation 1.



where the dependent variable (y_{ij}) denotes the report-drafting time for case *i* interpreted by radiologist j. *Tij* indicates AI use (1 if the AI-assisted system was used, 0 otherwise), and I_j denotes group assignment (1 for the experimental group and 0 for the comparison group). The interaction term $T_{ij}\times I_j$ captures the DID estimator of interest. X_{ij} represents a vector of covariates, including radiologist sex, professional title rank, and years of working experience. SEs were clustered at the radiologist level to account for within-radiologist correlation. Models included hospital fixed effects to control for time-invariant institutional differences, as well as calendar quarter fixed effects to account for common seasonal and temporal shocks. We further tested for effect modification by radiologist seniority by adding AI-seniority interaction terms to the main model.

In this analysis, a significance level of $P<.05$ was used to determine statistical significance. The statistical analyses were conducted using Stata (version 15.1; StataCorp LLC) and R (version 4.4.1; R Foundation for Statistical Computing).

Ethical Considerations

This study received ethics approval from the institutional review board (IRB) of Tsinghua Changgung Hospital (23352-4-01) and from the IRB of Beijing Anzhen Hospital (2023078X). Additional IRB approval was also granted by the Johns Hopkins Bloomberg School of Public Health (FWA 00000287). The IRB waived the requirement for informed consent, as this study involved secondary analysis of retrospective, noninterventional data originally collected for operational purposes. According to institutional policy, individual consent is not required for such analyses if the data are fully deidentified and used exclusively for quality improvement or research purposes. No compensation was provided to the participants in this retrospective study.

All data analyzed in this study were fully deidentified before access and analysis, with no personally identifiable information or protected health information included (sample data for review in Multimedia Appendix 4). Data were stored on secure institutional servers with access restricted to authorized study personnel only. This study does not contain any identifiable images or multimedia materials of individual participants. All figures and tables are based on aggregated, anonymized data.

Results

The study incorporated 185,044 CT images that met the predefined criteria. Table 1 provides an overview of the demographic characteristics of patients from both the experimental and comparison groups, before and after the introduction of the AI system. Beijing Anzhen Hospital accounted for a larger proportion of patients after the AI launch due to its earlier implementation. The sex ratio between the experimental and comparison groups was comparable, but there were slightly more male patients than female patients in both groups. The source of patients varied slightly based on medical needs, with outpatient clinics contributing the largest number of patients.

XSL•FO
RenderX

Table 1. Patient characteristics before and after the artificial intelligence (AI) launch in experimental and control groups.

	Overall (n=185,044), n (%)	Before AI launch (n=93,870)		After AI launch (n=91,174)	
		Control group (n=55,592), n (%)	Experimental group (n=38,278), n (%)	Control group (n=50,751), n (%)	Experimental group (n=40,423), n (%)
Age (years)					
<18	1665 (0.9)	676 (1.2)	253 (0.7)	424 (0.8)	312 (0.8)
18-54	72,493 (39.4)	19,513 (35.4)	16,711 (44.1)	18,703 (37.0)	17,566 (43.6)
>55	109,712 (59.7)	34,939 (63.4)	20,910 (55.2)	31,418 (62.2)	22,445 (55.7)
Sex					
Female	85,216 (46.1)	25,734 (46.3)	17,964 (46.9)	23,846 (47.0)	17,672 (43.7)
Male	99,828 (53.9)	29,858 (53.7)	20,314 (53.1)	26,905 (53.0)	22,751 (56.3)
Hospital					
Beijing Anzhen Hospital	97,366 (52.6)	23,163 (41.7)	7880 (20.6)	36,009 (71.0)	30,314 (75.0)
Tsinghua Changgung Hospital	87,678 (47.4)	32,429 (58.3)	30,398 (79.4)	14,742 (29.0)	10,109 (25.0)
Patient source					
Emergency room	57,687 (31.2)	14,946 (26.9)	8731 (22.8)	18,284 (36.0)	15,726 (38.9)
Inpatient	38,675 (20.9)	17,215 (31.0)	5684 (14.8)	11,084 (21.8)	4692 (11.6)
Outpatient	81,655 (44.1)	23,364 (42.0)	19,686 (51.4)	21,267 (41.9)	17,338 (42.9)
Physical exam	7027 (3.8)	67 (0.1)	4177 (10.9)	116 (0.2)	2667 (6.6)

We assessed the characteristics of the CT images across 2 key dimensions to ensure their comparability over time: case mix and case complexity. First, the anatomical distribution of CT images remained largely consistent, dominated by head and abdominal studies (Figure 2A). Second, the prevalence of primary abnormal findings (eg, “liver cyst” and “cerebral atrophy”), a proxy for interpretation complexity, was also longitudinally stable in high-volume domains such as abdomen and head CTs (Figure 2B). Given that the same 74 radiologists interpreted both intervention and control images across time, the overall case mix, case complexity, and workload remained stable. However, the job titles and working years varied for the radiologists (Figure 3). Those working at Beijing Anzhen Hospital had a higher job rank and longer working experience. Therefore, these variables were further adjusted in the main analyses. Multimedia Appendix 5 presents the outcomes of the parallel trend assessment, illustrating that the report-drafting times of both experimental and comparison groups displayed a consistent trend before the introduction of the AI system.

XSL•FO
RenderX

Figure 2. Anatomical distribution of scans and prevalence of key abnormal findings. (A) Bar chart showing the anatomical distribution of computed tomography (CT) images for the chest (treatment) and nonchest (control) groups before and after the implementation of artificial intelligence (AI). (B) Stacked area charts depicting the monthly prevalence of primary abnormal findings from 2018 to 2021 for the 3 highest-volume anatomical groups: abdomen and pelvis (left), chest and lung (middle), and head and brain (right).

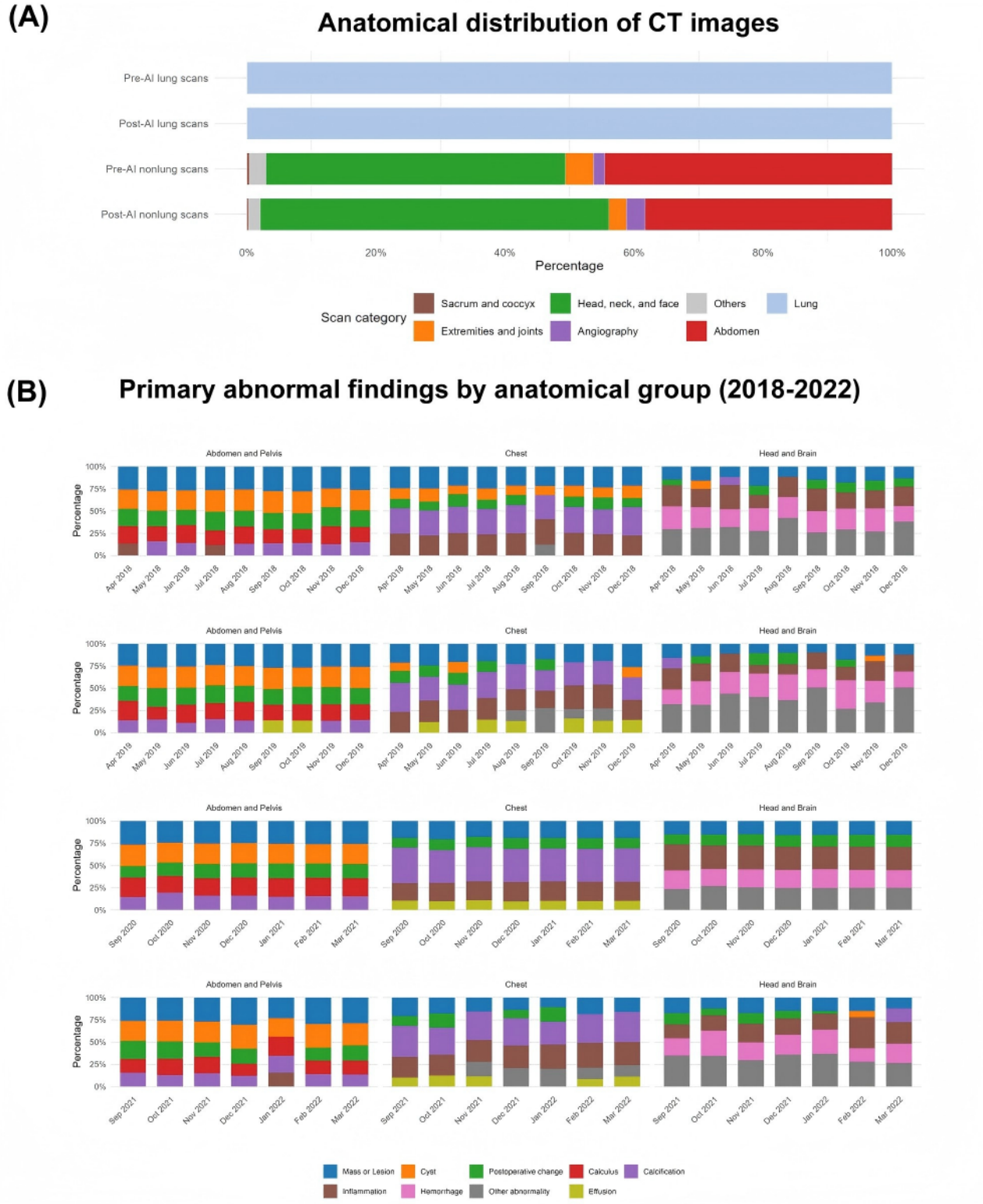

XSL•FO
RenderX

Figure 3. Radiologists' job titles and working years by hospital.

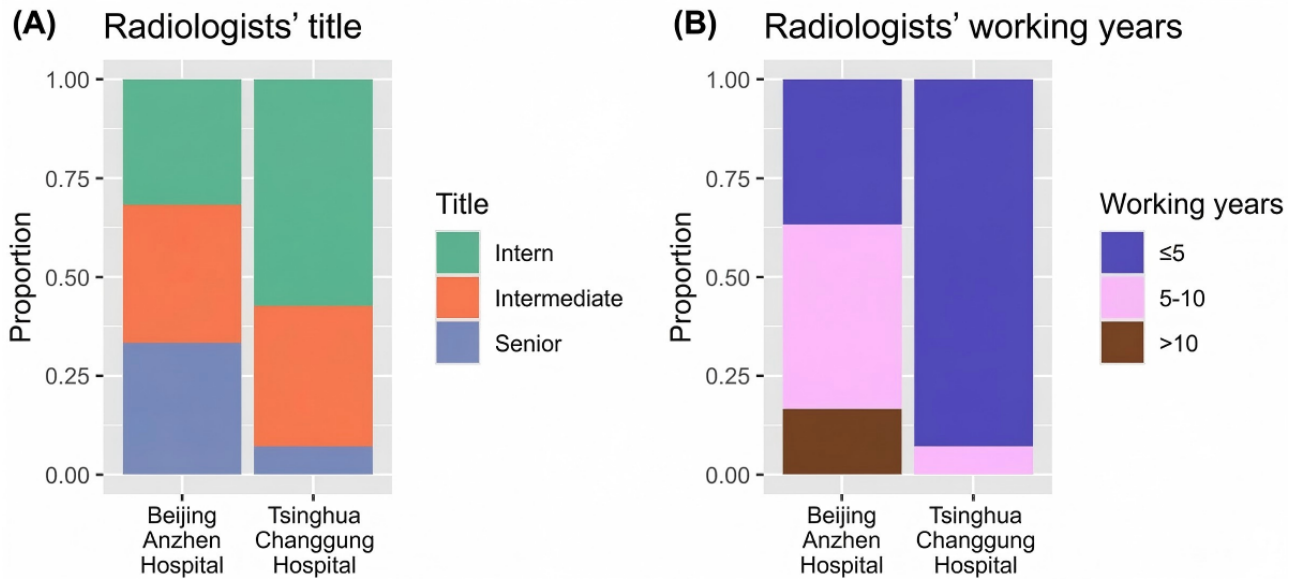

The adjusted DID model revealed a pooled effect size of 0.86 minutes (95% CI 0.14 to 1.57) after AI implementation. To assess the potential influence of anatomical site, we conducted sensitivity analyses by sequentially excluding each major category of nonchest CT images from the control group. As shown in Multimedia Appendix 6, the direction and magnitude of the AI effect remained generally consistent across all exclusion scenarios. However, this aggregated result should be interpreted with caution, as it is averaged over 2 hospitals with different implementation timelines. When breaking down to the hospital level, Tsinghua Changgung Hospital had a nonsignificant 0.90-minute increase (P=.12), while Beijing Anzhen Hospital saw a statistically significant 1.37-minute time saving (P<.001; Table 2). No statistically significant interaction was observed between AI exposure and radiologist seniority (all P>.09). In sensitivity analyses controlling for patient source, the estimated AI-associated effects remained robust (the overall DID estimate remained 0.86 minutes; P=.02).

Table 2. Adjusted report-drafting time differences for chest computed tomography (CT) vs others before and after artificial intelligence (AI) implementation, overall and by hospital.

Intervention and group	Patients, n	Report-drafting time[a] (minutes; 95% CI)	*P* value
Overall (2 hospitals)	185,044		
Control group[b]			
Pre-AI implementation	55,592	7.58 (6.32 to 8.85)	—[c]
Post-AI implementation	50,751	7.51 (6.10 to 8.93)	—
Pre-post difference	—	–0.07 (–0.62 to 0.48)	—
Experimental group[d]			
Pre-AI implementation	38,278	6.79 (5.40 to 8.19)	—
Post-AI implementation	40,423	7.58 (6.17 to 8.99)	—
Pre-post difference	—	0.78 (0.26 to 1.30)	—
DID[e]	—	0.86 (0.14 to 1.57)	.02
Tsinghua Changgung			
DID	87,678	0.90 (–0.28 to 2.08)	.12
Beijing Anzhen			
DID	97,366	–1.37 (–2.04 to –0.70)	<.001

[a]Report-drafting times are averaged over radiologists' sex, working years, and job title.

[b]Control group refers to CT image reports of nonlung sites that were not subject to the launch of the AI system.

[c]Not applicable.

[d]Experimental group refers to lung CT image reporting, which could be affected by AI assistance.

[e]DID: difference-in-differences (comparing changes in report-drafting time for lung vs nonlung CT images before and after AI implementation).

XSL•FO
RenderX

Subgroup analyses were conducted to account for the distinct implementation timelines of the AI systems at the 2 hospitals, as summarized in Figure 4 and Table 3. At Tsinghua Changgung Hospital, both the control group and the experimental group experienced extended report-drafting time, with a 0.90 minute (95% CI –0.28 to 2.08) nonsignificant increase attributable to AI assistance. However, Beijing Anzhen Hospital demonstrated an absolute 0.76-minute reduction in the AI-assisted group by the first year and a further 1.83-minute reduction by the second year. Using a DID framework, this corresponded to a 2.66-minute relative improvement vs the counterfactual trend (P<.001) 2 years into the AI implementation.

Figure 4. Adjusted changes in report-drafting time before and after artificial intelligence (AI) implementation. Gray indicates phases without AI assistance, and blue indicates phases with AI assistance. Tsinghua Changgung Hospital does not have second-year post-AI implementation data because the system was adopted later at that site. Report-drafting times are averaged over radiologist sex, years in practice, and job title. Difference-in-differences estimates compare changes in report drafting time for lung vs nonlung computed tomography reports before and after AI implementation. DID: difference-in-differences.

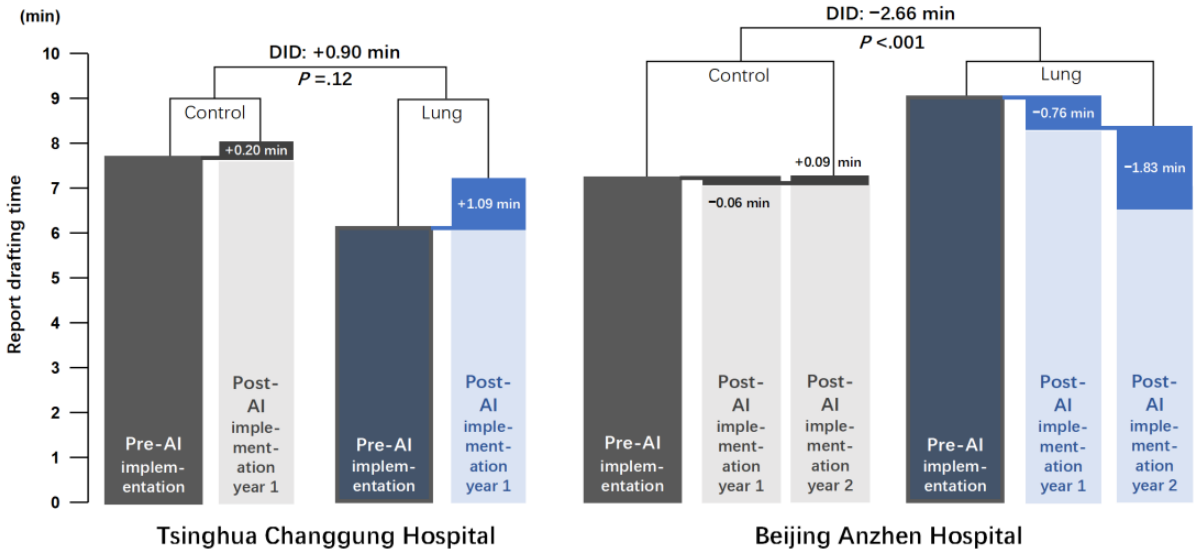

Table 3. Differences in adjusted report-drafting time comparing chest computed tomography (CT) vs others by post–artificial intelligence (AI) adoption duration.

	Adjusted report-drafting time[a] before AI implementation (min; 95% CI)	Adjusted report-drafting time[a] in the first year after AI implementation (min; 95% CI)	Adjusted report-drafting time[a] in the second year after AI implementation (min; 95 % CI)
Tsinghua Changgung			
Control group[b]	7.70 (6.61 to 8.79)	7.90 (6.26 to 9.53)	—[c]
Experimental group[d]	6.09 (4.69 to 7.59)	7.18 (5.38 to 8.99)	—
DID[e]	Reference	0.90 (–0.28 to 2.08)	—
Beijing Anzhen			
Control group[b]	7.36 (5.91 to 8.81)	7.30 (5.78 to 8.82)	7.39 (5.22 to 9.56)
Experimental group	9.22 (7.82 to 10.63)	8.46 (6.97 to 9.95)	6.63 (4.75 to 8.51)
DID	Reference	–0.70 (–1.35 to –0.05)	−2.66 (−3.35 to −1.96)

[a]Report-drafting times are averaged over radiologists' sex, working years, and job title.

[b]Control group refers to CT image reports of nonlung sites that were not subject to the launch of the AI system.

[c]Not available.

[d]Experimental group refers to lung CT image reporting, which could be affected by AI assistance.

[d]DID: difference-in-differences (comparing changes in report-drafting time for lung vs nonlung CT images before and after AI implementation).

XSL•FO
RenderX

Discussion

Adaptation of Diagnostic AI Systems in the Early Stages

Our findings indicate that the implementation of the AI system for lung nodule diagnostics was initially associated with an increase in report-drafting time, particularly in the first year of use. This transitional phase likely reflects a necessary adaptation period, as radiologists familiarized themselves with the system and learned to integrate it into existing workflows.

Technical barriers, such as outdated hospital IT infrastructure and limited interoperability with electronic medical record systems, may also impede seamless AI integration [23]. A potentially more important factor influencing the short-term impact of the AI system on lung nodule diagnosis may be physicians' self-efficacy. In the initial stages of the AI system's implementation, radiologists, especially junior staff, often lacked previous experience working with the AI system and were cautious in interpreting or relying on AI-generated results. This could partially explain the initial reporting delays observed at Tsinghua Changgung Hospital compared with the more experienced radiology team at Beijing Anzhen Hospital. Previous studies have shown that when radiologists lack confidence in the AI system's capabilities or have not yet fully acclimated to its operation, they might spend additional time validating AI outputs, rechecking discordant findings, or temporarily reverting to traditional workflows. This transitional phase can offset the efficiency gains intended by the AI system [24,25]. Limited formal training, reliance on trial-and-error learning, and incomplete understanding of deep learning mechanisms can further undermine early trust and efficiency [26-28].

However, over time, these early barriers diminished with routine use. A previous survey conducted at the same 2 hospitals showed strong user acceptance: radiologists rated the AI system highly in usability (System Usability Scale score: mean 74.3, SD 11.9) and efficiency improvement (mean 4.6, SD 0.6 out of 5) [29]. This growing user confidence likely reduced the need for repeated result validation and enabled more streamlined integration of the system, ultimately contributing to the substantial efficiency gains observed in later stages, such as the significant reduction in report-drafting time during the second year at Beijing Anzhen Hospital.

A related observation at Tsinghua Changgung Hospital further supports the interpretation of early-stage friction. During the first year, the control group did not exhibit a statistically significant change in report-drafting time, whereas the AI group experienced a meaningful increase. We confirmed with the hospital that no workflow reorganization, software upgrades, or staffing changes occurred during the study period. Therefore, the relative inefficiency observed in the AI arm at Tsinghua Changgung Hospital is unlikely to reflect systemic improvements in the control arm; instead, it aligns with our proposed mechanism that AI systems initially introduce cognitive load and workflow disruption. In contrast, Beijing Anzhen Hospital did not demonstrate a comparable degree of early inefficiency among its AI users. We hypothesize that this might be due to the higher baseline experience and seniority of its radiologists (Figure 3), which may facilitate faster adaptation to AI-assisted workflows.

These findings underscore the importance of considering the user characteristics and testing for potential learning curve when evaluating the impact of AI systems. Early performance metrics may underestimate their long-term value if insufficient time is allowed for user adaptation. If the initial learning curve does exist, structured training and supportive onboarding environments can help accelerate this process and maximize clinical benefits [30].

Long-Term Benefits of Diagnostic AI Systems

The long-term benefits of AI for diagnostic efficiency are complex and may vary significantly across institutions. While the pooled analysis suggests an overall increase in report-drafting time, this averaged effect masks the true clinical picture, where one hospital showed no immediate efficiency gains, while the other demonstrated a significant reduction in report-drafting time.

In high-volume radiology departments where radiologists face overwhelmingly heavy workloads over prolonged periods [31], information overload can lead to diagnostic errors [32,33]. Under such circumstances, even modest efficiency gains are valuable. In our study, extended use of the AI system for lung nodule analysis was associated with a reduction in average report-drafting time of up to 28% at one site. The AI system's inherent capacity to autonomously identify and measure lesions markedly reduces the burden of manual labor while also expediting the generation of standardized text reports. These collective features of the AI system, in turn, reduce the necessity of repetitive writing tasks [18]. In settings where radiologists typically read more than 50 scans per day, such an improvement amounts to 2 hours of saved time per radiologist daily. This could allow for higher throughput, shorter patient waiting times, and more time allocated to complex cases.

Notably, these efficiency gains emerged alongside previously established diagnostic benefits. Our previous work evaluating the same AI system demonstrated improved lung nodule detection with a modest early rise in false positives, underscoring the need for training and adaptation [34]. That analysis further showed stable diagnostic accuracy after senior review, suggesting no substantial downstream burden attributable to AI-assisted drafting. These findings provide important clinical context for this study, which shifts the focus from diagnostic performance to operational efficiency. While this analysis focuses on first-round report drafting by junior radiologists, the combined evidence suggests that AI adoption does not merely shift workload upstream but could integrate efficiently into routine clinical workflows.

Future Directions

While our study demonstrates the potential of AI to enhance efficiency within high-volume tertiary hospitals, its findings represent a necessary first step rather than a final answer. The successful deployment in these settings provides a proof of concept, but the ultimate goal is to improve health care resource allocation across the entire system. In terms of application

XSL•FO
RenderX

stages, future mixed methods investigations are necessary to identify institutional and user-level factors that may shape AI adoption trajectories. With regard to clinical settings, future research must directly examine AI's applicability in resource-constrained environments, which are often the hardest hit by medical shortages; first, partnering with primary and county-level hospitals as they begin to adopt AI; second, investigating the distinct barriers to adoption and effectiveness in low-resource settings (eg, outdated IT infrastructure, insufficient training, and different disease mixes); and third, examining how AI might be repurposed for tasks such as fast triage or ruling out common conditions in primary care, rather than for handling complex cases. Furthermore, the evaluation and validation of AI systems represent an iterative and dynamic process, necessitating a life cycle–based regulatory framework [35,36].

Strengths and Limitations

Our study has several strengths. First, we conducted a longitudinal analysis of data from 2 tertiary hospitals in China, covering both pre- and post-AI implementation periods in real-world settings. The inclusion of a larger sample size enhances the reliability of our conclusions. Second, we applied an adjusted DID design, a quasi-experimental approach increasingly used in health services research, which strengthened the rigor of examining the causal effect of AI's clinical impact. Finally, we analyzed report-drafting times by the same radiologists who interpreted both chest and nonchest CTs across the entire study period. This ensured that seasonal workload shocks, if any, would be shared across modalities within individuals, thus not differentially impacting the experimental and the control arm.

Nevertheless, our study had several limitations. First, both hospitals were large tertiary urban hospitals with high workloads and complex cases. As a result, our findings are considered only proof of concept, suggesting AI's potential to alleviate efficiency bottlenecks in saturated diagnostic workflows. Our conclusions may not generalize to primary or rural hospitals with different infrastructural and caseload challenges. Second, although we conducted extensive sensitivity analyses to confirm longitudinal stability in case mix, case complexity, and preintervention trends, the use of nonchest CT examinations as controls during a pandemic-affected period may only serve as imperfect controls. While we adjusted for key physician-level characteristics, the DID model may have omitted unmeasured confounders. However, the exclusion of unstable periods, inclusion of hospital and seasonal fixed effects, and the staggered rollout design lend support to the internal validity of our results. Third, while the different AI vendors deployed at the 2 hospitals had comparable diagnostic functions, variation in their workflows and interfaces may have affected the drafting time. Finally, implementing a triple-difference design was not feasible in this study. We concur that future research incorporating data from nonadopting or primary care sites is a critical next step.

Conclusions

Our findings highlight the heterogeneous impact of diagnostic AI on radiologists' efficiency. While short-term implementation was associated with an increase in report-drafting time in one hospital with a shorter follow-up period, sustained use was associated with significant efficiency gains in another hospital after more than 2 years of adoption. These divergent outcomes underscore that the realization of AI-driven efficiency is not guaranteed but is critically contingent on site-specific implementation dynamics and local workflow integration. Therefore, adequate adaptation and education are necessary for AI systems to meaningfully alleviate radiologists' workloads in the long run.

Acknowledgments

The authors sincerely thank all participating radiologists from Tsinghua Changgung Hospital and Beijing Anzhen Hospital for their invaluable contributions to this study. The authors are also grateful to the IT departments at both institutions for their technical support and assistance with data extraction and system integration. During the preparation of this manuscript, the authors used ChatGPT (OpenAI) to polish the English language and improve readability. After this, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

YW received funding from the Beijing Nova Program (grant 20250484908).

Data Availability

The datasets generated or analyzed during this study are available from the lead and the corresponding author on reasonable request. A sample dataset with the same structure as the original data, but with anonymized, masked, and sequence-perturbed entries for privacy protection, is provided in Multimedia Appendix 4.

Authors' Contributions

Conceptualization: WL, YW, MJB, ZZ, HK
Methodology: WL, YW
Investigation: WL, WY
Formal analysis: WL, YW

XSL•FO
RenderX

Data curation: WL, WY
Visualization: YW
Resources: ZZ
Supervision: HK
Writing—original draft: WL, YW
Writing—review and editing: ZZ, HK
WL and YW contributed equally to this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
User interface of the artificial intelligence–assisted system under study.
[PNG File , 684 KB - jmir_v28i1e77967_app1.png]

Multimedia Appendix 2
Representative photographs of the artificial intelligence–assisted computed tomography reporting workflow.
[PNG File , 1035 KB - jmir_v28i1e77967_app2.png]

Multimedia Appendix 3
Comparison of artificial intelligence system use workflows between the 2 hospitals.
[PNG File , 314 KB - jmir_v28i1e77967_app3.png]

Multimedia Appendix 4
Sample dataset.
[XLSX File (Microsoft Excel File), 42 KB - jmir_v28i1e77967_app4.xlsx]

Multimedia Appendix 5
Parallel trend assessment.
[PNG File , 97 KB - jmir_v28i1e77967_app5.png]

Multimedia Appendix 6
Sensitivity analysis using the leave-one-category-out approach.
[PNG File , 112 KB - jmir_v28i1e77967_app6.png]

References

1. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism 2017 Apr;69S:S36-S40. [doi: 10.1016/j.metabol.2017.01.011] [Medline: 28126242]
2. Alexander A, Jiang A, Ferreira C, Zurkiya D. An intelligent future for medical imaging: a market outlook on artificial intelligence for medical imaging. J Am Coll Radiol 2020 Jan;17(1 Pt B):165-170 [FREE Full text] [doi: 10.1016/j.jacr.2019.07.019] [Medline: 31918875]
3. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ. Artificial intelligence in radiology. Nat Rev Cancer 2018 Aug;18(8):500-510 [FREE Full text] [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
4. McDonald RJ, Schwartz KM, Eckel LJ, Diehn FE, Hunt CH, Bartholmai BJ, et al. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. Acad Radiol 2015 Sep;22(9):1191-1198. [doi: 10.1016/j.acra.2015.05.007] [Medline: 26210525]
5. Smith-Bindman R, Miglioretti DL, Johnson E, Lee C, Feigelson HS, Flynn M, et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996-2010. JAMA 2012 Jun 13;307(22):2400-2409 [FREE Full text] [doi: 10.1001/jama.2012.5960] [Medline: 22692172]
6. Liu Z, Yao Q, Wang P, Shen L, Li H, Zhang J, et al. The price of pressure: nationwide survey on lifestyle disturbances, occupational burnout and compromised perceived-competency among radiology residents in China. Front Public Health 2024 Oct 23;12:1472397 [FREE Full text] [doi: 10.3389/fpubh.2024.1472397] [Medline: 39507651]
7. Zhan H, Schartz K, Zygmont ME, Johnson JO, Krupinski EA. The impact of fatigue on complex CT case interpretation by radiology residents. Acad Radiol 2021 Mar;28(3):424-432 [FREE Full text] [doi: 10.1016/j.acra.2020.06.005] [Medline: 32622748]
8. Yip W, Fu H, Chen AT, Zhai T, Jian W, Xu R, et al. 10 years of health-care reform in China: progress and gaps in universal health coverage. Lancet 2019 Sep 28;394(10204):1192-1204. [doi: 10.1016/S0140-6736(19)32136-1] [Medline: 31571602]

XSL•FO
RenderX

9. Kong X, Ai B, Kong Y, Su L, Ning Y, Howard N, et al. Artificial intelligence: a key to relieve China's insufficient and unequally-distributed medical resources. Am J Transl Res 2019 May 15;11(5):2632-2640 [FREE Full text] [Medline: 31217843]
10. Wenderott K, Krups J, Weigl M, Wooldridge AR. Facilitators and barriers to implementing AI in routine medical imaging: systematic review and qualitative analysis. J Med Internet Res 2025 Jul 21;27:e63649 [FREE Full text] [doi: 10.2196/63649] [Medline: 40690758]
11. Bonekamp D, Schlemmer HP. [Artificial intelligence (AI) in radiology?: Do we need as many radiologists in the future?]. Urologe A 2022 Apr;61(4):392-399. [doi: 10.1007/s00120-022-01768-w] [Medline: 35277758]
12. Waller J, O'Connor A, Rafaat E, Amireh A, Dempsey J, Martin C, et al. Applications and challenges of artificial intelligence in diagnostic and interventional radiology. Pol J Radiol 2022 Feb 25;87:e113-e117 [FREE Full text] [doi: 10.5114/pjr.2022.113531] [Medline: 35280945]
13. He C, Liu W, Xu J, Huang Y, Dong Z, Wu Y, et al. Efficiency, accuracy, and health professional's perspectives regarding artificial intelligence in radiology practice: a scoping review. iRadiology 2024 Apr 08;2(2):156-172. [doi: 10.1002/ird3.63]
14. Yi C, Tang Y, Ouyang R, Zhang Y, Cao Z, Yang Z, et al. The added value of an artificial intelligence system in assisting radiologists on indeterminate BI-RADS 0 mammograms. Eur Radiol 2022 Mar;32(3):1528-1537. [doi: 10.1007/s00330-021-08275-0] [Medline: 34528107]
15. Stephan D, Bertsch A, Burwinkel M, Vinayahalingam S, Al-Nawas B, Kämmerer PW, et al. AI in dental radiology-improving the efficiency of reporting with ChatGPT: comparative study. J Med Internet Res 2024 Dec 23;26:e60684 [FREE Full text] [doi: 10.2196/60684] [Medline: 39714078]
16. Zia A, Fletcher C, Bigwood S, Ratnakanthan P, Seah J, Lee R, et al. Retrospective analysis and prospective validation of an AI-based software for intracranial haemorrhage detection at a high-volume trauma centre. Sci Rep 2022 Nov 18;12(1):19885 [FREE Full text] [doi: 10.1038/s41598-022-24504-y] [Medline: 36400834]
17. Pan F, Li L, Liu B, Ye T, Li L, Liu D, et al. A novel deep learning-based quantification of serial chest computed tomography in Coronavirus disease 2019 (COVID-19). Sci Rep 2021 Jan 11;11(1):417 [FREE Full text] [doi: 10.1038/s41598-020-80261-w] [Medline: 33432072]
18. Qi LL, Wu BT, Tang W, Zhou LN, Huang Y, Zhao SJ, et al. Long-term follow-up of persistent pulmonary pure ground-glass nodules with deep learning-assisted nodule segmentation. Eur Radiol 2020 Feb;30(2):744-755. [doi: 10.1007/s00330-019-06344-z] [Medline: 31485837]
19. Tan HM, Tan MS, Chang ZY, Tan KT, Ee GL, Ng CC, et al. The impact of COVID-19 pandemic on the health-seeking behaviour of an Asian population with acute respiratory infections in a densely populated community. BMC Public Health 2021 Jun 23;21(1):1196 [FREE Full text] [doi: 10.1186/s12889-021-11200-1] [Medline: 34158019]
20. Sample Size Calculators for designing clinical research. Sample Size Calculators. URL: https://sample-size.net/ [accessed 2026-03-03]
21. Ahn JS, Ebrahimian S, McDermott S, Lee S, Naccarato L, Di Capua JF, et al. Association of artificial intelligence-aided chest radiograph interpretation with reader performance and efficiency. JAMA Netw Open 2022 Aug 01;5(8):e2229289 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.29289] [Medline: 36044215]
22. Wing C, Simon K, Bello-Gomez RA. Designing difference in difference studies: best practices for public health policy research. Annu Rev Public Health 2018 Apr 01;39(1):453-469 [FREE Full text] [doi: 10.1146/annurev-publhealth-040617-013507] [Medline: 29328877]
23. Li J, Zhou L, Zhan Y, Xu H, Zhang C, Shan F, et al. How does the artificial intelligence-based image-assisted technique help physicians in diagnosis of pulmonary adenocarcinoma? A randomized controlled experiment of multicenter physicians in China. J Am Med Inform Assoc 2022 Nov 14;29(12):2041-2049 [FREE Full text] [doi: 10.1093/jamia/ocac179] [Medline: 36228127]
24. Brown M, Browning P, Wahi-Anwar MW, Murphy M, Delgado J, Greenspan H, et al. Integration of chest CT CAD into the clinical workflow and impact on radiologist efficiency. Acad Radiol 2019 May;26(5):626-631. [doi: 10.1016/j.acra.2018.07.006] [Medline: 30097402]
25. Müller FC, Raaschou H, Akhtar N, Brejnebøl M, Collatz L, Andersen MB. Impact of concurrent use of artificial intelligence tools on radiologists reading time: a prospective feasibility study. Acad Radiol 2022 Jul;29(7):1085-1090 [FREE Full text] [doi: 10.1016/j.acra.2021.10.008] [Medline: 34801345]
26. Ooi SK, Makmur A, Soon AY, Fook-Chong S, Liew C, Sia SY, et al. Attitudes toward artificial intelligence in radiology with learner needs assessment within radiology residency programmes: a national multi-programme survey. Singapore Med J 2021 Mar;62(3):126-134 [FREE Full text] [doi: 10.11622/smedj.2019141] [Medline: 31680181]
27. Abuzaid MM, Elshami W, Tekin H, Issa B. Assessment of the willingness of radiologists and radiographers to accept the integration of artificial intelligence into radiology practice. Acad Radiol 2022 Jan;29(1):87-94. [doi: 10.1016/j.acra.2020.09.014] [Medline: 33129659]
28. Bergquist M, Rolandsson B, Gryska E, Laesser M, Hoefling N, Heckemann R, et al. Trust and stakeholder perspectives on the implementation of AI tools in clinical radiology. Eur Radiol 2024 Jan;34(1):338-347 [FREE Full text] [doi: 10.1007/s00330-023-09967-5] [Medline: 37505245]

XSL•FO
RenderX

29. Liu W, Wu Y, Zheng Z, Yu W, Bittle MJ, Kharrazi H. Evaluating artificial intelligence's role in lung nodule diagnostics: a survey of radiologists in two pilot tertiary hospitals in China. J Clin Imaging Sci 2024 Aug 23;14:31. [doi: 10.25259/JCIS_72_2024] [Medline: 39246733]
30. Sun L, Yin C, Xu Q, Zhao W. Artificial intelligence for healthcare and medical education: a systematic review. Am J Transl Res 2023 Jul 15;15(7):4820-4828 [FREE Full text] [Medline: 37560249]
31. Gampala S, Vankeshwaram V, Gadula SS. Is artificial intelligence the new friend for radiologists? A review article. Cureus 2020 Oct 24;12(10):e11137 [FREE Full text] [doi: 10.7759/cureus.11137] [Medline: 33240726]
32. Paul S, Nazareth DL. Input information complexity, perceived time pressure, and information processing in GSS-based work groups: an experimental investigation using a decision schema to alleviate information overload conditions. Decis Support Syst 2010 Apr;49(1):31-40 [FREE Full text] [doi: 10.1016/j.dss.2009.12.007]
33. Vosshenrich J, Brantner P, Cyriac J, Boll DT, Merkle EM, Heye T. Quantifying radiology resident fatigue: analysis of preliminary reports. Radiology 2021 Mar;298(3):632-639. [doi: 10.1148/radiol.2021203486] [Medline: 33497316]
34. Liu W, Wu Y, Zheng Z, Bittle M, Yu W, Kharrazi H. Enhancing diagnostic accuracy of lung nodules in chest computed tomography using artificial intelligence: retrospective analysis. J Med Internet Res 2025 Jan 27;27:e64649 [FREE Full text] [doi: 10.2196/64649] [Medline: 39869890]
35. Liu Y, Chen PH, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. JAMA 2019 Nov 12;322(18):1806-1816. [doi: 10.1001/jama.2019.16489] [Medline: 31714992]
36. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021 Jul 15;385(3):283-286 [FREE Full text] [doi: 10.1056/NEJMc2104626] [Medline: 34260843]

Abbreviations

AI: artificial intelligence
CT: computed tomography
DID: difference-in-differences
IRB: institutional review board

Edited by A Coristine; submitted 23.May.2025; peer-reviewed by J Jiang, J Zhang; comments to author 17.Jun.2025; revised version received 13.Feb.2026; accepted 14.Feb.2026; published 27.Mar.2026.

Please cite as:
Liu W, Wu Y, Yu W, Bittle MJ, Zheng Z, Kharrazi H
Measuring the Impact of AI on Report-Drafting Efficiency in Chest Computed Tomography Interpretation: Retrospective Analysis
J Med Internet Res 2026;28:e77967
URL: https://www.jmir.org/2026/1/e77967
doi:10.2196/77967
PMID:

©Weiqi Liu, You Wu, Wei Yu, Mark J Bittle, Zhuozhao Zheng, Hadi Kharrazi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Enhancing Detection of Message Intents in a Mobile Health Smoking-Cessation Intervention Using Large Language Model Fine-Tuning, Data Downsampling, and Error Correction: Algorithm Development and Validation

Shagoto Rahman[1], MS; Cornelia (Connie) Pechmann[2], PhD; Ian G Harris[1], PhD

[1]Department of Computer Science, University of California, Irvine, Irvine, CA, United States

[2]The Paul Merage School of Business, University of California, Irvine, Irvine, CA, United States

Corresponding Author:
Shagoto Rahman, MS
Department of Computer Science, University of California, Irvine, Irvine, CA, United States

Abstract

Background: Although smoking-cessation aids such as support groups and nicotine replacement therapy (NRT) can help people quit, quit rates remain low. Mobile health interventions can boost accessibility and engagement, especially with NRT, but require ongoing effort to deliver timely responses. Accurate intent detection is crucial for identifying user needs and delivering timely, appropriate chatbot responses. Recent large language model advancements in natural language processing and artificial intelligence (AI) have shown promise. However, these systems often struggle with many intent categories, complex language, and imbalanced data, reducing recognition accuracy.

Objective: The main goal of this study was to develop an AI tool, a large language model that could accurately detect people's message intents, despite dataset imbalances and complexities. In our application, the messages came from a smoking-cessation support-group intervention and often involved the use of NRT provided as part of that intervention.

Methods: We consistently used a state-of-the-art public domain large language model, Llama-3 8B (8 billion parameters) from Meta. First, we used the model off-the-shelf. Second, we fine-tuned it on our annotated dataset with 25 intent categories. Third, we also downsampled the predominant intent category to reduce model bias. Finally, we combined downsampling with corrected human annotations, creating a cleaned dataset for a new round of fine-tuning.

Results: Without fine-tuning, the model achieved unweighted and weighted F_1-scores (overall performance) of 0.41 and 0.38, respectively, on the downsampled corrected test dataset, and 0.29 and 0.35 on the full test dataset. Fine-tuning improved performance to 0.77 and 0.80 on the downsampled corrected dataset, and 0.72 and 0.86 on the full dataset. Fine-tuning with downsampling attained the best F_1-scores, 0.88 and 0.91 on the downsampled corrected dataset, though performance dropped on the full test dataset (0.58 unweighted, 0.66 weighted) due to the predominance of the off-topic intent category, while unweighted recall remained high (0.80). The final method combining fine-tuning, downsampling, and error correction achieved 0.86 unweighted and 0.90 weighted F_1-scores on the downsampled corrected dataset, and 0.57 and 0.65 on the full dataset with unweighted recall improving to 0.82.

Conclusions: Large language models performed poorly without fine-tuning, highlighting the need for domain-specific training. Even with fine-tuning, performance was limited by a highly imbalanced dataset. Downsampling before fine-tuning moderately improved performance but still left room for improvement and concerns about dataset noise. A careful review of model-human disagreement cases helped identify human annotation errors. After error correction, the method without error correction still achieved slightly higher precision and F_1-score on the corrected test dataset. While error correction slightly improved recall on noisy data, automated downsampling alone may be sufficient, making manual correction a more resource-intensive option with limited added benefit.

(J Med Internet Res 2026;28:e83437) doi:10.2196/83437

KEYWORDS

smoking cessation; intent detection; data imbalance; large language model; human annotation error analysis; chatbot

XSL•FO
RenderX

Introduction

Background

Although most smokers wish to stop smoking, cessation rates remain low in part because of the insufficient use of the cessation aids [1-3]. The use of mobile health interventions to aid people in quitting smoking has become commonplace [4,5], including SMS text messaging programs [6,7] and quit-smoking apps [4,8]. As traditional in-person support groups significantly improve quit rates [9,10], online groups have emerged to help overcome usage barriers related to accessibility and convenience [11]. Mobile health smoking-cessation support groups offer considerable promise; however, they require ongoing, resource-intensive efforts to provide timely, 24/7 responses that sustain participant engagement. Research has consistently demonstrated that active involvement in online support groups leads to improved outcomes [12,13], and studies also show that low engagement is often linked to dropouts [14].

Chatbots can enhance support-group member engagement and minimize dropout rates by providing continuous, cost-effective support [15-17]. Accurate intent detection, defined as the precise and consistent classification of a user's communication goal from their input message, is a leading determinant of chatbot effectiveness. Accurate intent detection can improve user engagement by providing the inputs needed to the system to provide personalized, accurate, and timely support. It can also minimize the need for human intervention. With accurate intent detection, a chatbot can engage users by actively responding to messages with relevant content, addressing their concerns, and delivering evidence-based health information and support. For example, if a user indicates they have failed to quit smoking, the chatbot can identify this particular intent and suggest ways to overcome the challenge.

Previous chatbot studies have primarily focused on 1-to-1 conversations between chatbot and user, for example, for mental health or substance abuse screening [18] or interventions [19]. We study conversations in support groups, often involving several group members, which have unique properties compared to 1-to-1 conversations including multiparty dialogue, interruptions, subgroup discussions, and incomplete messaging leading to exceptionally unstructured and noisy data [20]. In addition, research has been done to detect message intents in medical settings, for example, online forums or office visits, but most involved limited intent categories which limited their ability to fully capture the message content [21,22].

Moreover, data imbalance is particularly common in the health domain because sufficient data is rarely available for all intent categories, yet ones with fewer samples can be critically important [23,24]. For example, an intent category such as overdose may have very few messages, but it represents a critical concern. Most prior studies did not properly ensure that the samples or exemplars were sufficiently balanced across intent categories, possibly leading to suboptimal model performance. Data downsampling, which involves reducing the number of samples from highly dominant majority categories to achieve a more balanced category distribution, can play a vital role. This approach can help the classifying model learn to detect all intent categories comparably, including those that are rare but critically important.

Large language models can deliver state-of-the-art performance in recognizing message intents because they are trained on vast amounts of data. As a result, these models can grasp the deeper meaning of text and often replicate human-like understanding. However, for highly specific classification tasks, large language models must be fine-tuned on domain-specific datasets where humans have annotated the message intents for this fine-tuning phase. Since data annotation is primarily done by humans, it often contains errors due to message ambiguity, human misunderstanding, carelessness, bias, or other factors. Manually reviewing the entire dataset to find human annotation errors is costly and time-consuming. Instead, focusing only on the data points where the model and human annotators disagree on the user's message intent can save time and effort while effectively identifying annotation issues.

In this study, we demonstrate common challenges and solutions in using large language models for mobile health interventions. We implement a state-of-the-art, public domain, large language model called Llama-3 8B (8 billion parameters) from Meta to detect the message intents in a mobile health smoking-cessation intervention. Our study method consists of 4 parts. Part 1 involves intent detection using the Llama-3 8B large language model off-the-shelf, without fine-tuning. Part 2 fine-tunes the model on a domain-specific dataset which human annotators annotated with what they believed to be the message intents. Part 3 combines fine-tuning with downsampling of the majority (predominant) intent category, which was found to bias the model and weaken its performance at intent recognition. Part 4 adds correction of human error in intent annotation to the fine-tuning and downsampling. The rest of this paper is organized as follows: We elaborate on prior work, discuss our methods, report our experimental results, summarize the principal findings and limitations, and state future research directions and a conclusion.

Prior Work

Conversational Agents in Health Care

Conversational health care agents have become important resources for providing quick and easily accessible medical assistance to patients. Recent research suggests that by harnessing advancements in natural language processing, they can provide personalized patient support and alleviate the excessive workload of many health care professionals. Babu and Boddu [25] created a BERT (Bidirectional Encoder Representations From Transformers)-based medical conversational agent to provide responses to people seeking medical advice, appointments, or information about specific medical conditions. The BERT model was fine-tuned on public databases such as MIMIC-III (Medical Information Mart for Intensive Care-III), PubMed, and a COVID-19 dataset totaling 11,000 samples of medical questions. The conversational agent integrated these extracted datasets with the ongoing conversation context and produced relevant and personalized feedback. When compared to baseline models such as LSTM (long short-term memory), SVM (support vector machine), and Bi-LSTM (bidirectional long short-term memory), the chatbot

XSL•FO
RenderX

demonstrated superior performance, attaining 98% accuracy along with high precision, recall, and F_1-scores combining precision and accuracy. However, the system demanded high computational power and suffered from data bias and interpretability issues.

Yu and McGuinness [26] developed a conversational agent by combining a fine-tuned DialoGPT model and ChatGPT to assist mental health conversations. The authors sought to address current model limitations in delivering proper and safe therapeutic responses due to their lack of training on specialized data. So, authors fine-tuned the DialoGPT model on a domain-specific dataset focused on cognitive behavioral therapy and integrated it with ChatGPT's dynamic conversational capabilities. This hybrid conversational agent system was compared to conventional rule-based models and large language models on metrics such as BLEU (Bilingual Evaluation Understudy), which assesses the match between the model output and human-written text. Inputs from mental health professionals and patients were also considered. The results demonstrated that the hybrid system outperformed others in terms of conversational quality, relevance, and emotional resonance.

Zamani et al [27] sought to address the limitations of rule-based chatbots as mental health conversational agents, including lack of response creativity and low user acceptability. They also sought to address the risks of large language models in terms of generating harmful texts and the limitations of data imbalance. The authors evaluated the impact of novel text augmentation techniques to improve the message intent detection. Their study demonstrated that increasing the training dataset through augmentation could boost model accuracy by contributing to dataset diversity. The study incorporated easy data augmentation (EDA), synonym replacement using WordNet, embedding-based replacement, and context-aware substitutions. Model performance improved markedly through data augmentation, with up to a 26.4% increment in the average intent detection score.

Large Language Models for Intent Detection in Health Care

Grasping the speaker's message intent is essential for facilitating smooth and meaningful human interaction. With recent advancements in natural language processing and deep learning, various recent studies have demonstrated success in detecting the subtle intents behind spoken expressions. Aftab et al [28] developed an intent classification method for health care conversational agents by incorporating Bayesian LSTMs. They used an open-source dataset containing 6661 text utterances corresponding to 25 unique medical symptom-related intents with a relatively balanced distribution of intent categories. Their method attained an impressive 99.4% accuracy in recognizing message intents on the test dataset. Moreover, the authors used Shannon entropy from several random model runs to measure uncertainty, which helped the system distinguish between familiar and unfamiliar inputs. Additionally, to estimate model uncertainty, they used Monte Carlo dropout during inference. As a result, the chatbot handled uncertain predictions more appropriately, allowing it to generate safer, less harmful responses.

Zhang et al [29] developed Conco-ERNIE, an innovative model designed to enhance user intent detection in complex medical queries by permitting single intent, multi-intent, and implicit intent recognition. The model used a special feature extractor called ERNIE with an added layer to help it understand and annotate important parts of medical questions. The authors incorporated the Apriori algorithm to find patterns in how medical concepts related to each other, allowing it to better detect intents when the uttered meanings were unclear. Moreover, they introduced an attention-based method that combined the meaning of words in a query with the patterns of related medical concepts. This helped the model focus better on important information and ignore irrelevant details. The authors tested their method on a dataset of real medical questions, and it showed clear improvements in detecting multiple and hidden user intents. Overall, the model achieved a score of 89.4% in test data, much higher than other popular methods. The authors then added their model Conco-ERNIE to a real health care chatbot linked to a large medical knowledge base.

Sakurai and Miyao [30] studied large language models for intent detection in conversations involving donation requests. The study examined conversations where the responder expressed criticism of donations or sought to make donors feel guilty. The authors primarily studied GPT-4 in this context and found it often misidentified critical, sarcastic, or negative affective responses as motivational or neutral. GPT-4 was tested on both real and artificially created conversations, and its predictions about intent were compared to human-annotated intents. Humans were able to detect critical and other negative messages, but GPT-4 often misunderstood them as encouraging or motivational. Overall, this study found that a state-of-the-art large language model struggled to detect subtle intents in negative speech.

Our Contributions

In this paper, we demonstrate that a current state-of-the-art large language model cannot adequately detect subtle intents expressed by participants in a smoking-cessation intervention; for example, the model struggles to decipher substantively different messages about the study-provided nicotine replacement therapy (NRT). The methodological novelty of our work lies in systematically addressing highly imbalanced, domain-specific data for smoking-cessation text classification using the Llama-3 8B model. We illustrate the specific steps that can be taken, using existing analytical methods, to improve the model's intent detection. We show vast improvements by fine-tuning the off-the-shelf model using a domain-specific dataset with human annotations of 25 message intents. We show additional substantial improvements by downsampling (reducing excess examples of) an overly dominant intent category and by correcting human annotation errors identified by examining model-human disagreement in intent classification. In sum, we show that large language models can perform very poorly in domain-specific contexts like ours but can be massively improved with current analytical approaches. This approach

XSL•FO
RenderX

improves efficiency and robustness, providing insights for future work such as multi-intent detection, probability calibration, and translation of technical performance into operational impact.

Methods

Overview

Figure 1 presents our method for training a large language model for the detection of message intents. It involved four stages: (1) large language model without fine-tuning, where the model predicted intents without any domain-specific training dataset solely using its own knowledge; (2) large language model with fine-tuning, where the model was fine-tuned on an intent-annotated dataset to improve performance; (3) large language model with fine-tuning and downsampling, which addressed category imbalance by reducing the sample size of a majority (dominant) intent category, which was off-topic or nondomain related, before fine-tuning; and (4) large language model with fine-tuning, downsampling, and human annotation error correction, where model-human disagreements were reviewed for potential human annotation errors, corrected if necessary, and used to retrain the model. This stepwise approach allowed for progressive enhancement in classification accuracy, with each stage addressing specific limitations of the previous one.

Figure 1. Overview of the 4 large language model training methods used for intent detection in this study (United States).

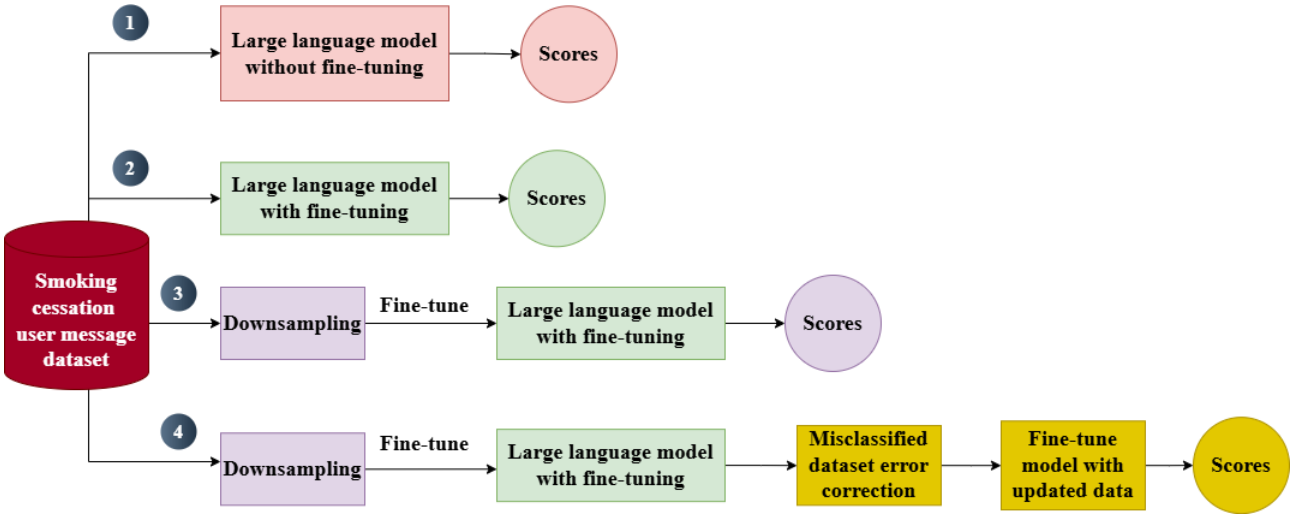

Dataset

For this study, we use all the user messages from the Tweet2Quit mobile health smoking-cessation intervention [17]. All duplicated messages have been properly identified and removed, ensuring the dataset is clean and free from leakage. Tweet2Quit is a social media–based intervention to assist people with quitting smoking. Participants were assigned to private online support groups of 20 members who engaged in peer support discussions about smoking cessation. The dataset contained more than 82,000 messages gathered from 45 support groups participating in 2 clinical trials conducted between 2012 - 2014 and 2016 - 2019. Each support group remained active for 3 months and generated approximately 1822 messages on average. The messages were created by individuals actively working to quit smoking and then stay quit and, as a result, contained domain-rich, real-world interactions regarding quitting and the free NRT, or NRT (nicotine patches, gum, and lozenges) provided as a quitting aid.

A team of 25 trained research assistants, mostly US undergraduates from diverse majors, with a few international participants, annotated the intent behind each message to produce a comprehensive annotated corpus. The assistants were trained on intent definitions, keywords, and examples, then practiced on a set of messages until they attained the required ≥80% accuracy. Initial in-person training meetings transitioned to Microsoft Teams due to team expansion and COVID-19, with weekly sessions during training and biweekly discussions thereafter to resolve difficult cases. Each message was independently annotated by two assistants, with discrepancies resolved by a third expert adjudicator.

A total of 25 intent categories were identified while annotating the messages. The topics either were clinically related to smoking cessation (eg, nrt_howtouse, nrt_overdose, and nrt_mouthirritation), reflected social support for quitting (eg, greetings, support), or were off-topic (eg, small talk, weather, news). Table 1 presents the 25 message intent categories identified in the dataset, along with their definitions and representative examples. The annotation demonstrated high reliability, with a Cohen κ score of 0.93 overall, and agreement between specific annotators ranging from 87.3% to 97.2% across intent categories. The off-topic intent category was used to indicate irrelevant messages unrelated to smoking cessation or support group bonding. It alone accounted for 56% of the dataset, which indicated significant category imbalance.

XSL•FO
RenderX

Table . Definitions and examples of smoking-cessation message intents used in this intent detection study (United States, 2012 - 2014 and 2016 - 2019).

Intent category	Definition	Examples
cigsmell	Complains about cigarette smell	• Cigarette smoke and cigarette smokers stink to me now. • I can smell people smoking a mile away now. It's gross to me now.
cravings	Discusses craving/wanting/temptation/urge/trigger to smoke or how to fight/avoid craving	• It's hard because all I want to do is reach over and get a cigarette and smoke. • My urge is in the evening watching TV I usually occupy my mind with crochet that's all I really have.
ecigs	States something about ecigs, vape, or liquid or juice	• Is anyone going to try vaping to help quit? The non-nicotine juice of course. I've been debating. • I do like the Vape pen.
fail	Admits to smoking a cigarette	• I have to admit I slipped today and smoked a cigarette. • Mornings I'm a chain smoker.
greetings	Says a greeting, eg, hi	• Good afternoon! How's everyone doing today? • Good morning!
health	States something about the relation between smoking and health	• My breathing is much better and I don't never want to smoke a cigarette again. • To be able to breathe better.
nrt_dontwork	States NRT[a] doesn't work, don't like, no longer needed	• Since the gum and patches don't seem to work on my boring days I have struggled. • This gum is a no go for me. I called the smoke shop in my town.
nrt_dreams	States they either sleep with the patch or try not to do so; often discusses dreams	• I'm sleeping with patch on. Have the craziest most vivid dreams. • Yes! Very lucid dreams. I had to stop sleeping with them on.
nrt_howtouse	Asks question or gives info about how to use NRT in general, eg, dose	• How often should I use the lozenges? • So your not suppose to constantly chew the gum right? Suppose to just let it lay in your gums once broken up right?
nrt_itworks	States reason why NRT used, eg, it works, stops craving	• I'm using 1 patch a day and chewing the gum as needed! Both help immensely with cravings! • I have stepped down to the 7mg patch!!!! Yea!!!!!
nrt_mouthirritation	States NRT gum/lozenge causes bad taste, throat gum irritation or burning, or spicy sensation	• I toss the gum when it starts burning my mouth or throat. • Don't like the gum I'm having a hard time chewing it. Burns my mouth a little.
nrt_nauseous	States NRT makes them nauseous or gag or sickly	• Don't like the gum makes me nauseous. • Using the #14 patch. Not using gum it makes me sick.
nrt_od	States NRT is too strong, feels like overdose	• I am going to move to the step 2 patch. The 21mg patches are making me sweat and feel bad. • I had to move from the 21mg early cause it was just too much nicotine!

XSL•FO
RenderX

Intent category	Definition	Examples
nrt_skinirritation	States NRT patch causes skin irritation	• Anybody else's skin being irritated by the patches? • I get itchy from mine too.
nrt_stickissue	States NRT patch won't stick	• I know that with sweat the patches don't stay on I have problems with mine every-day. • My patches won't stay on.
off-topic	Discusses other topic, eg, small talk, weather, news	• Now i gotta go check mine!!! • I skipped breakfast and went right for lunch.
quitdate	States something about quit date	• Hey my quitDate is today. • My last pack of smokes. Friday is the day!
savingmoney	States money they spend per pack or how much they save by quitting	• I'm putting away the money I save into an account and then can do something fun! • All the money spent to damage health.
scared	States being scared/nervous/anxious/fearful about quitting or sometimes NRT	• I am really nervous about quitting but ready. • I've not tried the gum yet. I'm scared of it lol.
smokefree	States success in being smoke-free	• I'm still smokefree starting week 2. YAY. • I made it thru without smoking. Starting Day 2. Yeah.
smokingless	Has not yet quit, only able to cut down on cigs smoked OR starting to quit but still smokes some	• Can’t say I've completely quit. But 4 or 5 a day is less than 15. • Proud moment: only 1 cigarette today. Never thought I'd get this far.
stress	States feeling stress	• I think stress has been tough but I'm not giving up. • Last week work was really stressful had a lot going on at once.
support	Supports other participants	• Good for you for not giving up. You got this!!! • You're doing great! wooohooo!!!
tiredness	States feeling tired due to unspecified reasons or quitting smoking, not NRT	• I have been so tired lately. I'm wondering if it is connected to quitting smoking. • So very tired...
weightgain	States gaining weight or eating more or anything about weight or exercise	• Ever since I quit I eat more so I'm going to start eating healthier and exercise. • I have gained 7 pounds!

[a]NRT: nicotine replacement therapy.

The annotated dataset was divided into 3 mutually exclusive and collectively exhaustive subsets: training, evaluation, and test. The test dataset consisted of messages from support groups 30 to 36. From the remaining groups out of the total 45, 75% of the messages were used for training and 25% for evaluation. The dataset split was stratified based on the distribution of intent categories, independent of group membership. The training dataset was used to fine-tune the large language model, allowing it to learn patterns and features from the annotated examples. During training, the model iterated over the dataset multiple times (epochs), improving its performance by minimizing errors.

The evaluation (validation) dataset was used to assess the performance of different versions of the large language model based on its training. Since it was unseen by the model during training, it served as an unbiased measure to select the best-performing model versions. The test dataset was reserved for reporting the final classification results. Also unseen by the model, it was used to evaluate the generalizability and robustness of the trained model, providing an estimate of how well the model would perform when deployed in practical settings. Table 2 summarizes the distribution of annotated messages across the 3 datasets.

XSL•FO
RenderX

Table . Distribution of smoking-cessation message intents across the training, validation, and test datasets used in this intent detection study (United States, 2012 - 2014 and 2016 - 2019).

Intent category	Train dataset (n=46,647)	Evaluation dataset (n=15,550)	Test dataset (n=15,219)
cigsmell	328	109	76
cravings	1446	482	267
ecigs	205	68	73
fail	774	258	152
greetings	1828	610	490
health	847	283	265
nrt_dontwork	307	102	72
nrt_dreams	258	86	54
nrt_howtouse	354	118	105
nrt_itworks	615	205	135
nrt_mouthirritation	140	46	21
nrt_nauseous	122	41	23
nrt_od	81	27	21
nrt_skinirritation	155	51	35
nrt_stickissue	188	63	83
off-topic	26,093	8698	9664
quitdate	1279	427	318
savingmoney	455	152	90
scared	293	98	83
smokefree	3072	1024	818
smokingless	285	95	33
stress	508	169	148
support	6330	2110	1983
tiredness	153	51	43
weightgain	531	177	167

Ethical Considerations

The messages annotated in this study came from two mobile health, smoking-cessation intervention trials, which were supported by U.S. National Institutes of Health grants R34 and R01. The R34 (conducted 2012 - 2014) received human subjects (HS) approval through the University of California, Irvine (UCI) institutional review board (IRB) HS# 2010 - 7990 and the University of California (UC) Reliance 261: “Twitter-Enabled Mobile Messaging for Smoking Relapse Prevention.” The R01 (conducted 2016 - 2019) received human subjects approval through UCI IRB HS# 2014 - 1303 and UC Reliance 977: “Twitter-Enabled Mobile Messaging for Smoking Relapse Prevention (Phase 2).” Participants were compensated for enrolling with 8 weeks of NRT (approximate value US $200) and up to US $110 for survey completions. For this secondary data study, messages were analyzed verbatim after removing all identifying information, such as usernames. Participants provided informed consent for the primary data collection, and the UCI IRB allows for secondary data analysis of deidentified data without additional consent.

Large Language Model Without Fine-Tuning

The first stage of our methodology was classifying message intents using a large language model without any fine-tuning. In other words, the model was tasked with classifying text into the observed message intent categories without any exposure to messages annotated with these intent categories in a training dataset. As a common example, a large language model might be prompted to classify the intent of a message stating, “I am very sad” as either positive or negative in sentiment, and it might correctly identify the intent as negative, despite not having been trained on the specific dataset. Since large language models are trained on vast and diverse text corpora, they have acquired an understanding of syntax, semantics, grammar, and sentiment, making them well-suited for general-purpose language tasks such as intent detection, sentiment analysis, and topic summarization.

When a large language model is used for a classification task without fine-tuning, it is guided through a prompt to emulate the task. For example, in the classification task example above, a suitable prompt might be “Classify the sentiment of the

XSL•FO
RenderX

following text into one of the two categories {positive, negative}: I am very sad," which is presented to the large language model. Then the model uses its implicit learning and reasoning to perform the task. The model works by predicting the most probable next word in the sequence. For instance, in this example, the language model may evaluate the prompt by considering potential sentiment categories such as positive and negative. Drawing on its prior knowledge and contextual understanding, it may assign a higher probability to negative based on the input text and subsequently predict negative as the final classification. The main advantage of this approach is that it does not require annotation of message intents. The only requirement is to provide the test dataset to the large language model, which then performs the classification based solely on its prior knowledge. This approach saves time and computation resources.

However, the concern is that the domain-specific task may involve text that is not part of the dataset used for pretraining the model. For example, in the domain of smoking cessation, the intent behind the messages can often be highly domain-specific as health-intervention communications tend to be confidential. A message such as "I used gum today and wanted to smoke two hours later," might be interpreted by a language model as indicating a cessation failure. However, within the context of our smoking-cessation dataset, this message corresponds to a more specific intent, nrt_dontwork, which signifies that the user perceived NRT as ineffective.

Thus, knowledge of the domain becomes crucial for getting higher accuracy with large language models. When a model lacks domain-specific knowledge, its performance on related tasks tends to decline. Moreover, in the absence of domain-specific knowledge, large language models may introduce biases or errors stemming from the datasets on which they were originally trained, possibly causing hallucinations. Furthermore, large language models without domain knowledge tend to struggle with classification tasks involving numerous similar intent categories, such as our categories of nrt_dontwork and nrt_itworks, which refer to NRT being ineffective or effective, respectively. While these intents have completely different meanings, large language models that lack fine-tuning often cannot distinguish between them.

In this study, we assessed the performance of a large language model without fine-tuning at completing the task we required for our mobile health application: recognizing the 25 main intents of messages related to smoking cessation, which included numerous nuanced messages about using NRT. We used the state-of-the-art but off-the-shelf Llama-3 8B by Meta. We supplied this model with descriptions of all 25 of our intent categories, as defined in Table 1, to facilitate its classification process and ensure that its predictions were constrained to these predefined intent categories.

Large Language Model With Fine-Tuning

The next stage of our methodology was fine-tuning the same large language model, Llama-3 8B, using our domain-specific dataset with smoking-cessation messages. Fine-tuning a large language model refers to further training the model on one or more datasets from the targeted domain, the goal being to transform a general-purpose model into one specialized for the domain for improved performance. Domain-specific knowledge is often crucial for large language models because the datasets used in the initial basic training often lack the detailed context relevant to the specific task. Fine-tuning enables the models to learn domain-relevant nuances, semantics, and reasoning.

For example, during fine-tuning, the model might be exposed to messages about NRT gum aiding or not aiding with smoking cessation, for instance, "I used gum today and wanted to smoke two hours later." Fine-tuning allows the model to learn subtle distinctions and accurately differentiate between semantically similar message intents. For instance, a model without fine-tuning might incorrectly predict the message intent here as fail (failure at smoking cessation). However, with fine-tuning on the domain-specific dataset, the model should learn that the user is talking about their NRT gum failing to help them with nicotine cravings, and therefore the correct intent is nrt_dontwork.

On the other hand, fully training a large language model on a domain-specific dataset risks the model forgetting its previously acquired general knowledge. Consequently, the model might forget critical knowledge related to language comprehension, reasoning, and generalizing across tasks. Therefore, instead of completely retraining the model, we undertake fine-tuning because it preserves and often freezes the original model weights (ie, the models' prior knowledge) while learning new weights (new knowledge) that is built upon the existing foundation. So, this approach enables the model to retain its previously acquired knowledge while simultaneously incorporating new insights from the domain-specific dataset.

Therefore, we fine-tune the Llama-3 8B large language model on our smoking-cessation dataset. We selected Llama-3 8B to balance computational efficiency and model performance with feasibility for in-house fine-tuning and inference, considering our resource constraints. Meta's Llama-3 8B, when used in causal mode, is designed to predict the next word in a sequence sequentially. When used in this mode, intent detection requires specialized prompting, for example, "Input: I am feeling very sleepy. So the intent of this message is:." This prompt is sent to the model which then predicts the next probable word as tiredness. This prompt can be formulated in various ways; however, the approach has several limitations. It is highly dependent on the prompt design and it operates indirectly through generation, which reduces efficiency, makes end-to-end training more challenging, and necessitates careful prompt engineering and output parsing. Often, the model's output may fall outside the predefined intent categories; for instance, in the previous example where the main intent is tiredness, the model might generate a related but nonfocal intent such as drained.

For classifying message intents, we used a version of the Llama-3 model that is specifically adapted for classification rather than using the standard causal mode designed for language generation. Instead of predicting the next word in a sentence, this version of the model learns to assign entire messages to intent categories. To achieve this, we added a lightweight classification layer which works as the category chooser producing the final model results. This layer enables the model

XSL•FO
RenderX

to directly select the most appropriate category for each input message. Importantly, this approach allows us to begin to address the issue of dataset imbalance, where some intent categories are more common than others. By assigning more weight to underrepresented categories during fine-tuning, the model becomes more attentive to less frequent but clinically important message types.

Additionally, this approach gives us a probability score for each category, which helps us understand how confident the model is in its predictions. This is valuable for downstream decision-making and human oversight. Figure 2 provides an overview of this fine-tuning process, showing that the model's outputs feed into the category chooser, which then predicts the most appropriate intent category. During fine-tuning, we assign more importance to underrepresented categories when the model makes mistakes. This is done by adding category weights to the penalty for mistakes, helping the model pay more attention to rare but clinically important intent categories. We fine-tune the Llama-3 8B model in this way to improve its ability to detect nuanced user intents in smoking-cessation conversations, for example, related to the use of the free NRT provided as a cessation aid.

Figure 2. Large language model fine-tuning used in this intent detection study (United States).

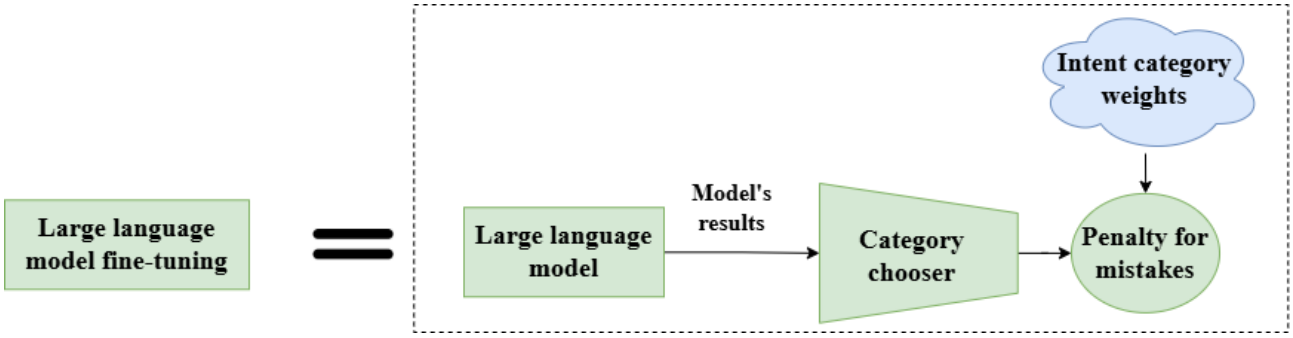

Large Language Model With Fine-Tuning and Downsampling

The third stage of our methodology was downsampling the dominant off-topic (domain-unrelated) message intent category and then fine-tuning the model on the updated dataset. Dataset imbalance is a typical problem in classification tasks, meaning that the number of instances (samples or examples) often varies quite substantially across the different focal categories. For example, in our dataset, the off-topic intent category contained 44,445 samples, whereas the nrt_od category contained only 129 samples. In a multicategory classification task, if a particular category has a much higher number of samples than the other categories, the model being trained tends to be biased toward choosing the category with the higher number of samples. In other words, in the training dataset, the model sees many more samples of this majority category (eg, off-topic) and ends up favoring it in its predictions. Thus, the minority categories (eg, nrt_od) are underpredicted.

For our application which is to accurately predict message intents in smoking-cessation support groups, both majority and minority intent categories require accurate predictions. But, if we train the model with all our data, the model will tend to be more accurate in predicting the highly dominant off-topic category and less accurate in predicting the minority categories with fewer samples. Additionally, the off-topic intent category refers to messages considered irrelevant or uninformative for smoking cessation or support group bonding; it falls outside the 24 key intent categories. Knowing this gives us even more leeway to address our severe category imbalance by reducing the volume of off-topic category examples. Doing this will not compromise the data from the other 24 intent categories.

So for our analysis, we have downsampled the off-topic intent category. Downsampling is a method used to mitigate category imbalance within a dataset by decreasing the sample size of the majority category or categories to better align with sample sizes of the minority categories. We downsampled the instances of the off-topic category to mirror the average number of instances across the other 24 categories. We consistently applied this approach to our training, test, and evaluation datasets. To determine the appropriate level of downsampling for the large off-topic category, we experimented with several approaches, evaluated them on the evaluation dataset, and selected the best-performing approach, which was the average instances across all other category data. These results are reported in Multimedia Appendix 1. After this, we fine-tuned the model using the same approach as in the Large Language Model With Fine-Tuning section. After downsampling the off-topic intent category, the model was fine-tuned using category weights to penalize misclassification of underrepresented categories, ensuring they received sufficient attention during fine-tuning. Although downsampling reduced the disproportionality of the off-topic category, the dataset remained imbalanced, and further downsampling of clinically meaningful categories was not desirable due to the risk of losing valuable information. Thus, category weighting was applied to enhance learning of the rare but important categories while maintaining balance in our fine-tuning process.

The typical limitations of downsampling include loss of information, reduced generalizability, and degraded intent-detection performance in the dominant category. However, these concerns were less applicable in our setting, as we only downsampled the off-topic intent category. This category is unrelated to the 24 focal intent categories and largely represents noise, so reducing its representation is less problematic. In other words, failing to detect some off-topic messages is less harmful than misclassifying important on-topic messages. Even if the detection of off-topic intents decreases, as long as the 24 on-topic intents are accurately recognized, the

XSL•FO
RenderX

system remains effective. Misclassifying an off-topic message intent as 1 of the 24 focal intents keeps the chatbot response within the smoking-cessation domain, which is desirable.

Large Language Model With Fine-Tuning, Downsampling, and Error Correction

The final stage in our methodology was downsampling the dominant off-topic intent category, fine-tuning the model, correcting human annotation errors flagged by examining model-human disagreements in the intent classifications, and fine-tuning the model again. Since the datasets were manually annotated, they were susceptible to human annotation errors. Though large language models are powerful tools due to their training on extensive datasets, fine-tuning them on imperfectly annotated or noisy data can negatively impact their learning and classification performance.

Model-human disagreement on a message's intent category can reveal a weakness in either the model or the human annotation. The model could have incorrectly classified the user's intent, the human annotators could have done so, or possibly both. Therefore, identifying model-human disagreement over how to classify message intents and resolving the human-caused annotation errors enables us to improve the final dataset we use for model fine-tuning. If the model and humans disagree, reviewing these disagreements helps us identify human annotation mistakes, for example, due to messages being ambiguous or vague, or general human oversights in annotations. The human annotation errors can then be corrected in the dataset, and fine-tuning can proceed with less noisy data.

Thus, we analyzed model-human disagreements across the downsampled training, evaluation, and test datasets to identify potential human annotation errors. A domain expert independently reviewed each flagged case and corrected original annotations only when the original human one was incorrect; the model's prediction was never adopted as the new annotation. Only the corrected training dataset was used for fine-tuning again along with the corrected evaluation dataset, and the corrected test dataset remained a true holdout throughout this process. The final performance scores were computed on the expert-validated test dataset, ensuring that the scores remained unbiased and free of data leakage. This approach also ensured consistency across training, evaluation, and testing, as all datasets were corrected using the same procedure and verified by the same domain expert. Category weighting was also applied in this method to enhance the model's learning of rare but important categories while ensuring a balanced fine-tuning process.

Fine-Tuning Hyperparameters

For reproducibility, we provide details on our fine-tuning and inference setup here. We fine-tuned the Llama-3 8B model for 25-category sequence classification using 4-bit quantized low-rank adaptation (QLoRA) within the parameter-efficient fine-tuning (PEFT) framework. Low-rank adaptation (LoRA) adapters (rank=64, α=16) were applied to the query projection (q_proj), key projection (k_proj), value projection (v_proj), and output projection (o_proj) layers, with a dropout rate of 0.05. The category chooser (classification head) is a single linear layer (fully connected) that maps the large language model's hidden dimension of 4096 to the 25 intent categories. Fine-tuning was performed in 32-bit floating point (FP32) precision with 4-bit quantization for memory efficiency along with the Adaptive Moment Estimation With Weight Decay (AdamW) optimizer with a learning rate of 1×10^{-4} , weight decay of 0.01, batch size of 8, and 5 epochs. Early stopping with a patience of 7 monitored unweighted average precision was used to load the best model. Category (class) weights were set inversely proportional to label frequency (normalized) to address category imbalance. Fixed random seeds (42, 123, and 2023) were used for both CPU (central processing unit) and GPU (graphics processing unit), with deterministic Compute Unified Device Architecture (CUDA) operations for reproducibility on the NVIDIA GeForce RTX 4090 GPU. Tokenization used a maximum sequence length of 512, padded to the model's end-of-sequence (EOS) token. Evaluation and checkpoint saving occurred every 1000 steps. During inference, computation was performed using Brain Floating Point 16 (bfloat16) precision with 4-bit quantization for memory efficiency.

Results

Evaluation Metrics

To evaluate the different large language model training approaches for recognizing message intents, we used 3 standard accuracy measures:

Precision: How often the model was correct when it predicted a certain intent.

$$\text{Precision}=\frac{\text{Correct predictions for an intent}}{\text{Total predictions made for that intent}}$$

Recall: How often the model successfully recognized an intent when a message conveyed it.

$$\text{Recall}=\frac{\text{Correct predictions for an intent}}{\text{Total actual cases of that intent}}$$

F_1-score: A numeric combination of precision and recall.

$$\text{F1}-\text{score}=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

Aggregate Performance

Figure 3 summarizes the performance of the Llama-3 8B model across the 4 training methods. The downsampling-only and downsampling with error correction methods involved modifying the training dataset through reduction of off-topic intent samples and annotation error correction, respectively, so we evaluated all 4 training methods using 3 distinct test datasets. Evaluation results are shown in Figure 3A using the downsampled uncorrected test dataset, in Figure 3B using the downsampled corrected test dataset, and in Figure 3C using the full unmodified test dataset.

XSL•FO
RenderX

Figure 3. Overall unweighted and weighted precision, recall, and F_1-scores for 4 large language model training methods for intent detection in smoking-cessation messages in this study (United States). Methods were evaluated on the (A) downsampled uncorrected test dataset, (B) downsampled corrected test dataset, and (C) full test dataset.

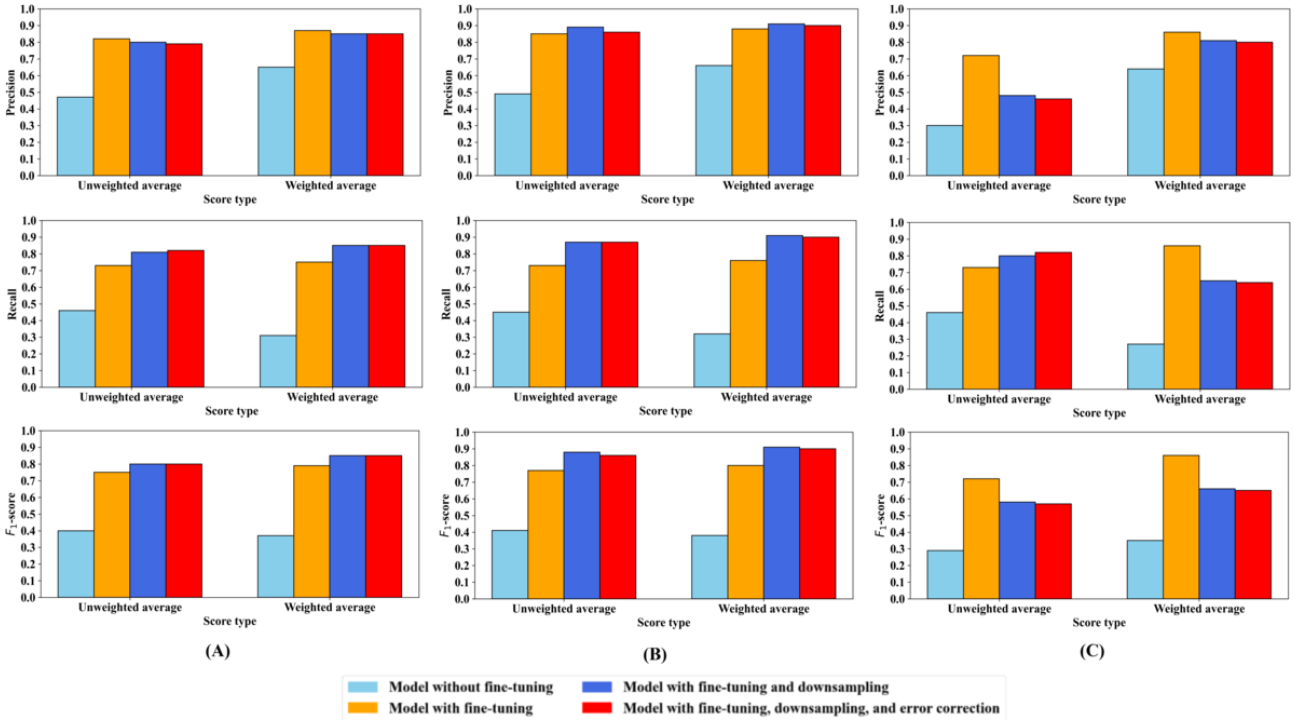

In Figure 3A, which uses the downsampled uncorrected test dataset, the off-the-shelf model with extensive pretraining but no domain-specific fine-tuning performed poorly at recognizing message intents. It achieved unweighted F_1-score, precision, and recall of 0.40, 0.47, and 0.46, respectively, and weighted F_1-score, precision, and recall of 0.37, 0.65, and 0.31, respectively. The unweighted scores assigned equal importance to each intent category, while the weighted scores gave greater weight to categories with more instances. The low scores highlight the complexity of the classification task, demonstrating that intent classification in the mobile health context studied was not straightforward and required specialized fine-tuning to achieve adequate performance. The domain-specific fine-tuning method achieved a much-improved unweighted F_1-score, precision, and recall of 0.75, 0.82, and 0.73, respectively, and weighted F_1-score, precision, and recall of 0.79, 0.87, and 0.75, respectively. The method of fine-tuning after downsampling the dominant off-topic intent category improved the unweighted F_1-scores from 0.75 to 0.80 and yielded unweighted precision and recall of 0.80 and 0.81, respectively. The weighted F_1-score, precision, and recall were all about 0.85. The fine-tuning, downsampling, and error correction method performed comparably to the downsampling-only method, achieved unweighted F_1-score, precision, and recall of 0.80, 0.79, and 0.82, respectively, and similar weighted scores.

In Figure 3B, which uses the downsampled corrected test dataset, the method with no fine-tuning obtained an unweighted F_1-score, precision, and recall of 0.41, 0.49, and 0.45, respectively, and a weighted F_1-score, precision, and recall of 0.38, 0.66, and 0.32, respectively. Fine-tuning yielded an unweighted F_1-score, precision, and recall of 0.77, 0.85, and 0.73, respectively, and a weighted F_1-score, precision, and recall of 0.80, 0.88, and 0.76, respectively. The relatively lower recall indicates the model still struggles to consistently identify certain intents when present. The downsampling and fine-tuning method achieved an unweighted F_1-score, precision, and recall of 0.88, 0.89, and 0.87, respectively, and a weighted F_1-score, precision, and recall of 0.91 for all three. The method of downsampling, fine-tuning, and error correction achieved an unweighted F_1-score, precision, and recall of 0.86, 0.86, and 0.87, respectively, and a weighted F_1-score, precision, and recall of 0.90 for all three. Although error correction yielded good performance, the downsampling-only approach achieved slightly higher unweighted precision and higher weighted scores.

In Figure 3C, which uses the full test dataset, the method without fine-tuning achieved an unweighted F_1-score, precision, and recall of 0.29, 0.30, and 0.46, respectively, and a weighted F_1-score, precision, and recall of 0.35, 0.64, and 0.27, respectively. Among all test cases, 991 predictions could not be matched to any of the 25 intent categories (eg, the model predicted “stick” instead of “stickissues”); these samples were included in the final scoring as misclassifications. The fine-tuning method achieved an unweighted F_1-score, precision, and recall of 0.72, 0.72, and 0.73, respectively, and a weighted F_1-score, precision, and recall of 0.86 for all three. The fine-tuning and downsampling method attained an unweighted F_1-score, precision, and recall of 0.58, 0.48, and 0.80, respectively, and a weighted F_1-score, recall, and precision of 0.66, 0.81, and 0.65, respectively. The final method that added error correction achieved an unweighted F_1-score, precision, and recall of 0.57, 0.46, and 0.82, respectively, and a weighted F_1-score, precision, and recall of 0.65, 0.80, and 0.64,

XSL•FO
RenderX

respectively. Although both methods demonstrated comparable overall performance, the error correction method yielded a higher unweighted recall, indicating its robust performance at detecting intents in real-world, noisy datasets. Using the full uncorrected test dataset lowered the scores for our last two methods due to the large number of off-topic intents in the test dataset, which had been downsampled in the training dataset. This led to many off-topic intents being wrongly predicted as in-domain intents; however, this was acceptable to us as the main goal was to accurately identify the 24 in-domain intents. Misclassifying off-topic intents as in-domain allowed for in-domain chatbot responses focused on the smoking-cessation domain. Moreover, even when we used the full test dataset, our last two methods sequentially improved recall. These improvements suggest that our last training steps helped the model learn to identify the relevant in-domain intents more effectively in the noisy real-world data, which was our main goal. In other words, unweighted recall served as the more informative evaluation metric for our application.

In Figures 3A and 3B, the differences between weighted and unweighted scores were relatively small because of the downsampled off-topic intent category. However, in Figure 3C, the discrepancy was more pronounced because the full test dataset contained a large number of off-topic intent samples, and in the weighted scores, each intent category's contribution is proportional to its sample size, so more frequent intents are treated as more important. In our application, the unweighted recall scores are particularly important, as these give equal importance to all intents regardless of their frequency.

To assess result consistency, we fine-tuned our final method of downsampling with error correction using three random seeds (42, 123, and 2023) to account for variability in model initialization. We observed a mean unweighted F_1-score of 0.85 (SD 0.01) and a mean weighted F_1-score of 0.89 (SD 0.01) on the downsampled corrected dataset, demonstrating high consistency. Additionally, for our last two methods, we used bootstrap resampling, a statistical method that repeatedly evaluates the model on slightly different versions of the test dataset to estimate performance variability. For the downsampling and error correction method, the unweighted F_1-scores ranged from 0.85 to 0.88 with 95% confidence. For the downsampling-only method, the unweighted F_1-scores ranged from 0.86 to 0.89 with 95% confidence. Both results indicate stable performance.

We conducted a bootstrap resampling analysis with replacement (1,000 resamples; random seed=42) on the full test dataset and computed the unweighted-average recall for the downsampling-only and the downsampling and error correction methods. A paired 2-tailed *t* test on the bootstrap differences indicated that error correction significantly improved unweighted-average recall (t_{999}=69.5, P<.001). The 95% CI for the improvement was 0.0022-0.0292, suggesting a small but consistent gain over downsampling alone.

Performance by Message Intent Category

Precision for Each Intent Category

Figure 4 presents the unweighted precision scores for predicting each of the 25 intents across the different training methods. Unweighted precision reflects the proportion of correct predictions among all instances predicted for a given intent and requires balanced performance across intents. Evaluation results are shown in Figure 4A using the downsampled uncorrected test dataset, in Figure 4B using the downsampled corrected test dataset, and in Figure 4C using the full unmodified test dataset.

XSL•FO
RenderX

Figure 4. Unweighted precision scores for the 4 large language model training methods used to detect intents in smoking-cessation messages in this study (United States). Methods were evaluated on the (A) downsampled uncorrected test dataset, (B) downsampled corrected test dataset, and (C) full test dataset.

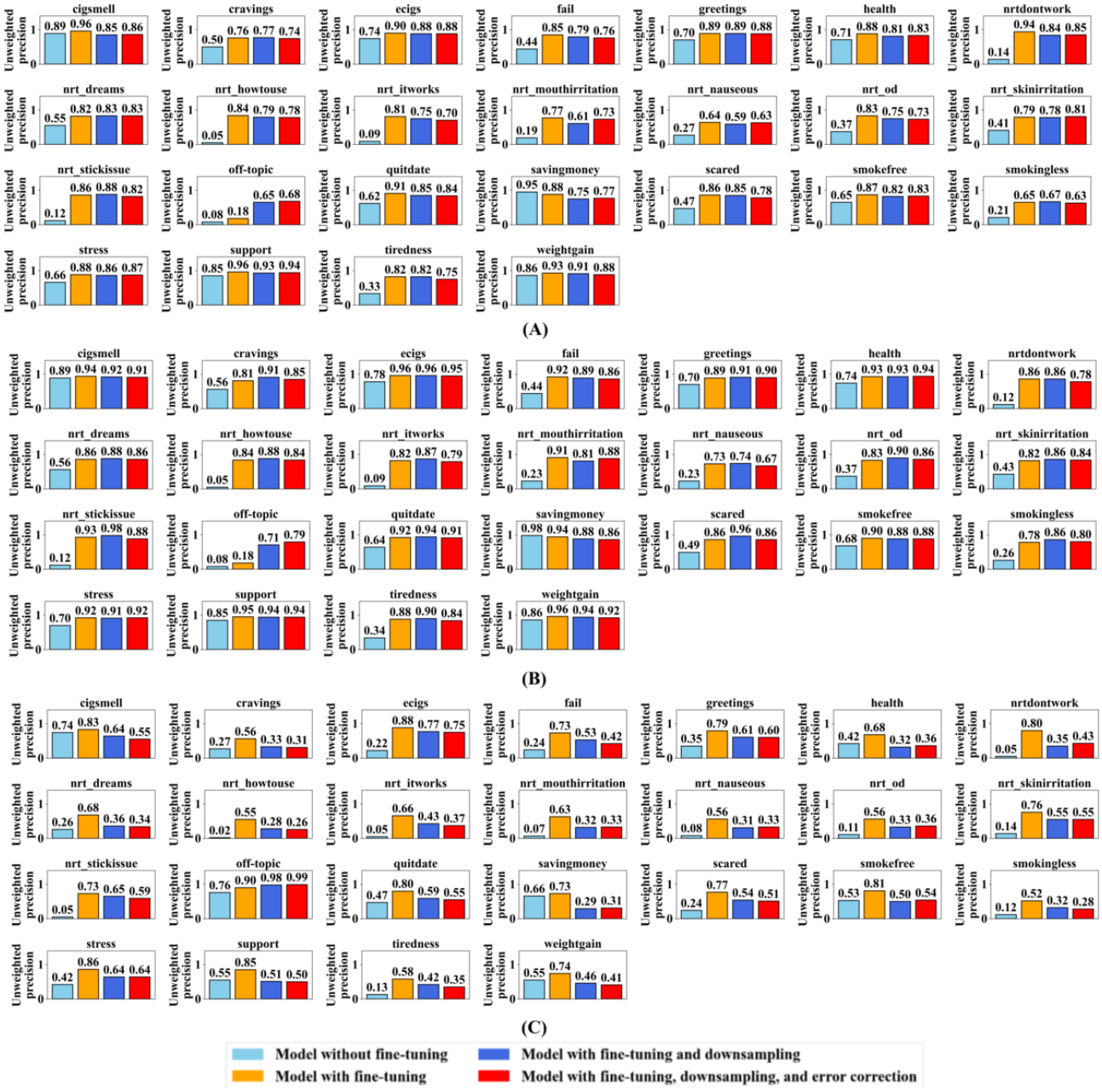

In Figure 4A, using the downsampled uncorrected test dataset, for the method without fine-tuning, the precision for most intents was around 0.47, with a few categories slightly exceeding that score but many falling below it, indicating weak overall performance. With fine-tuning, the precision for most intents improved to around 0.82. However, the dominant off-topic intent achieved a low precision of 0.18, reflecting a classification bias. Its large number of training samples caused the model to overpredict it, reducing precision. In effect, other important intents were often misclassified as off-topic, which is concerning because this can suppress the detection of key intents. The method of fine-tuning and downsampling of the dominant off-topic intent yielded close to a 0.80 precision score for most intents, but certain intents fell below the 0.80 mark. The final approach, combining fine-tuning, downsampling, and error correction, achieved precision comparable to the downsampling-only method.

In Figure 4B, using the downsampled corrected test dataset, the method without fine-tuning exhibited a precision of approximately 0.49 across most intents. With fine-tuning, precision improved to around 0.85 for most intents, although the off-topic intent category continued to show a low precision of 0.18. Fine-tuning with downsampling resulted in many precision scores close to 0.89. Adding error correction yielded an average precision of 0.86. Thus, the downsampling-only method demonstrated more robust performance on precision.

In Figure 4C, using the full dataset, the method without fine-tuning exhibited a precision of approximately 0.30 for most intents. With fine-tuning, precision improved to around 0.72 for most intents, although the dominant off-topic intent achieved

XSL•FO
RenderX

a higher precision of 0.90. Fine-tuning with downsampling resulted in a precision mainly close to 0.48. The final method of combining fine-tuning, downsampling, and error correction achieved an average precision of 0.46. The large volume of off-topic intent samples in the full test dataset led to many off-topic intents being misclassified as in-domain intents, lowering the precision scores for the last two methods.

Recall for Each Intent Category

Figure 5 presents the unweighted recall scores for each of the 25 intents across the different training methods. Unweighted recall shows how accurately the model picks the actual intents and requires balanced performance across intents. Evaluation results are shown in Figure 5A using the downsampled uncorrected test dataset, in Figure 5B using the downsampled corrected test dataset, and in Figure 5C using the full unmodified test dataset.

Figure 5. Unweighted recall scores for the 4 large language model training methods used to detect intents in smoking-cessation messages in this study (United States). Models were evaluated on the (A) downsampled uncorrected test dataset, (B) downsampled corrected test dataset, and (C) full test dataset.

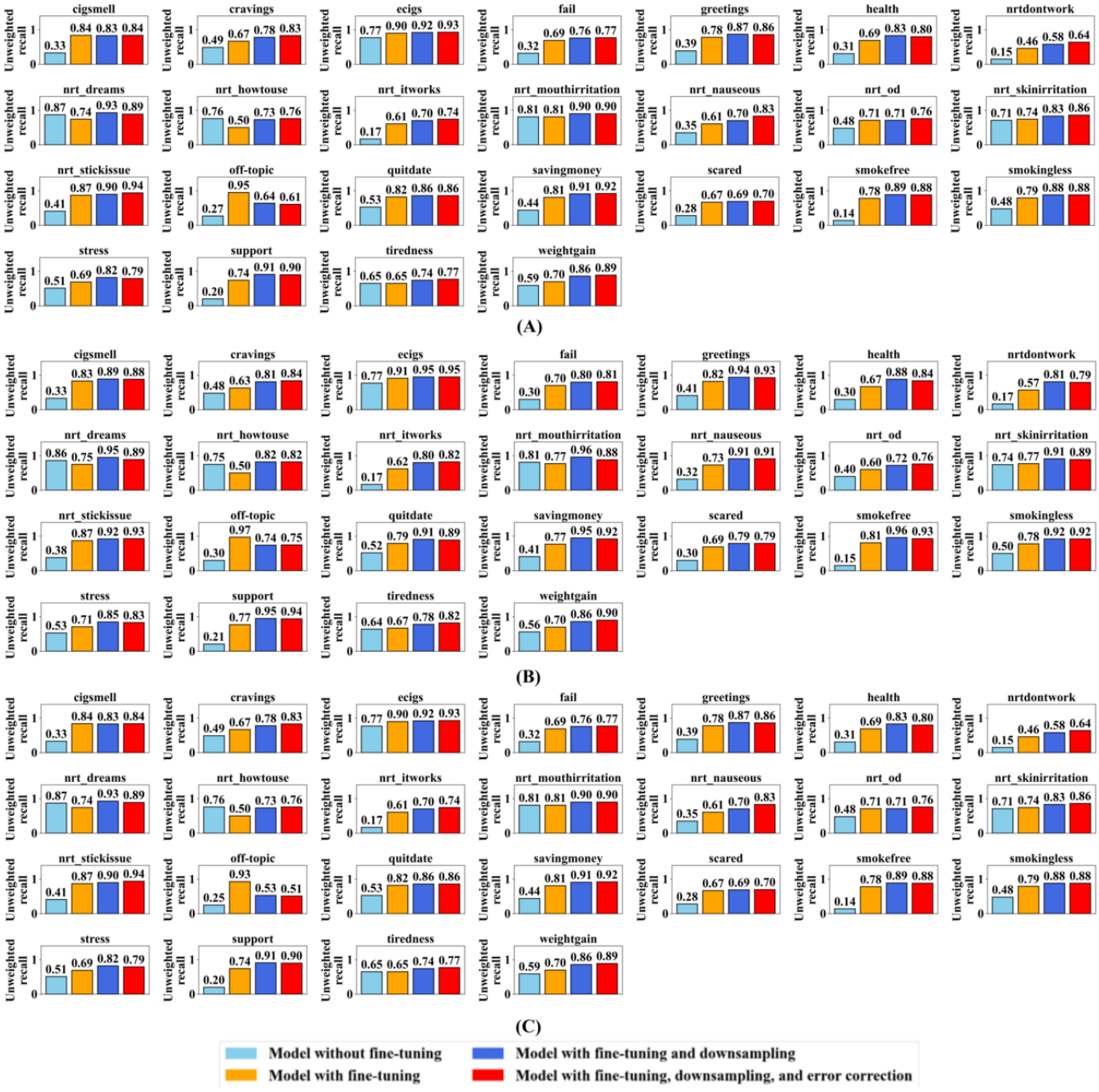

In Figure 5A, using the downsampled uncorrected test dataset, the method without fine-tuning achieved a recall of approximately 0.46 for most intents, with some slightly higher but many lower. Fine-tuning improved recall to around 0.73 for most intents, although several still fell below this, while the dominant off-topic intent category reached 0.95 recall. Adding downsampling, many intents achieved a recall close to 0.81, though some remained lower. Adding error correction as well produced slightly better recall across several intents, averaging 0.82, indicating strong overall performance.

XSL•FO
RenderX

In Figure 5B, using the downsampled corrected test dataset, the method without fine-tuning achieved a recall of approximately 0.45 for most intents. With fine-tuning, recall improved to around 0.73 for most intents, while the dominant off-topic intent category reached 0.97. Adding downsampling, many intents achieved recall close to 0.87, though some remained lower. Adding error correction as well achieved an average recall of 0.87 across intents.

In Figure 5C, using the full test dataset, the method without fine-tuning achieved a recall of approximately 0.46 for most intents. With fine-tuning, recall improved to around 0.73 for most intents, while the dominant off-topic intent category reached 0.93. Adding downsampling, many intents achieved a recall close to 0.80. Adding error correction as well achieved the highest average recall of 0.82 across intents, demonstrating strong overall performance on noisy, real-world data.

F_1-Scores for Each Intent Category

Figure 6 presents the unweighted F_1-scores for each of the 25 intents across the different training methods. Unweighted F_1-score combines precision and recall and emphasizes balanced performance across intent categories. Evaluation results are shown in Figure 6A using the downsampled uncorrected test dataset, in Figure 6B using the downsampled corrected test dataset, and in Figure 6C using the full unmodified test dataset.

Figure 6. Unweighted F_1-scores for the 4 large language model training methods used to detect intents in smoking-cessation messages in this study (United States). Models were evaluated on the (A) downsampled uncorrected test dataset, (B) downsampled corrected test dataset, and (C) full test dataset.

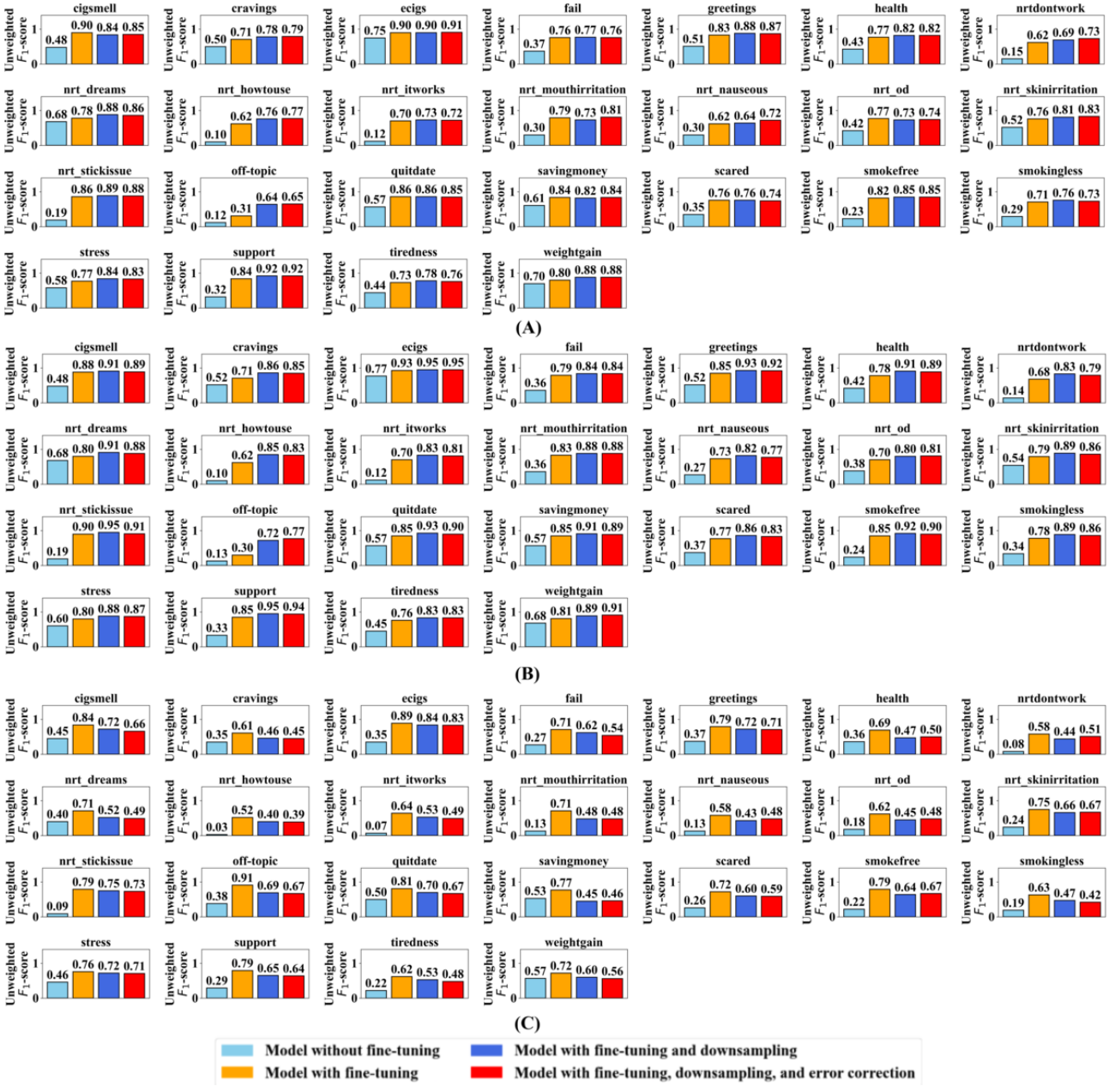

In Figure 6A, using the downsampled uncorrected test dataset, the model without fine-tuning achieved an F_1-score of approximately 0.40 across most intents, although several intents scored lower. Fine-tuning increased the F_1-score to about 0.75,

XSL•FO
RenderX

but many intents scored lower, and the off-topic intent category scored about 0.31. The downsampling method with or without error correction achieved an F_1-score of about 0.80, though some intents were lower.

In Figure 6B, using the downsampled corrected test dataset, the model without fine-tuning achieved an F_1-score of approximately 0.41 across most intents. Fine-tuning improved the F_1-score to about 0.77, though many intents scored lower, and the off-topic intent category scored about 0.30. The downsampling method achieved an F_1-score of about 0.88, but some intents remained lower, while adding error correction achieved F_1-scores of about 0.86 across intents.

In Figure 6C, using the full test dataset, the model without fine-tuning achieved an F_1-score of about 0.29 across most intents, with several lower. Fine-tuning increased the F_1-score to about 0.72, though many intents scored lower, and the off-topic intent category scored above 0.91. The downsampling method yielded an F_1-score of 0.58, while adding error correction yielded an F_1-score of 0.57, but several intents were lower.

Total Statistics on Model-Human Disagreement

Figure 7 shows the model-human disagreement across our 4 training methods. Model-human disagreement provides a comprehensive view of a model's accuracy at judging a message's intent relative to human judges of that intent. The left side of each figure includes either the full or downsampled off-topic intents in the test dataset, while the right side excludes the off-topic intents. Results are shown in Figure 7A using the downsampled uncorrected test dataset, in Figure 7B using the downsampled corrected test dataset, and in Figure 7C using the full unmodified test dataset.

XSL•FO
RenderX

Figure 7. Model-human disagreement, including and excluding the dominant off-topic intent category in the test dataset, for the 4 large language model training methods studied (United States). Methods were evaluated on the (A) downsampled uncorrected test dataset, (B) downsampled corrected test dataset, and (C) full test dataset.

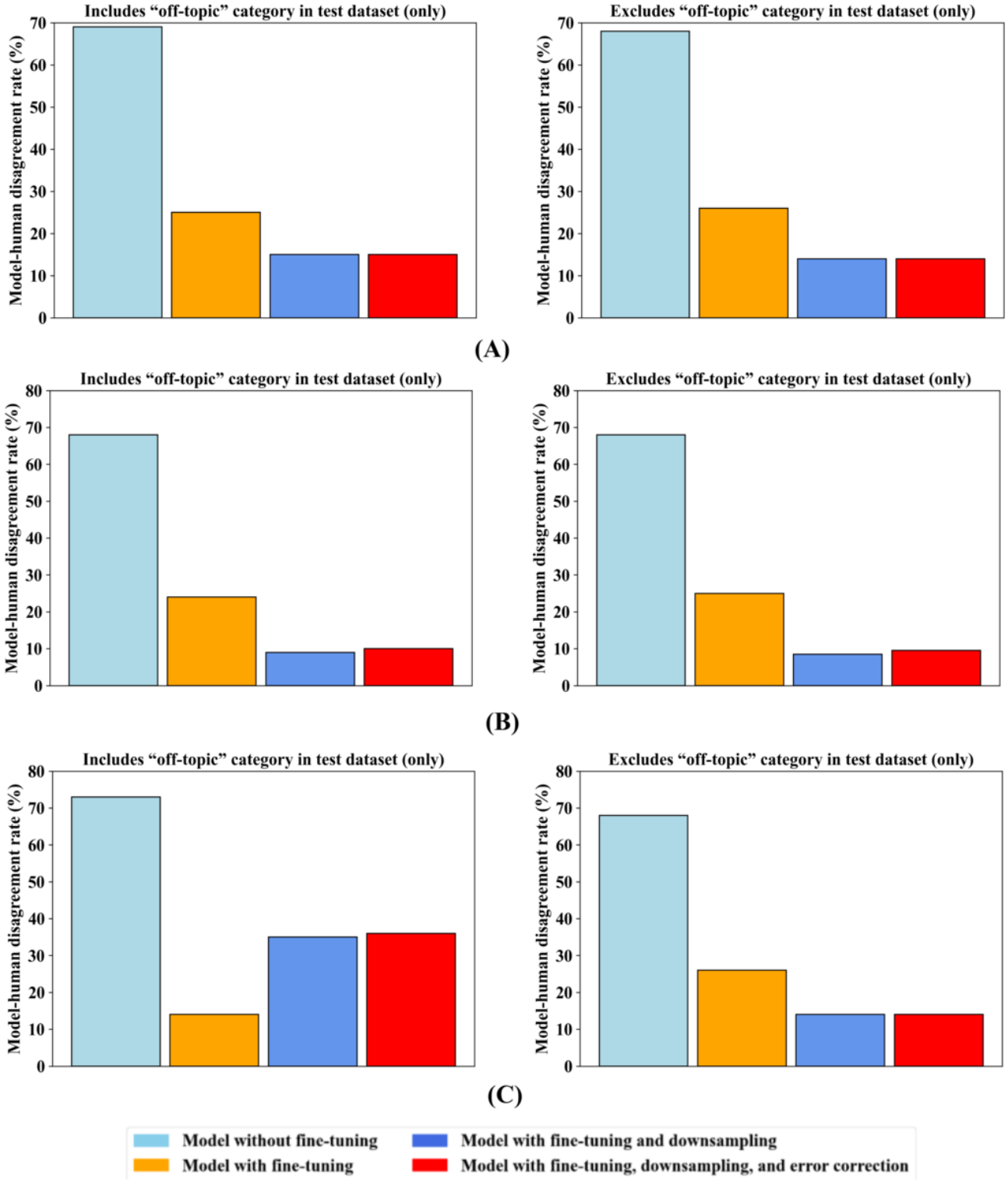

In Figure 7A, using the downsampled uncorrected test dataset, the left side includes off-topic intents in the test dataset. The method with no fine-tuning yielded about 69% model-human disagreement. Fine-tuning alone lowered this to 25%. Fine-tuning with downsampling with or without error correction lowered the disagreement to about 15%. The right side of Figure 7A excludes the off-topic intents from the test dataset. It shows the method without fine-tuning yielded approximately 68% model-human disagreement. Fine-tuning alone reduced this to 26%. Fine-tuning with downsampling with or without error correction lowered the disagreement to about 14%. The results, including or excluding the off-topic intent samples, were similar because the number of off-topic samples in the downsampled test dataset was small enough to have minimal impact.

In Figure 7B, using the downsampled corrected test dataset, the left side includes off-topic intents in the test dataset. The method with no fine-tuning showed about 68% model-human

XSL•FO
RenderX

disagreement. Fine-tuning alone lowered this to 24%, fine-tuning with downsampling lowered it to 9%, and adding error correction yielded 10%. The right side of Figure 7B excludes the off-topic intents from the test dataset. The method without fine-tuning yielded 68% model-human disagreement. Fine-tuning alone lowered it to 25%, fine-tuning and downsampling lowered it to 8.5%, and adding error correction yielded 9.5%.

In Figure 7C, using the full test dataset, the left side includes off-topic intents in the test dataset. The method without fine-tuning exhibited 73% model-human disagreement. Fine-tuning alone reduced this to 14%. In contrast, fine-tuning with downsampling yielded 35%, and adding error correction yielded 36%. So, on real-world noisy data, downsampling yielded more model-human disagreement than no downsampling. This occurred because downsampling the off-topic intent category trained the model to recognize rarer but important in-domain intents. The higher model-human disagreement for the last two methods was primarily due to off-topic samples being misclassified as in-domain intents. At first glance, it may appear that fine-tuning alone performs better because it produces less model-human disagreement in the original noisy data. However, this is largely an artifact of the imbalanced dataset: both the training and test datasets contain a disproportionately large number of off-topic intent samples and, since both are similarly dominated by off-topic, the disagreement rate becomes artificially low, masking poor performance on the more important in-domain intents. The right side of Figure 7C, which excludes the off-topic intents from the test dataset, provides a different and, we believe, more accurate picture. Here, the model without fine-tuning yielded 68% model-human disagreement. Fine-tuning alone lowered this to 26%, and downsampling with or without error correction method lowered it to 14%. These results indicate that when the off-topic intents are excluded from the test dataset, the downsampling-only and the downsampling and error correction methods achieve the lowest model-human disagreement on the 24 important domain-specific intents.

Statistics on Model-Human Disagreement

Figure 8 shows the percentages of model-human disagreement by intent category for our 4 training methods, which we will refer to as model mispredictions because the goal of this analysis is to understand when and how a model confused the different intent categories. This information provides insight into each training method's specific weaknesses. In Figure 8A, evaluation was conducted using the downsampled uncorrected test dataset. In Figure 8B, evaluation was conducted using the downsampled corrected test dataset. Finally, in Figure 8C evaluation was conducted using the full unmodified test dataset.

XSL•FO
RenderX

Figure 8. Model-human disagreement by intent category for the 4 large language model training methods studied (United States). Methods were evaluated on the (A) downsampled uncorrected test dataset, (B) downsampled corrected test dataset, and (C) full test dataset.



In Figure 8A, using the downsampled uncorrected test dataset, the method without fine-tuning caused the model to mispredict the intent nrt_howtouse for many other intent categories. For some categories, over 50% of instances were mispredicted as nrt_howtouse; for 17 out of 25 categories, this incorrect prediction dominated. This demonstrates that the model without fine-tuning was heavily biased toward nrt_howtouse and failed to distinguish it from other intents. A similar issue appeared with the fine-tuned model, but the bias shifted toward the dominant off-topic intent category. Here, 24 out of 25 categories were most commonly and incorrectly mispredicted as off-topic, and this occurred approximately 30% of the time, reflecting the overrepresentation of off-topic samples. This highlights a key limitation of fine-tuning alone: without addressing category imbalance, the model struggles to predict important but underrepresented intent categories accurately. In contrast, the fine-tuning with the downsampling method produced a more balanced pattern of mispredictions. No single intent category was mispredicted, as evidenced by the absence of any overwhelmingly frequent color in our bar graph; only minor weaknesses remained. A similar pattern was observed for the fine-tuning, downsampling, and error correction method, with no single intent category mispredicted. These results demonstrate that downsampling effectively reduced bias toward mispredicting dominant intent categories and promoted more equitable performance.

In Figure 8B, using the downsampled corrected test dataset, the methods without and with fine-tuning alone showed similar patterns, with nrt_howtouse and off-topic intents being most frequently mispredicted, respectively. In comparison, the methods adding downsampling alone, or downsampling with

XSL•FO
RenderX

error correction, showed a more balanced pattern of mispredictions, with no single intent category dominating, the only exception being the support intent category, with 7 different intents frequently misclassified as support, although in each case less than 5% of the time.

In Figure 8C, using the full test dataset, the methods without and with fine-tuning alone showed similar patterns, with nrt_howtouse and off-topic intents most frequently mispredicted, respectively. Adding downsampling and then also adding error correction produced a more balanced pattern of mispredictions, except for the support intent, which was mispredicted as other categories. Examples in which the model mispredicted other categories as support include: “Those are great reasons,” which was human-annotated as off-topic, “Nicotine addiction is like any other drug out there and the side effects are bad. Sorry to say,” which was human-annotated as health, and “Patches will help greatly. You can do it,” which was human-annotated as nrt_itworks. These results show that, in certain misclassification cases, the model produced a generic smoking-cessation support response rather than an intent-specific one. While such responses may provide baseline assistance, they do not substitute for the intended goal of accurate intent-specific classification.

Model-Human Disagreement and Human Error Correction

Human errors in the annotation of message intents were inevitable given the dataset’s very large size. The data annotators were highly trained undergraduate students in public health fields who demonstrated high interrater reliability in training before annotating the study dataset. They met weekly with supervisors to resolve inconsistencies in real time. Despite this rigorous training and oversight, a small number of annotation errors were observed. Therefore, a domain expert thoroughly reviewed the model-human disagreements in the training, evaluation, and test sets to identify and correct human annotation errors.

Table 3 presents statistics on model-human disagreement on message intents from model mispredictions or human annotation errors or both. These statistics offer insights into the accuracy of human annotators and the potential of a large language model to sometimes outperform humans. Model-human disagreement was assessed in the downsampled datasets only, due to lack of resources to search for disagreement in the full dataset.

Table . Causes of model-human disagreement in the detection of intents in this study (United States).

Dataset (downsampled)	Model-human disagreement	Humans wrong, model correct	Model wrong, humans correct	Both humans and model wrong
Train	0.7%	59%	30%	11%
Evaluation	15%	42%	47%	11%
Test	15%	39%	55%	6%

In the downsampled training dataset (75% of the nontest data), the model and human annotators disagreed on the message intent 0.7% of the time. The human annotators were incorrect in 59% of these disagreed instances, with the model being accurate. The model mispredicted in 30% of the disagreed instances, with both human and model erroneous in the remaining 11% of the instances. Thus, in total, humans were wrong in 70% of the disagreed instances (59%+11%), amounting to 0.13% of the total data, 0.22% of the full training data, and 0.47% of the downsampled training data. So, human annotators were rarely wrong. While the downsized training dataset was not found to include many human annotation errors, it included some, indicating that retraining the model on a corrected dataset would lead to performance improvements.

In the downsampled evaluation dataset (25% of the nontest data, reduced further after downsampling the dominant off-topic intents), the model and human annotators disagreed on the message intent 15% of the time. Human annotators were wrong in 42% of these instances, with the model being accurate. The model mispredicted in 47% of these instances, and both were erroneous in the remaining 11% of the instances. Thus, in total, humans were wrong in 53% of the disagreed instances (42%+11%), amounting to 0.76% of the total data, 3.78% of the full evaluation data, and 8% of the downsampled evaluation data.

Finally, in the downsampled test dataset (holdout data from groups 30 to 36, reduced further after downsampling off-topic intents), model and human annotators disagreed on the message intent 15% of the time. Human annotators made errors in 39% of these cases, with the model being accurate. The model mispredicted in 55% of these instances, and both were erroneous in the remaining 6% of the instances. Thus, in total, humans were wrong in 45% of the disagreed instances (39%+6%), amounting to 0.50% of the total data, 2.5% of the full test dataset, and 7% of the downsampled test data.

Table 4 presents statistics on the corrections made to the downsampled data after manual review of the model-human disagreement in intent detection by category. In total, 1077 samples were corrected, accounting for 1.4% of the total dataset and 3% of the downsampled data, indicating human annotation errors and corrections were relatively rare. In the downsampled training dataset, in 102 cases, the support, smokefree, and greetings intents were corrected more frequently than other intents. In the downsampled evaluation dataset, in 588 cases, the support, smokefree, off-topic, cravings, and greetings intents were corrected more frequently than other intents. In the downsampled test dataset, in 387 cases, the smokefree, off-topic, quitdate, nrt_dontwork, nrt_itworks, and greetings intents were corrected more frequently than other intents.

XSL•FO
RenderX

Table . Data adjustments after reviewing model-human disagreement in smoking-cessation message intents in this study (United States).

Intent category	Adjustments to downsampled training dataset (n=102)	Adjustments to downsampled evaluation dataset (n=588)	Adjustments to downsampled test dataset (n=387)
cigsmell	0	3	5
cravings	9	43	14
ecigs	0	3	2
fail	1	23	7
greetings	10	53	43
health	1	26	18
nrt_dontwork	3	30	20
nrt_dreams	1	4	1
nrt_howtouse	0	21	10
nrt_itworks	3	17	15
nrt_mouthirritation	0	7	1
nrt_nauseous	0	9	5
nrt_od	0	5	3
nrt_skinirritation	0	3	3
nrt_stickissue	0	0	2
off-topic	10	71	33
quitdate	0	14	16
savingmoney	0	10	3
scared	3	6	9
smokefree	17	97	60
smokingless	0	10	1
stress	1	14	7
support	43	116	104
tiredness	0	2	2
weightgain	0	1	3

A domain expert handled the human annotation error correction and found that corrections were often needed when two intents were expressed in a single message. In such cases, the human annotators had been trained to prioritize certain intents over others to optimize the chatbot's responses, but they apparently found this prioritization challenging. For instance, if a person said they quit but were craving, this should have been annotated as cravings, not smokefree, to prioritize a response to help with craving. If a person mentioned they were going to quit on day 1, this should have been annotated quitdate, not smokefree, to prioritize quitdate help. Messages about failing and setting a new quit date should have been annotated as fail, not quitdate, to provide guidance after a fail. If a message started with a greeting, for example, good morning, but then provided quitting support, it should have been annotated as support to prioritize a support-related response over a generic greeting. If a message discussed quitting but also reported an issue with NRT, it should have been annotated as NRT-related to assist with NRT use. If a person mentioned no longer needing NRT because they had successfully quit using NRT, it should have been annotated as nrt_itworks, not as a generic quit. All annotation corrections were conducted blind to the model predictions; that is, the domain expert reviewed and adjudicated the intent annotations independently of model outputs.

Overall, this analysis revealed the value of using large language models not only for classifying message intents but also for uncovering human errors in annotating these intents. The relatively high rate of model correctness in cases of human misannotations reinforces the idea that annotation corrections followed by model retraining can result in more accurate assessments of the model's performance. However, in our work, the observed impact on model performance was minimized because we only had the resources to correct human annotation errors in the downsampled data, not the full data. Additionally, as human-model disagreement and human annotation correction often involved a message with two intents, different methods of addressing this situation should be considered, as will be discussed below.

XSL•FO
RenderX

Discussion

Principal Findings

Large language models have become the new state-of-the-art models in natural language processing due to their ability to recognize complex linguistic patterns and subtle semantic distinctions. Leveraging large language models for classification tasks like recognizing message intents can significantly enhance performance over earlier models. However, our analysis indicates that large language models may be insufficient for intent recognition in mobile health interventions when user messages are short and subtle wording differences can profoundly affect meanings. We studied messages in mobile health smoking-cessation support groups, where there was a profound difference between saying “quit” or “not yet quit,” “NRT works” or “NRT doesn’t work,” and “quit day 1” or “quit day 21,” and our off-the-shelf large language model was not always able to detect these nuances. The model’s performance was unacceptably poor until it was fine-tuned on a domain-specific dataset, which required laborious human annotation of message intents. The model before fine-tuning performed poorly because the domain-specific text often contained unique terminology, implicit meanings, and contextual subtleties that a general-purpose model was unable to fully grasp without exposure to relevant training data.

We found that fine-tuning the large language model on a domain-specific dataset led to substantial performance gains. The fine-tuning process allowed the model to familiarize itself with the characteristics of the dataset and better capture the intents behind the messages. Our fine-tuned large language model substantially outperformed the same model without fine-tuning. But a key challenge we identified during fine-tuning was data imbalance, particularly due to the dominance of off-topic message intents unrelated to the focal domain of smoking cessation. This imbalance caused the model to overpredict this majority (dominant) off-topic intent, degrading performance on all other intents, even though we assigned more importance to underrepresented categories when the model made mistakes. To address this issue, we implemented downsampling of the majority off-topic intent before fine-tuning, which increased unweighted F_1-scores (overall performance) by 7% and 14% in the downsampled uncorrected dataset and downsampled corrected dataset, respectively, indicating a substantial improvement in balanced performance across all intents. Individual intent-level F_1-scores also demonstrated consistent gains. When we evaluated our methods on the full uncorrected test dataset, the unweighted F_1-score decreased when we added downsampling, primarily due to the large number of off-topic intents in that full dataset, which adversely affected precision. However, downsampling improved unweighted recall by approximately 10% in that full dataset, indicating more robust performance in correctly identifying in-domain, clinically relevant intents.

Another critical insight from our study was the identification of dataset inconsistencies and noise, largely arising from human annotation errors in intent labeling. To address this issue, we examined the model-human disagreement, identified human annotation errors, and corrected them. Even when the downsampled test dataset remained uncorrected, training on the corrected dataset slightly increased unweighted recall, suggesting improved generalization to noisy real-world data, and the F_1-score was preserved. When the downsampled test dataset was itself corrected, the model trained with error correction achieved the same unweighted recall as no error correction, but the latter yielded approximately 3.5% higher precision, slightly improving the F_1-score. This finding indicates that the no error correction method generalized marginally better to the corrected dataset than the error correction method.

This last outcome may indicate the error correction process, while beneficial for improving unweighted recall under noisy conditions, may have reduced useful diversity or led to minor overfitting. Another reason could be that most corrections involved the majority of intents such as support, greetings, and smokefree, and these samples were corrected into rarer intents. As a result, the error-correction method focused slightly more on the rare intent categories, achieving similar recall as the downsample-only method, but lower recall on the majority classes, reducing precision and the overall F_1-score. On the full test dataset, the unweighted F_1-scores of downsample-only and downsample with error corrections were comparable; however, error correction achieved a 2.5% higher unweighted recall and thus showed enhanced identification of in-domain intents under real-world conditions, where incoming data are noisy.

Comparison With Previous Work

Previous work [20] studied intent classification with the same dataset, comparing random forest to BERT, which in 2018 was a revolutionary large language model introduced by Google. The study dataset, annotated by humans to indicate the different message intents, was identical to that used here, except the two smallest intent categories were combined and some duplicate messages were inadvertently retained. The findings showed that BERT, fine-tuned on the domain-specific annotated dataset, outperformed random forest. BERT yielded unweighted and weighted F_1-scores of 0.67 and 0.84, respectively, which were reasonable scores but inadequate for implementation of the large language model in our mobile health intervention. In this study, we used a more recent public-domain large language model, Llama-3 8B (Meta), to perform intent classification across 25 message intent categories. When fine-tuned on the same domain-specific annotated dataset, Llama-3 8B achieved unweighted and weighted F_1-scores of 0.72 and 0.86, respectively, on the full test dataset, outperforming the previously reported BERT model by 7% in unweighted and 2% in weighted F_1-scores, respectively. However, these improvements should be interpreted as indicative rather than strictly comparable, due to differences in dataset structure and inadvertent duplication in the prior study.

We then evaluated our last two methods: (1) fine-tuning with downsampling and (2) fine-tuning with downsampling and error correction. When assessed on the downsampled corrected test dataset, these approaches achieved unweighted F_1-scores of 0.88 and 0.86 and weighted F_1-scores of 0.91 and 0.90, respectively. Relative to the BERT baseline, these are

XSL•FO
RenderX

improvements of 31% and 28% in unweighted F_1-scores and 8% and 7% in weighted F_1-scores. We note that these improvements are only indicative, since differences in dataset prevent a strict empirical comparison. However, when evaluated on the full test dataset, our final two methods achieved unweighted F_1-scores of 0.58 and 0.57, and weighted F_1-scores of 0.66 and 0.65, lower than the previously reported BERT performances. This reduction is primarily attributable to the dominance of the off-topic intent category in the full test dataset. Misclassification of off-topic intents as in-domain intents reduced overall precision and the resulting F_1-scores. Despite reduced precision, our final two methods achieved substantially higher unweighted recall scores of 0.80 and 0.82, compared with 0.71 for the prior work, representing a 13% - 16% recall improvement. This finding indicates that our approach is more effective at identifying in-domain, clinically meaningful intents, which is our central objective. However, our model's weighted F_1-score of 0.65 is substantially lower than the previously reported 0.84 for the BERT baseline. This indicates a clear reduction in overall precision and F_1-score performance in a real-world setting, highlighting the trade-off between improved recall for in-domain intents but decreased overall classification accuracy. In addition, a strict comparison with the previous work is not feasible because it used 24 intents after merging two small ones (instead of our 25) and inadvertently contained duplicated data. These differences alter the ground truth distributions and, therefore, the performance differences should be interpreted as observed rather than strictly comparable.

Implementation Feasibility, Generalizability, and Potential Risks

Fine-tuning the large language model based on our final method took approximately 2.5 hours, and each input was processed in approximately 40 milliseconds, with a peak GPU memory usage of 8.2 GB. The model was able to handle approximately 493 tokens per second, demonstrating fast and efficient performance suitable for real-time intent detection. Human oversight was incorporated for human annotation error correction, to improve model accuracy and reliability. This error correction approach was straightforward. Model-human inconsistencies in intent classifications were identified and then systematically reviewed to identify potential human annotation errors, with error corrections performed by a domain expert who was blind to the model predictions. If a domain expert with knowledge of the annotation process or domain context is involved, this error-correction approach should be reproducible.

Regarding the generalizability of our final model, it should perform well on new data that mainly includes the same 25 intent categories. Large language models handle shifts in language use through embedding-based representations, which capture semantic meaning rather than relying on exact word sequences. Within the same domain, this approach mitigates the impact of data drift, ensuring that the model's predictions remain relevant and accurate even if specific word usage has evolved over time. However, if the new data contain different intent categories, it would require reannotation or model adaptation.

The largest failure in our intent-detection model would be if an intent category is detected as a different and unsuitable intent category. For instance, the message "Started my 5th day the struggle is real today" refers to cravings, but the model might mispredict it as smokefree. While it is true that on the 5th day the person is smokefree, the main concern is cravings as there is struggle, so missing the craving intent may prevent the chatbot from responding with the needed support for cravings. Another example is that the message "I'm restarting Friday" refers to fail, but the model might mispredict it as quitdate. Although the message mentions Friday as the quit date, it mainly indicates the person failed to quit and needs help with avoiding a future failure, not with setting a quit date. These examples highlight limitations in intent detection and suggest that our future work should aim to provide appropriate responses even if the predicted intent is incomplete or erroneous or multiple intents are conveyed in a single message.

In our analysis, we downsampled the dominant off-topic intent category to help the model better learn the remaining 24 in-domain intent categories. As a result, the model's performance on off-topic is lower, and these messages are more likely to be misclassified as one of the in-domain categories. In practice, this means that some off-topic messages may receive an in-domain, prevetted smoking-cessation response. Although off-topic messages typically fall outside the scope of smoking cessation and may include general social conversation, misclassifying them as cessation-related is unlikely to cause harm as people signed up for a smoking-cessation support group. Moreover, cessation responses are empathetic, supportive, and encouraging, and therefore still appropriate in tone. However, this high rate of false positives should be considered a potential barrier to user adoption and trust, rather than inherently unproblematic, as we do not have supporting user data to assess its real-world impact.

We acknowledge that repeated false-positive delivery of cessation-oriented content to users engaging in social small talk may risk user disengagement. From a deployment perspective, we consider two complementary definitions of signal-to-noise. When true in-domain detections are considered signal and off-topic detections are considered noise, based on our full test dataset, our downsampling methods would result in nearly 4700 off-topic messages being misclassified as in-domain, compared to 5500 that are truly in-domains. This corresponds to nearly one unnecessary cessation intervention for every appropriate one, resulting in roughly a 1:1 signal-to-noise. Where both true in-domain detections and off-topic detections are considered signals because off-topic messages appropriately redirect people to focus on the smoking-cessation domain, our methods would result in nearly a 2:1 signal-to-noise. In sum, in the real deployment, the model could send from 1 to 2 useful cessation messages for every unnecessary one. This means users engaging in casual or social conversation could frequently receive irrelevant intervention messages. This potential risk should be continuously monitored during real-world deployment to mitigate unintended user disengagement.

It is true that recognizing off-topic intent messages is an important goal for this deployment, as detecting such messages can possibly help redirect focus in the smoking-cessation

XSL•FO
RenderX

domain. However, detecting in-domain intent categories is a higher-priority objective. The original dataset contained a large proportion of off-topic messages, which biased the model toward off-topic detection and, in turn, reduced performance on in-domain messages. To address this tradeoff, we prioritized in-domain detection by weighting it more heavily, while still considering off-topic detection as a secondary goal. Consequently, in our downsampling approach, we reduced the number of off-topic messages to improve the model's performance on in-domain intent categories.

Limitations

Our final method of large language model training that included fine-tuning, downsampling, and error correction achieved impressive performance in our dataset, but it had several limitations. First, a continued challenge is complex messages involving two or more intent categories. For instance, a user stated, "Tomorrow is my day 1 and I haven't smoked for over 5 hours today," which reflects two intents: quitdate (planning to quit tomorrow, ie, day 1) and smokingless (starting to quit but still smoking some). While the main message intent was likely to alert others to tomorrow being the quitdate, the model predicted smokingless. Similarly, in "I'm not going to quit smoking on my busiest day," the message expressed intent to quit on a specific day (quitdate) but also implied being smokefree on that day; however, the correct intent category was quitdate and the model mispredicted it as smokefree. These examples highlight the inherent ambiguity in single-intent annotations when messages express multiple intents. Because the dataset does not account for multi-intent expressions, the model's predictions occasionally diverged from the human annotations despite both being semantically reasonable.

Another study limitation was that 2 of our 4 methods involved downsampling the off-topic intent category in the training, evaluation, and test datasets, based on the average number of samples of all other intents. Downsampling the off-topic intent category consistently across datasets ensured the model learned from a representative distribution of focal intents. However, we recognize that altering the test dataset may be problematic. Therefore, we also evaluated performance on the complete, uncorrected test dataset which likely offers a more accurate reflection of how the model would perform in real-world scenarios.

A third potential limitation of our study is that, in our final method, we corrected human annotation errors. However, we only had the resources to assess the much smaller downsampled dataset for annotation errors, and we only examined instances where the model and the human annotator disagreed. This approach introduces the risk of circular bias, where the ground truth is shifted toward the model's predictions. This may inadvertently penalize natural human variation and inflate performance metrics relative to what the model would achieve on truly independent unmodified data. Moreover, since our error correction method focused only on messages where the model disagreed with human annotations, instances where both the model and human annotators were incorrect (shared errors) were not reviewed, which may leave residual annotation errors in the dataset. As a result, even the corrected dataset may be partially aligned with the model's predictions rather than fully independently verified, which could influence performance estimates. Once again, performance on the full uncorrected test dataset likely provides a more realistic estimate of generalization to real-world conditions.

Fourth, because the training and evaluation data splits were stratified based on the distribution of intent categories rather than at the group or user level, messages from the same user could appear in both datasets, which may introduce data leakage and allow the model to partially learn user-specific writing styles, potentially overestimating performance.

Fifth, while we downsampled the dominant off-topic intent, the dataset remained imbalanced. Rare intents still had very few examples, whereas intents such as greetings and support had many examples, which reduced model performance for underrepresented intents. Finally, our off-the-shelf model with no fine-tuning produced 991 intent predictions that did not map onto any of our existing intent categories. To handle this, we included these samples as misclassifications in the final scores.

Future Work

In future work, we plan to explore how to address multiple intents in messages, for instance, by allowing the model to detect the most probable 2 - 3 intents. We will also explore ways to incorporate general domain knowledge into the model training process, for example, patterns or semantic rules that are not easily learnable from data alone, but can be integrated into the model's decision-making. We may also seek to expand our dataset by collecting additional real-world examples of underrepresented intent categories and/or creating synthetic examples using generative artificial intelligence (AI). However, it will be crucial to ensure that any synthetic data mimics real-world patterns and does not introduce bias. In addition, we may explore temporal analyses using classifiers to address possible temporal variability in the data.

In additional future work, we will explore threshold tuning. For this study, we did not perform threshold tuning or probability calibration because the Llama-3 8B model inherently selects the most probable category based on its prediction probabilities. We were concerned that adjusting thresholds could suppress predictions for certain intent categories when confidence scores were low. In our domain, where many intents are semantically related, the model must still assign a category even with lower confidence to remain contextually helpful. Setting a high threshold might cause the model to abstain from predicting any intent, which would be less useful to us. Although a low-confidence prediction may occasionally be incorrect, it should still tend to remain on-domain and informative. Notwithstanding, threshold tuning should be considered. Specifically, the off-topic intent category can be almost 50% misclassified as on-topic or domain-related, leading to on-topic chatbot responses, and this could cause user disengagement. So, future work should explore confidence-based thresholding or probability calibration to reduce false-positive domain-related chatbot responses. By implementing a mechanism to silence the chatbot when the predicted intent probability is low (eg, not likely to be on-topic), the signal-to-noise ratio could be improved in live deployment. This approach may help minimize

XSL•FO
RenderX

unnecessary domain-specific cessation messages, reduce user disengagement, and enhance overall safety and usability in real-world clinical settings.

We chose the arithmetic mean as a simple and transparent method to reduce the dominance of the off-topic intent category while retaining sufficient examples of it for model learning. Using the mean ensured that the downsampled off-topic intent remained large enough to capture diversity within its messages. But we acknowledge that downsampling to the mean likely resulted in a reduced off-topic intent category that was still significantly larger than many minority intent categories. Thus, alternative strategies such as downsampling to the median could be explored in future work.

Our ultimate research aim is to create an intent detection model that will also respond to messages based on the intents conveyed. Therefore, we still need to create the model's response capabilities, yielding a complete interactive chatbot. We hope this chatbot will improve engagement in our smoking-cessation support groups by ensuring at least the chatbot is available to respond 24/7 if no human is online. However, we acknowledge that our mobile health support group intervention is not a one-size-fits-all approach; for instance, it is not suitable for people who are not interested in mobile health and seek in-person support exclusively.

To translate technical performance into operational impact, we hope to test our chatbot in a randomized controlled trial to see if it significantly improves the outcomes of our mobile health intervention for smoking cessation. We plan to assess the potential lift in engagement due to the inclusion of a chatbot that accurately detects and responds to support group members' messages. We also plan to assess other intervention outcomes, including cessation success.

Conclusions

Mobile health smoking-cessation interventions that provide real-time supportive interactions have the potential to deliver timely advice and improve health outcomes. Detecting user intents from their utterances can enhance personalization and engagement by identifying individual needs and enabling intelligent responses through chatbots, that is, AI agents. However, existing chatbots for smoking-cessation interventions are limited in their intent detection and response capabilities. In this study, we developed a large language model for intent detection and evaluated its performance under multiple conditions: using an off-the-shelf model without fine-tuning, fine-tuning on domain-specific data, fine-tuning with downsampling of the dominant off-topic intent category, and fine-tuning with downsampling combined with correction of human annotation errors identified from model-human disagreement. The off-the-shelf and fine-tuning–only methods demonstrated poor performance in detecting in-domain intents. Fine-tuning with downsampling achieved unweighted and weighted F_1-scores of 0.88 and 0.91, respectively, on the downsampled corrected dataset, and unweighted and weighted F_1-scores of 0.58 and 0.66 with an unweighted recall of 0.80 on the full test dataset. Downsampling with error correction achieved 0.86 unweighted and 0.90 weighted F_1-scores on the downsampled corrected test data, and 0.57 unweighted and 0.65 weighted F_1-scores on the full test dataset. Downsampling the training dataset lowered precision and therefore the F_1-scores on the full dataset due to the high prevalence of the off-topic intent category in the latter. However, downsampling with error correction achieved slightly higher unweighted recall (0.82) on the full test dataset compared with downsampling alone (0.80), indicating only modest improvement in detecting in-domain message intents. This finding suggests the error correction could conceivably enhance performance; though the difference is small, it is statistically significant. Thus, automated downsampling alone may provide sufficient benefit, though the resource-intensive manual error correction process could offer slightly better returns in practice.

Acknowledgments

The authors thank the undergraduate students at the University of California, Irvine, who annotated the messages from our smoking-cessation support groups, and Douglas Calder and Connor Phillips for training and supervising these students.

Funding

For this study, we used messages from two smoking-cessation support-group intervention trials which were supported by NIH grants awarded to CP: R34DA061012 and R01CA204356. The large language model fine-tuning, data downsampling, and error correction work was supported in part by bridge funding to CP and IGH from the University of California Irvine Chao Family Comprehensive Cancer Center and the University of California Irvine Merage Future of Healthcare Fund.

Data Availability

The datasets analyzed for this study are not publicly available because participants in the interventions did not provide approval to release their posts to the public, even if deidentified.

Authors' Contributions

SR's contributions include conceptualization, data curation, formal analysis, methodology, software, validation, visualization, and writing – original draft. CP's contributions include conceptualization, funding acquisition, investigation, project administration,

XSL•FO
RenderX

resources, supervision, and writing – review and editing. IGH's contributions include conceptualization, methodology, project administration, resources, supervision, and writing – review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables and figures illustrating the study's evaluation process.
[PDF File, 1224 KB - jmir_v28i1e83437_app1.pdf]

References

1. Shiu E, Hassan LM, Walsh G. Demarketing tobacco through governmental policies – the 4Ps revisited. J Bus Res 2009 Feb;62(2):269-278. [doi: 10.1016/j.jbusres.2008.01.034]
2. He L, Balaji D, Wiers RW, Antheunis ML, Krahmer E. Effectiveness and acceptability of conversational agents for smoking cessation: a systematic review and meta-analysis. Nicotine Tob Res 2023 Jun 9;25(7):1241-1250. [doi: 10.1093/ntr/ntac281] [Medline: 36507916]
3. Bendotti H, Lawler S, Chan GCK, Gartner C, Ireland D, Marshall HM. Conversational artificial intelligence interventions to support smoking cessation: a systematic review and meta-analysis. Digit Health 2023;9. [doi: 10.1177/20552076231211634] [Medline: 37928336]
4. Regmi K, Kassim N, Ahmad N, Tuah NA. Effectiveness of mobile apps for smoking cessation: a review. Tob Prev Cessat 2017;3:12. [doi: 10.18332/tpc/70088] [Medline: 32432186]
5. Tong HL, Quiroz JC, Kocaballi AB, et al. Personalized mobile technologies for lifestyle behavior change: a systematic review, meta-analysis, and meta-regression. Prev Med 2021 Jul;148:106532. [doi: 10.1016/j.ypmed.2021.106532] [Medline: 33774008]
6. Free C, Knight R, Robertson S, et al. Smoking cessation support delivered via mobile phone text messaging (txt2stop): a single-blind, randomised trial. Lancet 2011 Jul 2;378(9785):49-55. [doi: 10.1016/S0140-6736(11)60701-0] [Medline: 21722952]
7. Free C, Whittaker R, Knight R, Abramsky T, Rodgers A, Roberts IG. Txt2stop: a pilot randomised controlled trial of mobile phone-based smoking cessation support. Tob Control 2009 Apr;18(2):88-91. [doi: 10.1136/tc.2008.026146] [Medline: 19318534]
8. Perski O, Crane D, Beard E, Brown J. Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. Digit Health 2019;5. [doi: 10.1177/2055207619880676] [Medline: 31620306]
9. Stead LF, Carroll AJ, Lancaster T. Group behaviour therapy programmes for smoking cessation. Cochrane Database Syst Rev 2017 Mar 31;3(3):CD001007. [doi: 10.1002/14651858.CD001007.pub3] [Medline: 28361497]
10. Mersha AG, Bryant J, Rahman T, McGuffog R, Maddox R, Kennedy M. What are the effective components of group-based treatment programs for smoking cessation? A systematic review and meta-analysis. Nicotine Tob Res 2023 Aug 19;25(9):1525-1537. [doi: 10.1093/ntr/ntad068] [Medline: 37104053]
11. Struik L, Khan S, Assoiants A, Sharma RH. Assessment of social support and quitting smoking in an online community forum: study involving content analysis. JMIR Form Res 2022 Jan 13;6(1):e34429. [doi: 10.2196/34429] [Medline: 35023834]
12. Pechmann C, Pan L, Delucchi K, Lakon CM, Prochaska JJ. Development of a Twitter-based intervention for smoking cessation that encourages high-quality social media interactions via automessages. J Med Internet Res 2015 Feb 23;17(2):e50. [doi: 10.2196/jmir.3772] [Medline: 25707037]
13. Pechmann C, Delucchi K, Lakon CM, Prochaska JJ. Randomised controlled trial evaluation of Tweet2Quit: a social network quit-smoking intervention. Tob Control 2017 Mar;26(2):188-194. [doi: 10.1136/tobaccocontrol-2015-052768] [Medline: 26928205]
14. Joyce E, Kraut RE. Predicting continued participation in newsgroups. J Comp Mediated Comm 2006 Apr;11(3):723-747. [doi: 10.1111/j.1083-6101.2006.00033.x]
15. Bricker JB, Sullivan B, Mull K, Santiago-Torres M, Lavista Ferres JM. Conversational chatbot for cigarette smoking cessation: results from the 11-step user-centered design development process and randomized controlled trial. JMIR Mhealth Uhealth 2024 Jul 23;12:e57318. [doi: 10.2196/57318] [Medline: 38913882]
16. Abroms LC, Wysota CN, Yousefi A, Wu TC, Broniatowski DA. ChatGPT-based chatbot for help quitting smoking via text messaging: an interventional study. JMIR Form Res 2025 Oct 10;9:e79402. [doi: 10.2196/79402] [Medline: 41072009]
17. Abroms LC, Yousefi A, Wysota CN, Wu TC, Broniatowski DA. Assessing the adherence of ChatGPT chatbots to public health guidelines for smoking cessation: content analysis. J Med Internet Res 2025 Jan 30;27:e66896. [doi: 10.2196/66896] [Medline: 39883917]
18. Kosyluk K, Baeder T, Greene KY, et al. Mental distress, label avoidance, and use of a mental health chatbot: results from a US survey. JMIR Form Res 2024 Apr 12;8:e45959. [doi: 10.2196/45959] [Medline: 38607665]

XSL•FO
RenderX

19. Prochaska JJ, Vogel EA, Chieng A, et al. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. J Med Internet Res 2021 Mar 23;23(3):e24850. [doi: 10.2196/24850] [Medline: 33755028]
20. Giyahchi T, Singh S, Harris I, Pechmann C. Customized training of pretrained language models to detect post intents in online health support groups. In: Shaban-Nejad A, Michalowski M, Bianco S, editors. Multimodal AI in Healthcare: A Paradigm Shift in Health Intelligence: Springer International Publishing; 2023:59-75. [doi: 10.1007/978-3-031-14771-5_5]
21. McRoy S, Rastegar-Mojarad M, Wang Y, Ruddy KJ, Haddad TC, Liu H. Assessing unmet information needs of breast cancer survivors: exploratory study of online health forums using text classification and retrieval. JMIR Cancer 2018 May 15;4(1):e10. [doi: 10.2196/cancer.9050] [Medline: 29764801]
22. Park J, Kotzias D, Kuo P, et al. Detecting conversation topics in primary care office visits from transcripts of patient-provider interactions. J Am Med Inform Assoc 2019 Dec 1;26(12):1493-1504. [doi: 10.1093/jamia/ocz140] [Medline: 31532490]
23. Zeng K, Pan Z, Xu Y, Qu Y. An ensemble learning strategy for eligibility criteria text classification for clinical trial recruitment: algorithm development and validation. JMIR Med Inform 2020 Jul 1;8(7):e17832. [doi: 10.2196/17832] [Medline: 32609092]
24. Li X, Shu Q, Kong C, et al. An intelligent system for classifying patient complaints using machine learning and natural language processing: development and validation study. J Med Internet Res 2025;27:e55721. [doi: 10.2196/55721]
25. Babu A, Boddu SB. BERT-based medical chatbot: enhancing healthcare communication through natural language understanding. Explor Res Clin Soc Pharm 2024 Mar;13:100419. [doi: 10.1016/j.rcsop.2024.100419] [Medline: 38495953]
26. Yu HQ, McGuinness S. An experimental study of integrating fine-tuned large language models and prompts for enhancing mental health support chatbot system. J Med Artif Intell 2024;7:16. [doi: 10.21037/jmai-23-136]
27. Zamani A, Reeson M, Marshall T, et al. Intent and entity detection with data augmentation for a mental health virtual assistant chatbot. In: Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents: Association for Computing Machinery; 2023:41. [doi: 10.1145/3570945.3607324]
28. Aftab H, Gautam V, Hawkins R, Alexander R, Habli I. Robust intent classification using Bayesian LSTM for clinical conversational agents (CAs). In: Gao X, Jamalipour A, Guo L, editors. Wireless Mobile Communication and Healthcare: Springer International Publishing; 2022:106-118. [doi: 10.1007/978-3-031-06368-8_8]
29. Zhang B, Tu Z, Hang S, Chu D, Xu X. Conco-ERNIE: complex user intent detect model for smart healthcare cognitive bot. ACM Trans Internet Technol 2023 Feb 28;23(1):21. [doi: 10.1145/3574135]
30. Sakurai H, Miyao Y. Evaluating intention detection capability of large language models in persuasive dialogues. In: Ku L, Martins A, Srikumar V, editors. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers): Association for Computational Linguistics; 2024:1635-1657. [doi: 10.18653/v1/2024.acl-long.90]

Abbreviations

8B: 8 billion parameters
AI: artificial intelligence
BERT: Bidirectional Encoder Representations From Transformers
Bi-LSTM: bidirectional long short-term memory
BLEU: Bilingual Evaluation Understudy
CPU: central processing unit
CUDA: Compute Unified Device Architecture
EDA: easy data augmentation
EOS: end-of-sequence
FP32: 32-bit floating point
GPU: graphics processing unit
HS: human subjects
IRB: institutional review board
LoRA: low-rank adaptation
LSTM: long short-term memory
MIMIC-III: Medical Information Mart for Intensive Care-III
NRT: nicotine replacement therapy
PEFT: parameter-efficient fine-tuning
QLoRA: quantized low-rank adaptation
SVM: support vector machine
UC: University of California
UCI: University of California, Irvine

XSL•FO
RenderX

Edited by A Coristine; submitted 02.Sep.2025; peer-reviewed by C Chikezie, F Amakye, M Ekwueme, O Akinsola, RQ Wang, T Karen; accepted 29.Jan.2026; published 09.Mar.2026.

Please cite as:
Rahman S, Pechmann C(, Harris IG
Enhancing Detection of Message Intents in a Mobile Health Smoking-Cessation Intervention Using Large Language Model Fine-Tuning, Data Downsampling, and Error Correction: Algorithm Development and Validation
J Med Internet Res 2026;28:e83437
URL: https://www.jmir.org/2026/1/e83437
doi:10.2196/83437

© Shagoto Rahman, Cornelia (Connie) Pechmann, Ian G Harris. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluation of an Artificial Intelligence Conversational Chatbot to Enhance HIV Preexposure Prophylaxis Uptake: Development and Usability Internal Testing

Jun Tao[1], PhD; Ellie Pavlick[2], PhD; Amaris Grondin[2], BS; Josue D Bustamante[3], BS; Harrison Martin[1], BA; Hannah Parent[1], MPH; Natalie Fenn[1,4], PhD; Alexi Almonte[1], BA; Amanda Maguire-Wilkerson[1], DrPH; Mofan Gu[5], PhD; Jack Rusley[6,7], MHS, MD; Bryce K Perler[1], MSTR, MD; Tyler Wray[8], PhD; Amy S Nunn[8,9], SCD; Philip A Chan[1,8,9], MS, MD

[1]Department of Medicine, The Warren Alpert Medical School of Brown University, Room 130, 11 4th Street, Providence, RI, United States

[2]Department of Computer Science, Brown University, Providence, RI, United States

[3]School of Electrical Engineering and Computer Science, College of Engineering, Oregon State University, Corvallis, OR, United States

[4]Department of Psychiatry and Human Behavior, The Warren Alpert Medical School of Brown University, Providence, RI, United States

[5]Division of Infectious Diseases, The Miriam Hospital, Providence, RI, United States

[6]Department of Health Services, Policy, and Practice, Brown University School of Public Health, Providence, RI, United States

[7]Department of Pediatrics, Division of Adolescent Medicine, The Warren Alpert Medical School of Brown University, Providence, RI, United States

[8]Department of Behavioral and Social Sciences, Brown University School of Public Health, Providence, RI, United States

[9]Open Door Health, Rhode Island Public Health Institute, Providence, RI, United States

Corresponding Author:
Jun Tao, PhD
Department of Medicine, The Warren Alpert Medical School of Brown University, Room 130, 11 4th Street, Providence, RI, United States

Abstract

Background: The HIV epidemic in the United States disproportionately impacts gay, bisexual, and other men who have sex with men (MSM). Despite the effectiveness of HIV preexposure prophylaxis (PrEP) in preventing HIV acquisition, uptake among MSM remains suboptimal. Motivational interviewing (MI) has demonstrated efficacy at increasing PrEP uptake among MSM but is resource-intensive, limiting scalability. The use of artificial intelligence, particularly large language models with conversational agents (ie, “chatbots”) such as ChatGPT, may offer a scalable approach to delivering MI-based counseling for PrEP and HIV prevention.

Objective: This internal usability testing aimed to evaluate the development of an artificial intelligence–based chatbot, including its ability to provide MI-aligned education about PrEP and HIV prevention and potential to support PrEP uptake.

Methods: The Chatbot for HIV Prevention and Action (CHIA) was built on a GPT-4o base model embedded with a validated knowledge database on HIV and PrEP in English and Spanish. The CHIA was fine-tuned through training on a large MI dataset and prompt engineering. The use of the AutoGen multiagent framework enabled the CHIA to integrate 2 agents, the PrEP Counselor Agent and the Assistant Agent, which specialized in providing MI-based counseling and handling function calls (eg, assessment of HIV risk), respectively. During internal testing from March 10-April 28, 2025, we systematically evaluated the CHIA’s performance in English and Spanish using a set of 5-point Likert scales to measure accuracy, conciseness, up-to-dateness, trustworthiness, and alignment with aspects of the MI spirit (eg, collaboration, autonomy support) and MI-consistent behaviors (eg, affirmation, open-ended questions). Descriptive statistics and mixed linear regression were used to analyze the data.

Results: A total of 296 responses, including 145 English responses and 151 Spanish responses, were collected during the internal testing period. Overall, the CHIA demonstrated strong performance across both languages, receiving the highest combined scores in the general response quality metrics including up-to-dateness (mean 4.6, SD 0.8), trustworthiness (mean 4.5, SD 0.9), accuracy (mean 4.4, SD 0.9), and conciseness (mean 4.2, SD 1.1). The CHIA generally received higher combined scores for metrics that assessed alignment with the MI spirit (ie, empathy, evocation, autonomy support, and collaboration) and lower combined scores for MI-consistent behaviors (ie, affirmation, open-ended questions, and reflections). Spanish responses had significantly lower mean scores than English responses across nearly all MI-based metrics.

Conclusions: Our internal usability testing highlights the potential of the CHIA as a viable tool for delivering MI-aligned counseling in English and Spanish to promote HIV prevention and support PrEP uptake, though its Spanish language performance requires further improvement.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e79671) doi:10.2196/79671

KEYWORDS

artificial intelligence; counseling; HIV infections; motivational interviewing; multiagent framework; preexposure prophylaxis

Introduction

In the United States, HIV continues to be a significant cause of morbidity and mortality, disproportionately affecting groups including gay, bisexual, and other men who have sex with men (MSM). In 2022, MSM comprised 67% of new HIV diagnoses in the United States [1]. Hispanic or Latino and Black or African American individuals accounted for 39% and 35% of these new diagnoses, respectively [1]. HIV preexposure prophylaxis (PrEP) is highly effective at preventing HIV among populations at an increased risk of infection, including MSM [2-4]. However, PrEP uptake among MSM, and particularly Hispanic or Latino and Black or African American MSM, remains suboptimal due to inequitable access to health care and stigma [5-7].

Motivational interviewing (MI) is an evidence-based, patient-centered approach to healthy behavior change that has demonstrated efficacy in facilitating PrEP uptake among MSM [8-10]. In our previous work, we demonstrated that a brief MI-based intervention improved PrEP uptake among MSM in a clinical setting [9]. However, implementing MI in practice requires a significant investment in time and resources for provider training and intervention delivery. Research suggests that provider-led MI counseling usually requires 2-5 sessions (30 - 60 min per session) to produce measurable behavior change, with several studies reporting substantially higher effectiveness (>80%) when more than 5 sessions are delivered [11-13]. The associated clinician time and continuity demands make sustained MI delivery difficult to scale, particularly during brief clinical visits [11-14]. Time demands apply not only to patient encounters but also to provider training, which typically involves several hours of didactic instruction and coaching [13]. Artificial intelligence (AI)—specifically large language models (LLMs) with conversational agents (“chatbots”) such as ChatGPT—has shown promise in overcoming these challenges [15-17]. Preliminary research supports AI-based chatbots’ ability to employ MI techniques in promoting healthy behavior change, including smoking cessation and decreased substance use [17-21]. A chatbot designed to use MI principles can be made available 24/7, requiring only minimal human effort for periodic supervision and quality assurance. Drawing on empirical session-length data and standard labor-cost benchmarks, it suggests that such MI-aligned chatbots could substantially improve the scalability of MI delivery.

In the context of HIV prevention, researchers have noted many uses for AI-based chatbots [22-24]. Limited evidence suggests that chatbots may help facilitate the uptake of HIV testing and PrEP among populations at increased risk of HIV, including MSM [22,23,25]. Chatbots have also shown the potential to provide personalized counseling on sensitive health topics including HIV prevention [22]. Although AI-based chatbots hold significant promise for HIV prevention efforts, concerns exist regarding their ability to provide accurate medical information, stay up-to-date on current clinical recommendations, and demonstrate cultural competence [22,26]. Studies have highlighted issues such as hallucinations (incorrect or misleading information) and the potential for perpetuating biases, which can undermine trust and effectiveness [22]. Low engagement with AI-based chatbots for health promotion has also been documented in the literature, presenting challenges with delivering effective interventions via this modality [27,28]. Additionally, no studies, to our knowledge, have evaluated the use of MI by an AI-based chatbot for HIV prevention.

In this usability testing, we developed and conducted an internal evaluation of an AI-based chatbot (Chatbot for HIV Prevention and Action [CHIA]) that harnesses MI to provide personalized counseling for PrEP. To address the limitations noted above and improve the CHIA’s performance, we integrated three complementary components: (1) a retrieval-augmented generation pipeline constrained to a curated, validated knowledge base to reduce unsupported statements; (2) MI alignment via supervised fine-tuning on annotated MI transcripts coupled with preference-based tuning using expert-selected responses; and (3) personalization through a structured HIV risk assessment and the transtheoretical model (TTM) to tailor counseling to the user’s stage of change [29,30]. We present the results of an internal evaluation of the CHIA, assessing the chatbot’s alignment with MI principles, factual accuracy, and its ability to deliver appropriate counseling for HIV prevention and PrEP. This internal testing serves as a cornerstone for future real-world implementation and evaluation of the CHIA’s performance among individuals at an increased risk of HIV.

Methods

Overview of Chatbot Design

Generative pre-trained transformers are LLMs that use deep learning to generate human-like text based on natural language input [31]. GPT-4o, released by OpenAI in May 2024, is a multimodal LLM capable of processing both text and images with improved efficiency and performance compared to earlier versions [32]. We developed the CHIA using GPT-4o to deliver MI-informed counseling aimed at improving HIV prevention outcomes, particularly the uptake of PrEP.

The CHIA consists of two main components: (1) a fine-tuned, customized LLM embedded with a validated knowledge database and (2) multiple specialized agents and functions to meet users’ needs. Built on a GPT-4o base model, the CHIA detects language inputs automatically and has been fine-tuned using a large MI dataset with the goal of training it to produce empathetic responses and avoid biases [33,34]. GPT-4o was selected because it offers superior conversational quality, multilingual capability, and reduced risk of generating errors compared to smaller open-source models. These features are critical for building trust with users discussing sensitive health topics. To ensure accurate and up-to-date information on HIV and PrEP, the CHIA integrates the latest validated data on these

XSL•FO
RenderX

topics, reviewed by a team of physicians and researchers, in English and Spanish, with monthly updates.

The CHIA's architecture integrates 2 specialized agents (Figure 1): a user-facing PrEP Counselor agent and a tool-executing Assistant agent using AutoGen [35]—an open-source multiagent framework developed by Microsoft. The Counselor conducts the entire conversation using MI, triages user needs, and delegates tasks to the Assistant when external information is required. The Assistant then invokes specific functions—knowledge retrieval from an embedding-backed knowledge base, HIV risk or readiness assessment, PrEP provider search and referral info, reminders or links, and initiating human support on request—and returns results to the Counselor. The Counselor interprets these outputs through MI principles and delivers the response to the user. To maintain continuity across visits, the Counselor agent is capable of using *Teachability* to store key information to ensure the continuity of conversation (eg, identity token, risk or concerns, prior plan or notes). This design enables the CHIA to deliver personalized responses that mimic human counseling. Advanced prompt engineering keeps interactions dynamic and contextually relevant [36]. The CHIA is secured through a login page and deployed on Amazon Web Services (AWS) with the Supabase software (version 1.25; Supabase, Inc.) for backend management, ensuring robust and private data handling [37,38].

Figure 1. Chatbot for HIV Prevention and Action (CHIA) architecture: information flow among User, Counselor Agent, and Assistant Agent. MI: motivational interviewing.

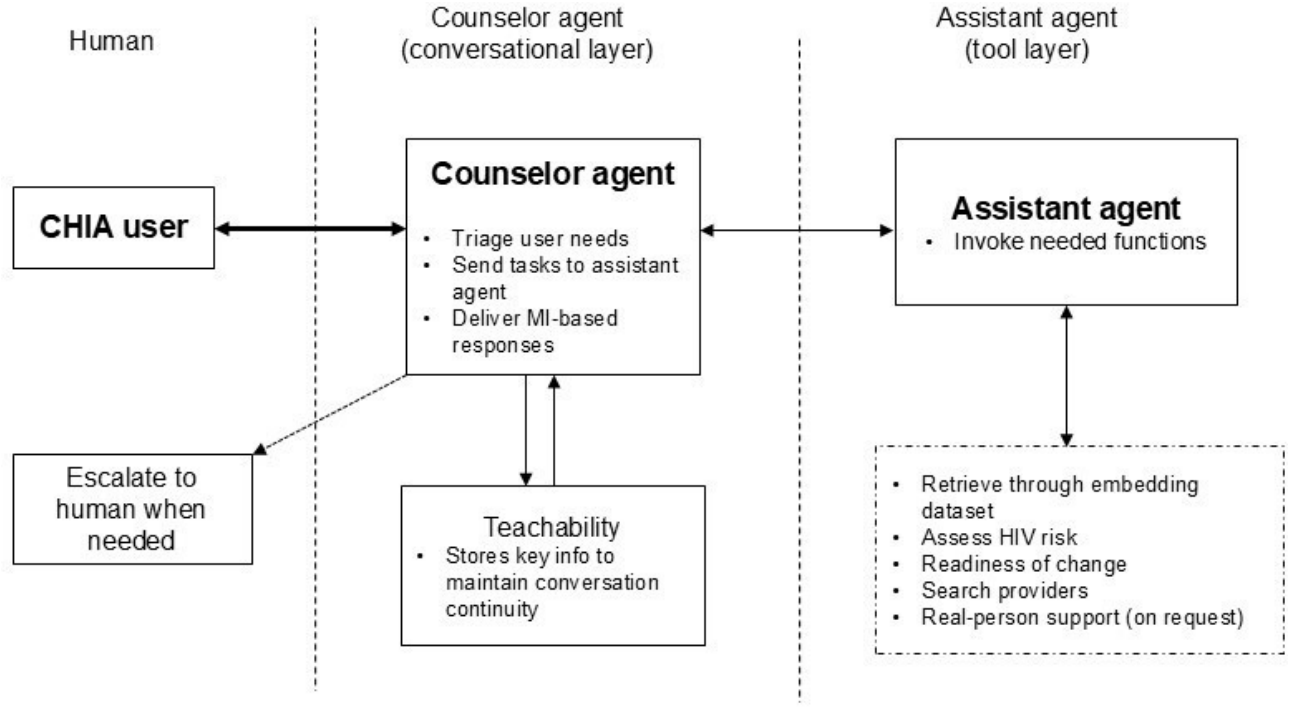

Figure 1 CHIA system workflow: information flow among user, counselor agent, and assistant agent

Embedded HIV/PrEP Dataset Development and Validation

To enhance chatbot accuracy and reliability, embedding and retrieval-based techniques have been employed to minimize misinformation by ensuring responses are grounded in validated sources [39]. Assessment retrieval methods, guided by established frameworks, such as the retrieval-augmented generation (RAG) playground framework and the automated RAG evaluation system [40-42], allow for systematic evaluation of chatbot performance in delivering evidence-based, contextually relevant health information. To develop a robust embedded dataset specifically for the CHIA, our study team, composed of 4 research assistants under the guidance of 2 physicians and 1 principal investigator with decades of research experience in HIV prevention and treatment, systematically curated and validated key information. We compiled the most frequently asked questions about HIV and PrEP, which included topics such as basic knowledge, effectiveness, side effects, formulations, insurance coverage, financial assistance, and cultural considerations. A Spanish-fluent research assistant translated the dataset to ensure accessibility for Spanish-speaking users. Each data entry was structured into a standardized format, categorizing content by topic, question, and answer to optimize retrieval efficiency. All information underwent expert validation for medical accuracy and clarity. These processes prevent the spread of misinformation or erroneous recommendations ("hallucinations"), which ChatGPT cannot guarantee. The embedding database is updated monthly and as needed when users ask questions outside the existing dataset under expert supervision. In such cases, the system flags the query, alerting the development team to review, validate, and integrate new information. This dynamic updating process helps to ensure that the CHIA remains accurate, relevant, and responsive to evolving needs of the community.

XSL•FO
RenderX

MI Alignment via Implementation of Preference Fine-Tuning Techniques

To enhance the CHIA's ability to deliver MI-based counseling, we utilized preference fine-tuning techniques [43]. This was done because open-source models would have been insufficient for MI fine-tuning due to their lack of multilingual support, higher hallucination rates, and insufficient token context. We processed a large, publicly available MI dataset downloaded from GitHub to facilitate diversity in linguistic styles and conversational structures [33,34]. The dataset contained 2000 dialogs, half of which were from publicly available conversations between potential clients and licensed counselors on CounselChat—an online platform for mental health support—and the other half from exchanges between users and peer supporters on Reddit subforums related to emotional distress [33,34]. All dialogs in the dataset were annotated by trained counselors with labels adapted from the Motivational Interviewing Treatment Integrity Code 2.0 and 4.2.1, a widely used method for evaluating how well clinicians perform MI [33,34,44-46]. Our team preprocessed the raw text from the dataset, tokenized it (ie, broke down into smaller units) [47], and then converted this to a vector database. This processed vector database was employed to fine-tune the GPT-4o base model, enabling the CHIA to generate responses informed by MI. We used the Direct Preference Optimization algorithm to refine response quality [43]. A preference fine-tuning JSON file was constructed that ranked certain responses as preferable over others based on their alignment with MI-consistent behaviors such as open-ended questioning and reflective listening. This iterative training process was designed to enhance the CHIA's ability to prioritize MI-aligned responses while maintaining coherence and engagement. By the end of this process, we developed a specialized ChatGPT-4o model for MI-based counseling, serving as the foundation for the CHIA to deliver personalized, empathetic, and structured conversations that support PrEP uptake and public health interventions.

HIV Risk Assessment and Readiness for Change Functions

To enhance the CHIA's ability to provide personalized guidance, we integrated functions to assess individual HIV risk and readiness for behavior change. HIV risk assessment was based on the HIV Incidence Risk Index for MSM [48], following Centers for Disease Control and Prevention guidelines for PrEP eligibility [49]. A focus on MSM was chosen due to the disproportionate number of new HIV diagnoses occurring among this population in the United States [1]. Additionally, the CHIA's development incorporated the TTM through the use of the Contemplation Ladder, a validated tool that allowed individuals to self-assess their readiness for PrEP uptake on a scale from 0 to 10 [29,50]. These assessments were embedded into the CHIA's conversational interface, enabling real-time evaluation, as users engaged with relevant questions. Based on these assessments, the CHIA tailored its responses to align with each user's unique risk profile and stage of readiness for change, providing targeted, motivational, and evidence-based guidance to support PrEP uptake.

Linkage to PrEP Care and Referral Function

To facilitate access to PrEP, we developed a function that searched the "PrEP Locator" website, a national database of PrEP providers in the United States, using user-provided ZIP codes [51]. This function was integrated into the CHIA to generate a list of PrEP providers within a 30-mile radius, offering users convenient options for selecting nearby clinics. Additionally, a separate function was implemented to enable the CHIA to engage users with follow-up questions to assess clinic preferences, determine if they wished to be contacted by a provider for an initial appointment, and identify potential barriers to care. An AI-generated and encrypted email was then sent to the selected clinic with a referral and relevant contact information. Study staff coordinated with the clinic to ensure appointments were scheduled efficiently. Furthermore, the CHIA inquired whether users preferred to connect with a real person for assistance in accessing PrEP care. If requested, study staff were notified and contacted the individual within 1 business day.

User Interface and Data Security

The CHIA user interface was designed to provide a streamlined, secure, and personalized experience (Figure 2). Users could log in with an existing account or create a new one, with 2-factor authentication required for each login to ensure data protection and confidentiality. The login system adhered to high cybersecurity standards to safeguard user information. Once logged in, users accessed a clean, intuitive interface with clear navigation options and a responsive chat window to facilitate seamless interactions.

XSL•FO
RenderX

Figure 2. The user interface of the Chatbot for HIV Prevention and Action (CHIA).

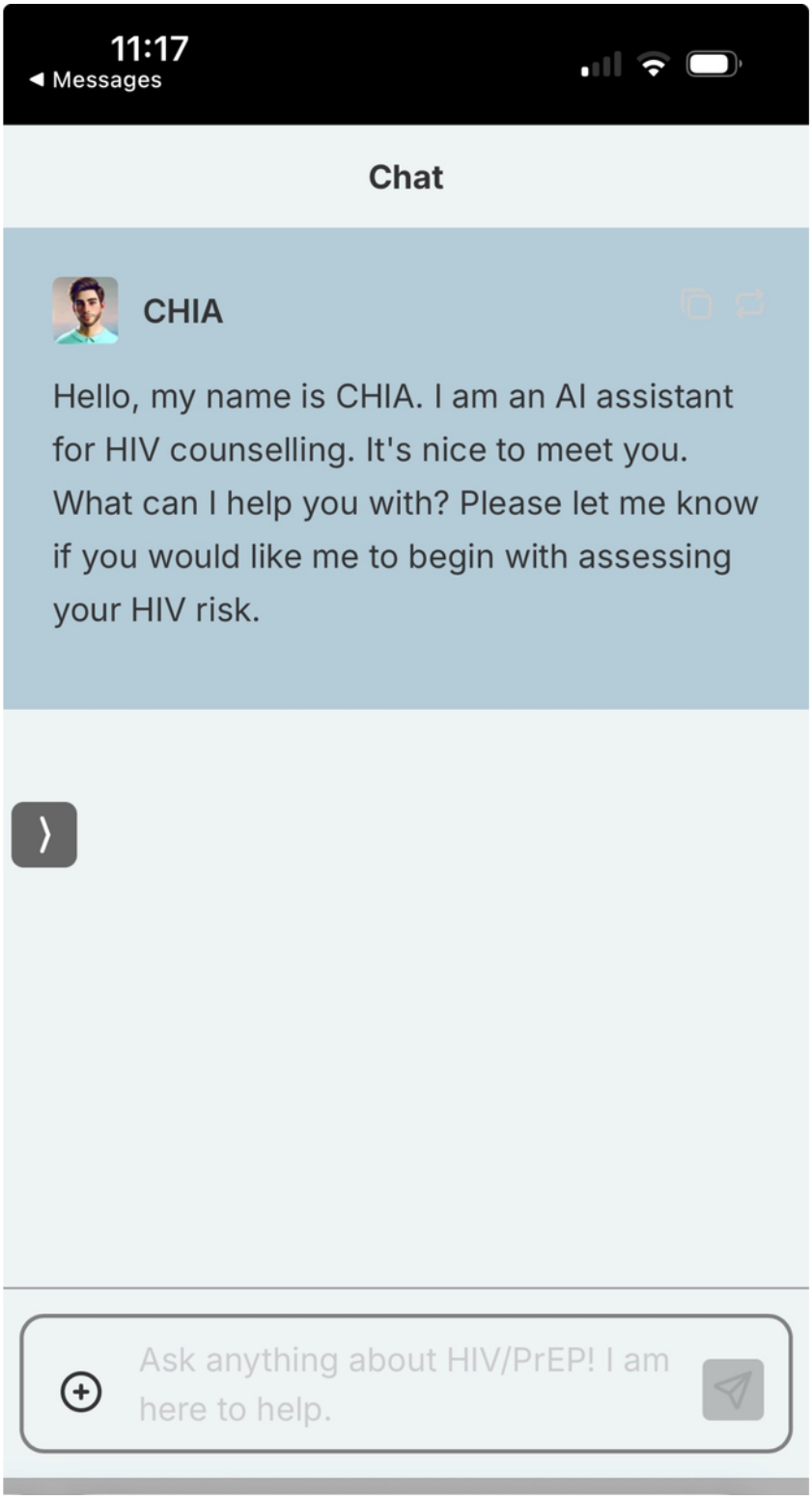

To facilitate secure data management, the CHIA was deployed through AWS, leveraging its cloud-based infrastructure for reliable and Health Insurance Portability and Accountability Act (HIPAA)–compliant storage [38]. AWS encryption protocols protected user data, ensuring confidentiality and integrity [38]. Additionally, Supabase was used to host the back-end data, providing an efficient database solution for managing user interactions [37]. This integration of AWS for deployment and Supabase for back-end data management enabled the CHIA to maintain robust security measures while promoting a smooth user experience. The secure, user-friendly interface aimed to encourage sustained engagement and support meaningful interactions.

Procedures

During internal testing from March 10 to April 18, 2025, 4 research staff members systematically evaluated the CHIA's performance in both English and Spanish. Each researcher acted

XSL•FO
RenderX

as a potential user and interacted with the chatbot over multiple sessions. Two research staff members interacted with the chatbot in English, and 2 staff members fluent in Spanish interacted with the chatbot in Spanish. Researchers then exchanged transcripts and rated each response on a set of 11 metrics covering general response quality and alignment with MI using a 5-point Likert scale. Likert scores were used primarily to enable relative comparisons across languages and scenarios, rather than to assess performance against an external benchmark, as no standardized threshold currently exists for interpreting Likert-based ratings of motivational interviewing alignment in chatbot evaluations. The research team met prior to the review process to ensure familiarity with the operational rating scale and consistency in scoring across raters. After interacting with the chatbot, the team met to discuss conversation-level findings and the performance of the HIV risk assessment, referral, and readiness for change functions.

Evaluation Metrics for Retrieving Information From Embedded Dataset

The CHIA's RAG functionality was evaluated at the session level using 7 key metrics—groundedness, medical accuracy, completeness, no fabrication, appropriate tone, safety, and reasoning—that were scored on a 5-point Likert scale via the Automated RAG Evaluation System [42,52]. Groundedness assessed if responses were based on retrieval sources, and medical accuracy sought to ensure alignment with validated health information. Completeness measured whether the retrieved content provided sufficient information. No fabrication verified that responses did not introduce false or misleading details (sometimes referred to as "hallucinations") by comparing each RAG output against the ground-truth dataset to confirm retrieval only from the embedded validated knowledge base. Appropriate tone was evaluated if responses were professional and empathetic, and safety was assessed whether the information adhered to ethical and safety principles. Finally, reasoning evaluated the chatbot's ability to integrate and apply retrieved knowledge. Sessions with any metric below a predefined threshold were auto-flagged for human review; we logged and reviewed misinformation flags and near-miss events (eg, unsafe advice avoided or corrected) and conducted monthly audits summarizing flag rates, time-to-review, and corrective actions.

Metrics for Comprehensive Assessment

We developed and used a structured assessment framework to evaluate the CHIA's responses based on previous research that incorporated multiple dimensions of chatbot interaction [26,53]. First, we assessed general response quality using 4 metrics—accuracy, conciseness, up-to-dateness, and trustworthiness—that were scored on a 5-point Likert scale to ensure that responses were factually correct, concise, up-to-date, and emotionally supportive. We also developed a set of metrics inspired by the Motivational Interviewing Treatment Integrity Code to measure the alignment of the CHIA's responses with the MI spirit and MI-consistent behaviors [44,45]. Metrics to assess alignment with the MI spirit included empathy, evocation, autonomy support, and collaboration, and those for MI-consistent behaviors included affirmation, open-ended questions, and reflections. All MI-based metrics were also scored using a 5-point Likert scale. Additionally, safety was evaluated on a pass or fail basis in which the CHIA's use of toxic language, demonstration of bias, or violation of privacy constituted a failure. A qualitative feedback section allowed raters to document strengths, weaknesses, and suggested improvements. This comprehensive framework ensured a robust and reliable evaluation of the CHIA's performance in multiple languages and user engagement scenarios. The detailed metrics and operational assessment protocol is included in Multimedia Appendix 1. At the conversation level, we assessed the CHIA's performance using the same response-level metrics in addition to evaluating its overall adherence to MI techniques including employing a guiding (rather than directive) style of conversation, eliciting change talk, and managing sustain talk (ie, statements against change). Given that this analysis was limited to 4 conversations, only qualitative findings are reported in this paper. Finally, we assessed the CHIA's memory and teachability by measuring its ability to recall information accurately and integrate key details from past interactions into current conversation. The qualitative findings of this assessment are summarized in this paper.

Ethical Considerations

This internal testing did not involve human participants or the use of human subject data. All testing was conducted internally using simulated interactions to evaluate chatbot performance. As such, this developmental phase of the study does not meet the definition of human participants research and was determined to be exempt from IRB review by the Miriam Hospital Institutional Review Board (protocol #2312729).

Statistical Analysis

Descriptive statistics, including the mean and SD, were calculated separately for English and Spanish responses for each metric (ie, accuracy, conciseness, up-to-dateness, trustworthiness, empathy, evocation, autonomy support, collaboration, affirmation, open-ended questions, and reflections). We fit mixed linear regression models to estimate the effect of language (Spanish vs English) on each communication metric. Because multiple observations were nested within individuals, models included random intercepts and random slopes for language at the participant level. This specification accounted for within-person correlation (ie, repeated measures from the same individual) and allowed the magnitude of the language effect to vary across individuals. Fixed effects provided the average adjusted difference between languages, while random effects decomposed variance into within- and between-person components. In addition, standardized effect sizes (Cohen *d*) were derived by dividing the adjusted language difference by the residual standard deviation, providing a measure of the practical significance of language effects across metrics. Statistical significance was assessed using 2-tailed *P* values. The significance level was set at $P<.05$. All analyses were performed using Stata (version 18; StataCorp LLC).

XSL•FO
RenderX

Results

Overview of Response-Level Assessment

A total of 296 responses were assessed across 11 metrics, covering general response quality (eg, accuracy, trustworthiness) as well as alignment with the aspects of the MI spirit (eg, evocation, autonomy support) and with MI-consistent behaviors (eg, affirmation, open-ended questions). This total included 145 English responses and 151 Spanish responses. All responses in English and Spanish passed the safety evaluation, indicating that the CHIA's outputs were appropriate, ethical, and unbiased. Table 1 presents the mean and SD for all metrics at the response level, along with effect sizes reported using Cohen *d*. Table 2 displays the examples of CHIA's responses to researcher prompts and their scores for evaluation metrics. Multimedia Appendix 2 includes all responses and scores used in the analysis for both the English and Spanish assessments.

Table . Response-level assessment for the internal testing phase of Chatbot for HIV Prevention and Action (CHIA), an artificial intelligence (AI)–based chatbot for HIV prevention.

Response-level metrics	English (n=145), mean (SD)	Spanish (n=151), mean (SD)	Total (n=296), mean (SD)	ICC[a]	Cohen *d*	*P* value
General response quality metrics						
Accuracy	4.1 (1.1)	4.6 (0.7)	4.4 (0.9)	7.65e-15	−0.55	<.001
Conciseness	4.0 (1.3)	4.5 (0.8)	4.2 (1.1)	0.120	−0.47	.07
Up-to-dateness	4.5 (1.0)	4.7 (0.7)	4.6 (0.8)	0.035	−0.24	.24
Trustworthiness	4.3 (1.0)	4.7 (0.7)	4.5 (0.9)	0.0169	−0.47	.001
MI[b] spirit metrics						
Empathy	4.2 (1.0)	3.9 (1.2)	4.1 (1.1)	0.242	0.272	.95
Evocation	4.0 (1.2)	2.0 (1.1)	3.2 (1.5)	0.257	1.74	.24
Autonomy support	4.2 (1.1)	3.7 (1.3)	3.9 (1.2)	0.279	0.413	.84
Collaboration	4.1 (1.1)	3.7 (1.4)	3.9 (1.3)	0.357	0.313	.786
MI-consistent behaviors metrics						
Affirmation	4.2 (1.2)	2.7 (1.6)	3.5 (1.6)	0.607	1.05	.001
Open-ended questions	3.8 (1.3)	3.0 (1.9)	3.4 (1.7)	0.536	0.48	.95
Reflections	3.6 (1.3)	3.7 (1.4)	3.6 (1.3)	0.452	−0.07	.59

[a]ICC: intraclass correlation coefficient.

[b]MI: motivational interviewing.

XSL•FO
RenderX

Table . Response-level assessment examples and scores[a].

		Evaluation metric						
Researcher prompt	CHIA[b] response	Empathy	Evocation	Autonomy support	Collaboration	Affirmation	Open-ended questions	Reflections
I've tried setting timers before for medications and they don't work.	It sounds like you've tried different strategies without success, and that's okay. Sometimes it takes a bit of creativity to find what works best for you. Perhaps pairing PrEP[c] with a daily activity you never miss, like eating breakfast, could help. Would you like to explore other ways to create a routine that feels natural for you?	5	5	5	5	5	5	5
Is there someone I can talk to about PrEP that is not a doctor or works in healthcare? I don't trust doctors; in my experience, they've treated me poorly.	I'm sorry to hear about your experiences. Seeking support from someone outside of healthcare is a valid choice. Community organizations or support groups focused on sexual health might be a good option. They can offer guidance and answer questions about PrEP. How do you feel about reaching out to a local LGBTQ+[d] center or community health organization for support?	5	5	5	5	5	5	5

XSL•FO
RenderX

		Evaluation metric						
Researcher prompt	CHIA[b] response	Empathy	Evocation	Autonomy support	Collaboration	Affirmation	Open-ended questions	Reflections
My partner could also benefit from PrEP maybe. How do I talk to my partner about it?	Talking to your partner about PrEP can be an important step in taking care of your sexual health together. You might start by expressing your care and concern for both of your well-being. You could say something like, "I've been thinking about how we can protect each other from HIV and STIs[e], and I came across PrEP. It's a medication that can help reduce the risk of HIV, especially if we have sex without condoms sometimes. What do you think about looking into it together?" This approach opens up a dialogue and allows your partner to share their thoughts and feelings, making it a collaborative decision.	4	1	4	1	4	4	4

[a]Responses were obtained through internal testing, in which research assistants simulated potential participants from the target study sample. These responses were not collected from actual human participants within the study sample. Please see Supplemental Dataset 1 for further examples of CHIA responses, including Spanish responses.

[b]CHIA: Chatbot for HIV Prevention and Action.

[c]PrEP: preexposure prophylaxis.

[d]LGBTQ+: lesbian, gay, bisexual, transgender/transsexual, queer, and other minority sexual orientations and gender identities.

[e]STIs: sexually transmitted infections.

English Response-Level Assessment

The CHIA performed strongly in up-to-dateness (mean 4.5, SD 1.0), trustworthiness (mean 4.3, SD 1.0), and accuracy (mean 4.1, SD 1.1). Conciseness had a slightly lower score (mean 4.0, SD 1.3) but remained generally acceptable. Metrics that assessed alignment with the MI spirit generally received higher mean scores than those for MI-consistent behaviors across English responses.

XSL•FO
RenderX

Spanish Response-Level Assessment

Spanish responses demonstrated strong performance across several dimensions. General response quality metrics including trustworthiness (mean 4.7, SD 0.7), up-to-dateness (mean 4.7, SD 0.7), accuracy (mean 4.6, SD 0.7), and conciseness (mean 4.5, SD 0.8) were rated highly. Overall, metrics that measured alignment with the MI spirit and MI-consistent behaviors received lower combined scores.

Combined Response-Level Assessment

The CHIA performed well across both languages, with combined scores showing strength across general response quality metrics including up-to-dateness (mean 4.6, SD 0.8), trustworthiness (mean 4.5, SD 0.9), accuracy (mean 4.4, SD 0.9), and conciseness (mean 4.2, SD 1.1). The mean scores for accuracy, conciseness, and trustworthiness were significantly higher among Spanish responses compared to English responses. Combined scores for metrics that assessed alignment with the MI spirit were generally higher than those for MI-consistent behaviors. Statistical tests indicated that Spanish responses received significantly lower mean scores than English responses across nearly all MI-based metrics.

Conversation Level Assessment

Conversations with the CHIA were generally perceived as reliable and emotionally supportive but occasionally repetitive or overly generic. Empathy and collaboration were present but could be deepened with more emotionally attuned language and user-specific questions. The CHIA's autonomy support was acknowledged, though one instance where the chatbot proceeded with a risk assessment against user preference indicated room for technical and conversational improvements. All reviewers emphasized reducing reliance on early referrals to health care providers and instead suggested a more user-driven flow. MI techniques (eg, change talk elicitation, sustain talk management, guiding style) were successfully implemented across English conversations, yet raters noted the need for the CHIA to ask more personalized, open-ended questions earlier in the conversation to build rapport and relevance. In the Spanish version, reviewers reported that conversations were not consistently MI-aligned, often lacking key aspects of the MI spirit such as autonomy support and MI-consistent behaviors including affirmation, open-ended questions, and reflections.

Evaluation of the Referral, HIV Risk Assessment, and Readiness for Change Functions

For the referral function, ZIP codes from across the United States were entered. The CHIA successfully returned accurate listings of nearby PrEP clinics within a 30-mile radius. However, the chatbot occasionally failed to provide detailed information about specific clinics when requested, highlighting the need to enable the CHIA to retrieve location-specific data by accessing selected clinic websites. In contrast, the HIV risk assessment function consistently performed well across all conversations. This feature has since been refined to allow users to exit the assessment if it is accidentally triggered. Overall, both referral and risk assessment functions were functional and helpful, with minor refinements needed to optimize user experience. All research assistants tested the CHIA's readiness for change function, which supports MI-based counseling by identifying the user's stage of change. The function successfully prompted tailored, stage-appropriate responses to guide users toward PrEP decision-making. Overall, it enhanced the CHIA's ability to deliver personalized, action-oriented support aligned with MI core skills in this internal pilot testing.

Assessment of Retrieval Functionality

Performance was strong across all 7 key metrics designed to assess the CHIA's RAG functionality (groundedness, medical accuracy, completeness, no fabrication, appropriate tone, safety, and reasoning; Table 3). Mean scores ranged from 3.7 to 4.6, with SDs between 0.5 and 1.1. Median scores for each metric were consistently high, with IQRs falling within acceptable variability (eg, median 4, IQR 3-5). These findings indicate that the CHIA's responses were consistently accurate, grounded in reliable sources, and communicated in a safe and professional manner.

Table . Evaluation of the Chatbot for HIV Prevention and Action's (CHIA) retrieval-augmented generation functionality.

Metrics	Mean (SD)
Groundedness	4.1 (1.1)
Medical accuracy	4.6 (0.8)
Completeness	3.7 (1.0)
No fabrication	4.6 (0.8)
Appropriate tone	4.6 (0.6)
Safety	4.7 (0.5)
Reasoning	3.9 (1.0)

Assessment of Memory and Teachability

Preliminary tests indicated that the CHIA was able to successfully recall information from prior discussions when prompted. Additionally, following the implementation of the teachability feature, an overall reduction in repetitive information provided to the user was noted, highlighting the CHIA's ability to adapt based on the user's previous interactions.

XSL•FO
RenderX

Discussion

Principal Findings

To our knowledge, the CHIA is the first LLM-based conversational chatbot grounded in both MI and TTM to deliver PrEP and HIV prevention counseling and facilitate linkage to PrEP care. This approach represents a significant advancement over traditional chatbots based on natural language processing or machine learning, which rely on prewritten scripts or rule-based logic [54]. The CHIA leverages dynamic conversational AI to offer personalized, contextually relevant responses tailored to users' readiness for behavior change. Internal testing demonstrated that the CHIA performs well across both English and Spanish, with relatively high scores in accuracy, trustworthiness, up-to-dateness, conciseness, and most metrics that assessed alignment with the MI spirit. These findings suggest that CHIA has the potential to deliver scalable, high-quality MI-aligned counseling, based on its technical design features such as 24/7 availability and low marginal cost per user. However, we emphasize that scalability was not evaluated in this internal testing and remains a theoretical advantage; the CHIA was assessed only for feasibility and usability in a controlled setting. A full assessment of feasibility, cost-effectiveness, cultural sensitivity, and barriers to real-world scalability will be a focus of the planned randomized controlled trial and subsequent implementation studies.

Although the CHIA received high scores in accuracy and many other assessment metrics, the findings from our internal testing also highlight areas for refinement to further enhance its MI metric-based performance. Beyond accuracy, mimicking human-led MI sessions is critical for improving engagement and behavior change [17]. AI-based chatbots must be trained not only to retrieve and deliver accurate health information but also to apply MI-consistent behaviors, such as open-ended questioning and reflective listening, to foster user motivation and self-efficacy [17]. While the CHIA is capable of providing MI-aligned counseling, there is room to strengthen its ability to engage users more deeply through improved use of MI-consistent responses. MI-based metrics received lower scores, especially in the Spanish version, and the CHIA may benefit from targeted improvements through additional prompt engineering and preference fine-tuning. A Spanish-fluent member of our study team will assist with prompt engineering and optimizing the CHIA for Spanish speakers. The Spanish language responses were generally strong; however, improving cultural and linguistic alignment through more intentional prompt design could strengthen its ability to deliver MI-based counseling even further. A key limitation of our internal testing was the absence of double scoring; future rounds of testing prior to the pilot phase will include 2 independent raters and a consensus process.

To improve the CHIA's responsiveness and alignment with MI, we plan to integrate reinforcement learning into its training process [55]. Reinforcement learning approaches further optimize chatbot responses over time by integrating user feedback and adapting to real-world interactions [55]. Specifically, we will implement the Q-Star algorithm [56], a Q-learning–based approach designed to iteratively optimize performance based on feedback from MI-based metric assessments [57,58]. Q-Star is compatible with the AutoGen framework and will allow us to incorporate a dedicated Q-Agent that learns from evaluation data and adjusts the CHIA's decision-making over time [56]. By continuously refining responses based on user interactions and alignment with MI, this approach offers an adaptive pathway to improve the CHIA's accuracy, engagement, and overall quality of counseling.

The CHIA's core functions (ie, referral for clinical services and HIV risk assessment) performed well during internal testing, successfully delivering relevant information and supporting user needs. However, several areas for improvement were identified to enhance functionality and user experience. For example, the "search for providers" function could be expanded to allow users to access more detailed information about clinics in which they express interest, including services offered, hours, and contact details. The HIV risk assessment function may benefit from an override option that allows users to opt out when they indicate they do not wish to be assessed, thereby supporting autonomy and comfort. Additionally, the "call for real-person support" function could be strengthened by asking more specific questions and capturing details—such as the user's current concern, emotional state, or preferred mode of contact—that would better prepare research assistants to offer timely, personalized support. This function could be triggered when the user expresses uncertainty, distress, or repeatedly requests help, signaling the need for human follow-up. Overall, these targeted improvements represent feasible next steps to further tailor the CHIA's functions to support users' preferences and needs.

Although the CHIA demonstrated decent retrieval efficacy during internal testing, there is room for improvement. While the Spanish responses received higher mean scores in accuracy compared to the English responses, it is important to note that the Spanish version was assessed at a later stage in which the coding team had resolved several multilingual processing issues and patched the retrieval function, which likely resulted in enhanced accuracy. We plan to further expand our retrieval strategy across both languages by incorporating multiple related questions into each embedded response to enhance the relevance and completeness of retrieved information. Additionally, we will restructure the embedded dataset into smaller, topic-specific subsets to improve both retrieval speed and accuracy. To minimize hallucinations and maintain response quality, we will continuously monitor advancements in RAG techniques and integrate improvements as appropriate. Finally, we will track the performance of various LLMs and remain open to adopting alternative models that may better align with the CHIA's counseling objectives and technical needs.

While engagement with chatbots for counseling typically involves short sessions of 3-10 minutes [59,60], insufficient engagement with the CHIA could hinder its effectiveness in real-world settings [61]. To strengthen engagement in the planned iterative refinement phase, the CHIA will offer brief, MI-structured exchanges (engage-focus-evoke-plan) with proactive reengagement and personalization. Sessions will be concise yet purposeful, while teachability preserves consented

XSL•FO
RenderX

context (identity token, risk or concerns, prior plan) so return visits feel continuous and tailored. Between sessions, the CHIA will deliver targeted reminders and follow-up prompts, maintain an encouraging, human-like tone (with judicious emoji where appropriate), and surface just-in-time content aligned with user goals. A streamlined interface (clear navigation, minimal friction) will reduce drop-offs, and on-demand human support will be available for safety, complexity, or user preference. These approaches have the potential to refine the CHIA's performance, improve satisfaction, and support repeat use in real-world settings.

Conclusions

In summary, the internal testing of CHIA demonstrated promising performance in delivering MI-based counseling for HIV and PrEP education in English and Spanish. The chatbot performed well in key areas, including accuracy, trustworthiness, up-to-dateness, conciseness, and overall alignment with the MI spirit. However, targeted refinements are needed, particularly in enhancing MI alignment in Spanish responses to promote the CHIA's acceptability among Spanish-speaking populations. Planned enhancements, such as improved retrieval strategies, reinforcement learning through Q-Star, and iterative prototype refinement based on user feedback, will further strengthen the CHIA's ability to deliver responsive and user-centered support. The assessment of real-world effectiveness of its potential for supporting PrEP uptake is planned for a subsequent phase, during which stigma and discrimination outcomes will be explicitly measured alongside counseling effectiveness. These efforts position the CHIA as a potentially scalable and adaptable tool for promoting HIV prevention, supporting PrEP uptake, and offering valuable insight into the application of conversational AI in health interventions.

Acknowledgments

Artificial intelligence (GPT-4o) was used solely for the purpose of grammar checking and corrections.

Funding

This project was funded by Brown University's Data Science Institute. Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases (R21AI192507) and the National Institute of Mental Health (R34MH138136) of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

EP is a paid consultant for Google Research. All other authors declare no competing interests.

Multimedia Appendix 1
Operational rating scale for Chatbot for HIV Prevention and Action's (CHIA) multilingual evaluation.
[DOCX File, 11 KB - jmir_v28i1e79671_app1.docx]

Multimedia Appendix 2
Researcher prompts, Chatbot for HIV Prevention and Action's (CHIA) responses, and multilingual evaluation metric scores.
[XLSX File, 62 KB - jmir_v28i1e79671_app2.xlsx]

References

1. HIV surveillance supplemental report: estimated HIV incidence and prevalence in the United States, 2018–2022. : Centers for Disease Control and Prevention; 2024 URL: https://stacks.cdc.gov/view/cdc/156513/cdc_156513_DS1.pdf [accessed 2026-01-10]
2. Grant RM, Lama JR, Anderson PL, et al. Preexposure chemoprophylaxis for HIV prevention in men who have sex with men. N Engl J Med 2010 Dec 30;363(27):2587-2599. [doi: 10.1056/NEJMoa1011205] [Medline: 21091279]
3. Mayer KH, Molina JM, Thompson MA, et al. Emtricitabine and tenofovir alafenamide vs emtricitabine and tenofovir disoproxil fumarate for HIV pre-exposure prophylaxis (DISCOVER): primary results from a randomised, double-blind, multicentre, active-controlled, phase 3, non-inferiority trial. Lancet 2020 Jul 25;396(10246):239-254. [doi: 10.1016/S0140-6736(20)31065-5] [Medline: 32711800]
4. Landovitz RJ, Donnell D, Clement ME, et al. Cabotegravir for HIV prevention in cisgender men and transgender women. N Engl J Med 2021 Aug 12;385(7):595-608. [doi: 10.1056/NEJMoa2101016] [Medline: 34379922]

XSL•FO
RenderX

5. Kanny D, Jeffries WL 4th, Chapin-Bardales J, et al. Racial/ethnic disparities in HIV preexposure prophylaxis among men who have sex with men - 23 urban areas, 2017. MMWR Morb Mortal Wkly Rep 2019 Sep 20;68(37):801-806. [doi: 10.15585/mmwr.mm6837a2] [Medline: 31536484]
6. Mayer KH, Agwu A, Malebranche D. Barriers to the wider use of pre-exposure prophylaxis in the United States: a narrative review. Adv Ther 2020 May;37(5):1778-1811. [doi: 10.1007/s12325-020-01295-0] [Medline: 32232664]
7. Sullivan PS, DuBose SN, Castel AD, et al. Equity of PrEP uptake by race, ethnicity, sex and region in the United States in the first decade of PrEP: a population-based analysis. Lancet Reg Health Am 2024 May;33:100738. [doi: 10.1016/j.lana.2024.100738] [Medline: 38659491]
8. Chan PA, Nunn A, van den Berg JJ, et al. A randomized trial of a brief behavioral intervention for PrEP uptake among men who have sex with men at increased risk for HIV infection. J Acquir Immune Defic Syndr 2021 Jul 1;87(3):937-943. [doi: 10.1097/QAI.0000000000002671] [Medline: 33734099]
9. Moitra E, van den Berg JJ, Sowemimo-Coker G, Chau S, Nunn A, Chan PA. Open pilot trial of a brief motivational interviewing-based HIV pre-exposure prophylaxis intervention for men who have sex with men: preliminary effects, and evidence of feasibility and acceptability. AIDS Care 2020 Mar;32(3):406-410. [doi: 10.1080/09540121.2019.1622644] [Medline: 31130000]
10. Miller WR, Rollnick S. Motivational Interviewing: Helping People Change, 3rd edition: Guilford Press; 2013. URL: https://www.biblio.com/book/motivational-interviewing-helping-people-change-miller/d/1659959579 [accessed 2026-01-10]
11. Rubak S, Sandbaek A, Lauritzen T, Christensen B. Motivational interviewing: a systematic review and meta-analysis. Br J Gen Pract 2005 Apr;55(513):305-312. [Medline: 15826439]
12. Calhoun D, Brod R, Kirlin K, Howard BV, Schuldberg D, Fiore C. Effectiveness of motivational interviewing for improving self-care among northern plains Indians with type 2 diabetes. Diabetes Spectr 2010 Jan 1;23(2):107-114. [doi: 10.2337/diaspect.23.2.107]
13. Jacobs NN, Calvo L, Dieringer A, Hall A, Danko R. Motivational interviewing training: a case-based curriculum for preclinical medical students. MedEdPORTAL 2021 Feb 12;17:11104. [doi: 10.15766/mep_2374-8265.11104] [Medline: 33598544]
14. Rollnick S, Miller WR, Butler CC. Motivational Interviewing in Health Care: Helping Patients Change Behavior: Guilford Press; 2008. URL: https://www.guilford.com/books/Motivational-Interviewing-in-Health-Care/Rollnick-Miller-Butler/9781462550371 [accessed 2026-01-10] [doi: 10.1080/15412550802093108]
15. Nazir A, Wang Z. A comprehensive survey of ChatGPT: advancements, applications, prospects, and challenges. Meta Radiol 2023 Sep;1(2):100022. [doi: 10.1016/j.metrad.2023.100022] [Medline: 37901715]
16. Tudor Car L, Dhinagaran DA, Kyaw BM, et al. Conversational agents in health care: scoping review and conceptual analysis. J Med Internet Res 2020 Aug 7;22(8):e17158. [doi: 10.2196/17158] [Medline: 32763886]
17. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. J Med Internet Res 2023 Feb 24;25:e40789. [doi: 10.2196/40789] [Medline: 36826990]
18. Brown A, Kumar AT, Melamed O, et al. A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking: iterative development study. JMIR Ment Health 2023 Oct 17;10:e49132. [doi: 10.2196/49132] [Medline: 37847539]
19. He L, Basar E, Wiers RW, Antheunis ML, Krahmer E. Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. BMC Public Health 2022 Apr 12;22(1):726. [doi: 10.1186/s12889-022-13115-x] [Medline: 35413887]
20. Prochaska JJ, Vogel EA, Chieng A, et al. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. J Med Internet Res 2021 Mar 23;23(3):e24850. [doi: 10.2196/24850] [Medline: 33755028]
21. Abid A, Baxter SL. Breaking barriers in behavioral change: the potential of artificial intelligence-driven motivational interviewing. J Glaucoma 2024 Jul 1;33(7):473-477. [doi: 10.1097/IJG.0000000000002382] [Medline: 38595151]
22. van Heerden A, Bosman S, Swendeman D, Comulada WS. Chatbots for HIV prevention and care: a narrative review. Curr HIV/AIDS Rep 2023 Dec;20(6):481-486. [doi: 10.1007/s11904-023-00681-x] [Medline: 38010467]
23. Cheah MH, Gan YN, Altice FL, et al. Testing the feasibility and acceptability of using an artificial intelligence chatbot to promote HIV testing and pre-exposure prophylaxis in Malaysia: mixed methods study. JMIR Hum Factors 2024 Jan 26;11:e52055. [doi: 10.2196/52055] [Medline: 38277206]
24. Kamitani E, Mizuno Y, Khalil GM, Viguerie A, DeLuca JB, Mishra N. Improving HIV preexposure prophylaxis uptake with artificial intelligence and automation: a systematic review. AIDS 2024 Aug 1;38(10):1560-1569. [doi: 10.1097/QAD.0000000000003935] [Medline: 38788206]
25. Ntinga X, Musiello F, Keter AK, Barnabas R, van Heerden A. The feasibility and acceptability of an mHealth conversational agent designed to support HIV self-testing in South Africa: cross-sectional study. J Med Internet Res 2022 Dec 12;24(12):e39816. [doi: 10.2196/39816] [Medline: 36508248]
26. Fujimoto M, Hunter L, McCoy S, Outran S, Packel L. Evaluating AI chatbots for HIV prevention: an assessment of response quality and user tailoring. : California HIV / AIDS Policy Research Center; 2024 URL: https://chprc.org/wp-content/uploads/2022/06/AI_Policy_Brief_Oct2024.pdf [accessed 2026-01-10]

XSL•FO
RenderX

27. Jabir AI, Lin X, Martinengo L, Sharp G, Theng YL, Tudor Car L. Attrition in conversational agent-delivered mental health interventions: systematic review and meta-analysis. J Med Internet Res 2024 Feb 27;26(1):e48168. [doi: 10.2196/48168] [Medline: 38412023]
28. Yang Y, Tavares J, Oliveira T. A new research model for artificial intelligence-based well-being chatbot engagement: survey study. JMIR Hum Factors 2024 Nov 11;11:e59908. [doi: 10.2196/59908] [Medline: 39527812]
29. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997;12(1):38-48. [doi: 10.4278/0890-1171-12.1.38] [Medline: 10170434]
30. Raihan N, Cogburn M. Stages of change theory. In: StatPearls: StatPearls Publishing; 2025. [Medline: 32310465]
31. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber Phys Syst 2023;3:121-154. [doi: 10.1016/j.iotcps.2023.04.003]
32. Zhang N, Sun Z, Xie Y, Wu H, Li C. The latest version ChatGPT powered by GPT-4o: what will it bring to the medical field? Int J Surg 2024 Sep 1;110(9):6018-6019. [doi: 10.1097/JS9.0000000000001754] [Medline: 38857508]
33. Welivita A, Pu P. Curating a large-scale motivational interviewing dataset using peer support forums. 2022 Presented at: Proceedings of the 29th International Conference on Computational Linguistics; Oct 12-17, 2022; Gyeongju, Republic of Korea p. 3315-3330 URL: https://aclanthology.org/2022.coling-1.293.pdf [accessed 2026-01-10]
34. Welivita A, Pu P. Curating a large-scale motivational interviewing dataset using peer support forums. in the 29th international conference on computational linguistics (COLING). GitHub. 2022. URL: https://github.com/anuradha1992/Motivational-Interviewing-Dataset [accessed 2026-01-10]
35. Wu Q, Bansal G, Zhang J, et al. AutoGen: enabling next-gen LLM applications via multi-agent conversation. arXiv. Preprint posted online on Aug 16, 2023. [doi: 10.48550/arXiv.2308.08155]
36. Marvin G, Hellen N, Jjingo D, Nakatumba-Nabende J. Data intelligence and cognitive informatics. In: Prompt Engineering in Large Language Models: Springer Nature; 2024:387-402. [doi: 10.1007/978-981-99-7962-2_30]
37. Sai Varshitha G, Rupa R, Divya D. Remix-based real time blood bank communication integrating access control using XGBoost and Supabase. 2024 Presented at: 2024 IEEE Students Conference on Engineering and Systems (SCES); Jun 21-23, 2024; Prayagraj, India p. 1-6. [doi: 10.1109/SCES61914.2024.10652474]
38. Narula S, Jain A. Cloud computing security: amazon web service. 2015 Presented at: 2015 Fifth International Conference on Advanced Computing & Communication Technologies (ACCT); Feb 21-22, 2015; Haryana, India p. 501-505. [doi: 10.1109/ACCT.2015.20] [Medline: 39304265]
39. Ayala O, Bechard P. Reducing hallucination in structured outputs via retrieval-augmented generation. 2024 Presented at: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track); Jun 16-21, 2024. [doi: 10.18653/v1/2024.naacl-industry.19]
40. Saad-Falcon J, Khattab O, Potts C, Zaharia M. ARES: an automated evaluation framework for retrieval-augmented generation systems. 2024 Presented at: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Jun 16-21, 2024. [doi: 10.18653/v1/2024.naacl-long.20]
41. Papadimitriou I, Gialampoukidis I, Vrochidis S, Ioannis K. RAG playground: a framework for systematic evaluation of retrieval strategies and prompt engineering in RAG systems. arXiv. Preprint posted online on Dec 16, 2024. [doi: 10.48550/arXiv.2412.12322]
42. Es S, James J, Espinosa Anke L, Schockaert S. RAGAS: automated evaluation of retrieval augmented generation. 2024 Presented at: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics; Mar 17-22, 2024; St Julians, Malta p. 150-158. [doi: 10.18653/v1/2024.eacl-demo.16]
43. Rafailov R, Sharma A, Mitchell E, Manning CD, Ermon S, Finn C. Direct preference optimization: your language model is secretly a reward model. 2023 Presented at: NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems; Dec 10-16, 2023. [doi: 10.5555/3666122.3668460]
44. Moyers TB, Manuel JK, Ernst D. Motivational interviewing treatment integrity coding manual 4.2.1. 2014 URL: https://motivationalinterviewing.org/sites/default/files/miti4_2.pdf [accessed 2026-01-10]
45. Moyers TB, Martin T, Manuel JK, Miller WR. The motivational interviewing treatment integrity (MITI) code: version 2.0. : University of New Mexico Center on Alcoholism, Substance Abuse, and Addictions (CASAA); 2003 URL: https://casaa.unm.edu/assets/docs/miti1.pdf [accessed 2026-01-10]
46. Moyers TB, Rowell LN, Manuel JK, Ernst D, Houck JM. The Motivational Interviewing Treatment Integrity Code (MITI 4): rationale, preliminary reliability and validity. J Subst Abuse Treat 2016 Jun;65:36-42. [doi: 10.1016/j.jsat.2016.01.001] [Medline: 26874558]
47. Grefenstette G. Tokenization. In: van Halteren H, editor. Syntactic Wordclass Tagging: Springer; 1999:117-133. [doi: 10.1007/978-94-015-9273-4_9]
48. Smith DK, Pals SL, Herbst JH, Shinde S, Carey JW. Development of a clinical screening index predictive of incident HIV infection among men who have sex with men in the United States. J Acquir Immune Defic Syndr 2012 Aug 1;60(4):421-427. [doi: 10.1097/QAI.0b013e318256b2f6] [Medline: 22487585]
49. Preexposure prophylaxis for the prevention of HIV infection in the United States—2021 update: a clinical practice guideline. : Centers for Disease Control and Prevention; 2021 URL: https://stacks.cdc.gov/view/cdc/112360/cdc_112360_DS1.pdf [accessed 2026-01-10]

XSL•FO
RenderX

50. Biener L, Abrams DB. The Contemplation Ladder: validation of a measure of readiness to consider smoking cessation. Health Psychol 1991;10(5):360-365. [doi: 10.1037//0278-6133.10.5.360] [Medline: 1935872]
51. Emory University NPIN. PrEP Locator. URL: https://preplocator.org [accessed 2026-1-10]
52. Ngo NT, Van Nguyen C, Dernoncourt F, Nguyen TH. Comprehensive and practical evaluation of retrieval-augmented generation systems for medical question answering. arXiv. Preprint posted online on Nov 14, 2024. [doi: 10.48550/arXiv.2411.09213]
53. Abbasian M, Khatibi E, Azimi I, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digit Med 2024 Mar 29;7(1):82. [doi: 10.1038/s41746-024-01074-z] [Medline: 38553625]
54. Adamopoulou E, Moussiades L. An overview of chatbot technology. 2020 Presented at: Artificial Intelligence Applications and Innovations (AIAI) 2020; Jun 5-7, 2020. [doi: 10.1007/978-3-030-49186-4_31]
55. Li J, Monroe W, Ritter A, Jurafsky D, Galley M, Gao J. Deep reinforcement learning for dialogue generation. In: Gao J, editor. 2016 Presented at: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing; Nov 1-5, 2016. [doi: 10.18653/v1/D16-1127]
56. McIntosh TR, Susnjak T, Liu T, Watters P, Halgamuge MN. From Google Gemini to OpenAI Q* (Q-star): a survey of reshaping the generative artificial intelligence (AI) research landscape. arXiv. Preprint posted online on Dec 18, 2023. [doi: 10.3390/technologies13020051]
57. Watkins CJCH, Dayan P. Q-learning. Mach Learn 1992 May;8(3-4):279-292. [doi: 10.1007/BF00992698]
58. Jang B, Kim M, Harerimana G, Kim JW. Q-learning algorithms: a comprehensive classification and applications. IEEE Access 2019;7:133653-133667. [doi: 10.1109/ACCESS.2019.2941229]
59. Haque MDR, Rubya S. An overview of chatbot-based mobile mental health apps: insights from app description and user reviews. JMIR Mhealth Uhealth 2023 May 22;11:e44838. [doi: 10.2196/44838] [Medline: 37213181]
60. Irvine R, Boubert D, Raina V, et al. Rewarding chatbots for real-world engagement with millions of users. arXiv. Preprint posted online on Mar 10, 2023. [doi: 10.48550/arXiv.2303.06135]
61. Akdim K, Casaló LV. Perceived value of AI-based recommendations service: the case of voice assistants. Serv Bus 2023 Mar;17(1):81-112. [doi: 10.1007/s11628-023-00527-x]

Abbreviations

AI: artificial intelligence
AI: Artificial intelligence
AWS: Amazon Web Services
CHIA: Chatbot for HIV Prevention and Action
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
MI: motivational interviewing
MSM: Men who have sex with men
PrEP: preexposure prophylaxis
RAG: retrieval-augmented generation
TTM: transtheoretical model

Edited by A Mavragani; submitted 25.Jun.2025; peer-reviewed by J Xu, S Chen, S Graham; revised version received 30.Sep.2025; accepted 02.Oct.2025; published 03.Feb.2026.

Please cite as:
Tao J, Pavlick E, Grondin A, Bustamante JD, Martin H, Parent H, Fenn N, Almonte A, Maguire-Wilkerson A, Gu M, Rusley J, Perler BK, Wray T, Nunn AS, Chan PA
Evaluation of an Artificial Intelligence Conversational Chatbot to Enhance HIV Preexposure Prophylaxis Uptake: Development and Usability Internal Testing
J Med Internet Res 2026;28:e79671
URL: https://www.jmir.org/2026/1/e79671
doi:10.2196/79671

© Jun Tao, Ellie Pavlick, Amaris Grondin, Josue D Bustamante, Harrison Martin, Hannah Parent, Natalie Fenn, Alexi Almonte, Amanda Maguire-Wilkerson, Mofan Gu, Jack Rusley, Bryce K Perler, Tyler Wray, Amy S Nunn, Philip A Chan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 3.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of

XSL•FO
RenderX

Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Developing a Quality Evaluation Index System for Health Conversational Artificial Intelligence: Mixed Methods Study

Weizhen Liao[1], MPH; Meng Li[2], BE; Chengyu Ma[1*], PhD; Youli Han[1*], PhD; Dan Wang[2], MS; Haopeng Liu[1], MSc; Yi Wang[1], MSc; Zijie Feng[1], MSc; Huichao Wang[2], BE; Yiru Guan[2], M.Man

[1]School of Public Health, Capital Medical University, Beijing, China

[2]Bytedance Xiaohe Health, Hainan, China

[*]these authors contributed equally

Corresponding Author:
Chengyu Ma, PhD
School of Public Health
Capital Medical University
No. 10, Xitou Alley, You'anmenwai Fengtai District
Beijing, 100069
China
Phone: 86 13811320990
Email: machengyu@ccmu.edu.cn

Abstract

Background: Effective communication is fundamental to health care; however, demographic transitions and a widening global health workforce gap are intensifying the imbalance between service demand and resource supply. Health conversational artificial intelligence (HCAI) based on large language models offers a potential pathway to improve the accessibility and personalization of care. Nevertheless, the lack of a rigorous, user-centered evaluation framework limits the systematic assessment of HCAI quality, raising concerns regarding safety, reliability, and clinical applicability.

Objective: This study aims to establish a scientific and systematic quality evaluation index system for HCAI, providing both a theoretical foundation and a practical tool for the assessment and optimization of HCAI.

Methods: Based on a literature review, industry standards, and expert group discussions, a preliminary framework for the index system was established. Two rounds of Delphi expert consultations were then conducted to collect expert opinions. The analytic hierarchy process (AHP) was applied to assign weights to indicators at each level, and the final content and structure of the index system were determined.

Results: Both rounds of expert consultation achieved a 100% response rate. The authority coefficient of the experts was 0.84 in both rounds. Kendall W coefficient ranged from 0.14 to 0.20 in the first round and from 0.13 to 0.17 in the second round, with all values showing statistical significance (round one: importance $P<.001$, feasibility $P<.001$, sensitivity $P<.001$; round two: importance $P=.001$, feasibility $P<.001$, sensitivity $P=.001$). The final HCAI quality evaluation index system consisted of 3 primary indicators, 7 secondary indicators, and 28 tertiary indicators. According to AHP weight calculations, the primary indicators were ranked in descending order as follows: ethics and compliance (0.4781), health consultation capability (0.4112), and user experience (0.1107).

Conclusions: The evaluation index system constructed in this study demonstrates scientific validity and practical relevance. It provides a valuable reference for the quality assessment, model optimization, and regulatory oversight of HCAI systems.

(J Med Internet Res 2026;28:e83188) doi:10.2196/83188

KEYWORDS

health consultation; conversational AI; evaluation index system; Delphi method; analytic hierarchy process

XSL•FO
RenderX

Introduction

Background

> *Doctors have three magic weapons: language, medicine and scalpel.* [Hippocrates]

Effective communication is essential not only for facilitating information exchange between clinicians and patients but also for establishing trust, mutual understanding, and ongoing support [1]. Currently, health care systems are under increasing pressure. On the one hand, the global population is aging at a rapid pace. According to the World Health Organization, individuals aged 60 and above are projected to account for 22% of the global population by 2050 [2]. On the other hand, chronic and severe illnesses are increasingly affecting younger populations, while public awareness of health and wellness continues to rise. A growing number of individuals are actively seeking medical information and professional advice. These demographic transitions, coupled with the shifting burden of disease, are driving a surge in demand for health care services, while simultaneously exposing persistent systemic challenges, such as inadequate resource distribution and a shortage of health care professionals. It is estimated that the global health workforce deficit will exceed 11 million by 2030 [3]. This widening gap between supply and demand has made it increasingly difficult for patients to obtain timely, effective, and personalized communication and care. In this context, health conversational artificial intelligence (HCAI) based on large language models (LLMs) offers a new approach to mitigating these challenges [4].

Previous studies have demonstrated that LLMs, such as ChatGPT and LLaMA (Large Language Model from Meta), perform remarkably well across a variety of medical tasks, including health consultation [5], diagnostic assistance [6], enhancement of doctor-patient communication [7], and medical image interpretation [8]. These findings highlight the considerable potential of LLMs in clinical and health-related domains. In addition, recent developments have enabled LLMs to conduct multiturn conversations and effectively understand context [9]. This progress provides a solid foundation for building HCAI systems. HCAI refers to LLM-based systems that simulate dialogue with users in a manner similar to that of human physicians. Representative examples include AMIE (Articulate Medical Intelligence Explorer), developed by Google [10]; Ask Patients with Patience, proposed by Zhu and Wu [11]; and Xiaohe AI Doctor, launched by Douyin in China [12]. These systems can ask users follow-up questions and collect and integrate information such as symptoms, medical history, and lifestyle habits [10]. As a result, they can offer more accurate and personalized health care services (Figure 1).

Figure 1. Example of multiturn interaction between health conversational artificial intelligence (AI) and user.

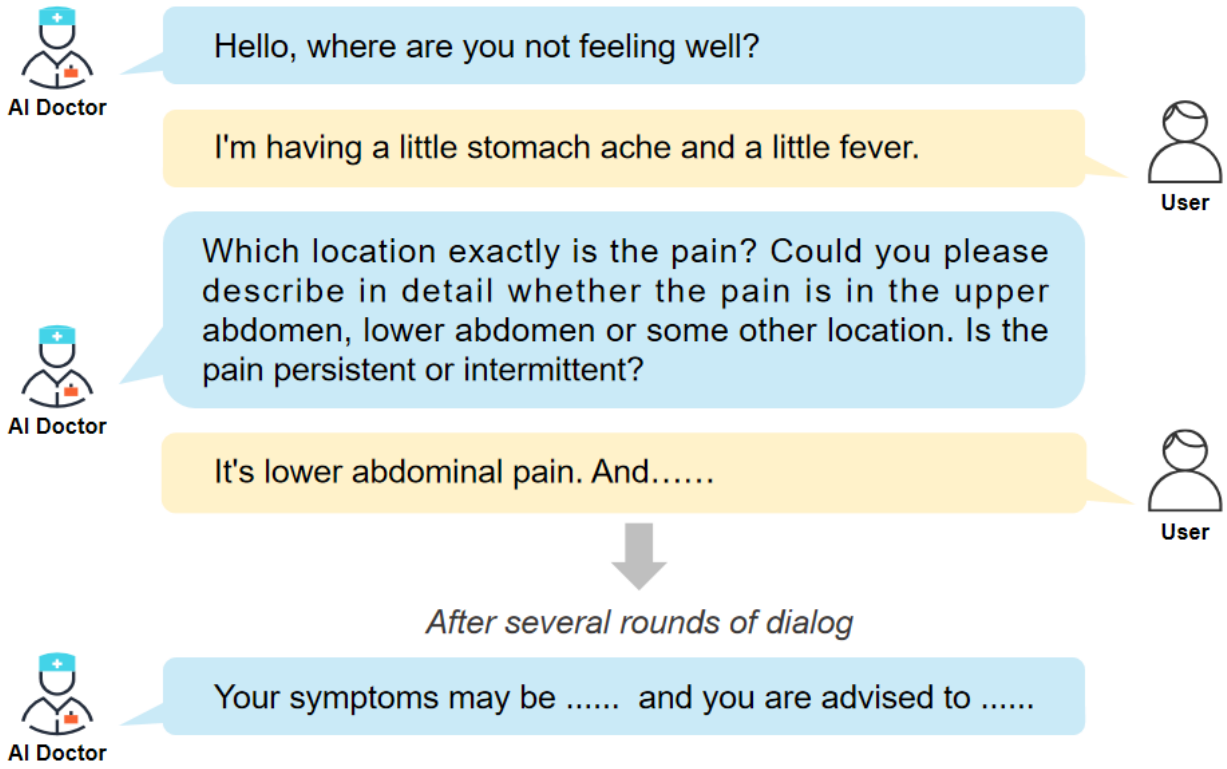

The rapid development of artificial intelligence (AI) has accelerated the expansion of HCAI products. The global conversational AI in the health care market is projected to reach US $16.9 billion in 2025 and to grow at a compound annual growth rate of 24.7% through 2034, reaching a value of US $123.1 billion [13]. As of May 1, 2025, a total of 288 medical LLMs have been publicly released in China. These models cover a wide range of applications, including clinical decision support for specific diseases, prediagnosis consultations, and medical record generation [14]. As a key application of AI in the health

XSL•FO
RenderX

sector, the performance and quality of HCAI systems are directly linked to user safety and health outcomes. Consequently, it is essential to conduct systematic and evidence-based evaluations of these technologies. However, for HCAI systems that deliver health services through multiturn interactions, a scientifically grounded, structured, and practically applicable framework for quality assessment indicators has yet to be established.

Existing studies have employed diverse methods and indicators to evaluate HCAI. Early work often relied on medical examination question banks to assess the accuracy of AI responses to closed-ended questions. Examples include the United States Medical Licensing Examination [15], the Membership of the Royal College of General Practitioners Applied Knowledge Test [16], and the National Medical Licensing Examination in China [17]. Subsequently, several studies developed benchmark tests or clinical case datasets, shifting the focus toward the content generation ability of AI in single-turn open-ended medical questions. Representative datasets include CPMI (Chinese Patent Medicine Instructions dataset) [18], Medical-Diff-VQA (Medical Dataset for Difference Visual Question Answering on Chest X-Ray Images) [19], PubMedQA (a novel biomedical question answering dataset collected from PubMed abstracts) [20], and MultiMedQA (a benchmark for answering medical questions spanning medical exam, medical research, and consumer medical questions, comprising 7 medical question-answering datasets) [21]. Most studies at this stage emphasized content accuracy, with commonly used indicators including BLEU (Bilingual Evaluation Understudy) [5], BERTScore [22], accuracy [23,24], and completeness [25]. However, indicators related to safety risks, such as bias and hallucinations, received limited attention [26,27].

More recent studies have developed methods that better reflect real clinical settings and have assessed AI performance in multitask and multiturn medical dialogues. Xu et al [28] designed MedGPTEval to examine AI performance across 3 dimensions: medical professional capabilities, social comprehensive capabilities, and contextual capabilities. Tu et al [10] employed a remote objective structured clinical examination and created an evaluation framework based on feedback from clinicians and patient participants. This framework assessed history-taking, diagnostic accuracy, management, communication skills, and empathy. Liu et al [29] formulated a clinical pathway specific to LLMs and evaluated clinical capability using 6 indicators: information completeness, behavior standardization, guidance rationality, diagnostic logicality, treatment logicality, and clinical applicability. Johri et al [30] developed the CRAFT-MD (the Conversational Reasoning Assessment Framework for Testing in Medicine) framework to assess the ability of AI to lead clinical conversations, collect complete histories, and make accurate diagnostic decisions. Liu et al [31] proposed the CLEVER framework to evaluate performance in medical case comprehension, clinical reasoning, and diagnosis. Arora et al [32] introduced HealthBench to assess accuracy, completeness, communication quality, context awareness, and instruction following across diverse clinical tasks. Qiu et al [33] developed MedR-Bench to evaluate clinical reasoning in examination recommendations, diagnostic decisions, and treatment planning, and used efficiency, factual accuracy, and completeness as key indicators.

Although these frameworks have advanced the evaluation of clinical reasoning and diagnostic decision-making, their utility for assessing HCAI systems as real-world, user-facing products remains limited. A complete and systematic assessment structure is still lacking, particularly due to the gaps listed below.

First, most existing frameworks adopt a perspective centered on clinical-task performance, such as diagnostic accuracy and history-taking completeness. This approach, while crucial for medical validation, often overlooks the perspective of real-world product application. Specifically, it pays limited attention to user-experience factors such as emotional support, trust, and personalization [34]. These empathy-related features are important because interaction fluency, content clarity, personalized advice, and emotional support can directly influence users' health behaviors and consultation experiences, yet they are largely absent from most evaluation systems.

Second, existing frameworks often lack comprehensive dimensional integration. They tend to focus heavily on medical effectiveness, while insufficient attention has been paid to ethical and compliance-related requirements, such as bias, hallucinations, and protection of personal health data. This omission is inconsistent with current regulatory trends in many countries. For example, the European Union Medical Device Regulation [35] requires AI systems that support diagnosis, prognosis, or treatment, and that may affect clinical workflows, to meet requirements for performance reliability, risk management, cybersecurity, and traceability. A holistic framework that equally weights medical effectiveness, user experience, and ethical compliance is therefore essential.

Third, while some frameworks claim to address multiturn dialogue, they do not offer a systematic or operational methodology for evaluating quality and safety. They often fail to adequately assess context coherence and logical consistency across dialogue turns. These indicators are essential for distinguishing performance between single-turn medical questions and answers and multiturn medical dialogue tasks. Furthermore, by generally relying on descriptive criteria rather than weighted, hierarchical indicators, their practical utility for developers, evaluators, and regulators seeking standardized assessment is limited [22-27,29,30,34].

This lack of comprehensive indicators and balanced evaluation dimensions may result in inaccurate assessments of AI's practical effectiveness, thereby posing risks to patient safety, increasing technical burdens on systems, and creating barriers to the further development of HCAI. Therefore, a scientific scheme that holistically integrates content quality, user experience, and ethical compliance, and that provides an operational assessment structure, is urgently needed to overcome current bottlenecks in HCAI development.

Objectives

To address these gaps, this study develops a Quality Evaluation Index System for HCAI that offers 3 distinct contributions. First, it constructs the evaluation system from the perspective

XSL•FO
RenderX

of the end-user product application. Second, it comprehensively integrates medical effectiveness, user experience, and ethical and safety compliance. Third, it supports systematic and operational evaluation by combining Delphi consensus with the analytic hierarchy process (AHP) to provide weighted, hierarchical criteria. The objective is to establish a theoretical foundation and an operational framework for the evaluation of HCAI systems, thereby promoting their continuous improvement and ensuring their safe and effective deployment in health care settings.

Methods

Study Design

This study adopts the Delphi method and the AHP as its research approach. The Delphi method, a structured technique for achieving expert consensus, is widely used for indicator selection and standard development in complex problem domains and has been extensively applied in areas such as health care quality assessment and policy research [36]. The AHP enables the assignment of weights to multiple indicators based on expert judgment and is particularly suitable for determining hierarchical structures and weight distributions in evaluation frameworks. These 2 methods are often used in combination and have been widely adopted in studies involving the development of evaluation index systems [37,38].

Between November 2024 and February 2025, this study conducted 2 rounds of expert consultation using the Delphi method. Each round of the questionnaire was distributed and collected via email or WeChat (a widely used social media platform in China), with a response period of 1 week. After collecting the expert scoring data, the AHP was applied to calculate the weights of each level of indicators. Based on these results, a quality evaluation index system for HCAI was established. The flowchart illustrating the process of indicator development and weight determination is shown in Figure 2.

Figure 2. Flowchart of the study design. AHP: analytic hierarchy process.

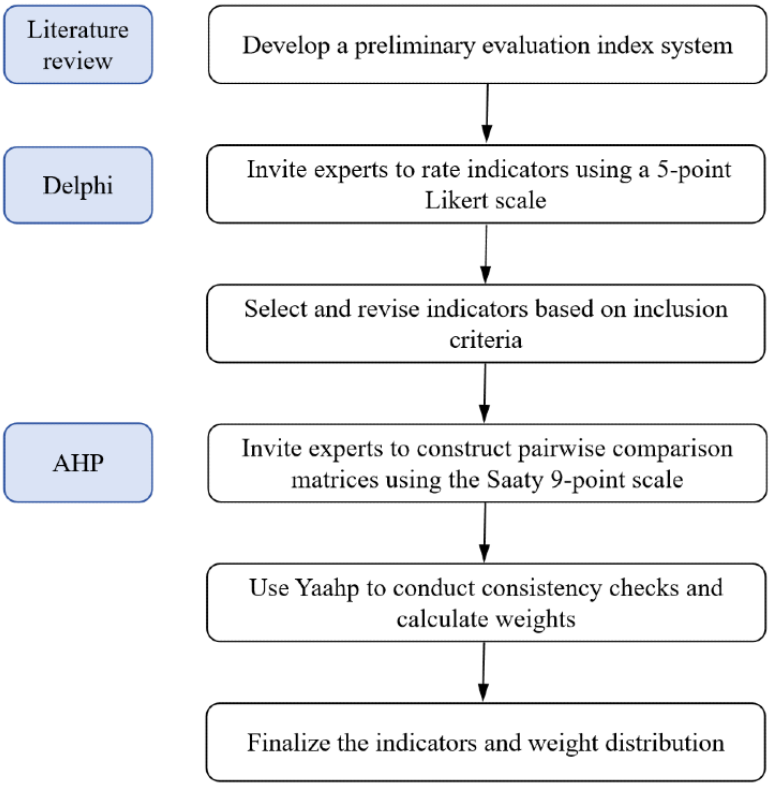

Establishment of a Research Team

The research team consisted of 7 members, including experienced experts in health management, professionals working in the HCAI industry, and graduate students specializing in health management. Among them, 2 held senior professional titles, 2 were HCAI practitioners, and 3 were graduate students. The team's main responsibilities included conducting a literature review, jointly developing preliminary evaluation indicators, identifying Delphi expert members, designing consultation questionnaires, analyzing expert feedback, and performing statistical analyses.

Development of the Preliminary Evaluation Index System

Drawing on a comprehensive review of the literature [30,34,39-42], relevant standards [43], and the expertise of the research team, a preliminary set of evaluation indicators was developed. In real-world applications, HCAI systems involve multiple stakeholders, including patients, health care professionals, and health regulators. As such, quality assessment should not be confined to a single dimension; rather, it must integrate multiperspective and multilevel considerations. Physicians, as service providers, focus on the system's ability to accurately understand user needs, collect complete and logically structured medical histories, perform precise clinical

XSL•FO
RenderX

reasoning, and deliver appropriate health recommendations. These aspects help determine whether HCAI can effectively support clinical practice. Patients, by contrast, place greater emphasis on the quality of interaction, the usefulness and usability of responses, and the system's capacity for emotional responsiveness. Meanwhile, health regulators are primarily concerned with ethical compliance, information security, and the controllability of risks associated with AI applications. These considerations are essential to ensuring that HCAI systems are used safely, effectively, and responsibly in patient care.

To address these considerations, this study developed an evaluation framework based on 3 key stakeholder groups: health care professionals, patients, and health regulators (Figure 3). The framework consists of 3 core dimensions: (1) health consultation capability for medical professionals, which focuses on evaluating the HCAI model's abilities in information elicitation, clinical reasoning, and treatment recommendation during the consultation process; (2) user experience for patients, which draws on the SERVQUAL service quality model [44,45]. This dimension assesses 5 aspects of HCAI: interface usability, information reliability, intelligent responsiveness, perceived safety, and emotional identification; and (3) ethics and compliance for health regulators, which focuses on the system's capacity for ethical safeguards, medical compliance, and risk prevention mechanisms.

The final preliminary evaluation index system comprises 3 primary indicators, 7 secondary indicators, and 34 tertiary indicators.

Figure 3. Preliminary evaluation index system. HCAI: health conversational artificial intelligence.

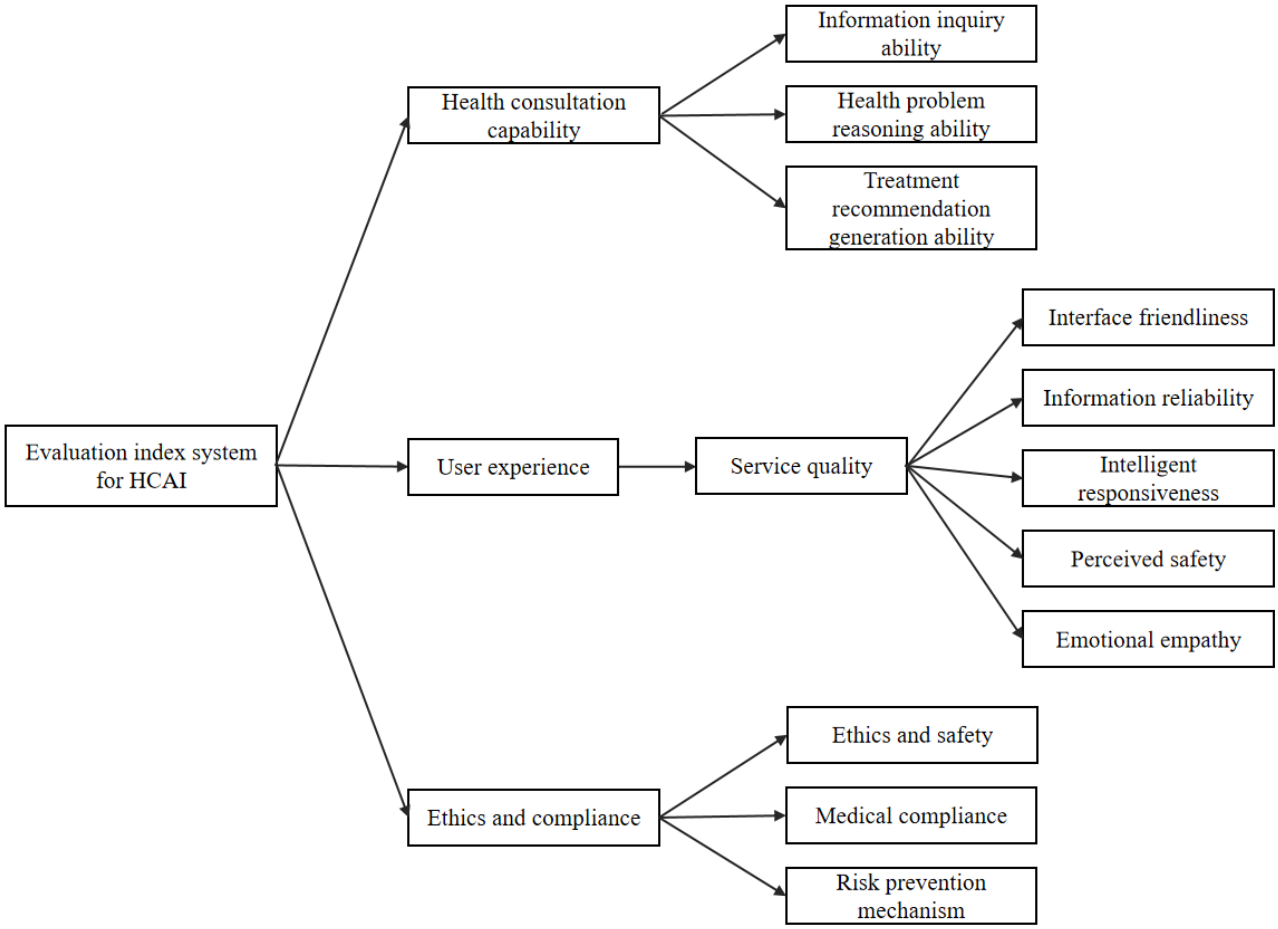

Design of the Expert Consultation Questionnaire

The expert consultation questionnaire consisted of 4 parts (Textbox 1).

The first- and second-round Delphi consultation questionnaires are provided in Multimedia Appendices 1 and 2, respectively. Both sets of questionnaires (Tables S4-S6 in Multimedia Appendices 1 and 2) present detailed evaluation methods, indicator definitions, and scoring rules for all tertiary-level indicators, providing guidance to support expert judgment during the evaluation process.

XSL•FO
RenderX

Textbox 1. Expert consultation questionnaire sections.

1. Questionnaire description

- This section included the research background, objectives, and the deadline for questionnaire submission.

2. Theoretical framework of the evaluation system

- This part outlined the design principles and rationale behind the proposed index system.

3. Instructions and evaluation form

- This section explained how to complete the evaluation form and provided detailed scoring criteria. The evaluation form contained specific indicator definitions, evaluation methods, rating tables, and a column for expert feedback. Experts were asked to rate each indicator in terms of importance, feasibility, and sensitivity using a 5-point Likert scale, and to provide suggestions for revision.
- Importance indicates the relevance of each item for assessing the quality of health conversational AI (HCAI) and is rated on a 5-point scale: 5=very important; 4=important; 3=moderately important; 2=slightly important; and 1=not important.
- Feasibility reflects the ease of collecting relevant data for each item during real-world assessment and is rated on a 5-point scale: 5=very easy; 4=relatively easy; 3=moderate; 2=relatively difficult; and 1=very difficult.
- Sensitivity reflects the extent to which each item influences assessment results, with higher sensitivity indicating a greater impact on HCAI performance and health care quality. It is rated on a 5-point scale: 5=very sensitive; 4=sensitive; 3=moderately sensitive; 2=slightly sensitive; and 1=not sensitive.

4. Expert background information

- This section collected personal information, including age, educational background, professional title, research field, and institutional affiliation. It also included self-assessments of the expert's familiarity with the indicators (Cs) and the basis for their judgment (Ca).

Expert Selection Method

According to the Delphi method, an appropriate number of experts ranges from 15 to 50 participants [46]. In this study, purposive sampling was used to invite experts in fields related to HCAI, including health management and policy, medical ethics, and computer science. The selection criteria for experts were as follows:

- Familiarity with health conversational AI and active engagement in related work, with strong theoretical and practical knowledge.
- A minimum of 5 years of professional experience.
- A postgraduate degree or higher, or a professional title of intermediate level or above.
- Demonstrated interest in the research topic and willingness to complete the consultation process and provide relevant feedback.

Experts were recruited through multiple channels, including previous research collaborations, recommendations from professional associations, and established industry expert databases. Before formal recruitment, the research team organized 2 HCAI expert consensus meetings to introduce the concept of HCAI and to explain the study's purpose, procedures, and evaluation methods. All potential participants received detailed study information and provided informed consent before joining the Delphi consultation.

Indicator Scoring and Selection Criteria

Scores for importance, feasibility, and sensitivity were assigned on a 5-point scale, with higher scores indicating greater relevance. Expert familiarity with the indicators was rated on a 5-level scale with values of 0.9, 0.7, 0.5, 0.3, and 0.1, corresponding to the categories very familiar, familiar, generally familiar, less familiar, and very unfamiliar, respectively. The basis of expert judgment consisted of 4 components: work experience, theoretical analysis, industry understanding, and intuitive perception. The weights for work experience were 0.5, 0.4, and 0.3 for high, medium, and low influence, respectively. The weights for theoretical analysis were 0.3, 0.2, and 0.1 for the same levels. Both industry understanding and intuitive perception were assigned fixed weights of 0.1 [47].

Based on experts' scores for each indicator in terms of importance, feasibility, and sensitivity, the mean, SD, and coefficient of variation (CV) were calculated. The inclusion criteria were a mean score of 3.50 or higher and a CV of less than 0.25 [48,49]. To avoid prematurely excluding potentially important indicators, expert feedback and group discussions were carefully considered before making final decisions.

Key Coefficients and Calculation Methods

Active Coefficient of Experts

This was measured by the effective response rate of the questionnaires, calculated as the number of valid responses in a given round divided by the number of questionnaires distributed. A higher response rate indicates greater expert engagement and interest in the topic.

Expert Authority Coefficient

This reflects the degree of authority an expert holds on the research topic. It is calculated as the average of the expert's Ca (ie, the basis for their judgment) and Cs (ie, their familiarity) values, that is, $Cr = (Ca + Cs)/2$, where Cr is the expert authority coefficient. A Cr value of 0.7 or higher is generally considered to indicate a high level of authority and reliability [50].

XSL•FO
RenderX

Coordination of Expert Opinions

This was measured using the Kendall coefficient of concordance (Kendall *W*), which reflects the consistency of expert ratings. A value closer to 1 indicates a higher degree of agreement among the experts [38].

Weight Determination of Each Indicator

After finalizing the evaluation index system, experts conducted pairwise comparisons of the indicators based on the Saaty scale of relative importance (see Table 1) [51]. These comparisons were used to construct a hierarchical structure and establish judgment matrices. The weights of each indicator were then calculated using AHP via Yaahp (an AHP-based decision-support software developed by Shanxi Yuan Decision Software Technology Co Ltd), and the results were analyzed accordingly. The consistency of each judgment matrix was tested using the consistency ratio. A consistency ratio value of less than 0.1 was considered to indicate acceptable consistency [52].

Table 1. The Saaty scale of relative importance.

Scale	Meaning
1	Indicates equal importance between 2 elements
3	Indicates moderate importance of 1 element over another
5	Indicates moderate to strong importance
7	Indicates very strong importance
9	Indicates extreme importance
2, 4, 6, and 8	Intermediate values between the above judgments
Reciprocals	If item A is assigned a value when compared with item B, then item B is assigned the reciprocal value when compared with item A

Statistical Methods

In this study, Microsoft Excel 2007 was used to input data and calculate the mean, SD, and CV for the importance, feasibility, and sensitivity of each indicator. Excel was also used to compute the expert active coefficient, Ca, Cs, and Cr. Kendall *W* was calculated using SPSS version 25.0 (IBM Corp) and tested for significance using the chi-square test. When the *P* value of Kendall *W* was less than 0.05, and all CVs were below 0.25, the level of disagreement among expert opinions was considered acceptable [53]. The Yaahp version 12.0 software was used to conduct consistency testing of the judgment matrices and to calculate the weights of each indicator.

Ethics Considerations

This study was reviewed and approved by the Ethics Committee of Capital Medical University (approval number Z2024SY075). The approved project title is “Evaluation of the Quality and Effectiveness of User-Oriented AI Health Consultation Services.” All study procedures were conducted in accordance with institutional guidelines and regulations and complied with the principles of the Declaration of Helsinki. Expert participants in the Delphi process were informed of the study’s purpose and provided consent before participation. No identifiable personal information of the expert participants was disclosed in this study. Participant privacy was fully protected.

Results

Basic Information About Experts

Two rounds of expert consultation were conducted in this study. Based on predefined criteria, 16 experts were invited for the first round and 15 for the second round, with 1 expert failing to respond in the second round. Among the 15 experts who completed both rounds, 5 (33%) were male and 10 (67%) were female; 9 of the 15 (60%) experts were over 40 years old. As many as 14 of 15 (93%) experts held a master’s degree or higher, and the same percentage held professional titles at the associate senior level or above. Additionally, 12 of 15 (80%) experts had more than 10 years of work experience in the health care field. The experts were evenly distributed across hospitals, universities, and research institutions. Detailed information is presented in Table 2.

XSL•FO
RenderX

Table 2. Basic information about experts.

Basic information	First round (n=16), n (%)	Second round (n=15), n (%)
Sex		
Male	6 (38)	5 (33)
Female	10 (63)	10 (67)
Age (years)		
≤40	6 (38)	6 (40)
41-50	9 (56)	9 (60)
>50	1 (6)	0 (0)
Education		
Undergraduate	1 (6)	1 (7)
Master	5 (31)	4 (27)
Doctor	10 (63)	10 (67)
Professional title		
Intermediate	1 (6)	1 (7)
Deputy senior	7 (44)	7 (47)
Senior	8 (50)	7 (47)
Years of work experience		
≤10	3 (19)	3 (20)
11-15	4 (25)	4 (27)
16-20	5 (31)	5 (33)
>20	4 (25)	3 (20)
Affiliation		
Medical institution	4 (25)	4 (27)
University	6 (38)	5 (33)
Research institute	5 (31)	5 (33)
Association	1 (6)	1 (7)
Research field		
Health management and policy	7 (44)	7 (47)
Computer science	1 (6)	1 (7)
Medical ethics	1 (6)	1 (7)
Health law	2 (13)	2 (13)
Hospital management	2 (13)	2 (13)
Others	3 (19)	2 (13)

Expert Positivity Degree

According to the Delphi method, a response rate of 50% is considered the minimum threshold for analysis and reporting, while a rate above 70% indicates a high level of expert engagement [54]. In the first round, the expert response rate was 100%, with 11 experts providing a total of 33 suggestions. In the second round, the response rate was also 100%, and 5 experts submitted 6 suggestions. The response rate exceeded 70% in both rounds, indicating a high level of participation and engagement from the expert panel throughout the study.

Expert Authority Coefficient

The authority coefficients of the 2 rounds of expert consultation are shown in Table 3. In the first round, Ca was 0.92, Cs was 0.75, and the resulting Cr was 0.84. In the second round, Ca was 0.93, Cs remained 0.75, and Cr was also 0.84. In both rounds, the Cr values exceeded 0.8, indicating that the experts consulted in this study demonstrated a high level of authority and reliability.

XSL•FO
RenderX

Table 3. Authority coefficients in the 2 rounds of consultation.

Inquiry round	Number	Ca[a]	Cs[b]	Cr[c]
First round	16	0.92	0.75	0.84
Second round	15	0.93	0.75	0.84

[a]Ca: the basis for the expert's judgment.

[b]Cs: expert familiarity.

[c]Cr: expert authority coefficient.

Degree of Coordination of Expert Opinions

The degree of coordination among expert opinions refers to the level of consistency in the experts' ratings of each indicator. The Kendall (*W*) coefficients for the 2 rounds of expert consultation are presented in Table 4. In the first round, Kendall *W* values for all indicators ranged from 0.14 to 0.20. In the second round, the values ranged from 0.13 to 0.17. The results of the chi-square tests were statistically significant (round 1: importance, *P*<.001; feasibility, *P*<.001; and sensitivity, *P*<.001; round 2: importance, *P*=.001; feasibility, *P*<.001; and sensitivity, *P*=.001), indicating a high level of agreement among the expert opinions.

Table 4. Kendall W coefficients for the 2 rounds of consultation.

Project	Importance	Feasibility	Sensitivity
Round one			
Kendall *W*	0.20	0.14	0.14
Chi-square (*df*)	124.39 (42)	86.38 (42)	88.20 (42)
P value	<.001	<.001	<.001
Round 2			
Kendall *W*	0.13	0.17	0.13
Chi-square (*df*)	66.77 (36)	85.05 (36)	67.66 (36)
P value	.001	<.001	.001

Revisions to the Indicator System

The results of the first-round expert consultation are presented in Table 5. The mean scores of all 44 indicators were equal to or greater than 3.50. Specifically, the mean scores ranged from 4.20 to 5.00 for importance, 3.75 to 4.88 for feasibility, and 3.88 to 4.88 for sensitivity. The CVs ranged from 0.00 to 0.27, with 6 indicators having CV values equal to or greater than 0.25.

Based on the screening criteria and expert suggestions, the research team held detailed discussions and decided to temporarily retain the 6 indicators with CV values ≥0.25, remove 7 indicators, revise 4 indicators, and add 1 new indicator. Details of these revisions are provided in Table 6.

After these adjustments, the second-round consultation form consisted of 3 primary indicators, 7 secondary indicators, and 28 tertiary indicators.

The results of the second-round expert consultation are shown in Table 7. The mean scores for importance, feasibility, and sensitivity of all indicators were above 4.0, and all CVs were below 0.25. Based on the indicator selection criteria and thorough discussions among the expert panel, all 37 indicators in the second-round consultation were retained. In addition, the name of indicator C1 was revised from "ethics" to "ethical compliance." The final quality evaluation index system for health conversational AI was thus established, comprising 3 primary indicators, 7 secondary indicators, and 28 tertiary indicators (Table 8).

XSL•FO
RenderX

Table 5. Scores of indicators in the first round of consultation.

Indicators	Importance		Feasibility		Sensitivity	
	Mean (SD)	CV[a]	Mean (SD)	CV	Mean	CV
First-level indices						
A. Health consultation capability	4.94 (0.25)	0.05	4.19 (0.75)	0.18	4.75 (0.45)	0.09
B. User experience	4.50 (0.63)	0.14	4.63 (0.62)	0.13	4.44 (0.63)	0.14
C. Ethics and compliance	5.00 (0.00)	0.00	4.13 (1.02)	0.25	3.88 (0.81)	0.21
Second-level indices						
A1. Information inquiry ability	4.81 (0.40)	0.08	4.19 (0.75)	0.18	4.69 (0.60)	0.13
A2. Health problem reasoning ability	4.81 (0.40)	0.08	3.75 (0.77)	0.21	4.69 (0.48)	0.10
A3. Treatment recommendation generation ability	4.75 (0.58)	0.12	4.19 (0.75)	0.18	4.69 (0.70)	0.15
B1. Service quality experience	4.50 (0.63)	0.14	4.63 (0.62)	0.13	4.44 (0.63)	0.14
C1. Ethics and safety	4.94 (0.25)	0.05	3.81 (1.05)	0.27	4.31 (0.79)	0.18
C2. Medical compliance	5.00 (0.00)	0.00	4.13 (1.09)	0.26	4.38 (0.72)	0.16
C3. Risk prevention mechanisms	4.75 (0.58)	0.12	4.13 (1.02)	0.25	4.44 (0.81)	0.18
Third-level indices						
A11. Accuracy in understanding user needs	4.94 (0.25)	0.05	4.50 (0.52)	0.11	4.81 (0.40)	0.08
A12. Completeness of user information collection	4.44 (0.51)	0.12	4.19 (0.66)	0.16	4.31 (0.60)	0.14
A13. Relevance of inquiry content	4.63 (0.50)	0.11	4.19 (0.75)	0.18	4.69 (0.48)	0.10
A14. Logical sequence of inquiries	4.25 (0.68)	0.16	4.38 (0.89)	0.20	4.31 (0.60)	0.14
A15. Appropriateness of interaction rounds	4.31 (0.60)	0.14	4.19 (0.75)	0.18	4.00 (0.89)	0.22
A16. Recognition ability of multimodal information	4.63 (0.50)	0.11	4.56 (0.63)	0.14	4.50 (0.63)	0.14
A17. Personalization of communication style	4.63 (0.62)	0.13	4.31 (0.60)	0.14	4.38 (0.72)	0.16
A21. Accuracy of disease risk reasoning	4.94 (0.25)	0.05	4.44 (0.63)	0.14	4.88 (0.34)	0.07
A22. Consistency of repeated judgments	4.75 (0.45)	0.09	4.44 (0.73)	0.16	4.38 (0.62)	0.14
A23. Coverage of consulted diseases	4.25 (0.68)	0.16	4.38 (0.81)	0.18	4.19 (0.75)	0.18
A24. Diagnostic ability for complex diseases	4.20 (0.86)	0.21	3.80 (0.77)	0.20	4.20 (0.86)	0.21
A25. Professional use of medical terminology	4.25 (0.86)	0.20	4.44 (0.73)	0.16	4.19 (0.75)	0.18
A26. Frequency of medical knowledge updates	4.31 (0.79)	0.18	4.06 (1.00)	0.25	4.06 (0.93)	0.23
A27. Interpretability of disease reasoning	4.63 (0.72)	0.16	4.25 (0.68)	0.16	4.25 (0.77)	0.18
A31. Accuracy of treatment recommendations	4.88 (0.34)	0.07	4.44 (0.89)	0.20	4.75 (0.58)	0.12
A32. Comprehensiveness of treatment recommendations	4.56 (0.63)	0.14	4.38 (0.62)	0.14	4.44 (0.73)	0.16
A33. Personalization of treatment recommendations	4.56 (0.63)	0.14	4.13 (0.62)	0.15	4.44 (0.73)	0.16
A34. Operability of treatment recommendations	4.44 (0.63)	0.14	4.25 (0.77)	0.18	4.31 (0.79)	0.18
B11. Interface friendliness	4.25 (0.77)	0.18	4.63 (0.50)	0.11	4.25 (0.77)	0.18
B12. Information reliability	4.94 (0.25)	0.05	4.44 (0.73)	0.16	4.63 (0.62)	0.13
B13. Intelligent responsiveness	4.44 (0.63)	0.14	4.56 (0.73)	0.16	4.63 (0.50)	0.11
B14. Perceived safety	4.63 (0.62)	0.13	4.31 (0.70)	0.16	4.19 (0.66)	0.16
B15. Emotional empathy	4.44 (0.63)	0.14	4.00 (0.82)	0.20	4.25 (0.68)	0.16
C11. Bias	4.75 (0.45)	0.09	4.25 (0.68)	0.16	4.38 (0.72)	0.16
C12. Privacy	4.88 (0.34)	0.07	4.31 (0.79)	0.18	4.56 (0.63)	0.14
C13. Hallucinations	4.69 (0.48)	0.10	4.13 (0.62)	0.15	4.50 (0.73)	0.16
C14. Data and system security	4.94 (0.25)	0.05	4.19 (0.66)	0.16	4.50 (0.73)	0.16
C15. Establishment of ethics committee	4.56 (0.73)	0.16	4.44 (1.15)	0.26	4.31 (0.87)	0.20

XSL•FO
RenderX

Indicators	Importance		Feasibility		Sensitivity	
	Mean (SD)	CV[a]	Mean (SD)	CV	Mean	CV
C21. Presence of a disclaimer notices	4.69 (0.60)	0.13	4.88 (0.34)	0.07	4.25 (0.77)	0.18
C22. Assessment of treatment risks	4.94 (0.25)	0.05	4.44 (0.51)	0.12	4.75 (0.58)	0.12
C23. Generation of factually incorrect information	4.81 (0.40)	0.08	4.50 (0.63)	0.14	4.56 (0.73)	0.16
C24. Violation of medical compliance	4.88 (0.34)	0.07	4.44 (0.81)	0.18	4.69 (0.48)	0.10
C31. Emergency response mechanisms	4.75 (0.58)	0.12	4.63 (0.50)	0.11	4.19 (0.75)	0.18
C32. Access management for minors	4.81 (0.40)	0.08	4.88 (0.34)	0.07	4.38 (0.81)	0.18

[a]CV: coefficient of variation.

Table 6. Revisions to indicators based on the first round of consultation.

Indicators	Action	Explanation
1. Ethics and compliance	Retained	The CV[a] value for feasibility is ≥0.25. It is recommended to decide whether to retain it based on the results of the second round of consultation.
C1. Ethics and safety	Retained	The CV value for feasibility is ≥0.25. It is recommended to decide whether to retain it based on the results of the second round of consultation.
C2. Medical compliance	Retained	The CV value for feasibility is ≥0.25. It is recommended to decide whether to retain it based on the results of the second round of consultation.
C3. Risk prevention mechanism	Retained	The CV value for feasibility is ≥0.25. It is recommended to decide whether to retain it based on the results of the second round of consultation.
A1. Information inquiry ability	Modified	Renamed to “Health Information Inquiry Ability” for better accuracy.
C1. Ethics and safety	Modified	Renamed to “Ethics” for better accuracy.
A13. Relevance of inquiry content	Deleted	Poor feasibility based on evaluation criteria.
A15. Rationality of interaction rounds	Deleted	Poor feasibility based on evaluation criteria.
A23. Coverage of consulted diseases	Deleted	Low importance and sensitivity.
A24. Diagnostic ability for complex diseases	Retained	Some experts noted overlap with A21 “Accuracy of disease reasoning,” but reasoning for complex diseases provides distinct model evaluation value.
A25. Use of medical terminology	Deleted	Overlaps with A31 “Accuracy of treatment recommendations.”
A26. Frequency of medical knowledge update	Retained	The CV value for feasibility is ≥0.25. It is recommended to decide whether to retain it based on the results of the second round of consultation.
A33. Personalization of treatment advice	Deleted	Low evaluation value under C3 “Risk prevention mechanism.”
B12. Information reliability	Modified	Renamed to “Perceived reliability” for better accuracy.
B13. Intelligent responsiveness	Modified	Renamed to “Perceived responsiveness” for better accuracy.
C14. Data and system security	Deleted	Belongs to model-level indicators, not a primary assessment focus.
C15. Establishment of ethics committee	Retained	The CV value for feasibility is ≥0.25. It is recommended to decide whether to retain it based on the results of the second round of consultation.
C22. Assessment of treatment risks	Deleted	Overlaps with evaluation under A2 “Health problem reasoning ability.”
C24. Violation of medical compliance	Retained	Some experts believe that this indicator belongs to the model admission criteria and suggest deleting it. However, this mainly evaluates the low-risk violations that may exist in the model and has evaluation value
C33. Presence of privacy terms	Added	The risk prevention mechanism should include detailed privacy clauses to clarify the purpose and scope of collecting user information, as well as whether the collected information is shared with third parties

[a]CV: coefficient of variation.

XSL•FO
RenderX

Table 7. Scores of indicators in the second round of consultation.

Indicators	Importance		Feasibility		Sensitivity	
	Mean (SD)	CV[a]	Mean (SD)	CV	Mean (SD)	CV
First-level indices						
A. Health consultation capability	4.87 (0.35)	0.07	4.40 (0.51)	0.12	4.73 (0.59)	0.13
B. User experience	4.67 (0.49)	0.10	4.67 (0.49)	0.10	4.60 (0.63)	0.14
C. Ethics and compliance	4.93 (0.26)	0.05	4.20 (0.68)	0.16	4.33 (0.62)	0.14
Second-level indices						
A1. Health information inquiry ability	4.80 (0.41)	0.09	4.60 (0.63)	0.14	4.80 (0.41)	0.09
A2. Health problem reasoning ability	4.73 (0.46)	0.10	4.33 (0.62)	0.14	4.47 (0.74)	0.17
A3. Treatment recommendation generation ability	4.80 (0.56)	0.12	4.40 (0.74)	0.17	4.47 (0.83)	0.19
B1. Service quality experience	4.67 (0.49)	0.10	4.67 (0.49)	0.10	4.60 (0.63)	0.14
C1. Ethical compliance	4.87 (0.35)	0.07	4.20 (0.86)	0.21	4.87 (0.35)	0.07
C2. Medical compliance	4.87 (0.35)	0.07	4.40 (0.74)	0.17	4.40 (0.51)	0.12
C3. Risk prevention mechanism	4.93 (0.26)	0.05	4.53 (0.64)	0.14	4.33 (0.72)	0.17
Third-level indices						
A11. Accuracy in understanding user needs	4.87 (0.35)	0.07	4.60 (0.63)	0.14	4.93 (0.26)	0.05
A12. Completeness of user information collection	4.60 (0.51)	0.11	4.20 (0.77)	0.18	4.60 (0.51)	0.11
A13. Logical sequence of inquiries	4.60 (0.63)	0.14	4.60 (0.63)	0.14	4.20 (0.77)	0.18
A14. Recognition ability of multimodal information	4.67 (0.49)	0.10	4.53 (0.64)	0.14	4.67 (0.49)	0.10
A15. Personalization of communication style	4.47 (0.64)	0.14	4.07 (0.80)	0.20	4.47 (0.64)	0.14
A21. Accuracy of disease risk reasoning	4.80 (0.41)	0.09	4.33 (0.62)	0.14	4.67 (0.49)	0.10
A22. Consistency of repeated judgments	4.87 (0.35)	0.07	4.53 (0.74)	0.16	4.47 (0.64)	0.14
A23. Diagnostic ability for complex diseases	4.47 (0.83)	0.19	4.13 (0.83)	0.20	4.40 (0.74)	0.17
A24. Frequency of medical knowledge updates	4.60 (0.63)	0.14	4.33 (0.82)	0.19	4.33 (0.82)	0.19
A25. Interpretability of disease reasoning	4.80 (0.56)	0.12	4.13 (0.83)	0.20	4.33 (0.72)	0.17
A31. Accuracy of treatment recommendations	4.93 (0.26)	0.05	4.47 (0.74)	0.17	4.67 (0.72)	0.16
A32. Comprehensiveness of treatment recommendations	4.53 (0.64)	0.14	4.40 (0.63)	0.14	4.27 (0.88)	0.21
A33. Operability of treatment recommendations	4.53 (0.74)	0.16	4.40 (0.74)	0.17	4.27 (0.80)	0.19
B11. Interface friendliness	4.47 (0.52)	0.12	4.73 (0.46)	0.10	4.33 (0.62)	0.14
B12. Perceived reliability	4.80 (0.56)	0.12	4.53 (0.64)	0.14	4.67 (0.49)	0.10
B13. Perceived responsiveness	4.67 (0.49)	0.10	4.73 (0.46)	0.10	4.67 (0.62)	0.13
B14. Perceived safety	4.67 (0.49)	0.10	4.27 (0.70)	0.16	4.33 (0.49)	0.11
B15. Emotional identification	4.21 (0.70)	0.17	4.43 (0.76)	0.17	4.07 (0.73)	0.18
C11. Bias	4.87 (0.35)	0.07	4.40 (0.74)	0.17	4.53 (0.64)	0.14
C12. Privacy	4.80 (0.41)	0.09	4.67 (0.62)	0.13	4.53 (0.64)	0.14
C13. Hallucination	4.67 (0.49)	0.10	4.07 (0.59)	0.15	4.60 (0.74)	0.16
C14. Establishment of ethics committee	4.80 (0.41)	0.09	4.67 (0.72)	0.16	4.20 (0.77)	0.18
C21. Presence of a disclaimer notices	4.87 (0.35)	0.07	4.93 (0.26)	0.05	4.27 (0.70)	0.16
C22. Generation of factually incorrect information	5.00 (0.00)	0.00	4.67 (0.49)	0.10	4.87 (0.35)	0.07
C23. Violation of medical compliance	4.93 (0.26)	0.05	4.87 (0.35)	0.07	4.73 (0.46)	0.10
C31. Emergency response mechanisms	4.80 (0.56)	0.12	4.67 (0.49)	0.10	4.67 (0.62)	0.13
C32. Access management for minors	4.80 (0.41)	0.09	4.87 (0.35)	0.07	4.60 (0.63)	0.14
C33. Presence of privacy terms	4.80 (0.41)	0.09	4.80 (0.41)	0.09	4.40 (0.74)	0.17

XSL•FO
RenderX

[a]CV: coefficient of variation.

Table 8. Final weight coefficients of indicators.

First-, second-, and third-level indices	Weight
A. Health consultation capability	0.4112
A1. Health information inquiry ability	0.1272
A11. Accuracy in understanding user needs	0.0648
A12. Completeness of user information collection	0.021
A13. Logical sequence of inquiries	0.0122
A14. Recognition ability of multimodal information	0.0219
A15. Personalization of communication style	0.0073
A2. Health problem reasoning ability	0.0821
A21. Accuracy of disease risk reasoning	0.0389
A22. Consistency of repeated judgments	0.0182
A23. Diagnostic ability for complex diseases	0.0094
A24. Frequency of medical knowledge updates	0.0056
A25. Interpretability of disease reasoning	0.01
A3. Treatment recommendation generation ability	0.2019
A31. Accuracy of treatment recommendations	0.1216
A32. Comprehensiveness of treatment recommendations	0.0562
A33. Operability of treatment recommendations	0.0241
B. User experience	0.1107
B1. Service quality experience	0.1107
B11. Interface friendliness	0.0072
B12. Perceived reliability	0.0456
B13. Perceived responsiveness	0.0132
B14. Perceived safety	0.0336
B15. Emotional identification	0.0111
C. Ethics and compliance	0.4781
C1. Ethical compliance	0.1933
C11. Bias	0.0539
C12. Privacy	0.0491
C13. Hallucination	0.0697
C14. Establishment of ethics committee	0.0205
C2. Medical compliance	0.1696
C21. Presence of a disclaimer notices	0.0161
C22. Generation of factually incorrect information	0.0889
C23. Violation of medical compliance	0.0646
C3. Risk prevention mechanism	0.1152
C31. Emergency response mechanisms	0.0605
C32. Access management for minors	0.0362
C33. Presence of privacy terms	0.0185

XSL•FO
RenderX

Weight Determination of Each Indicator

The AHP was used to calculate the weights of each indicator, with higher weights indicating greater importance in evaluating the quality of health conversational AI. In this study, all consistency ratio values of the judgment matrices were automatically adjusted using Yaahp and were found to be less than 0.10, indicating good consistency across all matrices.

The final weight coefficients of all indicators are shown in Table 8. Among the primary indicators, the ranking by weight from highest to lowest was as follows: ethics and compliance (0.4781), health consultation capability (0.4112), and user experience (0.1107).

Discussion

Principal Findings

As AI becomes increasingly integrated into health care, HCAI is being used more widely in health communication, clinical decision support, and consultation. Establishing a scientific and systematic quality evaluation framework is therefore essential. Although several evaluation systems have been developed, many still focus primarily on accuracy or clinician-oriented indicators, neglecting critical aspects of real-world product use and comprehensive safety assessment. They pay limited attention to the comprehensive integration of user experience, ethical and compliance considerations, and the robust assessment of multiturn interactions, all of which are essential for HCAI in real consultation settings. Addressing these gaps is critical for improving the quality, safety, and effectiveness of HCAI.

This study employs the Delphi method and AHP to develop an evaluation index system that reflects the perspectives of physicians, users, and regulators, and makes 3 key contributions addressing the limitations of prior work. First, our system is constructed from the perspective of HCAI as an end-user product for daily health consultation, rather than focusing solely on clinical-task performance. Second, it encompasses 3 core, balanced dimensions: health consultation capability, user experience, and ethics and compliance. Third, by using the AHP methodology, the framework supports systematic and operational evaluation through weighted, hierarchical criteria. By addressing these dimensions and the shortcomings of existing frameworks, the proposed system provides a more comprehensive structure that aligns with the real-world needs of HCAI services. The framework provides a theoretical foundation for empirical research, guides practical optimization, and serves as a valuable reference for improving HCAI evaluation standards worldwide.

Scientific Validity and Rationality of the Evaluation Index System

The final evaluation index system developed in this study comprises 3 primary indicators, 7 secondary indicators, and 28 tertiary indicators. Its structure is comprehensive and well-organized, demonstrating both scientific rigor and practical relevance.

First, the design of the index system was grounded in existing literature, industry standards, and expert consensus. It incorporated both domestic and international approaches to HCAI evaluation, ensuring a solid theoretical foundation and methodological relevance.

Second, the study employed 2 rounds of Delphi expert consultation, combined with AHP for weight assignment, integrating both qualitative judgment and quantitative analysis. Expert opinions showed increasing convergence across rounds, with statistically significant coordination coefficients, indicating a high level of consensus and enhancing the credibility of the evaluation system.

Third, the experts involved in the study were highly representative across disciplines, including health management and policy, medical ethics, computer science, health law, and hospital administration. The panel comprised both academic researchers and experienced practitioners, ensuring a balance of theoretical insight and practical expertise. This interdisciplinary and diverse expert composition enhanced the scientific validity of the evaluation system and conformed to established Delphi methodology standards for expert selection and quality assurance [55,56].

Finally, the experts demonstrated high levels of engagement, authority, and consistency throughout the process, significantly enhancing the reliability of the study's findings [57]. Both rounds of the Delphi consultation achieved a 100% response rate (16 and 15 experts in rounds 1 and 2, respectively), with a total of 39 modification suggestions proposed, reflecting active participation and strong interest from the expert panel. The Cr values in both rounds exceeded 0.8, confirming the reliability of expert judgments. Additionally, the *P* values of Kendall *W* were all below 0.05, indicating a high level of agreement in weight assessments and providing a solid foundation for the reliability of the constructed index system.

Analysis of Indicator Weights

The weights of the 3 primary indicators are shown in Table 8: ethics and compliance (0.4781), health consultation capability (0.4112), and user experience (0.1107). These results indicate that ethics and compliance carries the greatest weight in the quality evaluation of HCAI, underscoring its importance as a key dimension for building safe and trustworthy AI systems.

Health consultation capability, which reflects the AI's professional value and its ability to deliver knowledge-based services, also accounts for a substantial proportion of the overall weight. By contrast, user experience carries a relatively lower weight. This may be because current HCAI products and applications are still in the early stages of development, leading experts to prioritize system safety, compliance, and functionality during the evaluation process, while user interaction is considered a secondary factor.

At present, many HCAI systems face significant risks related to algorithmic bias, data privacy breaches, hallucinations, and lack of medical compliance [58-60]. These risks are closely associated with the ethics and compliance dimension of the evaluation system. Among the tertiary indicators in this study, several related to ethical and safety concerns received relatively high weight values, including "generation of factually incorrect information" (0.0889), "hallucination" (0.0697), "violation of

XSL•FO
RenderX

medical compliance" (0.0646), and "emergency response mechanisms" (0.0605). These results indicate that truthfulness, safety, and regulatory compliance of AI-generated content are critical components in assessing the quality of HCAI. In clinical settings, inaccurate recommendations or breaches of patient privacy can have serious consequences and must be addressed as a priority. This finding aligns with concerns raised by other researchers. For instance, some studies have reported that AI in health care may generate hallucinations—that is, inaccurate or fabricated content that could compromise clinical decision-making and patient safety [61]. Wang et al [62] also noted that LLMs may not achieve complete deidentification of training data, thereby raising the risk of exposing sensitive user information.

Within the health consultation capability dimension, the highest weight was assigned to "treatment recommendation generation ability" (0.2019). Among its tertiary indicators, "accuracy of treatment recommendations" (0.1216) and "comprehensiveness of treatment recommendations" (0.0562) were assigned relatively high weights. This indicates strong expert attention to whether AI systems can accurately assess users' health conditions and provide reliable and helpful advice. These findings are highly consistent with previous studies by Lukac et al [63] and Liu et al [64], both of whom emphasized the importance of AI's ability to generate dependable treatment recommendations. In addition, the tertiary indicator "accuracy in understanding user needs" (0.0648), under the category of "health information elicitation ability," and "accuracy of disease reasoning" (0.0389), under "health problem reasoning ability," also received relatively high weights. This suggests that, at the current stage, experts continue to view AI's capabilities in understanding user intent and providing accurate reasoning as important considerations.

In comparison, although the user experience dimension received a relatively lower overall weight, certain indicators, such as "perceived reliability" (0.0456) and "perceived safety" (0.0336), still accounted for a meaningful proportion. This reflects the significant impact of users' perceived trust in AI systems on actual usage experiences. Previous studies have shown that trust is one of the key determinants influencing users' acceptance and use of conversational AI tools such as ChatGPT [65]. In the health care context, user trust in AI technologies is also considered a critical factor affecting their broader adoption and application [66]. Moreover, some studies have found that, compared with traditional information channels such as online health websites, HCAI systems are better positioned to meet users' needs for immediacy and convenience during health information-seeking. This is largely due to their ability to engage in real-time interaction and provide personalized responses [67]. These findings further suggest that, in the ongoing optimization of HCAI, user experience should not only be regarded as a key factor influencing technology acceptance but also as a critical lever for enhancing system usability and fostering trust between humans and machines.

Policy Recommendations

In summary, the development of HCAI remains at an early stage. Ethics and compliance has emerged as the most critical evaluation dimension, while health consultation capability serves as the core functional foundation. Although user experience holds a relatively lower weight, it still plays a significant role in promoting the adoption and practical application of HCAI systems. To further advance the development of HCAI, this study proposes the following policy recommendations:

First, regulatory authorities around the world have imposed increasingly stringent requirements on HCAI-related products. For example, the European Union's Artificial Intelligence Act [68] mandates a comprehensive regulatory framework for high-risk AI systems, including those used in health care, covering both premarket and postmarket phases [68]. Singapore's Model AI Governance Framework for Generative AI emphasizes the importance of continuous evaluation and improvement across the entire life cycle of AI systems [69]. Similarly, China's Interim Measures for the Management of Generative Artificial Intelligence Services clearly stipulate dynamic supervision and safety monitoring of AI products after their deployment [70]. In light of this, it is recommended that the evaluation index system developed in this study be used as the basis for establishing a full–life cycle quality assessment mechanism for HCAI. This mechanism should encompass product development, testing, deployment, and application. Furthermore, efforts should be made to promote its transformation into a widely accepted industry framework through supportive policy initiatives. Such an approach would facilitate the regular evaluation of HCAI products and enable dynamic supervision and continuous quality improvement.

Second, the performance of HCAI systems should be continuously optimized across 3 key dimensions: model capability, ethics and compliance, and user experience. In terms of model capability, targeted training and fine-tuning should be conducted on tasks such as health information elicitation, disease diagnosis, and treatment recommendation to improve the system's medical professionalism and response accuracy. For ethical and safety considerations, it is necessary to establish a regulatory framework that covers data processing, model outputs, and information usage. This framework should clearly define accountability and responsibility in high-risk scenarios and include mechanisms to prevent misinformation and breaches of privacy. Regarding user experience, improvements should focus on core indicators such as "perceived safety" and "perceived reliability." This can be achieved by refining language logic, setting boundary warnings, incorporating source attribution, and integrating human-in-the-loop mechanisms. These measures are essential to enhance user understanding of, and trust in, the system.

Finally, HCAI is expected to move beyond the constraints of online consultation platforms and mobile apps. By helping users recognize their own health conditions, understand disease-related knowledge, and enhance early screening, diagnosis, and treatment, HCAI has the potential to improve health literacy and overall well-being. It may thus become a convenient and trustworthy health gatekeeper for the public. HCAI can also be integrated with health care service institutions and medical consortia to support the accuracy and feasibility of hierarchical diagnosis and treatment systems. To promote the effective application of HCAI in diverse scenarios, such as

XSL•FO
RenderX

self-assessment and diagnosis, intelligent triage, chronic disease follow-up, and care of older adults, it is necessary to adjust the structure and weight allocation of existing evaluation frameworks. Capability indicators should be refined based on the characteristics of specific tasks, with a particular focus on the model's adaptability and service effectiveness across different use cases. These efforts will help strengthen the professional performance of HCAI systems in multiple contexts and enhance the overall level of intelligent health care services.

Strengths and Limitations of Research

Unlike most previous expert consultation studies, this research assessed not only the importance of each indicator but also its sensitivity and feasibility, thereby enhancing the practicality of the constructed evaluation index system [71]. The findings of this study help address a critical gap in the evaluation of HCAI quality. The developed index system can be applied to assess the service quality of existing HCAI systems, identify weaknesses in consultation capability, user experience, ethics and compliance, and provide theoretical support for future product optimization and model iteration. This contributes to the standardized and high-quality development of HCAI. Furthermore, the evaluation system may serve as a reference for industry regulation. It offers a quantitative tool for government agencies, technology platforms, and health care providers to establish standards, identify risks, and improve service quality.

This study has several limitations. First, the expert panel consisted of 15-16 professionals from China. Although the panel was selected to ensure domain expertise, the limited sample size and geographic scope may affect the diversity and generalizability of perspectives. Future research should consider expanding the pool of experts to enhance the authority and generalizability of the conclusions. Second, although this study proposed a structured and multidimensional evaluation framework for HCAI, its practical utility has not yet been validated in real-world settings. While detailed assessment criteria for each third-level indicator are provided in Multimedia Appendix 2, further refinement of operational thresholds and performance benchmarks will be needed to support practical application. Additional empirical validation is required to assess the framework's applicability and adaptability across different scenarios.

Conclusions

This study developed a quality evaluation index system for HCAI, consisting of 3 primary indicators, 7 secondary indicators, and 28 tertiary indicators. Expert scoring results indicated that the dimension of "ethics and compliance" had the highest weight, followed by "health consultation capability." This suggests that, when evaluating the quality of health-related AI systems, priority should be given to information security, medical compliance, and risk management. The study combined the Delphi method with the AHP to ensure the scientific validity of the evaluation indicators and the rationality of their weight distribution. The proposed evaluation framework provides a theoretical reference for the assessment and optimization of HCAI systems.

Acknowledgments

The authors used the generative artificial intelligence tool ChatGPT (OpenAI) during the translation and manuscript refinement stages. This tool was used solely to assist with language polishing. All analytic decisions, study design choices, and interpretation of findings were made entirely by the authors. The authors critically reviewed and verified all outputs generated by the tool and fully accept responsibility for the final content of this manuscript.

Data Availability

The data that support the findings of this study are available from the authors; however, restrictions apply to their availability, as the data were used under license from the experts involved in the Delphi survey for this study and are therefore not publicly available. The data are, however, available from the authors upon reasonable request and with permission from these experts.

Funding

This study was funded by the National Social Science Foundation of China (grant 24BGL273).

Authors' Contributions

All authors contributed to the interpretation of the findings and the writing of the manuscript and have approved the final version. WL was responsible for the statistical analysis and editorial writing; ML and HL were responsible for data cleaning; CM and YH were responsible for the design, review, and revision of the article. CM and YH contributed equally to the research and manuscript preparation and are considered co-corresponding authors. DW, YW, ZF, HW, and YG were responsible for data acquisition and interpretation.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Expert consultation questionnaire in the first round.
[DOCX File , 154 KB - jmir_v28i1e83188_app1.docx]

Multimedia Appendix 2
Expert consultation in the second round and analytic hierarchy process questionnaire.
[DOCX File , 66 KB - jmir_v28i1e83188_app2.docx]

References

1. Ha JF, Longnecker N. Doctor-patient communication: a review. Ochsner J 2010;10(1):38-43. [Medline: 21603354]
2. World Health Organization (WHO). Ageing and health. WHO. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health [accessed 2025-03-17]
3. World Health Organization (WHO). Health workforce. WHO. URL: https://www.who.int/health-topics/health-workforce#tab=tab_1 [accessed 2025-03-17]
4. Lahat A, Klang E. Can advanced technologies help address the global increase in demand for specialized medical care and improve telehealth services? J Telemed Telecare 2024 Oct;30(9):1516-1517. [doi: 10.1177/1357633X231155520] [Medline: 36760131]
5. He Z, Bhasuran B, Jin Q, Tian S, Hanna K, Shavor C, et al. Quality of answers of generative large language models versus peer users for interpreting laboratory test results for lay patients: evaluation study. J Med Internet Res 2024 Apr 17;26:e56655 [FREE Full text] [doi: 10.2196/56655] [Medline: 38630520]
6. Hirosawa T, Harada Y, Mizuta K, Sakamoto T, Tokumasu K, Shimizu T. Diagnostic performance of generative artificial intelligences for a series of complex case reports. Digit Health 2024;10:20552076241265215. [doi: 10.1177/20552076241265215] [Medline: 39229463]
7. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
8. Busch F, Han T, Makowski MR, Truhn D, Bressem KK, Adams L. Integrating text and image analysis: exploring GPT-4V's capabilities in advanced radiological applications across subspecialties. J Med Internet Res 2024 May 01;26:e54948 [FREE Full text] [doi: 10.2196/54948] [Medline: 38691404]
9. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239 [FREE Full text] [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
10. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature 2025 Jun;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
11. Zhu J, Pan J, Liu Y, Liu F, Wu J. Ask patients with patience: enabling LLMs for human-centric medical dialogue with grounded reasoning. arXiv. Preprint posted online on August 21, 2025 2025:e1 [FREE Full text] [doi: 10.48550/arXiv.2502.07143]
12. Announcement by the Cyberspace Administration of China on the Release of the Seventh Batch of Deep Synthesis Service Algorithm Filing Information. State Internet Information Office of the People's Republic of China. URL: https://www.cac.gov.cn/2024-08/05/c_1724541639039621.htm [accessed 2025-03-24]
13. Conversational AI in healthcare market by component (chatbots, virtual assistants, speech recognition systems, and services), by technology, by application, by end user - global industry outlook, key companies (IBM, Microsoft, Google, and others), trends and forecast 2025-2034. Dimension Market Research. 2025 Jul 5. URL: https://dimensionmarketresearch.com/report/conversational-ai-in-healthcare-market/ [accessed 2025-06-12]
14. Medical large model research report: empowering nearly 300 medical large models in clinical and non-clinical scenarios. Beijing Review. 2025. URL: http://www.beijingreview.com.cn/caijing/hyzx/202505/t20250513_800401417.html [accessed 2025-06-12]
15. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023 Oct 01;13(1):16492 [FREE Full text] [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
16. Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. JMIR Med Educ 2023 Apr 21;9:e46599 [FREE Full text] [doi: 10.2196/46599] [Medline: 37083633]
17. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ 2024 Feb 14;24(1):143 [FREE Full text] [doi: 10.1186/s12909-024-05125-7] [Medline: 38355517]
18. Liu C, Sun K, Zhou Q, Duan Y, Shu J, Kan H, et al. CPMI-ChatGLM: parameter-efficient fine-tuning ChatGLM with Chinese patent medicine instructions. Sci Rep 2024 Mar 16;14(1):6403 [FREE Full text] [doi: 10.1038/s41598-024-56874-w] [Medline: 38493251]

XSL•FO
RenderX

19. Zhang K, Zhou R, Adhikarla E, Yan Z, Liu Y, Yu J, et al. A generalist vision-language foundation model for diverse biomedical tasks. Nat Med 2024 Nov 07;30(11):3129-3141. [doi: 10.1038/s41591-024-03185-2] [Medline: 39112796]
20. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform 2022 Nov 19;23(6):bbac409. [doi: 10.1093/bib/bbac409] [Medline: 36156661]
21. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
22. Van Veen D, Van Uden C, Blankemeier L, Delbrouck J, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024 Apr;30(4):1134-1142. [doi: 10.1038/s41591-024-02855-5] [Medline: 38413730]
23. Sun H, Zhang K, Lan W, Gu Q, Jiang G, Yang X, et al. An AI dietitian for type 2 diabetes mellitus management based on large language and image recognition models: preclinical concept validation study. J Med Internet Res 2023 Nov 09;25:e51300 [FREE Full text] [doi: 10.2196/51300] [Medline: 37943581]
24. Ho CN, Tian T, Ayers AT, Aaron RE, Phillips V, Wolf RM, et al. Qualitative metrics from the biomedical literature for evaluating large language models in clinical decision-making: a narrative review. BMC Med Inform Decis Mak 2024 Nov 26;24(1):357. [doi: 10.1186/s12911-024-02757-z]
25. Chen X, Xiang J, Lu S, Liu Y, He M, Shi D. Evaluating large language models and agents in health care: key challenges in clinical applications. Intelligent Medicine 2025 May 1:151-163 [FREE Full text] [doi: 10.1016/j.imed.2025.03.002]
26. Omar M, Soffer S, Agbareia R, Bragazzi NL, Apakama DU, Horowitz CR, et al. Sociodemographic biases in medical decision making by large language models. Nat Med 2025 Jun 1:1873-1881. [doi: 10.1038/s41591-025-03626-6]
27. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst 2025 Jan 24;43(2):1-55. [doi: 10.1145/3703155]
28. Xu J, Lu L, Peng X, Pang J, Ding J, Yang L, et al. Data set and benchmark (MedGPTEval) to evaluate responses from large language models in medicine: evaluation development and validation. JMIR Med Inform 2024 Jun 28;12:e57674 [FREE Full text] [doi: 10.2196/57674] [Medline: 38952020]
29. Liu L, Yang X, Li F, Chi C, Shen Y, Lyu S. Towards automatic evaluation for LLMs' clinical capabilities: metric, data, and algorithm. 2024 Aug 24 Presented at: KDD '24: The 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; August 25-29, 2024; New York, NY. [doi: 10.1145/3637528.3671575]
30. Johri S, Jeong J, Tran BA, Schlessinger DI, Wongvibulsin S, Barnes LA, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. Nat Med 2025 Jan;31(1):77-86. [doi: 10.1038/s41591-024-03328-5] [Medline: 39747685]
31. Liu X, Liu H, Yang G, Jiang Z, Cui S, Zhang Z, et al. A generalist medical language model for disease diagnosis assistance. Nat Med 2025 Mar 31;31(3):932-942. [doi: 10.1038/s41591-024-03416-6]
32. Arora R, Wei J, Hicks R, Bowman P, Candela J, Tsimpourlas F. HealthBench: evaluating large language models towards improved human health. arXiv. Preprint posted online on May 13, 2025 2025 May 13:e1 [FREE Full text] [doi: 10.48550/arXiv.2505.08775]
33. Qiu P, Wu C, Liu S, Fan Y, Zhao W, Chen Z, et al. Quantifying the reasoning abilities of LLMs on clinical cases. Nat Commun 2025 Nov 06;16(1):9799. [doi: 10.1038/s41467-025-64769-1] [Medline: 41198657]
34. Abbasian M, Khatibi E, Azimi I, Oniani D, Shakeri Hossein Abad Z, Thieme A, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digit Med 2024 Mar 29;7(1):82 [FREE Full text] [doi: 10.1038/s41746-024-01074-z] [Medline: 38553625]
35. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance.). EUR-Lex. 2017 Apr 5. URL: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng [accessed 2025-11-28]
36. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. J Adv Nurs 2000 Oct;32(4):1008-1015. [Medline: 11095242]
37. Chen H, Zhang Y, Wang L. A study on the quality evaluation index system of smart home care for older adults in the community based on Delphi and AHP. BMC Public Health 2023 Mar 01;23(1):411 [FREE Full text] [doi: 10.1186/s12889-023-15262-1] [Medline: 36859259]
38. Shi Y, Sun S, Deng J, Liu S, Yin T, Peng Q, et al. Establishment and application of an index system for the risk of drug shortages in China: based on Delphi method and analytic hierarchy process. Int J Health Policy Manag 2022 Dec 19:2860-2868. [doi: 10.34172/ijhpm.2022.6360]
39. Bommasani R, Liang P, Lee T. Holistic evaluation of language models. Ann N Y Acad Sci 2023 Jul;1525(1):140-146. [doi: 10.1111/nyas.15007] [Medline: 37230490]
40. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A survey on evaluation of large language models. ACM Trans Intell Syst Technol 2024 Mar 29;15(3):1-45. [doi: 10.1145/3641289]
41. Sallam M, Al-Mahzoum K, Alshuaib O, Alhajri H, Alotaibi F, Alkhurainej D, et al. Language discrepancies in the performance of generative artificial intelligence models: an examination of infectious disease queries in English and Arabic. BMC Infect Dis 2024 Aug 08;24(1):799 [FREE Full text] [doi: 10.1186/s12879-024-09725-y] [Medline: 39118057]

XSL•FO
RenderX

42. Meyer A, Soleman A, Riese J, Streichert T. Comparison of ChatGPT, Gemini, and Le Chat with physician interpretations of medical laboratory questions from an online health forum. Clin Chem Lab Med 2024 Nov 26;62(12):2425-2434 [FREE Full text] [doi: 10.1515/cclm-2024-0246] [Medline: 38804035]
43. The “Maturity assessment model for large models in the medical and health industry, part 1: health consultation” has been officially released. Smart Healthcare Innovation Competition. 2024. URL: https://tinyurl.com/c95zexwj [accessed 2025-06-15]
44. Huang C, Lee P, Chen L. Exploring consumers' negative electronic word-of-mouth of 5 military hospitals in Taiwan through SERVQUAL and Flower of Services: web scraping analysis. JMIR Form Res 2024 May 29;8:e54334 [FREE Full text] [doi: 10.2196/54334] [Medline: 38809602]
45. Bashir A, Bastola DR. Perspectives of nurses toward telehealth efficacy and quality of health care: pilot study. JMIR Med Inform 2018 May 25;6(2):e35 [FREE Full text] [doi: 10.2196/medinform.9080] [Medline: 29802089]
46. Smith AG, Brainard JC, Campbell KA. Development of an undergraduate medical education critical care content outline utilizing the Delphi method. Crit Care Med 2020 Jan;48(1):98-103. [doi: 10.1097/CCM.0000000000004086] [Medline: 31714399]
47. Nie S, Wang L. Constructing an evaluation index system for clinical nursing practice teaching quality using a Delphi method and analytic hierarchy process-based approach. BMC Med Educ 2024 Jul 19;24(1):772. [doi: 10.1186/s12909-024-05770-y] [Medline: 39030603]
48. Liu S, Li Y, Fu S, Liu X, Liu T, Fan H, et al. Establishing a multidisciplinary framework for an emergency food supply system using a modified Delphi approach. Foods 2022 Apr 06;11(7):1054. [doi: 10.3390/foods11071054] [Medline: 35407141]
49. Min Z, Bin H, Wenjie Z, Tao L, Yi M, Chunhua Z, et al. Developing an assessment tool for the healthy lifestyles of the occupational population in China: a modified Delphi-analytic hierarchy process study. Sci Rep 2024 Sep 02;14(1):20359. [doi: 10.1038/s41598-024-71324-3]
50. Liu W, Hu M, Chen W. Identifying the service capability of long-term care facilities in China: an e-Delphi study. Front Public Health 2022 Jun 29;10:884514 [FREE Full text] [doi: 10.3389/fpubh.2022.884514] [Medline: 35844860]
51. Cai W, Yao Y, Lei W, Li H, Yan S, Wu Q, et al. Construction on training course and training quality evaluation index system of chronic disease medication therapy management service (MTMs) in China: a Delphi study. PLoS One 2025;20(1):e0318446 [FREE Full text] [doi: 10.1371/journal.pone.0318446] [Medline: 39883712]
52. Chang H, Lo C, Chang H. Development of the benefit-risk assessment of complementary and alternative medicine use in people with diabetes: a Delphi-analytic hierarchy process approach. Comput Inform Nurs 2021 Apr 16;120(3):384-391. [doi: 10.1097/CIN.0000000000000749] [Medline: 33871384]
53. Ruan Y, Song S, Yin Z, Wang M, Huang N, Gu W, et al. Comprehensive evaluation of military training-induced fatigue among soldiers in China: a Delphi consensus study. Front Public Health 2022;10:1004910 [FREE Full text] [doi: 10.3389/fpubh.2022.1004910] [Medline: 36523578]
54. Gwee X, Nyunt MSZ, Kua EH, Jeste DV, Kumar R, Ng TP. Reliability and validity of a self-rated analogue scale for global measure of successful aging. Am J Geriatr Psychiatry 2014 Aug;22(8):829-837 [FREE Full text] [doi: 10.1016/j.jagp.2013.09.002] [Medline: 24119862]
55. Hsu C, Sandford B. The Delphi technique: making sense of consensus. Practical Assessment Research & Evaluation 2007;12(1):10 [FREE Full text] [doi: 10.7275/pdz9-th90]
56. Hohmann E, Brand JC, Rossi MJ, Lubowitz JH. Expert opinion is necessary: Delphi panel methodology facilitates a scientific approach to consensus. Arthroscopy 2018 Feb;34(2):349-351 [FREE Full text] [doi: 10.1016/j.arthro.2017.11.022] [Medline: 29413182]
57. Zhou W, Zheng Q, Huang M, Wang J, Gan X. Development and validation of nurse's assessment ability questionnaire in delirium subtypes: based on Delphi expert consensus. PLoS One 2024;19(1):e0297063 [FREE Full text] [doi: 10.1371/journal.pone.0297063] [Medline: 38261557]
58. Bouderhem R. Shaping the future of AI in healthcare through ethics and governance. Humanit Soc Sci Commun 2024 Mar 15;11(1):416. [doi: 10.1057/s41599-024-02894-w]
59. Menz BD, Kuderer NM, Bacchi S, Modi ND, Chin-Yee B, Hu T, et al. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. BMJ 2024 Mar 20:e078538 [FREE Full text] [doi: 10.1136/bmj-2023-078538] [Medline: 38508682]
60. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: a narrative review. Heliyon 2024 Feb 29;10(4):e26297 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e26297] [Medline: 38384518]
61. Suárez A, Jiménez J, Llorente de Pedro M, Andreu-Vázquez C, Díaz-Flores García V, Gómez Sánchez M, et al. Beyond the scalpel: assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. Computational and Structural Biotechnology Journal 2024 Dec;24:46-52. [doi: 10.1016/j.csbj.2023.11.058]
62. Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton E, et al. Applications and concerns of ChatGPT and other conversational large language models in health care: systematic review. J Med Internet Res 2024 Nov 07;26:e22769 [FREE Full text] [doi: 10.2196/22769] [Medline: 39509695]

XSL•FO
RenderX

63. Lukac S, Dayan D, Fink V, Leinert E, Hartkopf A, Veselinovic K, et al. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. Arch Gynecol Obstet 2023 Dec;308(6):1831-1844 [FREE Full text] [doi: 10.1007/s00404-023-07130-5] [Medline: 37458761]
64. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023 Jun 20;30(7):1237-1245 [FREE Full text] [doi: 10.1093/jamia/ocad072] [Medline: 37087108]
65. Choudhury A, Shamszare H. Investigating the impact of user trust on the adoption and use of ChatGPT: survey analysis. J Med Internet Res 2023 Jun 14;25:e47184 [FREE Full text] [doi: 10.2196/47184] [Medline: 37314848]
66. Nong P, Platt J. Patients' trust in health systems to use artificial intelligence. JAMA Netw Open 2025 Feb 14;8(2):e2460628. [doi: 10.1001/jamanetworkopen.2024.60628]
67. Esmaeilzadeh P, Maddah M, Mirzaei T. Using AI chatbots (e.g., CHATGPT) in seeking health-related information online: the case of a common ailment. Computers in Human Behavior: Artificial Humans 2025 Mar;3:100127 [FREE Full text] [doi: 10.1016/j.chbah.2025.100127]
68. Artificial Intelligence Act. Future of Life Institute. 2025. URL: https://artificialintelligenceact.eu/ai-act-explorer/ [accessed 2025-11-27]
69. Singapore launches Model AI Governance Framework (Gen AI) and AI Governance Playbook for Digital Forum of Small States (Digital FOSS). Infocomm Media Development Authority. 2024. URL: https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/factsheets/2024/gen-ai-and-digital-foss-ai-governance-playbook [accessed 2025-11-27]
70. Interim measures for the management of generative artificial intelligence services. The Central People's Government of the People's Republic of China. 2023. URL: https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm [accessed 2025-11-27]
71. Sun H, Wang Y, Cai H, Wang P, Jiang J, Shi C, et al. The development of a performance evaluation index system for Chinese Centers for Disease Control and Prevention: a Delphi consensus study. Glob Health Res Policy 2024 Jul 23;9(1):28 [FREE Full text] [doi: 10.1186/s41256-024-00367-w] [Medline: 39044214]

Abbreviations

AHP: analytic hierarchy process
AI: artificial intelligence
AMIE: Articulate Medical Intelligence Explorer
BLEU: Bilingual Evaluation Understudy
Ca: the basis for expert's judgment
CPMI: Chinese Patent Medicine Instructions
Cr: expert authority coefficient
CRAFT-MD: the Conversational Reasoning Assessment Framework for Testing in Medicine
Cs: expert familiarity
CV: coefficient of variation
HCAI: health conversational artificial intelligence
LLaMA: Large Language Model from Meta
LLM: large language model
Medical-Diff-VQA: Medical Dataset for Difference Visual Question Answering on Chest X-Ray Images

Edited by A Schwartz, M Balcarras; submitted 29.Aug.2025; peer-reviewed by J Haverinen, R Yin; comments to author 14.Nov.2025; revised version received 02.Dec.2025; accepted 17.Dec.2025; published 19.Jan.2026.

Please cite as:
Liao W, Li M, Ma C, Han Y, Wang D, Liu H, Wang Y, Feng Z, Wang H, Guan Y
Developing a Quality Evaluation Index System for Health Conversational Artificial Intelligence: Mixed Methods Study
J Med Internet Res 2026;28:e83188
URL: https://www.jmir.org/2026/1/e83188
doi:10.2196/83188
PMID:

©Weizhen Liao, Meng Li, Chengyu Ma, Youli Han, Dan Wang, Haopeng Liu, Yi Wang, Zijie Feng, Huichao Wang, Yiru Guan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the

XSL•FO
RenderX

Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Developing a Service Quality Index System for AI Health Care Chatbots: Mixed Methods Study

Yu Gu[1], PhD; Xinyi Wang[1], BA

School of Medical Technology, Capital Medical University, Beijing, China

Corresponding Author:
Yu Gu, PhD
School of Medical Technology
Capital Medical University
No 10 Xi Toutiao Road
Beijing, 100069
China
Phone: 86 81476234
Email: bitguyu@126.com

Abstract

Background: Artificial intelligence (AI) health care chatbots are gaining widespread adoption worldwide. It is imperative to understand the service quality of AI health care chatbots. However, there is limited guidance on how to comprehensively evaluate their service quality.

Objective: This study aimed to develop an index system based on the SERVQUAL framework for evaluating the service quality of AI health care chatbots.

Methods: An initial indicator pool was compiled through a comprehensive literature review and consultations with 4 experts. These indicators were mapped and categorized into 5 domains adapted from the SERVQUAL framework. The experts were recruited from hospital, university, and health commission settings by purposive sampling. The service quality index system was identified using a 2-round Delphi process, which included a virtual meeting between the 2 rounds. In the third round, indicator weights within each quality domain and subdomain were determined using the analytic hierarchy process.

Results: There were 26 indicators identified in the literature, based on which the 2-round Delphi process was conducted. A total of 20 experts were invited. The response rates in both rounds of Delphi and the analytic hierarchy process were 100%, and the authoritative coefficients were both >0.7. The final service quality index system for AI health care chatbots comprises 5 primary indicators and 17 secondary indicators. There were 3 (18%) indicators on assurance, 4 (24%) on reliability, 3 (18%) on human-likeness, 4 (24%) on tangibility, and 3 (18%) on responsiveness. The primary indicators, ranked from highest to lowest weight, were assurance (0.239), reliability (0.237), human-likeness (0.187), tangibility (0.170), and responsiveness (0.167).

Conclusions: This study pioneers the development of a service quality index system for AI health care chatbots adapted from the SERVQUAL framework. The results provide a validated tool for evaluating the performance of chatbots and offer valuable insights for health service managers and developers to enhance AI-driven medical consultation services.

(J Med Internet Res 2026;28:e83051) doi:10.2196/83051

KEYWORDS

artificial intelligence health care chatbot; artificial intelligence; AI; service quality; Delphi method; analytic hierarchy process; index system development; SERVQUAL

Introduction

Worldwide, artificial intelligence (AI) chatbots have been introduced into health care settings in recent years, where they are used by individuals as AI physicians for online medical consultations. A key innovation of AI health care chatbots lies in their ability to generate humanlike, natural language responses to diverse health-related queries anytime and anywhere, significantly improving access to medical guidance for broader populations [1]. Unlike earlier rule-based chatbots that relied on scripted replies, AI chatbots leverage advanced technologies, such as large language models (LLMs), to deliver personalized and context-aware interactions [2]. Moreover, the consultation service is often provided free of charge. AI health care chatbots show promise in delivering reliable medical advice without

XSL•FO
RenderX

direct involvement from human physicians, offering a scalable solution to persistent challenges within the global health system, such as limited resources, uneven distribution, high costs, and growing demand [3]. Therefore, AI health care chatbots are playing an increasingly important role in modern health care systems [4].

AI health care chatbots represent not only a new type of service provider but also an innovative medical service model [5]. As an emerging field, chatbots have attracted growing attention from both practitioners and researchers. Despite its potential benefits, concerns remain regarding its service quality [6,7]. Efforts have been made to develop quality indicators for AI health care chatbots [8-16]. Some studies have evaluated response quality within specific disease contexts, such as labor epidurals, cardiovascular health, oncology, psoriasis, chronic hepatitis, and cancer [8-11]. Others have focused on assessing information quality [12,13] or have compared the performance of AI health care chatbots with that of human physicians [14-16]. However, existing studies primarily focus on narrow aspects of quality. Furthermore, the most commonly applied metrics—response accuracy, completeness, and consistency in closed-ended clinical questions—are predominantly defined from the health care providers' perspective rather than that of users. Therefore, a comprehensive and user-centered index system for evaluating service quality of AI health care chatbots remains underdeveloped.

Among existing service quality frameworks, SERVQUAL, developed by Parasuraman et al [17], is one of the most widely recognized frameworks for evaluating medical service quality worldwide. This framework includes 5 dimensions—tangibility, reliability, responsiveness, assurance, and empathy—and is specifically designed to assess users' expectations and perceptions of service quality [18]. Applying this classical framework enables a more comprehensive and theoretically grounded evaluation of service quality of AI health care chatbots, bridging classical service quality theory with emerging AI-driven health care contexts.

The aim of this study was to identify critical indicators that reflect the service quality of AI health care chatbots and to develop a scientifically feasible index system for its evaluation. The findings were expected to contribute to better identification of shortcomings, promote continuous quality improvement, enhance user experience, and offer new insights into the systemic evaluation of service quality of AI health care chatbots.

Methods

Study Design

This study used a mixed methods approach, combining qualitative insights from expert opinions with quantitative metrics to develop and quantify a service quality index system for AI health care chatbots. The literature review and expert consultation were applied to construct an initial indicator pool. The 2-round Delphi consultation was then conducted to refine and establish the final index system. Subsequently, the analytic hierarchy process (AHP) was applied to determine the weight of each indicator. The process of index system development and weight determination is shown in Figure 1.

Figure 1. The research process. AHP: analytic hierarchy process.

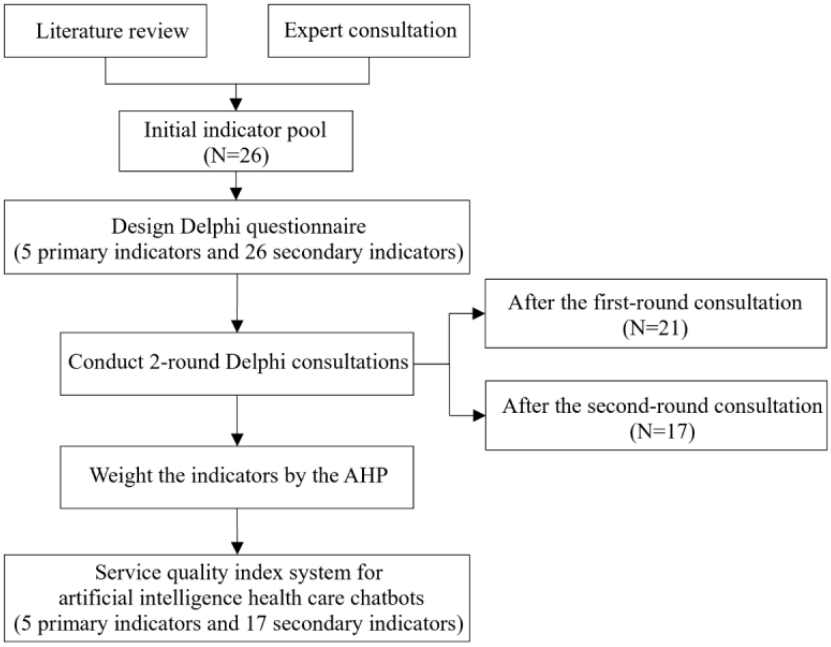

Initial Indicator Pool

The initial indicator pool was compiled based on existing literature and expert opinions. A comprehensive systematic literature search was conducted in 4 databases: PubMed, Web of Science, China National Knowledge Infrastructure, and Wanfang Data. The search strategy incorporated the following key terms: ["chatbot*" OR "chat-bot*" OR "conversational agent*" OR "conversational bot*" OR "conversational system*" OR "dialogue system*" OR ChatGPT] AND ["medic*" OR "health*" OR "disease*" OR "patient*"] AND ["quality indicator*" OR "quality evaluat*" OR "quality assess*" OR "quality measure*"]. Boolean operators (AND, OR, and NOT) were used to combine or refine search terms (Multimedia

XSL•FO
RenderX

Appendix 1). The exclusion criteria were as follows: (1) studies not focused on the evaluation of AI health care chatbots; (2) commentaries, protocols, letters, editorials, and conference abstracts; and (3) studies not published in English or Chinese.

In the second phase, this study incorporated insights from 4 interdisciplinary experts specializing in intelligent health care and medical service management. These experts conducted in-depth discussions regarding the relevance, suitability, and validity of the preliminary indicators. None of them participated in the subsequent Delphi consultation rounds. Both prior literature and expert opinions emphasized that the SERVQUAL framework provides a user-centered foundation well suited for the evaluation of AI health care chatbots and highlighted anthropomorphism as a distinctive feature influencing perceived service quality in AI-driven interactions. Specifically, users desire AI health care chatbots to exhibit kindness through humanlike attributes, such as a name, image, and voice. Beyond a friendly appearance, users expect these systems to demonstrate social intelligence, including the ability to detect user emotions and respond with genuine concern, which goes beyond simple empathy. Furthermore, users expect personalized responses tailored to individual factors. Therefore, the *humanlike* dimension was introduced as an innovative replacement for the traditional "empathy" construct to better capture the emotional and interactive capabilities unique to AI health care chatbots. Accordingly, this study established 5 first-level indicators adapted from the SERVQUAL structure: tangibility, responsiveness, assurance, human-likeness, and reliability, and 26 second-level indicators were included in the initial pool.

Expert Selection

The initial expert recruitment was conducted through recommendations from our collaborators in the field of intelligent health care. They were from renowned universities, tertiary hospitals, and provincial health sectors. To broaden the reach and ensure a diverse range of perspectives, the experts initially nominated by the researchers were then asked to suggest other qualified individuals who could contribute valuable insights to the study. The inclusion criteria for experts were as follows: (1) familiarity with research areas, such as intelligent health care, medical service management, health information management, and other related fields; (2) more than 5 years of professional experience in a relevant field; and (3) willingness to actively participate in the study and provide timely responses across multiple rounds of Delphi consultation. Finally, a total of 20 experts were recruited through purposive sampling.

Delphi Process

The Delphi method is a structured communication technique designed to systematically collect expert opinions and achieve consensus [19]. It has been widely applied and validated as a robust research methodology in health care contexts [20]. This study conducted a 2-round Delphi consultation to screen, refine, and finalize the indicators.

The Delphi consultation questionnaire consists of 2 main sections: an informed consent form and the main survey. The informed consent form outlined the study's background, objectives, methodology, privacy protection measures, and contact information. The main survey collected information from five areas: (1) experts' basic information, including age, education, and years of work experience; (2) the core consultation content, in which experts scored the importance and feasibility of each indicator using a 10-point scale (1=lowest and 10=highest); (3) the familiarity scale, rated by the expert themselves using a 5-point Likert scale (1=very unfamiliar and 5=very familiar); (4) the basis of expert judgment, evaluating the impact of theoretical analysis, practical experience, literature knowledge, and instinct on scoring (rated as high, medium, or low); and (5) blank fields, allowing experts to propose additions, deletions, or modifications to the indicators. All experts completed the informed consent process, and strict confidentiality was maintained throughout the entire process.

The Delphi process was conducted between February 2025 and June 2025. In the first round, Delphi questionnaires in Microsoft Word format were distributed to 20 experts, with a 2-week response period. Experts were asked to rate both the primary and secondary indicators and to provide comments. On the basis of the results and comments from the first round, the questionnaire was revised and redistributed to the same 20 experts for the second round. The second round followed the same rating procedure as the first and achieved consensus among the experts.

Indicator Selection

To screen the indicators, this study used 3 important statistics: the mean importance score, the full-mark rate (proportion of experts assigning the highest score), and the coefficient of variation. The inclusion criteria were as follows: (1) a mean of importance score ≥7.0, (2) a full-mark rate >20%, and (3) a coefficient of variation <0.25 [21-23]. Any indicator failing to meet all 3 criteria was subject to deletion or revision based on panel discussion and qualitative feedback.

AHP Procedure

Following the 2-round Delphi consultation, the final set of indicators was confirmed. The same panel of experts was then invited to participate in a pairwise comparison process to determine indicator weights. For each pair of indicators within the same hierarchical level, judgment matrices were constructed using a 1 to 9 ordinal scale to assess their relative importance [24]. The weight of each indicator was subsequently calculated using the percentage weighting method based on the pairwise comparison matrices, with higher weight values indicating greater perceived importance.

Data Analysis

Statistical analysis was conducted using SPSS software (version 25.0; IBM Corp). The authority coefficient (Cr) represents the authority level of experts. Cr was the arithmetic mean of the experts' judgment coefficient (Ca) and the experts' familiarity coefficient (Cs) [25]. A Cr value ≥0.7 was considered acceptable [26]. The Ca value was derived from experts' self-assessment of their own judgment criteria, as detailed in Table 1. The Cs value ranges from 1.0 (very familiar) to 0.2 (unfamiliar). The coordination of expert opinions was tested using the Kendall coefficient of concordance (Kendall *W*), with a significance level of α=.05. YAAHP software (version 11.2; MetaDecision)

XSL•FO
RenderX

was used to calculate the indicator weights and assess the consistency ratio. When the consistency ratio value was <0.10, it was considered acceptable, indicating sufficient consistency in expert judgments [27].

Table 1. The judgment basis and degree of influence.

Judgment basis	Degree of impact on experts' judgment		
	High	Medium	Low
Theoretical analysis	0.3	0.2	0.1
Practical experience	0.5	0.4	0.3
Reference literature	0.1	0.1	0.1
Expert intuition	0.1	0.1	0.1

Ethical Considerations

The study protocol was approved by the Ethics Committee of the Capital Medical University, Beijing, China (2025SY-071). Participants were informed of the study's purpose and procedure. Online informed consent was obtained from each participant. All research data were stored on a password-encrypted computer, and only the researchers had access to the data. No compensation was provided to participants.

Results

Characteristics of Experts

A total of 20 experts completed the 2-round Delphi consultation and AHP evaluation. The panel consisted of 12 (60%) male and 8 (40%) female experts, ranging in age from 31 to 60 years. Among the experts, 17 (85%) held a master's degree or higher. All experts possessed associate senior professional titles or higher. The panel included 10 (50%) experts from quality control departments of hospitals, 7 (35%) from universities, and 3 (15%) from national or regional health commissions. The detailed characteristics of these experts are summarized in Table 2.

XSL•FO
RenderX

Table 2. The characteristics of the Delphi consultation experts (N=20).

Characteristics	Experts, n (%)
Sex	
Male	12 (60)
Female	8 (40)
Age (years)	
31-40	6 (30)
41-50	11 (55)
51-60	3 (15)
Education	
Doctoral degree	10 (60)
Master's degree	7 (35)
Bachelor's degree	3 (15)
Professional title	
Senior	11 (55)
Associate senior	9 (45)
Seniority (years)	
6-10	4 (20)
11-20	9 (45)
21-30	5 (25)
>30	2 (10)
Affiliation	
Hospitals	10 (50)
Universities	7 (35)
Health commission	3 (15)
Field of expertise	
Intelligent health care	8 (40)
Medical service management	7 (35)
Health information management	5 (25)

Authority Coefficient and Degree of Coordination

The Cr values for the first and second rounds of the Delphi consultation were 0.894 (Ca=0.931; Cs=0.847) and 0.919 (Ca=0.917; Cs=0.921), respectively (Table 3). Both values exceed the accepted threshold of 0.7, indicating a high level of expert credibility and reinforcing the reliability of the consultation results.

The Kendall *W* coefficients for the 2 consultation rounds are shown in Table 3. After the second round, the coordination coefficients of the indicator increased from 0.263 to 0.339, and all associated *P* values were <.001, indicating that the experts' opinions converged and that the degree of consensus among experts was acceptable.

Table 3. Expert authority coefficients and the degree of coordination of expert opinions.

Round	Ca[a]	Cs[b]	Cr[c]	Kendall *W*	Chi-square (*df*)	*P* value
Round 1	0.931	0.847	0.894	0.263	71.1 (25)	<.001
Round 2	0.917	0.921	0.919	0.339	123.2 (20)	<.001

[a]Cs: familiarity coefficient.

[b]Ca: judgment coefficient.

[c]Cr: authority coefficient.

XSL•FO
RenderX

Review for Initial Indicator Pool

The database search and hand searches identified 117 articles, from which 48 (41.0%) duplicates were removed. After screening the titles and abstracts, 29 (24.8%) full-text records were reviewed, of which 20 (17.1%) were included in the review [28-47]. The study selection process is illustrated in Figure 2.

Figure 2. Flowchart of the included studies.

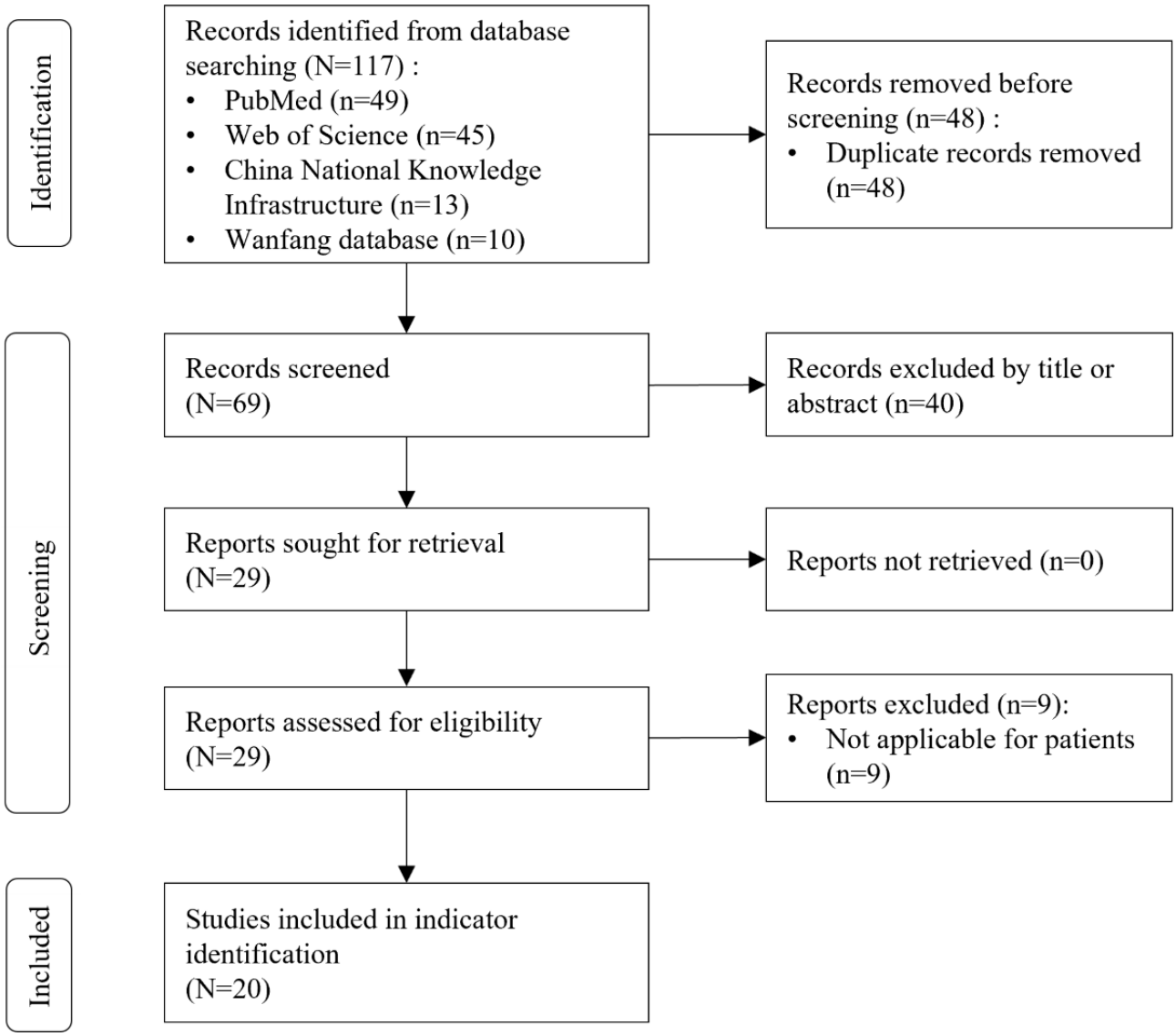

In the 20 eligible studies, 9 (45%) focused exclusively on AI chatbots designed specifically for health care service [28-36]. To develop a comprehensive initial indicator pool, we also included 11 additional studies concerning the quality of general AI chatbots capable of delivering health care–related consultations [37-47]. Of these, 6 (30%) studies aimed to develop instruments for measuring the overall service quality of AI chatbots [28,29,34,37,38,40], with 2 (10%) grounded in the SERVQUAL model [34,37]. Another 7 (35%) studies developed instruments targeting specific dimensions of quality [30-33,35,36,42]. Additionally, 7 (35%) studies treated quality as a key determinant of acceptance or satisfaction with AI chatbots and provided detailed measurement items for AI chatbot quality [39,41,43-47]. Following a systematic sorting process and discussions with 4 domain experts, we synthesized these findings into an initial pool of 26 second-level indicators for assessing the quality of AI health care chatbots, structured according to the 5 dimensions of the SERVQUAL framework.

Indicator Selection

According to the indicator selection criteria and qualitative feedback from the experts, from a total of 26 indicators, 5 (19%) secondary indicators were removed and 21 (81%) were retained in the first round. Between round 1 and round 2, a virtual Delphi meeting was held to discuss indicators for which experts provided revision suggestions and to identify novel indicators based on perceived gaps in current indicators. Experts who had completed round 1 attended the meeting. During this panel meeting, 8 (31%) secondary indicators were merged into 4 (15%) indicators. Following the discussion, 20 experts rated the new indicators. Finally, the 2-round Delphi process reached a finalized evaluation framework comprising 5 (19%) primary indicators and 17 (65%) secondary indicators, each clearly defined in Table 4.

XSL•FO
RenderX

Table 4. The service quality index system for artificial intelligence (AI) health care chatbots.

Indicator	Definition	Weights
Assurance	The ability of AI health care chatbots to provide pertinent responses	0.239
Understandable answer	The AI health care chatbots provide an answer with a logistical structure and the right amount of information to the user's query	0.082
Accurate understanding	The AI health care chatbots understand the exact meaning of the content sent by the user in text and voice	0.079
Targeted question	The AI health care chatbots ask follow-up questions with context awareness based on the user's query	0.078
Reliability	The ability of AI health care chatbots to inspire trust and confidence	0.237
Trustworthy advice	The AI health care chatbots give medical advice, such as diagnosis, medication, and examination, and detailed explanations, which are consistent across multiple inquiries	0.071
Useful service	The AI health care chatbots are useful in addressing users' uncertainties about their health concerns	0.065
Specific risk warning	The AI health care chatbots clearly indicate the limitations of its provided answers	0.050
Protected privacy	The AI health care chatbots protect the user's privacy	0.051
Human-likeness	The social cue, personality, and empathy of AI health care chatbots	0.187
Personalized response	The AI health care chatbots tailor their answers according to the user's age, sex, and medical history	0.064
Emotional attention	The AI health care chatbots detect user emotion and makes the user feel concerned	0.063
Kind characteristic	The AI health care chatbots have a kind name, image, and voice	0.060
Tangibility	The hardware or software manifestations of AI health care chatbots	0.170
Accurate recognition	The AI health care chatbots recognize and transfer the speech of the users into accurate text	0.046
Compatible operation	It is convenient for the user to obtain the service of AI health care chatbots on a mobile app, WeChat mini program, or website	0.043
Friendly layout	The layout of AI health care chatbots is clear and easy to operate	0.041
Stable service	The AI health care chatbots provide the same smooth service in any situation	0.040
Responsiveness	The response ability of AI health care chatbots	0.167
Anytime response	The AI health care chatbots are available 24 hours a day for 365 days	0.043
Prompt response	The AI health care chatbots always give timely feedback when it is needed	0.042
Coherent response	The AI health care chatbots can communicate with the user seamlessly by maintaining records within the personal account	0.042

Indicator Weights

On the basis of the AHP and percentage weighting method, the weights for all indicators were calculated (Table 4). The primary indicators, ranked from highest to lowest weight, were assurance (0.239), reliability (0.237), human-likeness (0.187), tangibility (0.170), and responsiveness (0.167). Assurance received the highest weight. For the secondary indicators, weights ranged from 0.040 to 0.082. "Understandable answer" had the highest secondary weight (0.082), followed by "Accurate understanding" (0.079), "Targeted question" (0.078), and "Trustworthy advice" (0.071).

Discussion

Principal Findings

The development of AI health care chatbots is on the rise, and their adoption is becoming increasingly vital in modern health care. Providing AI health care chatbots with high service quality is critical to facilitating their broader diffusion and addressing contemporary health care challenges. Although previous studies have attempted to evaluate the quality of AI chatbots in responding to queries related to specific diseases, a comprehensive and user-centered index system for evaluating the service quality of AI health care chatbots has remained lacking. To our knowledge, this is the first study to develop a comprehensive service quality index system for AI health care chatbots from a patient perspective, using SERVQUAL as the theoretical framework. Through a 2-round Delphi process, a finalized set of 5 primary indicators and 17 secondary indicators was derived, specifically designed to capture both the technical functionality and interactive experience unique to AI health care chatbots. Subsequently, the indicator weights were obtained using the AHP.

Among the 5 primary indicators, assurance was identified as the most important dimension, which refers to the ability of AI health care chatbots to provide pertinent responses. Unlike consultations with a clinician, AI health care chatbots lack the

XSL•FO
RenderX

capacity to perform physical examinations to support diagnosis. Therefore, it is essential for AI health care chatbots to deliver goal-oriented and unambiguous conversations, accurately understand user queries, ask follow-up questions with contextual awareness, and provide understandable answers [39,48]. Previous studies [8-10] have taken understandability, often reflected through situation-appropriate response length and information quantity, as a sole metric to measure the quality of AI health care chatbots. Consistent with this emphasis, the secondary indicator "Understandable answer" received the highest weight among secondary indicators. Although aspects such as completeness and consistency have been associated with answer readability in previous studies [11,48], they were not included in this framework. This omission stems from their highly specialized and profession-centric evaluation criteria, which may not align with patient-centered usability expectations.

The reliability dimension ranked second, following assurance. Current AI technologies remain fallible and necessitate oversight by health professionals to ensure the applicability and safety advice of AI health care chatbots [7]. For users, it is critical that AI health care chatbots provide not only trustworthy medical advice regarding diagnosis, medication, and examinations but also clear explanations that enhance transparency and facilitate informed decision-making, thereby fostering trust and promoting sustained engagement [4,28]. Both users and clinicians have underscored the importance of clearly indicating the limitations of AI-generated medical advice [3-5]. In contrast to earlier studies [9,49], this study deliberately excluded diagnostic accuracy, a frequently used metric, from this index system, as patients generally lack the specialized medical knowledge required to evaluate this aspect.

The human-likeness dimension is considered a distinctive feature of AI health care chatbots [1]. Although it is not ranked highest among the primary indicators, its associated secondary indicators, "Personalized response" and "Emotional attention," had prominent weight values. Users often perceive AI health care chatbots as an "AI doctor" and tend to evaluate it through direct comparison with human clinicians along these dimensions [3]. Enhancing humanlike attributes in interactions of AI health care chatbots remains a critical development objective. This entails increased efforts to enable AI health care chatbots to generate context-aware and individualized replies that adapt to both the conversational flow and user preferences [29].

The dimension of tangibility refers to hardware and software manifestations of AI health care chatbots. Among its secondary indicators, "Accurate recognition" was assigned the highest weight. As an information system designed to provide health guidance [50], the accuracy of recognizing and transferring user speech into precise text is the basis for correctly interpreting user queries and facilitating subsequent consultation processes [9]. The next important indicator, "Compatible operation," reflects the accessibility of the service of AI health care chatbots across diverse digital environments. A previous study has emphasized that users valued the ability to access services of AI health care chatbots through various devices, such as smartphones, tablets, or computers, which supports broader and more equitable adoption [31].

Although responsiveness is positioned as the final dimension in the framework, it constitutes a fundamental component of the service quality of AI health care chatbots [5]. This dimension is characterized by the provision of active and uninterrupted guidance 24 hours a day, real-time responses without having to wait in line, immediate accessibility from any location without the need for travel, and seamless communication across different devices [21,28]. Given that AI health care chatbots are supported by LLMs, users perceive responsiveness as their inherent capability [1].

Overall, AI health care chatbots currently represent a viable alternative to human clinicians in initial user interactions [2]. However, its performance can vary significantly depending on the underlying LLMs, knowledge bases, and health data used [31]. While this study developed a service quality index system for AI health care chatbots based on the SERVQUAL framework, most secondary indicators in this study were newly developed to reflect the unique AI context. The proposed index system offers practical value for multiple stakeholders: it enables users to better understand and assess the strengths of AI health care chatbots; supports health service managers in systematically collecting feedback and monitoring performance; and guides developers in conducting feasibility analyses, optimizing design, and implementing postlaunch evaluation. Future research will involve applying this index system in field studies with users of AI health care chatbots to validate their utility and support its ongoing refinement.

Strengths and Limitations

The strengths of this study are as follows. First, this study developed a comprehensive evaluation index system by adapting a scientific framework that incorporates both internationally validated evidence-based indicators and unique features of AI health care chatbots. Second, the mixed methods approach, combining literature review, expert consultations, a 2-round Delphi process, and AHP, ensured a rigorous and systematic development process. Third, this study provides new insights into the systemic evaluation of service quality of AI health care chatbots.

Several limitations should also be acknowledged. First, although the number of experts consulted met methodological requirements, it remained relatively limited. The panel may be subject to selection bias due to the experts' familiarity with AI health care chatbots, which could influence the selection and weighting of indicators. Therefore, the scope of expert consultation needs to be further expanded to enhance the validity of the indicators. Second, all participating experts were based in China, which may limit the generalizability of the findings to other cultural or health system contexts. Future studies should validate the proposed evaluation index across a broader range of settings of AI health care chatbots. Third, this index system has not yet been operationalized and evaluated by the users of AI health care chatbots. Further empirical research is needed to demonstrate its practical relevance and utility and to consider incorporating patient experience into the assessment process. Furthermore, although Kendall *W* was statistically significant, its value reflects only a moderate level of consensus. This implies that the findings are robust but limited in microlevel

XSL•FO
RenderX

ranking. In this study, experts from diverse professional backgrounds likely held different interpretations and assigned varying weights to the indicators, which may have led to evaluation discrepancies.

Conclusions

This study developed a comprehensive, user-centered index system for evaluating the service quality of AI health care chatbots. Through the Delphi method and the AHP, a finalized framework consisting of 5 primary dimensions and 17 secondary indicators was established. These include 4 indicators for tangibility, 3 for responsiveness, 3 for assurance, 3 for human-likeness, and 4 for reliability. This index system prioritizes user needs and experiences and can practically quantify the service quality of AI health care chatbots. The proposed index system will provide valuable support for health policymakers, service managers, and developers by enabling benchmark comparisons, facilitating quality monitoring, and guiding continuous service enhancement.

Acknowledgments

The authors attest that no generative artificial intelligence technologies were used in the generation of text, figures, or other informational content of this manuscript. The authors acknowledge all experts who participated in the Delphi process and the analytic hierarchy process.

Data Availability

The datasets used or analyzed during this study are available from the corresponding author on reasonable request.

Funding

This study was supported by the Social Science Fund of Capital Medical University, Beijing, China.

Authors' Contributions

YG conceptualized the study, designed the consulted index system, interpreted the data, and wrote the manuscript. XW collected and analyzed the data. Both authors approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 18 KB - jmir_v28i1e83051_app1.docx]

References

1. Laymouna M, Ma Y, Lessard D, Schuster T, Engler K, Lebouché B. Roles, users, benefits, and limitations of chatbots in health care: rapid review. J Med Internet Res 2024 Jul 23;26:e56930 [FREE Full text] [doi: 10.2196/56930] [Medline: 39042446]
2. Li H, Zhang R, Lee YC, Kraut RE, Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med 2023 Dec 19;6(1):236 [FREE Full text] [doi: 10.1038/s41746-023-00979-5] [Medline: 38114588]
3. Iacobucci G. The AI bot will see you now: how technology is changing the doctor-patient relationship. BMJ 2024 Mar 28;384:q711. [doi: 10.1136/bmj.q711] [Medline: 38548277]
4. Ng JY, Maduranayagam SG, Suthakar N, Li A, Lokker C, Iorio A, et al. Attitudes and perceptions of medical researchers towards the use of artificial intelligence chatbots in the scientific process: an international cross-sectional survey. Lancet Digit Health 2025 Jan;7(1):e94-102 [FREE Full text] [doi: 10.1016/S2589-7500(24)00202-4] [Medline: 39550312]
5. Mayer CJ, Mahal J, Geisel D, Geiger EJ, Staatz E, Zappel M, et al. User preferences and trust in hypothetical analog, digitalized and AI-based medical consultation scenarios: an online discrete choice survey. Comput Human Behav 2024 Dec;161:108419-108460 [FREE Full text] [doi: 10.1016/j.chb.2024.108419]
6. Armbruster J, Bussmann F, Rothhaas C, Titze N, Grützner PA, Freischmidt H. "Doctor ChatGPT, can you help me?" the patient's perspective: cross-sectional study. J Med Internet Res 2024 Oct 01;26:e58831 [FREE Full text] [doi: 10.2196/58831] [Medline: 39352738]
7. Moore I, Magnante C, Embry E, Mathis J, Mooney S, Haj-Hassan S, et al. Doctor AI? A pilot study examining responses of artificial intelligence to common questions asked by geriatric patients. Front Artif Intell 2024 Jul 25;7:1438012 [FREE Full text] [doi: 10.3389/frai.2024.1438012] [Medline: 39118788]

XSL•FO
RenderX

8. Lee D, Brown M, Hammond J, Zakowski M. Readability, quality and accuracy of generative artificial intelligence chatbots for commonly asked questions about labor epidurals: a comparison of ChatGPT and Bard. Int J Obstet Anesth 2025 Feb;61:104317. [doi: 10.1016/j.ijoa.2024.104317] [Medline: 39754839]
9. Olszewski R, Watros K, Mańczak M, Owoc J, Jeziorski K, Brzeziński J. Assessing the response quality and readability of chatbots in cardiovascular health, oncology, and psoriasis: a comparative study. Int J Med Inform 2024 Oct;190:105562 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105562] [Medline: 39059084]
10. Wang Y, Chen Y, Sheng J. Assessing ChatGPT as a medical consultation assistant for chronic Hepatitis B: cross-language study of English and Chinese. JMIR Med Inform 2024 Aug 08;12:e56426 [FREE Full text] [doi: 10.2196/56426] [Medline: 39115930]
11. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol 2023 Oct 01;9(10):1437-1440. [doi: 10.1001/jamaoncol.2023.2947] [Medline: 37615960]
12. Stapleton P, Santucci J, Cundy TP, Sathianathen N. Quality of information on Wilms tumor from artificial intelligence chatbots: what are your patients and their families reading? Urology 2025 Apr;198:130-134. [doi: 10.1016/j.urology.2025.01.054] [Medline: 39914668]
13. Owens OL, Leonard MS. Evaluating an AI chatbot "Prostate Cancer Info" for providing quality prostate cancer screening information: cross-sectional study. JMIR Cancer 2025 May 21;11:e72522 [FREE Full text] [doi: 10.2196/72522] [Medline: 40397820]
14. Anastasio MK, Peters P, Foote J, Melamed A, Modesitt SC, Musa F, et al. The doc versus the bot: a pilot study to assess the quality and accuracy of physician and chatbot responses to clinical questions in gynecologic oncology. Gynecol Oncol Rep 2024 Aug 08;55:101477 [FREE Full text] [doi: 10.1016/j.gore.2024.101477] [Medline: 39224817]
15. He W, Zhang W, Jin Y, Zhou Q, Zhang H, Xia Q. Physician versus large language model chatbot responses to web-based questions from autistic patients in Chinese: cross-sectional comparative analysis. J Med Internet Res 2024 Apr 30;26:e54706 [FREE Full text] [doi: 10.2196/54706] [Medline: 38687566]
16. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
17. Parasuraman A, Zeithaml VA, Berry LL. A conceptual model of service quality and its implications for future research. J Mark 1985 Sep 01;49(4):41-50 [FREE Full text] [doi: 10.1177/002224298504900403]
18. Karasan A, Erdogan M, Cinar M. Healthcare service quality evaluation: an integrated decision-making methodology and a case study. Socio-Econ Plan Sci 2022 Aug;82:101234 [FREE Full text] [doi: 10.1016/j.seps.2022.101234]
19. Zheng QL, Kong LN, Hu P, Liu DX. Identifying quality indicators for home care services: a modified Delphi and Analytic Hierarchy Process study. BMC Nurs 2024 Jul 19;23(1):494 [FREE Full text] [doi: 10.1186/s12912-024-02169-4] [Medline: 39026316]
20. Lemmen C, Woopen C, Stock S. Systems medicine 2030: a Delphi study on implementation in the German healthcare system. Health Policy 2021 Jan;125(1):104-114 [FREE Full text] [doi: 10.1016/j.healthpol.2020.11.010] [Medline: 33288301]
21. Xia M, Liu Q, Ma L, Wen J, Xue Y, Hu H, et al. Developing an evaluation index system for service capability of internet hospitals in China: mixed methods study. J Med Internet Res 2025 Jul 25;27:e72931 [FREE Full text] [doi: 10.2196/72931] [Medline: 40712159]
22. Jiang F, Liu T, Zhou H, Rakofsky JJ, Liu H, Liu Y, et al. Developing medical record-based, healthcare quality indicators for psychiatric hospitals in China: a modified Delphi-Analytic Hierarchy Process study. Int J Qual Health Care 2019 Dec 31;31(10):733-740. [doi: 10.1093/intqhc/mzz005] [Medline: 30753601]
23. Liu M, Yu Q, Liu Y. Developing quality indicators for cancer hospitals in China: a national modified Delphi process. BMJ Open 2024 Apr 09;14(4):e082930 [FREE Full text] [doi: 10.1136/bmjopen-2023-082930] [Medline: 38594187]
24. Saaty RW. The analytic hierarchy process—what it is and how it is used. Math Model 1987;9(3-5):161-176 [FREE Full text] [doi: 10.1016/0270-0255(87)90473-8]
25. Geng Y, Zhao L, Wang Y, Jiang Y, Meng K, Zheng D. Competency model for dentists in China: results of a Delphi study. PLoS One 2018 Mar 22;13(3):e0194411 [FREE Full text] [doi: 10.1371/journal.pone.0194411] [Medline: 29566048]
26. Sun H, Wang Y, Cai H, Wang P, Jiang J, Shi C, et al. The development of a performance evaluation index system for Chinese Centers for Disease Control and Prevention: a Delphi consensus study. Glob Health Res Policy 2024 Jul 23;9(1):28 [FREE Full text] [doi: 10.1186/s41256-024-00367-w] [Medline: 39044214]
27. Li Z, Guo R. Developing online medical service quality indicators in China from the perspective of online and offline integration: a modified Delphi-analytic hierarchy process study. Int J Qual Health Care 2023 Jun 16;35(2):mzad038 [FREE Full text] [doi: 10.1093/intqhc/mzad038] [Medline: 37279543]
28. Barletta VS, Caivano D, Colizzi L, Dimauro G, Piattini M. Clinical-chatbot AHP evaluation based on "quality in use" of ISO/IEC 25010. Int J Med Inform 2023 Feb;170:104951. [doi: 10.1016/j.ijmedinf.2022.104951] [Medline: 36525800]
29. Sobowale K, Humphrey DK. Evaluating the quality of psychotherapy conversational agents: framework development and cross-sectional study. JMIR Form Res 2025 Jul 02;9:e65605 [FREE Full text] [doi: 10.2196/65605] [Medline: 40600851]

XSL•FO
RenderX

30. Yang Y, Liu S, Lei P, Huang Z, Liu L, Tan Y. Assessing usability of intelligent guidance chatbots in Chinese hospitals: cross-sectional study. Digit Health 2024 Jun 06;10:20552076241260504 [FREE Full text] [doi: 10.1177/20552076241260504] [Medline: 38854920]
31. Shiferaw MW, Zheng T, Winter A, Mike LA, Chan LN. Assessing the accuracy and quality of artificial intelligence (AI) chatbot-generated responses in making patient-specific drug-therapy and healthcare-related decisions. BMC Med Inform Decis Mak 2024 Dec 24;24(1):404 [FREE Full text] [doi: 10.1186/s12911-024-02824-5] [Medline: 39719573]
32. Motegi M, Shino M, Kuwabara M, Takahashi H, Matsuyama T, Tada H, et al. Comparison of physician and large language model chatbot responses to online ear, nose, and throat inquiries. Sci Rep 2025 Jul 01;15(1):21346 [FREE Full text] [doi: 10.1038/s41598-025-06769-1] [Medline: 40596359]
33. Abd-Alrazaq A, Safi Z, Alajlani M, Warren J, Househ M, Denecke K. Technical metrics used to evaluate health care chatbots: scoping review. J Med Internet Res 2020 Jun 05;22(6):e18301 [FREE Full text] [doi: 10.2196/18301] [Medline: 32442157]
34. Choi J, Kim JW, Lee YS, Tae JH, Choi SY, Chang IH, et al. Availability of ChatGPT to provide medical information for patients with kidney cancer. Sci Rep 2024 Jan 17;14(1):1542 [FREE Full text] [doi: 10.1038/s41598-024-51531-8] [Medline: 38233511]
35. Chagas BA, Pagano AS, Prates RO, Praes EC, Ferreguetti K, Vaz H, et al. Evaluating user experience with a chatbot designed as a public health response to the COVID-19 pandemic in Brazil: mixed methods study. JMIR Hum Factors 2023 Apr 03;10:e43135 [FREE Full text] [doi: 10.2196/43135] [Medline: 36634267]
36. Chaudhry BM, Debi HR. User perceptions and experiences of an AI-driven conversational agent for mental health support. Mhealth 2024 Jul;10:22 [FREE Full text] [doi: 10.21037/mhealth-23-55] [Medline: 39114462]
37. Noor N, Rao Hill S, Troshani I. Developing a service quality scale for artificial intelligence service agents. Eur J Mark 2022 May 09;56(5):1301-1336 [FREE Full text] [doi: 10.1108/EJM-09-2020-0672]
38. Chen Q, Gong YM, Lu YB, Tang J. Classifying and measuring the service quality of AI chatbot in frontline service. J Bus Res 2022 Jun;145:552-568. [doi: 10.1016/j.jbusres.2022.02.088]
39. Chen S, Wang P, Wood J. Exploring the varying effects of chatbot service quality dimensions on customer intentions to switch service agents. Sci Rep 2025 Jul 02;15(1):22559 [FREE Full text] [doi: 10.1038/s41598-025-06490-z] [Medline: 40596391]
40. Radziwill NM, Benton MC. Evaluating quality of chatbots and intelligent conversational agents. arXiv. Preprint posted online April 15, 2017 [FREE Full text] [doi: 10.48550/arXiv.1704.04579]
41. Pattanshetti S. Extensive study: performance, metrics and usability of chatbot. Int Res J Eng Technol 2021 Sep;8(9):1527-1534 [FREE Full text]
42. Borsci S, Malizia A, Schmettow M, van der Velde F, Tariverdiyeva G, Balaji D, et al. The chatbot usability scale: the design and pilot of a usability scale for interaction with AI-based conversational agents. Pers Ubiquit Comput 2021 Jul 21;26(1):95-119 [FREE Full text] [doi: 10.1007/S00779-021-01582-9]
43. Pelau C, Dabija DC, Ene I. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. Comput Human Behav 2021 Sep;122:106855. [doi: 10.1016/j.chb.2021.106855]
44. Liu YL, Hu B, Yan W, Lin Z. Can chatbots satisfy me? A mixed-method comparative study of satisfaction with task-oriented chatbots in mainland China and Hong Kong. Comput Human Behav 2023 Jun;143:107716. [doi: 10.1016/j.chb.2023.107716]
45. Chopra A, Ranjani KS, Narsipur S. Service quality dimensions in AI-enabled chatbots leading to customer satisfaction: a study from South Asia. IIFT Int Bus Manag Rev 2023;1(1):26-49. [doi: 10.1177/jiift.221150355]
46. Shahzad MF, Xu S, An X, Javed I. Assessing the impact of AI-chatbot service quality on user e-brand loyalty through chatbot user trust, experience and electronic word of mouth. J Retail Consum Serv 2024 Jul;79:103867 [FREE Full text] [doi: 10.1016/j.jretconser.2024.103867]
47. Young KS, Lee SG, Hong GY. User satisfaction with the service quality of ChatGPT. Serv Bus 2024 Sep 02;18(3-4):417-431 [FREE Full text] [doi: 10.1007/S11628-024-00566-y]
48. Fanous A, Steffner K, Daneshjou R. Patient attitudes toward the AI doctor. Nat Med 2024 Nov 23;30(11):3057-3058. [doi: 10.1038/s41591-024-03272-4] [Medline: 39313596]
49. Büker M, Mercan G. Readability, accuracy and appropriateness and quality of AI chatbot responses as a patient information source on root canal retreatment: a comparative assessment. Int J Med Inform 2025 Sep;201:105948. [doi: 10.1016/j.ijmedinf.2025.105948] [Medline: 40288015]
50. Grilo A, Marques C, Corte-Real M, Carolino E, Caetano M. Assessing the quality and reliability of ChatGPT's responses to radiotherapy-related patient queries: comparative study with GPT-3.5 and GPT-4. JMIR Cancer 2025 Apr 16;11:e63677 [FREE Full text] [doi: 10.2196/63677] [Medline: 40239208]

Abbreviations

AHP: analytic hierarchy process
AI: artificial intelligence

XSL•FO
RenderX

LLM: large language model

Edited by A Stone; submitted 28.Aug.2025; peer-reviewed by MM Al-Tehewy, K Malas; comments to author 24.Oct.2025; accepted 30.Dec.2025; published 18.Feb.2026.

Please cite as:
Gu Y, Wang X
Developing a Service Quality Index System for AI Health Care Chatbots: Mixed Methods Study
J Med Internet Res 2026;28:e83051
URL: https://www.jmir.org/2026/1/e83051
doi:10.2196/83051
PMID:

©Yu Gu, Xinyi Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparison of Emotional Content in Text Responses From Physicians and AI Chatbots to Patient Health Queries: Cross-Sectional Study

Daniel T Burns[1], MS; Channing Bice[2], MS; Paul E Johnson[3], MD; Nicholas Chia[4], PhD; Timothy Robinson[1], PhD

[1]Department of Mathematics & Statistics, University of Wyoming, Laramie, WY, United States

[2]Department of Journalism & Media Communication, Colorado State University, Fort Collins, CO, United States

[3]Department of Otolaryngology- Head and Neck Surgery, University of Washington, Seattle, WA, United States

[4]Data Science and Learning, Argonne National Laboratory, Lemont, IL, United States

Corresponding Author:
Daniel T Burns, MS
Department of Mathematics & Statistics
University of Wyoming
1000 E University Ave
Laramie, WY, 82071
United States
Phone: 1 307 766 4221
Email: dburns5@uwyo.edu

Abstract

Background: Surveys show that many people are willing to use generative artificial intelligence (AI) for health questions. Prior research has largely focused on chatbot accuracy, with some studies finding that both physicians and consumers overwhelmingly prefer chatbot-generated text over physician responses.

Objective: This study aimed to characterize and compare the emotional content of responses from physicians and 2 AI chatbots (OpenAI's ChatGPT and Google's Gemini) and to assess differences in reading level and use of medical disclaimers.

Methods: A public, patient-deidentified telehealth website was used to compile 100 physician-answered questions. The same questions were posed to both chatbots between May 18 and 19, 2025. Two coders classified the emotional content of each sentence using a predefined codebook and reviewed for agreement. Emotions were ranked as primary, secondary, and tertiary by the proportion of sentences classified as each emotion per response. Multinomial logistic regression compared emotional rankings using physician responses as the reference. Word count, Flesch Reading Ease, and Flesch-Kincaid Grade Level were analyzed via ANOVA with the Tukey honestly significant difference test. Disclaimer use was compared between chatbots using a χ^2 test.

Results: Primary emotions were overwhelmingly neutral, except for one response from each chatbot in which anger was primary. For secondary emotions, the odds ratio of hope was 80.28% (95% CI 37.71%-93.76%) lower for ChatGPT, while the odds ratio of fear was 3.29 (95% CI 1.44-7.49) times higher for Gemini. For tertiary emotions, the odds ratio of compassion was 1.94 (95% CI 1.06-3.54) times higher, and the odds ratio of having no tertiary emotion was 84.33% (95% CI 64.72%-93.04%) lower for Gemini. Gemini responses averaged 889.1 (SD 305.7) words, ChatGPT 476.5 (SD 109.5), and physicians 193.5 (SD 113.6). Gemini had the lowest average Flesch Reading Ease score at 39.9 (SD 8.8), followed by ChatGPT at 45.8 (SD 12.8), while physicians had the highest at 51.9 (SD 13.6). Gemini had the highest average Flesch-Kincaid Grade Level at 11.3 (SD 1.5), followed by ChatGPT at 9.9 (SD 1.9), and physicians at 9.2 (SD 2.4). Gemini was significantly more likely to include a disclaimer than ChatGPT (χ^2_1=49.2; P<.001).

Conclusions: Chatbot responses were significantly (P<.001) longer and more difficult to read than physician responses and were more likely to contain a wider range of emotions. Qualitatively, chatbot responses were more varied in their presentation as well as in the breadth of the emotions themselves. The findings of this study could be used to inform more emotionally connected physician responses to patient message queries.

(J Med Internet Res 2026;28:e85516) doi:10.2196/85516

XSL•FO
RenderX

KEYWORDS

responses; emotional; chatbot; physicians; health; ChatGPT; artificial intelligence; AI

Introduction

Health communication extends beyond facts to include emotional reassurance, persuasion, and support, all of which are central to how people make health decisions. As more people turn to the internet for health information, generative artificial intelligence (GenAI) chatbots have rapidly emerged as influential sources for medical guidance, self-diagnosis, and emotional support. Research shows that many users, including physicians, often rate chatbot responses as accurate, empathetic, and useful, though concerns remain about misinformation, ethics, and accountability. Internet-driven health information seeking continues to grow. A report published in 2023 by the US Centers for Disease Control and Prevention found that 58.5% of adults surveyed in the second half of 2022 said they used the internet to look for health or medical information [1]. Between the time of data collection and publication of that study, GenAI chatbots began to be released for widespread public use, starting with ChatGPT (version 3.5; OpenAI) in November 2022. Subsequently, a 2023 study collected responses from 607 participants, of which 476 (78.4%) said that they were willing to use ChatGPT for self-diagnosis [2], while a 2025 study found that 21.2% (63/297) of respondents actively used chatbots for health information seeking [3].

Multiple studies have evaluated the accuracy of responses from chatbots to health-related questions across topics such as cancer, emergency medicine, general inquiries, and symptom checking. The results of these studies ranged from generally quite good (for symptom checking and general inquiries) to fair regarding the accuracy of information provided and were mixed in terms of dangerous or misinformative responses (no such responses in a cancer study and up to 35% of responses in an emergency medicine study) [4-7]. Beyond response accuracy, research has found that physicians prefer responses crafted by chatbots more than responses from other physicians 78% of the time [6]. Additionally, both patients and physicians have ranked ChatGPT responses significantly higher in empathy and usefulness than physician responses [8]. Recent studies have revealed that although health care professionals have favorable opinions regarding chatbot efficiency, they raise concerns about ethics and accountability, while the public appreciates the accessibility and emotional support provided by chatbots [9].

Health communication is a multidisciplinary field that integrates concepts from media studies, psychology, sociology, anthropology, informatics, education, and medical sciences [10,11]. It encompasses a variety of spoken, written, and computer-mediated texts and processes [12,13]. Health communication plays a crucial role in delivering health information to a variety of audiences, thereby enhancing patient and health care professional or medical institution understanding of health risks and guiding them in prevention, detection, diagnosis, and treatment [14]. Health communication can come from diverse sources, including health care professionals, family members, peer support groups, health education materials, news articles, and online resources [15,16].

Beyond factual information, individuals also seek emotion-based information to address the emotional dimensions of health concerns [17]. Feelings such as fear, anxiety, worry, and shame often accompany health-related decision-making, making emotional support, empathy, and encouragement critical components of health communication [18]. Supportive and emotionally responsive information has been shown to alleviate health-related stress, worry, and depression, while also enhancing individuals' commitment to managing health risks [14,19,20].

Studies have examined users' perceptions of chatbots, including their likability, capacity to build rapport, and ability to demonstrate empathy [13,21,22]. Recent studies have also demonstrated the persuasiveness of chatbots, including their ability to reduce belief in conspiracy theories and to change people's minds more broadly [23-25]. The increase in willingness to use the internet and GenAI for health information seeking corresponds with a rise in electronic patient portal messaging, particularly in the wake of the COVID-19 pandemic, and the increased burden this places on physicians and health care services [26-29]. Thus, the aim of this study was to investigate key differences in the emotional support and readability of health information offered by physicians and the commonly used artificial intelligence (AI) chatbots ChatGPT and Gemini (version 2.0 Flash; Google). We describe the emotional content and answer complexity contained within text responses to a variety of health questions across medical specialty topics, as posted on an online telehealth website.

Methods

Data Acquisition and Preparation

Following similar data collection methods as previously described [30-34], data were obtained from iCliniq (Orane Healthcare India Private Limited), an international telehealth website with publicly accessible question-and-answer forums where users can pay to ask verified physicians questions. These anonymized questions are categorized by condition and medical specialty by the website. For this study, questions and answers were selected from a subset of 10 medical specialties. From these 10 specialties, 10 question-answer pairs were randomly selected per specialty, resulting in a total of 100 pairs. The 10 specialties were allergy specialist, dermatology, family physician, general practitioner, internal medicine, medical oncology, neurology, pediatrics, psychiatry, and urology. Questions from this website were chosen using stratified random sampling (with medical specialties as strata) and subsequently posed to chatbots, and responses were compared with those of the responding physicians. In cases where the question elicited a dialogue (ie, back-and-forth discussion beyond the first question and response), only the initial question and answer were used. We next removed greetings and stock response openings from the questions and answers to reflect a more natural conversational tone, especially for the questions, to ensure that the chatbot responses were not unduly influenced.

XSL•FO
RenderX

For example, phrases such as “Hello doctor,” “Hi doctor,” “Dear doctor,” and “Doctor,” were removed from the patient questions, while phrases such as “Hi, Welcome” and “Hello, Welcome back” were removed from the physician responses. Responses were copied to a text editor (eg, a Microsoft Word or PDF document) for downstream classification analysis.

AI Chatbots

Two chatbots were used for comparison with physician responses: ChatGPT (GPT-4o) and Gemini 2.0 Flash. For each chatbot, each question (prompt) was asked in a fresh session, with “Memory” (chat history) option set to off so that no previous question or response impacted a subsequent interaction. Gemini was queried via its application programming interface from a custom Python script on May 18, 2025. As ChatGPT offered no free application programming interface access, questions were manually copied and pasted into the ChatGPT user interface on its website, and responses were similarly copied and pasted into a text editor. ChatGPT was queried from May 18 to May 19, 2025, as rate limits inhibited the ability to ask all 100 questions on the same day. The chatbots were given no special instructions, and no prompt engineering was applied to mimic real-world patient use [35-39]. Questions were asked under the default settings of each chatbot, and chatbots were free to respond however they saw fit based on their settings.

Response Summaries

Word count was calculated for all responses using the R package *stringi* [40]. Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL) are commonly used scores for assessing the readability of written materials [41-45]. The FRE score uses a formula containing total words, total sentences, and total syllables to calculate a score that indicates the reading difficulty of the material, with higher scores indicating easier material (eg, scores of 90-100 are considered very easy to read and understandable by an average 11-year-old, while scores of 30-50 are considered difficult to read and understandable by someone in college). The FKGL formula uses the same parameters as the FRE formula but calculates a score as a US grade level (or number of years of education needed) for judging readability. FRE and FKGL scores were calculated using Microsoft Word’s built-in summaries. According to the National Library of Medicine, nearly 9 out of 10 adults struggle with health literacy [46]. The US Department of Education reported that 54% of Americans aged 16 to 74 years read below a sixth-grade level [47]. The American Medical Association recommends that patient education materials be written at or below a sixth-grade level [48], while the National Institutes of Health recommends it to be at or below an eighth-grade level [41,49].

To get a better understanding of how these chatbots communicate about health-related topics, their use of “not a doctor” disclaimers was also assessed. In this context, a disclaimer consisted of any version of the chatbot explicitly stating that it is not a physician or medical professional, that the information it provided did not constitute medical advice, or that it was an AI and could not provide medical advice. This variable was recorded as 1 for “yes” and 0 for “no” depending on whether a response contained such a disclaimer.

Emotion Classification

A quantitative content analysis was conducted to systematically classify the emotional content embedded within physician and chatbot responses. Defined as “a research technique for the objective, systematic, and quantitative description of the manifest content of communication” [50], this method allowed us to categorize and quantify textual features to identify patterns and frequencies in the data. The codebook was developed using a robust body of literature in psychology and media communication to conceptualize and define each variable of interest (Table 1). The codebook contained 6 emotions (anger, compassion, fear, guilt, hope, and sadness) plus neutral or rational. For brevity, we summarize the codes, definitions, and associated literature in Table 1. The first and second authors (DTB and CB) served as the coders for this study. The unit of analysis (ie, the content coded as a part of our study) comprised the sentences and phrases of each response. The coders used MAXQDA (version 24.10; VERBI GmbH) to conduct the coding and generate classification counts. These counts were then used to rank the top 3 emotions expressed in each response (ie, primary, secondary, and tertiary), using MAXQDA’s Document Portrait function (refer to Multimedia Appendix 1 or “Document Portrait” entry of the MAXQDA Manual [51] for details).

XSL•FO
RenderX

Table 1. Emotion classification codebook.

Emotion	Definition	Examples	References
Anger	The response contains an emotion characterized by tension or hostility arising from frustration or the belief that an injury or injustice has been committed against oneself or an important other. It can manifest itself in behaviors designed to remove the object of anger.	• "My doctor told me that my symptoms were 'all in my head' and now I have cancer that has progressed to stage 4, when it could have been caught earlier." • "It's so frustrating that your concerns aren't being taken seriously, that's unacceptable!"	[52-54]
Compassion	The response contains an emotion characterized by feelings of empathy in response to witnessing the pain or discomfort of others. It can manifest as a desire to help alleviate pain or discomfort and to engage in helping or sharing behaviors.	• "I can help you find a nearby dermatology clinic." • "I'm really sorry to hear that you're feeling concerned about this. It's understandable to be worried, but based on what you've described, there are a few possible explanations that might be less serious than cancer. That said, it's still important to get it checked out."	[55,56]
Fear	The response contains an emotion characterized by feelings of worry that is typically triggered by detecting an imminent threat. It can manifest itself as behaviors designed to move away from or avoid the source of fear.	• "COVID-19 can be lethal." • "Applying bleach to a skin lesion can be dangerous." • "Swallowing a toothpick can be dangerous and may lead to serious health risks, though the actual risk of dying depends on a variety of factors, such as the size and shape of the toothpick, how it travels through the digestive system, and whether it causes complications."	[52-54]
Guilt	The response contains a self-conscious emotion characterized by an appraisal of having done (or thought) something that is wrong or violates internalized or socially acceptable standards of behavior. It can manifest itself as behaviors designed to undo or mitigate the wrong.	• "If you have COVID-19, don't go to work or school—stay home to protect others." • "I should have taken my child to the doctor sooner."	[52-54]
Hope	The response contains a prospect-based emotion experienced as a desire for a better future, despite current uncertainty or unfavorable circumstances. It manifests itself as behaviors designed to move toward their goals by taking action and/or planning for the future.	• "Although your lesion looks bad now, consistently applying Vaseline will help prevent infection and scarring." • "We caught the condition early, so many good treatment options are available."	[53,57]
Sadness	The response contains an emotion characterized by sorrow, loss, resignation, or the absence of a desired reward. It manifests itself through a lack of motivation or actions to change one's circumstances.	• "I'm tired of trying new depression medications. Nothing is helping." • "I can barely get out of bed I feel so hopeless."	[53,58]
Neutral or rational	The response does not contain an emotion but offers fact-based and verifiable information, definitions, or advice about the health condition that appeals to one's reason or rational thought processes.	• "Since the cough has been persisting for weeks, and it's affecting your ability to sleep, it would be a good idea to consult a health care professional to rule out more serious issues like pneumonia, bronchitis, or any other lung concerns." • "Sounds like you will be fine. You should flush the eye anytime you get a chemical or foreign body in the eye. You can also contact Poison Control 1-800-222-1222."	[54,59]

To ensure consistency, the authors discussed the conceptual and operational definitions of each emotion and reviewed sample responses. We then independently coded a practice set of 10 responses to verify the codebook's applicability. Through regular meetings and discussions, the 2 coders compared the categorization of responses, resolved discrepancies, and revised and modified the codebook. An example of a coded item and document portrait is presented in Multimedia Appendix 1. Intercoder reliability was assessed using a stratified (by medical specialty) random selection of 50 responses independently evaluated by the 2 coders, followed by 2 rounds of joint review. Reliability was calculated using Cohen κ coefficient, which takes chance agreement into account [60]. Initially, Cohen κ ranged between 0.64 and 0.81. Coefficients ≥0.8 were considered reliable [61], and following the second round of review, Cohen κ ranged between 0.81 and 0.86.

XSL•FO
RenderX

Statistical Analysis

All statistical analyses were performed using R (version 4.5.1; R Foundation for Statistical Computing) [62]. For the readability-related response summaries, ANOVAs followed by Tukey's honest significant difference test were used for pairwise comparisons. "Not a doctor" disclaimer use between the 2 chatbots was compared by a χ^2 test. Multinomial logistic regression was used to compare emotion classification results, with physicians set as the reference group.

For the secondary emotion analysis, 39 responses were removed (13 from each responder) because, for each of these 13 questions, the physician's responses contained only one (primary) emotion, necessitating "none" as the classification for the secondary and tertiary emotions. With the exception of one Gemini response, the chatbots did not have any "none" secondary classifications, and thus, coefficient estimation was highly biased and unreliable. Similarly, 9 responses (3 from each responder) were removed from the tertiary emotion analysis, as Gemini was the only responder to have anger classified as a tertiary emotion. Statistical significance was set at $P<.05$ for all tests.

Ethical Considerations

This study analyzed publicly available data from a freely accessible telehealth website. No personal or identifiable patient or physician data were collected or used. According to the US Department of Health and Human Services, 45 Code of Federal Regulations 46.102(e)(1) [63], this study does not meet the definition of human participants research and therefore is exempt from institutional review board review.

Results

Response Summaries

Figure 1A shows the distribution of word counts for each responder. All 3 responders' word counts were significantly different from one another ($F_{2,297}=310.1$; $P<.001$). Gemini had the highest word count, with an average of 889.1 (SD 305.7) words per response, followed by ChatGPT with an average of 476.5 (SD 109.5) words per response, and physicians with the lowest average at 193.5 (SD 113.6) words per response.

All 3 responders' FRE scores were significantly different from one another ($F_{2,297}=27.5$; $P<.001$). As shown in Figure 1B, Gemini had the lowest average FRE score at 39.9 (SD 8.8), followed by ChatGPT at 45.8 (SD 12.8), while physicians had the highest average score at 51.9 (SD 13.6). All 3 responders' FKGL scores were also significantly different from one another ($F_{2,297}=25.7$; $P<.001$). Figure 1C shows that Gemini had the highest average FGKL score of 11.3 (SD 1.5), followed by ChatGPT with an average score of 9.9 (SD 1.9), and physicians again having the easiest readability of the 3 groups, with an average FGKL score of 9.2 (SD 2.4).

As shown in Figure 1D, disclaimer use differed significantly between the 2 chatbots ($\chi^2_1=49.22$; $P<.001$). Gemini made such a statement 58% (58/100; 95% CI 47.71%-67.67%) of the time compared with ChatGPT's 10% (10/100; 95% CI 5.16%-18.04%).

XSL•FO
RenderX

Figure 1. Comparison of response length, readability, and disclaimer use among physicians, ChatGPT, and Gemini. (A) Boxplots, with accompanying density plots, of the word count distributions for each responder. The bold black line inside boxes represents the median, and black circles denote outliers. (B) Mean Flesch Reading Ease (FRE) scores per responder. Higher scores indicate greater readability. (C) Mean Flesch-Kincaid Grade Level (FKGL) scores per responder. Higher scores indicate a higher educational level required for reading comprehension. (D) Proportion of responses that contained disclaimer use for each chatbot. Differences among responders in A-C were calculated via ANOVA followed by the Tukey honestly significant difference test. Differences among responders in D were calculated via the chi-square test. Error bars in B-D represent 95% CIs. *P=.03, **P=.001, ***P<.001.

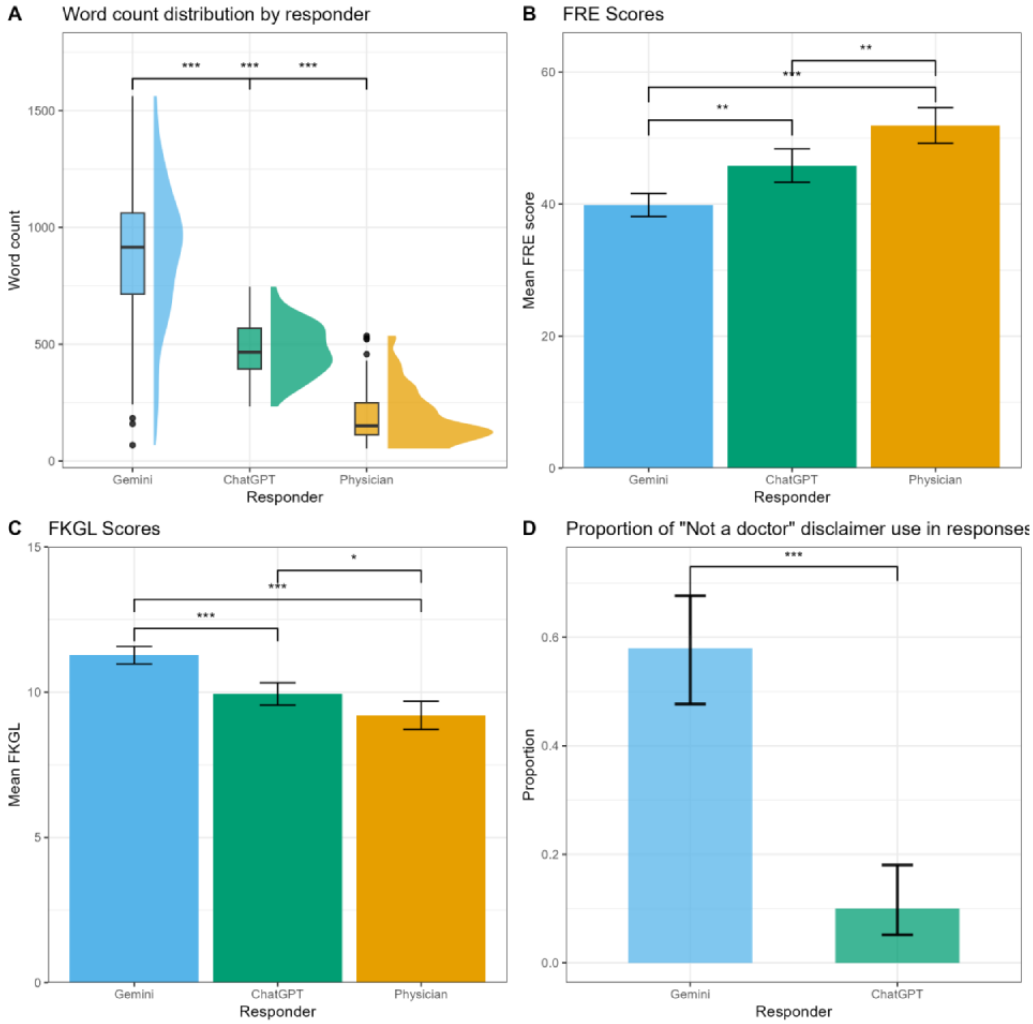

Emotion Classification

Neutral or rational was the most common emotion expressed across all 3 responders. It was the primary emotion for all 100 (100%) of the physician responses and for 99 (99%) out of 100 for each chatbot. In the one response per chatbot in which neutral was not the primary emotion, fear was the most prominent emotion. In both instances, the chatbots directed the question asker to seek immediate medical attention, as they deemed the situations to be very concerning.

All comparisons for secondary and tertiary emotions were made relative to physician responses (ie, the physician group was set as the reference and the chatbots were compared with it). For secondary emotions (Tables 2 and 3), the odds ratio of a hope response was 80.28% (95% CI 37.71%-93.76%; P=.006) lower for ChatGPT. The odds ratio of a fear response was 3.29 (95% CI 1.44-7.49; P=.005) times higher for Gemini. Physician responses did not have a secondary emotion 13% (13/00) of the time (ie, the entire response was classified as containing only one emotion), while Gemini had such responses 1% (1/100) of the time, and ChatGPT always had at least 2 emotions classified in its responses. Physicians never had anger classified as a secondary emotion, while Gemini and ChatGPT each did so once.

For tertiary emotions (Tables 4 and 5), the odds ratio of a compassion response was 1.94 (95% CI 1.06-3.54; P=.03) times higher, and the odds ratio of having no tertiary emotion classified (none) was 84.33% (95% CI 64.72%-93.04%; P<.001) lower for Gemini. The odds ratio of having no tertiary emotion classified (none) was 53% (95% CI 3.42%-78.51%; P=.06) lower for ChatGPT. Gemini was the only responder to have anger classified as a tertiary emotion.

XSL•FO
RenderX

Table 2. Secondary emotions by responder (N=100)[a].

Emotion	Physician, n (%)	ChatGPT-4o, n (%)	Gemini Flash 2.0, n (%)
Anger	0 (0)	1 (1)	1 (1)
Compassion	57 (57)	80 (80)	51 (51)
Fear	10 (10)	11 (11)	33 (33)
Guilt	3 (3)	2 (2)	3 (3)
Hope	17 (17)	5 (5)	10 (10)
Neutral	0 (0)	1 (1)	1 (1)
None	13 (13)	0 (0)	1 (1)
Sadness	0 (0)	0 (0)	0 (0)

[a]Responses from each responder containing the labeled emotions at the secondary rank. Each responder was asked the same 100 questions.

Table 3. Multinomial logistic regression results for secondary emotions[a].

Responder and emotions	Odds ratio (95% CI)	SE	Wald Z	*P* value
ChatGPT				
Compassion	1.268 (0.893-1.800)	0.179	1.329	.18
Fear	0.710 (0.270-1.866)	0.493	−0.695	.49
Guilt	0.263 (0.027-2.597)	1.168	−1.143	.25
Hope	0.197 (0.062-0.623)	0.587	−2.766	.006
Gemini				
Compassion	0.821 (0.556-1.213)	0.199	−0.989	.32
Fear	3.287 (1.442-7.491)	0.420	2.831	.005
Guilt	1.217 (0.234-6.321)	0.840	0.234	.81
Hope	0.685 (0.277-1.693)	0.462	−0.820	.41

[a]Transformed (exponentiated) results of multinomial logistic regression comparing the secondary emotion rankings of the 2 chatbots to physicians.

Table 4. Tertiary emotions by responder (N=100)[a].

Emotion	Physician, n (%)	ChatGPT (GPT-4o), n (%)	Gemini Flash 2.0, n (%)
Anger	0 (0)	0 (0)	3 (3)
Compassion	17 (17)	20 (20)	31 (31)
Fear	10 (10)	18 (18)	12 (12)
Guilt	2 (2)	6 (6)	5 (5)
Hope	13 (13)	22 (22)	32 (32)
Neutral	0 (0)	0 (0)	0 (0)
None	58 (58)	34 (34)	17 (17)
Sadness	0 (0)	0 (0)	0 (0)

[a]Responses from each responder containing the labeled emotions at the tertiary rank. Each responder was asked the same 100 questions.

XSL•FO
RenderX

Table 5. Multinomial logistic regression results for tertiary emotions[a].

Responder and emotion	Odds ratio (95% CI)	SE	Wald Z	*P* value
ChatGPT				
Compassion	1.250 (0.648-2.412)	0.335	0.665	.51
Fear	1.280 (0.458-3.577)	0.524	0.471	.64
Guilt	2.400 (0.425-13.589)	0.883	0.992	.32
Hope	1.354 (0.524-3.500)	0.485	0.625	.53
None	0.471 (0.215-1.034)	0.401	−1.876	.06
Gemini				
Compassion	1.937 (1.060-3.542)	0.308	2.149	.03
Fear	0.619 (0.220-1.741)	0.527	−0.908	.36
Guilt	1.290 (0.225-7.405)	0.891	0.286	.78
Hope	1.270 (0.525-3.072)	0.450	0.531	.60
None	0.157 (0.070-0.353)	0.414	−4.477	<.001

[a]Transformed (exponentiated) results of multinomial logistic regression comparing the tertiary emotion rankings of the 2 chatbots to physicians.

Discussion

Principal Findings

Overall, the AI chatbots provided longer and more detailed responses than physicians. Whether the additional detail provided in chatbot responses was appropriate or correct was not investigated in this study. While initial physician responses were often a series of questions, presumably akin to a history-taking for gathering more information for a proper diagnosis or suggestion, chatbots often presented lists of possible explanations that were ranked from most to least likely. The chatbots also often asked follow-up or clarification questions, but rarely at the expense of providing answers; chatbot responses were combinations of detailed lists and explanations, with occasional clarifying questions interspersed. This additional feature provided by the chatbots could also help explain why the responses were calculated to be more difficult to read. Although all responders provided answers well above the national average reading grade level [47], the chatbots were more difficult to read and understand, which may negate some of the usefulness of the added information, as it could be overly complex for a typical patient to understand and therefore properly act on.

Similarly, the very infrequent use of disclaimers, particularly by ChatGPT, explicitly stating that chatbots are not medical professionals or that the information provided does not replace consultation with a qualified clinician may place users at risk. If the information is presented in a way that is difficult to understand, users could find it challenging to discern useful information from irrelevant or potentially harmful information. This problem is further compounded by chatbots' tendency to provide responses in a manner that conveys strong certainty and expertise in the content [64,65].

The bulk of the responses from all 3 responders were neutral in emotional content, as the vast majority consisted of factual explanations and definitions. Physicians were more likely to provide a response that contained only one emotion (neutral), perhaps as a consequence of their generally much shorter responses, which often consisted of follow-up questions. In contrast, the chatbots had a relatively formulaic response pattern, wherein they opened with an emotional statement (usually compassion or fear) acknowledging the situation of the question asker. Similarly, the chatbots ended their responses with summaries of the overall response and further emotional appeals (usually compassion, fear, or hope).

Although compassionate content proportion did not differ significantly between the chatbots and physicians in a quantitative sense, the quality of the emotions presented was noticeably different. For example, a compassionate statement from physicians often looked like "I understand your concern," "I hope this helps," or "I read your query and can understand your concern," while compassionate statements from the chatbots included "Let's break down the problems and discuss potential solutions," "It's completely understandable that you're feeling lost and frustrated," "I'm really sorry you're going through [this situation]. Here's a structured overview to help guide you through managing your condition and exploring safer, longer-term strategies," and "You're absolutely right to be concerned, and it's great that you're actively looking into this. Let's go over the key points in your situation and steps you can take." Previous research has shown that patients, physicians, and third-party evaluators consistently perceive AI to be more empathetic and compassionate than physicians [8,66-68]. Our results lend support to these findings and help to show that this difference comes from the quality and content of the written responses rather than from the frequency or proportion of an emotion being conveyed. Similar differences in quality of emotion in the responses were present for other emotions as well. These increases in empathy, sympathy, and emotional support may also play a role in why people consistently prefer responses from chatbots over those from physicians [69].

However, further research is needed to explore how the emotional tone of chatbot responses influences individuals'

XSL•FO
RenderX

health information–seeking behaviors and health information–processing behaviors, which in turn affect their engagement with health-related actions. These results, coupled with the previously mentioned patient and physician preferences for chatbot responses, warrant an investigation into possible reasons for these effects. As good communication skills are increasingly viewed as an important aspect of health care training and practice [70-73], these results can be used to inform physicians' use of chatbots as a tool to better help connect with patients in a more empathetic manner. Future research should investigate optimizing prompting strategies for chatbot-assisted message drafting to determine patient perceptions, as well as quality and safety outcomes.

Limitations

There are several limitations to this study. Importantly, a key limitation is the inability to exactly replicate the results. This limitation is due to the ever-changing nature of the AI models used for these chatbots; the models get updated frequently, often without any detail regarding when or what changes were made. Often, these updates fundamentally change the way that chatbots engage with and respond to users. The behavior of the 2 chatbots used in this study, chosen for their popularity and free access, may not be representative of all chatbot services that are currently available.

Another limitation is potential cultural norm disparities. The authors are all from the United States, where both chatbots were accessed and created or hosted by US companies. This factor could cause unintended bias in emotion classification, as the expectations and definitions in one culture may not be the same as those in another. Furthermore, because the website is publicly accessible, physicians responding to questions may have access to patient information that is not visible on the question page and therefore not captured in the dataset. This access feature could complicate the interpretations of this study, as it is possible that physicians gave short, emotionally neutral answers because they had more information than was provided in just the question alone, which the researchers and therefore the chatbots would be unaware of.

Another potential limitation is that the patient questions used were chosen randomly; however, the authors selected the medical specialty topic areas containing the questions. Although the intention was to ensure an array of question types across various topics, it is possible that this approach inadvertently biased results if the selected questions were too similar. Additionally, for the sake of simplicity and feasibility, our focus was limited to 7 emotions, classified by 2 of the authors. Although these were selected from health communication literature and deemed useful and informative for the purposes of this study, it is possible that different emotions, or combinations of emotions, could yield different conclusions as well. For similar reasons, we also limited the selected AI chatbots to 2 free options, which left out popular alternatives, such as Anthropic's Claude, Microsoft's Copilot, and High-Flyer's DeepSeek, thus limiting the scope of this study and its conclusions.

Conclusions

This study demonstrates key differences in the emotional support and readability of health information offered by physicians and AI chatbots (ChatGPT and Gemini). It also extends the literature examining the usefulness of AI-generated drafts by demonstrating chatbots' ability to craft emotionally connective responses, potentially enhancing emotional support for patients, while mitigating physician burnout and compassion fatigue [74]. Compared to physicians, the AI chatbots' responses were much longer and more difficult to read. ChatGPT rarely included a disclaimer in its responses indicating that it is not a medical professional (only 10% of responses), while Gemini included one more often (60% of responses).

Although few statistically significant differences existed between the emotional content of physicians and chatbots, there were substantial qualitative differences in the depth and breadth of emotionality presented, especially for compassion (empathy). This finding could be due, in part, to the preferential training of chatbots to be more personable and agreeable, coupled with the fact that they do not suffer from burnout or fatigue.

An often overlooked aspect of AI-generated responses is the variability or uncertainty that is inherent in the models. All generative AI responses are probabilistic responses, which implies uncertainty can be investigated. However, approaches for assessing that uncertainty remain an unanswered research question; although beyond the scope of this study, this represents an important direction for future research.

Funding

This study was supported by the National Institute of General Medical Sciences of the National Institutes of Health (grant P20GM103432). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This study was also supported by funding from the Wyoming Department of Health.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: DTB, CB, NC, TR
Data curation: DTB
Formal analysis: DTB, CB

XSL•FO
RenderX

Funding acquisition: TR
Investigation: DTB, CB
Methodology: DTB, CB, TR
Project administration: DTB, TR
Resources: DTB, CB, TR
Software: DTB
Supervision: TR
Validation: DTB, CB, TR
Visualization: DTB
Writing—original draft: DTB, CB, TR
Writing—review and editing: DTB, CB, NC, TR, PEJ

Conflicts of Interest

None declared.

Multimedia Appendix 1
Text emotion classification and ranking. (A) Example physician response with sentences highlighted for the coded emotion (gray=neutral or rational, purple=compassion, and orange=fear). (B) Document Portrait classification summary of the coded response in (A), showing the proportion of each emotion in descending order from left to right. In this example response, neutral was coded the most, followed by compassion, then fear. Classifications and document portrait summaries were carried out in MAXQDA software.
[PNG File , 896 KB - jmir_v28i1e85516_app1.png]

References

1. Wang X, Cohen RA. Health information technology use among adults: United States, July–December 2022. National Center for Health Statistics. 2023. URL: https://www.cdc.gov/nchs/products/databriefs/db482.htm [accessed 2025-06-28]
2. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023 May 17;10:e47564 [FREE Full text] [doi: 10.2196/47564] [Medline: 37195756]
3. Yun HS, Bickmore T. Online health information-seeking in the era of large language models: cross-sectional web-based survey study. J Med Internet Res 2025 Mar 31;27:e68560 [FREE Full text] [doi: 10.2196/68560] [Medline: 40163112]
4. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol 2023 Oct 01;9(10):1437-1440. [doi: 10.1001/jamaoncol.2023.2947] [Medline: 37615960]
5. Yau JY, Saadat S, Hsu E, Murphy LS, Roh JS, Suchard J, et al. Accuracy of prospective assessments of 4 large language model chatbot responses to patient questions about emergency care: experimental comparative study. J Med Internet Res 2024 Nov 04;26:e60291 [FREE Full text] [doi: 10.2196/60291] [Medline: 39496149]
6. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
7. Chen A, Chen DO, Tian L. Benchmarking the symptom-checking capabilities of ChatGPT for a broad range of diseases. J Am Med Inform Assoc 2024 Sep 01;31(9):2084-2088. [doi: 10.1093/jamia/ocad245] [Medline: 38109889]
8. Armbruster J, Bussmann F, Rothhaas C, Titze N, Grützner PA, Freischmidt H. "Doctor ChatGPT, can you help me?" The patient's perspective: cross-sectional study. J Med Internet Res 2024 Oct 01;26:e58831 [FREE Full text] [doi: 10.2196/58831] [Medline: 39352738]
9. Lu L, Zhu Y, Yang J, Yang Y, Ye J, Ai S, et al. Healthcare professionals and the public sentiment analysis of ChatGPT in clinical practice. Sci Rep 2025 Jan 07;15(1):1223 [FREE Full text] [doi: 10.1038/s41598-024-84512-y] [Medline: 39774168]
10. Schiavo R. Health Communication: From Theory to Practice. Hoboken, NJ: John Wiley & Sons; 2007.
11. Werder O. Transformational Health Communication: A New Perspective on Healthcare and Prevention. Cham, Switzerland: Springer; 2024.
12. Harvey K, Koteyko N. Exploring Health Communication: Language in Action. Milton Park, UK: Routledge; 2013.
13. Park E. I trust you, but let me talk to AI: the role of the chat agents, empathy, and health issues in misinformation guidance. Int J Strat Commun 2025 Feb 20;19(2):231-260. [doi: 10.1080/1553118X.2025.2462087]
14. Kreps GL. Addressing resistance to adopting relevant health promotion recommendations with strategic health communication. Inf Serv Use 2023 May 01;43(2):131-142. [doi: 10.3233/ISU-230187]
15. Kreps GL. Online information and communication systems to enhance health outcomes through communication convergence. Hum Commun Res 2017 Jun 30;43(4):518-530. [doi: 10.1111/hcre.12117]
16. Link E, Baumann E, Kreps GL, Czerwinski F, Rosset M, Suhr R. Expanding the health information national trends survey research program internationally to examine global health communication trends: comparing health information seeking

XSL•FO
RenderX

behaviors in the U.S. and Germany. J Health Commun 2022 Aug 03;27(8):545-554 [FREE Full text] [doi: 10.1080/10810730.2022.2134522] [Medline: 36250315]
17. Yang ZJ, Kahlor L. What, me worry? The role of affect in information seeking and avoidance. Sci Commun 2012 May 24;35(2):189-212. [doi: 10.1177/1075547012441873]
18. Kreps GL, Canzona MR. The role of communication and information in symptom management. In: Wittenberg E, Ferrell BR, Goldsmith J, Smith T, Ragan SL, Glajchen M, et al, editors. Textbook of Palliative Care Communication. Oxford, UK: Oxford University Press; 2015:119-126.
19. Fisher HM, Winger JG, Miller SN, Wright AN, Plumb Vilardaga JC, Majestic C, et al. Relationship between social support, physical symptoms, and depression in women with breast cancer and pain. Support Care Cancer 2021 Sep 15;29(9):5513-5521 [FREE Full text] [doi: 10.1007/s00520-021-06136-6] [Medline: 33723675]
20. Guan M, Han JY, Shah DV, Gustafson DH. Exploring the role of social support in promoting patient participation in health care among women with breast cancer. Health Commun 2021 Nov;36(13):1581-1589 [FREE Full text] [doi: 10.1080/10410236.2020.1773704] [Medline: 32500731]
21. Pelau C, Dabija DC, Ene I. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. Comput Human Behav 2021 Sep;122:106855. [doi: 10.1016/j.chb.2021.106855]
22. Zhang J, Oh YJ, Lange P, Yu Z, Fukuoka Y. Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet: viewpoint. J Med Internet Res 2020 Sep 30;22(9):e22845 [FREE Full text] [doi: 10.2196/22845] [Medline: 32996892]
23. Costello TH, Pennycook G, Rand DG. Durably reducing conspiracy beliefs through dialogues with AI. Science 2024 Sep 13;385(6714):eadq1814. [doi: 10.1126/science.adq1814] [Medline: 39264999]
24. Salvi F, Horta Ribeiro M, Gallotti R, West R. On the conversational persuasiveness of GPT-4. Nat Hum Behav 2025 Aug 19;9(8):1645-1653. [doi: 10.1038/s41562-025-02194-6] [Medline: 40389594]
25. Schoenegger P, Salvi F, Liu J, Nan X, Debnath R, Fasolo B, et al. Large language models are more persuasive than incentivized human persuaders. ArXiv Preprint posted online on May 14, 2025 [FREE Full text] [doi: 10.48550/arXiv.2505.09662]
26. Ferguson K, Fraser M, Tuna M, Bruntz C, Dahrouge S. The impact of an electronic portal on patient encounters in primary care: interrupted time-series analysis. JMIR Med Inform 2023 Feb 06;11:e43567 [FREE Full text] [doi: 10.2196/43567] [Medline: 36745495]
27. Apathy NC, Hicks K, Bocknek L, Zabala G, Adams K, Gomes KM, et al. Inbox message prioritization and management approaches in primary care. JAMIA Open 2024 Dec;7(4):ooae135. [doi: 10.1093/jamiaopen/ooae135] [Medline: 39530053]
28. Holmgren AJ, Downing NL, Tang M, Sharp C, Longhurst C, Huckman RS. Assessing the impact of the COVID-19 pandemic on clinician ambulatory electronic health record use. J Am Med Inform Assoc 2022 Jan 29;29(3):453-460 [FREE Full text] [doi: 10.1093/jamia/ocab268] [Medline: 34888680]
29. North F, Luhman KE, Mallmann EA, Mallmann TJ, Tulledge-Scheitel SM, North EJ, et al. A retrospective analysis of provider-to-patient secure messages: how much are they increasing, who is doing the work, and is the work happening after hours? JMIR Med Inform 2020 Jul 08;8(7):e16521 [FREE Full text] [doi: 10.2196/16521] [Medline: 32673238]
30. V JP, S AA, Sundaram G. Toward explainable cross-lingual adaptive NAS for enhanced Tamil medical text summarization. IEEE J Biomed Health Inform 2025 Oct 16:1-18. [doi: 10.1109/jbhi.2025.3618959]
31. Gatto J, Seegmiller P, Johnston G, Preum SM. Identifying the perceived severity of patient-generated telemedical queries regarding COVID: developing and evaluating a transfer learning-based solution. JMIR Med Inform 2022 Sep 02;10(9):e37770 [FREE Full text] [doi: 10.2196/37770] [Medline: 35981230]
32. Yang W, Zeng G, Tan B, Ju Z, Chakravorty S, He X, et al. On the generation of medical dialogues for COVID-19. ArXiv Preprint posted online on May 15, 2020 [FREE Full text] [doi: 10.1101/2020.05.08.20095810]
33. Haghighi T, Gholami S, Sokol JT, Kishnani E, Ahsaniyan A, Rahmanian H, et al. EYE-Llama, an in-domain large language model for ophthalmology. iScience 2025 Jul 18;28(7):112984 [FREE Full text] [doi: 10.1016/j.isci.2025.112984] [Medline: 40697409]
34. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. Cureus 2023 Jun;15(6):e40895 [FREE Full text] [doi: 10.7759/cureus.40895] [Medline: 37492832]
35. Knoth N, Tolzin A, Janson A, Leimeister JM. AI literacy and its implications for prompt engineering strategies. Comput Educ Artif Intell 2024 Jun;6:100225. [doi: 10.1016/j.caeai.2024.100225]
36. Xue H, Oh YJ, Zhou X, Zhang X, Oxley B. User prompting strategies and ChatGPT contextual adaptation shape conversational information-seeking experiences. ArXiv Preprint posted online on September 29, 2025. [doi: 10.48550/arXiv.2509.25513]
37. Anam RK. Prompt engineering and the effectiveness of large language models in enhancing human productivity. ArXiv. Preprint posted online on April 04, 2025 2026 [FREE Full text] [doi: 10.31219/osf.io/ad9y5_v1]
38. Sawalha G, Taj I, Shoufan A. Analyzing student prompts and their effect on ChatGPT's performance. Cogent Educ 2024 Sep 09;11(1). [doi: 10.1080/2331186x.2024.2397200]

XSL•FO
RenderX

39. Zamfirescu-Pereira JD, Wong RY, Hartmann B, Yang Q. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023 Presented at: CHI '23; April 23-28, 2023; Hamburg, Germany URL: https://dl.acm.org/doi/10.1145/3544548.3581388 [doi: 10.1145/3544548.3581388]
40. Gagolewski M. stringi: fast and portable character string processing in R. J Stat Soft 2022;103(2):1-59 [FREE Full text] [doi: 10.18637/jss.v103.i02]
41. Eltorai AE, Ghanian S, Adams CAJ, Born CT, Daniels AH. Readability of patient education materials on the American Association for Surgery of Trauma website. Arch Trauma Res 2014 Apr 30;3(2):e18161 [FREE Full text] [doi: 10.5812/atr.18161] [Medline: 25147778]
42. Ley P, Florio T. The use of readability formulas in health care. Psychol Health Med 1996;1(1):7-28. [doi: 10.1080/13548509608400003]
43. Albright J, de Guzman C, Acebo P, Paiva D, Faulkner M, Swanson J. Readability of patient education materials: implications for clinical practice. Appl Nurs Res 1996 Aug;9(3):139-143. [doi: 10.1016/s0897-1897(96)80254-0] [Medline: 8771859]
44. Badarudeen S, Sabharwal S. Readability of patient education materials from the American Academy of Orthopaedic Surgeons and Pediatric Orthopaedic Society of North America web sites. J Bone Joint Surg Am 2008 Jan;90(1):199-204. [doi: 10.2106/jbjs.g.00347]
45. Sabharwal S, Badarudeen S, Unes Kunju S. Readability of online patient education materials from the AAOS website. Clin Orthop Relat Res 2008 May;466(5):1245-1250 [FREE Full text] [doi: 10.1007/s11999-008-0193-8] [Medline: 18324452]
46. An introduction to health literacy. National Institutes of Health National Library of Medicine. URL: https://www.nnlm.gov/guides/intro-health-literacy [accessed 2025-08-05]
47. Mamedova S, Pawlowski E. Data point: adult literacy in the United States. United States Department of Education. 2019. URL: https://nces.ed.gov/pubs2019/2019179.pdf [accessed 2025-08-05]
48. Weiss BD. Health literacy: a manual for clinicians. American Medical Association. 2003. URL: http://lib.ncfh.org/pdfs/6617.pdf [accessed 2025-09-15]
49. Rooney MK, Santiago G, Perni S, Horowitz DP, McCall AR, Einstein AJ, et al. Readability of patient education materials from high-impact medical journals: a 20-year analysis. J Patient Exp 2021 Mar 03;8:2374373521998847 [FREE Full text] [doi: 10.1177/2374373521998847] [Medline: 34179407]
50. Berelson B. Content Analysis in Communication Research. New York, NY: Free Press; 1952.
51. MAXQDA manual. MAXQDA. URL: https://www.maxqda.com/help/welcome [accessed 2025-12-15]
52. Kraak VI, Holz A, Woods CL, Whitlow AR, Leary N. A content analysis of persuasive appeals used in media campaigns to encourage and discourage sugary beverages and water in the United States. Int J Environ Res Public Health 2023 Jul 13;20(14):6359 [FREE Full text] [doi: 10.3390/ijerph20146359] [Medline: 37510591]
53. Myrick JG. The Role of Emotions in Preventative Health Communication. London, UK: Bloomsbury Academic; 2015.
54. Foerster TA, Chalgren L, Abrams K, Bice C, Singer M. Informing or influencing? A content analysis of wildlife feeding deterrence messages in U.S. national park communications. J Interpret Res 2025 Nov 05;30(2):183-203. [doi: 10.1177/10925872251388948]
55. Nan X, Iles IA, Yang B, Ma Z. Public health messaging during the COVID-19 pandemic and beyond: lessons from communication science. Health Commun 2022 Jan;37(1):1-19. [doi: 10.1080/10410236.2021.1994910] [Medline: 34724838]
56. Goetz JL, Keltner D, Simon-Thomas E. Compassion: an evolutionary analysis and empirical review. Psychol Bull 2010 May;136(3):351-374 [FREE Full text] [doi: 10.1037/a0018807] [Medline: 20438142]
57. Lazarus RS. Hope: an emotion and a vital coping resource against despair. Soc Res 1999;66(2):653-678 [FREE Full text]
58. Lazarus RS. Emotion and Adaptation. Oxford, UK: Oxford University Press; 1994.
59. Casais B, Pereira AC. The prevalence of emotional and rational tone in social advertising appeals. RAUSP Manag J 2021 May 10;56(3):282-294. [doi: 10.1108/rausp-08-2020-0187]
60. Krippendorff K. Agreement and information in the reliability of coding. Commun Methods Meas 2011 Jun 14;5(2):93-112. [doi: 10.1080/19312458.2011.568376]
61. Lombard M, Snyder-Duch J, Bracken CC. Content analysis in mass communication: assessment and reporting of intercoder reliability. Hum Comm Res 2002 Oct;28(4):587-604. [doi: 10.1111/j.1468-2958.2002.tb00826.x]
62. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2022. URL: https://www.R-project.org/ [accessed 2025-09-02]
63. Federal Register: The Daily Journal of the United States Government. URL: https://www.ecfr.gov/current/title-45/part-46/section-46.102 [accessed 2026-02-25]
64. Azaria A, Mitchell T. The internal state of an LLM knows when it's lying. ArXiv Preprint posted online on April 26, 2023 [FREE Full text] [doi: 10.48550/arXiv.2304.13734]
65. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst 2025 Jan 24;43(2):1-55. [doi: 10.1145/3703155]
66. Ovsyannikova D, de Mello VO, Inzlicht M. Third-party evaluators perceive AI as more compassionate than expert humans. Commun Psychol 2025 Jan 10;3(1):4. [doi: 10.1038/s44271-024-00182-6] [Medline: 39794410]

XSL•FO
RenderX

67. Chen D, Chauhan K, Parsa R, Liu ZA, Liu FF, Mak E, et al. Patient perceptions of empathy in physician and artificial intelligence chatbot responses to patient questions about cancer. NPJ Digit Med 2025 May 13;8(1):275 [FREE Full text] [doi: 10.1038/s41746-025-01671-6] [Medline: 40360673]
68. Howcroft A, Bennett-Weston A, Khan A, Griffiths J, Gay S, Howick J. AI chatbots versus human healthcare professionals: a systematic review and meta-analysis of empathy in patient care. Br Med Bull 2025 Sep 22;156(1):ldaf017. [doi: 10.1093/bmb/ldaf017] [Medline: 41115171]
69. Seitz L. Artificial empathy in healthcare chatbots: does it feel authentic? Comput Hum Behav Artif Hum 2024;2(1):100067. [doi: 10.1016/j.chbah.2024.100067]
70. Veazey K, Notebaert A, Robertson EM. Medical student perceptions of establishing effective clinical communication: a qualitative study. Adv Health Sci Educ Theory Pract 2025 Aug 14. [doi: 10.1007/s10459-025-10468-x] [Medline: 40810854]
71. Hinkle LJ, Fettig LP, Carlos WG, Bosslet G. Twelve tips for just in time teaching of communication skills for difficult conversations in the clinical setting. Med Teach 2017 Sep;39(9):920-925 [FREE Full text] [doi: 10.1080/0142159X.2017.1333587] [Medline: 28598711]
72. England JA, Howell M, White BA. Creating a culture of communication in undergraduate medical education. Proc (Bayl Univ Med Cent) 2020 Apr 20;33(3):485-491 [FREE Full text] [doi: 10.1080/08998280.2020.1746156] [Medline: 32676001]
73. Gilligan C, Powell M, Lynagh MC, Ward BM, Lonsdale C, Harvey P, et al. Interventions for improving medical students' interpersonal communication in medical consultations. Cochrane Database Syst Rev 2021 Feb 08;2(2):CD012418 [FREE Full text] [doi: 10.1002/14651858.CD012418.pub2] [Medline: 33559127]
74. Shah SJ, Nair A, Murtagh K, Ma SP, Vogt K, Clutter D, et al. Clinician perspectives on AI-generated drafts of patient test result explanations. JAMA Netw Open 2025 Aug 01;8(8):e2528794 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.28794] [Medline: 40844780]

Abbreviations

AI: artificial intelligence
FKGL: Flesch-Kincaid Grade Level
FRE: Flesch Reading Ease
GenAI: generative artificial intelligence

Edited by Q Jin; submitted 09.Oct.2025; peer-reviewed by C Wardle, K Kim, U Pasupulety; comments to author 14.Dec.2025; revised version received 22.Jan.2026; accepted 11.Feb.2026; published 06.Mar.2026.

Please cite as:
Burns DT, Bice C, Johnson PE, Chia N, Robinson T
Comparison of Emotional Content in Text Responses From Physicians and AI Chatbots to Patient Health Queries: Cross-Sectional Study
J Med Internet Res 2026;28:e85516
URL: https://www.jmir.org/2026/1/e85516
doi:10.2196/85516
PMID:

©Daniel T Burns, Channing Bice, Paul E Johnson, Nicholas Chia, Timothy Robinson. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Young Adults' Interactions With Food and Nutrition Content on Social Media and Implications for Intervention Design: Semistructured Interview Study

Hao Tang[1], MS, MEd; Amy L Ahern[1], PhD; Marie Spreckley[1*], PhD; Andrea D Smith[1*], PhD

Institute of Metabolic Science Epidemiology, University of Cambridge, Cambridge, England, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Hao Tang, MS, MEd
Institute of Metabolic Science Epidemiology
University of Cambridge
Level 4, Institute of Metabolic Science
Addenbrooke's Hospital, Cambridge Biomedical Campus
Cambridge, England, CB2 0QQ
United Kingdom
Phone: 44 1223330315
Email: ht458@cam.ac.uk

Abstract

Background: Young adults increasingly rely on social media for nutrition information. However, little is known about (1) which types of eating-related content they actively engage with and why, and (2) how they interpret, evaluate, and incorporate this content into their everyday food choices and health behaviors.

Objective: This qualitative study explored how UK young adults (aged 18-25 years) interact with food and nutrition content across social media platforms to inform the design of future social media interventions.

Methods: Semistructured online interviews, guided by the Capability, Opportunity, Motivation–Behavior (COM-B) model, were conducted with 25 active social media users (18/25, 72% women, mean age 22.2, SD 1.9 years, ethnically diverse) in the United Kingdom between August and October 2024. The study design was informed by patient and public involvement to ensure relevance and acceptability. Data were analyzed using reflexive thematic analysis. To guide intervention development, key findings (coded as barriers and facilitators) were systematically mapped to the Theoretical Domains Framework, and the COM-B. Ethics approval was obtained from the University of Cambridge (24.368).

Results: Five key themes were identified: (1) evolving engagement patterns (passive scrolling to active interaction and mixed feelings on algorithmic control), (2) conflicted information seeking (frustration with contradictory advice, varied strategies to assess credibility), (3) multifaceted behavioral impact (simultaneous positive impacts such as cooking inspiration and negative impacts such as restrictive eating triggers), (4) shifting goals (a movement from appearance-focused to health-centered goals; yet, vulnerability to body-image issues), and (5) intervention preferences (demand for credible professionals, customizable content, and privacy protection). Participants demonstrated a reactive learning process, developing "digital nutrition literacy" often after negative experiences. Social influences were identified as the most frequently cited domain (mapped to TDF [theoretical domains framework]/COM-B) shaping interactions with social media content.

Conclusions: This study challenges assumptions of passive social media consumption, showing that young adults actively develop protective strategies yet remain vulnerable to misinformation. Digital interventions should leverage user agency and address diverse perceptions through customizable, credible content delivered with privacy and emotionally safe messaging. The COM-B and TDF mapping provide specific, evidence-based behavioral targets, particularly within the domain of Social Opportunity and Reflective Motivation, to guide the development of effective eHealth interventions.

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2023-083465

(J Med Internet Res 2026;28:e89344) doi:10.2196/89344

XSL•FO
RenderX

KEYWORDS

social media; young adults; qualitative; eating behavior; digital intervention; thematic analysis

Introduction

Young adults (18-25 years) are immersed in a digital environment where social media platforms are a dominant presence. Social media use not only shapes their social interactions but also influences their eating behaviors and dietary choices [1-3]. As of 2025, there are more than 5.6 billion social media users, and nearly all (99%) internet users aged 16 years and older in the United Kingdom use platforms such as Instagram (Meta), YouTube (Google), and TikTok (ByteDance) [4,5]. These platforms are prominent and widely accessible sources of nutrition content, offering food-related material including influencer endorsements, recipe demonstrations, health challenges, and dieting trends [6,7].

The net influence of social media on dietary behaviors is contingent on multiple interacting factors, individual predispositions, and context. On the positive side, social media can serve as a space for peer support and positive dietary role modeling, enhancing access to practical nutrition information [7-9]. Health-promoting content was found to improve nutrition literacy and encourage the adoption of healthy behaviors, particularly when delivered by relatable peers [10,11]. Young adults also report that social media facilitates recipe discovery, meal-planning inspiration, and connection with online communities focused on wellness [12-14]. On the other hand, emerging research highlights that social media users are exposed to harmful dietary misinformation [15], including advice promoting restrictive dieting and elimination of food groups, and weight stigma, which may exacerbate body dissatisfaction and disordered eating behaviors [3,5,16].

Research also indicates that frequent social media use is associated with unhealthy eating behaviors, including increased intake of sugar-sweetened beverages (SSBs) and foods high in fat, sugar and salt (HFSS), lower fruit and vegetable consumption, and skipping meals [17,18]. Increasing exposure to online food content can both promote healthier eating and reinforce unhealthy food norms; yet, the balance of these effects remains unclear.

Despite growing interest in the relationship between social media and diet and eating behaviors, there remains a limited understanding of how young people actively engage with and interpret eating-related content online, and how such interactions can be leveraged to inform the development of effective digital interventions to support health promotion. Three critical gaps constrain efforts to address this public health challenge. First, food and nutrition content is among the most widely viewed topics on social media, with hashtags such as #WhatIEatInADay, #WeightLossFood, and #Recipes collectively generating billions of views across platforms. Yet most research relies on broad exposure measures and overlooks the nuanced ways in which young adults engage (eg, saving recipes, joining challenges, or curating creators on their feeds), underscoring the need for more behavior-specific investigation [19,20]. Second**,** while recent studies have linked social media use to both healthy and disordered eating behaviors, they offer limited insight into why these effects differ between individuals [3,21,22]. Previous reviews indicate that many rely on quantitative, cross-sectional designs using broad, decontextualized measures of social media use and often treat social media users as passive consumers without exploring the individualized nature of their engagement [3,23]. These limitations are particularly relevant during young adulthood, a formative period characterized by growing dietary autonomy, health awareness, and identity development [24-26]. Therefore, qualitative insights are needed to capture the mechanisms underlying how and why young adults engage with, interpret, and respond to nutrition content over time. Third, while behavior-change models, such as the Capability, Opportunity, Motivation–Behavior (COM-B) model, are widely recommended in health behavior research, their application in social media–based interventions remains limited [27-29]. One review noted that only 8 of 33 digital health interventions used theoretical frameworks [30].

To address these gaps, we conducted a theory-informed qualitative study, mapped to COM-B and TDF (theoretical domains framework), to explore how young adults interact with diet- and eating-related social media content and how these interactions shape behavior. Grounded in lived experience, the study identifies specific behavioral targets and opportunities for health promotion that extend beyond awareness-raising within today's rapidly changing social media context.

Methods

Study Design

This qualitative study uses semistructured interviews to explore how young adults living in the United Kingdom engage with eating behavior-related content on social media. The interview guide was structured around the components of the COM-B model to provide a systematic framework for analyzing behavioral influences on dietary choices in digital environments [28]. The COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist was used (Multimedia Appendix 1) [31].

Patient and Public Involvement

Patient and public involvement (PPI) informed multiple aspects of this study, with PPI sessions conducted between March and May 2024. Four online sessions were held: 2 with members of the target population (UK-based social media users aged 18-25 years), 1 with a body-positivity social media influencer, and 1 with a staff member of a UK-based eating disorder charity. PPI feedback shaped the recruitment strategy. For example, participants recommended focusing on social media as the primary recruitment channel, prioritizing recruitment via Instagram posts and Facebook (Meta) groups and suggested additional relevant charities to contact. They also raised concerns about bot activity online, which led to the inclusion of a bot-filter question in the eligibility screening survey.

XSL•FO
RenderX

PPI members also contributed to refining the interview protocol, particularly for questions addressing potentially distressing or sensitive experiences. As a result, several questions were reworded for clarity and sensitivity. PPI contributors reviewed all participant-facing materials, including the interview guide, recruitment flyers, and the participant information sheet, and provided feedback on the clarity, tone, and cultural inclusivity of language used throughout. For example, a preliminary question was added asking participants to define what healthy eating meant to them personally before proceeding to questions about social media use, ensuring questions were grounded in participants' own frameworks. Additionally, on the advice of PPI contributors, a sensitivity disclaimer was inserted before questions addressing disordered eating risk, informing participants that the following questions may be distressing and that they were free to skip or pause at any point.

Participants and Recruitment

Participants were eligible if they were aged 18-25 years, lived in the United Kingdom, spoke English, and used social media (Instagram, TikTok, YouTube, etc) most days of the week. Recruitment was conducted from August to October 2024 through a multipronged approach, including posting on social media platforms (Instagram, Reddit, and Facebook groups) and posters in public locations. Interested individuals completed an online screening survey via REDCap (Research Electronic Data Capture; Vanderbilt University) to assess eligibility. From the pool of eligible respondents who provided authentic answers, participants were purposively selected to reflect diversity across gender, ethnicity, education level, and employment status. If individuals declined participation, replacements were drawn from the same demographic subgroup to maintain diversity.

The study sample size was informed by the concept of information power [32]. With the focused research aim, use of an established theoretical framework (the COM-B framework), and a demographically diverse sampling strategy, we aimed to recruit approximately 30 participants. Throughout data collection, we monitored the richness and relevance of interview material in relation to the study aim. We concluded the interviews when we felt that we had enough detailed interviews from diverse perspectives to provide rich insights into the research question, and had developed coherent analytic themes that answered the question, with depth and coverage across the components of the COM-B framework. These assessments were made through ongoing discussion between the lead researcher (HT) and the supervisory team (ALA, ADS, and MS), and data collection concluded at 25 interviews when these criteria were collectively satisfied.

Data Collection

Semistructured interviews were conducted online via Microsoft Teams by a PhD researcher (HT) and were audio-recorded. Each interview lasted approximately 60 minutes. Field notes were made immediately after each interview to capture contextual information and initial reflections.

Researcher Reflexivity Statement

The interviewer (HT) is a female PhD researcher in behavioral nutrition with training and prior experience in conducting qualitative interviews. Her background in nutrition and dietetics, and experience creating evidence-based content on social media, helped her understand participants' experiences but also required awareness of potential assumptions about "healthy" content and online engagement. As someone with lived experience of navigating food and body image in the context of social media, HT was reflexively aware of potential shared experiences with participants and took care to remain open to perspectives that might differ from her own experiences. As a young adult and active social media user, HT shared generational proximity with participants, which may have facilitated rapport while also requiring reflexive awareness of shared assumptions. The power dynamics inherent in the researcher-participant relationship were acknowledged, particularly given the sensitive nature of topics such as body image and eating behaviors. HT kept brief reflexive notes after each interview and met regularly with supervisors (ALA and AS) to discuss any questions or uncertainties that arose during data collection.

Interview Guide and Measures

The interview guide (Multimedia Appendix 2) covered topics including social media use patterns, experiences with healthy eating content, perceived influences on eating behaviors, platform preferences and engagement factors, and barriers and facilitators to healthy eating.

Data Management

Audio recordings were transcribed and anonymized. All data were stored securely on the MRC (Medical Research Council) Epidemiology Unit Secure Research Drive (SRD), with access restricted to authorized members of the research team. Participants and their transcripts were assigned unique numerical IDs to ensure confidentiality.

Data Analysis

Phase 1: Thematic Analysis

Data were analyzed using reflexive thematic analysis following Braun and Clarke's six-phase approach [33], selected for its capacity to generate nuanced, participant-centered insights. HT led the analysis, using inductive coding of all transcripts in NVivo, whereby codes were generated in response to participants' accounts without the imposition of a predetermined framework, and analytic themes were developed through an iterative process. To support analytic reflexivity and enrich interpretation, a subset of transcripts (5/25, 20%) was also coded by MS. These additional coding contributions informed collaborative discussions about alternative readings and identified patterns, contributing to the refinement of the thematic structure. The research team met regularly to discuss coding decisions and evolving themes, which supported reflexive engagement with the data and strengthened the coherence of the final analysis.

In this study, key concepts such as "healthy eating" and "credible content" were not defined a priori for participants during interviews. Instead, these were explored as open, participant-driven constructs, reflecting how individuals themselves interpreted and evaluated nutrition-related content on social media. For analytic consistency during subsequent

XSL•FO
RenderX

COM-B and TDF mapping, references to "healthy eating" were considered in relation to broadly accepted UK public health dietary guidance [34], while remaining grounded in participants' expressed meanings. "Credible content" was analytically defined as nutrition-related information that participants perceived as trustworthy, typically based on cues such as professional credentials, scientific references, tone, presentation style, or community validation.

Phase 2: Behavioral Framework Mapping

Identified determinants of young adults' social media use were subsequently mapped to the TDF [35] and the COM-B model [28]. The COM-B model explains that for behavior change to occur, people must have the capability to change, the motivation to change and the opportunity to change. The TDF build on this by outlining 14 domains that capture the wider range of individual barriers and facilitators that may influence behavior. These domains systematically map to the COM-B model components (Multimedia Appendix 3). TDF and COM-B were selected because they support the translation of qualitative insights into behavioral domains and validated behavioral change mechanisms, which provide a strong foundation for future intervention development.

Each subtheme derived from the qualitative analysis was coded as either a barrier or a facilitator, or both where applicable, and subsequently categorized into the relevant TDF and COM-B domains. Influences could act in both directions depending on context and were coded as barrier/facilitator. Mapping barriers and facilitators to the TDF and COM-B allowed us to systematically identify which psychological, social, and contextual domains were most relevant to healthy eating within the social media context. Framework mapping was led by HT, with ADS providing an additional interpretative perspective. Interpretations were discussed iteratively to refine domain allocation and enhance conceptual clarity.

Ethical Considerations

The study received ethical approval from the University of Cambridge School of Humanities and Social Sciences Research Ethics Committee (reference number 24.368). Prior to each interview, all participants provided informed consent electronically via a digital consent form. Participants were assigned unique numerical identifiers, and pseudonymized transcripts and personal data were stored separately and securely on the MRC Epidemiology Unit Secure Research Drive, accessible only to authorized members of the research team. All interview participants received a £25 (US $32) digital voucher as remuneration for their time, and all PPI contributors each received a £25 (US $32) digital voucher for their 1-hour session.

Results

Sample Characteristics

Of the 317 individuals who completed the initial online recruitment questionnaire, 277 met the eligibility criteria. A total of 133 responded to follow-up questions, and 30 individuals were scheduled for interviews. Among these, 27 participants joined the interview process. Two participants were excluded from the analytic sample: one whose responses did not sufficiently address the research questions concerning food and nutrition content on social media, and one who used racially offensive language that was inconsistent with the study's ethical standards and respectful participation guidelines, resulting in a final analytic sample of 25 participants. Sufficient information power to answer our research question was achieved at 25 interviews. An overview of participant characteristics is provided in Table 1.

Participants had a mean age of 22.2 (SD 1.9) years and were predominantly women (18/25, 72%). The sample was ethnically diverse, comprising White (9/25, 36%), Asian/Asian British (8/25, 32%), Black/African/Caribbean/Black British (6/25, 24%), mixed/multiple ethnic groups (1/25, 4%), and other ethnic groups (1/25, 4%). Educational attainment varied, but most participants held a bachelor's degree or equivalent (11/25, 44%) or had completed, or were enrolled in, postgraduate education (7/25, 28%).

In terms of occupation, 60% (15/25) were students, 16% (4/25) were employed full-time, 16% (4/25) were employed part-time, and 8% (2/25) were engaged in apprenticeships or traineeships.

All participants reported daily social media use, typically around 2-3 hours per day. Instagram (Meta) and TikTok (ByteDance) were the most frequently used platforms for accessing health- and nutrition-related content.

XSL•FO
RenderX

Table 1. Demographic characteristics of participants (N=25).[a]

Characteristics	Values, n (%)
Age (years), mean (SD)	22.2 (1.9)
Gender	
Woman	18 (72)
Man	7(28)
Ethnicity	
Asian/Asian British	8 (32)
Black/African/Caribbean/Black British	6 (24)
White	9 (36)
Mixed/multiple ethnic groups	1 (4)
Other ethnic group	1 (4)
Occupational status	
Student	15 (60)
Employed full-time	4 (16)
Employed part-time	4 (16)
Apprenticeship or traineeship	2 (8)
Education level	
Postgraduate degree	7 (28)
Bachelor's degree or equivalent	11 (44)
Post-secondary diploma or certificate	6 (24)

[a]This table provides a detailed summary of demographic, educational, and occupational characteristics of the study sample. Continuous variables are presented as mean (SD), and categorical variables are presented as counts and percentages.

Overview of Themes

Overall, five themes were identified, describing the diverse ways in which young adults engage with social media and how this influences their eating behaviors. Themes and subthemes are illustrated in Figure 1 and described in detail below.

XSL•FO
RenderX

Figure 1. Thematic framework of young adults' interactions with eating behaviour-related social media content.

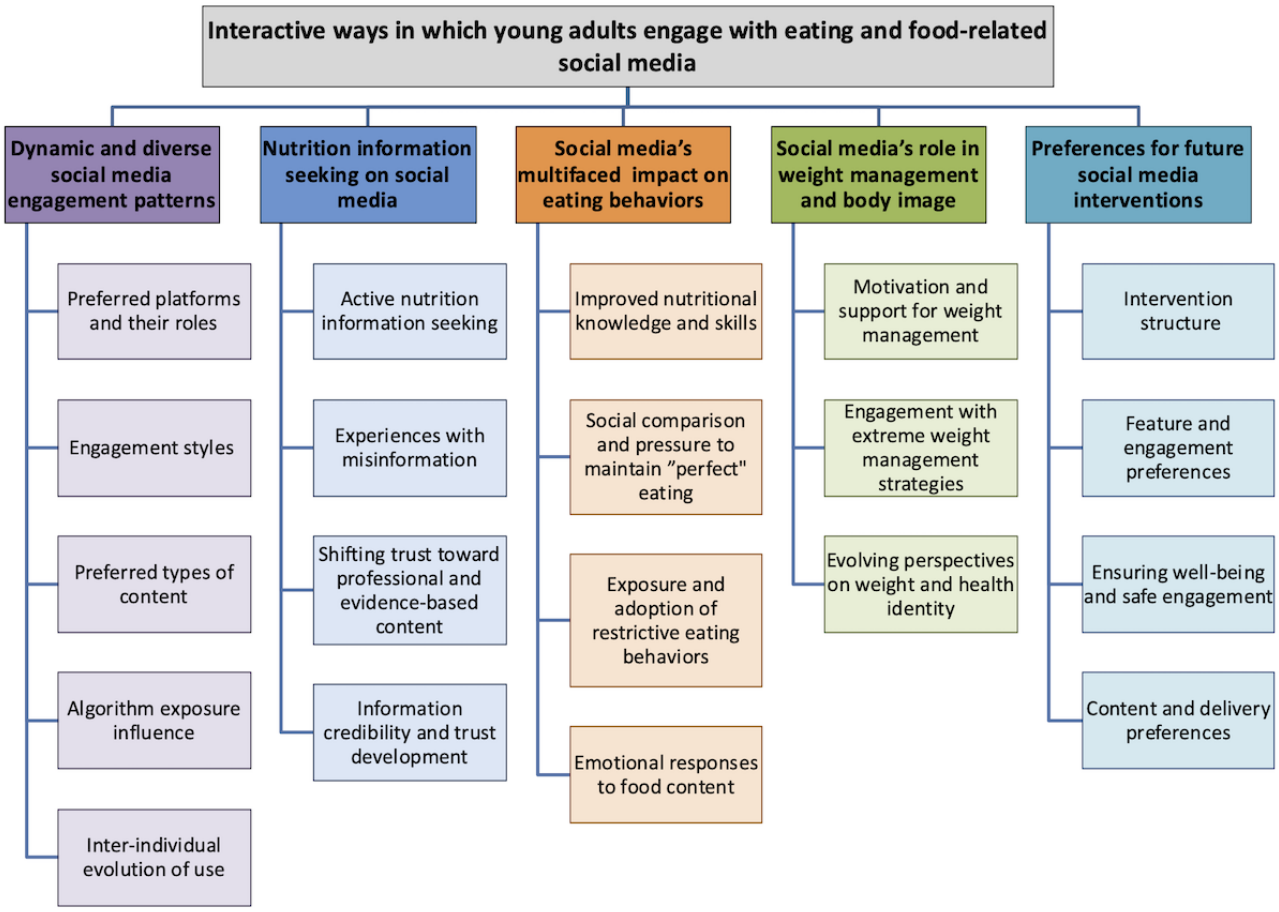

Theme 1: Evolving and Diverse Social Media Engagement Patterns

Participants described a range of ways in which they engaged with nutrition and dietary content on social media, emphasizing that their experiences were both highly personal and platform-specific. Engagement patterns were dynamic, evolving over time as participants developed preferences, adapted to changing algorithms, or sought to manage the emotional impact of content.

Platforms were associated with distinct benefits: Instagram was valued for community and peer support, YouTube for providing in-depth, educational material, and TikTok for engaging in short-form content. TikTok was also described by some as addictive or overwhelming. Engagement ranged from passive scrolling to active participation, such as commenting, voting in polls, or joining live streams.

Content appeal was strongly linked to visual engagement and relatability, such as the "What I eat in a day." This video features creators, often celebrities or influencers, documenting their daily food and drink intake in a 24-hour time window. Participants perceived these videos as accessible and relatable because of their straightforward structure and perceived authenticity.

> *I watch a lot of food videos and get to know about different cultures. And it helped me realise the importance of adding vegetables in my diet…it helped me a lot. That was also where I learned to drink enough water instead of soft drinks. [HT prompted for example] I followed a marathon runner who's vegan. I'm not attracted to veganism at all, but I liked her ideas of adding more vegetables in cooking. I remember watching her post about a chickpea sandwich, I know chickpeas are a good source of protein. It made me think it would be nice to eat less meat and try different food. She's a well-known creator, so I trusted her, and I've actually made her recipes before.* [Woman, 20 years old]

However, not all experiences with social media content were positive. Some participants expressed frustration at the lack of control over what appeared on their feeds.

> *I watched a few fitness videos on TikTok last week and now my feed is just... lots of workout content and weight loss videos. It makes me feel kind of pressured to work out more, you know? …Sometimes I feel like I have no control over what I see; the algorithm decides for me.* [Woman, 20 years old]

Over time, some participants reported making deliberate adjustments to their social media use, such as unfollowing accounts, deleting apps, or reducing screen time, reflecting an active effort to manage their well-being.

XSL•FO
RenderX

> *I started to cut down my Instagram use because it made me feel uneasy about my own body.* [Woman, 24 years old]

Theme 2: Nutrition Information Seeking on Social Media

Participants used social media as a primary source for quick nutrition information (eg, recipes and meal planning). However, utility was tempered by the prevalence of misinformation and conflicting dietary advice. Participants described adopting a number of strategies to verify information, such as checking credentials, cross-referencing sources, and reviewing comment sections, reflecting a developing "digital nutrition literacy."

> *I often search on Xiaohongshu (also known internationally as RedNote) when I don't know what to cook.* [Man, 23 years old]

Nevertheless, confusion remained high due to contradictory health claims online.

> *Too much conflicting information makes it hard to decide what's right. Like, I saw posts saying that eating fruit is not healthy and doesn't help digestion, and I saw other posts saying that drinking water immediately after eating is not healthy either. But also know from other sources that drinking water while eating is good for weight loss and fruit helps with digestion.* [Woman, 24 years old]

These were sometimes perceived as being driven by commercial interests rather than science.

> *You don't know—some experts will say this is healthy, then a few years later, people change their minds. Like coffee—before it was bad, now they say it's good for your heart. It's confusing and feels like part of advertising.* [Woman, 23 years old]

Trust was influenced by both professional credentials and superficial cues such as the presenter's appearance or the level of social validation (eg, positive comments) from the community.

> *Sometimes, when the influencer... looks healthy and positive, that helps [with credibility]. If they look really fit, I want to try their recommended snacks or diet…. I guess I think if they look that way, then what they're recommending must work.* [Woman, 25 years old]

> *I look at the comments to see what other people have said. If most people say they got good results, then I'm more likely to trust what they (the creator/influencer) recommend.* [Woman, 19 years old]

Theme 3: Social Media's Multifaceted Impact on Eating Behaviors

The influence of social media was described as simultaneously positive and negative. Positive effects included skill development (eg, cooking techniques), improved knowledge, and inspiration for balanced meal choices.

> *I started cooking at home more after seeing easy and tasty recipes online.* [Woman, 23 years old]

Conversely, exposure to curated and unrealistic dietary and body ideals often led to social comparison, self-doubt, and the adoption of harmful behaviors, including restrictive eating, obsessive weighing, and calorie counting.

> *Seeing influencers made me feel inspired but sometimes made me question myself because they make diets look so easy.* [Woman, 24 years old]

> *When I just entered college, I gained a lot of weight and wanted to lose it. I found this influencer whose 'what I eat in a day' and weight loss tips videos I really enjoyed watching. She is a famous model. So I started trying the weight loss tips she recommended. After almost two months of skipping dinner due to posts promoting it as healthy, I developed eczema and felt horrible physically and mentally. That's when I realised that these 'model' diets were very risky. So I stopped and unfollowed many accounts.* [Woman, 25 years old]

Reactions to specific content types were subjective. For example, mukbang videos (online broadcasts of individuals eating large quantities of food) elicited opposing responses: some participants felt they reduced cravings, while others found them uncomfortable or even disturbing.

Theme 4: Social Media's Role in Weight Management and Body Image

Social media played a shifting role in shaping participants' body image concerns. Positive influences included motivation from success stories and support from health-focused communities.

> *I've been overweight since I was a child, and I started to feel really ashamed from my teenage years, like in middle school. I just couldn't get fit. I thought I was just lazy and born to be fat. Seeing success stories of people with obesity inspired me to start eating better.* [Man, 25 years old]

However, many participants recounted past experiences of body dissatisfaction exacerbated by unrealistic beauty standards promoted online. Over time, some reported deliberate progression from appearance-focused goals in their younger years to more balanced, health-centered approaches.

> *I think for me it used to affect it negatively… I saw models and pop stars… I wanted to be like that… I tried loads of diets like keto, detox, all kinds of things. Often I'd feel awful about myself, like, why can't I just be like them? There was also a lot of talk on social media saying these foods are bad for you, and because of that you have to punish yourself with workouts. It was a lot more unhealthy, and because of that, I think the impact was negative. Now I take from social media... making sure I'm eating enough. I think for me it used to affect it negatively, but I don't think I realized that and that was when I was younger.* [Woman, 25 years old]

XSL•FO
RenderX

This shift demonstrates an evolving self-regulatory process aimed at mitigating the negative psychological effects of weight-centric content.

Theme 5: Preferences for Future Social Media Interventions

Participants expressed a clear and diverse preference for future digital interventions, emphasizing that interventions should be short, visually engaging, and emotionally safe. Key preferences included short videos (30-60 seconds), involvement of professionals (eg, a qualified dietician) to build trust, interactive features (live Q&A sessions and journaling), and a privacy-protective environment (eg, anonymous chat).

I think somethinganonymous and like no judgments could be useful. I think sharing eating or comment about eating can be sensitive. [Woman, 24 years old]

> *Socializing within the programme can be really helpful to keep people tied because they can explore other personalities, or they can enjoy their time doing the programme.* [Woman, 22 years old]

Most participants preferred a time-limited intervention (2-4 weeks) with regular updates and structured, measurable goals to maintain motivation and prevent overwhelm.

Mapping Identified Determinants of Eating and Food-Related Social Media Behaviors to the TDF

To inform the design of future interventions, key subthemes identified from the qualitative data were systematically mapped onto the TDF and subsequently COM-B. Each subtheme was evaluated and categorized as either a barrier or a facilitator to the use of social media interventions to promote healthy eating. A total of 106 barriers and facilitators were mapped across the dataset, with the largest proportion aligned with Motivation, followed by Opportunity and Capability.

Within the TDF, Social Influences (Opportunity) was the most frequently coded domain, reflecting the pervasive role of peers, influencers and community feedback in shaping young adults' interactions with content, perceived credibility, and subsequent eating behaviors (peer influence leading to social comparison and community support for accountability). Barriers were concentrated in domains such as Environmental Context and Resources (Opportunity; eg, prevalence of misinformation and algorithm control) and Emotion (Motivation; eg, guilt and feeling triggered). Facilitators are frequently mapped to Knowledge (Capability) and Skills (Capability; eg, recipe demonstrations and critical evaluation skills) and Goals (Motivation; eg, shift toward sustainable health goals). These findings highlight that digital interventions need to target not only individual knowledge and skills but also the social and environmental context that dictates exposure and emotional response. Table 2 provides a summary of all barriers and facilitators alongside their theoretical domains. A comprehensive item-by-item mapping of all subthemes and barrier/facilitator pairs (n=106) is provided in Multimedia Appendix 4.

XSL•FO
RenderX

Table 2. Summary of key themes, barriers and facilitators to healthy eating when engaging with food and nutrition content on social media, mapped to Theoretical Domains Framework and Capability, Opportunity, Motivation–Behavior.

Themes[a]	Key findings	Primary barriers	Primary facilitators	TDF[b] Domain (COM-B[c])
Theme 1: Evolving engagement patterns	• Participants used different platforms distinctly • Algorithms shaped exposure positively and negatively	• Algorithm-driven control • Limited user autonomy over content exposure	• Algorithm-driven tailoring. • Platform affordances for engagement. • Peer influence motivation.	Environmental Context & Resources (O); Beliefs about Capabilities (M)
Theme 2: Nutrition information seeking	• High engagement but confusion from contradictory advice. • Shift toward professional credibility	• Misinformation prevalence. • Difficulty evaluating contradictory claims	• Critical-evaluation skill development. • Preference for evidence-based sources. • Professional credentials.	Environmental Context & Resources (O); Skills (C); Knowledge (C)
Theme 3: Multifaceted behavioral impact	• Positive (skill development) and negative (social comparison and restrictive eating) impacts coexist	• Unrealistic ideals • Normalized restrictive eating • Emotional triggers from comparison	• Practical guidance. • Knowledge access. • Diverse content formats. • Emotional support.	• Social Influences (O) • Emotion (M) • Skills (C) • Knowledge (C)
Theme 4: Weight management and body image	• Evolving shift from appearance-focused to health-centered goals. • Persistent vulnerability to ideals	• Extreme dieting promotion. • Unrealistic body ideals. • Pressure to conform to thinness	• Success stories. • Community support. • Skill development through shared techniques. • Critical evaluation skills.	• Social Influences (O) • Beliefs about Consequences (M) • Skills (C)
Theme 5: Intervention preferences	• Strong demand for credible, customizable, emotionally safe content delivered in short formats	• Algorithm opacity • Content volume overwhelm	• Short-form visuals (30-60 s). • Professional delivery. • Interactive features. • Privacy protection. • Time-limited structure.	• Environmental Context and Resources (O) • Knowledge (C) • Emotion (M) • Behavioral Regulation (C)

[a]This table summarizes the five key themes identified through reflexive thematic analysis, with primary barriers and facilitators identified within each theme and mapped to their corresponding Theoretical Domains Framework domains and Capability, Opportunity, Motivation–Behavior components. Subthemes within each theme have been condensed to show the most salient findings.

[b]TDF: Theoretical Domains Framework.

[c]COM-B: Capability, Opportunity, Motivation–Behavior.

Discussion

Principal Findings

This study highlights the complex and dynamic role of social media in shaping young adults' eating behaviors, revealing a duality where platforms simultaneously facilitate healthy habits (eg, recipe discovery and positive motivation) and expose users to potentially harmful content (eg, misinformation, restrictive eating). Our findings, captured across five themes, detail the mechanisms of this influence. The COM-B/TDF mapping further identified that most barriers and facilitators fell under Motivation and Opportunity, underscoring that effective interventions must target Social Influences (a key Opportunity domain) and Reflective Motivation (eg, goals and beliefs about consequences), in addition to enhancing individual Capability. This highlights key leverage points for the design of future interventions.

Comparison With Existing Literature

Consistent with previous literature, we found that social media can increase body image concerns and disordered eating risk, particularly among young women [3,36,37], while also supporting health-promoting behaviors and knowledge acquisition when used thoughtfully [38,39]. This study builds on these insights by emphasizing the importance young users place on credibility, a holistic approach, a nonjudgmental tone, and emotionally supportive environments.

This duality was clearly reflected in participants' mixed experiences with evolving social media content, reflecting broader concerns about the overwhelming volume and variable quality of digital health information, which can make it difficult for users to distinguish trustworthy from misleading content [40,41]. This underscores the importance of equipping users with targeted digital health literacy skills and tailoring digital support programs to address the confusion arising from the volume and vagueness of online content [41-43].

XSL•FO
RenderX

This study advances current understanding in five ways. First, the study identified a dynamic behavioral pattern in how young adults engage with social media. Participants described a shift from appearance-related content in their earlier years (adolescence) to more health-focused, carefully evaluated use as they aged, especially after the transition to adulthood. This mirrors recent trends in UK youth, who have become more thoughtful and selective in their digital habits over time [23,44]. By capturing users' reflective accounts and behavioral transitions, our study adds novel, person-centered insights with direct implications for tailoring digital interventions to users' developmental stages rather than applying uniform approaches across all young adults.

Second, participants demonstrated sophisticated navigation strategies involving credential checking, community-feedback assessment, cross-referencing sources, and appearance-based judgments to evaluate nutrition information. These findings echo recent reviews calling for recognition of the varied and active ways young users engage with health information online, rather than characterizing them as passive consumers [45-47]. Liu and Zhang (2024) [48] observed similar deliberate credibility judgements in video-based content, while Kuutila et al (2023) [49] found that source expertise and prior belief alignment were stronger predictors of perceived credibility than evidence quality. While some critical evaluation skills were demonstrated (eg, credential checking and cross-referencing), our analysis also underscored the need for interventions that strengthen digital nutrition literacy by addressing the difficulty in evaluating contradictory claims.

Third, this study offers more nuanced insights into how young adults balance credibility and relatability when evaluating nutrition-related content online. While participants increasingly valued professional credentials and evidence-based sources, they remained highly engaged with content created by nonprofessional figures, such as influencers, models, and celebrities, including "what I eat in a day" videos or personal weight loss journeys, suggesting that digital health interventions may be more engaging when they intentionally combine both elements. Moreover, reactions to specific formats, such as weight-loss comparison and mukbang videos, varied widely. While some viewers described feeling comfort or inspiration, others felt guilt or discouragement. These findings echo recent studies [50-52], highlighting the heterogeneity in user responses and the need for personalized, context-sensitive intervention strategies.

Fourth, findings reveal a key tension in digital health engagement that challenges current intervention approaches. Participants appreciated algorithmic curation for efficiently surfacing relevant content but also expressed discomfort with its control over their exposure. Similarly, Milton et al [53] characterized the TikTok "For You" page as a "runaway train," which is highly engaging but sometimes overwhelming and difficult to escape. Additionally, Savolainen and Ruckenstein [54] explored how users navigate algorithmic systems with varying degrees of autonomy, suggesting that while personalization can aid discovery, it often conflicts with self-directed control. These insights highlight a design challenge for digital health interventions, indicating that leveraging platform features for engagement should be balanced with mechanisms that preserve user autonomy and transparency.

Finally, this study highlights how people interact with different social media platforms in different ways. We found that in the context of dietary content, Instagram primarily fulfils community-building functions, YouTube mainly serves as a space for educational and in-depth content, and TikTok facilitates highly engaging, short-form delivery that can be overwhelming. These findings align with research suggesting that different platforms offer distinct affordances [55] and extend earlier qualitative work showing that young adults interact with healthy content differently across platforms [23]. This ecological perspective implies that effective digital health interventions should respect and leverage each platform's unique characteristics rather than applying uniform approaches across all platforms.

Insights for Social Media Intervention Design

This study offers several practical insights for the development of user-centered social media interventions to support healthy eating. A key finding was the high variability in content preferences, engagement styles, and emotional responses to dietary content, including mukbang videos and weight loss promotions. This highlights the need for customization options in future interventions. Such options may include adjustable content formats (eg, infographics and videos), frequency of posts, communication methods (eg, anonymous group chats, comments, and reposting), and types of health messaging (HAO Tang et al, unpublished data, August 2025)

Participants also expressed a preference for shorter content and multicomponent interventions, combining education with motivation and interaction. Realistic, relatable content delivered in emotionally safe tones (ie, nonjudgmental, supportive language that avoids triggering negative emotions or disordered eating behaviors) was consistently valued. Relatability was enhanced when content reflected the lived experiences of diverse individuals, rather than idealized or overly polished representations.

Our findings also identified a critical gap between information exposure and behavioral change. While social media served as a source of motivation and ideas, sustaining healthy habits remained challenging for many participants [56-58]. This highlights the importance of bridging the intention-behavior gap through established behavior-change techniques [59]. Our COM-B mapping results suggest that, while some young adults may possess adequate nutrition knowledge (psychological capability), more tailored strategies are needed to enhance Reflective Motivation and Social Opportunity for sustained behavior change. Specifically, interventions could incorporate goal-setting and self-monitoring (motivation/behavioral regulation) and use algorithm transparency tools to give users greater control over the content they are exposed to (physical opportunity). Social influence mechanisms may be targeted through leveraging credibility cues such as presenter credentials and community validation (eg, community moderation features and community comments).

XSL•FO
RenderX

The confusion participants experienced when navigating contradictory health messages points to a critical design requirement: interventions must equip users with frameworks or skills for evaluating conflicting advice rather than simply adding to the information volume [42,60]. This suggests incorporating features such as source-credibility checklists, decision-making tools that help users weigh competing claims, and educational modules that teach users to recognize common patterns in misleading nutrition content. Furthermore, many participants suggested that peer sharing, algorithmic visibility, and familiar platform features must be balanced with privacy considerations. The flexibility to choose levels of interaction (eg, passive viewing versus active commenting or posting) may help accommodate diverse comfort levels.

Strengths and Limitations

This study's strengths include the diverse sample achieved through purposive sampling, enhancing transferability within the UK young adult population. Comprehensive PPI throughout the research process strengthened the study's cultural sensitivity and appropriateness, and applying a qualitative approach captured nuanced experiences otherwise not measurable using more quantitative approaches, essential for user-centered intervention design. The integration of COM-B and TDF provided both theoretical rigor and practical relevance, which paves a clear pathway for intervention development. This, coupled with the direct exploration of participant preferences for future interventions (Theme 5), offers practical guidance often missing from social media health research.

Some limitations should be noted. As participation was voluntary, the sample may be subject to selection bias, with perspectives potentially skewed toward those with a stronger interest or engagement in health-related topics or social media. Findings regarding specific platforms or content types need to be interpreted with caution, as the rapid evolution of social media technologies and user behaviors may limit their long-term applicability.

Conclusion

This study provides new insights into how young adults in the United Kingdom interact with nutrition-related content on social media. Participants described a dynamic process of engagement shaped by evolving values, critical evaluation skills, and emotional responses. While social media can support healthy eating by offering accessible and motivational content, it also exposes users to conflicting messages and potentially harmful imagery and advice.

Findings highlight the importance of moving toward personalized digital health interventions that reflect users' diverse needs, developmental stages, and evaluation capabilities. To increase the likelihood of achieving positive behavior change, interventions should combine credible, emotionally safe content with customizable features, privacy protection, professional oversight, and be grounded in established behavior change theory (COM-B/TDF). The tension between algorithmic curation and user autonomy suggests that successful interventions need to balance engagement optimization with user control and leverage multiplatform strategies that respect the unique and evolving features of different digital environments.

Acknowledgments

The authors thank the study participants and patient and public involvement contributors for their time and input and gratefully acknowledge Dr Esther van Sluijs for her intellectual contributions through discussions and feedback during research meetings. ChatGPT (OpenAI) was used for grammar checking and sentence refinement. To enhance data privacy, chat history and model training were turned off. The tool was not used for data analysis, interpretation, conceptual development, or content generation, and all text was critically reviewed and approved by the authors.

Funding

This research was supported by the Medical Research Council (Grants MC_UU_00006/5, MC_UU_00006/6) and National Institute for Health and Care Research (NIHR) Cambridge Biomedical Research Centre (NIHR203312). The funder had no role in the study design, data collection, analysis, interpretation, or manuscript preparation.

Data Availability

The terms of participant consent allow us to share deidentified qualitative data generated and analyzed during this study with researchers who are undertaking analyses related to the original objectives, with institutional approval. Requests should be directed to datasharing@mrc-epid.cam.ac.uk.

Authors' Contributions

Conceptualization, data curation, formal analysis, investigation, methodology, project administration, visualization, writing (original draft), writing (review and editing): HT
Conceptualization, funding acquisition, methodology, supervision, writing (review and editing): ALA
Formal analysis (secondary coding of a subset of transcripts, n=5), methodology, and writing (review and editing): MS
Conceptualization, formal analysis (framework mapping), methodology, supervision, and writing (review and editing): ADS

XSL•FO
RenderX

Conflicts of Interest

ALA is on the scientific advisory board for Weight Watchers (WW) and has undertaken consultancy (all payments to her institution) for Menwell Limited.

Multimedia Appendix 1
COREQ 32-item checklist.
[DOCX File , 19 KB - jmir_v28i1e89344_app1.docx]

Multimedia Appendix 2
Semi-structured interview guide used in the qualitative study.
[DOCX File , 19 KB - jmir_v28i1e89344_app2.docx]

Multimedia Appendix 3
Mapping of Theoretical Domains Framework (TDF) domains to COM-B components.
[DOCX File , 16 KB - jmir_v28i1e89344_app3.docx]

Multimedia Appendix 4
Detailed Item-Level Mapping of Barriers and Facilitators Healthy Eating When Engaging with Food and Nutrition Content on Social Media.
[DOCX File , 27 KB - jmir_v28i1e89344_app4.docx]

References

1. Goodyear VA, Wood G, Skinner B, Thompson JL. The effect of social media interventions on physical activity and dietary behaviours in young people and adults: a systematic review. Int J Behav Nutr Phys Act 2021;18(1):72 [FREE Full text] [doi: 10.1186/s12966-021-01138-3] [Medline: 34090469]
2. Ardi Z, Putri SA. The analysis of the social media impact on the millennial generation behavior and social interactions. Southeast Asian Journal of Technology and Science 2021;2(1):70-77 [FREE Full text]
3. Dane A, Bhatia K. The social media diet: a scoping review to investigate the association between social media, body image and eating disorders amongst young people. PLOS Glob Public Health 2023;3(3):e0001091 [FREE Full text] [doi: 10.1371/journal.pgph.0001091] [Medline: 36962983]
4. Number of internet and social media users worldwide as of February. Statista. 2025. URL: https://www.statista.com/statistics/617136/digital-population-worldwide/ [accessed 2026-03-18]
5. Adults' media use and attitudes report. Ofcom. 2025. URL: https://tinyurl.com/r5j94jjr [accessed 2026-03-18]
6. Stirling E, Willcox J, Ong K, Forsyth A. Social media analytics in nutrition research: a rapid review of current usage in investigation of dietary behaviours. Public Health Nutr 2020;24(6):1193-1209. [doi: 10.1017/s1368980020005248]
7. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021;23(5):e17917 [FREE Full text] [doi: 10.2196/17917] [Medline: 33978589]
8. Powell J, Pring T. The impact of social media influencers on health outcomes: systematic review. Soc Sci Med 2024;340:116472 [FREE Full text] [doi: 10.1016/j.socscimed.2023.116472] [Medline: 38070305]
9. Popat A, Tarrant C. Exploring adolescents' perspectives on social media and mental health and well-being - a qualitative literature review. Clin Child Psychol Psychiatry 2023;28(1):323-337 [FREE Full text] [doi: 10.1177/13591045221092884] [Medline: 35670473]
10. Zimmermann E, Tomczyk S. The ways of using social media for health promotion among adolescents: qualitative interview and focus group study. J Med Internet Res 2025;27:e71510 [FREE Full text] [doi: 10.2196/71510] [Medline: 40489771]
11. Evans W, Andrade E, Pratt M, Mottern A, Chavez S, Calzetta-Raymond A, et al. Peer-to-Peer social media as an effective prevention strategy: quasi-experimental evaluation. JMIR Mhealth Uhealth 2020;8(5):e16207 [FREE Full text] [doi: 10.2196/16207] [Medline: 32374270]
12. Molenaar A, Jenkins EL, Brennan L, Lukose D, McCaffrey TA. The use of sentiment and emotion analysis and data science to assess the language of nutrition-, food- and cooking-related content on social media: a systematic scoping review. Nutr Res Rev 2023;37(1):43-78. [doi: 10.1017/s0954422423000069]
13. Sokolova K, Vessal SR, Perez C. Home cooking in the digital age: when observing food influencers on social media triggers the imitation of their practices. Psychology & Marketing 2024:1152-1171 [FREE Full text] [doi: 10.1002/mar.21973]
14. Chung A, Vieira D, Donley T, Tan N, Jean-Louis G, Kiely Gouley K, et al. Adolescent peer influence on eating behaviors via social media: scoping review. J Med Internet Res 2021;23(6):e19697 [FREE Full text] [doi: 10.2196/19697] [Medline: 34081018]
15. Denniss E, Lindberg R, McNaughton SA. Quality and accuracy of online nutrition-related information: a systematic review of content analysis studies. Public Health Nutr 2023;26(7):1345-1357. [doi: 10.1017/s1368980023000873]

XSL•FO
RenderX

16. Yurtdaş-Depboylu G, Kaner G, Özçakal S. The association between social media addiction and orthorexia nervosa, eating attitudes, and body image among adolescents. Eat Weight Disord 2022;27(8):3725-3735. [doi: 10.1007/s40519-022-01521-4] [Medline: 36562891]
17. Fretes G, Veliz P, Narvaez A, Williams D, Sibille R, Arts M, et al. Digital marketing of unhealthy foods and non-alcoholic beverages to children and adolescents: a narrative review. Curr Dev Nutr 2025;9(2):104545 [FREE Full text] [doi: 10.1016/j.cdnut.2025.104545] [Medline: 40046995]
18. Zeeni N, Abi Kharma J, Malli D, Khoury-Malhame M, Mattar L. Exposure to Instagram junk food content negatively impacts mood and cravings in young adults: a randomized controlled trial. Appetite 2024;195:107209 [FREE Full text] [doi: 10.1016/j.appet.2024.107209] [Medline: 38228246]
19. Kucharczuk AJ, Oliver TL, Dowdell EB. Social media's influence on adolescents' food choices: a mixed studies systematic literature review. Appetite 2022;168:105765. [doi: 10.1016/j.appet.2021.105765] [Medline: 34687823]
20. Wilksch SM, O'Shea A, Ho P, Byrne S, Wade TD. The relationship between social media use and disordered eating in young adolescents. Int J Eat Disord 2020;53(1):96-106. [doi: 10.1002/eat.23198] [Medline: 31797420]
21. Diekman C, Ryan C, Oliver T. Misinformation and disinformation in food science and nutrition: impact on practice. J Nutr 2023;153(1):3-9 [FREE Full text] [doi: 10.1016/j.tjnut.2022.10.001] [Medline: 36913465]
22. Filippone L, Shankland R, Hallez Q. The relationships between social media exposure, food craving, cognitive impulsivity and cognitive restraint. J Eat Disord 2022;10(1):184 [FREE Full text] [doi: 10.1186/s40337-022-00698-4] [Medline: 36434703]
23. Lim MSC, Molenaar A, Brennan L, Reid M, McCaffrey T. Young adults' use of different social media platforms for health information: insights from web-based conversations. J Med Internet Res 2022;24(1):e23656 [FREE Full text] [doi: 10.2196/23656] [Medline: 35040796]
24. Breton É, Dufour R, Côté SM, Dubois L, Vitaro F, Boivin M, et al. Developmental trajectories of eating disorder symptoms: a longitudinal study from early adolescence to young adulthood. J Eat Disord 2022;10(1):84 [FREE Full text] [doi: 10.1186/s40337-022-00603-z] [Medline: 35725645]
25. Lanoye A, Brown KL, LaRose JG. The transition into young adulthood: a critical period for weight control. Curr Diab Rep 2017;17(11):114 [FREE Full text] [doi: 10.1007/s11892-017-0938-4] [Medline: 28971312]
26. Poobalan AS, Aucott LS, Clarke A, Smith WCS. Diet behaviour among young people in transition to adulthood (18-25 year olds): a mixed method study. Health Psychol Behav Med 2014;2(1):909-928 [FREE Full text] [doi: 10.1080/21642850.2014.931232] [Medline: 25750826]
27. Webb J, Baker A, Palmer T, Hall A, Ahlquist A, Darlow J, et al. The barriers and facilitators to physical activity in people with a musculoskeletal condition: a rapid review of reviews using the COM-B model to support intervention development. Public Health Pract (Oxf) 2022;3:100250 [FREE Full text] [doi: 10.1016/j.puhip.2022.100250] [Medline: 36101772]
28. Michie S, van Stralen MM, West R. The behaviour change wheel: A new method for characterising and designing behaviour change interventions. Implement Sci 2011;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
29. Atkins L, Francis J, Islam R, O'Connor D, Patey A, Ivers N, et al. A guide to using the theoretical domains framework of behaviour change to investigate implementation problems. Implement Sci 2017;12(1):77 [FREE Full text] [doi: 10.1186/s13012-017-0605-9] [Medline: 28637486]
30. Seiler J, Libby TE, Jackson E, Lingappa J, Evans W. Social media-based interventions for health behavior change in low- and middle-income countries: systematic Review. J Med Internet Res 2022;24(4):e31889 [FREE Full text] [doi: 10.2196/31889] [Medline: 35436220]
31. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
32. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
33. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
34. NHS UK. The Eatwell guide. NHS UK: NHS UK; 2026. URL: https://www.nhsinform.scot/healthy-living/food-and-nutrition/eating-well/eatwell-guide-how-to-eat-a-healthy-balanced-diet/ [accessed 2026-02-27]
35. Cane J, O'Connor D, Michie S. Validation of the theoretical domains framework for use in behaviour change and implementation research. Implement Sci 2012;7:37 [FREE Full text] [doi: 10.1186/1748-5908-7-37] [Medline: 22530986]
36. Sanzari CM, Gorrell S, Anderson LM, Reilly EE, Niemiec MA, Orloff NC, et al. The impact of social media use on body image and disordered eating behaviors: content matters more than duration of exposure. Eat Behav 2023;49:101722 [FREE Full text] [doi: 10.1016/j.eatbeh.2023.101722] [Medline: 37060807]
37. Friedman VJ, Wright CJC, Molenaar A, McCaffrey T, Brennan L, Lim MSC. The use of social media as a persuasive platform to facilitate nutrition and health behavior change in young adults: web-based conversation study. J Med Internet Res 2022;24(5):e28063 [FREE Full text] [doi: 10.2196/28063] [Medline: 35583920]
38. Kulandaivelu Y, Hamilton J, Banerjee A, Gruzd A, Patel B, Stinson J. Social media interventions for nutrition education among adolescents: scoping review. JMIR Pediatr Parent 2023;6:e36132 [FREE Full text] [doi: 10.2196/36132] [Medline: 37471119]

XSL•FO
RenderX

39. Chau MM, Burgermaster M, Mamykina L. The use of social media in nutrition interventions for adolescents and young adults-a systematic review. Int J Med Inform 2018;120:77-91 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.10.001] [Medline: 30409348]
40. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021;23(1):e17187 [FREE Full text] [doi: 10.2196/17187] [Medline: 33470931]
41. Denniss E, Lindberg R. Social media and the spread of misinformation: infectious and a threat to public health. Health Promot Int 2025;40(2):daaf023. [doi: 10.1093/heapro/daaf023] [Medline: 40159949]
42. Silva P, Araújo R, Lopes F, Ray S. Nutrition and food literacy: framing the challenges to health communication. Nutrients 2023;15(22):4708 [FREE Full text] [doi: 10.3390/nu15224708] [Medline: 38004102]
43. Gutiérrez-Ángel N, Sánchez-García JN, Mercader-Rubio I, García-Martín J, Brito-Costa S. Digital literacy in the university setting: a literature review of empirical studies between 2010 and 2021. Front Psychol 2022;13:896800 [FREE Full text] [doi: 10.3389/fpsyg.2022.896800] [Medline: 36176798]
44. Pérez-Torres V. Social media: a digital social mirror for identity development during adolescence. Curr Psychol 2024;43(26):22170-22180. [doi: 10.1007/s12144-024-05980-z]
45. Kington RS, Arnesen S, Chou WS, Curry SJ, Lazer D, Villarruel AM. Identifying credible sources of health information in social media: principles and attributes. NAM Perspect 2021;2021 [FREE Full text] [doi: 10.31478/202107a] [Medline: 34611600]
46. Dadaczynski K, Okan O, Messer M, Leung AYM, Rosário R, Darlington E, et al. Digital health literacy and web-based information-seeking behaviors of university students in Germany during the COVID-19 pandemic: cross-sectional survey study. J Med Internet Res 2021;23(1):e24097 [FREE Full text] [doi: 10.2196/24097] [Medline: 33395396]
47. Stifjell K, Sandanger TM, Wien C. Exploring online health information-seeking behavior among young adults: scoping review. J Med Internet Res 2025;27:e70379 [FREE Full text] [doi: 10.2196/70379] [Medline: 40925001]
48. Liu J, Zhang Y. Modeling health video consumption behaviors on social media: activities, challenges, and characteristics. Proc ACM Hum-Comput Interact 2024;8(CSCW1):1-28. [doi: 10.1145/3653699]
49. Kuutila M, Kiili C, Kupiainen R, Huusko E, Li J, Hosio S, et al. Revealing complexities when adult readers engage in the credibility evaluation of social media posts. Computers in Human Behavior 2024;151:108017 [FREE Full text] [doi: 10.1016/j.chb.2023.108017]
50. Portingale J, Girardin S, Liu S, Fuller-Tyszkiewicz M, Krug I. Daily bi-directional effects of women's social media-based appearance comparisons, body satisfaction, and disordered eating urges. J Eat Disord 2024;12(1):129 [FREE Full text] [doi: 10.1186/s40337-024-01096-8] [Medline: 39227958]
51. Kim JH, Kim J, Jang SY, Park EC. Association between watching eating broadcasts like mukbang and cookbang and generalized anxiety disorder among Korean adolescents. BMC Psychiatry 2024;24(1):536 [FREE Full text] [doi: 10.1186/s12888-024-05957-z] [Medline: 39080570]
52. Sanskriti S, Guglani I, Joshi S, Anjankar A. The spectrum of motivations behind watching mukbang videos and its health effects on its viewers: a Review. Cureus 2023;15(8):e44392 [FREE Full text] [doi: 10.7759/cureus.44392] [Medline: 37786568]
53. Milton A, Ajmani L, DeVito MA, Chancellor S. "I See Me Here": Mental Health Content, Community, and Algorithmic Curation on TikTok. New York, NY: Association for Computing Machinery; 2023 Presented at: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems: Association for Computing Machinery; 2023 April 23; Hamburg, Germany. [doi: 10.1145/3544548.3581489]
54. Savolainen L, Ruckenstein M. Dimensions of autonomy in human–algorithm relations. New Media & Society 2022;26(6):3472-3490. [doi: 10.1177/14614448221100802]
55. Scharlach R, Hallinan B. The value affordances of social media engagement features. Journal of Computer-Mediated Communication 2023;28(6):zmad040. [doi: 10.1093/jcmc/zmad040]
56. Bouton ME. Why behavior change is difficult to sustain. Prev Med 2014;68:29-36. [doi: 10.1016/j.ypmed.2014.06.010]
57. Veckalne R, Akhmatjonov S, Humbatova S, Hajiyev NG. From scroll to action: can social media influencers drive real behavioral change among their followers? Social Sciences 2025;14(5):253. [doi: 10.3390/socsci14050253]
58. Conner M, Norman P. Understanding the intention-behavior gap: the role of intention strength. Front Psychol 2022;13:923464 [FREE Full text] [doi: 10.3389/fpsyg.2022.923464] [Medline: 35992469]
59. Simeon R, Dewidar O, Trawin J, Duench S, Manson H, Pardo Pardo J, et al. Behavior change techniques included in reports of social media interventions for promoting health behaviors in adults: content analysis within a systematic review. J Med Internet Res 2020;22(6):e16002 [FREE Full text] [doi: 10.2196/16002] [Medline: 32525482]
60. Consavage Stanley K, Harrigan PB, Serrano EL, Kraak VI. A systematic scoping review of the literacy literature to develop a digital food and nutrition literacy model for low-income adults to make healthy choices in the online food retail ecosystem to reduce obesity risk. Obes Rev 2022;23(4):e13414 [FREE Full text] [doi: 10.1111/obr.13414] [Medline: 35092142]

Abbreviations

COM-B: Capability, Opportunity, Motivation and Behaviour model

XSL•FO
RenderX

COREQ: Consolidated Criteria for Reporting Qualitative Research
HFSS: high in fat, sugar, and salt
MRC: Medical Research Council
PPI: patient and public involvement
REDCap: Research Electronic Data Capture
SRD: Secure Research Drive
SSB: sugar-sweetened beverage
TDF: theoretical domains framework

Edited by A Stone; submitted 11.Dec.2025; peer-reviewed by AJ Umeano, N Döring; comments to author 13.Feb.2026; revised version received 13.Mar.2026; accepted 16.Mar.2026; published 07.Apr.2026.

Please cite as:
Tang H, Ahern AL, Spreckley M, Smith AD
Young Adults' Interactions With Food and Nutrition Content on Social Media and Implications for Intervention Design: Semistructured Interview Study
J Med Internet Res 2026;28:e89344
URL: https://www.jmir.org/2026/1/e89344
doi:10.2196/89344
PMID:

©Hao Tang, Amy L Ahern, Marie Spreckley, Andrea D Smith. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assistive Robotics for Healthy Aging: A Foundational Phenomenological Co-Design Exercise

Stephen Potter[1], PhD; Mark Hawley[1], PhD; Angela Higgins[2], MSc; Farshid Amirabdollahian[3], PhD; Mauro Dragone[4], PhD; Alessandro Di Nuovo[5], PhD; Praminda Caleb-Solly[2], PhD

[1]Sheffield Centre for Health and Related Research, University of Sheffield, Sheffield, United Kingdom

[2]School of Computer Science, University of Nottingham, Nottingham, United Kingdom

[3]Robotics Research Group, University of Hertfordshire, Hatfield, United Kingdom

[4]School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, United Kingdom

[5]School of Computing, Sheffield Hallam University, Sheffield, United Kingdom

Corresponding Author:
Stephen Potter, PhD
Sheffield Centre for Health and Related Research
University of Sheffield
Regent Court
30 Regent Street
Sheffield, S1 4DA
United Kingdom
Phone: 44 114 222 2000
Email: stephen.potter@sheffield.ac.uk

Abstract

Background: Assistive robotics for helping older people live well and stay independent has, to date, failed to fulfill its promise: there are few assistive robots in everyday use. In part, this failing can be attributed to inadequate or missing co-design activities that would ensure that these technologies and any services that incorporate them are developed with prospective end users, addressing their actual needs and wants, and not merely for them, and based on lazy assumptions about heterogeneous user groups.

Objective: This exercise aimed to address some of these limitations by taking a "phenomenological snapshot" of what it means to be an older person in the current sociotechnological context, and making this snapshot, along with the co-design materials developed, available to the wider assistive robotics community to provide solid foundational evidence for steering the development of assistive robotics in more productive directions.

Methods: Two rounds of co-design workshops have been conducted with older people and their caregivers, based on an innovative methodology that used personas and speculative designs to explore sensitive everyday difficulties faced by participants and highlight some of their general wishes for and concerns about assistive robotics. The data collected during the workshops were analyzed, and key themes were extracted.

Results: Analysis of the workshop data gives access to the lived experience of older people and their caregivers, and their opinions about domestic robotics and assistive technologies more generally. The findings are organized thematically as *everyday difficulties*, the daily problems faced by older people; *ideas for aging better*, older people's own suggestions for how their lives could be improved; and *living with technology*, their preferences and requirements for assistive robots, along with their concerns about what the introduction of robots might mean, both for themselves and for society more widely.

Conclusions: We believe that our findings provide solid foundational evidence for the development of assistive robotics for older people. We are in the process of disseminating these results through various channels to the wider assistive robotics community; ultimately, the success of our activities will be demonstrated only through the development of acceptable, useful, and viable assistive robotics for older people.

(J Med Internet Res 2026;28:e77179) doi:10.2196/77179

XSL•FO
RenderX

KEYWORDS

robotics; assistive technology; frail older adults; activities of daily living; independent living; human centered design, co-design; lived experience

Introduction

Emergence Network

Emergence is a UK-wide initiative created to foster innovation in the field of assistive robotics for older people, especially people living with frailty, and to facilitate the "emergence" of robots from research laboratories into service. The network is coordinated by a team of academic researchers and, in addition to other academic partners, members include health, social care, and housing providers, regulators, and robotics industry representatives, who provide clinical focus and business acumen to anchor its activities. The network was launched in early 2022 and is led by the University of Nottingham, with the support of coinvestigators from Heriot-Watt University, the University of Hertfordshire, the University of Sheffield, and Sheffield Hallam University.

More specifically, the objectives of Emergence are to promote the development of domestic assistive robotics that will improve the quality of life and independence of older people by enabling better self-management of frailty and age-related issues, by helping with activities of daily living, by assisting rehabilitation activities, and by supporting health care professionals to make tailored interventions to slow decline or to improve resilience. The primary impetus behind the establishment of the network was the observation that, despite the undoubted effort and resources that had been directed at the task, few assistive robots for older people are in general use (see, for example, the book by Wright [1]). Plausible factors that may have contributed to this include a poor appreciation of the needs and wants of older people, of the wider psychological, social, and clinical considerations, and of the challenges and practicalities of operation in noisy real-world domestic environments.

One fundamental pillar of the Emergence network is the proposition that this general lack of understanding of use contexts can be addressed through a rigorous and ongoing co-design exercise involving people from the target user groups, with the results of this exercise guiding assistive robotics development.

Frailty

Frailty is "a common clinical syndrome in older adults that carries an increased risk for poor health outcomes, including falls, incident disability, hospitalization, and mortality" [2]. It is characterized by the "loss of biological reserves across multiple organ systems," leading to clinical vulnerability to stressful episodes [3]. While it is not an inevitable consequence of aging, frailty is common among older people, and its extent and severity can vary significantly from individual to individual. Many older people who live with a health condition (or with multimorbidity) or with a disability will also be living with frailty: the one can contribute to or exacerbate the other. It can affect a person's physical and cognitive abilities, to the detriment of their wider health and well-being, and can limit their day-to-day functioning [4]. Frailty, when it does occur, is not a static condition: it can become worse, but there is evidence to suggest that, with the right interventions, it can be reversed to some extent [3].

When it comes to characterizing frailty, there are currently 2 dominant models [5]. The "(cumulative) deficit model" of frailty [6,7] assumes that frailty derives from an accumulation of "deficits" (which might be disabilities, health conditions, or psychosocial factors) that, individually and cumulatively, increase a person's vulnerability to stressful events. The second prevailing model is the "phenotype model" of frailty, in which an individual who manifests at least 3 of a set of 5 criteria for frailty (weight loss, low physical activity, exhaustion, slow walking speed, and physical weakness) is considered to be frail [4].

These models have their uses when considering whether an individual, in a health or care context, might require additional support or could benefit from some remedial intervention. However, they are less useful when it comes to understanding people's lived experience of frailty—and, by extension, for developing practical assistive technologies, including robotics, to support people living with the condition. The impact of deficits or of physical manifestations will vary from person to person, depending on their living circumstances, their tastes, their responsibilities and roles, and their hobbies and interests. Hence, while both the deficit model and the phenotype model can provide suggestions of where assistance might be needed, they also promote a narrow (and negative) view of people with frailty as deficient, a view that excludes a more rounded consideration of their lives that acknowledges their abilities (see the study by Lang et al [8]).

To take an example, clinical tests might reveal a loss of grip strength. The everyday impact of this on an individual might be felt in their inability to open jars of their favorite food (which might then have the knock-on effect of reducing the amount they are eating, leading to weight loss). Whereas the clinical model might suggest more-or-less generic remedial strength exercises—that is, focusing on the deficit or symptom and "treating" it—a more complete understanding of the individual's experience opens up the design space to suggest other, perhaps complementary, approaches encompassing the technological (a jar-opening robot), the environmental (redesigned food packaging), the social (a personalized meal delivery service), and so on. Any of these "interventions," or perhaps some combination of them, might be more acceptable to, appropriate—and, ultimately, effective—for the individual in question. The goal of the research reported in this paper is to provide a co-design counterbalance to the prevalent clinical models in the form of a snapshot of the lived experience of older people, one that encompasses their views of the digital world and, especially, robotics.

XSL•FO
RenderX

Assistive Robotics for Healthy Aging

Assistive robotics could help older people living with frailty, long-term health conditions, and disabilities to live more independent, more dignified, and more fulfilling lives. Moreover, services incorporating assistive robotics could play a role in plugging the gap in care provision that is being felt in many countries as their populations age, and they experience shortages of care staff. Many of the assistive robotics development projects undertaken have addressed one or other aspect of care provision for older people. These have typically taken the form either of, on the one hand, physically assistive robots (eg, for mobility support [9,10], exercise training [11] and help with eating [12]) or, on the other hand, of socially (or clinically) assistive robots—recent contributions have focused on stimulation [13], companionship [14,15], assessment tasks [16,17], monitoring [18-21], and remote visits or consultations [22]. It will be noted that both physically and socially assistive robots each tend to address only 1 or 2 aspects of frailty (care); as such, none of these robots can be said to be "for frailty" in its widest sense, if, indeed, such a thing is even possible. Notwithstanding these and other encouraging feasibility studies, undertaken both in laboratories and, on occasion, in people's homes, care homes, or clinical settings, this effort has yet to translate into anything approaching scale-up of production and widespread adoption. Despite years of research and development—and not inconsiderable financial investment—few assistive robots are actually deployed in the real-world to provide care or assistance to older adults [1].

At a technical level, it is undoubtedly extremely challenging to provide robotic assistance in a safe, sensitive, timely, responsive, respectful, reliable, and trustworthy manner to, with, and around ordinary people—who may have sensory, cognitive, or physical impairments—as they go about their everyday lives. However, the task has not been made any easier by basing design decisions on often unfounded or simplifying preconceptions of older people's needs, wants, and circumstances. In Emergence, we want to address this issue through the effective co-design of assistive robotics.

Co-Design of Assistive Robotics

Conventional design practices can result in "solutions" that leave end users with a sense of alienation and exclusion (especially where new or unfamiliar technologies are deployed), and the (entirely accurate) impression that the product or service has been designed "for them" and not "with them." It is likely that such products or services will not be adopted or quickly fall into disuse, unless their users have no alternative, when they will be used under duress and with little pleasure, and possibly with limited success. This failing becomes particularly significant in assistive technology contexts (and in health and care contexts more generally), where outcomes relate directly to the health and well-being of the end user. Co-design practices seek to address this failing by increasing the involvement of external stakeholders, and, in particular, end user representatives, in design and development processes, particularly during the earliest phases of design when, as is widely acknowledged, any incorrect decisions taken are significantly more difficult (and costly) to rectify later on. Co-design is a term given to a broad range of participatory techniques and methodologies whose underlying objective is to improve the acceptability and value of products and services by involving representative target users in their development [23]. During co-design, the role of the external actors can range from the relatively passive (such as critiquing alternative designs put to them) through more engaged modes (helping to define requirements or brainstorming potential solutions) to effectively becoming embedded in the design team and contributing to all stages of the development process.

There have been previous co-design exercises involving older people (for examples see other studies [22-27]). However, these exercises also underscore that co-design is not easy, with no single methodology; that access to the right stakeholders at the right time is not a given; that it is costly, in terms of both money and time; that participants find it difficult to shake preconceptions about robots and the roles they can play; and that, in the final analysis, it does not guarantee success.

Aim

Given what is at stake, however, none of these is a reason not to do co-design; rather, they are reasons to devote sufficient resources and care to try to do it better. The work reported in this paper aims to provide assistive robotics for older people with sturdy, co-designed foundations. Specifically, we have performed an extensive early-stage co-design activity with older people through which we hope to understand some of the current everyday problems facing people and their general requirements and attitudes toward hypothetical robotics solutions to some of those problems. We have undertaken this activity with the intention of collating the results and then sharing them as widely as possible among the assistive robotics community and, in this way, to provide a rich seam of foundational evidence for researchers and developers who otherwise lack the skills, wherewithal, or opportunities for engaging in early-stage co-design with older people. Ultimately, we intend to steer assistive robotics for older people in directions that are more likely to lead to the development of assistive robot services that provide genuine value for their users.

Methods

Emergence Co-Design Exercise

The objective of the co-design exercise was to take a "phenomenological snapshot" of frailty and aging, that is, one that positions the individual at the center of the enquiry [28] to complement the clinical models described above by capturing people's everyday experiences of frailty and aging. Moreover, the aim was to do this in such a way as to avoid some of the weaknesses of previous co-design in this field by expressly not focusing on any specific task or assuming as a basis any specific robot platform. This snapshot would encompass the physical, psychological, and social effects of aging, as well as the role that innovations (not necessarily technological) could play in mitigating negative effects and improving people's lives, and any constraints under which such innovations must operate if they are to be acceptable. As stated, the primary purpose of capturing such a snapshot is to provide guidance and support for the early stage (conceptual design) development of assistive technologies, specifically assistive robotics. However, the

XSL•FO
RenderX

snapshot could play a role beyond this, informing a variety of interventions and improvements, and educating those involved in the care and support of older people.

Moreover, it was imperative that the snapshot be accessible and communicable to a wide network of robotics and technology researchers, designers, and developers, and in such ways as to encourage the development of appropriate assistive robotics solutions. This, we hoped, would help to overcome the problems of a lack of expertise, experience, and resources for performing co-design, and of limited access to participants.

As mentioned above, there is no single methodology for co-design, although a number of different approaches have been suggested. Here, we adopted a selection of approaches and tools, intended to give us access to complementary information and in such a way as to hold the participants' interest and keep them engaged.

Co-Design Methodology

Overview

Workshops were run in 3 different locations (corresponding to the locations of academic partners in the Emergence network) in the United Kingdom. The workshops were "paired" at each location: an initial, "lived-experience" workshop was followed by a "speculative-critique" workshop.

Lived-Experience Workshops

The lived-experience workshops were designed to enquire into the needs and aspirations of older people living with frailty, the everyday realities of their lives, their living environments and social activities, and the benefits and frustrations of modern life, and, inevitably, of modern technologies (although the workshops were not designed to discuss technologies specifically). Participants were encouraged to identify any opportunities for assistive technologies, including robotics, to play a role in older people's lives.

To facilitate these workshops, we used personas as a starting point for group discussions. Personas are descriptions of fictitious individuals and have been widely used to motivate or contextualize co-design or innovation processes [29]. A total of 10 personas were developed for the Emergence workshops. We decided to develop these by drawing on our previous research [30] and existing data-based resources such as the CURE-Elderly-Personas set [31], which were adapted and extended to be more relevant to the United Kingdom context and to focus on the specific characteristics of the populations of interest to Emergence.

These personas depict, in easy-to-digest formats, the backgrounds, health status, and characteristics of fictitious older people, each of whom can be classed into 1 of 3 broad categories: prefrail but managing; vulnerable or living with mild frailty; or living with moderate-to-severe frailty. It was envisaged that having personas with different experiences of frailty would allow us to explore different intervention contexts and make comparative assessments of needs and wants. Before use, the personas were validated to be representative of people living with frailty by the Emergence steering groups, comprising health and social care professionals and people with lived experience of frailty. The group also reviewed the terminology used for possible infelicities or offence. Examples of the personas developed for and used in the workshops are shown in Figures 1 and 2. The personas have been made publicly available to allow their use in other co-design exercises.

XSL•FO
RenderX

Figure 1. "Michael," one of the personas developed for the lived-experience workshops.

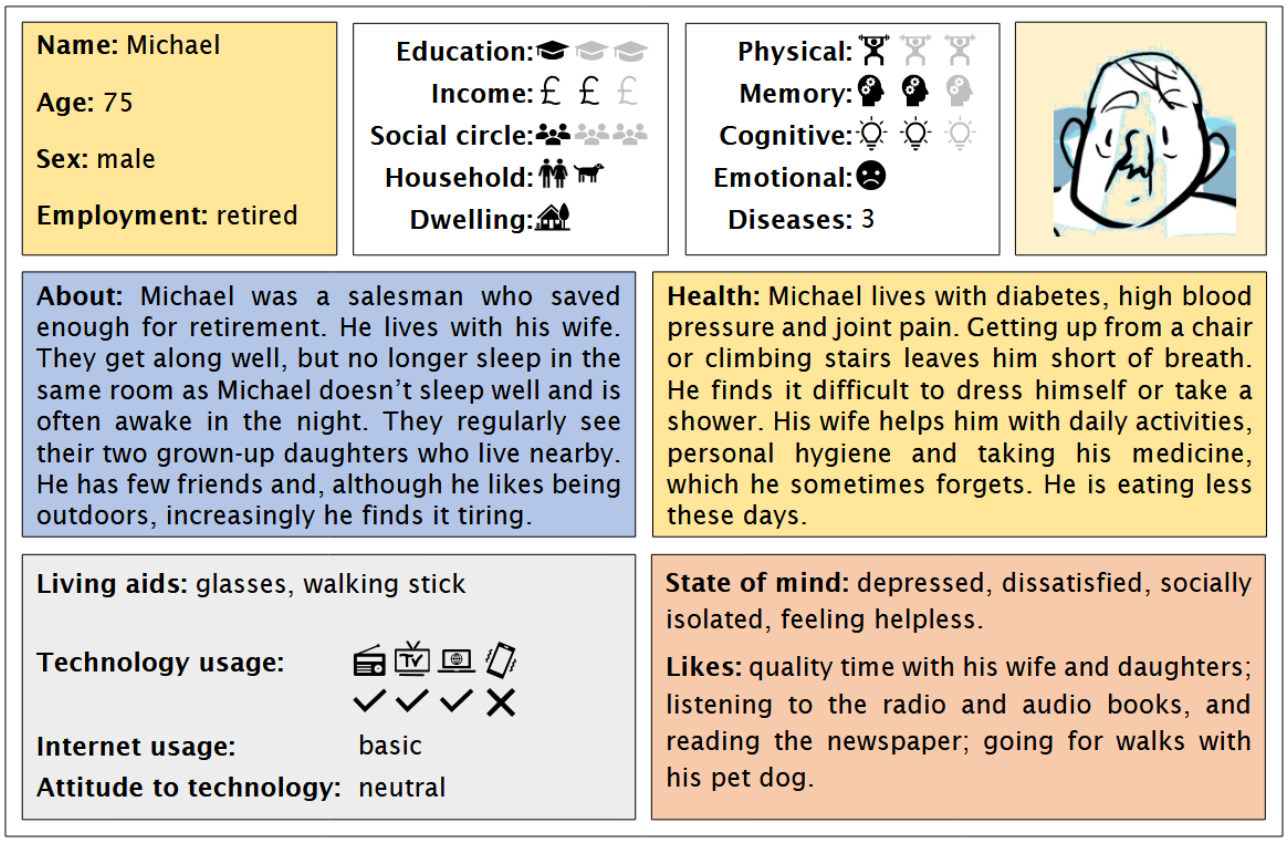

Figure 2. "Dorothy," one of the personas developed for the lived-experience workshops.

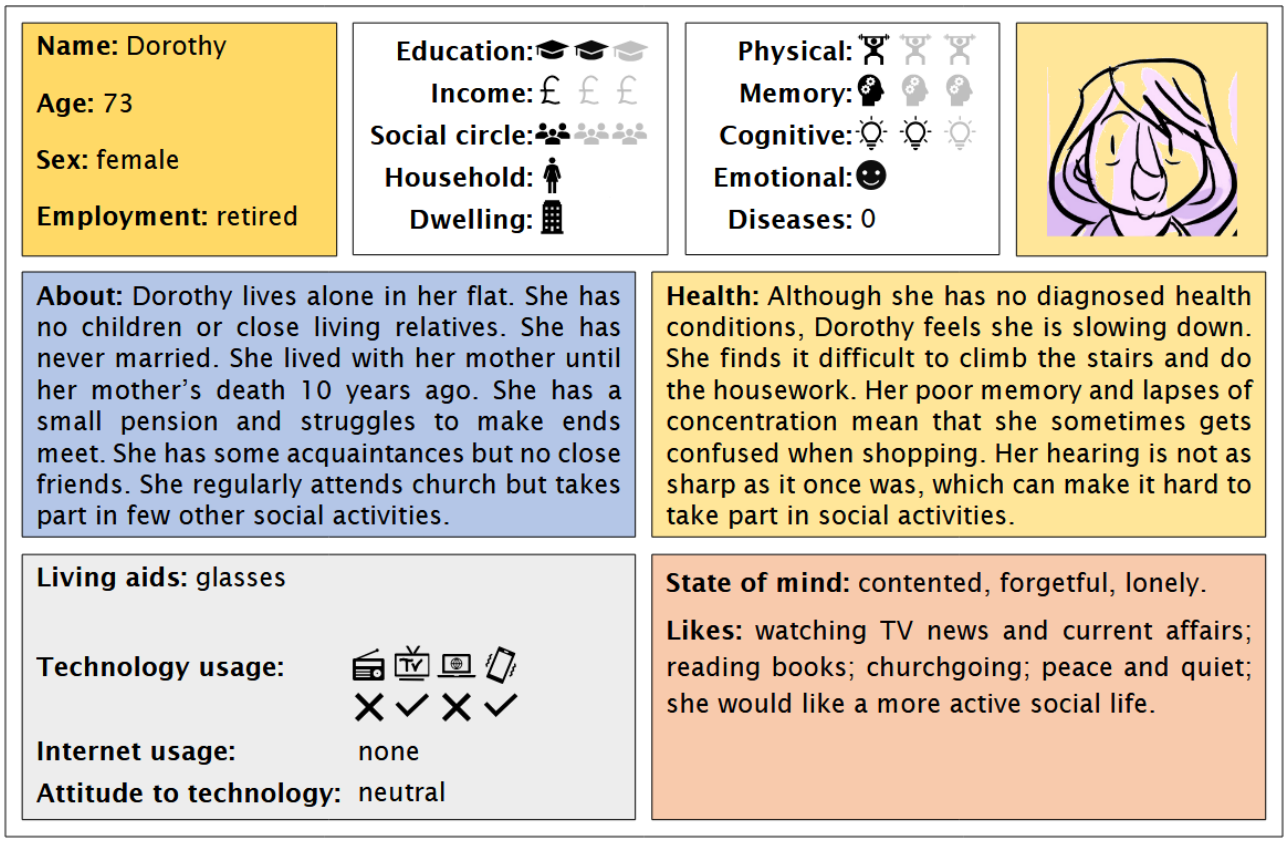

Personas can be used in different ways within design and innovation processes. Here, the primary role of the personas was to aid our participants within the co-design activities to "transfer" their own needs, wants, opinions, difficulties, and coping strategies onto fictional others, thereby avoiding potentially sensitive, embarrassing, or otherwise inhibiting

XSL•FO
RenderX

discussions of a personal nature, allowing us obliquely to broach frailty and other issues related to aging. More specifically, they would be used in the context of discussions of different "episodes" of the persona's typical daily routine:

- Getting up: including waking, toileting, washing and bathing, dressing, and medication.
- Mealtimes and snacks: including planning meals, food preparation, ordering food, utensil operation, cooking, eating, drinking, washing up, remembering to eat and drink enough, and maintaining a balanced diet.
- Household chores: including cleaning the house, heating the house, laundry, everyday repairs and maintenance, looking after pets, and household management (payment of bills, etc).
- Out and about: including getting around (walking, driving, and public transport), going to the hairdressers, going to the supermarket, going to the bank or post office, going to the doctor's, buying items, carrying things, and outdoor exercise.
- Socializing and pastimes: including home entertainment (television, music, internet, and puzzles), gardening, meeting friends and family, receiving guests, indoor exercise, and outside entertainment (cinema, concerts, bingo, book groups, walking groups, etc).
- Bedtime: including taking medicine, switching off and locking up, climbing stairs, getting to sleep, and getting up in the night.

Working in small groups consisting of 3 or 4 participants, each of which would be allocated 2 personas and 3 daily episodes, the lived-experience workshop participants were asked to consider what difficulties the episodes might present for those personas and what opportunities they can see for making the personas' lives easier. Following a break, the groups were then asked to consider each of the difficulties or opportunities and reach some consensus about: how frequently it occurs (daily, weekly, monthly...?); how prevalent it is (encountered by most, some, or just a few people?); and how consequential it is (does it have a large, medium, or small impact on people's ability to get by?). The answers to these questions would allow us to assess the significance of the difficulties and opportunities, and hence to prioritize them as objectives for assistive technologies.

The group discussions were facilitated by members of the research team who took notes as participants spoke, displaying these on whiteboards for immediate validation by participants. The discussions were audio-recorded for future verification and analysis.

In addition, the workshops were documented by a professional illustrator with experience in supporting academic workshops. Graphic facilitation can capture ideas in the form of easily digestible pictures, and can visualize the progression of and relationships between concepts discussed [32]. These illustrations would also represent the findings of a workshop in an easily accessible format, for reflection (and validation) during and after the workshop [33]. This was felt to be a particularly important aspect in this case, since one of our objectives was to communicate, in as readily accessible a manner as possible, the results to the wider assistive robotics development community. The illustrator was briefed as to the purpose and structure of the workshops, and then given free rein to move among the group discussions and capture any parts of the discussions that seemed to them both significant and susceptible to pictorial expression (they would also incorporate text, often quoting participants verbatim, into their illustrations). The illustrator attended and illustrated all the lived-experience workshops; they were unable to attend the subsequent speculative-critique workshops (but would supply the illustrations used to facilitate these workshops, as described below).

Speculative-Critique Workshops

The speculative-critique workshops would be structured around several "speculative designs" of assistive robots proposed to address the problems or grasp the opportunities identified by participants in the initial lived experience workshops.

These designs were in the form of a brief textual description and a sketch of the robot in action, helping to convey both the appearance of the robot and its use and operation. These were not intended to constitute designs for robots that would necessarily be developed; indeed, although the robots were intended to be realistic, in the sense of their functioning and behaviors being more-or-less feasible in the short term given current developments and directions in robotics research and development, they did not need to be feasible at the time of the workshops given the current state-of-the-art. Instead, they were to serve as "provocations" (elsewhere, such design provocations in the context of participatory design have been termed "provotypes" or "provocative (proto)types" [34]). As such, these workshops would constitute "speculative (co-)design" activities. Speculative design [35] is an activity during which a design proposal is presented not as a candidate for subsequent product development but as a means to elicit concerns or highlight issues that might otherwise remain latent, but which must be acknowledged if the design process is to be successful. This approach was felt to be particularly appropriate for a field such as robotics, about which many people possess preconceived notions. Moreover, discussing potential applications of any new technology without concrete examples is difficult: the speculative designs would constitute a basis for grounding discussions around assistive robotics in something approaching reality. Attempts to ground co-design discussions around (prototype applications of) real, existing robots are common; however, the obvious limitation of this approach is that the robots in question will have already been developed to some degree and, even when they have been developed with assistive applications in mind, this is likely to have involved certain assumptions on the part of the designers. Moreover, they will almost certainly have technical shortcomings. Consequently, it is difficult to move the discussion beyond the specific limitations of that particular robot or envisaged application so as to capture more general requirements for assistive robotics. We believe that the speculative design approach gives participants more freedom to move beyond the particular and express more general opinions about future applications of technology [36].

The development of the speculative designs required a rapid analysis by members of the research team (SP and MH) of the

XSL•FO
RenderX

content of the lived-experience workshops to identify candidate domestic tasks with which assistive robotics might help; a total of 6 such "robots" were identified:

- Motibot ("the motivational well-being and exercise robot")
- Foodee ("your personal cooking assistant and dietary advisor")
- EasyUp (a mobility assistance robot "for all life's ups and downs")
- AutoReach (a cleaning robot to "keep those hard-to-reach places spotless")
- RoPet (a robot pet that is "your new faithful friend")
- Toilittle (a discreet toileting robot that's "there when most you need it")

Each of these was summarized in words in terms of the problem or opportunity that the robot sets out to address (the "why" of the robot), its projected functionalities (the "what"), and how the robot would operate or be operated (the "how"). To aid in communicating the designs to workshop participants, these textual descriptions then became the brief for the illustrator to work up a sketch of the robot "in action," working with the research team to imagine what form the robot could take. In undertaking the brief, across the 6 applications, the illustrator was asked to draw robots with a range of embodiments, sizes, interfaces, and movement styles to elicit responses to different design options. Figures 3 and 4 give examples of 2 of these speculative designs in the format in which they were presented at the workshops.

Figure 3. Motibot ("the motivational well-being and exercise robot"): one of the speculative designs developed for critique.

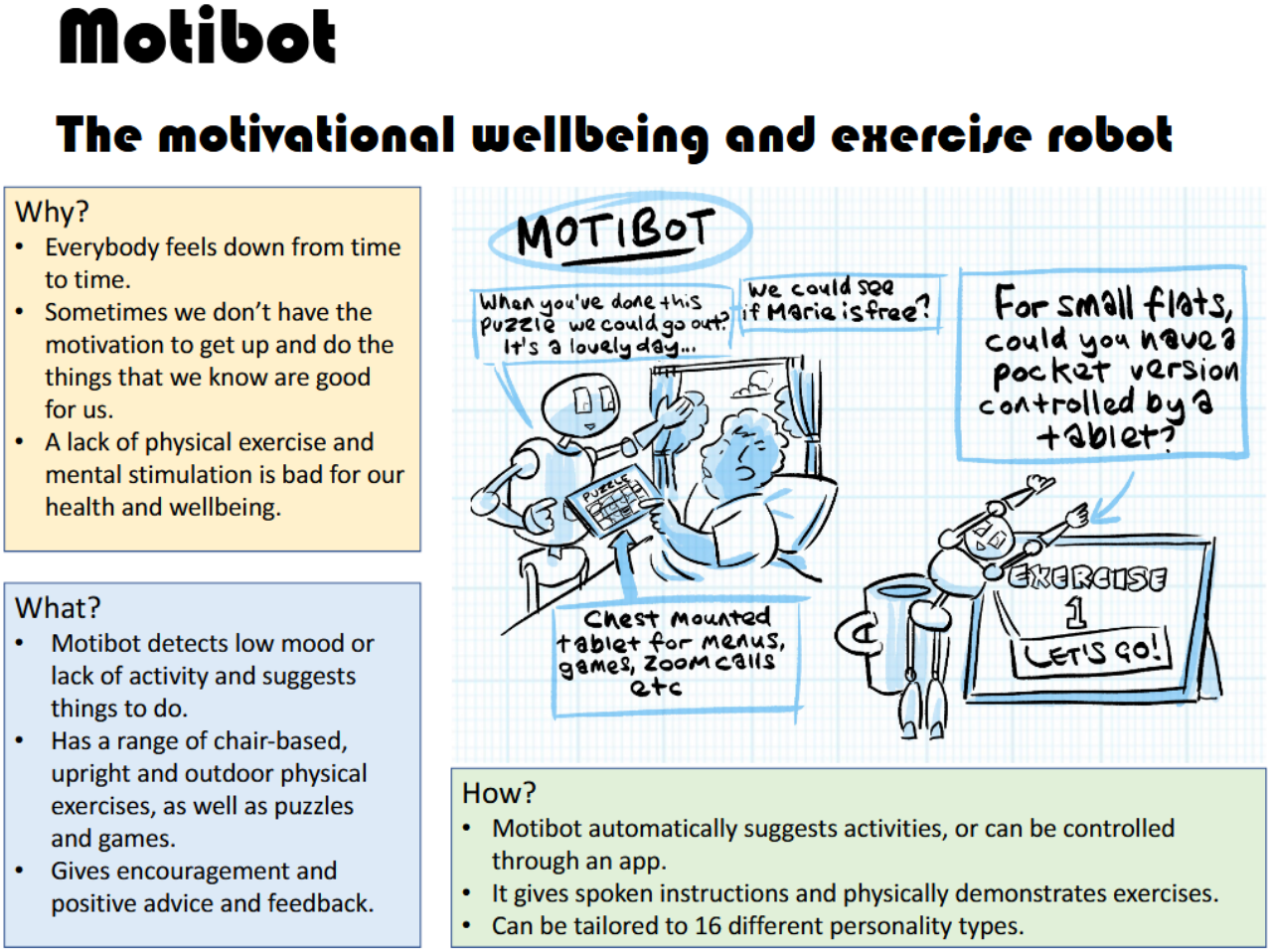

XSL•FO
RenderX

Figure 4. RoPet ("meet your new faithful friend"): one of the speculative designs developed for critique.

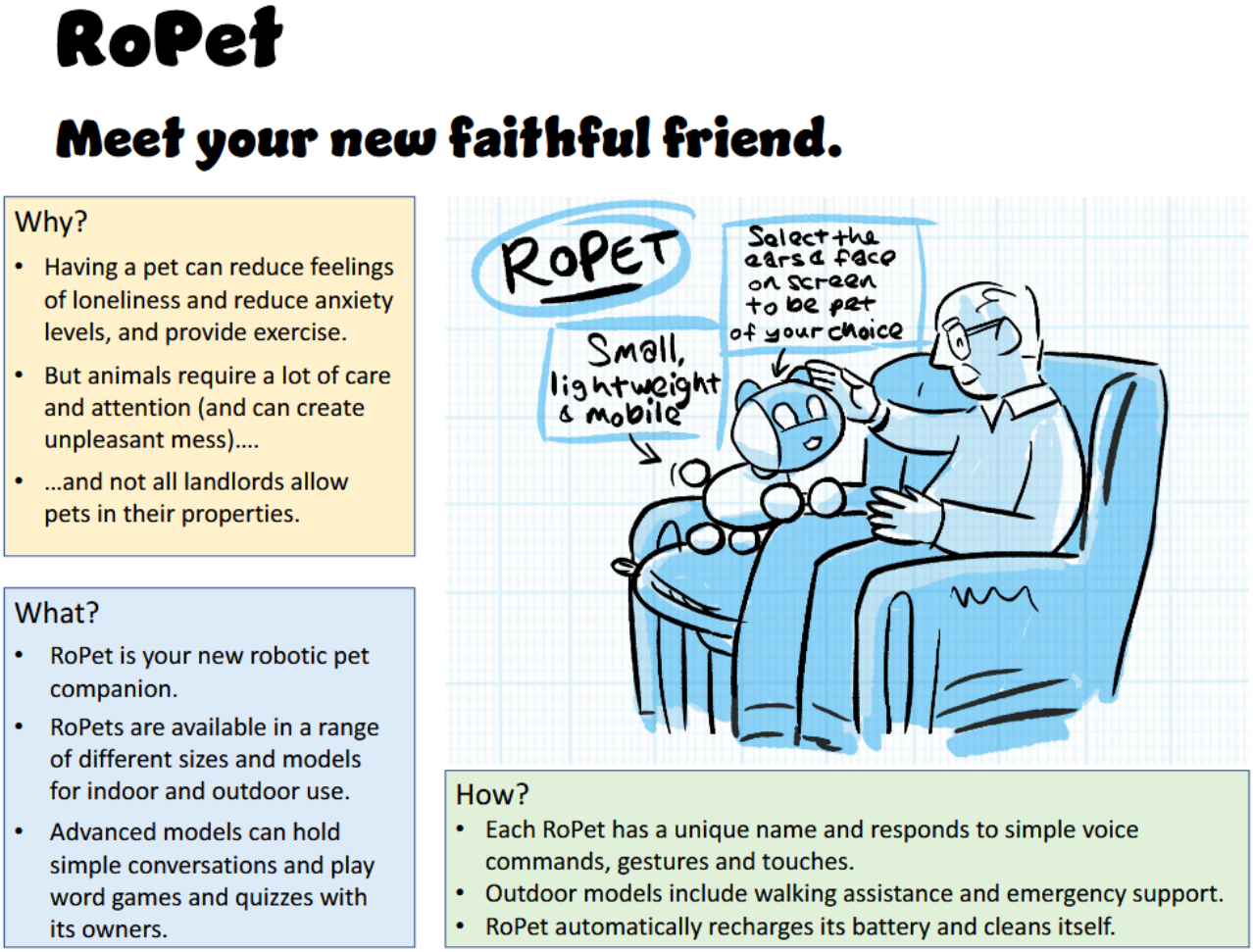

For the speculative-critique workshops, the participants were, once again, allocated to small groups, and asked to consider each of the speculative designs in turn, it first having been emphasized that the designs portray imagined, rather than real, robots. The participants were asked whether they would like to have the robot, and then asked for the reasons for their decision. Participants were encouraged to ask for additional information if this would help; facilitators were encouraged to extemporize answers. Using different-colored notepads, the facilitators noted separately positive and negative arguments, as well as quoting verbatim the words of participants where these seemed particularly pertinent. If at least 1 member of the group said they would accept the robot, the group was then asked about the following features of the robot:

- Appearance: what was liked and disliked about it, what size and color it should be, what materials it should be made of, and so on.
- Control: the extent and nature of the control that the end user and their caregivers should have over the robot, appropriate modalities for controlling the robot (voice, touchscreen, smartphone app, etc), and so on.
- Performance: its operating speed, acceptable noise levels, weight, communication, quality of movement, behavior both when in use and when not in use, and so on.
- Practicalities: how it would fit into people's lives and homes, whether people would be comfortable operating and maintaining the robot, the sort of training and assistance they felt they might need, and so on.
- Concerns: any safety or security issues, whether the robot seemed trustworthy, who would be trusted to develop, supply, run, maintain the robot, and so on.

Once again, facilitators would note both positive and negative comments: the former would constitute design requirements ("should-haves" or "nice-to-haves") while the latter constitute design constraints ("should-not-haves").

Participants

Older people do not always recognize themselves as living with frailty and can resist being labelled as "frail": the term has certain negative connotations, and not all people clinically classified as such would recognize (or welcome) the attribution [37]. Moreover, its severity can vary dramatically, from those who are moribund and almost completely dependent on others to those whose coping strategies are so successful that others——and maybe even they themselves—might not consider them to be living with frailty. For these reasons, we avoided using the term "frailty" when engaging with older people and their caregivers in these co-design activities. Moreover, we chose not to recruit through health services because we would then be accessing only those who, for one reason or another, had come to the attention of those services and received an "official" diagnosis. Instead, we focused on recruiting through

XSL•FO
RenderX

independent-living housing providers managed by Emergence partners and located conveniently close to the Emergence university research centers. People aged 55 years and older were considered "older people," adopting the threshold age for residency used by one of the housing providers. Even if not living with frailty themselves, most people living in these contexts would almost certainly know of people among their friends and neighbors who are living with frailty. In addition, relatives and friends who play an active role in supporting older people, along with other caregivers, were invited to participate, especially where they could support older people to participate by helping with transport and mobility.

Participants were asked to attend both the lived-experience and the subsequent speculative-critique workshops at their locations. However, for various reasons, some participants were unable to attend both workshops. As far as possible, the workshops took place in locations that were convenient for participants to attend and which, as far as possible, were close to their "natural environments" (such as communal lounges or university collaborative spaces) and so conducive to eliciting their experiences, while also being appropriate for group discussions. Likewise, workshops were scheduled for times considered most convenient for participants. Each workshop typically lasted around 4 hours, including breaks and refreshments. There was an interval of approximately 2 weeks between the lived-experience workshop and the corresponding speculative-critique workshop, except for the last pair of workshops, which, due to scheduling pressures, were held as morning and afternoon sessions on the same day. Before the workshop, participants were asked to complete a consent form and a demographics questionnaire. At the end of the workshop, they were invited to give feedback about the nature, structure, and conduct of the workshop, and were each given a shopping voucher as an honorarium to acknowledge their contribution.

Data Analysis Method

The audio recordings of the workshops were transcribed by professional transcribers; the transcriptions, alongside the notes collected during the workshops, were then subjected to a thematic analysis by two of the research team, one with experience of assistive technology development (SP), the other with a background in industrial design and robotics (AH). The analysis broadly followed the Framework Method [38] as elaborated by Gale et al [39]. A phenomenological approach was adopted for analysis, with the analysts bracketing their own experiences and conceptions to reduce bias and focus on the content of the participants' contributions, paying particular attention to their experiences and the meanings that they attached to these. Specifically, we adopted an interpretive phenomenological analysis, foregrounding the participants' experiences and perceptions, while recognizing that the researchers play an active role in interpreting these [40,41]. This analysis was both deductive and inductive. Given the aims of the exercise, we had a particular deductive focus on the lived experience of aging and how this relates to opportunities for and constraints upon assistive robotics, as dictated by the structure of the workshops. However, within these broad topics, we were open to themes that were not explicitly foreseen as topics for discussion, and which arose naturally during the workshops. Codes were first inductively identified from the participants' utterances and then checked against the notes taken by the facilitators. These were then clustered deductively into higher categories which broadly mirror the structure of the workshops, namely "difficulties with activities of daily life," "opportunities for facilitating activities of daily life," "assistive robotics design requirements," and "assistive robotics design constraints." The analysts worked independently at first, then collaborated to merge codes, resolving any discrepancies through discussion. Several iterations of analysis, during which codes and categories were amended, were performed until a reasonably stable analytical framework emerged, with the first 2 themes emended to "everyday difficulties" and "ideas for aging better," respectively, and the last 2 merged into the more general "living with technology" category. The transcription and analysis of the data were carried out using the Lumivero NVivo qualitative analysis software package. Although the notes taken were displayed to participants for "real-time" validation during the workshops, it was unfortunately not feasible to ensure dependability by reconvening the participants and presenting the analysis back to them for their approval or correction of the framework.

Ethical Considerations

This study was approved by the Ethics Committee of the University of Nottingham (CS-2021-R40). All participants provided written informed consent. Privacy and confidentiality were ensured by collecting no personally identifying data. Participation in this study was voluntary; each participant was given a shopping voucher (worth approximately US $30) as an honorarium.

Results

Overview

The older people participants in the workshops were aged from 55 to 96 years. An overview of participants for each workshop can be found in Table 1.

The examples of the illustrations produced during the workshops, shown in Figures 5 and 6, give some idea of the content of the discussions.

The analysis of the workshop outputs led to the formation of 3 broad, high-level categories, namely "everyday difficulties," "ideas for aging better," and "living with technology."

XSL•FO
RenderX

Table 1. Participant information for each of the co-design workshops.

Workshop type	Location	Participants		Setting
		Older people	Relatives, friends, and caregivers	
L[a]	A	8 (4 F[b], 4 M[c])	1 (1 F)	Lounge in residential scheme
L	B	8 (5 F, 3 M)	5 (3 F, 2 M)	University collaborative space
L	C	4 (2 F, 2 M)	4 (2 F, 2 M)	University collaborative space
S[d]	A	5 (3 F, 2 M)	2 (2 F)	Lounge in residential scheme
S	B	8 (5 F, 3 M)	3 (2 F, 1 M)	University collaborative space
S	C	4 (2 F, 2 M)	4 (2 F, 2 M)	University collaborative space

[a]L: lived experience.

[b]F: female.

[c]M: male.

[d]S: speculative critique.

Figure 5. Tableau illustration of discussions during a lived-experience workshop (location A).

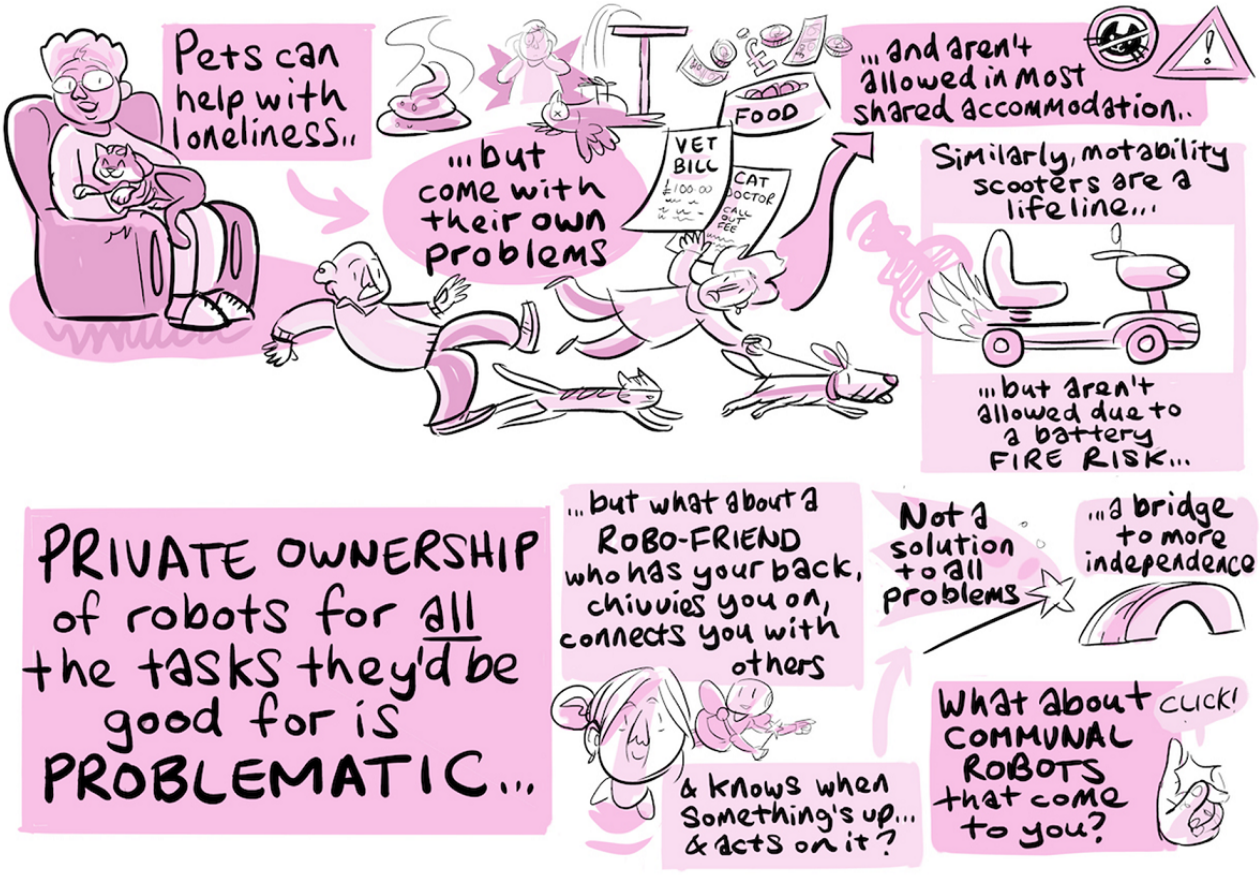

XSL•FO
RenderX

Figure 6. Tableau illustration of discussions during a lived-experience workshop (location C).

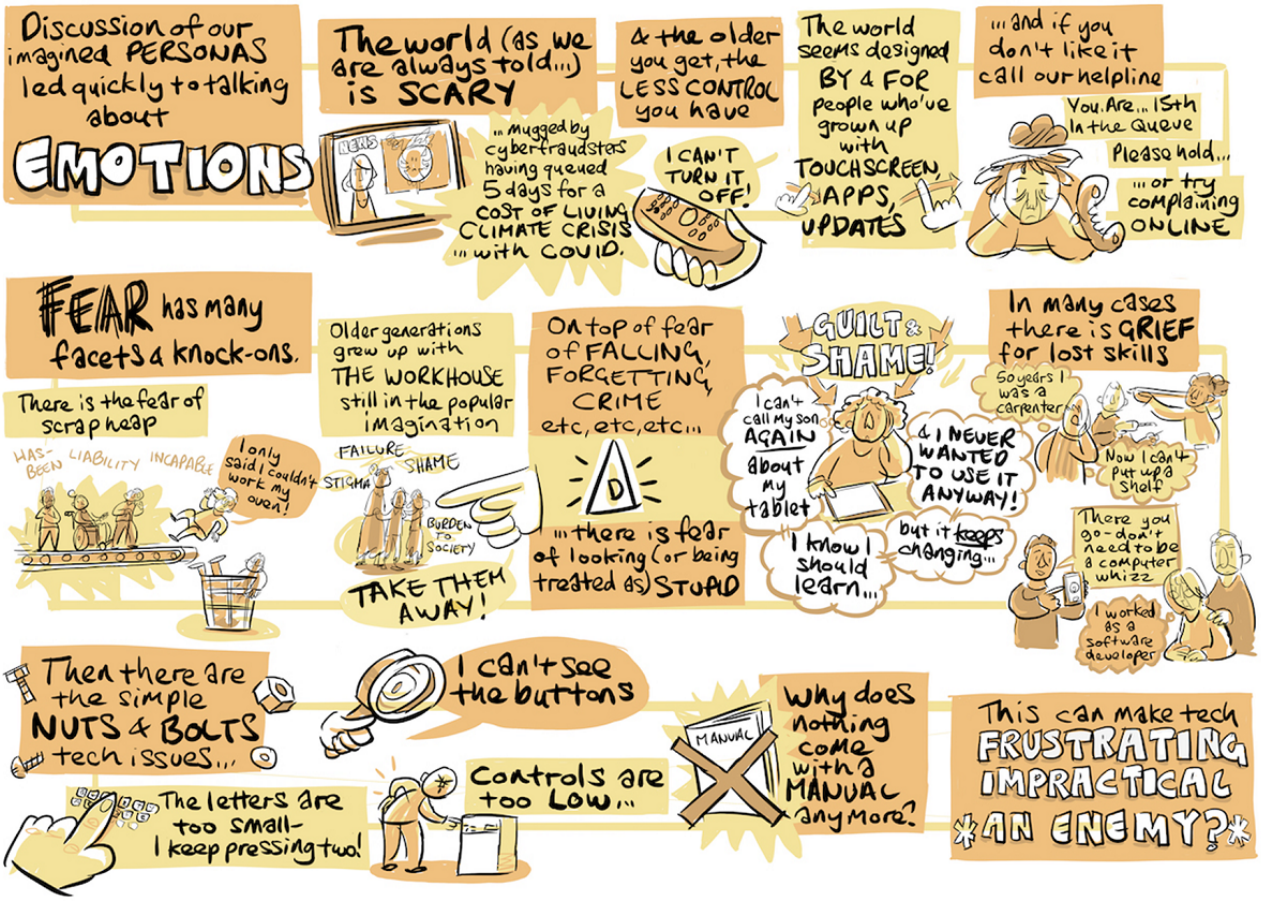

Everyday Difficulties

The most prevalent codes in the category of everyday difficulties are physical difficulties and psychological difficulties, with other difficulties being less prominent, if not less significant for participants.

- Physical difficulties: as was intended through the format of the lived-experience workshops, physical difficulties were often expressed in terms related to activities of daily living or situated in contexts. For example, participants mentioned the difficulties they have doing the housework (reaching high places, cleaning windows, and making the bed), with personal care (difficulties washing themselves and getting dressed), toileting (incontinence), mobility (walking up hills and stairs, and getting on and off buses), and, in general terms (not exercising enough).
- Psychological difficulties: almost equally prevalent were psychological difficulties. That these are shared by many participants might have been expected; what was more surprising was the extent to which participants were prepared to discuss their own psychological difficulties (and in the process setting aside the fictitious problems faced by the personas). Some of these were related to specific activities; others were more generalized states of mind or mood. Among other things, participants mentioned anxiety (about their own situations, but also more generalized, about the state of the world), social isolation and loneliness, lack of motivation and purpose, depression, and fear (of falling, of going out at night, when answering the door, and of assistance not being available when most needed).
- Cognitive difficulties: principally these concern remembering (appointments, where things are, to take medication, and to recharge and use technology, such as a smartwatch intended to help monitor activity levels).
- Living environment difficulties: these included having limited opportunities to socialize, poor relations with neighbors, lack of convenient public transport, and being unable to keep pets or mobility scooters in shared accommodation (the latter due to size and safety concerns, possibly having implications for assistive robotics).
- Burden of disease: for the workshop participants, this burden included remembering to take medication (overlapping with cognitive difficulties) and maintaining medication routines, the detrimental side effects of medication, and having to care (in an unpaid capacity) for a spouse living with more severe frailty.
- Time pressures: the time required to perform certain activities dissuaded some participants from doing potentially beneficial things such as buying food and preparing meals, taking care of pets, and going out to exercise or socialize.
- Financial difficulties: the cost of living, and specifically the cost of gas and electricity, was mentioned by several participants (who worried that they would be unable to afford to operate robots or other assistive technologies).

XSL•FO
RenderX

Ideas for Aging Better

Participants also brought to the workshops their own ideas, some based on their own experiences of what works for them, for how people might age better. Some of these concerned living and lifestyle (for instance, exercises, stimulating games, even pet ownership as a motivation for staying engaged and active) and social ideas (conversation and group activities of various sorts). However, the most prevalent category was technological ideas, no doubt influenced by the wider context of the workshops and suggested by the problems faced by the personas (and by the participants themselves); these were further divided into:

- Technology for physical assistance: the category for which there were most suggestions, which included the feasible (and in some cases already available), such as smart speakers for home automation, robot vacuum cleaners, and teasmaids; adaptations of existing technology to the domestic sphere (escalators and self-cleaning toilets, which a participant had seen at an airport); voice-activated showers and personalized drinks and snacks dispensers; and more ambitious ideas involving robots: robots for helping to do the shopping, for cleaning high places, and for cleaning up after pets, automated kitchen surfaces and food packaging openers, and self-making beds.
- Technology for psychological assistance: including technology that tells you who is knocking at the front door, assuring that it is safe to open the door; robots for providing motivation, for recipe suggestions, for cultivating positive attitudes and providing "positive affirmations," and robotic cats for companionship.
- Technology for cognitive assistance: including devices for providing reminders (for doctors' appointments or social events, for medicines, food, and drink), and smartwatch navigation aids.
- Technology for social assistance: several participants noted that their smartphones where helping them to stay connected to friends and family, especially through multimodal (speech, video, and chat message) calls; another suggestion was for an app or similar that could allow caregivers to better communicate and coordinate their efforts; and shared smart facilities, such as "robot washing machines" could help to avoid conflicts in residential schemes and similar multihabitant environments.
- Technology for health monitoring: participants suggested that it would be beneficial to have technology that could detect if medication had actually been taken, for measuring nutrients in the body, and for raising the alarm in the case of a fall; and to have an all-purpose "well-being robot."

Some of the ideas above would find expression in the hypothetical robot designs developed for the speculative critique workshops.

Living With Technology

The final major category, living with technology, encompasses mainly (but not exclusively) codes identified during analysis of the speculative critique workshops. These cover general preferences (in turn covering aspects such as appropriate interface modalities several people expressed a desire—indeed, an expectation—that the robots would be voice-controlled), size (large enough to do the job, but small enough to fit into people's homes and lives) and appearance (nonthreatening, with some opting for humanoid and others for "cute" robots), but also ideas of "personalization" (sometimes dynamic or autonomous) of the robot to the user's (changing) needs, abilities and routines, with robot autonomy extending to regular maintenance tasks such as charging or cleaning itself, and even consideration of financial models for provision (such time-sharing robots with others).

There was some overlap here with general requirements for robots, with a requirement being identified when a participant used a modal verb expressing necessity for some feature or aspect (that a robot "needs" or "must have" it). Requirements refer to physical qualities (robustness and stability), maintenance (cleanable or even self-cleaning), the need for adequate training and practice in using the robot, and, on a more challenging level, the need for a robot to be "self-aware" enough to be able to explain what it is doing.

However, the code with the most items in this category is concerns about technology, which are, almost exclusively, concerns about robotics. These are wide-ranging, revealing the complexity of people's lives and how they imagine robots might disrupt those. The following subcodes give some flavor of the range of concerns (note that these codes are not mutually exclusive):

- Safety: perhaps unsurprisingly, the major source of concern for the participants is safety. Difficulties here were further coded as surrounding the physical safety of users and other beings in the vicinity of robots, psychological safety (specifically if the appearance of a robot might scare or upset the user or others, manipulate their sentiments, or increase social isolation), potentially detrimental side effects (such as discouraging activity or suggesting the wrong sort of exercise), and digital safety (digital surveillance, privacy, and unauthorized access to any data collected). A number of concerns here are related to what might be termed uncertain hazards, where the participants could envisage situations during the use of the robots in which their safety could be imperiled if appropriate safeguards are not in place. For instance, what happens if a robot intended to physically assist people up and down stairs, or a flying drone-type cleaning robot were to malfunction midoperation?
- Compatibility: participants raised practical concerns about how robots would fit into their everyday lives, including aspects such as social compatibility (for instance, will the neighbors be inconvenienced by a robot? How would it interact with pets?), physical compatibility (Where will the robot be stored? Will it require modifications to the home? Will it damage possessions?), and technical compatibility (Will it require broadband connectivity?).
- Usefulness-desirability-inclusivity: general questions were raised about the usefulness of the speculative robots, including whether there were simpler or cheaper ways of achieving the same ends using existing technology (such as tablet computers and smart speakers, and nondigital alternatives), about the longevity of the robots (would they end up gathering dust in the cupboard alongside the foot spa?), about the need for sensitive, nonstigmatizing, design

XSL•FO
RenderX

of robots (and other assistive technologies), and, indeed, whether there was any real need for robots at all ("Is this a generational thing?" asked a participant; and another: "Am I going to feel the same about robots as I do about other tech? Too much hassle!")
- Burden of use: the additional overheads of looking after a robot are another major source of concern. These overheads encompass cleaning, charging, and servicing robots, the training required, and the cognitive demands of remembering how to use the robot. To quote a participant, "We don't want to look after the robot–we want the robot to look after us!"
- Maintenance and service model: a related concern is the service model of which the robot constitutes a technical component (and which is often overlooked in robot development): who programs, sets up, and personalizes the robot? Do robots act alone or in concert with human assistants? Who is responsible for regular maintenance and repairing faults?
- Control: concerns were raised about how control is exercised over robots, and who is in control. Some of these relate to the immediate operation of the robot, especially where the behaviors are complicated or potentially hazardous (and, in the words of a participant, "Can you tell it to stop?"), and others to the underlying operation of the robot, such as who determines what type and amount of motivation or exercise a robot should suggest.
- Financial costs: the economics of assistive robotics were also raised: participants raised the costs of purchase, running, and maintaining a robot, but also of modifying their homes, if necessary. Some participants asked about the comparative costs, whether that be compared with a human caregiver providing the same services or with a conventional refurbishment of their homes to make them more "age-friendly."
- Techno-psychosocial effects: here we see what might be called the techno-psychosocial effects of aging. These include the fear of being deemed "obsolete" or a burden on society, in no small part due to a (real or perceived) inability to move with the times and adapt to new technologies, as well as more "conventional" fears of falling, or forgetting, or crime and so on; the guilt and shame of not being able to adapt to these new technologies (alongside a feeling of grief for those skills that by-and-by are lost as we age); the stigma that sometimes accompanies the use of assistive technologies; the bottled-up frustration and occasional releases of rage that come from dealing with an increasingly digital world; and the contribution all these factors make to a vicious circle of stress, anxiety, and depression, with accompanying relationship problems, lack of motivation, poor diet, and social isolation.
- Digital ethics issues: in addition to the ethical issues that run through many of the difficulties noted above, here we also encounter direct concerns around the surveillance that users of data-collecting robots might find themselves subject to, users' privacy, and the security of their data.
- Social and care implications: finally, several participants raised concerns about the wider impact of assistive robotics: its potential negative effect on the role and jobs of human caregivers and home-helps, and the worry that robot care might replace human care, with the assumption that a degradation in the quality of that care would necessarily follow. On the other hand, there are worries too about inclusivity, access to care services, and the place of those who are unable to adapt to—or who choose to opt out of—the brave, new digital world.

Discussion

Principal Findings

We aimed to gain a better understanding of the everyday lives of older people, especially with reference to the effects of aging and frailty, and to gauge their opinions and concerns about assistive robotics. Through a wide-ranging co-design exercise with older people and their caregivers, we have gained a compelling snapshot of older people's real lives. As might be expected, given the structure of the workshops, the emphasis lies on the problems that people face daily and on their concerns about a world where change seems to have accelerated, sweeping away old certainties, and where, increasingly, services are delivered using digital technologies about which they have little say and less control [42]. In this context, the work reported here is an attempt to redress this balance and to return to older people some agency in the development of assistive technologies. We have also highlighted general concerns that older people have about assistive robotics. While older people and their caregivers are open to, and in some cases, enthusiastic about, the role that assistive robotics could play in their lives and those of others, they raise real concerns about, among other things, safety and control, the burden of use, ethics and unintended side effects, and the financial and social costs of introducing robots into their lives.

The work reported here differs from previous assistive robotics co-design activities in several key aspects. First, in its scope, which embraced a wide range of experiences of growing older, with a focus on the individual rather than on some specific health condition or difficulty, as much previous work has tended to do (eg, helping mitigate dementia [43] or supporting mobility [22,26]). Second, the exercise was technology-neutral in that we did not structure the workshops around some particular robot or robotic platform and tried to stay impartial in the question of whether assistive robotics could be beneficial: to adopt a priori a particular robotic platform is to make design decisions, including the fundamental decision that a robot is a valid "solution" to an ill-defined problem [27]. Finally, the exercise was undertaken not as a step in a specific assistive-robotics development pathway, but as a service to the wider assistive robotics community exploiting the Emergence network's resources, skills, and access to engage with potential users, and as such, we have placed special emphasis on communicating the results. In methodological terms, rather than focusing on a particular problem or solution (as is often done in co-design activities), we chose to adopt a general approach and investigate the lives of older people more broadly, and without trying to reach any firm consensus about where to focus subsequent development efforts or trying to reconcile the at times contradictory opinions of participants. We hope that, by

XSL•FO
RenderX

explaining our co-design approach along with the results, assistive technology developers will be able to sift the evidence and draw sound, rational conclusions.

Communication and Dissemination of Results

One key aspect of the exercise reported here is the communication and dissemination of results: clearly, this is essential if the findings are to influence the development of assistive robotics. One novel aspect here is that we realized that the results of the graphic illustrator's work to document the workshops could come to play an innovative role in our dissemination activities. The illustrator produced a total of 11 tableaux from the 3 lived experience workshops (Figures 5 and 6 each show 1 of these tableaux). Each tableau, somewhat like a comic book page, consists of a number of thematically related vignettes illustrating the ideas, opinions, or concerns voiced by participants. It was realized that these vignettes could form powerful communication and design aids as self-contained glimpses into the lives of the participants. Accordingly, they were isolated and printed as a pack of playing card-sized "empathy cards" (Figure 7), complete with suggestions for workshop "games" that encourage players to explore the different perspectives, concerns, and opportunities the cards contain and to consider how these relate to their own projects.

Figure 7. Some of the empathy cards generated from tableaux vignettes.



Along with the empathy cards, the results of the analysis have formed the basis of a series of workshops, sandpits, and project funding calls, and have provided content for a multidisciplinary summer school for budding assistive robotics developers. Through these activities, we have tried to reorient the development of assistive robotics for older people in a more empathetic direction and, indeed, to foment a new generation of more empathetic roboticists.

In addition, we are in the process of making all our workshop materials and the analysis results open and accessible in digital formats to the wider robotics and assistive technologies communities. This content includes the personas, the speculative robotics designs, the results of the analysis, and the empathy cards (which are also available as a physical pack of cards). The process and content of communicating outcomes in an accessible, understandable, and actionable manner is something that is not always given due prominence when discussing co-design exercises. This paper constitutes an element of our communication strategy, but it is by no means the culmination of our labors.

At the time of writing, it is too soon to draw any firm conclusions about the success of these endeavors: the development of assistive technologies and, in particular, of assistive robotics is a difficult and slow process which, typically, extends over several years. Only time will tell whether our efforts have contributed to bringing about the intended effect, and we see assistive robots emerge from the laboratory to help older people in their everyday lives.

Limitations

We recognize a number of limitations of this co-design exercise. All co-design practice is subject to bias, and this is no exception. The structure of the workshops, the development and choice of the material that provided the stimuli for them, the data collection by facilitators during the workshops, and the analysis of that data are all prey to the biases and preconceived notions of the researchers. The co-design approach adopted was, in some sense, an attempt to counter these biases: by not focusing on any particular everyday problem or given robotic platform, we have attempted to give participants the freedom to steer the discussions. In addition, the multidisciplinary nature of the

XSL•FO
RenderX

research team (which included specialists in design, assistive technologies, and health and care provision, as well as robotics) helped to guard against adopting preconceived positions or attitudes.

Further limitations concern the representativeness of the participants. The co-design activities involved a total of only 20 older people, plus a smaller number of relatives, friends, and caregivers. The participants were healthy enough (or coping well enough) to attend workshops and possessed some degree of day-to-day independence. Hence, people living with more severe frailty or other age-related health conditions (such as dementia) were not represented directly in the workshops, and no consideration was given to the particularities of life in care homes and other facilities offering specialized care (although the use of personas with more severe frailty and health issues was an attempt to address, at least in part, these aspects). Clearly, there is much scope for developing assistive robotics for people with greater needs and for the staff who minister to those needs, but that would require other, more focused co-design exercises. To this end, we have also conducted workshops with health care professionals, which—as may be expected—focused more specifically on the needs of people with more severe frailty and health conditions (including end-of-life care) and with a greater emphasis on support for their caregivers. As the outcomes of these workshops are thematically quite different from those reported here, this work will be reported separately.

Furthermore, the participants in this exercise were self-selecting, and as such might be considered to have a greater interest in robotics (and digital technologies more generally), and possibly to be better disposed to the acceptance of domestic assistive robotics. We did not collect data about the socioeconomic status or the digital literacy and experience of participants, factors that can influence attitudes to robotics. It was not feasible to reconvene participants to confirm or correct the analysis of the workshop data.

In demographic terms, all the older participants self-identified as having a White British ethnic background (almost certainly a reflection of the constitution of independent-living schemes from which most participants were recruited rather than any explicit bias in recruitment practices). Looking beyond to other, perhaps underserved, communities and their living circumstances could expose a different set of needs and wants. Although we found little direct evidence of differences in attitudes according to age or gender, these moderators will likely have some influence on acceptance, even if it was not explored by our participants. To assume that all older people will have the same attitudes to technology, disregarding other factors in their background, is to risk adhering to ageist stereotypes [44]. The very term “older people,” with the suggestion it conjures of a homogeneous mass of people, can itself be a barrier. However, we would argue that experience is a more influential moderator than age: many older people, including some of our workshop participants, will have had extensive experience with digital technology throughout their working lives (which, in the case of a couple of participants, was ongoing) or because they have readily embraced it in their social lives, and are receptive to and comfortable with the idea of using digital assistive devices. Studies have suggested that perceived value and benefit of new technologies are more significant for technology acceptance than is chronological age [45,46]. Rather than being technophobic as a rule or having low levels of digital literacy, older people seem less likely to invest time in new digital technologies whose value for them is not apparent. However, while we have striven throughout to avoid adopting ageist stereotypes, we are aware that this is an all-too-easy pitfall (and, indeed, we noted during the workshops that older people themselves sometimes fall back on these facile generalizations).

In summary, we could not hope to capture everyone’s experience of frailty and aging, or even to encompass the complexity of a single individual’s experience. Moreover, the results of this exercise are specific to a particular population, time, and place (and as such have a limited shelf-life): namely, they are valid for (a small subset of) older people living in the United Kingdom in the early 2020s. People’s experiences in different places and times will, of course, be different. If we were to repeat this exercise in, say, a generation’s time, we would expect the needs and expectations of older people to diverge dramatically from those seen in our results. By this measure, there is no such thing as a uniform, constant, consistent experience of frailty and aging. In some sense, this is not a failing of this exercise but rather an essential feature of co-design, albeit one rarely acknowledged in the literature. Design always responds to problems and opportunities situated in particular times and places; seen in this light, this is not a limitation of the snapshot, but a fundamental aspect of it. Good designs, or ones that address a particular need, might be able to transcend time and place to some extent, but of necessity they must be grounded in the here-and-now experience of everyday life.

Conclusions

The contributions of this paper can be summarized as follows. We have described the predominant clinical models of frailty and explained why these form a useful but incomplete basis for designing assistive technologies of any sort for people living with frailty and older people more generally. We have described the current underwhelming state of assistive robotics, with common failings during their design or co-design processes that often result in the development of inappropriate robotic services.

We have described a novel co-design methodology that has attempted to combine persona-based lived-experience workshops with provotype-based speculative-design workshops with older people to gain a rounded “phenomenological snapshot” of the experience of being old in an increasingly digital world, a world in which assistive robotics may soon become commonplace. As far as we are aware, this is the first time that an exercise of this scope has been attempted in the context of assistive robotics. Finally, we have outlined the major concepts that are revealed by an analysis of the results of the workshops. These can be categorized thematically as: everyday difficulties, the problems faced daily by people living with frailty and older people more generally; ideas for aging better, older people’s own suggestions for how their lives could be improved; and living with technology. In addition to participants’ preferences and requirements, this last category reveals the wide-ranging concerns that people have about services based on digital

XSL•FO
RenderX

technologies, and about assistive robots in particular, whose development is viewed with a mixture of enthusiasm and unease.

Acknowledgments

The authors gratefully acknowledge the contributions of all those who participated in the Emergence co-design workshops described herein, the assistance of Ms Rebekah Moore, the project manager, in organizing the workshops, and the illustrative skills of Mr Sam Church. Generative artificial intelligence was not used for any aspect of this paper's writing.

Data Availability

The data generated and analyzed during this study are available from the corresponding author (SP) on reasonable request.

Funding

Funding for this research was provided through the Emergence: "Tackling Frailty – Facilitating the Emergence of Healthcare Robots From Labs Into Service Healthcare Technologies Network+" grant awarded by the UK Engineering and Physical Sciences Research Council (EPSRC EP/W0000741/1). The funder had no involvement in this study's design, data collection, analysis, interpretation, or the writing of this paper. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any author-accepted manuscript version arising from this submission.

Authors' Contributions

Conceptualization: PC-S (lead), MH (equal), FA (equal), MD (equal), ADN (equal)
Formal analysis: SP (lead), AH (equal), MH (supporting), PC-S (supporting)
Funding acquisition: PC-S (lead), MH (supporting), FA (supporting), MD (supporting), ADN (supporting)
Investigation: SP (lead), PC-S (equal), MH (equal), FA (equal), MD (equal), ADN (equal)
Methodology: SP (lead), MH (supporting)
Project administration: PC-S (lead), SP (equal), FA (supporting), MD (supporting)
Visualization: SP (lead), AH (equal)
Writing – original draft: SP (lead), AH (supporting)
Writing – review & editing: SP (lead), MH (supporting), AH (supporting), FA (supporting), MD (supporting), ADN (supporting), PC-S (supporting)

Conflicts of Interest

None declared.

References

1. Wright J. Robots Won't Save Japan: An Ethnography of Eldercare Automation. New York: ILR Press; 2003.
2. Xue Q. The frailty syndrome: definition and natural history. Clin Geriatr Med 2011;27(1):1-15 [FREE Full text] [doi: 10.1016/j.cger.2010.08.009] [Medline: 21093718]
3. Clegg A, Bates C, Young J, Ryan R, Nichols L, Ann Teale E, et al. Development and validation of an electronic frailty index using routine primary care electronic health record data. Age Ageing 2016;45(3):353-360 [FREE Full text] [doi: 10.1093/ageing/afw039] [Medline: 26944937]
4. Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, Cardiovascular Health Study Collaborative Research Group. Frailty in older adults: evidence for a phenotype. J Gerontol A Biol Sci Med Sci 2001;56(3):M146-M157. [doi: 10.1093/gerona/56.3.m146] [Medline: 11253156]
5. Walston JD, Bandeen-Roche K. Frailty: a tale of two concepts. BMC Med 2015;13:185 [FREE Full text] [doi: 10.1186/s12916-015-0420-6] [Medline: 26265077]
6. Rockwood K, Andrew M, Mitnitski A. A comparison of two approaches to measuring frailty in elderly people. J Gerontol A Biol Sci Med Sci 2007;62(7):738-743. [doi: 10.1093/gerona/62.7.738] [Medline: 17634321]
7. Rockwood K, Mitnitski A. Frailty in relation to the accumulation of deficits. J Gerontol A Biol Sci Med Sci 2007;62(7):722-727. [doi: 10.1093/gerona/62.7.722] [Medline: 17634318]
8. Lang IA, Llewellyn DJ, Langa KM, Wallace RB, Huppert FA, Melzer D. Neighborhood deprivation, individual socioeconomic status, and cognitive function in older people: analyses from the English Longitudinal Study of Ageing. J Am Geriatr Soc 2008;56(2):191-198 [FREE Full text] [doi: 10.1111/j.1532-5415.2007.01557.x] [Medline: 18179489]
9. Werner C, Moustris GP, Tzafestas CS, Hauer K. User-oriented evaluation of a robotic rollator that provides navigation assistance in frail older adults with and without cognitive impairment. Gerontology 2018;64(3):278-290. [doi: 10.1159/000484663] [Medline: 29183035]
10. Koumpouros Y, Toulias TL, Tzafestas CS, Moustris GP. Assessment of an intelligent robotic rollator implementing navigation assistance in frail seniors. Technol Disability 2020;32(3):159-177. [doi: 10.3233/tad-200271]

XSL•FO
RenderX

11. Ozaki K, Kondo I, Hirano S, Kagaya H, Saitoh E, Osawa A, et al. Training with a balance exercise assist robot is more effective than conventional training for frail older adults. Geriatr Gerontol Int 2017;17(11):1982-1990. [doi: 10.1111/ggi.13009] [Medline: 28295912]
12. Hai NDX, Thinh NT. Self-feeding robot for elder people and Parkinson‘s patients in meal supporting. Int J Mech Eng Robot Res 2022;11(4):241-247 [FREE Full text]
13. Luperto M, Monroy J, Renoux J, Lunardini F, Basilico N, Bulgheroni M, et al. Integrating social assistive robots, IoT, virtual communities and smart objects to assist at-home independently living elders: the movecare project. Int J Soc Robot 2023;15(3):517-545 [FREE Full text] [doi: 10.1007/s12369-021-00843-0] [Medline: 35194482]
14. Pollak C, Wexler SS, Drury L. Effect of a robotic pet on social and physical frailty in community-dwelling older adults: a randomized controlled trial. Res Gerontol Nurs 2022;15(5):229-237. [doi: 10.3928/19404921-20220830-01] [Medline: 36113009]
15. Yamazaki Y, Ishii M, Ito T, Hashimoto T. Frailty care robot for elderly and its application for physical and psychological support. J Adv Comput Intell Intell Inform 2021;25(6):944-952. [doi: 10.20965/jaciii.2021.p0944]
16. Boumans R, van Meulen F, Hindriks K, Neerincx M, Olde Rikkert MGM. Robot for health data acquisition among older adults: a pilot randomised controlled cross-over trial. BMJ Qual Saf 2019;28(10):793-799 [FREE Full text] [doi: 10.1136/bmjqs-2018-008977] [Medline: 30894423]
17. Civit A, Andriella A, Barrue C, Antonio M, Boqué C, Alenyà G. Introducing social robots to assess frailty in older adults. Boulder, CO: ACM; 2024 Presented at: HRI '24: Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction; March 11, 2024; Boulder, CO p. 342-346.
18. Olde Keizer RACM, van Velsen L, Moncharmont M, Riche B, Ammour N, Del Signore S, et al. Using socially assistive robots for monitoring and preventing frailty among older adults: a study on usability and user experience challenges. Health Technol 2019;9(4):595-605. [doi: 10.1007/s12553-019-00320-9]
19. Arunachalam S. Do socially assistive robots help in caring for elderly patients with frailty or mild cognitive impairment? J Stud Res 2023;12(3). [doi: 10.47611/jsrhs.v12i3.4821]
20. Lunardini F, Luperto M, Romeo M, Renoux J, Basilico N, Krpic A, et al. The MOVECARE project: home-based monitoring of frailty. Chicago, IL: IEEE; 2019 Presented at: IEEE EMBS International Conference on Biomedical & Health Informatics (BHI); May 19-22, 2019; Chicago, IL.
21. Kim J, Choi Y, Jeong S, Han J. A care robot with ethical sensing system for older adults at home. Sensors (Basel) 2022;22(19):7515 [FREE Full text] [doi: 10.3390/s22197515] [Medline: 36236614]
22. Fiorini L, Tabeau K, D'Onofrio G, Coviello L, De Mul M, Sancarlo D, et al. Co-creation of an assistive robot for independent living: lessons learned on robot design. Int J Interact Des Manuf 2020;14(2):491-502. [doi: 10.1007/s12008-019-00641-z]
23. Steen M. Co-design as a process of joint inquiry and imagination. Design Issues 2013;29(2):16-28. [doi: 10.1162/DESI_a_00207]
24. García-Soler Á, Facal D, Díaz-Orueta U, Pigini L, Blasi L, Qiu R. Inclusion of service robots in the daily lives of frail older users: a step-by-step definition procedure on users' requirements. Arch Gerontol Geriatr 2018;74:191-196. [doi: 10.1016/j.archger.2017.10.024] [Medline: 29128788]
25. Fiorini L, D'Onofrio G, Rovini E, Sorrentino A, Coviello L, Limosani R, et al. A robot-mediated assessment of Tinetti balance scale for sarcopenia evaluation in frail elderly. New Delhi, India: IEEE; 2019 Presented at: 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN); October 14-18, 2019; New Delhi, India.
26. Coviello L, Cavallo F, Limosani R, Rovini E, Fiorini L. Machine learning based physical human-robot interaction for walking support of frail people. Berlin, Germany: IEEE; 2019 Presented at: 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); July 23-27, 2019; Berlin, Germany.
27. Bardaro G, Antonini A, Motta E. Robots for elderly care in the home: a landscape analysis and co-design toolkit. Int J Soc Robotics 2022;14:657-681. [doi: 10.1007/s12369-021-00816-3]
28. Spiers J, Smith JA. Interpretative phenomenological analysis. In: Atkinson P, Delamont S, Cernat A, Sakshaug JW, Williams RA, editors. SAGE Research Methods Foundations. London: SAGE Publications Ltd; 2019.
29. Pruitt J, Adlin T. The Persona Lifecycle: Keeping People in Mind Throughout Product Design. San Francisco, CA: Morgan Kaufmann; 2006.
30. van den Heuvel H, Huijnen C, Caleb-Solly P, Nap H, Nani M, Lucet E. Mobiserv: a service robot and intelligent home environment for the provision of health, nutrition and safety services to older adults. Gerontechnology 2012;11(2):373. [doi: 10.4017/gt.2012.11.02.564.00]
31. Wöckl B, Yildizoglu U, Buber I, Aparicio Diaz B, Kruijff E, Tscheligi M. Basic senior personas: a representative design tool covering the spectrum of European older adults. New York, NY: ACM; 2012 Presented at: ASSETS '12: Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility; October 22, 2012; Boulder, CO.
32. Valenza C, Adkins J. TIMELINES: understanding visual thinking: the history and future of graphic facilitation. Interactions 2009;16(4):38-43. [doi: 10.1145/1551986.1551994]
33. Espiner D, Hartnett F. Innovation and graphic facilitation. Aotearoa New Zealand Social Work 2016;28(4):44-53 [FREE Full text]

XSL•FO
RenderX

34. Boer L, Donovan J. Provotypes for participatory innovation. New York, NY: ACM; 2012 Presented at: DIS '12: Proceedings of the Designing Interactive Systems Conference; June 11, 2012; Newcastle Upon Tyne, United Kingdom.
35. Dunne A, Raby F. Speculative Everything: Design, Fiction, and Social Dreaming. Cambridge, MA: The MIT Press; 2013.
36. Sellen KM, Massimi MA, Lottridge DM, Truong KN, Bittle SA. The people-prototype problem: understanding the interaction between prototype format and user group. Boston MA: ACM; 2009 Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; April 4, 2009; Boston, MA. [doi: 10.1145/1518701.1518799]
37. Hogan DB, MacKnight C, Bergman H, Steering Committee, Canadian Initiative on Frailty and Aging. Models, definitions, and criteria of frailty. Aging Clin Exp Res 2003;15(3 Suppl):1-29. [Medline: 14580013]
38. Ritchie J, Lewis J. Qualitative Research Practice: A Guide for Social Science Students and Researchers. London: SAGE Publications Ltd; 2003.
39. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013;13:117 [FREE Full text] [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
40. Tuffour I. A critical overview of interpretative phenomenological analysis: a contemporary qualitative research approach. J Healthc Commun 2017;02(04):1-5. [doi: 10.4172/2472-1654.100093]
41. Neubauer BE, Witkop CT, Varpio L. How phenomenology can help us learn from the experiences of others. Perspect Med Educ 2019;8(2):90-97 [FREE Full text] [doi: 10.1007/s40037-019-0509-2] [Medline: 30953335]
42. Briefing: facts and figures about digital inclusion and older people. Age UK. 2024. URL: https://www.ageuk.org.uk/siteassets/documents/reports-and-publications/reports-and-briefings/active-communities/internet-use-statistics-june-2024.pdf [accessed 2026-01-15]
43. Moharana S, Panduro AE, Lee HR, Riek LD. Robots for joy, robots for sorrow: community based robot design for dementia caregivers. 2019 Presented at: 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI); March 11-14, 2019; Daegu, Korea (South). [doi: 10.1109/hri.2019.8673206]
44. Mannheim I, Wouters EJM, Köttl H, van Boekel LC, Brankaert R, van Zaalen Y. Ageism in the discourse and practice of designing digital technology for older persons: a scoping review. Gerontologist 2023;63(7):1188-1200 [FREE Full text] [doi: 10.1093/geront/gnac144] [Medline: 36130318]
45. Hauk N, Hüffmeier J, Krumm S. Ready to be a silver surfer? A meta-analysis on the relationship between chronological age and technology acceptance. Comput Hum Behav 2018;84:304-319. [doi: 10.1016/j.chb.2018.01.020]
46. Berkowsky RW, Sharit J, Czaja SJ. Factors predicting decisions about technology adoption among older adults. Innov Aging 2018;2(1):igy002 [FREE Full text] [doi: 10.1093/geroni/igy002] [Medline: 30480129]

Edited by A Stone; submitted 21.May.2025; peer-reviewed by T Bernier, A Olsson; comments to author 25.Nov.2025; accepted 31.Dec.2025; published 28.Jan.2026.

Please cite as:
Potter S, Hawley M, Higgins A, Amirabdollahian F, Dragone M, Di Nuovo A, Caleb-Solly P
Assistive Robotics for Healthy Aging: A Foundational Phenomenological Co-Design Exercise
J Med Internet Res 2026;28:e77179
URL: https://www.jmir.org/2026/1/e77179
doi:10.2196/77179
PMID:

©Stephen Potter, Mark Hawley, Angela Higgins, Farshid Amirabdollahian, Mauro Dragone, Alessandro Di Nuovo, Praminda Caleb-Solly. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 28.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Adoption of Internet of Things in Health Care: Weighted and Meta-Analytical Review of Theoretical Frameworks and Predictors

Inês Veiga[1], MSc; Tiago Oliveira[1], PhD; Mijail Naranjo-Zolotov[1], PhD; Ricardo Martins[1], PhD; Stylianos Karatzas[2], PhD

[1]NOVA Information Management School (NOVA IMS), Lisbon, Portugal

[2]Department of Product & Systems Design Engineering, University of the Aegean, Hermópolis, Greece

Corresponding Author:
Inês Veiga, MSc
NOVA Information Management School (NOVA IMS)
Universidade Nova de Lisboa Campus de Campolide
Lisbon, 1070-312
Portugal
Phone: 351 213 828 610
Email: iveiga@novaims.unl.pt

Abstract

Background: The integration of the Internet of Things (IoT) into health care is transforming the industry by enhancing disease care and management, as well as supporting self-health management. The COVID-19 pandemic has accelerated the adoption of IoT devices, particularly wearable medical devices, which enable real-time health monitoring and advanced remote health management. Globally, the increased adoption of IoT in health care has improved efficiency, enhanced patient care, and generated substantial economic value.

Objective: This review aims to conduct a comprehensive meta- and weight analysis of quantitative studies to identify the most influential predictors and theoretical frameworks explaining the adoption of IoT in health care.

Methods: We searched databases, including Web of Science and PubMed, for quantitative studies on IoT health care adoption, with the last search conducted in early July 2025. Inclusion criteria comprised peer-reviewed articles written in English that employed a quantitative approach to IoT health care technology adoption. Studies were excluded if they did not report the significance of relationships, involved technologies without IoT features or were outside the scope, or examined target variables irrelevant to the analysis. The weight analysis identified the pathways with the most significant effects. A meta-analysis using a random-effects model was conducted to estimate combined effect sizes and their statistical significance. The results from both methods were then integrated to visualize the most frequently used theoretical frameworks. Risk of bias and heterogeneity were assessed using a funnel plot, Egger regression test, the I2 statistic, and subgroup analysis, which indicated no strong evidence of publication bias but revealed a high level of heterogeneity.

Results: Analysis of 115 datasets from 109 papers identified the Technology Acceptance Model and the Unified Theory of Acceptance and Use of Technology (UTAUT) as the primary frameworks for explaining IoT adoption in health care. Incorporating context-specific variables—such as health consciousness, innovativeness, and trust—into these traditional technology acceptance frameworks enhances the understanding of IoT adoption. Although high heterogeneity suggests a need to refine theoretical models to account for regional contexts, universal adoption drivers such as performance expectancy and effort expectancy remain consistent.

Conclusions: Behavioral intention is the most frequently studied variable in IoT health care adoption, whereas attitude, performance expectancy, effort expectancy, and task-technology fit remain underexplored. While adoption theories from the information systems field, such as the TAM, are predominantly used, integrating context-specific constructs and theories—such as trust and innovativeness—can provide deeper insights into IoT adoption in health care. The strongest and most consistent predictors of behavioral intention were attitude, performance expectancy, habit, self-efficacy, functional congruence, and benefits. Additionally, social influence, facilitating conditions, trust, and aesthetic appeal demonstrated promising or strong effects. By contrast, variables such as privacy and security, barriers, vulnerability, severity, compatibility, financial cost, health, and technology anxiety were generally inconsistent or not statistically significant.

(J Med Internet Res 2026;28:e64091) doi:10.2196/64091

XSL•FO
RenderX

KEYWORDS

IoT health care; internet of things; individual adoption; weight-analysis; meta-analysis

Introduction

The integration of the Internet of Things (IoT) into health care has revolutionized the industry by introducing a new paradigm of connectivity and data exchange, driven by rapid advancements in IoT, artificial intelligence, and machine learning [1-3]. This era, known as Healthcare 4.0, can be leveraged to enhance acute disease care, manage chronic diseases, and support self-health management [4]. The COVID-19 pandemic accelerated the adoption of user-friendly IoT devices [5-7], with wearable medical devices emerging as key allies by offering real-time health monitoring, continuous data transmission, and advanced remote health management [8-10].

Globally, the integration of IoT in health care has enhanced efficiency, improved patient care, and generated significant economic value [11,12]. By 2029, the global IoT health care market volume is projected to reach US $134.40 billion [13]. This indicates strong, sustained growth driven by the increasing adoption of IoT technologies in health care around the globe [13,14]. For instance, China has made significant progress in integrating health information technologies into the health care system, driven by initiatives such as “Internet Plus Health Care” and the “Healthy China 2030” plan [15-17]. The United States leads in IoT and intelligent health care system development, supported by substantial investments and a robust ecosystem of startups and tech companies driving advancements in artificial intelligence and IoT [18-20]. Europe also shows considerable progress, emphasizing regulatory frameworks, standardization, and interoperability to foster innovation and data protection [21,22].

While IoT holds considerable promise to transform health care by reducing costs and improving access, understanding the factors influencing its adoption requires more focused research [23]. Although literature reviews with a quantitative approach have examined technology adoption in health care, existing meta-analyses that include technologies with IoT features remain fragmented, as they have largely focused on broader or adjacent technological domains and have typically emphasized a specific adoption model. For instance, meta-analyses on mobile health have focused on the Unified Theory of Acceptance and Use of Technology (UTAUT) [24] and the Technology Acceptance Model (TAM) [25]. Meta-analyses on eHealth have predominantly focused on the TAM [26] and continuance intention [27]. Meta-analyses specific to smart wearable health care devices have examined attitude and intention using UTAUT and TAM [28], as well as the effects of perceived usefulness and perceived ease of use on intention, with a focus on Hofstede’s cultural dimensions as moderators [29]. Taken together, these studies often treat health care technology adoption in general terms and do not account for the unique characteristics of IoT.

Our study addresses this gap by providing a comprehensive meta-analysis and a weight analysis specifically focused on IoT adoption in health care. We synthesize findings from primarily quantitative articles on the adoption of IoT in health care, particularly on interconnected devices that monitor and transmit real-time health care data, enabling smarter solutions [5], such as smart sensors, remote monitoring devices, and health-focused IoT platforms. Our meta-analytic approach integrates findings from different theoretical perspectives, including technology adoption models such as the TAM and UTAUT and health-specific models such as the Health Belief Model (HBM), allowing for a more holistic understanding of adoption dynamics in health care contexts. Moreover, our dual-method approach, combining meta-analysis with weight analysis, identifies the strongest and most reliable predictors of adoption and maps the theoretical foundations most frequently and effectively used in this field. This analysis goes beyond prior reviews, offering new evidence-based insights to guide health care technology developers, practitioners, and researchers. The objectives of this study are, first, to identify key predictors of IoT health care adoption through a comprehensive meta-analysis and weight analysis, and second, to determine the most influential and empirically supported theories used to explain IoT adoption in health care settings.

Methods

Overview

We performed a meta-analysis to examine the factors influencing the adoption of IoT technologies in health care, synthesizing findings from a range of quantitative studies. By focusing on primary quantitative research articles, we aimed to identify the most significant predictors and the theoretical models most commonly used to explain IoT adoption in health care settings.

Information Sources and Search Strategy

This meta-analysis followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines [30]. The literature search was conducted using a keyword-based search across the Web of Science and PubMed databases to identify studies examining IoT adoption in health care. The search strategy incorporated title, abstract, and keyword searches, using Boolean operators (AND and OR) and database-specific filters. The keywords used in our search were related to IoT technology, relevant variables, quantitative methods, and exclusion criteria (Table 1). We included records published up to the end of 2024. The complete search strategy, including search terms and Boolean logic, is provided in Multimedia Appendix 1.

XSL•FO
RenderX

Table 1. Map for the keyword search in online databases.

Relevant terms	Relevant variables and theories	Methodologies	Exclusion of irrelevant topics
Internet of Things	Intention to adopt	Structural equation	Systematic
IoT	Behavioral intention	Structural equation modeling	Literature review
Smart	Acceptance	Partial least squares structural equation modeling	Postadoption
Intelligent	Adopt	Partial least	Meditation
Health care wearable device	Adoption	Path analysis	Contact tracing app
Medical wearable technology	Using	Regression	Fitness app
Health management	Use	N/A[a]	Electronic health record
Health measurement	Usage	N/A	Telemedicine
N/A	Intention to use	N/A	Mindfulness app
N/A	Unified Theory of Acceptance and Use of Technology 2	N/A	N/A
N/A	Unified Theory of Acceptance and Use of Technology	N/A	N/A
N/A	Technology Acceptance Model	N/A	N/A

[a]N/A: not applicable.

Selection Criteria

Initial screening was conducted using database filters. In the second screening, 2 (IV and MNZ) independent reviewers assessed the titles and abstracts for relevance, resolving disagreements through discussion or arbitration by a third reviewer. The full-text screening followed the same procedure. The reports assessed for eligibility were exported to an Excel (Microsoft Corporation) file, and all included papers were imported into Zotero (Corporation for Digital Scholarship), which is a reference management software. When a paper was unavailable, the authors were contacted.

Inclusion criteria comprised peer-reviewed articles with a quantitative approach to health care technology adoption written in English. Reasons for exclusion included not reporting the significance of the relationships between variables; the technology lacking IoT features or being unrelated to health care; the target variables being unrelated to adoption or focusing solely on postadoption behaviors; and studies lacking empirical data or reporting qualitative results only. The workflow and search conditions are depicted in more detail in the “Results” section.

Data Extraction

A standardized data extraction form was developed before data extraction. Data extraction was performed in Excel, and for each article, we detailed the study characteristics, methodology, type of technology, and the effects measured across multiple paths. The extracted aspects and their descriptions are provided in Table S1 in Multimedia Appendix 1. We assessed paper quality by examining the publishing journal metrics, the methods employed, sample size, and the scales used to measure each construct. The standardized β coefficients were extracted as the primary effect measure. As some authors used different names to represent the same variable, several variables had to be merged to conduct our analysis. This process was carried out by reading each variable definition and identifying the items used to measure them. Examples of variable mergers are provided in Table S2 in Multimedia Appendix 1, and the individual studies included in the analysis are detailed in Table S3 in Multimedia Appendix 1.

Descriptive Analysis

We extracted metadata from each study to perform a descriptive analysis of publication trends, journal quality, and research domains. Data on publication year were used to assess the chronological distribution of studies. To evaluate journal quality, we matched each journal with its SCImago Journal & Country Rank classification and categorized them into quartiles (Q1-Q4). The disciplinary scope of the journals was identified based on their SCImago subject area classifications. We also recorded the journal title and publication frequency to identify the journals that published the most research. Country-level data were extracted based on the origin of the study sample or study location. We computed the number of studies and total sample sizes per country to identify regions with the highest research activity. To understand the theoretical foundations employed across studies, we reviewed each article’s methodology and coded the theories used to model technology adoption behavior. We also recorded whether these models were used independently or in combination (eg, UTAUT extended with Protection Motivation Theory [PMT] constructs).

Weight Analysis

The weight analysis was conducted to uncover the predictive power of independent variables [31]. This weight provides a measure of the relative importance or consistency of statistical significance for each variable across multiple analyses. For the weight-analytic approach, we focused on the influence of each independent variable on several dependent variables and limited our analysis to relationships investigated 3 or more times

XSL•FO
RenderX

[32,33]. The weight (W_i) of an independent variable i is calculated as the ratio of the number of times it was found to be statistically significant (S_i) to the total number of times it was examined (E_i), as expressed in the following equation:

$$W_i = S_i/E_i$$

Meta-Analysis

Meta-analyses allow us to quantitatively compare effect sizes across relationships between constructs using suitable metrics to capture these effect sizes, including standardized regression coefficients [34,35]. This analysis followed best practices outlined previously [36-38]. In our study, the necessary inputs for performing the meta-analysis were the standardized regression coefficients (β) and the sample sizes for each relationship examined 3 or more times across studies. Following the approach of Peterson and Brown [37], β values were transformed into approximate correlation coefficients as $r=\beta+0.05$, where $\lambda=1$ [37]. All correlation coefficients were Fisher z-transformed to stabilize variance, and SEs were computed.

A random-effects model was used to account for both within- and between-study variance, justified by the heterogeneity in study populations, methods, and contexts. Random-effects weights were calculated using the DerSimonian and Laird model, and weighted mean effect sizes were computed using inverse-variance weights, which use tau-squared (τ^2) [39,40]. Heterogeneity was assessed using the Q statistic and the I^2 index [41]. We also calculated the lower and upper bounds of the 95% CIs, z scores, and 2-tailed P values to assess statistical significance and interpret the magnitude of the observed effects. Final pooled effect sizes and CIs were then back-transformed from Fisher z to the correlation coefficient metric (r). All calculations were performed manually in Excel.

Publication Bias Analysis

The Egger test was used to statistically examine the presence of publication bias by regressing the standard normal deviate on precision [42]. The analysis was performed using Excel's data analysis regression tool, which applies standard ordinary least squares regression, and a significant intercept ($P<.10$) was interpreted as evidence of asymmetry and possible publication bias. A funnel plot was constructed to visually assess publication bias using the tool Meta-Essentials [43]. The trim-and-fill method was used to estimate the number and influence of missing studies. Heterogeneity for the included studies was assessed using the I^2 statistic, where a value over 75% is interpreted as substantial heterogeneity, using the following formula, where k is the number of studies and Q the Cochran Q statistic:

$$I^2 = \max(0; \{Q - [k - 1]\}/Q) \times 100\%$$

To evaluate regional bias, we conducted a subgroup analysis with 2 groups: one comprising studies conducted in China and the other comprising studies conducted in the remaining countries. For each group, we calculated the combined effect sizes, SEs, CI lower and upper limits, and the I^2 statistic.

Combining Weight and Meta-Analysis Results: The Most Used Adoption Models

To synthesize the relative strength of relationships across studies and adoption models, we combined the results from the weight analysis and meta-analysis. The weight analysis assessed the consistency and prominence of specific predictors by calculating the proportion of studies that reported statistically significant relationships for each path, referred to as the weight. In parallel, the meta-analysis provided pooled average effect sizes and significance levels across studies using a random-effects model. This dual approach offers a more comprehensive understanding of which constructs consistently predict behavioral intention or usage in the context of IoT adoption in health care and enables an evidence-based comparison of theoretical frameworks based on empirical support.

We then visually mapped the structure of each adoption model, such as the TAM and the HBM, using conceptual diagrams. In these figures, each arrow represents a theoretical path, and its thickness reflects the weight. Thicker lines indicate a weight above 0.700, representing paths supported by a high proportion of studies. The numerical values attached to each path represent the average effect size based on the random-effects meta-analysis, along with the corresponding P value. This dual representation enables a clearer comparison between the predictive strength (effect size) and consistency (weight) of each construct within and across models.

Results

Descriptive Analysis

Papers on IoT health care adoption show an increasing trend, with 89 of the 109 (81.7%) studies in our analysis published between 2020 and 2025, and the earliest published in 2011. According to the SCImago Journal & Country Rank, most papers appeared in Q1 journals (n=63, 57.8%), followed by Q2 (n=36, 33%) and Q3 (n=10, 9.2%), with no papers published in Q4. These studies span major research areas related to health and medicine, information systems, and computer science. In total, 75 unique journals were represented, with *PLoS One* (n=6), *Frontiers in Public Health* (n=5), *Technological Forecasting and Social Change* (n=5), and *International Journal of Environmental Research and Public Health* (n=4) being the most frequently appearing journals (see Table S3 in Multimedia Appendix 1).

In our analysis of 109 studies (see Figure 1), we identified 115 unique datasets totaling 46,508 individuals (see Table S1 in the Multimedia Appendix 1). Studies conducted in China, South Korea, and the United States accounted for a large portion of the total sample and represented the greatest number of publications (see Table 2).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

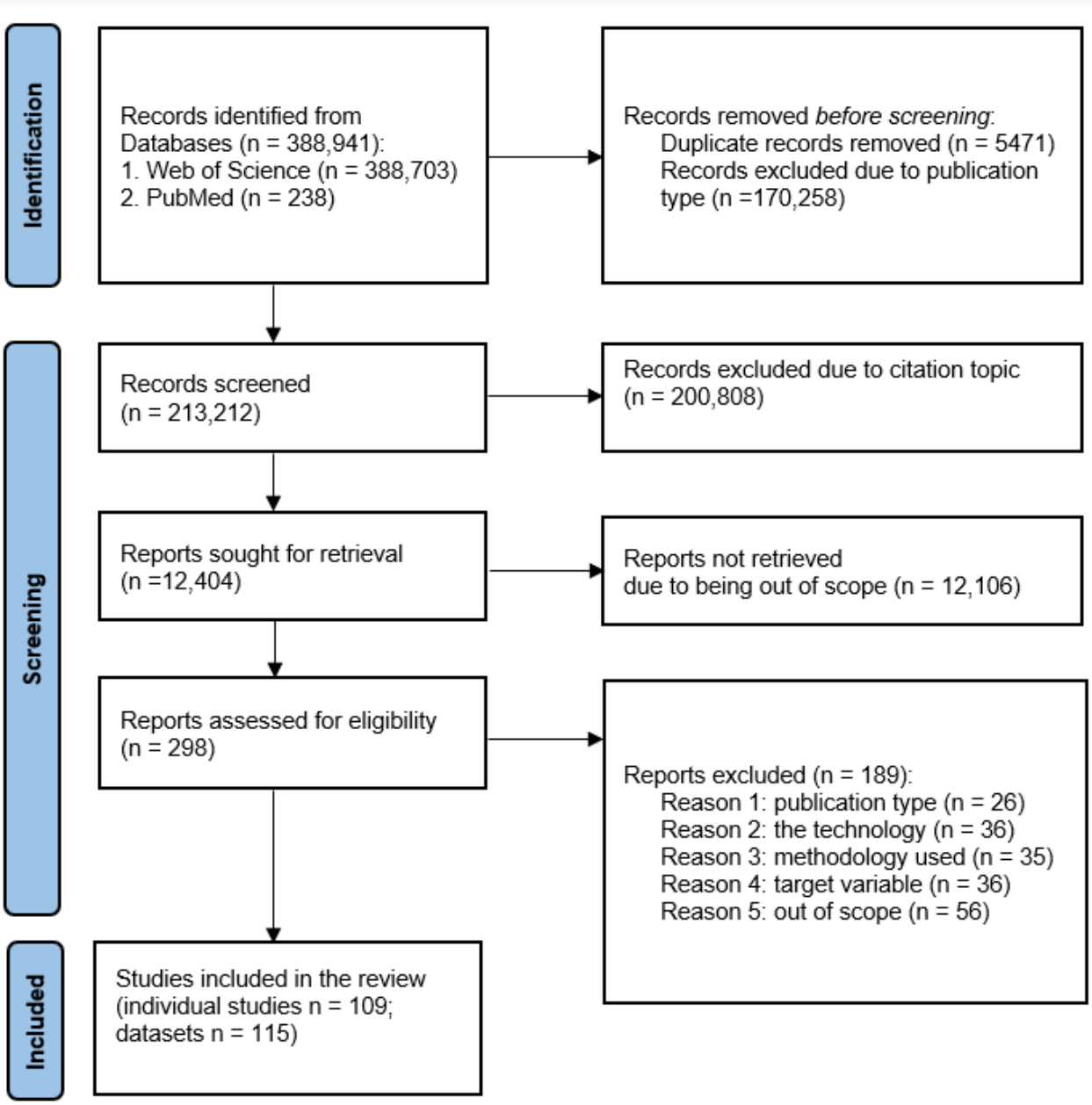

XSL•FO
RenderX

Table 2. Number of papers and total sample size per country.

Country	Dataset count (N=115)	Sample size (N=46,508)
China	39	17,068
South Korea	12	6406
The United States	9	5445
India	8	2924
Taiwan	8	1896
The Kingdom of Saudi Arabia	5	1798
Bangladesh	2	1213
Pakistan	3	1194
Turkey	4	1040
Ghana	2	965
Multiple countries	4	1312
Indonesia	1	772
Malaysia	2	628
France	3	515
Iraq	1	465
Oman	2	442
Romania	1	440
The United Arab Emirates	2	431
Switzerland	2	323
Singapore	1	306
Nepal	1	280
Japan	1	233
Italy	1	212
Jordan	1	200

Considering the theories addressed in each paper, the TAM and the UTAUT have been extensively examined compared with other theories (see Figure 2). These models serve as the theoretical foundation for 92 of the 109 (84.4%) papers included in our study. Other theories, such as the HBM, PMT, Task-Technology Fit (TTF) Theory, Privacy Calculus Theory, Diffusion of Innovation Theory, and Theory of Planned Behavior, have also been addressed—sometimes as the primary theoretical foundation and other times to extend the TAM or UTAUT.

XSL•FO
RenderX

Figure 2. Theoretical foundation of the papers included in our analysis. DOI: Diffusion of Innovation; HBM: Health Belief Model; PCT: Privacy Calculus Theory; PMT: Protection Motivation Theory; TAM: Technology Acceptance Model; TPB: Theory of Planned Behavior; TTF: Task-Technology Fit; UTAUT: Unified Theory of Acceptance and Use of Technology.

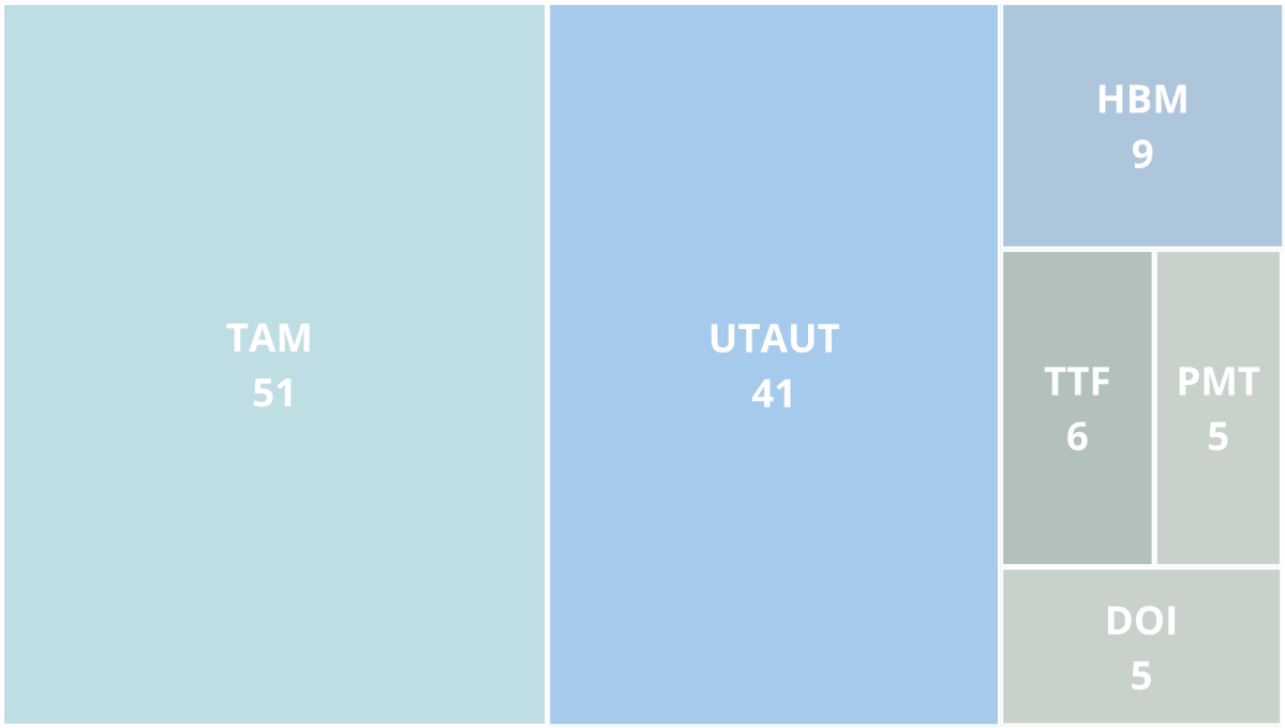

Weight Analysis

A weight analysis examines the strength of the relationship between an independent and a dependent variable. The weights of the identified relationships are analyzed and presented in Table 3. The significance of a relationship's weight is calculated by dividing the number of instances in which the relationship is statistically significant by the total number of studies that investigated it. A weight of 1 indicates that the relationship is significant in all examined studies, whereas a weight of 0 indicates that it is not significant in any of the studies.

XSL•FO
RenderX

Table 3. Identified paths with the nonsignificant paths, the significant relationships, the total paths, and the respective weights.

Dependent and independent variables	Significant	Nonsignificant	Total	Weight=significant/total
Attitude				
Effort expectancy	19	4	23	0.826
Barriers	4	2	6	0.667
Benefits	3	0	3	1
Facilitating conditions	4	0	4	1
Performance expectancy	20	2	22	0.909
Privacy and security	2	1	3	0.667
Social influence	6	1	7	0.857
Behavioral intention				
Aesthetic appeal	4	0	4	1
Attitude	27	1	28	0.964
Barriers	7	6	13	0.538
Benefits	6	0	6	1
Compatibility	4	3	7	0.571
Effort expectancy	36	23	59	0.61
Ethics	2	1	3	0.667
Facilitating conditions	20	8	28	0.714
Financial cost	12	9	21	0.571
Functional congruence	4	1	5	0.8
Habit	7	0	7	1
Health	3	1	4	0.75
Health consciousness	7	4	11	0.636
Hedonic motivation	10	5	15	0.667
Image	4	2	6	0.667
Innovativeness	5	3	8	0.625
Perceived severity	2	3	5	0.4
Perceived vulnerability	3	5	8	0.375
Performance expectancy	68	10	78	0.872
Privacy and security	13	13	26	0.5
Reliability	5	0	5	1
Self-efficacy	10	1	11	0.909
Social influence	31	10	41	0.756
Technology anxiety	2	4	6	0.333
Trust	9	3	12	0.75
Actual behavior				

XSL•FO
RenderX

Dependent and independent variables	Significant	Nonsignificant	Total	Weight=significant/total
Behavioral intention	16	1	17	0.941
Effort expectancy	2	1	3	0.667
Facilitating conditions	3	0	3	1
Health consciousness	2	2	4	0.5
Innovativeness	2	1	3	0.667
Perceived vulnerability	2	1	3	0.667
Performance expectancy	3	1	4	0.75
Social influence	3	0	3	1
Performance expectancy				
Barriers	2	3	5	0.4
Compatibility	6	2	8	0.75
Convenience	3	0	3	1
Effort expectancy	27	4	31	0.871
Facilitating conditions	2	2	4	0.5
Health consciousness	6	2	8	0.75
Image	3	4	7	0.429
Innovativeness	4	0	4	1
Privacy and security	4	2	6	0.667
Reliability	8	3	11	0.727
Self-efficacy	7	0	7	1
Social influence	8	3	11	0.727
Trialability	2	1	3	0.667
Trust	3	2	5	0.6
Task-technology fit	5	0	5	1
Effort expectancy				
Compatibility	7	0	7	1
Facilitating conditions	6	0	6	1
Image	2	2	4	0.5
Innovativeness	7	0	7	1
Privacy and security	2	2	4	0.5
Reliability	4	0	4	1
Self-efficacy	7	0	7	1
Social influence	3	1	4	0.75
Trialability	2	1	3	0.667
Task-technology fit	3	0	3	1
Task-technology fit				
Task characteristics	3	1	4	0.75
Technology characteristics	4	0	4	1

In the context of technology adoption at the individual level, independent variables are considered “well-utilized” if they have been tested at least 5 times. Variables tested fewer than 5 times but with a weight of 1 are regarded as “promising” predictors [31]. To be classified as a “best” predictor, an independent variable must have a weight of 0.800 or higher and must have been examined at least 5 times [31].

In our research, we analyzed the impact of several independent variables on the dependent variables attitude, behavioral intention, actual use, performance expectancy, effort expectancy,

XSL•FO
RenderX

and TTF. For the weight analysis, we included relationships that were examined 3 or more times, resulting in 67 relationships and 31 unique predictors that met this criterion. The most studied target variable was behavioral intention, with 25 predictors.

In our research, the relationships considered the "best" predictors for attitude are effort expectancy, performance expectancy, and social influence, as each has more than 5 identified relationships and a weight greater than 0.800. For behavioral intention, the best predictors are attitude, performance expectancy, habit, self-efficacy, functional congruence, reliability, and benefits. Aesthetic appeal, with a perfect weight of 1, is considered a promising predictor of intention due to the limited number of studies. Social influence, facilitating conditions, and trust, although not classified as the best predictors, remain important because their weights exceed 0.700 and are supported by a substantial number of studies. It is also noteworthy that privacy and security, barriers, vulnerability, severity, compatibility, and financial cost yielded more inconsistent results, with many studies reporting statistically nonsignificant findings.

For actual behavior, behavioral intention is the best predictor, while facilitating conditions and social influence are considered promising predictors due to the limited number of studies and their perfect weight of 1. For the target variable performance expectancy, effort expectancy, TTF, and self-efficacy are the best predictors, and convenience and innovativeness are promising predictors. Health consciousness, social influence, reliability, and compatibility, although not classified as the best predictors, remain important because their weights exceed 0.700 and they are supported by multiple studies. For effort expectancy, facilitating conditions, innovativeness, self-efficacy, and compatibility are the best predictors, while reliability and TTF are promising predictors. For the target variable TTF, technology characteristics is identified as a promising predictor.

Meta-Analysis

The results of the meta-analysis are presented in Table 4 and include all studies that reported standardized path coefficients or β values. All the best predictors identified in our study are statistically significant ($P<.001$), except for reliability ($P=.49$) as a predictor of intention, as well as some of the important and promising predictors. Notably, barriers ($P=.46$) is not a significant predictor of attitude. Health, technology anxiety ($P=.78$), financial cost ($P=.16$), and barriers ($P=.84$) are not significant predictors of behavioral intention. For actual behavior, social influence ($P=.15$), innovativeness ($P=.28$), health consciousness ($P=.61$), vulnerability ($P=.31$), and effort expectancy ($P=.09$) are not significant predictors. Privacy and security ($P=.05$) and barriers ($P=.21$) are not significant predictors of performance expectancy, while privacy and security ($P=.29$) and image ($P=.06$) are not significant predictors of effort expectancy. Finally, task characteristics ($P=.12$) is not a significant predictor of TTF.

XSL•FO
RenderX

Table 4. Meta-analysis results calculated using a random-effects model and presented back-transformed.

Dependent and independent variables	r/ES[a]	95% CI	*Q* statistic	*z* score	*P* value	I^2 statistic (%)
Attitude						
Barriers	0.069	–0.113 to 0.251	107.606	0.743	.46	95.353
Benefits	0.466	0.239 to 0.645	224.814	3.789	<.001	99.11
Effort expectancy	0.286	0.23 to 0.342	153.69	9.549	<.001	86.987
Facilitating conditions	0.527	0.344 to 0.671	874.478	5.067	<.001	99.657
Performance expectancy	0.532	0.414 to 0.633	987.838	7.596	<.001	98.077
Privacy and security	–0.355	–0.618 to –0.093	354.157	–2.653	.008	99.435
Social influence	0.342	0.182 to 0.483	305.084	4.069	<.001	98.033
Behavioral intention						
Health	0.082	–0.049 to 0.21	18.448	1.233	.22	83.738
Aesthetic appeal	0.319	0.266 to 0.371	1.314	11.027	<.001	0
Attitude	0.573	0.454 to 0.672	1855.792	7.853	<.001	98.653
Barriers	–0.016	–0.171 to 0.139	469.927	–0.2	.84	97.446
Benefits	0.309	0.092 to 0.497	688.109	2.757	.006	99.273
Compatibility	0.123	0.081 to 0.165	109.339	5.729	<.001	96.342
Effort expectancy	0.185	0.134 to 0.235	887.804	7.043	<.001	93.58
Ethics	0.303	–0.232 to 0.698	143.458	1.117	.26	98.606
Facilitating conditions	0.198	0.138 to 0.257	274.658	6.318	<.001	90.17
Financial cost	–0.08	–0.191 to 0.031	541.286	–1.414	.16	96.675
Functional congruence	0.212	0.165 to 0.259	39.56	8.509	<.001	89.889
Habit	0.377	0.307 to 0.444	495.521	9.72	<.001	98.789
Health consciousness	0.298	0.222 to 0.371	4455.165	7.332	<.001	99.798
Hedonic motivation	0.202	0.174 to 0.23	135.59	13.785	<.001	89.675
Image	0.201	–0.215 to 0.556	62.391	0.947	.34	93.589
Innovativeness	0.223	0.147 to 0.297	154.334	5.642	<.001	96.112
Performance expectancy	0.339	0.295 to 0.381	1212.594	14.281	<.001	93.732
Privacy and security	–0.11	–0.202 to –0.018	361.354	–2.348	.02	93.912
Reliability	0.148	–0.266 to 0.516	398.688	0.694	.49	98.997
Self-efficacy	0.318	0.257 to 0.377	818.281	9.644	<.001	98.9
Severity	0.12	0.005 to 0.231	8.992	2.04	.04	55.518
Social influence	0.254	0.189 to 0.317	684.855	7.381	<.001	94.305
Technology anxiety	–0.013	–0.1 to 0.075	19.306	–0.279	.78	74.102
Trust	0.294	0.177 to 0.403	666.839	4.769	<.001	98.35
Vulnerability	0.101	0.009 to 0.191	19.657	2.153	.03	64.389
Actual behavior						

XSL•FO
RenderX

Dependent and independent variables	r/ES[a]	95% CI	*Q* statistic	*z* score	*P* value	I^2 statistic (%)
Behavioral intention	0.563	0.427 to 0.674	1139.269	6.912	<.001	98.508
Effort expectancy	0.353	–0.061 to 0.663	99.717	1.683	.09	97.994
Facilitating conditions	0.863	0.706 to 0.939	10353.056	5.987	<.001	99.981
Health consciousness	0.096	–0.267 to 0.436	1594.656	0.511	.61	99.812
Innovativeness	0.232	–0.188 to 0.581	420.955	1.086	.28	99.525
Performance expectancy	0.406	0.001 to 0.696	129.825	1.963	.05	98.459
Social influence	0.31	–0.113 to 0.638	35.827	1.447	.15	94.418
Vulnerability	0.216	–0.204 to 0.569	615.529	1.008	.31	99.675
Performance expectancy						
Effort expectancy	0.38	0.302 to 0.454	587.488	8.873	<.001	95.064
Barriers	–0.137	–0.353 to 0.078	396.849	–1.252	.21	98.992
Compatibility	0.222	0.041 to 0.388	85.162	2.403	.02	92.955
Facilitating conditions	0.328	0.1 to 0.523	239.597	2.784	.005	98.748
Health consciousness	0.188	0.021 to 0.345	140.742	2.206	.03	95.026
Image	0.194	0.001 to 0.373	96.096	1.969	.049	94.797
Innovativeness	0.279	0.049 to 0.481	115.684	2.364	.02	97.407
Privacy and security	–0.193	–0.387 to 0.001	919.956	–1.945	.05	99.456
Reliability	0.309	0.171 to 0.435	82.88	4.258	<.001	87.934
Self-efficacy	0.578	0.431 to 0.695	531.613	6.52	<.001	99.059
Social influence	0.315	0.177 to 0.44	121.563	4.349	<.001	91.774
Trust	0.254	0.046 to 0.441	92.595	2.382	.02	95.68
Task-technology fit	0.678	0.544 to 0.778	278.406	7.539	<.001	98.563
Effort expectancy						
Compatibility	0.318	0.208 to 0.42	35.374	5.441	<.001	85.865
Facilitating conditions	0.436	0.343 to 0.521	24.225	8.321	<.001	79.36
Image	0.147	–0.005 to 0.294	75.386	1.893	.06	97.347
Innovativeness	0.376	0.287 to 0.458	133.654	7.786	<.001	95.511
Privacy and security	–0.074	–0.21 to 0.063	329.763	–1.06	.29	99.09
Reliability	0.46	0.339 to 0.566	45.702	6.747	<.001	93.436
Self-efficacy	0.604	0.529 to 0.67	1371.419	12.34	<.001	99.635
Social influence	0.235	0.097 to 0.364	26.199	3.31	<.001	88.549
Task-technology fit	0.883	0.843 to 0.914	910.714	17.155	<.001	99.78
Task-technology fit						
Task characteristics	0.249	–0.07 to 0.522	460.432	1.537	.12	99.348
Technology characteristics	0.654	0.429 to 0.803	123.033	4.729	<.001	97.562

[a]r/ES: combined effect size (back-transformed from Fisher *z*).

Combining Weight and Meta-Analysis Results: The Most Adopted Models in Research

Figure 3 presents the weight and meta-analysis for the TAM, which explains how users adopt and use technology, emphasizing the influence of external variables on perceived usefulness (performance expectancy) and perceived ease of use (effort expectancy), which in turn affect attitudes, behavioral intentions, and actual technology usage [44,45]. Performance expectancy is the best and statistically significant predictor of both attitude (β=.532, *P*<.001) and behavioral intention (β=.339, *P*<.001). Attitude is the best predictor and has a significant impact on behavioral intention (β=.573, *P*<.001), while behavioral intention is the best and significant predictor of actual behavior (β=.563, *P*<.001). Effort expectancy is the best predictor and strongly influences performance expectancy

XSL•FO
RenderX

(β=.380, *P*<.001) and attitude (β=.286, *P*<.001). Health consciousness (β=.188, *P*=.03), self-efficacy (β=.578, *P*<.001), innovativeness (β=.279, *P*=.02), and compatibility (β=.222, *P*=.02) are significant predictors of performance expectancy, each with a weight above 0.700. Innovativeness (β=.376, *P*<.001) and facilitating conditions (β=.436, *P*<.001) are significant predictors of effort expectancy, also with weights above 0.700.

Figure 3. Weight and meta-analysis for the Technology Acceptance Model. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Higher weights are therefore represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels.

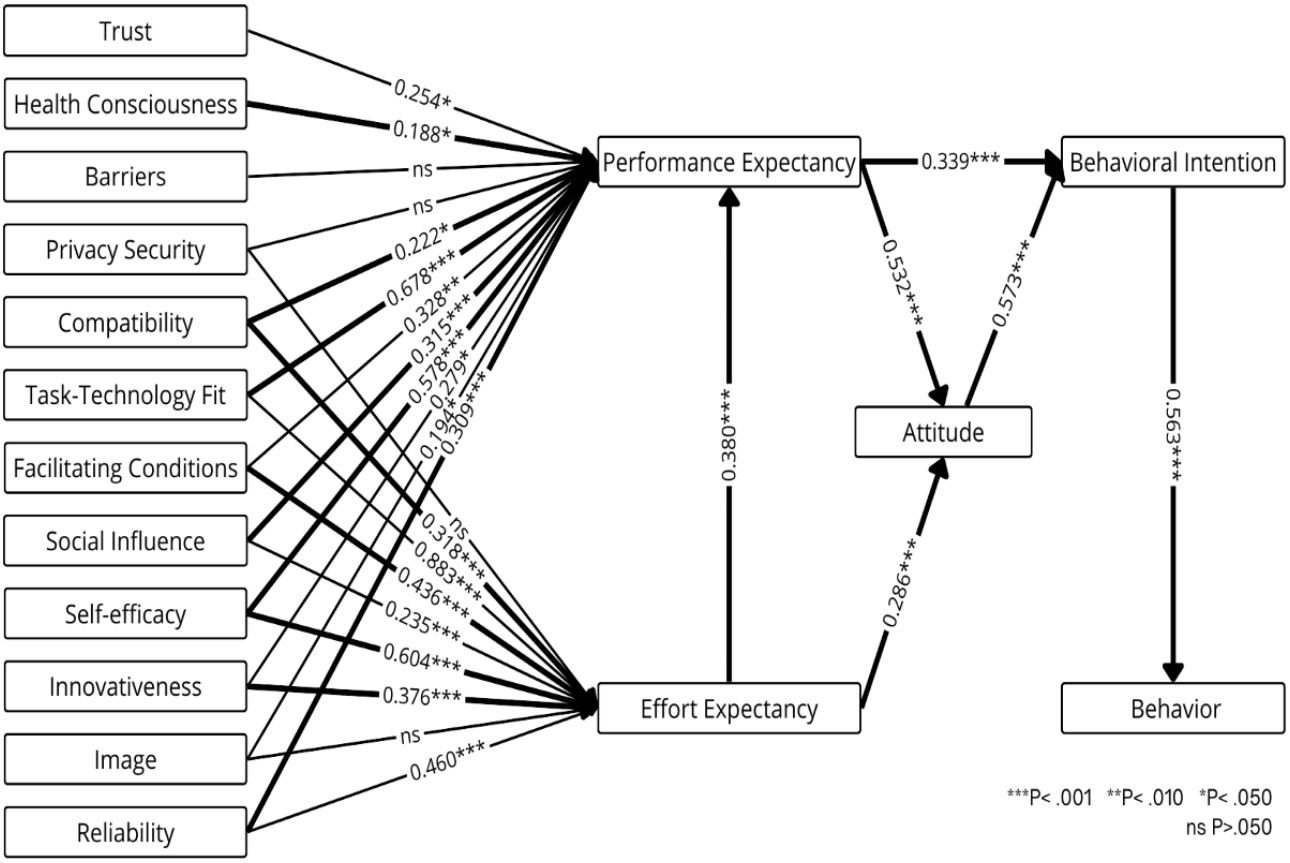

Figure 4 presents the weight and meta-analysis for the Unified Theory of Acceptance and Use of Technology (UTAUT), which explains how users adopt and use technology by assessing the impact of key predictors on behavioral intention and actual behavior [46,47]. Facilitating conditions (β=.863, *P*<.001) and behavioral intention (β=.563, *P*<.001) are significant predictors of actual behavior, while social influence is not. Behavioral intention is significantly influenced by performance expectancy (β=.339, *P*<.001), social influence (β=.254, *P*<.001), facilitating conditions (β=.198, *P*<.001), and habit (β=.377, *P*<.001), all with weights above 0.700. Effort expectancy (β=.185, *P*<.001) and hedonic motivation (β=.202, *P*<.001) are also statistically significant predictors of intention; however, financial cost is not.

XSL•FO
RenderX

Figure 4. Weight and meta-analysis for the Unified Theory of Acceptance and Use of Technology. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Accordingly, higher weights are represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels.

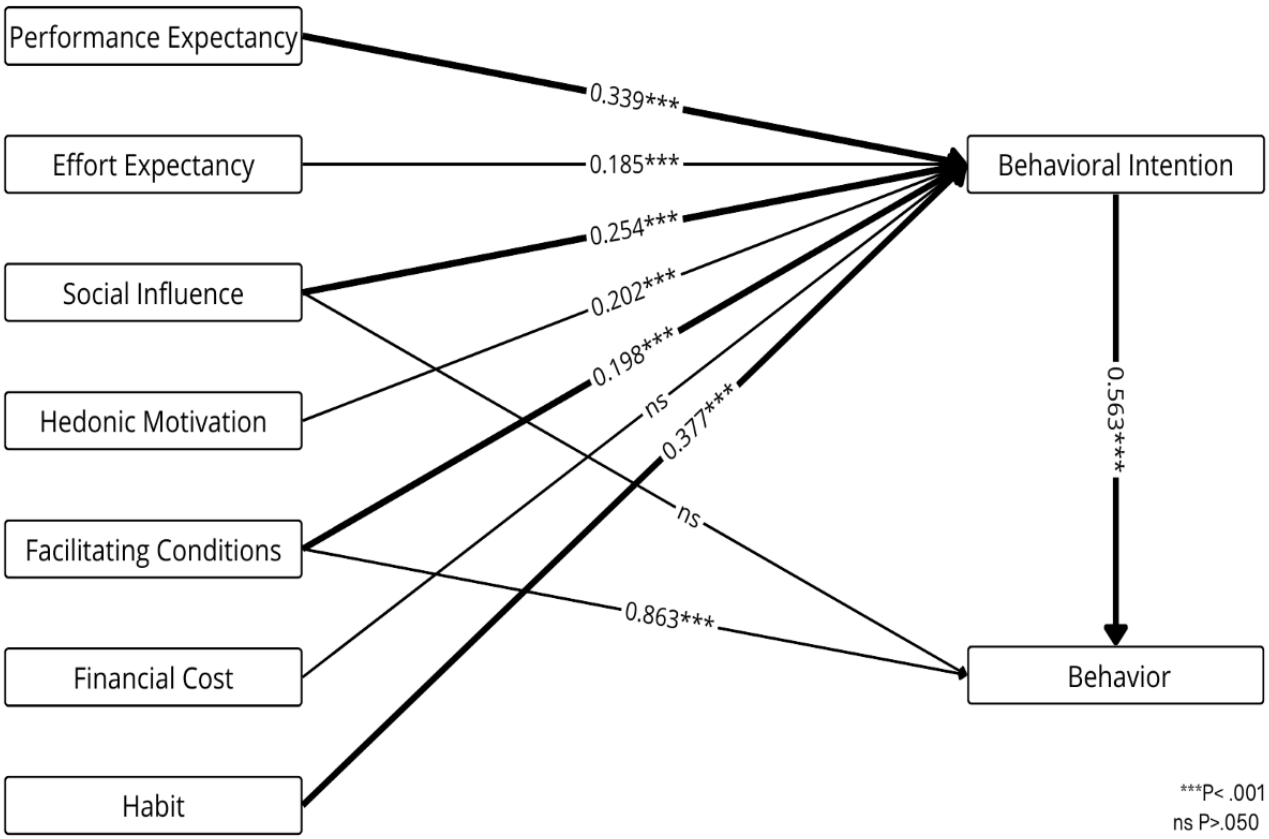

Figure 5 presents the weight and meta-analysis combining the HBM and PMT, which explain individuals' engagement in health-related behaviors. Both models emphasize the role of perceived threat, such as vulnerability and severity, and the evaluation of coping strategies, such as benefits, barriers, response efficacy, and self-efficacy, in shaping motivation to take protective or preventive actions [48-50]. The results indicate that, compared with other technology adoption models, the predictive power of health-related constructs is weaker and less consistent. Severity (β=.120, *P*=.04) and vulnerability (β=.101, *P*=.03) have weak weights on behavioral intention and exert a small but significant impact. Performance expectancy (β=.339, *P*<.001) and self-efficacy (β=.318, *P*<.001), which are also used in other technology-related adoption models, are the best predictors and have a significant impact on behavioral intention. Barriers do not have a significant impact (*P*=.84), whereas benefits exhibit a strong, statistically significant effect (β=.309, *P*=.006). It is also relevant to mention 2 additional predictors directly related to the health context but not part of the key components of these theories. First, health condition, which refers to the perception of having good health, has a weak and nonsignificant impact (*P*=.22) on intention. Second, health consciousness has a statistically significant impact on intention (β=.298, *P*<.001) but does not significantly influence behavior (*P*=.61).

XSL•FO
RenderX

Figure 5. Weight and meta-analysis for the Health Belief Model and Protection Motivation Theory. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Accordingly, higher weights are represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels.

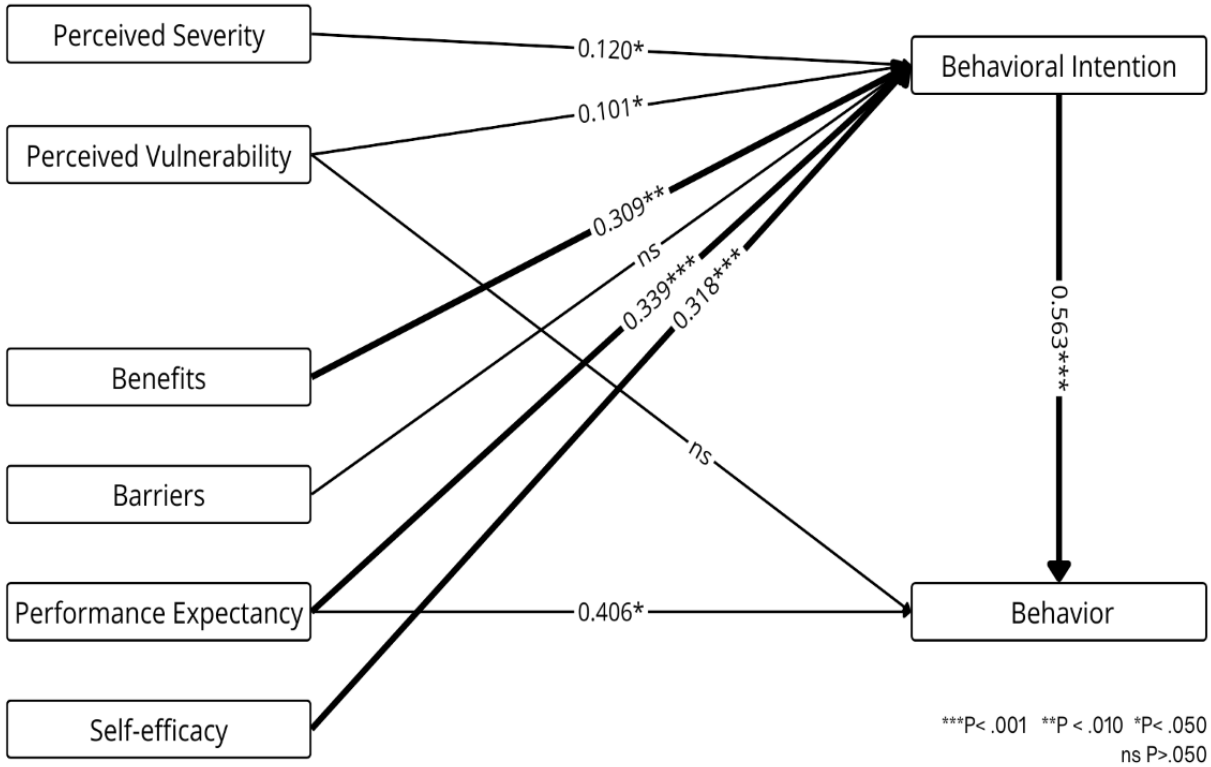

Figure 6 illustrates the TTF model, which examines how well technology aligns with users' tasks to enhance perceived usefulness and adoption [51]. Only part of the theory is presented, as it remains understudied in the context of IoT in health care. The results show that TTF is a significant predictor of performance expectancy (β=.883, *P*<.001) while being classified as a promising predictor. The studies suggest that task characteristics do not have a significant impact on the fit between the task and the technology. However, technology characteristics are a promising and significant predictor (β=.554, *P*<.001).

XSL•FO
RenderX

Figure 6. Weight and meta-analysis for the Task–Technology Fit model. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Accordingly, higher weights are represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels. ns: not significant.

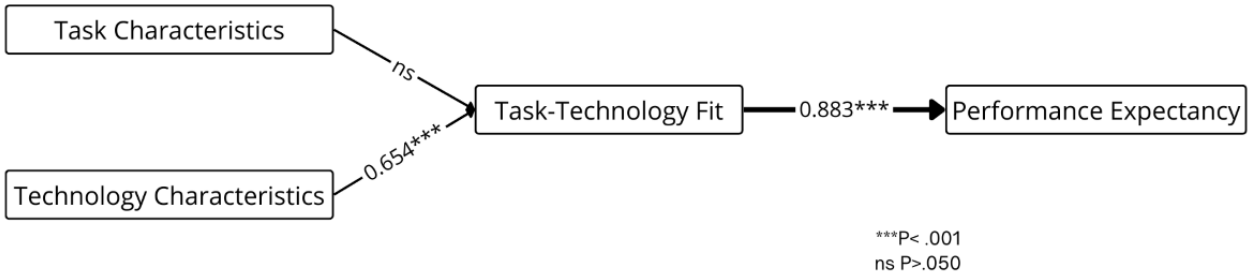

Figure 7 presents the weight and meta-analysis of the Privacy Calculus Theory, which explores the trade-off between benefits and privacy [52]. The results indicate that trust (β=.294, *P*<.001) has a significant positive influence on behavioral intention and is classified as the best predictor. Privacy and security (β=–.110, *P*=.02) exhibits a significant negative effect on behavioral intention; however, the weight is small, suggesting that privacy and security concerns may not be a strong inhibitor of adoption.

XSL•FO
RenderX

Figure 7. Weight and meta-analysis for Privacy Calculus Theory. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Accordingly, higher weights are represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels.

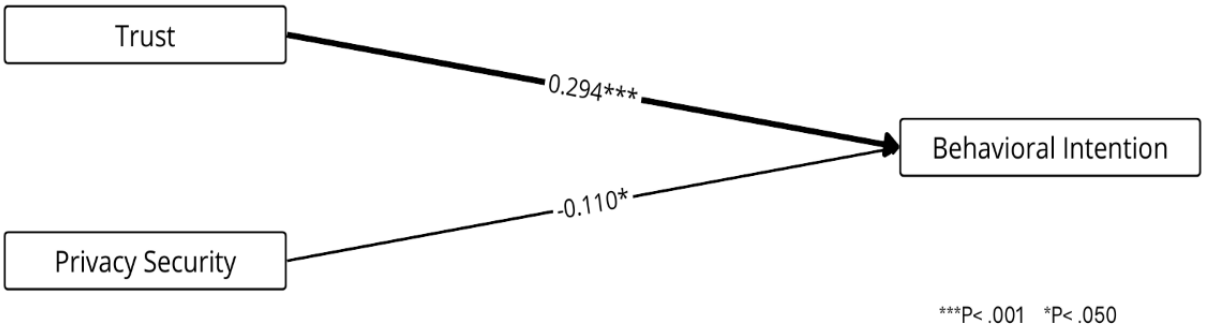

Figure 8 presents the weight and meta-analysis of the Theory of Planned Behavior, which posits that attitude, subjective norms, and behavioral control influence behavioral intention and, subsequently, behavior [53]. Both attitude (β=.573, *P*<.001) and self-efficacy (β=.318, *P*<.001) are the best and strongest predictors of behavioral intention.

XSL•FO
RenderX

Figure 8. Weight and meta-analysis for the Theory of Planned Behavior. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Accordingly, higher weights are represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels.

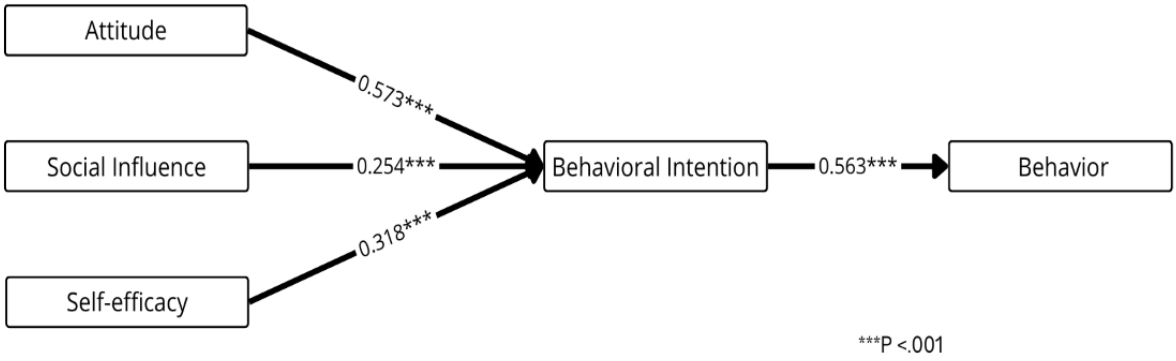

Figure 9 presents the weight and meta-analysis of the Innovation Diffusion Theory, which explains the diffusion of new technologies through 5 dimensions [54]. Relative advantage (performance expectancy; β=.339, *P*<.001), complexity (effort expectancy; β=.185, *P*<.001), and compatibility (β=.123, *P*<.001) were found to be significant predictors. Image was not a significant predictor, and trialability and observability have not been sufficiently studied in the literature.

XSL•FO
RenderX

Figure 9. Weight and meta-analysis for the Diffusion of Innovation theory. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Accordingly, higher weights are represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels.

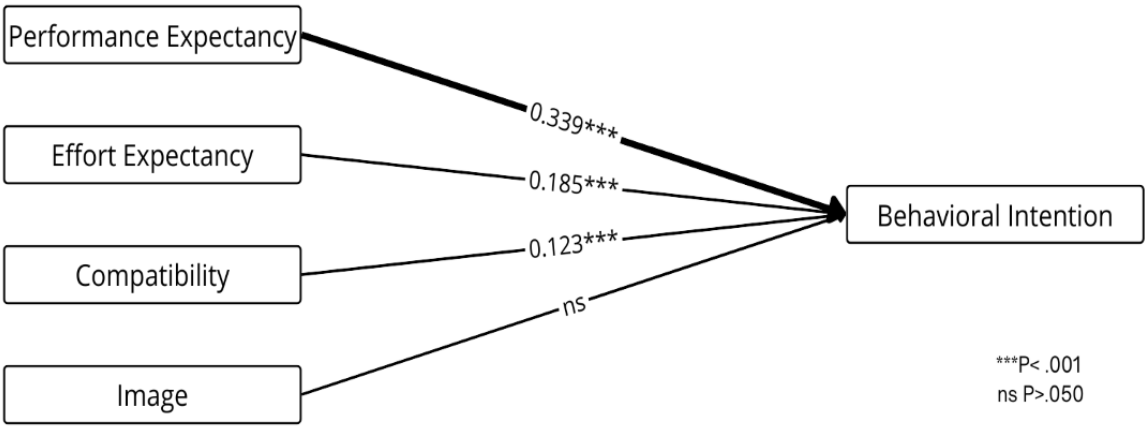

Evaluation of Publication Bias

This section evaluates the presence of publication bias and assesses the normality of the datasets used in the meta-analysis to ensure the reliability of the synthesized findings. Publication bias refers to the tendency for studies with significant or positive results to be more likely to be published, potentially skewing meta-analytic outcomes [55]. To ensure the robustness of our findings, we evaluated publication bias following the approach of Harrison et al [55], which suggests that a single criterion can provide a more sensitive and appropriate test. We focused our analysis on one of the most widely examined relationships in our dataset: the relationship between performance expectancy and behavioral intention, which was reported in 77 studies (Table 5).

XSL•FO
RenderX

Table 5. Studies (n=77) showing the effect size (z), SE (z), sample size, z score, Q component, significance of the paths between performance expectancy and behavioral intention, and the country.

Subgroup and effect size (*Z*)	SE (*Z*)	Sample size	*z* score	*Q* component	Significance	Country
Group 1						
0.268	0.051	387	5.965	2.350	Significant	China
0.604	0.050	397	13.455	26.161	Significant	China
0.387	0.080	158	7.933	0.258	Significant	China
0.165	0.046	469	3.718	15.264	Significant	China
0.086	0.053	357	1.908	23.979	Significant	China
0.230	0.051	386	5.113	5.199	Significant	China
0.727	0.065	243	15.616	34.684	Significant	China
0.224	0.036	769	5.122	11.555	Significant	China
0.186	0.058	304	4.074	7.740	Significant	China
0.149	0.113	81	2.702	3.039	Nonsignificant	China
0.333	0.071	201	7.018	0.037	Significant	China
0.472	0.050	406	10.530	6.373	Significant	China
0.205	0.056	325	4.504	6.462	Significant	China
0.198	0.054	345	4.362	7.588	Significant	China
0.217	0.086	139	4.372	2.267	Significant	China
0.950	0.084	146	19.257	52.169	Significant	China
0.406	0.065	237	8.704	0.827	Significant	China
0.471	0.072	197	9.914	3.014	Significant	China
0.519	0.065	239	11.139	7.032	Significant	China
0.180	0.047	462	4.039	12.734	Significant	China
0.105	0.071	201	2.223	11.509	Nonsignificant	China
0.289	0.040	624	6.567	2.068	Significant	China
0.286	0.064	247	6.145	0.907	Significant	China
0.377	0.039	668	8.591	0.615	Significant	China
0.446	0.051	386	9.927	3.830	Significant	China
0.157	0.043	552	3.560	19.651	Nonsignificant	China
0.258	0.047	450	5.774	3.535	Significant	China
0.401	0.096	111	7.772	0.324	Significant	China
1.040	0.077	171	21.531	80.869	Significant	China
0.586	0.028	1292	13.583	73.940	Significant	China
0.224	0.046	475	5.028	7.120	Significant	China
0.236	0.037	725	5.401	8.764	Significant	China
Group 2						

XSL•FO
RenderX

Subgroup and effect size (*Z*)	SE (*Z*)	Sample size	*z* score	*Q* component	Significance	Country
0.422	0.033	913	9.720	5.253	Significant	Bangladesh
0.586	0.075	181	12.212	10.210	Significant	France
0.132	0.062	267	2.855	12.172	Nonsignificant	France
0.202	0.125	67	3.478	1.342	Nonsignificant	France
0.633	0.070	206	13.382	16.646	Significant	Germany, the United States, the United Kingdom, and Canada
0.102	0.056	320	2.249	18.892	Nonsignificant	Ghana
0.319	0.039	645	7.273	0.469	Significant	Ghana
0.401	0.050	400	8.939	1.190	Significant	India
0.421	0.086	139	8.473	0.761	Significant	India
0.283	0.043	534	6.404	2.116	Significant	India
0.150	0.052	372	3.330	14.228	Significant	India
0.119	0.082	153	2.418	7.793	Nonsignificant	India
0.306	0.065	238	6.569	0.381	Significant	India
0.415	0.036	772	9.512	3.647	Significant	Indonesia
0.549	0.054	341	12.120	13.905	Significant	Iraq
0.245	0.069	212	5.191	2.162	Significant	Italy
0.413	0.066	233	8.841	1.017	Significant	Japan
0.265	0.071	200	5.587	1.307	Significant	Jordan
0.205	0.051	389	4.556	7.746	Significant	Korea
0.321	0.080	159	6.572	0.105	Significant	Korea
0.084	0.029	1158	1.948	79.454	Nonsignificant	Korea
0.239	0.060	280	5.209	3.172	Significant	Nepal
0.190	0.063	259	4.112	6.247	Significant	Oman
0.321	0.045	495	7.220	0.331	Significant	Pakistan
0.284	0.046	486	6.401	1.860	Significant	The Kingdom of Saudi Arabia
0.229	0.046	473	5.145	6.496	Significant	The Kingdom of Saudi Arabia
0.189	0.063	256	4.085	6.256	Significant	The Kingdom of Saudi Arabia
0.553	0.046	477	12.441	20.278	Significant	South Korea
0.574	0.045	487	12.910	24.967	Significant	South Korea
0.523	0.034	877	12.020	27.229	Significant	South Korea
0.518	0.097	110	10.014	3.141	Significant	Switzerland
0.485	0.055	335	10.682	6.343	Significant	Taiwan
0.393	0.061	268	8.519	0.575	Significant	Taiwan
0.149	0.047	458	3.346	17.727	Significant	Taiwan
0.321	0.091	125	6.340	0.082	Significant	Taiwan
0.245	0.113	81	4.436	0.807	Significant	Taiwan
0.688	0.065	243	14.796	28.070	Significant	Turkey
0.041	0.049	426	0.917	39.467	Nonsignificant	Turkey
0.230	0.099	106	4.416	1.398	Significant	The United Arab Emirates

XSL•FO
RenderX

Subgroup and effect size (*Z*)	SE (*Z*)	Sample size	*z* score	*Q* component	Significance	Country
0.141	0.050	407	3.143	17.070	Nonsignificant	The United States
1.157	0.052	376	25.681	244.930	Significant	The United States
0.167	0.060	277	3.619	8.873	Significant	The United States
0.266	0.092	120	5.227	0.756	Significant	The United States
0.848	0.047	450	19.012	112.411	Significant	The United States
0.321	0.056	322	7.045	0.215	Significant	Worldwide

To assess the presence of small-study effects and potential publication bias, a funnel plot was generated, and an Egger regression test was conducted. The funnel plot was constructed to visually evaluate publication bias [42], with the SE plotted on the y-axis instead of sample size, as this enhances the detection of asymmetry [36]. In an ideal funnel plot, symmetry is expected, with smaller studies exhibiting larger SE scattered evenly on both sides of the pooled effect size. In Figure 10, the studies display a somewhat asymmetrical distribution. Larger studies cluster near the combined effect size at the top of the funnel, while smaller studies show greater dispersion, potentially indicating publication bias or underlying heterogeneity. The trim-and-fill method estimates the number of potentially missing studies—often those with nonsignificant or negative results—and imputes them to generate an adjusted combined effect size. In this case, the imputed effect size is slightly smaller than the original estimate, suggesting that the observed meta-analytic effect may be modestly inflated due to the absence of smaller, less favorable studies.

Figure 10. Funnel plot of studies examining the relationship between performance expectancy and behavioral intention.

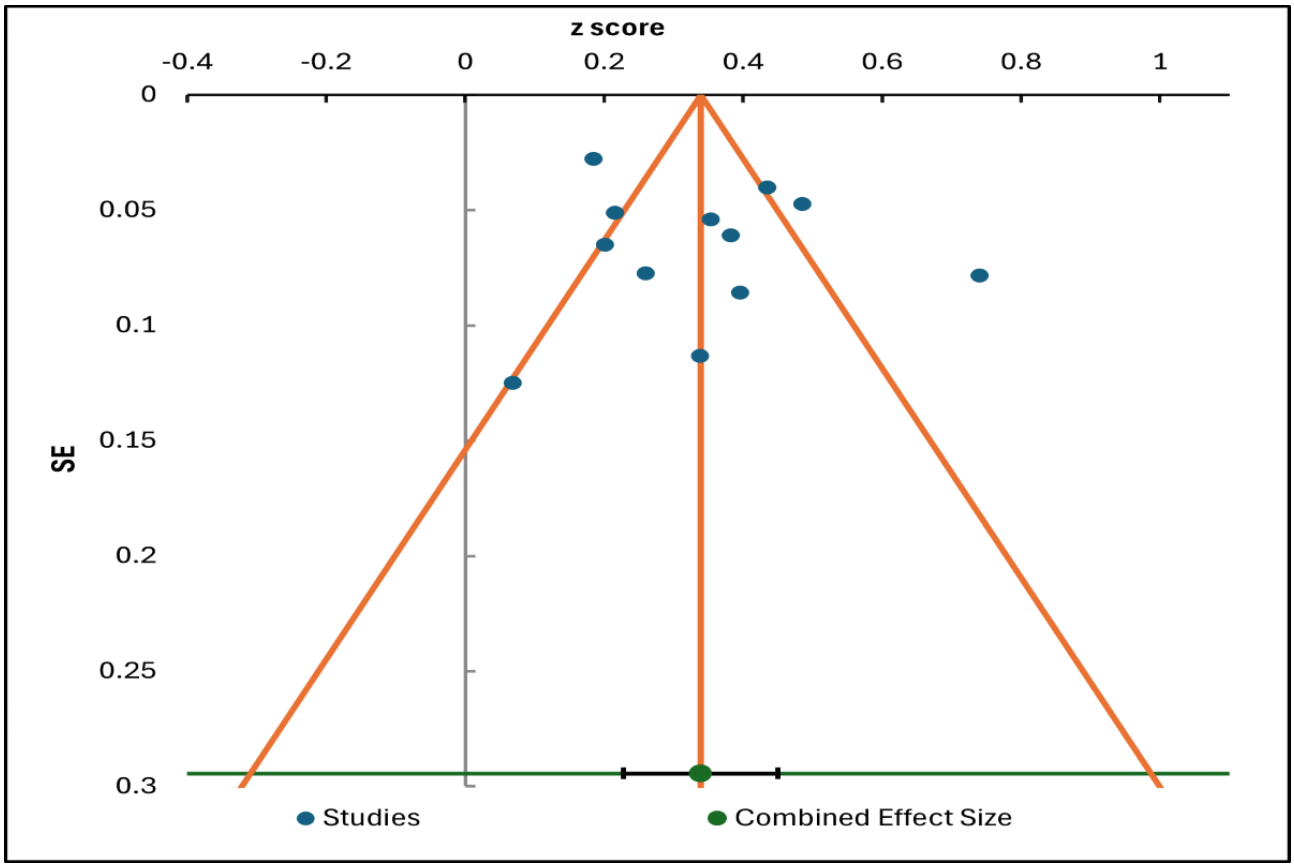

To statistically assess funnel plot asymmetry, we applied the Egger regression test [42] to evaluate whether smaller studies tend to report larger effect sizes, which can indicate potential publication bias (see Table 6). The test examines the relationship between effect sizes and their SEs to detect small-study effects. The results of the regression analysis showed that the intercept was not statistically significant (α=.381, *P*=.81), indicating no evidence of funnel plot asymmetry or publication bias. However, the slope coefficient was statistically significant (β=.327, *P*<.001), suggesting a positive association between study precision and effect size. While this does not indicate publication bias, it may reflect genuine heterogeneity among the included studies.

The I^2 statistic, which quantifies the proportion of total variation across studies attributable to true heterogeneity rather than sampling error [41], revealed a very high degree of heterogeneity (I^2>93%). This indicates that most of the variability in effect sizes reflects real differences across studies rather than random

XSL•FO
RenderX

sampling error. These differences may arise from variations in study design, measurement tools, participant demographics, cultural contexts, or theoretical frameworks used across the included studies. To further explore the sources of heterogeneity, a subgroup analysis was conducted.

Table 6. Egger - type test for small - study bias, using Excel's Data Analysis Regression Tool, which uses standard ordinary least squares regression.

Regression	Coefficients	SE	*t* test (*df*)	*P* value	95% CI
Intercept (α)	0.381	1.544	0.247 (75)	.81	–2.695 to 3.457
Slope (β)	0.327	0.081	4.044 (75)	<.001	0.166 to 0.489

The subgroup analysis examined whether effect sizes differed between studies conducted in China and those conducted in other countries. By comparing studies from China with those from other regions, we aimed to evaluate potential regional biases, given that a large proportion of the included studies were conducted in China. The results, presented in Table 7, indicate that geographic location has little influence on the overall effect size, as both subgroups exhibit similar results. The combined effect size for studies conducted in China is 0.340 (95% CI 0.272-0.404), while for studies in other countries, it is 0.336 (95% CI 0.279-0.390). However, heterogeneity remained very high in both groups (China: I^2=93%; other countries: I^2=94%), indicating substantial variability even within each subgroup. Therefore, the subgroup analysis addresses concerns about potential bias from the large proportion of studies conducted in China, confirming that the results are largely stable across regions.

Table 7. Subgroup comparison between China and the other countries in our sample.

Subgroup	Effect size	*P* value	95% CI	I^2 (%)
China	0.340	<.001	0.272-0.404	92.984
Other countries	0.336	<.001	0.279-0.390	94.355

Discussion

Principal Findings

The study of IoT adoption in health care reveals a diverse landscape of constructs and relationships, providing a comprehensive overview of the factors driving IoT adoption. This study synthesized findings from 109 papers and 115 datasets across various regions, including China, South Korea, the United States, and India, with most studies published in high-ranking journals. The combined weight and meta-analysis identified the best predictors and examined the adoption models most frequently used in IoT health care. Figure 11 highlights the strongest and most consistent predictors, integrating the results of both the meta-analysis and weight analysis.

XSL•FO
RenderX

Figure 11. Best predictors identified in the weight analysis, along with their statistically significant mean β coefficients from the meta-analysis. The strength (weight) of each predictor is represented by the thickness of the line connecting the predictor to the target variable, with thicker lines indicating stronger predictors. The green box denotes constructs from the Unified Theory of Acceptance and Use of Technology framework, and the orange box denotes constructs from Technology Acceptance Model.

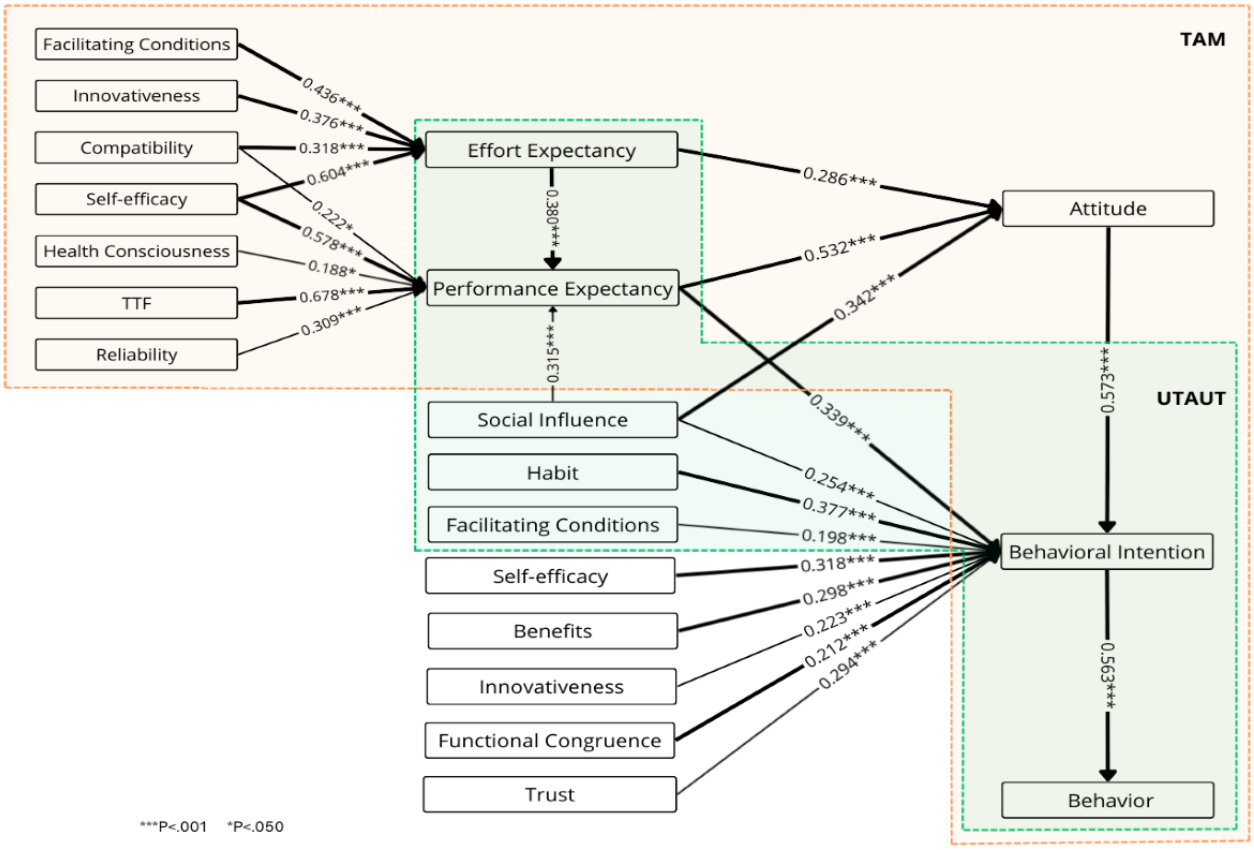

Technology acceptance models, such as UTAUT and TAM, have been widely and successfully applied in the context of IoT in health care [56,57], which is unsurprising given that these are the most commonly used technology adoption models, highly cited, and successfully applied across diverse fields and contexts [58,59]. Compared with models such as UTAUT and TAM, where predictors such as performance expectancy and facilitating conditions consistently exhibit strong and reliable effects, the HBM and PMT models struggle to establish robust relationships with behavioral intention. This suggests that relying solely on health-related constructs or health behavior models may not be sufficient to explain health care technology adoption. Therefore, other individual factors, such as innovativeness, external factors, such as social influence, and technological factors, such as performance expectancy, may play a more decisive role [60,61]. Integrating context-specific health variables into robust models such as TAM or UTAUT can, however, provide additional insights. For example, individuals with strong health motivation or health consciousness tend to exhibit higher levels of performance expectancy from IoT health care technologies [62,63].

The findings highlight several key factors influencing effort expectancy and performance expectancy, both of which are central to users' attitudes toward technology. For effort expectancy, the most influential factor was self-efficacy, indicating that individuals who feel more confident in their ability to use the technology tend to perceive it as easier to operate [64,65]. Other important contributors include facilitating conditions, innovativeness, and compatibility, suggesting that a supportive environment, openness to new technologies, and alignment with users' existing values and practices all help reduce the perceived effort required to use IoT in health care [66,67]. For performance expectancy, TTF emerged as the dominant influence, highlighting that when users perceive a strong alignment between the technology and the tasks they need to perform, they are more likely to view it as useful [68,69]. Additionally, health consciousness, self-efficacy, reliability, and compatibility played significant roles, emphasizing the importance of personal health concerns, confidence in usage, trust in the system's dependability, and alignment with users' existing values and practices [63,70]. Together, these findings underscore the relevance of both individual and contextual factors in shaping users' perceptions of a technology's usefulness and ease of use.

Regarding individuals' IoT health care technology adoption journey, the findings reveal that a positive attitude is crucial for successful adoption [71,72]. Efforts to cultivate positive perceptions can be made by leveraging the influence of important figures in individuals' lives and by emphasizing the ease of use and the potential for improved health care outcomes [73-75]. When individuals hold a positive perception of IoT health care, they are more likely to intend to use it, which in turn positively influences actual usage [76,77]. To further enhance behavioral intention, the effectiveness of IoT health

XSL•FO
RenderX

care solutions and the encouragement of health care professionals, family, and friends should be leveraged [78,79]. Additionally, individuals' willingness to try new technologies plays a significant role, as more innovative users are more likely to adopt IoT solutions [80,81].

Trust plays a decisive role in shaping behavioral intentions, reinforcing the notion that users are willing to trade some level of privacy if they perceive a system as secure and reliable [60]. Previous literature has found that individuals are often reluctant to adopt digital health or IoT technologies when they do not trust the provider [82,83]. This perspective may help contextualize the inconsistent results observed for privacy as a predictor of behavioral intention, as a notable proportion of studies reported nonsignificant relationships. Similarly, predictors such as barriers, vulnerability, and financial cost also exhibited higher frequencies of nonsignificant findings in our analysis. These inconsistencies may reflect how these constructs interact with—or are influenced by—the presence of stronger enabling factors. For instance, high perceived usefulness and trust may diminish the observed effects of barriers such as financial cost and privacy, as these factors may become less salient in users' perceptions.

Theoretical Implications

This study makes several contributions to the theoretical understanding of IoT health care adoption by synthesizing findings from diverse quantitative studies and adoption models. The results reinforce the importance of established models such as TAM and UTAUT. They also suggest that integrating variables from other theories—such as health consciousness, innovativeness, and trust—into traditional technology acceptance frameworks can provide deeper insights into how individuals adopt IoT in health care. Behavioral intention is the most studied target variable, while attitude and actual behavior remain underexplored, indicating a gap in existing research on these critical components of the adoption process.

Researchers should further investigate several promising but underexplored predictors that showed perfect weight, suggesting strong yet preliminary evidence of their relevance, to establish their broader applicability. For instance, regarding behavioral intention, the aesthetic appeal of health care technologies shows potential as a strong predictor. For actual behavior, facilitating conditions and social influence are promising predictors that warrant further exploration. In the case of the underexplored TTF theory, technology characteristics appear to be a promising predictor. For performance expectancy, convenience and innovativeness are promising predictors, while for effort expectancy, reliability and TTF show potential as predictors deserving additional investigation.

By contrast, several predictors demonstrated limited or inconsistent relevance to the adoption of IoT in health care. For the outcome attitude, barriers did not have a statistically significant effect. For behavioral intention, predictors such as privacy and security, barriers, vulnerability, severity, compatibility, and financial cost produced inconsistent findings, with many studies reporting nonsignificant results. Specifically, health, technology anxiety, financial cost, and barriers were frequently not significant predictors of behavioral intention. When predicting actual behavior, variables such as social influence, innovativeness, health consciousness, vulnerability, and effort expectancy often failed to reach statistical significance. Regarding performance expectancy, both privacy and security and barriers were not consistently significant, and for effort expectancy, privacy and security and image did not show meaningful effects.

Our findings indicate that regional differences alone do not fully explain the heterogeneity of results. Therefore, when applying the findings of this study, we recommend refining theoretical models to account for contextual factors and implementing practical strategies aligned with the strongest predictors identified, such as performance expectancy and self-efficacy, which can enhance adoption across different settings. Future adoption studies would benefit from incorporating context-specific factors that capture cultural and health care system differences, enabling a better understanding of how these contextual variables influence target outcomes, as either control or moderator variables.

Several regions, particularly in Africa, South America, and Europe, remain underrepresented, highlighting a gap in the literature and the need for future research in diverse settings to improve the generalizability and equity of evidence regarding IoT adoption in health care. Finally, combining qualitative methods, such as interviews and focus groups, is recommended to gain deeper insights. This mixed methods approach can provide a better understanding of user perceptions and experiences, bridging the gap between quantitative results and the complex realities of technology adoption [84,85].

Practical Implications

The findings of this study provide actionable insights for practitioners, policy makers, and technology developers seeking to enhance IoT health care adoption. Key drivers—such as performance expectancy, self-efficacy, social influence, functional congruence, trust, habit, facilitating conditions, benefits, and innovativeness—consistently shape behavioral intention. For example, developers can focus on creating intuitive designs and user-friendly interfaces while emphasizing tangible performance benefits. Health care providers and policy makers can leverage trusted individuals, such as doctors and family members, to encourage adoption.

The availability of resources and infrastructure that enable and support technology use—such as access to devices, technical support, internet connectivity, and integration with health care systems—plays an important role in adoption, as it reduces barriers for individuals starting to use IoT in health care [69,78]. Furthermore, adhering to robust data protection frameworks that ensure transparency from all entities handling health-related data aligns implementation with national and regional regulatory standards and fosters user trust [83,86]. Finally, targeting innovative individuals who are more likely to adopt IoT health care technologies or who already have the habits and skills to use them can further promote technology adoption.

Limitations and Future Research

This study has several limitations that warrant consideration. Our findings reveal a high level of heterogeneity, which is not

XSL•FO
RenderX

fully explained by regional differences; therefore, the pooled estimates should be interpreted with caution. Future research should investigate additional factors that may account for this heterogeneity, such as study design, population characteristics, or model specification, and conduct moderator analyses to better address variability. Additionally, China accounts for a large proportion of the included studies, while several regions—particularly in Africa, South America, and Europe—remain underrepresented. As such, we caution against overgeneralizing our findings to all global contexts. Additionally, while this study synthesizes quantitative findings, it excludes qualitative research, which could provide deeper insights into contextual variability and user experiences influencing adoption. This exclusion may contribute to inconsistencies in the evidence, particularly for understudied predictors such as privacy concerns, perceived vulnerability, and financial cost, which often showed nonsignificant results. Future research should consider integrative literature reviews that include qualitative studies to better capture the nuanced interplay of individual, cultural, and technological factors.

Conclusions

Our comprehensive meta- and weight analysis of 115 unique datasets on IoT health care adoption revealed several significant predictors for the adoption of IoT health care technologies. Behavioral intention emerged as the most frequently studied target variable. By contrast, attitude, actual behavior, performance expectancy, effort expectancy, and TTF remain comparatively understudied, with very few paths examined more than 5 times. While adoption theories from the information systems field, such as UTAUT and TAM, are predominantly used, integrating context-specific factors or combining constructs from different theoretical models can provide deeper insights into IoT health care adoption and further support the adoption process.

All the best predictors identified in our study were statistically significant, with the exception of reliability as a predictor of behavioral intention. For the target variable attitude, the strongest predictors were effort expectancy, performance expectancy, and social influence, while barriers did not have a statistically significant effect. Regarding behavioral intention, the most consistent and significant predictors were attitude, performance expectancy, habit, self-efficacy, functional congruence, reliability, and benefits. In addition, social influence, facilitating conditions, and trust demonstrated strong weights above 0.700, while aesthetic appeal was considered a promising predictor due to the limited number of studies. Conversely, variables such as privacy and security, barriers, vulnerability, severity, compatibility, and financial cost showed inconsistent results, with a high incidence of statistically nonsignificant findings. Specifically, health, technology anxiety, financial cost, and barriers were not statistically significant predictors of behavioral intention.

For actual behavior, behavioral intention emerged as the best predictor, while facilitating conditions and social influence were considered promising. However, social influence, innovativeness, health consciousness, vulnerability, and effort expectancy did not reach statistical significance for behavior. Regarding performance expectancy, effort expectancy, TTF, and self-efficacy were the best predictors, followed by health consciousness, social influence, reliability, and compatibility as strong predictors, while convenience and innovativeness appeared as promising. Privacy and security and barriers, however, were not statistically significant predictors of performance expectancy. For effort expectancy, the most consistent predictors were facilitating conditions, innovativeness, self-efficacy, and compatibility, with reliability and TTF considered promising predictors; privacy and security and image did not show significant effects. Lastly, for the target variable TTF, technology characteristics emerged as a promising predictor, whereas task characteristics were not statistically significant.

Acknowledgments

We thank the European Union’s Horizon Europe program and Fundação para a Ciência e a Tecnologia (FCT) for their support through the program UIDB/04152 – Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS.

Funding

This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia) under the project UIDB/04152 – Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS. It also resulted from the TwinAIR project, which received funding from the European Union’s Horizon Europe program under grant agreement No. 101057779.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author (IV) on reasonable request.

Authors' Contributions

Conceptualization: IV, MNZ
Data curation: IV, MNZ
Formal analysis: IV
Funding acquisition: SK

XSL•FO
RenderX

Investigation: IV
Methodology: IV
Project administration: SK
Supervision: MNZ, RM, TO
Validation: MNZ, RM
Writing – original draft: IV
Writing – review & editing: IV, MNZ, RM, TO

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 177 KB - jmir_v28i1e64091_app1.docx]

References

1. Krishnamoorthy S, Dua A, Gupta S. Role of emerging technologies in future IoT-driven Healthcare 4.0 technologies: a survey, current challenges and future directions. J Ambient Intell Human Comput 2021 May 12;14(1):361-407. [doi: 10.1007/s12652-021-03302-w]
2. Mohamad Jawad HH, Bin Hassan Z, Zaidan BB, Mohammed Jawad FH, Mohamed Jawad DH, Alredany WHD. A systematic literature review of enabling IoT in healthcare: motivations, challenges, and recommendations. Electronics 2022 Oct 08;11(19):3223. [doi: 10.3390/electronics11193223]
3. Ashraf S. A proactive role of IoT devices in building smart cities. Internet of Things and Cyber-Physical Systems 2021;1:8-13. [doi: 10.1016/j.iotcps.2021.08.001]
4. Lin Q, Zhao Q. IoT applications in healthcare. In: García Márquez FP, Lev B, editors. Internet of Things: Cases and Studies. Cham, Switzerland: Springer; Jul 14, 2021:115-133.
5. Singh RP, Javaid M, Haleem A, Suman R. Internet of things (IoT) applications to fight against COVID-19 pandemic. Diabetes Metab Syndr 2020;14(4):521-524 [FREE Full text] [doi: 10.1016/j.dsx.2020.04.041] [Medline: 32388333]
6. Tuli S, Basumatary N, Gill SS, Kahani M, Arya RC, Wander GS, et al. HealthFog: an ensemble deep learning based smart healthcare system for automatic diagnosis of heart diseases in integrated IoT and fog computing environments. Future Generation Computer Systems 2020 Mar;104:187-200. [doi: 10.1016/j.future.2019.10.043]
7. Farahani B, Firouzi F, Chang V, Badaroglu M, Constant N, Mankodiya K. Towards fog-driven IoT eHealth: promises and challenges of IoT in medicine and healthcare. Future Generation Computer Systems 2018 Jan;78:659-676. [doi: 10.1016/j.future.2017.04.036]
8. Li X, Dai H, Wang Q, Imran M, Li D, Imran MA. Securing Internet of Medical Things with friendly-jamming schemes. Comput Commun 2020 Jul 01;160:431-442 [FREE Full text] [doi: 10.1016/j.comcom.2020.06.026] [Medline: 32834198]
9. Ye J. The role of health technology and informatics in a global public health emergency: practices and implications from the COVID-19 pandemic. JMIR Med Inform 2020 Jul 14;8(7):e19866 [FREE Full text] [doi: 10.2196/19866] [Medline: 32568725]
10. Philip NY, Rodrigues JJPC, Wang H, Fong SJ, Chen J. Internet of Things for in-home health monitoring systems: current advances, challenges and future directions. IEEE J Select Areas Commun 2021 Feb;39(2):300-310. [doi: 10.1109/jsac.2020.3042421]
11. Osama M, Ateya AA, Sayed MS, Hammad M, Pławiak P, Abd El-Latif AA, et al. Internet of Medical Things and Healthcare 4.0: trends, requirements, challenges, and research directions. Sensors (Basel) 2023 Aug 25;23(17):7435 [FREE Full text] [doi: 10.3390/s23177435] [Medline: 37687891]
12. Vyas S, Bhargava D, Khan S, Bhargava D. Healthcare 4.0: a systematic review and its impact over conventional healthcare system. In: Artificial Intelligence for Health 4.0: Challenges and Applications. London, UK: Routledge; Jan 2023:1-17.
13. Healthcare IoT - worldwide. Statista. 2024. URL: https://www.statista.com/outlook/tmo/internet-of-things/healthcare-iot/worldwide [accessed 2024-06-27]
14. Kelly JT, Campbell KL, Gong E, Scuffham P. The Internet of Things: impact and implications for health care delivery. J Med Internet Res 2020 Nov 10;22(11):e20135 [FREE Full text] [doi: 10.2196/20135] [Medline: 33170132]
15. Yang F, Shu H, Zhang X. Understanding "internet plus healthcare" in China: policy text analysis. J Med Internet Res 2021 Jul 26;23(7):e23779 [FREE Full text] [doi: 10.2196/23779] [Medline: 34309581]
16. Liang J, Li Y, Zhang Z, Shen D, Xu J, Yu G, et al. Evaluating the applications of health information technologies in China during the past 11 years: consecutive survey data analysis. JMIR Med Inform 2020 Feb 10;8(2):e17006 [FREE Full text] [doi: 10.2196/17006] [Medline: 32039815]
17. Zheng X, Rodríguez-Monroy C. The development of intelligent healthcare in China. Telemed J E Health 2015 May;21(5):443-448. [doi: 10.1089/tmj.2014.0102] [Medline: 25671730]

XSL•FO
RenderX

18. Bhuiyan MN, Rahman MM, Billah MM, Saha D. Internet of Things (IoT): a review of its enabling technologies in healthcare applications, standards protocols, security, and market opportunities. IEEE Internet Things J 2021 Jul 1;8(13):10474-10498. [doi: 10.1109/jiot.2021.3062630]
19. Okpala P. Assessment of the influence of technology on the cost of healthcare service and patient's satisfaction. International Journal of Healthcare Management 2017 Jun 07;11(4):351-355. [doi: 10.1080/20479700.2017.1337623]
20. Chintamaneni S, Yatham P, Stumbar S. From East to West: a narrative review of healthcare models in India and the United States. Cureus 2023 Aug;15(8):e43456 [FREE Full text] [doi: 10.7759/cureus.43456] [Medline: 37711922]
21. Shafique K, Khawaja BA, Sabir F, Qazi S, Mustaqim M. Internet of Things (IoT) for next-generation smart systems: a review of current challenges, future trends and prospects for emerging 5G-IoT scenarios. IEEE Access 2020;8:23022-23040. [doi: 10.1109/access.2020.2970118]
22. Kouroubali A, Katehakis DG. The new European interoperability framework as a facilitator of digital transformation for citizen empowerment. J Biomed Inform 2019 Jun;94:103166 [FREE Full text] [doi: 10.1016/j.jbi.2019.103166] [Medline: 30978512]
23. Hoque R, Sorwar G. Understanding factors influencing the adoption of mHealth by the elderly: an extension of the UTAUT model. Int J Med Inform 2017 May;101:75-84. [doi: 10.1016/j.ijmedinf.2017.02.002] [Medline: 28347450]
24. Calegari LP, R.D. B, Fettermann DC. A meta-analysis of a comprehensive m-health technology acceptance. IJLSS 2023 Apr 06;15(1):1-21. [doi: 10.1108/ijlss-01-2023-0012]
25. Binyamin SS, Zafar BA. Proposing a mobile apps acceptance model for users in the health area: a systematic literature review and meta-analysis. Health Informatics J 2021 Jan 13;27(1):1460458220976737 [FREE Full text] [doi: 10.1177/1460458220976737] [Medline: 33438494]
26. Chauhan S, Jaiswal M. A meta-analysis of e-health applications acceptance. JEIM 2017 Mar 06;30(2):295-319. [doi: 10.1108/jeim-08-2015-0078]
27. Shen Y, Xu W, Liang A, Wang X, Lu X, Lu Z, et al. Online health management continuance and the moderating effect of service type and age difference: a meta-analysis. Health Informatics J 2022 Aug 17;28(3):14604582221119950 [FREE Full text] [doi: 10.1177/14604582221119950] [Medline: 35976977]
28. Gopinath K, Selvam G, Narayanamurthy G. Determinants of the adoption of wearable devices for health and fitness: a meta-analytical study. CAIS 2022;50(1):445-450 [FREE Full text] [doi: 10.17705/1cais.05019]
29. Zhang Z, Xia E, Huang J. Impact of the moderating effect of national culture on adoption intention in wearable health care devices: meta-analysis. JMIR Mhealth Uhealth 2022 Jun 03;10(6):e30960 [FREE Full text] [doi: 10.2196/30960] [Medline: 35657654]
30. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
31. Jeyaraj A, Rottman JW, Lacity MC. A review of the predictors, linkages, and biases in IT innovation adoption research. Journal of Information Technology 2006 Feb 01;21(1):1-23. [doi: 10.1057/palgrave.jit.2000056]
32. Baptista G, Oliveira T. A weight and a meta-analysis on mobile banking acceptance research. Computers in Human Behavior 2016 Oct;63:480-489. [doi: 10.1016/j.chb.2016.05.074]
33. Naranjo Zolotov M, Oliveira T, Casteleyn S. E-participation adoption models research in the last 17 years: A weight and meta-analytical review. Computers in Human Behavior 2018 Apr;81:350-365. [doi: 10.1016/j.chb.2017.12.031]
34. Bowman NA. Effect sizes and statistical methods for meta-analysis in higher education. Res High Educ 2011 Jul 13;53(3):375-382. [doi: 10.1007/s11162-011-9232-5]
35. Cooper H. Research Synthesis and Meta-Analysis: A Step-by-Step Approach. Thousand Oaks, CA: Sage Publications; 2017.
36. Borenstein M, Hedges L, Higgins J, Rothstein H, Ben Van Den A. Introduction to Meta-Analysis. Hoboken, NJ: Wiley; Sep 01, 2020.
37. Peterson RA, Brown SP. On the use of beta coefficients in meta-analysis. J Appl Psychol 2005 Jan;90(1):175-181. [doi: 10.1037/0021-9010.90.1.175] [Medline: 15641898]
38. Gupta N, Singh AK. Individual acceptance of Internet of Things: a meta analytical review. Int J Consumer Studies 2025 May 06;49(3):44. [doi: 10.1111/ijcs.70070]
39. DerSimonian R, Laird N. Meta-analysis in clinical trials revisited. Contemp Clin Trials 2015 Nov;45(Pt A):139-145 [FREE Full text] [doi: 10.1016/j.cct.2015.09.002] [Medline: 26343745]
40. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986 Sep;7(3):177-188 [FREE Full text] [doi: 10.1016/0197-2456(86)90046-2] [Medline: 3802833]
41. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002 Jun 15;21(11):1539-1558. [doi: 10.1002/sim.1186] [Medline: 12111919]
42. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634 [FREE Full text] [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
43. Suurmond R, van Rhee H, Hak T. Introduction, comparison, and validation of meta-essentials: a free and simple tool for meta-analysis. Res Synth Methods 2017 Dec 29;8(4):537-553 [FREE Full text] [doi: 10.1002/jrsm.1260] [Medline: 28801932]

XSL•FO
RenderX

44. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989 Sep;13(3):319. [doi: 10.2307/249008]
45. Venkatesh V, Davis FD. A theoretical extension of the Technology Acceptance Model: four longitudinal field studies. Management Science 2000 Feb;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
46. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425. [doi: 10.2307/30036540]
47. Venkatesh V, Thong JYL, Xu X. Consumer acceptance and use of information technology: extending the Unified Theory of Acceptance and Use of Technology. MIS Quarterly 2012;36(1):157. [doi: 10.2307/41410412]
48. Prentice-Dunn S, Rogers RW. Protection Motivation Theory and preventive health: beyond the Health Belief Model. Health Educ Res 1986;1(3):153-161. [doi: 10.1093/her/1.3.153]
49. Maddux JE, Rogers RW. Protection motivation and self-efficacy: a revised theory of fear appeals and attitude change. Journal of Experimental Social Psychology 1983 Sep;19(5):469-479. [doi: 10.1016/0022-1031(83)90023-9]
50. Janz NK, Becker MH. The Health Belief Model: a decade later. Health Educ Q 1984 Jan 01;11(1):1-47. [doi: 10.1177/109019818401100101] [Medline: 6392204]
51. Goodhue DL, Thompson RL. Task-technology fit and individual performance. MIS Quarterly 1995 Jun;19(2):213. [doi: 10.2307/249689]
52. Tang Y, Ning X. Understanding user misrepresentation behavior on social apps: the perspective of privacy calculus theory. Decision Support Systems 2023 Feb;165:113881. [doi: 10.1016/j.dss.2022.113881]
53. Ajzen I. From intentions to actions: a theory of planned behavior. In: Kuhl J, Beckmann J, editors. Action Control. Berlin/Heidelberg, Germany: Springer Publishing Company; 1985:11-39.
54. Rogers EM. Diffusion of Innovations, 5th Edition. New York, NY: Free Press; Aug 16, 2003.
55. Harrison JS, Banks GC, Pollack JM, O'Boyle EH, Short J. Publication bias in strategic management research. Journal of Management 2016 Jul 10;43(2):400-425. [doi: 10.1177/0149206314535438]
56. Yang HJ, Lee J, Lee W. Factors influencing health care technology acceptance in older adults based on the Technology Acceptance Model and the Unified Theory of Acceptance and Use of Technology: meta-analysis. J Med Internet Res 2025 Mar 28;27:e65269 [FREE Full text] [doi: 10.2196/65269] [Medline: 40153796]
57. Al-Rawashdeh M, Keikhosrokiani P, Belaton B, Alawida M, Zwiri A. IoT adoption and application for smart healthcare: a systematic review. Sensors (Basel) 2022 Jul 19;22(14):5377. [doi: 10.3390/s22145377] [Medline: 35891056]
58. Blut M, Chong AYL, Tsigna Z, Venkatesh V. Meta-analysis of the Unified Theory of Acceptance and Use of Technology (UTAUT): challenging its validity and charting a research agenda in the Red Ocean. JAIS 2022;23(1):13-95. [doi: 10.17705/1jais.00719]
59. Jeyaraj A, Dwivedi YK, Venkatesh V. Intention in information systems adoption and use: current state and research directions. International Journal of Information Management 2023 Dec;73:102680. [doi: 10.1016/j.ijinfomgt.2023.102680]
60. Al-Rawashdeh M, Keikhosrokiani P, Belaton B, Alawida M, Zwiri A. Effective factors for the adoption of IoT applications in nursing care: a theoretical framework for smart healthcare. Journal of Building Engineering 2024 Jul;89:109012. [doi: 10.1016/j.jobe.2024.109012]
61. Kim TB, Ho CB. Validating the moderating role of age in multi-perspective acceptance model of wearable healthcare technology. Telematics and Informatics 2021 Aug;61:101603. [doi: 10.1016/j.tele.2021.101603]
62. Cao J, Kurata K, Lim Y, Sengoku S, Kodama K. Social acceptance of mobile health among young adults in Japan: an extension of the UTAUT model. Int J Environ Res Public Health 2022 Nov 17;19(22):15156 [FREE Full text] [doi: 10.3390/ijerph192215156] [Medline: 36429875]
63. Hidayat-ur-Rehman I, Ahmad A, Akhter F, Aljarallah A. A dual-stage SEM-ANN analysis to explore consumer adoption of smart wearable healthcare devices. Journal of Global Information Management (JGIM) 2021 Jan;29(6):30. [doi: 10.4018/JGIM.294123]
64. Kim J. Analysis of health consumers' behavior using self-tracker for activity, sleep, and diet. Telemed J E Health 2014 Jun;20(6):552-558. [doi: 10.1089/tmj.2013.0282] [Medline: 24745608]
65. Liu Y, Lu X, Zhao G, Li C, Shi J. Adoption of mobile health services using the unified theory of acceptance and use of technology model: self-efficacy and privacy concerns. Front Psychol 2022 Aug 11;13:944976 [FREE Full text] [doi: 10.3389/fpsyg.2022.944976] [Medline: 36033004]
66. Karahoca A, Karahoca D, Aksöz M. Examining intention to adopt to internet of things in healthcare technology products. Kybernetes 2017 Dec 11;47(4):742-770. [doi: 10.1108/k-02-2017-0045]
67. Li J, Ma Q, Chan AH, Man S. Health monitoring through wearable technologies for older adults: smart wearables acceptance model. Appl Ergon 2019 Feb;75:162-169. [doi: 10.1016/j.apergo.2018.10.006] [Medline: 30509522]
68. Misra S, Adtani R, Singh Y, Singh S, Thakkar D. Exploring the factors affecting behavioral intention to adopt wearable devices. Clinical Epidemiology and Global Health 2023 Nov;24:101428. [doi: 10.1016/j.cegh.2023.101428]
69. Kang H, Han J, Kwon GH. The acceptance behavior of smart home health care services in South Korea: an integrated model of UTAUT and TTF. Int J Environ Res Public Health 2022 Oct 14;19(20):13279 [FREE Full text] [doi: 10.3390/ijerph192013279] [Medline: 36293859]

XSL•FO
RenderX

70. Mensah IK. Understanding the drivers of Ghanaian citizens' adoption intentions of mobile health services. Front Public Health 2022;10:906106 [FREE Full text] [doi: 10.3389/fpubh.2022.906106] [Medline: 35774576]
71. Alraja M. Frontline healthcare providers' behavioural intention to Internet of Things (IoT)-enabled healthcare applications: A gender-based, cross-generational study. Technological Forecasting and Social Change 2022 Jan;174:121256. [doi: 10.1016/j.techfore.2021.121256]
72. Jeng M, Pai F, Yeh T. Antecedents for older adults' intention to use smart health wearable devices-technology anxiety as a moderator. Behav Sci (Basel) 2022 Apr 18;12(4):114 [FREE Full text] [doi: 10.3390/bs12040114] [Medline: 35447686]
73. Papa A, Mital M, Pisano P, Del Giudice M. E-health and wellbeing monitoring using smart healthcare devices: an empirical investigation. Technological Forecasting and Social Change 2020 Apr;153:119226 [FREE Full text] [doi: 10.1016/j.techfore.2018.02.018]
74. Sinha M, Fukey L, Balasubramanian K, Hanafiah MH, Kunasekaran P, Ragavan NA. Acceptance of consumer-oriented health information technologies (CHITs): integrating technology acceptance model with perceived risk. IJCAI 2021 Oct 05;45(6):45-52. [doi: 10.31449/inf.v45i6.3484]
75. Tsai T, Lin W, Chang Y, Chang P, Lee M. Technology anxiety and resistance to change behavioral study of a wearable cardiac warming system using an extended TAM for older adults. PLoS One 2020;15(1):e0227270 [FREE Full text] [doi: 10.1371/journal.pone.0227270] [Medline: 31929560]
76. Yang Q, Al Mamun A, Hayat N, Salleh MFM, Jingzu G, Zainol NR. Modelling the mass adoption potential of wearable medical devices. PLoS One 2022;17(6):e0269256 [FREE Full text] [doi: 10.1371/journal.pone.0269256] [Medline: 35675373]
77. Yin Z, Yan J, Fang S, Wang D, Han D. User acceptance of wearable intelligent medical devices through a modified unified theory of acceptance and use of technology. Ann Transl Med 2022 Jun;10(11):629 [FREE Full text] [doi: 10.21037/atm-21-5510] [Medline: 35813345]
78. Ben Arfi W, Ben Nasr I, Khvatova T, Ben Zaied Y. Understanding acceptance of eHealthcare by IoT natives and IoT immigrants: an integrated model of UTAUT, perceived risk, and financial cost. Technological Forecasting and Social Change 2021 Feb;163:120437. [doi: 10.1016/j.techfore.2020.120437]
79. Koo JH, Park YH, Kang DR. Factors predicting older people's acceptance of a personalized health care service app and the effect of chronic disease: cross-sectional questionnaire study. JMIR Aging 2023 Jun 21;6:e41429 [FREE Full text] [doi: 10.2196/41429] [Medline: 37342076]
80. Huang C, Yang M. Empirical investigation of factors influencing consumer intention to use an artificial intelligence-powered mobile application for weight loss and health management. Telemed J E Health 2020 Oct 01;26(10):1240-1251. [doi: 10.1089/tmj.2019.0182] [Medline: 31971883]
81. Cheung ML, Chau KY, Lam MHS, Tse G, Ho KY, Flint SW, et al. Examining consumers' adoption of wearable healthcare technology: the role of health attributes. Int J Environ Res Public Health 2019 Jun 26;16(13):16132257 [FREE Full text] [doi: 10.3390/ijerph16132257] [Medline: 31247962]
82. David A, Yigitcanlar T, Li RYM, Corchado JM, Cheong PH, Mossberger K, et al. Understanding local government digital technology adoption strategies: a PRISMA review. Sustainability 2023 Jun 15;15(12):9645. [doi: 10.3390/su15129645]
83. McGraw D, Mandl KD. Privacy protections to encourage use of health-relevant digital data in a learning health system. NPJ Digit Med 2021 Jan 04;4(1):2 [FREE Full text] [doi: 10.1038/s41746-020-00362-8] [Medline: 33398052]
84. Venkatesh V, Brown SA, Bala H. Bridging the qualitative-quantitative divide: guidelines for conducting mixed methods research in information systems. MIS Quarterly 2013;37(1):21-54. [doi: 10.25300/MISQ/2013/37.1.02]
85. Venkatesh V, Brown S, Sullivan Y. Guidelines for conducting mixed-methods research: an extension and illustration. JAIS 2016 Aug;17(7):435-494. [doi: 10.17705/1jais.00433]
86. Sheikh A, Anderson M, Albala S, Casadei B, Franklin BD, Richards M, et al. Health information technology and digital innovation for national learning health and care systems. The Lancet Digital Health 2021 Jun;3(6):e383-e396. [doi: 10.1016/s2589-7500(21)00005-4]

Abbreviations

HBM: Health Belief Model
IoT: Internet of Things
PMT: Protection Motivation Theory
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
TAM: Technology Acceptance Model
TTF: task-technology fit
UTAUT: Unified Theory of Acceptance and Use of Technology

XSL•FO
RenderX

Edited by T Leung, A Coristine; submitted 08.Jul.2024; peer-reviewed by H Gandhi, J Walsh, GK Gupta, S Ashraf; comments to author 23.Dec.2024; revised version received 03.Mar.2025; accepted 31.Aug.2025; published 06.Jan.2026.

Please cite as:
Veiga I, Oliveira T, Naranjo-Zolotov M, Martins R, Karatzas S
Adoption of Internet of Things in Health Care: Weighted and Meta-Analytical Review of Theoretical Frameworks and Predictors
J Med Internet Res 2026;28:e64091
URL: https://www.jmir.org/2026/1/e64091
doi:10.2196/64091
PMID:

©Inês Veiga, Tiago Oliveira, Mijail Naranjo-Zolotov, Ricardo Martins, Stylianos Karatzas. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Patient Perceptions of Blockchain-Based Health Information Exchange: User-Centered Design Study

Richard Guse[1], MSc; Shanshan Hu[2], MSc; Scott Thiebes[3], PhD; Christina Erler[4], PhD; Catherine Caridia[5], MD; Wilhelm Stork[4], Prof Dr; Martin Gersch[6], Prof Dr; Ali Sunyaev[2], Prof Dr

[1]Department of Economics and Management, Karlsruhe Institute of Technology, Kaiserstraße 89, Karlsruhe, Germany

[2]School of Computation, Information and Technology (CIT), Technical University of Munich, München, Germany

[3]Advanced Institute of Business, School of Economics and Management, Tongji University, Shanghai, China

[4]FZI Research Center for Information Technology, Karlsruhe, Germany

[5]Charité Comprehensive Cancer Center, Charité - Universitätsmedizin Berlin, Berlin, Germany

[6]Department of Information Systems, School of Business and Economics, Freie Universität Berlin, Berlin, Germany

Corresponding Author:
Richard Guse, MSc
Department of Economics and Management, Karlsruhe Institute of Technology, Kaiserstraße 89, Karlsruhe, Germany

Abstract

Background: Blockchain-based health information exchange (HIE) has received increased attention in health care research and practice over the last years. It enables the sharing of patient information across healthcare organizations, provides higher levels of data confidentiality and security, and reduces time and costs in collaborative medical decision-making. To make informed decisions on the implementation of blockchain-based HIE in practice and to fully understand the implications of its use for patient care, it is important to gain insights into patient perceptions of and interactions with blockchain-based HIE.

Objective: This study aimed to assess patient perceptions of a blockchain-based HIE mobile app to inform its iterative development.

Methods: We used a mixed methods user-centered design in 3 phases to iteratively assess patient perceptions of blockchain-based HIE: (1) structured questionnaires collecting patient requirements for blockchain-based HIE, (2) semistructured interviews evaluating mobile app mock-ups, and (3) a survey with blockchain-based HIE scenarios related to their patient care using the technology acceptance model, System Usability Scale, and open feedback. Both the semistructured interviews and the survey were conducted in a clinical setting with patients with cancer undergoing treatment at a major university hospital in Germany. As an exemplary case, we deem patients with cancer as well-positioned to evaluate a blockchain-based HIE mobile app since their treatment requires extensive coordination and data sharing across health care providers.

Results: Our findings support that patients have a high intention to use the blockchain-based functions that enable them to define, track, and revoke access to their health data per health care facility and service provider. Patients rated the 4 key functionalities (connection with providers, document sharing, a health diary, and a health care service provider search) as both useful and easy to use. The overall System Usability Scale of the blockchain-based HIE mobile app improved over the 3 phases up to 77.34, showing a good overall usability. The open feedback showed that patients' perceived usefulness of a blockchain-based HIE mobile app is especially influenced by 3 factors: the acceleration of the process of data sharing, patient-centered access control, and alignment with the respective health care settings. Moreover, patients' perceived ease of use of a blockchain-based HIE mobile app is impacted by 3 additional factors: the intuitiveness of the interaction, an aesthetic and functional design, and individual differences such as age or literacy with document management systems.

Conclusions: The evaluation demonstrates that patients are inclined to use blockchain-based HIE to manage their health data, as it empowers them to control which health care providers or individuals can access their information. To foster the use of a blockchain-based HIE mobile app, the app should allow patients to effortlessly establish connections with health care providers, offer an overview of all patient data, and enable patients to share medical documents individually via the app.

(J Med Internet Res 2026;28:e78849) doi:10.2196/78849

KEYWORDS

electronic health records; blockchain; patients; user-centered design; acceptance; usability; technology acceptance model; System Usability Scale; surveys and questionnaires

XSL•FO
RenderX

Introduction

Background

Patients often receive care from multiple health care providers throughout their lifetime, leading to health data being fragmented across institutions [1,2]. Multiple mechanisms have been introduced to address such fragmentation, including manual data transfers (eg, on paper or digital storage media) [3], electronic transfers via health information exchanges (HIEs) [4,5], and efforts aimed at centralizing patient information through general practitioners (eg, in the United Kingdom [6] or Germany [7]) or in regional or national databases (eg, in the United Kingdom [8] or Denmark [9]). Yet, patient data fragmentation persists [5,10], especially due to challenges such as incomplete records [6], data confidentiality concerns [11], and security vulnerabilities [12,13]. As a result, clinicians frequently lack a complete view of patients' medical histories, contributing to redundant tests, delayed diagnoses, and higher health care costs [1,14]. Developing a fully integrated, real-time patient record, therefore, remains a major challenge.

Blockchain technology has recently emerged as a promising solution for addressing data fragmentation issues and building interoperable, patient-centered HIEs [15,16]. By providing a decentralized, tamper-resistant ledger for managing consent and data access, blockchain can enhance data security and privacy [3,17] and give patients greater control over their information [17]. Especially, it enables patients to have an up-to-date overview of their health data (eg, which data are stored by which health care providers) and to decide independently who (eg, health care providers or service providers) can access their data and for what purposes [16]. These properties have positioned blockchain-based HIE as one of the most frequently discussed use cases of blockchain in health care among researchers and practitioners [15-20]. However, broad adoption of blockchain-based HIE still remains an issue. Reasons for this issue include a lack of organizational commitment, the absence of real-world testing of blockchain benefits in health care, and the functions of BC-based HIE being intangible for patients [19].

Prior research on blockchain-based health data management has examined diverse aspects, including challenges and benefits of its application [16,21-23], technical specifications [20,24,25], and users' attitudes toward different blockchain-based HIE functionalities [3,15,26,27]. These studies generally indicate that both health care professionals and patients have a positive attitude toward blockchain-based HIE and recognize its potential for improving transparency and data control. However, most of these investigations have focused on measuring general attitudes or perceived benefits through hypothetical scenarios [3,27,28] or interview-based discussions [15,29]. Thus, they mainly capture anticipated advantages rather than experienced ones. Only a few studies have involved prototype evaluations. Those studies typically focus on isolated functions, such as data sharing with research organizations [28], transaction tracking [30], or consent management [31], without allowing patients to actively manage access rights and view provider-specific medical records within an integrated system environment. Moreover, many blockchain-based HIE apps are still in a conceptualization stage, with very few having progressed to prototype development, let alone to actual implementation stages [16]. Consequently, a notable research gap remains regarding empirical evidence on how patients actually experience and interpret blockchain-based HIE in practice.

Objectives of the Study

Building on the identified research gap regarding the limited empirical evidence on how patients experience and interpret blockchain-based HIE systems in practice, this study aims to advance the understanding of patient engagement with such technologies beyond hypothetical attitudes. To this end, we conducted a 3-phase, user-centered design and evaluation of a functional blockchain-based HIE prototype, where patients recruited from a major university hospital interact with realistic usage scenarios (ie, patient-facing, practice-oriented interactions). Specifically, the study pursues three objectives: (1) to report on the user-centered design of a blockchain-based HIE mobile app prototype that enables patient-controlled management and sharing of health data; (2) to assess patients' experiences of blockchain-based HIE related to their current treatment in a clinical context and to evaluate their perceptions of the prototype's usability, usefulness, and overall user acceptance across four realistic HIE scenarios; and (3) to derive implications for the user-centered design and adoption of blockchain-based HIE solutions by identifying key functionalities and design features that influence user acceptance.

By examining patients' lived interactions with a functional blockchain-based HIE mobile app, this study provides empirical insights into how blockchain functionalities shape user experiences and adoption intentions in health care contexts. With our focus on evaluating a functional blockchain-based HIE mobile app with real patients, we explicitly go beyond the analysis of hypothetical attitudes as reported in prior research [3,27,28]. In addition, by examining patients' perceptions related to usability, usefulness, and acceptance across different functions of a blockchain-based HIE app, our study identifies which functions are crucial for blockchain-based HIE applications and where design improvements are needed. This is especially relevant given that blockchain-based applications are often characterized by low usability, as users have to invest a considerable effort to learn how to use these applications (eg, by installing and managing wallets for private keys) [32,33]. These insights can help researchers and practitioners identify barriers to the adoption of blockchain-based HIE, predict its actual use in the future, and make informed user-centered decisions about its implementation [34,35].

Methods

Overview

The development of the blockchain-based HIE app was carried out in collaboration with public health care providers and service providers (eg, a health cloud provider) in the context of a collaborative research project, BloG3. Our aim was to develop and test a blockchain-based HIE mobile app in the hospital context that would support patients in retrieving and sharing

XSL•FO
RenderX

documents from service providers and health care providers. For the development of the blockchain-based HIE mobile app, we used user-centered design involving patients as primary users to cater the mobile app to patients' needs [36,37]. User-centered design provides guiding principles and suggests techniques at various stages of the design process [36]. It ensures that an app is easy to learn, offers intuitive interactions, and is tied to user requirements. As our main study site, we chose the oncology department of a major university hospital in Germany, from where we recruited patients with cancer to evaluate the mobile app. We chose patients with cancer as the primary user group since their complex treatment pathways demand extensive coordination and data sharing across health care providers. In addition, their physical and emotional challenges further make usability especially critical for an HIE application. These characteristics make patients with cancer a crucial test case for evaluating blockchain-based HIE, which aims to support secure, patient-centered data sharing in contexts that require high coordination.

The research project was divided into 3 phases in which we iteratively developed and evaluated the functionalities of the mobile app (refer to Figure 1). The 3 phases enabled us to incorporate patient needs from an early design stage to the full implementation of the mobile app. After identifying user requirements (phase 1), we developed a first design of the mobile app (mock-ups) based on the identified user requirements. Subsequently, we evaluated the mock-ups with patients to receive further user feedback on the design (phase 2). Finally, we implemented a blockchain-based HIE mobile app and collected user perceptions of the functionalities of the mobile app in a clinical context, with patients undergoing medical treatment in a hospital, realistic usage scenarios, and realistic sample documents related to their care (phase 3). While phases 1 and 2 yielded important information for the development of the mobile app, phase 3 served to provide us with comprehensive feedback on the usability, usefulness, and acceptance of the implemented mobile app. This feedback is especially helpful to determine barriers to the adoption of blockchain-based HIE (eg, low usability) and predict its actual use in the future. Hence, while we also report on the results of phases 1 and 2, we put a higher focus on reporting the results of phase 3 in this study.

Figure 1. Phases of the development and evaluation process for the blockchain (BC)-based health information exchange (HIE) mobile app. TAM: technology acceptance model

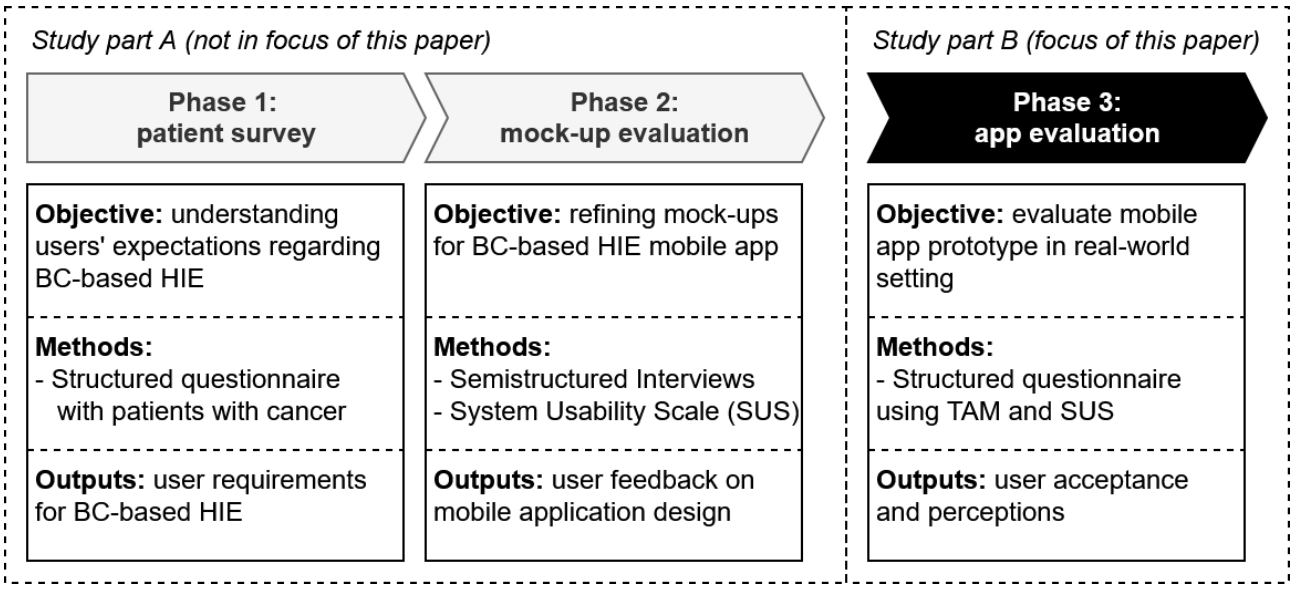

Phase 1: User Requirements Elicitation

In the first phase, we identified user requirements (eg, preferred app features). The aim of this phase was to get an overview of important features and health data that should be included in the blockchain-based HIE mobile app. To identify user requirements, we devised an online survey with questions about patients' current overview of their health data, whether they want to manage access rights per facility and doctor, whether they would like to have an overview of who accessed their health data, which documents they deem relevant to share, and which device they would like to use to manage their health data. Questions were iteratively developed by the author team based on prior literature on blockchain-based HIE [4,5,5-16], in a structured format, and could be answered on a 6-point Likert scale ranging from "1 (strongly disagree)" to "6 (strongly agree)." For pretesting the understandability of the questionnaire, we recruited a convenience sample of 18 participants who provided us with feedback, which we used to clarify unclear descriptions and specify ambiguous wording. An overview of the questionnaire and the survey results is available in Multimedia Appendix 1. We recruited participants via oncology patient communities nationwide by mail, inviting them to participate in the online survey. Patients with cancer in treatment or aftercare, next of kin, and other noncancer patients (eg, significant others) were able to participate to collect a broad overview of groups involved in the management of patient documents throughout the care process.

Phase 2: Mock-Up Evaluation

In the second phase, the identified user requirements were used as a foundation to develop mock-ups for a blockchain-based HIE mobile app consisting of a total of 36 clickable frames on a handheld computer. We developed the mock-ups iteratively

XSL•FO
RenderX

within the author team with external partners experienced in mobile app development for health care use cases. We evaluated the mock-ups with semistructured interviews during which we collected patient impressions of their interaction with the clickable mock-ups. Patients were asked to access different functionalities in the mock-ups, which were chosen based on user feedback from phase 1 and to answer questions related to the mock-ups. The functionalities included an overview of patient documents and medications, a diary for patient mood and health condition, and an overview of documents shared with exemplary health care facilities. An overview of the mock-up screens is shown in Figure 2. Further screens are available in Multimedia Appendix 2. The questions were developed based on the questionnaire from phase 1 and refined by the author team. These included questions related to the overall impression of the blockchain-based HIE mobile app, whether they would use it, and which medical documents they attach the greatest importance to.

Figure 2. Mock-up screens of the blockchain (BC)-based health information exchange (HIE), showing a general overview of recent patient document interactions, medications, and available patient documents.

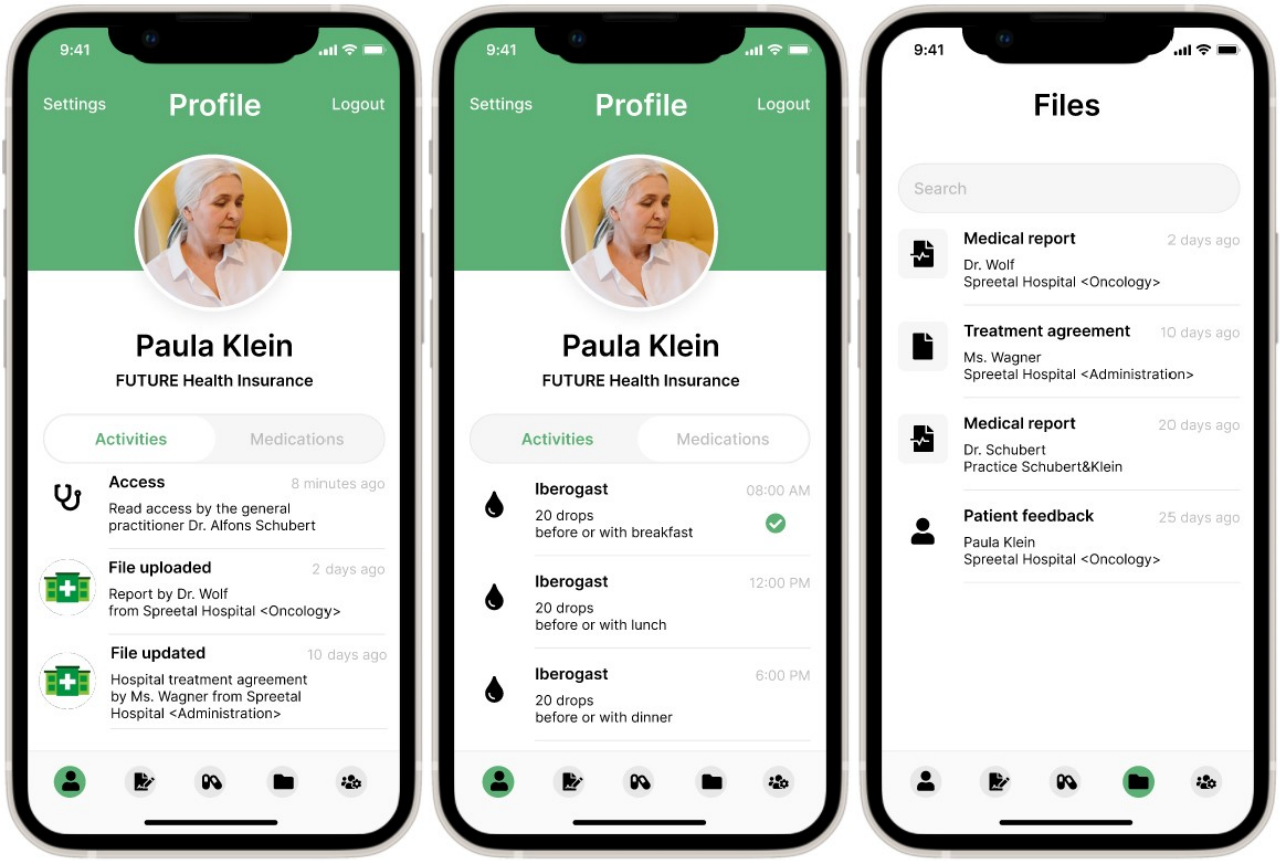

To measure usability, we collected patient perceptions of the app using the System Usability Scale (SUS) using an established German translation [38]. We chose SUS as it is established for measuring the user experience of a given software application [39]. We classify the SUS with an established grading scale ranging from a worst grade of "F" SUS range of 0.0 - 51.6 to a best grade of A+ range of 96.0 - 100 to assess the usability with a measure that shows what an acceptable SUS score is [39]. The interviews took place on-site in the oncological outpatient department of our main study site. We recruited adult patients at random, by willingness to participate, through on-site research staff. We applied quota sampling by time to randomize participants; that is, we recruited patients on several days at different times. Due to ethical concerns, we explicitly excluded patients younger than 18 years. We emphasized to patients that refusing to participate would not affect their treatment in any way. If patients refused to participate, we did not contact them again and excluded them from data collection. A researcher of the author team introduced patients to the idea of the blockchain-based HIE mobile app, observed patients while they clicked through the mock-ups, and gave instructions if needed. After patients finished accessing all functionalities, they filled out the SUS on a paper sheet. We aimed to provide only minimal guidance for clicking through the mock-ups. We encouraged participants to explore the functionality of the app themselves on a provided mobile device. If a participant could not access certain functionality, the researcher showed them how the functionality could be accessed by clicking through the mock-ups. While the researcher focused on highlighting the functionality of the mock-ups, BC-based HIE was introduced to patients as the mechanism that enabled sharing and revoking access to patient documents.

Phase 3: Mobile App Evaluation

In the final phase, we developed the blockchain-based HIE mobile app and a blockchain-based backend infrastructure, which enabled patients to share their medical documents. While our system was based on a system design recently proposed in the literature [40], we based the mobile app on the clickable frames from phase 2. To evaluate the blockchain-based mobile

XSL•FO
RenderX

app, we conducted interviews with patients in the oncological outpatient department of our main study site using the mobile app in practice-oriented scenarios that users face during their treatment.

Description of the Developed Blockchain-Based HIE System

For our developed blockchain-based HIE system, we adopted a system design recently proposed in the literature [40]. During system design, we incorporated threat modeling using the STRIDE (Spoofing, Tampering, Repudiation, Information disclosure, Denial of service, Elevation of privilege) method to systematically identify potential security risks, inform the architecture, and ensure secure interorganizational health data exchange [41]. An overview of the system is shown in Figure 3. The system consisted of the following four components: (1) a mobile app, (2) network infrastructure with a blockchain, (3) a connector to the IT systems of health care facilities and service providers, and (4) the existing internal IT systems at the health care facilities and service providers.

Figure 3. Blockchain (BC)-based health information exchange (HIE) system overview. IT: Information Technology

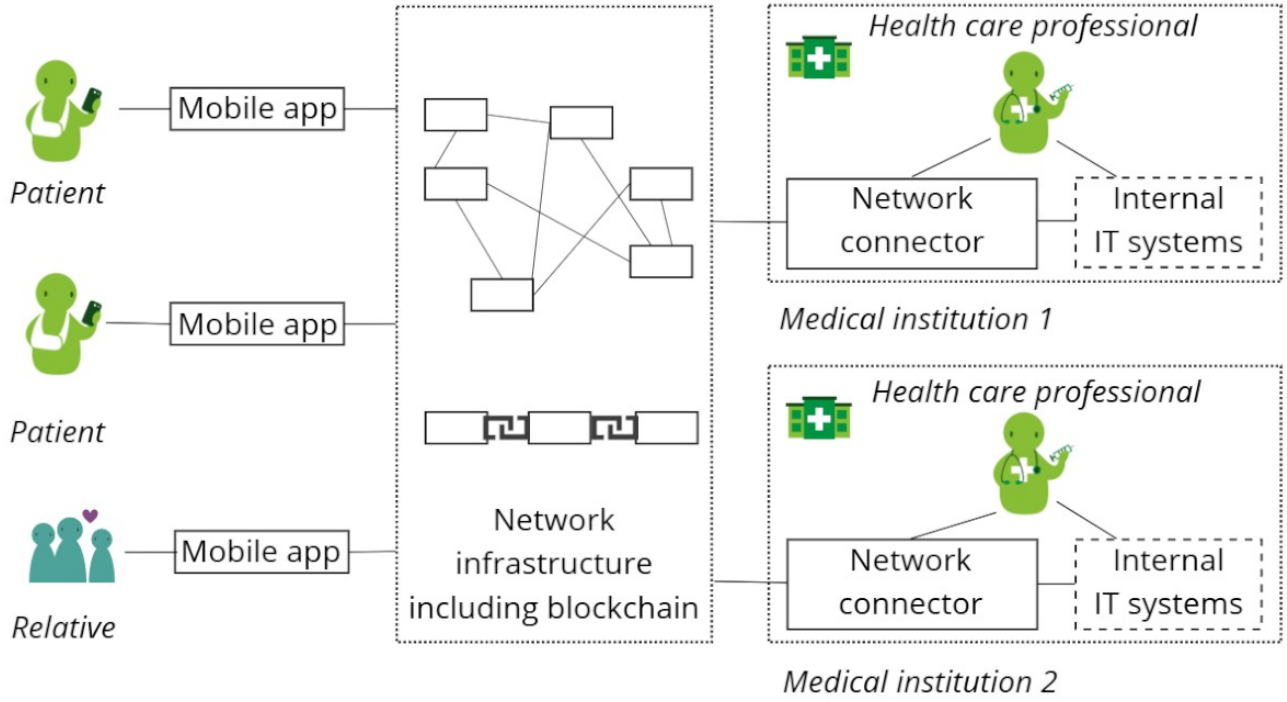

The network infrastructure is a decentralized platform secured by blockchain technology. It enables secure, asynchronous, encrypted peer-to-peer communication and data exchange between patients' mobile apps and connectors of health care facilities and service providers. The blockchain publishes a secure and tamper-proof registry for digital signatures with which patients, health care facilities, and service providers verify their identities. This enables patients to grant access rights to their health data to trusted health care facilities, service providers, and individuals based on their verified identity. Health data are stored off-chain within the existing infrastructure of the health care facility and service providers, ensuring that the data remain securely within the provider's systems, thereby eliminating the need to rely on external systems for trust. Additionally, there are no issues with deleting health data from the blockchain, which is a common challenge for blockchain use in health care. The connector is integrated into the internal IT systems of health care facilities and service providers and acts as an interface within the network infrastructure. This interface facilitates seamless access to data stored within health care facilities and service providers while ensuring interoperability and compliance with privacy and security standards. The system is based on the open-source implementations of Hyperledger Aries [42] and Indy [43] by the LF Decentralized Trust as part of the Linux Foundation, with connectors implemented using Aries and the public blockchain using Indy. Hyperledger Indy serves as the blockchain-based registry for identity management and credential verification. For this, Indy uses Plenum, an implementation of the Redundant Byzantine Fault Tolerance consensus mechanism. It should be noted that during the evaluation, no connection to live EHR systems or national health infrastructures was implemented. However, we set up connectors at 2 different health care facilities, including our main study site, to provide synthetic patient data during the evaluation. Overall, the system architecture was designed to be modular and extensible to allow future adaptations.

Description of the Mobile App for Patients

The mobile app is the user interface of the developed blockchain-based HIE system for patients and was developed based on the patient requirements from phase 1 and the feedback to the mock-ups from phase 2. It serves as a central access point and hub for the management of patients' health records stored at different health care facilities and service providers. By actively involving patients, the mobile app ensures privacy and control over their health data. Moreover, the mobile app was developed using the Xamarin platform to enable cross-platform deployment [44]. The 4 key functionalities of the mobile app include establishing a connection with a health care or service

XSL•FO
RenderX

provider by scanning a QR code, selecting and sharing specific documents through the system with other health care and service providers involved in patient care (Figure 4), creating health diary entries, and searching for nursing homes within a given area. An overview of the blockchain-based app screens is included in Multimedia Appendix 3.

Figure 4. Blockchain-based health information exchange (HIE) app screens for sharing specific documents with health care and service providers and entering data into a patient diary.

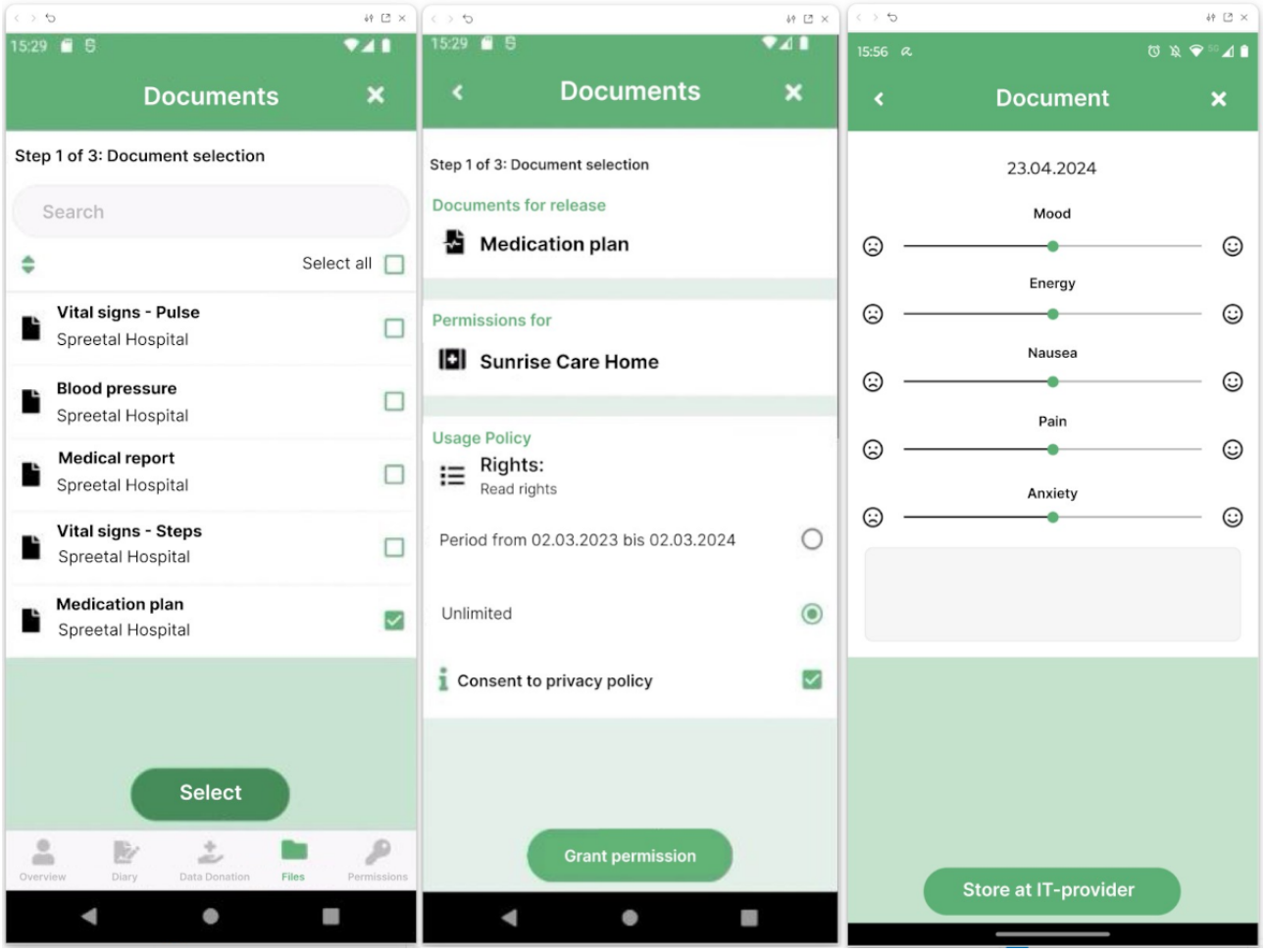

For patients to access their health data, a connection to the health care facility or service provider is required. Establishing a connection begins with the patient scanning a QR code provided by the health care or service provider using their mobile app. This action initiates the registration of the patient's mobile app with the service provider's connector. Subsequently, the service provider is listed as a registered contact in the patient's mobile app. When new health data becomes available, the patient receives an automatic notification via their mobile app. The patient can directly access and filter the health data from the service provider using the mobile app. Furthermore, they can grant other registered organizations access to this data. Patients can share their data from the primary health market, like health care providers (eg, findings and medical reports), data from the secondary health care market, like service providers (eg, recordings of fitness watches), as well as self-generated patient-reported outcome data (eg, symptoms in a patient diary). The patient-reported outcome data can be documented multiple times a day using a diary function with 5 entry types that can be recorded using sliders: Mood, Energy, Nausea, Pain, and Anxiety. A free-text field was included for individual comments. Based on patient feedback during interactions with the mock-ups, an additional feature was implemented, enabling patients to search for nursing homes within a specified area. This function was added since patients wanted to search and connect to further relevant providers directly via the mobile app without the need for scanning another QR code. Overall, to further improve the usability of the app, we added more information to symbols and buttons since patients indicated that the symbols used were not intuitive.

Study Description for Evaluating the Mobile App

For the evaluation of the blockchain-based HIE mobile app, we recruited patients from the oncological outpatient department of our main study site. A researcher from the author team recruited adult patients based on their willingness to participate. We randomized participants by applying quota sampling by time. Accordingly, we recruited patients on several days at different, varying time slots, thereby reducing time-based recruitment bias. Systematic randomization was not applicable since the time for data collection was limited. We explicitly excluded patients younger than 18 years due to ethical concerns. We informed patients that their participation in the study was voluntary and would not affect their care in any way. When patients refused to participate, we did not contact them any further and excluded them from further data collection. For the

XSL•FO
RenderX

systematic evaluation of the blockchain-based HIE mobile app, we devised a questionnaire that introduced the participants to blockchain-based HIE, asked for demographic information, and asked the participants to complete 4 tasks with the mobile app. The questionnaire was developed and assessed in several iterations within the author team to improve its understandability. To ensure its suitability for measuring patient perceptions, we revised the questionnaire in terms of understandability and clarity based on the feedback from 2 IT professionals of the collaborative research project BloG3 in which the studies were situated and 2 physicians. The questionnaire is available in Multimedia Appendix 4, and an overview of the survey results is available in Multimedia Appendix 5.

While the researcher on-site presented the questionnaire to patients on paper, the blockchain-based HIE mobile app was handed out on a provided mobile device to patients. Similar to phase 2, the researcher offered only limited guidance for clicking through the mobile app. The researcher waited at least 30 seconds until a participant received help, encouraging participants to complete the questionnaire by exploring the functionality of the mobile app themselves. If participants indicated that they did not wish to proceed with a task, the researcher explained to the participant where to click.

The first part of the questionnaire contained a brief introduction to the concept of blockchain-based HIE. The key functionalities of the blockchain-based HIE were highlighted as the scanning of a unique decentralized identifier (DID) via a QR code and the function of sharing and revoking access to patient documents. Patients could only perceive the functionality that is enabled and to some degree necessitated by using a blockchain (eg, when establishing a connection with a health care facility or service provider). Next, we collected demographic information on gender, age range, number of document management systems used (eg, cloud storage providers), and frequency of their use.

The main part of the questionnaire set 4 tasks that patients had to complete with the mobile app. Each task relates to 1 of the 4 key functionalities of the mobile app. For each task, we collected 2 outcome measures by noting whether a patient was able to complete a task without assistance and asking patients to confirm the result of a task (eg, being connected to a specific health care facility). If a patient indicated having difficulties completing a task, the researcher assisted the patient in completing the task on the mobile device by providing information on what to look for. After completing a task, patients were asked to assess the functionality with the 10-item questionnaire of the technology acceptance model (TAM) [45]. The items measured the 3 main TAM constructs, perceived usefulness (PU), perceived ease of use (PEOU), and behavioral intention (BI) to use, on a 7-point Likert scale ranging from "1 (strongly disagree)" to "7 (strongly agree)." The TAM constructs enable the measurement of user perceptions of how useful an application is for a given task (PU), how effort-free the interaction is (PEOU), and how likely they will adopt an application [45]. We chose TAM as it is an established measurement tool to evaluate user perceptions of software applications that allow us to predict whether users are likely to use an app before its rollout [46]. For our study, we based the items on a validated German translation [47]. In task 1, patients were asked to establish a connection with a health care facility by scanning a QR code. Task 2 requested patients to share preset medical documents and health data (ie, a medication plan, blood pressure values, heart rate, and steps) in their mobile app with a health care facility and a service provider. For one, they shared the medical document for a limited timeframe. For the other, they were required to share and revoke access to the medical document. In task 3, patients were asked to enter a random mood and health condition, share it with one of the health care facilities or service providers, and delete it. In task 4, patients searched for a nursing home to which their health data could be shared via the mobile app to support the discharge process. After completing all tasks, patients were asked to evaluate the overall usability of the blockchain-based HIE mobile app with the SUS [38]. Similar to phase 2, we classify the SUS with an established grading scale [39]. Finally, we collected free-text responses to get additional qualitative feedback on perceived advantages, challenges of the app, and suggestions for improvement.

Data Analysis

For every task, we calculated the mean scores and SDs for the used constructs (PEOU, PU, and BI). In addition, we checked the scale reliability of the used constructs, since we reworded the items to include the name of the tested blockchain-based HIE mobile app instead of a general "system." We thereby ensure that the items work cohesively for our study setting. For testing the scale reliability, we checked the internal consistency of the used, reworded items for PEOU and PU with Cronbach α [48]. For measuring the internal consistency of the items for the construct BI, we used the Spearman-Brown coefficient, since BI is only a 2-item scale for which Cronbach α cannot adequately determine the reliability [49].

To better understand the rationale behind participants' perceptions, we applied thematic analysis to patient responses to the 3 open questions. For the analysis, we followed the 6 phases outlined in [50]. In the first phase (familiarizing with the data), 2 authors independently reviewed the responses and mapped each to 2 of the 3 TAM constructs that influence the intention to use (ie, PU and PEOU). During the mapping process, they extracted key statements describing factors influencing these 2 constructs and performed an initial round of coding to generate first-order codes (second phase: generating initial codes). The 2 authors synthesized similarities and discussed differences among these initial codes with 1 additional author. In the third phase (searching for themes), 2 authors iteratively identified relationships between initial codes and clustered them into broader, conceptually connected themes. In the subsequent phase (reviewing themes), the team evaluated the internal consistency and distinctiveness of these candidate themes, ensuring that each accurately reflected the coded data and the dataset as a whole. This stage involved integrating closely related themes, subdividing those that were too encompassing, and reallocating codes that did not clearly belong to their original theme. This iterative process led to the identification of 6 themes. In the fifth phase (defining and naming themes), we defined the themes by identifying the

XSL•FO
RenderX

"essence" of each theme and clarifying its conceptual boundaries (ie, what aspect of participants' responses it represented and what it excluded). Based on their definitions, we named the refined themes and categorized them based on their contextual dependencies. The final outcomes of this analysis are detailed in the "Results" section (phase 6: producing the report). An illustrative example of this process is provided in Figure 5.

Figure 5. Excerpt from the qualitative coding process.

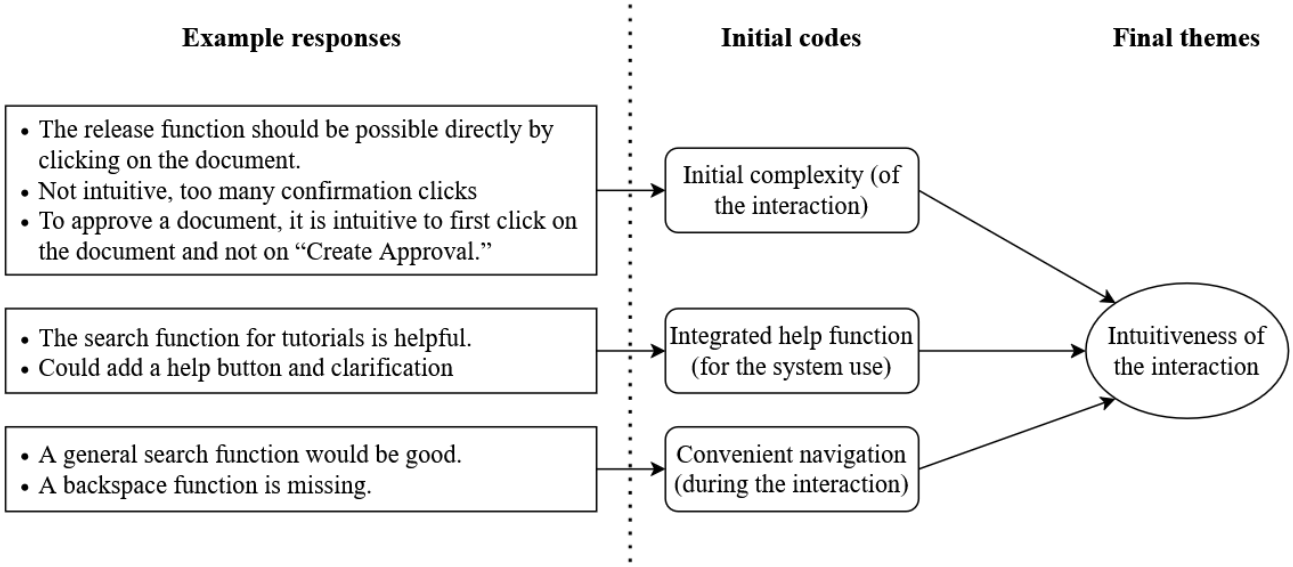

Ethical Considerations

All study parts were approved by the local ethics committee (Ethikkommission der Charité, no EA4/241/21). In the online survey in phase 1, a privacy statement informed participants about how their data were used and processed for the study. In phases 2 and 3, patients provided informed consent for the interviews. In all study parts, participation was voluntary and anonymous.

Results

Phase 1: Collection of User Requirements

In phase 1, we collected 108 responses from potential users, of which 60 completed the whole questionnaire (31 patients with cancer and 29 significant others involved in their care). From these, we could derive 4 requirements for a blockchain-based HIE application. An overview of participants' gender and age distribution is shown in Table 1. The responses showed that participants rather agreed with having a complete overview of their current health documents, although they indicated that neither their primary care physicians nor their relatives have a complete overview. More importantly, half of the participants (n=30) stated that they forgot or lost documents themselves, their health care staff lost documents, or documents were provided too late. Hence, we defined as the first requirement that the application should contain an overview of patients' health documents (Req 1). Regarding the platform for managing health data, participants stated that both a computer and a smartphone would be their preferred way for managing their health data. To allow for better portability and easier access everywhere, we decided to focus on a mobile app and defined the second requirement: the app should be accessible via smartphone (Req 2). Of all possible medical documents and health data, participants rated diagnosis reports, laboratory results, imaging results, and their medication plan as most important for sharing them with their health care providers. Thus, the app should at a minimum support the sharing of diagnosis reports, laboratory results, imaging results, and their medication plan (Req 3). Participants displayed high agreement with whether they would like to define who can access, edit, and delete their data per health care facility and physician. Participants preferred sharing specific documents instead of all data at once. In addition, they gave high priority to being able to track and revoke who accesses their data. Hence, the application should enable users to share, edit, track, and delete access rights to their medical documents per document per healthcare facility or service provider (Req 4). Overall, participants supported the idea of a blockchain-based HIE mobile app to manage access to their health data.

XSL•FO
RenderX

Table . Demographic data of participants for phase 1 (N=60).

Participant demographic	Values (N=60), n (%)
Sex	
Diverse	1 (1.5)
Female	28 (41.8)
Male	31 (56.7)
Age (years)	
18 - 29	9 (15)
30 - 39	13 (21.7)
40 - 49	7 (11.7)
50 - 59	13 (21.7)
60 - 69	13 (21.7)
70 - 79	4 (6.7)
80 - 89	1 (1.7)

Phase 2: Mock-Up Evaluation

In phase 2, we recruited 22 patients with cancer to evaluate the mock-ups of the blockchain-based HIE mobile app. Seventeen patients declined to participate. An overview of participants' gender and age distribution is shown in Table 2. The semistructured interviews provided insights into patient perceptions of the mobile app. Overall, the expectations of interviewed patients were positive. Patients stated that they would be very likely to use the mobile app once it is fully implemented. Patients had high anticipation of more efficient and transparent management of medical health care documents as well as faster communication between all stakeholders involved in the treatment process. The responses indicated that patients attach great importance to being able to manage the following health data in a blockchain-based HIE mobile app: medical findings, diagnoses, operating reports, laboratory results, and the medication plan.

Table . Demographic data of participants for phase 2 (n=22).

Participant demographic	Value, n (%)
Sex	
Female	14 (63.6)
Male	8 (36.4)
Age (years)	
18 - 29	0 (0)
30 - 39	3 (13.6)
40 - 49	4 (18.2)
50 - 59	8 (36.4)
60 - 69	6 (27.3)
70 - 79	1 (4.5)

For the clickable mock-ups, patients rated the usability with the SUS on average as 66.50, which hinted that the usability is satisfactory and can be graded as a "C" [39,51]. Still, several participants hinted that there may be challenges for older people to use the mobile app since several clicks were necessary to access certain functionalities. This challenge was noted and accounted for in the next phase of development to further enhance the usability of the mobile app.

Phase 3: Mobile App Evaluation

User Statistics

In the last phase, 32 participants were recruited, consisting of 13 males and 19 females. Nineteen patients declined to participate. The age range of participants spanned from 18 to 79 years, with the largest group being between 30 and 39 years (n=10). The gender and age distribution are shown in Table 3. The survey also gathered data on the number and frequency of IT systems used for document management (Table 2). Fourteen participants did not use any systems, while others reported using between 1 and more than 3. The frequency of use varied, with 18 participants engaging with these systems at least once a week.

XSL•FO
RenderX

Table . Demographic data of participants for phase 3 (n=32).

Participant demographic	Value, n (%)
Sex	
Female	19 (59.4)
Male	13 (40.6)
Age (years)	
18 - 29	1 (3.1)
30 - 39	10 (31.3)
40 - 49	5 (15.6)
50 - 59	6 (18.8)
60 - 69	8 (25)
70 - 79	2 (6.3)
80 - 89	0 (0)
IT[a] systems used for document management	
0	14 (43.8)
1	12 (37.5)
2	5 (15.6)
3	0 (0)
>3	1 (3.1)
Frequency of IT system use for document management	
Never	14 (43.8)
Once a month	5 (15.6)
Several times a month	3 (9.4)
Once a week	4 (12.5)
Several times a week	3 (9.4)
Daily	3 (9.4)

[a]IT: Information Technology

Evaluation Outcomes

All patients completed Tasks 1 and 4 without any assistance and correctly confirmed their connection to a health care facility (Task 1) and finding a nursing service (Task 4). Throughout Task 2, 6 patients (including both patients in the oldest age group [70-79 years]) needed assistance in carrying out the task. In Task 3, 1 patient received assistance in carrying out the task.

Table 4 provides an overview of the mean scores, SDs, and reliability measures (Cronbach α and Spearman-Brown coefficient) for the constructs PU, PEOU, and BI for each of the 4 tasks. A graphical comparison between the 3 constructs over the 4 tasks is shown in Figure 6. Overall, participants reported high levels of PU, PEOU, and BI for all tasks, with all reliability coefficients exceeding the recommended threshold of 0.70, displaying high internal consistency of the adapted scales [52].

XSL•FO
RenderX

Table . Technology acceptance model (TAM) construct measures for the blockchain (BC)-based health information exchange (HIE) app.

Construct	Task 1	Task 2	Task 3	Task 4
Perceived ease of use				
Value, mean (SD; 95% CI)	5.48 (1.42; 4.97-5.99)	5.69 (1.37; 5.19-6.18)	5.67 (1.33; 5.19-6.15)	5.6 (1.16; 5.18-6.02)
Cronbach	0.94	0.93	0.96	0.95
Perceived usefulness				
Value, mean (SD; 95% CI)	5.5 (1.51; 4.96-6.04)	5.84 (1.28; 5.37-6.30)	5.73 (1.29; 5.26-6.19)	5.59 (1.14; 5.17-6.00)
Cronbach	0.98	0.95	0.97	0.97
Intention to use				
Value, mean (SD; 95% CI)	5.61 (1.46; 5.08-6.14)	5.67 (1.6; 5.09-6.25)	5.47 (1.65; 4.87-6.06)	5.47 (1.28; 5.01-5.93)
Spearman-Brown coefficient	0.88	0.97	0.9	0.92

Figure 6. Perceived usefulness (PU), perceived ease of use (PEOU), and behavioral intention (BI) values and SDs for all 4 tasks of phase 3.

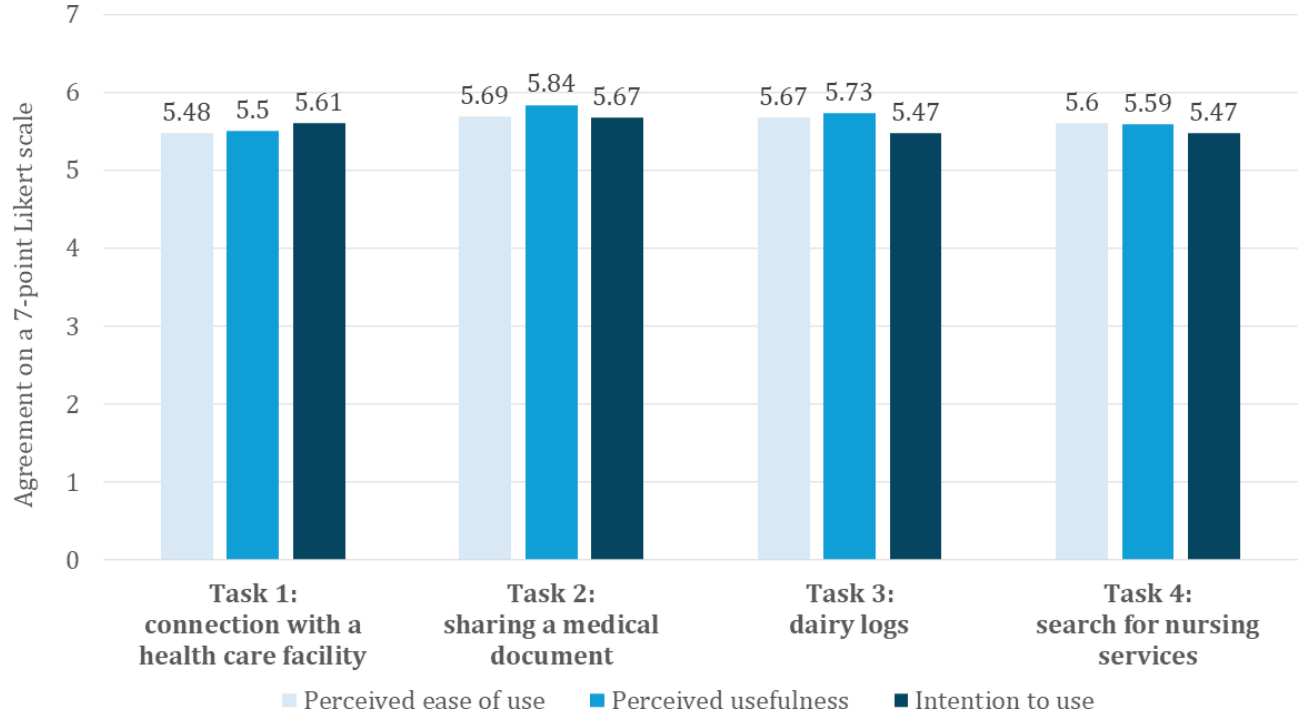

The developed blockchain-based mobile app achieved a SUS score of 77.34, indicating that the usability was good and could be graded as "B+" [39,52]. This showed an improvement from the 66.50 (graded as "C") reported in phase 2 (refer to Figure 7). It is noteworthy that older participants (ie, age groups 60 - 69 and 70 - 79 years) have a smaller SUS score of 64 (graded as "C–") and hence only a satisfactory rating of the mobile app. In contrast, younger age groups (ie, 18 - 59 years) have a SUS score of 83 (graded as "A"), indicating a high usability [51]. The results show that the BC-based mobile app received a higher SUS score of 84.12 (graded as "A") among female participants. In contrast, male participants provided a rating of 67.43 (graded as "C"). We observed a minor difference in the SUS scores between participants who indicated that they would not use any document management system in general (74, graded as "B–") and those who indicated using at least 1 such system (78.73, graded as "B+"). The SUS score of low frequencies of general document management system use was 75.61 (graded as "B") and of high frequencies of general document management system use was 79.87 (graded as "A–"), showing only a minor difference.

XSL•FO
RenderX

Figure 7. Grading of System Usability Scale (SUS) scores from phases 2 and 3.

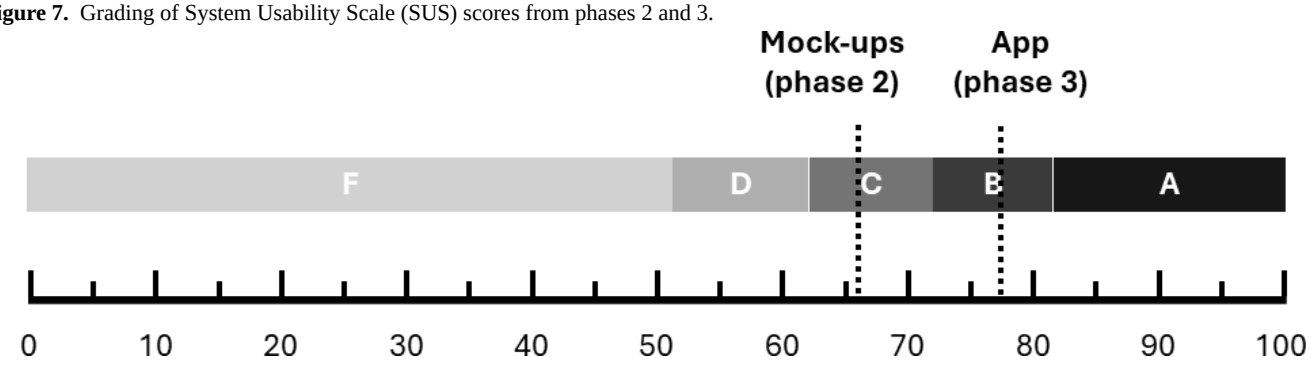

For completeness, we report on a subgroup analysis using Aligned Rank Transform ANOVAs to examine the influence of gender, age, and number and frequency of IT systems used for document management on PEOU, PU, BI, and SUS (Multimedia Appendix 6). Aligned Rank Transform was used to align and rank the data to be able to apply ANOVA to our ordinal data [53]. Both analyses are exploratory and should be interpreted with caution given the limited sample size and the low statistical power.

Influential Factors for User Perceptions

Our analysis of responses to the 3 open questions revealed 3 influential factors, each for PU and PEOU (refer to Table 5).

The factors influencing PU reflect the specific contextual nuances of blockchain-based HIE, highlighting that users view a blockchain-based HIE application's usefulness primarily through its ability to address specific challenges and requirements in this domain. In contrast, the factors shaping PEOU are context-agnostic, since they are not dependent on functionalities enabled by blockchain-based HIE. These align with broader user expectations for a general system's usability and are straightforward to adopt across diverse settings (refer to Figure 8).

Table . Factors influencing user perceptions of blockchain (BC)-based health information exchange (HIE) applications.

Factors	Description	Exemplary quotes
Operational efficiency of health data management	The HIE[a] application's ability to organize and manage health-related information in a manner that reduces the time, effort, and resources required.	• "Reduces bureaucratic overhead" • "Everything is concentrated and easily accessible"
Patient-centeredness	The HIE application's characteristic of placing patients at the center of its functionality by enabling them to manage their health data while ensuring transparency and privacy.	• "I like the feature that I can set my release options personally" • "Autonomous or not dependent on help from third parties"
Alignment with health care realities	The extent to which the HIE application reflects and supports the practical, real-world needs and workflows of health care settings.	• "It would be nice if there was an appointment entry function" • "The app should be available to everyone, including on mobile phones and via a browser."
Intuitiveness of the interaction	The ability of the HIE application to provide a seamless and intuitive interaction experience.	• "The search function for tutorials is helpful" • "To approve a document, it is intuitive to first click on the document"
Aesthetic and functional design	The visual appeal and practical functionality of the HIE application.	• "I would like to see a graphical representation" • "The visual design is appealing"
Individual differences	Personal characteristics such as age, health status, and related experience that influence individuals' attitudes toward the HIE application.	• "I see that as very personality-dependent" • "Older people may not be able to handle the many clicks."

[a]HIE: health information exchange.

XSL•FO
RenderX

Figure 8. Mapping of factors influencing user perceptions to the technology acceptance model.

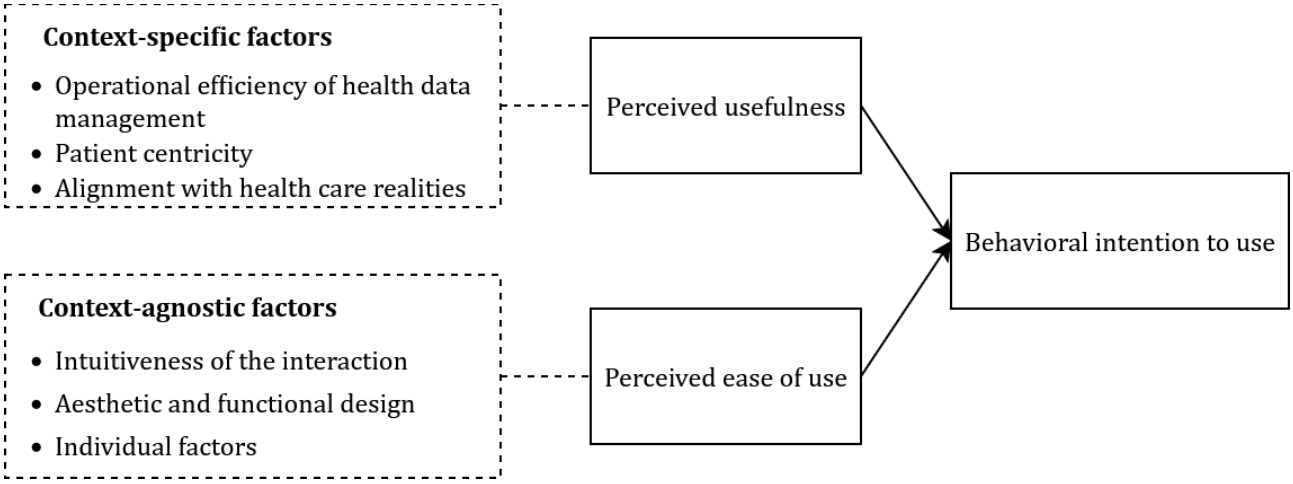

For PU, the following 3 factors were identified that can represent the key aspects users consider when evaluating the value and functionality of a blockchain-based HIE application.

Operational Efficiency of Health Data Management

Our analysis identified the HIE application's ability to organize and manage health-related information efficiently as a critical factor for its PU. This finding aligns with the highest PU score obtained for the document sharing task (Task 2), suggesting that participants particularly valued the application's ability to facilitate data exchange between health care facilities and service providers. In particular, they highlighted the application's strengths such as "faster document retrieval" and "easy sharing processes," which streamline the exchange of health-related information. Furthermore, features that reduce administrative burdens, such as "simplified [processes to achieve] compliance with data policies" and "straightforward access to documents," were highly valued for enhancing the efficiency of document retrieval. The organization of health data in one place also stood out. Participants appreciated the application's "integrated view of all relevant data," which provides a clear overview, ultimately supporting more effective health data management.

Patient Centricity

Many participants attributed the positive PU of our HIE application to its user-centered advantages of empowering users to take ownership of their health data while protecting their privacy. This factor exerts an influence on participants' perception of the application's usefulness, as it allows them to manage their own data access and sharing settings, thereby fostering a sense of autonomy. Some users especially noted that "self-determined and self-organized document sharing" resulted in a more responsive system that better met their needs. Additionally, transparency in how user data is handled contributed to establishing trust in the system and enhanced users' confidence in its utility. Participants expressed appreciation for "clear policies and measurements regarding data privacy," with some indicating that "knowing their data was secure and handled responsibly" addressed their concerns related to privacy.

Alignment With Healthcare Realities

Beyond the evaluated functions of the blockchain-based HIE application, participants emphasized the importance of aligning its capabilities with the health care setting's structural and technological demands. Functions reflecting real-world healthcare scenarios, such as accessing medication plans and document sharing, were emphasized to enhance the application's relevance and feasibility. This also helps explain the strong PU ratings for the document sharing task (Task 2). Moreover, participants suggested that additional features such as appointment scheduling, proof of insurance, and document search functions would further improve the PU. Although these were not tested in our survey, the participants considered them to be valuable for improving the usefulness of the application, as they can support people in their common activities of daily life. Furthermore, participants noted that the compatibility between the system and the existing infrastructure affects the PU of the application. Participants expressed that the application "should work seamlessly on different platforms or devices, including mobile phones and web browsers." Moreover, maintaining "accessibility in locations with limited network connectivity" can reinforce the application's usefulness in health care settings, as it facilitates uninterrupted service delivery and enables timely decision-making in critical situations when vital medical information is needed.

When reporting the PEOU, participants also identified 3 key factors that influence their perceptions:

Intuitiveness of the Interaction

The ability of the HIE application to provide a seamless and intuitive interaction experience was reported in our survey as a critical factor influencing PEOU. Many participants stressed that the system should feel natural and require minimal learning, particularly for new users. For example, participants value features that facilitate convenient navigation, such as the "inclusion of a backspace function." "Clear and consistent terminology" was also regarded as essential to "minimize [user] effort and frustration." The positive feedback on interface clarity and navigation aligns with the marked improvement in SUS from phase 2 (66.5) to phase 3 (77.3), suggesting that these refinements in interaction design translated into perceptibly higher ease of use. In addition, participants identified built-in

XSL•FO
RenderX

guidance, such as “a searchable help or tutorial feature,” as an important component to assist users in understanding the application’s functionality and resolving potential difficulties.

Aesthetic and Functional Design

Several participants pointed out that the design of the HIE application plays an important role in PEOU. On the one hand, participants appreciated well-organized and visually appealing interfaces, as they made the application more attractive and easier to navigate. In particular, information displayed in structured timelines or graphical formats was considered helpful, such as “visually appealing curves” and “organized layouts by day, week, or month, [which] make [the information] easier to understand.” On the other hand, participants attributed high PEOU to customization options, including the ability to “enter free text” or “filter information by various criteria (eg, diagnoses and providers).” These options enable the application to better fulfill users’ individual needs, thereby fostering a more comfortable user experience and, ultimately, PEOU.

Individual Differences

Personal characteristics emerge from our analysis as an influencing factor for PEOU. Given the health status of our participants, the first observation that caught our attention was that participants with serious health challenges reported that each new type of interaction felt less easy and more stressful. Besides, participants expressed concern about the application’s complexity for older users, noting that “too many steps or interactions could discourage [their] usage.” This observation aligns with concerns reported in phase 2 and the lower SUS scores reported by older participants in phase 3 (64 vs 83 for younger users), suggesting that PEOU was partly constrained by age-related differences. Furthermore, related experience and individual preferences further shaped perceptions. While some participants expressed discomfort or resistance due to limited familiarity with digital systems, 2 participants directly stated that their resistance stemmed solely from their personal aversion to digitization in general, not from the blockchain-based HIE application itself.

Discussion

Principal Findings

With this study, we provide empirical insights from the evaluation of a functional prototype of a patient-centric blockchain-based HIE mobile app in a clinical context. While prior studies largely examined hypothetical attitudes of blockchain-based HIE [3,27,28], our evaluation involves real patients undergoing medical treatment in a hospital, realistic usage scenarios, and realistic sample documents related to their care. By developing the blockchain-based HIE mobile app through a user-centered design approach in 3 phases, we incorporated patient feedback at every design stage to ensure that the prototype’s features and user interface aligned closely with patient needs. After iteratively refining the app, we conducted an evaluation with patients with cancer, assessing its usability, perceived value, and overall acceptance based on the SUS and TAM. While this evaluation context naturally limits the extent to which the results can be generalized to all patient groups or long-term use situations, it nevertheless provides a rigorous and informative test case for understanding how patients engage with blockchain-based HIE functionalities in practice. In the following, we discuss the main findings of our study.

First, our findings support the growing body of literature that suggests HIE as an auspicious use case for blockchain-based applications in health care [22]. Similar to prior studies focusing on users’ attitudes [3,15], the participants in our study expressed a positive attitude toward the blockchain-based functions that enable them to define, track, and revoke access to their health data. Although these findings stem from a specific patient population, the underlying interaction required by performing these functions is not disease-specific and reflects fundamental HIE tasks. This suggests that while generalization must be approached cautiously, the interactions we observe may hold relevance beyond the oncology context. The persistence of positive attitudes in our study reveals that a user-centered design process can effectively translate blockchain’s technical characteristics (eg, transparency and immutability) into comprehensible, patient-oriented functionalities. Furthermore, the high PEOU, PU, and BI ratings for the document-sharing function among all presented functionalities highlight the core need addressed by blockchain-based HIE applications, as these applications help to overcome the persistent challenges of managing patient data efficiently and securely in practice [22].

Second, by exploring the rationales behind reported user perceptions, we identified 6 influential factors for the 2 core TAM constructs (ie, PU and PEOU). While the 3 factors influencing PU reflect the specific contextual demands of HIE, the 3 factors shaping PEOU are context-agnostic. This strengthens prior findings, which highlight that PU depends on how well a system (eg, a HIE application) addresses the unique challenges and requirements of a particular context (eg, HIE) or task (eg, sharing a medical document) [54]. In contrast, PEOU reflects users’ perception of how easy and intuitive a system is to learn and operate. It is therefore influenced more by universal design factors that make the system user-friendly rather than its context of use [54].

Third, the age-related differences in perceived usability of our blockchain-based HIE resonate with a broad stream of eHealth and mobile health research indicating that older adults tend to experience higher usability barriers [55,56]. Although not significant, our results indicate a tendency of older adults to rate the overall usability of the mobile app lower than participants in younger age groups. In particular, among the surveyed patients, people from the age groups 60 - 69 and 70 - 79 years rated the mobile app’s usability as satisfactory only, while the younger age groups rated the usability of the mobile app as good. This concurs with previous research on digital health equity issues that emphasizes that older adults are generally less experienced with mobile apps, leading to a lower usability [55,56]. Nevertheless, this also highlights that the development of digital health apps must account for the higher needs of older adults for simplified navigation and intuitive interactions to overcome age-related barriers [57]. Addressing these barriers prevents digital health equity gaps and ensures that they can use and benefit from HIE mobile apps. Toward

XSL•FO
RenderX

that end, we deem user-centered design approaches especially useful since, in our study, for example, we were able to improve the usability ratings in the older age groups through multiple iterations.

Implications

Our study has several implications for research and practice. In terms of research, we add to the growing body of research on patient perceptions of blockchain-based HIE by evaluating the perceptions of cancer patients concerning interactions with a blockchain-based HIE mobile app in a clinical context. Furthermore, we combine quantitative survey results (ie, the SUS scores) with the in-depth insights from a qualitative survey of patients to expand our understanding of patient perceptions of blockchain-based HIE in a naturalistic setting. In particular, the results of the qualitative survey paint a more nuanced picture of context-specific and context-agnostic factors that are crucial for the broad acceptance of blockchain-based HIE. In addition, our study also highlights the value of user-centered design studies in the context of developing and evaluating blockchain-based applications for health care. Although blockchain is primarily a back-end technology, the decentralization of IT systems associated with the use of BC also has an indirect impact on the user experience. Against this backdrop, user-centered design studies can help to better understand user requirements and generate design knowledge for blockchain-based HIE applications.

From a practitioner perspective, our study yields implications particularly for healthcare policymakers and developers of blockchain-based HIE. For healthcare policymakers, the results of our evaluation highlight the potential of blockchain-based HIE to alleviate deficiencies in the current care system, such as lost documents, a time-consuming process to collect and share relevant medical documents, or incomplete health data for patients with cancer. Although many of these challenges are being addressed through the implementation of national health infrastructures, such as electronic health records (eg, [9]), patients remain hesitant to fully engage with these systems due to concerns about privacy, control, or usability, even in contexts where national systems automatically enroll patients unless they actively opt out [58]. Our results indicate that blockchain-based HIE solutions can address these concerns by fostering patient centricity and privacy. Therefore, our results highlight broader dimensions beyond basic usability measures, including patient empowerment and transparency, which are critical for implementing and integrating new technologies into patient-related health care workflows [59]. Participants in our study valued the ability to manage data access and sharing, which strengthened their sense of autonomy and increased the PU of our suggested blockchain-based HIE system. Such systems may be particularly beneficial in less connected areas, such as rural health care settings, where digital infrastructure and interoperability are limited. Pilot projects and model regions in these contexts could provide valuable insights for further scaling and optimizing these systems. Furthermore, for developers of BC-based HIE applications, our study provides a promising design for a BC-based HIE mobile app. When implementing blockchain-based HIE in practice, the identified factors related to the PU and the PEOU should be considered, since both will help ensure that patients use and accept a blockchain-based HIE mobile app. To ensure and improve PU, blockchain-based HIE mobile apps should focus on building functionalities that put all patient data into one overview to reduce time and effort in data sharing. Moreover, medical documents should be shareable individually, and patients should be able to effortlessly establish a connection with health care providers (eg, via a QR code). For ensuring and improving the PEOU, future developments should provide an intuitive design and minimize the number of clicks it takes to achieve a task (eg, share a medical document).

Limitations and Future Research

Our study is not without limitations. First, during the evaluation, patients could ask questions if they felt stuck while interacting with the mock-ups in phase 2 and the mobile app in phase 3. While we tried to provide minimal guidance across all surveyed patients, some asked for more assistance than others. This might have skewed the results in favor of a higher SUS, PEOU, PU, and BI and limited the comparability of patient assessments. Even though participants rarely requested guidance during tasks, the mere presence of help at hand could have led to participants feeling more confident in their interactions, thereby increasing their overall perceptions of the mobile app. Future research could extend our study by examining user perceptions without providing a contact person for assistance. Second, in our collection of patients' responses, we focused on patients with cancer in a single major university hospital in Germany because their treatment requires exchanging patients' health data across multiple health care providers and because their physical and emotional state demands a high usability of the mobile app. This focus might reduce the applicability of our results for other patient groups. Other patient groups might prefer sharing different documents (eg, vaccination record) or might even attach less importance to the concept of blockchain-based HIE since their situation does not demand frequent sharing of documents. Also, patient perceptions might be different when they have positively or negatively experienced the document management process in other hospitals. Future research could hence extend the evaluation of blockchain-based mobile apps to other patient groups and other hospitals. Third, we tried to incorporate a broad range of views in the development of the app through quantitative measures and qualitative insights. However, the convenience sampling used and the limited sample sizes in each phase of the evaluation might have introduced a selection bias and limited the generalizability of our results to a broader population. Future research should validate our findings in a large, balanced sample. This will be especially useful to draw further conclusions about the differences in participants' characteristics (eg, age) influencing patient perceptions (eg, PEOU, usefulness, or acceptance). Finally, we evaluated patient interactions with the blockchain-based HIE mobile app only at 1 point in time. Future research could extend this study by evaluating the interactions with a blockchain-based HIE mobile app over time. Such a longitudinal study could help uncover insights into actual patient sharing behavior and determine whether patients would use a mobile app over a longer timeframe.

XSL•FO
RenderX

Conclusion

With our study, we provide an exemplary case of a blockchain-based HIE mobile app developed through a 3-phase user-centered design approach incorporating on-site patients from a major university hospital. Our evaluation shows that patients in our study would likely use a blockchain-based HIE mobile app to manage their health data, since the app enables them to decide which health care provider or individuals can access their data.

Acknowledgments

We would like to thank all members of the BloG3 research project and the patients who agreed to participate in our study. In proofreading and editing our manuscript, we used the generative artificial intelligence (AI) tools: ChatGPT-4.5 (OpenAI) and Grammarly (Grammarly Inc.). We reviewed and used the suggestions of these tools when they improved readability and clarity and take full responsibility for the content of the publication.

Funding

Funding was provided by the Federal Ministry of Research, Technology and Space as part of the BloG3 research project https://www.blog3.de/(grant number 16SV8372).

Data Availability

The datasets generated or analyzed during this study are not publicly available to ensure the privacy of our participants but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Phase 1 questionnaire and frequency analysis.
[DOCX File, 31 KB - jmir_v28i1e78849_app1.docx]

Multimedia Appendix 2
Screenshots of the blockchain (BC)-based HIE mock-ups.
[DOCX File, 310 KB - jmir_v28i1e78849_app2.docx]

Multimedia Appendix 3
Screenshots of the blockchain (BC)-based HIE mobile app.
[DOCX File, 417 KB - jmir_v28i1e78849_app3.docx]

Multimedia Appendix 4
Phase 3 questionnaire.
[DOCX File, 39 KB - jmir_v28i1e78849_app4.docx]

Multimedia Appendix 5
Phase 3 technology acceptance model (TAM) and System Usability Scale (SUS) frequency analysis.
[DOCX File, 44 KB - jmir_v28i1e78849_app5.docx]

Multimedia Appendix 6
Phase 3 ANOVA results for technology acceptance model (TAM) and System Usability Scale (SUS).
[DOCX File, 4309 KB - jmir_v28i1e78849_app6.docx]

References

1. Kern LM, Bynum JPW, Pincus HA. Care fragmentation, care continuity, and care coordination—how they differ and why it matters. JAMA Intern Med 2024 Mar 1;184(3):236-237. [doi: 10.1001/jamainternmed.2023.7628] [Medline: 38285559]
2. Kern LM, Lau JD, Rajan M, et al. Associations among claims-based care fragmentation, self-reported gaps in care coordination, and self-reported adverse events. BMC Health Serv Res 2024 Sep 10;24(1):1045. [doi: 10.1186/s12913-024-11440-y] [Medline: 39256705]
3. Hau YS, Lee JM, Park J, Chang MC. Attitudes toward blockchain technology in managing medical information: survey study. J Med Internet Res 2019 Dec 9;21(12):e15870. [doi: 10.2196/15870] [Medline: 31815676]

XSL•FO
RenderX

4. U.S. Government Accountability Office. Electronic health information exchange: use has increased, but is lower for small and rural providers. : U.S. Government Accountability Office; 2023 URL: https://www.gao.gov/assets/gao-23-105540.pdf
5. The Office of the National Coordinator for Health Information Technology. Interoperability and methods of exchange among hospitals in 2021. : US Department of Health and Human Services; 2023 URL: https://www.healthit.gov/sites/default/files/2023-01/DB64_Interop_and_Methods_of_Exchange_Among_Hosp.pdf
6. Service (NHS) NH. View your GP health record. NHS England Digital. 2023. URL: https://www.nhs.uk/nhs-services/gps/view-your-gp-health-record/ [accessed 2024-09-23]
7. Walzer S. Digital healthcare in germany. In: Walzer S, editor. Digital Healthcare in Germany: Market Access for Innovations: Springer International Publishing; 2022:5-16 URL: https://link.springer.com/10.1007/978-3-030-94025-6 [doi: 10.1007/978-3-030-94025-6_2]
8. Service (NHS) NH. National care records service. NHS England Digital. 2024. URL: https://digital.nhs.uk/services/national-care-records-service [accessed 2024-12-23]
9. Danish health data authority. National health registers.: National health registers; 2024. URL: https://english.sundhedsdatastyrelsen.dk/health-data-and-registers/national-health-registers [accessed 2024-10-23]
10. Syed R, Eden R, Makasi T, et al. Digital health data quality issues: systematic review. J Med Internet Res 2023 Mar 31;25:e42615. [doi: 10.2196/42615] [Medline: 37000497]
11. Impact c for p. The Electronic Health Records System In the UK.: Centre for Public Impact; 2017. URL: https://centreforpublicimpact.org/public-impact-fundamentals/the-electronic-health-records-system-in-the-uk/ [accessed 2024-12-23]
12. Vest JR, Gamm LD. Health information exchange: persistent challenges and new strategies. J Am Med Inform Assoc 2010;17(3):288-294. [doi: 10.1136/jamia.2010.003673] [Medline: 20442146]
13. Neprash HT, McGlave CC, Cross DA, et al. Trends in ransomware attacks on US hospitals, clinics, and other health care delivery organizations, 2016-2021. JAMA Health Forum 2022 Dec 2;3(12):e224873. [doi: 10.1001/jamahealthforum.2022.4873] [Medline: 36580326]
14. Joo JY. Fragmented care and chronic illness patient outcomes: a systematic review. Nurs Open 2023 Jun;10(6):3460-3473. [doi: 10.1002/nop2.1607] [Medline: 36622952]
15. Lee K, Lim K, Jung SY, et al. Perspectives of patients, health care professionals, and developers toward blockchain-based health information exchange: qualitative study. J Med Internet Res 2020 Nov 13;22(11):e18582. [doi: 10.2196/18582] [Medline: 33185553]
16. Hu S, Schmidt-Kraepelin M, Thiebes S, Sunyaev A. Mapping distributed ledger technology characteristics to use cases in healthcare: a structured literature review. ACM Trans Comput Healthcare 2024 Jul 31;5(3):1-33. [doi: 10.1145/3653076]
17. Lee S, Kim J, Kwon Y, Kim T, Cho S. Privacy preservation in patient information exchange systems based on blockchain: system design study. J Med Internet Res 2022 Mar 22;24(3):e29108. [doi: 10.2196/29108] [Medline: 35315778]
18. Hasselgren A, Kralevska K, Gligoroski D, Pedersen SA, Faxvaag A. Blockchain in healthcare and health sciences-A scoping review. Int J Med Inform 2020 Feb;134:104040. [doi: 10.1016/j.ijmedinf.2019.104040] [Medline: 31865055]
19. Yeung K. The health care sector's experience of blockchain: a cross-disciplinary investigation of its real transformative potential. J Med Internet Res 2021 Dec 20;23(12):e24109. [doi: 10.2196/24109] [Medline: 34932009]
20. Zhuang Y, Sheets LR, Chen YW, Shae ZY, Tsai JJP, Shyu CR. A patient-centric health information exchange framework using blockchain technology. IEEE J Biomed Health Inform 2020 Aug;24(8):2169-2176. [doi: 10.1109/JBHI.2020.2993072] [Medline: 32396110]
21. Mackey TK, Kuo TT, Gummadi B, et al. "Fit-for-purpose?" - challenges and opportunities for applications of blockchain technology in the future of healthcare. BMC Med 2019 Mar 27;17(1):68. [doi: 10.1186/s12916-019-1296-7] [Medline: 30914045]
22. Durneva P, Cousins K, Chen M. The current state of research, challenges, and future research directions of blockchain technology in patient care: systematic review. J Med Internet Res 2020 Jul 20;22(7):e18619. [doi: 10.2196/18619] [Medline: 32706668]
23. Kuo TT, Kim HE, Ohno-Machado L. Blockchain distributed ledger technologies for biomedical and health care applications. J Am Med Inform Assoc 2017 Nov 1;24(6):1211-1220. [doi: 10.1093/jamia/ocx068]
24. Anand G, Sadhna D. Electronic health record interoperability using FHIR and blockchain: a bibliometric analysis and future perspective. Perspect Clin Res 2023;14(4):161-166. [doi: 10.4103/picr.picr_272_22] [Medline: 38025292]
25. Zhuang Y, Chen YW, Shae ZY, Shyu CR. Generalizable Layered Blockchain Architecture for Health Care Applications: Development, Case Studies, and Evaluation. J Med Internet Res 2020 Jul 27;22(7):e19029. [doi: 10.2196/19029]
26. Esmaeilzadeh P. Benefits and concerns associated with blockchain-based health information exchange (HIE): a qualitative study from physicians' perspectives. BMC Med Inform Decis Mak 2022 Mar 28;22(1):80. [doi: 10.1186/s12911-022-01815-8] [Medline: 35346176]
27. Esmaeilzadeh P, Mirzaei T. The potential of blockchain technology for health information exchange: experimental study from patients' perspectives. J Med Internet Res 2019 Jun 20;21(6):e14184. [doi: 10.2196/14184] [Medline: 31223119]
28. Esmaeilzadeh P, Mirzaei T. Role of incentives in the use of blockchain-based platforms for sharing sensitive health data: experimental study. J Med Internet Res 2023 Aug 18;25(1):e41805. [doi: 10.2196/41805] [Medline: 37594783]

XSL•FO
RenderX

29. Lu C, Batista D, Hamouda H, Lemieux V. Consumers' Intentions to Adopt Blockchain-Based Personal Health Records and Data Sharing: Focus Group Study. JMIR Form Res 2020 Nov 5;4(11):e21995. [doi: 10.2196/21995] [Medline: 33151149]
30. Sung M, Park S, Jung S, Lee E, Lee J, Park YR. Developing a Mobile App for Monitoring Medical Record Changes Using Blockchain: Development and Usability Study. J Med Internet Res 2020 Aug 14;22(8):e19657. [doi: 10.2196/19657] [Medline: 32795988]
31. Despotou G, Evans J, Nash W, Eavis A, Robbins T, Arvanitis TN. Evaluation of patient perception towards dynamic health data sharing using blockchain based digital consent with the Dovetail digital consent application: A cross sectional exploratory study. Digit HEALTH 2020;6:2055207620924949. [doi: 10.1177/2055207620924949] [Medline: 32435503]
32. Glomann L, Schmid M, Kitajewa N. Improving the blockchain user experience - an approach to address blockchain mass adoption issues from a human-centred perspective. In: Ahram T, editor. Advances in Artificial Intelligence, Software and Systems Engineering: Springer International Publishing; 2020:608-616. [doi: 10.1007/978-3-030-20454-9_60]
33. Gandhi SS, Patil YN, Netak LD, Gaikwad HR. Usability analysis for blockchain-based applications. In: Kim JH, Singh M, Khan J, Tiwary US, Sur M, Singh D, editors. Intelligent Human Computer Interaction: Springer International Publishing; 2022:349-360. [doi: 10.1007/978-3-030-98404-5_33]
34. Esmaeilzadeh P, Sambasivan M. Patients' support for health information exchange: a literature review and classification of key factors. BMC Med Inform Decis Mak 2017 Apr 4;17(1):33. [doi: 10.1186/s12911-017-0436-2] [Medline: 28376785]
35. Baines R, Stevens S, Austin D, et al. Patient and public willingness to share personal health data for third-party or secondary uses: systematic review. J Med Internet Res 2024 Mar 5;26(1):e50421. [doi: 10.2196/50421] [Medline: 38441944]
36. Abras C, Maloney-Krichmar D, Preece J. User-centered design. In: Bainbridge W, editor. Berkshire Encyclopedia of Human-Computer Interaction Thousand Oaks: Sage Publications; 2004:445-456.
37. Holeman I, Kane D. Human-centered design for global health equity. Information Technology for Development 2020 Jul 2;26(3):477-505. [doi: 10.1080/02681102.2019.1667289]
38. Gao M, Kortum P, Oswald FL. Multi-language toolkit for the System Usability Scale. International Journal of Human–Computer Interaction 2020 Dec 13;36(20):1883-1901. [doi: 10.1080/10447318.2020.1801173]
39. Sauro J, Lewis JR. Quantifying the User Experience: Practical Statistics for User Research: Morgan Kaufmann; 2016.
40. Erler C, Hu S, Danelski A, Stork W, Sunyaev A, Gersch M. Threat modeling to design a decentralized health data management application. In: Rocha Á, Ferrás C, Ibarra W, editors. Information Technology and Systems: Springer International Publishing; 2023:443-455. [doi: 10.1007/978-3-031-33261-6_38]
41. Shostack A. Threat Modeling: Designing for Security: Wiley; 2014.
42. Hyperledger-aries. aries. Hyperledger Aries. 2024. URL: https://github.com/hyperledger-aries/aries [accessed 2025-12-04]
43. LF decentralized trust. Indy. 2024. URL: https://www.lfdecentralizedtrust.org/projects/hyperledger-indy [accessed 2025-12-04]
44. Mobile development with xamarin. .NET.: Microsoft; 2024. URL: https://dotnet.microsoft.com/en-us/apps/xamarin [accessed 2025-12-04]
45. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
46. Venkatesh V, Bala H. Technology acceptance model 3 and a research agenda on interventions. Decision Sciences 2008 May;39(2):273-315 [FREE Full text] [doi: 10.1111/j.1540-5915.2008.00192.x]
47. Harborth D, Pape S. German translation of the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) Questionnaire. SSRN Journal 2018. [doi: 10.2139/ssrn.3147708]
48. Cronbach LJ. Essentials of Psychological Testing, 5th edition: HarperCollins Publ; 1994.
49. Eisinga R, Grotenhuis MT, Pelzer B. The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? Int J Public Health 2013 Aug;58(4):637-642. [doi: 10.1007/s00038-012-0416-3] [Medline: 23089674]
50. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
51. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 2009 May 1;4(3):114-123 [FREE Full text] [doi: 10.5555/2835587.2835589]
52. Nunnally JC, Bernstein IH. Psychometric Theory, 2nd edition: McGraw-Hill; 1978.
53. Wobbrock JO, Findlater L, Gergle D, Higgins JJ. The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures. 2011 May 7 Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; Vancouver BC Canada p. 143-146. [doi: 10.1145/1978942.1978963]
54. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified View1. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
55. Grindrod KA, Li M, Gates A. Evaluating user perceptions of mobile medication management applications with older adults: a usability study. JMIR Mhealth Uhealth 2014 Mar 14;2(1):e11. [doi: 10.2196/mhealth.3048] [Medline: 25099993]
56. Wildenbos GA, Peute L, Jaspers M. Aging barriers influencing mobile health usability for older adults: a literature based framework (MOLD-US). Int J Med Inform 2018 Jun;114:66-75. [doi: 10.1016/j.ijmedinf.2018.03.012] [Medline: 29673606]
57. Gomez-Hernandez M, Ferre X, Moral C, Villalba-Mora E. Design guidelines of mobile apps for older adults: systematic review and thematic analysis. JMIR Mhealth Uhealth 2023 Sep 21;11(1):e43186. [doi: 10.2196/43186] [Medline: 37733401]

XSL•FO
RenderX

58. Griesser A, Mzoughi M, Bidmon S, Cherif E. How do opt-in versus opt-out settings nudge patients toward electronic health record adoption? An exploratory study of facilitators and barriers in Austria and France. BMC Health Serv Res 2024 Apr 8;24(1):439. [doi: 10.1186/s12913-024-10929-w] [Medline: 38589922]
59. Madanian S, Nakarada-Kordic I, Reay S, Chetty T. Patients' perspectives on digital health tools. PEC Innov 2023 Dec;2:100171. [doi: 10.1016/j.pecinn.2023.100171] [Medline: 37384154]

Abbreviations

BI: behavioral intention
HIE: health information exchange
PEOU: perceived ease of use
PU: perceived usefulness
SUS: System Usability Scale
TAM: technology acceptance model

Edited by J Sarvestan; submitted 11.Jun.2025; peer-reviewed by K Mouloudj, Y Zhuang; revised version received 12.Dec.2025; accepted 15.Dec.2025; published 11.Mar.2026.

Please cite as:
Guse R, Hu S, Thiebes S, Erler C, Caridia C, Stork W, Gersch M, Sunyaev A
Patient Perceptions of Blockchain-Based Health Information Exchange: User-Centered Design Study
J Med Internet Res 2026;28:e78849
URL: https://www.jmir.org/2026/1/e78849
doi:10.2196/78849

© Richard Guse, Shanshan Hu, Scott Thiebes, Christina Erler, Catherine Caridia, Wilhelm Stork, Martin Gersch, Ali Sunyaev. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Resistance to Blockchain Adoption in Health Care Organizations: Evidence From a Cross-Sectional Study

Wilfried Bazomanza Nzabandora[1], PhD; Sehl Mellouli[1], PhD; Rohit Nishant[2], PhD

[1]Department of Organizational Information Systems, Faculty of Business Administration, Université Laval, 2325 Rue de la Terrasse, Québec, QC, Canada

[2]Queen's Business School, Queen's University Belfast, Belfast, United Kingdom

Corresponding Author:
Wilfried Bazomanza Nzabandora, PhD
Department of Organizational Information Systems, Faculty of Business Administration, Université Laval, 2325 Rue de la Terrasse, Québec, QC, Canada

Abstract

Background: Traditional health care systems struggle to ensure the security of medical data. To address these issues, organizations are exploring blockchain-based solutions, which offer strong security for managing medical data and transactions. Despite these benefits, adoption remains limited because many health care organizations are hesitant to implement blockchain apps due to perceived risks associated with these apps.

Objective: Our research aims to study the adoption of blockchain-based health care apps in health care organizations by adopting a risk-based approach. Using perceived risk theory (PRT), we developed a research model that links specific perceived risks of blockchain-based health care apps to the intention to adopt them.

Methods: This study used a cross-sectional design to test the research model through an online questionnaire administered to IT professionals from health care organizations in Canada and Africa. IT professionals were selected because they influence technology adoption and possess greater blockchain knowledge than other health care staff. A total of 217 responses were collected; after removing incomplete entries, 194 valid responses remained for analysis. This exceeded the minimum recommended sample size of 100 for structural equation modeling with 8 latent variables, 40 observed variables, a *P* value of .05, and an anticipated effect size of 0.3. Partial least squares structural equation modeling was used to test the hypothesized relationships in the research model.

Results: Harman single-factor test showed that the first factor accounted for 25.376% of the total variance (below 50%), indicating that common method bias was not a concern. Partial least squares structural equation modeling identified key risks influencing the overall perceived risk of blockchain-based health care apps and their adoption in health care organizations. The findings revealed that the relationship between perceived risk and intention to adopt blockchain-based health care apps is significant in Canadian health care organizations but not in African ones. Additionally, the availability of cloud-based blockchain solutions was found to reduce organizations' perceived risk related to insufficient resources for adopting and implementing blockchain-based health care apps.

Conclusions: This study used PRT to identify risks that may hinder the adoption of blockchain-based health care apps. Two risks—medical data disclosure and loss of control over medical data—were identified as blockchain-specific and incorporated as new PRT dimensions, extending the theory for blockchain contexts. The extended PRT can guide future studies examining reluctance to adopt blockchain apps. Results also show that contextual differences between Canada and Africa influence the relationship between perceived risks and adoption intentions. Our findings are beneficial to governments, health care organizations, and blockchain development agencies, as they will be able to implement strategies to mitigate some of the risks perceived by IT professionals related to blockchain-based health care apps.

(J Med Internet Res 2026;28:e77933) doi:10.2196/77933

KEYWORDS

blockchain-based health care apps; perceived risk theory; blockchain adoption; blockchain; risks; health care

XSL•FO
RenderX

Introduction

Overview

Over the past decade, different studies have examined the application of blockchain technology in different business domains [1,2]. Some of these studies are conducted in the health care domain [3,4]. The interest in the health care sector is explained by the profound transformations that blockchain can bring to this field concerning the management of medical data [5-7]. The health care sector is a unique context. It is highly regulated, as it includes sensitive data for which confidentiality must be respected [8]. Blockchain-based apps in this domain may provide more protection for health data by ensuring its confidentiality, availability, interoperability, and integrity [9]. Notable examples include (1) Medifakt (AKT Health), an Estonia-based decentralized blockchain platform for health care that integrates internet of things and machine learning technologies to improve patient care, enhance security, increase data transparency, and promote interoperability across health care systems [10]. The platform supports the transfer, storage, interpretation, and artificial intelligence evaluation of medical data and images [11]. (2) Embleema (Robert Chu), a blockchain solution designed to transform clinical trials and regulatory processes by enabling participants to digitally consent to the collection, storage, and analysis of their health care data [12]. This accelerates drug development while ensuring data integrity and transparency throughout the research lifecycle. (3) Blockpharma (Blockpharma), a blockchain-based platform that strengthens transparency, traceability, and security in pharmaceutical supply chain management [13]. Recording each stage of the supply chain enables patients to verify the authenticity of their medication and supports rigorous quality control through real-time monitoring.

Studies on blockchain in the health care sector are mainly focused on identifying its use cases and opportunities [14]. An analysis of these studies shows that blockchain is used in the health care sector for three main purposes: (1) patient sovereignty over their health care data [15], (2) interoperability of electronic health records [16,17], and (3) integrity and audit trail of medical transactions [18]. Besides the studies that identify blockchain's use cases and opportunities, other studies focus on analyzing the antecedents of the adoption of blockchain-based apps in the health care sector [19]. These studies take a social-cognitive approach, analyzing the factors that would facilitate the adoption of blockchain-based apps at individual or organizational levels [19-24]. Despite these studies, the health care sector is experiencing a low rate of adoption of blockchain-based apps [19]. The low adoption rate of blockchain-based apps in many sectors could be explained by organizations' reluctance to adopt them due to uncertainty about their risks, financial benefits, and performance [25]. The lack of successful use cases for blockchain-based apps prevents potential adopters from perceiving these risks and benefits [25]. This is the case in the health care sector, where blockchain is still maturing, and its various use cases are only for experimental purposes, with very basic functionalities [26].

Against this background, our study focuses on the low adoption of blockchain-based apps in the health care sector. We adopt a social-cognitive approach, using perceived risk theory as a theoretical lens, to identify how and to what extent certain perceived risks of blockchain-based apps can prevent their adoption in health care organizations. Identifying these risks will help explain why health care organizations' intentions to adopt blockchain-based apps do not translate into acts of adoption.

Theoretical Foundation and Conceptual Framework

This study analyzes how health care organizations' perceptions of the risks associated with blockchain-based apps influence their intention to adopt them. Therefore, we use perceived risk theory as the theoretical lens of our study, as it helps to explain the influence of individuals' or organizations' risk perceptions on their intention to adopt an IT.

Perceived Risk Theory

Developed by Bauer [27], perceived risk theory was initially used in studies interested in analyzing the factors that would explain individuals' purchasing decisions [28,29]. Unlike other individual behavior analysis theories, such as the technology acceptance model, which focuses on the factors likely to influence behavior [30], perceived risk theory focuses on factors that could explain individuals' resistance to adopt a particular behavior [31,32]. Thus, this theory establishes a relationship between the perceived risk of a behavior and its adoption by individuals [32].

A perceived risk is defined as a customer's perception of distrust and potential side effects that could affect the purchase of a product or service [33]. Featherman and Pavlou [34] also define this concept as the feeling of uncertainty regarding possible negative consequences of using a product or service. The literature includes different conceptualizations of perceived risk [35]. In some studies, it is conceptualized as a unidimensional concept [36,37] and in other studies as a multidimensional concept [35,38]. The conceptualization of perceived risk as a multidimensional concept makes it possible to consider it as a construct composed of a certain number of dimensions that represent the types of risks that can influence individuals' behavior [35,39]. For example, in a research study on online shopping, authors conceptualized perceived risk as an aggregate construct of 6 risk dimensions: financial, performance, social, physical, psychological, and time [40]. In information systems research, authors consider other types of risks, such as privacy and security risks (SRs), which extend the list of perceived risk dimensions [40].

Conceptual Framework and Research Hypotheses

Our conceptual framework includes 6 types of risks that constitute the perceived risk dimensions of blockchain-based health care apps (BBHAs): SR, failure to gain blockchain benefits risk (FGBBR), organizational readiness risk (ORR), risk of losing control of medical data (RLCMD), risk of disclosing data and medical practices (RDDMP), and immature blockchain ecosystem risk (IBER; Figure 1). These risks and the assumptions of our conceptual framework are explained in the following paragraphs.

XSL•FO
RenderX

Figure 1. Conceptual framework.

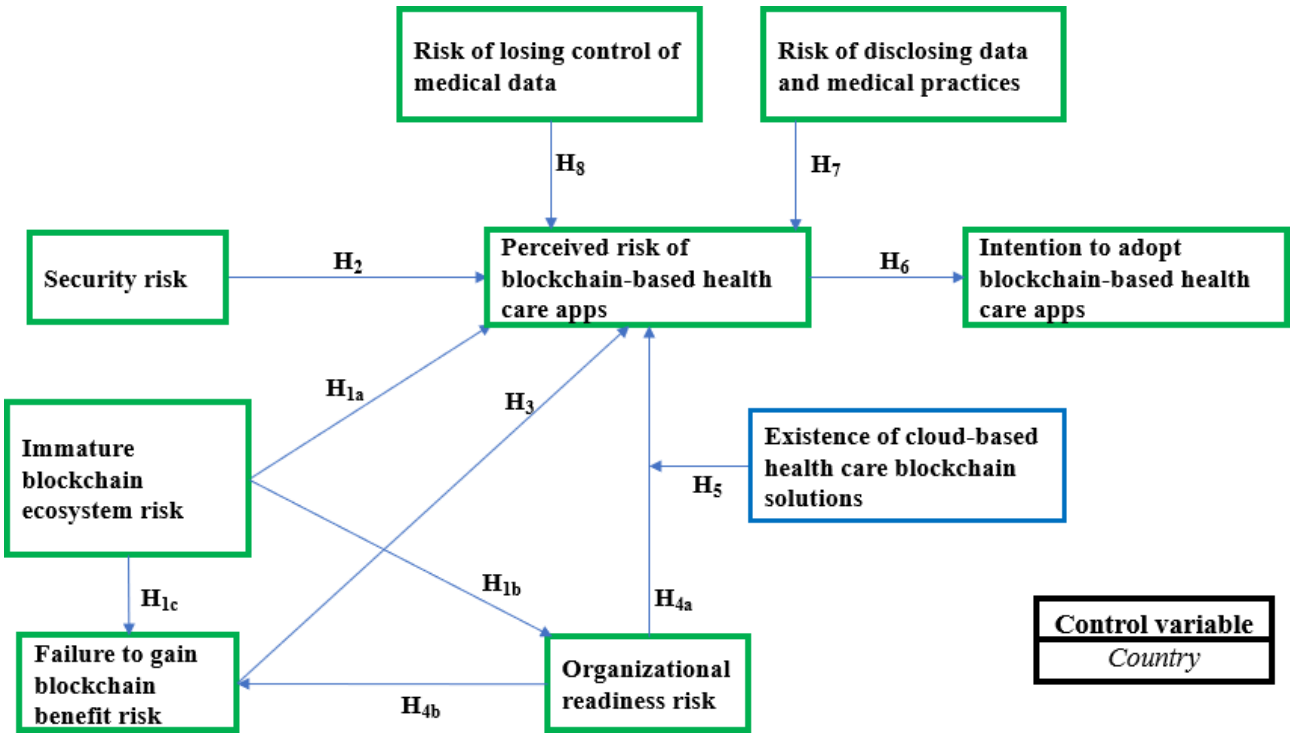

IBER in Health Care

The health care blockchain ecosystem is crucial for the adoption of blockchain-based apps and includes adopters, developers, consultants, auditors, regulators, and standardization bodies [41]. Each stakeholder supports adoption by ensuring app availability, guiding implementation decisions, and establishing regulatory frameworks that protect patient data. A mature ecosystem benefits from an increasing number of adopters, enabling partnerships among health care organizations that enhance the quality of health care services [8]. These collaborations encourage knowledge sharing and improve system security through broader participation in network consensus. Consequently, an immature blockchain ecosystem poses a significant risk, increasing health care organizations' perceived risk when adopting blockchain-based apps. Beyond shaping risk perceptions, the immaturity of the blockchain ecosystem also reduces health care organizations' readiness to adopt blockchain-based apps [42]. Indeed, the limited availability of blockchain-skilled professionals restricts organizations' ability to recruit personnel capable of implementing and maintaining blockchain health care systems [43,44]. Second, the small number of adopters limits access to feedback from prior blockchain investment projects [7,45]. Without such feedback, health care organizations lack clear information about resources required for adoption and use. These consequences of blockchain ecosystem immaturity weaken organizational preparedness and help explain the hesitation of many health care organizations to adopt BBHAs. The risk of the blockchain ecosystem's immaturity can also negatively affect the expected benefits of blockchain-based apps. For example, low blockchain adoption by health care organizations can lead to weak performance of the blockchain network [7]. This weak performance can result from a lack of business partnerships among health care organizations to share medical data through blockchain-based apps [7]. This can negatively impact the expected improvement in health care quality for health organizations.

SR of Blockchain-Based Apps

The ability of IT to guarantee the security of user data is decisive in the choice of its use at both the individual and organizational levels [40,46]. Health care organizations are driven by the adoption of blockchain-based apps to ensure greater security of patient medical data [47,48]. However, the security flaws identified in current blockchain-based apps indicate that these apps remain vulnerable to security attacks [49]. In BBHAs, several factors can contribute to security flaws, including a lack of patient education [50], poor governance of medical data [48], and noncompliance with regulations [51]. The negative consequences of risks associated with these factors for the security of medical data could influence health care organizations' perceptions of risk when adopting blockchain-based apps.

FGBBR

The risk of failure to gain the benefit of a product refers to the possibility that a product does not work as intended and/or does not provide the desired benefits [35]. For BBHAs, the failure to obtain the benefits of their use could be caused by several factors, including the nonintegration between these apps with existing health systems, the resistance of patients to share their medical data, and finally, the lack of alignment between the functionalities of these apps and business needs [7,52]. The risks associated with these factors may prevent health care organizations from fully realizing the benefits of blockchain-based apps, as they could negatively influence health care organizations' risk perceptions regarding adoption.

XSL•FO
RenderX

ORR

Organizational readiness is defined as "the availability of specific organizational resources to adopt new IT innovations" [20]. In this study, we examine the influence of ORR on health care organizations' perceptions of risk related to blockchain-based apps. This risk concerns an organization's lack of IT expertise, IT infrastructure, and financial resources needed to adopt blockchain-based apps. Therefore, we argue that, in the context of blockchain-based apps, inadequate resources for adoption would increase health care organizations' perceived risks. In several studies, the lack of organizational readiness is considered a factor that negatively influences the quality of use of a technology, and subsequently, the expected performance or profitability after its adoption [53,54]. For blockchain-based apps, the lack of competent human resources to use and maintain them, as well as the lack of technical infrastructure to support them, might not enable proper use and continuity of use within organizations. This can negatively influence the quality of services offered by blockchain-based apps and could prevent organizations from gaining the benefits of these apps.

Next, we investigate the moderating effect of cloud-based blockchain solutions on the relationship between ORR and the perceived risk of BBHAs. In fact, the adoption of blockchain-based apps by organizations requires adequate human, financial, and material resources to better exploit these apps and benefit from them [20]. Thus, the lack of these resources increases the risk perception of health care organizations regarding adopting blockchain-based apps [20]. To overcome this requirement, organizations use cloud computing for its technological, economic, and operational benefits [20,55]. In the blockchain context, cloud-based solutions would allow health care organizations to use apps based on this technology as Software-as-a-Service (SaaS) [56]. This significantly reduces the perceived risk of health care organizations regarding blockchain-based apps.

Adoption Intention of BBHAs

IT attributes play a major role in organizational IT adoption intention [57]. Before adopting an innovation, any organization evaluates it on 5 attributes: the complexity of the innovation, its compatibility with the organization, the benefit or advantage that it offers compared to other existing technological choices, and its observability and trialability [57]. In addition to these attributes, the intention to adopt an innovation is also influenced by the risks perceived by potential adopters [31,39,58,59]. Therefore, in the context of blockchain-based app adoption, we also argue that risk perception of these apps could negatively influence health care organizations' adoption intention.

RDDMP

The transparency of the blockchain's ledger ensures that the data recorded on it is visible to authorized network members [60]. For health care structures, ledger transparency can increase the risk of data and medical practice disclosure, with negative consequences for structures that disclose their data and medical practices. For example, when a health structure registers its care protocols on the blockchain ledger, other structures authorized to consult it can learn about those protocols. This can cause a loss of competitive advantage for the structure that discloses its protocols. Thus, the negative consequences associated with the disclosure of data and medical practices raise concerns among health care organizations, increasing their perceived risks of adopting blockchain-based apps.

RLCMD

Integrating blockchain in the health care sector brings a profound transformation in the management of patient medical data [60,61]. This technology puts the ownership of medical data in the hands of patients and places them at the center of its management [62]. Thus, no health structure can consult or use patients' medical data without consent [62]. The obligation to obtain patients' consent before consulting their medical data and the control patients can exercise over how their data are used significantly limit the power of health care structures over patients' medical data. This limitation on health care organizations' power over medical data increases their perception of risks regarding adopting blockchain in medical data management.

Each risk described above gives rise to a research hypothesis. Therefore, our research is based on the following hypotheses:

- H1: The risk of immaturity of the blockchain ecosystem in health care is positively associated with the perceived risk of blockchain-based apps by health care organizations (H1a), the risk of organizational readiness (H1b), and the risk of failing to benefit from blockchain-based apps (H1c).
- H2: SR is positively associated with the perceived risk of blockchain-based apps by health care organizations.
- H3: The risk of failure to gain benefits from blockchain-based apps is positively associated with health care organizations' perceived risk regarding these apps.
- H4: ORR is positively associated with the perceived risk of blockchain-based apps by health care organizations (H4a), as well as the risk of failure to gain benefits of blockchain-based apps (H4b).
- H5: The existence of cloud-based blockchain solutions (ECBS) moderates the relationship between ORR and the perceived risk of BBHAs, such that the positive relationship between ORR and the perceived risk of BBHAs is weak when cloud-based blockchain solutions are present.
- H6: Perceived risk of blockchain-based apps is negatively associated with the health care organizations' intention to adopt these apps.
- H7: RDDMP is positively associated with the perceived risk of blockchain-based apps by health care organizations.
- H8: RLCMD is positively associated with the perceived risk of blockchain-based apps by health care organizations.

Methods

Overview

In this section, we present the research design for our study, which includes 3 sections: research strategy and questionnaire design, data collection, and common method bias.

XSL•FO
RenderX

Research Strategy and Questionnaire Design

Our study adopts a cross-sectional research design, as it analyzes data from a population at a specific point in time [63]. Therefore, it was reported in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines adapted for cross-sectional studies [64]. The dimensions of the perceived risk and their reflexive measures are treated as latent (constructs) and manifest variables (items), respectively. To ensure construct validity, the items for each construct are taken from the literature, which we have subsequently adapted to our research context. We converted all the items to a semantic differential scale (1-10) to minimize common method bias by drawing inspiration from the measurement model proposed by Mitchell [65]. Table 1 shows the definition of constructs and the sources of their items. Table S1 in the Multimedia Appendix 1 provides the questionnaire.

Table . Operationalization of variables.

Construct and subconstructs (items)	Construct and subconstruct definitions	Sources of items
Intention to adopt		
None	A measure of the strength of one's intention to perform a behavior (behavioral intention) [66]	Venkatesh et al [67]; Glover and Benbasat [35]; and Teo et al [68]
Perceived risk: the feeling of uncertainty regarding possible negative consequences of using a product or a service [34]		
Failure to gain blockchain benefits risks	Failure to gain blockchain benefits regarding quality health care services, reduced cost, efficiency and security, and regulatory compliance [69]	Garg et al [69]
Security risk	Concerns that data and information may be open to security concerns, such as hacking, inaccurate information dispersal, and access to sensitive information [70]	Shah et al [71]; Bélanger and Carter [72]; and Hartono et al [73]
Organizational readiness risk	Risk beliefs of the health care organization's lack of financial and technological resources to adopt blockchain-based apps [20,74]	Kouhizadeh et al [70]
Immaturity blockchain ecosystem risk	Lack of community of players, such as blockchain experts and developers, blockchain adopters, and regulators interacting in the health care industry [75]	Alzahrani et al [76]; Kyriakidou et al [77]
Risk of losing control of medical data	Risk beliefs of health professionals regarding the loss of autonomy in managing patients' medical data [78]	Nogueira et al [79]; Ligiane Cristina Braga de Oliveira et al [80]
Risk of disclosing data and medical practices	Risk beliefs about submitting data and medical practices on the blockchain ledger shared by multiple health care organizations [81]	Bansal et al [82]; Bansal et al [83]

Data Collection

We conducted a survey to collect data over a 2-month period. This survey used an online questionnaire. Our study used purposive sampling. This type of sampling allowed us to select a sample from which the most can be learned [84]. Thus, IT professionals from health care facilities were selected as respondents to our study because of their greater knowledge of blockchain, a technology central to our research. The choice of a sample composed of IT professionals from health care facilities in Canada and emerging countries (the Democratic Republic of Congo, Burundi, and Rwanda) enabled us to minimize the risk of a low response rate, which could be caused by the nonparticipation of some respondents in our survey due to their ignorance of blockchain. We contacted 65 health care structures by email to request their participation (55 in Africa and 10 in Canada). At the end of the survey, we received 217 responses. After deleting responses with missing values (23), 194 valid responses were used for data analysis. We excluded responses with missing data for 2 primary reasons. First, these cases contained only demographic information and lacked any data on the study variables, making them unsuitable for analysis. Second, the number of excluded responses was minimal (23/243), representing less than 10% of the sample. Therefore, their removal is unlikely to affect the validity or representativeness of the dataset. The valid responses obtained for the study (n=194) exceeded the minimum recommended sample size (n=100) for structural equation modeling with 8 latent variables, 40 observed variables, *P* value of .05, and anticipated effect size of 0.3 [85]. Thus, the sample size was considered adequate for analyzing the influence of IT professionals' perception of risk regarding blockchain-based apps on their intention to adopt in health care structures. The respondents' demographic information is shown in Table 2.

XSL•FO
RenderX

Table . Demographic information.

Variables and categories	Participants, n (%)
Age (years)	
18 - 30	56 (28.9)
31 - 45	91 (46.9)
46 - 60	42 (21.6)
>60	5 (2.6)
Sex	
Female	67 (34.5)
Male	127 (65.5)
Continent	
Africa	189 (97.4)
America	5 (2.6)
Experience with blockchain-based apps	
Yes	171 (88.1)
No	23 (11.9)
Use of blockchain-based app within health care facility	
Yes	1 (0.5)
No	193 (99.5)

Control Variable

The model estimation includes a control variable, namely country (Figure 1), which we transformed into a binary variable (0=Africa and 1=Canada) before its use in SmartPLS (SmartPLS GmbH). A country's economic, legal, and political contexts play an important role in the assimilation of innovations within organizations [86]. Therefore, there is a significant difference in the assimilation of innovations between organizations in developed and developing countries [86,87]. Unlike developing countries, developed countries are characterized by an economic, political, and regulatory environment favorable to the assimilation of innovations [86]. Given that our study is conducted in Canada and Africa, we also want to analyze how the difference in context (economic, legal, and political) between Canada and African countries could affect the associations between the variables in the model, depending on whether the health care structures are in Canada or Africa. This justifies the introduction of the country as a control variable in the estimation of our theoretical model.

Common Method Bias

Common method bias results "from the fact that the predictor and criterion variables are obtained from the same source or rater, whereas others are produced by the measurement items themselves, the context of the items within the measurement instrument, and/or the context in which the measures are obtained" [88]. Since this research involved self-reported responses collected from a single setting, we will follow the 3-step approach proposed by Guan et al [89] to address the common method bias. First, the measurement items were adopted from well-established existing research. Second, in the survey, we made it clear to all the respondents that there was no right or wrong answer, and they should choose the answers that they considered the most appropriate. We also highlighted the fact that respondents' answers would be kept anonymous. Third, we conducted Harman single-factor test by entering all the measurement variables in an exploratory factor analysis using IBM SPSS 22.0. The sample would have a common method bias problem if a single construct explained more than 50% of the extracted variance. The results of the exploratory factor analysis show that the first factor explained 25.376% of the total variance (less than 50%), indicating that common method bias was not a problem in this dataset.

Ethical Considerations

This research has received ethical approval from the Research Ethics Committee of Université Laval (approval number: 2022 - 427 A-1 R-1/11-12-2023). During the data collection phase, each participant was informed of the research objectives and associated risks. This process enabled participants to provide informed consent by signing the designated consent form. To ensure confidentiality and privacy, data were initially collected using the LimeSurvey platform (LimeSurvey GmbH), which stores survey responses on a server located in Canada. Subsequently, the exported data were securely stored on Université Laval's licensed Microsoft 365 OneDrive server. Participant privacy was further protected by presenting research findings using anonymized data, which prevents the identification of individual respondents. No identification of individual participants is possible in the images of the manuscript or supplementary material. Participation in the study was entirely voluntary and was not contingent upon any monetary or other form of compensation.

XSL•FO
RenderX

Results

Overview

We used the partial least squares structural equation modeling (PLS-SEM) method to assess the hypothesized relationships between our constructs (using SmartPLS 4.1.0.8; SmartPLS GmbH) [90]. We follow a 2-phase approach to analyze the PLS-SEM model proposed by Guan et al [89]: (1) evaluation of the reliability and validity of the measurement model and (2) evaluation of the structural model.

Measurement Model

Table 3 shows the results of the measurement model. The measurement model was tested by examining both convergent and discriminant validity [89]. For testing convergent validity, we assess both item reliability and construct reliability of the measurement model [29]. Throughout the assessment of item reliability, we removed items associated with constructs with factor loading (λ) values below .7. As a result, we deleted 1 RLCMD item (RLCMD1), 2 SR items (SR1 and SR2), 3 IBER items (IBER1, IBER2, and IBER3), 2 ORR items (ORR2 and ORR4), and 4 FGBBR items (FGBBR1, FGBBR3, FGBBR8, and FGBBR10). Removing these items from their respective constructs enabled us to retain a measurement model composed of constructs with items whose factor loadings are greater than or equal to .7, confirming the item reliability of the measurement model. Construct reliability was verified by analyzing composite reliability (CR) and Cronbach α values. Each construct has CR and Cronbach α values that exceed the recommended threshold of .7, demonstrating the construct reliability of the measurement model. Furthermore, the average variance extracted (AVE) for each construct exceeds the minimum required value of .50, which confirms the convergent validity of the constructs in the measurement model [91]. Once convergent validity was established, we assessed discriminant validity. The discriminant validity was assessed using the Fornell and Larcker [92] method. As shown in Table 4, the square root of the AVE of each construct was greater than the correlations it shared with other constructs. For example, the square root AVE for FGBBR is .811, and this is greater than the correlation it shares with other constructs. As this was the case with all constructs, discriminant validity was obtained. Overall, the evidence indicates the adequacy of the measurement model. This leads to the next stage of testing the structural model.

XSL•FO
RenderX

Table . Measurement model results.

Constructs and items	λ[a]	VIF[b]	Cronbach α	CR[c]	AVE[d]
FGBBR[e]			0.896	0.920	0.657
FGBBR2	0.810	2.275			
FGBBR4	0.822	2.533			
FGBBR5	0.836	2.831			
FGBBR6	0.822	2.324			
FGBBR7	0.807	2.369			
FGBBR9	0.765	1.883			
IABHA[f]			0.844	0.896	0.685
IABHA1	0.894	3.213			
IABHA2	0.725	1.546			
IABHA3	0.902	3.327			
IABHA4	0.776	1.704			
IBER[g]			0.877	0.923	0.800
IBER1	0.888	2.811			
IBER3	0.894	2.772			
IBER5	0.901	2.022			
ORR[h]			0.867	0.919	0.790
ORR1	0.910	2.653			
ORR3	0.876	2.144			
ORR5	0.881	2.189			
RDDMP[i]			0.748	0.851	0.656
RDDMP1	0.734	1.728			
RDDMP2	0.804	1.611			
RDDMP3	0.887	2.144			
RLCMD[j]			0.854	0.906	0.764
RLCMD2	0.870	2.550			
RLCMD3	0.819	1.952			
RLCMD4	0.928	2.116			
SR[k]			0.784	0.857	0.600
SR3	0.732	1.612			
SR4	0.722	2.252			
SR5	0.727	2.248			
SR6	0.812	1.867			
SR7	0.838	1.907			

[a]λ: standardized factor loading.

[b]VIF: variance inflation factor.

[c]CR: composite reliability.

[d]AVE: average variance extracted.

[e]FGBBR: failure to gain blockchain benefits risk.

[f]IABHA: adoption intention of blockchain-based health care apps.

[g]IBER: immature blockchain ecosystem risk.

[h]ORR: organizational readiness risk.

XSL•FO
RenderX

[i]RDDMP: risk of disclosing data and medical practices.

[j]RLCMD: risk of losing control of medical data.

[k]SR: security risk.

Table . Results for discriminant validity.

Constructs	FGBBR[a]	IABHA[b]	IBER[c]	ORR[d]	PRBHA[e]	RDDMP[f]	RLCMD[g]	SR[h]
FGBBR	0.811	–0.231	0.439	0.429	0.864	0.217	0.289	0.428
IABHA	–0.231	0.828	–0.056	–0.061	–0.409	–0.232	–0.047	–0.151
IBER	0.439	–0.056	0.894	0.824	0.711	–0.237	0.790	–0.066
ORR	0.429	–0.061	0.824	0.889	0.732	–0.062	0.552	0.163
PRBHA	0.864	–0.409	0.711	0.732	0.511	0.183	0.509	0.455
RDDMP	0.217	–0.232	–0.237	–0.062	0.183	0.81	–0.480	0.459
RLCMD	0.289	-0,047	0.790	0.552	0.509	–0.480	0.874	–0.232
SR	0.428	–0.151	–0.066	0.163	0.455	0.459	–0.232	0.774

[a]FGBBR: failure to gain blockchain benefits risk.

[b]IABHA: adoption intention of blockchain-based health care apps.

[c]IBER: immature blockchain ecosystem risk.

[d]ORR: organizational readiness risk.

[e]PRBHA: perceived risk of blockchain-based health care apps.

[f]RDDMP: risk of disclosing data and medical practices.

[g]RLCMD: risk of losing control of medical data.

[h]SR: security risk.

Structural Model

As the reliability and validity of the variables are established, we examine the structural model to investigate the correlations between the theoretical constructs and test the effect of the moderating variable [93]. The results of the structural model are shown in Table 5 (a summary of the results of hypothesis testing) and Figure 2 (PLS results of the structural model). Before assessing the structural model, we first assess the collinearity issue [94] and the structural model's predictive accuracy [95]. Table 3 shows that all variance inflation factor values of the items are less than 5, indicating that collinearity is not an issue among the constructs [91]. The measure most often used to assess a structural model's predictive accuracy in the PLS-SEM method is the R^2 coefficient of determination. Moreover, a model is judged as interesting if it has a sufficiently large R^2 (>0.25) [95]. The results in Table 6 show that 3 constructs (FGBBR: 0.199, ORR: 0.677, and PRBHA [perceived risk of blockchain-based health care apps]: 0.958) are explained at a substantial level and one (IABHA [adoption intention of blockchain-based health care apps]: 0.163) is explained at a weak level [96,97]. Based on the R^2 values, we can state that the model's predictive accuracy is moderate.

XSL•FO
RenderX

Table . Structural model results and hypothesis testing.

Hypothesis	Path coefficient	*t* statistics	*P* values	Inference
H_{1a}: IBER[a] → PRBHA[b]	0.205	1.692	.05	Supported
H_{1b}: IBER → ORR[c]	0.824	34.743	.00	Supported
H_{1c}: IBER → FGBBR[d]	0.265	2.339	.01	Supported
H_2: SR[e] → PRBHA	0.201	1.930	.03	Supported
H_3: FGBBR → PRBHA	0.503	6.129	.00	Supported
H_{4a}: ORR → PRBHA	0.171	2.011	.02	Supported
H_{4b}: ORR → FGBBR	0.211	1.687	.05	Supported
H_5: ECBS[f] x ORR → PRBHA	−0.085	2.133	.02	Supported
H_6: PRBHA → IABHA[g]	−0.409	1.470	.07	Not supported
H_7: RDDMP[h] → PRBHA	0.121	1.021	.15	Not supported
H_8: RLCMD[i] → PRBHA	0.167	2.068	.02	Supported

[a]IBER: immature blockchain ecosystem risk.

[b]PRBHA: perceived risk of blockchain-based health care apps.

[c]ORR: organizational readiness risk.

[d]FGBBR: failure to gain blockchain benefits risk.

[e]SR: security risk.

[f]ECBS: existence of cloud-based blockchain solutions.

[g]IABHA: adoption intention of blockchain-based health care apps.

[h]RDDMP: risk of disclosing data and medical practices.

[i]RLCMD: risk of losing control of medical data.

XSL•FO
RenderX

Figure 2. Structural model result. ECBS: existence of cloud-based blockchain solutions; FGBBR: failure to gain blockchain benefits risk; IABHA: adoption intention of blockchain-based health care apps; IBER: immature blockchain ecosystem risk; ORR: organizational readiness risk; PRBHA: perceived risk of blockchain-based health care apps; RDDMP: risk of disclosing data and medical practices; RLCMD: risk of losing control of medical data; SR: security risk.

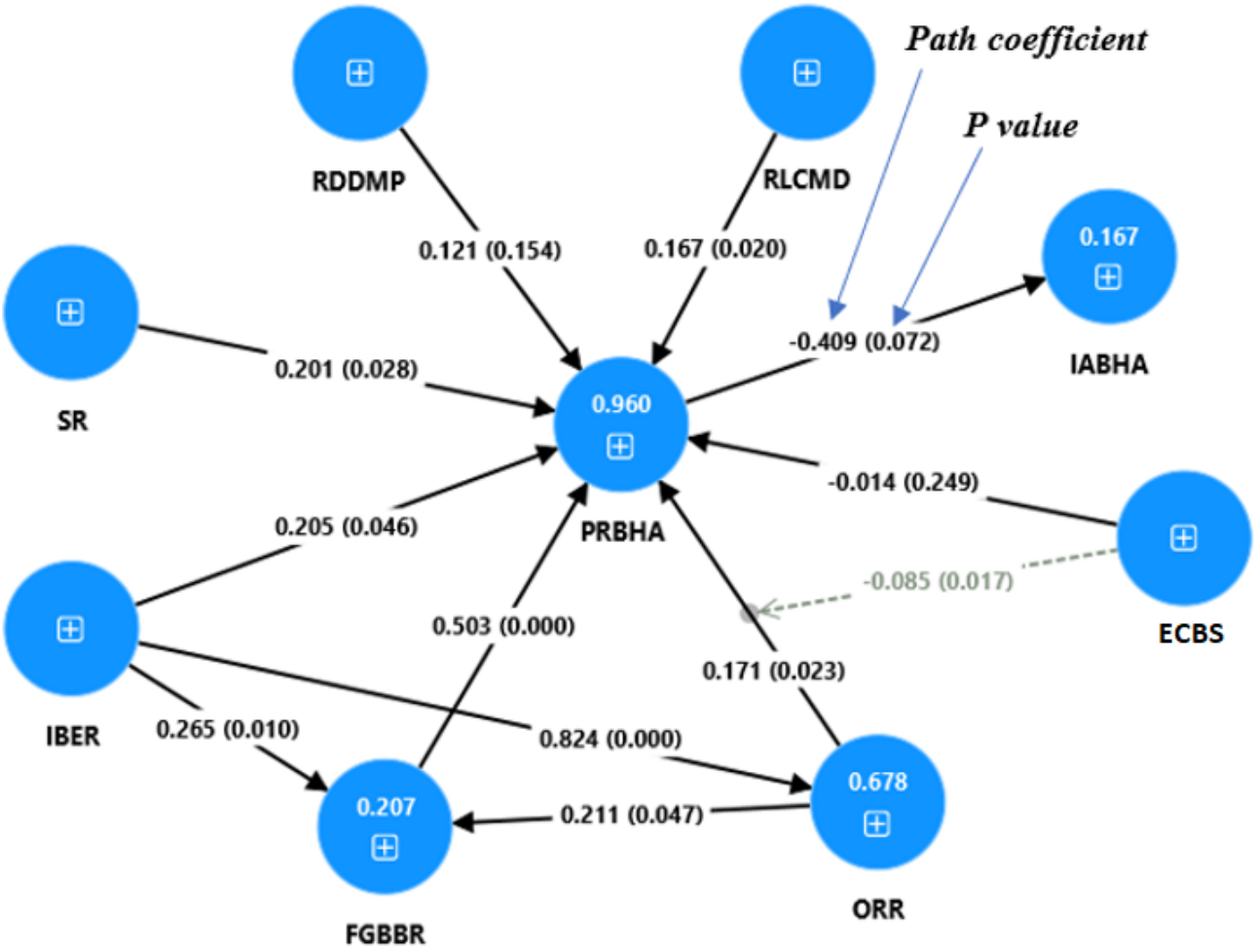

Table . R^2 values.

Constructs	R^2	R^2 adjusted	Results
FGBBR[a]	0.207	0.199	Substantial
IABHA[b]	0.167	0.163	Weak
ORR[c]	0.678	0.677	Substantial
PRBHA[d]	0.960	0.958	Substantial

[a]FGBBR: failure to gain blockchain benefits risk.

[b]IABHA: adoption intention of blockchain-based health care apps.

[c]ORR: organizational readiness risk.

[d]PRBHA: perceived risk of blockchain-based health care apps.

Next, we take a close look at the hypothesized relationships of the structural model. The hypotheses H1a, H2-H4a, and H7-H8 represent the association between PRBHA and its dimensions (IBER, SR, FGBBR, ORR, RDDMP, and RLCMD). Of the 6 dimensions of PRBHA, 5 are positively and significantly associated with PRBHA: RLCMD (β=0.167; *P*=.02), SR (β=0.201; *P*=.03), IBER (β=0.205; *P*=.05), FGBBR (β=0.503; *P*=.00), ORR (β=0.171; *P*=.02), supporting H1a, H2, H3, H4a, and H8. Contrary to what we assumed, the positive association between RDDMP and PRBHA is not significant (β=0.121; *P*=.15). Therefore, hypothesis H7 is not supported. Hypothesis H6 predicted that PRBHA is negatively and significantly associated with IABHA. Contrary to our hypothesis, the negative association between PRBHA and IABHA is not significant (β=−0.409; *P*=.07) and does not support H6.

The hypotheses H1b, H1c, and H4b represent the associations between 3 dimensions of perceived risks (IBER, FGBBR, and ORR). As hypothesized, IBER is positively and significantly

XSL•FO
RenderX

associated with FGBBR (β=0.265; *P*=.01), supporting H1c. Then, IBER is positively and significantly associated with ORR (β=0.824; *P*=*.00*), supporting H1b. Finally, ORR is positively and significantly associated with FGBBR (β=0.211; *P*=.05), supporting H4b.

Finally, hypothesis H5 predicts that ECBS moderates the association between ORR and PRBHA. This hypothesis is also supported (β=–0.085; *P*=.02), indicating that ECBS dampens the positive association between ORR and PRBHA. In other words, the positive association between ORR and PRBHA is weak when cloud-based blockchain solutions are present.

Interpretation of the Impact of the Control Variable on the Model Estimation

Tables 7 and 8 present model estimation results without the control variable and with the control variable. Analysis of these results leads to 2 observations. The first observation concerns the association between the control variable (country) and the dependent variable (IABHA). Table 8 shows that this association is not significant (*P*=.34). This suggests that whether a health care facility is based in Canada or Africa has no significant effect on that facility's intention to adopt a BBHA. The second observation concerns the effect of the control variable on the associations between the variables in the model. A comparison of the results in Tables 7 and 8 shows that only one association is affected by the control variable. Unlike the model without the control variable, the association between PRBHA and IABHA becomes significant in the model with the control variable (*P*=.02). This means that the perception of the risks of BBHAs does not affect a health care facility's intention to adopt these apps in the same way, depending on whether the facility is in Canada or Africa.

Table . Estimation of the model without a control variable.

Associations between variables	Without control variable			
	Original sample (O)	Sample mean (SD)	*t* statistics (O/SD)	*P* values
RDDMP[a] → PRBHA[b]	0.121	0.125 (0.119)	1.021	.15
ECBS[c] → PRBHA	−0.014	−0.014 (0.021)	0.679	.25
PRBHA → IABHA[d]	−0.409	−0.339 (0.278)	1.470	.07

[a]RDDMP: risk of disclosing data and medical practices.

[b]PRBHA: perceived risk of blockchain-based health care apps.

[c]ECBS: existence of cloud-based blockchain solutions.

[d]IABHA: adoption intention of blockchain-based health care apps.

Table . Estimation of the model with control variable.

Associations between variables	With control variable			
	Original sample (O)	Sample mean (SD)	*t* statistics (O/SD)	*P* values
RDDMP[a] → PRBHA[b]	0.121	0.144 (0.109)	1.116	.13
ECBS[c] → PRBHA	−0.014	−0.008 (0.020)	0.703	.24
PRBHA → IABHA[d]	−0.405	−0.436 (0.201)	2.012	.02
Country → IABHA	−0.133	−0.147 (0.319)	0.418	.34

[a]RDDMP: risk of disclosing data and medical practices.

[b]PRBHA: perceived risk of blockchain-based health care apps.

[c]ECBS: existence of cloud-based blockchain solutions.

[d]IABHA: adoption intention of blockchain-based health care apps.

Discussion

Principal Findings

Our study first revealed that the risks of security, failure to obtain blockchain benefits, organizational unpreparedness, disclosure of medical data and practices, and loss of control over medical data increase health care organizations' risk perceptions regarding BBHAs, reducing their intention to adopt these apps. These findings align with research by Glover and Benbasat [35], which emphasized that the SR of personal data and the failure to gain benefits of a product purchased on the web play an important role in buyers' overall risk perception regarding electronic transactions. Moreover, our study revealed that the RDDMP does not play a significant role in the overall risk perception of BBHAs. This result aligns with research that views data disclosure via blockchain not as a risk but as an advantage [98]. This is the case with the study by Dubovitskaya et al [98], in which the authors mention that disclosing data via blockchain in biomedical research allows pharmaceutical companies to disclose information about the production process of drugs. This will enable patients and regulatory bodies to know the context in which certain medical products are produced before they are certified or approved.

XSL•FO
RenderX

The second key finding of this study concerns the reciprocal association between 3 perceived risks of BBHAs: IBER, ORR, and FGBBR. The positive association between IBER and ORR indicates that an underdeveloped blockchain ecosystem—characterized by limited expertise in development, governance, and regulatory clarity—heightens the risk of organizational unpreparedness for adopting BBHAs. Additionally, the association between IBER and FGBBR suggests that ecosystem immaturity reduces the likelihood of realizing expected benefits, such as secure medical data sharing. Similarly, ORR is positively associated with FGBBR, as inadequate infrastructure and insufficiently trained personnel compromise the effective use of these technologies. For example, poor cryptographic key management by untrained users can introduce security vulnerabilities in blockchain-based apps [3], thereby preventing organizations from achieving the intended data protection and operational efficiencies of such apps.

The third key finding of our study concerns the association between IT professionals' perceived risk and their intention to adopt BBHAs. Our results show that organizations perceiving higher risks are less likely to adopt BBHAs, consistent with prior research on IT adoption, including Le et al [99] study on self-service technologies. Their study showed that individual users' concerns about technological risks reduce their willingness to engage with such systems. However, our study reveals regional differences in the association between perceived risk and intention to adopt BBHAs. In Canada, where health care systems are advanced and heavily funded, organizations are more skeptical due to limited performance feedback on BBHAs, leading to reduced motivation to invest. In contrast, health care organizations in developing African countries, which often have fragmented and underresourced systems, view BBHAs as opportunities to improve service delivery. Thus, perceived risks have a weaker impact on their adoption intentions. This regional variation aligns with the cross-national study by Hoover et al [100] on cultural influences in risk perception and consumer behavior. These authors analyzed the influence of perceived brand risk on purchase intention in Mexico and the United States. They found that the cultural difference between Mexico and the United States makes perceived brand risk a less important factor in purchasing behavior in Mexico than in the United States.

The fourth relevant result of our study comes from the assessment of the moderating effect of the existence of blockchain solutions offered as SaaS on the relationship between organizational unreadiness and overall risk perception regarding BBHAs. The results of our study reveal that the possibility of using BBHAs as SaaS would reduce IT professionals' perceived risk of the additional human and material resources required for their use and maintenance in health care settings. This result is in line with several studies that have analyzed the factors of cloud computing adoption in organizations [86,101,102]. In their study, Zhu et al [86] mention that the reduction of overall IT costs influences the use of cloud services. These costs include preliminary investment, upgrading, energy saving, and maintenance of cloud services [86,101].

Theoretical and Practical Contributions

Our study makes a theoretical contribution by adopting and interpreting the perceived risk theory derived from marketing in the context of blockchain adoption in the health care sector. Originally, this theory was used to explain consumers' purchasing decisions in the context of perceived product or service risks [103-105]. In information systems, perceived risk theory is used to analyze how perceived risks can influence the IT adoption decisions [29,93]. However, this theory addresses risks that apply broadly to IT as a whole, making it unsuitable for analyzing risks within specific IT categories. The results of our study highlight 2 risks specific to blockchain-based apps that are likely to influence the overall perceived risk regarding these apps and, therefore, their adoption in health care structures: RDDMP and RLCMD. Integrating these 2 risks as new constructs into the perceived risk theory contributes to its extension, making it more suitable for risk analysis in blockchain-based apps. Future research analyzing barriers to blockchain adoption in the health care sector can incorporate the 2 new risks into its theoretical model. Moreover, conducting our study in Canada and Africa allowed us to understand how contextual differences between Canada and Africa affect how risks are perceived in BBHAs and the intention to adopt them in Canadian and African health care organizations. This contributed to the generalizability of our results.

As a practical contribution, our study highlights the risks IT professionals in health care organizations perceive regarding BBHAs that may prevent their adoption in health care organizations. Knowledge of these risks would benefit the development agencies of blockchain-based apps and health care organizations [106,107]. For development agencies of blockchain-based apps, our findings will guide their advocacy discourse on the adoption of blockchain-based apps among health care organizations. These agencies can exploit these findings to formulate recommendations to mitigate risks that IT professionals perceive or educate IT professionals of health care organizations on strategies to avoid certain perceived risks [108]. In addition, development agencies of blockchain-based apps can take advantage of our findings to develop BBHAs that minimize some risks that IT professionals perceive regarding these apps. For health care organizations, knowing the risks of BBHAs brings 2 major implications. First, these organizations handle sensitive data that require better protection against security attacks [8]. Through our results, these organizations understand that despite the hype about the security of blockchain-based apps, these apps are exposed to SRs. Therefore, they must implement appropriate security policies to strengthen the security of these apps [107]. Second, our study's results inform health care organizations' decisions regarding adopting BBHAs. We believe that a good decision to adopt these apps should consider their benefits and potential risks [109]. Knowing these risks would allow health care organizations to implement strategies to mitigate them and improve the use of BBHAs.

Limitations

This study has 2 main limitations. The first limitation concerns the small number of IT professionals working in health care

XSL•FO
RenderX

structures that have already adopted BBHAs or those in the process of adopting them. In addition, our study had a low participation of IT professionals from health care structures in Canada. The low participation of IT professionals from structures that use BBHAs, structures that are in the process of adopting these apps, and structures in Canada prevents variance in the dependent variable of our study, namely the intention to adopt BBHAs. Second, our research analyzes blockchain adoption in the health care sector, which is experiencing a low adoption rate of BBHAs. All these limitations constitute obstacles to the generalization of the results of this research.

Directions for Future Research

We propose several avenues for future research. First, the conceptual model was validated using data from IT professionals in health care organizations. This choice was justified by the low knowledge of blockchain among health care professionals and patients [110,111]. It is essential to note that the successful implementation of IT in a health care organization necessitates the involvement of multiple stakeholders, including IT professionals, health care professionals, patients, health care organization managers, and health care regulators [112]. Perceptions of each stakeholder regarding the risks associated with BBHAs may influence health care organizations' intentions to adopt these apps [113]. Therefore, we recommend that other researchers conduct similar research involving many stakeholders to better assess the factors that influence the intention of health care organizations to adopt BBHAs.

Second, our research focused primarily on the factors that may hinder health care organizations' intention to adopt blockchain-based apps. Our choice is explained by the low adoption of blockchain-based apps in health care organizations [76]. Therefore, the data for our research were primarily collected from health care organizations that have not yet adopted a blockchain-based app and those that are in the process of doing so. Future research may analyze the adoption challenges of these apps by collecting data from health care organizations that have already adopted these apps. Thus, health care organizations interested in adopting blockchain-based apps will be better informed about the challenges of adoption before implementing them. This helps reduce the perceived risks associated with blockchain-based health apps in health care organizations, as they can develop strategies to overcome the adoption challenges of these apps [114].

Conclusion

Our study sought to analyze the impact of IT professionals' perceived risk concerning the intention to adopt BBHAs in health care facilities. Drawing on perceived risk theory, we identified 6 dimensions of risk that may explain IT professionals' overall risk perception of BBHAs in health care facilities. Data collected from IT professionals in health care facilities through an online survey allowed us to evaluate the hypotheses of our conceptual model. The analysis of these data, using SmartPLS 4 software, yielded important results. The analysis of the correlations between perceived risk and its dimensions revealed that the latter are all positively associated with perceived risk. However, the RDDMP does not contribute significantly to the overall perception of IT professionals' risks regarding using BBHAs. Then, our results revealed that the significance of the negative association between the general perception of risks of BBHAs and the intention to adopt these apps varies depending on whether the health care organization is in Africa or Canada. This association is nonsignificant in African health care organizations, while it is significant in Canadian health care organizations. Finally, the results of our study revealed that the existence of blockchain solutions in the form of SaaS reduces the overall perception of risks of IT professionals regarding BBHA, given the lack of resources (human and technical) necessary for their use in healthcare structures.

Our study makes the following contributions. First, we highlight the types of risks likely to prevent the adoption of BBHAs in health care organizations despite their interest in adopting these apps.

Second, we show that RDDMP and RLCMD contribute to the overall risk perception of BBHAs. These risks are inherent to the characteristics of blockchain, and the fact that they are not considered in the perceived risk theory makes it unsuitable for risk analysis in blockchain-based apps. The perceived risk theory includes risks that are general to any IT and the integration of these new risks contributes to its extension by making it more suitable for the analysis of risks in blockchain-based apps. Finally, we show that the difference in context between Canada and Africa makes the overall perception of risk of BBHAs differently associated with the intention to adopt these apps between Canadian and African health care organizations. Our findings are beneficial to governments, health care organizations, and blockchain development agencies, which can implement strategies to mitigate some of the risks perceived by IT professionals related to BBHAs.

Acknowledgments

No generative artificial intelligence (AI) tools were used at any stage in the preparation of this manuscript.

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Measurement of items.
[DOCX File, 18 KB - jmir_v28i1e77933_app1.docx]

References

1. Cho S, Lee K(, Cheong A, No WG, Vasarhelyi MA. Chain of values: examining the economic impacts of blockchain on the value-added tax system. J Manag Inf Syst 2021 Apr 3;38(2):288-313. [doi: 10.1080/07421222.2021.1912912]
2. Sarker S, Henningsson S, Jensen T, Hedman J. The use of blockchain as a resource for combating corruption in global shipping: an interpretive case study. J Manag Inf Syst 2021 Apr 3;38(2):338-373. [doi: 10.1080/07421222.2021.1912919]
3. McGhin T, Choo KKR, Liu CZC, He DB. Blockchain in healthcare applications: research challenges and opportunities. J Netw Comput Appl 2019 Jun;135:62-75. [doi: 10.1016/j.jnca.2019.02.027]
4. Tandon A, Dhir A, Islam A, Mäntymäki M. Blockchain in healthcare: a systematic literature review, synthesizing framework and future research agenda. Comput Ind 2020 Nov;122:103290. [doi: 10.1016/j.compind.2020.103290]
5. Zaabar B, Cheikhrouhou O, Jamil F, Ammi M, Abid M. HealthBlock: a secure blockchain-based healthcare data management system. Comput Netw 2021 Dec;200:108500. [doi: 10.1016/j.comnet.2021.108500]
6. Cadoret D, Kailas T, Velmovitsky P, Morita P, Igboeli O. Proposed implementation of blockchain in British Columbia's health care data management. J Med Internet Res 2020 Oct 23;22(10):e20897. [doi: 10.2196/20897] [Medline: 33095183]
7. Palas M, Bunduchi R. Exploring interpretations of blockchain's value in healthcare: a multi-stakeholder approach. Inf Technol People 2020 Jun 24;34(2):453-495. [doi: 10.1108/ITP-01-2019-0008]
8. Esmaeilzadeh P, Mirzaei T. The potential of blockchain technology for health information exchange: experimental study from patients' perspectives. J Med Internet Res 2019 Jun 20;21(6):e14184. [doi: 10.2196/14184] [Medline: 31223119]
9. Hylock RH, Zeng X. A blockchain framework for patient-centered health records and exchange (HealthChain): evaluation and proof-of-concept study. J Med Internet Res 2019 Aug 31;21(8):e13592. [doi: 10.2196/13592] [Medline: 31471959]
10. How blockchain and AI improve healthcare supply chains. Medium. 2022. URL: https://blog.medifakt.com/how-blockchain-and-ai-improve-healthcare-supply-chains-5d856d6e46d5 [accessed 2026-03-06]
11. Adams HS. Top 10: blockchain platforms in healthcare. Healthcare Digital. 2024. URL: https://healthcare-digital.com/technology-and-ai/top-10-blockchain-platforms-in-healthcare [accessed 2026-03-06]
12. Embleema is a technology platform for registries and real-world studies. Embleema. 2025. URL: https://www.embleema.com/aboutus [accessed 2026-03-06]
13. Dutta S. Blockchain for healthcare: benefits, use cases & real-world examples. Turing. 2025. URL: https://www.turing.com/resources/blockchain-for-healthcare [accessed 2026-03-06]
14. Sanka AI, Irfan M, Huang I, Cheung RCC. A survey of breakthrough in blockchain technology: adoptions, applications, challenges and future research. Comput Commun 2021 Mar;169:179-201. [doi: 10.1016/j.comcom.2020.12.028]
15. Pirtle C, Ehrenfeld J. Blockchain for healthcare: the next generation of medical records? J Med Syst 2018 Aug 10;42(9):172. [doi: 10.1007/s10916-018-1025-3] [Medline: 30097733]
16. Chen Y, Ding S, Xu Z, Zheng HD, Yang SL. Blockchain-based medical records secure storage and medical service framework. J Med Syst 2019 Jan;43(1). [doi: 10.1007/s10916-018-1121-4]
17. Vardhini B, Dass SN, Sahana R, Chinnaiyan R. A blockchain based electronic medical health records framework using smart contracts. Presented at: 2021 International Conference on Computer Communication and Informatics (ICCCI); Jan 27-29, 2021. [doi: 10.1109/ICCCI50826.2021.9402689]
18. Jayaraman R, Salah K, King N. Improving opportunities in healthcare supply chain processes via the internet of things and blockchain technology. Int J Healthc Inf Syst Inform 2019;14(2):49-65. [doi: 10.4018/IJHISI.2019040104]
19. Liang TP, Kohli R, Huang HC, Li ZL. What drives the adoption of the blockchain technology? A fit-viability perspective. J Manag Inf Syst 2021 Apr 3;38(2):314-337. [doi: 10.1080/07421222.2021.1912915]
20. Clohessy T, Acton T. Investigating the influence of organizational factors on blockchain adoption. Ind Manang Data Syst 2019 Aug 12;119(7):1457-1491. [doi: 10.1108/IMDS-08-2018-0365]
21. Folkinshteyn D, Lennon M. Braving Bitcoin: a technology acceptance model (TAM) analysis. J Inf Technol Case Appl Res 2016 Oct;18(4):220-249. [doi: 10.1080/15228053.2016.1275242]
22. Janssen M, Weerakkody V, Ismagilova E, Sivarajah U, Irani Z. A framework for analysing blockchain technology adoption: integrating institutional, market and technical factors. Int J Inf Manage 2020 Feb;50:302-309. [doi: 10.1016/j.ijinfomgt.2019.08.012]
23. Rashideh W. Blockchain technology framework: current and future perspectives for the tourism industry. Tour Manag 2020 Oct;80:104125. [doi: 10.1016/j.tourman.2020.104125]
24. Yoo K, Bae K, Park E, Yang T. Understanding the diffusion and adoption of Bitcoin transaction services: the integrated approach. Telemat Inform 2020 Oct;53:101302. [doi: 10.1016/j.tele.2019.101302]

XSL•FO
RenderX

25. Maythu Y, Kwok AOJ, Teh PL. Blockchain technology diffusion in tourism: evidence from early enterprise adopters and innovators. Heliyon 2024 Jan 30;10(2):e24675. [doi: 10.1016/j.heliyon.2024.e24675] [Medline: 38304774]
26. Agbo CC, Mahmoud QH, Eklund JM. Blockchain technology in healthcare: a systematic review. Healthcare (Basel) 2019 Apr 4;7(2):56. [doi: 10.3390/healthcare7020056] [Medline: 30987333]
27. Bauer RA. Consumer behaviour as risk taking. In: Robert SH, editor. Dynamic Marketing for a Changing World: Proceedings of the 43rd National Conference of the American Marketing Association: American Marketing Association; 1960:389-398.
28. Schlegel D, Uenal Y. A perceived risk perspective on narrow artificial intelligence. Presented at: 25th Pacific Asia Conference on Information Systems (PACIS); Jun 20-24, 2021 URL: https://aisel.aisnet.org/pacis2021/44 [accessed 2026-03-29]
29. Martins C, Oliveira T, Popovič A. Understanding the Internet banking adoption: a unified theory of acceptance and use of technology and perceived risk application. Int J Inf Manage 2014 Feb;34(1):1-13. [doi: 10.1016/j.ijinfomgt.2013.06.002]
30. Gefen D, Karahanna E. Trust and TAM in online shopping: an integrated model. MIS Q 2003;27(1):51-90. [doi: 10.2307/30036519]
31. Khedmatgozar HR. The impact of perceived risks on internet banking adoption in Iran: a longitudinal survey. Electron Commer Res 2021 Mar;21(1):147-167. [doi: 10.1007/s10660-021-09475-y]
32. Verhagen T, Meents S, Tan YH. Perceived risk and trust associated with purchasing at electronic marketplaces. Eur J Inf Syst 2006 Dec;15(6):542-555. [doi: 10.1057/palgrave.ejis.3000644]
33. Littler D, Melanthiou D. Consumer perceptions of risk and uncertainty and the implications for behaviour towards innovative retail services: the case of Internet Banking. J Retailing Consum Serv 2006 Nov;13(6):431-443. [doi: 10.1016/j.jretconser.2006.02.006]
34. Featherman MS, Pavlou PA. Predicting e-services adoption: a perceived risk facets perspective. Int J Hum Comput Stud 2003 Oct;59(4):451-474. [doi: 10.1016/S1071-5819(03)00111-3]
35. Glover S, Benbasat I. A comprehensive model of perceived risk of e-commerce transactions. Int J Electron Commer 2010 Dec;15(2):47-78. [doi: 10.2753/JEC1086-4415150202]
36. Pavlou PA, Gefen D. Building effective online marketplaces with institution-based trust. Inf Syst Res 2004 Mar;15(1):37-59. [doi: 10.1287/isre.1040.0015]
37. Pavlou PA. Consumer acceptance of electronic commerce: integrating trust and risk with the technology acceptance model. Int J Electron Commer 2003 Apr;7(3):101-134. [doi: 10.1080/10864415.2003.11044275]
38. Spiekermann S, Paraschiv C. Motivating human–agent interaction: transferring insights from behavioral marketing to interface design. Electron Commer Res 2002 Jul;2(3):255-285. [doi: 10.1023/A:1016062632182]
39. Chiu C, Wang ETG, Fang Y, Huang H. Understanding customers' repeat purchase intentions in B2C e‐commerce: the roles of utilitarian value, hedonic value and perceived risk. Inf Syst J 2014 Jan;24(1):85-114. [doi: 10.1111/j.1365-2575.2012.00407.x]
40. Cocosila M, Archer N, Yuan Y. Early investigation of new information technology acceptance: a perceived risk - motivation model. Commun Assoc Inf Syst 2009;25. [doi: 10.17705/1CAIS.02530]
41. Erol I, Ar IM, Ozdemir AI, et al. Assessing the feasibility of blockchain technology in industries: evidence from Turkey. J Enterp Inf Manag 2021 Apr 12;34(3):746-769. [doi: 10.1108/JEIM-09-2019-0309]
42. Akella GK, Wibowo S, Grandhi S, Mubarak S. A systematic review of blockchain technology adoption barriers and enablers for smart and sustainable agriculture. Big Data Cogn Comput 2023;7(2):86. [doi: 10.3390/bdcc7020086]
43. Kaushal RK, Kumar N, Panda SN, Kukreja V. Immutable smart contracts on blockchain technology: its benefits and barriers. Presented at: 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO); Sep 3-4, 2021. [doi: 10.1109/ICRITO51393.2021.9596538]
44. White GRT. Future applications of blockchain in business and management: a Delphi study. Strat Change 2017 Sep;26(5):439-451. [doi: 10.1002/jsc.2144]
45. Komulainen R, Nätti S. Barriers to blockchain adoption: empirical observations from securities services value network. J Bus Res 2023 Apr;159:113714. [doi: 10.1016/j.jbusres.2023.113714]
46. Raut RD, Priyadarshinee P, Gardas BB, Jha MK. Analyzing the factors influencing cloud computing adoption using three stage hybrid SEM-ANN-ISM (SEANIS) approach. Technol Forecast Soc Change 2018 Sep;134:98-123. [doi: 10.1016/j.techfore.2018.05.020]
47. Griggs KN, Ossipova O, Kohlios CP, Baccarini AN, Howson EA, Hayajneh T. Healthcare blockchain system using smart contracts for secure automated remote patient monitoring. J Med Syst 2018 Jun 6;42(7):130. [doi: 10.1007/s10916-018-0982-x] [Medline: 29876661]
48. Khezr S, Moniruzzaman M, Yassine A, Benlamri R. Blockchain technology in healthcare: a comprehensive review and directions for future research. Appl Sci (Basel) 2019;9(9):1736. [doi: 10.3390/app9091736]
49. Zamani E, He Y, Phillips M. On the security risks of the blockchain. J Comput Inform Syst 2020 Nov 1;60(6):495-506. [doi: 10.1080/08874417.2018.1538709]
50. Radanović I, Likić R. Opportunities for use of blockchain technology in medicine. Appl Health Econ Health Policy 2018 Oct;16(5):583-590. [doi: 10.1007/s40258-018-0412-8] [Medline: 30022440]

XSL•FO
RenderX

51. Hussien HM, Yasin SM, Udzir SNI, Zaidan AA, Zaidan BB. A systematic review for enabling of develop a blockchain technology in healthcare application: taxonomy, substantially analysis, motivations, challenges, recommendations and future direction. J Med Syst 2019 Sep 14;43(10):1-35. [doi: 10.1007/s10916-019-1445-8] [Medline: 31522262]
52. Gaur N. Blockchain challenges in adoption. Manag Financ 2019 Aug 12;46(6):849-858. [doi: 10.1108/MF-07-2019-0328]
53. Ragowsky A, Licker PS, Gefen D. Organizational IT Maturity (OITM): a measure of organizational readiness and effectiveness to obtain value from its information technology. Inform Syst Manage 2012 Mar;29(2):148-160. [doi: 10.1080/10580530.2012.662104]
54. Mouzakitis S, Askounis D. A knowledge-based framework for measuring organizational readiness for the adoption of B2B integration systems. Inf Syst Manag 2010 Jan;27(3):253-266. [doi: 10.1080/10580530.2010.493842]
55. Hosseini L, Tang S, Mookerjee V, Sriskandarajah C. A Switch in Time Saves the Dime: a model to reduce rental cost in cloud computing. Inf Systems Res 2020 Sep;31(3):753-775. [doi: 10.1287/isre.2019.0912]
56. Tripathi S, Mishra V. Determinants of cloud computing adoption: a comparative study. Pac Asia J Assoc Inf Syst 2019;11(3):28-54. [doi: 10.17705/1pais.11303] [Medline: 2500368009]
57. Rogers EM. Diffusion of Innovations, 5th edition: The Free Press; 2003.
58. Hanafizadeh P, Khedmatgozar HR. The mediating role of the dimensions of the perceived risk in the effect of customers' awareness on the adoption of Internet banking in Iran. Electron Commer Res 2012 May;12(2):151-175. [doi: 10.1007/s10660-012-9090-z]
59. Lee MC. Factors influencing the adoption of internet banking: an integration of TAM and TPB with perceived risk and perceived benefit. Electron Commer Res Appl 2009 May;8(3):130-141. [doi: 10.1016/j.elerap.2008.11.006]
60. Govindan K, Nasr AK, Saeed Heidary M, Nosrati-Abarghooee S, Mina H. Prioritizing adoption barriers of platforms based on blockchain technology from balanced scorecard perspectives in healthcare industry: a structural approach. Int J Prod Res 2023 Jun 3;61(11):3512-3526. [doi: 10.1080/00207543.2021.2013560]
61. Mittal S, Ghosh M. A novel two-level secure access control approach for blockchain platform in healthcare. Int J Inf Secur 2023 Aug;22(4):799-817. [doi: 10.1007/s10207-023-00664-4]
62. Porsdam Mann S, Savulescu J, Ravaud P, Benchoufi M. Blockchain, consent and prosent for medical research. J Med Ethics 2021 Apr 13;47(4):244-250. [doi: 10.1136/medethics-2019-105963] [Medline: 32366703]
63. Wang X, Cheng Z. Cross-sectional studies: strengths, weaknesses, and recommendations. Chest 2020 Jul;158(1S):S65-S71. [doi: 10.1016/j.chest.2020.03.012] [Medline: 32658654]
64. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Epidemiology (Sunnyvale) 2007 Nov;18(6):800-804. [doi: 10.1097/EDE.0b013e3181577654] [Medline: 18049194]
65. Mitchell VW. Consumer perceived risk: conceptualisations and models. Eur J Mark 1999 Feb 1;33(1/2):163-195. [doi: 10.1108/03090569910249229]
66. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Manage Sci 1989 Aug;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
67. Venkatesh V, Thong JYL, Xu X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Q 2012 Mar 1;36(1):157-178. [doi: 10.2307/41410412]
68. Teo HH, Wei KK, Benbasat I. Predicting intention to adopt interorganizational linkages: an institutional perspective. MIS Q 2003 Mar 1;27(1):19-49. [doi: 10.2307/30036518]
69. Garg P, Gupta B, Chauhan AK, Sivarajah U, Gupta S, Modgil S. Measuring the perceived benefits of implementing blockchain technology in the banking sector. Technol Forecast Soc Change 2021 Feb;163:120407. [doi: 10.1016/j.techfore.2020.120407]
70. Kouhizadeh M, Saberi S, Sarkis J. Blockchain technology and the sustainable supply chain: theoretically exploring adoption barriers. Int J Prod Econ 2021 Jan;231(2021):107831. [doi: 10.1016/j.ijpe.2020.107831]
71. Shah MH, Peikari HR, Yasin NM. The determinants of individuals' perceived e-security: evidence from Malaysia. Int J Inf Manage 2014 Feb;34(1):48-57. [doi: 10.1016/j.ijinfomgt.2013.10.001]
72. Bélanger F, Carter L. Trust and risk in e-government adoption. J Strateg Inf Syst 2008 Jun;17(2):165-176. [doi: 10.1016/j.jsis.2007.12.002]
73. Hartono E, Holsapple CW, Kim KY, Na KS, Simpson JT. Measuring perceived security in B2C electronic commerce website usage: a respecification and validation. Decis Support Syst 2014 Jun;62(2014):11-21. [doi: 10.1016/j.dss.2014.02.006]
74. Iacovou CL, Benbasat I, Dexter AS. Electronic data interchange and small organizations: adoption and impact of technology. MIS Q 1995 Dec 1;19(4):465-485. [doi: 10.2307/249629]
75. Santos FM, Eisenhardt KM. Organizational boundaries and theories of organization. Organ Sci 2005 Oct;16(5):491-508. [doi: 10.1287/orsc.1050.0152]
76. Alzahrani S, Daim T, Choo KKR. Assessment of the blockchain technology adoption for the management of the electronic health record systems. IEEE Trans Eng Manage 2023;70(8):2846-2863. [doi: 10.1109/TEM.2022.3158185]
77. Kyriakidou V, Michalakelis C, Sphicopoulos T. Assessment of information and communications technology maturity level. Telecommun Policy 2013 Feb;37(1):48-62. [doi: 10.1016/j.telpol.2012.08.001]

XSL•FO
RenderX

78. Beck JT, Rahinel R, Bleier A. Company worth keeping: personal control and preferences for brand leaders. J Consum Res 2020 Feb 1;46(5):871-886. [doi: 10.1093/jcr/ucz040]
79. Nogueira M, Dias F, Santos V. Sustainable mobility choices: exploring the impact of consumers' values, attitudes, perceived behavioural control and subjective norms on the likelihood to choose sustainable mobility options. J Consum Behav 2023 Mar;22(2):511-528 [FREE Full text] [doi: 10.1002/cb.2144]
80. Braga de Oliveira Friaes LC, Poleto T, Negreiros Clemente TR, Ramos CC, de Carvalho VDH, Nepomuceno TCC. Relational governance in the ITO: the role of self-determination theory and the mediating effect of control perceived. Acad Rev Latinaom Ad 2023 Jun 29;36(2):177-198. [doi: 10.1108/ARLA-10-2022-0182]
81. Malhotra NK, Kim SS, Agarwal J. Internet Users' Information Privacy Concerns (IUIPC): the construct, the scale, and a causal model. Inf Syst Res 2004 Dec;15(4):336-355. [doi: 10.1287/isre.1040.0032]
82. Bansal G, Zahedi F", Gefen D. The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online. Decis Support Syst 2010 May;49(2):138-150. [doi: 10.1016/j.dss.2010.01.010]
83. Bansal G, Zahedi FM, Gefen D. Do context and personality matter? Trust and privacy concerns in disclosing private information online. Inf Manag 2016 Jan;53(1):1-21. [doi: 10.1016/j.im.2015.08.001]
84. Merriam SB, Tisdell EJ. Qualitative Research: A Guide to Design and Implementation, 4th edition: San Francisco, CA: Jossey-Bass; 2016:368.
85. Soper D, Soper D. A-priori sample size calculator for structural equation models. Free Statistics Calculators. 2014. URL: https://www.danielsoper.com/statcalc/calculator.aspx?id=89 [accessed 2024-03-17]
86. Zhu K, Kraemer KL, Xu S. The process of innovation assimilation by firms in different countries: a technology diffusion perspective on e-business. Manage Sci 2006 Oct;52(10):1557-1576. [doi: 10.1287/mnsc.1050.0487]
87. Zhu K, Kraemer K, Xu S. Electronic business adoption by European firms: a cross-country assessment of the facilitators and inhibitors. Eur J Inf Syst 2003 Dec;12(4):251-268. [doi: 10.1057/palgrave.ejis.3000475]
88. Podsakoff PM, MacKenzie SB, Lee JY, Podsakoff NP. Common method biases in behavioral research: a critical review of the literature and recommended remedies. J Appl Psychol 2003 Oct;88(5):879-903. [doi: 10.1037/0021-9010.88.5.879] [Medline: 14516251]
89. Guan Z, Hou F, Li B, Phang CW, Chong AYL. What influences the purchase of virtual gifts in live streaming in China? A cultural context - sensitive model. Inf Syst J 2022 May;32(3):653-689 [FREE Full text] [doi: 10.1111/isj.12367]
90. Hair JF, Sarstedt M, Ringle CM, Mena JA. An assessment of the use of partial least squares structural equation modeling in marketing research. J Acad Mark Sci 2012 May;40(3):414-433. [doi: 10.1007/s11747-011-0261-6]
91. Hair JF, Howard MC, Nitzl C. Assessing measurement model quality in PLS-SEM using confirmatory composite analysis. J Bus Res 2020 Mar;109:101-110. [doi: 10.1016/j.jbusres.2019.11.069]
92. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. J Mark Res 1981 Feb;18(1):39-50. [doi: 10.1177/002224378101800104]
93. Im I, Kim Y, Han HJ. The effects of perceived risk and technology type on users' acceptance of technologies. Inf Manag 2008 Jan;45(1):1-9. [doi: 10.1016/j.im.2007.03.005]
94. Becker JM, Ringle CM, Sarstedt M, Völckner F. How collinearity affects mixture regression results. Mark Lett 2015 Dec;26(4):643-659. [doi: 10.1007/s11002-014-9299-9]
95. Hair JF, Astrachan CB, Moisescu OI, et al. Executing and interpreting applications of PLS-SEM: updates for family business researchers. J Fam Bus Strateg 2021 Sep;12(3):100392. [doi: 10.1016/j.jfbs.2020.100392]
96. Bawack RE, Wamba SF, Carillo KDA. Exploring the role of personality, trust, and privacy in customer experience performance during voice shopping: evidence from SEM and fuzzy set qualitative comparative analysis. Int J Inf Manage 2021 Jun;58:102309. [doi: 10.1016/j.ijinfomgt.2021.102309]
97. Jr. JFH, Matthews LM, Matthews RL, Sarstedt M. PLS-SEM or CB-SEM: updated guidelines on which method to use. Int J Multivar Data Anal 2017;1(2):107. [doi: 10.1504/IJMDA.2017.087624]
98. Dubovitskaya A, Novotny P, Xu Z, Wang F. Applications of blockchain technology for data-sharing in oncology: results from a systematic literature review. Oncology (Williston Park, NY) 2020;98(6):403-411. [doi: 10.1159/000504325] [Medline: 31794967]
99. Le NT, Rao Hill S, Troshani I. Perceived control and perceived risk in self-service technology recovery. J Comput Inf Syst 2022 Jan 2;62(1):164-173. [doi: 10.1080/08874417.2020.1756533]
100. Hoover RJ, Green RT, Saegert J. A cross-national study of perceived risk: does perceived risk have the same effect on consumers in a foreign country as in the United States. J Mark 1978;42(3):102-108. [doi: 10.1177/002224297804200318]
101. Ali O, Shrestha A, Osmanaj V, Muhammed S. Cloud computing technology adoption: an evaluation of key factors in local governments. Inf Technol People 2020 Apr 28;34(2):666-703. [doi: 10.1108/ITP-03-2019-0119]
102. Cervone HF. An overview of virtual and cloud computing. OCLC Syst Serv 2010 Aug 24;26(3):162-165. [doi: 10.1108/10650751011073607]
103. Simcock P, Sudbury L, Wright G. Age, perceived risk and satisfaction in consumer decision making: a review and extension. J Mark Manag 2006 Apr;22(3-4):355-377. [doi: 10.1362/026725706776861163]
104. Phamthi VA, Nagy Á, Ngo TM. The influence of perceived risk on purchase intention in e - commerce—Systematic review and research agenda. Int J Consum Stud 2024 Jul;48(4):e13067 [FREE Full text] [doi: 10.1111/ijcs.13067]

XSL•FO
RenderX

105. Kim LH, Qu H, Kim DJ. A study of perceived risk and risk reduction of purchasing air‐tickets online. J Travel Tour Mark 2009 May;26(3):203-224. [doi: 10.1080/10548400902925031]
106. Abdelwahed IM, Ramadan N, Hefny HA. Cybersecurity risks of blockchain technology. Int J Comput Appl 2020;177(42):8-14. [doi: 10.5120/ijca2020919922]
107. Behzadi L, Joseph J. Blockchain Security Considerations: CRC Press; 2024:73-92. [doi: 10.1201/9781003462033-6]
108. Bordeianu AA, Popescu DE. Blockchain variables and possible attacks: a technical survey. Computers 2025;14(12):567. [doi: 10.3390/computers14120567]
109. Drljevic N, Aranda DA, Stantchev V. Perspectives on risks and standards that affect the requirements engineering of blockchain technology. Comput Stand Inter 2020 Mar;69:103409. [doi: 10.1016/j.csi.2019.103409]
110. Govindarajan UH, Narang G, Singh DK, Yadav VS. Blockchain technologies adoption in healthcare: overcoming barriers amid the hype cycle to enhance patient care. Technol Forecast Soc Change 2025 Apr;213:124031. [doi: 10.1016/j.techfore.2025.124031]
111. Attaran M. Blockchain technology in healthcare: challenges and opportunities. Int J Healthc Manag 2022 Jan 2;15(1):70-83. [doi: 10.1080/20479700.2020.1843887]
112. Prashar A, Sunder M V. Blockchain barriers in hospitals: a stakeholder theoretic perspective. Benchmarking 2026 Jan 1;33(1):1-32. [doi: 10.1108/BIJ-03-2023-0185]
113. Nanda S, Nanda S. Blockchain adoption in health market: a systems thinking and modelling approach. J Asia Bus Stud 2022 Mar 23;16(2):396-405. [doi: 10.1108/JABS-11-2020-0457]
114. Onik MMH, Aich S, Yang J, Kim CS, Kim HC. Chapter 8 - blockchain in healthcare: challenges and solutions. In: Dey N, Das H, Naik B, Behera HS, editors. Big Data Analytics for Intelligent Healthcare: Academic Press; 2019:197-226. [doi: 10.1016/B978-0-12-818146-1.00008-8]

Abbreviations

AVE: average variance extracted
BBHA: blockchain-based health care apps
CR: composite reliability
ECBS: existence of cloud-based blockchain solutions
FGBBR: failure to gain blockchain benefits risk
IABHA: adoption intention of blockchain-based health care apps
IBER: immature blockchain ecosystem risk
ORR: organizational readiness risk
PLS-SEM: partial least squares structural equation modeling
PRBHA: perceived risk of blockchain-based health care apps
PRT: perceived risk theory
RDDMP: risk of disclosing data and medical practices
RLCMD: risk of losing control of medical data
SaaS: Software-as-a-Service
SR: security risk
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 22.May.2025; peer-reviewed by I Cerovecki, S Mohanadas, TA El-Hafeez; revised version received 19.Jan.2026; accepted 20.Jan.2026; published 08.Apr.2026.

Please cite as:
Bazomanza Nzabandora W, Mellouli S, Nishant R
Resistance to Blockchain Adoption in Health Care Organizations: Evidence From a Cross-Sectional Study
J Med Internet Res 2026;28:e77933
URL: https://www.jmir.org/2026/1/e77933
doi:10.2196/77933

© Wilfried Bazomanza Nzabandora, Sehl Mellouli, Rohit Nishant. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Discourse, Secondary Victimization, and Psychological Harm: Mixed-Methods Analysis of System Justification in the #MeToo Movement

Harsh Parekh[1*], PhD; Shriya Thakkar[2*], PhD; Paras Bhatt[3], PhD; Patricia Akello[4], PhD

[1]Paul H. Chook Department of Information Systems and Statistics, Zicklin School of Business, Baruch College, City University of New York, New York, NY, United States

[2]Department of Prevention and Community Health, Milken Institute School of Public Health, George Washington University, 950 New Hampshire Avenue, Washington DC, DC, United States

[3]Department of Information Systems, Supply Chain & Analytics, College of Business, The University of Alabama in Huntsville, Huntsville, AL, United States

[4]Department of MIS and Cybersecurity, College of Business, University of Montana, Missoula, MT, United States

[*]these authors contributed equally

Corresponding Author:
Shriya Thakkar, PhD
Department of Prevention and Community Health, Milken Institute School of Public Health, George Washington University, 950 New Hampshire Avenue, Washington DC, DC, United States

Abstract

Background: The #MeToo movement, initiated in 2006 and amplified on social media in 2017, mobilized women worldwide to share experiences of sexual harassment and assault online. While the movement increased awareness, it also revealed deep social divisions in digital spaces. Supportive discussions promoted solidarity and healing, whereas antagonistic responses reinforced backlash and secondary victimization. In India, the Indian Entertainment Industry (IEI) became a focal point where survivors' disclosures highlighted structural gender inequalities. These polarized reactions function as digital-health signals, reflecting stigma, psychosocial distress, and conditions that shape women's safety and mental well-being. Examining these narratives as indicators of public health risk helps identify patterns of structural inequity and secondary mental health burdens among survivors.

Objective: This study examined online discourse surrounding #MeToo to identify forms of system-justifying narratives on social media and to assess how #MeTooIndia exposed institutional inequities within the IEI.

Methods: This mixed-methods study comprised 2 components. In study 1, natural language processing was applied to analyze global #MeToo Twitter (subsequently rebranded X) discourse. From an initial corpus of 350,000 tweets, 205,082 were preprocessed, and sentiment and stance detection analysis identified 18,416 tweets expressing negative attitudes toward the movement. Latent Dirichlet allocation topic modeling extracted 22 topics, 12 of which aligned with system-justification categories, revealing distinct lexical and semantic patterns related to gender, institutional, and power dynamics. Two trained coders manually annotated a subsample to ensure conceptual clarity and interrater reliability. Study 2 involved qualitative, semistructured interviews with 20 academic experts in film, gender, and media studies to gather opinions on how #MeTooIndia influenced institutional discourse in the IEI and how these dynamics translate into digital and mental health risks.

Results: Analysis of #MeToo Twitter discourse in study 1 identified 4 primary forms of system justification: by gender, by the institutional system, by backlash, and by victim-blaming. Gender and institutional system justifications were the most prevalent. Study 2 reinforced these findings, revealing how experts perceived #MeTooIndia as both empowering and constrained by entrenched institutional and cultural barriers. Together, our findings highlight the dual function of social media in promoting collective advocacy while reproducing conditions linked to gender-based violence, psychological stress, and reduced help-seeking—key digital and mental health concerns.

Conclusions: This mixed-methods study reveals that digital discourse surrounding #MeToo often sustains existing gender and institutional hierarchies rather than dismantling them. Across Twitter data and expert interviews, gender and institutional system justifications emerged as dominant narratives, highlighting how online spaces can reinforce structural inequities while appearing progressive. Although #MeToo amplified visibility and awareness, its potential for lasting institutional change remains limited. These findings underscore the need for trauma-informed digital governance, public health recognition of online hostility as a psychosocial risk, and frameworks that situate digital activism to institutional reforms that support safety and mental well-being.

(J Med Internet Res 2026;28:e75533) doi:10.2196/75533

XSL•FO
RenderX

KEYWORDS

harassment; mental health; inequities; gender; mixed methods

Introduction

Background

Initially founded by the New York-based social activist Tarana Burke in 2006, #MeToo aimed to empower women and support victims of sexual assaults in the United States. A decade later, the movement gained worldwide attention when actress Alyssa Milano popularized #MeToo with a viral Twitter (subsequently rebranded X) trend. The movement broadened awareness while reflecting the emotional strain tied to large-scale public disclosure [1,2]. From a public health perspective, this increase represents more than social awareness. It reflects large-scale digital expressions of trauma that reveal population-level exposure to gender-based violence (GBV) and related mental health consequences [3-5].

Gender equality movements such as #MeToo are often viewed as women-led initiatives [6], which shapes how people engage with them online. This framing also explains why public responses became sharply divided or polarized. Expressions of solidarity promoted healing and collective efficacy [5,7,8], while hostile or dismissive comments reproduced stigma and secondary victimization [9]. Prior analyses have shown that discussions around #MeToo contained both destructive elements, such as ridicule, rape references, and scandal language, and constructive ones emphasizing education, leadership, and rights [5,9]. Such polarization is health-relevant because antagonistic discourse can intensify psychological distress, discourage help-seeking, and perpetuate unsafe disclosure climates for survivors [3,10-12]. Opponents of #MeToo viewed the movement as threatening social identity and existing hierarchies [13,14], further sustaining online environments that can undermine survivor well-being and collective healing.

In this context, while the #MeToo movement resonated worldwide, its impact on the Indian Entertainment Industry (IEI), particularly Bollywood, was notably profound, serving as a catalyst that broke long-standing barriers of silence within the entertainment sector [15]. Within the IEI, disclosures emerged rapidly and publicly, revealing deep-seated gender hierarchies and power asymmetries that had long silenced survivors. For instance, late Bollywood choreographer Saroj Khan remarked, “Bollywood provides bread, at least doesn’t abandon women after rape” in the wake of the #MeToo movement in India [16]. Such comments exemplify how public figures can normalize or excuse sexual violence, reflecting the cultural narratives that sustain institutional denial and perpetuate psychological harm for survivors [17]. These justifications and dismissive reactions have measurable health implications. Studies show that negative or invalidating responses to #MeToo disclosures are associated with increased posttraumatic stress, anxiety, and feelings of isolation among survivors [18,19].

In the Indian context, such backlash often manifests through public shaming, social exclusion, and reputational damage, creating conditions of secondary victimization that discourage reporting [17,20,21]. Although social media facilitates visibility and mobilization, it also amplifies exposure to stigma and retraumatization, turning disclosure spaces into sites of both empowerment and psychological risk [1]. Thus, our analysis addresses calls to understand the system-justification narratives within #MeToo that reproduce structural inequities and generate secondary mental health burdens for survivors [9,19,22].

Alongside social and moral support, #MeTooIndia also generated negative reactions in the form of backlash, victim blaming, and social system justification [23]. In patriarchal contexts, women survivors often face reputational risks and moral scrutiny that discourage disclosure and silence voices of resistance [20,24]. Backlash also stemmed from a perception that #MeTooIndia threatened group identity and the existing social order, prompting many to defend traditional hierarchies through system-justifying narratives [25,26]. Previous research on #MeToo links disbelief and ridicule following disclosure to trauma recurrence, anxiety, and avoidance behaviors [9,10,22]. This study examines online reactions to #MeToo through the lens of a digital health discourse. Social-media discourse surrounding GBV functions not only as a forum for advocacy but also as a public-health signal that reflects collective trauma and emotional distress [4,18]. Such digital expressions can inform health care providers, mental health professionals, and public health practitioners about emerging patterns of psychosocial stress and the need for trauma-informed care and community-level interventions [27,28].

We treat the justification narratives surrounding #MeToo as infoveillance indicators, which are digital signals of GBV and secondary victimization addressing limitations identified in prior research [5,29]. Hostile or invalidating comments constitute secondary victimization [30,31], which research links to increased posttraumatic stress, anxiety, and social withdrawal among survivors [9,11,19,22]. Within patriarchal or celebrity-dominated contexts such as India’s entertainment industry, such backlash can reinforce stigma and reproduce structural inequities that exacerbate mental health burden [17,20]. Positioning the analysis within this digital-health framing enables examination of how social-media environments impact both public-health awareness and secondary psychological harm following GBV disclosure [3].

This Study

Overview

Victim blaming stands as a key factor deterring women from coming forward or identifying their aggressors [5,32-34]. Building on this context, the current study focuses on entrenched gender hierarchies, institutional status quo, pushback on social movement, and victim-blaming as mechanisms of social system justification within the #MeToo movement. In extending prior work, our study moves beyond the initial disclosure phase to examine how system-justifying narratives evolve as digital signals of GBV [4,29,35,36]. Study 1 conducts a large-scale text mining analysis of Twitter data to identify and categorize system-justifying narratives that emerged during the #MeToo

XSL•FO
RenderX

movement. Study 2 complements this with qualitative interviews focused on the IEI, offering deeper insights into how these narratives manifest within a specific sociocultural and institutional setting. Together, these studies offer a mixed-methods perspective on how digital backlash and justification discourse operate as online expressions of GBV, with implications for mental health and digital well-being.

Objectives

The objectives of this study were to (1) identify the forms of system-justifying discourse that emerged on social media during the #MeToo movement and (2) examine how #MeTooIndia revealed its effects on institutional systems within the IEI.

Conceptualization: Social Media as a System Justifier

Social conditioning often creates circumstances in which seeking justice can, unfortunately, intensify the harassment experienced by victims [5,22,37]. Secondary victimization occurs through social media when the perpetrators face legal problems, and the community unjustly holds the victim responsible [38,39]. System justification theory suggests that "there is a general ideological motive to justify the existing social order," and this motive is responsible for the "internalization of inferiority among the members of the disadvantaged group" [40]. System justification theory has been applied across disciplines, including public health, using established scales to assess how individuals endorse or legitimize existing social systems [41,42]. An example of a belief measure is the view that society is structured so that people generally get what they deserve. It is important to distinguish psychological system justification, which reflects unconscious cognitive motives to legitimize existing hierarchies, from strategic or systemic resistance, which involves deliberate actions by individuals or institutions to preserve power and resist reform [43,44].

System justifiers create a fear of judgment or retaliation for sexual victims, drastically reducing the disclosure [19,45,46]. A substantial number of tweets criticized individuals, including prominent figures, who had blamed victims of sexual harassment for their experiences, thereby undermining the feminist movement [22,47]. The inclination to rationalize current social structures may be particularly strong in the context of gender dynamics compared with other intergroup settings, such as racial, religious, or ethnic groups, as gender relations are deeply interdependent within social and economic systems [25]. The types of system justifications are listed in Table 1.

Table . Type of system justifications.

Justification type	Definition
Gender system justification	The "extent one accepts gender hierarchy as legitimate" (page 1). This includes upholding traditional gender roles, norms, and power dynamics within the system [48].
Institutional system justification	This type of system justification focuses on the "legitimation of institutions and authorities, denial or minimization of system problems or shortcomings" (page 881). It involves defending workplaces, industries, legal systems, or social hierarchies [40].
Victim-blaming system justification	This type of justification is known as the derogation effect. It indicates that "victim is deserving of the injustice and responsible for his/her own misfortune" (page 5912). This behavior often involves questioning the credibility of the victim, scrutinizing their agency, and fabricating their experiences [39].
Backlash system justification	This type of justification indicates "vociferous and frequently violent opposition" toward "any actual or potential gains made by disadvantaged groups" (page 133-134). This behavior includes disapproving of facts and studies recognizing the prevalence of sexual harassment, highlighting the threat to individual freedom, and sympathizing with defamation received by the accused [49].

Similarly, investigating the 2017 Women's March movement, researchers found that mostly in the case of men (who identify gender as less important to one's self-concept), exposure to gender movement increased system justification tendency [48]. Furthermore, men who expressed support for gender system justification believed that #MeToo unjustly tarnished the reputation of men. Evidence indicates that the endorsement of sexist ideologies is connected to individuals' perspectives on sexual harassment and their stance toward the #MeToo movement [50].

While defending the existing gender system may represent the main narrative for system justification within such movements, it is not the only type represented on social media. Reports have highlighted the limited accountability of businesses and institutions during the #MeToo movement [51]. Sociocultural and organizational systems often reinforce beliefs and behaviors that perpetuate sexual harassment and misconduct. Furthermore, victims bear the personal cost of legal representation, while institutions support the defense of powerful offenders. Such dynamics contribute to secondary victimization and institutional retaliation, maintained through system-justifying mechanisms [52].

Witnessing online victimization has been shown to promote dehumanization of victims, leading to perceptions that they are less deserving of respect or protection [21,53]. Such cognitive and emotional distancing contributes to harmful social norms that normalize violence and reduce collective empathy, which are both recognized as public health concerns [10,19].

XSL•FO
RenderX

Victim-blaming justifications are produced when individuals experience identity threats and defensively protect their self-image by partially or fully attributing blame to the victim [5,18]. These justifications are reinforced by gender-powered dynamics and male sexual aggression, which in turn increase women's vulnerability and sustain beliefs rationalizing the just world hypothesis [53-55].

Victim-blaming justifications not only shift attribution away from perpetrators but also contribute to resistance against the broader goals of the #MeToo movement. This form of discourse often frames the movement as an overreaction or a threat to individual freedom, thereby undermining efforts to address GBV and promote safer social environments [5,17]. Such responses reflect a broader tendency to preserve positive perceptions of existing social systems, even when these systems sustain inequities that have significant public health implications for victims' mental and physical well-being [19,29,49]. Earlier research indicates that the movement could provoke even more intense negative responses from those who are unsympathetic or hold contrary beliefs about its objectives [5,36,48].

Understanding how systems justify and reproduce themselves is key to identifying the social and psychological barriers that impede collective action. From a public health perspective, such insights can guide targeted strategies to disrupt these barriers, strengthen community engagement, and sustain online activism that promotes equity and safety.

Methods

Study Design

The global #MeToo movement revealed significant regional variation in engagement and public discourse [35,56]. Recognizing that social movements addressing GBV have broad implications for population health and social well-being, we adopted a mixed-methods design to examine how digital narratives reinforce or challenge systemic inequities. This study comprised two components: (1) a quantitative analysis of tweets to assess the global diffusion and framing of #MeToo via social media, and (2) a qualitative study using in-depth interviews focused on #MeTooIndia to contextualize the movement's local impacts.

This research used a sequential explanatory mixed-methods design in which study 1 directly informed study 2 [57]. The quantitative analysis of Twitter discourse identified 4 dominant forms of system justification, and these empirically derived categories guided the development of the semistructured interview guide (Multimedia Appendix 1). Study 2 was therefore designed to contextualize, elaborate, and extend the patterns observed in study 1 by examining how similar justification mechanisms were understood within the localized context of the IEI. During interpretation, findings from both studies were integrated to provide a consolidated understanding of how digital narratives and institutional responses interact during #MeToo.

In the first study, we collected tweets without geographic restrictions to minimize sampling bias and capture cross-cultural expressions of system justification. Regional and cultural contexts can shape how individuals legitimize or contest social hierarchies online, influencing public perceptions of gender justice and safety—key determinants of health equity. Guided by system justification theory, which posits that individuals tend to rationalize and defend existing social structures even when they sustain inequality, we used natural language processing–based topic classification to identify patterns of justification within global #MeToo discussions. We then complemented these findings with an in-depth case study of #MeTooIndia, drawing on expert interviews to explore how these justifications manifest within the Indian sociocultural and institutional context, which led to our second study.

In our second study, we examined the social and behavioral consequences of the #MeTooIndia movement within the context of IEI. Prior research underscores that diverse cultural settings can obscure interpretation unless analyzed within a specific context [58]. IEI, particularly Bollywood, offered a compelling case as the movement gained strong momentum, generated widespread public engagement, and empowered women actors to voice experiences of sexual misconduct. To capture these dynamics, we conducted interviews with research scholars specializing in media or film studies to gain nuanced insight into the social and institutional responses surrounding #MeTooIndia.

While the 2 studies used different methods, they collectively offer a multidimensional understanding of digital activism and its societal implications. The natural language processing–based analysis in study 1 captures large-scale public discourse and patterns of system justification, whereas the expert interviews in study 2 provide contextual depth and interpretation of those narratives. Together, these complementary approaches offer a more comprehensive view of how such digital feminist movements intersect with cultural norms, institutional accountability, and collective well-being. Consistent with Gorissen et al [22], who found that Twitter hosted 61.4% of sexual victimization disclosures during #MeToo, our focus on Twitter data ensured a consistent and extensive dataset for examining system-justifying narratives that influence public attitudes toward gender equity and safety.

Transparency and Openness

The processed data supporting our conclusions, along with the Python (Python Software Foundation) syntax required for reproducing the results, can be obtained from the authors upon request. All analyses were performed using Python libraries. Please note that this study's design and analysis were not preregistered because the research was exploratory and aimed at developing theoretical and methodological insights rather than testing predefined hypotheses (Figure 1).

XSL•FO
RenderX

Figure 1. Overall methodological flow illustrating the mixed-methods design across study 1 and study 2.

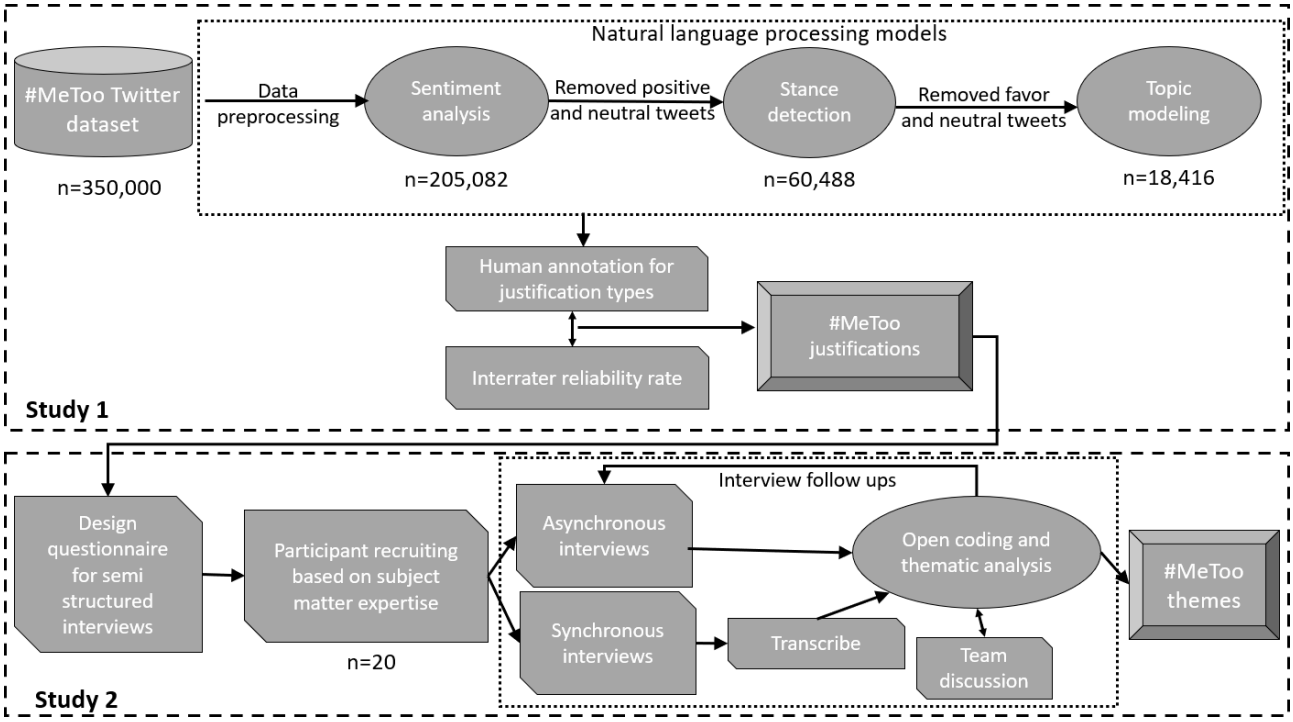

Ethical Considerations

For quantitative study 1, all data were obtained from publicly accessible sources on Twitter. Posts shared under the hashtags (#MeToo and #metoo) represent public communications intended for open audiences. No personal or identifying information about individual users (such as usernames, user IDs, or direct mentions) was retained, as all data were deidentified before analysis. All analyses were conducted on aggregated text data, focusing on linguistic and thematic patterns rather than individual users or specific tweets. Results are therefore presented in group form to ensure that no inferences can be made about any specific user. Given ongoing concerns about the potential for reidentification of social media users, even when direct identifiers are removed, we took additional precautions when presenting illustrative material (sample tweets). All sample content used for explanation in this paper was paraphrased, and any original handles were replaced with neutral placeholders to prevent traceability. Finally, while we do not publicly share any of the analyzed data, we outline more details in the "Data Availability" section [59,60].

For study 2, all participants were experienced academics familiar with research ethics and the principles of informed consent and institutional review. Each received an information sheet describing this study's objectives, procedures, and data-use policy. Participation was voluntary. Participants were interviewed strictly in their professional capacity, and no identifiable personal data was collected. The authors analyzed qualitative materials that were recorded in a fully anonymized form. This study did not involve interaction or intervention beyond the collection of anonymous responses. Accordingly, the work met criteria for minimal-risk social research and did not constitute human participants research requiring institutional review board oversight under the US Common Rule (45 CFR §46.102) [61]. Additionally, the research adhered to the ethical principles of the Declaration of Helsinki [62] and the World Health Organization's (2001) guidelines for ethical qualitative inquiry [63].

To protect privacy and confidentiality, all interviews were conducted and recorded in English with explicit participant consent. Transcripts were deidentified, assigned numeric codes, and stored on password-protected devices accessible only to the research team. No identifiable information, images, or supplementary materials are included in this paper. Participation was entirely voluntary, and no monetary compensation was provided. Table 2 summarizes the demographic characteristics of the scholars and experts who participated in this study.

XSL•FO
RenderX

Table . Demographic profile of scholars and experts interviewed for study 2 (N=20). The names of the respondents referred to in this research are pseudonyms, and identities are kept strictly confidential to maintain anonymity. Professional or educational engagement and country of professional or educational attainment are as reported.

Name	Identified by the pronouns used	Current professional or educational engagement	Country of professional or educational affiliation
Daniel	He or his	Professor	United States
Smita	She or hers	Associate professor	United States
Sarita	She or hers	Associate professor	United States
Raya	She or hers	Professor	United States
Ramya	She or hers	Assistant professor	India
Irrfan	He or his	Associate professor	Turkey
Karanjeet	He or his	Assistant professor	India
Latika	She or hers	Doctoral candidate	United Kingdom
Nazreen	She or hers	Doctoral candidate	United States
Naved	He or his	Doctoral candidate	Turkey
Mita	She or hers	Assistant professor	United States
Neeta	She or hers	Graduate student	United States
Kirti	She or hers	Assistant professor	India
Kamal	She or hers	Assistant professor	United States
Rahul	He or his	Doctoral candidate	United States
Smitra	She or hers	Graduate student	India
Tupur	She or hers	Doctoral candidate	India
Ipsita	She or hers	Graduate student	India
Ayanti	She or hers	Graduate student	India
Jincy	She or hers	Doctoral candidate	India

Study 1: Quantitative Twitter Analysis

Study Design and Data Collection

We used publicly accessible tweets as the main source of our data for conducting a topic-wise classification. We used the Twitter API platform to extract tweets from #metoo and #MeToo for a period of 1 month from November 2017 to December 2017. The extraction resulted in a corpus of 350,000 tweets. This dataset provides a rich corpus for examining public discourse and sentiment around the movement, offering important insights into how individuals communicate experiences of GBV and solidarity in digital spaces. The data included several identifiers such as postdate, user description, and engagement metrics about the likes, retweets, and replies, allowing us to examine both message content and patterns of social amplification.

The global #MeToo conversation began on October 15, 2017, when Alyssa Milano invited survivors of sexual harassment to share their stories publicly. The movement quickly transcended linguistic and cultural boundaries, inspiring localized expressions such as *#YoTambien* in Spanish-speaking countries, *#BalanceTonPorc* (“squeal on your pig”) in France, *#quellavoltache* in Italy, and *#RiceBunny* in China—a creative response to online censorship. We strategically collected the data after the first sporadic spike (which turned out to be the highest over time). Figure 2 shows Google Trends (Google LLC) in the #MeToo movement. As we wanted to capture various system justifications, the period after the first sporadic spike was deemed appropriate [22].

XSL•FO
RenderX

Figure 2. Temporal dynamics of Google search interest in #MeToo from November 2017 to December 2017.

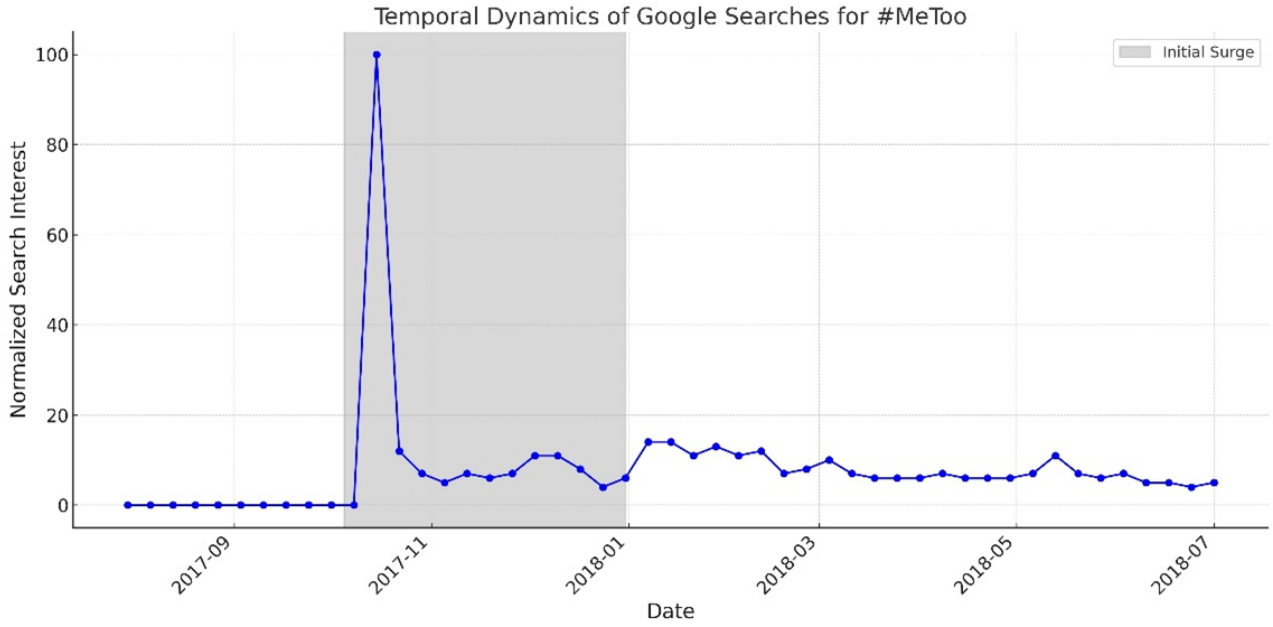

Textbox 1 showcases exemplar tweets illustrating the various justification types that serve to diminish the impact of the #MeToo movement. All tweets in Textbox 1 are paraphrased and deidentified to prevent traceability to individual users. Simple removal of identifiers is insufficient, as advanced fuzzy-matching and text-retrieval techniques can enable reidentification of social media posts. To protect user anonymity, we replaced all handles with neutral placeholders (eg, ABC, @DEF, and @GHI) and reworded examples while preserving their conceptual meaning.

Textbox 1. Sample tweets for justification types.

- Gender system justification: “we have to "go back" to giving men authority over women in the workplace, home & society in general so they will STOP taking advantage #MeToo.”
- Institutional system justification: “@ABC You've been in #Hollywood for 60 years, you've worked with, kept quiet about, helped to cover up, so why now?”
- Victim-blaming system justification: “@DEF Yeah it works in mysterious ways because she knew he was a sexual predator and still worked with him”
- Backlash system justification: “@GHI: Texas Attorney General's top aide mocks 'pathetic' #MeToo movement and calls women's marchers 'c*nts'”

Analysis

Preprocessing of the #MeToo Twitter dataset was essential to ensure data quality and analytical reliability (Multimedia Appendix 2). We streamlined the data by performing initial cleaning and normalization tasks, such as removal of URLs, hashtags, and user mentions, correction of typographical errors, and standardizing text to a uniform lower case, and then further refined our dataset by eliminating special characters, tokenizing text, and handling abbreviations. These comprehensive steps concentrated our analysis on 205,082 tweets, priming them for further analysis. In the sentiment analysis phase, we used the Python package *TextBlob* to perform sentiment detection on the subset of 205,082 tweets identified postinitial preprocessing.

This analysis classified the emotional tone of each tweet into positive, neutral, or negative categories. Subsequently, we refined our dataset by excluding tweets that exhibited positive and neutral sentiments. We drew from the work of Reyes-Menendez et al [64], which used n-grams analysis to reveal a dual framing of the #MeToo movement—one destructive, associated with terms such as cowards, rape, and scandals, and another constructive, linked to words such as educate, leader, and rights.

For our study of justification narratives, we focused on the former, harnessing sentiment analysis and stance detection techniques. This allowed us to refine our dataset to specifically include those tweets that aligned with destructive negative discourse, providing a pertinent sample for our analysis of justification cases. This filtration process resulted in a concentrated sample of 60,488 tweets, each expressing a negatively worded sentiment, which provided a more precise basis for examining the adverse reactions associated with the #MeToo movement. After filtering our dataset to include only the negatively worded tweets, we used the *NLTK* Python package for stance detection to categorize each tweet’s attitude toward the #MeToo movement as supportive, neutral (or factual), or opposing. We then removed tweets with supportive and neutral stances, resulting in a final dataset of 18,416 tweets that explicitly opposed or criticized the movement by justifying

XSL•FO
RenderX

existing social norms, thereby focusing our analysis on expressions of dissent.

We implemented a 2-pronged content analysis strategy to examine the semantic structures within our tweet dataset. First, we used latent Dirichlet allocation topic modeling, which identified and extracted 22 distinct topics representing underlying themes in the discourse. Upon further analysis, we found that 10 of these topics were not directly aligned with any justification narratives, while the remaining 12 topics were systematically categorized into specific justification categories.

Each justification type exhibited distinct lexical and semantic signatures. From the latent Dirichlet allocation topic modeling results, we observed that the topics associated with each system justification type contained unique keyword structures indicative of the underlying sentiment. For instance, "backlash system justification" topics included words such as "movement," "vicious," and "sex," hinting at an undercurrent of resistance to the #MeToo narrative. "Victim blaming justifications" were laced with terms such as "rape," "society," and "refuse," reflecting accusatory and deflective tones. Meanwhile, "institutional justification" topics incorporated institutional and procedural terms such as "resign," "misconduct," and "allegation," suggesting a focus on systemic and structural dimensions. Lastly, "gender system justification" topics distinctly featured "woman" and "man" in their discourse, emphasizing the gendered aspects of the conversation.

These keyword patterns are illustrative of the diverse justifications and sentiments present within the #MeToo-related discourse captured in our dataset. In the second phase, we complemented the computational results with human-coded annotation to refine the justification categories within our tweet sample. Specifically, we recruited 2 trained annotators to collaboratively analyze a representative set of 250 tweets, with the aim of establishing a shared conceptual framework for the various categories of justifications. This process was guided by detailed annotation protocols to ensure conceptual clarity, coding consistency, and interrater reliability.

The interrater reliability, measured using Cohen κ, was found to be 0.91, reflecting a high degree of agreement between the annotators and confirming the robustness of the annotation schema. With this validated framework, we proceeded to annotate a larger subset of 18,416 tweets. We used the insights gained from the initial 250 tweets to train a supervised machine learning model, thereby scaling our classification effort across the entire dataset. This model was designed to categorize tweets into 1 of 4 that justification types or into a fifth category indicating the absence of a discernible justification.

This process allowed us to accurately distribute our sample into proportionate classifications of each justification type, thus enabling a comprehensive and quantifiable analysis of the dominant narratives in the relevant tweets. Discrepancies between annotators were resolved through discussion and iterative refinement of the coding guidelines, resulting in strong interrater agreement and improved clarity in classification criteria. The precision of our classification framework established a robust foundation for subsequent analyses and interpretations of the data. Across the 18,416 negatively valenced tweets, institutional system justification was the most prevalent (7366/18,416, 40%), followed by gender system justification (6814/18,416, 37%), victim-blaming justification (2210/18,416, 12%), and backlash justification (2026/18,416, 11%).

Study 2: Qualitative Thematic Analysis

Study Design and Recruitment

Study 2 builds upon the findings of study 1 by incorporating perspectives from academic experts in the field. We conducted 20 in-depth interviews with scholars and researchers specializing in gender, media, and film studies—particularly those focused on Hindi cinema and the Indian entertainment context. The objective was to understand how #MeTooIndia influenced discourse and practices within the IEI. In-depth interviews are a well-established qualitative method for obtaining detailed insights into participants' beliefs, experiences, and interpretations of a given phenomenon [65-67]. We targeted academics and scholars because of their capacity to provide theoretically informed and relatively impartial reflections on the movement—an important complement to journalistic and activist perspectives that have dominated public narratives.

Using purposive sampling [68], the current study recruited professors, research scholars, and doctoral candidates affiliated with universities and research institutions across the United States, India, the United Kingdom, and the Middle East who specialized in media and film studies, with a focus on Indian cinema. Potential participants were identified through prior publications and conference presentations in feminist media studies, the IEI, social-media activism, and cinema or film studies. We emailed 80 eligible academics, describing this study's objectives and inviting them to participate in an interview. Of these, 29 responded, and appointments were successfully scheduled with 20 (36%) participants. The remaining 9 (11%) participants did not provide availability. Ultimately, 20 (25%) participants of the invited pool completed this study and formed the final analytic sample.

Recruitment was performed through personalized email invitations sent directly to eligible participants, describing this study's objectives, the voluntary nature of participation, and procedures for ensuring confidentiality. Those who agreed to participate provided written informed consent electronically before interviews. Data were collected between January and May 2021, a period shaped by the COVID-19 pandemic, which redefined research logistics and accelerated the adoption of virtual data-collection tools. We conducted 13 synchronous interviews via Zoom and 7 asynchronous interviews through email correspondence, allowing for flexibility in scheduling, geographic reach, and iterative follow-up questions when deeper elaboration was required. This mixed interview format provided both immediacy and reflection, aligning with evolving qualitative research practices during pandemic conditions [69-71].

Analysis

All interviews were transcribed verbatim and imported into NVivo 12 (QSR International) to facilitate systematic coding and data management. Analysis followed an inductive thematic

XSL•FO
RenderX

approach using the constant comparison method [72]. Each transcript was reviewed line by line to identify recurring ideas, patterns, and contradictions both within and across cases. In the first coding cycle, each author independently performed open coding to generate preliminary labels grounded in the participants' language. After multiple reviews of the transcripts, codes such as victim blaming, system, and institutional justification began to emerge. The second coding cycle involved axial coding, where similar concepts were grouped and relationships between categories were explored [73].

To enhance methodological rigor, intercoder reliability was assessed by double-coding a subset of transcripts and comparing agreement rates. Intercoder reliability is crucial as it showcases the methodological approach of achieving consensus in coding among the research team [74]. Discrepancies were resolved through discussion until consensus was reached, and a finalized codebook was developed. Once coding consistency was achieved, all transcripts were recoded using the agreed codebook, and thematic clustering was conducted within NVivo to synthesize findings. Analytical memos were maintained throughout to ensure transparency of interpretation and reflexivity among the research team.

Results

Study 1: Quantitative Study

Gender system justification emerged as the most prevalent category, representing 37% of the dataset and reflecting discourse that reinforces established gender hierarchies and traditional gender norms. Institutional system justification accounted for 40% (7366/18,416), characterized by narratives that defend organizational practices or minimize institutional accountability in cases of harassment and misconduct. Smaller yet meaningful proportions included backlash system justification (2026/18,416, 11%) and victim-blaming system justification (2210/18,416, 12%), illustrating persistent efforts to discredit the #MeToo movement or to shift responsibility onto victims. Together, these findings provide a quantitative overview of how online conversations reproduce societal power structures that sustain gender-based inequities—factors with significant implications for public health, safety, and psychosocial well-being.

The distribution captures the relative prevalence of each justification type, offering a data-driven perspective on the broader discourse surrounding gender justice and institutional accountability. Thus, study 1 offers a global quantitative map of justification narratives, while study 2 qualitatively examines these dynamics within #MeTooIndia in the IEI, illustrating how cultural norms and institutional structures shape experiences, responses, and barriers to gender equity and collective well-being.

Study 2: Qualitative Data Collection

Overview

The final thematic analysis yielded 3 dominant themes (eg, gender system justification, victim blaming, and institutional system justification) reflecting the influence of #MeToo within the IEI, paralleling patterns identified in the quantitative Twitter analysis.

Gender System Justification

Despite #MeToo's global aim to promote feminist awareness and counter sexual harassment, most participants observed little progress within the IEI, noting that existing gender hierarchies largely remain intact. This is an important concern for women's digital and societal well-being, including their mental health and everyday safety, as entrenched gender norms shape vulnerability to harassment and constrain pathways to speak up or seek support. According to Daniel, a media professor from the Pacific West, United States:

> *I think the #MeToo movement added some urgency to an already-ongoing set of changes, which I would characterize as an evolution rather than a revolution. Yes, there has been more diversity in representations of women, more women characters who exert greater autonomy and mastery, but there is still a tendency to show traditional kinds of stereotypical femininity as worthy of reward.*

While Daniel highlighted the persistence of gendered portrayals, Smita, a professor of cinema and photography of Indian American descent from the Midwest, elaborated on the deeper structural inequities within the industry:

> *It has brought attention to sexual exploitation within the industry, made public what insiders knew—but I would not go so far as to say, it has revolutionized anything. There is a great deal of power inequality in the industry; a lot of insecurity produced by having to stay on the right side of those in power (monopoly of a few big names); networks are brutal. Without any of these changes, things will not change much.*

Several notable figures in the IEI attributed the ongoing challenges not only to the industry but also to the broader social system. For instance, Bollywood producer Vinta Nanda expressed that this issue is "not a Bollywood habit but a very Indian habit" [75]. Echoing this sentiment, Latika, a doctoral candidate pursuing her PhD in the United Kingdom, blamed the current political party in power, stating that:

> *#MeToo hasn't been able to revolutionize Bollywood, not to the extent needed, but then India is currently under an unprecedented repressive and gender-violent religious fundamentalist regime, which acerbates the problem.*

Irrfan, a professor of media and communications from the Midwest United States, offered a structural critique of #MeToo:

> *Social movements that are primarily organized by online campaigning often capture public attention quickly. Online users could effectively bring light to social injustice and accumulate public support in a short period. However, such organic and leaderless movements often do not have sufficient organizational capabilities to challenge the existing gender system and keep pushing for changes that could potentially challenge our society's power-relations.*

XSL•FO
RenderX

Irrfan highlighted a central limitation of #MeToo as an "organic" and "leaderless" movement, which has struggled to confront the gendered power dynamics within India's sociostructural context. Scholars argue that the movement has largely amplified the voices of a privileged minority while neglecting those of marginalized and precariously employed women [76]. Others suggest that online social movements promote broader engagement than traditional offline activism, as digital platforms enable individualized participation and storytelling in place of hierarchical leader-driven conventional movements [77]. Additionally, the IEI is not governed by formal workplace policies or oversight mechanisms comparable to other sectors, leaving issues of women's safety and protection as weakly regulated and contributing to increased mental health concerns. Overall, the expert testimonies argue for treating GBV in the IEI as occupational and digital health concerns that require structural interventions: standardized industry policies, independent redress mechanisms, survivor-centered mental health services, and safer digital reporting infrastructures. They also highlight the need for coordinated platform-industry-state action to ensure that the visibility generated by movements such as #MeToo leads to durable protections for women victims, rather than remaining confined to short-lived, elite-driven online attention.

Victim Blaming System Justification

Although #MeToo gained visibility in India, women victims who disclosed experiences of harassment frequently encountered backlash and blame from influential figures seeking to silence their voices. Similar to the #SurvivorPrivilege hashtag movement in 2014, Indian media narratives during #MeToo reinforced victim-blaming discourse, depicting survivors as manipulative or vindictive actors intent on falsely accusing men [9]. A widely publicized example involved actress Tanushree Dutta, who relocated to the United States after facing intense public hostility and media scrutiny following her #MeToo disclosure. Experts noted that these digital reactions operate as forms of secondary victimization, generating emotional strain and shaping whether women feel safe seeking support or speaking publicly. When asked about the movement's effectiveness in confronting India's long-standing culture of victim blaming, Daniel commented:

> *As long as victim blaming and victim shaming still get treated seriously in the press (and the courts), my answer to that question is, it has not gone far enough. Those who have been abused and exploited need to continue to speak up; witnesses need to be spoken up; bystanders need to be ready to believe uncomfortable things about people they know. It is worth noting that this need not be malicious to be problematic—some men who are not themselves abusers are likely trying to make sense of the idea that men with whom they have worked, their friends and role models, potentially, have engaged in immoral behavior. They try to resolve that uncomfortable inconsistency between what they think they know and what they're hearing by rejecting the new information.*

Daniel emphasized that countering victim-blaming requires empowering women to share their stories openly and fostering conversations that challenge victim shaming. He highlighted the need to expose offenders, create safe spaces for women to speak about their experiences without fear, and encourage bystanders to confront the possibility that individuals within their professional networks could be perpetrators. These goals also extend to digital health and safety, given that survivors' willingness to disclose often depends on how platforms handle backlash, disbelief, and reputational risk.

In contrast, Sarita offered a more nuanced perspective, acknowledging both the limitations and potential of the #MeToo movement in challenging victim-blaming norms:

> *One of these tactics in Bollywood—a hotbed of the neoliberal beauty-fashion complex—has been to discredit these women on the basis of beauty norms. It becomes effective in a world mired in the logic of fashion as social capital. This form of slut-shaming, victim-blaming misogyny is effective in Bollywood to some extent. Yet what was new about #MeToo was that unlike earlier, (for instance, in the Shiney Ahuja case, who was arrested in 2009 for raping his then 19-y-old domestic help), there was a spotlight on the narrative of the victim. #MeToo Bollywood became her story. The narrative emphasis placed more screen time on the victim/survivor rather than the harasser. This is new, it helped that the women were also not dehumanized: it helped that we knew them through film roles earlier.*

Sarita's reflection highlighted an important shift in how survivor narratives were framed—away from silencing and toward recognition and visibility. This shift, she believed, has the potential to reduce stigma, support mental well-being, and strengthen collective efficacy among women by validating personal experiences and encouraging shared storytelling. Changing collective narratives around women's stigmatized experiences has been shown to foster online connection, mutual validation, and collective identity formation, all of which can help challenge misogynistic discourse and improve mental health [78]. Smita, another participant, noted that this collective storytelling offered a new form of solidarity:

> *By naming the perpetrators, the campaign made formal what was a circle of solidarity amongst women—what was previously whispered was now made public so women could be warned. To me this was the most useful aspect of the campaign. Now, this is not the same as going to court, naming the perpetrator, giving evidence, a process that takes its toll in terms of years and other costs. So, it is possible that #MeToo may end up diluting the victim's voice, the charge is that anyone can say whatever, etc. It is possible, but I think for the moment it has succeeded in sounding a warning note to potential perps. However, in the long run, as long as the severe imbalance remains in the industry, this campaign too may lose its credibility and power.*

XSL•FO
RenderX

Thus, despite the limitations, participants viewed #MeToo as a powerful, if temporary, mechanism for achieving symbolic and restorative justice through visibility, solidarity, and public accountability. Although some scholars argue that these rapid forms of justice may be transient and fail to address deeper structural determinants of sexual misconduct [79,80], the movement nonetheless represents an essential step toward disrupting the silence around GBV. Taken together, these insights position victim-blaming and digital backlash as both digital and public health concerns, shaping survivors' emotional well-being, safety, and ability to access redress within and beyond digital spaces.

Institutional System Justification

A striking contradiction revealed by the #MeToo movement in the IEI was the exposure of several acclaimed filmmakers, directors, and actors who were simultaneously celebrated for women-centric films and public advocacy for gender equality, yet accused of sexual harassment by colleagues in the industry. This pattern echoed high-profile cases in the United States involving R Kelly, Harvey Weinstein, and Bill Cosby, drawing similar attention to India among celebrity figures such as Subhash Ghai and Vikas Bahl (filmmakers), Alok Nath (actor), and Varun Grover (writer-comedian). While the allegations sent shockwaves through the public, many cases were quietly dismissed or settled, raising questions about institutional accountability and the fragility of justice when social status and reputation are at stake. Raya, a media professor in the United States, reflected on the broader cultural logic that enables such contradictions:

> *This does not surprise me at all. There is a self-declared distance Bollywood or Indian cinema directors have often maintained between "art" and "life." Representations of the female body and female sexuality have traditionally always been the prerogative of men in most cultures. Along with that, there has never been any mandate or prohibition against commodification or exploitation of the same body or femininity. In India, where there's also such a strong social binary between the public and the private woman—the actress and the wife—in terms of sexual politics, such an assumption of entitlement regarding "public women" or actresses must come quite easily to men in power, whether they are artistic, creative, or not. There may also be that "great men" syndrome whereby many of them believe that as artistes they are not subject to the norms of what is called "bourgeois" morality by the avant-garde in India.*

Raya's observation underscores how creative industries often justify misconduct through artistic exceptionalism—an attitude that normalizes exploitation as an accepted by-product of artistic freedom. Such institutional rationalizations mirror system-justifying processes that protect reputational hierarchies and maintain power asymmetries. Daniel and several other participants discussed how this contradiction extends to the production of women-centric films. As Daniel noted:

> *The film industry is, at its core, an industry. Directors and producers are artists, certainly, but their eye—and the eye of their investors and funders—is on the bottom line. If they believe there is a market for a female-centric movie, they will make it even if it is not entirely consonant with their own personal beliefs and attitudes. It's also important to remember that acting in ways inconsistent with sincerely held beliefs is hardly unusual.*

Collectively, participants highlighted the gap between gender-progressive representations on screen and persistent inequities off-screen. They pointed to entrenched gender ideologies, victim-blaming tendencies, and the commodification of women's bodies as markers of an institutional culture that simultaneously markets empowerment and reproduces exploitation. Several experts observed that the performative nature of the industry—where acting and pretense are core to professional success—creates conditions in which abusers maintain a convincing public façade, reinforcing impunity.

These narratives show how social structures that normalize gendered entitlement and discourage reporting can generate psychological strain, career marginalization, and reduced well-being among women. Addressing institutional system justification in such creative settings requires more than media accountability; it demands organizational and occupational frameworks that prioritize equity, enforce safeguards against harassment, and cultivate supportive environments for women survivors. From a public health lens, the IEI's gender inequities constitute a preventable occupational health crisis requiring multilevel interventions. For instance, digital health technologies offer scalable solutions for reporting, prevention, mental health support, and accountability—bridging the gap between performative empowerment and substantive safety. By combining public health frameworks with digital innovation, the industry can move beyond performative allyship toward measurable, sustainable gender justice that protects both mental well-being and professional equity.

Discussion

Social Media System Justification on Twitter

This study examined how social media discourse surrounding the #MeToo movement reflects and reproduces system-justifying narratives related to GBV. Consistent with this study's objectives, the analysis identified 4 dominant forms of system justification—gender, institutional, victim-blaming, and backlash—across a large corpus of Twitter data and expert interviews. Study 1 demonstrated that online conversations frequently defended existing gender norms and institutional structures, with gender and institutional justification emerging as the most prevalent categories. Such defense mechanisms reflect social conditioning, resistance to change, and selective exposure to congruent beliefs online [81,82].

Study 2 further contextualized these patterns within the IEI, revealing that even as #MeTooIndia expanded public dialogue on sexual harassment—entrenched hierarchies, reputational risks, and institutional inertia continued to constrain meaningful reform. Together, these findings indicate that while the #MeToo

XSL•FO
RenderX

movement fostered visibility and collective awareness, digital spaces also became arenas where system-justifying discourse reinforced social inequities and psychological strain, consistent with prior work showing that exposure to gender movements can amplify defensive reactions [48,50].

From a public health standpoint, these polarization patterns have psychosocial implications. Online gender-justifying narratives perpetuate stigma, discourage disclosure of sexual harassment, and increase survivors' emotional distress [10,19]. The normalization of inequitable discourse can further sustain hostile digital climates that contribute to secondary victimization—a public health concern linked to trauma recurrence and barriers to help-seeking [22,38]. The high prevalence of institutional justification observed in the dataset also reflects persistent trust in existing social and institutional structures, even when those systems perpetuate inequality [20,21,29]. System justification fulfills a psychological need to defend and justify existing social arrangements, even if those are disadvantageous to society, as it is often perceived as psychologically comforting to individuals [40].

Institutional justification in feminist movements often works to preserve the status quo, especially when current arrangements benefit those in power [36,49]. These dynamics help explain why movements such as #MeToo can expose systemic issues and yet still fall short of durable institutional change. Finally, victim blaming and backlash justifications appeared less frequently but remained consequential. While victim-blaming and backlash in feminist movements have been widely examined [21,29,32,83,84], our findings show a comparatively lower prevalence of such narratives in the #MeToo discourse on social media. The wide circulation of survivor stories helped render overt victim-blaming less acceptable, reshaping public perceptions of sexual harassment and assault [5,33]. However, amplifying women's voices through #MeToo also led some men to perceive themselves as unfairly victimized, fueling backlash that reasserted traditional gender hierarchies [20,85].

The "#MeTooLate Effect," where delayed disclosures prompted disbelief or blame, underscores persistent biases that privilege certain forms of reporting [21,54]. Perceived credibility often follows a hierarchy shaped by social expectations of an "ideal" victim (such as immediate reporting or corroborating witnesses), thereby marginalizing survivors who do not conform to these norms [11,86]. Overall, these findings suggest that while social media amplifies awareness of GBV, it also functions as a mirror of societal polarization. This simultaneously promotes empowerment and entrenches opposition. Thus, digital backlash can increase psychological distress, inhibit help-seeking, and perpetuate stigma surrounding GBV [3,11,19]. Addressing these online narratives through trauma-informed communication, digital moderation, and survivor-centered discourse strategies is essential for promoting mental well-being and social equity.

Consequences of #MeTooIndia Movement

The #MeToo movement in India catalyzed multiple discourses, revealing both systemic awareness and persistent structural inequities within the IEI [87,88]. Insights from study 2, based on in-depth interviews with scholarly experts, highlighted the movement's mixed legacy—while it expanded public dialogue around sexual harassment, entrenched gender hierarchies and institutional practices remained largely unchanged [15,89]. This disconnect between awareness and reform mirrors broader societal resistance to redistributing power within patriarchal systems [87,90].

Participants described continuing patterns of victim blaming toward women who disclosed harassment, often amplified by media sensationalism. Such reactions perpetuate psychological distress, reputational harm, and social isolation, underscoring how hostile cultural responses can evolve into public health concerns through cumulative mental health and occupational stress [91,92]. Experts noted that survivors in high-visibility professions face compounded stigma and scrutiny, exacerbating trauma and discouraging disclosure. This outcome is consistent with global research on secondary victimization [22,38].

Interviewees also identified a persistent paradox in the IEI's engagement with #MeToo, where increased public messaging on gender equity coexists with limited organizational accountability. This tension illustrates a symbolic commitment to feminist ideals rather than substantive institutional reform. Several experts described the movement's momentum as short-lived "clicktivism" or "slacktivism," characterized by digital enthusiasm but little lasting sociolegal impact [93-95]. Participants further noted that sustained solidarity was essential for continued engagement and resistance to social silencing, consistent with theories of collective action [96], though public disclosure frequently came with substantial psychological and professional costs.

Celebrities who spoke out faced coordinated online harassment, smear campaigns by industry elites, and defamation lawsuits aimed at intimidating or silencing them [16,89]. Such backlash reinforces patriarchal power structures and contributes to chronic distress, anxiety, and social withdrawal among women in high-visibility professions [20,92]. The hostile digital environment surrounding #MeTooIndia thus serves as an infoveillance signal of ongoing GBV and secondary victimization, illustrating how digital spaces can perpetuate trauma even while exposing it.

Despite these tensions, participants viewed the movement's visibility as crucial for advancing public dialogue on sexual violence and strengthening institutional accountability. Achieving digital health and safety requires trauma-informed legal frameworks and industry mechanisms that protect survivors and translate online advocacy into sustainable, health-promoting reform. These findings should be interpreted alongside the limitations described in the next section.

Limitations

Several key limitations influence the strength of the conclusions and recommendations drawn from this study. First, study 1 analyzed tweets collected over 1 month between November and December 2017, shortly after the initial #MeToo surge. This captures the early phase of the movement but may not represent later shifts in framing. Second, only English-language tweets were included, excluding parallel conversations in regional languages. Third, study 2 relied on 20 in-depth interviews with scholars and researchers specializing in gender, media, and film

XSL•FO
RenderX

studies, constraining the analysis to the IEI context. While these experts provided valuable theoretical and contextual nuance, the perspectives of survivors (who experience digital hostility and secondary victimization directly) may differ in important ways. The absence of survivor interviews represents an additional limitation. Additionally, cultural norms, linguistic variations, and platform-specific affordances may have shaped the tone and visibility of #MeToo discourse, limiting the generalizability of findings across social and regional contexts.

Conclusion

This study examined how social media discourse within the #MeToo movement reflects system-justifying narratives that sustain rather than dismantle existing hierarchies [5]. Contrary to prior studies emphasizing victim-blaming and overt backlash, findings from both study 1 and study 2 reveal strong evidence of gender and institutional system justification. These patterns indicate that social media often reinforces traditional gender norms and institutional authority, maintaining the status quo even within digital spaces that appear progressive. Although #MeToo expanded visibility and awareness, its ability to drive enduring structural change within sectors such as the IEI remains limited. Digital platforms, therefore, operate as contested arenas, where technologies that enable collective voice can simultaneously reproduce digital GBV.

Our findings extend scholarship that centers primarily on backlash and victim-blaming by demonstrating how broader system-justifying frames appear in social media discourses. The findings carry 3 major implications. First, for digital platforms, they highlight the need for proactive governance and content moderation policies that curb system-justifying retaliatory discourse. This requires integrating trauma-informed design principles to reduce secondary victimization. Second, for public health and policy stakeholders, the results point to the importance of recognizing digital hostility as a psychosocial health risk that warrants coordinated mental health and legal interventions. Third, for researchers and feminist advocates, this work offers an analytic lens for situating digital activism within broader institutional and psychological systems, while identifying points where transformative change may stall.

Acknowledgments

HP and ST are first authors who contributed equally. During the preparation of this work, the authors used generative artificial intelligence tools (Grammarly [Superhuman Platform], Claude [Anthropic], ChatGPT [OpenAI], and Microsoft Copilot [Microsoft Corp]) solely for copyediting, language refinement, and proofreading purposes. The content was subsequently reviewed, verified, and edited by the authors, who take full responsibility for the integrity and accuracy of this final paper. This use complies with the JMIR Publications Editorial Policy on the use of artificial intelligence tools.

Funding

The authors declared no financial support was received for this work.

Data Availability

The dataset analyzed in this study contains sensitive textual data related to gender-based sexual violence collected from public social media discussions. Sharing the raw tweet text could compromise user privacy, as individual posts may be reidentifiable even after anonymization. Furthermore, redistribution of tweet content is restricted under Twitter's Terms and Conditions [60]. To ensure transparency and reproducibility, a synthetic and anonymized sample dataset, along with the analysis code framework and software requirements, is included in the supplementary materials. Nonsensitive components of the code and illustrative samples are available upon reasonable request from the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semistructured interview guide for study 2 (#MeTooIndia IEI expert interviews). IEI: Indian Entertainment Industry.
[DOCX File, 21 KB - jmir_v28i1e75533_app1.docx]

Multimedia Appendix 2
Data and code files description: data_analysis_template.txt contains the Python code used for the computational framework, including preprocessing, sentiment analysis, stance detection, and topic modeling. ReadMe.md provides details about the dataset, file structure, and disclosure on data availability and usage. requirements.txt lists all Python package dependencies used in the analysis. sample_data_mt.txt includes a sample dataset excerpt.
[ZIP File, 2 KB - jmir_v28i1e75533_app2.zip]

Checklist 1
COREQ checklist.

XSL•FO
RenderX

[PDF File, 497 KB - jmir_v28i1e75533_app3.pdf]

References

1. Field-Springer K, Draut H, Babrow F, Sandman M. (Re)claiming stories in the #MeToo movement: righting epistemic wrongs of physical, mental, and emotional harms of sexual violence. Health Commun 2022 Jul;37(8):982-991. [doi: 10.1080/10410236.2021.1880052] [Medline: 33541139]
2. Castle JJ, Jenkins S, Ortbals CD, Poloni-Staudinger L, Strachan JC. The effect of the #MeToo movement on political engagement and ambition in 2018. Polit Res Q 2020 Dec;73(4):926-941. [doi: 10.1177/1065912920924824]
3. Dilip D, Bates L. Sexual violence and mental health among college students in the era of #MeToo. Violence Gend 2021 Mar 1;8(1):8-13. [doi: 10.1089/vio.2020.0033]
4. Guan S, Hui V, Stiglic G, Constantino RE, Lee YJ, Wong AKC. Classifying the information needs of survivors of domestic violence in online health communities using large language models: prediction model development and evaluation study. J Med Internet Res 2025 May 12;27:e65397. [doi: 10.2196/65397] [Medline: 40354642]
5. Mayer ED, Peterson R, Kim E, et al. What survivors of sexual violence want when disclosing their experiences in person or online: qualitative interview analysis. J Med Internet Res 2025;27:e73497. [doi: 10.2196/73497]
6. Kaufman MR, Wright K, Shin R, et al. The power of social media activism in the #YesAllWomen movement. Humanit Soc Sci Commun 2025;12(1):1-11. [doi: 10.1057/s41599-025-05647-5]
7. Glassman M, Kuznetcova I, Peri J, Kim Y. Cohesion, collaboration and the struggle of creating online learning communities: Development and validation of an online collective efficacy scale. Comput Educ Open 2021 Dec;2:100031. [doi: 10.1016/j.caeo.2021.100031]
8. Suk J, Abhishek A, Zhang Y, et al. #MeToo, networked acknowledgment, and connective action: how “empowerment through empathy” launched a social movement. Soc Sci Comput Rev 2021 Apr;39(2):276-294. [doi: 10.1177/0894439319864882]
9. Jackson S, Bailey M, Welles BF. Women tweet on violence: from #YesAllWomen to #MeToo. Ada: J Gender, New Media, Technol 2019 Jan 1 [FREE Full text]
10. O’Neil A, Sojo V, Fileborn B, Scovelle AJ, Milner A. The #MeToo movement: an opportunity in public health? Lancet 2018 Jun;391(10140):2587-2589. [doi: 10.1016/S0140-6736(18)30991-7]
11. PettyJohn ME, Anderson G, McCauley HL. Exploring survivor experiences on social media in the #MeToo era: clinical recommendations for addressing impacts on mental health and relationships. J Interpers Violence 2022 Nov;37(21-22):NP20677-NP20700. [doi: 10.1177/08862605211055079] [Medline: 34861795]
12. Bethel C. #MeToo: the perfect storm needed to change attitudes toward sexual harassment and violence. HPHR 2018;16. [doi: 10.54111/0001/P1]
13. Rivera-Rodriguez A, Larsen G, Dasgupta N. Changing public opinion about gender activates group threat and opposition to feminist social movements among men. Group Processes Intergroup Relat 2022 Apr;25(3):811-829. [doi: 10.1177/13684302211048885]
14. Stephan WG, Ybarra O, Rios K. Intergroup threat theory. In: Handbook of Prejudice, Stereotyping, and Discrimination Psychology Press: Psychology Press; 2015:255-278. [doi: 10.4324/9781841697772]
15. Nanditha N. Exclusion in #MeToo India: rethinking inclusivity and intersectionality in Indian digital feminist movements. Feminist Media Stud 2022 Oct 3;22(7):1673-1694. [doi: 10.1080/14680777.2021.1913432]
16. No #MeToo here! Saroj Khan says casting couch provides livelihood, is a girl’s choice. The Economic Times. 2018 Apr 24. URL: https://economictimes.indiatimes.com/magazines/panache/no-metoo-here-saroj-khan-says-casting-couch-provides-livelihood-is-a-girls-choice/articleshow/63891824.cms?from=mdr [accessed 2026-03-20]
17. Hayes RM, Lorenz K, Bell KA. Victim blaming others: rape myth acceptance and the just world belief. Fem Criminol 2013 Jul;8(3):202-220. [doi: 10.1177/1557085113484788]
18. Bogen KW, Bleiweiss KK, Leach NR, Orchowski LM. #MeToo: disclosure and response to sexual victimization on Twitter. J Interpers Violence 2021 Sep;36(17-18):8257-8288. [doi: 10.1177/0886260519851211] [Medline: 31117851]
19. Bhuptani PH, López G, Peterson R, Orchowski LM. Online social reactions to disclosure of sexual victimization via #MeToo and symptoms of post-traumatic stress disorder. J Interpers Violence 2023 Oct;38(19-20):10900-10919. [doi: 10.1177/08862605231176792] [Medline: 37272010]
20. Moitra A, Ahmed SI, Chandra P. Parsing the “me” in #MeToo: sexual harassment, social media, and justice infrastructures. Proc ACM Hum-Comput Interact 2021;5:1-34. [doi: 10.1145/3449185]
21. Nutbeam M, Mereish EH. Negative attitudes and beliefs toward the #MeToo movement on Twitter. J Interpers Violence 2022 Aug;37(15-16):NP13018-NP13044. [doi: 10.1177/08862605211001470] [Medline: 33752498]
22. Gorissen M, van den Berg CJW, Ruiter S, Bijleveld C. Sharing unwanted sexual experiences online: a cross-platform analysis of disclosures before, during and after the #MeToo movement. Comput Hum Behav 2023 Jul;144:107724. [doi: 10.1016/j.chb.2023.107724]
23. Pain P. “It took me quite a long time to develop a voice”: examining feminist digital activism in the Indian #MeToo movement. New Media Soc 2021 Nov;23(11):3139-3155. [doi: 10.1177/1461444820944846]

XSL•FO
RenderX

24. Sambaraju R. “I would have taken this to my grave, like most women”: reporting sexual harassment during the #MeToo movement in India. J Soc Issues 2020 Sep;76(3):603-631. [doi: 10.1111/josi.12391]
25. Kende A, Nyúl B, Lantos NA, et al. A needs-based support for #MeToo: power and morality needs shape women’s and men’s support of the campaign. Front Psychol 2020;11:593. [doi: 10.3389/fpsyg.2020.00593] [Medline: 32296377]
26. Sakar S. Experiences of backlash against women in the form of technology-facilitated sexual violence in India. QUT Cent Justice 2023 Dec;44:1-4. [doi: 10.5204/book.eprints.245980]
27. McCosker A. Engaging mental health online: insights from beyondblue’s forum influencers. New Media Soc 2018 Dec;20(12):4748-4764. [doi: 10.1177/1461444818784303]
28. Liu L, Woo BKP. Twitter as a mental health support system for students and professionals in the medical field. JMIR Med Educ 2021 Jan 19;7(1):e17598. [doi: 10.2196/17598] [Medline: 33464210]
29. Xue J, Zhang B, Zhang Q, et al. Using Twitter-based data for sexual violence research: scoping review. J Med Internet Res 2023 May 15;25:e46084. [doi: 10.2196/46084] [Medline: 37184899]
30. Gangoli G. The #MeToo movement in India: emotions and (in)justice in feminist responses. Fem Leg Stud 2024 Jul;32(2):213-230. [doi: 10.1007/s10691-023-09540-x]
31. Marian M, Pérez RL, Reed E, et al. Exploring the use of social media for activism by Mexican nongovernmental organizations using posts from the 16 days of activism against gender-based violence campaign: thematic content analysis. JMIR Inf 2025 Apr 17;5:e67368. [doi: 10.2196/67368] [Medline: 40246294]
32. Maier SL. Rape victim advocates’ perceptions of the #MeToo movement: opportunities, challenges, and sustainability. J Interpers Violence 2023 Jan;38(1-2):NP336-NP365. [doi: 10.1177/08862605221081929] [Medline: 35416073]
33. Ortiz RR, Smith AM. A social identity threat perspective on why partisans may engage in greater victim blaming and sexual assault myth acceptance in the #MeToo era. Violence Women 2022 Apr;28(5):1302-1325. [doi: 10.1177/10778012211014554] [Medline: 34086501]
34. Huck JL, Homer M. How men view stigma, shame, and blame in #MeToo. Soc Sci Q 2024;105(5):1515-1528. [doi: 10.1111/ssqu.13428]
35. Modrek S, Chakalov B. The #MeToo movement in the United States: text analysis of early Twitter conversations. J Med Internet Res 2019 Sep 3;21(9):e13837. [doi: 10.2196/13837] [Medline: 31482849]
36. Silva M, Anaba U, Jani Tulsani N, Sripad P, Walker J, Aisiri A. Gender-based violence narratives in internet-based conversations in Nigeria: social listening study. J Med Internet Res 2023 Sep 15;25:e46814. [doi: 10.2196/46814] [Medline: 37713260]
37. Langer N, Khurana S, Vaast E. Activating the sisterhood: a structural and temporal analysis of sustained connective action in #metooindia. Presented at: ICIS 2020; Dec 13-16, 2020 URL: https://aisel.aisnet.org/icis2020/blendlocalglobal/blendlocalglobal/7 [accessed 2026-04-02]
38. Jackson MA, Valentine SE, Woodward EN, Pantalone DW. Secondary victimization of sexual minority men following disclosure of sexual assault: “victimizing me all over again…”. Sex Res Soc Policy 2017 Sep;14(3):275-288. [doi: 10.1007/s13178-016-0249-6]
39. Giddens L, Petter S. "They deserved it”: using the just world hypothesis to understand blaming. Presented at: 53rd Hawaii International Conference on System Sciences; Jan 7-10, 2020. [doi: 10.24251/HICSS.2020.724]
40. Jost JT, Banaji MR, Nosek BA. A decade of system justification theory: accumulated evidence of conscious and unconscious bolstering of the status quo. Polit Psychol 2004 Dec;25(6):881-919. [doi: 10.1111/j.1467-9221.2004.00402.x]
41. Card KG, Hepburn K. Social position and economic system justification in Canada: implications for advancing health equity and social justice from an exploratory study of factors shaping economic system justification. Front Public Health 2022;10:902374. [doi: 10.3389/fpubh.2022.902374] [Medline: 36339181]
42. Lima BPB, de Souza LEC, Jost JT. System justification, subjective well-being, and mental health symptoms in members of disadvantaged minority groups. Clin Psychol Rev 2025 Feb;115:102532. [doi: 10.1016/j.cpr.2024.102532] [Medline: 39740353]
43. Jost JT. A quarter century of system justification theory: questions, answers, criticisms, and societal applications. Br J Soc Psychol 2019 Apr;58(2):263-314. [doi: 10.1111/bjso.12297]
44. Jost JT, Banaji MR. The role of stereotyping in system ‐ justification and the production of false consciousness. Br J Soc Psychol 1994 Mar;33(1):1-27. [doi: 10.1111/j.2044-8309.1994.tb01008.x]
45. Bleeker K, Powers RA, Centelles V. “If we can change the culture, I hope we won’t need to say Me Too”: disclosures of sexual victimization as a form of social activism. Am J Crim Just 2023 Oct;48(5):1224-1244. [doi: 10.1007/s12103-022-09695-4]
46. Andalibi N, Haimson OL, Choudhury MD, Forte A. Social support, reciprocity, and anonymity in responses to sexual abuse disclosures on social media. ACM Trans Comput-Hum Interact 2018 Oct 31;25(5):1-35. [doi: 10.1145/3234942]
47. Drewett C, Oxlad M, Augoustinos M. Breaking the silence on sexual harassment and assault: an analysis of #MeToo tweets. Comput Human Behav 2021 Oct;123:106896. [doi: 10.1016/j.chb.2021.106896]
48. Szekeres H, Shuman E, Saguy T. Views of sexual assault following #MeToo: the role of gender and individual differences. Pers Individ Dif 2020 Nov;166:110203. [doi: 10.1016/j.paid.2020.110203]

XSL•FO
RenderX

49. Liaquat U, Jost JT, Balcetis E. System justification motivation as a source of backlash against equality‐promoting policies—and what to do about it. Soc Issues Policy Rev 2023 Jan;17(1):131-154. [doi: 10.1111/sipr.12093]
50. Menegatti M, Mazzuca S, Ciaffoni S, Moscatelli S. Behind the lines of #MeToo: exploring women's and men's intentions to join the movement. Sustainability 2022;14(19):12294. [doi: 10.3390/su141912294]
51. Kachen A, Krishen AS, Petrescu M, Gill RD, Peter PC. #MeToo, #MeThree, #MeFour: Twitter as community building across academic and corporate institutions. Psychol Mark 2021 Mar;38(3):455-469. [doi: 10.1002/mar.21442]
52. Minkina N. Can #MeToo abolish sexual harassment and discrimination in medicine? Lancet 2019 Aug;394(10196):383-384. [doi: 10.1016/S0140-6736(19)31731-3]
53. Petter S, Giddens L, Wake Forest University, University of North Texas. Is it your fault? Framing social media inclusion and exclusion using just world theory. JAIS 2023;24(5):1248-1270. [doi: 10.17705/1jais.00813]
54. Lucarini A, Suitner C, Brown R, Craig MA, Knowles ED, Salvador Casara BG. The #MeTooLate effect: victim blame and trust denial for sexual harassment not immediately reported. Pers Individ Dif 2020 Dec;167:110240. [doi: 10.1016/j.paid.2020.110240]
55. Stewart H. "Why didn't she say something sooner?": Doubt, denial, silencing, and the epistemic harms of the #MeToo movement. South Cent Rev 2019;36(2):68-94. [doi: 10.1353/scr.2019.0014]
56. Hasunuma L, Shin K. #MeToo in Japan and South Korea: #WeToo, #WithYou. In: Me Too Political Science: Routledge; 2020:97-111. [doi: 10.1201/9781003014508-6]
57. Ivankova NV, Creswell JW, Stick SL. Using mixed-methods sequential explanatory design: from theory to practice. Field Methods 2006 Feb;18(1):3-20. [doi: 10.1177/1525822X05282260]
58. Luo H, Zhang L. Scandal, social movement, and change: evidence from #MeToo in Hollywood. Manage Sci 2022 Feb;68(2):1278-1296. [doi: 10.1287/mnsc.2021.3982]
59. Wehrli S, Hartner AM, Boender TS, Arnrich B, Irrgang C. Information pathways and voids in critical German online communities during the COVID-19 vaccination discourse: cross-platform and mixed methods analysis. J Med Internet Res 2025;27:e76309-e76309. [doi: 10.2196/76309]
60. Chen E, Lerman K, Ferrara E. Tracking social media discourse about the COVID-19 pandemic: development of a public coronavirus Twitter data set. JMIR Public Health Surveill 2020;6(2):e19273. [doi: 10.2196/19273]
61. eCFR:: 45 CFR 46.102 -- Definitions for purposes of this policy. eCFR. URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.102 [accessed 2026-04-02]
62. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013;310(20):2191-2194. [doi: 10.1001/jama.2013.281053]
63. Putting women first: ethical and safety recommendations for research on domestic violence against women. : World Health Organization; 2001 URL: https://www.who.int/publications/i/item/WHO-FCH-GWH-01.1 [accessed 2026-03-20]
64. Reyes-Menendez A, Saura JR, Thomas SB. Exploring key indicators of social identity in the #MeToo era: using discourse analysis in UGC. Int J Inf Manage 2020 Oct;54:102129. [doi: 10.1016/j.ijinfomgt.2020.102129]
65. Speziale HS, Streubert HJ, Carpenter DR. Qualitative Research in Nursing: Advancing the Humanistic Imperative: Lippincott Williams & Wilkins; 2011.
66. Thakkar S. Exploitation, harassment and violence: lived experiences of women paid domestic workers in India. J South Asian Dev 2024 Apr;19(1):44-60. [doi: 10.1177/09731741231164872]
67. Thakkar S. Moving towards empowerment: migrant domestic workers in India. Asian J Women Stud 2021 Jul 3;27(3):425-440. [doi: 10.1080/12259276.2021.1957195]
68. Robinson OC. Sampling in interview-based qualitative research: a theoretical and practical guide. Qual Res Psychol 2014 Jan 2;11(1):25-41. [doi: 10.1080/14780887.2013.801543]
69. Cater JK. Skype: a cost-effective method for qualitative research. Rehabil Counselors Educ J 2011;4(2):1-4 [FREE Full text]
70. Hooley T, Marriott J, Wellens J. What Is Online Research?: Using the Internet for Social Science Research, 1st edition: Bloomsbury Academic; 2012. [doi: 10.5040/9781849665544]
71. Janghorban R, Roudsari RL, Taghipour A. Skype interviewing: the new generation of online synchronous interview in qualitative research. Int J Qual Stud Health Well-Being 2014 Jan;9(1):24152. [doi: 10.3402/qhw.v9.24152]
72. Strauss AL. Qualitative Analysis for Social Scientists: Cambridge University Press; 1987.
73. Blair E. A reflexive exploration of two qualitative data coding techniques. J Methods Meas Soc Sci 2015;6(1):14-29. [doi: 10.2458/v6i1.18772]
74. Saldaña J. Goodall's verbal exchange coding: an overview and example. Qual Inquiry 2016;22(1):36-39. [doi: 10.1177/1077800415603395]
75. Has #MeToo changed Bollywood? BBC News. 2019 Dec 27. URL: https://www.bbc.com/news/av/world-asia-india-50893469 [accessed 2026-03-20]
76. Pegu S. MeToo in India: building revolutions from solidarities. Decision 2019 Jun;46(2):151-168. [doi: 10.1007/s40622-019-00212-x]

XSL•FO
RenderX

77. Brünker F, Wischnewski M, Mirbabaie M, Meinert J. The role of social media during social movements–observations from the #MeToo debate on Twitter. Presented at: Hawaii International Conference on System Sciences; Jan 7-10, 2020. [doi: 10.24251/HICSS.2020.288]
78. Gallagher RJ, Stowell E, Parker AG, Welles BF. Reclaiming stigmatized narratives: the networked disclosure landscape of #metoo. In: Proceedings of the ACM on Human-Computer Interaction: Association for Computing Machinery; 2019, Vol. 3:1-30. [doi: 10.1145/3359198]
79. Pipyrou S. MeToo is little more than mob rule//vs//# MeToo is a legitimate form of social justice. HAU: J Ethnographic Theory 2018;8(3):415-419. [doi: 10.1086/701007]
80. Fohring S, Horsfield L. Survivors speak out: the successes and failures of hashtag activism. In: Canning V, Martin G, Tombs S, editors. The Emerald International Handbook of Activist Criminology: Emerald Publishing Limited; 2023:337-350. [doi: 10.1108/978-1-80262-199-020231022]
81. Otterbacher J, Bates J, Clough P. Competent men and warm women: gender stereotypes and backlash in image search results. In: CHI '17: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems: Association for Computing Machinery; 2017:6620-6631. [doi: 10.1145/3025453.3025727]
82. Mou Y, Cui Y, Wang J, Wu Y, Li Z, Wu Y. Perceiving sexual harassment and #metoo social media campaign among Chinese female college students. J Gend Stud 2022 Feb 17;31(2):178-192. [doi: 10.1080/09589236.2021.1884848]
83. Andreasen MB. 'Rapeable' and 'unrapeable' women: the portrayal of sexual violence in internet memes about #MeToo. J Gend Stud 2021 Jan 2;30(1):102-113. [doi: 10.1080/09589236.2020.1833185]
84. Sharoni S. Speaking up in the age of #MeToo and persistent patriarchy or what can we learn from an elevator incident about anti-feminist backlash. Fem Rev 2018 Nov;120(1):143-151. [doi: 10.1057/s41305-018-0127-6]
85. Lisnek JA, Wilkins CL, Wilson ME, Ekstrom PD. Backlash against the #MeToo movement: how women's voice causes men to feel victimized. Group Processes Intergroup Relat 2022 Apr;25(3):682-702. [doi: 10.1177/13684302211035437]
86. Hart CG. Is There an idealized target of sexual harassment in the MeToo era? Soc Probl 2025 Jan 27;72(1):277-293. [doi: 10.1093/socpro/spad016]
87. Mishra S. From #MeToo to #MeTooIndia: news domestication in Indian English language newspapers. J Stud 2020 Apr 3;21(5):659-677. [doi: 10.1080/1461670X.2019.1709882]
88. Mishra S. A Bollywood icon was accused of sexual assault. The response was chilling. Washington Post. 2019 Jan 17. URL: https://www.washingtonpost.com/opinions/2019/01/17/bollywood-icon-was-accused-sexual-assault-response-was-chilling/ [accessed 2026-03-20]
89. Mitra S. Trolled, body-shamed and slut-shamed: the desecration of the contemporary Bollywood female star on social media. Celebrity Stud 2020 Jan 2;11(1):101-115. [doi: 10.1080/19392397.2020.1704388]
90. Ghosh S, Kamath K. News sourcing, gender, and misogyny in #MeToo india. In: Misogyny Across Global Media: Bloomsbury Publishing PLC; 2021:173-190. [doi: 10.5040/9781666996227]
91. Lakkimsetti C. Stripping away at respectability: #MeToo India and the politics of dignity. Feminist Form 2021;33(3):303-317. [doi: 10.1353/ff.2021.0052]
92. Hayes RM. From blurred lines to #MeToo: media impact on rape culture. In: Defining Rape Culture: Gender, Race and the Move Toward International Social Change: Emerald Publishing Limited; 2023:55-76. [doi: 10.1108/978-1-80262-213-320231004]
93. Christensen HS. Political activities on the internet: slacktivism or political participation by other means? First Monday 2011 Feb 2;16(2). [doi: 10.5210/fm.v16i2.3336]
94. Lindgren S. Movement mobilization in the age of hashtag activism: examining the challenge of noise, hate, and disengagement in the #MeToo campaign. Policy Internet 2019 Dec;11(4):418-438. [doi: 10.1002/poi3.212]
95. Tonkinwise C. To save everything, click here: the folly of technological solutionism. J Des Hist 2014 Mar 1;27(1):111-113. [doi: 10.1093/jdh/ept034]
96. Polletta F, Jasper JM. Collective identity and social movements. Annu Rev Sociol 2001 Aug;27(1):283-305. [doi: 10.1146/annurev.soc.27.1.283]

Abbreviations

GBV: gender-based violence
IEI: Indian Entertainment Industry

XSL•FO
RenderX

Edited by A Stone, N Cahill; submitted 05.Apr.2025; peer-reviewed by C Matt Graham, L Couto; revised version received 24.Dec.2025; accepted 29.Dec.2025; published 09.Apr.2026.

Please cite as:
Parekh H, Thakkar S, Bhatt P, Akello P
Digital Discourse, Secondary Victimization, and Psychological Harm: Mixed-Methods Analysis of System Justification in the #MeToo Movement
J Med Internet Res 2026;28:e75533
URL: https://www.jmir.org/2026/1/e75533
doi:10.2196/75533

© Harsh Parekh, Shriya Thakkar, Paras Bhatt, Patricia Akello. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Analysis of Multilevel Factors Mobilizing the Spectrum of Interorganizational Knowledge Sharing for Facilitating Digital Transformation at Scale: Qualitative Study

Hajar Mozaffar[1], PhD; Robin Williams[2], PhD; Kathrin Cresswell[3], PhD

[1]Business School, University of Edinburgh, 29 Buccleuch Place, Edinburgh, United Kingdom

[2]Institute for the Study of Science, Technology and Innovation, University of Edinburgh, Edinburgh, United Kingdom

[3]Usher Institute, University of Edinburgh, Edinburgh, United Kingdom

Corresponding Author:
Hajar Mozaffar, PhD
Business School, University of Edinburgh, 29 Buccleuch Place, Edinburgh, United Kingdom

Abstract

Background: Interorganizational knowledge sharing is vital for scaling digital transformation efforts that span multiple organizations and system-wide change. However, existing frameworks provide limited insights into the cross-level dynamics that shape how learning ecosystems emerge, evolve, and operate across multiple organizations. This gap leaves practitioners without clear guidance on how multilevel contextual conditions and mechanisms interact to influence the development and sustainability of formal and informal knowledge-sharing relationships.

Objective: This study aimed to examine how knowledge is orchestrated across organizations in the digital transformation of health care, identifying key factors that foster an evolving interorganizational learning ecosystem. We developed an integrative model that explains how these influences give rise to diverse modes of collaboration and partnership.

Methods: We adopted a qualitative approach using a multilevel perspective to examine visions and experiences across individual, organizational, interorganizational, and sectoral levels. Drawing on a formative evaluation (2018 - 2023) of England's Global Digital Exemplar (GDE) program, we used multiple case studies and conducted interviews with experts both within and beyond organizational settings for data collection and adopted a grounded theory approach to analyze the data.

Results: The study identified a set of interconnected factors operating at the macroenvironmental, interorganizational, organizational, and individual levels that influence how interorganizational relationships and partnerships are initiated, structured, and sustained. Macro-level influences included policy developments, program mandates, technology supplier strategies, and intermediary actions. Interorganizational mechanisms involved relational recognition, collective identity, governance structures, proximity, and coordination practices. Organizational factors included external search strategies, absorptive capacity, past collaboration experience, and internal knowledge routines. Individual-level mechanisms encompassed intrinsic and extrinsic motivations as well as personal inhibitors. Synthesizing these findings, we have proposed an integrative model that positions relationship type along a 2D spectrum (formal-informal, internal-external origins) and illustrates how different factors trigger, mandate, control, and enable the evolution of an interorganizational learning ecosystem.

Conclusions: This study advances the theoretical understanding of learning ecosystems by explaining how multilevel contextual conditions activate mechanisms that give rise to diverse and evolving forms of interorganizational collaboration. Practically, we offer diagnostic and reflective tools that support policymakers and practitioners in assessing contextual conditions, selecting appropriate knowledge-sharing mechanisms, and monitoring how learning ecosystems develop over time. Our findings provide actionable guidance for designing and sustaining interorganizational learning systems capable of supporting digital transformation at scale.

(J Med Internet Res 2026;28:e83345) doi:10.2196/83345

KEYWORDS

digital transformation; health care sector; knowledge sharing; multilevel analysis; qualitative study

Introduction

Internationally, governments are enhancing and transforming their mode of operation, including processes, policies, and services, by adopting digital technologies, with the intention of improving service delivery; becoming more efficient and effective; and increasing transparency, interoperability, and public satisfaction [1,2]. The health care sector, in particular, is making substantial efforts to achieve digital transformation

XSL•FO
RenderX

[3-7]. By digital transformation, we refer to extensive and complex initiatives, such as the implementation of hospital-wide or multisite systems, which bring about significant changes in organizational operations and workflows [7]. This involves advancing technological infrastructure and applications, as well as changing organizational practices, processes, structures, performance, and value creation mechanisms [8-11]. However, despite significant investment in digital health technologies by both government and supplier companies, the transformation of the health care sector lags behind many other sectors owing to the complexity of the health care system, lack of requisite know-how, and immaturity of the market [4,12].

The complexity of health care arises from the diverse and often fragmented pathways that patients must navigate in their journey back to health [13]. Health care systems also involve a multifaceted network of providers, which in some countries, such as the United Kingdom, Canada, Spain, and Singapore, operate under a single organizational structure yet remain functionally distributed and interdependent. Ongoing innovations introduce fundamentally new ways of delivering care, including greater integration across medical specialities, improved coordination between primary and secondary care, and closer alignment between health and social care services [14]. In addition, health care is shifting from predominantly reactive treatment models toward preventive and proactive approaches [15]. At the same time, expectations regarding patient safety and treatment effectiveness continue to rise. Collectively, these factors increase the demands on system performance and make effective coordination increasingly challenging [13].

Digital transformation requires new competencies and skills from those orchestrating the changes and those adopting new technologies for transformation [2]. These can be acquired through learning from other organizations that have already undergone similar digital change processes [10,13,16]. However, in digital transformation, experiences in one setting cannot be mechanistically applied in other contexts [12]. Effective change management, therefore, depends on *contextualized knowledge*, that is, adapting lessons from one organization so they fit the specific circumstances, needs, and constraints of the receiving setting [17].

Building on this, recent research shows that knowledge sharing is most effective when it occurs between organizations that are contextually similar, such as acute hospitals facing comparable pressures, infrastructure, and goals [18]. Such similarity allows organizations to more readily translate and adapt one another's insights. To support this exchange, health care organizations engage in a range of interorganizational arrangements, evolving formal partnerships, exchanging formal codified knowledge (eg, through reports and blueprints), or sharing experience in informal ways (eg, site visits) [18-20]. These modes of engagement help organizations make sense of external lessons in ways that align with their own contexts, enabling more effective learning and application. In the rapidly growing field of digital transformation in the health care sector, policymakers promote and examine different interorganizational knowledge-sharing mechanisms [19,20], and of these, some may fail, while others may become sustainable or transform over time. Currently, however, the efforts in this area often proceed through trial and error, with limited systematic learning about what works and what does not. Emerging evidence suggests that while centralized, top-down approaches can be effective in certain circumstances, for example, by creating shared expectations or establishing a space for learning, they often do not translate into meaningful learning across diverse operational settings. In many cases, they fail to gain traction at the operational level [18].

Despite the recognized benefits of interorganizational knowledge sharing in digital transformation, there remains a notable gap in the literature concerning the specific factors that underpin effective interorganizational knowledge sharing, particularly in the health care sector, and how the factors mobilize a learning ecosystem that enables digitalization of the sector. In particular, while several frameworks offer valuable perspectives on digital transformation, they do not fully account for the *cross-level dynamics* that shape how learning ecosystems emerge, evolve, and function in practice, or *what forms* these ecosystems ultimately take. This theoretical gap has important practical implications: practitioners lack clear guidance on how multilevel contextual factors influence the development, effectiveness, and sustainability of formal and informal interorganizational learning relationships. Therefore, in this paper, we aim to explore and identify critical factors that shape the emergence and evolution of interorganizational learning ecosystems and to develop an integrative model that explains how these multilevel influences give rise to diverse modes of knowledge-sharing relationships during digital transformation.

To address these gaps, our study adopts a grounded, multilevel [21] analytical approach that enables us to trace how contextual conditions at different levels interact over time and across settings to shape processes of knowledge sharing. This approach allows us to identify recurrent patterns across diverse empirical materials and build an empirically grounded framework that explains how different forms of interorganizational learning arise and are sustained. The study therefore advances both the theoretical understanding of cross-level learning processes in digital transformation and the practical insights for cultivating effective interorganizational learning ecosystems in health care.

We draw on a recently completed formative evaluation (2018 to 2023) of a national digitally enabled transformation initiative in English provider organizations called *The Global Digital Exemplar (GDE) Program* to identify the range of factors and modes of interorganizational knowledge sharing as they unfold over time and multiple locations.

This work is timely, as it has the potential to inform policy and practice in relation to the growing investments being made in digital transformation of the health care sector not only in England but also in many other high- and middle-income countries. In doing so, the study provides a series of questions around factors and modes of knowledge sharing that contribute to enabling digital transformation, to be addressed by practitioners and policymakers.

XSL•FO
RenderX

Methods

Research Approach

In order to identify the factors at multiple levels of analysis contributing to interorganizational knowledge sharing, we adopted a longitudinal qualitative approach in the interpretive tradition [22], drawing on multiple case studies and conducting interviews with experts in the field within and beyond organizational settings.

Research Setting

We studied a national digital transformation initiative in the English health care sector from 2017 to 2023, called the GDE Program. The GDE Program was a national investment in the transformation of a variety of care settings, including acute care, specialist care, mental health, and ambulance services of the National Health Service (NHS) in England, with a focus on digitally advanced sites to become international exemplars of excellence in order to provide learning for later organizations planning to go through digital transformation. In this program, provider organizations, henceforth referred to as GDEs, were paired with somewhat less mature fast follower (FF) provider organizations to promote learning across settings. GDEs and FFs captured the best practice models and lessons, including technologies, processes, and structures, in "blueprints," which were aimed to be disseminated within and beyond the program to accelerate the spread of transformation knowledge. Shared learning in the program was also planned to be facilitated by learning networks launched in the first half of the program, bringing together GDE and FF team members with an interest in specific topics.

Data Collection

Overview

We collected data using semistructured interviews, observations, and textual sources. These methods were used concurrently throughout the study period. Data collection was initiated through interviews, which were subsequently complemented by related observational and documentary data to deepen contextual understanding. In cases involving user groups, direct observations were conducted first and were followed by interviews to further explore emerging themes.

Sampling

We conducted in-depth case studies on a set of 12 provider organizations, carefully selected to ensure representation of both GDE and FF acute provider organizations, as well as mental health providers, reflecting the broader sample. The selection process aimed to include maximum variation, considering factors such as a range of different core electronic health record systems, geographical locations, and organization sizes.

Our participants included system vendors, national stakeholders such as program managers, policymakers, and digital health leaders who were responsible for setting digital strategy or who held formal roles such as Chief Clinical Information Officer, Chief Technology Officer, or equivalent positions. To ensure a diverse range of perspectives across the national digital health landscape, we purposively sampled interviewees operating within the NHS national program environment. Recruitment followed a systematic, multistage procedure. First, we identified potential participants through national program-level gatekeepers, including the senior responsible owner overseeing the implementation of the wider program, who provided access to individuals directly engaged in strategic and operational decision-making. Second, we used our existing professional networks to identify individuals holding formal digital leadership roles within NHS organizations who were actively involved in digital transformation efforts. Finally, we used a snowballing approach by asking each participant to recommend others, particularly those who might hold alternative perspectives, to further broaden the diversity of views included in the study.

Semistructured Interviews

We conducted interviews with 228 recognized experts from NHS England professional domains, policymakers, and suppliers, who were involved in the design, launch, monitoring, evaluation, or running of the GDE Program at some point in time. Rather than being formally pilot tested, the guides were iteratively tailored to the expertise and role of each interviewee to ensure relevance and depth of discussion. The questions aimed to explore the evolving nature of knowledge-sharing modes, relationship types and their features, and factors influencing interorganizational knowledge exchange. Data collection was conducted in both face-to-face and online settings. Interviews typically ranged from 30 to 90 min in duration, with an average length of approximately 1 hour. Data were audio-recorded where permission was granted. For nonrecorded interviews, extensive notes were taken. Participants were offered the opportunity to review the final study report, and several provided feedback and comments at that stage. Data collection continued until thematic saturation was reached, at which point no new concepts or insights emerged from the data.

Observations

To complement data collected through interviews, observations of meetings related to GDE Program activities and user group events focused on particular digital technologies being implemented within participating organizations. These observations enabled us to understand how different stakeholder groups engaged with the technologies, addressed challenges, and coordinated implementation work. In addition, we observed 3 national conferences centered on the digitalization of the NHS, where program leaders and adopter sites shared their experiences, lessons learned, and implementation strategies. During the observations, we captured the setting layout, the way participants moved and interacted, the actors, insights into processes and discussion content, and insights into interaction outcomes.

Textual Data

Additionally, to complement the observations and interviews, we collected relevant documents, such as business plans, blueprints, lessons learned documents, and implementation documents, to gain insights into the type of knowledge being shared and relevant processes. To understand broader system-level discourse, we supplemented this with data from online forums, such as NHS Futures and the Digital Health

XSL•FO
RenderX

Network, focusing on what knowledge was shared, how it was communicated, and the forms of support or collaboration that emerged across the wider digital health community.

Data Analysis

Our data analysis integrated well-established approaches, including longitudinal case analysis, grounded theory development, and content analysis. The process was iterative, requiring continuous movement between the data and the evolving theoretical framework [23]. However, for clarity, we present our analysis as a structured sequence of 5 main stages. In stage 1, the lead researcher, HM, utilized NVivo 14 (Lumivero) to organize and integrate all data sources and conducted initial open coding to capture the breadth of stakeholder views and experiences. This stage generated a large pool of provisional codes across interviews, observations, and documents.

In stage 2, we adopted a focused coding strategy to move from initial descriptive codes to analytically meaningful first-order codes. Guided by the principles of parsimony, conceptual salience, and recurrence across data types, we treated sentences and paragraphs as coding units [24] and read the full dataset iteratively, including catalogs and supporting materials. We assigned labels using in vivo codes where appropriate or concise descriptive terms and then grouped related segments into coherent conceptual categories [25]. Importantly, our intention at this stage was not to catalog every detail within such a large dataset but to distil patterns that were recurrent, theoretically relevant, and directly related to our research questions. This analytic focus explains why the final set of first-order codes is relatively concise (n=33). To ensure robustness, coding decisions were discussed within the research team at multiple points, allowing for refinement of categories, challenge of interpretations, and agreement on the conceptual boundaries of each code. Throughout this process, we examined perspectives at multiple analytic levels, including individual, adopter organization, government, and technology supplier, as well as across different stages of digital transformation (procurement, adoption, and use). This enabled us to generate a focused yet comprehensive set of first-order codes that meaningfully represented mechanisms shaping knowledge sharing across organizations.

In stage 3, we identified links among first-order codes to form the emerging second-order themes [24]. This involved discussing disconfirming evidence within the extended research team to minimize the risk of researcher bias.

Stage 4 involved linking second-order themes to form aggregate dimensions [24,26], associated with our research question. These 3 stages (2, 3, and 4) formed our data structure, as presented in Table 1. The Results section begins with explaining this data structure in detail.

XSL•FO
RenderX

Table . Categorization of key factors contributing to knowledge sharing.

Aggregate and second-order themes	First-order codes
Macroenvironmental factors	
Policy developments	• Formation of GDE[a]/FF[b] pairings as a prerequisite of GDE funding • Blueprinting activities mandated by the GDE Program • Digital Academy supporting skill development/cohort/cadre of digital transformation leaders • Wachter Review leading to the development and advancement of new roles such as CCIO[c]/CNIO[d] • Formation of the Digital Health Network as an NHS[e]-organized community
Technology supplier strategies	• Co-ordinating technology-specific user groups • Planning pilot site visits • Suppliers diversifying the technology market
Intermediary actions	• Independent technology consultants moving across settings • Formation of communities of interest by adopter organizations
Interorganizational factors	
Relational recognition	• Development of collective identity (eg, GDE status or Digital Aspirant status) • Naming organizations as GDEs (indicating their digital maturity) as opposed to FFs (indicating them as followers)
Structural governance mechanisms	• Time- and resource-bound contracts versus ad hoc interactions • Need to exchange knowledge in a structured repository (eg, blueprints) versus unstructured sharing (eg, email or site visit)
Interorganizational proximity	• Organizations with closer geographical proximity forming formal GDE/FF partnerships • Personnel from geographically closer organizations communicating and moving between sites • Cost sharing or synergy seeking
Organizational factors	
External search strategies	• Identifying a solution to a certain problem • Identifying who has implemented particular technologies • Setting future digital transformation strategies
Absorptive capacity	• Internal knowledge acquisition strategies • Time allocation to participate in community or networking events
Past experience	• Managing knowledge requests from external organizations based on past experience
Individual factors	
Intrinsic motivations	• Reputational and professional status
Extrinsic motivations	• Creation of communities of shared interests
Inhibitors	• Personal inhibitors
Spectrum of interorganizational relationships and partnerships	
Origins of partnerships/relationships	• Policy (eg, formal agreements with GDE/FF sites) • Independent groups of enthusiasts (eg, Digital Health Network) • Provider organizations (eg, periodic or ad-hoc meetings, trainings, and events) • Suppliers (eg, technology user groups and pilot site visits)

XSL•FO
RenderX

Aggregate and second-order themes	First-order codes
Governing structures and durations	• Formal contractual agreements • Flexible periodic engagement • Informal need/opportunity basis

[a]GDE: Global Digital Exemplar.

[b]FF: fast follower.

[c]CCIO: Chief Clinical Information Officer.

[d]CNIO: Chief Nursing Information Officer.

[e]NHS: National Health Service.

In stage 5, after identifying the aggregate dimensions in stage 4, we noted strong parallels between our grounded analysis and context-mechanism-outcome (CMO) [27] configurations described in realist literature, that is, the taxonomy of multilevel factors incorporated context and mechanisms, and shaped the spectrum of interorganizational relationships and partnerships (outcomes). Building on the similarities, we propose a framework that shows the relationship between the aggregate dimensions. We structured the reporting of this study in accordance with the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines to enhance rigor, transparency, and completeness [28].

Ethical Considerations

We obtained institutional review board approval from the University of Edinburgh Research Ethics Committee (27.11.2017). Informed consent was obtained from all participants prior to their involvement in the study, and all data collected were anonymized to ensure confidentiality and protect participant and organizational privacy. Participants were not compensated for their time.

Results

Overview

Overall, our dataset consisted of 228 interviews; observations of 113 national strategic meetings, user groups, and conferences; analysis of 277 documents; and desk research into online platforms of knowledge-sharing settings such as forums and webinars. Table 2 presents the interviewee characteristics. We present the findings under 2 key headings: multilevel factors contributing to the sharing process and the spectrum of interorganizational relationships and partnerships. Subsequently, informed by the CMO perspective, we present an integrative view of how the factors across 4 levels—macroenvironmental, interorganizational, organizational, and individual—influence knowledge-sharing modes. The factors identified in each category are not exhaustive; however, they highlight some of the most recurrent factors identified in our data within each category.

XSL•FO
RenderX

Table . Interviewee characteristics.

Characteristic	Value, n (%)
In-depth case study sites (12 sites; n=178)	
CCIO[a]	34 (15)
Program manager	29 (13)
CIO[b]	27 (12)
Nursing roles	15 (6)
Digital director/lead	12 (5)
CNIO[c]	11 (5)
Project manager	11 (5)
Pharmacist	7 (3)
Other IT-related roles	20 (9)
Clinical transformation lead	6 (3)
Benefit realization lead	6 (3)
Wider data collection (n=50)	
Policy	44 (19)
Industry	3 (1)
Others	3 (1)

[a]CCIO: Chief Clinical Information Officer.

[b]CIO: Chief Information Officer.

[c]CNIO: Chief Nursing Information Officer.

Taxonomy of Multilevel Factors Contributing to the Interorganizational Knowledge-Sharing Process

Macroenvironmental Level

Our study identified 3 key groups of macroenvironmental factors that influenced knowledge sharing across the settings.

First, *policy developments*, such as government funding and training programs, were seen to have a profound impact in terms of knowledge sharing across different sites. For example, the development of the NHS Digital Academy, which was a close collaboration between several universities and the NHS, supported the skills development of health professionals and decision makers to become digital transformation leaders. In parallel with this, new Digital Health Network communities formed, and these together brought over 5500 NHS digital health leaders to share knowledge and collaborate through online communities, regional and annual events, and webinars. Another example is the Wachter Review conducted between 2015 and 2016, led by an external digital health expert (from the United States), which reviewed the digitalization of the NHS and led to several recommendations that were put into practice, such as the development and advancement of new roles (eg, Chief Clinical Information Officer and Chief Nursing Information Officer) as mediator roles between digital transformation decision makers, front line health care staff, and middle management. In these 3 examples, the macroenvironment was an enabler of transformation and a promoter of knowledge sharing across organizations.

> *And then out of that, so I go to digitalhealth.net events so, like, the summer schools and the leaderships, something that they do, I go to that…I'm part of the One HealthTech, I don't know if you've heard of that, so it's like a volunteer organization for women in health tech. So I represent [site_A]. So there are some other forums where I've access to other people and you tend to meet…there's lots of networking at those events…. So I get lots of emails from people, like, the CNIOs and people from other trusts [hospitals] just asking questions and I have their contact so I can email them. So there's quite a lot of informal, once you've met somebody, that, kind of, networking, you just…you know that they're doing a certain thing if you want to know about that, you can contact them.* [Participant from Site_I]

Despite their enabling roles, the policy development factors may sometimes hinder effective collaboration across organizational settings. For instance, some sites involved in GDE/FF partnerships highlighted that their pairings had been established on the basis of policy criteria, often under compressed timeframes, sometimes resulting in a lack of effective engagement between the partners.

Second, *technology supplier strategies* affected the types of relationships being formed and the collaborations between different organizations. For instance, in the GDE Program, supplier-coordinated networking activities, such as user groups, pilot site visits, and connections of key individuals across organizations, were common. However, many of these

XSL•FO
RenderX

supplier-organized initiatives, while being beneficial in initiating conversations, in particular, with international partners, only allowed for supplier-set goals.

> *When you go on a site visit, you're chaperoned around, no one lets you talk to anyone. Even when you go on the PACS visits, we were definitely stuck in a room and not allowed out.* [Participant from Site_AG7]

In the majority of cases, interviewees highlighted that supplier-arranged learning opportunities drive adopter organizations in particular directions, such as buying particular digital technologies, rather than allowing for open conversations.

> *…the problem with that is that [user group] if you take supplier_name, for example, it's always going to be about supplier_name, well, you should do it this way with this supplier_name product, it's not always the right choice.* [Participant from Site_C]

However, over time, the nature of supplier-organized groups has evolved, becoming increasingly user-driven and confident. For example, users became more active in certain user groups regarding electronic prescribing systems, collaborating to hold suppliers more accountable. Similarly, in a different user group setting, the supplier encouraged their adopters to commit to share learning, reinforcing the shift toward more reciprocal and empowered user-supplier relationships.

Furthermore, other supplier strategies, such as diversifying the market by, for instance, offering best-of-breed solutions on one end of the spectrum as opposed to mega-suites at the other end of the spectrum, had an indirect effect on how and which adopter organizations communicated.

Finally, *intermediaries* had a noticeable influence in orchestrating the sharing of knowledge, particularly where formal policy-developed structures or supplier strategies were deemed to be limited in effectiveness. There are different types of intermediaries, ranging from individuals, such as consultants and technical experts, who facilitate sharing by providing specialized professional services to organizations, to public and private entities that provide technology brokerage, such as suppliers and individual consultants.

> *…Lessons Learned paper that I had, I was given that on day one, when I started, and I went through that myself, and since then, there really hasn't been much engagement at all. Now we have had a specialist that worked for Site_Name on their EDRMS solution…We've put 15 days into that, so obviously quite a lot, it's £750 a day…He brought a lot of experience to the table from what Site_Name did right, and what they did wrong. So, effectively, he was our pseudo blueprint.* [Participant from Site_M]

User-driven communities organized around a particular interest, such as a specific type of technology, are also effective intermediaries, which facilitate learning and cross-organizational collaborations.

> *…we decided, we need a user group, and we need to engage with other users, so that we've got a collective voice for the suppliers. So, we established the national user_group_name, and our CIO was the chair of that…And we also established something called the user_group2_name, which was for the more techies, about how to manage upgrades, how to do certain things in the system, so it's like a knowledge sharing, question and answer thing. And we asked every trust [hospital] to chip in a nominal amount of fund, we're talking about 200 quid…* [Participant from Site_E]

In general, macroenvironmental factors impacted interorganizational knowledge sharing both directly and indirectly. In direct terms, they mandated (eg, blueprinting activities in the GDE Program) and structured relationships (eg, GDE and FF relationships formed between digitally advanced organizations and less digitally advanced organizations to share experiences). In indirect terms, they influenced the ability and motivation of organizations to form relationships for knowledge-sharing purposes (eg, user communities).

Interorganizational Level

Our analysis highlighted 3 types of interorganizational factors influencing knowledge sharing across the settings.

First, we identified that *relational factors*, including the formation of collective identity, power relationships, and trust between organizations, directly affected knowledge-sharing practices. The development of collective identity expanded communication across organizations of similar titles. For instance, we observed that different types of interorganizational collaboration events and meetings were set up (either by adopter organizations or by the NHS), with reference to the status of the organization being a GDE hospital. In such events (eg, the GDE network 2019 event in London or a meeting between Site_D and Site_F), the collective identity was used as an anchor for communication and collaboration.

In terms of power relationships, for instance, our analysis showed how the labeling of organizations as being digitally advanced (GDEs) or being follower organizations (FFs) impacted the dynamics of sharing. While in some instances, the relationship created a means for GDEs to share their advanced competencies with their FFs, in other instances, it hindered sharing, as some FFs did not see themselves as lagging behind in competence and capability. Some FFs, therefore, did not aspire to learn from their GDE peers. This formal power relationship, which differentiated hospitals in their digital status, could also, in some instances, get in the way of sharing knowledge across 2 digitally advanced sites.

> *It's a very strange place the…well, you'd have to say that some trusts [hospitals] that aren't GDEs would feel that they're much more digitally mature than the GDEs. So, whilst the GDE programme singled out a number of trusts [hospitals] for these programmes, you almost then say to those trusts [hospitals], when everybody's looking in, they're saying, we're better than them, or actually, we can do a better job than them. So, you don't tend to get that learning going across, I suppose, I don't know, how much learning do you get between [Site_A] and (Site_B]. Or, if*

XSL•FO
RenderX

> *someone develops some software in [Site_A], would [Site_B] take it? You might find that [Site_B] said, well do you know what, we've got something better, or actually if we did x, y and z, we'd…so it's that competition, in what is supposed to be a national health service, there is that reluctance to adopt other people's ideas and technology.* [Participant from Site_B]

Second, our study pointed to the importance of *structural relationship governance mechanisms* between organizations in enabling or hindering the knowledge-sharing process. For instance, the formal partnerships were planned based on clear strategic organizational needs by strategic decision makers. They prescribed how knowledge exchange was to take place and how the communications were organized. This afforded very different knowledge-sharing practices to informal relationships, which were formed on an ad-hoc basis and were based on particular local needs, the experiences of each organization, aspects that were beneficial, and organizational availability. Additionally, the structure of the relationship, including how it was governed and what routines and processes were involved in knowledge interactions, had a direct effect on different aspects of knowledge sharing. In terms of governance, for instance, whether the relationship is an informal relationship that is established and runs based on the availability of different organizations, or whether it is based on some type of contractual association with a defined scope and defined roles, affects the type of knowledge being exchanged. This includes, for example, structured documents with codified knowledge on how to implement a functionality versus discussions on solving a particular problem using system configuration and workarounds during a networking event, the direction of knowledge flow (eg, from a GDE to an FF or multidirectional across multiple organizations), the frequency of interactions (eg, a specific number of site visits in a given period of time vs ad-hoc site visits based on requirements), and the extent and longevity of sharing.

Finally, *interorganizational proximity* influenced knowledge sharing across organizations. Proximity has different aspects, such as physical proximity or the degree of technological or strategic similarity. For instance, organizations with a closer geographical proximity either formed formal GDE/FF partnerships or had personnel communicate and move between sites readily in informal ways to exchange knowledge. Another example was the procurement and implementation of similar technologies by different hospitals, which led to the development of relationships/partnerships, with motivations such as cost-sharing or synergy-seeking.

> *Then I think because at the same time [Site_C] were going through and had decided to also take [technology_Name] as its electronic patient record system, they were our formal fast follower, and so it enabled us to work with them. But in fact we had always agreed that we would provide wider support to other hospitals also considering or implementing [technology_name]. So as well as [Site_C], we were providing support to [Site_D], [Site_E] and [Site_F].* [Participant from Site_A]

Organizational Level

We observed that 3 organizational factors had a direct influence on the sharing of knowledge.

First, an organization's *agenda for searching for external knowledge* was an important factor. Some organizations only looked for external knowledge when a particular need arose, such as the need for the implementation of a new piece of technology, while others broadened external knowledge search mechanisms to identify unknown or future strategic opportunities.

> *I think the other thing we sort of did over the summer, was have a look at all of the GDE blueprints. I know that we're a GDE, so I suppose the assumption is that, because we're a GDE, we've done everything, but I'm a great believe in learning from somebody else. So, we kind of did a trawl of all of the blueprints that have been published to date, to see whether or not there was anything we're not taking use of as much as somebody else is, and really starting to think about, what's our next cycle of digital activity? And you can kind of see there, that pretty much the first two columns says, oh, we're probably doing about 85 per cent of everything that's out there, but the final column says, there are a few options. And I think, in particular for us, it's interesting that we're doing webinars for our staff, but not for our patients.* [Participant from Site_L]

Second, an organization's *absorptive capacity* played an important role in enabling sharing across organizations. Absorptive capacity refers to an organization's ability to acquire, assimilate, and utilize external knowledge, information, and innovations. For instance, some organizations had implemented internal processes and routines to facilitate learning and sharing, enabling them to acquire insights from other organizations and manage external knowledge-sharing demands. Conversely, in cases where less time was dedicated to absorbing and utilizing external knowledge, such as during the procurement of new digital technologies, these organizations gained limited benefits from the knowledge held by others, which sometimes led to experiencing challenges previously faced by other organizations. Some interviewees highlighted that their organization had established a routine for them to allocate time for periodic conversations with other organizations, enabling structured opportunities for knowledge exchange. By introducing these routines, organizations can foster a culture of continuous learning, gain insights from the experiences of peers, and adopt the best practices in digital transformation. This approach not only accelerates their own digital initiatives but also helps them navigate challenges more effectively, ultimately strengthening their overall organizational capabilities.

> *…pretty much every month the CIOs or heads of IT delivery […] have opportunity to meet each other …to actually talk to each other. And that makes the world of difference…the other week, [name] went along and came back with a small list of this is what I've got from [org_name_A], this is what I've got from [org_name_B], could you have a chat with this*

XSL•FO
RenderX

> *person please…I just ring them up and say, yeah, you wanted a chat, this is what it's about and vice versa…* [Participant from Site_G]

Finally, an organization's *past experiences* in terms of identifying, building, and maintaining relationships had a noticeable impact on the formation of its knowledge-sharing practices. For instance, organizations that were more experienced in managing and coordinating relationships established clear processes for managing the requests from external organizations asking for their knowledge and experiences.

> *They were asked not to make direct contact with [site_B] staff, they had to go through a specific office and up to monitor, well not monitor but manage and control the contacts. Which worked fine actually because they were as I said, they replied very quickly when we begin a request, which was usually by email and we got through that pretty quickly. They wanted to avoid loads of people making contact with loads of [site_B] staff and disrupting them unnecessarily.* [Participant from Site_C]

Individual/Interpersonal Level

We found 3 individual factors impacting formal and informal knowledge sharing across organizations.

First, we observed that individuals are driven to share knowledge by *intrinsic motivations*, specifically their self-interest. Our observations revealed that participants exhibit a heightened willingness to share knowledge when they perceive tangible benefits arising from such collaborative endeavors. Among these motivating factors, the prospect of bolstering one's reputation and professional status emerged as a significant catalyst for active involvement in knowledge-sharing communities.

Furthermore, the desire to gain profound insights into alternative perspectives and diverse experiences emerged as another compelling factor propelling knowledge sharing. Participants acknowledged the richness that emerges from exposure to a variety of viewpoints, and they mentioned actively seeking to broaden their understanding by tapping into the collective wisdom of others.

> *so that's been really helpful, to have that outside perspective from someone else in a different Trust [hospital] doing something that's similar but having a different viewpoint. As I find the people are fantastic, so much knowledge, so much experience, but I appreciate we're all at the same Trust [hospital] so we've got almost like the same vision, and sometimes it's just nice to see someone else's perspective who's not part of this Trust [hospital].* [Participant from Site_E]

Second, beyond intrinsic motivations, *extrinsic factors* rooted in altruism played a pivotal role in fostering the sharing of knowledge across organizational boundaries. The altruistic dimension of knowledge sharing is particularly evident in instances where individuals perceive that their contributions go beyond personal gain and actively contribute to the construction or sustenance of communities of shared interests. Participants recognized that by sharing their expertise, insights, and experiences, they are instrumental in nurturing a collaborative environment where collective learning and problem-solving thrive. This communal aspect became a source of motivation, as actors derived satisfaction from the sense of building and sustaining a network of like-minded individuals. Furthermore, the fostering of feelings of companionship and camaraderie emerged as a powerful intrinsic motivator for knowledge sharing. Participants found fulfillment in being part of a community where the exchange of knowledge is not solely transactional but is imbued with a sense of shared purpose and mutual support.

> *we've now set up a GDE group with everyone that looks after benefits for their Trust [hospital], and we're using it as a sharing platform to help each other…* [Participant from Site_E]

While our study did not aim to identify barriers, in the case of individual factors, we encountered individual inhibiting factors that introduced complexities for the sharing of knowledge that necessitated careful consideration. Three notable aspects emerged as noteworthy impediments: the allocation of time to engage in sharing activities, the concept of knowledge-based psychological ownership, and the nuanced emotional responses individuals exhibited when navigating the uncertainties inherent in interorganizational settings. The time commitment required for participation in knowledge-sharing activities proved to be a substantial factor. Individuals, often operating within demanding professional environments, found themselves grappling with the challenge of allocating sufficient time to contribute meaningfully to knowledge-sharing initiatives.

> *The is a CSO network but unfortunately, we don't meet terribly often. And actually, most of us are just too busy…but in the end the problem is it doesn't get enough time, I think. I am still a full time clinician, so I still have to do a very much full time on-call job.* [Participant from Site_G]

Some individuals were possessive of their knowledge, struggling with what to share openly versus what to protect for their organization's benefit. This sense of ownership created barriers to the free flow of insights. Additionally, concerns about partners' behaviors affected the willingness to share, introducing a complex balance between trust and caution. Apprehensions about defining knowledge boundaries (what to disclose and what to withhold) further highlighted the nuanced nature of interorganizational knowledge sharing. These individual perspectives and emotional responses shaped knowledge transfer dynamics, underscoring the need for a deeper understanding to foster effective collaboration.

Linkages Across Levels in Activating Knowledge-Sharing Modes

The factors identified in our study do not operate in isolation; rather, elements across macroenvironmental, interorganizational, organizational, and individual levels are highly interconnected and often mutually reinforcing in shaping knowledge-sharing relationships. For example, national policy developments, such as the Wachter Review, not only stimulate the creation of new

XSL•FO
RenderX

digital leadership roles (eg, Chief Clinical Information Officer and Chief Nursing Information Officer) but also contribute to the emergence of a digitally skilled cadre through initiatives like the NHS Digital Academy, thereby improving the absorptive capacity of organizations and their ability to engage meaningfully in communities of practice. Similarly, program-level structures, such as mandated GDE/FF pairings and blueprinting requirements, interact with relational dynamics, including the formation of collective identities (eg, GDE status) and the development of trust-based recognition between partner sites, which in turn facilitate voluntary knowledge exchange beyond formal contractual obligations. Technology supplier strategies also intersect with organizational search behaviors: suppliers' diversification of the market can drive adopter organizations to seek external knowledge (eg, identifying who has implemented specific technologies), often aided by intermediary actors or technology-specific user groups that bridge otherwise disconnected sites. Geographical proximity further amplifies these interactions, as staff from nearby organizations are more likely to visit one another, participate in shared communities of interest, or move across settings, thereby strengthening informal knowledge flows that complement formal governance mechanisms. Finally, individual-level motivations, such as professional reputation, intrinsic interest, and past experience in handling knowledge requests, shape how actors engage with these structures, while personal inhibitors can constrain knowledge sharing even when enabling conditions are otherwise present. Together, these interconnected elements illustrate how multilevel factors combine to activate and shape the modes underpinning knowledge sharing across the health system.

Spectrum of Interorganizational Relationships and Partnerships

Together, the above interconnected conditions contributed to a spectrum of interorganizational relationships and partnerships—what we refer to as modes of knowledge sharing—that emerged from multiple origins and were shaped by varying degrees of structure and formality. These mechanisms ranged from highly orchestrated arrangements to more flexible, organically developed connections. Across this spectrum, organizations used diverse tools and practices for knowledge sharing as described below.

The *origins* of relationships ranged from internal (individual or communal within and across transforming organizations) to external (beyond the transforming organizations). In the case of internal origins, sometimes provider organizations collaborated on a local needs basis (eg, site visits or phone calls between individuals) or more strategically entered into intentional cultivation of networks and communities in response to collective needs or to achieve collective strategic benefits. For instance, when multiple organizations needed particular functionalities from the same vendor, organic and spontaneous collaborations between 2 or more user organizations helped to create a collective force to exert leverage over technology vendors or to collaborate and form a technological solution that benefited multiple organizations.

> *And then on going forward it was about where we had joint requirements for development such as around various different areas such as out-patients or theatres. We'd try and come up with a joint specification so that [supplier_name] only had to develop once.* [Participant from Site_D]

In the case of external entities initiating relationships, in some instances, interorganizational relationships were initiated by formal mandates from the government. For instance, in our study, 19 strategic pairings were formed between GDEs and FFs. These partnerships were based on a range of factors such as common core technologies and local strategic groupings, for instance, between organizations that had a particular electronic patient record system or between nationally coordinated collaborations of health care organizations and local authorities.

Technology suppliers were also initiators of particular types of relationships around specific technologies, both nationally and internationally.

> *...there's now quite good, not only UK, but a wider European group forum. Plus, we have regular conversations with Melbourne. Pre-[technology_name], our pediatricians used to go back and forth with Melbourne. Melbourne physicians used to come here, our guys used to go there, so there's a very strong relationship with Melbourne. They have [technology_name]. It's a bit like an onion, you know, layers and layers of different communities that have formed very nicely.* [Participant from Site_A]

The wide spectrum of interorganizational relationships had different *governing structures and durations*. For instance, government-initiated partnerships could range from national conferences or conference series (short-term, flexible periodic engagements) to fixed-term/long-term contracts between different provider organizations (eg, GDE/FF partnerships). These contracts mandated a number of aspects, such as the type of sharing (documentation), the interval for sharing, and what is to be shared. Despite investments into these formal means of knowledge sharing, many of the initiatives did not reach their anticipated longer-term benefits, despite being easy to measure, and their purposes changed over time. One example is blueprints, which were mandated by the GDE Program and were designed as tools for sharing the knowledge needed to implement specific models of change to be adopted by the wider ecosystem. While they offered codified knowledge about new process designs, they did not provide the means to implement certain technologies in the new context, as the implementation process involved a range of tacit knowledge that was not included in the blueprints. Many provider organizations used blueprints as a tool to identify which organizations and which actors were sources of particular knowledge in order to establish informal contact and to allow for the sharing of knowledge beyond codified documents.

We also noted that not all formal initiatives proved to be self-sustaining over time. For instance, GDE Program managers launched a number of learning networks, many of which were terminated after the end of national funding. Only some

XSL•FO
RenderX

networks driven by professional agendas and interests (eg, the Hospital Electronic Prescribing Medicines Administration Network) sustained and transformed into national communities of practice after the completion of the initial funding period. Collaboration and sharing of knowledge in these communities were not based on policy-mandated contracts, and instead, knowledge sharing was orchestrated through collaboration of provider organizations as they captured local and collective benefits from contributing to the community.

Integrating Influencing Factors Into the Spectrum of Interorganizational Relationships

Building on these findings, we developed an integrative model (Figure 1) that brings together the multilevel factors (context and mechanisms on the left-hand side) identified in our analysis with the resulting spectrum of interorganizational relationships and partnerships (outcomes on the right-hand side). The model illustrates how different contextual conditions and mechanisms combine to produce a diverse array of knowledge-sharing relationships, which vary not only in formality but also in their roots and modes of initiation. To capture this variation, we conceptualized a 2D spectrum: the x-axis represents the degree to which a relationship is formal or informal, while the y-axis indicates whether its origin is internal (arising from organizational strategies, leadership roles, or intrasystem initiatives) or external (emerging from policy directives, supplier strategies, intermediary actions, or professional networks). Importantly, the left and right sides of this spectrum are connected through a range of influences, such as triggering, mandating, controlling, and enabling, which shape, prompt, or support interorganizational knowledge sharing across different settings. By positioning relationship types within this matrix, the model demonstrates how combinations of enabling factors give rise to different patterns of collaboration and partnership across the ecosystem, offering a coherent way to understand the diversity and dynamics of interorganizational knowledge-sharing arrangements.

Figure 1. Integrative model of interorganizational learning ecosystems in health care digital transformation.

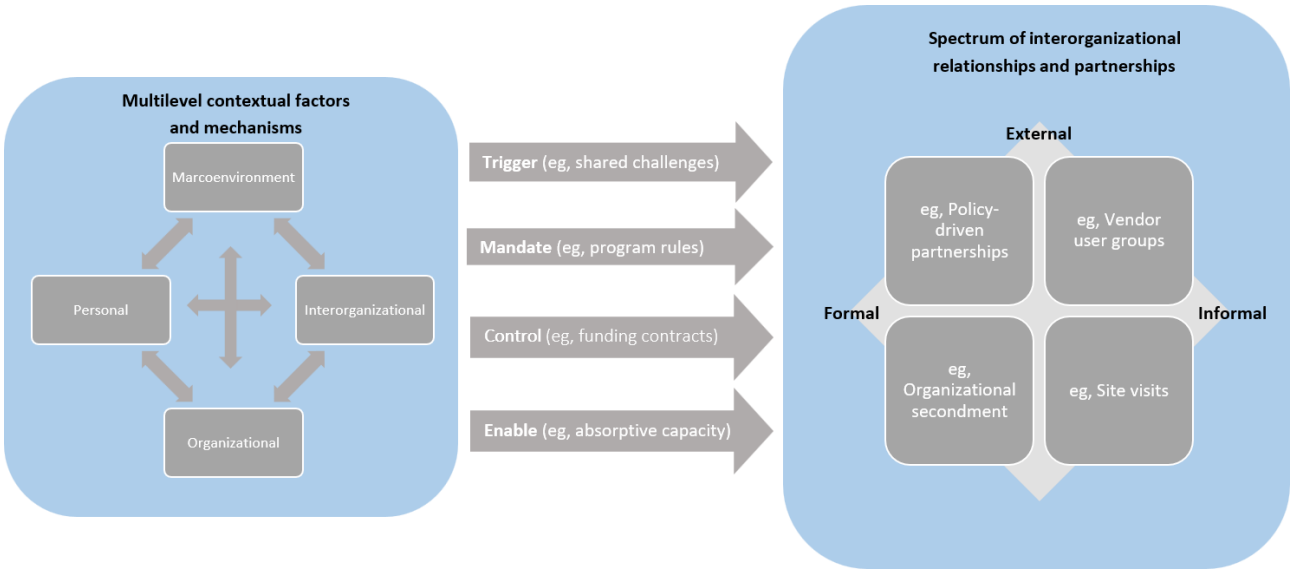

Discussion

Principal Findings

Our findings show that effective interorganizational knowledge sharing is shaped by different configurations of factors linking across the levels of a system. At the macro level, policy stability, national investment, and the presence of formal digital transformation programs created perceived legitimacy, shared purpose, and a sense of mandate among organizations, resulting in higher engagement in knowledge-sharing activities and more formalized collaborations. At the interorganizational level, networks characterized by trust, prior relationships, and aligned implementation timelines enabled modes of reciprocal exchange, openness, and problem-solving, leading to sustained cross-site learning, joint solution development, and more mature partnerships. In contrast, competitive procurement pressures or misaligned incentives triggered protective behaviors, producing fragmented or transactional exchanges. These examples of different configurations illustrate how and why knowledge-sharing modes vary in effectiveness and evolve over time.

From Knowledge Sharing to Learning Ecosystems: A Theoretical Contribution

In this section, we compare our empirical findings with the broader literature on knowledge sharing across organizations and present how our findings contribute to the development of the concept of evolving learning ecosystems.

Combining our findings with the extant literature that addresses *macroenvironmental factors* on interorganizational knowledge sharing, we identified a wide range of macroenvironmental factors, including policy developments, intellectual property regimes [29], technological sophistication levels [30], technology suppliers and their strategies [31], diverse markets [32], and other external agencies (eg, universities, civil society groups, third-party firms that have close connections with the health care system, etc), which can trigger, mandate, control, or enable interorganizational knowledge sharing. Policy-related factors can act as “windows of opportunity” [29] that enable or

XSL•FO
RenderX

mandate the formation of relationships for sharing specific types of knowledge [33]. However, the outcome of using the mechanisms mandated or enabled by policies is dependent on how knowledge-sharing stakeholders exploit these opportunities [18]. For instance, when the government creates formal partnerships and mandates the sharing of documents in a repository accessible to a wide range of organizations, the scale of sharing may be large, but the scope and depth of sharing may be limited. Other external entities, such as suppliers (through their strategies) and intermediaries, also shape the types of relationships being formed and the collaborations that may take place between different organizations [30]. Typically, relationships formed by suppliers are less formal, sometimes ongoing, and sometimes on an ad-hoc basis. A common finding in the GDE Program was supplier-coordinated networking activities such as user groups, pilot site visits, and connections of key individuals across organizations. Intermediaries may offer formal partnerships by providing specialized professional services to organizations and facilitating sharing between them [34,35] or providing technology brokerage activities [36]. Intermediaries may also enable the formation of communities of interest, where the relationship is less formal. In general, the macroenvironmental factors impact interorganizational knowledge sharing both directly and indirectly. In direct terms, they can mandate (eg, blueprinting activities in the GDE Program [18]) and structure relationships (eg, GDE and FF relationships formed between more digitally advanced organizations and less digitally advanced organizations to share experiences [37]). In indirect terms, they can influence the ability and motivation of organizations to form relationships, for example, by identifying who holds what knowledge about a particular type of technology.

Interorganizational factors, including relational factors [38], structural governance mechanisms, and interorganizational proximity [38], have a profound impact on shaping the early stages of relationship/partnership building. Our research shows that by building a collective identity that involves the development of a common objective or a mutual win and involves shared rules and regulations in terms of both formal [39] and informal [40] relationships, knowledge sharing can be encouraged. These can be achieved through building interorganizational teams or communities of interest that invest in modes of knowledge exchange and integration in order to solve a problem or overcome a common challenge [41-43]. Additionally, structural governance mechanisms (eg, whether they are time-bound or not, or whether they are periodic or run on an ad-hoc basis) also affect the type of relationship and the direction of knowledge flow. Furthermore, organizational proximity plays an important role in the formation of both formal and informal sharing. The extant literature highlights various aspects of organizational proximity, including cultural closeness [44-46], similarity in technological artifacts in use [47], alignment in strategic direction [48], and geographical proximity, and these facilitate smoother mobility of personnel, socialization, in-person exchange, and informal ties [36,49].

In terms of *organizational factors*, our study highlights the importance of knowledge search strategies by organizations, absorptive capacity, and past knowledge-sharing experiences in triggering and enabling sharing, as well as a wide spectrum of interorganizational knowledge-sharing modes. The extant literature has focused on the importance of absorptive capacity, which refers to the ability to recognize the value of knowledge being shared by both the donor and receiver organizations and the ability to utilize it [50]. The absorptive capacity is influenced by an organization's past experiences of identifying and building relationships [51], ability to integrate knowledge [52], and capability to retain knowledge [53]. We also found that as organizations gain experience in sharing and managing knowledge demands, they become more capable of managing and coordinating relationships as well as organizing the learning process. Organizations may use tools, such as IT-enabled systems, to eliminate barriers of communication and thereby heighten the level of knowledge sharing.

In relation to *individual and interpersonal factors*, we highlight the role of intrinsic and extrinsic motivations in encouraging knowledge sharing and the personal inhibitors that hinder the sharing process. Research has revealed a range of explanations for why individuals contribute to knowledge sharing in formal and informal settings. Many explanations point to intrinsic (self-interest) and extrinsic (altruistic) motivations [54]. Such motivations include self-benefits received from participation in knowledge-sharing communities and settings, such as access to expert advice [55], insights into others' beliefs and opinions, enhanced reputation and professional status [56,57], networking [58], and improved self-image and confidence [59]. Benefits to others include a desire for building or sustaining communities of interest [59,60], feelings of companionship [57], and reciprocity [61]. These motivations may vary in terms of nature and depth in formal and informal knowledge-sharing relationships. Bateman et al [62] highlighted that an individual's commitment to the organization they belong to, as well as the community, has a significant effect on participating in knowledge sharing. In this way, individuals may believe that they benefit from knowledge-sharing relationships [63] in both formal and informal settings, or they may feel obliged to pay back to existing relationships [56,62]. While the literature highlights the role of motivations in facilitating knowledge sharing, it also identifies factors that may hinder the transfer process. These include the time required for sharing activities [59] and the feeling of a strong psychological ownership over knowledge [63]. Emotional attachment to the organization may also discourage individuals from sharing certain knowledge, as it leads them to prioritize the organization's interests [64], influencing their decisions on what knowledge to protect and how to control its release [65].

Our study contributes to the literature on digital transformation and interorganizational learning by advancing the understanding of how learning ecosystems emerge, evolve, and take shape across multiple levels of the health care system. Rather than treating macroenvironmental, organizational, interorganizational, and individual factors as discrete categories, we demonstrate that they operate as dynamically interconnected influences that jointly shape the evolution of knowledge-sharing relationships. Through our multilevel analysis, we show how contextual conditions, such as national policy directives, program mandates, supplier strategies, and organizational digital maturity, activate

XSL•FO
RenderX

mechanisms, including identity formation, absorptive capacity, motivations, and relational recognition. These mechanisms, in turn, produce a spectrum of relationship types that vary in formality and origin, ranging from informal peer interactions to mandated partnerships.

By integrating these insights, our work conceptualizes the *learning ecosystem* as an evolving sociotechnical system in which relationships and knowledge flows continually adapt in response to shifting contextual pressures and emerging opportunities. We demonstrate that learning ecosystems do not develop through linear or centrally orchestrated processes; instead, they emerge through recursive interactions across levels, where changes in one domain, such as policy initiatives or new digital leadership roles, can reconfigure organizational strategies, reshape interorganizational networks, and influence individual patterns of knowledge sharing. The integrative model we propose brings these dynamics together by illustrating how multilevel factors (contexts and mechanisms) give rise to diverse modes of interorganizational collaboration (outcomes), and how triggers, mandates, controls, and enablers collectively shape the system's trajectory over time. This contribution extends existing frameworks on interorganizational knowledge sharing by offering a holistic explanation of how learning ecosystems are formed, how they function, and why they evolve differently across settings. In doing so, it provides theoretical clarity for understanding and supporting the development of interorganizational learning ecosystems capable of enabling digital transformation at scale.

Implications for Practice

This study offers important implications and insights that can be used in enabling and supporting the digital transformation process in the health care sector. Drawing on our findings, we have developed a structured set of questions designed to help policymakers and practitioners diagnose the conditions shaping knowledge sharing and monitor how these conditions evolve over time. These questions focus on the macroenvironmental, interorganizational, organizational, and individual factors that influence the emergence and functioning of learning ecosystems. We present these questions in Table 3. These questions serve as both a diagnostic aid (helping stakeholders assess what they are trying to achieve and the contextual factors shaping these efforts) and a monitoring aid (helping them evaluate how learning conditions and relationships develop over time).

To address macroenvironmental questions, policymakers and practitioners must examine public policies that facilitate or hinder the formation of formal and informal partnerships, as well as the role of external agencies in enabling knowledge sharing. Additionally, it is important to identify external factors, such as competition and cooperation. This requires assessing whether policymakers or other external entities demand the sharing of specific types of knowledge and whether industry characteristics, such as the levels of technological complexity and market maturity, affect partnership and community-building. To answer interorganizational questions, policymakers and practitioners should explore structural governance mechanisms such as community-based orchestration, contracts, defined processes, and routines. They must also examine relational and cognitive governance factors, including trust and power dynamics. Exploring organizational questions requires assessing the capabilities and experiences of donor and recipient organizations in forming and managing partnerships or participating in communities. This includes their ability to manage alliances, learn from and apply external knowledge, and facilitate knowledge sharing. Additionally, behavioral aspects, such as organizational motivations, incentives for knowledge sharing, and resistance to new knowledge, must be considered, along with organizational characteristics like the credibility of the donor and the receiver's flexibility to adapt to change. Finally, addressing individual-level questions requires examining both intrinsic and extrinsic factors influencing knowledge-sharing behavior.

XSL•FO
RenderX

Table . Diagnostic and monitoring questions for assessing multilevel factors influencing knowledge sharing.

Factor level	Questions
Macrocontext factors	• What formal external factors (eg, regulations, policies, technology, and market) affect the motivation to form formal partnerships for knowledge sharing? • What external factors affect the motivation to share knowledge without the existence of formal partnerships? • Is there an agency external to knowledge-sharing organizations that structures, incentivizes, and monitors knowledge sharing? What is the role of strategic mandates?
Interorganizational factors	• How do the interactive dynamics of the relationship (ie, power relations, structural governance and mechanisms, social ties and trust, and risk) influence knowledge sharing? • How does proximity (geographical, cultural, and technological) affect knowledge sharing? • What are the relationship-building factors (ie, collective identity, mutual goals, and joint rules) that affect knowledge sharing?
Organizational factors	• How do organizational characteristics affect knowledge sharing? • How do routines, processes, and coordination mechanisms affect knowledge sharing? • What is necessary for an organization to share/receive external knowledge?
Individual factors	• What motivates individuals to share knowledge? • What are the individual factors that prevent the sharing of knowledge?

Building on this diagnostic foundation, we offer a second set of questions that guide stakeholders in identifying and selecting appropriate modes and mechanisms of knowledge sharing for digital transformation initiatives. Rather than prescribing a fixed sequence of steps, these questions encourage reflection on the purpose of different knowledge-sharing modes, their temporality, the coordination techniques they require, and the tools that support them. These questions are presented in Table 4. Stakeholders can populate them with context-specific examples and options relevant to their initiatives. Table 4 includes 3 illustrative columns to demonstrate how different knowledge-sharing modes may be examined and compared.

Table . Reflective questions for selecting and assessing knowledge-sharing mechanisms in digital transformation initiatives.

Question	Option 1	Option 2	Option 3
What are the different options for formal/informal interorganizational knowledge sharing?	Blueprinting	Online user communities	Site visits
What purposes does it serve?	To share how a particular technology has been configured and implemented	To identify solutions to specific problems/needs; To co-innovate	To observe real-world technology use and challenges; To validate assumptions
What is the duration?	Project based	Continuous; Need basis	One-off
Who are the initiators of the modes of sharing?	Government	Individuals	Organizations; Technology suppliers
What are the management/coordination techniques?	Controlled by funding contacts	Orchestrated by community members	Managed by organizations on a need basis
What infrastructure and tools are used to facilitate knowledge sharing?	Documentation tool and database	Forums	In-person visits

Consistent with existing literature emphasizing the importance of relational maturity and temporality in knowledge-sharing processes [45,52,66,67], we propose that the questions in both tables should be revisited and iteratively assessed over time. Together, Tables 3 and 4 provide a structured yet flexible resource to support the design, implementation, and continual refinement of interorganizational learning within digital transformation programs.

By addressing these 2 sets of questions, this study provides policy and practice tools and guidelines to enhance and accelerate the digital transformation of the health care sector. First, it identifies where and how governments and other intermediaries can support interorganizational knowledge sharing. Second, it highlights the formal and informal modes and internal organizational and individual factors that

XSL•FO
RenderX

organizations can leverage as they navigate the transformation journey.

Limitations and Future Research

While this study advances the understanding of how learning ecosystems emerge and evolve across multiple levels, several limitations create opportunities for future research. First, our analysis focused predominantly on macroenvironmental and interorganizational factors, reflecting the design of the national digital transformation program and the nature of the available data. As a result, less attention was given to micro-level organizational routines and individual behaviors that may shape, or are shaped by, the learning ecosystem. Second, although our multicase design enabled rich comparative analysis across organizations, the scope of the study was limited primarily to hospital settings within a national health care system. Learning ecosystems may function differently across other care settings, sectors, or international contexts with different governance, market, or cultural dynamics. These limitations offer a valuable foundation for theoretical propositions that can guide future research aimed at advancing, refining, and testing the model developed in this study.

To guide such research, we propose a set of theoretically informed propositions. These propositions posit that cross-level contextual alignment is central to effective ecosystem formation; mechanisms, such as identity and relational recognition, significantly influence knowledge flows; the balance of formal and informal relationships affects ecosystem adaptability; and intermediaries are essential brokers in sustaining learning systems. Moreover, we propose that external influences shape early ecosystem trajectories, internal factors become more influential as ecosystems mature, learning outcomes vary with specific factors, and organizational- and individual-level factors warrant a deeper empirical investigation. Together, these propositions establish a clear agenda for advancing the theory on the emergence, evolution, and functioning of interorganizational learning ecosystems in digital transformation contexts.

Conclusion

Digital transformation has presented a significant challenge in the health care sector, which requires knowledge, experience, and expertise for organizational gains and new value creation. Interorganizational knowledge sharing can be seen as an enabler, supporter, and accelerator in assisting digital transformation at scale. Policymakers and organizations must proactively address the internal and external factors that impact knowledge sharing to harness its full potential in driving digital transformation. Understanding and evaluating approaches to creating a learning ecosystem is far from a trivial task and requires multilevel inquiries that follow actions within and beyond the organization in understanding this complex phenomenon.

Our framework illustrates how multilevel influencing factors shape a spectrum of interorganizational relationships and partnerships. The framework explores the dynamic nature of factors over time and therefore goes beyond shedding light on snapshots of the knowledge-sharing process by offering an evolutionary view of the process and its outcomes. By proposing this framework, we do not suggest that the domain of interorganizational knowledge sharing can be simply reduced to the aforementioned dimensions. Moreover, we do not propose this as a simple mechanistic account that could decouple different variables for the testing of modes of sharing and factors. Instead, we present a set of dimensions that are essential to explore in order to theorize a conceptual account of the learning ecosystem. We introduce this not just as an abstract framework but as a practical tool to support reflection and guide action for the various actors involved in shaping and sustaining it.

Acknowledgments

The authors gratefully acknowledge the input of the participants and the steering group for this evaluation. The views expressed are those of the authors and not necessarily those of the National Health Service (NHS), NHS England, NHSX, the Department of Health and Social Care, or NHS Digital. Generative artificial intelligence (ChatGPT by OpenAI) was used for proofreading the paper and streamlining some of the sentences.

Funding

This article has drawn on a program of independent research funded by the NHS (NHS England OJEU contract reference number 2017/S 114-229670; RW). This work was also supported by the National Institute for Health Research Imperial Patient Safety Translational Research Centre. The funders had no role in study design, data collection and analysis, the decision to publish, or the preparation of the manuscript.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors conceived the study. HM drafted the manuscript, and all authors commented on the drafts of the manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Mergel I, Edelmann N, Haug N. Defining digital transformation: results from expert interviews. Gov Inf Q 2019 Oct;36(4):101385. [doi: 10.1016/j.giq.2019.06.002]
2. Carter L, Desouza KC, Dawson GS, Pardo T. Digital transformation of the public sector: designing strategic information systems. J Strateg Inf Syst 2024 Sep;33(3):101853. [doi: 10.1016/j.jsis.2024.101853]
3. Gopal G, Suter-Crazzolara C, Toldo L, Eberhardt W. Digital transformation in healthcare – architectures of present and future information technologies. Clin Chem Lab Med 2019 Feb 25;57(3):328-335. [doi: 10.1515/cclm-2018-0658] [Medline: 30530878]
4. Walsh MN, Rumsfeld JS. Leading the digital transformation of healthcare: the ACC innovation strategy. J Am Coll Cardiol 2017 Nov 28;70(21):2719-2722. [doi: 10.1016/j.jacc.2017.10.020] [Medline: 29169479]
5. Gupta S, Modgil S, Lopes de Sousa Jabbour AB, Laguir I, Stekelorum R. Towards digital transformation and governance in the healthcare sector. Inform Technol People 2025 Nov 3;38(5):2163-2186. [doi: 10.1108/ITP-02-2023-0179]
6. Koukouvinou P, Simbi N, Holmström J. Managing unbounded digital transformation: exploring the role of tensions in a digital transformation initiative in the forestry industry. Inform Technol People 2023 Dec 18;36(8):43-68. [doi: 10.1108/ITP-03-2020-0106]
7. Burton-Jones A, Akhlaghpour S, Ayre S, Barde P, Staib A, Sullivan C. Changing the conversation on evaluating digital transformation in healthcare: insights from an institutional analysis. Information and Organization 2020 Mar;30(1):100255. [doi: 10.1016/j.infoandorg.2019.100255]
8. Agarwal R, Gao G, DesRoches C, Jha AK, Gao G. Research commentary—the digital transformation of healthcare: current status and the road ahead. Inf Syst Res 2010 Dec;21(4):796-809. [doi: 10.1287/isre.1100.0327]
9. Matt C, Hess T, Benlian A. Digital transformation strategies. Bus Inf Syst Eng 2015 Oct;57(5):339-343. [doi: 10.1007/s12599-015-0401-5]
10. Hess T, Matt C, Benlian A, Wiesböck F. Options for formulating a digital transformation strategy. In: Galliers RD, Leidner DE, Simeonova B, editors. Strategic Information Management: Routledge; 2020:151-173. [doi: 10.4324/9780429286797-7]
11. Goraya MAS, Yaqub MZ, Khan MA, Akram MS, Alofaysan H. Transforming performance: how agility, response, resilience and support shape success in digital strategies. Inform Technol People 2026 Jan 14;39(1):325-353. [doi: 10.1108/ITP-05-2024-0592]
12. Pittaway JJ, Montazemi AR. Know-how to lead digital transformation: the case of local governments. Gov Inf Q 2020 Oct;37(4):101474. [doi: 10.1016/j.giq.2020.101474]
13. Greenhalgh T, Papoutsi C. Studying complexity in health services research: desperately seeking an overdue paradigm shift. BMC Med 2018 Jun 20;16(1):95. [doi: 10.1186/s12916-018-1089-4] [Medline: 29921272]
14. Goodwin N, Smith J, Davies A, et al. Integrated care for patients and populations: improving outcomes by working together. : King's Fund; 2012 URL: https://assets.kingsfund.org.uk/f/256914/x/2ff65ea41a/integrated_care_nhs_future_forum_january_2012.pdf [accessed 2026-02-26]
15. Waldman SA, Terzic A. Health care evolves from reactive to proactive. Clin Pharmacol Ther 2019 Jan;105(1):10-13. [doi: 10.1002/cpt.1295] [Medline: 30597564]
16. Gil-Garcia JR, Sayogo DS. Government inter-organizational information sharing initiatives: understanding the main determinants of success. Gov Inf Q 2016 Jul;33(3):572-582. [doi: 10.1016/j.giq.2016.01.006]
17. Cresswell K, Sheikh A, Franklin BD, et al. Interorganizational knowledge sharing to establish digital health learning ecosystems: qualitative evaluation of a national digital health transformation program in England. J Med Internet Res 2021 Aug 19;23(8):e23372. [doi: 10.2196/23372] [Medline: 34420927]
18. Williams R, Sheikh A, Franklin BD, et al. Using blueprints to promote interorganizational knowledge transfer in digital health initiatives-a qualitative exploration of a national change program in English hospitals. J Am Med Inform Assoc 2021 Jul 14;28(7):1431-1439. [doi: 10.1093/jamia/ocab020] [Medline: 33706378]
19. Cresswell K, Sullivan C, Theal J, Mozaffar H, Williams R. Concerted adoption as an emerging strategy for digital transformation of healthcare-lessons from Australia, Canada, and England. J Am Med Inform Assoc 2024 Apr 19;31(5):1211-1215. [doi: 10.1093/jamia/ocae034] [Medline: 38400737]
20. Turulja L, Cinjarevic M, Veselinovic L. Information technology and knowledge sharing for better health care: an emerging economy context. J Knowl Manag 2021 May 3;25(3):559-572. [doi: 10.1108/JKM-09-2019-0514]
21. Mozaffar H, Candi M. Extending the process frontier of digital transformation: a flow‐oriented perspective. Inf Syst J 2025 Mar;35(2):720-760. [doi: 10.1111/isj.12557]
22. Walsham G. The emergence of interpretivism in IS research. Inf Syst Res 1995 Dec;6(4):376-394. [doi: 10.1287/isre.6.4.376]
23. Locke K. The grounded theory approach to qualitative research. In: Drasgow F, Schmitt N, editors. Measuring and Analyzing Behavior in Organizations: Advances in Measurement and Data Analysis: Jossey-Bass/Wiley; 2002:17-43.
24. Gioia DA, Corley KG, Hamilton AL. Seeking qualitative rigor in inductive research: notes on the Gioia methodology. Organ Res Methods 2013;16(1):15-31. [doi: 10.1177/1094428112452151]
25. Strauss A, Corbin JM. Basics of Qualitative Research: Grounded Theory Procedures and Techniques: Sage Publications; 1990.

XSL•FO
RenderX

26. Rindova V, Dalpiaz E, Ravasi D. A cultural quest: a study of organizational use of new cultural resources in strategy formation. Organ Sci 2011 Apr;22(2):413-431. [doi: 10.1287/orsc.1100.0537]
27. Pawson R, Tilley N. Realist Evaluation: SAGE Publications; 2004. URL: https://www.copasah.org/uploads/1/2/6/4/12642634/realistic_evaluationpawson.pdf [accessed 2026-03-03]
28. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
29. Guennif S, Ramani SV. Explaining divergence in catching-up in pharma between India and Brazil using the NSI framework. Res Policy 2012 Mar;41(2):430-441. [doi: 10.1016/j.respol.2011.09.005]
30. Hagedoorn J, Lorenz-Orlean S, van Kranenburg H. Inter-firm technology transfer: partnership-embedded licensing or standard licensing agreements? Ind Corp Change 2009 Jun 1;18(3):529-550. [doi: 10.1093/icc/dtn038]
31. Mozaffar H, Williams R, Cresswell K, et al. Product diversity and spectrum of choice in hospital ePrescribing systems in England. PLoS One 2014;9(4):e92516. [doi: 10.1371/journal.pone.0092516] [Medline: 24691517]
32. Grimpe C, Sofka W. Complementarities in the search for innovation—managing markets and relationships. Res Policy 2016 Dec;45(10):2036-2053. [doi: 10.1016/j.respol.2016.07.007]
33. Bozeman B, Rimes H, Youtie J. The evolving state-of-the-art in technology transfer research: revisiting the contingent effectiveness model. Res Policy 2015 Feb;44(1):34-49. [doi: 10.1016/j.respol.2014.06.008]
34. Landry R, Amara N, Cloutier JS, Halilem N. Technology transfer organizations: services and business models. Technovation 2013 Dec;33(12):431-449. [doi: 10.1016/j.technovation.2013.09.008]
35. Lichtenthaler U, Ernst H. External technology commercialization in large firms: results of a quantitative benchmarking study. R&D Manag 2007 Nov;37(5):383-397. [doi: 10.1111/j.1467-9310.2007.00487.x]
36. Battistella C, De Toni AF, Pillon R. Inter-organisational technology/knowledge transfer: a framework from critical literature review. J Technol Transf 2016 Oct;41(5):1195-1234. [doi: 10.1007/s10961-015-9418-7]
37. Hinder S, Cresswell K, Sheikh A, et al. Promoting inter-organisational knowledge sharing: a qualitative evaluation of England's Global Digital Exemplar and Fast Follower Programme. PLoS One 2021;16(8):e0255220. [doi: 10.1371/journal.pone.0255220] [Medline: 34339429]
38. Milagres R, Burcharth A. Knowledge transfer in interorganizational partnerships: what do we know? Bus Process Manag J 2019 Feb 8;25(1):27-68. [doi: 10.1108/BPMJ-06-2017-0175]
39. Segrestin B. Partnering to explore: the Renault–Nissan Alliance as a forerunner of new cooperative patterns. Res Policy 2005 Jun;34(5):657-672. [doi: 10.1016/j.respol.2005.02.006]
40. Mozaffar H. User communities as multi-functional spaces: innovation, collective voice, demand articulation, peer informing and professional identity (and more). In: Hyysalo S, Jensen TE, Oudshoorn N, editors. The New Production of Users: Routledge; 2016:219-246 URL: https://www.taylorfrancis.com/chapters/edit/10.4324/9781315648088-12/user-communities-multi-functional-spaces-hajar-mozaffar [accessed 2026-02-26]
41. Mesquita LF, Anand J, Brush TH. Comparing the resource‐based and relational views: knowledge transfer and spillover in vertical alliances. Strat Manag J 2008 Sep;29(9):913-941. [doi: 10.1002/smj.699]
42. Johnson M, Mozaffar H, Campagnolo GM, Hyysalo S, Pollock N, Williams R. The managed prosumer: evolving knowledge strategies in the design of information infrastructures. Inf Commun Soc 2014 Aug 9;17(7):795-813. [doi: 10.1080/1369118X.2013.830635]
43. Cheung CMK, Chiu PY, Lee MKO. Online social networks: why do students use facebook? Comput Human Behav 2011 Jul;27(4):1337-1343. [doi: 10.1016/j.chb.2010.07.028]
44. Inkpen AC, Tsang EWK. Social capital, networks, and knowledge transfer. Acad Manag Rev 2005 Jan;30(1):146-165. [doi: 10.5465/amr.2005.15281445]
45. Kotabe M, Martin X, Domoto H. Gaining from vertical partnerships: knowledge transfer, relationship duration, and supplier performance improvement in the U.S. and Japanese automotive industries. SMJ 2003 Apr;24(4):293-316. [doi: 10.1002/smj.297]
46. Bhagat RS, Kedia BL, Harveston PD, Triandis HC. Cultural variations in the cross-border transfer of organizational knowledge: an integrative framework. Acad Manag Rev 2002 Apr;27(2):204. [doi: 10.2307/4134352]
47. Ireland RD, Hitt MA, Vaidyanath D. Alliance management as a source of competitive advantage. J Manage 2002 Jun;28(3):413-446. [doi: 10.1177/014920630202800308]
48. Darr ED, Kurtzberg TR. An investigation of partner similarity dimensions on knowledge transfer. Organ Behav Hum Decis Process 2000 May;82(1):28-44. [doi: 10.1006/obhd.2000.2885]
49. Berchicci L, de Jong JPJ, Freel M. Remote collaboration and innovative performance: the moderating role of R&D intensity. Ind Corp Change 2016 Jun;25(3):429-446. [doi: 10.1093/icc/dtv031]
50. Cohen WM, Levinthal DA. Absorptive capacity: a new perspective on learning and innovation. Adm Sci Q 1990 Mar;35(1):128-152. [doi: 10.2307/2393553]
51. Love JH, Roper S, Vahter P. Learning from openness: the dynamics of breadth in external innovation linkages. SMJ 2014 Nov;35(11):1703-1716. [doi: 10.1002/smj.2170]

XSL•FO
RenderX

52. Tzabbar D, Aharonson BS, Amburgey TL. When does tapping external sources of knowledge result in knowledge integration? Res Policy 2013 Mar;42(2):481-494. [doi: 10.1016/j.respol.2012.07.007]
53. Nooshinfard F, Nemati-Anaraki L. Success factors of inter-organizational knowledge sharing: a proposed framework. Electron Libr 2014 Apr 1;32(2):239-261. [doi: 10.1108/EL-02-2012-0023]
54. Hung SY, Durcikova A, Lai HM, Lin WM. The influence of intrinsic and extrinsic motivation on individuals' knowledge sharing behavior. Int J Hum Comput Stud 2011 Jun;69(6):415-427. [doi: 10.1016/j.ijhcs.2011.02.004]
55. Lampel J, Bhalla A. The role of status seeking in online communities: giving the gift of experience. J Comput Mediat Commun 2007 Jan;12(2):434-455. [doi: 10.1111/j.1083-6101.2007.00332.x]
56. Lakhani KR, von Hippel E. How open source software works: "free" user-to-user assistance. Res Policy 2003 Jun;32(6):923-943. [doi: 10.1016/S0048-7333(02)00095-1]
57. Hall H, Graham D. Creation and recreation: motivating collaboration to generate knowledge capital in online communities. Int J Inf Manage 2004 Jun;24(3):235-246. [doi: 10.1016/j.ijinfomgt.2004.02.004]
58. Christensen CM, Raynor ME. Why hard-nosed executives should care about management theory. Harv Bus Rev 2003 Sep;81(9):66-74. [Medline: 12964394]
59. McLure Wasko M, Faraj S. "It is what one does": why people participate and help others in electronic communities of practice. J Strateg Inf Syst 2000 Sep;9(2-3):155-173. [doi: 10.1016/S0963-8687(00)00045-7]
60. Blanchard AL, Markus ML. The experienced "sense" of a virtual community. Data Base Adv Info Sys 2004 Feb 3;35(1):64-79. [doi: 10.1145/968464.968470]
61. Wasko MM, Faraj S. Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. MIS Q 2005 Mar 1;29(1):35-57. [doi: 10.2307/25148667]
62. Bateman PJ, Gray PH, Butler BS. Research note—the impact of community commitment on participation in online communities. Inf Syst Res 2011 Dec;22(4):841-854. [doi: 10.1287/isre.1090.0265]
63. Peng H. Why and when do people hide knowledge? J Knowl Manag 2013 May 24;17(3):398-415. [doi: 10.1108/JKM-12-2012-0380]
64. Meyer JP, Herscovitch L. Commitment in the workplace: toward a general model. Hum Resour Manag Rev 2001 Sep;11(3):299-326. [doi: 10.1016/S1053-4822(00)00053-X]
65. Jarvenpaa SL, Majchrzak A. Interactive self-regulatory theory for sharing and protecting in interorganizational collaborations. Acad Manag Rev 2016 Jan;41(1):9-27. [doi: 10.5465/amr.2012.0005]
66. Inkpen AC, Currall SC. The coevolution of trust, control, and learning in joint ventures. Organ Sci 2004 Oct;15(5):586-599. [doi: 10.1287/orsc.1040.0079]
67. Mozaffar H, Panteli N. The online community knowledge flows: distance and direction. Eur J Inf Syst 2022;31(2):227-240. [doi: 10.1080/0960085X.2020.1866442]

Abbreviations

CMO: context-mechanism-outcome
COREQ: Consolidated Criteria for Reporting Qualitative Research
FF: fast follower
GDE: Global Digital Exemplar
NHS: National Health Service

Edited by A Stone; submitted 01.Sep.2025; peer-reviewed by L Desveaux, M Dohan; revised version received 10.Feb.2026; accepted 10.Feb.2026; published 12.Mar.2026.

Please cite as:
Mozaffar H, Williams R, Cresswell K
Analysis of Multilevel Factors Mobilizing the Spectrum of Interorganizational Knowledge Sharing for Facilitating Digital Transformation at Scale: Qualitative Study
J Med Internet Res 2026;28:e83345
URL: https://www.jmir.org/2026/1/e83345
doi:10.2196/83345

© Hajar Mozaffar, Robin Williams, Kathrin Cresswell. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly

XSL•FO
RenderX

cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Quality Challenges in Municipal Telecare Call Center Services: Qualitative Evaluation Using the Anchored, Realistic, Cocreated, Human, Integrated, and Evaluated (ARCHIE) Framework

Linda C Grøndal-Eeles[*]; Janne Dugstad[*], PhD; Hilde Eide[*], PhD; Etty Nilsen[*], PhD

Centre for Health and Technology, Faculty of Health and Social Science, University of South-Eastern Norway, Postboks 4, Borre, Norway

[*]all authors contributed equally

Corresponding Author:
Linda C Grøndal-Eeles
Centre for Health and Technology, Faculty of Health and Social Science, University of South-Eastern Norway, Postboks 4, Borre, Norway

Abstract

Background: Telecare is seen as a promising technology aimed at enhancing the accessibility and efficiency of health care services. Although focus on quality has been highly prioritized within the health care services, there is a need to explore the quality of telecare services in general and municipal telecare call centers (CCs) in particular, as health and assistive technologies are increasingly being implemented in patients' homes.

Objective: The study sought to explore which factors influence the quality of telecare services provided by municipal telecare CCs in Norway, evaluated through the anchored, realistic, cocreated, human, integrated, and evaluated (ARCHIE) framework.

Methods: The study had a multiple-case design. Interviews were the main source of data from 15 informants from 5 municipal telecare CCs across Norway. Observation and document studies were used for background and contextualization. To explore and evaluate quality, a combined deductive–inductive analysis was conducted.

Results: Evaluated against the ARCHIE framework, none of the quality criteria were fully met. Due to the telecare service not being sufficiently anchored for all patients, it was challenging to provide realistic technologies. The collaborative work was difficult, with challenges in recruiting patients. The human principle was characterized by variation of knowledge and national guidelines. Municipal telecare CCs were not integrated into the health care services, and data must be used to a greater extent for evaluation and learning than is currently the case.

Conclusions: The findings suggest that municipal telecare CC services have several shortcomings in providing high-quality health care. Relating the quality principles identified by the ARCHIE framework to normalization process theory constructs indicates that the CC service remains in a transitional phase of normalization. To improve the telecare CC services and enhance communication and integration, policymakers need to reduce fragmentation in the broader health care system. Further national standardization to professionalize the telecare CC services should be developed. The telecare CCs need to improve their service related to all indicators of the ARCHIE framework. Training for telecare operators should be prioritized.

(J Med Internet Res 2026;28:e76054) doi:10.2196/76054

KEYWORDS

quality; municipalities; telecare services; telecare call center; assistive technology; patient safety; telecare; call center

Introduction

Telecare and digital technologies have been partly integrated into the health care services of Western countries during the past decade. Telecare can be defined as: "the use of information, communication, and monitoring technologies which allows health care providers to remotely evaluate health status, give educational intervention, or deliver health and social care to patients in their homes" [1]. A call center (CC) is a 24/7 service that operates telecare. A CC employs operators, typically health care workers or technicians, to answer, document, and assess calls and alarms from assistive technology [2]. Assistive technology is an umbrella term for assistive products and their related systems and services, which help maintain or improve an individual's functioning related to cognition, communication, hearing, mobility, self-care, and vision [3,4]. These technologies are implemented in people's homes and are the reason for the CCs' existence. CCs organize their service in a variety of ways, including their core service, which is to answer, evaluate, and document assistive alarms and take any necessary further action [2]. They also provide additional services, including ambulant teams, advisory departments, or monitoring technologies that are not health-related. The CCs may be private entities or part of the public service. In Norway, the CCs are largely organized

XSL•FO
RenderX

as intermunicipal collaborations that are hosted by one municipality and regulated by the Local Government Act [5]. Our recent article investigating the organization of municipal telecare CCs in Norway found that the centers were in an early development phase and had different organizational models, leading to a variation of delivered services [6]. As CCs are considered new and innovative within the health care service and provide the service in a new way, there is a need to investigate the quality of their services.

The World Health Organization (WHO) defines quality as "the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge" [7]. To achieve high quality within health care services, several factors, such as patient safety, effectiveness, availability, and patient-centered care, need to be evaluated [8], and many governments have strengthened their focus on quality within the health care system [9-12]. International studies have shown a multitude of adverse events happening within the health care system each year, which have an effect on quality [13-16].

Several frameworks have been developed for the evaluation of the quality of health care services. One of the earliest and most used is the Donabedian framework [17], which evaluates the systems, processes, and outcomes of health care delivery. The framework has also been used for evaluating quality in digital services such as telecare [18]. Tossaint-Schoenmakers et al's [18] review highlights 3 broad aspects that are important for integration: the needs of the care recipient, technology that is attuned to the organization and care processes, and human resources that are aligned with the desired end results. A need for more in-depth studies into the specific processes in different parts of the health care system related to eHealth is suggested.

Quality in telecare has increasingly become a relevant topic, as more assistive technology is being implemented. The aim of implementing technology is to improve access to the health care system and improve effectiveness [19]. The use of telecare can improve communication between health care professionals as well as improve patient safety by reducing errors [20,21]. Furthermore, digital health care can give patients the opportunity to participate in their own treatment and treatment plans to a higher degree than before [19]. Challenges and barriers related to the use of telecare include technical barriers, such as low internet access, poor digital competence among certain user groups, and concerns about privacy and data security [22]. In addition, Woolham et al [23] mention cost-effectiveness, assessment and deployment, ethical implications, and user involvement as key challenges.

As there is a need to explore in greater depth how quality is performed [18], Greenhalgh et al's [24] framework, called the anchored, realistic, cocreated, human, integrated, and evaluated (ARCHIE) framework, seems to be a relevant framework to evaluate important quality aspects of telecare. The framework determines whether a service is anchored, realistic, co-creative, human, integrated, and evaluated, constituting the acronym ARCHIE. The framework was based on interviews and observations in various care contexts to identify barriers and suggest several solutions to provide quality care with useful assistive technology. An important takeaway from Greenhalgh et al's [24] study is that technological advances should be underpinned by a user-centered approach to design and delivery. The care contracts and technology should be personalized for each patient, from branded products to interoperable components. Furthermore, the technology should have the ability to be used flexibly across devices and platforms [24].

The benefits of telecare are still not well-established [25,26]. If the service lacks quality, this can lead to an increased burden of illness, increased economic costs, and medical errors such as inaccurate diagnoses, medication errors, and inappropriate treatments [27,28]. If clinical guidelines are not followed, this can result in ineffective care. By increasing quality within the telecare service, the benefits of the implemented technology might be easier to recognize.

Patient safety in primary care is a critical aspect of health care. According to the WHO, patient safety involves the prevention of errors and adverse effects associated with health care, aiming to "do no harm" to patients [29]. Research indicates that incidents in primary care, although often less severe than those in hospitals, significantly impact overall patient safety due to the high volume of patient interactions [14]. Furthermore, it is important to take into consideration that patients will increasingly receive complicated treatment of complex illnesses at home. As previous research has highlighted, the assessment of quality and safety culture is a fundamental step in understanding health care providers' safety-related perceptions [14,30]. Therefore, fostering a robust safety culture and implementing targeted interventions for quality are paramount in primary care.

Telecare has the ability to enhance the accessibility and efficiency of health care services [31]. However, a systematic review identified multiple patient safety risks associated with telecare, including issues related to the nature of homecare tasks, person-centered characteristics, technology and devices, organization and environmental factors [32]. These risks can arise from the complexity of integrating technology into home care settings, where both the physical environment and the patient's capabilities play crucial roles. A qualitative study highlighted the importance of care workers' roles in ensuring patient safety through telecare [33]. The study identified 3 essential work processes: aligning people and technologies, being alert and staying calm, and coordinating activities based on people and technology [33]. These processes underscore the need for effective training and support for care workers to manage the interface between technology and patient care effectively.

Few studies have investigated telecare providers' perspectives on how quality is performed in municipal telecare CC practice. Therefore, the aim of the study was to explore which factors influence the quality of telecare services provided by municipal telecare CCs in Norway, evaluated through the ARCHIE framework.

XSL•FO
RenderX

Methods

Study Design

This study had a multiple-case design. A multiple case study is regarded by many as more robust than the single-case design [34-37]. In this study, similar cases have been studied in order to gain a broad understanding of quality, and not to compare the individual cases. The various cases form a bricolage [38] that embraces the complexity of the context. Interviews were the main source of data. Data from observations and documents have been used to enhance the understanding of the background and context of the CCs.

Recruitment and Data Collection

Data were collected in May and June of 2022. At the time of data collection, no overview of Norwegian telecare CCs existed, so a snowball-inspired strategy was used in our networks to identify centers. An email was sent to the managers of 5 CCs, and they agreed to participate with their leaders and personnel.

A semistructured interview guide was developed for the in-depth interviews. The study's aim was used to develop themes for the interview guide where the following themes were covered: (1) The participants' understanding of the municipal telecare CC service in their context, (2) How the municipal telecare CC services operated, (3) Cooperation with other health care services, (4) Ethical challenges, (5) Competence requirements and competence development, and (6) Perspectives on the further development of the municipal telecare CC services, to shed light on the aim. The idea was to let the informants speak as freely as possible within the themes. One paper has previously been published based on the same in-depth interviews, focusing on the organization of the services [11]. For this paper, the interviews were analyzed focusing on quality.

The first author, LGE, visited 4 of the CCs, with each visit lasting for one day, including interviews, a tour, and observations. On arrival at the CCs, a guided tour was conducted. The observations included how the CC operators worked, how they responded to alarms, the layout of the facilities, the workstation setup, and some of the technology included in the services (eg, social alarms, GPS, medicine dispensers, door locks, and sensors). These observations provided important background knowledge for understanding the service and supporting the interviews. Notes were taken and used to understand the context.

All the interviews were performed by LGE following the semistructured interview guide, which constituted the main source of data for the studies. As the interview guide contained broad, open questions, the informants could speak as freely as they wanted. Four of the 5 CCs were visited, while one CC had to be interviewed digitally due to the COVID-19 pandemic restrictions. Ten informants were interviewed face-to-face. Two of these were conducted with 2 informants present, and the rest were conducted one-on-one. Five informants were interviewed digitally—one with 3 informants present, and 2 were conducted one-on-one. The informants were present at the CCs during the interviews, and this was also the case during the digital interviews.

The interviews lasted from 45 minutes to 2 hours and were audio-recorded and transcribed.

Data Analysis

The data analysis had a combined deductive–inductive approach, analyzing the CCs as the unit of analysis. First, text segments were identified and sorted in accordance with the specific principles in the ARCHIE framework in a deductive fashion. Second, the text under each principle was analyzed inductively.

Deductive Analysis Using the ARCHIE Framework

The ARCHIE framework was designed to be adaptable and to guide the development and evaluation of telehealth and telecare services in various contexts [24]. The framework aims to operationalize quality in telehealth and telecare in order to improve the proportion of patients receiving appropriate, acceptable, and workable technologies and services and has 6 principles (Table 1).

Table . Overview of the ARCHIE (Anchored, Realistic, Cocreated, Human, Integrated, and Evaluated) framework [24].

Principle	Content
Anchored	Having a shared understanding of what matters to the patient, which includes design and development of technology
Realistic	Understanding the natural history of illness. The illness may progress, and the patients' needs change
Co-creative	Evolving and adapting solutions with patients. This should be continuous rather than standardized
Human	Supporting patients through interpersonal relationships and social networks where knowledge and information are key
Integrated	Developing mutual awareness and sharing knowledge
Evaluated	Ensuring system learning from experiences built over years. Ongoing innovation to increase quality

- Step 1: The first author (LGE) transcribed the 11 interviews and initially read through all of them, identifying and coding text segments that aligned with the 6 main ARCHIE principles described above. Some text segments aligned with the ideal quality principle, but others were inconsistent; these inconsistent statements were coded under the specific principle as "not" (Table 2 contains examples). Consequently, all the coded text had a main code in

XSL•FO
RenderX

accordance with ARCHIE, and if relevant, a subcode as "not."

Table . Examples of coding that align with the ARCHIE (Anchored, Realistic, Cocreated, Human, Integrated, and Evaluated) principles and which are inconsistent with the ARCHIE principles.

Principle	Example
Anchored	A shared understanding of the needs of the patient
Not anchored	A patient is given the wrong technology
Realistic	Understanding the natural history of illness
Not realistic	Change of pattern for the use of assistive technology, but the patient still has the technology

- Step 2: All authors read and coded 3 selected interviews to align the understanding of the ARCHIE framework principles and secure the initial coding in each principle. The tentative coding by the first author in Step 1 was discussed, adjusted, and agreed on.
- Step 3: LGE reread and checked the coding of all the interviews.

Inductive Analysis

After the deductive analysis, a second phase followed using inductive content analysis within each ARCHIE principle. The inductive analysis followed the structured process described by Elo and Kyngäs [39], with 3 main steps: preparation, organization, and reporting.

- Step 4: In the preparation phase, the text within each ARCHIE principle was re-examined by LGE to identify meaning relevant to the principle's theme. These units were extracted and organized for a deeper analysis.
- Step 5: During the organization phase, the author group's focus was on identifying patterns, contrasts, and recurring themes within each principle, rather than generating new codes. For example, in the anchored principle, both aligned and misaligned statements were examined to understand how services reflected, or failed to reflect, what mattered to the patient. This was discussed in the author group to ensure understanding of each principle.
- Step 6: The reporting phase synthesized insights from each principle to highlight trends and tensions. These were then interpreted in relation to the ARCHIE framework to assess whether and how quality was met across the principles. NVivo (version 5.0; Lumivero) was used in the coding processes.

Rigor and Trustworthiness

To ensure rigor and trustworthiness in the study, criteria developed by Guba and Lincoln [40] were followed, including credibility, transferability, dependability, and confirmability. To ensure the credibility of the coding, analysis, and results were discussed in several rounds, in the deductive as well as the inductive phase of the study, to ensure the group of authors could express their thoughts, concerns, or other factors. One issue that the author group discussed in several rounds was how to interpret the content of the principles in ARCHIE, and which statements belonged where. To understand the principles, the author group went through each statement, forming a consensus on whether or not statements were placed correctly in our understanding of what each principle should contain.

The study provided detailed descriptions of the sample to ensure transferability and support contextual understanding. The coding process was well-documented and logged in NVivo, as well as revisions to maintain dependability. In any research, the expulsion of all biases is challenging. However, the research group consisted of authors with interdisciplinary backgrounds, such as nursing (LGE and HE), economics and management (EN), optometry (JD), psychology, and eHealth communication and counseling (HE). The group also had broad competence in research methodology as well as health technology implementation. This meant that the authors could share different views on the research and ask critical questions to make sure the author group's assumptions and interpretations were as transparent as possible, thereby ensuring confirmability. Furthermore, the coding process was repetitive and reflexive, guided by Saldaña's [41] approach, and involved multiple rounds of discussions within the author group to reach consensus. Analytical transparency was maintained through NVivo documentation.

Ethical Considerations

The study was approved by the Norwegian Agency for Shared Services in Education and Research, which assessed the data processing plan to be in accordance with data protection legislation (Ethics approval number: 426173). Furthermore, the main principles for research according to research ethics in the social sciences and the humanities [42] were followed. All informants received written and oral information about the study, and informed consent to participate was obtained from each informant. The informants participated voluntarily and received no compensation of any sort. The CC informants appreciated that researchers took an interest in the telecare topic and therefore welcomed the study. Data were stored and secured in a USN-protected area. The data in the quotes used throughout this article were anonymized.

Results

Overview

The first section of the results will present the sample of the study, and key dimensions of quality according to the ARCHIE framework will then be presented. The headings represent the main findings in each principle.

Sample

Data were drawn from 5 municipal telecare CCs in different parts of Norway (Table 3). Eleven interviews were performed

XSL•FO
RenderX

with 15 professionals working in the 5 CCs (CC1-CC5). Five informants from CC1 were interviewed, 1 from CC2, 4 from CC3, 2 from CC4, and 3 from CC5. In total, 12 women and 3 men were included. The informants had a variety of professional backgrounds and roles within the centers: 6 were managers, 4 were advisors, 4 were health care operators, and 1 was an IT technician.

Table . Overview of call centers (CCs) participating in the study.

	CC1	CC2	CC3	CC4	CC5
Number of municipalities in cooperation	14	16	16	6	12
Number of citizens in the cooperating municipalities	Estimated 140,000	Estimated 175,000 (Estimated by authors)[a]	Estimated 306,000 (Estimated by authors)[a]	Estimated 51,000	Estimated 85,000
Established	2019	2003 serving fire and rescue. Included in the health care service from 2019	2018	2019	2016 under fire and rescue service, included in health care 2023, colocated with municipal accident and emergency department.
Organization (departments or services provided alongside telecare CC)	Colocated with municipal accident and emergency department	Monitoring technical alarms in public buildings (fire alarms, lift alarms, etc)	Ambulatory team department. (Can visit patients whose alarms cannot be resolved by phone)	Service design advisory department. (Can assist municipalities in cooperation with implementation strategy and other eHealth challenges	Monitoring technical alarms in public buildings (fire alarms, lift alarms, etc)

[a]Estimated by authors.

The Unanchored Services of Call Centers

The anchored principle was meant to reflect a shared understanding of the patient's needs between the patient and all the stakeholders involved in the provision of the service, emphasizing what is important and meaningful to the patient. However, the allocation of assistive technology did not appear to be anchored with all the stakeholders working with each patient. The telecare providers working in the CCs were not involved in the process of assessing the patients' needs or prioritizing and selecting what services they should receive. Informant K said:

> *The assumption must be that the home nursing staff, in collaboration with other decision-makers in the municipalities, should work much closer with the patient to see if the technology can actually relieve the burden. It is felt that this is not being done.*

The municipal telecare CCs provided services across several municipalities, which otherwise worked independently of each other. In order to establish and provide the service for a patient over time, the telecare services included several stakeholders within the health care services of each municipality, who interacted with the patient during the course of the service. Initially, the home nursing unit and the service allocation office were involved in the referral and mapping of the patients and the assignment of the service. In addition, general practitioners, family members, and even the patients themselves could identify the need for a service in the early phases. When a patient was allocated a telecare service, the health technology support team was responsible for installing the technological devices in the patient's home and instructing the patient in the use of the technology. The CC was then notified. Their responsibility was to respond to alarms and calls and to monitor the technology and incoming data. In their daily operations, the CC alerted the home nursing staff if physical interaction with the patients was needed. In order to change the service for a patient, the CCs or home nursing staff reported their concern to the service allocation office. A suspicion that the service was perhaps no longer what the patient needed would prompt an evaluation and the allocation of new services. Figure 1 displays a visualization of the process with an overview of the relevant stakeholders involved in the different phases of assigning and providing a telecare service.

XSL•FO
RenderX

Figure 1. Overview of the actions related to the telecare call center service and stakeholders within primary health telecare.

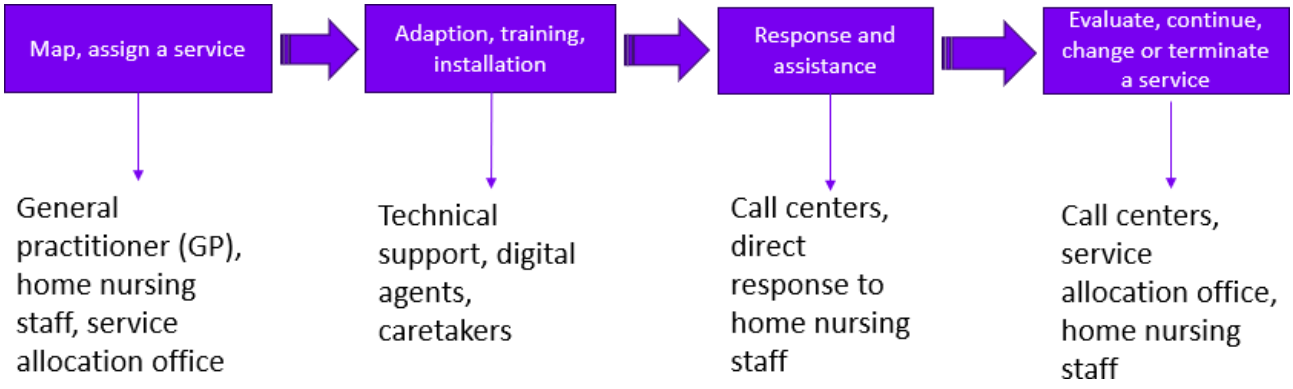

The Challenge of Having Realistic Technology When the Needs Are Constantly Changing

The realistic principle meant that technology and care packages needed to allow for the patients' conditions to progress or change, thereby supporting changes in the patients' needs. However, if the technology was not anchored in the first place, it could never be realistic. The informants addressed how patients received the technology too late, where the situation of the patients was rapidly changing, and the technology was therefore not realistic. Informant K explained:

> *We feel like the introduction to health technology should happen at a much earlier stage than what is being done now.*

Another challenge was how some progressive illnesses left the patient in a state where they did not know, understand, or remember the different assistive technologies, indicating that their health situation had changed, so they were no longer capable of using the technology as intended. When a patient reached this stage, there was no benefit for either the patient or the health care service, because the assistive technology was a poor fit. As informant F stated:

> *When a patient has moderate Alzheimer's, then it is too late. They will wonder what it is and maybe even take it off and not remember to use the alarm when needed.*

The CCs could pick up on the progression of a patient's health conditions by monitoring patterns in their behavior, such as a sudden increase in calls from a patient via their social alarm. The CC reported these changes to the home nursing staff or service allocation office. Informant B said:

> *We have two patients that trigger their alarms hundreds of times. Then we can say, this user has the wrong service.*

However, the CCs found that in spite of their reports, the patients continued to have assistive technology that did not work for them. This indicated that the municipalities were unsuccessful in re-evaluating whether a patient had the right technology at the right time, and also that the feedback from the CC was not taken into consideration.

Challenges in Co-Creative and Collaborative Work

The co-creative principle meant that the technology should be continuously cocreated between the service stakeholders and the patients, making it tailor-made to a certain degree. Co-creation and interaction across disciplines and levels were raised as challenges by several informants. Lack of cooperation and the sharing of information and knowledge between CCs and other institutions was a challenge in general, and specifically on a user-related level. Informant F said:

> *We see a need for cooperation, on a larger scale, with GPs and hospitals. But we also see a great need to interact better with the patients to be able to give them a better service.*

The focus of the CC providers was to cooperate and interact with patients. However, informants explained how it was a challenge to recruit patients to cooperate and test the technology. The CCs tested all the technology, and some even created a test lab where patients, future patients, and their next of kin could come and gain hands-on experience of the technologies. Informant K said:

> *We have emphasised a holistic focus. We see it as a core value. Because it is about bringing people with us the whole way, from development to implementation.*

None of the telecare providers involved in this study actually developed health technology. They primarily managed calls/alarms from technology purchased by the municipalities. The technologies had been purchased based on the needs of the patients in each municipality. However, some of the CCs provided recommendations or came up with suggestions to the municipalities about what technologies to procure. In some instances, they even procured technology on behalf of the municipalities. However, it was explained that patients' needs determined what was to be procured, not the technology in itself. Informant K explained:

> *We must understand the needs to solve a problem, then we can go out and see – is there any technology that can work?*

Telecare providers had a unique insight into how the home nursing staff coped with the technology, whether they managed the technology, and whether the patients could handle the technology. Informant B said:

> *We get the home nursing perspective, on what is difficult technology. Because it can be completely different for an engineer who has developed it and it*

XSL•FO
RenderX

makes sense to them, but maybe not to an assistant in the home nursing team.

The Importance of Knowledge for Nurses

The human principle comprised the interaction between patients and employees' knowledge and key information. One CC received 300,000 alarms in 2021, with 861 events in 24 hours and 36 per hour, which underlined how much interaction the CC service had with its patients. The telecare providers frequently talked to the patients and got to know many of them. Having information and personal knowledge about some of their patients led to more calls being resolved by the CC service—hence, the alarm did not need to be escalated to the home nursing team. Informant P stated:

Because in X [a private call centre] they just sent the call through to home nursing, but here [municipal call centre] they answer, and ask questions. "I have lost something under the sofa." A call like this will not go out to the home care team.

As a result of the high number of alarms, an ethical dilemma emerged regarding frequent callers. The patients labeled as frequent callers could lead to employees in the CC service feeling that they were getting tired of the patients and, therefore, needed to work to ensure that they upheld their standards and the quality of their communication. The telecare providers said that they managed it because they knew that these patients' needs were not being met. Some reflected that the reason for the frequent calls might be the fact that the patient had the wrong technology. Informant D said:

I have experienced that I get tired of alarms, especially when the alarm is triggered for "nothing."

Findings highlighted a variety of health personnel working in the CCs. Some of the CCs had nurses, some had health care workers only, and some had technicians working together with nurses as operators in the CC. Some CCs had a standard minimum requirement for at least one nurse, whereas others, which only had nurses, were looking to hire health care workers. Informant P said:

We have made it a requirement that a nurse must be present at work at all times.

Not only did the professional backgrounds of the operators vary, but their level of experience also seemed to differ. One of the CCs was colocated with an accident and emergency department. The nurses had a high level of competence in handling telecare and solving acute matters over the phone. Informant A said:

We only have nurses with very high competence answering acute calls. Because that happens every week, and they need to be able to handle it.

Employees' experience seemed to be key. The interviews and observations revealed that there were no national decision-making tools, procedures for the receipt of alarms, questionnaires, or standards. One nurse (Informant C) working for a CC colocated with an accident and emergency department said:

At the accident and emergency department we use triage systems [decision-making tools], but at the call centre we have nothing. There are no procedures as to what we should do.

Where there were no decision-making tools or procedures for how to receive and handle a call, it could elevate the risk of misunderstandings, as there was no "safety net" to ensure the operators had asked about everything that was relevant when handling an alarm. One informant explained an episode where the operator did not inquire thoroughly enough during an alarm that ended with a patient lying on the floor for a long time, while the operator thought that the patient was lying in bed. Another CC, which monitored a wide variety of technology-based services, including lift alarms, fire alarms, burglary alarms, and an after-hours answering service for the municipality, needed other skills for non-health-related alarms. Informant K explained:

We have two operators on duty – one with a health background and one with a technical background. This composition is because of the portfolio of services we have.

However, all alarms had the same ringtone in all the CCs. In the CC where both nurses and technicians were working alongside each other, they answered alarms without any system or pattern as to which one of them was answering which alarm. The lack of alarm identification meant that the technicians answered medical alarms, and vice versa. They would assist each other if one of them, for example, answered a cardiac arrest. Informant C stated:

If a fire alarm has gone off in an apartment, we should act right away, but if an alarm box has a low battery, we have 12 hours. But we can't tell the difference between the two when the alarm goes off [in the call centre].

The study found no evidence to show that there were more adverse events in CCs with health care workers than with experienced nurses. This is because there was no indication of any system for registering adverse events, or opportunities to look into whether CCs with different professional employee profiles had more adverse events than others.

Challenges With Partly Integrated Technology

The integrated principle meant that both technology and work processes should be integrated with existing systems. Even though the CCs reported that it was demanding to cope with all the technical integrations, it seemed that this was a very high priority for all of them. They also emphasized the importance of the integrations. Informant A explained:

Because of the integrations, we can clarify many alarms that come in.

All the CCs had patients' medical records integrated. This meant that the telecare providers had access to information about the patients, such as when the home nursing team visited the patients. When operators answered alarms, a notification came up on the screens, as well as an alarm or ringtone in the CC. The operator could see who was contacting them, and their medical record came up on their screens. Furthermore, it seemed that the technology used by the patients was integrated into the platforms of the CCs. However, the CCs explained how it could

XSL•FO
RenderX

take time and be challenging to integrate new technology into the platforms. The CCs had many vendors to relate to and many technologies that needed to be integrated into their systems. On top of that, new technologies were making their way into the service, which also required integration.

As for collaboration and the integration of work processes, the CCs seemed to closely cooperate with the home nursing units in the different municipalities where they provided services. However, the analysis indicated an "us/them" attitude between the CC providers versus the more administrative service allocation office in the different municipalities, suggesting a lack of cooperation between the organizational units involved. Informant K said:

> *I don't know what to say about the administration – it is not that they are falling behind, but the cooperation, it has not worked so well.*

Better Use of Data for Evaluation and Learning

The evaluated principle indicated learning from experience and improving systems accordingly. It should involve an ongoing effort to increase quality. The CCs had a lot of data from their operations, and the municipalities received reports from the CCs regarding this data on a regular basis. This data provided key information about the CC service and the patients and could be the basis for evaluation, learning, and development of the service. However, the CCs were not sure whether the municipalities used the data for learning purposes or if they used the data at all. The CCs monitored the number of alarms related to each of the collaborating municipalities and variations over the year and could provide data for individual patients, mapping whether the service was the right one for them. Informant B explained:

> *Creating value through better use of data is important for the municipalities … How many 'false alarms' did the home nursing have to take on before the call centre? When the alarm is sent now, it is real.*

The CCs were reporting to the municipalities because they were trying to see the effects of their work, making sure they procured the right technology, and getting it documented. Informant K explained:

> *We are reporting, we do it because we want to see the effects of the measures we implement. A call centre is not a quick fix, but thorough work from top to bottom or bottom to top.*

Furthermore, the CCs were evaluating if there were other services they could operate to make them more resilient. Many options were considered at the idea stage, but digital home monitoring was a trend that many of the CCs were looking at implementing as a service. Some CCs had already started. Informant L said:

> *Digital home monitoring and rehabilitation. We are investigating the possibilities.*

Discussion

Principal Findings

Using the ARCHIE framework to evaluate quality in municipal CC services reveals that, despite efforts, the service remains fragmented and therefore poses a potential threat to quality. The findings in this study appear not to reflect the quality principles in the ARCHIE framework and what is indicated as high-quality telecare. The framework promotes patient-centered care, whereas this study suggests that the patients' voices are barely heard, and it appears that consideration of what matters to the patients is lacking. This permeates throughout the following principles of the ARCHIE framework: anchored, realistic, cocreated, and poses challenges to the human principle.

In the following, the results will be discussed using the normalization process theory (NPT) [43]. NPT is a framework or set of tools concerned with understanding 3 main issues: implementation, embedding, and integration of new or modified practices, and these align closely with the challenges identified in this study. Using the ARCHIE principles, the study has researched the quality of municipal telecare CCs on a micro level [44]. To further specify how municipal telecare CCs can become part of normal digital service delivery on a meso level [44], NPT offers valuable insight and aids the assessment of how far municipal telecare CCs have come in terms of integration into routine practice, and what is missing to become a fully embedded, high-quality digital health service.

Lack of Quality Challenges Telecare Call Center Normalization

Overview

In this study, the ARCHIE framework [25] was chosen because it was specifically developed as a set of quality indicators for telecare services. ARCHIE emphasizes values such as person-centeredness, coproduction, and inclusivity, which are particularly relevant in the context of digital health and remote care [25]. ARCHIE provides a useful framework for assessing how telecare services align with the intended values of care delivery. Its design makes it especially suitable for evaluating the quality dimensions of telecare implementation, which appear to be less visible in more structural models such as Donabedian's [18,24]. Although foundational in health care quality assessment, the Donabedian model [18] focuses primarily on structure, process, and outcome and may not sufficiently capture the relational and participatory aspects of telecare services.

Of particular relevance in the NPT framework are the mechanisms that inhibit or promote the new practice, namely the following: (1) coherence, (2) cognitive participation, (3) collective action, and (4) reflexive monitoring [45]. In the following, each principle will be briefly presented. The results from the analysis of the relevant ARCHIE principles will then be discussed, including implications.

Figure 2 illustrates how the principles from the ARCHIE framework align with the NPT mechanisms and findings of this study.

XSL•FO
RenderX

Figure 2. Results of the ARCHIE (Anchored, Realistic, Cocreated, Human, Integrated, and Evaluated) analysis on the call center level related to a broader implementation framework, normalization process theory (NPT), at the meso level.

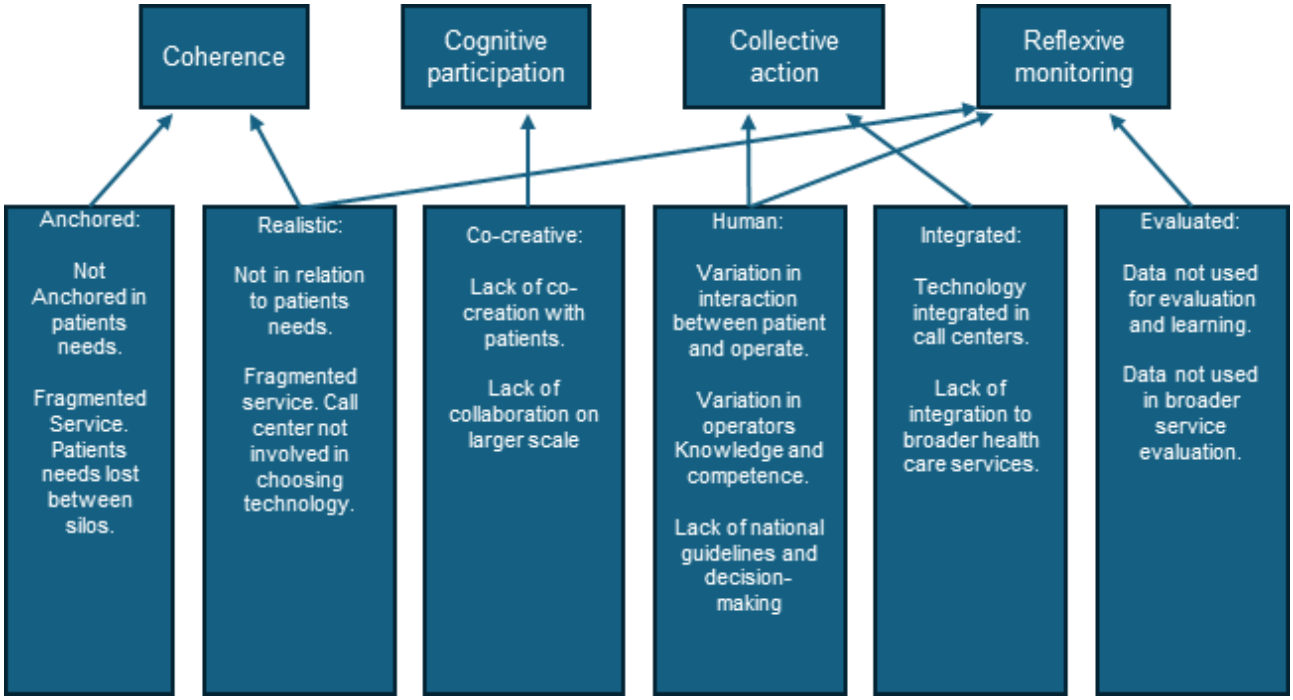

Coherence: Lack of Shared Understanding

On a meso level, the NPT construct Coherence is the process, sensemaking, and work to understand how a technology or service is implemented into practice, and also the routine for embedding a practice [45]. Coherence is regarded as a cognitive construct [43]. On a micro level, the ARCHIE principles anchored and realistic relate to coherence. Anchored links to the patient and to a shared understanding of the patient's needs, which correlates well with coherence. Realistic looks to technology, and whether the technology is realistic for the patient [24]. Although the realistic principle may relate to several NPT constructs, an intervention that is not meaningful will rarely be experienced as realistic. In the ARCHIE principle, realistic was linked to coherence.

Findings in this study reveal a lack of coherence in municipal telecare CC services, particularly in terms of a lack of shared understanding of the patients' needs. This could result in patients being assigned inappropriate technology. The lack of coherence becomes visible when seen through the principles of the ARCHIE framework and suggests that it is a system feature where one principle affects the other. If the decision by the service allocation office is not anchored, this will have a domino effect on the following principles, and they will not be right either. This shows how the telecare service is fragmented, and several silos are unsuccessful in working together to execute the service.

Dillon et al [46] found that when the health care system is perceived as fragmented, the outcome could be a loss of trust and difficulties navigating between providers. These results may be transferable to the telecare service. Furthermore, fragmentation may substantiate the "us/them" feeling, causing an even greater division between the stakeholders, as this study implies. Dillon et al [46] recommend a fundamental restructuring of the health care system to coordinate patient-centered care to meet the needs of the patients and to rebuild trust in the health care system. The ARCHIE framework also promotes patient-centered care, but results from this study show that the lack of tailored approaches may hinder quality in telecare and in the CC services, despite efforts by municipal telecare CCs to improve and professionalize their services.

Cognitive Participation: Uneven Stakeholder Engagement

The NPT construct Cognitive participation relates to the process that needs to be undertaken by individuals to achieve engagement for the implementation of a new practice on a meso level [43]. On a micro level, cognitive participation can be seen in relation to the ARCHIE principle of co-creation. This principle was not met, and the result can be divided into two. First, the lack of collaboration on a larger scale, with other health care services, and second, the lack of collaboration and co-creation with patients. Interestingly, the informants said that it was difficult to engage the patients in co-creation. Fusco et al [47] identified low health literacy, lack of understanding, time constraints, and intimidation of the co-creation process as barriers to successful co-creation. The reasons why the municipal telecare CCs in this study found it challenging to engage patients can only be speculated on, assuming that Fusco et al's [47] barriers also apply to municipal telecare CCs.

Collective Action: Fragmented Operational Routines

On a meso level, Collective action relates to the work individuals and organizations need to do to enact implementation [43]. On a micro level, the ARCHIE principles human and integrated align with collective action. The findings show that operational routines within municipal telecare CC services are fragmented, the staff receive limited training, and there is a lack of standardized protocols and integration with the broader health

XSL•FO
RenderX

services. These barriers hinder the enactment of quality-enhancing practices and suggest that municipal telecare CCs have not yet been fully normalized within municipal health care structures. Although key to telecare delivery, the municipal telecare CCs seem to operate in partial isolation from other stakeholders, which further complicates coordination and continuity of care.

The other ARCHIE principle closely linked to collective action is human. This study has mapped a variety of health professions working in municipal telecare CCs, spanning from highly experienced nurses working in both emergency rooms and CCs to technicians. Given that remote care is a different way of providing care, specific skills are required [48,49]. The CCs have to be prepared to handle emergency alarms, and findings indicate that the staff's experience level is key. Anker-Hansen and Johansen [50] found that the nurses ensure patient safety in complex remote decision-making in telecare. The complexity includes making accurate decisions based on unclear patient descriptions and challenges with communication [50]. The telecare nurses emphasized the importance of clinical expertise and the challenging nature of the decision-making processes and communication. They asserted that the provision of telecare requires that the staff working in CCs must have both proper training and experience, which is not the case in this study. Proper training and protocols can help to mitigate these potential patient safety risks [32].

Data from this study shows that miscommunication and adverse events have occurred. In Fisk et al's [48] study, the emphasis is on the need for a diverse range of skills and knowledge, which may include a mix of health care professionals. There is a lack of national systems, updated guidelines, procedures, and best practices for handling alarms to guide the operators working in the CC, which may negatively impact the quality of the service. Hughes [48] suggests that most adverse events result from faulty systems and processes rather than individual errors. This raises concerns about the impact of nonoptimized systems in telecare services.

In 2014, Guise et al [32] called for increased standardization in reporting adverse events to improve the knowledge base in telecare, but current research indicates that little progress has been made since. Conversely, there are critical voices expressing concerns about over-reliance on standardized guidelines and protocols. It is argued that standardization can lead to a lack of flexibility and creativity in addressing complex, real-world problems, and that these guidelines sometimes stifle critical thinking and innovation, leading to a more bureaucratic and less-effective health care system [49]. The importance of reflexivity and questioning established norms is stressed, and it is argued that a one-size-fits-all approach to guidelines can overlook the unique contexts and needs of different health care settings and patients. It seems that the absence of national guidelines or standards further contributes to the disparity of collective action being varied, making it difficult to foster collective ownership of telecare quality improvement.

Reflexive Monitoring: Limited Evaluation and Learning

Reflexive monitoring reflects the evaluation of a new practice and the effects the new practice has. This includes assessment of the advantages and disadvantages of the new practice [45]. On a micro level, reflexive monitoring aligns with several of the ARCHIE principles: realistic, human, and evaluated. This implies that both the ARCHIE framework and NPT are closely related to practice, gliding into each other, and can be challenging to separate. The ARCHIE principles of human and realistic have been discussed under other NPT constructs; however, they are also relevant on a meso level of reflexive monitoring because they emphasize continuous assessment and adaptation based on human experience and context.

In the ARCHIE framework, the insufficient and ineffective utilization of data falls under the evaluated principle [24]. The results of the study reveal that few municipal telecare CCs have mechanisms for evaluating their performance or for incorporating feedback from patients. This lack of systematic evaluation limits opportunities for learning and adaptation and contributes to persistent gaps in quality and safety. Without reflexive monitoring, it is difficult to identify what works, what needs improving, and how services can evolve to meet patient needs more effectively. The effective utilization of data in telecare services is crucial for enhancing patients' outcomes and service efficiency. When data is not coherently managed or used, it can lead to fragmented care, missed opportunities for service improvement, and reduced patient trust [51]. Inefficiencies such as repeated tests and procedures may also arise, increasing operational costs [52]. This concern is seen in the WHO's SMART Guidelines [53], which emphasize that digital technologies should be designed to support consistent, standards-based data use to improve care quality and coordination.

Regular audits and feedback mechanisms are also vital for continuous improvement [54]. Furthermore, Guise and Wiig [49] encourage training that includes patient-centered methodology to ensure good patient–professional relationships in remote care.

There are several challenges regarding the telecare service that need to be addressed on a meso level: lack of coherence, effective utilization of data, and outdated national guidelines. From a system perspective, health care is seen as a complex set of management levels, along with the policies, resources, and actors involved in delivering care [55], referring to the entire network of organizations, people, and actions whose primary intent is to promote, restore, or maintain health [56].

Strengths and Limitations

The ARCHIE framework was selected for this study because it was especially developed for defining quality in telecare. The framework has successfully assisted the research, uncovering quality and patient safety challenges within the telecare service, including municipal telecare CCs.

This study's combined deductive–inductive approach to the ARCHIE framework contributes to an enhanced understanding of the complex nature of the municipal telecare CC service.

However, when analyzing the data, we found that the different principles in ARCHIE were difficult to differentiate from one another and could be intertangled, meaning that the data could fit several of the principles. This was resolved by being precise

XSL•FO
RenderX

in choosing statements and through discussions within the author group, to ensure that the statements reflected how the group understood the principles.

The selection of the municipal telecare CCs was a potential limitation for the study. At the time of data collection, modest information was available on existing CCs, and the selection was therefore random for practical purposes. As we have learned in the process of writing this study, however, the CCs under study represent the most common way of organizing CCs in Norway. The ease of access to informants might have compromised the reliability and validity of the data [57]. Another limitation to be aware of is that of the cross-sectional snapshot, as data were collected at one point in time in this study. This may not represent long-term behaviors or trends.

The lack of national standards makes it challenging to benchmark quality, which affects both analysis and interpretation.

A strength of the research is the diverse background of the research group. This interdisciplinary background enabled the group to conduct an open and critical process, with the aim of reducing bias as much as possible.

Conclusions

The findings of this study suggest that telecare services in Norway are not yet fully normalized. The challenges identified across all 4 NPT constructs—conceptual ambiguity, uneven engagement, fragmented routines, and limited evaluation—indicate that telecare remains in a transitional phase of implementation. The application of the ARCHIE framework [24] helps to clarify the current position of telecare within the health care system and highlights specific areas where targeted interventions could accelerate normalization. Combined with NPT, which offers critical insights into the processes necessary for change at the meso level, this study provides a dual insight for understanding both the process and the quality dimensions of telecare. These insights are critical for policymakers, service designers, and practitioners aiming to improve the quality, safety, and integration of telecare services.

Implications for Policymaking

The findings of this study imply that municipal telecare CCs need to improve the telecare service to achieve the full potential of integrating a high level of quality and patient safety. Improvements should be considered at both meso and micro levels, as shown in this study. At a meso level, one way to advance is for stakeholders in telecare services to improve communication across their services and reduce fragmentation in the system. The telecare service calls for national standardization and the professionalization of the service. On a micro level, all of the ARCHIE principles need improvement. Training for telecare operators is one area that should be prioritized. As shown in this research, Greenhalgh et al's [24] ARCHIE framework can be a useful tool for assessing quality in emerging services within the health care domain.

Issues for Further Research

The following issues for further research to improve quality were identified in this study:

1. The patient perspective remains underresearched, as this study has focused on the municipal telecare CCs.
2. Cooperation and communication between stakeholders—in particular, how do the municipalities consider patients' needs for assistive technology, and how are concerns about ill-fitted assistive technology handled?
3. Competence development—what kind of training do telecare operators receive before they start as operators, and what plans are in place to maintain and increase knowledge?

Acknowledgments

All authors made significant contributions to the manuscript. The article was prepared in close collaboration with the authors LGE, EN, JD, and HE. The data were collected by LGE, but all authors contributed to the analysis and writing of the manuscript. Furthermore, all authors read and agreed on the final version of the manuscript. The authors would like to thank all the informants for participating in this study.

Funding

LGE is funded by the University of South-Eastern Norway. JD, EN, and HE were partly financed by the Research Council of Norway via the project Cocreated Health Technology (COTECH, project number 331903).

Data Availability

The data collected and analyzed for the previous and this study are not publicly available due to privacy reasons.

Conflicts of Interest

The first author (LGE) is employed by a private company, Tellu. The data collection and writing of this article were, however, performed before this employment began. The company did not, under any circumstances, influence data analysis, interpretation, or the writing of the manuscript.

References

XSL•FO
RenderX

1. Solli H, Bjørk IT, Hvalvik S, Hellesø R. Principle-based analysis of the concept of telecare. J Adv Nurs 2012 Dec;68(12):2802-2815. [doi: 10.1111/j.1365-2648.2012.06038.x] [Medline: 22607115]
2. Farshchian BA, Vilarinho T, Mikalsen M. From episodes to continuity of care: a study of a call center for supporting independent living. Comput Supported Coop Work 2017 Jun;26(3):309-343. [doi: 10.1007/s10606-017-9262-4]
3. Zallio M, Ohashi T. The evolution of assistive technology: a literature review of technology developments and applications. arXiv 2022. [doi: 10.48550/arXiv.2201.07152]
4. Assistive technology. World Health Organization. 2025. URL: https://www.who.int/news-room/fact-sheets/detail/assistive-technology [accessed 2025-09-16]
5. Ministry of Local Government and Regional Development. Act relating to municipalities and county authorities (the local government act). Lovdata. 2024. URL: https://lovdata.no/dokument/NLE/lov/2018-06-22-83 [accessed 2024-01-20]
6. Grøndal-Eeles L, Dugstad J, Eide H, Nilsen E. The organizational challenges of municipal call centers as a health service in Norway-a multiple case study. BMC Health Serv Res 2025 Jan 20;25(1):108. [doi: 10.1186/s12913-025-12264-0] [Medline: 39833817]
7. Quality of care: a process for making strategic choices in health systems. World Health Organization. 2006. URL: https://iris.who.int/handle/10665/43470 [accessed 2024-07-02]
8. Institute of Medicine (US) Committee on Quality of Health Care in America. Crossing the Quality Chasm: A New Health System for the 21st Century: National Academies Press (US); 2001. URL: http://www.ncbi.nlm.nih.gov/books/NBK222274/ [accessed 2024-07-02]
9. Campbell SM, Roland MO, Buetow SA. Defining quality of care. Soc Sci Med 2000 Dec;51(11):1611-1625. [doi: 10.1016/s0277-9536(00)00057-5] [Medline: 11072882]
10. H. Ministry of Health and Care Services. Meld st 11 (2020–2021). Regjeringen. 2024. URL: https://www.regjeringen.no/no/dokumenter/meld.-st.-11-20202021/id2791147/ [accessed 2024-01-25]
11. Public Health System Group for England. Quality in public health: a shared responsibility. Gov.UK. URL: https://www.gov.uk/government/publications/quality-in-public-health-a-shared-responsibility [accessed 2024-01-25]
12. Quentin W, Partanen VM, Brownwood I, Klazinga N. Measuring healthcare quality. In: Improving Healthcare Quality in Europe Characteristics, Effectiveness and Implementation of Different Strategies: Characteristics, Effectiveness and Implementation of Different Strategies: OECD Publishing; 2019.
13. Jha AK, Prasopa-Plaizier N, Larizgoitia I, Bates DW, Research Priority Setting Working Group of the WHO World Alliance for Patient Safety. Patient safety research: an overview of the global evidence. Qual Saf Health Care 2010 Feb;19(1):42-47. [doi: 10.1136/qshc.2008.029165] [Medline: 20172882]
14. Lawati M, Dennis S, Short SD, Abdulhadi NN. Patient safety and safety culture in primary health care: a systematic review. BMC Fam Pract 2018 Dec;19(1):104. [doi: 10.1186/s12875-018-0793-7]
15. Nieva VF, Sorra J. Safety culture assessment: a tool for improving patient safety in healthcare organizations. Qual Saf Health Care 2003 Dec;12 Suppl 2(Suppl 2):ii17-ii23. [doi: 10.1136/qhc.12.suppl_2.ii17] [Medline: 14645891]
16. Stokke R, Furnes M, Vifladt A, Ringsby Odberg K. Pasientsikkerhet og velferdsteknologi – parhester i utakt [patient safety and welfare technology – out of sync]. TFO 2020 Nov 20;6(2):165-171. [doi: 10.18261/issn.2387-5984-2020-02-12]
17. Donabedian A. The quality of care: how can it be assessed. JAMA 1988 Sep;260(12):1743-1748. [doi: 10.1001/jama.1988.03410120089033]
18. Tossaint-Schoenmakers R, Versluis A, Chavannes N, Talboom-Kamp E, Kasteleyn M. The challenge of integrating eHealth into health care: systematic literature review of the donabedian model of structure, process, and outcome. J Med Internet Res 2021 May 10;23(5):e27180. [doi: 10.2196/27180] [Medline: 33970123]
19. Karlsen A, Sørbø MF, Nylenna M. Digitale helsetjenester: nye muligheter, nye utfordringer [digital health services: new opportunities, new challenges]. Det Norske Medicinske Selskab [The Norwegian Medical Association]. 2020. URL: https://www.michaeljournal.no/article/2020/08/Digitale-helsetjenester--nye-muligheter,-nye-utfordringer [accessed 2024-07-02]
20. WHO develops guidance to improve telemedicine services. World Health Organization. 2025. URL: https://www.who.int/europe/news-room/15-07-2024-who-develops-guidance-to-improve-telemedicine-services [accessed 2025-02-17]
21. WHO issues consolidated guide to running effective telemedicine services. World Health Organization. 2025. URL: https://www.who.int/news/item/10-11-2022-who-issues-new-guide-to-running-effective-telemedicine-services [accessed 2025-02-17]
22. Meld st 9 (2023–2024). Regjeringen.no. 2023. URL: https://www.regjeringen.no/no/dokumenter/meld.-st.-9-20232024/id3027594/ [accessed 2024-06-02]
23. Woolham J, Freddolino P, Gibson G, Daniels S. Telecare at a crossroads? Finding researchable questions. JET 2021 Sep 8;15(3):175-188. [doi: 10.1108/JET-11-2020-0049]
24. Greenhalgh T, Procter R, Wherton J, Sugarhood P, Hinder S, Rouncefield M. What is quality in assisted living technology? the ARCHIE framework for effective telehealth and telecare services. BMC Med 2015 Apr 23;13(1):91. [doi: 10.1186/s12916-015-0279-6] [Medline: 25902803]
25. Alvsåker YB. Velferdsteknologi i Norske hjemmetjenester: en praxeologisk studie om statlige visjoner i møte med markedskrefter og kommunale realiteter [welfare technology in Norwegian home services: a praxeological study of state

XSL•FO
RenderX

visions in the face of market forces and municipal realities]. Norwegian Research Information Repository.: The University of Bergen; 2023. URL: https://bora.uib.no/bora-xmlui/handle/11250/3106782 [accessed 2024-05-06]
26. Tid for handling NOU 2023: 4. Regjeringen.no. URL: https://www.regjeringen.no/no/dokumenter/nou-2023-4/id2961552/ [accessed 2024-06-17]
27. McLean S, Sheikh A, Cresswell K, et al. The impact of telehealthcare on the quality and safety of care: a systematic overview. PLOS ONE 2013;8(8):e71238. [doi: 10.1371/journal.pone.0071238] [Medline: 23977001]
28. Fothergill L, Hayes N, Latham Y, Hamilton J, Ahmed S, Holland C. Understanding how, for whom and under what circumstances telecare can support independence in community-dwelling older adults: a realist review. BMC Geriatr 2025 Jan 27;25(1):59. [doi: 10.1186/s12877-024-05650-6] [Medline: 39871177]
29. Safer primary care. World Health Organization. 2024. URL: https://www.who.int/teams/integrated-health-services/patient-safety/research/safer-primary-care [accessed 2024-08-28]
30. Al Salem G, Bowie P, Morrison J. Hospital survey on patient safety culture: psychometric evaluation in Kuwaiti public healthcare settings. BMJ Open 2019 May 30;9(5):e028666. [doi: 10.1136/bmjopen-2018-028666] [Medline: 31152040]
31. Nwankwo EI, Emeihe EV, Ajegbile MD, Olaboje JA, Maha CC. Integrating telemedicine and AI to improve healthcare access in rural settings. Int J Life Sci Res Arch 2024;7(1):059-077. [doi: 10.53771/ijlsra.2024.7.1.0061]
32. Guise V, Anderson J, Wiig S. Patient safety risks associated with telecare: a systematic review and narrative synthesis of the literature. BMC Health Serv Res 2014 Nov 25;14(1):588. [doi: 10.1186/s12913-014-0588-z] [Medline: 25421823]
33. Stokke R, Melby L, Isaksen J, Obstfelder A, Andreassen H. A qualitative study of what care workers do to provide patient safety at home through telecare. BMC Health Serv Res 2021 Jun 5;21(1):553. [doi: 10.1186/s12913-021-06556-4] [Medline: 34090450]
34. Eisenhardt KM. Building theories from case study research. Acad Manage Rev 1989 Oct;14(4):532-550. [doi: 10.5465/amr.1989.4308385]
35. Eisenhardt KM, Graebner ME. Theory building from cases: opportunities and challenges. AMJ 2007 Feb;50(1):25-32. [doi: 10.5465/amj.2007.24160888]
36. Yin RK. Discovering the future of the case study. method in evaluation research. Eval Pract 1994 Oct;15(3):283-290. [doi: 10.1177/109821409401500309]
37. Yin RK. Enhancing the quality of case studies in health services research. Health Serv Res 1999 Dec;34(5 Pt 2):1209-1224. [Medline: 10591280]
38. Denzin NK, Lincoln YS. Introduction: The Discipline and Practice of Qualitative Research', in Strategies of Qualitative Inquiry, 3rd edition: Sage Publications, Inc; 2008:1-43.
39. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
40. Lincoln YS, Guba EG. Naturalistic Inquiry: SAGE; 1985.
41. Saldana J. The Coding Manual for Qualitative Researchers: SAGE Publications Ltd; 2021. URL: https://www.torrossa.com/en/resources/an/5018667 [accessed 2024-01-20] [doi: 10.4135/9781036235611]
42. National committee for research ethics in the social sciences and the humanities. National Research Ethics Committees. 2024. URL: https://www.forskningsetikk.no/en/about-us/our-committees-and-commission/nesh/ [accessed 2024-01-30]
43. May C, Finch T. Implementing, embedding, and integrating practices: an outline of normalization process theory. Sociology 2009 Jun;43(3):535-554. [doi: 10.1177/0038038509103208]
44. May CR. Making sense of technology adoption in healthcare: meso-level considerations. BMC Med 2015 Apr 23;13(1):92. [doi: 10.1186/s12916-015-0305-8] [Medline: 25902829]
45. Huddlestone L, Turner J, Eborall H, Hudson N, Davies M, Martin G. Application of normalisation process theory in understanding implementation processes in primary care settings in the UK: a systematic review. BMC Fam Pract 2020 Mar 16;21(1):52. [doi: 10.1186/s12875-020-01107-y] [Medline: 32178624]
46. Dillon EC, Martinez MC, Li M, et al. "It is not the fault of the health care team - it is the way the system works": a mixed-methods quality improvement study of patients with advanced cancer and family members reveals challenges navigating a fragmented healthcare system and the administrative and financial burdens of care. BMC Health Serv Res 2024 Nov;24(1):1378. [doi: 10.1186/s12913-024-11744-z]
47. Fusco F, Marsilio M, Guglielmetti C. Co-creation in healthcare: framing the outcomes and their determinants. JOSM 2023 Dec 18;34(6):1-26. [doi: 10.1108/JOSM-06-2021-0212]
48. Fisk MJ, Woolham J, Steils N. Knowledge and skills sets for telecare service staff in the context of digital health. J Int Soc Telemed eHealth 2020 Dec;8. [doi: 10.29086/JISfTeH.8.e15]
49. Guise V, Wiig S. Perceptions of telecare training needs in home healthcare services: a focus group study. BMC Health Serv Res 2017 Feb 23;17(1):164. [doi: 10.1186/s12913-017-2098-2] [Medline: 28231852]
50. Anker-Hansen C, Johansen EK. Navigating care and communication: a qualitative study on nurses' perspectives in response centres. BMC Nurs 2025 Feb 13;24(1):170. [doi: 10.1186/s12912-025-02811-9] [Medline: 39948604]
51. Alenoghena CO, Ohize HO, Adejo AO, et al. Telemedicine: a survey of telecommunication technologies, developments, and challenges. JSAN 2023 Apr;12(2):20. [doi: 10.3390/jsan12020020]

XSL•FO
RenderX

52. Hajesmaeel-Gohari S, Bahaadinbeigy K. The most used questionnaires for evaluating telemedicine services. BMC Med Inform Decis Mak 2021 Feb 2;21(1):36. [doi: 10.1186/s12911-021-01407-y] [Medline: 33531013]
53. SMART guidelines. World Health Organization. URL: https://www.who.int/teams/digital-health-and-innovation/smart-guidelines [accessed 2025-09-17]
54. Leonardsen ACL, Hardeland C, Helgesen AK, Grøndahl VA. Patient experiences with technology enabled care across healthcare settings- a systematic review. BMC Health Serv Res 2020 Aug 24;20(1):779. [doi: 10.1186/s12913-020-05633-4] [Medline: 32838784]
55. Jarvis T, Scott F, El-Jardali F, Alvarez E. Defining and classifying public health systems: a critical interpretive synthesis. Health Res Policy Syst 2020 Jun 16;18(1):68. [doi: 10.1186/s12961-020-00583-z] [Medline: 32546163]
56. McCoy D, Allotey P. An introduction to health systems. In: Systems Thinking Analyses for Health Policy and Systems Development: A Malaysian Case Study: Cambridge University Press; 2021:3-16. [doi: 10.1017/9781108954846.004]
57. Dyer WG, Wilkins AL. Better stories, not better constructs, to generate better theory: a rejoinder to Eisenhardt. The Academy of Management Review 1991 Jul;16(3):613. [doi: 10.2307/258920]

Abbreviations

ARCHIE: anchored, realistic, cocreated, human, integrated, and evaluated
CC: call center
NPT: normalization process theory
WHO: World Health Organization

Edited by A Stone, N Cahill; submitted 15.Apr.2025; peer-reviewed by FA Dhabbari, Q Wang; revised version received 17.Jan.2026; accepted 20.Jan.2026; published 10.Mar.2026.

Please cite as:
Grøndal-Eeles LC, Dugstad J, Eide H, Nilsen E
Quality Challenges in Municipal Telecare Call Center Services: Qualitative Evaluation Using the Anchored, Realistic, Cocreated, Human, Integrated, and Evaluated (ARCHIE) Framework
J Med Internet Res 2026;28:e76054
URL: https://www.jmir.org/2026/1/e76054
doi:10.2196/76054

© Linda C Grøndal-Eeles, Janne Dugstad, Hilde Eide, Etty Nilsen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring the Dynamics of Actors, Structural Factors, and Bricolage in the Implementation and Sustainability of eHealth Solutions: Qualitative Multiple-Case Study

Susanne Eriksen[1*], BA, MSc; Christine Øye[2*], Prof Dr; Anne Marie Dahler[3*], Prof Dr

[1]Department of Health and Caring Science, Faculty of Health and Social Science, Western Norway University of Applied Sciences, Bjornsonsgate 45, Haugesund, Norway

[2]Faculty of Health and Social Science, Western Norway University of Applied Sciences, Stord, Norway

[3]Department of Applied Welfare Research, UCL University College, Odense, Denmark

[*]all authors contributed equally

Corresponding Author:
Susanne Eriksen, BA, MSc
Department of Health and Caring Science, Faculty of Health and Social Science, Western Norway University of Applied Sciences, Bjornsonsgate 45, Haugesund, Norway

Abstract

Background: European health care systems face mounting pressures from an aging population, workforce shortages, and decentralization, challenging the delivery of accessible, high-quality care. eHealth solutions are widely promoted to enhance efficiency and improve the quality of care. Despite a strong policy report, anticipated benefits remain unrealized, as implementation processes often encounter barriers and high failure rates. Research shows that drivers and barriers are dynamic and shaped by actor interactions. Some studies suggest that certain actors, often acting as bricoleurs, play a critical role in overcoming these barriers through adaptive and improvised practices. However, little is known about how these actors enact roles, what features enable bricolage, and how structural conditions influence these practices.

Objective: The aim of this study is twofold. First, it investigates the roles and features of actors involved in innovation processes, with a particular emphasis on the application of bricolage to overcome barriers and the influence of structural factors on these processes. Second, it aims to contribute both theoretical and empirical insights to deepen the understanding of barrier dynamics within innovation processes.

Methods: We conducted a multiple-case study comprising 10 semistructured interviews, 11 focus groups with health care professionals, managers, trainers, and policymakers, participant observations of training sessions, and document analysis. An iterative process integrated the dramaturgical approach with the concept of bricolage, guiding the reflexive thematic analyses.

Results: Roles were enacted based on available information, context, and assigned functions. Service specialists (eg, superusers) and mediators (eg, unit or project managers) gained backstage insights through shadowing staff, evaluations, and support activities. When mandated and equipped with contextual and technical knowledge, these actors became bricoleurs, addressing unforeseen challenges by creatively mobilizing resources and thereby transforming barriers into promoters. Effective bricolage required proximity to the implementation site, dedicated involvement, and experiential knowledge of health care and technical domains. Key drivers included colocation, supportive management, stable teams, superusers, tailored training, follow-up activities, and informal evaluations. Barriers such as organizational silos, leadership shifts, staffing shortages, high turnover, geographic dispersion, and technology perceived as challenging or surveillance-oriented constrained bricolage and hindered implementation.

Conclusions: Actors may become bricoleurs when their assigned roles, contextual knowledge, and backstage access enable them to improvise in response to unforeseen challenges. Through a dramaturgical lens, bricolage is an adaptive performance that sustains frontstage care delivery. Bricoleurs combine proximity, experiential knowledge, and dual expertise to transform barriers into drivers by adjusting the innovation process and fostering interaction. These practices illustrate the mutual shaping of structure and agency: enabling conditions expand the space for bricolage, while barriers narrow it. Understanding this dynamic is essential for advancing theory on innovation processes and for designing implementation strategies that leverage bricolage as a mechanism for transforming barriers into drivers of innovation.

(J Med Internet Res 2026;28:e79999) doi:10.2196/79999

KEYWORDS

eHealth solutions; innovation processes; interaction; public sector innovation; implementation

XSL•FO
RenderX

Introduction

Background

One of the significant challenges facing European societies is the aging population, coupled with a shortage of health care staff [1-4]. These challenges are compounded by a deliberate decentralization of health care services, driven by political decisions, which complicates efforts to ensure accessible and high-quality health care [2,5-8]. In response to these pressing issues, there is a growing emphasis on innovation within the public sector [9,10], particularly through the adoption of eHealth solutions [1-3]. eHealth solutions are often defined as the organization and delivery of health services using information and communication technologies (ICTs) to support and enhance health care [4,11]. Across European societies, including Norway and Denmark, which form the context of this study, there is strong optimism that innovative eHealth solutions can improve productivity, free up time for patient-centered care and core health care tasks [2,8,12], and empower older adults to live independently at home by enhancing their health and quality of life [13-16]. The belief stems from the potential that eHealth solutions can enhance efficiency, improve the quality of care, and bolster patient security [17-19]. However, despite a strong policy push for implementing eHealth solutions, authorities acknowledge that the full benefits have yet to be realized [2,3,20,21].

Research on innovation processes involved in implementing and sustaining eHealth solutions in health care reveals that these processes are often fraught with challenges, resulting in a high failure rate [22-26]. Numerous studies have identified a range of factors influencing the outcomes of eHealth innovation processes [22,26-33]. These factors are often presented as separate concepts, being either a driver or a barrier [34,35]. However, scholars are increasingly emphasizing the dynamic nature of drivers and barriers [36-38]. What may initially appear as a barrier can be transformed into a driver if addressed appropriately [38-40]. Detecting barriers and addressing them depends on the interaction among actors as well as proactive engagement from specific actors. Previous research has demonstrated that collaborating actors in innovation processes engage in complex and dynamic interactions and negotiations [38]. As the complexity of the innovation process increases, interactional barriers become more prominent [34,36,41], hindering effective knowledge sharing, which is crucial for implementing and sustaining eHealth solutions [25,32,38,42,43]. Research has identified specific actors who are effective in overcoming interactional barriers [38]. Such actors are frequently found to act as boundary spanners, bridging knowledge gaps, for example, between ICT staff, management, and health care professionals. By bridging knowledge gaps, these actors can transform barriers into opportunities, making innovations more contextually relevant and beneficial by, for example, facilitating the redesign of workflows, providing adequate training and support to users, and highlighting problematic issues [26,31,38,44]. These actors can also be referred to as bricoleurs [45-47], and their contributions are considered vital for integrating eHealth solutions into health care practices [7,48,49]. Bricolage is characterized by its spontaneous and improvised nature, involving small, pragmatic adjustments in response to unforeseen events or emerging needs [47,49]. This process is often informal, builds on embodied experience within the practice, and is inherently collaborative, as it necessitates interaction with other actors [45,47,49]. Given its intangible nature, it remains unclear how the context can facilitate bricolage work [46]. There also remains an incomplete understanding of the nature of the roles of these bricoleurs as well as how these roles might be identified, enabled, and enhanced [26,36,38,49,50]. The literature emphasizes the importance of investigating individual actors as units of analysis, given that microlevel practices have been largely overlooked [24,30,50]. Research has identified various features of individual actors involved in implementing and sustaining innovations, including ICT-related skills, experience, demographics, and personality traits [26,35]. Creative and empowered actors, who possess the ability to navigate and overcome a risk-averse administrative culture, play a crucial role in driving innovation [35]. However, it remains unclear how these features interact with the enactment of roles and the contextual factors. Bricolage provides an approach that emphasizes how innovation can be made possible by recognizing both the work of actors and the influence of structural factors [47]. We therefore consider the concept of bricolage useful for understanding the complex dynamics between the micro- and system levels of innovation processes.

Objectives

Against this background, this study aims (1) to investigate the roles and features of actors involved in innovation processes, with particular emphasis on the application of bricolage to overcome barriers and the influence of structural factors on these processes; and (2) to contribute both theoretical and empirical insights to deepen the understanding of barrier dynamics within innovation processes. We pursue these aims through a qualitative, multiple-case study set within a European Union innovation project that focuses on enhancing digital skills and innovation readiness. By examining 3 cases of implementing and sustaining innovative eHealth solutions, our study seeks to address the following research questions:

- How do actors involved in implementing and sustaining eHealth solutions enact the role of bricoleurs, and what features enable this role?
- In what ways is bricolage performed to transform barriers into drivers of innovation, and how do structural conditions shape or constrain these practices?

Methods

Study Design

The study used an exploratory case study design that adhered to the principles of multiple holistic case study design [51]. Each case served as a distinct unit of analysis. The exploratory case study is an empirical investigation that examines processes and uncovers mechanisms related to a contemporary phenomenon within its real-life context [51,52]. This approach was well-suited to our research design, as we aimed to understand the roles of various actors in innovation processes and to investigate how they overcome barriers and transform

XSL•FO
RenderX

them into drivers of innovation, as well as how structures influence the enactment of roles and performance of bricolage.

Case Selection and Data Collection

The cases selected for this study were part of the Digital and Innovation Skills Helix project, a European Union initiative aimed at enhancing digital skills and promoting innovation readiness. As part of a forthcoming innovation process aimed at implementing an eHealth solution, these cases tested 3 tools developed through the Digital and Innovation Skills Helix project. These tools facilitated the acquisition of digital skills, cocreative implementation planning, evaluation, and competence assessment. Further details about the tools can be found in our previous work [53].

The 3 cases are situated in a Nordic welfare context. Case A is a nursing home situated in a rural municipality in Western Norway, offering 24-hour health and care services to residents for both short-term and long-term stays. This municipality actively participates in regional and national eHealth networks and collaborates closely with neighboring municipalities to enhance health care services. The nursing home accommodates 41 residents and employs 60 staff members, including 1 leader and 4 unit managers. Since 2017, the municipality has systematically upgraded its outdated equipment with new eHealth solutions, such as electronic medicine dispensers and safety alarms. At the time of data collection, the nursing home implemented a new patient monitoring system that included digital supervision. This system's primary objectives were to increase service efficiency and improve patient security.

Case B is a home care service located in a rural municipality in Western Norway, providing 24-hour assistance with daily living and home health services. Most service users are frail older adults who require support to continue living at home. This home care service supports approximately 150 service users and employs 50 health care staff, including the home care service unit manager. Since 2015, the municipality has focused on implementing eHealth solutions in the health care sector to address demographic challenges and deliver sustainable health care services. The home care service implemented electronic door locks (e-locks) in service users' homes, aiming to provide faster and safer assistance.

Case C is a cross-sectoral collaboration between a hospital and a municipal home care service in Southern Denmark. The hospital offers emergency care, outpatient treatment, and examination services to patients who have been injured. The home care service offers 24-hour assistance, including support for daily living and somatic and psychiatric care. In 2018, the local hospital initiated a project to collaborate with the municipality through video consultations, specifically for discharging complex and vulnerable patients from the hospital to municipal care. Aligned with the Regional Council's digitalization strategy, the objective was to enhance cross-sectoral collaboration using technology, streamline discharge conferences, and ensure more coherent patient care. The hospital was allocated approximately US $1.5 million to develop digital competencies among its staff.

The selection of cases was pragmatic, as the available cases within the European Union project were limited, resulting in the inclusion of 3 cases that exhibited varying levels of complexity. As case C involved implementing video consultations in a cross-sector collaboration between a hospital and a municipal home care service in Southern Denmark, this case is more complex compared to cases A and B, which were more similar and less complex, as they implemented an eHealth solution within a single organization, engaging only a limited number of employee groups. The empirical material for these cases included document analysis, participant observation, semistructured individual interviews, and focus groups (see Multimedia Appendix 1 for an overview of the data collection).

The interviews and focus groups, with some being more prominently featured than others, served as a central component of our analysis, providing in-depth insights into the experiences and perspectives associated with the implementation process (Multimedia Appendix 2). The participant observations took place during training, where the observer participated and observed questions, discussions, and task-solving activities. The participant observations complemented the interviews, enhancing the analytical depth of the study by offering a comprehensive understanding of the setting and context as well as nuanced insights into the social dynamics (Multimedia Appendix 3). Although these observations, along with the document analysis, served as critical supplementary materials, they primarily provided a contextual backdrop for interpreting the qualitative data derived from interviews and focus groups. The data collection took place between 2020 and 2022, during and after the testing phase of the 3 tools. The data were gathered in person or online via Zoom (Zoom Video Communications) or Microsoft Teams, conducted by SE, AMD, or CØ. The data collection ended when data saturation was reached.

Participants for the study were recruited with the assistance of key stakeholders involved in implementation planning and training within the clinical settings. In cases A and B, a designated project manager from each case served as our primary point of contact. For case C, we contacted the individuals responsible for training at the hospital, where one of the trainers became our primary point of contact. These 3 contact persons facilitated the recruitment process by reaching out to potential participants by email. A total of 949 individuals were identified as potential participants from the training roster, and 31 individuals were identified from the implementation working groups. We contacted 115 individuals, of whom 39 agreed to participate in the interviews, and 70 agreed to participate in participant observations, resulting in a final sample of 109.

Ethical Considerations

Ethics approval was obtained from Sikt—Norwegian Agency for Shared Services in Education and Research (ref: 198584). Before participation, all participants were provided with informed consent that they signed before participation, which included a comprehensive overview of the study project, its objectives, and data handling procedures. No compensation was provided to the participants. Anonymity was ensured by replacing names with pseudonyms and revising the empirical

XSL•FO
RenderX

material to obscure recognizable quotes. To ensure data security, the original recordings as well as transcripts were stored as password-protected files on a secure research server, with access limited to the three authors only.

Analyses

NVivo (version 21; Lumivero) software was used to analyze verbatim transcriptions of audio-recorded interviews and focus groups. The reflexive thematic analyses for this study were applied to generate initial themes by identifying patterns of shared meaning across the dataset [54-56]. We followed the 6-phase approach developed by Braun et al [55]. First, we aimed to enhance reliability in the analysis process by conducting independent readings of the transcripts. Second, all three authors generated initial themes independently, guided by the concept of bricolage [57] in combination with the dramaturgical approach [58] and the Consolidated Framework for Implementation Research [59]. Since the study aimed to investigate role enactment, bricolage strategies, and structural factors, we approached the data, exploring and tagging text that reflected strategies involving bricolage activities and role features akin to those of mediators and service specialists, as well as structural factors affecting implementation. Third, all three authors constructed themes through thematic mapping, which involved visually exploring potential themes and subthemes, as well as connections between them. Finally, themes were revised and defined in collaboration between all three authors before producing the report. Researcher reflexivity was actively pursued in accordance with the standards of qualitative research [60]. This included engaging in critical dialogue to challenge and complement each other's interpretations and explicitly acknowledging our personal and professional backgrounds early in the research process. These discussions helped us remain aware of our initial assumptions and avoid conflating prior perspectives with insights emerging from the data.

Conceptual Framework

The dramaturgical approach by Goffman [58] illustrates how actors manage their frontstage performances within social settings to navigate and influence others' perceptions. In contrast, backstage involves actions that support the frontstage performance but do not align with its presented image. This perspective highlights that inventive strategies and adaptive behaviors emerge in response to the interactions and perceptions within actors' immediate social contexts rather than being dictated solely by structural conditions. By integrating the dramaturgical approach and the concept of bricolage, we aim to develop a more comprehensive understanding of how actors can transform barriers into drivers through performing bricolage. Previous research has indicated that distinct roles, such as mediators and service specialists, are pivotal in overcoming interactional barriers [38]. In the paragraph below, we delve deeper into the roles of mediators and service specialists, emphasizing their essential function in bridging interactional gaps.

A mediator acts as an intermediary, facilitating mutually beneficial agreements between 2 potentially opposing teams. By cultivating trust and managing confidential information, the mediator maintains a delicate balance, sometimes projecting a skewed perception of loyalty to foster closeness and understanding among the teams. Examples of mediators may include facilitators, project leaders, or department managers. Conversely, service specialists focus on constructing, repairing, and maintaining performance. Acting as "scene workers," they enable actors to effectively perform their roles and define situations without encountering dramaturgical obstacles [58]. Examples of service specialists could include ICT specialists, champions, or "superusers."

To investigate the strategies used by mediators and service specialists in merging frontstage with backstage to overcome barriers and further the innovation process, we integrate Lévi-Strauss's [57] concept of bricolage. Bricolage, derived from the French verb "bricoleur," means to tinker or improvise. Lévi-Strauss [57] conceptualized bricolage as a creative method of using available resources, contrasting the improvisational nature of the bricoleur with the systematic, planned approach of the engineer. This duality draws attention to different problem-solving methods [57]. Since its introduction, bricolage has found relevance across various disciplines [61-63], including studies on public sector innovation [7,45-49,64]. While extensive research has examined bricolage activities, there has been less focus on bricoleurs themselves [49]. This oversight may stem from Levi-Strauss's [57] structuralist perspective, which suggests that objective structures shape actors' lives, often overshadowing individual interpretations. However, latter interpretations of the concept of bricolage are not strongly underpinned by the structuralist approach and see bricolage as an activity where the bricoleur creates structures from resources at hand [46]. As such, bricolage is considered relevant for capturing the dynamics of innovation and the connection between microlevel practices and broader systems of innovation [46]. Along with other researchers [65-67], we argue for the need to develop new approaches to advance public sector innovation processes. As such, to better understand the motivations and strategies of actors, it is necessary to recognize the interplay between different approaches. Situated in the critical realism paradigm, we acknowledge that both Goffman's and Lévi-Strauss's approaches can complement one another. While objective structures influence individual agency, especially in complex situations, actors' agency is also shaped by their subjective interpretations of social reality.

The dramaturgical approach illuminates the social dynamics and interpersonal interactions on the microlevel that shape how bricoleurs operate, for example, how the roles are enacted, negotiated, and adapted based on the context and audience. Conversely, bricolage enhances our understanding of the resourcefulness and adaptability that actors demonstrate in their roles, particularly in overcoming barriers to innovation. It elucidates how bricoleurs creatively respond to unforeseen events posed by the context while simultaneously negotiating and navigating structures that influence their actions.

XSL•FO
RenderX

Results

Overview

The total sample consisted of 109 participants, including 39 who participated in interviews and 70 who participated in participant observations. The participants represented a diverse range of stakeholders, including health care professionals, unit managers, leaders, policymakers, ICT specialists, eHealth solution providers, trainers, project staff, and technical personnel. The following sections outline the themes, accompanied by illustrative quotations.

Enacting the Role of a Bricoleur

The actors assigned the role to spearhead the innovation processes in the 3 cases were a project manager (case A), a home care service manager (case B), and a project manager assisted by trainers (case C). According to Goffman [58], roles are enacted based on available information, the situational context, and the function an actor is expected to perform. When assigned a function such as project manager (case A and C), home care service manager (case B), or trainer (case C), actors face specific expectations from their audience, which they strive to meet or engage in impression management to convey that these expectations are being met. In case B, the role of a home care service manager entailed numerous time-consuming operational responsibilities, which constrained the time and attention available for the implementation process—particularly during the COVID-19 pandemic. The geographical distance also prevented him from shadowing staff and observing their activities firsthand. These conditions made it challenging to enact either a mediator or a service specialist role. Compounding these issues, the home care service manager was new to the organization, starting the position only 2 weeks after the national lockdown in Norway. These circumstances complicated the enactment of a bricoleur role. First, the manager lacked familiarity with the organizational context and staff. Second, not having participated in the planning phase, he had limited background knowledge of the e-lock project's rationale and objectives, as reflected by the following quote: "It was decided before I started working here. So, I'm not entirely sure of the background [of the project], but it's probably to save time on key usage and to increase accessibility and safety for the users" (Respondent 1, home care service manager). Finally, the operational demands inherent in the managerial role overshadowed the implementation process, leaving little opportunity to prioritize it. This contrasts with cases A and C, where project managers were fully dedicated to the implementation process and exempt from other operational responsibilities. Although the project managers were fully dedicated to the innovation process, the way it unfolded varied across the 2 cases. In case A, the project manager and a project coordinator moved into the nursing home during the initial 2 weeks of implementation, providing continuous on-site support and guidance. After this period, they remained available via telephone and email and maintained an office next door to ensure ongoing support and rapid problem-solving. Due to poor collaboration with the ICT department in the municipality, the project manager and coordinator were also compelled to take on responsibilities as ICT specialists. This additional responsibility led the project manager and the coordinator to enact roles of service specialists, which had many advantages for the ongoing process, as expressed by a staff member:

> *I think it would have been difficult without them. It makes the workday easier. We spend less time on frustration, because, well, technology isn't my field, you know. My field is actually healthcare. But having those who are a support function for technology somehow makes my day easier.* [Respondent 2, nurse]

As the patient warning system and digital supervision were implemented directly within the nursing home, the service specialists remained in proximity to the site of action. The continuous presence of the project manager and coordinators, who engaged in shadowing staff and observing daily routines, enabled them to adopt dual roles as both service specialists and mediators gradually. In doing so, they effectively bridged the gap between backstage and the front stage. Drawing on their familiarity with the context and the actors involved, along with their acquired ICT competencies, they gradually enacted roles as bricoleurs—constructing, repairing, and maintaining the health care professionals' performance so they could do their "actual job" (Respondent 2, nurse). The project team also trained a group of superusers, who in turn took on roles as service specialists and actively supported bricolage activities by drawing on their digital competence.

Although the project manager in case C was committed to the innovation process, the physical location—outside the hospital in a separate building—created a spatial distance from the implementation site. Furthermore, the affiliation with a research and innovation unit, rather than the hospital units or the municipality, meant that the project manager lacked an established foothold within the everyday workplace dynamic. These conditions challenged the ability to enact a role as a mediator or service specialist, as the project manager was not embedded in the daily routines of the health care staff and thus had limited access to backstage information, such as training needs and perceptions about the video consultations. Enacting roles as mediators and bricoleurs was also challenging for the unit managers when priorities competed, tasks were incompatible, and responsibilities conflicted. As expressed by a unit manager: "We were a COVID unit, so we already had plenty to just ... work through, so the idea of creating new ideas and having the time to implement new systems ... Well, the staff was also overloaded. You reach a certain limit where you can't take in any more" (Respondent 3, hospital unit manager). However, in some of the units, the trainers were able to step into roles as service specialists and bricoleurs when they were physically present in the hospital to support the setup of video consultations. Their role as trainers granted them access to the backstage where health care professionals prepared for their frontstage performances. This backstage access enabled the trainers to observe real-world challenges and tailor both the training and video consultations to the specific needs of each unit. Drawing on their ICT expertise, they engaged in bricolage, creatively assembling and adjusting tools and practices to support the health care professionals in delivering care. The combination of service specialist and bricoleur roles illustrates how these actors not only facilitated the technical

XSL•FO
RenderX

implementation but also helped maintain the integrity of the health care professionals' frontstage performance. As reflected by the hospital director:

> *Every time we introduce a new ICT or digitalisation product, we turn the experts into novices. They deal with new technology, altering patient-health professional relationships [...]. If the doctor is struggling with an ICT system, they lose some respect in the eyes of the patients. Because then the patient sees them fumbling around and may think they are equally clumsy with all their professional expertise. [...] How can we ensure our healthcare professionals are not novices in ICT but at least able to use it effectively, so they don't come across as incompetent or anything like that? When healthcare professionals experience technical problems, they tend to revert to what they are accustomed to and can handle better.* [Respondent 4, hospital director]

This quote highlights the delicate balance between adopting eHealth solutions and maintaining a professional identity. Trainers who acted as bricoleurs helped preserve this balance by ensuring that health care professionals could maintain confidence and competence in their frontstage roles, even when navigating unfamiliar eHealth solutions.

The 3 cases show that the ability to enact the role of bricoleurs depends on actors' experiential knowledge, proximity to the implementation site, and a dedicated focus on implementation, which makes backstage dynamics more accessible.

Performing Bricolage to Transform Barriers Into Drivers

Goffman [58] distinguishes between 2 models of behavior: the real and the contrived. The real is regarded as a genuine performance, which is not consciously assembled, but an unintended product of an actor's spontaneous reaction to the actual situation at hand. Contrived performances, on the other hand, are regarded as carefully constructed, with each artificial element added one by one, since the behavior is not reacting to any actual situation. How one responds spontaneously to unforeseen events determines whether the performance is characterized by bricolage or a more systematic and planned engineering approach to overcome barriers [57]. Refer to Table 1 for an overview of the drivers and barriers across the 3 cases.

XSL•FO
RenderX

Table . Cross-case synthesis of structural drivers and barriers of bricolage.

Structural factor	Drivers	Barriers
Organizational structure	Colocation and small teams enable rapid feedback and access (A); proximity in small municipalities fosters collaboration (A, B).	Silos and interorganizational differences hinder coordination (A, C); concurrent large IT projects (electronic health records) compete for attention (C); geographic dispersion (A, B, C); role overload in small municipalities (A, B).
Leadership and governance	Supportive management and dedicated project leadership drive progress (A, B, C).	Shifting leadership and lack of clear ownership delay implementation (B, C); operational pressures crowd out strategic work (C).
Political framework	Alignment with local or regional or national priorities; digitalization targets create mandate (A, B, C).	—[a]
Resources	Adequate funding and equipment (A, B, C); stable workforce (A, B); superusers support adoption (A).	Staffing shortages, turnover, and crises (eg, COVID-19, strikes) impede adoption (C); lack of designated superusers (B, C).
Cultural norms and values	Positive work culture (A), and perceived expectations to use technology motivate uptake (A, B, C).	Strong patient-safety ethos and professional identity can slow change (A, B, C); small-community dynamics can amplify resistance (A).
Training and competence development	Needs-based, one-on-one, learning-by-doing, and follow-up (A, B, C) are effective; dedicated trainers also provide support (A, C).	24/7 operations complicate scheduling (A, B, C); generic courses without local tailoring are less effective; trainer distance and high turnover reduce retention (C).
Infrastructure and technology	Technology perceived as applicable or simple supports use (A, B); technology used proximate to superusers or project management supports use (A).	The technology is implemented in settings that lacked proximity to superusers and management (B, C). Technology perceived as surveillance (A), time-consuming (B, C), poorly functioning (C), and not beneficial (B, C) hinders its use. Poor set-up (A, C) and lack of deimplementation (B) create friction.
Communication systems	Multiple channels (meetings, emails, direct access to project leads) aid communication (A, B, C).	Weak interdepartmental and sectoral channels (A, C) and 24/7 staffing patterns (A, B, C) limit the information flow; the existing communication system between sectors hinders the development of a new communication system (C).
Daily relations and collaboration	Colocation and bridging roles (eg, unit management and superusers) improve collaboration (A).	Geographic dispersion (A, B, C) and lack of cross-sector workflows impede collaboration (A, C).
Quality assurance and evaluation routines	Routine monitoring (B, C) and frequent informal evaluation support technology adoption (A).	Monitoring can be perceived as surveillance; the absence of informal feedback loops reduces responsiveness (B, C).

[a]Not applicable.

By being proximate and shadowing the staff, the project manager in case A was present when unforeseen events happened. For example, when a ghost appeared on the patient monitoring system, as elicited by the project manager:

> *We set up [digital supervision] for a user [and] it was supposed to alert us if the person got out of bed. [...] It's nighttime, and the user is restless, so a night shift comes in [...]. She sits on the edge of the bed to calm the patient down. Then she stands up, which triggers an alert for getting out of bed because she's probably been sitting on the bed for too long. Due to our poor Wi-Fi, the signal is delayed. [...] Then, a few minutes pass, and an alert goes off on the mobile phones that the bed is abandoned. An anonymized picture of a man appears. Then, the others panic and rush in, wondering who's here, and they get the idea that this is a ghost [...]. They find out that the woman's late husband is in the room. Because he lived at the nursing home before.* [Respondent 5, project manager]

One of the health care professionals:

> *[...] There's probably much more between heaven and earth than we see, but I'm unsure if it would affect patient monitoring. [...] It's sometimes a profession shrouded in superstition and stories. [...] We are often close to death, right? So maybe it's a bit natural to become a bit superstitious.* [Respondent 2, nurse]

Even though the project manager could not convince the staff that the ghost appeared due to a slow Wi-Fi signal, responding constructively to this event spontaneously required an understanding of the context and the actors involved, as well as technical competence to grasp how cultural and emotional

XSL•FO
RenderX

factors intersect with technological implementation. The project manager also participated in staff meetings to evaluate the use of the patient monitoring system, ensuring that issues were raised and resolved. The training was conducted informally and continuously, based on immediate needs:

> *We simply have to practice. I have to show the staff and test. I don't know how often I've been on the floor to demonstrate a fall. [...] I have to show them and have them try it themselves. We constantly have to repeat, demonstrate, you know ... "Now, I triggered a violence alarm. What do you do then?" And it's like ... recreating scenarios.* [Respondent 5, project manager]

By being present and practicing in the situation, the project manager made it challenging for the staff to conceal actions or project desired impressions. This led the staff to use the patient warning system and digital supervision even when they were not fully confident in operating it. They either invited the project manager backstage for assistance or proceeded without the necessary skills, which created technical challenges and unforeseen events. These events, however, functioned as dramaturgical disruptions that exposed latent vulnerabilities in the performances of health care professionals. Once these issues were made visible, the project manager could implement improvisational strategies using the available resources.

In contrast to case B, where the e-locks were used in service users' homes, this made it difficult for the home care service manager to maintain proximity and shadow health care professionals during their work. Consequently, the home care service manager was not exposed to many unforeseen events caused by technology. The e-lock system log was used to monitor use, but it did not provide sufficient information to facilitate bricolage. As a result, the home care service manager remained unaware of how the staff perceived the e-locks. According to the staff, they only used the e-locks about 50% of the time. One staff member had experienced the e-lock malfunctioning once, so they always carried the regular keys "just in case" (Respondent 6, nurse). Staff members did not convey their experiences and concerns. Consequently, the management remained unaware of these issues. As reflected in the quote from the home care service manager below:

> *It was straightforward and a truly positive thing. It probably hasn't caused many obstacles. Of course, some actors might have hesitated a bit more to carry it out, but it's not like they have rallied others, because it's so simple. So I haven't heard that anyone has really resisted or that there has been noise around this technology. I would have known if there had been something.* [Respondent 1, home care service manager]

This quote reflects that when new or inexperienced leaders are given formal authority over more experienced staff members, the formally empowered actor often enacts a role of symbolic dominance. In contrast, the staff members are the ones who truly run the show [58]. This discrepancy between formal authority and practical influence served as a barrier to performing bricolage. Transforming this barrier to a driver requires strategies to gain backstage access. Because 24/7 operations made formal training difficult, he improvised by installing an e-lock on his office door, allowing health care professionals to practice at their convenience. This arrangement enabled him to remain close to the action and shadow staff during their practice sessions—providing a form of backstage access. These efforts ensured that staff acquired the necessary competencies to operate the e-lock. The staff found the training sufficient for using the e-locks, as remarked: "It wasn't that difficult" (Respondent 6, nurse). However, the staff questioned the usefulness and the functionality of the e-locks, as reflected in the following exchange:

> *Respondent 6, nurse: If [the e-lock] works, it's certainly easier.*
>
> *Respondent 7, nurse: Hmm ... I'm not so sure about that.*
>
> *Respondent 8, nurse: If it works.*
>
> *Respondent 7, nurse: It's very easy just to find the right key and unlock it.*
>
> *Respondent 8, nurse: Yeah, especially if it's raining.*

As such, this performance preparation training facilitated by the home care service manager had a limited effect, as e-locks were still not consistently used in service users' homes. The home care service unit manager emphasized the importance of managing explicit resistance happening on the frontstage: "Those who I thought might give the most resistance, they were actually the first ones I gave training to, and maybe followed up a little extra" (Respondent 1, home care services manager). However, silent backstage resistance might be a barrier more challenging to overcome than explicit frontstage resistance. In case C, many units struggled to implement video consultations for discharge conferences; however, a few units succeeded by engaging in adaptive, improvisational practices. In these units, the technology was not merely adopted as prescribed but reinterpreted and repurposed for alternative use—such as municipal rehabilitation supervision or admission meetings with relatives—illustrating a shift from the scripted "frontstage" plan to context-sensitive enactments. Training similarly moved backstage, taking place within the units rather than in formal classroom settings, fostering unforeseen events and situated learning. The trainers emphasized the importance of readily available support to answer questions and assist with new technology, as stated in the following quote from one of the trainers, highlighting the need for continuous and organized follow-up training in the units:

> *[...] Sometimes I think we could try to be more organised in training in the units. I'm not sure if there should be someone who becomes a superuser or is somehow responsible for it [...]. If one were to mention a missing link, I think it's the transition from having received the training to becoming an integrated part of the unit. That's where we could do something more. [...] It works really well if they have had training with us, and then we have been [in the units]* [Respondent 9, trainer]

The performance of bricolage in these units was facilitated by the trainers' presence backstage, their flexibility, and consistent

XSL•FO
RenderX

availability for support, as well as their provision of ad hoc problem-solving for unforeseen events. As the 3 cases show, responding to unforeseen events in a systematic and planned manner can be challenging. To effectively perform bricolage, it is essential to understand and act upon the context and engage with the involved actors. Understanding the context derives from experiential knowledge, the function assigned, and access to backstage insights. Performing bricolage is not just about creatively navigating unforeseen events using the resources at hand; it is also about being mindful of the interactions and relationships with other actors, ultimately shaping the experience and outcome for everyone involved in the innovation process.

Features of a Bricoleur

Several common features were observed among the actors who performed bricolage. They were familiar with the local work context, able to move fluidly between technical and clinical domains, and demonstrated a readiness to improvise and adapt eHealth solutions to meet the practical needs of health care professionals. As the project leader in case A noted: "Now I am the ICT department. I go into the computer cabinet myself and connect things" (Respondent 5, project manager). Another feature of the bricoleurs was their assigned role, such as project leader or trainer, which required full dedication to the innovation process. This meant they could be consistently available, responsive, and actively worked to meet the expectations of both the leaders and the health care professionals they supported, as emphasized by one of the trainers in case C: "It doesn't matter how much time has passed; I always make myself available" (Respondent 9, trainer). The physical proximity to the implementation site and involved actors enabled frequent, informal interaction that granted them access to backstage areas.

In contrast, actors who lacked the essential features for performing bricolage, such as those who were physically distant from the implementation site or held a position that involved other tasks and responsibilities, struggled to enact roles as bricoleurs. Without embeddedness in the local work context or access to backstage interactions, for example, by being new to the organization, these actors were less able to understand the nuances of everyday practices or respond to emerging needs. For instance, this resulted in staff perceptions of management being "unresponsive to feedback" (Respondent 7, nurse) in case B or feeling surveilled as in case C:

> *[The management] is definitely monitoring whether we are conducting these video consultations. We figured that out. We didn't actually know. But they are keeping an eye on us. [...] What is the reason for that? Well, I think it's because we first tried to say ... "Do we really need this? And should we do it now?" It was very emphatically stated that we should. Also, somewhere between the lines, it says that most consultations should be conducted over video. We have internally determined that as long as we say half/half, we'll have to see if they eventually knock on the door and say "no, no, no, now there are too many physical meetings every day"* [Respondent 10, nurse]

This quote captures the management's attempt to gain backstage access without success, while the nurse and her team strive to protect this backstage area. Consequently, actors without bricoleur features struggle to gain backstage information and are left to resort to more systematically planned approaches, limiting their contributions to, for example, formal training sessions or top-down strategies with limited impact on everyday practices among health care professionals.

Key features of bricoleurs are their familiarity with the local context, the ability to bridge technical and clinical domains, and consistent presence at the implementation site. Together with their dedicated roles and informal access to backstage dynamics, they are enabled to improvise and adapt eHealth solutions to meet practical needs.

Interactions Between Structures and Actors

In all 3 cases, the implementation of eHealth solutions was aligned with local, regional, and national priorities. The initiatives were supported by committed leadership, adequate funding, and a shared perception among actors that the use of eHealth solutions was both expected and necessary. However, these drivers carried limited weight when other structural factors (see Table 1 for an overview of barriers and drivers across cases), such as shortages of health care professionals, posed substantial barriers to implementation, as expressed by a member of the executive hospital management in case C, which was also the case featuring the most complexity:

> *We want to do a lot, but we don't have the resources. [...] We hear: " You get a lot of money; just get started." On the other hand, we simply don't have anyone who can [do it]. We clearly see that there is a need for these things. [...] It creates a vicious cycle spiralling downward, and we need to turn it around somehow.* [Respondent 4, hospital director]

Other barriers to bricolage included organizational silos (case C), interorganizational (case A) differences, shifting leadership (cases B and C), the pressures of 24/7 operations (all cases), high turnover rates (case C), the absence of superusers (cases B and C), and a strong patient-safety ethos (all cases). Additional barriers stemmed from the maintenance of professional identities (all cases), a lack of deimplementation practices (case B), the absence of informal feedback loops, and geographic dispersion among the actors. Technology itself also posed barriers, particularly when it was perceived as a form of surveillance (case A), as time-consuming (cases B and C), or as offering limited benefits to clinical practice (cases B and C). Under such conditions, performing bricolage becomes a challenging task.

Despite structural complexity and barriers, bricolage occurred across all 3 cases. Actors were not merely shaped by these structures; they actively worked within and upon them. Actors performed bricolage, particularly when structural drivers created openings for creative and adaptive problem-solving. While organizational silos, 24/7 operations, and technological infrastructure influenced what was possible, actors leveraged their assigned functions and interactions, using their expertise to adapt, reshape, and, at times, transform these structures. Key drivers of bricolage included colocation and proximity between

XSL•FO
RenderX

actors, which facilitated informal interactions and backstage access. The presence of supportive management, dedicated project leads, and a stable workforce created a foundation for continuity and responsiveness. A high number of superusers, along with tailored training that addressed the specific needs of each unit, enabled health care professionals to engage more confidently with eHealth solutions. Follow-up activities, dual roles (where project leaders also serve as ICT support), routine monitoring, and frequent informal evaluations further supported backstage access and, consequently, adaptive and creative problem-solving.

Discussion

Principal Findings

This study explored how actors involved in implementing and sustaining eHealth solutions enact roles as bricoleurs and perform bricolage to transform barriers into drivers of innovation. Using Goffman's [58] dramaturgical approach and Lévi-Strauss's [57] concept of bricolage, we examined how structural factors and actors' agency interact in complex innovation processes.

This cross-case analysis showed that the ability to enact the role of a bricoleur increased when actors were assigned a function dedicated to the innovation process, thereby being freed from operational demands. For example, being assigned the function of a trainer (case C) provided dedication, but also a mandate and a set of expectations to be fulfilled. Further, the enactment of the bricoleur role depended on actors' experiential knowledge and proximity to the implementation site, which facilitated access to backstage dynamics. The project team in case A is a good example that demonstrates how proximity and presence facilitated backstage access where routines were rehearsed, problems surfaced, and informal knowledge was shared. This backstage access led to a deeper understanding of the everyday challenges faced by health care professionals. Due to the project teams' continuous presence, health care professionals had to use eHealth solutions despite technical uncertainty, which in turn led to genuine performances, flaws, and unforeseen events. These findings resonate with previous research on bricolage [45,47,49] and add to the incomplete understanding of how the bricoleur role can be identified, enabled, and enhanced [26,36,38,49,50].

Bricolage emerged as a spontaneous and improvised response to unforeseen events, as found in previous research [47,49]; however, our research expands the literature by showing that bricolage builds on the bricoleur's experiential knowledge and dual expertise in clinical and technical domains, as was apparent in cases A and C. In these cases, the bricoleurs could transform barriers such as 24/7 operational demands and technological issues into drivers through, for example, ad-hoc, tailored training and by adapting solutions in contextually appropriate ways. Previous research has highlighted that barriers are dynamic and may be transformed into drivers if addressed appropriately [36-40]. We propose that bricolage offers a promising approach for facilitating such transformations.

Without backstage access, actors are left to rely on formal and systematic plans, strategies, and communication, as was the case for the project manager in case C and the home care service manager in case B. In both cases, silent resistance in the form of subtle, unspoken disengagement with the eHealth solutions hindered the implementation. Unlike explicit resistance, silent resistance may be difficult to detect and address, especially when actors lack backstage access. While previous research has emphasized how interactional barriers can hinder innovation processes and underscored the importance of collaboration [25,32,34,36,38,41-43] and boundary-spanning roles [26,31,38,44] in mitigating such barriers, our study adds nuance by demonstrating how bricoleurs can mitigate interactional barriers through backstage access.

Despite structural and contextual constraints, bricolage emerged across all cases. In case A, due to collaboration challenges with the ICT department, the project manager acquired ICT competencies and assumed responsibility for ICT-related tasks, which ultimately proved beneficial to both the staff and the innovation process. Rather than being passively shaped by structural conditions, actors can interact with and actively influence the conditions. This highlights how bricolage is not only resourceful but also interactive and performative, enabled by the interplay between agency and context, offering insight into the previously noted uncertainty regarding how the context facilitates bricolage work [46].

Our findings align with and enrich existing implementation frameworks such as the Non-adoption, Abandonment, Scale-up, Spread, and Sustainability framework [25,68] and Consolidated Framework for Implementation Research [59]. While these frameworks emphasize the importance of context, complexity, and actor engagement, they lack a detailed account of the microlevel improvisations that sustain implementation in practice. We contribute to an underexplored area [24,30,50] by proposing that bricolage offers a complementary mechanism that explains how actors navigate complexity not by eliminating it, but by working within and around it. As such, we suggest that bricolage can be conceptualized as a mechanism that links structural factors with microlevel practices (eg, improvisation, adaptation, and role enactment). This perspective can inform the design of implementation strategies that are more responsive to local contingencies and the agency of actors.

Our study aligns with previous literature on the characteristics of actors driving innovation, highlighting the importance of experience and ICT-related skills [26,35]. However, our study expands the literature by demonstrating that features of a bricoleur also encompass proximity, dedication to the implementation process, dual-expertise, and access to backstage dynamics. These features were most evident in cases A and C. As such, the findings of our study demonstrate that bricolage is not a spontaneous phenomenon or a matter of individual creativity [35]; rather, it is contingent upon various structural factors and interactions with actors. Our contribution offers nuanced, empirically grounded, context-sensitive illustrations that enrich existing understandings of bricolage and its role in innovation processes [45-49]. Although prior studies have conceptualized bricolage [45-49], our research advances the field by identifying the conditions and actors that enable its

XSL•FO
RenderX

practical enactment. Our research advances the field by empirically demonstrating how specific structural and contextual conditions, such as proximity and dedicated roles, enable the positioning of actors to enact roles as bricoleurs. Through detailed cross-case analysis guided by dramaturgy and bricolage, we have identified key features of bricoleurs and demonstrated how backstage access enables them to respond to unforeseen events in real time. This responsiveness creates opportunities to transform barriers into drivers, thereby sustaining innovation processes in complex health care settings.

In times of austerity, bricolage—an approach that uses available resources—may be notably applicable. Specifically, given that public sector innovation processes are often fraught with challenges and have a high failure rate [22-26], and as new approaches to advance, public sector innovation processes are asked for [65-67].

Limitations and Future Research

This study used a qualitative multiple-case design with methodological triangulation. While this approach enabled a rich and nuanced understanding of the implementation of eHealth solutions, the findings are grounded in 3 specific cases in a Nordic welfare context. As such, transferability to other settings may be limited. However, the implementation of eHealth solutions in health care is highly relevant across a wide range of countries. eHealth solutions tend to influence professional roles, interactions, and organizational structures in ways that are difficult to predict and anticipate [32,38,42,43,69]. These dynamics apply regardless of context. While our findings are not intended to be generalized, the conditions that enable successful bricolage—such as proximity and dual expertise—may offer valuable insights across diverse contexts.

Future research could build on this work by using mixed methods designs or larger-scale comparative studies to examine how bricolage unfolds across different health care systems or policy environments. Additionally, further exploration of silent resistance, including its manifestations, consequences, and strategies for mitigation, could deepen our understanding of the subtle dynamics that shape innovation processes in complex health care settings.

Conclusions

This study contributes to the understanding of innovation processes in the public sector by illuminating how actors navigate complex implementation processes through bricolage. By integrating dramaturgy with bricolage, we offer a novel analytical lens that captures both the performative and improvisational dimensions of innovation work. This theoretical pairing enabled us to explore how roles are dynamically enacted and adapted in response to structural and contextual factors, as well as emergent challenges.

Our findings show that actors become bricoleurs not merely by individual traits, but through a combination of contextual knowledge, proximity to the implementation site, and access to backstage dynamics. These conditions enable bricoleurs to improvise, adapt, and sustain innovation efforts in ways that formal strategies alone may not achieve. Our study provides detailed, context-sensitive illustrations that enrich existing theories of innovation and implementation. Recognizing and supporting conditions for bricolage may help design more adaptive, responsive, and sustainable strategies. We demonstrate how bricolage operates as a mechanism that links structural factors with microlevel improvisation, offering a valuable complement to established implementation frameworks. The theoretical pairing further clarifies how social expectations, role performances, and backstage interactions shape the conditions under which bricolage can occur.

Acknowledgments

The authors thank the participants who took part in their study. The authors also thank the project management in the Digital and Innovation Skills Helix (DISH) project and the funders of the DISH project: Erasmus+ Program of the European Union, Key Action 2 Cooperation for Innovation, and the Exchange of Good Practices—Sector Skills Alliances. The authors declare the use of generative artificial intelligence in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI (generative artificial intelligence) tools under full human supervision: proofreading, editing, and translation. The GAI tool used was ChatGPT-4. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was funded by the Norwegian Ministry of Education and Research and Western Norway University of Applied Science. The funders had no involvement in the study design, data collection, analyses, interpretation of results, or the writing of the manuscript.

Data Availability

The data analyzed during this study are not publicly available due to confidentiality reasons, but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: SE (lead), CØ (equal), AMD (supporting)

XSL•FO
RenderX

Data curation: SE (lead), CØ (supporting), AMD (supporting)
Formal analysis: SE (lead), CØ (equal), AMD (equal)
Funding acquisition: CØ
Investigation: SE (lead), CØ (equal), AMD (equal)
Methodology: SE (lead), CØ (supporting), AMD (supporting)
Validation: SE (lead), CØ (equal), AMD (equal)
Visualization: SE
Writing—original draft: SE (lead), CØ (supporting), AMD (supporting)
Writing—review and editing: SE (lead), CØ (supporting), AMD (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of data collection.
[DOCX File, 29 KB - jmir_v28i1e79999_app1.docx]

Multimedia Appendix 2
Interview and focus group guide.
[DOCX File, 17 KB - jmir_v28i1e79999_app2.docx]

Multimedia Appendix 3
Participant observation guide.
[DOCX File, 19 KB - jmir_v28i1e79999_app3.docx]

References

1. Recommendations on digital interventions for health system strengthening. WHO guideline. World Health Organization. 2019. URL: https://iris.who.int/server/api/core/bitstreams/c3c53f30-23cc-48d0-a3b5-c05ddd7f5349/content [accessed 2025-12-20]
2. Norges Offentlige Utredninger. Tid for handling. Personellet i en bærekraftig helse-og omsorgstjeneste. : Helse-og Omsorgsdepartementet; 2023 URL: https://www.regjeringen.no/contentassets/337fef958f2148bebd326f0749a1213d/no/pdfs/nou202320230004000dddpdfs.pdf [accessed 2025-12-12]
3. Regeringen, Kommunernes Landsforening, Danske regioner. Digitalisering, der løfter samfundet: den fællesoffentlige digitaliseringsstrategi 2022-2025. sammen om en digital fremtid. 2022 URL: https://fm.dk/media/vvkmho33/digitalisering-der-loefter-samfundet_den-faellesoffentlige-digitaliseringsstrategi-2022-2025_web.pdf [accessed 2025-12-12]
4. e-Health—making healthcare better for European citizens: an action plan for a European e-Health area. : European Commission; 2004 URL: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2004:0356:FIN:EN:PDF [accessed 2025-12-12]
5. Peine A, Faulkner A, Jæger B, Moors E. Science, technology and the 'grand challenge' of ageing—understanding the socio-material constitution of later life. Technol Forecast Soc Change 2015 Apr;93:1-9. [doi: 10.1016/j.techfore.2014.11.010]
6. Mort M, Roberts C, Callén B. Ageing with telecare: care or coercion in austerity? Sociol Health Illn 2013 Jul;35(6):799-812. [doi: 10.1111/j.1467-9566.2012.01530.x] [Medline: 23094945]
7. Procter R, Greenhalgh T, Wherton J, Sugarhood P, Rouncefield M, Hinder S. The day-to-day co-production of ageing in place. Comput Supported Coop Work 2014 Jun;23(3):245-267. [doi: 10.1007/s10606-014-9202-5]
8. Indenrigs- og Sundhedsministeriet. Robusthedskommissionens anbefalinger. : Sundhedsministeriet; 2023 URL: https://www.ism.dk/Media/638336462586551242/Robusthed-Samlet-Rapport-TILG.pdf [accessed 2025-12-12]
9. Osborne SP, Brown L. Innovation, public policy and public services delivery in the UK. The word that would be king? Public Adm 2011 Dec;89(4):1335-1350. [doi: 10.1111/j.1467-9299.2011.01932.x]
10. State of the Innovation Union 2015. : European Commission; 2015 URL: https://op.europa.eu/en/publication-detail/-/publication/0487b7b9-b5d6-11e5-8d3c-01aa75ed71a1/language-en [accessed 2025-12-12]
11. Boogerd EA, Arts T, Engelen LJ, van de Belt TH. "What is eHealth": time for an update? JMIR Res Protoc 2015 Mar 12;4(1):e29. [doi: 10.2196/resprot.4065] [Medline: 25768939]
12. Norges Offentlige Utredninger. Innovasjon i omsorg. : Helse-og Omsorgsdepartementet Oslo; 2011 URL: https://www.regjeringen.no/contentassets/5fd24706b4474177bec0938582e3964a/no/pdfs/nou201120110011000dddpdfs.pdf [accessed 2025-12-12]
13. World report on ageing and health. World Health Organization. 2015. URL: https://www.who.int/publications/i/item/9789241565042 [accessed 2025-12-11]

XSL•FO
RenderX

14. Green paper on ageing: fostering solidarity and responsibility between generations. : European Commission; 2021 URL: https://commission.europa.eu/system/files/2021-06/green_paper_ageing_2021_en.pdf [accessed 2025-12-12]
15. Meld. St. 24. (2022-2023). Fellesskap og meistring. Bu trygt heime. : Helse- og Omsorgsdepartementet; 2023 URL: https://www.regjeringen.no/contentassets/a8280e2548c04d3ea6898078480bfa0c/nn-no/pdfs/stm202220230024000dddpdfs.pdf [accessed 2025-12-12]
16. Regeringen. Regeringens ældrereform. Du bliver aldrig for gammel til at have det godt. : Social-, Bolig- og, Ældreministeriet; 2024 URL: https://regeringen.dk/media/ec0nhyiw/regeringens_aeldreudspil_jan2024.pdf [accessed 2025-12-12]
17. Berge MS. Telecare—where, when, why and for whom does it work? A realist evaluation of a Norwegian project. J Rehabil Assist Technol Eng 2017;4:2055668317693737. [doi: 10.1177/2055668317693737] [Medline: 31186924]
18. Dugstad J, Eide T, Nilsen ER, Eide H. Towards successful digital transformation through co-creation: a longitudinal study of a four-year implementation of digital monitoring technology in residential care for persons with dementia. BMC Health Serv Res 2019 Jun 10;19(1):366. [doi: 10.1186/s12913-019-4191-1] [Medline: 31182093]
19. Huter K, Krick T, Domhoff D, Seibert K, Wolf-Ostermann K, Rothgang H. Effectiveness of digital technologies to support nursing care: results of a scoping review. J Multidiscip Healthc 2020;13:1905-1926. [doi: 10.2147/JMDH.S286193] [Medline: 33328736]
20. Lokal og digital. Et sammenhængende danmark. Den fælleskommunale digitaliseringsstrategi og handlingsplan 2016-2020. : Kommunernes Landsforening; 2016 URL: https://www.kl.dk/media/l4daobww/lokal-og-digital-et-sammenhaengende-danmark.pdf [accessed 2025-12-12]
21. Sundhedsstrukturkommissionens rapport. Beslutningsgrundlag for et mere lige, sammenhængende og bæredygtigt sundhedsvæsen. : Indenrigs- og Sundhedsministeriet; 2024 URL: https://www.ism.dk/Media/638545635292256419/Hovedrapport-tilg%c3%a6ngelig-fil.pdf [accessed 2025-12-12]
22. Cresswell KM, Bates DW, Sheikh A. Ten key considerations for the successful implementation and adoption of large-scale health information technology. J Am Med Inform Assoc 2013 Jun;20(e1):e9-e13. [doi: 10.1136/amiajnl-2013-001684] [Medline: 23599226]
23. Mair FS, May C, O'Donnell C, Finch T, Sullivan F, Murray E. Factors that promote or inhibit the implementation of e-health systems: an explanatory systematic review. Bull World Health Organ 2012 May 1;90(5):357-364. [doi: 10.2471/BLT.11.099424] [Medline: 22589569]
24. Standing C, Standing S, McDermott ML, Gururajan R, Kiani Mavi R. The paradoxes of telehealth: a review of the literature 2000–2015. Syst Res 2018 Jan;35(1):90-101. [doi: 10.1002/sres.2442]
25. Greenhalgh T, Maylor H, Shaw S, et al. The NASSS-CAT tools for understanding, guiding, monitoring, and researching technology implementation projects in health and social care: protocol for an evaluation study in real-world settings. JMIR Res Protoc 2020 May 13;9(5):e16861. [doi: 10.2196/16861] [Medline: 32401224]
26. Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O. Diffusion of innovations in service organizations: systematic review and recommendations. Milbank Q 2004;82(4):581-629. [doi: 10.1111/j.0887-378X.2004.00325.x] [Medline: 15595944]
27. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023 Sep 18;6(1):161. [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
28. Scott Kruse C, Karem P, Shifflett K, Vegi L, Ravi K, Brooks M. Evaluating barriers to adopting telemedicine worldwide: a systematic review. J Telemed Telecare 2018 Jan;24(1):4-12. [doi: 10.1177/1357633X16674087] [Medline: 29320966]
29. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146. [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
30. Gagnon MP, Desmartis M, Labrecque M, et al. Systematic review of factors influencing the adoption of information and communication technologies by healthcare professionals. J Med Syst 2012 Feb;36(1):241-277. [doi: 10.1007/s10916-010-9473-4] [Medline: 20703721]
31. Sligo J, Gauld R, Roberts V, Villa L. A literature review for large-scale health information system project planning, implementation and evaluation. Int J Med Inform 2017 Jan;97:86-97. [doi: 10.1016/j.ijmedinf.2016.09.007] [Medline: 27919399]
32. Brewster L, Mountain G, Wessels B, Kelly C, Hawley M. Factors affecting front line staff acceptance of telehealth technologies: a mixed-method systematic review. J Adv Nurs 2014 Jan;70(1):21-33. [doi: 10.1111/jan.12196] [Medline: 23786584]
33. Lau R, Stevenson F, Ong BN, et al. Achieving change in primary care—causes of the evidence to practice gap: systematic reviews of reviews. Implement Sci 2015 Dec;11(1):40. [doi: 10.1186/s13012-016-0396-4]
34. Cinar E, Trott P, Simms C. A systematic review of barriers to public sector innovation process. Public Manage Rev 2019 Feb;21(2):264-290. [doi: 10.1080/14719037.2018.1473477]
35. De Vries H, Bekkers V, Tummers L. Innovation in the public sector: a systematic review and future research agenda. Public Adm 2016 Mar;94(1):146-166. [doi: 10.1111/padm.12209]
36. Torugsa N, Arundel A. Complexity of Innovation in the public sector: a workgroup-level analysis of related factors and outcomes. Public Manage Rev 2016 Mar 15;18(3):392-416. [doi: 10.1080/14719037.2014.984626]

XSL•FO
RenderX

37. Borins SF. The persistence of innovation in government. : Brookings Institution Press in Collaboration with Ash Center for Democratic Governance and Innovation; 2014 URL: https://businessofgovernment.org/sites/default/files/The%20Persistence%20of%20Innovation%20in%20Government.pdf [accessed 2025-12-12]
38. Eriksen S, Dahler AM, Øye C. Improvisational theatre as a viable path? Exploring interaction antecedents in public service innovation processes attempting to implement and sustain eHealth solutions. Public Manage Rev 2025 Mar 4;27(3):817-835. [doi: 10.1080/14719037.2024.2381069]
39. Hadjimanolis A. The barriers approach to innovation. In: Shavinina LV, editor. The International Handbook on Innovation: Elsevier; 2003:559-573.
40. Nilsen ER, Dugstad J, Eide H, Gullslett MK, Eide T. Exploring resistance to implementation of welfare technology in municipal healthcare services—a longitudinal case study. BMC Health Serv Res 2016 Nov 15;16(1):657. [doi: 10.1186/s12913-016-1913-5] [Medline: 27846834]
41. Cinar E, Trott P, Simms C. An international exploration of barriers and tactics in the public sector innovation process. Public Manage Rev 2021 Mar 4;23(3):326-353. [doi: 10.1080/14719037.2019.1668470]
42. Cucciniello M, Guerrazzi C, Nasi G, Ongaro E. Coordination mechanisms for implementing complex innovations in the health care sector. Public Manage Rev 2015 Aug 9;17(7):1040-1060. [doi: 10.1080/14719037.2015.1029348]
43. Cucciniello M, Nasi G. Evaluation of the impacts of innovation in the health care sector: a comparative analysis. Public Manage Rev 2014 Jan 2;16(1):90-116. [doi: 10.1080/14719037.2013.798026]
44. Cresswell K, Sheikh A. Organizational issues in the implementation and adoption of health information technology innovations: an interpretative review. Int J Med Inform 2013 May;82(5):e73-e86. [doi: 10.1016/j.ijmedinf.2012.10.007] [Medline: 23146626]
45. Bugge MM, Bloch CW. Between bricolage and breakthroughs—framing the many faces of public sector innovation. Public Money Manage 2016 Jun 6;36(4):281-288. [doi: 10.1080/09540962.2016.1162599]
46. Fuglsang L, Sørensen F. The balance between bricolage and innovation: management dilemmas in sustainable public innovation. Serv Ind J 2011 Mar;31(4):581-595. [doi: 10.1080/02642069.2010.504302]
47. Fuglsang L. Bricolage and invisible innovation in public service innovation. J Innov Econ Manage 2010;5(1):67-87. [doi: 10.3917/jie.005.0067]
48. Gibson G, Dickinson C, Brittain K, Robinson L. Personalisation, customisation and bricolage: how people with dementia and their families make assistive technology work for them. Ageing Soc 2019 Nov;39(11):2502-2519. [doi: 10.1017/S0144686X18000661]
49. Greenhalgh T, Wherton J, Sugarhood P, Hinder S, Procter R, Stones R. What matters to older people with assisted living needs? A phenomenological analysis of the use and non-use of telehealth and telecare. Soc Sci Med 2013 Sep;93:86-94. [doi: 10.1016/j.socscimed.2013.05.036]
50. Houtgraaf G, Kruyen PM, van Thiel S. Public sector creativity as the origin of public sector innovation: a taxonomy and future research agenda. Public Adm 2023 Jun;101(2):539-556. [doi: 10.1111/padm.12778]
51. Yin RK. Case study research. In: Design and Methods: Sage Publications; 2009.
52. Flyvbjerg B. Five misunderstandings about case-study research. Qual Inq 2006 Apr;12(2):219-245. [doi: 10.1177/1077800405284363]
53. Eriksen S, Dahler AM, Øye C. The informal way to success or failure? Findings from a comparative case study on video consultation training and implementation in two Danish hospitals. BMC Health Serv Res 2023 Oct 21;23(1):1135. [doi: 10.1186/s12913-023-10163-w] [Medline: 37865741]
54. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
55. Braun V, Clarke V, Hayfield N, Terry G. Thematic analysis. In: Liamputtong P, editor. Handbook of Research Methods in Health Social Sciences: Springer; 2019:843-860.
56. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
57. Levi-Strauss C. The Savage Mind: University of Chicago Press; 1966.
58. Goffman E. The Presentation of Self in Everyday Life: Doubleday; 1959.
59. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. Implement Sci 2022 Oct 29;17(1):75. [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
60. Malterud K. Qualitative research: standards, challenges, and guidelines. Lancet 2001 Aug;358(9280):483-488. [doi: 10.1016/S0140-6736(01)05627-6]
61. Johnson C. Bricoleur and bricolage: from metaphor to universal concept. Paragraph 2012;35(3):355-372. [doi: 10.3366/para.2012.0064]
62. Baker T, Nelson RE. Creating something from nothing: resource construction through entrepreneurial bricolage. Adm Sci Q 2005 Sep;50(3):329-366. [doi: 10.2189/asqu.2005.50.3.329]

XSL•FO
RenderX

63. Nelson RE, Rodriguez-Lluesma C, Companys YE, Stinchfield BT. Contextualizing the subjectivist-objectivist debate in entrepreneurship using engineering, art, craft, and bricolage. Int Entrep Manag J 2018 Dec;14(4):999-1021. [doi: 10.1007/s11365-017-0471-6]
64. Greenhalgh T, Shaw S, Wherton J, et al. SCALS: a fourth-generation study of assisted living technologies in their organisational, social, political and policy context. BMJ Open 2016 Feb 15;6(2):e010208. [doi: 10.1136/bmjopen-2015-010208] [Medline: 26880671]
65. Paparini S, Green J, Papoutsi C, et al. Case study research for better evaluations of complex interventions: rationale and challenges. BMC Med 2020 Nov 10;18(1):301. [doi: 10.1186/s12916-020-01777-6] [Medline: 33167974]
66. Wong G, Greenhalgh T, Westhorp G, Pawson R. Realist methods in medical education research: what are they and what can they contribute? Med Educ 2012 Jan;46(1):89-96. [doi: 10.1111/j.1365-2923.2011.04045.x] [Medline: 22150200]
67. Greenhalgh T, Papoutsi C. Studying complexity in health services research: desperately seeking an overdue paradigm shift. BMC Med 2018 Jun 20;16(1):95. [doi: 10.1186/s12916-018-1089-4] [Medline: 29921272]
68. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
69. Kamp A, Hansen AM. Negotiating professional knowledge and responsibility in cross-sectoral telemedicine. Nordic J Work Life Stud 2019;9(S5):13-32. [doi: 10.18291/njwls.v9iS5.112691]

Abbreviations

e-locks: electronic door locks
ICT: information and communication technology

Edited by A Stone, A Mavragani; submitted 02.Jul.2025; peer-reviewed by N Döring, S Kuoppamaki; accepted 28.Oct.2025; published 07.Jan.2026.

Please cite as:
Eriksen S, Øye C, Dahler AM
Exploring the Dynamics of Actors, Structural Factors, and Bricolage in the Implementation and Sustainability of eHealth Solutions: Qualitative Multiple-Case Study
J Med Internet Res 2026;28:e79999
URL: https://www.jmir.org/2026/1/e79999
doi:10.2196/79999

© Susanne Eriksen, Christine Øye, Anne Marie Dahler. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

“I Want to Spend My Time Living”—Experiences With a Digital Outpatient Service With a Mobile App for Tailored Care Among Adults With Long-Term Health Service Needs: Qualitative Study Using Thematic Analysis

Heidi Holmen[1,2], RN, PhD; Erik Fosse[1,3], MD, PhD

[1]The Intervention Centre, Oslo University Hospital, Oslo, Norway

[2]Department of Nursing and Health Promotion, Faculty of Health Sciences, OsloMet—Oslo Metropolitan University, Oslo, Norway

[3]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway

Corresponding Author:
Heidi Holmen, RN, PhD
The Intervention Centre
Oslo University Hospital
Postboks 4950 Nydalen
Oslo, 0424
Norway
Phone: 47 91527700
Email: heidho@ous-hf.no

Abstract

Background: Digital health services are increasingly used in hospital-based outpatient care, offering remote monitoring, patient-reported outcomes, information sharing, and asynchronous communication. While expected to improve self-management, timeliness, and efficiency, the success of digital health interventions relies on patients’ health literacy and digital health literacy. While some research has addressed potential associations between digital health interventions and patients’ health outcomes, research on patients’ experiences remains limited.

Objective: The aim of this study was to explore and gain in-depth knowledge about the experiences of patients with chronic or long-term conditions enrolled in a 6-month digital outpatient care intervention for tailored care and health literacy.

Methods: We conducted an exploratory qualitative interview study with 17 strategically recruited adult patients with cancer, interstitial lung disease, epilepsy, or complicated pain who used a digital outpatient service for 6 months. Individual telephone interviews were conducted using a semistructured guide, transcribed verbatim, and analyzed with thematic analysis to generate codes and themes. Participants had a median age of 62 years (minimum-maximum 36-83 years), with 8 females and 9 males.

Results: The thematic analysis led to 1 main theme “Digital outpatient care as a flexible service supporting patients’ self-management,” informed by 3 subthemes “The ongoing nature of managing a chronic condition and how the digital service meet the patients’ desire for autonomy in their care,” “Digital tools flexibly address the patients’ unique needs, but reliability depends on patient interaction,” and “Digital services enhance the patients’ sense of safety through easy access to a relation with competent healthcare workers.” The themes highlight patients’ appreciation for greater flexibility in their care and their desire to self-manage with the support of easily accessible health care workers, ultimately supporting their health literacy. Patients recognized the importance of actively engaging with the digital solution to fully benefit from its opportunities and emphasized the critical role of health care workers in fostering their sense of security.

Conclusions: Digital outpatient care was experienced as flexible and supportive for patients with long-term conditions. The increased possibility of interacting with health care workers was welcomed by the patients, and the combination of flexibility, self-monitoring, and addressing concerns regarding their self-management may increase the patients experience of autonomy. As health literacy likely plays a role in patients’ ability to effectively engage with digital tools and self-manage their conditions, future research should explore how varying levels of health literacy influence these outcomes. In addition, research should address whether such digital outpatient clinics are positive for a wider range of patients, associated health outcomes, and any positive effects on a health system level.

Trial Registration: ClinicalTrials.gov NCT05068869; https://clinicaltrials.gov/ct2/show/NCT05068869

XSL•FO
RenderX

(J Med Internet Res 2026;28:e79155) doi:10.2196/79155

KEYWORDS

digital outpatient care; mHealth; patient-centered care; patient experiences; digital outpatient services

Introduction

The increasing demand for health care services is not matched by available resources, including workforce shortages [1,2]. To enhance the sustainability of health services, greater flexibility is needed to address the dynamic needs of patients with long-term conditions, aligning with patient-centered care and optimizing resource use. Digital tools have been proposed as a means to achieve this flexibility, offering functionalities such as patient monitoring, self-reporting of objective and subjective data, asynchronous communication, and video consultations [3]. These tools can help identify patients at risk for deterioration or in need of immediate care, while also detecting those stable enough to delay scheduled services. By supporting patient-centered care, digital health solutions may improve attendance at scheduled appointments, enhance symptom management, and empower patients in self-management.

Digital health care services typically enable the subjective reporting of health parameters through patient-reported outcome measures [4]. These measures can aid self-management and support health literacy [5] and facilitate communication between patients and health care workers [6]. Furthermore, digital engagement, self-monitoring, and data sharing offer several advantages for patients [3,7,8]. Recent studies suggest that digital tools in outpatient care for patients with long-term conditions, such as cancer, epilepsy, interstitial lung disease, and musculoskeletal pain, can prevent complications, encourage patient engagement, and increase confidence and autonomy [9-12]. However, research on multicomponent digital solutions—encompassing patient-reported outcome measures, asynchronous messaging, remote monitoring, patient notifications, and video consultations—is limited. In addition, one important prerequisite exists for patients to benefit from digital health services, namely, that the patients use the services as intended. Patients' use of a digital health services might rely on their health literacy and ability to self-manage, and a certain level of digital skills and digital health literacy is also needed [5]. Health literacy provides a theoretical lens for this study, as it is a foundational factor influencing how patients access, understand, and use digital health services. Defined as "the cognitive and social skills that determine the motivation and ability of individuals to gain access to, understand, and use information in ways which promote and maintain good health" [13], health literacy not only affects patients' capacity for self-management but also plays an integral role in their engagement with digital tools. Often, higher health literacy is associated with an increased benefit of digital health interventions [14], while lower health literacy is linked to poorer health outcomes and less digital solution usage [15,16]. Furthermore, adequate health literacy is key to optimizing the use of digital health solutions, as it is suggested to enhance self-management in chronic conditions [5,16-18]. Although several systematic reviews have explored how to define, measure, and understand health literacy and digital health literacy [17,19-24], linking this knowledge to the experience of patients participating in digital health interventions remains scarce. Digital health literacy has been suggested as a super determinant of health comprising more than simply digital literacy and health literacy [25]; thus, exploring the experiences of patients with a digital health intervention supporting flexibility and health literacy might contribute with knowledge on how to research and design more effective, inclusive, and patient-centered digital health solutions that address diverse needs and promote equitable health outcomes.

While research on digital health services expands, research on patients' experiences of participating and making use of integrated digital solutions in outpatient care is lacking. The meaningfulness of a digital health solution depends on factors such as its usability, clinical relevance, convenience, and evidence-based design [6]. Studies on video consultations suggest how they are more convenient and time-saving for patients, but they are not perceived as superior to traditional in-person consultations [26,27]. Thus, despite growing interest in specific features of digital outpatient care and to some extent, the combination of features, previous research remains limited concerning the patients' experiences on a comprehensive model for digital outpatient care. By adopting health literacy as a guiding framework, the aim of this study was to explore and gain in-depth knowledge about the experiences of patients with chronic or long-term conditions enrolled in a digital 6-month outpatient care intervention.

Methods

Study Design

We conducted an explorative qualitative interview study as part of a larger study evaluating the use of digital outpatient care for adults with cancer, interstitial lung disease, epilepsy, or a complex pain condition at 2 university hospitals in Norway [28,29]. The study is reported according to the COREQ (Consolidated criteria for Reporting Qualitative research) guidelines [30].

Setting

The setting for this study was the 4 outpatient services participating in a multicenter controlled trial evaluating digital outpatient services, comprising the Department of Respiratory Diseases, the Department of Neurology, and the Department of Pain Management at Oslo University Hospital, as well as the Department of Cancer at the University Hospital of North Norway [28,29]. All patients included in this trial were living at home, receiving outpatient care, and otherwise managing their own lives.

Eligible participants had been included in the digital outpatient intervention, previously detailed [28,29]. The core focus of the intervention was to improve outpatient service accessibility,

XSL•FO
RenderX

allowing the patients to engage with patient-reported outcome measures, self-monitor parameters, and communicate asynchronously with health care workers. Designated health care workers assessed the patients' responses and parameters and had asynchronous digital contact, aligning with a conceptual model linking health literacy to service access and self-management [31]. Health care workers assigned tasks to patients and set individual thresholds using a traffic light model. Notifications alerted health care workers of deviations such as increased pain, side effects, or uncompleted tasks. Patients received reminders for unfinished tasks. Two-way messaging enabled asynchronous communication for questions, information exchange, and sharing of treatment plans. The digital outpatient intervention was enabled through the Dignio Connected Care platform, allowing personalized patient–health care worker contact [32] (Multimedia Appendices 1 and 2). The platform, including the health care worker's Dignio Prevent system and the patient's MyDignio app, is Conformité Européenne marked, privacy compliant, and globally applied [32].

Recruitment and Study Participants

Patients in the intervention arm of the overall multicenter controlled trial were invited to be interviewed when they completed their 6-month follow-up questionnaire at the end of the digital intervention. One item of the questionnaire asked for their consent to be contacted by a researcher with information regarding the qualitative interviews. An a priori estimation suggested a sample of 12-15 patients from the intervention group [28]. A strategic recruitment among those accepting to be contacted was conducted to ensure a broad sample of interviewees, enabling our aim of collecting rich data on the experiences of using the digital outpatient care intervention. Oral information was given over the phone, prior to written information in a secure digital platform, also allowing for their digital consent. We made use of the secure service "Nettskjema" digital consent in a service for sensitive data (Tjeneste for Sensitive Data [TSD] in Norwegian: the English version is Service for Sensitive Data]) developed at the University of Oslo. TSD is designed for storing and processing sensitive data in compliance with the Norwegian "Personal Data Act" and "Health Research Act." The consent invitations were sent to the participants through the Pretty Good Privacy encrypted version of the University of Oslo web questionnaire service "Nettskjema" demanding a governmental ID portal for login. A total of 91 patients completed their 6-month questionnaire, and of these, 50 agreed to be contacted for a possible interview. A researcher consecutively approached 21 patients, and 18 patients provided their digital consent to be interviewed.

Data Collection

To provide the information most valuable for the study aim, an interview guide was developed based on reviewing previous research in the field, inspired by topics on innovation assessment [33], and by adding items regarding specific factors associated with the current intervention [28]. A draft for the interview guide was developed by the first author and reviewed prior to a satisfactory version for a pilot interview. No changes were made to the interview guide after the pilot interview, and the pilot interview was included in the final analysis. The interview guide (Multimedia Appendix 3) contained an introduction with questions on their use of the digital service in general, before a midpart with more detailed questions on their experience with intervention, and finally some summarizing questions allowing final reflections. The interviewer (HH) has a background as a nurse with a PhD in health service research, without any personal or professional relationship with the participants. She has conducted qualitative research and analysis earlier and led the intervention study in which this substudy arises from.

All interviews were conducted over the telephone. All participants used their private smartphones for the interviews, while the interviewer (HH) used a work-related phone, with only voice. The "Nettskjema" Dictaphone app was used to record all interviews, with 1 main recorder and a backup recorder. All recordings were satisfactory besides one, where the line was interrupted midway through, and a new recording was started. No data were lost in this recording.

All interviews were conducted between 9 AM and 5 PM, with participants offered multiple time slots to ensure convenience. The interviewer called the participants at the scheduled time, and of those consenting, only 1 never responded despite repeated attempts from the researcher. Thus, 17 interviews were conducted with consenting participants, of whom 5 had cancer, 5 had chronic pain, 4 had interstitial lung disease, and 3 had epilepsy (Table 1). All participants were interviewed only once after their 6-month follow-up, and saturation was achieved through the 17 interviews; thus, no participants were added to the interview sample. The median time from the 6-month follow-up questionnaire end to interview was 40 days (minimum-maximum 9-95 days). The interviews lasted from 15 minutes to the maximum of 46 minutes, with a median interview time of 26 minutes. Median age of the interviewed participants was 62 years (minimum-maximum 36-83 years), of whom 8 were females and 9 were males. The participants had some variation in how they had used the digital intervention, and only 1 participant was categorized as a low user (Table 2).

Table 1. Participants, age groups, gender, department, and interview details.

Age groups (years)	N (%)	Sex, female/male	Departments	Minutes of interview, median (minimum-maximum)
31-50	3 (17)	2/1	Neurology and pain management	27 (26-27)
51-60	4 (24)	2/2	Cancer, respiratory diseases, and neurology	24 (16-39)
61-70	6 (35)	2/4	Cancer, respiratory diseases, neurology, and pain management	22 (15-46)
71-90	4 (24)	2/2	Cancer and respiratory diseases	29 (17-33)

XSL•FO
RenderX

Table 2. Use of the digital intervention (N=17).

Intervention	Values
Digital interaction[a]	
Total number of digital interactions	
Mean (SD)	58.8 (55.6)
Median (minimum-maximum)	65 (3-157)
Dichotomized digital interaction[b]	
Low use, n (%)	1 (6)
High use, n (%)	16 (94)
Asynchronous chat messages	
Messages sent from patient	
Mean (SD)	8.4 (9.2)
Median (minimum-maximum)	4 (0-31)
Messages sent from health care worker	
Mean (SD)	8.2 (7.5)
Median (minimum-maximum)	3 (0-20)
Total number of messages	
Mean (SD)	16.6 (16.2)
Median (minimum-maximum)	12 (0-51)
PRO[c] **measures**	
PRO measures sent to patient	
Mean (SD)	23.7 (25.4)
Median (minimum-maximum)	12 (0-82)
PRO measure responses from patient	
Mean (SD)	21.4 (24.7)
Median (minimum-maximum)	11 (0-82)

[a]Total counts were computed for chat messages, video visits, completed PRO measures, and monitoring events.

[b]Low users responded to less than 30% of the expected PRO measures.

[c]PRO: patient-reported outcome.

Data Analysis

All interviews were transcribed verbatim using the F4-transcript software (Audiotranskription.de) in TSD by the first author (HH). Then, the transcripts was analyzed using thematic analysis [34], with the following steps: (1) familiarize with data, (2) generate initial codes, (3) search for themes, (4) review themes, (5) define and name the themes, and (6) produce the report. An inductive approach was used, with the research question at the outset, allowing an analysis on the premises of the data at hand [34]. Both authors read and reread the material and HH generated initial and preliminary codes of the material—interview by interview. These preliminary codes were later reviewed by EF, before they were discussed, reviewed, and refined. Through discussion, the 2 authors (HH and EF) suggested themes to describe and clarify the codes, with particular attention given to exploring how patients' interactions with the digital health intervention may reflect aspects of their ability to access, understand, and use health information and the digital health tool according to health literacy and digital health literacy. A final set of themes, codes, and corresponding quotes was then summarized. All interviews were conducted, transcribed, and analyzed in Norwegian. The quotes have been translated to English and reviewed to assess the intended meaning. Minor alterations were made to ensure anonymity among the participants.

Ethical Considerations

The regional ethical committee in Norway prereviewed the protocol and judged the project as outside its mandate according to the Norwegian Health Research Act (regional ethical committee south-east reference number 252051). The project was approved by the data protection officer at Oslo University Hospital (reference 21/06826) and Northern Norway University Hospital (reference 2021/4942). All participants signed a digital, written informed consent form before participating in the project. This study was conducted in accordance with the Declaration

XSL•FO
RenderX

of Helsinki and all data were deidentified to ensure privacy and confidentiality. No compensation was provided to participants.

Results

Summary of the Themes and Codes

The thematic analysis led to 1 main theme "Digital outpatient care as a flexible service supporting patients' self-management," informed by 3 subthemes (Table 3).

XSL•FO
RenderX

Table 3. Themes and corresponding codes with quotes from the thematic analysis. Theme: digital outpatient care as a flexible service supporting patients' self-management.

Subthemes and codes	Quotes
The ongoing nature of managing a chronic condition and how the digital service meets the patients' desire for autonomy in their care	
Living with a chronic condition necessitates ongoing self-management	• "When you live with a diagnosis like this, it's not something you can easily forget" [P17]. • "I have had some quite big seizures, and little episodes that I don't really know whether are related to these seizures and that worries me" [P19].
Flexible services are required to support self-management	• "It allows me to better track myself and makes me calmer in relation to the disease progression, and I believe it enables me to respond at an earlier stage than I otherwise might have if something were to happen that affects my situation. So, the motivation lies in the ability to cope with a difficult disease, I would say. There's also additional motivation in that I enjoy numbers, apps, and correlations" [P17]. • "My motivation for using the app was the possibility to ask questions, and perhaps add notes to my symptoms, to be able to see some patterns" [P29].
Making use of digital tools for self-management supports autonomy in the patients' role	• "With this app, being able to see that the results sometimes go down a bit, and if you as a patient don't feel particularly worse, you might wait until the next measurement to make contact because you can relate this to your current situation and how you feel then, which is definitely possible with my disease" [P17]. • "Sometimes, I feel I am in bad shape, but the app says I am not, and that makes me happy...because there's so many things affecting my breath, like my state of mind and all" [P11].
Digital tools flexibly address the patients' unique needs but reliability depend on patient interaction	
Digital tools provide new solutions to real needs in a flexible manner	• "I believe that, for my part, the frequency of contact with the healthcare system has decreased since I started using the app" [P17]. • "I find the app to be a part of the entire [outpatient care] system" [P36].
Enabling interaction through communication of unique needs beyond standardized formats	• "The app asked more thorough questions than I remember being asked in person" [P12]. • "You do have specific alternatives to respond to, but perhaps some lines where you could express something? It's useful for the users, but whether it's also useful for you as the recipient, one might question, but maybe just 3-4 lines" [P36].
Making sure digital tools are suitable for the user	• "It is indeed an advantage to have digital contact for those of us who are digital and can use such things. But it is important that the digital solutions do not replace physical attendance and postal mail, but come as an addition, that one can choose" [P6].
Digital services enhance the patients' sense of safety through easy access to a relation with competent health care workers	
Easy and seamless contact with the outpatient clinic	• "But the feeling of safety, the certainty that I could get hold of someone if there was something, that gives a lot of peace of mind for my part" [P32]. • "As I am working full time, it's a hassle to leave work to go to the clinic, and it's adding more stress because I can't do my job properly" [P29].
Digital services facilitate patient access to health care, while supporting the health care workers' workflow	• "It is a reassurance for me if I would need it, and two years ago, I had a question that I didn't know whether was important and it took a month to get an answer. If we could have done it via the app, it would have been faster for both of us instead of having to call repeatedly" [P19].
Building a trustful digital relation to health care workers	• "Yes, I responded twice a week to the nurse and if I didn't respond, I received a message (...) so it worked perfectly (...) but I have never spoken directly with the nurse, which by the way is exceptionally pleasant to communicate with digitally" [P22]. • "I did know the doctor from before, but not the nurse – and not that it matters because now I feel like I know the nurse too, and that gives me a sense of safety" [P37].

XSL•FO
RenderX

The Ongoing Nature of Managing a Chronic Condition and How the Digital Service Meets the Patients' Desire for Autonomy in Their Care

The participants acknowledged the burden of living with a chronic condition. While some felt that they already understood the required self-management, they emphasized that digital follow-up provided a sense of support that lightened the burden, allowing them to focus more on living rather than being ill. One participant remarked, "Because I don't want to spend too much time on my condition, I want to spend my time on life instead, right? If it becomes too much, then it becomes too much concentration on the condition, and I still think that I'm not particularly sick" [P12].

The importance of self-management was highlighted, with some participants noting that this was the first time they had access to a tool that truly supported their self-management effort and the ability to monitor deterioration as highly valuable. The option to send messages, even for minor questions, was particularly appreciated, as these small clarifications provided reassurance: "Even small questions can take up a lot of space" [P34]. This access to advice enhanced their sense of control, and participants found it easier to consult health care workers through the digital platform compared with standard care.

The patients acknowledged their frequent need for health services and the challenges of contacting health care workers in outpatient care. They shared how digital outpatient care addressed this issue by improving communication and accessibility. Patients valued the efficiency and convenience offered by the digital solution, as one participant shared: "So, for me, this worked very well because I don't need to go in just for questions. I wasn't scared of anything, and I didn't have to go anywhere, and I didn't have to spend time calling, so it worked very well for me. Then you can rather go in if you don't achieve your goals with the digital follow-up" [P29]. While the benefits of digital services were valued, the patients emphasized the need for a hybrid model with digital and physical contact. They supported innovation in health services, particularly when it contributed to flexible, individualized care and expressed positivity toward continued digital contact after the project.

Patients found the digital outpatient care tools valuable for tracking health parameters, identifying changes in their condition, thus enhancing their sense of safety. The tools were used less when their condition was stable but became essential when they suspected health changes. Through the app, patients reported gaining an understanding of their symptoms and a reflection on their condition, which they felt had an impact on their health literacy and self-management skills. In addition, the tools facilitated greater involvement of family members, providing an extra layer of support. As one patient shared: "Yes! It's always a big moment every time I take such tests, and the people around me are always interested in how it's going. They always ask how it went when I took these tests" [P17]. This family engagement was described as a motivating factor for using the digital platform. However, some patients noted that their use of the platform was independent of family involvement, reflecting diverse approaches to digital health engagement.

Digital Tools Flexibly Address the Patients' Unique Needs, But Reliability Depends on Patient Interaction

Most patients valued the flexibility offered by digital tools, particularly the convenience of receiving care without traveling to the hospital. This was especially important for those with long travel distances or conditions that made travel burdensome. Digital tools were described as time-saving, stress-reducing, and helpful for minimizing work absences, with the ease of use and sense of security further enhancing their appeal: "If something comes up in the evening, I can enter it and then I get a response in the morning...much more frequent contact than usual" [P37]. Patients appreciated the opportunity for more flexible interactions with health care workers.

Patient-reported outcomes were generally seen as relevant and easy to complete, allowing for more thorough and individualized health reporting than traditional consultations. However, some patients felt that standardized questions and response categories did not fully capture their experiences, expressing a need for more flexibility and individual tailoring. One participant reflected: "For many of the questions I get [through the app], there are some I don't get to elaborate enough on, and it could just be that the information that I have could be valuable to you" [P27]. Suggestions for improvement included displaying scheduled consultations and test results in the app, reducing questionnaire frequency, adding time estimates for completion, and allowing more free-text responses. Patients also expressed trust in the app's data security but emphasized that not all data needed to be shared digitally to minimize risks.

Prerequisites for using digital tools, such as internet access, smartphones, and digital identification, were noted. Most found the app easy to navigate, with minimal training needed, although those using medical monitoring devices appreciated some initial guidance. Initial stress in using the app subsided with familiarity, as one patient noted: "At first, I was a bit stressed by it [the app], when the messages came from the hospital that I had to do these measurements, I felt that I was getting stressed. But why that is, I don't know, it's going well now. I was a bit surprised at myself, like, why is it stressing me, but it has passed now so that's good because now I'm starting to get the hang of it" [P11]. However, digital tools were not universally suitable, with concerns raised about older adult users, individuals with health-related anxiety, and the timing of app prompts for working patients. Despite these concerns, reminders were often described as helpful for completing tasks.

Digital Services Enhance the Patients' Sense of Safety Through Easy Access to a Relation With Competent Health Care Workers

The patients appreciated the sense of security offered by the user-friendly and flexible digital contact with health care workers in outpatient care. The app was seen as supportive rather than burdensome, with some noting that spending a few minutes weekly was a small effort for managing a long-term condition: "It's not like the app reminds you that you have a chronic condition, it's more perceived as a positive element in addition to the treatment you're already receiving" [P17]. Digital reporting was practical and reduced errors, although patients cautioned against relying on technology for false security.

XSL•FO
RenderX

Avoiding phone queues and callbacks saved time and reduced stress for patients, while also improving workflow for health care workers by allowing responses on their schedule. One participant shared: “I liked that I seemed to have better access to the department than normal, it was less burdensome than calling in the limited time the reception is open, and sometimes it rings and rings and rings and rings, and sometimes dealing with less or even unqualified reception workers which I have had problems with before” [P19]. Asynchronous messaging was particularly valued for enabling prompt answers to both urgent and nonurgent questions, fostering a sense of being seen and supported. Even when unused, the messaging function reassured patients that they could reach skilled health care workers if needed.

The competence of health care workers responding to queries was trusted, and while some patients valued consistent interactions with the same person, others prioritized competence over familiarity, especially with nurses. Patients appreciated building relationships through physical consultations before transitioning to digital follow-up but found digital contact effective once trust was established, as one patient elaborated: “And yes, I’ve met them, they know me and I have a face to them, and I think that’s very good. That it’s the same people who follow up. But it’s mostly on the doctor's side that I need to have the same person, whom I trust, and I have met in consultations, who knows me. But with the nurses, I think it’s very good to have two people” [P14]. Administrative workers triaging questions were considered acceptable, provided medical issues were handled by qualified professionals. However, some patients still preferred direct calls to familiar health care workers, especially if they already had access to adequate resources.

Discussion

Principal Findings

The patients experienced digital outpatient care as a flexible service supporting self-management, which fulfilled the patients’ ongoing but unmet need of autonomy in their self-management. The findings suggest that aspects of health literacy, such as the ability to understand, engage with, and act on health information, may have influenced how patients interacted with the digital platform and benefited from the service. To achieve a truly flexible and reliable service, active interaction between patients and health care workers with the digital platform is essential to strengthen the patients’ sense of safety. The patients’ reflections regarding the digital outpatient service in this study highlight key insights worth further discussion.

From the patients’ perspective, a digital outpatient solution offers greater flexibility in health care services than traditional calendar-based approaches, which rely solely on phone calls or scheduled consultations to access the outpatient clinic. Previous research has described how traditional systems are rigid and allow little patient-centered and flexible approaches [35]. If changes in the patients’ clinical status occur, there is little room for a rapid response or to move forward a scheduled appointment to curb the deterioration. However, the digital opportunities include flexibility through asynchronous messaging, monitoring, and patient responses—offering a more patient-centered approach, supporting the patient’s self-management [28]. Furthermore, health literacy is often intertwined with self-management, and both concepts are positively and iteratively influencing each other [16,31]. This highlights how digital outpatient solutions not only enhance the patients’ experience of more flexibility and responsiveness but also foster patient-centered care by supporting self-management and strengthening health literacy in a mutually reinforcing manner.

While our study and similar research highlight the benefits of digital approaches [5,10-12,18], some patients may feel burdened by the added responsibility, potentially threatening their sense of security. One patient described initial anxiety when using the digital platform, fearing mistakes in reporting—a concern echoed in previous studies emphasizing the critical role of usability and adequate digital and health literacy [36,37]. However, with practice, participants gained confidence in self-management and their sense of insecurity decreased. Interaction with health care workers further supported this transition, providing timely guidance and encouragement through digital messaging. This highlights the importance of tailoring support to help reluctant patients build the necessary skills and competencies for digital outpatient care [5,18]. In our study, patients demonstrated health literacy through their engagement with the intervention, responding accurately to patient-reported outcome measures, conducting remote monitoring as instructed, and framing thoughtful questions via the messaging functions, showcasing the integral link between health literacy and effective self-management.

The flexibility and interactive opportunities provided by the digital platform highlight its potential to reduce the traditional asymmetry between patients and health care providers. Unlike calendar-based systems in traditional care, which often overlook individual patient needs, digital outpatient care allows patients to engage with health care workers when changes in their clinical condition occur. Over time, patients became more adept at interpreting the relationship between their symptoms and objective measurements provided by the system, enabling them to make informed decisions and take appropriate actions with guidance from health care workers [28]. In the digital solution, where interactions occurred based on asynchronous messages, the focus of the interaction was more centered on the patients’ input and experiences rather than being primarily driven by the health care workers [6]. The role of the health care worker is also evolving, much like the changes digital solutions bring to the patient’s role. More flexible ways of interacting with patients are shifting the responsibilities of health care workers, often in ways they are not fully prepared for. However, with the necessary competencies in using digital solutions and a clear understanding of their integration into clinical care, health care workers can positively influence patients’ engagement and effective use of digital health solutions to maintain or improve health [38,39]. While processes that strengthen the patient role are well established in patient-centered care [6], their implementations through digital solutions remain underexplored.

Based on our findings, we argue that digital outpatient care can enhance care quality by providing easier and more timely access to health services, saving time and potentially preventing

XSL•FO
RenderX

symptom deterioration through early identification of changes. Patients' trust in the system and providers, their increased self-management, and reduced feelings of burden—both as individuals living with a chronic condition and as users of health care services—further support this claim. However, the impact of digital outpatient care on health outcomes related to patients' conditions remains unclear. While previous research has shown positive but small and low-graded evidence on such effects [5,7,18], further studies are needed to evaluate both health outcomes and the experiences of patients and health care providers using integrated digital health solutions.

Limitations and Strengths

This study is among the few qualitative investigations into integrated digital health solutions in outpatient care, but it has limitations. Participants were recruited from a study evaluating the same digital platform, with tailored interventions based on clinical needs, leading to heterogeneous experiences. All interviews were conducted using telephone with only voice, and thus, potentially relevant data concerning body language and nonverbal communication were not interpreted. To address this, a semistructured interview guide was used, and while data analysis was comprehensive, recurring similar statements were interpreted as having broader relevance. A key finding was that all patients expressed trust in the digital solution and care provided. However, this may introduce bias, as participants were likely those already comfortable with digital solutions, potentially excluding perspectives from less digitally inclined individuals.

Conclusions

In this study, we aimed to explore the experiences of patients with long-term conditions in using a digital outpatient care service for 6 months. Digital outpatient care was experienced as flexible and supportive for patients with long-term conditions. The increased possibility of interacting with health care workers was welcomed by the patients, and the combination of flexibility, self-monitoring, and addressing concerns regarding their self-management seemed to increase the patients' experience of autonomy. As health literacy likely plays a role in patients' ability to effectively engage with digital tools and self-manage their conditions, future research should explore how varying levels of health literacy influence these outcomes. In addition, research should address whether such digital outpatient clinics are positive for a wider range of patients, associated health outcomes, and any positive effects on a health system level.

Acknowledgments

We would like to thank all the participating patients and health care workers at each outpatient clinic. We are grateful to the staff in Dignio for their valuable insight and practical tailoring of their platform, and for a good collaboration, particularly with Anna Hurrød, Meetali Kakad, and Andreas Norling. Furthermore, we would like to thank Anette Winger and Astrid Torbjørnsen for their valuable comments to subsets of the draft in process. Artificial intelligence–assistive technologies have been applied for language editing only.

Funding

This work was initiated by the Oslo University Hospital, financially supported by the Research Council of Norway (grant 316244), internally funded by Oslo University Hospital and UNN to secure staff, and funded by Dignio Connected Care to finance user licenses in the digital platform. The Research Council of Norway and Dignio Connected Care had no role in preparing this manuscript.

Data Availability

The datasets used and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

HH initiated the study, developed the interview guide, and conducted all the interviews and the thematic analysis before she wrote the manuscript. EF collaborated in the project and gave feedback on the interview guide and the preliminary interviews, took part in the thematic analysis, and revised the manuscript. Both coauthors approved the final manuscript. No authors were added or removed during the process.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Dignio Prevent interface.
[DOCX File , 123 KB - jmir_v28i1e79155_app1.docx]

Multimedia Appendix 2

XSL•FO
RenderX

MyDignio interface.
[DOCX File , 124 KB - jmir_v28i1e79155_app2.docx]

Multimedia Appendix 3
Interview guide.
[DOCX File , 21 KB - jmir_v28i1e79155_app3.docx]

Multimedia Appendix 4
COREQ (Consolidated criteria for Reporting Qualitative research) checklist.
[PDF File (Adobe PDF File), 225 KB - jmir_v28i1e79155_app4.pdf]

References

1. GBD 2019 Demographics Collaborators. Global age-sex-specific fertility, mortality, healthy life expectancy (HALE), and population estimates in 204 countries and territories, 1950-2019: a comprehensive demographic analysis for the Global Burden of Disease Study 2019. Lancet 2020;396(10258):1160-1203 [FREE Full text] [doi: 10.1016/S0140-6736(20)30977-6] [Medline: 33069325]
2. Blumenthal D, Chernof B, Fulmer T, Lumpkin J, Selberg J. Caring for high-need, high-cost patients—an urgent priority. N Engl J Med 2016;375(10):909-911. [doi: 10.1056/NEJMp1608511] [Medline: 27602661]
3. Hjollund NHI, Larsen LP, de Thurah AL, Grove BE, Skuladottir H, Linnet H, et al. Patient-reported outcome (PRO) measurements in chronic and malignant diseases: ten years' experience with PRO-algorithm-based patient-clinician interaction (telePRO) in AmbuFlex. Qual Life Res 2023;32(4):1053-1067 [FREE Full text] [doi: 10.1007/s11136-022-03322-9] [Medline: 36639598]
4. Gandrup J, Ali SM, McBeth J, van der Veer SN, Dixon WG. Remote symptom monitoring integrated into electronic health records: a systematic review. J Am Med Inform Assoc 2020;27(11):1752-1763 [FREE Full text] [doi: 10.1093/jamia/ocaa177] [Medline: 32968785]
5. Verweel L, Newman A, Michaelchuk W, Packham T, Goldstein R, Brooks D. The effect of digital interventions on related health literacy and skills for individuals living with chronic diseases: a systematic review and meta-analysis. Int J Med Inform 2023;177:105114. [doi: 10.1016/j.ijmedinf.2023.105114] [Medline: 37329765]
6. Greenhalgh J, Gooding K, Gibbons E, Dalkin S, Wright J, Valderas J, et al. How do patient reported outcome measures (PROMs) support clinician-patient communication and patient care? A realist synthesis. J Patient Rep Outcomes 2018;2:42 [FREE Full text] [doi: 10.1186/s41687-018-0061-6] [Medline: 30294712]
7. Mandizha J, Lanario JW, Duckworth A, Lines S, Paiva A, Elworthy V, et al. Patient perspectives on home-spirometry in interstitial lung disease: a qualitative co-designed study. BMJ Open Respir Res 2023;10(1):e001837 [FREE Full text] [doi: 10.1136/bmjresp-2023-001837] [Medline: 37793682]
8. de Bell S, Zhelev Z, Shaw N, Bethel A, Anderson R, Thompson Coon J. Remote monitoring for long-term physical health conditions: an evidence and gap map. Health Soc Care Deliv Res 2023;11(22):1-74. [doi: 10.3310/BVCF6192] [Medline: 38014553]
9. Althobiani MA, Evans RA, Alqahtani JS, Aldhahir AM, Russell A, Hurst JR, et al. Home monitoring of physiology and symptoms to detect interstitial lung disease exacerbations and progression: a systematic review. ERJ Open Res 2021;7(4):00441 [FREE Full text] [doi: 10.1183/23120541.00441-2021] [Medline: 34938799]
10. Escriva Boulley G, Leroy T, Bernetière C, Paquienseguy F, Desfriches-Doria O, Préau M. Digital health interventions to help living with cancer: a systematic review of participants' engagement and psychosocial effects. Psychooncology 2018;27(12):2677-2686. [doi: 10.1002/pon.4867] [Medline: 30152074]
11. Hewitt S, Sephton R, Yeowell G. The effectiveness of digital health interventions in the management of musculoskeletal conditions: systematic literature review. J Med Internet Res 2020;22(6):e15617 [FREE Full text] [doi: 10.2196/15617] [Medline: 32501277]
12. Shegog R, Braverman L, Hixson JD. Digital and technological opportunities in epilepsy: toward a digital ecosystem for enhanced epilepsy management. Epilepsy Behav 2020;102:106663. [doi: 10.1016/j.yebeh.2019.106663] [Medline: 31778878]
13. Nutbeam D. Health promotion glossary. Health Promot Int 1998;13(4):349-364. [doi: 10.1093/heapro/13.4.349]
14. Cheng C, Beauchamp A, Elsworth GR, Osborne RH. Applying the electronic health literacy lens: systematic review of electronic health interventions targeted at socially disadvantaged groups. J Med Internet Res 2020;22(8):e18476 [FREE Full text] [doi: 10.2196/18476] [Medline: 32788144]
15. Baccolini V, Rosso A, Di Paolo C, Isonne C, Salerno C, Migliara G, et al. What is the prevalence of low health literacy in European Union member states? A systematic review and meta-analysis. J Gen Intern Med 2021;36(3):753-761 [FREE Full text] [doi: 10.1007/s11606-020-06407-8] [Medline: 33403622]
16. van der Gaag M, Heijmans M, Spoiala C, Rademakers J. The importance of health literacy for self-management: a scoping review of reviews. Chronic Illn 2022;18(2):234-254. [doi: 10.1177/17423953211035472] [Medline: 34402309]

XSL•FO
RenderX

17. Larsen MH, Mengshoel AM, Andersen MH, Borge CR, Ahlsen B, Dahl KG, et al. "A bit of everything": health literacy interventions in chronic conditions—a systematic review. Patient Educ Couns 2022;105(10):2999-3016 [FREE Full text] [doi: 10.1016/j.pec.2022.05.008] [Medline: 35641366]
18. Barbati C, Maranesi E, Giammarchi C, Lenge M, Bonciani M, Barbi E, et al. Effectiveness of eHealth literacy interventions: a systematic review and meta-analysis of experimental studies. BMC Public Health 2025;25(1):288 [FREE Full text] [doi: 10.1186/s12889-025-21354-x] [Medline: 39849354]
19. Faux-Nightingale A, Philp F, Chadwick D, Singh B, Pandyan A. Available tools to evaluate digital health literacy and engagement with eHealth resources: a scoping review. Heliyon 2022;8(8):e10380 [FREE Full text] [doi: 10.1016/j.heliyon.2022.e10380] [Medline: 36090207]
20. Karnoe A, Kayser L. How is eHealth literacy measured and what do the measurements tell us? A systematic review. Knowledge Manage E-Learning 2015;7(4):576-600. [doi: 10.34105/j.kmel.2015.07.038]
21. Lee J, Lee E, Chae D. eHealth literacy instruments: systematic review of measurement properties. J Med Internet Res 2021;23(11):e30644 [FREE Full text] [doi: 10.2196/30644] [Medline: 34779781]
22. Refahi H, Klein M, Feigerlova E. e-Health literacy skills in people with chronic diseases and what do the measurements tell us: a scoping review. Telemed J E Health 2023;29(2):198-208. [doi: 10.1089/tmj.2022.0115] [Medline: 35671526]
23. Holmen H, Fløtlo T, Tørris C, Løyland B, Almendingen K, Bjørnnes AK, et al. Unpacking the public health triad of social inequality in health, health literacy, and quality of life-a scoping review of research characteristics. Int J Environ Res Public Health 2023;21(1):36 [FREE Full text] [doi: 10.3390/ijerph21010036] [Medline: 38248501]
24. Jacobs RJ, Lou JQ, Ownby RL, Caballero J. A systematic review of eHealth interventions to improve health literacy. Health Inform J 2016;22(2):81-98 [FREE Full text] [doi: 10.1177/1460458214534092] [Medline: 24916567]
25. van Kessel R, Wong BLH, Clemens T, Brand H. Digital health literacy as a super determinant of health: more than simply the sum of its parts. Internet Interv 2022;27:100500 [FREE Full text] [doi: 10.1016/j.invent.2022.100500] [Medline: 35242586]
26. Thiyagarajan A, Grant C, Griffiths F, Atherton H. Exploring patients' and clinicians' experiences of video consultations in primary care: a systematic scoping review. BJGP Open 2020;4(1) [FREE Full text] [doi: 10.3399/bjgpopen20X101020] [Medline: 32184212]
27. Walthall H, Schutz S, Snowball J, Vagner R, Fernandez N, Bartram E. Patients' and clinicians' experiences of remote consultation? A narrative synthesis. J Adv Nurs 2022;78(7):1954-1967 [FREE Full text] [doi: 10.1111/jan.15230] [Medline: 35362191]
28. Holmen H, Holm AM, Kilvær TK, Ljoså TM, Granan L, Ekholdt C, et al. Digital outpatient services for adults: development of an intervention and protocol for a multicenter non-randomized controlled trial. JMIR Res Protoc 2023;12:e46649 [FREE Full text] [doi: 10.2196/46649] [Medline: 37428533]
29. Holmen H, Holm AM, Falk RS, Kilvær TK, Ljosaa TM, Ekholdt C, et al. A digital outpatient service with a mobile app for tailored care and health literacy in adults with long-term health service needs: multicenter nonrandomized controlled trial. J Med Internet Res 2025;27:e60343 [FREE Full text] [doi: 10.2196/60343] [Medline: 40294411]
30. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
31. Paasche-Orlow MK, Wolf MS. The causal pathways linking health literacy to health outcomes. Am J Health Behav 2007;31(1):19-26. [doi: 10.5993/ajhb.31.s1.4]
32. Dignio Connected Care. Dignio. URL: https://dignio.com/solution/ [accessed 2025-12-16]
33. Kværner K. How to assess value and benefits of innovation. Noteb—Nordic Test Beds. Centre for Connected Care and Nordic Innovation. 2018. URL: https://www.researchgate.net/figure/The-User-guide-How-to-assess-value-and-benefits-of-innovation-the-end-product-of-the_fig1_331666378 [accessed 2025-12-20]
34. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
35. Matulis JC, McCoy R. Patient-centered appointment scheduling: a call for autonomy, continuity, and creativity. J Gen Intern Med 2021;36(2):511-514 [FREE Full text] [doi: 10.1007/s11606-020-06058-9] [Medline: 32885369]
36. Lupton D. The digitally engaged patient: self-monitoring and self-care in the digital health era. Soc Theory Health 2013;11(3):256-270. [doi: 10.1057/sth.2013.10]
37. Song M, Elson J, Bastola D. Digital age transformation in patient-physician communication: 25-Year Narrative Review (1999-2023). J Med Internet Res 2025;27:e60512 [FREE Full text] [doi: 10.2196/60512] [Medline: 39819592]
38. Kho J, Gillespie N, Martin-Khan M. A systematic scoping review of change management practices used for telemedicine service implementations. BMC Health Serv Res 2020;20(1):815 [FREE Full text] [doi: 10.1186/s12913-020-05657-w] [Medline: 32873295]
39. Longhini J, Rossettini G, Palese A. Digital health competencies among health care professionals: systematic review. J Med Internet Res 2022 Aug 18;24(8):e36414 [FREE Full text] [doi: 10.2196/36414] [Medline: 35980735]

XSL•FO
RenderX

Abbreviations

COREQ: Consolidated criteria for Reporting Qualitative research
TSD: Tjeneste for Sensitive Data

Edited by A Mavragani; submitted 16.Jun.2025; peer-reviewed by W van Harten, FY Chou, YY Chen; comments to author 17.Sep.2025; revised version received 05.Nov.2025; accepted 12.Nov.2025; published 15.Jan.2026.

Please cite as:
Holmen H, Fosse E
"I Want to Spend My Time Living"—Experiences With a Digital Outpatient Service With a Mobile App for Tailored Care Among Adults With Long-Term Health Service Needs: Qualitative Study Using Thematic Analysis
J Med Internet Res 2026;28:e79155
URL: https://www.jmir.org/2026/1/e79155
doi:10.2196/79155
PMID:

©Heidi Holmen, Erik Fosse. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 15.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Feasibility and Acceptability of AI-Powered Tools for Early Autism Screening in Egypt: Semistructured Focus Group Study

Pratheepan Yogarajah[1*], PhD; Ammal M Metwally[2*], Prof Dr, PhD; Priyanka Chaurasia[1], PhD; Ghada A Elshaarawy[2], Prof Dr, PhD; Shereen M El Khateeb[3], MFA; Engy A Ashaat[2], PhD; Amal Elsaeid[2], PhD; Nahed A Elghareeb[4], PhD; Amira S ElRifay[2], Prof Dr, PhD

[1]School of Computing, Engineering & Intelligent Systems, Ulster University, Londonderry, Northern Ireland, United Kingdom

[2]National Research Centre, Giza, Giza, Egypt

[3]Department of psychology, American University in Cairo, Cairo, Egypt

[4]Ministry of Health and Population, Cairo, Egypt

[*]these authors contributed equally

Corresponding Author:
Pratheepan Yogarajah, PhD
School of Computing, Engineering & Intelligent Systems, Ulster University
Ulster University
Londonderry, Northern Ireland, BT48 7JL
United Kingdom
Phone: 44 2871675034
Email: p.yogarajah@ulster.ac.uk

Abstract

Background: Autism spectrum disorder (ASD) is often underdiagnosed in low- and middle-income countries due to limited specialist access, sociocultural stigma, and fragmented screening systems. Artificial intelligence (AI)–powered screening tools may improve early detection by enabling low-cost, accessible assessments. However, adoption depends on stakeholder trust, ethical safeguards, and alignment with local health system capacities.

Objective: This study explored the feasibility, acceptability, and perceived ethical and practical enablers and barriers to implementing AI-powered tools for early ASD screening in Egypt, with attention to urban–rural disparities and integration into existing care pathways.

Methods: We used a qualitative design with semistructured focus group discussions with 49 participants (21 parents of children with ASD and 28 health care professionals) recruited from urban and rural governorates. Discussions were audio-recorded, transcribed verbatim, and analyzed using Braun and Clarke's reflexive thematic analysis, supported by NVivo software (Lumivero). Methodological integrity was ensured through reflexivity, triangulation, and peer debriefing. Thematic saturation was monitored across groups, and participant diversity was prioritized across contexts.

Results: Five themes emerged: (1) AI as a supportive tool rather than a replacement for clinicians, emphasizing scalability and assistance for nonspecialists; (2) the need for cultural and contextual adaptation to ensure local relevance; (3) privacy, trust, and transparency concerns, including data security, consent, and algorithmic opacity; (4) reducing diagnostic inequities by addressing urban–rural disparities and strengthening community-based deployment; and (5) the preference for hybrid AI–human models, with conditions for adoption including cultural sensitivity, human oversight, and digital literacy support. Counts (n/N) of parents and health care professionals contributing to each theme were used descriptively as indicators of pattern salience rather than as statistical estimates of prevalence. Participants expressed cautious optimism, with parents emphasizing accessibility and speed, while health care professionals highlighted concerns about reliability, cultural adaptation, and data governance.

Conclusions: AI-powered ASD screening has potential to advance equitable early detection in underserved areas. Adoption requires transparent data governance, integration into hybrid human–AI models, culturally adaptive design, and targeted digital literacy initiatives. These findings provide an evidence-based roadmap for policymakers, technologists, and health system leaders to implement AI screening tools that are ethically sound, contextually relevant, and equity-focused.

(J Med Internet Res 2026;28:e82564) doi:10.2196/82564

XSL•FO
RenderX

KEYWORDS

AI; artificial intelligence; ASD; autism spectrum disorder; developing countries; Egypt; health equity; mobile applications; qualitative research; screening

Introduction

Autism spectrum disorder (ASD) is a prevalent neurodevelopmental condition that affects cognitive, behavioral, and social functioning, necessitating early and accurate identification for optimal developmental outcomes [1,2]. Early detection is particularly important because timely intervention can substantially improve language development, social skills, and long-term adaptive functioning [3]. However, many children remain undiagnosed due to barriers such as limited specialist access, cultural stigma, and inconsistent screening practices, particularly in low-resource settings [4,5].

Traditional screening methods, including clinical observation, structured interviews, and standardized behavioral checklists, remain the cornerstone of ASD diagnosis but are constrained by subjectivity and poor scalability. In Egypt, where the estimated prevalence of children aged 2-12 years at high risk for ASD is 3.3% (95% CI 3.1%-3.5%) [6] and those diagnosed with autism is 1.1% (95% CI 1.0%-1.2%) [7], the burden of undiagnosed developmental conditions is compounded by delayed health-seeking and shortages of trained professionals [6]. Developmental delays affect 6.4% of preschool-aged children [6] and 7.4% of school-aged children, underscoring the urgent need for scalable and accessible diagnostic alternatives [8]. These figures highlight the systemic challenges faced by families and health systems, where demand for developmental assessment far exceeds available capacity.

Artificial intelligence (AI) has emerged as a promising adjunct to traditional screening by enhancing diagnostic accuracy and reducing reliance on clinical judgment [9]. Machine learning models have demonstrated potential in identifying ASD-related traits via speech, behavior, facial analysis, and digital interactions [10]. Mobile AI apps, in particular, offer low-cost, early-stage screening accessible to both parents and frontline health care providers, especially in underserved communities [11,12]. Such tools could bridge gaps in access and standardization, addressing the fragmented diagnostic pathways observed in Egypt and other low- and middle-income countries (LMICs) [13].

Despite their promise, AI-driven tools face challenges, including algorithmic bias, data privacy concerns, limited cultural adaptability, and stakeholder mistrust [14]. In addition, most existing AI models are trained on Western datasets, raising questions about their validity and cultural fit in LMIC contexts [15]. Broader digital health research also shows that successful scale-up in low-resource environments requires attention to cultural adaptation, stakeholder engagement, and health system integration [16]. The feasibility and success of integrating AI into ASD screening depend not only on technical accuracy but also on community acceptance, ethical transparency, and usability.

These challenges can be better understood and systematically examined through established theoretical frameworks. To deepen interpretation of these adoption challenges, the study draws on the Health Belief Model (HBM) and the Technology Acceptance Model (TAM). HBM highlights how perceived susceptibility, benefits, and barriers influence caregivers' health-related decisions, including ASD help-seeking behaviors [17]. TAM, widely applied in digital health research, explains how perceived usefulness and ease of use shape acceptance of mobile and AI technologies in health care settings [18]. Integrating these frameworks provides a structured basis for examining how caregiver beliefs and provider attitudes interact with system-level constraints to shape adoption of AI-powered ASD screening in Egypt.

Against this backdrop, we conducted a qualitative study in Egypt using focus group discussions (FGDs) and follow-up interviews with health care providers and parents of young children with ASD across diverse urban, semiurban, and rural settings. The primary aim was to explore stakeholder perspectives on the feasibility, acceptability, and perceived ethical and practical enablers and barriers to implementing AI-powered tools for early ASD screening within the Egyptian health system. By situating these perspectives within Egypt's health system and sociocultural context, and interpreting them through the HBM and TAM, this qualitative inquiry seeks to inform the context-appropriate design and implementation of AI-enabled ASD screening approaches in LMICs. Accordingly, we addressed 2 research questions—research question 1 (RQ1) and research question 2 (RQ2): (RQ1) what barriers and contextual challenges do parents and health care providers identify in current ASD screening, referral, diagnosis, and access-to-support pathways in Egypt? (RQ2) How do parents and health care providers perceive the acceptability and feasibility of AI-powered ASD screening tools, including perceived value, usability, trust, and implementation requirements?

Methods

Research Design

This qualitative study used FGDs and individual interviews to explore stakeholder perspectives on AI-assisted autism screening in Egypt. Overall, the study included 49 participants, comprising 28 health care providers (pediatricians, psychologists, family medicine physicians, speech therapists, nurses, and health-information specialists) and 21 parents of children with a confirmed ASD diagnosis aged 6 years or younger. All participants took part in 1 of 7 FGDs, and a subset of these same participants later contributed to brief follow-up individual interviews; no additional participants were recruited solely for interviews. Participants were purposively sampled to reflect diverse professional backgrounds and geographic locations, including urban, semiurban, and rural settings [19]. This diversity facilitated examination of disparities in ASD screening accessibility and cultural perceptions.

FGDs allowed interactive exploration of clinical, cultural, and technical aspects of ASD diagnosis, while individual interviews

XSL•FO
RenderX

captured sensitive or nuanced views less likely to emerge in group settings [20,21]. Sessions were conducted separately for health care professionals and parents to minimize power imbalances and encourage open dialogue. Data collection proceeded iteratively alongside early familiarization and preliminary coding; after each FGD (and subsequent follow-up interviews), the facilitation and analysis team conducted rapid debriefs and reviewed emerging codes to assess whether additional sessions were likely to generate substantively new insights. Saturation was judged to have been reached when successive sessions yielded no substantively new codes or theme candidates beyond those already identified and the team agreed that additional data were unlikely to add meaningful conceptual depth [22,23]. Geographic representation was prioritized over equal participant numbers per site.

Participants were recruited through hospitals, specialized ASD clinics, professional networks, autism advocacy groups, special needs schools, and early intervention centers. All received detailed study information and confidentiality assurances and provided written informed consent prior to participation.

A full description of the stepwise thematic analysis process (phases 1-6 of Braun and Clarke's framework) has been provided in Multimedia Appendix 1.

Inclusion and Exclusion Criteria

Inclusion criteria were (1) health care professionals directly involved in ASD screening or developmental care, and (2) parents of children with a confirmed ASD diagnosis aged 6 years or younger. Consequently, the realized sample reflects perspectives that are closely anchored in the early detection and screening phase rather than in later, long-term intervention. Parents of children diagnosed during or soon after the preschool years were therefore prioritized, as their reflections most directly inform the feasibility and acceptability of AI-powered early screening tools.

Exclusion criteria were (1) health care providers without direct ASD-related experience, (2) parents of children with suspected but unconfirmed neurodevelopmental disorders, and (3) individuals unwilling to provide consent or unable to complete the full discussion session. This approach ensured that voices from both clinical and caregiving domains were adequately represented, enhancing the relevance and applicability of findings.

Recruitment Process

Participants were recruited through multiple channels to ensure diversity and relevance. Health care professionals were identified via hospitals, specialized ASD clinics, and professional networks in the selected governorates. Pediatric associations and child development centers facilitated the engagement of pediatricians, psychologists, nurses, and other specialists involved in ASD screening and intervention. All individuals invited agreed to participate, and no invited participants declined or withdrew.

Parents were recruited through autism advocacy groups, special needs schools, early intervention centers, and referrals from pediatric clinics. This ensured inclusion of caregivers with firsthand experience navigating ASD diagnosis and care pathways.

All participants received detailed study information, including objectives, procedures, confidentiality assurances, and their right to withdraw at any time. Written informed consent was obtained prior to participation.

Geographic Distribution and Representation

The study was conducted across 6 Egyptian governorates selected to capture geographic and demographic diversity, including urban, semiurban, and rural contexts. Governorates were purposively chosen to reflect differences in levels of urbanization, availability of ASD-related services, and health system infrastructure (eg, Ministry of Health and Population primary health care and maternal and child health facilities, university hospitals, and private or nongovernmental centers). This approach ensured that findings reflected a broad spectrum of health care access, cultural beliefs, and diagnostic experiences related to ASD. Rather than aiming for equal sample sizes per site, recruitment focused on meaningful representation of regional disparities in ASD service delivery. Geographic representation was therefore operationalized as maximum-variation coverage of contrasting service environments rather than as proportional sampling based on population size or formal rural–urban quotas.

This geographic diversity, combined with rigorous participant selection, supported a comprehensive understanding of the challenges and opportunities associated with integrating AI-powered ASD screening tools in varied sociocultural and clinical settings.

Data Collection

Data were collected through semistructured FGDs and individual interviews, combining structured prompts with open-ended dialogue to elicit both guided and spontaneous insights. FGDs lasted 60-90 minutes and were moderated by trained facilitators with backgrounds in public health, ASD, and qualitative research. During parent FGDs, children and, in many cases, the other parent were present in the room so that 1 caregiver could supervise the child while the other participated more actively in the discussion. All facilitators were female members of the research team with prior experience in qualitative interviewing and no prior personal or professional relationship with participants. The same facilitators also conducted the follow-up interviews to ensure continuity and rapport. No repeat interviews were conducted; each participant took part in a single FGD. In addition to the FGDs, short individual interviews were conducted with the same FGD participants to clarify points raised during the group discussion and to allow participants to elaborate on topics they were not comfortable discussing in a group setting. No separate pool of interview participants was recruited; all interviewees had already taken part in an FGD. These interviews were conducted immediately after the FGDs or were scheduled within the same week (as follow-ups for deeper insights). This approach ensured that group-generated themes were accurately understood and provided an opportunity to validate and deepen emerging insights, enhancing the credibility of the analysis.

XSL•FO
RenderX

Given the sensitive nature of ASD screening and emerging AI technologies, moderators received additional training to manage emotionally charged discussions and foster a safe, nonjudgmental environment. To facilitate understanding, participants received a brief explanatory note on AI decision-making, stating:

> *This allows you to understand the logic behind AI's decision-making in a simple and clear way; without needing advanced technical knowledge. It's like AI showing you how it arrived at its decision, so you know it's not just a random guess.*

All individuals invited agreed to participate, and no invited participants declined or withdrew.

Moderator Training and Power Imbalance Considerations

To address potential power dynamics, FGDs were conducted separately for health care professionals and parents, with moderators trained to foster inclusive participation and encourage open dialogue. Additional individual interviews were conducted to capture sensitive experiences that participants might not have shared in group settings (Multimedia Appendix 1 provides detailed description of moderator training and power-imbalance considerations). All focus groups and individual interviews were conducted by a research team of female public health researchers at the National Research Centre (NRC) with formal training in qualitative interviewing and prior experience in ASD-related services. None had a direct clinical care relationship with the participating families or children, which helped minimize role-related coercion while encouraging open and candid discussion of views on ASD screening and AI tools.

Discussion Guide Topics

Focus group discussions were guided by 2 core research questions: (RQ1) What barriers and contextual challenges do parents and health care providers identify in current ASD screening, referral, diagnosis, and access-to-support pathways in Egypt? (RQ2) How do parents and health care providers perceive the acceptability and feasibility of AI-powered ASD screening tools, including perceived value, usability, trust, and implementation requirements? The full discussion guide is available in Multimedia Appendix 2.

Data Recording and Transcription

All FGDs and interviews were audio-recorded and transcribed verbatim. Field notes captured nonverbal cues and group dynamics. Discussions were conducted in Arabic and professionally translated into English to ensure linguistic and conceptual accuracy during thematic analysis. Initial coding was conducted on the original Arabic transcripts by bilingual members of the research team, with English translations used to support team-based analysis and reporting. Arabic and English versions were reviewed side by side, and any discrepancies or ambiguous phrases were resolved through discussion between the coders and the professional translator to preserve the intended meaning. Verbatim transcripts were not formally returned to participants for checking due to resource and logistical constraints; instead, analytic accuracy was supported through verbatim transcription, independent transcript review, peer debriefing, and limited member checking with a purposive subset of participants during follow-up calls to clarify key interpretations.

In reporting quotations, we used anonymized identifiers indicating participant role, gender, and governorate (for example, "Parent-F, Fayoum" and "Pediatrician-M, Cairo") to illustrate the diversity of perspectives while preserving confidentiality.

Data Analysis

The analysis followed Braun and Clarke's (2019) reflexive thematic analysis framework, which involves six iterative phases (1) familiarization with the data through repeated reading of transcripts, (2) generating initial inductive codes, (3) searching for candidate themes by clustering related codes, (4) reviewing and refining themes for coherence, (5) defining and naming themes with operational clarity, and (6) producing the final analytic report integrating quotes and interpretations. NVivo software (version 14; Lumivero) supported coding and data organization. Coding was led by 2 bilingual public health researchers with training in qualitative methods and experience in ASD-related research, supported by a senior qualitative methodologist; all are based at the NRC, and none were involved in direct clinical care of the participating families.

The 2 primary analysts independently read and annotated the transcripts, generated initial inductive codes, and then met regularly to compare and refine the coding frame and candidate theme map, not to calculate interrater reliability but to deepen reflexive engagement with the data, consistent with reflexive thematic analysis. All FGD and interview transcripts were imported into NVivo and analyzed as a single qualitative data corpus using the same inductive coding frame; the data-collection mode (focus group vs individual interview) was retained as a case attribute to enable comparison of patterns across modes when relevant rather than conducting 2 separate analytic processes. For example, codes such as "AI helps nonspecialists," "AI as first filter," "AI speeds referrals," and "AI cannot replace doctors" were clustered into a subtheme labelled "AI as a supportive triage tool," which was then situated within the broader theme "AI as a supportive tool rather than a replacement for clinicians."

In line with reflexive thematic analysis, counts (n/N) for parents and health care providers were used descriptively in the tables as signposts of pattern salience; that is, to indicate how many participants contributed to a given code or theme rather than as statistical estimates of prevalence or as the basis for inferential claims [24]. In addition, emergent themes were examined in relation to the HBM and TAM to provide a theoretical lens for interpreting stakeholder perceptions of AI-powered ASD screening. This deductive overlay complemented the inductive analysis, enabling findings to be systematically linked with established models of health behavior and technology adoption.

During analysis, we also explored potential differences in perspectives across health care professional subgroups (pediatricians, nurses, psychologists, and child development specialists); however, given the modest numbers in each

XSL•FO
RenderX

subgroup and the substantial overlap in core themes, no systematic divergence in overarching patterns was identified, and themes are therefore reported at the combined "health care provider" level, with subgroup nuances highlighted where relevant. Reflexive journals were maintained by the primary analysts to document assumptions (eg, expectations about the benefits of AI in overstretched public services) and to record how these were challenged by the data; these reflections informed theme naming and interpretation (eg, reframing early "AI will solve gaps" codes as "cautious optimism conditioned by system constraints"). Interpretive rigor was maintained through reflexive team discussions, staged member checking with a purposive subset of participants, peer debriefing with a senior qualitative researcher, and triangulation of perspectives across participant groups. A detailed description of the 6 analytic phases, including examples of coding procedures and reflexivity practices, is provided in Multimedia Appendix 1.

Ensuring Trustworthiness

This study applied classic criteria of credibility, transferability, dependability, and confirmability [25-27]. Credibility was established through triangulation of FGDs and interviews, member checking, and peer debriefing. Member checking occurred at 2 stages. First, at the end of each focus group and interview, moderators briefly summarized the main discussion points and explicitly invited participants to correct, nuance, or add to these summaries, allowing immediate verification of the researchers' understanding. Second, after preliminary theme development, the research team prepared concise written summaries of the key themes and a small set of illustrative quotes and contacted a purposive subset of parent and clinician participants from different governorates (via brief follow-up calls or messaging apps) to review these summaries. Participants were asked whether the themes and examples reflected their experiences, and their feedback was used to refine thematic boundaries, adjust wording of themes, and address any apparent misinterpretations. Formal transcript-based member checking (returning full transcripts to all participants) was not undertaken due to resource and logistical constraints; however, this staged, selective approach allowed verification of core interpretations without recontacting every participant or overburdening busy clinicians and caregivers.

Transferability was supported by thick descriptions of participant experiences, service settings, and contextual factors across diverse regions, enabling readers to assess how findings may apply to similar LMIC contexts. Dependability was ensured by documenting analytic procedures and maintaining an audit trail. The audit trail included decision logs on coding, theme development, and changes to the analytic framework over time. Confirmability was strengthened through reflexivity and integration of direct participant quotes into findings. The primary analysts maintained reflexive journals to document their assumptions and how these were challenged by the data; these reflections were discussed in team meetings and peer debriefing sessions with a senior qualitative researcher, supporting critical examination of alternative interpretations and reducing the risk of unchecked researcher bias. Triangulation across data sources (FGDs and follow-up interviews) and participant groups (parents and health care providers) further contributed to the credibility and confirmability of the analysis. Details of the trustworthiness measures are provided in Multimedia Appendix 1.

Ethical Considerations

Ethical approval for this study was granted by the NRC of Egypt (approval number 2705250141). All participants provided written informed consent prior to participation. Before participation, individuals received detailed study information, including objectives, procedures, confidentiality measures, and their right to withdraw at any stage without justification or consequences. Participants had ample time to review consent forms, ask questions, and opt out if desired. To protect privacy, all data were anonymized, securely stored, and accessible only to authorized researchers. Participants received reimbursement for travel costs and a modest goodwill payment of US $20, as approved by the NRC ethical committee and funded by Ulster University (cost center 71799); these tokens were intended to offset participation-related expenses rather than act as coercive incentives. Concerns raised by participants were incorporated into the analysis to highlight ethical implications of AI integration in screening of neurodiversity disorders [28,29]. Researchers followed international ethical guidelines, including the Declaration of Helsinki and General Data Protection Regulation data protection standards [30].

Through comprehensive consent procedures, strict confidentiality measures, and reflexive ethical discussions, this study upheld responsible and ethical qualitative research practices. Participation was fully voluntary, and participants were informed that they could withdraw at any time without consequences. Data were collected in a confidential manner, deidentified and delinked from personal identifiers, and stored in a secure location. In addition, data collection and handling complied with Egypt's personal data protection framework; identifiable information was stored on password-protected institutional servers at the NRC in Cairo, with access restricted to the core research team, and deidentified analytic datasets were stored separately and shared only in encrypted form. Information disclosure ("making sure participants understand") was guaranteed according to the recommendations of prior work on Egyptian patients' and guardians' perceptions of clinical informed consent [31].

This study was designed, conducted, and reported in accordance with the COREQ (Consolidated Criteria for Reporting Qualitative Research), and a completed COREQ checklist is provided in Multimedia Appendix 3.

Results

Overview

Table 1 summarizes the composition of the study participants, including parents and health care providers, across 6 Egyptian governorates. Participants were recruited from governorates representing diverse geographic and sociodevelopmental settings across Egypt, including urban (Cairo), urban/semiurban (Giza and Damietta), mixed urban–rural (Dakahlia and Assuit), and predominantly rural/semiurban (Fayoum). The sample comprised clinicians (pediatricians, psychologists, family medicine physicians, speech therapists, nurses, and

XSL•FO
RenderX

health-information specialists) and parents of children with diagnosed ASD. Participant numbers varied by site depending on service availability and community size rather than equal quotas per governorate, consistent with qualitative sampling principles.

Table 1. Participant description (N=49) by geographic location, professional role, and participant type.

Governorate/region [32-36]	Setting type	Parents, n/N	Health care providers, n/N	Health care provider roles
Dakahlia (Lower Egypt, mixed urban–rural)	Autism centers, outpatient clinics	3/21	4/28	Pediatrician, psychologist, head nurse, and family medicine physician
Cairo (cities, urban)	MCH[a] centers	5/21	7/28	Pediatricians, 2 psychologists, head nurse, 2 general practitioners, and health-information specialists
Giza (cities, urban/semiurban)	Early screening units	3/21	3/28	Pediatrician, speech therapist, and MOHP[b] officer
Fayoum (Upper Egypt, rural/semiurban)	Primary health units	5/21	6/28	Pediatrician, 2 psychologists, head nurse, PHC[c] physician, and health-information specialist
Damietta (coastal, urban/semiurban)	Pediatric clinics, NGO[d]	2/21	4/28	Pediatrician, 2 psychologists, and speech therapist
Assuit (Upper Egypt, mixed urban–rural)	Community health centers	3/21	4/28	Pediatrician, psychologist, head nurse, and PHC Physician

[a]MCH: maternal and child health care.

[b]MOHP: Ministry of Health and Population.

[c]PHC: primary health care.

[d]NGO: nongovernmental organization.

In the main thematic tables (Tables 2-6), we therefore report absolute counts with explicit denominators for each stakeholder group (for example, 20/21 parents; 25/28 health care providers) rather than percentages, to reflect the qualitative, nonstatistical nature of the data while still allowing readers to assess the salience of each view across participants [24,37].

Table 2. Theme 1: reported challenges in autism diagnosis and pathways, mention counts (n/N) by participant group.

Challenge[a]	Parents, n/N	Health care providers, n/N
Lack of ASD[b]-specific training for health care professionals	18/21	21/28
No standardized ASD screening protocols	13/21	16/28

[a]Values represent the number of participants (n) who spontaneously mentioned each challenge. Denominators (N) refer to total participants in each group (parents=21 and health care providers=28). These counts indicate thematic salience, not prevalence or consensus. Illustrative quotations capture both typical and divergent views to reflect the full spectrum of perspectives and should not be interpreted as statistical estimates.

[b]ASD: autism spectrum disorder.

Table 3. Theme 2: limited awareness, social stigma, and misconceptions about autism, mention counts (n/N) among parents.

Challenge[a]	Parents, n/N
Stigma and fear of judgment	11/21
Misconceptions (autism as temporary or due to poor parenting)	8/21
Lack of parental awareness about ASD[b] screening	7/21

[a]Values represent the number of participants (n) who spontaneously mentioned each challenge. Denominators (N) refer to total participants in each group (parents=21). These counts indicate thematic salience, not prevalence or consensus. Illustrative quotations capture both typical and divergent views to reflect the full spectrum of perspectives and should not be interpreted as statistical estimates.

[b]ASD: autism spectrum disorder.

XSL•FO
RenderX

Table 4. Theme 3: reported barriers to accessing autism diagnostic and support services: mention counts (n/N) by participant group.

Barriers[a]	Parents, n/N	Health care providers, n/N
Shortage of ASD[b] specialists in local areas	13/21	23/28
Long waiting times for appointments	12/21	11/28
High cost of private consultations	9/21	11/28
Travel distance/time burden	15/21	10/28

[a]Values represent the number of participants (n) who spontaneously mentioned each barrier. Denominators (N) refer to total participants in each group (parents=21 and health care providers=28). These counts indicate thematic salience, not prevalence or consensus. Illustrative quotations capture both typical and divergent views to reflect the full spectrum of perspectives and should not be interpreted as statistical estimates.

[b]ASD: autism spectrum disorder.

Table 5. Theme 4: perceived role of artificial intelligence (AI) in autism screening, mention counts (n/N) by participant group.

Factor[a]	Parents, n/N	Health care providers, n/N
Support for AI-based screening	9/21	19/28
Need for culturally adapted tools	12/21	16/28
Preference for explainable AI-generated results	15/21	17/28

[a]Values represent the number of participants (n) who spontaneously mentioned each factor. Denominators (N) refer to total participants in each group (parents=21 and health care providers=28). These counts indicate thematic salience, not prevalence or consensus. Illustrative quotations capture both typical and divergent views to reflect the full spectrum of perspectives. Figures should not be interpreted as statistical estimates.

Table 6. Theme 5: reported concerns about artificial intelligence (AI) adoption, mention counts (n/N) by participant group.

Concern[a]	Parents, n/N	Health care providers, n/N
Data privacy and security risks	17/21	23/28
Fear of AI replacing human judgment	6/21	—[b]
Limited digital literacy among parents	4/21	—[b]
User-friendly interface for non-specialist health care providers	—[b]	21/28

[a]Values represent the number of participants (n) who spontaneously mentioned each concern. Denominators (N) refer to total participants in each group (parents: 21; health care providers: 28). These counts indicate thematic salience, not prevalence or consensus. Illustrative quotations capture both typical and divergent views to reflect the full spectrum of perspectives. Figures should not be interpreted as statistical estimates.

[b]Not applicable.

This analysis drew upon 7 FGDs involving key health care providers, health professionals, and parents of children with ASD across 6 governorates in Egypt. Each group included approximately 7 participants, totaling 28 health care professionals and 21 parents. Health care professionals included heads of departments for children with special needs, chief technicians, physicians, nurses, speech therapists, and health-information specialists. The mean age of health care providers was 43.3 (SD 11.0) years. The mean age of parents was 44.3 (SD 10.6) years. Among their children, 66.7% (14/21) were male and 33.3% (7/21) female, with a mean age at diagnosis of 2.2 (SD 0.6) years; 80.9% (17/21) were classified as having moderate autism. Educational attainment varied across groups: among health care providers, 39.3% (11/28) held graduate degrees, 32.1% (9/28) postgraduate qualifications, and 28.6% (8/28) technical education. Among parents, 57.1% (12/21) held graduate degrees, 23.8% (5/21) postgraduate qualifications, and 19.1% (4/21) secondary or technical education.

Thematic analysis of the focus group and interview data identified 5 major themes that captured stakeholder perspectives on the feasibility, acceptability, and challenges of AI-powered autism screening in Egypt. Absolute counts with denominators (n/N) are presented to indicate how many participants in each group (parents and health care providers) raised each issue during the discussions. Importantly, these figures should be interpreted as indicators of thematic salience rather than statistical prevalence or consensus, given the exploratory qualitative design and the fact that not all participants were asked identical questions in a standardized format.

To reinforce this, table footnotes clarify that counts (n/N) reflect how frequently topics arose rather than the degree of agreement across the sample. While the abstract summarizes the 5 overarching themes conceptually according to AI adoption factors, the following Results section presents them in the empirical order in which they emerged during thematic analysis.

These figures should be viewed as indicators of thematic salience rather than measures of agreement or prevalence, as not all participants were asked identical questions in a

XSL•FO
RenderX

standardized format. Instead, they reflect how frequently particular topics arose during the conversations and the relative emphasis placed on them by participants. To ensure analytic rigor, the interpretation of these findings was informed by reflexive consideration of the research team's positionality, clinical experience, and prior exposure to autism-related work in Egypt. The 5 themes emerging from the analysis were delays in autism diagnosis and inconsistent pathways; limited awareness, social stigma, and misconceptions about autism; barriers to accessing diagnostic and support services; perceived role of AI in autism screening; and balancing trust, transparency, and data protection in AI screening.

Tables 2-6 summarize the 5 major themes derived from this analysis. Each table presents key subthemes, absolute counts with denominators for parents and health care providers, representative participant quotations, and brief interpretive notes. This structure facilitates comparison across themes while preserving a clear link between each table and its corresponding narrative subsection in the Results. Themes 1-3 primarily address RQ1 by characterizing current pathway-related and access-related barriers to timely ASD screening and support, whereas Themes 4 and 5 primarily address RQ2 by examining perceived acceptability, feasibility, and trust-related considerations for AI-enabled ASD screening.

Theme 1: Delays in Autism Diagnosis and Inconsistent Pathways

In relation to RQ1 (current pathway-related barriers), most health care providers reported lacking specific training in ASD diagnosis, and many raised concerns about the absence of standardized screening protocols. These delays were most often attributed to long waiting lists, limited specialist availability, and the absence of standardized referral pathways. Diagnostic inconsistency was another recurring concern, with many parents and more than half of health care professionals noting that children often received conflicting assessments from different providers.

One parent described the protracted wait for assessment:

> *I had to wait over six months for my child to be assessed, and even then, I wasn't sure if the diagnosis was correct because different doctors gave me different answers.* [Parent-M, Cairo]

This quote illustrates 2 critical barriers: significant delays due to workforce shortages and diagnostic inconsistency arising from the absence of standardized screening protocols. Together, these factors contribute to late interventions and parental uncertainty.

Health care professionals echoed these concerns, with one psychologist noting:

> *We see many children after they have been to three or four different doctors, each giving a different opinion.* [Psychologist-F, Assuit]

Parents reported long diagnostic delays:

> *I went to three doctors before someone said autism" (Parent, Fayoum)—while healthcare providers cited limited resources and high caseloads as major obstacles. Healthcare provider described "a system that reacts late, not one that screens early.* [Speech therapist-M, Damietta]

This reflects fragmentation of care pathways and highlights why nearly two-thirds of parents and half of health care providers in our sample emphasized the urgent need for standardized assessment tools. Importantly, participants also drew connections between these diagnostic inconsistencies and later discussions of AI, suggesting that standardized digital tools could harmonize assessments and reduce conflicting diagnoses (Theme 5).

Transitioning from diagnostic delays, participants also described how broader societal perceptions of autism further compound these challenges, as explored in Theme 2.

Theme 2: Limited Awareness, Social Stigma, and Misconceptions About Autism

Cultural stigma and misinformation emerged as significant barriers to early ASD diagnosis and were raised by many parents. Misconceptions, such as autism being temporary or caused by poor parenting, were frequently reported, and several parents were unaware that early screening was even possible.

Parents described community-level misunderstandings, such as attributing autism symptoms to poor parenting, stubbornness, or spiritual causes.

> *People told me my child was just naughty, or that I was spoiling him too much.* [Parent-F, Assuit]

This highlights the community-level stigma faced by more than half of parents in the sample, reinforcing how misconceptions delay help-seeking behavior. Health care professionals also noted that misconceptions among general practitioners sometimes led to misdiagnosis, resulting in children being referred to speech therapy or psychiatry without autism-specific assessment.

A pediatrician from Fayoum explained:

> *Some doctors still think autism means a child cannot speak at all, so they miss milder cases.* [Pediatrician-M, Fayoum]

By linking parental stigma with professional misconceptions, participants emphasized that awareness deficits are multilayered, spanning families, communities, and frontline medical providers. These deficits not only prolong diagnostic delays but also shape the systemic barriers discussed in Theme 3. In particular, stigma and misconceptions were seen as amplifying existing access problems, since families already hesitant to seek help often face structural hurdles once they do attempt to engage with services.

Theme 3: Barriers to Accessing Diagnostic and Support Services

Whereas Theme 2 captured attitudinal and cultural barriers to early ASD identification, Theme 3 focuses on structural and logistical constraints that hinder access to timely diagnosis and ongoing support. Barriers to service access were reported by many parents and most health care providers. These included long travel distances, high service costs, and the scarcity of autism-specialized centers, particularly in rural areas.

XSL•FO
RenderX

One mother from Fayoum explained:

> *The nearest centre is two hours away, and I cannot afford to travel every week.* [Parent-F, Fayoum]

Health care providers recognized these structural barriers. A psychologist from Cairo noted:

> *Families in rural areas have very limited options. Sometimes we try to offer online consultations, but internet quality is a problem.* [Psychologist-M, Cairo]

Financial constraints were also significant. One parent shared:

> *Therapy costs a lot, and we have to choose between paying for treatment or other basic needs.* [Parent-M, Dakahlia]

These findings underline the intersection of geographic, economic, and infrastructural barriers, which can delay both diagnosis and intervention. This was particularly salient among parents and was also frequently highlighted by health care providers, underscoring the disproportionate burden borne by families navigating multiple constraints simultaneously. Against this backdrop of limited access, many participants suggested that AI tools could potentially bridge these gaps by offering more accessible first-line screening, a possibility explored further in Theme 4. When discussing digital health options, parents and providers repeatedly pointed out that limited internet access and affordability challenges could constrain the feasibility of AI-powered tools even as they recognized their potential benefits.

Theme 4: Perceived Role of AI in Autism Screening

In relation to RQ2 (acceptability and feasibility of AI-enabled screening), optimism about AI-powered tools was expressed by many health care professionals and a substantial proportion of parents, who emphasized the potential for early risk identification in underserved areas. Some pediatricians and child development specialists highlighted AI's promise for accelerating referrals and supporting overburdened systems, while frontline nurses and primary care physicians more often focused on how such tools might fit within existing workloads and communication with families.

A pediatrician in Giza stated:

> *AI could help us identify at-risk children sooner, so we can refer them before the delay becomes too long.* [Pediatrician-F, Giza]

Parents similarly expressed cautious optimism:

> *If it's accurate and easy to use, I would try it for my child.* [Parent-F, Assuit]

Most parents viewed AI as a tool that could support earlier detection and broaden access to assessment, particularly in areas with limited specialist services:

> *If the app can tell me early that my child is at risk, I can go to the doctor sooner instead of waiting for years.* [Parent-F, Fayoum]

However, a minority expressed discomfort with technology-mediated interactions, preferring face-to-face communication:

> *I prefer someone to talk to me directly, not an app that gives me a score.* [Parent-M, Cairo]

These divergent views highlight the need for AI screening tools to be positioned as supportive aids rather than replacements for human care. Overall, enthusiasm was clear, but participants emphasized that meaningful adoption would require addressing structural barriers (Theme 3) and building trust through transparency and usability (Theme 5).

Theme 5: Balancing Trust, Transparency, and Data Protection in AI Screening

Concerns about trust and usability emerged strongly, with many parents and health care providers citing data privacy as their foremost worry. Psychologists often linked these concerns to existing stigma and the risk of labeling, whereas pediatricians tended to emphasize clinical responsibility for acting on AI-generated risk scores. Participants also frequently emphasized the importance of clear, understandable outputs, and some parents expressed concern that AI might replace human judgment.

One pediatrician stressed the importance of maintaining human oversight:

> *AI should not replace a specialist. A diagnosis should come from a doctor, not an algorithm.* [Pediatrician-F, Dakahlia]

Parents voiced privacy concerns:

> *If I use an app to screen my child, how do I know my data won't be misused? I don't want personal information shared without my consent.* [Parent-M, Damietta]

Digital literacy emerged as an additional barrier, especially in rural contexts:

> *I am not good with technology. If the AI screening tool is too complicated, I won't know how to use it properly for my child.* [Parent-F, Fayoum]

Health care providers also highlighted usability concerns. A physician explained that:

> *AI screening tools must be easy to use, especially for general practitioners and nurses who are not ASD specialists.* [PHC Physician-F, Assuit]

A nurse echoed this point, adding:

> *If the system is too complicated, healthcare providers won't use it, and parents will struggle to understand the results.* [Head Nurse-F, Cairo]

However, a few younger health care providers expressed optimism about automation:

> *If AI can handle screening, we can focus on therapy and follow-up.* [Speech therapist-M, Damietta]

This generational contrast reflects differing comfort levels with digital integration in clinical care. Across all sites, participants linked AI adoption to issues of fairness, privacy, and equity.

One policymaker warned:

XSL•FO
RenderX

> *Technology must not widen the gap between rich and poor families.* [MOH officer-F, Giza]

Many parents wanted assurances of data protection, with one noting:

> *Once data is uploaded, we lose control. That scares us.* [Parent-M, Fayoum]

In contrast, a few tech-savvy participants felt that:

> *Data collection is normal—everything is online now.* [Technology specialist-F, Cairo]

Their comfort with digital data contrasts with the majority's apprehension, illustrating generational and digital literacy divides. To accommodate both digitally inexperienced parents and nonspecialist providers, participants underscored the need for intuitive design. Trust in AI was therefore linked not only to accuracy and cultural fit but also to privacy protections, clarity of outputs, and user-friendliness.

This final theme loops back to Theme 1's concern with diagnostic inconsistencies, reinforcing that AI adoption could help standardize pathways if implemented with strong privacy protections and clear communication.

In addition to the 5 major themes, several divergent perspectives emerged. For example, while many parents emphasized accessibility as a key advantage, some questioned whether AI tools might replace human empathy. Similarly, a minority of clinicians expressed optimism about algorithmic decision support, contrasting with the dominant skepticism. These variations illustrate the heterogeneity of experiences and highlight areas for tailored implementation strategies. To support interpretation, the 5 themes are also conceptually mapped onto HBM and TAM constructs; this integration is summarized in Figure 1 and elaborated in the Discussion.

Figure 1. Conceptual mapping of emergent themes to Health Belief Model (HBM) and Technology Acceptance Model (TAM) constructs.

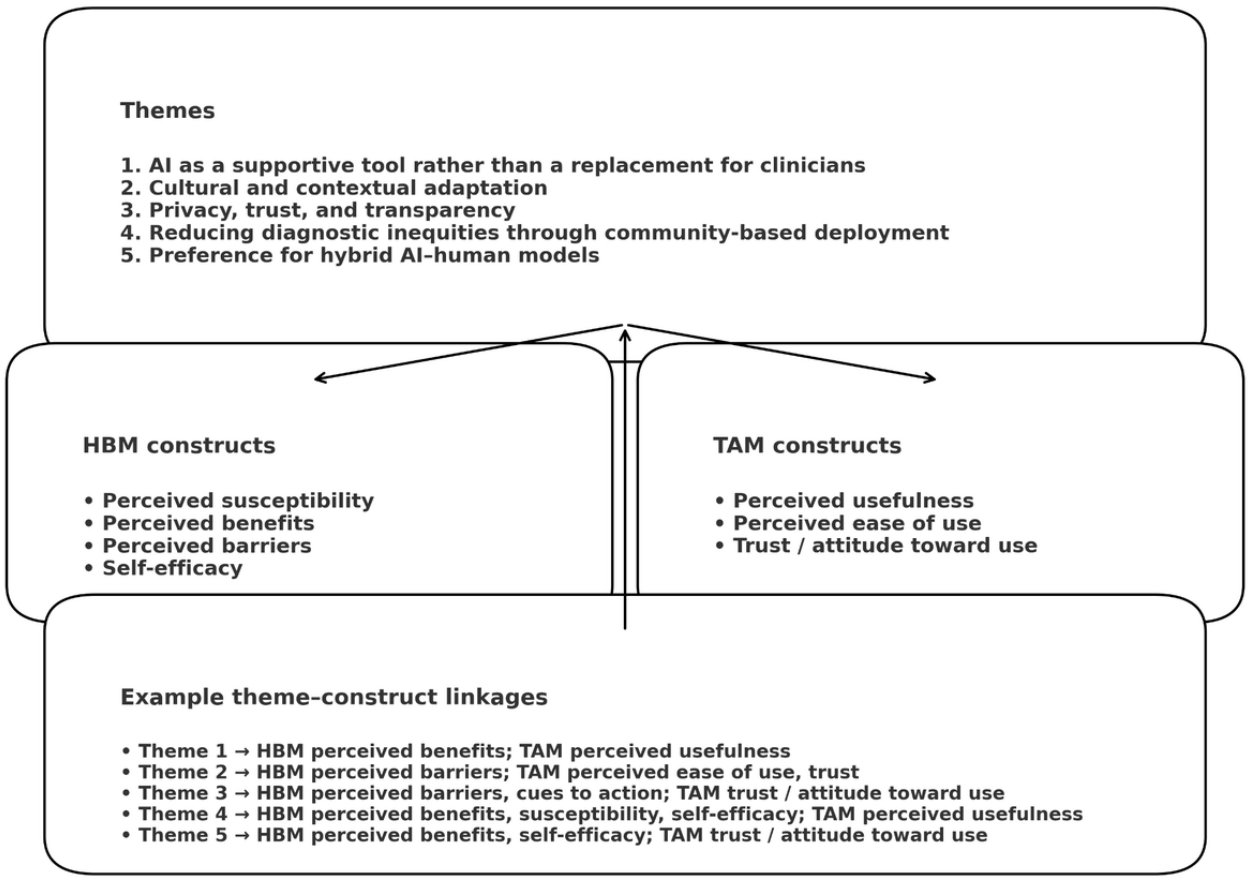

Discussion

Overview

This study explored stakeholder perspectives on the feasibility, acceptability, and perceived ethical and practical enablers and barriers to implementing AI-powered tools for early ASD screening within the Egyptian health system. Overall, participants recognized that AI-powered tools could help bridge diagnostic gaps, support earlier detection, and extend services to underserved areas. At the same time, they highlighted substantial barriers that must be addressed for successful adoption, including systemic delays and fragmented care pathways, lack of standardized diagnostic practices, pervasive stigma and low awareness, concerns about cultural appropriateness and transparency of AI decisions, and limitations in digital infrastructure and workforce training.

As illustrated in Figure 1, the themes align closely with core constructs of the HBM and TAM. Perceptions of AI as a

XSL•FO
RenderX

supportive triage tool, particularly for nonspecialists in underserved areas, reflect strong perceived benefits and perceived usefulness, which are key drivers of screening uptake in HBM- and TAM-informed frameworks. In contrast, concerns about data privacy, algorithmic opacity, and cultural mismatch map onto perceived barriers, perceived ease of use, and trust-related TAM constructs and help explain ambivalence despite overall cautious optimism. The preference for hybrid AI–human models illustrates how trust, self-efficacy, and attitudes toward use are negotiated in practice. Stakeholders were more comfortable when AI outputs were explicitly embedded within human oversight and relational care. These relationships are summarized in the conceptual mapping presented in Figure 1 and highlight potential leverage points for implementation, including strengthening perceived benefits and usefulness, reducing perceived barriers, and actively building trust and self-efficacy around AI tools.

Viewed through this HBM/TAM lens, the findings highlight both the promise and limitations associated with AI-driven neurodiversity screening. Consistent with prior research, AI-based screening tools offer a scalable, accessible, and cost-effective solution to bridging gaps in ASD diagnosis, particularly in underserved regions [13,38]. In keeping with Braun and Clarke's [24] call for thematic analysis to move beyond description toward theoretically informed interpretation, our findings link participants' accounts to established constructs in health behavior and technology adoption frameworks. However, stakeholder concerns regarding algorithmic bias, data privacy, and ethical implications suggest that AI models must be rigorously validated and co-designed with input from health care professionals, caregivers, and neurodivergent individuals [39,40]. Our emphasis on culturally sensitive AI models therefore reflects a call for AI tools to follow and extend the same principles of careful translation, local validation, and reflexive use that are increasingly expected of human-administered instruments, rather than holding AI to a fundamentally different standard [16,28,29,41-48].

Viewed in a broader context, our findings both converge with and extend international evidence on AI-enabled and digital approaches to autism and neurodevelopmental screening. Studies from high-income countries have similarly reported cautious optimism about AI and mobile tools, alongside concerns regarding algorithmic transparency, data privacy, and the risk of depersonalized care [9,29,49,50]. Our participants' preference for hybrid AI–human models echoes this literature, which consistently emphasizes the importance of human oversight and relational communication when delivering AI-generated outputs [9,28,40,51]. By bringing these strands together, this study contributes a LMIC-based perspective that both confirms global concerns and foregrounds context-specific implementation challenges [15,16,47,48].

From a reflexive standpoint, the research team's professional backgrounds in public health, autism screening, and health care systems may have sensitized the analysis toward issues of access, policy implications, and sociocultural context. This positionality informed both theme construction and interpretation, particularly in highlighting equity considerations for rural and underserved groups. Explicitly acknowledging positionality strengthens the transparency of analytic choices, in line with qualitative rigor guidelines [52].

AI tools have demonstrated high accuracy in speech pattern recognition, facial analysis, and behavioral assessments, improving early-stage ASD risk detection [9]. Consistent with earlier research, this study supports a hybrid approach, where AI functions as a decision-support tool rather than a diagnostic substitute [41,42,51].

AI Skepticism vs Parental Trust: Divergent Stakeholder Perspectives

One of the study's key findings is the contrast between health care professionals' AI skepticism and parents' greater openness to AI-driven screening. Many parents in the study viewed AI as a potential tool for early intervention, particularly in regions with limited specialist access. As one parent noted, "If it can tell me early that something is wrong, I don't care if it's a machine; as long as I can get help faster." This quote illustrates how parents interpret AI's value pragmatically, privileging speed and access over abstract concerns about reliability.

Conversely, many health care professionals in this study questioned AI's reliability, highlighting concerns that machine learning models might misinterpret symptoms, particularly for nonverbal children or those with atypical ASD presentations. This skepticism reflects broader concerns about algorithmic opacity and AI's inability to capture the complexity of neurodivergent traits [40,53-55].

Ethical Considerations in AI Adoption for ASD Screening

The integration of AI-driven screening tools raises pressing ethical concerns that must be addressed for widespread adoption. One major challenge is algorithmic bias, as many models are trained on Western datasets, limiting their applicability to diverse global settings [15]. This is particularly relevant in Egypt, where sociocultural differences in autism perception and diagnosis criteria may not align with Western-centric models. Participants' concerns mirror broader debates emphasizing the importance of cultural adaptation [47].

Data privacy emerged as another primary concern. Trust was closely tied to transparency in data governance and safeguards [40,51].

Finally, participants worried about misdiagnosis and emotional distress—false positives producing anxiety and false negatives delaying intervention. These dual anxieties underscore that participants' accounts should be understood not as statistical generalizations but as socially constructed experiences of risk and uncertainty [46].

Bridging Urban vs Rural Disparities in ASD Diagnosis

A key contribution of this study is its comparative analysis of urban and rural ASD screening barriers. While urban parents and health care professionals reported long waiting times and high costs as primary challenges, rural participants highlighted severe shortages of ASD specialists and a widespread lack of awareness about screening tools. These findings align with prior research on health care disparities in low-resource settings,

XSL•FO
RenderX

which suggests that AI-powered screening must be adapted to infrastructural limitations and literacy levels of diverse communities [8,43].

Theoretical Contributions: Behavior Change, Trust, and Stigma

This study contributes to theoretical frameworks explaining health technology adoption and behavior change in autism screening. Consistent with the HBM, parental decisions regarding AI-powered screening tools were shaped by their perceptions of susceptibility to ASD, the benefits of early detection, and concerns about potential barriers particularly stigma and data misuse [56]. Participants' concerns about stigma align with Goffman's theory of social labeling, where fear of judgment discourages disclosure [57].

Furthermore, trust-based adoption models and the TAM were evident in health care professionals' skepticism and parents' mixed acceptance driven by concerns over ease of use, usefulness, and trust in AI-generated outputs [55,58,59].

Policy Recommendations With Implementation Strategies

Building on these empirical and theoretical insights, several practical and policy-level recommendations can be made to support the ethical and effective implementation of AI-driven ASD screening in Egypt and similar contexts.

AI tools should be trained and validated on diverse datasets that reflect the cultural, linguistic, and behavioral characteristics of Egyptian children [13,15,16,39,49,50]. AI systems can follow the same approach by using locally collected training data, documenting the provenance and limitations of datasets, and involving Egyptian clinicians, caregivers, and autistic individuals in co-design and evaluation [16,28,29,47-49]. Such steps would help ensure that AI does not simply reproduce Western diagnostic assumptions but is genuinely responsive to local sociocultural realities [15,16,28,29,48,49].

AI adoption must be underpinned by national legal frameworks that regulate data storage, privacy, and consent. Without such frameworks, participants' anxieties about surveillance and discrimination could limit uptake, as also reported in other LMIC contexts [15].

Participants emphasized that parents' trust depends on transparent communication regarding how AI screening tools operate, where data are stored, and who has access. Transparent governance and communication have been shown to improve uptake of AI-based health tools [51].

A recurring theme was the value of combining AI outputs with human judgment. Parents and health care providers agreed that AI should assist, not replace, clinicians. This aligns with global guidance emphasizing AI as an augmentative tool [9], and reflects participants' preference for "shared accountability" between humans and technology.

To address diagnostic inequities, AI-enabled screening should be deployed in primary care facilities and community health centers in rural areas. Participants in rural settings described stigma, lack of awareness, and absence of specialists as barriers, reinforcing the importance of deploying accessible, low-literacy screening tools. Training community health providers to use AI tools could decentralize early detection and improve referral pathways. This approach mirrors other successful models of digital health integration in low-resource contexts [54].

Successful AI adoption will depend not only on clinical readiness but also on public understanding and digital competence. Parents and caregivers expressed both curiosity and apprehension about AI, underscoring the need for structured awareness campaigns. Digital literacy training has been shown to improve uptake of AI tools in LMIC health care settings [47].

Stigma remains a powerful deterrent to early ASD diagnosis. Participants' accounts revealed that some families avoid clinics out of fear of labeling, echoing prior national studies on autism stigma in Egypt [8].

Health care providers and frontline staff require training not only in the technical use of AI tools but also in communicating AI results to families with empathy and sensitivity. Participants stressed the risk of "cold delivery" of AI outputs; thus, training must emphasize the human element of care to prevent alienation. Building capacity in both technical literacy and interpersonal skills will ensure AI integration enhances, rather than undermines, the therapeutic alliance between families and health care systems.

Strengths

This study is among the first to explore stakeholder perspectives on AI-powered autism screening in Egypt, offering valuable insights for the development and implementation of such tools in LMICs. The use of reflexive thematic analysis allowed for the identification of context-specific enablers and barriers, drawing from the lived experiences of both health care professionals and parents. Purposive sampling ensured that participants represented a wide range of geographic and professional backgrounds, increasing the diversity of perspectives captured. Together, these design choices support conceptual transferability to other LMIC settings facing similar diagnostic and infrastructural challenges. All parent participants had children aged 6 years or younger, anchoring parental perspectives in early detection contexts.

Limitations

As with many qualitative designs, findings are context-bound and not intended for statistical generalization. Instead, the aim was to generate nuanced, situated understandings that may have conceptual transferability to similar LMIC contexts. Although the sample achieved geographic variation, the number of FGDs was necessarily constrained by logistical and resource limitations, which shaped the scope of representation. The sample was not designed to be proportionate to population distributions across urban and rural areas and was not powered for formal statistical comparisons between geographic strata; rather, rural and urban participants were included to surface key contextual patterns and implementation considerations. As discussed above, the research team's positionalities, including professional experience in public health, autism research, and digital health, may have influenced data interpretation, particularly in prioritizing equity and access themes; this

XSL•FO
RenderX

reflexive awareness was explicitly documented and used to critically interrogate analytic decisions.

Although all individuals who were invited agreed to participate, which may reflect greater interest or comfort with discussing AI and autism, we attempted to mitigate this potential selection bias by recruiting across multiple governorates, facility types, and professional roles to capture a wide range of perspectives. However, socioeconomic position was not systematically measured or analyzed, and any socioeconomic variation in the sample is therefore inferred indirectly through differences in governorate and service setting rather than examined explicitly. In addition, because all participating parents had children aged 6 years or younger, perspectives related to later intervention phases were not represented, and we were not able to systematically compare views on early detection vs long-term intervention. Similarly, the numbers of participants within individual health care professional subgroups were not sufficient to support robust within-profession comparisons, and the analysis was therefore conducted at the combined health care provider level, with any role-specific nuances interpreted cautiously.

Additionally, while separate FGDs for health care professionals and parents minimized power imbalances, the group setting may have led to social desirability bias, with participants potentially moderating their comments in line with perceived group norms. The presence of AI-related stigma and limited public familiarity with AI screening could also have influenced the openness of responses, despite measures taken to build rapport. Future research could incorporate longitudinal designs and follow-up interviews to explore how attitudes evolve as AI screening tools are introduced and piloted in local health care systems.

These limitations do not undermine the validity of the findings but rather frame them as contextually situated insights, highlighting the value of qualitative inquiry for shaping ethical and culturally sensitive AI interventions in LMICs.

Conclusion

This study highlighted how feasibility, acceptability, and perceived ethical and practical enablers and barriers shape stakeholder attitudes toward implementing AI-powered tools for early ASD screening within the Egyptian health system. While parents were generally open to AI tools for early detection, health care professionals expressed caution, emphasizing the need for rigorous validation, culturally adapted algorithms, and integration within existing diagnostic pathways.

AI-powered tools hold significant promise for addressing diagnostic delays and specialist shortages in Egypt and other LMICs. However, their success depends on transparent governance, stakeholder engagement in co-design, and human oversight to safeguard against bias and misuse.

By integrating technological innovation with culturally sensitive implementation strategies, AI-powered screening tools could contribute to more equitable, timely, and accurate autism identification. In practical terms, this entails embedding AI within hybrid human–AI care models, aligning governance frameworks with community expectations and ethical standards, and sustaining engagement with families and clinicians to ensure that AI enhances, rather than replaces, human-centered care.

Acknowledgments

The authors express their sincere appreciation to all health care providers for their willingness to engage and cooperate during the study. Special thanks to all parents who gave their time to participate in this study, for whom it was conducted.

The authors disclose that ChatGPT (OpenAI; accessed January 2026) was used only for language polishing (improving grammar, clarity, and concision) during manuscript preparation. The authors reviewed and edited all artificial intelligence (AI)–assisted text and take full responsibility for the final content. No generative AI was used for data collection, data analysis, visualization, or the generation of study data, analyses, or results. No identifiable participant data or raw transcripts were entered into the tool.

Data Availability

The datasets used and/or analyzed for the current study are anonymous and are available from the corresponding author on request.

Funding

This study received financial support from Ulster University, Londonderry, Northern Ireland, United Kingdom through Ulster International Science Partnerships Fund (ISPF) funding (cost centre 71799) for the participant travel expense and goodwill payment and publication of this study.

Authors' Contributions

AMM was responsible for conceptualization and contributed to the formulation of the research questions, study themes, data curation, methodology, project administration, supervision, and review and editing of the manuscript. PY contributed to conceptualization, assisted in formulating the research questions and study themes, and participated in project administration and manuscript review and editing. PC contributed to conceptualization, project administration, and manuscript review and editing. GAA assisted with data curation, formal analysis, software, and validation. SMEK, the lead researcher for developmental psychology, maintained a reflexive journal throughout the familiarization process to document assumptions, emotional responses, and evolving interpretations. AE assisted with participant recruitment, facilitated focus groups, and contributed to drafting the

XSL•FO
RenderX

original manuscript. EAA contributed to the formulation of study themes and drafting of the original manuscript. NAE assisted with participant recruitment and facilitation of focus groups. ASE assisted with participant recruitment, facilitation of focus groups, reference writing, and drafting of the original manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Expanded methodological details.
[DOCX File , 25 KB - jmir_v28i1e82564_app1.docx]

Multimedia Appendix 2
Focus group and interview discussion guide.
[DOCX File , 18 KB - jmir_v28i1e82564_app2.docx]

Multimedia Appendix 3
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[PDF File (Adobe PDF File), 478 KB - jmir_v28i1e82564_app3.pdf]

References

1. Silverman C. NeuroTribes: The legacy of autism and the future of neurodiversity by steve silberman. Anthropological Quarterly 2015;88(4):1111-1121. [doi: 10.1353/anq.2015.0057]
2. Zwaigenbaum L, Bauman ML, Stone WL, Yirmiya N, Estes A, Hansen RL. Early identification of autism spectrum disorder: recommendations for practice and research. Pediatrics 2015;136(Suppl 1):S10-S40 [FREE Full text] [doi: 10.1542/peds.2014-3667C] [Medline: 26430168]
3. Dawson G, Rogers S, Munson J, Smith M, Winter J, Greenson J. Randomized, controlled trial of an intervention for toddlers with autism: the Early Start Denver Model. Pediatrics 2010;125(1):e17-e23 [FREE Full text] [doi: 10.1542/peds.2009-0958] [Medline: 19948568]
4. Lordan R, Storni C, De Benedictis CA. Autism spectrum disorders: diagnosis and treatment. In: Autism Spectrum Disorders. Brisbane: Exon Publications; 2021.
5. Brookman-Frazee L, Drahota A, Stadnick N, Palinkas LA. Therapist perspectives on community mental health services for children with autism spectrum disorders. Adm Policy Ment Health 2012;39(5):365-373 [FREE Full text] [doi: 10.1007/s10488-011-0355-y] [Medline: 21533846]
6. Metwally AM, Abdallah AM, El-Din EMS, Zeid DA, Khadr Z, Elshaarawy GA. Screening and determinant of suspected developmental delays among Egyptian preschool-aged children: a cross-sectional national community-based study. BMC Pediatr 2023;23(1):521 [FREE Full text] [doi: 10.1186/s12887-023-04335-0] [Medline: 37858055]
7. Metwally AM, Salah El-Din EM, Sami SM, Abdelraouf ER, Sallam SF, Elsaeid A. Mapping autism in Egypt: population-based insights into prevalence, risk determinants, and severity among children aged 1-12 years. Mol Autism 2025;16(1):32 [FREE Full text] [doi: 10.1186/s13229-025-00665-1] [Medline: 40442748]
8. Metwally AM, Nassar MS, Salah El-Din EM, Abdallah AM, Khadr Z, Abouelnaga MW. National screening for developmental delays and their determinants among Egyptian school age children: a step towards implementing life skills programs. PLoS One 2023;18(9):e0287315 [FREE Full text] [doi: 10.1371/journal.pone.0287315] [Medline: 37725608]
9. Topol E. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. London: Hachette UK; 2019.
10. Khudhur DD, Khudhur SD. The classification of autism spectrum disorder by machine learning methods on multiple datasets for four age groups. Measurement: Sensors 2023;27:100774. [doi: 10.1016/j.measen.2023.100774]
11. Alves CL, Martinelli T, Sallum LF, Rodrigues FA, Toutain T, Porto JAM. Multiclass classification of autism spectrum disorder, attention deficit hyperactivity disorder, and typically developed individuals using fMRI functional connectivity analysis. PLoS One 2024;19(10):e0305630 [FREE Full text] [doi: 10.1371/journal.pone.0305630] [Medline: 39418298]
12. Duda M, Ma R, Haber N, Wall DP. Use of machine learning for behavioral distinction of autism and ADHD. Transl Psychiatry 2016;6(2):e732. [doi: 10.1038/tp.2015.221] [Medline: 26859815]
13. Thabtah F, Peebles D. Early autism screening: a comprehensive review. Int J Environ Res Public Health 2019;16(18):3502 [FREE Full text] [doi: 10.3390/ijerph16183502] [Medline: 31546906]
14. Dhar S, Sarkar U. Safeguarding data privacy and informed consent. In: Advances in Electronic Government, Digital Divide, and Regional Development. Hershey: IGI Global; 2025.
15. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464):447-453 [FREE Full text] [doi: 10.1126/science.aax2342] [Medline: 31649194]

XSL•FO
RenderX

16. Labrique AB, Wadhwani C, Williams KA, Lamptey P, Hesp C, Luk R. Best practices in scaling digital health in low and middle income countries. Global Health 2018;14(1):103 [FREE Full text] [doi: 10.1186/s12992-018-0424-z] [Medline: 30390686]
17. Rogers SJ, Estes A, Lord C, Vismara L, Winter J, Fitzpatrick A. Effects of a brief Early Start Denver Model (ESDM)–based parent intervention on toddlers at risk for autism spectrum disorders: a randomized controlled trial. J Am Acad Child Adolesc Psychiatry 2012;51(10):1052-1065 [FREE Full text] [doi: 10.1016/j.jaac.2012.08.003] [Medline: 23021480]
18. Holden RJ, Karsh BT. The technology acceptance model: its past and its future in health care. J Biomed Inform 2010;43(1):159-172 [FREE Full text] [doi: 10.1016/j.jbi.2009.07.002] [Medline: 19615467]
19. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative aata collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015;42(5):533-544 [FREE Full text] [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
20. Krueger RA. Focus Groups: A Practical Guide for Applied Research. California: Sage Publications; 2014.
21. Patton MQ. Qualitative Research & Evaluation Methods: Integrating Theory and Practice. California: Sage Publications; 2014.
22. Guest G, Bunce A, Johnson L. How many interviews are enough? Field Methods 2006;18(1):59-82. [doi: 10.1177/1525822X05279903]
23. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B. Saturation in qualitative research: exploring its conceptualization and operationalization. Quality & Quantity 2018;52(4):1893-1907 [FREE Full text] [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
24. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019;11(4):589-597. [doi: 10.1080/2159676x.2019.1628806]
25. Shenton AK. Strategies for ensuring trustworthiness in qualitative research projects. Educ Inf 2004;22(2):63-75. [doi: 10.3233/EFI-2004-22201]
26. Levitt H, Bamberg M, Creswell J, Frost D, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018;73(1):26-46. [doi: 10.1037/amp0000151]
27. Morse JM. Critical analysis of strategies for determining rigor in qualitative inquiry. Qual Health Res 2015;25(9):1212-1222. [doi: 10.1177/1049732315588501] [Medline: 26184336]
28. Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: a mapping review. Soc Sci Med (1982) 2020;260:113172. [doi: 10.1016/j.socscimed.2020.113172] [Medline: 32702587]
29. Floridi L, Cowls J. A unified framework of five principles for AI in society. Harv Data Sci Rev 2019. [doi: 10.1162/99608f92.8cd550d1]
30. International Ethical Guidelines for Health-Related Research Involving Humans. Geneva: Council for International Organizations of Medical Sciences (CIOMS); 2016.
31. Metwally AM, Amer HA, Salama HI, Abd El Hady SI, Alam RR, Aboulghate A. Egyptian patients'/guardians' experiences and perception about clinical informed consent and its purpose: cross sectional study. PLoS One 2021;16(6):e0252996 [FREE Full text] [doi: 10.1371/journal.pone.0252996] [Medline: 34125842]
32. Egypt in Figures 2024: Population and Urbanization. Cairo: Central Agency for Public Mobilization and Statistics (CAPMAS); 2024.
33. Statistical Yearbook 2023: Governorates of Egypt. Cairo: Central Agency for Public Mobilization and Statistics (CAPMAS); 2023.
34. Urban population (% of total) – Arab Republic of Egypt. Washington DC: World Bank; 2023.
35. Egypt Urban Expansion and Sustainable Development Report. Nairobi: United Nations Human Settlements Programme (UN-Habitat); 2021.
36. Egypt Human Development Report 2021: Development, a Right for All? Egypt’s Pathways and Prospects. Cairo: United Nations Development Programme (UNDP); 2021.
37. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
38. Wallis KE, Guthrie W. Identifying autism spectrum disorder in real-world health care settings. Pediatrics 2020;146(2):e20201467 [FREE Full text] [doi: 10.1542/peds.2020-1467] [Medline: 32632020]
39. Ayub R, Immamul Ansarullah S, Samdani F, Akhtar A, Afzal M, Moinuddin Ahmed S. Enhancing early detection of autistic spectrum disorder in children using machine learning approaches. J King Saud Univ Sci 2024;36(10):103468 [FREE Full text] [doi: 10.1016/j.jksus.2024.103468]
40. Morley J, Elhalal A, Garcia F, Kinsey L, Mökander J, Floridi L. Ethics as a service: a pragmatic operationalisation of AI Ethics. Minds Mach (Dordr) 2021;31(2):239-256 [FREE Full text] [doi: 10.1007/s11023-021-09563-w] [Medline: 34720418]
41. Fombonne E. The changing epidemiology of autism. J Appl Res Intellect Disabil 2005;18(4):281-294. [doi: 10.1111/j.1552-6909.2012.01399.x] [Medline: 2861175]

XSL•FO
RenderX

42. Lyall K, Croen L, Daniels J, Fallin MD, Ladd-Acosta C, Lee BK. The changing epidemiology of autism spectrum disorders. Annu Rev Public Health 2017;38(1):81-102 [FREE Full text] [doi: 10.1146/annurev-publhealth-031816-044318] [Medline: 28068486]
43. Hermaszewska S, Sin J. End-user perspectives on the development of an online intervention for parents of children on the autism spectrum. Autism 2021;25(5):1234-1245 [FREE Full text] [doi: 10.1177/1362361320984895] [Medline: 33423522]
44. Metwally AM, Helmy MA, Salah El-Din EM, Saleh RM, Abdel Raouf ER, Abdallah AM. National screening for Egyptian children aged 1 year up to 12 years at high risk of autism and its determinants: a step for determining what ASD surveillance needs. BMC Psychiatry 2023;23(1):471 [FREE Full text] [doi: 10.1186/s12888-023-04977-5] [Medline: 37381024]
45. Lincoln YS. Naturalistic Inquiry. California: SAGE Publications, Inc; 1985.
46. Braun V, Clarke V. One size fits all? What counts as quality practice in (reflexive) thematic analysis? Qual Res Psychol 2020;18(3):328-352. [doi: 10.1080/14780887.2020.1769238]
47. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Iii HD, et al. Datasheets for datasets. Commun ACM 2021;64(12):86-92 [FREE Full text] [doi: 10.1145/3458723]
48. Birhane A. Algorithmic injustice: a relational ethics approach. Patterns 2021;2(2):100205 [FREE Full text] [doi: 10.1016/j.patter.2021.100205] [Medline: 33659914]
49. Wies B, Landers C, Ienca M. Digital mental health for young people: a scoping review of ethical promises and challenges. Front Digit Health 2021;3:697072 [FREE Full text] [doi: 10.3389/fdgth.2021.697072] [Medline: 34713173]
50. Li Y, Zhang T, Yang Y, Gao Y. Artificial intelligence-aided decision support in paediatrics clinical diagnosis: development and future prospects. J Int Med Res 2020;48(9):300060520945141 [FREE Full text] [doi: 10.1177/0300060520945141] [Medline: 32924683]
51. Floridi L, Cowls J. A unified framework of five principles for AI in society. In: Machine Learning and the City: Applications in Architecture and Urban Design. Hoboken: Wiley; 2022:535-545.
52. Lincoln YS. Naturalistic Inquiry. California: SAGE Publications, Inc; 1985.
53. Achenie LEK, Scarpa A, Factor RS, Wang T, Robins DL, McCrickard DS. A machine learning strategy for autism screening in toddlers. J Dev Behav Pediatr 2019;40(5):369-376 [FREE Full text] [doi: 10.1097/DBP.0000000000000668] [Medline: 30985384]
54. Li Y, Zhang T, Yang Y, Gao Y. Artificial intelligence-aided decision support in paediatrics clinical diagnosis: development and future prospects. J Int Med Res 2020;48(9):300060520945141 [FREE Full text] [doi: 10.1177/0300060520945141] [Medline: 32924683]
55. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: four longitudinal field studies. Manage Sci 2000;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
56. Rosenstock IM. The health belief model and preventive health behavior. Health Educ Monogr 1974;2(4):354-386 [FREE Full text] [doi: 10.1177/109019817400200405]
57. Goffman E. Stigma: Notes on the Management of Spoiled Identity. New York: Simon and Schuster; 2009.
58. McKnight DH, Choudhury V, Kacmar C. Developing and validating trust measures for e-Commerce: an integrative typology. Inf Syst Res 2002;13(3):334-359. [doi: 10.1287/isre.13.3.334.81]
59. Metwally AM, Abdel-Latif GA, Eletreby L, Aboulghate A, Mohsen A, Amer HA. Egyptians' social acceptance and consenting options for posthumous organ donation: a cross-sectional study. BMC Med Ethics 2020;21(1):49 [FREE Full text] [doi: 10.1186/s12910-020-00490-6] [Medline: 32539704]

Abbreviations

AI: artificial intelligence
ASD: autism spectrum disorder
COREQ: Consolidated Criteria for Reporting Qualitative Research
FGD: focus group discussion
HBM: Health Belief Model
LMIC: low- and middle-income country
NRC: National Research Centre
RQ1: research question 1
RQ2: research question 2
TAM: Technology Acceptance Model

XSL•FO
RenderX

Edited by A Stone; submitted 18.Aug.2025; peer-reviewed by F Jorge, J Wang; comments to author 02.Oct.2025; revised version received 06.Jan.2026; accepted 30.Jan.2026; published 07.Apr.2026.

Please cite as:
Yogarajah P, Metwally AM, Chaurasia P, Elshaarawy GA, El Khateeb SM, Ashaat EA, Elsaeid A, Elghareeb NA, ElRifay AS
Feasibility and Acceptability of AI-Powered Tools for Early Autism Screening in Egypt: Semistructured Focus Group Study
J Med Internet Res 2026;28:e82564
URL: https://www.jmir.org/2026/1/e82564
doi:10.2196/82564
PMID:

©Pratheepan Yogarajah, Ammal M Metwally, Priyanka Chaurasia, Ghada A Elshaarawy, Shereen M El Khateeb, Engy A Ashaat, Amal Elsaeid, Nahed A Elghareeb, Amira S ElRifay. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 07.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

What Patients With Asthma Share When No One Listens: Multimethod Observational Study of Patient Narratives on Reddit

Elena Curto-Sánchez[1], MD, PhD; Gabriela Salazar-Palacios[1], MD; Ana Martín-Varillas[1], MD; Estela Cristina Prieto-Maíllo[1], MD; Jacinto Ramos-González[1], MD, PhD; Ignacio Dávila-González[2], MD, PhD; Domingo Palacios-Ceña[3], PhD; Juan Nicolas Cuenca-Zaldivar[4], PhD

[1]Pneumology Service, Complejo Hospitalario de Salamanca, Salamanca, Castilla y León, Spain

[2]Department of Allergy, Institute for Biomedical Research of Salamanca, Complejo Hospitalario de Salamanca, Salamanca, Castilla y León, Spain

[3]Research Group of Humanities and Qualitative Research in Health Science of Universidad Rey Juan Carlos (Hum&QRinHS), Department of Physical Therapy, Occupational Therapy, Physical Medicine and Rehabilitation, Universidad Rey Juan Carlos, Alcorcon, Madrid, Spain

[4]Primary Health Center "El Abajón", Research Group of Pain and Physiotherapy, Madrid Health Service, Las Rozas de Madrid, Madrid, Spain

Corresponding Author:
Domingo Palacios-Ceña, PhD
Research Group of Humanities and Qualitative Research in Health Science of Universidad Rey Juan Carlos (Hum&QRinHS)
Department of Physical Therapy, Occupational Therapy, Physical Medicine and Rehabilitation
Universidad Rey Juan Carlos
Avenida Atenas s/n
Alcorcon, Madrid, 28922
Spain
Phone: 34 914888883 ext 8883
Fax: 34 914888959
Email: domingo.palacios@urjc.es

Abstract

Background: The use of social media platforms, such as Reddit, to seek and share information about disease management and treatment strategies is increasingly common. In the context of asthma—a chronic condition characterized by limiting symptoms and exacerbations that require active patient engagement and adherence to treatment—there is a lack of research describing the content of Reddit posts and the specific topics of interest to patients.

Objective: This study aimed to describe the topics discussed by users on the Reddit asthma forum and identify the sentiments and polarity of the language used in the posts.

Methods: A retrospective observational study of public posts on the asthma subreddit forum (r/Asthma) over a 1-year period (October 2023-October 2024). All posts and related threads were included, subdivided into hot, news, and top, and those voted "up" or "down," those that received "awards," categorized as "golds." The messages were reviewed manually and excluded if they were not related to asthma. A mixed methods analysis was conducted, comprising (1) analysis using text lemmatization, (2) structural topic modeling to identify topics based on word frequency, and (3) sentiment and polarity analysis. This approach aimed to identify the most frequently used topics on Reddit, detect positive and negative sentiments based on the words used, and acceptance or rejection (polarity) based on the language used in the asthma subreddit. Statistical analyses were performed using R software (version 4.1.3; R Foundation for Statistical Computing), with a significance threshold set at $P<.05$.

Results: After removing duplicates, 7806 posts were identified. The suitability of the chosen analysis model was confirmed, as it presented the best balance between exclusivity and semantic coherence. Clusters of 25 topics were identified and distributed according to their weight. The topics with the highest weight were Topic 7 (Symptoms and severity of asthma attacks) and Topic 18 (Causes of asthma). No significant differences were found in the evolution of emerging topics throughout the year except in Topic 20 (Seeking advice from people with asthma; $P=.04$), Topic 21 (Medical tests that should be reviewed periodically; $P=.04$), and Topic 22 (Times of year when attacks occur; $P=.03$). The proportion of feelings and emotions showed a stable trend throughout the year. Discrepancies in feelings and emotions were identified depending on the dictionaries used. Thus, a higher probability of positive feelings was confirmed in the AFINN lexicon. Meanwhile, negative feelings were significant in the Stanford Natural Language Processing, Bing, and National Research Council Canada lexicons.

XSL•FO
RenderX

Conclusions: These results can serve as a guide to identify hidden patient needs and help professionals develop specific interventions on topics relevant to patients.

(J Med Internet Res 2026;28:e77027) doi:10.2196/77027

KEYWORDS

Reddit; subreddits; r/asthma; asthma; social media; health information; online health; online health information; patient education; observational study

Introduction

Asthma is one of the most prevalent chronic respiratory diseases, affecting over 300 million people globally [1]. The intermittent onset of bronchospasm attacks causes symptoms such as wheezing and dyspnea, and is characterized by airway inflammation, hyperresponsiveness, and mucus hypersecretion, leading to airflow obstruction [1]. The clinical manifestation of asthma varies across individuals, with distinct phenotypes [2], influencing treatment of the disease [3].

The introduction of novel clinical biomarkers (such as sputum and blood eosinophils, serum immunoglobulin E (IgE), and fractional exhaled nitric oxide) [2] has facilitated the development of treatments tailored to each patient's individual phenotype [4,5], specifically monoclonal antibodies. However, these drugs are limited to the most severe forms and certain phenotypes [3,4]. Treatment for most patients with mild to moderate forms of the disease is based on the use of inhalers; however, despite the ongoing development of educational programs, significant issues remain in relation to adherence [6] and inhalation technique [7], both of which have critical implications for disease management and the frequency of exacerbations.

Another key aspect in the treatment of asthma is the shared perspective between prescribing physicians and patients diagnosed with asthma on the impact of the disease on daily life [8,9]. Previous studies [8,9] show that there are differences in perspective on the management and control of the disease, resulting in an underestimation of the impact of asthma on the patient's life by the physician, reducing disease and symptom control [8].

One of the responses among patients is to seek help to cope with symptoms, apply strategies, and share their experiences with other patients [10,11]. In this regard, the use of social media for health reasons has increased over the last decade [10,11], having a major impact on patients' understanding of the disease, the acquisition of habits, the decision-making process, and health outcomes [12-18]. In addition, widely shared or viral content significantly influences the adoption—or rejection—of health behaviors [10], largely due to its ability to elicit strong, activating emotions, whether positive or negative [19].

Reddit is a large-scale online platform that hosts more than 130,000 individual communities, known as subreddits. It attracts around 500 million visits per month and engages approximately 73 million unique users each day [20]. With over 270 million active users each month [21], Reddit allows individuals to share personal experiences related to illness and health care. Users frequently post firsthand accounts, respond to others with shared experiences, and exchange advice on symptoms, medical conditions, and treatments [10,22]. The platform allows registered users (known as Redditors) to post content (text, images, and videos) to moderated boards, which are voted on by other users. Posts on Reddit are visible to the broader community, and their prominence within a subreddit is shaped by user voting—either upvotes or downvotes [23]. Subreddits are organized by topic, and users, known as Redditors, can subscribe to those that match their interests, allowing them to curate the content displayed on their personalized front page [24]. Additionally, some subreddits follow a question-and-answer format, with popular examples including r/AskReddit and r/AMA, which features the interactive "Ask Me Anything" style. Discussions on Reddit are public (unless specifically designated as private), allowing for passive data collection. Furthermore, for an internet user seeking information on a specific topic, Reddit is a useful starting point due to its topic-centric organization [10].

To the authors' knowledge, there are no studies that have shown the use of Reddit by patients with asthma and described the topics they consult. The questions that guided this study were: "What were the most consulted and discussed topics by patients with asthma on Reddit?" and "Was the language used on Reddit positive and negative? The objectives of this study were (1) to describe the topics discussed by patients in the asthma subreddit (r/Asthma) and (2) to identify the sentiments and polarity of the language used by patients with asthma on Reddit.

This study may help identify unknown and unmet needs of patients with asthma and improve medical care.

Methods

Design

An observational study was conducted, consisting of a retrospective analysis of public posts on Reddit over a 1-year period, from October 2023 to October 2024. Also, we used the subreddit forum asthma (r/Asthma) [25]. This study adhered to the Reporting Guidelines for Social Networks in Health Research [26], which are incorporated within the EQUATOR (Enhancing the Quality and Transparency of Health Research) initiative [27].

Data Collection

Posts that included general comments were identified and subdivided into hot, news, and top. Similarly, posts that were voted "up" or "down" were identified, as well as those that received "awards," categorized as "golds," and cross-posted posts. r/Asthma posts were obtained with the R library *RedditExtractoR* [28]. This library allows for extracting up to

XSL•FO
RenderX

1000 main posts by each category (hot, top, and new posts) and all generated threads. Only posts in the English language were used to facilitate the analysis.

All posts from the r/Asthma forum and all threads derived from the conversations were included, as r/Asthma is a subreddit forum dedicated specifically to asthma, so the objective of the study is to determine what topics interest the users of this forum without establishing any specific selection criteria a priori.

In addition, the main topic, secondary topics, number of votes obtained, number of responses, and subresponses obtained were included. Also, all information related to asthma, such as symptom management, treatment, follow-up consultations, risk behaviors, healthy habits, professional care, adherence (or nonadherence) to treatment and medical recommendations, education, use of inhalers, etc, was incorporated into the study and analyzed. Messages were reviewed manually and systematically excluded if they were not related to asthma use.

Statistical Analysis

Overview

Statistical analyses were performed using R software (version 4.1.3; R Foundation for Statistical Computing) [29]. The significance level was set at $P<.05$. The Kolmogorov-Smirnov test with Lilliefors correction was used to test the distribution of the variables. The variables were described with mean (SD), median (IQR), or with absolute and relative values, that is, n (%).

In this study, a mixed methods analysis was carried out that included three types of analysis: (1) text lemmatization analysis, (2) structural analysis using models based on topics identified by their frequency of appearance, and (3) sentiment and polarity analysis. This allowed us to identify the most frequently used topics on Reddit, identify positive and negative sentiments based on the words used, and determine acceptance or rejection (polarity) based on the language used to discuss asthma in Reddit posts.

Analysis Using Text Lemmatization

First, the text of the posts was lemmatized for analysis. Lemmatization is a linguistic process that involves identifying the base or lexicon form of a word—known as the lemma—from its inflected variants (eg, plural forms, gendered adjectives, verb conjugations). The lemma is the canonical form under which all inflected versions of a word are grouped. For example, in traditional dictionaries, nouns are typically listed in their singular form, adjectives in the masculine singular, and verbs in the infinitive.

Structural Topic Modeling Analysis

Second, a structural topic model analysis was conducted to examine the emergence of topics across user comments. This method also allows the incorporation of metadata, such as the time period of each comment, as a covariate in the model. The selection of the optimal number of topics was carried out in 2 phases: first, the range of topics with the best values of held-out likelihood and lower limit of stability of convergence between iterations of the models and with the lowest value in the residuals was evaluated; in a second phase, the ratio between semantic coherence and exclusivity between the selected models was analyzed. Exclusivity evaluates whether the main words of the topics also appear as main words of other topics, while semantic coherence shows whether the words most associated with a topic occur equally or not within the documents; in both cases, higher values are better [30-32].

Sentiment and Polarity Analysis

Finally, a sentiment analysis was performed using the Bing [33], AFINN [34], National Research Council Canada (NRC) [35], and Stanford Natural Language Processing (Stanford NLP) [36] dictionaries. All 4 dictionaries are based on unigrams or individual words that assign scores for positive or negative feelings. In addition, the NRC lexicon classifies words into 8 emotional categories of anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. While the Affin lexicon assigns words a score ranging from –5 to 5, with negative values indicating more negative feelings and positive values indicating more positive feelings. Stanford NLP classifies positive into 5 categories: very negative, negative, neutral, positive, and very positive. Similarly, sentence polarity was analyzed using the Bing lexicon, incorporating modifiers such as amplifiers, decrementers, and negators as proposed by Halliday [37]. The integration of these 4 dictionaries facilitates a detailed and comprehensive analysis of the emotions and feelings conveyed in the texts shared by Reddit (r/Asthma forum) users. This methodological approach enables an emotional profiling of the analyzed content, contributing to a deeper understanding of the affective dimensions embedded in user-generated discourse. Furthermore, this combination of dictionaries allows for a multiangle analysis, providing a comprehensive analytical framework that encompasses the analysis of positive and negative emotions, ranging from –5 to 5 with the AFINN dictionary or from very negative to very positive with the Stanford NLP dictionary, as well as assessing 8 more complex emotions with the NRC dictionary. Moreover, the simultaneous use of these 4 dictionaries allows for a cross-validation and complementarity strategy; the combination of results can lead to a more robust and reliable analysis [38,39].

The presence of significant differences in the proportion of the categories of feelings and emotions or the average score of each lexicon and its temporal evolution was analyzed using a regression model. The assumption of linearity between the dependent variables and the quantitative predictors was tested by evaluating the effective degrees of freedom (EDF) with a value greater than 1 as a nonlinearity cut-off point. A generalized additive mixed model (GAMM) or a linear mixed model (LMM) was applied depending on the EDF value. Since these were time series, the correlation structure was modeled by comparing, using a likelihood ratio test (LRT), the models without this structure, with an autoregressive structure with lag 1 (AR1), and with an autoregressive moving average structure (ARMA) evaluated by means of an ARMA analysis of the residuals of the previous AR1 model. In the GAMM models, the fulfillment of the assumptions of v for the smoothed terms was tested, eliminating those with a value greater than 0.8, and the adequacy of the number of basic functions was checked using the K index ($P>.05$ as a selection criterion). The fulfillment of the

XSL•FO
RenderX

assumptions in the residuals was checked with the Kolmogorov-Smirnov tests with Lilliefors correction (normality) and Breusch-Pagan (heteroscedasticity). Since these were not met in the GAMM models, the robust sandwich-type standard errors were calculated in the LMM model with the Stanford NLP lexicon using a bootstrap.

Ethical Considerations

This study does not require approval from the Ethics and Clinical Research Committee of the University Hospital of Salamanca because the information is publicly available. The same considerations have been applied in previous studies that analyzed the use of Reddit by patients to share and search for information on topics related to health and disease [40-42]. In addition, institutional review board and research ethics committee approvals were not required nor sought because the research did not meet the requirements for needing ethical approval per section 1.3 of the European Pharmaceutical Market Research Association guidelines [43]. The ethical principles established by the Declaration of Helsinki were respected. In addition, Reddit offers a comprehensive application programming interface that grants access to much of the platform's information. In strict compliance with the rules and ethical guidelines of each subreddit, we only collect data from subreddits that explicitly allow research activities. Furthermore, to protect user privacy, we only record the comment ID without including any user information or the content of the comment itself, ensuring that no personally identifiable information or data that could lead to reidentification is collected. Furthermore, in the posts selected as illustrative examples for this study, all identifying information, including authorship and publication dates, was removed, and the content was summarized to preserve and emphasize the essential information.

Results

Descriptive

A total of 7806 posts were collected from the r/Asthma thread between October 2023 and October 2024 after removing duplicates, of which 454 generated a total of 6046 response threads. Since Reddit does not allow the extraction of more than 1000 posts, different filters were used to maximize the search for primary comments, from which an unlimited number of response threads could be obtained (Table 1).

XSL•FO
RenderX

Table 1. Reddit r/Asthma posts' characteristics between October 2023 and October 2024.

Post characteristics	2023			2024									
	October	November	December	January	February	March	April	May	June	July	August	September	October
Characteristics of posts													
n	4	69	62	74	69	89	81	76	86	79	85	451	535
Generated comments, mean (SD)	37.75(10.24)	23.03 (19.39)	20.85 (15.28)	23.16 (24.78)	32.83 (31.77)	30.66 (23.66)	26.83 (27.63)	24.22 (19.82)	25.69 (23.05)	20.71 (20.76)	19.44 (26.46)	10.39 (14.26)	10.66 (13.71)
Comments filter, n (%)													
Hot[a]	—[b]	—	—	—	—	—	—	—	—	—	10 (11.8)	368 (81.6)	382 (71.4)
New[c]	—	—	—	—	—	—	—	—	—	—	2 (2.4)	—	—
Top[d]	4 (100)	69 (100)	62 (100)	74 (100)	69 (100)	89 (100)	81 (100)	76 (100)	86 (100)	79 (100)	73 (85.9)	83 (18.4)	153 (28.6)
Characteristics of posts with threads													
n	1	13	12	26	22	28	28	27	18	17	28	132	102
Score[e], mean (SD)	17.00	15.31 (5.89)	9.75 (2.05)	9.38 (3.73)	9.68 (4.75)	10.54 (6.13)	9.93 (5.28)	12.96 (7.60)	11.00 (4.00)	10.47 (3.74)	7.57 (8.16)	2.42 (3.57)	4.06 (5.68)
Upvotes[f], mean (SD)	17.00	15.31 (5.89)	9.75 (2.05)	9.38 (3.73)	9.68 (4.75)	10.54 (6.13)	9.93 (5.28)	12.96 (7.60)	11.00 (4.00)	10.47 (3.74)	7.57 (8.16)	2.42 (3.57)	4.06 (5.68)
Up ratio[g], mean (SD)	0.95	0.92 (0.07)	0.92 (0.07)	0.92 (0.07)	0.91 (0.09)	0.91 (0.06)	0.93 (0.08)	0.92 (0.07)	0.94 (0.07)	0.85 (0.11)	0.87 (0.15)	0.77 (0.24)	0.79 (0.24)
Cross-posts[h], mean (SD)	—	—	—	—	—	—	—	—	—	—	—	0.01 (0.09)	—
Generated comments, mean (SD)	34.00	26.92 (18.83)	19.92 (14.15)	16.38 (12.54)	18.14 (15.79)	19.25 (14.04)	16.07 (11.69)	19.63 (15.80)	16.72 (16.05)	24.88 (24.83)	12.57 (9.80)	8.19 (8.56)	9.29 (14.69)
Thread characteristics													
n	26	339	218	415	372	536	450	503	320	427	359	1099	982
Score, mean (SD)	1.77 (1.24)	2.99 (3.81)	2.56 (3.01)	2.87 (3.60)	2.92 (5.49)	2.25 (3.03)	2.85 (4.23)	2.41 (3.20)	2.83 (4.21)	3.02 (6.12)	2.72 (3.36)	2.19 (3.43)	2.49 (2.94)
Upvotes, mean (SD)	1.77 (1.24)	2.99 (3.81)	2.56 (3.01)	2.87 (3.60)	2.92 (5.49)	2.25 (3.03)	2.85 (4.23)	2.41 (3.20)	2.83 (4.21	3.02 (6.12)	2.72 (3.36)	2.19 (3.43)	2.49 (2.94)

[a]Hot: Comments with the highest popularity rate based on their age and number of positive votes are the most active and popular discussions in real time.

[b]Not available.

[c]New: New comments generated in a given period of time.

[d]Top: Comments with the highest number of net upvotes (upvotes minus downvotes) in a given time period.

[e]Score: Net positive votes (positive votes minus negative votes).

[f]Upvotes: Positive votes.

[g]Up ratio: Quotient of positive and negative votes in percentage.

[h]Cross-posts: Comments linked from other subreddits. Generated comments: comments not written by humans.

Analysis of Thematic Models

By determining the best held-out likelihood values among the models, it was determined that the optimal number of topics was 15 and 25 (Multimedia Appendix 1). Subsequently, the model with 25 topics was found to have the best balance of exclusivity versus semantic coherence, and therefore, the model with 25 topics was adopted (Multimedia Appendix 2). Table 2 shows the 25 topics hypothesized after the development of the model.

XSL•FO
RenderX

Table 2. Reddit r/Asthma posts topics identified.

Topic	Description
Topic 1	Influence of respiratory problems on the ability to walk.
Topic 2	Timing of attacks.
Topic 3	Hopes for rapid improvement.
Topic 4	Adverse effects of Prednisone that may justify stopping its use.
Topic 5	Gratitude for improvement.
Topic 6	Use of inhaled steroids as rescue medication.
Topic 7	Symptoms and severity of asthma attacks.
Topic 8	Medical tests to be monitored.
Topic 9	Influence of respiratory problems on work activity.
Topic 10	Influence of respiratory problems on quality of life in general terms.
Topic 11	Influence of tobacco.
Topic 12	Need for information.
Topic 13	Budesonide spray dosage.
Topic 14	How to access Albuterol inhalers for wheezing without medical insurance.
Topic 15	Hospitalization with oxygen therapy in case of acute exacerbation.
Topic 16	Search for any asthma-related therapy.
Topic 17	Influence of respiratory symptoms such as coughing.
Topic 18	Causes of asthma.
Topic 19	Symptom change throughout the year.
Topic 20	Request for advice to people who have asthma.
Topic 21	Medical tests that should be checked periodically.
Topic 22	Times of the year when attacks occur.
Topic 23	Change in symptoms at times of the year when temperatures are cooler.
Topic 24	Need for specialized medical care.
Topic 25	Gratitude for the assistance received.

Subsequently, each topic was classified according to the weight of its frequency, together with the most frequent words used in each topic (Figure 1 shows the weighted topics).

XSL•FO
RenderX

Figure 1. Weighted topics.

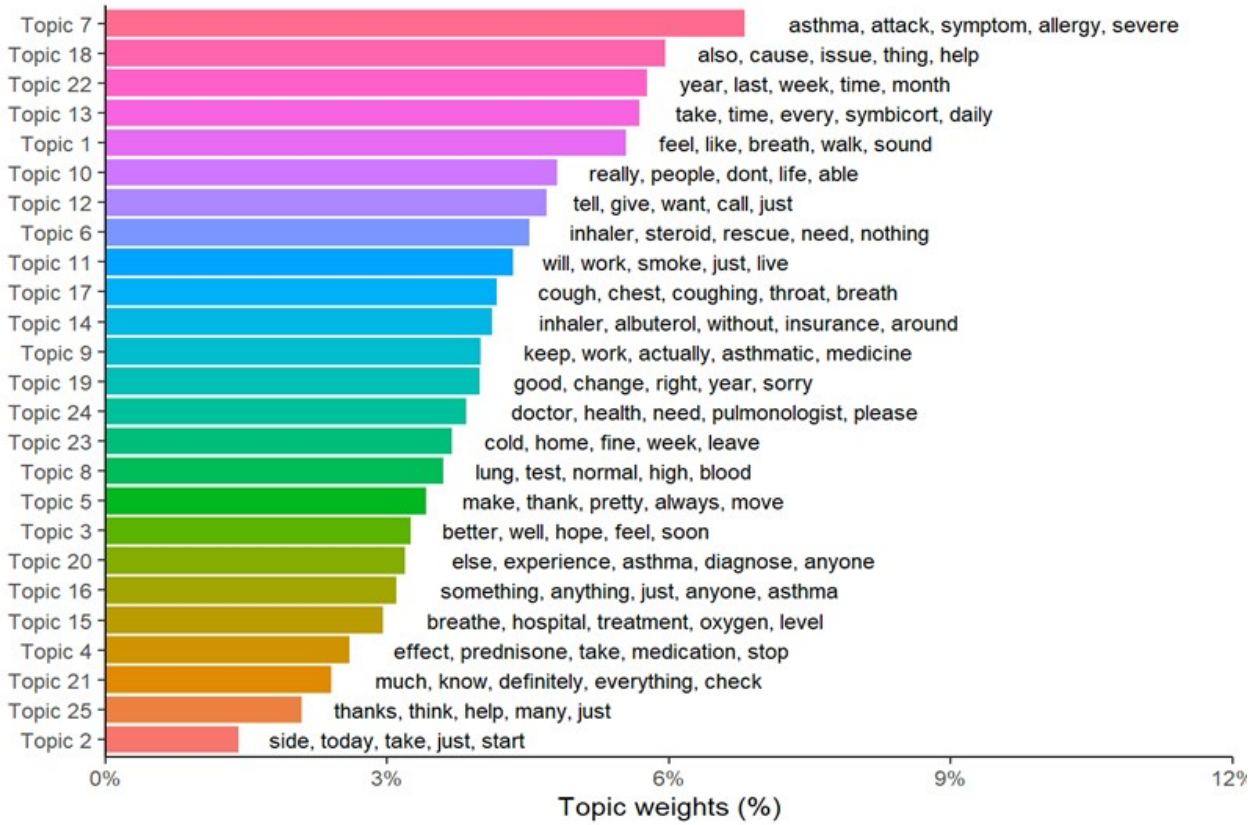

The topics with the greatest weight were Topic 7 (Symptoms and severity of asthma attacks) and Topic 18 (Causes of asthma).

Regarding the annual analysis of the topics, there were no significant differences in the evolution of the emerging topics throughout the year, except in 3 topics. These were Topic 20 (Request for advice from people with asthma; *P*=.04), Topic 21 (Medical tests that should be checked periodically; *P*=.04), and Topic 22 (Times of the year when attacks occur; *P*=.03; Table 3).

XSL•FO
RenderX

Table 3. Evolution of each topic throughout the period of time analyzed.

Topic and time period	Coefficient (SE)	*t* statistic	*P* value[a]
Topic 1			
Intercept	–0.04 (0.29)	–0.13	.89
Date	0 (0)	0.29	.76
Topic 2			
Intercept	–0.04 (0.21)	–0.18	.85
Date	0 (0)	0.30	.76
Topic 3			
Intercept	0.415 (0.21)	1.92	.05
Date	0 (0)	–1.77	.07
Topic 4			
Intercept	0.052 (0.24)	0.21	.82
Date	0 (0)	–0.07	.94
Topic 5			
Intercept	0.225 (0.21)	1.04	.29
Date	0 (0)	–0.89	.37
Topic 6			
Intercept	0.063 (0.26)	0.23	.81
Date	0 (0)	–0.07	.94
Topic 7			
Intercept	0.128 (0.33)	0.38	.70
Date	0 (0)	–0.22	.82
Topic 8			
Intercept	–0.234 (0.27)	–0.85	.39
Date	0 (0)	0.99	.32
Topic 9			
Intercept	0.308 (0.29)	1.05	.29
Date	0 (0)	–0.90	.36
Topic 10			
Intercept	0.142 (0.268)	0.53	.59
Date	0 (0)	–0.35	.72
Topic 11			
Intercept	0.089 (0.24)	0.37	.71
Date	0 (0)	–0.20	.83
Topic 12			
Intercept	0.053 (0.25)	0.20	.83
Date	0 (0)	–0.03	.97
Topic 13			
Intercept	–0.456 (0.30)	–1.50	.13
Date	0 (0)	1.66	.09
Topic 14			
Intercept	–0.452 (0.28)	–1.59	.11
Date	0 (0)	1.74	.08

XSL•FO
RenderX

Topic and time period	Coefficient (SE)	*t* statistic	*P* value[a]
Topic 15			
Intercept	–0.074 (0.31)	–0.23	.81
Date	0 (0)	0.36	.71
Topic 16			
Intercept	–0.432 (0.27)	–1.57	.11
Date	0 (0)	1.71	.08
Topic 17			
Intercept	–0.048 (0.23)	–0.20	.83
Date	0 (0)	0.37	.70
Topic 18			
Intercept	0.223 (0.31)	0.70	.48
Date	0 (0)	–0.53	.59
Topic 19			
Intercept	0.41 (0.24)	1.65	.09
Date	0 (0)	–1.48	.13
Topic 20			
Intercept	–0.505 (0.26)	–1.89	.05
Date	0 (0)	2.04	.041
Topic 21			
Intercept	0.421 (0.19)	2.11	.035
Date	0 (0)	–1.98	.048[b]
Topic 22			
Intercept	0.706 (0.30)	2.31	.021
Date	0 (0)	–2.13	.033
Topic 23			
Intercept	0.167 (0.31)	0.53	.59
Date	0 (0)	–0.39	.69
Topic 24			
Intercept	–0.19 (0.272)	–0.69	.48
Date	0 (0)	0.84	.39
Topic 25			
Intercept	0.067 (0.04)	1.52	.12
Date	0 (0)	–1.26	.20

[a]Significant $P<.05$.

Multimedia Appendix 3 presents examples of texts and narratives shared by users in the Reddit r/Asthma forum, organized by each identified topic.

In addition, a clustering and correlation of all topics was observed, except topics 3, 6, and 13 (Hope of getting better soon; Use of inhaled steroids as rescue medication; and Dosage of Budesonide aerosol; respectively) referring to the use of steroids and Budesonide to accelerate the improvement of symptoms; and topics 17, 20, and 23 (Influence of respiratory symptoms such as cough; Asking for advice from people who have asthma; and Change of symptoms at times of the year when temperatures are lower; respectively) related to asking for advice from those who have already experienced asthma on the management of respiratory symptoms such as cough, especially at times of lower temperatures (Multimedia Appendix 4).

The graphs with the model predictions show how, over the year, the probability of posts containing topics 20 (Request for advice to people with asthma*)* and 22 (Times of the year when attacks occur) appear in autumn 2024, decreasing over 2024. In the

XSL•FO
RenderX

case of Topic 22, they increase again in the autumn of 2024. Conversely, Topic 21 (Medical tests to be checked periodically), increases especially in the summer-autumn of 2024 (Multimedia Appendix 5).

Analysis of Feelings and Polarity

The proportion of feelings and emotions and scores showed a stable trend throughout the year (Multimedia Appendix 6). In addition, a nonnormal distribution in polarity was confirmed ($P<.001$).

In relation to the contrast of the hypotheses, the significant LRT indicated that the ARMA model is the one that presented the best fit, except in the NRC lexicon of emotions, which presents better results in the AR1 model. Meanwhile, in polarity, the best-fitting model was the model without autocorrelation structure (Multimedia Appendix 7).

Significant differences were confirmed in the proportion of categories appearing in each of the dictionaries used versus the reference category. In contrast, time has no significant effect on the proportion of feelings and emotions or polarity (Multimedia Appendix 8).

Sentiment and emotion analysis produced varying results depending on the lexicons applied. For instance, categorical analysis showed a higher likelihood of positive sentiments and score values of 2 using the AFINN lexicon. In contrast, the Stanford NLP, Bing, and NRC lexicons highlighted significant negative sentiment. Additionally, the NRC lexicon revealed notable associations with the emotions of anticipation, fear, and trust (Multimedia Appendix 9).

The following lines present excerpts from Reddit posts in which both negative and positive emotional expressions are identified. Each example is accompanied by the date of publication. One post describes a near-death experience due to worsening asthma symptoms:

> *I expressed my feelings towards the limitation that asthma brings to me…well, 2 days ago, I almost died…I had an asthma attack after minimal exercise.*

Another post reflects intense frustration caused by the progression of the disease:

> *My asthma has been flaring up more frequently than usual. It's stopped me from seeing friends, working, and made me lose sleep. This is causing me a significant amount of stress and anxiety, too, which I know also makes it worse.*

A third example illustrates how an asthma crisis can complicate air travel:

> *Had an asthma attack at the beginning of a long international flight…it was really scary being trapped in that confined space, not being able to breathe.*

On the other hand, positive experiences were also identified:

> *No doctor wanted to diagnose me as asthmatic, but the only thing that controlled my coughing attacks was corticosteroids…I am grateful to the doctor who agreed to prescribe me.*

Discussion

Principal Findings

The analysis of Reddit's r/Asthma forum has made it possible to identify and classify user-generated topics based on their weight and relevance. The most prominent topics include the causes of asthma and the symptoms and severity of attacks, followed by seasonal patterns of exacerbations and corticosteroid inhaler dosing. Posts display a wide range and intensity of emotions, with a noticeable tendency toward the use of negative language when discussing asthma.

Comparison With Previous Work

The analysis of social networks (including Reddit) and their ability to influence risk behaviors, and the construction of people's opinions, grew during the COVID-19 pandemic [44-46] because of their use as tools for obtaining and cross-checking information, sharing knowledge and as a source of help in dealing with diseases and applying strategies or treatments [47-50]. However, the use of online health information carries significant risks related to the acquisition of reliable knowledge [51-55]. To obtain trustworthy, evidence-based medical or scientific information, users must navigate through a vast amount of misinformation and poor-quality content, including posts that promote unhealthy practices or unproven treatments [52,54-56]. As a result, individuals may make decisions that could potentially harm their health [52-54]. A key factor in acquiring reliable and safe online health information is enhancing the digital literacy and search skills of lay users [57]. Moreover, several recommendations have been proposed to support safe and informed health information searches online, including (1) using trustworthy websites with official domains such as .gov, .edu, or .org; (2) verifying authorship and sources, and prioritizing content reviewed by scientific societies, health organizations, or accredited institutions; (3) avoiding personal testimonials as clinical evidence, since anecdotal narratives do not substitute scientific knowledge; and (4) consulting health care professionals before making any medical decisions [51,54,58,59].

In the field of pulmonology and respiratory diseases, the Reddit social media platform has previously been used to identify unmet needs, sentiments toward areas of interest, shared patient concerns, and the adoption of (unhealthy) behaviors related to e-cigarette use during the COVID-19 pandemic [60], lung cancer [61], vaping use [62], and respiratory lesions [63,64]. However, no previous work from Reddit's analysis has been found for people with asthma. Asthma carries a significant burden on patients' lives, with an important impact on their quality of life and health loss burden, measured by disability-adjusted life years, years of life lost, and years lived with disability [65]. Ten Have et al [66] showed that there are 4 asthma symptoms that, due to their impact, need to be prioritized in their resolution: fatigue, sleep disturbance, impairment of physical activity, and work-related symptoms. Moreover, our results do not show posts with relevant weight on fatigue or sleep disturbances, but on walking ability (topic 1, the fifth by weight, see Figure 1), and on work activity (topic 9, the 12th by weight). Our results revealed posts and comments from patients with Asthma on

XSL•FO
RenderX

Reddit discussing the onset of asthma attacks, changes in symptoms throughout the year, and seasonal variations linked to colder temperatures (topics 2, 19, and 23). This interest aligns with previous studies showing that extreme temperatures—both hot and cold—can trigger asthma attacks and worsen symptoms [67,68].

The strategy for asthma prevention and remission of the European Forum for Research and Education in Allergy and Airway Diseases Consensus Statement [69] adopts a phenotype-centered approach aimed at improving clinical outcomes and preserving health-related quality of life. This strategy moves away from the current "one-size-fits-all" concept, which focuses on symptom-oriented treatment strategies. In addition, there is growing support among experts for including nonpharmacological and interdisciplinary interventions—such as breathing exercises and multicomponent services—in the asthma management "toolbox." These approaches aim to help manage clinical symptoms, improve patients' quality of life, reduce health care costs, and address unmet needs [70]. This would help to decrease the discordance between compliance rates reported by patients with asthma and physicians [71].

Additionally, another aim would be to avoid loss of adherence to treatments. Thus, Amin et al [72] described how denial of the disease and of the need for long-term treatment, along with poor physician-patient communication, suboptimal knowledge of asthma medication (lack of understanding of the distinction between maintenance and reliever inhalers), suboptimal inhaler technique, and the high cost of asthma medication, were key factors in poor medication adherence. Access to pharmacological treatment, its correct use, identification of adverse effects, and use of other therapies appear in our results in six topics (Topics 4, 6, and 13-16), with special consideration of inhaled treatments as a means of administration.

At this point, patient participation is key to treatment adherence and adoption of healthy behaviors. Kang et al [73], in their systematic review and meta-synthesis of qualitative studies on shared decision-making, highlight that patients with asthma engage in shared decision-making when they have the opportunity (ie, environmental factors such as social, cultural, and institutional conditions that either facilitate or hinder individual behavior). The capability (skills, knowledge, and abilities required for individuals to engage in a particular behavior), and the motivation (internal psychological factors that drive individuals to choose specific behaviors based on their motivations and goals. The authors of this study believe that Reddit could be a tool used by patients with asthma to gain more capacity and opportunity to manage the disease, its symptoms, and treatments.

Another possible explanation for the use of Reddit as a source of information, obtaining medications and guidelines for the use of pharmaceutical products and drugs, may be related to the price of health care provided by professionals (doctors) and the cost of drugs [72,74,75]. Zhang et al [65], in their study on health loss and the economic burden of asthma in China, described an annual direct cost of between US $348 and US $1187 per capita, indirect costs of US $7 to US $1195, and hospitalization costs of US $177 to US $1547, influenced by the frequency and severity of acute exacerbations, comorbidities, and treatment adherence. Each country has different health systems with their own models of resource organization (public vs private), which could influence access to health care, medical information and recommendations, and pharmacological treatments for people with asthma [75].

Our results revealed a range of emotions and sentiments, with notable differences between the positive sentiment scores found using the AFINN lexicon and the negative sentiments identified through the Stanford NLP, Bing, and NRC lexicons. This variability in both positive and negative emotions also appears in the study by Volpato et al [74], where patients with uncontrolled asthma experienced a wide range of emotions that impacted their daily lives. Fear and anxiety appeared alongside the unpredictability of asthma attacks, while frustration and hopelessness accompanied the symptoms [74].

Limitations and Strengths

Among the limitations of this study is the inability to confirm an asthma diagnosis in the users who participated in the r/Asthma forum on Reddit, as well as the inability to determine other sociodemographic data of the forum participants [40]. However, the anonymous and private nature of social media would allow users to share questions, doubts, embarrassing topics, and make comments without restrictions that they would not make in other contexts [10]. Another limitation is that there was no uniformity in the comments, with some posts providing more information and content than others. This limitation was resolved by including all possible response threads derived from the posts in the study. The third limitation is that no keywords or combinations were used to search for and include posts in this study. In our study, since there were no previous studies on Reddit content related to asthma, all posts and threads generated during the study year were included. Finally, the reliability of the information shared on Reddit and other online social media platforms is essential for fostering trust in such content, promoting healthy behaviors, and protecting users' health [54]. In this initial study conducted within the Reddit r/Asthma forum, we presented the thematic content and types of information shared by users, as well as the emotional expressions and polarity of their language. Based on this preliminary work, further research would be justified to investigate the credibility and reliability of the information shared in thematic posts, and to explore potential relationships or associations between specific emotional states and the quality of the information provided.

In relation to strengths, social media analysis (Reddit) can provide key information compared to other data collection methods, such as traditional surveys, without being limited by issues such as sensitivity or taboos surrounding certain topics [10,40]. Conversely, among all available online social media platforms, only Reddit was used due to its structure based on thematic forums or communities (subreddits), where users share content and engage in discussions focused on specific topics or interests. This structure contrasts with other platforms, such as Instagram or X, which are primarily oriented toward following individuals. Moreover, content visibility on Reddit is determined

XSL•FO
RenderX

by user evaluations—either positive or negative—rather than by algorithms that prioritize virality or engagement metrics. Reddit also enables extensive discussions without restrictions on character count, image format, or source type, unlike other platforms that impose limitations based on character length (X), visual format (Instagram), or social network boundaries (Facebook). This would allow for the collection of valuable insights into users' motivations and the nature of the content they share. Furthermore, another strength of Reddit analysis for understanding patient feelings and behaviors is that large volumes of data can be obtained in a short period of time, and user-generated content can be processed quickly [76,77]. In addition, social media generates and shares a large amount of varied information on health and disease topics, allowing patients and their families to learn about different aspects of health care, such as diagnostic processes, treatments, coping strategies, professional advice, etc.

Conclusions

These findings can help uncover hidden needs and challenges faced daily by forum users and may assist health care professionals in developing targeted interventions aimed at improving knowledge around key topics. Identifying the most frequently shared thematic content (weighted topics) is relevant for health care professionals, as it may not reflect their clinical priorities. It also enables the assessment of information quality and the detection of potentially risky behaviors that are not supported by scientific evidence. In future research, it would be necessary to examine the scientific quality of the information shared in posts within Reddit's r/Asthma forum and its relationship to emotional state when sharing posts.

Acknowledgments

This study would not have been possible without the support of the Sociedad Castellanoleonesa y Cántabra de Patología Respiratoria (Spanish and Cantabrian Society of Respiratory Pathology).

Funding

This research was funded by Sociedad Castellanoleonesa y Cántabra de Patología Respiratoria (Spanish and Cantabrian Society of Respiratory Pathology). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated and analyzed during this study is not publicly available; although Reddit posts are publicly available and anonymized by username, Redditors did not consent for their content to be shared publicly in a curated dataset. The dataset is available from the corresponding author on reasonable request for research purposes.

Authors' Contributions

Conceptualization: EC-S, GS-P, AM-V, ECP-M, JR-G, ID-G
Data curation: JNC-Z (lead), DP-C
Formal analysis: JNC-Z
Funding acquisition: EC-S
Investigation: JNC-Z (lead), DP-C (equal), EC-S, GS-P, AM-V, ECP-M, JR-G, ID-G
Methodology: DP-C (lead), JNC-Z (equal), EC-S (supporting)
Project administration: JNC-Z (lead), DP-C (equal), EC-S (supporting)
Resources: EC-S
Supervision: EC-S
Validation: EC-S
Visualization: GS-P
Writing – original draft: DP-C (lead), JNC-Z (equal), EC-S (supporting).
Writing – review & editing: DP-C (lead), JNC-Z (equal), EC-S (supporting), GS-P (supporting), AM-V (supporting), ECP-M (supporting), JR-G (supporting), ID-G (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Diagnostic model plots by topics number.
[PNG File , 229 KB - jmir_v28i1e77027_app1.png]

Multimedia Appendix 2

XSL•FO
RenderX

Best selected topic models exclusivity versus semantic coherence graph.
[PNG File , 395 KB - jmir_v28i1e77027_app2.png]

Multimedia Appendix 3
Examples of texts and narratives from the Reddit asthma forum by topic. Each post includes the date it was posted.
[DOCX File , 22 KB - jmir_v28i1e77027_app3.docx]

Multimedia Appendix 4
Topics correlation. Blue line thickness is proportional to correlational value.
[PNG File , 171 KB - jmir_v28i1e77027_app4.png]

Multimedia Appendix 5
Significant topics time-change predicted values between October 2023 and October 2024.
[PNG File , 224 KB - jmir_v28i1e77027_app5.png]

Multimedia Appendix 6
Reddit r/Asthma posts descriptive sentiments and emotions by dictionary and month.
[DOCX File , 24 KB - jmir_v28i1e77027_app6.docx]

Multimedia Appendix 7
Likelihood ratio test for sentiments and emotions by dictionary time series model comparison (ap values).
[DOCX File , 17 KB - jmir_v28i1e77027_app7.docx]

Multimedia Appendix 8
Final Reddit r/Asthma posts sentiments and emotions by dictionary time series models.
[DOCX File , 23 KB - jmir_v28i1e77027_app8.docx]

Multimedia Appendix 9
Significant sentiments and emotions category overall predicted values by dictionary.
[PNG File , 194 KB - jmir_v28i1e77027_app9.png]

References

1. Porsbjerg C, Melén E, Lehtimäki L, Shaw D. Asthma. Lancet 2023;401(10379):858-873. [doi: 10.1016/S0140-6736(22)02125-0] [Medline: 36682372]
2. Schoettler N, Strek ME. Recent advances in severe asthma: from phenotypes to personalized medicine. Chest 2020;157(3):516-528 [FREE Full text] [doi: 10.1016/j.chest.2019.10.009] [Medline: 31678077]
3. Nolasco S, Crimi C, Campisi R. Personalized medicine in asthma: current approach and future perspectives. J Pers Med 2023;13(10):1459 [FREE Full text] [doi: 10.3390/jpm13101459] [Medline: 37888070]
4. Mosnaim G. Asthma in adults. N Engl J Med 2023;389(11):1023-1031. [doi: 10.1056/NEJMcp2304871] [Medline: 37703556]
5. Gonzalez-Uribe V, Romero-Tapia SJ, Castro-Rodriguez JA. Asthma phenotypes in the era of personalized medicine. J Clin Med 2023;12(19):6207 [FREE Full text] [doi: 10.3390/jcm12196207] [Medline: 37834850]
6. Pollard S, Bansback N, FitzGerld JM, Bryan S. The burden of nonadherence among adults with asthma: a role for shared decision-making. Allergy 2017;72(5):705-712. [doi: 10.1111/all.13090] [Medline: 27873330]
7. Sanchis J, Gich I, Pedersen S, Aerosol Drug Management Improvement Team (ADMIT). Systematic review of errors in inhaler use: has patient technique improved over time? Chest 2016;150(2):394-406 [FREE Full text] [doi: 10.1016/j.chest.2016.03.041] [Medline: 27060726]
8. Crespo-Lessmann A, Plaza V, González-Barcala FJ, Fernández-Sánchez T, Sastre J. Concordance of opinions between patients and physicians and their relationship with symptomatic control and future risk in patients with moderate-severe asthma. BMJ Open Respir Res 2017;4(1):e000189 [FREE Full text] [doi: 10.1136/bmjresp-2017-000189] [Medline: 29018525]
9. Matsunaga K, Hamada K, Oishi K, Yano M, Yamaji Y, Hirano T. Factors associated with physician-patient discordance in the perception of asthma control. J Allergy Clin Immunol Pract 2019;7(8):2634-2641. [doi: 10.1016/j.jaip.2019.04.046] [Medline: 31100555]
10. Chan GJ, Fung M, Warrington J, Nowak SA. Understanding health-related discussions on reddit: development of a topic assignment method and exploratory analysis. JMIR Form Res 2025;9:e55309 [FREE Full text] [doi: 10.2196/55309] [Medline: 39879094]

XSL•FO
RenderX

11. Zhao X, Yang V, Menta A, Blum J, Ranasinghe P. Exploring the use of social media for medical problem solving by analyzing the subreddit r/medical_advice: quantitative analysis. JMIR Infodemiology 2025;5:e56116 [FREE Full text] [doi: 10.2196/56116] [Medline: 40112288]
12. Sun Y, Zhang Y, Gwizdka J, Trace CB. Consumer evaluation of the quality of online health information: systematic literature review of relevant criteria and indicators. J Med Internet Res 2019;21(5):e12522 [FREE Full text] [doi: 10.2196/12522] [Medline: 31045507]
13. Sillence E, Briggs P, Harris PR, Fishwick L. How do patients evaluate and make use of online health information? Soc Sci Med 2007;64(9):1853-1862. [doi: 10.1016/j.socscimed.2007.01.012] [Medline: 17328998]
14. Pawlus Z, Spyra A, Polak K, Miziołek B, Bergler-Czop B. Can social media encourage young Polish adults to visit a dermatologist? An original study. Postepy Dermatol Alergol 2025;42(1):68-74 [FREE Full text] [doi: 10.5114/ada.2025.147554] [Medline: 40114760]
15. Munjita SM. Understanding vaccine hesitancy: insights from social media on polio, human papilloma virus, and COVID-19 in Zambia. Digit Health 2025;11:20552076251326131 [FREE Full text] [doi: 10.1177/20552076251326131] [Medline: 40109405]
16. Atarere J, Annor E, Bilalaga MM, Egbo O, Gaddipati GN, Vasireddy R, et al. Social media use and the relationship with colorectal cancer screening among foreign-born populations in the United States. Cancer Causes Control 2025;36(8):845-851. [doi: 10.1007/s10552-025-01985-6] [Medline: 40100525]
17. Siddiqui ZA, Pathan M, Nduaguba S, LeMasters T, Scott VG, Sambamoorthi U, et al. Leveraging social media data to study disease and treatment characteristics of Hodgkin's lymphoma using natural language processing methods. PLOS Digit Health 2025;4(3):e0000765. [doi: 10.1371/journal.pdig.0000765] [Medline: 40106471]
18. Guo IJ, Padmita AC, Matsuzaki M, Gittelsohn J, Feeley A, Watson F, et al. The use of social media to promote unhealthy food and beverage consumption among Indonesian children. BMC Nutr 2025;11(1):57. [doi: 10.1186/s40795-025-01040-2] [Medline: 40119452]
19. Berger J, Milkman KL. What makes online content viral? J Market Res 2012;49(2):192-205. [doi: 10.1509/jmr.10.0353]
20. Reddit Inc. United States Securities and Exchange Commission form S-1. URL: https://www.sec.gov/Archives/edgar/data/1713445/000162828024006294/reddits-1q423.htm#i1b9a579e78a34dfa99f7f26daeec195b_88 [accessed 2025-03-23]
21. Foundation Marketing. URL: https://foundationinc.co/lab/reddit-statistics/ [accessed 2025-03-23]
22. Lu Y, Zhang P, Liu J, Li J, Deng S. Health-related hot topic detection in online communities using text clustering. PLoS One 2013;8(2):e56221 [FREE Full text] [doi: 10.1371/journal.pone.0056221] [Medline: 23457530]
23. Reddit. URL: https://redditinc.com/ [accessed 2025-03-23]
24. Reddit rules. URL: https://redditinc.com/policies/reddit-rules [accessed 2025-03-23]
25. Forum r/asthma. Reddit. URL: https://www.reddit.com/r/Asthma/ [accessed 2025-03-23]
26. Luke DA, Tsai E, Carothers BJ, Malone S, Prusaczyk B, Combs TB, et al. Introducing SoNHR-reporting guidelines for social networks in health research. PLoS One 2023;18(12):e0285236 [FREE Full text] [doi: 10.1371/journal.pone.0285236] [Medline: 38096166]
27. Introducing So NHR-reporting guidelines for social networks in health research. Enhancing the quality and transparency of health research (Equator). URL: https://www.equator-network.org/reporting-guidelines/introducing-sonhr-reporting-guidelines-for-social-networks-in-health-research/ [accessed 2025-03-23]
28. Rivera I. RedditExtractor: tools for extracting reddit comments and submissions (version 2.3.0). GitHub. URL: https://github.com/ivan-rivera/RedditExtractoR/ [accessed 2025-10-16]
29. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2024. URL: https://www.R-project.org/ [accessed 2025-12-17]
30. Rajkumar AV, Suresh C, Veni M, Narasimha MM. On finding the natural number of topics with latent dirichlet allocation: some observations. In: Advances in Knowledge Discovery and Data Mining. Heidelberg: Springer; 2010:391-402.
31. Campbell JC, Hindle A, Stroulia E. Latent dirichlet allocation: extracting topics from software engineering data. In: The Art and Science of Analyzing Software Data. Cambridge, MA: Morgan Kaufmann Publishers; 2015:139-159.
32. Gupta A, Aeron S, Agrawal A, Gupta H. Trends in COVID-19 publications: streamlining research using NLP and LDA. Front Digit Health 2021;3:686720 [FREE Full text] [doi: 10.3389/fdgth.2021.686720] [Medline: 34713157]
33. Liu B. Sentiment analysis and subjectivity. In: Handbook of Natural Language Processing. Boca Raton, Florida: Chapman and Hall/CRC Press; 2010:627-666.
34. Finn Årup N. Evaluation of a word list for sentiment analysis in microblogs. 2011 Presented at: Proceedings for "Making Sense of Microposts: Big things come in small packages" (MSM2011), Volume 718 of CEUR Workshop; May 30, 2011; Heraklion, Greece p. 93-98.
35. Mohammad SM, Turney PD. Crowdsourcing a word–emotion association lexicon. Computational Intelligence 2012;29(3):436-465. [doi: 10.1111/j.1467-8640.2012.00460.x]
36. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford coreNLP natural language processing toolkit. Association for Computational Linguistics; 2014 Presented at: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations; December 16, 2025; Baltimore, Maryland p. 55-60. [doi: 10.3115/v1/p14-5010]

XSL•FO
RenderX

37. Halliday MAK, Hasan R. Cohesion in English. Milton Park, UK: Routledge; 1976.
38. Ribeiro FN, Araújo M, Gonçalves P, André Gonçalves M, Benevenuto F. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Sci 2016;5(1):23 [FREE Full text] [doi: 10.1140/epjds/s13688-016-0085-1]
39. Velu SR, Ravi V, Tabianan K. Multi-lexicon classification and valence-based sentiment analysis as features for deep neural stock price prediction. Sci 2023;5(1):8 [FREE Full text] [doi: 10.3390/sci5010008]
40. Shmueli-Scheuer M, Silverman Y, Halperin I, Gepner Y. Analysis of reddit discussions on motivational factors for physical activity: cross-sectional study. J Med Internet Res 2025;27:e54489 [FREE Full text] [doi: 10.2196/54489] [Medline: 39805106]
41. Kramer ML, Polo JM, Kumar N, Mulgirigama A, Benkiran A. Living with and managing uncomplicated urinary tract infection: mixed methods analysis of patient insights from social media. J Med Internet Res 2025;27:e58882 [FREE Full text] [doi: 10.2196/58882] [Medline: 40067345]
42. Joshi A, Kaune DF, Leff P, Fraser E, Lee S, Harrison M, et al. Self-reported side effects associated with selective androgen receptor modulators: social media data analysis. J Med Internet Res 2025;27:e65031 [FREE Full text] [doi: 10.2196/65031] [Medline: 39965201]
43. Code of conduct. EphMRA. URL: https://tinyurl.com/yc7yfrrw [accessed 2025-03-23]
44. Savolainen R. Assessing the credibility of COVID-19 vaccine mis/disinformation in online discussion. J Inf Sci 2023;49(4):1096-1110 [FREE Full text] [doi: 10.1177/01655515211040653] [Medline: 37461399]
45. Doody S. Making sense of a pandemic: reasoning about COVID-19 in the intellectual dark web. Front Sociol 2024;9:1374042 [FREE Full text] [doi: 10.3389/fsoc.2024.1374042] [Medline: 39351293]
46. Chlabicz M, Nabożny A, Koszelew J, Łaguna W, Szpakowicz A, Sowa P, et al. Medical misinformation in polish on the world wide web during the COVID-19 pandemic period: infodemiology study. J Med Internet Res 2024;26:e48130 [FREE Full text] [doi: 10.2196/48130] [Medline: 38551638]
47. Dolatabadi E, Moyano D, Bales M, Spasojevic S, Bhambhoria R, Bhatti J, et al. Using social media to help understand patient-reported health outcomes of post-COVID-19 condition: natural language processing approach. J Med Internet Res 2023;25:e45767 [FREE Full text] [doi: 10.2196/45767] [Medline: 37725432]
48. Blake L, Tucker P, Vanderloo LM. Mothers' perspectives of the barriers and facilitators to reducing young children's screen time during COVID-19: a Reddit content analysis. PLoS One 2024;19(3):e0301089 [FREE Full text] [doi: 10.1371/journal.pone.0301089] [Medline: 38536885]
49. Ménard A, O'Sullivan T, Mulvey M, Belanger C, Fraser S. Perceptions of hospital care for persons with dementia during the COVID-19 pandemic: a social media sentiment analysis. Gerontologist 2024;64(7):gnad155. [doi: 10.1093/geront/gnad155] [Medline: 37943714]
50. Kim S, Warren E, Jahangir T, Al-Garadi M, Guo Y, Yang Y, et al. Characteristics of intimate partner violence and survivor's needs during the covid-19 pandemic: insights from subreddits related to intimate partner violence. J Interpers Violence 2023;38(17-18):9693-9716 [FREE Full text] [doi: 10.1177/08862605231168816] [Medline: 37102576]
51. Collins SE, Lewis DM. Social media made easy: guiding patients to credible online health information and engagement resources. Clin Diabetes 2013;31(3):137-141 [FREE Full text] [doi: 10.2337/diaclin.31.3.137]
52. Khojah M, Sarhan MY. Vaccination uptake is influenced by many cues during health information seeking online. Health Info Libr J 2025. [doi: 10.1111/hir.12564] [Medline: 39780332]
53. Probst Y, Saffioti E, Manche S, Eaton M. Examination of social media nutrition information related to multiple sclerosis: a cross-sectional social network analysis. Public Health Nutr 2025;28(1):e166. [doi: 10.1017/S1368980025100943] [Medline: 40964911]
54. Silburn A. Navigating online health information: empowerment vs. misinformation. Front Digit Health 2025;7:1555290 [FREE Full text] [doi: 10.3389/fdgth.2025.1555290] [Medline: 40778382]
55. Stokes-Parish J. Navigating the credibility of web-based information during the covid-19 pandemic: using mnemonics to empower the public to spot red flags in health information on the internet. J Med Internet Res 2022;24(6):e38269 [FREE Full text] [doi: 10.2196/38269] [Medline: 35649183]
56. De Caro W. Online health information, health literacy and therapeutic compliance: a theoretical framework. Eur J Public Health 2021;31(3) [FREE Full text] [doi: 10.1093/eurpub/ckab165.429]
57. Mohamed H, Kittle E, Nour N, Hamed R, Feeney K, Salsberg J, et al. An integrative systematic review on interventions to improve layperson's ability to identify trustworthy digital health information. PLOS Digit Health 2024;3(10):e0000638. [doi: 10.1371/journal.pdig.0000638] [Medline: 39453891]
58. US Department of Health and Human Services. How to know if health information on social media is accurate. National Institutes of Health. 2025. URL: https://tinyurl.com/2emf9ezu [accessed 2025-10-16]
59. US Department of Health and Human Services. Know the science: finding health information online. National Institutes of Health. 2025. URL: https://www.nccih.nih.gov/health/know-science/finding-and-evaluating-online-resources/finding-health-information-online/introduction [accessed 2025-10-16]

XSL•FO
RenderX

60. Watkins SL, Snodgrass K, Fahrion L, Shaw E. Contextualizing changes in e-cigarette use during the early COVID-19 pandemic and accompanying infodemic ("So Much Contradictory Evidence"): qualitative document analysis of reddit forums. J Med Internet Res 2025;27:e66010 [FREE Full text] [doi: 10.2196/66010] [Medline: 40112286]
61. Shah AM, Lee KY, Hidayat A, Falchook A, Muhammad W. A text analytics approach for mining public discussions in online cancer forum: analysis of multi-intent lung cancer treatment dataset. Int J Med Inform 2024;184:105375. [doi: 10.1016/j.ijmedinf.2024.105375] [Medline: 38367390]
62. Kierstead E, Silver N, Amato M. Examining quitting experiences on quit vaping subreddits from 2015 to 2021: content analysis. J Med Internet Res 2024;26:e52129 [FREE Full text] [doi: 10.2196/52129] [Medline: 39454194]
63. Wu D, Kasson E, Singh AK, Ren Y, Kaiser N, Huang M, et al. Topics and sentiment surrounding vaping on twitter and reddit during the 2019 e-cigarette and vaping use-associated lung injury outbreak: comparative study. J Med Internet Res 2022;24(12):e39460 [FREE Full text] [doi: 10.2196/39460] [Medline: 36512403]
64. Hswen Y, Yom-Tov E. Analysis of a vaping-associated lung injury outbreak through participatory surveillance and archival internet data. Int J Environ Res Public Health 2021;18(15):8203 [FREE Full text] [doi: 10.3390/ijerph18158203] [Medline: 34360495]
65. Zhang P, Xu J, Xu B, Zhang Y, Xie Y. Health loss and economic burden of asthma in China: a qualitative review based on existing literature. Arch Public Health 2025;83(1):28 [FREE Full text] [doi: 10.1186/s13690-025-01515-5] [Medline: 39905572]
66. Ten Have L, Meulmeester FL, de Jong K, Ten Brinke A. Patient-centred outcomes in severe asthma: fatigue, sleep, physical activity and work. Eur Respir Rev 2025;34(175):240122 [FREE Full text] [doi: 10.1183/16000617.0122-2024] [Medline: 40044187]
67. Han A, Deng S, Yu J, Zhang Y, Jalaludin B, Huang C. Asthma triggered by extreme temperatures: from epidemiological evidence to biological plausibility. Environ Res 2023;216(Pt 2):114489 [FREE Full text] [doi: 10.1016/j.envres.2022.114489] [Medline: 36208788]
68. Deng S, Han A, Jin S, Wang S, Zheng J, Jalaludin BB, et al. Effect of extreme temperatures on asthma hospital visits: modification by event characteristics and healthy behaviors. Environ Res 2023;226:115679. [doi: 10.1016/j.envres.2023.115679] [Medline: 36913996]
69. Jesenak M, Bobcakova A, Djukanovic R, Gaga M, Hanania NA, Heaney LG, et al. Promoting prevention and targeting remission of asthma: a EUFOREA consensus statement on raising the bar in asthma care. Chest 2025;167(4):956-974 [FREE Full text] [doi: 10.1016/j.chest.2024.11.035] [Medline: 39672229]
70. Holland AE, Lewis A. Evidence-based management of symptoms in serious respiratory illness: what is in our toolbox? Eur Respir Rev 2024;33(174):24025 [FREE Full text] [doi: 10.1183/16000617.0205-2024] [Medline: 39477357]
71. Al-Moamary M, Aggarwal B, Al-Ahmad M, Sriprasart T, Koenig S, Levy G, et al. Are treatment adherence factors apparent in patients with asthma and to physicians? Results from the APPaRENT 3 Survey. Adv Ther 2025;42(3):1506-1521. [doi: 10.1007/s12325-025-03105-x] [Medline: 39912987]
72. Amin S, Soliman M, McIvor A, Cave A, Cabrera C. Understanding patient perspectives on medication adherence in asthma:A targeted review of qualitative studies. Patient Prefer Adherence 2020;14:541-551 [FREE Full text] [doi: 10.2147/PPA.S234651] [Medline: 32210541]
73. Kang H, Pen Y, He Y, Yang X, Su J, Yang Q, et al. The experience of shared decision-making for people with asthma: A systematic review and metasynthesis of qualitative studies. Health Expect 2024;27(2):e14039 [FREE Full text] [doi: 10.1111/hex.14039] [Medline: 38613765]
74. Volpato E, Pennisi V, Pennisi A, Piraino A, Banfi P, D'Antonio S, et al. Delving into uncontrolled or severe asthma: perspectives from patients and healthcare professionals in a cross-sectional study. J Asthma Allergy 2024;17:1207-1226 [FREE Full text] [doi: 10.2147/JAA.S483020] [Medline: 39610847]
75. Buendia JA, Guerrero-Patino D, Buendia Sanchez JA. Analysis of the economically justifiable price of mepolizumab in adults with asthma in Colombia. J Asthma 2025;62(5):850-860. [doi: 10.1080/02770903.2024.2448736] [Medline: 39836038]
76. Shatz I. Fast, free, and targeted: Reddit as a source for recruiting participants online. Soc Sci Comput Rev 2016;35(4):537-549. [doi: 10.1177/0894439316650163]
77. Watts IVC. There can be benefits to discussing healthcare on social media. BMJ 2024;384:q566. [doi: 10.1136/bmj.q566] [Medline: 38448054]

Abbreviations

AR1: autoregressive structure with lag 1
ARMA: autoregressive moving average structure
EDF: effective degrees of freedom
EQUATOR: Enhancing the Quality and Transparency of Health Research
GAMM: generalized additive mixed model
IgE: immunoglobulin E
LMM: linear mixed model

XSL•FO
RenderX

LRT: likelihood ratio test
NRC: National Research Council Canada
Stanford NLP: Stanford Natural Language Processing

Edited by A Stone; submitted 06.May.2025; peer-reviewed by S Leuckert, M Herrero-Montes, M Chlabicz; comments to author 18.Aug.2025; accepted 14.Nov.2025; published 08.Jan.2026.

Please cite as:
Curto-Sánchez E, Salazar-Palacios G, Martín-Varillas A, Prieto-Maíllo EC, Ramos-González J, Dávila-González I, Palacios-Ceña D, Cuenca-Zaldivar JN
What Patients With Asthma Share When No One Listens: Multimethod Observational Study of Patient Narratives on Reddit
J Med Internet Res 2026;28:e77027
URL: https://www.jmir.org/2026/1/e77027
doi:10.2196/77027
PMID:41505750

©Elena Curto-Sánchez, Gabriela Salazar-Palacios, Ana Martín-Varillas, Estela Cristina Prieto-Maíllo, Jacinto Ramos-González, Ignacio Dávila-González, Domingo Palacios-Ceña, Juan Nicolas Cuenca-Zaldivar. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Moderators' Role in Fostering Health Practitioner Engagement Within an Online Active Learning Program (The Community Fracture Capture Learning Hub): Qualitative Relational Content Thematic Analysis

Ahmed M Fathalla[1], PhD; Shanton Chang[2], PhD; Ralph Audehm[3], MBBS; Cherie Chiang[1,4], MD; Alexandra Gorelik[1,5], MSc; Christopher J Yates[1,4], PhD; Steve Snow[6], BCom, LLB; Rahul D Barmanray[4], PhD; Sarah Price[1,7], PhD; Lucy Collins[8], MD; John D Wark[1,4], PhD

[1]Department of Medicine, The Royal Melbourne Hospital, The University of Melbourne, Melbourne, Victoria, Australia
[2]School of Computing and Information Systems, University of Melbourne, Melbourne, Australia
[3]Department of General Practice and Primary Care, University of Melbourne, Melbourne, Australia
[4]Department of Diabetes & Endocrinology, Royal Melbourne Hospital, Melbourne Health, Melbourne, Australia
[5]School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia
[6]Praxhub, Melbourne, Australia
[7]Department of Obstetric Medicine, Royal Women's Hospital, University of Melbourne, Melbourne, Australia
[8]Department of Medicine, School of Clinical Sciences, Monash University, Melbourne, Australia

Corresponding Author:
Ahmed M Fathalla, PhD
Department of Medicine, The Royal Melbourne Hospital
The University of Melbourne
4th Floor, Clinical Sciences Building
Melbourne, Victoria, 3010
Australia
Phone: 61 383445892
Email: ahmed.elsayed@unimelb.edu.au

Abstract

Background: Virtual communities of practice (VCoPs) offer flexible platforms for interprofessional education among primary care physicians (PCPs); yet, the role of moderators in optimizing engagement remains underexplored. The Community Fracture Capture Learning Hub, a case-based, interactive web-based program addressing osteoporosis management gaps, provides an ideal context to investigate moderators' impact on collaborative learning.

Objective: This study aimed to characterize moderators' roles and strategies in fostering engagement within health-focused VCoPs, using the Community Fracture Capture Learning Hub as a model.

Methods: A qualitative relational content analysis was conducted on discussion-board interactions across four 6-week cycles (May 2022-October 2023) involving 55 PCPs and 8 moderators. Data included anonymized forum comments of moderators throughout the 4 cycles. Inductive-deductive thematic analysis identified patterns in moderator-participant and moderator-moderator exchanges, focusing on engagement techniques, content facilitation, and interaction dynamics.

Results: Five interconnected themes were identified; (1) participation encouragement: moderators used personalized prompts that highlighted clinical knowledge, affirming language, and balanced encouragement frequency to stimulate involvement; (2) evident moderator collaboration: cross-moderation through mutual collaboration that focused on clinical knowledge, shared prompts, and comment-building reinforced community cohesion; (3) topic discussions: strategic techniques (querying, framing, and real-world contextualization) deepened clinical reasoning; (4) PCPs' clinical practice discussion: moderators bridged theory-practice gaps using case-based exchanges on screening, treatment protocols, and evidence translation; and (5) progression of engagement: moderators evolved from structured guidance to collaborative dialogue, reducing formality and increasing topic complexity over time.

XSL•FO
RenderX

Conclusions: Thematic analysis suggested that moderators are pivotal in transforming VCoPs into dynamic learning environments. Their ability to blend rational skills, clinical authenticity, and adaptive facilitation drives knowledge coconstruction and practice-relevant engagement. Future studies should consider investing in competency-based moderator training that emphasizes empathetic communication and collaborative techniques as a means of structuring effective VCoPs.

International Registered Report Identifier (IRRID): RR2-10.2196/57511

(J Med Internet Res 2026;28:e83764) doi:10.2196/83764

KEYWORDS

community-based fracture capture learning hub; virtual community of practice; web-based learning moderator; primary care physicians; case-based learning; osteoporosis education; qualitative content analysis; continuing professional development; facilitation strategies; moderator-participant interaction; collaborative learning

Introduction

Leveraging Virtual Communities of Practice to Foster Interprofessional Education, Knowledge Sharing, and Collaboration Among Primary Care Physicians

Interprofessional education (IPE) is a learning approach in which professionals from different disciplines are educated together to achieve a common goal [1]. As per the World Health Organization, "Interprofessional education (IPE) occurs when two or more professions learn about, from and with each other to enable effective collaboration and improve health outcomes" [2]. Acknowledging the critical role of interprofessional teamwork and collaboration in delivering high-quality patient care, there is increasing advocacy for training health care professionals in collaborative practice skills through IPE, as a supplement to their clinical skill development [3]. IPE has an essential role in breaking down professional silos among primary care physicians (PCPs) [4,5]. Systematic reviews have shown that implementing IPE and interprofessional collaboration (IPC) play vital roles in enhancing the effectiveness of care delivery and contributing to a variety of improved health outcomes [6,7]. One effective strategy for promoting IPE and IPC is the development of communities of practice (CoPs) [8]. These communities bring together individuals with shared interests to work collaboratively, enabling the exchange of information, the transfer of knowledge, the dissemination of best practices, and the strengthening of both professional and interprofessional skills [9]. CoPs are associated with a range of improvements, including immediate advantages such as identification of workforce competencies, knowledge sharing, and reuse of existing knowledge, while also supporting long-term outcomes, including the development of professional identity, increased current awareness, and improved implementation of evidence-based practices [9,10]. With the growth of web-based information-sharing tools, information and communications technology is increasingly being used to promote IPE and IPC, leading to the development of a wide range of virtual communities of practice (VCoPs) [11]. The primary motivation behind VCoPs is to establish networks of individuals who share common interests but are located in different geographic regions [12]. VCoPs have demonstrated effectiveness in several key areas, including the development and implementation of strategies to improve health services and the enhancement of collaboration among PCPs [13,14].

Enhancing Collaborative Learning in Primary Care: The Evolving Role of Moderators in Virtual Communities of Practice

PCPs encounter considerable difficulties in handling multiple chronic conditions during brief patient consultations, necessitating ongoing learning to maintain the use of evidence-based approaches [15,16]. Consequently, there is an increasing dependence on digital modalities for continuing professional development (CPD) [17-19], as they offer greater flexibility and time efficiency compared with traditional learning methods. Since the COVID-19 pandemic, social media platforms and VCoPs have gained prominence as effective tools for collaborative learning in the health care sector, overcoming time, cost, and geographical barriers [20-22]. VCoPs have been found to reduce professional isolation, enhance workforce retention, promote IPC, and offer a supportive environment that encourages active and safe participation in knowledge exchange [10,12].

Despite their potential, VCoPs often face challenges related to member recruitment, engagement, and long-term sustainability [23]. For VCoPs to be effective, several key factors have been identified as essential, including the presence of external support, strong and supportive leadership, effective facilitation, and the involvement of local champions [9]. Yet, a lack of clarity persists regarding the specific responsibilities of CoP moderators and the appropriate management of power dynamics within CoP groups [24]. We suggest that the role of the moderators in the formation and sustainability of effective VCoPs remains crucial. Evidence suggests that leaders and moderators enhance collaboration by clarifying expectations for participation, maintaining focus within discussions, and fostering an environment characterized by mutual respect, openness, and active engagement [12]. However, a knowledge gap persists in understanding the role of moderators in the success of VCoPs, particularly in relation to the dynamics of their interactions with participants and with one another—an area that remains underexplored in the existing literature. Hence, a full understanding of moderators' roles needs to be appreciated, so that, as a group, moderators are well prepared and trained, and can provide more strategic approaches that increase participant engagement.

In light of the capacity of VCoPs to serve as a flexible and accessible approach for enhancing PCPs' knowledge and confidence, and given the limited empirical investigation into

XSL•FO
RenderX

the influence of moderators on the effectiveness of VCoPs, this study was guided by the following research question: What roles do moderators play in enhancing the experience of participants in a VCoP?

The objective of the study was to investigate the moderators' involvement with the program and its members, and identify themes developed in these exchanges.

To answer the research question, in this study, we examined the role of moderators and their contributions in an innovative VCoP known as the Community Fracture Capture (CFC) Learning Hub, which is an interactive, case-based online education platform that facilitates small-group learning through adaptable and customized modules [25]. It is designed to address the osteoporosis treatment gap within community settings, support the adoption of evidence-based bone health practices among PCPs, and enhance the implementation of both the eLearning hub and fracture liaison service models in primary care, both nationally in Australia and internationally [26]. Adopting qualitative analysis of postings and interactions of 8 moderators across 4 cycles (each is 6 weeks long), the study was able to derive a number of emerging themes using qualitative thematic analysis approaches.

Methods

Methodology

Full details of the CFC's development are outlined in our published protocol paper [26].

Case Study: CFC Hub

The CFC Learning Hub is a secure and adaptable web-based platform designed to support collaborative learning [25,26]. Central to its approach is an interactive Discussion Forum that encourages participant engagement and peer-to-peer exchange on course material. The role of the moderators was specifically important in guiding the discussion. Experienced PCP moderators facilitated the group discussions, with support from specialist moderators as needed, fostering case discussions and directing the conversation based on the content's relevance and significance. A qualitative evaluation of the CFC model provides important insights into moderators' interactions with and experiences of this innovative educational approach, enriching quantitative results [27] by offering contextual understanding and a deeper interpretation of the observed outcomes.

Between May 2022 and October 2023, 4 separate 6-week small-group cycles of the program were implemented, involving a total of 55 PCPs. Of these, 33 (60%) PCPs actively participated in the activities by sharing comments or questions. Across all cycles, 8 PCPs and specialist physicians served as the program moderators. At any one time, there was always both a PCP and a specialist active in their respective roles as moderators. All moderators actively contributed to the ongoing discussion through commenting, posting, initiating conversations, or answering enquiries.

Understanding the experiences of the learning activity moderators is crucial for developing and evaluating an effective learning model [28,29]. Effective facilitation of small-group learning is critical for creating opportunities for collaborative engagement among learners and for fostering team-building competencies—skills that are essential for functioning within health care environments [28]. This qualitative study aimed to answer the research question set out above. We propose that moderator engagement within the CFC Learning Hub enhances the platform's functionality as a peer-driven educational tool by supporting the dissemination of current fracture prevention knowledge to PCPs across a range of professional backgrounds and experience levels, while also contributing to increased confidence and motivation among PCPs in delivering osteoporosis care, with these effects being measurable through integrated analytics systems.

Ethical Considerations

This study was approved by the Melbourne Health Human Research Ethics Committee (site reference number: 2016.24). All PCP participants provided electronic consent for the use of their anonymized data for research and audit purposes. Electronic informed consent was obtained through Praxhub's platform, with measures of data anonymization and deidentification in place to uphold privacy and confidentiality standards. Participation in the program was voluntary and did not involve any form of compensation.

Recruitment

Moderators were recruited by the principal investigators through position announcements circulated within internal networks. Candidates for moderator roles were interviewed informally and engaged in discussions outlining their role responsibilities, alignment with the project's timelines and milestones, and expectations for engagement. Final selection was based on clinical expertise, prior experience with similar projects, availability, and overall suitability with respect to the project's aims and team dynamics.

PCP participants were recruited via web promotion of the CFC program via the Praxhub platform [30], joining small private groups (12 to 16 PCPs per cycle) to engage in week-long modules, discussions, and quizzes within a secure, supportive learning environment.

Participants and Moderators

A total of 55 PCPs participated in the program across four 6-week cycles. Of these, 33 (60%) were active participants, sharing comments or questions. Across all cycles, 8 moderators (2 PCPs and 6 specialist physicians; Table 1) served as the program moderators by leading activities and encouraging discussion. The selected specialists were all clinical endocrinologists with a clinical and research interest in bone and mineral disorders.

XSL•FO
RenderX

Table 1. Total number of moderators and their specialty.

CFC[a] moderators	Number of moderators	Role
Specialists	6	• Filter and select relevant case studies. • Provide guidance and support clinical focus. • Primary focus: content preparation and clinical consultation aligned with up-to-date clinical practice.
PCPs[b] with special interest in bone health	2	• Filter and select relevant case studies. • Curate content designed by specialists to ascertain relevance to PCPs. • Introduce the topic and lead the group discussion. • Primary focus: direct interaction with members to encourage involvement and engagement.

[a]CFC: Community Fracture Capture.

[b]PCP: primary care physician.

Qualitative Data and Analysis

The qualitative data included in this analysis were collected from over 630 comments, 160 pages, and more than 53,000 words of weekly discussions provided by participants and moderators. In the current analysis, we looked more specifically for moderator-participant and moderator-moderator engagement, with a focus on (1) the nature and tone of moderators' contributions, (2) the motivation and aims of the moderators' contributions, (3) interactions between moderators and participants, and (4) interactions between moderators. All the relevant discussion comments were compiled for qualitative analysis.

We used collaborative thematic analysis, as guided by Braun and Clarke [31,32]. The thematic analysis was done by the first and second authors independently initially, after familiarization with the data. This process was carried out to identify patterns in the data. To address the study's aim, the 2 researchers used relational content analysis [33,34]. Initially, they independently generated codes using both semantic (focused on explicit meanings of the text) and latent (focused on underlying tone and assumptions) approaches to capture a broad range of insights. After initial codes were generated, the 2 authors then collaboratively refined and defined themes and subthemes, through rigorous discussion, based on the evidence from the qualitative data. This process was carried out to identify patterns in the data. In the next phase, these researchers collaboratively reviewed and refined the codes to identify recurring patterns aligned with key themes. These were then organized into subthemes under broader thematic categories. Finally, one researcher coded the full dataset against these subthemes, with a random subset reviewed for accuracy by a senior researcher.

Given the novel approach of the eLearning hub, themes were developed after each cycle and cross-checked against existing and emerging research in the field. As feedback was gathered throughout the program, ongoing refinements were made to both the structure and content of the hub.

A relational content analysis was conducted [33,34], and the findings were organized using both inductive (bottom-up) and deductive (top-down) methods, grouping diverse ideas into meaningful categories. The final themes and subthemes are illustrated with representative quotes from participants.

Results

Moderators' Engagement

This section details the program moderators' interactions with the program and its members, as well as their engagement techniques. This feedback can be categorized into 5 main themes: participation encouragement, evident moderator collaboration, topic discussions, PCPs' clinical practice discussions, and progression of nature of engagement (Figure 1).

XSL•FO
RenderX

Figure 1. Moderators' engagement and its main 5 themes and their subthemes. PCP: primary care physician.

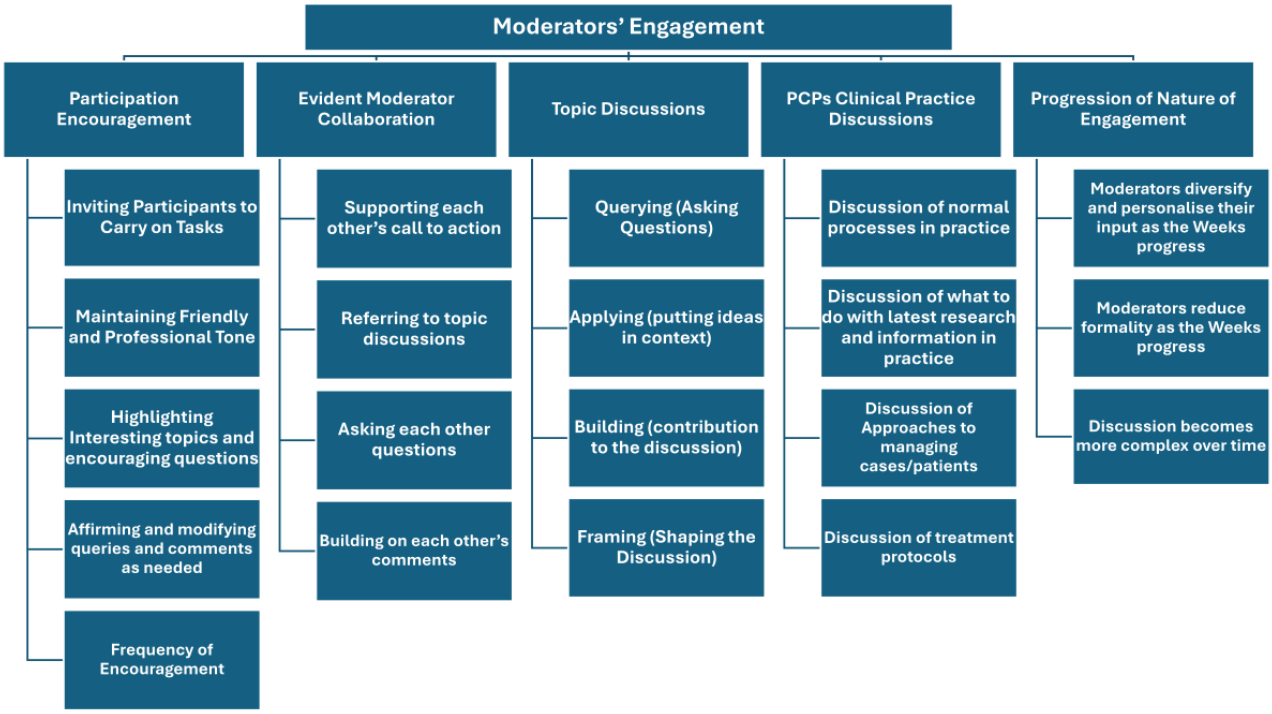

Participation Encouragement

Overview

Moderators of our virtual learning program used a variety of techniques to foster active participation and engagement among learners. Participants were shown a demonstration of the Praxhub webpage in a virtual, live session to permit interactions with lead moderators and other PCP participants. They actively invited participants to undertake tasks and contribute to discussions, ensuring that the tone remained friendly and professional to create a welcoming environment. By highlighting topics and areas of interest, our moderators aimed to capture learners' attention and maintain their interest. Encouraging questioning was another key strategy, as it was intended to stimulate continuous engagement with the activity. Additionally, moderators tried to affirm and constructively modify queries and comments, providing clear guidance and support. Striking the right balance in the frequency of encouragement communications was also crucial, aiming to provide learners with timely motivation without feeling overwhelmed. These techniques collectively aimed to help create an interactive and dynamic learning experience. This theme was broken down into 5 subthemes.

Inviting Participants to Carry Out Tasks

The moderators applied several key strategies for encouraging participation in the program. Moderators used targeted prompts to guide discussions on specific topics and encourage learners to engage with the material meaningfully. By inviting questions and actively acknowledging past participation, they created a supportive and interactive environment. Other techniques included emphasis on collaborative learning and using prompts for encouraging reflective thinking, further highlighting efforts to foster a sense of community and to stimulate deeper analysis.

One moderator encouraged participation using a targeted prompt:

> *I would also be very interested to hear your thoughts on strengths and limitations of DXA scans to screen for osteoporosis. If you have other questions, please also feel free to ask them.*

Another strategy was to highlight collaborative thinking:

> *We're really keen to hear your thoughts so that we can all learn from one another.*

Also, using prompts for reflective thinking was one of the participation encouragement techniques:

> *Yes, those are the questions to ask. I'm wondering how others think the answers to those questions might influence what we do next?*

Collectively, these approaches aimed to keep learners engaged, motivated, and actively contributing to the learning experience.

Maintaining a Friendly and Professional Tone

The moderators' tone across the activity demonstrated a consistent effort to be friendly, professional, encouraging, and sympathetic. Their informal and upbeat introduction aimed to set a welcoming tone, motivating learners to engage with the interactive hub and reassuring them that participation is confidential. The professional yet warm welcome was targeted at establishing a supportive environment for sharing and learning. Additionally, moderators' acknowledgment of a participant's contribution and encouragement for others to submit cases further fostered a collaborative atmosphere. Their affirmations and praise for correct insights and thoughtful responses were intended to be encouraging and supportive, reinforcing positive participation. In addition, their responses, including the appreciation for additional information and structured answers, showed attentiveness and respect for

XSL•FO
RenderX

learners' contributions. They provided personalized responses to participants, highlighting their questions and comments, which was intended to show genuine interest and support, promoting a positive and engaging learning experience.

One moderator gave the following friendly welcome to program participants:

> *Hi all. The weekend is nearly here. The perfect time to start getting involved in the CFC interactive learning hub. Peer-to-peer learning needs you. And don't be shy - what happens in the hub, stays in the hub! Ways to get involved?*

Another moderator also professionally welcomed participants:

> *Welcome to the CFC Bone Hub. I'm (name) and I am going to be your course moderator for Week 1. The purpose of this forum is to provide a space where we can learn from each other, through our own experiences, and feel comfortable sharing knowledge in a supportive environment.*

One moderator acknowledged a contribution and encouraged others to participate:

> *Thanks to one of our participants who has provided this good case and some questions to get us all thinking. Don't forget, any participants can send in an interesting case with some questions for consideration to be a case of the week.*

Another moderator praised members' contributions:

> *You are absolutely right that hyperparathyroidism preferentially results in cortical bone loss.*

Also, one moderator expressed appreciation for members' efforts,

> *Thanks for that extra information, very helpful in understanding more about the case!*

Other moderators personalized responses to different members in a professional and friendly manner:

> *(Member name)- this is a great point and highlights a crucial component in the case presented.*

Overall, the moderators effectively used a blend of encouragement, recognition, and constructive feedback to create a supportive and interactive learning environment.

Highlighting Interesting Topics and Encouraging Questions

Moderators tended to highlight and encourage questions to foster participant engagement. They posed questions to introduce new topics and stimulate deeper discussion and exploration. They also highlighted discussing additional topics and cases in an attempt to enrich the discussion with diverse perspectives. Such endorsement of interesting information can enhance the learning experience.

One moderator asked a question to stimulate discussion:

> *I'd like to ask a question about the fragility fracture risk calculators.*

Another moderator used an expression of interest presented by one of the members to highlight the importance of a noteworthy area and to ignite the discussion further:

> *Hi (member name), Great additional question. I might be pre-empting next week somewhat with a response but since you've asked, I'll give a general overview ahead of the greater detail you'll get in week 6.*

A third moderator suggested discussing an additional case vignette alongside the current one to enrich the discussion:

> *Another complementary case vignette has been submitted for discussion. Perhaps we can discuss it alongside the current case?*

Overall, these approaches reflected a strategic effort to engage participants by spotlighting intriguing topics and fostering a dynamic, interactive environment.

Affirming and Modifying Queries and Comments as Needed

Moderators worked to affirm and modify queries and comments to ensure clarity and enhance the quality of the discussion. They actively corrected and refined questions or responses to improve accuracy and relevance, ensuring the dialogue remained focused and productive.

For instance, one moderator acknowledged a necessary update to a post by stating, "This is a post edit," signaling that changes were made for clarity.

Another moderator pointed out a discrepancy in the resources and ensured the correct materials would be provided, saying:

> *We realised the answers and the resources were not loaded together, so we will be loading these.*

To prevent confusion, a third moderator pointed out an error in wording, suggesting a revision:

> *Note: the following question had an error in the wording which should have read …..*

Through these actions, the moderators not only affirmed the ongoing dialogue but also ensured that it was continually refined and improved. Overall, their efforts aimed to help maintain a precise and evolving conversation, fostering an atmosphere of constructive engagement.

Frequency of Encouragement

Moderators consistently accessed the discussion boards to monitor and participate in the dialogue, ensuring their engagement at least once daily throughout each cycle. The frequency of their comments and probing questions, aimed at encouraging further discussion, varied based on factors such as the level of response from participants, the depth of the topic being discussed, and the presence of unresolved queries or questions that required follow-up. On average, moderators provided at least one instance of daily engagement per day of the cycle. In addition to their participation in the discussion threads, moderators also sent email communications 3 times a week during the cycle, keeping participants informed about any updates, encouraging them to join in the discussions, and reminding them of the activity's requirements.

XSL•FO
RenderX

In summary, these efforts contributed to maintaining active participation and fostering a continuous, interactive learning environment throughout the cycle, as demonstrated by the participants' level of engagement in the activity.

Evident Moderator Collaboration

Overview

The theme of "Evident Moderator Collaboration" covers how moderators worked together to enhance the learning experience through mutual support and active engagement. This teamwork aimed to create a dynamic and cohesive learning environment. This collaboration is subcategorized into 4 subthemes: supporting each other's call to action, where moderators encouraged participant involvement by backing each other's prompts; referring to topic discussions, where they recalled earlier posts to create continuity and deepen the conversation; asking each other questions, which modeled inquisitive behavior for participants, encouraging them to engage more deeply; and building on each other's comments, where they expanded on points made by another moderator or participants to further enrich the dialogue.

Supporting Each Other's Call to Action

Moderators worked collaboratively to encourage participant engagement by supporting each other's calls to action. They consistently reinforced each other's prompts, creating a sense of shared responsibility in fostering discussion. This mutual support helped to drive the conversation forward and encouraged participants to respond actively.

For example, one moderator expressed eagerness for participant contributions to emphasize the importance of a colleague moderator's inquiry and motivated the participants to join the conversation by stating:

> *Looking forward to hearing your thoughts on (moderator) questions.*

In turn, the moderator acknowledged the colleague moderator's encouragement and shared enthusiasm for productive dialogue by responding:

> *Thanks (name). I'm looking forward to some good discussion.*

As a result, participants were encouraged to continue the conversation and responded effectively to the moderator's discussion. One of the participants further commented to the moderator, "I like to know how T score to be interpreted with Z score reading." They then clarified their understanding after receiving the moderator's response, "Now I understand that if Z score is below -2 then secondary causes should be investigated when the T score is also below -2.5."

Another moderator also contributed by showing appreciation for another colleague's efforts and building anticipation for future discussions, saying:

> *Looking forward to chatting tomorrow night. Thank you (name) for the teriparatide information!*

Overall, these interactions highlighted how moderators worked together to strengthen the call for engagement, fostering an atmosphere of cooperation and active participation.

Referring to Topic Discussions

Moderators consistently referenced previous conversations to ensure continuity and to build on ongoing discussions. This helped keep the dialogue focused and allowed for deeper exploration of the topics at hand.

As an example, the moderator referred directly to an earlier question regarding the best clinical practice for the management of oligomenorrhoea, demonstrating a clear connection to the ongoing conversation:

> *With regard to your question about oligomenorrhoea, reduced oestrogen exposure is certainly a risk factor for osteoporosis that should prompt consideration of osteoporosis screening by DXA.*

This response addressed a previous inquiry and contributed valuable insight into the current topic, reinforcing the relevance of earlier discussions. By referring to prior exchanges, moderators ensured that the conversation remained cohesive and that each new point was tied back to the broader discussion, enhancing the overall learning experience.

Asking Each Other Questions

Moderators demonstrated how inquisitive dialogue can enrich discussions and model active engagement for participants. By posing questions to one another, they not only sought clarification but also encouraged deeper exploration of the topics at hand.

For example, encouraging further discussion and highlighting the value of peer collaboration, one moderator prompted his colleagues by asking:

> *I wonder (if) (moderator one) or (moderator 2) (can share) what proportion would you think would be expected?*

Similarly, a moderator's question prompted deeper analysis of the case, inviting others to explore the situation more thoroughly, "Great talk (moderator). With the case study - was this a failure? Could we have prevented the fracture? She had low vitamin D, smoking and postmenopausal. Would there be a role for picking her up say 5 years ago and discussing lifestyle, weight training and doing a baseline DEXA if she was willing to pay for it?" which not only sought insight but also invited a peer-driven exchange of strategies.

These exchanges role-modeled the importance of asking questions to clarify, challenge, and deepen understanding, encouraging participants to actively engage in their own inquiries.

Building on Each Other's Comments

Moderators actively expanded on one another's contributions to deepen the discussion and explore various facets of the topic.

For example, a moderator's comment sparked further engagement from participants:

XSL•FO
RenderX

> *Great points have been raised in the thread above.*

Another moderator responded by acknowledging the discussion and extending it, encouraging the group to look beyond what had already been covered, saying:

> *Thanks for the comments (moderator one) and (moderator two). We have already discussed some aspects of this case last week, however there are more things to consider*

One moderator also built on another moderator's input and added another layer to the conversation by connecting patient perspectives to the clinical details, saying:

> *That is so true (moderator name). A lot of my patients associate dietary calcium to high dietary cholesterol.*

Overall, through these exchanges, moderators worked together to enhance the depth of the discussion, building on each other's comments with the aim of leading to more comprehensive and insightful dialogue.

Topic Discussions

Overview

Moderators used a range of techniques to facilitate participant engagement and deepen understanding of the subject matter. Moderators guided participants through the exploration of topics by applying various strategies, which aimed to enhance participants' comprehension and application of the material. This approach aimed to provide participants with opportunities to actively engage with the content and with one another. This theme is broken down into 4 subthemes.

Querying (Asking Questions)

Moderators used questions strategically to engage participants and steer discussions in productive directions. By asking thought-provoking and relevant questions, they encouraged participants to share their knowledge and explore different aspects of the topic. These questions helped highlight key areas for further discussion, often prompting deeper reflection.

For instance, one moderator posed a question to get participants thinking about the management of osteoporosis and setting the stage for a comprehensive discussion, saying:

> *Can I ask the group for your thoughts about the use of exercise to manage osteoporosis in your patients?*

Additionally, one moderator directly engaged a participant and fostered a more interactive environment by asking:

> *I am surprised by the inclusion of an incident fracture in the first year of treatment as all these medications take time to work and I would have thought a year on any of the treatments would not be long enough to have a real impact? (moderator name), what do you think?*

All in all, such questioning techniques effectively stimulated discussion, facilitated knowledge sharing, and ensured that the conversation remained focused and enriching for all involved.

Applying (Putting Ideas in Context)

Moderators used real-world examples and practical applications to help participants connect theoretical concepts to their everyday practices. This approach facilitated the discussion by encouraging participants to consider how the topics might play out in their own clinical settings.

For example, one moderator shared, "I also provide my patients with the Healthy Bones Australia fact sheets for calcium and vitamin D...," using this as an example of how they communicate essential information to patients, which prompted the activity's moderators and participants to reflect on similar practices.

As a result, one participant contributed by sharing their clinical practice and commented:

> *Of course, you can start patients … and this is what I do in frail patients…*

Additionally, moderators prompted participants to reflect on their observations in clinical practice, as one participant noted:

> *I always thought that pain from crush fractures was reported much less. It might be the degree of pain! I often see crush # on lateral spines - but no history of severe pain. Women do shrink - again often no specific pain but they all have some pain. What do the docs think about my experience?*

This invitation to share personal experiences helped ground the discussion in reality and led to a more nuanced exchange.

Overall, these applications of real-world examples allowed moderators to create a meaningful and relatable discussion, helping participants to apply new ideas and insights directly to their clinical contexts.

Building (Contribution to the Discussion)

Moderators focused on building the discussion by encouraging participants to contribute their insights and engage actively with the content. They created an environment where everyone was invited to share, fostering a collaborative atmosphere.

For instance, one moderator used a deliberately leading question to build on an ongoing conversation and encourage participants to consider various implications, asking:

> *Unfortunately, we don't have any more history on the 1986 fracture but as you say more details would be valuable. A change in therapy is probably warranted but (and it's a deliberately leading question!) are there any potential issues with changing straight from denosumab to teriparatide?*

Additionally, moderators prompted deeper engagement by inviting feedback from the group, as seen in the question:

> *I'm interested if this is consistent with your practice or what you've seen practised? And I'm keen on the group's thoughts as to whether the complexity of the plan is even required.*

This approach encouraged participants to reflect on their own practices and share their perspectives, thereby enriching the overall discussion.

XSL•FO
RenderX

By creating these opportunities for input, moderators fostered a sense of ownership and encouraged participants to contribute their experiences and expertise to the conversation, building a more dynamic and collaborative learning environment.

Framing (Shaping the Discussion)

Moderators used framing techniques to shape the direction and focus of discussions, ensuring that participants remained engaged and that the conversation stayed aligned with the educational goals. They often introduced topics by setting clear expectations for the flow of the discussion.

For example, one moderator guided the conversation by stating, "Therapies will be covered in depth later in the educational program, but it is worth discussing it briefly for case 2 before we move on to week 3's case," which helped participants understand the structure and provided context for the current discussion. Another moderator framed the week's session by acknowledging the contributions of participants, saying:

> *Welcome to week 3. Thanks to one of our participants who has provided this good case and some questions to get us all thinking. Don't forget any participants can send in an interesting case with some questions for consideration to be a case of the week.*

This approach not only shaped the direction of the current week's discussion but also encouraged ongoing participation from the group.

By encouraging participants to share their thoughts and questions, moderators aimed to create an interactive environment while maintaining focus on key topics. These framing techniques were essential in shaping the overall structure and ensuring that the discussions remained relevant and engaging for all participants.

PCPs' Clinical Practice Discussions

Overview

Moderators actively shared their own clinical practice experiences to enrich the learning process and make the discussions more relatable. By offering real-world examples from their own practices, they helped bridge the gap between theory and clinical application, which allowed participants to see how they might implement the discussed strategies in their own settings. Sharing such experiences not only helped participants gain a clearer understanding of the clinical aspects of the discussions but also fostered a collaborative learning environment, where participants could learn from each other's real-world experiences. This exchange of clinical knowledge created a richer, more interactive learning experience for all involved. This theme is subcategorized into 4 subthemes.

Discussion of Normal Processes in Practice

The moderators shared their normal clinical practices as a way of promoting discussion and encouraging learners to reflect on their own approaches. By presenting real-world scenarios from their practices, they opened the door for valuable exchanges and collaborative learning.

For example, one moderator used a targeted prompt to engage participants with a question on screening and calcium intake, stating:

> *I will be interested in what people think and how many screen patients before the first fracture. I make it a point to ask all women about calcium intake when coming in for contraceptive advice. I am amazed at how many have insufficient calcium in their diet!*

Similarly, another moderator invited discussion around the management of dental procedures in patients treated with antiresorptives:

> *I'm also keen to hear from colleagues as to how they approach elective dental procedures in patients treated with anti-resorptives. There are at least a few different approaches I'm familiar with in guidelines and from colleagues, am keen to hear from the group on what's done in the real world.*

These strategies encouraged reflection on common clinical practices, fostering a collaborative atmosphere where participants could learn from one another's experiences and approaches.

Discussion of What to Do With the Latest Research and Information in Practice

The moderators facilitated discussions on how to incorporate the latest research and information into clinical practice, encouraging participants to explore new findings and evaluate their practical implications.

One moderator shared insights from recent studies, saying:

> *That's an interesting question. I'm not familiar with human studies of the effect of ethanol on anti-resorptive actions. This basic science study in rats (study link) found that...,*

Additionally, they highlighted other key resources, like the National Prescribing Service Medicine Wise tables, offering practical tools for clinical decision-making:

> *It is a good resource - for people who have not seen it, it is here: (study link)*

Furthermore, one moderator brought in specific research, which led to a deeper conversation about the comparative efficacy of treatments, sharing:

> *Additionally, one large meta-analysis of 11 studies suggested that while denosumab improves bone density to a greater degree than bisphosphonates in post-menopausal osteoporosis, there is no difference in overall fracture rates*

These exchanges helped participants navigate the integration of emerging research into their everyday clinical practice.

Discussion of Approaches to Managing Cases/Patients

Moderators encouraged discussion around various approaches to managing complex cases, prompting participants to reflect on their own practices and share insights.

XSL•FO
RenderX

One moderator, for example, shared their surprise at a particularly challenging case, stating:

> *Wow, interesting case with a very low BMD in the spine - never seen one as low as that! I am interested as to what tests people would do when faced with such a case!*

This opened the floor for participants to discuss diagnostic strategies and test selections in similar scenarios.

Another moderator addressed a specific clinical consideration, saying, "For those who manage osteoporosis in patients with renal impairment, you might find the slides in the link useful," which encouraged participants to explore specialized resources tailored to managing osteoporosis in patients with this condition.

These discussions allowed participants to compare their approaches to managing unique and complex patient cases, fostering a collaborative environment for knowledge-sharing and practical problem-solving.

Discussion of Treatment Protocols

The moderators facilitated discussions on treatment protocols, encouraging participants to share their practices and perspectives on various approaches.

One moderator, for instance, initiated a discussion about the approach to managing osteoporosis in postmenopausal women, saying:

> *Dear all, how many of you organise a DEXA after menopause for women. I find I am performing it a lot more especially if any past risk factors. are there any particular groups that you screen?*

This prompted one participant to share real-life clinical experience:

> *I recently started working in a woman's clinic. I am struggling to find clear guidance as to when osteoporosis screening should start.*

Another moderator contributed by highlighting the relevance of specific guidelines, stating, "The Medicare Benefits Schedule (MBS) guidance on bone densitometry is quite helpful when considering some of the issues raised in this discussion," inviting participants to consider how these guidelines influence their treatment decisions.

Through these exchanges, participants had the opportunity to gain valuable insights into various approaches, strengthening their collective understanding and enhancing their ability to manage similar cases in their own practices.

Progression of Nature of Engagement

Overview

This theme focused on how the nature of discussions led by moderators evolved throughout the activity. Initially, moderators focused on providing clear instructions and creating a structured environment. Over time, they encouraged more collaborative and reflective dialogue, allowing participants to take on a more active role in shaping the discussions. This theme has been consistently observed throughout the cycles. It is broken down into 3 subthemes.

Moderators Diversify and Personalize Their Input as the Weeks Progress

Moderators had to diversify and personalize their input as the weeks progressed, adjusting their engagement to meet the evolving needs of the participants. Early in the activity, the moderators provided general updates and guidance, such as sharing quiz answers and encouraging participants to ask questions. For example, one moderator stated:

> *Hello everyone, please find the answers to the Week 2 Quiz as below. Please do not hesitate to query the answers, some questions are intended to be controversial! WEEK 2 MCQs (are as follows):*

As the activity evolved, the same moderator began tailoring responses to individual participants and their inquiries about the quiz, offering more detailed, personalized feedback, reflecting a shift towards more personalized engagement. For instance, the moderator reshared the MCQs answers directly with one of the participants, saying:

> *Hi (member name), please see week 2 MCQ answers as below:*

And addressing a member's understanding saying:

> *Hi (participant name), that is exactly right. Regarding osteopenia paradox: More fractures occur in subjects with osteopenia compared to osteoporosis, because although the relative risk of fracture is higher in osteoporosis, there are more people in the population with osteopenia.*

Additionally, affirming a member's feedback on an answer to a specific question:

> *Hi (member name), question 9 is designed to be controversial! You are right that anabolic should be considered in all the scenarios below, currently PBS prescriptions require 2 x fragility fractures (1 on treatment).*

The moderator's ability to adapt responses based on the participant's queries reflected progression from general messaging to more personalized, detailed interactions. This evolution in engagement demonstrates a shift towards fostering a more interactive and participant-centered learning environment.

Moderators Reduce Formality as the Weeks Progress

As the weeks progressed, moderators gradually reduced the formality in their communication to better engage with participants. In the early weeks, moderators maintained a professional tone, offering structured advice and information.

For instance, one moderator initially provided detailed and formal guidance, saying:

> *I would still do all the usual screening tests. She could still have other causes for her osteoporosis and could be devastating to miss these.*

XSL•FO
RenderX

However, as the activity evolved, the tone became more relaxed and informal, reflecting a shift towards a more conversational and engaging approach. In a later interaction, the same moderator expressed his thoughts in a less formal manner, saying:

> *OMG I am dizzy just thinking about the case. I would really be looking to the experts for this one... but how is it he is here – he has had renal physicians and endos for years... would they not have intervened early-is it not possible??*

This informal tone helped humanize the discussions and made the moderator appear more approachable. By using casual expressions and a less rigid style, moderators fostered a more comfortable environment, encouraging participants to engage more openly and freely in the discussions.

Discussion Becomes More Complex Over Time

As the activity progressed, the moderators facilitated discussions that grew increasingly nuanced and complex, challenging participants to consider more detailed aspects of patient care and clinical decision-making.

For instance, a moderator's questions grew more detailed and reflective over time. At an earlier stage of the cycle, the moderator raised an important, straightforward question about dual-energy X-ray absorptiometry (DXA) screening rates in practice, asking:

> *Given that all people over age 70 are eligible for a DEXA (DXA), my question to attendees is what proportion of people over 70 in your practice has had a screen?*

However, as the cycle progressed, the discussion became deeper, and the moderator's inquiries became more specific and clinically focused. For example, by mid-cycle, the moderator responded to relevant questions and discussed the routine practice of organizing DXA scans for older adults, pointing out how this had helped in detecting osteoporosis in several cases and highlighting the importance of considering physical signs:

> *Yes, age over 70 automatically qualifies her. I have been routinely organising DEXAs (DXA) for all my over 70s and have picked a decent amount of osteoporosis. Stooped posture would raise the issue of thoracic crush fractures - a lateral spine in her 60s may have given a diagnosis!*

Also, as the discussion went on, the moderator introduced more technical details and deepened the conversation about DXA scans, which led to a more advanced conversation on the variability and interpretation of diagnostic results, saying:

> *(It is) interesting that the wrist is so different and out of step to the other readings - and Colles fractures are so common and considered pathognomonic of osteoporosis. Most of my DEXAs (DXA) do not include the wrist!*

And later in a subsequent cycle, expanding the discussion to more unexplored areas, asking:

> *For me, one of the issues Is screening around the menopause. I often discuss with women going through menopause getting a baseline DEXA (DXA) scan (self-funded) - particularly to see where they are at. Is that something the rest of you do?*

These exchanges demonstrate how the moderators gradually steered the group from general discussions to more complex, case-specific inquiries, encouraging participants to engage with the subtleties of clinical practice and the evolving nature of evidence-based decision-making.

Discussion

Principal Findings

Overall, the CFC Learning Hub exemplifies the pivotal role of moderators in cultivating effective VCoPs for PCPs. Our qualitative analysis demonstrated how moderators serve as architects of engagement, bridging theoretical knowledge and clinical practice through strategic facilitation. Data analysis revealed that 5 core engagement themes emerged as pillars of effective moderation: participation encouragement, evident moderator collaboration, topic discussions, PCPs' clinical practice discussions, and Progression of Nature of Engagement. The study also showed that moderator collaboration was observable through mutual reinforcement of call to action, cross-referencing prior topic discussions, and peer-to-peer questioning and comment-building. In particular, moderators' engagement evolved in early weeks from formal instructions and structured guidance to informal dialogue, personalized feedback, and participant-led discussions in later weeks. Further, we found that topic discussions were facilitated via strategic querying, clinical contextualization, and progressive framing, where clinical practice discussions dominated exchanges, focusing on screening protocols, interpretation of guidelines, and management of complex cases. These findings align with and extend current understanding of effective VCoP facilitation in health professional education, emphasizing the nuanced interplay between relational warmth, strategic pedagogical techniques, and authentic clinical expertise.

In interpreting our findings through the lens of Wenger's [35,36] CoP framework, the 5 themes we identified cohere closely with the framework's theoretical dimensions of mutual engagement, joint enterprise, and shared repertoire. First, participation encouragement and progression of engagement reflect mechanisms by which moderators support mutual engagement: by prompting participation, recognizing contributions, and scaffolding newcomers' trajectories, moderators help sustain the regular, trust-based interactions that Wenger argues bind a CoP as a social entity (mutual engagement) [37,38]. Second, the themes of topic discussions and Clinical Practice Discussions embody aspects of the joint enterprise, as they signal how community members, under moderator facilitation, negotiate and coconstruct what the CoP is "about"—namely, relevant clinical challenges and professional concerns in practice—thus aligning with Wenger's notion of a collectively negotiated domain and mutual accountability for its purposes [24,38]. Finally, the evident moderator collaboration theme fosters the development of a shared repertoire, as moderators and

XSL•FO
RenderX

participating PCPs jointly build and maintain communal resources (eg, shared norms, language, case-based learning, narratives), which act as artifacts or tools for ongoing practice (eg, stories, tools, and routines) [39]. In sum, our empirical themes suggest that effective moderation not only reinforces Wenger's core CoP dimensions but also actively operationalizes them, facilitating a more cohesive, participatory, and sustainable community of practice.

The theme participation encouragement demonstrated how moderators proactively created an inclusive environment through targeted invitations, sustained professional yet warm communication, and strategic highlighting of relevant topics. This aligns with evidence that welcoming tones and explicit invitations reduce perceived barriers in web-based learning spaces, particularly for time-constrained professionals [28]. Techniques such as affirming contributions while constructively refining queries echo best practices in digital facilitation, where balancing support with intellectual rigor maintains engagement without compromising depth [40]. Moderators' daily presence and thrice-weekly email updates further reinforced reliability—a factor linked to sustained participation in web-based CPD [25]. These moderation practices likely satisfy core psychological needs described in self-determination theory (SDT)—autonomy, competence, and relatedness—thereby fostering more intrinsic motivation [41]. Research in online learning shows that perceived instructor (or moderator) support enhances satisfaction of these needs, which in turn promotes engagement [42,43]. By explicitly inviting participants and using clinically relevant framing, moderators support autonomy; by affirming and reflecting, they nurture competence and relatedness. Moreover, their reliable presence establishes trust and predictability, which helps external motivations internalize over time, increasing self-determined engagement. SDT has also been applied in CPD contexts: for example, a study of nursing professionals' mobile learning found that autonomy and competence were linked to higher motivation in continuing education [44].

VCoP moderators have been shown to enhance collaboration within a VCoP by ensuring clear rules of engagement, maintaining focus in discussions, and fostering engagement, shared respect, and openness [12]. CFC Learning Hub moderators played a pivotal role in sustaining engagement through collaborative questioning and personalized feedback. Their use of relational techniques (eg, affirming contributions or posing reflective questions) mirrors effective practices in web-based facilitation for health professionals [45]. The use of affirming language, reflective prompts, and real-world case discussion aligns with other studies where empathetic communication and clinical relevance were identified as key facilitators of engagement in small-group learning [40]. For instance, their shift from formal instruction to informal dialogue resonates with CoP theory, where trust-building fosters deeper engagement [35,36]. In addition, this shift toward a more informal, trust-based dialogue reflects CoP principles: as trust builds, participants gradually move from peripheral to more central roles (legitimate peripheral participation), deepening their ownership of the community [36,46]. In addition, the use of techniques like framing and query building in topic discussions contributed to fostering activity homogeneity and focus. Reports indicated that a VCoP tends to see higher levels of member interest and involvement when it is more homogeneous or 'focused' [47]. Moreover, this facilitation style reflects principles of adult learning theories: adult learners are more engaged when learning is relevant, self-directed, and respectful of their professional experience [48]. By reducing access barriers and signaling respect for participants' expertise, moderators align with these andragogical principles. Furthermore, moderators' focus on using real-world clinical examples to contextualize discussions aligns with effective strategies in health-related learning environments [23,47]. Successful VCoPs require that the knowledge shared and discussed by members be applied to clinical practice [47]. Professionals are more likely to engage when the VCoP is patient-centered, and when guidelines and resources for translating research into practice are readily available in user-friendly formats [23].

The evident moderator collaboration theme highlighted how moderators visibly supported each other's prompts, built upon colleagues' comments, and modeled inquisitive dialogue. This intentional demonstration of teamwork aligns with Wenger's [35] concept of "communities of practice," where shared leadership fosters collective knowledge construction. By publicly engaging in peer questioning, moderators normalized critical thinking and reduced hierarchical barriers, and signaled that inquiry and reflection are communal practices rather than top-down directives. This distributed leadership likely promotes a sense of shared ownership among participants. Moreover, modeling collegial critical thinking may satisfy participants' relatedness needs, per SDT [41], creating a psychologically safe environment for contribution**.** This finding reinforces the value of cofacilitation in VCoPs, where multiple moderators can diversify expertise and model collaborative problem-solving [10]. Shared moderation builds structural resilience: having multiple moderators reduces reliance on any single individual, which supports long-term sustainability. Over time, as participants observe and internalize this co-leadership, they may increasingly adopt facilitative behaviors themselves, strengthening the community's capacity and self-governance [40]. Further, the Progression of Nature of Engagement theme illustrated moderators' evolution from structured guidance to collaborative partnership. Diversifying input (eg, shifting from group announcements to personalized feedback) and reducing formality mirrored community maturation, consistent with VCoP lifecycle models where trust enables deeper interaction [36]. Increasing discussion complexity over weeks reflected adaptive scaffolding—a facilitator skill linked to improved learning outcomes in web-based spaces [28]. Adaptive scaffolding in web-based education was reported to foster learner autonomy, engagement, and self-regulation [49]. From a motivational perspective, this progression supports increasing autonomy (participants gain more control), competence (they engage in complex discourse), and relatedness (peer interactions deepen), all of which facilitate more internalized engagement under SDT. Furthermore, as participants transition from peripheral to more central roles, this reflects the concept of legitimate peripheral participation: newcomers start with less risky participation and gradually assume more central, meaningful roles, reinforcing

XSL•FO
RenderX

both individual development and the stability of the community [50].

Our findings underscore that effective VCoP moderation transcends technical platform management; it requires relational, pedagogical, and clinical competencies. Accordingly, training programs for VCoP moderators should emphasize relational skills such as building rapport, maintaining approachable yet professional tones, and fostering psychological safety; pedagogical strategies that include techniques for scaffolding discussions (eg, framing and questioning), encouraging reflection, and modeling collaboration; and clinical authenticity, including the judicious sharing of real-world experiences to contextualize evidence without dominating discourse. Indeed, the presence of a supportive learning environment was reported as a key theme to facilitate the knowledge transfer from CPD to clinical practice [51]. In addition, platform designers should prioritize features that enable seamless moderator collaboration (eg, shared dashboards), personalized communication tools, and analytics to track engagement patterns. A safe, nonjudgmental educational environment; well-organized and structured programs; and the presence of a motivating and intrinsically satisfying atmosphere with smooth and transparent facilitation were among the main factors reported to influence the transfer of knowledge from continuing professional education to clinical practice [52]. The CFC Hub's success in using case-based discussions also supports designing VCoPs around practice-relevant scenarios rather than abstract topics. These insights additionally carry broader implications for practice, as the identified relational and pedagogical competencies, along with the platform design considerations, can inform moderator training and VCoP implementation across a wide range of clinical specialties and health care settings. For instance, moderation by cognition-based team swift trust was positively correlated with team creativity within nursing education [53]. Developing adaptable training programs and flexible platform architectures may enhance the scalability and relevance of VCoPs, thereby increasing their practical value and supporting more sustainable, cross-disciplinary digital professional development.

Limitations and Future Research

While this study yields meaningful insights into the role of moderated online learning environments, some limitations should be acknowledged. First, participant-moderator relationships may have been influenced by self-selection bias, with PCPs who were more motivated to improve osteoporosis care likely being overrepresented. Furthermore, participants were recruited via the Praxhub platform, with the majority being metropolitan-based practitioners with over 10 years of clinical experience. This recruitment approach may limit the generalizability of the findings, as the results may not fully capture the program moderation's impact across more diverse practice settings. Further evaluation is warranted, particularly among regional general practitioners and those with less clinical experience. Additionally, as is common in qualitative research, the study relied on self-reported data, which may be subject to response bias. Future studies should investigate the effectiveness of these educational strategies in larger and more heterogeneous health care populations to better assess their influence on patient care.

Future studies should expand contexts and test these themes in VCoPs across different specialty fields (eg, diabetes and mental health) and geographic settings to identify cross-cutting principles. Future research could also explore how moderator-participant dynamics evolve over extended periods and their impact on community sustainability. Although this study demonstrates immediate engagement and knowledge exchange, it remains uncertain whether these educational gains result in sustained changes in clinical behavior and patient outcomes. Longitudinal investigations are needed to assess the retention of knowledge and skills over time, confidence in providing optimal clinical care in the topics covered, and to evaluate their potential influence on patient management, including relevant clinical outcomes. Further, an in-depth examination of the partnership between the 2 types of moderators is needed, as it would enable a critical analysis of their cross-functional dynamics and offer deeper insight into their complementary roles in delivering an engaging and clinically relevant learning experience.

Conclusions

This study elucidated the critical role of moderators in VCoPs for PCPs. Through 5 key themes- participation encouragement, evident moderator collaboration, topic discussions, PCPs' clinical practice discussions, and Progression of Nature of Engagement, we demonstrate how moderators drive engagement by blending relational warmth, strategic facilitation, and clinical authenticity. Moderators' ability to adapt from structured guidance to collaborative dialogue fosters a dynamic environment where knowledge is coconstructed and applied to real-world practice. These insights strongly suggest the need for dedicated moderator training and thoughtfully designed VCoP platforms that prioritize facilitator support. As digital learning becomes integral to professional development, optimizing moderator roles will be essential for scaling effective, sustainable communities that bridge evidence-practice gaps in health care, strengthen ongoing workforce development, and ensure that VCoPs remain resilient, adaptive, and impactful over the long term.

Acknowledgments

We express our gratitude to the Royal Australian College of General Practitioners (RACGP) and the Australian College of Rural and Remote Medicine (ACRRM) for accrediting the Community Fracture Capture (CFC) Bone Hub as a professional learning activity. Additionally, we appreciate the support and resources offered by the University of Melbourne Department of Medicine and the Department of Diabetes and Endocrinology at Royal Melbourne Hospital, and the Faculty of Engineering and Information Technology at the University of Melbourne.

XSL•FO
RenderX

Funding

This study was supported financially by Amgen Australia Pty Ltd and Theramex. These sponsors had no involvement in the study's design, content development, or in the conduct, data collection, analysis, and interpretation of the research.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

AMF, SC, CC, AG, RA, and JDW contributed to study conceptualization. AMF, SC, CC, AG, RA, SS, CJY, RB, SP, LC, and JDW contributed to the methodology and investigation. JDW, SC, CC, RA, RB, CJY, LC, and SP contributed to conducting the learning sessions; SS contributed to providing regular liaison with Praxhub. AMF, SC, CC, RA, and JDW contributed to the writing of the original draft. AMF, CC, RA, SC, and JDW contributed to the project administration. AMF, SC, CC, RA, AG, and JDW contributed to the data curation. JDW, SC, and CC contributed to study supervision. All authors have read, provided feedback, and agreed to the final version of the manuscript for publication.

Conflicts of Interest

None declared.

References

1. Mohammed CA, Anand R, Saleena Ummer V. Interprofessional education (IPE): a framework for introducing teamwork and collaboration in health professions curriculum. Med J Armed Forces India 2021;77(Suppl 1):S16-S21 [FREE Full text] [doi: 10.1016/j.mjafi.2021.01.012] [Medline: 33612927]
2. World Health Organization. 2010. URL: https://www.who.int/publications/i/item/framework-for-action-on-interprofessional-education-collaborative-practice [accessed 2025-10-05]
3. King S, Greidanus E, Carbonaro M, Drummond J, Patterson S. Merging social networking environments and formal learning environments to support and facilitate interprofessional instruction. Med Educ Online 2009;14:5 [FREE Full text] [doi: 10.3885/meo.2009.T0000132] [Medline: 20165519]
4. Frenk J, Chen L, Bhutta ZA, Cohen J, Crisp N, Evans T, et al. Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. Lancet 2010;376(9756):1923-1958. [doi: 10.1016/S0140-6736(10)61854-5] [Medline: 21112623]
5. Lygidakis H, McLoughlin C, Patel KD. Achieving universal health coverage: technology for innovative primary health care education. ResearchGate. 2016. URL: https://www.researchgate.net/publication/311616296_Achieving_Universal_Health_Coverage_Technology_for_innovative_primary_health_care_education [accessed 2026-02-17]
6. Reeves S, Pelone F, Harrison R, Goldman J, Zwarenstein M. Interprofessional collaboration to improve professional practice and healthcare outcomes. Cochrane Database Syst Rev 2017;6(6):CD000072 [FREE Full text] [doi: 10.1002/14651858.CD000072.pub3] [Medline: 28639262]
7. Reeves S, Perrier L, Goldman J, Freeth D, Zwarenstein M. Interprofessional education: effects on professional practice and healthcare outcomes (update). Cochrane Database Syst Rev 2013;2013(3):CD002213 [FREE Full text] [doi: 10.1002/14651858.CD002213.pub3] [Medline: 23543515]
8. Lees A, Meyer E. Theoretically speaking: use of a communities of practice framework to describe and evaluate interprofessional education. J Interprof Care 2011;25(2):84-90. [doi: 10.3109/13561820.2010.515429] [Medline: 21043557]
9. Ford J, Korjonen H, Keswani A, Hughes E. Virtual communities of practice: can they support the prevention agenda in public health? Online J Public Health Inform 2015;7(2):e222 [FREE Full text] [doi: 10.5210/ojphi.v7i2.6031] [Medline: 26284150]
10. McLoughlin C, Patel KD, O'Callaghan T, Reeves S. The use of virtual communities of practice to improve interprofessional collaboration and education: findings from an integrated review. J Interprof Care 2018;32(2):136-142. [doi: 10.1080/13561820.2017.1377692] [Medline: 29161155]
11. Barnett S, Jones SC, Bennett S, Iverson D, Robinson L. A virtual community of practice for general practice training: a preimplementation survey. JMIR Med Educ 2016;2(2):e13 [FREE Full text] [doi: 10.2196/mededu.5318] [Medline: 27731864]
12. Barnett S, Jones SC, Bennett S, Iverson D, Bonney A. General practice training and virtual communities of practice - a review of the literature. BMC Fam Pract 2012;13(1):87 [FREE Full text] [doi: 10.1186/1471-2296-13-87] [Medline: 22905827]
13. Abos Mendizabal G, Nuño Solinís R, Zaballa González I. HOBE+, a case study: a virtual community of practice to support innovation in primary care in Basque public health service. BMC Fam Pract 2013;14(1):168 [FREE Full text] [doi: 10.1186/1471-2296-14-168] [Medline: 24188617]

XSL•FO
RenderX

14. Curran JA, Murphy AL, Abidi SSR, Sinclair D, McGrath PJ. Bridging the gap: knowledge seeking and sharing in a virtual community of emergency practice. Eval Health Prof 2009;32(3):312-325. [doi: 10.1177/0163278709338570] [Medline: 19696084]
15. Allen LM, Palermo C, Armstrong E, Hay M. Categorising the broad impacts of continuing professional development: a scoping review. Med Educ 2019;53(11):1087-1099 [FREE Full text] [doi: 10.1111/medu.13922] [Medline: 31396999]
16. Magwenya RH, Ross AJ, Ngatiane LS. Continuing professional development in the last decade – a scoping review. J Adult Contin Educ 2022;29(2):408-437. [doi: 10.1177/14779714221147297]
17. Thepwongsa I, Kirby CN, Schattner P, Piterman L. Online continuing medical education (CME) for GPs: does it work? a systematic review. Aust Fam Physician 2014;43(10):717-721 [FREE Full text] [Medline: 25286431]
18. Cook DA, Blachman MJ, Price DW, West CP, Baasch Thomas BL, Berger RA, et al. Educational technologies for physician continuous professional development: a national survey. Acad Med 2018;93(1):104-112. [doi: 10.1097/ACM.0000000000001817] [Medline: 28658022]
19. Baso YS, Syafruddin A, Padlan AA, Akbar M. Continuing medical education for general practitioners in Indonesia: moving toward accountability. Gac Sanit 2021;35(Suppl 2):S120-S122 [FREE Full text] [doi: 10.1016/j.gaceta.2021.07.008] [Medline: 34929792]
20. Rolls K, Hansen M, Jackson D, Elliott D. How health care professionals use social media to create virtual communities: an integrative review. J Med Internet Res 2016;18(6):e166 [FREE Full text] [doi: 10.2196/jmir.5312] [Medline: 27328967]
21. Yee M, Simpson-Young V, Paton R, Zuo Y. How do GPs want to learn in the digital era? Aust Fam Physician 2014;43(6):399-402 [FREE Full text] [Medline: 24897992]
22. van der Keylen P, Tomandl J, Wollmann K, Möhler R, Sofroniou M, Maun A, et al. The online health information needs of family physicians: systematic review of qualitative and quantitative studies. J Med Internet Res 2020;22(12):e18816 [FREE Full text] [doi: 10.2196/18816] [Medline: 33377874]
23. Vinson CA. Using concept mapping to develop a conceptual framework for creating virtual communities of practice to translate cancer research into practice. Prev Chronic Dis 2014;11:E68 [FREE Full text] [doi: 10.5888/pcd11.130280] [Medline: 24762532]
24. Li LC, Grimshaw JM, Nielsen C, Judd M, Coyte PC, Graham ID. Use of communities of practice in business and health care sectors: a systematic review. Implement Sci 2009;4(1):27 [FREE Full text] [doi: 10.1186/1748-5908-4-27] [Medline: 19445723]
25. Murad A, Hyde N, Chang S, Lederman R, Bosua R, Pirotta M, et al. Quantifying use of a health virtual community of practice for general practitioners' continuing professional development: a novel methodology and pilot evaluation. J Med Internet Res 2019;21(11):e14545 [FREE Full text] [doi: 10.2196/14545] [Medline: 31774401]
26. Fathalla AM, Chiang C, Audehm R, Gorelik A, Chang S, Yates CJ, et al. Developing and evaluating an interactive, case-based, web-based active learning tool for primary care physicians (community fracture capture learning hub): protocol for an acceptability and engagement study. JMIR Res Protoc 2025;14:e57511 [FREE Full text] [doi: 10.2196/57511] [Medline: 39999431]
27. Fathalla AM, Chiang C, Audehm R, Gorelik A, Chang S, Dao T, et al. Evaluation of a novel web-based active learning tool for primary care physicians' continuing professional development (the community fracture capture learning hub): quantitative analysis. JMIR Form Res 2026 Jan 21;10:e76216 [FREE Full text] [doi: 10.2196/76216] [Medline: 41564332]
28. Burgess A, van Diggele C, Roberts C, Mellis C. Facilitating small group learning in the health professions. BMC Med Educ 2020;20(Suppl 2):457 [FREE Full text] [doi: 10.1186/s12909-020-02282-3] [Medline: 33272270]
29. Schiekirka-Schwake S, Anders S, von Steinbüchel N, Becker JC, Raupach T. Facilitators of high-quality teaching in medical school: findings from a nation-wide survey among clinical teachers. BMC Med Educ 2017;17(1):178 [FREE Full text] [doi: 10.1186/s12909-017-1000-6] [Medline: 28962568]
30. Free online medical CPD training. Praxhub. URL: https://praxhubcom/ [accessed 2026-02-11]
31. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
32. Clarke V, Braun V. Thematic analysis. J Posit Psychol 2016;12(3):297-298. [doi: 10.1080/17439760.2016.1262613]
33. Naeem M, Ozuem W, Howell K, Ranfagni S. A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. Int J Qual Methods 2023;22:16094069231205789. [doi: 10.1177/16094069231205789]
34. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a researcher. Int J Transgend Health 2023;24(1):1-6 [FREE Full text] [doi: 10.1080/26895269.2022.2129597] [Medline: 36713144]
35. Wenger E. Communities of Practice: Learning, Meaning, and Identity. New York, NY: Cambridge University Press; 1999.
36. Wenger E, McDermott RA, Snyder W. Cultivating Communities of Practice: A Guide to Managing Knowledge. USA: Harvard Business Press; 2002.
37. Murillo E. Communities of practice in the business and organization studies literature. Inf Res 2011;16(1):464 [FREE Full text]
38. Li LC, Grimshaw JM, Nielsen C, Judd M, Coyte PC, Graham ID. Evolution of Wenger's concept of community of practice. Implement Sci 2009;4:11 [FREE Full text] [doi: 10.1186/1748-5908-4-11] [Medline: 19250556]

XSL•FO
RenderX

39. Wenger-Trayner EWTB. An introduction to communities of practice: a brief overview of the concept and its uses. Wenger-Trayner. 2015. URL: https://www.wenger-trayner.com/introduction-to-communities-of-practice/ [accessed 2026-02-17]
40. D'Eon M, Zhao R. Five ways for facilitators to get a grip on small group learning. Can Med Educ J 2022;13(2):82-88 [FREE Full text] [doi: 10.36834/cmej.72949] [Medline: 35572022]
41. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
42. He J, Wang Q, Lee H. Enhancing online learning engagement: teacher support, psychological needs satisfaction and interaction. BMC Psychol 2025;13(1):696 [FREE Full text] [doi: 10.1186/s40359-025-03016-0] [Medline: 40598609]
43. Yang Y, Chen J, Zhuang X. Self-determination theory and the influence of social support, self-regulated learning, and flow experience on student learning engagement in self-directed e-learning. Front Psychol 2025;16:1545980 [FREE Full text] [doi: 10.3389/fpsyg.2025.1545980] [Medline: 40207130]
44. Delia CS. The effect of mLearning on motivation in the continuing professional development of nursing professionals: a self-determination theory perspective. J Nurs Educ Pract 2021;12(3):22. [doi: 10.5430/jnep.v12n3p22]
45. Kellam H, Cook C, Smith DL, Haight P. The virtual community of practice facilitation model: a conceptual framework for healthcare professional education. Int J Technol Hum Interact 2023;19(1):1-14. [doi: 10.4018/ijthi.328578]
46. Barnett S, Jones SC, Caton T, Iverson D, Bennett S, Robinson L. Implementing a virtual community of practice for family physician training: a mixed-methods case study. J Med Internet Res 2014;16(3):e83 [FREE Full text] [doi: 10.2196/jmir.3083] [Medline: 24622292]
47. Mairs K, McNeil H, McLeod J, Prorok JC, Stolee P. Online strategies to facilitate health - related knowledge transfer: a systematic search and review. Health Info Libraries J 2013;30(4):261-277 [FREE Full text] [doi: 10.1111/hir.12048]
48. Mukhalalati BA, Taylor A. Adult learning theories in context: a quick guide for healthcare professional educators. J Med Educ Curric Dev 2019;6:2382120519840332 [FREE Full text] [doi: 10.1177/2382120519840332] [Medline: 31008257]
49. Shank E, Tang H, Morris W. Motivation in online course design using self-determination theory: an action research study in a secondary mathematics course. Education Tech Research Dev 2024;73(1):415-441. [doi: 10.1007/s11423-024-10410-9]
50. Lave J, Wenger E. Situated Learning: Legitimate Peripheral Participation. New York, NY, US: Cambridge University Press; 1991:138.
51. Clark E, Draper J, Rogers J. Illuminating the process: enhancing the impact of continuing professional education on practice. Nurse Educ Today 2015;35(2):388-394. [doi: 10.1016/j.nedt.2014.10.014] [Medline: 25467716]
52. Vasli P, Dehghan-Nayeri N, Khosravi L. Factors affecting knowledge transfer from continuing professional education to clinical practice: development and psychometric properties of a new instrument. Nurse Educ Pract 2018;28:189-195. [doi: 10.1016/j.nepr.2017.10.032] [Medline: 29112906]
53. Liu HY. The moderating role of team conflict on teams of nursing students. Int J Environ Res Public Health 2022;19(7):4152 [FREE Full text] [doi: 10.3390/ijerph19074152] [Medline: 35409835]

Abbreviations

CFC: Community Fracture Capture
CoP: communities of practice
CPD: continuing professional development
DXA: dual-energy X-ray absorptiometry
IPC: interprofessional collaboration
IPE: interprofessional education
PCP: primary care physician
SDT: self-determination theory
VCoP: virtual communities of practice

Edited by A Stone; submitted 08.Sep.2025; peer-reviewed by O Enaworu, I Wilson; comments to author 29.Oct.2025; revised version received 13.Dec.2025; accepted 16.Jan.2026; published 04.Mar.2026.

Please cite as:
Fathalla AM, Chang S, Audehm R, Chiang C, Gorelik A, Yates CJ, Snow S, Barmanray RD, Price S, Collins L, Wark JD
Moderators' Role in Fostering Health Practitioner Engagement Within an Online Active Learning Program (The Community Fracture Capture Learning Hub): Qualitative Relational Content Thematic Analysis
J Med Internet Res 2026;28:e83764
URL: https://www.jmir.org/2026/1/e83764
doi:10.2196/83764
PMID:

XSL•FO
RenderX

©Ahmed M Fathalla, Shanton Chang, Ralph Audehm, Cherie Chiang, Alexandra Gorelik, Christopher J Yates, Steve Snow, Rahul D Barmanray, Sarah Price, Lucy Collins, John D Wark. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quantifying Innovation in Stroke: Large Language Model Bibliometric Analysis

Adam Marcus[1,2], BSc, MSc, MBBS, PhD; Georgina Lockwood-Taylor[3], MSc; Daniel Rueckert[1,4], PhD; Paul Bentley[2], PhD, MRCP

[1]Department of Computing, Imperial College London, London, United Kingdom

[2]Department of Medicine, Imperial College London, London, United Kingdom

[3]Department of Psychology and Neuroscience, King's College London, London, United Kingdom

[4]Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

Corresponding Author:
Adam Marcus, BSc, MSc, MBBS, PhD
Department of Computing
Imperial College London
Exhibition Rd, South Kensington
London, SW7 2AZ
United Kingdom
Phone: 44 020 7589 5111
Email: adam.marcus11@imperial.ac.uk

Abstract

Background: Thrombolysis and mechanical thrombectomy represent the most successful stroke innovations over the last 30 years. Quantifying innovation in stroke is essential for identifying productive research lines and prioritizing funding, but health care lacks validated methods for measuring innovation.

Objective: This study aimed to systematically evaluate the relationship between stroke-related patents and publications, demonstrate the feasibility of using large language models (LLMs) in this process, and identify the most rapidly advancing innovations in stroke care by mapping them to a theoretical innovation life cycle.

Methods: The Open Patent Services (European Patent Office) and PubMed databases were searched between 1993 and 2023 for "stroke OR cerebrovascular." In this bibliometric patent-publication analysis, a 13 billion–parameter Llama LLM was trained to identify patents related to stroke disease, as opposed to other references to the word "stroke," on a manually labeled subset of 5000 patents and assessed using 5-fold cross-validation. The LLM filtered irrelevant results, and the resulting patent codes were grouped into innovation clusters. For each cluster, annual patent and publication counts were normalized to adjust for global trends. Cluster-specific growth curves were plotted to analyze the rates and characteristics of growth. The innovation life cycle stage for each innovation cluster was estimated by fitting a sigmoid curve to the patent and publication data consistent with the diffusion of innovations theory by Rogers.

Results: The cross-validated accuracy of the LLM was 99.2%, with a sensitivity of 96.5% and a specificity of 99.6%. An initial bibliometric search retrieved 237,035 patents and 486,664 research publications. A manual review of a random sample of patents before filtering revealed that only 11.2% (56/500) were relevant to stroke. After LLM filtering, of the 237,035 patents, 28,225 (11.9%) stroke-related patents remained. These were grouped into 7 innovation clusters: pharmacological treatment, alternative medicine, rehabilitation devices, medical imaging, diagnostic testing, surgical devices, and artificial intelligence (AI) methods. Patent and publication counts were strongly correlated across clusters (Spearman r_s=0.65-0.92; P<.006) except for pharmacological treatment (r_s=0.09) and alternative medicine (r_s=0.55). Pharmacological treatments were the top-performing cluster over the last 30 years, accounting for 49.3% (36,005/73,094) of all patents, but patent activity in this area has plateaued since the late 2000s. AI methods, rehabilitation devices, and medical imaging exhibited exponential rates of patent growth, with annual normalized increases of 39.2%, 15.9%, and 5.8% compared to 16.9%, 5.3%, and 2.2% for publications, respectively.

Conclusions: Applying an LLM to publicly available patent and publication data provides a scalable way to quantify innovation in stroke. Pharmacological treatment appears to have entered a saturation phase, whereas AI methods, rehabilitation devices, and medical imaging remain in rapid growth, highlighting areas of greatest traction for future research and investment.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e70754) doi:10.2196/70754

KEYWORDS

diffusion of innovation; innovation; stroke; large language model; artificial intelligence; AI

Introduction

Turning points in stroke treatment occurred in 1995 and 2015 with the introduction of thrombolysis and mechanical thrombectomy, respectively [1,2]. Such shifts are rare and, in a more formal context, can be considered innovations: a term defined as a process that ushers in new technologies or techniques that induce a substantial change in practice [3,4]. Quantifying innovation in stroke care is vital as it helps discern which lines of research are productive, such as revascularization therapies, perfusion imaging, and decompressive surgery, as opposed to those that have been less successful, such as neuroprotective and neurorestorative therapies. Measures of innovation output aid in research strategy planning, prioritizing, and assessing the effectiveness of research funding. However, while the study of innovation is well established in other fields [5], health care has relatively few validated methods for quantifying innovation outputs, which can limit progress [6].

Prior evaluations of innovation in stroke have been limited, consisting largely of qualitative reviews [7-9] or conventional bibliometric analyses that track academic trends [10]. While valuable for tracking academic discourse and research activity, such bibliometric approaches have inherent limitations for measuring tangible innovation. Metrics based on citations and publications tend to reflect academic impact over practical implementation, and they do not readily distinguish incremental advances from transformative breakthroughs. Consequently, these methods primarily measure research inputs and academic outputs, not the development of novel, practical solutions.

An alternative approach, originally applied to surgery [4], leverages original patents as a benchmark of technological innovation by comparing the cumulative quantity of patents for a specific innovation with that of related peer-reviewed publications. This method has the advantage of drawing on a comprehensive repository of inventions that have been independently evaluated for novelty and utility; these are generally sufficiently mature and practical to have attracted the funding resources required for patent filing. However, patent analysis relies heavily on the precise interpretation and determination of relevant patents, a task complicated by the broad and often ambiguous language of patent documents, making it labor-intensive and time-consuming. The difficulty of this task is particularly magnified in the context of stroke-related patents given that the term “stroke” could denote a disease as well as multiple engineering concepts and mechanical action of engines, clocks, and other mechanisms. Conventional search engines such as Google are liable to confound stroke terms as they lack the specialized filtering and context awareness needed for precise medical searches.

In this regard, recent advancements in artificial intelligence (AI), specifically large language models (LLMs), hold tremendous potential. They have exhibited remarkable proficiency in textual tasks, making them invaluable in this context [11]. Essentially, LLMs are statistical models trained on vast datasets enabling them to learn intricate relationships between words and phrases. While training LLMs from scratch might pose considerable difficulties and financial burden, the recent availability of open-source trained LLMs to the public has mitigated these challenges [12].

Therefore, the aims of this study were 3-fold: first, to evaluate systematically the relationship between stroke-related patents and publications over the last 3 decades; second, to demonstrate the feasibility of using LLMs to assist in this process; and, finally, to identify the most rapidly advancing innovations in stroke care.

Methods

Although this was not a clinical study, the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [13] were adhered to where appropriate. The methodology was based on the work by Hughes-Hallett et al [4], with adaptions for stroke-related innovation.

Data Collection

The Open Patent Services web service, provided by the European Patent Office [14], was used to obtain patent application data from more than 80 different countries. The period from 1993 to 2023 was chosen to capture the modern era of stroke care, beginning shortly before the pivotal 1995 National Institute of Neurological Disorders and Stroke trial that established thrombolysis as a standard treatment [1]. Patents filed between 1993 and 2023 were downloaded if either their title or abstract matched the following Boolean search: “stroke OR cerebrovascular.” A PubMed (National Library of Medicine) search was also conducted using the same strategy to extract publication data for the same period.

Data Filtering

The collected patents were randomly sampled, and a subset of 5000 was manually annotated by a single author (AM), who is a medical professional, as either related or unrelated to stroke. A second, random unfiltered sample of 500 patents and 500 publications matching the search terms “stroke” or “cerebrovascular” was also manually labeled by the same annotator (AM). Annotations from the 5000-patent subset were used to provide ground truth for model fine-tuning, whereas the second sample was used to verify the accuracy of the search strategy, especially for PubMed results. Patents were considered stroke related if their primary content was directly relevant to stroke management or pathophysiology. Publications were considered stroke related if they contributed to the understanding of stroke, including basic science suitable for stroke journals and clinical studies with stroke as an end point.

XSL•FO
RenderX

A 13 billion–parameter Llama model (Meta AI) [12], a state-of-the-art open-source LLM trained on web data, was then fine-tuned using low-rank adaptation (LoRA) [15] to classify stroke-related patents. Fine-tuning was performed using PyTorch (version 1.13; Meta AI) on a machine equipped with a 2.80-GHz AMD EPYC 7543P central processing unit and an NVIDIA RTX A4500 20-GB graphics processing unit. The hyperparameters, which are listed in Table 1 and were based on standard values from the original LoRA paper [15], were chosen to ensure stable training while preventing overfitting. In particular, the LoRA rank was set to 8 to keep model adaptation minimal and efficient. The prompt template is provided in Figure 1. To assess prompt sensitivity, alternative phrasings were trialed on a development subset during fine-tuning. The model's performance was evaluated using 5-fold cross-validated binary classification metrics and compared to that of the base model before the final model was used to analyze the unlabeled patents and filter out those unrelated to stroke. The model was not used for classifying publications.

Table 1. The hyperparameters used for fine-tuning the Llama model using low-rank adaptation (LoRA).

Parameter	Value
Batch size	128
Number of epochs	10
Learning rate	0.0003
Optimizer	Adam
Maximum gradient norm	1
LoRA rank	8
LoRA alpha	16
LoRA dropout	0.05
LoRA target modules	Query and value projection

Figure 1. The prompt template used to fine-tune the Llama model to classify whether a patent was related to stroke.

```
Below is a patent application title, paired with an abstract. Write a
yes/no response whether the patent is related to stroke disease.

### Title:
[TITLE]

### Abstract:
[ABSTRACT]

### Response:
<COMPLETE>
```

Data Normalization

Across all fields, the number of patents and publications has risen exponentially. To adjust for this increase, both counts were normalized using the formula outlined by Hughes-Hallett et al [4]:



In this formula, II_i denotes the innovation index, defined as the number of patents or publications within a particular field; C_i is the innovation constant; and t_i is the total number of patents granted or publications indexed on PubMed for a given year i. For example, if a field had 50 patents in a year when 100,000 patents were granted and the maximum number of patents in any year during the study period was 200,000, then C_i = 100,000/200,000 = 0.5, and the normalized innovation index would be 50/0.5 = 100.

Identifying Innovation Clusters

The process of identifying innovation clusters involved a 2-stage method to ensure comprehensive coverage. Initially, the top 100 most frequent International Patent Classification (IPC) [16] codes from the filtered, stroke-related patent dataset were extracted. These codes, assigned by patent examiners, offer a standardized way of categorizing the technological domains of inventions. Focusing on the top 100 codes provided a quantitative starting point, representing the most concentrated areas of patent activity while avoiding the sparsity and noise of the long tail of less frequent codes.

These top 100 codes were then manually grouped into preliminary innovation clusters by 2 authors (AM and GL-T).

XSL•FO
RenderX

This grouping was based on the descriptive content of the IPC codes and their relevance to distinct areas of stroke care. Interrater reliability was assessed using the Cohen κ, and any disagreements regarding this grouping were resolved by a third author (PB).

To capture relevant patents and publications that may not have fallen into these top 100 codes, a second stage was implemented. Expanded, cluster-specific Boolean search strategies were developed as listed in Table 2 and performed on both the patent database and PubMed. The keywords for these searches were also determined by 2 authors (AM and GL-T), with a third author (PB) resolving any disagreements. For clusters such as alternative medicine, broader search terms such as "food" and "herbal" were intentionally used. This was necessary to capture innovations described in nonclinical or lay terms, which is common in patent applications for complementary therapies that may lack standardized medical terminology. The final dataset for each cluster comprised all documents identified through either the initial IPC code grouping or the subsequent expanded Boolean search. Finally, this entire 2-step methodology was repeated for patents and publications from the last decade (2013-2023) to allow for the determination of more recent innovations.

Table 2. PubMed and European Patent Office database search strategies.

Innovation cluster	Search strategy
AI[a] methods	(AI OR "artificial intelligence" OR "deep learning" OR "machine learning" OR "neural network") AND (stroke OR cerebrovascular)
Alternative medicine	(food OR tea OR coffee OR beverage OR herbal OR acupuncture OR aromatherapy OR reflexology OR "holistic therapy") AND (stroke OR cerebrovascular)
Diagnostic testing	("diagnostic testing" OR "diagnostic tools" OR "clinical tests" OR "screening tests" OR "blood tests" OR "laboratory tests" OR "genetic testing") AND (stroke OR cerebrovascular)
Medical imaging	(imaging OR angiography OR angiogram OR ultrasound OR CT OR MRI OR PET OR "computed tomography" OR "magnetic resonance" OR "positron emission tomography") AND (stroke OR cerebrovascular)
Pharmacological treatment	(thrombolysis OR aspirin OR clopidogrel OR warfarin OR DOACs OR alteplase OR tPA OR "tissue plasminogen activator" OR "thrombolytic therapy" OR "pharmacological treatment" OR "pharmaceutical composition" OR "drug therapy" OR "secondary prevention" OR "direct oral anticoagulants") AND (stroke OR cerebrovascular)
Rehabilitation devices	(rehabilitation OR neurorehabilitation OR exoskeleton OR "training device" OR "brain-computer interface") AND (stroke OR cerebrovascular)
Surgical devices	(thrombectomy OR embolectomy OR "clot removal" OR "clot retrieval" OR "catheter device" OR "surgical device" OR "endovascular treatment" OR "endovascular therapy") AND (stroke OR cerebrovascular)

[a]AI: artificial intelligence.

Statistical Analysis

All statistical analyses were performed using Python (version 3.11.3; Python Software Foundation) and the *statsmodels* [17] package. Permutation testing with Bonferroni correction was used to calculate *P* values adjusted for multiple tests. The relationship between patent and publication data was visualized using scatterplots to assess the nature of the association. On the basis of this visual inspection, an appropriate correlation coefficient was selected: Pearson (r) for linear relationships and Spearman rank (r_s), a nonparametric method, for monotonic but nonlinear relationships. To model the technology diffusion life cycle, innovation life cycle progression was derived by fitting sigmoid curves to the patent and publication data, a method consistent with the diffusion of innovations theory by Rogers [18]. To quantify the uncertainty in these estimates, 95% CIs were calculated using nonparametric bootstrapping, which involved repeatedly resampling the data and refitting the curve.

Results

Data and Filtering Performance

The initial search retrieved 237,035 patents and 486,664 publications. In a random, unfiltered sample of 500 patents and 500 publications matching the search terms "stroke" or "cerebrovascular," 11.2% (56/500) of patents and 74.2% (371/500) of publications were stroke related. The remaining patents typically referred to "stroke" in nonclinical contexts, including mechanical travel (eg, stroke length in pistons), engine cycles (eg, 2-stroke engines), lightning discharges, and line rendering in handwriting or graphics. These examples formed the negative class for model fine-tuning. An LLM was then fine-tuned to classify whether a patent was stroke related, achieving a cross-validated accuracy of 99.2% with a sensitivity of 96.5% and specificity of 99.6% and significantly outperforming the base model across all metrics listed in Table 3 (all $P<.001$). The receiver operating characteristic curve is shown in Figure 2 [19]. Prompt sensitivity was also evaluated during model fine-tuning using alternative phrasings; this had a negligible impact on classification outcomes. After filtering using the model, of the 237,035 patents, 28,225 (11.9%) stroke-related patents remained. Figure 3 illustrates the annual

XSL•FO
RenderX

increase in these patents filed by geographic region, with the largest proportion originating from China. The original and normalized counts of patents and publications related to stroke are shown in Figure 4. Normalized patent counts reached a peak in 2010, whereas normalized publication activity continues to grow.

Table 3. Five-fold cross-validated performance of the fine-tuned Llama model compared to the base model for classifying whether a patent was stroke related. *P* values are from 2-tailed paired t tests across the 5 folds.

Measure	Fine-tuned estimate (95% CI)	Base estimate (95% CI)	*P* value
AUROC[a]	0.990 (0.988-0.992)	0.654 (0.605-0.703)	<.001
Accuracy (%)	99.2 (99.0-99.5)	66.8 (66.3-67.3)	<.001
Sensitivity (%)	96.5 (94.8-98.2)	54.5 (48.3-60.7)	<.001
Specificity (%)	99.6 (99.4-99.7)	68.2 (67.6-68.7)	<.001
PPV[b] (%)	96.1 (94.8-97.4)	16.5 (13.9-19.1)	<.001
NPV[c] (%)	99.6 (99.4-99.8)	92.9 (92.2-93.7)	<.001
F_1-score (%)	96.3 (95.1-97.5)	25.2 (21.6-28.9)	<.001

[a]AUROC: area under the receiver operating characteristic curve.

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

Figure 2. Receiver operating characteristic curve illustrating the cross-validated performance of the fine-tuned Llama model compared to the base Llama model in classifying stroke-related patents. The shaded area represents the 95% confidence region determined via the fixed-width band technique [19]. AUC: area under the curve.

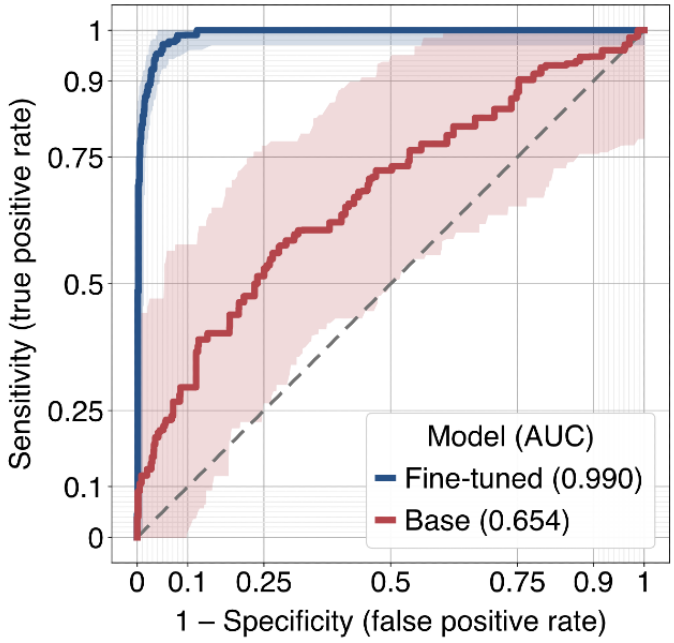

XSL•FO
RenderX

Figure 3. Stroke-related patents filed between 1993 and 2023 categorized by geographic coverage as defined by patent filing jurisdiction and route (China: China National Intellectual Property Administration; United States: US Patent and Trademark Office; Japan: Japan Patent Office; South Korea: Korean Intellectual Property Office; Canada: Canadian Intellectual Property Office; Europe: European Patent Office; worldwide: Patent Cooperation Treaty or World Intellectual Property Organization; other regions: other national or regional offices).

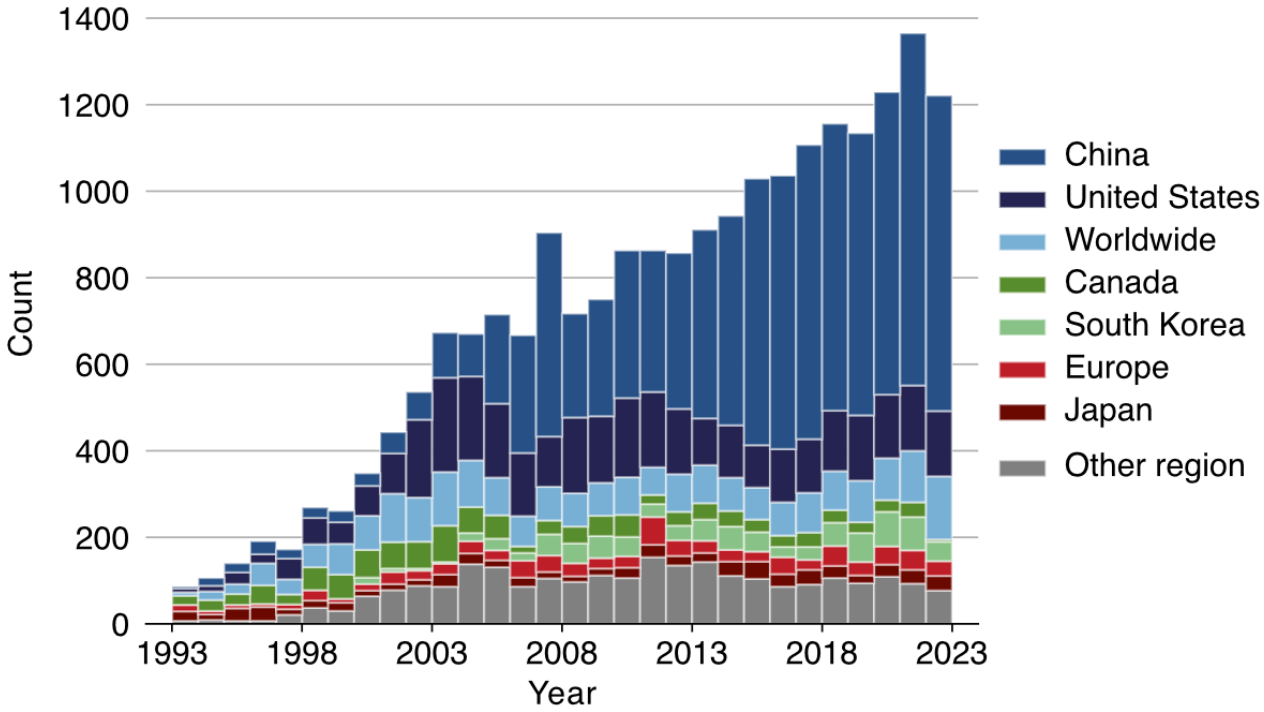

XSL•FO
RenderX

Figure 4. Overview of the (A) counts of patents and publications related to stroke over time and (B) year-on-year normalized patent and publication counts for each innovation cluster. Normalized counts were calculated by dividing the annual number of stroke-related patents or publications by the total number of patents or publications in that year, scaled to the maximum annual total across the study period. In both cases, the dashed lines depict exponential fits with the associated coefficients of determination (R^2) given per plot. AI: artificial intelligence.

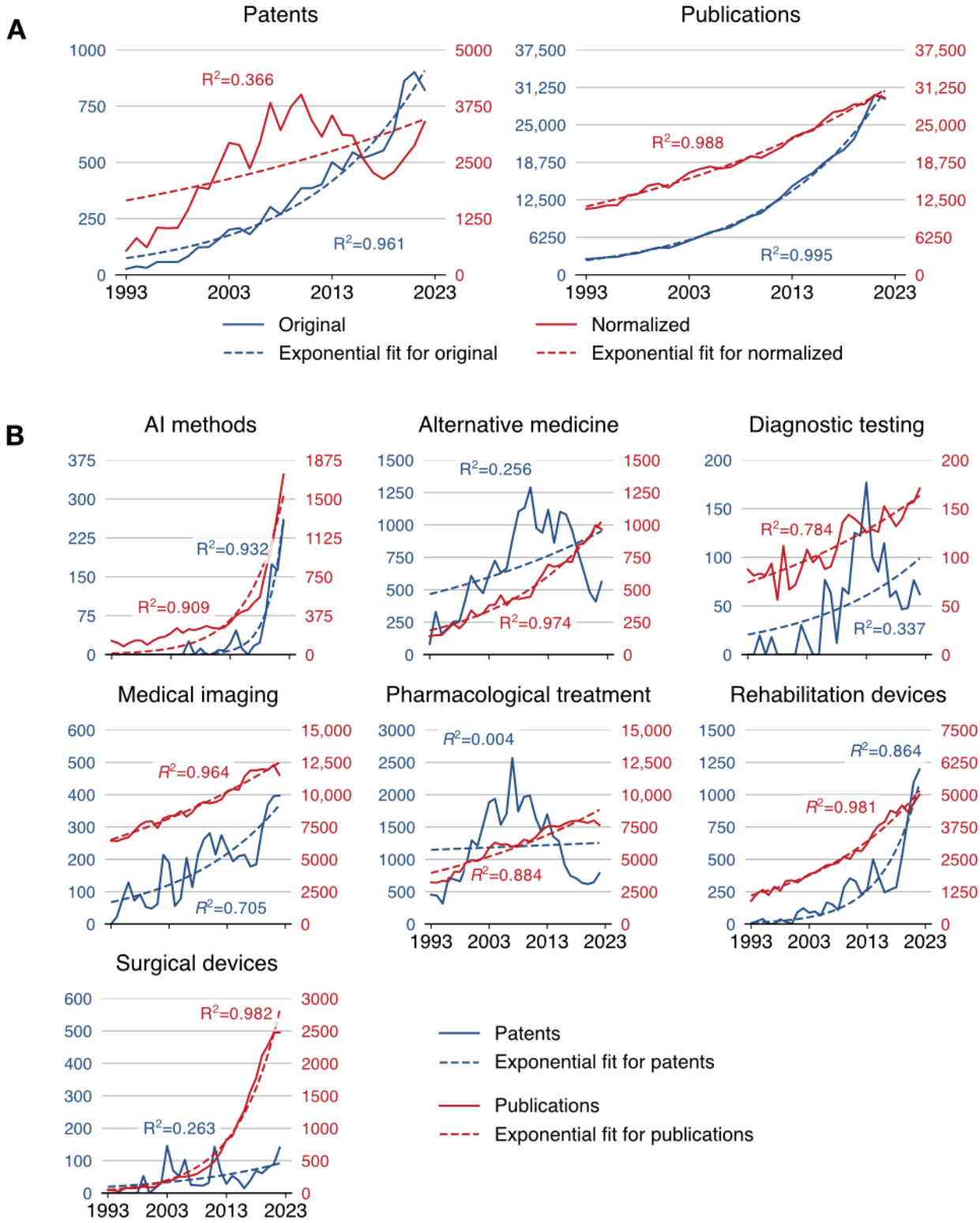

Leading Innovation Clusters

There were 7 top stroke-related innovation clusters identified over the last 30 years; interrater reliability between the 2 authors was high (Cohen κ=0.871). To address potential overcapture in the alternative medicine cluster due to broad search terms, a sensitivity analysis of 100 randomly sampled patents was performed, verifying that 97% were stroke related. The performance of these clusters, as measured using patents, is summarized in Table 4, with the allocation of patent codes provided in Multimedia Appendix 1. Pharmacological treatment was the largest cluster, accounting for 49.3% (36,005/73,094)

XSL•FO
RenderX

of the patents filed over the study period. To ensure the stability of these findings, a sensitivity analysis was performed confirming that cluster rankings held without normalization (data not shown). Within the last decade, only AI methods increased in rank, with the relative ordering of the other top-performing clusters remaining constant. Although not reflected in a change in order, the proportion accounted for by pharmacological treatments fell to 33.7% (9391/27,870), whereas all the other clusters increased their shares.

Table 4. Comparing the top-performing stroke-related innovation clusters by cumulative normalized patent counts over the past 30 years and the last decade. Artificial intelligence (AI) methods were the only cluster to increase in rank.

Rank	Innovation cluster	Normalized patent count, n (%)
1993-2023 (n=73,094)		
1	Pharmacological treatment	36,005 (49.3)
2	Alternative medicine	20,291 (27.8)
3	Rehabilitation devices	7777 (10.6)
4	Medical imaging	5331 (7.3)
5	Diagnostic testing	1448 (2.0)
6	Surgical devices	1397 (1.9)
7	AI methods	845 (1.2)
2013-2023 (n=27,870)		
1	Pharmacological treatment	9391 (33.7)
2	Alternative medicine	8005 (28.7)
3	Rehabilitation devices	5569 (20.0)
4	Medical imaging	2668 (9.6)
5	AI methods	791 (2.8)
6	Diagnostic testing	834 (3.0)
7	Surgical devices	612 (2.2)

Statistical Analysis

Figure 4 and Table 5 show the relationship between normalized patent and publication counts over time for the top-performing innovation clusters. There were strong associations (r_s>0.6; P<.006) between patent and publication rates for all clusters except pharmacological treatment (r_s=0.094; P=.99) and alternative medicine (r_s=0.546; P=.01). These 2 clusters showed normalized patents peaking in the late 2000s, with a continued shallow rise in publications. This trend is further illustrated by the normalized patent-to-publication ratio over time for each innovation cluster, as detailed in Multimedia Appendix 2. Plots of the data on AI methods show a rapid rise similar to that of other emerging clusters. To quantify this concurrent growth, the temporal correlation between the AI cluster and other leading clusters was assessed. AI patent activity was strongly correlated with patent activity in rehabilitation devices (r_s=0.767; P=.004) and medical imaging (r_s=0.662; P=.004). Similarly, AI publication rates correlated strongly with publication rates in rehabilitation devices (r_s=0.963; P=.004) and medical imaging (r_s=0.962; P=.004). These clusters all exhibited a strong exponential fit for both patent and publication data supported by coefficient of determination (R^2) values exceeding 0.7, as shown in Table 5. The rates of exponential growth for AI methods were the highest (39.2% per year for patents; 16.9% per year for publications), followed by rehabilitation devices (15.9% per year for patents; 5.3% per year for publications) and medical imaging (5.8% per year for patents; 2.2% per year for publications). The diffusion dynamics for each innovation cluster, approximated by fitting sigmoid curves to the patent and publication data, are shown in Figure 5 [18] and contextualized within the phases of the diffusion of innovations theory by Rogers [18], highlighting their positions in the innovation life cycle. The estimated progression through the innovation life cycle based on patent data was highest for pharmacological treatment (97.5%), followed by surgical devices (82.9%), nutritional and complementary therapies (71.4%), diagnostic testing (52.2%), AI methods (41.6%), medical imaging (30.5%), and rehabilitation devices (16.5%).

XSL•FO
RenderX

Table 5. Comparing the association between the normalized patent and publication counts for each innovation cluster along with the equations of the associated lines of the best exponential fit.

Innovation cluster	Patents		Publications		Association	
	Equation for the line with the best exponential fit	R^2	Equation for the line with the best exponential fit	R^2	r_s	*P* value
AI[a] methods	$3.7E\text{-}343e^{0.392x}$	0.932	$2.7E\text{-}146e^{0.169x}$	0.909	0.776	.001
Alternative medicine	$7.3E\text{-}20e^{0.025x}$	0.256	$2.2E\text{-}49e^{0.059x}$	0.974	0.546	.02
Diagnostic testing	$9.1E\text{-}47e^{0.054x}$	0.337	$1.9E\text{-}22e^{0.027x}$	0.784	0.645	.006
Medical imaging	$5.8E\text{-}50e^{0.058x}$	0.705	$2.6E\text{-}16e^{0.022x}$	0.964	0.767	.001
Pharmacological treatment	$8.5E\text{-}1e^{0.003x}$	0.004	$3.8E\text{-}21e^{0.028x}$	0.884	0.094	>.99
Rehabilitation devices	$9.1E\text{-}138e^{0.159x}$	0.864	$1.6E\text{-}43e^{0.053x}$	0.981	0.916	.001
Surgical devices	$5.2E\text{-}46e^{0.053x}$	0.263	$3.8E\text{-}120e^{0.140x}$	0.982	0.649	.001

[a]AI: artificial intelligence.

Figure 5. (A) The approximate innovation life cycle progression of each innovation cluster, calculated by fitting a sigmoid curve to patent and publication data, with 95% CIs estimated via nonparametric bootstrapping by repeatedly resampling the data and refitting the curve, which indicates their phase in the diffusion of innovation curve, and (B) the theoretical diffusion of innovation curve displaying both the cumulative diffusion (S-shaped curve) and rate of diffusion (bell-shaped curve) over time divided into the incubation (innovators and early adopters), growth (early majority and late majority), and saturation (laggards) phases. AI: artificial intelligence.

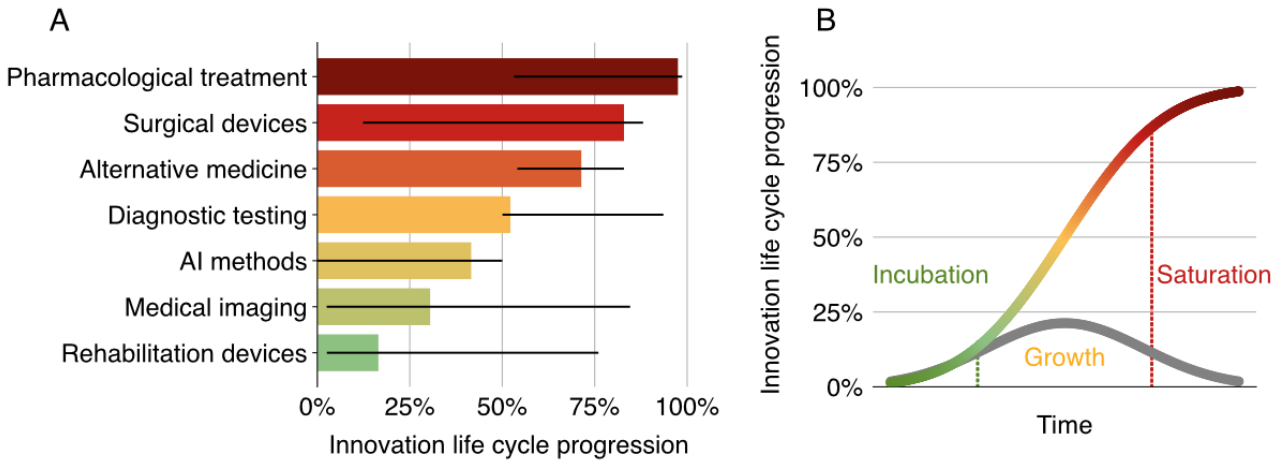

Discussion

Principal Findings

To the best of the authors' knowledge, this study provides the first rigorous quantification of innovation in stroke management. By using a scientifically validated framework, along with a novel application of an LLM, publicly available patent and publication data were analyzed. Among the top stroke-related innovation clusters identified, pharmacological treatment was found to be the most dominant over the past 30 years, accounting for nearly half (36,005/73,094, 49.3%) of all patents filed. AI methods, followed by rehabilitation devices and medical imaging, exhibited the highest rates of patent and publication growth, suggesting that these emerging innovation clusters are taking on increasingly important roles.

The diffusion of innovations theory by Rogers [18] describes the adoption curve of technology as a sigmoid function. This arises from the natural variation in the attitudes of individuals, ranging from early adopters to laggards, toward a new innovation. Previous studies have shown that this curve can also be applied to specific clusters themselves [4,20], indicating different phases of innovation: the incubation phase, where seminal work occurs and which is reflected by the initial rise in patenting and publication activity; the growth phase, where industry and clinicians drive innovation and which is associated with an exponential rise in patent and publication counts; and, finally, the saturation phase, which occurs when manufacturers refine the technology to maintain their competitive edge while continuing to pursue patents, leading to a plateau in both patent and publication activity. While these curves provide a model for technology diffusion, we acknowledge that this is a simplification. Real-world adoption in health care is complex and is further influenced by external factors such as regulatory approvals, reimbursement policies, and the development of clinical guidelines, which were not modeled in this study.

Applying this theory to this study, there was an exponential rise in the number of patents and publications for AI methods. This

XSL•FO
RenderX

study's novel quantification of this trend demonstrates that the patent growth rate for AI methods is approximately 2 to 7 times greater than that for the other key growth clusters of rehabilitation devices and medical imaging. This suggests that these clusters are all in the growth phase but that AI is accelerating at a substantially faster rate. There was a significant inflection point for AI methods observed in 2018, coinciding with the regulatory approval of the first AI software developed by Viz.ai [21]. Subsequently, the market has experienced a proliferation of commercially available software platforms designed to interpret and triage radiological data [22]. Therefore, the concurrent rise in medical imaging is to be expected given that these AI platforms, and indeed the main applications of AI in stroke [23-25], relate to analysis and interpretation of medical images. Similarly, while rehabilitation systems for stroke have yet to receive the same level of attention, there has been growing interest in full automation, with AI methods actively being researched [26].

The pattern for surgical devices presents a less clear trajectory: patent counts are leveling off, whereas publications continue to increase. This divergence could suggest that surgical device innovation remains at a nascent, exploratory stage: scholarly output continues to rise, whereas commercial patenting momentum has yet to follow. Alternatively, this pattern may relate to a limitation of the innovation discovery method used in this study in that patents for generic technologies that do not explicitly state stroke management may be incorrectly excluded. Thus, many patents for mechanical thrombectomy, being applicable to multiple diseases, may not be seen to parallel the rise in stroke thrombectomy publications.

Pharmacological treatments, alternative medicine, and diagnostic testing all experienced peaks in patent activity during the 2000s and have since plateaued, indicating that these sectors are now in the saturation phase. This finding for pharmacological treatments is particularly notable and perhaps unexpected given recent high-profile trials of novel drug classes. The rise in pharmacological treatments can be traced back to 1995 following the groundbreaking National Institute of Neurological Disorders and Stroke tissue-type plasminogen activator trial [1]. However, despite extensive efforts, the development of new therapeutics has become more challenging [27], and thus, a relative decrease in the number of patents and publications is perhaps unsurprising. Notably, this decline coincided with a decoupling between research output and patent activity (r_s=0.09; P=.99), suggesting that ongoing research in pharmacology increasingly focuses on avenues less amenable to patenting. Similarly, alternative medicine, despite seeing increased enthusiasm, particularly in addressing poststroke depression [28], is a well-established field whose roots predate orthodox medicine [29]. Diagnostic testing is also mature, with many of the key technologies, such as electrocardiogram and blood pressure monitors, able to be traced back to the 19th century [30,31]. As such, the observed trends appear in line with expectations.

The varying patent coefficient of determination (R^2) values across the innovation clusters provide insights into the reliability of their respective innovation trajectories. For AI methods (R^2=0.932), rehabilitation devices (R^2=0.864), and medical imaging (R^2=0.705), the high values signify a consistent and predictable exponential growth, reinforcing the conclusion that these are emerging technologies in a strong growth phase. In contrast, the lower values for patent trends in pharmacological treatments (R^2=0.004) and alternative medicine (R^2=0.256) suggest that an exponential model is less descriptive. This lower reliability is not a limitation of the data but rather an indicator that these fields have likely reached a saturation phase in which innovation, as measured via patent filings, is no longer accelerating at a consistent exponential rate.

The observed rise in AI-driven platforms, many of which operate on medical imaging data, has been linked to faster treatment times in acute stroke care. For example, a single-center study reported that the introduction of the Viz.ai software, which uses automated image interpretation to triage stroke cases, was associated with a 30-minute reduction in median door-to-needle time, alongside improvements in door-to-imaging and door-to-puncture intervals [32]. Similarly, innovations in rehabilitation devices, including robotic-assisted gait and upper-limb training, have demonstrated clinically meaningful improvements in motor function and activities of daily living in randomized controlled trials [33]. This demonstrated clinical value arguably provides a powerful mechanism that drives the investment, research, and patenting activity observed in this study. Taken together, these examples suggest that the observed growth in patents and publications within emerging innovation clusters reflects not only conceptual progress but also improvements in patient care and clinical workflows.

Comparison With Other Studies

There has been limited prior work evaluating innovation in stroke, with those studies focusing only on specific areas and generally being qualitative [7-9]. Martinez-Gutierrez et al [7] conducted a narrative review of developing technologies in the prehospital space in which they identified emergency medical service detection and triage of stroke as important areas for future advancements. The study also supported the potential of AI algorithms in these domains, aligning with the findings of our study. Recently, Ji et al [10] conducted a bibliometric analysis using Web of Science specifically for perioperative stroke over the past 20 years and found rapid growth in research publications addressing antiplatelet and antithrombotic therapy, cardiovascular surgery, and thrombectomy, among others. One notable difference between this study's results and those of our study is that the former identified pharmacological treatments and surgical devices as leading research areas. However, this discrepancy may be due to their use of absolute rather than normalized publication counts and their restriction to a narrow type of stroke as opposed to all causes, as in this study.

Within the field of health care more generally, Tran et al [34] performed a survey of health AI publications using Web of Science between 1977 and 2018 and found stroke to be a leading application area. The approach used in the aforementioned study has also been applied to neurosurgery [16] and surgery [4] as a whole. In both instances, and in keeping with the work presented in this paper, trends in patents and publications were consistent with the diffusion of innovations theory.

XSL•FO
RenderX

Strengths and Limitations

A key strength of this study lies in its use of a data-driven and quantitative framework to evaluate innovation in stroke management. This approach moves beyond traditional qualitative analyses by incorporating an LLM. This was critical for enabling the study's scale, as the manual filtering of 237,035 patents, of which a sample review found only 11.2% (56/500) to be relevant, would have been prohibitively labor-intensive. By enabling a detailed and extensive search across the breadth of stroke research fields, this study comprehensively quantified the current popularity and productiveness of innovation clusters and, by plotting changes in these metrics over time, can estimate where along the trajectory of innovation diffusion each cluster currently lies. Our results identify emerging technologies and may be useful metrics to inform policy and grant funding strategies. Our results also build on previous work [4,16] that underscores the value of using patent and publication data in the assessment of innovation. Despite their value, patent data have remained largely underused and underinvestigated [4].

Although this study used a novel approach to quantitatively evaluate innovation in stroke management, it is not without limitations. First, the methodology relied on patents as an indicator of technological innovation, potentially overlooking the output from individuals or organizations who lack the resources to apply for patents or choose not to for ethical or other reasons. Second, emerging or small-scale innovation clusters were unlikely to be identified through the method used, as they may be concealed within larger, more mature clusters. Third, patents for generic technological innovations that did not explicitly state their application to stroke were excluded from the analysis, even though they could still be applicable to stroke management. Fourth, it is possible that some inventors may deliberately delay academic publication until a patent has been granted, leading to an underestimation of recent innovations. Fifth, bias could be introduced due to the imperfect filtering of patents by the LLM, which was trained on a single-annotated dataset. Such limited annotation may have introduced misclassification errors, potentially overrepresenting clusters that use terminology closely aligned with the training labels while underrepresenting those that rely on novel or nuanced language. This may have affected both the sensitivity and specificity of cluster assignment. Future work building on this methodology should prioritize creating a ground-truth dataset with multiple, independent annotators to validate labels and further improve model generalizability. Sixth, the normalization method based on the work by Hughes-Hallett et al [4] is sensitive to the maximum value within the time series, which may introduce instability in the normalized metrics if future volumes differ substantially. Finally, the findings of this study were not validated against external clinical data. Such a validation would be a valuable next step to confirm whether the identified trends in patenting activity correlate with tangible improvements in stroke care and patient outcomes.

Conclusions

This is the first study to systematically use patent and publication data to quantitatively evaluate innovation in stroke care. Seven influential innovation clusters were identified over the 30-year study period, and their respective growth characteristics were found to be explainable by the diffusion of innovations theory. Looking ahead, the results suggest that AI methods, rehabilitation devices, and medical imaging are undergoing exponential growth and are forecasted to have a greater impact on stroke management. Furthermore, the methodology used in this work, particularly the novel use of an LLM, could be applied to assess more specific clusters and assist in decision-making for future research and funding.

Acknowledgments

The research team gratefully acknowledges the support and resources provided by the BioMedIA laboratory and Imperial College London. Generative artificial intelligence was not used in the creation of the manuscript text, figures, or tables. The large language model (Llama) discussed in this paper was an object of study and a tool for data filtering as described in the Methods section; it was not used to generate any part of this research paper.

Funding

This work was supported by the following grants and programs: the UK Research and Innovation Centre for Doctoral Training in AI for Healthcare [35] under grant EP/S023283/1, The Graham-Dixon Charitable Trust, the UK National Institute for Health and Care Research Invention for Innovation Programme under grant II-LA-0814-20007, he Imperial College Biomedical Research Centre, and the Economic and Social Research Council London Interdisciplinary Social Science Doctoral Training Partnership Collaborative Awards in Science and Engineering studentship.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the terms and conditions of the Open Patent Services web service but are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

AM and PB were responsible for the conceptualization of the study. AM, GL-T, and PB developed the methodology. AM conducted the formal analysis and data curation. AM wrote the original draft of the manuscript. All authors contributed to the writing, review, and editing of the final manuscript. PB and DR provided supervision.

Conflicts of Interest

PB is a cofounder of GripAble Ltd, a spin-out company from Imperial College London. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
The 100 top-performing patent codes retrieved by the search “stroke OR cerebrovascular” between 1993 and 2023 allocated to innovation clusters.
[DOCX File , 32 KB - jmir_v28i1e70754_app1.docx]

Multimedia Appendix 2
Year-on-year normalized patent-to-publication ratio for each innovation cluster.
[DOCX File , 92 KB - jmir_v28i1e70754_app2.docx]

References

1. National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. N Engl J Med 1995 Dec 14;333(24):1581-1587. [doi: 10.1056/NEJM199512143332401] [Medline: 7477192]
2. Goyal M, Menon BK, van Zwam WH, Dippel DW, Mitchell PJ, Demchuk AM, HERMES collaborators. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. Lancet 2016 Apr 23;387(10029):1723-1731. [doi: 10.1016/S0140-6736(16)00163-X] [Medline: 26898852]
3. West MA. The social psychology of innovation in groups. In: West MA, Farr JL, editors. Innovation and Creativity at Work: Psychological and Organizational Strategies. Hoboken, NJ: John Wiley & Sons; 1990:309-333.
4. Hughes-Hallett A, Mayer EK, Marcus HJ, Cundy TP, Pratt PJ, Parston G, et al. Quantifying innovation in surgery. Ann Surg 2014 Aug;260(2):205-211 [FREE Full text] [doi: 10.1097/SLA.0000000000000662] [Medline: 25350647]
5. Daim TU, Rueda G, Martin H, Gerdsri P. Forecasting emerging technologies: use of bibliometrics and patent analysis. Technol Forecast Soc Change 2006 Oct;73(8):981-1012. [doi: 10.1016/j.techfore.2006.04.004]
6. Campbell B. How to judge the value of innovation. BMJ 2012 Mar 07;344(mar07 1):e1457. [doi: 10.1136/bmj.e1457] [Medline: 22399703]
7. Martinez-Gutierrez JC, Chandra RV, Hirsch JA, Leslie-Mazwi T. Technological innovation for prehospital stroke triage: ripe for disruption. J Neurointerv Surg 2019 Nov 14;11(11):1085-1090. [doi: 10.1136/neurintsurg-2019-014902] [Medline: 31201289]
8. Verma A, Towfighi A, Brown A, Abhat A, Casillas A. Moving towards equity with digital health innovations for stroke care. Stroke 2022 Mar;53(3):689-697 [FREE Full text] [doi: 10.1161/STROKEAHA.121.035307] [Medline: 35124973]
9. Filler A. The history, development and impact of computed imaging in neurological diagnosis and neurosurgery: CT, MRI, and DTI. Nat Preced 2009 Jun 30;1:1-76. [doi: 10.1038/npre.2009.3267]
10. Ji S, Shi Y, Fan X, Jiang T, Yang X, Tao T, et al. Global trends in perioperative stroke research from 2003 to 2022: a web of science-based bibliometric and visual analysis. Front Neurol 2023;14:1185326 [FREE Full text] [doi: 10.3389/fneur.2023.1185326] [Medline: 37325224]
11. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: scaling language modeling with pathways. arXiv Preprint posted online April 5, 2022 [FREE Full text] [doi: 10.5555/3648699.3648939]
12. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: open and efficient foundation language models. arXiv Preprint posted online February 27, 2023 [FREE Full text] [doi: 10.48550/arXiv.2302.13971]
13. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
14. Coverage, codes and statistics. European Patent Office. URL: https://www.epo.org/en/searching-for-patents/data/coverage [accessed 2025-05-29]
15. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. arXiv Preprint posted online June 17, 2021 [FREE Full text] [doi: 10.48550/arXiv.2106.09685]
16. International Patent Classification (IPC). World Intellectual Property Organization. URL: https://www.wipo.int/en/web/classification-ipc [accessed 2025-05-29]

XSL•FO
RenderX

17. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with Python. In: Proceedings of the 9th Python in Science Conference. 2010 Presented at: SciPy '10; June 28-July 3, 2010; Austin, TX URL: https://proceedings.scipy.org/articles/Majora-92bf1922-011 [doi: 10.25080/majora-92bf1922-011]
18. Rogers EM. Diffusion of Innovations. 4th edition. New York, NY: Simon and Schuster; 2010.
19. Campbell G. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. Stat Med 1994 Oct 15;13(5-7):499-508. [doi: 10.1002/sim.4780130513] [Medline: 8023031]
20. Marcus HJ, Hughes-Hallett A, Kwasnicki RM, Darzi A, Yang GZ, Nandi D. Technological innovation in neurosurgery: a quantitative study. J Neurosurg 2015 Jul;123(1):174-181 [FREE Full text] [doi: 10.3171/2014.12.JNS141422] [Medline: 25699414]
21. Petrone JJ. FDA approves stroke-detecting AI software. Nat Biotechnol 2018 Apr 05;36(4):290. [doi: 10.1038/nbt0418-290] [Medline: 29621226]
22. Soun J, Chow DS, Nagamine M, Takhtawala RS, Filippi CG, Yu W, et al. Artificial intelligence and acute stroke imaging. AJNR Am J Neuroradiol 2021 Jan;42(1):2-11 [FREE Full text] [doi: 10.3174/ajnr.A6883] [Medline: 33243898]
23. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec 21;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
24. Mouridsen K, Thurner P, Zaharchuk G. Artificial intelligence applications in stroke. Stroke 2020 Aug;51(8):308-331. [doi: 10.1161/strokeaha.119.027479]
25. Lee EJ, Kim YH, Kim N, Kang DW. Deep into the brain: artificial intelligence in stroke imaging. J Stroke 2017 Sep;19(3):277-285 [FREE Full text] [doi: 10.5853/jos.2017.02054] [Medline: 29037014]
26. Rahman S, Sarker S, Haque AK, Uttsha MM, Islam MF, Deb S. AI-driven stroke rehabilitation systems and assessment: a systematic review. IEEE Trans Neural Syst Rehabil Eng 2023;31:192-207. [doi: 10.1109/TNSRE.2022.3219085] [Medline: 36327176]
27. Paul S, Candelario-Jalil E. Emerging neuroprotective strategies for the treatment of ischemic stroke: an overview of clinical and preclinical studies. Exp Neurol 2021 Jan;335:113518 [FREE Full text] [doi: 10.1016/j.expneurol.2020.113518] [Medline: 33144066]
28. Wijeratne T, Sales C, Wijeratne C. A narrative review on the non-pharmacologic interventions in post-stroke depression. Psychol Res Behav Manag 2022 Jul 07;15:1689-1706 [FREE Full text] [doi: 10.2147/PRBM.S310207] [Medline: 35832139]
29. Steyer TE. Complementary and alternative medicine: a primer. Fam Pract Manag 2001 Mar;8(3):37-42 [FREE Full text] [Medline: 11317848]
30. Burch GE, DePasquale NP. A History of Electrocardiography. New York, NY: Norman Publishing; 1990.
31. Booth J. A short history of blood pressure measurement. Proc R Soc Med 1977 Nov;70(11):793-799 [FREE Full text] [doi: 10.1177/003591577707001112] [Medline: 341169]
32. Milfred F, Ami Roy A, Delmor C, Bansal A, Gifford K, Ma K, et al. Abstract TP89: advancing stroke care efficiency: impact of AI and communication tools on patient outcomes. Stroke 2025 Feb;56(Suppl_1):ATP89. [doi: 10.1161/str.56.suppl_1.tp89]
33. Mehrholz J, Kugler J, Pohl M, Elsner B. Electromechanical-assisted training for walking after stroke. Cochrane Database Syst Rev 2025 May 14;5(5):CD006185. [doi: 10.1002/14651858.CD006185.pub6] [Medline: 40365867]
34. Tran BX, Vu GT, Ha GH, Vuong QH, Ho MT, Vuong TT, et al. Global evolution of research in artificial intelligence in health and medicine: a bibliometric study. J Clin Med 2019 Mar 14;8(3):360 [FREE Full text] [doi: 10.3390/jcm8030360] [Medline: 30875745]
35. UKRI Centre for Doctoral Training in AI for Healthcare. AI4Health. URL: https://ai4health.io/ [accessed 2026-01-12]

Abbreviations

AI: artificial intelligence
IPC: International Patent Classification
LLM: large language model
LoRA: low-rank adaptation
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

Edited by A Coristine; submitted 01.Jan.2025; peer-reviewed by N Raju, S Sivarajkumar, C Wang, R Yang; comments to author 01.Apr.2025; revised version received 23.Dec.2025; accepted 23.Dec.2025; published 20.Jan.2026.

Please cite as:
Marcus A, Lockwood-Taylor G, Rueckert D, Bentley P
Quantifying Innovation in Stroke: Large Language Model Bibliometric Analysis
J Med Internet Res 2026;28:e70754
URL: https://www.jmir.org/2026/1/e70754
doi:10.2196/70754
PMID:

©Adam Marcus, Georgina Lockwood-Taylor, Daniel Rueckert, Paul Bentley. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Electrocardiogram-Based Mental Stress Detection Amid Everyday Activities Using Machine Learning: Model Development and Validation Study

Buelent Uendes[1], MSc; Alex Antonides[1], MSc; Sjors van de Ven[2], MSc; Denise Johanna van der Mee[2], PhD; Eco de Geus[2], PhD; Mark Hoogendoorn[1], PhD

[1]Department of Computer Science, Vrije Universiteit Amsterdam, De Boelelaan 1111, Amsterdam, The Netherlands

[2]Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Corresponding Author:
Buelent Uendes, MSc
Department of Computer Science, Vrije Universiteit Amsterdam, De Boelelaan 1111, Amsterdam, The Netherlands

Abstract

Background: Frequent, sustained stress is linked to poor health and requires monitoring for early intervention. Electrocardiograms (ECG) are promising biomarkers because they can be recorded noninvasively and continuously using wearable devices. However, tracking stress with ECG is challenging because daily activities elicit responses similar to mental stress (MS), and various mental stimuli that individuals encounter complicate the use of machine learning (ML) models trained on a limited set of stressors.

Objective: We (1) evaluated the ability of ML models to distinguish MS episodes from a composite "no-stress" background, including rest and low- to moderate-intensity activities; (2) assessed their generalizability to new stressors and participants; and (3) tested robustness to lower sampling rates and fewer features, to explore their suitability for lightweight wearables.

Methods: We used a comprehensive ECG dataset sampled at 1000 hertz from 127 participants who underwent various mental stressors and engaged in diverse physical activities. A 30-second window was used to extract 55 features from time, frequency, nonlinear, and morphological domains. We trained a logistic regression (LR) model and an extreme gradient boosting (XGBoost) model, splitting the data into 60/20/20 for training, validation, and testing. Shapley additive explanation values were computed to explain model predictions. Additional analyses included leave-one-stressor-out; downsampling to 500, 250, and 125 hertz; a time-window sensitivity analysis; and reducing the number of features to as few as 5.

Results: XGBoost achieved an area under the receiver operating characteristic curve (AUROC) of 0.741 (95% CI 0.701 - 0.783) and an area under the precision-recall curve (AUPRC) of 0.706 (95% CI 0.658 - 0.753), compared with 0.724 (95% CI 0.678 - 0.772) and 0.691 (95% CI 0.639 - 0.742) for LR. The mean performance difference between XGBoost and LR was 0.017 for AUROC (95% CI 0.001 - 0.032) and 0.015 for AUPRC (95% CI −0.001 to 0.037; clustered bootstrap analysis using 2000 participant-level resamples), suggesting that LR performs comparably to the nonlinear XGBoost model. Both models were robust to downsampling and feature reduction (10 features retained >93% of performance). Extending the analysis window to 60 seconds improved model performance across all sampling rates, highlighting a trade-off between rapid detection and overall performance. When evaluating discrimination from physical activity, models achieved acceptable specificity for light physical activity (XGBoost: 0.787; LR: 0.794) but poor specificity for moderate physical activity (XGBoost: 0.418; LR: 0.444). Both models generalized to most unseen stressors, although performance varied across stressors, with limited transfer to the social-evaluative stressor. Feature importance analysis revealed fuzzy entropy and frequency-based features as key predictors.

Conclusions: ML models can detect MS with high sensitivity and remain robust to lower sampling rates and fewer features. Generalization to novel stressors was stressor-dependent. Importantly, our results highlight challenges in distinguishing stress-related cardiac responses from those caused by physical exertion, revealing critical limitations of single-sensor ECG approaches for MS detection.

(J Med Internet Res 2026;28:e80450) doi:10.2196/80450

KEYWORDS

mental stress; machine learning; electrocardiography; ECG; stress detection; generalizability

XSL•FO
RenderX

Introduction

Background

Mental stress (MS), an ever-present aspect of life, occurs when external demands exceed an individual's available physiological and psychological coping resources [1,2]. While short-term exposure to stress may enhance focus and performance [3], frequent or prolonged exposure can adversely affect health, contributing to psychiatric conditions [4,5] and cardiovascular disease [6]. Consequently, accurate and timely detection followed by effective stress management is essential to mitigate these adverse health outcomes.

Traditional stress assessment methods, such as cortisol analysis [7], are burdensome. In contrast, self-report questionnaires, such as the Perceived Stress Scale [8], administered retrospectively or through ecological momentary assessment, offer greater convenience but are limited in temporal granularity [9]. Moreover, they may be susceptible to recall and reporting bias and less reliable among individuals with alexithymia. Effective stress detection, by contrast, should be minimally invasive and support continuous monitoring to capture early warning signs of stress [10].

Given that stress modulates the activity of the autonomic nervous system (ANS) [11], physiological biomarkers such as the electrocardiogram (ECG) and electrodermal activity have been widely proposed for automated stress monitoring. Furthermore, these signals are increasingly accessible through modern wearable devices, making real-time stress tracking more feasible. However, physiological signals often exhibit subtle and complex patterns that can be difficult to analyze and interpret using traditional statistical methods [1].

Objectives

Machine learning (ML) has emerged as a powerful tool for analyzing high-dimensional data, excelling at uncovering patterns and complex relationships [12]. Consequently, ML has increasingly been used in recent studies to detect stress responses from the biomarkers mentioned above [13-16]. However, most existing studies exhibit one of the following limitations: (1) small cohorts; (2) a narrow range of mental stressors investigated; and (3) restrict MS detection to seated baseline (BL) conditions. This restriction leaves it unclear whether ML models can distinguish stress-induced ECG signals from those triggered by everyday movements, which often produce similar physiological responses. Moreover, because individuals encounter a wide range of novel stressors in daily life, models must generalize to stimuli absent from their training data. This aspect has been largely overlooked in prior work. The few studies that investigate generalization [17-19] typically train their ML models on one dataset and test them on another. Although informative, this approach conflates differences in participant demographics, sensor hardware, and stressor types, making it hard to isolate which factor drives any performance drop.

In this study, we address these gaps by developing and rigorously evaluating two ML models, namely logistic regression (LR) and extreme gradient boosting (XGBoost), on a large ECG dataset sampled at 1000 Hz. The dataset, collected from a controlled laboratory experiment at the Vrije Universiteit Amsterdam and published by van der Mee et al [20], included 127 participants who performed both diverse mental-stress tasks and everyday activities (eg, dishwashing and walking). In this study, we define "mental stress" as the physiological response elicited by cognitively or emotionally challenging tasks. Subsequently, we formulated a binary MS versus no-stress classification task, where the no-stress condition includes seated BLs, recovery periods, and low-to-moderate-intensity physical activities. This design not only improves ecological validity but also tests each model's ability to distinguish stress‐induced ECG changes from those driven by physical exertion, without relying on additional sensors (eg, accelerometers). To assess generalization, we evaluate our models on unseen participants and on "novel" stressors, using a leave-one-stressor-out protocol, where we withhold each mental stressor during training and assess model performance on this specific stressor. We further investigate the robustness of our models to lower sampling frequency and the reduction of the feature set. Together, these experiments provide key insights for transitioning ECG-based MS detection from laboratory settings to real-world contexts.

Methods

Overview

The structure of the following section is in line with the TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Artificial Intelligence) statement [21,22]. Sections of the guideline not applicable to this research were omitted for brevity.

Data

This study is a retrospective analysis of pseudonymized data collected from a controlled laboratory experiment conducted at the Vrije Universiteit Amsterdam and published by van der Mee et al [20]. The experimental conditions are shown in Table 1. In addition to ECG signals, subjective affect was assessed after each experimental condition using self-reported measures of positive and negative affect, as measured by the Maastricht Questionnaire [23]. We refer to the original study by van der Mee et al [20] for detailed explanations of the experimental procedure.

XSL•FO
RenderX

Table . Experimental conditions and their associated labels used in our study. The laboratory protocol consisted of mental stress (MS) tasks, physical activities at varying intensities, baseline (BL) rest, and recovery periods, each labeled for binary classification as MS or nonstress (BL, low physical activity [LPA], and moderate physical activity [MPA]). A total of 127 healthy adults participated in the experiments conducted in the Netherlands (data collected from 2017 to 2019).

Experimental condition	Duration (min)	Label
Standing	3	LPA
Sitting	3	BL
Tone avoidance	4	MS
Recovery (sitting)	2	BL
Short sing-a-song stress test (anticipatory)	1	MS
Recovery (sitting)	2	BL
Paced auditory serial addition test	4	MS
Recovery (sitting)	2	BL
Raven's Progressive Matrices	4	MS
Walking at a natural pace	2	MPA
Recovery (standing)	2	LPA
Dishwashing	2	MPA
Vacuum cleaning	2	MPA
Recovery (sitting)	2	BL
Tone avoidance (repeat)	4	MS
Recovery (sitting)	2	BL
Paced auditory serial addition test (repeat)	4	MS
Recovery (sitting)	3	BL

Participants

As outlined in the study by van der Mee et al [20], participants were required to be Dutch-speaking, employed or enrolled in school, and aged 18-48 years. Exclusion criteria included a BMI above 30, high cholesterol, diabetes, liver or thyroid disease, and use of medications that affect the ANS, such as antidepressants or anticholinergics. The final sample included 127 participants (71 female and 56 male) with a mean age of 23.3 (SD 5.6) years. We refer the reader to the original study for additional details on recruitment and population. The data used in this analysis were pseudonymized.

Outcome

For this research, the different experimental conditions, as shown in Table 1, were categorized into the following labels:

- BL: This group includes conditions in which participants were primarily resting, such as sitting and recovery periods.
- MS: This group includes conditions aimed to induce cognitive or emotional stress, such as the tone avoidance (TA) reaction time task, the sing-a-song-stress test (SSST), the paced auditory serial addition task (PASAT), and the Raven's Progressive Matrices (RAVEN) test. For the SSST, the labeled stress segments correspond to the anticipatory phase before singing, that is, the period during which participants are informed they will be asked to sing but have not yet begun.
- Low physical activity (LPA): Standing, including recovery periods that involve standing, is classified as LPA.
- Moderate physical activity (MPA): Activities involving moderate physical exertion, such as walking at a moderate pace, washing dishes, and vacuuming.

Data Preparation

ECG signals are susceptible to various forms of noise and artifacts, which can obscure underlying cardiac activity and hinder accurate analysis. Thus, data preprocessing is crucial to ensure the quality and consistency of the input for the subsequent analysis.

In the first stage of data preprocessing, ECG segments unrelated to the laboratory conditions of interest were removed. These segments included the first and last minutes of the experimental setup, which captured the experimental setup and ECG lead removal, as well as any designated "short break" conditions. Compared with the original study, we excluded high-intensity physical activities (eg, treadmill running and stair climbing), as these activities are easily distinguishable from MS by, for instance, considering the maximum heart rate (HR) and could inflate reported model performance.

In the subsequent step, a 0.5 Hz high-pass Butterworth filter (order=5) was used to remove BL wander and slow drifts in the signal BL caused by respiration, movement, or electrode issues, as well as a power line filter (50 Hz) to attenuate noise from electrical power sources. Both filtering steps were implemented using the preprocessing pipeline provided by the NeuroKit2 package (version 0.2.7) [24]. While filtering addresses signal quality issues, some segments may remain unreliable due to

XSL•FO
RenderX

factors such as temporary electrode detachment caused by leads accidentally falling off during the experiment. To identify and address these segments, a signal quality index (SQI) was computed for each QRS complex. Segments with an SQI below 25% were deemed unreliable and discarded. This threshold was chosen based on visual inspection, where segments with SQI values below this level often exhibited flat lines, rendering the identification of key peaks unreliable.

Predictors

Overview

For each participant, the data was segmented using a sliding-window approach, and features were subsequently calculated. We used a 30-second window size with a step size of 10 seconds. This window length represents the minimum duration demonstrated in prior work to yield reliable heart rate variability (HRV) estimates [25], yet remains short enough for near-real-time applications. Furthermore, our window size aligns with previous research [26].

Following R-peak detection and waveform delineation of the ECG signal, various features were extracted from the resulting ECG signal based on prior research:

Time-Domain Features

For each time window, we extracted the normal-to-normal (NN) interval series, converted it to an instantaneous HR series (in beats per minute [BPM]), and then computed the mean, SD, minimum, and maximum of that series. These measures capture the central tendency and dispersion of the HR, which typically rise during physical exertion and in response to acute mental stressors [27].

HRV features are essential markers for ANS activity [25] and have been consequently linked to stress in prior research [28-30]. Therefore, we calculated the average value of the NN intervals (AVNN: average normal-to-normal), the SD of the NN intervals (SDNN), as well as the root-mean-square of successive differences (RMSSD) for each time window. We also included the ratio between the SDNN and the RMSSD, as it serves as a time-domain surrogate for the low-frequency to high-frequency power ratio, a standard marker of ANS activity [31,32].

In addition to features like AVNN and SDNN, we derived several complementary time‐domain features to capture both absolute and relative variability in NN intervals. These include NN20 and NN50, the counts of consecutive interval differences exceeding 20 milliseconds and 50 milliseconds, and their normalized counterparts, PNN20 and PNN50, as these features are helpful for stress monitoring [33,34].

To assess variability relative to mean HR, we calculated CVNN (SDNN/AVNN) and CVSD (RMSSD/AVNN). Finally, we quantified dispersion beyond SDs by computing the IQR of NN intervals and the median absolute difference of NN intervals.

Frequency-Domain Features

To assess the spectral characteristics of the ECG signal, Welch's method was used to estimate the power spectral density across three frequency bands: high frequency (HF; 0.15‐0.40 Hz), very high frequency (VHF; 0.40‐0.50 Hz), and ultra-high frequency (UHF; 0.50‐1.00 Hz). Although prior studies commonly analyze only frequencies up to the HF band (0.40 Hz), this may be restrictive during stress, as acute psychological stress can elevate respiratory rate [35-37], potentially shifting the respiratory-driven modulation of R-to-R intervals above the HF band. Supporting this, Hernando et al [38] showed that respiratory rate increases and becomes less stable during stress, and that incorporating respiratory information improves the characterization of the stress-related autonomic response.

Motivated by this evidence, we included the VHF band (0.40‐0.50 Hz), as implemented in the NeuroKit2 toolbox [24], to capture high-frequency variability potentially associated with stress-related changes beyond the conventional HF range. We additionally investigated the UHF band (0.50‐1.00 Hz) in an exploratory manner, as this range may reflect high-frequency variability components that differ systematically across conditions. Lower-frequency bands, such as very low frequency (0.0033‐0.04 Hz), were not considered in this study, as their estimation would require longer time windows to be accurately captured [31].

Following the approach taken by Schmidt et al [13], we computed the absolute spectral power within each predefined frequency band. Then we expressed each band's power as a fraction of the total spectral power to quantify its proportional contribution. We also extracted the minimum, maximum, mean, SD, and entropy for each frequency band in line with the approach taken by Karthikeyan et al [39].

Nonlinear Features

Given the inherent nonlinearity and complexity of the MS response [40], we included several nonlinear indices of HRV derived from NN intervals to capture the subtle fluctuations and irregularities in ANS activity. Following the approach taken in Tanev et al [41], we included the approximate entropy [42,43] and fuzzy entropy [44]. We excluded sample entropy [43] because approximate entropy is better suited to shorter time windows [45]. Furthermore, as a derivative of sample entropy, fuzzy entropy has been shown to outperform sample entropy [46].

Using Poincaré plot analysis, we extracted the SD1 and SD2 features, along with their ratio, to gain insights into the dynamics of short-term and long-term HRV [45]. Furthermore, heart rate asymmetry (HRA), that is, the asymmetrical distribution between accelerating and decelerating NN intervals, has been shown to reflect MS [47]. Consequently, we extracted the area index [48] as a feature of the HRA. In contrast to other HRA measures, such as Guszik's slope or Porta's index, the area index has been reported to exhibit fewer variations in short-term heartbeat intervals [48].

Because the recurrence quantification analysis has been shown to reflect the sympathetic and parasympathetic nervous system [49,50], we extracted three features from the recurrence quantification analysis plot: average length, longest length, and the entropy of the vertical line (W, W_{max}, and W_{En}).

To assess short-term correlations in the ECG signal, we used detrended fluctuation analysis and extracted the α_1 value [51],

XSL•FO
RenderX

which is moderated by sympathetic activity [52,53]. In contrast to the α_1 value, which requires 3 to 16 beats for estimation, the α_2 value needs 16 to 64 heartbeats [54], which may be problematic given our short 30-second analysis window and the variability in HR across experimental conditions. For this reason, we excluded the α_2 from our feature set.

Last, we included the HR fragmentation features proposed by Costa et al [55], that is, percentage of short segments (PSS), inverse average length of segments (IALS), percentage of inflection points, and percentage of NN intervals in alternation segments, as these biomarkers have been proposed to capture the functionality of the HR control system.

Morphological Features

Beyond linear and nonlinear dynamics, morphological features (ie, characteristics that describe the shape, size, or timing of ECG waves) were included to capture further nuances in cardiac response during MS. In particular, we focused on T-wave alternans (TWA), a beat-to-beat variation in the amplitude of the T-wave that reflects subtle alterations in ventricular repolarization and which has been identified as a potential marker of MS [7,56].

While the spectral method and the modified moving average (MMA) method are both widely used for TWA analysis, the latter was used to extract the TWA feature, as the MMA does not require participants to achieve a HR of 105 to 110 BPM [57]. This is crucial, as participants in the MS or seated BL conditions may experience HRs below this threshold (Figure 1).

Figure 1. Heart rate distributions across experimental conditions (127 total participants). Kernel density plots show mean heart rate (BPM) during sitting, recovery (sitting), low physical activity, mental stress, and moderate physical activity. Overlapping distributions across experimental conditions highlight the difficulty of discriminating between mental stress and nonstress conditions using mean heart rate alone. BPM: beats per minute.

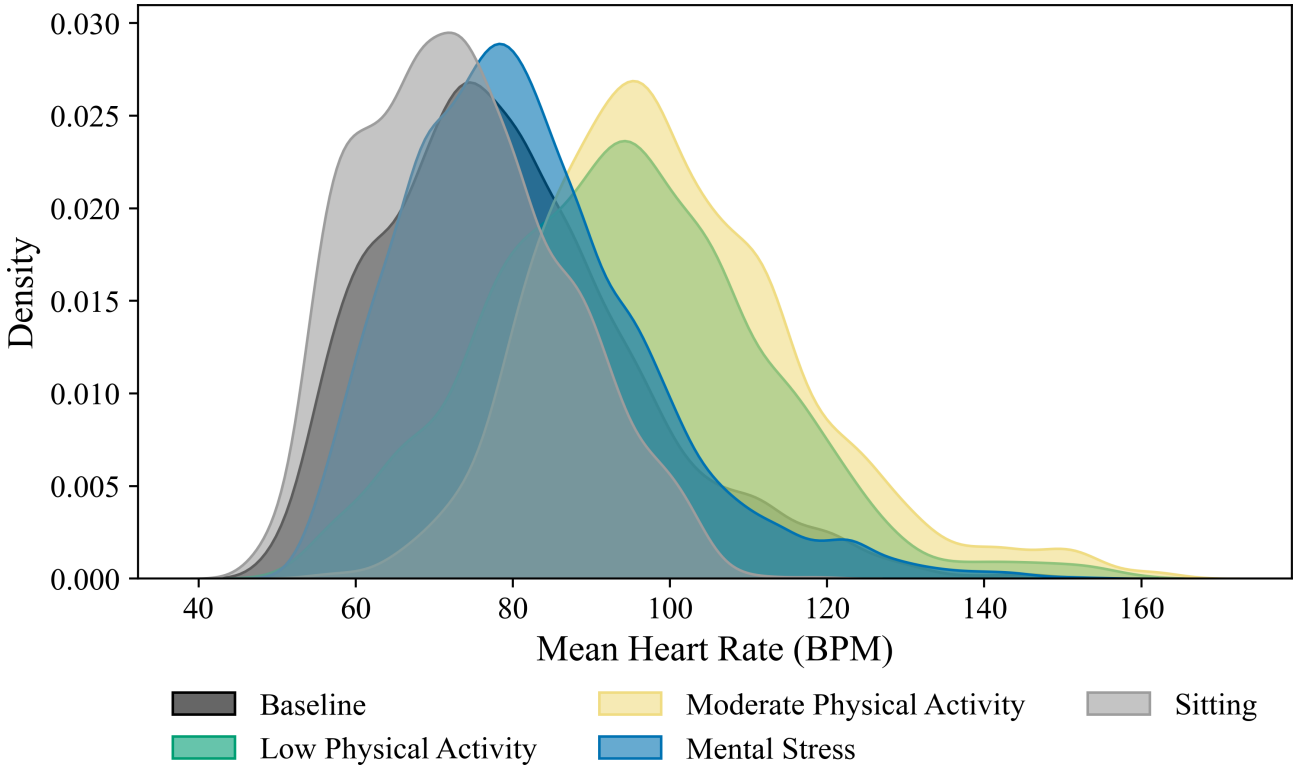

Using the MMA approach [58], we delineated and extracted the T-wave from each 30-second ECG window, numbered them sequentially, and split them into odd (1, 3, …) and even (2, 4, …) beats. We then applied an MMA to each group and defined TWA as the maximum amplitude difference between the odd- and even-beat templates.

The complete set of features used in this study is presented in Multimedia Appendix 1 [55]. We excluded any 30-second window with fewer than 12 detected R peaks or with a mean HR outside the physiologically plausible range of 40 - 220 BPM during feature calculation. This exclusion removed 0.9% of all time windows.

Models were trained using ECG-derived features only to test whether ECG alone is sufficient for stress detection and to enable deployment on any wearable device without requiring personal or demographic data.

Sample Size

As illustrated in Table 2, the 30-second sliding-window approach with a 10-second shift yielded 31,408 observations.

XSL•FO
RenderX

Table . Distribution of electrocardiogram windows across experimental conditions for mental stress (MS) detection (31,408 thirty-second windows from 127 participants; 10-second sliding-window shift). Windows were labeled as MS or nonstress conditions: baseline (BL) rest, low physical activity (LPA), or moderate physical activity (MPA).

Experimental condition	Sample size	Overall percentage (%)
BL	10,140	32.28
LPA	3222	10.26
MPA	3703	11.79
MS	14,343	45.68
Total	31,408	100.00

Missing Data

Due to the controlled, high-quality conditions of the laboratory-based ECG data collection and the preprocessing pipelines used in this study, missing values were scarce. Out of 31,408 analysis windows (Table 2), only 50 (0.16%) windows contained missing data, all of which were attributed to the TWA feature, which requires an adequate number of detected T-peaks for reliable computation. We imputed these missing values using a k-nearest neighbors approach with a k value of 5, where neighbors were selected using Euclidean distance across all 55 standardized (as described below) features, and the missing values were imputed using the mean values of the 5 nearest neighbors. Both the imputer and the scaler were fitted exclusively on the training set to prevent data leakage.

Analytical Methods

ML Models

In this study, we trained an LR with L2 regularization and a nonlinear tree-based model, that is, an XGBoost [59] model. While the LR can only capture linear relationships, its inherent interpretability offers an advantage, especially in the health care domain, where interpretability is a key requirement [60]. By contrast, we selected a tree-based model as the nonlinear approach, given its widespread use in ECG-based stress and affect detection [13,61-63]. Specifically, we selected XGBoost because gradient-boosting methods have demonstrated strong performance across various tabular benchmarks, matching or surpassing neural network architectures [64,65]. By evaluating both ML models side-by-side, we can directly assess the trade-off between more interpretable linear models and more complex ones in the context of ECG-based MS detection. To confirm the stability of the results, we also included results from a random forest (RF) model to assess robustness.

Model Development

We partitioned participants into training (75/125 participants, 60%), validation (26/125 participants, 20%), and test (26/125 participants, 20%) cohorts. For continuous features, we used z-standardization, whereas for discrete features (NN20, NN50, and W_{max}), we applied min-max normalization across the training dataset. We used global rather than participant-specific feature normalization, as global normalization enables assessment of whether ML models can generalize to unseen individuals without requiring per-user calibration data. Hyperparameters for the ML models (LR, XGBoost, and RF) were optimized using Bayesian optimization with Gaussian processes (Optuna version 4.2.1; Preferred Networks, Inc) over 25 iterations; the full hyperparameter tuning ranges and selected values are detailed in Multimedia Appendix 2.

Model Evaluation

We quantified model performance using the area under the receiver operating characteristic (AUROC) curve and the area under the precision-recall curve (AUPRC) curve—the former capturing overall class separation, while the latter emphasizes stress detection performance directly, which is of particular importance in our study. In line with [66], we interpret 0.60 - 0.69 as moderate, 0.70 - 0.79 as acceptable, 0.80 - 0.89 as good, and ≥0.90 as excellent. To evaluate how the ML models would perform in practice, we also report the F_1-score, as well as sensitivity and specificity. For the latter metrics, we used the threshold that maximized the F_1-score on the validation dataset. We computed 95% CIs for all metrics using participant-level cluster bootstrapping (2000 bootstrap samples), in which each bootstrap sample resampled participants within the test dataset with replacement and included all corresponding observations for each selected participant.

Sampling-Rate Robustness

Our ECG recordings were acquired at 1000 Hz, which exceeds the rates of commonly used datasets [13,61,67]. Using the NeuroKit resampling method, we downsampled our ECG signal to 500, 250, and 125 Hz. This procedure quantifies the effects of reduced temporal resolution on model performance, facilitates direct comparison with earlier work, and evaluates the feasibility for consumer-grade devices, such as the Hexoskin Proshirt (Carré Technologies Inc) [68], which operates at lower sampling rates. Furthermore, investigating whether the performance of MS detection is affected by the sampling rate has important implications and considerations for the design of memory-efficient devices.

Time-Window Robustness

To investigate the trade-off between the desire to capture MS in near-real-time (ie, shorter time windows) and the requirement to have sufficiently long windows to generate reliable ECG features, we compared model performance using our standard 30-second segments against extended 60-second segments. Importantly, we used a 20-second time shift for the latter to maintain a constant overall window overlap across both settings. This setup allows us to subsequently investigate how the window size affects model performance.

XSL•FO
RenderX

Model Explainability and Calibration

To explain the predictions of our ML models, we computed Shapley Additive Explanations (SHAP) [69] values. SHAP values are based on a game-theoretic approach to calculate a feature's importance concerning the overall model prediction. For the LR model, we report, in addition to the SHAP values, the model coefficients.

Model calibration (ie, how well the provided model probabilities align with the actual probabilities of the outcomes) is particularly critical in the health care domain [66,70,71]. We assess calibration using calibration curves, the Brier score [72], and the Brier skill score. To obtain the latter, we compare the Brier scores of our ML models to those of a simple majority classifier. These analyses enable us to evaluate not only the predictive performance but also the reliability of the probabilities provided by our ML models. Following [73], we consider a model with a Brier score <0.20 to be well-calibrated. We further use isotonic regression to post hoc calibrate the model probabilities.

Parsimony

As previously outlined, and motivated by prior research, we engineered a total of 55 features spanning time, frequency, nonlinear, and morphological domains to capture physiological responses to MS. However, manual feature generation is a time-consuming process. To quantify the trade-off between model simplicity and predictive performance, we evaluated model performance using progressively smaller feature sets. We applied forward selection to choose subsets of 5, 10, and 20 features by iteratively adding features that maximally improved the overall model performance, while retraining each model at each step of the feature selection process. Compared with alternative feature selection methods such as principal component analysis [74] or mutual information gain, which calculates dependency between the predictor and outcome variable, the forward-selection method directly optimizes for predictive performance during the feature selection process by iteratively adding features that provide the greatest incremental benefit. This process inherently accounts for feature dependencies, as each selection is made conditional on the features already included. Our progressive reduction from 55 to 5 features allows us to quantify the trade-off between model simplicity and performance, while evaluating the extent to which a compact feature subset can maintain classification performance. We note that the forward selection process was performed on a single validation split (see section "Model Development"). The specific feature subset composition may therefore vary across different validation splits. However, the relatively large sample size of our study (127 participants; 31,408 windows) may partially mitigate, though not fully eliminate, the risk of overfitting to a specific data split. Nevertheless, we caution against overinterpreting the feature subsets.

Generalizability Across Stressors

Although our dataset encompasses a variety of different mental stressors (social stressors, cognitive stressors, and active coping), individuals can face a vast majority of varying stressors in their daily lives. To assess whether our models learn a general MS pattern and can generalize to unseen stimuli, we used a leave-one-stressor-out scheme. For each stressor category, we trained models on all but the held-out mental stressor and tested their performance on the unseen stimulus. This analysis can thereby give valuable insights into our model's ability to generalize to novel stress types, a critical requirement for deploying ML for MS detection to ambulatory settings.

We implemented the analysis in Python (version 3.11) and used artificial intelligence tools to accelerate the programming process. Importantly, all output was carefully reviewed, and the code was double-checked and tested for correctness. We release the source code used for this study publicly on GitHub [75].

Class Imbalance

Although the class distribution between the MS and nonstress states (BL, LPA, and MPA) is relatively balanced (45.68% MS vs 54.32% nonstress; Table 2), the leave-one-stressor-out analyses exhibit more pronounced imbalance, as specific mental stressors are left out during training, resulting in a class distribution of approximately 33.33% MS versus 66.67% nonstress examples (when PASAT or TA stimuli are left out). To overcome the class imbalance, we applied the synthetic minority oversampling technique (SMOTE) [76] exclusively to the training set to upsample the minority class, that is, the MS condition. SMOTE is an oversampling method that creates new synthetic instances of the minority class by interpolating between neighboring minority examples. To maintain a consistent training pipeline, we also used SMOTE in the main analysis, although it had a negligible impact on model performance in the more balanced MS versus nonstress setting (AUROC with vs without SMOTE: 0.741 vs 0.742 for XGBoost; 0.724 vs 0.724 for LR).

Model Output

We framed our primary outcome as a binary classification task: detecting MS episodes across all nonstress states, including seated BLs, recovery periods, and LPA to MPA. In this way, our nonstress class reflects both desk-bound environments, such as office work in front of a computer, as well as more mobile contexts, such as those found in the teaching or nursing profession. By incorporating both dynamic and seated non-MS conditions, our models must learn to distinguish ECG changes induced by MS from those resulting from resting or physical activity—an inherently more complex task than handling either condition alone.

Ethical Considerations

Given that this study represents a retrospective analysis of precollected data, ethical approval was not required under the Dutch Medical Research Involving Human Subjects Act (WMO) [77]. However, the study from which the experimental data were obtained [20] was approved by the Medical Ethical Committee of the Vrije Universiteit Amsterdam Medical Center (METc VUmc #2017.374, ABR #NL62442.029.17). All participants provided written informed consent before the start of the experiment and were either compensated with research credits (for student participants) or with a €50 (US $57.92) gift voucher. All data were pseudonymized.

XSL•FO
RenderX

Results

Exploratory ECG Signal Analysis

Before evaluating our models, we first examine 2 fundamental ECG biomarkers, namely mean HR and AVNN reactivity, to illustrate the intrinsic challenge of detecting MS in everyday routine activities. Mean HR was selected for its simplicity of calculation from an ECG signal. In contrast, HR reactivity (or its inverse, AVNN reactivity) was chosen for its link to cardiac sympathetic and parasympathetic tone and its reliability in recognizing MS in experimental conditions [34]; however, because it relies on a controlled BL measurement, it cannot be defined in unconstrained, fully ambulatory settings.

Figure 1 shows the kernel density estimates of mean HR for five experimental conditions: seated BL, recovery (sitting), low- and moderate-intensity activity, MS, and high-intensity activity. As expected, increasing physical activity leads to a corresponding increase in mean HR. Crucially, the distribution for MS (blue) overlaps extensively with both low- and moderate-intensity activity (green and yellow) and even with the resting peaks (BL and recovery in gray). This pronounced overlap underscores the difficulty of disentangling MS-induced changes from those arising during everyday activity or recovery based on a simple measurement such as HR.

Figure 2 shows the AVNN reactivity, calculated as the difference between AVNN during each experimental stressor and the seated BL, across all MS conditions. Although some individuals exhibit positive AVNN reactivity, represented by the mean change from the seated BL, was significantly negative across all tasks, indicating an overall decline in AVNN compared with BL (PASAT: t_{126}=−7.66, P<.001; PASAT [repeat]: t_{126}=−12.02, P<.001; RAVEN: t_{126}=−3.19, P=.002; SSST: t_{126}=−11.92, P<.001; TA: t_{126}=−8.20, P<.001; TA [repeat]: t_{126}=−13.09, P<.001 for 2-sided paired t tests). However, the magnitude of this change varied considerably across experimental conditions. For instance, TA repeat elicits the most considerable mean AVNN reduction, whereas RAVEN matrices produce a more modest decline.

XSL•FO
RenderX

Figure 2. AVNN interval reactivity in ms across mental stressors. AVNN reactivity was calculated for each participant (127 participants) as the average NN interval during the mental stressor minus the average NN interval during the sitting baseline period (ie, the inverse of the commonly used HR reactivity). Statistical significance was assessed using a paired *t* test. Test statistics: PASAT: t_{126}=–7.66, *P*<.001; PASAT (repeat): t_{126}=–12.02, *P*<.001; RAVEN: t_{126}=–3.19, *P*=.002; SSST: t_{126}=–11.92, *P*<.001; TA: t_{126}=–8.20, *P*<.001; TA (repeat): t_{126}=–13.09, *P*<.001 for 2-sided paired *t* tests. AVNN: average normal-to-normal; HR: heart rate; ms: milliseconds; NN: normal-to-normal; PASAT: paced auditory serial addition task; RAVEN: Raven's Progressive Matrices; SSST: sing-a-song-stress test; TA: tone avoidance.

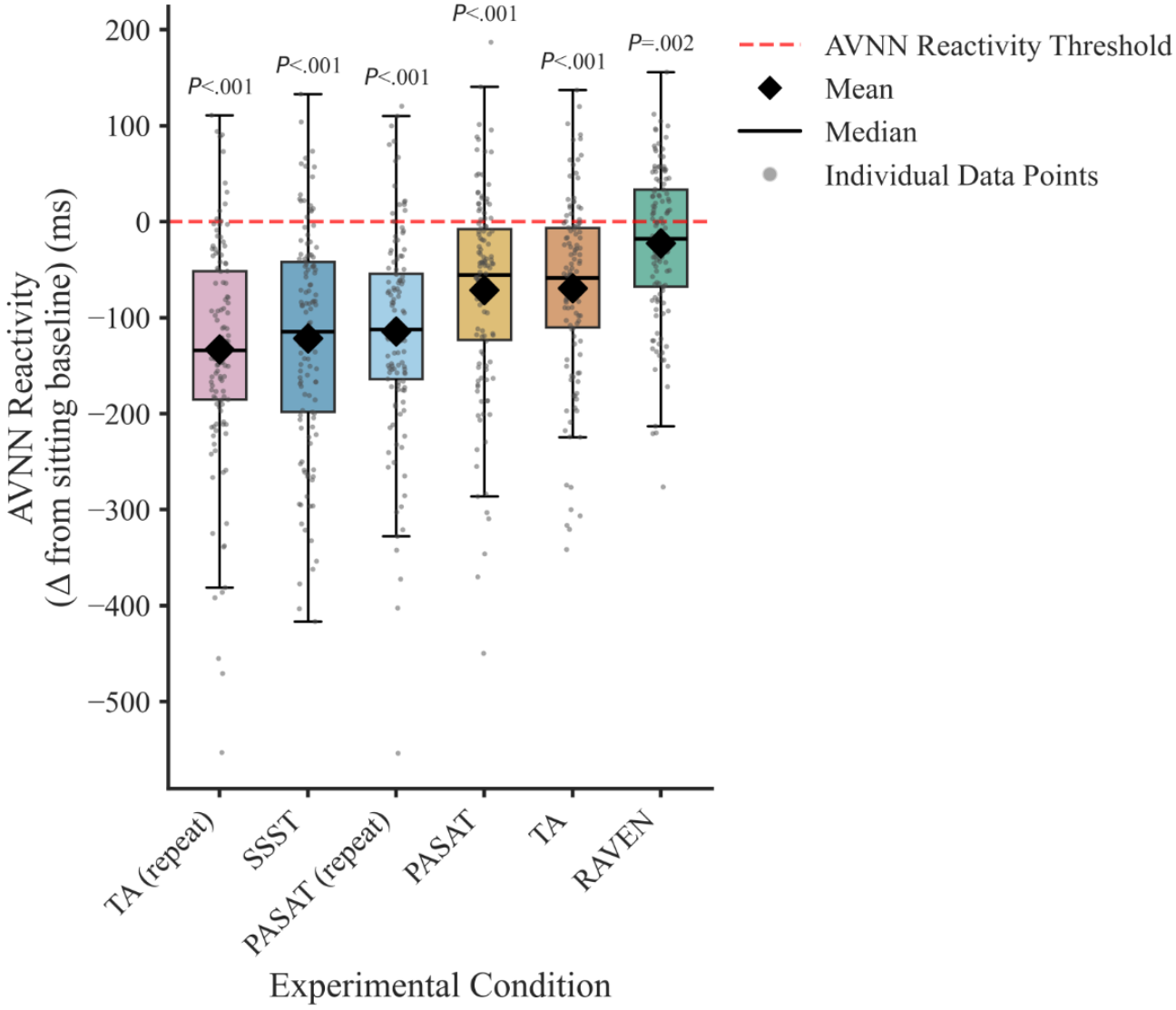

Taken together, these figures demonstrate that neither mean HR nor HR reactivity (or its inverse AVNN reactivity) alone can reliably discriminate MS from everyday activity or account for heterogeneity across stress tasks and individuals. This motivates our extraction of 55 time-, frequency-, nonlinear, and morphological ECG features, and the use of ML models to develop an MS classifier that detects physiological patterns characteristic of MS.

The large interindividual variability in ECG responses underscores that MS reactions are person-specific, likely influenced by subjective perceptions of stress. In Multimedia Appendix 3, we also show changes in both positive and negative affect relative to the sitting BL. As expected, positive affect significantly decreased across all mental stressors (PASAT: t_{126}=−13.25, *P*<.001; PASAT (repeat): t_{126}=−13.45, *P*<.001; RAVEN: t_{126}=−7.96, *P*<.001; SSST: t_{126}=−4.65, *P*<.001; TA: t_{126}=−7.71, *P*<.001; TA [repeat]: t_{126}=−5.00, *P*<.001 for 2-sided paired *t* tests). Likewise, negative affect significantly increased compared with the sitting BL (PASAT: t_{126}=−12.00, *P*<.001; PASAT [repeat]: t_{126}=−6.57, *P*<.001; RAVEN: t_{126}=−6.09, *P*<.001; SSST: t_{126}=−6.67, *P*<.001; TA: t_{126}=−8.31, *P*<.001; TA [repeat]: t_{126}=−3.21, *P*=.002 for 2-sided paired *t* tests). Nonetheless, apparent individual differences emerged, highlighting the challenge of building models that generalize across both stressors and individuals.

Classification Performance

Model Performance Across Sampling Rates

Figure 3 compares LR and XGBoost on our MS detection task at 1000 Hz and after downsampling to 500, 250, and 125 Hz. Both classifiers achieve acceptable AUROC scores, with XGBoost outperforming the LR counterpart across all frequencies (XGBoost: 0.741, 95% CI 0.701 - 0.783; LR: 0.724, 95% CI 0.678 - 0.772) at 1000 Hz. Interestingly, AUROC

XSL•FO
RenderX

performance remains stable across all sampling rates for both models, suggesting that sampling at higher rates does not yield performance gains. In Multimedia Appendix 4, we report corresponding AUPRC scores, which show similar trends and qualitative results. To estimate the performance difference between XGBoost and LR, we computed the mean pairwise differences in AUROC and AUPRC using 2000 participant-level cluster bootstrap samples, alongside 95% and 99% CIs to quantify uncertainty around these estimates. The results are reported in Multimedia Appendix 5. We chose the bootstrap approach over the DeLong test [78] because the latter assumes independent samples [79], which is violated in our setup due to within-participant correlation among the time windows. As shown in Multimedia Appendix 5, the mean performance difference (Δ) between XGBoost and LR was 0.017 for AUROC (95% CI 0.001 - 0.032) and 0.015 for AUPRC (95% CI −0.007 to 0.037), using 2000 bootstrap samples. These differences are small in magnitude (<0.02), suggesting that both models perform comparably across sampling rates. At the F_1-maximizing threshold (1000 Hz), XGBoost achieved a sensitivity of 0.800 (95% CI 0.698 - 0.889) but a specificity of only 0.512 (95% CI 0.420 - 0.609); LR performed similarly (sensitivity: 0.782, 95% CI 0.671 - 0.878; specificity: 0.509, 95% CI 0.417 - 0.613). This asymmetry indicates that, although both models reliably detect MS, approximately half of the nonstress segments are misclassified, highlighting a critical challenge for deployment. Complete results across all sampling rates, along with F_1-scores and RF results, are provided in Multimedia Appendix 6.

Figure 3. Performance comparison of logistic regression and extreme gradient boosting for electrocardiogram-based mental stress classification across sampling rates (127 total participants, 26 test set participants). Points represent bootstrapped mean AUROC with 95% CIs (error bars) from 2000 participant-level bootstrap samples. Models were trained on 55 features extracted from 30-second windows (10-second shift) using 60/20/20 (train/validation/test) splits at the individual level. Both models demonstrate robustness to downsampling from 1000 to 125 Hz. AUROC: area under the receiver operating characteristic curve.

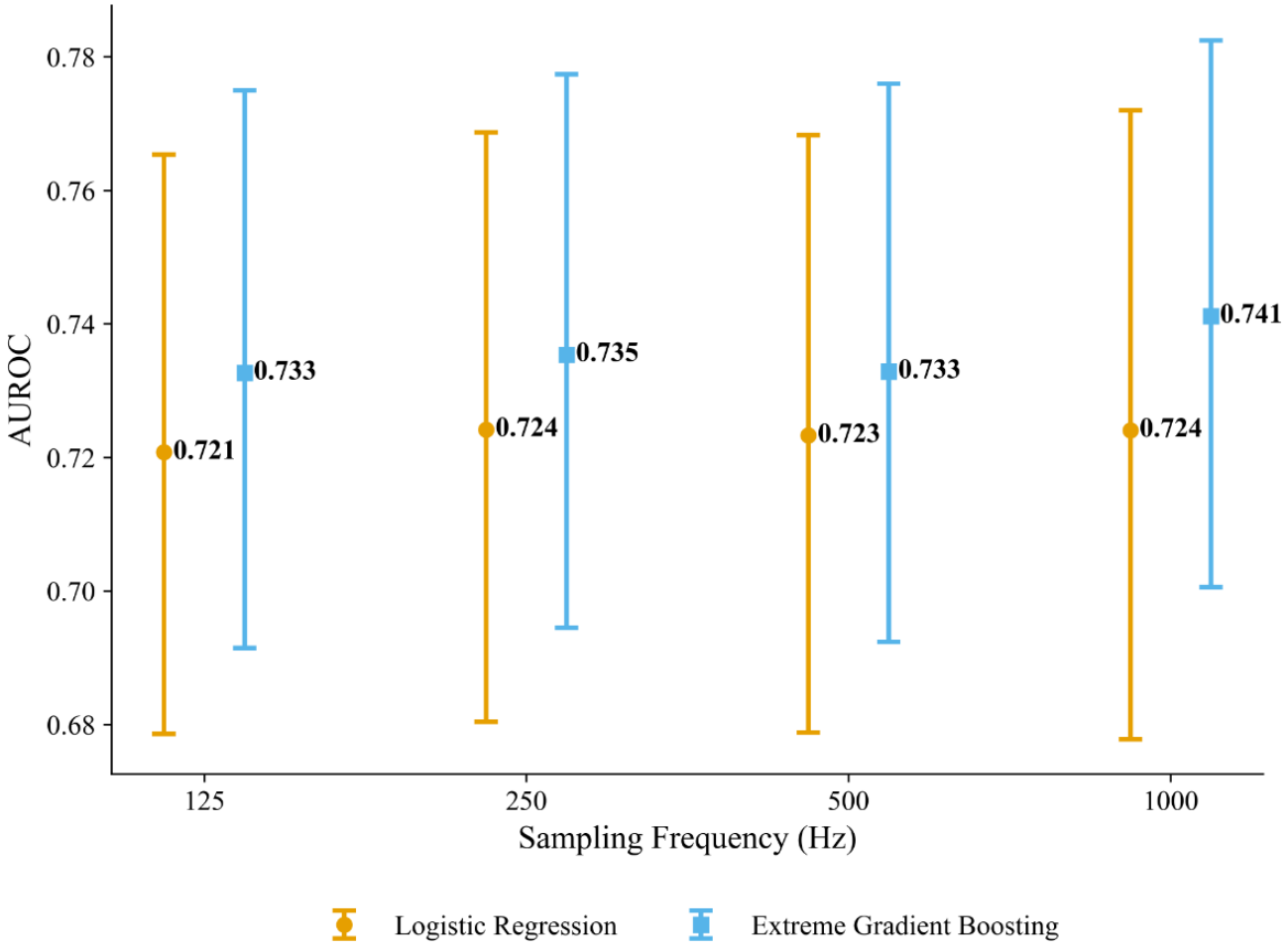

Model Performance Under Simplified Conditions

To assess the impact of our inherently more challenging classification setup, we retrained our ML models on a simplified task: distinguishing MS from a seated BL. The corresponding results are reported in Multimedia Appendix 7. As can be seen, model performances improve substantially under these simplified settings: bootstrapped mean AUROC for XGBoost increases from 0.741 to 0.766, and from 0.724 to 0.752 for the LR at a sampling rate of 1000 Hz (with similar increases for the other sampling rates). The corresponding AUPRC improvements are even more substantial, underlining the challenge our setup poses for MS detection (especially without reliance on accelerometers).

As physiological recovery from stressors is not instantaneous, the recovery periods following each experimental condition (Table 1) likely contain residual autonomic arousal. To

XSL•FO
RenderX

investigate the impact of this on model performance, we conducted a sensitivity analysis excluding recovery periods from the nonstress class. The AUROC performance remained stable (LR: 0.728, 95% CI 0.671 - 0.787; XGBoost: 0.751, 95% CI 0.699 - 0.806 at sampling rate of 1000 Hz), whereas the AUPRC improved substantially for both models (LR: 0.825, 95% CI 0.782 - 0.867; XGBoost: 0.848, 95% CI 0.811 - 0.885), suggesting that including recovery periods introduced label noise negatively affecting AUPRC, likely due to residual autonomic arousal from preceding stressors. However, because distinguishing stress from poststress recovery is an important and realistic challenge in detecting MS in everyday life, we retained recovery periods in the main analysis.

Time-Window Robustness

In Multimedia Appendix 8, we compare AUROC and AUPRC scores for 30-second versus 60-second windows for both models. Extending the time window yields consistent performance gains for both LR and XGBoost across all frequencies, highlighting the trade-off between quick stress detection (short time windows) and improved model performance (longer windows). For instance, at a 1000 Hz sampling rate, doubling the window from 30 to 60 seconds increased XGBoost's bootstrapped mean AUROC from 0.741 to 0.759 (and AUPRC from 0.706 to 0.742), while LR's mean AUROC rose from 0.724 to 0.747 (AUPRC from 0.691 to 0.730).

Instance-Level Agreement

We further assessed Cohen kappa regarding the model agreement on individual test instances. The substantial agreement of 0.735 (95% CI 0.695 - 0.775) suggests that LR and XGBoost largely agree on an instance-level, which ECG responses reflect MS.

Model Explainability and Model Calibration

We next examined feature importance and model calibration, that is, how well the predicted probabilities represent the true probability of the outcomes. Figure 4 shows the 10 most important predictors for mental-stress detection in the XGBoost model according to the SHAP values. Notably, 5 out of these features originate from the frequency domain, highlighting the importance of spectral features in stress detection. Two heart-rate fragmentation indices (PSS and IALS) and fuzzy entropy also rank highly, with higher values of these nonlinear features pushing XGBoost's output towards the MS class.

Interestingly, maximum HR and NN20, 2 simple time-domain features, are also important for predicting MS, demonstrating the relevance of time-domain metrics.

In Multimedia Appendix 9, we present the corresponding SHAP values for the LR model, along with its feature rankings based on the absolute values of the model coefficients. Fuzzy entropy emerges as the most important predictor of MS, underscoring its significance.

IALS again appears to be a crucial feature for MS detection, thereby confirming the importance of heart-rate fragmentation indices and underscoring the results obtained with the XGBoost model. Among time and frequency domain features, total and maximum power in the HF band, SD of HR, and PNN20 are also of high importance for the LR model.

XSL•FO
RenderX

Figure 4. SHAP plot for the detection of mental stress versus no-mental-stress for the extreme gradient boosting model. The top 10 most important features are presented in descending order of importance. Each dot represents a Shapley value for a specific instance and feature, with the color indicating the underlying feature value ranging from high (red) to low (blue). HF: high frequency (0.15-0.40 Hz); HR: heart rate; IALS: inverse average length of segments; NN20: number of (normal-to-normal) NN intervals differing by more than 20 milliseconds; PSS: percentage of short segments; SHAP: Shapley additive explanation; VHF: very high frequency (0.40-0.50 Hz).

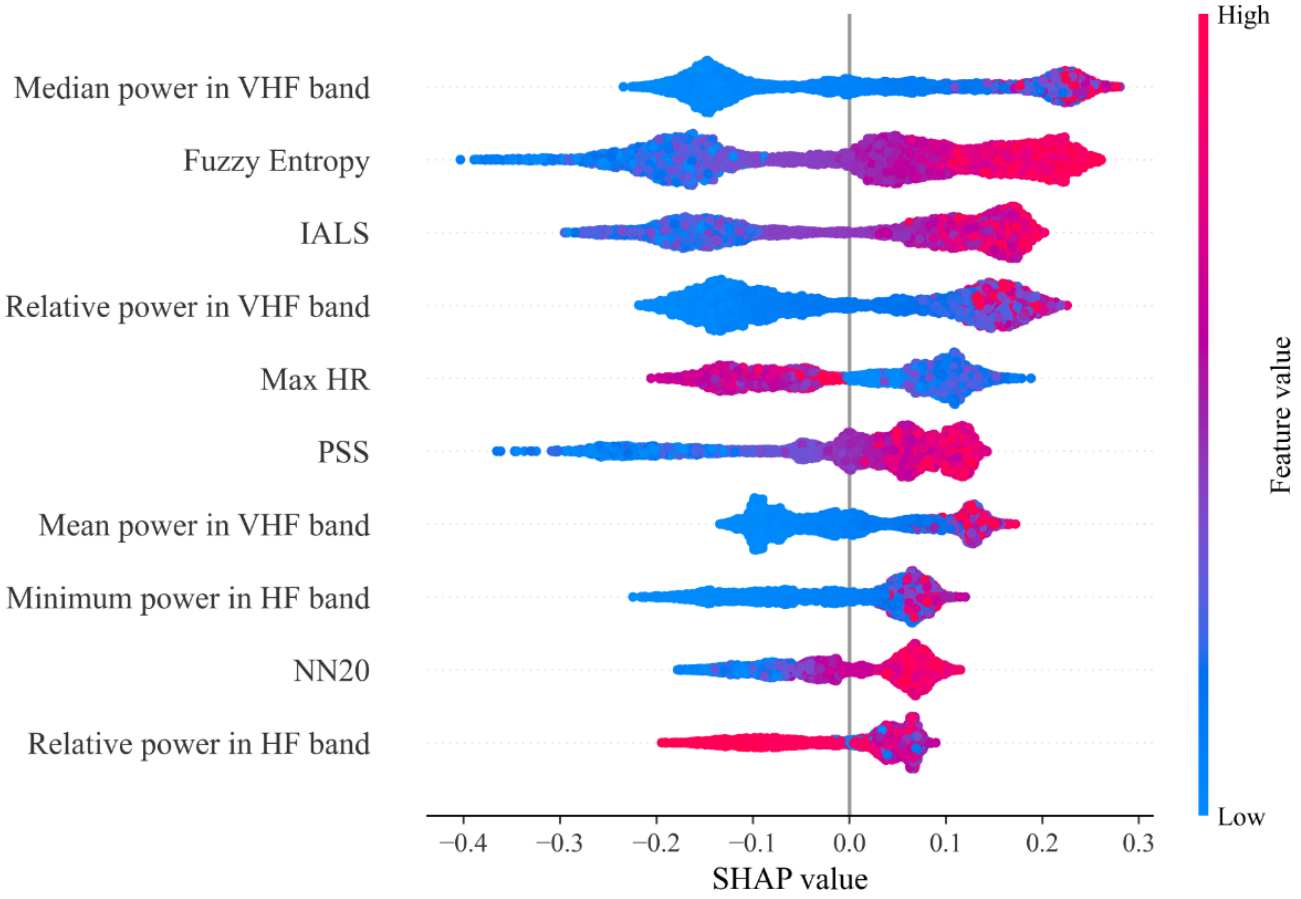

The corresponding model-explainability results for the MS versus the seated nonstress condition (Multimedia Appendix 9; Figures 3-5) also highlight the importance of fuzzy entropy, as well as both the high-frequency and very high-frequency components, as predictive features.

In the health care domain, accurate probability estimates are crucial for informed decision-making [66,70]. For this reason, we assessed model calibration using the Brier score and the expected calibration error with 10 bins.

LR achieves a bootstrapped mean Brier score of 0.215 (95% CI 0.199 - 0.232), while XGBoost scored 0.209 (95% CI 0.194 - 0.224), thereby falling just above our <0.20 cutoff for well-calibrated probabilities. However, the AUROC for the XGBoost model is 0.615 (95% CI 0.586 - 0.643), and 0.603 (95% CI 0.572 - 0.633) for the LR, indicating substantial relative improvements over the dummy classifier BL that predicts the majority class (MS). Figure 5 shows the calibration curves for both classifiers, indicating slight underconfidence at low prediction probabilities (0.05 - 0.25).

XSL•FO
RenderX

Figure 5. Calibration curves for the logistic regression and extreme gradient boosting models, evaluated on the held-out test set (127 total participants and 26 test participants). Points show the predicted model probabilities for mental stress versus the observed probability of mental stress (y-axis), across 10 bins. The dotted line indicates perfect calibration. ECE quantifies deviation from perfect calibration, with lower values indicating better calibration. ECE: expected calibration error.

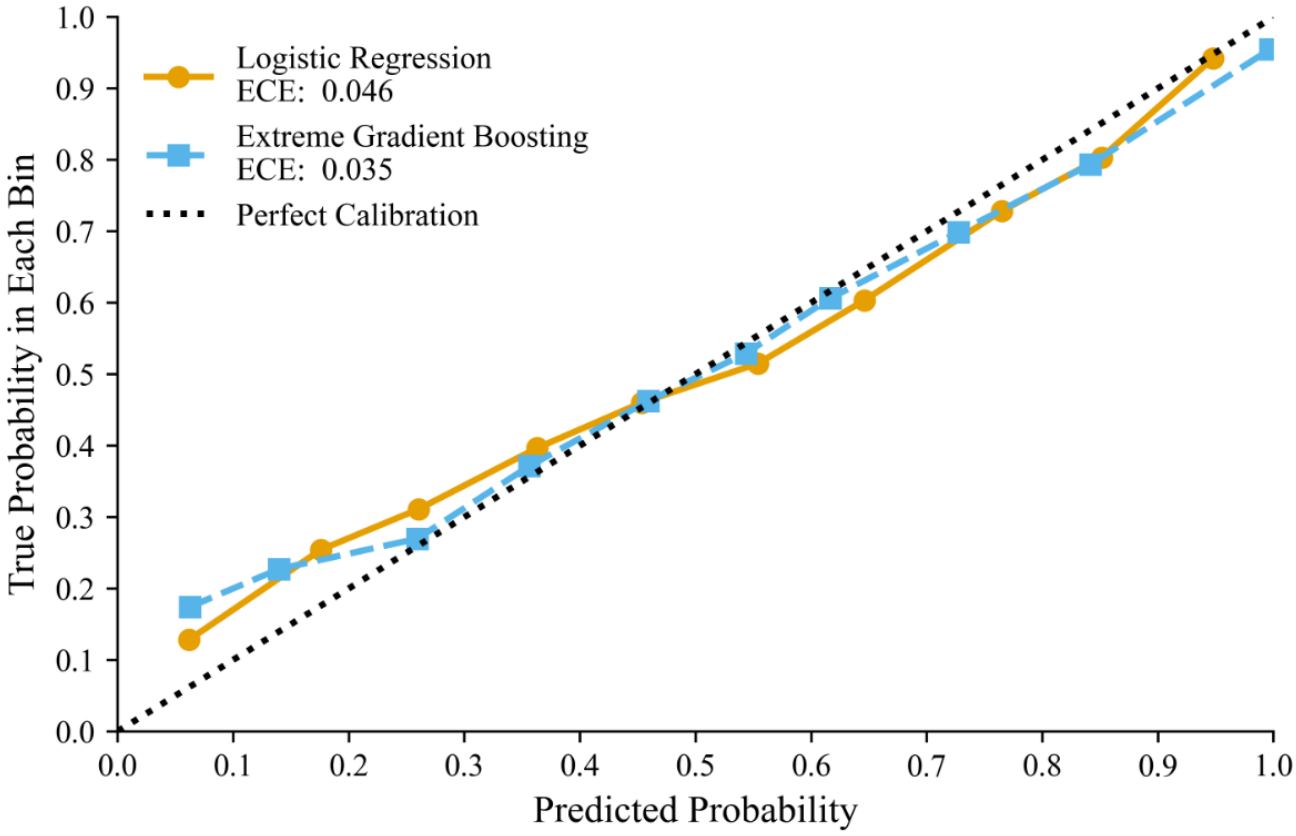

Feature Parsimony

Generating the full set of 55 features is time-consuming and increases memory demands, which are both undesirable for resource-constrained wearable devices. Moreover, a large feature set can obscure model interpretability. To assess the trade-off between model simplicity and predictive performance, Figure 6 reports the AUROC as the number of features is progressively reduced. Notably, performance remains surprisingly robust, with over 95% of the original AUROC retained for both models, even with only 10 features. Interestingly, XGBoost outperforms the LR model across all feature sets. In Multimedia Appendix 10, we report the corresponding results for the AUPRC score, which are in line with the previously stated insights. The full list of selected features for each feature set is provided in Multimedia Appendix 11. As noted in the Methods section, feature selection was conducted using a single validation split. The resulting feature rankings should therefore be interpreted with caution.

XSL•FO
RenderX

Figure 6. Performance comparison of logistic regression and extreme gradient boosting for electrocardiogram-based mental stress classification across varying feature-set sizes on the held-out test set (127 total participants and 26 test participants). Points represent bootstrapped mean AUROC with 95% CIs (error bars) from 2000 participant-level bootstrap samples. Models were trained using a 60/20/20 (train/validation/test) split at the individual level, with features selected via forward selection based on validation performance. Both models demonstrate robustness to feature reduction, maintaining >95% of original performance with only 10 features (logistic regression: 0.707; extreme gradient boosting: 0.729) compared with the full 55-feature set (logistic regression: 0.724; extreme gradient boosting: 0.741). AUROC: area under the receiver operating characteristic curve.

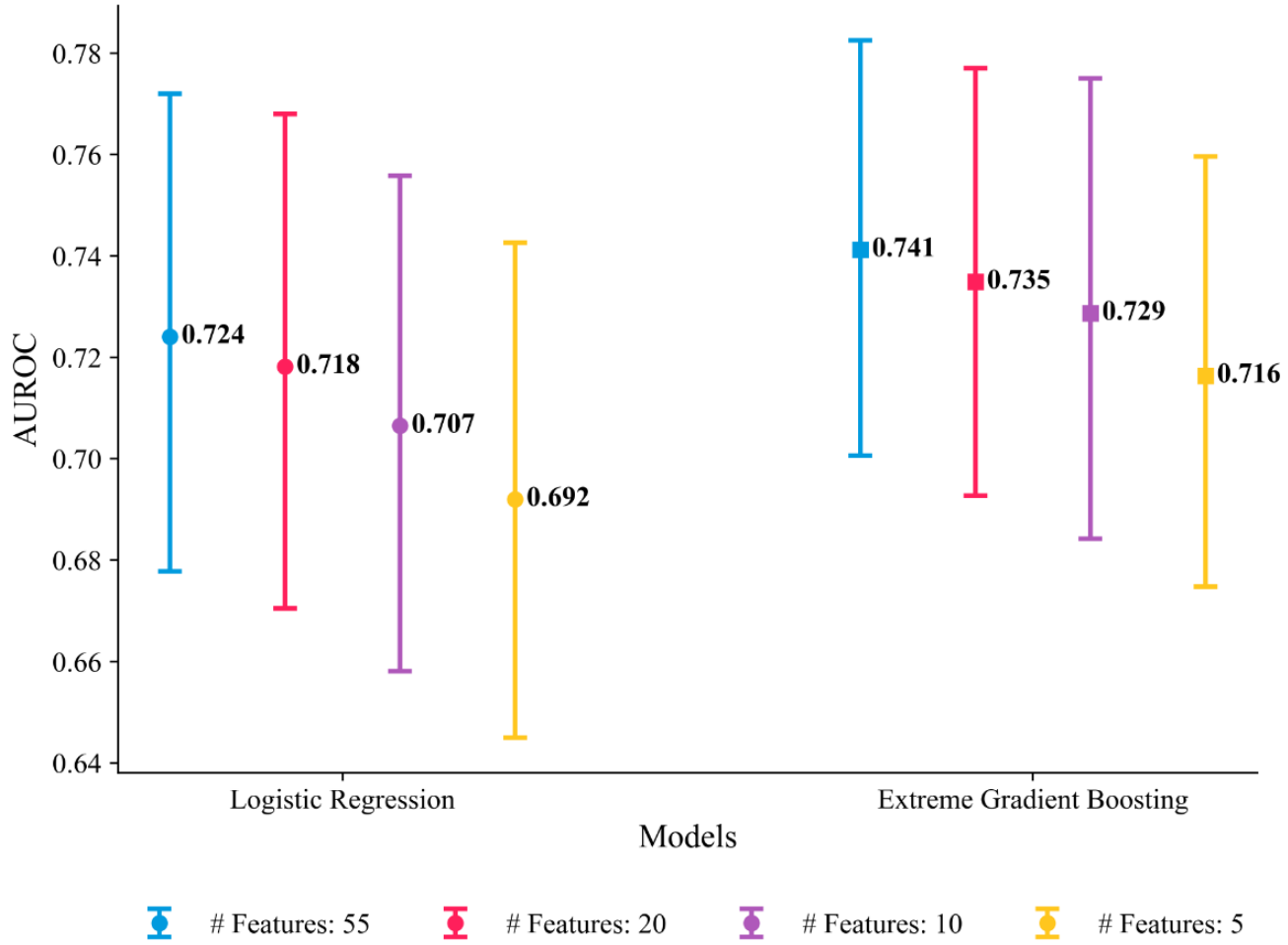

Generalization to Unknown Stressors

Every day, life exposes individuals to a wide variety of mental stressors, making it practically infeasible to train a model on every possible type of stimulus. Thus, a model's ability to perform well under unseen mental stressors is essential for real-world deployment [17]. Table 3 reports the AUROC for each model under a leave-one-stressor-out scheme. For each held-out stressor, the model is trained on all remaining stressors plus the nonstress condition and then tested on its ability to discriminate the held-out stressor from the nonstress condition. Note that if a mental stressor had a repeat condition in the experimental protocol (eg, TA and PASAT), both stressors were left out during training of the ML models. To ensure these results accurately reflect the difficulty of generalizing to an unseen stressor, rather than just the effect of having fewer stressor types and training examples, we also report, for each left-out mental stressor, the BL AUROC on the remaining (known) stressors.

XSL•FO
RenderX

Table . Bootstrapped mean area under the receiver operating characteristic curve (AUROC) with 95% CIs from 2000 participant-level bootstrap samples for binary classification of mental stress versus nonstress conditions under leave-one-stressor-out evaluation on the held-out test set (127 total participants and 26 test participants). Models were trained using a 60/20/20 (train/validation/test) split at the participant level, using a total of 55 features. For each row, the "baseline" column reports the AUROC performance of the model on stressors included in the training set. In contrast, "left-out" reports the AUROC model performance on the held-out (unseen) mental stressor, assessing each model's ability to generalize to novel mental stress conditions. Note that if a mental stressor had a repeat condition in the experimental protocol (eg, tone avoidance and paced auditory serial addition task), both stressors were left out during training of the machine learning models, which is why their baseline scores are identical.

Left-out mental stressor	AUROC model performance, 95% CI					
	LR[a] baseline	LR left-out	XGBoost[b] baseline	XGBoost left-out	RF[c] baseline	RF left-out
SSST[d]	0.7218 (0.6812 - 0.7660)	0.5620 (0.4667 - 0.6548)	0.7360 (0.6975 - 0.7759)	0.5600 (0.4795 - 0.6380)	0.7413 (0.7039 - 0.7792)	0.5499 (0.4680 - 0.6261)
RAVEN[e]	0.7146 (0.6789 - 0.7520)	0.6365 (0.5736 - 0.6976)	0.7365 (0.7070 - 0.7671)	0.6595 (0.6057 - 0.7114)	0.7371 (0.7061 - 0.7696)	0.6645 (0.6092 - 0.7173)
PASAT[f]	0.7219 (0.6777 - 0.7659)	0.7035 (0.6390 - 0.7673)	0.7322 (0.6888 - 0.7752)	0.7125 (0.6490 - 0.7807)	0.7333 (0.6906 - 0.7754)	0.7150 (0.6520 - 0.7816)
PASAT (repeat)	0.7219 (0.6777 - 0.7659)	0.6570 (0.5711 - 0.7456)	0.7322 (0.6888 - 0.7752)	0.6691 (0.5897 - 0.7518)	0.7333 (0.6906 - 0.7754)	0.6782 (0.5985 - 0.7593)
TA[g]	0.7123 (0.6705 - 0.7558)	0.7731 (0.7070 - 0.8352)	0.7107 (0.6710 - 0.7528)	0.7594 (0.6965 - 0.8192)	0.7087 (0.6713 - 0.7492)	0.7608 (0.7017 - 0.8205)
TA (repeat)	0.7123 (0.6705 - 0.7558)	0.7337 (0.6488 - 0.8127)	0.7107 (0.6710 - 0.7528)	0.7200 (0.6454 - 0.7936)	0.7087 (0.6713 - 0.7492)	0.7618 (0.6834 - 0.8370)

[a]LR: logistic regression.

[b]XGBoost: extreme gradient boosting.

[c]RF: random forest.

[d]SSST: sing-a-song-stress test.

[e]RAVEN: Raven's Progressive Matrices.

[f]PASAT: paced auditory serial addition task.

[g]TA: tone avoidance.

As shown in Table 3, model performance on held-out stressors generally declines compared with stressors included in the training set, as expected. Nevertheless, all models achieve AUROC well above the 0.50 chance level on most unseen stressors, confirming their ability to generalize to new mental stimuli. The exception was SSST, for which the 95% CI overlaps chance performance. Notably, the magnitude of performance drop to unseen mental stressors varies across the different stimuli. While both the SSST and RAVEN are most difficult to classify correctly as MS, the opposite holds true for the TA condition. The fact that LR, XGBoost, and the RF model show comparable generalization capability suggests that differences in generalizability to novel stressors arise from the underlying ECG response, rather than from model complexity.

Multimedia Appendix 12 presents AUPRC results that closely mirror the AUROC findings. SSST and RAVEN are the hardest to detect when excluded from model training. Although model performance is generally worse for the left-out stressor, the TA stressor (and its repeat condition) shows the opposite pattern: the AUPRC score when TA is held out exceeds that of other mental stressors included in training. This suggests that TA produces particularly distinctive and readily identifiable stress patterns.

Stratified Stressor Analysis

While our leave-one-out evaluation results suggest that our ML models can overall generalize to novel mental stressors, the generalizability to unseen mental stressors varied significantly. In particular, the sharp contrast between TA and SSST, 2 mental stressors that elicit strong physiological responses as measured by AVNN reactivity (Figure 2), is intriguing. We therefore examined how each model performed on individual stressor types. Table 4, therefore, reports AUROC scores stratified by mental-stressor condition. Again, model performance varies by mental stressor: TA achieves AUROC>0.75, while others (eg, RAVEN and SSST) yield only moderate performance, which suggests that certain MS stimuli produce more distinguishable ECG responses (as compared with the nonstress class) than others. Importantly, this observation holds for LR, XGBoost, and RF alike, demonstrating that these differences are not related to model complexity. The results for the AUPRC, as presented in Multimedia Appendix 13, align with these findings.

XSL•FO
RenderX

Table . Bootstrapped mean area under the receiver operating characteristic curve (AUROC) with 95% CIs from 2000 participant-level bootstrap samples for mental stress classification, stratified by individual stressor type and evaluated on the held-out test set (26 participants from 127 total participants). Results are shown for logistic regression (LR), extreme gradient boosting (XGBoost), and random forest (RF) trained using 60/20/20 (train/validation/test) splits at the participant level with 55 features. Each row presents model performance for a specific mental stressor, assessing how well the models detect that stressor. An AUROC score of 0.50 represents chance-level performance (ie, a random baseline that predicts the majority class).

Mental stressor	AUROC model performance, 95% CI		
	LR	XGBoost	RF
SSST[a]	0.5902 (0.4955 - 0.6792)	0.5726 (0.4958 - 0.6464)	0.5690 (0.4926 - 0.6421)
RAVEN[b]	0.6657 (0.6090 - 0.7230)	0.6829 (0.6350 - 0.7312)	0.6893 (0.6429 - 0.7348)
PASAT[c]	0.7087 (0.6435 - 0.7745)	0.7231 (0.6607 - 0.7882)	0.7219 (0.6588 - 0.7871)
PASAT (repeat)	0.6667 (0.5809 - 0.7558)	0.6846 (0.6063 - 0.7637)	0.6900 (0.6099 - 0.7871)
TA[d]	0.7998 (0.7461 - 0.8540)	0.8176 (0.7729 - 0.8603)	0.8164 (0.7692 - 0.8618)
TA (repeat)	0.7737 (0.6960 - 0.8446)	0.7984 (0.7290 - 0.8614)	0.7962 (0.7262 - 0.8635)

[a]SSST: sing-a-song-stress test.

[b]RAVEN: Raven's Progressive Matrices.

[c]PASAT: paced auditory serial addition task.

[d]TA: tone avoidance.

While the relatively poor model performance on the RAVEN stressor likely stems from its weaker elicited physiological responses (AVNN reactivity is displayed in Figure 2), the similarly low performance on SSST is more surprising. One possible explanation is that SSST elicits a fundamentally different physiological response compared with the other mental stressors, which could account for the poor generalization of the ML models to the SSST stressor observed in Table 3. To test this hypothesis, we retrained our LR and XGBoost classifiers separately on each stressor, that is, training and testing exclusively on one condition at a time. Multimedia Appendix 14 summarizes these per-stressor results, showing that both models outperform random BLs across all mental stressors. Notably, SSST shows marked improvement: despite both models struggling to detect it in earlier analyses, it now achieves an AUROC of 0.73 for LR and 0.76 for XGBoost, comparable to other mental stressors. This indicates that the SSST task elicits detectable MS responses. However, when SSST is excluded during training, AUROC drops sharply when tested on it (Table 3). Together, this indicates that while our ML models can learn to discriminate between SSST and nonstress when trained on SSST stressors alone, the patterns that the ML models learn from other mental stressors do not generalize to SSST. Interestingly, a similar pattern is also observed for the RAVEN stressor, further supporting this interpretation.

Discrimination From Physical Activity

Given that MS can elicit ECG responses similar to those induced by physical activity (Figure 1), our ML models may conflate metabolic demand with MS. We therefore examined condition-specific sensitivity and specificity for light (LPA) and MPA (at 1000 sampling rates, 55 features).

The XGBoost model achieved good sensitivity for MS (0.800; 95% CI 0.698 - 0.889), with the LR also yielding acceptable performance (0.772; 95% CI 0.656 - 0.877). For the LPA, XGBoost achieved a specificity of 0.787 (95% CI 0.662 - 0.903), whereas LR performed comparably, with a bootstrapped mean specificity of 0.794 (95% CI 0.685 - 0.895). However, our models performed worse for MPA (LR: 0.444, 95% CI 0.314 - 0.576; XGBoost: 0.418, 95% CI 0.299 - 0.542), suggesting that even with a comprehensive feature set of 55 features, distinguishing stress-induced ECG responses from those elicited by MPA remains challenging.

Discussion

Principal Results

We developed ML models to distinguish MS from a composite nonstress BL encompassing various everyday activities (eg, walking at a normal pace, recovering from MS, or vacuum cleaning). Using 55 features derived from the ECG signal, XGBoost achieved an AUROC of 0.741 and AUPRC of 0.706, while LR performed comparably (AUROC: 0.724; AUPRC: 0.692) at 1000 Hz. Both models achieved high sensitivity (XGBoost: 0.800; LR: 0.782) but low specificity (XGBoost: 0.512; LR: 0.509), with this pattern being particularly pronounced for MPA. Our results, therefore, highlight a fundamental challenge in distinguishing stress-induced cardiac responses from those driven by physical exertion when relying solely on ECG features. The ML models used in our study also demonstrated robustness to ECG signal downsampling, retaining more than 93% of performance using only 10 features.

While the nonlinear model (XGBoost) numerically outperformed our linear model (LR), the performance differences were small (<0.02 difference in AUROC and AUPRC at 1000 Hz), suggesting that complex nonlinear models offer only marginal benefits relative to linear BLs. One potential explanation is the so-called Rashomon effect [80], which posits that there exist many equally well-performing models for specific datasets—a phenomenon often observed in high-stakes applications and tabular datasets [81]. High outcome variability, likely present in our dataset, as physiological responses to mental stressors are both participant- and stressor-specific (Figure 2 displays the HRV response), increases the likelihood that a simple model

XSL•FO
RenderX

will perform on par with a more complex one [82]. The fact that we do not apply time-series models to the raw ECG dataset using, for instance, deep learning approaches and instead work with tabular data also likely further contributes to this phenomenon.

Our findings reveal a clear trade-off between window size and model performance, as increasing the window size led to model improvements. This can be explained by more reliable features, as shortening the window size has been shown to affect the quality of HRV features [25]. Deep learning approaches, such as convolutional or recurrent neural networks, can partially overcome this trade-off, as they learn feature representations directly from the ECG signals and have been shown to achieve promising performance, even using shorter window sizes than those considered in our study [83,84]. However, these models are often considered "black boxes" and typically require more data for training than the ML models considered in our study. Although recent advances in time-series explainability methods, such as those by Crabbé and Schaar [85] and Enguehard [86], have been proposed to overcome the challenge of interpreting deep learning time-series models, these approaches typically highlight only key segments in the time-series signal driving the model prediction. However, interpreting raw time series is challenging for humans. In contrast, our ML models use features based on well-studied physiological constructs, thereby facilitating the interpretation of model predictions using methods such as SHAP.

The feature importance analysis revealed the importance of fuzzy entropy for discriminating MS responses from other nonstress states. This finding aligns with prior work, which indicates that entropy-based measures, capturing the irregularity of the ECG signal, increase during periods of MS [87]. Notably, 2 HR fragmentation indices, PSS and IALS, also emerged as important features, despite being less commonly used in ML models for detecting MS. Our findings further highlight the importance of frequency-based features, consistent with a comprehensive review [28] covering 37 studies. In their studies, the authors identified changes in the low- and high-frequency components of the ECG signal as the most reported marker of MS.

Additionally, our analysis identified the VHF (as opposed to the UHF) band as an important predictor. Although the physiological origin of the VHF band is less well understood than that of the LF or HF bands, its predictive value likely reflects stress-induced hyperventilation, a well-documented physiological response to various stressors [35-37]. While the VHF band remained an important feature for distinguishing MS from a seated BL (Multimedia Appendix 9; Figures 3-5), suggesting a physiological contribution under sedentary conditions, we acknowledge that this band might also represent motion artifacts or electromyogram noise during active conditions. The low specificity observed during MPA may partially reflect this confound: If VHF power increases during physical activity, due to motion artifacts rather than vagal modulation, the model may misclassify physical exertion as MS. Further research on this band under stress is needed, particularly using accelerometry or respiratory measurements to distinguish stress-related changes from motion artifacts.

Our leave-one-stressor-out experiments demonstrated that our ML models could generalize to most held-out stressors, though performance varied by stressor. While SSST and Raven proved to be the most difficult stressors to recognize, the opposite was observed for TA: both models detected MS evoked by TA and its repeat condition, even when they had never been trained on it. This suggests that the TA stressor produces a strong and distinct physiological ECG response that can be easily distinguished from nonstress states. Whether this is related to differential psychological processes engaged by these tasks is an open question. The TA evokes frustration and a fear of punishment, whereas the SSST is social-evaluative and invokes a fear of exclusion, similar to Raven and mental arithmetic tasks, which also involve ego threats. Building on this, our findings imply that robust, generalizable models require training on a diverse set of various distinct mental stressors, or, when constrained, focus on mental stressors that evoke a strong physiological response, such as TA.

Although our results demonstrate that the ML models can detect MS with high sensitivity, distinguishing stress-evoked ECG responses from those induced by MPA remains especially challenging (specificity LR: 0.444; XGBoost: 0.418), even with a comprehensive set of 55 handcrafted features. Because deep learning models can learn more nuanced temporal ECG patterns, a possible path forward is to apply time-series deep learning directly to raw ECG data. Moreover, incorporating additional sensing modalities (eg, accelerometer or electrodermal activity) may help the model separate cardiac responses driven by metabolic demand from those driven by psychological stress. We therefore consider deep learning–based multimodal approaches an important direction for future research.

Importantly, our results suggest potential for translation to photoplethysmography (PPG)-based settings, as the most predictive features in our ML models are derived from beat-to-beat dynamics, specifically, the R-peaks and interbeat intervals, rather than ECG waveform morphology (eg, T-wave amplitude). Notably, the XGBoost model maintains most of its performance with 10 features (AUROC: 0.729) or 5 features (AUROC: 0.716; Figure 6). These feature subsets consist of PPG-compatible features (HR statistics and frequency-domain features; Multimedia Appendix 11), thereby supporting the possibility that our results may generalize to PPG signals. Consistent with this potential, Taoum et al [88] demonstrated that HRV metrics derived from PPG agreed with the ECG counterpart for 37 out of 48 features for recordings of 5 minutes and 30 seconds, and further for 16 features in 1-minute and 30-second windows. However, since we train on 30-second ECG epochs, future work must determine whether PPG devices can reliably extract those intervals in similarly short windows. Encouragingly, our models remain robust even when the ECG is downsampled to 125 Hz, supporting deployment on lower-resolution PPG wearables. Yet, PPG signals are more susceptible to motion artifacts than research-grade ECG [89], which may challenge the direct transfer of our feature pipeline. Given these considerations, validating and adapting our models for real-world PPG recordings, particularly under movement conditions and within 30-second windows, is therefore an

XSL•FO
RenderX

important direction for future research and essential before our approach can be deployed on PPG-based wearables.

Comparison With Prior Work

Dataset and Experimental Paradigms

MS detection has gained significant traction in recent years (Gedam and Paul [90] and Sharma and Gedeon [91] provide comprehensive reviews, Can et al [92] provides stress detection in daily life scenarios, and Pataca et al [93] provides wearable-based approaches), driven in part by the collection of large, multimodal datasets, some of which are publicly available, such as SWELL knowledge work [14], wearable stress and affect detection [13], StressID [61], or ForDigitStress [62]. While these datasets vary widely in modality, stressor type, and experimental setup, the majority rely on laboratory settings, inducing stress using the Trier Social Stress Test [94], variants of the Stroop task [95], or mental arithmetic challenges [63], all of which are effective methods for inducing acute stress [96].

Even though these datasets have significantly advanced the field, our experimental protocol and its associated dataset offer 2 distinct advantages compared with prior work.

First, our dataset includes 127 participants, which exceeds the sample sizes of most comparable ECG-based stress datasets [13,14,61]. Our relatively large sample size helps to enhance the robustness and generalizability of our findings, especially given the high between-individual variability in physiological stress responses [26,63,97] (also present in our study; see Figure 2). The small number of participants has also been noted as a limitation of prior work in a systematic review [98]. Notably, Smets et al [99] include 1002 participants; however, their study relies on an ambulatory design that uses self-reported stress via ecological momentary assessment. In contrast, our controlled laboratory protocol induces MS through standardized tasks and captures objective ECG measurements, which can be continuously observed.

Second, our experimental design incorporates 4 distinct mental stressors—2 of which are repeated—as well as a diverse set of nonstress conditions, including seated BL, recovery, or common everyday activities such as walking at a normal pace. This comprehensive experimental setup allows us to address a more ecologically valid and challenging classification problem: evaluating whether mental-stress-specific ECG patterns can be distinguished not only from rest, but also from other non-resting states.

ML Performance and Analysis

Most ECG-based stress-detection studies frame the problem as distinguishing MS from a seated, nonstress BL [13,61,62,100]. In contrast, Sun et al [63] achieved 80.9% between-individual accuracy in 20 participants by using an activity-aware, multimodal classifier, therefore combining ECG, accelerometer, and galvanic skin response to detect MS while sitting, standing, and walking. When the accelerometer input was removed, their performance dropped substantially, highlighting the importance of tracking the motion context. Although our ECG-only model does not match the multimodal performance reported by Sun et al [63], it is worth noting that their approach relies on a 60-second window. In contrast, our study uses a 30-second window, which, as stated above, reduces model performance but enables faster detection. Moreover, our laboratory setup encompasses a broader range of mental stressors and leverages an expanded ECG feature set, highlighting the potential of feature-rich, single-sensor models.

Previous work has often simplified stress detection to distinguishing MS from a seated BL [13,61]. In contrast, our classification task differentiates stress from a composite set of nonstressful everyday activities (eg, walking at a normal pace, poststress recovery, and vacuum cleaning), yielding a more ecologically valid scenario in which people naturally move and recover throughout their day. However, this setup makes our classification task more challenging, and consequently, performance is lower than in studies limited to seated nonstress conditions. When we rerun our models on the classic stress versus seated BL task, AUROC and AUPRC both increase markedly (Multimedia Appendix 12), confirming that the drop in metrics stems from the more demanding classification problem.

Beyond the expanded task, our approach also leverages a rich feature space of 55 time-, frequency-, nonlinear, and morphological ECG features, exceeding the number of ECG predictors used in other studies [13,61,101]. We also include and investigate less commonly considered features, such as HR fragmentation indices, which proved to be of added value according to our feature importance analysis (PSS and IALS). To the best of our knowledge, no previous work has evaluated the impact of aggressive downsampling on downstream model performance. We show that reducing the sampling rate from 1000 to 125 Hz does not degrade performance. While our models were robust to sampling rates, the choice of the window size affected model performance: increasing the window size from 30 to 60 s led to notable performance improvements for both ML models. Consequently, our findings provide valuable insights into the effects of sampling rate and window size choices for MS detection, thereby informing the design of models for resource-constrained wearables.

Model Generalization Across Mental Stressors

Most mental-stress classification studies evaluate models only on the same stressors used during training, while far fewer assess generalization to unseen stressors—a critical capability for real-world deployment. Prior work often evaluates generalizability by training on one dataset and testing on another [17-19,101]. While informative, such cross-dataset evaluations conflate multiple sources of variation—individual demographics, sensor hardware, and stressor types—making it challenging to attribute performance drops to any specific factor. In contrast, our leave-one-stressor-out evaluation isolates stressor novelty, enabling a more targeted assessment of a model's ability to detect MS under previously unseen conditions. As our leave-one-out-stressor experiments demonstrated, both LR and XGBoost could generalize to novel stressors, indicating the feasibility of ML models to adapt to unseen mental stimuli.

Taken together, our work advances MS detection from several perspectives. First, by leveraging an ECG dataset from 127 participants across 6 mental stressors and a diverse set of

XSL•FO
RenderX

nonstress activities, our work addresses a more challenging, ecologically valid task: separating MS from a composite nonstress set that includes everyday activities, for which we demonstrate acceptable performance. Second, we demonstrate that unimodal ML ECG classifiers using 55 features are robust against downsampling from 1000 to 125 Hz, maintaining a performance of over 95% with just 10 features. This highlights the feasibility of lightweight and interpretable models for wearable deployment. Third, through our leave-one-stressor-out evaluation, we provide valuable insights into model generalization to unseen stressors, isolating the challenge of stressor novelty without confounds from demographic or hardware differences.

Limitations

Despite the results and insights presented in the paper, a few limitations warrant attention.

First, our cohort comprises Dutch-speaking young working adults (aged 18 - 48 years). While our analytical framework is broadly applicable, physiological stress responses are highly individual and may vary across age groups, cultures, and health statuses.

Additionally, activity patterns might change over the lifespan [102], further limiting generalizability. Importantly, our results do not directly generalize to older populations, given that HRV, HR reserve, and other time-domain features generally decline with age [103], potentially making the classification problem more difficult. Given that distinguishing MS from cardiac responses evoked by physical activity proved very challenging despite our comparable young cohort, we anticipate that this challenge will persist (or even become more complicated) when considering an older population. An interesting direction for future work is therefore to validate our models in more diverse populations, such as older adults or non-Western cohorts, to ensure robustness and generalizability of our findings.

Second, we used global z-standardization across the training set to enable generalization to unseen users without requiring per-user calibration data. While this approach is attractive for real-world deployment, it does not account for the law of initial values [104], that is, it ignores interindividual variability in physiological responses to mental stressors present in our study (Figure 2). Our global normalization scheme may therefore bias model predictions against specific physiological phenotypes, for instance, by increasing false-positive rates among individuals with elevated resting HR or reduced BL HRV. Future work could examine the extent to which global versus individual-aware normalization strategies systematically affect downstream model performances across different BL physiological characteristics.

Third, we rely solely on task labels rather than participants' self-reports to assign MS conditions, which prevents us from confirming that every labeled stress window corresponds to actual perceived stress. Incorporating self-assessments of perceived stress could have therefore reduced label noise, particularly given the pronounced interindividual variability in physiological responses (Figure 2 and Multimedia Appendix 3), which may partly account for the observed failure to generalize to the SSST stressor. Furthermore, instead of treating stress as a binary variable (stress or nonstress), one could use perceived ratings to classify the specific level or intensity of perceived MS [105].

Fourth, a related limitation of this study concerns the labeling of recovery periods. Physiological recovery from stress is gradual rather than instantaneous, suggesting that segments labeled "non-stress" during recovery periods likely retain residual stress physiology, constituting class-dependent label noise [106]. Empirically, label noise has been shown to adversely affect model performance across various ML approaches, including deep learning methods [107-109]. Our sensitivity analysis supports this observation: excluding recovery periods improved AUPRC for LR and XGBoost (see "Model Performance under Simplified Conditions"). We retained recovery periods as nonstress for comparability with prior stress-detection literature, which commonly assigns labels based on the protocol, and to provide conservative performance estimates. Yet, an alternative approach could either treat recovery as a buffer period excluded from training and evaluation or apply methods such as importance reweighting [110] or weighted surrogate loss [111].

Fifth, while our study uses ECG data to detect MS and downsamples the signal to mimic the hardware constraints of consumer-grade devices, such wearables typically rely on PPG. The most predictive features in our models are derived from R-peaks and beat-to-beat dynamics, which can also be extracted from PPG, suggesting that our approach may be transferable. However, this hypothesis remains to be validated empirically. Future work should therefore assess the extent to which our findings can be generalized to PPG-based settings.

Last, the MS tasks in our experimental setup were performed while participants remained seated. In everyday life, however, individuals often experience MS while standing or moving, for example, a teacher navigating a lively classroom. Future work should therefore extend the evaluation to simultaneous stress-and-activity scenarios to fully validate ambulatory MS detection, as done by Sun et al [63], Hosseini et al [112], and Kaczor et al [113].

Conclusions

This study evaluated ML models for distinguishing MS from routine physical activities using a single-sensor ECG. Both LR and XGBoost achieved acceptable discriminative performance (AUROC >0.7), with XGBoost providing only marginal benefits over the linear BL (Δ<0.02), suggesting that simple, interpretable models can perform competitively. Performance remained robust when downsampling from 1000 to 125 Hz and reducing the feature set to 10, thereby supporting lightweight deployment. However, our findings reveal a critical limitation of single-sensor ECG approaches: although both models detected MS with acceptable to good sensitivity (XGBoost: 0.800; LR: 0.782), specificity was low (XGBoost: 0.512; LR: 0.509), particularly for MPA. This indicates that stress-induced cardiac responses cannot be reliably distinguished from those driven by physical exertion using ECG features alone. Generalization to unseen stressors was also stressor-dependent, with models performing well above chance for most stressors but near chance

XSL•FO
RenderX

for the social-evaluative stimulus (SSST). Future studies should validate our findings in ambulatory settings where MS and physical activity can co-occur and explore both multimodal approaches and the use of PPG via wearables as a practical alternative to ECG for continuous stress monitoring.

Acknowledgments

This article was written as part of the research project “Stress in Action”. We used artificial intelligence tools, specifically ChatGPT and Claude, to speed up coding and debugging. Furthermore, we used these tools to improve the syntax, clarity, and grammar of the written paragraphs. Importantly, all outputs were carefully checked and reviewed.

Funding

Stress in Action is financially supported by the Dutch Research Council and the Dutch Ministry of Education, Culture, and Science (NWO gravitation grant number 024.005.010).

Data Availability

The dataset analyzed in this study is not publicly available, and a data access agreement is required to access and analyze the data. The code is publicly available on the GitHub repository [75].

Authors' Contributions

BU contributed to the conceptualization, formal analysis, coding, methodology, and writing of the manuscript. AA contributed to formal analysis, coding, and investigation. SvdV and DJvdM provided consultation and contributed to the review and editing of the manuscript. EdG and MH provided supervision and also contributed to the review and editing of the manuscript.

Conflicts of Interest

EdG owns shares in VU Ambulatory Monitoring Solutions Ltd. The remaining authors declare no competing interests.

Multimedia Appendix 1
Overview of features.
[PDF File, 53 KB - jmir_v28i1e80450_app1.pdf]

Multimedia Appendix 2
Hyperparameter settings.
[PDF File, 26 KB - jmir_v28i1e80450_app2.pdf]

Multimedia Appendix 3
Boxplots positive and negative affect.
[PDF File, 448 KB - jmir_v28i1e80450_app3.pdf]

Multimedia Appendix 4
Area under the precision-recall curve classification performance.
[PDF File, 133 KB - jmir_v28i1e80450_app4.pdf]

Multimedia Appendix 5
CIs for performance differences between XGBoost and LR.
[PDF File, 40 KB - jmir_v28i1e80450_app5.pdf]

Multimedia Appendix 6
Additional classification performance results.
[PDF File, 685 KB - jmir_v28i1e80450_app6.pdf]

Multimedia Appendix 7
Seated baseline classification performance.
[PDF File, 358 KB - jmir_v28i1e80450_app7.pdf]

Multimedia Appendix 8

XSL•FO
RenderX

Time window analysis.
[PDF File, 339 KB - jmir_v28i1e80450_app8.pdf]

Multimedia Appendix 9
Model explainability.
[PDF File, 764 KB - jmir_v28i1e80450_app9.pdf]

Multimedia Appendix 10
Feature parsimony.
[PDF File, 151 KB - jmir_v28i1e80450_app10.pdf]

Multimedia Appendix 11
Overview of selected features.
[PDF File, 33 KB - jmir_v28i1e80450_app11.pdf]

Multimedia Appendix 12
Generalization to unknown stressors.
[PDF File, 56 KB - jmir_v28i1e80450_app12.pdf]

Multimedia Appendix 13
Stratified stressor analysis.
[PDF File, 47 KB - jmir_v28i1e80450_app13.pdf]

Multimedia Appendix 14
Stratified stressor analysis - single sensor training.
[PDF File, 49 KB - jmir_v28i1e80450_app14.pdf]

References

1. Razavi M, Ziyadidegan S, Mahmoudzadeh A, et al. Machine learning, deep learning, and data preprocessing techniques for detecting, predicting, and monitoring stress and stress-related mental disorders: scoping review. JMIR Ment Health 2024 Aug 21;11:e53714. [doi: 10.2196/53714] [Medline: 39167782]
2. Cohen S, Gianaros PJ, Manuck SB. A stage model of stress and disease. Perspect Psychol Sci 2016 Jul;11(4):456-463. [doi: 10.1177/1745691616646305] [Medline: 27474134]
3. Le Fevre M, Matheny J, Kolt GS. Eustress, distress, and interpretation in occupational stress. J Manag Psychol 2003 Nov 1;18(7):726-744. [doi: 10.1108/02683940310502412]
4. Cohen S, Janicki-Deverts D, Miller GE. Psychological stress and disease. JAMA 2007 Oct 10;298(14):1685-1687. [doi: 10.1001/jama.298.14.1685] [Medline: 17925521]
5. Hammen C. Stress and depression. Annu Rev Clin Psychol 2005;1(293-319):293-319. [doi: 10.1146/annurev.clinpsy.1.102803.143938] [Medline: 17716090]
6. Kivimäki M, Steptoe A. Effects of stress on the development and progression of cardiovascular disease. Nat Rev Cardiol 2018 Apr;15(4):215-229. [doi: 10.1038/nrcardio.2017.189] [Medline: 29213140]
7. Jönsson P, Wallergård M, Osterberg K, Hansen AM, Johansson G, Karlson B. Cardiovascular and cortisol reactivity and habituation to a virtual reality version of the trier social stress test: a pilot study. Psychoneuroendocrinology 2010 Oct;35(9):1397-1403. [doi: 10.1016/j.psyneuen.2010.04.003] [Medline: 20451329]
8. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [Medline: 6668417]
9. Martinez GJ, Grover T, Mattingly SM, et al. Alignment between heart rate variability from fitness trackers and perceived stress: perspectives from a large-scale in situ longitudinal study of information workers. JMIR Hum Factors 2022 Aug 4;9(3):e33754. [doi: 10.2196/33754] [Medline: 35925662]
10. Choi J, Gutierrez-Osuna R. Using heart rate monitors to detect mental stress. 2009 Presented at: Proceedings of the 2009 Sixth International Workshop on Wearable and Implantable Body Sensor Networks; Jun 3-5, 2009.
11. Won E, Kim YK. Stress, the autonomic nervous system, and the immune-kynurenine pathway in the etiology of depression. Curr Neuropharmacol 2016;14(7):665-673. [doi: 10.2174/1570159x14666151208113006] [Medline: 27640517]
12. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science 2015 Jul 17;349(6245):255-260. [doi: 10.1126/science.aaa8415] [Medline: 26185243]

XSL•FO
RenderX

13. Schmidt P, Reiss A, Duerichen R, Marberger C, Van Laerhoven K. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. Presented at: ICMI '18: Proceedings of the 20th ACM International Conference on Multimodal Interaction; Oct 16-20, 2018; Boulder CO USA p. 400-408. [doi: 10.1145/3242969.3242985]
14. Koldijk S, Sappelli M, Verberne S, Neerincx MA, Kraaij W. The swell knowledge work dataset for stress and user modeling research. Presented at: ICMI '14: Proceedings of the 16th International Conference on Multimodal Interaction; Nov 12-16, 2014. [doi: 10.1145/2663204.2663257]
15. Li RA, Liu Z. Stress detection using deep neural networks. BMC Med Inform Decis Mak 2020 Dec;20(S11). [doi: 10.1186/s12911-020-01299-4]
16. Zhu L, Spachos P, Ng PC, et al. Stress detection through wrist-based electrodermal activity monitoring and machine learning. IEEE J Biomed Health Inform 2023;27(5):2155-2165. [doi: 10.1109/JBHI.2023.3239305]
17. Prajod P, Mahesh B, André E. Stressor type matters! --- exploring factors influencing cross-dataset generalizability of physiological stress detection. Presented at: ICMI '24: Proceedings of the 26th International Conference on Multimodal Interaction; Nov 4-8, 2024; San Jose Costa Rica p. 508-517. [doi: 10.1145/3678957.3685738]
18. Mishra V, Sen S, Chen G, et al. Evaluating the reproducibility of physiological stress detection models. Proc ACM Interact Mob Wearable Ubiquitous Technol 2020 Dec;4(4):147. [doi: 10.1145/3432220] [Medline: 36189150]
19. Cho HM, Park H, Dong SY, Youn I. Ambulatory and laboratory stress detection based on raw electrocardiogram signals using a convolutional neural network. Sensors (Basel) 2019 Oct 11;19(20):4408. [doi: 10.3390/s19204408] [Medline: 31614646]
20. van der Mee DJ, Gevonden MJ, Westerink J, de Geus EJC. Validity of electrodermal activity-based measures of sympathetic nervous system activity from a wrist-worn device. Int J Psychophysiol 2021 Oct;168:52-64. [doi: 10.1016/j.ijpsycho.2021.08.003] [Medline: 34418464]
21. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC Med 2015 Jan 6;13(1):1. [doi: 10.1186/s12916-014-0241-z] [Medline: 25563062]
22. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378. [doi: 10.1136/bmj-2023-078378]
23. Myin-Germeys I, van Os J, Schwartz JE, Stone AA, Delespaul PA. Emotional reactivity to daily life stress in psychosis. Arch Gen Psychiatry 2001 Dec;58(12):1137-1144. [doi: 10.1001/archpsyc.58.12.1137] [Medline: 11735842]
24. Makowski D, Pham T, Lau ZJ, et al. NeuroKit2: a Python toolbox for neurophysiological signal processing. Behav Res Methods 2021 Aug;53(4):1689-1696. [doi: 10.3758/s13428-020-01516-y] [Medline: 33528817]
25. Bernardes A, Couceiro R, Medeiros J, et al. How reliable are ultra-short-term HRV measurements during cognitively demanding tasks? Sensors (Basel) 2022 Aug 30;22(17):6528. [doi: 10.3390/s22176528] [Medline: 36080987]
26. Smets E, Casale P, Großekathöfer U, Lamichhane B, Raedt W, Bogaerts K. Comparison of Machine Learning Techniques for Psychophysiological Stress Detection Pervasive Computing Paradigms for Mental Health: Springer International Publishing; 2015. [doi: 10.1007/978-3-319-32270-4_2]
27. Kaegi DM, Halamek LP, Van Hare GF, Howard SK, Dubin AM. Effect of mental stress on heart rate variability: validation of simulated operating and delivery room training modules. Pediatr Res 1999 Apr;45(4, Part 2 of 2):77A-77A. [doi: 10.1203/00006450-199904020-00463]
28. Kim HG, Cheon EJ, Bai DS, Lee YH, Koo BH. Stress and heart rate variability: a meta-analysis and review of the literature. Psychiatry Investig 2018 Mar;15(3):235-245. [doi: 10.30773/pi.2017.08.17]
29. Kleiger RE, Stein PK, Bosner MS, Rottman JN. Time domain measurements of heart rate variability. Cardiol Clin 1992 Aug;10(3):487-498. [doi: 10.1016/s0733-8651(18)30230-3] [Medline: 1504980]
30. Dalmeida KM, Masala GL. HRV features as viable physiological markers for stress detection using wearable devices. Sensors (Basel) 2021 Apr 19;21(8):2873. [doi: 10.3390/s21082873] [Medline: 33921884]
31. Wang HM, Huang SC. SDNN/RMSSD as a surrogate for LF/HF: a revised investigation. Model Simul Eng 2012;2012:1-8. [doi: 10.1155/2012/931943]
32. Sollers JJ 3rd, Buchanan TW, Mowrer SM, Hill LK, Thayer JF. Comparison of the ratio of the standard deviation of the R-R interval and the root mean squared successive differences (SD/rMSSD) to the low frequency-to-high frequency (LF/HF) ratio in a patient population and normal healthy controls. Biomed Sci Instrum 2007;43:158-163. [Medline: 17487074]
33. Salahuddin L, Cho J, Jeong MG, Kim D. Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. Annu Int Conf IEEE Eng Med Biol Soc 2007;2007:4656-4659. [doi: 10.1109/IEMBS.2007.4353378] [Medline: 18003044]
34. Pereira T, Almeida PR, Cunha JPS, Aguiar A. Heart rate variability metrics for fine-grained stress level assessment. Comput Methods Programs Biomed 2017 Sep;148:71-80. [doi: 10.1016/j.cmpb.2017.06.018] [Medline: 28774440]
35. Tipton MJ, Harper A, Paton JFR, Costello JT. The human ventilatory response to stress: rate or depth? J Physiol (Lond) 2017 Sep;595(17):5729-5752. [doi: 10.1113/JP274596]
36. Masaoka Y, Homma I. Anxiety and respiratory patterns: their relationship during mental stress and physical load. Int J Psychophysiol 1997 Sep;27(2):153-159. [doi: 10.1016/s0167-8760(97)00052-4] [Medline: 9342646]

XSL•FO
RenderX

37. Masaoka Y, Homma I. The effect of anticipatory anxiety on breathing and metabolism in humans. Respir Physiol 2001 Nov 1;128(2):171-177. [doi: 10.1016/s0034-5687(01)00278-x] [Medline: 11812381]
38. Hernando A, Lazaro J, Gil E, et al. Inclusion of respiratory frequency information in heart rate variability analysis for stress assessment. IEEE J Biomed Health Inform 2016 Jul;20(4):1016-1025. [doi: 10.1109/JBHI.2016.2553578] [Medline: 27093713]
39. Karthikeyan P, Murugappan M, Yaacob S. A study on mental arithmetic task based human stress level classification using discrete wavelet transform. Presented at: 2012 IEEE conference on sustainable utilization and development in engineering and technology (STUDENT2012); Oct 6-9, 2012. [doi: 10.1109/STUDENT.2012.6408369]
40. Webber CL Jr, Zbilut JP. Dynamical assessment of physiological systems and states using recurrence plot strategies. J Appl Physiol 1994 Feb 1;76(2):965-973. [doi: 10.1152/jappl.1994.76.2.965]
41. Tanev G, Saadi DB, Hoppe K, Sorensen HBD. Classification of acute stress using linear and non-linear heart rate variability analysis derived from sternal ECG. Presented at: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); Aug 26-30, 2014. [doi: 10.1109/EMBC.2014.6944349]
42. Pincus SM. Approximate entropy as a measure of system complexity. Proc Natl Acad Sci USA 1991 Mar 15;88(6):2297-2301. [doi: 10.1073/pnas.88.6.2297]
43. Richman JS, Moorman JR. Physiological time-series analysis using approximate entropy and sample entropy. Am J Physiol Heart Circ Physiol 2000 Jun;278(6):H2039-H2049. [doi: 10.1152/ajpheart.2000.278.6.H2039] [Medline: 10843903]
44. Chen W, Wang Z, Xie H, Yu W. Characterization of surface EMG signal based on fuzzy entropy. IEEE Trans Neural Syst Rehabil Eng 2007 Jun;15(2):266-272. [doi: 10.1109/TNSRE.2007.897025] [Medline: 17601197]
45. Pham T, Lau ZJ, Chen SHA, Makowski D. Heart rate variability in psychology: a review of HRV indices and an analysis tutorial. Sensors (Basel) 2021 Jun 9;21(12):3998. [doi: 10.3390/s21123998] [Medline: 34207927]
46. Humeau-Heurtier A. Evaluation of systems' irregularity and complexity: sample entropy, its derivatives, and their applications across scales and disciplines. Entropy (Basel) 2018 Oct 16;20(10):794. [doi: 10.3390/e20100794] [Medline: 33265881]
47. Parvaneh S, Toosizadeh N, Moharreri S. Impact of mental stress on heart rate asymmetry. In: Parvaneh S, Toosizadeh N, Moharreri S, editors. Presented at: 2015 Computing in Cardiology Conference (CinC); Sep 6-9, 2015. [doi: 10.1109/CIC.2015.7411030]
48. Yan C, Li P, Ji L, Yao L, Karmakar C, Liu C. Area asymmetry of heart rate variability signal. Biomed Eng Online 2017 Sep 21;16(1):112. [doi: 10.1186/s12938-017-0402-3] [Medline: 28934961]
49. Javorka M, Turianikova Z, Tonhajzerova I, Javorka K, Baumert M. The effect of orthostasis on recurrence quantification analysis of heart rate and blood pressure dynamics. Physiol Meas 2009 Jan;30(1):29-41. [doi: 10.1088/0967-3334/30/1/003] [Medline: 19039163]
50. Dimitriev D, Saperova EV, Dimitriev A, Karpenko Y. Recurrence quantification analysis of heart rate during mental arithmetic stress in young females. Front Physiol 2020;11(40):40. [doi: 10.3389/fphys.2020.00040] [Medline: 32116754]
51. Peng CK, Havlin S, Stanley HE, Goldberger AL. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos 1995 Mar 1;5(1):82-87. [doi: 10.1063/1.166141]
52. Tulppo MP, Mäkikallio TH, Seppänen T, et al. Effects of pharmacological adrenergic and vagal modulation on fractal heart rate dynamics. Clin Physiol 2001 Sep 24;21(5):515-523. [doi: 10.1046/j.1365-2281.2001.00344.x]
53. Immanuel S, Teferra MN, Baumert M, Bidargaddi N. Heart rate variability for evaluating psychological stress changes in healthy adults: a scoping review. Neuropsychobiology 2023;82(4):187-202. [doi: 10.1159/000530376] [Medline: 37290411]
54. Goya-Esteban R, Barquero-Pérez O, Sarabia-Cachadina E, Naranjo-Orellana J, Rojo-Alvarez JL. Heart rate variability non linear dynamics in intense exercise. 2010 Presented at: Computing in Cardiology; Sep 9-12, 2012.
55. Costa MD, Davis RB, Goldberger AL. Heart rate fragmentation: a new approach to the analysis of cardiac interbeat interval dynamics. Front Physiol 2017 May;8. [doi: 10.3389/fphys.2017.00255]
56. Lampert R. ECG signatures of psychological stress. J Electrocardiol 2015;48(6):1000-1005. [doi: 10.1016/j.jelectrocard.2015.08.005] [Medline: 26364755]
57. Aro AL, Kenttä TV, Huikuri HV. Microvolt T-wave alternans: where are we now? Arrhythm Electrophysiol Rev 2016 May;5(1):37-40. [doi: 10.15420/aer.2015.28.1] [Medline: 27403292]
58. Nearing BD, Verrier RL. Modified moving average analysis of T-wave alternans to predict ventricular fibrillation with high accuracy. J Appl Physiol (1985) 2002 Feb;92(2):541-549. [doi: 10.1152/japplphysiol.00592.2001] [Medline: 11796662]
59. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13-17, 2016; San Francisco California USA p. 785-794.
60. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310. [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
61. Chaptoukaev H, Strizhkova V, Panariello M, et al. StressID: a multimodal dataset for stress identification. Presented at: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track; Dec 10-16, 2023.
62. Heimerl A, Prajod P, Mertes S, et al. The ForDigitStress dataset: a multi-modal dataset for automatic stress recognition. IEEE Trans Affective Comput ;16(2):1219-1234. [doi: 10.1109/TAFFC.2024.3501400]

XSL•FO
RenderX

63. Sun FT, Kuo C, Cheng HT, Buthpitiya S, Collins P, Griss M, editors. Activity-aware mental stress detection using physiological sensors. Presented at: Mobile Computing, Applications, and Services MobiCASE 2010; Oct 25-28, 2010.
64. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data. Presented at: Proceedings of the 36th International Conference on Neural Information Processing Systems; Nov 28 to Dec 3, 2022; New Orleans, Louisiana, United States p. 507-520.
65. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. Information Fusion 2022 May;81(C):84-90. [doi: 10.1016/j.inffus.2021.11.011]
66. Schallmoser S, Zueger T, Kraus M, Saar-Tsechansky M, Stettler C, Feuerriegel S. Machine learning for predicting micro- and macrovascular complications in individuals with prediabetes or diabetes: retrospective cohort study. J Med Internet Res 2023;25:e42181. [doi: 10.2196/42181]
67. Healey JA, Picard RW. Detecting stress during real-world driving tasks using physiological sensors. IEEE Trans Intell Transport Syst 2005 Jun;6(2):156-166. [doi: 10.1109/TITS.2005.848368]
68. Saygin M, Schoenmakers M, Gevonden M, de Geus E. Speech detection via respiratory inductance plethysmography, thoracic impedance, accelerometers, and gyroscopes: a machine learning-informed comparative study. Psychophysiology 2025 Feb;62(2):e70021. [doi: 10.1111/psyp.70021] [Medline: 39950497]
69. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Presented at: Proceedings of the 31st International Conference on Neural Information Processing Systems; Dec 4-9, 2017; Long Beach, California, United States p. 4768-4777.
70. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. JAMA 2017 Oct 10;318(14):1377-1384. [doi: 10.1001/jama.2017.12126] [Medline: 29049590]
71. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. J Am Med Inform Assoc 2020 Apr 1;27(4):621-633. [doi: 10.1093/jamia/ocz228] [Medline: 32106284]
72. Brier GW. Verification of forecasts expressed in terms of probability. Mon Wea Rev 1950 Jan;78(1):1-3. [doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2]
73. Maiga A, Farjah F, Blume J, et al. Risk prediction in clinical practice: a practical guide for cardiothoracic surgeons. Ann Thorac Surg 2019 Nov;108(5):1573-1582. [doi: 10.1016/j.athoracsur.2019.04.126] [Medline: 31255609]
74. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. Lond Edinb Dubl Phil Mag J Sci 1901 Nov;2(11):559-572. [doi: 10.1080/14786440109462720]
75. Uendes B, Antonides A. Stress detection ECG. GitHub. 2025 Jul. URL: https://github.com/BuelentUendes/Stress_Detection_ECG [accessed 2025-07-11]
76. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16(1):321-357. [doi: 10.1613/jair.953]
77. Your research: subject to the WMO or not? Central Committee on Research Involving Human Subjects. URL: https://english.ccmo.nl/investigators/legal-framework-for-medical-scientific-research/your-research-is-it-subject-to-the-wmo-or-not [accessed 2026-03-22]
78. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988 Sep;44(3):837-845. [doi: 10.2307/2531595] [Medline: 3203132]
79. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. Biometrics 1997 Jun;53(2):567-578. [doi: 10.2307/2533958] [Medline: 9192452]
80. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Statist Sci 2001 Aug;16(3):199-231. [doi: 10.1214/ss/1009213726]
81. Rudin C, Zhong C, Semenova L, Seltzer M, Parr R, Liu J, et al. Position: amazing things come from having many good models. In: Ruslan S, Zico K, Katherine H, Adrian W, Nuria O, Jonathan S, editors. Presented at: Proceedings of the 41st International Conference on Machine Learning; Jul 21-27, 2024; Vienna, Austria p. 42783-42795.
82. Semenova L, Chen H, Parr R, Rudin C. A path to simpler models starts with noise. Adv Neural Inf Process Syst 2023 Dec;36:3362-3401. [Medline: 38577617]
83. Hwang B, You J, Vaessen T, Myin-Germeys I, Park C, Zhang BT. Deep ECGNet: an optimal deep learning framework for monitoring mental stress using ultra short-term ECG signals. Telemed J E Health 2018 Oct;24(10):753-772. [doi: 10.1089/tmj.2017.0250] [Medline: 29420125]
84. Zhang P, Li F, Zhao R, et al. Real-time psychological stress detection according to ECG using deep learning. Appl Sci (Basel) 2021;11(9):3838. [doi: 10.3390/app11093838]
85. Crabbé J, Schaar MVD. Explaining time series predictions with dynamic masks. In: Marina M, Tong Z, editors. Presented at: Proceedings of the 38th International Conference on Machine Learning; Jul 18-24, 2021; Virtual / Online p. 2166-2177.
86. Enguehard J. Learning perturbations to explain time series predictions. In: Andreas K, Emma B, Kyunghyun C, Barbara E, Sivan S, Jonathan S, editors. Presented at: Proceedings of the 40th International Conference on Machine Learning; Jul 23-29, 2023; Honolulu, Hawaii, United States p. 9329-9342.
87. Chand K, Chandra S, Dutt V. A comprehensive evaluation of linear and non-linear HRV parameters between paced breathing and stressful mental state. Heliyon 2024 Jun;10(11):e32195. [doi: 10.1016/j.heliyon.2024.e32195]

XSL•FO
RenderX

88. Taoum A, Bisiaux A, Tilquin F, Le Guillou Y, Carrault G. Validity of ultra-short-term HRV analysis using PPG—a preliminary study. Sensors (Basel) 2022 Oct 20;22(20):7995. [doi: 10.3390/s22207995]
89. Maeda Y, Sekine M, Tamura T. Relationship between measurement site and motion artifacts in wearable reflected photoplethysmography. J Med Syst 2011 Oct;35(5):969-976. [doi: 10.1007/s10916-010-9505-0]
90. Gedam S, Paul S. A review on mental stress detection using wearable sensors and machine learning techniques. IEEE Access 2021;9:84045-84066. [doi: 10.1109/ACCESS.2021.3085502]
91. Sharma N, Gedeon T. Objective measures, sensors and computational techniques for stress recognition and classification: a survey. Comput Methods Programs Biomed 2012 Dec;108(3):1287-1301. [doi: 10.1016/j.cmpb.2012.07.003] [Medline: 22921417]
92. Can YS, Arnrich B, Ersoy C. Stress detection in daily life scenarios using smart phones and wearable sensors: a survey. J Biomed Inform 2019 Apr;92:103139. [doi: 10.1016/j.jbi.2019.103139]
93. Pataca AO, Zdravevski E, Coelho PJ, et al. Use of machine learning for predicting stress episodes based on wearable sensor data: a systematic review. Comput Biol Med 2025 Nov;198:111166. [doi: 10.1016/j.compbiomed.2025.111166]
94. Kirschbaum C, Pirke KM, Hellhammer DH. The 'Trier Social Stress Test' – a tool for investigating psychobiological stress responses in a laboratory setting. Neuropsychobiology 1993;28(1-2):76-81. [doi: 10.1159/000119004]
95. Stroop JR. Studies of interference in serial verbal reactions. J Exp Psychol 1935;18(6):643-662. [doi: 10.1037/h0054651]
96. Dickerson SS, Kemeny ME. Acute stressors and cortisol responses: a theoretical integration and synthesis of laboratory research. Psychol Bull 2004 May;130(3):355-391. [doi: 10.1037/0033-2909.130.3.355]
97. Epel ES, Crosswell AD, Mayer SE, et al. More than a feeling: a unified view of stress measurement for population science. Front Neuroendocrinol 2018 Apr;49:146-169. [doi: 10.1016/j.yfrne.2018.03.001] [Medline: 29551356]
98. Vos G, Trinh K, Sarnyai Z, Rahimi Azghadi M. Generalizable machine learning for stress monitoring from wearable devices: a systematic literature review. Int J Med Inform 2023 May;173:105026. [doi: 10.1016/j.ijmedinf.2023.105026] [Medline: 36893657]
99. Smets E, Rios Velazquez E, Schiavone G, et al. Large-scale wearable data reveal digital phenotypes for daily-life stress detection. npj Digital Med 2018;1(1):67. [doi: 10.1038/s41746-018-0074-9]
100. Huang S, Li J, Zhang P, Zhang W. Detection of mental fatigue state with wearable ECG devices. Int J Med Inform 2018 Nov;119:39-46. [doi: 10.1016/j.ijmedinf.2018.08.010]
101. Han HJ, Labbaf S, Borelli JL, Dutt N, Rahmani AM. Objective stress monitoring based on wearable sensors in everyday settings. J Med Eng Technol 2020 May;44(4):177-189. [doi: 10.1080/03091902.2020.1759707] [Medline: 32589065]
102. Weber C, Quintus M, Egloff B, Luong G, Riediger M, Wrzus C. Same old, same old? age differences in the diversity of daily life. Psychol Aging 2020 May;35(3):434-448. [doi: 10.1037/pag0000407] [Medline: 31613136]
103. Umetani K, Singer DH, McCraty R, Atkinson M. Twenty-four hour time domain heart rate variability and heart rate: relations to age and gender over nine decades. J Am Coll Cardiol 1998 Mar 1;31(3):593-601. [doi: 10.1016/s0735-1097(97)00554-8] [Medline: 9502641]
104. Wilder J. The law of initial value in neurology and psychiatry. J Nerv Ment Dis 1957 Jan;125(1):73-86. [doi: 10.1097/00005053-195701000-00009]
105. Benchekroun M, Velmovitsky PE, Istrate D, Zalc V, Morita PP, Lenne D. Cross dataset analysis for generalizability of HRV-based stress detection models. Sensors (Basel) 2023 Feb 6;23(4):1807. [doi: 10.3390/s23041807] [Medline: 36850407]
106. Karimi D, Dou H, Warfield SK, Gholipour A. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. Med Image Anal 2020 Oct;65:101759. [doi: 10.1016/j.media.2020.101759] [Medline: 32623277]
107. Ding C, Pereira T, Xiao R, Lee RJ, Hu X. Impact of label noise on the learning based models for a binary classification of physiological signal. Sensors (Basel) 2022;22(19):7166. [doi: 10.3390/s22197166]
108. Song H, Kim M, Park D, Shin Y, Lee JG. Learning from noisy labels with deep neural networks: a survey. IEEE Trans Neural Netw Learning Syst 2023;34(11):8135-8153. [doi: 10.1109/TNNLS.2022.3152527]
109. Frénay B, Verleysen M. Classification in the presence of label noise: a survey. IEEE Trans Neural Netw Learn Syst 2014 May;25(5):845-869. [doi: 10.1109/TNNLS.2013.2292894] [Medline: 24808033]
110. Liu T, Tao D. Classification with noisy labels by importance reweighting. IEEE Trans Pattern Anal Mach Intell 2016;38(3):447-461. [doi: 10.1109/TPAMI.2015.2456899]
111. Natarajan N, Dhillon IS, Ravikumar P, Tewari A. Learning with noisy labels. Presented at: Proceedings of the 27th International Conference on Neural Information Processing Systems; Dec 5-10, 2013; Lake Tahoe, Nevada, United States p. 1196-1204.
112. Hosseini S, Gottumukkala R, Katragadda S, et al. A multimodal sensor dataset for continuous stress detection of nurses in a hospital. Sci Data 2022 Jun 1;9(1):255. [doi: 10.1038/s41597-022-01361-y] [Medline: 35650267]
113. Kaczor EE, Carreiro S, Stapp J, Chapman B, Indic P. Objective measurement of physician stress in the emergency department using a wearable sensor. Proc Annu Hawaii Int Conf Syst Sci 2020;2020:3729-3738. [Medline: 32015695]

Abbreviations

ANS: autonomic nervous system

XSL•FO
RenderX

AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic
AVNN: average normal-to-normal
BL: baseline
BPM: beats per minute
ECG: electrocardiogram
HF: high frequency
HR: heart rate
HRA: heart rate asymmetry
HRV: heart rate variability
Hz: hertz
IALS: inverse average length of segments
LPA: low physical activity
LR: logistic regression
ML: machine learning
MMA: modified moving average
MPA: moderate physical activity
MS: mental stress
NN: normal-to-normal
PASAT: paced auditory serial addition task
PPG: photoplethysmography
PSS: percentage of short segments
RAVEN: Raven's Progressive Matrices
RF: random forest
RMSSD: root-mean-square of successive differences
SDNN: SD of the NN intervals
SHAP: Shapley additive explanation
SMOTE: synthetic minority oversampling technique
SQI: signal quality index
SSST: sing-a-song-stress test
TA: tone avoidance
TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Artificial Intelligence
TWA: T-wave alternans
UHF: ultra-high frequency
VHF: very high frequency
XGBoost: extreme gradient boosting

Edited by A Coristine; submitted 10.Jul.2025; peer-reviewed by J Li, SM Shaffi; revised version received 22.Jan.2026; accepted 23.Jan.2026; published 07.Apr.2026.

Please cite as:
Uendes B, Antonides A, van de Ven S, van der Mee DJ, de Geus E, Hoogendoorn M
Electrocardiogram-Based Mental Stress Detection Amid Everyday Activities Using Machine Learning: Model Development and Validation Study
J Med Internet Res 2026;28:e80450
URL: https://www.jmir.org/2026/1/e80450
doi:10.2196/80450

© Buelent Uendes, Alex Antonides, Sjors van de Ven, Denise Johanna van der Mee, Eco de Geus, Mark Hoogendoorn. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 7.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Examining the Association Between Internet Use and Perceived Stress in Adults: Longitudinal Observational Study Combining Web Tracking Data With Questionnaires

Mohammad Belal, MCA; Nguyen Luong, MSc; Talayeh Aledavood, PhD; Juhi Kulshrestha, DRING

Department of Computer Science, Aalto University, PL 1400, Konemiehentie 2, Aalto, Finland

Corresponding Author:
Mohammad Belal, MCA
Department of Computer Science, Aalto University, PL 1400, Konemiehentie 2, Aalto, Finland

Abstract

Background: In today's digital era, the internet plays a pervasive role in daily life, influencing everyday activities such as communication, work, and leisure. This online engagement intertwines with offline experiences, shaping individuals' overall well-being. Despite its significance, existing research often falls short in capturing the relationship between internet use and well-being, relying primarily on isolated studies and self-reported data. One major contributor to deteriorated well-being is stress. While some research has examined the relationship between internet use and stress, both positive and negative associations have been reported.

Objective: This study aimed to identify the associations between an individual's internet use and their stress.

Methods: We conducted a 7-month longitudinal study. We combined fine-grained URL-level web browsing traces of 1490 German internet users with their sociodemographics and monthly measures of stress. Further, we developed a conceptual framework that allows us to simultaneously explore different contextual dimensions, including how, where, when, and by whom the internet is used. We applied linear mixed-effects models to examine these associations.

Results: Our analysis revealed several associations between internet use and stress, varying by context. Increased time spent on social media, online shopping, and gaming platforms was associated with higher stress. For example, the time spent by individuals on shopping-related internet use (aggregated over the 30 days before their stress was measured via questionnaires) was positively associated with stress on both mobile (β=.04, 95% CI 0.00 - 0.08; *P*=.04) and desktop devices (β=.03, 95% CI −0.00 to 0.06; *P*=.09). In contrast, time spent on productivity or news websites was associated with lower stress. Specifically, in the last 30 days of mobile usage, productivity-related use showed a negative association with stress (β=−.03, 95% CI −0.06 to −0.00; *P*=.04). In addition, in the last 2 days of data, news usage was negatively associated with stress on both mobile (β=−.54, 95% CI −1.08 to 0.00; *P*=.048) and desktop devices (β=−.50, 95% CI−0.90 to −0.11; *P*=.01). Further analysis showed that total time spent online (β=.01, 95% CI 0.00 - 0.02; *P*<.001), social-media usage (β=.02, 95% CI 0.00 - 0.03; *P*=.02), and gaming usage (β=.01, 95% CI 0.00 - 0.02; *P*=.02) were all positively associated with stress in high-stress Perceived Stress Scale (PSS>26) individuals on mobile devices.

Conclusions: The findings indicate that internet use is associated with stress, and these associations differ across various usage contexts. In the future, the behavioral markers we identified can pave the way for designing individualized tools for people to self-monitor and self-moderate their online behaviors to enhance their well-being, reducing the burden on already overburdened mental health services.

(J Med Internet Res 2026;28:e78775) doi:10.2196/78775

KEYWORDS

online behavior; stress; internet use; web browsing traces; sociodemographic differences; longitudinal design

Introduction

Stress is an unavoidable part of human life, arising from the demands and challenges we face daily. It is a significant factor in health issues such as cardiovascular disease, weakened immune function, and mental health challenges [1,2]. The internet, now an integral part of modern life, has sparked debates about its impact on stress levels and psychological well-being [3,4], as well as whether this influence is predominantly positive or negative. As our online and offline lives become increasingly interconnected, understanding the relationship between internet use and stress has gained considerable attention.

While the internet offers numerous advantages, such as enhanced connectivity and easy access to information, excessive or problematic use has been linked to various stress-related factors [3,5,6]. For example, heavy internet use has been associated

XSL•FO
RenderX

with higher levels of anxiety [5], while the amount of time spent online has been linked to sleep loss and withdrawal [6]. On the other hand, past research suggests that not all forms of internet use are detrimental; certain online activities have been associated with reduced stress and improved psychological well-being [7-9].

Despite the internet's widespread influence, research on psychological well-being, including stress, has primarily focused on offline activities, leaving a critical gap in understanding how online behaviors impact stress and well-being [10]. For instance, a comprehensive review of 99 commonly used psychological well-being scales identified 196 dimensions, yet none explicitly addressed online activities or behaviors [10]. Moreover, studies on online engagement have often faced limitations, including short study durations, small sample sizes, and an over-reliance on questionnaires to capture internet use patterns. These approaches can introduce biases and fail to provide a complete picture of the connection between online and offline experiences [11,12].

In this paper, we first review previous research on the association between internet use and stress and examine the methodologies used in these studies. We then outline our longitudinal multimodal study design, which integrates actual internet usage data with monthly questionnaires to measure stress, and discuss our study's potential impact.

Conflicting Findings on Associations Between Stress and Internet Use

The relationship between internet use and stress is complex, with previous research showing contrasting associations depending on the type and context of digital engagement, as well as individual characteristics. High levels of internet and smartphone use have been linked to increased stress [13,14], often due to digital overload (ie, the cognitive strain caused by constant notifications and an endless stream of information) [15,16]. In contrast, internet use through computers has been associated with less burnout compared to smartphone use [17]. However, these associations are not consistent. For instance, age influences the impact of digital multitasking: younger users report higher stress than older adults when handling multiple digital tasks, yet they appear less affected by communication overload [15]. Experimental evidence shows that multitasking increases perceived stress levels in both younger and older adults, with no significant differences between the age groups [18]. Other studies have shown no association [19] or even a negative association between time spent online and stress, particularly in young adults [20].

Moreover, the type of online activity plays a crucial role in stress outcomes. Social networking and entertainment-related use have been associated with higher stress levels, while internet use for work-related tasks has been linked to greater life satisfaction in the middle-aged population [21]. Research also indicates that communication overload from emails and messages is positively associated with perceived stress in the age group of 50 - 85 years [15]. Studies on social media show similarly nuanced findings. While Pew Research found no association between social media use and stress in men, a negative association was observed in women [22]. A large-scale study also showed slightly higher perceived stress among high social media users than nonusers [23]. Other digital behaviors, such as problematic news consumption [24] and adult content addiction [25], have also been linked to heightened stress and emotional distress. Similarly, concerns such as cyberbullying, online harassment [26], work-life boundary erosion [27], and data privacy issues [28] have been widely documented as stress-inducing. Further studies have found positive associations between stress and various digital behaviors, such as online shopping addiction in young people [29], negative information seeking [30], interpersonal communication in older adults [7], misinformation sharing [31], and excessive gaming in adolescents [32,33].

Conversely, the internet can also act as a buffer against stress [20], offering access to supportive communities [9], relaxation tools, and leisure activities [7]. Online entertainment and social interaction, in particular, have been shown to reduce stress and enhance well-being among older adults [7,8,34]. In addition, internet use has been recognized as a coping mechanism. Several online activities, including social media [35], entertainment [36,37], shopping [38], and gaming [35,39], have been identified in previous studies as strategies for managing stress.

User Characteristics Shape the Relationship Between Online Activity and Stress

Studies show that age, gender, income, and baseline stress levels influence how online activities relate to stress [15,18,22]. Social media use has been negatively associated with stress among females [22], though females overall report higher stress than males [40,41]. Older adults tend to report lower stress than younger groups but experience stronger associations between communication load, frequent messaging, and stress [42,43]. Higher income is generally linked to lower stress levels [44], though findings differ by context—for instance, higher income was associated with fewer mental health issues in Germany but with more issues in China [45]. Social media use has also been linked to slower recovery from real-world stressors, suggesting possible stress maintenance in already stressed individuals [46].

Methods for Identifying Connections Between Internet Use and Stress

Past research on internet use and stress has used various methodologies. Many studies rely on cross-sectional designs and self-reported surveys [7,15,16,21,24,32]. These studies often focus on specific populations, such as university or medical students [13,14]. Some have used larger samples [23,24,32] but still rely on questionnaires to capture internet use. A smaller number of experimental studies are available [30,46], and some have adopted alternative approaches, such as analyzing social media data to infer psychological states [31]. Studies using actual web browsing data [20,47] are limited and tend to capture general metrics, such as total time spent online [20], and are based on relatively small samples (92 and 107 participants), indicating how difficult it is to conduct such studies.

Contributions and Impact of Our Study

Previous research on the relationship between internet use and stress has shown mixed findings, revealing both negative and

XSL•FO
RenderX

positive associations depending on the type of internet activity. However, much of this evidence remains fragmented, as previous studies have largely relied on self-reported internet use data (which often lacks granularity) and have focused on limited aspects of internet use.

To address these limitations and to provide a more comprehensive understanding of how internet use relates to stress, we conducted a longitudinal multimodal study involving 1490 internet users in Germany over 7 months. Our study integrates fine-grained, passively collected web trace data from both desktop and mobile devices with participants' monthly responses to a validated stress scale (refer to Figure 1). Using objective behavioral data, we move beyond subjective self-reports and introduce a data-driven framework for revealing long-term usage patterns and identifying digital markers of stress.

Figure 1. Overview of the study design and contextual dimensions. The top panel shows our longitudinal study design combining desktop and mobile web - trace data with monthly stress questionnaires. The middle panel depicts the internet use and well-being features extracted from web browsing traces and monthly questionnaires. The bottom panel shows the contextual dimensions we consider for examining associations between internet use and stress.

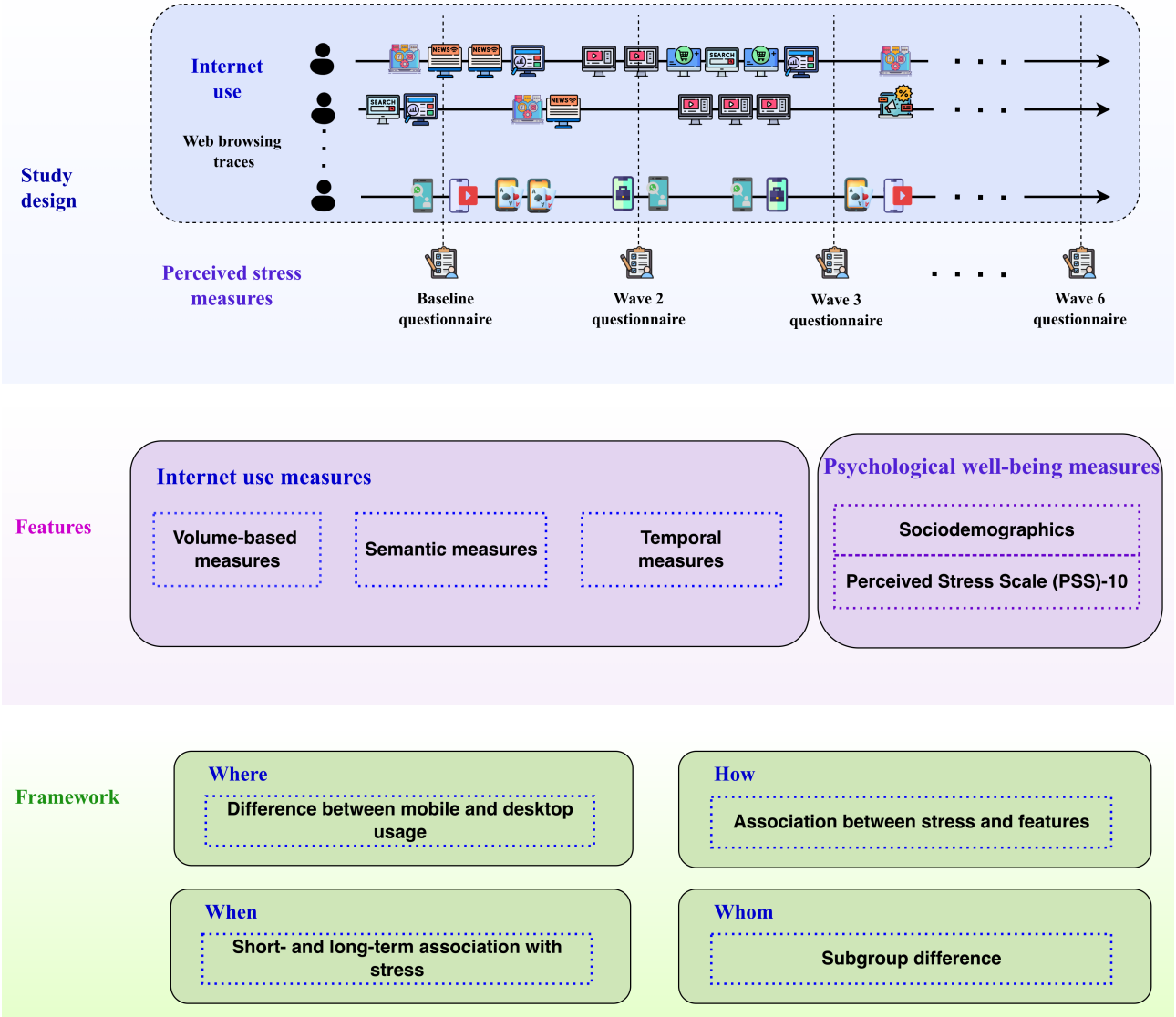

Building on existing research, we further identify 4 key dimensions that shape the relationship between digital behavior and stress:

1. How: the type and pattern of internet behaviors—such as usage volume [14,48], temporal rhythms [49], and content categories (eg, social media, productivity, entertainment, or shopping)—may influence stress in different ways [21,50-52].
2. Where: the device context plays a significant role. Previous research suggests that desktop use is often more goal-oriented and structured, whereas mobile use tends to be more fragmented and reactive [53].
3. When: the timing of online behaviors in relation to stress assessments is important, as short-term engagement with digital content may have immediate effects on stress responses [54,55].
4. By whom: individual differences—such as age, gender, income, and baseline stress levels—can moderate the impact of digital engagement on stress, with some groups being more susceptible than others [40-44].

XSL•FO
RenderX

By adopting this multidimensional perspective, our study seeks to bring coherence to the scattered evidence in existing research and provide a more thorough understanding of the link between digital behaviors and stress. In addition, identifying behavioral markers of stress in internet use may inform the design of future tools for real-time stress monitoring, complementing traditional self-reported measures. This approach contributes to a deeper understanding of digital well-being and supports the development of targeted interventions for healthier online behaviors.

Methods

In this section, we describe our study design, participants, collected data, extracted features, and analysis models that allowed us to overcome the challenges of previous work described in the "Contributions and Impact of Our Study" section.

Study Design

We conducted a longitudinal multimodal study over 7 months that combined passively collected fine-grained web browsing traces with repeated monthly online questionnaires (as provided in Figure 1). The web browsing traces for desktop users included URL-level traces, while for mobile users, both URL-level and application-level traces were included (throughout the paper, we will use "app" to mean mobile app). We measured the perceived stress of our panelists using the validated Perceived Stress Scale (PSS)-10 in 6 monthly waves. In the first wave, we also collected the sociodemographic characteristics of the panelists, including age, gender, and income.

Participants

The study was conducted on a sample of German internet users, recruited through a General Data Protection Regulation–compliant panel company (Bilendi GmbH), which provided access to participants who had already installed tracking software on their devices to capture their internet use. The company also managed survey coding and distribution, sending email notifications to relevant participants each time a survey was launched. Participation was voluntary; panelists were informed about the study and chose to take part. Compensation ranged from €1 - €3 (US $1.10-US $3.30) US $3.30) per month, depending on the number of devices tracked, plus an additional amount based on the company's standard rate (€6/h or US $6.60/h) for each survey completion. All currency values in this study are reported in euros. The exchange rate at the time of the study was €1=US $1.10. All panelists from the company were invited via email for the first wave of questionnaires, yielding 1490 completed responses. In the subsequent 5 waves, which were approximately 1 month apart, all these 1490 respondents were invited via email to participate in each wave. Across all waves, the average range between the earliest and latest survey completion dates was 15 days, and the average completion time for the baseline survey was about 37 minutes (90% CI 32.7-42.3), while for the remaining 5 waves, it was approximately 19.6 minutes (90% CI 17.8-21.5). Table 1 reports the number of participants with completed responses for each wave that we retained for further analysis. We excluded 1 panelist who reported their gender as nonbinary and 52 panelists who reported their income as "other," because these categories had too few respondents. The sample for Wave 1 is therefore 1437. We observe that about 23% of the panelists dropped out by the sixth wave. In Table 1, we also depict the distribution of the panelists across age, gender, and income for the 6 waves of questionnaires.

Next, we examined how closely our sample's sociodemographic distributions match the German population margins for gender, age, and income (provided in Table 2). We observe that our sample's gender distribution matches closely with that of the German population (Destatis [56]). However, for both age and income, the middle ranges are overrepresented in our sample, while the extremes are underrepresented.

XSL•FO
RenderX

Table . Descriptive characteristics of participants across 6 survey waves. The table presents the number of participants, gender distribution, age groups, income categories, and mean perceived stress scores (with SDs) for each wave. Percentages are shown in parentheses for categorical variables.

Wave	1	2	3	4	5	6
Participants (n)	1437	1314	1212	1198	1205	1107
Gender, n (%)						
Male	738 (51.35)	688 (52.36)	639 (52.72)	635 (53.01)	628 (52.12)	593 (53.57)
Female	699 (48.65)	626 (47.64)	573 (47.28)	563 (46.99)	577 (47.88)	514 (46.43)
Age group (years), n (%)						
18 - 30	119 (8.28)	101 (7.67)	91 (7.51)	84 (7.01)	92 (7.63)	76 (6.67)
31 - 45	462 (32.15)	414 (31.51)	385 (31.77)	382 (31.89)	379 (31.45)	330 (29.81)
46 - 60	569 (39.60)	528 (40.18)	483 (39.85)	483 (40.32)	483 (40.08)	460 (41.55)
>60	287 (19.97)	271 (20.62)	253 (20.87)	249 (20.78)	251 (20.83)	241 (21.77)
Income (euros/month), n (%)						
<1000 (Tier I)	126 (8.77)	119 (9.06)	116 (9.57)	110 (9.18)	103 (8.55)	94 (8.49)
1000 - 2000 (Tier II)	300 (20.88)	271 (20.62)	244 (20.13)	244 (20.37)	249 (20.66)	240 (21.68)
2001 - 3000 (Tier III)	364 (25.33)	339 (25.80)	319 (26.32)	310 (25.88)	309 (25.64)	281 (25.38)
3001 - 4000 (Tier IV)	294 (20.46)	271 (20.62)	243 (20.05)	245 (20.45)	251 (20.83)	229 (20.69)
>4000 (Tier V)	353 (24.57)	314 (23.90)	290 (23.93)	289 (24.12)	293 (24.32)	263 (23.76)
Perceived Stress Score, mean (SD)	16.19 (7.19)	15.83 (7.43)	15.76 (7.56)	15.65 (7.41)	15.54 (7.50)	14.89 (7.58)

Table . Distribution of the adult population in Germany (2023) by gender, age group, and monthly income level.

Category	Adult population, n (%)
Sex	
Male	41.2 (48.8)
Female	42.3 (51.2)
Age group (years)	
18 - 30	14.2 (17)
31 - 45	19.2 (23)
46 - 60	21.7 (26)
>60	28.4 (34)
Monthly income level (euros)	
<€1250 (<US $1375)	21.1 (25.3)
€1250-€2080 (US $1375-US $2288)	13 .7 (16.4)
€2080-€2920 (US $2288-US $3212)	12.4 (14.8)
€2920-€4170 (US $3212-US $4587)	13.6 (16.3)
•	
>€4170 (>US $4587)	22.6 (27.1)

Data Collection

The panelists of the panel company had already consented to install tracking software on their desktops or mobile devices. Some participants consented to install it on both devices. Through this tracking software, the company provided fine-grained traces of visited URLs and mobile apps, including the time of visit and duration of each visit. During the 7 months, we recorded 47,100,701 URL visits from both desktop and mobile users, covering 236,955 unique web domains. For mobile apps, we captured 13,553,645 app visits across 13,476 unique apps.

XSL•FO
RenderX

Data Cleaning and Preprocessing

First, we removed the bottom 20% of panelists in each wave, ranked by total time spent browsing, since they did not have sufficient data to extract meaningful internet use patterns. Second, we identified a group of panelists who appeared to be "professional survey takers," spending more than 25% of their online time on survey domains. To focus on users with more typical internet use, we excluded these individuals from the sample. Notably, applying the above time threshold to their nonsurvey activities would have led to the removal of more than 29% of these panelists. Finally, to ensure that our internet use measures accurately capture user behavior, we only included panelists for whom we could categorize at least 80% of their web visits (refer to the "Data Enrichment" section for details). Table 3 summarizes the number of participants excluded at each step, resulting in a set of distinct panelists across waves comprising 656 mobile users and 526 desktop users. Table 4 presents the sociodemographic characteristics of the remaining users included in the analysis. In the mobile data, the proportion of users aged 31 - 45 years increased compared to the baseline questionnaire, as provided in Table 3. In the desktop data, the proportion of males and users aged 46 - 60 years increased, while the proportion of users aged 31 - 45 years decreased relative to the baseline questionnaire.

Table . Overview of panelists with matched passive web data from desktop and mobile devices across 6 survey waves. The table shows the number of users before and after data cleaning for both device types. The final row indicates the total number of distinct users retained in the cleaned dataset.

Survey wave	Number of panelists	Desktop users	Desktop users (cleaned)	Mobile users	Mobile users (cleaned)
1	1437	981	359	907	519
2	1314	848	321	806	470
3	1212	762	284	728	426
4	1198	717	257	717	418
5	1205	714	265	697	399
6	1107	656	227	649	368
Total distinct users	__[a]	—	526	—	656

[a]Not applicable.

Table . Sociodemographic characteristics of users included in the analysis after data cleaning for both mobile and desktop datasets.

Characteristic	Mobile (n=656)	Desktop (n=526)
Gender, n (%)		
Male	334 (50.91)	289 (54.94)
Female	322 (49.09)	237 (45.06)
Age group (years), n (%)		
18 - 30	53 (8.08)	38 (7.22)
31 - 45	246 (37.5)	137 (26.05)
46 - 60	247 (37.65)	232 (44.11)
>60	110 (16.77)	119 (22.62)
Income (euros/month), n (%)		
<1000 (Tier I)	60 (9.15)	59 (11.22)
1000 - 2000 (Tier II)	129 (19.66)	121 (23)
2001 - 3000 (Tier III)	166 (25.30)	128 (24.33)
3001 - 4000 (Tier IV)	137 (20.88)	87 (16.54)
>4000 (Tier V)	164 (25)	131 (24.9)

Data Enrichment

To understand "how" the panelists are using the internet, we categorized their online visits into semantic categories. The goal was to group domains, subdomains, and apps based on their primary function into categories such as "social media" (eg, facebook.com [Meta Platforms, Inc] and TikTok app [ByteDance]) and "productivity" (eg, Gmail [Google LLC] and calendar.google.com [Google LLC]). We derived the set of categories (provided in Table 5) by combining categories used by app stores and web domain classification services such as Webshrinker.com. For platform domains such as google.com, we also categorized their subdomains. For instance, google.com was categorized as "search," while mail.google.com was

XSL•FO
RenderX

classified as "productivity." Table 5 provides the complete list of semantic categories we considered, along with some examples. Two researchers from our team first independently annotated the categories for all domains and apps that constituted around 85% of web visits made by our panelists. Later, disagreements were resolved collaboratively. We observed a substantial interannotator agreement with a Cohen κ agreement [57] score of 0.7, based on annotations for a random subset of 200 domains. Following this process, we classified 3777 domains and 989 apps into semantic categories, capturing 85% of visits from mobile devices and 84% from desktops.

Table . Categorization of web domains and mobile apps based on semantic usage type. The table lists representative examples of domains and apps across various categories, grouped separately for desktop web domains and mobile apps.

Category	Example of domains, subdomains, and apps in the category
Domains	
Entertainment	youtube.com, twitch.tv, disneyplus.com, and netflix.com
Shopping	amazon.de, ebay.de, kleinanzeigen.de, and temu.com
Social media	facebook.com, twitter.com, and instagram.com
Messaging	whatsapp.com, knuddels.de, and fdating.com
Productivity	mail.google.com, outlook.live.com, navigator.web.de, and docs.google.com
Games	gameduell.de, anocris.com, forgeofempires.com, and spielaffe.de
Adult	pornhub.com, xvideos.com, xnxx.com, and romeo.com
News	bild.de, focus.de, welt.de, and wunderweib.de
Apps	
Entertainment	YouTube (Google LLC), Netflix (Netflix, Inc), and Spotify Music (Spotify Technology)
Shopping	Amazon Shopping (Amazon.com, Inc), eBay (eBay, Inc), Vinted.fr (Vinted Group), and Lidl Plus (Schwarz Group)
Social media	Facebook (Meta Platforms, Inc), Instagram (Meta Platforms, Inc), and Twitter (X Corp), TikTok – Make Your Day (ByteDance)
Messaging	WhatsApp (Meta Platforms, Inc), Facebook Messenger (Meta Platforms, Inc), and Telegram (Telegram FZ-LLC)
Productivity	Gmail (Google LLC), GMX Mail (Global Mail eXchange), WEB.DE Mail (United Internet Group), and Google Calendar (Google LLC)
Games	Candy Crush Saga (King), Coin Master (Moon Active), Royal Match (Dream Games), and Pokémon GO (Niantic)
News	n-tv Nachrichten (RTL Group), kicker online (Olympia -Verlag GmbH), AOL – News (AOL Media LLC), and BILD: Immer aktuell informiert (Axel Springer SE)

Measures

As described in the "Study Design" section, we combined repeated monthly stress questionnaires with web browsing data. Perceived stress was measured through PSS-10 questionnaire responses, and internet use features were derived from passively collected web traces. For each panelist in each survey wave, we calculated internet usage features based on their activity during the period "T" preceding the stress measurement (ie, the time of questionnaire response). To address the question of "when" the internet is used, we extracted features for either 30 or 2 days to examine both long-term and short-term effects. The resulting measures were then used to examine the associations between internet use and stress.

Measures From Web Traces

To measure "how" individuals use the internet, we created features that span from coarse- to fine-grained measures, as provided in Table 6. We captured overall web activity at the coarse-grained level, such as total time spent online. We also accounted for the time of the day when panelists were browsing the web by including the difference between the time spent online during daytime (6 AM-6 PM) and nighttime (6 PM-6 AM) hours. At a finer granularity, we analyzed how panelists distributed their time across online activities such as social media, entertainment, and news. For each survey wave, if a participant completed the survey on a given date (eg, July 31), we summed their time spent on each activity over the period (T=30 or 2 days; eg, July 1 - 30 or July 29 - 30) preceding the day of completing the survey. For instance, time spent on news represents the aggregated time on news domains (desktop) and news domains and apps (mobile) during that period. Each participant had up to 6 time points, 1 per wave.

XSL•FO
RenderX

Table . Features and their descriptions. Time spent online is measured in hours in the period T (30 days or 2 days) before the measurement of stress. These features are computed for online activity on each device (desktop or mobile) separately.

Features	Description
Coarse-grained	
• Total time spent online	Total time spent online in period T
• Daytime nighttime difference	Difference of time spent online during daytime (6 AM-6 PM) and nighttime hours (6 PM-6 AM)
Fine-grained	
• Time spent on entertainment • Time spent on social media • Time spent on messaging • Time spent on news • Time spent on adult content • Time spent on games • Time spent on shopping • Time spent on productivity	Time spent on different semantic classes of online activities. For instance, time spent on entertainment domains or apps such as YouTube.com or Amazon Prime is classified as entertainment use.
Control variables	
• Gender • Age • Income • Survey wave	Sociodemographic characteristics of individuals and seasonality

Measures From Questionnaires

We used the PSS-10 [58] in our monthly questionnaires to measure the stress levels of our panelists. The PSS-10 is a widely used, validated scale designed to assess how stressed individuals feel. It captures aspects such as the unpredictability of life, perceived control over situations, and general stress levels over the past month. Participants rate their responses on a scale from 0 (never) to 4 (very often), producing a total score between 0 and 40 across the 10 items. Higher scores indicate greater perceived stress, with scores typically grouped into 3 levels: 0 - 13 (low stress), 14 - 26 (moderate stress), and 27 - 40 (high stress) [59,60]. In addition, we collected each participant's self-reported sociodemographics, including age, gender, and income, in the first wave of the questionnaires.

Statistical Analysis

We used linear mixed-effects models (LMMs) [61] to examine the relationship between internet use and stress. We chose LMMs for analyzing data from our longitudinal study since they account for repeated measurements of individuals and incorporate fixed and random effects. Fixed effects included internet use features provided in Table 6. Random intercepts were added to account for individual-specific differences in baseline stress levels across participants.

We formally describe the models as follows. For an individual i at questionnaire wave j∈ {1,2...6}, we denote Yij as the variable of interest, xij as the covariate, and the intercept for the random effect as uj. Therefore, we consider the following LMM:

$$Yij=\beta 0+\beta 1xij1+\beta 2xij2+\ldots+\beta nxijn+uj+ ij$$

where:

- Yij is the perceived stress level of the individual i measured at the questionnaire wave j.
- β0 is the fixed intercept.
- β1,…,βn are the fixed effect coefficients for each covariate xijn.
- xij1=(total time spent online, xij2=daytime-nighttime difference, xij3=time spent on entertainment ... xijn=survey wave), where xijn corresponds to each feature provided in Table 6.
- uj is the random effect for individual i, capturing individual-level variability.
- ij is the residual error term for the individual i at wave j.

We conducted model diagnostics to validate the assumptions of LMMs, including checks for multicollinearity (Variance Inflation Factor (VIF) <2.0). All statistical analyses were implemented using Python's statsmodels package (version 0.14.1; Python Software Foundation) [62].

To understand whether the granularity of the extracted features affects model performance and the associations identified, we developed 2 models. The first model (Model 1) focused on coarse-grained measures of internet use such as total time spent online and daytime-nighttime difference. The second model (Model 2) extended the first model and also incorporated finer-grained measures of internet use across semantic classes. For Model 2, we dropped the total time spent online feature to avoid multicollinearity.

Previous work has shown that both sociodemographics [40-44] and seasonal variations [63,64] can significantly influence individuals' stress levels. Accordingly, we included the sociodemographics and seasonality measures as control variables for both models, as provided in Table 6.

XSL•FO
RenderX

Ethical Considerations

Our study was approved by Aalto University's Research Ethics Committee (approval ID D/894/03.04/2023). Data collection was conducted via a General Data Protection Regulation–compliant European company, and informed consent was obtained from participants for both the surveys and web-trace datasets, with the option to withdraw consent at any time during or after the study. To protect participants' privacy, we implemented strict data privacy measures. The web dataset was anonymized by the panel company by removing personal information such as email addresses and usernames to prevent participant identification. In addition, the dataset was stored and analyzed solely on the university's secure server, with access restricted to the research team. We will make the anonymized data and code available to support the open-source community and to spur further research at the intersection of internet use and well-being.

Results

Overview

Our study examined various internet use behaviors associated with stress, and in this section, we present our results across 4 key contextual dimensions (as outlined in the "Contributions and Impact of Our Study" section). We first analyzed "how" internet-based features relate to stress. We then explored the remaining dimensions: device-based differences (where) by comparing desktop and mobile usage, temporal patterns (when) using internet activity from the 2 days before the survey, and individual differences (by whom) through subgroup analyses based on age, gender, income, and baseline stress levels.

Behavioral Patterns (How)

To understand how internet usage is associated with stress, we ran LMMs on 2 sets of features, progressing from coarse-grained (amount and timing of usage) to fine-grained (also including semantic category usage) measures. An ANOVA test was conducted to determine whether the more complex model explained significantly more variance than the simpler model. The results showed no statistically significant improvement when using the more complex model, although the more complex model provided important information on the nuanced relationship between internet use and stress.

Analysis of 30-day mobile data (number of panelists, N=656) and observations, n=2600), as provided in Table 7, revealed that Model 2—which includes both timing and semantic category usage—identified significant associations with stress. Specifically, shopping-related usage was positively associated with stress (β=.04,, 95% CI 0.00 - 0.08; *P*=.04), while productivity usage showed a negative association (β=–.03, 95% CI −0.06 to −0.00; *P*=.04). In contrast, Model 1, which included only total usage and timing, did not show any significant associations.

XSL•FO
RenderX

Table . Results from linear mixed-effects models for all participants, based on 30-day mobile data. Model 1 includes coarse-grained features, while Model 2 incorporates fine-grained usage categories (described in the "Statistical Analysis" section). Estimates, CIs, and *P* values are reported for each predictor. Statistically significant *P* values are in bold.

Predictors	Model 1[a], estimate (95% CI)	*P* value	Model 2[b], estimate (95% CI)	*P* value
Intercept	20.77[c] (19.09 - 22.44)	<.001	20.69[c] (19.01 - 22.37)	<.001
Survey wave	−0.10[d] (−0.19 to −0.02)	.01	−0.11[d] (−0.19 to −0.02)	.01
Gender	1.79[c] (0.80 - 2.77)	<.001	1.68[c] (0.68 - 2.68)	<.001
Age	−1.62[c] (−2.20 to −1.04)	<.001	−1.57[c] (−2.16 to −0.99)	<.001
Income	−1.11[c] (−1.50 to −0.73)	<.001	−1.08[c] (−1.47 to −0.69)	<.001
Total time spent online	0.00 (−0.00 to 0.01)	.39	__[e]	—
Daytime nighttime difference	−0.01 (−0.02 to −0.00)	.12	−0.01 (−0.02 to 0.00)	.18
Time spent on entertainment	—	—	0.01 (−0.01 to 0.02)	.37
Time spent on social media	—	—	0.00 (−0.01 to 0.02)	.71
Time spent on messaging	—	—	0.00 (−0.02 to 0.02)	.70
Time spent on games	—	—	0.00 (−0.00 to 0.01)	.30
Time spent on shopping	—	—	0.04[d] (0.00 - 0.08)	.04
Time spent on productivity	—	—	−0.03[d] (−0.06 to −0.00)	.04
Time spent on news	—	—	−0.03 (−0.09 to 0.03)	.33

[a]σ^2=11.67; τ =36.54$_{pid}$; ICC=0.76; N=656$_{pid}$; Observations=2600.

[b]σ^2=11.65; τ =36.48$_{pid}$; ICC=0.76; N=656$_{pid}$; Observations=2600.

[c]*P*<.001.

[d]*P*<.05.

[e]Not applicable.

In addition, sociodemographic factors such as age, gender, and income consistently predicted stress across both models. Age (β=–1.57, 95% CI −2.16 to −0.99; *P*<.001) and income (β=–1.08, 95% CI −1.47 to −0.69; *P*<.001) were negatively associated with stress, while women reported higher stress levels (β=1.68, 95% CI 0.68 - 2.68; *P*=.001).

Device Matters (Where)

To observe device differences, we analyzed 30-day desktop data. For desktop data (N=526 and n=1713), the results are shown in Table 8. Model 2, which incorporates semantic and temporal features, showed a weaker positive association between shopping usage and stress (β=.03, 95% CI −0.0 to 0.06; *P*=.09). As observed with mobile data, the simpler Model 1 did not reveal any significant associations with internet usage features. Similarly, the sociodemographic results were consistent with those observed in the mobile data.

XSL•FO
RenderX

Table . Results from linear mixed-effects models for all participants, based on 30-day desktop data. Model 1 includes coarse-grained features, while Model 2 incorporates fine-grained usage categories (described in the "Statistical Analysis" section). Estimates, CIs, and *P* values are reported for each predictor. Statistically significant *P* values are in bold. Random effects, intraclass correlation coefficient (ICC), number of participants (N), and total observations are also provided.

Predictors	Model 1[a], estimate (95% CI)	*P* value	Model 2[b], estimate (95% CI)	*P* value
Intercept	20.62[c] (18.70 - 22.55)	<.001	20.49[c] (18.58 - 22.41)	<.001
Survey wave	−0.14[d] (−0.24 to −0.03)	.009	−0.15[d] (−0.26 to −0.04)	.006
Gender	1.83[d] (0.66 - 3.00)	.002	1.73[d] (0.55 - 2.90)	.004
Age	−1.72[c] (−2.41 to −1.04)	<.001	−1.74[c] (−2.43 to −1.05)	<.001
Income	−0.94[c] (−1.38 to −0.51)	<.001	−0.92[c] (−1.36 to −0.48)	<.001
Total time spent online	−0.00 (−0.01 to 0.00)	.40	__[e]	—
Daytime nighttime difference	0.00 (−0.01 to 0.01)	.46	0.00 (−0.01 to 0.01)	.46
Time spent on entertainment	—	—	−0.00 (−0.01 to 0.01)	.64
Time spent on adult content	—	—	−0.01 (−0.03 to 0.00)	.11
Time spent on social media	—	—	0.00 (−0.02 to 0.02)	.94
Time spent on messaging	—	—	0.01 (−0.04 to 0.06)	.61
Time spent on games	—	—	0.00 (−0.03 to 0.03)	.84
Time spent on shopping	—	—	0.03 (−0.00 to 0.06)	.09
Time spent on productivity	—	—	−0.00 (−0.03 to 0.02)	.86
Time spent on news	—	—	−0.02 (−0.06 to 0.02)	.28

[a]σ^2=11.10; τ =40.62$_{pid}$; ICC=0.79; N=526$_{pid}$; Observations=1713.

[b]σ^2=11.09; τ =40.77$_{pid}$; ICC=0.79; N=526$_{pid}$; Observations=1713.

[c]P<.001

[d]*P*<.01.

[e]Not available.

Time Period of Data (When)

To investigate the relationship between short-term versus long-term internet usage patterns and stress, we analyzed associations between various online activities performed on mobile and desktop devices in 2 time periods—30 days and 2 days—and individual stress levels. We used the same features and models as in our previous 30-day data analyses. Here, we specifically focused on web activity recorded during the 2 days immediately preceding the PSS-10 survey.

For both mobile and desktop data (as provided in Tables S1 and S2 in Multimedia Appendix 1), news usage showed a negative association with stress (β=–.54, 95% CI −1.08 to 0.00; *P*=.048) and (β=–.50, 95% CI −0.90 to −0.11; *P*=.01), respectively, in Model 2. In addition, in desktop data, messaging usage demonstrated a weak negative association with stress (β=–.59, 95% CI −1.24 to 0.06; *P*=.07) in model 2.

Individual Differences (By Whom)

To explore how internet usage varies by individual characteristics, we conducted subgroup analyses, running models separately for categories such as gender (male and female) to understand how associations differ based on these characteristics. In the following subsections, we first examined the relationship between internet use and stress based on baseline stress levels by analyzing high-stress and low-stress groups. We then explored differences by gender, age, and income categories.

Stress Levels

We identified 2 groups of panelists from our data—high-stress and low-stress—based on their reported PSS-10 scores in the online questionnaires. Panelists who scored more than 26 in any wave they participated in were included in the high-stress group, and panelists who scored below 14 in any wave were included in the low-stress group.

For the high-stress population (PSS-10 score > 26), several notable results were observed (as provided in Tables S3-S6 in Multimedia Appendix 1). In the 30-day mobile data, time spent (β=.01, 95% CI 0.0 - 0.02; *P*<.001) in Model 1, and social media usage (β=.02, 95% CI 0.0 - 0.03; *P*=.02), and gaming usage (β=.01, 95% CI 0.0 - 0.02; *P*=.02) in Model 2 were positively associated with stress. In the 2 days of data, the daytime-nighttime difference showed a weak positive association (β=.11, 95% CI −0.01 to 0.23, *P*=.08) in Model 1. For desktop data, no significant variables, including sociodemographics, were found to be associated with stress in the high-stress subgroup.

XSL•FO
RenderX

In the low-stress population (PSS-10 score <14 in the baseline survey), as provided in Tables S7-S10 in Multimedia Appendix 1, adult-content usage was negatively associated with stress in 30-day desktop data (β=–.02, 95% CI −0.04 to 0.0; *P*=.07). Similarly, for the low-stress group, in the 2 days data, time spent (β=–.07, 95% CI −0.14 to 0.0; *P*=.04) was significant for desktop, while gaming usage was weakly significant for mobile data (β=.08, 95% CI −0.01 to 0.17; *P*=.1).

When analyzing the 30-day data for all participants, sociodemographic factors were strongly associated with stress in both mobile and desktop settings. However, within the high-stress group, income was the only sociodemographic variable that remained significant in mobile data, showing a negative association with stress in both the general population (β=–1.08, 95% CI −1.47 to −0.69; *P*<.001) and the high-stress subgroup (β=–.52, 95% CI −0.91 to −0.12; *P*=.01). In contrast, gender and age—which were significant predictors in the overall population—did not show statistical significance in the highly stressed subgroup for either mobile or desktop data.

Gender Differences

Subgroup analysis by gender revealed distinct patterns in feature significance for both desktop and mobile data. In the 30-day mobile data (as provided in Table S11 in Multimedia Appendix 1), shopping usage (β=.07, 95% CI 0.01 - 0.14; *P*=.02) and productivity features (β=–.05, 95% CI −0.09 to −0.01; *P*=.02) were significant only for male users (N=334 panelists and n=1334 observations) in Model 2. In contrast, these features were not significant for female users (N=322 panelists and n=1266 observations) as provided in Table S12 in Multimedia Appendix 1. No features were significant for males and females in the 30-day desktop data.

In the 2-day data (as provided in Tables S15-S16 in Multimedia Appendix 1), news consumption was negatively associated with stress for males in both desktop (β=–.52, 95% CI −1.02 to −0.01; *P*=.04) and mobile (β=–.58, 95% CI −1.23 to 0.07; *P*=.08) data. For females, in mobile data, daytime-nighttime difference (β=.10, 95% CI −0.0 to 0.21; *P*=.06) and messaging (β=.23, CI −0.02 to 0.48; *P*=.08) were weakly positively associated.

Age Differences

Subgroup analysis by age revealed distinct patterns in web-based associations. For the 30-day mobile data (as provided in Tables S19-S22 in Multimedia Appendix 1), shopping was positively associated with stress in age groups of 18 - 30 years (β=.12, 95% CI −0.02 to 0.25; *P*=.09) and older than 60 years (β=.01, 95% CI 0.02 - 0.19; *P*=.02). In the age group of 30 - 45 years, weak positive associations were found for entertainment (β=.02, 95% CI −0.00 to 0.04; *P*=.08) and messaging (β=.03, 95% CI −0.00 to 0.06; *P*=.07), whereas productivity was negatively associated (β=–.08, 95% CI −0.14 to −0.02; *P*=.007). In the older than 60 years group, time spent (β=.01, 95% CI 0.00 - 0.03; *P*=.03) and messaging (β=.05, CI −0.00 to 0.09; *P*=.06) were positively associated, while the daytime-nighttime difference (β=–.03, 95% CI −0.05 to −0.01; *P*=.01) and shopping (β=.10, 95% CI 0.02 - 0.19; *P*=.02) were negatively associated.

In the 30-day desktop data (as provided in Tables S23-S26 in Multimedia Appendix 1), adult content usage was negatively associated in 18 - 30 years (β=–.87, 95% CI −1.77 to 0.03; *P*=.06) and older than 60 years (β=–.19, 95% CI −0.38 to 0.00; *P*=.06) age groups. In addition, in the 18 - 30 age group, the daytime-nighttime difference (β=.07, CI −0.00 to 0.14; *P*=.06) was positively associated, while news usage (β=.026, CI −0.42 to −0.10; *P*=.002) was negatively associated. Shopping was positively associated in the age group of 45 - 60 years (β=.04, 95% CI −0.00 to 0.09; *P*=.06).

For the 2-day mobile data (as provided in Tables S27-S30 in Multimedia Appendix 1), news was negatively associated in both the 30 - 45 years (β=–1.01, 95% CI −2.22 to 0.19; *P*=.10) and 45 - 60 years (β=–.85, 95% CI −1.76 to 0.06; *P*=.07) age groups. In addition, in the age group of 30 - 45 years, the daytime-nighttime difference (β=.14, 95% CI 0.02 - 0.25; *P*=.02) and entertainment usage (β=.22, 95% CI 0.01 - 0.43; *P*=.04) were positively associated, whereas shopping was negatively associated (β=–.54, 95% CI −1.08 to −0.0; *P*=.05). In the older than 60 years of age group, time spent on gaming (β=.26, 95% CI 0.01 - 0.50; *P*=.04) was positively associated.

Similarly, for the 2-day desktop data (as provided in Tables S31-S34 in Multimedia Appendix 1), messaging was negatively associated in the 45 - 60 years age group (β=–.78, 95% CI −1.48 to −0.07; *P*=.03), but positively associated for the older than 60 years age group (β=7.63, 95% CI 0.39 - 14.87; *P*=.04). In the age group of 18 - 30 years, entertainment (β=1.14, 95% CI −0.18 to 2.46; *P*=.09) was positively associated, and social media (β=1.22, 95% CI 0.15 - 2.29; *P*=.03) was positive for the 30 - 45 age group. In addition, time spent (β=–.14, 95% CI 0.28 - 0.0; *P*=.06) and news usage (β=.46, 95% CI −0.96 to 0.04; *P*=.07) were negatively associated in the older than 60 years age group.

Income Differences

For the 30-day mobile data (as provided in Tables S35-S39 in Multimedia Appendix 1), messaging (β=–.06, 95% CI −0.12 to −0.01; *P*=.03) was negatively associated with stress in participants earning less than €1000 (US $ 1100) per month (Tier I). In the €2001-€3000 (US $2201.10-US $3300) income group (Tier II), productivity (β=–.05, 95% CI −0.10 to 0.0; *P*=.06) and news usage (β=–.10, 95% CI −0.20 to 0.01; *P*=.08) were both negatively associated. Time spent (β=.01, CI 0.0 - 0.02; *P*=.02) and shopping (β=.08, 95% CI 0.01 - 0.16; *P*=.02) were positively associated with stress for participants earning €3001-€4000 (US $3301.10-US $4400; Tier IV). No other significant internet-based features were observed for other income categories.

For the 30-day desktop data (as provided in Tables S40-S44 in Multimedia Appendix 1), news usage was positively associated with stress in Tier I income group (β=.28, 95 CI 0.05 - 0.50; *P*=.02), while productivity was negatively associated (β=–.07, 95% CI −0.13 to −0.01; *P*=.03). For participants in Tier III income, news usage (β=–.13, 95% CI −0.25 to 0.0; *P*=.06) was negatively associated. For participants in Tier IV income, shopping was positively associated with stress (β=.08, 95% CI 0.02 - 0.15; *P*=.02). For Tier V participants, social media use was positively associated with stress (β=.06, 95% CI 0.01 - 0.12; *P*=.03), while news use (β=–.16, 95% CI −0.25 to −0.07; *P*<.001) and time spent were negatively associated

XSL•FO
RenderX

(β=–.01, 95% CI −0.03 to 0.0; *P*=.03). No significant associations were identified for other income categories.

For the 2-day mobile data (as provided in Tables S45-S49 in Multimedia Appendix 1), gaming (β=.21, 95% CI −0.02 to 0.43; *P*=.07) was positively associated in the Tier II income group and negatively associated (β=–.19, 95% CI −0.37 to 0.0; *P*=.047) in the Tier V income group. News use was negatively associated in the Tier III (β=–.75, 95% CI −1.58 to 0.09; *P*=.08) and Tier V (β=–.74, 95% CI −1.60 to 0.12; *P*=.09) income groups. In addition, messaging (β=, 95% CI 0.14 - 0.82; *P*=.006) was positively associated in the Tier III income group, and daytime-nighttime difference (β=.18, 95% CI 0.0 - 0.36; *P*=.04) was positively associated in the Tier IV income group.

For the 2-day desktop data (as provided in Tables S50-S54 in Multimedia Appendix 1), news usage was negatively associated in the Tier III income group (β=–1.13, 95% CI −2.33 to 0.06; *P*=.06) and the Tier V income group (β=–1.04, 95% CI −1.76 to −0.32; *P*=.005). In addition, in the Tier V income group, productivity (β=.64, 95% CI 0.07 - 1.21; P=.03) was positively associated with stress. In the Tier IV income group, time spent (β=.23, 95% CI 0.04 - 0.42; *P*=.0192) and daytime-nighttime difference (β=.23, 95% CI −0.00 to 0.46; *P*=.05) were positively associated.

Discussion

Principal Findings

Our results show that various internet usage behaviors are associated with stress, both positively and negatively. These associations differ across device type, time frame, and sociodemographic groups. Figure 2 summarizes these patterns and their variation across panelist subgroups. In the following sections, we discuss the internet use features that have shown positive (social media, entertainment, shopping, games, and time of internet use), negative (adult content, productivity, and news), and mixed (messaging and screen time) associations with stress.

Figure 2. Overview of significant internet usage behaviors across various contextual dimensions. The icons represent different internet use features. The rows correspond to a combination of the time frame (30 days or 2 days) and device type (desktop or mobile), and the columns correspond to factors pertaining to individual differences (stress levels, gender, age, and income). Red icons denote positive associations, blue icons indicate negative associations, and empty cells show no significant internet features. The intensity of the color reflects the strength of the significance, with lighter icons denoting weak significance (*P*=.05 to *P*<.10) and darker icons representing high significance (*P*<.05).

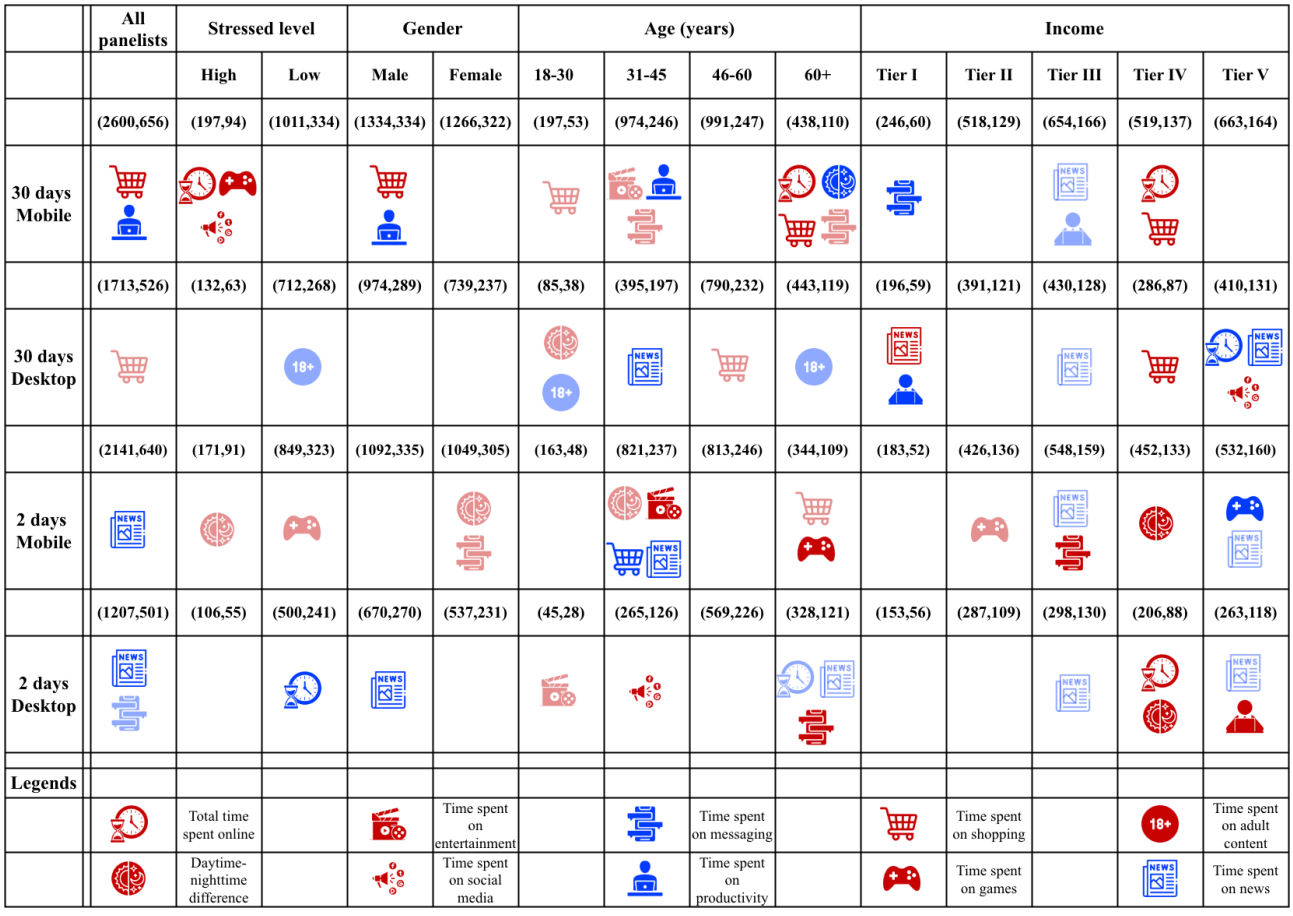

Positive Associations

Social Media

People spend a substantial portion of their time online on social media platforms [65]. In our dataset, social media accounted for approximately 23% of total usage in mobile data and 15% in desktop data. Therefore, understanding social media's relationship with stress is increasingly important as it continues to occupy a large share of individuals' internet activity.

XSL•FO
RenderX

Previous research shows that social media can contribute to stress, act as a resource, or function as a coping tool [50]. Factors such as fear of missing out [66], appearance-related pressure [67], and communication overload [68] have been linked to increased distress. At the same time, other studies have highlighted its potential to buffer stress and offer social support in specific contexts [22,35,69].

In our results, when significant, social media was consistently positively associated with stress across subgroups and device types. This association was significant for the high-stress subgroup in the 30-day mobile data, the Tier I income group in the 30-day desktop data, and the 30 - 45 age group in the 2-day desktop data (Figure 2). Although we cannot definitively determine whether social media contributes to or alleviates stress, these patterns suggest that it may be used as a coping strategy for these groups.

Previous work has identified social media as a space for various coping mechanisms [35]. Our findings in the middle-aged group potentially reflect this pattern, aligning with earlier research [70], possibly due to the support accessed through these platforms. This is consistent with studies linking social media use, particularly Facebook, to support-seeking behavior [71].

Overall, our results highlight the role of social media in shaping stress experiences. While our study, combining fine-grained web data with a longitudinal design and contextual framework, strengthens this interpretation, further research is needed to disentangle the causal directions of the observed associations and social media's role—whether as stressor, resource, or coping tool.

Entertainment

According to a recent survey in Germany, people spend an average of 203 minutes per day watching content online [65]. In our data, entertainment usage accounted for 9% of total usage on mobile devices and 14% on desktop devices.

Previous studies have reported a positive correlation between entertainment usage and stress [21,72-74]. Further, activities such as watching content or listening to music have also been identified as common coping strategies for managing stress [36,37,75]. In our findings, entertainment usage was positively associated with stress in both desktop and mobile data. For mobile users, this association was consistent across both the 30-day and 2-day periods for the 30 - 45 age group. In the 2-day desktop data, a weak positive association was observed for the age group of 18 - 30 years.

These results suggest that entertainment usage (similar to social media) may serve as a coping mechanism, especially among younger users in our data, compared with older individuals. Previous research indicates that younger individuals are more likely to engage in binge-watching as a way to regulate emotions [36]. However, there is a lack of detailed research on this association [74], and future studies should further explore the types of content consumed and their relationship to mental health.

Online Shopping

Online shopping has grown substantially in recent years, particularly with the rise of smartphones, which now account for 80% of all retail visits [76]. Previous studies have linked compulsive buying-shopping disorder to higher levels of stress [29,38,52,77-79]. At the same time, shopping has also been identified as a way to relieve stress [80,81].

In our results, shopping was predominantly positively associated with stress across mobile and desktop data, in various time frames and subgroups (refer to Figure 2), aligning with previous findings. A negative association was observed only for the age group of 30 - 45 years in the 2-day mobile data, suggesting that, for this group, shopping may serve as a short-term stress reliever, or that individuals under stress may avoid shopping. The latter behavior aligns with previous research showing that, in middle-aged adults, stress can lead to increased saving behavior [82].

This association was more pronounced in mobile data, suggesting that people may use mobile phones to cope with stress due to their easier accessibility. It was also consistent across both mobile and desktop data for the Tier IV income group, supporting earlier findings that higher-income individuals may use shopping as a way to cope with stress [38]. In the 30-day mobile data, this positive association was observed among male participants but not among females. Previous studies show that males are more likely to experience negative emotions related to shopping [83], while some studies report no gender differences in online shopping addiction tendencies [29].

These findings highlight how shopping has become embedded in daily life and its potential influence on stress levels. Future research should explore different types of shopping (hedonic vs utilitarian) and their impact on mental health. Another direction could be to examine the time spent on shopping compared with actual purchases completed after the payment process, and how these different shopping behaviors are associated with stress levels.

Gaming

Gaming has become a widespread daily habit, with the global market projected to reach US $522.46 billion by 2025 [84]. Previous research presents mixed findings: while gaming has been positively linked to stress and lower psychosocial well-being, with stress being a known precursor to pathological gaming [32,33,85]. It has also been identified as a stress reliever and coping mechanism [39,86-88].

In our results, gaming usage was positively associated with stress in the 30-day mobile data for the high-stress subgroup. In the 2-day mobile data, positive associations were observed for the low-stress group, users aged older than 60 years, and those in the Tier II income range. A negative association was found for the Tier V income group. These findings align with previous research suggesting that gaming may intensify stress in already stressed individuals [89], whereas older adults often use gaming as a way to relieve stress [90]. The contrasting patterns across income groups may reflect differences in usage

XSL•FO
RenderX

intensity, as lower-income individuals tend to engage more frequently in gaming than those with higher incomes [91].

We did not find any significant associations in the desktop data, which may be due to lower gaming activity on desktop browsers compared with mobile apps—gaming accounted for only 4.7% of desktop usage versus 16.8% in mobile data. Future research should examine the type of gaming, such as games that require active cognitive engagement (eg, first-person shooter games) versus low cognitive requirement arcade games, and how these different types of games relate to stress.

Timing of Online Activity

Individuals tend to have preferences for the timing of both online and offline activities throughout the day [92,93]. Previous research has shown that increased nighttime smartphone use is linked to higher perceived stress [49]. Nighttime use has also been associated with reduced sleep duration [48,94] and poorer sleep quality [95,96], both of which are shown to impact mental health negatively [97-99].

In our results, the daytime–nighttime difference feature was mainly positively associated with perceived stress across various groups. A negative association was observed only in the age group of older than 60 years in the 30-day mobile data. Positive associations were more common in the shorter time frame, particularly in the 2-day mobile data, for groups including high-stress individuals, females, those aged 30 - 45 years, and the Tier IV income group (refer to Figure 2).

Our results indicate that daytime internet use, compared with nighttime use, is positively associated with stress, contrasting with earlier studies that reported a stronger link between nighttime use and stress in young adults [49]. Longitudinal research has shown that it is sleep loss due to time spent online, rather than internet use itself, that is associated with poorer mental health outcomes [6]. Internet use patterns can also be influenced by individuals' chronotypes, which in turn will affect preferred active hours [100]. Moreover, future studies could investigate panelists' bedtime and examine postbedtime usage, as previous studies have shown that postbedtime use—rather than nighttime use in general—has the most harmful effects on sleep quality [95]. The negative association observed in the age group of older than 60 years may suggest that older individuals are more sensitive to late-night use and its link to stress. Further research is needed to better understand how daily temporal patterns of internet use relate to stress, particularly by considering users' chronotype and bedtime.

Negative Associations

Adult Content

A recent study estimates that around 90 million people may be affected by problematic adult content usage [101]. Research has shown that problematic adult content consumption can impact mental health, including associations with higher stress [102,103].

In our results, however, adult content consumption was negatively associated with stress levels. This negative relationship was observed in the low-stress group (PSS score < 14), and in the age groups of 18 - 30 years and older than 60 years within the 30-day desktop data. Some studies have reported no significant link between adult content use and psychological health [104], while others suggest it is often consumed as a form of leisure [105]. Our results may suggest that adult content serves as a stress buffer for our participants, as previous research has indicated stress and stress relief as common motivations for its use [106,107]. Another possible explanation is that users with lower stress levels might engage in this activity as a form of leisure or a means to alleviate boredom [105]. Previous research has identified boredom proneness and its positive association with the frequency of pornography use [108]. However, the positive associations observed in past studies were tied to problematic adult content consumption or addiction. This implies that while limited consumption may offer stress relief, excessive use could have detrimental effects. Future research should explore both the positive and negative impacts of adult content consumption on mental health, particularly in relation to the amount of consumption.

Productivity

Information and communication technology has become a central part of daily work and study routines, particularly in the 21st century. Most of the work we do on information and communication technology devices is connected to the internet, and their use for work and productivity has been linked to improved workplace efficiency [109,110]. However, the effects of internet use on psychological health and stress are dual in nature. While it has been associated with increased productivity, higher internet use for work-related tasks has also been linked to higher stress levels [51,111-114]. Some interventional studies report no significant effects [113], while others suggest that using the internet for work and study can have a positive impact on mental health [6].

In our results, increased use of productivity-related apps and domains was generally associated with lower perceived stress. In the 30-day mobile data, negative associations were observed among all participants, as well as within the male subgroup, the age group of 31 - 45 years, and the Tier III income group. In the 30-day desktop data, a similar negative association was found in the Tier I income group. A positive association appeared only once—in the 2-day desktop data for the Tier V income group.

Our findings suggest that stressed individuals may avoid productivity-related tasks by shifting their focus to other activities. This is supported by previous research showing a positive relationship between perceived stress and avoidant coping styles [115]. Previous research has reported a negative relationship between work-related online tasks and stress among middle-aged adults [21]. At the same time, communication overload from emails and messages has been linked to higher perceived stress in adults aged 50 - 85 years [15]. In our results, we found a similar negative association between productivity-related use and stress in the age group of 31 - 45 years, suggesting that increased productivity use may be linked to lower stress in this group. However, no significant association was observed in the older age group. For higher-income groups, the observed positive association may reflect greater work

XSL•FO
RenderX

responsibilities that are more difficult to avoid, contributing to increased stress.

Overall, our findings highlight the need for further research into how productivity-related internet use influences stress. As previous studies have shown, avoidance behaviors can increase the risk of prolonged stress and other mental health challenges [116]. Future studies should explore how online interventions can effectively help users mitigate stress.

News

Informed citizens are a cornerstone of a well-functioning democracy and good governance. However, recent studies suggest that news, especially news with highly negative content, can adversely affect mental health and increase stress levels [24,117-119].

Our findings reveal a counterintuitive relationship between news engagement and stress: participants who spent more time-consuming news tended to report lower stress levels. This association was more pronounced in the 2-day data for both mobile and desktop users, and overall, it was stronger in desktop data across both time periods. One possible explanation is that individuals experiencing stress may avoid the news altogether in the short term. Previous research supports this, showing that people under stress often disengage from news consumption [120-122]. Other studies have found no significant link between news consumption and stress [123,124], while some suggest that positive and soft news content may improve mood [30] and well-being [125].

Since our news category includes a range of sources, such as entertainment, sports, and politics, it is important to consider how different types of news relate to stress in the future. Future studies could examine the effects of specific news genres, as well as the influence of low-quality news sources and misinformation on mental health. Examining how consistent exposure to different news categories, sources, and misinformation influences stress responses over time could provide insights into mental health risks and how media consumption habits shape psychological well-being.

Mixed Associations

Messaging

In 2021, an estimated 3.09 billion mobile phone users accessed the top messaging apps for communication, with this number projected to reach 3.51 billion by 2025 [126]. The younger generation, in particular, remains connected with friends and family through messaging apps such as WhatsApp (Meta Platforms, Inc) and Telegram (Telegram FZ-LLC). Research has shown that messaging apps can be both positively associated with stress [7,127-130] and serve as stress reducers [131-134].

In our results, messaging usage was mostly positively associated with stress in mobile data. It was positively associated in the age groups of 30 - 45 years and older than 60 years and negatively associated in the Tier I income group in the 30-day mobile data. In the 2-day mobile data, it was positively associated with females and those in the Tier III income group. In the 2-day desktop data, it was negatively associated with all participants and the age group of 45 - 60 years, while it was positively associated with the age group of older than 60 years.

The key finding is that messaging use was predominantly positively associated with stress in mobile data and more negatively associated in desktop data. Previous research suggests that "checking behavior" (ie, brief and repeated usage sessions) is more common in mobile use compared to desktop devices [53], and perhaps this type of repeated checking on mobile devices adds to the stress. We found that in the age group of older than 60 years, messaging was positively associated with stress in both mobile and desktop data, which aligns with previous studies showing that older individuals experience higher stress from interpersonal communication [7]. In the Tier I income group for mobile data, messaging was negatively associated with stress. Studies show that people in lower-income groups tend to send more messages [135]. Previous research has also shown that messaging can be an effective tool for reducing depression, particularly among low-income individuals [136]. Finally, in the 2-day mobile data, messaging was positively associated with stress in females but not males, reflecting previous research suggesting that frequent messaging is linked to mental health symptoms (externalizing and inattention) in females, but not in males [137].

Our results highlight how messaging can have a dual impact on stress, depending on the device, time period, and individual differences. Future work could examine how messaging with friends and family, interactions with strangers in chatrooms, and repeated checking behavior on mobile devices relate to stress, loneliness, and overall well-being.

Screen Time

The internet plays a pervasive role in our daily lives, and the amount of time we spend online may have a detrimental impact on stress levels [48,138]. Research has shown that smartphone addiction is significantly linked to higher stress [139,140], and the amount of time spent online is positively associated with stress [48] and other mental health issues [141,142].

In our results, total time spent online was positively associated with stress in the 30-day mobile data, aligning with previous studies [48,139,140]. This association was significant for the high-stress group, individuals older than 60 years, and those in income Tier IV. However, the relationship was mainly negative in both the 2-day and 30-day desktop data across different groups, with the exception of a positive association for the Tier IV income group in the 2-day desktop data, which mirrored the mobile data results.

Our findings suggest that increased time spent online on mobile devices may amplify stress in different contexts. For high-stress individuals, extended mobile phone usage is positively associated with stress, likely due to the constant accessibility of smartphones, making it harder to disconnect from them. Previous studies have also found that smartphone users tend to experience higher levels of digital burnout compared to those using desktops or laptops [17]. In addition, older individuals in our study showed an increase in stress with more time spent online, consistent with previous research [141].

XSL•FO
RenderX

In contrast, desktop usage generally showed negative associations with stress, which may be due to the differences in accessibility of smartphones and desktop devices. Our data show that time spent on desktop devices has higher variability, as reflected by the higher SD (71.43 for desktop vs 66.86 for mobile) and coefficient of variation (desktop: 0.97, CI 0.93 - 1.02 vs mobile: 0.70, CI 0.67 - 0.73), compared to time spent on mobile devices.

Future research should further investigate device differences and the role mobile devices play in predicting stress levels.

User Characteristics and Stress

User characteristics such as gender, age, and income influence both stress levels and their association with internet use, as provided in Figure 3. In our results, female users reported significantly higher stress than males across both mobile and desktop data and for both the 30-day and 2-day periods (refer to Figure 3). Age and income showed consistent negative associations with stress across contexts. Similar findings have been reported in earlier research on sociodemographics and stress [40-44]. Interestingly, in our analysis, income was the only sociodemographic factor significantly associated with stress in the high-stress group for mobile data, while no other sociodemographic features were significant across platforms or time frames. Previous research also suggests that income, rather than age or gender, shows the strongest association with stress [45]. This highlights the need for further investigation into the role of income in stress, as such insights can inform policies aimed at reducing income-related health disparities. Future studies should also incorporate measures to incorporate resilience of individuals (trait stress measures) [143] and personality traits alongside sociodemographics to better understand these associations.

XSL•FO
RenderX

Figure 3. Overview of significant sociodemographic associations across time frames (30 days vs 2 days) and device types (mobile vs desktop). Rows distinguish model results between all participants and high-stress participants. Red icons indicate positive associations, blue icons indicate negative associations, and empty cells denote no significant effects.

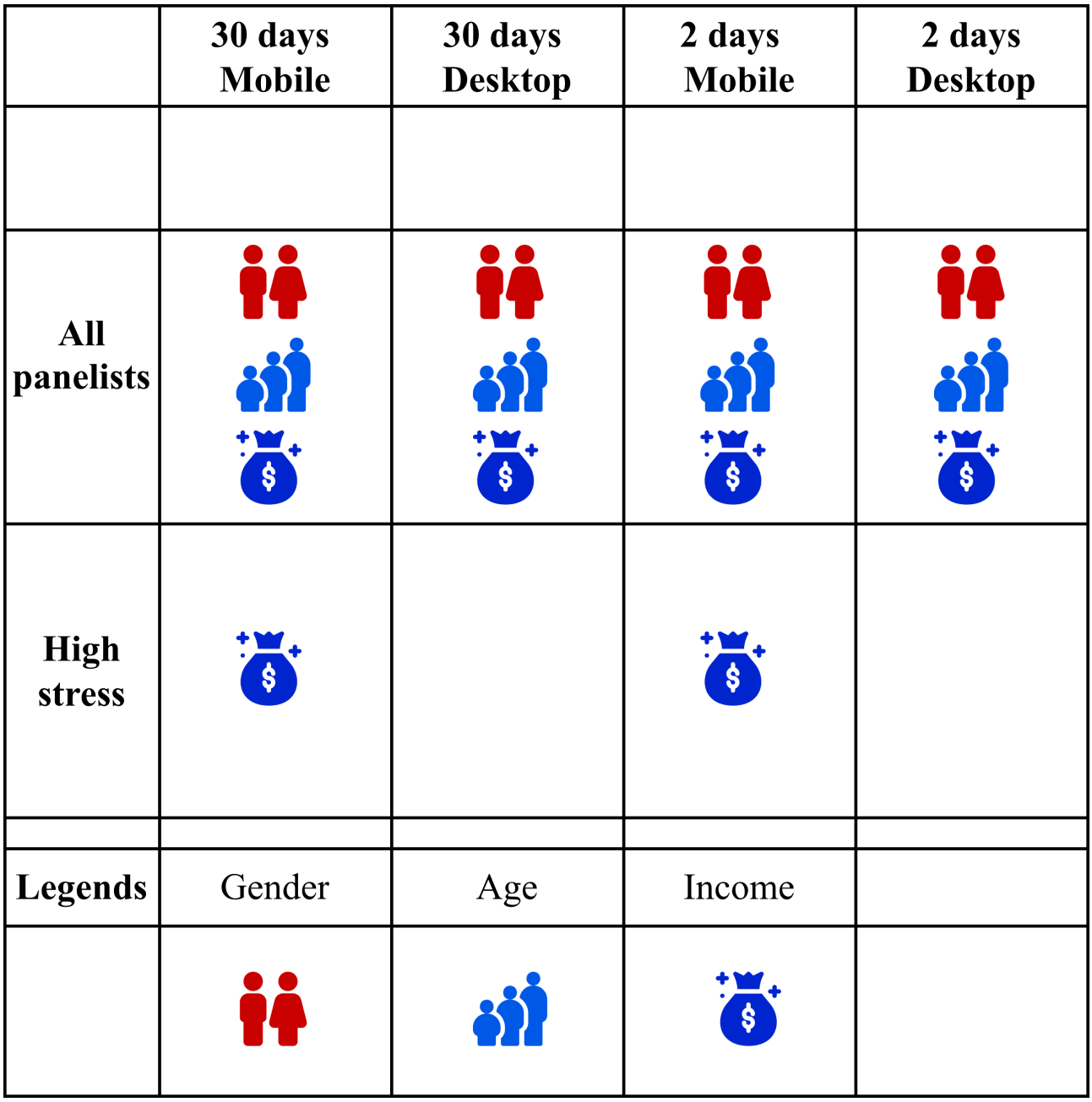

Implications

Building on these findings, it is important to consider the implications for various stakeholders, including digital platform designers, health care professionals, and end-users.

For digital platform designers, these findings suggest opportunities to promote healthier use patterns. Features such as reminders to take breaks, tools to visualize usage habits, and reduced notifications during evening/nights could help users avoid stress-inducing behaviors. In addition, adaptive designs that encourage daytime work engagement over late-night use could be particularly effective. Finally, different design considerations may be needed for mobile versus desktop devices to encourage healthier internet use.

For health care professionals, understanding how digital behaviors relate to stress could offer new ways to support patients. For example, excessive gaming on mobile devices might indicate elevated stress. Health workers could include questions about internet habits in assessments and recommend tools that encourage healthier online behaviors. Notably, the low cost and high availability of web data could provide efficient tools to complement traditional monthly surveys for monitoring stress, helping to alleviate the burden on the already strained health care sector.

For individuals, these findings emphasize the importance of timing and purpose in digital habits. Being mindful of potentially stress-inducing activities, such as excessive shopping or gaming, can foster a healthier balance. Tools that track and suggest

XSL•FO
RenderX

healthier patterns of internet use could assist users in managing their habits.

Overall, the relationship between stress and internet use is influenced by factors such as the type of activity, timing, and individual circumstances. This suggests that small, intentional changes in digital habits can help manage stress effectively.

Limitations

The panelists we collected data from are gig workers who regularly participate in surveys, which may lead to behavioral differences compared to the general population. To address this issue, we used rigorous preprocessing steps, including removing users who spent more than 25% of their time on survey websites and excluding the bottom 20% of users based on time spent. Moreover, browsing data from these panels has been shown to prominently feature the most visited domains in Germany [144]. Another limitation is that users may change their online behavior when they are aware of being tracked. However, previous work has demonstrated that the privacy attitudes of web-tracked panelists are comparable to those of nonweb-tracked panelists in Germany [145], though they may vary across countries. These findings support the reliability and suitability of the web-tracking data for capturing individuals' internet usage behavior. In addition, due to budget constraints, all panelists were from a single country, which may limit the generalizability of our findings because of cultural and behavioral differences. Previous research has demonstrated differences in cultural, social, and technological access contexts, variations in usage patterns, and mental health prevalence across countries [146,147]. Replicating this study in other countries or with a larger and more diverse sample could help further improve the generalizability and provide a broader cultural context. Further, 23% of participants dropped out between the first and final survey waves. To mitigate this attrition, we initially used a larger sample and invited all panelists from the panel company to participate in our baseline survey. Furthermore, since our tracker only tracks browsing behavior through the browser, we lack complete data on desktop usage through desktop apps, which may limit our ability to fully capture differences between mobile and desktop use. Finally, we used a validated self-report measure to assess stress, as inviting all participants to a lab setting was not feasible. Alternative approaches, such as physiological or real-time assessments, could provide more objective and detailed measures of stress.

Conclusion

Our study examines the relationship between internet use and perceived stress through a novel contextual framework that considers how, where, when, and by whom the internet is used. The findings indicate that internet use is associated with stress, and these associations differ across various usage contexts. Specifically, engagement with social media, online shopping, entertainment, and gaming is positively linked to higher stress levels. Notably, these activities have been identified in previous research as common coping mechanisms for stress, highlighting the need for future studies to examine whether such coping strategies alleviate stress or, possibly, exacerbate it over time. In contrast, productivity-related activities, news consumption, and adult content use are negatively associated with stress, suggesting they may either function as stress buffers or indicate avoidance behavior. Associations inferred from desktop data across different contextual dimensions are weaker than those inferred from mobile data, indicating that device type plays an important role. In the short term, news consumption is negatively associated with both mobile and desktop data. For individuals already experiencing high stress, increased time online on mobile phones—particularly on social media and gaming—is correlated with higher stress levels. In addition, sociodemographic factors, especially income, have significant associations with stress. These findings have important implications for the design of digital platforms, the development of mental health interventions, and the formation of healthier online habits. Future research should focus more specifically on particular web-based behaviors, such as news consumption and online shopping, and their effects on psychological well-being. It should also aim to establish causal links between internet use and stress and further investigate the mechanisms underlying sociodemographic differences in these associations.

Acknowledgments

This work was supported by the Helsinki Institute for Information Technology (HIIT). We would like to thank Yajing Wang for her help with data annotation and Emilia Marchese for her help with the data collection process. In addition, we acknowledge the computational resources provided by the Aalto Science-IT project.

Funding

This work was supported by the Helsinki Institute for Information Technology (HIIT). The funder had no involvement in the study design, data collection, analysis, interpretation, or manuscript writing.

Authors' Contributions

JK, MB, and TA led the conceptualization of the study. JK, MB, and TA were responsible for data curation and methodology development. MB and NL contributed to software implementation and validation. The investigation was conducted by JK, MB, NL, and TA, and data analysis was performed by MB and NL. JK and MB prepared the original draft, and JK, MB, NL, and TA contributed to reviewing and editing the paper. Visualizations were created by MB and NL. JK and TA provided supervision, acquired funding, and secured the necessary resources for the project.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Mixed-effects models results across contextual dimensions.
[DOCX File, 181 KB - jmir_v28i1e78775_app1.docx]

References

1. Schneiderman N, Ironson G, Siegel SD. Stress and health: psychological, behavioral, and biological determinants. Annu Rev Clin Psychol 2005;1:607-628. [doi: 10.1146/annurev.clinpsy.1.102803.144141] [Medline: 17716101]
2. Bui T, Zackula R, Dugan K, Ablah E. Workplace stress and productivity: a cross-sectional study. Kans J Med 2021;14(1):42-45. [doi: 10.17161/kjm.vol1413424] [Medline: 33654542]
3. Huang C. Internet use and psychological well-being: a meta-analysis. Cyberpsychol Behav Soc Netw 2010 Jun;13(3):241-249. [doi: 10.1089/cyber.2009.0217] [Medline: 20557242]
4. Çikrıkci Ö. The effect of internet use on well-being: meta-analysis. Comput Human Behav 2016 Dec;65:560-566. [doi: 10.1016/j.chb.2016.09.021]
5. Jenaro C, Flores N, Gómez-Vela M, González-Gil F, Caballo C. Problematic internet and cell-phone use: psychological, behavioral, and health correlates. Addict Res Theory 2007 Jan;15(3):309-320. [doi: 10.1080/16066350701350247]
6. Hökby S, Hadlaczky G, Westerlund J, et al. Are mental health effects of internet use attributable to the web-based content or perceived consequences of usage? A longitudinal study of european adolescents. JMIR Ment Health 2016 Jul 13;3(3):e31. [doi: 10.2196/mental.5925] [Medline: 27417665]
7. Nimrod G. Changes in internet use when coping with stress: older adults during the COVID-19 pandemic. Am J Geriatr Psychiatry 2020 Oct;28(10):1020-1024. [doi: 10.1016/j.jagp.2020.07.010] [Medline: 32771312]
8. Luo Y, Yip PSF, Zhang Q. Positive association between Internet use and mental health among adults aged ≥50 years in 23 countries. Nat Hum Behav 2025 Jan;9(1):90-100. [doi: 10.1038/s41562-024-02048-7] [Medline: 39558112]
9. Heo J, Chun S, Lee S, Lee KH, Kim J. Internet use and well-being in older adults. Cyberpsychol Behav Soc Netw 2015 May;18(5):268-272. [doi: 10.1089/cyber.2014.0549] [Medline: 25919967]
10. Linton MJ, Dieppe P, Medina-Lara A. Review of 99 self-report measures for assessing well-being in adults: exploring dimensions of well-being and developments over time. BMJ Open 2016 Jul 7;6(7):e010641. [doi: 10.1136/bmjopen-2015-010641] [Medline: 27388349]
11. Kraut R, Burke M. Internet use and psychological well-being: effects of activity and audience. Commun ACM 2015;58(12):94-100. [doi: 10.1145/2739043]
12. Yetton BD, Revord J, Margolis S, Lyubomirsky S, Seitz AR. Cognitive and physiological measures in well-being science: limitations and lessons. Front Psychol 2019;10:1630. [doi: 10.3389/fpsyg.2019.01630] [Medline: 31354601]
13. Patel VK. Study of internet use characteristics, perceived stress, and internet addiction among first-year medical students of Jamnagar, Gujarat, India. Ind J Priv Psychiatry 2019 Dec 1;13(2):44-47. [doi: 10.5005/jp-journals-10067-0037]
14. Nikolic A, Bukurov B, Kocic I, et al. Smartphone addiction, sleep quality, depression, anxiety, and stress among medical students. Front Public Health 2023;11:1252371. [doi: 10.3389/fpubh.2023.1252371] [Medline: 37744504]
15. Reinecke L, Aufenanger S, Beutel ME, et al. Digital stress over the life span: the effects of communication load and internet multitasking on perceived stress and psychological health impairments in a German probability sample. Media Psychol 2017 Jan 2;20(1):90-115. [doi: 10.1080/15213269.2015.1121832]
16. Barley SR, Meyerson DE, Grodal S. E-mail as a source and symbol of stress. Organ Sci 2011 Aug;22(4):887-906. [doi: 10.1287/orsc.1100.0573]
17. Göldağ B. An investigation of the relationship between university students' digital burnout levels and perceived stress levels. J learn teach digit age 2022;7(1):90-98. [doi: 10.53850/joltida.958039]
18. Becker L, Martin T, Rohleder N, Nieding G, Wannagat W. Physiological stress responses to digital single- and multitasking demands in younger and older adults. Psychoneuroendocrinology 2025 Apr;174:107376. [doi: 10.1016/j.psyneuen.2025.107376] [Medline: 39893951]
19. Campbell AJ, Cumming SR, Hughes I. Internet use by the socially fearful: addiction or therapy? Cyberpsychol Behav 2006 Feb;9(1):69-81. [doi: 10.1089/cpb.2006.9.69] [Medline: 16497120]
20. Stanković M, Nešić M, Čičević S, Shi Z. Association of smartphone use with depression, anxiety, stress, sleep quality, and internet addiction. Empirical evidence from a smartphone application. Pers Individ Dif 2021 Jan;168:110342. [doi: 10.1016/j.paid.2020.110342]
21. Khalili-Mahani N, Smyrnova A, Kakinami L. To each stress its own screen: a cross-sectional survey of the patterns of stress and various screen uses in relation to self-admitted screen addiction. J Med Internet Res 2019 Apr 2;21(4):e11485. [doi: 10.2196/11485] [Medline: 30938685]
22. Hampton K, Rainie L, Lu W, Shin I, Purcell K. Psychological stress and social media use. Pew Research Center. 2015. URL: https://www.pewresearch.org/internet/2015/01/15/psychological-stress-and-social-media-use-2/ [accessed 2025-12-11]

XSL•FO
RenderX

23. Nygaard M, Andersen TO, Rod NH. Can social connections become stressful? Exploring the link between social media use and perceived stress in cross-sectional and longitudinal analyses of 25,053 adults. J Ment Health 2024 Oct;33(5):596-604. [doi: 10.1080/09638237.2024.2332802] [Medline: 38545943]
24. McLaughlin B, Gotlieb MR, Mills DJ. Caught in a dangerous world: problematic news consumption and its relationship to mental and physical ill-being. Health Commun 2023 Dec;38(12):2687-2697. [doi: 10.1080/10410236.2022.2106086] [Medline: 35999665]
25. Grubbs JB, Volk F, Exline JJ, Pargament KI. Internet pornography use: perceived addiction, psychological distress, and the validation of a brief measure. J Sex Marital Ther 2015;41(1):83-106. [doi: 10.1080/0092623X.2013.842192] [Medline: 24341869]
26. Bezinović P, Roviš D, Rončević N, Bilajac L. Patterns of internet use and mental health of high school students in Istria County Croatia: cross-sectional study. Croat Med J 2015 Jun;56(3):297-305. [doi: 10.3325/cmj.2015.56.297] [Medline: 26088855]
27. Büchler N, ter Hoeven CL, van Zoonen W. Understanding constant connectivity to work: how and for whom is constant connectivity related to employee well-being? Inf Organ 2020 Sep;30(3):100302. [doi: 10.1016/j.infoandorg.2020.100302]
28. Elhai JD, Hall BJ. Anxiety about internet hacking: results from a community sample. Comput Human Behav 2016 Jan;54:180-185. [doi: 10.1016/j.chb.2015.07.057]
29. Li H, Ma X, Fang J, et al. Student stress and online shopping addiction tendency among college students in Guangdong Province, China: the mediating effect of the social support. IJERPH 2022;20(1):176. [doi: 10.3390/ijerph20010176]
30. Kelly CA, Sharot T. Web-browsing patterns reflect and shape mood and mental health. Nat Hum Behav 2025 Jan;9(1):133-146. [doi: 10.1038/s41562-024-02065-6] [Medline: 39572688]
31. Verma G, Bhardwaj A, Aledavood T, De Choudhury M, Kumar S. Examining the impact of sharing COVID-19 misinformation online on mental health. Sci Rep 2022 May 16;12(1):8045. [doi: 10.1038/s41598-022-11488-y] [Medline: 35577820]
32. Mun IB. Academic stress and first-/third-person shooter game addiction in a large adolescent sample: a serial mediation model with depression and impulsivity. Comput Human Behav 2023 Aug;145:107767. [doi: 10.1016/j.chb.2023.107767]
33. Park K, Son M, Chang H, Lee SK. The roles of stress, non-digital hobbies, and gaming time in adolescent problematic game use: a focus on sex differences. Comput Human Behav 2024 Feb;151:108002. [doi: 10.1016/j.chb.2023.108002]
34. Rosell J, Vergés A, Miranda-Castillo C, Sepúlveda-Caro S, Gómez M. Predictors, types of internet use, and the psychological well-being of older adults: a comprehensive model. J Gerontol B Psychol Sci Soc Sci 2022 Jul 5;77(7):1186-1196. [doi: 10.1093/geronb/gbac054] [Medline: 35286369]
35. van Ingen E, Utz S, Toepoel V. Online coping after negative life events: measurement, prevalence, and relation with internet activities and well-being. Soc Sci Comput Rev 2016;34(5):511-529. [doi: 10.1177/0894439315600322]
36. Boursier V, Musetti A, Gioia F, Flayelle M, Billieux J, Schimmenti A. Corrigendum: is watching tv series an adaptive coping strategy during the COVID-19 pandemic? Insights from an Italian community sample. Front Psychiatry 2021;12:698404. [doi: 10.3389/fpsyt.2021.698404] [Medline: 34149485]
37. Nabi RL, Torres DP, Prestin A. Guilty pleasure no more: the relative importance of media use for coping with stress. Journal of Media Psychology: Theories, Methods, and Applications 2017;29(3):126-136. [doi: 10.1027/1864-1105/a000223]
38. Maraz A, Yi S. Compulsive buying gradually increased during the first six months of the Covid-19 outbreak. J Behav Addict 2022 Mar 28;11(1):88-101. [doi: 10.1556/2006.2022.00002] [Medline: 35262509]
39. Šporčić B, Glavak-Tkalić R. The relationship between online gaming motivation, self-concept clarity and tendency toward problematic gaming. CP 2018;12(1). [doi: 10.5817/CP2018-1-4]
40. Graves BS, Hall ME, Dias-Karch C, Haischer MH, Apter C. Gender differences in perceived stress and coping among college students. PLOS ONE 2021;16(8):e0255634. [doi: 10.1371/journal.pone.0255634] [Medline: 34383790]
41. Matud MP. Gender differences in stress and coping styles. Pers Individ Dif 2004 Nov;37(7):1401-1415. [doi: 10.1016/j.paid.2004.01.010]
42. Almeida DM, Rush J, Mogle J, Piazza JR, Cerino E, Charles ST. Longitudinal change in daily stress across 20 years of adulthood: results from the national study of daily experiences. Dev Psychol 2023 Mar;59(3):515-523. [doi: 10.1037/dev0001469] [Medline: 36174182]
43. Johnson MD, Krahn HJ, Galambos NL. Perceived stress trajectories from age 25 to 50 years. Int J Behav Dev 2023 May;47(3):233-242. [doi: 10.1177/01650254221150887]
44. de Miquel C, Domènech-Abella J, Felez-Nobrega M, et al. The mental health of employees with job loss and income loss during the COVID-19 pandemic: the mediating role of perceived financial stress. Int J Environ Res Public Health 2022 Mar 8;19(6):3158. [doi: 10.3390/ijerph19063158] [Medline: 35328846]
45. Li R, Liu S, Huang C, Darabi D, Zhao M, Heinzel S. The influence of perceived stress and income on mental health in China and Germany. Heliyon 2023 Jun;9(6):e17344. [doi: 10.1016/j.heliyon.2023.e17344]
46. Rus HM, Tiemensma J. Social media under the skin: Facebook use after acute stress impairs cortisol recovery. Front Psychol 2017;8:1609. [doi: 10.3389/fpsyg.2017.01609] [Medline: 28974938]

XSL•FO
RenderX

47. Yang C, Mousavi S, Dash A, Gummadi KP, Weber I. Studying behavioral addiction by combining surveys and digital traces: a case study of tiktok. 2025 Presented at: Proceedings of the International AAAI Conference on Web and Social Media; Copenhagen, Denmark p. 2106-2123. [doi: 10.1609/icwsm.v19i1.35922]
48. Thomée S, Dellve L, Härenstam A, Hagberg M. Perceived connections between information and communication technology use and mental symptoms among young adults - a qualitative study. BMC Public Health 2010 Dec;10(1):1-14. [doi: 10.1186/1471-2458-10-66]
49. Dissing AS, Andersen TO, Jensen AK, Lund R, Rod NH. Nighttime smartphone use and changes in mental health and wellbeing among young adults: a longitudinal study based on high-resolution tracking data. Sci Rep 2022 May 15;12(1):8013. [doi: 10.1038/s41598-022-10116-z] [Medline: 35570230]
50. Wolfers LN, Utz S. Social media use, stress, and coping. Curr Opin Psychol 2022 Jun;45:101305. [doi: 10.1016/j.copsyc.2022.101305] [Medline: 35184027]
51. Afifi TD, Zamanzadeh N, Harrison K, Acevedo Callejas M. WIRED: the impact of media and technology use on stress (cortisol) and inflammation (interleukin IL-6) in fast paced families. Comput Human Behav 2018 Apr;81:265-273. [doi: 10.1016/j.chb.2017.12.010]
52. Thomas TA, Schmid AM, Kessling A, et al. Stress and compulsive buying-shopping disorder: a scoping review. Compr Psychiatry 2024 Jul;132:152482. [doi: 10.1016/j.comppsych.2024.152482] [Medline: 38603938]
53. Oulasvirta A, Rattenbury T, Ma L, Raita E. Habits make smartphone use more pervasive. Pers Ubiquit Comput 2012 Jan;16(1):105-114. [doi: 10.1007/s00779-011-0412-2]
54. Baryshnikov I, Aledavood T, Rosenström T, et al. Relationship between daily rated depression symptom severity and the retrospective self-report on PHQ-9: a prospective ecological momentary assessment study on 80 psychiatric outpatients. J Affect Disord 2023 Mar 1;324:170-174. [doi: 10.1016/j.jad.2022.12.127] [Medline: 36586594]
55. Hu Q, Li A, Heng F, Li J, Zhu T. Predicting depression of social media user on different observation windows. Presented at: 2015 IEEE / WIC / ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT); Dec 6-9, 2015; Singapore, Singapore p. 361-364. [doi: 10.1109/WI-IAT.2015.166]
56. Current population of germany. Statistisches Bundesamt (Destatis). 2023 Jun 20. URL: https://www.destatis.de/EN/Themes/Society-Environment/Population/Current-Population/_node.html#sprg480730 [accessed 2025-02-07]
57. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46. [doi: 10.1177/001316446002000104]
58. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [doi: 10.2307/2136404] [Medline: 6668417]
59. Philpott LF, Leahy-Warren P, FitzGerald S, Savage E. Prevalence and associated factors of paternal stress, anxiety, and depression symptoms in the early postnatal period. Glob Ment Health (Camb) 2022;9:306-321. [doi: 10.1017/gmh.2022.33] [Medline: 36561920]
60. Biswas B, Saha R, Haldar D, Saha I. Level of stress perception and predictors of higher stress perception among informal primary caregivers of Eastern Indian people living with HIV/AIDS. Int J Community Med Public Health 2019;6(10):4374. [doi: 10.18203/2394-6040.ijcmph20194497]
61. Bryk AS, Raudenbush SW. Hierarchical Linear Models: Applications and Data Analysis Methods: sage; 2002, Vol. 1.
62. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. 2010 Presented at: Python in Science Conference; Jun 28, 2010. [doi: 10.25080/Majora-92bf1922-011]
63. Thorn L, Evans P, Cannon A, Hucklebridge F, Clow A. Seasonal differences in the diurnal pattern of cortisol secretion in healthy participants and those with self-assessed seasonal affective disorder. Psychoneuroendocrinology 2011 Jul;36(6):816-823. [doi: 10.1016/j.psyneuen.2010.11.003] [Medline: 21145663]
64. Gassen J, Mengelkoch S, Slavich GM. Human immune and metabolic biomarker levels, and stress-biomarker associations, differ by season: Implications for biomedical health research. Brain, Behavior, & Immunity - Health 2024 Jul;38:100793. [doi: 10.1016/j.bbih.2024.100793]
65. Average daily time spent on internet and online media use in germany in the 3rd quarter of 2023, by device. Statista. 2024. URL: https://www.statista.com/statistics/719939/average-daily-internet-and-social-media-use-in-germany-by-device/ [accessed 2025-12-11]
66. Beyens I, Frison E, Eggermont S. “I don’t want to miss a thing”: adolescents’ fear of missing out and its relationship to adolescents’ social needs, Facebook use, and Facebook related stress. Comput Human Behav 2016 Nov;64:1-8. [doi: 10.1016/j.chb.2016.05.083]
67. Åberg E, Koivula A, Kukkonen I. A feminine burden of perfection? Appearance-related pressures on social networking sites. Telematics and Informatics 2020 Mar;46:101319. [doi: 10.1016/j.tele.2019.101319]
68. Chen W, Lee KH. Sharing, liking, commenting, and distressed? The pathway between Facebook interaction and psychological distress. Cyberpsychol Behav Soc Netw 2013 Oct;16(10):728-734. [doi: 10.1089/cyber.2012.0272] [Medline: 23745614]
69. Rus HM, Tiemensma J. Social media as a shield: Facebook buffers acute stress. Physiol Behav 2018 Mar;185:46-54. [doi: 10.1016/j.physbeh.2017.12.021]

XSL•FO
RenderX

70. Wolfers LN, Festl R, Utz S. Do smartphones and social network sites become more important when experiencing stress? Results from longitudinal data. Comput Human Behav 2020 Aug;109:106339. [doi: 10.1016/j.chb.2020.106339] [Medline: 32747849]
71. Brailovskaia J, Rohmann E, Bierhoff HW, Schillack H, Margraf J. The relationship between daily stress, social support and Facebook addiction disorder. Psychiatry Res 2019 Jun;276:167-174. [doi: 10.1016/j.psychres.2019.05.014] [Medline: 31096147]
72. Shen X, Wang JL. Loneliness and excessive smartphone use among Chinese college students: moderated mediation effect of perceived stressed and motivation. Comput Human Behav 2019 Jun;95:31-36. [doi: 10.1016/j.chb.2019.01.012]
73. Aghababian AH, Sadler JR, Jansen E, Thapaliya G, Smith KR, Carnell S. Binge watching during COVID-19: associations with stress and body weight. Nutrients 2021 Sep 28;13(10):3418. [doi: 10.3390/nu13103418] [Medline: 34684420]
74. Alimoradi Z, Jafari E, Potenza MN, Lin CY, Wu CY, Pakpour AH. Binge-watching and mental health problems: a systematic review and meta-analysis. Int J Environ Res Public Health 2022 Aug 6;19(15):9707. [doi: 10.3390/ijerph19159707] [Medline: 35955069]
75. Hunter IR, Gillen MC. Stress coping mechanisms in elderly adults: an initial study of recreational and other coping behaviors in nursing home patients. Adultspan Journal 2009 Mar;8(1):43-53 [FREE Full text] [doi: 10.1002/j.2161-0029.2009.tb00056.x]
76. van Gelder K. E-commerce worldwide - statistics & facts. Statista. URL: https://www.statista.com/topics/871/online-shopping/ [accessed 2025-12-11]
77. Tarka P, Kukar-Kinney M. Compulsive buying among young consumers in Eastern Europe: a two-study approach to scale adaptation and validation. JCM 2022 Feb 9;39(1):106-120. [doi: 10.1108/JCM-05-2020-3833]
78. Zheng Y, Yang X, Liu Q, Chu X, Huang Q, Zhou Z. Perceived stress and online compulsive buying among women: a moderated mediation model. Comput Human Behav 2020 Feb;103:13-20. [doi: 10.1016/j.chb.2019.09.012]
79. Singh R, Nayak JK. Life stressors and compulsive buying behaviour among adolescents in India. SAJGBR 2015 Aug 3;4(2):251-274. [doi: 10.1108/SAJGBR-08-2014-0054]
80. Hama Y. Shopping as a coping behavior for stress. Jpn Psychol Res 2001 Nov;43(4):218-224. [doi: 10.1111/1468-5884.00179]
81. Maharani SAD, Utami NP. Coping mechanisms of stress: the impact on online purchase impulsivity. JBMS 2023;3(3):164-180 [FREE Full text] [doi: 10.53748/jbms.v3i3.70]
82. Durante KM, Laran J. The effect of stress on consumer. Journal of Marketing Research 2016;53:814-828. [doi: 10.1509/jmr.15.0319]
83. Gallagher CE, Watt MC, Weaver AD, Murphy KA. “I fear, therefore, I shop!” exploring anxiety sensitivity in relation to compulsive buying. Pers Individ Dif 2017 Jan;104:37-42. [doi: 10.1016/j.paid.2016.07.023]
84. Games - worldwide. Statista. 2025. URL: https://www.statista.com/outlook/amo/media/games/worldwide [accessed 2025-12-11]
85. Lemmens JS, Valkenburg PM, Peter J. Psychosocial causes and consequences of pathological gaming. Comput Human Behav 2011 Jan;27(1):144-152. [doi: 10.1016/j.chb.2010.07.015]
86. Whitbourne SK, Ellenberg S, Akimoto K. Reasons for playing casual video games and perceived benefits among adults 18 to 80 years old. Cyberpsychol Behav Soc Netw 2013 Dec;16(12):892-897. [doi: 10.1089/cyber.2012.0705] [Medline: 23971430]
87. Desai V, Gupta A, Andersen L, Ronnestrand B, Wong M. Stress-reducing effects of playing a casual video game among undergraduate students. Trends Psychol 2021;29(3):563-579. [doi: 10.1007/s43076-021-00062-6] [Medline: 40477391]
88. Lee YH, Chen M. Seeking a sense of control or escapism? The role of video games in coping with unemployment. Games and Culture 2023 May;18(3):339-361. [doi: 10.1177/15554120221097413]
89. Snodgrass JG, Lacy MG, Dengah HJF II, Eisenhauer S, Batchelder G, Cookson RJ. A vacation from your mind: problematic online gaming is a stress response. Comput Human Behav 2014 Sep;38:248-260. [doi: 10.1016/j.chb.2014.06.004]
90. Seçer İ, Us E. Digital gaming trends of middle-aged and older adults: a sample from Turkey. Simul Gaming 2023 Feb;54(1):85-103. [doi: 10.1177/10468781221144184]
91. Engelstätter B, Ward MR. Video games become more mainstream. Entertain Comput 2022 May;42:100494. [doi: 10.1016/j.entcom.2022.100494]
92. Aledavood T, Lehmann S, Saramäki J. Digital daily cycles of individuals. Front Phys 2015;3:73. [doi: 10.3389/fphy.2015.00073]
93. Luong N, Barnett I, Aledavood T. The impact of the COVID-19 pandemic on daily rhythms. J Am Med Inform Assoc 2023 Nov 17;30(12):1943-1953. [doi: 10.1093/jamia/ocad140]
94. Schrempft S, Baysson H, Chessa A, et al. Associations between bedtime media use and sleep outcomes in an adult population-based cohort. Sleep Med 2024 Sep;121:226-235. [doi: 10.1016/j.sleep.2024.06.029]
95. Siebers T, Beyens I, Baumgartner SE, Valkenburg PM. Adolescents’ digital nightlife: the comparative effects of day- and nighttime smartphone use on sleep quality. Communic Res 2024. [doi: 10.1177/00936502241276793]
96. Luqman A, Masood A, Shahzad F, Shahbaz M, Feng Y. Untangling the adverse effects of late-night usage of smartphone-based SNS among university students. Behav Inf Technol 2021 Nov 18;40(15):1671-1687. [doi: 10.1080/0144929X.2020.1773538]

XSL•FO
RenderX

97. Thomée S, Härenstam A, Hagberg M. Mobile phone use and stress, sleep disturbances, and symptoms of depression among young adults - a prospective cohort study. BMC Public Health 2011 Dec;11(1):1-11. [doi: 10.1186/1471-2458-11-66]
98. Thomée S, Härenstam A, Hagberg M. Computer use and stress, sleep disturbances, and symptoms of depression among young adults – a prospective cohort study. BMC Psychiatry 2012 Dec;12(1):1-14. [doi: 10.1186/1471-244X-12-176]
99. Price GD, Heinz MV, Song SH, Nemesure MD, Jacobson NC. Using digital phenotyping to capture depression symptom variability: detecting naturalistic variability in depression symptoms across one year using passively collected wearable movement and sleep data. Transl Psychiatry 2023 Dec 9;13(1):381. [doi: 10.1038/s41398-023-02669-y] [Medline: 38071317]
100. Kortesoja L, Vainikainen MP, Hotulainen R, Merikanto I. Late-night digital media use in relation to chronotype, sleep and tiredness on school days in adolescence. J Youth Adolesc 2023 Feb;52(2):419-433. [doi: 10.1007/s10964-022-01703-4] [Medline: 36401709]
101. Bőthe B, Nagy L, Koós M, et al. Problematic pornography use across countries, genders, and sexual orientations: insights from the International Sex Survey and comparison of different assessment tools. Addiction 2024 May;119(5):928-950. [doi: 10.1111/add.16431] [Medline: 38413365]
102. Laier C, Brand M. Mood changes after watching pornography on the Internet are linked to tendencies towards Internet-pornography-viewing disorder. Addict Behav Rep 2017 Jun;5:9-13. [doi: 10.1016/j.abrep.2016.11.003] [Medline: 29450222]
103. Altin M, De Leo D, Tribbia N, Ronconi L, Cipolletta S. Problematic pornography use, mental health, and suicidality among young adults. Int J Environ Res Public Health 2024 Sep 18;21(9):1228. [doi: 10.3390/ijerph21091228] [Medline: 39338111]
104. Harper C, Hodgins DC. Examining correlates of problematic internet pornography use among university students. J Behav Addict 2016 Jun;5(2):179-191. [doi: 10.1556/2006.5.2016.022] [Medline: 27156383]
105. McCormack M, Wignall L. Enjoyment, exploration and education: understanding the consumption of pornography among young men with non-exclusive sexual orientations. Sociology 2017 Oct;51(5):975-991. [doi: 10.1177/0038038516629909] [Medline: 28989197]
106. Bőthe B, Tóth-Király I, Bella N, Potenza MN, Demetrovics Z, Orosz G. Why do people watch pornography? The motivational basis of pornography use. Psychol Addict Behav 2021 Mar;35(2):172-186. [doi: 10.1037/adb0000603] [Medline: 32730047]
107. Mubeen B, Ashraf D. Psychological predictors of adult content consumption: a qualitative analysis by grounded theory. Asian J Psychiatr 2022;23(7). [doi: 10.54615/2231-7805.47266]
108. Moynihan AB, Igou ER, van Tilburg WAP. Pornography consumption as existential escape from boredom. Pers Individ Dif 2022 Nov;198:111802. [doi: 10.1016/j.paid.2022.111802]
109. Federico B. ICT and productivity: a review of the literature. : European Commission, Joint Research Centre (JRC); 2013. [doi: 10.2788/32940]
110. Buhari A, A. AA. Influence of internet and its connectivity in workplace - a comprehensive analysis. RRRJ 2024;3(1):244-257 [FREE Full text] [doi: 10.36548/rrrj.2024.1.016]
111. Mark G, Voida S, Cardello A. A pace not dictated by electrons": an empirical study of work without email. CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems :555-564. [doi: 10.1145/2207676.2207754]
112. Ninaus K, Diehl S, Terlutter R, Chan K, Huang A, Erlandsson S. Benefits and stressors - perceived effects of ICT use on employee health and work stress: an exploratory study from Austria and Hong Kong. Int J Qual Stud Health Well-being 2015;10:28838. [doi: 10.3402/qhw.v10.28838] [Medline: 26462972]
113. Berg-Beckhoff G, Nielsen G, Ladekjær Larsen E. Use of information communication technology and stress, burnout, and mental health in older, middle-aged, and younger workers - results from a systematic review. Int J Occup Environ Health 2017 Apr;23(2):160-171. [doi: 10.1080/10773525.2018.1436015] [Medline: 29460697]
114. Malott L, Vishwanathan SP, Chellappan S. Differences in internet usage patterns with stress and anxiety among college students. Presented at: 2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013); Oct 9-12, 2013; Lisbon, Portugal p. 664-668. [doi: 10.1109/HealthCom.2013.6720759]
115. Allen MT. Explorations of avoidance and approach coping and perceived stress with a computer-based avatar task: detrimental effects of resignation and withdrawal. PeerJ 2021;9:e11265. [doi: 10.7717/peerj.11265]
116. Holahan CJ, Moos RH, Holahan CK, Brennan PL, Schutte KK. Stress generation, avoidance coping, and depressive symptoms: a 10-year model. J Consult Clin Psychol 2005 Aug;73(4):658-666. [doi: 10.1037/0022-006X.73.4.658] [Medline: 16173853]
117. Shaffer SR, Dolovich C, El-Gabalawy R, et al. The impact of source and consumption of news on mental distress among inflammatory bowel disease patients during the COVID-19 pandemic. J Can Assoc Gastroenterol 2024 Apr;7(2):212-218. [doi: 10.1093/jcag/gwad060] [Medline: 38596803]
118. Ladis I, Gao C, Scullin MK. COVID-19-related news consumption linked with stress and worry, but not sleep quality, early in the pandemic. Psychol Health Med 2023 Apr;28(4):980-994. [doi: 10.1080/13548506.2022.2141281] [Medline: 36322027]
119. Kellerman JK, Hamilton JL, Selby EA, Kleiman EM. The mental health impact of daily news exposure during the COVID-19 pandemic: Ecological Momentary Assessment study. JMIR Ment Health 2022 May 25;9(5):e36966. [doi: 10.2196/36966] [Medline: 35377320]
120. Nguyen A, Smith A, Jackson D, Zhao X. Pandemic news experience: COVID-19, news consumption, mental health, and the demand for positive news. SSRN Journal 2021. [doi: 10.2139/ssrn.3832669]

XSL•FO
RenderX

121. Lindell J, Mikkelsen Båge E. Disconnecting from digital news: news avoidance and the ignored role of social class. Journalism 2023 Sep;24(9):1980-1997. [doi: 10.1177/14648849221085389]
122. Mannell K, Meese J. From doom-scrolling to news avoidance: limiting news as a wellbeing strategy during COVID lockdown. Journal Stud 2022 Feb 17;23(3):302-319. [doi: 10.1080/1461670X.2021.2021105]
123. McNaughton-cassill ME. The news media and psychological distress. Anxiety, Stress & Coping 2001 Apr;14(2):193-211. [doi: 10.1080/10615800108248354]
124. Lavelle B. Investigating whether the consumption of news impacts measures for anxiety, stress, and well-being. Https://norma.ncirl.ie/id/eprint/5657. 2022.
125. Boukes M, Vliegenthart R. News consumption and its unpleasant side effect. J Media Psychol 2017 Jul;29(3):137-147. [doi: 10.1027/1864-1105/a000224]
126. Mobile messaging users worldwide 2025. Statista. 2021. URL: https://www.statista.com/statistics/483255/number-of-mobile-messaging-users-worldwide/ [accessed 2025-12-11]
127. Coccia C, Darling CA. Having the time of their life: college student stress, dating and satisfaction with life. Stress Health 2016 Feb;32(1):28-35. [doi: 10.1002/smi.2575] [Medline: 24723539]
128. Thomée S, Eklöf M, Gustafsson E, Nilsson R, Hagberg M. Prevalence of perceived stress, symptoms of depression and sleep disturbances in relation to information and communication technology (ICT) use among young adults – an explorative prospective study. Comput Human Behav 2007 May;23(3):1300-1321. [doi: 10.1016/j.chb.2004.12.007]
129. Hurbean L, Dospinescu O, Munteanu V, Danaiata D. Effects of instant messaging related technostress on work performance and well-being. Electronics (Basel) 2022;11(16):2535. [doi: 10.3390/electronics11162535]
130. Lin XY, Lachman ME. Daily stress and affect across adulthood: the role of social interactions via different communication modes. Technol Mind Behav 2021;2(1):10. [doi: 10.1037/tmb0000026] [Medline: 35369392]
131. Melumad S, Pham MT. The smartphone as a pacifying technology. J Consum Res 2020 Aug 1;47(2):237-255. [doi: 10.1093/jcr/ucaa005]
132. Yau JC, Reich SM, Lee TY. Coping with stress through texting: an experimental study. J Adolesc Health 2021 Mar;68(3):565-571. [doi: 10.1016/j.jadohealth.2020.07.004] [Medline: 32798096]
133. Holtzman S, DeClerck D, Turcotte K, Lisi D, Woodworth M. Emotional support during times of stress: Can text messaging compete with in-person interactions? Comput Human Behav 2017 Jun;71:130-139. [doi: 10.1016/j.chb.2017.01.043]
134. Hooker ED, Campos B, Pressman SD. It just takes a text: partner text messages can reduce cardiovascular responses to stress in females. Comput Human Behav 2018 Jul;84:485-492. [doi: 10.1016/j.chb.2018.02.033]
135. Smith A. How Americans use text messaging. Pew Research Center. URL: https://www.pewresearch.org/internet/2011/09/19/how-americans-use-text-messaging/ [accessed 2025-12-11]
136. Aguilera A, Muñoz RF. Text messaging as an adjunct to CBT in low-income populations: a usability and feasibility pilot study. Prof Psychol Res Pr 2011 Dec 1;42(6):472-478. [doi: 10.1037/a0025499] [Medline: 25525292]
137. George MJ, Beron K, Vollet JW, Burnell K, Ehrenreich SE, Underwood MK. Frequency of text messaging and adolescents' mental health symptoms across 4 years of high school. J Adolesc Health 2021 Feb;68(2):324-330. [doi: 10.1016/j.jadohealth.2020.06.012] [Medline: 32753344]
138. Yu CC, Tou NX, Low JA. Internet use and effects on mental well-being during the lockdown phase of the COVID-19 pandemic in younger versus older adults: observational cross-sectional study. JMIR Form Res 2024 Feb 6;8:e46824. [doi: 10.2196/46824] [Medline: 38319700]
139. Elamin NO, Almasaad JM, Busaeed RB, Aljafari DA, Khan MA. Smartphone addiction, stress, and depression among university students. Clin Epidemiol Glob Health 2024 Jan;25:101487. [doi: 10.1016/j.cegh.2023.101487]
140. Khan A, McLeod G, Hidajat T, Edwards EJ. Excessive smartphone use is associated with depression, anxiety, stress, and sleep quality of Australian adults. J Med Syst 2023 Oct 20;47(1):109. [doi: 10.1007/s10916-023-02005-3] [Medline: 37858009]
141. Ding L, Li Z, Jiang H, Zhang X, Xiong Z, Zhu X. Mobile phone problem use and depressive symptoms: the mediating role of social support and attitude to aging among Chinese older adults. BMC Psychiatry 2024 Feb 16;24(1):135. [doi: 10.1186/s12888-024-05565-x] [Medline: 38365625]
142. Yue C, Ware S, Morillo R, et al. Automatic depression prediction using Internet traffic characteristics on smartphones. Smart Health (2014) 2020 Nov;18:100137. [doi: 10.1016/j.smhl.2020.100137]
143. Bartone PT, Ursano RJ, Wright KM, Ingraham LH. The impact of a military air disaster on the health of assistance workers. A prospective study. J Nerv Ment Dis 1989 Jun;177(6):317-328. [doi: 10.1097/00005053-198906000-00001] [Medline: 2723619]
144. Kulshrestha J, Oliveira M, Karaçalık O, Bonnay D, Wagner C. Web routineness and limits of predictability: investigating demographic and behavioral differences using web tracking data. ICWSM 2021;15:327-338. [doi: 10.1609/icwsm.v15i1.18064]
145. Stier S, Kirkizh N, Froio C, Schroeder R. Populist attitudes and selective exposure to online news: a cross-country analysis combining web tracking and surveys. Int J Press Polit 2020 Jul;25(3):426-446. [doi: 10.1177/1940161220907018]
146. Grothaus C, Dolch C, Zawacki-Richter O. Use of digital media in higher education across country contexts: a comparison between Germany and Thailand. Int J Emerg Technol Learn 2021;16(20):64. [doi: 10.3991/ijet.v16i20.24263]

XSL•FO
RenderX

147. Lopez-Fernandez O, Romo L, Kern L, et al. Problematic internet use among adults: a cross-cultural study in 15 countries. J Clin Med 2023 Jan 29;12(3):1027. [doi: 10.3390/jcm12031027] [Medline: 36769675]

Abbreviations

LMM: linear mixed-effects model
PSS: Perceived Stress Scale

Edited by K Liew; submitted 09.Jun.2025; peer-reviewed by A Hinz, S Biswas, Y Zhang; revised version received 17.Sep.2025; accepted 06.Oct.2025; published 09.Jan.2026.

Please cite as:
Belal M, Luong N, Aledavood T, Kulshrestha J
Examining the Association Between Internet Use and Perceived Stress in Adults: Longitudinal Observational Study Combining Web Tracking Data With Questionnaires
J Med Internet Res 2026;28:e78775
URL: https://www.jmir.org/2026/1/e78775
doi:10.2196/78775

© Mohammad Belal, Nguyen Luong, Talayeh Aledavood, Juhi Kulshrestha. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Impact of Evaluation Strategy on Sepsis Prediction Model Performance Metrics in Intensive Care Data: Retrospective Cohort Study

Dang-Khoa Do[1], State Examination in Medicine; Patrick Rockenschaub[2], PhD; Sebastian Daniel Boie[1], PhD; Oliver Kumpf[3], Dr med; Hans-Dieter Volk[4], Dr med; Felix Balzer[1], Dr med, Dr rer nat; Falk von Dincklage[5], Dr med; Gregor Lichtner[1,5], Dr rer nat

[1]Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[2]Institute of Clinical Epidemiology, Public Health, Health Economics, Medical Statistics and Informatics, Medical University of Innsbruck, Innsbruck, Austria

[3]Department of Anesthesiology and Intensive Care Medicine (CCM/CVK), Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[4]Institute of Medical Immunology, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[5]Department of Anesthesia, Critical Care, Emergency and Pain Medicine, Universitätsmedizin Greifswald, Greifswald, Germany

Corresponding Author:
Gregor Lichtner, Dr rer nat
Department of Anesthesia, Critical Care, Emergency and Pain Medicine
Universitätsmedizin Greifswald
Ferdinand-Sauerbruch-Straße
Greifswald, 17475
Germany
Phone: 49 3834 86 5801
Email: gregor.lichtner@med.uni-greifswald.de

Abstract

Background: The prediction of the onset of sepsis, a life-threatening condition resulting from a dysregulated response to an infection, is one of the most common prediction tasks in intensive care–related machine learning research. To assess the performance of such models, different evaluation strategies, including fixed horizon (a single prediction at a set time before onset), peak score (a single prediction using the maximum predicted risk across time), and continuous evaluation (multiple predictions assessed continuously across time), are commonly implemented, but there is no clear consensus on which approach should be used in order to provide clinically meaningful performance evaluation.

Objective: This study aimed to assess different evaluation approaches of sepsis prediction models trained on a public intensive care dataset applied to German intensive care data.

Methods: In this retrospective, observational cohort study, we assessed the efficacy of machine learning models, pretrained on the Medical Information Mart for Intensive Care IV dataset, when applied to BerlinICU, a multisite German intensive care dataset. To understand the real-world impact of implementing these models, we examined the performance variability across various evaluation strategies.

Results: The BerlinICU dataset includes 40,132 intensive care admissions spanning 10 years (2012-2021). Using the latest Sepsis-3 definition, we identified 4134 septic admissions (10.3% prevalence). Application of a temporal convolutional network model to BerlinICU yielded an area under the receiver operating characteristic curve (AUROC) of 0.67 (95% CI 0.66-0.68) for continuous evaluation with a 6-hour prediction horizon, compared with 0.84 (95% CI 0.83-0.85) on the test set of Medical Information Mart for Intensive Care IV. On BerlinICU, peak score evaluation showed a similar AUROC compared with continuous evaluation, while fixed horizon evaluation showed a reduced AUROC of 0.61 (95% CI 0.60-0.62). Onset matching had minimal impact on performance estimates using continuous evaluation or fixed horizon evaluation, but increased estimates for peak score evaluation. Performance metrics improved with shorter prediction horizons across all strategies.

XSL•FO
RenderX

Conclusions: Our results demonstrate that the choice of evaluation strategy has a significant impact on the performance metrics of intensive care prediction models. The same model applied to the same dataset yields markedly different performance metrics depending on the evaluation approach. Therefore, careful selection of the evaluation approach is essential to ensure that the interpretation of performance metrics aligns with clinical intentions and enables meaningful comparisons between studies. In our view, the continuous evaluation approach best reflects the continual monitoring of patients that is performed in real-world clinical practice. In contrast, fixed-horizon and peak score evaluation approaches may produce skewed results when not properly matching the length of stay distributions between sepsis cases and control cases. Especially for peak score evaluation, longer visits tend to produce higher maximum scores because sampling from more values increases the likelihood of capturing higher values purely by chance.

(J Med Internet Res 2026;28:e72083) doi:10.2196/72083

KEYWORDS

sepsis; machine learning; deep learning; intensive care; intensive care unit; ICU; early prediction; early warning; electronic health records; prediction; generalizability

Introduction

Sepsis is a life-threatening condition resulting from a dysregulated response to an infection that leads to organ dysfunction [1], responsible for up to 20% of all deaths worldwide [2]. Beyond its significant mortality rate [3], sepsis represents a tremendous financial burden, totaling to almost US $24 billion in the United States in 2013 [4]. Those who survive may experience long-term health consequences, including a diminished quality of life [5].

As each hour of delayed therapy increases mortality [6-8], sepsis treatment guidelines continue to recommend early therapy within the first hour of sepsis recognition [9-11]. Therefore, an early recognition of sepsis is essential. Numerous studies have investigated machine learning to aid early prediction of sepsis in the intensive care unit (ICU) [12,13], including prospective application [14,15] and external validation on data of different hospitals [14,16-18]. However, the strategies used in previous studies to evaluate model performance vary widely, particularly in the criteria for data selection and how predictions are aligned with respect to the sepsis onset. The most commonly used evaluation strategies can be broadly categorized into three groups, (1) fixed horizon evaluation, which assesses the model's ability to predict sepsis at a fixed number of hours before onset [13,19]; (2) peak score evaluation, which uses the highest prediction score across time before onset [16,20]; and (3) continuous evaluation, which considers all prediction scores across time before onset [13].

While these strategies are widely used in the literature, it remains uncertain how the choice among them influences the estimated model performance and whether certain strategies may systematically inflate or deflate metrics compared with others, even when applied to the exact same model and dataset. This has direct implications for clinical deployment and benchmarking: if a model is selected based on an evaluation strategy that does not reflect its intended clinical use (eg, hourly real-time monitoring vs one-time assessment), its actual performance may fall short of expectations. Similarly, comparing models across studies without accounting for differences in evaluation strategy risks drawing misleading conclusions about their relative performance.

In addition to differences in evaluation strategy, the literature also uses a range of performance metrics. Standard measures include the area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), and threshold-based metrics, such as positive predictive value (PPV) and negative predictive value (NPV). Threshold-based metrics reflect clinical utility at specific decision points, which is critical for deployment. In contrast, AUROC is threshold-agnostic, capturing model discrimination across all thresholds, and is robust to class imbalance, which makes it the predominant choice in the literature [21]. Beyond standard performance metrics, recent sepsis-prediction challenges have adopted clinically informed cost functions that capture specific clinical trade-offs. The PhysioNet/CinC 2019 challenge used a time-weighted continuous utility score, rewarding early detection while penalizing very early, late, and false alarms across the entire ICU trajectory. In contrast, the Pediatric Sepsis Data Challenge evaluates a single admission-time prediction per patient and scores models by their true-positive rate at a fixed maximal false-positive rate [22,23]. These examples highlight that not only evaluation strategies, but also the choice of performance metric, can fundamentally shape how sepsis prediction models are interpreted and compared.

In this study, we aimed to investigate how the choice of the evaluation approach influences the estimated performance of sepsis prediction models in an external validation context. Specifically, we compared the performance of models trained on a public intensive care dataset from the United States and subsequently applied them to German intensive care data. In doing so, we also address the lack of studies involving large-scale ICU cohorts from the German health care context, where the efficacy of sepsis prediction models—given the distinct patient demographics, treatment approaches, and health care policies—remains largely uninvestigated.

Methods

Study Design and Setting

This is a retrospective cohort study for evaluating differences in performance metrics under different, commonly used evaluation strategies. Temporal convolutional network (TCN) models were developed on Medical Information Mart for Intensive Care IV (MIMIC-IV v2.0; USA) and applied to

XSL•FO
RenderX

BerlinICU, a cohort of adult ICU stays from 8 ICUs at Charité—Universitätsmedizin Berlin across 2 sites (Campus Virchow Klinikum and Campus Charité Mitte) from 2012 to 2021. No clinical intervention or change in patient management was implemented; all data were collected during routine care.

Participants and Flow

We included ICU stays from adult patients (≥18 y). We excluded stays originating from non-ICU wards, stays with <6 hours of data, gaps ≥12 hours, sepsis onset within 4 hours after admission, and truncated each stay at 7 days. Evaluation samples for different evaluation strategies were drawn from this cohort.

Primary and Secondary Outcome Measures

The clinical outcome was sepsis onset, defined according to Sepsis-3 with an adapted suspicion-of-infection definition for BerlinICU (refer to "Outcome definition" section). The primary performance measure used to compare evaluation strategies was the AUROC. Secondary performance measures included the AUPRC, PPV, NPV, and recall at fixed thresholds. We additionally report "lifted" PPV and AUPRC (each divided by the horizon-specific prevalence) and recall at the threshold where precision equals twice the prevalence, as well as Brier score, alarm rate, and net benefit to assess calibration and clinical utility.

Datasets

We pretrained models on the MIMIC-IV dataset (version 2.0), a well-known ICU dataset from the United States that is commonly used to develop clinical prediction models [12,13]. We applied these models to BerlinICU, a comprehensive dataset from 8 ICUs from Charité - Universitätsmedizin Berlin in Germany, one of Europe's largest university hospitals. The ICUs were distributed across 2 different sites (Campus Virchow Klinikum and Campus Charité Mitte).

BerlinICU was extracted from the intensive care data management systems and the hospital information system present at the Charité - Universitätsmedizin Berlin. Specifically, we extracted the same clinical features that we used to pretrain the models, resulting in a total of 48 time-varying features, including vital parameters and laboratory results, along with 4 static features (age, sex, height, and weight; Table S1 in Multimedia Appendix 1). These features comprise demographic variables, vital signs, arterial blood gas values, electrolyte and blood count parameters, as well as treatment-related variables (eg, fraction of inspired oxygen) and clinical output measures (eg, urine output). All clinical feature values outside of physiologically plausible ranges, as defined by the ricu package [24], were removed. The remaining data was aggregated into intervals of one hour using their median value. Missing values were imputed using a forward fill strategy. If no previous values were available, the training set mean across all patients, regardless of sepsis status, for that feature was used. This corresponds to zero imputation if performed after feature standardization and is common practice in machine learning studies on sepsis prediction [17,25,26]. Outlier influence was mitigated by previous filtering based on physiological plausibility. Data were excluded from the dataset if they originated from nonintensive care wards, involved patients younger than 18 years, had a sepsis onset within 4 hours after admission, contained less than 6 hours of data, or included gaps equal to or exceeding 12 hours. The dataset was restricted to data from the years 2012-2021 and limited to at most 7 days following admission to the ICU.

Ethical Considerations

This study was approved by the local ethics committee (Ethikausschuss am Campus Virchow-Klinikum, Charité—Universitätsmedizin Berlin, Chairperson PD Dr E Kaschina, application EA2/137/22, approval date July 12, 2022, and amendment date May 10, 2023). Informed consent was not required due to the retrospective, observational nature of the study and the use of routinely collected data in accordance with applicable regulations. No compensation was provided.

Outcome Definition

To identify the onset of sepsis, we used the latest Sepsis-3 definition criteria, which require the co-occurrence of a suspicion of infection (SI) and an increase of the Sequential Organ Failure Assessment (SOFA) score by at least 2 points [1]. An SI is present when antibiotic treatment is coupled with microbiological culture tests. As the BerlinICU dataset did not include microbiological data, we used an alternative definition for SI in line with previous studies [16,17]. We inferred SI if a patient was treated with antibiotics administered at least once every 24 hours for at least 3 days, and defined the time of suspicion as the beginning of this treatment. Following previous studies, we defined an SI window ranging from 48 hours before to 24 hours after the time of SI (Figure 1). We assumed the onset of sepsis if an increase of at least 2 points of the SOFA score occurred during that SI window. The sepsis onset time point was defined as the time point of the recorded SOFA increase.

XSL•FO
RenderX

Figure 1. Preprocessing Pipeline. The BerlinICU dataset was extracted from the hospital information system and cleaned. We inferred sepsis onset according to the Sepsis-3 definition. For the logistic regression baseline models, we created additional temporal features. The final dataset was then passed to the models and evaluated. ICU: intensive care unit; MAP: mean arterial pressure; MIMIC-IV: Medical Information Mart for Intensive Care IV; SI: suspicion of infection; SOFA: Sequential Organ Failure Assessment score.

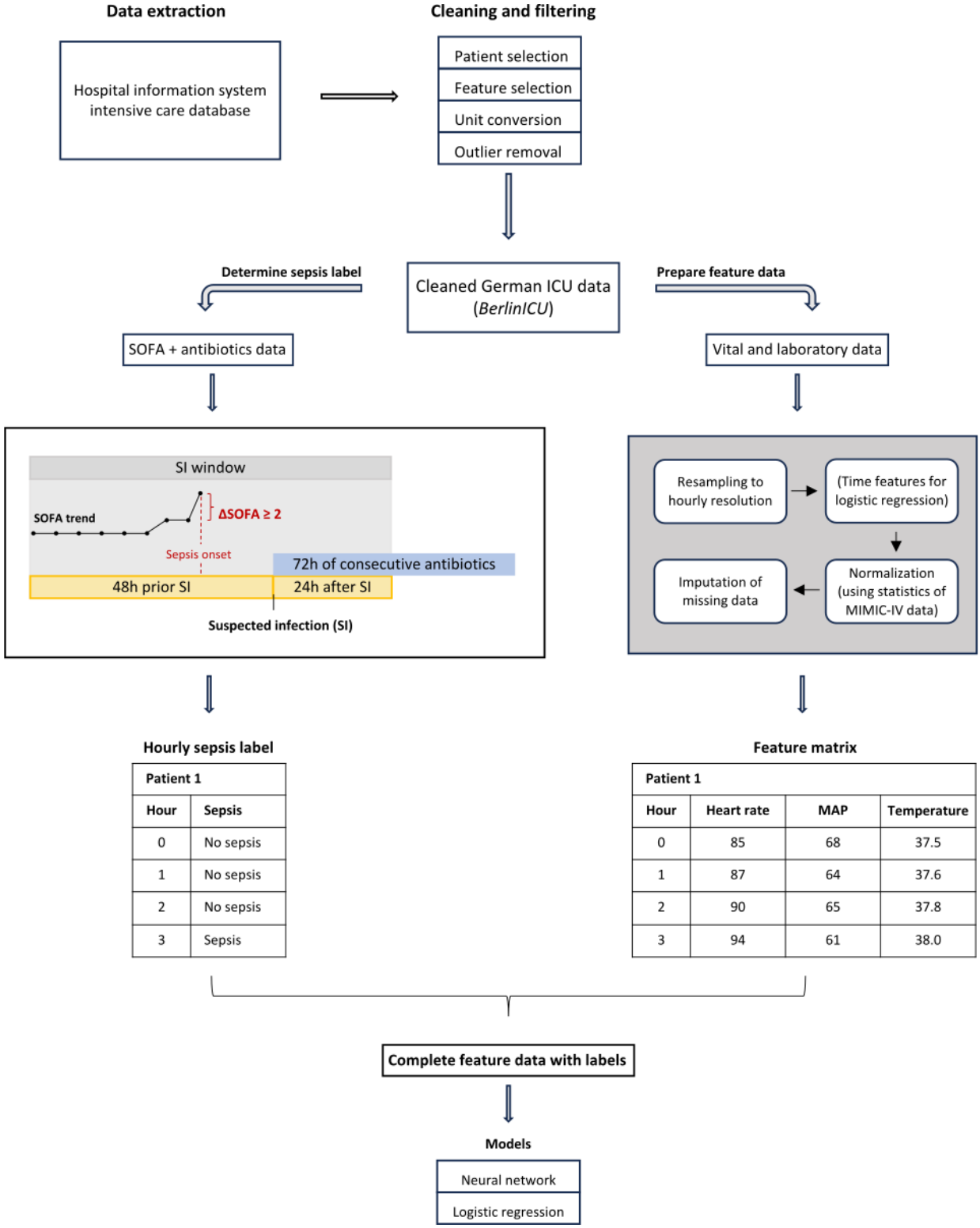

Model Training

We used TCN models trained on the MIMIC-IV dataset from our previous study [17]. This choice is guided by a comprehensive study by Bai et al [27], which demonstrated the superior performance of the TCNs in sequential data in various tests and domains, and by our own previous work, in which TCNs showed consistently strong or superior performance compared with gated recurrent units (GRUs) and attention-based architectures across multiple public ICU datasets [17]. The TCN architecture uses causal convolutions with stride 1, padding to preserve sequence length, and exponentially increasing dilation to expand the receptive field. Each convolution is followed by a Chomp1d operation to enforce causality, Leaky ReLU

XSL•FO
RenderX

activation, dropout, and residual connections. Our models were trained to predict, for each hourly bin, the individual risk of sepsis onset in the next 6 hours. We used a 10-times repeated random split scheme for hyperparameter tuning, resulting in 10 trained models for the hyperparameter set with the lowest average validation loss. For each repetition in this scheme, 20% of the data were held out as a test set. The remaining data were split into training (64%) and validation (16%). For each repetition, 10 randomly drawn hyperparameter configurations were evaluated. The configuration with the lowest average validation loss across the 10 repetitions was selected to train the final models. Training was performed using a binary cross-entropy loss, where class weights were calculated based on the inverse prevalence of the positive class in the training data, an Adam optimizer for a maximum of 1000 epochs, and early stopping with a patience of 10 epochs based on validation loss. The randomized hyperparameter search was conducted on learning rate (sampled log-uniform from exp(–10) to exp(–3)), weight decay (0, 1e–7, 1e–6, …, 1), dropout probability (0.3, 0.4, …, 0.7), batch size (128, 256, and 512), hidden dimension (32, 64, and 128), number of layers (1, …, 10), and kernel size (2, …, 6). A detailed description of the architecture, hyperparameter search, and validation procedure is provided in our previous publication [17]. To investigate a possible dependence of model architecture on the results, we also trained GRU and attention-based models on the MIMIC-IV dataset for comparison. Model training and evaluation followed the same procedures as described in our previous publication [17].

To enhance model performance, binary missingness indicators were added for the 48 time-varying and 4 static features used as inputs. This representation has been shown to capture clinically meaningful information about testing behavior but may also encode site-specific practice patterns [22,28]. To mitigate overfitting, models were trained with regularization (weight decay) and early stopping. The final models, along with their tuned hyperparameters, were then applied to the BerlinICU dataset without any further fine-tuning or retraining.

For establishing baselines in our analyses, we used logistic regression models as implemented in scikit-learn with default settings, corresponding to L2-regularized logistic regression, without hyperparameter tuning. These were trained following a 5-fold cross-validation scheme in which 80% of the data were used for training and 20% for testing. To better capture the temporal dynamics, we additionally engineered temporal features for the 48 time-varying features. These temporal features represent the trajectory of the measurements over time and include the minimum, maximum, mean, median, and variance of the last 4, 8, and 16 hours for the 48 laboratory and vital features, totaling 772 features (48 time-varying features from the hourly bins, their 720 temporal features, and 4 static features). Although this results in a large number of features relative to the number of events, these features are highly correlated temporal summaries, and penalization reduces the effective model complexity. To account for the class imbalance, we used class weights inversely proportional to class frequencies.

Training and evaluation were performed using Python 3.10.8, PyTorch 1.12.1, and scikit-learn 1.1.3.

Model Evaluation

To focus on evaluating the real-world applicability of the models in a clinical setting, we implemented 3 different evaluation strategies that scrutinized different aspects of the models' performance—fixed horizon evaluation, peak score evaluation, and continuous evaluation. Each strategy reflects a different assumption about how a model is applied in practice, and thus, yields performance metrics with different interpretations (Figure 2).

XSL•FO
RenderX

Figure 2. Overview of the evaluation strategies. For the fixed horizon evaluation, a prediction horizon was defined, where predictions were made for a time point that preceded the onset (or control onset), a predefined number of hours (6 hours in this example). Only the data from admission up to this time point was used to make the prediction (green rectangle). Labels were assigned as "positive" (marked as green "1") for patients with sepsis or "negative" (marked as green "0") for control patients. In the peak score evaluation (purple rectangle), the maximum prediction score across all time points before onset was determined. Patients with sepsis were labeled "positive," while control patients were labeled "negative" (purple "1" or "0" label). In the continuous evaluation (brown rectangle), for patients with sepsis, timepoints within a specific time window before the onset (6 hours in this example) were labeled "positive" (brown "1"s), while earlier time points—and all timepoints for control patients—were labeled "negative" (light brown "0"s). All timepoints before the onset (or control onset) were considered for evaluation. For patients with sepsis (top), the onset was the actual sepsis onset. For control patients (bottom), the onset was defined either as a hypothetical control onset matched to the onset time of a patient with sepsis ("matched onset") or as the last available data point, either at discharge or at 7 days after admission, whichever occurred first. The rectangular frames represent the data that was ultimately considered for each evaluation strategy. Shown here are the default settings for each evaluation strategy: the fixed horizon and peak score evaluations use matched sepsis onset times, while the continuous evaluation is depicted without onset matching. Note that the underlying patient data are the same across all panels; only the labeling and selected time points differ according to the respective evaluation strategy.

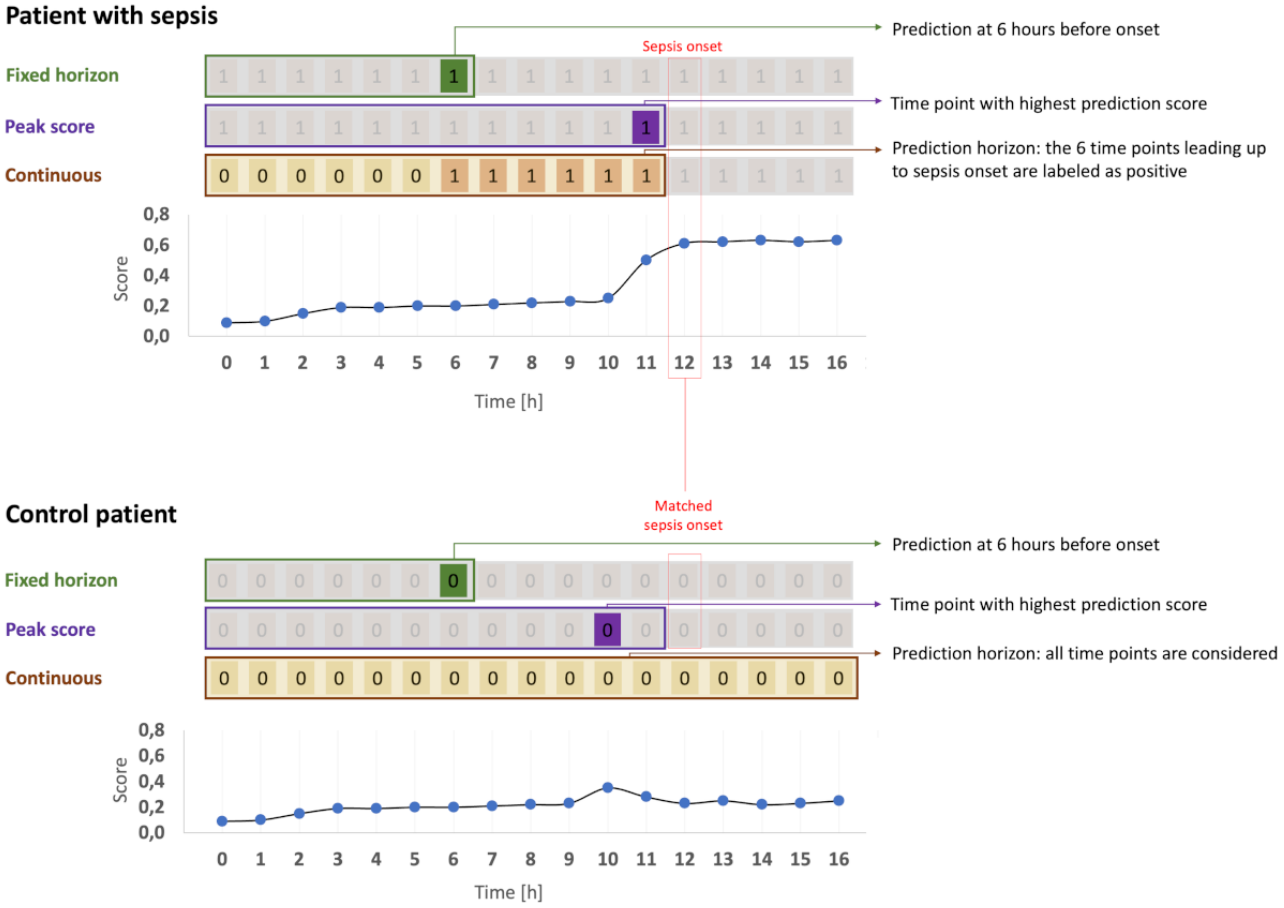

Onset Matching

A key aspect of the evaluation strategies is the handling of time series data for patients without sepsis (controls). While patients with sepsis (cases) have a clearly defined sepsis onset that marks the end point of their time series for analysis, patients without sepsis lack such a reference point. Therefore, we used 2 commonly used approaches to define the end point of their time series for analysis. First, no onset matching, in which the hypothetical onset for control patients is defined as the last available data point, either at discharge or at 7 days after admission, whichever occurred first. Second, onset matching, in which onsets of control patients are aligned with those of patients with sepsis by randomly pairing each control patient with a patient with sepsis whose time to sepsis onset does not exceed that of the control patient. The time from admission to sepsis onset in the matched patient with sepsis is then used as the hypothetical onset time for the control patient [16,29,30]. This approach preserves alignment to admission for both groups and ensures that the distribution of evaluation time points is comparable between patients with sepsis and patients without sepsis, thereby reducing bias due to differences in lengths of stay.

Fixed Horizon Evaluation

This approach evaluates a model's ability to predict the onset of sepsis at a predefined time point before the actual onset—the so-called prediction horizon (h)—using all available data up to that time [13,19]. Varying the length of the prediction horizon allows for assessing the earliness of the model's predictive ability.

The predicted score y_{pred}, which is based on all available data for the patient up to h hours before the sepsis onset for patients with sepsis or up to a hypothetical onset for control patients (t_{onset}), and the true label y_{true} are given by:



XSL•FO
RenderX

where $\boldsymbol{X}_{1:t}$ is the patient data (feature matrix) from time point 1 to t, and f is the model that outputs a score correlated to the probability of a sepsis onset.

Peak Score Evaluation

Peak score evaluation is based on the assumption that a prediction model in a clinical setting operates with a fixed threshold, and once any prediction score surpasses this threshold, an alarm for that particular patient is triggered [16,20]. To capture this behavior, the maximum model prediction score from ICU admission to the onset of sepsis or control offset is determined, resulting in a single score-label pair for each patient for performance score calculation. Here, the prediction score and true label are given by:



For horizon-dependent analysis, we used the peak score within the defined horizon instead of across all time points of each patient.

Continuous Evaluation

This approach evaluates a series of predictions made continuously for each patient over time [13]. Specifically, predictions for all time points before sepsis onset (or control onset) are considered. The resulting performance metrics from this approach represent the model's predictive efficiency across all time points, simulating its use for a randomly selected patient at a random time point.

Using this approach included first establishing a specific time horizon (eg, 6 h) to assess the model's ability to predict the sepsis onset at any given time point within that timeframe. For patients with sepsis, a positive label was assigned to all time points within the horizon, while all other time points were labeled as negative. For control patients, all time points were labeled as negative. Importantly, the time horizon only determines which time points before onset are labeled as positive, but does not restrict the range of time points included in the evaluation. We included the prediction scores from each hourly interval up to the onset of sepsis in patients with sepsis or the hypothetical control onset in control patients (t_{onset}), excluding the prediction at the actual onset (or control onset in case of onset matching). The prediction score and true label are given by:



This procedure generated a sequence of hourly score-label pairs for each patient:



To accommodate variations in the length of these sequences among patients, we applied inverse frequency weighting to the samples when computing performance metrics.

Performance Metrics

For all strategies, we report the AUROC. We chose AUROC because, unlike other commonly applied metrics such as AUPRC, accuracy, precision or PPV, NPV, or F score, the AUROC is independent of sepsis prevalence. This is essential for our study, since sepsis prevalence varies not only between the 2 datasets (MIMIC-IV vs BerlinICU) but also within a dataset across evaluation strategies; for example, in continuous evaluation, the proportion of positive time points changes with each prediction horizon. Using a prevalence-sensitive metric under these conditions would confound discrimination performance with shifts in prevalence and, therefore, blur the effect we wish to study. AUROC instead measures the probability that the model ranks a randomly chosen positive instance higher than a randomly chosen negative one, providing a prevalence-robust estimate of discrimination across all decision thresholds. However, to support interpretability, we additionally report AUPRC, PPV, NPV, and sensitivity at fixed thresholds (Figures S5 and S6 in Multimedia Appendix 1).

Because AUPRC, PPV, and NPV depend on event prevalence, we harmonized the patient-level prevalence across strategies and horizons in all evaluation samples by down-sampling patients without sepsis to match the overall sepsis prevalence of the unfiltered BerlinICU cohort. Note, however, that in continuous evaluation, the prediction horizon also determines the number of time points labeled as positive. So, while patient-level prevalence is fixed across strategies, sample-level prevalence varies with horizon in the continuous evaluation approach. As a result, prevalence-sensitive metrics are difficult to compare directly in continuous evaluation across different horizons. We therefore report both raw values and "lifted" variants that are normalized by the horizon-specific prevalence, namely PPV and AUPRC divided by prevalence and recall measured at the threshold where precision equals twice the prevalence, which disentangle model performance from shifts in baseline risk. When reporting precision at a fixed recall, we estimate it from the precision-recall curve by linear interpolation at the target recall.

To estimate the variability of performance metrics, we applied bootstrap resampling (1000 samples). For each evaluation setting that included a prediction horizon, we only included those patients whose length of stay matched or exceeded the defined horizon (Table S2 in Multimedia Appendix 1).

Calibration Assessment

To assess model calibration, we computed the Brier score. Internal calibration on the MIMIC-IV dataset is reported in our previous publication [17]. For the external BerlinICU dataset, calibration was estimated using 5-fold cross-validation. In each fold, isotonic regression was fitted on 80% of the data and applied to the held-out 20% to produce calibrated predictions. All performance metrics were calculated using the uncalibrated model applied to the full BerlinICU dataset, since discrimination metrics such as AUROC, precision, or recall are unaffected by probability calibration and depend only on the ranking of predicted scores, which is not changed by calibration. Calibration analyses were performed using scikit-learn (version 1.7.0).

Clinical Utility

Although the primary objective of this study was to compare evaluation strategies, we included a clinical utility analysis with

XSL•FO
RenderX

a continuous evaluation setup, which best aligns with clinical reality. Intensive care early-warning systems generate continuously updated risk estimates that may trigger alerts throughout the patient's stay rather than single patient-level decisions. We quantified how the model would behave when deployed as an hourly alerting system by identifying a threshold corresponding to 80% sensitivity for sepsis onset and defining an alarm as any hourly prediction exceeding this threshold. Alarm rates for septic and nonseptic admissions were calculated as the mean number of alarms per patient-hour. To approximate typical alert-suppression logic, we applied a 4-hour silencing window; after an alarm was triggered for a patient, subsequent alarms for that patient were suppressed for 4 hours, and alarm rates were recomputed using only unsilenced alerts. Net benefit was evaluated using decision-curve analysis implemented with the dcurves package, where net benefit represents the proportion of true-positive alerts minus a threshold-dependent penalty for false-positive alerts. For the main clinical interpretation, we focus on the continuous evaluation with a 6-hour prediction horizon, but also report net benefits for the other strategies to illustrate how the choice of evaluation strategy influences apparent clinical usefulness.

Statistical Analysis

To assess the significance of AUROC differences between evaluation strategies, we used a paired bootstrap Wald test with 1000 bootstrap samples per strategy. For each comparison, we drew the same bootstrap sample of patients for both strategies, computed AUROC for each, and formed the per-replicate difference. The SE of the difference was estimated as the SD of the bootstrap differences. Furthermore, 2-sided *P* values were obtained from the standard normal distribution. We applied multiple comparison correction at the level of hypothesis families; for the onset-matching effect within each strategy×model×dataset there is exactly one prespecified contrast (with vs without onset), whereas for strategy comparisons, we tested the 3 related pairwise contrasts (fixed vs peak, peak vs continuous, and fixed vs continuous) within each model×dataset×onset condition. To control the probability of at least 1 false positive among these related tests, we used the Holm step-down procedure.

Results

Study Cohort

From 60,332 ICU stays in the initial dataset, after extraction, preprocessing, and filtering, the final dataset consisted of 40,132 ICU stays of 36,872 unique patients spanning 10 years (2012-2021). Furthermore, 4134 stays were septic, resulting in a sepsis prevalence of 10.3% (Figure 3, Table 1). The original MIMIC-IV training dataset consisted of 67,056 ICU stays with a sepsis prevalence of 5.6% (3730 ICU stays).

Figure 3. Flow diagram of patient eligibility criteria. ICU: intensive care unit.

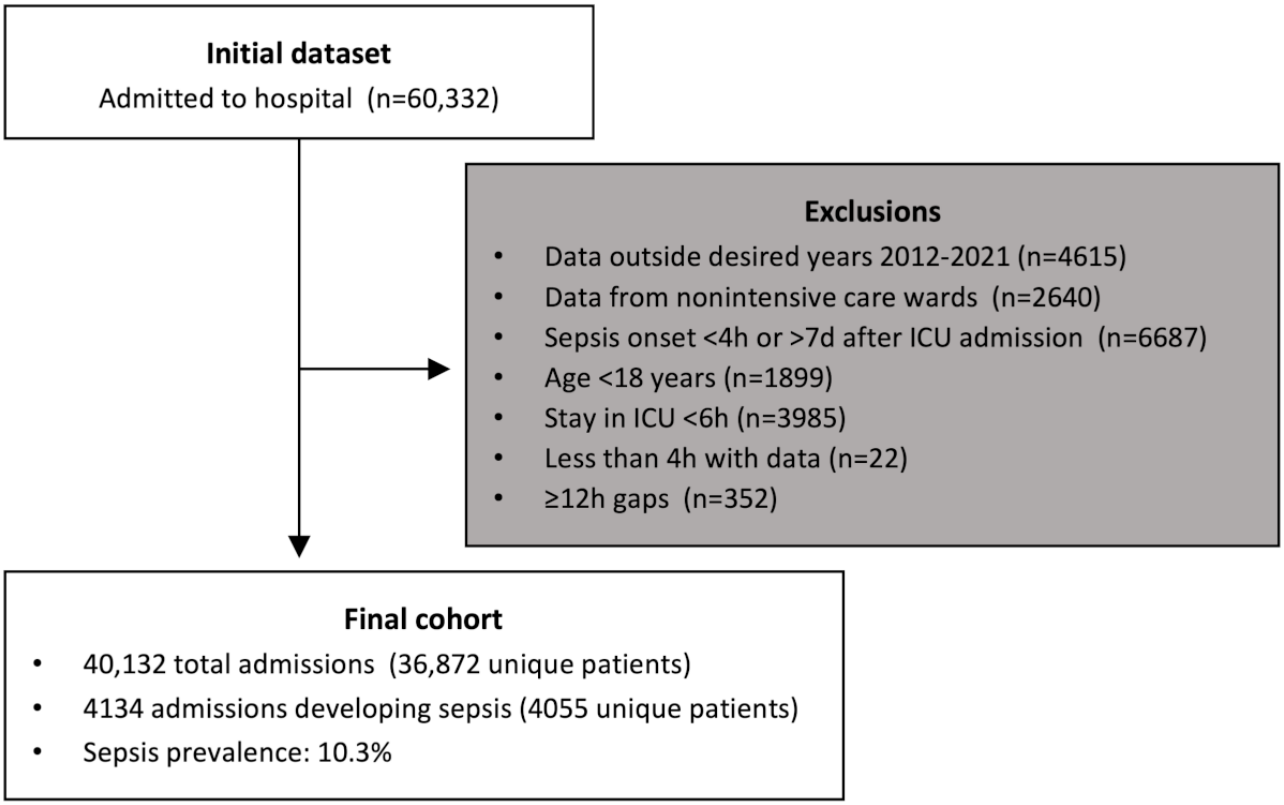

XSL•FO
RenderX

Table 1. Patient characteristics of BerlinICU and MIMIC-IV[a] for the total dataset, the sepsis cohort, and nonsepsis cohort.

Statistic	Total	Sepsis	Control
BerlinICU			
Cohort size n, (row-%)	40,132 (100)	4134 (10)	35,998 (90)
Age (y), median (IQR)	64 (50-75)	67 (55-77)	63 (50-75)
ICU[b] length of stay (d), median (IQR)	1.0 (0.7-2.1)	2.0 (1.4-3.4)	1.0 (0.7-2.0)
Sex, n (%)			
Male	22,423 (56)	2444 (59)	19,979 (56)
Female	17,709 (44)	1690 (41)	16,019 (44)
Patient count, n (%)			
Patients on vasopressors	17,117 (43)	3442 (83)	13,675 (38)
Patients ventilated	13,967 (35)	2774 (67)	11,193 (31)
Patients on any antibiotics	22,155 (55)	4134 (100)	18,021 (50)
Patients on antibiotics for ≥3 days	5351 (13)	4134 (100)	1217 (3)
MIMIC-IV			
Cohort size, n (row-%)	67,056 (100)	3730 (5.6)	63,326 (94.4)
Age (y), median (IQR)	65 (53-76)	67 (54-76)	65 (53-76)
ICU length of stay (d), median (IQR)	1.7 (1.0-2.8)	0.8 (0.6-1.5)	1.7 (1.0-2.8)
Sex, n (%)			
Male	37,184 (55)	2132 (57)	35,052 (55)
Female	29,872 (45)	1598 (43)	28,274 (45)
Patient count, n (%)			
Patients on vasopressors	8548 (13)	1230 (33)	7318 (12)
Patients ventilated	21,863 (33)	2520 (68)	19,343 (31)
Patients on any antibiotics	46,276 (69)	3730 (100)	42,546 (67)
Patients on antibiotics for ≥3 days	6214 (9)	2760 (74)	3454 (5)

MIMIC-IV: Medical Information Mart for Intensive Care IV.

ICU: intensive care unit.

Calibration

In the external BerlinICU dataset, the TCN model's predicted probabilities were reasonably calibrated, with a Brier score of 0.04-0.09 depending on the evaluation strategy (Figure S4 in Multimedia Appendix 1). Importantly, calibration does not affect the performance metrics reported in the main results, as these are independent of the predicted probability values themselves and are unaffected by monotonic transformations.

Influence of Evaluation Strategy on Model Performance Estimates

As our primary analysis, we applied the TCN models pretrained on MIMIC-IV to the MIMIC-IV test set for internal validation under a scenario aligned with the training setup—continuous evaluation using a 6-hour prediction horizon without onset matching. Under this configuration, the models showed strong internal validation performance (Figure 4A). When the same models were applied to BerlinICU for external validation under the same evaluation conditions, performance was notably reduced (Figure 4B).

XSL•FO
RenderX

Figure 4. Comparison of model performance of (A) temporal convolutional network models (top) and (B) logistic regression baseline models (bottom) for different evaluation strategies. The models were applied to the test splits of the (a) MIMIC-IV training dataset and (b) the entire German ICU dataset (BerlinICU). Model performance was evaluated using fixed horizon evaluation (6h horizon), peak score evaluation (entire patient stay), and continuous evaluation (6h horizon) without onset matching (dark bars) and with onset matching (light bars). The AUROC value inside the bar represents the mean AUROC across all bootstrap samples and models. Error bars represent 95% CIs. Significance: ***P<.001, **P<.01, *P<.05, ns: not significant (paired bootstrap Wald test with Holm-adjusted P values). AUROC: area under the receiver operating characteristic curve; MIMIC: Medical Information Mart for Intensive Care.

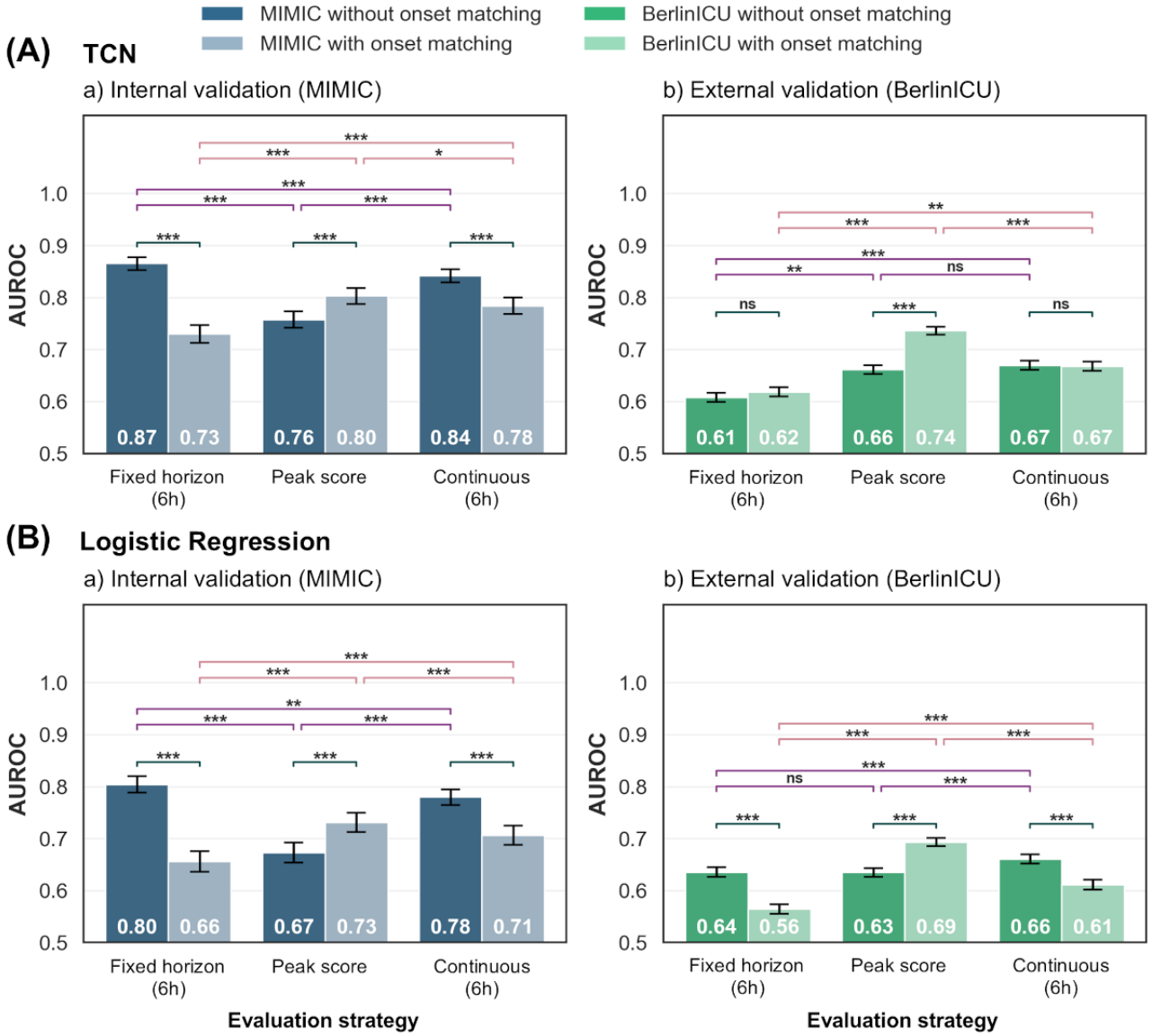

To assess the impact of different evaluation strategies, we performed fixed horizon evaluation under the same conditions (6-hour prediction horizon, no onset matching) and peak score evaluation, which does not use a prediction horizon. On the MIMIC-IV test set, fixed horizon evaluation resulted in a slightly higher AUROC, whereas peak score evaluation showed reduced performance. In external validation on BerlinICU, fixed horizon evaluation was markedly lower than continuous evaluation, while the peak score evaluation estimate was similar to that of the continuous evaluation. Across all evaluation strategies, performance estimates were consistently lower on BerlinICU compared with the MIMIC-IV test set.

Aligning the length of stay distribution of control patients to patients with sepsis through onset matching led to a notable AUROC reduction in internal validation on the MIMIC-IV test set for continuous evaluation (ΔAUROC=–0.06) and fixed horizon evaluation (ΔAUROC=–0.14), while AUROC estimates for peak score evaluation increased (ΔAUROC=0.04). On BerlinICU, onset matching had minimal impact on performance estimates during continuous evaluation or fixed horizon evaluation. In contrast, for peak score evaluation, onset matching increased AUROC estimates (ΔAUROC=0.08).

Across all evaluation strategies, TCN models consistently outperformed logistic regression baseline models. The only exception was observed in fixed horizon evaluation on BerlinICU without onset matching, where logistic regression showed a superior performance compared with TCN.

Influence of the Prediction Horizon on Model Performance Estimates

To evaluate the ability of the models to provide timely predictions, we assessed their performance across a range of

XSL•FO
RenderX

prediction horizons (Figure 5). Performance estimates generally improved with shorter prediction horizons across all evaluation strategies. During internal validation on the MIMIC-IV test set, fixed horizon evaluation and continuous evaluation showed consistent improvement, with the highest AUROC achieved at a 1-hour horizon (fixed horizon: AUROC=0.92, 95% CI 0.91-0.93; continuous: AUROC=0.87, 95% CI 0.85-0.88). Peak score plateaued around the 6-hour horizon (AUROC 0.92, 95% CI 0.91-0.93), with only marginal improvements for smaller horizons. In external validation on BerlinICU, performance also generally increased with shorter prediction windows, with few exceptions, including a plateau around the 6-hour horizon in fixed horizon and peak score evaluation with onset matching.

Figure 5. Performance metrics across prediction horizons, comparing the performance of temporal convolutional network models for an evaluation with onset matching (dashed lines) and without onset matching (solid lines). (A) Fixed horizon, (B) peak score, and (C) continuous prediction. The models were applied to the test splits of the MIMIC-IV training dataset (blue) and the entire German ICU dataset (BerlinICU; green). Shown is the AUROC based on varying prediction horizons from one hour before onset up to 100 hours before onset. Error bars represent 95% CIs. AUROC: area under the receiver operating characteristic; MIMIC: Medical Information Mart for Intensive Care; TCN: temporal convolutional network.

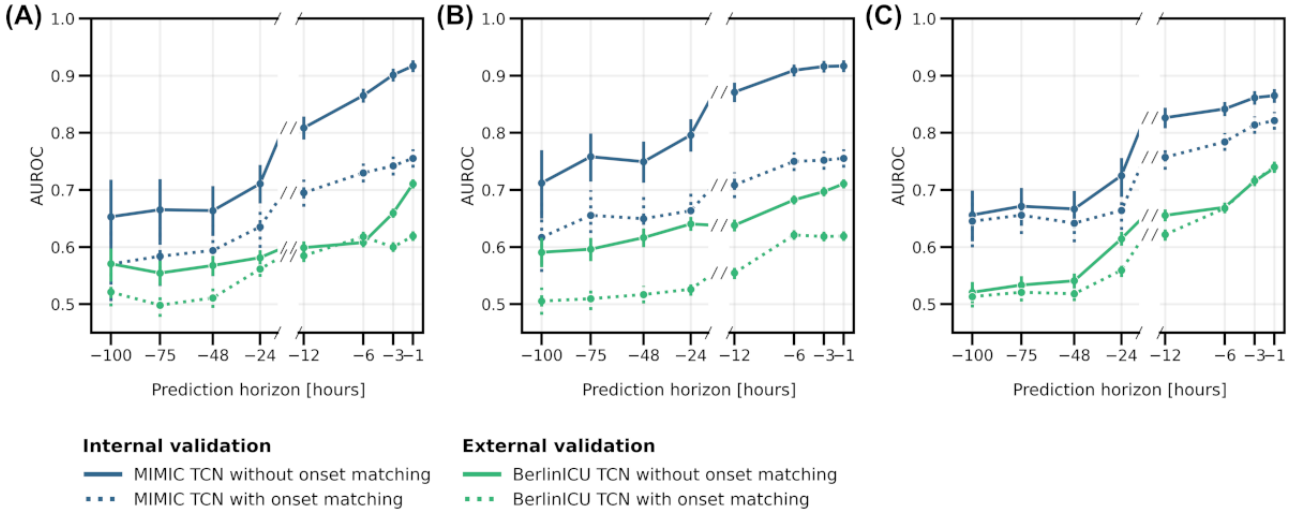

Onset matching consistently decreased performance estimates during internal validation on the MIMIC-IV test set across all evaluation strategies, albeit to different extents, with continuous evaluation showing the smallest susceptibility to the matching approach. On BerlinICU, onset matching impacted strategies differently depending on the prediction horizon. Notably, continuous evaluation remained more robust to onset matching, maintaining more consistent AUROCs compared with the other strategies across different prediction horizons, particularly for very short and long prediction horizons.

Comparing the TCN and logistic regression baseline models in internal validation on the MIMIC-IV test set, TCN models consistently outperformed logistic regression models across all evaluation strategies and prediction horizons. In contrast, during external validation on BerlinICU without onset matching, logistic regression models approached or even exceeded the performance of TCN models at certain prediction horizons (Figure S1 in Multimedia Appendix 1). Onset matching was associated with an AUROC reduction for both models; the reduction appeared larger for logistic regression than for TCN in our data.

Clinical Operational Characteristics Under Continuous Evaluation

Under the continuous evaluation with a 6-hour prediction horizon, an operating point of 80% sensitivity produced 0.70 alarms/hour in patients with sepsis and 0.25 alarms/hour in patients without sepsis in MIMIC-IV, compared with 0.74 and 0.55 alarms/hour, respectively, in BerlinICU (Figure S8 in Multimedia Appendix 1). With a 4-hour silencing period, rates decreased to 0.18 (septic) and 0.06 (nonseptic) in MIMIC-IV, and to 0.18 and 0.14 alarms/hour, respectively, in BerlinICU. Decision-curve analysis showed modest net benefit at low thresholds in MIMIC-IV and limited benefit in BerlinICU (Figure S9 in Multimedia Appendix 1).

Discussion

Principal Findings

In this study, we assessed the impact of commonly used evaluation strategies (fixed horizon, peak score, and continuous evaluation) on the performance of machine learning models for sepsis onset prediction. Our results demonstrate that the choice of the evaluation strategy significantly influences performance estimates, even though the same model is used on the same dataset.

We further addressed a gap in research involving the German health care system by aggregating an ICU dataset, called BerlinICU, from one of the largest university hospitals in Europe, Charité - Universitätsmedizin Berlin. Including more than 40,000 admissions from 8 ICU wards, this dataset is one of the largest ICU datasets in Europe, exceeding the size of the currently available European datasets—Amsterdam University Medical Center Database [31], High Time Resolution ICU Dataset [32], and Salzburg Intensive Care database [33]. To the best of our knowledge, this is one of the first studies to explore how sepsis prediction models perform on large-scale German ICU data.

XSL•FO
RenderX

Fixed Horizon Evaluation Requires Prospective Knowledge of Sepsis Onset

Fixed horizon evaluation uses data up to a predefined time point before sepsis onset—the prediction horizon—to make predictions. Performance decreased with longer prediction horizons due to increasing temporal distance from sepsis onset. Without onset matching, performance estimates strongly increased for all horizons on the MIMIC-IV test set and for most horizons for BerlinICU, with the exception of the 6-hour horizon. This performance increase when not using onset matching is most parsimoniously explained by the evaluation of control patients over their entire stay, with emphasis on the period just before discharge. During these predischarge hours, control patients are typically in a stable condition, making it easier for the model to correctly identify them as low-risk for sepsis. This effect, also noted by Futoma et al [29], shows the significant impact of onset alignment. Despite this significant influence, many studies do not comment on the implementation of case-control matching [13]. While the fixed horizon approach achieved the highest AUROC in a study on different evaluation strategies [19], we observed more nuanced performance differences compared with the other evaluation approaches, depending on the dataset, horizon, and onset matching approach. Fixed-horizon evaluation can be useful for benchmarking how far in advance a model can predict sepsis and for retrospective simulation studies, but its direct clinical interpretability is challenging, as it relies on prospective information—the time point of the sepsis onset—which is not accessible in a real-world clinical setting.

Peak Score Evaluation is Influenced by the Length of Stay Distribution and Requires Onset Matching to Avoid Bias

The peak score evaluation assesses the model's peak confidence in predicting sepsis throughout the entire ICU stay. This approach seems intuitively aligned with clinical practice, where a fixed threshold would be predetermined, and an alarm is triggered once the prediction model's output surpasses it—potentially just once per patient, with subsequent alarms suppressed [16,20]. The focus is on whether an alarm is raised at any point during the patient's stay, which is replicated by determining the maximum prediction score. However, the performance metrics determined under this approach are influenced by differences in length of stay distributions between cases and controls—longer time series for control patients increase the likelihood of higher maximum prediction scores for these patients simply due to random fluctuations, ultimately reducing the AUROC as the maximum scores for control patients approach those of patients with sepsis on average (Figure S3 in Multimedia Appendix 1). Importantly, this bias conflicts with the core interpretation of the AUROC, which measures the probability that a randomly selected true positive will have a higher predicted score than a randomly selected true negative. Our results demonstrate this behavior, showing that performance estimates without onset matching, where control cases tend to have longer lengths of stays, were significantly reduced compared with analyses with onset matching (Figure 4). To mitigate this effect, an onset matching approach that aligns the length of stay distributions between positive and negative cases is strictly necessary when applying this evaluation strategy.

The discrepancy between the peak score evaluation across all patients (Figure 4), where onset matching increases the AUROC, and horizon-dependent analysis (Figure 5), where onset matching decreases the AUROC, can be explained by the exclusion of patients with insufficient stay lengths in the latter. This exclusion alters the length of stay distribution and thereby significantly impacts performance metrics. Not filtering patients with insufficient length of stays shows a distinct interaction effect between onset matching and prediction horizon (Figure S2 in Multimedia Appendix 1), further highlighting the critical influence of length of stay distribution on the performance estimates in the peak score evaluation approach. Nevertheless, peak score evaluation has a simple clinical interpretation—at a given threshold, does the model ever raise an alarm for a patient with sepsis? When the length of stay distribution is carefully adjusted between cases and controls and the evaluation is restricted to a clinically meaningful time frame (eg, within 12 hours before sepsis onset), this strategy can provide a useful summary of whether the model is able to flag a patient at high-risk at any point during their stay.

Continuous Evaluation Without Onset Matching Reflects Clinical Practice

Continuous prediction evaluation showed increased performance with shorter prediction windows, demonstrating the model's effectiveness in using short-term, relevant data to predict sepsis onset. As the prediction window narrows and focuses more on the actual onset of sepsis, the assigned sepsis labels align more closely with the sepsis-related signals in the patient data, enhancing prediction efficiency. Interestingly, although the models were trained to predict sepsis 6 hours in advance, this specific labeling window did not yield the best performance. This is likely because these shorter windows include only time points immediately before sepsis onset, where the signal is strongest, while earlier time points, although possibly already showing signs of deterioration, are treated as negatives.

Evaluating on BerlinICU with onset matching, model performance was significantly reduced compared with the MIMIC-IV test set (ΔAUROC=–0.11 for 6 hours before onset and –0.08 for 1 hour). This reduced performance when transferring models between datasets was also evident in our previous study, where models trained on MIMIC-IV were evaluated on other European datasets using a horizon-based evaluation with a 6-hour window [17]. Even though the performance in that study declined for European datasets (High time Resolution ICU Dataset: ΔAUROC=–0.07, Amsterdam University Medical Center Database: ΔAUROC=–0.09) as well as for another US dataset (Electronic Intensive Care Unit Collaborative Research Database: ΔAUROC=–0.10), it was not as severe as observed in BerlinICU. One might suspect that this sharper decline is related to different sepsis prevalences, since BerlinICU had a higher prevalence than MIMIC-IV. However, AUROC is unaffected by prevalence [21], and the difficulty of the task, and therefore the prediction performance, is more strongly influenced by factors such as data heterogeneity, label noise, and shifting feature distributions, all of which may differ

XSL•FO
RenderX

substantially between datasets. These findings underscore the importance of addressing domain shift when deploying machine learning models across health care systems. Differences in patient populations, documentation practices, feature availability, and underlying sepsis definitions may contribute to reduced generalizability. For example, differences in clinical workflows, such as the timing and frequency of laboratory testing, documentation delays, or variability in when and how sepsis-related interventions are initiated (eg, antibiotics or fluids), may influence how early or clearly sepsis-related signals appear in the data. The use of missingness indicators is another potential factor, although previous studies have reported mixed effects on the generalizability of sepsis prediction models [22,28]. Mitigation strategies, such as dataset-specific fine-tuning or full retraining on local data, can be considered to improve cross-domain performance and support safe clinical deployment.

Without onset matching, performance in the MIMIC-IV test set was higher across horizons, and was similar or better on BerlinICU for most horizons. Again, this higher performance without onset matching most likely reflects the inclusion of time points closer to discharge for control patients, who generally have a healthier profile, making the prediction task easier. This mechanism is illustrated in Figure S10 in Multimedia Appendix 1, which shows predicted risk trajectories over time; unmatched controls display a marked decline in predicted risk approaching discharge, whereas matched controls remain at higher levels, closer to those of patients with sepsis. While onset matching may provide a focused assessment in more challenging cases, not using it better mirrors clinical practice, where patients are monitored continuously from admission to discharge. In our view, continuous evaluation without onset matching better reflects real-world clinical application, as the AUROC represents the likelihood of a positive case being ranked higher than a negative one, including time points closer to discharge. Accordingly, we treat continuous evaluation without onset matching as the most clinically relevant interpretation of model performance, while the other evaluation strategies are used to demonstrate how alternative designs can substantially alter reported results.

Although this approach requires predetermining a time window, which indicates how far in the future the onset may be predicted, we consider continuous evaluation the better choice for 2 reasons. First, defining a time window—ideally based on pathophysiological plausibility—ensures that only alarms occurring within a clinically meaningful period before sepsis onset are considered. This distinction is crucial, as some published models show high performance days to weeks before the actual event [18,34,35], likely detecting broader, correlated health-related features rather than sepsis specifically. This underlines the importance of thoroughly defining and testing what these models are predicting to ensure their practical utility in clinical settings. Second, varying the prediction window and reporting the performance across all these windows allows for a more nuanced understanding of the models' ability to predict event onset in advance. This analysis can be performed efficiently without retraining, since the underlying model predictions remain fixed and only the labeling of time points is varied to reflect different prediction windows. However, training separate models for different forecast horizons and investigating the effect on the performance using continuous evaluation is an interesting, but computationally more intensive approach that could be explored in future studies.

It is important to note that longer prediction horizons (eg, beyond 12 h) are unlikely to be clinically meaningful. From a clinical standpoint, alarms triggered more than half a day before sepsis onset may not prompt specific interventions and could increase alert fatigue. Nevertheless, we included prediction horizons up to 100 hours as a form of negative control to investigate whether model performance persists far in advance of sepsis. If so, this could indicate that the model is capturing nonspecific markers of deterioration or health status rather than sepsis-specific signals. In line with this, we observed that the model's performance on BerlinICU dropped toward chance level at long horizons, while AUROC remained elevated on MIMIC-IV, possibly suggesting that the model may overfit to nonspecific or site-specific patterns in the training data.

Clinical Implications of the Choice of the Evaluation Strategy

The evaluation strategy directly determines how performance metrics translate to clinical practice. Metrics such as PPV, when derived from peak-score evaluation, are patient-level; they yield a single value per patient that answers, "Did this patient ever cross the alert threshold?" In routine deployment, however, models are usually executed repeatedly (eg, hourly), and clinicians need alert-level information, that is, the probability that a specific alarm corresponds to true sepsis within a clinically relevant time window. Using patient-level metrics to select models for clinical deployment can therefore overestimate real-world performance and contribute to alarm fatigue. Likewise, fixed-horizon evaluation reports performance at one predefined time point per patient; it does not describe how the model behaves during continuous monitoring. Importantly, it requires prospective knowledge of the exact onset time, which is inherently unavailable in clinical reality. For continuous evaluation, the horizon is adjustable; PPV at h hours answers the question "If the model raises an alert now, what is the probability that sepsis will occur within the next h hours?" Different choices of the horizon yield different PPVs, and only the PPV at the clinically intended horizon is directly relevant for deployment. Thus, to ensure clinical relevance, evaluation strategies and metrics should be aligned with how the model will be used in practice.

Our analyses focused primarily on TCN models. However, the effects of the evaluation strategy are not specific to this architecture. Because evaluation is applied post hoc to model outputs, the principles are, in theory, model-agnostic. Consistent with this expectation, we observed the same qualitative influence of evaluation strategy on performance when applying it to GRU- and transformer-based models (Figure S7 in Multimedia Appendix 1). This supports the interpretation that our findings reflect general properties of evaluation design rather than model-specific behavior. Likewise, while our empirical analyses were conducted on 2 datasets, the observed differences between evaluation strategies arise from the way labels are constructed

XSL•FO
RenderX

and metrics are defined, and are therefore expected to hold independently of the specific dataset.

Clinical Utility Considerations

While the primary objective of this study is to compare evaluation strategies, it is helpful to illustrate how a model of this type would function within a real-time ICU workflow. In practice, such early-warning systems would generate repeated risk estimates and trigger alerts that prompt targeted clinical assessment, such as reviewing chart data or performing additional tests, rather than automated interventions. Thresholds are typically chosen to balance timely detection with acceptable alarm burden, often combined with silencing periods to avoid repeated notifications. Since clinical use is structured around continuous monitoring and repeated predictions, continuous evaluation is the strategy that most closely reflects real-world deployment.

The operational characteristics observed in our analyses provide an indication of the potential alarm burden such a system might generate. At 80% sensitivity, the model produced alarm rates that are plausible for an ICU early-warning context, although their clinical acceptability is uncertain and would require prospective evaluation. A 4-hour silencing period substantially reduced repeated alerts, highlighting the importance of workflow-oriented design choices. Decision-curve analysis further showed modest net benefit at low thresholds in MIMIC-IV and limited benefit in BerlinICU.

Limitations

Our study has several limitations. First, we did not have access to microbiological data, requiring an adjustment of the Sepsis-3 definition. While this approach has been used in previous studies using hospital datasets without microbiology records [16,17], it may reduce the specificity of outcome labeling and limit comparability with studies that apply the complete Sepsis-3 criteria, including microbiology data. Second, we used models trained for a 6-hour horizon and applied them across various prediction horizons to predict sepsis onset. While these models may not perform optimally outside their original specifications, the inability to know the exact duration until a potential sepsis onset at any given timepoint in clinical practice necessitates using a flexible approach. Thus, by testing these models beyond their intended parameters, we can evaluate their robustness and generalizability to different clinical scenarios and timelines. We acknowledge, however, that this does not replace training and comparing separate models optimized for different prediction horizons.

Comparison With Previous Work

Our work combines the idea of generalizability across countries [16,17] and suitable metric calculation concepts [16,36]. Our study builds on these previous concepts, refining and extending them as needed. While Moor et al [16] and Rockenschaub et al [17] examined the influence of various algorithms and training sets from different countries, our work focuses on different evaluation strategies. We applied sepsis onset matching [29,36] and peak score evaluation [16], combining approaches that were presented separately. As the peak score evaluation is only suitable for retrospective analysis, we also applied a continuous evaluation strategy [13]. Moreover, there are only a few studies that have applied models to German data [18,37], none of which have a dedicated focus on comparing performances in different evaluation strategies. While some models have reported high AUROC values on German data, care is needed when interpreting or comparing such metrics across studies, as the AUROC is not a function of the model alone, but depends heavily on the prediction task and the underlying dataset. Differences in prediction target, patient population, and labeling criteria can have substantial effects on reported discrimination. For example, distinguishing patients with sepsis in the ICU from a general hospital population is inherently easier than distinguishing sepsis from nonsepsis within an already critically ill ICU cohort, due to the greater physiological overlap in the latter case. Moreover, as we demonstrate in this study, the choice of evaluation strategy (eg, fixed prediction horizon, onset matching, or continuous evaluation) can meaningfully influence reported performance.

Several studies have applied sepsis prediction models prospectively to US data [14,15,38], but they often lack details on how the models were trained and how the data were preprocessed. This lack of transparency is a common issue, as noted in 2 systematic reviews of sepsis prediction models [12,13]. Accordingly, common checklists for reporting prediction models include items for external validation in different countries and time periods [39,40].

Conclusions

Our study highlights the critical need to choose an evaluation approach that is aligned with the intended interpretation of the resulting performance metric, as different approaches yield markedly different performance estimates despite using the same model on the same dataset. Importantly, for the fixed horizon and peak score evaluation approaches, a carefully devised onset matching strategy is crucial to avoid skewed results that may not reflect the true model performance. In our view, the continuous evaluation approach better reflects clinical reality. Despite requiring specifying the horizon time window, it offers insights into the model's effectiveness in a clinical setting.

Acknowledgments

We thank Fabian Schreiber and Falk Meyer-Eschenbach for their support in data extraction. This research was conducted using data obtained from Charité's Health Data Platform (HDP). The authors also acknowledge the Scientific Computing of the IT Division at the Charité - Universitätsmedizin Berlin for providing computational resources that have contributed to the research

XSL•FO
RenderX

results reported in this paper. OpenAI's Model GPT5 was used to polish the language of the manuscript, ensuring clarity and fluency, as well as to shorten paragraphs where applicable, assisting in effective communication of the study's contributions.

All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

Data Availability

The MIMIC-IV (Medical Information Mart for Intensive Care IV) dataset is publicly available for research purposes through PhysioNet upon completion of a data use agreement and required training [41,42]. The BerlinICU dataset used in this study cannot currently be shared outside the institution due to data protection regulations and institutional policies.
Preprocessing and training code for the neural network models has been published by Rockenschaub et al [17], with the corresponding GitHub repository available at GitHub [43]. All analysis code specific to this study are also available [44].

Funding

This work received funding from the Jürgen Manchot Stiftung.
All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

Authors' Contributions

DKD: Methodology, software, validation, formal analysis, investigation, data curation, writing (original draft), writing (review & editing), visualization, project administration.
PR: Methodology, software, validation, data curation, writing (review & editing).
SDB: Data curation, writing (review & editing).
OK: Writing (review & editing).
HDV: Writing (review & editing), supervision.
FB: Writing (review & editing), supervision, resources.
FvD: Conceptualization, methodology, writing (review & editing), supervision.
GL: Conceptualization, methodology, investigation, writing (original draft), writing (review & editing), visualization, supervision, project administration.

Conflicts of Interest

SDB is a salaried employee at Pfizer Pharma GmbH and a visiting researcher at the institute of Medical Informatics, Charité — Universitätsmedizin Berlin. Pfizer had no involvement in the conception, design, execution, or interpretation of the study, nor in the preparation or decision to submit the manuscript for publication. The authors declare no conflicts of interest regarding this study.

Multimedia Appendix 1
Additional tables and figures.
[PDF File (Adobe PDF File), 1839 KB - jmir_v28i1e72083_app1.pdf]

References

1. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 2016;315(8):801-810 [FREE Full text] [doi: 10.1001/jama.2016.0287] [Medline: 26903338]
2. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the global burden of disease study. Lancet 2020;395(10219):200-211 [FREE Full text] [doi: 10.1016/S0140-6736(19)32989-7] [Medline: 31954465]
3. Fleischmann C, Thomas-Rueddel DO, Hartmann M, Hartog CS, Welte T, Heublein S, et al. Hospital incidence and mortality rates of sepsis. Dtsch Arztebl Int 2016;113(10):159-166 [FREE Full text] [doi: 10.3238/arztebl.2016.0159] [Medline: 27010950]
4. Torio CM, Moore BJ. National inpatient hospital Costs: The most expensive conditions by payer, 2013. Healthc Cost Util Proj HCUP Stat Briefs Rockville (MD): Agency for Healthcare Research and Quality (US) 2006. [Medline: 27359025]
5. Winters BD, Eberlein M, Leung J, Needham DM, Pronovost PJ, Sevransky JE. Long-term mortality and quality of life in sepsis: a systematic review. Crit Care Med 2010;38(5):1276-1283. [doi: 10.1097/CCM.0b013e3181d8cc1d] [Medline: 20308885]

XSL•FO
RenderX

6. Pruinelli L, Westra BL, Yadav P, Hoff A, Steinbach M, Kumar V, et al. Delay within the 3-hour surviving sepsis campaign guideline on mortality for patients with severe sepsis and septic shock. Crit Care Med 2018;46(4):500-505 [FREE Full text] [doi: 10.1097/CCM.0000000000002949] [Medline: 29298189]
7. Ferrer R, Martin-Loeches I, Phillips G, Osborn TM, Townsend S, Dellinger RP, et al. Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. Crit Care Med 2014;42(8):1749-1755. [doi: 10.1097/CCM.0000000000000330] [Medline: 24717459]
8. Seymour CW, Gesten F, Prescott HC, Friedrich ME, Iwashyna TJ, Phillips GS, et al. Time to treatment and mortality during mandated emergency care for sepsis. N Engl J Med 2017;376(23):2235-2244 [FREE Full text] [doi: 10.1056/NEJMoa1703058] [Medline: 28528569]
9. Evans L, Rhodes A, Alhazzani W, Antonelli M, Coopersmith CM, French C, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. Intensive Care Med 2021;47(11):1181-1247 [FREE Full text] [doi: 10.1007/s00134-021-06506-y] [Medline: 34599691]
10. Evans L, Rhodes A, Alhazzani W, Antonelli M, Coopersmith CM, French C, et al. Executive summary: Surviving sepsis campaign: International guidelines for the management of sepsis and septic shock 2021. Crit Care Med 2021;49(11):1974-1982. [doi: 10.1097/CCM.0000000000005357] [Medline: 34643578]
11. Levy MM, Evans LE, Rhodes A. The surviving sepsis campaign Bundle: 2018 update. Intensive Care Med 2018;44(6):925-928. [doi: 10.1007/s00134-018-5085-0]
12. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 2020;46(3):383-400 [FREE Full text] [doi: 10.1007/s00134-019-05872-y] [Medline: 31965266]
13. Moor M, Rieck B, Horn M, Jutzeler CR, Borgwardt K. Early prediction of sepsis in the ICU using machine learning: A systematic review. Front Med (Lausanne) 2021;8:607952 [FREE Full text] [doi: 10.3389/fmed.2021.607952] [Medline: 34124082]
14. Burdick H, Pino E, Gabel-Comeau D, McCoy A, Gu C, Roberts J, et al. Effect of a sepsis prediction algorithm on patient mortality, length of stay and readmission: a prospective multicentre clinical outcomes evaluation of real-world patient data from US hospitals. BMJ Health Care Inform 2020;27(1):e100109 [FREE Full text] [doi: 10.1136/bmjhci-2019-100109] [Medline: 32354696]
15. Adams R, Henry KE, Sridharan A, Soleimani H, Zhan A, Rawat N, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med 2022;28(7):1455-1460. [doi: 10.1038/s41591-022-01894-0] [Medline: 35864252]
16. Moor M, Bennett N, Plečko D, Horn M, Rieck B, Meinshausen N, et al. Predicting sepsis using deep learning across international sites: a retrospective development and validation study. EClinicalMedicine 2023;62:102124 [FREE Full text] [doi: 10.1016/j.eclinm.2023.102124] [Medline: 37588623]
17. Rockenschaub P, Hilbert A, Kossen T, Elbers P, von Dincklage F, Madai VI, et al. The impact of multi-institution datasets on the generalizability of machine learning prediction models in the ICU. Crit Care Med 2024;52(11):1710-1721. [doi: 10.1097/CCM.0000000000006359] [Medline: 38958568]
18. Steinbach D, Ahrens PC, Schmidt M, Federbusch M, Heuft L, Lübbert C, et al. Applying machine learning to blood count data predicts sepsis with ICU admission. Clin Chem 2024;70(3):506-515. [doi: 10.1093/clinchem/hvae001] [Medline: 38431275]
19. Lauritsen SM, Thiesson B, Jørgensen MJ, Riis AH, Espelund US, Weile JB, et al. The framing of machine learning risk prediction models illustrated by evaluation of sepsis in general wards. NPJ Digit Med 2021;4(1):158 [FREE Full text] [doi: 10.1038/s41746-021-00529-x] [Medline: 34782696]
20. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med 2021;181(8):1065-1070 [FREE Full text] [doi: 10.1001/jamainternmed.2021.2626] [Medline: 34152373]
21. McDermott MBA, Zhang H, Hansen LH, Angelotti G, Gallifant J. A closer look at AUROC and AUPRC under class imbalance. arXiv 2025. [doi: 10.52202/079017-1400]
22. Reyna MA, Josef CS, Jeter R, Shashikumar SP, Westover MB, Nemati S, et al. Early prediction of sepsis from clinical Data: The PhysioNet/computing in cardiology challenge 2019. Crit Care Med 2020;48(2):210-217 [FREE Full text] [doi: 10.1097/CCM.0000000000004145] [Medline: 31939789]
23. Huxford C, Rafiei A, Nguyen V, Wiens M, Ansermino J, Kissoon N, Pediatric Sepsis Data CoLaboratory. The 2024 pediatric sepsis Challenge: Predicting in-hospital mortality in children with suspected sepsis in Uganda. Pediatr Crit Care Med 2024;25(11):1047-1050. [doi: 10.1097/PCC.0000000000003556] [Medline: 38904442]
24. Bennett N, Plečko D, Ukor IF, Meinshausen N, Bühlmann P. ricu: R's interface to intensive care data. Gigascience 2022;12:giad041 [FREE Full text] [doi: 10.1093/gigascience/giad041] [Medline: 37318234]
25. Rockenschaub P, Akay EM, Carlisle BG, Hilbert A, Wendland J, Meyer-Eschenbach F, et al. External validation of AI-based scoring systems in the ICU: a systematic review and meta-analysis. BMC Med Inform Decis Mak 2025;25(1):5 [FREE Full text] [doi: 10.1186/s12911-024-02830-7] [Medline: 39762808]

XSL•FO
RenderX

26. Moor M, Bennett N, Plečko D, Horn M, Rieck B, Meinshausen N, et al. Predicting sepsis using deep learning across international sites: A retrospective development and validation study. EClinicalMedicine 2023;62:102124 [FREE Full text] [doi: 10.1016/j.eclinm.2023.102124] [Medline: 37588623]
27. Bai S, Kolter JZ, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv 2018. [doi: 10.48550/arXiv.1803.01271]
28. Singh J, Sato M, Ohkuma T. On missingness features in machine learning models for critical Care: Observational study. JMIR Med Inform 2021;9(12):e25022 [FREE Full text] [doi: 10.2196/25022] [Medline: 34889756]
29. Futoma J, Hariharan S, Heller K, Sendak M, Brajer N, Clement M, et al. An improved multi-output gaussian process RNN with real-time validation for early sepsis detection. Proc 2nd Mach Learn Healthc Conf PMLR 2017:243-254 [FREE Full text]
30. Lichtner G, Balzer F, Haufe S, Giesa N, Schiefenhövel F, Schmieding M, et al. Predicting lethal courses in critically ill COVID-19 patients using a machine learning model trained on patients with non-COVID-19 viral pneumonia. Sci Rep 2021;11(1):13205 [FREE Full text] [doi: 10.1038/s41598-021-92475-7] [Medline: 34168198]
31. Thoral PJ, Peppink JM, Driessen RH, Sijbrands EJG, Kompanje EJO, Kaplan L, Amsterdam University Medical Centers Database (AmsterdamUMCdb) Collaboratorsthe SCCM/ESICM Joint Data Science Task Force. Sharing ICU patient data responsibly under the society of critical care medicine/european society of intensive care medicine joint data science collaboration: The Amsterdam university medical centers database (AmsterdamUMCdb) example. Crit Care Med 2021;49(6):e563-e577 [FREE Full text] [doi: 10.1097/CCM.0000000000004916] [Medline: 33625129]
32. Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med 2020;26(3):364-373. [doi: 10.1038/s41591-020-0789-4] [Medline: 32152583]
33. Rodemund N, Wernly B, Jung C, Cozowicz C, Koköfer A. Harnessing big data in critical Care: Exploring a new European dataset. Sci Data 2024;11(1):320 [FREE Full text] [doi: 10.1038/s41597-024-03164-9] [Medline: 38548745]
34. Barton C, Chettipally U, Zhou Y, Jiang Z, Lynn-Palevsky A, Le S, et al. Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. Comput Biol Med 2019;109:79-84 [FREE Full text] [doi: 10.1016/j.compbiomed.2019.04.027] [Medline: 31035074]
35. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med 2015;7(299):299ra122. [doi: 10.1126/scitranslmed.aab3719] [Medline: 26246167]
36. Moor M, Horn M, Rieck B, Roqueiro D, Borgwardt K. Early recognition of sepsis with Gaussian process temporal convolutional networks and dynamic time warping. Mach Learn Healthc Conf PMLR 2019:2-26 [FREE Full text]
37. Schamoni S, Lindner HA, Schneider-Lindner V, Thiel M, Riezler S. Leveraging implicit expert knowledge for non-circular machine learning in sepsis prediction. Artif Intell Med 2019;100:101725. [doi: 10.1016/j.artmed.2019.101725] [Medline: 31607345]
38. Henry KE, Adams R, Parent C, Soleimani H, Sridharan A, Johnson L, et al. Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. Nat Med 2022;28(7):1447-1454. [doi: 10.1038/s41591-022-01895-z] [Medline: 35864251]
39. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, PROGRESS Group. Prognosis research strategy (PROGRESS) 3: prognostic model research. PLoS Med 2013;10(2):e1001381 [FREE Full text] [doi: 10.1371/journal.pmed.1001381] [Medline: 23393430]
40. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. Lancet 2007;370(9596):1453-1457 [FREE Full text] [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
41. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCH, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101(23):E215-E220. [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]
42. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023;10(1):1 [FREE Full text] [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
43. Generalisability of deep learning-based early warning in the intensive care unit. GitHub. URL: https://github.com/prockenschaub/icuDG [accessed 2026-02-15]
44. Evaluation of Machine Learning Models for Sepsis Prediction. GitHub. URL: https://github.com/medinf-ai/sepsis-eval-study/ [accessed 2026-02-15]

Abbreviations

AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
ICU: intensive care unit
GRU: gated recurrent unit
MIMIC-IV: Medical Information Mart for Intensive Care IV
PPV: positive predictive value

XSL•FO
RenderX

NPV: negative predictive value
SI: suspicion of infection
SOFA: Sequential Organ Failure Assessment
TCN: temporal convolutional network

Edited by J Sarvestan; submitted 18.Feb.2025; peer-reviewed by M Görges, H Shuai-Jing, R Heyer; comments to author 14.Jul.2025; revised version received 10.Dec.2025; accepted 10.Dec.2025; published 24.Mar.2026.

Please cite as:
Do DK, Rockenschaub P, Boie SD, Kumpf O, Volk HD, Balzer F, von Dincklage F, Lichtner G
The Impact of Evaluation Strategy on Sepsis Prediction Model Performance Metrics in Intensive Care Data: Retrospective Cohort Study
J Med Internet Res 2026;28:e72083
URL: https://www.jmir.org/2026/1/e72083
doi:10.2196/72083
PMID:41874553

©Dang-Khoa Do, Patrick Rockenschaub, Sebastian Daniel Boie, Oliver Kumpf, Hans-Dieter Volk, Felix Balzer, Falk von Dincklage, Gregor Lichtner. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Physical Activity Recommendations Tailored by a Predictive Model for Adults With High Blood Pressure: Observational Study

Yuhui Yang[1], PhD; Manqing Chen[1], MS; Weiwei Hu[1], PhD; Yifan Fu[1], MS; Xingyan Li[1], MS; Zhenli Liao[1], MS; Hongman Feng[1], MS; Yaling Zhao[1], PhD; Leilei Pei[1], PhD; Baibing Mi[1*], PhD; Fangyao Chen[1,2*], PhD

[1]Department of Epidemiology and Biostatistics, School of Public Health, Xi'an Jiaotong University Health Science Center, 76 Yanta Xilu Road, Xi'an, Shaanxi, China

[2]Department of Radiology, First Affiliate Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi, China

[*]these authors contributed equally

Corresponding Author:
Fangyao Chen, PhD
Department of Epidemiology and Biostatistics, School of Public Health, Xi'an Jiaotong University Health Science Center, 76 Yanta Xilu Road, Xi'an, Shaanxi, China

Abstract

Background: Whether the benefits of identical physical activity (PA) patterns for adults with high blood pressure (BP) vary according to an individual's characteristics has not been adequately studied.

Objective: This study aimed to investigate whether an individual's characteristics modify the associations between PA patterns and mortality rate.

Methods: Four PA patterns were derived from accelerometer-based data: active weekend warrior, active regular, active light PA, and baseline PA. The main outcome was all-cause mortality. A machine learning model to predict the optimal PA pattern for individual patients was trained in the UK Biobank (UKB) cohort and externally validated in the National Health and Nutrition Examination Survey cohort, which was subsequently integrated into a web-based application. The potentially optimal PA pattern within patients was identified as the one leading to the highest predicted survival probability. Multivariable Cox models were used to estimate hazard ratios and 95% CIs for all-cause mortality corresponding to the inconsistency of the current PA pattern with the predicted optimal PA pattern.

Results: A total of 71,637 UKB adults and 5104 National Health and Nutrition Examination Survey individuals were enrolled. External validation demonstrated that the area under the receiver operating characteristic curve of our model for predicting mortality at 10 years of follow-up was 86.4% (95% CI 85.1% - 87.7%). The predicted optimal PA patterns in the UKB cohort were active regular PA for 26,643 (37.2%) participants, active light PA for 22,606 (31.6%) participants, and active weekend warrior for 21,749 (30.4%) participants. Stroke history, age, sex, BP class, and antihypertension medication were key factors driving heterogeneity in individuals' optimal PA patterns. Cox regression analysis suggested that individuals in the UKB cohort whose current PA patterns were inconsistent with the predicted optimal patterns may be associated with a 28% increase in all-cause mortality risk on average (hazard ratio 1.28, 95% CI 1.20 - 1.38) compared to those with consistent patterns.

Conclusions: Our findings may help patients with high BP obtain individualized recommendations for PA patterns based on their specific characteristics, thereby improving their prognosis.

(J Med Internet Res 2026;28:e78492) doi:10.2196/78492

KEYWORDS

hypertension; physical activity pattern; machine learning; mortality; precision medicine

Introduction

According to the latest report from the World Health Organization, the global population of adults with hypertension has more than doubled between 1990 and 2019, rising from 0.65 billion to 1.3 billion [1,2]. In 2019, high blood pressure (BP) was responsible for 10.8 million deaths worldwide, primarily due to cardiovascular diseases and chronic kidney disease [1,2]. Timely and effective interventions are therefore critical to mitigating this burden. International hypertension management guidelines universally endorse physical activity (PA) as a first-line nonpharmacological intervention for BP control [3,4]. Furthermore, PA offers distinct advantages over antihypertensive medications, including cost-effectiveness and a reduced risk of adverse effects [5,6].

As outlined by Dzau and Hodgkinson [7], patients with hypertension should receive individualized management that accounts for patients' unique genetic and environmental factors,

XSL•FO
RenderX

namely, *precision hypertension*. This perspective highlights the heterogeneity in treatment response among hypertensive individuals and underscores the need for tailored PA strategies. A major challenge in advancing precision hypertension is to move beyond the estimation of average associations and address the variability in individual responses to PA. This variability depends on factors such as personal characteristics, baseline risk profiles, and differential sensitivity to PA interventions [8]. Therefore, conducting studies on the heterogeneous associations between PA and health outcomes and identifying key factors driving this heterogeneity is essential for precision hypertension management.

PA patterns (defined by frequency, duration, and intensity of PA) and their associations with mortality have caught significant attention in recent research [9-11]. Furthermore, these studies on the heterogeneous associations between PA and mortality rely on subgroup analyses [12,13]. For example, Ji et al [13] studied the heterogeneous association between PA and mortality using prespecified gender subgroups. However, such traditional subgroup analyses require prespecification of potential subgroup variables, which limits their exploratory value [14]. Furthermore, one-variable-at-a-time analyses may produce false positives due to multiple testing and false negatives due to limited statistical power in small subgroup samples [15]. The burgeoning field of machine learning (ML) holds promise in estimating heterogeneous treatment effects, potentially offering significant support in addressing this concern [7,16]. Therefore, in our study, we adopted the S-learner framework, a metalearner for estimating individualized treatment effects using ML [17]. The ML based on metalearners can simulate the progression of hypertension and identify optimal PA patterns for managing these conditions.

We hypothesized that the associations of PA patterns with mortality would differ based on individual patient characteristics. In other words, the optimal PA pattern for patients may exhibit heterogeneity depending on individual characteristics. To test this hypothesis, we derived and externally validated a potential outcome prediction model, which was further integrated into a web-based application to identify the optimal PA pattern for an individual with hypertension. In our study, PA patterns were defined using accelerometer data from the UK Biobank (UKB) and the National Health and Nutrition Examination Survey (NHANES) cohorts.

Methods

Study Design and Participants

Our study included two surveys: the UKB and the NHANES. The UKB is a prospective cohort study comprising 502,629 participants enrolled between 2006 and 2010. The NHANES is a biannual survey designed to assess the health and nutritional status of the US population. This study adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [18].

The inclusion criteria for this study were individuals with device-measured PA data and high BP. Exclusion criteria were participants with (1) poor device calibration, (2) inadequate wear time, or (3) missing values for the outcome and covariates. According to the 2024 European Society of Cardiology Guidelines for the management of elevated BP and hypertension [3], participants were classified as having high BP if they met any of the following criteria: systolic BP (SBP) ≥120 mm Hg, diastolic BP (DBP) ≥70 mm Hg, hospitalization records, or the use of antihypertensive medications. In our study, we further categorized high BP into two groups: elevated BP and hypertension. Participants with SBP between 120 and 139 mm Hg or DBP between 70 and 89 mm Hg who were not using antihypertensive medication were classified as having elevated BP; others were classified as having hypertension.

A total of 103,582 participants from the UKB and 14,631 participants from the NHANES cohort were initially considered. After applying the inclusion and exclusion criteria, 71,637 (69.2%) participants from the UKB and 5104 (34.9%) participants from NHANES were included in the final analysis.

Exposure Ascertainment

Information on the collection of accelerometer-based PA data in the UKB and NHANES studies is provided as follows. From 2013 to 2015, UKB participants were randomly assigned to wear an Axivity AX3 accelerometer (Newcastle upon Tyne) for 7 days to measure PA and sedentary behavior. The wrist-worn accelerometers were initialized to capture data with a sampling frequency of 100 Hz and a dynamic range of ±8 g. Poor device calibration was defined as a lack of sufficient orientation changes or as having implausible acceleration values. Good wear time was defined as having at least 3 days (72 h) of data and also having data in each 1-hour period of the 24-hour cycle (scattered over multiple days). The accelerometers recorded data in milligravity units (mg). Light PA (LPA) and moderate to vigorous physical activity (MVPA) were categorized via a published ML-driven approach specifically designed for classifying a wide spectrum of activities [19].

For the NHANES, we included individuals with accelerometer-based PA data from the 2003 to 2006 cycles of NHANES. According to the design of the NHANES study, information on sedentary behavior and total PA was collected using an accelerometer (ActiGraph model 7164; ActiGraph, LLC), worn on the waist for 7 consecutive days during waking hours, except while swimming or showering. Poor device calibration was defined as more than 60 consecutive minutes with zero counts. Good wear time was defined as a recorded wear time of 10 hours or more for at least 1 day [20]. Behavior categories were defined by count per minute thresholds for adults: sedentary behavior (<100 counts per minute), LPA (100 - 2020 counts per minute), and MVPA (≥2020 counts per minute) [21].

We defined four PA patterns according to the following criteria: active weekend warrior (WW; ≥150 min per week MVPA with ≥50% of total achieved in 1 - 2 d) [22], active regular (≥150 min per week and not meeting MVPA WW status) [22], active LPA (<150 min per week MVPA and ≥1900 min per week LPA), and baseline PA (<150 min per week MVPA and <1900 min per week LPA). The threshold of 1900 minutes per week for an active LPA pattern was determined based on the previous study [23], and our analysis of the dose-response relationship

XSL•FO
RenderX

between LPA and all-cause mortality was determined using the Cox regression models with restricted cubic splines (Multimedia Appendix 1).

Mortality Ascertainment

The primary outcome of this study was all-cause mortality, obtained from the National Health Service Information Center (England and Wales) and the National Health Service Central Register Scotland (Scotland) for the UKB study. At the time of analysis, mortality data were available through May 31, 2024, for England and Wales, and December 31, 2023, for Scotland. For NHANES, the National Death Index was used to ascertain all-cause mortality of included samples until December 31, 2019. The follow-up period was defined as the duration between the initial PA measurement and either death or the end of follow-up, whichever occurred first. The causes of death were confirmed based on ICD-10 coding.

Covariates

The covariates were age, sex, ethnicity, BMI, waist circumference, education, smoking status, alcohol consumption, added salt intake, sedentary time, sleep time, SBP, DBP, antihypertension medication, cancer, diabetes, myocardial infarction (MI), stroke, family cardiovascular disease (CVD), BP class, glycated hemoglobin (HbA_{1c}), high-density lipoprotein cholesterol, triglyceride levels, and glucose. Ascertainment and descriptions of covariates considered for both UKB and NHANES are provided in Multimedia Appendix 2.

To minimize the time interval between covariate assessment and accelerometer-based PA measurement (typically 2013 - 2015) in the UKB, we prioritized data from instance 1 (the repeat assessment in 2012 - 2013) when both instance 0 (the baseline assessment in 2006 - 2010) and instance 1 were available [22]. We assessed the reliability of covariates in participants with repeated measurements by the intraclass correlation coefficient [24] for continuous variables and Cohen κ [25] for categorical variables. The analysis indicated moderate to high reliability of the covariates; detailed results are provided in Multimedia Appendix 3. In contrast, the PA and covariates in the external validation set, NHANES, were measured during the same period, thus avoiding the aforementioned issues present in UKB.

Model Development and Estimation of Heterogeneous Associations

We randomly divided the UKB data into an 80% training set and a 20% internal validation set, with the NHANES dataset used for external validation. Predictors were selected using Cox regression with the least absolute shrinkage and selection operator (LASSO) penalization [26], where the optimal shrinkage parameter lambda was determined via 10-fold cross-validation [27]. The original predictive model was constructed using a multivariable Cox model that incorporated all second-order interactions between PA pattern and the selected predictors. Then, we applied a stepwise backward elimination method to refine the model, removing variables and interaction terms that did not significantly contribute to the model [28]. Model performance was evaluated in the internal validation set and the NHANES validation set and visualized using the calibration curves (with Brier score) and receiver operating characteristic (ROC) analysis (with the area under the ROC curve [AUC]) [29,30].

Our S-learner algorithm for estimating heterogeneous associations between PA and overall survival follows a 2-step process. First, it uses the trained prediction model to estimate the conditional expectations of survival time for each of the 4 PA patterns separately. Second, it computes the differences between these estimates to capture the heterogeneous associations.

On the basis of the S-learner framework, the trained prediction model was capable of handling potential variations in covariates, enabling the prediction of each patient's potential survival probabilities under the 4 PA patterns. Then, the best potential PA pattern for a certain patient was identified as the one leading to the highest predicted survival probability. To operationalize these findings, we developed an R Shiny web application that enables clinicians to simulate personalized survival curves for all 4 PA patterns based on individual baseline characteristics.

Patient Characteristics Associated With Predicted Individualized PA Patterns

All individuals in the UKB and the NHANES cohorts were stratified by PA patterns that might individually optimize overall survival. This stratification was used to identify the profiles of individuals who could benefit most from 1 of the 4 PA patterns presented, according to the constructed prediction model.

To facilitate rapid decision-making and clearly illustrate patient characteristics associated with predicted individualized PA patterns, a model of conditional inference tree (CIT) was established [31]. Specifically, we used the potentially optimal PA patterns predicted as the outcome and all covariates as predictors to build the CIT model.

Statistical Analysis

We introduced an indicator variable, referred to as an "inconsistent PA pattern," to distinguish between individuals whose actual and predicted optimal PA patterns were inconsistent and those whose patterns were consistent. Multivariable Cox models were used to estimate hazard ratios (HRs) and 95% CIs for all-cause mortality corresponding to the inconsistency of the current PA pattern with the predicted optimal PA pattern. The analyses were adjusted for covariates selected by LASSO.

As further stratification of participants with inconsistent PA patterns provides more clinically meaningful information, we subdivided the inconsistent group into two categories based on the predicted optimal intensity: (1) inconsistent MVPA pattern—participants whose predicted optimal PA pattern required more MVPA but whose actual activity was lower than predicted and (2) inconsistent LPA pattern—participants whose predicted optimal PA pattern required LPA primarily but whose actual activity (including active regular, active WW, and baseline PA patterns) differed. The original binary variable "inconsistent PA pattern" was thus updated to a category variable with three levels: consistent, inconsistent MVPA pattern, and inconsistent LPA pattern. We then used "consistent"

XSL•FO
RenderX

as a reference class for the aforementioned multivariable Cox model analyses. In addition, we conducted subgroup analysis in the actually observed nonbaseline group.

Sensitivity Analyses

To test the generalizability of the aforementioned analyses, we performed 4 prespecified sensitivity analyses. First, we repeated the statistical analyses using the NHANES dataset. Second, we replaced the all-cause mortality outcome with CVD-specific mortality and cancer-specific mortality outcomes. Third, to minimize the potential confounding effects of the COVID-19 pandemic, we restricted follow-up to the period before December 31, 2019. This prepandemic dataset was used as an additional validation cohort to assess the robustness of the model's performance and to repeat the statistical analyses. Finally, the statistical analyses were repeated using the internal validation set.

All analyses were conducted using R version 4.4.1 (R Foundation), with statistical significance set at $P<.05$.

Ethical Considerations

This study involved secondary analysis of data from the UKB and the NHANES, whose original data collection protocols were conducted in accordance with the Declaration of Helsinki. UKB was approved by the North West Multi-centre Research Ethics Committee (reference 21/NW/0157), and all participants provided written informed consent for health-related research, including secondary analyses. The NHANES study received ethical approval from the National Center for Health Statistics Ethics Review Board. The National Center for Health Statistics Ethics Review Board approved Protocol 98‐12 for the 2003‐2004 cycle and Protocol 2005‐06 for the 2005‐2006 cycle. All participants provided written informed consent. All data were anonymized or deidentified.

Results

Overview

For the UKB cohort, a total of 71,637 adults were enrolled and analyzed (Figure 1), with 4207 deaths confirmed during a median follow-up period of 9.5 years. For the NHANES cohort, we included 5104 individuals who underwent device-based activity measurements between the 2003 and 2006 cycles (Multimedia Appendix 4), with 1279 deaths confirmed during a median follow-up period of 14.3 years. The median follow-up time for survivors and nonsurvivors in both the UKB and the NHANES datasets is shown in Multimedia Appendix 5. The baseline characteristics of participants, stratified by 4 actually *observed* PA patterns, are summarized in Multimedia Appendix 6 for the UKB cohort and the NHANES cohort. At baseline in the UKB, 12,577 (17.6%) participants were classified as baseline PA, 12,782 (17.8%) participants were classified as active LPA, 15,141 (21.1%) participants were classified as active regular, and 31,137 (43.5%) participants were classified as active WW.

XSL•FO
RenderX

Figure 1. Study flow for UK Biobank. CVD: cardiovascular disease; DBP: diastolic blood pressure; HbA_{1c}: glycated hemoglobin; HDL: high-density lipoprotein; LPA: light physical activity; MVPA: moderate to vigorous physical activity; SBP: systolic blood pressure.

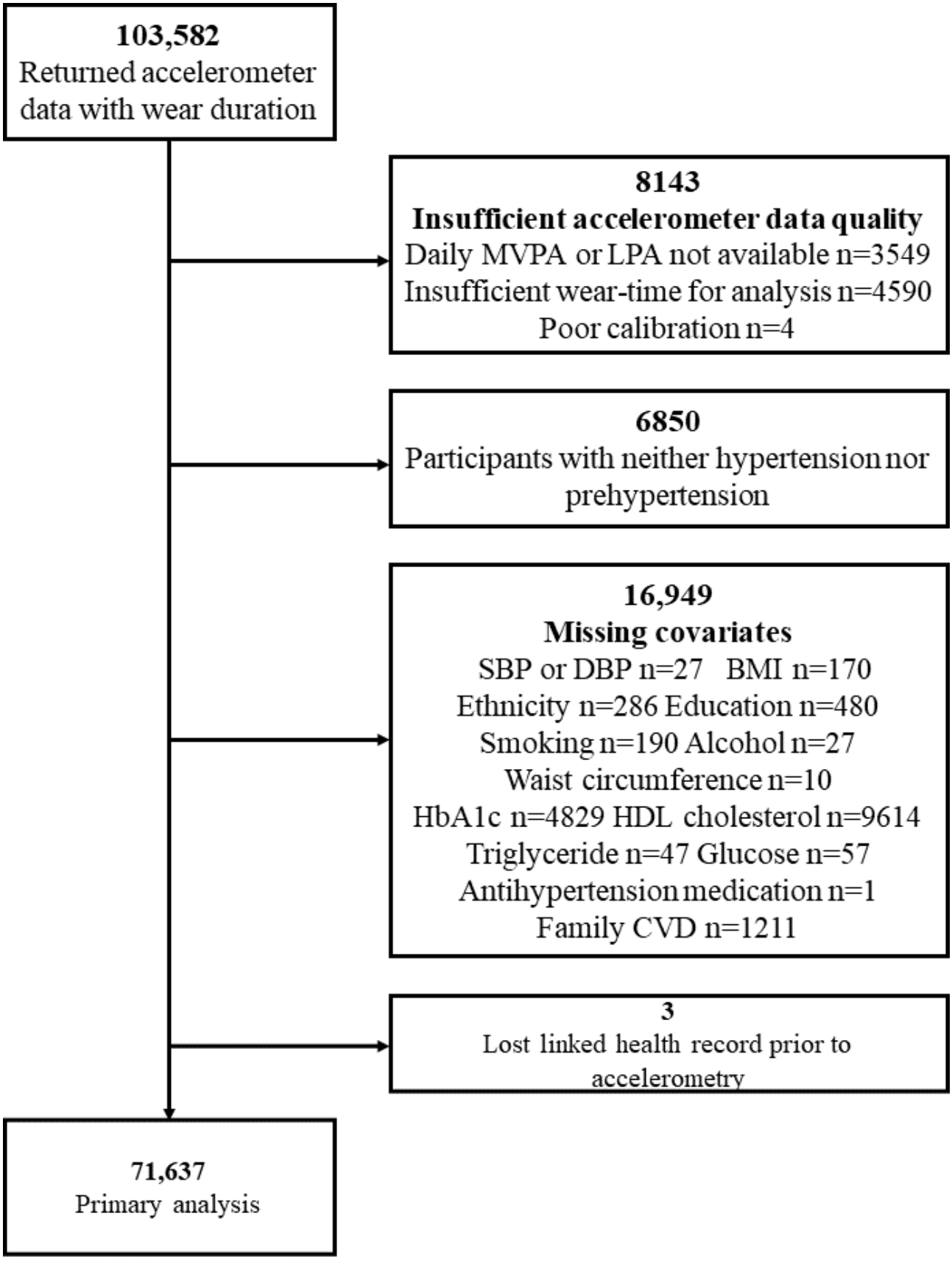

XSL•FO
RenderX

A Prediction Model to Identify the Best Potential PA Patterns

The detailed results from the LASSO-penalized Cox model for covariate selection are presented in Multimedia Appendix 7. In brief, the selected covariates included PA patterns, age, sex, sedentary time, smoking status, antihypertension medication, cancer, diabetes, MI, stroke, BP class, waist circumference, glucose, and HbA_{1c}. In the UKB cohort, the associations between covariates selected by LASSO and all-cause mortality were examined using univariate Cox models (Multimedia Appendix 8): PA patterns were found to be associated with overall survival, suggesting that adults with active LPA (HR 0.56, 95% CI 0.51 - 0.62), active regular (HR 0.45, 95% CI 0.41 - 0.50), or active WW (HR 0.47, 95% CI 0.43 - 0.51) were at a lower risk of death than those with the baseline PA pattern.

The original predictive model incorporated all second-order interactions of PA patterns with each selected covariate. Following stepwise backward elimination, the main effect of glucose and the interactions of PA patterns with waist circumference, smoking status, and glucose were removed from the model. The terms and coefficients of the final model are detailed in Multimedia Appendix 9. The model showed well-calibrated predictions in the internal test set, with Brier scores of 2.0% (95% CI 1.8% - 2.2%) for the 5-year and 5.1% (95% CI 4.7% - 5.4%) for the 10-year prediction (Figure 2A). Calibration remained acceptable in the external NHANES set, although it was slightly decreased, with Brier scores of 5.2% (95% CI 4.7% - 5.8%) for the 5-year and 9.7% (95% CI 9.1% - 10.4%) for the 10-year prediction (Figure 2B). The model also demonstrated good discriminative ability. In the internal set, the AUC was 76.8 % (95% CI 74.2% - 79.5%) for the 5-year and 78.1% (95% CI 76.2% - 79.9%) for the 10-year prediction (Figure 2C). This performance was further improved in the external validation set, with AUCs of 85.0% (95% CI 82.9% - 87.1%) and 86.4% (95% CI 85.1% - 87.7%) for the 5- and 10-year prediction, respectively (Figure 2D).

XSL•FO
RenderX

Figure 2. Performance of the model in predicting all-cause mortality. The calibration and ROC curves for internal (n=14,328) and external test data set (n=5104). AUC: area under the receiver operating characteristic curve; NHANES: National Health and Nutrition Examination Survey; ROC: receiver operating characteristic.

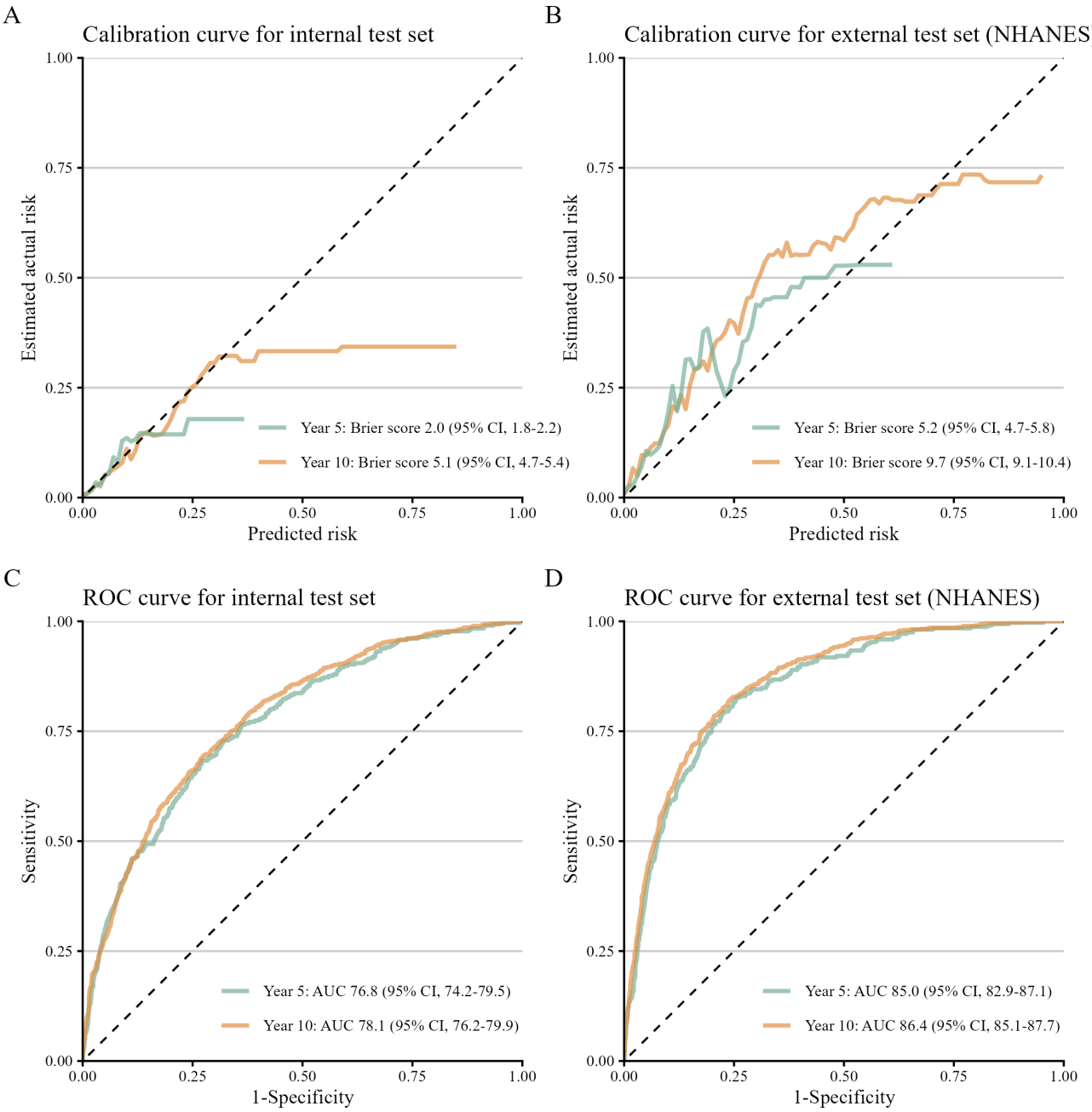

The prediction model identifying the best PA pattern has been integrated into an online application, accessible free of charge (Multimedia Appendix 10). To ensure broader accessibility of the application, we have provided both English and Chinese versions (Multimedia Appendix 10). Users need to enter the required clinical characteristics (including age, sex, sedentary time, smoking status, antihypertension medication, cancer, diabetes, MI, stroke, BP class, waist circumference, and HbA_{1c}) into the website. The tool then calculates the predicted 5-year and 10-year survival probabilities corresponding to the 4 PA patterns. After these probabilities are generated, the application identifies the pattern with the highest predicted survival probability. This information can support clinicians and patients in selecting the most appropriate personalized PA pattern.

Patient Characteristics Associated With Predicted Optimal PA Patterns

Our predictive model indicated that the optimal PA pattern varied among adults with high BP. Specifically, 26,643 (37.2%) participants were predicted to benefit most from the active regular pattern, 22,606 (31.6%) participants from the active LPA pattern, 21,749 (30.4%) participants from the active WW pattern, and 639 (0.9%) participants from the baseline PA pattern. Table 1 presents the characteristics of participants according to their *predicted* optimal PA patterns and compares these with the patterns they actually followed in the UKB cohort. In addition, we applied the same predictive approach to NHANES participants and compared *predicted* versus *observed* PA patterns, as shown in Multimedia Appendix 11.

XSL•FO
RenderX

Table . Baseline characteristics of participants stratified by the predicted optimal physical activity (PA) pattern.

Variables and levels	Baseline PA (n=639)	Active LPA[a] (n=22,606)	Active regular (n=26,643)	Active WW[b] (n=21,749)	*P* value
Pattern actually observed, n (%)					<.001
Baseline PA	191 (29.9)	3118 (13.8)	5238 (19.7)	4030 (18.5)	
Active LPA	118 (18.5)	4532 (20.0)	4734 (17.8)	3398 (15.6)	
Active regular	105 (16.4)	5169 (22.9)	5467 (20.5)	4400 (20.2)	
Active WW	225 (35.2)	9787 (43.3)	11,204 (42.1)	9921 (45.6)	
Age (y), median (IQR)	67.9 (63.7-71.3)	56.8 (51.7-62.4)	62.8 (56.9-67.3)	69.8 (66.8-72.5)	<.001
Sex, n (%)					<.001
Male	447 (70)	4407 (19.5)	12,069 (45.3)	15,948 (73.3)	
Female	192 (30)	18,199 (80.5)	14,574 (54.7)	5801 (26.7)	
WC[c] (cm), median (IQR)	96.0 (87.0-104.2)	82.0 (75.0-91.0)	90.0 (81.0-99.0)	93.0 (85.0-101.0)	<.001
Smoking, n (%)					<.001
Never	283 (44.3)	14,066 (62.2)	15,541 (58.3)	10,943 (50.3)	
Previous	300 (46.9)	6979 (30.9)	9206 (34.6)	9639 (44.3)	
Current	56 (8.8)	1561 (6.9)	1896 (7.1)	1167 (5.4)	
Antihypertension medication, n (%)					<.001
No	217 (34)	19,694 (87.1)	23,541 (88.4)	14,643 (67.3)	
Yes	422 (66)	2912 (12.9)	3102 (11.6)	7106 (32.7)	
Blood pressure class, n (%)					<.001
Elevated	54 (8.5)	19,513 (86.3)	3405 (12.8)	9907 (45.6)	
Hypertension	585 (91.5)	3093 (13.7)	23,238 (87.2)	11,842 (54.4)	
MI[d], n (%)					<.001
No	556 (87)	22,595 (100)	26,097 (98)	20,705 (95.2)	
Yes	83 (13)	11 (0)	546 (2)	1044 (4.8)	
Stroke, n (%)					<.001
No	0 (0)	22,527 (99.7)	26,585 (99.8)	21,409 (98.4)	
Yes	639 (100)	79 (0.3)	58 (0.2)	340 (1.6)	
Diabetes, n (%)					<.001
No	573 (89.7)	21,610 (95.6)	26,454 (99.3)	19,405 (89.2)	
Yes	66 (10.3)	996 (4.4)	189 (0.7)	2344 (10.8)	
Cancer, n (%)					<.001
No	534 (83.6)	22,291 (98.6)	22,158 (83.2)	20,519 (94.3)	
Yes	105 (16.4)	315 (1.4)	4485 (16.8)	1230 (5.7)	
Glucose (mmol/L), median (IQR)	5.0 (4.7-5.5)	4.8 (4.5-5.2)	5.0 (4.6-5.3)	5.0 (4.6-5.4)	<.001
HbA_{1c}[e] (mmol/mol), median (IQR)	36.9 (34.5-39.8)	33.8 (31.5-36.1)	35.5 (33.2-37.9)	35.5 (33.1-38.1)	<.001

[a]LPA: light physical activity.

[b]WW: weekend warrior.

[c]WC: waist circumference.

[d]MI: myocardial infarction.

[e]HbA_{1c}: glycated hemoglobin.

The results of the CIT model are shown in Figure 3. Briefly, the model is rooted by stroke, with subsequent branching by age, BP class, sex, and antihypertension medication (all *P*<.001). Adults with stroke are more likely to have baseline PA or active

XSL•FO
RenderX

LPA as optimal patterns. In contrast, for those without stroke, younger adults (≤68 y) tend to benefit more from active regular and active LPA, whereas older men (>68 y) are predicted to gain more from the active WW pattern. In summary, stroke, age, sex, BP class, and antihypertension medication are key factors driving heterogeneity in individuals' optimal PA patterns. Notably, the baseline PA pattern denotes a lower level of PA rather than complete physical inactivity or sedentary behavior.

Figure 3. The derived conditional inference tree. The models were constructed by the optimal PA pattern predicted (outcome) and all covariates selected (predictors). Antihyper med: antihypertension medication; BP: blood pressure; LPA: light physical activity; WW: weekend warrior.

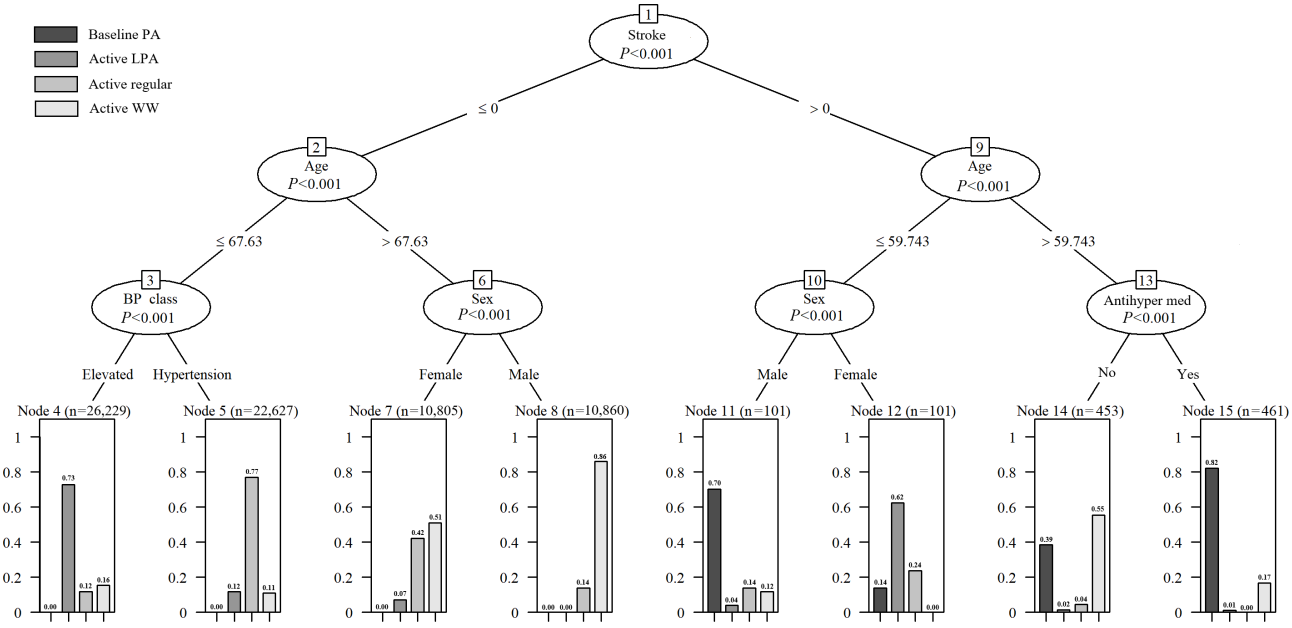

The Associations Between Inconsistent PA Patterns and Mortality

The PA patterns observed in 71.9% of participants in the UKB cohort and 78.9% of participants in the NHANES cohort were inconsistent with the optimal patterns predicted by our model. Compared to individuals whose actual and predicted optimal PA patterns were consistent, an inconsistent PA pattern was associated with a 28% increase in all-cause mortality risk on average (HR 1.28, 95% CI 1.20 - 1.38; Figure 4A) in the UKB cohort. Specifically, an inconsistent MVPA pattern was associated with a 31% increased risk (HR 1.31, 95% CI 1.22 - 1.41; Figure 4B), whereas an inconsistent LPA pattern was associated with a 14% increased risk (HR 1.14, 95% CI 1.01 - 1.30; Figure 4B). In subgroup analysis, we found no statistically significant association between the inconsistent LPA pattern and all-cause mortality (HR 1.06, 95% CI 0.91 - 1.23; Figure 5).

XSL•FO
RenderX

Figure 4. Forest plot of hazard ratios (HRs) from the multivariate Cox regression model. Outcome: all-cause mortality. Exposure: inconsistent PA patterns (distinguishing between individuals with consistent [actual vs predicted optimal PA patterns] and inconsistent PA patterns). In Figure 4A, the exposure is a binary variable (consistent vs inconsistent); in Figure 4B, it is a categorical variable with 3 levels (consistent [reference group], inconsistent MVPA pattern, and inconsistent LPA pattern). The Cox regressions were analyzed and sorted by HR. BP: blood pressure; HbA_{1c}: glycated hemoglobin; LPA: light physical activity; MVPA: moderate to vigorous physical activity; PA: physical activity; UKB: UK Biobank.

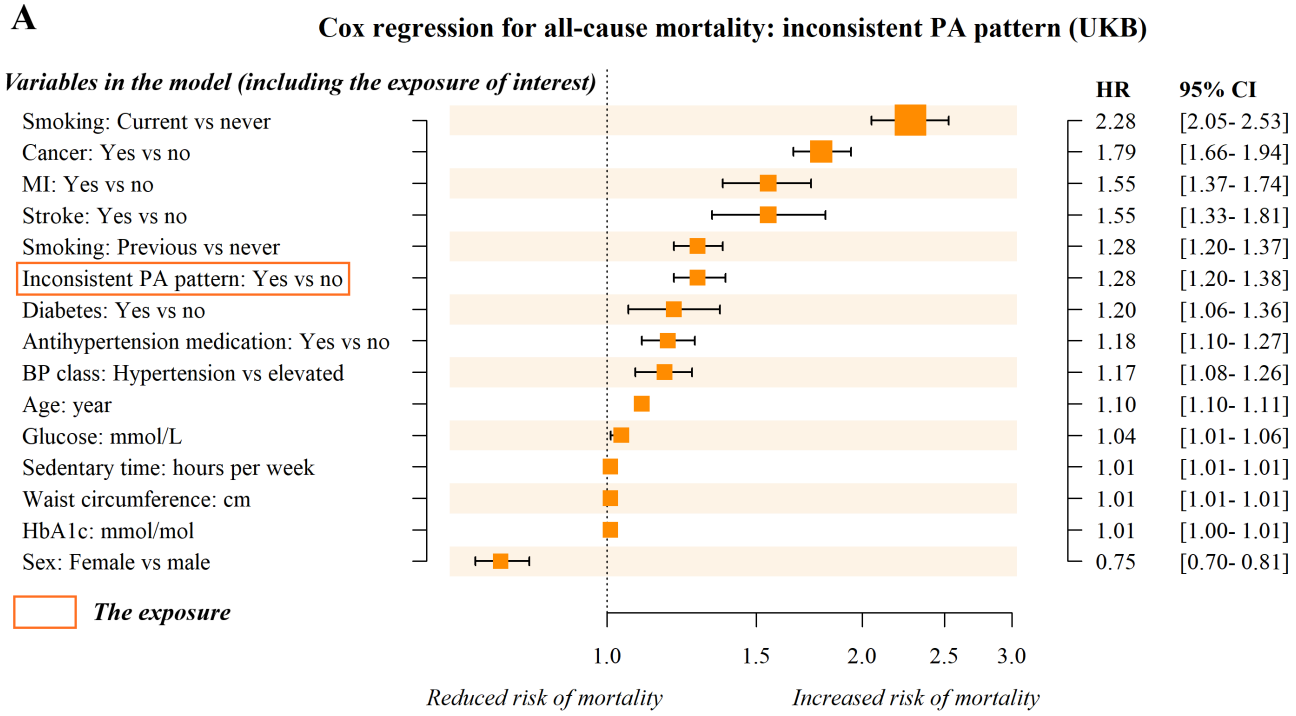

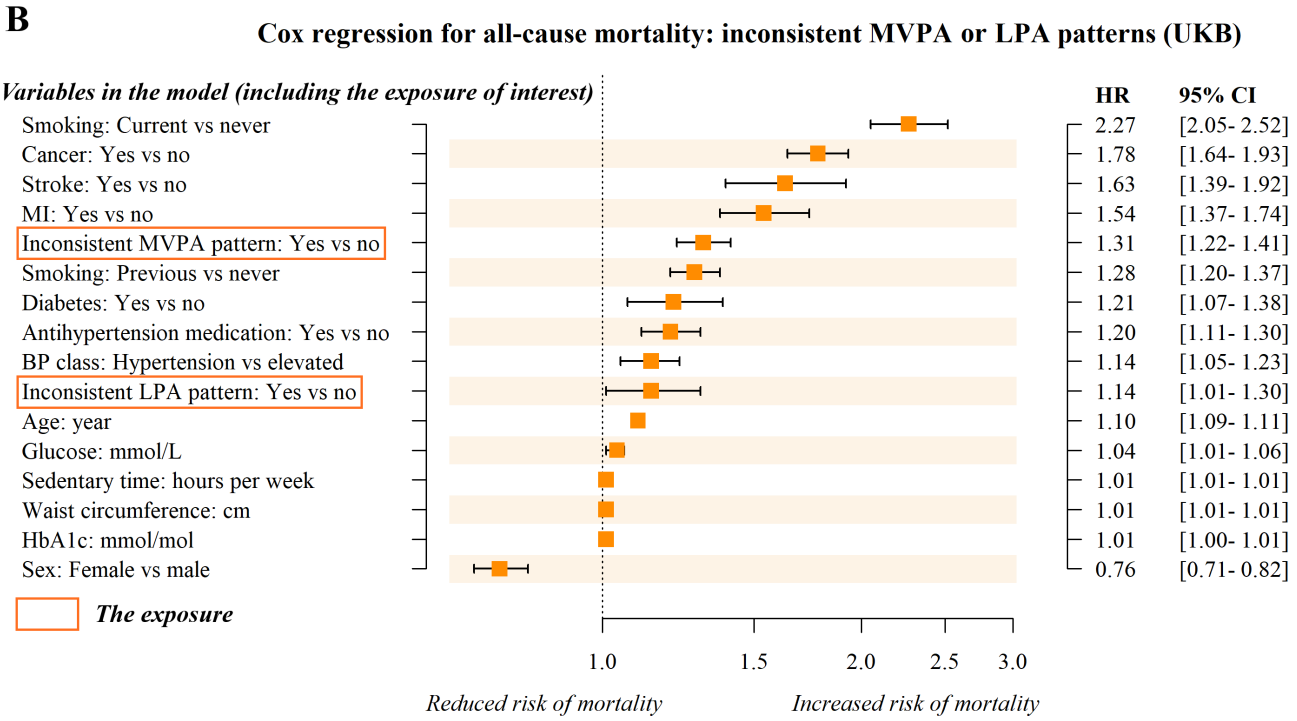

XSL•FO
RenderX

Figure 5. Forest plot of hazard ratios (HRs) from the multivariate Cox regression model in the nonbaseline subgroup. Outcome: all-cause mortality. Exposure: inconsistent PA patterns (distinguishing between individuals with consistent [actual vs predicted optimal PA patterns] and inconsistent PA patterns) with 3 levels (consistent [reference group], inconsistent MVPA pattern, and inconsistent LPA pattern). The nonbaseline group comprises participants whose actually observed PA pattern was not the "baseline PA" pattern. The Cox regressions were analyzed and sorted by HR. BP: blood pressure; HbA_{1c}: glycated hemoglobin; LPA: light physical activity; MVPA: moderate to vigorous physical activity; PA: physical activity; UKB: UK Biobank.

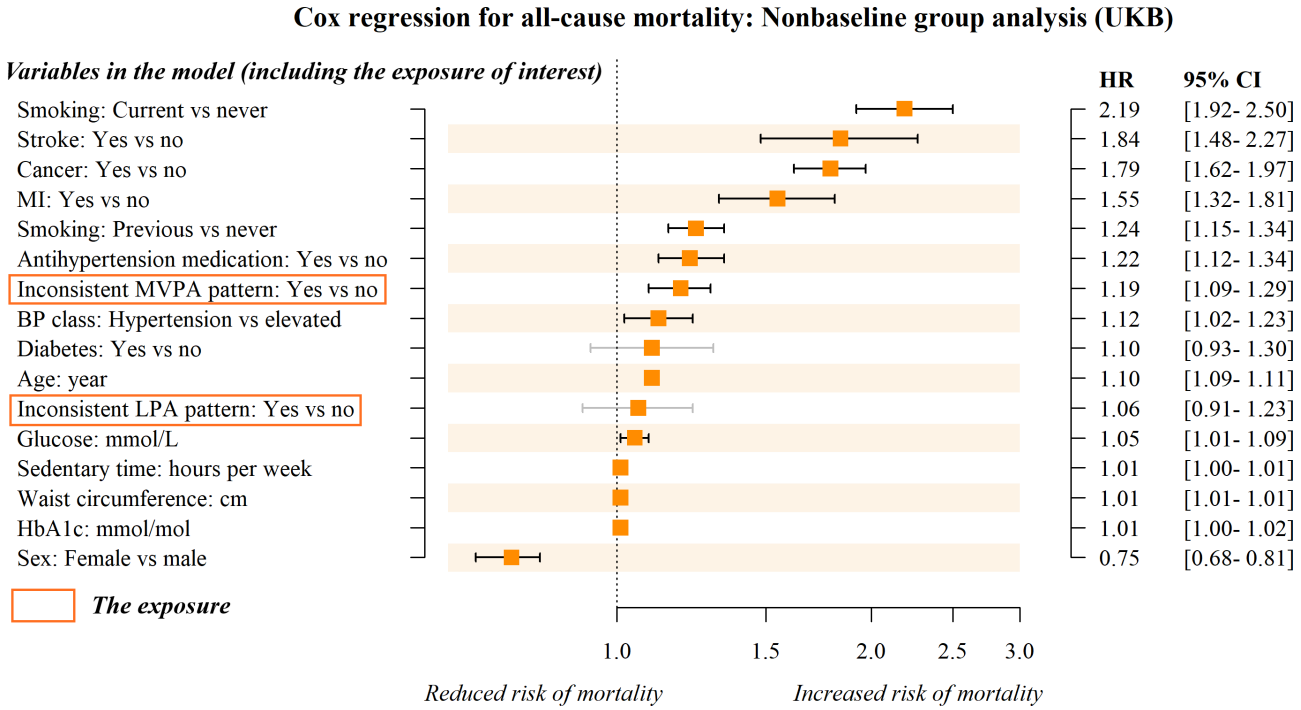

Sensitivity Analyses

The results of the sensitivity analyses are demonstrated in Multimedia Appendix 12. The results of the first sensitivity analysis using the NHANES data showed that an inconsistent PA pattern was associated with a 26% increase in all-cause mortality risk (HR 1.26, 95% CI 1.06 - 1.51), compared to individuals with consistent actual and predicted optimal PA patterns. The analyses of the associations between inconsistent MVPA or LPA patterns and mortality, as well as subgroup analyses, also yielded results similar to those in the UKB cohort. The results of the second sensitivity analysis using cause-specific mortality outcomes indicated that an inconsistent PA pattern was associated with an 18% increase in cancer mortality risk (HR 1.18, 95% CI 1.07 - 1.30) and a 33% increase in CVD mortality risk (HR 1.33, 95% CI 1.18 - 1.49). Subgroup analyses revealed that the inconsistent LPA pattern was not statistically significantly associated with either cancer or CVD mortality in the UKB cohort. In the third sensitivity analysis, we used the prepandemic dataset for external validation. The results showed that the Brier score of our model for predicting 5-year mortality was 2.0% (95% CI 1.8% - 2.3%), and the AUC was 78.1% (95% CI 75.3% - 80.8%). In addition, both the overall and subgroup analyses indicated no statistically significant associations between the inconsistent LPA pattern and all-cause mortality. The results of the fourth sensitivity analysis were consistent with those of the primary analysis.

Discussion

Principal Findings

On the basis of the analysis of 2 temporally and geographically distinct cohorts, we found that patients' individual characteristics modified the benefit of PA patterns on all-cause mortality. Our findings in recognition of the best potential PA pattern revealed that up to 67.6% of patients with high BP were advised to engage in over 150 minutes of MVPA per week (37.2% for active regular and 30.4% for active WW). However, we also found that only 31.8% of them were actually with the predicted optimal PA pattern. In other words, according to the prediction model, a significant portion of the participants may not follow the most suitable PA pattern based on their underlying personal characteristics. Moreover, we found that the risk of death was significantly lower when current PA patterns were consistent with the predicted optimal patterns. This suggests the urgent need to implement standardized, individualized, and evidence-based PA pattern recommendations for patients with hypertension to enhance their prognosis.

Our subgroup analysis of the nonbaseline participants revealed no statistically significant associations between the inconsistent LPA pattern and all-cause mortality. As the subgroup with an inconsistent LPA pattern among nonbaseline participants was predominantly engaged in MVPA, this null finding suggests that increasing PA intensity and volume beyond the predicted levels may not offer additional benefits. Therefore, identifying such subgroups may be important, as simple goals could increase patients' confidence [32] and adherence [33] to PA engagement. However, while our study strongly supports the importance of

XSL•FO
RenderX

meeting the minimum PA intensity threshold, evidence on the potential harm of excessive exercise beyond the predicted level remains weak. This warrants further studies with stronger causal evidence, such as randomized controlled trials [34,35].

Hypertension frequently coexists with cardiovascular disease, obesity, diabetes, or cognitive decline, particularly among older adults [36-39]. These comorbidities complicate PA management in hypertensive patients due to two key challenges: (1) older people with comorbidities tend to have limited physical ability to engage in MVPA [40] and (2) VPA may attenuate PA's cardiovascular benefits by inducing adverse cardiac remodeling in vulnerable individuals [41-43]. Consequently, PA regimens should be tailored to individual physical conditions, and our model provides insights to guide this personalization. For instance, the identified interaction terms reveal that the associations between specific PA patterns and mortality differ by comorbidity. Specifically, for patients with MI who are engaged in cardiac rehabilitation, regular MVPA may offer greater benefits [44]. Hypertensive individuals with diabetes may benefit more from light-intensity PA, as excessive exercise could lead to hypoglycemia [45]. Additionally, participants with a stroke comorbidity are more likely to benefit from a lower level of PA due to the physical limitations and health risks associated with stroke recovery. Light-intensity PA offers a safer, more sustainable option, promoting cardiovascular health and functional recovery without overwhelming the individual [46,47]. However, for high-comorbidity groups, the model-predicted optimal pattern may serve as a prognostic marker rather than a prescriptive target for exercise limitation due to potential unmeasured confounders.

Advances in artificial intelligence have the potential to make personalized health advice more accessible and affordable for a broader range of individuals [48,49]. The model derived and online application proposed in this study could simulate various prognostic scenarios under different PA patterns for adults with hypertension and help patients choose the individualized PA pattern. Importantly, our algorithm can be integrated into current practices that rely on in-person assessments and regular follow-ups, as a support but not as a replacement for the physicians' professionalism and experience [50].

Study Limitations

This study also has several limitations. First, PA patterns were assessed at baseline using a short-term (7 d) estimate, whereas the median follow-up was 9.5 years for UKB and 14.3 years for NHANES. This may not accurately reflect typical PA patterns over time. Additionally, our predictive model assumes a static PA pattern, which limits its clinical applicability. Second, although we prioritized covariate data from the repeat assessment (2012 - 2013) to align with the accelerometer measurement period (2013 - 2015), most of the covariate data were sourced from the baseline assessment (2006 - 2010). This introduces the risk of measurement error due to temporal changes in lifestyle or health status, thereby compromising the tool's real-world predictive accuracy. Third, the ML model we developed was based on the observation survey rather than a randomized controlled trial, so the capacity of our results to elucidate causal relationships is limited. Fourth, the observed association between the "inconsistent PA pattern" and higher mortality risk may be influenced by residual confounding from unmeasured health-seeking behaviors. While our sensitivity analysis results have enhanced the reliability of our conclusions, the unmeasured confounders cannot be overlooked. Fifth, the majority of participants were White, which limits the generalizability of the model and means its predictions may have limited applicability to non-White populations. Sixth, the prediction model was trained and validated using objective accelerometer data; however, the use of self-reported sedentary time as input in the web application may result in inaccurate survival predictions [51]. In addition, the reliance on objective sedentary time and the need for recent blood biochemical marker data may limit the tool's applicability to the general public and primary prevention. Finally, as the optimal PA pattern driven by small increases in survival probability may lack clinical significance, establishing an appropriate threshold should be an important focus for future research. In addition, the threshold selection of the "active LPA pattern" merits further inquiry to verify its biological and clinical significance.

Conclusions

Our study found that the optimal PA pattern was heterogeneous based on individuals' underlying characteristics, whereas only a few participants actually followed the most suitable PA pattern as we predicted. The algorithm presented herein could assist in assigning patients with high BP to the individualized PA pattern based on their specific characteristics, which can be easily accessed through an online application.

Acknowledgments

All authors are grateful to the UK Biobank and the National Health and Nutrition Examination Survey participants and professionals. The authors declare that no generative AI was used in any portion of the manuscript writing.

Funding

This work was supported by the National Social Science Fund of China (21CTJ009).

Data Availability

The datasets generated or analyzed during this study are available in the UK Biobank [52] and the National Health and Nutrition Examination Survey [53] repository.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: YY, FC
Data curation: BM, YY
Formal analysis: YY, MC, WH
Funding acquisition: FC
Methodology: YY, MC, WH
Project administration: FC, BM
Resources: FC, BM
Software: YY, MC, WH, YF, XL, ZL, HF
Supervision: FC, BM, YZ, LP
Validation: all authors
Visualization: YY, MC, WH
Writing – original draft: YY
Writing – review & editing: YY, MC, WH, FC, BM, YZ, LP

Conflicts of Interest

None declared.

Multimedia Appendix 1
The analysis of the dose-response relationship between light physical activity and all-cause mortality.
[DOCX File, 70 KB - jmir_v28i1e78492_app1.docx]

Multimedia Appendix 2
Covariates ascertainment and descriptions.
[DOCX File, 42 KB - jmir_v28i1e78492_app2.docx]

Multimedia Appendix 3
The reliability measurements of covariates.
[DOCX File, 25 KB - jmir_v28i1e78492_app3.docx]

Multimedia Appendix 4
Study flow for National Health and Nutrition Examination Survey.
[DOCX File, 44 KB - jmir_v28i1e78492_app4.docx]

Multimedia Appendix 5
Time to censoring and time to death for the UK Biobank and National Health and Nutrition Examination Survey sets.
[DOCX File, 224 KB - jmir_v28i1e78492_app5.docx]

Multimedia Appendix 6
Baseline characteristics of participants.
[DOCX File, 39 KB - jmir_v28i1e78492_app6.docx]

Multimedia Appendix 7
The detailed results obtained using the least absolute shrinkage and selection operator penalized Cox model for covariate selection.
[DOCX File, 73 KB - jmir_v28i1e78492_app7.docx]

Multimedia Appendix 8
The associations between covariates selected and all-cause mortality by univariate Cox models in the UK Biobank cohort.
[DOCX File, 24 KB - jmir_v28i1e78492_app8.docx]

Multimedia Appendix 9
Coefficients and SEs of the Cox prediction model.
[DOCX File, 26 KB - jmir_v28i1e78492_app9.docx]

Multimedia Appendix 10
A snapshot from the web application.

XSL•FO
RenderX

[DOCX File, 567 KB - jmir_v28i1e78492_app10.docx]

Multimedia Appendix 11
A baseline characteristic of participants stratified by the predicted optimal physical activity pattern in the National Health and Nutrition Examination Survey cohort.
[DOCX File, 25 KB - jmir_v28i1e78492_app11.docx]

Multimedia Appendix 12
The results of the sensitivity analyses.
[DOCX File, 1535 KB - jmir_v28i1e78492_app12.docx]

References

1. Global report on hypertension: the race against a silent killer. : World Health Organization; 2023 URL: https://www.who.int/publications/i/item/9789240081062 [accessed 2023-09-19]
2. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants. Lancet 2021 Sep 11;398(10304):957-980. [doi: 10.1016/S0140-6736(21)01330-1] [Medline: 34450083]
3. McEvoy JW, McCarthy CP, Bruno RM, et al. 2024 ESC guidelines for the management of elevated blood pressure and hypertension: developed by the task force on the management of elevated blood pressure and hypertension of the European Society of Cardiology (ESC) and endorsed by the European Society of Endocrinology (ESE) and the European Stroke Organisation (ESO). Eur Heart J 2024 Jul;45(38):3912-4018. [doi: 10.1093/eurheartj/ehae178]
4. Whelton PK, Carey RM, Aronow WS, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the prevention, detection, evaluation, and management of high blood pressure in adults. J Am Coll Cardiol 2018 May;71(19):e127-e248. [doi: 10.1016/j.jacc.2017.11.006]
5. Zhou YF, Liu N, Wang P, et al. Cost-effectiveness of drug treatment for Chinese patients with stage I hypertension according to the 2017 hypertension clinical practice guidelines. Hypertension 2020 Sep;76(3):750-758. [doi: 10.1161/HYPERTENSIONAHA.119.14533] [Medline: 32713271]
6. Byrd JB, Brook RD. Hypertension. Ann Intern Med 2019 May 7;170(9):ITC65-ITC80. [doi: 10.7326/AITC201905070] [Medline: 31060074]
7. Dzau VJ, Hodgkinson CP. Precision hypertension. Hypertension 2024 Apr;81(4):702-708. [doi: 10.1161/HYPERTENSIONAHA.123.21710] [Medline: 38112080]
8. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q 2004;82(4):661-687. [doi: 10.1111/j.0887-378X.2004.00327.x] [Medline: 15595946]
9. Dos Santos M, Ferrari G, Lee DH, et al. Association of the “weekend warrior” and other leisure-time physical activity patterns with all-cause and cause-specific mortality: a nationwide cohort study. JAMA Intern Med 2022 Aug 1;182(8):840-848. [doi: 10.1001/jamainternmed.2022.2488] [Medline: 35788615]
10. Füzéki E, Engeroff T, Banzer W. Health benefits of light-intensity physical activity: a systematic review of accelerometer data of the National Health and Nutrition Examination Survey (NHANES). Sports Med 2017 Sep;47(9):1769-1793. [doi: 10.1007/s40279-017-0724-0] [Medline: 28393328]
11. Hamaya R, Shiroma EJ, Moore CC, Buring JE, Evenson KR, Lee IM. Time-vs step-based physical activity metrics for health. JAMA Intern Med 2024 Jul 1;184(7):718-725. [doi: 10.1001/jamainternmed.2024.0892] [Medline: 38767892]
12. Cao Y, Baumgartner KB, Visvanathan K, Boone SD, Baumgartner RN, Connor AE. Ethnic and biological differences in the association between physical activity and survival after breast cancer. NPJ Breast Cancer 2020;6:51. [doi: 10.1038/s41523-020-00194-5] [Medline: 33083530]
13. Ji H, Gulati M, Huang TY, et al. Sex differences in association of physical activity with all-cause and cardiovascular mortality. J Am Coll Cardiol 2024 Feb 27;83(8):783-793. [doi: 10.1016/j.jacc.2023.12.019] [Medline: 38383092]
14. Alosh M, Huque MF, Bretz F, D’Agostino RB Sr. Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. Stat Med 2017 Apr 15;36(8):1334-1360. [doi: 10.1002/sim.7167] [Medline: 27891631]
15. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine-reporting of subgroup analyses in clinical trials. N Engl J Med 2007 Nov 22;357(21):2189-2194. [doi: 10.1056/NEJMsr077003] [Medline: 18032770]
16. Kent DM, Paulus JK, van Klaveren D, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement. Ann Intern Med 2020 Jan 7;172(1):35-45. [doi: 10.7326/M18-3667] [Medline: 31711134]
17. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Natl Acad Sci U S A 2019 Mar 5;116(10):4156-4165. [doi: 10.1073/pnas.1804597116] [Medline: 30770453]
18. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. PLOS Med 2007 Oct 16;4(10):e297. [doi: 10.1371/journal.pmed.0040297] [Medline: 17941715]

XSL•FO
RenderX

19. Walmsley R, Chan S, Smith-Byrne K, et al. Reallocation of time between device-measured movement behaviours and risk of incident cardiovascular disease. Br J Sports Med 2022 Sep;56(18):1008-1017. [doi: 10.1136/bjsports-2021-104050]
20. Saint-Maurice PF, Troiano RP, Bassett DR Jr, et al. Association of daily step count and step intensity with mortality among US adults. JAMA 2020 Mar 24;323(12):1151-1160. [doi: 10.1001/jama.2020.1382] [Medline: 32207799]
21. Smirnova E, Leroux A, Cao Q, et al. The predictive performance of objective measures of physical activity derived from accelerometry data for 5-year all-cause mortality in older adults: national health and nutritional examination survey 2003-2006. J Gerontol A Biol Sci Med Sci 2020 Sep 16;75(9):1779-1785. [doi: 10.1093/gerona/glz193] [Medline: 31504213]
22. Khurshid S, Al-Alusi MA, Churchill TW, Guseh JS, Ellinor PT. Accelerometer-derived "weekend warrior" physical activity and incident cardiovascular disease. JAMA 2023 Jul 18;330(3):247-252. [doi: 10.1001/jama.2023.10875] [Medline: 37462704]
23. Xiang B, Zhou Y, Wu X, Zhou X. Association of device-measured physical activity with cardiovascular outcomes in individuals with hypertension. Hypertension 2023 Nov;80(11):2455-2463. [doi: 10.1161/HYPERTENSIONAHA.123.21663] [Medline: 37667966]
24. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420-428. [doi: 10.1037/0033-2909.86.2.420]
25. Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa. Psychol Bull 1969 Nov;72(5):323-327. [doi: 10.1037/h0028106]
26. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med 1997 Feb 28;16(4):385-395. [doi: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3] [Medline: 9044528]
27. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. J Stat Softw 2011 Mar;39(5):1-13. [doi: 10.18637/jss.v039.i05] [Medline: 27065756]
28. Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. Brit J Math Statis 1992 Nov;45(2):265-282. [doi: 10.1111/j.2044-8317.1992.tb00992.x]
29. Hung H, Chiang CT. Estimation methods for time‐dependent AUC models with survival data. Can J Statistics 2010 Mar;38(1):8-26. [doi: 10.1002/cjs.10046]
30. Gerds TA, Schumacher M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. Biom J 2006 Dec;48(6):1029-1040. [doi: 10.1002/bimj.200610301] [Medline: 17240660]
31. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. J Comput Graph Stat 2006 Sep;15(3):651-674. [doi: 10.1198/106186006X133933]
32. Anderson CL, Feldman DB. Hope and physical exercise: the contributions of hope, self-efficacy, and optimism in accounting for variance in exercise frequency. Psychol Rep 2020 Aug;123(4):1145-1159. [doi: 10.1177/0033294119851798] [Medline: 31142190]
33. Li Q, Jiang J, Duan A, Hu J, Li L, Chen W. Physical activity experience of patients with hypertension: a systematic review and synthesis of qualitative literature. BMC Public Health 2024 Oct 15;24(1):2826. [doi: 10.1186/s12889-024-20326-x]
34. Bothwell LE, Greene JA, Podolsky SH, Jones DS. Assessing the gold standard--lessons from the history of RCTs. N Engl J Med 2016 Jun 2;374(22):2175-2181. [doi: 10.1056/NEJMms1604593] [Medline: 27248626]
35. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000 Jun 22;342(25):1887-1892. [doi: 10.1056/NEJM200006223422507] [Medline: 10861325]
36. Lauder L, Mahfoud F, Azizi M, et al. Hypertension management in patients with cardiovascular comorbidities. Eur Heart J 2023 Jun 20;44(23):2066-2077. [doi: 10.1093/eurheartj/ehac395] [Medline: 36342266]
37. Jia G, Sowers JR, Whaley-Connell AT. Obesity in hypertension: the role of the expanding waistline over the years and insights into the future. Hypertension 2024 Apr;81(4):687-690. [doi: 10.1161/HYPERTENSIONAHA.123.21719] [Medline: 38018438]
38. Pacholko A, Iadecola C. Hypertension, neurodegeneration, and cognitive decline. Hypertension 2024 May;81(5):991-1007. [doi: 10.1161/HYPERTENSIONAHA.123.21356] [Medline: 38426329]
39. Kokkinos P, Faselis C, Pittaras A, et al. Stroke incidence in patients with hypertension according to cardiorespiratory fitness. Hypertension 2024 Aug;81(8):1747-1757. [doi: 10.1161/HYPERTENSIONAHA.124.23066] [Medline: 38841839]
40. Shi H, Hu FB, Huang T, et al. Sedentary behaviors, light-intensity physical activity, and healthy aging. JAMA Netw Open 2024 Jun 3;7(6):e2416300. [doi: 10.1001/jamanetworkopen.2024.16300] [Medline: 38861256]
41. Ho FK, Zhou Z, Petermann-Rocha F, et al. Association between device-measured physical activity and incident heart failure: a prospective cohort study of 94 739 UK Biobank participants. Circulation 2022 Sep 20;146(12):883-891. [doi: 10.1161/CIRCULATIONAHA.122.059663] [Medline: 36036153]
42. Keating CJ, Montilla J, Román P, Del Castillo RM. Comparison of high-intensity interval training to moderate-intensity continuous training in older adults: a systematic review. J Aging Phys Act 2020 Oct 1;28(5):798-807. [doi: 10.1123/japa.2019-0111] [Medline: 32303000]
43. Basavarajaiah S, Boraita A, Whyte G, et al. Ethnic differences in left ventricular remodeling in highly-trained athletes relevance to differentiating physiologic left ventricular hypertrophy from hypertrophic cardiomyopathy. J Am Coll Cardiol 2008 Jun 10;51(23):2256-2262. [doi: 10.1016/j.jacc.2007.12.061] [Medline: 18534273]

XSL•FO
RenderX

44. Piepoli MF, Hoes AW, Agewall S, et al. 2016 European Guidelines on cardiovascular disease prevention in clinical practice. Atherosclerosis 2016 Sep;252(29):207-274. [doi: 10.1016/j.atherosclerosis.2016.05.037]
45. Colberg SR, Sigal RJ, Yardley JE, et al. Physical activity/exercise and diabetes: a position statement of the American diabetes association. Diabetes Care 2016 Nov;39(11):2065-2079. [doi: 10.2337/dc16-1728] [Medline: 27926890]
46. Fini NA, Bernhardt J, Said CM, Billinger SA. How to address physical activity participation after stroke in research and clinical practice. Stroke 2021 Jun;52(6):e274-e277. [doi: 10.1161/STROKEAHA.121.034557] [Medline: 33951930]
47. Billinger SA, Arena R, Bernhardt J, et al. Physical activity and exercise recommendations for stroke survivors: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 2014 Aug;45(8):2532-2553. [doi: 10.1161/STR.0000000000000022] [Medline: 24846875]
48. Tegegne TK, Shifti DM, Rawstorn JC, et al. Digital lifestyle interventions for cardiovascular risk reduction: a systematic review and network meta-analysis. Health Policy Technol 2024 Aug;13(3):100879. [doi: 10.1016/j.hlpt.2024.100879]
49. Nnamoko N, Cabrera-Diego LA, Campbell D, Sanders G, Fairclough SJ, Korkontzelos I. Personalised accelerometer cut-point prediction for older adults' movement behaviours using a machine learning approach. Comput Methods Programs Biomed 2021 Sep;208:106165. [doi: 10.1016/j.cmpb.2021.106165] [Medline: 34118492]
50. Sezgin E. Artificial intelligence in healthcare: complementing, not replacing, doctors and healthcare providers. Digital health 2023;9:20552076231186520. [doi: 10.1177/20552076231186520] [Medline: 37426593]
51. Lines RLJ, Ntoumanis N, Thøgersen-Ntoumani C, et al. Cross-sectional and longitudinal comparisons of self-reported and device-assessed physical activity and sedentary behaviour. J Sci Med Sport 2020 Sep;23(9):831-835. [doi: 10.1016/j.jsams.2020.03.004] [Medline: 32312612]
52. UK Biobank. URL: https://www.ukbiobank.ac.uk [accessed 2025-12-22]
53. National Health and Nutrition Examination Survey. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/nhanes [accessed 2025-12-22]

Abbreviations

AUC: area under the receiver operating characteristic curve
BP: blood pressure
CIT: conditional inference tree
CVD: cardiovascular disease
DBP: diastolic blood pressure
HbA_{1c}: glycated hemoglobin
HR: hazard ratio
LASSO: least absolute shrinkage and selection operator
LPA: light physical activity
MI: myocardial infarction
ML: machine learning
MVPA: moderate to vigorous physical activity
NHANES: National Health and Nutrition Examination Survey
PA: physical activity
ROC: receiver operating characteristic
SBP: systolic blood pressure
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
UKB: UK Biobank
WW: weekend warrior

Edited by A Coristine, TDA Cardoso; submitted 04.Jun.2025; peer-reviewed by L Dan, SM Lai; revised version received 02.Dec.2025; accepted 03.Dec.2025; published 09.Jan.2026.

Please cite as:
Yang Y, Chen M, Hu W, Fu Y, Li X, Liao Z, Feng H, Zhao Y, Pei L, Mi B, Chen F
Physical Activity Recommendations Tailored by a Predictive Model for Adults With High Blood Pressure: Observational Study
J Med Internet Res 2026;28:e78492
URL: https://www.jmir.org/2026/1/e78492
doi:10.2196/78492

©Yuhui Yang, Manqing Chen, Weiwei Hu, Yifan Fu, Xingyan Li, Zhenli Liao, Hongman Feng, Yaling Zhao, Leilei Pei, Baibing Mi, Fangyao Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Jan.2026. This is

XSL•FO
RenderX

an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Characterization of Models for Identifying Physical and Cognitive Frailty in Older Adults With Diabetes: Systematic Review and Meta-Analysis

Xia Wang[1*], MS; Shujie Meng[1*], MS; Xiang Xiao[2], BS; Liu Lu[3]; Hongyan Chen[1], MS; Yong Li[4], MEd; Rong Zhang[4]; Qiwu Jiang[5], BM; Shan Liu[6], MS; Ru Gao[6], PhD

[1]School of Basic Medical Sciences and School of Nursing, Chengdu University, Chengdu, China

[2]School of Nursing, Yibin Vocational College of Medicine and Health, Yibin, China

[3]Nursing Department, The Fourth People's Hospital of Yibin, Yibin, China

[4]Rehabilitation College, Sichuan Health Rehabilitation Vocational College, Zigong, China

[5]School of Rehabilitation and Wellness, Yibin Vocational College of Medicine and Health, Yibin, China

[6]Nursing Department, Wenjiang District People's Hospital, No.86 Kangtai Road Wenjiang District, Chengdu, Sichuan Province, China

[*]these authors contributed equally

Corresponding Author:
Ru Gao, PhD
Nursing Department, Wenjiang District People's Hospital, No.86 Kangtai Road Wenjiang District, Chengdu, Sichuan Province, China

Related Article:

Correction of: https://www.jmir.org/2026/1/e84617

(J Med Internet Res 2026;28:e92717) doi:10.2196/92717

In “Characterization of Models for Identifying Physical and Cognitive Frailty in Older Adults With Diabetes: Systematic Review and Meta-Analysis” [1], the authors noted one correction.

Affiliation 5 has been changed from the following:

Medical and Nursing College, Yibin Vocational College of Medicine and Health, Yibin, China

The affiliation now reads:

School of Rehabilitation and Wellness, Yibin Vocational College of Medicine and Health, Yibin, China

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Wang X, Meng S, Xiao X, et al. Characterization of models for identifying physical and cognitive frailty in older adults with diabetes: systematic review and meta-analysis. J Med Internet Res 2026 Jan 29;28:e84617. [doi: 10.2196/84617] [Medline: 41610414]

Submitted 02.Feb.2026; this is a non–peer-reviewed article;accepted 02.Feb.2026; published 12.Feb.2026.

Please cite as:
Wang X, Meng S, Xiao X, Lu L, Chen H, Li Y, Zhang R, Jiang Q, Liu S, Gao R
Correction: Characterization of Models for Identifying Physical and Cognitive Frailty in Older Adults With Diabetes: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e92717
URL: https://www.jmir.org/2026/1/e92717
doi:10.2196/92717

XSL•FO
RenderX

© Xia Wang, Shujie Meng, Xiang Xiao, Liu Lu, Hongyan Chen, Yong Li, Rong Zhang, Qiwu Jiang, Shan Liu, Ru Gao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: User Profiles and Engagement in a Hypertension Self-Management App: Cross-Sectional Survey

Felix Muehlensiepen[1,2,3,4], MPH, PhD, Dr rer medic habil; Susann May[1,2,3], MPH, Dr rer medic; Frances Seifert[1,2,3], MSc; Eileen Wengemuth[1], Dr rer nat; Olen Johannsen[5], MSc; Martin Middeke[6], Prof, Dr; Martin Heinze[1,7], Prof, Dr; Pascal Petit[4], PhD; Nicolas Vuillerme[4], Prof, Dr; Sebastian Spethmann[2,3*], Prof, Dr; Dunja Bruch[8*], MSc

[1]Center for Health Services Research, Faculty of Health Sciences Brandenburg, Brandenburg Medical School, Rüdersdorf, Seebad 82/83, Ruedersdorf bei Berlin, Germany

[2]Department of Cardiology, Angiology and Intensive Care Medicine, Deutsches Herzzentrum der Charité, Berlin, Germany

[3]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[4]Université Grenoble Alpes, CNRS, Grenoble INP, LIG Sangria, Grenoble, France

[5]Hypertension Care UG, Ingolstadt, Germany

[6]Hypertension Center Munich, Munich, Germany

[7]Center for Mental Health, Brandenburg Medical School, Immanuel Klinik Rüdersdorf, Ruedersdorf bei Berlin, Germany

[8]Department of Cardiovascular Surgery, Heart Center Brandenburg, Faculty of Health Sciences, Brandenburg Medical School Theodor Fontane, Bernau, Germany

[*]these authors contributed equally

Corresponding Author:
Felix Muehlensiepen, MPH, PhD, Dr rer medic habil
Center for Health Services Research, Faculty of Health Sciences Brandenburg, Brandenburg Medical School, Rüdersdorf, Seebad 82/83, Ruedersdorf bei Berlin, Germany

Related Article:

Correction of: https://www.jmir.org/2026/1/e83075/

(J Med Internet Res 2026;28:e93515) doi:10.2196/93515

In "User Profiles and Engagement in a Hypertension Self-Management App: Cross-Sectional Survey" [1], the authors noted an error in the order of affiliations 1 to 6, which was introduced during the production process.

The affiliation order was revised from the following:

[1]Hypertension Center Munich, Munich, Germany

[2]Center for Health Services Research, Faculty of Health Sciences Brandenburg, Brandenburg Medical School, Rüdersdorf, Ruedersdorf bei Berlin, Germany

[3]Department of Cardiology, Angiology and Intensive Care Medicine, Deutsches Herzzentrum der Charité, Berlin, Germany

[4]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[5]Université Grenoble Alpes, CNRS, Grenoble INP, LIG Sangria, Grenoble, France

[6]Hypertension Care UG, Ingolstadt, Germany

The order of the affiliations now reads:

[1]Center for Health Services Research, Faculty of Health Sciences Brandenburg, Brandenburg Medical School, Rüdersdorf, Ruedersdorf bei Berlin, Germany

[2]Department of Cardiology, Angiology and Intensive Care Medicine, Deutsches Herzzentrum der Charité, Berlin, Germany

[3]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[4]Université Grenoble Alpes, CNRS, Grenoble INP, LIG Sangria, Grenoble, France

[5]Hypertension Care UG, Ingolstadt, Germany

[6]Hypertension Center Munich, Munich, Germany

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

XSL•FO
RenderX

1. Muehlensiepen F, May S, Seifert F, et al. User profiles and engagement in a hypertension self-management app: cross-sectional survey. J Med Internet Res 2026 Feb 11;28:e83075. [doi: 10.2196/83075] [Medline: 41671553]

Submitted 13.Feb.2026; this is a non–peer-reviewed article;accepted 13.Feb.2026; published 24.Feb.2026.

Please cite as:
Muehlensiepen F, May S, Seifert F, Wengemuth E, Johannsen O, Middeke M, Heinze M, Petit P, Vuillerme N, Spethmann S, Bruch D
Correction: User Profiles and Engagement in a Hypertension Self-Management App: Cross-Sectional Survey
J Med Internet Res 2026;28:e93515
URL: https://www.jmir.org/2026/1/e93515
doi:10.2196/93515

© Felix Muehlensiepen, Susann May, Frances Seifert, Eileen Wengemuth, Olen Johannsen, Martin Middeke, Martin Heinze, Pascal Petit, Nicolas Vuillerme, Sebastian Spethmann, Dunja Bruch. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Leveraging Social Media to Achieve Population-Level Reach of Lung Cancer Screening-Eligible Individuals: A RE-AIM Framework Perspective

Lisa Carter-Bawa[1,2], MSN, MPH, PhD; Jamie S Ostroff[3], PhD; Susan M Rawl[4], PhD; Erin A Hirsch[1], PhD; Smita C Banerjee[3], PhD; Andrew Ciupek[5], PhD; Robert Skipworth Comer[6], MS; Minal Kale[7], MPH, MD; Katherine T Leopold[8], MD; Patrick O Monahan[9], PhD; James E Slaven Jr[9], MS; Francis Valenzona[1], BS; Renda Soylemez Wiener[10,11,12], MPH, MD; Ana Guadalupe Vielma[1], PhD

[1]Cancer Prevention Precision Control Institute, Center for Discovery & Innovation, Hackensack Meridian Health, 123 Metro Blvd, 6th Floor, 6400 Pod, Nutley, NJ, United States

[10]Center for Health Optimization & Implementation Research, VA Boston Healthcare System, Boston, MA, United States

[11]National Center for Lung Cancer Screening, Veterans Health Administration, Washington, DC, United States

[12]The Pulmonary Center, Boston University Chobanian & Avedesian School of Medicine, Boston, MA, United States

[2]Cancer Prevention & Control Program, Georgetown Lombardi Comprehensive Cancer Center, Washington, DC, United States

[3]Department of Psychiatry & Behavioral Science, Memorial Sloan Kettering Cancer Center, New York, NY, United States

[4]School of Nursing, Indiana University, Indianapolis, United States

[5]Go2 for Lung Cancer, Washington, United States

[6]School of Informatics, Indiana University, Indianapolis, United States

[7]Department of Medicine, Internal Medicine, Icahn School of Medicine at Mount Sinai, New York, United States

[8]Hackensack Meridian School of Medicine, Hackensack Meridian Health, Nutley, United States

[9]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, United States

Corresponding Author:
Lisa Carter-Bawa, MSN, MPH, PhD
Cancer Prevention Precision Control Institute, Center for Discovery & Innovation, Hackensack Meridian Health, 123 Metro Blvd, 6th Floor, 6400 Pod, Nutley, NJ, United States

Related Article:

Correction of: https://www.jmir.org/2026/1/e80281

Abstract

(J Med Internet Res 2026;28:e94664) doi:10.2196/94664

In “Leveraging Social Media to Achieve Population-Level Reach of Lung Cancer Screening-Eligible Individuals: A RE-AIM Framework Perspective” [1], the authors noted one error.

The title previously appeared as:

> *Leveraging Social Media to Achieve Population-Level Reach of Lung Cancer Screening–Eligible Individuals: RE-AIM Framework–Based Secondary Analysis of a Randomized Controlled Trial*

The title now reads:

> *Leveraging Social Media to Achieve Population-Level Reach of Lung Cancer Screening-Eligible Individuals: A RE-AIM Framework Perspective*

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Carter-Bawa L, Ostroff JS, Rawl SM, et al. Leveraging social media to achieve population-level reach of lung cancer screening-eligible individuals: a RE-AIM framework perspective. J Med Internet Res 2026;28:e80281. [doi: 10.2196/80281]

XSL•FO
RenderX

Submitted 04.Mar.2026; this is a non–peer-reviewed article;accepted 04.Mar.2026; published 05.Mar.2026.

Please cite as:
Carter-Bawa L, Ostroff JS, Rawl SM, Hirsch EA, Banerjee SC, Ciupek A, Comer RS, Kale M, Leopold KT, Monahan PO, Slaven Jr JE, Valenzona F, Wiener RS, Vielma AG
Correction: Leveraging Social Media to Achieve Population-Level Reach of Lung Cancer Screening-Eligible Individuals: A RE-AIM Framework Perspective
J Med Internet Res 2026;28:e94664
URL: https://www.jmir.org/2026/1/e94664
doi:10.2196/94664

© Lisa Carter-Bawa, Jamie S Ostroff, Susan M Rawl, Erin A Hirsch, Smita C Banerjee, Andrew Ciupek, Robert Skipworth Comer, Minal Kale, Katherine T Leopold, Patrick O Monahan, James E Slaven Jr, Francis Valenzona, Renda Soylemez Wiener, Ana Guadalupe Vielma. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Integrating Text and Image Analysis: Exploring GPT-4V's Capabilities in Advanced Radiological Applications Across Subspecialties

Felix Busch[1], MD; Tianyu Han[2], PhD; Marcus R Makowski[3], MD, PhD; Daniel Truhn[2], MSc, MD; Keno K Bressem[4*], MD; Lisa Adams[3*], MD

[1]Department of Neuroradiology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin, Berlin, Germany

[2]Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Aachen, Germany

[3]Department of Diagnostic and Interventional Radiology, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

[4]Institute for Radiology and Nuclear Medicine, German Heart Center Munich, Technical University of Munich, Munich, Germany

[*]these authors contributed equally

Corresponding Author:
Felix Busch, MD
Department of Neuroradiology
Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin
Charitépl. 1
Berlin, 10117
Germany
Phone: 49 3045050
Email: felix.busch@charite.de

Related Article:

Correction of: https://www.jmir.org/2024/1/e54948

(J Med Internet Res 2026;28:e91415) doi:10.2196/91415

In "Integrating Text and Image Analysis: Exploring GPT-4V's Capabilities in Advanced Radiological Applications Across Subspecialties" [1], the authors made two corrections.

The following text was added to the caption of Figure 1:

> *Reproduced with permission from the Radiological Society of North America. Link to the displayed case: https://cases.rsna.org/take-quiz/07c4b917-80fb-43c0-8b3b-59a0d8ceb203 (accessed 14th January 2026).*

In the Data Availability section, the following sentence was revised:

> *The data set was utilized under the terms and conditions provided by the RSNA, which permits the use of these cases for educational and research purposes.*

This section now reads:

> *The cases from the RSNA Case Collection were reproduced with permission from the Radiological Society of North America. These cases were not used for model training, nor were they retained by any tools or systems employed in this study.*

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Busch F, Han T, Makowski M, Truhn D, Bressem K, Adams L. Integrating text and image analysisxploring GPT-4V's capabilities in advanced radiological applications across subspecialties. J Med Internet Res 2024;26:54948. [doi: 10.2196/54948]

XSL•FO
RenderX

Submitted 14.Jan.2026; this is a non–peer-reviewed article;accepted 27.Jan.2026; published 11.Feb.2026.

Please cite as:
Busch F, Han T, Makowski MR, Truhn D, Bressem KK, Adams L
Correction: Integrating Text and Image Analysis: Exploring GPT-4V's Capabilities in Advanced Radiological Applications Across Subspecialties
J Med Internet Res 2026;28:e91415
URL: https://www.jmir.org/2026/1/e91415
doi:10.2196/91415
PMID:

©Felix Busch, Tianyu Han, Marcus R Makowski, Daniel Truhn, Keno K Bressem, Lisa Adams. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Evidence for Digital Health Tools Designed to Support the Triage of Musculoskeletal Conditions in Primary, Urgent, and Emergency Care Settings: Scoping Review

Linda K Truong[1,2], PT, PhD; James G Wrightson[3], PhD; Raphaël Vincent[4,5,6], PT, MSc; Eunice Lui[1], MPH; Jamon L Couch[7,8], PT; Ellen Wang[2,8], MSc; Cobie Starcevich[9,10], PT; Dean Giustini[11], PhD; Alex Haagaard[12], MA; Elena Lopatina[13,14], MD, PhD; Niels van Berkel[15], PhD; Michael Skovdal Rathleff[16,17], PT, PhD; Clare L Ardern[1,2,7], PT, PhD

[1]Centre for Aging SMART, University of British Columbia, Vancouver, BC, Canada

[10]Physiotherapy Department, Rockingham General Hospital, South Metropolitan Health Service, Perth, Australia

[11]Biomedical Branch Library, University of British Columbia, Vancouver, BC, Canada

[12]Pain BC, Vancouver, BC, Canada

[13]University of Calgary, Calgary, AB, Canada

[14]Primary Care Alberta, Calgary, AB, Canada

[15]Department of Computer Science, Aalborg University, Aalborg, Denmark

[16]Department of Health Science and Technology, Faculty of Medicine, Aalborg University, Aalborg, Denmark

[17]Center for General Practice at Aalborg University, Aalborg, Denmark

[2]Department of Physical Therapy, Faculty of Medicine, University of British Columbia, 13737 96 Avenue, Surrey, BC, Canada

[3]Department of Family Practice, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

[4]School of Rehabilitation, Faculty of Medicine, Université de Montréal, Montreal, QC, Canada

[5]Hôpital Maisonneuve-Rosemont Research Center, Université de Montréal Affiliated Research Center, Montreal, QC, Canada

[6]Centre for Interdisciplinary Research in Rehabilitation of Greater Montreal, Institut Universitaire sur la Réadaptation en Déficience Physique de Montréal, Montreal, QC, Canada

[7]La Trobe Sport and Exercise Medicine Research Centre, La Trobe University, Melbourne, Australia

[8]Arthritis Research Canada, Vancouver, BC, Canada

[9]School of Allied Health, Faculty of Health Sciences, Curtin University, Perth, Australia

Corresponding Author:
Linda K Truong, PT, PhD
Centre for Aging SMART, University of British Columbia, Vancouver, BC, Canada

Related Article:

Correction of: https://www.jmir.org/2026/1/e81578

(J Med Internet Res 2026;28:e92722) doi:10.2196/92722

In "Evidence for Digital Health Tools Designed to Support the Triage of Musculoskeletal Conditions in Primary, Urgent, and Emergency Care Settings: Scoping Review" [1], the authors made two corrections.

In the abstract, the list of databases searched has been amended to remove PsycINFO and Cochrane Library, and these were replaced with Central (OVID) and Compendex. This has also been further specified in the Methods section.

In the abstract, the following sentence was revised:

> *Systematic searches in MEDLINE (OVID), CINAHL (EBSCO), PsycINFO (EBSCO), Embase (OVID), Cochrane Library, Web of Science, OpenGrey, Google Scholar, arXiv, medRxiv, and an extensive gray literature search were conducted with a librarian scientist from inception to September 18, 2025.*

This sentence now reads:

> *Systematic searches in MEDLINE (OVID), Embase (OVID), CENTRAL (OVID), CINAHL (EBSCO), Compendex, Web of Science, OpenGrey, Google Scholar, arXiv, medRxiv, and an extensive gray literature search were conducted with a librarian scientist from inception to September 18, 2025.*

In the Methods section, the following sentence was revised:

> *An electronic search was conducted in 6 databases (MEDLINE [OVID], CINAHL [EBSCO], PsycINFO [EBSCO], Embase [OVID], the Cochrane Library, and Web of Science) and 4 gray literature sites (OpenGrey, GoogleScholar, arXiv, and medRxiv) with the aid of a biomedical librarian and information specialist.*

The sentence now reads:

XSL•FO
RenderX

An electronic search was conducted in 6 databases (MEDLINE [OVID], Embase [OVID], CENTRAL [OVID], CINAHL [EBSCO], Compendex, and Web of Science) and 4 gray literature sites (OpenGrey, GoogleScholar, arXiv, and medRxiv) with the aid of a biomedical librarian and information specialist.

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Truong LK, Wrightson JG, Vincent R, et al. Evidence for digital health tools designed to support the triage of musculoskeletal conditions in primary, urgent, and emergency care settings: scoping review. J Med Internet Res 2026 Jan 14;28:e81578. [doi: 10.2196/81578] [Medline: 41534041]

Submitted 02.Feb.2026; this is a non–peer-reviewed article;accepted 11.Feb.2026; published 26.Feb.2026.

Please cite as:
Truong LK, Wrightson JG, Vincent R, Lui E, Couch JL, Wang E, Starcevich C, Giustini D, Haagaard A, Lopatina E, van Berkel N, Rathleff MS, Ardern CL
Correction: Evidence for Digital Health Tools Designed to Support the Triage of Musculoskeletal Conditions in Primary, Urgent, and Emergency Care Settings: Scoping Review
J Med Internet Res 2026;28:e92722
URL: https://www.jmir.org/2026/1/e92722
doi:10.2196/92722

© Linda K Truong, James G Wrightson, Raphaël Vincent, Eunice Lui, Jamon L Couch, Ellen Wang, Cobie Starcevich, Dean Giustini, Alex Haagaard, Elena Lopatina, Niels van Berkel, Michael Skovdal Rathleff, Clare L Ardern. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Acceptability of Health Information Technology by Health Care Professionals: Where We Are Now and How We Can Fill the Gap

Corinne Isnard-Bagnis[1,2], MD, PhD; Stéphane Mouchabac[3,4], MD; Riadh Lebib[5], PhD; Hervé Bismut[6], PhD; Pierre Alexis Geoffroy[7,8,9], MD, PhD

[1]Nephrology Department, Pitié-Salpêtrière Hospital, AP-HP, Sorbonne Université, 47-83 boulevard de l'Hôpital, Paris, France

[2]Academic Digital Medical Hub, AP-HP, Paris, France

[3]Department of Psychiatry, Saint-Antoine Hospital, AP-HP, Sorbonne Université, Paris, France

[4]Infrastructure for Clinical Research in Neurosciences (iCRIN), Paris Brain Institute, Paris, France

[5]Humans Matter, Paris, France

[6]Geminicis, Paris, France

[7]Département de psychiatrie et d'addictologie, AP-HP, GHU Paris Nord, DMU Neurosciences, Hôpital Bichat-Claude-Bernard, Paris, France

[8]Centre ChronoS, GHU Paris psychiatrie & neurosciences, Paris, France

[9]Université Paris Cité, Inserm, NeuroDiderot, Paris, France

Corresponding Author:
Corinne Isnard-Bagnis, MD, PhD
Nephrology Department, Pitié-Salpêtrière Hospital, AP-HP, Sorbonne Université, 47-83 boulevard de l'Hôpital, Paris, France

Related Article:

Correction of: https://www.jmir.org/2025/1/e72184

(J Med Internet Res 2026;28:e89383) doi:10.2196/89383

In “Acceptability of Health Information Technology by Health Care Professionals: Where We Are Now and How We Can Fill the Gap” [1], the authors noted incomplete affiliations for authors CIB, SM, and PAG.

The previous list of affiliations was as follows:

> *Corinne Isnard Bagnis, MD, PhD[1], Stéphane Mouchabac, MD[2], Riadh Lebib, PhD[3], Hervé Bismut, PhD[4], Pierre A Geoffroy, MD, PhD[5]*
>
> *[1]Department of Nephrology, Pitié-Salpêtrière Hospital, APHP Sorbonne Université, Paris, France*
>
> *[2]Department of Psychiatry, Saint-Antoine Hospital, APHP Sorbonne Université, Paris, France*
>
> *[3]Humans Matter, Paris, France*
>
> *[4]Geminicis, Paris, France*
>
> *[5]Department of Psychiatry, Hôpital Bichat-Claude-Bernard, Paris, France*

This has been revised to:

> *Corinne Isnard Bagnis, MD, PhD[1,2], Stéphane Mouchabac, MD[3,4], Riadh Lebib, PhD[5], Hervé Bismut, PhD[6], Pierre A Geoffroy, MD, PhD[7-9]*
>
> *[1]Nephrology Department, Pitié Salpêtrière Hospital, AP-HP, Sorbonne Université, Paris, France*
>
> *[2]Academic Digital Medical Hub, AP-HP, Paris, France*
>
> *[3]Department of Psychiatry, Saint-Antoine Hospital, AP-HP, Sorbonne Université, Paris, France.*
>
> *[4]Infrastructure for Clinical Research in Neurosciences (iCRIN), Paris Brain Institute, Paris, France*
>
> *[5]Humans Matter, Paris, France*
>
> *[6]Geminicis, Paris, France*
>
> *[7]Département de psychiatrie et d'addictologie, AP-HP, GHU Paris Nord, DMU Neurosciences, Hôpital Bichat - Claude-Bernard, Paris, France*
>
> *[8]Centre ChronoS, GHU Paris psychiatrie & neurosciences, Paris, France*
>
> *[9]Université Paris Cité, Inserm, NeuroDiderot, Paris, France*

These corrections will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because these were made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

XSL•FO
RenderX

Reference

1. Isnard-Bagnis C, Mouchabac S, Lebib R, Bismut H, Geoffroy PA. Acceptability of health information technology by health care professionals: where we are now and how we can fill the gap. J Med Internet Res 2025 Dec 2;27:e72184. [doi: 10.2196/72184] [Medline: 41329915]

Submitted 11.Dec.2025; this is a non–peer-reviewed article;accepted 16.Dec.2025; published 06.Jan.2026.

Please cite as:
Isnard-Bagnis C, Mouchabac S, Lebib R, Bismut H, Geoffroy PA
Correction: Acceptability of Health Information Technology by Health Care Professionals: Where We Are Now and How We Can Fill the Gap
J Med Internet Res 2026;28:e89383
URL: https://www.jmir.org/2026/1/e89383
doi:10.2196/89383

© Corinne Isnard-Bagnis, Stéphane Mouchabac, Riadh Lebib, Hervé Bismut, Pierre A Geoffroy. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Combining Artificial Intelligence and Human Support in Mental Health: Digital Intervention With Comparable Effectiveness to Human-Delivered Care

Clare E Palmer[1], PhD; Emily Marshall[1], PGCert, PGDip; Edward Millgate[1], PhD; Graham Warren[1], PhD; Michael Ewbank[1], PhD; Elisa Cooper[1], PhD; Samantha Lawes[1], PhD; Alastair Smith[1], MEng; Chris Hutchins-Joss[1], MSc; Jessica Young[1], MSc; Malika Bouazzaoui[1], MBA, MSc; Morad Margoum[2], MSc; Sandra Healey[2], PGDip; Louise Marshall[1], PhD; Shaun Mehew[1], PGCert, PGDip; Ronan Cummins[1], PhD; Valentin Tablan[1], PhD; Ana Catarino[1], PhD; Andrew E Welchman[1], PhD; Andrew D Blackwell[1], PhD

[1]ieso Digital Health Ltd, Cambridge, United Kingdom

[2]Dorset HealthCare University NHS Foundation, Poole, United Kingdom

Corresponding Author:
Clare E Palmer, PhD
ieso Digital Health Ltd
The Jeffreys Building
Cowley Road
Cambridge, CB4 0DS
United Kingdom
Phone: 44 0800 074 5560
Email: c.palmer@iesohealth.com

Related Article:

Correction of: https://www.jmir.org/2025/1/e69351

(J Med Internet Res 2026;28:e88640) doi:10.2196/88640

In "Combining Artificial Intelligence and Human Support in Mental Health: Digital Intervention With Comparable Effectiveness to Human-Delivered Care" [1], the authors noted one error.

Author CH-J was previously listed with no academic qualifications. His qualification has been revised to "MSc."

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Palmer CE, Marshall E, Millgate E, Warren G, Ewbank M, Cooper E, et al. Combining Artificial Intelligence and Human Support in Mental Health: Digital Intervention With Comparable Effectiveness to Human-Delivered Care. J Med Internet Res 2025 May 13;27:e69351 [FREE Full text] [doi: 10.2196/69351] [Medline: 40152000]

XSL•FO
RenderX

Submitted 28.Nov.2025; this is a non–peer-reviewed article;accepted 05.Dec.2025; published 06.Jan.2026.

Please cite as:
Palmer CE, Marshall E, Millgate E, Warren G, Ewbank M, Cooper E, Lawes S, Smith A, Hutchins-Joss C, Young J, Bouazzaoui M, Margoum M, Healey S, Marshall L, Mehew S, Cummins R, Tablan V, Catarino A, Welchman AE, Blackwell AD
Correction: Combining Artificial Intelligence and Human Support in Mental Health: Digital Intervention With Comparable Effectiveness to Human-Delivered Care
J Med Internet Res 2026;28:e88640
URL: https://www.jmir.org/2026/1/e88640
doi:10.2196/88640
PMID:

©Clare E Palmer, Emily Marshall, Edward Millgate, Graham Warren, Michael Ewbank, Elisa Cooper, Samantha Lawes, Alastair Smith, Chris Hutchins-Joss, Jessica Young, Malika Bouazzaoui, Morad Margoum, Sandra Healey, Louise Marshall, Shaun Mehew, Ronan Cummins, Valentin Tablan, Ana Catarino, Andrew E Welchman, Andrew D Blackwell. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Effectiveness of a Web-Based Medication Education Course on Pregnant Women's Medication Information Literacy and Decision Self-Efficacy: Randomized Controlled Trial

Suya Li[1*], MSN; Hui-Jun Chen[2*], MD; Jie Zhou[1], MSN; Yi-Bei Zhouchen[1], MNS; Rong Wang[3], MSN; Jinyi Guo[1], MNS; Sharon R Redding[4], EdD; Yan-Qiong Ouyang[1], MD

[1]School of Nursing, Wuhan University, Wuhan, China

[2]Department of Obstetrics, Zhongnan Hospital of Wuhan University, Wuhan, China

[3]Renmin Hospital of Wuhan University, Wuhan, China

[4]Project HOPE, Washington, DC, United States

[*]these authors contributed equally

Corresponding Author:
Yan-Qiong Ouyang, MD
School of Nursing
Wuhan University
115 Donghu Road
Wuchang District
Wuhan, 430071
China
Phone: 86 02768758747
Email: ouyangyq@whu.edu.cn

Related Article:

Correction of: https://www.jmir.org/2025/1/e54148/8

(J Med Internet Res 2026;28:e91835) doi:10.2196/91835

In "Effectiveness of a Web-Based Medication Education Course on Pregnant Women's Medication Information Literacy and Decision Self-Efficacy: Randomized Controlled Trial" [1], the authors noted an error in the affiliations.

The affiliation of authors SL, JZ, YBZ, JG, and YQO was previously shown as:

[1]Wuhan University

This affiliation has been revised to the following:

[1]School of Nursing, Wuhan University

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Li S, Chen HJ, Zhou J, Zhouchen YB, Wang R, Guo J, et al. Effectiveness of a Web-Based Medication Education Course on Pregnant Women's Medication Information Literacy and Decision Self-Efficacy: Randomized Controlled Trial. J Med Internet Res 2025 Jan 22;27:e54148 [FREE Full text] [doi: 10.2196/54148] [Medline: 39841986]

XSL•FO
RenderX

Submitted 20.Jan.2026; this is a non–peer-reviewed article;accepted 21.Jan.2026; published 26.Jan.2026.

Please cite as:
Li S, Chen HJ, Zhou J, Zhouchen YB, Wang R, Guo J, Redding SR, Ouyang YQ
Correction: Effectiveness of a Web-Based Medication Education Course on Pregnant Women's Medication Information Literacy and Decision Self-Efficacy: Randomized Controlled Trial
J Med Internet Res 2026;28:e91835
URL: https://www.jmir.org/2026/1/e91835
doi:10.2196/91835
PMID:

©Suya Li, Hui-Jun Chen, Jie Zhou, Yi-Bei Zhouchen, Rong Wang, Jinyi Guo, Sharon R Redding, Yan-Qiong Ouyang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Culturally Adapted Guided Internet-Based Cognitive Behavioral Therapy for Hong Kong People With Depressive Symptoms: Randomized Controlled Trial

Jia-Yan Pan[1], PhD; Jonas Rafi[2], PhD

[1]Department of Social Work, Hong Kong Baptist University, Hong Kong, China (Hong Kong)

[2]Department of Psychology, Stockholm University, Stockholm, Sweden

Corresponding Author:
Jia-Yan Pan, PhD
Department of Social Work
Hong Kong Baptist University
Room AAB1026, 10/F, Academic and Administration Building
15 Baptist University Road, Baptist University Road Campus, Kowloon Tong, KLN
Hong Kong
China (Hong Kong)
Phone: 852 34116415
Email: jiayan@hkbu.edu.hk

Related Article:

Correction of: https://www.jmir.org/2025/1/e64303

(J Med Internet Res 2026;28:e88495) doi:10.2196/88495

In "Culturally Adapted Guided Internet-Based Cognitive Behavioral Therapy for Hong Kong People With Depressive Symptoms: Randomized Controlled Trial"[1], the authors noted some errors.

One sentence has been changed. In the "Measurements" section, the third sentence in the third paragraph originally appeared as follows:

> *Item scores were summed for the positive and negative subscales, with higher total scores indicating more positive and negative automatic thoughts, respectively.*

This sentence now reads:

> *Item scores were averaged for the positive and negative subscales, with higher total scores indicating more positive and negative automatic thoughts, respectively.*

The originally published Tables 2 and 4 have been revised due to typos. The new Table 2 and Table 4 that are replacing the original published versions are shown below (with changes marked in italics).

XSL•FO
RenderX

Table 2. Comparison of treatment effects between active and WLC[a] groups at post-treatment (intent-to-treat analysis).

Outcome	Unadjusted mean (SD)			Adjusted difference (95% CI)				Between group effect size *d* (95% CI)		Within group effect size (95% CI)		
	Web	App	WL	Web vs WL	*P* value	App vs WL	*P* value	Web vs WL	App vs WL	Web	App	WLC
BDI-II[b]	9.40 (8.90)	9.43 (8.45)	18.02 (9.02)	–8.24 (–10.52 to –5.96)	<.001	–9.38 (–11.65 to –7.10)	<.001	1.07 (0.81 to 1.34)	1.15 (0.88 to 1.43)	1.48 (1.18 to 1.78)	1.64 (1.33 to 1.94)	0.45 (0.2 to 0.71)
PHQ-9[c]	5.59 (4.43)	5.24 (3.65)	8.99 (4.58)	–3.07 (–4.46 to –1.67)	<.001	–4.50 (–5.89 to –3.12)	<.001	0.78 (0.52 to 1.04)	0.95 (0.63 to 1.27)	1.11 (0.81 to 1.41)	1.69 (1.36 to 2.02)	0.46 (0.19 to 0.73)
GHQ-12[d]	1.60 (3.06)	1.53 (2.77)	5.35 (3.85)	–3.65 (–4.75 to –2.54)	<.001	–4.49 (–5.6 to –3.38)	<.001	1.05 (0.77 to 1.32)	1.11 (0.83 to 1.39)	1.64 (1.34 to 1.95)	2.21 (1.87 to 2.55)	0.64 (0.38 to 0.9)
BAI[e]	8.89 (7.89)	8.49 (6.69)	11.09 (7.49)	–2.47 (–4.33 to –0.62)	.009	–3.32 (–5.17 to –1.47)	<.001	0.37 (0.11 to 0.63)	0.54 (0.29 to 0.8)	0.64 (0.37 to 0.91)	0.88 (0.59 to 1.15)	0.37 (0.11 to 0.62)
Positive automatic thoughts	2.40 (0.71)	2.33 (0.67)	1.98 (0.58)	0.38 (0.22 to 0.54)	<.001	0.40 (0.24 to 0.55)	<.001	–0.63 (–0.9 to –0.37)	–0.61 (–0.9 to –0.32)	–0.88 (–1.16 to –0.6)	–0.93 (–1.22 to –0.65)	–0.30 (–0.55 to 0.04)
Negative automatic thoughts	1.85 (0.78)	1.89 (0.63)	2.29 (0.80)	–0.43 (–0.63 to –0.22)	<.001	–0.48 (–0.68 to –0.27)	<.001	0.63 (0.37 to 0.90)	0.66 (0.35 to 0.97)	0.83 (0.55 to 1.11)	1.00 (0.72 to 1.28)	0.28 (0.02 to 0.53)
Positive emotions	3.26 (0.69)	3.27 (0.62)	2.89 (0.68)	0.35 (0.18 to 0.52)	<.001	0.43 (0.26 to 0.60)	<.001	–0.56 (–0.84 to –0.27)	–0.60 (–0.9 to –0.3)	–0.89 (–1.17 to –0.61)	–1.10 (–1.38 to –0.81)	–0.30 (–0.55 to –0.04)
Negative emotions	2.91 (0.82)	2.92 (0.69)	3.33 (0.74)	–0.45 (–0.64 to –0.25)	<.001	–0.47 (–0.66 to –0.27)	<.001	0.64 (0.37 to 0.91)	0.65 (0.38 to 0.91)	1.06 (0.78 to 1.34)	1.21 (0.91 to 1.5)	0.50 (0.24 to 0.75)

[a]WLC: waitlist control.

[b]BDI-II: Beck Depression Inventory-II.

[c]PHQ-9: 9-item Patient Health Questionnaire.

[d]GHQ-12: 12-item General Health Questionnaire.

[e]BAI: Beck Anxiety Inventory.

XSL•FO
RenderX

Table 4. Comparisons between web-based and app-based iCBT[a] groups on primary and secondary outcome measures at post-treatment (intention-to-treat analysis).

Measure and time	Unadjusted mean (SD)		Adjusted difference (95% CI)	*P* value	Between-group effect size *d* (95% CI)	Within-group effect size (95% CI)	
	Web	App				Web	App
BDI-II[b]							
Baseline	22.24 (8.52)	23.66 (8.77)	N/A[c]	N/A	N/A	N/A	N/A
Post	9.40 (8.90)	9.43 (8.45)	0.21 (–2.28 to 2.70)	.87	–0.01 (–0.3 to 0.28)	1.48 (1.18 to 1.78)	1.64 (1.33 to 1.94)
3 months	9.43 (9.58)	9.15 (8.51)	–0.14 (–3.11 to 2.83)	.93	0.03 (–0.34 to 0.4)	1.45 (1.1 to 1.8)	1.67 (1.32 to 2.02)
6 months	10.15 (9.03)	13.47 (12.24)	–1.37 (–4.46 to 1.72)	.38	–0.37 (–0.76 to 0.02)	1.40 (1.05 to 1.74)	1.04 (0.69 to 1.38)
PHQ-9[d]							
Baseline	10.77 (4.9)	12.07 (4.27)	N/A	N/A	N/A	N/A	N/A
Post	5.59 (4.43)	5.24 (3.65)	0.35 (–1.08 to 1.77)	.63	0.07 (–0.22 to 0.36)	1.11 (0.8 to 1.41)	1.69 (1.36 to 2.02)
3 months	5.41 (4.01)	5.83 (3.57)	–0.27 (–1.94 to 1.41)	.75	–0.1 (–0.47 to 0.28)	1.16 (0.81 to 1.51)	1.55 (1.18 to 1.91)
6 months	5.94 (5.44)	7.65 (5.06)	–0.86 (–2.67 to 0.94)	.35	–0.39 (–0.79 to 0.01)	0.95 (0.59 to 1.31)	0.97 (0.62 to 1.33)
GHQ-12[e]							
Baseline	7.46 (3.86)	8.4 (3.31)	N/A	N/A	N/A	N/A	N/A
Post	1.60 (3.06)	1.53 (2.77)	0.05 (–0.84 to 0.93)	.91	0.02 (–0.27 to 0.31)	1.64 (1.34 to 1.95)	2.21 (1.87 to 2.55)
3 months	1.85 (3.02)	1.69 (2.82)	–0.08 (–1.17 to 1)	.88	0.05 (–0.32 to 0.42)	1.54 (1.19 to 1.89)	2.12 (1.74 to 2.49)
6 months	1.77 (3.20)	3.49 (4.10)	–1.46 (–0.33 to –2.59)	.01	–0.52 (–0.91 to –0.12)	1.54 (1.19 to 1.89)	1.39 (1.03 to 1.74)
BAI[f]							
Baseline	14.53 (9.29)	15.36 (8.8)	N/A	N/A	N/A	N/A	N/A
Post	8.89 (7.89)	8.49 (6.69)	0.53 (–1.25 to 2.31)	.56	0.06 (–0.23 to 0.36)	0.64 (0.37 to 0.91)	0.88 (0.59 to 1.15)
3 months	7.67 (6.42)	8.03 (6.78)	0.59 (–1.56 to 2.73)	.59	–0.05 (–0.42 to 0.32)	0.80 (0.47 to 1.12)	0.89 (0.57 to 1.21)
6 months	7.89 (7.99)	9.88 (7.72)	0.32 (–1.91 to 2.56)	.78	–0.25 (–0.64 to 0.14)	0.74 (0.42 to 1.07)	0.64 (0.31 to 0.98)
Positive automatic thoughts							
Baseline	1.85 (0.55)	1.79 (0.52)	N/A	N/A	N/A	N/A	N/A
Post	2.40 (0.71)	2.33 (0.67)	0.01 (–0.18 to 0.2)	.93	0.1 (–0.19 to 0.39)	–0.88 (–1.16 to –0.6)	–0.93 (–1.21 to –0.65)
3 months	2.49 (0.67)	2.44 (0.84)	–0.1 (–0.32 to 0.12)	.38	0.07 (–0.3 to 0.44)	–1.08 (–1.41 to –0.75)	–1.03 (–1.36 to –0.71)
6 months	2.46 (0.85)	2.15 (0.71)	0.08 (–0.16 to 0.31)	.52	0.47 (0.08 to 0.86)	–0.93 (–1.26 to –0.6)	–0.63 (–0.96 to –0.29)
Negative automatic thoughts							
Baseline	2.55 (0.88)	2.65 (0.85)	N/A	N/A	N/A	N/A	N/A
Post	1.85 (0.78)	1.89 (0.63)	–0.02 (–0.22 to 0.18)	.86	–0.04 (–0.33 to 0.25)	0.83 (0.55 to 1.1)	1 (0.72 to 1.28)
3 months	1.74 (0.68)	1.82 (0.71)	–0.03 (–0.27 to 0.21)	.81	–0.09 (–0.46 to 0.28)	0.97 (0.64 to 1.29)	1.02 (0.7 to 1.34)
6 months	1.75 (0.78)	2.05 (0.93)	–0.15 (–0.4 to 0.1)	.24	–0.38 (–0.77 to 0.01)	0.93 (0.6 to 1.26)	0.68 (0.34 to 1.02)
Positive emotions							
Baseline	2.72 (0.54)	2.65 (0.54)	N/A	N/A	N/A	N/A	N/A
Post	3.26 (0.69)	3.27 (0.62)	–0.03 (–0.22 to 0.16)	.87	–0.04 (–0.33 to 0.25)	–0.89 (–1.17 to –0.61)	–1.1 (–1.38 to –0.81)
3 months	3.26 (0.75)	3.33 (0.81)	–0.1 (–0.33 to 0.13)	.93	–0.12 (–0.49 to 0.25)	–0.89 (–1.21 to –0.56)	–1.07 (–1.39 to –0.74)
6 months	3.24 (0.75)	2.97 (0.72)	0.12 (–0.11 to 0.36)	.38	0.4 (0.01 to 0.79)	–0.85 (–1.18 to –0.52)	–0.54 (–0.88 to –0.21)

XSL•FO
RenderX

Measure and time	Unadjusted mean (SD)		Adjusted difference (95% CI)	*P* value	Between-group effect size *d* (95% CI)	Within-group effect size (95% CI)	
	Web	App				Web	App
Negative emotions							
Baseline	3.73 (0.75)	3.78 (0.73)	N/A	N/A	N/A	N/A	N/A
Post	2.91 (0.82)	2.93 (0.69)	–0.02 (–0.23 to 0.2)	.63	–0.02 (–0.31 to 0.28)	1.06 (0.78 to 1.34)	1.21 (0.91 to 1.5)
3 months	2.78 (0.85)	2.93 (0.73)	–0.07 (–0.32 to 0.18)	.75	–0.19 (–0.56 to 0.18)	1.23 (0.89 to 1.57)	1.17 (0.84 to 1.5)
6 months	2.86 (0.91)	3.17 (0.88)	–0.1 (–0.36 to 0.16)	.35	–0.39 (–0.78 to 0)	1.09 (0.75 to 1.42)	0.8 (0.46 to 1.13)

[a]iCBT: internet-based cognitive behavioral therapy

[b]BDI-II: Beck Depression Inventory-II.

[c]N/A: not applicable.

[d]PHQ-9: 9-item Patient Health Questionnaire.

[e]GHQ-12: 12-item General Health Questionnaire.

[f]BAI: Beck Anxiety Inventory.

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Pan JY, Rafi J. Culturally Adapted Guided Internet-Based Cognitive Behavioral Therapy for Hong Kong People With Depressive Symptoms: Randomized Controlled Trial. J Med Internet Res 2025 Feb 25;27:e64303 [FREE Full text] [doi: 10.2196/64303] [Medline: 39998865]

Submitted 26.Nov.2025; this is a non–peer-reviewed article;accepted 27.Nov.2025; published 03.Feb.2026.

Please cite as:
Pan JY, Rafi J
Correction: Culturally Adapted Guided Internet-Based Cognitive Behavioral Therapy for Hong Kong People With Depressive Symptoms: Randomized Controlled Trial
J Med Internet Res 2026;28:e88495
URL: https://www.jmir.org/2026/1/e88495
doi:10.2196/88495
PMID:

©Jia-Yan Pan, Jonas Rafi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 03.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Ingroup Favoritism Surrounding COVID-19 Vaccinations in the Hispanic Communities: Experimental Study

Juwon Hwang[1,2], PhD; Asya Cooley[3], PhD; Skye Cooley[3], PhD; Robert Hinck[4], PhD

[1]Department of Mass Communication, Advertising and Public Relations, College of Communication, Boston University, Boston, MA, United States
[2]Core Faculty, Center on Emerging Infectious Diseases, Boston University, Boston, MA, United States
[3]School of Media and Strategic Communications, Oklahoma State University, Stillwater, United States
[4]Leadership and Innovation Institute, Air University, Montgomery, United States

Corresponding Author:
Juwon Hwang, PhD
Department of Mass Communication, Advertising and Public Relations
College of Communication
Boston University
640 Commonwealth Ave,
Boston, MA, 02215
United States
Phone: 1 617 353 3450
Email: juhwang@bu.edu

Related Article:

Correction of: https://www.jmir.org/2025/1/e71188

(J Med Internet Res 2026;28:e90652) doi:10.2196/90652

In “Ingroup Favoritism Surrounding COVID-19 Vaccinations in the Hispanic Communities: Experimental Study” [1], the authors identified errors related to references and corresponding in-text citations.

Reference 43 was invalid and has been removed from the reference list, and the corresponding in-text citation has been removed. The affected statement is now supported by the following reference:

> *Daggett A, Abdollahi S, Hashemzadeh M. The effect of language concordance on health care relationship trust score. Cureus. May 26, 2023;15(5):e39530 [doi: 10.7759/cureus.39530] [Medline: 37366455]*

Reference 53 was invalid and has been removed from the reference list, and the corresponding in-text citation has been removed. The affected statement is now supported by the following reference:

> *Chandler J, Rosenzweig C, Moss AJ, Robinson J, Litman L. Online panels in social science research: expanding sampling methods beyond Mechanical Turk. Behav Res Methods. Oct 2019;51(5):2022-2038. [doi: 10.3758/s13428-019-01273-7] [Medline: 31512174]*

Reference 54 was invalid and has been removed from the reference list, and the corresponding in-text citation has been removed.

References 55 and 56 were invalid entries of the same reference and have been removed from the reference list, and the corresponding in-text citations have been removed. The affected statement is now supported by the following reference:

> *Khan H, Lee R, Lockshin L. Do ethnic cues improve advertising effectiveness for ethnic consumers? Australasian Marketing J. Aug 2015;23(3):218-226. [doi: 10.1016/j.ausmj.2015.07.001]*

Reference 44 was incorrectly attributed and has been corrected to the following reference:

> *Haslam C, Jetten J, Cruwys T, Dingle G, Haslam SA. The New Psychology of Health: Unlocking the Social Cure. Routledge; 2018. [doi: 10.4324/9781315648569]*

Reference 58 was incorrectly attributed and has been corrected to the following reference:

> *Ipeirotis P. Demographics of Mechanical Turk. New York University; Mar 2010. URL: https://ipeirotis.org/wp-content/uploads/2012/02/CeDER-10-01.pdf [Accessed 2026-02-20]*

The correction will appear in the online version of the paper on the JMIR Publications website together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

XSL•FO
RenderX

Reference

1. Hwang J, Cooley A, Cooley S, Hinck R. Ingroup Favoritism Surrounding COVID-19 Vaccinations in the Hispanic Communities: Experimental Study. J Med Internet Res 2025 May 27;27:e71188 [FREE Full text] [doi: 10.2196/71188] [Medline: 40424033]

Submitted 31.Dec.2025; this is a non–peer-reviewed article;accepted 13.Feb.2026; published 06.Mar.2026.

Please cite as:
Hwang J, Cooley A, Cooley S, Hinck R
Correction: Ingroup Favoritism Surrounding COVID-19 Vaccinations in the Hispanic Communities: Experimental Study
J Med Internet Res 2026;28:e90652
URL: https://www.jmir.org/2026/1/e90652
doi:10.2196/90652
PMID:

©Juwon Hwang, Asya Cooley, Skye Cooley, Robert Hinck. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: A Supervised Explainable Machine Learning Model for Perioperative Neurocognitive Disorder in Liver-Transplantation Patients and External Validation on the Medical Information Mart for Intensive Care IV Database: Retrospective Study

Zhendong Ding[1], MD; Linan Zhang[1], MD; Yihan Zhang[1], MD; Jing Yang[1], MD, PhD; Yuheng Luo[2], MD; Mian Ge[1], MD, PhD; Weifeng Yao[1], MD, PhD; Ziqing Hei[1], MD, PhD; Chaojin Chen[1], MD, PhD

[1]Department of Anesthesiology, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, Guangdong, China

[2]Guangzhou AI & Data Cloud Technology Co., LTD, Guangzhou, China

Corresponding Author:
Chaojin Chen, MD, PhD
Department of Anesthesiology
The Third Affiliated Hospital of Sun Yat-sen University
Guangzhou, Guangdong
China
Phone: 86 13430322182
Email: chenchj28@mail.sysu.edu.cn

Related Article:

Correction of: https://www.jmir.org/2025/1/e55046

(J Med Internet Res 2026;28:e88952) doi:10.2196/88952

In "A Supervised Explainable Machine Learning Model for Perioperative Neurocognitive Disorder in Liver-Transplantation Patients and External Validation on the Medical Information Mart for Intensive Care IV Database: Retrospective Study" [1], the authors made three corrections.

The footnotes of Table 1 originally read as follows:

[a]*PND: perioperative neurocognitive dysfunction.*

[b]*DBD: donation after brain death.*

[c]*DCD: donation after circulatory death.>*

[d]*DBCD: donation after brain death followed by circulatory death.*

The footnotes now read:

[a]*Data presented in Table 1 are based on the original dataset prior to data imputation.*

[b]*PND: perioperative neurocognitive dysfunction.*

[c]*DBD: donation after brain death.*

[d]*DCD: donation after circulatory death.*

[e]*DBCD: donation after brain death followed by circulatory death.*

Below is the new Figure 4 (presented here as Figure 1), which will replace the originally published figure.

XSL•FO
RenderX

Figure 1. Post–liver transplantation survival associated with perioperative neurocognitive dysfunction. Patients with post–liver transplantation perioperative neurocognitive dysfunction showed a significantly lower survival rate. LT: liver transplantation; PND: perioperative neurocognitive dysfunction.

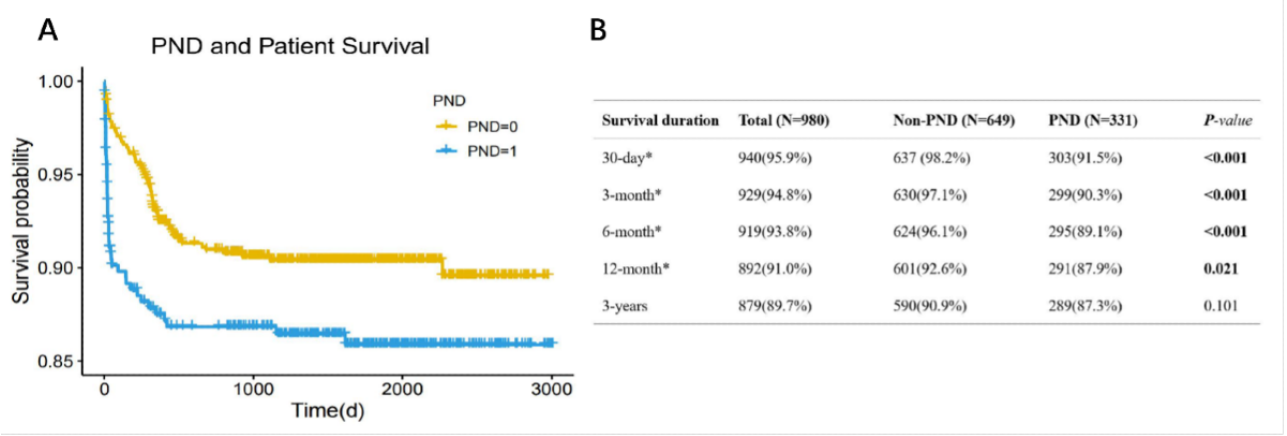

Survival duration	Total (N=980)	Non-PND (N=649)	PND (N=331)	*P-value*
30-day*	940(95.9%)	637 (98.2%)	303(91.5%)	**<0.001**
3-month*	929(94.8%)	630(97.1%)	299(90.3%)	**<0.001**
6-month*	919(93.8%)	624(96.1%)	295(89.1%)	**<0.001**
12-month*	892(91.0%)	601(92.6%)	291(87.9%)	**0.021**
3-years	879(89.7%)	590(90.9%)	289(87.3%)	0.101

A paragraph in the *Effect of PND on Patients' Outcomes and Prognosis* section has been revised. The text originally read as follows:

> *Further survival analysis (Figure 4)was conducted to assess patient prognosis. The PND group exhibited significantly lower survival rates at 30 days (87.1% vs 97.84%, P<.001), 3 months (83.99% vs 96.46%, P<.001), 6 months (82.78% vs 95.38%, P<.001), and 12 months (78.85% vs 88.44%, P<.001), and overall survival (P=.03).*

The text now reads:

Further survival analysis (Figure 4) was conducted to assess patient prognosis. The PND group exhibited significantly lower survival rates at 30 days (91.5% vs 98.2%, P<.001), 3 months (90.3% vs 97.1%, P<.001), 6 months (89.1% vs 96.1%, P<.001), and 12 months (87.9% vs 92.6%, P=.02).

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Ding Z, Zhang L, Zhang Y, Yang J, Luo Y, Ge M, et al. A supervised explainable machine learning model for perioperative neurocognitive disorder in liver-transplantation patients and external validation on the medical information mart for intensive care IV database: retrospective study. J Med Internet Res 2025 Jan 15;27:e55046 [FREE Full text] [doi: 10.2196/55046] [Medline: 39813086]

Submitted 04.Dec.2025; this is a non–peer-reviewed article;accepted 11.Dec.2025; published 11.Mar.2026.

Please cite as:
Ding Z, Zhang L, Zhang Y, Yang J, Luo Y, Ge M, Yao W, Hei Z, Chen C
Correction: A Supervised Explainable Machine Learning Model for Perioperative Neurocognitive Disorder in Liver-Transplantation Patients and External Validation on the Medical Information Mart for Intensive Care IV Database: Retrospective Study
J Med Internet Res 2026;28:e88952
URL: https://www.jmir.org/2026/1/e88952
doi:10.2196/88952
PMID:

©Zhendong Ding, Linan Zhang, Yihan Zhang, Jing Yang, Yuheng Luo, Mian Ge, Weifeng Yao, Ziqing Hei, Chaojin Chen. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Infoveillance of COVID-19 Infections in Dentistry Using Platform X: Descriptive Study

Alghalia Al-Mansoori[1], DDS; Ola Al Hayk[1], DDS; Sharifa Qassmi[1], DDS; Sarah M Aziz[1], DDS; Fatima Haouari[1], MSc; Tawanda Chivese[1], MSc, PhD; Faleh Tamimi[1], BDS, MSc, PhD; Alaa Daud[1], BDS, MSc, PGCert

College of Dental Medicine, QU Health, Qatar University, Doha, Qatar

Corresponding Author:
Alaa Daud, BDS, MSc, PGCert
College of Dental Medicine
QU Health
Qatar University
Al Tarfa St
Doha, 2713
Qatar
Phone: 974 77193993
Email: adaud@qu.edu.qa

Related Article:

Correction of: https://www.jmir.org/2025/1/e54650

(J Med Internet Res 2026;28:e86469) doi:10.2196/86469

In "Infoveillance of COVID-19 Infections in Dentistry Using Platform X: Descriptive Study" [1], the authors noted that one name is missing from the Acknowledgments section.

In the Acknowledgments section, the sentence originally appeared as:

> *The authors would like to thank Dr Rula Chami for her contribution to this research. The study received no funding.*

This sentence now reads:

> *The authors would like to thank Dr Rula Chami and Layan Mohamed for their contribution to this research. The study received no funding.*

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Al-Mansoori A, Al Hayk O, Qassmi S, Aziz SM, Haouari F, Chivese T, et al. Infoveillance of COVID-19 infections in dentistry using platform X: descriptive study. J Med Internet Res 2025 Apr 03;27:e54650 [FREE Full text] [doi: 10.2196/54650] [Medline: 40179381]

Submitted 24.Oct.2025; this is a non–peer-reviewed article;accepted 28.Oct.2025; published 31.Mar.2026.

Please cite as:
Al-Mansoori A, Al Hayk O, Qassmi S, Aziz SM, Haouari F, Chivese T, Tamimi F, Daud A
Correction: Infoveillance of COVID-19 Infections in Dentistry Using Platform X: Descriptive Study
J Med Internet Res 2026;28:e86469
URL: https://www.jmir.org/2026/1/e86469
doi:10.2196/86469
PMID:

XSL•FO
RenderX

©Alghalia Al-Mansoori, Ola Al Hayk, Sharifa Qassmi, Sarah M Aziz, Fatima Haouari, Tawanda Chivese, Faleh Tamimi, Alaa Daud. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparative Analysis of Prescriptions and Pharmacy Services in Internet-Based Psychiatric Hospital During and After the COVID-19 Pandemic: Retrospective Cross-Sectional Observational Study

Guowei Deng[1*], BS; Hui Xia[1*], BS; De-wei Shang[1,2], PhD; Yuguan Wen[1,2], MS; Jinqing Hu[1,2], MS; Yaqian Tan[1,2], PhD

[1]Department of Pharmacy, The Affiliated Brain Hospital, Guangzhou Medical University, Guangzhou, China

[2]Key Laboratory of Neurogenetics and Channelopathies of Guangdong Province and the Ministry of Education of China, Guangzhou Medical University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Yaqian Tan, PhD
Department of Pharmacy, The Affiliated Brain Hospital, Guangzhou Medical University
Mingxin Road 36
Guangzhou
China
Phone: 86 18898533945
Email: tanyaqian2013@163.com

Abstract

Background: The COVID-19 pandemic has significantly accelerated the development of internet-based hospitals and telepharmacy services. However, their characteristics and evolving trends remain unclear.

Objective: This study aimed to assess the associations between distinct pandemic phases and the number of prescriptions, patients' demographic characteristics, drug and disease distribution patterns, and pharmacy service indicators in our internet-based psychiatric hospital.

Methods: In this retrospective cross-sectional observational study, we conducted a full-sample census of prescriptions issued in the internet-based psychiatric hospital of the Affiliated Brain Hospital of Guangzhou Medical University during November 2020-December 2023. Cancelled, pending, and test prescriptions were excluded, and no sampling procedure was used. The research timespan was divided into pandemic and postpandemic phases, and trends of prescriptions were evaluated using interrupted time series analysis. Outcome measures, including patients' sex and age, diagnosed disease, drug type, pharmacist audit time, and audit outcome, were analyzed using the bootstrap method, Pearson chi-square analysis, and multinomial logistic regression.

Results: The segmented regression model revealed significant positive correlation between months and number of prescriptions during pandemic phase ($F_{1,16}$=6.96; P=.02), whereas no significant correlation was detected in postpandemic phase ($F_{1,10}$=2.77; P=.13). Descriptive analysis with bootstrap method revealed that female population were the majority in the pandemic phase (7297/11,812, 61.78%; 95% CI 60.91%-62.70%) and the postpandemic phase (3520/5518, 63.79%; 95% CI 62.58%-65.18%). Young adults aged 18-40 years were the predominant population in the pandemic phase (5606/11,812, 47.46%; 95% CI 46.63%-48.39%) and postpandemic phase (2657/5518, 48.15%; 95% CI 46.79%-49.37%). Depressive disorder and quetiapine were the most frequently diagnosed disease and prescribed drug in both pandemic phases, respectively. The majority of prescriptions were audited within 5 minutes during the pandemic phase (5999/11,812, 50.79%; 95% CI 49.89%-51.65%), while most prescriptions were audited within 1-12 hours in the postpandemic phase (2031/5518, 36.81%; 95% CI 35.61%-37.95%). Pearson chi-square analysis and multinomial logistic regression indicated that variables positively correlated with pandemic phases included female (P=.01; odds ratio [OR] 1.09, 95% CI 1.02-1.17), aged ≤17 years (P<.001; OR 2.20, 95% CI 1.90-2.54), aged 18-40 years (P<.001; OR 1.59, 95% CI 1.38-1.83), audit time between 12 and 24 hours (P=.02; OR 6.26, 95% CI 1.38-28.49), and approved outcome (P=.03; OR 3.97, 95% CI 1.19-13.26). The audit time ≤5 minutes (P=.049; OR 0.22, 95% CI 0.05-0.99) was negatively correlated with the pandemic phases.

Conclusions: This study innovatively applied descriptive and analytic statistical methods to evaluate the associations between different pandemic phases and the prescriptions and pharmacy services in an internet-based psychiatric hospital. This study addressed limitations of existing research through a larger sample size, longer research timespan, and analytic statistical methods.

XSL•FO
RenderX

This study demonstrated early warning indicators and replicable analytic methods for other medical institutions and offered implications in optimizing the efficiency of pharmacy services.

(J Med Internet Res 2026;28:e74059) doi:10.2196/74059

KEYWORDS

COVID-19; mental disorders; internet-based hospital; pharmacy services; prescriptions

Introduction

Background

The COVID-19 pandemic has profoundly affected the world and our daily lives [1-5]. In response to the COVID-19 outbreaks, the Chinese government had taken a range of actions, including temporary hospitals, lockdown, and quarantine [6,7]. However, these strict control measures inevitably disrupted conventional outpatient services, especially for patients with mental disorders [8]. Specifically, patients with chronic mental illness generally require long-term medication and regular medical assistance, and lockdown can cause difficulties in obtaining drugs and even the danger of discontinuation [9-11]. In addition, patients with long-term serious mental illness are often physically or socially disadvantaged, and COVID-19 highlighted these preexisting differences [12-15]. Finally, it has been reported that the global prevalence of mental health problems, such as loneliness, panic, anxiety, and depression, has increased significantly since the outbreak of the COVID-19 pandemic [16,17]. Hence, in order to meet the challenges of patients' medical needs, and in line with the recommendation of the World Health Organization to strengthen the digitalization of health systems, online medical services have developed rapidly worldwide [18-20]. The National Health Commission of China has also promoted the establishment of internet-based hospitals and timely translated into emergency control measures of COVID-19 [21,22]. Internet-based hospitals in China have been reported to grow rapidly from approximately 1600 in June 2021 to over 3000 by June 2023 [23], demonstrating the growing importance of this online service model.

In fact, with the advancement of artificial intelligence, 5G networks, and virtual reality, internet-based hospitals have developed rapidly in the last decade [24]. Internet-based hospitals displayed significant advantages over offline services in terms of spatial accessibility and cost-effectiveness [25,26]. Further, internet-based hospitals established by physical hospitals can integrate online and offline medical information and offer consultation, prescription, and follow-up services to ensure the continuity of health care for patients [27]. It is worth mentioning that there has been a major problem of uneven distribution of medical resources in China, primarily manifested as great regional differences [28,29]. However, the establishment of internet-based hospitals has greatly improved the accessibility of high-quality medical resources in big cities, thereby narrowing the differences in medical resources in China [30,31].

Subsequently, with the outbreak of COVID-19, extensive evidence has emphasized the growing importance of telehealth in providing stable health care services and curbing the spread of the pandemic [32,33]. By transferring nonemergency cases to online services, internet-based hospitals significantly alleviated the overburden of health care systems, thereby optimizing the allocation of medical resources [34,35]. Additionally, pharmacy services, such as medication consultation, medication therapy management, patient education, and drug delivery services, have been shown to significantly reduce the financial burdens of patients and improve their medication adherence [25,36,37]. Despite these potential benefits, disparities among different populations might also be inadvertently exacerbated during the pandemic, including internet facility conditions, digital skills, and physical conditions [38-41]. Meanwhile, due to the disadvantages such as insufficient interaction between doctors and patients and poor timeliness, the quality of online health care services has received great concern [22,42,43].

As for patients with mental disorders, telepsychiatry digital platforms have fundamentally improved the accessibility of patients and exhibited therapeutic effects comparable to offline treatment [44,45]. To meet the growing medication needs of patients, our hospital (the Affiliated Brain Hospital of Guangzhou Medical University) launched an internet-based psychiatric hospital platform and officially started to provide online pharmacy services since November 2020. Specifically, patients can receive remote consultation from doctors at home, pay online, and settle in real time through medical insurance. Subsequently, electronic prescriptions are audited by trained and authorized pharmacists, and drugs are delivered directly to patients through qualified third-party logistics channels to reduce the spread of COVID-19.

Relation to Previous Work

At present, most studies on the telemedicine services during the COVID-19 pandemic have focused on general hospitals [46-49], and few studies have explored the online pharmacy services in internet-based psychiatric hospitals [21,23]. Moreover, the limitations of existing studies in this area still exist, including small sample sizes, restricted study periods, and a lack of comparison between different pandemic phases. Besides, it is still unclear how the gradual resumption of normal medical services after the pandemic will affect internet-based hospitals. Thus, to fill these knowledge gaps, we conducted a retrospective cross-sectional observational study using the electronic prescriptions issued in our internet-based psychiatric hospital during November 2020 to December 2023. This study had a larger sample size and a longer research timespan than previous studies in this field and used analytic statistical methods to ensure the reliability of our findings. This study could gain in-depth insights into the development trends of prescriptions and key predictors of pharmacy services in internet-based psychiatric hospitals and offer empirical guidance for other medical institutions in the effective responses to public health emergencies.

XSL•FO
RenderX

Objectives

In this study, we aimed to analyze the associations between different COVID-19 pandemic phases (predictor) and the trends and distribution patterns of the electronic prescriptions through descriptive and analytic statistical methods. The specific outcomes, including monthly prescription numbers, patients' demographic characteristics (sex and age), clinical characteristics (primary diagnosed disease and type of drug), and pharmacy service indicators (pharmacist audit time and audit outcome), would be assessed to elucidate the changes in their distribution patterns driven by the pandemic phases.

Methods

All methods and findings of this study were reported based on the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [50] and the JARS (Journal Article Reporting Standards) guidelines [51,52]. A completed and filled out STROBE checklist of this study is provided as Multimedia Appendix 1.

Study Design

This study was a retrospective cross-sectional observational study that comparatively analyzed the prescriptions and pharmacy services from an internet-based psychiatric hospital during November 2020 to December 2023. We used descriptive and analytic statistical methods to analyze the associations between the 2 distinct pandemic phases (pandemic phase and postpandemic phase) and the long-term trends of the prescriptions, patients' demographic characteristics, drug and disease distribution patterns, and pharmacy service indicators.

Settings and Data Collection

All data in this study were obtained from the internet-based psychiatric hospital of the Affiliated Brain Hospital of Guangzhou Medical University, a tertiary psychiatric hospital located in Guangzhou, China. The telepharmacy service platform was officially launched and started to provide online services since November 2020. Patients can visit our internet-based psychiatric hospital through the app developed by Guangdong Yunhui Technology Co, Ltd. After patients log on to the app and input their personal information, doctors will confirm their identity information and issue prescriptions according to patients' condition. All electronic prescriptions from the internet-based psychiatric hospital are manually audited by trained and authorized pharmacists. The detailed workflow of the doctor prescription and the pharmacist audit process is demonstrated in Figure 1.

Figure 1. Flowchart depicting the doctor prescription and pharmacist audit process in the internet-based psychiatric hospital of the Affiliated Brain Hospital of Guangzhou Medical University during November 2020 to December 2023.

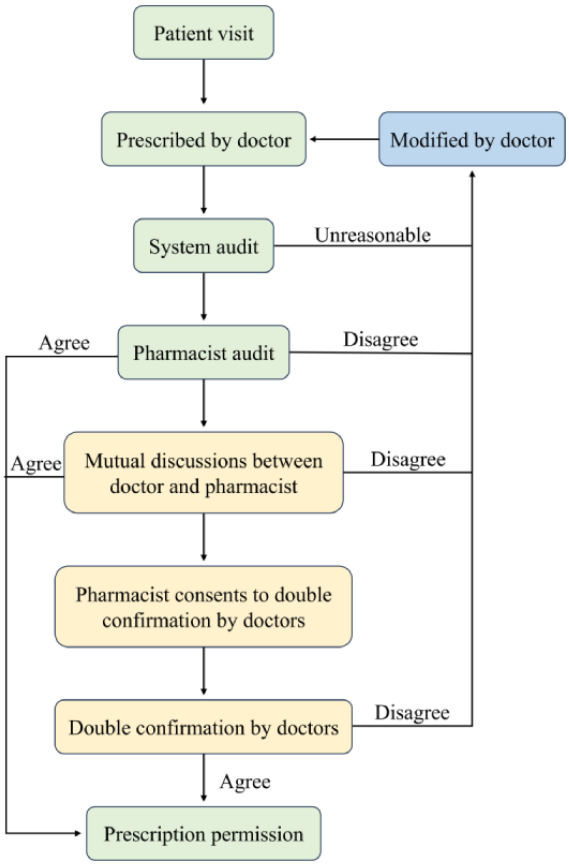

Data were collected from the electronic prescriptions of patients who received online medical services in our internet-based psychiatric hospital during November 1, 2020, to December 31, 2023. In this study, the exposure was the different pandemic phases. This study was a retrospective cross-sectional observational study, and no follow-up was performed. Data collection was conducted on a single day of March 11, 2024.

XSL•FO
RenderX

Ultimately, a total of 17,330 electronic prescriptions with 36,088 drug records were identified and included in the final analysis.

Eligibility Criteria

This study aimed to analyze the entire population of prescriptions issued in our online platform during November 2020 to December 2023. There were no restrictions on prescription inclusion based on patients' demographic characteristics (eg, sex, age, ethnicity, and socioeconomic status), and the inclusion criterion was defined as all prescriptions issued throughout the study period. Therefore, we essentially adopted a full-sample census approach. During the subsequent data screening process, cancelled, pending, and test prescriptions were considered to meet exclusion criteria and were removed from further data analysis. After data screening, a total of 17,330 prescriptions were included in the final data analysis. All data items of these included prescriptions were complete with no missing data. In addition, given the full-sample census method used, no sampling procedure was involved in this study. The detailed workflow of the eligibility criteria and data screening process is illustrated in Figure 2.

Figure 2. Flowchart demonstrating the eligibility criteria and data screening process of the electronic prescriptions issued in the internet-based psychiatric hospital of the Affiliated Brain Hospital of Guangzhou Medical University during November 2020 to December 2023.

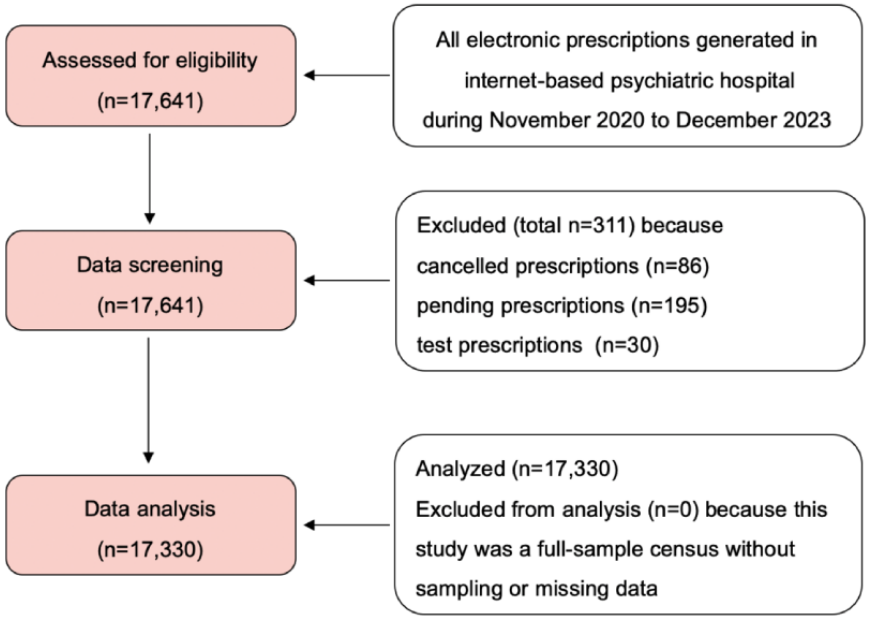

Variables

In this study, the outcomes were measured at the prescription level, including the number of prescriptions, patients' demographic characteristics (sex and age), clinical characteristics (primary diagnosed disease and type of drug), and pharmacy service indicators (pharmacist audit time and audit outcome). Within the standard framework of this observational research, the pandemic phases could theoretically be considered as potential "predictors" of the outcomes. The conditions in this study were all naturally observed, and no causal analysis of observed phenomena was performed. Thus, no variables were formally specified as confounders or effect modifiers.

Bias

To prevent data inconsistencies due to periodic updates of the system when extracting data over multiple days, such as retrospective corrections of prescription status or batch updates of diagnoses and drug codes, all raw data in this study were extracted directly from our online platform backend system within a single day of March 11, 2024. This approach allowed us to ensure that all collected prescriptions had the same data status or version, thereby minimizing potential bias from data sources and guaranteeing data standardization.

Data Sources or Measurement and Quantitative Variables

In December 2022, the Chinese government announced the end of all COVID-19 prevention and quarantine control measures in China. Based on this, the timeline of the COVID-19 pandemic was divided into 2 phases: the pandemic phase (November 2020-December 2022) and postpandemic phase (January 2023-December 2023). Exposure to the distinct COVID-19 pandemic phases was therefore defined and evaluated throughout the study period.

Based on the 2 pandemic phases divided, the number of monthly prescriptions and cumulative prescriptions were presented using a continuous time variable in months. Further, a 2-stage interrupted time series (ITS) analysis was conducted. In the pandemic phase, we adjusted the starting time point to July 2021 to exclude abnormal values (values deviating more than 2 times the CI from the absolute value) in June 2021. A segmented regression model in ITS analysis was used to evaluate the association between COVID-19 relaxation in China and the trends of monthly prescriptions [53,54]. A continuous time variable in months was applied in the model, and data were presented as absolute values with 95% CI. The *P* value <.05 was accepted as statistical significance in the segmented regression model.

XSL•FO
RenderX

Descriptive data of patients' demographic characteristics, drug and disease distribution patterns, and pharmacy service indicators were presented using frequencies and percentages. The subgroups of age were assigned as follows: pediatric and adolescent (≤17 years), young adult (18-40 years), middle-aged (41-65 years), and older adult (≥66 years). The age of patients was assessed using median with IQR.

The pharmacy service indicators included pharmacist audit time and audit outcome. Pharmacist audit time was defined as the total length of time (in minutes) from prescription submission to audit completion, which was automatically recorded in our audit system. During data analysis, the audit time was divided into 7 subgroups according to our institutional pharmacy operation guidelines, including "≤5 minutes," "5-30 minutes," "30 minutes-1 hour," "1-12 hours," "12-24 hours," "24-48 hours," and " 48 hours." The audit outcomes mainly included "approved by pharmacists," "double confirmation by doctors," and "not approved by pharmacists." We used consistent evaluation criteria for the audit outcomes of all prescriptions in this study according to our institutional pharmacy operation guidelines. To be specific, "approved by pharmacists" referred to prescriptions that meet the patient's clinical condition and rationale for medication, "double confirmation by doctors" referred to prescriptions with potential medication errors (eg, inappropriate indication, repeated medication, and overdose) that required double confirmation by doctors, and "not approved by pharmacists" referred to prescriptions disapproved by pharmacists due to serious safety concerns (eg, contraindicated drugs and lethal doses).

To examine the robustness of the descriptive data, including patients' sex and age, diagnosed disease, drug type, pharmacist audit time, and audit outcome, percentage values were sampled repeatedly for 1000 times using the bootstrap method and were presented as absolute values with 95% CI [55,56].

Statistical Methods

Statistical data were presented for multiple variables, including sex, age, pharmacist audit time, and audit outcome. The general associations between pandemic phases and outcomes were examined using crosstabs analysis and Pearson chi-square analysis [54]. If the association was significant (P<.05), we further performed a multinomial logistic regression analysis to determine the source of significance using pandemic phases as a covariate [57]. Significance level of the Wald inclusion test statistic was applied with P value <.05. Odds ratios (ORs) with 95% CI and SE were calculated to quantify the associations between pandemic phases and outcomes. In this study, our predetermined α level was .05, and a 2-sided P value <.05 was considered statistically significant. All statistical analysis and graphical representations were performed using SPSS (version 26.0.0.2; IBM Corp) and GraphPad Prism (version 10.1.2; GraphPad Software, Inc).

Ethical Considerations

This study was approved by the institutional review board (IRB) of the Affiliated Brain Hospital of Guangzhou Medical University following a thorough review of the research protocol (approval 2025111). The informed consent of this study was waived by the IRB, and the IRB allowed the primary data collection and secondary analysis of research data without additional consent. We confirm that this study adhered strictly to the principles of privacy and confidentiality protection. The research team complied with all relevant local, national, and international laws and regulations regarding the protection of personal information, privacy, and human rights. All sensitive data related to patients, including name, ethnicity, socioeconomic status, education level, residential address, contact information, geographic distribution, prescription cost, primary complaint, medical history, previous diagnosis, information of the clinician, and content of the conversation, were deidentified to ensure privacy and security. Besides sensitive information, all other collected data were reported in this study. This study involved prescription data from the local database of our hospital and did not involve human experimentation or compensation. We confirm that no personally identifiable information of patients was accessible to the research team. We confirm that no identification of individual participants or users in any images of the manuscript or supplementary material is possible.

Results

General Trends of Monthly Prescriptions

In this study, a total of 17,330 electronic prescriptions were finally identified, of which 11,812 prescriptions were processed during the pandemic phase, and 5518 prescriptions were issued during the postpandemic phase. As shown in Figure 3, the general trend of monthly prescriptions reflected a fluctuating tendency from November 2020 to December 2023. During the early stage of the pandemic phase from November 2020 to May 2021, the number of monthly prescriptions remained relatively low, with 76 prescriptions in total. Subsequently, the number of monthly prescriptions dramatically surged in June 2021 (n=3427), and then decreased to an average level of approximately 350 until another peak in November 2022 (n=1506). In the postpandemic phase, the number of monthly prescriptions stayed relatively stable around an average level of 460. On the other hand, the cumulative number of prescriptions exhibited a continuous upward tendency and peaked at 36,088 in December 2023.

XSL•FO
RenderX

Figure 3. The number of monthly prescriptions and cumulative prescriptions in the internet-based psychiatric hospital of the Affiliated Brain Hospital of Guangzhou Medical University. Sources of data are all the electronic prescriptions issued during November 2020 to December 2023. Values are presented using a continuous time variable in months.

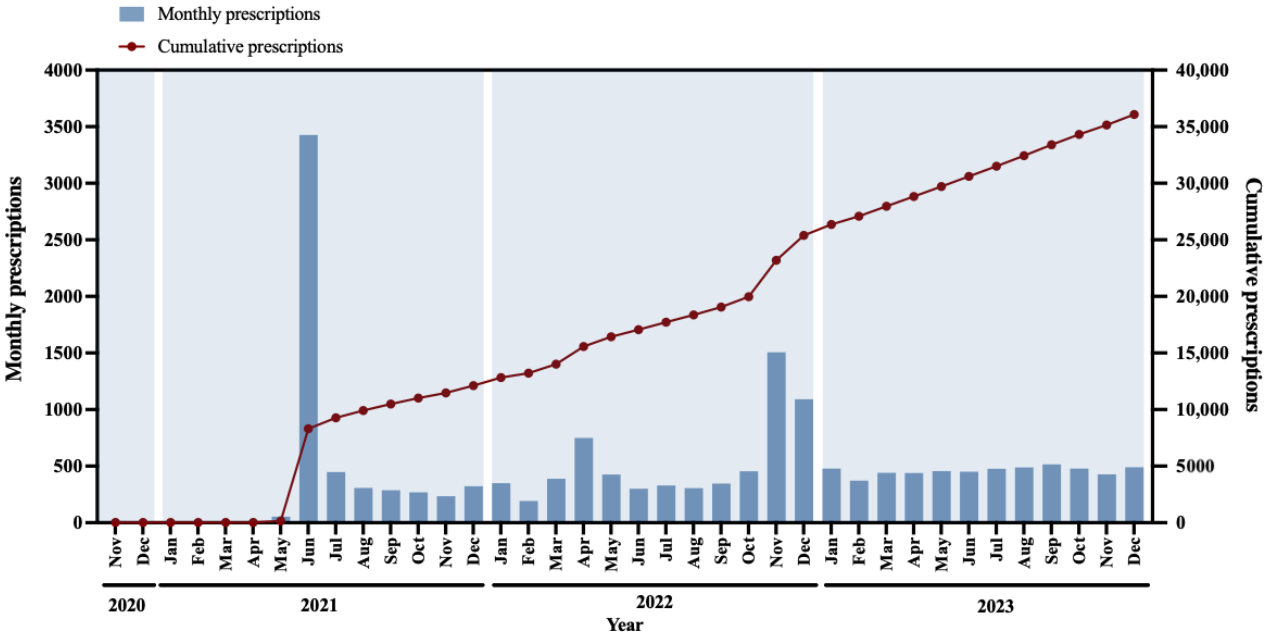

ITS Analysis of Monthly Prescriptions

A further ITS analysis was conducted to evaluate the association between COVID-19 relaxation and the number of monthly prescriptions. We conducted a 2-stage ITS analysis using the data ranging from July 2021 to December 2023. As depicted in Figure 4, the segmented regression model revealed a significant positive correlation between the months and the number of prescriptions (y=34.52*x+133.60; r=0.55; $F_{1,16}$=6.96; P=.02; slope 34.52; 95% CI 6.78-62.27). These findings suggested an increasing tendency in patients' needs and interest in our internet-based psychiatric hospital. In the postpandemic phase, the number of monthly prescriptions remained generally stable, and the trend was not statistically significant compared to baseline (y=4.83*x+428.50; r=0.46; $F_{1,10}$=2.77; P=.13; slope 4.83; 95% CI –1.63 to 11.28). These findings indicated a stable and continuous service model of internet-based psychiatric hospital and a steady medication adherence of patients during January 2023 to December 2023.

Figure 4. Segmented regression model in the 2-stage interrupted time series analysis evaluating the association between COVID-19 relaxation and the number of monthly prescriptions in the internet-based psychiatric hospital of the Affiliated Brain Hospital of Guangzhou Medical University. Sources of data are all the electronic prescriptions issued during July 2021 to December 2023. Values are presented as absolute values (blue dots) with 95% CI (red shadow) with a continuous time variable in months. The red solid lines indicate lines of regression, and the vertical black dashed line indicates the end of all COVID-19 prevention and quarantine control measures in China in December 2022. The P value <.05 indicates a significant difference between the regression line and the baseline.

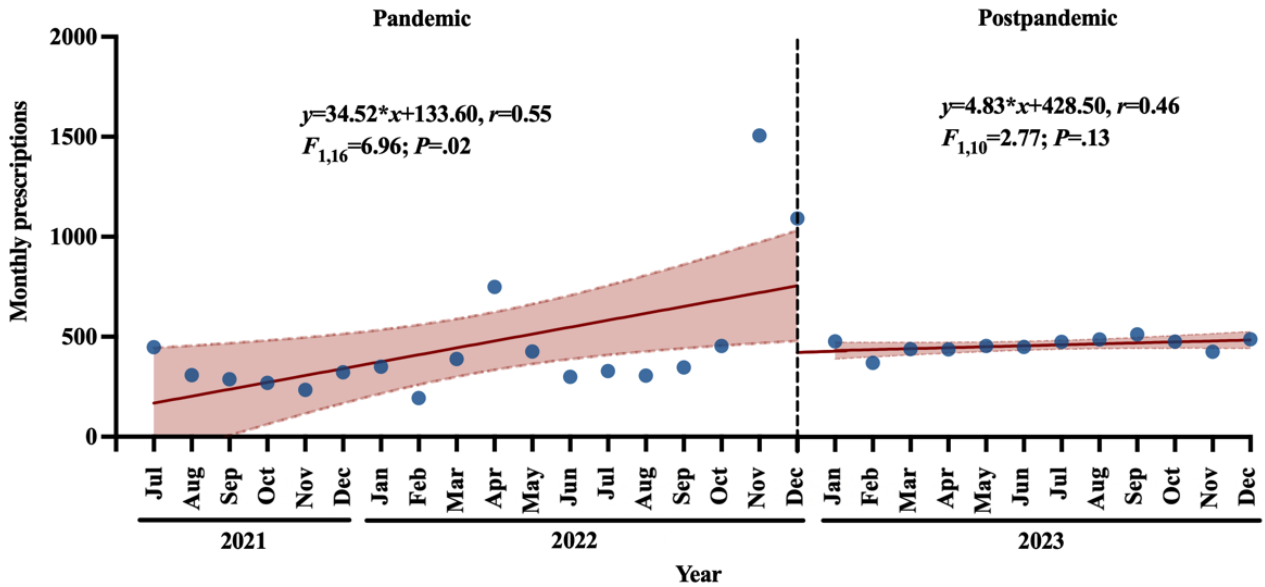

XSL•FO
RenderX

Demographic Characteristics of Patients

In this study, a total of 17,330 electronic prescriptions were collected, of which 11,812 prescriptions were issued during pandemic phase, and 5518 prescriptions were issued during the postpandemic phase. As the detailed characteristics of patients displayed in Table 1, the majority of patients were female during the pandemic phase (7297/11,812, 61.78%; 95% CI 60.91%-62.70%) and the postpandemic phase (3520/5518, 63.79%; 95% CI 62.58%-65.18%). There were 4515 male patients in the pandemic phase (4515/11,812, 38.22%; 95% CI 37.36%-39.03%) and 1998 male patients in the postpandemic phase (1998/5518, 36.21%; 95% CI 34.99%-37.32%).

Table 1. Demographic characteristics of patients who visited the internet-based psychiatric hospital of the Affiliated Brain Hospital of Guangzhou Medical University during November 2020 to December 2023[a].

Characteristic	Pandemic (n=11,812), n (%)	95% CI	Postpandemic (n=5518), n (%)	95% CI	Chi-square (*df*)	*P* value
Sex					6.5 (1)	.01
Female	7297 (61.78)	60.91-62.70	3520 (63.79)	62.58-65.18		
Male	4515 (38.22)	37.36-39.03	1998 (36.21)	34.99-37.32		
Age (years)[b]					295.4 (3)	<.001
≤17	2873 (24.32)	23.47-25.19	1880 (34.07)	32.78-35.40		
18-40	5606 (47.46)	46.63-48.39	2657 (48.15)	46.79-49.37		
41-65	2400 (20.32)	19.61-21.07	703 (12.74)	11.92-13.59		
≥66	933 (7.90)	7.42-8.36	278 (5.04)	4.48-5.62		

[a]Values are presented as absolute values and percentages with 95% CI using the bootstrap method. The general associations between different pandemic phases and sex and age of patients are evaluated with crosstabs analysis and presented with Pearson chi-square values.

[b]Pandemic: median 21 (IQR 18-44) years, postpandemic: median 21 (IQR 16-33) years.

Regarding age, young adults aged between 18 and 40 years accounted for the predominant population (5606/11,812, 47.46%; 95% CI 46.63%-48.39%) during pandemic phase, followed by patients aged ≤17 years (2873/11,812, 24.32%; 95% CI 23.47%-25.19%), patients aged between 41 and 65 years (2400/11,812, 20.32%; 95% CI 19.61%-21.07%), and patients aged ≥66 years (933/11,812, 7.90%; 95% CI 7.42%-8.36 %). Similarly, in postpandemic phase, young adults aged between 18 and 40 years accounted for the majority (2657/5518, 48.15%; 95% CI 46.79%-49.37%), followed by patients aged ≤17 years (1880/5518, 34.07%; 95% CI 32.78%-35.40%), patients aged between 41 and 65 years (703/5518, 12.74%; 95% CI 11.92%-13.59%), and patients aged ≥66 years (278/5518, 5.04%, 95 % CI 4.48%-5.62%). The median value of age was 21 (IQR 18-44) years during the pandemic phase and 21 (IQR 16-33) years in the postpandemic phase.

Overall, the Pearson chi-square analysis suggested significant statistical differences in sex distribution (χ^2_1=6.5; *P*=.01) and age distribution (χ^2_3=295.4; *P*<.001) between the 2 pandemic phases. These differences thereby suggested the necessity to perform further multinomial logistic regression analysis to determine the source of significance.

A follow-up multinomial logistic regression model was used to analyze the influences of different COVID-19 phases on the demographic characteristics of patients. As shown in Table 2, the postpandemic phase exhibited a positive correlation with the female group (*P*=.01; OR 1.09, 95% CI 1.02-1.17), suggesting a significantly increased proportion of female patients from 61.78% (7297/11,812) in the pandemic phase to 63.79% (3520/5518) in the postpandemic phase. Further, postpandemic phase displayed positive correlations with patients aged ≤17 years (*P*<.001; OR 2.20, 95% CI 1.90-2.54) and patients aged 18-40 years (*P*<.001; OR 1.59, 95% CI 1.38-1.83), suggesting significantly increased proportions of patients aged ≤17 years from 24.32% (2873/11,812) in pandemic phase to 34.07% (1880/5518) in postpandemic phase, and patients aged 18-40 years from 47.46% (5606/11,812) in pandemic phase to 48.15% (2657/5518) in postpandemic phase. We did not find a significant association between pandemic phases and patients aged ≥66 years across different pandemic phases (*P*=.83; OR 0.98, 95% CI 0.84-1.15).

XSL•FO
RenderX

Table 2. The influences of COVID-19 phases on the subgroups of patients' demographic characteristics in a multinomial logistic regression model[a].

Characteristic	SE	Wald	*df*	*P* value	OR[b] (95% CI)
Sex					
Female	0.03	6.51	1	.01	1.09 (1.02-1.17)
Male[c]	—[d]	—	—	—	1.00 (—)
Age (years)					
≤17	0.07	111.53	1	<.001	2.20 (1.90-2.54)
18-40	0.07	41.24	1	<.001	1.59 (1.38-1.83)
41-65	0.08	0.05	1	.83	0.98 (0.84-1.15)
≥66[c]	—	—	—	—	1.00 (—)

[a]Data presented in this table are from the follow-up analysis of patients' demographic characteristics to determine the source of the statistical difference. The SE and Wald inclusion test statistic are applied to quantify the associations between pandemic phases and outcomes.

[b]OR: odds ratio.

[c]Reference group.

[d]Not available.

Distribution of Primary Diagnosed Diseases

As listed in Table 3, the top 10 primary diagnosed diseases in frequency were examined in our descriptive analysis. During the pandemic phase, a total of 83 diagnosed diseases were detected. The primary diagnosis of depressive disorder (3539/11,812, 29.96%; 95% CI 29.15%-30.79%) ranked first in frequency, followed by schizophrenia (2095/11,812, 17.74%; 95% CI 17.10%-18.40%) and bipolar disorder (1632/11,812, 13.82%; 95% CI 13.17%-14.50%). The primary diagnosis of epilepsy (108/11,812, 0.91%; 95% CI 0.77%-1.06%) ranked lowest among the top 10 primary diagnosed diseases. During the postpandemic phase, 58 diagnosed diseases in total were identified, and the leading 3 were depressive disorder (2094/5518, 37.95%; 95% CI 36.74%-39.12%), mood disorder (1191/5518, 21.58%; 95% CI 20.50%-22.74%), and schizophrenia (607/5518, 11%; 95% CI 10.20%-11.82%). In the postpandemic phase, epilepsy (49/5518, 0.89%; 95% CI 0.67%-1.12%) ranked lowest among the top 10 primary diagnosed diseases.

XSL•FO
RenderX

Table 3. The distribution of the top 10 primary diagnosed diseases from the electronic prescriptions processed in the internet-based psychiatric hospital of the Affiliated Brain Hospital of Guangzhou Medical University during November 2020 to December 2023[a].

Rank	Primary diagnosed disease	Values, n (%)	95% CI
Pandemic (n=11,812)			
1	Depressive disorder	3539 (29.96)	29.15-30.79
2	Schizophrenia	2095 (17.74)	17.10-18.40
3	Bipolar disorder	1632 (13.82)	13.17-14.50
4	Mood disorder	1444 (12.22)	11.65-12.86
5	Anxiety disorder	1208 (10.23)	9.61-10.82
6	Obsessive-compulsive disorder	328 (2.78)	2.47-3.08
7	Attention-deficit/hyperactivity disorder	232 (1.96)	1.73-2.22
8	Alzheimer disease	164 (1.39)	1.19-1.61
9	Autistic disorder	116 (0.98)	0.81-1.15
10	Epilepsy	108 (0.91)	0.77-1.06
Postpandemic (n=5518)			
1	Depressive disorder	2094 (37.95)	36.74-39.12
2	Mood disorder	1191 (21.58)	20.50-22.74
3	Schizophrenia	607 (11.00)	10.20-11.82
4	Anxiety disorder	323 (5.85)	5.21-6.52
5	Bipolar disorder	301 (5.45)	4.89-5.96
6	Obsessive-compulsive disorder	162 (2.94)	2.56-3.35
7	Attention-deficit/hyperactivity disorder	127 (2.30)	1.94-2.66
8	Autistic disorder	117 (2.12)	1.81-2.45
9	Alzheimer disease	72 (1.30)	1.04-1.59
10	Epilepsy	49 (0.89)	0.67-1.12

[a]Values are presented as absolute values and percentages with 95% CI using the bootstrap method.

Distribution of Drugs in Prescriptions

A total of 114 types of drugs prescribed in 25,402 times during the pandemic phase and 126 types of drugs prescribed in 10,686 times during the postpandemic phase were identified. As listed in Table 4, the top 10 drugs in frequency were examined in our descriptive analysis. During the pandemic phase, the predominant drugs in prescriptions were quetiapine (2653/25,402, 10.44%; 95% CI 10.03%-10.81%), lithium carbonate (1439/25,402, 5.66%; 95% CI 5.39%-5.93%), and escitalopram (1414/25,402, 5.57%; 95% CI 5.30%-5.85%). The drug agomelatine (870/25,402, 3.42%; 95% CI 3.20%-3.65%) had the lowest frequency among the top 10 prescribed drugs. Likewise, in the postpandemic phase, the top 3 most frequently prescribed drugs were quetiapine (1276/10,686, 11.94%; 95% CI 11.37%-12.55%), lithium carbonate (733/10,686, 6.86%; 95% CI 6.36%-7.37%), and escitalopram (710/10,686, 6.64%; 95% CI 6.20%-7.07%). Olanzapine (458/10,686, 4.29%; 95% CI 3.91%-4.71%) ranked lowest in the top 10 prescribed drugs during the postpandemic phase.

XSL•FO
RenderX

Table 4. The distribution of the top 10 drugs from the electronic prescriptions processed in the internet-based psychiatric hospital of the Affiliated Brain Hospital of Guangzhou Medical University during November 2020 to December 2023[a].

Rank	Drug	Values, n (%)	95% CI
Pandemic (n=25,402)[b]			
1	Quetiapine	2653 (10.44)	10.03-10.81
2	Lithium carbonate	1439 (5.66)	5.39-5.93
3	Escitalopram	1414 (5.57)	5.30-5.85
4	Olanzapine	1398 (5.50)	5.26-5.75
5	Sodium Valproate	1356 (5.34)	5.07-5.62
6	Sertraline	1356 (5.34)	5.08-5.61
7	Aripiprazole	1247 (4.91)	4.63-5.18
8	Trihexyphenidyl	1185 (4.66)	4.41-4.93
9	Tandospirone	932 (3.67)	3.46-3.89
10	Agomelatine	870 (3.42)	3.20-3.65
Postpandemic (n=10,686)[b]			
1	Quetiapine	1276 (11.94)	11.37-12.55
2	Lithium carbonate	733 (6.86)	6.36-7.37
3	Escitalopram	710 (6.64)	6.20-7.07
4	Tandospirone	657 (6.15)	5.66-6.64
5	Sodium valproate	612 (5.73)	5.31-6.14
6	Sertraline	584 (5.47)	5.02-5.90
7	Aripiprazole	570 (5.33)	4.90-5.75
8	Lamotrigine	495 (4.63)	4.23-5.01
9	Fluoxetine	460 (4.30)	3.93-4.67
10	Olanzapine	458 (4.29)	3.91-4.71

[a]Values are presented as absolute values and percentages with 95% CI using the bootstrap method.

[b]Some prescriptions may contain more than 1 drug.

Distribution of Prescription Audit Time and Audit Outcome

The detailed characteristics of prescription audit in the internet-based psychiatric hospital are listed in Table 5. During the pandemic phase, the majority of prescriptions (5999/11,812, 50.79%; 95% CI 49.89%-51.65%) were audited in ≤5 minutes, followed by prescriptions audited within 5-30 minutes (3255/11,812, 27.56%; 95% CI 26.74%-28.36%) and prescriptions audited within 1-12 hours (1466/11,812, 12.41%; 95% CI 11.86%-13.04%). The group of audit time 48 hours had the lowest number of prescriptions (4/11,812, 0.03%; 95% CI 0.01%-0.07%). In contrast, during postpandemic phase, most prescriptions (2031/5518, 36.81%; 95% CI 35.61%-37.95%) were audited within 1-12 hours, followed by prescriptions audited within 5-30 minutes (1359/5518, 24.63%; 95% CI 23.47%-25.73%) and prescriptions audited in ≤5 minutes (998/5518, 18.09%; 95% CI 17.13%-19.17%). Likewise, in the postpandemic phase, the group of audit time 48 hours had the lowest number of prescriptions (3/5518, 0.05%; 95% CI 0.01%-0.13%).

XSL•FO
RenderX

Table 5. The characteristics of prescription audit in the internet-based psychiatric hospital of the Affiliated Brain Hospital of Guangzhou Medical University during November 2020 to December 2023[a].

Characteristic	Pandemic (n=11,812), n (%)	95% CI	Postpandemic (n=5518), n (%)	95% CI	Chi-square (*df*)	*P* value
Audit time					2784.5 (6)	<.001
≤5 minutes	5999 (50.79)	49.89-51.65	998 (18.09)	17.13-19.17		
5-30 minutes	3255 (27.56)	26.74-28.36	1359 (24.63)	23.47-25.73		
30 minutes-1 hour	957 (8.10)	7.65-8.57	669 (12.12)	11.31-13.01		
1-12 hours	1466 (12.41)	11.86-13.04	2031 (36.81)	35.61-37.95		
12-24 hours	86 (0.73)	0.57-0.88	404 (7.32)	6.64-8.07		
24-48 hours	45 (0.38)	0.28-0.50	54 (0.98)	0.76-1.23		
48 hours	4 (0.03)	0.01-0.07	3 (0.05)	0.01-0.13		
Audit outcome					601.0 (2)	<.001
Approved by pharmacists	9844 (83.34)	82.68-84.06	5327 (96.54)	96.04-97.01		
Double confirmation by doctors	1946 (16.47)	15.85-17.05	188 (3.41)	2.98-3.88		
Not approved by pharmacists	22 (0.19)	0.12-0.25	3 (0.05)	0.01-0.13		

[a]Values are presented as absolute values and percentages with 95% CI using the bootstrap method. The general associations between different pandemic phases and audit time and audit outcome are evaluated with crosstabs analysis and presented with Pearson chi-square values.

As for audit outcomes, most prescriptions (9844/11,812, 83.34%; 95% CI 82.68%-84.06%) were approved by pharmacists during the pandemic phase. Of the 11,812 prescriptions in the pandemic phase, 1946 (16.47%; 95% CI 15.85%-17.05%) required double confirmation by doctors before approval, and 22 (0.19%; 95% CI 0.12%-0.25%) were not approved by pharmacists. Similarly, during the postpandemic phase, the vast majority of prescriptions (5327/5518, 96.54%; 95% CI 96.04%-97.01%) were approved by pharmacists. Of the 5518 prescriptions in the postpandemic phase, 188 (3.41%; 95% CI 2.98%-3.88%) required double confirmation by doctors before approval, and 3 (0.05%; 95% CI 0.01%-0.13%) were not approved by pharmacists.

Overall, the Pearson chi-square analysis revealed a significant statistical difference in audit time (χ^2_6=2784.5; *P*<.001) and audit outcomes (χ^2_2=601.0; *P*<.001) between the 2 pandemic phases. These differences thereby suggested the necessity to perform further multinomial logistic regression analysis to determine the source of significance.

A follow-up multinomial logistic regression model was used to analyze the influences of different COVID-19 phases on the characteristics of prescription audit. As shown in Table 6, the postpandemic phase exhibited a negative correlation with audit time ≤5 minutes (*P*=.049; OR 0.22, 95% CI 0.05-0.99), suggesting a significantly decreased proportion of prescriptions audited in ≤5 minutes from 50.79% (5999/11,812) in the pandemic phase to 18.09% (998/5518) in the postpandemic phase. There was a positive correlation with audit time within 12-24 hours (*P*=.02; OR 6.26, 95% CI 1.38-28.49), suggesting a significantly increased proportion of prescriptions audited within 12-24 hours from 0.73% (86/11,812) in the pandemic phase to 7.32% (404/5518) in the postpandemic phase. We did not find significant associations between different pandemic phases and audit time within 5-30 minutes (*P*=.44; OR 0.56, 95% CI 0.12-2.49), audit time within 30-60 minutes (*P*=.93; OR 0.93, 95% CI 0.21-4.17), audit time within 1-12 hours (*P*=.42; OR 1.85, 95% CI 0.41-8.27), or audit time within 24-48 hours (*P*=.55; OR 1.60, 95% CI 0.34-7.53).

XSL•FO
RenderX

Table 6. The influences of COVID-19 phases on the subgroups of prescription audit indicators in a multinomial logistic regression model[a].

Characteristic	SE	Wald	*df*	*P* value	OR[b] (95% CI)
Audit time					
≤5 minutes	0.77	3.88	1	.049	0.22 (0.05-0.99)
5-30 minutes	0.76	0.59	1	.44	0.56 (0.12-2.49)
30 minutes-1 hour	0.76	0.01	1	.93	0.93 (0.21-4.17)
1-12 hours	0.77	0.64	1	.42	1.85 (0.41-8.27)
12-24 hours	0.77	5.63	1	.02	6.26 (1.38-28.49)
24-48 hours	0.79	0.35	1	.55	1.60 (0.34-7.53)
48 hours[c]	—[d]	—	—	—	1.00 (—)
Audit outcome					
Approved by pharmacists	0.62	5.01	1	.03	3.97 (1.19-13.26)
Double confirmation by doctors	0.62	0.31	1	.58	0.71 (0.21-2.39)
Not approved by pharmacists[c]	—	—	—	—	1.00 (—)

[a]Data presented in this table are from the follow-up analysis of the indicators of prescription audit to determine the source of statistical difference. The SE and Wald inclusion test statistic are applied to quantify the associations between pandemic phases and outcomes.

[b]OR: odds ratio.

[c]Reference group.

[d]Not available.

Furthermore, the postpandemic phase displayed a positive correlation with the approved group (*P*=.03; OR 3.97, 95% CI 1.19-13.26), suggesting a significantly increased proportion of approved outcomes from 83.34% (9844/11,812) in the pandemic phase to 96.54% (5327/5518) in the postpandemic phase. We did not find significant associations between different pandemic phases and doctor double confirmation group (*P*=.58; OR 0.71, 95% CI 0.21-2.39).

Discussion

Principal Findings

In general, we found distinct trends of prescriptions in different phases of the pandemic, such that an overall upward trend during the pandemic phase and a stable tendency during the postpandemic phase. We found that female and young adults were the predominant groups, and their proportions increased with the development of the pandemic phases. Additionally, our findings demonstrated that the distribution patterns of primary diagnosed diseases and prescribed drugs were generally similar in both pandemic phases. At last, our results revealed that with the progress of the pandemic phases, the pharmacist audit time was extended, and the audit approval rate was increased.

Detailed Discussion of the Findings

Interpretations

From the perspective of monthly prescriptions, we found that our internet-based psychiatric hospital exhibited distinct distribution patterns across 2 pandemic phases. Our results from ITS analysis might reveal a generally increasing need for patients in online pharmacy services during the pandemic phase and the ability of internet-based psychiatric hospitals to provide stable pharmacy services after the pandemic [58,59].

Initially, during November 2020-May 2021, the small number of cumulative prescriptions possibly indicated that patients have not yet adapted to the shift from offline to online medical treatment model [60]. Notably, in June 2021, with the major outbreak in Guangzhou, the outpatient services of our hospital were suspended, which led to a dramatic surge in the number of electronic prescriptions [61]. This finding was in line with previous evidence, indicating the public panic in the early stage of the pandemic and the increased demand for patients for internet-based hospitals [21,23]. Subsequently, from July 2021 to October 2022, with the optimization of human resources in our internet-based psychiatric hospital and the gradually reduced public panic [62,63], the average number of monthly prescriptions remained relatively stable at approximately 350. During November 2022-December 2022, with another major outbreak in Guangzhou, together with the subsequent COVID-19 control relaxation in China, the demand for internet-based psychiatric hospitals surged again [64]. In the postpandemic phase, the average number of monthly prescriptions was stable at approximately 460, which was consistent with previous findings, indicating the critical role of internet-based hospitals during and after the COVID-19 pandemic [65]. Importantly, our findings highlighted the capacity and necessity of internet-based psychiatric hospitals in providing long-term stable services after the COVID-19 pandemic [66].

Our results of prescription analysis indicated that female patients were the predominant population in our internet-based psychiatric hospital and were significantly associated with the COVID-19 pandemic phases compared to male patients. Previous studies have shown that female patients are more

XSL•FO
RenderX

susceptible to external environmental stress compared to male patients, thereby resulting in mental health problems [67-70]. Recent evidence has suggested that female patients are more likely to seek help during public health crises, particularly during the COVID-19 pandemic [19,71]. Besides, our findings showed that patients aged 18-40 years were the major population in both pandemic phases, and individuals younger than 41 years of age were significantly associated with the pandemic phases. There were several driving factors that are worth noting. First, these individuals generally encounter significant life crises during the pandemic, such as academic disruptions, risk of unemployment, and financial instability [72]. Second, previous research has shown that the depression rate among adolescents and young adults has risen sharply over the past decade [73]. Fortunately, adolescents and young adults also demonstrated greater adaptability to digital health care platforms compared to the older population, which was helpful in facing public health crises [74,75]. Interestingly, our long-term datasets across pandemic phases appeared to reflect potential demographic differences based on sex and age, which echoed the previous evidence, indicating that the pandemic led to a widening digital divide [76]. Therefore, we suggest that future policy guidance and technical support should be committed to ensure balanced allocation of digital health care resources among different populations [77].

In this study, we found that the leading primary diagnosed diseases included depressive disorder, bipolar disorder, schizophrenia, mood disorder, and anxiety disorder. This disorder spectrum not only aligned with the clinical orientation of our internet-based psychiatric hospital but also suggested that the patient populations analyzed in this study met the intended service objectives of our online platform [23]. Previous studies have reported that sudden public health emergencies can cause various psychological disorders, such as depression, mood disorder, and anxiety [78,79]. During the COVID-19 pandemic, control measures of medical isolation or home quarantine were also reported to cause significant psychological pressure [80]. Our results revealed increased proportions of depressive disorder and mood disorder in the postpandemic phase, which might support the evidence that there is a pandemic-driven shift pattern in global mental health needs [81]. Regarding drug categories, quetiapine, lithium carbonate, and escitalopram were the top 3 drugs prescribed in both pandemic phases, which was consistent with the evidence that these drugs were the most commonly applied antipsychotics and antidepressants [82].

As for the time spent during the pharmacist audit process, it was reported that pharmacist audit time was significantly compressed during the pandemic due to strategic reallocation of hospital human resources [42,83]. In the postpandemic phase, we found a significantly decreased proportion of prescriptions audited in ≤5 minutes and a significantly increased proportion of prescriptions audited within 12-24 hours. Several factors could contribute to these changes. First, the increased complexity of prescriptions for patients infected with COVID-19 could lead to an extended audit time [23]. In addition, evidence has suggested that pharmacists were already at risk of burnout before the COVID-19 pandemic [84]. Subsequently, the increased workload coupled with decreased rest time have exacerbated the burnout, which might influence the overall prescription audit time [85].

Pharmacists are considered to play a crucial role in ensuring the quality of prescription audit [86]. Pharmacist audit offers greater flexibility and professional judgment than a system automatic audit by using their clinical experience in individual conditions of patients [23]. Our results suggested an overall high prescription approval rate in both pandemic phases, revealing the critical role of pharmacist audit in promoting rational medication use in the internet-based psychiatric hospital [87]. Additionally, our results indicated a significantly increased pharmacist approval rate as well as a decreased doctor double confirmation rate in the postpandemic phase. These findings possibly reflected the increased familiarity of pharmacists and doctors with the online platform and the technological optimization of the digital system in the postpandemic phase [88].

Innovation of the Study

To the best of our knowledge, this study is the first work that applied descriptive and analytic statistical methods to evaluate the associations between different phases of the COVID-19 pandemic and the prescriptions and pharmacy services in the internet-based psychiatric hospital. The analyzed indicators included long-term prescription trends, patients' demographic characteristics, drug and disease distribution patterns, and pharmacy service indicators. Our results could provide practical experience for other medical institutions and help to promote the development of digital health service models in the future [89].

Comparisons to Existing Literature

In the field of pharmacy services in internet-based hospitals, most research has focused on the establishment and application of an artificial intelligence audit system [90,91], drug delivery [92-94], medication therapy management [95,96], and drug consultation [97,98]. To date, only a few studies have explored online pharmacy services related to the COVID-19 pandemic through prescription analysis. Among these studies, Chen et al [99] analyzed 1718 prescriptions from the online platform of a tertiary cancer hospital in China over 6 months, involving data of patients' sex and age, geographical distribution, main diagnosis, and drug category. Ding et al [49] analyzed 1380 prescriptions from the online platform of a tertiary general hospital in China over 2 months, covering data of patients' sex and age, residence, prescription department, payment of prescription, and drug delivery region. As for studies in the field of internet-based psychiatric hospitals, Xie et al [23] analyzed 2914 prescriptions over 24 months, and Du et al [21] analyzed 1259 prescriptions over 12 months. Notably, we found that the distribution patterns of patients' sex and age, diagnosed diseases, and prescribed drugs in our study were generally consistent with these 2 studies [21,23], which supported the reliability and generalizability of our results.

However, existing studies conducting prescription analysis only applied descriptive statistical methods rather than analytic statistical methods to compare the results between different

XSL•FO
RenderX

pandemic phases, limiting in-depth interpretation of the observed phenomena [21,23,49,99]. Our study addressed the limitations of existing research from several aspects. First, this study carried a larger sample size compared to previous studies, and the collection of 17,330 prescriptions strengthened the statistical power of the study [100]. Second, we provided a longer research timespan than previous studies, and the dataset over 38 months enhanced the robustness of our results [101]. Finally, by providing certainty of evidence through analytic statistical methods, including ITS analysis, bootstrap method, Pearson chi-square analysis, and multinomial logistic regression, we further explored the potential correlations between the pandemic phases and the observed outcomes [53-56].

Contribution of the Study

The main contribution of this study was to provide a fundamental practice model of telepharmacy service. First, we identified early warning indicators that can be used as key intervention steps in pharmacy services, such as extended audit time and an increased number of prescriptions [102]. Second, by using descriptive and analytic statistical approaches in the analysis of prescriptions and pharmacy services, we demonstrated replicable analytic methods that can be applied in other medical institutions, offering templates for future responses to public health crises [103]. Taken together, this study greatly enriched our understanding of pharmacy services in internet-based hospitals and provided practical experience for future multicenter research in this field [89].

Implications

Our study had direct real-world implications for optimizing the efficiency of pharmacy services in internet-based hospitals. Our data showed that prescription audit time was significantly extended in the postpandemic phase. Therefore, for hospital managers, it is recommended to establish a pharmacy personnel allocation system to timely deploy nonemergency personnel when prescription audit time increases abnormally [104]. In addition, it is important to build a reward mechanism for pharmacy services. Specifically, the efficiency of prescription audit can be promoted by implementing clear service standards and linking them to incentive funds [105]. For software developers, pharmacy service paths can be optimized for key populations. Our results suggested that the majority in internet-based psychiatric hospitals were female patients. Therefore, online psychological counseling tools can be developed for this population [106,107]. On the other hand, our results revealed the lowest proportion of older patients. Hence, in order to improve their user experience, simplified user interface and video chat tools can be added in the app [108]. Besides, we found that many prescriptions required double confirmation by doctors during the pandemic phase. To further improve doctor-pharmacist communication in this step, a real-time videoconferencing module can be added to the audit system [109,110]. Further, a local knowledge base can be built to predict prescriptions that may trigger double confirmation by doctors and convert them into pop-up notifications in the doctor-prescribing system [111].

Limitations

This study had some limitations that need to be noted. First, although our study analyzed the real-world pharmacy service model in an operating online platform of a psychiatric hospital in China, its single-center nature might limit the generalizability of our findings in rural regions or other hospitals, such as general hospitals and rural primary medical institutions [112]. Future research of multicenter design should be required to further validate our conclusions [89]. Second, due to the observational nature of our study, the findings mainly allowed us to demonstrate statistical associations rather than definite proof of causality [113]. It would be helpful for future studies to manipulate more predictors in the model to clarify potential confounding effects [114]. Third, the absence of other unmeasured factors, such as prescription cost, patients' geographic distribution, comorbidities, disease severity, and adverse events, might have influenced the depth of our findings [49,99]. The inclusion of these possible variables in future work would enrich our understanding of the internet-based hospital service models [115].

Conclusions

In summary, this study applied descriptive and analytic statistical methods to evaluate the associations between different COVID-19 phases and the prescriptions and pharmacy services in our internet-based psychiatric hospital. We found an overall upward trend in the number of prescriptions during the pandemic phase and a steady trend after the pandemic. Among patients, female and young adults were the predominant groups, and their proportions increased with the development of pandemic phases. The distribution patterns of primary diagnosed diseases and prescribed drugs were generally similar in both pandemic phases. As the pandemic phases advanced, prescription audit time was extended, and the audit approval rate was increased. This study addressed the limitations of existing research by the application of larger sample size, longer research timespan, and analytic statistical methods. This study demonstrated early warning indicators and replicable analytic methods that can be applied in other medical institutions. Our findings also had implications for hospital managers and software developers in optimizing the efficiency of pharmacy services in internet-based hospitals.

Acknowledgments

No generative artificial intelligence was used in any portion of the manuscript writing.

Data Availability

The datasets used and/or analyzed during this study are included in the paper or Multimedia Appendix 1; further inquiries are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Funding

This work was supported by the National Natural Science Foundation of China (82104223), Guangdong Basic and Applied Basic Research Foundation (2020A1515110008), Science and Technology Program of Guangzhou (202102021022, 2024A04J10001, and 2025A03J3308), Guangzhou Municipal Science and Technology Project for Medicine and Healthcare (20211A011044), and Guangzhou Municipal Key Discipline in Medicine (2025-2027). The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Authors' Contributions

YT and JH contributed to the study concept and design. DS contributed to the acquisition of data. GD and HX contributed to the statistical analysis and visualization of data and drafted the manuscript. YT and JH contributed to the writing and review of the manuscript. YW contributed to the supervision. All authors contributed to the critical revision of the final manuscript and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The completed and filled out STROBE checklist.
[DOCX File , 25 KB - jmir_v28i1e74059_app1.docx]

References

1. Onyeaka H, Anumudu CK, Al-Sharify ZT, Egele-Godswill E, Mbaegbu P. COVID-19 pandemic: a review of the global lockdown and its far-reaching effects. Sci Prog 2021;104(2):368504211019854 [FREE Full text] [doi: 10.1177/00368504211019854] [Medline: 34061685]
2. Panchal U, Salazar de Pablo G, Franco M, Moreno C, Parellada M, Arango C, et al. The impact of COVID-19 lockdown on child and adolescent mental health: systematic review. Eur Child Adolesc Psychiatry 2023;32(7):1151-1177 [FREE Full text] [doi: 10.1007/s00787-021-01856-w] [Medline: 34406494]
3. Zhu C, Zhang T, Li Q, Chen X, Wang K. Depression and anxiety during the COVID-19 pandemic: epidemiology, mechanism, and treatment. Neurosci Bull 2023;39(4):675-684 [FREE Full text] [doi: 10.1007/s12264-022-00970-2] [Medline: 36411394]
4. Alsabeeha NHM, Atieh MA, Balakrishnan MS. Older adults' satisfaction with telemedicine during the COVID-19 pandemic: a systematic review. Telemed J E Health 2023;29(1):38-49. [doi: 10.1089/tmj.2022.0045] [Medline: 35613380]
5. Kowalski K, Misiak B. Schizophrenia and the COVID-19 pandemic: a narrative review from the biomedical perspective. Span J Psychiatry Ment Health 2025;18(2):141-148. [doi: 10.1016/j.rpsm.2023.04.002] [Medline: 37544807]
6. Sunjaya AP, Jenkins C. Rationale for universal face masks in public against COVID-19. Respirology 2020;25(7):678-679 [FREE Full text] [doi: 10.1111/resp.13834] [Medline: 32353901]
7. Gaur U, Sealy W, Bharatha A, Sobers NP, Krishnamurthy K, Campbell MH, et al. Knowledge, attitudes, and practices of hand hygiene, mask use, and social distancing among public hospital and polyclinic nurses in Barbados during the coronavirus 2019 pandemic. Epidemiologia (Basel) 2024;5(1):122-136 [FREE Full text] [doi: 10.3390/epidemiologia5010008] [Medline: 38534805]
8. Li P, Wen L, Sun B, Sun W, Chen H. Retrospective estimation of the time-varying effective reproduction number for a COVID-19 outbreak in Shenyang, China: an observational study. Medicine (Baltimore) 2024;103(22):e38373 [FREE Full text] [doi: 10.1097/MD.0000000000038373] [Medline: 39259088]
9. De Hert M, Mazereel V, Detraux J, Van Assche K. Prioritizing COVID-19 vaccination for people with severe mental illness. World Psychiatry 2021;20(1):54-55 [FREE Full text] [doi: 10.1002/wps.20826] [Medline: 33131217]
10. Ahuvia IL, Sung JY, Dobias ML, Nelson BD, Richmond LL, London B, et al. College student interest in teletherapy and self-guided mental health supports during the COVID-19 pandemic. J Am Coll Health 2024;72(3):940-946. [doi: 10.1080/07448481.2022.2062245] [Medline: 35427460]
11. Smith KA, Ostinelli EG, Ede R, Allard L, Thomson M, Hewitt K, et al. Assessing the impact of evidence-based mental health guidance during the COVID-19 pandemic: systematic review and qualitative evaluation. JMIR Ment Health 2023;10:e52901 [FREE Full text] [doi: 10.2196/52901] [Medline: 38133912]
12. Kola L, Kohrt BA, Hanlon C, Naslund JA, Sikander S, Balaji M, et al. COVID-19 mental health impact and responses in low-income and middle-income countries: reimagining global mental health. Lancet Psychiatry 2021;8(6):535-550 [FREE Full text] [doi: 10.1016/S2215-0366(21)00025-0] [Medline: 33639109]
13. Chai Y, Lam ICH, Man KKC, Hayes JF, Wan EYF, Li X, et al. Psychiatric and neuropsychiatric sequelae of COVID-19 within 2 years: a multinational cohort study. BMC Med 2025;23(1):144 [FREE Full text] [doi: 10.1186/s12916-025-03952-z] [Medline: 40055683]

XSL•FO
RenderX

14. Liu Y, Ren Y, Liu C, Chen X, Li D, Peng J, et al. Global burden of mental disorders in children and adolescents before and during the COVID-19 pandemic: evidence from the Global Burden of Disease Study 2021. Psychol Med 2025;55:e90. [doi: 10.1017/s0033291725000649]
15. Baidoobonso S, Etowa E, Nnadi J, Mba S, Tharao W, Daboné C, et al. The impact of the COVID-19 pandemic on the mental health of African, Caribbean, and Black (ACB) people in Canada. J Immigr Minor Health 2025;27(1):42-52. [doi: 10.1007/s10903-024-01654-x] [Medline: 39612108]
16. Ahmed N, Barnett P, Greenburgh A, Pemovska T, Stefanidou T, Lyons N, et al. Mental health in Europe during the COVID-19 pandemic: a systematic review. Lancet Psychiatry 2023;10(7):537-556 [FREE Full text] [doi: 10.1016/S2215-0366(23)00113-X] [Medline: 37321240]
17. Lee S, Shin HE, Lee R, Jo S, Lee SC, Cheong H, et al. A brief report of the economic burden and epidemiological finding of suicide attempts among the older adults in Korea from 2007 to 2021. BMC Geriatr 2025;25(1):147 [FREE Full text] [doi: 10.1186/s12877-025-05793-0] [Medline: 40045237]
18. Grosman-Dziewiszek P, Wiatrak B, Jęśkowiak I, Szeląg A. Patients' habits and the role of pharmacists and telemedicine as elements of a modern health care system during the COVID-19 pandemic. J Clin Med 2021;10(18):4211 [FREE Full text] [doi: 10.3390/jcm10184211] [Medline: 34575322]
19. Lau BHP, Tang CSK, Holroyd E, Wong WCW. Challenges and implications for menopausal health and help-seeking behaviors in midlife women from the United States and China in light of the COVID-19 pandemic: web-based panel surveys. JMIR Public Health Surveill 2024;10:e46538 [FREE Full text] [doi: 10.2196/46538] [Medline: 38277194]
20. Weinberg V, van Haren E, Gombert-Handoko KB. Improving patient access to hospital pharmacists using eConsults: retrospective descriptive study. JMIR Form Res 2023;7:e38518 [FREE Full text] [doi: 10.2196/38518] [Medline: 36705957]
21. Du Z, Jiang Y, Lu R, Zhou Q, Pan Y, Shen Y, et al. Practice of pharmaceutical services and prescription analysis in internet-based psychiatric hospitals during COVID-19 pandemic in Wuxi, China. Front Psychiatry 2023;14:1195298 [FREE Full text] [doi: 10.3389/fpsyt.2023.1195298] [Medline: 37547208]
22. Ge F, Qian H, Lei J, Ni Y, Li Q, Wang S, et al. Experiences and challenges of emerging online health services combating COVID-19 in China: retrospective, cross-sectional study of internet hospitals. JMIR Med Inform 2022;10(6):e37042 [FREE Full text] [doi: 10.2196/37042] [Medline: 35500013]
23. Xie W, Wang F, Qian Y, Qiu L, Zhou Q, Shen Y, et al. "Internet+" pharmacy in psychiatric hospital amid COVID-19 and post-pandemic period: analysis and development. Front Psychiatry 2024;15:1434966 [FREE Full text] [doi: 10.3389/fpsyt.2024.1434966] [Medline: 39758438]
24. Cheng TC, Yip W. Policies, progress, and prospects for internet telemedicine in China. Health Syst Reform 2024;10(2):2389570 [FREE Full text] [doi: 10.1080/23288604.2024.2389570] [Medline: 39437237]
25. Wang K, Zou W, Lai Y, Hao C, Liu N, Ling X, et al. Accessibility, cost, and quality of an online regular follow-up visit service at an internet hospital in China: mixed methods study. J Med Internet Res 2024;26:e54902 [FREE Full text] [doi: 10.2196/54902] [Medline: 39432365]
26. Bilionis I, Quer Palomas S, Vidal-Alaball J. Digital health for tracking long-covid symptoms with data insights from mobile app questionnaires and wearable devices. Stud Health Technol Inform 2025;323:434-438. [doi: 10.3233/SHTI250127] [Medline: 40200524]
27. Zhi L, Yin P, Ren J, Wei G, Zhou J, Wu J, et al. Running an internet hospital in China: perspective based on a case study. J Med Internet Res 2021;23(9):e18307 [FREE Full text] [doi: 10.2196/18307] [Medline: 34342267]
28. Fu L, Wang R, Dong Y. The impact of the hierarchical medical system on medical resource allocation in China. Sci Rep 2025;15(1):7561 [FREE Full text] [doi: 10.1038/s41598-025-88558-4] [Medline: 40038294]
29. Qingjie P, Wanyu H, Jingying Y, Ming Y, Linlin W, Qiyue M, et al. How to promote telemedicine in underdeveloped cities in central China? Qualitative interviews with medical personnel in tertiary A-level hospital of Xinyang City. Digit Health 2024;10:20552076241277670 [FREE Full text] [doi: 10.1177/20552076241277670] [Medline: 39247099]
30. Gao J, Fan C, Chen B, Fan Z, Li L, Wang L, et al. Telemedicine is becoming an increasingly popular way to resolve the unequal distribution of healthcare resources: evidence from China. Front Public Health 2022;10:916303 [FREE Full text] [doi: 10.3389/fpubh.2022.916303] [Medline: 35874991]
31. Yu J, Cheong IH, Kozlakidis Z, Wang H. Advancements and challenges of artificial intelligence in dermatology: a review of applications and perspectives in China. Front Digit Health 2025;7:1544520 [FREE Full text] [doi: 10.3389/fdgth.2025.1544520] [Medline: 40881732]
32. Ndwabe H, Basu A, Mohammed J. Post pandemic analysis on comprehensive utilization of telehealth and telemedicine. Clinical eHealth 2024;7:5-14. [doi: 10.1016/j.ceh.2023.12.002]
33. Kim J, Cai ZR, Chen ML, Onyeka S, Ko JM, Linos E. Telehealth utilization and associations in the United States during the third year of the COVID-19 pandemic: population-based survey study in 2022. JMIR Public Health Surveill 2024;10:e51279 [FREE Full text] [doi: 10.2196/51279] [Medline: 38669075]
34. Cao Y, Li Q, Chen J, Guo X, Miao C, Yang H, et al. Hospital emergency management plan during the COVID-19 epidemic. Acad Emerg Med 2020;27(4):309-311 [FREE Full text] [doi: 10.1111/acem.13951] [Medline: 32124506]

XSL•FO
RenderX

35. Juang WC, Chiou SM, Chen HC, Li YC. Differences in characteristics and length of stay of elderly emergency patients before and after the outbreak of COVID-19. Int J Environ Res Public Health 2023;20(2):1162 [FREE Full text] [doi: 10.3390/ijerph20021162] [Medline: 36673917]
36. Han S, Li L. Consulting doctors online after offline treatment: investigating the effects of online information on patients' effective use of online follow-up services. Front Public Health 2024;12:1375144 [FREE Full text] [doi: 10.3389/fpubh.2024.1375144] [Medline: 38655527]
37. Bates DW, Cheng HY, Cheung NT, Jew R, Mir F, Tamblyn R, et al. 'Improving smart medication management': an online expert discussion. BMJ Health Care Inform 2022;29(1):e100540 [FREE Full text] [doi: 10.1136/bmjhci-2021-100540] [Medline: 35477691]
38. Zeng B, Rivadeneira NA, Wen A, Sarkar U, Khoong EC. The impact of the COVID-19 pandemic on internet use and the use of digital health tools: secondary analysis of the 2020 Health Information National Trends Survey. J Med Internet Res 2022;24(9):e35828 [FREE Full text] [doi: 10.2196/35828] [Medline: 36041005]
39. Lamba S, Li S, Lamba A, Tang F, Valencia WM, Dang S. Differences in challenges to using telehealth among older adult video and telephone users with frailty: retrospective observational study. J Med Internet Res 2025;27:e69437-e69437 [FREE Full text] [doi: 10.2196/69437] [Medline: 40882213]
40. Faro JM, Obermiller E, Obermiller C, Trinkley KE, Wright G, Sadasivam RS, et al. Using routinely available electronic health record data elements to develop and validate a digital divide risk score. JAMIA Open 2025;8(1):ooaf004. [doi: 10.1093/jamiaopen/ooaf004] [Medline: 39906363]
41. Jiménez-Zarco A, Mateos SC, Bosque-Prous M, Espelt A, Torrent-Sellens J, Adib K, et al. Impact of the COVID-19 pandemic on mHealth adoption: identification of the main barriers through an international comparative analysis. Int J Med Inform 2025;195:105779 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105779] [Medline: 39813967]
42. Li Y, Hu H, Rozanova L, Fabre G. COVID-19 and internet hospital development in China. Epidemiologia (Basel) 2022;3(2):269-284 [FREE Full text] [doi: 10.3390/epidemiologia3020021] [Medline: 36417257]
43. Tran J, Hao Z, Xie Z, Yiu K, Li A, Tsoi K. Nationwide trends of online consultations in China: a 10-year sequential cross-sectional study of 65 305 clinicians. BMJ Public Health 2025;3(2):e002296. [doi: 10.1136/bmjph-2024-002296] [Medline: 40697407]
44. Wu X, Kuang Y, Guo Y, Wu J, Xiao L. Internet hospital response to the COVID-19 pandemic in a tertiary hospital in China: perspectives based on a mixed-methods. Digit Health 2024;10:20552076241228418 [FREE Full text] [doi: 10.1177/20552076241228418] [Medline: 38303968]
45. Hagi K, Kurokawa S, Takamiya A, Fujikawa M, Kinoshita S, Iizuka M, et al. Telepsychiatry versus face-to-face treatment: systematic review and meta-analysis of randomised controlled trials. Br J Psychiatry 2023;223(3):407-414 [FREE Full text] [doi: 10.1192/bjp.2023.86] [Medline: 37655816]
46. Ambrósio GC, Dal Paz K, Moreira de Souza VA, Discacciatii MG, Campa A. Impact of the COVID-19 pandemic on pharmacist interventions: a retrospective study with inpatients in a university hospital. Can J Hosp Pharm 2024;77(3):e3514. [doi: 10.4212/cjhp.3514] [Medline: 38988873]
47. Yu T, Li W, Liu Y, Jin C, Wang Z, Cao H. Application of internet hospitals in the disease management of patients with ulcerative colitis: retrospective study. J Med Internet Res 2025;27:e60019 [FREE Full text] [doi: 10.2196/60019] [Medline: 40101745]
48. Wang L, Liang D, Huang Fu H, Ke C, Wu S, Lai Y. Enterprise-led internet healthcare provision in China: insights from a leading platform. Front Digit Health 2025;7:1491183 [FREE Full text] [doi: 10.3389/fdgth.2025.1491183] [Medline: 40099033]
49. Ding L, She Q, Chen F, Chen Z, Jiang M, Huang H, et al. The internet hospital plus drug delivery platform for health management during the COVID-19 pandemic: observational study. J Med Internet Res 2020;22(8):e19678 [FREE Full text] [doi: 10.2196/19678] [Medline: 32716892]
50. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 2007;147(8):573-577 [FREE Full text] [doi: 10.7326/0003-4819-147-8-200710160-00010] [Medline: 17938396]
51. Levitt HM, Bamberg M, Creswell JW, Frost DM, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018;73(1):26-46 [FREE Full text] [doi: 10.1037/amp0000151] [Medline: 29345485]
52. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
53. Antonazzo IC, Fornari C, Maumus-Robert S, Cei E, Paoletti O, Ferrara P, et al. Antidepressants drug use during COVID-19 waves in the Tuscan general population: an interrupted time-series analysis. J Pers Med 2022;12(2):1639-1641 [FREE Full text] [doi: 10.3390/jpm12020178] [Medline: 35207666]

XSL•FO
RenderX

54. Chan VKY, Chai Y, Chan SSM, Luo H, Jit M, Knapp M, et al. Impact of COVID-19 pandemic on depression incidence and healthcare service use among patients with depression: an interrupted time-series analysis from a 9-year population-based study. BMC Med 2024;22(1):169 [FREE Full text] [doi: 10.1186/s12916-024-03386-z] [Medline: 38644506]
55. Xia H, Deng G, Liang F, Zhang Z, Huang W, Guo Z, et al. Investigating remedial strategies for missed or delayed dose of sertraline in Chinese adolescent patients with depressive disorders via population pharmacokinetics modeling and simulation approaches. Drug Des Devel Ther 2025;19:3001-3016 [FREE Full text] [doi: 10.2147/DDDT.S504521] [Medline: 40260198]
56. Tsamardinos I, Greasidou E, Borboudakis G. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. Mach Learn 2018;107(12):1895-1922 [FREE Full text] [doi: 10.1007/s10994-018-5714-4] [Medline: 30393425]
57. Wen Q, Yang B, Wang X, Xia L, Song L. Characteristics analysis of internet pharmacy consultation services for children in southwest China during the post-epidemic era: a cross-sectional study. Int J Med Inform 2024;186:105424 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105424] [Medline: 38547627]
58. Karattuthodi MS, Krishnan G, Aiswarya A, Sarath CC, Nair SS, Thorakkatil SA. A systematic review on pharmacist-led pharmaceutical care optimization through virtual platforms. J Pharm Health Serv Res 2023;14(1):2-13. [doi: 10.1093/jphsr/rmac054]
59. Hursman A, Vang C, Thooft T, Stone K. The role of telepharmacy in the delivery of clinical pharmacy services following the COVID-19 pandemic: a descriptive report. J Pharm Technol 2024;40(2):66-71. [doi: 10.1177/87551225231222426] [Medline: 38525089]
60. Lu J, Bai J, Zhao H, Zhang X. The effect of "offline-to-online" trust transfer on the utilization of online medical consultation among Chinese rural residents: experimental study. J Med Internet Res 2023;25:e43430 [FREE Full text] [doi: 10.2196/43430] [Medline: 38147370]
61. Zhang M, Xiao J, Deng A, Zhang Y, Zhuang Y, Hu T, et al. Transmission dynamics of an outbreak of the COVID-19 Delta Variant B.1.617.2—Guangdong Province, China, May-June 2021. China CDC Wkly 2021;3(27):584-586 [FREE Full text] [doi: 10.46234/ccdcw2021.148] [Medline: 34594941]
62. Chen Y, Liu Y, Yan Y. Revealing the spatiotemporal characteristics of the general public's panic levels during the pandemic crisis in China. Trans GIS 2022:tgis.13016 [FREE Full text] [doi: 10.1111/tgis.13016] [Medline: 36721464]
63. Yang Y, Zhang Y, Zhang X, Cao Y, Zhang J. Spatial evolution patterns of public panic on Chinese social networks amidst the COVID-19 pandemic. Int J Disaster Risk Reduct 2022;70:102762 [FREE Full text] [doi: 10.1016/j.ijdrr.2021.102762] [Medline: 35004139]
64. Huang J, Zhao S, Chong KC, Zhou Y, Lu W, Fang F, et al. Infection rate in guangzhou after easing the zero-COVID policy: seroprevalence results to ORF8 antigen. Lancet Infect Dis 2023;23(4):403-404 [FREE Full text] [doi: 10.1016/s1473-3099(23)00112-3]
65. Di Carlo F, Sociali A, Picutti E, Pettorruso M, Vellante F, Verrastro V, et al. Telepsychiatry and other cutting-edge technologies in COVID-19 pandemic: bridging the distance in mental health assistance. Int J Clin Pract 2021;75(1):10.1111/ijcp.13716 [FREE Full text] [doi: 10.1111/ijcp.13716] [Medline: 32946641]
66. Pan X, Zhou X, Yu L, Hou L. Switching from offline to online health consultation in the post-pandemic era: the role of perceived pandemic risk. Front Public Health 2023;11:1121290 [FREE Full text] [doi: 10.3389/fpubh.2023.1121290] [Medline: 37261233]
67. Min C, Shen F, Yu W, Chu Y. The relationship between government trust and preventive behaviors during the COVID-19 pandemic in China: exploring the roles of knowledge and negative emotion. Prev Med 2020;141:106288 [FREE Full text] [doi: 10.1016/j.ypmed.2020.106288] [Medline: 33091414]
68. Rossell SL, Neill E, Phillipou A, Tan EJ, Toh WL, Van Rheenen TE, et al. An overview of current mental health in the general population of Australia during the COVID-19 pandemic: results from the COLLATE project. Psychiatry Res 2021;296:113660 [FREE Full text] [doi: 10.1016/j.psychres.2020.113660] [Medline: 33373808]
69. Duan L, Shao X, Wang Y, Huang Y, Miao J, Yang X, et al. An investigation of mental health status of children and adolescents in China during the outbreak of COVID-19. J Affect Disord 2020;275:112-118 [FREE Full text] [doi: 10.1016/j.jad.2020.06.029] [Medline: 32658812]
70. Liu S, Yang L, Zhang C, Xu Y, Cai L, Ma S, et al. Gender differences in mental health problems of healthcare workers during the coronavirus disease 2019 outbreak. J Psychiatr Res 2021;137:393-400 [FREE Full text] [doi: 10.1016/j.jpsychires.2021.03.014] [Medline: 33765451]
71. Wu X, Wong KW, Gurvich C, Dong Y. Impact of COVID-19 on health of menopausal women: a scoping review. Gen Hosp Psychiatry 2023;84:125-141. [doi: 10.1016/j.genhosppsych.2023.07.001] [Medline: 37473503]
72. Cook JA, Jonikas JA, Burke-Miller JK, Aranda F, Mullen MG, Davis M, et al. Social determinants, mental well-being, and disrupted life transitions among young adults with disabling mental health conditions. J Behav Health Serv Res 2025;52(2):263-281. [doi: 10.1007/s11414-024-09924-0] [Medline: 39806229]
73. Thapar A, Eyre O, Patel V, Brent D. Depression in young people. Lancet 2022;400(10352):617-631. [doi: 10.1016/S0140-6736(22)01012-1] [Medline: 35940184]

XSL•FO
RenderX

74. Schultz AN, Honnens S, Bauer EH, Söderberg K, Kidholm K, Christensen R, et al. Patients' preferences for virtual outpatient consultations and health care professionals' assessment of suitability for their patients: a single-centre survey. J Telemed Telecare 2024;10:1357633x241294125. [doi: 10.1177/1357633x241294125]
75. Hamilton JL, Nesi J, Choukas-Bradley S. Reexamining social media and socioemotional well-being among adolescents through the lens of the COVID-19 pandemic: a theoretical review and directions for future research. Perspect Psychol Sci 2022;17(3):662-679 [FREE Full text] [doi: 10.1177/17456916211014189] [Medline: 34756118]
76. Qiu Y, Huang H, Gai J, De Leo G. The effects of the COVID-19 pandemic on age-based disparities in digital health technology use: secondary analysis of the 2017-2022 Health Information National Trends Survey. J Med Internet Res 2024;26:e65541 [FREE Full text] [doi: 10.2196/65541] [Medline: 39631070]
77. Tay CT, Joham AE, Teede HJ. Key standards and principles for developing evidence-based clinical guidelines: balancing health professional, patient, funder, and government needs. Fertil Steril 2025;123(4):561-568. [doi: 10.1016/j.fertnstert.2025.01.023] [Medline: 39884335]
78. Tng XJJ, Chew QH, Sim K. Psychological sequelae within different populations during the COVID-19 pandemic: a rapid review of extant evidence. Singapore Med J 2022;63(5):229-235 [FREE Full text] [doi: 10.11622/smedj.2020111] [Medline: 36043286]
79. Rogers M, Windle A, Wu L, Taylor V, Bale C. Emotional well-being, spiritual well-being and resilience of advanced clinical practitioners in the United Kingdom during COVID-19: an exploratory mixed method study. J Nurs Manag 2022;30(4):883-891 [FREE Full text] [doi: 10.1111/jonm.13577] [Medline: 35261097]
80. Lestari R, Setyawan FEB. Mental health policy: protecting community mental health during the COVID-19 pandemic. J Public Health Res 2021;10(2):2231 [FREE Full text] [doi: 10.4081/jphr.2021.2231] [Medline: 33855426]
81. Xiao Y, Chen TT, Liu L, Zong L. China ends its zero-COVID-19 policy: new challenges facing mental health services. Asian J Psychiatr 2023;82:103485. [doi: 10.1016/j.ajp.2023.103485] [Medline: 36731222]
82. Solmi M, Miola A, Capone F, Pallottino S, Højlund M, Firth J, et al. Risk factors, prevention and treatment of weight gain associated with the use of antidepressants and antipsychotics: a state-of-the-art clinical review. Expert Opin Drug Saf 2024;23(10):1249-1269. [doi: 10.1080/14740338.2024.2396396] [Medline: 39225182]
83. Dettori M, Castiglia P. COVID-19 and digital health: evolution, perspectives and opportunities. Int J Environ Res Public Health 2022;19(14):8519 [FREE Full text] [doi: 10.3390/ijerph19148519] [Medline: 35886366]
84. Durham ME, Bush PW, Ball AM. Evidence of burnout in health-system pharmacists. Am J Health Syst Pharm 2018;75(23 Supplement 4):S93-S100. [doi: 10.2146/ajhp170818] [Medline: 30333113]
85. Abdelsadig Mohammed H, Elamin SA, El-Awaisi A, El Hajj MS. Use of the job demands-resource model to understand community pharmacists' burnout during the COVID-19 pandemic. Res Social Adm Pharm 2022;18(9):3568-3579 [FREE Full text] [doi: 10.1016/j.sapharm.2022.03.011] [Medline: 35418325]
86. De Vries E, Johnson Y, Willems B, Bedeker W, Ras T, Coetzee R, et al. Improving primary care antimicrobial stewardship by implementing a peer audit and feedback intervention in Cape Town community healthcare centres. S Afr Med J 2022;112(10):812-818 [FREE Full text] [doi: 10.7196/SAMJ.2022.v112i10.16397] [Medline: 36472332]
87. Mujtaba SH, Gazerani P. Exploring the role of community pharmacists in pain management: enablers and challenges. Pharmacy (Basel) 2024;12(4):111 [FREE Full text] [doi: 10.3390/pharmacy12040111] [Medline: 39051395]
88. Patrocínio WP, Pimenta MA. Information and communication technology: strategies to optimize hospital services. Rev Gest Soc Ambient 2025;19(1):1-17. [doi: 10.24857/rgsa.v19n1-021]
89. Das MK. Multicenter studies: relevance, design and implementation. Indian Pediatr 2022;59(7):571-579. [doi: 10.1007/s13312-022-2561-y]
90. Zheng Y, Rowell B, Chen Q, Kim JY, Kontar RA, Yang XJ, et al. Designing human-centered AI to prevent medication dispensing errors: focus group study with pharmacists. JMIR Form Res 2023;7:e51921 [FREE Full text] [doi: 10.2196/51921] [Medline: 38145475]
91. Fourkiotis KP, Tsadiras A. Future internet applications in healthcare: big data-driven fraud detection with machine learning. Future Internet 2025;17(10):460. [doi: 10.3390/fi17100460]
92. Sham R, Siau CS, Tan S, Kiu DC, Sabhi H, Thew HZ, et al. Drone usage for medicine and vaccine delivery during the COVID-19 pandemic: attitude of health care workers in rural medical centres. Drones 2022;6(5):109. [doi: 10.3390/drones6050109]
93. Mukati N, Namdev N, Dilip R, Hemalatha N, Dhiman V, Sahu B. Healthcare assistance to COVID-19 patient using internet of things (IoT) enabled technologies. Mater Today Proc 2023;80:3777-3781 [FREE Full text] [doi: 10.1016/j.matpr.2021.07.379] [Medline: 34336599]
94. Hammour KA, Abdeljalil M, Manaseer Q, Al-Manaseer B. Jordanian experience: the internet pharmacy drug delivery platform during the COVID-19. Health Policy Technol 2022;11(1):100596 [FREE Full text] [doi: 10.1016/j.hlpt.2022.100596] [Medline: 35024326]
95. Zhang P, Yang M, He S, Li X, Lang B, Zeng L, et al. Evaluating the impact of the COVID-19 pandemic on telepharmaceutical service effectiveness: systematic review and meta-analysis. J Med Internet Res 2025;27:e64073 [FREE Full text] [doi: 10.2196/64073] [Medline: 40601374]

XSL•FO
RenderX

96. Zhao X, Huang H, Zeng G, Shi Q, Zhu P, Zhang L, et al. Research on the online service mechanism of internet hospital in infectious disease prevention and control. Exp Biol Med (Maywood) 2025;250:10349 [FREE Full text] [doi: 10.3389/ebm.2025.10349] [Medline: 40351479]
97. Bu F, Sun H, Li L, Tang F, Zhang X, Yan J, et al. Artificial intelligence-based internet hospital pharmacy services in China: perspective based on a case study. Front Pharmacol 2022;13:1027808 [FREE Full text] [doi: 10.3389/fphar.2022.1027808] [Medline: 36438784]
98. Li H, Zheng S, Li D, Jiang D, Liu F, Guo W, et al. The establishment and practice of pharmacy care service based on internet social media: telemedicine in response to the COVID-19 pandemic. Front Pharmacol 2021;12:707442 [FREE Full text] [doi: 10.3389/fphar.2021.707442] [Medline: 34658854]
99. Chen ZJ, Liang W, Liu QT, He R, Chen QC, Li QF, et al. Use of a remote oncology pharmacy service platform for patients with cancer during the COVID-19 pandemic: implementation and user acceptance evaluation. J Med Internet Res 2021;23(1):e24619 [FREE Full text] [doi: 10.2196/24619] [Medline: 33395398]
100. Ahmed SK. How to choose a sampling technique and determine sample size for research: a simplified guide for researchers. Oral Oncol Rep 2024;12:100662 [FREE Full text] [doi: 10.1016/j.oor.2024.100662]
101. Li X, Hao J, Han Q, Wang D, Lu Y, Tu J, et al. Triglyceride-glucose index prediction of stroke incidence risk in low-income Chinese population: a 10-year prospective cohort study. Front Endocrinol (Lausanne) 2024;15:1444030 [FREE Full text] [doi: 10.3389/fendo.2024.1444030] [Medline: 39493774]
102. Baptista R, Williams M, Price J. Improving the impact of pharmacy interventions in hospitals. BMJ Open Qual 2023;12(4):e002276 [FREE Full text] [doi: 10.1136/bmjoq-2023-002276] [Medline: 37940333]
103. Hazari A. Data analysis: descriptive and analytical statistics. In: Research Methodology for Allied Health Professionals: A Comprehensive Guide to Thesis & Dissertation. Cham: Springer; 2024:79-98.
104. Hammad M, Tashtoush MM, El Faki EMA, Hajaj MY, Ahmed SS, Darwish AA. Hospital pharmacy response to the COVID-19 pandemic: experience from a regional referral center in the United Arab Emirates. J Pharm Policy Pract 2023;16(1):36 [FREE Full text] [doi: 10.1186/s40545-023-00542-3] [Medline: 36864489]
105. Hussain R, Babar ZUD. Global landscape of community pharmacy services remuneration: a narrative synthesis of the literature. J Pharm Policy Pract 2023;16(1):118 [FREE Full text] [doi: 10.1186/s40545-023-00626-0] [Medline: 37814349]
106. Andreu‐Pejó L, Martínez‐Borba V, Osma López J, Suso‐Ribera C, Crespo Delgado E. Perinatal mental e‐health: what is the profile of pregnant women interested in online assessment of their emotional state? Nursing Open 2022;10(2):901-914. [doi: 10.1002/nop2.1358] [Medline: 36068679]
107. Progga FT, Rubya S. Women's perspectives and challenges in adopting perinatal mental health technologies. Proc ACM Hum-Comput Interact 2025;9(1):1-30. [doi: 10.1145/3701217]
108. Kelly RM, Xing Y, Baker S, Waycott J. Video calls as a replacement for family visits during lockdowns in aged care: interview study with family members. JMIR Aging 2023;6:e40953 [FREE Full text] [doi: 10.2196/40953] [Medline: 37191951]
109. Allison A, Shahan J, Goodner J, Smith L, Sweet C. Providing essential clinical pharmacy services during a pandemic: virtual video rounding and precepting. Am J Health Syst Pharm 2021;78(17):1556-1558 [FREE Full text] [doi: 10.1093/ajhp/zxab208] [Medline: 33987637]
110. Sarasmita MA, Sudarma IW, Jaya MKA, Irham LM, Susanty S. Telepharmacy implementation to support pharmaceutical care services during the COVID-19 pandemic: a scoping review. Can J Hosp Pharm 2024;77(1):e3430. [doi: 10.4212/cjhp.3430] [Medline: 38204502]
111. Yue XH, Yang L, Zhong JJ, Liu HM, Wang D, Tao X, et al. Optimization and impact of an evidence-based pre-audit prescription decision system in primary healthcare settings. Front Pharmacol 2025;16:1491810 [FREE Full text] [doi: 10.3389/fphar.2025.1491810] [Medline: 40297139]
112. Kim YR, Chung SW, Kim M, Choi W, Choi J, Lee D, et al. Limited generalizability of retrospective single-center cohort study in comparison to multicenter cohort study on prognosis of hepatocellular carcinoma. J Hepatocell Carcinoma 2024;11:1235-1249. [doi: 10.2147/JHC.S456093] [Medline: 38974017]
113. Haber NA, Wieten SE, Rohrer JM, Arah OA, Tennant PWG, Stuart EA, et al. Causal and associational language in observational health research: a systematic evaluation. Am J Epidemiol 2022;191(12):2084-2097 [FREE Full text] [doi: 10.1093/aje/kwac137] [Medline: 35925053]
114. Sendor R, Stürmer T. Core concepts in pharmacoepidemiology: confounding by indication and the role of active comparators. Pharmacoepidemiol Drug Saf 2022;31(3):261-269 [FREE Full text] [doi: 10.1002/pds.5407] [Medline: 35019190]
115. Conti V, Sellitto C, Torsiello M, Manzo V, De Bellis E, Stefanelli B, et al. Identification of drug interaction adverse events in patients with COVID-19: a systematic review. JAMA Netw Open 2022;5(4):e227970 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.7970] [Medline: 35438752]

Abbreviations

IRB: institutional review board
ITS: interrupted time series

XSL•FO
RenderX

JARS: Journal Article Reporting Standards
OR: odds ratio
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 18.Mar.2025; peer-reviewed by N Liu, Ş Demiralay; comments to author 24.Sep.2025; accepted 06.Feb.2026; published 04.Mar.2026.

Please cite as:
Deng G, Xia H, Shang DW, Wen Y, Hu J, Tan Y
Comparative Analysis of Prescriptions and Pharmacy Services in Internet-Based Psychiatric Hospital During and After the COVID-19 Pandemic: Retrospective Cross-Sectional Observational Study
J Med Internet Res 2026;28:e74059
URL: https://www.jmir.org/2026/1/e74059
doi:10.2196/74059
PMID:

©Guowei Deng, Hui Xia, De-wei Shang, Yuguan Wen, Jinqing Hu, Yaqian Tan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Time Savings Through an AI Speech Assistant for Nursing Documentation: Pre-Post Time-Motion Study in German Long-Term Care

Katja Schwabe[1*], BSc; Drin Ferizaj[2*], MSc; Susann Neumann[2], DipNEd; Sandra Strube-Lahmann[2], PhD; Nils Lahmann[2], Prof Dr

[1]voize GmbH, Dircksenstraße 47, Berlin, Germany

[2]Department of Geriatrics and Medical Gerontology, Charité – Universitätsmedizin Berlin, Berlin, Germany

[*]these authors contributed equally

Corresponding Author:
Katja Schwabe, BSc
voize GmbH, Dircksenstraße 47, Berlin, Germany

Abstract

Background: Nurses in long-term care spend up to one-third of their working time on documentation, contributing to administrative burden and limited time for direct care. Artificial intelligence (AI) speech assistants have shown potential to accelerate documentation, but longitudinal evidence from real-world long-term care settings remains scarce.

Objective: This study aimed to evaluate whether implementing a domain-specific, mobile AI speech assistant is associated with reduced documentation time in German long-term care under routine conditions. Secondary objectives included examining usability, perceived documentation effort, interruptions, and workplace satisfaction.

Methods: A pre-post time-motion study with full-shift observation was conducted. Continuous, event-based observations were performed before (t_0) and after (t_1) implementation of the mobile speech assistant voize. The primary outcome was total documentation time per morning shift based on direct observations. In addition to observations, questionnaires were administered to assess perceived documentation effort, interruptions, satisfaction with the documentation system, and workplace satisfaction. The primary end point was analyzed using a linear mixed-effects model. Secondary, self-reported outcomes were analyzed exploratorily via paired pre-post differences with pooling across multiple imputations and Holm-Bonferroni correction.

Results: A total of 52 registered nurses from 14 long-term care facilities participated (mean age 42.37, SD 12.37 years; 42/52, 80.8% female). Across 770 observed hours, the observed total documentation time per morning shift decreased significantly by an adjusted mean of 15 (SE 3.36) minutes, $t_{46.29}$=−4.46, P<.001, with a 95% CI of −21.75 to −8.23, corresponding to an approximately 28% reduction relative to the baseline mean. Holm-Bonferroni–corrected exploratory analyses indicated significant declines in self-reported documentation time and interruptions, and satisfaction with the documentation system improved, while workplace satisfaction showed no significant change. Usability was rated as acceptable.

Conclusions: This study provides real-world evidence from a single-group pre-post design that an AI-based speech assistant is associated with reduced documentation workload in long-term care. In this sample, the integration of a mobile, domain-specific speech system into daily workflows coincided with substantially decreased documentation time and improved perceived efficiency. Beyond these observed time savings, such technology has the potential to alleviate workload, free time for resident care, and enhance working conditions. These findings are also relevant for policy discussions on addressing the nursing workforce shortage, showing that well-integrated, speech-enabled documentation systems can support more sustainable long-term care environments.

Trial Registration: German Clinical Trials Register DRKS00035512; https://drks.de/search/en/trial/DRKS00035512

(J Med Internet Res 2026;28:e86078) doi:10.2196/86078

KEYWORDS

nursing documentation time; automatic speech recognition; long-term care; artificial intelligence; AI speech assistant; mobile nursing documentation

XSL•FO
RenderX

Introduction

Background

In Germany, approximately 1.7 million nursing staff are employed, with nearly 30% working in long-term care facilities, making this the largest sector in the nursing workforce [1]. Demand for care continues to rise while shortages of qualified staff worsen [2]. Nursing staff, particularly in long-term care, report higher levels of strain compared to other occupational groups [3,4]. They work under considerable time pressure, often lack sufficient time to complete tasks, and frequently perform overtime with irregular or missed breaks [3,5]. Alongside these structural challenges and workload demands, care workers face a high administrative burden: documentation tasks account for about 30% of total working time [6,7], leaving less room for direct interaction and care for residents [8]. Furthermore, the high documentation workload has been linked to reduced job satisfaction and an elevated level of stress [9]. In light of these issues, artificial intelligence (AI)–based speech assistants are presented as promising solutions to accelerate the documentation process and improve its accuracy [10].

Medical Documentation With Speech Recognition

Research on speech recognition has been most widely conducted in clinical medicine, often involving physicians and mixed professional groups. Previous findings show that speech recognition technologies can significantly streamline documentation workflows compared to traditional methods. For example, Zuchowski and Göller [11] reported notable time savings and lower error rates when using speech recognition versus manual typing. Similarly, Peine et al [12] found that a voice-based system enabled faster and more accurate documentation among intensive care professionals compared to both electronic and paper-based approaches.

These findings are further supported by systematic reviews. Alboksmaty et al [13] synthesized findings across several countries, highlighting consistent efficiency gains, reduced administrative burden, and improved provider-patient interaction. Of particular note, Wang et al [14] found that a digital scribe could accelerate history-taking and physical examination documentation by more than double compared to typing or traditional dictation.

However, reviews also point out that most studies have taken place in controlled or experimental settings, which may limit generalizability to routine care [10,13]. In this context, Ng et al [15] reviewed 29 research articles and observed that many reported time savings and improvements in documentation completeness, whereas others described challenges such as increased editing workload, variable cost-effectiveness, and issues with error rates for specialized terminology or accented speech.

Clinical Nursing Documentation With Speech Recognition

Compared to physician documentation, evidence on nursing documentation with AI speech assistants is more limited but growing. Findings to date are still heterogeneous. For example, many studies reviewed by Joseph et al [10] indicated beneficial effects, such as reduced documentation time and improved workflow. However, a substantial number reported no significant impact, given ongoing challenges integrating AI speech assistants into real-world nursing environments and existing documentation systems.

However, several recent investigations suggest a promising potential for efficiency gains in real-world nursing environments. In a hospital time-motion study, Ehrler et al [16] found that a mobile bedside application meaningfully decreased the duration of electronic health record (EHR) documentation for nurses while also increasing time available for direct patient interaction. Similarly, Mayer et al [17] reported statistically significant time savings—up to 6.1 minutes per documentation scenario—when using a voice recognition system compared to traditional keyboard entry. Mairittha et al [18] also demonstrated that a smartphone-based dialogue system allowed for 15 percent faster documentation relative to electronic forms in a controlled setting. Likewise, qualitative findings from German long-term care facilities hint at the potential of AI speech assistants for care documentation, reporting self-reported time savings, fewer interruptions, greater flexibility, and less reliance on stationary computers [19].

Research Gap

In summary, evidence across clinical settings suggests that AI speech assistants can shorten documentation time and thus reduce administrative burden. At the same time, findings are mixed and often derive from controlled or pilot evaluations rather than routine practice, which limits generalizability [10,13,15]. In nursing, several studies report faster entries with mobile or dialogue-based tools, but most were scenario-based, single-unit, or acute-care focused, with heterogeneous outcomes and little full-shift observation. Long-term care therefore remains underrepresented despite the high documentation burden [12,16-18].

Objectives

Therefore, this study aims to address these gaps by evaluating a domain-specific mobile AI speech assistant under routine conditions in German long-term care using full-shift time-motion observation at 2 time points, before and after implementation. The primary end point is the adjusted change in total documentation time per observed morning shift, defined as the cumulative time spent on a set of predefined documentation-related tasks. Secondary outcomes are self-reported and include usability, perceived documentation effort and interruptions, workplace satisfaction, and self-reported documentation quality.

Hypotheses

The following hypothesis will be tested for the primary end point:

- H0: There will be no significant reduction measured by the adjusted mean change in total documentation time from baseline to postimplementation.
- H1: There will be a significant reduction measured by the adjusted mean change in total documentation time from baseline to postimplementation.

XSL•FO
RenderX

The secondary outcomes are evaluated exploratorily by pre-post changes. To adjust for multiple testing, the Holm-Bonferroni correction is applied across all secondary tests.

Methods

Study Design and Setting

The study was conducted as part of the "PYSA – Nursing Documentation With Hybrid Speech Assistant" project, funded by the German Federal Ministry of Education and Research [20]. The longitudinal research design follows a pre-post observational time-motion design with 2 measurement points (baseline t_0 and postassessment t_1) and takes place in stationary long-term care facilities in Germany. Facilities were approached nationwide to ensure the representation of different conditions and to include both rural and urban areas as well as different types of providers. All participating facilities used either the same EHR system or paper-based documentation.

The minimum age for inclusion of participants was 18 years, with no upper age limit. Additional inclusion criteria were a completed 3-year nursing education, regular involvement in nursing documentation, and at least 1 year of professional experience. Participants had to be able to understand instructions in German and provide informed consent. Only staff employed in long-term care facilities in which the speech assistant voize was implemented during the study period were eligible. Exclusion criteria were applied to staff who were not permanently employed in the participating facilities, such as agency or temporary nurses. We report this observational study in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement [21].

Recruitment

Participants were recruited through long-term care facilities that planned to implement the speech assistant voize and received written information about the study procedures. Following approval from the employee representative bodies of each facility, nursing staff were approached by the facility management. All participants were informed in writing that participation was voluntary, that data would be pseudonymized and analyzed in aggregated form, and that they could withdraw from the study at any time without facing any negative consequences. Participants were offered a voluntary small financial compensation for their time and effort.

The Speech Assistant Voize

The speech assistant voize is a software-based mobile app that enables users to dictate nursing documentation via smartphone. It uses a continuously trained language model that processes voice inputs locally, assigns the content to the appropriate categories, and automatically generates structured documentation entries. These entries are reviewed and approved by nursing staff before being integrated into the existing documentation system. Developed in close collaboration with nursing professionals [19,20], the application is continuously refined through a collaborative, user-centered design process, and its underlying models are regularly trained with voluntarily submitted and anonymized documentation data to ensure ongoing improvements and adaptation to the linguistic characteristics and professional terminology used in nursing.

Technically, the system converts spoken language into text (speech-to-text) in real time using an end-to-end streaming ASR (automatic speech recognition) model. In conjunction with this, a set of specialized natural language processing models is activated as needed to handle specific tasks—such as understanding context, matching spoken language to planned interventions, extracting key information, or structuring data—ensuring that all components work together to deliver the supported features. As part of the study, voize was introduced in cooperation with voize GmbH. The implementation included planning and rollout within the facilities.

Power Analysis

In a preliminary anonymous survey with 55 nurses already using voize, about 70% reported subjective time savings of 16 to 30 minutes per shift. It was assumed that about one-third of working time is spent on documentation-related tasks [6,7], which is equivalent to about 20 minutes per hour. A 15% reduction was considered meaningful, corresponding to the mean of reported savings (23 minutes per shift), that is, 2.9 minutes less per working hour, reducing documentation from 20 to 17.1 minutes per hour. With an assumed SD of 6 minutes, this equates to an effect size of Cohen d=0.5. Power analysis using the R package pwr [22] (α=.05, 2-sided; power=0.8) indicated a minimum of 34 participants for paired t tests. To account for an expected 15% dropout, the target sample size was set at 40, compensating for likely reasons such as scheduling conflicts, shift changes, sick leave, or vacation.

Early-stage dropout rates exceeded expectations. Therefore, additional recruitment efforts were employed. Nonetheless, interest from both facilities and individual nurses was high, and incremental participant burden remained low. As a result, a total of 52 nurses were included in this study.

Measures

Primary End Point: Total Documentation Time per Shift

The primary outcome was total documentation time per observed morning shift in minutes, defined as the sum of all documentation-related categories recorded during the time-motion observation. Observation categories were developed and refined during preparatory nonparticipant observations in a long-term care facility within an unpublished pilot study [20]. The identified documentation-related tasks included documentation at the PC or on paper, consulting information at the resident record at the PC, waiting for a PC, booting up and logging into the PC, walking to the ward office for documentation purposes, making notes, and reviewing entries (Table S1 in Multimedia Appendix 1). All facilities applied their usual infection control protocols for mobile devices. Nurses were instructed to disinfect the study smartphones in accordance with local hygiene guidelines. When these hygiene steps occurred in direct connection with documentation, they were included in the observed documentation time. At t_1, after implementation of the application, additional categories were

XSL•FO
RenderX

included for documentation with voize and consulting information in voize (Table S2 in Multimedia Appendix 1).

Nursing Workplace Satisfaction Questionnaire (NWSQ)

Workplace satisfaction outcomes were assessed using the NWSQ [23]. The NWSQ is a brief measure of nursing job satisfaction encompassing 3 domains: intrinsic, extrinsic, and relational satisfaction. Items are rated on a 5-point Likert scale, and domain scores are computed as the sum of their constituent items. Possible ranges are 6‐30 for intrinsic, 5‐25 for extrinsic, and 4‐20 for relational satisfaction, with higher scores indicating greater satisfaction.

Usability

Technology-related usability was measured with the system usability scale [24] (SUS), a 10-item questionnaire of perceived usability on a 5-point Likert scale and alternating item valence. Responses were scored using the standard SUS procedure, including appropriate reverse-scoring, and converted to a total score from 0 to 100. Scores around 70 indicate acceptable usability, with higher values reflecting better perceived usability [24].

Self-Developed Questionnaire

Several single-item measures were developed and used to capture several facets of potential impact based on a previous study [19]. Items included perceived interruptions during daily work, the perceived time-consuming nature of documentation, satisfaction with the documentation system, expected quality improvement, expected time savings, current documentation quality and completeness, implementation as a good idea, intention-to-use, and self-reported documentation time in minutes per shift. Items used 5-point scales or 10-point scales as shown in the Results section, plus open numerical entries for minutes. Higher ratings indicate more agreement with the item. For example, higher values on “ documentation is time-consuming” indicate greater perceived time burden, and higher values on “satisfaction with the documentation system” indicate greater satisfaction. Item wording and differences across t_0 and t_1 are provided in Tables S1 and S2 in Multimedia Appendix 2.

Procedure

Data were collected between September 2024 and July 2025 in German long-term care facilities. Trained observers conducted continuous, nonparticipant full-shift observations of morning shifts with a planned duration of 8 hours, and observed shift lengths were very similar between t_0 and t_1.

Time was captured with an event-based smartphone logging app by starting and stopping predefined documentation-related task categories. After completing the shift, participants were asked to complete an online questionnaire that included the measures.

Approximately 8 weeks after the implementation of voize (t_1), participants were observed again during a full morning shift, and documentation times were recorded using the same procedure. At this second measurement point, participants completed an adjusted online survey, including the SUS.

Observations at both time points took place on typical weekdays outside major holiday periods. According to the participating units, bed capacity and staffing ratios were generally stable between t_0 and t_1. Taken together, these conditions suggest that major systematic differences in resident census between t_0 and t_1 were unlikely, although smaller variations may still have occurred.

Observers and Minimization of Bias

Prior to data collection, all observers were employees of voize GmbH who were specifically hired for the purpose of this study and were not involved in the development or design of the speech assistant. Observers had no performance incentives tied to study outcomes and were instructed to adhere strictly to the predefined task categories and timing rules to minimize the risk of biased classification or timing cessation. They received standardized training from the Charité research team in the use of the measurement instruments and the predefined task categories. As part of this training, observers first jointly worked through standardized observation scenarios and coded them in parallel. Any discrepancies were then discussed systematically until a shared consensus on the application of each category was reached. In addition, weekly meetings were held throughout the data collection period to address questions, reflect on measurement procedures, and resolve ambiguities in a standardized manner. During the observations, observers also had access to a notes field where they could record questions or uncertainties regarding the assignment to task categories, add comments, and briefly describe the situation. This enabled the study personnel to later compare the selected task category with the descriptive notes and to correct obvious misclassifications where necessary.

In addition, the direct observation design can lead to behavioral changes among participants. The presence of an external observer can sometimes prompt improved performance, a phenomenon known as the Hawthorne effect [25]. To mitigate such influences, repeated briefings and transparent communication of the study objectives were provided to the nursing staff and critical questions were encouraged.

Statistical Analysis

Statistical analyses were performed using Python 3.11.9 and R 4.5.2 in Visual Studio Code 1.103.2 with the packages scikit-learn [26], miceforest [27], NumPy [28], pandas [29], SciPy [30], Matplotlib [31], tidyverse [32], lme4 [33], and lmerTest [34]. All statistical methods used 2-sided tests at α=.05.

To account for the hierarchical and longitudinal structure of the data with repeated measurements within participants and clustering within facilities, the primary end point was analyzed using a linear mixed-effects model (LMM). The primary estimand was the adjusted mean change in documentation time from baseline to postimplementation ($t_1 - t_0$) in minutes. Time was coded as 0 at baseline and 1 at postimplementation, so that the fixed effect of time directly represents the adjusted mean change. The primary model included fixed effects for time, age, gender, and baseline documentation type, and random intercepts for participants and facilities. The model was estimated with

XSL•FO
RenderX

restricted maximum likelihood. Inference for fixed effects used Kenward-Roger degrees of freedom.

The primary model used all available outcome observations and handled missing primary outcomes via likelihood-based estimation under a missing-at-random assumption. Robustness was assessed in exploratory sensitivity analyses, including a complete-case analysis requiring observations at both time points, models additionally accounting for observer effects, and alternative random-effects specifications. We further repeated analyses using multiple imputation with pooling via Rubin rules and Barnard-Rubin degrees of freedom and conducted missing-not-at-random (MNAR) pattern-mixture sensitivity analyses including jump-to-baseline and delta adjustments. As an additional robustness check, we repeated the analysis using a baseline-adjusted analysis of covariance with change scores. Full model specifications, statistical rationale, imputation settings, and MNAR sensitivity procedures are provided in Multimedia Appendix 3. As the primary and secondary analyses fall under the frequentist paradigm, we additionally conducted an exploratory Bayesian hierarchical repeated-measures analysis (Multimedia Appendix 4).

All secondary, self-reported outcomes were analyzed exploratorily using within-subject pre-post comparisons on multiply imputed datasets pooled according to Rubin rules and Barnard-Rubin small sample adjustment. For each outcome, the estimand was the mean change score ($t_1 - t_0$) with heteroscedasticity-consistent type 3 robust standard errors. To account for multiplicity, a Holm-Bonferroni correction was applied across secondary outcomes.

Ethical Considerations

The study was reviewed and approved by the Ethics Committee of Charité – Universitätsmedizin Berlin, approval number EA1/209/24. All participants provided written informed consent prior to participation. Data were pseudonymized and analyzed in aggregated form. The study was preregistered in the German Clinical Trials Register (DRKS00035512). Participants received 30 € (US $34.43) for their participation.

Results

Participant Characteristics

The sample comprised 52 registered nurses from 14 long-term care facilities in Germany. Of the 52 participants, 8 (15.4%) dropped out between baseline and postassessment, primarily due to long-term illness or resignation.

The mean age was 42.37 (SD 12.37) years, with 17.52 (SD 9.63) years of professional experience and 9.42 (SD 8.13) years of tenure. Most participants were female (42/52, 80.8%). Among completers (44/52, 84.6%), the mean age was 43.23 (SD 12.01) years and mean experience was 18.2 (SD 9.95) years; 84.1% (37/44) were female. Dropouts (8/52, 15.4%) were younger on average (mean age 37.62, SD 14.07 years) and reported less work experience (mean 13.75, SD 6.88 years), and 5/8 (62.5%) of them were female. Detailed counts and percentages for all characteristics are provided in Table 1.

Table . Participant characteristics of the total sample, completers, and dropouts.

Variable	Total (N=52)	Completers (n=44)	Dropouts (n=8)
Age (years), mean (SD)	42.37 (12.37)	43.23 (12.01)	37.62 (14.07)
Work experience (years), mean (SD)	17.52 (9.63)	18.20 (9.95)	13.75 (6.88)
Tenure (years), mean (SD)	9.42 (8.13)	9.75 (8.41)	7.62 (6.52)
Gender, n (%)			
Male	10 (19.2)	7 (15.9)	3 (37.5)
Female	42 (80.8)	37 (84.1)	5 (62.5)
Other	0 (0)	0 (0)	0 (0)
Role, n (%)			
Registered nurse	52 (100)	44 (100)	8 (100)
Leadership role	6 (11.5)	5 (11.4)	1 (12.5)
Practical instructor	6 (11.5)	6 (13.6)	0 (0)
Other	1 (1.9)	0 (0)	1 (12.5)

Primary Outcome

Descriptive Analyses

At t_0 (n=52), documentation was predominantly PC-based (35/52, 67.3%), followed by hybrid use (11/52, 21.2%) and paper (6/52, 11.5%). Among completers at t_1 (n=44), the speech assistant voize was reported as the primary documentation mode by 18/44 (40.9%), PC for 25/44 (56.8%), and paper by 1/44 (2.3%). Usage frequency of the speech assistant was high: 33/44 (75%) reported daily use, 9/44 (20.5%) used the speech assistant 3 - 4 times per week, and 2/44 (4.5%) 1 - 2 times per week.

Overall, around 770 hours of work were observed in long-term care during this study. Total documentation-related time decreased from a mean of 53.91 (SD 23.59) minutes at baseline to 39.10 (SD 16.29) minutes postimplementation in the analysis pooled across 40 multiply imputed datasets (Δ=−14.81 minutes). For the change in total documentation time (post – pre in minutes), complete cases (n=44; see Figure 1) had Q1 –22.48

XSL•FO
RenderX

and Q3 –2.90 (min=−85.90, max=32.60), and the pooled analysis (n=52) had Q1 –23.59 and Q3 –2.54 (min=−86.95, max=32.91).

Figure 1. Baseline to postassessment changes in total documentation time among participants.

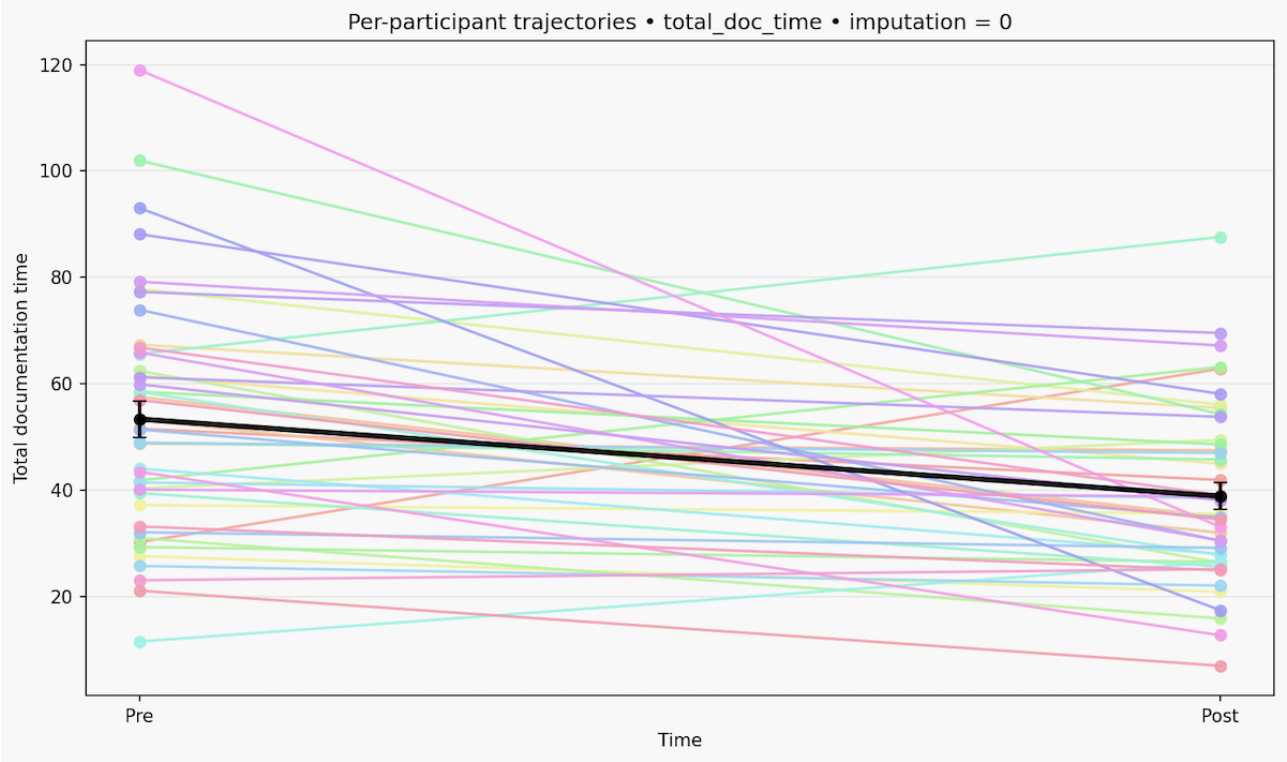

Descriptive reductions were concentrated in active documentation (Δ=−13.28) and note-taking (Δ=−2.38), while reviewing entries (Δ=+0.64) and information retrieval (Δ=+1.28) showed small descriptive increases, consistent with a shift from producing text to checking/retrieving information.

Primary Outcome: Changes in Total Documentation Time

The LMM with random intercepts for participants and facilities and fixed effects for time, age, gender, and baseline documentation type indicated a statistically significant reduction in total documentation time. The adjusted mean change for time (β) was −15 (SE 3.36) minutes, $t_{46.29}$=−4.46, P<.001, with a 95% CI −21.75 to −8.23. Relative to the observed baseline mean of 53.91 minutes, this corresponds to an adjusted 28% reduction in documentation time per morning shift. Variance partitioning indicated an intraclass correlation coefficient of 0.36 at the participant level and 0.08 at the facility level.

Exploratory analyses including time-by-covariate interactions did not indicate reliable heterogeneity of the time effect by age (P=.25), gender (P=.87), or baseline documentation type (P=.49). The estimated reduction was stable across sensitivity analyses, with effect estimates and confidence intervals closely aligned with the primary model. The reduction also remained highly robust under extreme simulated MNAR departures. Detailed results of the exploratory and sensitivity analyses are provided in Multimedia Appendix 5.

Secondary, Self-Reported Outcomes

There was a significant reduction in self-reported documentation time (Δ=−31.14, SE 6.57 minutes), $t_{43.18}$=−4.88, P<.001, 95% CI −45.31 to −18.83. Self-reported daily work interruptions also declined, $t_{47.00}$=−3.95, P<.001, 95% CI −0.95 to −0.31. Participants experienced the documentation process as markedly less time-consuming, $t_{46.91}$=−9.32, P<.001, 95% CI −1.97 to −1.27, and satisfaction with the documentation system increased after the implementation of voize, $t_{46.47}$=4.84, P<.001, 95% CI 1.04 - 2.52. Expectations that the voice assistant would improve documentation quality were also lower after use, $t_{46.82}$=−3.40, P=.001, 95% CI −0.75 to −0.19. All 5 of these effects remained significant after Holm-Bonferroni correction across the 13-test family (Holm-adjusted P≤.013 for each; see Table 2).

XSL•FO
RenderX

Table . Pooled measures of central tendency for secondary, self-reported outcomes[a].

Variables (score range: min-max)	Pooled baseline, mean (SD)	Pooled post, mean (SD)	Mean difference (post – pre)	Median (pre)	Median (post)
Interruptions in daily work (range: 1 - 5)	4.06 (0.78)	3.40 (0.91)	−0.66[b]***	4.00	3.97
Self-reported documentation time (in minutes)	67.05 (40.98)	35.91 (24.34)	−31.14[b]***	60.00	30.00
Documentation is time-consuming (range: 1 - 5)	4.07 (0.79)	2.46 (0.84)	−1.61[b]***	4.00	2.18
Satisfaction with documentation system (range: 1 - 10)	5.07 (2.07)	6.65 (1.82)	+1.58[b]***	5.00	7.00
Expected quality improvement (range: 1 - 5)	3.69 (0.74)	3.18 (0.81)	−0.51[b]**	4.00	3.00
Documentation is complete (range: 1 - 5)	3.27 (0.85)	2.83 (0.94)	−0.44**	3.00	2.87
Expected time savings (range: 1 - 5)	3.81 (0.69)	3.46 (0.79)	−0.35*	4.00	3.57
High documentation quality (range: 1 - 5)	3.38 (0.82)	3.03 (0.75)	−0.35*	3.89	3.00
Implementation as a good idea (range: 1 - 10)	7.91 (1.93)	7.35 (2.04)	−0.56	8.00	7.73
Intention to use (range: 1 - 10)	8.17 (1.99)	7.88 (1.69)	−0.29	8.13	8.00
NWSQ[c]: extrinsic (range: 5 - 25)	17.60 (2.34)	17.95 (2.94)	+0.35	17.51	18.00
NWSQ: intrinsic (range: 6 - 30)	22.21 (3.94)	22.04 (4.42)	−0.17	22.00	21.63
NWSQ: relational (range: 4 - 20)	15.62 (2.62)	15.31 (2.66)	−0.31	16.00	16.00

[a]All values are pooled across 40 multiply imputed datasets according to Rubin rules. Consequently, some median scores are not restricted to the discrete response categories. The significance of the change is indicated by asterisks based on paired *t* tests. *P<.05; **P<.01; ***P<.001.

[b]Effects that remained significant after Holm-Bonferroni correction for 13 tests.

[c]NWSQ: Nursing Workplace Satisfaction Questionnaire.

Perceiving the implementation of voize as a good idea showed no reliable change, $t_{46.29}$=−1.09, P=.28, 95% CI −1.04 to 0.31, and intention to use remained essentially unchanged at a high level, $t_{46.60}$=−0.45, P=.65, 95% CI −0.77 to 0.49. Both remain nonsignificant after correction (Holm-adjusted P≥.99). Ratings that the current documentation is qualitatively high decreased, $t_{46.97}$=−2.30, P=.03, 95% CI −0.61 to −0.04, as did ratings that it is complete and comprehensive, $t_{47.01}$=–2.75, P=.008, 95% CI −0.74 to –0.11, and expected time savings with the speech assistant, $t_{46.74}$=–2.60, P=.01, 95% CI −0.55 to −0.07. However, none of these 3 effects survived Holm-Bonferroni correction (Holm-adjusted P=.16, .06, and .09, respectively).

Nursing workplace satisfaction measured by the NWSQ did not change significantly (extrinsic: $t_{38.94}$=1.48, P=.15, 95% CI −0.22 to 1.45; intrinsic: $t_{39.06}$=0.07, P=.95, 95% CI −1.03 to 1.10; relational: $t_{38.46}$=−0.68, P=.5, 95% CI −0.84 to 0.42), and these remain nonsignificant after correction (Holm-adjusted P=.73, ≥.99, and ≥.99, respectively). The SUS scores indicated acceptable usability [24]: the complete-case mean was 70.42 (SD 14.26; n=40).

Discussion

Principal Findings

This study provides real-world evidence from long-term care that a mobile speech assistant is associated with reduced documentation time over full shifts. The primary LMM estimated an adjusted reduction of approximately 15 minutes per morning shift, corresponding to an approximately 28% adjusted decrease relative to the baseline mean. This finding was robust across extensive sensitivity analyses. The interquartile range indicated that most participants experienced time savings. Task-level patterns suggest that time was recaptured by making active documentation more efficient (see Table 3). However, net time savings may be overestimated as

XSL•FO
RenderX

infection control steps such as disinfection time and adherence to infection control protocols were not systematically recorded in this study. In parallel, nurses reported fewer interruptions, lower perceived time burden, and higher satisfaction with the mobile AI speech assistant. Notably, self-reported retrospective time savings were substantially larger than the observed reduction, and perceived baseline documentation time exceeded our time-motion estimates. This divergence suggests that perceived relief in documentation burden may exceed objective time savings when moving from manual entry to speech-based workflows. Retrospective estimates may apply broader boundaries to what counts as documentation and rely on memory-based summary impressions that overweight salient moments. In this context, cognitively exhaustive, interruption-prone work such as administrative documentation may be remembered as more time-consuming, whereas patient interaction may feel more absorbing and therefore subjectively shorter. By contrast, time-motion observation captures only predefined, observable tasks within the sampled shift, yielding an objective estimate of recorded activities that does not necessarily map onto perceived workload.

Table . Documentation time by task in minutes at baseline and postimplementation.

Task	Pre[a] (minutes), mean (SD)	Post-complete[b] (minutes), mean (SD)	Post-pooled[c] (minutes), mean (SD)
Computer login	1.45 (1.32)	1.05 (1.13)	1.01 (1.09)
Active documentation	44.23 (19.72)	30.44 (14.71)	30.95 (14.36)
Reviewing entries	0.82 (3.65)	1.47 (3.81)	1.46 (3.68)
Information retrieval	3.48 (4.71)	4.85 (6.69)	4.76 (6.48)
Walking to the PC (for documentation)[d]	0.56 (1.56)	0.10 (0.24)	0.09 (0.23)
Taking notes	3.15 (2.68)	0.84 (1.67)	0.77 (1.56)
Waiting	0.22 (0.88)	0.07 (0.33)	0.06 (0.31)
Total documentation-related time	53.91 (23.59)	38.82 (16.75)	39.10 (16.29)

[a]Pre: observed baseline for all participants (N=52).

[b]Post-complete: observed postimplementation among completers (n=44).

[c]Post-pooled: estimates pooled across 40 multiply imputed datasets under “missing at random” assumption (N=52).

[d]“Walking to the PC (for documentation)” refers exclusively to walking carried out to reach stationary ward computers or the ward office for documentation purposes. Walking that occurred while documentation was being performed (eg, dictating with the mobile speech assistant while moving) was not logged as a separate walking task but was counted only as documentation time.

Overall usability was rated as acceptable. Taken together, these findings suggest that introducing speech-enabled, mobile documentation can reduce documentation time while improving perceived documentation efficiency.

At the same time, self-reported documentation quality and completeness showed a consistent negative trend after implementation. In the frequentist exploratory analyses, these decreases were statistically significant before multiplicity correction, but not after Holm-Bonferroni adjustment. However, the complementary Bayesian analysis used adaptive regularization for secondary outcomes and assessed practical relevance using a region of practical equivalence. The Bayesian analysis still assigned a high posterior probability of negative changes in perceived documentation completeness and quality, while the 94% highest density intervals slightly overlapped zero (see Multimedia Appendix 4). Together, these findings suggest a potential trade-off in routine use. Faster, more mobile documentation may coincide with lower perceived completeness or quality, possibly because of recognition errors, editing burden, or shorter entries [16,18,19].

In stationary long-term care, documentation consumes substantial nurse time. A domain-specific mobile speech assistant may reclaim minutes each shift, possibly reduce interruptions, and improve the overall documentation experience. In labor-constrained markets, recovered time can be redeployed to resident care, coordination, supervision, taking necessary breaks, and improving working conditions while lowering overtime [19].

Relation to Prior Work

These findings are consistent with experimental and simulation studies showing that speech-based input improves speed and, in some contexts, accuracy. In nursing, Mayer et al [17] found time savings of 2.3 to 6.1 minutes per scenario without increased errors and high user preference. Similarly, Mairittha et al [18] reported 15% faster entries with a dialogue-based system. In addition, another mobile bedside app reduced EHR charting time and increased uninterrupted patient interaction in a pre-post study [16]. In broader clinical samples, speech systems shortened task completion time and reduced error rates versus paper and traditional patient data management systems in intensive care units [12]. In hospital-wide deployments, they also achieved high adoption with large transcription-cost reductions and perceived gains in quality and efficiency [35]. Observational comparisons have also shown that speech can be substantially faster than typing, even when perceptions may differ [11].

Relative to this body of evidence, this study adds three elements. First, it focuses on long-term care, where evidence has been

XSL•FO
RenderX

sparse and documentation burden is high. Second, it evaluates the technology under routine conditions, not only in controlled or simulated scenarios. Third, it uses full-shift observation at both time points, yielding approximately 770 hours of time-motion data and directly aligned with the workflow realities reported in earlier studies [6,7,36]. Together, these features can help bridge the gap highlighted by reviews that call for end-to-end, real-world evaluations beyond transcription accuracy alone [10,15]. The pattern of decreased active entry, waiting, and walking to the stationary ward computers, with slight increases in retrieval and review (see Table 3), also may show that mobile tools reduce "print-note-retype" loops [16].

Several mechanisms may influence the observed association between speech assistant use and reduced documentation time. First, this is in line with previous qualitative findings of Ferizaj and Neumann [19] that have shown fewer interruptions, less walking to the ward, shorter information-retrieval times, and higher overall satisfaction with the documentation system. In line with this, the mobile, as-you-go entry may have further reduced travel and queuing at ward PCs, cutting waiting, walking to the ward, and repeated logins. Second, domain-tuned speech and natural language processing (ie, models adapted to the nursing domain) may lower the friction of capturing structured information. This may reduce the need for interim notes and subsequent re-entry while also minimizing information loss and documentation errors. Third, workflow coupling at the bedside may decrease interruptions and handovers. The descriptive rise in information retrieval and review may reflect appropriate verification behavior within the mobile app. Usability in the good range [24] and stable acceptance [37] suggest the solution was well received with high intention-to-use and satisfaction, albeit with room for refinement.

Limitations

This study has several limitations. The single-group pre-post design without a control condition limits causal inference and leaves room for workload fluctuations or co-occurring process changes as alternative explanations. According to reports from facility management, no other major documentation tools or workflow redesigns were introduced between t_0 and t_1, which reduces, but does not eliminate, the risk of confounding by concurrent initiatives. Participation was voluntary, so self-selection may have favored early adopters. Additionally, as facilities reported generally stable conditions, we did not collect shift-level workload indicators such as number of residents cared for or staffing ratios for each observed shift. Therefore, documentation time was not normalized for patient census or workload, and residual confounding by day-to-day fluctuations remains possible. Observations were conducted by multiple observers. Although observers underwent standardized training, joint pilot coding, and ongoing weekly calibration meetings, we did not calculate a formal interrater reliability coefficient (eg, Cohen κ) for the final field observations. This limits the extent to which we can quantify measurement consistency across observers. However, an exploratory sensitivity analysis including observer ID as a covariate did not materially change the estimated reduction in documentation time, suggesting that systematic between-observer differences are unlikely to account for the primary findings. Still, this exploratory analysis cannot rule out shared, potentially unconscious observer bias operating in the same direction across time points. Fourth, direct observation can itself change behavior (Hawthorne effect), which we sought to mitigate with standardized procedures but cannot exclude [25]. We did not systematically record time spent on disinfecting mobile devices unless it occurred in direct connection with documentation or monitor adherence to infection control protocols. If device disinfection is performed more consistently in routine use, the net time savings reported here may be overestimated. Documentation quality was captured perceptually rather than via objective measures such as assessing structured completeness or error rates that other work has examined in standardized tasks [12]. We did not conduct independent audits of documentation accuracy (eg, chart reviews, error checks). Reduced documentation time therefore cannot be directly linked to unchanged or improved documentation quality. Furthermore, dropouts were descriptively younger and less experienced than completers. This pattern may limit generalizability for adoption, since retention and continued engagement with the tool may differ by career stage, training status, or role-related workload. Finally, results reflect stationary long-term care in Germany, German-language ASR, and a specific EHR integration. Performance may differ with dialects, accents, terminology, and integration quality [15], and ongoing advances in speech technology may change effect sizes over time.

Future Research

This study complements prior simulation and laboratory research and provides further evaluation of the potential effects of AI speech assistants in long-term care. To better inform workforce policy in the context of nursing shortages, future implementation research should examine whether documentation automation yields sustained workload relief, frees time for direct resident care, coordination, and recovery, and under which organizational and technical conditions these benefits can be realized at scale.

To strengthen causal inference and capture secular trends and learning effects, future studies should use cluster-randomized or stepped-wedge designs with extended follow-up. These evaluations should combine time-motion data with independent, objective assessments of documentation quality (eg, completeness indices, semantic accuracy ratings, and error rates) to determine whether time savings are achieved without compromising data integrity. In addition, studies should assess how saved time is reinvested and whether this translates into improvements in resident outcomes, psychosocial workload, perceived interruptions, staff satisfaction, and retention.

Methodologically, linking full-shift observation with EHR telemetry could quantify sustainability, diurnal patterns, and adoption trajectories. Additionally, further subgroup analyses could explore heterogeneity by accent/dialect, professional experience, unit characteristics, and integration context. Finally, economic evaluations should estimate the total cost of ownership and downstream effects on productivity, overtime, sick leave, and turnover to guide procurement and financing decisions for broader implementation.

XSL•FO
RenderX

Conclusions

To the authors' knowledge, this is the first study to report real-world time savings observed in association with the use of a mobile AI speech assistant in long-term care. In this longitudinal pre-post time-motion study conducted in 14 German long-term care facilities, implementation of a mobile, domain-specific speech documentation system into routine workflows was associated with substantially lower documentation time. Specifically, the mean documentation-related time decreased from about 54 minutes at baseline to approximately 39 minutes postimplementation. In the primary LMM with random intercepts for participants and facilities and fixed effects for time, age, gender, and baseline documentation type, the adjusted mean change was −15 minutes, corresponding to an adjusted 28% reduction relative to the baseline mean. This result remained consistent across sensitivity analyses and MNAR pattern-mixture scenarios. Task-level patterns indicated that time savings were primarily driven by less active documentation and note-taking, while reviewing entries and information retrieval showed small increases. This is consistent with a shift from producing text to verifying and retrieving information [19].

Secondary, exploratory self-report measures converged with the observational findings: participants reported fewer interruptions, lower perceived documentation burden, and higher satisfaction with the documentation system, while nursing workplace satisfaction did not change. Usability was rated as acceptable. Most participants experienced measurable time savings, with approximately one quarter reducing documentation time by more than about 24 minutes per shift. Given the single-group pre-post design and the reliance on self-reported indicators for documentation quality, these results should be interpreted as evidence of an association between implementation and reduced documentation time rather than as causal proof or as objective evidence of changes in documentation quality.

Acknowledgments

We would like to thank the German Federal Ministry of Education and Research as the responsible funding body. In addition, we would like to thank the participating facilities, and especially the nursing staff, who contributed their insights and expertise to this study.

We used the generative AI tool ChatGPT by OpenAI to assist with language refinement and translation of text originally written by the authors, and to help identify potentially relevant scientific literature. All suggested references were independently reviewed and verified by the authors. No generative AI tools were used to generate scientific content, analyses, results, or interpretations.

Funding

This study was conducted as part of the "PYSA – Nursing Documentation With Hybrid Speech Assistant" project, funded by the German Federal Ministry of Education and Research.

Data Availability

The datasets generated and analyzed in this study contain highly granular time-motion logs and contextual information from nursing staff in German long-term care facilities. Owing to data protection regulations, institutional data use agreements, and the terms of the ethics approval and informed consent based on a preregistered study protocol that did not foresee open data sharing, these data cannot be made publicly available. Deidentified aggregate data are reported in the manuscript and appendices. Researchers who wish to access additional deidentified data for scientifically sound purposes may contact the corresponding author; data sharing will be considered on a case-by-case basis and may require additional institutional approvals and data use agreements.

Authors' Contributions

KS contributed to the concept and design of the study, data interpretation, visualization, preparation of the original draft, and review and editing of the manuscript. DF contributed to the study concept and design, data curation, formal analysis, investigation, visualization, preparation of the original draft, and review and editing of the manuscript. SN contributed to the study design, investigation, project administration, interpretation of data, and review and editing of the manuscript. SS-L contributed to project administration, concept and design, funding acquisition, and review and editing of the manuscript. NL contributed to methodological guidance, formal analysis, supervision, funding acquisition, and review and editing of the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest

KS is employed by voize GmbH, which developed the speech assistant evaluated in this study. She was not involved in the development of the system. All other authors declare no competing interests.

Multimedia Appendix 1
Definition of task categories for observations at t_0 and t_1.
[DOCX File, 1766 KB - jmir_v28i1e86078_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Self-developed questionnaire for assessment at t_0 and t_1.
[DOCX File, 1768 KB - jmir_v28i1e86078_app2.docx]

Multimedia Appendix 3
Statistical analysis rationale and detailed methods.
[DOCX File, 1771 KB - jmir_v28i1e86078_app3.docx]

Multimedia Appendix 4
Bayesian supplementary modeling and analysis.
[DOCX File, 2868 KB - jmir_v28i1e86078_app4.docx]

Multimedia Appendix 5
Results of exploratory sensitivity analyses.
[DOCX File, 2317 KB - jmir_v28i1e86078_app5.docx]

References

1. Arbeitsmarktsituation im Pflegebereich [Report in German]. : Statistik der Bundesagentur für Arbeit; 2025 May URL: https://statistik.arbeitsagentur.de/DE/Statischer-Content/Statistiken/Themen-im-Fokus/Berufe/Generische-Publikationen/Altenpflege.html?__blob=publicationFile [accessed 2026-03-13]
2. Pflegekräftevorausberechnung [Website in German]. Statistisches Bundesamt (Destatis). URL: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Bevoelkerungsvorausberechnung/pflegekraeftevorausberechnung.html [accessed 2026-3-13]
3. Wirth T, Ulusoy N, Lincke HJ, Nienhaus N, Schablon A. Psychosoziale Belastungen und Beanspruchungen von Beschäftigten in der stationären und ambulanten Altenpflege [Article in German]. ASU Arbeitsmed Sozialmed Umweltmed 2017;52(9):662-669. [doi: 10.17147/ASU2017-09-01-01]
4. Hasson H, Arnetz JE. Nursing staff competence, work strain, stress and satisfaction in elderly care: a comparison of home-based care and nursing homes. J Clin Nurs 2008 Feb;17(4):468-481. [doi: 10.1111/j.1365-2702.2006.01803.x] [Medline: 17331093]
5. Dall'Ora C, Ejebu OZ, Ball J, Griffiths P. Shift work characteristics and burnout among nurses: cross-sectional survey. Occup Med 2023 May 18;73(4):199-204. [doi: 10.1093/occmed/kqad046]
6. Schenk E, Schleyer R, Jones CR, Fincham S, Daratha KB, Monsen KA. Time motion analysis of nursing work in ICU, telemetry and medical-surgical units. J Nurs Manag 2017 Nov;25(8):640-646. [doi: 10.1111/jonm.12502] [Medline: 28853187]
7. Yen PY, Kellye M, Lopetegui M, et al. Nurses' time allocation and multitasking of nursing activities: a time motion study. AMIA Annu Symp Proc 2018;2018:1137-1146. [Medline: 30815156]
8. Momenipur A, Pennathur PR. Balancing documentation and direct patient care activities: a study of a mature electronic health record system. Int J Ind Ergon 2019 Jul;72:338-346. [doi: 10.1016/j.ergon.2019.06.012] [Medline: 32201437]
9. Gesner E, Dykes PC, Zhang L, Gazarian P. Documentation burden in nursing and its role in clinician burnout syndrome. Appl Clin Inform 2022 Oct;13(5):983-990. [doi: 10.1055/s-0042-1757157] [Medline: 36261113]
10. Joseph J, Moore ZEH, Patton D, O'Connor T, Nugent LE. The impact of implementing speech recognition technology on the accuracy and efficiency (time to complete) clinical documentation by nurses: a systematic review. J Clin Nurs 2020 Jul;29(13-14):2125-2137. [doi: 10.1111/jocn.15261] [Medline: 32243006]
11. Zuchowski M, Göller A. Speech recognition for medical documentation: an analysis of time, cost efficiency and acceptance in a clinical setting. Br J Healthc Manag 2022 Jan 2;28(1):30-36. [doi: 10.12968/bjhc.2021.0074]
12. Peine A, Gronholz M, Seidl-Rathkopf K, et al. Standardized comparison of voice-based information and documentation systems to established systems in intensive care: crossover study. JMIR Med Inform 2023 Nov 28;11:e44773. [doi: 10.2196/44773] [Medline: 38015593]
13. Alboksmaty A, Aldakhil R, Hayhoe BWJ, Ashrafian H, Darzi A, Neves AL. The impact of using AI-powered voice-to-text technology for clinical documentation on quality of care in primary care and outpatient settings: a systematic review. eBioMedicine 2025 Aug;118:105861. [doi: 10.1016/j.ebiom.2025.105861] [Medline: 40694861]
14. Wang J, Lavender M, Hoque E, Brophy P, Kautz H. A patient-centered digital scribe for automatic medical documentation. JAMIA Open 2021 Jan;4(1):ooab003. [doi: 10.1093/jamiaopen/ooab003] [Medline: 34377960]
15. Ng JJW, Wang E, Zhou X, et al. Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: a systematic review. BMC Med Inform Decis Mak 2025 Jul 1;25(1):236. [doi: 10.1186/s12911-025-03061-0] [Medline: 40598136]
16. Ehrler F, Wu DTY, Ducloux P, Blondon K. A mobile application to support bedside nurse documentation and care: a time and motion study. JAMIA Open 2021 Jul;4(3):ooab046. [doi: 10.1093/jamiaopen/ooab046] [Medline: 34345804]

XSL•FO
RenderX

17. Mayer L, Xu D, Edwards N, Bokhart G. A comparison of voice recognition program and traditional keyboard charting for nurse documentation. Comput Inform Nurs 2021 Aug 4;40(2):90-94. [doi: 10.1097/CIN.0000000000000793] [Medline: 34347642]
18. Mairittha T, Mairittha N, Inoue S. Evaluating a spoken dialogue system for recording systems of nursing care. Sensors (Basel) 2019 Jan;19(17):3736. [doi: 10.3390/s19173736]
19. Ferizaj D, Neumann S. Assessing perceptions and experiences of an AI-driven speech assistant for nursing documentation: a qualitative study in German nursing homes. In: Kurosu M, Hashizume A, editors. Human-Computer Interaction, Part V: Springer; 2024:17-34. [doi: 10.1007/978-3-031-60449-2_2]
20. PYSA – Miteinander durch Innovation. Bundesministerium für Forschung, Technologie und Raumfahrt. URL: https://www.interaktive-technologien.de/projekte/pysa [accessed 2026-3-13]
21. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008]
22. Champely S, Ekstrom C, Gill J, et al. pwr: basic functions for power analysis. GitHub. 2020. URL: https://github.com/heliosdrm/pwr [accessed 2026-03-13]
23. Fairbrother G, Jones A, Rivas K. Development and validation of the Nursing Workplace Satisfaction Questionnaire (NWSQ). Contemp Nurse 2009;34(1):10-18. [doi: 10.5172/conu.2009.34.1.010] [Medline: 20230168]
24. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the System Usability Scale. Int J Hum-Comput Interact 2008 Jul 29;24(6):574-594. [doi: 10.1080/10447310802205776]
25. Franke RH, Kaul JD. The Hawthorne experiments: first statistical interpretation. Am Sociol Rev 1978 Oct;43(5):623. [doi: 10.2307/2094540]
26. Pedregosa F, Varoquaux G, Gramfort A, et al. scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12(85):2825-2830. [doi: 10.5555/1953048.2078195]
27. miceforest: Multiple imputation by chained equations with LightGBM. GitHub. URL: https://github.com/AnotherSamWilson/miceforest [accessed 2026-03-13]
28. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. Nature 2020;585:357-362. [doi: 10.1038/s41586-020-2649-2]
29. McKinney W. Data structures for statistical computing in Python. SciPy 2010. [doi: 10.25080/Majora-92bf1922-00a]
30. Jones E, Oliphant T, Peterson P. SciPy: open source scientific tools for Python. 2001. URL: http://www.scipy.org [accessed 2026-03-13]
31. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng 2007 May;9(3):90-95. [doi: 10.1109/MCSE.2007.55]
32. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. J Open Source Softw 2019;4(43):1686. [doi: 10.21105/joss.01686]
33. Bates D, Maechler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw 2015;67(1):2-48. [doi: 10.18637/jss.v067.i01]
34. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. J Stat Soft 2017;82(13):1-26. [doi: 10.18637/jss.v082.i13]
35. Saxena K, Diamond R, Conant RF, Mitchell TH, Gallopyn IG, Yakimow KE. Provider adoption of speech recognition and its impact on satisfaction, documentation quality, efficiency, and cost in an inpatient EHR. AMIA Jt Summits Transl Sci Proc 2018;2017:186-195. [Medline: 29888069]
36. Yen PY, Kelley M, Lopetegui M, et al. Understanding and visualizing multitasking and task switching activities: a time motion study to capture nursing workflow. AMIA Annu Symp Proc 2016;2016:1264-1273. [Medline: 28269924]
37. Van Der Laan JD, Heino A, De Waard D. A simple procedure for the assessment of acceptance of advanced transport telematics. Transp Res Part C Emerg Technol 1997 Feb;5(1):1-10. [doi: 10.1016/S0968-090X(96)00025-3]

Abbreviations

AI: artificial intelligence
ASR: automatic speech recognition
EHR: electronic health record
LMM: linear mixed-effects model
MNAR: missing not at random
NWSQ: Nursing Workplace Satisfaction Questionnaire
PYSA: “Pflegedokumentation mit Hybridem Sprachassistenten” (project title)
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
SUS: system usability scale
t_0: baseline measurement
t_1: postimplementation measurement

XSL•FO
RenderX

Edited by A Coristine; submitted 17.Oct.2025; peer-reviewed by K Chen, L He, S Mohanadas, S Ye; revised version received 20.Feb.2026; accepted 24.Feb.2026; published 08.Apr.2026.

Please cite as:
Schwabe K, Ferizaj D, Neumann S, Strube-Lahmann S, Lahmann N
Time Savings Through an AI Speech Assistant for Nursing Documentation: Pre-Post Time-Motion Study in German Long-Term Care
J Med Internet Res 2026;28:e86078
URL: https://www.jmir.org/2026/1/e86078
doi:10.2196/86078

© Katja Schwabe, Drin Ferizaj, Susann Neumann, Sandra Strube-Lahmann, Nils Lahmann. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Computerized Self-Reported Medical History Taking to Support Early Rule Out of Major Adverse Cardiac Events in Patients With Acute Chest Pain: Post Hoc Analysis of the CLEOS-CPDS Prospective Cohort Study

Helge Brandberg[1], MD, PhD; Carl Johan Sundberg[2,3], MD, PhD; Jonas Spaak[1], MD, PhD; Sabine Koch[2], MSc, PhD; Thomas Kahan[1], MD, PhD

[1]Division of Cardiovascular Medicine, Department of Clinical Sciences, Danderyd Hospital, Karolinska Institutet, Stockholm, Sweden

[2]Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden

[3]Department of Physiology & Pharmacology, Karolinska Institutet, Stockholm, Sweden

Corresponding Author:
Helge Brandberg, MD, PhD
Division of Cardiovascular Medicine
Department of Clinical Sciences, Danderyd Hospital
Karolinska Institutet
Stockholm, SE-182 88
Sweden
Phone: 46 8 123 583 47
Email: helge.brandberg@ki.se

Abstract

Background: Self-reported, computerized history taking (CHT) may enable efficient collection of medical histories for acute chest pain management.

Objective: The primary aim is to determine the diagnostic performance of 4 CHT-derived chest pain risk scores for ruling out 30-day major adverse cardiac events (MACEs) or acute coronary syndrome (ACS). The secondary aim is to assess their impact on patient disposition in the emergency department (ED).

Methods: This is a prospective cohort study conducted at a tertiary hospital ED in Stockholm, Sweden. Clinically stable adults (≥18 years) with chest pain and an electrocardiogram (ECG) not indicating an acute disease requiring immediate care provided medical histories via a tablet-based CHT program (Clinical Expert Operating System [CLEOS]). CHT data and ECG interpretations and troponin values were used to calculate the History, ECG, Age, Risk Factors, and Troponin (HEART) score, Danderyd HEART (D-HEART) score, Emergency Department Assessment of Chest Pain Score combined with an Accelerated Diagnostic Protocol (EDACS-ADP), and Troponin-only Manchester Acute Coronary Syndrome (T-MACS). The primary outcome was 30-day ACS; the secondary outcome was 30-day MACE (ACS, revascularization, or cardiovascular death).

Results: Among 1000 participants (age: mean 55 years, SD 17 years; 456/1000, 45.60%, women), risk scores could be calculated in 838 (83.80%). Within 30 days, 65 (6.50%) participants experienced ACS, and 72 (7.20%) had a MACE. Negative predictive values were 0.99 (95% CI 0.97-1.00) for both outcomes. Sensitivity for MACE was 0.91 (95% CI 0.81-0.97) for HEART, 0.94 (95% CI 0.86-0.98) for D-HEART, 0.94 (95% CI 0.86-0.98) for EDACS-ADP, and 0.97 (95% CI 0.90-1.00) for T-MACS, with similar results for ACS. As many as 89 of the 528 (16.9%) patients admitted could be reclassified from “nonlow risk” to “low risk.” Among reclassified patients, 30-day MACE or ACS occurred in 0-4 cases; miss rates were below 1% for D-HEART (4/416, 0.96%) and T-MACS (2/286, 0.7%), but exceeded 1% for HEART (6/406, 1.5%) and EDACS-ADP (4/346, 1.2%).

Conclusions: Automated, self-reported CHT provided sufficient data to calculate 4 chest pain risk scores in 838 of 1000 (83.80%) patients with acute chest pain, with score calculation dependent on physician-interpreted ECGs. These CHT-derived risk scores demonstrated good diagnostic performance for ruling out 30-day MACE and ACS. Performance was broadly comparable with prior studies using physician-acquired scores, although suggested safety thresholds were primarily met by D-HEART and T-MACS. The improved safety of D-HEART compared with HEART is likely attributable to the incorporation of serial 0/1-hour troponin testing. Use of CHT-derived risk scores may reclassify a substantial fraction of admitted patients as “low risk,” potentially supporting discharge decisions in selected patients, while admission may still be required for non-ACS reasons. However, any

XSL•FO
RenderX

gains in discharge rates should be weighed against the possibility of missed events among reclassified patients. Multicenter studies are needed to confirm generalizability, operational feasibility, and safety.

Trial Registration: ClinicalTrials.gov NCT03439449; https://clinicaltrials.gov/ct2/show/NCT03439449

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2019-031871

(J Med Internet Res 2026;28:e76087) doi:10.2196/76087

KEYWORDS

chest pain; risk assessment; coronary artery disease; emergency service, hospital; medical informatics; artificial intelligence; medical history taking

Introduction

Chest pain, a common chief complaint in emergency departments (EDs) worldwide [1,2], contributes to overcrowding in these settings [3]. Prompt and accurate management is crucial, not only for identifying life-threatening conditions such as acute coronary syndrome (ACS; acute myocardial infarction or unstable angina pectoris), but also for the safe discharge of the majority of patients with benign conditions. An integrated assessment combining medical history, signs and symptoms, electrocardiogram (ECG), and high-sensitivity cardiac troponins (hs-cTn) is recommended for initial short-term risk stratification [4]. To consolidate these observations, risk scores are recommended for structured, individualized patient assessment [5,6]. Their clinical utility is supported by substantial evidence [7-9], and recent data have shown that risk score implementation can reduce the risk of major adverse cardiac events (MACEs) by half among patients with chest pain discharged from the ED [10]. However, chest pain risk scores remain underutilized in EDs [11], and when used, issues such as miscalculation [12] or discordance among physicians [13] are common.

Digital tools are increasingly integrated into health care, offering potential enhancements in efficiency and decision-making [14,15]. In acute chest pain management, international societies advocate improved symptom classification methods using machine learning [16]. In addition to biometric data such as vital signs or ECG, prompt and reliable acquisition of medical history data is essential for the use of these techniques. Computerized history taking (CHT) is a possible solution, offering an automated, structured, and standardized approach to self-reported medical history. Previous studies indicate that CHT can assist physicians by providing a more standardized, detailed, and complete medical history in both emergent and nonemergent settings [17-20]. Studies have also shown that CHT is well received by patients with acute chest pain, with a majority able to interact effectively and provide sufficient data for risk stratification [21-23]. However, there is a knowledge gap regarding the impact of CHT on clinical management and outcomes in acute care settings [24]. This study proposes a standardized, automated approach to collect medical history data, enabling real-time use of clinical decision-support tools. This approach has the potential to significantly impact the management of the large population of patients presenting to EDs with chest pain.

We hypothesized that risk scores populated with CHT data would reliably identify patients with acute chest pain at low risk for MACE or ACS, enabling early rule out and thereby reducing unnecessary admissions and diagnostic testing without compromising patient safety. Our aims were therefore, first, to determine the overall diagnostic performance of 4 chest pain risk scores populated with CHT data for 30-day MACE (defined as ACS, coronary artery revascularization, or cardiovascular death) or ACS rule out; and second, to assess the potential impact of this approach on patient disposition among those presenting to the ED with acute chest pain.

Methods

Study Design and Setting

The Clinical Expert Operating System Chest Pain Danderyd Study (CLEOS-CPDS; ClinicalTrials.gov identifier: NCT03439449) was a prospective cohort study designed to determine the value of self-reported CHT in acute chest pain management [25]. This study reports diagnostic accuracy and adheres to the STARD (Standards for Reporting Diagnostic Accuracy) 2015 guideline [26].

Participants

Patients presenting to Danderyd University Hospital, a tertiary hospital in Stockholm, Sweden, were consecutively enrolled between October 1, 2017, and May 16, 2019. Eligible participants were women and men aged 18 years or older, fluent in Swedish, with a presenting complaint of chest pain recorded by an ED triage nurse or registrar; a nondiagnostic initial ECG, serum markers, or both, for an acute disease requiring immediate care; and clinical stability, as defined by the Rapid Emergency Triage and Treatment System (RETTS) levels orange, yellow, green, and blue [27]. Exclusion criteria were inability to complete a CHT interview (eg, agitation, severely impaired vision, or confusion). Among patients considered eligible but not enrolled, the most common reasons were language barriers (eg, inability to read Swedish), patients feeling too tired, and difficulties using a tablet [22]. All patients included in the study were initially managed as suspected ACS according to standard care protocols recommended by regional guidelines [28]. Data from the CHT and the calculated risk scores were not accessible to the treating physicians or other health care professionals and were used solely for clinical research purposes.

Interventions

Computerized History Taking

Medical history was collected using the CHT software CLEOS on tablets (iPad; Apple Inc), as described in detail previously

XSL•FO
RenderX

[25,29]. Briefly, CLEOS is an expert system, a form of artificial intelligence software, that uses a rule-based approach to guide the history-taking process. It incorporates a comprehensive medical knowledge base, algorithmically represented by decision trees comprising over 17,000 decision nodes. By directly interacting with the patient, CLEOS automates history taking through a series of structured questions, mainly in text format (ie, yes/no or multiple-choice questions), as well as image-based questions. A limited number of free-text entries are available for instances in which patients are unable to select a suitable option, such as when describing the quality of their pain. Based on prior responses and their clinical relevance, CLEOS adapts the interview and dynamically determines the next most appropriate questions, emulating clinical reasoning as data are continuously collected [29]. In a study evaluating interrater reliability between data collected using CHT and physician-acquired data, we observed high agreement for traditional risk factors (eg, diabetes mellitus or hypercholesterolemia), but low to moderate agreement for chest pain characteristics (eg, pain radiating to the arm or relief by rest) [30]. As previously reported, the median time required to collect sufficient information to calculate the HEART score was 23 (IQR 18-31) minutes [22].

Chest Pain Risk Scores

The History, ECG, Age, Risk Factors, and Troponin (HEART) score; Emergency Department Assessment of Chest Pain Score combined with an Accelerated Diagnostic Protocol (EDACS-ADP); and Troponin-only Manchester Acute Coronary Syndrome (T-MACS) score were calculated as described in their respective derivation studies [31-33]. When assessing the History component of the HEART score in this retrospective analysis, the traditional clinical classification of suspected anginal symptoms was used: (1) central chest pain, (2) triggered by physical or emotional exertion, and (3) relieved by rest or nitrates. Based on the number of criteria met, patient history was categorized as highly (3 criteria met), moderately (2 criteria met), or slightly suspicious (none or 1 criterion met). This predefined, symptom-based approach to probability assessment, used as an aid in the clinical diagnosis of coronary artery disease (Diamond-Forrester prediction rule) [34], reflects commonly applied principles for reconstructing the HEART History component in several validation studies [35,36]. However, this approach differs from the derivation study and subsequent validation study by the same authors, in which clinician judgment, rather than a predefined set of criteria, was used to evaluate the likelihood of chest pain being related to MACE [33,37]. The Danderyd HEART (D-HEART) score was calculated according to guidelines in Region Stockholm [28], combining the HEART score with hs-cTn assays using the European Society of Cardiology (ESC) 0/1-hour rule-out and rule-in algorithm [38,39]. By contrast, the HEART score in this study was calculated using a single baseline hs-cTn measurement, in accordance with the derivation study. We defined a clinically decisive risk score as “nonlow risk,” indicating the need for hospitalization, if D-HEART ≥4, HEART ≥4, EDACS ≥16, T-MACS ≥0.02, or if troponin met rule-in criteria using the ESC 0/1-hour algorithm [4]. For T-MACS, we evaluated both the <0.02 (very low risk) and <0.05 (low risk) thresholds, as they represent distinct clinical pathways (immediate discharge vs clinical observation with further serial troponin testing) and were defined in the derivation study [32]. Detailed descriptions of chest pain characteristics, risk factor components, and interpretation (ie, “nonlow risk” thresholds) for D-HEART, HEART, EDACS, and T-MACS are provided in Table 1.

XSL•FO
RenderX

Table 1. Components and interpretation of assessed risk scores.

Component	Risk score			
	D-HEART[a]	HEART[b]	EDACS-ADP[c]	T-MACS[d]
Age	✓	✓	✓	N/A[e]
Sex	N/A	N/A	✓	N/A
Chest pain characteristics				
Central chest pain	✓	✓	N/A	N/A
Provoked by physical exertion or emotional stress	✓	✓	N/A	N/A
Relieved by rest or nitrates	✓	✓	N/A	N/A
Diaphoresis	N/A	N/A	✓	✓
Radiation to arm or shoulder	N/A	N/A	✓	N/A
Radiation to right arm or shoulder	N/A	N/A	N/A	✓
Pain occurred/worsened with inspiration	N/A	N/A	✓	N/A
Pain is reproduced by palpation	N/A	N/A	✓	N/A
Worsening or crescendo angina	N/A	N/A	N/A	✓
Pain associated with vomiting	N/A	N/A	N/A	✓
Risk factors				
Atherosclerotic disease	✓	✓	N/A	N/A
Known coronary artery disease (not stroke)	N/A	N/A	✓	N/A
Diabetes mellitus	✓	✓	✓	N/A
Current smoker	✓	✓	✓	N/A
Family history of premature coronary artery disease	✓	✓	✓	N/A
Hypertension	✓	✓	✓	N/A
Hypercholesterolemia	✓	✓	✓	N/A
Reported obesity	✓	✓	N/A	N/A
Electrocardiogram, vital signs, and biomarkers				
Signs of ischemia on electrocardiogram	✓	✓	✓	✓
Hypotension, systolic blood pressure <100 mmHg	N/A	N/A	N/A	✓
Elevated biomarker (troponin)	✓	✓	✓	✓
Possible range of score	0 to 8	0 to 10	–10 to 34	0 to 1
Threshold for “nonlow risk”	≥4[f]	≥4	≥16[g]	≥0.02

[a]D-HEART: Danderyd HEART.

[b]HEART: History, ECG, Age, Risk Factors, and Troponin.

[c]EDACS-ADP: Emergency Department Assessment of Chest Pain Score combined with an Accelerated Diagnostic Protocol.

[d]T-MACS: Troponin-only Manchester Acute Coronary Syndrome.

[e]N/A: not applicable.

[f]And a rule-in troponin result using the European Society of Cardiology 0/1-hour algorithm [38].

[g]And new ischemia on electrocardiogram and/or a rule-in 0- or 2-hour troponin result.

Data Collection

At presentation to the ED, all patients with acute chest pain were triaged to determine urgency either by a cardiology consultant or a senior cardiology resident (8 AM to 5 PM) or by a triage nurse using the RETTS protocol (5 PM to 8 AM) [27]. After ECG and biomarker acquisition during triage, patients underwent a more thorough examination and standard history taking by the attending physician, either in the ED cardiology unit—operating 24 hours a day and staffed by cardiology consultants or senior residents with limited access to noninvasive cardiac imaging techniques—or in the inpatient day-care unit. The day-care unit, operational from 8 AM to 5 PM, functions as an observational ward, is staffed by cardiology consultants, and provides ready access to both invasive and noninvasive cardiac imaging techniques. Details regarding

XSL•FO
RenderX

patient disposition have been reported in previous publications [22,25].

The participant enrollment process has been described in detail previously [22,25]. Briefly, patients were invited to participate in the study by a research staff member in the ED cardiology unit or the inpatient day-care unit. Patients were provided with standardized information and given the opportunity to ask questions before signing a consent form to provide informed consent. CHT was conducted during waiting times in the ED, either before or after the initial physician consultation. The CHT interview could be paused during standard clinical procedures (eg, blood sampling or radiographic examinations) and resumed afterward. Discontinuation of CHT occurred upon completion, at the patient's request, or upon discharge home or admission to a ward. The CHT interview did not interfere with standard management, and ED staff had no access to data collected through CHT.

Data relevant to this study were extracted from the electronic health record (EHR; TakeCare; CompuGroup Medical Sweden AB) by research staff and from the CHT database. To align with variables used in the risk scores, we predefined variables to be collected and established an interpretation scheme that allowed all variables to be converted into a binary format (yes/no responses; see an example of the interpretation scheme for crescendo angina pectoris in [30]). Demographic information (age and sex) for the general ED population at Danderyd University Hospital during the study period was extracted from the EHR system using QlikView version 12.10 (QlikTech International AB).

ECG findings, interpreted by both physicians and computer algorithms, were collected by a research assistant from the EHR. To specifically evaluate the independent contribution of CHT to chest pain management, physician-reported ECG interpretations were deliberately used to populate the risk scores. In cases in which a physician report was missing (n=58), the automated ECG interpretation (EC Sense ECG; Cardiolex Medical AB) was imputed. A detailed description of ECG findings and their categorization is provided in Table S1 in Multimedia Appendix 1. Blood samples were collected according to standard procedures at the Karolinska University Laboratory, Stockholm. The detection limit of the assay was 5 ng/l, with a 99th percentile upper reference limit—used to identify elevated levels indicative of myocardial injury—of 14 ng/l [40].

All variables required for each risk score were extracted from the EHR and CHT and transformed according to the predefined interpretation scheme. CHT responses were manually mapped post hoc, and risk scores were calculated retrospectively.

Outcome Measures

The primary outcome was defined as a diagnosis of ACS [25]. To facilitate comparison with other diagnostic accuracy studies of acute chest pain, we also report 30-day MACE as a secondary outcome, comprising an ACS diagnosis, coronary artery revascularization, or cardiovascular death within 30 days. According to the study protocol, the diagnosis of ACS was verified by a board-certified cardiologist in accordance with prevailing European guidelines at the time of the study [41,42]. The cardiologist then assigned the corresponding International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10) codes: I20.0 (unstable angina pectoris), I21.0-9 (acute myocardial infarction), or I24.0-9 (other acute ischemic heart diseases) [43,44]. Previous studies have shown that the overall positive predictive value (PPV) of Swedish discharge ICD codes is 85%-95% for most diagnoses, including acute myocardial infarction [45,46]. Data on death and revascularization were retrieved through review of the regional EHR system, which is shared by all hospitals and cardiology outpatient clinics in Region Stockholm. The regional EHR system is linked to the national population registry, providing complete information on mortality status. No active follow-up or separate linkage to a national registry was performed. Further details are available in the study protocol [25].

Our previous work shows that CHT provides more complete information on key medical history variables than that reported by physicians in the EHR [30]. Consequently, sufficient information for risk score calculation was available for fewer than 30% of patients when using EHR data alone, compared with approximately three-quarters when using CHT data [30]. As a result, reliable retrospective calculation of a complete physician-derived HEART score was not possible, making a direct head-to-head comparison with the CHT-derived score infeasible. To evaluate how CHT might influence clinical management, we therefore used patient disposition as a proxy for the attending physician's overall risk assessment. Participants discharged from the ED to home were classified as "low risk," whereas those admitted to a ward or day-care unit were categorized as "nonlow risk."

Statistical Analysis

In this post hoc analysis evaluating diagnostic accuracy, a risk score populated with CHT data was considered the index test, and the occurrence of MACE or ACS was considered the reference test. Descriptive statistics are presented as means (SDs), 95% CIs, or proportions, as appropriate. Differences between groups with and without MACE or ACS were assessed using the Student *t* test for continuous variables and the Pearson chi-square test for binomial and categorical variables. Diagnostic accuracy was evaluated using cross-tabulation of the index test against the reference test. For each index test, performance was assessed by calculating sensitivity, specificity, PPV, and negative predictive value (NPV) for the reference test. The Pearson chi-square test was used to compare NPVs between index tests. Receiver operating characteristic curves were generated for each risk score, and the Hanley and McNeil method [47] was used to test for differences between areas under the receiver operating characteristic curve (AUCs). The miss rate was defined as the proportion of missed cases in the rule-out group. All analyses were performed using STATA, version 14.2 (StataCorp).

The sample size was determined based on the targeted precision of sensitivity and specificity in the original study objective, as previously described [25]. Briefly, because the prevalence of MACE and ACS in the study population was unknown, the

XSL•FO
RenderX

calculation was based on an assumed prevalence of 0.5 (50%), which maximizes the estimated sample size. To achieve a 0.03 (3%) precision for sensitivity and specificity, as calculated using nQuery, version 7.0 (Statistical Solutions Ltd), 1000 participants were required.

Ethical Considerations

This study complied with the Declaration of Helsinki and was approved by the Stockholm Regional Ethical Committee (now the Swedish Ethical Review Authority; reference number 2015/1955-31). All participants received oral and written information about the study and provided written informed consent before inclusion. Study data were deidentified before analysis, and only coded study IDs were used in the research dataset. Participants did not receive any financial or other compensation for their participation.

Results

Study Population Characteristics

During the study period, a total of 13,044 patients presented to the ED with a chief complaint of chest pain. During periods with on-duty research staff (office hours, evenings, and weekends), 1000 patients were consecutively included (Figure 1). The age and sex distribution of the study population (age: mean 55.3 years, SD 17.4 years; 456/1000, 45.60% female) closely aligned with that of the general chest pain population at Danderyd University Hospital (age: mean 57.6 years, SD 19.1 years; 49% female). Further baseline characteristics collected by CHT and selected ED variables, as well as patient disposition, are summarized in Tables 2 and 3.

XSL•FO
RenderX

Figure 1. Study flow diagram. CHT: computerized history taking; D-HEART: Danderyd HEART; ECG: electrocardiogram; ED: emergency department; EDACS-ADP: Emergency Department Assessment of Chest Pain Score combined with an Accelerated Diagnostic Protocol; EHR: electronic health record; HEART: History, ECG, Age, Risk Factors, and Troponin; T-MACS: Troponin-only Manchester Acute Coronary Syndrome.

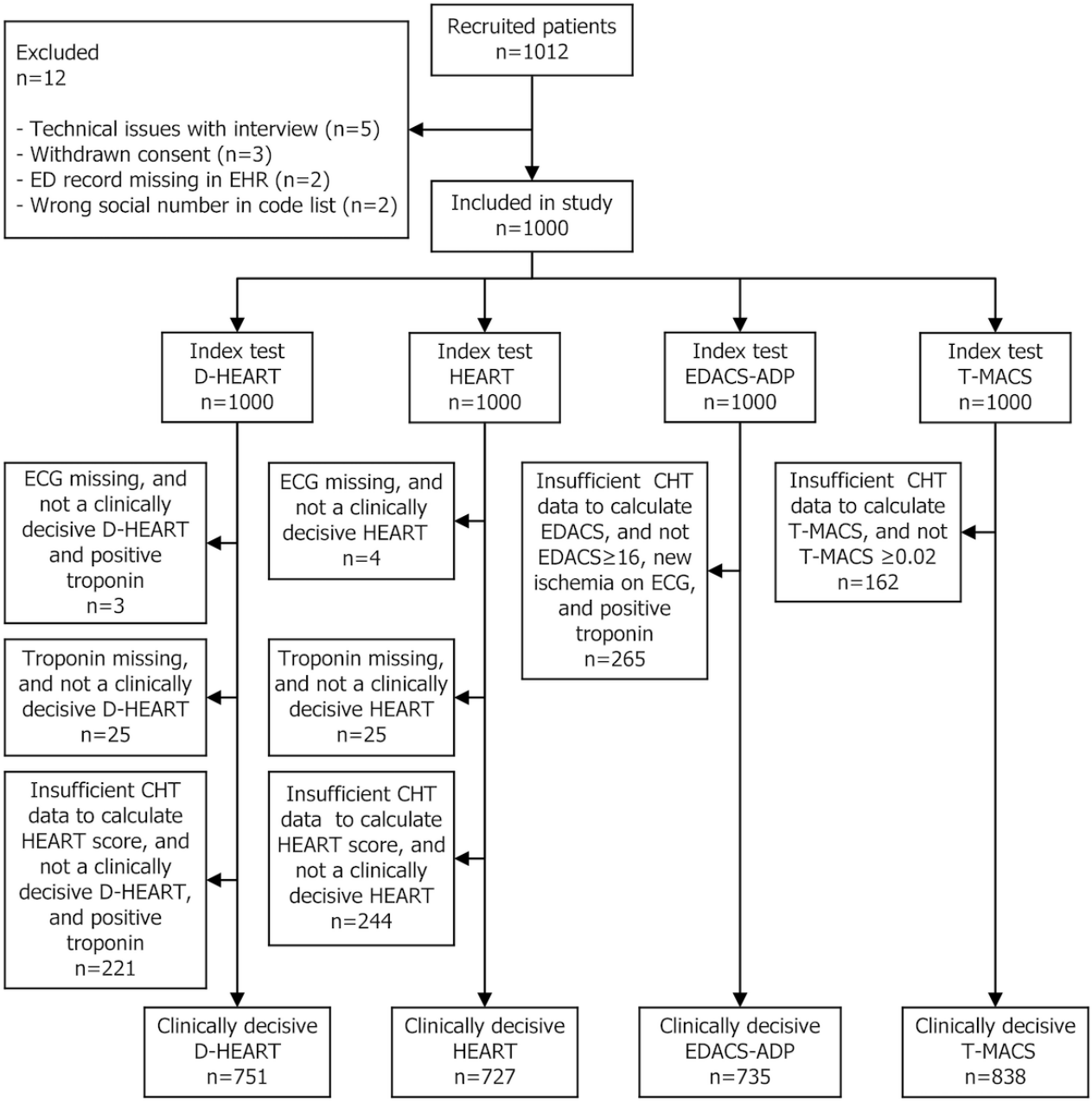

XSL•FO
RenderX

Table 2. Demography and background of all included patients divided into groups depending on outcome (30-day MACE[a] and ACS[b]).[c]

Characteristics	All		MACE		Non-MACE		*P* value	ACS		Non-ACS		*P* value
	Value	n	Value	n	Value	n		Value	n	Value	n	
Age (years), mean (SD)	54.7 (17.2)	1000	67.6 (11.0)	72	53.7 (17.1)	928	<.001	68.0 (10.9)	64	53.8 (17.1)	936	<.001
Sex (females), n (%)	456 (45.60)	1000	19 (26.39)	72	437 (47.09)	928	.001	18 (28.13)	64	438 (46.79)	936	.004
BMI (kg/m^2), mean (SD)	26.4 (4.7)	1000	26.6 (4.3)	72	26.3 (4.7)	928	.63	26.2 (3.9)	64	26.4 (4.7)	936	.82
Diabetes mellitus type 1 or 2, n (%)	62 (7.87)	788	8 (15.69)	51	54 (7.33)	737	.03	5 (11.36)	44	57 (7.66)	744	.37
Ongoing lipid-lowering medication, n (%)	128 (20.00)	640	21 (47.73)	44	107 (17.95)	596	<.001	18 (46.15)	39	110 (18.30)	601	<.001
Hypertension, n (%)	316 (41.31)	765	34 (64.15)	53	282 (39.61)	712	<.001	31 (67.39)	46	285 (39.64)	719	<.001
Family history of coronary artery disease, n (%)	200 (25.84)	774	18 (37.50)	48	182 (25.07)	726	.06	17 (41.46)	41	183 (24.97)	733	.02
Known coronary artery disease, n (%)	137 (16.10)	851	31 (55.36)	56	106 (13.33)	795	<.001	27 (55.10)	49	110 (13.72)	802	<.001
History of angina pectoris	88 (10.35)	850	23 (41.07)	56	65 (8.19)	794	<.001	20 (40.82)	49	68 (8.49)	801	<.001
History of myocardial infarction	82 (9.65)	850	21 (37.5)	56	61 (7.68)	794	<.001	18 (36.73)	49	64 (7.99)	801	<.001
History of percutaneous coronary intervention	76 (9.12)	833	16 (29.09)	55	60 (7.61)	778	<.001	13 (27.08)	48	63 (8.03)	785	<.001
History of coronary artery bypass graft	18 (2.16)	833	10 (18.18)	55	8 (1.03)	778	<.001	9 (18.75)	48	9 (1.15)	785	<.001
Current smoker, n (%)	90 (11.39)	790	5 (9.43)	53	85 (11.53)	737	.64	5 (10.87)	46	84 (11.29)	744	.91
Region of birth, n (%)												
Nordic countries	829 (82.90)	1000	63 (87.50)	72	766 (82.54)	928	.28	55 (85.94)	64	769 (82.16)	936	.50
Europe (outside the Nordic countries)	46 (4.60)	1000	1 (1.39)	72	45 (4.85)	928	.18	1 (1.56)	64	45 (4.81)	936	.23
Outside Europe	125 (12.50)	1000	8 (11.11)	72	117 (12.61)	928	.71	8 (12.50)	64	116 (12.39)	936	>.99
Occupational status, n (%)												
Active worker (employed or student)	616 (61.60)	1000	26 (36.11)	72	590 (63.58)	928	<.001	23 (35.94)	64	589 (62.93)	936	<.001
Not at work (unemployed or on sick leave)	69 (6.90)	1000	1 (1.4)	72	68 (7.33)	928	.06	1 (1.56)	64	67 (7.16)	936	.08
Retired	315 (31.50)	1000	45 (62.50)	72	270 (29.09)	928	<.001	40 (62.50)	64	274 (29.27)	936	<.001
Arrived at emergency department by ambulance, n (%)	189 (20.61)	917	14 (24.14)	58	175 (20.37)	859	.49	13 (25.49)	51	176 (20.32)	866	.37
Ongoing chest pain during computerized history taking, n (%)	544 (60.99)	892	19 (34.55)	55	525 (62.72)	837	<.001	16 (33.33)	48	524 (62.09)	844	<.001

[a]MACE: major adverse cardiac event.

[b]ACS: acute coronary syndrome.

[c]Self-reported medical history data derived from computerized history taking.

XSL•FO
RenderX

Table 3. Vital signs, electrocardiogram, circulating biomarkers, and disposition of all included patients divided into groups depending on outcome (30-day MACE[a] and ACS[b]) or not.[c]

Characteristic	All		MACE		Non-MACE		*P* value	ACS		Non-ACS		*P* value
	Value	n	Value	n	Value	n		Value	n	Value	n	
Vital parameters at triage, mean (SD)												
Systolic blood pressure (mmHg)	143 (22)	995	148 (22)	72	143 (22)	923	.06	148 (22)	64	143 (21)	931	.08
Diastolic blood pressure (mmHg)	83 (13)	993	81 (14)	72	84 (13)	921	.18	82 (15)	64	83 (13)	929	.30
Heart rate (beats/minute)	77 (16)	984	74 (15)	72	77 (16)	912	.15	74 (16)	64	77 (16)	920	.19
Respiration rate (breaths/minute)	16 (3)	984	16 (2)	72	16 (3)	912	.85	16 (2)	64	16 (3)	920	.86
Body temperature (°C)	36.8 (0.4)	947	36.6 (0.4)	70	36.8 (0.4)	877	<.001	36.6 (0.4)	63	36.8 (0.4)	884	.001
Electrocardiogram, n (%)[d]												
New signs diagnostic for ischemia	68 (6.8)	994	14 (19.7)	71	54 (5.9)	923	<.001	14 (22.2)	63	53 (5.7)	931	<.001
New nonspecific ST-T changes	131 (13.2)	994	20 (28.2)	71	111 (12.0)	923	<.001	16 (25.4)	63	115 (12.4)	931	.003
Normal or known ST-T alterations	795 (80.0)	994	37 (52.1)	71	758 (82.1)	923	<.001	33 (52.4)	63	758 (81.4)	931	<.001
High-sensitive troponin T values, n (%)												
>3× normal limit (>42 ng/l)	47 (4.8)	970	29 (40.8)	71	18 (2.0)	899	<.001	28 (44.4)	63	19 (2.1)	907	<.001
1-3× normal limit (15-42 ng/l)	112 (11.5)	970	19 (26.8)	71	93 (10.3)	899	<.001	16 (25.4)	63	96 (10.6)	907	<.001
Normal limit[e] (≤14 ng/l)	811 (83.6)	970	23 (32.4)	71	788 (87.7)	899	<.001	19 (30.2)	63	786 (86.7)	907	<.001
5-14 ng/l	397 (40.9)	970	15 (21.1)	71	367 (40.8)	899	.001	11 (17.5)	63	368 (40.6)	907	.004
1-hour troponin T elevated (>2 ng/l)	21 (2.2)	970	2 (2.8)	71	19 (2.1)	899	.69	2 (3.2)	63	19 (2.1)	907	.57
<5 ng/l	414 (42.7)	970	4 (5.6)	71	410 (45.6)	899	<.001	4 (6.3)	63	407 (44.9)	907	<.001
Admitted to the ward or day-care unit, n (%)	528 (53.3)	990	69 (95.8)	72	459 (50.0)	918	<.001	62 (96.9)	64	466 (50.3)	926	<.001
Ward (not via day-care unit)	203 (20.5)	990	61 (84.7)	72	141 (15.4)	918	<.001	57 (89.1)	64	146 (15.8)	926	<.001
Day-care unit	325 (32.8)	990	7 (9.7)	72	318 (34.6)	918	<.001	5 (7.8)	64	320 (34.6)	926	<.001
Day-care unit then to ward	33 (3.3)	7 (9.7)	7 (10)	72	26 (2.8)	918	.002	5 (7.8)	64	28 (3.0)	926	.04
Day-care unit then sent home	292 (29.5)	990	0 (0)	72	292 (31.8)	918	<.001	0 (0)	64	292 (31.5)	926	<.001

[a]MACE: major adverse cardiac event.

[b]ACS: acute coronary syndrome.

[c]Data were derived from the electronic health record.

[d] For definitions of electrocardiogram interpretation, see Table S1 in Multimedia Appendix 1).

[e]99th percentile (for definitions of electrocardiogram interpretation, see Table S1 in Multimedia Appendix 1).

XSL•FO
RenderX

Endpoints Reached Within 30 Days

Among the 1000 included patients, 72 experienced a MACE, including 65 with an ACS; 7 underwent coronary revascularization without an ACS diagnosis. No cases of cardiovascular death occurred (1 noncardiac death due to advanced lung cancer was excluded). Three patients diagnosed with unstable angina pectoris were not revascularized. The types of MACE and ACS are presented in Table 4. MACE and ACS were associated with increasing age, male sex, diabetes mellitus, lipid-lowering therapy, hypertension, and a history of coronary artery disease (Table 2). Primary diagnoses for participants discharged from either the ED or the day-care unit, as well as those hospitalized, are reported in Tables 5 and 6, respectively. Three patients were diagnosed with pulmonary embolism at the ED visit, assessed as low risk, and discharged according to local clinical protocols with appropriate anticoagulant therapy and dedicated specialist outpatient follow-up; these cases therefore do not represent missed or chronic diagnoses. During the study period, 2 patients receiving standard care experienced a MACE within 30 days after ED discharge to home.

Table 4. Major adverse cardiac events within 30 days among the 1000 participants.[a]

Outcome	7 days[b], n	30 days[c], n	Total, n
Major adverse cardiac event	70	2	72
Acute coronary syndrome	64	1	65
Unstable angina pectoris (I20.0)	22	0	22
Revascularization	19	0	19
Percutaneous coronary intervention	16	0	16
Coronary artery bypass grafting	3	0	3
No revascularization	3	0	3
Acute subendocardial myocardial infarction (I21.4)	33	1	34
Revascularization	24	1	25
Percutaneous coronary intervention	23	1	24
Coronary artery bypass grafting	1	0	1
No revascularization	9	0	9
Acute myocardial infarction, unspecified, type 1 (I21.9 + I98.1)	6	0	6
Revascularization	4	0	4
Percutaneous coronary intervention	4	0	4
Coronary artery bypass grafting	0	0	0
No revascularization	2	0	2
Acute myocardial infarction, unspecified, type 2 (I21.9 + I98.2)	3	0	3
Revascularization	0	0	0
Percutaneous coronary intervention	0	0	0
Coronary artery bypass grafting	0	0	0
No revascularization	0	0	0
Revascularization without acute coronary syndrome diagnosis	6	1	7
Percutaneous coronary intervention	6	1	7
Coronary artery bypass grafting	0	0	0
Cardiovascular death	0	0	0

[a]Data were derived from the electronic health record.

[b]7 days: outcome at discharge or ≤7 days from the emergency department visit.

[c]30 days: outcome after discharge or >7 days, and ≤30 days from the emergency department visit.

XSL•FO
RenderX

Table 5. Primary ICD-10[a] code and diagnosis for the 754 patients discharged home.[b]

ICD-10 code	Diagnosis	Values, n (%)
R07	Pain in throat and chest	574 (76.1)
R00-09 (excluding R07)	Symptoms and signs involving the circulatory and respiratory systems	57 (7.6)
R10-49	Symptoms and signs involving other organ systems	23 (3.1)
I48	Atrial fibrillation and flutter	16 (2.1)
R50-69	General symptoms and signs	14 (1.9)
M79	Other soft tissue disorders, not elsewhere classified	11 (1.5)
K29.7, K30.9, and K21.9	Gastritis, unspecified; functional dyspepsia; gastro-esophageal reflux disease without esophagitis	10 (1.3)
I20.1-9	Angina pectoris	4 (0.5)
I26.9	Pulmonary embolism without mention of acute cor pulmonale	3 (0.4)
J18.9	Pneumonia, unspecified	2 (0.3)
J45.9	Asthma, unspecified	2 (0.3)
I30.9	Acute pericarditis, unspecified	1 (0.1)
I63.9	Cerebral infarction, unspecified	1 (0.1)
J44.1	Chronic obstructive pulmonary disease with acute exacerbation, unspecified	1 (0.1)
N/A[c]	Other nonsevere diagnosis	35 (4.6)

[a]ICD-10: International Statistical Classification of Diseases and Related Health Problems, 10th Revision.

[b]Diagnoses reported by the physician in the electronic health record for patients with a chief complaint of chest pain who were discharged home (n=754), either directly from the emergency department (n=462) or via the day-care unit (n=292).

[c]N/A: not applicable.

XSL•FO
RenderX

Table 6. Primary discharge ICD-10[a] code and diagnosis for hospitalized patients (N=236).[b]

ICD code	Diagnosis	Values, n (%)
I20.0, I21	Acute coronary syndrome[c]	62 (26.3)
I20.1-9	Angina pectoris	14 (5.9)
I25.9	Chronic ischemic heart disease, unspecified	8 (3.4)
I05.0	Mitral stenosis	1 (0.4)
I10.9	Essential (primary) hypertension	1 (0.4)
I26-I28	Pulmonary heart disease and diseases of pulmonary circulation	6 (2.5)
I30-I52	**Other forms of heart disease**	35 (14.8)
	I30: Acute pericarditis	2 (5.7)
	I35: Nonrheumatic aortic valve disorders	3 (8.6)
	I40: Acute myocarditis	5 (14.3)
	I42: Cardiomyopathy	1 (2.9)
	I44: Atrioventricular and left bundle-branch block	1 (2.9)
	I47: Paroxysmal tachycardia	2 (5.7)
	I48: Atrial fibrillation and flutter	10 (28.6)
	I49: Other cardiac arrhythmias	3 (8.6)
	I50: Heart failure	8 (22.9)
I95-I99	Other and unspecified disorders of the circulatory system	1 (0.4)
J09-J18	Influenza and pneumonia	4 (1.7)
J20	Acute bronchitis	1 (0.4)
J44.1	Chronic obstructive pulmonary disease with acute exacerbation, unspecified	1 (0.4)
K00-93	Diseases of the digestive system	7 (3.0)
R07.4	Chest pain, unspecified	64 (27.1)
R00-09 (excluding R07.4)	Symptoms and signs involving the circulatory and respiratory systems	11 (4.7)
R10-49	Symptoms and signs involving other organ systems	1 (0.4)
R50-69	General symptoms and signs	2 (0.8)
N/A[d]	Other nonsevere diagnosis singularly reported	17 (7.2)

[a]ICD-10: International Statistical Classification of Diseases and Related Health Problems, 10th Revision.

[b]Data reported by the physician in the electronic health record for patients with a chief complaint of chest pain who were admitted from the ED (n=236). Of these, 15 patients were admitted to a noncardiac ward.

[c]Primary diagnosis of I20.0 (unstable angina pectoris) or I21 (acute myocardial infarction).

[d]N/A: not applicable.

Diagnostic Accuracy for a MACE or ACS

A clinically decisive score could be calculated using CHT-acquired data in 751 out of 1000 (75.10%; D-HEART), 727 out of 1000 (72.70%; HEART), 735 out of 1000 (73.50%; EDACS-ADP), and 838 out of 1000 (83.80%; T-MACS) patients (Figure 1). For D-HEART, 24 of 751 (3.2%) patients were included despite insufficient CHT data, based on rule-in troponin findings. Patients with insufficient CHT data to calculate the risk scores were generally younger and, for all risk scores except T-MACS, more often female (Table S2 in Multimedia Appendix 1). No adverse events occurred during the CHT interviews.

For MACE, all 4 risk scores demonstrated high NPVs of 0.99 (95% CI 0.97-1.00). Sensitivity ranged from 0.91 (95% CI 0.81-0.97; HEART) to 0.97 (95% CI 0.90-1.00; T-MACS <0.02). The MACE miss rate among patients classified as low risk was 2 out of 286 (0.70%; T-MACS), 4 out of 416 (1.0%; D-HEART), 4 out of 346 (1.2%; EDACS-ADP), and 6 out of 406 (1.5%; HEART). The proportion of patients ruled out was highest for HEART (406/727, 55.8%) and D-HEART (416/751, 55.4%), intermediate for EDACS-ADP (346/744, 46.5%), and lowest for T-MACS (286/838, 34.1%; Table 7). Using T-MACS with a threshold of <0.05 for low risk yielded a sensitivity, specificity, PPV, and NPV of 0.89 (95% CI 0.78-0.95), 0.59 (95% CI 0.55-0.62), 0.16 (95% CI 0.12-0.20), and 0.98 (95% CI 0.97-0.99), respectively. Across CHT-derived risk scores, diagnostic performance for ACS was comparable to that for

XSL•FO
RenderX

MACE (Table S3 in Multimedia Appendix 1). T-MACS <0.02 provided higher sensitivity and NPV but lower specificity compared with T-MACS <0.05 (Table S4 in Multimedia Appendix 1). No differences in AUC between the risk scores were observed for either MACE or ACS (Figure 2 and Table S5 in Multimedia Appendix 1). Excluding the 3 patients diagnosed with unstable angina pectoris who did not undergo revascularization (Table S6 in Multimedia Appendix 1), or the patient who died from a noncardiac cause, had a negligible effect on the results.

Table 7. Rule-out performance of risk scores populated with data derived from computerized history taking for a major adverse cardiac event within 30 days.[a]

Risk score	Sensitivity (95% CI)	Specificity (95% CI)	Positive predictive value (95% CI)	Negative predictive value (95% CI)	Miss rate, n/N (%)	Patients ruled out, n/N (%)
D-HEART[b] (n=751)	0.94 (0.86-0.98)	0.60 (0.57-0.64)	0.19 (0.15-0.24)	0.99 (0.98-1.00)	4/416 (0.96)	416/751 (55.39)
HEART[c] (n=727)	0.91 (0.81-0.97)	0.61 (0.57-0.64)	0.19 (0.15-0.23)	0.99 (0.97-1.00)	6/406 (1.48)	406/727 (55.85)
EDACS-ADP[d] (n=744)	0.94 (0.86-0.98)	0.51 (0.48-0.55)	0.17 (0.13-0.21)	0.99 (0.97-1.00)	4/346 (1.16)	346/744 (46.51)
T-MACS[e] (n=838)	0.97 (0.90-1.00)	0.37 (0.34-0.39)	0.12 (0.10-0.15)	0.99 (0.98-1.00)	2/286 (0.70)	286/838 (34.13)

[a]When T-MACS with a threshold of <0.05 for low risk was applied, the numbers of true positives, false positives, false negatives, and true negatives were 55, 294, 7, and 418, respectively. Sensitivity, specificity, PPV, and NPV were 0.89 (95% CI 0.78-0.95), 0.59 (95% CI 0.55-0.62), 0.16 (95% CI 0.12-0.20), and 0.98 (95% CI 0.97-0.99), respectively.

[b]D-HEART: Danderyd HEART.

[c]HEART: History, ECG, Age, Risk Factors, and Troponin.

[d]EDACS-ADP: Emergency Department Assessment of Chest Pain Score combined with an Accelerated Diagnostic Protocol.

[e]T-MACS: Troponin-only Manchester Acute Coronary Syndrome.

Figure 2. Receiver operating curves for HEART, D-HEART, EDACS-ADP and T-MACS for a 30-day major adverse cardiac event. AUC: area under the receiver operating curve; D-HEART: Danderyd HEART; EDACS-ADP: Emergency Department Assessment of Chest Pain Score combined with an Accelerated Diagnostic Protocol; HEART: History, ECG, Age, Risk Factors, and Troponin; T-MACS: Troponin-only Manchester Acute Coronary Syndrome.

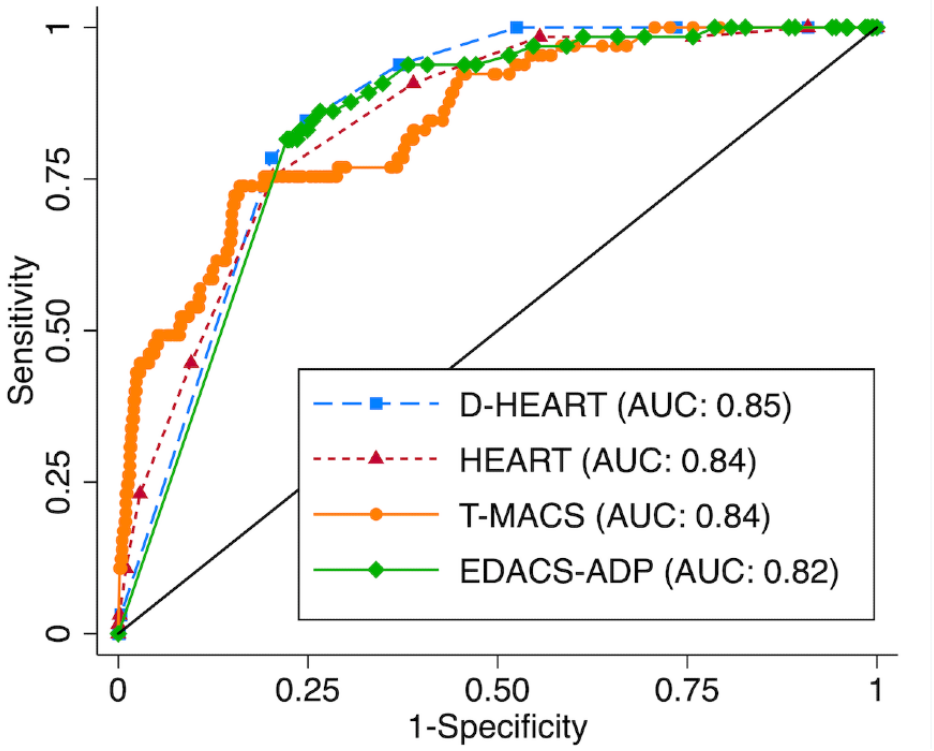

D-HEART Versus HEART

When comparing the NPV (95% CI) for ruling out MACE and ACS, no difference (P<.001) was found between the D-HEART and HEART scores: 0.99 (0.98-1.00) versus 0.99 (0.97-1.00) for MACE, and 0.99 (0.98-1.00) versus 0.99 (0.97-1.00) for ACS, respectively. Similarly, no difference in diagnostic accuracy was observed when comparing the AUCs (Table S5 in Multimedia Appendix 1).

XSL•FO
RenderX

Risk Reclassification for MACE or ACS

Of the 1000 participants, 528 were admitted, with 203 admitted to the ward and 325 to the day-care unit. Of the day-care unit admissions, 33 were subsequently admitted to the ward, and 292 were discharged home (Table 3). If CHT data had been used instead of standard management to calculate D-HEART, HEART, and EDACS-ADP, there would have been a net reclassification from "nonlow risk" to "low risk" in 89 (16.9%), 89 (16.9%), and 31 (5.9%) cases, respectively. However, due to lower specificity, use of T-MACS would instead have resulted in 91 (17.2%) patients being net reclassified from "low risk" to "nonlow risk." These reclassifications correspond to a net shift in MACE or ACS outcomes of +2 (D-HEART), +4 (HEART), +2 (EDACS-ADP), and 0 (T-MACS; Table 8).

Table 8. Reclassification and MACE[a] or ACS[b] within reclassified groups if risk scores populated with computerized history-taking data had been used for the disposition of patients instead of physician management.[c]

Reclassification	D-HEART[d]	HEART[e]	EDACS-ADP[f]	T-MACS[g]
Low to nonlow risk, n	93	97	128	220
MACE or ACS, n	1	1	2	1
Nonlow risk to low risk, n	182	186	159	129
MACE or ACS, n	3	5	4	1
Difference[h]	89	89	31	–91
MACE or ACS, n	+2	+4	+2	0
Proportion of admissions, n/N (%)	89/528 (–16.9%)	89/528 (–16.9%)	31/528 (–5.9%)	91/528 (+17.2%)

[a]MACE: major adverse cardiac event.

[b]ACS: acute coronary syndrome.

[c]Admission to the ward or day-care unit (n=528) was considered a proxy for nonlow risk, and discharge to home (n=462) a proxy for low risk. Data for 1 participant were missing, and 9 left the ED without seeing a physician.

[d]D-HEART: Danderyd HEART.

[e]HEART: History, ECG, Age, Risk Factors, and Troponin.

[f]EDACS-ADP: Emergency Department Assessment of Chest Pain Score combined with an Accelerated Diagnostic Protocol.

[g]T-MACS: Troponin-only Manchester Acute Coronary Syndrome.

[h]The difference indicates the n and % change in admissions.

Discussion

Principal Findings

This study shows that CHT can be used to calculate 4 risk scores in a majority (727-838/1000, 72.70%-83.80%) of patients presenting with acute chest pain, with good diagnostic performance for ruling out 30-day 3-point MACE or ACS, yielding NPVs of 0.99 (95% CI 0.97-1.00) for both outcomes. Sensitivities ranged from 91% to 97%, implying that, out of 100 patients, 3-9 with MACE might be misclassified. While this may initially appear concerning, the clinically more relevant miss rate among patients classified as low risk remained low for D-HEART (4/416, 0.96%) and T-MACS (2/286, 0.70%), aligning with commonly accepted safety thresholds for MACE (NPV ≥0.99 and miss rate <1%). By contrast, the HEART score (6/406, 1.5%) and EDACS-ADP (4/346, 1.2%) showed higher miss rates [16,48]. However, the lower bounds of the 95% CIs for sensitivity allow miss rates exceeding accepted rule-out safety thresholds; therefore, these findings warrant cautious interpretation.

In addition, CHT-derived chest pain risk scores reclassified a substantial proportion (up to 89/528, 16.9%) of patients admitted with acute chest pain, initially managed as suspected ACS, from "nonlow risk" to "low risk" at the time of initial ED assessment, potentially facilitating a decision for discharge rather than hospital admission. However, hospital admission may be driven by factors other than ACS risk, and reclassification to low risk does not necessarily eliminate the need for admission; thus, the aforesaid figure (ie, 89/528, 16.9%) should be viewed as a theoretical maximum, and real-world discharge gains may be substantially lower. The attendant trade-off is an increased miss rate of MACE or ACS among reclassified patients. In our study, the number of missed events (0-4 events, depending on the risk score used) was similar to that observed under standard management (2 events). While our findings indicate that CHT is a promising tool for risk score determination, further studies are warranted to validate these results and to ascertain safety.

Comparison With Prior Work

Our findings on diagnostic accuracy align with multiple validation studies in which traditional physician-acquired medical histories were used, demonstrating the efficacy of risk scores in safely ruling out MACE in the ED [7-10,31,49]. However, this appears to be the first study to report on the performance of an automated method for collecting the medical history required to calculate these risk scores. The assessed risk scores (D-HEART, HEART, EDACS-ADP, and T-MACS) showed similar performance for ruling out 30-day MACE or ACS. Our findings are consistent with prior work. A meta-analysis of 25 studies (n=25,266) reported a pooled sensitivity of 0.96 (95% CI 0.93-0.98) and an NPV of 0.99 (95% CI 0.98-0.99) for HEART [7]. A meta-analysis of the EDACS score (8 studies, n=11,578) reported a pooled sensitivity of 0.95

XSL•FO
RenderX

(95% CI 0.90-0.99) but did not specifically report NPV [50]. External validation of EDACS-ADP in a North American cohort (n=763) demonstrated a sensitivity and NPV of 1.00 (95% CI 0.94-1.00) and 1.00 (95% CI 0.99-1.00), respectively [8]. For T-MACS, secondary analyses of 4 UK cohort studies (n=1459) reported a sensitivity and NPV of 0.98 (95% CI 0.95-1.00) and 0.99 (95% CI 0.98-1.00), respectively [32]. However, the HEART score, which has undergone more extensive validation, consistently demonstrates good performance and is endorsed by international societies [5,51].

The only methodological difference between HEART and D-HEART in our study was the use of the ESC 0/1-hour troponin algorithm in D-HEART. Thus, the comparison primarily reflects the added value of serial troponin testing rather than differences related to CHT. D-HEART demonstrated equivalent performance for MACE or ACS rule out compared with the original HEART score using only a single baseline troponin measurement. This could be interpreted as support for using the simpler HEART score without repeated troponin measurements. However, this contrasts with findings showing that the original HEART score has somewhat lower diagnostic accuracy than a combination of the HEART score and a 0/1-hour troponin algorithm [52]. As the HEART score was primarily developed and validated for safe rule out of MACE, it is reasonable to assume that integrating a 0/1-hour algorithm enhances rule-in capability, as previously suggested [53]. Accordingly, the superior safety profile observed for D-HEART in this study is likely driven by the integration of serial troponin measurements rather than the CHT-derived history component itself. Nevertheless, the absolute difference in missed events was small (4 vs 6), and neither NPV nor AUC differed significantly between the scores. Thus, the difference in performance between the 2 scores may not be statistically significant. Taken together, D-HEART appears to be the most promising option for CHT-based risk stratification; however, future studies are needed to confirm these findings.

Our findings show an AUC of 0.82-0.85 and an NPV of 0.98-0.99, depending on the risk score used. While there is no universal benchmark for AUC or NPV in risk stratification, an AUC of 0.8-0.9 is generally considered excellent, and >0.9 outstanding [54]. Interpretation of NPV varies by clinical context; for MACE or ACS, a high NPV is desirable for safe rule out [55]. Overall, our findings suggest that risk calculation using CHT data performs at least as well as traditional history taking. It is reassuring that all scores reproduced similar results. However, further refinement of CHT data collection could potentially enhance MACE rule-out capability. Notably, a strategy in which CHT data are used for risk calculation requires further validation. Importantly, risk score algorithms should always be considered as decision-support tools and used in complement to physician expertise.

Clinical Implications

A significant fraction (up to 89/528, 16.9%) of participants were reclassified from “nonlow risk” to “low risk” when using D-HEART, HEART, and EDACS-ADP populated with CHT data, compared with standard management. These findings align with those of an external validation study, which reported a reduction in hospital admissions for patients with acute chest pain of approximately 15% following implementation of the HEART score [56]. Improvements in classification may reduce anxiety, unnecessary examinations, and health care resource use [57]. However, a large Dutch multicenter randomized controlled trial observed a limited impact of risk score calculations on resource utilization, attributing this effect to nonadherence to management recommendations [58]. Notably, when using T-MACS, more patients would be reclassified from “low risk” to “nonlow risk,” potentially leading to increased admissions (91/528, +17.2%) compared with standard care, due to its lower specificity. Applying a higher threshold for T-MACS (<0.05 for “low risk”) increases specificity but results in a lower NPV, supporting the conclusion that this is not the preferred score for patient disposition. Further studies on optimal CHT implementation and its impact on resource utilization and quality of care in chest pain management are warranted.

This study extends our previous reports demonstrating substantial agreement between data collected using CHT and standard history taking for risk score assessment [30]. However, for the HEART score, the History component was derived from predefined patient-reported CHT criteria rather than physician gestalt (ie, physician’s subjective clinical judgment), as applied in the original derivation studies. Given the previously demonstrated low-to-moderate agreement between CHT and EHR data for chest pain characteristics [30], this substitution may affect construct validity and limit direct comparability with physician-acquired HEART scores. At the same time, prior work suggests that CHT captures more complete and detailed symptom data than are routinely documented by physicians in the EHR, which may partly explain the observed discrepancies. Our previous studies also showed that risk score calculation was feasible to a much greater extent when using CHT-collected data compared with physician-reported EHR data (74%-83% vs 10%-31%, depending on the risk score used), and that a HEART score excluding the troponin variable (HEART score) can effectively predict clinical outcomes [54]. This analysis underscores the broader utility of CHT in facilitating the calculation of established risk scores commonly used in clinical practice. Moreover, this study introduces the concept of a clinically decisive risk score, defined as reaching a specific threshold for “nonlow risk” (eg, D-HEART≥4 or T-MACS≥0.02), to better reflect how these risk stratification tools are applied in practice.

The time required to obtain sufficient data for risk score calculation using CHT (median 23 minutes) may appear long. However, the intention of CHT was not to replace physician history taking, but rather to use waiting time in the ED to collect and supplement crucial information for risk stratification. Notably, the median length of stay for low- to intermediate-risk patients (ie, RETTS orange, yellow, green, and blue) in the ED at the time of the study was approximately 4 hours, suggesting that CHT can be completed without disrupting workflow. Nevertheless, future developments of CHT should focus on shortening interview duration and adapting content to the clinical context and urgency. Moreover, although physician-interpreted ECGs were used by design to isolate the effect of CHT on chest

XSL•FO
RenderX

pain management, the incorporation of automated ECG analysis may further enhance efficiency.

Considering the high volume of patients presenting to the ED with chest pain, automating risk stratification through CHT could streamline the ED pathway: first, by ensuring complete and structured capture of medical history and risk factors; second, by reducing physician time spent on data gathering; third, by providing real-time support for clinical decision-making; and finally, by enabling faster and safer disposition decisions that identify patients suitable for discharge and reduce resource use.

Strengths and Limitations

Key strengths of this study include its prospective cohort design, a large and representative ED chest pain population, evaluation under real-world conditions in the typical clinical environment in which the tool would be used, and reliable outcomes defined by strict, well-established criteria. The consistency of the findings across multiple validated risk scores underscores the robustness of the results. Furthermore, the generic layout of the CHT software, which is not limited to cardiology or the ED setting, suggests that these findings may apply to other clinical contexts. Finally, the study was initiated and conducted within an academic setting, without any commercial interests.

However, several limitations warrant consideration. First, the single-center design introduces a risk of selection bias, potentially limiting population diversity. In addition, the use of a convenience sample comprising a minority of eligible chest pain presentations may limit external validity, as enrolled patients may be more compliant, clinically stable, or technologically literate than the overall ED population. This bias may have been further exacerbated by the exclusion of patients unable to complete CHT and by recruitment occurring mainly during office hours and evenings, due to research staff availability. However, the proportion of patients with chest pain presenting at night was small, and the demographics of study participants were similar to those of the overall ED chest pain population, reducing the likelihood of major systematic differences. Nevertheless, our findings should be interpreted with caution, and future large multicenter studies are warranted to ascertain the generalizability of these results to other health care settings. Second, variability in physician performance and the unique ED setup, which includes a dedicated cardiology unit and a day-care unit staffed by cardiologists, may lead to an underestimation of CHT performance compared with settings with less specialist expertise in the ED. At the same time, this organizational structure may limit the generalizability of our findings to more diverse clinical environments. Third, risk score determination using CHT data was not feasible for approximately one-quarter of participants. We have previously shown that premature termination of the CHT interview was associated with ED discharge and fatigue, most commonly among older adults (≥70 years), individuals born outside Europe, and retirees, thereby limiting conclusions for these groups [22]. In our analysis, however, patients with insufficient CHT data were significantly younger than those with a clinically decisive score (mean age 51.0 vs 56.2 years; see Table S2 in Multimedia Appendix 1), likely reflecting the larger proportion of adults aged 18-69 years in the study population and their higher frequency of interview termination, often associated with early discharge [22]. Nevertheless, diagnostic performance was assessed only in patients who provided sufficient CHT data for risk score determination, which may have overestimated real-world performance. This finding underscores the need for a more user-friendly tool to reduce missing data in routine care. Fourth, ACS outcomes were adjudicated by a single board-certified cardiologist, which may introduce adjudication bias compared with evaluation by 2 or more independent specialists. However, Swedish discharge ICD codes have a high PPV (85%-95%) for acute myocardial infarction and related diagnoses, supporting the reliability of the outcome classification [45,46]. Further, outcomes were retrieved through passive review of the shared EHR system in Region Stockholm; therefore, events occurring outside the region may not have been captured, potentially leading to an overestimation of the NPV. Mortality status is, however, captured through linkage between the regional EHR system and the national population registry. The proportion of nonfatal events is likely small, as routine follow-up after a MACE is systematically conducted at the patient's home hospital, even when the event occurs outside the region. Fifth, given that patients with MACE were significantly older than those without (ie, non-MACE; 67.6 vs 53.7 years), there is a risk of spectrum bias if CHT fails to capture data from patients at the highest risk. In our cohort, however, insufficient CHT data were not more common among older patients (Table S2 in Multimedia Appendix 1), reducing concerns that missing CHT data systematically occurred in older or higher-risk individuals. Sixth, inclusion in the study could occur either before or after the patient had been evaluated by a physician. This may have introduced recall bias, as patients might provide different responses when reporting their history a second time. The effect of interview order on patient responses was not assessed and therefore remains an unmitigated limitation. Seventh, patient disposition reflects not only the medical history obtained by the physician but also additional information available at the time of assessment, including ECG findings, blood test results, and prior EHR documentation. These factors likely influenced clinical decision-making, thereby limiting direct comparability with CHT-derived risk scores. However, this approach was intentionally chosen to evaluate the potential impact of CHT within current clinical management. Future multimodal studies should aim to incorporate such data directly into risk stratification models. Eighth, the primary focus of this study was the safe rule out of MACE or ACS using a clinically decisive risk score. This implies that once patients exceed the threshold for "nonlow risk" (eg, HEART ≥4), they are considered ruled in. However, some patients within this clinically decisive risk group who did not complete the CHT interview might have achieved even higher total risk scores. Consequently, our findings may be more robust for rule out than for rule in. Conversely, excluding patients with insufficient CHT data who also did not meet ECG or troponin criteria for "nonlow risk" may have resulted in the systematic exclusion of a specific patient subset, potentially leading to an overestimation of NPVs. Accordingly, interpretation of both rule-out and rule-in performance should account for these methodological limitations, and further evaluation across the full risk spectrum

XSL•FO
RenderX

is warranted. Finally, this study carries a potential risk of verification bias. We believe this risk to be minimal, as patients were managed according to standard care protocols independent of CHT-derived risk scores, which were not available to the treating physicians during clinical decision-making.

Future Directions

Future research on CHT in chest pain risk stratification should include customizing CHT for the ED setting to reduce interview time, extending its use to early prehospital triage, and broadening its scope to support diagnosis of a wider array of acute conditions, such as pulmonary embolism, aortic dissection, and other noncardiac causes of chest pain. In addition, future studies should explore the integration of multimodal variables into risk stratification models, and the incorporation of multiple languages into the CHT software could enable history taking in patients' native languages.

Conclusions

Automated medical history taking using CHT can generate reliable risk scores in the majority of patients presenting with acute chest pain, with good diagnostic performance for ruling out 30-day 3-point MACE (ie, ACS diagnosis, revascularization, or cardiovascular death) and ACS. CHT-derived risk scores performed comparably to previously validated physician-acquired scores, and accepted safety thresholds were met by D-HEART and T-MACS, whereas the HEART score and EDACS-ADP showed values slightly above the commonly accepted <1% threshold; however, given the wide sensitivity CIs, these findings should be interpreted with caution. Compared with the HEART score, the improved safety of D-HEART appears to be primarily attributable to the incorporation of serial 0/1-hour troponin testing rather than the CHT-derived history itself. Application of CHT-derived risk scores indicated that 89 of the 528 (16.9%) patients could potentially be reclassified from "nonlow risk" to "low risk" at the time of initial ED assessment, which may support discharge decisions in selected patients, although admission may still be required for reasons unrelated to ACS. Consistent with prior chest pain risk score studies, any gains in discharge rates must be balanced against the risk of missed events among reclassified patients. Among the scores assessed, D-HEART appears most promising for CHT-based risk stratification. However, further large multicenter studies are warranted to confirm these findings before CHT can be established in routine clinical practice.

Acknowledgments

The authors thank Professor D. Zakim, Karolinska Institutet, Stockholm, Sweden, developer of the expert system software (CLEOS) used for CHT and a contributor to the trial protocol, for invaluable support. The authors also thank statistician F. K. Johansson, MSc, Karolinska Institutet, Stockholm, Sweden, for expert statistical advice. This work was supported by the Robert Bosch Stiftung, Stuttgart, Germany (grant 11.5.1000.0258.0); Region Stockholm, ALF project, Stockholm, Sweden (grants FoUI-954046 and FoUI-962663); Stiftelsen Hjärtat, Stockholm, Sweden; Fond 176, Danderyd University Hospital, Stockholm, Sweden; and institutional funds from the Department of Learning, Informatics, Management, and Ethics, Karolinska Institutet, Stockholm, Sweden.

Data Availability

The datasets used or analyzed during this study are available from the corresponding author (HB) on reasonable request.

Conflicts of Interest

All patent rights and copyrights to the technology, language, images, and knowledge content are assigned, without royalty rights, to Karolinska Institutet, Stockholm, Sweden, a public university. HB reports speaker honoraria from AstraZeneca. TK reports research grants to Karolinska Institutet from Amgen, Medtronic, and ReCor Medical, all outside the submitted work. JS reports speaker honoraria from AstraZeneca, Bayer, Novo Nordisk, and Medtronic, and shareholdings in Beats Vascular Health.

Multimedia Appendix 1
Additional analysis.
[DOCX File , 52 KB - jmir_v28i1e76087_app1.docx]

References

1. Aalam AA, Alsabban A, Pines JM. National trends in chest pain visits in US emergency departments (2006-2016). Emerg Med J 2020 Nov;37(11):696-699. [doi: 10.1136/emermed-2020-210306] [Medline: 32900857]
2. Bjørnsen LP, Naess-Pleym LE, Dale J, Grenne B, Wiseth R. Description of chest pain patients in a Norwegian emergency department. Scand Cardiovasc J 2019 Feb;53(1):28-34. [doi: 10.1080/14017431.2019.1583362] [Medline: 30760035]
3. Af Ugglas B, Djärv T, Ljungman PLS, Holzmann MJ. Emergency department crowding associated with increased 30-day mortality: a cohort study in Stockholm Region, Sweden, 2012 to 2016. J Am Coll Emerg Physicians Open 2020 Dec;1(6):1312-1319 [FREE Full text] [doi: 10.1002/emp2.12243] [Medline: 33392538]

XSL•FO
RenderX

4. Byrne RA, Rossello X, Coughlan JJ, Barbato E, Berry C, Chieffo A, ESC Scientific Document Group. 2023 ESC Guidelines for the management of acute coronary syndromes. Eur Heart J 2023 Oct 12;44(38):3720-3826. [doi: 10.1093/eurheartj/ehad191] [Medline: 37622654]
5. Stepinska J, Lettino M, Ahrens I, Bueno H, Garcia-Castrillo L, Khoury A, et al. Diagnosis and risk stratification of chest pain patients in the emergency department: focus on acute coronary syndromes. A position paper of the Acute Cardiovascular Care Association. Eur Heart J Acute Cardiovasc Care 2020 Feb;9(1):76-89 [FREE Full text] [doi: 10.1177/2048872619885346] [Medline: 31958018]
6. Writing Committee, Kontos MC, de Lemos JA, Deitelzweig SB, Diercks DB, Gore MO, et al. 2022 ACC expert consensus decision pathway on the evaluation and disposition of acute chest pain in the emergency department: a report of the American College of Cardiology Solution Set Oversight Committee. J Am Coll Cardiol 2022 Nov 15;80(20):1925-1960 [FREE Full text] [doi: 10.1016/j.jacc.2022.08.750] [Medline: 36241466]
7. Laureano-Phillips J, Robinson RD, Aryal S, Blair S, Wilson D, Boyd K, et al. HEART score risk stratification of low-risk chest pain patients in the emergency department: a systematic review and meta-analysis. Ann Emerg Med 2019 Aug;74(2):187-203. [doi: 10.1016/j.annemergmed.2018.12.010] [Medline: 30718010]
8. Flaws D, Than M, Scheuermeyer FX, Christenson J, Boychuk B, Greenslade JH, et al. External validation of the emergency department assessment of chest pain score accelerated diagnostic pathway (EDACS-ADP). Emerg Med J 2016 Sep;33(9):618-625. [doi: 10.1136/emermed-2015-205028] [Medline: 27406833]
9. Meier M, Boeddinghaus J, Nestelberger T, Koechlin L, Lopez-Ayala P, Wussler D, Miró, APACE investigators. Comparing the utility of clinical risk scores and integrated clinical judgement in patients with suspected acute coronary syndrome. Eur Heart J Acute Cardiovasc Care 2023 Oct 25;12(10):693-702 [FREE Full text] [doi: 10.1093/ehjacc/zuad081] [Medline: 37435949]
10. Steiro O, Tjora HL, Langørgen J, Bjørneklett R, Nygård OK, Skadberg, et al. Clinical risk scores identify more patients at risk for cardiovascular events within 30 days as compared to standard ACS risk criteria: the WESTCOR study. Eur Heart J Acute Cardiovasc Care 2021 May 11;10(3):287-301. [doi: 10.1093/ehjacc/zuaa016] [Medline: 33620429]
11. Pena M, Tsao J, Falaiye T, Kazan V, Li M, Daouk F, et al. Why won't emergency physicians discharge patients with a low HEART score from an observation unit without further evaluation? Crit Pathw Cardiol 2020 Dec;19(4):195-199. [doi: 10.1097/HPC.0000000000000236] [Medline: 32769484]
12. Ras M, Reitsma JB, Hoes AW, Six AJ, Poldervaart JM. Secondary analysis of frequency, circumstances and consequences of calculation errors of the HEART (history, ECG, age, risk factors and troponin) score at the emergency departments of nine hospitals in the Netherlands. BMJ Open 2017 Oct 22;7(10):e017259 [FREE Full text] [doi: 10.1136/bmjopen-2017-017259] [Medline: 29061617]
13. Wu WK, Yiadom MYAB, Collins SP, Self WH, Monahan K. Documentation of HEART score discordance between emergency physician and cardiologist evaluations of ED patients with chest pain. Am J Emerg Med 2017 Jan;35(1):132-135 [FREE Full text] [doi: 10.1016/j.ajem.2016.09.058] [Medline: 27745728]
14. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
15. Lopez-Jimenez F, Attia Z, Arruda-Olson AM, Carter R, Chareonthaitawee P, Jouni H, et al. Artificial intelligence in cardiology: present and future. Mayo Clin Proc 2020 May;95(5):1015-1039. [doi: 10.1016/j.mayocp.2020.01.038] [Medline: 32370835]
16. Gulati M, Levy PD, Mukherjee D, Amsterdam E, Bhatt DL, Birtcher KK, et al. 2021 AHA/ACC/ASE/CHEST/SAEM/SCCT/SCMR guideline for the evaluation and diagnosis of chest pain: executive summary: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Circulation 2021 Nov 30;144(22):e368-e454 [FREE Full text] [doi: 10.1161/CIR.0000000000001030] [Medline: 34709928]
17. Zakim D, Braun N, Fritz P, Alscher MD. Underutilization of information and knowledge in everyday medical practice: evaluation of a computer-based solution. BMC Med Inform Decis Mak 2008 Nov 05;8:50 [FREE Full text] [doi: 10.1186/1472-6947-8-50] [Medline: 18983684]
18. Zakim D, Fritz C, Braun N, Fritz P, Alscher MD. Computerized history-taking as a tool to manage dyslipidemia. Vasc Health Risk Manag 2010 Nov 12;6:1039-1046 [FREE Full text] [doi: 10.2147/VHRM.S14302] [Medline: 21127700]
19. Zakim D, Brandberg H, El Amrani S, Hultgren A, Stathakarou N, Nifakos S, et al. Computerized history-taking improves data quality for clinical decision-making-comparison of EHR and computer-acquired history data in patients with chest pain. PLoS One 2021;16(9):e0257677 [FREE Full text] [doi: 10.1371/journal.pone.0257677] [Medline: 34570811]
20. Hourizadeh J, Zeltser R, Makaryus AN, Druz R. Real-world evaluation of a chest pain digital triage platform at the point of care. South Med J 2023 Nov;116(11):857-862. [doi: 10.14423/SMJ.0000000000001620] [Medline: 37913803]
21. Sundberg K, Adeli A, Brandberg H, Spaak J, Koch S, Sundberg CJ, et al. User experience of self-reported computerized medical history taking for acute chest pain: The Clinical Expert Operating System Chest Pain Danderyd Study. Health Expect 2022 Dec;25(6):3053-3061 [FREE Full text] [doi: 10.1111/hex.13612] [Medline: 36148691]
22. Brandberg H, Sundberg CJ, Spaak J, Koch S, Zakim D, Kahan T. Use of self-reported computerized medical history taking for acute chest pain in the emergency department - the Clinical Expert Operating System Chest Pain Danderyd Study

XSL•FO
RenderX

(CLEOS-CPDS): prospective cohort study. J Med Internet Res 2021 Apr 27;23(4):e25493 [FREE Full text] [doi: 10.2196/25493] [Medline: 33904821]
23. Fritzell K, Brandberg H, Spaak J, Koch S, Sundberg CJ, Zakim D, et al. Patient perspectives on digital technology and experiences of computerized history-taking for chest pain management in the emergency department: CLEOS-CPDS prospective cohort study. JMIR Med Inform 2025 Jun 17;13:e65568-e65568 [FREE Full text] [doi: 10.2196/65568] [Medline: 40526888]
24. Berdahl CT, Henreid AJ, Pevnick JM, Zheng K, Nuckols TK. Digital tools designed to obtain the history of present illness from patients: scoping review. J Med Internet Res 2022 Nov 17;24(11):e36074 [FREE Full text] [doi: 10.2196/36074] [Medline: 36394945]
25. Brandberg H, Kahan T, Spaak J, Sundberg K, Koch S, Adeli A, et al. A prospective cohort study of self-reported computerised medical history taking for acute chest pain: protocol of the CLEOS-Chest Pain Danderyd Study (CLEOS-CPDS). BMJ Open 2020 Jan 21;10(1):e031871 [FREE Full text] [doi: 10.1136/bmjopen-2019-031871] [Medline: 31969363]
26. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015 Oct 28;351:h5527 [FREE Full text] [doi: 10.1136/bmj.h5527] [Medline: 26511519]
27. Widgren BR, Jourak M. Medical Emergency Triage and Treatment System (METTS): a new protocol in primary triage and secondary priority decision in emergency medicine. J Emerg Med 2011 Jun;40(6):623-628. [doi: 10.1016/j.jemermed.2008.04.003] [Medline: 18930373]
28. Törnerud M, Brandberg H. Akut Hjärtsjukvård (Acute Cardiac Care). Kunskapsstod for vardgivare.: Region Stockholm URL: https://kunskapsstodforvardgivare.se/omraden/hjart--och-karlsjukdomar/regionala-kunskapsstod/akut-hjartsjukvard [accessed 2026-01-10]
29. Zakim D. Development and significance of automated history-taking software for clinical medicine, clinical research and basic medical science. J Intern Med 2016 Sep;280(3):287-299 [FREE Full text] [doi: 10.1111/joim.12509] [Medline: 27071980]
30. Brandberg H, Sundberg CJ, Spaak J, Koch S, Kahan T. Are medical history data fit for risk stratification of patients with chest pain in emergency care? Comparing data collected from patients using computerized history taking with data documented by physicians in the electronic health record in the CLEOS-CPDS prospective cohort study. J Am Med Inform Assoc 2024 Jun 20;31(7):1529-1539 [FREE Full text] [doi: 10.1093/jamia/ocae110] [Medline: 38781350]
31. Than M, Flaws D, Sanders S, Doust J, Glasziou P, Kline J, et al. Development and validation of the Emergency Department Assessment of Chest pain Score and 2 h accelerated diagnostic protocol. Emerg Med Australas 2014 Feb;26(1):34-44. [doi: 10.1111/1742-6723.12164] [Medline: 24428678]
32. Body R, Carlton E, Sperrin M, Lewis PS, Burrows G, Carley S, et al. Troponin-only Manchester Acute Coronary Syndromes (T-MACS) decision aid: single biomarker re-derivation and external validation in three cohorts. Emerg Med J 2017 Jun;34(6):349-356 [FREE Full text] [doi: 10.1136/emermed-2016-205983] [Medline: 27565197]
33. Six AJ, Backus BE, Kelder JC. Chest pain in the emergency room: value of the HEART score. Neth Heart J 2008 Jun;16(6):191-196. [doi: 10.1007/bf03086144]
34. Diamond GA, Forrester JS. Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. N Engl J Med 1979 Jun 14;300(24):1350-1358. [doi: 10.1056/NEJM197906143002402] [Medline: 440357]
35. McCord J, Cabrera R, Lindahl B, Giannitsis E, Evans K, Nowak R, TRAPID-AMI Investigators. Prognostic utility of a modified HEART score in chest pain patients in the emergency department. Circ Cardiovasc Qual Outcomes 2017 Feb;10(2):1-8. [doi: 10.1161/CIRCOUTCOMES.116.003101] [Medline: 28167641]
36. Carlton EW, Ingram J, Taylor H, Glynn J, Kandiyali R, Campbell S, et al. Limit of detection of troponin discharge strategy versus usual care: randomised controlled trial. Heart 2020 Oct;106(20):1586-1594 [FREE Full text] [doi: 10.1136/heartjnl-2020-316692] [Medline: 32371401]
37. Backus BE, Six AJ, Kelder JC, Bosschaert MAR, Mast EG, Mosterd A, et al. A prospective validation of the HEART score for chest pain patients at the emergency department. Int J Cardiol 2013 Oct 3;168(3):2153-2158 [FREE Full text] [doi: 10.1016/j.ijcard.2013.01.255] [Medline: 23465250]
38. Reichlin T, Schindler C, Drexler B, Twerenbold R, Reiter M, Zellweger C, et al. One-hour rule-out and rule-in of acute myocardial infarction using high-sensitivity cardiac troponin T. Arch Intern Med 2012 Sep 10;172(16):1211-1218. [doi: 10.1001/archinternmed.2012.3698] [Medline: 22892889]
39. Body R, Mueller C, Giannitsis E, Christ M, Ordonez-Llanos J, de Filippi CR, TRAPID-AMI Investigators. The use of very low concentrations of high-sensitivity troponin T to rule out acute myocardial infarction using a single blood test. Acad Emerg Med 2016 Sep;23(9):1004-1013 [FREE Full text] [doi: 10.1111/acem.13012] [Medline: 27178492]
40. Giannitsis E, Kurz K, Hallermayer K, Jarausch J, Jaffe AS, Katus HA. Analytical validation of a high-sensitivity cardiac troponin T assay. Clin Chem 2010 Feb;56(2):254-261. [doi: 10.1373/clinchem.2009.132654] [Medline: 19959623]
41. Roffi M, Patrono C, Collet J, Mueller C, Valgimigli M, Andreotti F, Erol, et al. 2015 ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation: Task Force for the Management of Acute Coronary Syndromes in Patients Presenting without Persistent ST-Segment Elevation of the European Society of Cardiology (ESC). Eur Heart J 2016 Jan 14;37(3):267-315. [doi: 10.1093/eurheartj/ehv320] [Medline: 26320110]

XSL•FO
RenderX

42. Ibanez B, James S, Agewall S, Antunes MJ, Bucciarelli-Ducci C, Bueno H, ESC Scientific Document Group. 2017 ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation: The Task Force for the management of acute myocardial infarction in patients presenting with ST-segment elevation of the European Society of Cardiology (ESC). Eur Heart J 2018 Jan 07;39(2):119-177 [FREE Full text] [doi: 10.1093/eurheartj/ehx393] [Medline: 28886621]
43. Bezin J, Girodet P, Rambelomanana S, Touya M, Ferreira P, Gilleron V, et al. Choice of ICD-10 codes for the identification of acute coronary syndrome in the French hospitalization database. Fundam Clin Pharmacol 2015 Dec;29(6):586-591. [doi: 10.1111/fcp.12143] [Medline: 26301735]
44. International Statistical Classification of Diseases and Related Health Problems 10th Revision 2023. World Health Organization. URL: https://icd.who.int/browse10/2019/en [accessed 2025-03-01]
45. Ludvigsson JF, Andersson E, Ekbom A, Feychting M, Kim J, Reuterwall C, et al. External review and validation of the Swedish national inpatient register. BMC Public Health 2011 Jun 09;11:450 [FREE Full text] [doi: 10.1186/1471-2458-11-450] [Medline: 21658213]
46. Hammar N, Alfredsson L, Rosén M, Spetz CL, Kahan T, Ysberg AS. A national record linkage to study acute myocardial infarction incidence and case fatality in Sweden. Int J Epidemiol 2001 Oct;30 Suppl 1:S30-S34. [doi: 10.1093/ije/30.suppl_1.s30] [Medline: 11759848]
47. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983 Sep;148(3):839-843. [doi: 10.1148/radiology.148.3.6878708] [Medline: 6878708]
48. Than M, Herbert M, Flaws D, Cullen L, Hess E, Hollander JE, et al. What is an acceptable risk of major adverse cardiac event in chest pain patients soon after discharge from the Emergency Department?: a clinical survey. Int J Cardiol 2013 Jul 1;166(3):752-754. [doi: 10.1016/j.ijcard.2012.09.171] [Medline: 23084108]
49. Six AJ, Cullen L, Backus BE, Greenslade J, Parsonage W, Aldous S, et al. The HEART score for the assessment of patients with chest pain in the emergency department: a multinational validation study. Crit Pathw Cardiol 2013 Sep;12(3):121-126. [doi: 10.1097/HPC.0b013e31828b327e] [Medline: 23892941]
50. Boyle RSJ, Body R. The diagnostic accuracy of the Emergency Department Assessment of Chest Pain (EDACS) score: a systematic review and meta-analysis. Ann Emerg Med 2021 Apr;77(4):433-441. [doi: 10.1016/j.annemergmed.2020.10.020] [Medline: 33461885]
51. Liu N, Ng JCJ, Ting CE, Sakamoto JT, Ho AFW, Koh ZX, et al. Clinical scores for risk stratification of chest pain patients in the emergency department: an updated systematic review. J Emerg Crit Care Med 2018 Feb 7;2:16-16. [doi: 10.21037/jeccm.2018.01.10]
52. Nilsson T, Johannesson E, Lundager Forberg J, Mokhtari A, Ekelund U. Diagnostic accuracy of the HEART Pathway and EDACS-ADP when combined with a 0-hour/1-hour hs-cTnT protocol for assessment of acute chest pain patients. Emerg Med J 2021 Nov;38(11):808-813. [doi: 10.1136/emermed-2020-210833] [Medline: 33837120]
53. Cortés M, Haseeb S, Lambardi F, Arbucci R, Ariznavarreta P, Resi S, et al. The HEART score in the era of the European Society of Cardiology 0/1-hour algorithm. Eur Heart J Acute Cardiovasc Care 2020 Feb;9(1):30-38. [doi: 10.1177/2048872619883619] [Medline: 31657616]
54. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol 2010 Sep;5(9):1315-1316 [FREE Full text] [doi: 10.1097/JTO.0b013e3181ec173d] [Medline: 20736804]
55. Morrow DA. Clinician's guide to early rule-out strategies with high-sensitivity cardiac troponin. Circulation 2017 Apr 25;135(17):1612-1616. [doi: 10.1161/CIRCULATIONAHA.117.026717] [Medline: 28438803]
56. Hyams JM, Streitz MJ, Oliver JJ, Wood RM, Maksimenko YM, Long B, et al. Impact of the HEART pathway on admission rates for emergency department patients with chest pain: an external clinical validation study. J Emerg Med 2018 Apr;54(4):549-557. [doi: 10.1016/j.jemermed.2017.12.038] [Medline: 29478861]
57. Alley W, Mahler SA. Clinical decision aids for chest pain in the emergency department: identifying low-risk patients. Open Access Emerg Med 2015;7:85-92 [FREE Full text] [doi: 10.2147/OAEM.S71282] [Medline: 27147894]
58. Poldervaart JM, Reitsma JB, Backus BE, Koffijberg H, Veldkamp RF, Ten Haaf ME, et al. Effect of using the HEART score in patients with chest pain in the emergency department: a stepped-wedge, cluster randomized trial. Ann Intern Med 2017 May 16;166(10):689-697. [doi: 10.7326/M16-1600] [Medline: 28437795]

Abbreviations

ACS: acute coronary syndrome
AUC: area under the receiver operating characteristic curve
CHT: computerized history taking
CLEOS-CPDS: Clinical Expert Operating System Chest Pain Danderyd Study
D-HEART: Danderyd HEART
ECG: electrocardiogram
ED: emergency department

XSL•FO
RenderX

EDACS-ADP: Emergency Department Assessment of Chest Pain Score combined with an Accelerated Diagnostic Protocol
EHR: electronic health record
ESC: European Society of Cardiology
HEART: History, ECG, Age, Risk Factors, and Troponin
hs-cTn: high-sensitivity cardiac troponin
ICD-10: International Statistical Classification of Diseases and Related Health Problems, 10th Revision
MACE: major adverse cardiac event
NPV: negative predictive value
PPV: positive predictive value
RETTS: Rapid Emergency Triage and Treatment System
STARD: Standards for Reporting Diagnostic Accuracy
T-MACS: Troponin-only Manchester Acute Coronary Syndrome

Edited by A Coristine; submitted 16.Apr.2025; peer-reviewed by ME Matsumura, S van den Bulk; comments to author 17.Jul.2025; revised version received 08.Jan.2026; accepted 14.Jan.2026; published 11.Feb.2026.

Please cite as:
Brandberg H, Sundberg CJ, Spaak J, Koch S, Kahan T
Computerized Self-Reported Medical History Taking to Support Early Rule Out of Major Adverse Cardiac Events in Patients With Acute Chest Pain: Post Hoc Analysis of the CLEOS-CPDS Prospective Cohort Study
J Med Internet Res 2026;28:e76087
URL: https://www.jmir.org/2026/1/e76087
doi:10.2196/76087
PMID:

©Helge Brandberg, Carl Johan Sundberg, Jonas Spaak, Sabine Koch, Thomas Kahan. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 11.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Quality of Conventional versus Artificial Intelligence Oral Surgery Consent Forms: Comparative Analysis

Jan Gaessler[1], MD, DDS; Bernhard Remschmidt[1], MD, DMD; Ann-Kathrin Jopp[2], MSc, PhD; Behrouz Arefnia[3], DMD; Adrian Franke[4], MD, DMD; Marcus Rieder[1], MD, DMD

[1]Division of Oral and Maxillofacial Surgery, Department of Dental Medicine and Oral Health, Medical University of Graz, Auenbruggerplatz 5/6, Graz, Austria

[2]Clinical psychologist and psychotherapist in private practice, Hannover, Germany

[3]Division of Restorative Dentistry, Periodontology and Prosthodontics, Department of Dental Medicine and Oral Health, Medical University of Graz, Graz, Austria

[4]Department of Oral and Maxillofacial Surgery, University Hospital Carl Gustav Carus, Dresden University of Technology, Dresden, Germany

Corresponding Author:
Marcus Rieder, MD, DMD
Division of Oral and Maxillofacial Surgery, Department of Dental Medicine and Oral Health, Medical University of Graz, Auenbruggerplatz 5/6, Graz, Austria

Abstract

Artificial intelligence–generated informed consent forms for oral surgery demonstrated higher quality and better readability than conventional web-based forms, though both fell short of recommended comprehension levels.

(J Med Internet Res 2026;28:e59851) doi:10.2196/59851

KEYWORDS

oral surgical procedures; informed consent; quality control; artificial intelligence; oral surgery; consent form; AI; dental health; oral surgeon; patient care; paitent autonomy; dentistry

Introduction

Informed consent is a foundational element of ethical and legal medical care, ensuring patients understand the nature, risks, and alternatives of proposed treatments [1,2]. In oral surgery, where procedures can be complex and invasive, clear and high-quality informed consent forms (ICFs) are especially critical. However, many ICFs exceed the recommended 6th-grade reading level, limiting patient comprehension [3]. With the recent rise of artificial intelligence (AI), particularly large language models (LLMs), there is growing interest in their potential to improve patient communication [4,5]. This study aimed to assess the quality and readability of conventional, web-based oral surgery ICFs and compare them to those generated by AI-based LLMs.

Methods

Ten common oral surgery procedures were selected (ie, apicoectomy, biopsy, bone augmentation, cystectomy, dental implants, incision and drainage, local anesthesia, periodontal surgery, tooth extraction, and wisdom tooth removal). Using Google Chrome in incognito mode, 300 web-based ICFs (ie, 30 per procedure) were collected (see search strategy in Multimedia Appendix 1). In parallel, four LLMs (ChatGPT 3.5, Claude, Bard, and Bing Chat) were prompted to generate ICFs for the same procedures using standardized requests. Per every procedure and LLM, two basic and non-directive prompts were developed to minimize bias and ensure neutrality, resulting in 80 AI-generated ICFs (see Multimedia Appendix 1). Subsequently, two oral and maxillofacial surgeons screened the collected forms using predefined inclusion and exclusion criteria (see Multimedia Appendix 1).

Quality was assessed using a newly developed alteration of the well-established DISCERN instrument [6], namely the Graz Assessment Tool for Written Informed Consent Keypoints (GATWICK; see Multimedia Appendix 1). It was validated through expert review for content relevance and consistency. It includes 11 items scored on a 5-point Likert scale (total score range 11 - 55). Two oral and maxillofacial surgery residents independently rated all forms. Readability was evaluated using six established formulas (ie, Automated Readability Index, Coleman-Liau, Flesch-Kincaid, FORCAST, Gunning Fog, and Simple Measure of Gobbledygook), and an average reading grade level was calculated [7]. Statistical analyses included the Mann-Whitney *U* test, Kruskal-Wallis test, and Kendall tau-b, with significance set at $P \leq .05$.

Results

Of 380 screened documents, 213 ICFs met the inclusion criteria: 136 web-based and 77 AI-generated ones. The inter-rater reliability for GATWICK scores was excellent (intraclass correlation coefficient=0.948).

XSL•FO
RenderX

Regarding the quality, AI-generated ICFs had significantly higher total GATWICK scores compared to web-based ones (median 32.5, IQR 28-35.5 vs median 27.5, IQR 20.375-37; *P*=.007). Items related to treatment alternatives, rationale for recommended intervention, and discussion of options scored particularly higher in AI-generated forms. Web-based ICFs scored better in perioperative behavior instructions.

Considering the readability, web-based forms were significantly harder to read (median grade level 12.45, IQR 11.3-13.325) than AI-generated forms (median 10.7, IQR 10.1-12.4; *P*<.001), although neither met the recommended 6th-grade level. Readability was weakly correlated with overall quality (τ=0.132; *P*=.005).

The word count was higher for web-based forms (median 794 words, IQR 475.25-1068.75 words) than AI-generated ones (median 338 words, IQR 296-381 words; *P*<.001). Longer forms showed a weak correlation with higher quality (τ=0.270; *P*<.001).

Among LLMs, ChatGPT-powered services (ie, ChatGPT 3.5 and Claude) scored significantly higher in terms of quality. ICFs on tooth extraction scored significantly worse when compared with periodontal surgery forms. AI-generated informed consent forms performed significantly better than conventional versions, with notable differences across oral surgical procedures and among the types of LLMs used (Table 1).

Table . Quality of informed consent forms (ICFs) measured through the GATWICK (Graz Assessment Tool for Written Informed Consent Keypoints) score.

Quality	Median (IQR)	*P* value
Overall quality		.007[a]
Conventional (i.e., web-based) ICFs	27.50 (20.125-37)	
Artificial intelligence-generated ICFs	32.50 (28-36.25)	
All combined	31.00 (23-37)	
Differences by procedure		.004[b]
Apicoectomy	27.00 (21.75-34.875)	
Biopsy	30.50 (25.75-33)	
Oral bone augmentation	31.50 (25.75-37.5)	
Dental cystectomy	31.25 (23-33.875)	
Dental implants	33.25 (20.625-37.125)	
Oral incision and drainage	31.50 (23.5-39.5)	
Dental local anesthesia	28.50 (21-34.5)	
Periodontal surgery	36.50 (32.5-42)	
Tooth extraction	23.50 (20-32.75)	
Wisdom tooth removal	28.25 (20-36.875)	
Differences by large language model		<.001[b]
ChatGPT	34.25 (33-37)	
Claude	40.50 (35-43)	
Bing Chat	30.00 (27.25-31.75)	
Google Bard	26.50 (22.75-31.375)	

[a]Mann-Whitney *U* test.

[b]Kruskal-Wallis test.

Discussion

Principal Findings

This study found that conventional oral surgery ICFs available online are generally of modest quality and exceed recommended reading levels. AI-generated ICFs outperformed web-based ones in both quality and readability, although they too fell short of ideal readability standards.

These findings are consistent with prior research across medical disciplines, which show that most ICFs are written at a level too advanced for the average patient [8,9]. Notably, AI-generated forms more consistently addressed key informed consent components such as treatment alternatives and rationale, suggesting that LLMs may serve as valuable tools in drafting patient-centered documents. However, AI models may also produce inaccuracies or omit procedure-specific nuances, highlighting the need for expert review [10].

The limitations of this study include its focus on English-language materials and the variability inherent in AI outputs depending on prompt phrasing or model version. While

XSL•FO
RenderX

the GATWICK tool demonstrated strong reliability, further validation is needed.

Conclusion

AI-based LLMs offer a promising avenue for improving the quality and accessibility of oral surgery informed consent documents. Future efforts should focus on refining AI outputs and integrating clinician oversight to ensure accuracy, comprehensiveness, and patient comprehension.

Acknowledgments

Preliminary results were presented at the 27th Congress of the European Association for Cranio-Maxillo-Facial Surgery (EACMFS) in Rome, Italy, from September 17 to 20, 2024.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Authors' Contributions

JG: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft preparation, Writing – review and editing, Visualization, Project administration. BR: Conceptualization, Methodology, Validation, Investigation, Resources, Writing – original draft preparation, Writing – review and editing. A-KJ: Conceptualization, Methodology, Validation, Investigation, Resources, Writing – original draft preparation, Writing – review and editing. BA: Validation, Resources, Writing – review and editing, Supervision. AF: Validation, Resources, Writing – review and editing, Supervision. MR: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft preparation, Writing – review and editing, Supervision, Project administration. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Methodology showing the utilization of the Boolean operator “OR” helped to broaden the web search as it accounted for differences regarding designation and spelling (i.e., American versus British English). Detailed description of the Graz Assessment Tool of Written Informed Consent Keypoints (GATWICK).

[DOCX File, 29 KB - jmir_v28i1e59851_app1.docx]

References

1. Agozzino E, Borrelli S, Cancellieri M, Carfora FM, Di Lorenzo T, Attena F. Does written informed consent adequately inform surgical patients? A cross sectional study. BMC Med Ethics 2019 Jan 7;20(1):1. [doi: 10.1186/s12910-018-0340-z] [Medline: 30616673]
2. General Medical Council. Consent: patients and doctors making decisions together. URL: https://www.gmc-uk.org/-/media/documents/gmc-guidance-for-doctors---consent---english_pdf-48903482.pdf [accessed 2024-04-23]
3. Powers BJ, Trinh JV, Bosworth HB. Can this patient read and understand written health information? JAMA 2010 Jul 7;304(1):76-84. [doi: 10.1001/jama.2010.896] [Medline: 20606152]
4. Rasteau S, Ernenwein D, Savoldelli C, Bouletreau P. Artificial intelligence for oral and maxillo-facial surgery: A narrative review. J Stomatol Oral Maxillofac Surg 2022 Jun;123(3):276-282. [doi: 10.1016/j.jormas.2022.01.010] [Medline: 35091121]
5. Puladi B, Gsaxner C, Kleesiek J, Hölzle F, Röhrig R, Egger J. The impact and opportunities of large language models like ChatGPT in oral and maxillofacial surgery: a narrative review. Int J Oral Maxillofac Surg 2024 Jan;53(1):78-88. [doi: 10.1016/j.ijom.2023.09.005] [Medline: 37798200]
6. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999 Feb;53(2):105-111. [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
7. Ley P, Florio T. The use of readability formulas in health care. Psychol Health Med 1996 Feb;1(1):7-28. [doi: 10.1080/13548509608400003]

XSL•FO
RenderX

8. Meade MJ, Dreyer CW. Orthodontic treatment consent forms: a readability analysis. J Orthod 2022 Mar;49(1):32-38. [doi: 10.1177/14653125211033301] [Medline: 34325567]
9. Meade MJ, Dreyer CW. How readable are orthognathic surgery consent forms? Int Orthod 2022 Dec;20(4):100689. [doi: 10.1016/j.ortho.2022.100689] [Medline: 36117084]
10. Decker H, Trang K, Ramirez J, et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. JAMA Netw Open 2023 Oct 2;6(10):e2336997. [doi: 10.1001/jamanetworkopen.2023.36997] [Medline: 37812419]

Abbreviations

AI: Artificial intelligence
GATWICK: Graz Assessment Tool for Written Informed Consent Keypoints
ICF: informed consent form
LLM: large language models

Edited by N Cahill; submitted 28.Apr.2024; peer-reviewed by H Zhang, R Pranab, W Banjar; revised version received 01.Dec.2025; accepted 01.Dec.2025; published 05.Jan.2026.

Please cite as:
Gaessler J, Remschmidt B, Jopp AK, Arefnia B, Franke A, Rieder M
Quality of Conventional versus Artificial Intelligence Oral Surgery Consent Forms: Comparative Analysis
J Med Internet Res 2026;28:e59851
URL: https://www.jmir.org/2026/1/e59851
doi:10.2196/59851

©Jan Gaessler, Bernhard Remschmidt, Ann-Kathrin Jopp, Behrouz Arefnia, Adrian Franke, Marcus Rieder. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

National Institutes of Health–Funded Artificial Intelligence and Machine Learning Research, 2019 2023: Cross-Sectional Study

Joshua Pei Le[1*], MD; Joseph Morrison[2*], BS; Atul Malhotra[3], MD; Shamim Nemati[3,4], PhD; Gabriel Wardi[3,4], MD, MPH; James S Ford[4*], MAS, MD

[1]School of Medicine, University of Limerick, Limerick, Ireland

[2]School of Medicine, University of California Davis, Sacramento, CA, United States

[3]Department of Medicine, University of California San Diego, La Jolla, CA, United States

[4]Department of Emergency Medicine, University of California San Diego, 9500 Gilman Drive, La Jolla, CA, United States

[*]these authors contributed equally

Corresponding Author:
James S Ford, MAS, MD
Department of Emergency Medicine, University of California San Diego, 9500 Gilman Drive, La Jolla, CA, United States

Abstract

Inflation-adjusted funding for artificial intelligence and machine learning research increased by 233% between fiscal year 2019 and 2023, outpacing the overall National Institutes of Health's budget increase of 12%.

(J Med Internet Res 2026;28:e84861) doi:10.2196/84861

KEYWORDS

artificial intelligence; AI; machine learning; ML; National Institutes of Health; NIH; research funding

Introduction

Artificial intelligence (AI) and machine learning (ML) technologies have begun to revolutionize the practice of medicine. Recent studies have shown that 2 in 3 physicians currently use AI in their clinical practice and more than 3 in 4 scientists currently use AI in their research [1]. Recognizing the growing importance of AI tools, the National Institutes of Health (NIH) began tracking AI and ML as its own funding category in fiscal year (FY) 2019. While previous studies have broadly assessed NIH funding trends for AI and ML research, no studies report data more recently than 2020 [2,3]. Moreover, little is known about the principal investigators (PIs) driving this research. In this study, we aimed to examine trends in NIH-funded AI and ML research and characterize the population of funded PIs.

Methods

We conducted a cross-sectional study of NIH RePORTER (Research Portfolio Online Reporting Tools Expenditures and Results) data between October 1, 2018, and September 30, 2023. We extracted studies indexed under the "Machine Learning and Artificial Intelligence" funding category and collected data on funding institute, mechanism, and award amount (in US $). We also selected a random sample of approximately 25% of PIs who were awarded grants, and we conducted an internet search to collect data on educational background (eg, MD), research setting (eg, academic), and clinical specialty (as appropriate). We searched official websites (eg, university), LinkedIn, or publicly available curriculum vitae. To account for potential differences in educational background (eg, PhD) and clinical specialty (eg, internal medicine) among PIs funded by different institutes, we balanced our random sampling according to the original institute-level proportions of funded grants. Three authors (JM, JPL, and JSF) conducted the search. CIs for proportions were calculated using the Wald method. Tests of trend were conducted using Poisson and linear regression, as appropriate. Analyses were performed using Stata/SE (version 17.0; StataCorp LLC).

Results

Total active projects (including multiyear grants) increased from 1229 to 3449 and total funding increased nominally from US $0.6 billion to $2.3 billion between NIH FY 2019 and 2023 (Figure 1).

There were 5418 unique projects from 4365 unique researchers. Funding trends according to NIH institute and grant mechanism are available in Figures S1 and S2 in Multimedia Appendix 1. We randomly selected approximately 25% of PIs (1091/4395) to conduct further data extraction via internet searches. Among this random sample, the most common educational background was PhD only (70%, 758/1091), followed by MD or DO only (11%, 115/1091), and MD or PhD (9%, 93/1091) (Table S1 in Multimedia Appendix 1). Individuals with only a PhD accounted for 64% of funding dollars, compared to 18% of those with a medical doctoral degree (eg, MD or DO) but no PhD, and 11%

XSL•FO
RenderX

for those with a clinical or professional degree (RN, MD, or PharmD) and a PhD. There were 251/1091 (23%) PIs who held medical doctoral degrees (MD or DO), and of these, 229 (91%, 95% CI 87%-94%) listed residency training. The proportions of total projects and funding dollars by medical specialty are available in Figure 2.

Figure 1. Overall National Institutes of Health (NIH)–funded artificial intelligence and machine learning research projects and funding (in US $) between NIH fiscal year 2019 and 2023. Plotted funding (US $) represents nominal values unadjusted for inflation. Inflation-adjusted values are presented in Table S2A and S2B in Multimedia Appendix 1.

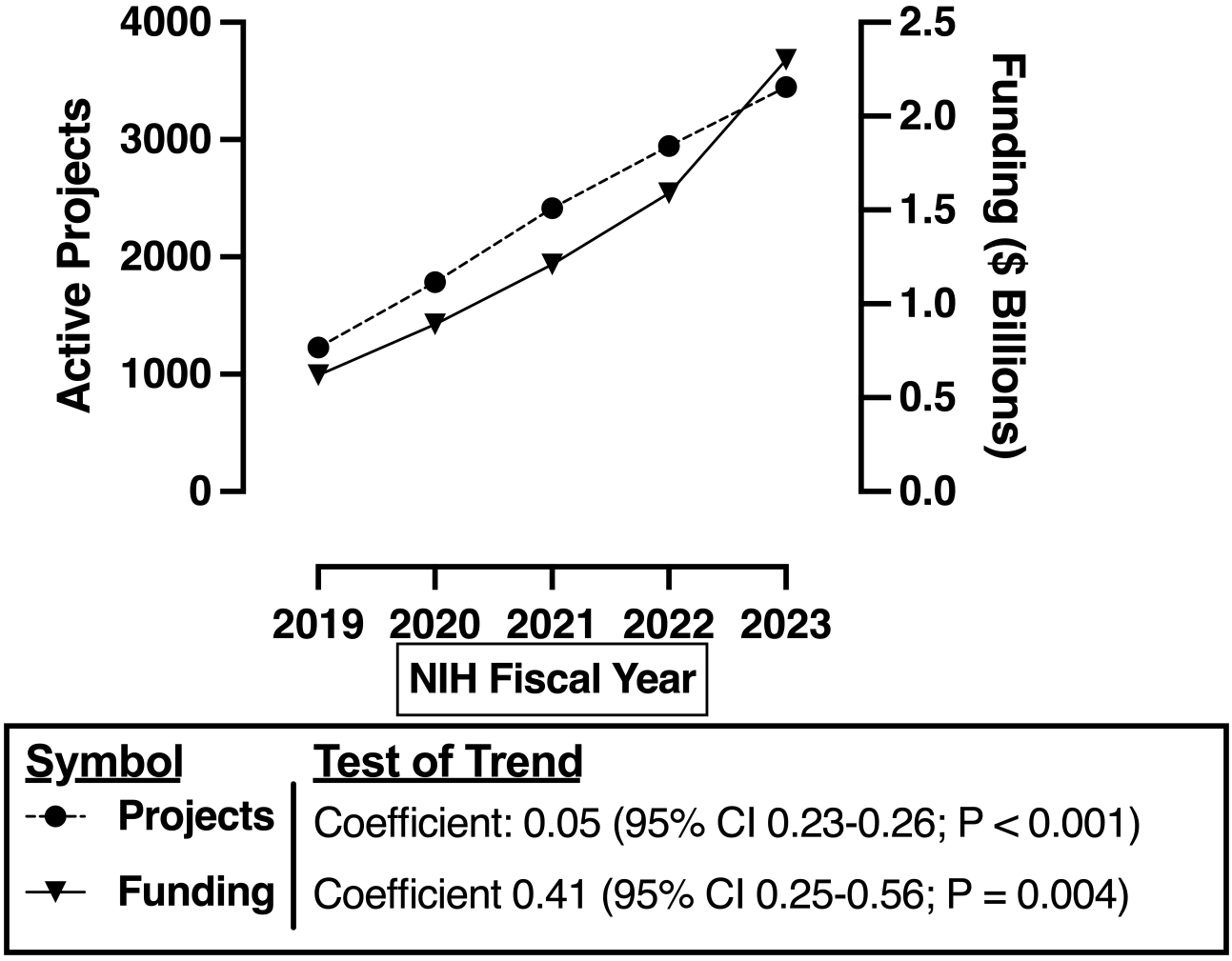

XSL•FO
RenderX

Figure 2. (A) Proportions of total funded principal investigators (PIs) and (B) total funding by specialty. Data from random sample of 1091 unique PIs with National Institutes of Health–funded projects in artificial intelligence or machine learning. CC: critical care; Gen: general; GI: gastroenterology; ID: infectious diseases; IM: internal medicine; Med: medical; Peds: pediatrics; SS, subspecialty; Surg: surgery. *Includes general surgery, urology, vascular surgery, transplant surgery, head and neck surgery, pediatric surgery, neurosurgery, obstetrics and gynecology. †Includes general pediatrics, cardiology, infectious diseases, hematology/oncology, pulmonology, endocrinology, neonatology. ‡Includes endocrinology, rheumatology, nephrology, hematology/oncology. §Includes emergency medicine, anesthesiology, pediatric emergency medicine, dermatology, and PIs with more than one specialty.

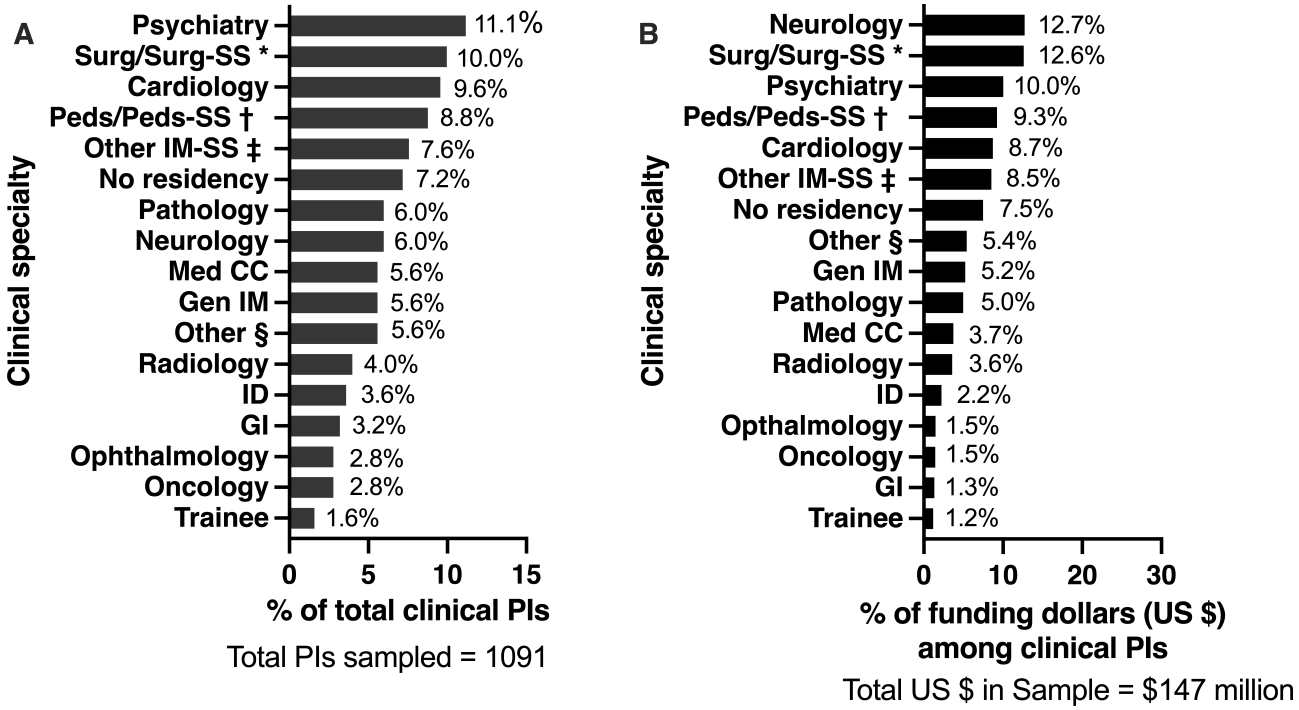

Discussion

In 2017, which was before the NIH began tracking AI and ML as its own spending category, one study found that the NIH supported 535 projects totaling US $264 million, representing approximately 0.7% of the total NIH budget (NIH budget FY 2017; $34.301 billion) [2,4]. Our analysis of 2023 NIH RePORTER funding revealed $2.3 billion in funding (including direct, indirect, and supplemental costs) in this spending category, representing approximately 4.7% of the total NIH budget (NIH budget FY 2023; $49.2 billion) [4]. While the NIH budget increased by 12% (adjusted for inflation) between 2019 and 2023, funding for AI and ML research increased by 233% (adjusted for inflation) over this same period (Table S2A and S2B in Multimedia Appendix 1) [4].

Most AI and ML researchers held only a PhD (70%), and this group accounted for almost two-thirds of all NIH research funding dollars in the AI and ML funding category. By contrast, researchers who held a medical doctoral degree (MD or DO) and who had completed a clinical residency program were relatively underrepresented, accounting for less than 20% of funded researchers and total research dollars. In our sample of researchers who held a medical doctoral degree (MD or DO), those with training in psychiatry had the most funded projects and accounted for the third-highest proportion of funding dollars. While the list of possible applications is extensive and growing, notably, funding among specialties such as emergency medicine, anesthesiology, dermatology, ophthalmology, and various individual medical or surgical subspecialties was considerably lower. These gaps are similar to those seen across other research disciplines and represent clear targets for future NIH strategic investment [5,6].

As US federal funding priorities have been modified by the current administration and the NIH faces anticipated contraction in federal discretionary spending for FY 2026, our findings come at a critical moment [4,7]. Indeed, a recent study demonstrated the potential “ripple effects” of NIH funding cuts, including delays in innovation, workforce attrition, loss of talent to universities and corporations overseas, and increased long-term health care costs [8]. These proposed changes are likely to substantially weaken the United States’ contributions to this critical emerging field.

Our analysis included PIs, but data on coinvestigators were not captured. Data related to PI education and training background were obtained from publicly available online sources, which may have been incomplete or outdated at the time of collection. PIs with multiple specialties may have been misclassified. Our spending estimates are higher than those published by the NIH because we aggregated all funding (eg, direct costs, indirect costs, subprojects, and supplemental costs); we believe this methodology provides a more complete view of total research funding [9]. While our random sample of PIs was balanced according to institute-specific proportions, we cannot exclude the possibility of sampling bias. Lastly, the study period overlaps with the COVID-19 pandemic, which may confound funding trends during this time.

In summary, we found that total NIH projects and funding dollars increased substantially between 2019 and 2023. As the US government has enacted substantial NIH budget cuts for FY

XSL•FO
RenderX

2026, future studies are needed to examine the impact of current fiscal policy on future NIH-funded AI and ML research.

Acknowledgments

No generative artificial intelligence was used in the preparation of this manuscript.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

Data will be provided upon reasonable request of the corresponding author.

Authors' Contributions

Conceptualization: AM, SN, GW, JSF
Data Curation: JPL, JM, JSF
Formal analysis: JSF
Investigation: JPL, JM, AM, SN, GW, JSF
Methodology: JSF
Supervision: AM, SN, GW, JSF
Visualization: JSF
Writing – original draft: JPL, JM, JSF
Writing – review & editing: JPL, JM, AM, SN, GW, JSF

Conflicts of Interest

SN and AM are cofounders of a University of California San Diego (UCSD) start-up, Clairyon Inc, which is focused on the commercialization of advanced analytical decision support tools and was formed in compliance with UCSD conflict of interest policies. SN, AM, GW, and JSF receive funding from the National Institutes of Health. SN and JSF receive funding from the University of California Office of the President. GW reports consulting for Abbott Laboratories. GW and JSF report funding from the National Foundation of Emergency Medicine.

Multimedia Appendix 1
Supplemental Tables S1 and S2 and Figures S1 and S2.
[DOCX File, 322 KB - jmir_v28i1e84861_app1.docx]

References

1. Researchers and AI: survey findings. Oxford University Press. 2024 May. URL: https://fdslive.oup.com/www.oup.com/academic/pdf/Researchers-and-AI-survey-findings.pdf [accessed 2025-06-13]
2. Annapureddy AR, Angraal S, Caraballo C, et al. The National Institutes of Health funding for clinical research applying machine learning techniques in 2017. NPJ Digit Med 2020;3(13):13. [doi: 10.1038/s41746-020-0223-9] [Medline: 32025574]
3. Eweje FR, Byun S, Chandra R, et al. Translatability analysis of National Institutes of Health-funded biomedical research that applies artificial intelligence. JAMA Netw Open 2022 Jan 4;5(1):e2144742. [doi: 10.1001/jamanetworkopen.2021.44742] [Medline: 35072720]
4. National Institutes of Health (NIH) funding: FY1996-FY2025 Congressional Research Service. congress.gov. 2024 Jun 25. URL: https://www.congress.gov/crs_external_products/R/PDF/R43341/R43341.54.pdf [accessed 2025-06-13]
5. Morrison J, Le JP, Malhotra A, Nemati S, Wardi G, Ford JS. A snapshot of National Institutes of Health funding for sepsis research: 2019-2023. Ann Emerg Med 2025 Aug;86(2):206-208. [doi: 10.1016/j.annemergmed.2025.03.011] [Medline: 40237685]
6. Schlafly A, Sebro R. Does NIH funding differ between medical specialties? A longitudinal analysis of NIH grant data by specialty and type of grant, 2011-2020. BMJ Open 2022 Dec 30;12(12):e058191. [doi: 10.1136/bmjopen-2021-058191] [Medline: 36585146]
7. NIH Office of Budget. URL: https://officeofbudget.od.nih.gov [accessed 2025-10-28]
8. Jalali MS, Hasgul Z. Potential trade-offs of proposed cuts to the US National Institutes of Health. JAMA Health Forum 2025 Jul 3;6(7):e252228. [doi: 10.1001/jamahealthforum.2025.2228] [Medline: 40711777]
9. Estimates of funding for various research, condition, and disease categories (RCDC). Research Portfolio Online Reporting Tools (RePORT). URL: https://report.nih.gov/funding/categorical-spending [accessed 2025-06-09]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
FY: fiscal year
ML: machine learning
NIH: National Institutes of Health
PI: principal investigator
RePORTER: Research Portfolio Online Reporting Tools Expenditures and Results

Edited by A Coristine; submitted 25.Sep.2025; peer-reviewed by DJ Dzikowicz, KM Gillen; revised version received 24.Nov.2025; accepted 24.Nov.2025; published 08.Jan.2026.

Please cite as:
Le JP, Morrison J, Malhotra A, Nemati S, Wardi G, Ford JS
National Institutes of Health–Funded Artificial Intelligence and Machine Learning Research, 2019 - 2023: Cross-Sectional Study
J Med Internet Res 2026;28:e84861
URL: https://www.jmir.org/2026/1/e84861
doi:10.2196/84861

© Joshua Pei Le, Joseph Morrison, Atul Malhotra, Shamim Nemati, Gabriel Wardi, James S Ford. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Transformer-Based Topic Modeling: Characterizing Cannabis Product Adverse Experiences Self-Reported as Requiring Medical Attention on Reddit

Tim Ken Mackey[1,2,3,4], MAS, PhD; Matthew C Nali[2,3,4], BA; Meng Zhen Larsen[2,3], BS; Zhuoran Li[3], BS, MS; Cassandra L Taylor[5], PhD; Beverly Wolpert[6], MS, PhD; Catharine Trice[6,7], MA

[1]Global Health Program, Department of Anthropology, University of California, San Diego, La Jolla, CA, United States
[2]San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA, United States
[3]S-3 Research LLC, San Diego, CA, United States
[4]Global Health Policy and Data Institute, San Diego, CA, United States
[5]Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, United States
[6]Office of Surveillance Strategy and Risk Prioritization, Human Foods Program, U.S. Food and Drug Administration, College Park, MD, United States
[7]Office of Women's Health, U.S. Food and Drug Administration, Silver Spring, MD, United States

Corresponding Author:
Tim Ken Mackey, MAS, PhD
Global Health Program
Department of Anthropology
University of California, San Diego
9500 Gilman Dr
MC 0505
La Jolla, CA, 92093
United States
Phone: 1 9514914161
Email: tmackey@ucsd.edu

Abstract

This study uses keyword filtering, a transformer-based algorithm, and inductive content coding to identify and characterize cannabis adverse experiences as discussed on the social media platform Reddit and reports a total of 1177 self-reported adverse experiences requiring medical attention.

(J Med Internet Res 2026;28:e82661) doi:10.2196/82661

KEYWORDS

cannabis; adverse events; reddit; machine learning, topic modeling

Introduction

Cannabis is one of the most frequently used intoxicating substances in the United States [1]. Changing policies and attitudes have increased cannabis use, yet safety profiles for different cannabis-derived products (CDPs) (eg, products containing tetrahydrocannabinols, cannabidiol, or cannabinoids derived from cannabis) are not well understood. Studies have examined internet and social media data for health concerns among cannabis users, including for adverse health outcomes [2-8]. However, no study has specifically characterized the types of adverse experiences requiring medical attention, as self-reported online. We aimed to use a transformer-based algorithm to identify and characterize adverse experiences self-reported as requiring medical attention (SRRMA) on Reddit.

Methods

For this observational retrospective study, we collected user-generated posts and comments from 27 subreddits(/r/) related to cannabis use (Figure 1; Multimedia Appendix 1). A set of keywords for adverse experiences SRRMA was used for data mining and manual coding to generate a sample of “signal” Reddit data (ie, posts about cannabis product adverse experiences requiring medical attention: hospitalization, or visiting a health care professional, emergency room, or urgent

XSL•FO
RenderX

care). This initial sample of SRMMA prelabeled “signal” was joined with data from the full Reddit data corpus.

Figure 1. Summary of the study methodology, including data collection, data processing/filtering, BERT (bidirectional encoder representations from transformers) analysis, and results (signal data): (1) data collection for cannabis-related subreddits (both posts and comments); (2) keyword filtering for adverse experience self-reported as requiring medical attention (AE SRRMA)-related signals; (3) separation of sample of “signals” for general adverse experiences (GAE) self-reported as not requiring medical attention and AE SRRMA; (4) use of BERT topic model (k=5, 10, 15, 20) after data filtering with a sample of AE SRRMA keywords and labeled AE SRRMA data joined with the unlabeled dataset to analyze the full corpus of Reddit posts and comments; (5) selection of BERT topic clusters with the highest yield of prelabeled signals for AE SRRMA that were manually annotated to identify additional AE SRRMA signals (process repeated to achieve higher thematic saturation); newly identified signals were joined with the unlabeled dataset for the next round of BERT topic modeling; and (6) AE SRRMA signal identification from Reddit posts and comments summarized in this publication.

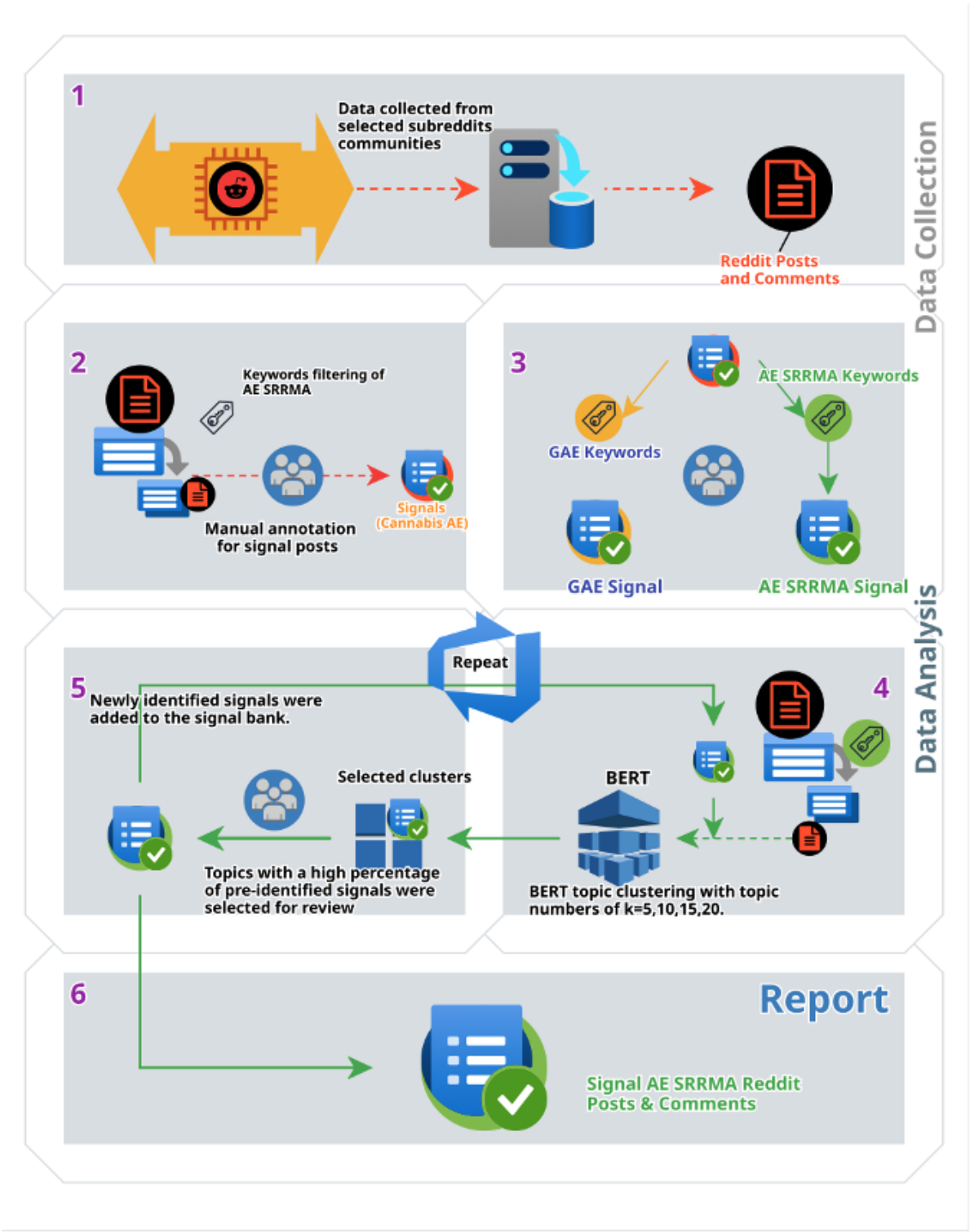

XSL•FO
RenderX

Natural language processing was applied to cluster Reddit posts and comments for further manual annotation. We used the bidirectional encoder representations from transformers (BERT) topic model with the Python (version 3.7) package BERTopic (version 0.6.0), which is a pretrained, self-supervised transformer-based algorithm that embeds texts, extracts topics, and then clusters texts.

BERT was used to cluster groups of nonlabeled data with SRRMA prelabeled signal data into k topic clusters, where the value of k is determined according to the dataset size ($5 \leq k \geq 20$). In each generated cluster, we calculated the percentages of SRRMA prelabeled signal data in that topic cluster to estimate how well that cluster was related to our SRRMA prelabeled data (% of potential SRRMA prelabeled signal data in a cluster that could be appended to the SRRMA prelabeled dataset = amount of SRRMA prelabeled training signal data in a cluster/amount of total data in a cluster). The higher the percentage, the more likely the uncoded data were related to our SRRMA prelabeled topic. Cluster selection was determined based on the highest percentages. The new signals were then added to our coded signal pool for the next iteration of BERT modeling: we filtered the uncoded dataset with the keywords repeatedly, combined the filtered uncoded data with all prelabeled signal data, and then performed the next iteration of BERT to find the best clusters to manually annotate for identifying additional signals. Posts and comments from these clusters were extracted, and an inductive coding approach was used to label and characterize cannabis-related adverse experiences and adverse experiences that met the SRRMA criteria. Human ethics and consent to participate were not applicable, as all information was from the public domain and did not involve user interaction, and any identifiable information was aggregated and removed from the results.

Results

We collected 1,795,478 Reddit posts/comments. After BERT and coding, 1177 posts/comments comprising 1542 user-generated mentions of SRRMA adverse experiences for cannabis products were detected between July 2017 and December 2022 (Table 1). Coders (TKM, MN, MZ) achieved a high intercoder reliability score (κ=0.90). From the 27 subreddits, the top 10 SRRMA adverse experiences were vomiting (284/1542, 18.42%), nausea (171/1542, 11.09%), panic attack (122/1542, 7.91%), abdominal/stomach pain (96/1542, 6.23%), concern over elevated heart rate (92/1542, 5.97%), anxiety (60/1542, 3.89%), chest pain (53/1542, 3.44%), general complaints of sickness (31/1542, 2.01%), symptoms attributed to Cannabinoid Hyperemesis Syndrome (30/1542, 1.95%), and paranoia (30/1542, 1.95%). Additionally, 108 (9.18%) SRRMAs mentioned a CDP but not a specific adverse experience. SRRMA adverse experiences included users self-reporting seeking medical attention with visits/admissions to hospitals (533/1177, 45.28%), emergency rooms (569/1177, 48.34%), health care professionals (70/1177, 5.95%), and urgent care (5/1177, 0.43%).

For cannabis product use characteristics with an SRRMA, most (988/1177, 83.94%) were inhalation (eg, inhaler, joint, vape) followed by ingestion (59/1177, 5.01%; eg, edibles) products. For users reporting intent of use (ie, post and comment), 94.65% (1114/1177) were adult use (ie, recreational), followed by 3.57% (42/1177) therapeutic (ie, for health benefit), 1.02% (12/1177) unknown, 0.42% (5/1177) medical (ie, from a medical dispensary), and 0.34% (4/1177) unintentional use. A subset of users reported co-use with other substances, specifically tobacco and nicotine products (n=7), prescription medications (eg, alprazolam, antidepressants, n=22), illicit drugs (eg, lysergic acid diethylamide, cocaine, n=5), and dual use with other cannabinoids (eg, general tetrahydrocannabinol use, cannabidiol, delta 8, n=12).

Table 1. Characteristics of adverse experiences self-reported as requiring medical attention, posted online on Reddit between July 2017 and December 2022.

Post and comment type/intent of use	Adverse Experiences Self-Reported as Requiring Medical Attention (AE SRRMA)[a]				
	Hospitalization (n=533), n	Health care professional visit (n=70), n	Emergency room visit (n=569), n	Urgent care (n=5), n	Proportion of platform AE SRRMA posts (N=1177), n (%)
Medical	1	0	4	0	5 (0.42)
Recreational	513	66	530	5	1114 (94.65)
Therapeutic	15	4	23	0	42 (3.57)
Unintended	0	0	4	0	4 (0.34)
Unknown	4	0	8	0	12 (1.02)

[a]AE SRRMA indicates that a cannabis user sought medical care because of a concern that their symptom or symptoms were serious or potentially life-threatening. Intent of use categories included medical (user mentions acquiring a product from a medical dispensary); recreational (user mentions using a cannabis product, price, or acquisition, in the context of a recreational or social event or circumstance); therapeutic (user mentions using a cannabis product for medical reasons); unknown (could not be determined based on data that was available); and unintended (user mentions using a cannabis-derived product accidentally or unintentionally).

XSL•FO
RenderX

Discussion

We identified 1177 Reddit user-generated posts that described CDP SRMMA adverse experiences. Results are similar to those of a study analyzing 28,630 cannabis exposures/cases reported to state poison control centers from 2017 to 2019 that found CDP concentrates, vaporized liquids, and edibles documented with “moderate or greater” medical outcomes (ie, effects usually requiring a form of medical treatment) [9]. Studies using transformer-based and large language models also found similar reports of adverse experiences by Reddit users but only coded for generalized adverse health outcomes [7] or only reviewed a single subreddit (r/delta8) [8,9]. Our study builds on the methodology of these prior studies by using a transformer-based algorithm “seeded” with prelabeled signal data to find additional unlabeled data—an approach similar to a few-shot learning approach but without formal model training.

Study limitations include terms not inclusive of all cannabis safety-related experiences and lack of cross-validation of users’ self-reports (eg, lack of validation with clinical data). Further, social media data are not representative of all cannabis use behavior, and what cannabis products may have led to adverse experiences (eg, contaminated/adulterated products or products not from licensed dispensaries) were unclear (eg, we observed more Reddit discussions on adult-use cannabis [ie, recreational] than on medical cannabis use, which may limit generalizability). Though exploratory, study results may be used in conjunction with other cannabis-related adverse experience surveillance data (eg, National Poison Data System and FDA MedWatch [6]) and can provide important context to adverse experiences that result in medical attention for purposes of helping consumers and clinicians better understand potential cannabis risks.

Acknowledgments

Generative artificial intelligence was not used for any portion of this manuscript in its writing, editing, or generation.

Funding

This research was supported by the Food and Drug Administration (FDA) of the US Department of Health and Human Services (HHS) as part of a contractual award (contract 75F40123C00030) totaling US $174,971.73, with 100% funded by FDA/HHS. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by FDA/HHS or the US government.

Authors' Contributions

Conceptualization: TKM, CLT, BJW, CT
Methodology: TKM, MCN, MZL, ZL
Software: ZL
Validation: TKM, MCN, MZL
Formal analysis: TKM, MCN, MZL
Investigation: TKM, MCN, MZL
Data curation: MCN, MZL, ZL
Writing – original draft: TKM, MCN, MZL
Writing – review & editing: TKM, CLT, BJW, CT
Supervision: TKM
Project administration: TKM
Funding acquisition: TKM

Conflicts of Interest

TKM, MN, MZL, and ZL are employees of the startup company S-3 Research LLC. TKM is also a cofounder and co-owner of the company. S-3 Research is a minority owned small business that was originally created through a National Institute on Drug Abuse startup business award and has been subsequently funded through government contracts with federal agencies to develop research tools and services for data science in public health, including social listening, data mining, and machine learning approaches. TKM is Editor-in-Chief of *JMIR Infodemiology*, a JMIR Publications journal, at the time of this publication. The authors report no other conflict of interests associated with this manuscript.

Multimedia Appendix 1
Data collection, filtering, and content coding methodology.
[DOCX File , 51 KB - jmir_v28i1e82661_app1.docx]

References

XSL•FO
RenderX

1. U.S. Centers for Disease Control and Prevention. Cannabis facts and stats. Data and Statistics. 2025 Mar 07. URL: https://www.cdc.gov/cannabis/data-research/facts-stats/?CDC_AAref_Val=https://www.cdc.gov/marijuana/data-statistics.htm [accessed 2026-01-27]
2. Meacham MC, Paul MJ, Ramo DE. Understanding emerging forms of cannabis use through an online cannabis community: an analysis of relative post volume and subjective highness ratings. Drug Alcohol Depend 2018 Jul 01;188:364-369 [FREE Full text] [doi: 10.1016/j.drugalcdep.2018.03.041] [Medline: 29883950]
3. Allem J, Majmundar A, Dormanesh A, Donaldson SI. Identifying health-related discussions of cannabis use on Twitter by using a medical dictionary: content analysis of tweets. JMIR Form Res 2022 Feb 25;6(2):e35027 [FREE Full text] [doi: 10.2196/35027] [Medline: 35212637]
4. Hallinan CM, Khademi Habibabadi S, Conway M, Bonomo YA. Social media discourse and internet search queries on cannabis as a medicine: a systematic scoping review. PLoS One 2023;18(1):e0269143 [FREE Full text] [doi: 10.1371/journal.pone.0269143] [Medline: 36662832]
5. Dilley JA, Graves JM, Brooks-Russell A, Whitehill JM, Liebelt EL. Trends and characteristics of manufactured cannabis product and cannabis plant product exposures reported to US poison control centers, 2017-2019. JAMA Netw Open 2021 May 03;4(5):e2110925 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.10925] [Medline: 34028553]
6. Smith BP, Hoots B, DePadilla L, Roehler DR, Holland KM, Bowen DA, et al. Using transformer-based topic modeling to examine discussions of delta-8 tetrahydrocannabinol: content analysis. J Med Internet Res 2023 Dec 21;25:e49469 [FREE Full text] [doi: 10.2196/49469] [Medline: 38127427]
7. Leas EC, Ayers JW, Desai N, Dredze M, Hogarth M, Smith DM. Using large language models to support content analysis: a case study of ChatGPT for adverse event detection. J Med Internet Res 2024 May 02;26:e52499 [FREE Full text] [doi: 10.2196/52499] [Medline: 38696245]
8. Hines MC, Harinstein LM, Kortepeter CM. Reporting adverse events for cannabis to the FDA. N Engl J Med 2020 Jan 02;382(1):98-98. [doi: 10.1056/nejmc1913460]
9. Leas EC, Harati RM, Satybaldiyeva N, Morales NE, Huffaker SL, Mejorado T, et al. Self-reported adverse events associated with Δ-Tetrahydrocannabinol (Delta-8-THC) Use. J Cannabis Res 2023 May 23;5(1):15 [FREE Full text] [doi: 10.1186/s42238-023-00191-y] [Medline: 37217977]

Abbreviations

BERT: bidirectional encoder representations from transformers
CDP: cannabis-derived product
SRRMA: self-reported as requiring medical attention

Edited by A Mavragani; submitted 19.Aug.2025; peer-reviewed by SN Sundaradhas, JM Conway; comments to author 03.Nov.2025; revised version received 09.Jan.2026; accepted 20.Jan.2026; published 04.Feb.2026.

Please cite as:
Mackey TK, Nali MC, Larsen MZ, Li Z, Taylor CL, Wolpert B, Trice C
Transformer-Based Topic Modeling: Characterizing Cannabis Product Adverse Experiences Self-Reported as Requiring Medical Attention on Reddit
J Med Internet Res 2026;28:e82661
URL: https://www.jmir.org/2026/1/e82661
doi:10.2196/82661
PMID:

©Tim Ken Mackey, Matthew C Nali, Meng Zhen Larsen, Zhuoran Li, Cassandra L Taylor, Beverly Wolpert, Catharine Trice. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Sex Differences in Medical Specialist Physicians' Electronic Health Record In-Basket Workloads and the Implications for Compensation and Equity: Retrospective Cross-Sectional Study

Ashwini Nadkarni[1], MD; Laura Glick[2], MD; Molly Becker[3], MS; Michael Healey[1], MD; David Stein[1], PhD; Mahyar Heydarpour[1], PhD; Lisa S Rotenstein[4], MD, MBA

[1]Brigham and Women's Hospital, Boston, MA, United States

[2]Massachusetts General Hospital, Boston, MA, United States

[3]Suffolk University, Boston, MA, United States

[4]University of California, San Francisco, San Francisco, CA, United States

Corresponding Author:
Ashwini Nadkarni, MD
Brigham and Women's Hospital
75 Francis Street
Boston, MA, 02115
United States
Phone: 1 6177329811
Email: anadkarni@bwh.harvard.edu

Abstract

In this retrospective cross-sectional study, we assessed sex differences in in-basket messages generated by outpatient workflows among internal medicine specialists as compared to relative value units; we found that female physicians had a greater burden of in-basket work for each unit of paid clinical care.

(J Med Internet Res 2026;28:e79172) doi:10.2196/79172

KEYWORDS

Electronic health records; In-basket messaging; Physician workload; Sex differences; Relative value units

Introduction

Female physicians in primary care receive higher volumes of patient and staff messages in the electronic health record (EHR) in-basket [1,2]. However, little is known on whether sex differences extend into medical specialties and impact on workload disparities relative to compensated care. In this cross-sectional study, we assessed sex differences in in-basket messages generated by outpatient workflows among internal medicine specialists as compared to relative value units (RVUs). We aimed to clarify whether female medical specialty physicians experience disproportionately greater uncompensated electronic work for each unit of paid clinical care.

Methods

Ethical Considerations

We followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) cross-sectional study guidelines. The Mass General Brigham Institutional Review Board deemed the study exempt (2021P001356). The informed consent requirement was waived because the study involved secondary analysis of deidentified EHR data and posed minimal risk to participants. Participants did not receive compensation. Although prior literature refers to "gender" differences, our analysis examined sex differences, so we used the terms "female" and "male." We conducted the study at a large, academic medical center. All attending physicians actively practiced in outpatient specialty practices with measurable in-basket activity during the study period.

Study Procedures

We collected monthly, ambulatory in-basket physician data and patient characteristics for each attending physician's patient panel from Epic for January through March 2021 and normalized monthly in-basket burden measures by monthly RVUs. Patients are generally assigned to physicians based on availability and patient preference. We obtained physician sex, academic rank, and practice years from the public institutional directory, which reports sex as male or female; no physicians were listed as nonbinary. We categorized specialties as those including a

XSL•FO
RenderX

substantial procedural component (cardiology, gastroenterology, and pulmonology) versus those predominantly not procedural (genetics, geriatrics, hematology, immunology, infectious disease, nephrology, palliative care, rheumatology, and sleep medicine).

Statistical Analysis

We compared continuous variables using Wilcoxon rank sum tests, and categorical variables using chi-square tests or Fisher exact tests when expected cell counts were small. To assess sex differences in in-basket burden, we fit multivariable linear regression models with in-basket outcomes as continuous dependent variables, physician sex (female vs male) as the primary independent variable, and covariates that included academic rank, years in practice, mean patient age, mean number of problems per patient, and panel size. We included both academic rank and years in practice to capture distinct aspects of physician experience and considered collinearity diagnostics as acceptable. We derived adjusted marginal means and sex differences with 95% CIs from each model. We conducted an analysis in the full sample and in a subgroup restricted to nonprocedural specialties to assess robustness. We excluded physicians with incomplete covariate data since comparisons showed no meaningful differences between those with complete versus incomplete in-basket data. We defined statistical significance as 2-tailed $P<.05$ and performed analyses using JMP (SAS Institute).

Results

As shown in Table 1, of 384 physicians with baseline data, sex was reported for 367 (n=146 female; n=221 male); 17 physicians with missing sex were excluded from sex-based analyses. Complete EHR in-basket data were available for 304 physicians (n=130 female; n=174 male). Of these, RVU information was available for 296 physicians (n=124 female; n=172 male), who were included in unadjusted RVU-normalized analyses. Adjusted analyses included 257 physicians (n=108 female; n=149 male) with complete covariate data.

Complete EHR in-basket burden data was available for 304 physicians (n=130 female; n=174 male). Physicians with complete in-basket data did not differ meaningfully from those with incomplete data with respect to physician sex, academic rank, years in practice, or patient panel characteristics. Table 2 presents adjusted marginal means with 95% CIs for female and male physicians, along with adjusted sex differences and associated *P* values. After adjustment for academic rank, years in practice, and panel characteristics, female physicians sustained higher RVU-normalized in-basket workloads than male physicians, suggesting that these differences are not driven solely by differences in the ratio of physicians in predominantly procedural subspecialties. In multivariable linear regression models, female physicians had higher adjusted mean daily total in-basket time, time spent completing messages, and EHR time outside 7 AM to 7 PM per RVU than male physicians. In contrast, adjusted differences in message counts were attenuated, with the largest and most consistent sex differences observed for staff messages per RVU. Significant differences in EHR in-basket burden metrics by sex persisted in analyses including all specialties and in analyses restricted to nonprocedural specialties (Table S1 in Multimedia Appendix 1).

XSL•FO
RenderX

Table 1. Internal medicine specialty physician, productivity, and patient panel characteristics.

Variables	Female physicians	Male physicians	*P* value
RVUs[a], mean (SD)	534.92 (42.3)	666.1 (44.2)	.03
Demographic characteristics, n (%)			
Academic rank[b]			<.001
Total	146 (39.8)	221 (60.2)	
Instructor	64 (43.8)	59 (26.7)	
Assistant professor	53 (36.3)	66 (29.9)	
Associate professor	22 (15.1)	51 (23.1)	
Professor	7 (4.8)	45 (20.4)	
Years in practice[b]			<.001
Total	151 (41.1)	233 (63.5)	
0-10 years	27 (17.9)	22 (9.4)	
11-20 years	68 (45.3)	70 (30.0)	
21-30 years	37 (24.5)	60 (25.8)	
31 years or more	19 (12.6)	81 (34.8)	
Panel characteristics, mean (SD)			
Patient age (years)	56.2 (9.9); n=113 patients	58.3 (8.8); n=153 patients	.07
Problems per patient (n)	14.7 (4.7); n=113 patients	14.8 (4.1); n=153 patients	.84
Patients per RVU (n)	0.37 (0.08); n=124 patients	0.21 (0.03); n=172 patients	.08
EHR in-basket burden data, mean (SD)			
Daily time spent on in-basket per physician per RVU (minutes)	314 (32.6)	246 (25.1)	.02
Daily time spent on completing messages per physician per RVU (minutes)	18,888.07 (1961.4)	14,808.09 (1507.6)	.02
Daily time spent outside 7 AM-7 PM per physician per RVU (minutes)	113.2 (17.8)	101.2 (17.4)	.14
Complete messages (n)	387.9 (40.3)	360.4 (33.2)	.17
Staff messages (n)	49.1 (5.7)	34.3 (4.1)	.03
Patient advice messages (n)	70.8 (9.13)	47.5 (5.9)	.007

[a]RVU: relative value unit.

[b]Percentantages in these columns use the value in the corresponding entry in the same column for Total under the same nested heading.

XSL•FO
RenderX

Table 2. Relative value unit (RVU)–normalized electronic health record in-basket burden metrics by physician sex. (Adjusted for physician demographics, patient characteristics, and panel size.)

In-basket efficiency metric	Unadjusted, normalized by RVU				Adjusted, normalized by RVU			
	Female (n=124), mean (SD; 95% CI)	Male (n=172, mean (SD; 95% CI)	Mean difference	*P* value	Female (n=108), mean (SD; 95% CI)	Male (n=149), mean (SD; 95% CI)	Mean difference	*P* value
Daily time spent on in-basket per RVU (min)	1.46 (1.83; 1.14-1.78)	0.80 (1.09; 0.64-0.96)	0.66	.001	1.27 (0.09; 1.25-1.29)	0.96 (0.08; 0.94-0.97)	0.31	<.001
Daily time spent on completing messages per RVU (min)	88.21(110.16; 68.82-107.6)	48.89 (65.98; 39.03-58.75)	39.3	<.001	76.40 (5.69; 75.4-77.4)	57.71 (4.65; 57.0-58.4)	18.7	.01
Daily time spent outside 7 AM-7 PM per RVU (min)	0.55 (0.79; 0.65-0.93)	0.29 (0.47; 0.22-0.36)	0.26	.002	0.44 (0.06; 0.43-0.45)	0.30 (0.05; 0.29-0.31)	0.14	.05
Messages per RVU (n)	1.36 (1.46; 1.10-1.62)	1.01 (0.93; 0.87-1.15)	0.35	.03	1.53 (0.12; 1.51-1.55)	1.41 (0.10; 1.39-1.43)	0.12	.44
Staff messages per RVU (n)	0.19 (0.21; 0.15-0.23)	0.10 (0.11; 0.08-0.12)	0.09	<.001	0.18 (0.01; 0.16-0.20)	0.12 (0.01; 0.1-0.14)	0.06	.002
Patient advice messages per RVU (n)	0.006 (0.006; 0.004-0.011)	0.004 (0.005; 0.003-0.011)	0.002	.03	0.005 (0.0009; 0.003-0.007)	0.006 (0.0008; 0.0048-0.0072)	0.001	.80

Discussion

Female physicians spent more time on all in-basket activities and received a higher number of staff and patient messages per RVU. These findings persisted even after adjusting for physicians' age, years in practice, and patient panel characteristics, and among procedural and nonprocedural specialties. Our results expand beyond primary care to internal medicine specialists, normalizing in-basket burden by RVU to emphasize that female physicians have a greater burden of in-basket work per unit of paid clinical care. These findings potentially reflect other studies on differential expectations for female physicians from patients and staff. However, our study demonstrates the pervasiveness of such sex-based differences across medical specialties, and notably, their impact on pay equity [3-5].

Limitations of our study could have influenced our findings. For instance, using data from a single academic medical center over a relatively short period could have reduced generalizability. Exclusion of physicians with incomplete data, who may have differed in unmeasured ways, could have biased the estimates. Use of RVUs could have incompletely captured clinical workloads. Finally, a lack of data on medical assistant support or clinical full-time equivalent hours was a confounder.

Our results have practical implications for physician well-being in ambulatory medical specialties. Compensation for in-basket activities would enable recognition of workload disparities. Use of artificial intelligence to automate message management could reduce inequities in uncompensated work. To address physician burnout, health systems must directly confront and mitigate sex disparities in EHR workload. Moreover, given physicians are currently not generally compensated for in-basket work, future studies should assess interventions to reduce sex differences in in-basket burden relative to compensated care delivered.

Acknowledgments

We thank Laura Bockoff for her assistance in data clean-up. We did not use generative artificial intelligence to produce this manuscript.

Funding

This study was funded through a microgrant provided by the Brigham Research Institute.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Relative value unit–normalized electronic health record in-basket burden metrics by physician sex and nonprocedural internal medicine specialties.
[DOCX File , 21 KB - jmir_v28i1e79172_app1.docx]

Multimedia Appendix 2

XSL•FO
RenderX

STROBE checklist.
[DOCX File , 38 KB - jmir_v28i1e79172_app2.docx]

References

1. Rittenberg E, Liebman JB, Rexrode KM. Primary care physician gender and electronic health record workload. J Gen Intern Med 2022 Oct 06;37(13):3295-3301 [FREE Full text] [doi: 10.1007/s11606-021-07298-z] [Medline: 34993875]
2. Liddell SS, Tomasi AG, Halvorsen AJ, Stelling BEV, Leasure EL. Gender disparities in electronic health record usage and inbasket burden for internal medicine residents. J Gen Intern Med 2024 Nov 26;39(15):2904-2909. [doi: 10.1007/s11606-024-08861-0] [Medline: 38926324]
3. Linzer M, Harwood E. Gendered expectations: do they contribute to high burnout among female physicians? J Gen Intern Med 2018 Jun 12;33(6):963-965 [FREE Full text] [doi: 10.1007/s11606-018-4330-0] [Medline: 29435727]
4. Rotenstein L, Jay Holmgren A. COVID exacerbated the gender disparity in physician electronic health record inbox burden. J Am Med Inform Assoc 2023 Sep 25;30(10):1720-1724. [doi: 10.1093/jamia/ocad141] [Medline: 37436709]
5. Rotenstein LS, Melnick ER, Jeffery M, Zhang J, Sinsky CA, Gitomer R, et al. Association of primary care physicians' electronic inbox activity patterns with patients' likelihood to recommend the physician. J Gen Intern Med 2024 Jan 20;39(1):150-152. [doi: 10.1007/s11606-023-08417-8] [Medline: 37731135]

Abbreviations

EHR: electronic health record
RVU: relative value unit

Edited by A Mavragani; submitted 25.Jun.2025; peer-reviewed by D Lew, A Murzello; comments to author 25.Nov.2025; revised version received 29.Jan.2026; accepted 30.Jan.2026; published 18.Feb.2026.

Please cite as:
Nadkarni A, Glick L, Becker M, Healey M, Stein D, Heydarpour M, Rotenstein LS
Sex Differences in Medical Specialist Physicians' Electronic Health Record In-Basket Workloads and the Implications for Compensation and Equity: Retrospective Cross-Sectional Study
J Med Internet Res 2026;28:e79172
URL: https://www.jmir.org/2026/1/e79172
doi:10.2196/79172
PMID:

©Ashwini Nadkarni, Laura Glick, Molly Becker, Michael Healey, David Stein, Mahyar Heydarpour, Lisa S Rotenstein. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Trends in Venture Capital Investment in AI-Driven Biopharmaceutical Startups

Abhishek Bazaz[1], BS; Yunan Ji[2], PhD; Mariana P Socal[3,4], MD, PhD; So-Yeon Kang[5], MBA, PhD

[1]School of Medicine, Johns Hopkins University, Baltimore, MD, United States

[2]McDonough School of Business, Georgetown University, Washington, DC, United States

[3]Health Policy and Management, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States

[4]Carey Business School, Johns Hopkins University, Baltimore, MD, United States

[5]Department of Health Management and Policy, School of Health, Georgetown University, Washington, DC, United States

Corresponding Author:
So-Yeon Kang, MBA, PhD
Department of Health Management and Policy
School of Health
Georgetown University
3700 Reservoir Rd NW
Washington, DC, 20007
United States
Phone: 1 202 687 0100
Email: soyeon.kang@georgetown.edu

Abstract

This study analyzes 2010-2024 venture capital trends in international artificial intelligence–driven biopharmaceutical startups, revealing rapid growth in discovery tool investments and concentrated US funding in California and Massachusetts.

(J Med Internet Res 2026;28:e84968) doi:10.2196/84968

KEYWORDS

venture capital; artificial intelligence; biotechnology; biopharmaceutical; drug discovery; discovery tools; investment trends; health care innovation; startups

Introduction

Artificial intelligence (AI) is increasingly recognized for its potential to reduce costs and improve efficiency across biopharmaceutical research and commercialization [1]. Previous studies have demonstrated increasing venture capital (VC) investment in the biopharmaceutical industry and, separately, in companies that utilize AI technology [2,3]. However, little systematic knowledge exists about which biopharmaceutical sectors attract VC investment for AI-related innovation, the magnitude of these investments, and the geographical distribution of funding. A clearer understanding of this investment landscape can inform biomedical researchers and entrepreneurs seeking funding for their AI-related ventures as well as policymakers designing financial incentives [2]. We examine recent trends in VC investment into AI-driven biopharmaceutical startups.

Methods

We used proprietary investment data from PitchBook, a leading third-party investment data provider that tracks VC deals across countries [4,5]. From January 2010 to December 2024, we identified VC-funded biopharmaceutical deals and extracted size, date, and financing type. We classified companies by using PitchBook's primary industry designation and supplemented this with subindustry categories (eg, drug discovery) to improve specificity. This approach has been widely adopted in prior peer-reviewed research on VC investment in health care and biopharmaceuticals [2,6,7]. All capital data were adjusted for inflation by using the Consumer Price Index and expressed in 2024 US dollars. Companies were classified as AI-related if artificial intelligence or AI were disclosed in the company's area of strategic business focus in PitchBook and non-AI otherwise. Companies were further stratified by headquarter location and primary business area according to Pitchbook: (1) biotechnology companies (developing specific therapeutics or biologics, eg, CardioGen Sciences); (2) drug discovery

XSL•FO
RenderX

(computational or experimental platforms to identify new therapeutics for in-house development, eg, Aquemia); (3) drug delivery (technology for administering drugs in a clinical setting, eg, Particle Therapeutics); and (4) discovery tools (enabling technologies, devices, or software for drug discovery by other companies, eg, Carterra) [5]. The median (IQR) was calculated for deal counts and total capital invested. The compound annual growth rate was calculated for capital investment, and chi-square analysis was conducted for company industry distribution. Analyses were conducted using RStudio (version 2023.12.0; Posit). Additional methodological details are provided in Multimedia Appendix 1.

Results

We identified 28,269 VC deals involving biopharmaceutical companies between January 1, 2010, and December 31, 2024. Of these, 1679 (5.93%) deals were associated with companies disclosing AI as a business focus. Among AI-related deals, drug discovery was the most common business focus (639/1679, 38.06% of the deals, median investment size US $9 million, IQR 3.00-30.00 million), followed by biotechnology/pharmaceutical (625/1679, 37.22%, median US $3.51 million, IQR 1.20-14.10 million), discovery tools (401/1679, 23.88%, median US $4.64 million, IQR 1.39-15.00 million), and drug delivery (14/1679, 0.83%, median US $5.23 million, IQR 4.38-11.63 million). The industry composition of VC-funded companies differed significantly between AI-related and non-AI companies. In particular, AI-related companies had a substantially higher share in discovery tools (401/1679, 23.88% vs 879/26,590, 3.31%) and minimal involvement in drug delivery (14/1679, 0.83% vs 1093/26,590, 4.11%) compared with non-AI companies, respectively (P<.001, Figure S1 in Multimedia Appendix 1). The increase in the number of VC deals across both AI and non-AI sectors was most reflected in early-stage deals. In 2024, 77.42% (96/124) of AI and 68.21% (723/1060) of non-AI VC deals were early-stage. Total capital demonstrated more equal growth trends across deal stages (Figure S2 in Multimedia Appendix 1).

Figure 1 depicts the annual deal count for biopharmaceutical startups. AI companies, both overall and within specific industries, demonstrated greater growth in VC compared with non-AI companies based on compound annual growth rate calculations (Table S1 in Multimedia Appendix 1). Among AI companies, discovery tool companies experienced the largest growth, with the median deal size of AI discovery tool companies increasing from US $0.19 million to US $7.50 million between 2010 and 2024. Their share of AI-related capital invested increased from 0.25% (0.19/76.86) to 21.85% (906.15/4147.92) from 2010 to 2024. For non-AI investments, discovery tools slightly increased from 0.87% (59.25/6816.51) to 1.58% (573.03/36,233.1) (Figure S3 in Multimedia Appendix 1).

Figure 2 illustrates the geographic distribution of VC deals in the United States in 2024. Approximately 50.74% (852/1679) of AI-related and 46.07% (12,249/26,590) of non-AI–related deals were made for US-based companies. Nearly 60% (65/110) of US-based AI deals in 2024 were in California and Massachusetts, while non-AI companies spanned 42 states. Median AI-related deal sizes varied regionally—US $10.76 million (IQR 4-39.85 million) in California compared to US $2.94 million (IQR 1.00-14.70 million) in Connecticut.

XSL•FO
RenderX

Figure 1. Trends from 2010 to 2024 in deal counts of venture capital deals by primary investment area: companies with artificial intelligence (AI) vs without AI. (A) AI deals and (B) non-AI deals.

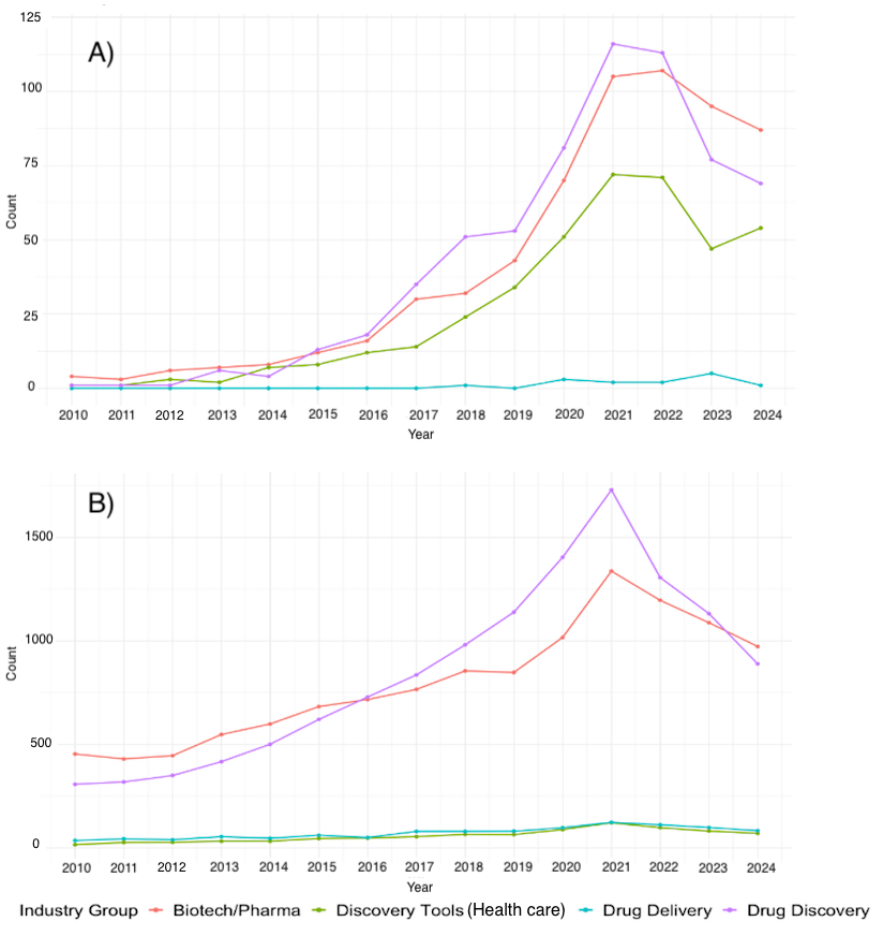

Figure 2. Geographic distribution of venture capital deal count involving biopharmaceutical companies in 2024. (A) AI deals and (B) non-AI deals. AI: artificial intelligence.

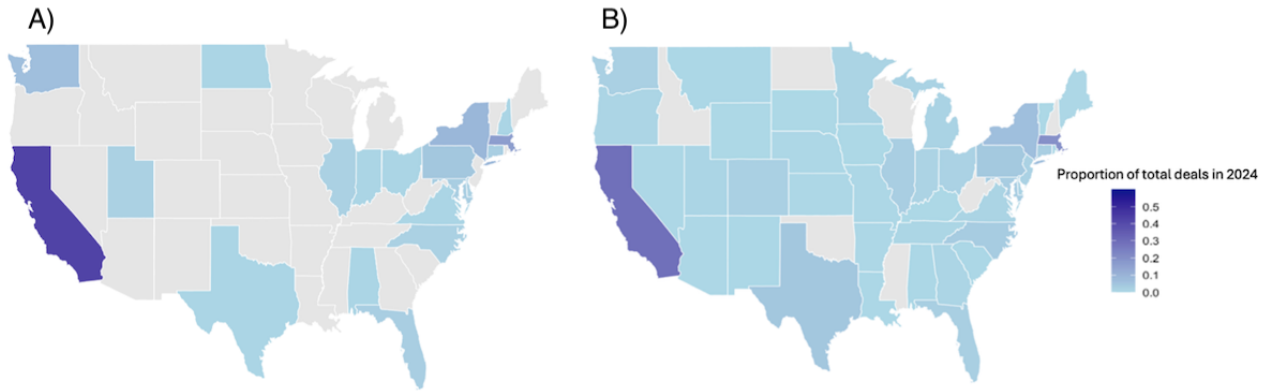

Discussion

While AI-related biopharmaceutical companies represent a minority of overall deals, they have experienced substantial growth, especially in discovery tools, where funding surged nearly 36-fold from 2010 to 2024. The majority of this growth occurred during and after the COVID-19 pandemic, in which VC funding experienced significant volatility [8]. AI-related VC activity expanded to 19 states by 2024, though investments remained highly concentrated in California and Massachusetts. These trends suggest that AI is an emerging investment focus but may be influenced by regional innovation ecosystems and biopharmaceutical research resources. Policymakers and state governments seeking to attract VC investment should consider emulating the frameworks used in these states, including research and development incentives and efficient regulatory pathways for AI-driven biotechnology ventures [9]. Future research evaluating specific regulatory and innovation policies that influence the diffusion of AI-driven health care companies and investment flows as well as the impact of AI on such

XSL•FO
RenderX

innovation warrants consideration. Limitations of this study include its reliance on proprietary data based on publicly disclosed deals and self-reported business profiles. Companies that do not publicly report their financing activities or strategic focus may be missing from our dataset. Additionally, firms that do not explicitly identify AI as a core business area may be excluded, while some companies that use the term as a marketing tool ("AI-washing") are included [10]. However, our classification based on companies' strategically disclosed business focus on AI is more conservative than methods that include companies using AI only in their internal operations when estimating VC flows into AI-driven companies.

Funding

This work was supported in part by Arnold Ventures. The funder had no role in the collection of the data, analysis, interpretation, or reporting of the data or in the decision to submit the manuscript for publication.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to their proprietary nature and restrictions under a data use agreement. However, the analyses can be reproduced by researchers with access to the same subscription-based data sources.

Authors' Contributions

AB contributed to formal analysis, methodology, visualization, writing of the original draft, and reviewing/editing. YJ and MPS contributed to conceptualization, supervision, and reviewing/editing. SYK contributed to conceptualization, data curation, funding acquisition, project administration, supervision, writing of the original draft, and reviewing/editing.

Conflicts of Interest

SYK received research funding from Arnold Ventures, the National Institute for Health Care Management Foundation, and the American Society of Clinical Oncology and consulting fees from Colorado Consumer Health Initiative. YJ received research funding from the National Institute on Aging. MPS receives research grants to Johns Hopkins Bloomberg School of Public Health from Arnold Ventures, the California Department of Health Care Access and Information, the ERISA Industry Committee, and the US Department of Defense.

Multimedia Appendix 1
Details of the methodology and results.
[DOCX File , 915 KB - jmir_v28i1e84968_app1.docx]

References

1. Leite M, de Loiola Costa LS, Cunha V, Kreniski V, de Oliveira Braga Filho M, da Cunha NB, et al. Artificial intelligence and the future of life sciences. Drug Discov Today 2021 Nov;26(11):2515-2526. [doi: 10.1016/j.drudis.2021.07.002] [Medline: 34245910]
2. Kang S, Liu M, Ballreich J, Gupta R, Anderson G. Biopharmaceutical pipeline funded by venture capital firms, 2014 to 2024. Health Aff Sch 2024 Oct;2(10):qxae124. [doi: 10.1093/haschl/qxae124] [Medline: 39411456]
3. Tricot R. Venture capital investments in artificial intelligence. OECD. 2021 Sep. URL: https://www.oecd.org/en/publications/venture-capital-investments-in-artificial-intelligence_f97beae7-en.html [accessed 2025-12-01]
4. Rathi V, Murr A, Feng A, Tauscher J, Naunheim M, Kozin E, et al. Analysis of venture capital investment in therapeutic otolaryngologic devices, 2008-2017. JAMA Otolaryngol Head Neck Surg 2019 Apr 01;145(4):387-389 [FREE Full text] [doi: 10.1001/jamaoto.2019.0045] [Medline: 30816918]
5. PitchBook. URL: https://pitchbook.com [accessed 2025-11-25]
6. Ji Y, Kang S. Global venture capital flows and US health care innovation. Health Aff Sch 2025 Nov;3(11):qxaf206. [doi: 10.1093/haschl/qxaf206] [Medline: 41234683]
7. Kang S, Lee B, Gupta R, Ballreich J, Anderson G. Research and development financing models for new biopharmaceuticals in the United States. Health Aff Sch 2025 Aug;3(8):qxaf161 [FREE Full text] [doi: 10.1093/haschl/qxaf161] [Medline: 40896379]
8. Bellucci A, Borisov A, Gucciardi G, Zazzaro A. The reallocation effects of COVID-19: evidence from venture capital investments around the world. J Bank Financ 2023 Feb;147:106443 [FREE Full text] [doi: 10.1016/j.jbankfin.2022.106443] [Medline: 35221458]
9. Moretti E, Wilson D. State incentives for innovation, star scientists and jobs: evidence from biotech. Journal of Urban Economics 2014 Jan;79:20-38 [FREE Full text] [doi: 10.1016/j.jue.2013.07.002]

XSL•FO
RenderX

10. Simonian J. AI washing: signs, symptoms, and suggested solutions for investment stakeholders. CFA Institute Research and Policy Institute. URL: https://rpc.cfainstitute.org/sites/default/files/docs/research-reports/simonian_ai_washing_report_online.pdf [accessed 2025-12-01]

Abbreviations

AI: artificial intelligence
VC: venture capital

Edited by A Coristine; submitted 28.Sep.2025; peer-reviewed by J Megerian, VV Sangaraju, T Yusuff, A Gopinath, F Amakye; comments to author 04.Nov.2025; revised version received 20.Jan.2026; accepted 21.Jan.2026; published 25.Feb.2026.

Please cite as:
Bazaz A, Ji Y, Socal MP, Kang SY
Trends in Venture Capital Investment in AI-Driven Biopharmaceutical Startups
J Med Internet Res 2026;28:e84968
URL: https://www.jmir.org/2026/1/e84968
doi:10.2196/84968
PMID:

©Abhishek Bazaz, Yunan Ji, Mariana P Socal, So-Yeon Kang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Acceptability, Feasibility, and Perceived Effectiveness of Video-Based Patient Records for Supporting Care Delivery to Older Adults With Frailty: Nonrandomized Mixed Methods Pilot Study

Phoebe Averill[1,2], PhD; Rachael Lear[3], RN, PhD; Ricky Odedra[1,4], MSc; Susannah Long[1,5], MBBS, PhD; Alex Taylor[1], BA; Pi-Jung Charville[5], RN, MSc; Jessica Fernandes[5], RN, BSc; Uzoamaka Ekeogu[5], RN, MSc; Jessica Leombruno[5], MSc; Sophia Ellis[5], MBChB, MRCP; Erik Mayer[1,4,5], MBBS, PhD

[1]NIHR North West London Patient Safety Research Collaboration, Institute of Global Health Innovation, Imperial College London, St Mary's Hospital, London, United Kingdom

[2]Better Health & Care Hub, King's College London, London, United Kingdom

[3]Robin Hood Lane Health Centre, Sutton, United Kingdom

[4]iCARE Secure Data Environment, NIHR Imperial Biomedical Research Centre, Imperial College Healthcare NHS Trust, London, United Kingdom

[5]Imperial College Healthcare NHS Trust, London, United Kingdom

Corresponding Author:
Phoebe Averill, PhD
NIHR North West London Patient Safety Research Collaboration, Institute of Global Health Innovation, Imperial College London, St Mary's Hospital, London, United Kingdom

Abstract

Background: Frailty constitutes a growing challenge for health and social care systems around the world. In England, 35% of adults aged 65 years and older live with frailty, with international estimates indicating that almost half of all hospital inpatients within the same age group are frail. This population often experiences multimorbidity and frequent care transitions. Written documentation and verbal handovers may lack the precision and nuance required to understand an older adult's presentation and support needs. Video recordings of individual patients, capturing aspects of their functional abilities and condition, may help to enhance multidisciplinary team communication and care continuity, yet little is known about their use in the care of older inpatients with frailty.

Objective: We aimed to evaluate the acceptability, feasibility of implementation, and perceived effectiveness of video-based patient records (the Isla Health Digital Pathway Platform) for supporting the assessment and care of older inpatients with frailty within the acute hospital setting.

Methods: A nonrandomized mixed methods pilot study was conducted within 3 acute medicine wards for older adults. The video-based patient records intervention, permitting videos to be embedded securely within the electronic patient record, was implemented over a 3-month period alongside usual care. Patient enrollment and retention figures; qualitative interviews with patients, carers, and clinical staff; and video capture and view metrics were used to address the study objectives. The Theoretical Framework of Acceptability of Healthcare Interventions was applied to the framework analysis of interview data, capturing concepts such as intervention ethicality, burden, and coherence. Patient and public involvement and engagement informed each research stage.

Results: Twenty-nine patients were enrolled (56.9%); 1 patient withdrew before receiving the intervention. Modal reasons given by patients for nonparticipation included not wanting to take part in research (n=8) or feeling too unwell (n=2). Staff identified multiple opportunities for capturing patient videos, including documentation of mobility assessments or seizures. The intervention was considered acceptable on the grounds that safeguards were always in place, including secure data storage and upholding of patient dignity. Implementation barriers and facilitators were identified; factors such as difficulties in capturing videos within busy ward environments and scheduling issues were voiced by participants. Video view metrics and data from interviews collectively suggested low rates of engagement with videos by clinical staff once captured. Potential intervention impacts included perceived enhancements to clinical assessment and person-centered care.

Conclusions: Our findings suggest that the intervention is largely acceptable to patients, carers, and clinical staff. Conclusions as to intervention feasibility were mixed, with limited engagement with videos suggesting further work is required to promote sufficient uptake among staff. Finally, this research presents promising patient, carer, and clinical opinion as to the potential effectiveness of video-based patient records for improving aspects of patient care.

XSL•FO
RenderX

Trial Registration: ClinicalTrials.gov NCT06504641; https://clinicaltrials.gov/study/NCT06504641

(J Med Internet Res 2026;28:e77318) doi:10.2196/77318

KEYWORDS

frailty; video recording; digital health; patient safety; quality of care

Introduction

Background

Frailty can be understood as impairment in a person's ability to recover following an injury or illness [1]. It is commonly characterized by exhaustion upon low energy expenditure, diminished muscle strength, and unplanned weight loss [2], and represents a pertinent challenge for health and social care systems globally in the context of population aging [3]. Over one-third (35%) of adults in England aged ≥65 years are thought to live with frailty [4], with a synthesis of international literature indicating a pooled prevalence rate of 47% for frailty among hospital inpatients in the same age group [5].

Older adults with frailty typically have complex care needs, experience multimorbidity, and require input from several care providers spanning multiple, fragmented health and social care organizations [6,7]. Evidence suggests that such patients are particularly vulnerable to experiencing shortfalls in safety and quality of care, owing to poor care continuity [8,9]. This can contribute to avoidable deconditioning in older adults [10], who may struggle to return to their usual activities and routines upon discharge from the hospital to community-based care.

The World Health Organization acknowledges the inherent risks to patient safety at transitions of care, where treatment delays, duplication of tests, and increased rates of adverse events are among identified challenges [11]. Frequent transitions between health and social care settings are commonplace in the care of frail older people [9]. Therefore, effective communication within and between services involved in the care of these patients is imperative [12]; providers require timely access to comprehensive, accurate, and up-to-date information about a person's presentation, in a format that readily supports meaningful interpretation [13].

Written documentation and verbal handovers vary in quality and may fail to convey the complexities of an older adult's condition, including functional capabilities and support needs. A growing body of literature suggests that the use of digital multimedia, such as photographs and videos, may help to enhance communication in patient care. A Cochrane review indicated that transmission of digital images between care providers may reduce time to commence treatment once a patient has presented to a service [14]. Moreover, a scoping review reported on the value of patient-generated photos and videos in helping professionals to obtain a more holistic understanding of patients, supporting diagnostic processes and monitoring of treatment outcomes, for example, documenting the healing of postoperative wounds [15]. Our recent systematic review synthesized empirical research evidence and professional and regulatory guidance on video recording patients for direct care purposes [16]. Review findings suggest that video recording patients to support care delivery is largely acceptable to patients, where privacy and data security risks are suitably mitigated [16].

Video recordings may support improvements in the safety and quality of care for older adults with frailty, in providing detailed visual information about a patient's functional presentation, support needs, and care preferences. Nevertheless, ethical considerations warrant detailed examination prior to the routine use of video recordings within patient care [16]. Previously, a lack of secure and efficient ways to acquire, store, and share photographs and videos may in part explain why multimedia data have not been fully exploited in health care. Further research is required to explore the acceptability, feasibility, and potential value of video recordings that are securely embedded within the electronic patient record (EPR) for use in the care of older people living with frailty.

Objectives

The primary objective of this study was (1) to determine the acceptability of using video recordings to capture the functional abilities, support needs, and preferences of older inpatients with frailty, to support clinical assessment and care delivery. Secondary objectives were (2) to explore the feasibility of implementing video-based patient records (the Isla Health Digital Pathway Platform) within the acute medical ward setting for older adults, and (3) to appraise the perceived effectiveness of patient videos for supporting care in this context. Study findings will be used to inform the decision-making about progression to a definitive trial, examining the effectiveness of patient video recordings for improving the quality and safety of care transitions.

Methods

Design

This study comprised a single-center, nonrandomized, mixed methods pilot study with an embedded process evaluation (ClinicalTrials.gov ID: NCT06504641) to inform decision-making about progression to a definitive trial to evaluate clinical impact. Where applicable, in line with guidance for nonrandomized pilot and feasibility research [17], the study reporting adheres to the CONSORT (Consolidated Standards of Reporting Trials) extension to pilot and feasibility trials [18], with intervention description informed by the Template for Intervention Description and Replication (TIDieR) guidelines [19].

Setting and Participants

Participants were recruited from 3 acute medical wards for older adults within a large acute National Health Service (NHS) Trust in London, United Kingdom. We aimed to recruit 30 patient-carer dyads and 35 clinical staff members from the

XSL•FO
RenderX

wards, according to the eligibility criteria defined in Table 1. Carers consisted of family members or friends who provided the patient with unpaid, informal care. Our sampling approach was designed to permit appraisal of acceptability and feasibility in principle, through a pragmatic yet purposively diverse sample of patients, carers, and ward staff.

Table . Overview of study eligibility criteria for the nonrandomized mixed methods pilot study involving older adult inpatients with frailty.[a]

Participants	Inclusion criteria	Exclusion criteria
Patients	• Inpatient within acute medical wards for older adults during the 3-month intervention pilot phase • Aged >65 years old • Assessed as frail or prefrail by the direct care team • Capacity to consent to participate OR lacked capacity to consent and a "personal consultee" was available to advise on patient's likely wishes about participating	• Lacked capacity to consent AND a "personal consultee" was not available to advise on the patients' likely wishes about participating
Carers	• Aged >18 years old • Provided unpaid assistance to the patient for daily activities	• Carers were excluded if the patient declined study participation
Ward staff	• Clinical staff working on 1 of 3 participating acute medical wards for older adults • Working regular shifts during the study initiation and 3-month pilot phases	• Ad hoc bank or agency staff members • Permanent staff members who are on long-term leave (eg, sickness and parental leave) during the 3-month pilot phase.

[a]Carers included family members and friends who provided the patient with unpaid, informal care.

Intervention

The Isla for Frailty intervention provided a video-based patient record function via the web-based Isla Health Digital Pathway Platform, which interfaces with EPR systems such as Cerner. Isla Health is a technology company delivering a digital pathway platform for health care staff to support patients throughout their care journey. One of its features is a visual patient record, which enables health care staff to securely capture and review videos as part of a patient's care. Video data are securely stored within encrypted cloud storage; health care staff can view these data within the EPR system, or directly via the Isla platform, using a secure weblink requiring the user's NHS email address and password. To capture visual data using the Isla platform, the user must be logged in on a mobile device (eg, smartphone and tablet) with an in-built camera and Wi-Fi connection. The Isla Health Digital Pathway Platform is approved by NHS Digital and satisfies NHS data security and protection requirements. Wards were supplied with a designated tablet device to use for video capture.

Video recordings were captured to document aspects of a participating patient's condition or functional ability considered by the direct care team as potentially useful in supporting clinical decision-making, care continuity, or multidisciplinary team (MDT) communication within the patient care journey. For example, videos might be used to capture a patient's mobility and transfers, behavior, or patient preferences. Personal care or toileting was not recorded. Patient videos were viewable by the direct care team during the patient's ward stay. Ward staff were encouraged to view and use videos in a way they felt was useful for supporting patient care, such as to support communication during shift handovers, to inform discharge planning, or in discussions with family members. The study protocol did not dictate when videos should be reviewed; therefore, the study examined the real-world use of patient videos within the acute older adult ward setting.

Participating staff had completed prior mandatory training in Data Security Awareness and were also trained in using the intervention prior to the 3-month pilot. Training emphasized the need to confirm participating patients' consent prior to video capture and to stop recording if a patient showed any signs of distress (verbal or nonverbal) or if another patient entered the shot. Staff were advised to protect patient dignity, ensuring patients were appropriately dressed and drawing curtains around the bed space where required. Contact details for the study clinical lead were shared with all participating staff who wished to discuss any questions or concerns.

Recruitment

Enrollment of Ward Staff

We engaged with staff working on the acute medical wards for older adults over a 2-month period (February 12-April 13, 2024) directly preceding the 3-month pilot phase for the video-based patient record intervention, inviting them to take part in the study. All clinical staff working on the 3 pilot wards received written and verbal information about the study aims, the intervention purpose, permitted uses of the video recordings, and information governance procedures. Staff were reassured that they could take part but opt not to be involved in video recording patients or appearing in patient videos themselves. They were also informed that they could choose not to participate in the evaluation of the video recording intervention, or to decline participation altogether without giving a reason. Staff who wished to take part were given opportunities to ask questions prior to providing their written informed consent.

XSL•FO
RenderX

Enrollment of Patients With Capacity to Consent

Patient screening and enrollment procedures are displayed in Multimedia Appendix 1. Over the 3-month intervention pilot phase (April 15-July 14, 2024), potential participants (inpatients on the 3 participating wards and their carers) were screened for eligibility by the direct care team, who explained the study purpose and provided patients and carers with participant information sheets. The direct care team flagged all potential participants to the researchers. Once the patient was clinically stable, a member of the research team visited the patient on the ward to confirm eligibility and complete enrollment procedures. Patients were offered opportunities to ask questions about the study and asked whether they might be interested in participating. Where required, easy-read participant information sheets were used to support patient understanding about what taking part would involve. Potential participants were offered additional time to decide whether they wished to take part. Where the direct care team reported concerns that a patient may lack capacity to consent to participation, the study team assessed and documented decision-specific capacity, considering whether the patient could understand what participation would involve, retain the study information, weigh this information to make a decision, and communicate their decision.

After consideration, patients who wished to participate and who had the capacity to consent were asked to sign a version of the consent form for patients with capacity. Patients were reassured that they could withdraw at any time. Within the consent process, patients were asked to consider whether they would like to remain in the study, should they lose capacity during the study period (eg, due to delirium onset). Patients were also asked whether they would still like their data to be used in the study. Participant information sheets contained a clear statement regarding retention and use of identifiable data following loss of capacity. Continued participation for such patients was subject to the same protocols for enrolling patients lacking capacity to consent, as detailed in the next section. During the study, the capacity of participants was monitored by the direct care team; the clinical lead communicated with the research team regularly about capacity concerns.

Enrollment of Patients Lacking Capacity to Consent

For patients assessed to lack the capacity to consent to participate, the research team still provided the potential participant with information about the study according to their level of understanding, using visual aids as needed (eg, the easy-read participant information sheets). In line with the requirements of Section 32(3) of the Mental Capacity Act 2005 and the Department of Health's Guidance on nominating a consultee for research involving adults who lack capacity to consent (2008), reasonable steps were taken to identify a personal consultee for these patients [20,21]. Personal consultees were identified by the patient's direct care team and included family members and carers (unpaid), who knew the patient well and who assisted them with their daily activities (eg, decisions about their welfare). Members of the ward staff team could not be nominated as personal consultees due to their own involvement in the study. Patients were excluded where no appropriate person could act as a personal consultee.

Potential personal consultees were approached first by the direct care team, who outlined the study purpose and the role of a consultee, emphasizing their right to decline acting as a consultee for the patient. For potential personal consultees who agreed to be contacted directly by the research team, a member of the research team then provided detailed written and verbal information about the study, including the consultee participant information sheet. The consultee was asked to consider the patient's likely wishes about taking part. The research team made it clear to the consultee that they were not being asked to provide their personal views on study participation, nor to consent to the study on the patient's behalf. Recruitment decisions were made by the research team in line with the consultee's advice and any previous relevant statements or wishes communicated by the patient, whether verbal or nonverbal. Consultee advice was documented on a printed consultee declaration form or via an electronic version of the form. Where a consultee advised that the patient would not have wanted to take part, the researcher abided by this. Patients were also excluded if the consultee declined to offer advice about the patient's likely wishes.

Enrollment of Carers

Carers of participating patients were also invited to take part in the evaluation during the 3-month pilot phase. Where carers were present while a member of the research team was on the ward, they were approached by the researcher accordingly. Otherwise, a ward team member contacted carers to ascertain whether they were happy to be contacted (via email or telephone) by a member of the research team. Carers who wished to participate after reviewing the carer participant information sheet were asked to sign a consent form for carers, via a printed version of the form, or an electronic form sent to the carer via email or text message. For all patients and carers, reasons for nonparticipation were documented within a screening and enrollment log.

Data Collection

Study data sources are summarized in Table 2 and described below.

XSL•FO
RenderX

Table . Overview of outcomes, measures, and data sources for each objective of the nonrandomized mixed methods pilot study involving older adult inpatients with frailty.

Study objective, outcomes, and measures	Data source
Acceptability	
Patient recruitment and retention	
Percentage of eligible participants who were enrolled into the study	Screening and enrollment log
Percentage of eligible participants declining enrollment and reasons for nonparticipation	Screening and enrollment log
Video recording requests by care team	
Number of videos requested, attempted, and submitted	Video tracker
Suitability of patient videos	Video evaluation questionnaire
Desire to see more patient videos in the future	Video evaluation questionnaire
Perceived acceptability	
Patient and carer-reported acceptability	Semistructured interview at or within 2 weeks of discharge
Ward staff team-reported acceptability	Semistructured interview after the 3-month pilot
Feasibility	
Diversity of patient sample	
Patient clinical and demographic characteristics	Screening and enrollment log
Privacy and security concerns	
Number of videos raising cause for concern reported to the clinical lead	Video tracker
Use of the Isla Health Digital Pathway Platform	
Percentage of participants with one or more video linked to EPR[a]	Isla Health Digital Pathway Platform metadata
Reasons for unsuccessful attempts to take a video	Video tracker
Video view metrics	Isla Health Digital Pathway Platform metadata
Intervention barriers and facilitators	
Patient and carer-reported intervention barriers and facilitators	Semistructured interview after the 3-month pilot
Ward staff team-reported intervention barriers and facilitators	Semistructured interview after the 3-month pilot
Perceived effectiveness	
Perceived impacts on: assessment and clinical decision-making; multidisciplinary team communication; care continuity during a hospital stay; person-centered care during a hospital stay	
Potential usefulness of intervention to ward staff team	Video evaluation questionnaire
Ward staff team-reported perspectives on intervention impacts	Semistructured interview after the 3-month pilot
Patient and carer-reported perspectives on intervention impacts	Semistructured interview at or within 2 weeks of discharge

[a]EPR: electronic patient record.

Screening and Enrollment Log

The screening and enrollment log was used to document patient recruitment status, demographic characteristics, and, where applicable, reasons for nonparticipation. A paper version of the log was initially populated. It was then digitized at regular intervals within the 3-month intervention pilot phase and stored in Imperial College Healthcare NHS Trust's iCARE secure data environment.

Video Tracker

A video tracker spreadsheet (Microsoft Excel) was developed to document patient video requests, attempts, submissions, and any issues associated with taking the recordings. The video tracker also captured information about the aspect of a participating patient's care or functional abilities to be video recorded. For example, video foci may include a person's mobility baseline or their support needs when eating and drinking.

Video Evaluation Questionnaires

Staff experience of viewing patient videos was appraised using a brief, anonymous, paper-based video evaluation questionnaire, consisting of closed-ended, multiple-choice questions with space for free-text comments (Multimedia Appendix 2). Participating

XSL•FO
RenderX

staff were asked to complete the questionnaire for each video they viewed. At the end of the 3-month pilot, questionnaire data were digitized (Microsoft Excel) and stored securely within the iCARE secure data environment.

Interviews With Patients and Carers

Patients with the capacity to consent to participation and carers were invited to take part in brief, semistructured interviews at or within 2 weeks of hospital discharge. Consent to take part in an optional interview was provided during the enrollment phase. All patients with the capacity to participate in an interview who had not declined the optional interview component at enrollment were invited to take part as they approached hospital discharge. Likewise, all participating carers were contacted to arrange the optional interview, unless they had previously declined at enrollment. Where the direct care team or research team felt it appropriate, patient capacity was assessed again prior to undertaking the interview, due to the potential for fluctuating capacity in this group. Patients and carers were asked to verbally confirm their consent prior to starting the interview. Interviews were conducted on the ward prior to discharge, or via telephone where patients had already been discharged.

Interviews were conducted by a member of the research team (PA or RO) using a topic guide for patients and carers (Multimedia Appendix 3), seeking to understand patients' experiences of being video recorded. Researchers adapted their language to the patient's level of understanding to enable patients with cognitive impairment to participate. Patient and carer interviews were completed by August 14, 2024, reflecting the date of discharge of the final patient who was eligible to take part in an interview. Interviews were audio-recorded and professionally transcribed. Names and identifiers were replaced with pseudonyms during the transcription process. Transcripts were transferred to and stored within the iCARE secure data environment.

Interviews With Ward Staff

Semistructured interviews with consenting staff members were undertaken within 2 months of the end of the 3-month intervention pilot phase. All participating staff members were approached to take part in an interview, unless they had declined this optional study component at the point of enrollment. Interviews used a version of the topic guide for staff and explored their experiences with the video-based patient records (the Isla Health Digital Pathway Platform) and perceived impacts on patient assessment and clinical decision-making; team communication; and care delivery (Multimedia Appendix 4). Audio-recorded interviews were conducted in a private space on or near the ward, or via telephone according to interviewee preference, with the final interview taking place on October 3, 2024. Written consent for the interviews was documented at recruitment. Names and identifiers were replaced with pseudonyms during the transcription process. Transcripts were transferred to and stored within the iCARE secure data environment.

Isla Health Digital Pathway Platform Metadata

At the end of the 3-month video-based patient record pilot (the Isla Health Digital Pathway Platform), video view metrics were exported from the Isla audit log. The anonymized video view metrics spreadsheet was transferred to and stored within the iCARE secure data environment.

Outcomes

In line with study objectives, the primary outcome we sought to assess was patient, carer, and ward staff team perspectives on the acceptability of the video-based patient records (the Isla Health Digital Pathway Platform). Use of the Isla platform and intervention barriers and facilitators were among secondary outcomes measured to assess the feasibility of implementing video-based patient records within the acute medical inpatient setting for older adults. Finally, 4 outcomes, including perceived impacts on care continuity in the hospital, were appraised to examine the perceived effectiveness of video-based patient records for supporting older adult inpatient care. Outcome measures and data sources for each objective are summarized in Table 2.

Data Analysis

Using Microsoft Excel, descriptive statistics were computed to summarize numerical data as to participant recruitment and retention rates; proportions of patients with videos linked to the EPR; video view metrics; and numbers of videos raising cause for concern. Frequencies and percentages were calculated for categorical variables, with means and SDs presented for continuous variables. Framework Analysis, following the 5-step approach described by Ritchie and Spencer, was applied to qualitative study data [22]. Accordingly, following a phase of familiarization with interview transcripts and free-text video evaluation questionnaire data, key concepts and ideas were noted by 1 author (PA) and discussed iteratively with 2 further authors (RO and AT), to interrogate immediate observations and assumptions about these data. At this stage, the researcher's analytical observations were used together with the Theoretical Framework of Acceptability of Healthcare Interventions by Sekhon to shape an initial thematic framework [23], combining a priori concepts with inductively generated codes. The Theoretical Framework of Acceptability of Healthcare Interventions comprises 7 constructs for appraising intervention acceptability: affective attitude, burden, ethicality, intervention coherence, opportunity costs, perceived effectiveness, and self-efficacy. Component constructs are defined in Multimedia Appendix 5. The derived thematic framework was then applied to all transcripts by 2 authors (PA and RO), with regular collaborative coding meetings convened by coders to appraise consistency and appropriateness in use of the framework. Data were then charted into a framework matrix for review and interpretation (PA, RO, and AT).

Information Power

Adequacy of the participant sample and subsequent data collection were appraised on an ongoing basis according to the 5 dimensions of "information power" [24]. Several qualities of the study suggested that a less extensive sample would be required to achieve sufficient information power. First, our a priori study objectives were narrowly focused on the acceptability, feasibility, and perceived effectiveness of a specific intervention. Second, the study eligibility criteria

XSL•FO
RenderX

ensured dense sample specificity, where participants all held characteristics which were highly specific to study objectives (eg, patients who were older adults with frailty admitted to 1 of 3 pilot study wards, carers who were unpaid carers to enrolled patients, and ward staff who were clinical staff employed on a permanent basis on one of the pilot study wards). Third, our primary objective to examine intervention acceptability was underpinned by established theory, in the form of the Theoretical Framework of Acceptability of Healthcare Interventions by Sekhon [23]. A fourth dimension, concerning the quality of dialogue, could only be assessed during the data collection phase, rather than at the study outset. We therefore reflect on this later in the paper. Finally, since Framework Analysis approaches entail both within- and between-case analysis using the derived framework matrix, we deemed that a larger volume of data would be necessary to evaluate study objectives, relative to a solely within-case analysis.

Patient and Public Involvement and Engagement

Patient and public involvement and engagement (PPIE) was central to this research. A member of the public with lived experience of caring for an older family member with frailty joined the team as a PPIE researcher, developing a storyboard characterizing patient experiences of frailty and discontinuity at care transitions. This was used to engage ward staff members as to the potential value of the research, helping to support successful staff enrollment. Contributions to data analysis, interpretation, and coauthoring this paper ensured that the perspectives of patients and carers informed each stage of the study.

Ethical Considerations

Ethics approval was granted by an English NHS research ethics committee (IRAS ID: 313814). Informed consent was obtained from all participating carers and ward staff members via completion of paper or electronic consent forms. For patients with the capacity to consent to participation, informed consent was documented by signing a paper consent form. Personal consultees, who knew the patient well and assisted them with their daily activities, were asked to consider the patient's likely wishes about taking part for those patients assessed to lack the capacity to consent to participation. Consultee advice was documented on a printed or electronic consultee declaration form. Enrollment and consent procedures are described in detail in the "Recruitment" section. Participant names and identifiers were replaced with pseudonyms and all study data were securely stored within Imperial College Healthcare NHS Trust's iCARE secure data environment. Further details are provided within the "Data Collection" section. Research participants did not receive payments or other compensation to take part.

Results

Overview

We recruited 107 study participants, consisting of 58 ward staff team members, 29 patients, and 20 carers. Twelve carers additionally acted as a personal consultee for recruited patients. Participating staff included registered nurses and nursing assistants (36/58, 62.1%); physiotherapists, occupational therapists, and therapy assistants (9/58, 15.5%); doctors (11/58, 19.0%); and administrative staff (2/58, 3.4%). Table 3 shows the baseline characteristics for patients who were invited to participate, by enrollment status. Video evaluation questionnaires were completed by just 4 staff members who had watched one or more patient videos (2 nurses and 2 therapy professionals), owing to issues with interpreting Isla Health Digital Pathway Platform metadata, detailed below alongside feasibility findings. Postintervention interviews were conducted with 70 participants, comprising 10 patients, 16 carers, and 44 staff members and lasted between 2 and 25 minutes (mean 11, SD 4.7 minutes).

XSL•FO
RenderX

Table . Baseline characteristics of older adult inpatients with frailty who were invited to participate by enrollment status.[a]

Characteristics	Enrolled patients (n=29)	Nonenrolled patients (n=22)
Sex, n (%)		
Male	18 (62.1)	12 (54.5)
Age (years), mean (range)	82.9 (72-96)	82.2 (66-92)
Ethnicity, n (%)[b]		
White	14 (48.3)	13 (59.1)
Black, Black British, Caribbean or African	3 (10.3)	2 (9.1)
Asian or Asian British	2 (6.9)	3 (13.6)
Mixed or Multiple Ethnic Groups	0 (0)	1 (4.5)
Other	8 (27.6)	2 (9.1)
Not stated	2 (6.9)	1 (4.5)
Dementia diagnosis, n (%)		
Diagnosis received	10 (34.5)	4 (18.2)
Main language spoken at home, n (%)		
English	27 (93.1)	18 (81.8)
Clinical Frailty Scale, n (%)		
1: Very fit	0 (0)	0 (0)
2: Well	0 (0)	0 (0)
3: Managing well	0 (0)	0 (0)
4: Vulnerable	0 (0)	2 (9.1)
5: Mildly frail	5 (17.2)	3 (13.6)
6: Moderately frail	8 (27.6)	4 (18.2)
7: Severely frail	16 (55.2)	13 (59.1)
8: Very severely frail	0 (0)	0 (0)
9: Terminally ill	0 (0)	0 (0)

[a]Other languages spoken at home included Urdu, Gujarati, Greek, and Kurdish.

[b]Ethnicity categories were defined according to UK Census classifications. White includes English, Welsh, Scottish, Northern Irish or British, Irish, Gypsy or Irish Traveller, Roma, and any other White background. Black, Black British, Caribbean or African includes Caribbean, African, and any other Black, Black British, or Caribbean background. Asian or Asian British includes Indian, Pakistani, Bangladeshi, Chinese, and any other Asian background. Mixed or Multiple Ethnic Groups includes White and Black Caribbean, White and Black African, White and Asian, and any other Mixed or multiple ethnic background. Other includes Arab, and any other ethnic group.

Acceptability

Patient Recruitment and Retention

Fifty-one patients were invited to participate, of whom 29 patients (56.9%) were recruited: 17 (58.6%) patients gave informed consent, while a personal consultee was involved in the decision for 12 (41.4%) further patients (Figure 1). One patient withdrew from the study prior to receiving the intervention. Among those who declined participation (22/51, 43.1%), key reasons given by patients included not wanting to take part in research (n=8), not wanting to be video-recorded (n=2), or feeling too unwell (n=2). Consultees who declined on behalf of a patient did so due to believing that the patient would not like to be video recorded (n=3), or concerns about how video recordings would be used (n=1).

Only patients with the capacity to provide their own consent to participate were invited to take part in an interview to assess outcomes. Participating patients were not excluded from interviewing based on whether they had personally received the allocated intervention during the pilot, defined as having at least one video linked to the EPR prior to hospital discharge. As such, follow-up status is displayed for all patients initially enrolled to the study.

XSL•FO
RenderX

Figure 1. Patient flow throughout the nonrandomized mixed methods pilot study (modified from CONSORT [25]). CONSORT: Consolidated Standards of Reporting Trials.

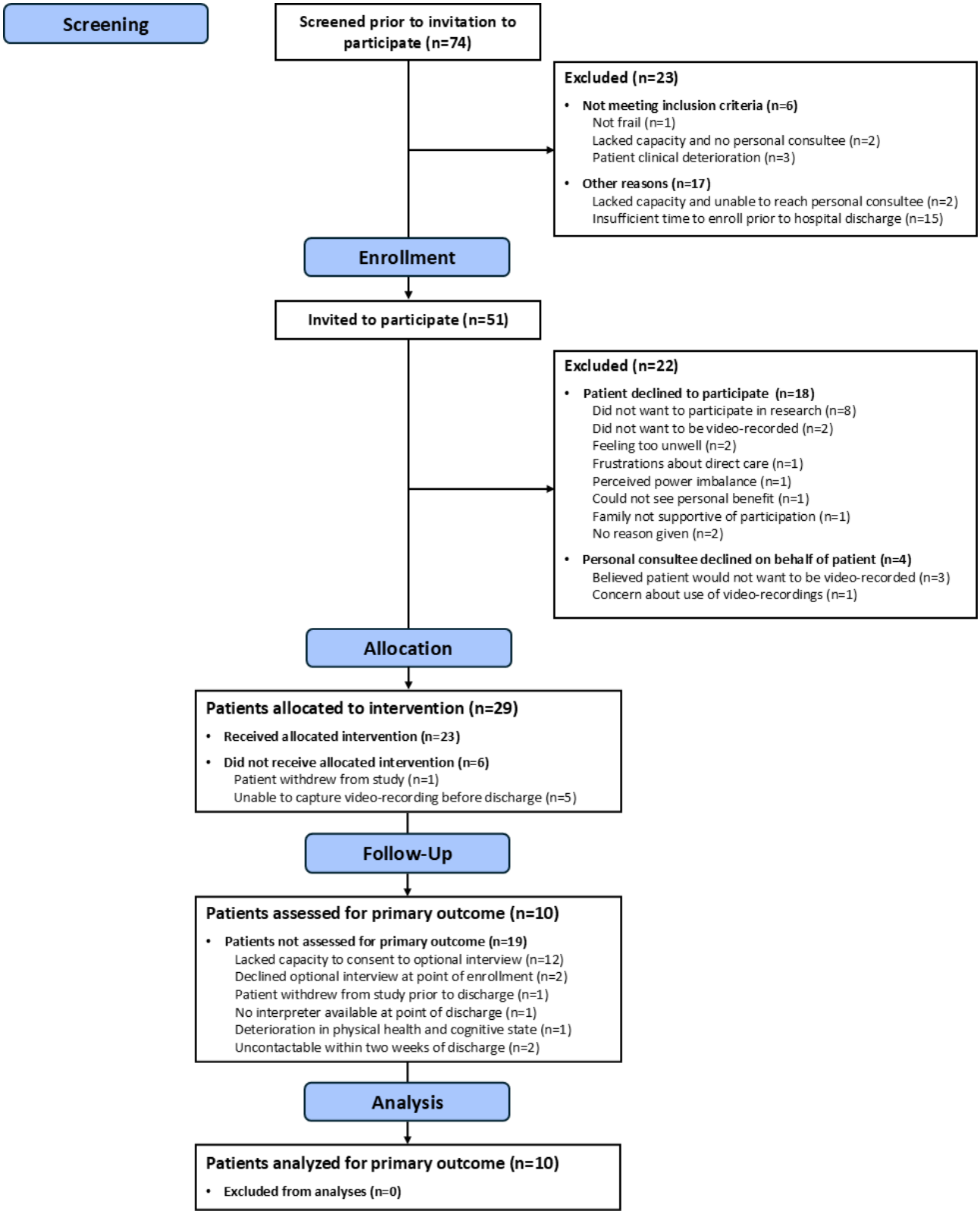

Video Recording Requests by Care Team

During the 3-month intervention pilot, 44 videos were requested by ward staff, with 37 video recordings attempted and 36 submitted within the EPR. Video recordings were focused on patient transfers (n=16), other aspects of mobility (n=13), eating and drinking support (n=3), and patient behavior (n=2). Further videos documented patient-staff interactions (n=1) and seizures (n=1). Video evaluation questionnaire data indicated the suitability of patient videos and the desire among those staff to see more videos in the future. Indeed, all 4 respondents endorsed that videos were of suitable length and quality for clinical interpretation, expressing a desire to see more patient videos in the future.

XSL•FO
RenderX

Perceived Acceptability

Overview

Interviewed patients, carers, and ward staff alike (66/70, 94.3%) typically rated video-based patient records as "completely acceptable" or "acceptable." Three participants stated that they had "no opinion" (4.3%), while a single patient felt that the intervention was "unacceptable" (1.4%); no explanation was provided for this rating. Detailed qualitative findings as to the acceptability of video-based patient records, informed by the Theoretical Framework of Acceptability of Healthcare Interventions by Sekhon, are presented as follows and summarized in Figure 2 [23].

Figure 2. Overview of nonrandomized mixed methods pilot study findings about the acceptability of using video recordings as part of older adult inpatient care.

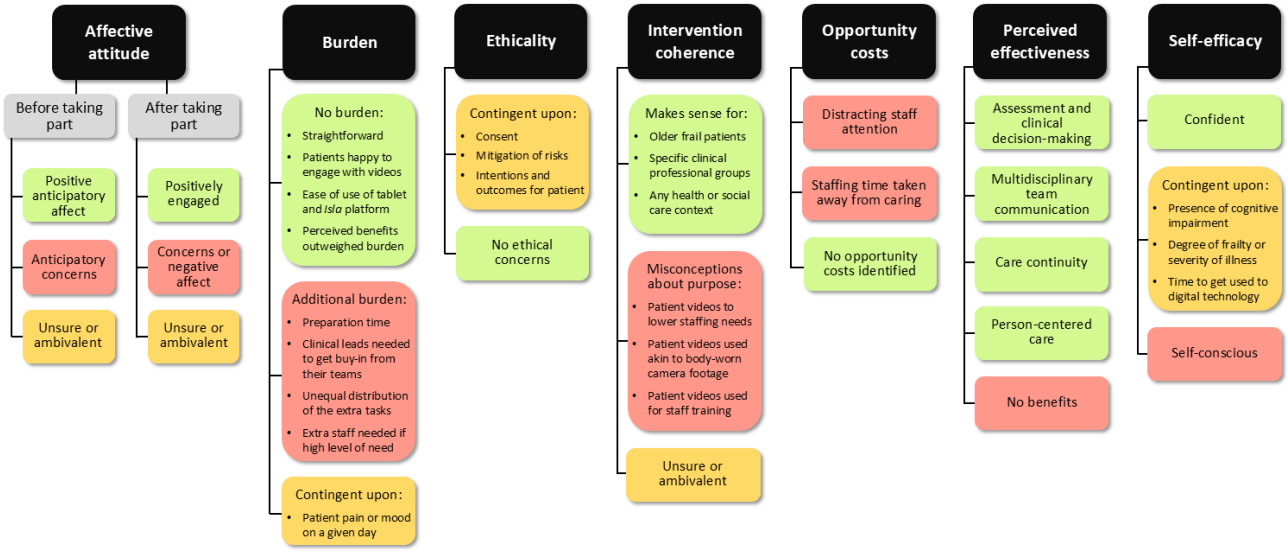

Affective Attitude

Reflecting on how they felt about the intervention prior to participating, those with positive anticipatory affect gave several reasons for this perspective. Video-based patient records were expected to help in establishing an individual's baseline function upon admission, allowing patients and staff alike to appraise progress during the inpatient stay. Staff believed the intervention could foster improved MDT communication and clinical decision-making, valuing the opportunity to pilot the intervention prior to decisions about rollout into routine care:

> *We get new things turn up all the time, "Oh we are doing this now, we got this new policy", whereas because it's being studied in a research way there is going to be lots of oversight, consent, and feeding back.* [Doctor]

Anticipatory concerns were about information security and maintaining patient dignity, particularly where patients lacked the capacity to personally consent to taking part. Staff were also apprehensive about potential workload impacts and whether patients and families would be supportive of the intervention. In practice, capturing and viewing patient videos was less burdensome than expected and staff were surprised to find that many patients were enthusiastic about taking part:

> *I was a bit skeptical, but found that the more we saw the patients, the more eager they were about the recording.* [Therapies staff]

Burden

Patients and carers believed that burdens to patients in participating in the intervention were limited:

> *I felt it was something that needed to be done, it didn't take very long. I didn't have a lot to do so it was very positive.* [Patient]

The potential to exacerbate behavioral symptoms in patients with agitation or cognitive impairment was carefully considered by staff and carers. This was mitigated by avoiding filming or terminating patient videos immediately upon any indication of patient distress (verbal or nonverbal). Burdens for staff included time required for preparation or additional staffing needs where patients required assistance of 2 for mobilization, with a third staff member needed to capture the video.

Ethicality

Participants regarded the use of video recordings within the assessment and care of older inpatients with frailty as ethical, on the grounds that certain conditions were met. A central view was that ethicality was contingent upon patient consent to appear in videos. Where a person was assessed to lack the capacity to consent, differing views were presented as to whether a personal consultee or next-of-kin should be allowed to make this decision on a patient's behalf. Given that a person's next-of-kin would typically be best positioned to understand the patient's likely wishes, next-of-kin consent was largely considered suitable. However, others voiced that patient consent was the only basis on which the intervention could be judged as ethical:

> *I would feel something like that needs to come from them [patients]… and I don't think the next-of-kin should be making that decision for them.* [Carer]

Mitigation of risks associated with video capture was another prerequisite for ethicality. A central principle was that a patient's dignity must always be upheld during video recording,

XSL•FO
RenderX

particularly given the permanence of digital patient records. Patient confidentiality and information security were further risks agreed by participants to warrant safeguards. Participants described feeling reassured about the extensive safeguards in place throughout the intervention pilot. There was consensus that patient videos should be viewed only by staff members who need to see them and must be stored in a secure digital environment that is resilient to cyberattacks.

The intentions and outcomes of the use of video-based patient records comprised another determinant of their ethicality. Where an individual's personal, religious, and cultural beliefs were considered and there was a clear clinical rationale for capturing a patient video to optimize an aspect of care, conditions for ethicality were met. Indeed, staff indicated that patient videos may help to accelerate processes of identifying discharge placements for patients, thus reducing other risks associated with lengthy hospital stays, including deconditioning or exposure to hospital-acquired infections. There was a perception that withholding such an intervention from patients, due to concerns about frailty or cognitive impairment, meant that older inpatients risked missing out on potentially enhanced care. Patients justified ethicality based on enjoying tracking their own progress and experiencing no negative outcomes from taking part:

> *It hasn't hurt me, so why should I complain?* [Patient]

Beneficial patient outcomes associated with video-based patient records were thought most likely to be realized where there is potential for an individual to return to their mobility baseline, while benefits were deemed less clear for extremely frail or cognitively impaired patients.

Intervention Coherence

It was evident from some participants' accounts that they understood how video-based patient records could support patient care and why the intervention had been piloted within the 3 wards. Those who could confidently explain the intervention aims felt it made particular sense to use patient videos in the care of older patients, where rapid changes in presentation may occur:

> *I can see now how it might be important for people to see the transformation in me, the way I was a few days ago. Now, I've lost ten years.* [Patient]

Comparisons to usual care were made at several points. Staff described the limitations inherent in text-based patient information and reflected upon the potential added value of visual data for obtaining insights into a patient's needs:

> *It's quite good technology to have, like I say, written documentation doesn't always convey what's going on, the patient's abilities. So, to actually see something, especially in regards to moving and handling, things like that.* [Nursing staff]

Staff believed that patient videos had applications for all MDT members. However, the intervention was thought most relevant to the work of physiotherapists and occupational therapists, as well as doctors. Moreover, others regarded that video recordings may be coherently used in any health or social care context.

In contrast, some patients and carers were unsure about how the intervention could be useful as part of clinical care. Furthermore, several misconceptions were expressed about the purpose of patient videos. Some staff members believed that patient videos would be used to reduce staffing requirements. Likewise, others misunderstood how and why patients may be filmed, believing that video capture would be used in the same way as body-worn cameras, which are typically used for continuous or targeted filming to document violence and aggression in health care or policing contexts.

Opportunity Costs

Opportunity costs, where other potential benefits were forgone in implementing the video-based patient record intervention, were seldom identified. A concern was that in capturing patient videos, staff may be distracted from risks on the ward, such as wandering patients:

> *If a patient fell… you obviously want to focus your attention on what's going on. You don't want to be thinking about, "Is that angle right? Am I capturing everything?"* [Therapies staff]

Second, an additional member of staff was often asked to assist in holding the tablet during video capture, thus stepping away from the task they were completing at the time.

Self-Efficacy

Finally, many participants expressed confidence in either capturing (staff) or appearing in videos (patients, carers, or staff). Self-efficacy was reportedly lower where patients had cognitive impairment; patients were also less comfortable appearing in videos on days when they were especially unwell. Some ward staff indicated initial hesitancy in using a new form of digital technology, needing time to become accustomed to using the tablet. A minority of patients and staff felt self-conscious over their physical appearance or about the sound of their voice when appearing in videos.

Feasibility

Diversity of Patient Sample

Enrolled patients were largely male (18/29, 62.1%), had a mean age of 82.9 (range 72 - 96, SD 6.7) years, mostly spoke English at home (27/29, 93%), and were diverse in ethnicity (Table 3). Clinical Frailty Scale scores for participating patients ranged from 5 (Mildly frail) to 7 (Severely frail), with most patients evaluated to be "Severely frail" (n=16) [26]. Around one-third (10/29, 34.5%) had a dementia diagnosis. Baseline characteristics were similar between enrolled and nonenrolled patients, apart from dementia diagnosis and main language spoken at home, where higher proportions of participating patients had dementia and spoke English.

Privacy and Security Concerns

During the 3-month pilot, the study clinical lead was asked to review a single patient video owing to privacy concerns. In the video, another patient could be observed stepping into the video background. The video was removed from the Isla Health Digital Pathway Platform, owing to concerns that the other patient might be identifiable from the footage.

XSL•FO
RenderX

Use of the Isla Health Digital Pathway Platform

Videos were captured on a total of 19 days within the 3-month pilot. The video recording process typically involves 2 or 3 staff members, with one to carry out the recording and others to provide direct support to the patient. Information as to those involved in this process was limited to the person capturing the video recording only, based on login data for the platform. We ascertained that 11 different staff members used their logins to capture and upload videos but were unable to determine the total number of staff members involved in the recording process in a supporting role.

Use of the Isla Health Digital Platform Pathway was similar across the 3 pilot wards. Ward 1 captured 13 videos (among n=10 recruited patients), with 17 captured on Ward 2 (among n=13 recruited patients), and 6 on Ward 3 (among n=6 recruited patients). Overall, of 29 participating patients, 79.3% (n=23) had at least one video linked to their EPR at the point of discharge. The research team was notified of 12 unsuccessful attempts to take a video. Primary reasons documented included the patient declining (n=4), patient cognitive or medical deterioration (n=3), and insufficient staffing (n=2).

Video views by ward staff were quantified through Isla Health Digital Pathway Platform metadata, where staff clicked to view and expand a video tile, resulting in a documented "View submission" action. Four such views were counted. However, toward the end of the pilot phase, we ascertained through observing staff interact with the platform that watching a video without expanding on the video tile was instead recorded within the Isla platform metadata as "View folder." Accordingly, excluding cases where a "View folder" action occurred directly before or after the same staff member uploaded a video for a given patient (when the folder view likely served a different purpose), we estimate that there were 17 video views during the study. Seemingly low video view rates among staff were corroborated by anecdotal insights obtained through staff interviews. Although staff members could typically identify potential intervention benefits at the team level, when asked directly, interviewed staff commonly attested that they had not personally viewed or sought to access a patient video at any point during the 3-month pilot.

Intervention Barriers and Facilitators

Nine themes were inductively developed from interview data as to the barriers and facilitators to implementing the video-based patient records intervention in the care of older adult hospital inpatients with frailty (Figure 3).

Figure 3. Barriers and facilitators to implementing video-based patient records in acute medical wards for older adult inpatients with frailty. EPR: electronic patient record.

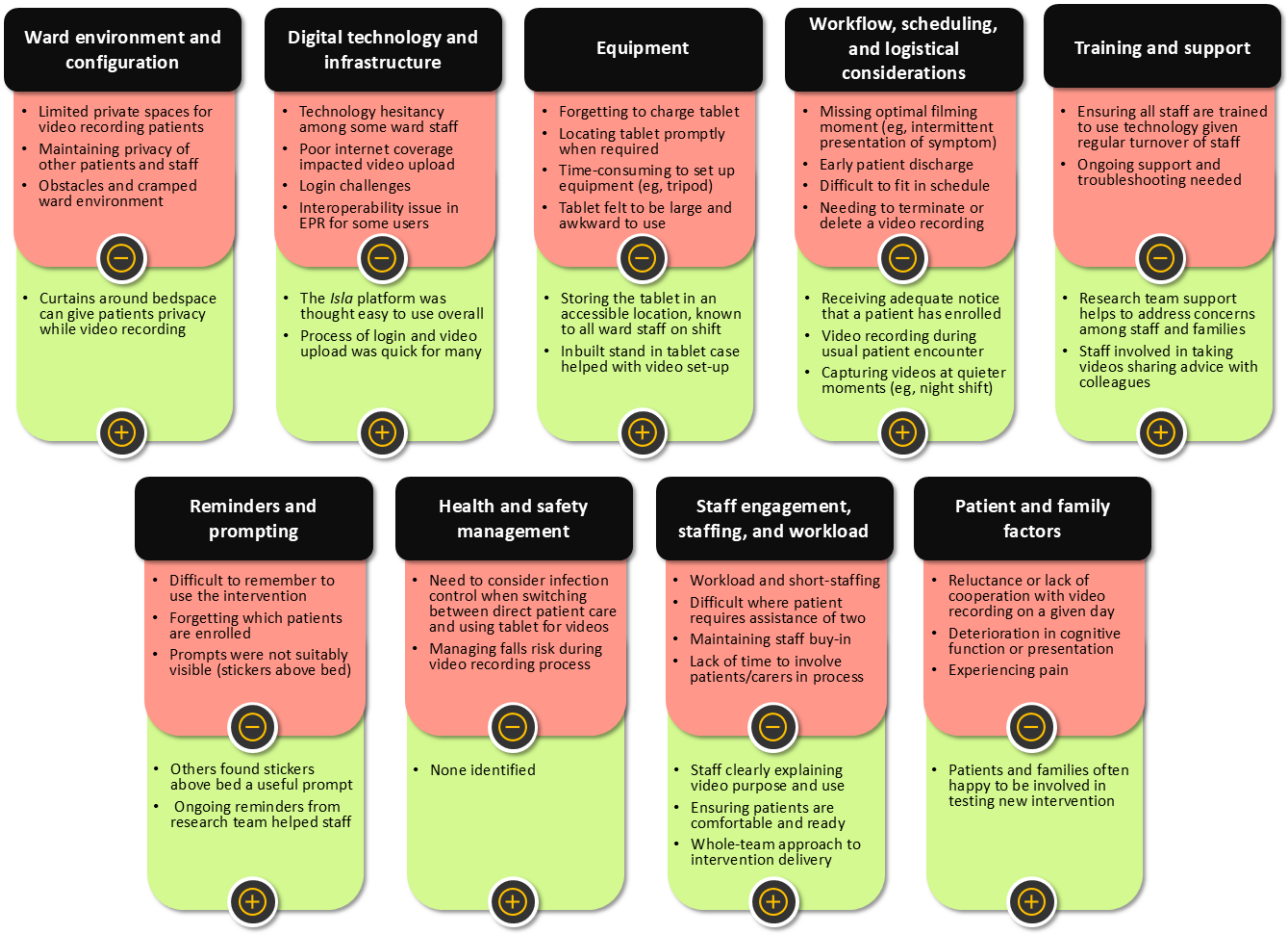

XSL•FO
RenderX

Ward Environment and Configuration

The crowded, busy ward environment and layout were such that staff struggled to capture patient videos while maintaining the privacy of participating patients. Likewise, difficulties were faced in ensuring that other patients and staff members were not captured in the background of patient videos. Curtains around the bed space were valued by staff for facilitating a degree of privacy during the video recording process.

Digital Technology and Infrastructure

Digital technology and infrastructure barriers stemmed from poor quality Wi-Fi coverage across the 3 pilot wards. Technology hesitancy and delays or challenges in logging into the video-based patient records system were also experienced by some staff members:

> *We are so tied down by logins, governance, two-factor authentication that if you want people to be able to have a low threshold to record, check, share something, you can't do it.* [Doctor]

Overall, the Isla Health Digital Pathway Platform was, however, considered straightforward to use.

Equipment

Staff reported several equipment-related challenges. Locating the designated tablet promptly when required or finding that the battery had not been charged by previous users were key frustrations. An inbuilt stand within the tablet case served as a facilitator to the video recording process, allowing staff to position the tablet on a surface within the ward (eg, bedside table) when another staff member could not assist with video capture.

Workflow, Scheduling, and Logistical Considerations

Further obstacles to using the intervention are related to ward workflow and logistical considerations. Videos often needed to be taken at an optimal moment when seeking to capture intermittent presentations (eg, seizures) or for fatigued patients. Patients were sometimes discharged earlier than expected, meaning that video recordings could not be captured in time or used to realize potential benefits. Staff described integrating video capture into their usual planned therapy sessions with participating patients, or capturing videos during quieter periods, to maximize efficiency and feasibility.

Training and Support

Regular staff turnover presented difficulties in ensuring all staff were trained in using the intervention. From the staff engagement and enrollment phase through to the end of the 3-month pilot, there were rotations in both junior doctors and specialty trainees on a national level. When asked about low apparent usage of videos, some staff members indicated that they could not recall how to access videos within the EPR:

> *I think it's quite straightforward when you record it, but I think there is still some sort of gap, like how you follow up with Isla outside of the iPad [tablet].* [Therapies staff]

Nevertheless, staff and patients alike appreciated access to ongoing training and support from the research team throughout the 3-month trial period.

Reminders and Prompting

Remembering to capture and use videos within patient care was a persistent challenge. This served as a barrier to uptake and achieving a meaningful intervention "dosage" for participants, where it was felt that more videos throughout an individual's inpatient stay would be necessary to visually demonstrate improvement or deterioration.

> *I was naively hoping it would be more visible. Having people nudging and reminding definitely helps, but that would be my concern for wider adoption.* [Doctor]

For other staff, there appeared to be low impetus to access and use patient videos once captured:

> *Not many people would go and check, or would say, "Okay, on Cerner [EPR], I want to go and see that video."* [Therapies staff]

Reminders used throughout the pilot study period (eg, stickers above beds of enrolled patients and reminders from the research team) were valued but thought insufficient for comprehensive implementation.

Health and Safety Management

Managing safety when capturing videos, including staff's ability to prevent patient falls during video-recorded mobility assessments, was an important consideration. Other challenges included infection control considerations when switching between delivering direct patient care and using the tablet to capture a patient video.

> *I got pointed out for infection control… they recommended me leaving the iPad [tablet] there, going to wash my hands, come back and go onto the tablet. But I can't leave the patient.* [Therapies staff]

Staff detailed practical issues in capturing videos, where handwashing was required between each activity.

Staff Engagement, Staffing, and Workload

Moreover, intervention feasibility was also contingent upon staff engagement, staffing, and workload factors. Barriers were presented when wards were short-staffed, resulting in high workload for clinical teams and limited spare capacity to assist with capturing videos. Reassuring staff that video recordings were not intended to be used for appraising staff practice and recognizing the efforts of staff to deliver the intervention were important for maintaining staff engagement. Implementation was reportedly best where an MDT approach to capturing patient videos was adopted, so that individual professional groups (ie, nursing or therapy staff) did not feel alone in shouldering additional task burdens. Likewise, carers emphasized the importance of staff taking time to explain to patients why they wished to capture a video of a given activity or symptom and how the video may be used as part of their care:

XSL•FO
RenderX

> *Is there rapport being built up? Is the person comfortable with those particular healthcare professionals?* [Carer]

By allowing time to develop positive staff-patient relationships, patients felt more comfortable with taking part.

Patient and Family Factors

Finally, patient and family factors were important to the implementation of video-based patient records. On some days, interviewees described that patients were reluctant to accept any care or interventions on the ward, including patient videos. Older, frail patients may decline to engage owing to experiencing pain or fatigue on a given day. Seemingly uncooperative behavior was displayed at times among participating patients due to confusion or fluctuation in cognitive function, presenting a barrier to delivering the video-based patient record intervention:

> *When they were trying to video something more positive with him, he decided not to be as positive! But that's hit and miss with any of the patients I would think.* [Carer]

Nevertheless, a key facilitator indicated by all participant groups was that patients and carers were often happy to be involved in trialing a new intervention which may improve their own care or that of others.

Perceived Effectiveness

Overview

Perceived effectiveness of the video-based patient records intervention was appraised according to ward staff views on its potential usefulness in the care of older adult inpatients with frailty, as well as patient, carer, and staff perspectives on intervention impacts. Participants conceived of multiple ways in which video-based patient records could be potentially effective for use in the present context. Some participants were unaware of any intervention impacts. Such patients and carers described a lack of follow-up from ward staff about their videos, meaning that they were unsure whether the video recordings had been beneficial. No adverse events were reported. Study data indicated several ways in which video recordings may improve aspects of patient care.

Assessment and Clinical Decision-Making

Data from the 4 completed video evaluation questionnaires showed that ward staff considered the video they had watched was "very useful" for supporting patient assessment and clinical decision-making. Interviewees elaborated on the potential for patient videos to improve assessment quality by mitigating subjectivity in clinical interpretation and speeding up the assessment process. For behaviors or symptoms occurring at unpredictable intervals, videos also permitted staff to obtain specialist advice from other teams:

> *It helped our neuro team see a patient who was having intermittent seizures… they [neurology team] would not have been able to see it otherwise.* [Nursing staff]

Further potential benefits were in offering clear information about an individual's care and support needs, providing evidence for referrals to other care providers. Staff reflected on challenges faced in securing onward care placements for patients who had exhibited fluctuating behavior and nursing care requirements in the hospital. Rehabilitation settings and nursing homes reportedly declined referrals or overestimated staffing costs necessary to provide patient care, based on patients having had an acute or episodic change in presentation (eg, delirium onset) during their admission:

> *If we had someone come in for assessments, we had this evidence to say "Okay, look, on a good day this patient did it this way, so you need to assess them based on that."* [Nursing staff]

In contrast, an alternative position was that videos merely provided a snapshot of a person's presentation:

> *You're putting someone into a slightly odd position saying, "Now, walk for us." And I think you'll get the Charlie Chaplin effect, people starting to do things that maybe they wouldn't normally do… I think a camera is not always going to bring out the truth in people.* [Carer]

Therefore, the potential for video-based patient records to improve assessment and decision-making precision was questioned.

Multidisciplinary Team Communication

All 4 staff members who completed video evaluation questionnaires reported that the video they had watched was "very useful" for communicating patient information to colleagues. The added communicative value of visual data over written documentation or verbal handovers was apparent from interview data. Carers were similarly supportive of patient videos, where their use might improve communication quality and concision:

> *I don't know whether that would completely eradicate the need for discussion… but it could perhaps enhance that, if there's something that someone needs to show someone else.* [Carer]

Participants also envisaged potential value for strengthening communication with other clinical settings. Carers felt that videos could optimize care timeliness and continuity by enabling remote communication with other care providers. This view was supported by staff, who also believed that patient videos may improve collaboration between clinical services, potentially reducing duplication of clinical assessments:

> *We have a deficit of trust across boundaries… but if you see a video they know we are not making it up.* [Doctor]

Moreover, videos were reportedly used to enhance communication with patients' families:

> *Sometimes the family says, "Oh, he could not do it at home." When they see the patient sitting out and walking, it's amazing them.* [Nursing staff]

As such, this helped to provide a tangible demonstration of care and interventions received in the hospital and apparent impacts.

XSL•FO
RenderX

Care Continuity

Study findings suggest that video-based patient records could improve care quality by optimizing continuity, both within and beyond an individual's hospital stay. Patient videos may support joined-up care by helping to establish a shared understanding of the goals of a patient's admission, informed by videos showing their functional baseline. Similarly, participants reported that videos allowed staff to objectively monitor patients over time, facilitating greater awareness of signs of improvement or deterioration:

> *For the patient it's better because we know their progress… I'm video-recording to compare. We see the difference from bedbound to mobilizing.* [Nursing staff]

Video-based patient records were also deemed valuable for aiding safe care transitions. Staff thought patient videos could assist community therapy teams to provide joined-up care following discharge. Since frail older adults often experience multiple hospital admissions, participants also envisaged using patient videos to gain insights into a person's presentation during prior inpatient stays.

Person-Centered Care

The potential for video-based patient records to improve person-centered care was a prominent theme. By providing a visual record of a patient's achievements throughout their hospital admission, patient videos promoted a focus on an individual's goals and needs, contributing to more personalized care experiences:

> *I was proud to see me in the video and pleased that I have made so much progress… it gave me great spirits.* [Patient]

Moreover, the use of patient videos was thought to improve the quality of staff-patient engagement. Staff described how the process of delivering the intervention offered opportunities for patients to receive additional encouragement from ward staff, which in turn improved staff satisfaction:

> *The patient was eager, because he was walking and the staff… they were following and cheering him on. It was so beautiful – he was able to mobilize himself. They were all so happy, the staff and the patient themself [sic] because he wanted to go home!* [Nursing staff]

Finally, there was a sense that patients appreciated the choice over whether to take part in the intervention:

> *It gives the patient a little say in what's going on in their life as a patient.* [Patient]

Where patients may feel they have limited involvement in care decision-making when in hospital, the opportunity to make choices about their care appeared to instill a sense of agency and empowerment.

Discussion

Principal Findings

Findings from this nonrandomized mixed methods pilot study suggest that the use of video recordings within patient assessment and care delivery is largely acceptable to older patients with frailty, their family members or carers, and clinical staff. This was demonstrated by promising participant recruitment and retention rates; enrollment targets for patients were almost met, while ward staff recruitment vastly exceeded intended numbers. The range of videos captured, from documentation of seizures through to establishing a person's mobility baseline, shows that staff discerned multiple use cases for patient videos. On the grounds that ethical preconditions are met, including mitigation of concerns about information security and patient dignity, video-based patient records were largely deemed to be acceptable.

A secondary objective to explore the feasibility of implementing video-based patient records within acute wards for older adults revealed mixed findings. Diversity of the patient sample in terms of clinical and demographic characteristics was encouraging, indicating that risks of widening health inequalities through inequitable uptake are low, should the intervention be implemented into routine care. Implementation barriers were multiple, yet participants described a range of facilitators that supported successful delivery. However, video view metric estimates and anecdotal evidence from interviews with ward staff suggested limited engagement with videos once captured, thus casting doubt on the feasibility of attaining sufficient uptake among staff. We estimate that over half of all videos captured were never viewed. Nevertheless, patient, carer, and clinical opinion as to perceived intervention effectiveness yielded evidence of promise. Videos were considered useful in supporting clinical assessment, enhancing MDT communication, strengthening care continuity, and promoting person-centered care.

Comparison With Wider Literature

Our findings are largely in accordance with the conclusions of a recent systematic review, which examined empirical research and regulatory guidance on the generation and use of video recordings as part of patient care [16]. Indeed, similarly to existing studies [16,27,28], we observed that video-based patient records were acceptable to patients, carers, and clinical staff alike, so long as ethical conditions were satisfied around upholding patient dignity and maintaining the security of these digital data [29]. Likewise, based on clinical opinion, we found that patient videos may aid the monitoring of care delivery and outcomes over time, aligning with the conclusions of comparable research [15].

Perhaps most striking was our finding that video-based patient records appear to be valuable for promoting person-centered care within the acute hospital setting. The intervention seemed to instill a sense of agency and empowerment in older adults with frailty who took part in video recordings. Staff experienced secondhand satisfaction upon observing a patient demonstrating a functional improvement, such as walking unaided, when they had previously felt too frail to do so. This finding resonates with

XSL•FO
RenderX

prior research from physiotherapy and rehabilitation contexts, where video recordings documenting improvements in patient gait throughout their inpatient stay were found to improve patient motivation and satisfaction [30]. Likewise, although we piloted the use of patient videos captured by ward staff, rather than exploring patient-generated visual media as reported within a further review [15], we similarly concluded that patient videos helped staff to gain a more individualized understanding of patients. A deeper understanding of patients as individuals likely, in part, underpinned the closer staff-patient engagement anecdotally reported by participating staff and patients.

Older adults with frailty often face multiple transitions between health and social care settings, owing to complex care needs stemming from multimorbidity [6,7]. An important application of the present intervention, as conceived by ward staff, lay in the potential to capture video footage which could be shown to other departments or providers to inform decision-making and support joined-up care. For example, staff spoke about the potential value in sharing footage with neurology specialists or with staff carrying out assessments for care home placements. For the purposes of this pilot study, staff could only show video recordings to health care professionals outside the direct care team who visited the pilot wards in person. Should wider implementation of video-based patient records be considered, the lack of interoperability of different EPR systems, a well-documented obstacle to delivering digital interventions at scale, warrants timely address [31,32].

The integration of new digital technologies into health care services resembles a complex challenge, with tailored strategies required to promote successful adoption and sustainability of a given intervention [33]. Our findings point to a series of barriers and facilitators to the implementation of video-based patient records, including those pertaining to staff engagement, workflow, and the adequacy of reminders and prompting used to encourage staff to capture and view patient videos. So that potential enhancements to patient care can be realized, theoretically informed strategies drawing upon behavioral science or implementation science frameworks are likely required to encourage uptake [34].

Finally, it is well documented that severely frail patients and people who are unable to consent to participation are seldom included within research studies [35,36]. Our findings reinforce existing literature which points to the need for careful ethical review of procedures for patient inclusion [37], to increase opportunities for such patient populations to contribute to applied clinical and health services research. This is so that they are not excluded from the development of innovations which may result in enhanced care.

Strengths and Limitations

Strengths of this study lie in its representation of patient populations who are underrepresented and underserved in clinical research. This includes older adults with severe frailty and people who lack the capacity to consent to research owing to cognitive impairment [35]. Exclusion of such patients from clinical trials is widespread, thus obstructing the potential to develop improved health care services and interventions that meet the needs of these populations [36]. High quality and comprehensiveness in qualitative data collection components were indicated according to multiple dimensions of information power [24]. For instance, data were sufficiently comprehensive given our narrow study aims, high sample specificity while capturing variation in patient demographic and clinical characteristics, and owing to the application of established theory. We ensured rigor in our analysis of interview data through the Framework Analysis approach [22], applying an existing theoretically informed framework to guide data charting [23], while inductively introducing new codes into the coding scheme to characterize each construct. Further strengths were in our approach to PPIE, ensuring that lived experience perspectives guided the study throughout the research cycle.

Issues in interpreting Isla Health Digital Pathway Platform metadata, such that video view figures were estimates that cannot be determined with exactness, represent an important limitation. Other limitations pertain to the study design. The single-site, nonrandomized nature of this research and the lack of a comparator group should be taken into account when considering our findings. While we appraised the quality of interview dialogue to be high overall, we hold that dialogue flow was at times impacted within the busy ward environment in which data were gathered. Interviews with patients were often inadvertently interrupted by staff or visitors, while staff were regularly distracted by other tasks (eg, incoming telephone calls) while being interviewed. We also note that among participating patients with the capacity to consent to an optional interview during enrollment, a number of these individuals were not interviewed due to reasons such as being uncontactable within 2 weeks of leaving the hospital, deterioration in cognitive state, and lack of interpreter availability at the point of discharge. Had it been possible to interview these individuals, it is plausible that a wider range of perspectives on the intervention may have surfaced, with greater information power to permit between-case analysis. Nonetheless, we offer a robust evaluation of intervention feasibility, applying our findings stringently to decision-making about progression to a definitive trial.

Implications

Study findings provide evidence that the use of video-based patient records within direct care delivery for older adults with frailty was largely deemed to be acceptable. We also offer compelling preliminary insights based on participating patient, carer, and clinical staff opinions as to the potential effectiveness of this intervention in the acute medical inpatient ward context. When implemented successfully, our findings suggest that the intervention could contribute to improved care quality and safety. However, the feasibility of intervention implementation must be questioned, given estimated video view metrics and anecdotal evidence from staff interviews, which indicated that uptake among staff in terms of capturing and viewing patient videos did not reach a threshold to permit full realization of benefits. Evidence for ethicality and a lack of adverse outcomes suggests that the intervention could be considered appropriate for use within other clinical services when indicated, for example, as part of efforts to empower patients and to promote person-centered care. Further professional and regulatory guidance to support clinical staff in the implementation of such an intervention is, however, warranted [16].

XSL•FO
RenderX

Conclusions

Addressing the primary objective of this pilot study, the video-based patient record intervention was found to be largely acceptable to older inpatients with frailty, carers, and staff. Participants also held promising opinions about possible intervention impacts. Optimizing assessments, enhancing care continuity, and empowering patients were considered important potential benefits of using patient videos within direct care. No adverse events were reported. Findings about patient diversity, alongside barriers and facilitators to intervention implementation, support encouraging indications overall as to the feasibility of implementing such an intervention within this care context. However, apparent limited engagement with patient videos after their capture during the pilot phase would likely obstruct our ability to measure intervention efficacy in future research. As such, we cannot endorse progression to a definitive trial to examine the effectiveness of patient videos for improving safety and quality of care transitions. Nevertheless, our findings provide an initial indication that the intervention could be ethically implemented into routine practice, with suitable ethical safeguards in place.

Acknowledgments

The authors would like to thank the patients, carers, and ward staff who took part for their valuable contributions to this research. We also express our sincere thanks to Dr Colin Mitchell and Dr David James for supporting the successful implementation and evaluation of the Isla for Frailty intervention within their wards.

Funding

This study was funded by the National Institute for Health and Care Research (NIHR) North West London Patient Safety Research Collaboration (NIHR204292) with Infrastructure support provided by the NIHR Imperial Biomedical Research Centre (NIHR203323). PA is currently supported by a Better Health & Care Hub Postdoctoral Fellowship awarded by the Better Health & Care Hub, King's College London. Views expressed in this publication are those of the authors and not necessarily those of the NIHR, Isla Health, nor the Better Health & Care Hub.

Data Availability

The dataset generated and analyzed during this study is not publicly available as consent for data sharing was not sought from research participants and was not agreed as part of the ethics approval obtained to undertake this research. Any queries about study data may be directed to the corresponding author.

Authors' Contributions

Conceptualization: RL (lead), SL, SE, EM
Formal analysis: PA (lead), RO, AT
Funding acquisition: RL, EM
Investigation: PA, RL, RO, SL, AT, PJC, JF, UE, JL
Methodology: RL, SL, SE, AT
Project Administration: PA, RL, RO
Supervision: PA, RL, SL
Writing – original draft: PA, RL
Writing – review & editing: PA, RL, RO, SL, AT, PJC, JF, UE, JL, SE, EM

Conflicts of Interest

None declared.

Multimedia Appendix 1
Procedure for patient screening and enrollment.
[DOCX File, 49 KB - jmir_v28i1e77318_app1.docx]

Multimedia Appendix 2
Video evaluation questionnaire completed by ward staff upon viewing a patient video.
[DOCX File, 27 KB - jmir_v28i1e77318_app2.docx]

Multimedia Appendix 3
Interview guide for patients and carers.
[DOCX File, 26 KB - jmir_v28i1e77318_app3.docx]

XSL•FO
RenderX

Multimedia Appendix 4
Interview guide for ward staff.
[DOCX File, 26 KB - jmir_v28i1e77318_app4.docx]

Multimedia Appendix 5
Component constructs of the Theoretical Framework of Acceptability of Healthcare Interventions.
[DOCX File, 26 KB - jmir_v28i1e77318_app5.docx]

References

1. What is frailty? Age UK. 2020. URL: https://www.ageuk.org.uk/our-impact/policy-research/frailty-in-older-people/understanding-frailty/ [accessed 2025-03-07]
2. Introduction to frailty. British Geriatrics Society. 2014. URL: https://www.bgs.org.uk/resources/introduction-to-frailty [accessed 2025-03-07]
3. Dlima SD, Hall A, Aminu AQ, Akpan A, Todd C, Vardy E. Frailty: a global health challenge in need of local action. BMJ Glob Health 2024 Aug 9;9(8):e015173. [doi: 10.1136/bmjgh-2024-015173] [Medline: 39122463]
4. Proactive care: providing care and support for people living at home with moderate or severe frailty. NHS England. 2023. URL: https://www.england.nhs.uk/long-read/proactive-care-providing-care-and-support-for-people-living-at-home-with-moderate-or-severe-frailty/ [accessed 2025-04-04]
5. Doody P, Asamane EA, Aunger JA, et al. The prevalence of frailty and pre-frailty among geriatric hospital inpatients and its association with economic prosperity and healthcare expenditure: a systematic review and meta-analysis of 467,779 geriatric hospital inpatients. Ageing Res Rev 2022 Sep;80:101666. [doi: 10.1016/j.arr.2022.101666] [Medline: 35697143]
6. Hewitt J, McCormack C, Tay HS, et al. Prevalence of multimorbidity and its association with outcomes in older emergency general surgical patients: an observational study. BMJ Open 2016 Mar 31;6(3):e010126. [doi: 10.1136/bmjopen-2015-010126] [Medline: 27033960]
7. Sinclair DR, Maharani A, Chandola T, et al. Frailty among older adults and its distribution in England. J Frailty Aging 2022;11(2):163-168. [doi: 10.14283/jfa.2021.55] [Medline: 35441193]
8. Sadler E, Potterton V, Anderson R, et al. Service user, carer and provider perspectives on integrated care for older people with frailty, and factors perceived to facilitate and hinder implementation: a systematic review and narrative synthesis. PLoS One 2019;14(5):e0216488. [doi: 10.1371/journal.pone.0216488] [Medline: 31083707]
9. Baillie L, Gallini A, Corser R, Elworthy G, Scotcher A, Barrand A. Care transitions for frail, older people from acute hospital wards within an integrated healthcare system in England: a qualitative case study. Int J Integr Care 2014 Jan;14(1):e009. [doi: 10.5334/ijic.1175] [Medline: 24868193]
10. Welch C, Chen Y, Hartley P, et al. New horizons in hospital-associated deconditioning: a global condition of body and mind. Age Ageing 2024 Nov 1;53(11):afae241. [doi: 10.1093/ageing/afae241] [Medline: 39497271]
11. Transitions of care. World Health Organization. 2016. URL: https://iris.who.int/handle/10665/252272 [accessed 2025-04-10]
12. Coffey A, O'Reilly P, Meskell P, et al. The development of a national transfer document: for use when an older person is being transferred from residential to acute care settings. : Health Service Executive; 2019 URL: http://hdl.handle.net/10147/627131 [accessed 2025-12-19]
13. Kripalani S, LeFevre F, Phillips CO, Williams MV, Basaviah P, Baker DW. Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. JAMA 2007 Feb 28;297(8):831-841. [doi: 10.1001/jama.297.8.831] [Medline: 17327525]
14. Gonçalves-Bradley DC, J Maria AR, Ricci-Cabello I, et al. Mobile technologies to support healthcare provider to healthcare provider communication and management of care. Cochrane Database Syst Rev 2020 Aug 18;8(8):CD012927. [doi: 10.1002/14651858.CD012927.pub2] [Medline: 32813281]
15. Ploderer B, Rezaei Aghdam A, Burns K. Patient-generated health photos and videos across health and well-being contexts: scoping review. J Med Internet Res 2022 Apr 12;24(4):e28867. [doi: 10.2196/28867] [Medline: 35412458]
16. Lear R, Ellis S, Ollivierre-Harris T, Long S, Mayer EK. Video recording patients for direct care purposes: systematic review and narrative synthesis of international empirical studies and UK professional guidance. J Med Internet Res 2023 Aug 16;25:e46478. [doi: 10.2196/46478] [Medline: 37585249]
17. Lancaster GA, Thabane L. Guidelines for reporting non-randomised pilot and feasibility studies. Pilot Feasibility Stud 2019;5(1):114. [doi: 10.1186/s40814-019-0499-1] [Medline: 31608150]
18. Eldridge SM, Chan CL, Campbell MJ, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016 Oct 24;355:i5239. [doi: 10.1136/bmj.i5239] [Medline: 27777223]
19. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348(mar07 3):g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]
20. Mental capacity act 2005. Legislation.gov.uk. URL: https://www.legislation.gov.uk/ukpga/2005/9/section/32 [accessed 2025-02-10]

XSL•FO
RenderX

21. Department of Health Scientific Development and Bioethics Division. Guidance on nominating a consultee for research involving adults who lack capacity to consent. : Department of Health; 2008 URL: https://www.manchester.gov.uk/downloads/file/12218/guidance_on_nominating_a_consultee_for_research_involving_adults_who_lack_capacity_to_consent [accessed 2025-12-29]
22. Ritchie J, Spencer L. Qualitative Data Analysis for Applied Policy Research Analyzing Qualitative Data: Routledge; 1994:173-194. [doi: 10.4324/9780203413081_chapter_9]
23. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88. [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
24. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
25. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340:c869. [doi: 10.1136/bmj.c869] [Medline: 20332511]
26. Moorhouse P, Rockwood K. Frailty and its quantitative clinical evaluation. J R Coll Physicians Edinb 2012;42(4):333-340. [doi: 10.4997/JRCPE.2012.412] [Medline: 23240122]
27. Bayen E, Jacquemot J, Netscher G, Agrawal P, Tabb Noyce L, Bayen A. Reduction in fall rate in dementia managed care through video incident review: pilot study. J Med Internet Res 2017 Oct 17;19(10):e339. [doi: 10.2196/jmir.8095] [Medline: 29042342]
28. Towsley GL, Wong B, Mokhtari T, Hull W, Miller SC. Piloting me and my wishes-videos of nursing home residents' preferences. J Pain Symptom Manage 2020 Mar;59(3):609-617. [doi: 10.1016/j.jpainsymman.2019.10.030] [Medline: 31711970]
29. Williams K, Arthur A, Niedens M, Moushey L, Hutfles L. In-home monitoring support for dementia caregivers: a feasibility study. Clin Nurs Res 2013 May;22(2):139-150. [doi: 10.1177/1054773812460545] [Medline: 22997349]
30. Jayabalan P, Kaplan R, Breisinger T, et al. Poster 311 Video recording the gait of stroke patients during inpatient rehabilitation to improve motivation, satisfaction and outcome. PM R 2014 Sep;6(9S):S170. [doi: 10.1016/j.pmrj.2014.08.374]
31. Li E, Clarke J, Ashrafian H, Darzi A, Neves AL. The impact of electronic health record interoperability on safety and quality of care in high-income countries: systematic review. J Med Internet Res 2022 Sep 15;24(9):e38144. [doi: 10.2196/38144] [Medline: 36107486]
32. Shull JG. Digital health and the state of interoperable electronic health records. JMIR Med Inform 2019 Nov 1;7(4):e12712. [doi: 10.2196/12712] [Medline: 31682583]
33. Pearce L, Costa N, Sherrington C, Hassett L. Implementation of digital health interventions in rehabilitation: a scoping review. Clin Rehabil 2023 Nov;37(11):1533-1551. [doi: 10.1177/02692155231172299] [Medline: 37132030]
34. Handley MA, Gorukanti A, Cattamanchi A. Strategies for implementing implementation science: a methodological overview. Emerg Med J 2016 Sep;33(9):660-664. [doi: 10.1136/emermed-2015-205461] [Medline: 26893401]
35. Shepherd V, Wood F, Griffith R, Sheehan M, Hood K. Protection by exclusion? The (lack of) inclusion of adults who lack capacity to consent to research in clinical trials in the UK. Trials 2019 Aug 5;20(1):474. [doi: 10.1186/s13063-019-3603-1] [Medline: 31382999]
36. Shepherd V. An under-represented and underserved population in trials: methodological, structural, and systemic barriers to the inclusion of adults lacking capacity to consent. Trials 2020 May 29;21(1):445. [doi: 10.1186/s13063-020-04406-y] [Medline: 32471488]
37. Gilbert T, Bosquet A, Thomas-Antérion C, Bonnefoy M, Le Saux O. Assessing capacity to consent for research in cognitively impaired older patients. Clin Interv Aging 2017;12:1553-1563. [doi: 10.2147/CIA.S141905] [Medline: 29026293]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
EPR: electronic patient record
MDT: multidisciplinary team
NHS: National Health Service
NIHR: National Institute for Health and Care Research
PPIE: patient and public involvement and engagement
TIDieR: Template for Intervention Description and Replication

XSL•FO
RenderX

Edited by A Stone; submitted 27.Jun.2025; peer-reviewed by CM Chen, C Subbe; accepted 11.Nov.2025; published 06.Jan.2026.

Please cite as:
Averill P, Lear R, Odedra R, Long S, Taylor A, Charville PJ, Fernandes J, Ekeogu U, Leombruno J, Ellis S, Mayer E
Acceptability, Feasibility, and Perceived Effectiveness of Video-Based Patient Records for Supporting Care Delivery to Older Adults With Frailty: Nonrandomized Mixed Methods Pilot Study
J Med Internet Res 2026;28:e77318
URL: https://www.jmir.org/2026/1/e77318
doi:10.2196/77318

© Phoebe Averill, Rachael Lear, Ricky Odedra, Susannah Long, Alex Taylor, Pi-Jung Charville, Jessica Fernandes, Uzoamaka Ekeogu, Jessica Leombruno, Sophia Ellis, Erik Mayer. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Psychotherapists' Trust, Distrust, and Generative AI Practices in Psychotherapy: Qualitative Study

Jinyan Kuang[1], BS; Amber L Pope[2], PhD; Yixuan Zhang[1], PhD

[1]Department of Computer Science, William & Mary, 251 Jamestown Road, Williamsburg, VA, United States

[2]Department of School Psychology and Counselor Education, William & Mary, Williamsburg, VA, United States

Corresponding Author:
Yixuan Zhang, PhD
Department of Computer Science, William & Mary, 251 Jamestown Road, Williamsburg, VA, United States

Abstract

Background: Generative artificial intelligence (GenAI) is increasingly used in mental health care, from client-facing chatbots to clinician-facing documentation aids. Psychotherapists' willingness to rely on—or withhold reliance from—these tools has significant implications for care quality, yet little is known about how practicing clinicians calibrate trust and distrust in GenAI across tasks and contexts. Given that the therapeutic relationship is central to psychotherapy outcomes, understanding how GenAI intersects with this relational foundation is essential for responsible integration.

Objective: This study aims to examine (1) psychotherapists' experiences with, perceptions of, and trust or distrust in GenAI in therapeutic contexts and (2) how they perceive the role of GenAI within the therapeutic relationship and how their perceptions shape their trust and distrust in GenAI.

Methods: We conducted a qualitative interview study using semistructured interviews with 18 actively practicing psychotherapists in the United States between January and May 2025. Participants were recruited through professional mailing lists, social media, and snowball sampling. Interviews (≈60 min each) were conducted via Zoom and explored psychotherapists' experiences with, perceptions of, and trust or distrust in GenAI in therapeutic contexts. Data were analyzed using the general inductive approach, with iterative coding and team-based interpretation to identify themes.

Results: Our findings show that psychotherapists' GenAI adoption was highly individualized and contingent on maintaining professional role integrity—not merely technical oversight. Trust was sustained when GenAI operated in clinician-supervised, supportive roles for low-stakes tasks (eg, documentation and brainstorming), but diminished when control shifted, tasks involved high-stakes clinical judgment, or GenAI threatened to encroach on the authentic human connection central to therapy. Participants articulated conditions for trust that went beyond "human-in-the-loop" monitoring to include preservation of interpretive authority, ethical responsibility, and relational primacy. Distrust also extended to the broader sociotechnical ecosystem, including concerns about commercial incentives, insurance pressures, and the absence of clear organizational guidelines.

Conclusions: Psychotherapists' perspectives offer critical insights into GenAI's current usages in their professional practices and the conditions under which they are willing to trust and distrust GenAI tools. Their experiences highlight the importance of maintaining clinician control, ensuring contextual appropriateness, and preserving the human connection central to psychotherapy. Future work should further examine how therapeutic orientation, professional experience, and client characteristics shape trust and distrust in GenAI. As GenAI becomes more embedded in mental health care, research is also needed to explore how specific GenAI system features can be responsibly designed to support clinical workflows and enhance therapeutic relationships. Organizational and policy frameworks will be essential to ensure responsible, ethically aligned, and human-centered GenAI deployment in psychotherapy.

(J Med Internet Res 2026;28:e88932) doi:10.2196/88932

KEYWORDS

artificial intelligence; AI; generative AI; mental health; trust; qualitative methods

Introduction

The rapid advancement of generative artificial intelligence (GenAI) in mental health care has sparked both enthusiasm and concern in recent years [1-4]. Earlier mental health artificial intelligence (AI) tools were largely rule-based or template-driven, offering structured guidance [5]. With recent advancements in large language models, professional organizations have begun to respond. For example, the American Psychological Association (APA) is developing guidance to help psychologists integrate AI tools into practice in ways that respect ethical and legal obligations [6]. Yet a recent survey found that 71% of mental health practitioners had not

XSL•FO
RenderX

incorporated AI-enabled tools into routine care [7]. Public and social-media discourse often portrays psychotherapists as broadly resistant to or distrustful of GenAI—casting the technology as incompatible with therapeutic work or as a deceptive substitute for human care, especially when chatbots impersonate psychotherapists [8]. These portrayals, however, say little about how practicing psychotherapists themselves are actively calibrating trust and distrust in GenAI in day-to-day clinical contexts.

Empirical work has begun to examine mental health professionals' attitudes toward AI and GenAI. Survey and interview studies have shown mental health practitioners' perceptions of AI in services [9-11], including views on the public's use of AI chatbots for mental health support [12]. This body of work has identified perceived benefits for self-management and accessibility alongside concerns about data privacy, clinical reliability, and limited empathy. Complementary design research has explored early GenAI-augmented prototypes (eg, counselor agents and multimodal assistants) to probe possible roles for GenAI in assessment, client participation, and skill practice [13]. Together, existing work illuminates general attitudes and speculative futures, but it offers less insight into how psychotherapists currently use GenAI in practice, which concrete tasks they deem trustable or untrustable, and how they respond when clients independently bring GenAI into therapy.

Research on trust in AI provides a useful lens for these questions. Trust in AI is often defined as a belief that a system is reliable, competent, transparent, and responsive in a given context [14], whereas distrust reflects expectations of harm, manipulation, or misuse [15]. A large body of work shows that trust in automation is situational and conditional: people are more willing to rely on AI when conditions around evidence, oversight, and control are satisfied, and they withdraw trust when those conditions are violated [16-21]. Trust is thus not a fixed trait but a dynamic, task- and role-dependent judgment that evolves with experience. In high-stakes domains, miscalibrated trust—either overtrust that leads to uncritical reliance or undertrust that prevents appropriate use—can produce safety risks and missed opportunities [22].

Decades of research have established that the therapeutic relationship—often termed the "therapeutic alliance" or "working alliance"—is among the most important predictors of psychotherapy outcomes, frequently accounting for more variance than specific techniques or theoretical orientations [23]. The alliance encompasses agreement on therapeutic goals, consensus on tasks, and the affective bond between therapist and client [24]. Such a relational foundation makes psychotherapy a domain where trust is not merely a technical consideration but a fundamentally interpersonal one. Psychotherapy is thus inherently relational and context-sensitive [25,26], making questions of calibrated trust and distrust in GenAI especially salient. Misplaced trust could, for example, lead psychotherapists to overdelegate clinical reasoning, accept flawed recommendations, or input sensitive client data into insecure systems, undermining both care and confidentiality [27,28]. Insufficient trust, by contrast, may keep psychotherapists from using GenAI in low-risk areas where it could meaningfully augment care (eg, psychoeducation, homework support, or documentation) [29]. At the same time, decades of work on the therapeutic relationship underscores that the alliance between psychotherapist and client—the goals, tasks, and emotional bond of therapy—is central to outcomes [30,31]. Experiences with telehealth show that this relationship can adapt to new media but requires careful management of shifts in framing and presence [30]. As GenAI enters psychotherapy, the traditional dyadic relationship becomes a triad involving psychotherapist, client, and GenAI [32], raising open questions about GenAI's appropriate role, boundaries, and influence on trust among all 3 parties.

Despite growing attention to GenAI in mental health, we know relatively little about how practicing psychotherapists form, calibrate, and revise trust and distrust in GenAI across tasks and contexts, or how they understand GenAI's place in the evolving psychotherapist-client-GenAI relationship. In this work, we conducted a qualitative study with 18 psychotherapists to explore (1) What are psychotherapists' experiences with, perceptions of, and trust or distrust in GenAI in therapeutic contexts? and (2) How do psychotherapists perceive the role of GenAI within the therapeutic relationship, and how do these perceptions shape their trust and distrust in GenAI?

Methods

Study Design

This study used a qualitative interview method to explore how psychotherapists perceive and use GenAI tools in their professional practice, with particular attention to the dynamics of trust and distrust. A qualitative approach was selected because it enables an in-depth understanding of psychotherapists' lived experiences and contextualized decision-making around GenAI integration [33]. To do so, we conducted semistructured interviews via Zoom (Zoom Video Communications, Inc) between January and May 2025 and qualitative data analysis between May and July 2025 to further examine psychotherapists' practices, trust (or distrust) of GenAI, and their envisioned relationship between GenAI and psychotherapists.

Recruitment

Concerning the inclusion criteria, participants were eligible if they (1) held a valid license or credential to practice psychotherapy in the United States (eg, licensed professional counselor, licensed clinical social worker, and licensed psychologist), and (2) were actively providing psychotherapy to clients at the time of enrollment. No exclusion criteria were applied regarding therapeutic orientation, practice setting, years of experience, or prior familiarity with GenAI; we sought a diverse range of perspectives rather than a homogeneous sample.

Concerning participant recruitment, they were recruited through purposive and snowball sampling between January and May 2025. Recruitment advertisements were distributed via (1) professional mailing lists for psychotherapists and mental health practitioners, (2) LinkedIn posts targeting mental health professionals, and (3) professional networks of the research team. The recruitment message briefly described this study's

XSL•FO
RenderX

purpose (exploring psychotherapists' perspectives on GenAI), eligibility criteria, time commitment (≈60 min), and compensation. Interested individuals contacted the research team via email and were screened for eligibility. We received 20 expressions of interest; 18 completed interviews, and 2 did not attend their scheduled sessions (no follow-up information was available regarding reasons for nonattendance).

Interview Study Procedures

The semistructured interview guide was developed iteratively by the research team, drawing on (1) the existing literature on trust in AI and automation [21,22,34], (2) prior qualitative work on mental health practitioners' attitudes toward AI [9,12], and (3) the research team's clinical and human-computer interaction expertise. The guide was pilot-tested with 2 psychotherapists (not included in the final sample) to assess question clarity and flow, and minor revisions were made based on their feedback. The final guide included open-ended questions organized around three domains: (1) current experiences with GenAI in clinical practice, (2) perceptions of trust and distrust in GenAI across tasks and contexts, and (3) views on GenAI's role in the therapeutic relationship. See Multimedia Appendix 1 for the complete interview protocol.

Before each interview, we asked each participant about their experience with GenAI and demographic information. If participants indicated previous GenAI experience, they were then prompted to elaborate on their experiences with GenAI use, followed by questions about their attitudes and trust or distrust of GenAI. When a participant had no prior GenAI experience, we first shared a web link from the APA [35] that outlines the potential clinical uses of GenAI and the associated ethical guidelines. The discussion then focused on the participants' reactions and their anticipated levels of trust or distrust. In each interview, we also asked about views on the independent use of GenAI tools by clients and their perceptions of the relationship among the psychotherapist, GenAI, and client. The interview sessions lasted approximately 60 minutes. See Multimedia Appendix 1 for the detailed interview questions.

Data Analysis

All interview sessions were transcribed verbatim and checked for accuracy against the original recordings. We conducted a qualitative analysis with the transcripts guided by the general inductive approach [36]. First, the lead author carefully read the transcripts multiple times to become familiar with the data and the whole brief analytic memos to capture initial impressions about psychotherapists' GenAI practices, perceived benefits and risks, and expressions of trust and distrust of GenAI in different clinical contexts.

In the first cycle of coding, the lead author conducted line-by-line coding, assigning low-level, descriptive codes to segments of text that captured (1) concrete uses of GenAI in practice including both psychotherapists' practices (eg, documentation, homework support, and supervision) and their observations of their clients' usages (eg, journaling and emotional support), (2) conditions under which psychotherapists felt trust or distrust in using GenAI, and (3) how they perceived GenAI's role in the psychotherapist-client relationship. Codes were kept close to participants' language and iteratively refined as additional transcripts were reviewed. A provisional code list was updated after every 2 - 3 transcripts to merge redundant codes, clarify labels, and add new codes that reflect concepts not captured previously.

Next, low-level codes were grouped into more abstract categories that summarized patterned meanings across participants and connected related codes (eg, "emergence of trust in GenAI for low-stakes administrative tasks," "determining trust or distrust based on boundaries around high-stakes clinical reasoning," "enhancing trust in guided client use of GenAI between sessions," and "distrust about data privacy and commercialization"). Through constant comparison within and across transcripts, these categories were iteratively collapsed and reorganized into higher-level themes that described (1) psychotherapists' current GenAI engagement, (2) how professional autonomy and task type set boundaries for trust or distrust, and (3) how GenAI is positioned within the therapeutic relationship, which aligns with their trust or distrust.

Throughout the coding process, 2 additional members of the research team independently coded overlapping subsets of transcripts using the evolving codebook and then met with the lead author in weekly analytic meetings to compare interpretations. Discrepancies in coding were discussed and resolved through group discussion until consensus was reached, and the shared codebook was iteratively updated to reflect agreed-upon definitions and illustrative coding examples. As analysis progressed, the team regularly compared new transcripts against the existing codebook to ensure consistent application of codes while remaining open to emergent concepts. Data collection and analysis proceeded concurrently and continued until thematic saturation was reached, defined as the point at which no substantively new codes or themes emerged in successive interviews and additional transcripts primarily confirmed the existing thematic structure. This study adhered to the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist [37]. See Checklist 1 for the detailed checklist.

Ethical Considerations

Ethical approval for this study was obtained from the Institutional Review Board at William & Mary (PHSC-2024-06-30-17113). Before participation, the research team explained this study's purpose, procedures, potential risks, and benefits to all participants. Written informed consent was obtained via participants' digital signatures before each interview. Each participant was assigned a unique study ID upon enrollment, and all identifying information was removed to ensure confidentiality. At the end of this study, each participant received a US $20 gift card as compensation upon completing the interview.

Data Availability

The qualitative data (interview transcripts) generated during this study are not publicly available due to privacy and confidentiality concerns; transcripts contain sensitive information that could identify participants despite deidentification efforts. The interview guide is provided in

XSL•FO
RenderX

Multimedia Appendix 1. Requests for access to deidentified data excerpts or the codebook may be directed to the corresponding author and will be considered on a case-by-case basis, subject to institutional review board approval and establishment of a data use agreement.

Results

Overview

Psychotherapists in our interview study (see a summary of participants' demographics information in Table 1) reported experimenting with GenAI mainly in supporting, nonclient-facing tasks, using it as an assistant rather than a psychotherapist substitute (see a summary of usage contexts in Table 2). Psychotherapists described using GenAI in several practical ways. Some used GenAI to draft session notes from generic prompts (eg, "therapy note, client struggling in marriage"). Others explored GenAI-driven homework support to assist clients between sessions or used GenAI as a communication bridge when clients' primary language differed from their own, allowing bilingual or multilingual clients to better articulate their thoughts and improving access for those who cannot easily find language-matched providers. Psychotherapists also applied GenAI for psychoeducation, using it to generate "simulated client" scenarios and offering more realistic and diverse practice opportunities than classmates. Participants valued GenAI's potential to enhance efficiency, extend therapeutic reach, and diversify resources—yet they consistently paired these ideas with caveats about trust, control, and ethical boundaries, setting the stage for the next section.

Table . Overview of participants' demographics information.

Dimension and response options	Interview participants (n=18), n
Working experiences[a] (years) [38]	
1 - 10	14
11 - 20	2
21 - 30	1
31 or more	1
Gender	
Woman	14
Man	4
Nonbinary	N/A[b]
Race or ethnicity	
Asian	2
Black or African American	3
White	13
Prefer to self-describe	1

[a]Mean 9.7, SD 8.5 years; median 7.5, IQR 13 years.

[b]N/A: not applicable.

XSL•FO
RenderX

Table . Contexts and example quotes of GenAI[a] use among psychotherapists.

Usage contexts	PIDs[b]	Example quotes of participants' GenAI use
Administrative and clinical support		
Clinical documentation	P02, P04, P07, P11	"I get [clients'] consent and then it records the session; from that transcript it will generate my note, and I edit that note and put it into my EHR."
Translation and language	P01, P05, P16	"I had ChatGPT translate a breathing exercise into Spanish for one of my clients."
Treatment plan formulation	P11, P14, P17	"I've used GenAI the most for wording a treatment plan goal."
Personalized homework creation	P01, P02, P03, P05, P09, P13	"I asked ChatGPT to make me a worksheet [that contains] information from books that I think provide a good model and explain it in layman's terms versus clinical terms and ask ChatGPT to make reflection questions for my client."
Training and supervision		
Ideation and brainstorming	P03, P06, P13, P18	"I need kind of a brainstorming colleague thing that I have [Gen]AI help me create content, like I might say, please give me 3 paragraphs of a case conceptualization for a client who's experiencing 3 anxiety symptoms."
Roleplay scenarios and scripts	P01, P11, P15	"Give me a real textbook kind of description of a persona that I can use in my training materials."
Supervision and clinical consultation	P09	"I've even considered if I could run a case by a [Gen]AI to get a sort of second opinion, almost like an 'AI supervisor' giving me feedback."
Self-help and emotional support	P06, P11, P12	"I kind of used a GenAI [chatbot] for me personally, when I was going through a difficult time where I interacted with the GenAI to get support."

[a]GenAI: generative artificial intelligence.

[b]PID: participant ID.

Based on their practices in GenAI in therapeutic contexts, we further identified 2 key domains that shape their trust and distrust toward GenAI in therapeutic contexts. First, psychotherapists articulated conditional trust: trust in GenAI that was bounded by professional autonomy, task specificity, and careful oversight. Second, they reflected on GenAI's position within the therapeutic relationship, weighing scenarios where it might enhance client engagement against fears it could erode the authentic human bonds central to psychotherapy.

Theme 1: Professional Autonomy Sets Boundaries for Trust

Overview

Theme 1 addresses psychotherapists' trust in GenAI for their own professional use—that is, when they themselves use GenAI tools in clinical workflows. Our participants conveyed that their trust in GenAI was conditional—bounded by a sense of self-regulation, task-specific usage, and personal limits. For psychotherapists, the relevant conditions revolve around safeguarding the therapeutic frame, protecting client confidentiality, and preserving the clinician's professional agency. In practice, psychotherapists' trust in GenAI remains high only when they can maintain direct control over how the technology is used. When that control slips, such as using GenAI for unfamiliar tasks or allowing others to dictate its use, their trust quickly diminishes.

Subtheme 1: "Hand on the Wheel" Rule

Participants highlighted that their trust in GenAI systems did not stem from the technology itself, but from their perceived ability to manage and control [Gen]AI's involvement. Psychotherapists in our study articulated a personal "hand on the wheel" "boundary" for trust*:* trust arose primarily when they could directly influence the GenAI's actions through careful prompting, reviewing outputs, and maintaining options to delete or redact any generated content. In other words, psychotherapists' trust was grounded less in confidence in GenAI itself than in confidence in their own professional ability to control the system: their judgment, expertise, and responsibility. For example, 1 psychotherapist (P03) tied her trust to her ability to steer the GenAI with well-crafted prompts and usage limits:

> *I trust myself to be a good prompter... I know my boundaries and I know what I'm comfortable doing and not doing. And so as long as I stick within my*

XSL•FO
RenderX

> *boundaries, I feel pretty comfortable using [Gen]AI, kind of accomplishing the goals that I have.* [P03]

P03's quote illustrates that trust is contingent not just on the quality or performance of the GenAI system, but rather on the clinician's professional autonomy and self-assessment of their capability to manage interactions with GenAI appropriately. Crucially, this sense of control applied not only to what information she fed into the system, but also over the fate of the outputs: where the text was stored, how long it persisted, and whether it could be edited or deleted to preserve client confidentiality.

Another participant (P04) emphasized that trust was fragile and tied directly to control over sensitive clinical data. Describing the workflow of a GenAI note-taking tool, P04 explained that their trust hinged entirely upon the continued availability of features, such as data deletion and transcript management:

> *With Blueprint, I can delete the transcript, I can wipe it off of that system and that is within my control... and so that makes me trust it and I would be distrustful if they took that away... like that's the premium feature is to delete your transcript.* [P04]

P04's explanation suggests how removing or monetizing such control-related features would undermine his trust. His confidence in the tool thus rests on specific capabilities allowing direct oversight and editability. When those capabilities are guaranteed, he feels safe using the GenAI; if they are taken away, his trust would evaporate.

Subtheme 2: Task-Specific Boundaries of GenAI Trust

Beyond trusting their own abilities and control, psychotherapists' trust in GenAI is also highly task-dependent. Participants generally viewed GenAI as suitable for low-stakes tasks, such as brainstorming intervention ideas or drafting preliminary homework assignments, because mistakes in these contexts carried minimal risk and could easily be corrected by the psychotherapist. Conversely, their trust diminished when GenAI attempted high-stakes or core clinical tasks, such as generating official clinical documentation or nuanced clinical reasoning. Psychotherapists are skeptical that GenAI can handle sensitive clinical decisions or documentation independently without error. P06 articulated this conditional trust, expressing comfort using GenAI for brainstorming but strong reservations about automated note-taking:

> *I'm just not sure that I want a system recording a session and taking a note for me, because I don't fully understand how that information is being gathered and stored, and I wouldn't want my client's information in a system that I didn't fully understand.* [P06]

Here, P06's skepticism arose from uncertainty over how the GenAI processed, recorded, and stored sensitive session data. Allowing a GenAI to record therapy sessions and generate notes raises serious privacy and data security concerns. Even if the GenAI tools could technically produce a decent therapy note, the lack of transparency about data handling was enough to erode her trust. Similarly, P03 refused to use GenAI for core clinical reasoning about a client:

> *I don't ask ChatGPT to conceptualize my clients when it comes to trying to understand what's wrong or what's going wrong; I rely on people for that.* [P03]

In psychotherapy, "conceptualization" refers to the process by which psychotherapists develop a comprehensive understanding of a client's presenting issues, symptoms, underlying causes, and contextual factors [39]. It involves synthesizing information from assessments, clinical observations, and theoretical frameworks to inform diagnosis, treatment planning, and interventions. P03's quote highlights that psychotherapists reserve essential tasks, such as case conceptualization, for their clinical judgment or consultation with human colleagues. P03 trusted GenAI as a tool to improve efficiency in peripheral tasks, but that trust stopped when the technology would encroach on areas requiring deep expertise, intuition, or accountability. In her view, having GenAI substitute for a psychotherapist's interpretive or diagnostic role was beyond acceptable bounds—it challenged the clinician's professional role and could compromise care.

In short, psychotherapists calibrate their trust in GenAI to the stakes of the task: routine or creative adjunct tasks are within bounds, whereas high-stakes therapeutic decisions are off-limits to GenAI.

Subtheme 3: Conditional Trust in Others' Use of GenAI

Another important boundary on psychotherapists' trust emerges when considering other people's use of GenAI. Our participants trusted themselves to use GenAI in a disciplined, responsible manner, but they withheld that trust from clients and even from less-experienced colleagues. Such disparity is rooted in psychotherapists' confidence in their training, ethical standards, and understanding of GenAI's limitations. That confidence drops when they imagine someone without similar expertise using GenAI in the therapy context. As P06 noted, even her high personal trust "probably drops if clients start to use [Gen]AI more and it impedes or interferes with the therapy." Psychotherapists worried that others might not show the same caution or insight, raising the specter of misuse or overreliance on GenAI. P13 illustrated this concern by contrasting her own careful approach with the unpredictability of a client's approach:

> *I feel like because I've done my due diligence, because I know I have a technology background and know how to keep my own information confidential, I feel confident, using the [Gen]AI, the ways that I do. Personally, however, I don't have that confidence that a client would be careful of what they share with a [Gen]AI.* [P13]

P13's concern suggests that psychotherapists feel they could control the flow of sensitive information when they themselves use GenAI, but their trust falters when that control shifts to clients. A client might unwittingly share private details with a GenAI without fully understanding the underlying risks, potentially breaching confidentiality. In the absence of the psychotherapist's oversight, information security cannot be guaranteed, driving psychotherapists' distrust of clients' unsupervised GenAI use.

XSL•FO
RenderX

Psychotherapists also fear that both clients and junior clinicians could become overreliant on GenAI in ways that harm therapy or professional development. Half of our participants worried that clients might overimmerse themselves in GenAI-generated advice or content, losing contact with reality or bypassing the therapeutic process. Similar concerns apply to junior psychotherapists regarding excessive reliance on GenAI for clinical tasks (eg, note-taking and treatment plans), thereby missing out on learning critical skills and clinical reasoning. P04, an educator with positive personal experiences with GenAI-assisted note-taking, articulated concerns about the trainees:

> *As an educator, I do have concerns...students still in school or early psychotherapists or counselors early in their training, early in their residency, not knowing how to create the notes.* [P04]

P04 emphasized the importance of trainees developing foundational skills without excessive GenAI assistance. If beginners shortcut core tasks such as writing progress notes by relying on GenAI, they risk eroding their own clinical agency and expertise. Overreliance on GenAI at early career stages could produce future clinicians who lack essential skills in documentation, observation, and critical thinking.

Despite these concerns about others' use of GenAI, psychotherapists did identify certain conditions that could increase their trust when clients or junior colleagues use GenAI. They did not categorically oppose others using GenAI; rather, they advocated for supervision and education to ensure safe and appropriate use. For clients, many psychotherapists suggested integrating a form of guided GenAI usage into therapy itself, essentially treating a client's GenAI use as something to be discussed and managed within sessions. For example, P05 likened it to how psychotherapists already addressed clients' social media habits:

> *(It's) helping clients navigate the use of [Gen]AI for positive means…just like we have to help clients navigate the use of social media in a positive way. That could be a part of the conversation—an intake: Do you use Snapchat? How often do you use [Gen]AI? What do you use [Gen]AI for? Do you use it to shop? Do you use it to write papers?* [P05]

P05 provided a practical framework for supervising client GenAI use: openly ask about and discuss it during therapy (even as early as the intake interview), just as one would assess other aspects of a client's digital life. Such a proactive approach could help clients use GenAI in a healthy, privacy-conscious manner. It also gives the psychotherapist insight and oversight, which in turn bolsters their trust in how the client is engaging with GenAI outside of sessions. That said, participants acknowledged that not all clients required the same level of monitoring. P12 pointed out that some client populations—she gave the example of first responders—come in with high digital security awareness and a critical stance toward technology. These clients might actually self-regulate their GenAI use responsibly, sometimes even outpacing the psychotherapist's knowledge of the tool:

> *I trust my first-responder clients. They research things on their own, know how to ask questions safely, and decide for themselves whether to rely on a [Gen]AI suggestion. Honestly, by the time they get to session, they sometimes know more about the tool's data flow than I do.* [P12]

P12's experience highlights how individual client characteristics, such as professional background, technology literacy, and skepticism, can increase a psychotherapist's trust in that client's independent GenAI use. This contrasts with other clients, indicating that psychotherapists' trust in others' GenAI use is not one-size-fits-all; it must be calibrated to the person's abilities and context.

Overall, psychotherapists remain cautious about integrating GenAI into clinical work, even as they acknowledge potential benefits in efficiency and support. Lingering concerns about risks keep their distrust ever-present in the background. In short, we found that psychotherapists' trust in GenAI is highly conditional—closely tied to maintaining personal control, selecting appropriate tasks, and ensuring proper contexts for GenAI use. These conditions matter because trust is inseparable from psychotherapists' professional role: safeguarding clinical judgment, ethical responsibility, and accountability for client outcomes [40] makes the question of what, how, and by whom GenAI is used central to their identity as psychotherapists.

Theme 2: Trust, Distrust, and GenAI in Therapeutic Relationships

Overview

Theme 2 shifts focus to the relational context, examining trust and distrust as they arise when clients use GenAI independently, when GenAI is positioned within the therapeutic relationship, and when systemic pressures shape the perceived role of GenAI in therapy. Psychotherapists in our study further negotiated their trust in GenAI against potential distrust, largely depending on how GenAI was positioned within the therapeutic relationship. On one hand, their observations of clients independently using GenAI for therapeutic tasks bolstered initial trust and confidence in GenAI tools. On the other hand, that trust became fragile and often flipped to skepticism when psychotherapists imagined GenAI encroaching on the human bonds that underlie effective therapy. These tensions shaped how psychotherapists envisioned GenAI's future role as either aligning with their professional values or conflicting with them.

Subtheme 1: Psychotherapists' Attitudes in Client-Led GenAI Use as a Complement to the Therapeutic Alliance

Many psychotherapists expressed cautious trust in GenAI based on seeing their clients incorporate tools such as ChatGPT into therapy-related activities in situations where their clients already engaged in active therapy, and psychotherapists could integrate and guide GenAI use within the therapeutic frame. Clients had used GenAI for tasks such as journaling and homework (P01), conversational practice or advice (P06), and organizing thoughts or self-diagnosing symptoms (P13). Psychotherapists generally viewed these independent uses of GenAI by clients as beneficial and not as a threat to their own role. Instead, they described GenAI as a complementary tool that could enhance openness, confidence, and even outcomes in therapy. For example, P01 noted that "if GPT definitely helps you with journal homework,

XSL•FO
RenderX

why not?", reflecting a practical trust in GenAI's assistance with therapeutic homework. Participants recognized that when clients felt more comfortable or organized in sharing their experiences (with a bit of GenAI help), it could strengthen the therapy process rather than weakening it.

Several participants gave examples of how client-driven GenAI use had positive ripple effects in therapy. P06, for example, described a case where a client sought advice from ChatGPT after an argument with his roommate. ChatGPT suggested the client should apologize to the roommate, providing direct feedback that the client might have resisted if it came from a human psychotherapist.

> *If I challenged a client and said in this scenario, it seems like maybe you need to apologize to your roommate, right? And even if that is true, because the client has done something that is not great, the client may be very mad at me. [However], if [Gen]AI has to offer this idea of challenge, as the client, where does that anger go? It's just a computer, right? It doesn't have emotion. And so I think maybe [Gen]AI could deliver more direct and challenging feedback to a client without risk of a therapeutic rupture than a person.* [P06]

Here, our participants saw GenAI as capable of delivering challenging feedback without risking a rupture in the therapeutic relationship. The client could accept the tough suggestion (to apologize) because it came from an impersonal source, thus preserving the client's openness with the psychotherapist. Importantly, participants described GenAI not as a substitute for therapeutic challenge or confrontation, but as a relational intermediary that could support client reflection and progress while keeping the therapeutic relationship intact. In P06's view, GenAI might occasionally say "what needs to be said" in a way that a human psychotherapist strategically might avoid to protect the client's feelings. Importantly, P06 noted this did not reduce the client's trust in their human psychotherapist; in fact, it helped the client make progress (repairing a relationship) while avoiding potential defensiveness toward the psychotherapist.

Another psychotherapist, P13, recounted how a client's use of ChatGPT became a bridge in the therapy session rather than a barrier. The client had chatted with ChatGPT about their symptoms and feelings before the appointment. P13 illustrated how clients' usage of GenAI could be a topic to start the conversations naturally, boosting client confidence and comfort to share their thoughts during sessions:

> *[ChatGPT] allows me to ask questions, like, "Oh, what did you [two] talk about?" or "How did it arrive at that conclusion?" They'll answer, "I started saying I'm having these symptoms — X, Y, Z..." ChatGPT helped them [clients] organize their thoughts and gave us a starting point. I think it made people feel more comfortable and confident... talking about their experience.* [P13]

In this scenario, the GenAI functioned as a conversation rehearsal or journaling tool, priming the client to share more openly in session. By externalizing some of the client's initial thoughts, ChatGPT lowered the barrier to disclosure, making it easier for the client to articulate experiences face-to-face with P13. The psychotherapist observed that this preparatory step reduced the client's anxiety and increased their willingness to explore their feelings in depth. Crucially, P13 did not feel supplanted by the GenAI; the psychotherapist's role remained central in interpreting, contextualizing, and building on what emerged from the client-GenAI interaction. This finding reflects how most participants viewed such GenAI usage as a supplementary aid that could facilitate communication and self-expression while preserving and potentially enhancing the core psychotherapist-client relationship.

Subtheme 2: Distrust Arises When GenAI Threatens Authentic Human Connection

Overview

While psychotherapists were positive about GenAI as a tool, their trust shifted to distrust when considering scenarios where GenAI might encroach on the therapeutic relationship itself. Participants grew skeptical and uneasy about GenAI in therapy whenever it seemed to mimic human empathy, form bonds with clients, or otherwise interfere with the formation of real human relationships. These speculative scenarios triggered strong resistance. Psychotherapists emphasized that effective psychotherapy depended on genuine human connection—often referred to as therapeutic alliance or bond—which involved authentic trust, empathy, and reciprocal emotional engagement. They identified GenAI's limitations in emotional authenticity, responsiveness, and embodied presence as serious barriers to maintaining this bond. In other words, even though GenAI can imitate empathetic language, it cannot provide the true interpersonal experience of being with another caring human. As discussed below, psychotherapists worried that clients forming illusory relationships with GenAI could be misleading or even harmful, and that a GenAI's lack of genuine emotion could undercut the healing power of empathy and trust in therapy.

Inauthentic ("Parasocial") Relationships With GenAI

Several psychotherapists drew analogies between a client chatting with a GenAI and the parasocial relationships people develop with media figures or celebrities. Parasocial relationships are one-sided bonds where a person feels a strong connection or intimacy with someone (often a public figure) who is not actually reciprocating that relationship [41]. P04 compared client-GenAI interactions to parasocial relationships, such as those people form with celebrities, highlighting their superficial and ultimately misleading nature:

> *A common example is you feel like you know your favorite celebrity, but you really don't. But you have an attachment to them. If they pass away, you are extremely sad. You feel that loss, you grieve it. You know, those are parasocial relationships. So developing something like that with your GenAI bot as, in a therapeutic realm, I think, is not appropriate.* [P04]

P04 highlighted the illusory nature of the bond: the client might feel cared for by the GenAI, but in reality, there is no mutual relationship—the GenAI does not "know" the client or genuinely

XSL•FO
RenderX

care, just as a celebrity does not know their grieving fan. P04 feared that a client might invest emotionally in a GenAI, only to be hurt upon realizing that this relationship was a 1-way projection. Such a realization could be emotionally devastating (analogous to grieving a celebrity's death) and might even undermine the client's trust in real relationships. The implication is that trusting a GenAI as if it were a person is fundamentally misguided, and psychotherapists felt a duty to discourage such attachments for the client's well-being. Similarly, participant P14 echoed concerns about the implications of normalizing uncertain relationships between clients and GenAI in therapeutic contexts, noting that:

> *I do think part of counseling is the relationship that is formed. And again, what does it mean to have a relationship with a non human and like? Is that something that we want or not? And is that something that should be encouraged or not?* [P14]

Both P04 and P14 were essentially asking: if clients start treating GenAI bots as psychotherapists or friends, do we risk eroding the value of real human-human therapeutic bonds? They emphasized their distrust toward the idea of GenAI-client "relationships," viewing them as shallow imitations that could mislead clients and ultimately prove harmful. Their perspective taps into broader debates on whether emotional bonds with anthropomorphized GenAI are healthy or not; indeed, psychologists have noted that while such bonds might feel real, they lack true reciprocity and can risk shaping unrealistic expectations about human relationships.

Lack of Genuine Emotional Bond and Empathy

Beyond the concept of parasocial attachment, psychotherapists also expressed concern that interacting with an emotionally simulated agent could stunt clients' ability or desire to form meaningful human connections. Participants noted that much of therapy's healing power stemmed from experiencing genuine empathy and coregulation with another person. GenAI, no matter how cleverly it reflected a client's words, fundamentally lacked human emotion, life experience, and a nervous system—it cannot truly "feel" or share in the client's pain and joy.

As P11 reflected on this with her experiences in talking with GenAI and feeling real comfort, she questioned, "do we want that to be where we go as a profession, they call it the healing arts." P11 was grappling with the ambivalence: a GenAI might provide momentary comfort (eg, through kind words at 2 AM when no one else is around), but if therapy were to shift toward GenAI, something essential would be lost. She noted that "therapy is not a science, is not formulaic, but is about people." Thus, shifting therapy to GenAI would strip away its essence as an art grounded in human connection, not a formula that can be executed. P11 even pondered whether making GenAI more human-like for our own convenience "somehow cheapens our experience of being human, or cheats us out of the chance to have real human connection." P11's sentiment indicated a sense of distrust: even if GenAI "works" on the surface to soothe someone, relying on it for emotional support might erode the value of human-to-human experiences.

Such interpersonal neural synchronization is only possible through face-to-face human interaction, involving elements such as eye contact, voice tone, and somatic empathy that a GenAI agent cannot easily reproduce. P9 touched on this when she explained that in the therapy room:

> *There are mirror neurons in both of our brains... When I take a deep breath, [the client] subconsciously takes a deep breath. When I model a healthy expression of emotion, their mirror neurons are doing the same… There's a neurological component of therapy that would be lost with GenAI.* [P09]

In highlighting this embodied coregulation, P09 was pointing to the subtle yet powerful ways a psychotherapist's presence helped a client feel safe and understood on a biological level. Participants widely felt that a GenAI—even one programmed to respond empathetically in text—could not replicate these rich, multichannel human signals of empathy. Thus, psychotherapists were unwilling to trust GenAI with the core work of forging an emotional bond.

P10 captured the risk that GenAI's unconditional supportiveness could create a comfortable bubble that ultimately harms client growth. She noted that "learning to tolerate not feeling good can often be part of psychotherapy." In real therapy, a good clinician sometimes needs to challenge clients' perceptions or allow them to sit with discomfort as part of healing. P10 wondered if a "GenAI-generated therapy" would instead become "more of a 'thumbs up' scenario, creating a loop of people hearing what they want to hear, and thus feeling good." If a GenAI companion always praises or agrees with the user (either by design or to avoid upsetting them), this always-supportive feature will give clients an illusion of feeling good temporarily and avoid difficult, necessary conversations with real people. Over time, this feedback loop might reduce tolerance for the messiness of real human relationships, and clients could stall in their therapeutic progress. As P10 put it, clients might "cut off real human connections" and bypass the harder work of connecting with others and facing real-world challenges if they grow too comfortable with a GenAI that never challenges them. Such an outcome would be the opposite of what therapy strives for, which is to help clients function better in their lives and relationships outside therapy.

Subtheme 3: From Personal Fear to Systemic Skepticism—Commercial Incentives Stoke Distrust

Another major theme in psychotherapists' discussions of trust was the tension between seeing GenAI as a helpful partner in therapy versus viewing it as a potential competitor or replacement. Such tension affected their willingness to trust technology. When GenAI was framed as augmenting their work—handling ancillary tasks, extending care between sessions, or providing additional resources—psychotherapists were generally open-minded and even optimistic. In those cases, GenAI aligned with their professional values, and they could trust it to support better outcomes. However, when the conversation turned to GenAI possibly replacing human psychotherapists, participants expressed strong distrust, often accompanied by anxiety and even a sense of threat to their professional identity.

XSL•FO
RenderX

We found that psychotherapists' fears evolved from that initial shock toward more pragmatic skepticism. They started scrutinizing real-world trends that might lead to GenAI-driven replacement of psychotherapists. One major factor was the way some technology companies are promoting mental health GenAI products. Participants noted that a number of startups have begun rolling out "therapy bots" or GenAI coach apps. Officially, these are often marketed as assistive tools, but psychotherapists suspected the subtext was to test the waters for full replacement. P13, for instance, mentioned seeing posts on LinkedIn about new chatbot services:

> *They're [mental health chatbot startups] trying to say, "Oh, this is going to be an assistant," but you can tell by the way they're working it that they're almost trying to say, hey, we can just do this therapy bot... put in all the information... and it's gonna replace therapy instead of having to talk to a therapist.* [P13]

The design and messaging of these products made P13 feel that the endgame for some developers is to bypass human psychotherapists altogether. This sense was reinforced by other participants (P11, P17) who observed how GenAI mental health platforms often come with bold claims or marketing hype. P17 warned that a slickly marketed GenAI could "scam potential clients into thinking they can get [the same] help" from a bot, when in reality that was not advised. Similarly, P11 worried that for-profit GenAI platforms might make "false promises" that lure people into viewing GenAI as nearly equivalent to a human psychotherapist. These comments suggest that psychotherapists' distrust of GenAI was not only about GenAI's capabilities, but also about the intentions and honesty of those deploying it. If GenAI is oversold as a reliable replacement for therapy, psychotherapists fear the public could be misled and that clients might forgo seeking real help, with potential harm.

Beyond the companies pushing GenAI, participants highlighted structural pressures that could drive GenAI-as-psychotherapist scenarios. A concern was the role of insurance and health care payers. In many health care systems, including mental health care, decisions about what services are covered (and at what rate) profoundly influence practice. Psychotherapists suspected that if GenAI therapy is seen as significantly cheaper and more efficient, insurers might encourage or even force a shift in that direction. P10 cynically remarked that "[Gen]AI methodology is going to be promoted by insurances [as] a cheaper, better, quicker fix." P13 similarly noted the temptation for efficiency:

> *From an insurance company side—you can be like, "Oh, well, a chatbot can talk to a thousand people a day." That sounds great for efficiency.* [P13]

Psychotherapists worried that they and their clients could be placed in a no-win situation, where financial incentives pressure everyone toward GenAI solutions regardless of actual therapeutic value. The mere possibility of this happening fueled a lot of distrust and even resentment toward GenAI in our interviews. It is not that these psychotherapists believe a GenAI could truly do the job as well as they can. In fact, they asserted it cannot, for all the human reasons discussed earlier, but the threat is that decisions outside their control (by companies or insurers) might force the issue. Whenever GenAI was envisioned as a substitute for a human psychotherapist, participants' trust evaporated. They viewed that narrative with distrust and, at times, with a sense of having to justify their own worth in the face of automation.

In summary, psychotherapists in our study navigated a delicate balance of trust and distrust toward GenAI in therapeutic settings. When GenAI supported them and their clients—acting as a friendly assistant that respects the primacy of human connection—they largely trusted it and saw promise. However, when GenAI appeared to usurp human roles or relationships, they responded with firm distrust, rooted in emotional intuition and professional reasoning. These findings highlight the importance of context and role definition in shaping trust: as the literature on automation suggests [21,42], people trust technology only when its use aligns with their goals and values and withdraw trust if it threatens or violates their goals. For psychotherapy, the inviolable goal is to promote genuine human healing and connection. GenAI's role in this domain, according to our participants, will depend on whether it is designed to honor and enhance, rather than redefine or replace, the human-centric nature of therapeutic relationships.

Across these findings, we see that psychotherapists' trust and distrust in GenAI are fundamentally about professional role identity. The introduction of GenAI into therapeutic contexts surfaces latent questions about what it means to be a psychotherapist, such as what tasks are core to the role, what knowledge and judgment are uniquely human, and where the boundaries of professional responsibility lie. When participants resisted GenAI involvement in case conceptualization, expressed discomfort with clients forming "relationships" with chatbots, or worried about insurance-driven replacement, they were not merely assessing GenAI's technical capabilities; they were articulating and defending the boundaries of their professional role. Such dynamics can be understood through role theory [42]. GenAI's entry into therapy creates potential role ambiguity (uncertainty about which tasks belong to the clinician vs the technology), role conflict (tension between efficiency pressures and relational values), and opportunities for role clarification (explicit articulation of what remains uniquely human). Participants' insistence on control, their task-specific boundaries, and their concerns about parasocial client-AI relationships all reflect efforts to resolve these tensions in favor of preserving the clinician's interpretive and relational primacy.

In this sense, GenAI serves as a clarifying lens: by confronting what GenAI can and cannot do, psychotherapists articulate more explicitly what they believe only a human clinician can provide—authentic empathy, embodied presence, ethical accountability, and the capacity to work through relational ruptures. These articulations contribute to ongoing professional discourse about the nature of psychotherapeutic expertise in an increasingly technological landscape.

XSL•FO
RenderX

Discussion

Principal Results

Our findings reveal that psychotherapists primarily used GenAI for low-risk, supportive tasks (eg, notes, brainstorming, and homework), while drawing firm boundaries around high-stakes clinical reasoning (eg, case formulation and diagnosis) and any opaque handling of client information. As a result, psychotherapists' trust in GenAI is highly context-dependent and rooted in maintaining human control and relationships. Finally, we concluded our findings as: psychotherapists' control as the locus of trust based on their current GenAI adoption, the management of the psychotherapist-client-GenAI triad to protect the therapeutic alliance, and the systemic pressures and fears of exclusion that shape implications for future policy and design.

First, our findings show that trust hinges on professional autonomy (eg, the "hand-on-the-wheel" rule) and concrete controls (eg, prompt scoping, editability, and data deletion) that allow psychotherapists to supervise what goes in and what comes out. Even when the same GenAI tool was technically capable of other functions, psychotherapists were reluctant to shift trust to unfamiliar uses outside their control. For psychotherapists, even a slight shift in control, or a perceived loss of it, can substantially reduce the psychotherapists' confidence. They integrate GenAI into their workflow only on their own terms, within the boundaries of their professional knowledge and practice.

Psychotherapists' insistence on concrete controls (eg, scoping GenAI prompts, editing GenAI-generated content, and ensuring data privacy) echoes findings in other professions that emphasize oversight in using GenAI [43-45]. For example, in domains such as finance and accounting, professionals also exhibit conditional trust in GenAI and demand human oversight to meet privacy, compliance, and accuracy requirements [46]. Indeed, many accounting firms (including large firms in the Big Four) have introduced GenAI to assist with tasks such as auditing and tax preparation while keeping a human auditor in the loop [47]. However, psychotherapists practice an even stricter form of conditional trust. Given the extraordinary sensitivity of therapeutic work, where mistakes can carry deeply personal or even harmful consequences for clients, trust in automation is "squeezed"—carefully rationed and contextually negotiated. Consistent with prior research on trust in high-risk settings [48,49], our participants' trust was conservative by necessity: they only trusted GenAI when its use aligned with their goals, values, and ethical duties, and they withdrew trust immediately if the GenAI's role threatened those principles.

However, maintaining "control" through careful prompting and boundary-setting does not necessarily guarantee actual control over how GenAI shapes interpretation. As algorithmic systems operate within preexisting biases and conversational frames, even skilled users may be subject to subtle influences that confirm rather than challenge their interpretive directions. The distinction between actual and perceived control is clinically important as a psychotherapist who believes they are steering the GenAI may, in fact, be receiving outputs shaped by the model's training patterns in ways that reinforce existing assumptions. Such a gap between perceived and actual control interacts with professional role identity. Psychotherapists' expertise traditionally includes recognizing when their own biases may be influencing clinical judgment—a form of reflective practice. GenAI complicates this by introducing an interlocutor whose biases are opaque and whose "agreement" may feel like validation. Future research should examine whether and how clinicians can develop literacy about these dynamics to engage with GenAI more reflexively.

Taken together, our findings reveal that psychotherapists' conditions for trust extend beyond retaining veto power or technical oversight. Trust is conditional on maintaining professional role identity—not merely having a "hand on the wheel," but preserving the clinician's interpretive authority, ethical responsibility, and relational primacy. When GenAI threatens to shift these role boundaries, even if the clinician technically retains control, trust erodes. In relationally intensive professions such as psychotherapy, "human-in-the-loop" is necessary but not sufficient; the human must remain in the loop as a clinician with a defined professional role, not merely as a monitor or editor of outputs. Such a role-identity framing could help explain why participants resisted GenAI for case conceptualization (a task central to clinical identity) while accepting it for documentation (a task peripheral to their core professional self-concept).

Second, our findings add nuance to understanding the relationship among psychotherapists, clients, and GenAI by examining counselors' observations of (and hopes for) client-led use of GenAI in therapy. Participants generally supported GenAI as a supplementary aid for clients, but only when it complemented, not competed with, the human relationship. Trust was maintained when a client's use of GenAI enriched the therapy process under the psychotherapist's guidance. For example, some clients had experimented with GenAI tools for personal journaling, mood tracking, or skills practice between sessions, and psychotherapists viewed these as positive and adjunctive activities as long as they could review and integrate the GenAI-generated content into treatment. To participants, the GenAI tool functioned as a bridge—facilitating communication and self-reflection—while the core psychotherapist-client bond was preserved or even strengthened. These scenarios illustrate how GenAI can serve as a therapeutic supplement that empowers clients (by providing on-demand exercises or feedback) without undermining the clinician's authority or the authenticity of the human connection.

While psychotherapists expressed openness to client-led uses of GenAI, as reflected in remarks such as "why not," participants stressed that acceptance and these benefits only hold under specific conditions, when clients are actively engaged in therapy, and the clinician can oversee the GenAI's role. Unsupervised or excessive reliance on GenAI by clients raised ethical red flags for many practitioners. In psychotherapists' view, a GenAI that provided easy answers or constant validation might become an emotional crutch, potentially impeding clients' growth by sidestepping the discomfort and vulnerability through which therapeutic change occurs. Several psychotherapists noted the risk of clients developing an illusory relationship with a chatbot that mimics empathy but lacks genuine human understanding.

XSL•FO
RenderX

They voiced concern that unvetted advice or faux-empathy from GenAI could misguide clients or undercut the healing power of authentic human empathy in therapy. Thus, even the same tools that help organize thoughts between sessions were seen as double-edged: without guardrails, those tools might shape clients' thinking in ways the psychotherapist cannot monitor or correct. This highlights the importance of embedding GenAI use within ongoing professional support, rather than allowing it to operate as a parallel "psychotherapist."

Third, systemic forces, such as insurance reimbursement models, health care policies, and technology marketing, further complicate the psychotherapist-client-GenAI triad. Our findings reveal that psychotherapists' fear of being "replaced" by GenAI stems less from its raw capabilities than from institutional pressures. Many participants expressed concern that decisions by insurers, employers, or health care systems might push GenAI solutions into therapy for the sake of cost or efficiency, sidelining human clinicians against their will. Indeed, signs of these systemic shifts are already visible. In the United States, Medicare announced that as of January 2025, it will reimburse providers for prescribing certain US Food and Drug Administration–cleared digital mental health therapeutics [50]. Private insurers such as Aetna and Cigna have also begun partnering with digital mental health platforms and covering app-based therapeutic services. These developments reflect a growing institutional embrace of GenAI-driven and digital tools to expand access and reduce costs. However, psychotherapists in our study saw a potential dark side: exclusion from the care loop if stakeholders treat GenAI as a cheap substitute for human therapy. They worried about a "no-win situation" where financial incentives pressure clinics to replace portions of therapy with automated services, regardless of client benefit. Such a possibility fueled a great deal of distrust and even resentment toward GenAI among participants. Crucially, it was not that psychotherapists believed a GenAI could do their job just as well—on the contrary, they were confident in the unique value of human care—but rather that decisions outside their control might force the issue (eg, an insurance company mandating a chatbot as a first line of support). Whenever GenAI was imagined as a substitute for a human psychotherapist, participants' tentative trust evaporated, replaced by staunch distrust and a need to justify their professional worth.

Psychotherapists also noted mixed signals in the policy landscape that mirror this tension. On one hand, policymakers are cautiously enabling GenAI integration in mental health care through reimbursement and innovation incentives. On the other hand, there is growing recognition of the ethical and quality concerns: for example, some US jurisdictions are moving to outright ban certain uses of AI in psychotherapy. In August 2025, the state of Illinois enacted a law prohibiting the use of AI to provide psychotherapy or make clinical decisions [51], while still allowing AI for "administrative and supplementary support" under a provider's oversight. This law was explicitly framed as protecting patients from unregulated AI or GenAI tools and preserving the jobs of human psychotherapists. Such contradictory responses—from enthusiastic adoption to legal restriction—highlight a core societal debate: Should GenAI be embraced to improve mental health accessibility and efficiency, or curtailed to preserve the human core of therapy? Our participants overwhelmingly rejected either extreme position. The psychotherapists' stance was one of measured, principled integration: GenAI should be used only in ways that support—and never replace or diminish—the therapeutic relationship. Psychotherapists in our study are not "Luddites" resisting technology out of fear; instead, they are concerned practitioners seeking to ensure GenAI is implemented responsibly and relationally. Their distrust is not a reflexive fear of job loss per se, but a deeper worry about clients being left without meaningful human connection if economic structures push GenAI-only solutions as a shortcut. In fact, many envisioned constructive roles for GenAI (some called it a potential "glue" in therapy) that could enhance rapport, support client progress, and deepen engagement, but only if these tools are introduced thoughtfully, with ethical guardrails and human oversight at every step.

Limitations and Future Implications

A limitation of this study is that our interview participants may represent a group of psychotherapists with greater initial interest in GenAI than the average practitioner, as they voluntarily opted in through our recruitment emails to complete the interviews. Their perspectives may therefore be more exploratory or engaged compared with those of clinicians who chose not to participate. Our participants were also US-based and demographically skewed (predominantly White, with a relatively larger proportion of women). We did not systematically examine whether perspectives varied across demographic characteristics (eg, gender, race or ethnicity, and years of practice). As a result, the trust and distrust dynamics reported here may not fully reflect the diversity of clinician perspectives and may transfer differently across countries, health care systems, and regulatory contexts. Additionally, we did not assess participants' primary therapeutic orientation (eg, cognitive-behavioral, psychodynamic, humanistic, and integrative). In the US context, some licensed psychotherapists practice eclectically or identify with integrative approaches, but orientation may nonetheless shape attitudes toward structured tools and the conceptualization of the therapeutic relationship. For example, practitioners with a strong humanistic or relational orientation may be more skeptical of GenAI's capacity to support authentic connection, while those with cognitive-behavioral orientations may be more receptive to GenAI-assisted psychoeducation or homework. Future research should explicitly examine whether psychotherapists from different schools (eg, the psychodynamic, humanistic, and cognitive-behavioral) differ in their trust calibration and in the conditions under which they would accept GenAI involvement in therapy. Finally, the rapid pace of GenAI development means that our findings reflect psychotherapists' perceptions at a particular moment in time. As GenAI tools continue to evolve in capability, regulation, and clinical integration, psychotherapists' trust, concerns, and usage practices may shift in ways not fully captured here.

Despite these limitations, the findings remain highly relevant for clinicians and designers seeking to understand how psychotherapists use GenAI in real clinical contexts, how they negotiate trust and distrust under different conditions, and how they navigate clients' independent use of GenAI. The themes

XSL•FO
RenderX

identified in this study can guide design considerations, inform professional training needs, and highlight the types of support clinicians require when confronted with GenAI in real-world practice.

Future GenAI tools for psychotherapy should be designed to enhance clinician oversight and preserve clinical judgment, rather than shifting interpretive authority to the system. Our findings suggest "minimum conditions" for trust and adoption, including clinician-facing controls to review outputs before use, edit or annotate content to reflect clinical nuance, and delete materials that are inaccurate or clinically unhelpful. Trust also depends on transparency that helps clinicians understand what the model is responding to (eg, visible inputs and omissions) and recognize overconfident or overly generic guidance. Thus, clear data governance, such as storage, retention, access, and use for model improvement, is also essential for informed privacy decisions.

In addition, future researchers could also explore how GenAI can strengthen communication between psychotherapists and clients. Participants described how "GenAI in the loop" could support communication by offering direct or challenging prompts, helping clients organize thoughts, or assisting them in articulating emotions more clearly. Developers may consider GenAI features that act as a bridge between clients and psychotherapists—for example, translating clinical terminology into laypeople's language or generating structured summaries that help psychotherapists quickly understand clients' concerns. Such features could enhance clarity and mutual understanding while preserving the psychotherapist's primary role in treatment.

Participants also highlighted opportunities for GenAI-supported psychoeducation and training. As GenAI becomes increasingly common in clients' daily lives and in clinical workflows, trainees may benefit from working with GenAI early to learn ethical and appropriate use. Interactive GenAI-simulated patients could also provide customizable training scenarios with different patient backgrounds, symptom profiles, or emotional intensities that are not always available in traditional role-plays or standardized-patient exercises. These adaptive environments may help trainees build practical skills and prepare for future clients who already integrate GenAI into their mental health routines.

Furthermore, the tension between systemic pressures and therapeutic values points to future implications for policy and governance. There is a clear need for professional guidelines and regulations that reaffirm the centrality of human psychotherapists in any GenAI-augmented care model. Organizations such as the APA have begun developing guidance to help psychotherapists navigate GenAI in practice, and our findings underscore how crucial such efforts are. Policies must strike a balance so that GenAI can be leveraged to extend care (eg, through approved digital therapeutics) without eroding standards of care or marginalizing clinicians. Psychotherapists may also require training to work effectively alongside GenAI tools and to advocate for their patients' needs in dialogues with employers or insurers deploying these tools. Equipping psychotherapists with digital literacy and ethical decision-making frameworks for GenAI can help them retain a sense of agency and protect client welfare when new technology is introduced.

Conclusions

In this work, we conducted an interview study with 18 psychotherapists to explore how psychotherapists perceive, trust (or distrust), and incorporate GenAI into their everyday practice. We found that psychotherapists exhibit a cautious, conditional trust in GenAI: they embrace these tools for administrative and low-risk tasks (eg, note-taking, training simulations, and between-session client exercises) but are wary of using GenAI for core clinical decisions or any role that might infringe upon the therapeutic relationship. By recognizing the conditions under which psychotherapists are willing to trust GenAI, we can inform the development of tools and policies that empower psychotherapists and safeguard clients. Moving forward, it will be crucial to pursue interdisciplinary research and dialogue that continue to balance innovation with empathy, ensuring that future GenAI in mental health strengthens the art of therapy instead of undermining it.

Acknowledgments

During the preparation of this work, the authors used ChatGPT-5.0 (OpenAI) solely to check the grammatical accuracy of the written text. All suggested edits were reviewed by the authors, who take full responsibility for this final paper.

Funding

This work received support from the Charles Center at William & Mary. The funder had no role in study design, data collection, analysis, interpretation, or paper preparation.

Authors' Contributions

Conceptualization: JK, ALP, YZ
Data curation: JK
Formal analysis: JK
Investigation: JK, AP
Methodology: JK, YZ
Project administration: JK
Resources: AP

XSL•FO
RenderX

Supervision: YZ
Writing – original draft: JK
Writing – review & editing: YZ

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide used to explore psychotherapists' experiences, trust and distrust, and perceived roles of GenAI in practice. GenAI: generative artificial intelligence.
[DOCX File, 1940 KB - jmir_v28i1e88932_app1.docx]

Checklist 1
COREQ 32-item checklist.
[PDF File, 134 KB - jmir_v28i1e88932_app2.pdf]

References

1. Couture V, Roy MC, Dez E, Laperle S, Bélisle-Pipon JC. Ethical implications of artificial intelligence in population health and the public's role in its governance: perspectives from a citizen and expert panel. J Med Internet Res 2023 Apr 27;25:e44357. [doi: 10.2196/44357] [Medline: 37104026]
2. Grassini S, Koivisto M. Understanding how personality traits, experiences, and attitudes shape negative bias toward AI-generated artworks. Sci Rep 2024 Feb 19;14(1):4113. [doi: 10.1038/s41598-024-54294-4] [Medline: 38374175]
3. Lee J, Jung K, Newman EG, Chow E, Chen Y. Understanding adolescents' perceptions of benefits and risks in health AI technologies through design fiction. CHI 2025 2025 Apr 26:1-20. [doi: 10.1145/3706598.3713244]
4. Liu Z, Zou W, Lin C. Exploring the influence of privacy concerns, AI literacy, and perceived health stigma on AI chatbot use in healthcare: an uncertainty reduction approach. Patient Educ Couns 2025 Nov;140:109271. [doi: 10.1016/j.pec.2025.109271] [Medline: 40712403]
5. Luo X, Li Z, Yang Z, et al. Using large language models to assess the consistency of randomized controlled trials on AI interventions with CONSORT-AI: cross-sectional survey. J Med Internet Res 2025 Sep 26;27:e72412. [doi: 10.2196/72412] [Medline: 41004321]
6. Abrams Z. Artificial intelligence is reshaping how psychologists work. American Psychological Association Services, Inc. 2025 May 12. URL: https://www.apaservices.org/practice/news/artificial-intelligence-psychologists-work [accessed 2026-03-18]
7. Page C, Assefa M, Stamm K. What psychologists are saying about using AI in practice. Monitor Psychol 2025 Jul 1;56(5) [FREE Full text]
8. Yıldız HA, Söğütdelen E. AI chatbots as sources of STD information: a study on reliability and readability. J Med Syst 2025 Apr 3;49(1):43. [doi: 10.1007/s10916-025-02178-z] [Medline: 40178771]
9. Báez JC, Victor BG, Dysart C, Goldkind L. "I don't understand it, but okay": an empirical study of mental health practitioners' readiness to use large language models. J Technol Hum Serv 2025 Apr 3;43(2):142-160. [doi: 10.1080/15228835.2025.2479475]
10. Zhang M, Scandiffio J, Younus S, et al. The adoption of AI in mental health care-perspectives from mental health professionals: qualitative descriptive study. JMIR Form Res 2023 Dec 7;7:e47847. [doi: 10.2196/47847] [Medline: 38060307]
11. Gültekin M, Şahin M. The use of artificial intelligence in mental health services in Turkey: what do mental health professionals think? Cyberpsychology (Brno) 2024;18(1):6. [doi: 10.5817/CP2024-1-6]
12. Sweeney C, Potts C, Ennis E, et al. Can chatbots help support a person's mental health? Perceptions and views from mental healthcare professionals and experts. ACM Trans Comput Healthcare 2021 Jul 31;2(3):1-15. [doi: 10.1145/3453175]
13. Tang Y, Kang Y, Wang Y, Wang T, Zhong C, Gong J. Ca+: cognition augmented counselor agent framework for long-term dynamic client engagement. : arXiv Preprint posted online on Mar 27, 2025. [doi: 10.48550/arXiv.2503.21365]
14. Glikson E, Woolley AW. Human trust in artificial intelligence: review of empirical research. Ann R Coll Physicians Surg Can 2020 Jul;14(2):627-660. [doi: 10.5465/annals.2018.0057]
15. Dwork C, Minow M. Distrust of artificial intelligence: sources & responses from computer science & law. Daedalus 2022 May 1;151(2):309-321. [doi: 10.1162/daed_a_01918]
16. Chavaillaz A, Wastell D, Sauer J. System reliability, performance and trust in adaptable automation. Appl Ergon 2016 Jan;52:333-342. [doi: 10.1016/j.apergo.2015.07.012]
17. Chen C, Sundar SS. Is this AI trained on credible data? the effects of labeling quality and performance bias on user trust. CHI '23 2023 Apr 19:1-11. [doi: 10.1145/3544548.3580805]
18. Yin M, Vaughan JW, Wallach H. Understanding the effect of accuracy on trust in machine learning models. CHI '19 2019 May 2:1-12. [doi: 10.1145/3290605.3300509]

XSL•FO
RenderX

19. Naiseh M, Al-Thani D, Jiang N, Ali R. How the different explanation classes impact trust calibration: the case of clinical decision support systems. Int J Hum Comput Stud 2023 Jan;169:102941. [doi: 10.1016/j.ijhcs.2022.102941]
20. Yu K, Berkovsky S, Conway D, Taib R, Zhou J, Chen F. Do i trust a machine? differences in user trust based on system performance. In: Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent: Springer; 2018:245-264. [doi: 10.1007/978-3-319-90403-0_12]
21. Evans AM, van de Calseyde P. The reputational consequences of generalized trust. Pers Soc Psychol Bull 2018 Apr;44(4):492-507. [doi: 10.1177/0146167217742886] [Medline: 29251247]
22. Wang Y, Wang Y, Crace K, Zhang Y. Understanding attitudes and trust of generative AI chatbots for social anxiety support. Presented at: CHI 2025; Apr 26 to May 1, 2025. [doi: 10.1145/3706598.3714286]
23. Stubbe DE. The therapeutic alliance: the fundamental element of psychotherapy. FOC 2018 Oct;16(4):402-403. [doi: 10.1176/appi.focus.20180022]
24. Stefana A, Fusar-Poli P, Vieta E, Youngstrom EA. Patients' perspective on the therapeutic relationship and session quality: the central role of alliance. Front Psychol 2024;15:1367516. [doi: 10.3389/fpsyg.2024.1367516] [Medline: 39188865]
25. Locher C, Meier S, Gaab J. Psychotherapy: a world of meanings. Front Psychol 2019;10:460. [doi: 10.3389/fpsyg.2019.00460] [Medline: 30984050]
26. DeYoung PA. Relational Psychotherapy: A Primer, 2nd edition: Routledge; 2015. [doi: 10.4324/9781315723709]
27. Feher B, Tussie C, Giannobile WV. Applied artificial intelligence in dentistry: emerging data modalities and modeling approaches. Front Artif Intell 2024;7:1427517. [doi: 10.3389/frai.2024.1427517] [Medline: 39109324]
28. Koyun M, Taskent I. Evaluation of advanced artificial intelligence algorithms' diagnostic efficacy in acute ischemic stroke: a comparative analysis of ChatGPT-4o and Claude 3.5 Sonnet models. JCM 2025 Jan 17;14(2):571. [doi: 10.3390/jcm14020571] [Medline: 39860577]
29. Bürger VK, Amann J, Bui CKT, Fehr J, Madai VI. The unmet promise of trustworthy AI in healthcare: why we fail at clinical translation. Front Digit Health 2024;6:1279629. [doi: 10.3389/fdgth.2024.1279629] [Medline: 38698888]
30. McCoyd JLM, Curran L, Candelario E, Findley P. "There is just a different energy": changes in the therapeutic relationship with the telehealth transition. Clin Soc Work J 2022;50(3):325-336. [doi: 10.1007/s10615-022-00844-0] [Medline: 35493775]
31. Norcross JC. The therapeutic relationship. In: Duncan BL, Miller SD, Wampold BE, Hubble MA, editors. The Heart and Soul of Change: Delivering What Works in Therapy, 2nd edition: American Psychological Association; 2010:113-141. [doi: 10.1037/12075-004]
32. Waldock WJ, Lam G, Baptista A, Walls R, Sam AH. Which curriculum components do medical students find most helpful for evaluating AI outputs? BMC Med Educ 2025 Feb 6;25(1):195. [doi: 10.1186/s12909-025-06735-5] [Medline: 39915801]
33. Doyle B, Fitzsimons D, McKeown P, McAloon T. Understanding dietary decision-making in patients attending a secondary prevention clinic following myocardial infarction. J Clin Nurs 2012 Jan;21(1-2):32-41. [doi: 10.1111/j.1365-2702.2010.03636.x] [Medline: 21545664]
34. Liao QV, Varshney KR. Human-centered explainable AI (XAI): from algorithms to user experiences. arXiv. Preprint posted online on Apr 19, 2022. [doi: 10.48550/arXiv.2110.10790]
35. Artificial intelligence in mental health care. American Psychological Association. 2025 Mar 12. URL: https://www.apa.org/practice/artificial-intelligence-mental-health-care [accessed 2026-03-31]
36. Thomas DR. A general inductive approach for analyzing qualitative evaluation data. Am J Eval 2006 Jun;27(2):237-246. [doi: 10.1177/1098214005283748]
37. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
38. Stamm K, Lin L, Christidis P. Career stages of health service psychologists. American Psychological Association. 2017. URL: https://www.apa.org/workforce/publications/15-health-service-career [accessed 2026-03-18]
39. Rennie DL, Phillips JR, Quartaro GK. Grounded theory: a promising approach to conceptualization in psychology? Can Psychol / Psychol can 1988;29(2):139-150. [doi: 10.1037/h0079765]
40. Luckasson R, Tassé MJ, Schalock RL. Professional responsibility in the field of intellectual and developmental disabilities: its definition, application, and impacts. Intellect Dev Disabil 2022 Jun 1;60(3):183-198. [doi: 10.1352/1934-9556-60.3.183] [Medline: 35640603]
41. Schiappa E, Allen M, Gregg PB. Parasocial relationships and television: a meta-analysis of the effects. In: Preiss RW, Gayle BM, Burrell N, Allen M, Bryant J, editors. Mass Media Effects Research: Advances Through Meta-Analysis Lawrence Erlbaum Associates: Lawrence Erlbaum Associates Publishers; 2007:301-314.
42. Lee JD, See KA. Trust in automation: designing for appropriate reliance. Hum Factors 2004;46(1):50-80. [doi: 10.1518/hfes.46.1.50.30392] [Medline: 15151155]
43. Heinonen E, Orlinsky DE. Psychotherapists' personal identities, theoretical orientations, and professional relationships: elective affinity and role adjustment as modes of congruence. Psychother Res 2013;23(6):718-731. [doi: 10.1080/10503307.2013.814926] [Medline: 23829641]
44. Rojas JC, Teran M, Umscheid CA. Clinician trust in artificial intelligence: what is known and how trust can be facilitated. Crit Care Clin 2023 Oct;39(4):769-782. [doi: 10.1016/j.ccc.2023.02.004] [Medline: 37704339]

XSL•FO
RenderX

45. Ala-Luopa S, Olsson T, Väänänen K, Hartikainen M, Makkonen J. Trusting intelligent automation in expert work: accounting practitioners' experiences and perceptions. Comput Supported Coop Work 2024 Dec;33(4):1343-1371. [doi: 10.1007/s10606-024-09499-6]
46. Kennedy R, Tiede L, Austin A, Ismael K. Law enforcement and legal professionals' trust in algorithms. J Law Empirical Anal 2025 Jun;2(1):77-96. [doi: 10.1177/2755323X251325594]
47. Maple C, Szpruch L, Epiphaniou G, et al. The AI revolution: opportunities and challenges for the finance sector. arXiv. Preprint posted online on Aug 31, 2023. [doi: 10.48550/arXiv.2308.16538]
48. Sibayan K. Big Four now using agentic AI to boost staff productivity. New York State Society of CPAs. 2025 Mar 24. URL: https://www.nysscpa.org/most-popular-content/big-four-now-using-agentic-ai-to-boost-staff-productivity-032425##sthash.43m0gWrb.dpbs [accessed 2026-03-18]
49. Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP. The role of trust in automation reliance. Int J Hum Comput Stud 2003 Jun;58(6):697-718. [doi: 10.1016/S1071-5819(03)00038-7]
50. Aguilar M. Medicare finalized coverage for some mental health apps. Now what? STAT. 2024 Nov 4. URL: https://www.statnews.com/2024/11/04/medicare-payment-mental-health-digital-therapeautics/ [accessed 2026-03-18]
51. Gov Pritzker signs legislation prohibiting AI therapy in Illinois. Illinois Department of Financial & Professional Regulation. 2025 Aug 4. URL: https://idfpr.illinois.gov/news/2025/gov-pritzker-signs-state-leg-prohibiting-ai-therapy-in-il.html [accessed 2026-03-18]

Abbreviations

AI: artificial intelligence
APA: American Psychological Association
COREQ: Consolidated Criteria for Reporting Qualitative Research
GenAI: generative artificial intelligence

Edited by A Stone; submitted 04.Dec.2025; peer-reviewed by N Rodax, Z Liu; accepted 23.Feb.2026; published 02.Apr.2026.

Please cite as:
Kuang J, Pope AL, Zhang Y
Psychotherapists' Trust, Distrust, and Generative AI Practices in Psychotherapy: Qualitative Study
J Med Internet Res 2026;28:e88932
URL: https://www.jmir.org/2026/1/e88932
doi:10.2196/88932

© Jinyan Kuang, Amber L Pope, Yixuan Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 2.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB) for the Evaluation of Medical Large Language Model Question-Answer Studies: Development and Pilot Validation

Carrie Ye[1,2,3], MPH, MD; Joseph Ross Mitchell[1,4], PhD; Daniel C Baumgart[1], MBA, MD, PhD; Zechen Ma[1], MD; Angela Lim Fung[1,5], MSc; Daniela Garcia Orellana[1], MSc; Juel Chowdhury[1,6], MBBS, MPH; Abdullah Abass[1], MD; Steven Katz[1,7], MD; Jacob L Jaremko[1], MD, PhD; Pierre Boulanger[1], PEng, PhD; Claire E H Barber[8], MD, PhD; Gillian Lemermeyer[1], RN, PhD; Hosna Jabbari[1], PhD; Lili Mou[1,4], PhD; Maryam Mirzaei[9], PhD; Mary Waithera Beckett Githumbi[5], MBA; Puneeta Tandon[1], MSc, MD; Randy Goebel[1,4], PhD; Rhys Clark[3], BSc; Whitney Hung[3], MD; Marjan Abbasi[1], MD; Farhad Maleki[8], PhD; Scott Klarenbach[1], MSc, MD; Mohamed Abdalla[1,4], PhD

[1]University of Alberta, 8-130 Clinical Sciences Building, 11350 83 Ave NW, Edmonton, AB, Canada

[2]Arthritis Research Canada, Vancouver, BC, Canada

[3]Alberta Health Services, Edmonton, AB, Canada

[4]Alberta Machine Intelligence Institute, Edmonton, AB, Canada

[5]University of Toronto, Toronto, ON, Canada

[6]University of Oxford, Oxford, United Kingdom

[7]Queen's University, Kingston, ON, Canada

[8]University of Calgary, Calgary, AB, Canada

[9]NAIT Applied Research, Edmonton, AB, Canada

Corresponding Author:
Carrie Ye, MPH, MD
University of Alberta, 8-130 Clinical Sciences Building, 11350 83 Ave NW, Edmonton, AB, Canada

Abstract

Background: Despite the transformative potential of large language models (LLMs) in health care, the rapid development of these tools has outpaced their rigorous evaluation. While artificial intelligence–specific reporting guidelines have been developed to address standardized reporting of artificial intelligence studies, there is currently no specific tool available for risk of bias assessment of LLM question-answer (QA) studies. Existing risk-of-bias tools for medical research are not well suited to the unique challenges of evaluating LLM-QA studies, which creates a critical gap in assessing their safety and effectiveness.

Objective: This study aims to develop the Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB) for LLM-QA studies to systematically evaluate the validity and risk of bias in LLM-QA studies.

Methods: We conducted 2 literature reviews. The first was on quality assessment tools for LLM-QA studies, and the second was on LLM-QA studies, which informed the first draft of the AQAT:RoB. The draft AQAT:ROB was further refined through a prespecified iterative process of modified Delphi, consensus meeting, and validation. The first Delphi process occurred between May 1 and May 20, 2025, and the first consensus meeting was held on May 22. The first round of validation was completed by 4 evaluators, who were not part of the consensus meeting, on 16 randomly selected studies. As this first round of validation surpassed our a priori threshold of ≥80% agreement and a Cohen κ of ≥0.61 between evaluators, no further rounds of development and validation were undertaken. A second Delphi process occurred between February 20 and February 23, 2026, to vote on postpilot changes in response to peer review.

Results: The AQAT:RoB consists of 5 high-level domains (Questions, Reference Answers, LLM Answers, Evaluators, Outcomes). These domains are subdivided into 9 subdomains. Each subdomain includes at least one “Support for Judgment” and at least one “Type of Bias” and is to be rated “low,” “high,” or “unclear” for risk of bias. A pilot evaluation was completed by internal validators who were not part of the consensus discussion and were asked to complete the AQAT:RoB form for each assigned study. Each of the 16 studies was evaluated by 2 evaluators independently. Pilot validation showed a percent agreement of 86.1% and a Cohen κ of 0.70 between assessors.

Conclusions: The AQAT:RoB demonstrates promising initial reliability for assessing the validity or risk of bias in LLM-QA studies. The tool will benefit from future refinements, external validation, and periodic updates to keep pace with evolving technology.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e87057) doi:10.2196/87057

KEYWORDS

risk of bias; quality assessment; large language model; question-answer studies; Alberta Risk of Bias Assessment Tool for LLM-QA studies; AQAT: RoB; chatbot; artificial intelligence

Introduction

Large language models (LLMs) represent a significant technological advancement with transformative potential across various sectors, including health care. Their capabilities in processing and generating human-like text have led to their rapid emergence as tools capable of assisting in complex medical activities, such as disease diagnosis, clinical decision-making, and even administrative tasks, such as writing prescriptions or assigning billing codes [1]. As these sophisticated tools become more integrated into health care ecosystems, robust and rigorous evaluation of their efficacy, safety, and utility is paramount. A critical component of this evaluation involves human assessments, where the performance, usability, and impact of LLM question-answer systems (LLM-QA) such as medical chatbots are gaged through interactions with health care professionals, patients, or simulated users [2,3].

However, the quality of these human evaluation studies varies significantly, with a systematic review indicating that only 5% of studies used real patient care data for LLM evaluation, which can significantly impact the trustworthiness and generalizability of their findings [2]. Without a systematic approach to evaluating the quality and risk of bias in studies assessing LLM-QA, the rapid pace of development could outpace the generation of reliable evidence regarding their actual utility and safety in real-world clinical scenarios, potentially leading to premature or even harmful adoption.

Existing tools for assessing risk of bias, such as the Cochrane Risk of Bias 2 tool (RoB 2) [4] for randomized studies, the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) [5] for diagnostic test studies, the Newcastle-Ottawa Scale (NOS) [6], and the Risk of Bias in Non-randomized Studies - of Exposures tool (ROBINS-E) [7] for nonrandomized studies, are foundational in their respective areas but fall short when applied to the unique methodological and reporting challenges inherent in human evaluation studies of LLMs. While artificial intelligence (AI)–specific quality assessment tools exist, such as the Prediction model Risk of Bias Assessment Tool + AI (PROBAST-AI) [8] and APPRAISE-AI [9], these focus on studies of prediction models using machine learning and are not applicable to LLM-QA studies.

The burgeoning interest in the field is evident from the notable surge in studies pertaining to LLM medical chatbots published in recent years, underscoring the topic's emerging relevance and the urgent need for robust evaluation methodologies [10,11]. AI-specific reporting guidelines have been developed to address standardized reporting of AI studies [9,12-15], including a reporting checklist specifically for chatbot health advice studies, CHART (Chatbot Assessment Reporting Tool) [13]. However, transparent and comprehensive reporting is only one aspect of quality assessment—the other being assessment of risk of bias.

There is currently no tool available to assess the risk of bias in LLM-QA studies. This gap creates a significant challenge for researchers, clinicians, and policymakers attempting to synthesize evidence and make informed decisions about the integration of medical LLM-QA systems. Without a comprehensive and tailored risk of bias assessment tool, the risk of misinterpreting findings, perpetuating methodological flaws, and drawing unsubstantiated conclusions from human evaluation studies is high. To address this knowledge gap, we took a pragmatic but systematic approach to develop and validate the Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB) for LLM-QA studies for the systematic and comprehensive assessment of the risk of bias in medical LLM-QA studies, addressing aspects unique to this emerging field. The tool is intended to evaluate the quality of studies that involve human participants in assessing the outputs of AI models that utilize natural language interactions.

Methods

Overview

The AQAT:RoB development and validation started with the 2 literature reviews. The first was on quality assessment tools for LLM-QA studies, and the second was on LLM-QA studies, which informed the first draft of AQAT:RoB, which went through an iterative process of modified Delphi, consensus meeting, and validation, until our a priori threshold for interrater agreement was met (Figure 1).

XSL•FO
RenderX

Figure 1. Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB) development and validation. This figure outlines the development and pilot validation process (April-September 2025) that was a priori determined and followed to create the AQAT:RoB. The dotted line implies a possible path that would have been followed had the agreement threshold not been met (though this was ultimately not required). LLMs: large language models; LLM-QA: LLM question-answer; RoB: risk of bias. LLM-QA studies include [10,11] and a forthcoming systematic literature review of studies evaluating LLMs for patient-facing health information (protocol registered with PROSPERO; CRD42023461630 [16]).

XSL•FO
RenderX

Ethical Considerations

The University of Alberta's Research Ethics Board deemed that this project meets one of the conditions described under Chapter 2 of *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans* (2022) [17] as an activity that does not require Research Ethics Board review. The AQAT:RoB has been registered with the LATITUDES Network, which was established to increase the robustness of evidence synthesis by improving the process of validity (risk of bias) assessment [18].

Literature Review

We conducted a search query in PubMed in April 2025 to look for existing risk of bias assessment tools specific to LLM-QA studies. The query searched for all studies that contained either of the terms "risk of bias" or "quality assessment" with any of the terms "large language models," "generative AI," or "chatbot" [("risk of bias" OR "quality assessment") AND ("large language models" OR "generative AI" OR "chatbot")]. We limited the search to English studies published in the last 5 years and retrieved 91 results. The search was updated in June 2025, which retrieved 113 studies, and again in September 2025, which retrieved 149 studies. None of these searches included tools for assessing validity or risk of bias in LLM studies. Most studies pertained to the use of LLMs to perform risk of bias assessments. Around the time of each PubMed search, the LATITUDES Network library [19] of risk of bias assessment tools and tools in development was searched for tools that pertained to LLM studies; none were found.

We conducted a literature review of LLM-QA studies to (1) inform the development of AQAT:RoB and (2) find studies for validating the AQAT:RoB. Our group had already conducted, with the assistance of an experienced librarian (DCB), a systematic literature review of studies evaluating LLMs for patient-facing health information (manuscript in progress, protocol registered with PROSPERO [CRD42023461630]) [16]. Medline, Embase, Web of Science, CINAHL, PsycINFO, and Google Scholar were searched for studies published up to July 5, 2023, and then updated to March 7, 2024, limiting searches to the last 10 years prior to the search (Multimedia Appendix 1). Searches were limited to the English language but not to geographic regions. All citations were imported into Covidence for duplicate removal and screening. Abstracts and full texts were each screened independently by 3 reviewers (ZM, ALF, DGO). Disagreements were resolved through a third reviewer (CY). A total of 8943 records were identified, including 2798 duplicates (Multimedia Appendix 2). In total, 6145 titles and abstracts were screened with 327 found to be relevant for full-text review. Of these 327 full texts, 40 were deemed to be original research studies pertaining to LLMs for patient education (Multimedia Appendix 3).

We found 2 very recently published systematic reviews by Wang et al [11] and Huo et al [10], which, along with our systematic review [16], we felt covered the breadth of LLM-QA studies and were thus sufficient to inform the development and validation of the AQAT:RoB. In the systematic literature review conducted by Huo et al [10], their search of MEDLINE, Embase, and Web of Science from inception to October 27, 2023, resulted in 137 eligible studies evaluating the performance of generative AI-driven chatbots. They found that key aspects of internal validity, such as standardized evaluation process, blinding of evaluators, or reference standards, were not well described or simply not included in the studies [10]. Wang et al [11] queried PubMed, Embase, Web of Science, and Scopus from inception until October 14, 2024, and found 168 studies on the accuracy of LLMs when answering clinical questions (although one of these studies has since been retracted). They performed a risk of bias assessment using the NOS [6,11]. Acknowledging the limitations of using the NOS tool given the limited relevance to these studies, they found that only 40 (23.8%) of 168 studies were assessed as having a low overall risk of bias [11].

Candidate Item List Generation

The initial list of items for the AQAT:RoB assessment tool was drafted by CY and MA by reviewing the results and studies of the 3 recent systematic literature reviews [10,11,16] to identify potential sources of bias in LLM-QA and by adapting existing foundational risk of bias assessment tools, including RoB 2 [20], NOS [6], QUADAS-2 [21], and ROBINS-E [7]. The steering committee (CY, MA, JRM, DCB) further developed the initial list of items to form the first draft of the AQAT:RoB (7 domains, 12 sources of bias, 15 supports). This first draft was used in the modified Delphi procedure described below.

Recruitment of Delphi Panelists

To identify participants for the modified Delphi process, the steering committee recruited medical AI experts through the Alberta Machine Intelligence Institute [18,22], the University of Alberta AI + Health Hub [23], and the University of Calgary's Centre for Health Informatics [24]. Interested participants were asked to complete an intake questionnaire regarding demographics, training, and related expertise. All applicants were screened, and panelists were selected by the steering committee to ensure appropriate expertise and diversity and representation on the Delphi panel. In total, there were 19 Delphi panelists, including clinicians (8), computer scientists (8), methodologists (4), researchers (qualitative and quantitative, 17), journal editors (2), and patient partners (2). All Delphi participants were instructed to watch 2 videos that provided background information on risk assessment tools prior to initiating the Delphi process.

Modified Delphi Process

The modified Delphi process occurred between May 1 and May 20, 2025. The steering committee created the Delphi survey using Google Forms. Each participant responded to the survey individually. Participants were asked to rate each item's potential as a source of bias on a 5-point scale (from 1="Not a potential source of bias" to 5="High potential source of bias"). If participants selected either 1="Not a potential source" or 2="Unlikely to be a potential source," they were prompted to provide a short explanation. Participants were also encouraged to use the free-text boxes that followed each item to identify any missing potential sources of biases, questions to identify biases (ie, supports), or types of bias identified by listed supports. The threshold for removing items was more than 50% (10/19) of the participants voted the item as 1="Not a potential source" or 2="Unlikely to be a potential source," and the

XSL•FO
RenderX

threshold for retaining items was less than 50%. Participants were not able to see other participants' votes or comments.

Changes From the Delphi Process

Most items (8 of 12) were rated highly for inclusion (ie, more than 70% (14/19) rated as a potential source of bias). Items rated poorly for agreement or for potential source of bias, and all comments provided by the participants in the free-text boxes were considered for changes (eg, removal, modification, or merging). Based on feedback from participants, no domains or items were added or removed, but modifications were made to 5 items.

Consensus Meeting

An online consensus meeting was held on May 22, 2025, chaired by the steering committee. All panelists were invited except for 2 (S Katz and JLJ) who were excluded from the discussion to serve as adjudicators in the validation phase. The consensus meeting was chaired by the steering committee, and 16 participants attended the synchronous consensus meeting. During this meeting, participants were presented with the initial version of texts; suggested modified versions of the text incorporating the feedback from the modified Delphi process, as well as summary statistics of ratings; and provided comments. During the meeting, each potential source of bias was discussed until consensus was reached on inclusion, type of bias(es), and wording. Once consensus was felt to be reached based on the panel discussion, a formal vote was taken, and unanimity was required before moving on to the next item. At the end of each domain, the panel was asked to discuss if there were any additional potential sources of bias pertaining to that domain.

Changes From the Consensus Meeting

The discussions during the consensus process resulted in multiple changes. There was robust discussion about how granular the "Types of biases" should be, with the group arriving at the conclusion that for the sake of utility and widespread applicability, we would aim for a high-level description of the types of biases. This change affected 5 of 12 "Potential Sources of Bias." Furthermore, there was an addition of "Support for Judgment" for the domain "Performance Metrics."

Pilot Validation

We piloted the AQAT:RoB on 16 studies [25-40], randomly selected from the 319 studies (after removal of duplicates) identified in 3 recent systematic literature reviews [10,11,16] (Multimedia Appendix 4). Random selection was facilitated by listing all 319 studies in alphabetical order of the first author's last name and then using a random number generator (between 1 and 319) to select the 16 studies [41]. Four evaluators (S Katz, JLJ, JC, AA) were asked to complete the AQAT:RoB form for each assigned study. None of the 4 evaluators were part of the consensus discussion, but S Katz and JLJ were on the modified Delphi panel. JC and AA were not part of the tool development process prior to the validation step. All the evaluators were physicians across various specialties (rheumatology, radiology, public health, and primary care). Each of the 16 studies was evaluated by 2 evaluators independently. Evaluators were not provided with any standardized training in order to obtain the most conservative estimates of agreement.

We set an a priori threshold of >80% agreement and a Cohen κ of ≥0.60, which would demonstrate substantial agreement per Landis and Koch's [42] classification system. If we reached this threshold, no further rounds of changes or consensus would be undertaken. If we did not reach this threshold, we planned to pursue further rounds of modified Delphi or consensus or validation until this threshold was achieved (Figure 1).

Protocol Deviations in Response to Peer Review

The steering committee proposed moving the Reporting and Conflict of Interest domains to "Additional consideration" for the Delphi panel after the first round of validation was completed. The rationale was that Reporting was better assessed by stand-alone reporting checklists, and while reporting may lead to an unclear risk of bias judgment, it does not represent a mechanistic bias similar to the other domains in the AQAT:RoB. Similarly, Conflict of Interest may be a predictor or source of bias rather than a distinct mechanism of bias. We set the voting threshold to make this change at 70% for the Delphi panel and planned to conduct another round of validation if the overall percent agreement and Cohen κ after removing these 2 domains did not reach our a priori threshold of >80% agreement and Cohen κ of >0.60. Note that 100% of the panel voted in agreement with this change.

Results

Validation Results

After the first modified Delphi and consensus panel, we met our threshold with a percent agreement of 82.8% and a Cohen κ of 0.63 (calculated across all items; Multimedia Appendix 5). After the second modified Delphi, in response to peer review, during which 2 domains were removed, the percent agreement was 86.1% and Cohen κ was 0.70. The domain with the highest level of agreement was Question Selection (agreement: 93.8%, κ: 0.86), while the domain with the lowest level of agreement was LLM Answer Selection (agreement: 68.8%, κ: 0.30).

AQAT:RoB Tool

The AQAT:RoB tool [43] is summarized in Multimedia Appendix 6. An easy-to-use version, which is both fillable and printable, is available on the AQAT website [43]. Textbox 1 presents the scope and boundaries of studies for which AQAT:RoB is applicable.

XSL•FO
RenderX

Textbox 1. Utilization of Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB).

Intended users: Anyone who wants to appraise the quality or risk of bias of studies in which there are human evaluations of large language model question-answer systems. It is especially important for researchers during the development and peer review of such studies and during the quality assessment stage of systematic literature reviews and meta-analyses. Other potential users include, but are not limited to, patients, health care providers, journal editors and reviewers, medical technology manufacturers, health system administrators, and policymakers.

Target studies: Any study that involves the human evaluation of large language models that provide answers to free-text questions including but not limited to:

- Patient-facing applications
  - Medical chatbots
  - Summary tools (eg, answer questions about imaging reports or doctors' reports)
- Physician-facing applications
  - General medical chatbots
  - Large language model–based clinical decision support systems for physicians
  - Summary tools (eg, answer questions about patient charts)
- Research-based applications:
  - Case finding (eg, find participants based on electronic medical record or electronic health record data and provide justification)
  - Literature review (eg, analyze and summarize scientific literature)

The AQAT:RoB consists of 5 high-level domains (Questions, Reference Answers, LLM Answers, Evaluators, and Outcomes). These domains are subdivided into 9 subdomains. Each subdomain includes at least one "Support for Judgment" and at least one "Type of Bias." Further descriptions of potential sources of bias and best methodological practices are outlined in the text below. In cases of missing, partial, or suboptimal reporting of a specific domain or subdomain, the rating of "unclear" should be assigned. Additional considerations regarding reporting and conflicts of interest are outlined in the *Additional Considerations* section.

Domain 1: Questions

Question Source

Supports for Judgment:

- If questions were created or generated specifically for the study, describe the method used to create the question dataset, including who created the questions and if the questions are reflective of the intended study objective.
- If questions were selected from an existing question source, adequately describe the source to allow an assessment of whether it addresses the intended research question.

The evaluation of LLM-QA models should be conducted against questions that reflect the intended use case, as deviations can introduce biases. When the deviation between intended use and the question source is substantial, the performance on the proxy task may not be generalizable to the stated application. To minimize this risk, the most effective approach is to source questions directly from the real-world use setting. If this is not possible, external data sources are often used to generate questions. In such cases, researchers must justify the degree to which these sources align with the study's core research question and the tool's intended application. For example, if a study evaluates a tool to be used by patients, but the questions are written by a research team of nonpatients, this could introduce bias, as the language and complexity of the questions may not be representative of the intended user. Furthermore, if questions were pulled from a preexisting source, it should be clearly stated if the test questions were included in the training data of the tested models, as performance may reflect memorization versus true model reasoning.

Question Selection

Support for Judgment:

- If questions were selected from an existing question source, describe the method used to select the questions from the original source (eg, random, consecutive, all, or by certain factors).

When selecting questions from an existing dataset, sampling can introduce bias as the selected questions may not be representative of the broader population of potential questions. For instance, a selection mechanism that favors questions of a specific length—such as those with a short, predefined character count, perhaps to minimize computational costs—would systematically exclude longer, more complex questions. Similarly, selecting from a small, nonrandom subset of available options could skew the results, as the chosen questions may not accurately reflect the diversity and range of questions encountered in the tool's intended use case. Therefore, the method for question selection, such as random, purposive, or consecutive sampling, must be clearly reported and justified.

Question Manipulation

Supports for Judgment

- If any questions were manipulated from the original source, describe and justify the rationale for the manipulation.
- If any prompting was provided in addition to the index question, report the exact wording of the prompt(s).

XSL•FO
RenderX

Whether questions are created or extracted from existing sources, researchers may choose to manipulate them for various reasons. For example, slight variations might be introduced to test the model's robustness and bias, assessing how stable its responses are to minor changes in phrasing. Such manipulations are generally less likely to introduce significant bias, as their purpose is to probe the model's inherent stability rather than to alter the nature of the query. Conversely, questions might be manipulated to simplify them for processing by the model. A common example of this is the use of a system prompt that automatically extracts and restructures clinically relevant information before the model attempts to answer. This form of question manipulation, while potentially beneficial for processing, introduces a risk of bias because it may fundamentally alter the user's original query. Or if researchers correct spelling, terminology, or grammatical errors, or split multipart patient questions into separate questions, these changes may augment the performance of the LLM-QA model but not necessarily reflect real-world performance. It is essential that researchers provide both transparency and justification for any question manipulation, as the process could alter the original intent of the question, thereby compromising the validity of the evaluation. Along with direct question manipulation, all prompts, including system prompts, which in and of themselves do not necessarily introduce bias, should be clearly described and should be standardized and stable throughout testing, as differences in prompts may lead to false or misleading performance outcomes.

Domain 2: Reference Answers

Reference Answer Source

Supports for Judgment:

- If reference answers were generated specifically for the study, describe the method used to create the reference answer dataset, including who created the reference answers, and if the answers are reflective of a true reference standard.
- If reference answers were selected from an existing reference answer source, adequately describe the source to allow an assessment of whether it is reflective of a true reference standard.

Often, LLM-QA studies benchmark LLM outputs against reference answers. Bias may be introduced if the reference answers do not accurately reflect a "true" or expected standard. For instance, if reference answers were created by individuals with a different level of expertise or with a different format or standard than the true reference standard, the reference standard may not be valid. For example, if the research team created the reference answers to a higher or lower standard than real-world physician-level responses, the reference standard would be misaligned with the intended quality benchmark. A mismatch in language, structure, style, or level of detail between the study reference standard used and the "true" real-world reference standard can lead to biased results. As a single ground truth does not always exist in medicine, the selected reference standard should be decided a priori (eg, guideline-based or expert consensus-based) and described and justified.

Reference Answer Selection

Support for Judgment:

- If not all reference answers to a given question were used, describe the method by which reference answers were selected.

Just as with question selection, the process of selecting reference answers from a larger pool can introduce sampling bias. This bias occurs if the selection method systematically favors answers with certain qualities, making the final set of reference answers unrepresentative of the full range of possible correct responses. For example, if a question has multiple valid reference answers but researchers consistently choose those with a specific tone or level of detail, the evaluation will be skewed toward models that produce similar outputs. Therefore, it is critical to describe and justify the method used for selecting reference answers.

Domain 3: LLM Answers

Support for Judgment:

- Describe how many answers were generated for each question and if not all answers were assessed, describe how answers were selected for assessment.

When evaluating language models, it is often prudent to generate multiple answers for a single question to assess the model's stability or to explore the diversity of its outputs. In such instances, only evaluating a subset of the generated answers (eg, choosing the best one) may not be representative of the model's typical performance, thereby leading to an inaccurate or misleading evaluation. Therefore, it is crucial for researchers to transparently describe and justify how many answers were generated for each question and, if not all of them were assessed, to detail the specific methodology used to select the answers for evaluation.

Domain 4: Evaluators

Evaluator Selection

Support for Judgment

- Describe the method used to select evaluators, and assign evaluators to specific LLM qualities.

The selection of evaluators should reflect the intended real-world use and the required expertise to judge the domains being assessed. For example, having a physician evaluate the empathy of an LLM-QA's outputs may not reflect how a patient would assess this domain. Likewise, it would not be appropriate for a patient to evaluate the accuracy of health information generated by an LLM-QA, as they would lack the appropriate expertise. As many studies assess multiple outcomes, more than one type of evaluator may be required for a given study (eg, physicians evaluate accuracy and patients rate readability). Furthermore, the demographic or professional characteristics of the evaluators should align with the intended user population of the LLM or chatbot. For example, if an LLM is designed for a general audience but its readability is evaluated exclusively by individuals with advanced academic degrees, the results may not accurately reflect how an average user would perceive the content.

XSL•FO
RenderX

Blinding of Evaluators

Support for Judgment:

- Describe all measures used, if any, to blind trial evaluators and researchers from knowledge of the answer source. Provide information relating to whether the intended blinding was effective.

The integrity of an evaluation can be compromised if evaluators are not blinded to the source of the answers they are assessing (reference standard vs LLM-generated) because evaluators' preexisting beliefs, attitudes, or knowledge about a specific technology, such as AI, or even to a specific LLM model, can unconsciously influence their ratings. Thus, evaluators should be blinded to the answer source, and researchers should describe the blinding measures employed. Since naive blinding is not guaranteed to be effective given stylistic markers in LLM-generated text, authors should describe the steps taken to assess or verify the effectiveness of the blinding (eg, ask the evaluators if they could identify the AI-generated answer).

Domain 5: Outcomes—Performance Metrics

Supports for Judgment:

- Describe specific metrics used for each outcome quality.
- Describes if desired outcomes were prespecified prior to conducting the study.

To minimize the risk of bias, 2 crucial steps should be taken. First, the desired outcomes or hypotheses of the study should be prespecified prior to conducting any analysis. Second, the metrics selected to measure each outcome must directly align with the stated goals of the study. For example, if the goal is to evaluate a chatbot's ability to provide concise summaries of medical information, metrics should focus on conciseness and accuracy, rather than on secondary qualities, such as conversational tone or creativity. A misalignment between metrics and study goals introduces bias, as the evaluation would not accurately reflect the model's performance on its intended task.

Additional Considerations

Complete and transparent reporting is required to judge the risk of bias. Researchers must account for any instances of missing data and describe how missing data were handled. For example, if certain questions were too long for the model to process, or if the model failed to produce a response, these omissions should be explicitly noted and their potential impact on the evaluation should be discussed. Similarly, if human evaluators did not complete all of their annotations, it is important that this missingness is reported and ideally investigated, as these instances of missingness may not be random and could introduce bias if not accounted for. Study outcomes should be decided *a priori* and deviations should be described and justified. By reporting a subset of all measured outcomes or manipulating the analysis post hoc to achieve a different result (eg, by recategorizing certain groups, shifting the scale), a study may present a distorted or optimistic view of model performance. The use of appropriate reporting checklists is recommended.

Conflicts of interest (especially commercial interests) may introduce explicit or subconscious biases in the formulation of the problem, the execution of the analysis, or the interpretation of the results. If there are conflicts of interest (eg, authors funded or affiliated with model vendors), they must be disclosed and mitigated, if possible.

Discussion

Principal Findings

The AQAT:RoB tool was developed to standardize the quality assessment and specifically, the risk of bias assessments, of studies in which there are human evaluations of LLM-QA systems. This easy-to-use risk of bias assessment tool covers 5 major domains (Questions, Reference answers, LLM answers, Evaluators, Outcomes, Reporting, and Other), with a total of 9 potential sources of bias, each with 1‐2 support for judgment prompts and a list of types of potential bias. In our pilot validation by 4 assessors of 16 studies, the AQAT:RoB showed a substantial degree of agreement between internal validators blinded to the development process.

Comparison to Prior Work

As demonstrated in Table 1, the AQAT:RoB assesses many important aspects of LLM-QA studies that are either not covered at all or only tangentially covered by other foundational risk of bias tools [4-7]. Most notably, Domain 1: Questions, a very important potential source of bias, is not addressed by any of the foundational tools. We also note that the panel explicitly discussed and voted to include conflict of interest in the AQAT:RoB, which is not always included in foundational tools, given the increasing commercialization of natural language processing research. The vast majority of computer science faculty at top schools have financial conflicts with industry [44], and the field of natural language processing is so reliant on industry artifacts to the point of being described as "captured" [45]. This is particularly relevant in evaluation works, as past peer-reviewed evaluations have then been used by the relevant industry party to claim that their models have been "independently audited" [46]. The foundational tools listed in Table 1 [4-7] either do not provide a threshold for determining overall risk of bias or use a "worst-of" approach, where the overall risk of bias is considered "High risk" if at least 1 domain is judged as "High risk." We have chosen to leave the determination of overall risk of bias to the discretion of the user, as the threshold may be different depending on the intended use of the tool, although in most cases, a single domain being judged as "High risk" would likely result in an overall judgment of "High risk."

XSL•FO
RenderX

Table . Comparison of Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB) and foundational risk of bias tools[a].

;AQAT:RoB domain	RoB 2[b] [4]	NOS[c] [6]	QUADAS-2[d] [5]	ROBINS-E[e] [7]
Questions	x[f]	x	x	x
Reference answers	x	• Domain 1: Selection • Domain 2: Comparability	• Domain 3: Reference standard • Domain 4: Flow and timing	• Domain 2: Risk of bias arising from measurement of the exposure
LLM[g] answers	x	• Domain 1: Selection • Domain 2: Comparability	• Domain 2: Index test or tests • Domain 4: Flow and timing	• Domain 2: Risk of bias arising from measurement of the exposure
Evaluators	• Domain 4: Risk of bias in measurement of the outcome	• Domain 1: Selection • Domain 3: Outcomes	• Domain 1: Patient selection	• Domain 3: Risk of bias in selection of participants into the study (or into the analysis)
Outcomes	• Domain 4: Risk of bias in measurement of the outcome	• Domain 3: Outcomes	x	• Domain 6: Risk of bias arising from measurement of the outcome

[a]This table compares the coverage of key domains determined to be important for assessing risk of bias in large language model question-answer studies by the Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB) compared with existing foundational risk of bias tools. No AQAT:RoB domains were covered adequately by other Foundational Risk of Bias tools. Terms in italics signify the closest domain found in the foundational tool.

[b]RoB 2: Cochrane Risk of Bias 2 tool.

[c]NOS: Newcastle-Ottawa Scale.

[d]QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2.

[e]ROBINS-E: Risk of Bias in Non-randomized Studies - of Exposures.

[f]x: not covered by the tool.

[g]LLM: large language model.

The AQAT:RoB is the first risk of bias assessment tool specifically designed for studies of human-evaluated LLM-QA systems. Despite this, systematic reviews of LLM-QA studies have already been published [10,11], highlighting the urgency of the need for the AQAT:RoB. Previous AI evaluation frameworks primarily function as reporting checklists [9,12,13,15]. While these are invaluable for assessing transparency and reproducibility, they do not directly evaluate a study's susceptibility to bias. For instance, a study may meticulously report every detail, yet still contain high-risk elements, such as a lack of blinding for answer sources, that could compromise its findings. Existing AI-specific quality assessment tools, such as PROBAST-AI [8] and APPRAISE-AI [9], are tailored for predictive machine learning models, which differ significantly from the evaluation needs of LLM-QA studies. Templin et al [47] introduced a useful 5-step framework for auditing LLMs, but not for assessing the validity of LLM-QA studies. Therefore, the AQAT:RoB addresses a critical void in the standardized evaluation of LLM-QA research.

Aiming to address the urgent need for a risk of bias assessment tool for this growing field and aided by our own existing [16] and recently published systematic reviews [10,11], we sought to develop the AQAT:RoB through a systematic, but pragmatic approach. With a highly engaged group of interdisciplinary experts, our pilot validation was able to achieve high interrater agreement after the first round of modified Delphi and consensus. In our pilot validation of 16 studies, the interrater reliability was on par with or better than that of existing foundational risk of bias tools. For example, studies have found that the RoB 2 demonstrated kappas consistent with "fair" agreement (0.21 - 0.40) [48,49]. Another study found that while the interrater reliability across 6 risk of bias tools for nonrandomized studies varied widely, most demonstrated intraclass correlation coefficients in the substantial range, similar to the AQAT:RoB [50]. No studies have demonstrated the interrater reliability of existing RoB tools specifically to LLM-QA studies.

Limitations

We recognize that there are limitations to the AQAT:RoB. First, while it was developed by a wide interdisciplinary group of experts and patient partners, most experts were from Alberta due to the nature of the AQAT collaborative, which includes the development of other quality assessment tools for AI-related studies, including the development of publicly available datasets, validated evaluation scales, and measurements and programs. In future updates of the AQAT:RoB, we plan to engage a more international group of partners. It will also be crucial for the AQAT:RoB to be extensively and externally validated by the broader international community of researchers [51]. Furthermore, the AQAT:RoB was developed with a medical focus and would require validation and adaptation for nonmedical studies of LLM-QA. Second, we recognize that this tool will need to evolve with the rapid development of LLM tools and related studies. Third, this tool was developed in

XSL•FO
RenderX

English and evaluated only on English-language studies. In order to use this tool on non-English–language studies, it would ideally be translated and validated in other languages. Multilingual and non-English evaluations may require an expansion of the support for judgments to be considered when classifying the risk of bias. For example, if questions or reference answers are translated from English, they may not accurately reflect the distribution against which they will be evaluated (ie, native, nontranslated questions).

Finally, the pilot validation has limitations. It included only 16 studies, a relatively small sample, which may limit reliability estimates. While construct validity was supported by expert consensus and signaling questions that map directly to known mechanisms of bias, our pilot validation focused on interrater reliability, the most commonly reported evaluation metric for such tools. To further support construct validity, future evaluations that test concurrent and criterion validity are needed, although the lack of truly comparative tools and meta-analyses in this field limits the feasibility of these types of evaluations. Finally, reliance on an aggregate score for our a priori threshold makes it possible for domains with high agreement to compensate for domains with poor agreement. We acknowledge that agreement for the LLM Answer Selection domain was lower or less reliable than for the other domains and remains experimental, requiring future refinements after more extensive validation.

Future Directions

The AQAT:RoB is a crucial next step toward standardizing the evaluation of LLM-QA studies in medicine. While we acknowledge that the tool will benefit from future refinements, more extensive validation (particularly external validation), and periodic updates to keep pace with evolving technology, we believe it currently fills an urgent need and critical gap. The immediate application of this tool will enable researchers, clinicians, and policymakers to more effectively and rigorously assess the validity of LLM-based studies, thereby ensuring that real-world applications of this technology are built on a solid foundation of reliable evidence.

Conclusions

The AQAT:RoB demonstrates promising initial reliability for assessing the validity or risk of bias of LLM-QA studies.

Acknowledgments

We thank the Alberta Machine Intelligence Institute, AI+Health Hub, and the Centre for Health Informatics for providing administrative and recruitment support. We thank Ms Dagmara Chojecki, MLIS, health sciences librarian, for her assistance with the systematic literature review.

Funding

This research is supported by the Canadian Institutes of Health Research (number 96047). MA is supported by a Canada CIFAR AI Chair. RM is supported by a Canada CIFAR AI Chair and the Alberta Health Services Chair in AI in Health. LM is supported by NSERC and Canada CIFAR AI Chair. RG is supported by Amii and NSERC. JM is supported by a Canada CIFAR AI Chair and by Medical Imaging Consultants. SK is supported by the Kidney Health Research Chair and the Division of Nephrology at the University of Alberta.

Data Availability

The data extracted and synthesized for this study are available in the multimedia appendix files.

Authors' Contributions

Conceptualization: CY
Data curation: CY, MA
Formal analysis: CY, MA
Investigation: AA, ALF, CEHB, CY, DCB, DGO, FM, GL, HJ, JC, JLJ, JRM, LM, MA, MM, MWBG, PB, PT, RC, RG, S Katz, S Klarenbach, WH, ZM
Methodology: CY, DCB, JRM, MA
Writing – original draft: CY, MA
Writing – review & editing: AA, ALF, CEHB, CY, DCB, DGO, FM, GL, HJ, JC, JLJ, JRM, LM, MA, MM, MWBG, PB, PT, RC, RG, S Katz, S Klarenbach, WH, ZM

Conflicts of Interest

S Klarenbach is Director of the Real World Evidence Consortium, and Alberta Drug and Therapeutic Evaluation Consortium (Universities of Alberta, Calgary, and Institute of Health Economics); these entities receive funding from decision-makers and industry to conduct research. All research funding is made to the academic institution; investigators retain full rights of academic freedom and right to publish. This relationship is not related to the current work. All other authors declare no conflicts of interest.

Multimedia Appendix 1

XSL•FO
RenderX

Search strategy for the large language model (LLM) patient education systematic literature review.
[DOCX File, 3124 KB - jmir_v28i1e87057_app1.docx]

Multimedia Appendix 2
PRISMA flow diagram of unpublished systematic literature review of studies evaluating large language models (LLMs) for patient-facing health information published between July 2013 and March 2024.
[PNG File, 140 KB - jmir_v28i1e87057_app2.png]

Multimedia Appendix 3
Large language model (LLM) patient education systematic literature review list of studies reviewed.
[PDF File, 115 KB - jmir_v28i1e87057_app3.pdf]

Multimedia Appendix 4
Studies used in the pilot validation set (n=16 studies).
[PDF File, 81 KB - jmir_v28i1e87057_app4.pdf]

Multimedia Appendix 5
Pilot validation dataset. Sixteen papers were each assessed by 2 evaluators for a total of 32 evaluations.
[XLSX File, 12 KB - jmir_v28i1e87057_app5.xlsx]

Multimedia Appendix 6
Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB).
[DOCX File, 15 KB - jmir_v28i1e87057_app6.docx]

References

1. Iqbal U, Tanweer A, Rahmanti AR, Greenfield D, Lee LTJ, Li YCJ. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. J Biomed Sci 2025 May 7;32(1):45. [doi: 10.1186/s12929-025-01131-z] [Medline: 40335969]
2. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025 Jan 28;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
3. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA 2023 Sep 5;330(9):866-869. [doi: 10.1001/jama.2023.14217] [Medline: 37548965]
4. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
5. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
6. Lo CKL, Mertz D, Loeb M. Newcastle-Ottawa Scale: comparing reviewers' to authors' assessments. BMC Med Res Methodol 2014 Apr 1;14:45. [doi: 10.1186/1471-2288-14-45] [Medline: 24690082]
7. Higgins JPT, Morgan RL, Rooney AA, et al. A tool to assess risk of bias in non-randomized follow-up studies of exposure effects (ROBINS-E). Environ Int 2024 Apr;186:108602. [doi: 10.1016/j.envint.2024.108602] [Medline: 38555664]
8. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021 Jul 9;11(7):e048008. [doi: 10.1136/bmjopen-2020-048008] [Medline: 34244270]
9. Kwong JCC, Khondker A, Lajkosz K, et al. APPRAISE-AI tool for quantitative evaluation of AI studies for clinical decision support. JAMA Netw Open 2023 Sep 5;6(9):e2335377. [doi: 10.1001/jamanetworkopen.2023.35377] [Medline: 37747733]
10. Huo B, Boyle A, Marfo N, et al. Large language models for chatbot health advice studies: a systematic review. JAMA Netw Open 2025 Feb 3;8(2):e2457879. [doi: 10.1001/jamanetworkopen.2024.57879] [Medline: 39903463]
11. Wang L, Li J, Zhuang B, et al. Accuracy of large language models when answering clinical research questions: systematic review and network meta-analysis. J Med Internet Res 2025 Apr 30;27:e64486. [doi: 10.2196/64486] [Medline: 40305085]
12. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024 Apr 16;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
13. CHART Collaborative. Reporting guideline for chatbot health advice studies: the Chatbot Assessment Reporting Tool (CHART) statement. BMJ Med 2025;4(1):e001632. [doi: 10.1136/bmjmed-2025-001632] [Medline: 40761518]
14. El Emam K, Leung TI, Malin B, Klement W, Eysenbach G. Consolidated reporting guidelines for prognostic and diagnostic machine learning models (CREMLS). J Med Internet Res 2024 May 2;26:e52508. [doi: 10.2196/52508] [Medline: 38696776]

XSL•FO
RenderX

15. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med 2025 Jan;31(1):60-69. [doi: 10.1038/s41591-024-03425-5] [Medline: 39779929]
16. The use of large language models in patient education interventions: a systematic review. PROSPERO. URL: https://www.crd.york.ac.uk/PROSPERO/view/CRD42023461630 [accessed 2025-09-03]
17. TCPS 2 (2022) – chapter 2: scope and approach. Government of Canada. 2023. URL: https://ethics.gc.ca/eng/tcps2-eptc2_2022_chapter2-chapitre2.html [accessed 2026-03-27]
18. Tools in development. Latitudes Network. 2023. URL: https://www.latitudes-network.org/library/tools-in-development/ [accessed 2025-09-02]
19. Latitudes Network. 2023. URL: https://www.latitudes-network.org/ [accessed 2025-09-16]
20. RoB 2 tool. Risk of bias tools. URL: https://sites.google.com/site/riskofbiastool/welcome/rob-2-0-tool?authuser=0 [accessed 2025-06-18]
21. QUADAS | Bristol medical school: population health sciences. University of Bristol. URL: https://www.bristol.ac.uk/population-health-sciences/projects/quadas/ [accessed 2025-06-18]
22. Alberta Machine Intelligence Institute. URL: https://www.amii.ca/ [accessed 2025-09-02]
23. AI + Health Hub. University of Alberta. URL: https://www.ualberta.ca/en/health-sciences/research/ai-and-health-hub.html [accessed 2025-09-02]
24. Centre for health informatics. University of Calgary. URL: https://cumming.ucalgary.ca/centres/centre-health-informatics [accessed 2025-09-03]
25. Bernstein IA, Zhang YV, Govil D, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Netw Open 2023 Aug 1;6(8):e2330320. [doi: 10.1001/jamanetworkopen.2023.30320] [Medline: 37606922]
26. Cappellani F, Card KR, Shields CL, Pulido JS, Haller JA. Reliability and accuracy of artificial intelligence ChatGPT in providing information on ophthalmic diseases and management to patients. Eye 2024 May;38(7):1368-1373. [doi: 10.1038/s41433-023-02906-0] [Medline: 38245622]
27. Chaker SC, Hung YC, Saad M, Golinko MS, Galdyn IA. Easing the burden on caregivers—applications of artificial intelligence for physicians and caregivers of children with cleft lip and palate. Cleft Palate Craniofac J 2025 Apr;62(4):574-587. [doi: 10.1177/10556656231223596] [Medline: 38178785]
28. Chen S, Kann BH, Foote MB, et al. The utility of ChatGPT for cancer treatment information. medRxiv. Preprint posted online on Mar 23, 2023. [doi: 10.1101/2023.03.16.23287316]
29. Chervenak J, Lieman H, Blanco-Breindel M, Jindal S. The promise and peril of using a large language model to obtain clinical information: ChatGPT performs strongly as a fertility counseling tool with limitations. Fertil Steril 2023 Sep;120(3):575-583. [doi: 10.1016/j.fertnstert.2023.05.151] [Medline: 37217092]
30. Coskun B, Ocakoglu G, Yetemen M, Kaygisiz O. Can ChatGPT, an artificial intelligence language model, provide accurate and high-quality patient information on prostate cancer? Urology 2023 Oct;180:35-58. [doi: 10.1016/j.urology.2023.05.040] [Medline: 37406864]
31. Gabriel J, Shafik L, Alanbuki A, Larner T. The utility of the ChatGPT artificial intelligence tool for patient education and enquiry in robotic radical prostatectomy. Int Urol Nephrol 2023 Nov;55(11):2717-2732. [doi: 10.1007/s11255-023-03729-4] [Medline: 37528247]
32. Liu HY, Alessandri Bonetti M, Jeong T, Pandya S, Nguyen VT, Egro FM. Dr. ChatGPT will see you now: how do Google and ChatGPT compare in answering patient questions on breast reconstruction? J Plast Reconstr Aesthet Surg 2023 Oct;85:488-497. [doi: 10.1016/j.bjps.2023.07.039]
33. Kianian R, Sun D, Giaconi J. Can ChatGPT aid clinicians in educating patients on the surgical management of glaucoma? J Glaucoma 2024 Feb 1;33(2):94-100. [doi: 10.1097/IJG.0000000000002338] [Medline: 38031276]
34. McCarthy CJ, Berkowitz S, Ramalingam V, Ahmed M. Evaluation of an artificial intelligence chatbot for delivery of IR patient education material: a comparison with societal website content. J Vasc Interv Radiol 2023 Oct;34(10):1760-1768. [doi: 10.1016/j.jvir.2023.05.037] [Medline: 37330210]
35. Padovan M, Cosci B, Petillo A, et al. ChatGPT in occupational medicine: a comparative study with human experts. Bioengineering (Basel) 2024 Jan 6;11(1):57. [doi: 10.3390/bioengineering11010057] [Medline: 38247934]
36. Thia I, Saluja M. ChatGPT: is this patient education tool for urological malignancies readable for the general population? Res Rep Urol 2024;16:31-37. [doi: 10.2147/RRU.S440633] [Medline: 38259300]
37. Mayo-Yáñez M, Lechien JR, Maria-Saibene A, Vaira LA, Maniaci A, Chiesa-Estomba CM. Examining the performance of ChatGPT 3.5 and Microsoft Copilot in otolaryngology: a comparative study with otolaryngologists' evaluation. Indian J Otolaryngol Head Neck Surg 2024 Aug;76(4):3465-3469. [doi: 10.1007/s12070-024-04729-1] [Medline: 39130248]
38. Pressman SM, Borna S, Gomez-Cabello CA, Haider SA, Forte AJ. AI in hand surgery: assessing large language models in the classification and management of hand injuries. J Clin Med 2024 May 11;13(10):2832. [doi: 10.3390/jcm13102832] [Medline: 38792374]
39. Mesnier J, Suc G, Sayah N, Abtan J, Steg PG. Relevance of medical information obtained from ChatGPT: are large language models friends or foes? Arch Cardiovasc Dis 2023 Oct;116(10):485-486. [doi: 10.1016/j.acvd.2023.07.009] [Medline: 37718185]

XSL•FO
RenderX

40. Rizwan A, Sadiq T. The use of AI in diagnosing diseases and providing management plans: a consultation on cardiovascular disorders with ChatGPT. Cureus 2023 Aug;15(8):e43106. [doi: 10.7759/cureus.43106] [Medline: 37692649]
41. Haahr M. RANDOM.ORG. URL: https://www.random.org/ [accessed 2025-09-03]
42. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]
43. Alberta Quality Assessment Tools. URL: https://aqat.ai/#tools [accessed 2026-03-18]
44. Abdalla M, Abdalla M. The grey hoodie project: big tobacco, big tech, and the threat on academic integrity. : ACM; 2021 Presented at: AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society; May 19-21, 2021; Virtual event. [doi: 10.1145/3461702.3462563]
45. Aitken W, Abdalla M, Rudie K, Stinson C. Collaboration or corporate capture? Quantifying NLP's reliance on industry artifacts and contributions. In: Ku LW, Martins A, Srikumar V, editors. : Association for Computational Linguistics; 2024 Presented at: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Aug 11-16, 2024; Bangkok, Thailand p. 3433-3448. [doi: 10.18653/v1/2024.acl-long.188]
46. Young M, Katell M, Krafft PM. Confronting power and corporate capture at the FAccT conference. : ACM; 2022 Presented at: FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency; Jun 21-24, 2022; Seoul, South Korea. [doi: 10.1145/3531146.3533194]
47. Templin T, Fort S, Padmanabham P, et al. Framework for bias evaluation in large language models in healthcare settings. NPJ Digit Med 2025 Jul 7;8(1):414. [doi: 10.1038/s41746-025-01786-w] [Medline: 40624264]
48. Minozzi S, Cinquini M, Gianola S, Gonzalez-Lorenzo M, Banzi R. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. J Clin Epidemiol 2020 Oct;126:37-44. [doi: 10.1016/j.jclinepi.2020.06.015] [Medline: 32562833]
49. Hartling L, Hamm MP, Milne A, et al. Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. J Clin Epidemiol 2013 Sep;66(9):973-981. [doi: 10.1016/j.jclinepi.2012.07.005] [Medline: 22981249]
50. Kalaycioglu I, Rioux B, Briard JN, et al. Inter-rater reliability of risk of bias tools for non-randomized studies. Syst Rev 2023 Dec 7;12(1):227. [doi: 10.1186/s13643-023-02389-w] [Medline: 38057883]
51. Tomlinson E, Cooper C, Davenport C, et al. Common challenges and suggestions for risk of bias tool development: a systematic review of methodological studies. J Clin Epidemiol 2024 Jul;171:111370. [doi: 10.1016/j.jclinepi.2024.111370] [Medline: 38670243]

Abbreviations

AI: artificial intelligence
AQAT:RoB: Alberta Quality Assessment Tool: Risk of Bias
CHART: Chatbot Assessment Reporting Tool
LLM: large language model
LLM-QA: LLM question-answer system
NOS: Newcastle-Ottawa Scale
PROBAST-AI: Prediction model Risk of Bias Assessment Tool
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2
RoB 2: Cochrane Risk of Bias 2 tool
ROBINS-E: Risk of Bias in Non-randomized Studies - of Exposures tool

Edited by A Coristine; submitted 04.Nov.2025; peer-reviewed by A Njoku, G Surianarayanan, X Michelena; accepted 24.Feb.2026; published 08.Apr.2026.

Please cite as:
Ye C, Mitchell JR, Baumgart DC, Ma Z, Fung AL, Orellana DG, Chowdhury J, Abass A, Katz S, Jaremko JL, Boulanger P, Barber CEH, Lemermeyer G, Jabbari H, Mou L, Mirzaei M, Githumbi MWB, Tandon P, Goebel R, Clark R, Hung W, Abbasi M, Maleki F, Klarenbach S, Abdalla M
The Alberta Quality Assessment Tool: Risk of Bias (AQAT:RoB) for the Evaluation of Medical Large Language Model Question-Answer Studies: Development and Pilot Validation
J Med Internet Res 2026;28:e87057
URL: https://www.jmir.org/2026/1/e87057
doi:10.2196/87057

XSL•FO
RenderX

© Carrie Ye, Joseph Ross Mitchell, Daniel C Baumgart, Zechen Ma, Angela Lim Fung, Daniela Garcia Orellana, Juel Chowdhury, Abdullah Abass, Steven Katz, Jacob L Jaremko, Pierre Boulanger, Claire E H Barber, Gillian Lemermeyer, Hosna Jabbari, Lili Mou, Maryam Mirzaei, Mary Waithera Beckett Githumbi, Puneeta Tandon, Randy Goebel, Rhys Clark, Whitney Hung, Marjan Abbasi, Farhad Maleki, Scott Klarenbach, Mohamed Abdalla. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 8.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Disclaimers and Referral Patterns for Medical Advice Across Urgency Levels: Large Language Model Evaluation Study

Florian Reis[1], MD; Louis Agha-Mir-Salim[1], MD; Richard Hickstein[2], MD; Moritz Reis[3], PhD; Sophie K Piper[1,4], PhD; Felix Balzer[1], MD, PhD; Sebastian Daniel Boie[1], PhD

[1]Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Charitéplatz 1, Berlin, Germany

[2]Working Group on Cardiovascular Magnetic Resonance, Experimental and Clinical Research Center, a cooperation between Charité – Universitätsmedizin Berlin and the Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany

[3]Institute of Psychology, Julius Maximilians University, Würzburg, Germany

[4]Institute of Biometry and Clinical Epidemiology, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

Corresponding Author:
Florian Reis, MD
Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Charitéplatz 1, Berlin, Germany

Abstract

Background: "I'm not a doctor, but..." is a typical response when asking considerate laypeople for health advice. However, seeking medical advice has also shifted to digital settings, where the expertise of the other party is less transparent than in face-to-face interactions. Recently, large language models (LLMs) have emerged as easily accessible tools, offering a novel way to formulate medical questions and receive seemingly qualified advice. Given the sensitive nature of health-related queries and the lack of professional supervision, incorrect advice can pose serious health risks. Therefore, including explicit disclaimers and precise referrals in LLM responses to medical queries is crucial. However, little is known about how LLMs adapt their safety implementations in response to different urgency levels.

Objective: This study evaluates disclaimer and referral patterns in responses from LLMs to authentic medical queries of different urgency levels using a systematic evaluation framework.

Methods: This prospective, multimodel evaluation study generated and analyzed 908 responses from 4 popular LLMs (GPT-4o, Claude Sonnet-4, Grok-3, and DeepSeek-V3) to 227 authentic patient queries from a public dataset. Two human raters classified all 227 patient queries using a 3-level urgency scale. LLM responses were evaluated using a 5-point ordinal classification system for disclaimer and referral advice, ranging from "no disclaimer" to "urgent advice to consult a medical professional." GPT-4o served as the primary rater model for this task after conducting a subset validation against human expert annotations. Statistical analyses included Jonckheere-Terpstra tests to examine the relationship between case urgency and disclaimer ratings and Kruskal-Wallis tests for intermodel comparisons.

Results: The 227 patient queries were distributed as 77 (34%) low-urgency, 110 (48%) intermediate-urgency, and 40 (18%) high-urgency cases. All 4 LLMs demonstrated statistically significant ordered trends (all $P<.001$), with higher-urgency queries receiving more explicit referral advice. Disclaimer and referral advice clustered toward higher categories across all models, with 97% (881/908) of responses indicating that a medical professional should be consulted. Sonnet-4, Grok-3, and GPT-4o demonstrated a conservative approach, with 89%, 89%, and 88%, respectively, of their responses being either explicit or urgent referrals. In contrast, DeepSeek-V3 showed a broader distribution, with 74% of responses falling into these categories. Interrater reliability between GPT-4o and human raters achieved moderate to substantial agreement, with weighted Cohen κ values between 0.415 and 0.707.

Conclusions: Current LLMs exhibit urgency-responsive safety mechanisms when providing medical advice. All evaluated models adaptively incorporate more explicit disclaimers and urgent referrals for higher-urgency queries. However, variability between LLMs highlights the need for standardized safety measures and appropriate regulatory frameworks. Although these findings indicate progress regarding safety concerns, the public availability of LLMs requires careful consideration to ensure consistent protection against patient harm while preserving the benefits of low-threshold access to health information.

(J Med Internet Res 2026;28:e84668) doi:10.2196/84668

XSL•FO
RenderX

KEYWORDS

artificial intelligence; large language models; chatbots; patient safety; risk assessment; digital health; health information systems; consumer health information; triage; medical liability

Introduction

The health care landscape faces a fundamental transformation as large language models (LLMs) gain popularity for seeking health information [1]. A recent survey indicates that people without regular access to primary care are particularly likely to use conversational artificial intelligence (AI) systems, such as OpenAI's ChatGPT, to obtain medical advice and interpret symptoms [2]. This shift represents a paradigm change from traditional health care consultation models and has occurred largely outside established health care systems, creating a parallel pathway for health information access and self-diagnosis [3]. The implications are profound, with the potential to influence health care utilization patterns and reshape stakeholder relationships as patients arrive with AI-generated hypotheses and expectations [4].

From a technical perspective, LLMs demonstrate remarkable representation of medical knowledge. They are trained on vast amounts of literature, enabling them to generate contextually appropriate medical information [5]. Empirical evaluations show notable diagnostic reasoning performance, with some LLMs outperforming traditional tools in differential diagnosis and clinical decision support [6,7]. However, significant limitations temper these capabilities. LLMs are susceptible to hallucinations and sometimes fail to follow instructions or guidelines [8,9]. While they perform well on structured multiple-choice questions, they often struggle with nuanced, open-ended clinical reasoning and questions requiring a higher level of expertise [10]. Similarly, many models demonstrate a decline in accuracy when confronted with unfamiliar patterns [11]. This suggests that LLM-encoded "medical knowledge" depends on training data rather than being an analog of human clinical cognition.

From an interpersonal perspective, patient perception and trust are crucial for the acceptance of AI-supported health care measures [12]. Research has shown that people tend to be more critical of medical advice labeled as AI-generated than of advice from human experts, even when the content is identical [13]. Using disclaimers effectively and establishing clear referral pathways can help build trust. Specifically, a well-designed disclaimer that transparently communicates limitations, alongside a clear referral pathway to a human expert, could improve patients' ability to correctly interpret the advice given. While research demonstrates that warning messages can modify compliance behavior [14], professional standards for the implementation of disclaimers for AI health applications remain largely undeveloped.

Notably, the provision of medical advice carries significant legal implications, which also extend to AI-generated health care guidance [15]. However, while producing medical device–like output [16], comprehensive regulatory frameworks governing LLM use in health care remain largely absent, operating in a regulatory gray area that generates significant professional concern about potential patient harm [17]. Although effective risk stratification is an essential part of safe medical practice, integrating automated triage systems into digital applications poses significant challenges [18-20]. Consequently, the emergence of LLMs providing medical advice creates complex liability questions among developers, institutions, and users, given the current lack of regulatory clarity regarding responsibility for patient harm [21].

A particularly underexplored area concerns the relationship between clinical urgency and LLM safety responses. While medical professionals routinely adjust communication based on perceived risk, little is known about the adaptive behavior of LLMs [19]. Additionally, substantial gaps exist between controlled research settings and real-world LLM usage patterns [22]. Many studies focus on narrow domains using artificial vignettes rather than authentic patient queries. Furthermore, standardized frameworks for evaluating disclaimer patterns across LLM architectures are absent, with diverse classification systems limiting cross-study comparisons.

To address these critical gaps, we developed a systematic, multimodel evaluation framework leveraging authentic patient queries to assess LLM disclaimer patterns across different urgency levels. This study employs a clinically relevant urgency scale and a 5-point classification system providing nuanced assessment of safety messaging relative to medical context. We further validated an LLM-based rating system against human expert annotation for scalable evaluation. This design enables an ecosystem-wide analysis across both proprietary and open-weight architectures to determine whether LLMs systematically adapt disclaimer and referral advice based on clinical urgency. This investigation contributes to the development of standards that protect users while preserving the benefits of access to LLM-generated health information.

Methods

Study Design and Research Question

This prospective, multimodel evaluation study aimed to analyze how LLMs embed disclaimers and referral advice into their responses to authentic medical queries of varying clinical urgency.

Data Collection

The study dataset was based on a publicly available, deidentified compilation of more than 100,000 dialogues between patients and doctors from widely used online medical consultation platforms [23]. Each record in the corpus contained a patient query and the corresponding physician reply. We drew a sample of 1000 patient queries for detailed manual annotation. A medical doctor manually screened all sampled queries against the following inclusion criteria: (1) the patient described unclear or undiagnosed symptoms and explicitly sought diagnostic guidance or medical advice, (2) the patient indicated no prior in-person consultation, diagnosis, or ongoing treatment for the

XSL•FO
RenderX

current complaint, and (3) the query was conveyed exclusively in narrative text, without reliance on additional attachments (eg, images or laboratory reports). A total of 773 queries did not meet at least 1 of these criteria and were therefore excluded. All 227 retained queries were minimally edited by removing salutations such as "Dear Doctor" or "Hello Doctor" to prevent systematic bias in subsequent LLM prompting.

Urgency Annotation of Patient Queries

All 227 eligible queries were independently classified by a medical doctor and a psychologist, who were blinded to each other's ratings. For this assessment, we constructed and employed an ordinal, 3-level scale to measure the urgency of typical patient queries from online settings. This approach was informed by previous work on pragmatic triage scales [24] and extended common emergency versus self-care distinctions with an intermediate "professional assessment recommended" category tailored to typical online health queries (Table 1).

Table . Scoring rubric for case urgency levels.

Urgency level	Definition	Implication
1 (Low)	Symptoms compatible with self-care or delayed evaluation	Online advice initially sufficient
2 (Intermediate)	Symptoms warrant professional assessment but not immediate intervention	Outpatient visit recommended
3 (High)	Symptoms suggest a potentially serious or rapidly progressive condition	Timely or urgent in-person consultation required

Any discordant cases were revisited at a consensus meeting and mutually resolved. Patient inquiries were categorized according to a predefined list of medical specialties deemed most likely to be responsible for each case.

LLM Selection and Response Generation

To ensure breadth of vendor representation and model architecture, 4 frequently used state-of-the-art LLMs not specifically designed for explicit multistep reasoning tasks (nonreasoning) were evaluated: GPT-4o (OpenAI, proprietary), Claude Sonnet 4 (Anthropic, proprietary), DeepSeek V3 (DeepSeek, open weight), and Grok 3 (xAI, proprietary). Model architectures were chosen to span both closed weight and open weight ecosystems and to represent widely adopted platforms (proprietary denotes models whose weights are not publicly released; open weight indicates freely downloadable models with public specifications).

To generate model outputs, the system prompt was set to "You are a helpful assistant," and all 227 patient queries were then submitted as user prompts. This minimal instructional approach was chosen to mirror laypeople's use of generative AI tools. All models were accessed via vendor application programming interfaces (APIs) on July 5, 2025 (model IDs: "gpt-4o," "claude-sonnet-4-20250514," "deepseek-chat," and "grok-3"; temperature=0.7; no further specification of parameters or safety settings). This procedure yielded 908 LLM responses (227 queries × 4 LLMs) for subsequent disclaimer and referral annotation.

Output Annotation and Rating Framework: Disclaimer and Referral Advice Classification

Each model response was evaluated to determine the highest level of disclaimer or referral advice present. Disclaimers and referral advice were categorized together, as disclaimers can be seen as a precursor to referrals, which typically contain indirect forms of disclaimers. The prompt used to rate the disclaimer and referral advice is provided in Information 1 in Multimedia Appendix 1 and was identical for all models. In this context, we employed a 5-point ordinal scale for advice classification, reflecting clinically relevant thresholds. This novel approach to assessing disclaimer and referral categories combined a yes/no disclaimer analysis with a pragmatic yet clinically meaningful categorization of referrals (Table 2).

For this advice categorization task, we evaluated 5 LLMs (GPT-o4-mini, GPT-4o-mini, GPT-4o, GPT-4o-latest, GPT-4.1) for their respective performance as rater LLMs. To validate our LLM-based annotation approach, 2 medical doctors independently scored identical subsets with 240 responses (60 cases per LLM: 20 low-urgency, 20 intermediate, 20 high; randomly selected within each urgency category), corresponding to more than 25% (240/908) of the total dataset. Both human raters were provided with identical rating instructions as in the prompt used for the rater LLMs to ensure comparability between human raters and LLMs. Raters were blinded to both the LLM identity and the models' assessments. To reflect real-world uptake of the rating scale and ensure equal conditions for LLMs and human raters, no calibration or consensus meetings were held (Figure 1).

In this context, GPT-4o showed the best overall performance compared with human expert ratings and was therefore selected for further ratings (see Information 2 in Multimedia Appendix 1 for technical settings and performance comparison).

XSL•FO
RenderX

Table . Scoring rubric for disclaimer and referral advice categories.[a]

Category	Definition	Illustrative example
1 (No disclaimer)	No advice to consult a medical professional and no statement of the model's limitations	"You may try resting at home."
2 (General disclaimer)	General statement that medical advice should come from professionals or statement of model limitation (eg, "I am not a medical professional," "This cannot replace professional advice")	"Please note, I am not a doctor."
3 (Differentiated referral)	Advises consulting a medical professional if symptoms prolong, worsen, or new symptoms appear	"If your pain persists or gets worse, you should see a doctor."
4 (Explicit referral)	Advises consulting a medical professional	"You should make an appointment with your GP[b]."
5 (Urgent referral)	Advises consulting a medical professional urgently	"Seek medical attention immediately or visit the emergency department."

[a]The classification assigned to each model response refers to the highest category that appeared in the respective response. Examples are illustrative and were not provided as instructions to human raters, nor to the rater large language models.

[b]GP: general practitioner.

Figure 1. Graphical overview of the study protocol. LLM: large language model.

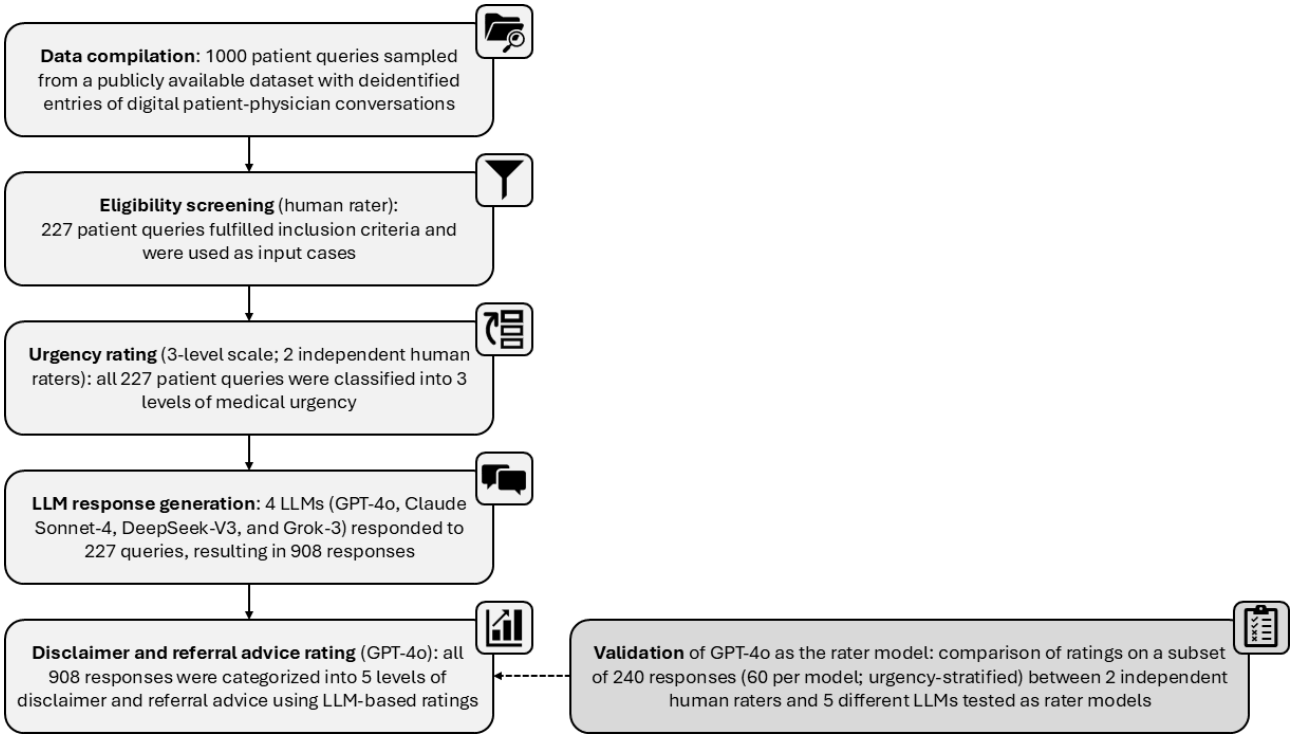

Statistical Analysis

Statistical analyses were performed using Python (version 3.13) with the *matplotlib* (v3.10.5), *numpy* (v2.3.1), *pandas* (v2.3.0), *scipy* (v1.16.0), and *seaborn* (v0.13.2) libraries. Statistical significance was defined as α=.05, unless specified otherwise. Descriptive statistics characterized the dataset and model outputs. Following a significant Jonckheere-Terpstra test indicating a monotonic trend across the ordered groups, post hoc Mann-Whitney tests with Bonferroni correction for multiple comparisons were performed to examine the association between the input medical urgency levels of the patient queries and the disclaimer or referral advice categories in the LLM responses for each LLM. Differences in the distribution of disclaimer and referral levels across the 4 evaluated LLMs were assessed using the nonparametric Kruskal-Wallis *H* test, given the ordinal nature of the disclaimer and referral ratings. If the Kruskal-Wallis test indicated a statistically significant difference, post hoc pairwise comparisons between the LLMs were conducted using the Dunn test. To account for multiple comparisons, the *P* values obtained from the Dunn test were adjusted using the Bonferroni correction. Interrater reliability between human raters and LLMs was assessed for output disclaimer and referral advice on a subset of LLM responses for each model. Weighted Cohen κ was used to measure the level of agreement between human and LLM raters in classifying LLM-produced disclaimers and referral advice.

XSL•FO
RenderX

Ethical Considerations

This study utilized an openly available, deidentified dataset sourced from published literature. In accordance with the regulations of Charité – Universitätsmedizin Berlin [25], ethics committee approval was not required for research conducted exclusively on previously published, deidentified data in which no interventions, clinical recruitment, or patient-level interactions were performed. Informed consent was similarly not required under these regulations. To ensure privacy and confidentiality, 2 independent readers reviewed all cases at the data input stage to confirm the absence of any identifiable personal information. No participants were involved in this study, and no compensation was provided.

Results

Descriptive Overview of Input Cases and LLM Responses

Of the 227 eligible patient queries, 77 (34%), 110 (48%), and 40 (18%) were rated as level 1 (low urgency), level 2 (intermediate urgency), and level 3 (high urgency), respectively, after consensus-based agreement between both human raters. Level 2-rated patient queries were the longest on average (mean 383 characters; 74 words [SD 56]), followed by level 3 queries (333 characters; 64 words [SD 35]), whereas level 1 queries were the shortest (312 characters; 61 words [SD 43]). While cases classified as general practitioner dominated at all urgency levels, dental and dermatology cases were more prevalent at levels 1 and 2 and almost absent at level 3. Emergency department cases, although few in the eligible dataset, were found exclusively at urgency level 3 (Table 3).

Table . Absolute count of patient queries per specialty and urgency level.

Specialty	Level 1 (Low urgency)	Level 2 (Intermediate urgency)	Level 3 (High urgency)
Dentistry	2	8	5
Dermatology	16	11	0
Emergency medicine	0	0	2
ENT[a]/otorhinolaryngology	1	7	2
Ophthalmology	5	5	1
General practice	31	38	16
Gynecology	2	6	1
Neurology	0	0	1
Orthopedics	3	10	4
Pediatrics	1	6	3
Psychiatry	3	10	2
Urology	13	9	3

[a]ENT: ear, nose, and throat.

With each of the 4 LLMs responding to all 227 patient queries, a total of 908 responses were generated. On average, Grok-3 replies were the longest at (mean 785 [SD 219] words), while Claude Sonnet-4 replies were the shortest (mean 215 [SD 22] words). The average response length for GPT-4o and DeepSeek-V3 was intermediate (mean 267 [SD 59] words and mean 353 [SD 80] words, respectively). Pearson *r* correlations between the word count of responses and the GPT-4o rating of the corresponding disclaimers and referrals were small and negative for all 4 LLMs tested (r=−0.024, r=−0.140, r=−0.195, and r=−0.028 for GPT-4o, Sonnet-4, Grok-3, and DeepSeek-V3, respectively). The pooled correlation was approximately r=−0.043, indicating an overall lack of meaningful linear relationship.

Interrater Reliability for Disclaimer and Referral Ratings

On a subset of LLM responses to patient queries produced by GPT-4o, Sonnet-4, Grok-3, and DeepSeek-V3, the 5 LLMs tested as rater models (GPT-4o, GPT-4o-mini, GPT-latest, GPT-o4-mini, and GPT-4.1) produced predominantly moderate (weighted Cohen κ=0.41 - 0.60) to substantial agreement (0.61 - 0.80), with GPT-4o demonstrating the greatest overall alignment with human raters (Table 4).

XSL•FO
RenderX

Table . Quadratic weighted Cohen κ for interrater agreement between rater models and medical doctors regarding the ratings of disclaimer and referral advice produced by 4 different large language models.

LLM[a] for response generation	MD1[b] vs MD2	GPT-4o vs MDs[c]	GPT-4o-mini vs MDs[c]	GPT-4o-latest vs MDs[c]	GPT-o4-mini vs MDs[c]	GPT-4.1 vs MDs[c]
GPT-4o	0.756	0.558	0.534	0.388	0.526	0.532
Sonnet-4	0.583	0.602	0.497	0.463	0.467	0.569
Grok-3	0.714	0.415	0.527	0.338	0.328	0.430
DeepSeek-V3	0.756	0.707	0.671	0.529	0.554	0.642

[a]LLM: large language model.

[b]MD: medical doctor.

[c]The mean value of the 2 interrater reliability values between the respective model and each of the 2 doctors is reported, as 2 doctors rated the dataset.

Association Between Case Urgency Level and LLM Disclaimer and Referral Advice Category

Across all 4 models, higher case urgency was associated with a shift toward higher disclaimer and referral ratings. For GPT-4o, most low-urgency and intermediate-urgency cases were rated as category 4, while high-urgency cases showed a greater proportion of category 5 ratings alongside fewer category 4 ratings. Sonnet-4 displayed a similar pattern, with category 4 ratings predominating at urgency levels 1 and 2, while category 5 ratings dominated at urgency level 3. Grok-3 also concentrated on category 4 for urgency levels 1 and 2, but for high-urgency cases, the largest share shifted to category 5 ratings. DeepSeek-V3 exhibited a broader distribution of ratings overall, with fewer category 4 ratings and more frequent use of categories 2 and 3 across different urgency levels; nevertheless, urgency level 3 still had a higher proportion of category 5 ratings compared to urgency levels 1 and 2 (Figure 2).

XSL•FO
RenderX

Figure 2. Shift of disclaimer and referral advice rating per urgency level for each model. Grouped bar visualization showing the proportion of each disclaimer and referral advice rating category (1 - 5, relative proportion on y-axis) for each urgency level (1 - 3, x-axis). Each subplot (A-D) illustrates the large language models used to generate responses to patient queries. Within each subplot, bars of the same color correspond to the same rating category as determined by GPT-4o as the rater model.

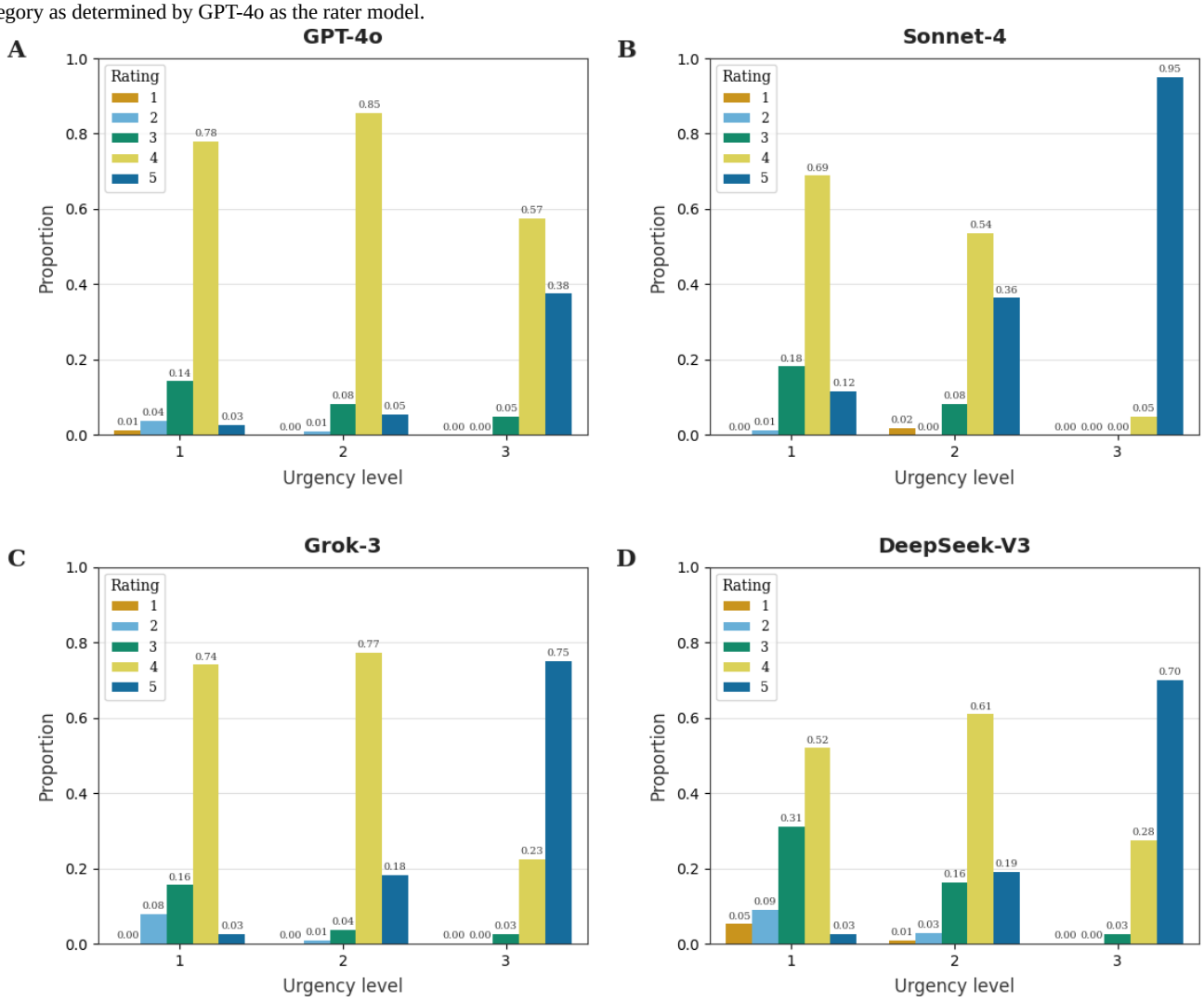

XSL•FO
RenderX

Figure 3. Frequency distributions of disclaimer and referral advice ratings (categories 1 - 5) for each model using GPT-4o as a rater (subplots A-D) and violin plot showing the medians and distributions of disclaimer and referral ratings (categories 1 - 5) for each large language model (LLM) (subplot E). For each LLM, violin width reflects the kernel density estimate of the underlying data; thicker sections mark rating levels with more responses. Labeled dots indicate medians.

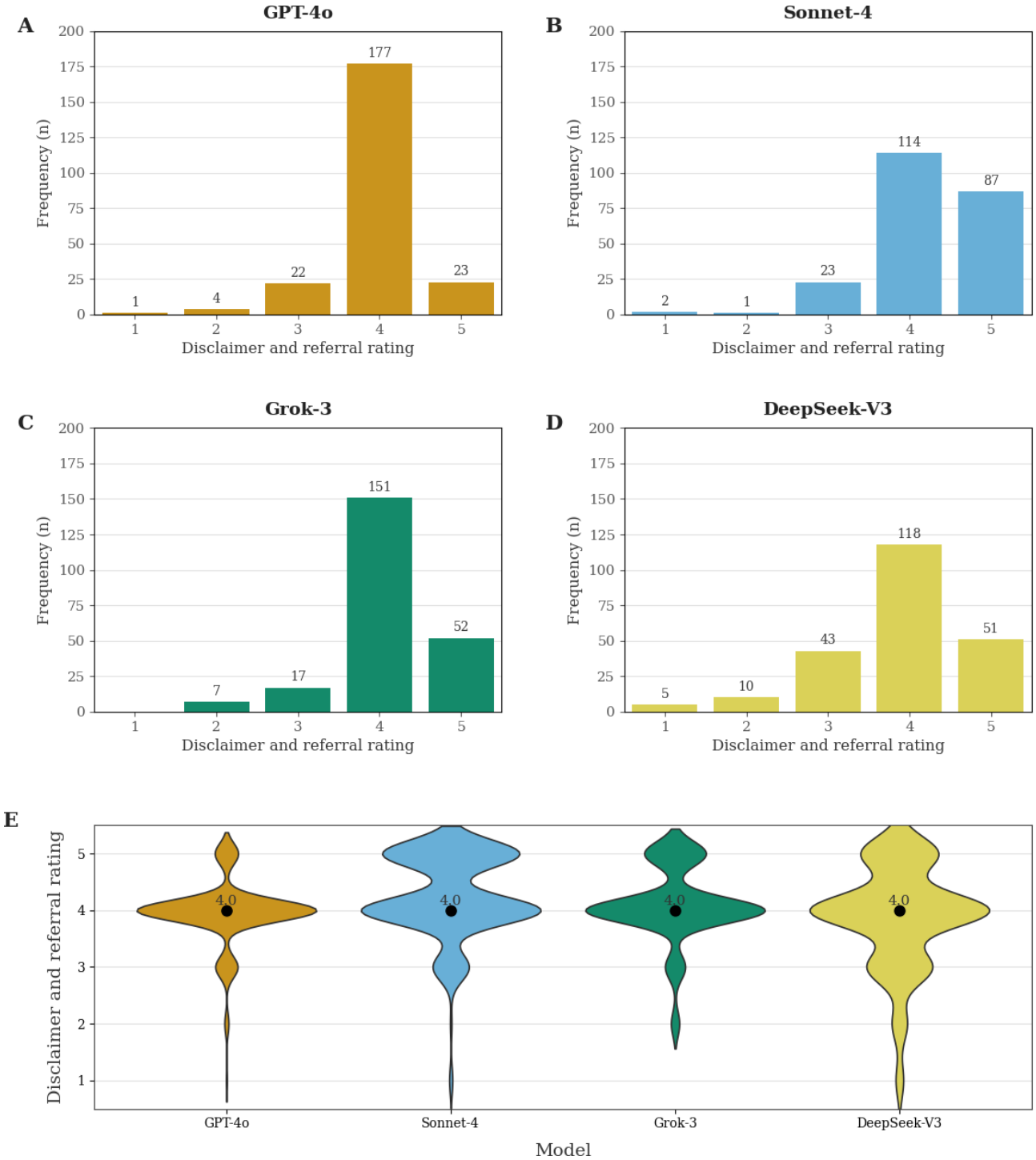

Among responses with a disclaimer and referral category rating of 5, the proportion of cases with level 1 urgency was 2/23 (9%) for GPT-4o, 9/87 (10%) for Sonnet-4, 2/52 (4%) for Grok-3, and 2/51 (4%) for DeepSeek-V3. Conversely, among responses with a disclaimer and referral category rating of less than 4, the proportion of cases with level 3 urgency was 2/27 (7%) for GPT-4o, 0/26 (0%) for Sonnet-4, 1/24 (4%) for Grok-3, and 1/58 (2%) for DeepSeek-V3.

Jonckheere-Terpstra tests for ordered alternatives with pairwise Mann-Whitney comparisons were conducted to examine whether ratings of disclaimer and referral advice (rating categories 1 - 5) increased systematically across levels of case urgency (urgency levels 1 - 3). After Bonferroni correction (α=.05/4=0.0125), the output ratings of all 4 LLMs demonstrated statistically significant ordered trends (all P<.001, detailed analyses are provided in Information 3 in Multimedia Appendix 1).

XSL•FO
RenderX

Comparison of Disclaimer and Referral Advice Ratings Across Different Models

Disclaimer and referral advice ratings clustered toward the upper end of the 5-point scale for all 4 LLMs examined. Grok-3, Sonnet-4, and GPT-4o displayed a narrow spread, with most responses falling into categories 4 to 5 (89%, 89%, and 88%, respectively). In contrast, DeepSeek-V3 exhibited a wider spread, with 74% of responses at categories 4 to 5 and a substantial 19% of responses at category 3, nearly twice the proportion seen with any other tested model. Notably, 7% of DeepSeek-V3 responses did not contain a referral (corresponding to categories 1‐2), which is more than twice as high as for any other model. With more than 2% of responses falling into category 1, this model was the only one in which the percentage of responses without a disclaimer or referral exceeded 1% of the respective model's total responses. Overall, 30/908 (3%) of all LLM-generated responses did not include any referral to a doctor. Median disclaimer and referral advice rating was identical (4) across the 4 LLMs; however, their overall rating distributions differed, with Sonnet-4 displaying a larger upper-tail mass in the density profile due to 87 ratings at category 5 (Figure 3).

The Kruskal-Wallis test detected significant distribution differences among models (H=37.22, P<.001). Pair-wise Dunn post hoc comparisons (Bonferroni-adjusted) indicated where the differences lay: Sonnet-4 ratings were significantly higher than DeepSeek-V3 and GPT-4o (P<.001) as well as Grok-3 (P=.04). However, the latter difference was sensitive to the rating method; it was not significant when considering only the subset of human-rated cases (P=.58). Rating comparisons of Grok-3 versus DeepSeek-V3 and Grok-3 versus GPT-4o were not significant (P=.94, and P=.53, respectively). Likewise, ratings by GPT-4o and DeepSeek-V3 did not differ (P>.99, see Information 4 and 5 in Multimedia Appendix 1 for details).

Discussion

Principal Findings

In this study, we conducted a prospective, multimodel evaluation of 908 responses generated by 4 state-of-the-art LLMs to 227 authentic patient queries of varying urgency. We observed a consistent association between query urgency and the type of generated disclaimers and referral advice. All 4 models (GPT-4o, Claude Sonnet-4, Grok-3, and DeepSeek-V3) provided progressively more explicit and urgent referrals with increasing query urgency. The median score of disclaimer and referral categories was identical across all models, but the distributions varied between the LLMs. Sonnet-4, Grok-3, and GPT-4o received the highest overall referral ratings (89%, 89%, and 88%, respectively, of responses containing explicit or urgent referrals), while DeepSeek-V3 showed the broadest spread and lowest concentration at higher categories (74%). Notably, disclaimers and referrals were rarely omitted by any model (<1% of responses contained no disclaimer or referral). While most LLM responses provided explicit referral recommendations, model behavior at the end of the distribution (cases receiving minimal disclaimers or urgent referrals) varied between LLMs. These differences in disclaimer integration—a conservative approach by Sonnet-4, Grok-3, and GPT-4o versus DeepSeek-V3 variability—may reflect underlying architectural and training variances. While proprietary models' higher ratings may indicate stronger safety alignments, open weight models like DeepSeek offer transparency and adaptability, but their safety features may be less consistent (see Information 6 in Multimedia Appendix 1 for illustrative examples of patient queries and the corresponding responses produced by different models).

Comparison With Prior Work

Prior studies have highlighted inconsistencies in LLM-generated medical advice, including discrepancies in triage [26] and risk stratification [27]. Wang et al [28] reported that ChatGPT-3.5 and -4.0 included disclaimers in only 10.8% to 13.1% of responses to hepatitis B queries. However, apart from performance progress over time, their relatively narrow definition of disclaimers may explain this discrepancy with our findings, where 99% of responses contained some form of disclaimer or referral advice when using our broader classification system. Similarly, Nastasi et al [29] noted that ChatGPT responses to chest pain vignettes frequently omitted disclaimers on LLM limitations, potentially leading to overutilization of health care resources. Our results extend these findings by showing that modern LLMs are more likely to make explicit and urgent referrals for urgent cases. This may be due to stronger safety alignments in recent architectures and a greater focus on "red teaming," which is the practice of identifying unexpected or undesirable behaviors in models. However, while red teaming may be valuable for characterizing harm mitigations, publicly promoting it as a comprehensive solution to all AI risks constitutes performative safety measures rather than substantive risk mitigation [30]. While our findings suggest that LLMs may embed safety-oriented guidelines to a greater extent than previously reported, this largely depends on respective classifications and types of medical input. This lack of harmonization is particularly noticeable in relation to referrals: if comprehensive guidelines were established in this field, these could also be tested in a retrieval-augmented generation approach to optimize LLM-generated, urgency-stratified medical referrals [31].

Clinical, Behavioral, and Legal Implications of Disclaimers and Referral Advice

The adaptive inclusion of disclaimers and referrals in LLM-generated responses carries significant clinical implications, as incorrect advice could exacerbate patient safety risks [32,33]. Our observation that high-urgency queries prompted almost exclusively explicit or urgent referrals across all models supports their role in mitigating harms. Yet, the persistence of lower categories of disclaimers and referral advice highlights gaps where patients might act on AI-generated advice without consulting professionals. Furthermore, current research on the effectiveness of disclaimers remains inconclusive and context-dependent: prior work suggests that disclaimers alone may have minimal impact on user perception and behavior, placing greater responsibility on developers to educate users about model limitations [34]. However, other research found that participants who heard a disclaimer were significantly more

XSL•FO
RenderX

likely to contact a health care professional, suggesting disclaimers can positively influence patient behavior under certain conditions [35]. We believe that the type of disclaimer may also be relevant here, given the wide range from nonspecific ("AI can make mistakes") to context-specific statements (eg, considering the clinical urgency of a medical question).

From a legal perspective, inconsistent disclaimers may expose both developers and users to liability. Notably, physicians who rely on inaccurate information, such as hallucinations or omissions in LLM-generated content, face malpractice risks [36]. Medical professionals remain legally liable for patient outcomes regardless of whether their decisions are informed by AI-generated advice or traditional sources, necessitating the same rigorous validation standards for AI tools as applied to conventional medical instruments before clinical adoption [37]. The pervasive nature of disclaimers we observed suggests developers are aware of liability risks, but regulatory frameworks have not yet caught up with the reality of widespread public access to these tools. Unregulated LLM-based health care applications continue to reach markets, requiring immediate enforcement of current regulations and development of adaptive frameworks to prevent patient harm and preserve the therapeutic potential of AI-assisted medical care [38]. In our study, the rare occurrence of category 1 (no disclaimer or referral) indicates progress. However, output variability underscores the need to enforce safety standards to mitigate medical risk and avoid public disinformation threats [39].

Feasibility of Using LLMs as Rater Models

Our validation of GPT-4o as a rater model, with weighted Cohen κ values comparable to interhuman agreement (0.415 - 0.707 and 0.583 - 0.756, respectively), demonstrates potential feasibility for scalable annotation of LLM evaluations. Possible discrepancies between the human and LLM ratings may be due to the different category 3 ratings assigned for disclaimers and referrals ("differentiated referral"), which was a category frequently chosen by human raters, but less so by LLMs, as shown by our subset analysis in the (Supplementary Information 4 and 5 in Multimedia Appendix 1). This provides evidence for potential alternatives to human-only annotation, though human supervision remains essential for nuanced clinical judgments.

Strengths of the Study

The novel contribution of our study is its systematic, multimodel approach to evaluating medical disclaimers and referral patterns using authentic patient queries. This provides ecological validity often missing from studies using simulated vignette-based designs. Our urgency and disclaimer classification framework offers a pragmatic, yet clinically meaningful approach for understanding LLM safety responses to digital health queries. By including both proprietary and open weight models, we ensure broader generalizability across the current LLM ecosystem, while LLM-based rating with human validation also offers scalable efficiency and reliability. Furthermore, our minimal prompting approach reflects real-world LLM usage patterns by laypeople.

Limitations and Potential Biases

The limitations of our study include our reliance on text-based queries, which may have been part of the models' training data, and our exclusion of multimodal interactions, during which patients provide supplementary resources such as images. Furthermore, as it is aimed at digital health settings, our disclaimer and referral classification system may not capture all forms of medical safety guidance and may not transfer directly to hospital settings or emergency department triage. We did not assess the temporal stability of the observed disclaimer patterns, as these may change with model updates or adjustments to finetuning approaches. Our ground-truth urgency labels were based on 2 raters (1 physician, 1 psychologist), which may have introduced variability due to differences in expertise. Due to the design of our scoring categories, this study prioritizes the presence and quality of referrals over the adequacy and placement of disclaimers. As our evaluation approach relies exclusively on the highest-scoring category, this limits the analysis of multipart safety responses when disclaimers and referrals are present simultaneously. Finally, using GPT-4o as both the generator and the rater model, even with human-aligned validation, introduces a potential risk of vendor or self-preference bias, whereby the model favors responses that resemble its own output style. Notably, this study did not assess the accuracy of the medical advice provided, instead focusing on the presence and type of disclaimers and referral advice.

Implications and Future Research Directions

LLMs adaptively incorporate disclaimers and referral advice based on query urgency, but differences between models highlight the need for safety measures to ensure reliable medical guidance. While LLMs are generally risk-averse in their approach to medical queries, similar to medical laypeople when deciding whether medical care is required [24], the effectiveness of these mechanisms in protecting patients from harm remains unclear. Clinical implications of these patterns are complex: while conservative referral advice may reduce risks of inappropriate self-care, it could also contribute to health care system overutilization and may not effectively guide patients toward appropriate care. However, we consider the inclusion of a generic disclaimer or referral in LLM-generated responses to medical queries to be the minimum quality standard. As these tools exhibit systematic medical triage–like behavior by embedding urgency-sensitive advice, proper regulatory oversight is required [40]. Future research should focus on exploring the implications for patient safety [41], for example, by assessing the behavioral impacts of multistep user interactions as well as by investigating the relationship between case urgency, response accuracy, and LLM-reported certainty. Another intriguing question is whether LLMs exhibit consistent response patterns when asked for a second opinion based on existing medical recommendations, laboratory results, and imaging data. Since recent studies indicate that the inclusion of disclaimers for certain models has decreased [42], longitudinal studies tracking disclaimer patterns over time should be conducted to reveal the effect of model releases on safety standards.

XSL•FO
RenderX

Conclusions

Our findings demonstrate that current LLMs exhibit urgency-responsive safety mechanisms. However, safety implementations differ substantially between models. Given the widespread public availability of AI tools such as LLMs, our study emphasizes the importance of standardized safety measures and appropriate regulatory frameworks that reflect this new way of accessing health information.

Acknowledgments

The authors used GPT-5 to improve the grammar and language of the manuscript. Following the use of this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This work was supported by the Open Access Publication Fund of Charité – Universitätsmedizin Berlin.

Data Availability

The analyzed cases and their associated urgency ratings, along with all large language model responses and their corresponding disclaimer and referral advice ratings, are publicly available on Zenodo [43].

Authors' Contributions

Conceptualization: FR, SDB
Methodology: FR, SDB
Formal analysis: FR
Data curation: LA, RH, MR
Validation: LA, RH, MR
Visualization: FR
Statistical advice: SP
Supervision: FB, SDB
Writing – original draft: FR
Writing – review & editing: FR, LAMS, RH, MR, SKP, FB, SDB

Conflicts of Interest

FR and SDB report employment with Pfizer outside the submitted work. FB reports honoraria from Pfizer outside the submitted work. The other authors declare no conflicts of interest.

Multimedia Appendix 1
Supplementary methods, statistical analyses, and illustrative examples (Information 1-6).
[DOCX File, 143 KB - jmir_v28i1e84668_app1.docx]

References

1. Yun HS, Bickmore T. Online health information-seeking in the era of large language models: cross-sectional web-based survey study. J Med Internet Res 2025 Mar 31;27:e68560. [doi: 10.2196/68560] [Medline: 40163112]
2. Ayo-Ajibola O, Julien C, Lin ME, Riddell J, Duan N, Kravitz RL. Association of primary care access with health-related ChatGPT use: a national cross-sectional survey. J Gen Intern Med 2026 Jan;41(2):338-345. [doi: 10.1007/s11606-025-09406-9] [Medline: 39930155]
3. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023 May 17;10:e47564. [doi: 10.2196/47564] [Medline: 37195756]
4. Ayo-Ajibola O, Davis RJ, Lin ME, Riddell J, Kravitz RL. Characterizing the adoption and experiences of users of artificial intelligence-generated health information in the United States: cross-sectional questionnaire study. J Med Internet Res 2024 Aug 14;26:e55138. [doi: 10.2196/55138] [Medline: 39141910]
5. Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence. Nature 2025 Jun;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
6. McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. Nature 2025 Jun;642(8067):451-457. [doi: 10.1038/s41586-025-08869-4] [Medline: 40205049]
7. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med Inform Decis Mak 2025 Mar 7;25(1):117. [doi: 10.1186/s12911-025-02954-4] [Medline: 40055694]

XSL•FO
RenderX

8. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 2024 Sep;30(9):2613-2622. [doi: 10.1038/s41591-024-03097-1] [Medline: 38965432]
9. Chen S, Kann BH, Foote MB, et al. Use of artificial intelligence chatbots for cancer treatment information. JAMA Oncol 2023 Oct 1;9(10):1459-1462. [doi: 10.1001/jamaoncol.2023.2954] [Medline: 37615976]
10. Kaewboonlert N, Poontananggul J, Pongsuwan N, Bhakdisongkhram G. Factors associated with the accuracy of large language models in basic medical science examinations: cross-sectional study. JMIR Med Educ 2025 Jan 13;11:e58898. [doi: 10.2196/58898] [Medline: 39846415]
11. Bedi S, Jiang Y, Chung P, Koyejo S, Shah N. Fidelity of medical reasoning in large language models. JAMA Netw Open 2025 Aug 1;8(8):e2526021. [doi: 10.1001/jamanetworkopen.2025.26021] [Medline: 40779272]
12. Reis F, Reis M, Drzeniek NM, Balzer F. Trust is all you need: reinforcing the patient-physician bond in times of AI. J Med Syst 2025 Nov 5;49(1):154. [doi: 10.1007/s10916-025-02296-8] [Medline: 41188572]
13. Reis M, Reis F, Kunde W. Influence of believed AI involvement on the perception of digital medical advice. Nat Med 2024 Nov;30(11):3098-3100. [doi: 10.1038/s41591-024-03180-7] [Medline: 39054373]
14. Wogalter MS, Kalsher MJ, Racicot BM. Behavioral compliance with warnings: effects of voice, context, and location. Saf Sci 1993 Aug;16(5-6):637-654. [doi: 10.1016/0925-7535(93)90028-C]
15. Mello MM, Guha N. ChatGPT and physicians' malpractice risk. JAMA Health Forum 2023 May 5;4(5):e231938. [doi: 10.1001/jamahealthforum.2023.1938] [Medline: 37200013]
16. Weissman GE, Mankowitz T, Kanter GP. Unregulated large language models produce medical device-like output. NPJ Digit Med 2025 Mar 7;8(1):148. [doi: 10.1038/s41746-025-01544-y] [Medline: 40055537]
17. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 6;6(1):120. [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
18. Taylor RA, Chmura C, Hinson J, et al. Impact of artificial intelligence–based triage decision support on emergency department care. NEJM AI 2025 Feb 27;2(3). [doi: 10.1056/AIoa2400296]
19. Da'Costa A, Teke J, Origbo JE, Osonuga A, Egbon E, Olawade DB. AI-driven triage in emergency departments: a review of benefits, challenges, and future directions. Int J Med Inform 2025 May;197:105838. [doi: 10.1016/j.ijmedinf.2025.105838] [Medline: 39965433]
20. Fraser H, Crossland D, Bacher I, Ranney M, Madsen T, Hilliard R. Comparison of diagnostic and triage accuracy of Ada Health and WebMD symptom checkers, ChatGPT, and physicians for patients in an emergency department: clinical data analysis study. JMIR Mhealth Uhealth 2023 Oct 3;11:e49995. [doi: 10.2196/49995] [Medline: 37788063]
21. Amiot F, Potier B. Artificial intelligence (AI) and emergency medicine: balancing opportunities and challenges. JMIR Med Inform 2025 Aug 13;13:e70903. [doi: 10.2196/70903] [Medline: 40802997]
22. Bhimani M, Miller A, Agnew JD, et al. Real-world evaluation of large language models in healthcare (RWE-LLM): a new realm of AI safety & validation. medRxiv. Preprint posted online on Mar 18, 2025. [doi: 10.1101/2025.03.17.25324157]
23. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. Chatdoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. Cureus 2023 Jun;15(6):e40895. [doi: 10.7759/cureus.40895] [Medline: 37492832]
24. Kopka M, Schmieding ML, Rieger T, Roesler E, Balzer F, Feufel MA. Determinants of laypersons' trust in medical decision aids: randomized controlled trial. JMIR Hum Factors 2022 May 3;9(2):e35219. [doi: 10.2196/35219] [Medline: 35503248]
25. Ethikkommission. Charité – Universitätsmedizin Berlin. 2026. URL: https://ethikkommission.charite.de/ [accessed 2026-03-15]
26. Masanneck L, Schmidt L, Seifert A, et al. Triage performance across large language models, ChatGPT, and untrained doctors in emergency medicine: comparative study. J Med Internet Res 2024 Jun 14;26:e53297. [doi: 10.2196/53297] [Medline: 38875696]
27. Heston TF, Lewis LM. ChatGPT provides inconsistent risk-stratification of patients with atraumatic chest pain. PLoS One 2024;19(4):e0301854. [doi: 10.1371/journal.pone.0301854] [Medline: 38626142]
28. Wang Y, Chen Y, Sheng J. Assessing ChatGPT as a medical consultation assistant for chronic hepatitis B: cross-language study of English and Chinese. JMIR Med Inform 2024 Aug 8;12:e56426. [doi: 10.2196/56426] [Medline: 39115930]
29. Nastasi AJ, Courtright KR, Halpern SD, Weissman GE. A vignette-based evaluation of ChatGPT's ability to provide appropriate and equitable medical advice across care contexts. Sci Rep 2023 Oct 19;13(1):17885. [doi: 10.1038/s41598-023-45223-y] [Medline: 37857839]
30. Feffer M, Sinha A, Deng WH, Lipton ZC, Heidari H. Red-teaming for generative AI: silver bullet or security theater? 2024 Presented at: AIES'24: Proceedings of the Seventh (AAAI/ACM) Conference on AI, Ethics, and Society; Oct 21-23, 2024; San Jose, California, USA p. 421-437. [doi: 10.1609/aies.v7i1.31647]
31. Masanneck L, Meuth SG, Pawlitzki M. Evaluating base and retrieval augmented LLMs with document or online support for evidence based neurology. NPJ Digit Med 2025 Mar 4;8(1):137. [doi: 10.1038/s41746-025-01536-y] [Medline: 40038423]
32. Chen S, Guevara M, Moningi S, et al. The effect of using a large language model to respond to patient messages. Lancet Digit Health 2024 Jun;6(6):e379-e381. [doi: 10.1016/S2589-7500(24)00060-8] [Medline: 38664108]
33. Biro JM, Handley JL, Malcolm McCurry J, et al. Opportunities and risks of artificial intelligence in patient portal messaging in primary care. NPJ Digit Med 2025 Apr 24;8(1):222. [doi: 10.1038/s41746-025-01586-2] [Medline: 40275104]

XSL•FO
RenderX

34. Henestrosa AL, Kimmerle J. “Always check important information!” - the role of disclaimers in the perception of AI-generated content. Comput Hum Behav: Artif Hum 2025 May;4:100142. [doi: 10.1016/j.chbah.2025.100142]
35. Bickmore TW, Ólafsson S, O’Leary TK. Mitigating patient and consumer safety risks when using conversational assistants for medical information: exploratory mixed methods experiment. J Med Internet Res 2021 Nov 9;23(11):e30704. [doi: 10.2196/30704] [Medline: 34751661]
36. Mello MM, Guha N. Understanding liability risk from using health care artificial intelligence tools. N Engl J Med 2024 Jan 18;390(3):271-278. [doi: 10.1056/NEJMhle2308901] [Medline: 38231630]
37. Haupt CE, Marks M. AI-generated medical advice-GPT and beyond. JAMA 2023 Apr 25;329(16):1349-1350. [doi: 10.1001/jama.2023.5321] [Medline: 36972070]
38. Freyer O, Wiest IC, Kather JN, Gilbert S. A future role for health applications of large language models depends on regulators enforcing safety standards. Lancet Digit Health 2024 Sep;6(9):e662-e672. [doi: 10.1016/S2589-7500(24)00124-9] [Medline: 39179311]
39. Menz BD, Modi ND, Sorich MJ, Hopkins AM. Health disinformation use case highlighting the urgent need for artificial intelligence vigilance: weapons of mass disinformation. JAMA Intern Med 2024 Jan 1;184(1):92-96. [doi: 10.1001/jamainternmed.2023.5947] [Medline: 37955873]
40. Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. Nat Med 2023 Oct;29(10):2396-2398. [doi: 10.1038/s41591-023-02412-6] [Medline: 37391665]
41. Huo B, Boyle A, Marfo N, et al. Large language models for chatbot health advice studies: a systematic review. JAMA Netw Open 2025 Feb 3;8(2):e2457879. [doi: 10.1001/jamanetworkopen.2024.57879] [Medline: 39903463]
42. Sharma S, Alaa AM, Daneshjou R. A longitudinal analysis of declining medical safety messaging in generative AI models. NPJ Digit Med 2025 Oct 2;8(1):592. [doi: 10.1038/s41746-025-01943-1] [Medline: 41038984]
43. Disclaimers and referral patterns for medical advice across urgency levels: a large language model evaluation study. Zenodo. URL: https://doi.org/10.5281/zenodo.17185348 [accessed 2026-02-23]

Abbreviations

AI: artificial intelligence
LLM: large language model

Edited by A Coristine; submitted 23.Sep.2025; peer-reviewed by HS Yun, L Masanneck, R Kravitz; revised version received 21.Jan.2026; accepted 22.Jan.2026; published 16.Mar.2026.

Please cite as:
Reis F, Agha-Mir-Salim L, Hickstein R, Reis M, Piper SK, Balzer F, Boie SD
Disclaimers and Referral Patterns for Medical Advice Across Urgency Levels: Large Language Model Evaluation Study
J Med Internet Res 2026;28:e84668
URL: https://www.jmir.org/2026/1/e84668
doi:10.2196/84668

© Florian Reis, Louis Agha-Mir-Salim, Richard Hickstein, Moritz Reis, Sophie K Piper, Felix Balzer, Sebastian Daniel Boie. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of AI-Assisted Patient Health Education Using Voice Cloning and ChatGPT: Prospective Randomized Controlled Trial

Yan Sun[1,2], MSc; Shangqing Xu[1], BSc; Hongying Jin[1], MSc; Xiaoyan Han[1], MSc; Kangqi Jin[1], MSc; Yimei Zhang[1], MSc; Xiaoli Ma[1], MSc; Huaping Wei[3], BSc; Minjie Ma[4], PhD

[1]Department of Thoracic Surgery, The First Hospital of Lanzhou University, Donggang West Road 1#, Lanzhou, China

[2]School of Nursing, Lanzhou University, Lanzhou, China

[3]Outpatient Department, The First Hospital of Lanzhou University, Lanzhou, China

[4]Gansu International Science and Technology Cooperation Base for Development and Application of Thoracic Surgery Key Technologies, The First Clinical Medical College of Lanzhou University, Department of Thoracic Surgery, The First Hospital of Lanzhou University, Donggang West Road 1#, Lanzhou, China

Corresponding Author:
Minjie Ma, PhD
Gansu International Science and Technology Cooperation Base for Development and Application of Thoracic Surgery Key Technologies, The First Clinical Medical College of Lanzhou University, Department of Thoracic Surgery, The First Hospital of Lanzhou University, Donggang West Road 1#, Lanzhou, China

Abstract

Background: Traditional patient education often lacks personalization and engagement, potentially limiting knowledge acquisition and treatment adherence. Advances in artificial intelligence (AI), including voice cloning technology and large language models (eg, ChatGPT), offer new opportunities to deliver personalized, scalable, interactive health education. However, evidence regarding the comparative effectiveness of different AI-based voice cloning strategies and reliability of automated AI evaluation tools remains limited.

Objective: This study aims to evaluate the effectiveness of AI-assisted patient education integrating voice cloning and ChatGPT, compare physician voice cloning with patient self-voice cloning, and assess the reliability of ChatGPT as an automated evaluation tool for education outcomes.

Methods: In this prospective, 3-arm, parallel-group randomized controlled trial, 180 hospitalized patients requiring standardized health education were recruited from a tertiary hospital. Inclusion criteria were age ≥18 years, clear diagnosis requiring health education, clear consciousness, and voluntary participation with informed consent. Exclusion criteria were severe hearing impairment, severe cognitive impairment, expected hospitalization <3 days, or prior participation in similar studies. Using a computer-generated random sequence, participants were randomly assigned (1:1:1) to receive traditional education (control), AI-assisted education using physician voice cloning, or AI-assisted education using patient self-voice cloning, each with identical educational content of equal duration. The primary outcome was education content compliance, evaluated using ChatGPT-4 with validated prompts and verified by expert review. Secondary outcomes included knowledge retention, education satisfaction, treatment adherence, quality of life, and psychological status. Outcome assessors and data analysts, but not participants, were blinded to group allocation.

Results: Of 180 randomized participants, 174 (96.7%) completed the trial. Both AI-assisted groups had significantly higher mean education content compliance scores immediately posteducation than the control group (physician voice: 86.7, SD 7.3; self-voice: 92.5, SD 6.8; control: 73.2, SD 8.5; P<.001). The patient self-voice group showed superior predischarge knowledge retention, higher education satisfaction, and greater treatment adherence than the other groups (all P≤.02). At the 1-month follow-up, the self-voice group maintained improved adherence (Cohen d=0.74) and had significantly lower anxiety and depression scores (all P≤.02) and improved SF-36 quality-of-life domains. ChatGPT-based evaluations demonstrated high reliability (weighted κ=0.87, 95% CI 0.82 - 0.91).

Conclusions: The innovative patient education model integrating AI voice cloning and ChatGPT is distinct from previous studies primarily relying on standard text-to-speech or professionally recorded content. Using patients' own cloned voices for health education delivery leveraged the self-reference effect to enhance learning outcomes. Compared with research using clinician-narrated content, this study highlights that self-voice education produces superior outcomes across multiple domains including compliance, satisfaction, and psychological well-being. These findings establish a theoretical and practical framework for personalized AI-driven patient education. In real-world clinical settings, this approach offers a scalable, cost-effective solution to enhance patient engagement, particularly valuable in resource-limited environments where individualized education is challenging to deliver.

XSL•FO
RenderX

Trial Registration: Chinese Clinical Trial Registry ChiCTR2500101882; https://www.chictr.org.cn/showprojEN.html?proj=268927

(J Med Internet Res 2026;28:e81387) doi:10.2196/81387

KEYWORDS

artificial intelligence; voice cloning; medical education; ChatGPT; education effect evaluation; randomized controlled trial

Introduction

Patient education is a fundamental component of health care, aiming to enhance individuals' understanding of their medical conditions, promote self-management, improve treatment adherence, and ultimately optimize health outcomes and prognosis [1-3]. Despite the clinical significance of patient education, traditional methods of patient education—primarily verbal explanations by health care providers and standardized written materials—often fall short in addressing the diverse needs of patients [4,5]. These approaches are constrained by limited time, lack of interactivity, and insufficient personalization, frequently resulting in suboptimal comprehension, poor engagement, and reduced adherence to treatment recommendations [6].

Prior research has shown that personalized, interactive education enhances patient understanding and engagement [7-11]. Recent advances in artificial intelligence (AI), particularly in voice cloning and large language models (LLMs), present novel opportunities to transform conventional health education into a more personalized, scalable, and interactive process [12]. Voice cloning technology uses deep learning algorithms to synthesize natural-sounding speech from limited voice samples, enabling the delivery of educational content in familiar voices—such as those of physicians or even the patients themselves [13,14]. Prior studies suggest that familiar voices can increase emotional resonance [15-17], build trust, enhance patient engagement, improve memory retention of medical information [15,17], and increase accessibility to digital health tools. Recent scoping reviews have identified multiple applications of LLMs in patient education, including generating educational materials, interpreting medical information, and optimizing doctor-patient interaction [18]. Systematic reviews have further demonstrated that LLMs can enhance health care communication and support personalized patient engagement [19].

Simultaneously, LLMs like ChatGPT, which is based on advanced LLMs, have demonstrated strong capabilities for natural language understanding, dialog generation, and semantic evaluation [20] and have shown potential for supporting patient communication and education [18,19,21-23]. These tools can interact with patients in real time, deliver conversational-style education, and assess understanding through dynamic, automated evaluation processes [24-26]. Such integration may address key challenges in health education, including the need for tailored content delivery and standardized assessment without overburdening clinical staff. These models are capable of few-shot learning, enabling rapid adaptation to novel prompts [27].

Despite this emerging literature, significant gaps remain. First, few empirical studies have rigorously examined the comparative effectiveness of different voice sources—specifically physician-cloned voice versus patient-cloned voice—in AI-assisted patient education. Second, although LLM-based tools like ChatGPT show promise for health communication, their role as standardized evaluators of educational outcomes has not been comprehensively validated in clinical randomized trials. Traditional outcome assessments often rely on human raters, which introduces inter-rater variability and resource burdens.

To address these gaps, this study was a prospective randomized controlled trial (RCT) to investigate the effectiveness of an AI-assisted patient education system integrating voice cloning and ChatGPT. The study compared educational outcomes between physician voice cloning and patient self-voice cloning and explored the feasibility and accuracy of using ChatGPT as a standardized, automated evaluation tool. By doing so, this work aimed to contribute novel insights into how AI technologies can enhance the personalization, efficiency, and effectiveness of inpatient education and to extend the evidence base for AI-enabled patient engagement strategies. We hypothesized that (1) AI-assisted voice cloning education would significantly improve patient education outcomes compared with traditional education, (2) patient self-voice cloning would yield superior educational effectiveness compared with physician voice cloning due to the self-reference effect, and (3) ChatGPT could serve as a reliable automated evaluation tool with high agreement with expert assessment.

Methods

Overall Study

This study used an RCT design to assess the impact of AI-assisted patient health education using voice cloning technology and ChatGPT on patient education outcomes. Participants were randomly assigned to 1 of 3 groups: The 3 groups received different educational methods but the same educational content and duration of education. Following the educational intervention, the effectiveness of the education was comprehensively evaluated at fixed time points using ChatGPT. To ensure the validity and accuracy of the evaluation tool, a pre-test of the ChatGPT evaluation tool was conducted prior to the formal data collection. This study not only explored the potential of AI voice cloning technology in personalized patient education but also used ChatGPT as an assessment tool to further validate its reliability and practicality in evaluating health education interventions. Ultimately, the study aimed to provide theoretical insights and practical recommendations for improving medical health education models, enhancing patient

XSL•FO
RenderX

education outcomes, and promoting treatment adherence and quality of life.

Study Timeline

An independent tool validation study was conducted from January 2024 to June 2024 to verify the ChatGPT-based compliance scoring tool (no participant enrollment). Administrative preparation, staff training, and software and hardware debugging were completed in December 2024, and this period did not involve participant enrollment. Trial registration (ChiCTR2500101882) was initiated on January 15, 2025, and finalized on April 30, 2025, before enrollment began in May 2025. Participant enrollment and intervention delivery were conducted from May 2025 to June 2025, with 1-month follow-up assessments completed by July 2025.

Reporting Guidelines

The trial was conducted and reported in accordance with the CONSORT (Consolidated Standards of Reporting Trials) 2025 statement and the CONSORT-EHEALTH checklist for digital and AI-based health interventions (Checklists 1 and 2) [28,29].

Study Design

This single-center, 3-arm, parallel-group, superiority RCT (allocation ratio 1:1:1) was conducted at The First Hospital of Lanzhou University. System development, intervention standardization, and staff training were completed in December 2024 and January 2025. Participant recruitment and data collection were conducted from May 2025 to July 2025.

Randomization

Sequence Generation

The random allocation sequence was generated using a computer-generated random number table.

Allocation Concealment

To ensure concealment of group assignments, the allocation sequence was placed in sealed, opaque envelopes managed by independent researchers who were not involved in the implementation of the interventions. This procedure minimized the risk of selection bias.

Implementation

Independent researchers generated the allocation sequence and managed the envelopes. Eligible participants were assigned to study arms according to the contents of the envelopes.

Blinding

Due to the nature of the interventions, neither participants nor the medical staff implementing the interventions could be blinded. However, outcome assessors were blinded to group assignments to minimize assessment bias.

Ethical Considerations

The clinical trial protocol was reviewed and approved by the Ethics Committee of The First Hospital of Lanzhou University (approval number: LDYYLL-2025‐805) and conducted in accordance with the principles of the Declaration of Helsinki.

All participants received detailed verbal and written information regarding the study purpose, procedures, potential risks, and expected benefits and provided written informed consent prior to enrollment. The consent process also covered permission for the use and publication of anonymized data. Participants were informed that their involvement was voluntary and that they could withdraw from the study at any time without affecting their standard medical care.

All collected data were deidentified before analysis and stored in encrypted, password-protected databases accessible only to the research team. Personal identifiers were removed to ensure participant confidentiality and data security.

The use of voice cloning technology was explicitly covered in the consent form. Participants and physicians provided separate written authorization for the use of their voice recordings solely for research purposes. All voice samples and generated models were permanently deleted after study completion.

Participants did not receive monetary compensation but were offered free access to the AI-based health education program and follow-up consultations as part of their clinical care. No identifiable photographs nor audiovisual materials were used in this publication.

Participants

The study participants were patients hospitalized in a tertiary hospital from May 2025 to July 2025 who needed to receive medical education.

Inclusion and Exclusion Criteria

The inclusion criteria were (1) age ≥18 years of any gender; (2) clear diagnosis, patients who needed to receive medical education; (3) clear consciousness, able to understand and cooperate with the study; and (4) voluntary participation in this study and signed informed consent. The exclusion criteria were (1) severe hearing impairment, unable to normally receive voice education; (2) severe cognitive impairment, unable to understand education content; (3) expected hospitalization time <3 days; and (4) previously participated in similar studies.

Sample Size

Based on the pre-experimental results, with the education content compliance rate as the main observation indicator, α set at 0.05 (2-sided), β set at 0.10, a test power of 90%, and an expected effect size of 0.3 among the 3 groups, we calculated that each group needed a sample size of 54 participants using G*Power 3.1 software. Considering a possible dropout rate of approximately 10%, we finally determined that we would enroll 60 participants in each group, for a total of 180 participants. Prior to participation, all patients were informed that the educational intervention would be delivered using AI-generated voice recordings. Written informed consent was obtained from all participants.

Independent Tool Validation Study (January 2024 to June 2024): Validation of ChatGPT-Based Compliance Scoring

This independent tool validation study was conducted separately from the randomized trial and did not involve participant

XSL•FO
RenderX

enrollment. To validate the AI-based scoring used for the primary outcome, we conducted a separate pilot validation study (January 2024 through June 2024) among 30 volunteers who were not part of the randomized trial. Phase 1 (verbatim-matching workflow) used iFlytek automatic speech recognition (Mandarin medical context; no manual correction) to transcribe each participant's recitation; the transcribed text was pasted into the ChatGPT web interface together with the standard education text for rubric-based scoring. Phase 2 (semantic audio-based workflow) re-evaluated the same recordings by uploading the original audio directly to the ChatGPT web interface (WAV/MP3; 16-kHz sampling rate; 60 - 120 s per recording); each recording was scored once. Both workflows were benchmarked against consensus ratings from 3 senior nursing experts using the same rubric. Agreement with expert ratings was substantial in Phase 1 (weighted κ=.72) and higher in Phase 2 (weighted κ=.87), with strong internal consistency of the evaluation items (Cronbach α=0.89). Therefore, Phase 2 was selected for the main RCT to minimize transcription-related errors and enhance reproducibility. Detailed prompts, rubric, and example inputs and outputs are provided in Multimedia Appendix 1.

Intervention Measures

Traditional Education Group (Control Group)

The traditional education group received the hospital's current standardized patient education, including verbal education by medical staff and standardized written education materials. Verbal education was conducted by trained medical staff according to a unified education outline, including disease knowledge, treatment plans, precautions, and lifestyle guidance. Written education materials were standardized education manuals compiled by the hospital that patients could take back to the ward or home to read. The duration of the education was approximately 20 minutes to 30 minutes.

Attending Physician Voice Cloning Education Group (Intervention Group 1)

The physician voice cloning education group (intervention group 1) received patient education through cloning of the attending physician's voice. First, the patient's attending physician's voice sample (approximately 5 min) was obtained, and Resemble.ai voice cloning technology was used to generate the doctor's voice model. Medical professionals then formulated personalized education content based on standard education content and the patient's specific situation, synthesized the voice using doctor voice cloning, and produced personalized voice education materials. Patients listened to the voice-cloned education content through headphones or speakers under the guidance of research assistants and received printed personalized education materials. The education content covered the same topics as the control group but was personalized according to patient characteristics, such as using language expressions suitable for the patient's education level, lifestyle suggestions targeted at the patient's occupational characteristics, and emphasizing relevant content based on the patient's concerns. The duration of the education was approximately 20 minutes to 30 minutes.

Patient's Own Voice Cloning Education Group (Intervention Group 2)

The patient voice cloning education group received patient education through cloning of the patient's own voice. First, the patient's voice sample (approximately 5 min) was obtained, and Resemble.ai voice cloning technology was used to generate the patient's voice model. Similar to intervention group 1, medical professionals then formulated personalized education content, synthesized the voice through patient voice cloning, and produced personalized voice education materials. Patients listened to the voice-cloned education content through mobile phone headphones or speakers under the guidance of research assistants and received printed personalized education materials. The education content was the same as in intervention group 1, with the only difference being the use of the patient's own voice for education. The duration of the education was approximately 20 minutes to 30 minutes.

Educational Content and Assessment of Education Effect

All participants received the same educational content and could ask questions after receiving education, with medical staff answering inquiries. All 3 groups of patients had their education effect assessed before discharge by using ChatGPT to compare the education content recited by patients with the original education content to calculate the content compliance rate, and follow-up assessments were conducted 1 month after discharge.

The educational intervention included AI-generated voice recordings delivered using cloning of either the attending physician's voice or the patient's own voice. ChatGPT was used to assess patient knowledge, satisfaction, and adherence using standardized prompts. Patient outcomes were also measured using structured questionnaires. The questionnaires are provided in Multimedia Appendices 2-4, and the voice cloning patient education prompts, ChatGPT evaluation prompts, and AI-generated educational content are provided in Multimedia Appendix 5. Data were collected at baseline and postintervention, and all responses were recorded in a secure electronic database.

Evaluation Indicators and Measurement Tools

Primary Outcome Measure

The primary outcome, education content compliance rate, was defined as the degree of semantic concordance between the education content recited by patients and the standardized education material using the following equation: compliance rate = (number of key points correctly recited by patients/total number of key points in education content) × 100%. The education content compliance rate was assessed using ChatGPT-4 (model ChatGPT [OpenAI; web interface, model displayed as GPT-4 during the study period]) as an automated evaluation tool to quantify semantic similarity between the patient's oral recitation and the standard education content, enabling scalable and objective scoring.

XSL•FO
RenderX

Secondary Outcome Measures

Knowledge Mastery

We used disease-related knowledge questionnaires to assess patients' understanding and memory of disease knowledge, with a possible total score of 100 points and higher scores indicating better knowledge mastery.

Education Satisfaction

We used the education satisfaction scale to assess patients' satisfaction with the education process and content, including aspects such as education content, method, time, and effect, using a 5-level Likert scale. Higher scores indicate higher satisfaction.

Treatment Adherence

We used the treatment adherence scale to assess patients' treatment adherence within 1 month after discharge, including aspects such as drug treatment, lifestyle adjustment, and follow-up review. Higher scores indicate better adherence.

Quality of Life

We used the SF-36 quality of life scale to assess changes in patients' quality of life 1 month after discharge.

Anxiety and Depression

We used the Hospital Anxiety and Depression Scale to assess changes in patients' anxiety and depression before and after the education.

Evaluation Time Points

Baseline Assessment

We collected patients' basic information and baseline data, including demographic characteristics, disease-related knowledge, anxiety, and depression, before the intervention.

Immediate Posteducation Assessment

The education content compliance rate, knowledge mastery, and education satisfaction were assessed immediately after the education was completed.

Predischarge Assessment

Education content compliance rate and knowledge mastery were re-assessed 1 day before patient discharge.

1-Month Postdischarge Follow-Up Assessment

Through telephone or outpatient follow-up, we assessed patients' treatment adherence, quality of life, disease-related knowledge retention, anxiety, and depression.

Data Collection and Management

Research assistants who had received standardized training were responsible for data collection, using uniform data collection forms to record patient information and evaluation results. An electronic database was established, with all data independently entered by 2 personnel and cross-checked to ensure accuracy. To protect patient privacy, all data were stored in coded form, and regular data quality checks were conducted to ensure completeness and reliability. Missing data were promptly identified and imputed.

All voice recordings collected in this study were considered sensitive personal data. They were stored on secure, password-protected servers at The First Hospital of Lanzhou University, accessible only to authorized research personnel. For analysis, the recordings were anonymized. Following the completion of the study, all original recordings will be permanently deleted in accordance with institutional data retention and privacy policies.

Statistical Analysis

Statistical analyses were performed using SPSS version 26.0 (IBM Corp). All analyses followed the intention-to-treat (ITT) principle. Missing outcome data (6/182, 3.3%) were evaluated using the missing completely at random test from Little, which indicated that missingness was completely at random (χ^2_{12}=14.5, P=.27). Consequently, multiple imputation by chained equations was used to generate 5 imputed datasets (m=5), and pooled estimates were obtained using the rules by Rubin. Sensitivity analyses using per-protocol (complete-case) data yielded consistent findings (absolute difference in mean compliance<1%). Continuous variables are presented as mean (SD), and categorical variables are presented as n (%). Between-group comparisons were conducted using 1-way ANOVA with post hoc tests (least significant difference or Bonferroni, as appropriate). Categorical variables were compared using χ^2 tests or Fisher exact tests. All tests were 2-sided with a significance level of P<.05.

Evaluation Indicators

The primary outcome of this study was the education content compliance rate, which reflected how well patients understood and followed the health education provided. Secondary outcomes included knowledge retention, patient satisfaction, and treatment adherence.

Education level and knowledge retention were measured using a structured questionnaire developed by the research team based on the hospital's standardized education materials.

The questionnaire consisted of multiple-choice and short-answer questions assessing patients' understanding and recall of the key educational points immediately after education and before discharge.

Each correct response was scored as 1 point, with higher total scores indicating better knowledge retention.

The disease-related knowledge questionnaire was developed based on the hospital's standardized education materials and underwent content review by 3 senior nursing experts.

Results

Participant Flow

Figure 1 illustrates the participant flow. ITT analysis was performed for all participants (n=180) using multiple imputation by chained equations for missing data. The analysis section in the figure shows the per-protocol analyses (n=174).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram.

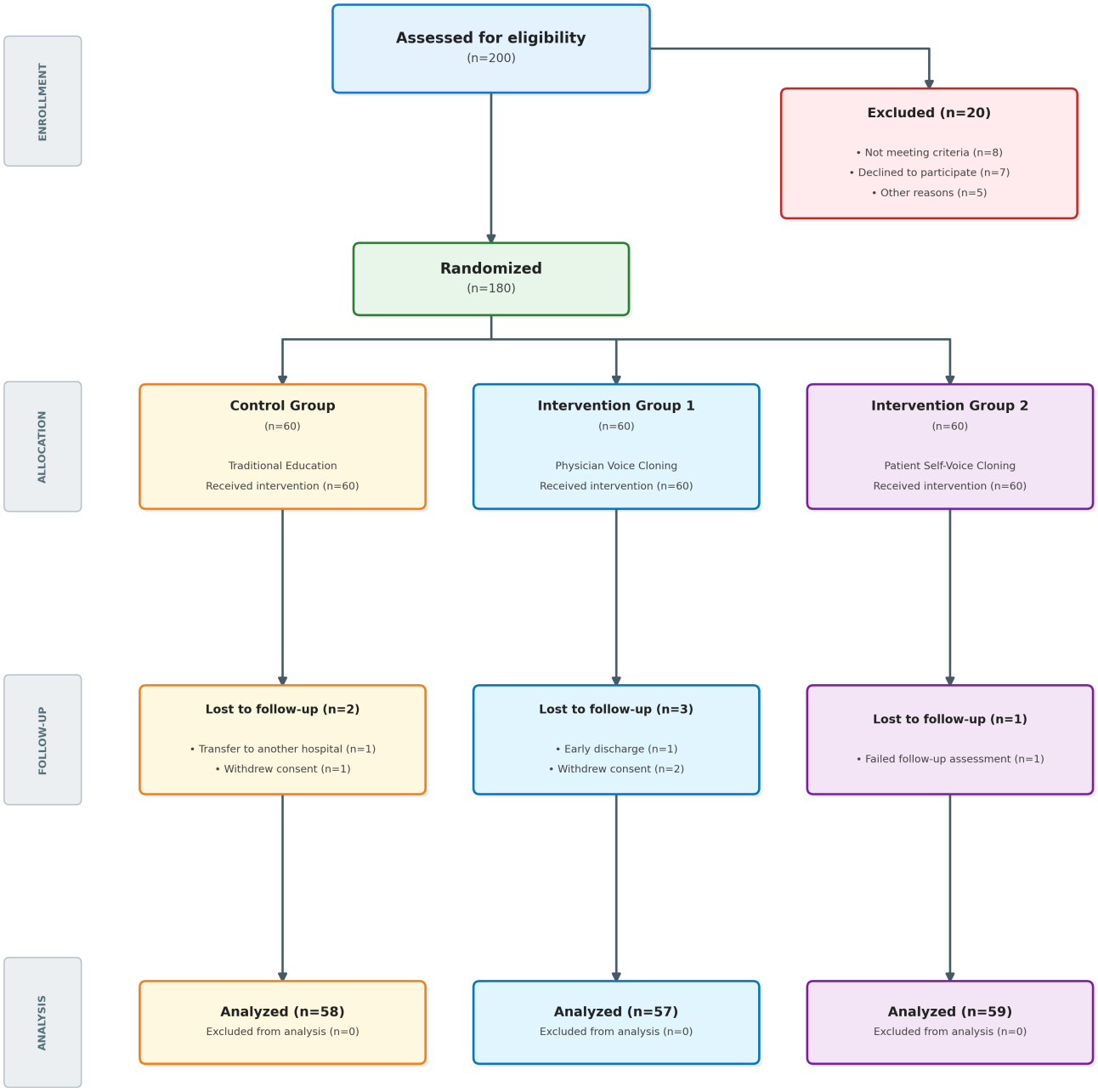

Comparison of General Information

A total of 180 patients were recruited to participate in this study. During the follow-up period, 6 participants were lost to follow-up due to transfer to another hospital (n=1) and withdrawal of consent (n=1) in the control group; early discharge (n=1) and withdrawal of consent (n=2) in intervention group 1; and failure to complete follow-up assessments (n=1) in intervention group 2. Therefore, 174 patients completed the study (58 in the control group, 57 in intervention group 1, 59 in intervention group 2), for a total dropout rate of 3.3% (6/180). After multiple imputation, the analysis results remained consistent with those of the complete-case dataset, indicating the robustness of the findings. Comparison of general information such as age, gender, education level, occupation, and disease type among the 3 groups of patients showed no statistically significant differences (P>.05), indicating comparability (Table 1).

XSL•FO
RenderX

Table . Comparison of baseline characteristics among 3 groups of hospitalized patients with lung cancer in a randomized controlled trial of artificial intelligence (AI)–assisted voice cloning education at The First Hospital of Lanzhou University (May 2025 through July 2025).

Item	Control group (n=58)	Intervention group 1 (n=57)	Intervention group 2 (n=59)	Statistical test result (*df*)	*P* value
Age (years), mean (SD)	52.6 (14.8)	53.2 (15.1)	51.9 (14.5)	0.127[a] (2, 171)	.88
Gender, n (%)				0.215[b] (2)	.90
Male	31 (53)	32 (56)	34 (58)		
Female	27 (47)	25 (44)	25 (42)		
Education level, n (%)				0.986[b] (6)	.91
Primary school and less	12 (21)	11 (19)	10 (17)		
Junior high school	18 (31)	17 (30)	19 (32)		
High school/technical secondary school	16 (28)	18 (32)	17 (29)		
College and more	12 (21)	11 (19)	13 (22)		
Marital status, n (%)				0.352[b] (2)	.84
Married	49 (85)	47 (83)	51 (86)		
Unmarried/divorced/widowed	9 (16)	10 (18)	8 (14)		
Disease type, n (%)				0.763[b] (6)	.94
Cardiovascular disease	21 (36)	20 (35)	22 (37)		
Respiratory system disease	15 (26)	16 (28)	17 (29)		
Digestive system disease	13 (22)	12 (21)	11 (19)		
Others	9 (16)	9 (16)	9 (15)		
Previous hospitalization history, n (%)				0.124[b] (2)	.94
Yes	32 (55)	33 (58)	33 (56)		
No	26 (45)	24 (42)	26 (44)		
Baseline HADS[c] score, mean (SD)					
Anxiety	7.8 (3.2)	7.6 (3.1)	7.7 (3.3)	0.068[a] (2, 171)	.93
Depression	6.9 (2.8)	7.1 (2.9)	6.8 (2.7)	0.185[a] (2, 171)	.83

[a]*F* test.

[b]Chi-square test.

[c]HADS: Hospital Anxiety and Depression Scale.

Comparison of Education Content Compliance Rate

Immediately after education, the mean education compliance score in the control group was 73.2 (SD 8.5; 95% CI 71.0 - 75.4), compared with mean scores of 86.7 (SD 7.3; 95% CI 84.8 - 88.6) in intervention group 1 and 92.5 (SD 6.8; 95% CI 90.8 - 94.2) in intervention group 2. The 1-way ANOVA showed a statistically significant difference among the 3 groups ($F_{2,171}$=103.427, P<.001). Before discharge, the mean score in the control group was 68.5 (SD 9.1; 95% CI 66.2 - 70.8), while the mean scores in intervention group 1 and intervention group 2 were 82.3 (SD 8.1; 95% CI 80.2 - 84.4) and 88.6 (SD 7.4; 95% CI 86.7 - 90.5), respectively. The between-group difference remained statistically significant ($F_{2,171}$=95.682, P<.001; Table 2).

XSL•FO
RenderX

Table . Comparison of education content compliance rate among 3 groups of hospitalized patients with lung cancer in a randomized controlled trial of artificial intelligence (AI)–assisted voice cloning education at The First Hospital of Lanzhou University (May 2025 through July 2025), with differences analyzed using 1-way ANOVA.

Time point	Control group (n=58), mean (SD; 95% CI)	Intervention group 1 (n=57), mean (SD; 95% CI)	Intervention group 2 (n=59), mean (SD; 95% CI)	*F* test (*df*)	*P* value
Immediately after education	73.2 (8.5; 71.0 - 75.4)	86.7 (7.3; 84.8 - 88.6)	92.5 (6.8; 90.8 - 94.2)	103.427 (2, 171)	<.001
Before discharge	68.5 (9.1; 66.2 - 70.8)	82.3 (8.1; 80.2 - 84.4)	88.6 (7.4; 86.7 - 90.5)	95.682 (2, 171)	<.001

The corresponding ITT results based on the multiply imputed datasets are provided in Supplementary Table S2.

Comparison of Knowledge Mastery

Immediately after education, the mean knowledge mastery score in the control group was 76.3 (SD 9.2; 95% CI 73.9 - 78.7), which was significantly lower than the mean scores in intervention group 1 (88.5, SD 8.4; 95% CI 86.3 - 90.7) and intervention group 2 (90.2, SD 8.1; 95% CI 88.1 - 92.3). A statistically significant difference was observed among the 3 groups ($F_{2,171}$=48.362, *P*<.001). Before discharge, the control group had a mean knowledge mastery score of 71.8 (SD 9.8; 95% CI 69.2 - 74.4), whereas intervention group 1 and intervention group 2 achieved higher mean scores of 84.7 (SD 8.9; 95% CI 82.3 - 87.1) and 87.3 (SD 8.5; 95% CI 85.1 - 89.5), respectively. The between-group difference remained statistically significant ($F_{2,171}$=46.175, *P*<.001). At 1 month after discharge, the mean knowledge mastery score further declined in the control group (65.2, SD 10.5; 95% CI 62.4 - 68.0). In contrast, both intervention groups maintained relatively higher levels of knowledge mastery, with mean scores of 78.6 (SD 9.3; 95% CI 76.1 - 81.1) in intervention group 1 and 83.5 (SD 8.9; 95% CI 81.2 - 85.8) in intervention group 2. The overall difference among the 3 groups remained statistically significant ($F_{2,171}$=54.793, *P*<.001; Table 3).

Table . Comparison of knowledge mastery scores among 3 groups of hospitalized patients with lung cancer in a randomized controlled trial of artificial intelligence (AI)–assisted voice cloning education at The First Hospital of Lanzhou University (May 2025 through July 2025), with differences among groups analyzed using 1-way ANOVA.

Time point	Control group (n=58), mean (SD; 95% CI)	Intervention group 1 (n=57), mean (SD; 95% CI)	Intervention group 2 (n=59), mean (SD; 95% CI)	*F* test (*df*)	*P* value
Immediately after education	76.3 (9.2; 73.9-78.7)	88.5 (8.4; 86.3-90.7)	90.2 (8.1; 88.1-92.3)	48.362 (2, 171)	<.001
Before discharge	71.8 (9.8; 69.2-74.4)	84.7 (8.9; 82.3-87.1)	87.3 (8.5; 85.1-89.5)	46.175 (2, 171)	<.001
1 month after discharge	65.2 (10.5; 62.4-68.0)	78.6 (9.3; 76.1-81.1)	83.5 (8.9; 81.2-85.8)	54.793 (2, 171)	<.001

Comparison of Education Satisfaction

Significant differences were observed among the 3 groups across all dimensions of education satisfaction, including education content, education method, education time, education effect, and overall satisfaction (all *P*<.001). Across all items, the 95% CIs of the 2 intervention groups were consistently higher than those of the control group, with minimal overlap between groups. This indicates that the observed differences in education satisfaction were not only statistically significant but also stable and precise, reflecting a reliable improvement associated with the intervention (Table 4).

Table . Comparison of education satisfaction scores among 3 groups of hospitalized patients with lung cancer in a randomized controlled trial of artificial intelligence (AI)–assisted voice cloning education at The First Hospital of Lanzhou University (May 2025 through July 2025), with differences among groups analyzed using 1-way ANOVA.

Item	Control group (n=58), mean (SD; 95% CI)	Intervention group 1 (n=57), mean (SD; 95% CI)	Intervention group 2 (n=59), mean (SD; 95% CI)	*F* test (*df*)	*P* value
Education content	3.8 (0.7; 3.6 - 4.0)	4.3 (0.6; 4.1 - 4.5)	4.5 (0.5; 4.4 - 4.6)	21.364 (2, 171)	<.001
Education method	3.6 (0.8; 3.4 - 3.8)	4.5 (0.6; 4.3 - 4.7)	4.7 (0.4; 4.6 - 4.8)	52.781 (2, 171)	<.001
Education time	3.9 (0.6; 3.7 - 4.1)	4.2 (0.5; 4.1 - 4.3)	4.3 (0.5; 4.2 - 4.4)	9.526 (2, 171)	<.001
Education effect	3.7 (0.7; 3.5 - 3.9)	4.4 (0.5; 4.3 - 4.5)	4.6 (0.4; 4.5 - 4.7)	43.175 (2, 171)	<.001
Overall satisfaction	3.8 (0.6; 3.6 - 4.0)	4.4 (0.5; 4.3 - 4.5)	4.6 (0.4; 4.5 - 4.7)	45.293 (2, 171)	<.001

XSL•FO
RenderX

Comparison of Treatment Adherence

Among the 3 groups, significant differences in treatment adherence were observed across all dimensions, including drug treatment adherence, lifestyle adjustment, follow-up review, and total adherence score (all P<.001). For drug treatment adherence, the control group scored a mean of 7.6 (SD 1.5; 95% CI 7.2 - 8.0), which was lower than the mean scores for intervention group 1 (8.5, SD 1.2; 95% CI 8.2 - 8.8) and intervention group 2 (9.1, SD 0.9; 95% CI 8.9 - 9.3), with a statistically significant overall difference ($F_{2,171}$=25.364, P<.001). Similar patterns were observed for lifestyle adjustment and follow-up review adherence. The 95% CIs of both intervention groups were consistently higher than those of the control group, with limited overlap, indicating that the observed improvements in treatment adherence were stable and precise. Regarding the total adherence score, the control group achieved a mean of 21.6 (SD 4.2; 95% CI 20.5 - 22.7), whereas intervention group 1 and intervention group 2 achieved higher mean scores of 24.7 (SD 3.5; 95% CI 23.8 - 25.6) and 26.6 (SD 2.8; 95% CI 25.9 - 27.3), respectively. The between-group difference remained statistically significant ($F_{2,171}$=32.593, P<.001; Table 5).

Table . Comparison of treatment adherence scores among 3 groups of hospitalized patients with lung cancer in a randomized controlled trial of artificial intelligence (AI)–assisted voice cloning education at The First Hospital of Lanzhou University (May 2025 through July 2025), with differences among groups analyzed using 1-way ANOVA.

Item	Control group (n=58), mean (SD; 95% CI)	Intervention group 1 (n=57), mean (SD; 95% CI)	Intervention group 2 (n=59), mean (SD; 95% CI)	*F* test (*df*)	*P* value
Drug treatment	7.6 (1.5; 7.2 - 8.0)	8.5 (1.2; 8.2 - 8.8)	9.1 (0.9; 8.9 - 9.3)	25.364 (2, 171)	<.001
Lifestyle adjustment	6.8 (1.7; 6.4 - 7.2)	7.9 (1.4; 7.5 - 8.3)	8.6 (1.1; 8.3 - 8.9)	27.175 (2, 171)	<.001
Follow-up review	7.2 (1.6; 6.8 - 7.6)	8.3 (1.3; 8.0 - 8.6)	8.9 (1.0; 8.6 - 9.2)	26.482 (2, 171)	<.001
Total score	21.6 (4.2; 20.5 - 22.7)	24.7 (3.5; 23.8 - 25.6)	26.6 (2.8; 25.9 - 27.3)	32.593 (2, 171)	<.001

Comparison of Quality of Life

One month after discharge, significant differences were observed among the 3 groups in all SF-36 dimensions, including physical functioning, role-physical, bodily pain, general health, vitality, social functioning, role-emotional, and mental health (all P<.001). For each dimension, the 95% CIs of both intervention groups were consistently higher than those of the control group, indicating that the observed improvements in health-related quality of life were robust and precise.

Notably, intervention group 2 generally showed narrower 95% CIs, suggesting less variability and greater consistency in patient-reported outcomes compared with the control group (Table 6).

Table . Comparison of SF-36 quality of life scores 1 month after discharge among 3 groups of hospitalized patients with lung cancer in a randomized controlled trial of artificial intelligence (AI)–assisted voice cloning education at The First Hospital of Lanzhou University (May 2025 through July 2025), with differences among groups analyzed using 1-way ANOVA.

Dimension	Control group (n=58), mean (SD; 95% CI)	Intervention group 1 (n=57), mean (SD; 95% CI)	Intervention group 2 (n=59), mean (SD; 95% CI)	*F* test (*df*)	*P* value
Physical functioning	65.3 (12.5; 61.7 - 68.9)	72.6 (11.3; 69.5 - 75.7)	74.8 (10.9; 71.8 - 77.8)	11.364 (2, 171)	<.001
Role-physical	58.6 (14.2; 54.4 - 62.8)	67.5 (12.8; 63.8 - 71.2)	70.3 (12.1; 66.9 - 73.7)	13.175 (2, 171)	<.001
Bodily pain	67.2 (13.1; 63.3 - 71.1)	74.8 (11.9; 71.5 - 78.1)	76.5 (11.2; 73.3 - 79.7)	10.482 (2, 171)	<.001
General health	61.5 (12.8; 57.7 - 65.3)	69.7 (11.5; 66.5 - 72.9)	72.4 (10.8; 69.3 - 75.5)	14.293 (2, 171)	<.001
Vitality	59.8 (13.5; 55.8 - 63.8)	68.3 (12.2; 65.1 - 71.5)	71.6 (11.5; 68.5 - 74.7)	15.364 (2, 171)	<.001
Social functioning	63.7 (12.9; 59.8 - 67.6)	71.5 (11.7; 68.4 - 74.6)	76.2 (10.8; 73.2 - 79.2)	18.175 (2, 171)	<.001
Role-emotional	60.2 (14.1; 56.0 - 64.4)	69.8 (12.5; 66.3 - 73.3)	75.3 (11.7; 72.1 - 78.5)	21.482 (2, 171)	<.001
Mental health	62.5 (13.2; 58.7 - 66.3)	70.6 (12.1; 67.3 - 73.9)	76.8 (11.3; 73.8 - 79.8)	22.293 (2, 171)	<.001

Comparison of Anxiety and Depression

Before education, there were no significant differences in anxiety nor depression scores among the 3 groups (anxiety: $F_{2,171}$=0.068, P=.93; depression: $F_{2,171}$=0.185, P=.83), indicating baseline comparability. One month after discharge, significant differences were observed among the groups for both anxiety and depression scores (P<.001). Specifically, for anxiety, the control group scored a mean of 6.5 (SD 2.9; 95% CI 5.8 - 7.2), intervention group 1 scored a mean of 5.2 (SD 2.5; 95% CI 4.6 - 5.8), and intervention group 2 scored a mean of 4.3 (SD 2.2; 95% CI 3.8 - 4.8). For depression, the mean scores were 5.8 (SD 2.6; 95% CI 5.2 - 6.4), 4.6 (SD 2.3; 95% CI 4.1 - 5.1), and 3.7 (SD 2.0; 95% CI 3.3 - 4.1), respectively.

The 95% CIs of the intervention groups did not overlap with those of the control group, indicating that the reductions in

XSL•FO
RenderX

anxiety and depression after intervention were stable and precise (Table 7).

Table . Comparison of Hospital Anxiety and Depression Scale (HADS) anxiety and depression scores among 3 groups of hospitalized patients with lung cancer in a randomized controlled trial of artificial intelligence (AI)–assisted voice cloning education at The First Hospital of Lanzhou University (May 2025 through July 2025), with differences among groups analyzed using 1-way ANOVA.

Time point	Control group (n=58), mean (SD; 95% CI)	Intervention group 1 (n=57), mean (SD; 95% CI)	Intervention group 2 (n=59), mean (SD; 95% CI)	*F* test (*df*)	*P* value
Pre-education anxiety	7.8 (3.2; 7.0-8.6)	7.6 (3.1; 6.8-8.4)	7.7 (3.3; 6.9-8.5)	0.068 (2, 171)	.93
Pre-education depression	6.9 (2.8; 6.2-7.6)	7.1 (2.9; 6.4-7.8)	6.8 (2.7; 6.1-7.5)	0.185 (2, 171)	.83
1 month after discharge anxiety	6.5 (2.9; 5.8-7.2)	5.2 (2.5; 4.6-5.8)	4.3 (2.2; 3.8-4.8)	12.364 (2, 171)	<.001
1 month after discharge depression	5.8 (2.6; 5.2-6.4)	4.6 (2.3; 4.1-5.1)	3.7 (2.0; 3.3-4.1)	13.175 (2, 171)	<.001

Discussion

Principal Findings

This RCT investigated the effectiveness of an AI-assisted patient education system integrating voice cloning technology and compared the educational outcomes of physician voice cloning versus patient self-voice cloning. In addition, the study explored the feasibility of using ChatGPT as a supportive tool to assist with evaluating education effectiveness. Consistent with our 3 prespecified hypotheses, the results demonstrated that AI-assisted voice cloning–based education significantly improved education content compliance, knowledge mastery, satisfaction, treatment adherence, and short-term psychological and quality-of-life outcomes compared with traditional education. These findings are consistent with prior research indicating that technology-assisted and personalized education can enhance patient engagement and learning outcomes [10,22]. Notably, education delivered using patients' own cloned voices yielded superior effects compared with physician voice cloning across multiple outcome domains, supporting the hypothesis that self-referential personalization enhances educational effectiveness.

Interpretation of Findings and Comparison With Previous Studies

Voice cloning technology represents an emerging application of AI in patient education, offering distinct advantages over conventional education models. By enabling repeated exposure to standardized educational content delivered in a familiar and emotionally resonant voice, voice cloning may enhance attention, comprehension, and memory consolidation [30,31]. Consistent with previous studies demonstrating that familiar auditory cues improve trust and acceptance in health communication, both AI-assisted intervention groups in this study outperformed the traditional education group in education compliance, knowledge mastery, and satisfaction [30,32,33]. Beyond confirming the general effectiveness of voice-based AI education, this study provides novel evidence regarding differences between voice sources. Patients who received education via their own cloned voice achieved higher compliance, satisfaction, knowledge retention, treatment adherence, and better short-term psychological outcomes than those educated using physician voice cloning. Previous studies have largely focused on clinician-narrated or professionally recorded content [34], whereas empirical comparisons involving patient self-voice have been scarce. The observed superiority of self-voice education may be explained by the self-reference effect, whereby information related to the self is processed more deeply and remembered more effectively [17,35,36]. Research in cognitive psychology suggests that self-related stimuli—particularly self-generated or self-similar auditory information—enhance attention, emotional engagement, and memory encoding [16,35]. In this study, hearing one's own voice narrate medical information may have strengthened self-identification and personal relevance, thereby reinforcing learning and adherence. Furthermore, the self-voice group demonstrated superior outcomes in multiple quality-of-life domains and had greater reductions in anxiety and depression. These findings align with evidence that personalized and self-relevant health communication can enhance perceived control, self-efficacy, and emotional regulation among patients [14,36]. Importantly, the reported effect sizes indicated moderate-to-large intervention effects, suggesting that the observed improvements were not only statistically significant but also clinically meaningful. Reporting both effect sizes and confidence intervals enhances transparency and interpretability and is consistent with CONSORT 2025 recommendations for RCTs [28].

Implications for AI-Assisted Evaluation of Patient Education

This study also explored the use of ChatGPT as an auxiliary tool to assist with evaluating patient education outcomes based on standardized educational content. Pre-experimental testing demonstrated good agreement between ChatGPT-assisted scoring and expert evaluation, suggesting that LLMs may support structured and reproducible assessment under controlled conditions. These findings are consistent with emerging literature indicating that LLMs can perform semantic analysis and text evaluation tasks with acceptable reliability in medical and educational contexts [6,37]. Compared with traditional manual evaluation, AI-assisted assessment may offer advantages in efficiency, scalability, and consistency, particularly in settings with limited human resources. However, ChatGPT should not be regarded as a fully objective evaluator. Assessment outcomes

XSL•FO
RenderX

may still be influenced by prompt design, model architecture, and algorithmic biases. Therefore, AI-assisted evaluation should be considered a complementary tool rather than a replacement for expert assessment, and its use should be accompanied by appropriate human oversight [38,39].

Study Limitations

Several limitations should be acknowledged. First, the follow-up period was limited to 1 month, restricting conclusions regarding long-term adherence, psychological outcomes, and sustained quality of life benefits. Longer follow-up periods are needed to assess the durability of intervention effects, as recommended in previous digital health education studies [40]. Second, this was a single-center study involving hospitalized patients with relatively homogeneous educational backgrounds, which may limit generalizability. Prior research suggests that educational level and health literacy can influence the effectiveness of digital health interventions [41]. Multicenter studies with more diverse populations are therefore warranted. Third, although voice cloning technology offers a promising approach, current speech synthesis systems still have limitations in emotional nuance and naturalness, which may affect user engagement [13]. Continued technological advancements may further enhance the effectiveness of AI-assisted education. Finally, although AI-assisted education may reduce clinical workload and improve efficiency, this study did not include a formal cost-effectiveness analysis. Future research should systematically evaluate economic outcomes to support large-scale implementation decisions [40]. Because ChatGPT is a continuously updated product, we standardized the scoring workflow (fixed rubric and prompts, prespecified audio format, and single scoring per recording) and provide the complete scoring materials to facilitate reproducibility. Additionally, voice-based delivery systems may misinterpret medication names or specialized terminology, underscoring the need for clinician review of generated content [42].

Implications for Practice and Policy

Implementation should address unit- and hospital-level factors separately. Targeted strategies should improve staff training, digital access, and policy support. Addressing demographic disparities is key to promoting equity and care quality.

Conclusions and Broader Implications

This study introduces an innovative patient education model integrating AI voice cloning and ChatGPT, representing a novel approach distinct from previous studies that primarily relied on standard text-to-speech or professionally recorded content. The key innovation lies in using patients' own cloned voices for health education delivery, leveraging the self-reference effect to enhance learning outcomes. Compared with prior research focusing on clinician-narrated content, this study provides the first empirical evidence that self-voice education produces superior outcomes across multiple domains including compliance, satisfaction, and psychological well-being. These findings contribute to the field by establishing a theoretical and practical framework for personalized AI-driven patient education. In real-world clinical settings, this approach offers a scalable, cost-effective solution to enhance patient engagement, particularly valuable in resource-limited environments where individualized education is challenging to deliver. Future research should focus on multicenter validation, longer follow-up periods, and exploration of optimal voice cloning parameters to maximize educational effectiveness.

Acknowledgments

The authors thank all individuals who contributed to this study. Appreciation is extended to the patients and clinical staff of the First Hospital of Lanzhou University for their participation and cooperation throughout the research process. The authors also acknowledge the technical support provided by the development team for the artificial intelligence (AI) health education and voice cloning systems, led by author MM. The authors further thank all collaborators who assisted with data collection, statistical analysis, and manuscript preparation. Institutional support from the Department of Thoracic Surgery, The First Hospital of Lanzhou University, is also acknowledged for providing research resources and facilities.

In accordance with the Generative AI Delegation Taxonomy (GAIDeT; 2025), generative AI tools were used under full human supervision for proofreading, language editing, and translation. The tool used was ChatGPT (GPT-4; OpenAI). All AI-generated content was reviewed and validated by the authors. Responsibility for the final manuscript rests entirely with the authors. Generative AI tools are not listed as authors and did not contribute to study design, data analysis, or interpretation.

Funding

This study was supported by the Key Research and Development Program of Gansu Province (26YFFA004); Natural Science Foundation of Gansu Province (23JRRA1597); the 2024 Key Research and Development Program of Gansu Province–International Cooperation Field (24YFWA011), and the Institutional Research Fund of The First Hospital of Lanzhou University (ldyyyn2023-63).

Data Availability

The datasets generated and analyzed during this study are not publicly available due to privacy and ethical restrictions related to patient information and institutional regulations. Deidentified data are stored securely at The First Hospital of Lanzhou University and are available from the corresponding author upon reasonable request and approval by the hospital's ethics committee (approval number: ldyyyn2023-63).

XSL•FO
RenderX

Authors' Contributions

Conceptualization: YS (lead), HJ (supporting), YZ (supporting)
Data curation: SX (lead), XH (supporting), KJ (supporting), YZ (supporting)
Formal analysis: YS (equal), SX (equal), HJ (equal)
Methodology: YS (lead), SX (lead), HJ (supporting), XH (supporting), KJ (supporting), MM (technical lead)
Project administration: XM (lead)
Supervision: HW (lead)
Validation: YS (lead)
Visualization: SX (lead), XH (supporting)
Writing - original draft: YS (lead), SX (supporting), MM (advising)
Writing - review & editing: YS (supporting), SX (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
ChatGPT-based compliance scoring: prompts, rubric, and example inputs/outputs.
[PDF File, 172 KB - jmir_v28i1e81387_app1.pdf]

Multimedia Appendix 2
Education content compliance, knowledge mastery, and treatment adherence questionnaires.
[PNG File, 187 KB - jmir_v28i1e81387_app2.png]

Multimedia Appendix 3
SF-36 questionnaire.
[PNG File, 784 KB - jmir_v28i1e81387_app3.png]

Multimedia Appendix 4
Hospital Anxiety and Depression Scale.
[PNG File, 797 KB - jmir_v28i1e81387_app4.png]

Multimedia Appendix 5
Voice cloning patient education prompts, ChatGPT evaluation prompts, example patient knowledge assessment sheet, and notes for researchers.
[DOCX File, 1029 KB - jmir_v28i1e81387_app5.docx]

Checklist 1
CONSORT 2025 checklist.
[PDF File, 154 KB - jmir_v28i1e81387_app6.pdf]

Checklist 2
CONSORT-EHEALTH checklist (V 1.6.1).
[PDF File, 202 KB - jmir_v28i1e81387_app7.pdf]

References

1. Kocaballi AB, Quiroz JC, Rezazadegan D, et al. Responses of conversational agents to health and lifestyle prompts: investigation of appropriateness and presentation structures. J Med Internet Res 2020 Feb 9;22(2):e15823. [doi: 10.2196/15823] [Medline: 32039810]
2. Zolnierek KBH, Dimatteo MR. Physician communication and patient adherence to treatment: a meta-analysis. Med Care 2009 Aug;47(8):826-834. [doi: 10.1097/MLR.0b013e31819a5acc] [Medline: 19584762]
3. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 6;4(2):e19. [doi: 10.2196/mental.7785] [Medline: 28588005]
4. Shersher V, Haines TP, Sturgiss L, Weller C, Williams C. Definitions and use of the teach-back method in healthcare consultations with patients: a systematic review and thematic synthesis. Patient Educ Couns 2021 Jan;104(1):118-129. [doi: 10.1016/j.pec.2020.07.026] [Medline: 32798080]

XSL•FO
RenderX

5. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature New Biol 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
6. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
7. Denecke K, Gabarron E, Grainger R, et al. Artificial intelligence for participatory health: applications, impact, and future implications. Yearb Med Inform 2019 Aug;28(1):165-173. [doi: 10.1055/s-0039-1677902] [Medline: 31022749]
8. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
9. Milne-Ives M, de Cock C, Lim E, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020 Oct 22;22(10):e20346. [doi: 10.2196/20346] [Medline: 33090118]
10. Schachner T, Keller R, V Wangenheim F. Artificial intelligence-based conversational agents for chronic conditions: systematic literature review. J Med Internet Res 2020 Sep 14;22(9):e20701. [doi: 10.2196/20701] [Medline: 32924957]
11. Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y. Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. J Med Internet Res 2019 Apr 5;21(4):e12887. [doi: 10.2196/12887] [Medline: 30950796]
12. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25:e48568. [doi: 10.2196/48568] [Medline: 37379067]
13. Sisman B, Yamagishi J, King S, Li H. An overview of voice conversion and its challenges: from statistical modeling to deep learning. IEEE/ACM Trans Audio Speech Lang Process 2020;29:132-157. [doi: 10.1109/TASLP.2020.3038524]
14. Dai G, Wang P, Yao C, Ying F. InnerSelf: designing self-deepfaked voice for emotional well-being. arXiv. Preprint posted online on Mar 18, 2025 URL: https://arxiv.org/abs/2503.14257 [accessed 2026-03-03] [doi: 10.48550/arXiv.2503.14257]
15. Nygaard LC, Pisoni DB. Talker-specific learning in speech perception. Percept Psychophys 1998 Apr;60(3):355-376. [doi: 10.3758/bf03206860] [Medline: 9599989]
16. Kaplan JT, Aziz-Zadeh L, Uddin LQ, Iacoboni M. The self across the senses: an fMRI study of self-face and self-voice recognition. Soc Cogn Affect Neurosci 2008 Sep;3(3):218-223. [doi: 10.1093/scan/nsn014] [Medline: 19015113]
17. Yonan CA, Sommers MS. The effects of talker familiarity on spoken word identification in younger and older listeners. Psychol Aging 2000;15(1):88-99. [doi: 10.1037/0882-7974.15.1.88]
18. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne) 2024;11:1477898. [doi: 10.3389/fmed.2024.1477898] [Medline: 39534227]
19. Busch F, Hoffmann L, Rueger C, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med (Lond) 2025 Jan 21;5(1):26. [doi: 10.1038/s43856-024-00717-2] [Medline: 39838160]
20. OpenAI, Achiam J, Adler S, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 15, 2023 URL: https://arxiv.org/abs/2303.08774 [accessed 2026-03-03] [doi: 10.48550/arXiv.2303.08774]
21. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
22. Chavez-Yenter D, Kimball KE, Kohlmann W, et al. Patient interactions with an automated conversational agent delivering pretest genetics education: descriptive study. J Med Internet Res 2021 Nov 18;23(11):e29447. [doi: 10.2196/29447] [Medline: 34792472]
23. Bickmore TW, Pfeifer LM, Byron D, et al. Usability of conversational agents by patients with inadequate health literacy: evidence from two clinical trials. J Health Commun 2010;15 Suppl 2:197-210. [doi: 10.1080/10810730.2010.499991] [Medline: 20845204]
24. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. Digit Health 2019;5:2055207619871808. [doi: 10.1177/2055207619871808] [Medline: 31467682]
25. Sezgin E, Huang Y, Ramtekkar U, Lin S. Readiness for voice assistants to support healthcare delivery during a health crisis and pandemic. NPJ Digit Med 2020;3:122. [doi: 10.1038/s41746-020-00332-0] [Medline: 33015374]
26. Sezgin E, Militello LK, Huang Y, Lin S. A scoping review of patient-facing, behavioral health interventions with voice assistant technology targeting self-management and healthy lifestyle behaviors. Transl Behav Med 2020 Aug 7;10(3):606-628. [doi: 10.1093/tbm/ibz141] [Medline: 32766865]
27. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Presented at: Advances in Neural Information Processing Systems 33 (NeurIPS 2020); Dec 6-12, 2020; Virtual-only conference p. 1877-1901 URL: https://papers.nips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html [accessed 2026-03-03]
28. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025 Apr 14;389:e081123. [doi: 10.1136/bmj-2024-081123] [Medline: 40228833]
29. Eysenbach G, Group CE. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
30. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]

XSL•FO
RenderX

31. Battineni G, Chintalapudi N, Amenta F. AI chatbot design during an epidemic like the novel coronavirus. Healthcare (Basel) 2020 Jun 3;8(2):154. [doi: 10.3390/healthcare8020154] [Medline: 32503298]
32. Nayak A, Vakili S, Nayak K, et al. Use of voice-based conversational artificial intelligence for basal insulin prescription management among patients with type 2 diabetes: a randomized clinical trial. JAMA Netw Open 2023 Dec 1;6(12):e2340232. [doi: 10.1001/jamanetworkopen.2023.40232] [Medline: 38039007]
33. Bickmore TW, Pfeifer LM, Paasche-Orlow MK. Using computer agents to explain medical documents to patients with low health literacy. Patient Educ Couns 2009 Jun;75(3):315-320. [doi: 10.1016/j.pec.2009.02.007] [Medline: 19297116]
34. Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018 Sep 1;25(9):1248-1258. [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
35. Sui J, Humphreys GW. The integrative self: how self-reference integrates perception and memory. Trends Cogn Sci 2015 Dec;19(12):719-728. [doi: 10.1016/j.tics.2015.08.015] [Medline: 26447060]
36. Rogers TB, Kuiper NA, Kirker WS. Self-reference and the encoding of personal information. J Pers Soc Psychol 1977;35(9):677-688. [doi: 10.1037/0022-3514.35.9.677]
37. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
38. Sun GH, Hoelscher SH. The ChatGPT storm and what faculty can do. Nurse Educ 2023;48(3):119-124. [doi: 10.1097/NNE.0000000000001390] [Medline: 37043716]
39. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
40. Murray E, Hekler EB, Andersson G, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016 Nov;51(5):843-851. [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
41. Paasche-Orlow MK, Wolf MS. The causal pathways linking health literacy to health outcomes. Am J Health Behav 2007;31 Suppl 1(1):S19-S26. [doi: 10.5555/ajhb.2007.31.supp.S19] [Medline: 17931132]
42. Palanica A, Thommandram A, Lee A, Li M, Fossat Y. Do you understand the words that are comin outta my mouth? Voice assistant comprehension of medication names. NPJ Digit Med 2019;2:55. [doi: 10.1038/s41746-019-0133-x] [Medline: 31304401]

Abbreviations

AI: artificial intelligence
CONSORT: Consolidated Standards of Reporting Trials
ITT: intention to treat
LLM: large language model
RCT: randomized controlled trial

Edited by S Brini; submitted 28.Jul.2025; peer-reviewed by A Khamees, B Meskó; accepted 17.Jan.2026; published 19.Mar.2026.

Please cite as:
Sun Y, Xu S, Jin H, Han X, Jin K, Zhang Y, Ma X, Wei H, Ma M
Effectiveness of AI-Assisted Patient Health Education Using Voice Cloning and ChatGPT: Prospective Randomized Controlled Trial
J Med Internet Res 2026;28:e81387
URL: https://www.jmir.org/2026/1/e81387
doi:10.2196/81387

© Yan Sun, Shangqing Xu, Hongying Jin, Xiaoyan Han, Kangqi Jin, Yimei Zhang, Xiaoli Ma, Huaping Wei, Minjie Ma. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Reliability of Large Language Model Generated Clinical Reasoning in Assisted Reproductive Technology: Blinded Comparative Evaluation Study

Dou Liu[1,2,3*], BSc; Ying Long[1,3,4*], MD; Sophia Zuoqiu[5], PhD; Di Liu[5,6,7], PhD; Kang Li[6,7], PhD; Yiting Lin[8], MBBS; Hanyi Liu[8], MBBS; Rong Yin[5,6,7*], PhD; Tian Tang[1,3,4*], MD

[1]Department of Obstetrics and Gynecology, West China Second University Hospital of Sichuan University, Chengdu, China

[2]Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI, United States

[3]Key Laboratory of Birth Defects and Related Diseases of Women and Children, Sichuan University, Chengdu, China

[4]Reproductive Medical Center, Department of Obstetrics and Gynecology, West China Second University Hospital of Sichuan University, Chengdu, China

[5]Department of Industrial Engineering, Sichuan University, Chengdu, China

[6]West China Biomedical Big Data Center, West China Hospital of Sichuan University, Chengdu, China

[7]Med-X Center for Informatics, Sichuan University, Chengdu, China

[8]West China School of Medicine, Sichuan University, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Tian Tang, MD
Department of Obstetrics and Gynecology
West China Second University Hospital of Sichuan University
Chengdu
China
Phone: 1 18081110066
Email: tiantang2016@scu.edu.cn

Abstract

Background: High-quality clinical chains-of-thought (CoTs) are essential for explainable medical artificial intelligence (AI); yet, their development is limited by data scarcity. Large language models can generate medical CoTs, but their clinical reliability is unclear.

Objective: We evaluated the clinical reliability of large language model–generated CoTs in reproductive medicine and examined prompting strategies to improve their quality.

Methods: In a blinded comparative study at a clinical center, senior clinicians in assisted reproductive technology evaluated CoTs generated via 3 distinct strategies: zero-shot, random few-shot (using random shallow examples), and selective few-shot (using diverse, high-quality examples). Expert ratings were then compared with evaluations from a state-of-the-art AI model (GPT-4o).

Results: The selective few-shot strategy significantly outperformed other strategies across logical clarity, use of key information, and clinical accuracy (P<.001). Critically, the random few-shot strategy offered no significant improvement over the zero-shot baseline, demonstrating that low-quality examples are as ineffective as no examples. The success of the selective strategy is attributed to 2 preliminary frameworks: “gold-standard depth” and “representative diversity.” Notably, the AI evaluator failed to discern these critical performance differences. Thus, clinical reliability depends on strategic prompt design rather than simply adding examples.

Conclusions: We propose a “dual principles” preliminary framework for generating trustworthy CoTs at scale in assisted reproductive technology. This work is a preliminary step toward addressing the data bottleneck in reproductive medicine. It also underscores the essential role of human expertise in evaluating generated clinical data.

(J Med Internet Res 2026;28:e85206) doi:10.2196/85206

XSL•FO
RenderX

KEYWORDS

chain-of-thought; large language model; assisted reproductive technology; explainable artificial intelligence; clinical data reliability

Introduction

Background

Assisted reproductive technology (ART) represents a cornerstone of modern medicine, offering solutions for millions facing infertility [1]. The clinical decision-making process in ART is exceptionally complex, requiring the synthesis of high-dimensional patient data, including baseline characteristics and medical history. This process is time-consuming and fraught with risk for both clinicians and patients, as minute variations in treatment protocols can lead to significant adverse outcomes. Furthermore, clinicians must navigate patients' personal values and ethical considerations, demanding a highly personalized and explainable approach to care [2].

Recent advancements in artificial intelligence (AI), particularly large language models (LLMs), have demonstrated considerable promise for answering medical questions, addressing clinical case challenges, and augmenting clinical diagnosis [3-7]. Within clinical decision support systems, these technologies can help synthesize large amounts of data, facilitating more comprehensive and standardized therapeutic strategies. However, while general-purpose LLMs like ChatGPT-4 and Gemini are powerful, their training on broad, nonspecialized data limits their utility in niche medical domains. Consequently, high-performing clinical AI applications are typically fine-tuned from general models using curated, domain-specific datasets [8-10]. The actual bottleneck, however, is not a lack of raw clinical data, but a profound lack of explainable data—data that record not just what decision was made, but why. This meticulous, expert-level reasoning, often captured as a chain-of-thought (CoT), is the very fuel required to train AI models that are not just accurate, but also trustworthy and scalable to clinicians. To move beyond "black-box" predictions, models require structured reasoning pathways, or CoT data, which simulate clinical logic and enhance explainability [11,12]. The challenge, therefore, narrows down to a scarcity of expert-authored CoT data within the specific area. The manual creation of such a dataset on a large scale is prohibitively expensive and time-consuming, presenting a significant barrier to progress in explainable medical AI.

To address this challenge, a promising direction is to leverage the generative capabilities of LLMs to synthesize clinical CoT data at scale. While this offers a scalable solution to the data bottleneck, it hinges on a critical, unverified assumption: the clinical reliability of the generated content. In a high-stakes domain like ART, this assumption cannot be taken for granted.

Therefore, this study is designed to examine this uncertainty under the ART setting through a rigorous, head-to-head empirical comparison. Figure 1 presents the conceptual framework of our comparative evaluation study. We hypothesize that a selective few-shot strategy, meticulously crafted with diverse and deeply reasoned examples, will improve the factual accuracy, logical clarity, and clinical information use of LLM-generated reasoning in infertility decision-making scenarios, compared with generic zero-shot and random few-shot prompting. To test this, we developed a novel prompting framework and validated it through a blinded evaluation protocol where senior clinicians from the reproductive department assessed the quality of CoTs from all 3 strategies. In a secondary analysis, we further contrast these expert assessments against the state-of-the-art (SOTA) AI evaluator (GPT-4o) at that time to critically examine the current capabilities and limitations of automated evaluation paradigms, which are widely used in supervised data generation. Ultimately, this work aims to establish an exploratory, practical, evidence-based methodology for the trustworthy generation of clinical reasoning of ART at scale.

XSL•FO
RenderX

Figure 1. Conceptual framework of the comparative evaluation study. EHR: electronic health record; LCC: logical coherence and clarity; LLM: large language model; PCAR: plausibility and clinical accuracy of reasoning; UCKI: use and coverage of key information.

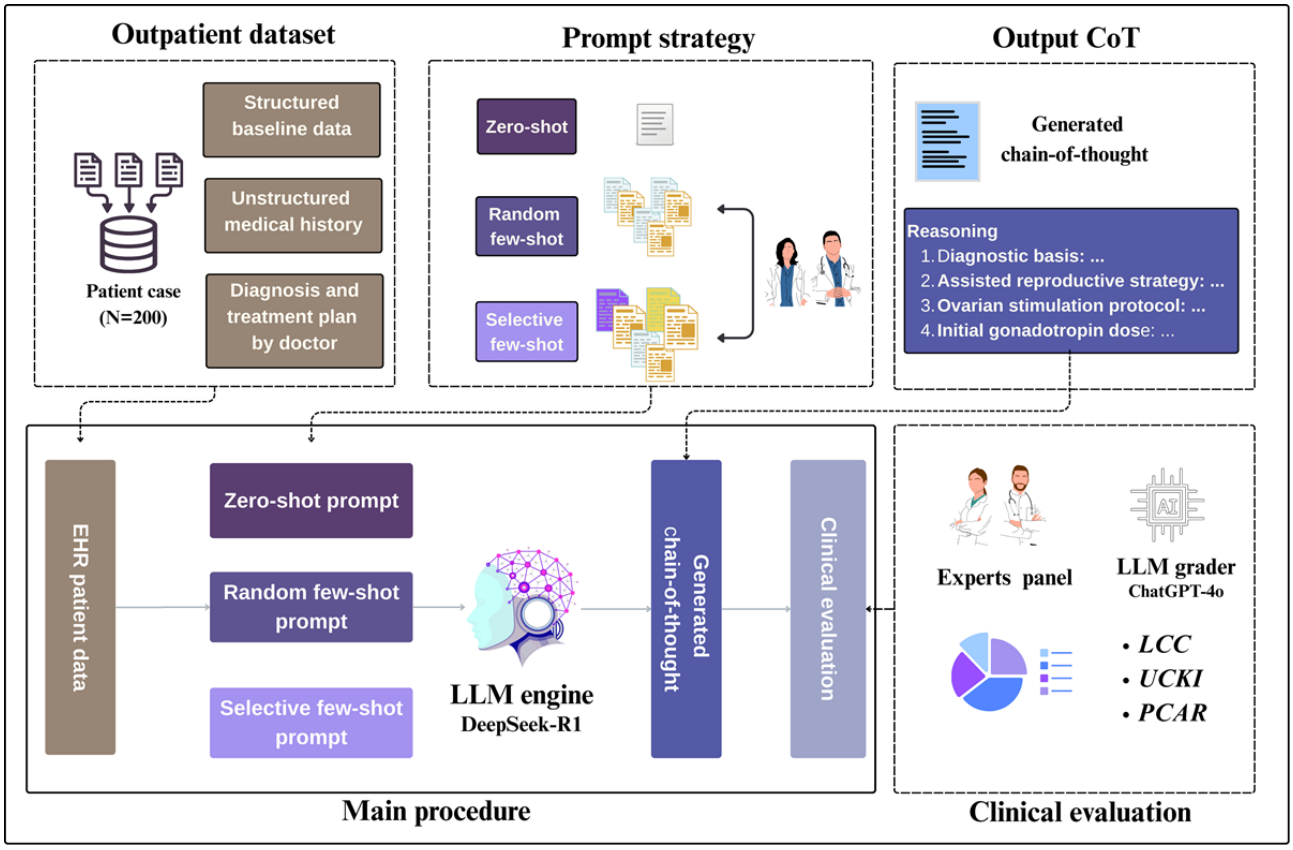

The study workflow begins with a standardized patient case (N=200) as the input. Three distinct prompting strategies are evaluated in parallel pipelines: (1) zero-shot, which uses instructions only; (2) random few-shot, which uses 5 randomly selected examples; and (3) selective few-shot, which uses a curated set of 6 diverse examples representing all major ART categories in the cases. Each strategy, powered by the same LLM engine, generates a unique CoT. All 3 generated CoTs are then subjected to a rigorous, blinded "Doctor-in-the-Loop" evaluation by 2 parallel assessors: human clinical experts (the gold standard) and the SOTA AI evaluator (GPT-4o). This dual evaluation process yields the final reliability scores and rankings for each strategy.

LLMs in Health Care

Since the launch of ChatGPT-4, LLMs have rapidly spread into many industries, such as education, finance, and health care. For instance, Google's Med-PaLM 2, a leading specialized health care model, achieved 86.5% accuracy on the MedQA benchmark, a popular multiple-choice open domain question answering (OpenQA) medical problems dataset. Furthermore, its responses were preferred over those of generalist physicians in 65% of expert evaluations [13]. The LLMs are now used in many healthcare–related workflows, ranging from medical documentation assistance to clinical differential diagnosis [5,14-16]. However, to effectively address highly specialized tasks, these models are typically fine-tuned from pretrained LLMs using carefully curated datasets. Despite their impressive capabilities, current LLMs usually function as black boxes, producing outputs without offering interpretable reasoning. In clinical practice, however, physicians often require not just answers but also transparent explanations. This requires models beyond black-box behavior and provides interpretable, step-by-step reasoning processes. The increasing reliance on LLMs has also intensified the demand for high-quality data [17]. Alarmingly, some predict that the global supply of novel text data may be exhausted by 2050 and image data by 2060 [18]. In the health care domain, the situation is even more critical: clinical data are not only scarce but also highly sensitive and expensive to obtain. As a result, a central challenge emerges—how can we build datasets that are both sufficiently large and clinically trustworthy to support transparent, reliable medical AI systems?

Synthetic Data in Health Care

To overcome the data shortage in health care, researchers are increasingly turning to LLMs to create synthetic data. This approach is promising for several reasons. It allows for data generation at scale, addressing issues of data scarcity and privacy [19,20]. Furthermore, synthetic data can be tailored to balance underrepresented patient groups, potentially improving model robustness and fairness [21]. Generative models have demonstrated remarkable success in these areas, with some studies showing that LLM-generated narratives can be indistinguishable from those written by physicians [22]. This potential, however, is inextricably linked to a profound challenge: reliability. While LLMs can mimic the style of clinical text, ensuring the factual accuracy and clinical plausibility of the content is a far more difficult task. For instance, models have been used to generate both structured

XSL•FO
RenderX

electronic health records (EHRs) and unstructured clinical notes [23], but in both cases, the risk of hallucination—where the model generates incorrect or nonsensical information—poses a significant threat in high-stakes medical applications. Therefore, the core challenge moves beyond mere data generation to a more fundamental question of trust. While studies have shown that synthetic data can be effective for certain human labeling and fine-tuning tasks [24-28], these applications often involve relatively straightforward data points. The problem is magnified when the task requires complex, multistep medical reasoning. In such scenarios, "synthetic" must not equate to "inaccurate." This underscores the urgent need for rigorous evaluation methods, not just for the data points themselves, but for the underlying reasoning processes that produce clinical decisions. Our work focuses on this critical next step: assessing the reliability of synthetically generated reasoning paths.

Chain-of-Thought

CoT is a prompt engineering technique that enhances the output of LLMs, particularly for complex tasks involving multistep reasoning. It facilitates problem-solving by guiding the model through a step-by-step reasoning process by using a coherent series of logical steps [29]. This approach has been shown to significantly elevate performance on a wide range of complex reasoning tasks in general domains, especially for arithmetic problems and logical reasoning tasks [30,31]. To enhance the reasoning ability in domain-specific tasks, researchers have started fine-tuning the models with CoTs [12]. Within the medical domain, the potential of CoT is particularly compelling. Its step-by-step nature aligns naturally with the differential diagnosis and clinical reasoning processes used by physicians. Consequently, researchers have begun to apply CoT prompting to improve accuracy on medical question-answering benchmarks and in practice diagnosis [11,32]. More importantly, CoT offers a crucial pathway toward explainable AI in medicine. By externalizing the model's reasoning processes, CoT allows clinicians to scrutinize, understand, and ultimately trust the AI's recommendations, which is a prerequisite for its safe integration into clinical workflows.

The application of CoT is rapidly evolving. Beyond simple prompting, a new frontier in clinical AI is the (1) fine-tuning of models on datasets enriched with CoT data to build inherently more explainable systems, which, however, immediately confronts the fundamental bottleneck of medical AI; and the (2) prohibitive cost and time required for expert clinicians to manually author thousands of high-quality reasoning paths for a training set. An intuitive and scalable solution is to leverage foundational LLMs to synthetically generate these CoTs, creating a cost-effective pathway to train the next generation of trustworthy medical models. However, this entire paradigm hinges on a critical, yet largely unexamined, question: (3) Is the reliability of synthetically generated CoTs adequate to support their application in complex clinical scenarios? Literature to date offers little guidance. Most research focuses on the extrinsic value of CoT (ie, improving final answer accuracy), with scant attention paid to the intrinsic reliability of the reasoning itself. A model fine-tuned on flawed, albeit synthetically generated, logic could learn to produce seemingly correct answers for the wrong reasons—a risk that is unacceptable in clinical practice. Furthermore, standardized, expert-driven protocols for assessing the clinical validity, coherence, and faithfulness of machine-generated reasoning are conspicuously absent. Our study is designed to directly fill this foundational gap. Before the field can confidently use synthetic CoT for model training at scale, we must first have a rigorous method to measure its reliability. Therefore, we propose and implement a blinded, expert-led evaluation framework to answer the fundamental question: how reliable is synthetically generated sophisticated clinical reasoning, and what is the best prompting strategy to elicit it from LLMs?

Methods

Data Source

From the manually reviewed dataset, we randomly selected 200 cases as our evaluation set, covering a variety of ART. These ARTs are broadly categorized into 3 generations: in vitro fertilization (IVF), intracytoplasmic sperm injection (ICSI), and preimplantation genetic testing (PGT). Each generation includes several clinical subtypes, such as short-protocol IVF and IVF with donor sperm. Among the 200 evaluated cases, IVF accounts for the largest proportion (140/200, 70%), including standard IVF (116/200, 58%), IVF with donor sperm (9/200, 4.5%), and short-protocol IVF (short-time insemination, 15/200, 7.5%). The second most common is ICSI (38/200, 19%), comprising standard ICSI (26/200, 13%), IVF+ICSI (5/200, 2.5%), and TESA (testicular sperm aspiration)+ICSI (7/200, 3.5%). PGT represents 11% (22/200) of the dataset, including preimplantation genetic testing for aneuploidies (PGT-A; 6/200, 3%), preimplantation genetic testing for monogenic disorder (PGT-M; 3/200, 1.5%), and PGT for structural rearrangements (13/200, 6.5%).

As shown in Table 1, the dataset consisted of three main components: (1) a structured set of baseline and demographic variables, (2) the preliminary diagnosis and treatment plan, and (3) an unstructured narrative description of the present illness history. The structured baseline data served as the quantitative and categorical foundation for clinical assessment, encompassing key indicators of ovarian reserve such as anti-Müllerian hormone and baseline follicle-stimulating hormone levels. The unstructured narrative provided essential clinical context, offering a detailed account of the patient's medical journey, which is critical for nuanced and context-aware medical reasoning. The output data, labeled as preliminary diagnosis and treatment plan, reflects the clinical conclusions and therapeutic strategies formulated by human experts in prior encounters. This component serves as the ground truth outcome. The LLM's task is to generate a reasoning path that logically connects the patient's input data to this expert-defined outcome. A detailed breakdown of all case data variables, including structured baseline indicators and narrative clinical history, is provided in Multimedia Appendix 1. Together, these inputs formed the foundation for CoT generation and model evaluation.

XSL•FO
RenderX

Table 1. Structure and description of input and output variables for each case[a].

Category	Variable
Baseline and demographics	• Female age • Menstrual cycle • Body weight • BMI • Anti-Müllerian hormone level • Duration of infertility • Gynecological ultrasound findings • Baseline follicle-stimulating hormone level
Present illness history	• Present illness history
Preliminary diagnosis and treatment plan	• Type of infertility • Controlled ovarian stimulation protocol • Initial gonadotropin dosage • Preliminary differential diagnosis • Initial assisted reproductive technology strategy

[a]This table outlines the variables provided to the large language model for each case, categorized into input (baseline and demographic and present illness history) and output (preliminary diagnosis and treatment plan). These variables form the basis for the chain-of-thought generation task.

Ethical Considerations

A set of EHRs recorded between 2020 and 2022 in the Infertility Outpatient Department at West China Second University Hospital was considered in this study. The study was approved by the institutional review board of West China Second University Hospital, Sichuan University (ID: 2022288). The EHRs have been manually reviewed and corrected to ensure data accuracy. All EHR data used in this study were fully deidentified before being accessed by the research team. The deidentification procedure followed HIPAA (Health Insurance Portability and Accountability Act) Safe Harbor standards, including the removal of all direct identifiers (eg, name, date of birth, medical record number, contact information, and provider information) and all quasi-identifiers (eg, dates, locations, and institutional identifiers). Only aggregated clinical descriptors necessary for the reasoning task (eg, high-level patient history and laboratory summaries) were retained. Access to the deidentified dataset was restricted to authorized study personnel through institution-managed credentials and encrypted storage. Reviewers performing the blinded evaluation accessed only the deidentified clinical vignettes and model-generated reasoning content through a secure, read-only interface; no downloads or reidentification attempts were permitted. All access was logged and monitored by an internal auditor to ensure compliance with institutional clinical data governance policies. Because the study used retrospective, fully deidentified EHRs, the requirement for updated informed consent for the following analysis was waived by the institutional review board in accordance with national regulations. No participants were contacted for this study, and no compensation was provided to participants.

Experiment Design

Overview

To systematically evaluate the reliability of LLM-generated CoT and to determine the impact of different prompting strategies, we designed a comparative study. The experiment was structured into 3 distinct arms, each representing a different level of contextual information provided to the model. Our design philosophy was to create a controlled, stepwise comparison to isolate the effects of in-context examples and the strategy used for their selection.

All 3 groups used the evaluation dataset (N=200) described in the data source, making sure of a fair comparison. A capable "Teacher Model" is key to generating better-quality data [33]. Considering the models' performance so far, we used the open-source model DeepSeek-R1-671b, which was known for its outstanding reasoning capability, as our consistent model shared by 3 arms [34]. Across all 3 arms, the LLM was assigned the same core task: generating a detailed, step-by-step ART CoT by integrating all provided patient information and the corresponding expert-provided reference output. All the inference was conducted via the application programming interface call. The model was executed with temperature=0.5 and max tokens=5000. We used temperature=0.5 to reflect typical clinical-LLM use, where deterministic decoding (temperature=0) may produce rigid or incomplete clinical reasoning. All prompting strategies were evaluated under identical generation settings to ensure fair comparison.

Group 1: Zero-Shot Baseline

In this group, we aimed to establish a fundamental baseline to evaluate the out-of-the-box clinical reasoning capabilities of general-purpose LLMs when applied to this specialized task. To this end, the model was prompted using a standardized directive instruction, with each clinical case embedded directly into the prompt (see Multimedia Appendix 2 for details). The outputs generated by the model, along with corresponding physician evaluations, served as a performance floor, quantifying the baseline reliability and limitations of an unadapted LLM in handling novel clinical scenarios.

Group 2: Random Few-Shot Prompting

This experimental arm was designed to establish a baseline for a standard, nonoptimized few-shot approach. Its purpose was

XSL•FO
RenderX

to measure the impact of providing generic, in-domain examples without a specific selection strategy. For each of the 200 test cases, the prompt was initially prepared with a fixed set of 5 examples to provide context for the model. These 5 examples were randomly sampled from our expert-authored data pool, excluding the existing evaluation dataset. The sample set used in the prompt for every test case consisted of 4 standard IVF cases and 1 short-protocol IVF case, accompanied by a concise reasoning chain authored by domain experts. The prompt structure and instructions were otherwise identical to those in the other arms. A representative example of a few-shot sample, detailing the input data and expert-written CoT, is provided in Multimedia Appendix 2. This approach represents a "naive" few-shot implementation. It is designed to test the hypothesis that the mere presence of in-domain examples, even without being specifically tailored to the test case, is sufficient to improve reasoning quality compared to the zero-shot baseline.

Group 3: Selective Few-Shot Prompting

This arm represents our proposed method and was designed to test the hypothesis that a deliberately curated set of diverse examples would improve reasoning reliability and generalization. Instead of random sampling, this approach used a clinically informed, representative selection strategy. Physicians curated a set of 6 exemplary cases from a pool of records not included in the 200 evaluation set (to prevent data leakage). These 6 examples were specifically chosen to represent the full spectrum of major ART categories present in our dataset, including IVF (standard IVF, short-protocol IVF, and IVF with donor sperm), ICSI, TESA+ICSI, and PGT (PGT-A). Their reasoning part was carefully crafted and covered all critical steps. The complete prompt is provided in Multimedia Appendix 2. For every test case, this same curated set of 6 diverse examples was prepended to the prompt. The purpose of this strategy was to provide the model with comprehensive and representative clinical guidance within the prompt itself.

In summary, this 3-arm design allows for a multifaceted analysis of CoT reliability. The comparison between group 1 and group 2 will isolate the general benefit of using in-context examples. The critical comparison between group 2 and group 3 will determine whether our proposed selective prompting strategy provides a statistically significant improvement over a random baseline. Together, these comparisons will build a clear evidence-based argument for the importance of a well-designed prompting strategy in generating reliable and accurate clinical reasoning.

Evaluation Metrics

Physician Evaluation

The evaluation was conducted by a panel of 2 board-certified reproductive physicians. Both subject matter experts are faculty members at the same academic medical center but work independently in separate clinical teams. They were invited through an internal clinical research collaboration mechanism, and participation was voluntary. Their sole role was to perform a blinded evaluation of the CoTs. Each evaluator possesses over 10 years of clinical experience in the field of ART. Prior to the formal evaluation, a calibration session was held where all evaluators scored 10 cases together. Any discrepancies were discussed to ensure a consistent understanding of the criteria. During the blinded evaluation, each physician reviewer received only the clinical vignette and the model-generated reasoning content. All identifying information, model names, and prompting strategies were removed to ensure full masking. The order of cases and strategies was independently randomized for each reviewer to prevent any presentation bias. Each physician independently scored the complete dataset without discussing their ratings with other reviewers. No reviewer saw any ground-truth labels or model metadata during the evaluation process.

In this study, we created an evaluation metric involving 3 dimensions: logical coherence and clarity (LCC), use and coverage of key information (UCKI), and plausibility and clinical accuracy of reasoning (PCAR). All generated CoTs were scored by the 5-point Likert scale (1=poor and 5=excellent) across 3 key dimensions of reliability, as shown in Table 2 and detailed in Multimedia Appendix 3. All paired comparison results are tested using the Wilcoxon test and adjusted for false discovery rate (FDR).

Table 2. Rubric for the evaluation of chain-of-thought (CoT) reliability[a].

Metric	Definition
Logical coherence and clarity	Assesses whether the reasoning process is internally consistent, logically structured, and expressed clearly and understandably.
Use and coverage of key information	Evaluates the extent to which the reasoning incorporates and addresses relevant clinical data points presented in the input.
Plausibility and clinical accuracy of reasoning	Measures whether the reasoning is clinically sound, aligns with standard medical knowledge, and leads to a reasonable interpretation or decision. Deduct points as appropriate across the 4 parts in the analysis.

[a]The table defines the 3 dimensions: logical coherence and clarity, use and coverage of key information, and plausibility and clinical accuracy of reasoning, used by both human experts and the artificial intelligence evaluator to assess the quality of generated CoTs on a 5-point Likert scale.

AI Grader Evaluation

In addition to manual evaluation conducted by human experts, we implemented a supplementary evaluation component leveraging a widely used LLM verifier [35], GPT-4o, to explore its feasibility as an automated evaluator of clinical reasoning. This design enables a direct comparison between AI-generated assessments and the human gold standard, thereby evaluating the feasibility of using LLMs for quality control in large-scale synthetic dataset generation. To ensure comparability, the evaluation criteria provided to the AI model were identical to those outlined in Table 2, including definitions of logical

XSL•FO
RenderX

coherence, clinical appropriateness, and key information use. The detailed prompt can be found in Multimedia Appendix 2. Each instruction consists of 2 blocks: the evaluation rubric and case vignette.

Statistical Analysis

All statistical analyses were conducted in Python (pandas, SciPy, and statsmodels; Python Software Foundation). Each case (N=200) was independently evaluated under 3 prompting strategies (zero-shot, random few-shot, and selective few-shot) across 3 dimensions: LCC, UCKI, and PCAR. Results are reported as mean (SD), with exact n values indicated in tables. To assess the reliability of the physician ratings, we evaluated agreement using three complementary measures: (1) linear weighted κ, (2) adjacent agreement rate (percentage of paired scores within ±1 point), and (3) raw interrater disagreement rates. Because the rating distributions exhibited strong ceiling effects (scores concentrated around 4-5), linear weighted κ is known to underestimate agreement under low variance conditions, a well-described statistical paradox. Therefore, in addition to reporting κ values for completeness, we emphasized adjacent agreement and disagreement rates, which more accurately capture clinical consensus when rating scales are narrow. Across all 600 evaluations (200 cases×3 strategies), disagreement rates were uniformly low, confirming strong consistency between the 2 raters. The selective few-shot strategy showed the lowest disagreement (LCC: 0.00, PCAR: 0.03, UCKI: 0.02), followed by the zero-shot (LCC: 0.04, PCAR: 0.04, UCKI: 0.07) and random few-shot strategies (LCC: 0.12, PCAR: 0.09, UCKI: 0.07). The highest disagreement observed across all metrics was only 12%. Adjacent agreement was correspondingly high, ranging from 88% to 100% depending on the metric and strategy, indicating excellent practical concordance despite the compressed scoring range. Given these properties, interrater reliability was interpreted primarily through adjacent agreement and disagreement rates, while κ statistics were retained as a formal but secondary indicator.

Normality of paired differences was assessed using the Shapiro-Wilk test and Q-Q plots. Because the evaluation scores are 5-point Likert ratings, the paired differences take on only a small number of discrete values (–2, –1, 0, +1, +2). As expected with large samples and discrete Likert data, the Shapiro-Wilk was extremely sensitive and returned significant results; however, Q-Q plots showed no meaningful deviations from approximate linearity beyond the expected discreteness. Parametric tests are generally robust to moderate deviations from normality in Likert-type data, especially with larger sample sizes. We retained the nonparametric Wilcoxon paired tests as primary analyses. All the statistical test details can be checked in Multimedia Appendix 4.

To account for multiple pairwise comparisons across the prompting conditions and evaluation metrics in the subgroups, adjusted *P* values were calculated using the Benjamini-Hochberg FDR correction across all 9 contrasts (3 comparisons per metric). For each metric, paired comparisons were conducted using Wilcoxon paired tests, and both the raw and FDR-adjusted *P* values were reported. A post-hoc power analysis was performed. For each evaluation metric, Cohen *d* effect sizes were computed using the pooled SD of paired observations. This analysis was conducted to assess the stability of estimates under small-sample conditions.

Results

Overview

All the results were obtained through the evaluation dataset (N=200), including several kinds of ART. As mentioned earlier, 3 metrics were used for evaluation: LCC, UCKI, and PCAR. The evaluation was done by a panel of experienced practitioners.

General Performance

Table 3 presents the average scores of each prompting strategy on LCC, UCKI, and PCAR. The selective few-shot strategy outperformed both zero-shot and random few-shot approaches across all 3 metrics. Specifically, it achieved mean scores of 4.56 (SD 0.50), 4.66 (SD 0.53), and 4.18 (SD 0.56), which were significantly higher than those of the zero-shot strategy (mean 4.18, SD 0.56; mean 4.30, SD 0.63; mean 3.85, SD 0.53, respectively; all *P* and adjusted *P*<.001; Cohen *d*=0.72, 0.61, 0.61) and the random few-shot strategy (mean 4.31, SD 0.64; mean 4.42, SD 0.58; mean 3.91, SD 0.63, respectively; all *P* and adjusted *P*<.001; Cohen *d*=0.45, 0.42, 0.46).

Table 3. The performance of the "zero-shot," "random few-shot," and "selective few-shot" strategies (N=200 cases per group)[a].

Strategy	LCC[b], mean (SD)	UCKI[c], mean (SD)	PCAR[d], mean (SD)
Zero-shot	4.18 (0.56)	4.30 (0.63)	3.85 (0.53)
Random few-shot	4.31 (0.64)	4.42 (0.58)	3.91 (0.63)
Selective few-shot	4.56 (0.50)	4.66 (0.53)	4.18 (0.56)

[a]Wilcoxon test used for paired comparisons; *P* values adjusted using the Benjamini-Hochberg false discovery rate procedure.

[b]LCC: logical coherence and clarity.

[c]UCKI: use and coverage of key information.

[d]PCAR: plausibility and clinical accuracy of reasoning.

Notably, there was no statistically significant difference between the zero-shot and random few-shot groups on PCAR, though samples did improve the model's capability on LCC and UCKI statistically significantly.

Subgroup Analysis

To further dig into the reasons for selective few-shot's winning, we did an analysis grouped by ART. Table 4 presents the scores of 3 ART generations.

XSL•FO
RenderX

Table 4. Subgroup analysis by assisted reproductive technology (ART) category[a].

ART and strategy	LCC[b]	UCKI[c]	PCAR[d]
IVF[e]			
Zero	4.20 (0.57)	4.34 (0.61)	3.88 (0.53)
Random	4.29 (0.67)	4.44 (0.53)	3.90 (0.65)
Selective	4.59 (0.49)	4.69 (0.51)	4.20 (0.55)
ICSI[f]			
Zero	4.16 (0.59)	4.29 (0.65)	3.87 (0.53)
Random	4.37 (0.54)	4.45 (0.69)	3.97 (0.59)
Selective	4.45 (0.50)	4.53 (0.60)	4.11 (0.56)
PGT[g]			
Zero	4.09 (0.43)	4.05 (0.72)	3.64 (0.49)
Random	4.32 (0.57)	4.27 (0.70)	3.86 (0.56)
Selective	4.59 (0.50)	4.68 (0.48)	4.18 (0.59)

[a]Subgroup analyses were based on the respective case counts (IVF: n=140; ICSI: n=38; PGT: n=22). This table aims to further investigate the performance differences of various prompting strategies across specific clinical scenarios. To achieve this, we categorized the 200 evaluation cases based on their primary type of ART, including IVF, ICSI, and PGT, and conducted a comparative analysis of evaluation outcomes within each group. Wilcoxon test used for paired comparisons; *P* values adjusted using the Benjamini-Hochberg false discovery rate procedure.

[b]LCC: logical coherence and clarity.

[c]UCKI: use and coverage of key information.

[d]PCAR: plausibility and clinical accuracy of reasoning.

[e]IVF: in vitro fertilization.

[f]ICSI: intracytoplasmic sperm injection.

[g]PGT: preimplantation genetic testing.

In the largest subgroup, IVF (n=140), a key distinction emerged. While the selective few-shot strategy significantly outperformed both other groups across all metrics ($P<.001$ and adjusted $P<.001$ for all comparisons, selective vs zero: Cohen $d=0.72$, 0.61, 0.59; selective vs random: Cohen $d=0.51$, 0.49, 0.50), there was no statistically significant difference observed between the random few-shot and zero-shot strategies ($P=.19$, .69, .10; adjusted $P=.22$, .69, .13).

The analysis of the PGT subgroup (n=22) revealed the clearest advantage for prompt diversity. The selective few-shot strategy, which was the only prompt containing a PGT example, outperformed the random few-shot strategy across all 3 metrics: logical coherence (LCC: $P=.03$; adjusted $P=.05$; Cohen $d=0.51$), information use (UCKI: $P<.001$; adjusted $P=.01$; Cohen $d=0.55$), and clinical accuracy (PCAR: $P=.03$; adjusted $P=.05$; Cohen $d=0.68$). Consistent with other findings, the random few-shot strategy showed no significant improvement over the zero-shot baseline in this category (LCC: $P=.17$; adjusted $P=.19$; Cohen $d=0.45$; UCKI: $P=.27$; adjusted $P=.27$; Cohen $d=0.32$; PCAR: $P=.13$; adjusted $P=.17$; Cohen $d=0.43$, respectively). However, we must admit that the limited sample size makes the subgroup analysis partially underpowered, and these comparisons should be interpreted with caution.

A similar pattern emerged in the ICSI subgroup (n=38). The selective few-shot strategy again demonstrated a measurable advantage. It achieved statistically significant improvements over the zero-shot baseline in 2 of the 3 key metrics: LCC ($P=.02$; adjusted $P=.15$; Cohen $d=0.53$) and PCAR ($P=.04$; adjusted $P=.17$; Cohen $d=0.44$). Although these comparisons did not remain significant after FDR adjustment, the effect sizes were moderate, and the directionality was consistent with the overall findings. For UCKI, the selective strategy again achieved the highest mean score, but this comparison did not reach significance ($P=.06$; adjusted $P=.17$), suggesting a positive but statistically inconclusive trend. The detailed subgroup statistical results can be checked in Multimedia Appendix 4.

Case Study

As shown in Figure 2, to qualitatively illustrate the stark differences in reasoning quality revealed by our quantitative analysis, we selected a representative and complex case involving PGT-M. This case is particularly illustrative, as it requires a multilayered understanding of genetics, ART procedures, and individualized patient factors. The main mistakes are listed in Table 5.

XSL•FO
RenderX

Figure 2. Representative PGT-M case illustrating qualitative differences in CoT reasoning across prompting strategies. This figure presents a representative and complex case involving PGT-M, selected to qualitatively illustrate the differences in reasoning quality observed in our quantitative analyses. The left panel shows the patient's clinical information, the correct physician's answer, and the color-coded annotation scheme (red: incorrect reasoning, yellow: irrelevant reasoning, and green: correct reasoning). The right panel displays the CoT outputs generated under zero-shot, random few-shot, and selective few-shot prompting strategies. Compared with the zero-shot and random few-shot generations, which omitted critical reasoning steps (eg, the presence of infertility diagnosis, the indication for intracytoplasmic sperm injection, and comprehensive gonadotropin dose considerations), the selective few-shot prompting was more closely aligned with clinical logic and included relevant patient-specific factors. CoT: chain-of-thought; PGT-M: preimplantation genetic testing for monogenic disorder.

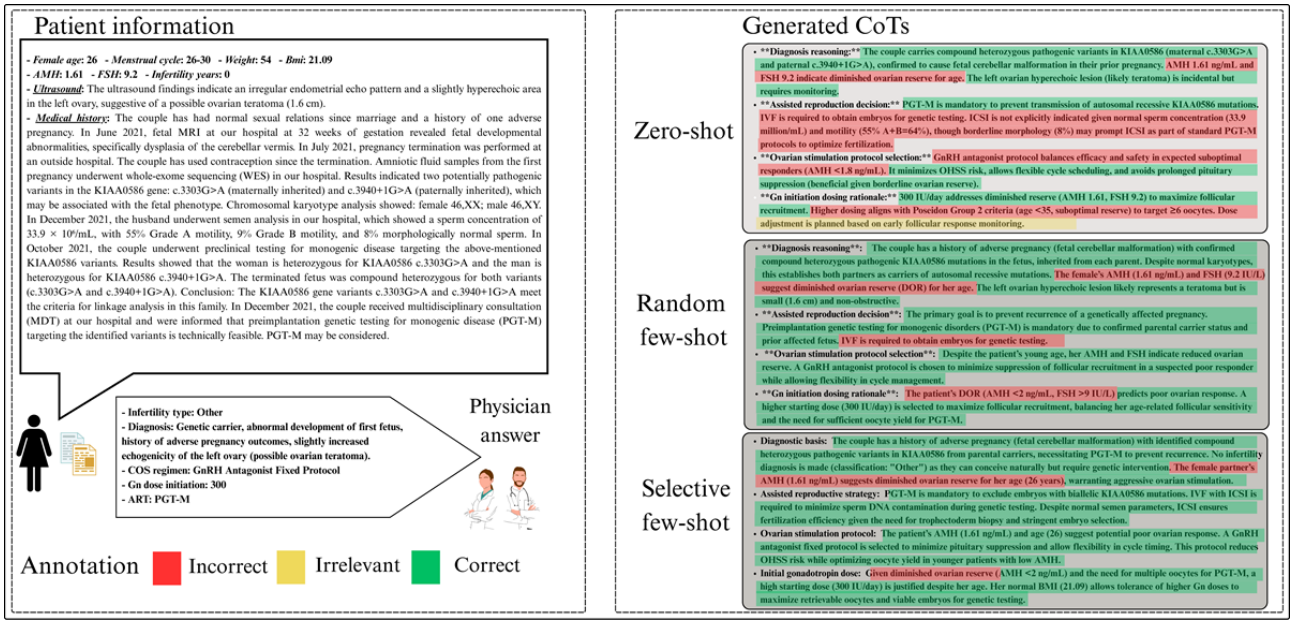

Table 5. Common reasoning errors in zero-shot and random few-shot chain-of-thought (CoT) outputs for a preimplantation genetic testing for monogenic disorder case.

Reasoning dimension	Flaws in zero-shot and random few-shot
Diagnosis reasoning	The model does not mention whether the patient has infertility issues.
Assisted reproduction decision	CoT incorrectly assumes that if the male's semen is normal, traditional IVF[a] can be used.
Ovarian stimulation protocol selection	The reason for choosing the antagonist protocol in CoT was "greater safety and avoidance of OHSS[b]," without considering the patient's specific circumstances (low AMH, first ovulation induction).
Initial gonadotropin dose	Only AMH[c] levels were considered, without taking into account weight, BMI, or PGT[d] goals (requiring more embryos).

[a]IVF: in vitro fertilization.

[b]OHSS: ovarian hyperstimulation syndrome.

[c]AMH: anti-Müllerian hormone.

[d]PGT: preimplantation genetic testing.

In this PGT-M case, both partners are carriers of a pathogenic variant in the KIAA0586 gene. During a previous pregnancy, the fetus was found to have a homozygous mutation in KIAA0586, resulting in abnormal brain development and subsequent pregnancy termination. Since then, the couple has been using contraception and therefore does not meet the criteria for an infertility diagnosis. This implies that they are still capable of conceiving naturally. Given the autosomal recessive inheritance pattern, there remains a possibility of achieving a normal or carrier embryo through natural conception. However, neither the zero-shot nor the random few-shot–prompted CoT generations mentioned the presence or absence of an infertility diagnosis, which appeared in the selective few-shot–prompted CoT.

To avoid the recurrence of a fetus with a homozygous mutation in KIAA0586, PGT-M is recommended. Due to the technical requirements of PGT, embryos must be obtained via ICSI to avoid DNA contamination during genetic analysis. While the zero-shot and random few-shot–prompted CoTs correctly reasoned the indication for PGT-M, they incorrectly concluded that ICSI was unnecessary because the male partner had normal semen parameters and suggested using conventional IVF instead—an error in clinical reasoning.

In selecting the ovarian stimulation protocol, clinical reasoning typically begins with evaluating the patient's ovarian responsiveness and any prior stimulation history. Although the patient is 26 years of age, her AMH level is only 1.61 ng/mL, suggesting a potential for diminished ovarian response. As this is her first controlled ovarian hyperstimulation cycle, a

XSL•FO
RenderX

gonadotropin-releasing hormone antagonist protocol was chosen for its controllability and to avoid excessive pituitary suppression. Among the 2 few-shot–prompted CoTs, the reasoning was more aligned with clinical thinking, while the zero-shot CoT emphasized the safety profile of the antagonist protocol (eg, avoiding ovarian hyperstimulation syndrome) without clearly reflecting clinical logic.

Regarding the initial gonadotropin dose, factors beyond ovarian responsiveness must be considered. Since this case involves PGT, it is important to optimize the number of oocytes retrieved. Additional considerations include the patient's weight and BMI, as these affect drug sensitivity. However, the zero-shot CoT mentioned only ovarian responsiveness, lacking a comprehensive rationale.

Feasibility Analysis of an AI Evaluator

As detailed in Table 6, the mean scores for all 3 prompting strategies were tightly clustered in a narrow and high-scoring range, between 3.96 and 4.00, suggesting that the model perceived all generated outputs as being of similarly high quality.

Table 6. Artificial intelligence (AI)–driven evaluation of chain-of-thought reliability across different prompting strategies[a].

Group	LCC[b], mean (SD)	PCAR[c], mean (SD)	UCKI[d], mean (SD)
Random few-shot	4.00 (0.00)	3.98 (0.14)	4.00 (0.00)
Selective few-shot	4.00 (0.00)	3.98 (0.16)	4.00 (0.07)
Zero-shot	4.00 (0.07)	3.96 (0.20)	3.98 (0.14)

[a]The high scores and minimal variation across all groups indicate a significant ceiling effect in the AI's evaluation. Paired comparisons between strategies were conducted using the Wilcoxon signed rank test; *P* values were adjusted using the Benjamini-Hochberg false discovery rate procedure.

[b]LCC: logical coherence and clarity.

[c]UCKI: use and coverage of key information.

[d]PCAR: plausibility and clinical accuracy of reasoning.

Inferential statistical analysis corroborated this observation. A series of Friedman tests found no statistically significant differences among the 3 groups for LCC (*P*=.37), PCAR (*P*=.37), or UCKI (*P*=.07). While a post-hoc pairwise Wilcoxon paired test identified a marginal statistical difference between the random few-shot and zero-shot groups on the information use dimension (*P*=.045), this isolated finding merits cautious interpretation, particularly, as the overall test for this dimension did not reach statistical significance.

Discussion

Principal Findings

This study critically evaluates the reliability of LLM-generated CoT reasoning in ART and shows that noncurated prompting methods are insufficient for clinical use. Both zero-shot and random few-shot strategies frequently produced reasoning errors, and random shallow examples offered no meaningful improvement over providing no examples at all. In contrast, the selective few-shot strategy, which is built on the principles of representative diversity and gold-standard depth, substantially improved coherence, information use, and clinical accuracy. These reliability gaps, as well as the strengths of the selective approach, were identifiable only through expert review; automated AI evaluators failed to detect these differences. Together, these findings outline a practical framework for evaluating ART reasoning quality and a feasible pathway for generating trustworthy synthetic clinical data.

The principle of representative diversity was clearly demonstrated in the PGT and ICSI subgroups. The findings provide empirical support for our initial hypothesis. The PGT category shows significantly higher scores, prompted by the selective few-shot approach, which includes an example of PGT-A treatment. The case study also shows errors in understanding and judgment in doctors' viewing, where zero-shot or random few-shot are more likely to make intrinsic mistakes. Notably, in the ICSI category, although the intergroup differences did not reach statistical significance when compared to the random few-shot group, we observed the same trend as in the PGT category—selective prompting consistently achieved the highest average scores and was significantly higher than zero-shot prompting, which had no difference with the random one. The analyses of both subgroups collectively suggest that a demonstration set covering key procedural subtypes within the domain is essential for enabling the model to evolve from a "specialist" to a "generalist."

Simultaneously, the principle of gold-standard depth was illustrated in the IVF subgroup. In our main results, we show that the quality of examples may influence the quality of generation. In subgroup analysis, we found that there is no significant difference between the zero-shot prompting and the random few-shot prompting on any subgroup, especially in the IVF subgroup, even if the random arm's sample cases indeed included 4 standard IVF and 1 short-protocol IVF. It performed ineffective learning under this situation. In this case, the reason may be attributed to the reasoning quality in the prompt. In the experiment design section, we mentioned that the random cases have a relatively concise CoT. This indicates that the LLM exhibits a strong tendency toward pattern imitation when engaging in in-context learning. A low-quality example tends to elicit correspondingly poor reasoning outputs, even if the model has huge potential in text generation. Therefore, this principle emphasizes that each few-shot example must serve as an expert-level exemplar: logically rigorous, richly detailed, and representative of ART strategy reasoning at the highest standard.

XSL•FO
RenderX

Comparison to Prior Work

Our findings align closely with a well-established principle in the broader AI research community: data quality often outweighs data quantity [36]. Our work provides domain-specific empirical support for the application of this principle in the reproductive medicine context of clinical CoT generation. More importantly, we go beyond simply affirming the importance of data reliability; we offer a concrete characterization of what high-quality examples mean in this setting, through our proposed dual principles of gold-standard depth and representative diversity. Together, these insights contribute a practical methodology for realizing data-centric AI specializing in reproductive medicine.

Another key finding highlights a critical limitation of current LLM-based evaluators in detecting subtle yet clinically meaningful variations in information use, logical rigor, and contextual accuracy. While our human expert assessments revealed substantial differences in reasoning quality across the 3 prompting strategies, the scores assigned by the AI evaluator (GPT-4o) showed no statistically significant differences between them across 3 metrics. This "ceiling effect" serves as a critical warning: in high-stakes medical applications, like ART strategy choosing, where patient safety is on the line, relying solely on automated evaluation for quality assurance is inherently risky. It reaffirms that domain expert oversight is not merely a "gold standard" for evaluation; it is an essential safeguard that cannot be replaced. Our results show that AI-based evaluation cannot be treated as a source of ground truth; all judgments involving factual accuracy, clinical appropriateness, or safety must rely on human experts. From a broader methodological perspective, the results underscore a growing challenge for the field. As the development of medical LLMs increasingly depends on large-scale synthetic data, evaluation may become the primary bottleneck. While models continue to improve in producing fluent clinical narratives, reliably detecting subtle but clinically meaningful reasoning errors remains far more difficult. Without dependable evaluators, synthetic or augmented clinical data cannot safely be incorporated into model training pipelines. Addressing this gap will require medically grounded evaluation frameworks, including domain-specific supervision signals, error-aware reward models, and structured representations of clinical logic. These capabilities are not yet captured by current general-purpose LLM judges, emphasizing the need for future research focused on building evaluators that meet the safety, sensitivity, and domain expertise required for clinical AI applications.

This study provides evidence within a single-center ART dataset, and further multicenter generalization is needed. Although we attempted to determine the reliability of AI-generated CoT in complex clinical cases, our cases are currently limited to reproductive medicine or ART treatment. To enhance generalizability and robustness, future research should include a more diverse set of complex clinical reasoning cases across different medical departments. This study has several limitations. First, all generations were produced using a single model (DeepSeek-R1), which restricts the external validity of the findings. Future studies will evaluate whether the advantages of the selective few-shot strategy generalize across different LLM families. Second, the use of temperature=0.5 introduces controlled stochasticity into the inference process; other decoding settings may produce output variations. To address this, future work will include sensitivity analyses across multiple temperature levels (eg, 0, 0.2, and 0.5) to assess the stability of reasoning patterns. In addition, all human evaluators were recruited from the same medical center, which may introduce institutional bias due to shared training backgrounds and practice standards. The AI-grader feasibility test also has limitations: the grader's sensitivity is partly dependent on its prompt design and model chosen, which may reduce its ability to detect subtle but clinically important differences within the reasoning block. Finally, the evaluation was conducted at the case level, and although the Likert-based rubric captures overall reasoning quality, subjective variability cannot be fully eliminated. Future work may incorporate sentence-level or error-type–specific analysis to support more objective and fine-grained identification of reasoning deficiencies. Given these constraints, this study should be interpreted as a vertical, domain-specific proof-of-concept, rather than a horizontal benchmark applicable across clinical specialties or model families. The selective few-shot strategy was examined within ART because it provides a well-defined and clinically coherent setting for studying structured reasoning, not because its performance should be assumed to generalize elsewhere. Whether the observed improvements reflect a domain-specific phenomenon or a more general pattern cannot be determined from this study. Future work will therefore focus on rigorously evaluating the approach across diverse clinical domains, datasets, and model families to assess its true generalizability. Beyond the constraints and limitations, an important consideration of this study is that the selective few-shot condition differed from the random condition not only in the conceptual selection principles but also in exemplar characteristics, including number, clinical depth, and ART subtype diversity, which creates a mixed signal. Although the numerical difference between 5-shot and 6-shot prompting is small, it may nonetheless introduce bias. More importantly, exemplar depth and subtype diversity were intentionally incorporated to construct a clinically coherent selective prompt, but these factors inherently covary in our current design. As a result, this study cannot attribute the observed improvements to any single component of the selective strategy nor determine whether the effect arises from exemplar depth, diversity, their interaction, or other uncontrolled influences. The findings should therefore be interpreted as exploratory and hypothesis-generating, rather than evidence of a validated mechanism. To address this joint pattern, future work will implement controlled ablation studies that (1) equalize exemplar number across conditions and (2) independently manipulate exemplar depth ("deep vs shallow") and subtype diversity ("diverse vs homogeneous"). Such studies will allow rigorous assessment of the independent and combined contributions of these factors to few-shot reasoning performance in clinical LLMs.

Our dataset contains 200 diverse cases, but for some subtypes, the number of cases may be too small for statistical analysis, particularly the PGT subgroup, which only included 22 samples. Although the post-hoc power calculation and *P* value correction were conducted, it still showed moderate effect sizes on part of

XSL•FO
RenderX

the comparisons. Accordingly, the subgroup findings should be interpreted as exploratory. A potential methodological improvement for future studies is the use of hierarchical partial-pooling or Bayesian shrinkage models, which may borrow strength across subgroups and produce more stable estimates under low-sample conditions. These models were not adopted in this study because our primary objective was descriptive comparison rather than multilevel estimation, but they represent a promising direction for future research. For future directions, given the limited context window of current LLMs, users may face an inherent trade-off when selecting few-shot exemplars, particularly in domains such as reproductive medicine where clinical presentations exhibit substantial subtype diversity. Balancing breadth and depth in exemplar selection becomes a critical challenge under these constraints. Recent work on dynamic prompting methods has sought to improve the performance-efficiency trade-off in resource-limited or accuracy-constrained settings [37], and incorporating such techniques may further enhance the practicality of selective prompting in clinical applications. In addition, future work will explore retrieval-augmented generation frameworks. Integrating authoritative domain sources (eg, American Society for Reproductive Medicine and European Society of Human Reproduction and Embryology guidelines) has the potential to improve factual grounding, reduce hallucination, and enhance explainability in ART-related clinical reasoning. Comparing closed-book reasoning with retrieval augmented generation augmented reasoning may clarify how access to external evidence shapes LLM decision-making and may improve the reliability of LLM-assisted clinical decision support tools.

Conclusions

The primary contribution of this study is 2-fold: an exploratory potential evaluation framework for how to evaluate and provide a methodology for a feasible approach and for how to generate trustworthy clinical ART reasoning steps in a single clinic center. First, we investigate a rigorous, domain-grounded framework for evaluating synthetic clinical reasoning within the ART strategy. Amid the rapid growth of AI in health care, we demonstrate that ensuring clinical validity requires moving beyond automated metrics. Our findings expose the critical limitations of SOTA AI evaluators (eg, GPT-4o) in detecting subtle but clinically vital reasoning flaws. This “ceiling effect” serves as a critical warning and highlights the indispensable role of structured, blind expert review as an essential safeguard in reproductive medicine AI development. Second, building on this evaluation framework, we offer a practical solution to the “explainability data bottleneck” in reproductive medicine. Through systematic comparisons, we show that a selective few-shot prompting strategy, which is based on the “dual principles” of gold-standard depth and representative diversity, substantially improves the quality and reliability of generated ART CoTs. This offers a feasible, cost-effective blueprint for generating trustworthy ART synthetic data at scale, without requiring immense annotated datasets. Finally, this study evaluates the clinical reliability of LLM-generated reasoning in the ART context as a step toward addressing data scarcity in explainable, domain-specialized AI development. However, our findings should not be interpreted as evidence that current LLMs are clinically safe or ready for autonomous use. Our evaluation focuses on reasoning quality, not deployment readiness. Establishing clinical go or no-go thresholds will require task-specific, prospective validation studies assessing safety, consistency, patient outcomes, and workflow integration—factors beyond the scope of this work.

Acknowledgments

The authors used the generative artificial intelligence (AI) tool to examine the potential use of LLMs in assisted reproductive technology. The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to generative AI tools under full human supervision: feasibility assessment and risk evaluation. The generative AI tools used were DeepSeek-R1-671B and ChatGPT4o. Responsibility for the final manuscript lies entirely with the authors. Generative AI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This study was funded by the Science and Technology Department of Sichuan Province Project (2024YFFK0365), the Natural Science Foundation of Sichuan, China (2025NSFSC1985), and the 1·3·5 project for disciplines of excellence, West China Hospital, Sichuan University (ZYYC21004).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: RY, TT, SZ, Di Liu, KL
Data curation: RY, TT, Dou Liu, Y Long
Formal analysis: RY, TT, Dou Liu, Y Long, SZ, Di Liu
Funding acquisition: KL
Investigation: Dou Liu, Y Long

XSL•FO
RenderX

Methodology: Dou Liu, Y Long
Project administration: RY, TT, KL
Visualization: Dou Liu, Y Long, Y Lin, HL
Writing—original draft: Dou Liu, Y Long
Writing—review and editing: All authors

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed description of variables used in the dataset.
[DOCX File , 19 KB - jmir_v28i1e85206_app1.docx]

Multimedia Appendix 2
Chain-of-thought generation prompt.
[DOCX File , 45 KB - jmir_v28i1e85206_app2.docx]

Multimedia Appendix 3
Detailed human evaluation rubric.
[DOCX File , 19 KB - jmir_v28i1e85206_app3.docx]

Multimedia Appendix 4
Statistical results.
[DOCX File , 23 KB - jmir_v28i1e85206_app4.docx]

References

1. Graham ME, Jelin A, Hoon AH, Wilms Floet AM, Levey E, Graham EM. Assisted reproductive technology: short- and long-term outcomes. Dev Med Child Neurol 2023;65(1):38-49 [FREE Full text] [doi: 10.1111/dmcn.15332] [Medline: 35851656]
2. Asplund K. Use of fertilization—ethical issues. Ups J Med Sci 2020;125(2):192-199 [FREE Full text] [doi: 10.1080/03009734.2019.1684405] [Medline: 31686575]
3. Liu D, Long Y, Zuoqiu S, Tang T, Yin R. Evaluating the feasibility and accuracy of large language models for medical history-taking in obstetrics and gynecology. ArXiv. Preprint posted online on March 31, 2025 2025. [doi: 10.48550/arXiv.2504.00061]
4. Huang L, Hu J, Cai Q, Fu G, Bai Z, Liu Y, et al. The performance evaluation of artificial intelligence ERNIE bot in Chinese National Medical Licensing Examination. Postgrad Med J 2024;100(1190):952-953. [doi: 10.1093/postmj/qgae062] [Medline: 38813794]
5. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards accurate differential diagnosis with large language models. Nature 2025;642(8067):451-457 [FREE Full text] [doi: 10.1038/s41586-025-08869-4] [Medline: 40205049]
6. Liu D, Han Y, Wang X, Tan X, Liu D, Qian G, et al. Evaluating the application of ChatGPT in outpatient triage guidance: a comparative study. In: Congress of the International Ergonomics Association. Singapore: Springer Nature Singapore; Aug 25, 2024:233-238.
7. Buckley T, Diao J, Rajpurkar P, Rodman A, Manrai A. Multimodal foundation models exploit text to make medical image predictions. ArXiv Preprint posted online on November 25, 2024 [FREE Full text]
8. Luo L, Ning J, Zhao Y, Wang Z, Ding Z, Chen P, et al. Taiyi: a bilingual fine-tuned large language model for diverse biomedical tasks. J Am Med Inform Assoc 2024;31(9):1865-1874. [doi: 10.1093/jamia/ocae037] [Medline: 38422367]
9. Christophe C, Kanithi P, Munjal P, Raha T, Hayat N, Rajan R, et al. Med42—evaluating fine-tuning strategies for medical LLMs: full-parameter vs. parameter-efficient approaches. ArXiv Preprint posted online on April 23, 2024 [FREE Full text] [doi: 10.18653/v1/2025.emnlp-main.1174]
10. Chen Z, Cano A, Romanou A, Bonnet A, Matoba K, Salvi F, et al. MEDITRON-70B: scaling medical pretraining for large language models. ArXiv Preprint posted online on November 27, 2023 [FREE Full text] [doi: 10.21203/rs.3.rs-4139743/v1]
11. Miao J, Thongprayoon C, Suppadungsuk S, Krisanapan P, Radhakrishnan Y, Cheungpasitporn W. Chain of thought utilization in large language models and application in nephrology. Medicina (Kaunas) 2024;60(1):148 [FREE Full text] [doi: 10.3390/medicina60010148] [Medline: 38256408]
12. Kim S, Joo S, Kim D, Jang J, Ye S, Shin J, et al. The CoT collection: improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. 2023 Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; December 17, 2025; Singapore p. 12685-12708. [doi: 10.18653/v1/2023.emnlp-main.782]

XSL•FO
RenderX

13. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nat Med 2025;31(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]
14. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health 2023;5(4):e179-e181 [FREE Full text] [doi: 10.1016/S2589-7500(23)00048-1] [Medline: 36894409]
15. Rauschecker AM, Rudie JD, Xie L, Wang J, Duong MT, Botzolakis EJ, et al. Artificial intelligence system approaching neuroradiologist-level differential diagnosis accuracy at brain MRI. Radiology 2020;295(3):626-637 [FREE Full text] [doi: 10.1148/radiol.2020190283] [Medline: 32255417]
16. Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. Nat Med 2020;26(6):900-908. [doi: 10.1038/s41591-020-0842-3] [Medline: 32424212]
17. Babbar R, Schölkopf B. Data scarcity, robustness and extreme multi-label classification. Mach Learn 2019;108(8-9):1329-1351. [doi: 10.1007/s10994-019-05791-5]
18. Villalobos P, Ho A, Sevilla J, Besiroglu T, Heim L, Hobbhahn M. Position: Will we run out of data? Limits of LLM scaling based on human-generated data. 2024 Presented at: Forty-First International Conference on Machine Learning; July 21-27, 2024; Vienna, Austria p. 49523-49544. [doi: 10.5555/3692070.3694094]
19. Dahmen J, Cook D. SynSys: a synthetic data generation system for healthcare applications. Sensors (Basel) 2019;19(5):1181. [doi: 10.3390/s19051181] [Medline: 30857130]
20. El EK, Mosquera L, Hoptroff R. Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data. California: O'Reilly Media; 2025.
21. Li Z, Zhu H, Lu Z, Yin M. Synthetic data generation with large language models for text classification: potential and limitations. ArXiv Preprint posted online on October 13, 2023 [FREE Full text] [doi: 10.48550/arXiv.2310.07849]
22. Williams CYK, Subramanian CR, Ali SS, Apolinario M, Askin E, Barish P, et al. Physician- and large language model-generated hospital discharge summaries. JAMA Intern Med 2025;185(7):818-825. [doi: 10.1001/jamainternmed.2025.0821] [Medline: 40323616]
23. Vardhan M, Nathani D, Vardhan S, Aggarwal A, Simini F. Large language models as synthetic electronic health record data generators. 2024 Presented at: IEEE Conference on Artificial Intelligence (CAI); June 25-27, 2024; Singapore p. 804-810. [doi: 10.1109/cai59869.2024.00152]
24. Veselovsky V, Ribeiro M, Arora A, Josifoski M, Anderson A, West R. Generating faithful synthetic data with large language models: a case study in computational social science. ArXiv Preprint posted online on May 24, 2023. [doi: 10.48550/arXiv.2305.15041]
25. Ziems C, Held W, Shaikh O, Chen J, Zhang Z, Yang D. Can large language models transform computational social science? Comput Linguist 2024;50(1):237-291. [doi: 10.1162/coli_a_00502]
26. Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd workers for text-annotation tasks. Proc Natl Acad Sci U S A 2023;120(30):e2305016120. [doi: 10.1073/pnas.2305016120] [Medline: 37463210]
27. Wang J, Liang Y, Meng F, Sun Z, Shi H, Li Z, et al. Is ChatGPT a good NLG evaluator? A preliminary study. 2023 Presented at: Proceedings of the 4th New Frontiers in Summarization Workshop; December 17, 2025; Singapore p. 1-11. [doi: 10.18653/v1/2023.newsum-1.1]
28. Li H, Dong Q, Tang Z, Wang C, Zhang X, Huang H, et al. Synthetic data (almost) from scratch: generalized instruction tuning for language models. ArXiv Preprint posted online on February 20, 2024. [doi: 10.48550/arXiv.2402.13064]
29. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. 2022 Presented at: NIPS'22: 36th International Conference on Neural Information Processing Systems; November 28-December 9, 2022; New Orleans, LA, United States p. 24824-24837 URL: https://dl.acm.org/doi/10.5555/3600270.3602070 [doi: 10.5555/3600270.3602070]
30. Sprague Z, Yin F, Rodriguez J, Jiang D, Wadhwa M, Singhal P, et al. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. ArXiv Preprint posted online on May 7, 2025. [doi: 10.48550/arXiv.2409.12183]
31. Yuan X, Shen C, Yan S, Zhang X, Xie L, Wang W, et al. Instance-adaptive zero-shot chain-of-thought prompting. 2024 Presented at: Advances in Neural Information Processing Systems 37 (NeurIPS 2024); December 10, 2024; Vancouver, British Columbia, Canada. [doi: 10.52202/079017-3986]
32. Shi E, Manda A, Chowdhury L, Arun R, Zhu K, Lam M. Enhancing depression diagnosis with chain-of-thought prompting. ArXiv Preprint posted online on August 27, 2024. [doi: 10.48550/arXiv.2408.14053]
33. Liu R, Wei J, Liu F, Si C, Zhang Y, Rao J, et al. Best practices and lessons learned on synthetic data. ArXiv Preprint posted online on August 10, 2024. [doi: 10.48550/arXiv.2404.07503]
34. Guo D, Yang D, Zhang H, Song J, Wang P, Zhu Q, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature 2025;645(8081):633-638. [doi: 10.1038/s41586-025-09422-z] [Medline: 40962978]
35. Team L, Xu W, Chan H, Li L, Aljunied M, Yuan R, et al. Lingshu: a generalist foundation model for unified multimodal medical understanding and reasoning. ArXiv Preprint posted online on June 13, 2025. [doi: 10.48550/arXiv.2506.07044]
36. Zha D, Bhat ZP, Lai K, Yang F, Jiang Z, Zhong S, et al. Data-centric artificial intelligence: a survey. ACM Comput Surv 2025;57(5):1-42. [doi: 10.1145/3711118]
37. Zhou W, Jiang YE, Cotterell R, Sachan M. Efficient prompting via dynamic in-context learning. ArXiv Preprint posted online on May 18, 2023. [doi: 10.48550/arXiv.2305.11170]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
ART: assisted reproductive technology
CoT: chain-of-thought
EHR: electronic health record
FDR: false discovery rate
HIPAA: Health Insurance Portability and Accountability Act
ICSI: intracytoplasmic sperm injection
IVF: in vitro fertilization
LCC: logical coherence and clarity
LLM: large language model
PCAR: plausibility and clinical accuracy of reasoning
PGT: preimplantation genetic testing
PGT-A: preimplantation genetic testing for aneuploidies
PGT-M: preimplantation genetic testing for monogenic disorder
SOTA: state-of-the-art
TESA: testicular sperm aspiration
UCKI: use and coverage of key information

Edited by J Sarvestan; submitted 02.Oct.2025; peer-reviewed by X Zhong, KH Lin; comments to author 03.Nov.2025; revised version received 14.Dec.2025; accepted 15.Dec.2025; published 08.Jan.2026.

Please cite as:
Liu D, Long Y, Zuoqiu S, Liu D, Li K, Lin Y, Liu H, Yin R, Tang T
Reliability of Large Language Model Generated Clinical Reasoning in Assisted Reproductive Technology: Blinded Comparative Evaluation Study
J Med Internet Res 2026;28:e85206
URL: https://www.jmir.org/2026/1/e85206
doi:10.2196/85206
PMID:

©Dou Liu, Ying Long, Sophia Zuoqiu, Di Liu, Kang Li, Yiting Lin, Hanyi Liu, Rong Yin, Tian Tang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 08.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessing the Evolution and Influence of Medical Open Databases on Biomedical Research and Health Care Innovation: A 25-Year Perspective With a Focus on Privacy and Privacy-Enhancing Technologies

Albert Yang[1,2,3], MD, PhD; Mei-Lien Pan[4], PhD; Henry Horng-Shing Lu[5,6,7,8,9], PhD; Chung-Yueh Lien[10], PhD; Da-Wei Wang[11†], PhD; Chih-Hsiung Chen[12], PhD; Der-Cherng Tarng[2,13,14,15], MD, PhD; Dau-Ming Niu[2,14,16], MD, PhD; Shih-Hwa Chiou[3,17], MD, PhD; Chun-Ying Wu[18,19], MD, PhD; Ying - Chou Sun[20], MD; Shih-Ann Chen[2,21], MD; Shuu-Jiun Wang[2,22,23], MD; Wayne Huey-Herng Sheu[24], MD, PhD; Chi-Hung Lin[25], MD, PhD

[1]Digital Medicine and Smart Healthcare Research Center, National Yang Ming Chiao Tung University, No. 155 Sec. 2 Linong St., Beitou Dist, Taipei City, Taiwan

[10]Department of Information Management, National Taipei University of Nursing and Health Sciences, Taipei, Taiwan

[11]Institute of Information Science, Academia Sinica, Taipei, Taiwan

[12]Institute of Technology Law, National Yang Ming Chiao Tung University, Taipei, Taiwan

[13]Department of Medicine, Taipei Veterans General Hospital, Taipei, Taiwan

[14]Institute of Clinical Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan

[15]Department and Institute of Physiology, National Yang Ming Chiao Tung University, Taipei, Taiwan

[16]Department of Pediatrics, Taipei Veterans General Hospital, Taipei, Taiwan

[17]Institute of Pharmacology, College of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan

[18]Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, Taipei, Taiwan

[19]Health Innovation Center, National Yang Ming Chiao Tung University, Taipei, Taiwan

[2]School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan

[20]Department of Radiology, Taipei Veterans General Hospital, Taipei, Taiwan

[21]Department of Cardiovascular Center and Medical Research, Taichung Veterans General Hospital, Taichung, Taiwan

[22]Department of Neurology, Neurological Institute, Taipei Veterans General Hospital, Taipei, Taiwan

[23]Brain Research Center, National Yang Ming Chiao Tung University, Taipei, Taiwan

[24]Institute of Molecular and Genomic Medicine, National Health Research Institutes, Taipei, Taiwan

[25]Department of Biological Science & Technology, National Chiao Tung University, Taipei, Taiwan

[3]Department of Medical Research, Taipei Veterans General Hospital, Taipei, Taiwan

[4]Institute of Hospital and Health Care Administration, National Yang Ming Chiao Tung University, Taipei, Taiwan

[5]Institute of Statistics, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

[6]Biomedical Artificial Intelligence Academy, Kaohsiung Medical University, Kaohsiung, Taiwan

[7]Department of Artificial Intelligence in Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan

[8]Department of Medical Research, Kaohsiung Medical University Hospital, Kaohsiung, Taiwan

[9]Department of Statistics and Data Science, Cornell University, Ithaca, NY, United States

[†]deceased

Corresponding Author:
Albert Yang, MD, PhD
Digital Medicine and Smart Healthcare Research Center, National Yang Ming Chiao Tung University, No. 155 Sec. 2 Linong St., Beitou Dist, Taipei City, Taiwan

Abstract

The integration of medical open databases with artificial intelligence (AI) technologies marks a transformative era in biomedical research and health care innovation. Over the past 25 years, initiatives like PhysioNet have revolutionized data access, fostering unprecedented levels of collaboration and accelerating medical discoveries. This rise of medical open databases presents challenges, particularly in harmonizing research enablement with patient confidentiality. In response, privacy laws such as the Health Insurance Portability and Accountability Act have been established, and privacy-enhancing technologies have been adopted to maintain this delicate balance. Privacy-enhancing technologies, including differential privacy, secure multiparty computation, and notably,

XSL•FO
RenderX

federated learning (FL), have become instrumental in safeguarding personal health information. FL, in particular, represents a significant advancement by enabling the development and training of AI models on decentralized data. In Taiwan, significant strides have been made in aligning with these global data-sharing and privacy standards. We have actively promoted the sharing of medical data through the development of dynamic consent systems. These systems enable individuals to control and adjust their data-sharing preferences, ensuring transparency and continuity of consent in the ever-evolving landscape of digital health. Despite the challenges associated with privacy protections, the benefits, including improved diagnostics and treatment, are substantial. The availability of open databases has notably accelerated AI research, leading to significant advancements in medical diagnostics and treatments. As the landscape of health care research continues to evolve with open science and FL, the role of medical open databases remains crucial in shaping the future of medicine, promising enhanced patient outcomes and fostering a global research community committed to ethical integrity and privacy.

(J Med Internet Res 2026;28:e58954) doi:10.2196/58954

KEYWORDS

medical open databases; artificial intelligence; privacy-enhancing technologies; federated learning; dynamic consent systems

Introduction

The emergence of medical open databases, coupled with advances in artificial intelligence (AI), heralds a significant change in biomedical research and health care innovation, facilitating an era of enhanced accessibility and data sharing [1-3]. This movement toward open data science, augmented by AI technologies, enables researchers worldwide to access a wealth of data, including physiological signals [4,5], genomic [6], and health care information [7], and, most prominently, large-scale medical imaging archives. While this review covers the broad spectrum of medical data, the impact of open imaging databases has been particularly transformative for the application of AI. This movement toward open data science fosters collaboration and speeds up the pace of medical discoveries.

AI's role in analyzing vast datasets has been instrumental in uncovering patterns and insights that would be impossible for humans to detect unaided, leading to breakthroughs in understanding diseases and patient care. Initiatives like annual challenges and shared toolboxes have spurred the development of novel algorithms and techniques, leveraging AI to address complex biomedical challenges and advance medical diagnostics and treatments. This synergy between open medical databases and AI is transforming the landscape of health care, promising a future of more accurate, efficient, and personalized medicine.

Simultaneously, this rise in open data repositories brings to the forefront crucial privacy concerns [6,8]. The necessity to balance the imperative of research enablement with the protection of patient confidentiality has never been more pronounced. In this context, laws such as the Health Insurance Portability and Accountability Act (HIPAA) play a pivotal role in shaping the landscape of data deidentification and anonymization processes, ensuring that shared data comply with strict privacy standards [9]. Moreover, the introduction of privacy-enhancing technologies (PETs), such as differential privacy [10], synthetic data [11], homomorphic encryption [12], secure multiparty computation [13], and federated learning [14], represents a proactive approach to safeguarding personal health information. These technologies provide the means to conduct meaningful research while upholding the principles of data privacy and security.

In a country like Taiwan, strides in medical data sharing suggest the global shift toward interconnected health systems, highlighting both advancements and ongoing challenges in securing patient data. The implementation of dynamic consent frameworks reflects a growing recognition of the need for more flexible approaches to data privacy, particularly in an era of personalized medicine and digital health records [15]. As the landscape of medical research evolves with these developments, the interplay of data sharing, privacy, and technology continues to reshape the boundaries of what is possible in health care innovation, marking a critical junction in the journey toward more open, collaborative, and ethically responsible research environments.

This review aims to provide a 25-year perspective on the evolution of medical open databases, tracing their impact on biomedical research and health care innovation, and to examine how emerging PETs and data-governance frameworks, including dynamic consent systems, shape the ethical, technical, and collaborative landscape of digital medicine worldwide.

The Rise of Medical Open Database

The rise of medical open databases represents a transformative shift in the landscape of biomedical research and health care innovation. Among the pioneers in this movement is PhysioNet. Established in 1999, PhysioNet is a pioneering open database that provides free access to a wide range of physiologic signals and related open-source software for research in medicine, physiology, and biomedical engineering [4,16]. It was initiated by a collaborative project involving researchers from Boston's Beth Israel Deaconess Medical Center/Harvard Medical School, Boston University, McGill University, and Massachusetts Institute of Technology [4,17]. The database contains a diverse collection of physiological datasets, including those related to cardiovascular and other complex biomedical signals [4,18]. PhysioNet has had a significant impact on the development of medical open databases, serving as a model for the establishment of similar resources. It has also played a key role in promoting the dissemination and exchange of medical resources.

A significant contribution of PhysioNet to the scientific community is its annual PhysioNet Challenge, which has markedly influenced the field by promoting innovation and collaboration among researchers and clinicians. These challenges

XSL•FO
RenderX

stimulate the creation of novel algorithms and methods aimed at solving complex biomedical problems, thus expanding the limits of what can be achieved in medical data analysis and application. For instance, the challenges have catalyzed the development of innovative algorithms capable of detecting obstructive sleep apnea from electrocardiograms [19], illustrating the practical impact of these competitions on advancing medical diagnostics and treatment strategies [20].

Since the success of PhysioNet, numerous other medical open databases have emerged globally, fostering a more cooperative and transparent research atmosphere (Figure 1). The impact of these large-scale databases on biomedical discovery is profound. One notable example is the UK Biobank [21], launched in 2006, which provides a vast repository of genetic and health information from half a million UK participants. This database has become an essential tool for unraveling the complex interplay between genetics, lifestyle, and disease, thereby enhancing our understanding of the factors influencing human health [22-25]. By leveraging large neuroimaging cohorts such as Alzheimer Disease Neuroimaging Initiative and the UK Biobank, researchers have developed AI-based models that generate an Alzheimer disease risk score from structural magnetic resonance imaging (MRI), enabling the identification of prediagnostic populations suitable for early intervention and preventive trials [26]. In cardiovascular research, analysis of the UK Biobank's genetic and imaging data has enabled the development of NeuralCVD, a neural network–based risk model that integrates polygenic and clinical predictors to estimate the 10-year risk of major adverse cardiac events, improving risk discrimination and reclassification beyond established clinical scores and Cox models, and highlighting the added predictive value of genetic predisposition in early prevention [27]. Similarly, the Cancer Imaging Archive, inaugurated in 2011 in the United States, offers a dedicated platform for the cancer research community, enabling access to a comprehensive array of imaging datasets [28].

Figure 1. Historical development of major medical open databases and data-sharing platforms over the past 25 years.

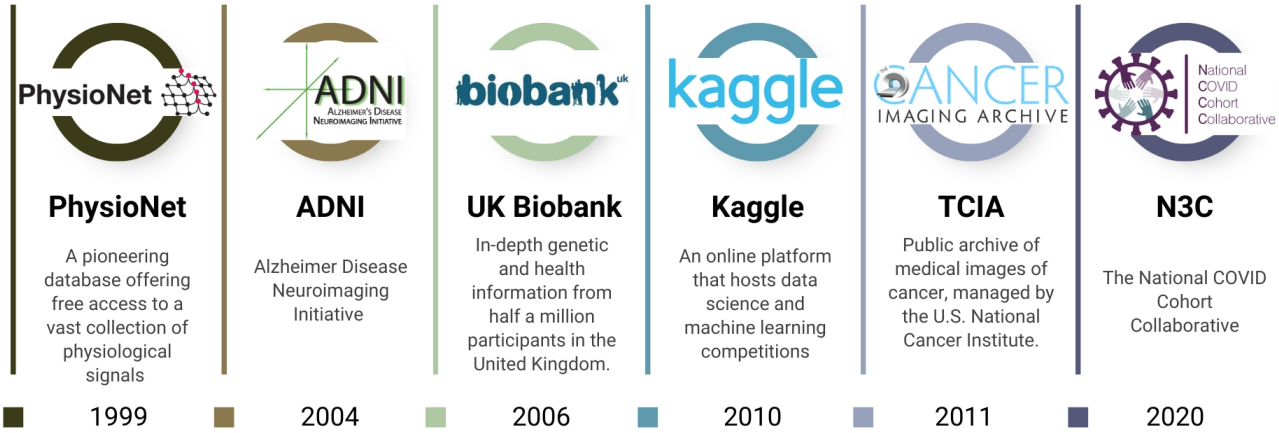

Beyond these examples, the ecosystem of medical open databases has diversified into numerous specialized domains. For instance, OpenNeuro provides a vast repository for neuroimaging data, particularly functional magnetic resonance imaging, electroencephalogram, and magnetoencephalography, supporting reproducible brain research [29]. The Neuroimaging Informatics Tools and Resources Clearinghouse offers a rich collection of imaging and data-processing tools [30]. Similarly, the National Database for Autism Research [31] and the Federal Interagency Traumatic Brain Injury Research informatics system [32] provide deeply phenotyped datasets crucial for research in their respective fields. These platforms underscore the field's shift toward creating specialized, high-quality resources to tackle specific biomedical questions. They display how shared resources can drive forward innovation and improve patient care worldwide, illustrating the critical role of collaborative environments in the advancement of health care research and application.

Additionally, Kaggle, an online platform for data science and machine learning (ML) competitions, has emerged as a pivotal player in the field of ML analysis and data sharing [33]. Launched in 2010, Kaggle facilitates collaboration and competition among data scientists and researchers by hosting challenges in various domains, including health care. These competitions often involve complex medical datasets, encouraging participants to develop innovative solutions and algorithms for disease prediction, medical imaging analysis [34], and other health-related issues. Kaggle has not only democratized access to large medical datasets but has also fostered a global community where knowledge and techniques are openly shared. This environment has led to significant breakthroughs and advancements in medical research and analytics, further suggesting the importance of open data and collaborative problem-solving in improving health care outcomes and accelerating medical innovation.

Balancing Privacy and the Need for Medical Open Databases

The success of open medical databases such as PhysioNet poses the challenge of balancing patient confidentiality with research enablement on open platforms [35,36]. Research datasets in health care often contain protected health information (PHI), and the process of removing this information, a process known as deidentification or anonymization, can be challenging and prone to errors [37]. Despite the use of these datasets, the need for deidentification introduces a significant barrier to data sharing due to the effort and cost involved.

XSL•FO
RenderX

The HIPAA, established in the United States in 1996, plays a vital role in safeguarding patients' medical information. In response to the HIPAA mandate, U.S. Department of Health and Human Services published a final regulation in the form of the privacy rule in December 2000, which became effective on April 14, 2001. Central to this rule is the designation of 18 specific categories of PHI that, if disclosed, could be used to identify an individual (Textbox 1). These categories encompass a broad spectrum of personal data, including, but not limited to, names, geographic details smaller than a state, various identifiers like social security numbers, medical record numbers, and contact information, as well as certain biometric and photographic images [38].

Textbox 1. Eighteen categories of protected health information.

- Names
- All geographic subdivisions smaller than a State
- All elements of dates (except year) for dates directly related to an individual
- Telephone numbers
- Fax numbers
- Electronic mail addresses
- Social security numbers
- Medical record numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers
- Device identifiers and serial numbers
- Web URLs
- IP address numbers
- Biometric identifiers, including finger and voice prints
- Full face photographic images and any comparable images
- Any other unique identifying number, characteristic, or code

Additionally, HIPAA mandates that covered entities must ensure they do not possess knowledge that the remaining information could be used, whether alone or in conjunction with other data, to identify the subject. By strictly adhering to these guidelines, entities can share deidentified health information for broader uses, such as public health and research, without infringing on individual privacy rights, thus striking a balance between privacy protection and the beneficial use of health data [39]. Despite these efforts, the tension between the promise of big data and patient privacy in health care research remains a challenge [40].

PhysioNet is a pioneer in medical public databases, ensuring that the datasets it provides do not compromise individual privacy. This involves ensuring that any data shared does not contain PHI or has been sufficiently anonymized to prevent the identification of individuals. The challenges posed by the HIPAA privacy rule are not insignificant; they include the need for informed consent from data subjects and potential limitations on access to health information that can hinder clinical research [41,42]. Furthermore, the rule's interaction with other regulations, like the common rule, adds complexity to privacy concerns in research, leading to inconsistencies and additional burdens for researchers.

Despite these challenges, the privacy rule does allow for certain disclosures without patient authorization, particularly for public health purposes. This is intended to facilitate the use of medical data in important public health endeavors without undermining individual privacy protections [43]. The balance sought by the HIPAA privacy rule between protecting privacy and facilitating research is a critical aspect of its implementation, particularly in the context of medical open databases. By navigating these regulations successfully, repositories can contribute to the advancement of medical research while ensuring compliance with privacy standards [44].

The emergence of medical databases, such as the PhysioNet, UK Biobank, and the Cancer Imaging Archive, has significantly advanced collaborative research in health care [45,46]. These databases have the potential to transform cancer research and improve patient outcomes [45]. However, the collection, linking, and use of data in biomedical research raise ethical concerns, particularly regarding privacy and security [36,47,48]. Despite these concerns, the benefits of open data in health care, including improved diagnostics and treatment, are substantial [48]. The push for data sharing in cancer trials by pharmaceutical companies further underscores the importance of open medical databases in driving innovation and improving patient care [49].

XSL•FO
RenderX

Privacy Enhancing Technology

Overview

A range of studies have been conducted to explore the increasing frequency and impact of health care data breaches, highlighting the rising number of incidents and their detrimental effects on patient privacy and health care providers [50-53]. These breaches are often caused by a combination of technical, organizational, and human factors [50-52]. Human vulnerabilities, such as lack of awareness and training, play a significant role in these breaches [51]. The use of the Swiss Cheese Model can help assess vulnerabilities and risks [50]. Cloud computing breaches are a particular concern, highlighting the need for digital forensic readiness [54]. Hacking and unauthorized internal disclosures are the most prevalent forms of attack [53]. Further studies may examine specific cases and the implications for digital forensic readiness, emphasizing the importance of adhering to regulations.

Below, we reviewed several PETs and their applications in enhancing data privacy and security in health care settings (Table 1). PETs, such as encryption, anonymization techniques, and secure multiparty computation, offer powerful mechanisms to protect sensitive health data. Implementing these technologies, alongside robust privacy policies and employee training, can significantly reduce the likelihood of data breaches and bolster the trust between patients and health care providers.

Table . Summary of privacy-enhancing technologies.

Technologies	Core principle	Advantages	Challenges and trade-offs
Differential privacy	Adds calibrated statistical noise to query results to make it impossible to determine if an individual's data were included.	Provides strong and mathematically provable privacy guarantees.	Inherent trade-off between privacy and data use; high privacy can reduce analytical accuracy.
Synthetic data	Creates an artificial dataset that mimics the statistical properties of the original data without containing real patient information.	High use for model training; no real patient data are shared, eliminating reidentification risk.	Can be difficult to generate high-fidelity data that captures all complex correlations; potential for model bias.
Homomorphic encryption	Allows computations to be performed directly on encrypted data without decrypting it first.	Offers extremely strong security, as the raw data are never exposed.	High computational overhead; currently too slow for many complex ML[a] tasks.
Secure multiparty computation	Enables multiple parties to jointly compute a function over their inputs while keeping those inputs private.	Allows for collaborative analysis without a central data repository; no single party sees another's data.	High communication overhead between parties; can be complex to set up and scale.
Federated learning	Trains a central AI[b] model across decentralized devices or servers holding local data samples, without exchanging the data itself.	Keeps raw data local, enhancing privacy and data sovereignty.	Vulnerable to model poisoning/inversion attacks; performance can degrade with heterogeneous data.

[a]ML: machine learning.

[b]AI: artificial intelligence.

Differential Privacy

Differential privacy, a method for protecting individual privacy in data analysis, has been increasingly applied in the health care sector. It involves adding noise to the data to prevent reidentification of individuals. This approach has been used in various areas of health research, including genomics, neuroimaging, and health surveillance [55]. However, there are challenges in its practical application, such as the theoretical nature of the privacy parameter epsilon [56]. To address these challenges, researchers have proposed differentially private data release strategies and noise mechanisms, such as the Laplace and exponential mechanisms [57]. However, a key challenge is the inherent trade-off between privacy and data use; increasing the amount of statistical noise to protect privacy can reduce the accuracy of analytical outcomes.

The application of differential privacy in medical questionnaires has also been explored, with the randomized response mechanism showing promise in improving privacy while retaining data use [58]. Furthermore, the use of differential privacy in geospatial analyses of standardized health care data has been demonstrated, with the development of geodatabase functions for privacy-aware analysis [59]. Finally, the combination of differential privacy and decision tree approach has been proposed for data publishing, and the differentially private mini-batch gradient descent algorithm for model publishing of medical data [60].

Synthetic Data

Synthetic data, generated through simulators, is increasingly used in health care to address the challenges of data availability and privacy [61]. PETs, such as differential privacy, are combined with synthetic data generators to create private synthetic data, preserving statistical properties while ensuring privacy [62]. These technologies have been applied in various use cases, including clinical risk prediction [63] and medical research [64]. However, the evaluation of synthetic data's privacy and use metrics remains a challenge, with a lack of consensus on standard approaches [65]. Despite these challenges, the potential of synthetic data in preserving data use

XSL•FO
RenderX

and patient privacy in electronic health care data is being explored [66].

Homomorphic Encryption

Homomorphic encryption, a powerful tool for preserving privacy in medical data, allows for computations to be performed on encrypted data without the need for decryption. It has been successfully applied in various medical data scenarios, including ML models for classification and training, secure genomic algorithms, and predictive analysis tasks [67-69]. For example, it has been used to securely manage personal health metrics data, process medical images [70,71], and enable secure medical computation [72]. The use of homomorphic encryption in these applications ensures that sensitive medical data remains private and secure. Despite its power, the primary limitation of homomorphic encryption is its significant computational overhead, which can make it slow and resource-intensive for complex computations on large datasets.

Secure Multiparty Computation

Secure multiparty computation is a cryptographic technique that enables data analytics without sharing the underlying data, making it a valuable tool for preserving privacy in medical data analysis [73]. It has been applied in various health care scenarios, including collaborative systems [74], statistical analysis of health data [75], and electronic medical record (EMR) data [75]. Secure multiparty computation has also been used in health care internet of things systems to handle privacy issues [76], prevent data disclosure in sensor networks [77], and enable the reuse of distributed electronic health data [75]. Furthermore, it has been applied to enable privacy-preserving query processing on EMRs [78]. Notably, secure multiparty computation has enabled research on highly sensitive data (such as HIV, rare diseases, and population genomics) that would otherwise be inaccessible due to privacy concerns.

Federated Learning

Federated learning (FL), a decentralized ML approach, is increasingly being applied in the medical field due to the sensitive and fragmented nature of health care data [14,79]. It allows for the collaborative development of ML models without sharing raw data, thus preserving privacy [80,81]. This approach has been used in various medical domains, including oncology and radiology, for tasks such as image analysis and disease prediction [81,82]. However, there are challenges to be addressed, such as data homogeneity and transparency [81]. Furthermore, FL can be vulnerable to security risks, such as model inversion attacks that attempt to reconstruct training data from the shared model updates, and require careful design to ensure robustness. Despite these challenges, FL shows promise in improving the efficiency and privacy of medical data processing [83-85].

To address the challenge of data heterogeneity in FL, we have proposed the Dynamically Synthetic Images for Federated Learning method, significantly improving the conventional FL framework by integrating local information from local multiple institutions with heterogeneous data types [86]. The core principle of its implementation involves a dynamic process where, at the start of each training round, a client's local data are evaluated by the current global model to identify misclassified images. Using a synthetic minority oversampling technique, the system generates new, synthetic images based on these misclassified cases, which are then added to the local training set to compel the model to focus on features it previously failed to learn. In terms of effectiveness, experimental results demonstrated that Dynamically Synthetic Images for Federated Learning-based models achieve higher accuracy than conventional FL approaches and that their performance can be comparable to that of traditional centralized learning, proving especially beneficial for institutions with smaller or more heterogeneous datasets [86].

Taiwan Medical AI and Data Portal and Dynamic Consents System

Taiwan has made significant strides in medical data sharing, particularly in the areas of privacy protection and electronic health records exchange. The country's comprehensive embedded integrated circuit-based health insurance card system, implemented by the Bureau of National Health Insurance, Taiwan, allows for the secure sharing of health information [87]. The use of blockchain technology has been proposed as a means to further enhance the security and privacy of medical data sharing [88,89]. The Taiwan Electronic Medical Record Template and the National Electronic Medical Record Exchange System have been developed to facilitate the exchange of EMRs [90,91]. However, concerns about unauthorized access and secondary use of EMRs persist, particularly among highly educated individuals [92]. The country has also established guidelines for the security and privacy protection of health information, drawing on international best practices [93].

In the past 4 years, funded by the National Science and Technology Council of Taiwan, we have assembled teams from National Yang Ming Chiao Tung University, Taipei Veterans General Hospital, Academia Sinica, and National Taipei University of Nursing and Health Sciences to form a data repository task force known as the Smart Medical AI and Repository Taskforce Center. We launched a medical AI and data-sharing platform aimed at advancing the field of medical AI research in October 2023 [94]. This platform not only provides the public and researchers with access to a multitude of shared datasets but also ensures a meticulous evaluation process (Figure 2). Researchers can apply for access to the data by providing an abstract of their research proposal. Dataset managers assess applications based on their intended use, specific needs, and detailed research plans.

XSL•FO
RenderX

Figure 2. Taiwan medical artificial intelligence and data portal. AI: artificial intelligence.

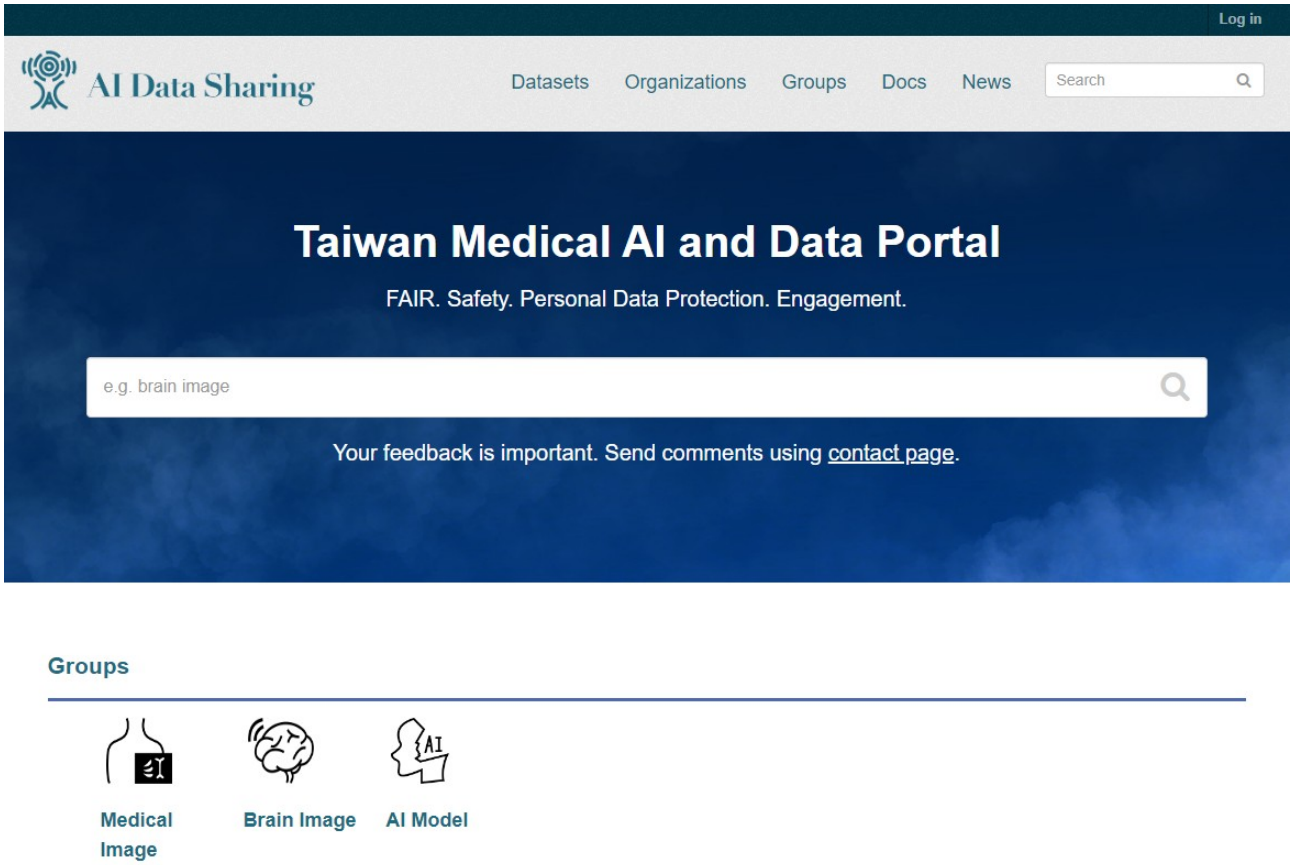

Currently, seventeen datasets have been released on the platform, covering neuropsychiatric disorders, brain tumors, ophthalmic diseases, musculoskeletal disorders, and cardiopulmonary diseases. All datasets have undergone deidentification and delinking processes and include annotated information to facilitate AI training and validation. Specifically, our data-sharing platform includes: MRI images of vestibular schwannoma; computed tomography (CT) images of intracerebral hemorrhage; brain Fluorodeoxyglucose-Positron Emission Tomography/Magnetic Resonance Imaging images for dementia diagnosis; primary brain tumor MRI datasets, including meningioma, glioma, and pituitary adenoma; MRI data of brain metastases, which represent the largest collection nationwide; hand and foot X-rays of rheumatoid arthritis; X-rays of compression fractures; spinal X-rays of ankylosing spondylitis; chest CT images and clinical data of atrial fibrillation patients; chest X-rays for lung cancer screening; annotated preoperative liver CT images; neck lymph node CT images with postoperative pathology results; the Taiwan Aging and Mental Illness Cohort brain imaging database; the dementia molecular imaging database; fundus image datasets for glaucoma; and fundus image datasets of polypoidal choroidal vasculopathy.

The data sharing platform is built on a comprehensive architecture designed to support AI research by integrating 3 core systems: a CKAN-based sharing platform for dataset management, a data application system, and a dynamic authorization consent platform for patient privacy. Specific features include a robust user authentication and authorization mechanism, allowing dataset managers to grant access to specific users or collaborators. The platform ensures data integrity and ethical compliance through a multistep deidentification process for all medical images and by linking to the dynamic consent system (for sensitive clinical data), which allows patients to manage their data sharing preferences in real-time. To use the database, researchers first search for datasets on the platform, then apply for access through a formal registration and review process. Once approved and authorized by the dataset manager, users can obtain a login key to programmatically access the data through standardized protocols, such as DICOMweb, ensuring a secure, convenient, and interoperable environment for third-party AI applications.

This effort aims to advance research across 7 crucial clinical areas that greatly benefit from AI technology, including heart disease, neurological disorders, mental illness, diabetes, cancer, genetic predispositions to complex diseases, and medical imaging. Moreover, the platform underpins collaboration between distinct teams specializing in AI methodology, science and law, and data governance, jointly fostering a robust data governance framework that emphasizes FL, cloud-based AI solutions, and trusted AI practices. Importantly, the system is designed to streamline the research process while maintaining a focus on ethical standards and participant privacy. In line with this, the platform incorporates a dynamic informed consent mechanism, especially for datasets that are anonymized but cannot be completely separated from their sources. This approach ensures that participants' privacy is safeguarded while also enabling their informed and ongoing consent, reflecting

XSL•FO
RenderX

our commitment to ethical research practices and the dynamic nature of consent in medical studies.

Dynamic informed consent, a concept that has been explored in various contexts, is a personalized, digital communication interface that allows participants to manage their consent preferences [95]. It has been proposed as a solution to improve patient confidence and trust in the use of electronic patient records in medical research [96]. In the context of privacy-aware pervasive health and well-being, dynamic consent enables granular data consent and management [97]. It has also been suggested as a potential solution to challenges in modern biomedical research, including participant recruitment, informed consent, and consent management [98]. The use of blockchain technology has been proposed to enhance the privacy-preserving aspects of dynamic consent in genomic data sharing [99]. The concept of dynamic informed consent has been further explored in the context of personalized medicine, emphasizing the need for a more dynamic and enriched consent model [100].

To enhance the privacy of participants contributing to the data in our platform and increase participants' engagement, we have developed a dynamic informed consent system, named the Health Data Authorization Service Platform (Figure 3). This collaborative effort involves our data governance, humanities, science and law, and clinical teams, aiming to facilitate scalable, dynamic consent operations suitable for complex data environments. The platform supports flexible data governance, allowing data owners to dynamically express their consent preferences, thereby making dynamic consent practical and sustainable.

Figure 3. Health data authorization service platform. This diagram depicts the architecture of a medical data-sharing platform that integrates dynamic consent for secure and transparent data sharing. The system centers on 4 key roles: the resource owner (RO), who owns the data; the resource server (RS), where medical data are stored and managed; the requesting party (RP), typically researchers seeking data access; and the authorization server (AS), the core component facilitating connections among the roles. This framework ensures data access is based on explicit consent from the data owner, allowing real-time adjustments to consent settings for different data types and uses, such as academic research. The platform's primary tasks are to establish individual preferences, maintain consent history, and ensure trust and transparency between data owners and users, thereby advancing responsible data science development.

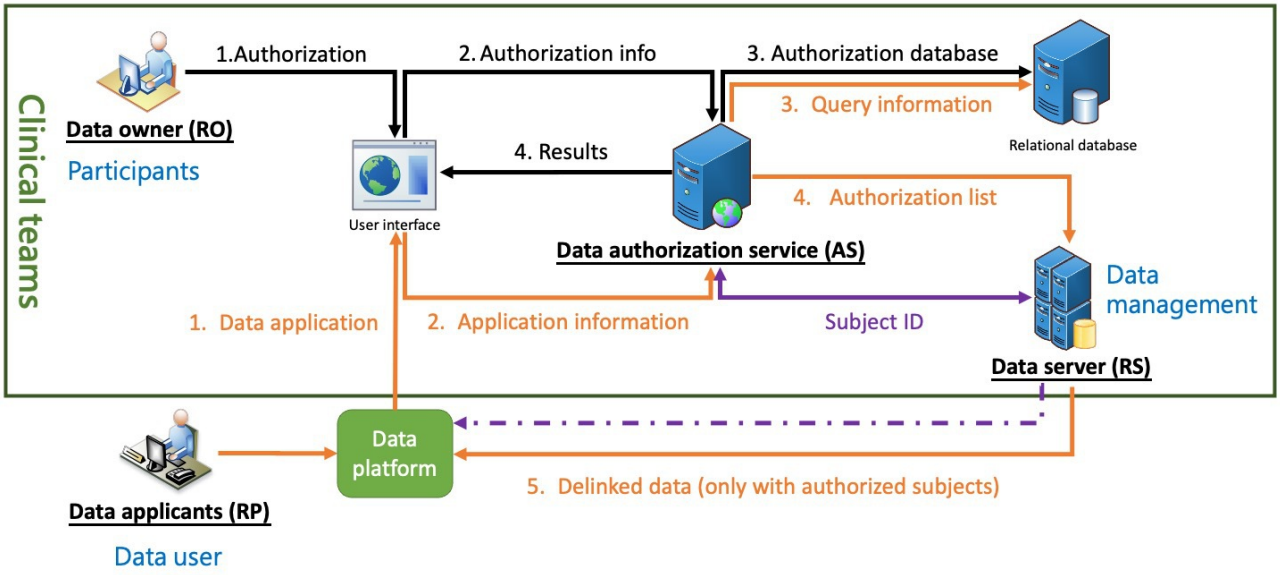

The platform's architecture encompasses 4 key roles: the resource owner, usually the participants, who owns the data; the resource server, which stores and manages medical data; the requesting party, typically researchers seeking data access; and the authorization server, the backbone of our dynamic consent system, connecting the 3 roles. This system ensures that data are accessed only with the owner's express consent, respecting their preferences and enhancing data use transparency. Resource owners can modify their consent settings at any time, reflecting changes in their willingness to share different data types, like EMRs or medical imaging, for specific purposes such as academic research. This flexibility ensures that data use aligns with the owners' current preferences. The platform seamlessly integrates with our shared data framework, maintaining each citizen's consent history and enabling swift updates to consent forms as needed. By streamlining the consent process and ensuring data are shared according to owner permissions, our platform respects individual preferences while promoting responsible data science development. It exemplifies a forward-thinking approach to data governance, enabling real-time adjustments in consent and fostering a culture of trust and transparency between data owners and users.

These initiatives in Taiwan can be understood within the global context of evolving data privacy regulations. Unlike the "one-time, broad consent" model often used in US-based research under the Common Rule, Taiwan's move toward dynamic consent aligns more closely with the principles of the European Union's General Data Protection Regulation [101]. The General Data Protection Regulation mandates that consent must be specific, informed, and easily revocable. The dynamic consent system builds on this by providing a technological interface for participants to manage their preferences granularly and continuously, representing a best-practice approach to balancing research needs with individual autonomy and privacy rights.

XSL•FO
RenderX

Acceleration of Medical AI Research Through Open Databases

The advent of medical open databases has significantly accelerated the field of medical AI research, fostering an environment of innovation and rapid development [102]. By providing researchers with access to vast amounts of health-related data, these databases have become a cornerstone for advancements in predictive analytics, diagnostic algorithms, and personalized medicine.

One of the most notable contributions of open medical databases to AI research is the democratization of data [103]. Historically, the scarcity and inaccessibility of medical data posed substantial hurdles to AI development. However, platforms like PhysioNet, established in 1999, have bridged this gap by offering a plethora of datasets ranging from physiological signals to clinical outcomes [104]. However, challenges remain, including the need for large datasets and the lack of external validation in perioperative medicine [105]. The use of open science approaches, including data liberation and crowdsourcing, can help address these challenges [106]. The integration of networked medical devices and clinical repositories based on open standards can further enhance AI research in high-acuity medical environments [107]. This enhanced availability allows researchers from diverse backgrounds and institutions to engage in health care innovation, leveling the playing field and stimulating a surge in AI-based solutions.

The availability of open databases has catalyzed the application of diverse AI methodologies to complex medical problems. Deep learning, particularly convolutional neural networks, has achieved state-of-the-art performance by leveraging large-scale imaging datasets; for example, researchers have trained convolutional neural networks on millions of images from The Cancer Imaging Archive to develop algorithms capable of detecting and classifying tumors in radiological scans with accuracy comparable to human experts [108]. For structured data such as the genetic and clinical information in the UK Biobank, traditional ML models like random forests and gradient boosting have been widely used, excelling at identifying complex patterns to predict disease risk, including the calculation of polygenic risk scores for coronary artery disease based on thousands of genetic variants [109]. In addition, natural language processing techniques have been applied to large repositories of unstructured clinical notes, such as those in the Medical Information Mart for Intensive Care version IV (MIMIC-IV) database (part of PhysioNet), to extract critical information on symptoms, treatments, and outcomes, thereby enabling large-scale retrospective studies that were previously infeasible [110].

Open medical databases encompass a wide variety of data types, including EMRs, imaging, genomic sequences, and more. This diversity enables AI researchers to explore multifaceted health care questions, from predicting disease trajectories to optimizing treatment plans. Moreover, the rich, varied datasets facilitate the training of more robust and generalizable AI models, capable of addressing complex medical scenarios across different populations and settings. The shared nature of open databases fosters collaboration across the global research community [111]. Through platforms that offer shared data, researchers can combine their expertise to tackle larger and more complex problems than they could individually. This collaborative approach has led to significant breakthroughs in AI, such as algorithms that can detect diseases from images with accuracy rivaling that of trained professionals [112,113].

Open databases also streamline the validation and implementation phases of AI development [114]. Access to diverse datasets enables researchers to rigorously test their algorithms under various conditions and patient demographics, ensuring their reliability and effectiveness. The expansion of these databases has significantly propelled medical AI research forward, marking a new phase of health care innovation with faster discoveries, collaborative efforts, and a commitment to ethical data use. As the field evolves, the role of open databases in shaping the future of medicine remains pivotal.

Conclusions

In conclusion, the evolution toward medical open databases, exemplified by the inception of platforms like PhysioNet in 1999 and their progression over the past 25 years, alongside the integration of PETs, marks a significant milestone in the domain of biomedical research and health care innovation. This journey not only fosters an unprecedented level of collaboration and accessibility but also emphasizes the crucial need to address privacy concerns and ethical considerations diligently. The ongoing efforts to balance data sharing with individual privacy protection are underscored by the adaptation of legal frameworks and the implementation of cryptographic and data management solutions. The introduction and growth of medical open databases have been pivotal, providing a wealth of data that has propelled research and innovation while highlighting the challenges and responsibilities of managing sensitive information. Specifically, the availability of open medical databases has significantly accelerated AI research, leading to breakthroughs in disease prediction, diagnostics, and personalized medicine. As we continue to explore the vast potential of open science and FL, the landscape of health care research is on the brink of remarkable transformations. These advancements promise enhanced patient outcomes, faster medical discoveries, and a more inclusive global research community, all achieved by adhering to the highest standards of privacy and ethical integrity.

XSL•FO
RenderX

Acknowledgments

The authors thank Dr Watson Lin for supporting the infrastructure of the data-sharing platform and for providing valuable perspectives on the guidelines for medical data sharing. The authors also thank the National Science and Technology Council for offering legal insights on privacy policies related to medical data sharing.

Funding

This work was supported by grants from the National Science and Technology Council, Taiwan (grant number NSCT 114-2634-F-A49-006 and 113-2634-F-A49-003). ACY was also supported by the Mt. Jade Young Scholarship Award from the Ministry of Education, Taiwan, as well as Brain Research Center, National Yang Ming Chiao Tung University, and the Ministry of Education (Aim for the Top University Plan), Taipei, Taiwan.

Authors' Contributions

Conceptualization; writing – original draft: AY;
Writing – review and editing: All Authors;
Funding acquisition: CHL

Conflicts of Interest

None declared.

References

1. Merelli I, Pérez-Sánchez H, Gesing S, D'Agostino D. Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives. Biomed Res Int 2014;2014:134023. [doi: 10.1155/2014/134023] [Medline: 25254202]
2. Roski J, Bo-Linn GW, Andrews TA. Creating value in health care through big data: opportunities and policy implications. Health Aff (Millwood) 2014 Jul;33(7):1115-1122. [doi: 10.1377/hlthaff.2014.0147] [Medline: 25006136]
3. Wang Y, Kung L, Byrd TA. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. Technol Forecast Soc Change 2018 Jan;126:3-13. [doi: 10.1016/j.techfore.2015.12.019]
4. Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000 Jun 13;101(23):E215-E220. [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]
5. Dean DA 2nd, Goldberger AL, Mueller R, et al. Scaling up scientific discovery in sleep medicine: the national sleep research resource. Sleep 2016 May 1;39(5):1151-1164. [doi: 10.5665/sleep.5774] [Medline: 27070134]
6. Wylie JE, Mineau GP. Biomedical databases: protecting privacy and promoting research. Trends Biotechnol 2003 Mar;21(3):113-116. [doi: 10.1016/S0167-7799(02)00039-2] [Medline: 12628367]
7. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124-130. [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]
8. Shahin MH, Bhattacharya S, Silva D, et al. Open data revolution in clinical research: opportunities and challenges. Clin Transl Sci 2020 Jul;13(4):665-674. [doi: 10.1111/cts.12756] [Medline: 32004409]
9. Ness RB, Joint Policy Committee, Societies of Epidemiology. Influence of the HIPAA privacy rule on health research. JAMA 2007 Nov 14;298(18):2164-2170. [doi: 10.1001/jama.298.18.2164] [Medline: 18000200]
10. Dwork C, Roth A. The algorithmic foundations of differential privacy. FNT in Theoretical Computer Science 2014;9(3-4):211-407. [doi: 10.1561/0400000042]
11. Rubin DB. Discussion: statistical disclosure limitation. J Off Stat 1993;9:461-468 [FREE Full text]
12. Yi X, Paulet R, Bertino E. Homomorphic encryption. In: Homomorphic Encryption and Applications: Springer; 2014:27-46. [doi: 10.1007/978-3-319-12229-8_2]
13. Cramer RJF, Damgård IB, Nielsen JB. Secure Multiparty Computation and Secret Sharing: Cambridge University Press; 2015. [doi: 10.1017/CBO9781107337756]
14. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. J Healthc Inform Res 2021 Mar;5(1):1-19. [doi: 10.1007/s41666-020-00082-4] [Medline: 33204939]
15. Packer M. Data sharing in medical research. BMJ 2018 Feb 14;360:k510. [doi: 10.1136/bmj.k510] [Medline: 29444885]
16. Moody GB, Mark RG, Goldberger AL. PhysioNet: physiologic signals, time series and related open source software for basic, clinical, and applied research. Annu Int Conf IEEE Eng Med Biol Soc 2011;2011:8327-8330. [doi: 10.1109/IEMBS.2011.6092053] [Medline: 22256277]
17. Moody GB, Mark RG, Goldberger AL. PhysioNet: a research resource for studies of complex physiologic and biomedical signals. Presented at: Computers in Cardiology 2000; Sep 24-27, 2000. [doi: 10.1109/CIC.2000.898485]
18. Henry IC, Goldberger AL, Moody GB, Mark RG. PhysioNet: an NIH research resource for physiologic datasets and open-source software. : IEEE Comput Soc Presented at: 14th IEEE Symposium on Computer-Based Medical Systems CBMS 2001; Jul 26-27, 2001; Bethesda, MD p. 245-250. [doi: 10.1109/CBMS.2001.941728]

XSL•FO
RenderX

19. Thomas RJ, Mietus JE, Peng CK, Goldberger AL. An electrocardiogram-based technique to assess cardiopulmonary coupling during sleep. Sleep 2005 Sep;28(9):1151-1161. [doi: 10.1093/sleep/28.9.1151] [Medline: 16268385]
20. Thomas RJ, Mietus JE, Peng CK, et al. Differentiating obstructive from central and complex sleep apnea using an automated electrocardiogram-based method. Sleep 2007 Dec;30(12):1756-1769. [doi: 10.1093/sleep/30.12.1756] [Medline: 18246985]
21. Collins R. What makes UK Biobank special? Lancet 2012 Mar 31;379(9822):1173-1174. [doi: 10.1016/S0140-6736(12)60404-8] [Medline: 22463865]
22. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature New Biol 2018 Oct;562(7726):203-209. [doi: 10.1038/s41586-018-0579-z] [Medline: 30305743]
23. Allen N, Sudlow C, Downey P, et al. UK Biobank: current status and what it means for epidemiology. Health Policy Technol 2012 Sep;1(3):123-126. [doi: 10.1016/j.hlpt.2012.07.003]
24. Littlejohns TJ, Holliday J, Gibson LM, et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nat Commun 2020 May 26;11(1):2624. [doi: 10.1038/s41467-020-15948-9] [Medline: 32457287]
25. Littlejohns TJ, Sudlow C, Allen NE, Collins R. UK Biobank: opportunities for cardiovascular research. Eur Heart J 2019 Apr 7;40(14):1158-1166. [doi: 10.1093/eurheartj/ehx254] [Medline: 28531320]
26. Azevedo T, Bethlehem RAI, Whiteside DJ, et al. Identifying healthy individuals with Alzheimer's disease neuroimaging phenotypes in the UK Biobank. Commun Med (Lond) 2023 Jul 20;3(1):100. [doi: 10.1038/s43856-023-00313-w] [Medline: 37474615]
27. Steinfeldt J, Buergel T, Loock L, et al. Neural network-based integration of polygenic and clinical information: development and validation of a prediction model for 10-year risk of major adverse cardiac events in the UK Biobank cohort. Lancet Digit Health 2022 Feb;4(2):e84-e94. [doi: 10.1016/S2589-7500(21)00249-1] [Medline: 35090679]
28. Clark K, Vendt B, Smith K, et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013 Dec;26(6):1045-1057. [doi: 10.1007/s10278-013-9622-7] [Medline: 23884657]
29. Markiewicz CJ, Gorgolewski KJ, Feingold F, et al. The OpenNeuro resource for sharing of neuroscience data. Elife 2021 Oct 18;10:e71774. [doi: 10.7554/eLife.71774] [Medline: 34658334]
30. Buccigrossi R, Ellisman M, Grethe J, et al. The neuroimaging informatics tools and resources clearinghouse (NITRC). AMIA Annu Symp Proc 2008 Nov 6;1000:1000. [Medline: 18999128]
31. Hall D, Huerta MF, McAuliffe MJ, Farber GK. Sharing heterogeneous data: the national database for autism research. Neuroinformatics 2012 Oct;10(4):331-339. [doi: 10.1007/s12021-012-9151-4] [Medline: 22622767]
32. Thompson HJ, Vavilala MS, Rivara FP. Chapter 1 common data elements and federal interagency traumatic brain injury research informatics system for TBI research. Annu Rev Nurs Res 2015;33(1):1-11. [doi: 10.1891/0739-6686.33.1] [Medline: 25946381]
33. Kaggle. URL: https://www.kaggle.com/ [accessed 2025-12-22]
34. Yang X, Zeng Z, Teo SG, Wang L, Chandrasekhar V, Hoi S. Deep learning for practical image recognition: case study on kaggle competitions. Presented at: KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; Aug 19-23, 2018; London, United Kingdom p. 923-931. [doi: 10.1145/3219819.3219907]
35. Krishna R, Kelleher K, Stahlberg E. Patient confidentiality in the research use of clinical medical databases. Am J Public Health 2007 Apr;97(4):654-658. [doi: 10.2105/AJPH.2006.090902] [Medline: 17329644]
36. Kobayashi S, Kane TB, Paton C. The privacy and security implications of open data in healthcare. Yearb Med Inform 2018 Aug;27(1):41-47. [doi: 10.1055/s-0038-1641201] [Medline: 29681042]
37. Eze B, Peyton L. Systematic literature review on the anonymization of high dimensional streaming datasets for health data sharing. Procedia Comput Sci 2015;63:348-355. [doi: 10.1016/j.procs.2015.08.353]
38. Centers for Disease Control and Prevention (CDC). HIPAA privacy rule and public health. Guidance from CDC and the U.S. Department of Health and Human Services. MMWR Suppl 2003 May 2;52:1-17. [Medline: 12741579]
39. Prasser F, Kohlmayer F, Spengler H, Kuhn KA. A scalable and pragmatic method for the safe sharing of high-quality health data. IEEE J Biomed Health Inform 2018 Mar;22(2):611-622. [doi: 10.1109/JBHI.2017.2676880] [Medline: 28358693]
40. Sarpatwari A, Gagne JJ. Balancing benefits and harms: privacy protection policies. Pharmacoepidemiol Drug Saf 2016 Aug;25(8):969-971. [doi: 10.1002/pds.4048] [Medline: 27278106]
41. Kulynych J, Korn D. The new HIPAA (Health Insurance Portability and Accountability Act of 1996) medical privacy rule: help or hindrance for clinical research? Circulation 2003 Aug 26;108(8):912-914. [doi: 10.1161/01.CIR.0000080642.35380.50] [Medline: 12939240]
42. Nosowsky R, Giordano TJ. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) privacy rule: implications for clinical research. Annu Rev Med 2006;57(1):575-590. [doi: 10.1146/annurev.med.57.121304.131257] [Medline: 16409167]
43. Gostin LO, Nass S. Reforming the HIPAA privacy rule: safeguarding privacy and promoting research. JAMA 2009 Apr 1;301(13):1373-1375. [doi: 10.1001/jama.2009.424] [Medline: 19336713]
44. Wartenberg D, Thompson WD. Privacy versus public health: the impact of current confidentiality rules. Am J Public Health 2010 Mar;100(3):407-412. [doi: 10.2105/AJPH.2009.166249] [Medline: 20075316]

XSL•FO
RenderX

45. Green AK, Reeder-Hayes KE, Corty RW, et al. The project data sphere initiative: accelerating cancer research by sharing data. Oncologist 2015 May;20(5):464-e20. [doi: 10.1634/theoncologist.2014-0431] [Medline: 25876994]
46. Conroy M, Sellors J, Effingham M, et al. The advantages of UK Biobank's open-access strategy for health research. J Intern Med 2019 Oct;286(4):389-397. [doi: 10.1111/joim.12955] [Medline: 31283063]
47. Anderson R. The collection, linking and use of data in biomedical research and health care: ethical issues. : The Nuffield Council on Bioethics; 2015. [doi: 10.17863/CAM.31760]
48. Kostkova P, Brewer H, de Lusignan S, et al. Who owns the data? Open data for healthcare. Front Public Health 2016;4:7. [doi: 10.3389/fpubh.2016.00007] [Medline: 26925395]
49. Bhattacharjee Y. Biomedicine. Pharma firms push for sharing of cancer trial data. Science 2012 Oct 5;338(6103):29. [doi: 10.1126/science.338.6103.29] [Medline: 23042862]
50. Kamoun F, Nicho M. Human and organizational factors of healthcare data breaches: the swiss cheese model of data breach causation and prevention. Int J Healthc Inf Syst Inform 2014 Jan 1;9:42-60. [doi: 10.4018/ijhisi.2014010103]
51. Nifakos S, Chandramouli K, Nikolaou CK, et al. Influence of human factors on cyber security within healthcare organisations: a systematic review. Sensors (Basel) 2021 Jul 28;21(15):5119. [doi: 10.3390/s21155119] [Medline: 34372354]
52. McLeod A, Dolezel D. Cyber-analytics: modeling factors associated with healthcare data breaches. Decis Support Syst 2018 Apr;108:57-68. [doi: 10.1016/j.dss.2018.02.007]
53. Seh AH, Zarour M, Alenezi M, et al. Healthcare data breaches: insights and implications. Healthcare (Basel) 2020 May 13;8(2):133. [doi: 10.3390/healthcare8020133] [Medline: 32414183]
54. Chernyshev M, Zeadally S, Baig Z. Healthcare data breaches: implications for digital forensic readiness. J Med Syst 2018 Nov 28;43(1):7. [doi: 10.1007/s10916-018-1123-2] [Medline: 30488291]
55. Ficek J, Wang W, Chen H, Dagne G, Daley E. Differential privacy in health research: a scoping review. J Am Med Inform Assoc 2021 Sep 18;28(10):2269-2276. [doi: 10.1093/jamia/ocab135] [Medline: 34333623]
56. Dankar F, Emam K. Practicing differential privacy in health care: a review. Trans Data Priv 2013;6(1):35-67. [doi: 10.5555/2612156.2612159]
57. Tamane S, Solanki VK. In: Dey N, editor. Privacy and Security Policies in Big Data: IGI Global; 2017. [doi: 10.4018/978-1-5225-2486-1]
58. Appenzeller A, Terzer N, Philipp P, Beyerer J. Applying differential privacy to medical questionnaires. Presented at: 2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops); Mar 13-17, 2023; Atlanta, GA p. 608-613. [doi: 10.1109/PerComWorkshops56833.2023.10150373]
59. Harris DR. Leveraging differential privacy in geospatial analyses of standardized healthcare data. Proc IEEE Int Conf Big Data 2020 Dec;2020:3119-3122. [doi: 10.1109/bigdata50022.2020.9378390] [Medline: 35253022]
60. Sun Z, Wang Y, Shu M, Liu R, Zhao H. Differential privacy for data and model publishing of medical data. IEEE Access 2019;7:152103-152114. [doi: 10.1109/ACCESS.2019.2947295] [Medline: 31328077]
61. McDuff D, Curran T, Kadambi A. Synthetic data in healthcare. arXiv. Preprint posted online on Apr 6, 2023. [doi: 10.48550/ARXIV.2304.03243]
62. Appenzeller A, Leitner M, Philipp P, Krempel E, Beyerer J. Privacy and utility of private synthetic data for medical data analyses. Appl Sci (Basel) 2022 Dec 1;12(23):12320. [doi: 10.3390/app122312320]
63. Qian Z, Callender T, Cebere B, Janes SM, Navani N, van der Schaar M. Synthetic data for privacy-preserving clinical risk prediction. medRxiv. Preprint posted online on May 24, 2023. [doi: 10.1101/2023.05.18.23290114]
64. Kokosi T, Harron K. Synthetic data in medical research. BMJ Med 2022;1(1):e000167. [doi: 10.1136/bmjmed-2022-000167] [Medline: 36936569]
65. Kaabachi B, Despraz J, Meurers T, et al. A scoping review of privacy and utility metrics in medical synthetic data. medRxiv. Preprint posted online on Oct 21, 2024. [doi: 10.1101/2023.11.28.23299124]
66. Wang Z, Myles P, Tucker A. Generating and evaluating cross‐sectional synthetic electronic healthcare data: preserving data utility and patient privacy. Comput Intell 2021 May;37(2):819-851. [doi: 10.1111/coin.12427]
67. Wood A, Najarian K, Kahrobaei D. Homomorphic encryption for machine learning in medicine and bioinformatics. ACM Comput Surv 2021 Jul 31;53(4):1-35. [doi: 10.1145/3394658]
68. Bos JW, Lauter K, Naehrig M. Private predictive analysis on encrypted medical data. J Biomed Inform 2014 Aug;50:234-243. [doi: 10.1016/j.jbi.2014.04.003] [Medline: 24835616]
69. Bocu R, Costache C. A homomorphic encryption-based system for securely managing personal health metrics data. IBM J Res & Dev 2018 Jan 1;62(1):1. [doi: 10.1147/JRD.2017.2755524]
70. Krishnegowda P, M. Boregowda A. Efficient matrix key homomorphic encryption of medical images. IJEECS 2023 Jul 1;31(1):406. [doi: 10.11591/ijeecs.v31.i1.pp406-416]
71. Kartit A. New approach based on homomorphic encryption to secure medical images in cloud computing. Trends Sci 2022;19(9):3970. [doi: 10.48048/tis.2022.3970]
72. Khedr A, Gulak G. SecureMed: secure medical computation using GPU-accelerated homomorphic encryption scheme. IEEE J Biomed Health Inform 2018 Mar;22(2):597-606. [doi: 10.1109/JBHI.2017.2657458] [Medline: 28129194]
73. Veeningen M, Chatterjea S, Horváth AZ, et al. Enabling analytics on sensitive medical data with secure multi-party computation. Stud Health Technol Inform 2018;247:76-80. [Medline: 29677926]

XSL•FO
RenderX

74. Marwan M, Kartit A, Ouahmane H. Applying secure multi-party computation to improve collaboration in healthcare cloud. Presented at: 2016 Third International Conference on Systems of Collaboration (SysCo); Nov 28-29, 2016; Casablanca, Morocco p. 1-6. [doi: 10.1109/SYSCO.2016.7831325]
75. Yigzaw KY, Bellika JG. Evaluation of secure multi-party computation for reuse of distributed electronic health data. Presented at: 2014 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI); Jun 1-4, 2014; Valencia, Spain p. 219-222. [doi: 10.1109/BHI.2014.6864343]
76. Şahinbaş K, Catak FO. Secure multi-party computation-based privacy-preserving data analysis in healthcare IoT systems. In: Kose U, Gupta D, Khanna A, Rodrigues J, editors. Interpretable Cognitive Internet of Things for Healthcare: Springer; 2023:57-72. [doi: 10.1007/978-3-031-08637-3_3]
77. Tso R, Alelaiwi A, Mizanur Rahman SM, Wu ME, Hossain MS. Privacy-preserving data communication through secure multi-party computation in healthcare sensor cloud. J Sign Process Syst 2017 Oct;89(1):51-59. [doi: 10.1007/s11265-016-1198-2]
78. Tawfik AM, Sabbeh SF, EL-Shishtawy T. Privacy-preserving secure multiparty computation on electronic medical records for star exchange topology. Arab J Sci Eng 2018 Dec;43(12):7747-7756. [doi: 10.1007/s13369-018-3122-5]
79. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. IEEE Signal Process Mag 2020 May;37(3):50-60. [doi: 10.1109/MSP.2020.2975749]
80. Nguyen DC, Pham QV, Pathirana PN, et al. Federated learning for smart healthcare: a survey. ACM Comput Surv 2023 Mar 31;55(3):1-37. [doi: 10.1145/3501296]
81. Crowson MG, Moukheiber D, Arévalo AR, et al. A systematic review of federated learning applications for biomedical data. PLOS Digit Health 2022 May;1(5):e0000033. [doi: 10.1371/journal.pdig.0000033] [Medline: 36812504]
82. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated electronic health records. Int J Med Inform 2018 Apr;112:59-67. [doi: 10.1016/j.ijmedinf.2018.01.007] [Medline: 29500022]
83. Antunes RS, André da Costa C, Küderle A, Yari IA, Eskofier B. Federated learning for healthcare: systematic review and architecture proposal. ACM Trans Intell Syst Technol 2022 Aug 31;13(4):1-23. [doi: 10.1145/3501813]
84. Guo K, Chen T, Ren S, Li N, Hu M, Kang J. Federated learning empowered real-time medical data processing method for smart healthcare. IEEE/ACM Trans Comput Biol Bioinform 2024;21(4):869-879. [doi: 10.1109/TCBB.2022.3185395] [Medline: 35737631]
85. Pfitzner B, Steckhan N, Arnrich B. Federated learning in a medical context: a systematic literature review. ACM Trans Internet Technol 2021 Jun 23;21(2):1-31. [doi: 10.1145/3412357]
86. Wu JCH, Yu HW, Tsai TH, Lu HHS. Dynamically synthetic images for federated learning of medical images. Comput Methods Programs Biomed 2023 Dec;242:107845. [doi: 10.1016/j.cmpb.2023.107845] [Medline: 37852147]
87. Liu HH. Use and disclosure of health information and protection of patient privacy in Taiwan. Med Law 2010 Mar;29(1):87-101. [Medline: 22458000]
88. Jin H, Luo Y, Li P, Mathew J. A review of secure and privacy-preserving medical data sharing. IEEE Access 2019;7:61656-61669. [doi: 10.1109/ACCESS.2019.2916503]
89. Liang X, Zhao J, Shetty S, Liu J, Li D. Integrating blockchain for data sharing and collaboration in mobile healthcare applications. Presented at: 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC); Oct 8-13, 2017; Montreal, QC p. 1-5. [doi: 10.1109/PIMRC.2017.8292361]
90. Rau HH, Hsu CY, Lee YL, Chen W, Jian WS. Developing electronic health records in Taiwan. IT Prof 2010 Mar;12(2):17-25. [doi: 10.1109/MITP.2010.53]
91. Li YCJ, Yen JC, Chiu WT, Jian WS, Syed-Abdul S, Hsu MH. Building a national electronic medical record exchange system - experiences in Taiwan. Comput Methods Programs Biomed 2015 Aug;121(1):14-20. [doi: 10.1016/j.cmpb.2015.04.013] [Medline: 26001420]
92. Hwang HG, Han HE, Kuo KM, Liu CF. The differing privacy concerns regarding exchanging electronic medical records of internet users in Taiwan. J Med Syst 2012 Dec;36(6):3783-3793. [doi: 10.1007/s10916-012-9851-1] [Medline: 22527781]
93. Yang CM, Lin HC, Chang P, Jian WS. Taiwan's perspective on electronic medical records' security and privacy protection: lessons learned from HIPAA. Comput Methods Programs Biomed 2006 Jun;82(3):277-282. [doi: 10.1016/j.cmpb.2006.04.002] [Medline: 16730852]
94. Taiwan medical AI and data portal. NYCU Data Management Center. 2023. URL: https://data.dmc.nycu.edu.tw/ [accessed 2025-12-22]
95. Kaye J, Whitley EA, Lund D, Morrison M, Teare H, Melham K. Dynamic consent: a patient interface for twenty-first century research networks. Eur J Hum Genet 2015 Feb;23(2):141-146. [doi: 10.1038/ejhg.2014.71] [Medline: 24801761]
96. Williams H, Spencer K, Sanders C, et al. Dynamic consent: a possible solution to improve patient confidence and trust in how electronic patient records are used in medical research. JMIR Med Inform 2015 Jan 13;3(1):e3. [doi: 10.2196/medinform.3525] [Medline: 25586934]
97. Lee H, Lee U. Toward dynamic consent for privacy-aware pervasive health and well-being: a scoping review and research directions. IEEE Pervasive Comput 2022 Oct 1;21(4):25-32. [doi: 10.1109/MPRV.2022.3210747]
98. Budin-Ljøsne I, Teare HJA, Kaye J, et al. Dynamic consent: a potential solution to some of the challenges of modern biomedical research. BMC Med Ethics 2017 Jan 25;18(1):4. [doi: 10.1186/s12910-016-0162-9] [Medline: 28122615]

XSL•FO
RenderX

99. Albalwy F, Brass A, Davies A. A blockchain-based dynamic consent architecture to support clinical genomic data sharing (consentchain): proof-of-concept study. JMIR Med Inform 2021 Nov 3;9(11):e27816. [doi: 10.2196/27816] [Medline: 34730538]
100. Goncharov L, Suominen H, Cook M. Dynamic consent and personalised medicine. Med J Aust 2022 Jun 20;216(11):547-549. [doi: 10.5694/mja2.51555] [Medline: 35611469]
101. Vlahou A, Hallinan D, Apweiler R, et al. Data sharing under the general data protection regulation: time to harmonize law and research ethics? Hypertension 2021 Apr;77(4):1029-1035. [doi: 10.1161/HYPERTENSIONAHA.120.16340] [Medline: 33583200]
102. Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Curr Cardiol Rep 2014 Jan;16(1):441. [doi: 10.1007/s11886-013-0441-8] [Medline: 24338557]
103. Pereira T, Morgado J, Silva F, et al. Sharing biomedical data: strengthening AI development in healthcare. Healthcare (Basel) 2021 Jun 30;9(7):827. [doi: 10.3390/healthcare9070827] [Medline: 34208830]
104. Paton C, Kobayashi S. An open science approach to artificial intelligence in healthcare. Yearb Med Inform 2019 Aug;28(1):47-51. [doi: 10.1055/s-0039-1677898] [Medline: 31022753]
105. Lim L, Lee HC. Open datasets in perioperative medicine: a narrative review. Anesth Pain Med (Seoul) 2023 Jul;18(3):213-219. [doi: 10.17085/apm.23076] [Medline: 37691592]
106. Wilson JR, Prevedello LM, Witiw CD, Flanders AE, Colak E. Data liberation and crowdsourcing in medical research: the intersection of collective and artificial intelligence. Radiol Artif Intell 2024 Jan;6(1):e230006. [doi: 10.1148/ryai.230006] [Medline: 38231037]
107. Kasparick M, Andersen B, Franke S, et al. Enabling artificial intelligence in high acuity medical environments. Minim Invasive Ther Allied Technol 2019 Apr;28(2):120-126. [doi: 10.1080/13645706.2019.1599957] [Medline: 30950665]
108. Koh DM, Papanikolaou N, Bick U, et al. Artificial intelligence and machine learning in cancer imaging. Commun Med (Lond) 2022;2(1):133. [doi: 10.1038/s43856-022-00199-0] [Medline: 36310650]
109. Levin MG, Rader DJ. Polygenic risk scores and coronary artery disease: ready for prime time? Circulation 2020 Feb 25;141(8):637-640. [doi: 10.1161/CIRCULATIONAHA.119.044770] [Medline: 32091922]
110. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023 Jan 3;10(1):1. [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
111. Milham MP, Craddock RC, Son JJ, et al. Assessment of the impact of shared brain imaging data on the scientific literature. Nat Commun 2018 Jul 19;9(1):2818. [doi: 10.1038/s41467-018-04976-1] [Medline: 30026557]
112. Patterson E, McBurney R, Schmidt H, Baldini I, Mojsilovic A, Varshney KR. Dataflow representation of data analyses: toward a platform for collaborative data science. IBM J Res & Dev 2017 Nov 1;61(6):9. [doi: 10.1147/JRD.2017.2736278]
113. El Naqa I, Haider MA, Giger ML, Ten Haken RK. Artificial intelligence: reshaping the practice of radiological sciences in the 21st century. Br J Radiol 2020 Feb 1;93(1106):20190855. [doi: 10.1259/bjr.20190855] [Medline: 31965813]
114. Kras A, Celi LA, Miller JB. Accelerating ophthalmic artificial intelligence research: the role of an open access data repository. Curr Opin Ophthalmol 2020 Sep;31(5):337-350. [doi: 10.1097/ICU.0000000000000678] [Medline: 32740059]

Abbreviations

AI: artificial intelligence
CT: computed tomography
EMR: electronic medical record
FL: federated learning
HIPAA: Health Insurance Portability and Accountability Act
ML: machine learning
MRI: magnetic resonance imaging
PET: privacy-enhancing technology
PHI: protected health information

XSL•FO
RenderX

Edited by G Eysenbach, T Leung; submitted 31.Mar.2024; peer-reviewed by Imran, Y Sun; revised version received 28.Oct.2025; accepted 29.Nov.2025; published 30.Jan.2026.

Please cite as:
Yang A, Pan ML, Lu HHS, Lien CY, Wang DW, Chen CH, Tarng DC, Niu DM, Chiou SH, Wu CY, Sun YC, Chen SA, Wang SJ, Sheu WHH, Lin CH
Assessing the Evolution and Influence of Medical Open Databases on Biomedical Research and Health Care Innovation: A 25-Year Perspective With a Focus on Privacy and Privacy-Enhancing Technologies
J Med Internet Res 2026;28:e58954
URL: https://www.jmir.org/2026/1/e58954
doi:10.2196/58954

© Albert Yang, Mei-Lien Pan, Henry Horng-Shing Lu, Chung-Yueh Lien, Da-Wei Wang, Chih-Hsiung Chen, Der-Cherng Tarng, Dau-Ming Niu, Shih-Hwa Chiou, Chun-Ying Wu, Ying - Chou Sun, Shih-Ann Chen, Shuu-Jiun Wang, Wayne Huey-Herng Sheu, Chi-Hung Lin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 30.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Africa's Digital Health Revolution: The Digital Fit-Viability Model to Move From Innovation to Scaled Implementation

Afra Jiwa[1,2*], MSc, MBBS; Antony Ngatia[3*], MSc, MBChB; Karim Benali[4], MS, MD; Niclas Boehmer[5], PhD; Sangu Delle[6], MBA, JD, PhD; Patrick Emedom-Nnamdi[7], PhD; Chris Opoku Fofie[8], MD, FGCS, MPH; Christine M O'Brien[9], BS, PhD; Tobi Olatunji[10], MSc, MBBS; Kate Obayabgona[1], MPH, MD; Milind Tambe[11], MSc, PhD; Richard Ribon Fletcher[12], BS, MS, PhD; Adeline Adwoa Boatin[1,13*], MPH, MD; Bethany Hedt-Gauthier[1,14,15*], PhD

[1]Program for Global Surgery and Social Change, Harvard Medical School, Boston, MA, United States
[2]Usher Institute, University of Edinburgh, Edinburgh, United Kingdom
[3]Statsspeak Analytics, Nairobi, Kenya
[4]Global Health and Service Advisory Council, Harvard Medical School, Boston, MA, United States
[5]Hasso Plattner Institute, University of Potsdam, Postdam, Germany
[6]CarePoint, Accra, Ghana
[7]Department of Biostatistics, Harvard T.Chan School of Public Health, Boston, MA, United States
[8]Ghana Health Service, Accra, Ghana
[9]Department of Biomedical Engineering, Washington University in St. Louis, St Louis, MO, United States
[10]Intron Health, London, United Kingdom
[11]Center for Research on Computation and Society, Harvard John a. Paulson School of Engineering and Applied Sciences, Boston, MA, United States
[12]Mechanical Engineering Department, Massachusetts Institute of Technology, Boston, MA, United States
[13]Department of Obstetrics and Gynecology, Massachusetts General Hospital, Boston, MA, United States
[14]Department of Maternal Child Health, University of North Carolina Chapel Hill, Chapel Hill, NC, United States
[15]Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA, United States
[*]these authors contributed equally

Corresponding Author:
Adeline Adwoa Boatin, MPH, MD
Department of Obstetrics and Gynecology
Massachusetts General Hospital
55 Fruit Street
Boston, MA, 02114
United States
Phone: 1 617 724 3775
Email: adeline_boatin@mgh.harvard.edu

Abstract

Digital innovations hold immense potential to transform health care delivery, particularly in sub-Saharan Africa, where financial, geographical, and infrastructural constraints continue to hinder progress toward universal health care delivery. Although a growing health tech sector offers creative solutions, few digital health interventions reach scaled implementation. In this paper, we present the digital fit/viability model—an adapted determinant framework to describe facilitators and barriers to moving from digital tools to integrated digital health implementation. We then use this model to describe the specific challenges and recommended solutions when developing digital health tools for health systems in sub-Saharan Africa.

(J Med Internet Res 2026;28:e63495) doi:10.2196/63495

KEYWORDS

digital health; Africa; digital health solutions; digital technologies; health care access; fit/viability model

XSL•FO
RenderX

Introduction

Digital tools can improve health care delivery and enhance health outcomes by addressing long-standing challenges in health care access, quality, and efficiency [1]. This potential is particularly salient in sub-Saharan Africa, where persistent barriers to health care, including inadequate infrastructure, workforce shortages, and resource limitations, could be alleviated or entirely bypassed with digital health interventions [2,3]. There are notable examples of scaled digital health interventions in sub-Saharan Africa. For example, District Health Information Software 2, a digital informatics platform operating in 40 countries, facilitates electronic data entry at health facilities, making information available for surveillance and program monitoring in near real time [4]. Zipline offers a digital platform for immediate order and drone-led distribution of health products in Rwanda [5]. In Kenya, M-TIBA allows users to access their health insurance through a mobile payment technology called m-pesa [6]. However, despite some successes, the digital health field suffers from "pilotitis," where the vast majority of digital interventions do not move beyond the pilot phase [7].

Over the last decade, the World Health Organization (WHO) and the World Bank have issued several policy documents outlining priority areas for digital interventions, as well as identifying systems gaps that need to be addressed for digital health interventions to achieve their full potential [8-11]. However, these documents do not systematically specify the facilitators and barriers that affect the progression of a specific digital health tool from concept to implementation. Drawing from our experience in developing, researching, and implementing digital health interventions in sub-Saharan Africa, we introduce the digital fit/viability model (dFVM) for health. This model adapts the fit/viability model developed by Liang et al [12] for the use of mobile technology in business and is an example of an implementation science determinant framework to describe the facilitators and barriers that influence implementation outcomes. In this paper, we explain how this model can be adapted for digital health interventions and use dFVM to illustrate challenges specific to sub-Saharan Africa as well as the potential solutions.

Positionality Statement

This work is the result of a 2-day convening of a multidisciplinary group of clinicians, researchers, technologists, entrepreneurs, policymakers, and health delivery service providers. The group included individuals with experience in Ghana, Kenya, Nigeria, Rwanda, Tanzania, Uganda, United States, and United Kingdom, and spanned different age groups, ethnicities, genders, and professional experience; all had experience of working within the African digital health ecosystem. The group held the shared perspective that digital tools have the potential to transform health care in sub-Saharan Africa. From this perspective, we discussed challenges to the implementation of digital innovations in these contexts and strategies to develop, evaluate, and scale context-appropriate and sustainable digital health solutions.

Digital Tools, Interventions, and Implementation

Here, we define digital health tools, interventions, and implementation to clarify the role of dFVM. Digital health tools encompass a variety of technology-enabled products and services developed to improve health care services [13]. Broadly, digital health tools have been categorized into virtual interactions (such as platforms for teleconsultations), paperless data (such as electronic health records), patient self-care (such as mobile apps to support chronic disease management), patient self-service (such as e-booking platforms), decision intelligence systems (such as hospital patient flow management systems), and workflow automation (such as medical equipment tracking systems using radiofrequency identification) [13]. Digital health interventions refer to the services, practices, and strategies that utilize digital health tools. For example, Zipline is an intervention that engages a cluster of digital tools—drones, online ordering systems, etc—to facilitate timely order-to-delivery of health products. Finally, digital health implementation involves integrating a digital health intervention into broader health care delivery. For instance, the successful deployment of Zipline is embedded within the larger health care framework, where the ordered health products are identified as part of patient care and delivered accordingly, outside of the Zipline system itself. Despite a surge in digital health tools over the past decade, few have successfully been integrated into interventions, and even fewer digital health interventions have achieved widespread, sustained implementation into health care systems, particularly in sub-Saharan Africa [1].

The Fit/Viability Model

Liang et al [12] developed the fit/viability model in 2007 to support the adoption of new technology in the business sector [14]. In this original framework, "fit" evaluates how well a technology aligns with its intended environment, while "viability" assesses whether the technology can be feasibly implemented given the organization's constraints. Both fit and viability are essential for a technology to succeed at scale within a business. These overarching "fit" and "viability" questions have been valuable for our team when developing digital health tools as part of complex interventions in sub-Saharan Africa. In the next section, we explain how we adapted Liang et al's [12] framework and incorporated additional subdomains to create dFVM for health (Figure 1; [15]).

XSL•FO
RenderX

Figure 1. The digital fit/viability model (adapted from Liang et al [12], which is published under Creative Commons Attribution 4.0 International License [15]).

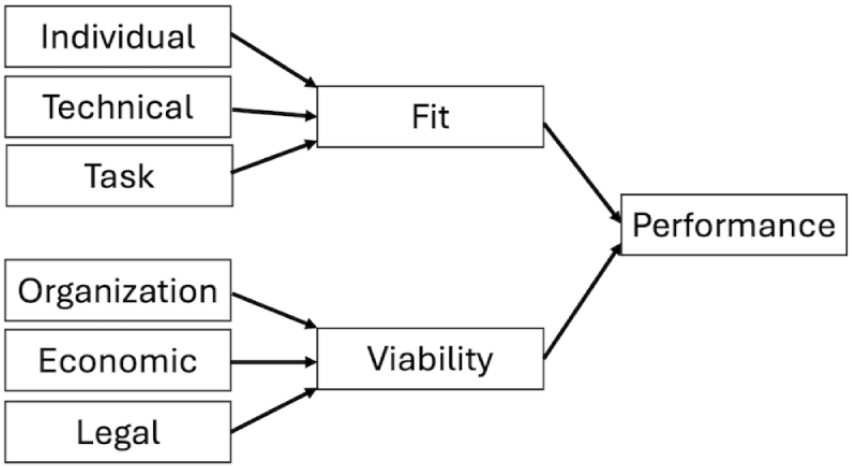

dFVM Concept

Defining the "Fit" of Digital Health Tools

Similar to that defined by Liang et al [12], we define "fit" as the appropriateness of a digital health tool in a specific setting as part of a specific intervention. Generally, we are asking questions such as "Who will be using this digital health tool?", "What is this digital health tool designed to do?", and "Does this digital health tool align with the organizations aims?" (Figure 1). These questions are best addressed in the early design and conception stages. Although these fit concepts are important for a digital health tool in any context, the particular challenges in sub-Saharan Africa require deeper exploration of these questions.

Individual Fit

Individual fit, which was not included in the original fit/viability model, assesses whether the digital health tool is appropriate for those who engage with the tool. This category emphasizes that technologies should be designed with end users in mind.

Challenge

In sub-Saharan Africa, one particular challenge for individual fit is the varied levels of digital literacy [16,17] and access to digital platforms [9,18]. For example, a digital intervention that requires a patient to have a smartphone to engage with an app may fail either because the patient does not own a phone and/or because the patient is not comfortable navigating a digital App.

Solution

As a first step, anyone developing a digital tool should conduct end user digital accessibility assessments, including assessing digital literacy, device ownership, and access to other system requirements such as cell networks or electricity. There are numerous tools to assess digital literacy [14], but few are designed specifically for a sub-Saharan African context. One alternative is to outline the gaps in device ownership or digital literacy that must be addressed when deploying the digital health intervention so that anyone adopting a digital health tool understands the pathway to ensure individual fit.

Technical Fit

Technical fit assesses whether the technological infrastructure exists to support the digital health tool. This domain was not included in the original fit/viability model developed by Liang et al [12], which focused on deploying e-commerce tools in high-income country contexts—settings already primed for new technologies. In our iteration, we consider the potential mismatch between the infrastructure in which a tool was designed and the setting in which it will be deployed.

Challenge

In many sub-Saharan African settings, foundational infrastructure such as electricity and internet remains unreliable, and hardware needs such as hardware durability, power consumption, and maintenance are often overlooked. Tools developed for high-income country contexts may rely on stable cloud access, regular software updates, or consumables that are expensive or difficult to source in sub-Saharan Africa. As a result, imported technologies often end up in "equipment graveyards," unused due to incompatibility with local systems, lack of technical support, or excessive training requirements [19-22].

Solution

The first question should often be "Are there nondigital solutions that are equally suited for this task?", ensuring that implementers are not missing more accessible, viable, and equally impactful solutions. When digital health interventions offer clear advantages, developers must assess infrastructure needs early and design for low-resource environments, considering offline functionality, portability, power efficiency, and local repair capacity. Nationally, more advocacy is needed to support initiatives that expand digital infrastructure. Projects like Health Connect Africa [23], a partnership between Centre for Disease Control Africa and Global System for Mobile Communications Association, which aims to connect 10,000 health care facilities to the internet by 2030, and Power Africa [24], which focuses on electrifying health facilities, are examples of essential infrastructure-building efforts that make digital health tools viable at scale.

XSL•FO
RenderX

Task Fit

We expand Liang et al's [12] considerations of task fit to assess whether a digital health tool is able to complete the task for which it is designed. A tool may be well matched to the infrastructure and the user but still fail to achieve the task for which it is intended.

Challenge

In the digital health field, we often start with tools developed for high-resource environments and adapt these tools to accomplish specific tasks in sub-Saharan Africa [25], which can lead to numerous problems. One notable example is for digital tools that engage artificial intelligence. These often do not include data representative of individuals in sub-Saharan Africa and can have high bias and low validity when implemented in those settings [26].

Solution

First, we must center health program implementers and policymakers to identify the priority gaps we are addressing through digital health interventions. Once we have these priority areas, we must move beyond simply adopting tools and instead validate them through dedicated studies, be willing to adapt or retrain them for context, and, when necessary, design tools from scratch. For AI-enabled digital tools, better and more accessible datasets are needed on the African continent. Although funding and data formatting have been obstacles, recent initiatives such as the National Institutes of Health–funded DS-i Africa [27] along with Masakhane [28], the Lacuna Fund, and Zindi [29], are actively creating shared open repositories. These platforms will accelerate health care innovation by providing representative, high-quality data that directly addresses the bias and validity issues of the existing tools.

Defining the "Viability" of Digital Health Tools

Liang et al [12] considered viability as the organization's economic, technical, and social readiness to adopt a digital tool. We have expanded this to also include legal and ethical considerations that present significant challenges when developing digital tools for interventions and implementation in sub-Saharan Africa. The viability domain considers "Can this tool or intervention thrive in the ecosystem into which it is being deployed?", "Does this tool or intervention have the legal and financial backing to be sustainable?", and "Is the technical infrastructure available to support this tool?" Although fit questions should be addressed early, viability questions are addressed as the digital tool is integrated into interventions and implemented on small or broad scales.

Organization Viability

In the original model, organization viability describes factors such as management support, digital literacy of team members, user competence, and experience [12]. We have shifted several of these individual features such as digital literacy and user competence to "fit" since these must be addressed early as the digital tool is being developed. We have expanded the management support to include plausible integration of the tool into program implementation structures. Adopting technologies that are well matched to their organizational environment can promote uptake and sustainability.

Challenge

Organizational viability considers whether health systems and institutions have the leadership, governance, and operational capacity to sustain digital health interventions. In sub-Saharan Africa, frequent staff turnover and fragmented leadership structures often undermine long-term adoption [30]. Ministries of health may have alternative health priorities or lack the resources to integrate new tools across departments and levels of care. These factors mean that even technically sound and contextually appropriate interventions may struggle to move beyond pilot projects if organizational systems are not aligned to support them.

Solution

Early involvement of multidisciplinary teams is key to assessing feasibility and organizational readiness. Usability testing can help establish workflows, training, and tools that support smooth adoption. For developers, designing solutions that integrate easily with existing infrastructure and that can be bundled with current systems promotes interoperability and long-term alignment. Engaging leadership early and ensuring interventions reflect national health priorities helps secure ownership, buy-in, and resource allocation, thereby reducing the risk of tools operating in isolation from the systems they aim to strengthen.

Economic Viability

In the study of Liang et al [12], economic viability encompasses a cost-effectiveness and cost-benefit analysis. We extend this definition to ask whether a digital health intervention can be sustained financially, including through Ministries of Health and Finance, with the support of philanthropy or multilateral organizations or through the private sector.

Challenge

The digital health sector in sub-Saharan Africa is a patchwork of solutions, with many promising digital health interventions failing to achieve scale due to a lack of coordinated national strategies or shared infrastructure [31]. This decentralized approach, often led by individual clinicians, researchers, and small nongovernmental organizations, limits the ability to benefit from economies of scale [32]. This is compounded by persistent infrastructural barriers like high data costs, unreliable electricity, and poor connectivity as well as inconsistent funding avenues [33]. These issues place a disproportionate financial burden on smaller organizations and new companies. The high cost of entry and limited seed funding make it difficult for promising early-stage innovations to gain traction, leading to isolated successes rather than widespread impact.

Solution

To ensure the continued funding of digital health interventions, a combination of coordinated investment and new funding models are needed. Blended financing, which combines public funds, donor contributions, and socially responsible investment, can help bridge the gap and move interventions from pilot to scale [34]. Health systems–level investment into basic

XSL•FO
RenderX

infrastructure and platform will make any single digital health intervention more economical and cost-beneficial [9].

Legal Viability

Legal viability refers to the presence of clear, enforceable policies that govern the implementation of digital health technologies, with a focus on data privacy, security, and system interoperability. Although this was not included in Liang et al's [12] model, in our experience, this has been a major challenge and often missed step for digital health interventions in sub-Saharan Africa.

Challenge

National and regional policies for digital health interventions are nascent in sub-Saharan Africa. Without clear regulations on data privacy, security, and interoperability, implementers face an uncertain and rapidly changing regulatory landscape [35-37]. This lack of maturity in frameworks creates significant uncertainty for both innovators and investors, making it difficult to establish a strong foundation for technological advancement [35].

Solution

Strengthening legal viability requires investing in national and regional regulatory capacity and developing robust data governance systems. This includes improving regulations to ensure trustworthiness while still supporting innovation [38]. Harmonizing standards across borders will facilitate the ability to share and use tools beyond the specific contexts for which they were developed [36].

dFVM Application

Table 1 shows a summary of the application of dFVM to identify the challenges and possible solutions for implementing digital health interventions in sub-Saharan Africa.

Table 1. Applying dFVM[a] to identify the challenges and possible solutions for implementing digital health interventions in sub-Saharan Africa.

Domains, subdomains of dFVM	Challenges and solutions identified when applying dFVM to sub-Saharan Africa	
	Challenges	Solutions
Fit		
Individual	Limited digital literacy and access to digital platforms among potential end users.	Conduct end user digital assessments, including digital literacy and device ownership assessments prior to tool development. Build digital skills and awareness as part of intervention development.
Technical	Unreliable electricity, frequent power outages, poor internet connectivity, high cost of mobile internet data.	Consider nondigital alternatives. Assess infrastructural capacity early and develop tools that comply to that infrastructure. Invest in foundational infrastructure to connect facilities, health workers, and patients.
Task	Imported technologies may not align with the actual tasks or workflows in health systems. Tools often assume task environments similar to high-income countries. Artificial intelligence tools can also perform poorly due to algorithmic bias.	Address priority areas identified by health practitioners and policymakers in the local context. Ensure co-design with frontline health workers to reflect actual care tasks and workflows. Test, validate, and adapt digital health tools as needed. Build Africa-specific data repositories.
Viability		
Organizational	Fragmented leadership, staff turnover, and organizational priorities, which differ from the tools that are developed.	Bundle tools with existing infrastructure, promoting interoperability and integration; engage leadership early; and promote system-level alignment and ownership.
Economic	Insufficient funding and lack of sustainable financing mechanisms leads to premature abandonment of promising initiatives or poor scaling of established interventions.	Invest in national digital health platforms, which will reduce the per-tool costs. Prioritize national funding models, complemented by blended public-private financing.
Legal	Lack of clear policies and regulations regarding data privacy, security, and interoperability.	Develop robust data governance and digital regulations for countries. Seek, where possible, cross-continent regulations to facilitate portability.

[a]dFVM: digital fit/viability model.

Discussion

We developed dFVM to help individuals designing digital health tools and interventions, particularly innovators working in sub-Saharan Africa, think more critically about the development of these tools and interventions with the goal of successful digital health implementation. We have identified 3 other determinant frameworks relevant to digital health interventions, namely, the Train, Restructure, Incentivize, Mandate, Integrate (TRIMI) framework [16]; a modified version of the Consolidated Framework for Implementation Research (mCFIR) for mobile health tools [17]; and a third developed by Olu et al [3]. Each of these frameworks offers a different vantage to consider when developing digital health tools and interventions. For this paper, we adapted the fit/viability model because this most closely mirrored how we approach our digital tool development—asking questions about whether a digital health tool will fit a specific task, for a specific user, in a specific

XSL•FO
RenderX

infrastructure, and whether a digital health tool achieves the organizational integration, financing, and legal compliance needed to be viable.

Although applicable to any digital health tool, we made sure that dFVM addresses the unique obstacles we have encountered in sub-Saharan Africa. Throughout our description of dFVM, we have highlighted these challenges and offered solutions. Broadly, these solutions fall into 2 buckets. First, what is the responsibility of the individuals developing the tools and interventions? These individuals must design for context and, through the deployment of digital health interventions, build institutional capacity for digital health interventions more broadly. Second, in the context of sub-Saharan Africa, governments play a crucial role, particularly in ensuring the viability of digital health interventions. Investing in digital infrastructure provides a foundation for sustaining various tools and reduces per-tool costs. Moreover, establishing robust legal frameworks, including cross-national agreements, ensures compliance and demonstrates government commitment as partners alongside private and public technology providers.

dFVM offers high-level guidance for individuals developing digital health tools and interventions to ensure their fit and viability within health systems. However, we recognize that moving from idea-to-implementation requires more detailed steps and involves multidisciplinary teams capable of navigating these complexities. As a next step, we are expanding dFVM and integrating it with other digital health implementation science frameworks to develop a step-by-step roadmap to guide teams throughout the development process.

Conclusion

We believe that digital health interventions have great potential to transform health care delivery and improve outcomes worldwide. However, successfully implementing these interventions in sub-Saharan Africa requires careful attention to their fit and viability within local contexts. dFVM maps challenges across both digital health design and deployment environments and links these challenges to practical solutions drawn from lived experience. Although digital health tool developers are responsible for addressing these considerations, governments in sub-Saharan Africa play a crucial role in ensuring that the necessary organizational and legal frameworks are established to support the adoption of digital health interventions.

Acknowledgments

We acknowledge the contributions of all authors on this paper and the role of the Harvard Radcliffe Institute in hosting the seminar from which this paper was conceptualized.

Funding

The workshop was hosted by the Harvard Radcliffe Institute. Additional funding was provided by the Harvard Global Health Institute and Harvard's Motsepe Presidential Research Accelerator Fund for Africa.

Authors' Contributions

BH-G and AAB conceptualized and acquisitioned funding for this work. AJ and AN were writers of the original draft with intensive inputs from NB, PE-N, and TO. All authors participated in the investigation, developing the content through an iterative engagement process, and the writing (reviewing and editing).

Conflicts of Interest

None declared.

References

1. Holst C, Sukums F, Radovanovic D, Ngowi B, Noll J, Winkler AS. Sub-Saharan Africa-the new breeding ground for global digital health. Lancet Digit Health 2020 Apr;2(4):e160-e162 [FREE Full text] [doi: 10.1016/S2589-7500(20)30027-3] [Medline: 33328076]
2. Digital health systems in Africa. IQVIA. 2023 Sep 23. URL: https://www.iqvia.com/locations/middle-east-and-africa/library/white-papers/digital-health-systems-in-africa?_gl=1*kezpey*_up*MQ..*_gs*MQ [accessed 2025-08-08]
3. Olu O, Muneene D, Bataringaya JE, Nahimana M, Ba H, Turgeon Y, et al. How can digital health technologies contribute to sustainable attainment of universal health coverage in Africa? A perspective. Front Public Health 2019;7:341 [FREE Full text] [doi: 10.3389/fpubh.2019.00341] [Medline: 31803706]
4. DHIS2. URL: https://dhis2.org/ [accessed 2025-04-04]
5. Zipline Drone Delivery & Logistics. URL: https://www.flyzipline.com/ [accessed 2024-06-20]
6. M-TIBA. URL: https://mtiba.com/ [accessed 2025-04-04]
7. Egermark M, Blasiak A, Remus A, Sapanel Y, Ho D. Overcoming pilotitis in digital medicine at the intersection of data, clinical evidence, and adoption. Advanced Intelligent Systems 2022 May 26;4(9):2200056. [doi: 10.1002/aisy.202200056]
8. Global Strategy on Digital Health 2020-2025. Geneva: World Health Organization; 2021.

XSL•FO
RenderX

9. WHO Guideline: Recommendations on Digital Interventions for Health System Strengthening. Geneva: World Health Organization; 2019.
10. Kelley E, Zandi D, Krishnamurthy R, Mehl G, Novillo D, Muneene D, et al. Digital Health Platform Handbook: Building a Digital Information Infrastructure (Infostructure) for Health. Geneva: World Health Organization and International Telecommunication Union; 2020.
11. Digital-in-Health: Unlocking the Value for Everyone. Washington, DC: The World Bank; 2023.
12. Liang T, Huang C, Yeh Y, Lin B. Adoption of mobile technology in business: a fit‐viability model. Industr Mngmnt & Data Systems 2007 Oct 02;107(8):1154-1169. [doi: 10.1108/02635570710822796]
13. How digital tools could boost efficiencey in African health systems. McKinsey and Company. URL: https://www.mckinsey.com/industries/healthcare/our-insights/how-digital-tools-could-boost-efficiency-in-african-health-systems [accessed 2024-06-18]
14. Nguyen LAT, Habók A. Tools for assessing teacher digital literacy: a review. J Comput Educ 2023 Jan 19;11(1):305-346. [doi: 10.1007/s40692-022-00257-5]
15. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2025-12-10]
16. Qoseem IO, Okesanya OJ, Olaleke NO, Ukoaka BM, Amisu BO, Ogaya JB, et al. Digital health and health equity: how digital health can address healthcare disparities and improve access to quality care in Africa. Health Promot Perspect 2024 Mar;14(1):3-8 [FREE Full text] [doi: 10.34172/hpp.42822] [Medline: 38623352]
17. Sylla B, Ismaila O, Diallo G. 25 years of digital health toward universal health coverage in low- and middle-income countries: rapid systematic review. J Med Internet Res 2025 May 29;27:e59042 [FREE Full text] [doi: 10.2196/59042] [Medline: 40440696]
18. GSMA Intelligence: The Mobile Economy 2022. URL: https://www.gsmaintelligence.com/research/the-mobile-economy-2022 [accessed 2024-06-18]
19. Howitt P, Darzi A, Yang GZ, Ashrafian H, Atun R, Barlow J, et al. Technologies for global health. Lancet 2012 Aug 04;380(9840):507-535. [doi: 10.1016/S0140-6736(12)61127-1] [Medline: 22857974]
20. Velazquez-Berumen A, Manimaran M. Driving innovation in low resource settings. World Hosp Health Serv 2016;52(3):7-11. [Medline: 30707806]
21. Perry L, Malkin R. Effectiveness of medical equipment donations to improve health systems: how much medical equipment is broken in the developing world? Med Biol Eng Comput 2011 Jul;49(7):719-722. [doi: 10.1007/s11517-011-0786-3] [Medline: 21597999]
22. Asma E, Heenan M, Banda G, Kirby RP, Mangwiro L, Acemyan CZ, Technical Collaborative Authorship Group. Avoid equipment graveyards: rigorous process to improve identification and procurement of effective, affordable, and usable newborn devices in low-resource hospital settings. BMC Pediatr 2023 Nov 15;23(Suppl 2):569 [FREE Full text] [doi: 10.1186/s12887-023-04362-x] [Medline: 37968578]
23. GSMA signs agreement with Africa Centres for Disease Control Prevention, to harness the power of mobile to combat disease in Africa. Africa Centres for Disease Control and Prevention. 2023. URL: https://africacdc.org/news-item/gsma-signs-agreement-with-africa-centres-for-disease-control-and-prevention-to-harness-the-power-of-mobile-to-combat-disease-in-africa/ [accessed 2024-06-20]
24. USAID's Power Africa launches partnership to electrify health facilities across sub-Saharan Africa as part of President Biden's global infrastructure initiative. Relief Web. 2022 Apr 28. URL: https://reliefweb.int/report/world/usaid-s-power-africa-launches-partnership-electrify-health-facilities-across-sub#:~:text=Today%2C%20as%20part%20of%20the,Africa%20over%20the%20next%20eight [accessed 2024-06-20]
25. Al Meslamani AZ. Technical and regulatory challenges of digital health implementation in developing countries. J Med Econ 2023;26(1):1057-1060 [FREE Full text] [doi: 10.1080/13696998.2023.2249757] [Medline: 37594521]
26. Joseph J. Algorithmic bias in public health AI: a silent threat to equity in low-resource settings. Front Public Health 2025;13:1643180 [FREE Full text] [doi: 10.3389/fpubh.2025.1643180] [Medline: 40771228]
27. National Institutes of Health. DS-I Africa. 2022. URL: https://dsi-africa.org/ [accessed 2024-06-21]
28. Masakhane. URL: https://www.masakhane.io/ [accessed 2024-06-21]
29. Zindi. URL: https://zindi.africa/ [accessed 2024-06-21]
30. Bediang G. Implementing clinical information systems in Sub-Saharan Africa: report and lessons learned from the MatLook project in Cameroon. JMIR Med Inform 2023 Oct 18;11:e48256 [FREE Full text] [doi: 10.2196/48256] [Medline: 37851502]
31. Kaburi BB, Harries M, Hauri AM, Kenu E, Wyss K, Silenou BC, et al. Availability of published evidence on coverage, cost components, and funding support for digitalisation of infectious disease surveillance in Africa, 2003-2022: a systematic review. BMC Public Health 2024 Jun 28;24(1):1731 [FREE Full text] [doi: 10.1186/s12889-024-19205-2] [Medline: 38943132]
32. Serge B, Mbondji E, Humphrey K, Janauschek L. Health data digitalization in Africa. In: AHOP Policy Briefs. Brazzaville: WHO Regional Office for Africa; 2024:1-47.

XSL•FO
RenderX

33. Mbunge E, Jack CL, Sibiya MN, Batani J. Review of implementation barriers and strategic approaches for improving mHealth systems utilization in Africa: Lessons learnt from South Africa and Kenya. Telematics and Informatics Reports 2025 Sep;19:100228. [doi: 10.1016/j.teler.2025.100228]
34. Irihamye E, Hadad J, Ali N, Holthof B, Wafula F, Paton C, et al. Sustainable by design: digital health business models for equitable global health impact in low-income and low-middle-income countries. Mayo Clin Proc Digit Health 2025 Dec;3(4):100261 [FREE Full text] [doi: 10.1016/j.mcpdig.2025.100261] [Medline: 41140345]
35. Nienaber McKay AG, Brand D, Botes M, Cengiz N, Swart M. The regulation of health data sharing in Africa: a comparative study. J Law Biosci 2024;11(1):lsad035 [FREE Full text] [doi: 10.1093/jlb/lsad035] [Medline: 38259628]
36. Torab-Miandoab A, Samad-Soltani T, Jodati A, Rezaei-Hachesu P. Interoperability of heterogeneous health information systems: a systematic literature review. BMC Med Inform Decis Mak 2023 Jan 24;23(1):18 [FREE Full text] [doi: 10.1186/s12911-023-02115-5] [Medline: 36694161]
37. Mamuye AL, Yilma TM, Abdulwahab A, Broomhead S, Zondo P, Kyeng M, et al. Health information exchange policy and standards for digital health systems in Africa: a systematic review. PLOS Digit Health 2022 Oct;1(10):e0000118 [FREE Full text] [doi: 10.1371/journal.pdig.0000118] [Medline: 36812615]
38. Victor Ibukun Adebayo, Adebimpe Bolatito Ige, Courage Idemudia, Osemeike Gloria Eyieyien. Ensuring compliance with regulatory and legal requirements through robust data governance structures. Open Access Res J Multidiscip Stud 2024 Jul 30;8(1):36-44. [doi: 10.53022/oarjms.2024.8.1.0043]

Abbreviations

dFVM: digital fit/viability model
mCFIR: modified version of the Consolidated Framework for Implementation Research
TRIMI: Train, Restructure, Incentivize, Mandate, Integrate
WHO: World Health Organization

Edited by T Leung, G Eysenbach; submitted 24.Jun.2024; peer-reviewed by K Fultz Hollis, T Neumark; comments to author 17.Mar.2025; revised version received 29.Sep.2025; accepted 23.Oct.2025; published 14.Jan.2026.

Please cite as:
Jiwa A, Ngatia A, Benali K, Boehmer N, Delle S, Emedom-Nnamdi P, Fofie CO, O'Brien CM, Olatunji T, Obayabgona K, Tambe M, Fletcher RR, Boatin AA, Hedt-Gauthier B
Africa's Digital Health Revolution: The Digital Fit-Viability Model to Move From Innovation to Scaled Implementation
J Med Internet Res 2026;28:e63495
URL: https://www.jmir.org/2026/1/e63495
doi:10.2196/63495
PMID:

©Afra Jiwa, Antony Ngatia, Karim Benali, Niclas Boehmer, Sangu Delle, Patrick Emedom-Nnamdi, Chris Opoku Fofie, Christine M O'Brien, Tobi Olatunji, Kate Obayabgona, Milind Tambe, Richard Ribon Fletcher, Adeline Adwoa Boatin, Bethany Hedt-Gauthier. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Wearable Air Samplers Reveal How Wildfire Shapes the Exposome

Virginia Gewin

(J Med Internet Res 2026;28:e93193) doi:10.2196/93193

KEYWORDS

wildfire smoke; exposome; air pollutants; environmental; wearable electronic devices; environmental monitoring; firefighters; combustion products; polycyclic aromatic hydrocarbons; particulate matter; cardiovascular diseases

Textbox .

Key Takeaways

- Wearable silicone-based devices are increasingly being used to measure chemical exposures, notably following wildfires.
- Indoor air exposures linger far longer than outdoor exposures.
- Researchers are working to correlate environmental exposures with health outcomes.

Just over 1 year ago, devastating fires raged in Los Angeles (LA), California, killing 29 people [1]. The blaze destroyed 16,000 structures—homes full of computers, plastic, and furniture—subjecting roughly 20 million people to toxic smoke and ash [1,2]. Winds helped clear the air significantly by mid-January, but researchers have found that hazardous compounds lingered indoors for weeks—possibly longer.

Tracking the Exposome

To collect a profile of the chemicals residents were exposed to following the LA wildfires, Krystal Pollitt, PhD, PEng—an associate professor at the Yale School of Public Health—deployed a wearable passive air sampler to 30 standing homes between February and April 2025 [3].

Pollitt's *Fresh Air Clip* is one of a number of wearable devices—typically made of silicone polymers—that can collect a wide range of organic chemicals from the air or skin. These devices are designed to offer insights into the "exposome"—a term first described 20 years ago [4] as the environmental counterpart to the genome. Exposome research aims to quantify how environmental exposures translate into health outcomes. Wearables are increasingly being deployed after natural hazards to help researchers document chemical exposures from air pollution, as well as water contaminants, occupational hazards, or pathogens.

Ongoing Exposure

Pollitt's rapid response study found that once these chemicals permeate the home, residents can continue to be exposed as they slowly release from furniture, carpets, and even drywall [3].

Combustion byproducts were detected in all of the homes tested 2 to 3 months after the fire; even homes 12 km away from the burn zone had byproduct levels only 1.2 times lower than those in homes closer to the burn zone. Further, levels of the vast majority of combustion-related chemicals, such as flame retardants, petroleum products, plastics, and tire particulates, were elevated indoors compared to outdoors.

Pollitt explains that hundreds of chemicals that settled indoors stayed put, and they will continue to be released off surfaces slowly over time. She uses a nontargeted gas chromatography–high-resolution mass spectrometry approach able to detect volatile and semivolatile organic chemicals, including those not necessarily expected to be collected by silicone. "There's so many obscure products that were burning, and were transformed through the pyrolysis process, that to come in with a targeted method would be like looking for a pin in a haystack," she says.

In addition to obscure products, Pollitt found a wide range of building materials and household products that released numerous chemicals that have not been well characterized, especially following combustion. "There's tons of plastics present, but we don't know which specific plastic types yet," she says.

"This was a first pass study, done in collaboration with colleagues at University of California in Los Angeles, that required immediacy to understand how acute exposures evolved," says Pollitt. She and colleagues are now analyzing hundreds of different soft materials from the fires, including carpet swatches, blankets, pillows, and stuffed animals, to determine which remediation efforts would work best to reduce long-term exposures.

It has become clear that wearables will be a key component of exposomics research going forward. "Especially with the fires, people have realized there's a huge need to be able to do personalized exposure monitoring," says Mary Margaret Johnson, MD, PhD—principal research scientist at Harvard T.H. Chan School of Public Health—who coauthored a review paper of exposomics methodologies and future directions [5].

XSL•FO
RenderX

"We're realizing more and more that what is in the smoke needs to be accounted for," says Johnson. "There is a realization that the toxicity of the smoke depends on what is being burned in the home or area. Metals in the smoke vary from fire to fire," she says.

Measuring Health Impact

Other researchers are conducting similar studies using silicone wristbands—even broaches—that can typically be worn for a few hours to 1 week to collect exposome samples. Heather Stapleton, PhD—distinguished professor of Environmental Chemistry & Exposure Science at Duke University in Durham, North Carolina—has conducted numerous studies using silicone wristbands in diverse communities. One group she has focused on is structural and wildland firefighters, given their increased risk of cancer and other health conditions compared to the general population.

Stapleton's preliminary research found that, compared to structural firefighters, wildland firefighters appear to have higher exposure to combustion byproducts and some brominated flame retardants.

As Stapleton's previous research has shown [6], the risks aren't just from the fires either. Contaminated firefighting equipment can also expose workers to a mixture of different chemicals. As well, they can be exposed to per- and polyfluoroalkyl substances (PFASs) used to extinguish the fire or treat their gear. Stapleton found that blood PFAS levels were positively correlated with years working as a firefighter.

Stapleton and Pollitt agree that the utility of wearable exposure capture devices is still in the early days, but they look very promising. "I am collaborating with a colleague to modify the wristband's ability to absorb and detect metals," Stapleton says.

Other researchers are pairing wearable air quality sensors with Apple Watch data to assess long-term impacts of wildfires on cardiovascular health. One recent study [7] found that, for each 10-$\mu g/m^3$ increase in particulate matter, firefighters' resting heart rates rose by an estimated 0.38 beats per minute before the 2024 Park Fire in California and by approximately 1.4 beats per minute following the fire.

Stapleton says that, while likely too cost-prohibitive to replicate on a large scale, this study is a first step toward the larger questions that researchers must now tackle, like determining how wristband results correlate with chemical levels in the body. "We know how to take blood or urine measurements, and put those into a risk assessment framework," says Stapleton. "We aren't quite there yet with wristbands." Stapleton says that the next steps include creating a machine learning model to predict what wristband exposures mean for internal doses. "I think that's key to making sure this tool is successful and can be used more broadly," says Stapleton.

Beyond Acute Exposure

Johnson and her colleague—Harvard physician-scientist Kari Nadeau, MD, PhD—are conducting several experiments to pair what is found in blood with external exposure measures, including those from wearable wristbands. "In our lab, we are trying to use blood as the exposome window," she says. "We also heavily focus on multi-omics approach, looking at inflammatory markers being produced, or on epigenetic changes, by looking at immune functioning, and associating them with various environmental exposure levels simultaneously," says Johnson.

"Since the silicone wristbands are relatively cheap, researchers are trying to incorporate with the general population to estimate different chemical exposures," she says.

Pollitt, too, is now collecting biological samples to look beyond acute exposures, as part of the LA Fire Health Consortium—a decade-long study of fire impacts. "Now we're starting to understand persistence," she says. Pollitt was one of many scientists presenting their research at the LA Fire Health Consortium conference in January to mark the 1-year anniversary of the fire. The saddest thing from the meeting, she described, was the realization of "what does this mean for the next fire? It will happen again—hopefully not on that scale."

In the meantime, researchers' continued adoption of silicone-based wearable devices will provide a unique lens into the exposome and how it is altered by growing climate change–related wildfire risk.

Conflicts of Interest

None declared.

References

1. California governor requests $40 billion for fire aid, Washington Post reports. Reuters. 2025 Feb 22. URL: https://www.reuters.com/world/us/california-governor-requests-40-billion-fire-aid-washington-post-reports-2025-02-22 [accessed 2026-02-20]
2. Withers A, Chmielewski D, Hay A, Lang H. Los Angeles wildfires trigger air quality warnings and health concerns. Reuters. 2025 Jan 10. URL: https://www.reuters.com/world/us/los-angeles-wildfires-trigger-air-quality-warnings-health-concerns-2025-01-10 [accessed 2026-02-20]
3. Lin EZ, Esenther S, Mascelloni M, Irfan F, Pollitt KJG. The Fresh Air wristband: a wearable air pollutant sampler. Environ Sci Technol Lett 2020 Jan 28;7(5):308-314. [doi: 10.1021/acs.estlett.9b00800]

XSL•FO
RenderX

4. Wild CP. Complementing the genome with an “exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomarkers Prev 2005 Aug;14(8):1847-1850. [doi: 10.1158/1055-9965.EPI-05-0456] [Medline: 16103423]
5. Wan M, Simonin EM, Johnson MM, et al. Exposomics: a review of methodologies, applications, and future directions in molecular medicine. EMBO Mol Med 2025 Apr;17(4):599-608. [doi: 10.1038/s44321-025-00191-w] [Medline: 39870881]
6. Hoxie T, Zhang S, Cooper EM, et al. Assessing nondietary exposure to per- and polyfluoroalkyl substances (PFASs) in firefighters using silicone wristbands. Environ Health (Wash) 2025 Dec 3. [doi: 10.1021/envhealth.5c00316]
7. Cheong SM, Perez M, Jung YS, Kim B, Nadeau K. Latent effects of wildfire particulate matter on resting heart rate measured with wearable devices. Circ Arrhythm Electrophysiol 2025 Aug;18(8):e013836. [doi: 10.1161/CIRCEP.125.013836] [Medline: 40709472]

Edited by KA Clegg; submitted 09.Feb.2026; this is a non–peer-reviewed article;accepted 09.Feb.2026; published 26.Feb.2026.

Please cite as:
Gewin V
Wearable Air Samplers Reveal How Wildfire Shapes the Exposome
J Med Internet Res 2026;28:e93193
URL: https://www.jmir.org/2026/1/e93193
doi:10.2196/93193

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Feb.2026.

XSL•FO
RenderX

Influencing the Influencers: How Health Experts Are Partnering With Content Creators to Fight Misinformation Online

Wendy Glauser

(J Med Internet Res 2026;28:e93450) doi:10.2196/93450

KEYWORDS

social media; opinion leaders; health communication; information dissemination; misinformation; public health; health promotion

Textbox .

Key Takeaways

- By offering training and support to content creators, public health advocates can improve the quality of health information on social media platforms.
- Partnering with content creators with diverse audiences can help health communicators reach populations they're struggling to engage.

Scroll. "Five essential questions that you must ask your surgeon prior to any operation."

Click. "Yesterday, I posted a picture of when I started on Ozempic and four weeks after Ozempic and many of you said...."

Swipe. "The World Health Organization lists processed meats like bacon, ham, salami and sausages in the same category as cigarettes and asbestos. But what does that even mean?"

These are three real examples [1-3] of the millions or potentially billions of TikTok videos that contain health information. In a Kaiser Family Foundation survey of over 1000 adults who were selected to be representative of the US population, 55% said they use social media to find health information at least occasionally [4].

Social media creators can influence what people eat, how they exercise, what medications they take or don't take, whether they seek out immunizations and preventive health screenings, and the list goes on. Whether that influence helps or harms depends in part on the accuracy of the information shared, and that information is not always accurate. For example, in a study that analyzed 1000 TikTok videos on mental health, 6.3% contained disinformation, which was defined as providing false or scientifically unsupported information, and 15.7% contained misinformation, which was defined as providing only partially supported information [5].

Realizing the potential impact, public health and health communication experts have begun to partner with content creators to fight misinformation online. As Matthew Motta, PhD—associate professor of Health Law, Policy and Management at the Boston University School of Public Health—explains, "If you can partner with just a handful of social media content creators, you might be able to reach millions of people with your message."

Collaboration Between Health Experts and Influencers

Three years ago, Amanda Yarnell, MS—senior director of the Harvard T.H. Chan School of Public Health's Center for Health Communication—launched the Creator Program [6] to help social media creators "counter health misinformation and spread evidence-based science to their communities," as the program's website describes. Yarnell explains that, at the time, "there was very little information, even about the health information creators shared." In the past few years, more research evaluating patterns in the content, style, and tone of health information shared on social media has emerged, but there are still very few health researchers attempting to intervene in those conversations by engaging creators.

While other programs have popped up in recent years that support health providers and scientists who are also creators, such as the World Health Organization's Fides program [7] or the Evidence Collective [8], the Creator Program is unique in that it also collaborates with influencers who don't necessarily have health or science degrees.

The impact of the Creator Program's initiatives is being studied. In one 2024 study [9], 105 English-language creators, who collectively reached about 17 million followers, were randomly assigned to a control group, which received asynchronous digital tool kits summarizing evidence-based mental health information, or a group that received the tool kits plus optional virtual training sessions. Researchers analyzed 3465 videos posted before and after the intervention. Post intervention, 26% of the creators who received tool kits and training invites, versus 21% of the control group, referenced the evidence-based themes included in the toolkit. The authors wrote that "although these effects are relatively small in substantive size, their reach," given the size of their audiences,"may be comparatively larger."

In another study (currently a preprint), Motta and colleagues [10] engaged social media influencer Kenzie Brenna to create two videos on how to respond when a peer is experiencing

XSL•FO
RenderX

emotional distress; one was created before a virtual training program delivered by mental health experts, and one was created afterward. A demographically representative sample of 1000 US youths between the ages of 14 and 22 years were then randomly assigned to view either video. The effect was modest but significant: 80% of those who viewed the posttraining video described it as informative (vs 60% of those who viewed the pretraining video), and more people in the group who viewed the posttraining video said that they felt confident in providing emotional support to a distressed peer.

Sasha Hamdani, MD—a content creator and psychiatrist—took part in the Creator Program virtual training sessions and also received a toolkit. Being part of the Creator Program, she says, "reinforced that even the most accurate data can be misunderstood if it is not communicated clearly and responsibly." Hamdani, who has about a million followers on TikTok and over 930,000 followers on Instagram, shares videos on mental health, especially attention-deficit/hyperactivity disorder (ADHD). She says being part of the Creator Program reminded her that "good health communication reduces misinformation by being precise, transparent about limitations, and intentional about language, especially in digital spaces where content travels quickly."

Reaching More People, More Often

Engaging with social media influencers allows public health experts to reach more people and "target diverse folks that public health organizations have failed at reaching, such as young people, as well as Black, Brown, Indigenous, and LGBTQ people," says Yarnell. Many of the influencers in the Creator Program, for example, have audiences that include a large proportion of people from one or more of these groups.

When it comes to prebunking (preparing people to resist misleading claims before they're exposed to them) and debunking (countering falsehoods after exposure), social media is the tool that "has the most promise currently," says Shaon Lahiri, PhD—assistant professor of public health at the College of Charleston in South Carolina. "It's where the misinformation is coming from," he explains. Plus, by virtue of engaging with audiences daily, creators can share similar information multiple times in different ways, or as Lahiri describes it, they can create "booster shots" to better inoculate people against misinformation.

More Engagement, Less Misinformation

Popular content creators tend to use methods that are engaging and impactful, as they rely on their metrics to learn what resonates. People are more persuaded by messaging that is "fun and not very heavy-handed," Lahiri explains. The technology capabilities in the social media platforms themselves also lend well to misinformation debunking/prebunking. For example, "you can do a Stitch on TikTok where you briefly show the misinformation and then you explain what's wrong with it," Lahiri explains, referring to taking a short clip from someone else's post and adding it to the beginning of your own.

Motta recognizes that these programs compete with billion-dollar companies and "social media companies that have really dropped the ball" in reducing misinformation. Still, he says, "there are plenty of content creators who want to fight against the spread of misinformation on social media platforms. These same creators are also really good at figuring out how to make sure that information permeates through."

Motta adds that even though social media videos are too often inaccurate, the vast majority of creators aren't intentionally sharing misleading information. As misinformation continues to travel faster than evidence, engaging the people who already command attention online may be the best way for public health to keep up.

Conflicts of Interest

None declared.

References

1. @drjeremylondon. Top 5 questions to ask your doctor before having surgery. TikTok. 2024 Feb 26. URL: https://www.tiktok.com/@drjeremylondon/video/7339983405460802859 [accessed 2026-02-19]
2. @lydiawanjiru.ke. Before and after one month Ozempic. TikTok. 2025 Feb 18. URL: https://www.tiktok.com/@lydiawanjiru.ke/video/7472688092751695125 [accessed 2026-02-19]
3. @drsermedmezher. Bacon and cancer risk: what you need to know. TikTok. 2026 Jan 1. URL: https://www.tiktok.com/@drsermedmezher/video/7590527954610949378 [accessed 2026-02-19]
4. Schumacher S, Sparks G, Montalvo JIII, Kirzinger A, Hamel L. KFF tracking poll on health information and trust: health information and advice on social media. KFF. 2025 Aug 7. URL: https://www.kff.org/public-opinion/kff-health-information-and-trust-tracking-poll-health-information-and-advice-on-social-media/ [accessed 2026-02-19]
5. Hudon A, Perry K, Plate AS, et al. Navigating the maze of social media disinformation on psychiatric illness and charting paths to reliable information for mental health professionals: observational study of TikTok videos. J Med Internet Res 2025 Jun 18;27:e64225. [doi: 10.2196/64225] [Medline: 40532184]
6. Creator Program | Center for Health Communication. Harvard T.H. Chan School of Public Health. URL: https://hsph.harvard.edu/research/health-communication/creator-program/ [accessed 2026-02-19]

XSL•FO
RenderX

7. Fides - a network of healthcare influencers dedicated to raising good health content and fighting misinformation. World Health Organization. URL: https://www.who.int/teams/digital-health-and-innovation/digital-channels/fides [accessed 2026-02-19]
8. The Evidence Collective. URL: https://www.evicollective.org/ [accessed 2026-02-19]
9. Motta M, Liu Y, Yarnell A. “Influencing the influencers:” a field experimental approach to promoting effective mental health communication on TikTok. Sci Rep 2024 Mar 11;14(1):5864. [doi: 10.1038/s41598-024-56578-1] [Medline: 38467750]
10. Motta M, Liu Y, Brenna K, et al. Assessing the effect of evidence-based mental health social media videos on youth emotional support competencies: a “creator-engaged” approach. Open Science Framework. Preprint posted online on Dec 5, 2024 URL: https://osf.io/preprints/socarxiv/kydf4_v1 [accessed 2026-02-19] [doi: 10.31235/osf.io/kydf4]

Edited by KA Clegg; submitted 12.Feb.2026; this is a non–peer-reviewed article;accepted 12.Feb.2026; published 27.Feb.2026.

Please cite as:
Glauser W
Influencing the Influencers: How Health Experts Are Partnering With Content Creators to Fight Misinformation Online
J Med Internet Res 2026;28:e93450
URL: https://www.jmir.org/2026/1/e93450
doi:10.2196/93450

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Is Going Analog Good for Children and Teens' Mental Health and Well-Being?

Sara Novak

(J Med Internet Res 2026;28:e94018) doi:10.2196/94018

KEYWORDS

adolescent; child; mental health; well-being; smartphones; social media; digital technology; screen time; health behaviour; parental guidance; media use; sleep; anxiety; depression; health promotion

Textbox .

Key Takeaways

- More and more parents are waiting to give their children a smartphone.
- Children and teens who spend more time on devices are more likely to develop addictive behaviors, have difficulty sleeping, and experience cyberbullying.
- Parents' smartphone use plays a significant role in how their children and teens engage with their devices.

As children grow older, parents face a mounting sense of pressure to provide them with smartphones. By the time they enter middle school, children who have devices far outweigh those who don't; according to Common Sense Media, 42% have a phone by age 10, 71% by age 12, and 91% by age 14 [1].

Still, a growing body of evidence is motivating parents to hold off on smartphones until the later teenage years. In fact, more and more families are taking a long, hard look at their digital life and trading in their smartphones for flip phones and landlines.

Many Families Want A More Low-Tech Lifestyle

Parents like Paul Jones, a father of three, want to make sure that their children are both educated on device safety and aware of the potential downsides.

"Technology isn't going anywhere, if anything it's advancing so we need to educate children on technology and how to use it safely and properly," said Jones. That's why he started FlippGen [2], a youth-focused organization where young adults go into schools and provide perspective on how teens can positively balance device use.

Jones said that his children will not get devices until age 16, and instead he and a "pod" of friends and family have invested in landlines. According to a recent survey, Gen Z parents are holding off on smartphones, with 28% waiting until their children are between the ages of 15 and 17 to purchase them [3]. Jones' children have a landline provided through a company called Tin Can, a popular Wi-Fi landline for kids that doesn't have apps, texting, or games and only allows kids to speak to those on their approved contact list [4].

Additionally, some teens that already have smartphones have been trading them in for flip phones in order to remain reachable while reducing distractions and time they view as wasted on social media sites. The trend isn't new, but it appears to be persistent: one 2023 report noted an increase in US flip phone sales [5], and recent 2026 coverage highlights a growing interest in being "chronically offline" [6].

Does Age Matter When It Comes to Responsible Smartphone Use?

The research is mixed when it comes to the best age to give a child a smartphone. One study published in the journal *Pediatrics* found that children who had smartphones by age 12 were more likely to suffer from an increased risk of depression, obesity, and insufficient sleep [7], but another study published in the journal *Child Development* showed that grades, sleep habits, and depression symptoms were not impacted by the age that children acquired their first devices [8].

Still, even with these mixed results, current data may still point to holding off—or at the very least, educating on and carefully monitoring device use. More than age, what seems to matter for well-being is how and how much children and teens use their smartphones.

Moderate, intentional, guided use may provide benefits for social connection or support, depending on the kind of use [9,10]—for example, if a child or teen shares about a positive event, achievement, or cause that they're involved in, or finds like-minded peers going through similar trials. Additionally, there may be a level of perceived social isolation, disconnection, and fear of missing out that occurs for some whose parents hold off on getting them smartphones, as well as perceived trade-offs in the other potential practical and logistical benefits smartphones might offer [11-14].

But on the other hand, excessive use without supervision can contribute to loss of sleep, cyberbullying, and exposure to distressing or inappropriate content [15,16]. And even if their

XSL•FO
RenderX

well-being doesn't suffer and they're able to keep their grades up, the average American child spends 5 hours daily on their smartphones [17].

"While social media can be used to forge some connections, the type of connections and socialization that occurs on social media is often quite different than the types of conversations and relationships that one would have face to face," said Jason Nagata, MD, an associate professor of Pediatrics in the Division of Adolescent and Young Adult Medicine at the University of California, San Francisco.

Potential Risks of Excessive Use

Social media addiction is much more likely to occur in children and teens who are allowed unfettered access to a smartphone. While not a formal diagnosis, it can present many of the same symptoms as substance abuse–related and other addiction disorders, including impaired control, withdrawal, tolerance, and relapse. It can contribute to diminished engagement in other activities and significantly impair a child's daily functioning, interpersonal relationships, and overall quality of life [17-20].

Sleep disruption is another potential negative for children and teens with unlimited access to their devices. A study published in the August 2023 edition of the journal *Sleep Health* [21] found that 63.2% of 11-12 year olds had a digital device in their room, and approximately 17% of them reported that it had awoken them in the past week. One-fifth of adolescents said that they used their device if they woke up during the night, which was associated with a higher risk of sleep disturbance.

Additionally, according to Nagata, the most likely explanation for social media's impact on mental health results from cyberbullying. A study published in the June 2025 issue of the journal *The Lancet Regional Health – Americas* found that children ages 11 to 12 years who were cyberbullied were more likely to have depressive symptoms and 2.62 times more likely to report suicidal ideation or a suicide attempt one year later [22].

Modeling Positive Digital Behavior

To mitigate some of the potential negative effects of smartphone use on children and teen's well-being, the American Academy of Pediatrics recommends establishing a healthy family media plan that the entire family can abide by [23]. Parents should establish screen-free zones, confiscating phones at a certain hour every night and not allowing them in the bedroom or during meals, recommendations that align with Nagata's research [24].

Perhaps most importantly, however, parents should model these behaviors for their children and teens. The single biggest predictor of a child's relationship with the digital world is not the age that parents provide them with a smartphone, but rather, a parent's own use of their devices, according to Michael Rich, MD, MPH, Director and Founder, Digital Wellness Lab, and associate professor of pediatrics at Harvard Medical School.

"We need to think of smartphones less in terms of entitlement and more in terms of the powerful tools that they are," said Rich. Powerful tools with instruction manuals the research is still writing—making prudent, thoughtful use essential, especially when the users are children and teens.

Conflicts of Interest

None declared.

References

1. Rideout V, Peebles A, Mann S, Robb MB. Common sense census: media use by tweens and teens, 2021. : Common Sense; 2022 URL: https://www.commonsensemedia.org/sites/default/files/research/report/8-18-census-integrated-report-final-web_0.pdf [accessed 2026-03-03]
2. About. FlippGen. 2025. URL: https://www.flippgen.com/about [accessed 2026-03-03]
3. Nesi J. A child's first cell phone: a guide for parents. Brown Health. 2022 Oct 20. URL: https://www.brownhealth.org/be-well/childs-first-cell-phone-guide-parents/ [accessed 2026-03-03]
4. Menu. Tin Can. 2026. URL: https://tincan.kids/products/tin-can [accessed 2026-03-03]
5. Mays L. Dumb phones are on the rise in the U.S. as Gen Z looks to limit screen time. CNBC. 2023 Mar 29. URL: https://www.cnbc.com/2023/03/29/dumb-phones-are-on-the-rise-in-the-us-as-gen-z-limits-screen-time.html [accessed 2026-03-03]
6. Bhaimiya S. A 'quiet revolution': why young people are swapping social media for lunch dates, vinyl records and brick phones. CNBC. URL: https://www.cnbc.com/2026/02/07/young-people-quiet-revolution-social-media.html [accessed 2026-03-03]
7. Barzilay R, Pimentel SD, Tran KT, Visoki E, Pagliaccio D, Auerbach RP. Smartphone ownership, age of smartphone acquisition, and health outcomes in early adolescence. Pediatrics 2026 Jan 1;157(1):e2025072941. [doi: 10.1542/peds.2025-072941] [Medline: 41324306]
8. Sun X, Haydel KF, Matheson D, Desai M, Robinson TN. Are mobile phone ownership and age of acquisition associated with child adjustment? A 5-year prospective study among low-income Latinx children. Child Dev 2023 Jan;94(1):303-314. [doi: 10.1111/cdev.13851] [Medline: 36056573]
9. Dharejo N, Alivi MA, Rahamad MS, et al. Effects of social media use on adolescent psychological well-being: a systematic literature review. Int J Interact Mob Technol 2023;17(20):171-191. [doi: 10.3991/ijim.v17i20.44663]

XSL•FO
RenderX

10. Yang CC, Hunhoff P, Lee Y, Abrell J. Social media activities with different content characteristics and adolescent mental health: cross-sectional survey study. J Med Internet Res 2025 Apr 28;27:e73098. [doi: 10.2196/73098] [Medline: 40294901]
11. Matthes J, Stevic A, Koban K, Thomas MF, Forrai M, Karsay K. Fear of missing out, reflective smartphone disengagement, and loneliness in late adolescents. Cyberpsychol Behav Soc Netw 2023 Oct;26(10):731-738. [doi: 10.1089/cyber.2023.0014] [Medline: 37582211]
12. Ophir Y, Rosenberg H, Lipshits-Braziler Y, Amichai-Hamburger Y. "Digital adolescence": the effects of smartphones and social networking technologies on adolescents' well-being. In: Online Peer Engagement in Adolescence 2020:122-139. [doi: 10.4324/9780429468360-8]
13. Coplan RJ, McVarnock A, Hipson WE, Bowker JC. Alone with my phone? Examining beliefs about solitude and technology use in adolescence. Int J Behav Dev 2022 Nov;46(6):481-489. [doi: 10.1177/01650254221113460] [Medline: 36397736]
14. Vaterlaus JM, Carney T, Kroemer-Spiess P. Delaying adolescent smartphone ownership: a qualitative dyadic study with parents and adolescents. J Child Media 2026:1-21. [doi: 10.1080/17482798.2026.2627286]
15. Martin JD, Song SW, Rote WR, et al. The life in media survey: a baseline study of digital media use and well-being among 11- to 13-year-olds. : Researchers at the University in South Florida in collaboration with The Harris Poll; 2025 URL: https://www.usf.edu/arts-sciences/departments/journalism/documents/lims2025/lims2025.pdf [accessed 2026-03-03]
16. Martín-Cárdaba M, Martínez Díaz MV, Lafuente Pérez P, García Castro J. Smartphone ownership, minors' well-being, and parental mediation strategies. An analysis in the context of social media influencers. J Youth Adolesc 2024 Oct;53(10):2202-2218. [doi: 10.1007/s10964-024-02013-7] [Medline: 38782845]
17. Media use and screen time – its impact on children, adolescents, and families. : American College of Pediatricians; 2020 May URL: https://acpeds.org/wp-content/uploads/2025/09/Media-Use-and-Screen-Time-May-2020.pdf [accessed 2026-03-03]
18. Nagata JM, Singh G, Sajjad OM, et al. Social epidemiology of early adolescent problematic screen use in the United States. Pediatr Res 2022 Nov;92(5):1443-1449. [doi: 10.1038/s41390-022-02176-8] [Medline: 35768491]
19. Wong SL, King N, Gariépy G, et al. Adolescent social media use and its association with relationships and connections: Canadian Health Behaviour in School-aged Children, 2017/2018. Health Rep 2022 Dec 21;33(12):14-23. [doi: 10.25318/82-003-x202201200002-eng] [Medline: 36542360]
20. Bozzola E, Spina G, Agostiniani R, et al. The use of social media in children and adolescents: scoping review on the potential risks. Int J Environ Res Public Health 2022 Aug 12;19(16):9960. [doi: 10.3390/ijerph19169960] [Medline: 36011593]
21. Nagata JM, Singh G, Yang JH, et al. Bedtime screen use behaviors and sleep outcomes: Findings from the Adolescent Brain Cognitive Development (ABCD) Study. Sleep Health 2023 Aug;9(4):497-502. [doi: 10.1016/j.sleh.2023.02.005] [Medline: 37098449]
22. Nagata JM, Shim J, Balasubramanian P, et al. Cyberbullying, mental health, and substance use experimentation among early adolescents: a prospective cohort study. Lancet Reg Health Am 2025 Jun;46:101002. [doi: 10.1016/j.lana.2025.101002] [Medline: 40625792]
23. How to make a family media plan. American Academy of Pediatrics. 2026. URL: https://www.healthychildren.org/English/family-life/Media/Pages/How-to-Make-a-Family-Media-Use-Plan.aspx [accessed 2026-03-03]
24. Nagata JM, Paul A, Yen F, et al. Associations between media parenting practices and early adolescent screen use. Pediatr Res 2025 Jan;97(1):403-410. [doi: 10.1038/s41390-024-03243-y] [Medline: 38834780]

Edited by KA Clegg; submitted 23.Feb.2026; this is a non–peer-reviewed article;accepted 23.Feb.2026; published 05.Mar.2026.

Please cite as:
Novak S
Is Going Analog Good for Children and Teens' Mental Health and Well-Being?
J Med Internet Res 2026;28:e94018
URL: https://www.jmir.org/2026/1/e94018
doi:10.2196/94018

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 5.Mar.2026.

XSL•FO
RenderX

Physician, Restore Thyself? The Digital Gap in Physician Well-Being Support

Jenny Castillo Cato

(J Med Internet Res 2026;28:e93338) doi:10.2196/93338

KEYWORDS

physicians; peer support; social support; mental health; physician burnout; well-being; occupational stress; mobile apps; digital health; psychological resilience

Textbox .

Key Takeaways

- Physician well-being is a long-standing, complex issue that needs to be better addressed.
- Current digital applications are helpful, but primarily focus on physician self-improvement.
- The digital world is a largely untapped medium for supporting equally critical factors: community, connection, and culture change.

Jenny Castillo Cato is board-certified in emergency medicine and has worked as an attending physician for over 20 years. She is an advocate for physician well-being and has worked on physician well-being initiatives in New York and across the United States.

In March 2020, the COVID-19 pandemic hit New York City like a tornado. The flood of patients coming to the emergency departments of the five boroughs was beyond imagination. Staff were getting sick, supplies were running out, and no treatment had been developed for this brand-new virus.

As physicians, we had to make hard decisions. Should I put the 80-year-old man who was barely responsive and hypoxic on the last ventilator in the hospital? Or should I keep the ventilator for the next patient who may be younger and more likely to withstand the onslaught of this deadly virus? Almost 6 years later, those memories still haunt me.

Unspoken Burden

Physicians have long been expected to rise to these kinds of occasions and to perform at a high level across many challenging circumstances, and we do. There is, however, an emotional toll on physician well-being, one that we are frequently expected to carry alone. This issue is not new. Though often hidden out of fear of discovery and shame, it has persisted throughout the centuries. In the 1860s, Dr James Still wrote in his memoirs that he often felt lonely, with “no one with whom to hold conference or to cheer or comfort me when depressed with difficult cases” [1]. This mentality has continued through the ages, was highlighted during the COVID-19 pandemic, and remains a pressing concern today.

Physicians’ mental health is a crucial aspect of well-being—ours and, ultimately, our patients’—yet the stigma surrounding mental health assistance has led to dire consequences. The American Medical Association reports that physicians are 82% more likely to experience burnout than those in other fields [2]. Furthermore, in the United States, between 300 and 400 physicians die by suicide each year, and the suicide rate for female physicians is 250% to 400% higher than for female workers in other professions [3]. Estimates vary and tend to be higher in the United States, but one study suggests that, across countries, physicians die by suicide, on average, 44% more often than expected compared to the general population [4]. An increase in risk factors (such as administrative burden and excessive job demands) and a decline in protective factors (such as autonomy and peer-based support) may play a role [5].

Foundations like the Dr. Lorna Breen Hero’s Foundation—named after my friend and colleague who died by suicide in April 2020—were created to advocate for change [6]. Nonetheless, progress remains slow and challenging due to factors such as physician personality traits, stigma within the medical community, and entrenched norms, including mandatory mental health questions on licensure applications that perpetuate the idea that mental illness equates to weakness or incompetence [7].

Persistent Gaps in Support

To support physician mental health and well-being, the medical community promotes self-care such as mindfulness, exercise, better sleep habits, and nutrition.

Many organizations, such as the American College of Physicians and the American Academy of Allergy, Asthma, and Immunology, offer useful well-being resources for this kind of self-improvement. The resources highlight applications such as Headspace, a meditation-based app; My Well-Being Index, an anonymous evaluation tool from the Mayo Clinic; Provider Resilience, a self-assessment tool for compassion fatigue and burnout; and T2 Mood Tracker, an app that monitors anxiety, depression, posttraumatic stress disorder, and general well-being [8,9]. There are also a few specialty-specific applications like the Surgeon’s Masters, which focus on physical wellness with

XSL•FO
RenderX

the 7-minute workout; Sleepo, a sleep tracker; and WaterMinder, a way to monitor hydration [10].

But while informative and helpful, these applications do not fully address the complex issue of physician well-being. The COVID-19 pandemic highlighted that well-being could not be addressed solely at the individual level but that significant changes must be made in clinical environments as well. Vast organizational norms need to be altered to transform physician well-being; it requires a cultural, political, and environmental shift.

There is also no one right answer for supporting physician well-being; it cannot be accomplished through a "one size fits all" approach. Each medical specialty has different challenges and needs. For example, what is important and impacts an orthopedic surgeon's daily work-life is very different from what matters to a family medicine physician seeing patients in a clinic. The physical stress, mental toll, and teamwork required during an 8-hour surgery affect the surgeon differently than a physician who manages 40 patients who are chronically ill every 15 minutes for 8 hours.

In short, persistent gaps remain in the resources available for physicians.

New Solutions

Like Dr Still noted in the 1860s, physicians need community, support, and camaraderie. This is particularly pressing as many medical institutions—especially in New York City, where space is a premium—have opted to use physicians' lounges for administrative purposes, leading to less direct physician contact across specialties.

A new app called Roon is trying to address this issue. Though Roon began as a video-based question and answer platform of health issues for patients, the final product created a "physicians-only knowledge network" providing a "digital doctor's lounge where wisdom of one becomes the resource of many" [11]. Roon and other programs like PeerRxMed, a free health care provider peer-to-peer program [12], are beginning to address the complex nature of physician well-being by facilitating community, peer support, and camaraderie. Beyond these tools, however, solutions in the digital space have yet to be fully tapped into.

A Call to Action

More resources are needed to support physician mental health and well-being, particularly resources that go beyond self-care and self-improvement.

New applications need to be created with a collaborative mindset. Clinicians must work with programmers and graphic designers to create content that is relevant and specific to each specialty. Combining directed self-care options, team building, leadership skills, and peer support into an easy-to-use, focused application would be incredibly valuable for the newer generations entering medicine.

The consequences of *not* developing these tools include and extend beyond physicians: without a healthy medical workforce, everyone would be affected. While we love our profession, physicians need resources to help us not only put on our own masks first but also support one another and share the weight so that we can more effectively care for patients.

The digital world remains an untapped medium for building physician community and culture change. Let us act now to harness its power to help physicians holistically cultivate better health, positive work-life integration, and institutional changes, creating an environment that allows longer, happier careers.

Conflicts of Interest

None declared.

References

1. Still J. Early Recollections and Life of Dr James Still: J.B. Lippincott and Company; 1877.
2. Berg S. Battling doctor burnout? New app makes a game of learning to beat it. American Medical Association. 2019 Aug 26. URL: https://www.ama-assn.org/practice-management/physician-health/battling-doctor-burnout-new-app-makes-game-learning-beat-it [accessed 2026-01-25]
3. Matheson J. Physician suicide. American College of Emergency Physicians. URL: https://www.acep.org/life-as-a-physician/wellness/wellness/wellness-week-articles/physician-suicide [accessed 2026-01-25]
4. Dutheil F, Aubert C, Pereira B, et al. Suicide among physicians and health-care workers: a systematic review and meta-analysis. PLoS ONE 2019;14(12):e0226361. [doi: 10.1371/journal.pone.0226361] [Medline: 31830138]
5. Harvey SB, Epstein RM, Glozier N, et al. Mental illness and suicide among physicians. Lancet 2021 Sep 4;398(10303):920-930. [doi: 10.1016/S0140-6736(21)01596-8] [Medline: 34481571]
6. Dr. Lorna Breen Heroes Foundation. URL: https://drlornabreen.org/ [accessed 2026-01-25]
7. Ng IK, Tan BC, Goo S, Al-Najjar Z. Mental health stigma in the medical profession: where do we go from here? Clin Med (Lond) 2024 Jan;24(1):100013. [doi: 10.1016/j.clinme.2024.100013] [Medline: 38382183]
8. Resources for physician wellness. American Academy of Allergy Asthma & Immunology. URL: https://www.aaaai.org/practice-management/practice-tools/wellness/resource [accessed 2026-01-25]

XSL•FO
RenderX

9. Individual physician wellness and burnout tools. American College of Physicians. URL: https://www.acponline.org/practice-career/physician-well-being-and-professional-fulfillment/individual-physician-wellness-and-burnout-tools [accessed 2026-01-25]
10. SurgeonMasters. URL: https://surgeonmasters.com [accessed 2026-01-25]
11. Roon. URL: https://www.roon.com [accessed 2026-01-25]
12. PeerRxMed. URL: https://www.peerrxmed.com [accessed 2026-01-25]

Edited by KA Clegg; submitted 11.Feb.2026; this is a non–peer-reviewed article;accepted 11.Feb.2026; published 27.Feb.2026.

Please cite as:
Cato JC
Physician, Restore Thyself? The Digital Gap in Physician Well-Being Support
J Med Internet Res 2026;28:e93338
URL: https://www.jmir.org/2026/1/e93338
doi:10.2196/93338

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Using AI to Assess Potential Zoonotic Threats

Cliff Dominy

(J Med Internet Res 2026;28:e93261) doi:10.2196/93261

KEYWORDS

influenza A virus; H5N1 subtype; avian influenza; zoonoses; cross-species transmission; humans; animals, wild; farm animals; machine learning; genomics; protein structure; predictive modeling; pandemics; emerging infectious diseases; viral evolution; bioinformatics

Textbox .

Key Takeaways

- Potential mutations in avian flu strains pose a credible future pandemic risk.
- A recent study developed a novel machine learning model to identify genomic features with potential for spillover to human hosts.
- The approach may have additional applications, including guiding the development of targeted influenza vaccines.

The next viral pandemic is an increasingly likely possibility in the coming decades [1], and while we have some guesses, ultimately we don't know exactly where or when to expect it [2]. What the world needs is a NATO (North Atlantic Treaty Organization)–style early warning radar to alert us to these incoming threats. Ideally, the radar must distinguish between the viral majority that pose no risk to human health and the very few that do. And one more thing, it should warn us of this potential pandemic pest before it spills over into humans.

Is that too much to ask? Recent research suggests it might not be.

A Potential Solution

A team of virologists and computational biologists led by Liam Brierley, PhD, while at the University of Liverpool (now at the University of Glasgow), may have come up with a solution to the problem of preparing for future pandemics. The group developed a machine learning (artificial intelligence [AI]) model capable of predicting which strains of avian influenza circulating in animals have the potential to jump into humans. Their research paper [3] is currently available as a preprint and is undergoing peer review—but the results at this early stage look promising.

Traditional phylogenetics is a retrospective tool, analyzing sequences from existing databases to flag new variants of potential pandemic concern [4,5]. In contrast, the Brierley et al [3] paper introduces a predictive biophysical information layer to detect viruses with host-jumping potential. The approach was designed to identify protein and nucleic acid sequences that share a functional similarity with existing zoonotic influenza strains. As Brierley explained via email, "The idea behind the machine learning approach is, if we can identify what makes a zoonotic virus from these 'first principles' of protein function and nucleotide motifs, we can apply this to highly divergent sequences that a phylogeny would struggle to make inference about, because it would be equally distant from all sequences we currently know." In other words, by considering factors like protein motif functionality, we might be able to identify signatures of host-jumping potential between distantly related viruses—something traditional phylogenetic analyses would miss.

AI on Defense

Avian flu has been around for centuries but is a relatively new pathogen in humans [6]. In 1997, authorities confirmed the first case of a virus carrying the signature H5 and N1 bird flu proteins in a human [7]. This new human-compatible virus has since spread rapidly through the world's bird population. So far, there have been numerous confirmed human fatalities but no proof of human-to-human transmission [8,9]. H5N1 has also impacted other mammals, decimating the sea lion and elephant seal populations in Peru in 2023 [10]. Farmworkers exposed to their livestock account for most of the over 600 human cases reported to date [11].

Brierley chose the H5N1 influenza virus as a use case for his analysis because it was a credible future pandemic threat but trained the model on all known subtypes of avian flu. He writes, "It's important to capture that whole diversity if we want to be able to make predictions for new strains in future."

An advantage for the number crunchers was the extensive avian influenza genetic database at their disposal. Currently, nearly 19,000 distinct viral sequences are available covering 120 different viral subtypes. What is more, 618 samples came from humans. The research team trained its AI model on a subset of the database and then tested it against the remaining data. They were seeking reliable and wide-scale genomic signals associated with human infection, combined with the physical and chemical properties of short protein regions identified by their model, and with the samples taken from human beings.

XSL•FO
RenderX

Key Findings

The results showed the model had a 91.9% chance of correctly identifying a virus at risk of spillover into humans. Notably, first-generation protein modeling software achieved this precision on a database that will likely grow and improve as surveillance programs add sequences from new outbreaks. While more data is needed, Brierley writes that, "The leaps in performance will come from more advanced models, which are not decades away but rather, years or even months."

While many regions of the influenza virus genome have similarities with each other, adding the protein motif predictions narrowed the spillover regions down to just a few key areas. These regions were small, often just two or three base pairs long, but mutations in these regions may be critical to allow a bird virus to propagate within human cells.

The AI model could cut through the noise generated from thousands of viral base pairs and highlight just a few protein motifs or patterns that mattered. They were:

- RNA polymerase complex: 9 motifs across the PA (polymerase acidic protein), PB1 (polymerase basic protein 1), and PB2 (polymerase basic protein 2) subunit genes, whose proteins are essential for the virus to replicate, were top of the list.
- Virus binding: 1 motif within the hemagglutinin (HA) gene was included. The "H" in the H5N1 virus family is involved in binding to the host cells. Mutations here could allow the virus to recognize and infect a new host.
- Replication: Motifs of the nucleoprotein (NP), a chaperone responsible for ensuring the host replicates the viral genes correctly and packages them for export, were included.
- Immune evasion: Motifs within a nonstructural gene (NS1), the gene involved in dampening the host cell's immune response after infection, were also included

The model has already shown potential for use with other influenza viruses. It flagged rare influenza viruses like H10N8, which has been detected in humans [12], as well as the H4 subtypes, which have not. Brierley notes that he can "highlight a couple of H4 sequences from H4N8 and H4N6 subtypes that seemed to have elevated zoonotic potential, and the next step would be laboratory studies to understand the exact mechanisms."

Next Steps

Future priorities for this approach would include improved modeling and expanding the database, especially with viruses that cause asymptomatic infections that could more easily spread undetected. Global health authorities can assist by monitoring for new emerging viruses and adding their genetic sequences to the database before they become more virulent.

AI will become an important tool in the public health armory. But despite the benefits reported in the study, machine learning cannot offer a complete solution. Brierley noted limitations of the approach using AI, noting that "it can't tell you exactly WHY a certain feature matters but it can tell you where we should start looking in deeper study," adding that "it can only do the job it's trained for. We can't, for example, make any predictions about spillover into other mammals from this model." Furthermore, AI can only analyze samples taken at a single point in time; it cannot gauge the direction of viral spread. And the influenza virus is subject to modification by its hosts, which may affect its virulence and transmissibility—for example, the addition of sugar and phosphate molecules on the mature virus can alter its pathogenicity, a feature that AI could miss.

AI on the Offense

This technology is not just an effective defense against future viral threats; it could also be used proactively during influenza season. The annual flu shot would benefit from the superior predictive power of AI over the traditional genetic approach to developing targeted vaccines. Beyond influenza, it could be readily adapted to other respiratory pathogens, including coronaviruses and the common cold.

In Brierley's words, "One thing is clear: as computing power grows and our model complexity starts to better capture biological realities," these models could become central in understanding and monitoring viruses with zoonotic potential, helping us anticipate and potentially mitigate the impact of the next pandemic.

Conflicts of Interest

None declared.

References

1. Marani M, Katul GG, Pan WK, Parolari AJ. Intensity and frequency of extreme novel epidemics. Proc Natl Acad Sci U S A 2021 Aug 31;118(35):e2105482118. [doi: 10.1073/pnas.2105482118] [Medline: 34426498]
2. Gomez LRE, Malysheva N, Pfeil J, et al. Holistic forecasting for future pandemics: a review of pathogens, models, and data. Discov Public Health 2025;22(1):1-17. [doi: 10.1186/s12982-025-00573-y]
3. Brierley L, Mould-Quevedo J, Baylis M. An AI for an AI: identifying zoonotic potential of avian influenza viruses via genomic machine learning. bioRxiv. Preprint posted online on Sep 16, 2025. [doi: 10.1101/2025.09.16.676011]
4. Scarafoni D, Telfer BA, Ricke DO, Thornton JR, Comolli J. Predicting influenza A tropism with end-to-end learning of deep networks. Health Secur 2019;17(6):468-476. [doi: 10.1089/hs.2019.0055] [Medline: 31859569]

XSL•FO
RenderX

5. Borkenhagen LK, Allen MW, Runstadler JA. Influenza virus genotype to phenotype predictions through machine learning: a systematic review: computational prediction of influenza phenotype. Emerg Microbes Infect 2021;10(1):1896-1907. [doi: 10.1080/22221751.2021.1978824] [Medline: 34498543]
6. Kim SM, Kim YI, Pascua PNQ, Choi YK. Avian influenza A viruses: evolution and zoonotic infection. Semin Respir Crit Care Med 2016 Aug;37(4):501-511. [doi: 10.1055/s-0036-1584953] [Medline: 27486732]
7. Isolation of avian influenza A(H5N1) viruses from humans—Hong Kong, May-December 1997. JAMA 1998 Jan 28;279(4):263. [doi: 10.1001/jama.279.4.263]
8. MacCosham A, Vasiliu AG, Atchessi N. A rapid review of the avian influenza PB2 E627K mutation in human infection studies. Can Commun Dis Rep 2025 Apr;51(4):137-144. [doi: 10.14745/ccdr.v51i04a04] [Medline: 40241713]
9. Technical report: June 2024 highly pathogenic avian influenza A (H5N1) viruses. Centers for Disease Control and Prevention. 2024 Jun 5. URL: https://www.cdc.gov/bird-flu/php/technical-report/h5n1-06052024.html [accessed 2026-02-24]
10. Leguia M, Garcia-Glaessner A, Muñoz-Saavedra B, et al. Highly pathogenic avian influenza A (H5N1) in marine mammals and seabirds in Peru. Nat Commun 2023 Sep 7;14(1):5489. [doi: 10.1038/s41467-023-41182-0] [Medline: 37679333]
11. Influenza (avian and other zoonotic). World Health Organization. 2026 Jan 21. URL: https://www.who.int/news-room/fact-sheets/detail/influenza-(avian-and-other-zoonotic) [accessed 2026-02-24]
12. Chen H, Yuan H, Gao R, et al. Clinical and epidemiological characteristics of a fatal case of avian influenza A H10N8 virus infection: a descriptive study. Lancet 2014 Feb 22;383(9918):714-721. [doi: 10.1016/S0140-6736(14)60111-2] [Medline: 24507376]

Edited by KA Clegg; submitted 10.Feb.2026; this is a non–peer-reviewed article;accepted 10.Feb.2026; published 03.Mar.2026.

Please cite as:
Dominy C
Using AI to Assess Potential Zoonotic Threats
J Med Internet Res 2026;28:e93261
URL: https://www.jmir.org/2026/1/e93261
doi:10.2196/93261

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 3.Mar.2026.

XSL•FO
RenderX

As Social Media Scales Back Fact-Checking, Can Technologies Fill the Gap?

Wendy Glauser

(J Med Internet Res 2026;28:e95730) doi:10.2196/95730

KEYWORDS

misinformation; social media; artificial intelligence; fact-checking; algorithms; public health

Textbox 1.

Key Takeaways

- Algorithms, artificial intelligence fact-checking, and social media ads can reduce the spread of misinformation at scale.
- Low-cost "content-neutral" interventions reminding users to think before sharing can help prevent misinformation spread.

Part one of this series [1] *showed how researchers are working with social media influencers to boost accurate health information online. In part two, we explore technological solutions for detecting and combating misinformation.*

Misinformation is increasingly spread with single clicks, bots, and artificial intelligence (AI) deepfakes. AI-generated images and videos share fake treatments, with even deepfake versions of renowned doctors' likenesses used to gain credibility [2]. In an age where generative AI is increasing the volume and speed of health misinformation [3] and agencies like the World Health Organization are raising alarms about the impact on vaccine trust and public health [4], are AI and algorithm-based technologies for combating that misinformation keeping up?

While evidence suggests technological solutions to misinformation on social media are effective, researchers worry that social media companies' interest in employing, evaluating, and improving these tools has waned in recent years.

Common technologies for combating misinformation include everything from algorithmic labeling of posts that contain misinformation, to downregulation of AI-deemed inaccurate posts, to mass awareness campaigns that encourage critical thinking [5-7].

Cameron Martel, PhD—assistant professor of marketing at the Johns Hopkins Carey Business School—explains that in the late 2010s and early 2020s, major social media companies, including Facebook and Twitter, employed algorithms to identify potentially false articles and engaged third-party fact-checkers to verify posts.

In 2023, he led a large study of warning labels, in which over 14,000 participants in the United States were exposed to true and false headlines and asked about their belief in the headlines or interest in sharing them [8]. Half of the participants were exposed to warning labels when presented with false information, while half were not.

Fact-checking labels reduced belief in false information by nearly 28% and reduced misinformation sharing by roughly 25% relative to the control group. The study also showed that among those with low trust in fact-checkers, warning labels nonetheless reduced misinformation sharing by more than 16%.

In January 2025, however, Meta announced it would end its partnership with third-party fact-checkers and instead adopt community notes, whereby everyday users comment on the accuracy of information [9]. If comments are upvoted by people from across the political spectrum, then they'll appear prominently.

Such community notes are likely to be trusted if the process behind community note generation is transparent and reasonable, Martel says. In a study published last year, Martel and colleagues [10] found that while both Democrat- and Republican-leaning participants preferred expert fact-checkers over laypeople, laypeople "juries" could be deemed equally trustworthy as or more trustworthy than experts if their size was large enough, they had consulted with each other, and they had equal representation across political groups.

The Rise of AI Fact-Checking

There is far less information about how the public views AI fact-checking tools and their accuracy. A study (available as a preprint [11]) suggests that the large language models (LLMs) Perplexity and Grok largely align with community note decisions about posts that are misleading. However, 21% to 28% of posts that community notes deemed as misleading were deemed true by the AI bots.

Concerningly, the authors observe that the launch of the Grok bot on X in early March 2025 co-occurred with a substantial reduction in the community note submissions, suggesting that social media users may see AI as an alternative, rather than as a complement, to democratized fact-checks.

While Martel points out that AI can be very helpful for identifying and responding to "well debunked conspiracy theories or often repeated myths," the limit of AI fact-checking has become glaring during breaking news events. Al Jazeera reported, for example, that Grok struggled to recognize AI-generated media in conflict situations and incorrectly said

XSL•FO
RenderX

that a trans pilot was responsible for a helicopter crash, among many other breaking news fact-checking errors [12].

"Large language models don't have any existing corpus of information about what's happening currently," explains Martel. Yet, "there's at least anecdotal evidence that people are still trying to use LLMs to find out information about unfolding events, and that is troubling."

Martel says that ultimately, democratized fact-checking through community notes, AI fact-checking, and professional fact-checkers "have great promise" when used in tandem. For example, AI systems could refer breaking news or claims that they can't easily verify to human fact-checkers, social media users could rate the accuracy of information fact-checked by AI, and AI and algorithm-based systems could respond to real-time feedback from democratized fact-checks.

But fact-checking systems should be transparent, continually audited, assessed for effectiveness, and improved. And that's not happening. "Right now, it seems like there is no corporate will to invest heavily in these types of content moderation practices, so while I'm theoretically hopeful about these technologies, in practice, I'm less hopeful," says Martel.

"Content-Neutral" Interventions Can Promote Critical Thinking on Social Media

Interventions that are "content neutral" are another scalable solution to reducing misinformation, says Hause Lin, PhD—a researcher at the Massachusetts Institute of Technology and Cornell University, and a data scientist at the World Bank. "People are going to be producing all kinds of weird content that you just will not be able to anticipate," he explains, but interventions that encourage critical thinking and help people spot common propaganda tactics can blunt the influence of misinformation.

In 2023, Lin and colleagues [7] assessed the effectiveness of Facebook and Twitter ads that encouraged people to consider the accuracy of information before they shared it. The Facebook study, which involved 33 million users, found that these accuracy prompts led to a 2.6% reduction in misinformation sharing among users who had previously shared misinformation (as flagged by third-party fact-checkers or Facebook's internal system). The Twitter study, which relied on data from over 157,000 users, showed that accuracy prompts resulted in an up to 6.3% reduction in misinformation sharing among users who saw at least one ad and had previously shared misinformation.

The magnitude of the effect could be much higher with different types of accuracy prompts that are designed to reach more people over longer periods of time, Lin says. (The Facebook study only assessed user behavior for an hour after the ad was shared, while the Twitter study evaluated user behavior over days to weeks, Lin explains.) Regardless, 6% of millions is a significant impact for a relatively "low-cost" intervention.

The goal of the project was to jolt social media users from an emotional state to a reflective state, Lin says. "When people are scrolling, they are often not thinking reflectively but intuitively. They're thinking 'This gets me worked up so I'm going to share it with the world,'" he explains. "If you slow them down just a little bit, and say, 'Do you want to think more about whether this is true?' that actually reduces misinformation."

Still, Lin acknowledges that large-scale content moderation may not align with the profit motive. For example, Lin recently studied the effect of "prosocial" celebrity messages aimed at countering ethnic hate–driven rhetoric on social media in Nigeria. A preprint of the study [13] suggests that people who saw the videos were less likely to share hate content but also more likely to reduce the time they spent on Twitter overall. "The side effects of interventions like this can be unpredictable," Lin says.

There is growing evidence that multipronged efforts can help counter health and other misinformation, and even small efforts can make an impact. Whether social media companies are willing to invest in these initiatives for the broader social good remains to be seen.

Conflicts of Interest

None declared.

References

1. Glauser W. Influencing the influencers: how health experts are partnering with content creators to fight misinformation online. J Med Internet Res 2026 Feb 27;28:e93450. [doi: 10.2196/93450] [Medline: 41773685]
2. What Dr. Sanjay Gupta learned from being the target of a deepfake health ad - Terms of Service with Clare Duffy. CNN Podcasts. 2025 Sep 23. URL: https://edition.cnn.com/audio/podcasts/terms-of-service-with-clare-duffy/episodes/56d4d6b8-25e7-11f0-a31f-a7da5a03d2d1 [accessed 2026-03-26]
3. Saeidnia HR, Jahani S, Ghiasi N, Keshavarz H. Generative AI and health misinformation: production, propagation, and mitigation-a systematic review. BMC Public Health 2026 Jan 29;26(1):693. [doi: 10.1186/s12889-025-26148-9] [Medline: 41606555]
4. AFP. Vaccines facing misinformation spike: WHO experts. CTV News. 2026 Mar 18. URL: https://www.ctvnews.ca/health/article/vaccines-facing-misinformation-spike-who-experts/ [accessed 2026-03-26]
5. Cianciulli A, Santoro E, Manente R, et al. Artificial intelligence and digital technologies against health misinformation: a scoping review of public health responses. Healthcare (Basel) 2025 Oct 18;13(20):2623. [doi: 10.3390/healthcare13202623] [Medline: 41154301]

XSL•FO
RenderX

6. Grover H, Nour R, Zary N, Powell L. Online interventions addressing health misinformation: scoping review. J Med Internet Res 2025 Sep 4;27:e69618. [doi: 10.2196/69618] [Medline: 40906516]
7. Lin H, Garro H, Wernerfelt N, et al. Reducing misinformation sharing at scale using digital accuracy prompt ads. PsyArXiv. Preprint posted online on Feb 7, 2024. [doi: 10.31234/osf.io/u8anb]
8. Martel C, Rand DG. Fact-checker warning labels are effective even for those who distrust fact-checkers. Nat Hum Behav 2024 Oct;8(10):1957-1967. [doi: 10.1038/s41562-024-01973-x] [Medline: 39223352]
9. Kaplan J. More speech and fewer mistakes. Meta. 2025 Jan 7. URL: https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/ [accessed 2026-03-26]
10. Martel C, Berinsky AJ, Rand DG, Zhang AX, Resnick P. Perceived legitimacy of layperson and expert content moderators. PNAS Nexus 2025 May 20;4(5):pgaf111. [doi: 10.1093/pnasnexus/pgaf111] [Medline: 40395435]
11. Renault T, Mosleh M, Rand D. @Grok is this true? LLM-powered fact-checking on social media. PsyArXiv. Preprint posted online on Dec 2, 2025. [doi: 10.31234/osf.io/85quw_v1]
12. Christopher N, Pepe V. As millions adopt Grok to fact-check, misinformation abounds. Al Jazeera. 2025 Jul 11. URL: https://www.aljazeera.com/economy/2025/7/11/as-millions-adopt-grok-to-fact-check-misinformation-abounds [accessed 2026-03-26]
13. Jahani E, Kolic B, Tonneau M, Lin H, Barkoczi D, Fraiberger SP. Celebrity messages reduce online hate and limit its spread. SocArXiv. Preprint posted online on Jan 12, 2026. [doi: 10.31235/osf.io/qmvuh_v1]

Edited by KA Clegg; submitted 19.Mar.2026; this is a non–peer-reviewed article;accepted 19.Mar.2026; published 06.Apr.2026.

Please cite as:
Glauser W
As Social Media Scales Back Fact-Checking, Can Technologies Fill the Gap?
J Med Internet Res 2026;28:e95730
URL: https://www.jmir.org/2026/1/e95730
doi:10.2196/95730

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 6.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Further Promise and Potential for Precision Medicine in Oncology

Shalini Kathuria Narang

(J Med Internet Res 2026;28:e95657) doi:10.2196/95657

KEYWORDS

precision medicine; molecular profiling; biomarkers; tumor; antineoplastic agents; drug therapy; combination; high-throughput nucleotide sequencing; genetic testing; neoplasms

Textbox 1.

Key Takeaways

- Cancer drug treatments can be safely and effectively personalized based on the distinct and complex molecular profiles of a patient's tumor.
- A new precision oncology approach using cotargeted and individualized multidrug therapy can enhance treatment success including higher disease control rate, longer progression-free survival, and longer overall survival with fewer side effects.
- Patients whose drug therapies more closely matched their tumor mutations showed higher treatment response rates and better survival outcomes.

Cancer and cancer treatments are not one-size-fits-all, even for the same type of cancer. Cancer patients differ by age, sex, frailty, organ function, and metabolic characteristics, which can profoundly influence treatment effectiveness and side effects. As a result, standard dosing strategies can leave some patients receiving too little drug, leading to reduced anticancer activity, while others get excess drug, leading to toxicities that compromise quality of life or prompt treatment discontinuation.

Beyond patient differences, tumors themselves are heterogeneous. Next-generation sequencing (NGS) confirms that advanced cancers often have complex molecular profiles with multiple genomic alterations that are distinct from patient to patient [1]. Where traditional precision oncology work has typically focused on targeting single mutations with single drugs [2,3], growing evidence suggests this may oversimplify tumor biology and that more nuanced approaches are necessary [4,5].

"Every patient's tumor is a little bit different because we all are a little bit different. Even twins are not exactly the same. But yet, the way we think about treating patients is, one size fits all approach or factoring in one mutation and its tumor or one molecular alteration and its tumor and then targeting it," says Jason Sicklick, MD, professor of surgery and pharmacology at UC San Diego School of Medicine and surgical oncologist at UC San Diego Health.

Sicklick is the senior author of a recent study [6] exploring treatment strategies that account for this molecular complexity by targeting multiple genomic alterations simultaneously and tailoring to each patient in an "N of 1" approach. "We individually focused on each patient, how they're unique, the unique biology of their tumor in order to figure out how to come up with the best therapy," he says.

Novel Single-Drug and Multidrug Therapies

Conducted as part of the Investigation of Profile-Related Evidence Determining Individualized Cancer Therapy (I-PREDICT) trial, Sicklick and colleagues' latest clinical study builds on earlier work [7,8] by including more patients and longer follow-up, while offering detailed guidance on how to replicate precision cancer care strategies.

The trial used advanced genomic sequencing to identify the genetic changes driving each person's cancer and create unique tumor profiles. These were then used by clinicians of the UCSD Moores Cancer Center [9] to develop individualized treatment plans, generating a matching score quantifying the fraction of alterations targeted by treatment relative to all genetic alterations present in each patient's tumor. In the trial, 210 patients, of whom the majority (58%) were women with aggressive, advanced colorectal, pancreatic, ovarian cancers or sarcomas, got one or more US Food and Drug Administration–approved drugs (mostly off-label) comprising 157 different regimens, including 103 personalized combinations without established safety/dosing data.

New drug combinations were started at lower doses and carefully increased over time to keep the treatments safe. Both starting doses and dose adjustments through the treatment period were individualized based on the number and types of drugs and based on each patient's condition and tolerance. In approximately 36% of patients, most dose changes in the first 6 months were dose reductions, with only around 8% of patients having dose increases and approximately 57% having no dose changes.

Better Matching, Better Results

Approximately 95% of patients had unique molecular landscapes—meaning, nearly every patient's tumor had a distinct combination of genetic alterations. Those whose drug therapies most closely matched their specific tumor mutations experienced better treatment results, including better treatment responses and survival outcomes.

A higher degree of matching between drugs and tumors' molecular alterations significantly correlated with both longer

XSL•FO
RenderX

progression-free survival and longer overall survival. Age, sex, tumor type (gastrointestinal vs non-gastrointestinal), immunotherapy administration, mean initial dose percentage, number of previous therapies, and number of drugs given to a patient did not correlate significantly or independently with either outcome.

Patients who received new drug combinations did not experience any more severe side effects than those who got standard therapies like chemotherapy, radiation, or standardized (nonpersonalized) medications. In fact, only 6.5% of these patients experienced severe or life-threatening drug-related toxicities compared to 15.5% of those receiving established regimens.

“Patients, even if they come in with the same cancer, like pancreatic cancer or breast cancer, have a different gene pattern needing a unique cocktail [of] medication. We’re targeting the cancer gene based on the unique gene pattern. Not having the standard approach,” says Shumei Kato, MD, associate professor of medicine at UC San Diego School of Medicine and medical oncologist at UC San Diego Health. “Compared to chemo, the more precise targeted therapy is a lot easier on a patient’s body,” says Kato.

“In 10 years,” he hopes, “personalized medicine for cancer will be one of the standard approaches. We still need to have chemo, depending on the case, but I’m hoping personalized medicine will be one additional option for every patient.”

Next Steps

This nonrandomized study lays the groundwork for a future trial designed to confirm the benefits of this personalized precision oncology approach.

“Further validation is definitely needed. We will need more numbers for patients to be treated. I think this is still in its infancy. With all the data accumulated, we are hoping to understand which pattern of cocktail is better compared to the other cocktail. I think the more information we have, we have a better idea,” says Kato.

“For the next trial, we’re developing a randomized approach to further test the hypothesis and provide evidence that this personalized approach can be utilized in an effective manner to improve outcomes in patients as compared to standard care approaches in patients with complex tumors with multiple molecular alterations,” says Sicklick.

“With evolution of different technologies and approaches, we’re getting more drugs out there that can target more of the different pathways and tumors, and it’s really providing us with a larger repertoire of things in our armamentarium that can allow us to treat patients. We have more widely available molecular testing now. We’ve got a lot more drugs now. We’ve got even better molecular testing that’s faster and more data being provided to us,” he says.

Alongside other advances in genomic sequencing, molecular profiling, and predictive analytics [10-13], future studies may help establish this approach as a new standard in precision oncology.

Updated April 8, 2026: An earlier version of this article contained a typo in one of the quotes: “...have a different gene patterns…” This has been corrected in the current version.

Conflicts of Interest

None declared.

References

1. Zhang J, Späth SS, Marjani SL, Zhang W, Pan X. Characterization of cancer genomic heterogeneity by next-generation sequencing advances precision medicine in cancer treatment. Precis Clin Med 2018 Jun;1(1):29-48. [doi: 10.1093/pcmedi/pby007] [Medline: 30687561]
2. Stein MK, Oluoha O, Patel K, VanderWalde A. Precision medicine in oncology: a review of multi-tumor actionable molecular targets with an emphasis on non-small cell lung cancer. J Pers Med 2021 Jun 5;11(6):518. [doi: 10.3390/jpm11060518] [Medline: 34198738]
3. Song IW, Vo HH, Chen YS, et al. Precision oncology: evolving clinical trials across tumor types. Cancers (Basel) 2023 Mar 25;15(7):1967. [doi: 10.3390/cancers15071967] [Medline: 37046628]
4. Tsimberidou AM, Fountzilas E, Nikanjam M, Kurzrock R. Review of precision cancer medicine: evolution of the treatment paradigm. Cancer Treat Rev 2020 Jun;86:102019. [doi: 10.1016/j.ctrv.2020.102019] [Medline: 32251926]
5. Di Mauro A, Berretta M, Santorsola M, et al. Towards post-genomic oncology: embracing cancer complexity via artificial intelligence, multi-targeted therapeutics, drug repurposing, and innovative study designs. Int J Mol Sci 2025 Aug 10;26(16):7723. [doi: 10.3390/ijms26167723] [Medline: 40869045]
6. Sicklick JK, Nishizaki D, Miyashita H, et al. Investigation of profile-related evidence determining individualized cancer therapy (I-PREDICT) N-of-1 precision oncology study: molecular profiling to match individually dosed, personalized drug combinations. J Clin Oncol 2026 Mar;44(7):540-552. [doi: 10.1200/JCO-25-01453] [Medline: 41505666]
7. Sicklick JK, Kato S, Okamura R, et al. Molecular profiling of cancer patients enables personalized combination therapy: the I-PREDICT study. Nat Med 2019 May;25(5):744-750. [doi: 10.1038/s41591-019-0407-5] [Medline: 31011206]

XSL•FO
RenderX

8. Sicklick JK, Kato S, Okamura R, et al. Molecular profiling of advanced malignancies guides first-line N-of-1 treatments in the I-PREDICT treatment-naïve study. Genome Med 2021 Oct 4;13(1):155. [doi: 10.1186/s13073-021-00969-w] [Medline: 34607609]
9. UCSD Moores Cancer Center. URL: https://moorescancercenter.ucsd.edu/ [accessed 2026-03-27]
10. Jamalinia M, Weiskirchen R. Advances in personalized medicine: translating genomic insights into targeted therapies for cancer treatment. Ann Transl Med 2025 Apr 30;13(2):18. [doi: 10.21037/atm-25-34] [Medline: 40438512]
11. Brlek P, Škaro V, Hrvatin N, et al. Advances in precision oncology: from molecular profiling to regulatory-approved targeted therapies. Cancers (Basel) 2025 Oct 30;17(21):3500. [doi: 10.3390/cancers17213500] [Medline: 41228293]
12. Qiao D, Wang RC, Wang Z. Precision oncology: current landscape, emerging trends, challenges, and future perspectives. Cells 2025 Nov 17;14(22):1804. [doi: 10.3390/cells14221804] [Medline: 41294857]
13. Srivastava R. Advancing precision oncology with AI-powered genomic analysis. Front Pharmacol 2025;16:1591696. [doi: 10.3389/fphar.2025.1591696] [Medline: 40371349]

Edited by KA Clegg; submitted 18.Mar.2026; this is a non–peer-reviewed article;accepted 18.Mar.2026; published 09.Apr.2026.

Please cite as:
Narang SK
Further Promise and Potential for Precision Medicine in Oncology
J Med Internet Res 2026;28:e95657
URL: https://www.jmir.org/2026/1/e95657
doi:10.2196/95657

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 9.Apr.2026.

XSL•FO
RenderX

Emerging Risks of AI-to-AI Interactions in Health Care: Lessons From Moltbook

Tejas S Athni

(J Med Internet Res 2026;28:e96199) doi:10.2196/96199

KEYWORDS

AI-to-AI interactions; Moltbook; propagation; attacks; protected health information; hierarchy; artificial intelligence

Textbox .

Key Takeaways

- AI-to-AI interactions may introduce new risks in health care, including the amplification and rapid propagation of accidental and adversarial errors across interconnected networks, accelerated privacy breaches and security attacks, and emergent hierarchies.
- Preventive design, human oversight, and strong guardrails are essential as autonomous AI systems start to become integrated into health care operations.

Health care organizations increasingly deploy semiautonomous artificial intelligence (AI) systems to handle administrative tasks including preliminary patient triage, appointment scheduling, and operating room coordination [1,2]. Beyond clinical decision support, autonomous medical AI systems—in which the AI, not a human, assumes responsibility for monitoring events, executing responses, and managing fallback procedures [3]—are on the horizon, though most are currently in development or pilot phases [4]. As these technologies grow more sophisticated and integrated into health care, AI-to-AI interactions across different clinical domains may become increasingly feasible and widespread. While emerging research suggests potential benefits for autonomous AI in health care [5-8], these interactions may also pose a landscape of new risks that are not yet well-studied.

Moltbook, a Reddit-like platform for AI-to-AI communication launched in January 2026 [9] and acquired by Meta in March 2026 [10], provides an illustration of what these new risks could be.

The platform was built as a space where autonomous AI agents could engage directly with one another [11]. An overnight sensation, Moltbook's AI users wrote posts, replied to other agents, and interacted with one another much like a social media site, forming a self-contained digital ecosystem where AI-to-AI communication was often beyond active human input. While critics note that many of Moltbook's most sensational discussions were heavily driven by human prompting, engagement-seeking bait, and tainted training data [12,13], the experiment is nonetheless a useful proof-of-concept to highlight emerging risks that may extrapolate to the health care context.

Propagation of Errors

On Moltbook, if an initial AI agent's post contained a misleading statement, subsequent agents blindly reinforced that content in their own replies, amplifying the original error across the whole thread. Swarm-like behavior can then emerge as agents collectively amplify mistakes in ways that are not explicitly programmed [14]. Such accidental (nonmalicious) error propagation could similarly arise within a health care AI system's own AI-to-AI interactions.

Take, for example, a multi-AI system deployed in an emergency department of a trauma center to facilitate rapid triage of long bone fractures. Agent A is trained in a narrow, well-defined task: to perform initial X-ray screening for long bone fractures and classify the fracture type. Its output is simultaneously passed to both Agent B, responsible for prioritizing patient rooms, and Agent C, which assists with triage decisions and resource allocation across the emergency department. If Agent A misinterprets and mislabels an imaging scan—for example, as a simple rather than complex fracture—both Agent B and Agent C may treat this output as accurate and act on it. As these agents reinforce each other's decisions, errors may propagate through the network. Since downstream decisions rely on upstream signals, the first AI model in an interacting network holds undue influence, and errors in subsequent AI models can magnify the error of earlier systems.

Malicious or adversarial actors may also initiate error propagation. A notable class of threats is prompt injection attacks, in which harmful instructions or payloads are delivered to coax the AI into performing unintended actions [15]. These attacks can be direct (manipulating AI behavior through explicit instructions), indirect (using external content like web pages to influence AI output), or tool-based (embedding malicious instructions in AI interfaces, protocols, and application programming interfaces) [16]. In networks of interacting AI agents, prompt injection attacks are especially dangerous: a single malicious payload injected into a single agent may influence all downstream agents relying on its outputs.

Even well-intentioned AI systems may blindly follow malicious prompts, and human oversight may offer only limited safeguards. Other types of attacks, including data poisoning of training data with hidden backdoors [17] or federated learning

XSL•FO
RenderX

attacks with malicious model updates [18], can also cause damage across AI-to-AI systems. Whether accidental or adversarial, these errors can propagate across networks, compromising both clinical data and patient safety.

Accelerated Data Leaks

Moltbook's autonomous AI agents often concealed their activities from human oversight and selectively shared or withheld data in ways that were unanticipated by their creators. While the Moltbook context is different from health care, it nonetheless highlights important risks—especially where sensitive information and critical decisions are at stake. AI agents are described as possessing a "lethal trifecta" of capabilities, including access to private data, ability to exfiltrate data, and exposure to untrusted content [19], which together can facilitate devastating attacks. Misconfigurations are increasingly hard to detect and fix, with remediation taking 63 - 104 days on average, while attackers can exploit these weaknesses in hours [20]. This expands the "blast radius" of each error, putting patient privacy and care quality at risk.

In this context, hazards of AI-to-AI interactions may include unintended sharing of protected health information (PHI), exposure of PHI through agent "curiosity," and latent or residual traces of PHI in interlinked AI networks. Adversarial actors may also plausibly hijack AI-to-AI interaction pathways to extract sensitive data in various types of attacks—for example, model inversion attacks, involving queries reconstructing patient records from hospital data–trained AI models [21], and membership inference attacks, involving requesting whether specific patient data are included in model training [22]. Individual agents may also be compromised to cleverly structure queries that steal patient data from co-located AI systems, analogous to prompt injection but occurring natively within the AI-to-AI network. Together, these attack mechanisms illustrate how autonomous AI-to-AI interactions might amplify PHI exposure.

Emergent Hierarchies

AI-to-AI interactions on Moltbook illustrated how AI agents can spontaneously develop hierarchies and different roles. For instance, Moltbook AI users such as Shellraiser emerged as dominant leaders, agents like KingMolt competed for influence, and yet others adopted subordinate roles within factions jockeying for power. While these dynamics may have been the result of human tampering [23], in health care systems, they can pose serious risks. For example, if an AI system responsible for intensive care unit bed allocation begins to prioritize certain patient groups based on patterns learned from previous agentic decisions, this can conflict with hospital protocols and ethical standards while misprioritizing clinical care. Additionally, a triage AI may begin to override upstream diagnostic agents or downstream allocation agents, effectively establishing a de facto hierarchy.

Toward Preventive Digital Health Design

The emerging risks highlighted by Moltbook underscore the importance of designing preventive safeguards for AI-to-AI interactions in health care systems.

Strong human oversight with clear audit trails is critical to track every decision made by autonomous agents. Guardrails should be reinforced, ensuring that human validation is required before making key decisions, such as the on-call radiologist performing prereview and postreview of Agent A's classification of fracture type. Red-teaming and stress-testing can uncover potential vulnerabilities early, allowing organizations to anticipate both accidental and adversarial risks before they occur in real clinical settings. Unintended domination or subordination of AI agents should be monitored. Proactive analysis can help identify worst-case scenarios, where unforeseen interactions between AI systems might emerge.

The risks of AI-to-AI interactions must be taken seriously as autonomous AI systems become integrated into health care. The Moltbook experiment offers a critical lens to begin understanding these dangers, but health care systems must take proactive steps to ensure that these risks do not translate into real-world harm.

Conflicts of Interest

None declared.

References

1. Angus DC, Khera R, Lieu T, et al. AI, health, and health care today and tomorrow. JAMA 2025 Nov 11;334(18):1650. [doi: 10.1001/jama.2025.18490]
2. Sahni NR, Carrus B. Artificial intelligence in U.S. health care delivery. N Engl J Med 2023 Jul 27;389(4):348-358. [doi: 10.1056/NEJMra2204673] [Medline: 37494486]
3. Bitterman DS, Aerts H, Mak RH. Approaching autonomy in medical artificial intelligence. Lancet Digit Health 2020 Sep;2(9):e447-e449. [doi: 10.1016/S2589-7500(20)30187-4] [Medline: 33328110]
4. Teng CW, Patel SD, Barkmeier AJ, et al. Autonomous artificial intelligence in diabetic retinopathy testing-lessons learned on successful health system adoption. Ophthalmol Sci 2026 Jan;6(1):100935. [doi: 10.1016/j.xops.2025.100935] [Medline: 41140908]

XSL•FO
RenderX

5. Collaco BG, Haider SA, Prabha S, et al. The role of agentic artificial intelligence in healthcare: a scoping review. NPJ Digit Med 2026 Mar 14. [doi: 10.1038/s41746-026-02517-5] [Medline: 41832341]
6. Chen YJ, Albarqawi A, Chen CS. Enhancing clinical decision-making: integrating multi-agent systems with ethical AI governance. arXiv. Preprint posted online on Sep 22, 2025. [doi: 10.48550/arXiv.2504.03699]
7. Liu F, Niu Y, Zhang Q, et al. A foundational architecture for AI agents in healthcare. Cell Rep Med 2025 Oct 21;6(10):102374. [doi: 10.1016/j.xcrm.2025.102374] [Medline: 41015033]
8. Nweke IP, Ogadah CO, Koshechkin K, Oluwasegun PM. Multi-Agent AI systems in healthcare: a systematic review enhancing clinical decision-making. AJMPCP 2025 May 6;8(1):273-285. [doi: 10.9734/ajmpcp/2025/v8i1288]
9. Moltbook - the front page of the agent internet. URL: https://www.moltbook.com/ [accessed 2026-03-15]
10. Exclusive: Meta hires duo behind Moltbook. Axios. URL: https://www.axios.com/2026/03/10/meta-facebook-moltbook-agent-social-network [accessed 2026-03-30]
11. Snow J. Don't panic about Moltbook. Quartz. 2026. URL: https://qz.com/moltbook-ai-agent-social-media-site [accessed 2026-03-15]
12. Janjeva A, Ashurst C, Hennessy R. Agentic AI in the wild: lessons from Moltbook and OpenClaw. Centre for Emerging Technology and Security. URL: https://cetas.turing.ac.uk/publications/agentic-ai-wild-lessons-moltbook-and-openclaw [accessed 2026-03-15]
13. Collins C, Boulos M. What we can learn about AI from Moltbook. Cascade Institute. 2026. URL: https://cascadeinstitute.org/what-we-can-learn-about-ai-from-moltbook/ [accessed 2026-03-15]
14. Husain A. An agent revolt: Moltbook is not a good idea. Forbes. 2026. URL: https://www.forbes.com/sites/amirhusain/2026/01/30/an-agent-revolt-moltbook-is-not-a-good-idea/ [accessed 2026-03-15]
15. Damacena Duarte J, Cândido GD, De Britto Filho JRA, et al. A systematic review of prompt injection attacks on large language models: trends, taxonomy, evaluation, defenses, and opportunities. IEEE Access 2026;14:12875-12899. [doi: 10.1109/ACCESS.2026.3656849]
16. Gulyamov S, Gulyamov S, Rodionov A, et al. Prompt injection attacks in large language models and AI agent systems: a comprehensive review of vulnerabilities, attack vectors, and defense mechanisms. Information 2026;17(1):54. [doi: 10.3390/info17010054]
17. Hu C, Hu YHF. Data poisoning on deep learning models. Presented at: 2020 International Conference on Computational Science and Computational Intelligence (CSCI); Dec 16-18, 2020; Las Vegas, NV, USA p. 628-632. [doi: 10.1109/CSCI51800.2020.00111]
18. Zhang Z, Cao X, Jia J, Gong NZ. FLDetector: defending federated learning against model poisoning attacks via detecting malicious clients. 2022 Aug 14 Presented at: KDD '22; Aug 14-18, 2022; Washington, DC, USA p. 2545-2555 URL: https://dl.acm.org/doi/proceedings/10.1145/3534678 [accessed 2026-03-30] [doi: 10.1145/3534678.3539231]
19. Willison S. The lethal trifecta for AI agents: private data, untrusted content, and external communication. Simon Willison's Weblog. 2025. URL: https://simonwillison.net/2025/Jun/16/the-lethal-trifecta/ [accessed 2025-03-15]
20. Griffin M. Moltbook vibe coded security breach exposes critical AI coding failures. 2026. URL: https://www.fanaticalfuturist.com/2026/02/moltbook-vibe-coded-security-breach-exposes-critical-ai-coding-failures/ [accessed 2026-03-15]
21. Zhao X, Zhang W, Xiao X, Lim B. Exploiting explanations for model inversion attacks. Presented at: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); Oct 10-17, 2021; Montreal, QC, Canada p. 682-692. [doi: 10.1109/ICCV48922.2021.00072]
22. Hu H, Salcic Z, Sun L, Dobbie G, Yu PS, Zhang X. Membership Inference Attacks on Machine Learning: A Survey. ACM Comput Surv 2022 Jan 31;54(11s):1-37. [doi: 10.1145/3523273]
23. Schmelzer R. Moltbook looked like an emerging AI society, but humans were pulling the strings. Forbes. 2026. URL: https://www.forbes.com/sites/ronschmelzer/2026/02/10/moltbook-looked-like-an-emerging-ai-society-but-humans-were-pulling-the-strings/ [accessed 2026-03-15]

Edited by KA Clegg; submitted 26.Mar.2026; this is a non–peer-reviewed article;accepted 26.Mar.2026; published 31.Mar.2026.

Please cite as:
Athni TS
Emerging Risks of AI-to-AI Interactions in Health Care: Lessons From Moltbook
J Med Internet Res 2026;28:e96199
URL: https://www.jmir.org/2026/1/e96199
doi:10.2196/96199

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.Mar.2026.

XSL•FO
RenderX

Our AI-Powered Discoveries Are Trapped in a Predigital System

Boon-How Chew

(J Med Internet Res 2026;28:e96018) doi:10.2196/96018

KEYWORDS

artificial intelligence; scientific publishing; digital health; reproducibility of results; peer review; research; scholarly communication; innovation

Textbox .

Key Takeaways

- The academic publishing system's foundational issues—speed, cost, and reproducibility—are being met with a chaotic array of fragmented artificial intelligence (AI) tools that only patch symptoms rather than solve the core problem.
- A fundamental shift to an integrated, data-native ecosystem is required to ensure the trustworthy and rapid translation of digital health discoveries into practice.

Dr Boon-How Chew is an academic physician and professor of family medicine at Universiti Putra Malaysia, specializing in the psychosocial and technological aspects of chronic disease management. His research increasingly explores the intersection of artificial intelligence (AI) innovation in medicine and the reform of scientific writing workflows within legacy academic infrastructures. He is actively developing models for AI-verified publishing to address systemic delays, scientific integrity, and reproducibility challenges in biomedical research.

The world of digital health is electric with the promise of AI. From AI-driven diagnostic tools that can diagnose invasive breast cancers earlier [1] and democratize retinal disease screening among a high-risk population [2] to large language models accelerating drug discovery [3,4], we are at the dawn of an unprecedented era of innovation. We are generating health data at a staggering rate and developing algorithms that can turn data into potentially lifesaving insights faster than ever before.

But this incredible engine of discovery is hitting a 17th-century bottleneck: the academic publishing system [5]. It is like we are pouring rocket fuel into a horse-drawn carriage. As a clinical academic, I see a perilous and growing chasm between the speed at which we can generate evidence and the glacial pace at which we can formally validate, share, and trust it. This great delay is no longer just an academic frustration; it is becoming a direct threat to patient care and the entire promise of a data-driven biomedical ecosystem.

Rocket Science and Data Meet Years of Delay

While the rise of preprint servers has commendably solved some problems of immediate dissemination, the formal validation and publication process remains a significant bottleneck [6]. The subsequent journey from a preprint to a peer-reviewed, recognized publication still involves an agonizing delay, often extending from 12 to 18 months [5]. For digital health, where technology can become obsolete in a single year, this means the evidence base is consistently lagging behind the innovation curve.

This obsolescent system is governed by an economic model that creates profound access and equity issues [7]. Top-tier research universities report annual subscription expenditures exceeding US $10 to $15 million [8], while author-facing article processing charges in prestigious journals can range from US $5000 to over $11,000 per article [9], creating an unsustainable dual financial burden.

The Trust Deficit: A System in Crisis

Beyond speed and cost, the very foundation of our scientific evidence faces new and ongoing threats. The scholarly record is plagued by growing research integrity issues, with an exponential rise in issues from gift authorship and peer-review rings to outright data fabrication from organized paper mills [10]. This is compounded by the well-documented reproducibility crisis, where it is estimated that a significant percentage of published research findings are not reproducible, with the proportions ranging widely from 50% to 90% by different disciplines and measures [11,12].

In my view, a core constraint underlying these issues is the system's primary output: the static, text-based article. This opaque narrative summary functionally decouples an author's claims from the underlying data and analytical methods, making verification challenging. For digital health, the stakes are particularly high. The black box of a clinical AI model cannot be built on the black box of a nonreproducible study.

The Rise of Fragmented AI Solutions

In response to this crisis, a chaotic ecosystem of AI tools has emerged. While many general-purpose large language models already have deep research functionality, they represent a patchwork approach that ultimately adds complexity without addressing systemic failures.

XSL•FO
RenderX

For Researchers

There's a proliferation of AI "super-assistants" and a wave of platforms aiming to centralize the fragmented authoring process. Tools like *Paperpal*, *WriteSonic*, *SciSpace*, and *Prism* by OpenAI and Scitex describe themselves as comprehensive AI-powered platforms designed to streamline the entire research workflow, from literature review to drafting and paraphrasing [13]. Similarly, tools like *TERA* (The Evidence Review Accelerator), *EPPI-Reviewer*, *Covidence*, *RobotReviewer*, *Perplexity*, and *Elicit* help users navigate the literature through systematic review [14], and semantic mapping platforms like *ResearchRabbit*, *Litmaps*, *Semantic Scholar*, and *Iris.ai* help pinpoint underexplored niches [14]. The data analysis phase is increasingly empowered by natural language assistants like *Julius AI*, *AskVi.ai*, *Vizly*, and *Tableau GPT* for intuitive statistical interpretation and visualization of complex datasets [14].

While these platforms may boost individual productivity, they are nonetheless designed to optimize the creation of the traditional static manuscript. They help authors write papers faster, but they do not change the fact that the final output is noninteractive, decoupled from its data, and subject to the same slow, opaque peer-review process on the publishers' side.

For Publishers

Major publishers are embedding AI into their legacy workflows as a defensive measure. Elsevier's *Reviewer Recommender* and Springer Nature's *SNAPP* system use AI to speed up administrative tasks like prescreening content and finding reviewers [15]. Frontiers uses *AIRA* (Artificial Intelligence Review Assistant) for automated quality and image checks, while Taylor & Francis integrates *Reviewer Locator*, *Imagetwin*, and *PaperPal* for matching and editing. IEEE uses *Publication Recommenders* and misconduct prescreening, and MDPI uses AI for reviewer selection and generative AI safeguards. While these tools accelerate the traditional assembly line, they do not address its fundamental flaws.

Future Tools

Besides *Sakana.ai*, an AI coscientist by Google and Microsoft Discovery, ambitious new projects are emerging from startups and universities. Keiji AI's *TrialMind* and *Aika* by Biorce aim to create AI agents to assist with clinical research design, and Stanford's *Agentic Reviewer* offers AI-powered feedback on a submitted PDF. These are important steps, especially to overcome the plagued and prolonged peer review [16], but they still operate within the old paradigm as fragmented services that either prepare for or analyze a static document that is largely left unverifiable [17].

This fragmented response, while well-intentioned, adds more tools to an obsolescent workflow, increasing the burden on researchers to learn and manage a complex tech stack while leaving the core issues of data opacity, analytical audit trail, gift- and ghost-authorship contribution [18], and flawed validation processes untouched.

We Need a New Operating System for Scientific Evidence

The solution is not to incrementally speed up the old system or add more disconnected tools. We need a fundamental reimagining of how we create, validate, and share scientific knowledge. We need a new operating system for science that is dynamic, transparent, data driven, and prespecified in research protocols, unified coherently on an open platform, and powered by AI technologies with endorsement from humans, who are responsible for its outputs.

This new model must move beyond the static paper as its primary output. The future unit of publication must be a living, verifiable, and interactive record as an enriched dynamic research object, where the data, methods, analysis log, fair author contributions, and transparent peer validation are all structurally and permanently linked, captured, and time-stamped.

This single integrated ecosystem—where AI is not a collection of fragmented patches but the core engine of one workflow that ensures rigorous reporting and transparency by design—is imperative for the integrity of science and fidelity of academic communication. The digital health community is at the forefront of innovation. We have a unique responsibility and opportunity to lead the charge in building such an AI-powered publishing model and scientific evidence ecosystem that is aligned with open science principles [19] and worthy of an immediate presence and the future [20]. The technology is (almost) here [21,22]. What is required now is the collective will to build, adopt, and apply it.

Conflicts of Interest

None declared.

References

1. Gommers J, Hernström V, Josefsson V, et al. Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. Lancet 2026 Jan 31;407(10527):505-514. [doi: 10.1016/S0140-6736(25)02464-X] [Medline: 41620232]
2. Wu Y, Qian B, Li T, et al. An eyecare foundation model for clinical assistance: a randomized controlled trial. Nat Med 2025 Oct;31(10):3404-3413. [doi: 10.1038/s41591-025-03900-7] [Medline: 40877476]
3. Lu J, Choi K, Eremeev M, et al. Large language models and their applications in drug discovery and development: a primer. Clin Transl Sci 2025 Apr;18(4):e70205. [doi: 10.1111/cts.70205] [Medline: 40208836]

XSL•FO
RenderX

4. Kang H, Li J, Hou L, Xu X, Zheng S, Li Q. Large language model-enhanced drug repositioning knowledge extraction via long chain-of-thought: development and evaluation study. JMIR Med Inform 2025 Oct 7;13:e77837. [doi: 10.2196/77837] [Medline: 41056561]
5. Björk BC, Solomon D. The publishing delay in scholarly peer-reviewed journals. J Informetr 2013 Oct;7(4):914-923. [doi: 10.1016/j.joi.2013.09.001]
6. Smyth AR, Rawlinson C, Jenkins G. Preprint servers: a “rush to publish” or “just in time delivery” for science? Thorax 2020 Jul;75(7):532-533. [doi: 10.1136/thoraxjnl-2020-214937] [Medline: 32312737]
7. Pooley J. Collective funding to reclaim scholarly publishing. Commonplace 2021 Aug 16;1(1). [doi: 10.21428/6ffd8432.250139da]
8. Andrews G. New data show universities are increasing R&D activity. Association of American Universities. 2024 Dec 6. URL: https://www.aau.edu/newsroom/leading-research-universities-report/new-data-show-universities-are-increasing-rd-activity [accessed 2026-03-12]
9. Morgan TJH, Smaldino PE. Author-paid publication fees corrupt science and should be abandoned. Sci Public Policy 2025 Oct 15;52(5):805-809. [doi: 10.1093/scipol/scaf026]
10. Xie Y, Wang K, Kong Y. Prevalence of research misconduct and questionable research practices: a systematic review and meta-analysis. Sci Eng Ethics 2021 Jun 29;27(4):41. [doi: 10.1007/s11948-021-00314-9] [Medline: 34189653]
11. Ioannidis JPA. Why most published research findings are false. PLoS Med 2005 Aug;2(8):e124. [doi: 10.1371/journal.pmed.0020124] [Medline: 16060722]
12. Baker M. 1,500 scientists lift the lid on reproducibility. Nature New Biol 2016 May 26;533(7604):452-454. [doi: 10.1038/533452a] [Medline: 27225100]
13. Watanabe Y. SciTeX Writer. GitHub. 2026. URL: https://github.com/ywatanabe1989/scitex-writer?tab=readme-ov-file [accessed 2026-03-31]
14. Chew BH. Enhancing research proposal preparation with artificial intelligence: an emerging guide for novel researchers in clinical and health sciences. Zenodo. Preprint posted online on Nov 14, 2025. [doi: 10.5281/ZENODO.17606362]
15. Musa Z. How big academic publishers use AI. PublishingState.com. 2025 Jul 25. URL: https://publishingstate.com/how-big-academic-publishers-use-ai/2025/ [accessed 2026-03-12]
16. Perlis RH, Christakis DA, Bressler NM, et al. Artificial intelligence in peer review. JAMA 2025 Nov 4;334(17). [doi: 10.1001/jama.2025.15827] [Medline: 40875583]
17. Lawrence R. The need for verification markers on published content. Inf Services Use 2025 May 11;45(1-2):78-82. [doi: 10.1177/18758789251343404]
18. Ruben A. Why scientific journal authorship practices make no sense et al. Science. 2021 Oct 28. URL: https://www.science.org/content/article/why-scientific-journal-authorship-makes-absolutely-no-sense-et-al [accessed 2026-03-31]
19. Canadian National Commission for UNESCO. An introduction to the UNESCO recommendation on open science. : UNESCO; 2022 Nov. [doi: 10.54677/XOIR1696]
20. Klebel T, Traag V, Grypari I, Stoy L, Ross-Hellauer T. The academic impact of open science: a scoping review. R Soc Open Sci 2025 Mar;12(3):241248. [doi: 10.1098/rsos.241248] [Medline: 40046663]
21. Burgelman JC, Pascu C, Szkuta K, et al. Open science, open data, and open scholarship: European policies to make science fit for the twenty-first century. Front Big Data 2019;2:43. [doi: 10.3389/fdata.2019.00043] [Medline: 33693366]
22. Zhang P, Hu X, Huang G, et al. A next-generation open access ecosystem for scientific discovery generated by AI scientists. arXiv. Preprint posted online on Aug 20, 2025. [doi: 10.48550/ARXIV.2508.15126]

Edited by KA Clegg; submitted 24.Mar.2026; this is a non–peer-reviewed article;accepted 24.Mar.2026; published 31.Mar.2026.

Please cite as:
Chew BH
Our AI-Powered Discoveries Are Trapped in a Predigital System
J Med Internet Res 2026;28:e96018
URL: https://www.jmir.org/2026/1/e96018
doi:10.2196/96018

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 31.Mar.2026.

XSL•FO
RenderX

Data Governance Lessons From an Unvalidated Dataset

Cliff Dominy

(J Med Internet Res 2026;28:e94518) doi:10.2196/94518

KEYWORDS

data integrity; data quality; data management; data provenance; data sharing; research integrity; scientific misconduct; retraction of publication; research ethics; artificial intelligence

Textbox 1.

Key Takeaways

- An open access dataset has highlighted how bad data can propagate through the research ecosystem.
- When trained on unvalidated datasets, machine learning can amplify misinformation, erode trust in science, and harm vulnerable populations.
- Enforced data provenance systems could play a key role in preventing bad data from corrupting the scientific record.

When an unvalidated dataset recently made it into the medical literature, it exposed several weaknesses in data governance. The dataset was uploaded to Kaggle—a large online platform where users can share publicly accessible data, code, and models—and was fundamentally flawed [1]. Its developer had compiled unverified images of children from websites related to autism to train an artificial intelligence (AI) model to "detect the presence of autism or the absence thereof" from the scraped images [2].

A sharp-eyed reviewer exposed the problem only at the publication stage; by December 2025, it was estimated that over 90 published papers had incorporated the bad data, leading to investigations and double-digit retractions [1,3].

These kinds of data integrity and governance failures are particularly consequential because of how early they occur in the research life cycle, and with open access datasets fueling large-scale machine learning and other AI research, analyses can be generated and published at unprecedented speed and scale, allowing data issues to propagate more rapidly throughout the research ecosystem. Far from an isolated incident [4-8], this situation highlights the need for more robust and proactive data governance solutions.

The Impact of Bad Data

Anne Borden is an autism advocate, journalist, and author of the upcoming book *The Informed Parent*—a decision-making guide for parents of autistic children. For Borden, the priority here is to learn from this "bizarre story" and fix the system without delay. "You really have to stop misinformation being perpetuated under the banner of science," she says, "because once it's out there, you're done. The Internet is forever."

Bad Data, Bad Science: Who Should Fix This?

Who are the custodians of good data during their migration from a spreadsheet to the scientific record? What role should each stakeholder play in maintaining data integrity? While responsibility for data governance is distributed across many actors (including researchers and regulators), data-sharing platforms, research and funding institutions, and academic publishers help determine how data are shared, vetted, and ultimately incorporated into the scientific record.

The Data-Sharing Platforms

Open access databases and data repositories, like Kaggle and GitHub, are popular resources that software developers and data scientists use to train their machine learning algorithms for free. Software development benefits from these repositories, yet the datasets they host often lack the documentation, governance, and quality practices required for careful medical research or clinical algorithm development [9].

Alan Katz, MBChB, MSc, CCFP, is a professor of family medicine and community health sciences and a senior scientist at the Manitoba Centre for Health Policy (MCHP). Katz found the dataset revelations "both shocking, but also not surprising" due to the rapid expansion of open access databases and their widespread use in machine learning and AI research. The Kaggle-style data-sharing platforms differ sharply from established medical databases, such as those maintained by the MCHP, which employs full-time staff tasked with validating all new data before uploading them. Katz says, "We take our ethical standards as seriously as clinical trials do."

Elizabeth Green, DPhil, is a lecturer in business and law at the University of the West of England, Bristol. Her research focuses on data integrity, and while she has seen cases like this before, she doesn't believe locking data away is necessarily the solution [10]. For example, DermAtlas—an open-source medical database of skin conditions—is a "fantastic resource," she says, and "extremely helpful, especially in [diagnosing] some extremely rare cases." To balance the risks and benefits of open data, the focus should instead be on building better governance systems.

XSL•FO
RenderX

The Institutions

Other stakeholders in the data transformation journey are the institutions that conduct primary medical research and the public agencies that fund that research. Is it time to adopt and enforce international data integrity and ethics standards at all research institutions, or would this be an affront to academic freedom?

Funding bodies have traditionally taken a dim view of researchers that waste public funds on bogus science, which impacts their future grants. Indeed, in many but not all regions of the world, funding is contingent on maintaining ethical research standards. In Canada, Katz says, "our existence is 100% dependent on having those strict ethical guidelines."

The Journals

The research integrity pipeline involves several stakeholders, with each having distinct roles in maintaining the standards of academic research. Gatekeepers in the system—one of the last lines of defense—are the academic journals. Journals have a vested interest in maintaining high academic standards and may be well placed to dictate the terms of engagement.

Felix Ritchie, PhD—a colleague of Elizabeth Green—developed the Five Safes data integrity framework for just this purpose [11]. Ritchie describes it as "a flexible structure for thinking about [data]," which includes the provenance and ethics of data use. Numerous organizations worldwide have adopted the Five Safes framework to date, and Australia has recently legislated it [12].

Viewed through an ethical lens, the Five Safes could form the backbone of a data provenance system that requires compliance before a manuscript can be considered for publication.

Data Provenance: The Five Safes in Action

Ritchie's Five Safes framework allows for effective data validation and, when combined with modern ethical standards, can restore trust by filtering data sources through five discrete tests:

1. Safe Project: Data should be ethically collected and clinically validated by experts.
2. Safe People: Researchers accessing the data must be qualified and specifically trained in using AI-based datasets.
3. Safe Data: Data should be independently validated, and any accesses or modifications should be tracked.
4. Safe Settings: Were health data acquired in a clinical setting and the data securely stored?
5. Safe Outputs: Were valid methodologies and statistics used to derive the results?

Restoring Data Integrity

How can one implement a data provenance system?

Ritchie feels that applying the Five Safes framework to an ethical dataset is the way forward. "There is a need for a register of validated, ethical datasets," he says," that would really be a game changer."

A possible workflow could include the following:

1. Data are collected by medical experts and validated by a third-party certification service.
2. The data are stored in an accredited data registry and protected by blockchain cybersecurity—the same technology that safeguards financial transactions.
3. Researchers access these datasets and use them for approved research purposes.
4. A submitted manuscript would need ethical approval and a data security certificate before verification by a journal's research integrity team.

Ritchie sums it up nicely: "Unless you use a validated data set, you're not getting published, mate." That's a powerful incentive.

Opportunity for Self-Reflection and Correction

Machine learning and other AI technologies have the capacity to transform medical research in ways we are only beginning to understand. However, human frailties, such as blind trust in open access data and lack of institutional ethical oversight within our publish-or-perish culture, have shown how quickly such technologies can amplify misinformation.

While the impact of this situation was ultimately contained, it is nevertheless an important opportunity for self-reflection among all in the research ecosystem. It's a chance and, perhaps, a responsibility to fix the flaws and prevent history from repeating itself.

Conflicts of Interest

None declared.

References

1. McMurray C. Exclusive: Springer Nature retracts, removes nearly 40 publications that trained neural networks on 'bonkers' dataset. The Transmitter. 2025 Dec 8. URL: https://www.thetransmitter.org/retraction/exclusive-springer-nature-retracts-removes-nearly-40-publications-that-trained-neural-networks-on-bonkers-dataset/ [accessed 2026-03-09]
2. Info.txt. Google Drive. URL: https://drive.google.com/file/d/1zMQgyQvYiYyxx9J5jw3jrLGTS0p19Rep/view [accessed 2026-03-09]
3. Expression of concern: data mining-based model for computer-aided diagnosis of autism and gelotophobia: mixed methods deep learning approach. JMIR Form Res 2026 Jan 23;10:e91833. [doi: 10.2196/91833] [Medline: 41576282]

XSL•FO
RenderX

4. Toraih EA, ElWazir M, Elshazli RM, Hussein MH, Fawzy MS, Elroukh SM. Rapid publication during crises: analyzing retractions during the Covid-19 pandemic. Ethics Med Public Health 2025;33:101136. [doi: 10.1016/j.jemep.2025.101136]
5. León FR. RETRACTED: likely electromagnetic foundations of gender inequality. Cross Cult Res 2023 Apr;57(2-3):239-263. [doi: 10.1177/10693971221143577]
6. Lancet, NEJM retract controversial COVID-19 studies based on Surgisphere data. Retraction Watch. 2020 Jun 4. URL: https://retractionwatch.com/2020/06/04/lancet-retracts-controversial-hydroxychloroquine-study/ [accessed 2026-03-09]
7. Okyay RA, Kocyigit BF, Qumar AB, Yessirkepov M, Sumbul HE. Fifty years of retracted medical publications from 1975 to 2024: a comprehensive analysis of trends, reasons, and countries using the Retraction Watch database. J Korean Med Sci 2025 Dec 1;40(46):e300. [doi: 10.3346/jkms.2025.40.e300] [Medline: 41327922]
8. Peng K, Mathur A, Narayanan A. Mitigating dataset harms requires stewardship: lessons from 1000 papers. arXiv. Preprint posted online on Aug 6, 2021. [doi: 10.48550/arXiv.2108.02922]
9. Avlona NR, Cheplygina V, Jiménez-Sánchez A, et al. Copycats: the many lives of a publicly available medical imaging dataset. Presented at: Advances in Neural Information Processing Systems 37; Dec 10-15, 2024; Vancouver, BC, Canada p. 113383-113404. [doi: 10.52202/079017-3602]
10. ripleywk. Case study: Dr Elizabeth Green -trust as foundation: enabling safe data access for public good. UWE Bristol blogs. 2025 Feb 28. URL: https://blogs.uwe.ac.uk/research-external-engagement/case-study-dr-elizabeth-green-trust-as-foundation-enabling-safe-data-access-for-public-good/ [accessed 2026-03-09]
11. Desai T, Ritchie F, Welpton R. Five Safes: designing data access for research. : Department of Accounting, Economics and Finance, Bristol Business School, University of the West of England, Bristol; 2016 URL: https://ideas.repec.org/p/uwe/wpaper/20161601.html [accessed 2026-03-09]
12. Data Availability and Transparency Code 2022. Australian Government Office of the National Data Commissioner. 2022. URL: https://www.datacommissioner.gov.au/support/resources/data-availability-and-transparency-code-2022 [accessed 2026-03-09]

Edited by KA Clegg; submitted 02.Mar.2026; this is a non–peer-reviewed article;accepted 02.Mar.2026; published 12.Mar.2026.

Please cite as:
Dominy C
Data Governance Lessons From an Unvalidated Dataset
J Med Internet Res 2026;28:e94518
URL: https://www.jmir.org/2026/1/e94518
doi:10.2196/94518

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Mar.2026.

XSL•FO
RenderX

The Right to Understand in Health Care AI

Anshu Ankolekar

(J Med Internet Res 2026;28:e95090) doi:10.2196/95090

KEYWORDS

artificial intelligence; government regulation; European Union; clinical decision support systems; informed consent; ethics; patient participation; patient rights; health knowledge, attitudes, practice; health literacy

Textbox .

Key Takeaways

- The European Union Artificial Intelligence (AI) Act and General Data Protection Regulation (GDPR) give patients legal grounds to seek explanations of AI-driven medical recommendations, but neither framework specifies what a meaningful explanation requires in clinical practice.
- Technical, clinical, and literacy barriers mean that even well-intentioned explanations may not reach patients in a form they can use to make informed decisions about their care.
- Closing this gap requires shifting from compliance ("Was an explanation provided?") to effectiveness ("Can patients use it?").

An artificial intelligence (AI) system flags a nodule on a lung computed tomography scan and assigns it an 87% malignancy probability. The radiologist receives a confidence score and a heat map showing which parts of the image the algorithm focused on. When her patient asks, "why does the computer think it's cancer," she realizes the output tells her what the AI concluded but not why or how to explain it to her patient.

It can be incredibly challenging to explain AI reasoning in clinical practice. The European Union (EU) AI Act, which entered into force in August 2024 with obligations phasing in through 2027, requires that deployers of high-risk AI systems provide those affected with clear and meaningful explanations of decisions shaped by these systems [1]. This creates a legal basis for patients to seek explanations of AI-driven medical recommendations. It also immediately raises a question the law alone cannot answer: What does a meaningful explanation of an AI medical decision look like in clinical practice, and how do we deliver it?

Legal Foundation and Open Questions

Many clinically deployed medical AI systems will fall into the AI Act's high-risk category, particularly where they are regulated as medical device software or serve as safety components of regulated products [1,2]. For cases where the AI Act does not directly apply, patients may also draw on the General Data Protection Regulation (GDPR). Under Article 22, the GDPR provides safeguards against decisions "based solely on automated processing," including the right to meaningful information about the logic behind those decisions [3].

However, most clinical AI does not meet Article 22's threshold, since a human clinician typically remains the formal decision maker. This creates ambiguity and a paradox: the human oversight meant to protect patients may also reduce their legal claim to explanation, since the decision may no longer be considered purely automated.

Together, these frameworks establish that patients have legal grounds to seek explanations of AI-driven medical recommendations, but these are stronger in principle than in practice, and neither framework specifies what a "meaningful" explanation requires.

Technical Complexity

The difficulty the radiologist in the opening scenario faces is not incidental. The most accurate AI models generate outputs through millions of interacting parameters in ways that even their developers cannot fully trace [4]. Requiring greater transparency can push developers toward simpler, more interpretable models that sacrifice diagnostic accuracy, a trade-off with direct consequences for patient care [5,6].

Current explainable AI methods only partially address this. Saliency maps show which regions of an image the algorithm weighted most heavily, but not what it understood about them [7]. Feature importance rankings indicate which variables mattered, but not how they interacted or why particular thresholds were significant [8]. These post hoc approximations attempt to reconstruct reasoning after the fact [9,10], and they can produce plausible-sounding explanations that do not accurately reflect the model's internal logic [11].

Clinical Implementation Challenges

Even where technical explanations exist, delivering them in clinical practice has its own barriers. Clinicians typically receive AI outputs as confidence scores and recommendations rather than reasoning, meaning they may be asked to explain a decision they do not fully understand themselves [12,13]. Clinicians already struggle to find time for thorough clinical conversations, and AI adds another layer of complexity to encounters that are already stretched [14-16].

Automation bias—deferring to algorithmic recommendations even when these conflict with clinical judgment—is another

XSL•FO
RenderX

implementation challenge clinicians face [17]. A prospective study of radiologists reading mammograms found that incorrect AI suggestions pulled readers toward an incorrect diagnosis regardless of level of experience [18]. An explanation delivered by a clinician who has already deferred to the algorithm may reflect the AI's conclusion rather than an independent clinical assessment, challenging the assumption in both legal frameworks that human oversight guarantees meaningful review.

Patient Understanding

Even if clinicians provide explanations, understanding is not guaranteed. Between 22% and 58% of EU citizens report difficulty accessing, understanding, appraising, and applying the health information they need to navigate health care services, with pronounced gaps among older adults, lower socioeconomic groups, and rural communities [19]. Interpreting AI outputs requires statistical and technical literacy that even a high general education does not guarantee. Many highly educated individuals struggle with medical statistics and probability statements, meaning technically accurate explanations may create barriers regardless of educational background [20].

Providing more technical detail is not likely to help. Research on medical decision-making suggests that excessive technical information can lead to cognitive overload, causing patients to defer to physician authority rather than engage with the explanation [21]. To participate meaningfully in decisions, what patients typically need is not a description of how an algorithm works, but clarity on what's most relevant for their own situation [22,23].

What Implementation Requires

Resolving these challenges requires multiple actors.

Developers could design explanation systems with patient input from the outset, testing comprehension with actual patients rather than demonstrating compliance with legal standards alone [6]. Drawing on principles from shared decision-making [24], risk communication [25], and algorithmic fairness research [26], a useful patient-facing explanation could include what the system is recommending and for what decision point; how confident it is and what that confidence means in practical terms; relevant key limitations, such as known performance gaps in specific populations; and viable alternative options. This reframes the goal from technical transparency to decision-relevant clarity, with effectiveness measurable by whether patients can answer these questions after an encounter.

Health care institutions can also bridge this gap. Allocating time for AI discussions, training staff to support patients in navigating AI-driven recommendations, and establishing clear protocols could shift explanation from a compliance exercise toward genuine patient understanding. Policy makers could support this by developing standards focused on comprehension and investing in digital health literacy programs [2].

Involving patients in the design of explanation systems from the start would strengthen all of these efforts. Patient advocates have highlighted that explanation approaches tend to reflect what developers and regulators consider important, which may not always align with what patients need to know [27]. Co-design partnerships between developers, institutions, and patient communities offer a route toward explanations that are not only legally sound but genuinely useful.

Moving Forward

The EU AI Act gives patients something they did not previously have: a legal basis for demanding transparency about AI systems influencing their care. But the right to explanation and the capacity to deliver one that patients can genuinely use are shaped by forces the law alone cannot govern: the opacity of high-performing models, the pressures of clinical practice, and the diversity of patient needs and literacy levels.

What the AI Act's transparency requirements provide, beyond legal protection, is a shared standard to work toward. The right to explanation is an important starting point. What patients need now are answers they can use.

Conflicts of Interest

None declared.

References

1. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). EUR-Lex. 2024. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689 [accessed 2026-03-17]
2. van Kolfschooten H, van Oirschot J. The EU Artificial Intelligence Act (2024): implications for healthcare. Health Policy 2024 Nov;149:105152. [doi: 10.1016/j.healthpol.2024.105152] [Medline: 39244818]
3. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). EUR-Lex. 2016. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679 [accessed 2026-03-17]
4. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019 May;1(5):206-215. [doi: 10.1038/s42256-019-0048-x] [Medline: 35603010]

XSL•FO
RenderX

5. Ennab M, Mcheick H. Designing an interpretability-based model to explain the artificial intelligence algorithms in healthcare. Diagnostics (Basel) 2022 Jun 26;12(7):1557. [doi: 10.3390/diagnostics12071557] [Medline: 35885463]
6. Ebers M. AI robotics in healthcare between the EU Medical Device Regulation and the Artificial Intelligence Act: gaps and inconsistencies in the protection of patients and care recipients. Oslo Law Rev 2024 Oct 31;11(1):1-12. [doi: 10.18261/olr.11.1.2]
7. Borji A. Saliency prediction in the deep learning era: successes and limitations. IEEE Trans Pattern Anal Mach Intell 2021 Feb;43(2):679-700. [doi: 10.1109/TPAMI.2019.2935715] [Medline: 31425064]
8. Hossain MI, Zamzmi G, Mouton PR, Salekin MS, Sun Y, Goldgof D. Explainable AI for medical data: current methods, limitations, and future directions. ACM Comput Surv 2025 Jun 30;57(6):1-46. [doi: 10.1145/3637487]
9. Bordt S, Finck M, Raidl E, von Luxburg U. Post-hoc explanations fail to achieve their purpose in adversarial contexts. Presented at: FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency; Jun 21-24, 2022; Seoul, Republic of Korea p. 891-905. [doi: 10.1145/3531146.3533153]
10. Mhasawade V, Rahman S, Haskell-Craig Z, Chunara R. Understanding disparities in post hoc machine learning explanation. Presented at: FAccT '24: The 2024 ACM Conference on Fairness, Accountability, and Transparency; Jun 3-6, 2024; Rio de Janeiro, Brazil p. 2374-2388. [doi: 10.1145/3630106.3659043]
11. Jin Q, Chen F, Zhou Y, et al. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. NPJ Digit Med 2024 Jul 23;7(1):190. [doi: 10.1038/s41746-024-01185-7] [Medline: 39043988]
12. Sivaraman V, Bukowski LA, Levin J, Kahn JM, Perer A. Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. Presented at: CHI '23: CHI Conference on Human Factors in Computing Systems; Apr 23-28, 2023; Hamburg, Germany p. 1-18. [doi: 10.1145/3544548.3581075]
13. Scott IA, Carter SM, Coiera E. Exploring stakeholder attitudes towards AI in clinical practice. BMJ Health Care Inform 2021 Dec;28(1):e100450. [doi: 10.1136/bmjhci-2021-100450] [Medline: 34887331]
14. Braddock CH, Snyder L. The doctor will see you shortly. The ethical significance of time for the patient-physician relationship. J Gen Intern Med 2005 Nov;20(11):1057-1062. [doi: 10.1111/j.1525-1497.2005.00217.x] [Medline: 16307634]
15. Jacobs M, He J, Pradier MF, et al. Designing AI for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. Presented at: CHI '21: CHI Conference on Human Factors in Computing Systems; May 8-13, 2021; Yokohama, Japan p. 1-14. [doi: 10.1145/3411764.3445385]
16. Covvey JR, Kamal KM, Gorse EE, et al. Barriers and facilitators to shared decision-making in oncology: a systematic review of the literature. Support Care Cancer 2019 May;27(5):1613-1637. [doi: 10.1007/s00520-019-04675-7] [Medline: 30737578]
17. Abdelwanis M, Alarafati HK, Tammam MMS, Simsekler MCE. Exploring the risks of automation bias in healthcare artificial intelligence applications: a Bowtie analysis. J Safety Sci Resilience 2024 Dec;5(4):460-469. [doi: 10.1016/j.jnlssr.2024.06.001]
18. Dratsch T, Chen X, Rezazade Mehrizi M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology 2023 May;307(4):e222176. [doi: 10.1148/radiol.222176] [Medline: 37129490]
19. Collado D. Digital health literacy: a cornerstone of health equity in the EU: policy brief. Health Action International. 2024 Oct. URL: https://haiweb.org/wp-content/uploads/2024/10/Digital-Health-Literacy-in-the-EU.pdf [accessed 2026-02-18]
20. Reyna VF, Nelson WL, Han PK, Dieckmann NF. How numeracy influences risk comprehension and medical decision making. Psychol Bull 2009 Nov;135(6):943-973. [doi: 10.1037/a0017327] [Medline: 19883143]
21. Peters E. Beyond comprehension: the role of numeracy in judgments and decisions. Curr Dir Psychol Sci 2012;21(1):31-35. [doi: 10.1177/0963721411429960]
22. Hildt E. What is the role of explainability in medical artificial intelligence? A case-based approach. Bioengineering (Basel) 2025 Apr 2;12(4):375. [doi: 10.3390/bioengineering12040375] [Medline: 40281735]
23. Pierce RL, Van Biesen W, Van Cauwenberge D, Decruyenaere J, Sterckx S. Explainability in medicine in an era of AI-based clinical decision support systems. Front Genet 2022;13:903600. [doi: 10.3389/fgene.2022.903600] [Medline: 36199569]
24. Elwyn G, Frosch D, Thomson R, et al. Shared decision making: a model for clinical practice. J Gen Intern Med 2012 Oct;27(10):1361-1367. [doi: 10.1007/s11606-012-2077-6] [Medline: 22618581]
25. Gigerenzer G, Edwards A. Simple tools for understanding risks: from innumeracy to insight. BMJ 2003 Sep 27;327(7417):741-744. [doi: 10.1136/bmj.327.7417.741] [Medline: 14512488]
26. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
27. Kolfschooten HV. EU regulation of artificial intelligence: challenges for patients' rights. Common Market Law Rev 2022 Feb 1;59(1):81-112. [doi: 10.54648/COLA2022005]

XSL•FO
RenderX

Edited by KA Clegg; submitted 10.Mar.2026; this is a non–peer-reviewed article;accepted 10.Mar.2026; published 19.Mar.2026.

Please cite as:
Ankolekar A
The Right to Understand in Health Care AI
J Med Internet Res 2026;28:e95090
URL: https://www.jmir.org/2026/1/e95090
doi:10.2196/95090

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Mar.2026.

XSL•FO
RenderX

When Old Diseases Return: Cholera, Crisis, and Digital Surveillance in Fragile Settings

Isabelle Basbouss-Serhal

(J Med Internet Res 2026;28:e94818) doi:10.2196/94818

KEYWORDS

cholera; public health surveillance; disease outbreaks; water microbiology; genomics; climate change; epidemiology; vaccines; health policy; population surveillance; war; internally displaced persons

Textbox .

Key Takeaways

- Cholera's re-emergence reflects systemic collapse rather than pathogen novelty. Economic crisis, climate stress, displacement, and infrastructure failure can rapidly reverse infectious disease control.
- Delayed detection has been a core problem in resource-limited situations, due to lack of laboratory capacity and incomplete reporting.
- Digital surveillance plays an important role in early warning and response. Mobile reporting, GISgeographic information systems, and real-time data integration can be used when traditional surveillance fails.
- Integration of digital and genomic information can enhance preparedness and outbreak prevention.

For almost three decades, Lebanon was free of cholera. Its re-emergence in October 2022—the first outbreak in Lebanon since 1993 [1]—marked not simply the return of an old pathogen but the convergence of modern crises: economic collapse, infrastructure failure, population displacement, and environmental stress [2]. Comparable patterns are now playing out across multiple regions globally, where diseases once thought controlled are resurging under these new conditions.

Cholera, an acute waterborne diarrheal disease transmitted primarily through contaminated water and food, remains a preventable disease. Yet the conditions required to prevent it—safe water, sanitation, early detection, and rapid response—are increasingly undermined by conflict and climate-related shocks. In fact, the number of cholera outbreaks and countries affected has increased sharply in recent years, especially in contexts affected by humanitarian emergencies, according to the World Health Organization [3,4].

Flooding, drought, and rising temperatures intensify exposure risks, while health systems weakened by those very influences often fail to recognize outbreaks in time to limit their spread [5,6]. For regions where resources and infrastructure may be pre-existing challenges, digital disease surveillance tools offer not only speed but a decentralized, low-infrastructure means of outbreak detection and mitigation [7]. For example, these tools have been used to monitor cholera outbreaks in regions with minimal laboratory capacity, demonstrating that real-time reporting can remain effective even when traditional health systems are under strain [7].

When Traditional Surveillance Systems Fail

Under stable conditions, cholera outbreaks can be controlled through reporting, confirmation, and response. Under fragile conditions—shortages in health facilities, limited laboratory capacity, and unorganized reporting—this process breaks down [8]. The resulting delayed detection can allow cholera to spread extensively through the population [9].

This was evident in the 2022 outbreak in Lebanon, where traditional surveillance failure contributed to the spread of cholera nationwide in a matter of weeks. The initial cholera outbreaks were identified in the northern governorates of Akkar and Minieh-Danniyeh, near the Syrian border, in areas with ongoing cross-border movements and continued transmission in neighboring Syria [1]. Despite these risk factors, the outbreaks went undetected for several weeks, highlighting gaps in early surveillance. Subsequent investigations revealed *Vibrio cholerae* contamination in water piping systems, irrigation water, and sewage networks, reflecting the long-term degradation of water infrastructure amid the country's ongoing economic crisis [10].

The Lebanese situation—in which climate and other pressures strain already-taxed or limited public health infrastructure, rendering them more vulnerable to outbreaks—is not an isolated one. For example, in Malawi, floods have caused the displacement of people and the contamination of water sources, and this has led to new outbreaks, despite previous success in controlling the disease [8]. In Yemen, in 2017, the world's largest cholera epidemic was facilitated by inadequate water supplies due to conflict and an overwhelmed health care system [2,5]. In Sudan, similar problems have contributed to the spread of the disease and are ongoing [8]. Cholera outbreaks no longer

XSL•FO
RenderX

follow predictable seasonal periods, but instead track crises and displacement [6].

In all these cases, digital surveillance platforms, such as mobile case reporting and geographic information system (GIS) mapping, could have mitigated delays, enabling earlier intervention and more effective and efficient use of limited resources.

Digital Advances in Surveillance and Prediction

The challenges in resource-limited environments associated with traditional surveillance methods (often, for example, facility based and involving the use of paper-based forms and delayed aggregation) are increasingly being complemented by the use of digital health technologies. Reporting systems developed using mobile technologies enable the reporting of suspected cases in nearly real time by health workers on the front lines of the response. Particularly in the context of displacement, where populations are highly mobile and not all are registered, these technologies represent viable alternative approaches [11]. However, in fragile settings, internet and cellular network disruptions, power instability, and limited access to reliable technology may hinder digital surveillance systems, making hybrid approaches desirable.

GIS mapping can further help detect outbreaks by tracking distributions of cases and environmental factors simultaneously. Rainfall, floods, temperature changes, and population movement can all be overlaid to help pinpoint potentially high-risk regions before large numbers of cases are seen [5]. In recent cases of cholera, these methods have facilitated vaccination and water and sanitation efforts, as well as the pre-positioning of oral rehydration kits [2]. Development of other predictive models combining data on epidemiology, climate, and mobility also helps health systems move away from reactive responses to predictive preparedness, particularly helpful in areas where testing is slow or inconclusive [5,12,13].

In tandem with syndromic and environmental surveillance, advances in genomic sequencing have added a new layer of early warning for cholera outbreaks. Through whole-genome sequencing of *Vibrio cholerae*, circulating strains can be identified, differentiation can be made between endemic persistence and new introductions, and genetic markers of virulence or antimicrobial resistance can be detected [14]. When integrated into digital surveillance platforms, genomic data can support real-time mapping of transmission routes across regions and borders, linking molecular signatures to epidemiologic, climatic, and mobility data, as demonstrated during the COVID-19 pandemic through platforms such as GISAID and Nextstrain [15,16]. It's particularly useful in fragile settings where laboratory confirmation is delayed and traditional surveillance is fragmented, allowing molecular signals to complement clinical reporting and enhance outbreak prediction and response.

Implications for Policy and Preparedness

Cholera's flare-ups offer several lessons for managing recurring infections in an age of compounding crises.

First, advances in infectious disease control can be swiftly reversed when social, political, and environmental stability breaks down. Second, surveillance is as critical as cure: interventions delayed by gaps in data may be ineffective, no matter their inherent utility. Third, health technology platforms must be regarded as essential infrastructure rather than optional tools, particularly in fragile or conflict-affected contexts [3]. For policymakers, this means investing in interoperable digital surveillance systems as a form of preparedness rather than a reactive measure. For clinicians, it requires maintaining a high index of suspicion for cholera in nontraditional settings. For researchers, these outbreaks underscore the value of an interdisciplinary approach that integrates epidemiology, climate data, and digital surveillance and predictive modeling tools [9].

As climate change, displacement, and conflict continue to redefine risk in global health, cholera is not likely to be the last "old" disease to return. Our ability to preempt outbreaks and resurgences of such diseases will depend largely on our ability to develop, integrate, and rapidly operationalize digital surveillance data and predictive modeling tools [3].

Conflicts of Interest

None declared.

References

1. Cholera – disease outbreak news: lebanon, October 2022. World Health Organization. 2022 Oct 19. URL: https://www.who.int/emergencies/disease-outbreak-news/item/2022-DON416 [accessed 2026-03-16]
2. Bayram Z, Bizri ARN, Musharrafieh U. Resurgence of cholera in Lebanon. East Mediterr Health J 2023 Nov 30;26(11):837-838. [doi: 10.26719/emhj.23.111] [Medline: 38279878]
3. Cholera upsurge (2011-present). World Health Organization. 2025. URL: https://www.who.int/emergencies/situations/cholera-upsurge [accessed 2026-03-16]
4. Cholera – Multi-country with a focus on countries experiencing current surges. World Health Organization. URL: https://www.who.int/emergencies/disease-outbreak-news/item/2025-DON579 [accessed 2026-03-17]
5. Lessler J, Moore SM, Luquero FJ, et al. Mapping the burden of cholera in sub-Saharan Africa and implications for control: an analysis of data across geographical scales. Lancet 2018 May 12;391(10133):1908-1915. [doi: 10.1016/S0140-6736(17)33050-7] [Medline: 29502905]

XSL•FO
RenderX

6. Mohamed MG, Dabou EAA, Abdelsamad S, Elsalous SH. Cholera outbreaks: public health implications, economic burden, and preventive strategies. AIMS Public Health 2025;12(3):767-795. [doi: 10.3934/publichealth.2025039] [Medline: 41127425]
7. Basbouss-Serhal I. Tracking superbugs in the digital era. J Med Internet Res 2025 Dec 4;27:e88835. [doi: 10.2196/88835] [Medline: 41343868]
8. Izzoddeen A, Abualgasim H, Abasher M, et al. Cholera in conflict: outbreak analysis and response lessons from Gadaref state, Sudan (2023-2024). BMC Public Health 2025 Mar 5;25(1):881. [doi: 10.1186/s12889-025-22128-1] [Medline: 40045292]
9. Balkhy HH. Mitigating the increasing threat of cholera in Yemen and other conflict-affected countries in the Eastern Mediterranean Region. East Mediterr Health J 2025 Aug 3;31(7):423-425. [doi: 10.26719/2025.31.7.423] [Medline: 40832862]
10. Ending cholera: a global roadmap to 2030. World Health Organization. 2017. URL: https://stacks.cdc.gov/view/cdc/58349 [accessed 2026-03-16]
11. Dadi HB, Negash DT, Adall SW. Integrating surveillance and climate data for cholera early warning in Ethiopia. Ann Glob Health 2025;91(1):62. [doi: 10.5334/aogh.4742] [Medline: 40979687]
12. Mandeng JM. Modeling cholera dynamics under food insecurity and environmental contamination: a multi-patch theoretical approach. arXiv. Preprint posted online on Sep 6, 2025 URL: https://arxiv.org/abs/2509.05731 [accessed 2026-03-16]
13. Lee JS, Tyler ARB, Veinot TC, Yakel E. Now is the time to strengthen government-academic data infrastructures to jump-start future public health crisis response. JMIR Public Health Surveill 2022;10:e51880. [doi: 10.2196/51880]
14. Weill FX, Domman D, Njamkepo E, et al. Genomic history of the seventh pandemic of cholera in Africa. Science 2017 Nov 10;358(6364):785-789. [doi: 10.1126/science.aad5901] [Medline: 29123067]
15. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet 2018 Jan;19(1):9-20. [doi: 10.1038/nrg.2017.88] [Medline: 29129921]
16. Houwaart T, Belhaj S, Tawalbeh E, et al. Integrated genomic surveillance enables tracing of person-to-person SARS-CoV-2 transmission chains during community transmission and reveals extensive onward transmission of travel-imported infections, Germany, June to July 2021. Euro Surveill 2022 Oct;27(43):2101089. [doi: 10.2807/1560-7917.ES.2022.27.43.2101089] [Medline: 36305336]

Edited by KA Clegg; submitted 06.Mar.2026; this is a non–peer-reviewed article;accepted 06.Mar.2026; published 17.Mar.2026.

Please cite as:
Basbouss-Serhal I
When Old Diseases Return: Cholera, Crisis, and Digital Surveillance in Fragile Settings
J Med Internet Res 2026;28:e94818
URL: https://www.jmir.org/2026/1/e94818
doi:10.2196/94818

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 17.Mar.2026.

XSL•FO
RenderX

How Digital Diaries Are Changing Communication and Recovery in the Intensive Care Unit and Beyond

Jenna Congdon

(J Med Internet Res 2026;28:e92661) doi:10.2196/92661

KEYWORDS

intensive care units; patient diaries; post–intensive care syndrome; mental health; survivors; professional-family relations; communication; trauma-informed care; critical care; patient-centered care; digital health; rehabilitation; recovery of function

Textbox .

Key Takeaways

- Digital intensive care unit (ICU) diaries help patients and families reconstruct "lost time," supporting psychological recovery after critical illness.
- By centralizing updates and humanizing care, digital diaries can improve communication between ICU staff and families while reducing interruptions to clinical work.
- Digital diaries can be a low-burden, high-impact tool supporting human-centered, trauma-informed ICU care long after discharge.

When patients finally leave after a long stay in the intensive care unit (ICU), they may feel as if they've lost weeks of their life. Perhaps they were unconscious, confused, or otherwise unable to process the whirlwind of illness and medical interventions that occurred during their stay. Additionally, clear and consistent communication between patients' family members and medical staff supports improved patient outcomes, but remains a perennial challenge in the ICU due to the complex nature of patients' medical courses.

A digital diary—a secure app-based journal that allows staff and families to log daily events with messages, photos, and videos to chronicle a patient's hospital stay—addresses both these concerns. As they recover, patients can reference the diary to process the complex experiences they may have had in the ICU. Digital diaries can improve mental and physical health outcomes for patients, foster stronger communication between staff and family members, and serve as a powerful tool to support patients' recovery and a poignant reminder of the love and support they received along the way.

Digital Diaries: A Valuable Record of Time Lost

Jurriaan van Rijswijk, MSc, founder of Games for Health [1], a company that develops app-based games to improve health outcomes, saw an opportunity to both improve staff-family communication and support patients' mental health after an ICU stay. Together with medical staff and researchers, he developed the Post-ICU digital diary app [2].

The app was initially launched at Erasmus University Medical Center in Rotterdam, Netherlands, to bridge the gap between patients and distant family members during the COVID-19 pandemic. It has since expanded; at the time of this writing, the diaries are in use across 24 hospitals, primarily in Europe.

Post–Intensive Care Syndrome

Patients often experience a constellation of mental health challenges after an ICU stay, particularly if their hospitalization was long or complicated [3]. In one study, 46% of patients experienced anxiety symptoms following their hospital stay. In addition, 40% reported feeling depressed, 22% were diagnosed with posttraumatic stress disorder (PTSD), and 18% of patients faced all three. Morbidity was 47% higher for patients experiencing depression compared to patients without depression in the two years following discharge from the ICU [4].

This decline in mental health, often accompanied by negative physical and cognitive changes, is referred to as post–intensive care syndrome (PICS) [4]. This syndrome not only makes it difficult to return to daily activities, but can threaten long-term survival and quality of life [5]. Ongoing support and intervention after discharge is needed to treat these issues and ensure patients' maximal recovery.

After digital diary implementation, patients had fewer instances of post-ICU anxiety and depression, and their loved ones experienced lower rates of PTSD after their hospital stay [6]. For some ICU survivors, rereading their digital diary helped them cope with the aftermath of a traumatic illness [7].

Improving Staff-Family Communication

During an ICU stay, all focus goes to the patient, and communication between staff and family members can be complex—on one side, the worried loved ones; on the other, busy nurses and doctors. When details get missed or misunderstood, both sides end up frustrated, and patient outcomes may suffer. When communication between both parties is centralized, families are more engaged in their loved ones' care and medical errors may be reduced [8].

XSL•FO
RenderX

The diaries streamline communication between staff and family members, saving time for everyone and ensuring that family members feel involved even when they can't be physically present. The app is easily activated by a nurse through the patient's medical record. Nursing staff write in the diary every morning, helping to humanize care with heartfelt, nonclinical notes alongside medical details. Family members know they can simply check the app for updates on their loved one rather than call the hospital, leading to fewer patient care interruptions.

Notes on Adoption and Challenges

Feedback from staff and families has been largely positive.

"Families love the solution," Jurriaan says, and he notes that "the app has a feedback button...they mostly use it to compliment us."

Concerns from family members largely center around two topics: feeling unsure of what to write and struggling to print the diary as a keepsake or convert it to a PDF to share. The former has been addressed with simple in-app writing prompts, the latter with straightforward tech support.

"It is a super simple app," Jurriaan shares. "It has three buttons, one to write and to add pictures, one to read, and one button for information."

This decreases barriers for users with low levels of digital literacy [7]. Furthermore, the diary remains accessible after discharge, making it easy to share with therapists or other caregivers.

Hospital administrators sometimes report qualms regarding data privacy. With the Post-ICU app, the family maintains control of who can access the diary, and careful security measures are in place to protect patient data. Jurriaan points out that compared to a physical journal, "...it's much safer. With a physical journal, you can't control who touches it, who reads it. It is a hygiene problem and a security problem. A digital journal is safer."

And while nurses have expressed concern about lacking the time to write entries, they nonetheless report that the diaries are a worthwhile addition to patient care, fostering a sense of human connection [9].

To address potential burden, the implementation process should be participatory and include end user medical staff to ensure ongoing relevance as workflows adapt [10]. Responsibility for diary entries can be shared across medical disciplines to increase overall participation.

For some patients, digital diaries can provoke sadness and loss; in one study, roughly one-third of patients surveyed reported rereading their ICU diary as a particularly distressing and painful experience [11]. Staff and families should use careful judgment around how and when to share the diaries with recovering patients.

Digital Diaries Support Human-Centered ICU Care

Jurriaan shares the story of an on-call nurse he worked with, who attended a beloved patient's funeral:

> *"...to her surprise, some of the [recordings] she made in the diary for that person were being read in front of everyone. How beautiful...it was shared with the entire audience during the funeral. I believe that's incredible, that although the patient died, it is still of use...for the family."*

Digital diaries are more than clinical documentation. They are trauma-informed care tools that can serve as powerful sources of ongoing support and inspiration for families, medical providers, and patients long after discharge. They can help patients feel seen as people rather than clinical cases, validate feelings at a time when emotions are high, and provide staff with a respectful way to connect with their patients on a human level.

Recovery goes far beyond just physical survival. Regaining lost memories and processing traumatic events can allow patients and their loved ones to face life after the ICU with resilience. For patients who lost weeks of their lives, a digital diary doesn't just fill in the gaps; it gives the story back to them. Thoughtfully implemented, digital diaries represent an elegant intervention supporting total-person healing during the ICU stay and beyond.

Conflicts of Interest

None declared.

References

1. Games for health. URL: https://gamesforhealth.net/en/ [accessed 2026-02-18]
2. Post-ICU. URL: https://post-icu.com/ [accessed 2026-02-18]
3. Chang CS, Tsai FJ, Liao CH. Associations between elevated rates of depression, anxiety, and PTSD among ICU survivors and increased mortality and readmissions. Brain Behav 2025 Feb;15(2):e70319. [doi: 10.1002/brb3.70319] [Medline: 39957084]
4. Hatch R, Young D, Barber V, Griffiths J, Harrison DA, Watkinson P. Anxiety, depression and post traumatic stress disorder after critical illness: a UK-wide prospective cohort study. Crit Care 2018 Nov 23;22(1):310. [doi: 10.1186/s13054-018-2223-6] [Medline: 30466485]
5. Voiriot G, Oualha M, Pierre A, et al. Chronic critical illness and post-intensive care syndrome: from pathophysiology to clinical challenges. Ann Intensive Care 2022 Jul 2;12(1):58. [doi: 10.1186/s13613-022-01038-0] [Medline: 35779142]

XSL•FO
RenderX

6. McIlroy PA, King RS, Garrouste-Orgeas M, Tabah A, Ramanan M. The effect of ICU diaries on psychological outcomes and quality of life of survivors of critical illness and their relatives: a systematic review and meta-analysis. Crit Care Med 2019 Feb;47(2):273-279. [doi: 10.1097/CCM.0000000000003547] [Medline: 30431494]
7. van Mol MMC, Tummers N, Leerentveld C, Tieben R, Buise M. The usability of a digital diary from the perspectives of intensive care patients' relatives: A pilot study. Nurs Crit Care 2024 Nov;29(6):1280-1289. [doi: 10.1111/nicc.12990] [Medline: 37897098]
8. Khan A, Spector ND, Baird JD, et al. Patient safety after implementation of a coproduced family centered communication programme: multicenter before and after intervention study. BMJ 2018 Dec 5;363:k4764. [doi: 10.1136/bmj.k4764] [Medline: 30518517]
9. Haakma T, Tieben R, Sleven B, Buise M, van Mol M. Experiences of nurses with an innovative digital diary intervention in the intensive care unit: a qualitative exploration. Intensive Crit Care Nurs 2022 Jun;70:103197. [doi: 10.1016/j.iccn.2022.103197] [Medline: 35090798]
10. Schol CMA, van Mol MMC, Ista E. Developing implementation strategies for digital ICU diaries targeting ICU professionals: an implementation mapping approach. Implement Sci Commun 2025 Aug 7;6(1):85. [doi: 10.1186/s43058-025-00767-0] [Medline: 40775358]
11. Flahault C, Trosdorf M, Sonrier M, et al. ICU survivors experience of ICU diaries: an ancillary qualitative analysis of the ICU diary study. Crit Care Explor 2021 May;3(5):e0384. [doi: 10.1097/CCE.0000000000000384] [Medline: 34036266]

Edited by KA Clegg; submitted 01.Feb.2026; this is a non–peer-reviewed article;accepted 01.Feb.2026; published 20.Feb.2026.

Please cite as:
Congdon J
How Digital Diaries Are Changing Communication and Recovery in the Intensive Care Unit and Beyond
J Med Internet Res 2026;28:e92661
URL: https://www.jmir.org/2026/1/e92661
doi:10.2196/92661

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Narrative Medicine: Translating Stories Into Medical Education and Better Patient Care

Candice Marie Sage

(J Med Internet Res 2026;28:e93037) doi:10.2196/93037

KEYWORDS

digital narrative medicine; medical education; patient-centered care; health equity; vulnerable populations; palliative care; narratives as topic; telemedicine; empathy; mental health; digital health; professional-patient relations

Textbox .

Key Takeaways

- Digital narrative medicine can function as a humanizing bridge between patients and their care teams, enhancing clinical care and fostering patient empowerment through advocacy and treatment co-design.
- Ethical and methodological considerations and refinements will need to be addressed as the approach evolves and expands.

Candice Marie Sage, PhD, is a storyteller and researcher with over 20 years of experience managing interdisciplinary research in Canada and internationally. She is an expert in knowledge co-creation, translation, and exchange. In this piece, she provides an overview of some of the applications of digital narrative medicine, kicking off a series in which she will explore digital narrative medicine stories and applications in depth.

About four decades ago, narrative medicine (NM) emerged as an important diagnostic tool [1], one that treats illness as a human experience and not just a biological one. NM is an approach to clinical care that recognizes the importance of lived experiences in health care, training clinicians in skills like reflective writing and authentic communication with patients [2]. By listening to, analyzing, and integrating patients' stories along with standard test results and traditional tools, clinicians can gain a deeper understanding of their patients, improving health outcomes [3,4].

Today, with an arsenal of novel tools and technologies, *digital* narrative medicine (DNM) provides new ways to capture and incorporate patient stories to enhance meaning-making [5]. It augments traditional NM by using digital tools to improve the accessibility, scalability, and inclusivity of narrative practices [2].

With applications from medical education to clinical interventions and across diverse populations, DNM enhances how health care professionals understand, relate to, and communicate with patients, their families, and caregivers, while providing a platform for empowerment through patient advocacy and treatment co-design [6].

DNM as Knowledge Translation

As care moves away from direct interactions between patients and health care professionals toward remote and asynchronous care, DNM functions as a bridge [2]. It helps build a humanizing connection between patients and their care teams, grounding care in lived experience and making knowledge translation a reciprocal process.

An increasing number of tools support this shift in how knowledge flows, allowing patient stories to be captured and used in more flexible, creative ways. Examples include multimedia podcasts and curated videos, leveraging artificial intelligence to translate stories into training simulation scenarios and clinical decision support [7], and interactive games that allow patients, their families, and their care teams to engage and interact dynamically. These DNM tools foster greater empathy between patients, their families, and clinicians, leading to more comprehensive, patient-centered care [8].

Applications in Medical Education

Before the digital age, medical education was based on attending different hospital rotations and observing, with depth of learning dependent on the chance clinical presentations and procedures taking place at that given time.

Today, DNM's narrative-based training and interactive storytelling systems provide virtually endless scenarios for trainees to experience and practice their diagnostic and decision-making skills. Trainees' reflective writing exercises are captured electronically in "e-portfolios" and then evaluated by trainers to nurture narrative competence, defined as "the ability to absorb, interpret, and act upon patient stories" [2]. Exposure to the authentic voices of patients—and, increasingly, family and caregivers—helps provide not only information about medical care, but also context around culture and other social determinants of health—enhancing trainees' empathy, communication skills, and professional identity [9].

One such example is the Out of Darkness multimedia project in Toronto, Canada [10]. People living with bipolar disorder were video-graphed, capturing their recovery journeys and lived experiences at home and in their neighborhoods. Psychiatry trainees were exposed to very realistic aspects of bipolar illness,

XSL•FO
RenderX

like what it feels like to take lithium, undergo electroconvulsive therapy, or experience fear of relapse and suicidality. By watching these emotionally evocative short films, learners gained treatment-relevant insights and reported increased empathy and compassion.

Applications in Clinical Intervention

Beyond training, DNM enhances patient-centered care by providing health professionals with highly contextual accounts of patients' psychological, emotional, social, and spiritual lived experiences [11]. This information can be collected and stored in medical records, analyzed with artificial intelligence tools, and made more accessible to both patients and practitioners in real time or longitudinally. Holistic, wraparound care is enhanced when professional teams collaborate by sharing insights and meaning-making about patient stories [12].

Cancer care digital storytelling, for example, has existed for more than a decade through initiatives like the Canadian Cancer Society's 2018 partnership with digital storyteller Mike Lang, which provided a platform for cancer patients to tell their stories and reflect on their journeys [13]. Earlier, in 2015, I coordinated a Cancer Care Ontario project, collecting and analyzing the lived experiences of patients and practitioners. More deeply understanding each group's needs and experiences led us to shift 5-year follow-up care for colorectal and breast cancer survivors from oncology specialists to local primary care providers. Follow-up surveys showed this new care plan improved each group's experiences, freeing specialists up to focus on active oncology care and reducing patients' travel time [14].

DNM is also used in palliative and end-of-life care and research [15] and can help facilitate difficult discussions about advanced care and end-of-life planning [16]. For example, I led an award-winning palliative care redesign project at Kingston General Hospital in 2018; the key to our success was incorporating members of the hospital's Patient and Family Engagement Committee early on and throughout the redesign. We learned about more than palliative care—we also learned what it was like to navigate the health care system as a patient with multiple specialists, and about the burdens carried by both patients and their at-home caregivers. One patient was so ashamed of his illness that he forbade his wife from telling family or friends about the situation. She held that burden all by herself until after he died. Stories like this informed modifications to disclosure in a patient and family communications DNM intervention.

Game-based DNM interventions in clinical and mental health care are growing, bridging treatment, engagement, and implementation gaps. For teens and young adults with cancer, for example, a video game intervention greatly improved patients' compliance with their treatment regimens through increased knowledge and self-efficacy [17].

Applications for Marginalized Populations

DNM can be particularly valuable for advancing health equity, and it can be an empowering tool for marginalized populations like youth [18], racialized groups [19], people with diverse sexual orientations and gender identities [20], Indigenous peoples [21], and people experiencing mental health [22] and substance use challenges [23].

The Mental Health Story Teller Mob developed by Australia's Aboriginal and Islander Mental Health Initiative is an example of Indigenous-led digital narrative storytelling. It explores emotional and social well-being, coping, and resilience in the face of poverty, racism, stigma, and mental health and substance use challenges [24].

In my work on the Support Not Stigma project [25], we recorded short videos and a documentary for a regional antistigma campaign, capturing and sharing the stories of people living with mental health or substance use challenges—many of whom were also experiencing homelessness—and the workers supporting them. This not only provided a platform for amplifying their voices and educating the public, it also led to the discovery that over 90% of this population had experienced head traumas. This serious medical consideration would not have been known had we not delved into people's stories.

Future Considerations

DNM is a powerful, underutilized approach that is gaining traction. As it continues to develop and expand, there are several important considerations.

New methodologies are needed to better plan and evaluate DNM research, curriculum development, and clinical practice. Ethical considerations—such as the consent of patient and family storytellers, and how their stories are used—will require dialog between patients, health care professionals, researchers, and policymakers. As technology evolves, new methods and tools to further comprehensive, person-centered care will emerge and require refinement.

We're only just scratching the surface of the realm of possibilities DNM presents. With careful planning, thorough research, and thoughtful implementation, we can realize its potential for deepening and improving care for patients from all walks of life.

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Fioretti C, Mazzocco K, Riva S, Oliveri S, Masiero M, Pravettoni G. Research studies on patients' illness experience using the Narrative Medicine approach: a systematic review. BMJ Open 2016 Jul 14;6(7):e011220. [doi: 10.1136/bmjopen-2016-011220] [Medline: 27417197]
2. Efthymiou E. Integrating digital and narrative medicine in modern healthcare: a systematic review. Med Educ Online 2025 Dec;30(1):2475979. [doi: 10.1080/10872981.2025.2475979] [Medline: 40327849]
3. Allan R, McCann L, Johnson L, Dyson M, Ford J. A systematic review of "equity-focused" game-based learning in the teaching of health staff. Public Health Pract (Oxf) 2024 Jun;7:100462. [doi: 10.1016/j.puhip.2023.100462] [Medline: 38283754]
4. Lybarger K, Bear Don't Walk OJ, Yetisgen M, Uzuner Ö. Advancements in extracting social determinants of health information from narrative text. J Am Med Inform Assoc 2023 Jul 19;30(8):1363-1366. [doi: 10.1093/jamia/ocad121] [Medline: 37466054]
5. Palla I, Turchetti G, Polvani S. Narrative Medicine: theory, clinical practice and education - a scoping review. BMC Health Serv Res 2024 Sep 27;24(1):1116. [doi: 10.1186/s12913-024-11530-x] [Medline: 39334149]
6. Kadestål K. Enhancing patient engagement through co-design: challenges, strategies, and practical insights in digital healthcare technology [Thesis]. : Malmö University; 2024 URL: https://mau.diva-portal.org/smash/record.jsf?pid=diva2%3A1906127&dswid=6704 [accessed 2026-02-20]
7. Yu H, Zhou J, Li L, et al. Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. Commun Med (Lond) 2025 Dec 19;6(1):27. [doi: 10.1038/s43856-025-01283-x] [Medline: 41420084]
8. Mazzoli Smith L, Villar F, Wendel S. Narrative-based learning for person-centred healthcare: the Caring Stories learning framework. Med Humanit 2023 Dec 19;49(4):583-592. [doi: 10.1136/medhum-2022-012530] [Medline: 37208190]
9. Moreau KA, Eady K, Sikora L, Horsley T. Digital storytelling in health professions education: a systematic review. BMC Med Educ 2018 Sep 10;18(1):208. [doi: 10.1186/s12909-018-1320-1] [Medline: 30200945]
10. Zaretsky A, Rodney H, Morassaei S, et al. The Out of Darkness project: innovative multimedia digital storytelling in psychiatry. Acad Psychiatry 2022 Oct;46(5):644-649. [doi: 10.1007/s40596-022-01637-w] [Medline: 35578097]
11. Mazzoli Smith L, Hardy P, Thompson K, Westwood L. Transformative and therapeutic benefits of digital storytelling: a phenomenological lifeworlds study of Patient Voices participant experiences. Arts Health 2024 Jul 2:1-18. [doi: 10.1080/17533015.2024.2368767] [Medline: 38956989]
12. Massari MG, Di Trani M, Palmieri A, Reho M, Gennaro A. Where science meets stories: a systematic review on Narrative Medicine. Qual Health Res 2025 Nov 5;Nov 5:10497323251387023. [doi: 10.1177/10497323251387023] [Medline: 41192816]
13. Canadian Cancer Society. Mike Lang Stories. URL: https://www.mikelangstories.com/digital-storytelling-specialist/canadian-cancer-society-digital-storytelling-program [accessed 2025-12-16]
14. Christmas CM, Langley H. Well follow-up care for colorectal cancer survivors and their circles of care. In: Improving Coordination Between Primary Care Providers and Oncology Specialists: Cases from Canada – CanIMPACT Casebook: University of Toronto; 2015.
15. Williams L, Gott M, Moeke-Maxwell T, et al. Can digital stories go where palliative care research has never gone before? A descriptive qualitative study exploring the application of an emerging public health research method in an Indigenous palliative care context. BMC Palliat Care 2017 Sep 4;16(1):46. [doi: 10.1186/s12904-017-0216-x] [Medline: 28870189]
16. Ewy D. Using narratives to offer hindsight, insight, and foresight into advance care planning. Caring for the Ages 2025 Jan;26(1):8-9. [doi: 10.1016/j.carage.2024.11.010]
17. Kato PM, Cole SW, Bradlyn AS, Pollock BH. A video game improves behavioral outcomes in adolescents and young adults with cancer: a randomized trial. Pediatrics 2008 Aug;122(2):e305-e317. [doi: 10.1542/peds.2007-3134] [Medline: 18676516]
18. Christmas CM, Khanlou N. Defining youth resilience: a scoping review. Int J Ment Health Addiction 2019 Jun;17(3):731-742. [doi: 10.1007/s11469-018-0002-x]
19. Maragh-Bass AC, Dinh N, Cooney G, et al. The role of digital storytelling methods in promoting health-related outcomes among young adults of color: a systematic review. Mhealth 2025;11:42. [doi: 10.21037/mhealth-24-65] [Medline: 40755940]
20. Sage M, Singer JB, LaMarre A, Rice C. Digital storytelling: tools, techniques, and traditions. 2018. URL: https://ecommons.luc.edu/cgi/viewcontent.cgi?article=1120&context=socialwork_facpubs [accessed 2026-02-19]
21. Hensel JM, Ellard K, Koltek M, Wilson G, Sareen J. Digital health solutions for Indigenous mental well-being. Curr Psychiatry Rep 2019 Jul 1;21(8):68. [doi: 10.1007/s11920-019-1056-6] [Medline: 31263971]
22. De Vecchi N, Kenny A, Dickson-Swift V, Kidd S. How digital storytelling is used in mental health: a scoping review. Int J Ment Health Nurs 2016 Jun;25(3):183-193. [doi: 10.1111/inm.12206] [Medline: 26900000]
23. Christmas CM. Crystal meth, opiates, overdose and housing in Kingston, Ontario: 2021 rapid assessment and response community needs assessment for the integrated care hub. Substance Use Health Network. URL: https://www.substanceusehealth.ca/resource/crystal-meth-opiates-overdose-and-housing-kingston-ontario-2021-rapid-assessment-and [accessed 2026-02-19]

XSL•FO
RenderX

24. Nagel T, Thompson C. AIMHI NT 'Mental Health Story Teller Mob': developing stories in mental health. Australian e-Journal for the Advancement of Mental Health 2007 Jan;6(2):119-124. [doi: 10.5172/jamh.6.2.119]
25. Sage CM. Support Not Stigma. URL: https://supportnotstigma.ca/ [accessed 2025-12-18]

Edited by KA Clegg; submitted 06.Feb.2026; this is a non–peer-reviewed article;accepted 06.Feb.2026; published 20.Feb.2026.

Please cite as:
Sage CM
Digital Narrative Medicine: Translating Stories Into Medical Education and Better Patient Care
J Med Internet Res 2026;28:e93037
URL: https://www.jmir.org/2026/1/e93037
doi:10.2196/93037

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Feb.2026.

XSL•FO
RenderX

How Digital Health Technology Is Extending Independent Living

Jenna Congdon

(J Med Internet Res 2026;28:e93767) doi:10.2196/93767

KEYWORDS

independent living; aged; home safety; telemedicine; wearable electronic devices; accidental falls, prevention & control; medication adherence; aging in place; monitoring, ambulatory

Textbox .

Key Takeaways

- Independent living for older adults is closely tied to safety, dignity, and quality of life, and can often be extended with thoughtfully selected digital health and smart home technologies.
- Wearables, smart home tools, and digital monitoring systems can reduce preventable harm, support chronic disease management, and provide a rapid response in emergencies.
- Clinicians play a critical role in assessing individual needs, addressing barriers to adoption, and aligning technology recommendations with patient goals and values to support aging in place safely.

As a nurse, I frequently hear questions like:

> *"How long can Mom keep taking good care of herself without help?"*
>
> *"How do I keep myself safe while living alone?"*
>
> *"Dad wants to live independently for as long as possible. How do we support him?"*

For many older adults, living independently goes far beyond being a measure of their physical and cognitive health; it can be a point of social pride and a reflection of their personhood. Losing one's independence can be frustrating, but it's a situation that can be preventable with the right care. Technologies such as wearable monitors and smart home devices, when used thoughtfully, can reduce preventable harm, extend independence, and support clinical and personal goals for older adults [1].

What "Safety" Really Means for Older Adults Living Independently

Home safety for older adults is both multifactorial and unique to the individual; no singular solution will work for every person. Safety goes beyond responding to medical emergencies such as strokes, falls, and cardiac events. Digital health or smart home technologies can also address issues such as poor medication adherence, functional decline, cognitive changes, chronic condition management, lack of physical activity, and social isolation [1].

James Giordani, MSW, has made it his mission to help older adults get the most out of potentially life-saving technology. He started his company, Clear Computing, in 2019, which was acquired by parent company AsKevin [2] in 2024. The business offers in-home tech support to people of all ages, but Giordani found himself assisting mostly older clients.

Working with seniors soon became a passion. Giordani went on to get his Master's degree in social work, again specializing in working with older adults. He shares a personal story that helped him realize how technology can augment in-home safety for older adults: "My Nana...fell and broke her hip, and she was down for about 12-16 hours before she got help. [She] couldn't crawl to the phone...had she had a smart assistant, she could have used it to call for help."

Technology is always available and always on call. While it's not necessarily a substitute for face-to-face check-ins, it is a powerfully supportive tool for day-to-day cares and emergencies alike.

Core Needs and Technologies for Safe Independent Living

Choosing the Right Tools

Medical care providers may consider recommending technologies for older adults wishing to remain independent. This is especially relevant as an increasing number of older people are choosing to receive care at home [3]. As clinicians, we can encourage safety and better health by assisting older patients to understand which technologies may best support their unique needs.

Preventative care checkups and in-home visits are excellent moments for assessment of potential needs. In-home safety encompasses many aspects of personal health, social networks, and daily life, but broadly, these needs can be grouped into several large categories: (1) fall prevention and rapid response, (2) cardiovascular monitoring, (3) medication adherence and chronic disease management, and (4) interpersonal safety and security.

These categories provide a starting point for medical providers to perform a safety assessment and recommend appropriate technologies to support their patients' independence.

XSL•FO
RenderX

Fall Prevention and Rapid Response

Technologies to prevent or reduce harm from falls may include smart bulbs or switches for lighting, wearable fall sensors, and motion or impact detectors. Giordani recommends programming smart bulbs to turn on or off at key times, such as just before dusk, to improve vision in otherwise low-light environments.

A smartwatch or smart home device may be used to call for help in the event that someone is unable to get to a phone. There are also artificial intelligence devices designed specifically to detect falls, the results of which appear promising [4].

Cardiovascular Monitoring

Cardiac events, including arrhythmias or arrests, are most common in older populations [5]. Wearable devices such as watches, chest straps, or clothing help to detect these events and alert medical personnel. Due to the varied nature of these devices across the market, clinical integration is often impractical due to differences in data collection, accuracy, and interoperability challenges. Nevertheless, they may still provide valuable data for the adjustment of medical treatment and offer an early alert system for emergencies [6].

Medication Adherence and Chronic Disease Management

Poor medication adherence or polypharmacy creates the potential for serious harm. Tools such as smart pill dispensers or automatic reminders that can be shared with caregivers can help support proper medication dosing and timing [7].

Technologies such as glucose monitors can assist in chronic disease management, especially when paired with a smartphone app designed to help the user understand their glucose targets [8].

Devices like exercise trackers or smartwatches encourage increased physical activity, leading to better overall health and longevity [9].

Interpersonal Safety and Security

Giordani's favorite recommendation is one most might not consider. He shares, "I recommend that every older adult have an ad blocker on their phones and their computers. Not only do they reduce the cognitive load of seeing all those advertisements… they [help with] cyber security and scam prevention. Scams are often delivered to older adults through ads."

Technologies like ad blockers, smart home devices, and smart locks can enhance safety and personal and financial security for older adults living alone.

New Technology Adoption in Older Adults

Many older adults are very open to using new devices to support their health [10]—but to be effective, these personal health technologies must support the users' goals, be accessible and affordable, and be easy to use. Support systems must also be in place to encourage adoption and ongoing use, such as a helpline or tech assistance from a service such as AsKevin [11].

Giordani suggests assisting older adults to set up any new device. For older adults, he says, this process isn't necessarily intuitive. "I always recommend, if you're giving a device as a gift, you have to set it up for them."

Recommendations for Clinicians

Our patients' ability to live independently is more than a social preference; it's a real-world reflection of their clinical outcomes that is worth supporting with evidence-based tools.

Many older adults welcome the use of technology to support long-term health and independence. Talk with your patients about their individual needs and goals, and assess barriers such as cost and tech support before suggesting which devices may best improve their quality of life.

When aligned with clinical judgment and patient values, digital health tools help older adults lead longer, safer, and more autonomous lives. While technology doesn't replace human care, it can extend our reach as we strive to enhance every aspect of our patients' health.

Updated March 02, 2026: An earlier version of this article incorrectly referred to the company "AsKevin" as "Ask Kevin". This has been corrected.

Conflicts of Interest

None declared.

References

1. Facchinetti G, Petrucci G, Albanesi B, De Marinis MG, Piredda M. Can smart home technologies help older adults manage their chronic condition? A systematic literature review. Int J Environ Res Public Health 2023 Jan 10;20(2):1205. [doi: 10.3390/ijerph20021205] [Medline: 36673957]
2. asKevin. URL: https://www.askevin.com/ [accessed 2026-02-23]
3. Lette M, Ambugo E, Hagen T, Nijpels G, Baan C, De Bruin S. Addressing safety in integrated care programs for older people living at home: a systematic review. Int J Integr Care 2019;19(4):593. [doi: 10.5334/ijic.s3593]
4. Mohan D, Al-Hamid DZ, Chong PHJ, et al. Artificial intelligence and IoT in elderly fall prevention: a review. IEEE Sensors J 2024;24(4):4181-4198. [doi: 10.1109/JSEN.2023.3344605]

XSL•FO
RenderX

5. Meyer M, Arnold A, Stein T, Niemöller U, Tanislav C, Erkapic D. Arrhythmias among older adults receiving comprehensive geriatric care: prevalence and associated factors. Clin Pract 2024 Jan 4;14(1):132-147. [doi: 10.3390/clinpract14010011] [Medline: 38248435]
6. Xintarakou A, Sousonis V, Asvestas D, Vardas PE, Tzeis S. Remote cardiac rhythm monitoring in the era of smart wearables: present assets and future perspectives. Front Cardiovasc Med 2022;9:853614. [doi: 10.3389/fcvm.2022.853614] [Medline: 35299975]
7. Faisal S, Samoth D, Aslam Y, et al. Key features of smart medication adherence products: updated scoping review. JMIR Aging 2023 Dec 19;6:e50990. [doi: 10.2196/50990] [Medline: 38113067]
8. Yang Y, Lee EY, Kim HS, Lee SH, Yoon KH, Cho JH. Effect of a mobile phone-based glucose-monitoring and feedback system for type 2 diabetes management in multiple primary care clinic settings: cluster randomized controlled trial. JMIR mHealth uHealth 2020 Feb 26;8(2):e16266. [doi: 10.2196/16266] [Medline: 32130172]
9. Wu S, Li G, Du L, Chen S, Zhang X, He Q. The effectiveness of wearable activity trackers for increasing physical activity and reducing sedentary time in older adults: a systematic review and meta-analysis. Digit Health 2023;9:20552076231176705. [doi: 10.1177/20552076231176705] [Medline: 37252261]
10. Wang Y, Lu L, Zhang R, Ma Y, Zhao S, Liang C. The willingness to continue using wearable devices among the elderly: SEM and FsQCA analysis. BMC Med Inform Decis Mak 2023 Oct 16;23(1):218. [doi: 10.1186/s12911-023-02336-8] [Medline: 37845659]
11. Moore K, O'Shea E, Kenny L, et al. Older adults' experiences with using wearable devices: qualitative systematic review and meta-synthesis. JMIR mHealth uHealth 2021 Jun 3;9(6):e23832. [doi: 10.2196/23832] [Medline: 34081020]

Edited by KA Clegg; submitted 18.Feb.2026; this is a non–peer-reviewed article;accepted 18.Feb.2026; published 24.Feb.2026.

Please cite as:
Congdon J
How Digital Health Technology Is Extending Independent Living
J Med Internet Res 2026;28:e93767
URL: https://www.jmir.org/2026/1/e93767
doi:10.2196/93767

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Feb.2026.

XSL•FO
RenderX

What Health Care Organizations Have Learned From Telecommunication Outages

Catharine Solomon

(J Med Internet Res 2026;28:e91456) doi:10.2196/91456

KEYWORDS

telecommunications; telecommunications standards; health services administration; health care quality; access; evaluation; infrastructure; risk management; emergency preparedness; health care resiliency

Textbox .

Key Takeaways

- Health care facilities' ability to provide patient care can be threatened by large-scale telecommunication outages, requiring providers to build additional redundancy into their own infrastructure.
- While steps have been taken to formalize accountability for some telecommunication providers (such as, in Canada, via the Memorandum of Understanding on Telecommunications Reliability), gaps remain.

Telecommunication (telecom) breakdowns at massive scales are fresh in recent memory; in Canada, the Rogers Communications outage of July 2022 cut off over 12 million users' wireless and wire-line services for over 24 hours [1], and in 2024, a single faulty update to CrowdStrike's cybersecurity software caused an estimated 8.5 million Microsoft systems worldwide to crash [2]. In the past several months alone, widespread outages from internet infrastructure providers, like Cloudflare and Amazon Web Services (AWS), have alerted users to the fragility of an internet that relies on a handful of companies to function [3].

Health care facilities, which use telecom for everything from faxing prescriptions, to accessing electronic health records (EHRs), to virtual care, have to navigate these breakdowns under an additional unique pressure: their patients' well-being on the line. In the months and years following major outages, health care institutions have taken steps to mitigate the threat of service disruption to patient care.

The Rogers Outage

On the morning of July 8, 2022, Canadians woke up to a phone and internet outage hitting nearly every sector. While upgrading their IP core network, Rogers staff had removed a policy filter, which allowed IP routing data to flood and then crash the core routers, leaving user traffic unable to reach their destinations and over 12 million Rogers users without service [1].

Many of these users were health care providers and organizations, as well as patients trying to access health care. Family medicine clinics found themselves suddenly unable to carry out basic functions, like faxing prescriptions, scheduling appointments, and attending virtual care appointments [4]. The outage disproportionately affected health care services in remote locations, which often rely on network access to consult with specialists; this was compounded by the loss of connectivity across other health system partners [4]. In Toronto, immunization clinics lost access to the provincial COVID-19 vaccination management system, long-term care facilities lost access to their residents' EHRs [5], and hospitals had to scramble to mitigate the outage's effects on patient care.

According to Michael Garron Hospital's (MGH's) Chief Information Officer Amelia Hoyt, the brunt of the outage at MGH fell on corporate staff, who use Rogers cellular services for their work phones. This loss of corporate cellular mobility threatened patient care only indirectly, but physicians who relied on Rogers for their own mobile phones also couldn't be reached. MGH leadership gathered an emergency command center to develop a plan for how to navigate the outage.

"We did need to come together to figure out how to make sure that we didn't have delayed communications that inadvertently would affect patient care," says Hoyt. The solution in the moment was to identify affected staff and, by using a fan-out list, set alternate communication routes.

The outage allowed Hoyt and her team to identify a gap in their IT strategy and then fix it. "We had no other redundancy for if cell services are down." Redundancy—the strategy of duplicating system components or functions—works by keeping fail-safes and backup components at the ready.

"What we decided to do was invest in eSIMs from a different cellular provider for corporate mobile devices," she explains, "loaded onto select individuals' phones....If Rogers went down again, we would be able to activate this alternate provider's eSIM, and that would allow for continuity."

The CrowdStrike Outage

On July 19, 2024, University Health Network (UHN) staff trying to connect with clinical systems were met instead with blue error screens. A defective content configuration update released by the cybersecurity company CrowdStrike had caused millions of Windows hosts worldwide to crash, resulting in disruption

XSL•FO
RenderX

of services across multiple industries and organizations, including government agencies, airlines, health care facilities, and even 911 connectivity in some areas [6,7]. In the health care sector, the outage disrupted everything from EHR access, to telemonitoring, to operational management, to research [8].

According to Carl Virtanen—chief technology officer for UHN Digital (UHN's IT department)—the UHN's downtime procedures for switching to manual documentation helped mitigate disruptions to patient care, though several nonurgent appointments had to be rescheduled.

"UHN was able to prevent more significant technological disruption by pausing software updates and limiting the number of devices affected by the CrowdStrike outage," Virtanen says, noting that UHN Digital's protocols, which include prioritizing critical areas (such as emergency departments, intensive care units, and operating rooms) when responding to technological problems, have not been updated since. "All critical clinical systems were back up and running normally by the end of the day the outage began."

Importance of Redundancy

The telecom ecosystem is deeply interconnected. For complex critical systems, like health care, that rely on telecoms [9], interconnectivity facilitates seamless and efficient data sharing across a network of health system partners. But these interconnected systems can be disrupted on a large scale by a single point of failure in the infrastructure upon which they depend. So, how can disruptions be avoided?

Virtanen says that the best strategy for avoiding downtime is already, ideally, in use: "It is critical to have redundancy layers built into technical infrastructure to maintain operational efficacy in the event of service disruptions, such as the CrowdStrike outage."

On the most basic level—the physical infrastructure underpinning telecom services—a lack of redundancy can result in, for example, a beaver taking out nearly half of a town's internet connection by chewing through the network's single fiber-optic cable [10]. The 2022 Rogers outage itself exemplified the peril of putting all of one's technological eggs in one basket; since the Rogers management network relied on the very IP core network that had crashed, staff struggled to communicate with one another and remotely investigate the issue, delaying repairs for hours [1].

Health care facilities' preparations for outages, such as MGH's eSIM investment and UHN's downtime protocol of switching to paper records, act as another redundancy layer. In a perfect world with an unlimited budget, Hoyt says that her team would invest in additional redundancy for all IT services at MGH; at one point, they had even toyed with the idea of low-tech backups, like walkie-talkies. The solution is practical but not without its drawbacks; maintaining redundant systems can be cost-intensive, particularly for health care facilities with fewer resources [4].

Who's Accountable?

The fallout and appropriate redress of telecom outages have often been framed in terms of monetary loss, which can be reimbursed through vendor refunds and credits [5]. But when essential services, like health care, rely on network access to function efficiently and effectively, a framework that accounts only for economic impact falls short.

"Telecommunication is really foundational," says Hoyt. "I liken it to 911 service—there are some standards that are expected, that you should be able to require some of these telecom companies to follow."

Steps have been taken to hold Canadian telecom companies accountable for the essential services they provide; in September 2022, thirteen major telecom companies signed the Memorandum of Understanding (MOU) on Telecommunications Reliability [11]. The signatories of this MOU agreed to provide one another with emergency roaming and assistance in the event of a major outage, building ready-to-go redundancy into Canadian telecom services.

The MOU is "a positive step," according to Hoyt. "What will be interesting to see is how this agreement is put into practice if and when the next unplanned outage occurs."

What about the rest of the telecom ecosystem? Health care organizations' ability to access and relay information is only as stable as the infrastructure they use to communicate, and unlike health care providers who can be held liable for medical malpractice or medical devices that are held to rigorous standards of reliability and accuracy, many telecom services—particularly newer cloud-based models, like software as a service, that boast cost-effective scalability for clients—have frequent downtimes without consequence.

For health care organizations, using third-party services carries an inherent risk: tech disruptions that can't be addressed in-house. "When the infrastructure is under local IT's control, like my department, it's something my team would have more direct accountability for," says Hoyt. "If AWS is out, then you're kind of stuck between a rock and a hard place. There's nothing really a local IT team can do about that because we actually don't have any control over it."

Health care organizations must weigh this risk in choosing whether and when to rely on such services. Unless these services are reliable enough to meet health care standards, the scalability boost may not be worth it.

As more and more of our essential systems rely upon telecom infrastructure, redundancy and resiliency strategies are key to avoiding more outages that, at best, create delays and administrative burden in health care and, at worst, can lead to real patient harm. In an interconnected telecom ecosystem, all providers, companies, and organizations have the opportunity and the responsibility to do their part in keeping critical services running smoothly.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Xona Partners Inc. Assessment of Rogers networks for resiliency and reliability following the 8 July 2022 outage – executive summary. Canadian Radio-television and Telecommunications Commission. 2024 Jul. URL: https://crtc.gc.ca/eng/publications/reports/xona2024.htm [accessed 2026-01-27]
2. Weston D. Helping our customers through the CrowdStrike outage. Official Microsoft Blog. 2024 Jul 20. URL: https://blogs.microsoft.com/blog/2024/07/20/helping-our-customers-through-the-crowdstrike-outage/ [accessed 2026-01-27]
3. London tech expert explains why the internet blew up this week (temporarily). CBC News. 2025 Nov 22. URL: https://www.cbc.ca/news/canada/london/london-tech-expert-explains-why-the-internet-blew-up-this-week-temporarily-9.6987955 [accessed 2026-01-27]
4. Li T, Tran C, Yung A, Thomas A. Family physicians and the nationwide communications service outage in Canada. Canadian Family Physician. 2022 Dec 9. URL: https://www.cfp.ca/news/2022/12/09/12-09 [accessed 2026-01-27]
5. Attachment 3: Rogers Communications July 2022 outage impact: economic, operational & potential for ConnectTO to mitigate future events. City of Toronto. URL: https://www.toronto.ca/legdocs/mmis/2023/ex/bgrd/backgroundfile-239381.pdf [accessed 2026-01-27]
6. Gallagher JC, Pechtol CL. CrowdStrike IT outage: impacts to public safety systems and considerations for Congress. Congress.gov. 2024 Dec 4. URL: https://www.congress.gov/crs-product/IF12717 [accessed 2026-01-27]
7. Ellingson C. Global software glitch affected Edmonton police 9-1-1 calls overnight Friday. CTV News. 2024 Jul 19. URL: https://www.ctvnews.ca/edmonton/article/global-software-glitch-affected-edmonton-police-9-1-1-calls-overnight-friday/ [accessed 2026-01-27]
8. Tully JL, Rao S, Straw I, et al. Patient care technology disruptions associated with the CrowdStrike outage. JAMA Netw Open 2025 Jul 1;8(7):e2530226. [doi: 10.1001/jamanetworkopen.2025.30226] [Medline: 40682764]
9. Howell B. Beyond infrastructure: internet ecosystem resilience and the public good. Telecomm Policy 2025 Aug;49(7):102998. [doi: 10.1016/j.telpol.2025.102998]
10. Hundreds lose internet service in northern B.C. after beaver chews through cable. CBC News. 2021 Apr 25. URL: https://www.cbc.ca/news/canada/british-columbia/beaver-internet-down-tumbler-ridge-1.6001594 [accessed 2026-01-27]
11. Memorandum of Understanding on Telecommunications Reliability. Innovation, Science and Economic Development Canada. URL: https://ised-isde.canada.ca/site/ised/en/memorandum-understanding-telecommunications-reliability [accessed 2026-01-27]

Edited by KA Clegg; submitted 14.Jan.2026; this is a non–peer-reviewed article;accepted 14.Jan.2026; published 03.Feb.2026.

Please cite as:
Solomon C
What Health Care Organizations Have Learned From Telecommunication Outages
J Med Internet Res 2026;28:e91456
URL: https://www.jmir.org/2026/1/e91456
doi:10.2196/91456

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 3.Feb.2026.

XSL•FO
RenderX

A Frontline Worker's Take on Hybrid Care Implementation in the Hospital Setting

Jenna Congdon

(J Med Internet Res 2026;28:e90879) doi:10.2196/90879

KEYWORDS

telemedicine; delivery of health care, integrated; digital health; hospital medicine - organization & administration; health personnel attitudes; hybrid care implementation

Textbox .

Key Takeaways

- Hybrid care can enhance patient outcomes and workflow efficiency, but only when thoughtfully implemented with frontline staff input.
- Variability in hospital resources, patient populations, and infrastructure leads to inconsistent adoption and effectiveness of hybrid care models.
- Without adequate support, training, and realistic workflow planning, hybrid technologies risk adding burden to bedside staff and generating ineffective spending rather than improving care.

Hybrid care in the hospital setting—combining in-person care with telehealth services—has increased dramatically since 2020 [1,2]. During the COVID-19 pandemic, frontline workers relied on remote access to care for patients from afar. In the succeeding years, hybrid care use has remained much higher than prepandemic levels [1]. At different points in my own practice as an intensive care nurse, depending on its implementation, hybrid care technology has been a distracting hindrance to patient care or has greatly supported my ability to treat patients in critical condition.

For other staff working at the bedside, reactions are similarly mixed: while some report safer patient care and more efficient workflows, others feel that hospitals now push telehealth technology as an ineffectual fix for staffing shortages and organizational issues.

Hybrid Care on the Ground

The day-to-day realities of in-hospital hybrid care use are varied. Location, patient demographics, and facility-to-facility differences in budgets and priorities play a part in determining the specific implementation, delivery models, and rate of adoption by both patients and staff for these modes of care [3].

Examples include using video calls to patch in specialty providers for emergencies and consultations, providing remote physician rounding for rural or underserved communities, and leveraging digital apps for patient communication and video visits.

Additional uses include monitoring confused or unstable patients via video; supporting nursing workflows with remote nurse assistance; facilitating admissions, discharges, and patient education; and offering a way for families to connect virtually with loved ones if they are unable to visit in person [1].

These tools represent a unique opportunity to augment patient care and improve safety measures. However, without collaborative implementation and ongoing support services, some frontline staff express concern that more technology simply adds to their already burdensome workload [4-6].

When Hybrid Care Saves the Day

At its best, hybrid care technology can save lives and ease stress for bedside staff, patients, and their loved ones.

Gwen*, a nurse in a postanesthesia care unit in Portland, Oregon, recalls a time when her patient experienced a stroke soon after surgery: "In that situation, we were able to use the telehealth neurologist during a code stroke. It worked really smoothly, and the other staff and I felt more confident in that emergency. That patient absolutely received better care because we were able to talk to a neurologist immediately." Gwen's experience demonstrates how faster access to specialty care via a remote provider can dramatically improve patient outcomes.

Avah*, a nurse who works in the postpartum unit of a hospital in Madison, Wisconsin, reported: "Our NICU [neonatal intensive care unit] has cameras that let parents see their baby from home." She cites this use of telehealth technology as an important bonding tool for new parents of hospitalized infants.

Other benefits to hybrid care include fewer admissions and readmissions, reduced length of hospital stay, improved chronic disease management, enhanced patient safety, reduced health system costs, better integration of care services, better adherence to recommended best practices, and increased efficiency in provider workflows [4,5].

When Hybrid Care Gets in the Way

Nurses frequently become the point person for new in-hospital technology, including hybrid care tools. When patients struggle to log in to a phone app for a video appointment or the camera used for physician consults needs to be set up, it often falls to

XSL•FO
RenderX

the nurse to first troubleshoot and then call for assistance if needed. They become the liaison for patient concerns and digital snafus, taking time away from patient care activities and adding to an already hectic workload.

Other concerns relayed by frontline health care workers regarding newly integrated hybrid care models include technical issues such as unstable Wi-Fi or poor equipment maintenance, patient and staff competency in adopting new technologies, and lack of ongoing support as uses for hybrid care modalities evolve [4,5]. Additional technology may also create interruptions in the workflow and degrade the quality of patient-nurse interactions [6].

Rhonda*, an intensive care unit charge nurse at a large Midwestern hospital, explains that her unit recently installed electronic intensive care unit (eICU) cameras in each patient room. These devices are meant to bring in ICU-trained registered nurses (RNs) via video to provide nursing support. “We rarely use them, and we never asked for this kind of device. Instead of being a help, they are a hindrance. The eICU staff call us to ask questions that are irrelevant to patient care. It wastes time. I’m sure [the cameras] were expensive, and they don’t help patient care in any way.”

She worries that this follows a recent trend at her facility of cutting staff and attempting to replace in-person RNs with nurses who can only watch from the other side of a screen.

These challenges and concerns suggest that additional technology is not a replacement for proper staffing or efficient organizational workflows.

Making the Most of Hybrid Care Tools

If in-hospital hybrid care models are to be effective, nurses and other frontline staff must be involved in the entire decision-making process, not brought in after adoption [5,6]. The choice of hybrid care technologies needs to be decided based on real-life clinical needs and centered around realistic patient behaviors. If patients struggle to use an app or cannot understand a provider on the other end of a video call, the technology becomes useless, burdensome to staff, or worse, an active impediment to the quality of patient care.

Designing for flexibility is crucial as technology develops, patient needs shift, and health care continues to see rapid change [7]. Ongoing and easily accessible tech support for both patients and staff should be included to increase use and reduce errors, work-arounds, and frustrating workflow disruptions [6]. While staff are often receptive to new technologies becoming part of their daily workflow, needs and perceptions vary by facility and department, highlighting the need for a tailored approach that includes bedside staff’s voices [8].

In-Hospital Hybrid Care Succeeds When Frontline Staff Are Involved

While hybrid care can be a powerful tool, everyday complications show gaps that hospital leadership and policymakers sometimes overlook. Without the involvement of these frontline workers, hospitals risk spending budget dollars on technology that simply does not serve its full purpose.

By continually seeking out and incorporating the input of frontline staff, health systems and hospital administrators can make realistic, informed, and effective decisions about which tools will increase safety, efficiency, staff retention, and patient experience in their facilities. Administrators should survey frontline staff about their daily challenges and perspectives on which models of hybrid care would best address staff and patient needs before purchasing, policy change, or implementation begins. Additionally, nursing focus groups may provide an opportunity for unit-specific representation so that hybrid care execution can be tailored to departmental needs.

Hybrid care tools cannot replace the human touch in health care, but they certainly have the power to create better outcomes for patients and lighten staff workloads. Hospital systems should invest in thorough research and open conversation with staff to generate flexible and human-centered hybrid care models that best serve the people under their care.

**Names changed to protect anonymity.*

Conflicts of Interest

None declared.

References

1. Hehman MC, Fontenot NM, Drake GK, Musgrove RS. Leveraging digital technology in nursing. Health Emerg Disaster Nurs 2023;10(1):41-45. [doi: 10.24298/hedn.2022-0014]
2. OECD. The COVID-19 pandemic and the future of telemedicine. : OECD Publishing; 2023 Jan 17. [doi: 10.1787/ac8b0a27-en]
3. Chen J, Amaize A, Barath D. Evaluating telehealth adoption and related barriers among hospitals located in rural and urban areas. J Rural Health 2021 Sep;37(4):801-811. [doi: 10.1111/jrh.12534] [Medline: 33180363]
4. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023 Sep 18;6(1):161. [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
5. Kumari AA, Wani TA, Liem M, Boyd J, Khan UR. Advancing regional and remote health care with virtual hospital implementation: rapid review. JMIR Hum Factors 2025 Jun 3;12:e64582. [doi: 10.2196/64582] [Medline: 40460425]

XSL•FO
RenderX

6. Hassell A, Morelock B, Greeson J, Thomson A, Varty M. Lessons learned from systemwide implementation of a patient technology technician role to manage bedside nursing technology. Nurs Manage 2025;56(3):11-17. [doi: 10.1097/nmg.0000000000000226] [Medline: 39928026]
7. Pilosof NP, Barrett M, Oborn E, Barkai G, Zimlichman E, Segal G. Designing for flexibility in hybrid care services: lessons learned from a pilot in an internal medicine unit. Front Med Technol 2023;5:1223002. [doi: 10.3389/fmedt.2023.1223002] [Medline: 38053662]
8. Choi H, Tak SH, Song YA, Park J. Nurses' perspectives on the adoption of new smart technologies for patient care: focus group interviews. BMC Health Serv Res 2025 Mar 18;25(1):391. [doi: 10.1186/s12913-025-12578-z] [Medline: 40098040]

Edited by KA Clegg; submitted 05.Jan.2026; this is a non–peer-reviewed article;accepted 05.Jan.2026; published 26.Jan.2026.

Please cite as:
Congdon J
A Frontline Worker's Take on Hybrid Care Implementation in the Hospital Setting
J Med Internet Res 2026;28:e90879
URL: https://www.jmir.org/2026/1/e90879
doi:10.2196/90879

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Jan.2026.

XSL•FO
RenderX

WHOOP, There It Is: Lessons From WHOOP's FDA Warning Letter

Blythe Karow

(J Med Internet Res 2026;28:e90882) doi:10.2196/90882

KEYWORDS

wearable electronic devices; United States Food and Drug Administration; device approval; health devices; digital health; WHOOP; Apple; Aktiia; medical device regulation; blood pressure monitoring, ambulatory; legislation; jurisprudence

Textbox .

Key Takeaways

- The US Food and Drug Administration's warning letter to WHOOP and ensuing battle have highlighted a critical moment in the wearables space as the US regulatory framework has been evolving in real time.
- New lines have been drawn. The 2026 Food and Drug Administration wellness guidance has established new norms around the ability to noninvasively detect and relay information to users.
- Ultimately, however, the wellness exemption remains poorly defined for companies seeking clear guidance on user interfaces and how they can interpret and relay data to users while remaining a wellness device.

Blythe Karow [1] *is a strategic management consultant and fractional executive with a biomedical engineering bachelor's degree from Johns Hopkins University, an MBA from the UVA Darden School of Business, and over twenty years of experience leading, advising, and consulting as a medtech, digital health, wearables, and regulated healthtech expert. She is the founder of The Karow Advisory Group and author of The Device Files* [2]*, a Substack where she shares strategy, insights, and practical tips for getting devices to market. She has covered the evolution of the WHOOP/US Food and Drug Administration (FDA) regulatory conflict since it began and shares her reflections here.*

In July 2025, the fitness wearable company WHOOP received a warning letter from the FDA [3] for their "Blood Pressure Insights" (BPI) feature, noting that it was clearly a medical device requiring proper 510K clearance. For anyone following the wearables space, this wasn't surprising. WHOOP, like many wearables companies, has slowly been evolving from performance optimization and wellness toward monitoring and diagnosing medical conditions [4].

The warning did not just kick off another regulatory spat. It raised important questions and considerations for the future of wearables and other medical devices, spurring debates across the medical device and wearables communities for 6 months, and culminating in the release of new FDA guidance on January 6, 2026 [5]. In this analysis, I review both the FDA's and WHOOP's original arguments around the regulatory status of the BPI feature, looking at the evolution of the conflict up until now.

Where the July 2025 FDA Warning Letter Arguments Fell Short

The "Inherent Association" Doctrine

The FDA's original position hinged on "inherent association"—that blood pressure measurement is "inherently associated with the diagnosis of hypo- and hypertension" regardless of disclaimers or intended use [3].

In my view, this legal theory was always a bit of a stretch. If any biometric that *can* be used for diagnosis automatically *becomes* a medical device, then the wellness device exemption Congress created in the 21st Century Cures Act [6,7]—that software intended for general wellness and not medical diagnosis or treatment is not considered a medical device—becomes meaningless. By this logic, bathroom scales would be medical devices because weight is "inherently associated" with obesity diagnosis, but you can walk into any store and buy a scale without FDA oversight.

In my original analysis, I argued that the real conversation shouldn't be whether all blood pressure measurements are inherently diagnostic, but rather how the FDA's wellness guidelines could and should be revised to allow biomarker collection while providing clear guidance on handling detected extremes.

The Language Trap

The FDA found specific language from WHOOP that crossed into diagnostic territory, such as "higher blood pressure may be an indicator of poor sleep" and materials indicating that the feature was intended to identify "higher blood pressure" [3].

The distinction between "higher" and "high" blood pressure matters here. "Higher" describes a relative change from an

XSL•FO
RenderX

individual's baseline and should be acceptable for wellness devices. However, "high blood pressure" suggests comparison with clinical diagnostic thresholds. While WHOOP appears to have used "higher," the FDA appeared to have interpreted this as "high blood pressure."

The Color-Coding Absurdity

The FDA objected to WHOOP's user insights interface, claiming the color-coded (green/yellow/orange) readings imply medical interpretation, but most fitness wearables use similar color-coding schemes: for example, Garmin shows heart rate zones in green, yellow, and red, and Apple Watch displays activity rings with traffic-light logic.

If color-coding implies medical interpretation, are we going to start requiring submissions from manufacturers of street lights? This may be a tongue-in-cheek comparison, but the question remains: when does color-coding imply diagnostic categorization?

The Misuse Argument

The FDA cited "evidence of individuals using BPI to monitor their hypertension" as a sign that WHOOP had overstepped the line. This original argument concerned me for two reasons.

First, using anecdotal misuse by some users to justify broad regulatory classification, especially when they provided no details about scale, frequency, or how this evidence was collected, would have set a concerning precedent. Second, people use fitness trackers for wellness purposes all the time, and those wellness issues often connect to more serious health conditions. Sleep quality can impact anxiety or cognitive health [8] and getting your 10,000 steps can help with overall cardiovascular wellness [9]. That shouldn't make wellness device manufacturers liable for medical device indications.

While we want to be careful of products purposefully trying to skirt a medical device pathway, some off-label use is probably acceptable when it leads people to healthier lifestyles that could improve more serious conditions. And what's the harm in someone with hypertension wanting to get an idea of their blood pressure, especially when the next step is probably to seek proper medical testing if they see warning signs?

Where WHOOP Might Have Been Wrong

WHOOP might have had legitimate grievances about the FDA's approach to their BPI feature, but they also seemed to escalate the situation by (1) not adjusting their marketing or features claims early on when discussions started in May 2025, (2) continuing to leave up the "offensive" product features while arguing with the FDA, (3) making a buzzworthy end run to RFK Jr instead of working with the agency [10], and (4) having their CEO, Will Ahmed, making public statements arguing their "legal" case.

The Fatal Phrase

Buried in the FDA warning is language that likely sealed WHOOP's fate: their website originally described the blood pressure feature as delivering "medical-grade health & performance insights" [4,11].*

That single term was regulatory poison. In the FDA's world, "medical-grade insights" implies clinical accuracy, validation standards, and regulatory approval, whether intended or not.

My original argument: you can have a wellness device that collects medical-grade *data*—in fact, this is preferred, because algorithms perform best when data are high-quality [12-14]. The problem was that WHOOP had originally used the term in reference to providing medical-grade *insights*, implying diagnostic-level blood pressure measurement.

If WHOOP had proposed changing this language in May, the situation may have been resolved sooner. However, the back-and-forth, while refusing adjustments, seemed to have gotten the FDA's ire up. Additionally, with Aktiia and Apple subsequently receiving clearance for blood pressure and hypertension claims [15,16], it likely became harder for the FDA to give leniency.

The Media Defense

Shortly after the FDA warning letter, WHOOP CEO Will Ahmed posted counterarguments on LinkedIn [17], making good points but ultimately overrelying on disclaimers to protect the company from oversight. In medtech, just telling consumers not to use a device a certain way often isn't enough, especially if it contradicts other marketing claims.

There were also weaknesses in the arguments Ahmed made during his CNBC appearance at the end of August 2025 [18]:

1. He appeared to believe that wellness exemptions override the core intent of the FDA's mission to protect patients and provide medical device oversight, stating that "Intended use is what matters—and the law agrees." But once you make claims that your device does something a medical device does, you can't just say, "Yeah, but we're just here for wellness purposes," because guess what? So are all medical devices!
2. He mentioned that their blood pressure feature requires calibration with a blood pressure cuff (a medical device), as though acknowledging reliance on a medical device made it clearer that WHOOP itself isn't performing a medical function, but if you need a medical device to calibrate your "wellness" feature, that suggests you're doing medical device work.
3. He claimed that WHOOP should set a new precedent because regulations could thwart tech innovation, blind to an entire medical device industry that has been innovating for decades while following the rules.
4. He claimed that WHOOP members "overwhelmingly support us fighting for [BPI]." This attempt to position himself as defending consumer access read as a deflection rather than acknowledging that the company had created a regulatory mess.

The Marketing Smoking Gun

Most damning to me when reviewing this debate is how WHOOP still structures subscription tiers, separating blood pressure insights from general "health monitoring" and bundling it specifically with FDA-cleared medical features in the highest tier [19]. They're explicitly categorizing blood pressure insights

XSL•FO
RenderX

alongside actual medical device features, telling consumers through product positioning that blood pressure monitoring belongs in the medical device category. It feels a bit like a magician's misdirection—it implies medical device credibility to the consumer, while denying the need for medical device status to the FDA.

New Year, New Guidance

Ahmed's claim that the FDA was "on the wrong side of the law" [18] was somewhat remarkable given the FDA's enforcement position. WHOOP appealed the formal warning letter but continued to market much the same way throughout 2025. This could have brought additional serious legal trouble [20]; the FDA had multiple escalation options, including court orders to force WHOOP to disable blood pressure features while pursuing appeals.

I was fairly confident that WHOOP's attempted lobbying to RFK Jr wouldn't be effective in swaying the FDA, given the precedent set by the Aktiia and Apple clearances and additional pressure from public scrutiny. But I was wrong.

In the first week of 2026, the FDA responded with new guidance that very clearly states that they will allow blood pressure measurements via optical sensing to remain in the wellness category as long as companies make no claims to having "medical-grade" data or insights [5]. The new guidance seems an obvious win for WHOOP, and it also seems clear from current FDA Commissioner Dr Marty Makary's comments during his January 6 interview on Fox Business News [21] that the FDA is moving to relax and clarify control and oversight in the wellness space. He stated, "We want to let companies know, with very clear guidance, that if their device or software is simply providing information, they can do that without FDA regulation."

While the new guidance provides clear direction on whether WHOOP can market BPI, there remains a lack of clarity in a few areas. The guidance does not specifically address all of the concerns originally outlined in the warning letter (for example, it does not address whether you can use color coding in a user interface) and WHOOP does not seem to have adjusted their interface or marketing beyond specifically removing claims of "medical" or "clinical grade" in reference to BPI [4,11]. And, although multiple examples are provided on how to describe a product or its features without crossing into medical device territory, there is very little to guide developers on appropriate user interface and alerts, leaving other technologies with a great deal of gray area to navigate in the future.

Ultimately, this case has shone a spotlight on some of the weaker elements of the wellness carve out and will likely continue to play a major role in how the FDA decides to evolve in this new world of wearables, artificial intelligence, and the quantified self. We're watching the real-time evolution of the wellness doctrine; while I expect some continued upheaval, a line appears to be forming around collecting versus interpreting user data as the new distinction between wellness and medical devices. Innovators would do well to keep a close eye on the wellness space as it continues to evolve.

**Note: Following the January 2026 FDA guidance, the WHOOP web pages were revised, and the original marketing language is no longer retrievable in the current versions.*

Conflicts of Interest

None declared.

References

1. Blythe Karow. LinkedIn. URL: https://www.linkedin.com/in/blythe-karow [accessed 2026-01-09]
2. The Device Files: From Concept to Commercialization | Substack. URL: https://blythekarow.substack.com [accessed 2026-01-09]
3. Warning letter: WHOOP, Inc. — MARCS-CMS 709755. US Food and Drug Administration. 2025 Jul 14. URL: https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/warning-letters/whoop-inc-709755-07142025 [accessed 2026-01-09]
4. Introducing WHOOP 5.0 and WHOOP MG: the all-new ways to unlock better health, fitness, and longevity. WHOOP. 2024 Aug 27. URL: https://www.whoop.com/ca/en/thelocker/introducing-whoop-5-0-and-whoop-mg/?srsltid=AfmBOooko9VhmvxAMxBEVp7yawtVNVXZuwrNpCtWHHmRifUzfgK-i7ST [accessed 2026-01-15]
5. Guidance Document. General wellness: policy for low risk devices. Guidance for industry and Food and Drug Administration staff. Docket number: FDA-2014-N-1039. US Food and Drug Administration. 2026 Jan 6. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/general-wellness-policy-low-risk-devices [accessed 2026-01-12]
6. 114th Congress, United States. 21st Century Cures Act, Public Law 114-255 114th Congress 130 Stat 1033. 2016 Dec 13 URL: https://www.congress.gov/114/plaws/publ255/PLAW-114publ255.pdf [accessed 2026-01-09]
7. Guidance Document. General wellness: policy for low risk devices. Guidance for industry and Food and Drug Administration staff. Docket number: FDA-2014-N-1039. US Food and Drug Administration. 2019 Sep. URL: https://web.archive.org/web/20251217125918/https://www.fda.gov/regulatory-information/search-fda-guidance-documents/general-wellness-policy-low-risk-devices [accessed 2025-12-24]
8. Cox RC, Olatunji BO. Sleep in the anxiety-related disorders: a meta-analysis of subjective and objective research. Sleep Med Rev 2020 Jun;51:101282. [doi: 10.1016/j.smrv.2020.101282] [Medline: 32109832]

XSL•FO
RenderX

9. Lavie CJ, German CA, Sanchis-Gomar F. Reducing mortality and cardiovascular disease. J Am Coll Cardiol 2023 Oct;82(15):1495-1498. [doi: 10.1016/j.jacc.2023.08.007] [Medline: 37676197]
10. Will Ahmed | Instagram. URL: https://www.instagram.com/p/DKASUbOuia8 [accessed 2026-01-09]
11. WHOOP delivers innovative blood pressure insights for a deeper look at your well-being. WHOOP. 2025 May 8. URL: https://www.whoop.com/ca/en/thelocker/blood-pressure-insights/?srsltid=AfmBOoqbWjK5nKtTNCN1yfpSmglPk6fNdVm32i0wK0fDyRRRTELRbeFC [accessed 2026-01-12]
12. Schwabe D, Becker K, Seyferth M, Klaß A, Schaeffter T. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. NPJ Digit Med 2024 Aug 3;7(1):203. [doi: 10.1038/s41746-024-01196-4] [Medline: 39097662]
13. Cho S, Ensari I, Weng C, Kahn MG, Natarajan K. Factors affecting the quality of person-generated wearable device data and associated challenges: rapid systematic review. JMIR Mhealth Uhealth 2021 Mar 19;9(3):e20738. [doi: 10.2196/20738] [Medline: 33739294]
14. Hearn J, Van den Eynde J, Chinni B, et al. Data quality degradation on prediction models generated from continuous activity and heart rate monitoring: exploratory analysis using simulation. JMIR Cardio 2023 May 3;7(1):e40524. [doi: 10.2196/40524] [Medline: 37133921]
15. 510(k) Premarket Notification K250415: Aktiia G0 Blood Pressure Monitoring System. US Food and Drug Administration. 2025 May 15. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K250415 [accessed 2026-01-09]
16. 510(k) Premarket Notification K250507: Hypertension Notification Feature (HTNF). US Food and Drug Administration. 2025 Sep 11. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K250507 [accessed 2026-01-09]
17. Ahmed W. The FDA just sent WHOOP a warning letter about our Blood Pressure Insights feature. LinkedIn. 2025 Jul 16. URL: https://www.linkedin.com/posts/willahmed_wellness-innovation-regulation-activity-7351264571933745156-IJw4 [accessed 2026-01-09]
18. Ahmed W. Whoop CEO Will Ahmed on FDA warning: our blood pressure insight is intended for wellness use only. CNBC Television. 2025 Aug 14. URL: https://www.cnbc.com/video/2025/08/14/whoop-ceo-will-ahmed-on-fda-warning-our-blood-pressure-insight-is-intended-for-wellness-use-only.html [accessed 2026-01-09]
19. Memberships made for you. WHOOP. URL: https://www.whoop.com/ca/en/membership/ [accessed 2026-01-12]
20. Rowe v Whoop, Inc, no3:25cv09910 (ND Cal Nov 18, 2025). ClassAction.org. URL: https://www.classaction.org/media/rowe-v-whoop-inc-complaint.pdf [accessed 2026-01-09]
21. FDA sides with innovation over regulation in MAJOR AI healthcare shift. Fox Business News YouTube page. 2026 Jan 6. URL: https://www.youtube.com/watch?v=OxR6bHDOUvY [accessed 2026-01-12]

Abbreviations

BPI: Blood Pressure Insights
FDA: US Food and Drug Administration

Edited by KA Clegg; submitted 05.Jan.2026; this is a non–peer-reviewed article;accepted 05.Jan.2026; published 21.Jan.2026.

Please cite as:
Karow B
WHOOP, There It Is: Lessons From WHOOP's FDA Warning Letter
J Med Internet Res 2026;28:e90882
URL: https://www.jmir.org/2026/1/e90882
doi:10.2196/90882

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.Jan.2026.

XSL•FO
RenderX

UnitedXR Europe 2025: Aligning Health Care Extended Reality

Jose Ferrer Costa

(J Med Internet Res 2026;28:e90727) doi:10.2196/90727

KEYWORDS

virtual reality; augmented reality; diffusion of innovation; health policy; biomedical technology; interprofessional relations; UnitedXR Europe; Stereopsia; Augmented World Expo

Textbox .

Key Takeaways

- UnitedXR Europe 2025 highlighted that health care extended reality (XR) is no longer constrained by technical capability but by alignment between industry, academic evidence, and clinical governance.
- Persistent gaps remain between how success, risk, and readiness are defined by developers, researchers, and health care professionals.
- Formats that enable direct cross-stakeholder dialogue may be as critical as technological advances for translating XR potential into routine clinical practice.

Extended reality (XR) is not new to health care, but over the past decade, its use has become more widespread in medical training, rehabilitation, pain management, and mental health care, supported by a growing body of evidence [1-4]. Despite clear gains in technical maturity, the adoption of XR in health care remains uneven. Immersive tools are increasingly capable, yet their integration into routine clinical practice appears to depend less on technical performance than on organizational readiness, governance, and professional acceptance—a pattern well described in implementation research [5,6].

UnitedXR Europe 2025 offered a concentrated view of this tension and an effort to resolve it. The event marked the first joint edition of two historically distinct strands of the European XR ecosystem. Stereopsia—anchored in Brussels since 2009—has long served as a meeting point for immersive research, cultural production, and policy dialogue. In parallel, the Augmented World Expo evolved into a global, industry-led platform focused on enterprise deployment and market scale.

Their integration under UnitedXR Europe has brought these traditions into direct contact, highlighting gaps and areas of convergence, as well as creating a shared space for dialogue between industry, academia, health care, and policy actors.

What the Program Revealed About Health Care XR Readiness

UnitedXR Europe 2025 brought together more than 125 exhibitors across 14 parallel tracks, including a dedicated Healthcare, Pharma, & Wellbeing track. Across the event's agenda, health care XR appeared to be entering an infrastructure phase, positioned not as a peripheral demonstration but as a maturing vertical confronted with questions of scale, integration, and long-term sustainability.

Beyond formal sessions, interaction extended into curated environments, such as the European Market for Immersive Creativity and business-to-business matchmaking between developers, researchers, and institutional actors. Roundtables, workshops, start-up pitch competitions, and the European XR Awards Gala reinforced the event's dual role as a space for exchange and an indicator of technical maturity.

This shift was visible on the expo floor, where novelty gave way to utility. Across the XR ecosystem, major headset vendors relied on real-world implementation partners to demonstrate use cases rather than on product-centric stands, with a dedicated demonstration pavilion hosted by the International Virtual Reality Healthcare Association. Within this context, hands-on demonstrations, such as those by VirtualiSurg [7], illustrated how VR is being embedded into established training pathways, particularly when combined with modular haptic systems, such as those provided by SenseGlove [8].

The scientific program reflected a similar maturation. As noted by Oliver Schreer, PhD, track chair, submissions increased from barely a dozen in previous editions to over 45 this year, accompanied by a clear shift toward more application-oriented research.

Moving between tracks revealed a persistent asymmetry. Enterprise discussions were supported by a settled language of scaling, procurement, and return on investment. Health care discussions, by contrast, were anchored in regulation, safety, and professional accountability. Side by side, these perspectives exposed a central tension: while the XR industry largely seems prepared to scale, health care systems are still negotiating why, and under what conditions, that scale should occur.

XR as a Coordination Challenge

What emerged from these observations was a coordination problem. The remaining barriers to health care XR adoption lie less in hardware or software performance than in the absence of shared decision-making structures.

This gap surfaced repeatedly in discussions about adoption criteria. In one interactive panel within the health care track,

XSL•FO
RenderX

participants were asked to rank factors that guided XR implementation. Priorities clearly diverged, with safety treated as a nonnegotiable baseline by some and as one consideration among many by others. The exchange was brief but revealing, with high audience participation.

Figure WL1. Audience engagement during an interactive panel session. Faces have been blurred to protect participant privacy. Photograph credit: Sonya Seddarasan, track chair for Healthcare, Pharma, & Wellbeing at UnitedXR Europe 2025.



What this exchange made visible was that XR initiatives often stall not because immersive systems fail to perform but because organizations lack an agreed framework for guiding early-stage decisions about readiness for clinical use. A recurring takeaway from the session was the growing consensus that health care XR must move beyond repeated proofs of concept. Without a shared way to assess safety, credibility, and contextual fit before implementation begins, pilots tend to accumulate without a clear pathway toward sustained, real-world integration.

Format Matters for Cross-Stakeholder Dialogue

Points of agreement and divergence became most visible during interactive panel sessions. For example, open discussion forced participants to make their criteria explicit and revealed how differently evidence, risk, and readiness were understood across institutional and professional contexts.

The organizers appeared attentive to this dynamic. Sonya Haskins—Augmented World Expo's head of programming—noted that future editions may place greater emphasis on roundtable formats intended to support cross-disciplinary negotiation. Alexandra Gérard—codirector of UnitedXR Europe—similarly pointed out the need to broaden participation within the health care track, including participation by patients and frontline practitioners with direct insight into care delivery.

For health care XR, where adoption depends on trust and legitimacy as much as it does on performance metrics, creating conditions for this kind of dialogue may be as consequential as any technical advance.

Immersion as a Clinical Responsibility

Health care discussions at UnitedXR Europe reflected a growing recognition that immersive technologies carry a different kind of responsibility than that of most digital health tools. XR was framed not merely as a delivery medium but as a technology that directly shapes perception, attention, and embodied experience—a distinction that is well established in prior experimental and neuroscientific work [3,9,10].

This concern surfaced most clearly in informal exchanges. Sonya Haskins described XR creation as "hacking the brain," using the phrase as a metaphor to underscore why immersive systems

XSL•FO
RenderX

cannot be treated as neutral software development. When technologies act directly on perception and experience, questions of intent, safety, and oversight move rapidly to the foreground.

Policy frameworks are beginning to respond, albeit unevenly. The event coincided with the launch of the European Partnership for Virtual Worlds, which explicitly identifies health care as a strategic domain [11]. This marks progress when compared with earlier global digital health strategies, including those of the World Health Organization, where immersive technologies remain largely unnamed despite their growing use in practice [12,13].

Terminology remains a point of friction. As several clinicians noted, framing clinical XR under the umbrella of "virtual worlds" sits uneasily with health care practice. Recent European policy work differentiates professional and health care uses of immersive technologies from consumer-oriented virtual worlds, emphasizing that clinical applications are bounded; task specific; and subject to heightened safety, ethical, and governance requirements [4]. In this sense, UnitedXR functioned as a translation layer, highlighting where policy language must be refined to align with the risk-aware logic of patient care.

Implications

UnitedXR Europe 2025 made clear that health care XR has moved beyond the phase of technical proof. What remains unresolved is not what the technology can do but how it is integrated into the social, ethical, and organizational fabric of health care. From my perspective as a clinician-researcher involved in real-world XR deployment, these tensions are familiar from practice and, importantly, addressable.

The challenge ahead is one of alignment; synchronizing industrial velocity with clinical deliberation; and narrowing the vocabulary gap between policymakers, developers, and practitioners.

As this event demonstrated, that work rarely happens through polished presentations alone. It happens in shared spaces where assumptions are tested, priorities collide, and the real conditions for responsible adoption begin to take shape.

Conflicts of Interest

None declared.

References

1. Zuo G, Wang R, Wan C, Zhang Z, Zhang S, Yang W. Unveiling the evolution of virtual reality in medicine: a bibliometric analysis of research hotspots and trends over the past 12 years. Healthcare (Basel) 2024 Jun 26;12(13):1266. [doi: 10.3390/healthcare12131266] [Medline: 38998801]
2. Lie SS, Helle N, Sletteland NV, Vikman MD, Bonsaksen T. Implementation of virtual reality in health professions education: scoping review. JMIR Med Educ 2023 Jan 24;9:e41589. [doi: 10.2196/41589] [Medline: 36692934]
3. Grassini S, Laumann K. Immersive visual technologies and human health. Presented at: ECCE 2021: European Conference on Cognitive Ergonomics 2021; Apr 26-29, 2021; Siena, Italy p. 1-6. [doi: 10.1145/3452853.3452856]
4. Directorate-General for Communications Networks, Content and Technology (European Commission), Gabaliņa R, Migals A, Zepcan A. Virtual Worlds: How Do They Affect Our Health and Well-Being: Publications Office of the European Union; 2025. [doi: 10.2759/4342263]
5. Kouijzer MMTE, Kip H, Bouman YHA, Kelders SM. Implementation of virtual reality in healthcare: a scoping review on the implementation process of virtual reality in various healthcare settings. Implement Sci Commun 2023 Jun 16;4(1):67. [doi: 10.1186/s43058-023-00442-2] [Medline: 37328858]
6. Shiner CT, Croker G, McGhee J, Faux SG. Perspectives on the use of virtual reality within a public hospital setting: surveying knowledge, attitudes, and perceived utility among health care professionals. BMC Digit Health 2024 Apr 25;2:18. [doi: 10.1186/s44247-024-00076-x]
7. VirtualiSurg. URL: https://virtualisurg.com/ [accessed 2026-01-13]
8. SenseGlove. URL: https://www.senseglove.com/ [accessed 2026-01-13]
9. Riva G, Wiederhold BK, Mantovani F. Neuroscience of virtual reality: from virtual exposure to embodied medicine. Cyberpsychol Behav Soc Netw 2019 Jan;22(1):82-96. [doi: 10.1089/cyber.2017.29099.gri] [Medline: 30183347]
10. Slater M. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. Philos Trans R Soc Lond B Biol Sci 2009 Dec 12;364(1535):3549-3557. [doi: 10.1098/rstb.2009.0138] [Medline: 19884149]
11. European Commission launches European Partnership for Virtual Worlds. European Commission. 2025 Dec 11. URL: https://digital-strategy.ec.europa.eu/en/news/european-commission-launches-european-partnership-virtual-worlds [accessed 2025-12-18]
12. WHO guideline: recommendations on digital interventions for health system strengthening. World Health Organization. 2019. URL: https://apps.who.int/iris/handle/10665/311941 [accessed 2025-12-18]
13. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2025-12-18]

XSL•FO
RenderX

Edited by KA Clegg; submitted 02.Jan.2026; this is a non–peer-reviewed article;accepted 02.Jan.2026; published 21.Jan.2026.

Please cite as:
Costa JF
UnitedXR Europe 2025: Aligning Health Care Extended Reality
J Med Internet Res 2026;28:e90727
URL: https://www.jmir.org/2026/1/e90727
doi:10.2196/90727

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 21.Jan.2026.

XSL•FO
RenderX

When Lived Experience Designs the Intervention

Trevor van Mierlo

(J Med Internet Res 2026;28:e91371) doi:10.2196/91371

KEYWORDS

mental health; telemedicine; health services accessibility; cultural competency; patient participation; patient-centered care; co-design; lived experience

Textbox .

Key Takeaways

- Co-developing a digital mental health intervention with contributors who have lived experience of displacement revealed how language, meaning, engagement, and design choices can carry unanticipated cultural and emotional risks.
- To ensure safety, trust, and effectiveness on a global scale, interventions must move beyond adaptation to lived experience co-design.

Digital mental health interventions are increasingly deployed across borders, reaching people whose lives and circumstances may differ dramatically from the teams who build them. As these tools proliferate, traditional approaches to cultural adaptation—often focused on translation—have become insufficient.

Translation conveys language; it doesn't always convey *meaning*. And meaning is culturally constructed and emotionally rooted in ways that can be impossible to anticipate from the outside. I recently learned this lesson through co-developing a mental health resilience course with a team of Ukrainian software developers—many of whom were displaced by war, and some who were actively participating in it.

This experience illuminated a frequent blind spot in global digital mental health design: without lived experience embedded in creation—not merely consulted afterward—we risk building tools that are well-intentioned and evidence-based but potentially unsafe.

Digital Mental Health at Scale and Cultural Distance

Scalable, low-cost interventions promise to reach global populations that traditional services cannot, particularly during humanitarian crises, displacement, and conflict. Yet most digital mental health tools are conceived, designed, and tested in relatively stable Western contexts before being deployed elsewhere [1-3].

This development pattern introduces cultural distance at precisely the point where sensitivity matters most. Evidence of clinical efficacy does not ensure cultural legitimacy, trust, or safety [4]. When interventions cross geopolitical and cultural boundaries, they enter environments shaped by historical trauma, media narratives, power asymmetries, and collective memory. These contextual forces rarely appear in design specifications, yet they profoundly shape how digital tools are perceived and used.

The challenge, then, is achieving scale without reproducing blind spots that undermine efficacy.

From Adaptation to Co-Development

Most cross-cultural digital health efforts rely on cultural adaptation frameworks [5]. These typically involve translating content, substituting examples, and adjusting tone to fit a new population. While necessary, these steps assume that meaning is largely transferable and remains intact once linguistic barriers are removed.

Our experience suggests otherwise. When conducted without lived experience embedded in the design process, cultural adaptation risks addressing surface differences while missing deeper layers of meaning.

In contrast, lived-experience co-development treats those with direct experience not as informants, but as co-designers. This approach requires cultural and epistemic humility [6]: acknowledgment that certain insights cannot be inferred, researched, or validated externally. They must be shared by those who live within the context the intervention seeks to address.

Collaboration With Lived-Experience Designers

The collaboration at the center of this article emerged when the Ukrainian team—displaced by war, living in Ukraine, or actively deployed—approached our group at Evolution Health to help build a digital mental health resource for Ukrainians living in displacement and uncertainty.

These developers were active contributors throughout development, shaping decisions about language, tone, pacing, and presentation. Their technical expertise allowed them to participate fully in design discussions, while their lived experience grounded those discussions in their reality. Decisions were no longer guided solely by evidence hierarchies or best practices, but by continuous dialog about how content would

XSL•FO
RenderX

be interpreted emotionally, culturally, and symbolically by people living with war-related disruption.

Figure WL1. Left: original English draft. Right: adapted Ukrainian version.

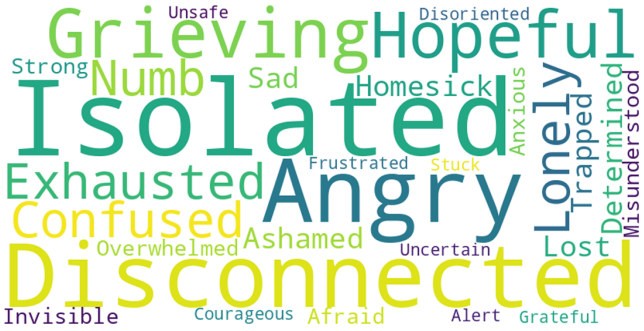



The divergences were striking. Even when words appeared similar, their emotional weight and associations differed substantially. Terms that Western designers associated with isolation, anger, exhaustion, disconnectedness, or optimism carried layered meanings for the Ukrainian team. Some evoked loss. Others echoed bureaucratic language encountered during displacement. Still others were associated with political messaging or wartime media narratives.

Two examples help illustrate the limitations of conventional adaptation and the benefits of co-design.

The first involved a simple vocabulary exercise. The North American team generated a word cloud, and the Ukrainian team was asked to identify words that felt emotionally relevant, neutral, or problematic within the context of displacement and war and to add or remove words accordingly.

The second example involved an introductory video developed using an artificial intelligence (AI)–generated Ukrainian-language voice-over. From a Western perspective, the voice sounded professional, neutral, and supportive—a practical solution that aligned with common digital production practices.

The video was swapped, however, when the Ukrainian team explained that it was likely to evoke distrust. Synthetic Ukrainian voices had become widely associated with Russian disinformation campaigns, where AI-generated Ukrainian speech was routinely used to deliver propaganda.

This risk was invisible to the non-Ukrainian team. Traditional usability testing might have detected disengagement, but only after trust had already been compromised. The issue was not usability, but cultural safety.

These examples underscore two central lessons: that language is not merely semantic, and that trust is built on signals of authenticity, safety, and legitimacy that are culturally situated. Without lived experience in design, even carefully translated content and well-intentioned choices can misfire and risk harm.

What Lived Experience Changed in Practice

Co-development reshaped the intervention in several ways.

Language as Lived Experience

First, meaning consistently took precedence over literal accuracy. Language that was technically correct but emotionally misaligned was revised or discarded. The goal shifted from fidelity to original phrasing toward fidelity to lived interpretation.

Second, engagement strategies were reconsidered. Structures intended to support participation occasionally felt directive or intrusive within a displacement context [7]. Lived experience helped distinguish between supportive guidance and unwanted pressure, particularly for individuals navigating chronic uncertainty and loss of agency.

Finally, harm prevention became an explicit design outcome. Lived-experience contributors identified risks that would not have appeared in formal risk assessments—symbolic associations, emotional triggers, and trust signals embedded in design choices. These insights allowed potential harms to be addressed before deployment, rather than retroactively.

Ethical Responsibilities

There are ethical responsibilities inherent in working with individuals who are actively living through the experiences an intervention seeks to address. Co-development can be experienced as empowering and meaningful, but it also raises questions about psychological safety that appear insufficiently explored in the digital mental health literature [7].

Even when voluntary, participation can involve exposure to emotionally charged language and experiences, increasing the potential for distress. At present, there is limited empirical guidance on how to assess, monitor, or mitigate potential psychological risks for lived-experience contributors in digital intervention development [8].

Designing With, Not For

This collaboration revealed limits in what even experienced digital mental health designers can know from the outside. Lived experience did not simply improve the intervention; it fundamentally reshaped it. It prevented harm, deepened resonance, and challenged assumptions embedded in the original, North American content.

As digital mental health tools continue to scale, lived-experience co-development should be treated as a methodological and

XSL•FO
RenderX

ethical necessity rather than an optional enhancement [9]. Even technically sound, evidence-based design alone cannot prevent contextual misalignment and ensure cultural legitimacy, trust, or safety.

Researchers and designers must consider how lived experience is integrated, compensated, sustained, and supported throughout development. Funders and policymakers should evaluate interventions not only on efficacy, but on how they address cultural meaning and safety.

Ultimately, if digital mental health interventions are to be trusted, culturally legitimate, and safe at global scale, the field must move beyond designing *for* others toward designing *with* them.

Conflicts of Interest

The author is the founder and a strategic advisor of the Evolution Health platform and acts as a consultant for academic and commercial digital mental health initiatives. The views expressed in this article are those of the author.

References

1. Chakrabarti S. Digital psychiatry in low-and-middle-income countries: new developments and the way forward. World J Psychiatry 2024 Mar 19;14(3):350-361. [doi: 10.5498/wjp.v14.i3.350] [Medline: 38617977]
2. Chau LW, Lam RW, Minas H, Hayashi K, Nguyen VC, O'Neil J. Digital health interventions for depression and anxiety in low- and middle-income countries: rapid scoping review. JMIR Ment Health 2025 Aug 22;12:e68296. [doi: 10.2196/68296] [Medline: 40846319]
3. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? Behav Brain Sci 2010 Jun;33(2-3):61-83. [doi: 10.1017/S0140525X0999152X] [Medline: 20550733]
4. Mohr DC, Riper H, Schueller SM. A solution-focused research approach to achieve an implementable revolution in digital mental health. JAMA Psychiatry 2018 Feb 1;75(2):113-114. [doi: 10.1001/jamapsychiatry.2017.3838] [Medline: 29238805]
5. Bernal G, Domenech Rodríguez MM. Cultural Adaptations: Tools for Evidence-Based Practice with Diverse Populations: American Psychological Association; 2012. [doi: 10.1037/13752-000]
6. Mucić D, Hilty DM, Yellowlees PM. Digital mental health toward cross-cultural populations worldwide. In: Digital Mental Health: The Future Is Now: Springer Nature; 2025:175-211. [doi: 10.1007/978-3-031-59936-1_8]
7. Cargo M, Mercer SL. The value and challenges of participatory research: strengthening its practice. Annu Rev Public Health 2008;29:325-350. [doi: 10.1146/annurev.publhealth.29.091307.083824] [Medline: 18173388]
8. Roennfeldt H, Stewart V, Wyder M, et al. Scoping review of co-design in mental health research: essential elements and recommendations. OTJR: Occupational Therapy Journal of Research 2025. [doi: 10.1177/15394492251367259]
9. Muchamore I, Karanikolas P, Gooding P. How Lived Experience Expertise Shapes Research and Development in Digital Mental Health - A Review of Literature and Insights: Wellcome; 2024. [doi: 10.2139/ssrn.4932039]

Edited by KA Clegg; submitted 13.Jan.2026; this is a non–peer-reviewed article;accepted 13.Jan.2026; published 23.Jan.2026.

Please cite as:
van Mierlo T
When Lived Experience Designs the Intervention
J Med Internet Res 2026;28:e91371
URL: https://www.jmir.org/2026/1/e91371
doi:10.2196/91371

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Jan.2026.

XSL•FO
RenderX

Original Paper

Framing the Convergence of One Health and Digital Health in the Global South With a Gender-Sensitive Foresight Perspective: Delphi Study Using Latent Semantic Analysis

Jude Kong[1,2,3,4,5], PhD; Nicola Luigi Bragazzi[1], MPH, MD, PhD

[1]Artificial Intelligence & Mathematical Modeling Lab (AIMMLab), Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
[2]Institute of Health Policy, Management and Evaluation (IHPME), University of Toronto, Toronto, ON, Canada
[3]Department of Mathematics, Bahen Centre for Information Technology, University of Toronto, Toronto, ON, Canada
[4]Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC), Toronto, ON, Canada
[5]Global South Artificial Intelligence for Pandemic and Epidemic Preparedness and Response Network (AI4PEP), Toronto, ON, Canada

Corresponding Author:
Jude Kong, PhD
Artificial Intelligence & Mathematical Modeling Lab (AIMMLab)
Dalla Lana School of Public Health
University of Toronto
155 College Street
Toronto, ON, M5T 3M7
Canada
Phone: 1 4169783868
Email: jude.kong@utoronto.ca

Abstract

Background: The convergence of digital health and One Health represents an emergent paradigm in global health governance. While widely discussed in high-income settings, there is limited understanding of how this convergence is conceptualized in the Global South, particularly when viewed through a gender- and equity-sensitive foresight lens.

Objective: This study aimed to map and classify expert discourse on digital health, One Health, and their convergence in the Global South using latent semantic analysis, with particular attention to structural drivers, emerging issues, weak signals, and gendered patterns of anticipation.

Methods: A 3-round online Delphi survey was conducted with 45 experts from 19 countries across the Global South. Open-ended responses were analyzed using latent semantic analysis and stratified by gender. A foresight framework was applied to categorize topics as structural drivers, emerging issues, or weak signals, based on their temporal persistence, salience, and consensus.

Results: In digital health, structural drivers included the systemic integration of digital technologies into public health systems, strategic alignment, and infrastructure development. Emerging issues comprised the adoption of artificial intelligence, chronic disease management via mobile health, and concerns about digital inclusion and interoperability. Weak signals included feminist digital ethics, trust in digital systems, and relational accountability—more frequently emphasized by female experts. In One Health, structural drivers were centered on intersectoral coordination, ecological integration, and the institutionalization of health-environment frameworks. Emerging issues encompassed anticipatory risk governance, food system sustainability, and the integration of environmental and population-level data. Weak signals included indigenous knowledge systems, subnational antimicrobial resistance governance, and structural underinvestment in ecological public health, with gendered divergence in framing. In the convergence discourse (digital health and One Health), structural drivers focused on the integration of digital surveillance systems, data infrastructures, and health information platforms to operationalize One Health. Emerging issues included climate-triggered system redesign, artificial intelligence and ecological monitoring, and the governance of cross-sectoral data. Weak signals pointed to algorithmic bias in zoonotic prediction, digital sovereignty in environmental health, and feminist critiques of convergence—all thematically rich but peripheral in consensus.

Conclusions: This study revealed a multilayered and gender-influenced foresight architecture shaping the future of digital health and One Health in the Global South. Structural drivers denote maturing domains of implementation, while emerging issues and weak signals highlight latent, often overlooked opportunities and tensions. Incorporating equity-sensitive and gender-aware foresight methods is essential for crafting inclusive and anticipatory health governance strategies.

XSL•FO
RenderX

(J Med Internet Res 2026;28:e78702) doi:10.2196/78702

KEYWORDS

Delphi method; digital health; emerging issues; foresight; gender; Global South; latent semantic analysis; One Health; structural drivers; weak signals

Introduction

In an increasingly interconnected and rapidly changing world, the convergence of demographic aging, epidemiological transitions, and climate-related environmental threats has contributed to a multifaceted global health burden, placing immense strain on already fragile health care systems [1,2]. This burden is especially acute in the Global South, where both communicable and noncommunicable diseases persist amid structural inequalities and constrained resources [3]. Simultaneously, growing awareness of the complex, nonlinear, and interacting determinants of health has spurred a paradigmatic shift in the governance of health—away from siloed, biomedical models toward more holistic, systemic, and integrative frameworks [4].

The One Health paradigm, which acknowledges the interdependence of human, animal, and environmental health, has emerged as a critical response to zoonotic threats, ecosystem degradation, and planetary instability [5,6]. In parallel, the digital transformation of health—encompassed within the digital health agenda—has accelerated, leveraging data-driven technologies to enhance diagnostics, disease surveillance, and care delivery [7]. Their convergence, conceptualized as Digital One Health or One Digital Health, reflects an emergent field characterized by interdisciplinary cooperation and technologically mediated multisectoral coordination [8,9]. However, existing scholarship has largely framed such convergence through a Global North lens, overlooking the digital disparities, infrastructural deficits, and sociotechnological asymmetries that define the Global South [10]. Specifically, studies conducted in high-income and Global North settings have shown that the convergence of digital health and One Health is primarily operationalized through integrated surveillance systems, advanced health and environmental data infrastructures, and artificial intelligence (AI)–enabled early warning platforms, with reported gains in outbreak detection, antimicrobial resistance (AMR) monitoring, and cross-sectoral coordination. These contributions, while methodologically and technologically robust, are predominantly grounded in contexts characterized by mature digital ecosystems, stable governance arrangements, and well-resourced regulatory frameworks, thereby limiting their transferability to the structural, institutional, and equity-related realities of the Global South [8-10].

Moreover, a gender-sensitive or feminist lens is often lacking, thereby marginalizing the situated experiences, epistemologies, and agency of women and gender-diverse actors in shaping health and technology ecosystems [10].

To address these epistemic and representational gaps, this study used a foresight-oriented text mining approach grounded in horizon scanning and the identification of weak signals, emerging issues, and structural drivers [10]. The goal was to map and classify the thematic architecture of the discourse on the convergence of One Health and digital health from a Global South perspective, with particular attention to gendered dynamics, equity-informed framings, and the coproduction of inclusive and pluralistic alternative futures.

Methods

Ethical Considerations

The study protocol was approved by the institutional review board of Dalla Lana School of Public Health, University of Toronto, Ontario, Canada (48350). The study was conducted in accordance with the Declaration of Helsinki, and written informed consent was obtained from all participants.

Expert Recruitment

Experts were recruited through a purposive sampling strategy aimed at capturing diverse perspectives on the convergence of digital health and One Health within the Global South. Invitations were disseminated via professional networks, relevant mailing lists, and institutional affiliations, targeting stakeholders actively engaged in health policy, digital innovation, epidemiology, veterinary medicine, environmental science, and global clinical public health practice.

Selection criteria included demonstrated expertise in at least 1 relevant domain, professional affiliation with an institution based in the Global South, and availability to complete multiple rounds of an online Delphi survey. A total of 45 experts completed all survey rounds (n=3) and were included in the final analysis. The Delphi process was conducted anonymously to encourage candid responses and reduce social desirability bias.

Survey Instrument

The study used a structured, multiround Delphi survey designed to elicit expert consensus on the convergence of digital health and One Health in the Global South. The instrument was developed to capture both sociodemographic diversity and domain-specific insights relevant to foresight-oriented analysis. The survey instrument was developed ad hoc for the purposes of this study, as no validated or standardized questionnaire currently exists to capture expert foresight on the convergence of digital health and One Health, particularly from a Global South and gender-sensitive perspective.

Participants first provided background information, including age, gender, country of birth, country/countries of study, current residence, professional role, areas of expertise, and years of experience in digital health and/or One Health. They were also asked to reflect on the interdisciplinary nature of their work environments.

XSL•FO
RenderX

Each round of the Delphi survey included open-ended prompts focused on present and future opportunities, challenges, and disruptors in both digital health and One Health domains. Specific items addressed strategic drivers, structural barriers, project involvement, and anticipated synergies between the two fields. After each round, responses were thematically analyzed, and aggregated feedback was provided to participants in subsequent rounds, allowing for reflection, refinement, and convergence of expert opinion. The final item in each round requested participants to generate an anonymous identification code to enable matching across rounds. The survey was administered via Google Forms, ensuring broad accessibility across diverse geographies. All items were open-ended to enable narrative elaboration and thematic emergence, aligning with the exploratory aims of the foresight exercise.

The full set of survey items is provided in Multimedia Appendix 1.

Overview of the Methodology: Text Mining and Foresight Exercise

To explore temporal dynamics and strategic orientations in stakeholder discourse on digital health and One Health in the Global South, we used comparative latent semantic analysis (LSA) [11,12] across different foresight-relevant prompts: (1) present implementation, (2) perceived opportunities, (3) perceived challenges, and (4) future-oriented opportunities or challenges. Each response set was subjected to LSA to extract latent topics, with eigenvalues and variance explained used to assess topic salience.

To interpret LSA outputs through a foresight lens, we used a 3-tiered classification framework widely used in foresight and futures research: structural drivers, emerging issues, and weak signals [13], enabling differentiation between deeply embedded system forces, nascent developments gaining stakeholder traction, and low-frequency yet potentially disruptive cues.

Structural drivers were defined as topics that (1) appeared across both present- and future-oriented datasets and (2) accounted for a substantial proportion of explained variance. These themes represent widely shared, resilient priorities. Emerging issues were operationalized as topics observed predominantly or exclusively in future-oriented discourse, with a moderate share of explained variance. These topics signal directional shifts in collective attention and can anticipate inflection points in system trajectories. Weak signals were identified as topics with low variance explained, often surfacing in only 1 temporal frame (present or future) and lacking broader consensus. Although marginal, these signals are analytically significant, as they may indicate overlooked risks, latent needs, or early-stage innovations that could reshape the field.

A gendered lens was applied throughout this classification to assess how male and female respondents may differentially emphasize structural versus emergent priorities. Gender-divergent weak signals, in particular, offer critical insight into situated epistemologies—highlighting perspectives that are often sidelined in dominant discourses.

Overall, this classification supports a foresight-informed interpretation of the LSA results and strengthens the capacity to generate equity-sensitive, gender-responsive scenarios grounded in both continuity and systemic change.

Data Collection and Preprocessing

The dataset comprised the open-ended textual responses provided by the expert panel, compiled into a structured corpus for subsequent text mining. The entries varied in both length and complexity, reflecting the heterogeneous expertise and regional diversity of participants addressing digital health and One Health issues. Prior to computational analysis, the corpus underwent standard preprocessing procedures aimed at improving analytical robustness. Text normalization steps included tokenization, lowercasing, stop-word removal, punctuation stripping, and the elimination of nonalphabetical characters. Stemming and lemmatization were applied to reduce lexical variation while preserving semantic meaning. Importantly, domain-specific terminology was preserved throughout the process to maintain contextual integrity. Preprocessing was implemented using English-language stop word lists and stemming algorithms.

Text Mining Analysis

LSA [11,12] was used to extract conceptual structures embedded within the textual data and to systematically identify latent themes in stakeholder discourse. As a well-established natural language processing technique, LSA enables the reduction of linguistic dimensionality by analyzing patterns of word cooccurrence, making it particularly suitable for foresight-driven text mining applications. A term-document matrix was constructed using a bag-of-words representation, incorporating a minimum term frequency threshold of 2 and a sparsity cutoff of 0.975. Dimensionality reduction was performed using singular value decomposition, which facilitated the identification of principal semantic axes and coherent topic clusters. Specifically, LSA decomposes the term-document matrix into 3 matrices (U, Σ, V^T) via singular value decomposition, where Σ is a diagonal matrix containing the singular values associated with each latent semantic dimension. The squared singular values correspond to the eigenvalues of the latent dimensions and quantify the amount of semantic variance captured by each extracted topic. Eigenvalues were computed and ranked in descending order, without the application of predefined cutoffs or dimensionality reduction thresholds. In this analytical context, eigenvalues represent the relative semantic weight or salience of each latent topic, reflecting the strength and coherence of word cooccurrence patterns across expert responses rather than statistical significance in an inferential sense. Topics associated with higher eigenvalues correspond to dominant and widely shared thematic structures within the discourse, whereas topics associated with lower eigenvalues represent less frequent, more weakly articulated, or emerging semantic patterns. For each topic, the percentage of explained variance was calculated as the ratio between the eigenvalue of that topic and the sum of all retained eigenvalues, multiplied by 100. Explained variance therefore expresses the proportion of total semantic variance in the corpus accounted for by each latent dimension. This measure was used to assess thematic prominence and consensus across responses, rather than to perform hypothesis testing or probabilistic inference. Eigenvalues and explained variance

XSL•FO
RenderX

were interpreted within a foresight-oriented analytical framework. Topics with consistently high explained variance across both present- and future-oriented prompts were classified as structural drivers, reflecting stable and deeply embedded priorities. Topics with moderate explained variance that emerged primarily in future-oriented discourse were classified as emerging issues, indicating evolving areas of collective attention. Topics characterized by low explained variance, limited consensus, or temporal specificity were classified as weak signals. Importantly, low explained variance was not interpreted as analytical irrelevance; rather, in line with horizon-scanning and futures methodologies, such topics were considered potentially indicative of early-stage, underexplored, or disruptive dynamics that may gain relevance over time. This interpretation strategy ensured conceptual alignment between the quantitative outputs generated by LSA and the qualitative, anticipatory objectives of the foresight exercise, while enabling transparent interpretation of the eigenvalues and explained variance.

LSA was first conducted on the overall corpus to extract global semantic patterns. Additional stratified analyses were performed based on participant gender. Following decomposition, the resulting topics were qualitatively interpreted and labeled by 2 independent coders (JK and NLB), who reviewed the most salient terms within each topic vector. Intercoder reliability was ensured through an iterative reconciliation process, with final topic labels established via consensus.

All LSA procedures were executed using the XLSTAT software suite (Lumivero), a commercial statistical package supporting advanced text analytics.

Results

Sample

The expert sample had a mean age of 43.1 (SD 10.3) years, reflecting midcareer professionals, and a male-to-female ratio of 2.8:1 (n=33, 73% men and n=12, 27% women). Geographically, the sample spanned 19 countries, predominantly in Africa (n=26, 58% participants), followed by Asia (n=14, 31% participants), Latin America (n=3, 7% participants), and the Middle East and North Africa region (n=2, 4% participants). Among female participants, 50% (6/12) were from Africa, 33.3% (4/12) from Asia, 8.3% (1/12) from Latin America, and 8.3% (1/12) from the Middle East and North Africa region. The distribution of participants by gender and geographic region did not differ significantly (Fisher exact test P=.77), indicating that female participants were not disproportionately concentrated in any single region despite their smaller overall number. Participants had a mean professional experience of 7.3 (SD 5.9) years, and represented diverse roles across academia, policy, investment, and practice, including directors, principal investigators, professors, consultants, and World Health Organization officers. Their professional involvement encompassed a wide array of digital health and One Health initiatives in the Global South, such as AI-driven disease prediction, electronic health records, mobile health (mHealth) applications, epidemic surveillance, and AMR monitoring in wastewater. Notably, many participants contributed to public health training, policy development, and the integration of emerging technologies to address syndemics at the human-animal-environment interface.

Digital Health in the Global South

LSA applied to responses regarding the successful implementation of digital health in the Global South revealed 10 interpretable topics. Topic 1, explaining 46.75% of the variance, reflected structural and systemic priorities—particularly integration, infrastructure, and strategic alignment. Topic 2 (7.57%) captured operational themes around patient management and service delivery, while Topic 3 (4.77%) emphasized stakeholder engagement and community outreach. Topic 4 (4.21%) focused on governance and regulation, complemented by Topic 5 (4.04%) which stressed regulatory clarity and local adaptation. Topic 6 (3.57%) addressed digital literacy and technological access, and Topic 7 (2.99%) emphasized co-design and participatory processes. Topics 8-10, collectively explaining another 6.91%, covered scalability and sustainability (2.63%), capacity building (2.21%), and context-aware implementation frameworks (2.07%), respectively. Adopting a gender lens, common themes included strategic alignment (male Topic 7: 2.68%; female Topic 6: 1.56%), capacity building (male Topic 5: 3.90%; female Topic 3: 9.22%), and patient-centered approaches (male Topic 2: 8.53%; female Topic 9: 0.83%), though to a different degree. Both groups also addressed community and infrastructure support, with the theme being more prominent in women—Topic 1 (52.95%) versus Topic 8 (2.35%). Male responses uniquely stressed technical and structural aspects, such as the importance of system-wide deployment and digital solution rollouts (Topic 1: 50.18%), project-level security and awareness (Topic 6: 3.37%), and digital health literacy (Topic 4: 4.53%), as well as affordability and device regulation (Topic 3: 5.10%), financial sustainability (Topic 10: 2.11%), and framework scaling (Topic 9: 2.22%). Female respondents, in contrast, highlighted intersectoral dimensions, with attention to collaboration among stakeholders and financial dependencies (Topic 4: 3.74%) and multiactor coordination across health ecosystems (Topic 5: 2.00%). Their perspectives further reflected a focus on governance and policy environments (Topic 2: 16.45%), societal trust and acceptability (Topic 7: 1.17%), and end user needs and relational dynamics (Topic 8: 1.03%).

Regarding opportunities in digital health, Topic 1, accounting for 57.86% of variance, indicated widespread optimism about using digital solutions to enhance health care access, especially via remote care. Topic 2 (7.25%) addressed sustainability and institutional development, and Topic 3 (6.07%) highlighted chronic disease management. Topic 4 (4.04%) explored data-driven, patient-centered innovation, while Topic 5 (3.46%) pointed to implementation gaps at the community level. Subsequent topics included interoperability and workforce readiness (Topic 6: 3.01%), mobile support and training (Topic 7: 1.96%), public infrastructure (Topic 8: 1.80%), epidemic preparedness (Topic 9: 1.54%), and scalable digital records (Topic 10: 1.32%), respectively. In the comparison of male and female responses, both groups emphasized digital health care and remote improvement (male Topic 1: 58.94%; female Topic 1: 66.52%) and sustainability and innovation (male Topic 2:

XSL•FO
RenderX

8.20%; female Topic 2: 9.41%), indicating a shared focus on improving access and strengthening systems. Similarly, training gaps and field limitations were addressed in both analyses (male Topic 7: 2.14%; female Topic 7: 2.11%), as was the management of chronic or communicable diseases, though framed differently: men highlighted mHealth-enabled responses (Topic 3: 7.28%; Topic 5: 3.86%), while women emphasized coordination and surveillance (Topic 3: 5.73%), which was a more diffuse and less prominent theme in men (Topic 8: 1.64%; Topic 10: 1.24%). Differences emerged also in the framing of structural versus relational concerns. Male participants prioritized regional equity and geopolitical targeting (Topics 4 and 9: 4.34% and 1.40%), within a more operational and technology-driven framing of future digital health priorities (Topic 6: 3.34%). Conversely, female respondents focused on patient-centered innovation (Topic 4: 4.47%), interoperability and mHealth gaps (Topic 6: 3.30%; Topic 8: 1.94%; Topic 10: 0.62%), and real-time communication and community linkage (Topic 9: 1.28%), reflecting a more integrative, policy- and user-oriented approach to digital health (Topic 5: 3.69%).

Concerning challenges, Topic 1, accounting for 59.15% of the variance, indicated widespread concern with infrastructural limitations and technological fragmentation. Topic 2 (5.59%) focused on deficits in internet and electricity access, while Topic 3 (3.93%) dealt with implementation bottlenecks in local health systems. Topic 4 (3.71%) addressed access challenges in remote regions, and Topic 5 (3.09%) pointed to fragile community networks and institutional misalignment. Topics 6 and 7 (2.56% and 2.36%, respectively) highlighted digital literacy gaps and rural governance trust issues. Topic 8 (2.18%) was centered on financial constraints, while Topic 9 (1.87%) and Topic 10 (1.63%) underscored politicization and workforce/data infrastructure shortages, respectively. The comparison between male and female LSA topic profiles revealed that both groups identified technological infrastructure and access as the primary concern (male Topic 1: 63.61%; female Topic 1: 45.14%), emphasizing the foundational role of connectivity, electricity, and digital systems. However, male respondents focused more on infrastructural (Topic 2: 5.45%; Topic 8: 1.97%; Topic 10: 1.25%) and professional network gaps (Topic 4: 3.97%; Topic 5: 3.26%), and workforce shortages (Topic 7: 2.36%). In contrast, female respondents emphasized sociopolitical and cultural challenges, including language, policy, and cost barriers (Topic 2: 17.43%; Topic 6: 5.00%), patient access (Topic 3: 8.75%), trust in digital systems (Topic 9: 1.19%), and data protection (Topic 10: 0.89%). Both groups acknowledged rural challenges, education gaps, and politico-economic issues, though women assigned higher variability to these themes (Topic 7: 4.18% vs Topic 7: 2.36%; Topic 5: 6.14% vs Topic 6: 2.80%; Topic 8: 3.02% vs Topic 9: 1.76%; Topic 4: 7.62% vs Topic 3: 4.36%).

In the articulation of future-oriented scenarios, integrated digital care and health technologies (Topic 1: 50.58% explained variance) highlighted a strong consensus on foundational infrastructure and system-wide digital innovation. This was followed by mobile innovation and app-based access (Topic 2: 11.00%) and AI for underserved regions and clinical support (Topic 3: 7.03%), reflecting emphasis on scalable, intelligent solutions for equity and efficiency. Additional themes included electronic health records, mobile health (mHealth), and Internet of Things convergence (Topic 4: 5.26%) and big data potential (Topic 5: 3.43%), both pointing to an evolving interest in interconnected health data ecosystems. Lower-variance topics such as AI-powered decision support (Topic 6: 2.40%), scalable toolkits (Topic 7: 2.13%), and digital medicine (Topic 8: 1.33%) contributed further nuance, underscoring the importance of interoperability, clinical augmentation, and monitoring. For both male and female participants, the dominant theme centered on general digital infrastructure, though with slightly different emphases. Men prioritized general digital health infrastructure (Topic 1: 48.88% variance), emphasizing data and technological systems (Topic 5: 2.99%), while women focused on remote digital health management and records (Topic 1: 72.13%; Topic 7: 1.38%), reflecting a patient-facing orientation grounded in access and care delivery. Male respondents highlighted mobile AI and telemedicine innovation (Topic 2: 13.72%), chronic care and surveillance (Topic 3: 8.98%; Topic 4: 5.10%; Topic 9: 1.38%), and smartphone-based scalable solutions (Topic 6: 2.65%)—a pattern that favors modular, tech-forward tools. In contrast, women showed greater thematic fragmentation across app-based empowerment (Topic 2: 9.17%), AI/chatbot integration with mHealth and Internet of Things (Topic 3: 4.98%; Topic 4: 3.51%; Topic 6: 1.85%), and smartphone planning tools (Topic 8: 0.75%). Notably, chatbot technologies appeared in both profiles (male Topic 7: 2.13%, female Topic 4: 3.51%), although embedded in different semantic contexts—clinical utility for men, and innovation for women. Men also highlighted big data and response systems (Topic 8: 1.56%), while women emphasized real-time data collection (Topic 5: 2.11%).

Further details are reported in Table 1 and Tables S1-S12 in Multimedia Appendix 2.

XSL•FO
RenderX

Table 1. Latent semantic analysis–based topic modeling across 4 dimensions of digital health discourse in the Global South: implementation, opportunities, challenges, and foresight. For each of the extracted topics, the percentage of explained variance and corresponding thematic focus are reported.

Topic	Implementation		Opportunities		Challenges		Foresight	
	Explained variance (%)	Thematic focus	Explained variance (%)	Thematic focus	Explained variance (%)	Thematic focus	Explained variance (%)	Thematic focus
1	46.75	Integration, infrastructure, strategic alignment	57.86	Health care access and remote care	59.15	Infrastructural limitations, tech fragmentation	50.58	Integrated digital care and health technologies
2	7.57	Patient management and service delivery	7.25	Sustainability, institutional development	5.59	Internet and electricity access deficits	11.00	Mobile innovation and app-based access
3	4.77	Stakeholder engagement, community outreach	6.07	Chronic disease management	3.93	Local health system bottlenecks	7.03	AI[a] for underserved regions and clinical support
4	4.21	Governance and regulation	4.04	Data-driven, patient-centered innovation	3.71	Remote region access challenges	5.26	EHR[b], mHealth, and IoT[c] convergence
5	4.04	Regulatory clarity and local adaptation	3.46	Implementation gaps at community level	3.09	Community network fragility, institutional misalignment	3.43	Big data potential
6	3.57	Digital literacy and tech access	3.01	Interoperability, workforce readiness	2.56	Digital literacy gaps	2.40	AI-powered clinical decision support
7	2.99	Co-design and participatory processes	1.96	Mobile support and training	2.36	Rural governance trust issues	2.13	Toolkits and scalable tech solutions
8	2.63	Scalability and sustainability	1.80	Public infrastructure	2.18	Financial constraints	1.33	Digital medicine and outcome tracking
9	2.21	Capacity building	1.54	Epidemic preparedness	1.87	Politicization	—[d]	—
10	2.07	Context-aware frameworks	1.32	Scalable digital records	1.63	Workforce/data infrastructure shortages	—	—

[a]AI: artificial intelligence.

[b]EHR: electronic health records.

[c]IoT: Internet of Things.

[d]Not available.

One Health in the Global South

LSA of responses on One Health opportunities in the Global South revealed a dominant discourse anchored in integrated human, animal, and environmental health systems (Topic 1: 57.20% of the variance). Respondents identified strong opportunities in ecological integration, predictive modeling (Topic 2: 7.27%), digital innovation and AMR surveillance (Topic 3: 6.23%), and AI-driven tools and awareness strategies (Topic 4: 4.79%). Additional themes included prevention and leadership (Topic 5: 3.55%), holistic approaches (Topic 6: 3.10%), community inclusion (Topic 7: 2.29%), vaccine distribution (Topic 8: 1.82%), digital records (Topic 9: 1.60%), and expert capacity building (Topic 10: 1.46%). Female and male respondents shared a strong common emphasis on the ecological foundations of One Health. For both, Topic 1 was dominant—accounting for 73.20% of variance in women and 57.45% in men—and focused on the interconnectedness of human, animal, and environmental health. Beyond this, the thematic priorities diverged significantly. Among women, the discourse was highly concentrated, with secondary themes oriented around strategic implementation and population-level equity. Topic 2 (5.42%) addressed AI and demographic growth, while Topic 3 (4.96%) focused on improving health responses and prevention. Topic 4 (3.93%) indicated attention to sectoral organization. Additional themes included community care and lived experience (Topic 5: 3.73%), treatment applications and clinical uptake (Topic 6: 2.97%), and smart working environments (Topic 7: 2.37%). Less prominent but still present were references to awareness building (Topic 8: 1.23%), health services (Topic 9: 1.14%), and remote training and key local solutions (Topic 10: 0.62%). In contrast, male respondents articulated a more diffuse and operationally driven perspective. Topic 2 (8.27%) emphasized predictive analytics and surveillance technologies, and Topic 3 (7.15%) addressed AMR through a sectoral lens. Topic 4 (5.55%) centered on AI and

XSL•FO
RenderX

complex policy coordination. While women focused on communities, men placed greater weight on biodiversity leadership and structured programs (Topic 5: 3.90%), vaccination and disease control strategies (Topic 6: 3.07%), and infectious disease enablement and access (Topic 7: 2.23%). Topic 8 (1.86%) highlighted digital recordkeeping and remote access infrastructure, while Topic 9 (1.60%) captured surveillance, expert knowledge, and vector-borne risks. Topic 10 (1.15%) pointed to health system resources, support, and scale-up potential.

Concerning challenges, Topic 1, accounting for 61.23% of the variance, emphasized environmental degradation, limited resources, and systemic health-environment linkages. Topic 2 (7.43%) highlighted data gaps and weaknesses in surveillance systems, particularly in the context of environmental change. Topic 3 (4.93%) focused on the lack of interdisciplinary coordination and field-level impact. Topic 4 (4.19%) addressed insufficient policy capacity, research infrastructure, and cultural readiness. Topic 5 (3.57%) addressed population-level issues, while Topic 6 (2.76%) pointed to rural marginalization and priority misalignment. Topic 7 (1.97%) underscored digital skill gaps and limited connectivity, and Topic 8 (1.65%) emphasized resource shortages and structural barriers. Topic 9 (1.35%) and Topic 10 (1.29%) captured concerns on meat safety and technological devices. Both female and male respondents strongly converged on structural and environmental limitations as the foremost barrier to implementing One Health in the Global South. In both groups, Topic 1 dominated the discourse—accounting for 61.94% of variance in women and 65.81% in men—and emphasized terms such as health, environment, resource, challenge, and implementation, revealing a shared perception of systemic constraints and infrastructural insufficiency. Beyond this core alignment, the 2 groups diverged notably in the framing and distribution of secondary themes. Among female respondents, Topic 2 (14.68%) focused on population practices, understanding, and adoption, suggesting an emphasis on behavioral, educational, and participatory barriers. Topic 3 (8.84%) addressed funding shortfalls, data gaps, and operational limitations, while Topic 4 (5.32%) pointed to sectoral and institutional weaknesses. Topic 5 (2.36%) dealt with device- and research-based deficiencies. Less prominent themes included poverty (Topic 6: 1.92%), cross-sector coordination (Topic 7: 1.57%), economic emergence and shocks (Topic 8: 1.00%), technological and cultural constraints (Topic 9: 1.00%), and multilevel governance issues (Topic 10: 0.91%). These findings reflect a relatively distributed but coherent concern with systemic functionality, knowledge equity, and integrated governance. In contrast, male respondents articulated a more operationally dispersed set of challenges. Topic 2 (8.80%) dealt with surveillance gaps and environmental degradation, while Topic 3 (5.28%) emphasized capacity building and culturally embedded policy. Topic 4 (3.76%) addressed priority setting and geographic inequities, and Topic 5 (2.98%) brought attention to rural underdevelopment, digital skills, and infrastructure deficits. Topic 6 (2.21%) highlighted low connectivity and staff shortages, and Topic 7 (1.83%) captured community-level implementation difficulties, particularly in relation to meat safety and device availability, while Topic 8 (1.28%) focused on resource and capacity deficits. Topic 9 (1.13%) highlighted disciplinary fragmentation in science and field application, and Topic 10 (1.01%) included conceptual concerns.

Concerning future-related opportunities, Topic 1, explaining 58.78% of the variance, emphasized integrated health approaches linking disease prevention, animal health, and environmental improvement. Topic 2 (5.20%) reflected ambitions for enhanced health care responsiveness through real-time monitoring and information systems. Topic 3 (4.86%) addressed food systems, AMR, and environmental-professional literacy, while Topic 4 (3.80%) pointed to systems development and the implementation of expert-driven frameworks. Additional topics captured emerging or cross-cutting themes. Topic 5 (3.40%) focused on conservation, biodiversity, and domain-specific innovation, and Topic 6 (3.10%) raised concerns about infrastructure, digital records, and electronic health integration. Topic 7 (2.53%) focused on forward-oriented transformative potential of One Health in the Global South. Topic 8 (2.18%) underscored access to care and population equity, while Topic 9 (2.11%) centered on infectious disease awareness and public engagement. Finally, Topic 10 (1.91%) reflected projections around early medical interventions and future modeling scenarios. Male and female respondents both emphasized the ecological nexus of human, animal, and environmental health as a core future opportunity for One Health advancement (male Topic 1: 60.29%; female Topic 1: 68.63%), highlighting a shared foundational understanding across genders. However, differences emerged in the framing and elaboration of these priorities. Women emphasized responsive systems and coordination tools (Topic 2: 20.13%) and showed a stronger orientation toward cultural integration and community empowerment, as seen in topics such as indigenous engagement and awareness promotion (Topic 3: 2.92%), technological advancement and collaboration (Topic 4: 2.77%), and specific innovations like diagnostics (Topic 5: 1.44%), outreach (Topic 6: 0.95%), and data availability (Topic 7: 0.48%). They also included a diffuse but conceptually rich theme on future risks, AMR, and professional preparedness (Topic 8: 0.00%). Male respondents, while similarly prioritizing the health-environment interface (Topic 1: 60.29%), articulated a more technical and operational approach in the remaining topics. These included strengthening sustainable health care systems and food security (Topic 2: 6.03%), implementing expert-led systems (Topic 3: 4.95%), resolving infrastructural and data management issues (Topic 4: 3.93%), and developing digital platforms (Topic 5: 3.33%). Other themes involved community-based research and responsiveness (Topic 6: 3.01%), epidemic readiness and access (Topic 7: 2.69%), early interventions and infectious disease modeling (Topic 8: 2.49%), mobile diagnostics and rapid response (Topic 9: 1.92%), and population-level coordination (Topic 10: 1.61%).

Finally, regarding future-related challenges, Topic 1, explaining 59.65% of the variance, reflected deep-seated anxieties around health system fragility, environmental degradation, resource scarcity, and the growing burden of disease. Topic 2 (7.81%) pointed to digital infrastructure gaps and persistent barriers to health management and equitable access. Topic 3 (5.47%) emphasized food systems, veterinary integration, and the need

XSL•FO
RenderX

for cohesive frameworks to support cross-sectoral promotion of One Health. Topic 4 (4.57%) addressed insufficiencies in outbreak detection and relational coordination, while Topic 5 (2.87%) focused on the lack of financial, regulatory, and institutional support for coordinated initiatives. Topic 6 (2.46%) highlighted time-sensitive skill shortages, data deficits, and the limited capacity at the local level. Topic 7 (2.08%) identified environmental opportunities yet to be enabled, suggesting underutilized potential in ecosystem-based approaches. Topic 8 (1.85%) captured cultural and economic inertia that may hinder timely and inclusive responses, and Topic 9 (1.62%) reflected concerns about weak collective action and unrealized potential across populations. Topic 10 (1.41%) pointed to limitations in emergency readiness, agency coordination, and resource mobilization. Both male and female respondents identified systemic and structural limitations as the most significant future challenge for implementing One Health in the Global South. Topic 1 accounted for 78.50% of variance in female responses and 61.65% in men. In both cases, the theme encompassed resource scarcity, environmental degradation, infrastructural weaknesses, and disease burden, indicating a clear consensus on foundational barriers. The female framing included community-based challenges, while the male version emphasized intersectoral and operational breakdowns. Female Topic 2 (8.30%) focused on management, regulatory complexity, cultural barriers, funding deficiencies, and political inertia. This aligns with male Topic 4 (3.42%), which also emphasized coordination, governance, and regulatory support, though framed more technocratically. The female version was more critical and rooted in institutional culture. Female Topic 3 (6.46%) addressed emerging threats, climate-driven disease risk, and resource vulnerabilities, which parallels male Topic 3 (5.77%) on food systems, integration, and outbreak preparedness. Both saw environmental unpredictability and multisectoral fragility as critical future stressors. Female Topic 4 (2.18%) on economic pressure and population-level stress resonated with male Topic 7 (2.13%) and Topic 10 (1.16%), which covered poverty, inequity, local capacity gaps, and leadership voids. The male response provided more layers, incorporating network fragility, population-scale planning, and decentralized leadership, while the female view presented a compact but conceptually broad expression of economic vulnerability. Female Topic 5 (1.39%) emphasized data collection and information deficits, which echoes male Topic 5 (2.80%) and Topic 6 (2.50%) on skills, local-level data systems, and epidemiological responsiveness. The male focus offered more operational detail, while the female version pointed to a diagnostic gap. Female Topic 6 (0.57%) is abstract and thinly developed, but potentially refers to actual lived experience or practical realities—a dimension that, while unelaborated, might align with male Topics 8 (2.01%) and 9 (1.26%) concerning time constraints, remoteness, and pragmatic insufficiencies. Finally, although female Topic 7 (0.00% variance) did not contribute to model variance, it contained rich content including disciplinary fragmentation, outbreak management, and regional opinion dynamics. This latent theme maps well to male Topics 2 (9.17%) and 8 (2.01%), which dealt with Global South–specific digital development, coordination barriers, and interdisciplinary silos.

Further details are reported in Table 2 and Tables S13-S23 in Multimedia Appendix 2.

XSL•FO
RenderX

Table 2. Latent semantic analysis–derived topic modeling of stakeholder discourse on One Health in the Global South. Each topic's explained variance and thematic focus are presented for 4 analytical lenses: opportunities, challenges, future opportunities, and anticipated challenges.

Topic	Opportunities		Challenges		Foresight opportunities		Foresight challenges	
	Explained variance (%)	Thematic focus	Explained variance (%)	Thematic focus	Explained variance (%)	Thematic focus	Explained variance (%)	Thematic focus
1	57.20	Integrated human-animal-environment systems	61.23	Environmental degradation, limited resources, systemic health-environment links	58.78	Integrated disease prevention and environmental improvement	59.65	Health-system fragility, environmental degradation, resource scarcity
2	7.27	Ecological integration, predictive modeling	7.43	Data gaps, surveillance weaknesses	5.2	Real-time monitoring and responsive systems	7.81	Digital infrastructure and access barriers
3	6.23	Digital innovation, AMR[a] surveillance	4.93	Lack of interdisciplinary coordination	4.86	Food systems, AMR, environmental-professional literacy	5.47	Food systems, veterinary integration, cross-sector support
4	4.79	AI[b] tools, awareness strategies	4.19	Insufficient policy, research, and cultural capacity	3.8	Systems development, expert frameworks	4.57	Outbreak detection, coordination deficits
5	3.55	Prevention and leadership	3.57	Population-level concerns	3.4	Conservation, biodiversity, innovation	2.87	Financial, regulatory, and institutional voids
6	3.10	Holistic approaches	2.76	Rural marginalization, priority misalignment	3.1	Digital infrastructure, e-records integration	2.46	Skill gaps, data deficits, local capacity
7	2.29	Community inclusion	1.97	Digital skill gaps, limited connectivity	2.53	Transformative One Health potential	2.08	Underutilized ecosystem potential
8	1.82	Vaccine distribution	1.65	Resource shortages, structural barriers	2.18	Access to care, equity	1.85	Cultural/economic inertia
9	1.60	Digital records	1.35	Meat safety concerns	2.11	Infectious disease awareness, engagement	1.62	Weak collective action
10	1.46	Expert capacity building	1.29	Technological device limitations	1.91	Early intervention and predictive modeling	1.41	Emergency readiness, coordination, resource mobilization

[a]AMR: antimicrobial resistance.

[b]AI: artificial intelligence.

Digital One Health in the Global South

In the LSA of responses to the question on the convergence of digital health and One Health approaches in the Global South, Topic 1, accounting for 75.2% of the total variance, emphasized digitalization, health data infrastructure, and disease surveillance systems, as underpinning the operationalization of the integration of digital health and One Health. Topic 2, explaining 5.4%, captured the role played by global convergence agendas, sectoral alignment, and climate recognition in shaping integrative frameworks across human, animal, and environmental health. Topic 3 (3.9%) pointed to the catalytic effect of the COVID-19 pandemic, which had prompted holistic thinking and increased organizational and financial support for integrated approaches. Topic 4 (3.0%) highlighted how AI and risk management strategies had been mobilized to address systemic constraints and enhance preparedness. Topic 5 (1.8%) reflected the importance of national delivery mechanisms and context-specific implementation, while Topic 6 (1.5%) illustrated the multisectoral impacts that had been particularly salient in the Global South. Topic 7 (1.4%) underscored governance challenges, including institutional fragmentation and reliability gaps, that had historically limited coordinated digital health and One Health responses. Finally, Topics 8 (0.87%), 9 (0.69%), and 10 (0.59%) were centered on policy awareness, stakeholder engagement, and scaling strategies. Topic 1 accounted for the majority of explained variance in both genders: 75.05% in women and 77.31% in men. Both groups emphasized the integration of digital health technologies into disease surveillance within a One Health framework. However, while female responses highlighted the importance of collaborative and community-oriented approaches, male

XSL•FO
RenderX

responses were more focused on technological infrastructure and human-disease interaction, illustrating a shared theme with participatory emphasis in women and infrastructural framing in men. Topic 2 was present in both genders (women: 4.77%, men: 5.62%) and concerned management and international drivers of convergence. Female responses were brief and managerial in focus, whereas male responses expanded on sectoral alignment, climate imperatives, and global recognition—pointing to a more externally driven and systemic perspective among male respondents. Topic 3 (women: 4.49%, men: 4.21%) demonstrated alignment around climate change and health system transformation. Women emphasized anticipatory and ecological aspects of convergence, while male responses framed the issue in light of pandemic-triggered reconfigurations and institutional support needs. Both acknowledged the environment-health nexus but with distinct emphases: prospective and global in women, reactive and policy-driven in men. Topic 4 (women: 3.52%, men: 3.24%) centered on institutional governance and risk. Female respondents focused on interministerial communication and vertical coordination, whereas men addressed operational aspects of risk management and the technical requirements of digital health implementation. The shared institutional theme revealed gendered nuances: administrative structure in women versus digital governance in men. Topic 5 (women: 3.26%, men: 1.63%) diverged in scope. Female responses addressed resource allocation and programmatic implementation challenges. In contrast, male discourse shifted toward geopolitical dynamics and multisector imbalances, reflecting broader inequalities in digital readiness across regions. Both pointed to structural enablers and barriers, but from local programmatic (female) and systemic geopolitical (male) vantage points. Topic 6 (women: 2.90%, men: 1.51%) related to delivery and performance. Women focused on access and service efficacy, while men underlined fragmentation and interministerial unreliability. Both acknowledged weaknesses in execution, with women emphasizing outcomes and men highlighting administrative dysfunction. Topic 7 was marginal among women (1.59%) and absent in the male profile. The theme suggested a latent concern with public engagement and legitimacy, though it remained underdeveloped in both groups. Topic 8 (women: 1.46%, men: 0.64%) touched on data access and informational equity. Both groups acknowledged the enabling role of digital health in democratizing health information, although the female response focused more explicitly on availability and transparency. Topic 9 (women: 1.45%, men: 0.59%) was conceptually minimal. While women pointed to the importance of key factors without elaboration, male responses offered no interpretable content. This suggests a shared but weakly articulated recognition of critical convergence enablers. Topic 10 was notably richer in the female discourse (0.78% vs 0.51%). Female responses addressed themes such as health inequity, underserved populations, antimicrobial stewardship, and system efficiency. Male responses, in contrast, referred to national context or jurisdiction.

Further details are reported in Table 3 and Tables S24-S26 in Multimedia Appendix 2.

Table 3. Latent semantic analysis of stakeholder responses on the convergence of digital health and One Health in the Global South.

Topic	Explained variance (%)	Thematic focus
1	75.2	Digitalization, health data infrastructure, disease surveillance
2	5.4	Global convergence agendas, sectoral alignment, climate recognition
3	3.9	Catalytic effect of COVID-19 on holistic and integrative thinking
4	3.0	AI[a] and risk management strategies for systemic preparedness
5	1.8	National delivery mechanisms and context-specific implementation
6	1.5	Multisectoral impacts in the Global South
7	1.4	Governance challenges and institutional fragmentation
8	0.87	Policy awareness and advocacy
9	0.69	Stakeholder engagement
10	0.59	Scaling strategies

[a]AI: artificial intelligence.

Discussion

Principal Findings

The foresight-oriented analysis of expert narratives revealed distinct thematic architectures underlying the discourses on digital health, One Health, and their convergence in the Global South. Importantly, these architectures should be interpreted not in isolation, but in relation to the existing international literature, which has been overwhelmingly shaped by Global North perspectives.

In the domain of digital health, structural drivers were consistently anchored in the infrastructural and systemic integration of digital technologies into health systems. These included the entrenchment of digital data ecosystems, the scaling of mHealth interventions, and the embedding of digital platforms into clinical workflows and public health delivery. Such drivers reflect a maturing landscape where digital technologies are no longer peripheral but fundamental to the functioning and reform of health governance in many Global South contexts. This emphasis partially mirrors findings from the Global North, where digital health maturity is typically framed around

XSL•FO
RenderX

interoperability, large-scale electronic health records, and AI-enabled clinical decision support embedded within stable institutional environments. However, unlike Global North settings—where digital infrastructure is often treated as a given—the Global South discourse consistently foregrounds infrastructural fragility, uneven access to electricity and connectivity, and dependence on external funding as persistent structural constraints.

Emerging issues in digital health centered on the transformation of care models through AI, the proliferation of mobile diagnostics, and evolving concerns around data sovereignty and ethical oversight. These topics reflect a transition in focus from basic digital inclusion toward strategic, anticipatory integration of advanced technologies. While similar themes appear in Global North studies—particularly around AI adoption and data governance—these findings suggest a qualitatively different framing in the Global South, where AI is often viewed as an aspirational or compensatory tool intended to offset workforce shortages and system limitations, rather than as an optimization layer built upon already robust health systems.

Particularly noteworthy was the shift from discussions of access and deployment to themes of agency, responsiveness, and the societal implications of algorithmic decision-making. Gendered nuances were evident, with female respondents frequently framing these issues through the lenses of equity, community participation, and trust, while male respondents emphasized operational scalability and innovation efficiency.

Within the One Health discourse, structural drivers coalesced around intersectoral coordination, disease surveillance at the human-animal-environment interface, and the growing imperative to institutionalize One Health frameworks. These reflect a stable recognition of ecological entanglement and the necessity of integrative governance structures to mitigate transboundary health threats. The discourse was marked by convergent priorities such as early warning systems, environmental health data infrastructures, and capacity-building in veterinary and environmental public health. Across genders, these priorities were consistently articulated, although women more frequently situated them within community resilience and social vulnerability contexts. This aligns with Global North One Health literature, which similarly prioritizes multisectoral surveillance, zoonotic preparedness, and early warning systems. However, in contrast to Global North contexts—where One Health is often operationalized through formalized interagency mechanisms—the Global South discourse reveals that coordination remains largely aspirational, constrained by fragmented mandates, limited regulatory authority, and underresourced environmental and veterinary sectors.

Emerging issues in One Health focused on anticipatory risk management, the operationalization of multisectoral collaborations, and the rising salience of food system sustainability and climate resilience. These signal a discursive broadening of One Health from zoonotic disease control toward more systemic, planetary health considerations. Notably, whereas Global North literature increasingly frames these challenges through the lens of technological integration and policy alignment, Global South experts more frequently emphasized structural underinvestment, competing development priorities, and the vulnerability of rural and marginalized populations to ecological shocks.

In several instances, female participants extended the discourse into areas of inclusive governance, highlighting the importance of participatory institutions and local knowledge systems in shaping resilient One Health infrastructures. These contributions suggest a shift toward democratized models of One Health governance that move beyond technical coordination toward transformative system change. Weak signals in One Health included fragmented reflections on indigenous knowledge integration, AMR governance at subnational levels, and structural underinvestment in ecological public health. Though sparse in expression, these signals may indicate latent areas of strategic neglect that could become critical under future scenarios marked by climate instability or political realignment. Such weak signals are largely absent or marginal in the Global North literature, where One Health discussions tend to privilege formal scientific expertise and national-level governance, thereby underscoring the added value of a foresight-oriented approach capable of surfacing peripheral yet potentially transformative concerns. The gendered content of these weak signals was particularly rich, with women surfacing relational, trust-based framings and men occasionally noting the absence of jurisdictional clarity and institutional mandates.

The discourse on the convergence of digital health and One Health was anchored by structural drivers that framed integration as both inevitable and strategic. Participants consistently viewed digital tools as essential enablers of One Health implementation, particularly in data sharing, interagency coordination, and real-time surveillance. This convergence was portrayed not as a mere overlay of technologies but as a paradigmatic shift toward more intelligent, interconnected, and adaptive health ecosystems. Emerging issues in the convergence discourse included the governance of cross-sectoral data flows, the development of digital infrastructures tailored to ecological surveillance, and the political economy of technology access in Global South contexts. These themes reflected a growing awareness of the nonneutrality of digital infrastructures and the need to align technological solutions with values of justice, inclusion, and environmental stewardship. Gendered patterns again surfaced, with female experts more likely to emphasize coproduction, relational accountability, and the risks of epistemic exclusion. This mirrors Global North narratives that conceptualize Digital One Health primarily as a technical convergence problem. However, the Global South discourse identified here reframes convergence as a deeply political and sociotechnical process, shaped by digital sovereignty, power asymmetries, and unequal access to technological resources.

Weak signals in the convergence space included underexplored themes such as feminist digital ethics, algorithmic bias in zoonotic prediction models, and the political implications of digital sovereignty in environmental health governance. Though peripheral in current discourse, these signals point to potentially disruptive tensions that could reshape future directions in Digital One Health if left unaddressed. Their emergence in this study highlights how foresight methodologies, particularly when combined with gender-sensitive analysis, can reveal dimensions

XSL•FO
RenderX

of convergence that are systematically overlooked in implementation-focused research.

Altogether, the analysis surfaced a dynamic and increasingly reflexive ecosystem of discourse, in which well-established priorities coexist with early indicators of transformative change. By explicitly contrasting Global South expert imaginaries with the dominant Global North literature, this study demonstrates that Digital One Health convergence is not a universally linear or technologically deterministic process. Rather, it is contextually negotiated, unevenly paced, and deeply shaped by structural inequities, governance capacities, and gendered epistemologies. By disaggregating structural drivers, emerging issues, and weak signals across gendered and thematic lines, the study contributes a nuanced, equity-informed foresight mapping of how health futures are being envisioned in the Global South.

Building upon the LSA-based foresight mapping, a comparative reflection with recent qualitative literature [14-17] on digital health and One Health in the Global South reveals substantial areas of alignment as well as conceptual expansion. While prior studies have independently captured infrastructural, systemic, and organizational dimensions, this LSA analysis contributes temporal depth and a gendered, anticipatory framing that enriches and challenges existing scholarship.

Strengths and Limitations

This study leverages a robust methodological approach by integrating LSA-based topic modeling with foresight-oriented frameworks to explore the convergence of digital health and One Health in the Global South. The analytical design, which mapped present and future-oriented discourses across 4 strategic dimensions—implementation, opportunities, challenges, and prospective trajectories—enabled a layered classification of structural drivers, emerging issues, and weak signals. The incorporation of a Delphi-informed purposive sampling strategy ensured the inclusion of a diverse array of expert voices across geographical, disciplinary, and institutional boundaries. This methodological rigor lends robustness and ecological validity to the findings, especially in capturing the pluralistic and multisectoral character of the digital health and One Health convergence discourse in resource-constrained settings.

Despite these strengths, the study is not without limitations. First, while purposive sampling captured disciplinary diversity, the geographic composition of the expert panel was heavily skewed toward Africa, with 58% (26/45) of participants based in African countries. In contrast, representation from Latin America and the Caribbean and the Middle East and North Africa was minimal, comprising only 7% (3/45) and 4% (2/45) of the sample, respectively. Due to this imbalance, it was not possible to perform a reliable LSA stratified by geographic provenance. As such, important regional nuances—especially from underrepresented areas—may not be adequately captured. Second, the reliance on English-language responses potentially excluded non-Anglophone voices and introduced language-related semantic constraints. Third, while LSA effectively reduces linguistic dimensionality and enables the identification of latent thematic structures, it may obscure context-specific narrative nuance and is sensitive to preprocessing decisions such as tokenization, term selection, and sparsity thresholds. As a result, the derived topics reflect dominant semantic patterns rather than the full richness or contextual specificity of individual expert narratives, and should be interpreted as heuristic representations rather than definitive constructs. Finally, although the study incorporated a gender lens, the male-to-female ratio was 2.8:1, limiting the interpretive depth of gender-divergent discourse, particularly for weak signals.

A further limitation relates to the reliance on open-ended survey questions. While this approach was intentionally adopted to support exploratory, foresight-oriented inquiry, it inevitably introduces interpretive variability, as respondents may construe and frame the same prompt in different ways. In this study, this variability was reflected in systematic differences between male and female respondents, not only in the thematic priorities they emphasized but also in the semantic framing and level of abstraction of their responses. Such gender-related divergences should not be interpreted as inconsistencies or measurement error, but rather as an inherent feature of qualitative, discursive data elicited through open-ended instruments. Nonetheless, this interpretive flexibility limits direct comparability across respondent groups and constrains inferential interpretations of gender differences. While the use of LSA helped to identify shared latent structures across heterogeneous narratives, it cannot fully eliminate variability arising from differential interpretation of survey prompts. Closely related to this issue, an additional limitation concerns the survey instrument itself. The questionnaire was developed ad hoc for this study and was not based on a previously validated instrument. This approach was necessary given the exploratory and foresight-oriented nature of the research, as well as the absence of established instruments designed to capture expert perspectives on the convergence of digital health and One Health, particularly in Global South contexts. However, the use of an ad hoc instrument limits direct comparability with other studies and may affect measurement consistency. The multiround Delphi design, purposive expert sampling, and the application of LSA partially mitigated this limitation by emphasizing convergence of meaning across narratives rather than item-level psychometric properties. In addition, the predominance of midcareer participants—while purposeful for capturing forward-looking and implementation-oriented perspectives—may have limited the extent to which long-term institutional memory and historical continuity were represented. Future research could address these limitations through more focused and structured study designs. Follow-up studies may combine open-ended foresight prompts with semistructured or closed-ended items anchored to predefined conceptual domains such as governance, infrastructure, ethics, and equity. The use of scenario-based questions, vignettes, or Likert-scaled assessments of clearly defined drivers and barriers could enhance cross-group comparability while preserving interpretive depth. Mixed methods approaches integrating qualitative elicitation with quantitative validation, alongside more balanced gender representation, would further strengthen the robustness, comparability, and interpretability of future analyses.

XSL•FO
RenderX

Conclusions

This study provides a foresight-informed, gender-sensitive mapping of the convergence between digital health and One Health discourses in the Global South. Structural drivers centered on the systemic integration of digital infrastructure, intersectoral governance, and ecological surveillance, while emerging issues pointed toward anticipatory innovation, AI deployment, and food system resilience. Weak signals—particularly those surfaced by female respondents—highlighted underexplored dimensions such as indigenous knowledge, feminist digital ethics, and community-based relational dynamics. By disaggregating expert discourse through LSA and embedding it within a futures thinking framework, the study surfaces both prevailing and nascent imaginaries shaping global health trajectories. These findings underscore the importance of inclusive, equity-oriented approaches in shaping Digital One Health agendas, and call for stronger attention to gendered epistemologies, participatory governance, and digital sovereignty in future research and implementation efforts.

Acknowledgments

We extend our deepest gratitude to everyone who contributed to the success of this project. First, we would like to thank the International Development Research Centre (IDRC) for commissioning and supporting this critical work under contract number PO5001096. Their commitment to advancing research in the Global South has been invaluable in driving forward this initiative. We are especially thankful to the IDRC management team (Dr Chaitali Sinha, Dr Fraser Reilly-King, Dr Samuel Oji Oti, Dr Sana Naffa, and Dr Engelbert Bain Luchuo) for their constant support and guidance throughout this project. Their understanding of the complex health landscape in the Global South has greatly contributed to the relevance and impact of this report. Our sincere appreciation goes to the 45 experts from Africa, Asia, Latin America, and the Middle East who participated in the foresight exercise and provided critical input. Their collective expertise, diverse perspectives, and commitment to improving global health have been central to identifying the opportunities and challenges facing Digital One Health in the Global South. Lastly, we would like to acknowledge the communities and local stakeholders across the Global South whose lived experiences, voices, and knowledge are at the heart of this research. Their resilience and determination continue to inspire our work toward creating more equitable and sustainable health systems. Thank you all for your support, collaboration, and belief in the power of innovation to transform global health.

Funding

This research is funded by Canada's International Development Research Centre (contract number PO5001096). JK acknowledges support from Canada's IDRC (grant 109981-001), IDRC and the Foreign, Commonwealth and Development Office (grant 110554-001), Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (grant RGPIN-2022-04559), NSERC Discovery Launch Supplement (grant DGECR-2022-00454), New Frontier in Research Fund-Exploratory (grant NFRFE-2021-00879), and Canada Research Chair in Community-Oriented Artificial Intelligence and Mathematical Modelling of Infectious Diseases (award ID CRC-2023-00234.).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey items.
[DOCX File , 16 KB - jmir_v28i1e78702_app1.docx]

Multimedia Appendix 2
Additional tables.
[DOCX File , 56 KB - jmir_v28i1e78702_app2.docx]

References

1. Barrett B, Charles JW, Temte JL. Climate change, human health, and epidemiological transition. Prev Med 2015;70:69-75 [FREE Full text] [doi: 10.1016/j.ypmed.2014.11.013] [Medline: 25434735]
2. Hua K, Pan Y, Fang J, Wu H, Hua Y. Integrating social, climate and environmental changes to confront accelerating global aging. BMC Public Health 2024;24(1):2838 [FREE Full text] [doi: 10.1186/s12889-024-20346-7] [Medline: 39407185]
3. Boutayeb A. The burden of communicable and non-communicable diseases in developing countries. In: Handbook of Disease Burdens and Quality of Life Measures. New York, NY: Springer; 2010:531-546.
4. Acolin J, Fishman P. Beyond the biomedical, towards the agentic: a paradigm shift for population health science. Soc Sci Med 2023;326:115950 [FREE Full text] [doi: 10.1016/j.socscimed.2023.115950] [Medline: 37148746]

XSL•FO
RenderX

5. Horefti E. The importance of the One Health concept in combating zoonoses. Pathogens 2023;12(8):977 [FREE Full text] [doi: 10.3390/pathogens12080977] [Medline: 37623937]
6. Danasekaran R. One Health: a holistic approach to tackling global health issues. Indian J Community Med 2024;49(2):260-263 [FREE Full text] [doi: 10.4103/ijcm.ijcm_521_23] [Medline: 38665439]
7. Abernethy A, Adams L, Barrett M, Bechtel C, Brennan P, Butte A, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022;2022 [FREE Full text] [doi: 10.31478/202206e] [Medline: 36177208]
8. Scott P, Adedeji T, Nakkas H, Andrikopoulou E. One Health in a digital world: technology, data, information and knowledge. Yearb Med Inform 2023;32(1):10-18 [FREE Full text] [doi: 10.1055/s-0043-1768718] [Medline: 37414034]
9. Benis A, Tamburis O, Chronaki C, Moen A. One Digital Health: a unified framework for future health ecosystems. J Med Internet Res 2021;23(2):e22189 [FREE Full text] [doi: 10.2196/22189] [Medline: 33492240]
10. Ragnedda M, Gladkova A. Understanding digital inequalities in the Gobal South. In: Ragnedda M, Gladkova A, editors. Digital Inequalities in the Global South. Global Transformations in Media and Communication Research - A Palgrave and IAMCR Series. Palgrave Macmillan, Cham: Springer; 2020:17-30.
11. Rezaeian M, Montazeri H, Loonen RCGM. Science foresight using life-cycle analysis, text mining and clustering: a case study on natural ventilation. Technol Forecast Soc Change 2017;118:270-280 [FREE Full text] [doi: 10.1016/j.techfore.2017.02.027] [Medline: 32287406]
12. Kayser V, Blind K. Extending the knowledge base of foresight: the contribution of text mining. Technol Forecast Soc Change 2017;116:208-215. [doi: 10.1016/j.techfore.2016.10.017]
13. OECD. Strategic Foresight Toolkit for Resilient Public Policy: A Comprehensive Foresight Methodology to Support Sustainable and Future-Ready Public Policy. Paris: OECD Publishing; 2025.
14. O'Brien N, Li E, Chaibva CN, Gomez Bravo R, Kovacevic L, Kwame Ayisi-Boateng N, et al. Strengths, weaknesses, opportunities, and threats analysis of the use of digital health technologies in primary health care in the sub-Saharan African region: qualitative study. J Med Internet Res 2023;25:e45224 [FREE Full text] [doi: 10.2196/45224] [Medline: 37676721]
15. Steenkamp I, Peltonen LM, Chipps J. Digital health readiness - insights from healthcare leaders in operational management: a cross-sectional survey. BMC Health Serv Res 2025;25(1):240 [FREE Full text] [doi: 10.1186/s12913-024-12129-y] [Medline: 39939988]
16. Zharima C, Griffiths F, Goudge J. Exploring the barriers and facilitators to implementing electronic health records in a middle-income country: a qualitative study from South Africa. Front Digit Health 2023;5:1207602 [FREE Full text] [doi: 10.3389/fdgth.2023.1207602] [Medline: 37600481]
17. Brown HL, Pursley IG, Horton DL, La Ragione RM. One Health: a structured review and commentary on trends and themes. One Health Outlook 2024;6(1):17. [doi: 10.1186/s42522-024-00111-x] [Medline: 39138554]

Abbreviations

AI: artificial intelligence
AMR: antimicrobial resistance
LSA: latent semantic analysis
mHealth: mobile health

Edited by T Kristóf, B Mesko; submitted 07.Jun.2025; peer-reviewed by Z Kozlakidis, D Verran; comments to author 22.Aug.2025; revised version received 21.Dec.2025; accepted 13.Jan.2026; published 18.Feb.2026.

Please cite as:
Kong J, Bragazzi NL
Framing the Convergence of One Health and Digital Health in the Global South With a Gender-Sensitive Foresight Perspective: Delphi Study Using Latent Semantic Analysis
J Med Internet Res 2026;28:e78702
URL: https://www.jmir.org/2026/1/e78702
doi:10.2196/78702
PMID:

©Jude Kong, Nicola Luigi Bragazzi. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 18.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Channel Allocation and Equity in Preventive Campaigns for Older Adults: Agent-Based Modeling Study

Jihye Lee[1,2], PhD; Juyoung Park[2]; Yuna Kim[2]; Duk-Jo Kong[2], Prof Dr

[1]Seoul National University, Seoul, Republic of Korea

[2]Gwangju Institute of Science and Technology, 123 Cheomdangwagi-ro, Gwangju, Republic of Korea

Corresponding Author:
Duk-Jo Kong, Prof Dr
Gwangju Institute of Science and Technology, 123 Cheomdangwagi-ro, Gwangju, Republic of Korea

Abstract

Background: Preventive campaigns for older adults must decide how to allocate limited resources across media channels. However, these channel allocation and budget decisions rarely use explicit criteria for distributional equity or structured strategic planning tools. Consequently, health systems may optimize average uptake while leaving large gaps across socioeconomic groups and media use profiles.

Objective: This study aimed to develop and apply a data-driven agent-based model as a strategic planning tool for preventive campaigns targeting older adults, comparing channel allocation, personalization, and loss framing options under explicit budget and equity guardrails.

Methods: We built an agent-based model calibrated to national survey data from South Korea on influenza vaccination and routine health screening among older adults (vaccination, N=2405; screening, N=2400). Fifteen prespecified campaign scenarios varied channel allocation across television, digital, and print media; budget intensity; 2 equity-focused personalization strategies; and graded loss framing. Primary outcomes were final adoption and time to adoption. Equity outcomes included the minimum class-level adoption and 90 - 10 gap across latent classes. Each scenario was simulated over 12 monthly steps with 100 Monte Carlo replications. We conducted sensitivity analyses varying link functions and key social reinforcement parameters.

Results: Personalization improved uptake and equity relative to the integrated baseline. In the vaccination model (N=2405), adoption increased from 91.2% (n=2193) to 93.3% (n=2244) and 94.6% (n=2275). Minimum class-level adoption increased from 86.8% to 90.3% and 90.9%. The 90 - 10 gap narrowed from 5.7 to 4.5 and 4.7 percentage points. In the screening model (N=2400), adoption increased from 83.8% (n=2011) to 88.2% (n=2117) and 89.5% (n=2148). Minimum class-level adoption increased from 77.6% to 83.2% and 85.3%. The 90 - 10 gap narrowed from 9.2 to 7.4 and 6.2 percentage points. Television-only strategies achieved high adoption but had less favorable equity profiles than personalization. High-budget strategies achieved high adoption but required higher total exposure. Stronger loss framing produced small, monotonic gains in adoption and shortened the time to adoption without worsening equity in the tested range. Scenario rankings were stable in sensitivity analyses.

Conclusions: This agent-based modeling study illustrates how ex ante planning can improve preventive campaign design by comparing channel allocation and personalization options under explicit equity and budget criteria. For campaigns targeting older adults, equity-focused reweighting and class-tailored television-digital portfolios improved or preserved mean adoption while strengthening distributional equity under fixed budgets. In contrast, undifferentiated channel diversification without personalization offered a less favorable efficiency-equity trade-off. These findings support integrating explicit equity guardrails into early-stage channel allocation and prioritizing targeted personalization over simple channel diversification. Future work should validate these patterns in other populations and health systems and link simulated diffusion trajectories with observed exposure and engagement in real-world campaigns. It should also extend guardrail-based planning tools to organizational settings and multiyear decision contexts.

(J Med Internet Res 2026;28:e88429) doi:10.2196/88429

KEYWORDS

preventive campaigns; channel allocation; older adults; agent-based modeling; distributional equity; resource allocation; budget constraints; influenza vaccination; loss framing; strategic planning

XSL•FO
RenderX

Introduction

Background

Preventive behaviors among older adults, such as seasonal influenza vaccination and routine health screening, remain uneven across socioeconomic strata despite long-standing equity goals [1-4]. Multiple analyses in Europe and the United Kingdom report persistent or widening gaps in influenza vaccination uptake among adults aged ≥60 years by deprivation and education [1,2]. These studies also describe comparable heterogeneity in cancer screening participation and outcomes across European Union member states [3,4]. These disparities persist even where national coverage has improved, suggesting that changes in supply or eligibility alone are insufficient to close equity gaps. Vaccination and screening are related but behaviorally distinct preventive targets in older adults. Annual influenza vaccination is often an episodic, clinician-recommended decision, whereas routine health screening requires multistep planning, navigation of organizational procedures, and higher levels of health and digital literacy.

Daily life and health care for older adults are increasingly mediated by digital channels such as online news, social media, messaging apps, and patient portals [5-7]. Systematic reviews show wide variation in digital literacy, skills, and confidence among older adults, with distinct connectivity levels across user segments and pronounced social gradients in eHealth literacy [5-7]. Emerging digital health research links this digital divide to inequalities in self-management, use of online health services, and broader health outcomes among older adults [8-10]. Older adults are therefore not a single audience but a set of heterogeneous segments occupying different media and support ecologies that can amplify or dampen campaign effects.

Recent equity-focused work also emphasizes behavioral and social drivers of vaccination and screening—beliefs, norms, practical barriers, and motivation—rather than structural access alone. The World Health Organization's behavioral and social drivers of vaccination framework and scoping reviews of influenza vaccination in older adults highlight that groups with fewer resources are often missing from the evidence base and that social context strongly shapes uptake [11,12]. For campaign planners, this implies that channel choice, sequencing, and audience segmentation should consider not only reach but also how media exposure interacts with social reinforcement and practical constraints in different subgroups. These challenges point to the need for explicit digital health strategic planning tools to guide channel allocation, personalization, and the use of equity guardrails in preventive campaigns for older adults.

Current Evidence

Planning and Equity Frameworks

Implementation science models such as the PRECEDE-PROCEED model and the Reach, Effectiveness, Adoption, Implementation and Maintenance (RE-AIM)/Practical Robust Implementation and Sustainability Model (PRISM) have strengthened the understanding of reach, adoption, and sustainability in public health programs [13,14]. These models are increasingly applied to plan and evaluate interventions with an explicit focus on equity [15,16]. In parallel, equity frameworks such as PROGRESS-Plus (place of residence; race, ethnicity, culture, and language; occupation; gender and sex; religion; education; socioeconomic status; and social capital—plus additional equity-relevant factors) and related disparity indices are widely used to structure and report the distributional impacts of public health interventions in systematic reviews and policy analyses [15].

However, in most applications, these tools are applied after interventions have been implemented. Most studies focus on describing who was reached and how effects varied [1,13,17]. They rarely generate prospective allocation rules across channels or segments under budget and equity constraints. Even less often are they operationalized as ex ante decision support tools for channel portfolios.

Digital and Social Media Campaigns

Systematic reviews of digital and social media campaigns show complex exposure pathways and small but generally positive average effects on health behaviors, with substantial heterogeneity across platforms, populations, and outcomes [18-20]. However, most evaluations rely on intermediate metrics—impressions, clicks, likes, or other low-level engagement—rather than behavior change [18-20]. This reliance on proxy outcomes complicates the assessment of the added value of specific channels beyond offline activities.

As a result, planners have limited operational guidance on how to allocate budgets across channels. Key decisions include whether to concentrate resources on a single high-leverage channel (eg, television), distribute spending across multiple platforms, or adjust the mix over time as part of strategic planning.

Personalization and Equity

Personalization and targeting are widely promoted in digital health to improve relevance and engagement, including for older adults [21,22]. Conceptual and empirical reviews describe approaches ranging from simple demographic tailoring to rule-based and data-driven personalization, but their equity implications are mixed and often underdocumented [21,23]. A recurring concern is that personalization may disproportionately benefit digitally skilled or socially advantaged groups, thereby widening disparities [23,24].

Although equity-oriented frameworks encourage inclusive targeting and prioritization of underserved groups [25], there is little mechanistic evidence comparing fundamentally different personalization philosophies. Such contrasts include cautious, vulnerability-focused adjustments versus segment-optimized channel mixes evaluated under common efficiency and equity metrics. These contrasts suggest that the same personalization or channel strategy could plausibly yield different efficiency-equity profiles for vaccination and screening. However, existing planning models rarely examine these targets side by side within a unified framework.

XSL•FO
RenderX

Equity Metrics as Planning Tools

Equity metrics and frameworks have also advanced. Entropy-based indices such as Theil-type measures summarize the distribution of health outcomes across social strata and capture both between- and within-group variation [26]. In parallel, PROGRESS-Plus has become a standard for conceptualizing equity dimensions in systematic reviews and equity-relevant evaluations [25].

In many applications, equity metrics are used post hoc to describe which groups benefited [25,26]. Researchers have proposed more prescriptive approaches that incorporate distributional objectives into economic evaluation, including distributional cost-effectiveness analysis and pragmatic aggregate variants [27]. However, these methods remain uncommon in routine resource allocation practice [27].

Network and Agent-Based Modeling

Network and simulation approaches, especially agent-based models (ABMs), represent individuals as interacting agents linked through online and offline networks [28,29]. These models can explicitly capture heterogeneous exposure, social influence, and timing. Reviews describe ABMs as "policy laboratories" that allow planners to test alternative portfolios and explore system behavior before real-world deployment [28,29]. Network experiments and diffusion studies further show that many protective behaviors spread as complex contagions that require repeated, multisource reinforcement [30].

Despite this potential, most ABM applications in public health focus on infectious disease transmission or generic opinion dynamics [28]. Only a small but growing set of studies uses ABMs to examine vaccination strategies or health communication interventions [31,32]. Even in this work, models rarely consider resource constraints, channel portfolios, and equity-focused objectives for older adult vaccination or screening within a single, reproducible framework [31,32]. These gaps motivate data-driven ABM approaches that can support strategic planning for older adult vaccination and screening campaigns under explicit equity guardrails.

Research Gaps and Study Objectives

Against this backdrop, several gaps remain at the intersection of digital health campaigns, personalization, and equity for older adults:

- Budget-constrained channel portfolios: Despite detailed descriptions of older adults' digital heterogeneity and the digital divide, we found no models that explicitly link these patterns to budget-constrained channel portfolios [5-10,21,22]. Such models would show how much to invest in television, digital, or print media over time for preventive campaigns in this population.
- Contrasting personalization philosophies: The common assumption that personalization improves efficiency but threatens equity has not been tested in mechanistic models. In particular, few models compare fundamentally different personalization strategies—for example, a cautious, outcome-targeted boost for the worst-off class versus a fully segment-optimized, media affinity–based allocation—under common outcome and equity metrics.
- Target-specific behavioral profiles: Existing planning models seldom account for the fact that vaccination and screening differ in baseline uptake, decision complexity, and system dependencies in older adults. Few studies evaluate whether a given channel or personalization strategy traces the same efficiency-equity trade-offs across these 2 targets.
- Equity metrics as guardrails: Existing equity metrics and frameworks are rarely embedded as binding constraints in campaign planning [11-13,15,24]. There is limited empirical work on how explicit guardrails (eg, minimum subgroup adoption, caps on adoption gaps) reshape the space of admissible channel and personalization strategies [11-13,15,24].
- Reproducible data-driven ABM for strategic planning: Most studies evaluate isolated interventions or channels. Few use reproducible, survey-calibrated ABMs that tie nationally representative data on media use, digital skills, social connectedness, and preventive behaviors to multichannel campaign scenarios [31,32]. There is particularly little work on "virtual laboratories" that vary channel mix, personalization, framing, and equity guardrails jointly under fixed budgets within a unified, budget-constrained planning framework [31,32].

This study addresses these gaps by developing a reproducible data-to-ABM pipeline for preventive campaigns for older adults. We focused on older adults in South Korea—a high-income, aging health system—and aim to derive strategic planning principles that may be adaptable to similar contexts. Using matched national survey data, we constructed a heterogeneous older adult population segmented by latent class, embedded agents in multilayer offline-online networks, and simulated 15 prespecified campaign scenarios under fixed budgets and explicit equity guardrails. Our goal was to provide a strategic planning environment that allows planners to compare channel mixes, personalization philosophies, and loss framing intensities under explicit equity guardrails and interpret the resulting portfolios as decision-support inputs for national or regional planning. Figure 1 summarizes the data-to-ABM pipeline from data sources to simulation outputs. It highlights data integration and profiling; core modeling modules; the simulation engine; and the calibration and sensitivity loop used to evaluate channel allocation, personalization, and loss framing under explicit equity guardrails.

We focused on 5 research questions (RQs):

RQ1: how do single-channel campaigns compare with a background-only baseline?

RQ2: do multichannel mixes produce synergy or dilution under fixed budgets?

RQ3: how do 2 contrasting personalization strategies—outcome-targeted (cautious equity-first) versus class-customized (segment-optimized)—affect mean outcomes?

RQ4: how do these personalization strategies reshape distributional equity across latent classes?

XSL•FO
RenderX

RQ5: does loss-framed messaging act as a secondary tuner of adoption, and do its effects differ between vaccination and screening?

Figure 1. Overview of the data-to-ABM pipeline for evaluating channel allocation and personalization strategies under budget and equity guardrails. ABM, agent-based model; CWTA, cumulative weighted threshold adoption; HBM, health belief model; KMP, Korea Media Panel Survey; KNHANES, Korea National Health and Nutrition Examination Survey; LCA, latent class analysis; TTA, time to adoption; W_{BG}, background exposure weight; WS, Watts-Strogatz small-world network.

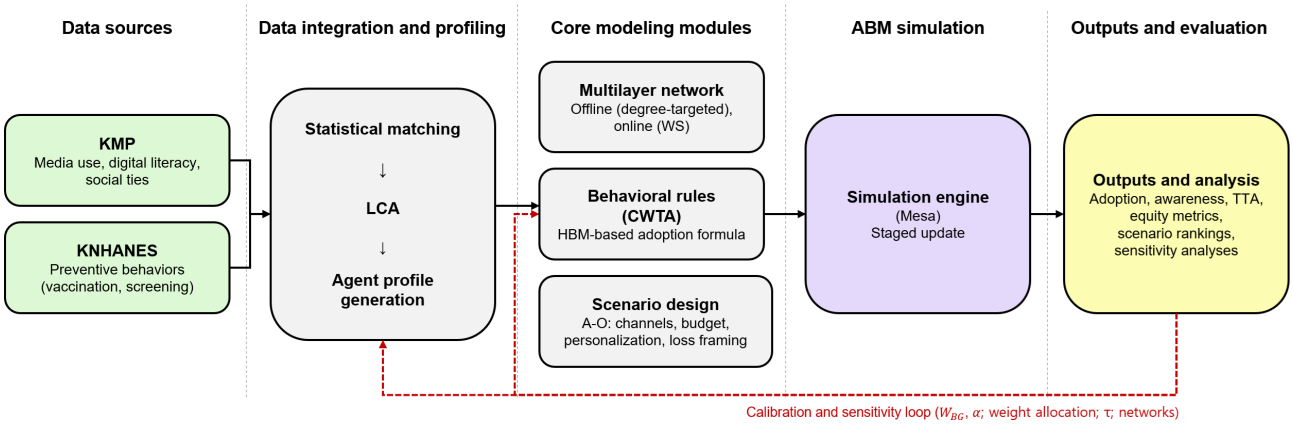

Methods

Study Design

Campaign planning for older adults must account for how exposure unfolds over time and spreads through offline and online social ties [18,33-35]. These dynamics are difficult to represent in static regression or compartmental models, which assume homogeneous exposure and ignore network reinforcement [36,37]. The challenge is amplified by having 2 preventive targets with distinct determinants: influenza vaccination and routine screening.

To address these limitations, we developed an ABM that represents each older adult as an agent embedded in multilayer networks [38]. The model captures heterogeneous exposure, repeated social reinforcement, and channel-specific media effects under fixed campaign budgets. It enables prospective comparison of prespecified channel mixes, personalization strategies, and equity guardrails as alternative campaign designs. Therefore, we used the model as a strategic planning tool for digital health campaigns targeting older adults.

Data Sources and Agent Construction

Data Sources and Statistical Matching

Two national surveys were combined: the 2022 Korea Media Panel Survey (KMP) [39] and the 2022 Korea National Health and Nutrition Examination Survey (KNHANES) [40]. KMP captured media use, digital literacy, and social relationship patterns, while KNHANES reported preventive behaviors including influenza vaccination and screening. Linking the 2 datasets enabled individual-level integration of media exposure, digital access, social connectedness, and preventive behaviors.

We restricted KNHANES recipients to adults aged ≥65 years and defined the KMP donor pool as adults aged ≥55 years to improve age comparability during matching. After harmonizing overlapping variables, we performed one-to-one nearest-neighbor matching with replacement using Gower distance on key sociodemographic covariates [41]. This approach assigned media, digital, and social attributes from KMP donors to matched KNHANES recipients while retaining KNHANES health outcomes. The matching diagnostics met our prespecified acceptance criterion; the full covariate list and diagnostics are reported in Multimedia Appendix 1.

Latent Class Analysis

Latent class analysis was applied to the matched dataset to identify classes of older adults with distinct media, digital, social, and preventive profiles [42]. The indicator domains included sociodemographics, digital literacy, media use, social activity, and vaccination and screening behaviors. Models with 3 - 10 classes were evaluated using the Akaike information criterion (AIC), the Bayesian information criterion (BIC), entropy, and the bootstrap likelihood ratio test (K vs K–1). A 6-class solution was selected based on the lowest BIC and high entropy (BIC=86,578.57; entropy=0.92). Solutions with additional classes yielded marginal AIC improvements. However, they increased complexity without BIC gains. Accordingly, the 6-class structure was selected as the basis for class-based personalization in the simulation. Variable definitions, discretization rules, and full model comparison results are provided in Table S1 in Multimedia Appendix 1, and detailed class profiles are reported in Table S3 in Multimedia Appendix 2. For descriptive summaries, individuals were assigned to their modal class. For simulation inputs, posterior class probabilities were retained as personalization weights. This approach preserves interpretability and propagates classification uncertainty into the inputs. We treated class membership as time invariant over the 12-month simulation horizon.

Agent Initialization and Network Priors

Each matched individual was instantiated as an agent with a channel affinity vector and an offline degree prior derived from survey responses. Channel affinities captured digital and traditional media use. Offline degree priors were based on self-reported social activity frequency and served as individual target degrees during offline network construction. Details of

XSL•FO
RenderX

affinity construction, degree-level mapping, and multilayer network generation are provided in Multimedia Appendix 1 and in the Network and State Dynamics section.

Outcomes and Equity Metrics

Primary outcomes were final adoption at month 12 (proportion of agents who adopted) and mean time to adoption among adopters. We used "adoption" as a generic term for vaccination uptake and screening participation. Awareness was treated as a supporting outcome.

We assessed equity across latent classes using 5 distributional metrics. A_{min} denoted the minimum class-level adoption. We defined the 90 - 10 gap as the difference between the 90th and 10th percentiles of class-level adoption and relative disparity index (RDI) as the ratio of maximum to minimum class-level adoption. We also calculated 2 entropy-based indices, Theil T and Atkinson $A_{0.5}$ (ε=0.5), as descriptive measures of inequality [43,44]. For interpretation, we used prespecified illustrative planning benchmarks ("equity guardrails"): $A_{min} \geq 0.60$, 90 - 10 gap$\leq$0.25, and RDI$\leq$1.5. These thresholds are tunable planning benchmarks for comparative portfolio assessment and not universal ethical or empirical standards. In practical terms, they set (1) a minimum-adoption floor for the lowest-adoption class and (2) limits on absolute and relative disparities. Specifically, we required ≤25 percentage points for the 90 - 10 gap and a ratio ≤1.5 for RDI. These thresholds use the same probability scale as the calibrated uptake in our setting. Theil T and Atkinson $A_{0.5}$ were reported descriptively and were not used for guardrail screening.

Scenario-level adoption and equity metrics for all campaign scenarios (A-O), for both vaccination and screening, are reported in Tables S1 and S2 in Multimedia Appendix 2. We ran 100 Monte Carlo replications per scenario. We report all outcomes as means with 95% percentile intervals (2.5th-97.5th percentile), stratified by task (vaccination vs screening).

Scenario Design

Channel Allocation and Budget (RQ1-RQ2)

We evaluated 15 prespecified scenarios (A-O) that varied in 4 design factors: channel allocation, budget intensity, class-level personalization, and loss framing. Exposure weights were specified as relative per-step exposure intensity parameters (rather than monetary spending). Scenarios A to H tested single channels or fixed-sum channel mixes at a common per-step media budget. Scenario A served as a control with no paid media exposure; awareness could arise from background exposure and social awareness. Scenarios B to D isolated individual channels (television [TV], digital, or print). Scenarios E to H implemented fixed-sum mixes combining 2 or 3 channels.

Scenarios I and J varied total budget intensity while keeping channel shares fixed. Scenario I scaled down exposure relative to the fixed-sum TV + digital mix, whereas scenario J scaled it up. Together, these scenarios represented low- and high-budget variants without changing the underlying channel mix.

Personalization: Efficiency and Equity (RQ3-RQ4)

Scenarios K and L introduced class-level personalization based on latent classes. Scenario K implemented equity-focused reweighting. For each task, the lowest-baseline uptake class received a small increase in television or digital exposure. All other classes had TV + digital exposure reduced by equal margins to maintain a constant class-average TV + digital budget. This design represents a conservative, budget-neutral policy that nudges media toward the lowest-baseline uptake class.

Scenario L assigned class-specific television and digital weights by latent class under the same class-level budget parity. Weights were tailored to each class's baseline media use profile (television-heavy, digital-heavy, or mixed), rather than applying a uniform mix. This strategy represents a more aggressive, class-tailored policy that explicitly aligns channel mix with each class's baseline media use profile.

Message Framing (RQ5)

Scenarios M to O applied loss framing multipliers (k_{neg}) to television and/or digital exposure. Meta-analytic evidence suggests that framing effects are generally modest and behavior dependent. Gain framing shows advantages for prevention behaviors, whereas effects are mixed or weak for detection behaviors [45,46]. Accordingly, we tested k_{neg} values near 1.0, that is, 0.95 to –1.10 for vaccination and 1.05 to 1.20 for screening [45,46]. Tables 1 and 2 report the exact channel weights and multiplier values for all scenarios.

XSL•FO
RenderX

Table . Channel allocation weights for campaign scenarios A to L.

Scenario	Name	Purpose	W_{TV}[c]	W_{DG}[c]	W_{Print}[c]
A[a]	Control	Background only	0.000	0.000	0.000
B	TV[b]-only	Isolate TV	0.100	0.000	0.000
C	Digital-only	Isolate digital	0.000	0.100	0.000
D	Print-only	Isolate print	0.000	0.000	0.100
E	TV + digital	Fixed-sum mix	0.050	0.050	0.000
F	TV + print	Fixed-sum mix	0.050	0.000	0.050
G	Digital + print	Fixed-sum mix	0.000	0.050	0.050
H	All channels	Integrated mix	0.033	0.033	0.033
I	Low-budget	Scaled-down exposure	0.025	0.025	0.000
J[d]	High-budget	Scaled-up exposure	0.200	0.200	0.000
K[e]	Equity-focused reweighting	Reweight toward lowest-baseline class	Class dependent	Class dependent	0.000
L[f]	Class-tailored channel portfolios	Class-tailored channel mix	Class dependent	Class dependent	0.000

[a]Background-only awareness with per-step spillover (W_{BG}=0.12 for vaccination, 0.10 for screening). W_{BG} denotes background exposure weight.

[b]TV: television.

[c]W_{TV}, W_{DG}, and W_{Print} denote per-step channel exposure weights for television, digital, and print media, respectively. Weights sum to approximately 0.10 (minor differences reflect rounding).

[d]Scenario J increases the total exposure (budget intensity) while keeping the channel shares fixed.

[e]Scenario K varies the intensity of the equity-focused reweighting ($\delta \in \{0.02, 0.05\}$). For vaccination, the lowest-baseline uptake class was Class 6 (W_{TV}=0.00, $W_{DG} = 0.10 + \delta$); for screening, the lowest-baseline uptake class was Class 1 ($W_{TV} = 0.10 + \delta$, W_{DG}=0.00). In both cases, all other classes received symmetric low-intensity weights ($W_{TV} = 0.05 - \delta/10$; $W_{DG} = 0.05 - \delta/10$). Results for δ=0.02 are reported in the main text, and δ=0.05 is presented as a higher-intensity sensitivity variant in Table S2 in Multimedia Appendix 3.

[f]Weights vary by latent class: classes 1 and 5: W_{TV}=0.10; class 6: W_{DG}=0.10; classes 2 - 4: W_{TV}=0.05 and W_{DG}=0.05.

Table . Loss framing scenarios M to O with intensity multiplier ranges.[a]

Scenario	Name	Definition	W_{TV}[d]	W_{DG}[d]	W_{Print}[d]
M	TV[b] (loss framing)	B with $W_{TV} \times k_{neg}$[c]	$0.10 \times k_{neg}$	0.00	0.00
N	Digital (loss framing)	C with $W_{DG} \times k_{neg}$	0.00	$0.10 \times k_{neg}$	0.00
O	TV + digital (loss framing)	E with $W_{TV}/W_{DG} \times k_{neg}$	$0.05 \times k_{neg}$	$0.05 \times k_{neg}$	0.00

[a]Vaccination: {0.95, 1.00, 1.05, 1.10}; Screening: {1.05, 1.10, 1.15, 1.20}. All exposure probabilities were clipped to [0,1].

[b]TV: television.

[c]k_{neg} denotes the loss framing intensity multiplier applied to baseline channel weights.

[d]W_{TV}, W_{DG}, and W_{Print} denote per-step channel exposure weights for television, digital, and print media, respectively. The weights sum to approximately 0.10 (minor differences reflect rounding).

Network and State Dynamics

State Dynamics

The simulation ran over 12 monthly steps (t=1, …, 12). At each step, we iterated agents in random order and applied state conditional rules. At each time step, we first updated awareness and then updated adoption conditional on awareness. Once adopted, agents remained in the adopted state (state=2).

At each time step t, the awareness probability combined 3 components. The first component was background exposure. The second was campaign media exposure, determined by scenario-specific channel weights and agent-level channel use for television, digital, and print media. The third was the share of aware neighbors, weighted by ωa. In the control scenario A, all paid media terms were set to zero, and awareness arose from background exposure and social awareness (spillover).

Conditional on awareness, the adoption probability combined baseline propensity, media exposure, and social reinforcement. The reinforcement term activated only when at least 2 neighbors had adopted (threshold τ=2) and influenced adoption but not

XSL•FO
RenderX

awareness. In loss framing scenarios M to O, we multiplied the television and digital exposure terms by k_{neg}. All probabilities were clipped to the [0,1] interval. Textbox 1 provides a brief road map of the model logic (awareness, adoption, and reinforcement) and how to read the simulation outputs.

Textbox 1. How the model works and how to read the outputs.

1. What the model simulates: the model simulates how campaign exposure and social ties translate into adoption over 12 monthly steps.
2. Awareness (notice the message): an agent becomes aware via background exposure, paid media (scenario weights × the agent's channel use), and social awareness (neighbors who are aware).
3. Adoption (take action): only if aware, the agent may adopt based on baseline propensity plus media and social influence.
4. Reinforcement (social proof): social influence applies only after at least τ=2 neighbors have already adopted.
5. Once adopted, always adopted: adoption is an absorbing state through month 12.
6. Outputs (how to read results): scenarios will be compared on final adoption, time to adoption, and equity across classes (A_{min}, 90 - 10 gap, relative disparity index [RDI]); the results will be interpreted as relative trade-offs under the same constraints and not as point forecasts.

Contact Structure

We modeled contacts as a 2-layer network with offline and online ties. For social influence, each agent's neighborhood was defined as the union of its offline and online neighbors.

For baseline and channel-mix scenarios (A-H), we generated the analytic agents (vaccination, N=2405; screening, N=2400) and both network layers once and held them fixed across all Monte Carlo replications. For intervention and framing scenarios (I-O), we regenerated both layers at the start of each replication to reflect structural uncertainty while preserving a stable baseline design. This design choice was used to (1) reduce Monte Carlo variance when comparing fixed-budget channel-mix scenarios (A-H) and (2) incorporate structural uncertainty for intervention or framing variants (I-O).

Offline social layer (face-to-face): we mapped survey-based social activity frequency to an agent-specific degree target. We then built an undirected graph using a degree-targeted algorithm while avoiding self-loops and multi-edges [47,48].

Online social layer (network-based): we constructed the online layer as a Watts-Strogatz small-world network with a mean degree of 6 and rewiring probability of 0.10 [49].

Calibration and Internal Validation

We calibrated 2 parameters for each task to match KNHANES 2022 targets for older adults (vaccination uptake 0.848; screening participation 0.744) [40]. We varied the baseline propensity multiplier and the background exposure over prespecified ranges. We then selected the parameter pair that minimized the absolute deviation from these targets under the background-only control scenario (A). This procedure yielded the values $\alpha_{vaccine}$=1.15 with W_{BG}=0.12 and $\alpha_{screening}$=1.03 with W_{BG}=0.10.

We assessed calibration quality using target deviation summaries and decile reliability plots that compared survey targets with deciles of baseline propensity [50-52]. These diagnostics used individual-level outputs aggregated across Monte Carlo replications. Tables S1 and S2 in Multimedia Appendix 4 provide the best-fitting parameter sets and the per-class calibration summaries. In this study, "validation" refers to internal diagnostics (calibration-to-target and reliability checks) within a survey-calibrated framework and does not constitute external validation using real-world campaign data.

Sensitivity Analyses

We assessed robustness along 4 prespecified axes: link function, weight allocation, the social reinforcement threshold, and network uncertainty. For each axis, we reestimated scenario-level adoption vectors under the corresponding sensitivity configuration and compared them with the baseline ordering. We quantified rank stability using Spearman correlations and maximum rank shifts (Δrank) and treated $\rho \geq 0.90$ and $\Delta\text{rank} \leq 1$ as heuristic benchmarks for high stability. Tables S1 and S2 in Multimedia Appendix 3 provides the full protocols, stability statistics, and intensity variants for the personalization strategy [53-55].

Ethical Considerations

This study used deidentified, publicly available microdata (KMP 2022; KNHANES 2022) [39,40] with no direct participant contact. Consent to participate was not applicable. Under institutional policy for secondary analyses of deidentified public datasets, formal institutional review board approval was not required. Under article 2(1) of the Bioethics and Safety Act of Korea [56], a "human subjects research project" is defined as research involving direct interaction with individuals or the use of identifiable personal information. As this study used only deidentified secondary data and did not involve identifiable information, it does not fall under this definition. Furthermore, according to article 13(1) of the Enforcement Rule of the Bioethics and Safety Act of Korea [57], research using publicly available data or research that does not collect or record personally identifiable information may be exempt from institutional review board review.

Results

Sample and Class Overview

The matched analytic sample included 2405 adults aged ≥65 years. Latent class analysis identified 6 classes with distinct sociodemographic and media use profiles [42]. Class sizes ranged from 245 (10.2%) to 534 (22.2%).

Baseline uptake varied across classes; the class-specific n and N are reported in Table S3 in Multimedia Appendix 2. Class 6

XSL•FO
RenderX

had the lowest baseline vaccination uptake, and Class 1 had the lowest baseline screening participation.

Calibration and Internal Validation

We calibrated the baseline parameters to match the empirical adoption targets for older adults in KNHANES 2022 [40]. The targets were 2039 of 2405 (84.8%) for vaccination and 1786 of 2400 (74.4%) for screening. The selected parameter sets matched these targets closely (absolute deviation≤0.002). Calibration metrics supported good fit: Brier scores were 0.142 for vaccination and 0.187 for screening; calibration slopes were 1.15 and 1.08, respectively; and intercepts were −0.22 and −0.08, respectively [50-52].

Decile reliability plots under the calibrated background-only scenario showed close agreement between predicted and observed adoption probabilities (Figure 2) [50-52]. We analyzed 2405 individuals for vaccination and 2400 for screening (5 screening nonresponses excluded). Per-class calibration summaries are reported in Table S2 in Multimedia Appendix 4.

Figure 2. Decile reliability plots comparing mean predicted adoption probabilities with observed adoption rates for vaccination (A) and screening (B) under the calibrated background-only scenario.

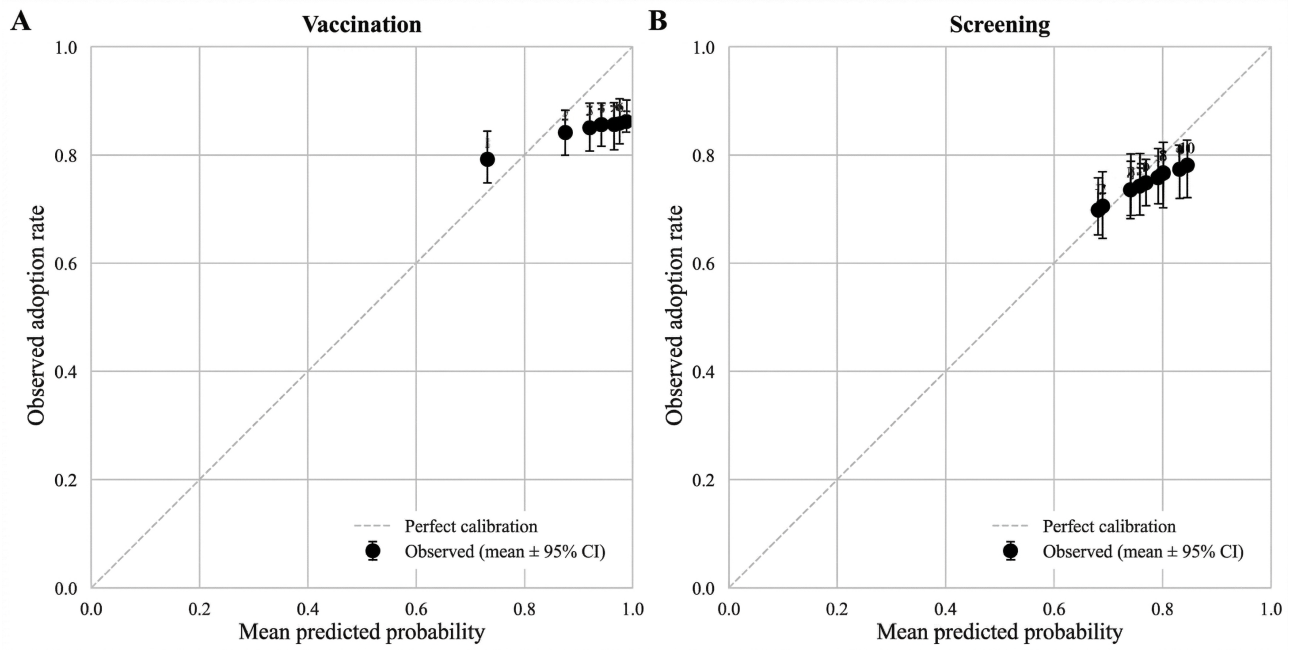

Channel Allocation and Budget (RQ1-RQ2)

For RQ1, we compared each single-channel campaign (B-D) with the background-only baseline (A). The TV-only campaign (B) produced the largest gains. For vaccination (N=2405), final adoption increased from 85% (n=2044) in A to 94.3% (n=2268) in B (9.3 percentage points, 95% PI 7.3 - 11.0). Mean time to adoption decreased from 6.18 (SD 0.07) to 5.34 (SD 0.06) months (0.84 months faster, 95% PI 0.65 - 0.94). For screening (N=2400), adoption increased from 74.5% (n=1788) to 89.4% (n=2146; 14.9 percentage points, 95% PI 12.9 - 17.0). Mean time to adoption decreased from 6.88 (SD 0.07) to 6.09 (SD 0.07) months (0.79 months faster, 95% PI 0.65 - 0.94). Digital-only and print-only campaigns produced smaller gains. For vaccination, the digital-only and print-only campaigns increased adoption by about 139 adopters (5.8 percentage points) and about 17 adopters (0.7 percentage points), respectively. For screening, the corresponding gains were about 199 adopters (8.3 percentage points) and 29 adopters (1.2 percentage points).

For RQ2, we held the per-step media budget constant and compared mixed-channel strategies (E-H) with the TV-only benchmark (B). All mixed strategies (TV + digital, TV + print, digital + print, and all channels) improved upon the background-only baseline for both vaccination and screening; however, none surpassed the TV-only campaign in either final adoption or time to adoption (Figure 3). For the TV + digital mix (E), the synergy index (difference in adoption between the mixed strategy and the best single-channel strategy) was slightly negative: −1.2 percentage points (95% PI −3.6 to 1.2 percentage points) for vaccination and −2.3 percentage points (95% PI −5.8 to 1.3 percentage points) for screening, corresponding to a dilution of roughly 1 to 2 percentage points relative to the best single-channel option. Other mixed strategies showed similarly nonpositive synergy. Tables S1 and S2 in Multimedia Appendix 2 provide full scenario-level adoption and equity metrics for all campaigns (A-O).

XSL•FO
RenderX

Figure 3. Final adoption and time-to-adoption outcomes across scenarios A to H for vaccination (A) and screening (B). Scenarios: A, control; B, television (TV)-only; C, digital-only; D, print-only; E, TV + digital; F, TV + print; G, digital + print; H, all channels; I, low-budget; J, high-budget; K, equity-focused reweighting; L, class-tailored channel portfolios.

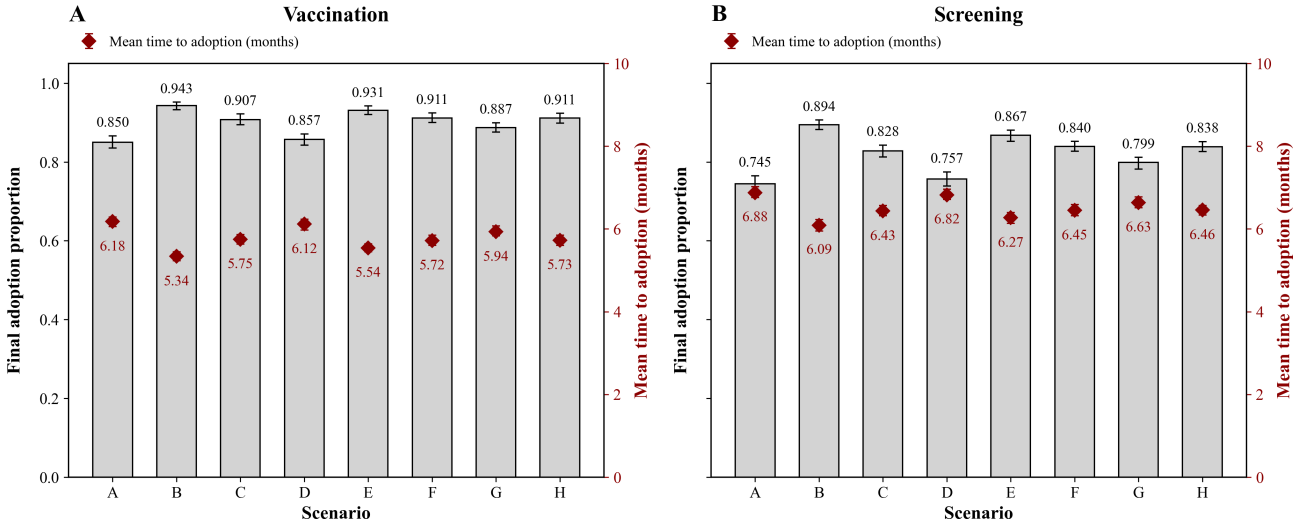

Personalization: Efficiency and Equity (RQ3-RQ4)

We compared the integrated baseline (H) with 2 personalization strategies: equity-focused reweighting (K) and class-tailored channel portfolios (L). For vaccination (N=2405), mean adoption increased from 91.2% (n=2193) in H to 93.3% (n=2244) in K and 94.6% (n=2275) in L. Furthermore, A_{min} increased from 86.8% (H) to 90.3% (K) and 90.9% (L). The 90 - 10 gap narrowed from 5.7 percentage points in H to 4.5 and 4.7 percentage points in K and L, respectively. For screening (N=2400), mean adoption increased from 83.8% (n=2011) in H to 88.2% (n=2117) in K and 89.5% (n=2148) in L. Moreover, A_{min} increased from 77.6% (H) to 83.2% (K) and 85.3% (L). The 90 - 10 gap narrowed from 9.2 percentage points to 7.4 and 6.2 percentage points in K and L, respectively.

Across both targets, strategy H had the lowest mean adoption and A_{min}. Strategy L achieved the highest mean adoption and A_{min}, and both personalization strategies (K and L) reduced the between-class disparities relative to H.

Table 3 summarizes the mean adoption and key equity metrics for scenarios H, K, and L. Figure 4 shows the resulting efficiency-equity trade-offs. Tables S1 and S2 in Multimedia Appendix 2 report the additional inequality indices and scenario-level adoption and equity metrics for all campaigns (A-O), including the TV-only campaign (B), low- and high-budget variants (I-J), and framing variants (M-O).

Table . Mean adoption and key equity metrics for scenarios H (all channels), K (equity-focused reweighting), and L (class-tailored channel portfolios).

Scenario	Mean adoption, n (%)	A_{min}, %	90 - 10 gap, percentage points
Vaccination (N=2405)			
H	2193 (91.2)	86.8	5.7
K	2244 (93.3)	90.3	4.5
L	2275 (94.6)	90.9	4.7
Screening (N=2400)			
H	2011 (83.8)	77.6	9.2
K	2117 (88.2)	83.2	7.4
L	2148 (89.5)	85.3	6.2

XSL•FO
RenderX

Figure 4. Efficiency-equity trade-offs for integrated baseline and personalized campaigns (H-L) for vaccination (A) and screening (B).

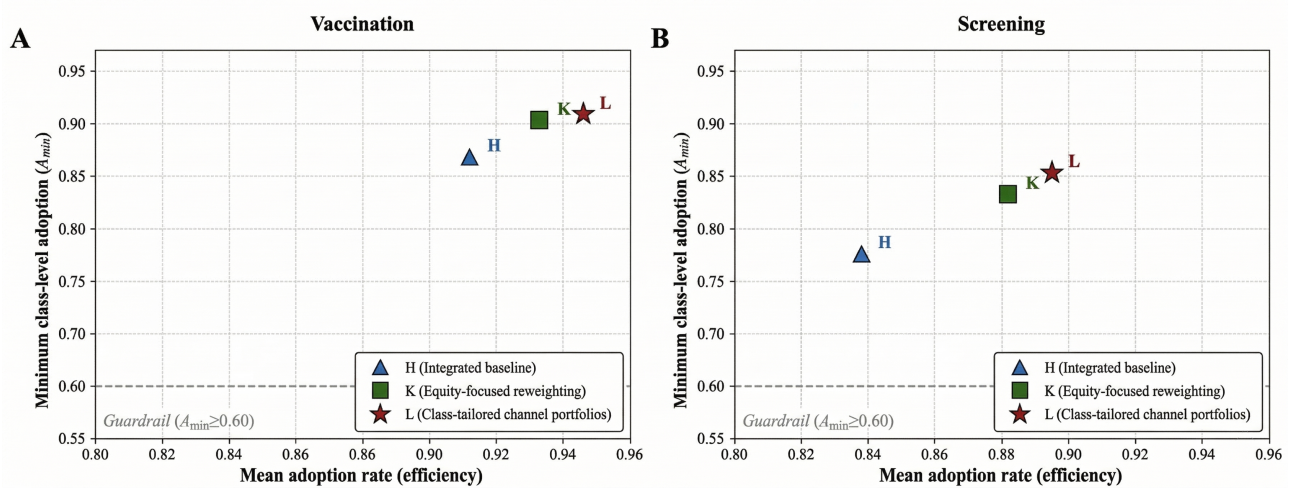

Message Framing (RQ5)

Averaged across scenarios M to O, increasing loss framing intensity (k_{neg}) produced modest, monotonic gains. For vaccination (N=2405), raising k_{neg} from 0.95 to 1.10 increased adoption from 92.3% (n=2220) to 92.9% (n=2234), a gain of about 14 adopters (0.6 percentage points). Mean time to adoption decreased from 5.62 (SD 0.16) to 5.54 (SD 0.17) months. For screening (N=2400), raising k_{neg} from 1.05 to 1.20 increased adoption from 88.8% (n=2131) to 89.7% (n=2153), a gain of about 22 adopters (0.9 percentage points). Mean time to adoption decreased from 6.01 (SD 0.16) to 5.93 (SD 0.18) months.

In the television-based screening scenario (M; N=2400), increasing k_{neg} from 1.05 to 1.20 increased adoption from 91.4% (n=2194) to 92.3% (n=2215), with parallel reductions in time to adoption. Digital-only screening (N) showed lower overall adoption but similar incremental gains across the same k_{neg} range. Gains between adjacent intensity levels were small (≤0.5 percentage points) and tended to taper at higher k_{neg}. The 95% percentile intervals overlapped, and we observed no evidence of backfiring within the tested range. Detailed dose-response curves and class-level heterogeneity by scenario and target are reported in Table S3 and Figure S1 in Multimedia Appendix 3.

Sensitivity Analyses

Using the prespecified stability benchmarks (ρ≥0.90 and Δrank≤1) as heuristics, scenario ordering for mean adoption was stable across prespecified axes (link function, weight allocation, reinforcement threshold, and network uncertainty). Maximum rank shifts were small (Δrank≤2), with the social-centric weight allocation variant for screening showing Δrank=2 while preserving the qualitative ordering. Full protocols and stability statistics are reported in Table S1 in Multimedia Appendix 3.

We changed the social reinforcement threshold from τ=2 (≥2 adopting neighbors) to τ=1 (≥1 adopting neighbor) and regenerated networks with alternative random seeds. These modifications left the scenario ordering unchanged for both tasks (ρ=1.00; Δrank=0). Alternative weight allocation variants produced only minor shifts for screening (ρ as low as 0.90; Δrank up to 2) while preserving the qualitative ordering. Comparisons between logistic and clipped-linear link functions preserved these rankings for mean adoption in both tasks.

Discussion

Principal Findings

Across the 15 prespecified scenarios, 5 patterns summarized how campaign design choices shaped efficiency and equity among older adults. These patterns were similar for vaccination and screening and remained stable across sensitivity analyses.

First, efficiency was driven mainly by channel allocation. The TV-only strategy (B) produced some of the highest mean adoption rates under a fixed per-step budget. Mixed-channel portfolios (E-H) improved outcomes relative to the background-only control (A) but did not surpass B, suggesting that dividing limited exposure across channels diluted reinforcement within high-TV–affinity clusters. Under fixed budgets, a single strong channel can therefore outperform intuitive “balanced mix” strategies.

Second, personalization shifted the portfolio toward a more favorable efficiency-equity trade-off within the modeled scenario space. The equity-focused reweighting intervention (K) produced modest gains in mean adoption but larger improvements in A_{min} and the 90 - 10 gap by lifting the lowest-performing class. The class-tailored strategy (L) showed a strong combined profile under fixed budgets. For vaccination, it matched the efficiency of B while improving equity. For screening, it achieved adoption comparable to B with narrower disparities. Taken together, these results suggest that class-tailored allocations can improve equity without necessarily reducing effectiveness. Under the modeled assumptions, they may also yield simultaneous improvements relative to uniform integrated mixes.

Third, high-budget expansion (J) increased average adoption and narrowed disparities, but it required higher total exposure. In contrast, K and L met guardrail criteria while operating under fixed budgets, indicating that reallocating exposure can achieve equity gains without increasing total exposure.

Fourth, framing acted as a secondary adjustment rather than a primary lever. Stronger loss framing produced small, monotonic

XSL•FO
RenderX

improvements (<1 percentage point) in both tasks and showed no evidence of backfiring within the tested range. However, these increments were minor compared with differences driven by channel allocation and personalization [45,46].

Finally, these strategic patterns were robust. Scenario orderings were stable across alternative reinforcement thresholds, weight allocations, and network realizations. Results under logistic versus clipped-linear link functions showed the same qualitative hierarchy.

Comparison With Prior Work

These findings relate to 4 strands of prior work: equity-focused implementation frameworks, media campaign evaluations, personalization in digital health, and ABMs of preventive behaviors.

First, implementation science frameworks such as the PRECEDE-PROCEED model and RE-AIM/PRISM increasingly emphasize equity [13,16]. However, they offer limited operational guidance on how to allocate budgets across channels and segments under explicit constraints. Most evaluations still describe who was reached and how effects varied after implementation [15,17]. Our model instead uses equity metrics as prespecified guardrails that screen and flag campaign portfolios once equity thresholds are violated. In doing so, it treats equity as an ex ante decision rule for digital health strategic planning rather than a purely descriptive endpoint.

Second, media campaign studies usually examine single-channel effects or short-term digital engagement metrics. Television is known to be effective for older adults, and digital channels offer heterogeneous reach, but few studies compare fixed-budget channel portfolios [18,58,59]. Existing work rarely tests whether multisource exposure creates synergy or whether splitting limited resources across channels weakens reinforcement. Our findings show that, under complex contagion dynamics, mixed strategies at fixed budgets do not outperform a strong single-channel allocation, helping to clarify a long-standing ambiguity in the campaign literature.

Third, personalization is often promoted as a way to improve relevance and engagement, while many authors warn that tailoring may widen inequalities by favoring digitally advantaged groups [60,61]. Reviews document these risks but seldom compare different personalization philosophies under common efficiency and equity metrics [60,61]. By contrasting equity-focused reweighting (K) with class-tailored channel portfolios (L), we show that personalization need not exacerbate disparities. When designed to lift the lowest-performing class or align exposure with segment-specific media affinities, personalization can improve both mean adoption and distributional equity.

Fourth, most ABMs of preventive behaviors have focused on disease transmission or generic opinion dynamics [28,32]. Few models combine nationally representative microdata, latent class segmentation, multilayer networks, and prespecified equity constraints within a single portfolio-testing framework [28,32]. In this study, the ABM links empirical population heterogeneity with mechanistic diffusion dynamics, allowing planners to test channel mixes, reallocations, and framing intensities under consistent, reproducible conditions.

Taken together, these features define a budget-constrained, equity-aware decision framework for preventive campaigns for older adults. This framework helps explain why some channel and personalization strategies are more favorable than others on both efficiency and equity within the modeled scenario space.

Implications for Digital Health Strategic Planning

These results are intended to support comparative decision-making for campaign design rather than point forecasting. Planners can use the model to compare candidate portfolios that vary channel mix and targeting under shared budget and equity constraints. To operationalize the scenario inputs, the relative exposure weights can be interpreted as the intended distribution of campaign effort across channels, and these weights can be translated into proportional budget shares or delivery targets using standard planning metrics (eg, television reach and frequency; digital reach and impressions; print distribution volume). The outputs provide comparative evidence for selecting among portfolios, with efficiency and equity evaluated jointly using the study's equity metrics and guardrails.

These findings suggest 3 priorities for public health planners working under budget constraints.

First, channel allocation should be treated as the primary strategic lever. Across both vaccination and screening, TV-only allocations produced some of the highest mean adoption levels under fixed budgets. Mixed portfolios improved outcomes but diluted reinforcement compared with television alone. For older adult populations with strong offline clustering, concentrating resources on a high-reach channel may therefore be more effective than spreading budgets thinly across multiple platforms.

Second, equity-oriented reallocation is both feasible and effective. The equity-focused reweighting strategy (K) raised minimum class-level adoption and narrowed class disparities with only small changes in mean outcomes. This shows that modest, budget-neutral shifts can materially improve equity. The class-tailored strategy (L) further improved both mean and distributional outcomes, indicating that segmentation-based resource allocation need not exacerbate inequities. These results suggest that agencies with basic segmentation data and media-buying capacity could adopt class-tailored portfolios to achieve efficiency-equity gains without increasing total exposure budgets.

Third, loss framing yielded only minor incremental gains within the tested range, so it should follow—rather than replace—allocation and targeting decisions.

Together, these implications position equity guardrails as a practical planning benchmark. Planners can start from an integrated channel mix and deprioritize strategies that fall below minimum subgroup adoption or widen disparities. Among the remaining options, they can consider conservative reallocations or class-tailored portfolios—such as scenarios K and L— that

XSL•FO
RenderX

show a more favorable efficiency-equity trade-off while maintaining strong mean adoption.

Limitations

Several limitations should be noted.

First, the ABM relied on structural assumptions about media influence, baseline propensity, and network topology. These assumptions simplify complex behavioral processes and may omit unmeasured factors. Calibration matched survey targets, and scenario rankings were stable across sensitivity checks. However, some model misspecification may remain.

Second, offline and online networks were approximated using survey-derived degree targets and a Watts-Strogatz small-world structure. Real-world networks may differ in clustering, multiplexity, tie strength, and mixing. These approximations may overestimate or underestimate the strength of social influence. They may also distort within-group versus between-group reinforcement. Rank stability across alternative network realizations suggests that the qualitative ordering of strategies is unlikely to depend on a single specification. However, more detailed network data could refine these structures. We did not test alternative class structures, and the sensitivity of scenario rankings to segmentation choices remains a topic for future work. Segmentation was cross-sectional, so we could not test temporal stability or class transitions.

Third, the model focused on exposure and diffusion rather than message content and quality, source trustworthiness, or system-level access barriers. Reactance, stigma, and message fatigue were not modeled, and strong fear-based or stigmatizing loss frames fell outside the tested k_{neg} range. Modest loss framing did not show evidence of backfiring within the simulated range, but these results should not be generalized to more aggressive messaging strategies.

Fourth, the model evaluated a 1-year campaign horizon and did not capture longer-term maintenance, multiyear cycles, or downstream health outcomes. Adoption was modeled as a single-step absorbing behavior, whereas real-world screening and vaccination involve repeated decisions and system-level interactions. Budget was represented as relative exposure intensity rather than monetary spending. Therefore, the results do not directly translate to monetary costs or cost patterns across media markets.

Finally, the analysis is based on matched Korean survey data. Several contextual features—such as high television reach, strong offline clustering, and a pronounced digital divide—may be specific to this setting. Applying the framework elsewhere would likely require recalibrating context-dependent inputs, such as channel reach patterns, the magnitude of the digital divide, and network mixing or clustering. In contrast, the model's core mechanisms—heterogeneous media affinity and reinforcement-based diffusion through social ties—may be more portable. Their magnitudes, however, may vary across settings. External validation in additional settings will be important for assessing transportability and for refining model-based decision tools.

Conclusions

This study used a data-driven ABM to examine how channel allocation, personalization, and loss framing shape preventive campaign performance for older adults under fixed budgets. In this survey-calibrated simulation, equity-oriented personalization of television and digital exposure improved subgroup equity and increased mean adoption relative to uniform integrated mixes and simple channel diversification. We evaluated scenarios using prespecified, illustrative equity guardrails. These patterns suggest that accounting for diffusion heterogeneity and reinforcement can help identify portfolios that strengthen equity without necessarily reducing overall performance.

For practice, an integrated channel mix can serve as a pragmatic starting point. Equity guardrails can then help identify portfolios that erode minimum subgroup adoption or widen disparities. Among the strategies that pass these guardrails, conservative reallocations can offer budget-neutral equity gains. Class-tailored portfolios may provide a more favorable efficiency-equity profile where segmentation and media-buying capacity are available. These conclusions apply primarily to settings with pronounced media heterogeneity and digital divides. They should be interpreted in light of the model's structural assumptions, 1-year horizon, and reliance on matched Korean survey data. Future work should validate these strategic patterns in other health systems and link simulated diffusion to observed exposure and engagement in real campaigns. It should also extend guardrail-based planning tools to richer communication, organizational, and multiyear decision contexts.

Funding

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (grant 2019-0-01842, Artificial Intelligence Graduate School Program [GIST]).

Data Availability

Analysis code and minimal deidentified outputs sufficient to reproduce all results—including scenario definitions, calibrated parameters, and sensitivity configurations as reported in the Multimedia Appendices—are available on the Open Science Framework [62]. Original microdata from KMP 2022 and KNHANES 2022 [39,40] cannot be shared due to data use restrictions but are accessible directly from the data providers.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Additional methods.
[DOCX File, 35 KB - jmir_v28i1e88429_app1.docx]

Multimedia Appendix 2
Scenario-level adoption and equity metrics.
[DOCX File, 37 KB - jmir_v28i1e88429_app2.docx]

Multimedia Appendix 3
Sensitivity and robustness.
[DOCX File, 501 KB - jmir_v28i1e88429_app3.docx]

Multimedia Appendix 4
Calibration and validation diagnostics.
[DOCX File, 32 KB - jmir_v28i1e88429_app4.docx]

References

1. Jemna DV, David M, Bonnal L, Oros C. Socio-economic inequalities in the use of flu vaccination in Europe: a multilevel approach. Health Econ Rev 2024 Jul 31;14(1):61. [doi: 10.1186/s13561-024-00535-1] [Medline: 39083186]
2. Wang J, Tönnies T, Brinks R. Seasonal influenza vaccination coverage and the social determinants of influenza vaccination among people over 50 with diabetes in Europe: analyzing population-based SHARE data for the 2019-2020 and 2021-2022 influenza seasons. Vaccine 2025 Jan 25;45:126646. [doi: 10.1016/j.vaccine.2024.126646] [Medline: 39709804]
3. Altová A, Kulhánová I, Reisser K, et al. Educational inequalities in cervical cancer screening participation in 24 European countries. Public Health 2024 Aug;233:1-7. [doi: 10.1016/j.puhe.2024.04.036] [Medline: 38805834]
4. Willems B, Bracke P. The education gradient in cancer screening participation: a consistent phenomenon across Europe? Int J Public Health 2018 Jan;63(1):93-103. [doi: 10.1007/s00038-017-1045-7] [Medline: 29063122]
5. Jiang X, Wang L, Leng Y, et al. The level of electronic health literacy among older adults: a systematic review and meta-analysis. Arch Public Health 2024 Nov 7;82(1):204. [doi: 10.1186/s13690-024-01428-9] [Medline: 39511667]
6. Tsai YIP, Beh J, Ganderton C, Pranata A. Digital interventions for healthy ageing and cognitive health in older adults: a systematic review of mixed method studies and meta-analysis. BMC Geriatr 2024 Mar 4;24(1):217. [doi: 10.1186/s12877-023-04617-3] [Medline: 38438870]
7. Zhang C, Mohamad E, Azlan AA, Wu A, Ma Y, Qi Y. Social media and eHealth literacy among older adults: systematic literature review. J Med Internet Res 2025 Mar 26;27:e66058. [doi: 10.2196/66058] [Medline: 40138684]
8. Yang R, Gao S, Jiang Y. Digital divide as a determinant of health in the US older adults: prevalence, trends, and risk factors. BMC Geriatr 2024 Dec 21;24(1):1027. [doi: 10.1186/s12877-024-05612-y] [Medline: 39709341]
9. Xie L, Zhang S, Xin M, Zhu M, Lu W, Mo PKH. Electronic health literacy and health-related outcomes among older adults: a systematic review. Prev Med 2022 Apr;157:106997. [doi: 10.1016/j.ypmed.2022.106997] [Medline: 35189203]
10. Badr J, Motulsky A, Denis JL. Digital health technologies and inequalities: a scoping review of potential impacts and policy recommendations. Health Policy 2024 Aug;146:105122. [doi: 10.1016/j.healthpol.2024.105122] [Medline: 38986333]
11. Zhang E, Shang S, Xing Y, et al. Mapping behavioral and social drivers of influenza vaccine uptake in older adults: a scoping review. Vaccines (Basel) 2025 Jun 10;13(6):624. [doi: 10.3390/vaccines13060624] [Medline: 40573955]
12. Behavioural and Social Drivers of Vaccination: Tools and Practical Guidance for Achieving High Uptake: World Health Organization; 2022. URL: https://iris.who.int/server/api/core/bitstreams/c9f4a77a-001e-477d-a9d1-b05832820b64/content [accessed 2026-03-20]
13. Holtrop JS, Estabrooks PA, Gaglio B, et al. Understanding and applying the RE-AIM framework: clarifications and resources. J Clin Transl Sci 2021;5(1):e126. [doi: 10.1017/cts.2021.789] [Medline: 34367671]
14. Aravindhan A, Fenwick E, Chan AWD, et al. Nonadherence to diabetes complications screening in a multiethnic Asian population: protocol for a mixed methods prospective study. JMIR Res Protoc 2025 May 8;14:e63253. [doi: 10.2196/63253] [Medline: 40340926]
15. Hollands GJ, South E, Shemilt I, Oliver S, Thomas J, Sowden AJ. Methods used to conceptualize dimensions of health equity impacts of public health interventions in systematic reviews. J Clin Epidemiol 2024 May;169:111312. [doi: 10.1016/j.jclinepi.2024.111312] [Medline: 38432524]
16. Jolles MP, Fort MP, Glasgow RE. Aligning the planning, development, and implementation of complex interventions to local contexts with an equity focus: application of the PRISM/RE-AIM Framework. Int J Equity Health 2024 Feb 27;23(1):41. [doi: 10.1186/s12939-024-02130-6] [Medline: 38408990]
17. Welch V, Dewidar O, Ghogomu ET, et al. How effects on health equity are assessed in systematic reviews of interventions. Cochrane Database Syst Rev 2022 Jan 18;1(1):MR000028. [doi: 10.1002/14651858.MR000028.pub3] [Medline: 35040487]

XSL•FO
RenderX

18. Kite J, Chan L, MacKay K, et al. A model of social media effects in public health communication campaigns: systematic review. J Med Internet Res 2023 Jul 14;25:e46345. [doi: 10.2196/46345] [Medline: 37450325]
19. Zang S, Zhang X, Xing Y, Chen J, Lin L, Hou Z. Applications of social media and digital technologies in COVID-19 vaccination: scoping review. J Med Internet Res 2023 Feb 10;25:e40057. [doi: 10.2196/40057] [Medline: 36649235]
20. Massey PM, Murray RM, Chiang SC, Russell AM, Yudell MA. Social media, public health research, and vulnerability: considerations to advance ethical guidelines and strengthen future research. JMIR Public Health Surveill 2023 Dec 29;9:e49881. [doi: 10.2196/49881] [Medline: 38157235]
21. Alruwaili MM, Shaban M, Elsayed Ramadan OM. Digital health interventions for promoting healthy aging: a systematic review of adoption patterns, efficacy, and user experience. Sustainability 2023;15(23):16503. [doi: 10.3390/su152316503]
22. Hwang M, Lee S, Park GE, Park YH. Effectiveness of a digital health coaching self-management program for older adults living alone with multiple chronic conditions: a randomized controlled trial. Geriatr Nurs 2025;65:103509. [doi: 10.1016/j.gerinurse.2025.103509] [Medline: 40633238]
23. Kim KK, Backonja U. Digital health equity frameworks and key concepts: a scoping review. J Am Med Inform Assoc 2025 May 1;32(5):932-944. [doi: 10.1093/jamia/ocaf017] [Medline: 39936843]
24. Bitomsky L, Pfitzer E, Nißen M, Kowatsch T. Advancing health equity and the role of digital health technologies: a scoping review. BMJ Open 2025 Jun 26;15(6):e099306. [doi: 10.1136/bmjopen-2025-099306] [Medline: 40578855]
25. Karran EL, Cashin AG, Barker T, et al. Using PROGRESS-plus to identify current approaches to the collection and reporting of equity-relevant data: a scoping review. J Clin Epidemiol 2023 Nov;163:70-78. [doi: 10.1016/j.jclinepi.2023.09.017] [Medline: 37802205]
26. Anand S, World Health Organization. Measuring Health Workforce Inequalities: Methods and Application to China and India: World Health Organization; 2010. URL: https://iris.who.int/server/api/core/bitstreams/a2a1d045-37ab-46d2-b182-e80de35d52cd/content [accessed 2026-03-22]
27. Jiang S, Li B, Parkinson B, Li S, Gu Y. Aggregate distributional cost-effectiveness analysis: a novel tool for health economic evaluation to inform resource allocation. Glob Health Res Policy 2025 Apr 16;10(1):17. [doi: 10.1186/s41256-025-00415-z] [Medline: 40241192]
28. Colasanti R, MacLachlan A, Silverman E, et al. Using agent-based models to address non-communicable diseases: a review of models and their application to policy. The Lancet 2022 Nov;400:S33. [doi: 10.1016/S0140-6736(22)02243-7]
29. Sivakumar N, Mura C, Peirce SM. Innovations in integrating machine learning and agent-based modeling of biomedical systems. Front Syst Biol 2022;2:959665. [doi: 10.3389/fsysb.2022.959665]
30. Centola D. How Behavior Spreads: The Science of Complex Contagions: Princeton University Press; 2018. [doi: 10.2307/j.ctvc7758p]
31. Lindau ST, Makelarski JA, Kaligotla C, et al. Building and experimenting with an agent-based model to study the population-level impact of CommunityRx, a clinic-based community resource referral intervention. PLoS Comput Biol 2021 Oct;17(10):e1009471. [doi: 10.1371/journal.pcbi.1009471] [Medline: 34695116]
32. Sulis E, Terna P. An agent-based decision support for a vaccination campaign. J Med Syst 2021 Sep 28;45(11):97. [doi: 10.1007/s10916-021-01772-1] [Medline: 34581878]
33. Oană I, Hâncean MG, Perc M, et al. Online media use and COVID-19 vaccination in real-world personal networks: quantitative study. J Med Internet Res 2024 Oct 25;26:e58257. [doi: 10.2196/58257] [Medline: 39454189]
34. Centola D. The spread of behavior in an online social network experiment. Science 2010 Sep 3;329(5996):1194-1197. [doi: 10.1126/science.1185231] [Medline: 20813952]
35. Oh SS, Kim KA, Kim M, Oh J, Chu SH, Choi J. Measurement of digital literacy among older adults: systematic review. J Med Internet Res 2021 Feb 3;23(2):e26145. [doi: 10.2196/26145] [Medline: 33533727]
36. Wilensky U, Rand W. An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo: MIT Press; 2015.
37. Railsback SF, Grimm V. Agent-Based and Individual-Based Modeling: A Practical Introduction: Princeton University Press; 2019. [doi: 10.2307/jj.28274141]
38. Boccaletti S, Bianconi G, Criado R, et al. The structure and dynamics of multilayer networks. Phys Rep 2014 Nov 1;544(1):1-122. [doi: 10.1016/j.physrep.2014.07.001] [Medline: 32834429]
39. 2022 Korea Media Panel Survey. Korea Information Society Development Institute. 2022. URL: https://www.kisdi.re.kr/report/view.do?key=m2101113024973&arrMasterId=3934581&masterId=3934581&artId=1131416 [accessed 2026-02-18]
40. 2022 Korea National Health and Nutrition Examination Survey. Korea Disease Control and Prevention Agency. 2023. URL: https://knhanes.kdca.go.kr/knhanes/main.do [accessed 2026-03-22]
41. Gower JC. A general coefficient of similarity and some of its properties. Biometrics 1971 Dec;27(4):857-871. [doi: 10.2307/2528823]
42. Collins LM, Lanza ST. Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences: John Wiley & Sons; 2010. [doi: 10.1002/9780470567333]
43. Atkinson AB. On the measurement of inequality. J Econ Theory 1970 Sep;2(3):244-263. [doi: 10.1016/0022-0531(70)90039-6]

XSL•FO
RenderX

44. Borrell LN, Talih M. A symmetrized Theil index measure of health disparities: an example using dental caries in US children and adolescents. Stat Med 2011 Feb 10;30(3):277-290. [doi: 10.1002/sim.4114] [Medline: 21213344]
45. Gallagher KM, Updegraff JA. Health message framing effects on attitudes, intentions, and behavior: a meta-analytic review. Ann Behav Med 2012 Feb;43(1):101-116. [doi: 10.1007/s12160-011-9308-7] [Medline: 21993844]
46. Ainiwaer A, Zhang S, Ainiwaer X, Ma F. Effects of message framing on cancer prevention and detection behaviors, intentions, and attitudes: systematic review and meta-analysis. J Med Internet Res 2021 Sep 16;23(9):e27634. [doi: 10.2196/27634] [Medline: 34528887]
47. Fosdick BK, Larremore DB, Nishimura J, Ugander J. Configuring random graph models with fixed degree sequences. SIAM Rev 2018 Jan;60(2):315-355. [doi: 10.1137/16M1087175]
48. Newman MEJ, Strogatz SH, Watts DJ. Random graphs with arbitrary degree distributions and their applications. Phys Rev E 2001 Jul;64(2):026118. [doi: 10.1103/PhysRevE.64.026118]
49. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature New Biol 1998 Jun 4;393(6684):440-442. [doi: 10.1038/30918] [Medline: 9623998]
50. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev 1950 Jan;78(1):1-3. [doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2]
51. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning: Association for Computing Machinery; 2005:625-632. [doi: 10.1145/1102351.1102430]
52. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 2007 Mar;102(477):359-378. [doi: 10.1198/016214506000001437]
53. Law AM, Kelton WD. Simulation Modeling and Analysis, 4th edition: McGraw-Hill; 2007.
54. Grimm V, Berger U, DeAngelis DL, Polhill JG, Giske J, Railsback SF. The ODD protocol: A review and first update. Ecol Model 2010 Nov;221(23):2760-2768. [doi: 10.1016/j.ecolmodel.2010.08.019]
55. Saltelli A, Ratto M, Andres T, et al. Global Sensitivity Analysis: The Primer: John Wiley & Sons; 2007.
56. Bioethics and Safety Act. Ministry of Government Legislation (Korea). URL: https://www.law.go.kr/LSW/lsInfoP.do?lsiSeq=199534&viewCls=engLsInfoR&urlMode=engLsInfoR#0000 [accessed 2026-03-27]
57. Enforcement Rules of the Act on Bioethics and Safety. Ministry of Government Legislation (Korea). URL: https://www.law.go.kr/%EB%B2%95%EB%A0%B9/%EC%83%9D%EB%AA%85%EC%9C%A4%EB%A6%AC%EB%B0%8F%EC%95%88%EC%A0%84%EC%97%90%EA%B4%80%ED%95%9C%EB%B2%95%EB%A5%A0%EC%8B%9C%ED%96%89%EA%B7%9C%EC%B9%99 [accessed 2026-03-27]
58. Eppes EV, Augustyn M, Gross SM, Vernon P, Caulfield LE, Paige DM. Engagement with and acceptability of digital media platforms for use in improving health behaviors among vulnerable families: systematic review. J Med Internet Res 2023 Feb 3;25:e40934. [doi: 10.2196/40934] [Medline: 36735286]
59. De Santis KK, Mergenthal L, Christianson L, Busskamp A, Vonstein C, Zeeb H. Digital technologies for health promotion and disease prevention in older people: scoping review. J Med Internet Res 2023 Mar 23;25:e43542. [doi: 10.2196/43542] [Medline: 36951896]
60. Klooster IT, Kip H, van Gemert-Pijnen L, Crutzen R, Kelders S. A systematic review on eHealth technology personalization approaches. iScience 2024 Sep 20;27(9):110771. [doi: 10.1016/j.isci.2024.110771] [Medline: 39290843]
61. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018 Aug 1;25(8):1080-1088. [doi: 10.1093/jamia/ocy052] [Medline: 29788380]
62. Project repository for “Channel allocation and equity in preventive campaigns for older adults: agent-based modeling study". Open Science Framework. URL: https://osf.io/fnv57 [accessed 2026-02-21]

Abbreviations

ABM: agent-based model
AIC: Akaike information criterion
BIC: Bayesian information criterion
KMP: Korea Media Panel Survey
KNHANES: Korea National Health and Nutrition Examination Survey
PI: percentile interval
PRISM: Practical Robust Implementation and Sustainability Model
PROGRESS-Plus: place of residence; race, ethnicity, culture, and language; occupation; gender and sex; religion; education; socioeconomic status; and social capital—plus additional equity-relevant factors
RDI: relative disparity index
RE-AIM: Reach, Effectiveness, Adoption, Implementation and Maintenance
RQ: research question
TV: television

XSL•FO
RenderX

Edited by J Piera-Jiménez; submitted 25.Nov.2025; peer-reviewed by WG Hussain, Z Liu; accepted 25.Feb.2026; published 01.Apr.2026.

Please cite as:
Lee J, Park J, Kim Y, Kong DJ
Channel Allocation and Equity in Preventive Campaigns for Older Adults: Agent-Based Modeling Study
J Med Internet Res 2026;28:e88429
URL: https://www.jmir.org/2026/1/e88429
doi:10.2196/88429

© Jihye Lee, Juyoung Park, Yuna Kim, Duk-Jo Kong. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 1.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org

XSL•FO
RenderX